UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

PRONÓSTICO DE VENTAS DE PRODUCTOS EN LA INDUSTRIA DEL RETAIL EN BASE A SIMILITUD DE SERIES TEMPORALES

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

DIEGO IGNACIO SALAZAR LAZO

PROFESOR GUÍA:
JUAN MANUEL BARRIOS NÚÑEZ

PROFESOR CO-GUÍA:
DAVID VALENZUELA URRUTIA

MIEMBROS DE LA COMISIÓN:
NELSON BALOIAN TATARYAN
ADOLFO CARRASCO ACOSTA

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN
POR: DIEGO IGNACIO SALAZAR LAZO
FECHA: 2022
PROF. GUÍA: JUAN MANUEL BARRIOS NÚÑEZ
PROF. CO-GUÍA: DAVID VALENZUELA URRUTIA

PRONÓSTICO DE VENTAS DE PRODUCTOS EN LA INDUSTRIA DEL RETAIL EN BASE A SIMILITUD DE SERIES TEMPORALES

La industria del *retail* es una de las de mayor relevancia económica, tanto en Chile, como en el resto del mundo. De hecho, se estima que esta industria concentra aproximadamente el 20 % del PIB chileno. En este ámbito, las empresas se ven enfrentadas a tomar decisiones con respecto a cómo planificar eficientemente la distribución de los productos desde sus bodegas o centros de distribución hacia las tiendas donde finalmente se venden sus productos.

Elaborar una buena planificación de distribución está directamente relacionado con la necesidad de realizar una buena predicción de demanda de productos en sus tiendas. El riesgo de tener una distribución ineficiente se traduce en dos posibles escenarios subóptimos de ventas, en los cuales, por un lado se arriesga a no destinar suficientes productos a una tienda de mayor demanda, y por otro, se arriesga a destinar una mayor cantidad de productos por sobre lo demandado, teniendo que incurrir en ofertas y liquidaciones no planeadas. Ambos escenarios representan pérdidas sensibles para los *retailers* y es por este motivo que surge la necesidad de encontrar una mejor herramienta para la predicción de ventas, para mejorar la distribución de productos.

El presente trabajo busca investigar la posibilidad de realizar pronósticos de ventas en base a las similitudes en los comportamientos de ventas de los productos. Para lograr esto, se realizó una extensa exploración inicial de los datos y una posterior etapa de preprocesamiento, tras la cual se definió una heurística para simular los pronósticos de ventas de las series temporales que tuvieran el menor valor de la distancia Euclidiana, basado en predecir las ventas de un producto para la segunda mitad del año, a partir de las ventas de otro producto que presentara un comportamiento de venta similar en la primera mitad del año, medido por la métrica de distancia mencionada. Esta heurística fue replicada en los 4 *datasets* disponibles, Ropa Interior, Lavadoras, Pantalones y Microondas y fueron realizados tanto para una tienda en específico, como para la agrupación de las tiendas de una misma zona geográfica.

Se concluye que el modelo predictivo propuesto es más robusto en la medida que se tenga un *dataset* final con una variedad del orden de cientos de productos y sus datos históricos de ventas de al menos 3 años atrás. Se observa que las mejores predicciones fueron obtenidas para productos de la categoría de vestuario, con un MAE de 0.0101 para el par de series más cercanas. Por otro lado, los productos correspondientes a electrodomésticos presentaron resultados menos favorables para la heurística de predicción propuesta, con un MAE de 0.0562 para el par de series más cercanas.

Estos resultados permitieron sentar las bases para la creación de un modelo predictivo basado en la similitud de las series temporales de ventas.

"Lo que sabemos es una gota de agua; lo que ignoramos es un océano."
— Isaac Newton

Agradecimientos

Sin duda alguna, optar por el camino de la Ingeniería ha sido uno de los más arduos y gratificantes desafíos de mi vida, y que hoy, por fin, se acerca a su cierre (o a su nuevo comienzo). Ha sido un camino bastante largo, mucho más de lo que hubiera pensado cuando lo inicié, sin embargo, no hubiera sido capaz de llegar a este punto sin contar con la participación y ayuda de todas las personas que me acompañaron de una u otra forma en este proceso.

Quiero comenzar dándole gracias infinitas a mis padres, primero por transmitirme a través del milagro de la genética, el gusto por esta apasionante área de conocimiento que es la Ingeniería y las Ciencias de la Computación, y segundo por ser los mejores padres con los que hubiera podido contar en la vida. Quiero agradecerles por su amor incondicional, por ser siempre mi guía, mi apoyo, por transmitirme sus buenos valores, por educarme de la mejor forma posible. Gracias por quedarse trasnochando conmigo desde pequeño, para ayudarme a aprenderme las disertaciones, o para realizar los más sofisticados trabajos manuales. Gracias porque nunca me faltara nada, por darme todo lo que he querido o necesitado. Gracias por permitirme estudiar en la Universidad más prestigiosa del país y perdón por atrasarme tanto en la carrera. También quiero agradecer a mis abuelos, Julio, Violeta y Ester por haberme inculcado con sus propias historias de vida, el ímpetu por la superación personal y el trabajo duro para progresar en la vida. El esfuerzo colectivo de toda mi familia me ha dado la fortaleza para poder hoy en día concretar logros como este.

Quiero agradecer a Macarena Osorio, por haber llegado a mi vida y darme el amor más bello que pudiera imaginar. Gracias por haber estado conmigo en todas, por haberme hecho reír como nadie, por haber compartido tu vida conmigo y por haber formado parte de la mía. También quiero agradecerle por la gran guía y apoyo que me dio durante este proceso de titulación, por haberme acompañado todas esas noches trabajando hasta tarde, por los refrigerios que me enviaba para avanzar en mi trabajo y por siempre haberme motivado a dar lo mejor de mí, como estudiante, como profesional y como persona.

Quiero agradecer a mi gran amigo de toda la vida, Camilo Gallyas, por estar siempre ahí, de una u otra manera, cuando lo necesito para conversar o jugar o cuando quiero reír hasta llorar. Quiero agradecer a Camila Araya por todos estos años de amistad, por su lealtad y por su apoyo constante, por saber que siempre estará presente si la necesito. También quiero agradecer a grandes amistades que hice en la Universidad, y que hicieron que mi paso por esta fuera mucho más ameno. Gracias a Hans, a Rodrigo, a Tomás, a Francisco, a Diego, a Joaquín, a Lucas, a Eduardo, a Catalina, a Luis, a Jean Pierre, a Sebastián y por último al equipo iDream, con el cual compartí mis breves tiempos como Industrial. La vida adulta

puede habernos separado por uno u otro motivo, pero siempre quedarán en mis recuerdos las jornadas intensas de estudio, de trabajo, de risas y, por supuesto, de mucho alcohol.

Quiero agradecer a mis profesores guía, Juan Manuel y David, por su infinita paciencia a lo largo de este prolongado proceso. Por no haberse rendido conmigo, y haberme dado más oportunidades, aún cuando quizás hasta yo mismo pensé en retirarme. Gracias por tener fe en mí. Además, les agradezco mucho por haber sido de gran utilidad en la realización de este trabajo, a través de sus valiosas sugerencias y guías durante el desarrollo. Me siento afortunado por lo que logré aprender de ustedes y de este trabajo.

Por último, quiero agradecer a la Facultad de Ciencias Físicas y Matemáticas y a la Universidad de Chile, por la calidad de la educación entregada. Por la formación que dan, por los espacios de aprendizaje y de crecimiento que ofrecen y por todo el conocimiento que pude adquirir a lo largo de mi carrera universitaria.

Tabla de Contenido

1. Introducción **1**
- 1.1. Motivación . . . 1
  - 1.1.1. Contexto general . . . 1
  - 1.1.2. Problema a abordar . . . 2
  - 1.1.3. Impacto del Trabajo de Memoria . . . 4
- 1.2. Objetivos . . . 4
  - 1.2.1. Objetivo general . . . 4
  - 1.2.2. Objetivos específicos . . . 4
- 1.3. Solución . . . 5
  - 1.3.1. Hipótesis . . . 5
  - 1.3.2. Alternativas analizadas . . . 5
- 1.4. Estructura de la Memoria . . . 5

2. Marco Teórico y Estado del Arte **6**
- 2.1. Ciencia de Datos . . . 6
- 2.2. Series de tiempo . . . 8
- 2.3. *Time Series Smoothing* . . . 10
- 2.4. Medidas de distancia . . . 11
  - 2.4.1. Distancia Euclidiana . . . 11
  - 2.4.2. Dynamic Time Warping (DTW) . . . 12
  - 2.4.3. *Cosine Similarity* . . . 14
  - 2.4.4. Angular Metric for Shape Similarity (AMSS) . . . 15
  - 2.4.5. Otras medidas de distancia . . . 17
- 2.5. Métricas de desempeño . . . 19
  - 2.5.1. Mean Absolute Error (MAE) . . . 19
  - 2.5.2. Mean Absolute Percentage Error (MAPE) . . . 20
  - 2.5.3. Root Mean Squared Errors (RMSE) . . . 20

3. Predicción de Demanda **21**
- 3.1. Situación actual de "La Empresa" . . . 21
  - 3.1.1. Problema y oportunidad . . . 21
  - 3.1.2. Modelo predictivo vigente y sus defectos . . . 22
  - 3.1.3. Relevancia . . . 24
  - 3.1.4. Especificaciones y alcances de la solución . . . 24

4. Descripción de los Datos y Análisis Exploratorio **26**

4.1. Descripción de los *datasets* y sus atributos 26
4.2. Confidencialidad 28
4.3. Distribución de los datos y estadísticas descriptivas 28
4.4. Análisis por tiendas y zonas 34
4.4.1. Gráficos comparativos de los volúmenes de ventas en las zonas geográficas 34
4.4.2. Gráficos comparativos de los volúmenes de ventas por cada tienda . . 37
4.5. Series temporales de ventas y gráficos preliminares 41

5. Diseño y Desarrollo de la Solución 48
5.1. Metodología 48
5.1.1. Etapas del proceso 48
5.1.2. Heurística de predicción de ventas 49
5.1.3. Herramientas computacionales 49
5.2. Prueba de concepto 50
5.2.1. Objetivos 50
5.2.2. Resultados y observaciones 50
5.3. Preprocesamiento de los datos 55
5.4. Cálculo de similitud entre series temporales 58
5.5. Simulación de pronóstico de series temporales 58

6. Resultados y Análisis 60
6.1. Resultados e interpretación de los experimentos 60
6.1.1. Pronóstico de ventas en una tienda específica del *dataset* de Ropa Interior 61
6.1.2. Pronóstico de ventas en una zona geográfica del *dataset* de Ropa Interior 67
6.1.3. Pronóstico de ventas en una tienda específica del *dataset* de Lavadoras 71
6.1.4. Pronóstico de ventas en una zona geográfica del *dataset* de Lavadoras 77
6.2. Discusión 80
6.3. Validación 81

7. Conclusiones y Trabajo Futuro 82
7.1. Resumen del trabajo realizado 82
7.2. Revisión de objetivos y conclusiones 83
7.3. Trabajo futuro 84

Bibliografía 85

Anexos 89

A. Resultados de la simulación de pronóstico de ventas con el *dataset* de Pantalones 90
A.1. Experimentos realizados para una tienda específica 91
A.1.1. Resultados obtenidos 91
A.2. Experimentos realizados para una zona específica 92
A.2.1. Resultados obtenidos 92

B. Resultados de la simulación de pronóstico de ventas con el *dataset* de Microondas 93
B.1. Experimentos realizados para una tienda específica 94

B.1.1. Resultados obtenidos 94
B.2. Experimentos realizados para una zona específica 95
B.2.1. Resultados obtenidos 95

C. Nomenclaturas e información relevante **96**
C.1. Equivalencia entre semanas comerciales y rangos de fechas 96

Índice de Tablas

1.1. Problema de distribución de productos, caso subóptimo 2
1.2. Problema de distribución de productos, caso óptimo 3

4.1. Variables presentes en cada conjunto de datos utilizado 27

6.1. Distancias Euclidianas de los 3 pares de estilos de Ropa Interior más cercanos en la tienda 63
6.2. Comparación cualitativa de los 3 pares de estilos de Ropa Interior más cercanos en la tienda 64
6.3. Distancia Euclidiana del par de estilos de Ropa Interior más lejano en la tienda 65
6.4. Comparación cualitativa del par de estilos de Ropa Interior más lejano en la tienda 65
6.5. Cálculo del MAE en la simulación de pronóstico de ventas de Ropa Interior en la tienda 65
6.6. Euclidianas de los 3 pares de estilos de Lavadoras más cercanos en la tienda 73
6.7. Comparación cualitativa de los 3 pares de estilos de Lavadoras más cercanos en la tienda 74
6.8. Distancia Euclidiana del par de estilos de Lavadoras más lejano en la tienda 75
6.9. Comparación cualitativa del par de estilos de Lavadoras más lejano en la tienda 75
6.10. Cálculo del MAE en la simulación de pronóstico de ventas de Lavadoras en la tienda 75

C.1. Intervalos de fechas de cada semana de los años 2019, 2020 y 2021. 97

Índice de Ilustraciones

1.1. Termómetro semanal de ventas minoristas del año 2021[11]. 2

2.1. Diagrama de Venn de la interdisciplinariedad de la Ciencia de Datos[3]. . . . 7
2.2. *Pipeline* del desarrollo de proyectos de Ciencia de Datos[27]. 8
2.3. Ejemplo de series de tiempo regulares e irregulares[12]. 8
2.4. Comportamientos presentes en las series de tiempo[25]. 9
2.5. Ejemplo de suavizado de series temporales[9]. 11
2.6. Comparación de distancia Euclidiana y DTW para dos series de tiempo[6]. . 13
2.7. Ejemplo de *Cosine Similarity* aplicado a documentos de texto[26]. 14
2.8. Visualización de dos series de tiempo, C y Q, representadas como secuencias de vectores[22]. 15
2.9. Ángulo entre dos vectores pertenecientes a series de tiempo[23]. 16
2.10. Comparación de dos series de tiempo mediante la AMSS[23]. 17
2.11. Comparación de las distancias de Manhattan y Euclidiana para dos puntos[21]. 17
2.12. Ejemplo de *Minkowski Distance* con distintos valores de p[33]. 18

3.1. Ejemplo de jeans similares vendidos en temporadas distintas. 22
3.2. Esquema del modelo predictivo actual de "La Empresa". 23
3.3. Esquema del algoritmo *XGBoost* para el pronóstico de demanda de "La Empresa". 23

4.1. Esquema de jerarquías de los productos. 26
4.2. Gráfico comparativo del volumen de datos presente en cada *dataset* utilizado. 29
4.3. Gráfico comparativo de la distribución de estilos y SKUs únicos en los *datasets*. 30
4.4. Histogramas de precios de los *datasets* de vestuario. 31
4.5. Histogramas de precios de los *datasets* de electrodomésticos. 32
4.6. Diagrama de caja de los precios de los *datasets* de vestuario. 32
4.7. Diagrama de caja de los precios de los *datasets* de electrodomésticos. 33
4.8. Comparación de la distribución de ventas de Ropa Interior según cada zona geográfica. 34
4.9. Comparación de la distribución de ventas de Lavadoras según cada zona geográfica. 35
4.10. Comparación de la distribución de ventas de Pantalones según cada zona geográfica. 35
4.11. Comparación de la distribución de ventas de Microondas según cada zona geográfica. 36
4.12. Comparación del volumen de ventas en las tiendas de cada zona del *dataset* de Ropa Interior. 37

4.13. Comparación del volumen de ventas en las tiendas de cada zona del *dataset* de Lavadoras. 38
4.14. Comparación del volumen de ventas en las tiendas de cada zona del *dataset* de Pantalones. 39
4.15. Comparación del volumen de ventas en las tiendas de cada zona del *dataset* de Microondas. 40
4.16. Comparación de series temporales de ventas semanales porcentuales de Ropa Interior para los años 2019, 2020 y 2021. 42
4.17. Comparación de series temporales de ventas semanales porcentuales de Lavadoras para los años 2019, 2020 y 2021. 43
4.18. Comparación de series temporales de ventas semanales porcentuales de Pantalones para los años 2019, 2020 y 2021. 44
4.19. Comparación de series temporales de ventas semanales porcentuales de Microondas para los años 2019, 2020 y 2021. 45
4.20. Comparación de series temporales de ventas semanales porcentuales promedio en los *datasets* de vestuario y electrodomésticos. 46
4.21. Comparación de series temporales de ventas semanales porcentuales promedio en todos los *datasets*. 47

5.1. Gráfico de series temporales con ventas escaladas de pantalones en un contexto diario. 51
5.2. Gráfico de dos series temporales de ventas semanales visualmente parecidas en el *dataset* de Pantalones. 52
5.3. Comparación de matrices de distancias para algunas series temporales. . . . 53
5.4. Pareo de series temporales según distancia DTW. 53
5.5. Par de series temporales más cercanas según distancia DTW. 54
5.6. Ejemplo del formato final de los datos de un producto. 57

6.1. Primer par de series temporales más cercanas en el *dataset* de Ropa Interior en la tienda Arauco Maipú. 61
6.2. Segundo par de series temporales más cercanas en el *dataset* de Ropa Interior en la tienda Arauco Maipú. 62
6.3. Tercer par de series temporales más cercanas en el *dataset* de Ropa Interior en la tienda Arauco Maipú. 63
6.4. Par de series temporales más lejanas en el *dataset* de Ropa Interior en la tienda Arauco Maipú. 64
6.5. Evolución de la diferencia del MAE en las predicciones para el *dataset* de Ropa Interior en la tienda Arauco Maipú. 66
6.6. Primer par de series temporales más cercanas en el *dataset* de Ropa Interior en la zona Centro Poniente. 67
6.7. Segundo par de series temporales más cercanas en el *dataset* de Ropa Interior en la zona Centro Poniente. 68
6.8. Tercer par de series temporales más cercanas en el *dataset* de Ropa Interior en la zona Centro Poniente. 69
6.9. Evolución de la diferencia del MAE en las predicciones para el *dataset* de Ropa Interior en la zona Centro Poniente. 70

6.10. Primer par de series temporales más cercanas en el *dataset* de Lavadoras en la tienda Plaza Oeste. 71
6.11. Segundo par de series temporales más cercanas en el *dataset* de Lavadoras en la tienda Plaza Oeste. 72
6.12. Tercer par de series temporales más cercanas en el *dataset* de Lavadoras en la tienda Plaza Oeste. 73
6.13. Par de series temporales más lejanas en el *dataset* de Lavadoras en la tienda Plaza Oeste. 74
6.14. Evolución de la diferencia del MAE en las predicciones para el *dataset* de Lavadoras en la tienda Plaza Oeste. 76
6.15. Primer par de series temporales más cercanas en el *dataset* de Lavadoras en la zona Centro Poniente. 77
6.16. Segundo par de series temporales más cercanas en el *dataset* de Lavadoras en la zona Centro Poniente. 78
6.17. Tercer par de series temporales más cercanas en el *dataset* de Lavadoras en la zona Centro Poniente. 79
6.18. Evolución de la diferencia del MAE en las predicciones para el *dataset* de Lavadoras en la zona Centro Poniente. 80

A.1. Pares de series temporales más cercanas en el *dataset* de Pantalones en la tienda Plaza Vespucio. 91
A.2. Evolución de la diferencia del MAE en las predicciones para el *dataset* de Pantalones en la tienda Plaza Vespucio. 91
A.3. Pares de series temporales más cercanas en el *dataset* de Pantalones en la zona Centro Poniente. 92
A.4. Evolución de la diferencia del MAE en las predicciones para el *dataset* de Pantalones en la zona Centro Poniente. 92

B.1. Pares de series temporales más cercanas en el *dataset* de Microondas en la tienda Plaza Vespucio. 94
B.2. Evolución de la diferencia del MAE en las predicciones para el *dataset* de Microondas en la tienda Plaza Vespucio. 94
B.3. Pares de series temporales más cercanas en el *dataset* de Microondas en la zona Centro Poniente. 95
B.4. Evolución de la diferencia del MAE en las predicciones para el *dataset* de Microondas en la zona Centro Poniente. 95

Capítulo 1

Introducción

1.1. Motivación

1.1.1. Contexto general

En el ámbito del comercio de productos, el *retail* consiste en la venta de productos al por menor, concepto que también es conocido como comercio minorista. Su objetivo es vender productos a una gran cantidad de clientes, en contraste con el comercio mayorista, donde lo que se busca es vender cantidades masivas de productos a unos pocos clientes. Los productos que se comercializan en esta industria son almacenados en bodegas o centros de distribución, que suelen estar presentes en distintos puntos del país y son a partir de donde se envían los productos a diversas sucursales para ser vendidos a los clientes.

Dependiendo de los propósitos y las características de los productos, estos suelen ser categorizados en grupos y subgrupos, los cuales, para efectos del presente trabajo, son, de más general a más específico, línea, sublínea, clase, subclase, estilo y SKU (*Stock Keeping Unit*).

En el contexto nacional, la industria del *retail* sufrió una importante caída en el año 2020, principalmente debido a que la crisis social vivida en el país a finales del año 2019, sumada al inicio de la pandemia del COVID-19 a principio del año 2020, empeoraron las condiciones económicas de los chilenos, provocando una reducción en el ingreso disponible de los hogares relacionada con el aumento del desempleo[5].

A pesar de este negativo escenario, durante la última semana del año 2021, se registró un alza anual de 12.9 % en el termómetro semanal de ventas del *retail*[11] con respecto al año pasado, mientras que en el mes de diciembre, las semanas marcaron un alza promedio de 19.2 % anual, tal como se muestra en la figura1.1. Además, en esta figura se aprecia que con respecto al año anterior, la industria del retail tuvo un fuerte crecimiento durante el 2021.

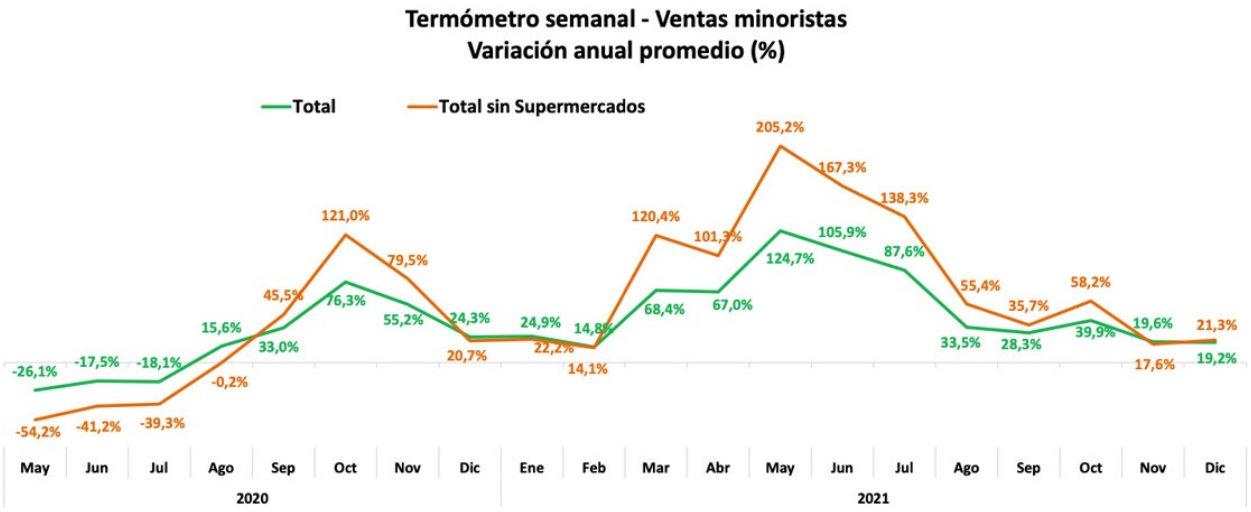


Figura 1.1: Termómetro semanal de ventas minoristas del año 2021[11].

1.1.2. Problema a abordar

En el presente trabajo se abordó un problema altamente relevante en esta industria que aqueja a la gran mayoría de los *retailers*. Este corresponde al problema de la asignación óptima de productos desde los centros de distribución hacia las sucursales de ventas. Esto ocurre cuando, por ejemplo, una tienda tiene un *stock* de 50 unidades de un producto en particular que necesita distribuir a 5 de sus sucursales presentes a lo largo del país para ser vendidas. Si la tienda decide hacer una distribución equitativa y envía 10 productos a cada una de esas sucursales, eventualmente habrán sucursales con una demanda superior a 10 unidades a las cuales les faltarán productos para poder vender, así como habrán otras sucursales con una demanda inferior a 10 unidades a las que les sobrarán productos.

Ambos escenarios están lejos del óptimo y representan pérdidas para la tienda, ya que si en una sucursal hay una demanda superior a la estipulada, significa que en esta tienda se estará desaprovechando la oportunidad de vender más productos; de igual forma, una sucursal con una demanda inferior a la esperada representa menores ganancias para la tienda, ya que al sobrar productos, probablemente será necesario hacer una liquidación y aplicarle un descuento sobre el precio original del producto, pudiendo haberlos vendido a precio normal en otra sucursal con mayor demanda.

En las siguientes tablas se muestra un ejemplo concreto del escenario recién descrito, correspondiente a un caso de venta de un producto de la empresa con la cual se realizó este trabajo.

Caso subóptimo

	Semana 1 ($10,000)		**Semana 2 ($5,000)**	
	Stock	Ventas	Stock	Ventas
Sucursal 1	80	50	30	30
Sucursal 2	20	20	0	0

$\Rightarrow \$850,000$

Tabla 1.1: Problema de distribución de productos, caso subóptimo

Caso óptimo

	Semana 1 ($10,000)		**Semana 2 ($5,000)**	
	Stock	Ventas	Stock	Ventas
Sucursal 1	50	50	0	0
Sucursal 2	50	50	0	0

$\Rightarrow \$1,000,000$

Tabla 1.2: Problema de distribución de productos, caso óptimo

En la tabla 1.1 bajo un escenario de distribución subóptima, se aprecia que durante la primera semana, en la sucursal 1 sobraron productos por vender a su precio original ($\$10,000$), motivo por el cual fue necesario venderlos con un importante descuento ($\$5,000$) en la semana siguiente, lo cual se tradujo en ganancias de $\$850,000$. Por otro lado, con una distribución óptima como la de la tabla 1.2, las ventas fueron realizadas todas durante la primera semana, sin sobrar productos para tener que vender a menor precio. Con esto, las ganancias fueron de $\$1,000,000$. Entonces, queda en evidencia que al no tener una buena predicción de distribución de productos a las sucursales de ventas, las tiendas están perdiendo la oportunidad de maximizar sus ganancias.

En la actualidad, la abundancia de datos es aprovechada cada día con mayor frecuencia mediante diversas técnicas de ciencia de datos. El problema a abordar en este trabajo se hace inminente de tratar de dicha forma, debido a la gran cantidad de datos con los que se suele contar en estos casos. El trabajo se realizó en este contexto de *Data Science*, con enfoque en el análisis de series de tiempo de productos con comportamientos de ventas similares.

El presente trabajo se llevó a cabo en Santiago de Chile, con una empresa que, por motivos de confidencialidad de datos, será mencionada de aquí en adelante simplemente como "La Empresa". Esta empresa lleva más de 50 años en el mercado del *retail* y cuenta con más de 40 sucursales a lo largo del país, por lo que su volumen de datos históricos de ventas es enorme. Esto mismo les motivó a querer aprovechar esa gran cantidad de datos para hacer predicciones de demanda mucho más realistas y efectivas. Si bien, "La Empresa" cuenta con un sistema de predicción de ventas que les permite tener una ventana de 4 semanas al futuro, estas predicciones son realizadas tomando en consideración únicamente las características relacionadas directamente con la categorización de los productos, vale decir, productos de una misma línea, sublínea, clase, subclase y estilo.

El principal problema que esto presenta es el hecho de restarle relevancia a la data de ventas históricas de los productos, perdiendo así una valiosa oportunidad de aprovechar esa riqueza de datos. Otro punto importante a considerar es que la gran mayoría de productos tienen periodos de venta acotados a solo un año, ya que los productos se venden en una temporada y después se retiran. Esto se traduce en tener que realizar predicciones para productos que no tienen un largo historial de ventas, lo cual deberá realizarse en base a los datos históricos de productos con comportamientos de ventas semejantes.

Dicho esto, se detectó una gran oportunidad para aplicar técnicas de ciencia de datos y así poder simular y evaluar predicciones para productos cuyas curvas de ventas estén más interrelacionadas entre sí.

1.1.3. Impacto del Trabajo de Memoria

El escenario recién planteado afecta desde pequeñas a grandes empresas del sector del *retail* chileno y también a nivel mundial. Por esta razón, en principio, la relevancia del presente trabajo de memoria podría generar un impacto en todas aquellas empresas del sector del *retail* que tengan suficiente riqueza de datos como para poder llevar a cabo un correcto pronóstico de ventas en base a la similitud de los comportamientos de ventas de sus productos.

Sin embargo, el proceso de análisis y predicción de series de tiempo es una tarea que se desarrolla como un trabajo de relojería, ya que, como en todo proyecto de ciencia de datos, no existen recetas prefabricadas en las cuales podamos tomar una solución y aplicarla directamente sobre datos de un mismo tipo. En consecuencia, puede que la heurística para la simulación de pronóstico de ventas propuesta en este trabajo no sea necesariamente la más adecuada para otra empresa, puesto que la naturaleza de los datos puede variar mucho incluso entre *retailers* similares.

No obstante, el trabajo llevado a cabo, en términos del desarrollo y la heurística con la que finalmente se realizaron y evaluaron los pronósticos de ventas, puede ser replicado en otras empresas y ser de gran utilidad para llegar a encontrar una solución óptima al problema planteado.

1.2. Objetivos

1.2.1. Objetivo general

El objetivo general de este trabajo es evaluar la factibilidad de realizar predicciones de ventas a partir de la demanda histórica de productos de una misma categoría con comportamientos similares en el tiempo.

1.2.2. Objetivos específicos

1. Realizar un análisis exploratorio de los datos para obtener estadísticas y gráficos que permitan descubrir filtros apropiados para aplicar a los *datasets* y así acotar la cantidad inicial de datos disponibles.
2. Generar un nuevo *dataset* con un formato adecuado para el análisis de series temporales.
3. Definir una métrica adecuada para comparar cuantitativamente la similitud entre las series temporales de ventas.
4. Comprobar la existencia de series temporales con comportamientos de ventas similares de productos que no sean cualitativamente parecidos.
5. Definir una heurística para simular predicciones de ventas para los pares de series temporales más similares.
6. Comparar las simulaciones de pronóstico de ventas mediante métricas, tablas y gráficos, con el fin de evaluar la calidad predictiva del modelo propuesto.
7. Replicar este estudio para los 4 *datasets* disponibles, tanto a nivel de tiendas como de zonas geográficas.

1.3. Solución

1.3.1. Hipótesis

La solución propuesta implica principalmente las siguientes hipótesis:

- Existen productos para los que un comportamiento de venta similar en su primera mitad del año implican que sus comportamientos de ventas en la segunda mitad del año también sean similares.
- Es posible que dos productos no sean cualitativamente semejantes pero que sí presenten un comportamiento de ventas similar que permita realizar la predicción de ventas mediante la metodología propuesta.

1.3.2. Alternativas analizadas

En cuanto a las mediciones de similitud entre series temporales, dentro de la gran variedad de medidas de distancia existentes, las alternativas estudiadas fueron: Distancia Euclidiana, *Dynamic Time Warping* (DTW), *Cosine Similarity*, *Angular Metric for Shape Similarity* (AMSS), *Manhattan Distance* y *Minkowski Distance*.

Por otro lado, para evaluar las simulaciones de los pronósticos de venta, se utilizaron la métricas de *Mean Absolute Error* (MAE, 2.5.1), *Mean Absolute Percentage Error* (MAPE, 2.5.2) y *Root Mean Squared Errors* (RMSE, 2.5.3).

1.4. Estructura de la Memoria

Los próximos capítulos del presente trabajo tienen la siguiente estructura:

- El capítulo 2 contiene el marco teórico y el estado del arte, donde se presentan y detallan los principales conceptos involucrados, así como las principales soluciones existentes, aportando toda la base necesaria para comprender a cabalidad el presente trabajo.
- Siguiendo con el capítulo 3, se introduce el problema de la predicción de demanda y se describe de manera más profunda dicho problema en el contexto de La Empresa, se habla de la relevancia de hallar una solución y se describe cómo debe ser esta.
- Continuando con el 4, se elabora una exploración preliminar de los datos disponibles, sus atributos, distribuciones, estadísticas y gráficos relevantes, con el fin de describir y entender los datos con los que se realizó el trabajo.
- Luego, en el capítulo 5 se habla en detalle sobre el diseño de la solución implementada, incluyendo la prueba de concepto realizada y las etapas de preprocesamiento y manejo de datos, junto con el desarrollo para la simulación de pronóstico de ventas.
- A continuación, en el capítulo 6 se muestran, interpretan y discuten los resultados de la solución implementada y cómo esta resuelve el problema planteado.
- Por último, en el capítulo 7 se resume brevemente el trabajo realizado, junto a unas conclusiones respecto a los resultados, en relación con los objetivos y las hipótesis. Se termina con una discusión sobre posibles trabajos futuros.

Capítulo 2

Marco Teórico y Estado del Arte

Para poder tener un buen entendimiento del presente trabajo, es necesario contar con una fuerte base teórica. Por lo mismo, en este capítulo se proporcionarán los conceptos más relevantes implicados. Se comenzará hablando de la ciencia de datos, luego se hablará sobre las series de tiempo y de por qué es necesario tener especial cuidado cuando se utilizan en proyectos de ciencia de datos. Se continuará mencionando las principales medidas de distancia que se utilizan para identificar datos cercanos o "similares" en este contexto. Por último, se explicarán las métricas de desempeño que se utilizaron para evaluar las predicciones.

2.1. Ciencia de Datos

En la actualidad, la ciencia de datos, en inglés *Data Science*, ha demostrado ser de extrema utilidad en sectores públicos, privados y académicos, en prácticamente cualquier rubro y área de investigación, resultando ser especialmente valiosa en esta era de abundancia de datos.

En la década de 1990, el término se conocía como *Knowledge Discovery in Databases* (KDD), traducido como "Descubrimiento de Conocimiento en las Bases de Datos", y aludía a un concepto que había surgido hacía algunos años, cuando la abundancia de datos comenzaba a hacerse cada vez más inminente. KDD se refiere al proceso general de descubrimiento de conocimiento útil a partir de los datos[14]; concepto que solía confundirse con Minería de Datos (*Data Mining*), englobado por KDD, el cual consiste en aplicar algoritmos para encontrar patrones en los datos, y también se confundía con (*Big Data*), que hace referencia a almacenar, procesar, consultar y analizar datos de un volumen masivo y exponencialmente creciente mediante técnicas no "convencionales".

Con el tiempo, KDD evolucionó a lo que hoy en día se conoce como *Data Science*, una ciencia interdisciplinaria conformada por múltiples aristas, en la que se ven envueltos conceptos y campos como ciencias de la computación, desarrollo de software, matemáticas, estadística, aprendizaje de máquinas, conocimiento del dominio e investigación, tal como se puede apreciar en la figura 2.1.

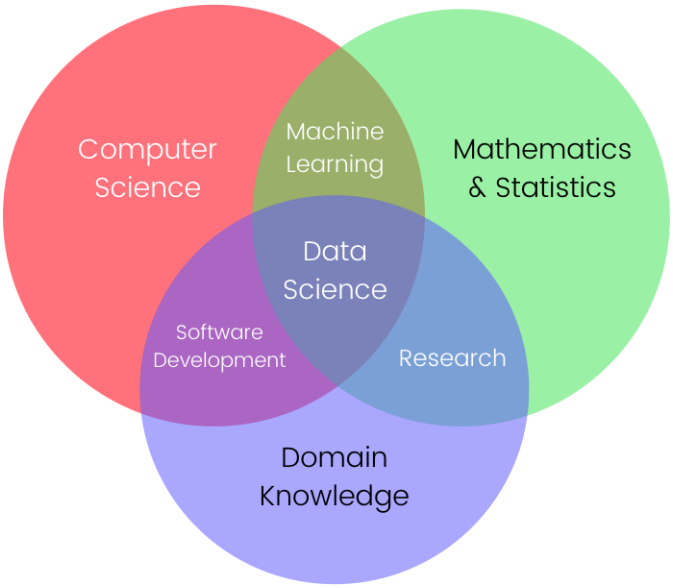


Figura 2.1: Diagrama de Venn de la interdisciplinariedad de la Ciencia de Datos[3].

Desarrollar proyectos de ciencia de datos es un trabajo minucioso en el que no hay recetas prefabricadas, ya que cada uno implica sus propios datos, que si bien pueden ser semejantes a los datos de otros casos, no será posible implementar una solución previa directamente sobre los datos. En estos proyectos se ven involucradas, a grandes rasgos, las siguientes etapas, las cuales forman parte de lo que se conoce como el *pipeline* de la ciencia de datos (ver figura 2.2):

- **Adquisición de datos**: luego de un trabajo inicial de conocimiento y entendimiento del dominio del problema, es necesario recolectar una gran cantidad de datos pertinentes que permitan generar modelos confiables. Aquí se genera un *dataset* conformado, usualmente, a partir de múltiples fuentes de datos.
- **Preprocesamiento**: tener abundancia de datos permitirá crear modelos precisos, pero antes de generar modelos con ellos es necesario preparar y limpiar estos datos para remover anomalías (también conocidas como *outliers*), eliminar datos duplicados, normalizar formatos, entre otras tareas.
- **Exploración**: teniendo los datos listos para ser usados, viene la etapa de Análisis Exploratorio de Datos (EDA, por sus siglas en inglés *Exploratory Data Analysis*), generalmente mediante visualizaciones, donde se podrá tener una mejor comprensión de los datos y también se podrán encontrar patrones o tendencias en ellos que se quieran modelar en la siguiente etapa.
- **Modelamiento**: a continuación se lleva a cabo la creación y perfeccionamiento de un modelo representativo de los datos, el cual será usado posteriormente con distintos propósitos, como hacer predicciones, clasificaciones, entre otros. Es en esta etapa donde se prueban múltiples algoritmos y técnicas.
- **Interpretación**: habiendo obtenido resultados, lo que sigue es interpretarlos, lo cual suele realizarse creando visualizaciones capaces de ilustrar y comunicar claramente los conocimientos descubiertos (también conocidos como *insights*). Para obtener la mayor utilidad de esta etapa se requiere un fuerte conocimiento del dominio.

- **Revisión**: por último, luego de una efectiva interpretación de los datos, y generalmente después de haber pasado a una fase productiva del modelo creado, viene una etapa de revisión y actualización de este modelo. Esto se debe a la naturaleza cambiante y evolutiva de los datos que puede "desafinar" y degradar el modelo.

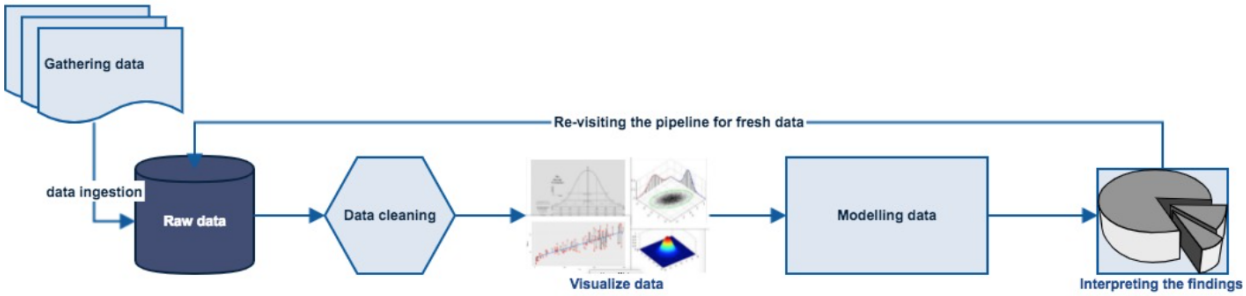


Figura 2.2: *Pipeline* del desarrollo de proyectos de Ciencia de Datos[27].

Si bien a lo largo de este trabajo de memoria se trabajó en la mayoría de las etapas de la ciencia de datos señaladas previamente, el foco y énfasis del presente informe estará puesto en las etapas de modelamiento e interpretación de los resultados de las predicciones de series temporales.

2.2. Series de tiempo

En un proyecto de Ciencia de Datos, es necesario tener claridad del tipo de dato con el que se estará trabajando, ya que será relevante a la hora de elegir un algoritmo para la creación del modelo. Los algoritmos de Ciencia de Datos operan con diversos tipos y formatos de datos, como por ejemplo, de tipo categórico, de tipo nominal, de tipo ordinal, de tipo numérico con valores que pueden ser discretos, o de naturaleza continua (discretizados para ser analizados), entre otros tipos. Un formato de dato en particular son las series de tiempo. Al ser graficadas, por convención, se utiliza el eje X para el tiempo, como en la siguiente figura.

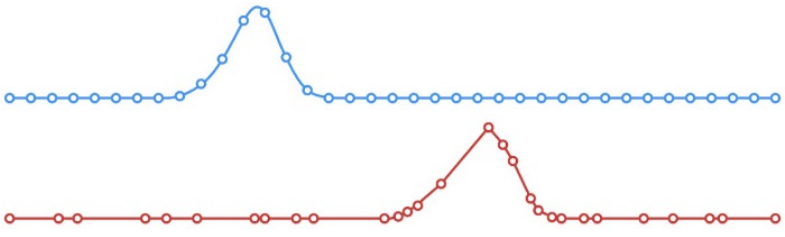


Figura 2.3: Ejemplo de series de tiempo regulares e irregulares[12].

Las series de tiempo son mediciones o eventos que se registran a lo largo del tiempo. Pueden clasificarse en dos tipos: regulares e irregulares. Se consideran regulares si las mediciones son obtenidas en intervalos de tiempo constantes, como en la serie superior de la figura 2.3; y son irregulares si los intervalos de tiempo no son constantes, es decir, presentan una frecuencia de muestreo irregular, tal como en la serie inferior de la misma figura.

Debido a que el tiempo está siempre presente durante cualquier tipo de observación, las series de tiempo tienen numerosas aplicaciones, tales como registros climáticos de datos ob-

tenidos a cada hora, diariamente o semanalmente, seguimientos de cambios en el rendimiento de aplicaciones, registro de datos en tiempo real mediante dispositivos médicos (como un electrocardiograma, por ejemplo), seguimientos de estabilidad de conexiones a redes, entre otras aplicaciones[12]. Las industrias y campos de estudio en los que se utilizan las series de tiempo son variados: estadísticas, econometría, reconocimiento de patrones, procesamiento de señales, matemáticas financieras, pronóstico del tiempo, astronomía, entre muchos otros.

Las series de tiempo pueden manifestar diversos aspectos de comportamientos tales como tendencia, estacionalidad y ciclos. La tendencia es la dirección a largo plazo de una serie; puede ser ascendente, descendente, como también puede no haber una tendencia claramente distinguible.

La estacionalidad ocurre cuando hay un comportamiento repetido en los datos que sucede en intervalos regulares y está relacionada con el comportamiento estacional natural o humano.

Los ciclos ocurren cuando una serie presenta un patrón ascendente y descendente que no es estacional. Los ciclos pueden una duración variable, lo cual implica que son más difíciles de detectar que la estacionalidad.

Otros comportamientos que pueden presentar las series de tiempo son la variación y la irregularidad. La variación aleatoria está presente en todos los datos, y en las series de tiempo, también lo está en mayor o menor medida.

Las irregularidades ocurren debido a fenómenos puntuales que pueden provocar alteraciones inesperadas, como "depresiones" o "saltos" en las series. En la figura 2.4 se muestran los comportamientos mencionados.

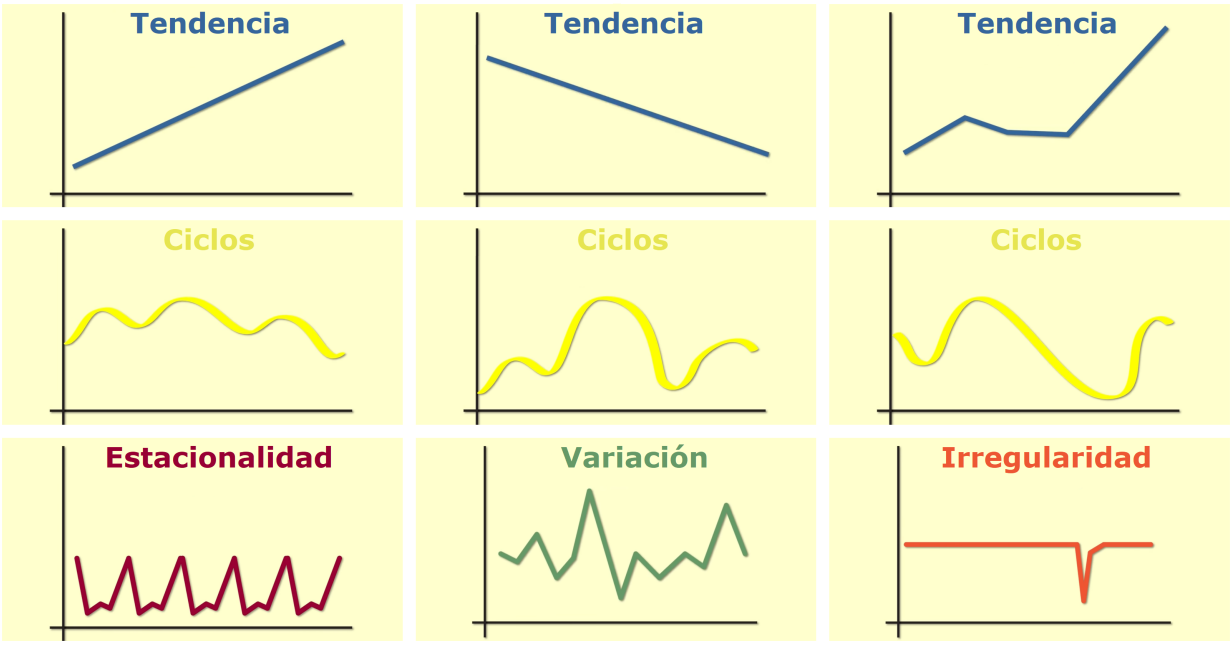


Figura 2.4: Comportamientos presentes en las series de tiempo[25].

En el análisis de las series de tiempo, donde se busca comprender o modelar la autoco-

rrelación en los datos de las series, principalmente con el fin de realizar predicciones, además de los distintos comportamientos mencionados, existen múltiples desafíos en cada una de las diversas aplicaciones posibles de series de tiempo[20]. En el presente trabajo se estudiaron las series de tiempo correspondientes a los datos históricos de ventas de productos con el fin de evaluar predicciones de ventas en base a la similitud de los comportamientos de ventas. Los desafíos y dificultades encontradas durante el trabajo se detallarán y discutirán en los capítulos 5 y 6.

2.3. *Time Series Smoothing*

En el análisis de series de tiempo, el *Smoothing*, o suavizado, de las series es una transformación habitual. Es una técnica de procesamiento de datos que ayuda a reducir el ruido en las series temporales, lo cual permite detectar información más valiosa con respecto a las series sin tratar. Por ejemplo, en el análisis de *marketing*, esta transformación es ampliamente utilizada, ya que permite identificar de manera los cambios en la economía[9].

El propósito del suavizado de series es disminuir el impacto de los altos *peaks* y *off-peaks* en las series, que corresponden a *outliers* (ruido) en el contexto del análisis de series de tiempo. En esta forma es mucho más fácil apreciar tendencias, estacionalidades y otros patrones en las series de tiempo.

Existen múltiples formas de aplicar *smoothing* a las series de tiempo, tales como: *Moving average smoothing*, *Exponential smoothing*, *Double exponential smoothing* y *Triple exponential smoothing*. En el presente trabajo se utilizó el *Centered Moving Average smoothing*, ya que es simple de implementar, además de ser uno de los más ocupados. El *Centered Moving Average smoothing* se calcula de la siguiente manera:

Definición 2.1 *Cálculo de Centered Moving Average smoothing.*

Para cada periodo t y cada uno de los puntos, X_t, que conforman a una serie de tiempo, el cálculo de su valor suavizado centrado, S_t se obtiene mediante la siguiente fórmula, para k impar:

$$CMA_t = \frac{(X_{t-\frac{k-1}{2}} + ... + X_t + ... + X_{t+\frac{k-1}{2}})}{k} \tag{2.1}$$

Luego de aplicar este cálculo sobre cada uno de los puntos que conforman una serie de tiempo, se puede obtener un resultado como el que se aprecia en la figura a continuación.

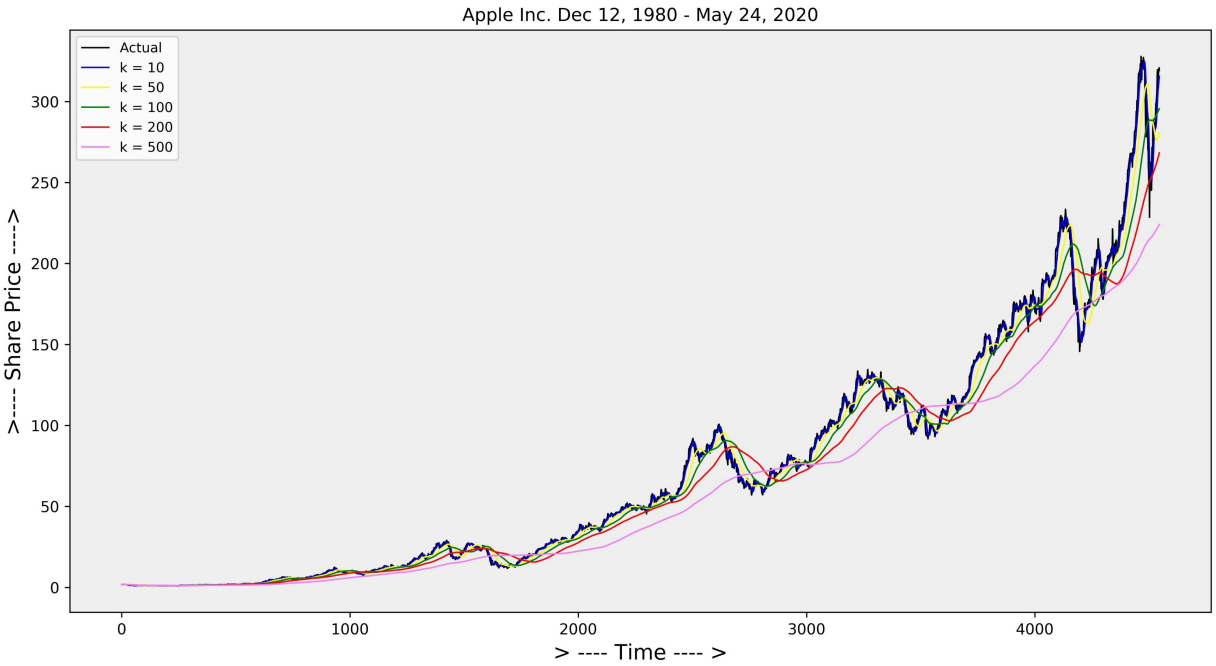


Figura 2.5: Ejemplo de suavizado de series temporales[9].

En la figura 2.5 se puede apreciar el efecto de "aplanado" que se produce al aumentar la ventana de periodos k a utilizar en el *smoothing*.

2.4. Medidas de distancia

Es fundamental tener una medida cuantitativa que permita determinar qué tan cerca están los elementos que se van a comparar, o, dicho de otro modo, qué tan similares son entre sí. Existen medidas de distancia que son más apropiadas bajo ciertos casos o bajo ciertos tipos de datos, y tener conocimiento sobre ellas ayudará a tener una buena elección de medida y así mejorar la precisión de los modelos. En esta sección se mencionarán las principales medidas de distancia estudiadas.

2.4.1. Distancia Euclidiana

La distancia Euclidiana es una de las más comunes. En términos simples, se define como el largo del segmento que conecta dos puntos. Su aplicación se extiende a datos con n variables, ya que se puede utilizar en un espacio de n dimensiones[15]. Su fórmula deriva directamente del teorema de Pitágoras, por lo que también se le conoce como la distancia de Pitágoras, y se define en un espacio euclídeo n-dimensional para dos puntos x e y, como sigue.

Definición 2.2 *Distancia Euclidiana entre dos puntos en un espacio n-dimensional.*

$$\mathrm{d}(x, y) = \sqrt{\sum_{\mathrm{i}=1}^{n} (x_{\mathrm{i}} - y_{\mathrm{i}})^2} \tag{2.2}$$

Debido a su uso intuitivo y a la simplicidad de su implementación, es una de las distancias más utilizadas, mostrando buenos resultados para múltiples casos y diversos algoritmos. Sin embargo, a medida que incrementa la dimensionalidad de los datos sobre los cuales se calcula, la distancia Euclidiana se vuelve menos útil y se recomienda utilizar otras medidas; esto se debe a lo que se conoce como la "Maldición de la Dimensión"[17]. Por otra parte, para evitar mediciones sesgadas, se recomienda normalizar los datos antes de usar esta medida de distancia.

2.4.2. Dynamic Time Warping (DTW)

Desde su publicación en 1978, donde Sakoe et al. propusieron esta novedosa medida para mejorar el reconocimiento de palabras habladas[28], la *Dynamic Time Warping* (DTW) ha sido ampliamente aplicada en la medición de distancias para series de tiempo, convirtiéndose en la más utilizada para este formato de dato.

La distancia Euclidiana, a pesar de tener buenos resultados en la mayoría de los casos, cuando se trata de series de tiempo, al ignorar la dimensión temporal de los datos no es tan apropiada, ya que por ejemplo, si hay dos series de tiempo que están altamente correlacionadas, pero con un mínimo desfase temporal, con la distancia Euclidiana se medirían como si fueran poco similares. La DTW toma en cuenta este caso, siendo una medida de distancia más apropiada para las series de tiempo, al poder medir similitud entre secuencias temporales que no estén necesariamente alineadas en el tiempo, largo o velocidad[1].

La definición matemática del cálculo de la DTW es la siguiente.

Definición 2.3 *Distancia DTW entre dos series temporales[31].*

Dadas las series $X = \{x_0, ..., x_n\}$ *e* $Y = \{y_0, ..., y_m\}$*, la distancia DTW desde* X *a* Y *se formula como el problema de optimización:*

$$DTW(X, Y) = \min_{\pi} \sqrt{\sum_{(\mathrm{i},j)\in\pi} \mathrm{d}(x_{\mathrm{i}}, y_{\mathrm{i}})^2} \tag{2.3}$$

donde $\pi = [\pi_0, ..., \pi_k]$ *es un camino que satisface las siguientes propiedades:*

- *Es una lista de pares de índices* $\pi_k = (\mathrm{i}_k, j_k)$ *con* $0 \leq \mathrm{i}_k < n$ *y* $0 \leq j_k < m$.
- $\pi_0 = (0, 0)$ *y* $\pi_k = (n-1, m-1)$.
- *Para todo* $k > 0$, $\pi_k = (\mathrm{i}_k, j_k)$ *está relacionado con* $\pi_{k-1} = (\mathrm{i}_{k-1}, j_{k-1})$ *de la siguiente manera:*
  - $\mathrm{i}_{k-1} \leq \mathrm{i}_k \leq \mathrm{i}_{k-1} + 1$.
  - $j_{k-1} \leq j_k \leq j_{k-1} + 1$.

Es decir, la DTW se calcula como la raíz cuadrada de la suma de las distancias al cuadrado entre cada elemento en X y su punto más cercano en Y. También se destaca que en

la definición matemática, d es una distancia entre dos puntos, usualmente la distancia Euclidiana; y que $DTW(X,Y) \neq DTW(Y,X)$. De esta forma, el algoritmo de DTW comparará cada elemento en X con cada elemento en Y, haciendo $N \cdot M$ comparaciones, por lo que la complejidad del algoritmo es $O(N \cdot M)$, y en su implementación se utiliza programación dinámica.

En la figura 2.6 se realiza una comparación entre el resultado del cálculo de la distancia Euclidiana para dos series de tiempo, y también el cálculo de la DTW para las mismas series.

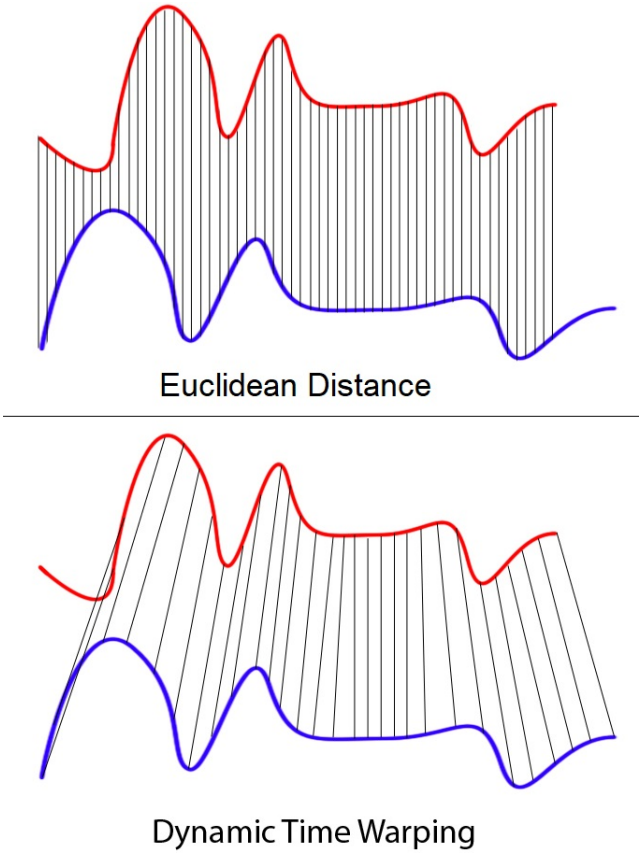


Figura 2.6: Comparación de distancia Euclidiana y DTW para dos series de tiempo[6].

Por último, es relevante mencionar una variante de la DTW conocida como *soft-DTW*[7]. DTW no es diferenciable con respecto a sus *inputs* debido a que la operación mín no es diferenciable. Para que sea diferenciable, y así pueda utilizarse en optimizaciones en base al método del gradiente, la variante *soft-DTW* utiliza la operación *soft-min*, la cual se define de la siguiente manera.

$$\text{mín}^{\,\gamma}(a_1, ..., a_n) = -\gamma \log \sum \mathrm{e}^{-a_\mathrm{i}/\gamma} \tag{2.4}$$

Con esto, la distancia *soft-DTW* queda definida como sigue.

$$soft\text{-}DTW^{\gamma}(X,Y) = \min_{\pi}{}^{\gamma} \sum_{(\mathrm{i},j)\in\pi} \sqrt{\mathrm{d}(x_\mathrm{i}, y_j)^2} \tag{2.5}$$

De esta forma, γ se convierte en un hiperparámetro que afectará en el funcionamiento de la *soft-DTW*, por lo que será necesario probar con distintos valores de este hiperparámetro para poder encontrar su valor óptimo tal que su funcionamiento sea el más consistente y efectivo a la hora de medir distancias entre series de tiempo.

2.4.3. *Cosine Similarity*

Cosine Similarity (Similitud Coseno) es otra medida de distancia, enfocada a medir la similitud entre dos vectores. Su aplicación más común es en el área de análisis de texto, y se utiliza para medir la similitud entre dos documentos, donde cada documento es representado por un vector con la frecuencia de sus términos (*term-frecuency vector*)[16].

En esta medida, la similitud se calcula mediante el coseno del ángulo entre dos vectores, y determina si dichos vectores apuntan en la misma dirección.

La definición matemática de esta medida de similitud deriva directamente del producto escalar (o producto punto) entre dos vectores dentro de un espacio euclídeo, de la siguiente manera.

Definición 2.4 *Cosine Similarity entre dos vectores en un espacio n-dimensional.*

Dados los vectores $X = \{x_1, ..., x_n\}$ *e* $Y = \{y_1, ..., y_n\}$*, que tienen entre sí un ángulo* θ*, la Cosine Similarity entre* X *e* Y *es:*

$$\text{similarity}(X, Y) = cos(\theta) = \frac{X \cdot Y}{\|X\| \, \|Y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2.6}$$

Si esta medida da un valor cercano a 0, significa que los vectores están en un ángulo de 90 grados entre sí (ortogonales), y quiere decir que no son muy similares. Mientras más se acerque a 1 el valor, significa que el ángulo entre los vectores es más pequeño y por ende, tienen una mayor coincidencia, o similitud.

En la figura 2.7 se muestra un ejemplo de *Cosine Similarity* aplicada en un espacio tridimensional, donde cada dimensión representa una palabra, y cada vector representa una frase compuesta por dichas palabras.

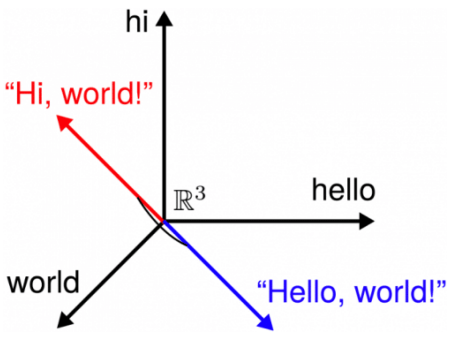


Figura 2.7: Ejemplo de *Cosine Similarity* aplicado a documentos de texto[26].

La ventaja de *Cosine Similarity* es que a pesar de que dos vectores estén distantes, según la medida de la distancia Euclidiana, puede que aún sean considerados como similares, si es que el ángulo entre ellos es pequeño. A menor ángulo entre los vectores, mayor será la similitud entre ellos.

2.4.4. Angular Metric for Shape Similarity (AMSS)

Otra medida de distancia, o similitud, es la *Angular Metric for Shape Similarity* (AMSS), propuesta en 2012 por Nakamura et al.[23], y fue diseñada exclusivamente para aplicarse a series de tiempo.

A diferencia de otras medidas de distancia que tratan individualmente los puntos pertenecientes a las series de tiempo, la AMSS trata las series de tiempo como secuencias de vectores que las representan, como se puede apreciar en la figura 2.8.

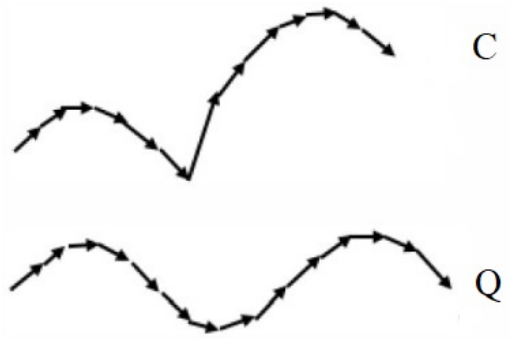


Figura 2.8: Visualización de dos series de tiempo, C y Q, representadas como secuencias de vectores[22].

De esta forma, la AMSS se centra en la forma de las series y las compara mediante una variante de *Cosine Similarity*, descrita anteriormente.

La definición matemática del cálculo de la AMSS se realiza de manera recursiva, como se muestra a continuación:

Definición 2.5 *Definición de la Angular Metric for Shaped Similarity para dos secuencias de vectores[23].*

Sean Q y C dos series de tiempo, representadas por las secuencias de vectores $Q_n = (q_1, ..., q_n)$ y $C_m = (c_1, ..., c_m)$, respectivamente, la AMSS se define de la siguiente manera:

$$AMSS(Q_n, C_m) = \text{máx} \left\{ \begin{array}{c} AMSS(Q_{n-1}, C_{m-1}) + 2sim(q_n, c_m), \\ AMSS(Q_{n-2}, C_{m-1}) + 2sim(q_{n-1}, c_m) + sim(q_n, c_m), \\ AMSS(Q_{n-1}, C_{m-2}) + 2sim(q_n, c_{m-1}) + sim(q_n, c_m) \end{array} \right\} \quad (2.7)$$

con las siguientes consideraciones:

- *$AMSS(Q_1, C_1)$ se define como $sim(q_1, c_1)$, para terminar la recursión.*
- *Para $n = 0$ o $m = 0$, $AMSS(Q_n, C_m) = -\infty$, para evitar caminos inválidos que involucren los vectores indefinidos q_0 o c_0.*
- *La función para la similitud coseno que se utiliza, se modifica ligeramente de la siguiente manera:*

$$sim(q_n, c_m) = \begin{cases} 0 & \text{si } \theta > \pi/2 \\ cos\,\theta (= \frac{q_n \cdot c_m}{\|q_m\| \|c_m\|}) & caso\ contrario \end{cases} \tag{2.8}$$

donde θ corresponde al ángulo entre los vectores q_n y c_m, como se muestra en la figura 2.9, parte b).

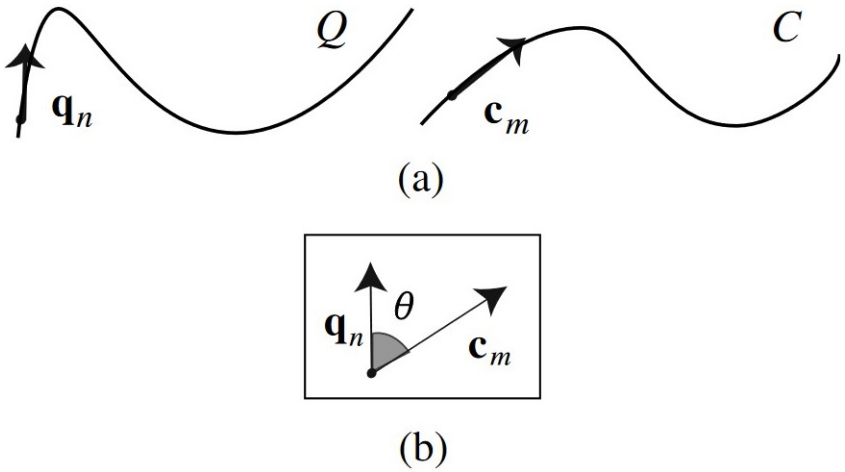


Figura 2.9: Ángulo entre dos vectores pertenecientes a series de tiempo[23].

Esta medida de distancia o similitud se calcula sobre las secuencias de vectores correspondientes a dos series de tiempo, en base a las direcciones de los vectores, en vez de las ubicaciones de los datos en su espacio n-dimensional.

De esta manera, es posible comparar dos series de tiempo con respecto a su forma, como se puede apreciar en la figura 2.10, donde se ve que es posible comparar tramos de las series de tiempo y así poder determinar tramos distintos, tal como se puede apreciar en los tramos c_7 y q_7.

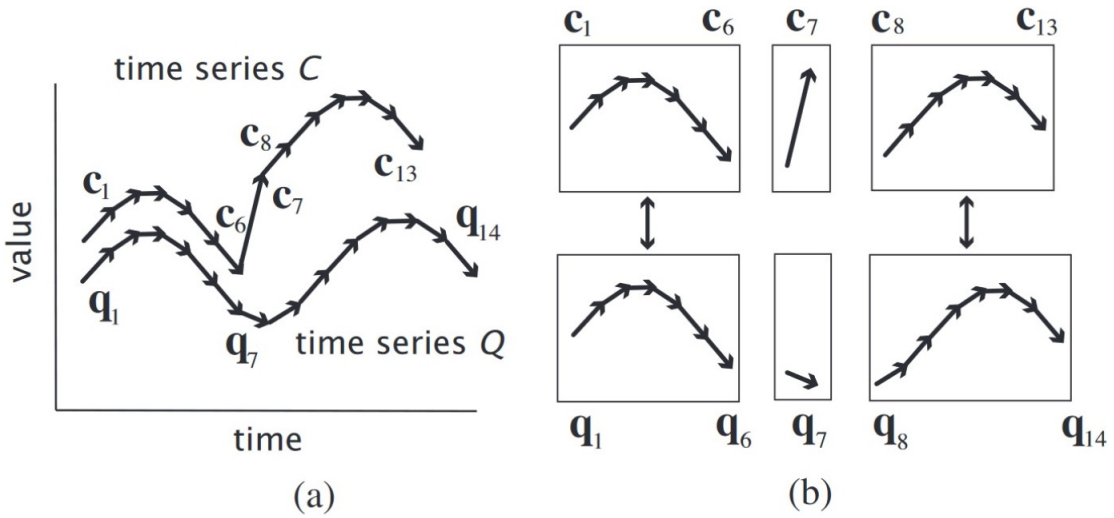


Figura 2.10: Comparación de dos series de tiempo mediante la AMSS[23].

La AMSS fue desarrollada de tal forma que sea robusta a las alteraciones temporales en la series, ya sea de amplitud o escalado. Por otra parte, también es "sensible" ante oscilaciones de corto plazo.

2.4.5. Otras medidas de distancia

A continuación se mencionarán otras medidas de distancia que fueron estudiadas, incluyendo medidas poco comunes para el análisis de series de tiempo, propuestas de medidas novedosas y comentarios sobre estudios comparativos de distancias.

Manhattan Distance. Esta medida, también conocida como la "geometría del taxista", calcula la distancia entre dos puntos mediante la suma de las diferencias absolutas de sus coordenadas.

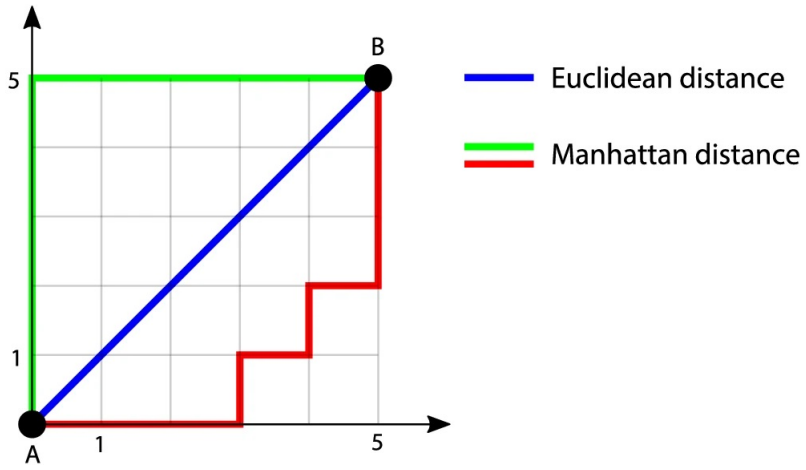


Figura 2.11: Comparación de las distancias de Manhattan y Euclidiana para dos puntos[21].

Su nombre está inspirado en la ciudad de Manhattan (EE.UU.), ya que la mayoría de sus calles tienen un diseño de cuadrícula, y la distancia entre un punto A y un punto B, es equivalente a sumar las calles por las que se transita para llegar desde A hasta B, tal como

se muestra en la figura 2.11, donde además se hace una comparación gráfica con la distancia Euclidiana.

Como se puede apreciar, en el cálculo de la *Manhattan Distance* no hay movimientos diagonales como en la distancia Euclidiana, y el camino superior tiene una distancia equivalente a la del camino inferior.

Formalmente, se define de la siguiente manera.

Definición 2.6 *Manhattan Distance entre dos puntos en un espacio n-dimensional.*

$$\mathrm{d}(x, y) = \sum_{\mathrm{i}=1}^{n} |x_{\mathrm{i}} - y_{\mathrm{i}}| \tag{2.9}$$

Esta distancia se puede utilizar en espacios n-dimensionales y suele ser apropiada en conjuntos de datos con atributos discretos o binarios.

Minkowski Distances. Esta corresponde a una familia de medidas de distancia que dependen del parámetro p, a partir del cual se obtienen distintos cálculos. Su expresión matemática es la siguiente.

Definición 2.7 *Minkowski Distance entre dos puntos en un espacio n-dimensional.*

$$\mathrm{d}(x, y) = (\sum_{\mathrm{i}=1}^{n} |x_{\mathrm{i}} - y_{\mathrm{i}}|^{p})^{\frac{1}{p}} \tag{2.10}$$

Esta distancia puede considerarse una generalización de la *Manhattan Distance* y la distancia Euclidiana, ya que para $p = 1$, la expresión anterior coincide con *Manhattan Distance* y para $p = 2$, coincide con la distancia Euclidiana.

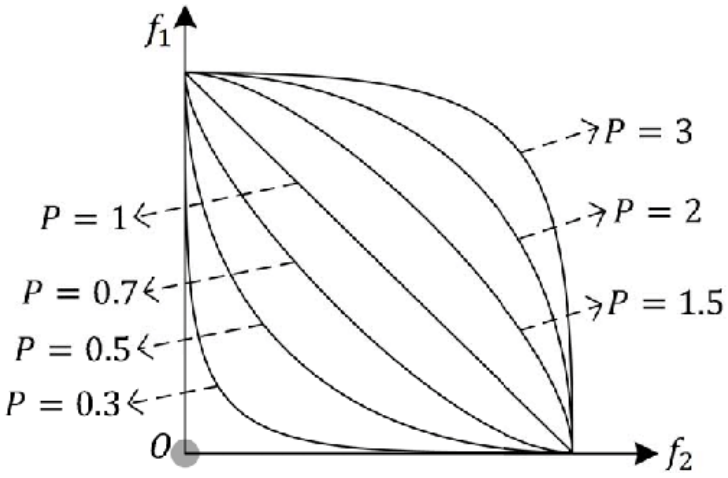


Figura 2.12: Ejemplo de *Minkowski Distance* con distintos valores de p[33].

En la figura 2.12 se muestra un ejemplo de las líneas de contorno correspondientes a la *Minkowski Distance* entre dos puntos, con distintos valores de p.

En esta medida, p es un hiperparámetro, por lo que sería necesario probar distintas configuraciones para encontrar un valor apropiado para este parámetro. Tener un hiperparámetro agrega cierto grado de dificultad, sin embargo, la posibilidad de probar distintas configuraciones para el parámetro p otorga gran flexibilidad para utilizar esta medida en diversos casos de uso.

Propuestas novedosas de medidas de distancias. Para el caso de medir distancias entre series de tiempo, Wang et al. propusieron una medida de distancia basada en área[32], que permite medir la similitud entre dos series de tiempo unidimensionales mediante la comparación de sus formas. Por otra parte, Jiang et al. propusieron la *Maximum Shifting Correlation Distance* (MSCD)[18] como una medida que pretende lidiar con las alteraciones de las series de tiempo, tanto en su fase, como en amplitud, de manera más precisa y eficiente con respecto a las tradicionales medidas de distancia.

Estas y otras novedosas medidas podrían ser estudiadas a futuro para aplicarlas al caso del presente trabajo, sin embargo, se utilizarán únicamente las medidas de distancia previamente mencionadas, ya que tienen mayor antigüedad y han sido ampliamente aplicadas en diversos campos de estudio.

Estudios comparativos. Por último, cabe señalar que se han hecho numerosos estudios comparativos entre las distintas medidas, como por ejemplo, el estudio de Kianimajd et al., donde compararon medidas de distancia para series de tiempo, en el área de la psicología[19].

Si bien estos estudios aportan una buena base para comprender las aplicaciones de cada medida de distancia, cada nuevo proyecto de ciencia de datos es distinto a los demás, como se señaló anteriormente. Por ende, los resultados y conclusiones encontradas probablemente varíen para cada caso, ya que no se puede llegar a una conclusión definitiva y generalizada para todos los casos.

2.5. Métricas de desempeño

Para evaluar los distintos modelos de *forecasting* y poder comparar entre ellos, existen métricas de desempeño que ayudarán a escoger aquel modelo que presente el mejor resultado entre todos.

2.5.1. Mean Absolute Error (MAE)

El MAE o error absoluto medio es una métrica típica para evaluar el *forecasting* de series de tiempo. Se calcula de la siguiente manera:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |x_i - \hat{x}_i|}{N} \tag{2.11}$$

2.5.2. Mean Absolute Percentage Error (MAPE)

El MAPE o error medio de porcentaje absoluto es otra métrica típica para evaluar el *forecasting* de series de tiempo. Es similar al MAE, pero su valor expresa un porcentaje. Se calcula de la siguiente manera:

$$\mathrm{MAPE} = \frac{100}{\mathrm{N}} \times \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{|\mathrm{x_i} - \hat{\mathrm{x}}_\mathrm{i}|}{\mathrm{x_i}} \tag{2.12}$$

2.5.3. Root Mean Squared Errors (RMSE)

Para evaluar modelos que realizan pronósticos numéricos existe la métrica conocida como *Root Mean Squared Errors* (RMSE), que corresponde a la desviación estándar de los errores de las predicciones. La RMSE da una noción de qué tan concentrados están los datos reales observados en relación con la línea de mejor ajuste (pronóstico) y se calcula de la siguiente manera:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{\mathrm{i}=1}^{n}(\hat{y}_\mathrm{i} - y_\mathrm{i})^2} \tag{2.13}$$

Donde $\hat{y}_\mathrm{i}$ corresponde al valor pronosticado, y_i corresponde al valor real observado y n es el número total de observaciones disponibles para analizar.

Capítulo 3

Predicción de Demanda

Este capítulo servirá al propósito de comprender en mayor detalle el problema de la predicción de demanda (conocido en la literatura como *demand forecasting*), presentando este problema en el contexto de "La Empresa".

3.1. Situación actual de "La Empresa"

A continuación se entrará en detalle en la situación actual de "La Empresa" que fue mencionada en la sección 1.1.2. Se describirá el problema y la forma en la que se ha intentado resolver, mencionando las características del modelo predictivo utilizado hasta ahora, sus falencias y las oportunidades de mejora que presenta ante la disponibilidad de los datos históricos de venta de productos. También se hablará sobre la relevancia del problema planteado y cómo se espera que sea la solución.

3.1.1. Problema y oportunidad

La distribución ineficiente de productos a sus tiendas de venta genera dos situaciones subóptimas: *sobre-stock*, cuando después de una temporada de venta sobraron productos en las tiendas; y quiebre de *stock*, cuando la cantidad de productos en el inventario no bastan para cumplir con la demanda real en las tiendas. Las consecuencias derivadas de estas planificaciones de distribución ineficientes implican, por una parte, en el escenario del *sobre-stock*, incurrir en liquidaciones y vender a precios menores que los originales, así como desaprovechar la oportunidad de realizar ventas más ajustadas en menor tiempo; y por otra parte, en el caso del quiebre de *stock*, además de desperdiciar la oportunidad de vender más en una tienda con mayor demanda, incluso se arriesga la imagen de la empresa al no tener disponibilidad de productos y que sus compradores deban satisfacer sus necesidades en otros lugares. Por este motivo, es necesario que una empresa de *retail* cuente con una planificación de reposición y distribución de productos eficiente. Ahora, para poder tener una adecuada planificación es necesario contar con una adecuada predicción de demanda de productos, lo cual es el foco de investigación del presente trabajo.

En el contexto de "La Empresa", esta distribuye sus ventas a más de 40 tiendas presentes

en el país, abarcando no solo una notable concentración en la Región Metropolitana, sino también una gran presencia en regiones, desde Arica a Punta Arenas. Para afrontar este gran volumen de ventas, existe todo un equipo de personas trabajando para realizar la mejor planificación de distribución posible, lo que corresponde al Área de *planning*.

Dicho equipo ha desarrollado un modelo predictivo que tiene capacidades limitadas, ya que realiza su pronóstico de ventas en base a una agrupación manual de productos y a otras variables externas. Por otro lado, existen componentes temporales que son desaprovechadas: los datos de venta históricos y la semejanza de sus comportamientos. Se ha podido observar que existen productos que presentan una similitud en sus comportamientos de venta históricos, lo cual puede deberse, por una parte, a que sean productos complementarios, es decir, que se compran simultáneamente en una misma visita a la tienda, tales como calzoncillos y calcetines, y por otra parte, puede deberse a la semejanza de las características físicas de los productos, tal como se ilustra en la figura 3.1.

Figura 3.1: Ejemplo de jeans similares vendidos en temporadas distintas.

Esta imagen muestra un caso típico y característico de los productos: existen productos muy similares a simple vista, pero que representan SKUs distintos en la base de datos, ya que sólo por el hecho de ser un producto de una nueva temporada o una nueva colección, se crea un código nuevo, a pesar de que a ojos de un consumidor promedio, sean el mismo producto. Aún siendo SKUs distintos, muchas veces esta similitud de productos implica una similitud en su comportamiento de ventas. Además de esto, otra característica de los productos es que "viven" y "mueren" en un mismo periodo o año comercial, debido a que entran en vigencia y luego quedan obsoletos en un mismo año comercial, a veces reapareciendo con ligeras variaciones, como un borde de distinto color, por ejemplo.

La semejanza en los comportamientos de venta y la abundancia de datos históricos de venta presenta una gran oportunidad para ser aprovechada en el problema del pronóstico de demanda. Esta oportunidad es explorada y analizada en los capítulos 5 y 6.

3.1.2. Modelo predictivo vigente y sus defectos

La solución actual que se utiliza en "La Empresa" corresponde a un modelo para hacer predicciones (*forecasting*) de ventas. La idea general de este modelo se ilustra en la siguiente figura.

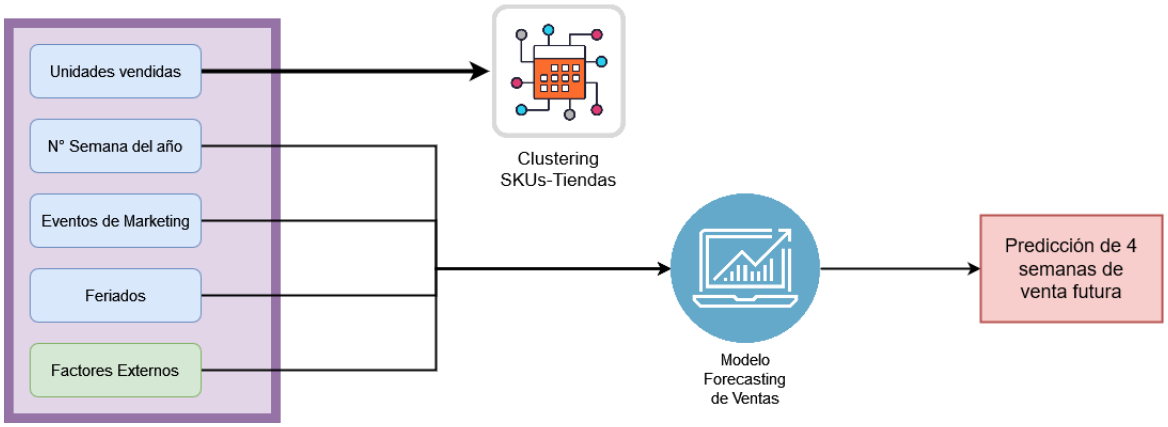


Figura 3.2: Esquema del modelo predictivo actual de "La Empresa".

En la figura 3.2 se aprecia que este modelo contempla, por una parte, una serie de variables independientes de los productos, tales como el número de semana del año, el calendario de marketing (también conocido como año comercial, el cual homologa las fechas actuales a su equivalente en las semanas comerciales estándar), fechas de feriados en Chile y condiciones externas a las ventas, por ejemplo, condiciones ambientales como la temperatura máxima y mínima y la cantidad de lluvia.

Por otra parte, este modelo también utiliza una agrupación o *clustering* **manual** de productos basada completamente en la jerarquización de los productos, considerando sus tipos de conjuntos (por ejemplo, *tops* o *outerwear*), su segmento (por ejemplo, hombre, mujer o niño) y su marca (ya sean externas o propias), y no se basa en un análisis más detallado del comportamiento de ventas de productos similares. Estos componentes entran al modelo para que este arroje un pronóstico de ventas con una ventana de 4 semanas en el futuro, el cual se realiza a nivel SKU-tienda.

En cuanto al algoritmo utilizado por el modelo predictivo, este corresponde a *XGBoost*[1], un algoritmo popularmente utilizado en predicciones de ventas[8], debido a que es una implementación rápida y eficiente de *grandient boosting* que sirve para mejorar la capacidad predictiva y el rendimiento de estos modelos. El esquema de este modelo predictivo se aprecia en la figura 3.3.

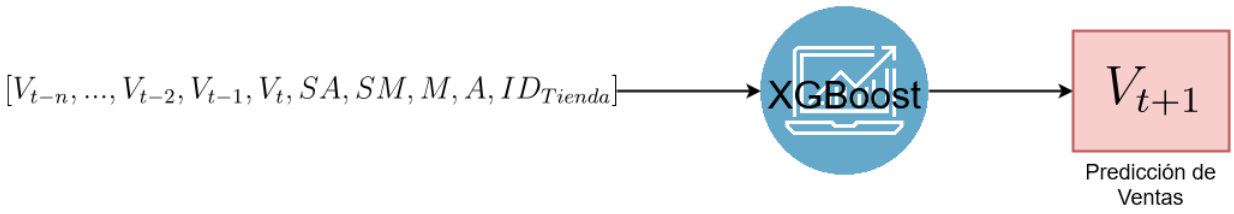


Figura 3.3: Esquema del algoritmo *XGBoost* para el pronóstico de demanda de "La Empresa".

Las variables del vector de características que recibe el algoritmo corresponden a:

[1]XGBoost: e**X**treme **G**radient **B**oosting. Algoritmo ampliamente usado en problemas de clasificación, regresión y *forecasting*. Está basado en árboles de decisión y en *boosting*.

- V_t: ventas del SKU en el periodo actual.
- SA: número de semana del año.
- SM: número de semana del mes.
- M: número del mes.
- A: número del año.
- ID_{Local}: identificador de la tienda donde se vende el SKU.

Además, el *output* V_{t+1} corresponde a la predicción para el siguiente periodo.

Uno de los defectos de este modelo es que realiza un *forecast* demasiado agrupado, el cual se basa en la jerarquía de los productos y no en casos específicos de SKUs, y tampoco se hace tomando en consideración otros SKUs con comportamientos de ventas similares.

Por último, con respecto al desempeño de este modelo, se ha registrado MAPE menor al 20 %, por lo que realiza un pronóstico de ventas aceptable en líneas generales, pero que no llega al nivel de detalle, SKU, que se requiere y que se aborda en el presente trabajo, por lo tanto no son resultados comparables y cada pronóstico cumple con su objetivo específico dentro de "La Empresa".

3.1.3. Relevancia

El problema planteado de planificación eficiente a partir de predicción de demanda es un tema que se enfrenta de manera global en el sector del *retail*. Una distribución ineficiente de los productos desde los centros de distribución hacia las tiendas donde se venden no solo impacta en las ganancias de un empresa, sino también otros problemas colaterales como ver afectada la imagen de la empresa por la falta de disponibilidad de productos. De hecho, un estudio[13] realizado en Europa indica que ante un quiebre de *stock*, un 37 % de los consumidores optará por comprar otra marca del producto que busca y no encuentra, un 21 % preferirá ir a otra tienda y un 9 % decidirá no comprá nada (el restante 33 % opta por otro producto similar al que busca dentro de la misma tienda, lo cual no representa pérdidas). Además este estudio estima que en la industria del *retail* europeo se pierden aproximadamente 4.000 millones de euros solo considerando el 9 % de consumidores que desiste de su compra.

El estudio llevado a cabo en el presente trabajo contribuye a la comprensión y posible mejora de las soluciones existentes en el problema de *forecasting* de demanda bajo un enfoque de predicción en base a la cercanía entre series temporales de ventas de productos con comportamientos similares. Este estudio podría ser replicado en cualquier empresa del *retail* que cuente con abundancia de datos históricos de venta: cientos o miles de productos y sus registros de ventas con una componente temporal que abarque al menos 3 años de ventas, ingresadas de manera diaria.

3.1.4. Especificaciones y alcances de la solución

El presente trabajo busca realizar un estudio que permita validar la idea de que es posible mejorar las predicciones de ventas mediante un agrupamiento y/o emparejamiento de series temporales de los datos de ventas históricos de productos. En este sentido, el requisito de la solución a construir es que esta permita generar nuevo conocimiento sobre el comportamiento

de las series temporales de ventas y su potencial predictivo en el problema de la predicción de demanda, en particular, para corroborar la factibilidad de hacer predicciones a partir de los datos históricos de ventas de productos de una misma categoría y comportamientos similares en el tiempo. En este trabajo no se pretende generar una nueva herramienta de pronóstico de demanda, ni tampoco busca mejorar directamente el modelo de *forecast* actual de "La Empresa".

Las características deseadas de esta solución contemplan una serie de gráficos, métricas y visualizaciones que impliquen directamente conclusiones claras y categóricas sobre la similitud de las series temporales y de cómo esto se relaciona con el pronóstico de ventas en la industria del *retail*. Se espera también que la solución conlleve un análisis de las características y cualidades de los productos para los cuales este estudio obtuvo resultados favorables, así como una descripción de los productos para los cuales este estudio no resulta factible. También se espera que la solución permita responder si es que es mejor implementar este estudio a nivel de SKU o de estilo y si es mejor implementarlo para una tienda en específico o para una zona geográfica que agrupe las tiendas.

Capítulo 4

Descripción de los Datos y Análisis Exploratorio

Como todo proyecto de ciencia de datos, fue necesario realizar en primer lugar un análisis exploratorio de los datos, con el objetivo de entender y conocer los datos con los cuales se trabaja. Esto permite comprender mejor el problema y el contexto para restringir y acotar el estudio y la solución propuesta a aquel segmento de los datos disponibles que presentara la mayor cantidad de información útil.

A continuación se mostrará la información principal de los *datasets* ocupados, así como los gráficos relevantes y estadísticas descriptivas de los mismos. Los datos usados son registros de ventas de SKUs en las categorías de vestuario y electrodomésticos; y fueron obtenidos consultando a la base de datos que posee "La Empresa" hospedada en la *nube*, a través de *BigQuery*[1] y procesadas en el presente trabajo como archivos CSV.

4.1. Descripción de los *datasets* y sus atributos

Para recordar la categorización de los productos mencionada brevemente en la Introducción (1.1.1), a continuación se muestra un esquema que resume la estructura de los productos.

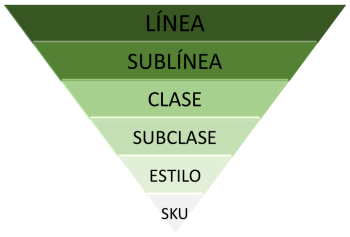


Figura 4.1: Esquema de jerarquías de los productos.

En el esquema de la figura 4.1, vemos que el nivel más general para describir a un producto es la línea, por ejemplo, "ELECTRO HOGAR", y el nivel más específico es SKU, lo cual

[1] *BigQuery* es el almacén de datos multinube de alta escalabilidad, rentable y sin servidor de *Google Cloud Platform*. Más información disponible en `https://cloud.google.com/bigquery/`.

corresponde a un código. Por otro lado, no es posible categorizar los productos en base a sus características físicas, tales como el material del que están hechos, sus medidas, colores, entre otros, ya que esta información no está presente en todos los *datasets*, ni está disponible para todos los productos, por lo tanto, esos datos no serán considerados.

A continuación, se presenta en la tabla 4.1 el detalle de las variables presentes en cada uno de los registros de los *datasets* con los que se trabajó.

Variable	**Tipo**	**Descripción**	**Observaciones**
id_boleta	Numérica	Identificador de la boleta en la que se vendió el SKU	Variable no usada
id_dia_bol	Fecha	Fecha en la que se vendió el SKU	Usada para comparar los productos cuantitativamente
id_tienda_bol	Categórica	Código de la tienda donde fue realizada la venta	
desc_tienda_bol	Texto	Nombre de la tienda donde fue realizada la venta	Usada para filtrar los *datasets*
id_sku	Categórica	Código del SKU que fue vendido	Usada para identificar las series temporales de SKUs
cod_temporada	Categórica	Código de la temporada en la que se originó el SKU	No corresponde necesariamente a la temporada en la que se vendió
id_estilo	Categórica	Código del estilo al que pertenece el SKU	Usada para identificar las series temporales de estilos
desc_estilo	Texto	Descripción del estilo al que pertenece el SKU	Usada para comparar los productos cualitativamente
desc_produc	Texto	Descripción del SKU	Usada para comparar los productos cualitativamente
id_marca	Categórica	Nombre de la marca a la que pertenece el SKU	Usada para comprar los productos cualitativamente
id_linea	Categórica	Código de la línea a la que pertenece el SKU	
desc_linea	Texto	Descripción de la línea a la que pertenece el SKU	
id_sublinea	Categórica	Código de la sublínea a la que pertenece el SKU	
desc_sublinea	Texto	Descripción de la sublínea a la que pertenece el SKU	
id_clase	Categórica	Código de la clase a la que pertenece el SKU	
desc_clase	Texto	Descripción de la clase a la que pertenece el SKU	Usada para comparar los productos cualitativamente
id_subclase	Categórica	Código de la subclase a la que pertenece el SKU	
desc_subclase	Texto	Descripción de la subclase a la que pertenece el SKU	Usada para comparar los productos cualitativamente
unidades_vendidas	Numérica	Cantidad del SKU que fue vendida en una misma transacción	Mayor o igual a 1
precio_venta	Numérica	Valor al que fue vendido el SKU	Expresado en pesos chilenos

Tabla 4.1: Variables presentes en cada conjunto de datos utilizado

Se utilizaron 4 *datasets*, los cuales se pueden separar conceptualmente en dos: vestuario y electrodomésticos. Para vestuario, se tienen los *datasets* de Ropa Interior y de Pantalones, mientras que para electrodomésticos se cuenta con los *datasets* de Lavadoras y de Microondas.

Cada uno de estos *datasets* abarca fechas que van desde "2019-01-01" hasta "2021-12-31" y por tanto, contienen la información de 3 años completos de ventas: 2019, 2020 y 2021.

Por último, cabe señalar que de los 4 *datasets* disponibles, la prueba de concepto fue realizada con el *dataset* de Pantalones, mientras que el resto del desarrollo y la simulación de predicción de ventas se realizó utilizando los 4 *datasets*.

4.2. Confidencialidad

A partir de este punto y durante el resto del presente trabajo, cada vez que se hable con respecto a los productos, se verán anonimizados los datos referentes a los códigos identificadores de SKUs, de Estilos, las marcas y sus precios, lo cual se realizará ya sea aplicando un *hash* para enmascarar, o bien ocultando información ocupando el símbolo "#", o bien reemplazando por un nombre genérico. Además, cada vez que se hable de cantidades, como por ejemplo, unidades vendidas o número de SKUs distintos, estas serán expresadas en términos de proporciones sobre un intervalo de valores, ocupando sobre los datos originales un escalador de tipo Mín-Máx[2].

Cada vez que estas modificaciones hayan sido aplicadas, se reportará en el presente informe como **s.e.p.c.**, significando "sin especificar, por confidencialidad".

La finalidad de estas transformaciones es ocultar la información sensible para "La Empresa" y velar por la confidencialidad de su información. Es importante destacar que a raíz de estos cambios de nomenclaturas y escalas, las observaciones y conclusiones obtenidas no se verán alteradas.

4.3. Distribución de los datos y estadísticas descriptivas

Para tener una idea del volumen de los datos de ventas, se presenta a continuación un gráfico que muestra la cantidad total de productos vendidos en cada *dataset*.

[2]Usando la herramienta de preprocesamiento *MinMaxScaler*, de la librería *Scikit-Learn*: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`.

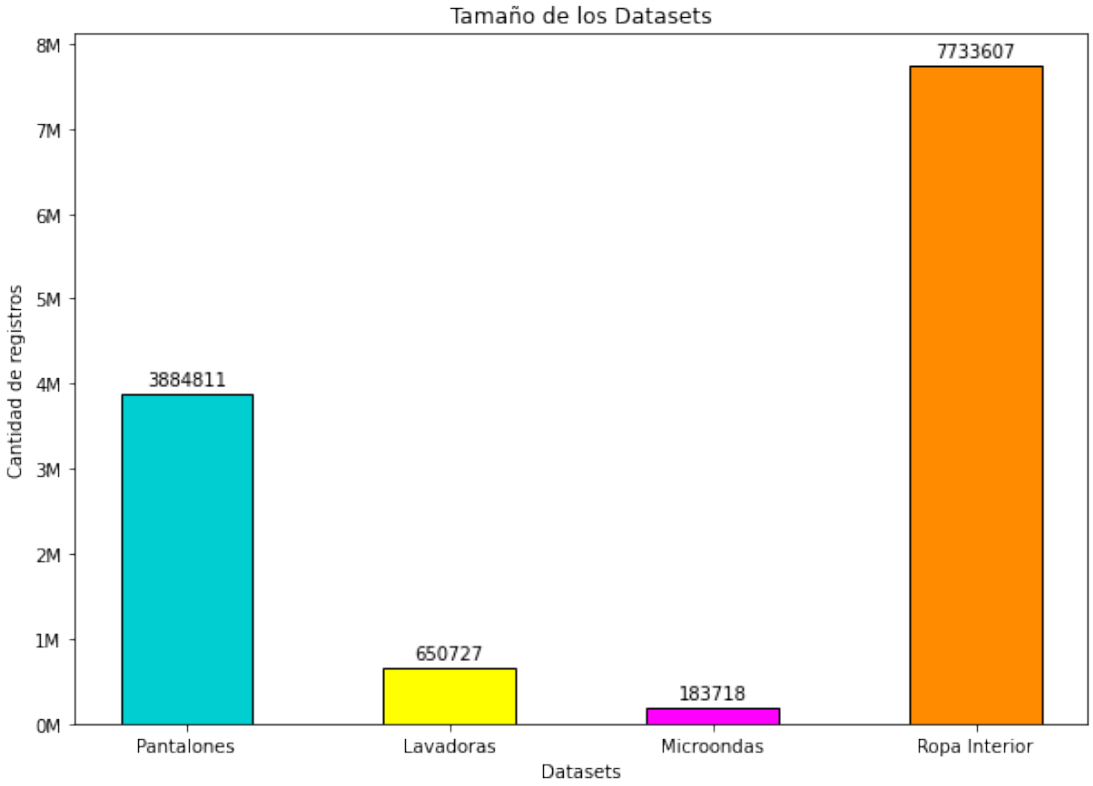


Figura 4.2: Gráfico comparativo del volumen de datos presente en cada *dataset* utilizado.

En la figura 4.2 se aprecia que el *dataset* de Ropa Interior es el más grande en datos de ventas, teniendo más de 7.7 millones de registros, razón por la cual fue escogido para realizar los estudios finales y las predicciones de ventas, como representante de los *datasets* de vestuario. De igual manera, como representante de la categoría de electrodomésticos, el *dataset* más grande resultó ser el de Lavadoras, con más de 650,000 registros. Cabe señalar que los colores usados en este gráfico para identificar a cada *dataset* se mantienen en las próximas visualizaciones, de tal manera de mantener la coherencia y facilitar la asociación de los gráficos a sus *datasets* respectivos.

En cuanto a las variables que sirven para categorizar y comparar cualitativamente los productos, las variables de `desc_linea` y `desc_sublinea` fueron utilizadas para acotar los datos originales y generar los *datasets* a partir de toda la información disponible en la base de datos, y por tanto, en cada *dataset* estas variables solo presentan un valor por cada uno: para Ropa Interior la línea y sublínea son, respectivamente, "HOMBRES" y "ROPA INTERIOR"; para Lavadoras la línea es "ELECTRO HOGAR" y la sublínea es "LAVADO"; para Pantalones, la línea es "JUVENIL HOMBRES" y la sublínea es "JEANS JUVENIL HOMBRES"; por último, para Microondas, la línea es "ELECTRO HOGAR" y la sublínea es "ELECTRODOMESTICOS". Dado que no hay una interacción directa entre los *datasets*, estas variables son únicamente referenciales y no fueron utilizadas.

Continuando con las variables `desc_clase` y `desc_subclase`, estas también sirven a un propósito referencial, sin embargo de todas maneras fueron utilizadas para comparar cualitativamente los productos, ya que presentan valores distintos para cada *dataset*: en Ropa Interior existen 7 clases y 25 subclases (tales como "PIJAMAS" y "PIJAMA CORTO HOMBRE",

respectivamente); en Lavadoras existen 8 clases y 13 subclases (por ejemplo, "LAVADORAS CARGA SUPERIOR" y "DE 7KG/16 LBS HASTA 10KG/22 LBS"); en Pantalones existen solo 2 clases y 2 subclases, ambas son "JEANS JUVENIL HOMBRES" y "JEANS MODA", por lo que no son de gran utilidad a la hora de comparar productos; por último en Microondas solo existe la clase "MICROONDAS" y existen 4 subclases (por ejemplo, "H.MICROONDAS 21-27 LTS"), por lo que, nuevamente, no representan gran utilidad.

La variable `desc_estilo` fue particularmente útil, ya que engloba productos de un mismo tipo, con pequeñas variaciones, tales como la talla. Por ejemplo, en Pantalones, algunos de los SKUs asociados al estilo "JEANS JUV LEE 134418" son: "JEANS LEM CHICAGO L42CGO0. 510. 44", "JEANS LEM CHICAGO L42CGO0. 510. 46" y "JEANS LEM CHICAGO L42CGO0. 510. 52", por lo que se entiende que, hasta cierto punto, representan el "mismo" producto, lo cual sugiere que tiene sentido realizar predicciones de ventas a nivel de estilo más que a nivel de SKU, tal como veremos en el próximo capítulo. La relación entre la cantidad de estilos y de SKUs únicos presentes en los *datasets* se aprecia en el siguiente gráfico.

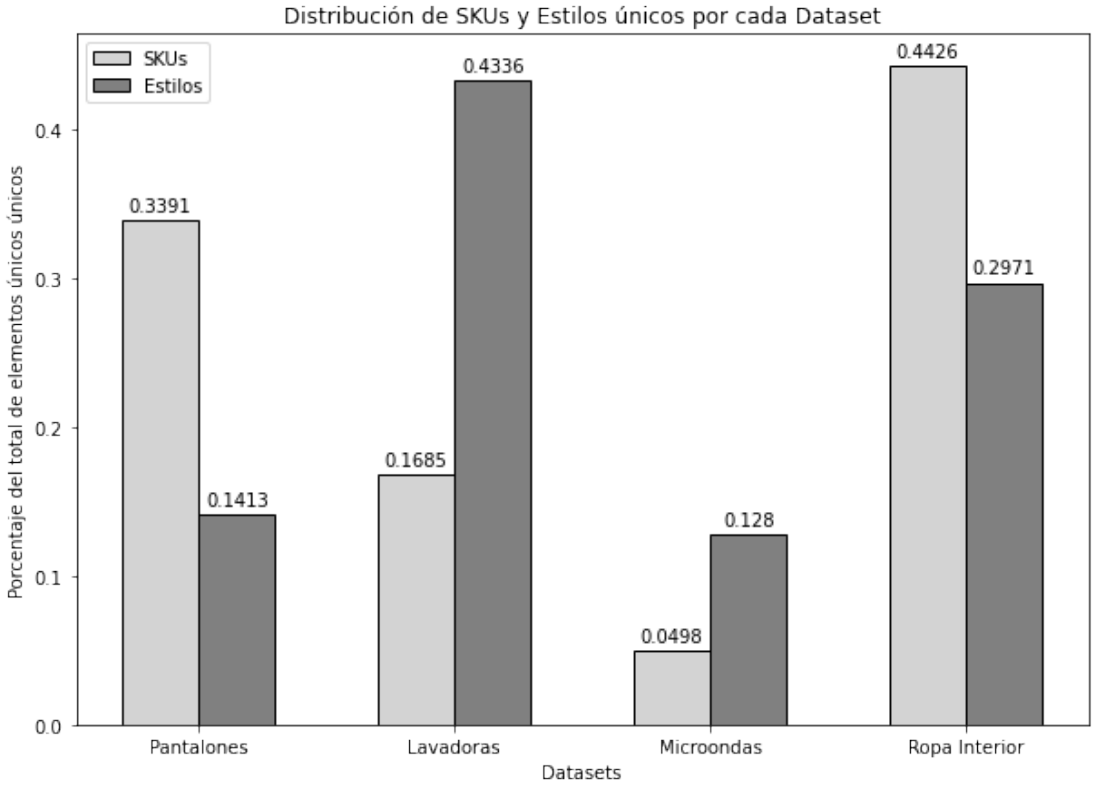


Figura 4.3: Gráfico comparativo de la distribución de estilos y SKUs únicos en los *datasets*.

En la figura 4.3 se observan los valores porcentuales por *dataset* sobre el total de SKUs distintos, cuyo valor total está entre 35,000 y 45,000 (s.e.p.c.) y también sobre el total de estilos distintos, cuyo valor se encuentra entre 10,000 y 20,000 (s.e.p.c.). Se aprecia que los *datasets* en los que se puede sacar mayor provecho de la variable de estilo son los de vestuario, no así como en los *datasets* de electrodomésticos, donde vemos que existe la misma cantidad de estilos y de SKUs únicos, lo que implica que en la práctica sea indiferente referirse a estos productos ya sea por su `id_estilo` o su `id_sku`.

La variable `id_marca` posee gran diversidad de valores: en Ropa Interior se tienen 137 marcas distintas; en Lavadoras existen 44 marcas; en Pantalones existen 45 marcas; y en Microondas se tienen 33 marcas. Si bien esta variable puede no decir mucho de un producto físicamente, igual tiene un impacto indirecto en el comportamiento de ventas de un producto, razón por la cual esta variable es utilizada para realizar comparaciones cualitativas.

Por otro lado, la variable `desc_tienda_bol` es de suma relevancia, ya que implica un componente geográfico en las ventas de los productos que impacta directamente en los comportamientos de ventas. Por ejemplo, el volumen y la frecuencia de las de ventas de parkas es absolutamente distinto en la zona norte en comparación con la zona sur del país. De esta manera, esta variable fue ampliamente utilizada, sobre todo para filtrar los *datasets* y acotar los estudios realizados. En la siguiente sección se entrará en detalle en la distribución de ventas de cada *dataset* según cada tienda y zona geográfica.

Finalmente, la variable `precio_venta` sirvió para comparar cuantitativamente entre los productos de un mismo *dataset*, sin embargo, existen numerosos casos en los que en un mismo día y en una misma tienda, un mismo SKU se vendió a diferentes precios. Esto ocurre cuando, por ejemplo, un consumidor realiza una compra con una tarjeta de crédito que presenta descuento para una cierta marca o tipo de producto. Esta variación de precios de un mismo SKU sugiere la idea de utilizar el promedio del valor de venta del SKU en un mismo día y tienda, y también ocupar el valor modal, tal como se verá en el siguiente capítulo.

Para explorar y comparar los distintos valores del precio de venta en los distintos *datasets*, se generaron los siguientes histogramas de precios para cada conjunto de datos de las categorías de vestuario y electrodomésticos, cada uno en sus propias escalas.

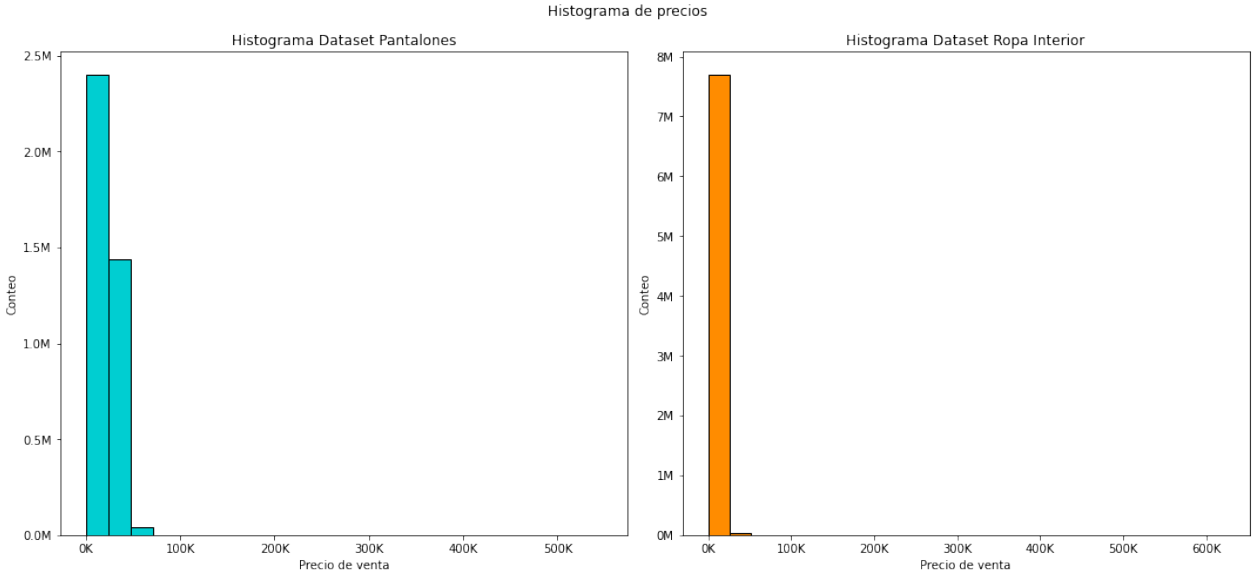


Figura 4.4: Histogramas de precios de los *datasets* de vestuario.

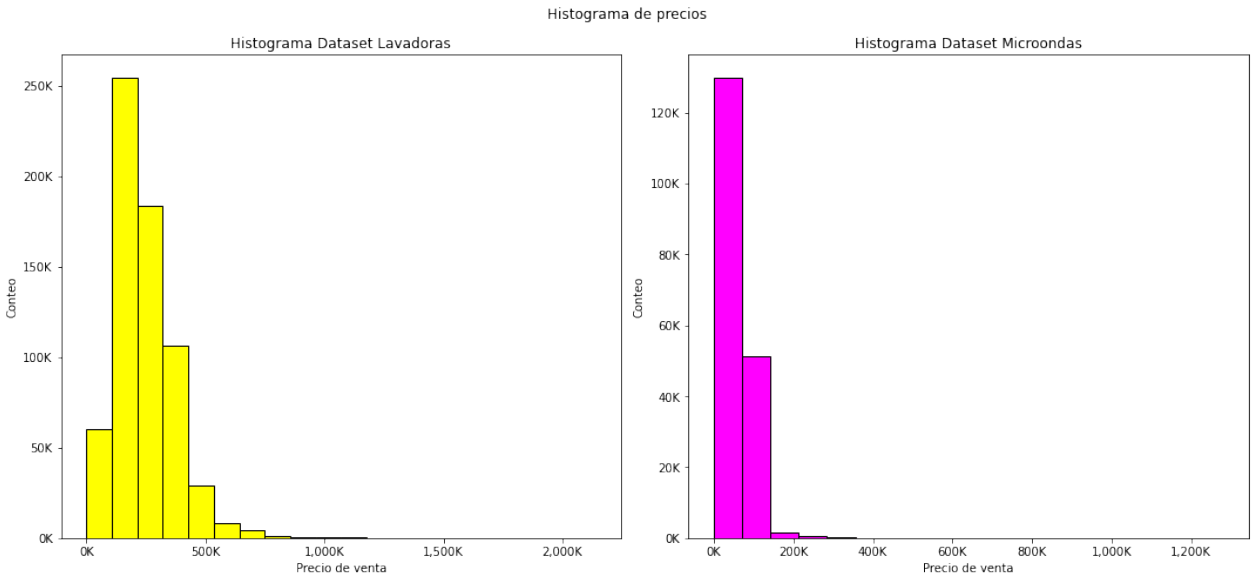


Figura 4.5: Histogramas de precios de los *datasets* de electrodomésticos.

En la figura 4.4 vemos que los *datasets* de Pantalones y de Ropa Interior, concentran sus precios en el rango de 0 a 75,000 pesos. Los histogramas se ven como tal ya que existe una pequeña porción de datos anómalos (*outliers*) con valores muy alejados del resto. Por este motivo, se escoge el valor de 75,000 como punto de corte para la variable `precio_venta` en los *datasets* de vestuario. Aplicado este corte, se observa que en Pantalones se pierden solo 80 productos, correspondientes al 0.002 % de los registros de ventas, mientras que en Ropa Interior, se pierden tan solo 31 SKUs, correspondientes al 0.0004 % de registros de ventas.

Luego de aplicar el corte de precio indicado, se generó el siguiente diagrama de caja para visualizar mejor las estadísticas de esta variable.

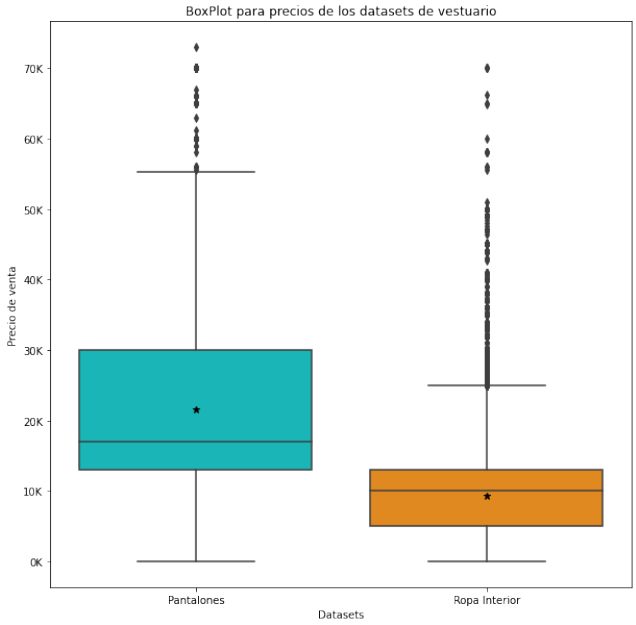


Figura 4.6: Diagrama de caja de los precios de los *datasets* de vestuario.

En el diagrama de la figura 4.6, vemos que en el *dataset* de Pantalones, el último cuartil llega hasta aproximadamente 55,000 pesos, cifra a partir de la cual existe una pequeña cantidad de valores anómalos, i.e. cuyo valor está considerablemente alejado de la gran mayoría de los datos. Por otro lado, en el *dataset* de Ropa Interior, el último cuartil llega hasta aproximadamente 25,000, con *outliers* a partir de dicho valor. La media o promedio del precio de venta de Pantalones es 21,594 y su mediana se encuentra por debajo de este valor; mientras que la media del precio de Ropa interior es 9,259 y su mediana se encuentra por arriba de este valor.

Con respecto a los *datasets* de electrodomésticos, en la figura 4.5 vemos que estos concentran sus precios en el rango de 0 a 1,000,000 de pesos para Lavadoras y de 0 a 400,000 pesos para Microondas. Nuevamente, los histogramas se ven como tal debido a la presencia de (*outliers*). Por este motivo, se escoge el valor de 1,000,000 como punto de corte para la variable `precio_venta` en el *datasets* de Lavadoras y 400,000 como punto de corte para los precios de Microondas. Luego de aplicar estos cortes, se aprecia que se pierden 608 Lavadoras, correspondientes al 0.09 % de los registros de ventas, mientras que en Microondas, se pierden solo 107 SKUs, correspondientes al 0.06 % de registros de ventas. Aplicando estos cortes de precios, se generó el diagrama de caja que se muestra a continuación.

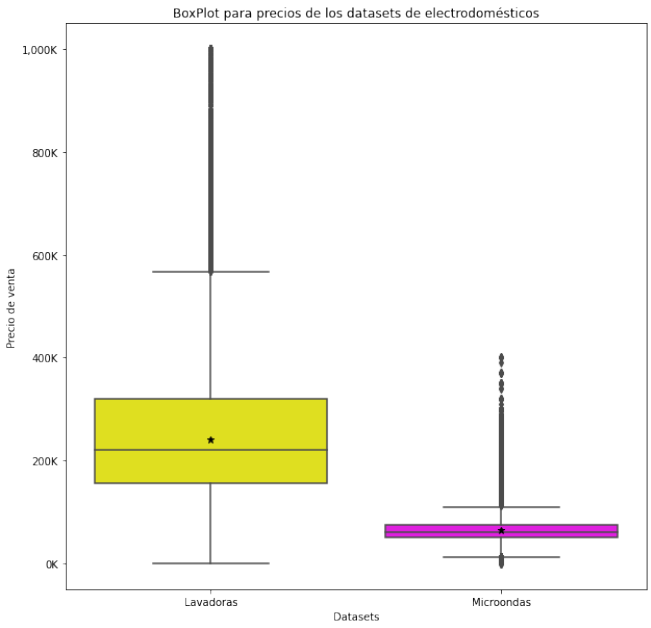


Figura 4.7: Diagrama de caja de los precios de los *datasets* de electrodomésticos.

En el diagrama de la figura 4.7, vemos que en el *dataset* de Lavadoras, el último cuartil llega hasta aproximadamente 600,000 pesos, con una gran cantidad de valores anómalos a partir de esa cifra, mientras que en el *dataset* de Microondas, el último cuartil llega hasta aproximadamente 110,000, con muchos *outliers* a partir de dicho valor. La media o promedio del precio de venta de Lavadoras es 240,093 y su mediana se encuentra por debajo de este valor; mientras que la media del precio de Microondas es 64,268 y su mediana también se encuentra por debajo de este valor.

4.4. Análisis por tiendas y zonas

Continuando con el análisis exploratorio, se investiga la distribución de ventas para las 4 zonas geográficas en las que se reparten las diferentes tiendas de "La Empresa". Las tiendas asociadas a cada zona son las siguientes, mencionadas en orden geográfico, de norte a sur del país:

- **Zona Norte**: Arica, Iquique, Calama, Mall Antofagasta, Mall Copiapó, La Serena, Ovalle, San Felipe, La Calera, Viña del Mar, Quilpué y Valparaíso.
- **Zona Centro Oriente**: La Dehesa, Alto Las Condes, Parque Arauco, Los Dominicos, Costanera Center, Lyon y Plaza Egaña.
- **Zona Centro Poniente**: Plaza Norte, Independencia, Plaza Puete, Estación Central, Arauco Maipú, Centro, Plaza Vespucio, Plaza Oeste, Tobalaba, San Bernardo, Melipilla, Rancagua y San Fernando.
- **Zona Sur**: Curicó, Talca, Chillán, El Trébol, Concepción Centro, Los Ángeles, Temuco, Pucón, Valdivia, Osorno, Puerto Montt, Castro y Punta Arenas.

4.4.1. Gráficos comparativos de los volúmenes de ventas en las zonas geográficas

A continuación se mostrarán gráficos de barra con la distribución porcentual de unidades vendidas por cada *dataset*.

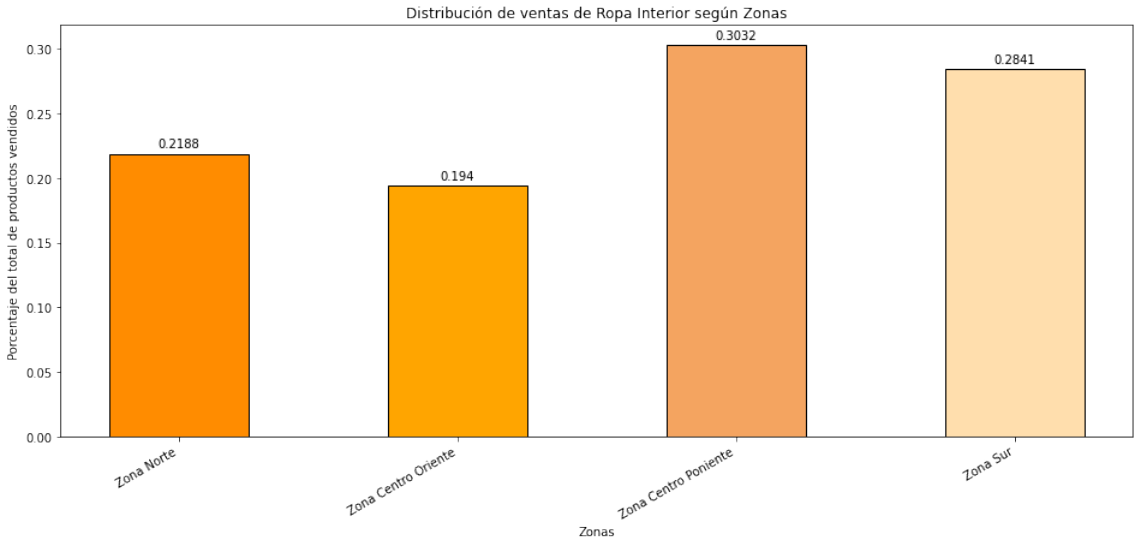


Figura 4.8: Comparación de la distribución de ventas de Ropa Interior según cada zona geográfica.

En el gráfico 4.8 notamos que para Ropa Interior, la zona Centro Oriente fue la que tuvo el menor volumen de ventas, con un 0.194 % del total productos vendidos, cuya suma se encuentra entre 5,000,000 y 7,000,000 unidades (s.e.p.c.). Por otro lado, el mayor volumen de ventas se obtuvo en la zona Centro Poniente, con un 0.3032 % del total de SKUs vendidos. Por esta razón, se ocupará la Zona Centro Poniente para realizar los estudios y predicciones de ventas.

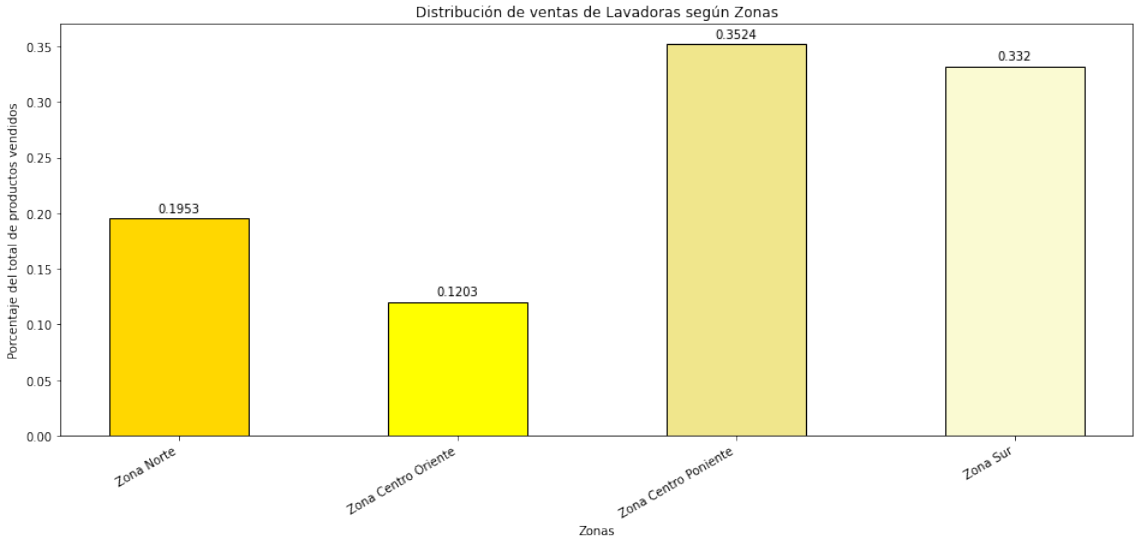


Figura 4.9: Comparación de la distribución de ventas de Lavadoras según cada zona geográfica.

Se puede apreciar en la figura 4.9 que para el *dataset* de Lavadoras, la Zona Centro Oriente fue la que tuvo el menor volumen de ventas, con un 0.1203 % del total de productos vendidos, cuya suma se encuentra entre 150,000 y 300,000 unidades (s.e.p.c.); mientras que el mayor volumen de ventas se obtuvo en la zona Centro Poniente, con un 0.3524 % del total de SKUs vendidos. Por esta razón, se ocupará la zona Centro Poniente para el desarrollo del trabajo.

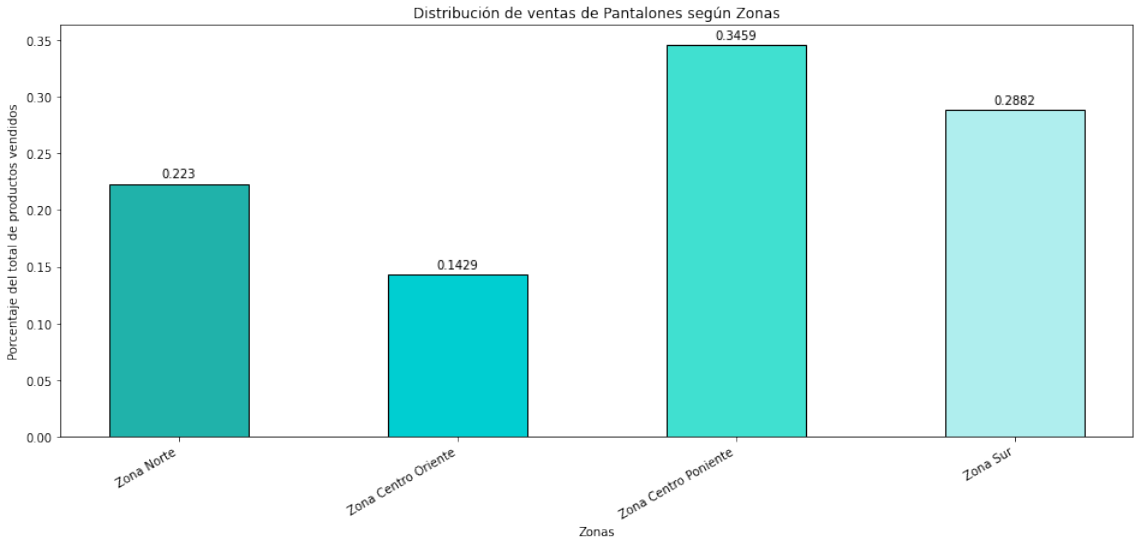


Figura 4.10: Comparación de la distribución de ventas de Pantalones según cada zona geográfica.

En el gráfico 4.10 notamos que para los Pantalones, la Zona Centro Oriente fue la que tuvo el menor volumen de ventas, con un 0.1429 % del total de productos vendidos, cuya suma se encuentra entre 2,000,000 y 4,000,000 unidades (s.e.p.c.); mientras que el mayor volumen de ventas se obtuvo en la zona Centro Poniente, con un 0.3459 % del total de SKUs vendidos. Por esta razón, se ocupará la zona Centro Poniente.

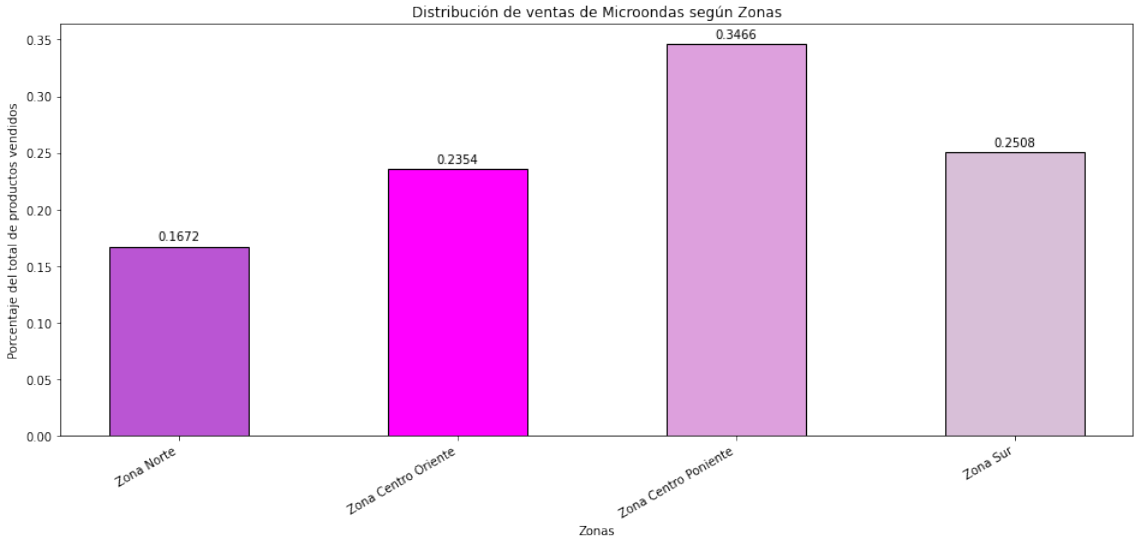
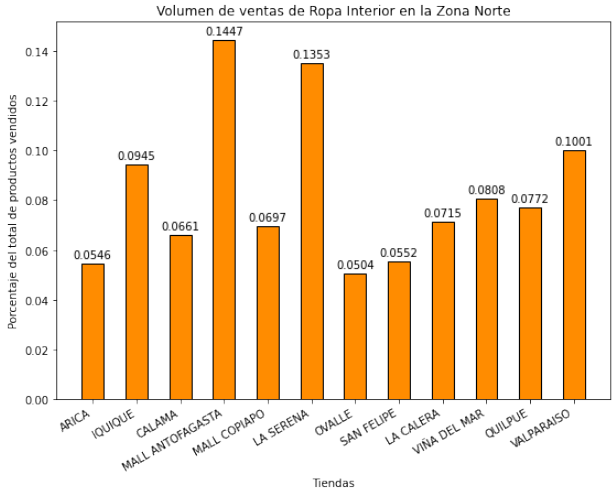


Figura 4.11: Comparación de la distribución de ventas de Microondas según cada zona geográfica.

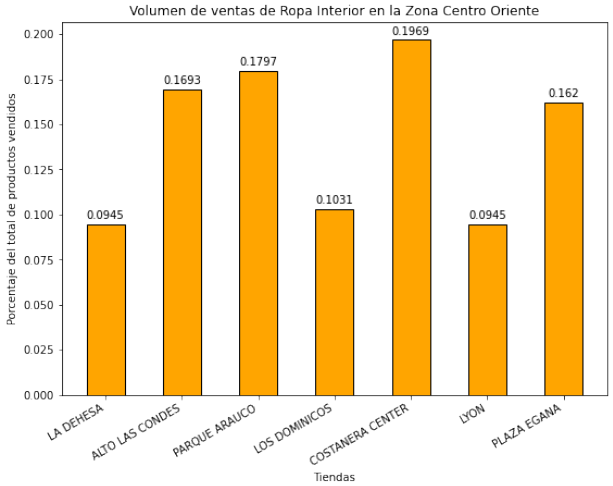
En la figura 4.11 se observa que para los Microondas, la Zona Norte fue la que tuvo el menor volumen de ventas, con un 0.1672 % del total de productos vendidos, cuya suma se encuentra entre 50,000 y 150,000 unidades (s.e.p.c.); mientras que el mayor volumen de ventas se obtuvo en la zona Centro Poniente, con un 0.3466 % del total de SKUs vendidos. Por esta razón, se ocupará la zona Centro Poniente.

4.4.2. Gráficos comparativos de los volúmenes de ventas por cada tienda

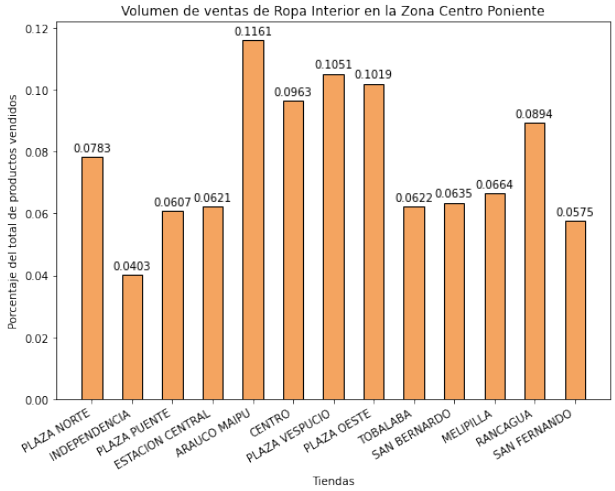
A continuación se mostrarán gráficos de barra con la distribución porcentual de unidades vendidas en cada una de las tiendas de cada zona geográfica de los *datasets*. Nuevamente, las tiendas se muestran ordenadas geográficamente.



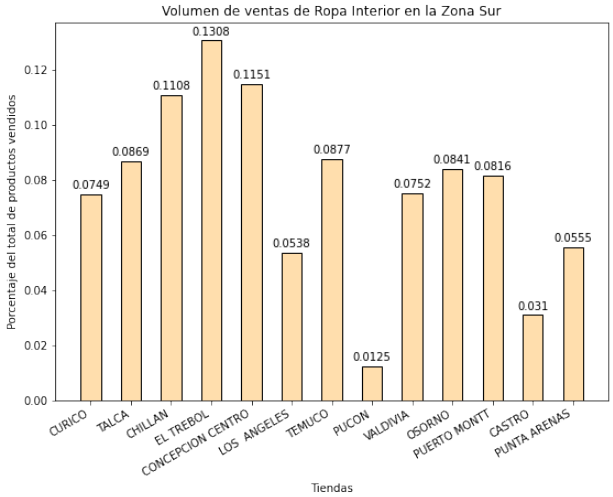
(a) Zona Norte.



(b) Zona Centro Oriente.



(c) Zona Centro Poniente.



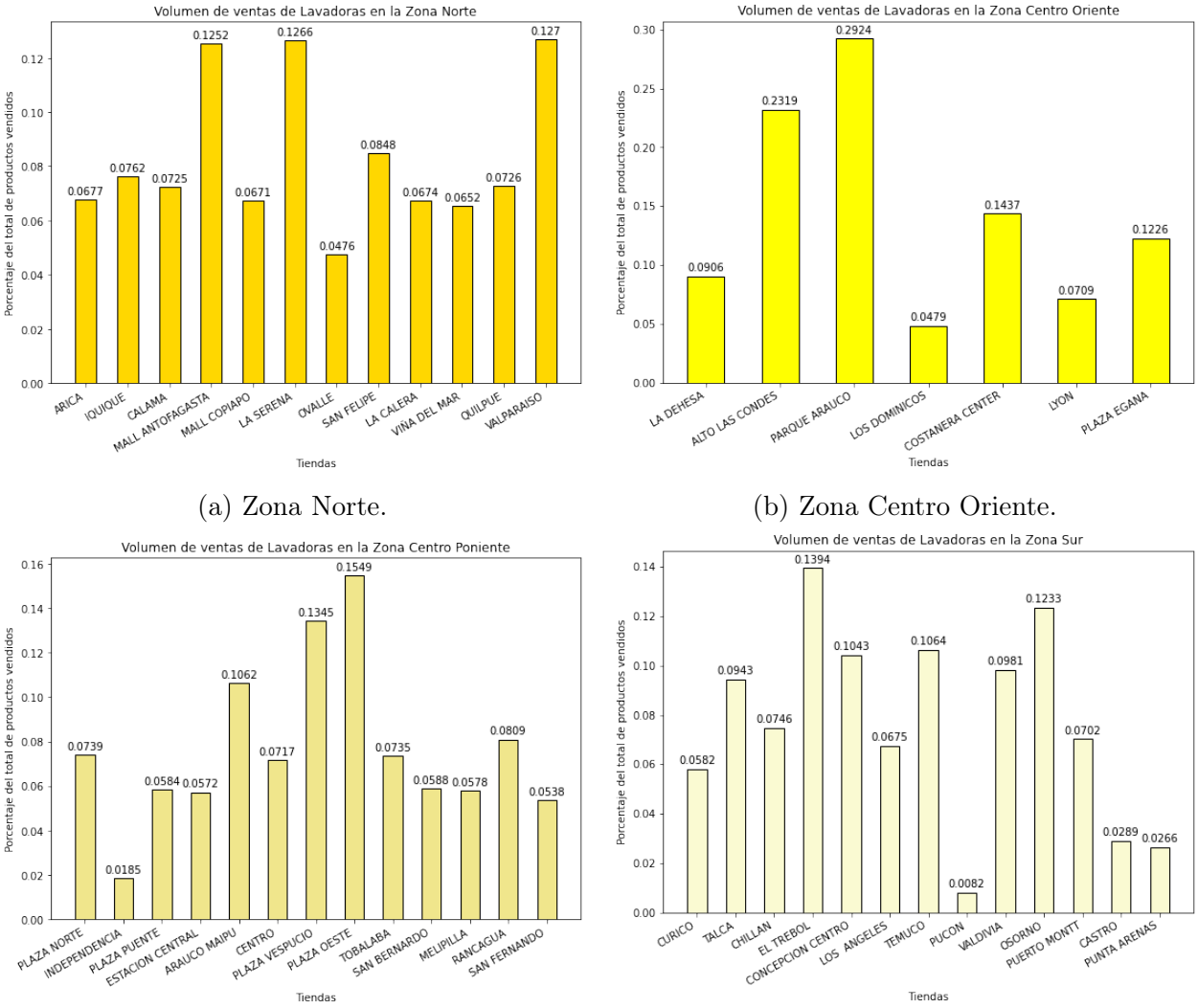
(d) Zona Sur.

Figura 4.12: Comparación del volumen de ventas en las tiendas de cada zona del *dataset* de Ropa Interior.

En la figura 4.12 vemos que en el *dataset* de Ropa Interior, en la Zona Norte, la tienda que más vendió corresponde al Mall Antofagasta con un 0.1447 % del total de productos vendidos, mientras que la que menos vendió fue la tienda Ovalle con un 0.0504 % del total de SKUs vendidos. En la Zona Centro Oriente, la tienda que más vendió corresponde al Costanera Center con un 0.1969 % del total de productos vendidos, mientras que la que menos vendió fue la tienda La Dehesa con un 0.0945 % del total de SKUs vendidos. En la Zona Centro Poniente, la tienda que más vendió corresponde a la tienda Arauco Maipú con un 0.1161 % del total de productos vendidos, y como esta corresponde a la zona de mayor cantidad de

ventas, se usará esta tienda para el desarrollo del presente estudio y predicciones a nivel de tienda. En esta Zona, el menor volumen de ventas lo tuvo la tienda Independencia con un 0.0403 % del total de SKUs vendidos. Por último, en la Zona Sur, la tienda que más vendió corresponde a la tienda El Trébol con un 0.1308 % del total de productos vendidos, mientras que la que menos vendió fue la tienda Pucón con un 0.0125 % del total de SKUs vendidos.

Continuamos con las tiendas del *dataset* de Lavadoras.



(a) Zona Norte.

(b) Zona Centro Oriente.

(c) Zona Centro Poniente.

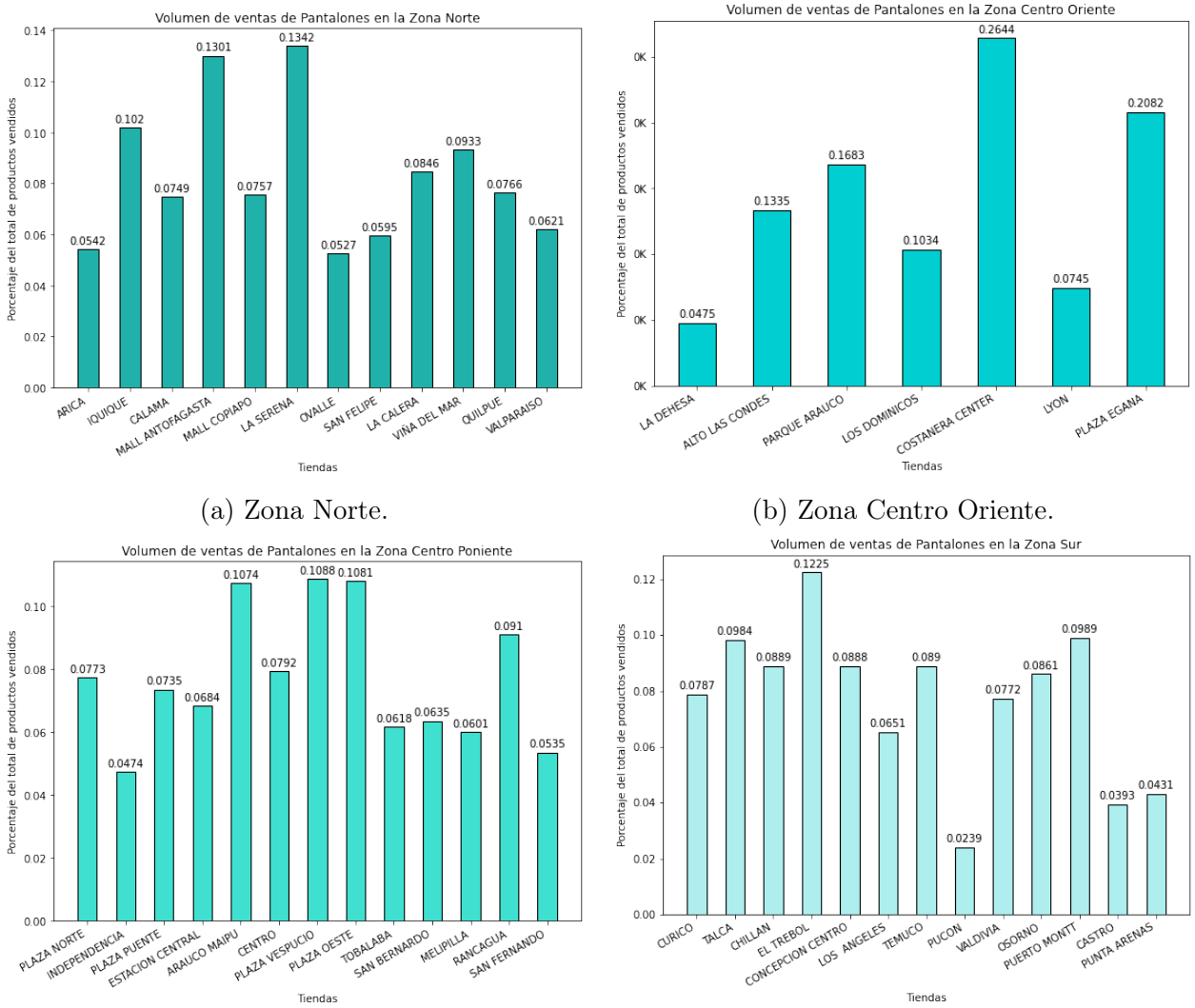
(d) Zona Sur.

Figura 4.13: Comparación del volumen de ventas en las tiendas de cada zona del *dataset* de Lavadoras.

En la figura 4.13 vemos que en el *dataset* de Lavadoras, en la Zona Norte, la tienda que más vendió corresponde a la tienda Valparíso con un 0.127 % del total de productos vendidos, mientras que la que menos vendió fue la tienda Ovalle con un 0.0476 % del total de SKUs vendidos. En la Zona Centro Oriente, la tienda que más vendió corresponde al Parque Arauco con un 0.2924 % del total de productos vendidos, mientras que la que menos vendió fue la tienda Los Dominicos con un 0.0479 % del total de SKUs vendidos. En la Zona Centro

Poniente, la tienda que más vendió corresponde al Plaza Oeste con un 0.1549 % del total de productos vendidos, y como esta corresponde a la zona de mayor cantidad de ventas, se usará esta tienda para el desarrollo del presente estudio y predicciones a nivel de tienda. En esta Zona, el menor volumen de ventas lo tuvo la tienda Independencia con un 0.0185 % del total de SKUs vendidos. Por último, en la Zona Sur, la tienda que más vendió corresponde a la tienda El Trébol con un 0.1394 % del total de productos vendidos, mientras que la que menos vendió fue la tienda Pucón con un 0.0082 % del total de SKUs vendidos.

Continuamos con las tiendas del *dataset* de Pantalones.



(a) Zona Norte.

(b) Zona Centro Oriente.

(c) Zona Centro Poniente.

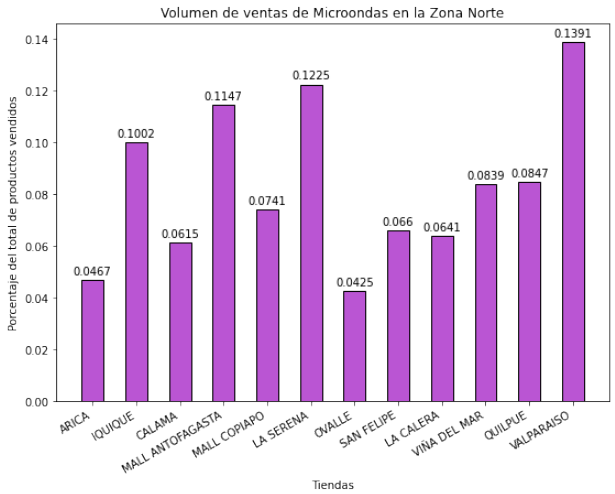
(d) Zona Sur.

Figura 4.14: Comparación del volumen de ventas en las tiendas de cada zona del *dataset* de Pantalones.

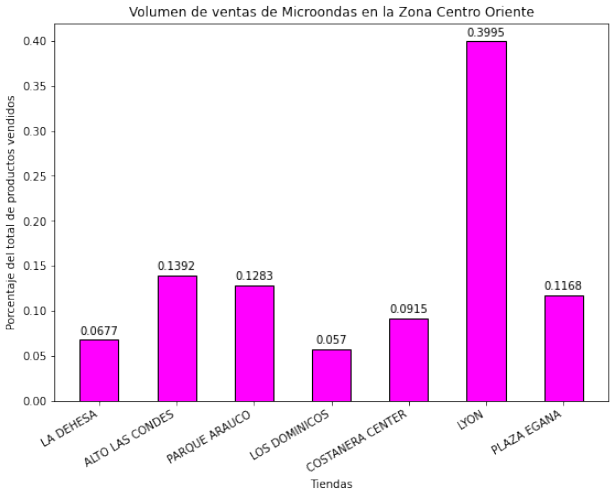
En la figura 4.14 vemos que en el *dataset* de Pantalones, en la Zona Norte, la tienda que más vendió corresponde a la tienda La Serena con un 0.1342 % del total de productos vendidos, mientras que la que menos vendió fue la tienda Ovalle con un 0.0527 % del total de SKUs vendidos. En la Zona Centro Oriente, la tienda que más vendió corresponde al Costanera

Center con un 0.2644 % del total de productos vendidos, mientras que la que menos vendió fue la tienda La Dehesa con un 0.0475 % del total de SKUs vendidos. En la Zona Centro Poniente, la tienda que más vendió corresponde al Plaza Vespucio con un 0.1088 % del total de productos vendidos, y como esta corresponde a la zona de mayor cantidad de ventas, se usará esta tienda para el desarrollo del presente estudio y predicciones a nivel de tienda. En esta Zona, el menor volumen de ventas lo tuvo la tienda Independencia con un 0.0474 % del total de SKUs vendidos. Por último, en la Zona Sur, la tienda que más vendió corresponde a la tienda El Trébol con un 0.1225 % del total de productos vendidos, mientras que la que menos vendió fue la tienda Pucón con un 0.0239 % del total de SKUs vendidos.

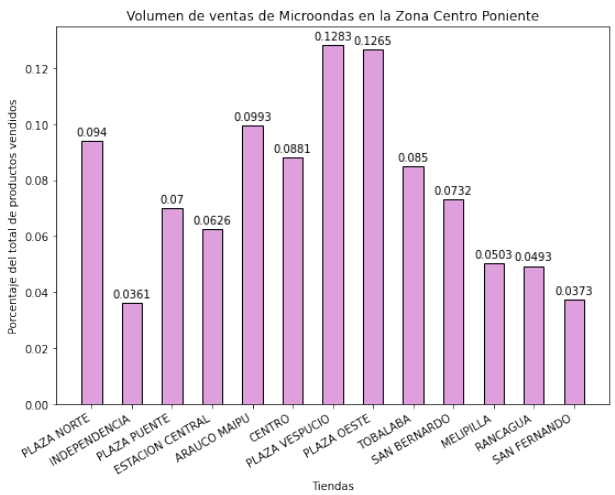
Finalmente, seguimos con las tiendas del *dataset* de Microondas.



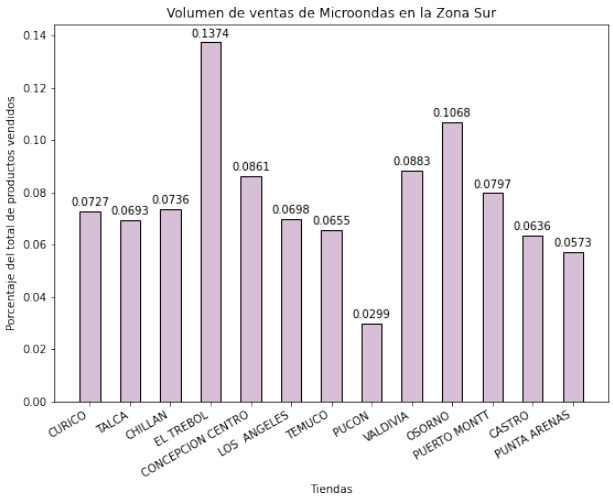
(a) Zona Norte.



(b) Zona Centro Oriente.



(c) Zona Centro Poniente.



(d) Zona Sur.

Figura 4.15: Comparación del volumen de ventas en las tiendas de cada zona del *dataset* de Microondas.

En la figura 4.15 vemos que en el *dataset* de Microondas, en la Zona Norte, la tienda que más vendió corresponde a la tienda Valparaíso con un 0.1391 % del total de productos

vendidos, mientras que la que menos vendió fue la tienda Ovalle con un 0.0425 % del total de SKUs vendidos. En la Zona Centro Oriente, la tienda que más vendió corresponde a la tienda Lyon con un 0.3995 % del total de productos vendidos, mientras que la que menos vendió fue la tienda Los Dominicos con un 0.057 % del total de SKUs vendidos. En la Zona Centro Poniente, la tienda que más vendió corresponde al Plaza Vespucio con un 0.1283 % del total de productos vendidos, y como esta corresponde a la zona de mayor cantidad de ventas, se usará esta tienda para el desarrollo del presente estudio y predicciones a nivel de tienda. En esta Zona, el menor volumen de ventas lo tuvo la tienda Independencia con un 0.0361 % del total de SKUs vendidos. Por último, en la Zona Sur, la tienda que más vendió corresponde a la tienda El Trébol con un 0.1374 % del total de productos vendidos, mientras que la que menos vendió fue la tienda Pucón con un 0.0299 % del total de SKUs vendidos.

4.5. Series temporales de ventas y gráficos preliminares

En esta sección se presentarán las primeras visualizaciones de las ventas históricas de productos en formato de series temporales.

En primer lugar, se mostrarán gráficos que comparan los datos de ventas de cada *dataset* y su distribución por semanas, a lo cual de ahora en adelante se le llamará serie de tiempo o serie temporal de ventas. Para lograr esto, se realiza una agregación de los datos de venta, en este caso, la suma total, por cada semana, sin restricciones de tienda, ni de otros filtros. El detalle de este proceso de transformación y generación de datos se encuentra en la sección 5.3. Además, en el apéndice C.1, se encuentra una tabla que será de utilidad para estos y los próximos gráficos de series temporales a lo largo del presente trabajo.

Si bien los siguientes gráficos aglomeran los datos de venta de todos los productos de los *datasets*, estos sirven para obtener una visión general de las tendencias de ventas y, en especial, del comportamiento de las series de tiempo. En cada serie temporal que veremos a continuación se aprecian fluctuaciones, en las cuales algunas están notoriamente marcadas por altos *peaks* de ventas, o bien por periodos de muy pocas ventas. Esto se debe, en parte, a factores geográficos y sus climas asociados, el estilo de vida, la cultura, la edad, el ingreso, entre otros[29]. Además, existen eventos que impactan directamente en el comportamiento de todos los compradores. A veces estos eventos son sucesos puntuales y, a priori, irrepetibles y/o difíciles de anticipar, tales como un desastre natural o, como ha ocurrido en la industria del *retail* en los últimos años, con comportamientos de compra alterados por el estallido social[24] y el inicio de la pandemia COVID-19[10][30].

Por otra parte, existen numerosos eventos[2] y fechas conmemorativas relacionadas con días de asueto, eventos nacionales e internacionales, las estaciones, festividades de origen religioso y otras efemérides. Dichos eventos implican campañas de *marketing* masivas que influyen en el comportamiento de compra de la gente. Los principales eventos asociados a campañas comerciales y ofertas especiales en Chile corresponden a: San Valentín, Día Internacional de la Mujer, Semana Santa, Día de la Madre, *CyberDay*, Día del Padre, Día del Niño, Fiestas Patrias, Halloween, Día del Soltero (*a.k.a* Ofertas del 11:11), *Black Friday*, *Cyber Monday*, Navidad y Año Nuevo. Todas estas fechas alteran las tendencias de compras, lo cual se verá reflejado en el comportamientos de las series temporales de las ventas de productos, como veremos en las próximas visualizaciones.

Los siguientes gráficos se realizaron para cada uno de los *datasets* disponibles y permiten observar tendencias de ventas de manera general para cada categoría de productos. Las ventas se expresan como un porcentaje sobre la cantidad total, por confidencialidad.

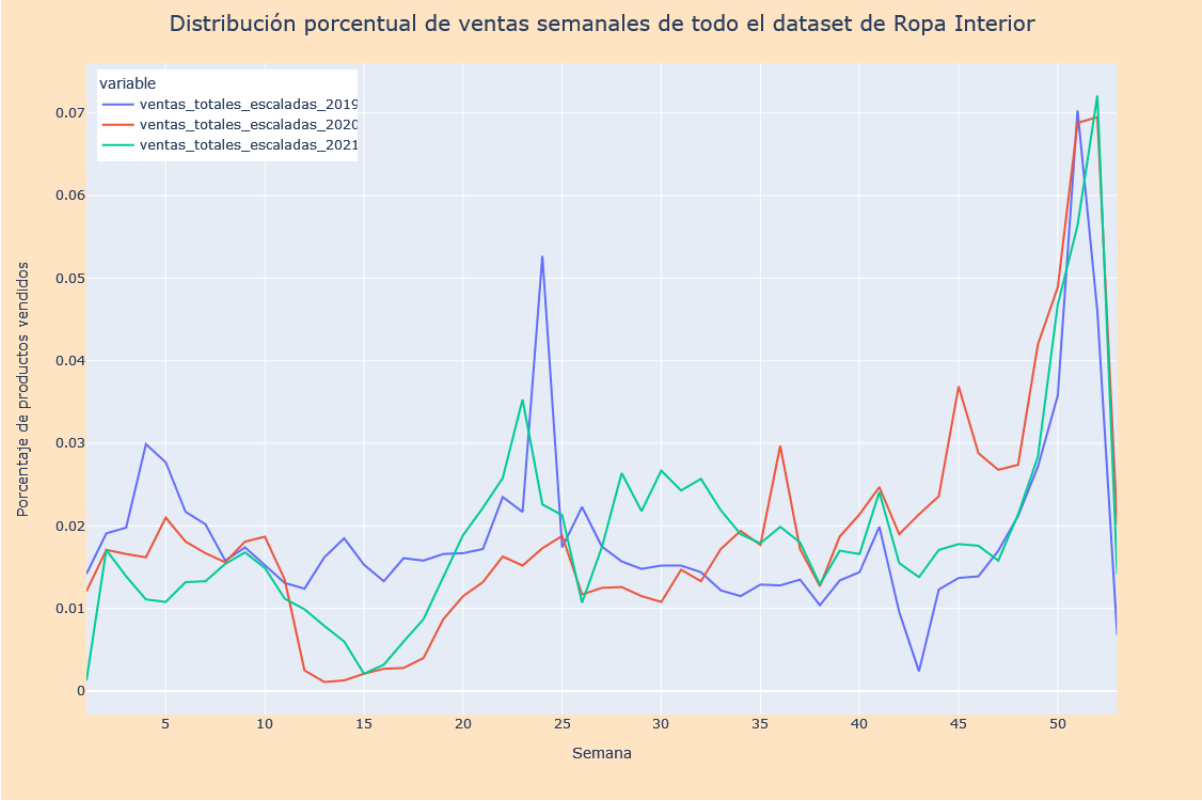


Figura 4.16: Comparación de series temporales de ventas semanales porcentuales de Ropa Interior para los años 2019, 2020 y 2021.

En la figura 4.16 se observan interesantes aspectos del comportamiento de ventas en el *dataset* de Ropa Interior:

- El año 2019 tuvo una marcada disminución de ventas hacia la semana 42, lo cual se explicar por el estallido social que vivió Chile a partir del 18 de octubre del mismo año.
- El año 2019, hasta la semana 41, es el único de los años presentes en el *dataset* cuyas ventas no se vieron afectadas por el estallido social ni por la situación sanitaria producto de la pandemia del coronavirus.
- Los años 2021 y 2021 presentan una notable baja de ventas entre las semanas 12 y 16, probablemente debido a la pandemia.
- Las series correspondientes a cada año muestran un gran *peak* de ventas entre las semana 20 y 24, la cuales se explicar por los eventos del Día de la Madre y el *CyberDay*.
- La víspera de Navidad generó un enorme *peak* de ventas en las últimas de diciembre de cada año.

Continuamos con el gráfico de ventas históricas del *dataset* de Lavadoras.

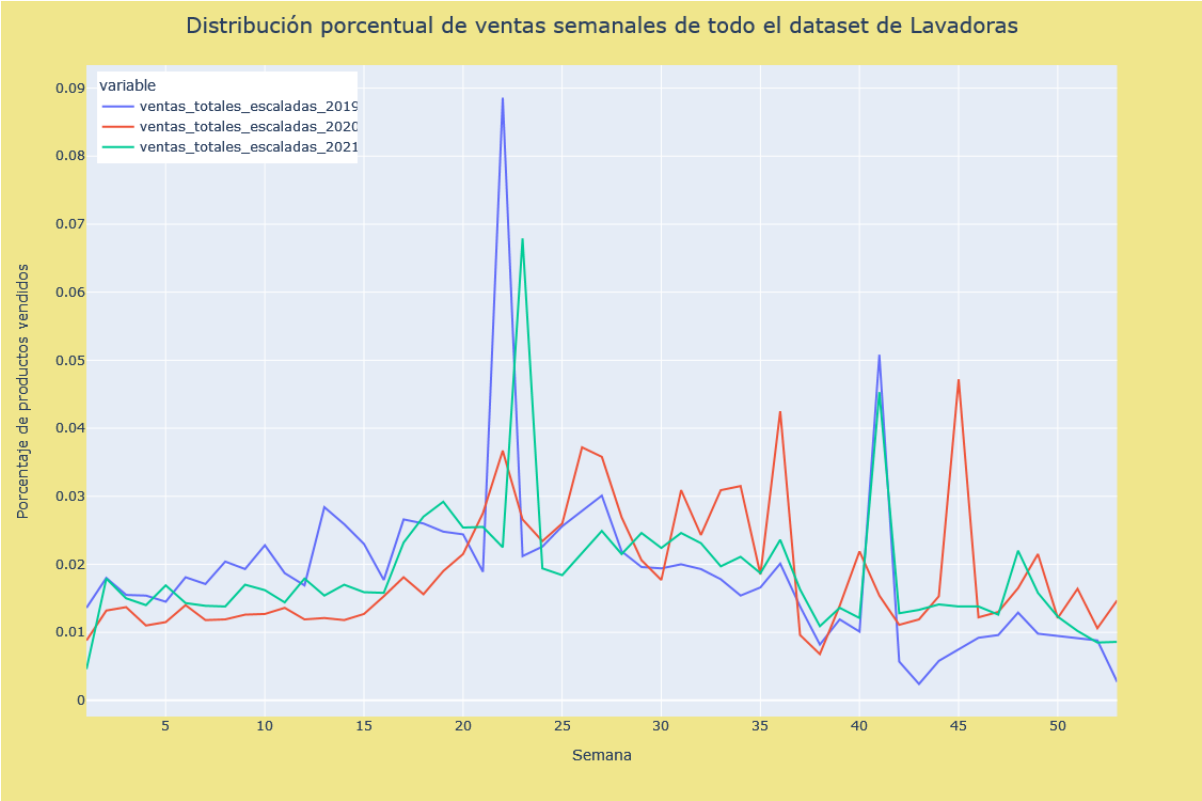


Figura 4.17: Comparación de series temporales de ventas semanales porcentuales de Lavadoras para los años 2019, 2020 y 2021.

Con respecto a la figura 4.17 se realizan las siguientes observaciones sobre el comportamiento de ventas:

- Nuevamente se evidencia el impacto del estallido social al ver considerablemente disminuidas las ventas en la semana 42 del año 2019.
- En los años 2020 y 2021 se aprecia un pronunciado *peak* de ventas en las semanas 22 y 23, debido al *CyberDay*, sin embargo este *peak* no se dio de forma tan marcada en el año 2019.
- Se observan *peaks* de venta en la semana 45 del año 2019 y en la semana 41 de los años 2020 y 2021.
- La víspera de Navidad no influyó en las ventas de Lavadoras, probablemente porque este tipo de productos no es tan usual comprarlos para darlos como un regalo navideño, a diferencia de los productos de vestuario.

Continuamos con el gráfico de ventas históricas del *dataset* de Pantalones.

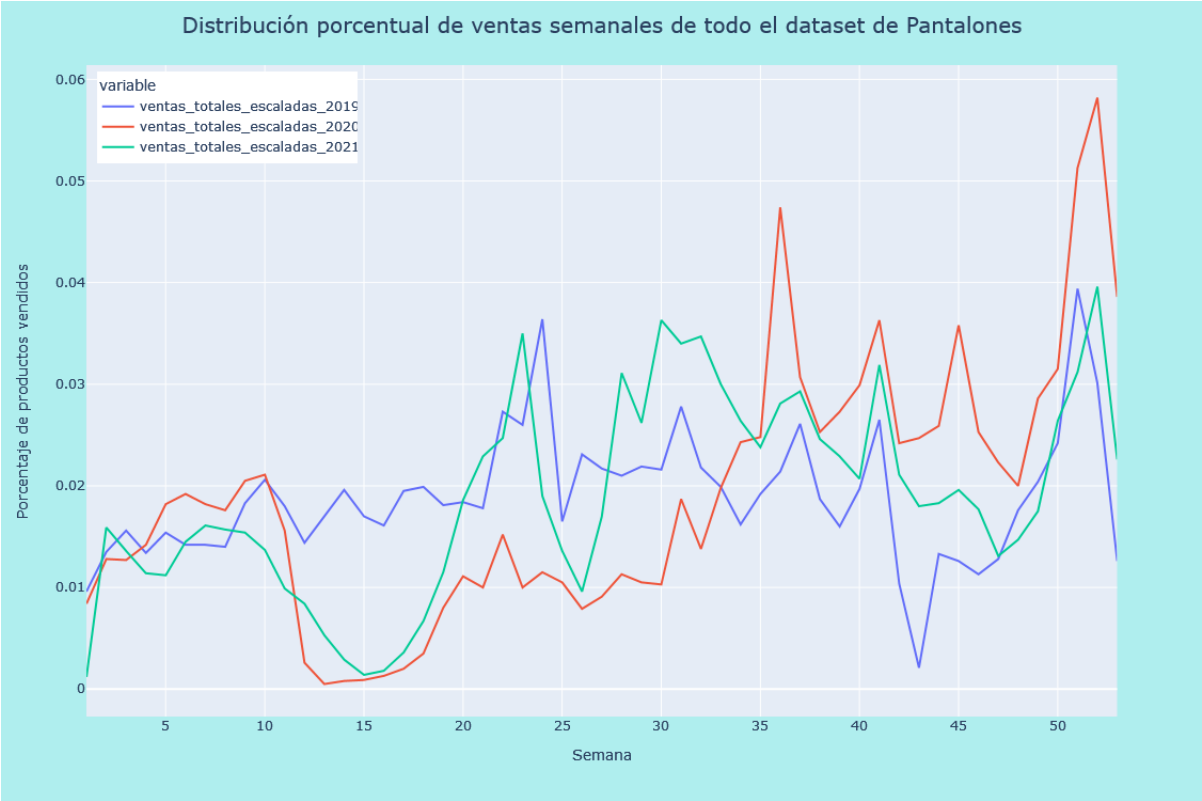


Figura 4.18: Comparación de series temporales de ventas semanales porcentuales de Pantalones para los años 2019, 2020 y 2021.

Con respecto a la figura 4.18 se realizan las siguientes observaciones similares a las hechas en el *dataset* de Ropa Interior (figura 4.16):

- Nuevamente se evidencia el impacto del estallido social al ver considerablemente disminuidas las ventas en la semana 42 del año 2019.
- De igual forma, los años 2020 y 2021 presentan una notoria baja de ventas entre las semanas 12 y 16, probablemente por la pandemia.
- Se aprecia un gran *peak* de ventas entre semanas 20 y 24 debido asl Día de la Madre y el Cyber Day, sobre todo en el año 2021, donde además se observa un alto *peak* en la semana 41, en los 3 años.
- Nuevamente se ve un enorme *peak* de ventas en las últimas semanas del año, debido a Navidad, confirmando que las ventas de productos de la categoría de vestuario son altamente sensibles a los días de la víspera de Navidad.

Finalmente, seguimos con el gráfico de ventas históricas del *dataset* de Microondas.

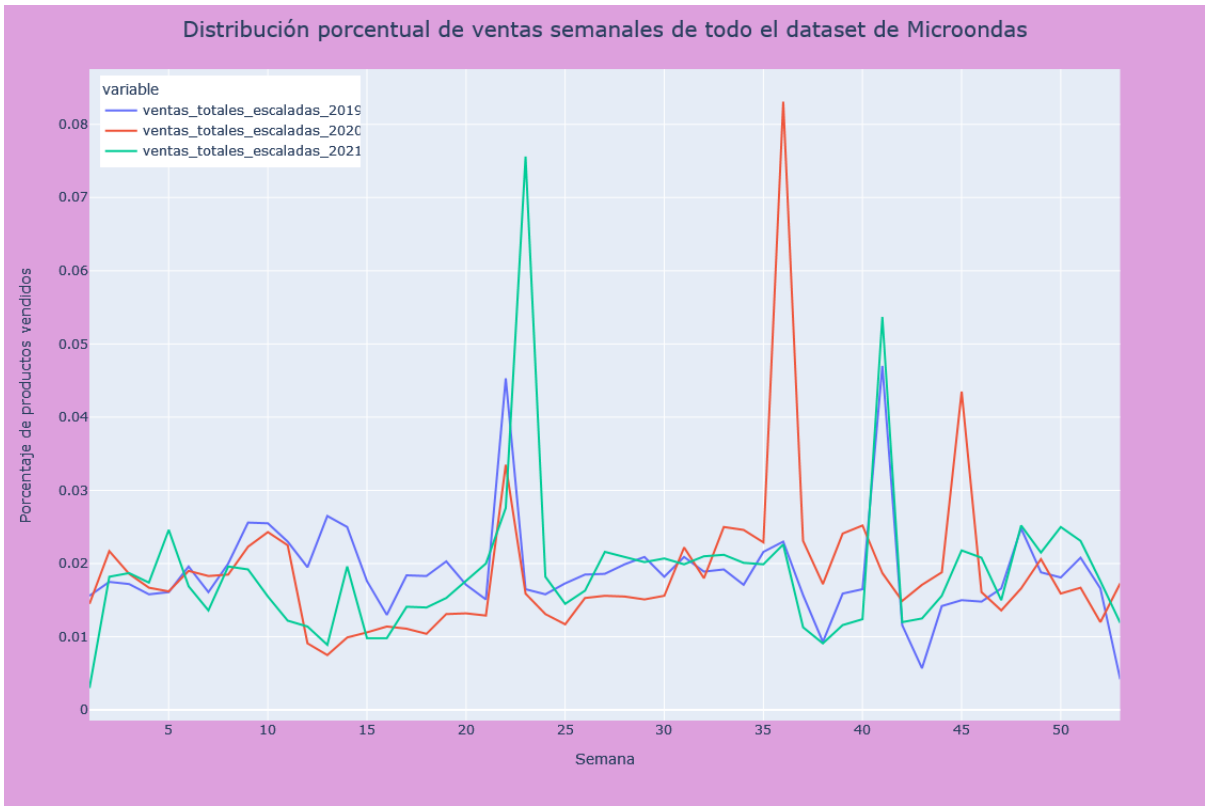


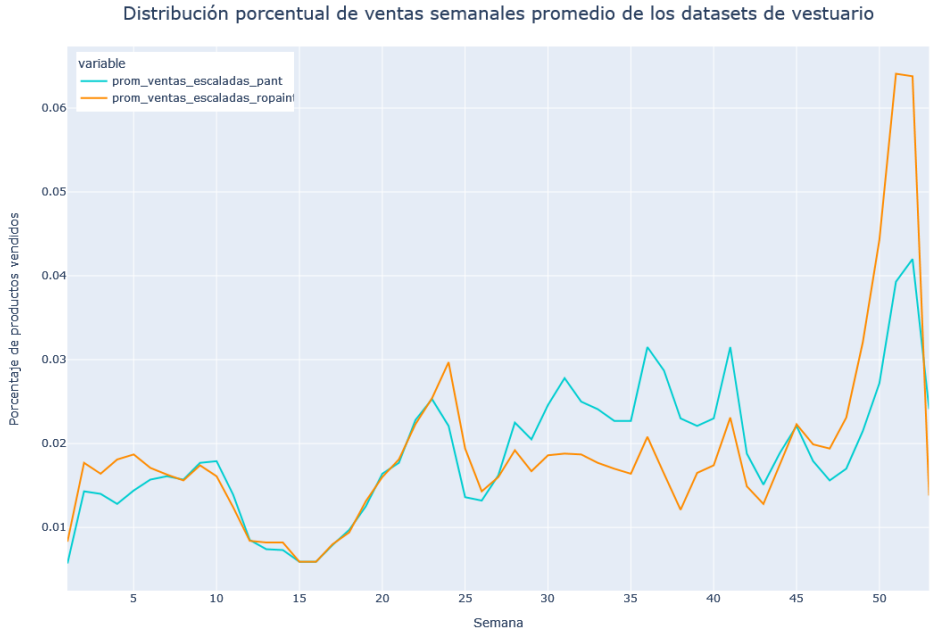
Figura 4.19: Comparación de series temporales de ventas semanales porcentuales de Microondas para los años 2019, 2020 y 2021.

En la figura 4.19 se encontraron observaciones similares a las realizadas en el *dataset* de Lavadoras (figura 4.17):

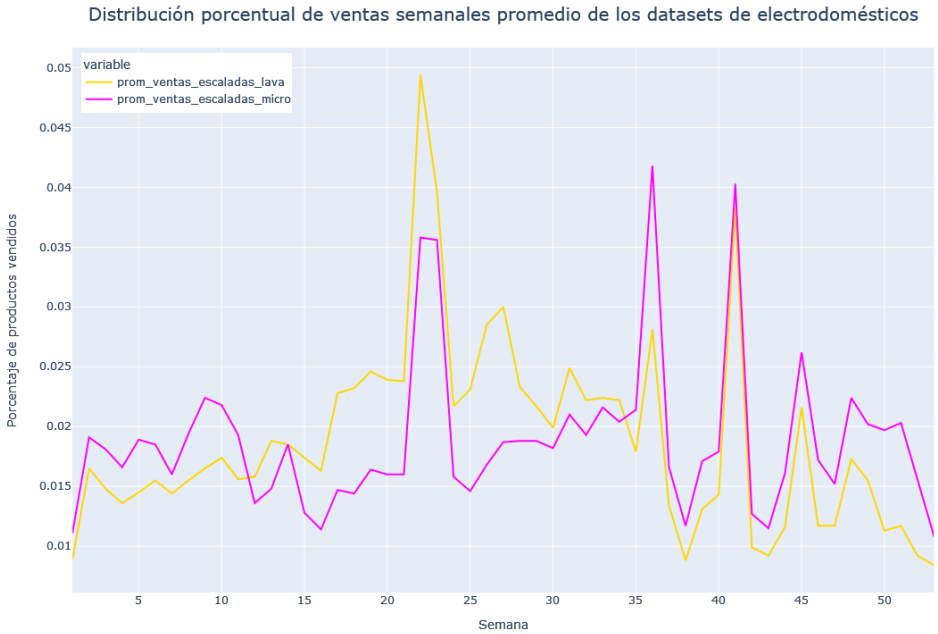
- Nuevamente se evidencia el impacto del estallido social al ver considerablemente disminuidas las ventas en la semana 42 del año 2019, comprobando que este suceso marcó de manera transversal las ventas de todos los *datasets* estudiados.
- De igual forma, en los años 2020 y 2021 se aprecia un pronunciado *peak* de ventas en las semanas 22 y 23, debido al *CyberDay*, sin embargo este *peak* no se dio de forma tan marcada en el año 2019.
- También se observan *peaks* de venta en la semana 41 de los años 2020 y 2021 y en la semana 45 del año 2019, además de un enorme *peak* de ventas en la semana 36 del mismo año.
- Nuevamente la víspera de Navidad no influyó en las ventas de Lavadoras, confirmando que Lavadoras y Microondas son dos segmentos de electrodomésticos que no se ven alterados por estas fechas.

Como notamos, se presentaron observaciones muy similares al comparar *datasets* dentro de la misma categoría, vestuario o electrodomésticos. Por esta razón, se realizaron gráficos adicionales para comparar lado a lado estos comportamientos de series temporales.

Estos gráficos se hicieron sacando el promedio de ventas semanales del 2019, 2020 y 2021.



(a) *Datasets* de vestuario



(b) *Datasets* de electrodomésticos

Figura 4.20: Comparación de series temporales de ventas semanales porcentuales promedio en los *datasets* de vestuario y electrodomésticos.

Como se ve en la figura 4.20a, los *datasets* de vestuario, si bien corresponden a productos distintos, tanto Ropa Interior como Pantalones presentaron *peaks* y disminuciones de ventas en las mismas semanas, notando que las ventas de Ropa Interior presentan estas fluctuaciones de manera mucho más marcada. Lo mismo sucede en la figura 4.20b al comparar entre los

datasets de electrodomésticos, destacando el de Lavadoras, que presentó sus fluctuaciones de manera más notoria que en el de Microondas.

Por último, se muestra el gráfico de las ventas semanales promedio juntando los 4 *datasets* disponible.

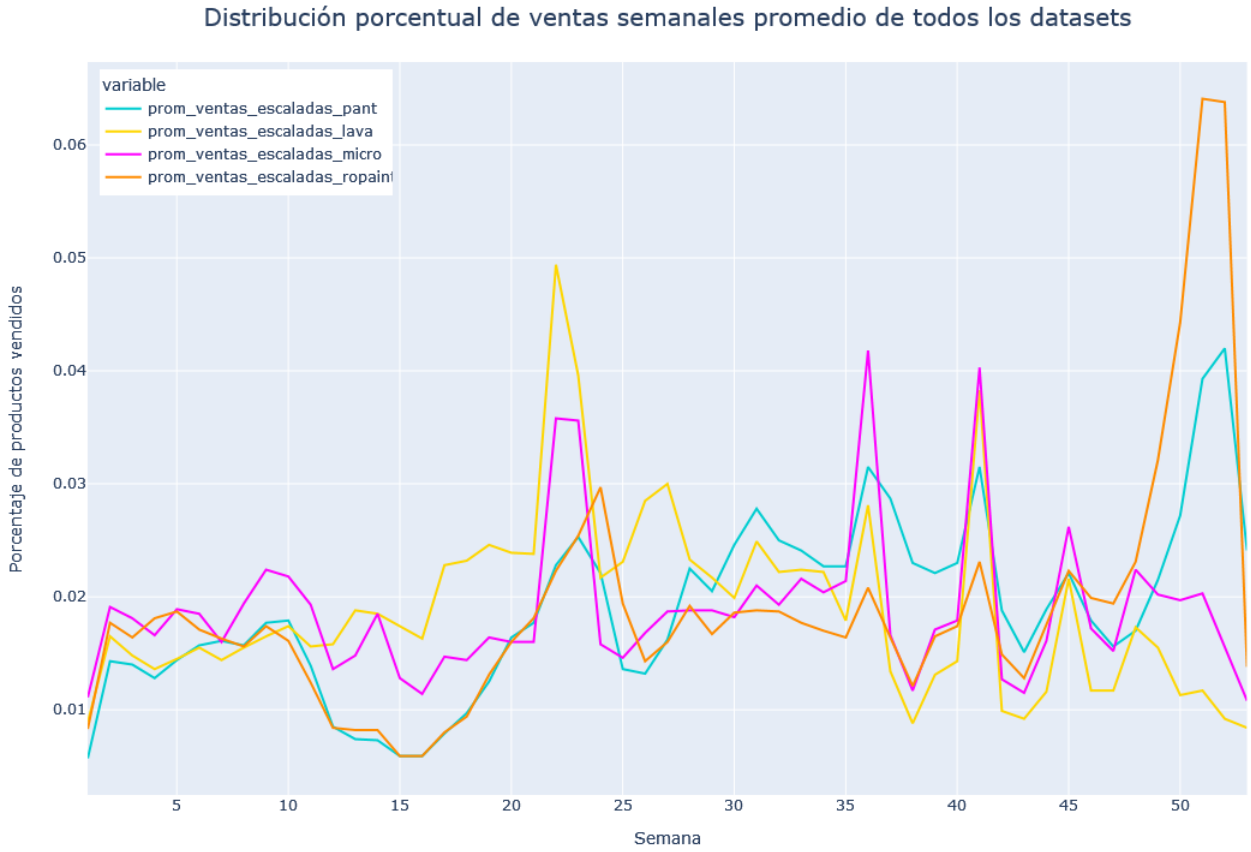


Figura 4.21: Comparación de series temporales de ventas semanales porcentuales promedio en todos los *datasets*.

En la figura 4.21 notamos que todos los *datasets* presentan en mayor o menor medida los *peaks* de venta encontrados entre las semanas 20 y 24 y en la semana 41. Además, queda en evidencia que el *dataset* de Ropa Interior es el que presentó un mayor incremento de sus ventas en la víspera de Navidad y Año Nuevo.

Capítulo 5

Diseño y Desarrollo de la Solución

En este capítulo se explicará cuál fue la metodología del desarrollo del presente trabajo. Se hablará en profundidad de las etapas más relevantes y se reportará la realización de la prueba de concepto.

5.1. Metodología

En esta sección se mencionará la metodología del proceso que llevó a poder realizar las simulaciones de predicción de ventas en base a la similitud de las series temporales de las ventas de los productos. Parte de esta metodología fue realizada durante la prueba de concepto, pero también otra parte fue diseñada a partir de las conclusiones obtenidas en ella, como veremos en la sección 5.2.

5.1.1. Etapas del proceso

A continuación se mencionarán y detallarán brevemente todas las etapas involucradas en el desarrollo de la solución. Estas etapas se llevaron a cabo en el orden indicado y de manera individual para cada *dataset*.

1. **Preprocesamiento de datos**: aquí se realizó una limpieza de los datos, se aplicaron filtros para seleccionar los datos y se realizaron transformaciones para obtener el *dataset* final para realizar el trabajo. Esta etapa se detallará en profundidad en la sección 5.3.
2. **Previsualización de series temporales**: esta etapa fue necesaria para validar que los datos estén en el formato deseado. También sirvió para identificar visualmente la existencia de series temporales similares.
3. **Aplicar la heurística de predicción de ventas**: en esta etapa se llevó a cabo el proceso de simulación del pronóstico de ventas, cuya heurística se define en la sección 5.1.2. El objetivo de esta etapa es encontrar y reunir los pares de series temporales más cercanas y luego simular una predicción de ventas.
4. **Evaluar resultados**: se calcularon las métricas más relevantes para evaluar el pronóstico de venta realizado, además se generaron visualizaciones que permitieran obtener conclusiones sobre los experimentos realizados.

5. **Replicar proceso en los demás *datasets***: se repitieron las etapas anteriores para realizar este estudio en las principales tiendas y zonas que concentraran el mayor volumen de ventas de cada *dataset*.

5.1.2. Heurística de predicción de ventas

La metodología para realizar el pronóstico de ventas entre series temporales más cercanas consta de las siguientes etapas:

1. **Calcular matriz de distancias entre las series temporales**: comparar cuantitativamente la similitud entre cada una de las primeras mitades de las series de tiempo, i.e. semanas 1 a 24 de cada serie, mediante el cálculo de la distancia Euclidiana entre ellas, generando un matriz de distancias. El detalle de este procedimiento se profundiza en la sección 5.4.
2. **Reunir los pares de series temporales más cercanas**: a partir de la matriz de distancias obtenida en la etapa anterior, se buscan los pares más cercanos al obtener los valores de distancia más bajos y reuniendo las series temporales asociadas a dichos valores. Se genera un arreglo de pares más cercanos, ordenados del más cercano al más lejano, para realizar con ellos la simulación de predicción.
3. **Simular el pronóstico de ventas**: se aplica el principio de que la primera mitad de las series temporales determina su cercanía y la segunda simulará la predicción de ventas. Con el arreglo obtenido anteriormente, para cada uno de sus pares de series más cercanas, se simulará el pronóstico de ventas, utilizando las segundas mitades de las series y estableciendo una de ellas como los datos originales y la otra como la predicción. Esta etapa se detallará en profundidad en la sección 5.5.

5.1.3. Herramientas computacionales

Las tecnologías utilizadas para llevar a cabo el trabajo fueron las siguientes que se mencionan a continuación:

Hardware

Las siguientes especificaciones de *hardware* corresponden a una instancia de cómputo en el entorno de desarrollo de *Google Colaboratory*[1].

- **CPU**: procesador *Intel*(R) *Xeon*(R) de dos núcleos a 2.20GHz.
- **RAM**: 12 GB.
- **Sistema Operativo**: *Ubuntu* 18.04.6 LTS.

Software

Al utilizar el entorno de desarrollo de *Google Colab*, se siguió el principio de uso de *Jupyter Notebook*[2], el cual utiliza el lenguaje de programación *Python*[3] en su versión 3.7.14.

[1]Conocido como *Google Colab*. Más información en el enlace `https://colab.research.google.com/`.
[2]Más información en el enlace `https://docs.jupyter.org/en/latest/`.
[3]Más información en el enlace `https://docs.python.org/release/3.7.14/`.

Las principales librerías utilizadas fueron las siguientes.

- **Manejo de datos**: *Pandas*[4].
- **Procesamiento y cálculos matemáticos**: *NumPy*[5], *SciPy*[6] y *Scikit-Learn*[7].
- **Visualizaciones**:*Matplotlib*[8], *Plotly*[9] y *Seaborn*[10].

5.2. Prueba de concepto

Antes de poder llevar a cabo el estudio final y la simulación de predicción de ventas, fue necesario realizar pruebas y visualizaciones preliminares con un enfoque en el análisis de series temporales. Para esto, se hizo una prueba de concepto utilizando el *dataset* de Pantalones, cuyo propósito fue sentar las bases para los desarrollos finales. A continuación se hablará de lo que fue esta prueba de concepto, sus objetivos, las visualizaciones más relevantes y las observaciones y conclusiones preliminares que se pudieron obtener al realizar esta prueba.

5.2.1. Objetivos

Parte de los objetivos mencionados a continuación están estrechamente relacionados con los objetivos específicos indicados en la sección 1.2.2.

- Graficar los datos de ventas de productos en el formato de series temporales.
- Encontrar visualmente series de tiempo similares.
- Determinar la métrica más apropiada para evaluar la similitud de las series temporales.
- Descubrir las mejores transformaciones sobre los datos para elaborar los *datasets* definitivos para las próximas simulaciones de ventas.

5.2.2. Resultados y observaciones

En primera instancia, se pensó que lo correcto era realizar el estudio de estas series temporales a nivel de SKUs y en un espacio temporal a nivel de días. El siguiente gráfico ilustra los resultados bajo dichos principios.

[4]Más información en el enlace `https://pandas.pydata.org/docs/`.
[5]Más información en el enlace `https://numpy.org/doc/stable/`.
[6]Más información en el enlace `https://docs.scipy.org/doc/scipy/`.
[7]Más información en el enlace `https://scikit-learn.org/stable/user_guide.html`.
[8]Más información en el enlace `https://matplotlib.org/stable/index.html`.
[9]Más información en el enlace `https://plotly.com/python/`.
[10]Más información en el enlace `https://seaborn.pydata.org/`.

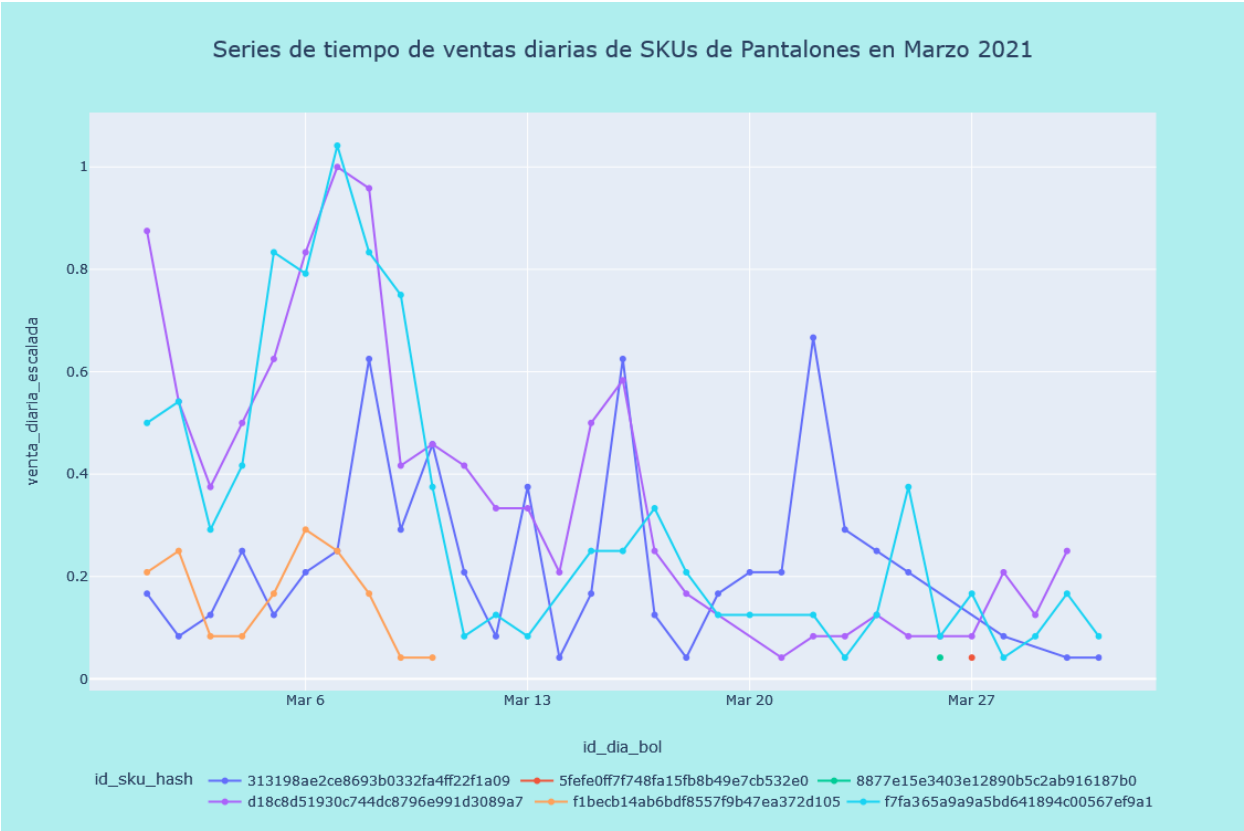


Figura 5.1: Gráfico de series temporales con ventas escaladas de pantalones en un contexto diario.

A partir del gráfico 5.1, podemos observar que existen series, como la de color naranja y las de color verde y rojo (que corresponden realmente a puntos), que tienen muy pocos datos de venta en la escala diaria y se ven truncadas o discontinuas en la ventana temporal observada.

Esto motivó la adición de dos transformaciones sobre los datos: una agrupación de ventas semanales en vez de una diaria, para tener una mayor cantidad de ventas en otra escala de tiempo; y la adición de los días o semanas sin ventas realizadas, para darle continuidad a las series "cortadas". Además para aprovechar aún más el nivel de agrupación, se decidió en este punto agrupar las ventas de los años 2019, 2020 y 2021 en una sola serie representativa con la suma de las ventas a nivel semanal, en vez de individualizar series para cada año.

Luego de aplicar estos cambios, se visualizaron un conjunto de series temporales aleatorias, para encontrar relaciones interesantes. Al hacer esto, se observó que era posible encontrar series temporales *aparentemente*, al menos visualmente, tal como se puede apreciar a continuación.

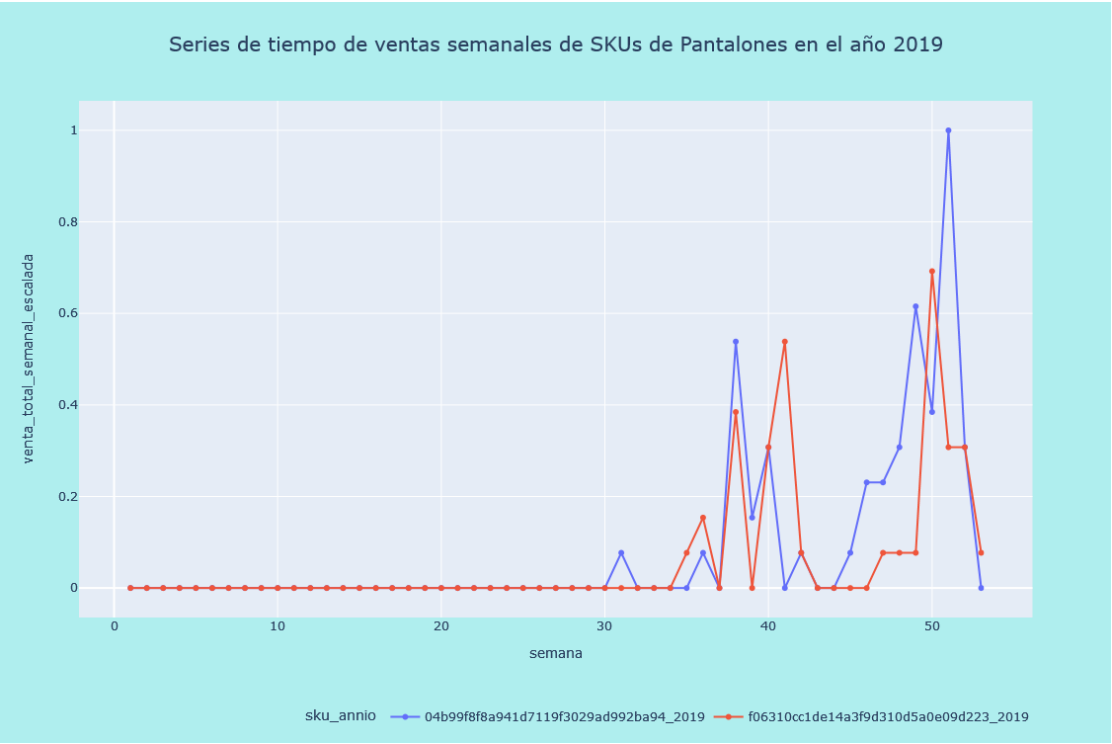
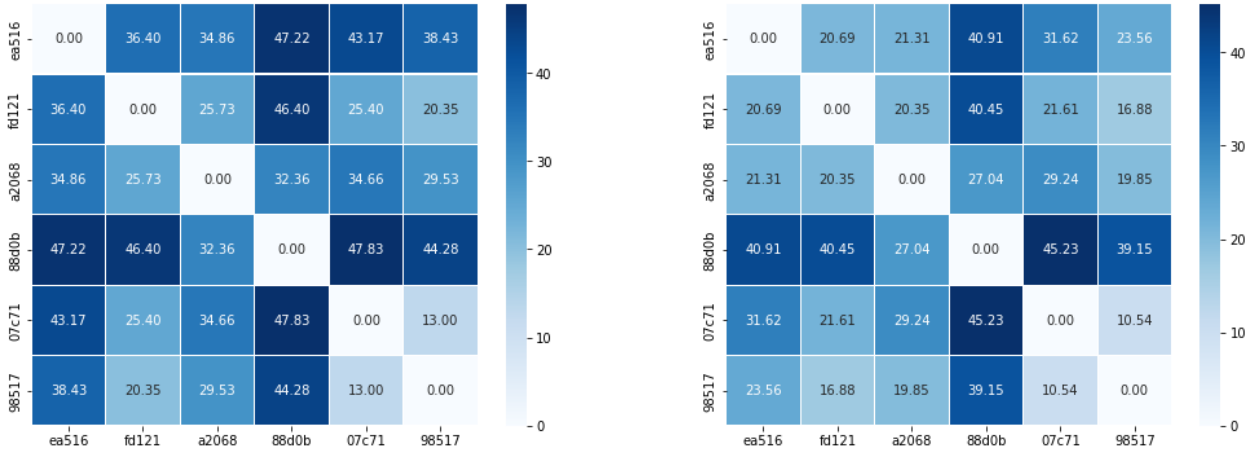


Figura 5.2: Gráfico de dos series temporales de ventas semanales visualmente parecidas en el *dataset* de Pantalones.

Al observar el gráfico 5.2, surgieron nuevas inquietudes que motivaron transformaciones adicionales sobre los datos: tanto estas series, como muchas otras, experimentan una explosión de ventas a partir de la semana 49 en adelante. Esto se explica por los eventos de Navidad y Año Nuevo, por lo que se estimó pertinente acortar la ventana temporal a las semanas 1 a 48, puesto que las últimas semanas no aportaban datos especialmente útiles, i.e. es lógico esperar un aumento de ventas en dichas fechas.

Por otra parte, se observó que efectivamente existen series temporales similares, y con esto, se hizo necesario medir esta similitud de manera cuantitativa.

Las medidas de distancia más apropiadas para comparar la similitud de series temporales son la distancia Euclidiana y la distancia DTW[4], por lo que se procedió a calcular estas medidas para obtener las distancia entre cada uno de los pares de series temporales que se podían formar. Con esto, se generaron las siguientes matrices de distancias.



(a) Matriz de distancias Euclidianas.

(b) Matriz de distancias DTW.

Figura 5.3: Comparación de matrices de distancias para algunas series temporales.

Con las matrices de la figura 5.3 fue posible observar cuáles eran los pares de series más cercanos y más lejanos, sin embargo, esto no fue suficiente para concluir cuál de las dos medidas de distancia era la más apropiada. Por lo que se procedió a visualizar cómo se realizaba el pareo (*matching*) de series según el cálculo de la DTW, mediante el gráfico a continuación.

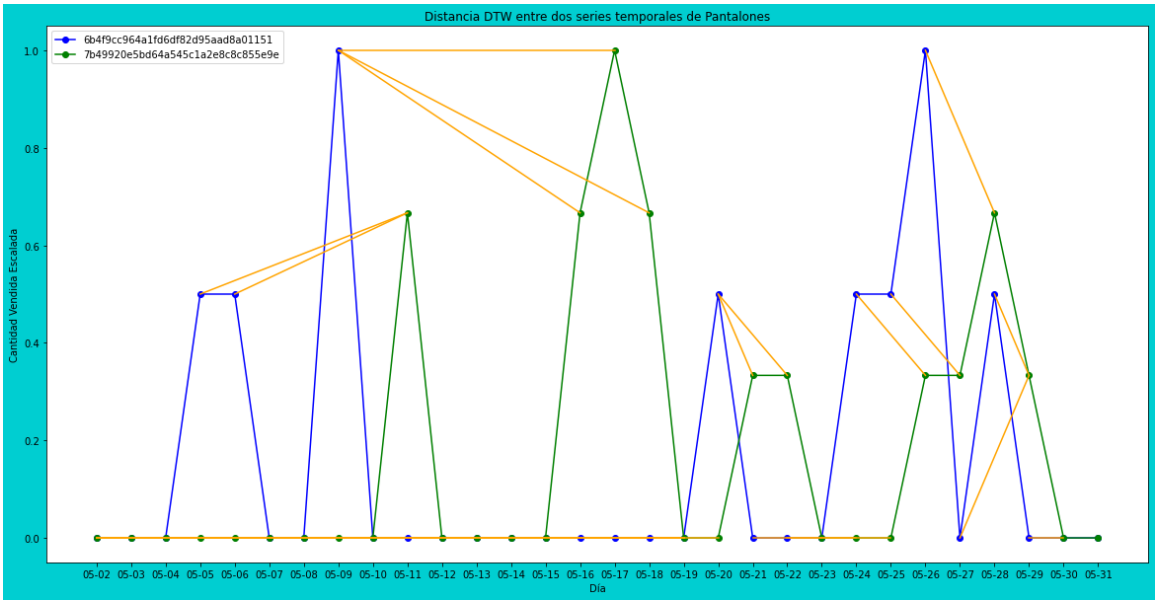
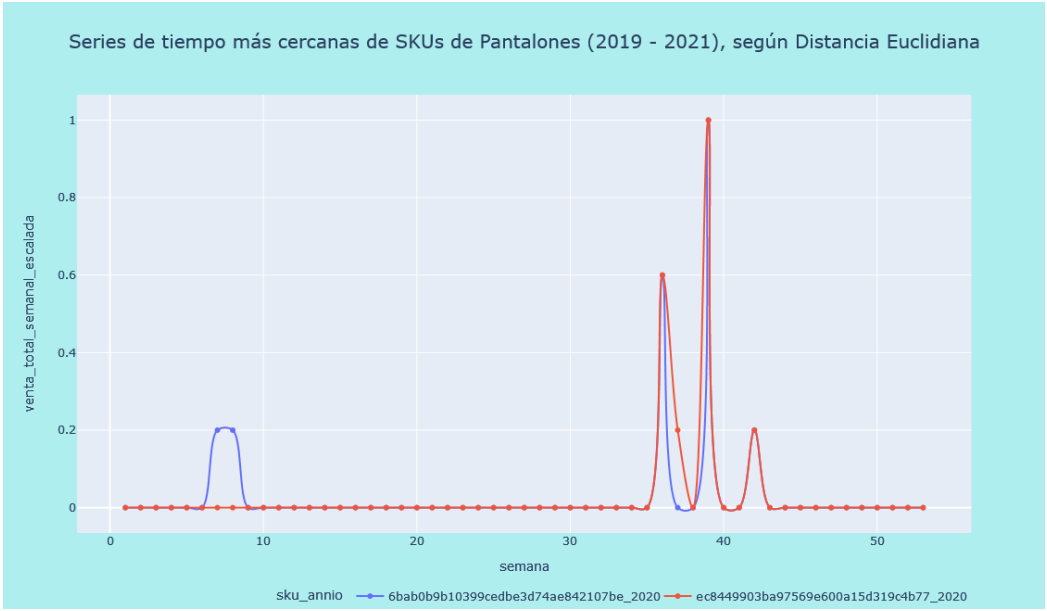
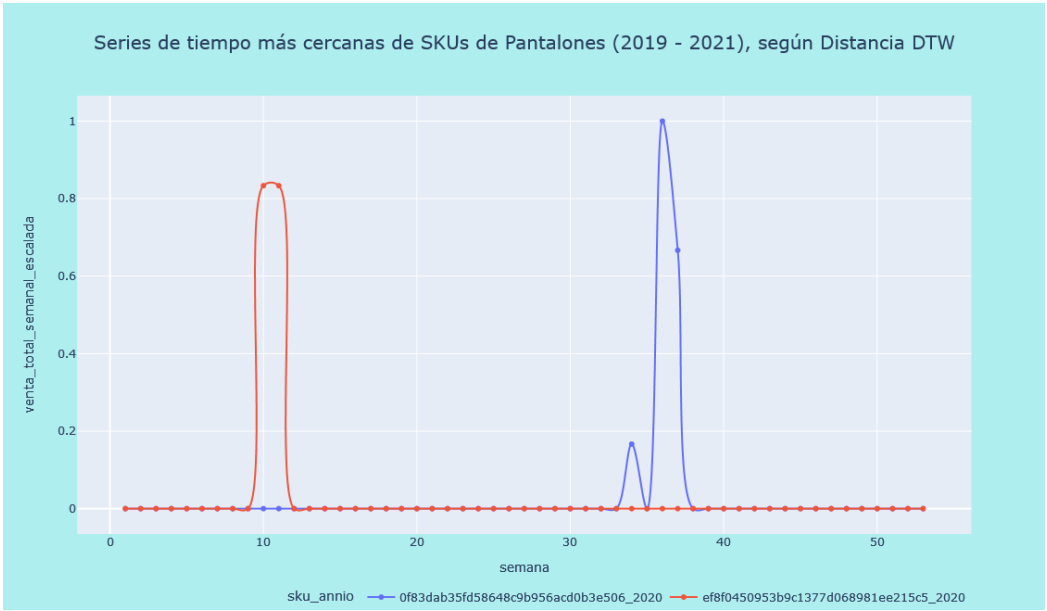


Figura 5.4: Pareo de series temporales según distancia DTW.

Gracias a este gráfico de la figura 5.4 fue posible observar que el pareo de la distancia DTW no era el deseado para este estudio, ya que este se hacía respetando la forma de la serie de tiempo, dejando en segundo plano la componente estacionaria, vale decir, el comportamiento semana a semana, y por ende, desaprovechando las tendencias de distribución de ventas en el tiempo. Esto se evidencia de manera más clara en la siguiente figura.



(a) Par de series temporales más cercanas según distancia Euclidiana.



(b) Par de series temporales más cercanas según distancia DTW.

Figura 5.5: Par de series temporales más cercanas según distancia DTW.

En la figura 5.5b vemos que las series más cercanas según la distancia DTW tienen una forma muy similar, sin embargo, sus ventas corresponden a periodos muy lejanos entre sí, lo cual no es de gran utilidad, ya que si bien esto ayudaría a determinar cómo serían los *peaks* de ventas, se aleja del objetivo de predecir unidades a vender de manera semanal. Esto no sucede con la distancia Euclidiana, ya que esta realiza el cálculo punto a punto (semana a semana), por lo que resulta ser más adecuada para lo que se busca predecir. De hecho, vemos que en el gráfico 5.5a, la similitud de las series encontradas concuerda con lo esperado en este estudio. Por esta razón, se decidió ocupar la distancia Euclidiana como medida definitiva para calcular similitudes entre series temporales.

Como última observación de esta prueba de concepto, se determinó que agrupar los SKUs de un mismo estilo y representarlos como una sola serie temporal tenía mucho más sentido desde el punto de vista de negocio de "La Empresa", ya que así se podían agrupar productos que tuvieran diferencias despreciables, tales como las diferentes tallas de un producto de vestuario.

5.3. Preprocesamiento de los datos

En esta sección se detallarán en profundidad todas las subetapas realizadas en la fase de preprocesamiento de los datos. Esta fase fue necesaria para transformar el *dataset* original y generar el conjunto de datos final en el formato de series temporales para poder llevar a cabo las simulaciones de predicciones de ventas.

Limpieza de datos. Primero fue necesario regular el formato de las fechas, ya que venían como texto y se necesitaban en formato *DateTime* para agrupar las ventas de cada día. Además, se aprovechó esta información para obtener y agregar a las columnas del *dataset* las variables de semana, utilizada en la prueba de concepto y en el trabajo final, y año, que solo fue usada en la prueba de concepto.

Selección de datos. Para filtrar y acotar el *dataset* a una proporción de datos que conllevara a resultados interpretables y útiles bajo el punto de vista de negocio de "La Empresa", se utilizó la variable de tienda para generar dos subconjuntos de datos, uno filtrando las ventas a una misma tienda y otro filtrando las ventas a las tiendas pertenecientes a una misma zona geográfica. Las tiendas y zonas ocupadas fueron aquellas que presentaran el mayor volumen de ventas del *dataset*, lo que, según se vio en las secciones 4.4.1 y 4.4.2, corresponden a la tienda Arauco Maipú para Ropa Interior, la tienda Plaza Oeste para Lavadoras, la tienda Plaza Vespucio para Pantalones y Microondas y la zona geográfica Centro Poniente para todos los *datasets*. Además de esto, se realizó un filtro en base a la cantidad de ventas de un producto, el cual será explicado más adelante.

Transformación del dataset. La finalidad de esta parte es modificar los datos a un formato más cómodo y apropiado para posteriormente ser visualizados y trabajados como series temporales. Para esto, primero fue necesario agregar la información correspondiente a los días que no figuraban en el *dataset* debido a ausencia de ventas de un producto. Esto se hizo para generar series de tiempo continuas, es decir, sin cortes por días sin ventas. Esto se realizó revisando cada uno de los SKUs presentes en el *dataset*. Luego, dado que para un mismo SKU podían haber varias boletas de venta asociadas, se realizó una agregación de estas, sumando la cantidad total vendida en un día y agregando variables adicionales con la información del promedio y la moda del valor al que se vendió el SKU en ese día, que como se mencionó anteriormente, podía no ser siempre el mismo. En primera instancia esta agrupación se realizó por días, y luego se realizó por semanas.

La última transformación del *dataset* fue agrupar los datos de ventas de SKUs de un mismo estilo, ya que se determinó que era más útil estudiar las predicciones de ventas a nivel de estilos y no de SKUs. Llegado a este punto, se aplicó el filtro de ventas mencionado anteriormente, ya que se observó que existían muchos productos que tuvieron una cantidad muy baja de ventas, lo cual no aportaría información suficiente para realizar el estudio. Por

este motivo, se establecieron cotas inferiores de ventas: se eliminaron todos los productos, identificados por su estilo, que tuvieron menos de 30 unidades vendidas en todo el año y también aquellos productos que en las primeras 20 semanas tuvieran menos de 10 unidades vendidas. Por último, se decidió excluir por completo las semanas correspondientes a la víspera de Navidad, es decir, se conservaron únicamente las semanas 1 a 48, ya que todos los productos experimentaban alzas de ventas en este periodo, y, debido a que el enfoque está en las series con comportamientos distinguibles, se estimó pertinente descartar la información de ventas de semanas posteriores a la semana 48.

Normalización por mitades. Debido a la heurística definida para la simulación de predicción de ventas, fue necesario aplicar normalizaciones a las series de tiempo. Para cada producto se separó la serie temporal en dos mitades, la primera con las ventas de la semana 1 a la 24, y la segunda con las ventas de la semana 25 a la 48. Luego, para cada mitad, se calculó la suma total de productos vendidos y se procedió a dividir las ventas de cada semana por dicho valor. Esto generó una normalización "por mitades", la cual expresa finalmente la distribución del porcentaje de la cantidad total vendida en cada mitad, quedando cada valor semanal entre 0 y 1.

Suavizado de series temporales. Este es un proceso típico que se realiza en el trabajo de series temporales. El objetivo es disminuir los impactos de los *peaks* y *off-peaks* de las series en la comprensión de su comportamiento. El suavizado o *smoothing* (*a.k.a. rolling window*) aplicado en este caso fue de una ventana de 3 y 5 periodos, pero finalmente, dado que se consideran series temporales de 24 semanas, se estimó que era más útil ocupar la ventana de 3 periodos, ya que 5 periodos abarcaría una ventana demasiado grande en consideración con el tamaño de la serie temporal. Después de haber calculado los datos suavizados, se realizó la normalización mencionada anteriormente sobre dichos datos.

Luego de este preprocesamiento de datos, a modo de ilustrar el formato final generado, se muestra en la imagen 5.6 los datos de un producto, a modo de ejemplo. Este corresponde al formato para ser tratado como serie de tiempo al que se quería llegar.

semana	venta_semanal_total	venta_normalizada	smooth3	smooth3_normalizado	smooth5	smooth5_normalizado	precio_venta_prom	precio_venta_moda
1	7	0.0493	7.000000	0.0508	5.666667	0.0416	##90.0	##90.0
2	7	0.0493	5.666667	0.0412	5.000000	0.0367	##90.0	##90.0
3	3	0.0211	4.333333	0.0315	4.600000	0.0338	##90.0	##90.0
4	3	0.0211	3.000000	0.0218	3.800000	0.0279	##90.0	##90.0
5	3	0.0211	3.000000	0.0218	3.600000	0.0264	##90.0	##90.0
6	3	0.0211	4.000000	0.0291	3.200000	0.0235	##90.0	##90.0
7	6	0.0423	3.333333	0.0242	3.400000	0.0250	##90.0	##90.0
8	1	0.0070	3.666667	0.0266	3.400000	0.0250	##90.0	##90.0
9	4	0.0282	2.666667	0.0194	3.400000	0.0250	##90.0	##90.0
10	3	0.0211	3.333333	0.0242	3.000000	0.0220	##90.0	##90.0
11	3	0.0211	3.333333	0.0242	3.600000	0.0264	##90.0	##90.0
12	4	0.0282	3.666667	0.0266	3.600000	0.0264	##90.0	##90.0
13	4	0.0282	4.000000	0.0291	4.400000	0.0323	##90.0	##90.0
14	4	0.0282	5.000000	0.0363	5.000000	0.0367	##90.0	##90.0
15	7	0.0493	5.666667	0.0412	5.000000	0.0367	##90.0	##90.0
16	6	0.0423	5.666667	0.0412	5.000000	0.0367	##14.0	##90.0
17	4	0.0282	4.666667	0.0339	5.000000	0.0367	##90.0	##90.0
18	4	0.0282	4.000000	0.0291	4.000000	0.0294	##90.0	##90.0
19	4	0.0282	3.333333	0.0242	4.800000	0.0353	##90.0	##90.0
20	2	0.0141	5.333333	0.0387	6.000000	0.0441	##90.0	##90.0
21	10	0.0704	7.333333	0.0533	6.600000	0.0485	##90.0	##90.0
22	10	0.0704	9.000000	0.0654	12.400000	0.0911	##90.0	##90.0
23	7	0.0493	16.666667	0.1211	15.000000	0.1102	##90.0	##90.0
24	33	0.2324	20.000000	0.1453	16.666667	0.1224	##07.0	##90.0
25	8	0.0741	9.000000	0.0831	6.666667	0.0630	##15.0	##90.0
26	10	0.0926	6.666667	0.0615	6.500000	0.0614	##90.0	##90.0
27	2	0.0185	6.000000	0.0554	6.600000	0.0624	##90.0	##90.0
28	6	0.0556	5.000000	0.0462	5.800000	0.0548	##90.0	##90.0
29	7	0.0648	5.666667	0.0523	4.800000	0.0454	##90.0	##90.0
30	4	0.0370	5.333333	0.0492	4.800000	0.0454	##90.0	##90.0
31	5	0.0463	3.666667	0.0338	4.000000	0.0378	##90.0	##90.0
32	2	0.0185	3.000000	0.0277	3.200000	0.0302	##90.0	##90.0
33	2	0.0185	2.333333	0.0215	2.800000	0.0265	##90.0	##90.0
34	3	0.0278	2.333333	0.0215	2.400000	0.0227	##90.0	##90.0
35	2	0.0185	2.666667	0.0246	2.400000	0.0227	##90.0	##90.0
36	3	0.0278	2.333333	0.0215	3.200000	0.0302	##24.0	##90.0
37	2	0.0185	3.666667	0.0338	3.400000	0.0321	##90.0	##90.0
38	6	0.0556	4.000000	0.0369	4.400000	0.0416	##90.0	##90.0
39	4	0.0370	5.666667	0.0523	5.600000	0.0529	##90.0	##90.0
40	7	0.0648	6.666667	0.0615	5.400000	0.0510	##90.0	##90.0
41	9	0.0833	5.666667	0.0523	4.800000	0.0454	##79.0	##90.0
42	1	0.0093	4.333333	0.0400	4.400000	0.0416	##90.0	##90.0
43	3	0.0278	2.000000	0.0185	3.600000	0.0340	##57.0	##90.0
44	2	0.0185	2.666667	0.0246	2.400000	0.0227	##90.0	##90.0
45	3	0.0278	2.666667	0.0246	3.600000	0.0340	##57.0	##90.0
46	3	0.0278	4.333333	0.0400	4.400000	0.0416	##57.0	##90.0
47	7	0.0648	5.666667	0.0523	5.000000	0.0472	##62.0	##90.0
48	7	0.0648	7.000000	0.0646	5.666667	0.0535	##33.0	##90.0

Figura 5.6: Ejemplo del formato final de los datos de un producto.

5.4. Cálculo de similitud entre series temporales

Para encontrar de manera rigurosa las series temporales más cercanas cuantitativamente, se estudió una gran variedad de medidas de distancia, mencionadas en el Marco Teórico (sección 2.4). Mediante esta investigación se concluyó que las dos métricas más apropiadas para realizar las pruebas son la distancia Euclidiana y la DTW. Con los resultados de la prueba de concepto (sección 5.2.2), se determinó que la distancia Euclidiana asocia de mejor forma las series temporales similares.

Con esto, se procedió a calcular la distancia Euclidiana entre las primeras mitades (datos de ventas de las semanas 1 a 24) de las series de tiempo de cada uno de los productos de un mismo *dataset*. Los cálculos se realizaron tanto para los datos de ventas normalizados y suavizados con ventana de 3 periodos, así como también para los datos sin normalizar, también para los sin suavizar y para los suavizados con ventana de 5 periodos, pero lo que se utilizó finalmente fueron los cálculos de distancia sobre los datos de venta normalizados con ventana de suavizado de 3 periodos.Teniendo estas distancias calculadas, dichos valores fueron dispuestos en lo que se conoce como matriz de distancias, que justamente permitió relacionar cada posible par de series temporales. Debido a que los *datasets* contienen muchos productos, no fue posible generar una visualización coherente de la mencionada matriz de distancias, sin embargo, esta matriz quedó almacenada en una tabla que sí fue de utilidad.

Contando con esta matriz de distancias, se procedió a buscar en cada experimento los N pares de series temporales más cercanas en su primera mitad, i.e. que tuvieran el menor valor de la distancia Euclidiana calculada. Para lograr esto, se convirtió la matriz en una lista, que luego se ordenó de menor a mayor valor de distancia y luego se obtuvieron los primeros N pares y los identificadores de dichas series temporales, que servirían para seleccionar las series completas (primera y segunda mitad) y continuar al siguiente paso de la simulación de predicción de ventas.

5.5. Simulación de pronóstico de series temporales

El enfoque adoptado para la realización de la simulación del pronóstico de ventas explicado a continuación, está enmarcado bajo el supuesto inicial de que dos series temporales que tengan un comportamiento de ventas similar en la primera mitad del año, también tendrán un comportamiento similar en la segunda mitad del año.

Dicho esto, y a partir de lo obtenido en la etapa anterior, se procedió a usar los identificadores, de los productos de los N pares de series temporales con la menor distancia calculada en su primera mitad, para obtener sus series temporales completas. Luego se ocuparon los datos de ventas de las semanas 25 a 48, para considerar únicamente la segunda mitad del año para simular las predicciones.

La idea general, explicada para un solo par de series temporales, es la siguiente: se toman los datos de las ventas del segundo producto para considerarlos como la predicción de ventas del primer producto. Dicho de otra forma, el producto 1 del par se asignó como la serie original (y_true), mientras que el producto 2 del par se estableció como la predicción de ventas (y_pred).

En esta simulación de pronóstico de ventas, fue necesario utilizar métricas clásicas para evaluar predicciones, tales como el MAE, el MAPE y el RMSE, de las cuales se habló en el Marco Teórico (sección 2.5). De dichas métricas, la más apropiada para evaluar este modelo resultó ser el MAE, puesto que sus resultados permitieron obtener mejores conclusiones y presentaban un mayor grado de interpretabilidad con respecto al MAPE y el RMSE. Para comparar la eficiencia de esta simulación de predicciones, se realizaron, para cada uno de los N pares seleccionados, simulaciones de predicciones con otras 10 series temporales de productos aleatorios del *dataset*. La finalidad de hacer esto es poder demostrar que la heurística de predicción propuesta tiene mejores resultados que al realizar predicciones con series aleatorias. Para esto, se calcularon nuevamente las métricas del MAE, MAPE y RMSE para estas 10 predicciones adicionales y se sacaron sus promedios para representar de manera unificada la simulación de predicción aleatoria.

Teniendo las métricas de evaluación para los N pares más cercanos (MAE_Pred) y sus respectivos promedios de los pares aleatorios (MAE_Rand_Prom), se procedió a utilizar la diferencia de los MAEs, MAE_Dif = MAE_Rand_Prom − MAE_Pred para evaluar la calidad del modelo con respecto a la predicción aleatoria como sigue: mientras más grande fue la diferencia de MAEs, MAE_Dif, se consideró que el modelo predecía mejor, ya que, para esto, MAE_Rand_Prom sería mayor que MAE_Pred.

Por último, se realizan experimentos variando el valor de N para considerar un diferente número de pares de series temporales cercanas a considerar en la simulación de pronóstico de ventas. Se obtienen diferentes valores de MAE_Dif para cada valor de N, los cuales son graficados para apreciar la evolución de la diferencia de MAEs a medida que se consideran más pares de series temporales. Con todo lo anterior realizado, ya se estuvo en condiciones para obtener conclusiones con respecto a la heurística propuesta, las cuales se reportarán y comentarán en profundidad en las secciones 6.1 y 6.2.

Capítulo 6

Resultados y Análisis

En este capítulo se reportan todos los resultados propios de la heurística propuesta para la simulación de pronóstico de ventas en base a la similitud de las series temporales. Se verán los gráficos más relevantes, junto a una interpretación y análisis de dichos resultados.

También se presentará una discusión sobre la calidad de los experimentos realizados y finalmente se hablará sobre cómo esta solución propuesta es validada con respecto a los criterios de aceptación previamente establecidos.

6.1. Resultados e interpretación de los experimentos

A continuación se mostrarán los principales resultados obtenidos en cuanto a las categorías de vestuario y electrodomésticos, así como una interpretación y análisis de los mismos. El *dataset* de Ropa Interior representará a la categoría de vestuario, mientras que el *dataset* de Lavadoras representará a la categoría de electrodomésticos. Los estudios realizados para los *datasets* de Pantalones y de Microondas respaldan los análisis y conclusiones realizadas, ya que arrojaron resultados similares, los cuales se pueden consultar en los apéndices A y B, respectivamente.

La forma en la que se presentarán los resultados será primero para una tienda en específico, la de mayor volumen ventas en su respectiva zona principal, y luego para la zona geográfica de mayor cantidad de ventas.

6.1.1. Pronóstico de ventas en una tienda específica del *dataset* de Ropa Interior

La tienda seleccionada corresponde a Arauco Maipú y, luego de aplicar los diversos filtros de ventas y transformaciones, su conjunto de datos contempla 626 estilos de productos. A continuación de presentan los resultados más relevantes.

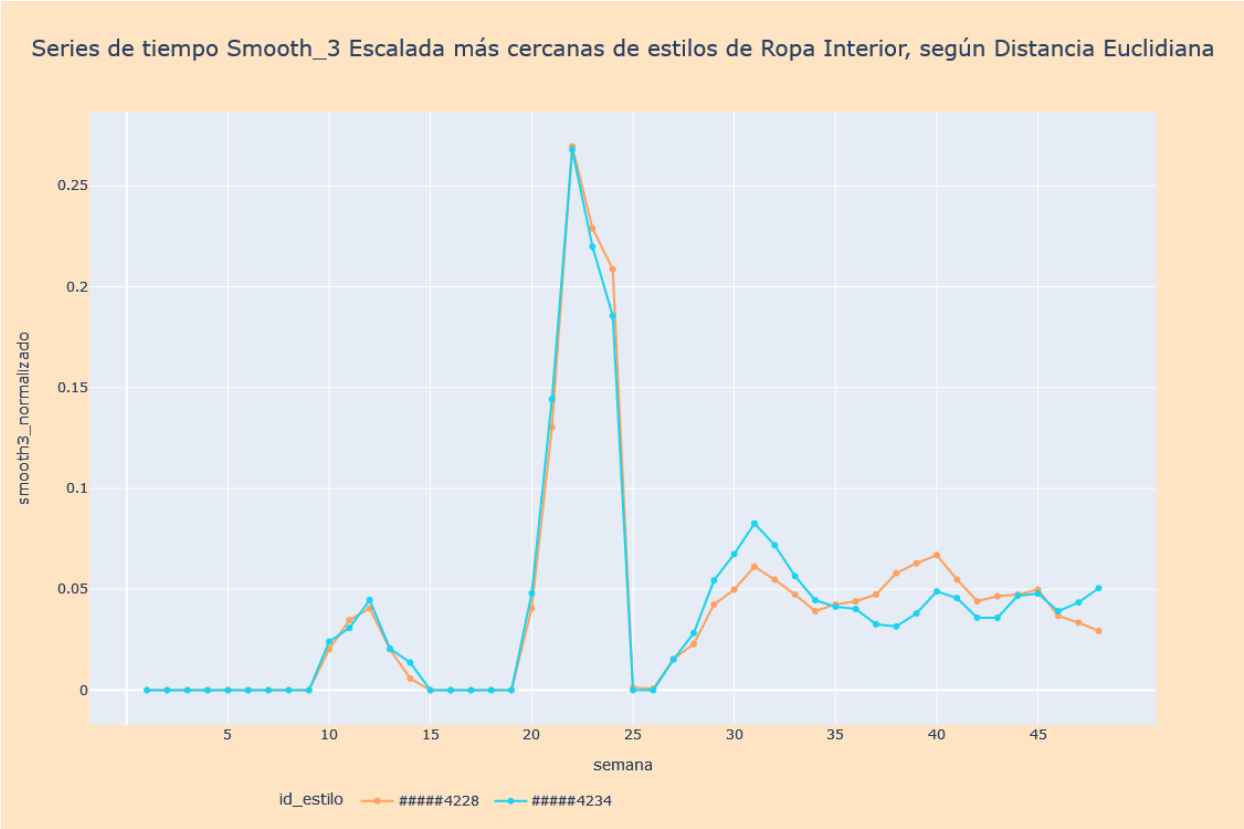


Figura 6.1: Primer par de series temporales más cercanas en el *dataset* de Ropa Interior en la tienda Arauco Maipú.

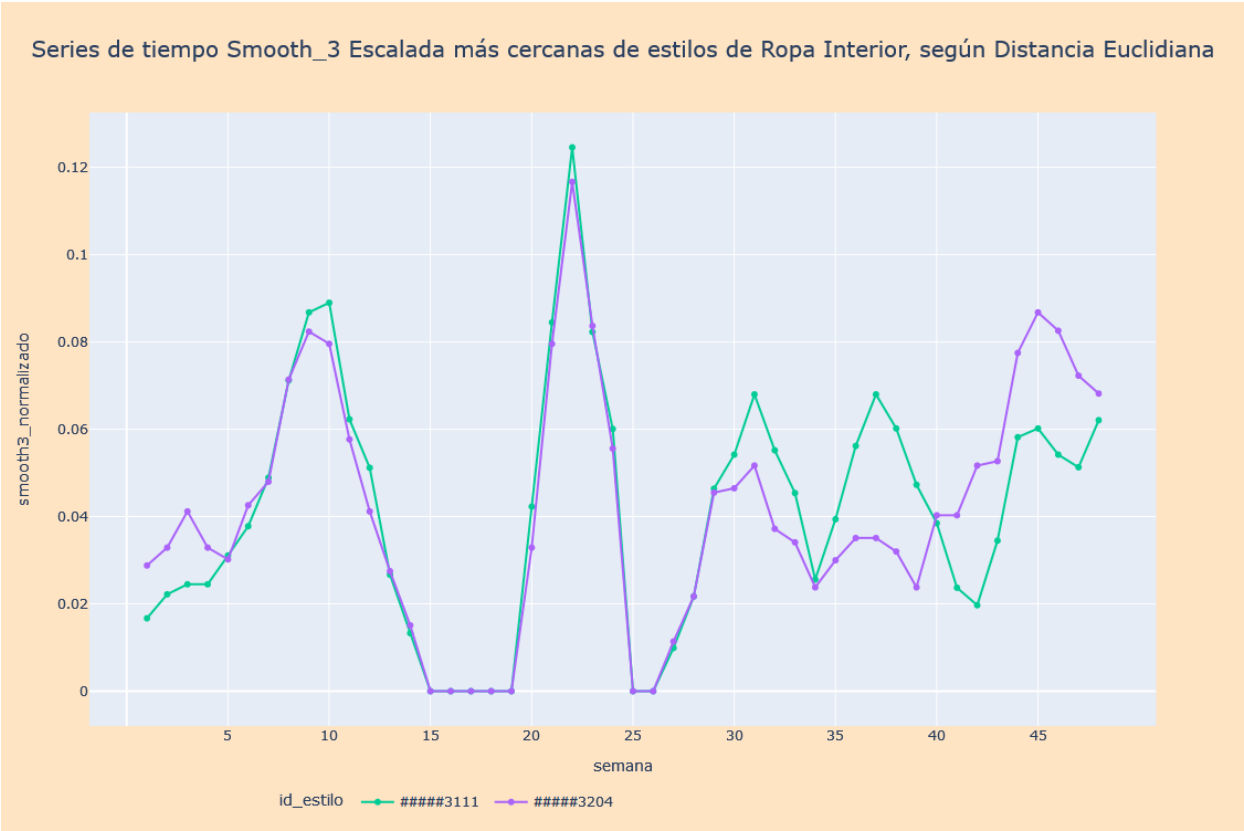


Figura 6.2: Segundo par de series temporales más cercanas en el *dataset* de Ropa Interior en la tienda Arauco Maipú.

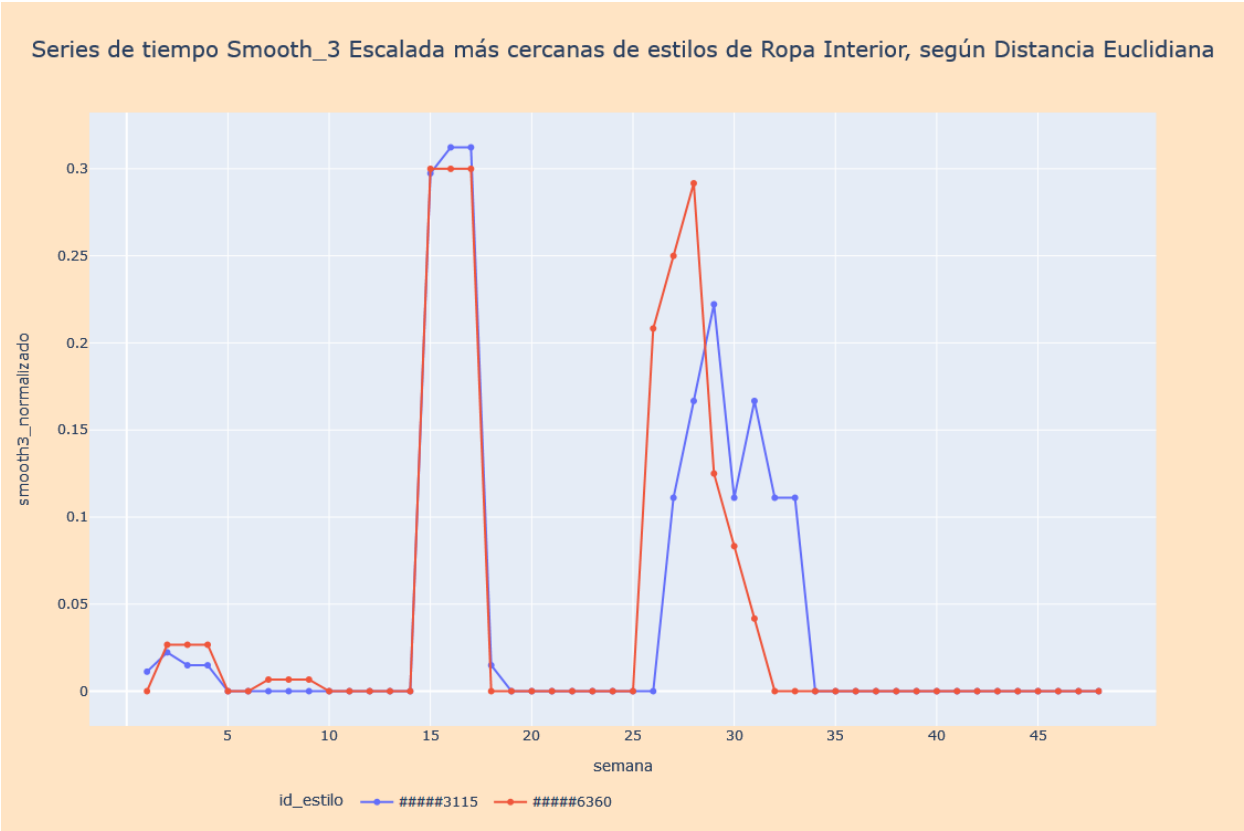


Figura 6.3: Tercer par de series temporales más cercanas en el *dataset* de Ropa Interior en la tienda Arauco Maipú.

A partir de los gráficos 6.1, 6.2 y 6.3, se observa que los 3 pares de series más cercanas son notoriamente similares entre sí, tanto en su primera mitad (semanas 1 a 24), como en su segunda mitad (25 a 48), mas no se parecen entre los pares, sino solo entre sí mismos. Esto apoya la idea de que es posible hacer predicciones de venta a partir de la similitud en la primera mitad del año.

En cuanto a los valores de distancia Euclidiana obtenidos entre los pares, se tienen los siguientes resultados.

Estilo 1	**Estilo 2**	**Distancia**
#####4228	#####4234	0.031263
#####3111	#####3204	0.032631
#####3115	#####6360	0.033005

Tabla 6.1: Distancias Euclidianas de los 3 pares de estilos de Ropa Interior más cercanos en la tienda

En la tabla 6.1 notamos que todos los pares se encuentran a distancias muy parecidas entre sí, a pesar de tener distintas formas y tendencias. Para investigar con respecto a las características físicas de los productos, se consultó su información relevante, dispuesta en la siguiente tabla.

Estilo	Clase	Sub-Clase	Descripción	Marca	Temporada	Precio Venta Promedio
#####3115	CALZONCILLOS	BOXER TEJIDO PLANO	BOXER TEJI BOX MIC BL	MARCA-1	S/T	##79.42
#####6360	CALZONCILLOS	BOXER TEJIDO PLANO	BOXER TEJIDO PU BOX CL GLO	MARCA-1	S/T	##00.77
#####3111	CALZONCILLOS	BOXER TEJIDO PUNTO	BOXER TEJIDO PU BX 3 PR7 B	MARCA-2	S/T	##17.54
#####3204	CALZONCILLOS	BOXER TEJIDO PUNTO	BOXER TEJIDO PU BX 3 PR8 B	MARCA-2	S/T	##85.31
#####4228	CALZONCILLOS	BOXER TEJIDO PUNTO	BOXER TEJIDO PU BX 3 PR1 A	MARCA-3	S/T	###09.52
#####4234	CALZONCILLOS	BOXER TEJIDO PUNTO	BOXER TEJIDO PU BX 3 QLTY2	MARCA-3	S/T	###87.07

Tabla 6.2: Comparación cualitativa de los 3 pares de estilos de Ropa Interior más cercanos en la tienda

Mediante la tabla 6.2 , fue posible comparar cualitativamente los pares de productos más cercanos. Observamos que en este caso, los productos son muy parecidos entre sí, e incluso, los pares más cercanos corresponden a la misma marca y rango de precios. Aún así, se evidencian comportamientos de ventas muy distintos entre pares.

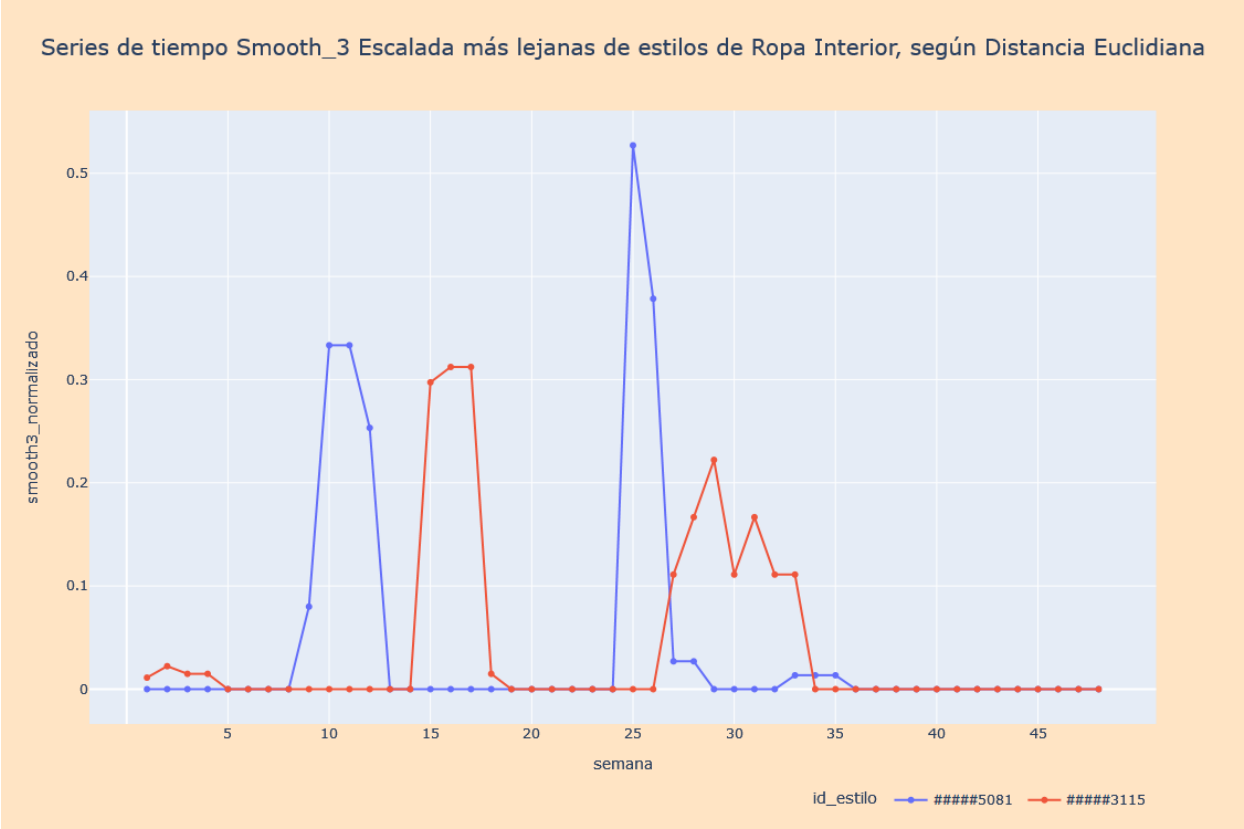


Figura 6.4: Par de series temporales más lejanas en el *dataset* de Ropa Interior en la tienda Arauco Maipú.

Por otra parte, en la imagen 6.4 vemos el par de series que obtuvo el mayor valor de distancia Euclidiana calculado y que, visualmente, se aprecia notoriamente lejano, representando comportamientos de venta distintos, ya que poseen *peaks* de ventas en periodos distintos. Esto queda en mayor evidencia al ver el valor de distancia en la siguiente tabla.

Estilo 1	Estilo 2	Distancia
#####5081	#####3115	0.759958

Tabla 6.3: Distancia Euclidiana del par de estilos de Ropa Interior más lejano en la tienda

A partir de la tabla 6.3, si comparamos el valor obtenido, 0.759, y lo comparamos con el de la distancia entre las series más cercanas, 0.031, notamos que la diferencia es abismal, ya que estamos trabajando en escalas porcentuales. Este mismo par de series hubiera sido catalogado como cercana, bajo el cálculo de la distancia DTW, debido a la aparente similitud en las formas de las series, ratificando la buena decisión de haber optado por la medida de distancia Euclidiana. Con la siguiente tabla podemos comparar cualitativamente estos productos.

Estilo	Clase	Sub-Clase	Descripción	Marca	Temporada	Precio Venta Promedio
#####5081	PIJAMAS	PIJAMA CORTO HOMBRE	PIJAMA ALG 180MAR 19	MARCA-4	V2020	##49.83
#####3115	CALZONCILLOS	BOXER TEJIDO PLANO	BOXER TEJI BOX MIC BL	MARCA-1	S/T	##79.42

Tabla 6.4: Comparación cualitativa del par de estilos de Ropa Interior más lejano en la tienda

En la tabla 6.4 podemos observar que los productos más lejanos en términos de similitud de series temporales de ventas, son productos muy distintos entre sí, ya que pertenecen a distintas clases, subclases, marcas y precios.

Luego de observar los pares de series más cercanos y los más lejanos, se procede con la heurística de simulación de pronóstico de ventas descrita en la sección 5.1.2. Los resultados de esta simulación, para los $N = 10$ pares más cercanos son los siguientes, evaluados por la métrica del MAE.

Num Par Cercano	ID TS Orig	ID TS Pred	MAE Pred	MAE Rand Prom	MAE Dif
1	#####4228	#####4234	0.0101	0.0317	0.0216
2	#####3111	#####3204	0.0143	0.0435	0.0292
3	#####3115	#####6360	0.0394	0.0568	0.0174
4	#####5997	#####7495	0.0126	0.0312	0.0186
5	#####0333	#####7525	0.0159	0.0423	0.0264
6	#####7742	#####0333	0.0163	0.0455	0.0292
7	#####4914	#####6360	0.0213	0.0785	0.0572
8	#####6012	#####7841	0.0213	0.0485	0.0272
9	#####3167	#####3215	0.0397	0.0478	0.0081
10	#####9191	#####3100	0.0284	0.0289	0.0005

Tabla 6.5: Cálculo del MAE en la simulación de pronóstico de ventas de Ropa Interior en la tienda

En la tabla 6.5, podemos ver los resultados ordenados de mayor a menor cercanía entre las series de los pares *TS_ Orig* y *TS_ Pred*. Observamos que en todo momento, la predicción de ventas de las semanas 25 a 48 fue con respecto a la serie de tiempo más cercana fue siempre mejor en comparación al promedio de simulaciones de predicciones con respecto a series aleatorias del mismo *dataset*. Esto evidencia que para Ropa Interior, este modelo predictivo arroja resultados favorables.

Para evaluar los efectos de variar la cantidad de pares a considerar en la simulación de pronóstico de ventas, se generó el siguiente gráfico.

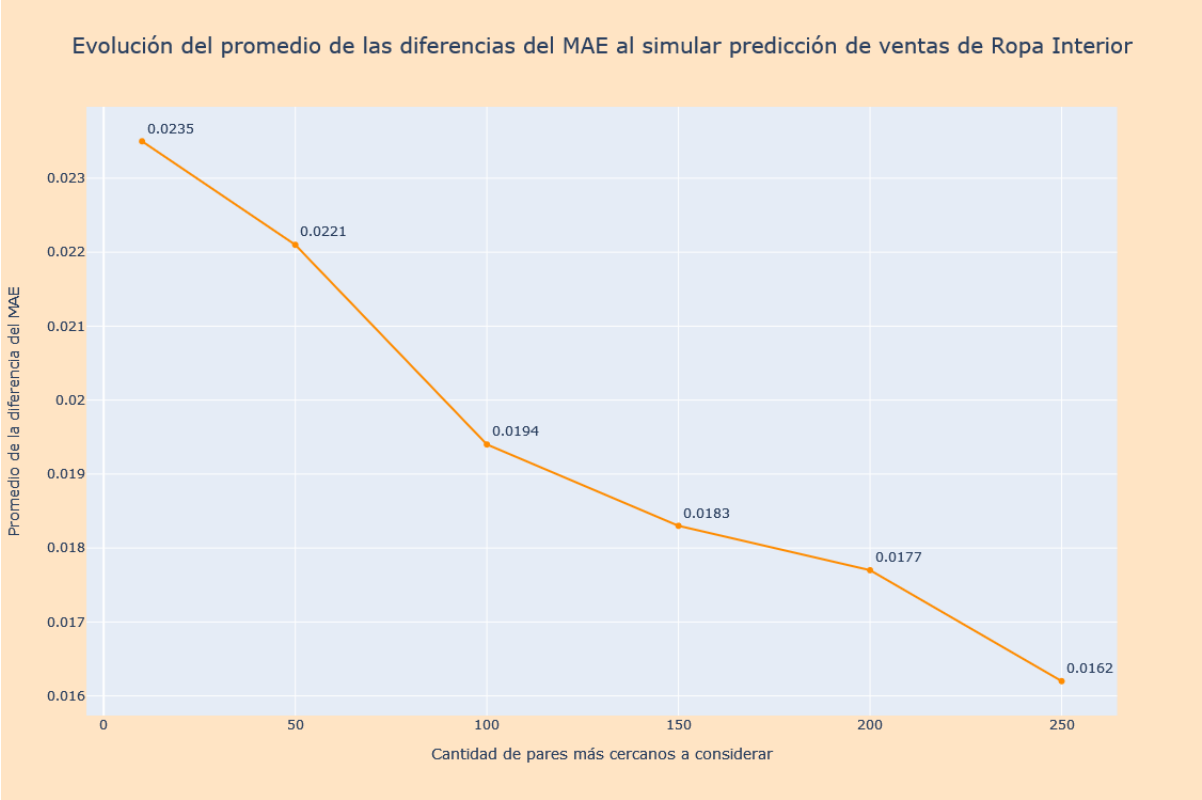


Figura 6.5: Evolución de la diferencia del MAE en las predicciones para el *dataset* de Ropa Interior en la tienda Arauco Maipú.

En la figura 6.5 se observa que a medida que aumenta la cantidad de series a considerar, N, el promedio de las diferencias entre el MAE de la predicción aleatoria y la predicción con la serie más cercana va disminuyendo, lo que quiere decir que, en promedio, la calidad de las predicciones va disminuyendo. Esto calza con lo esperado, ya que al aumentar el N, se empiezan a considerar pares de series que no están tan cercanos entre sí, como lo están los primeros pares encontrados, lo cual va impactando en el promedio de las diferencias del MAE.

6.1.2. Pronóstico de ventas en una zona geográfica del *dataset* de Ropa Interior

La zona de mayor suma de ventas corresponde a la Centro Poniente. Luego de agrupar las ventas de las tiendas de esta zona y aplicar los diversos filtros de ventas y transformaciones, su conjunto de datos contempla 1,088 estilos de productos. Los resultados más relevantes, en comparación con el mismo estudio aplicado a nivel de tiendas, son los siguientes.

Al pasar a replicar el estudio a nivel de agrupación de tiendas de una misma zona, la cantidad de series temporales aumentó considerablemente, lo cual arrojó resultados aún más favorables, como veremos a continuación.

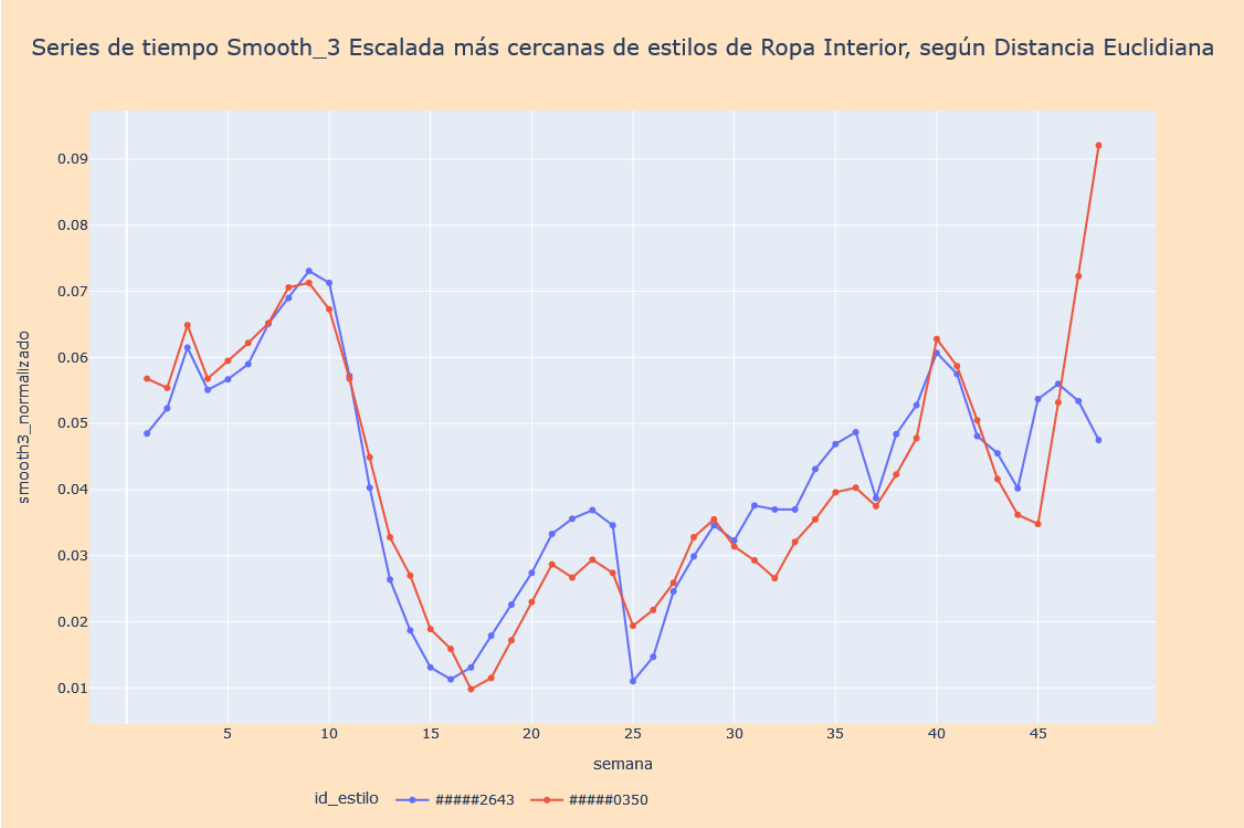


Figura 6.6: Primer par de series temporales más cercanas en el *dataset* de Ropa Interior en la zona Centro Poniente.

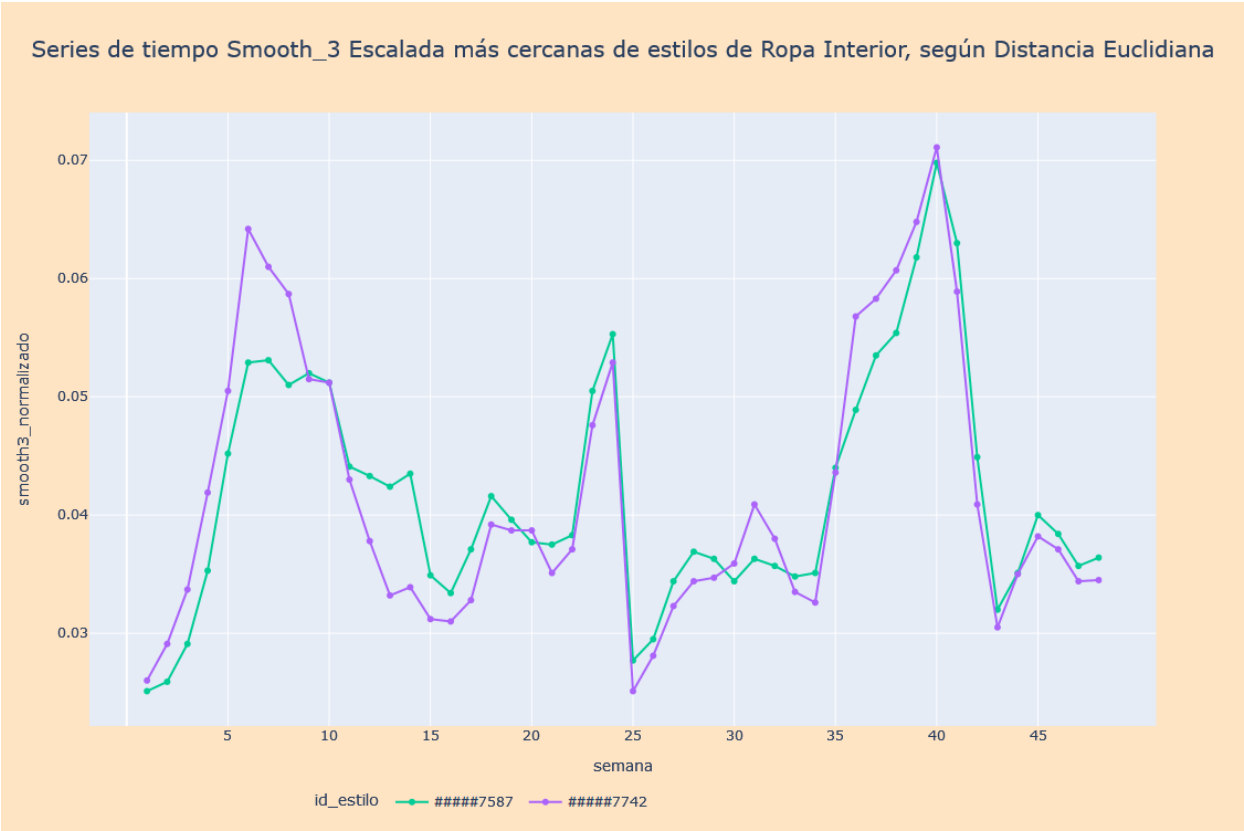


Figura 6.7: Segundo par de series temporales más cercanas en el *dataset* de Ropa Interior en la zona Centro Poniente.

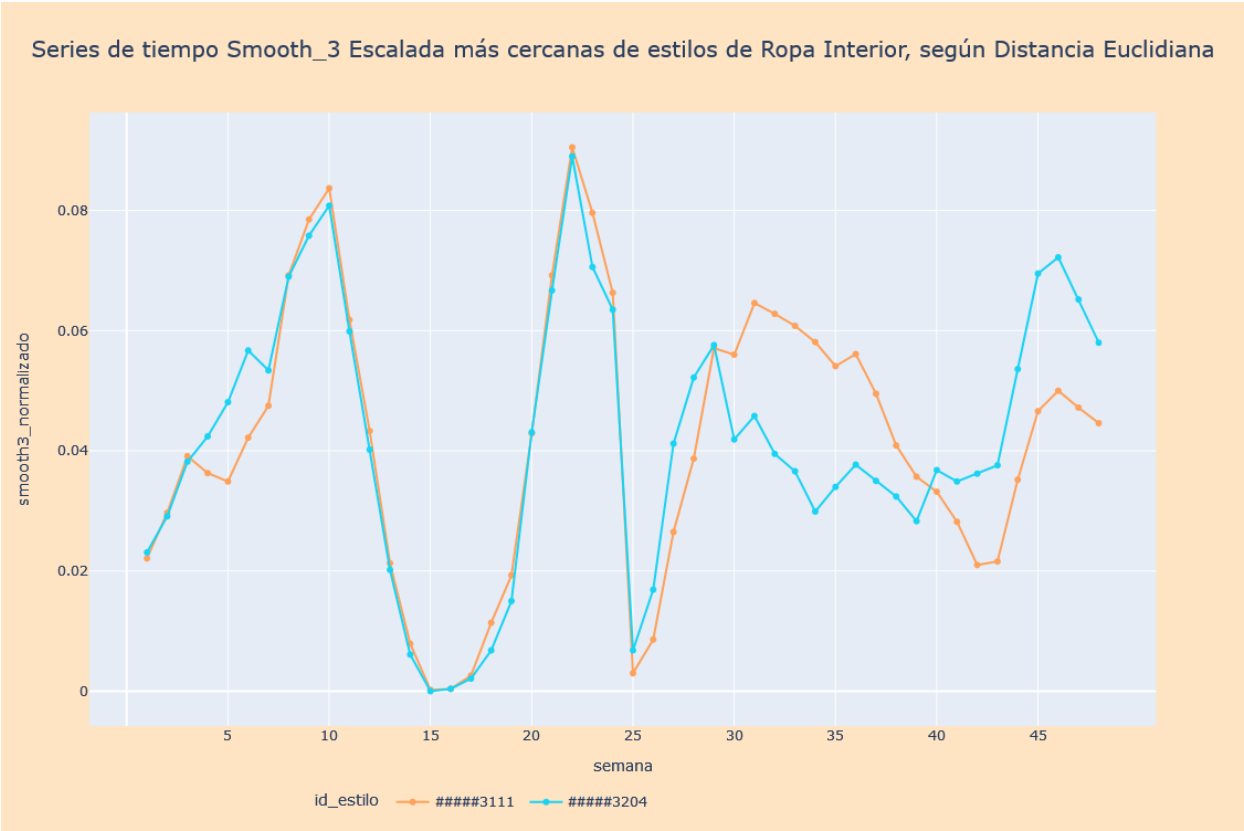


Figura 6.8: Tercer par de series temporales más cercanas en el *dataset* de Ropa Interior en la zona Centro Poniente.

En los gráficos 6.6, 6.7 y 6.8, se mantiene lo observado anteriormente, ya que los pares de series más cercanos se parecen tanto en su primera mitad, como en su segunda mitad. Algo muy particular que se dio es que el par cercano de la figura 6.6 corresponde a productos que no son necesariamente iguales, ya que uno es un pack de calcetines de marca MARCA-5 y el otro corresponde a unos calzoncillos tipo *boxer* de marca MARCA-4, ambos en rangos de precios distintos. Esto fue especialmente revelador en un principio, ya que se aprecia que dos productos distintos pueden presentar comportamientos de ventas muy similares. A su vez, este descubrimiento es de gran relevancia para "La Empresa", ya que esto permite encontrar relaciones no triviales entre productos de distinta clase y subclase.

Por último, se mostrará a continuación un gráfico que permite comparar entre el nivel de tienda y de zona los resultados del efecto de variar el número N de pares a considerar en la simulación de pronóstico de ventas.

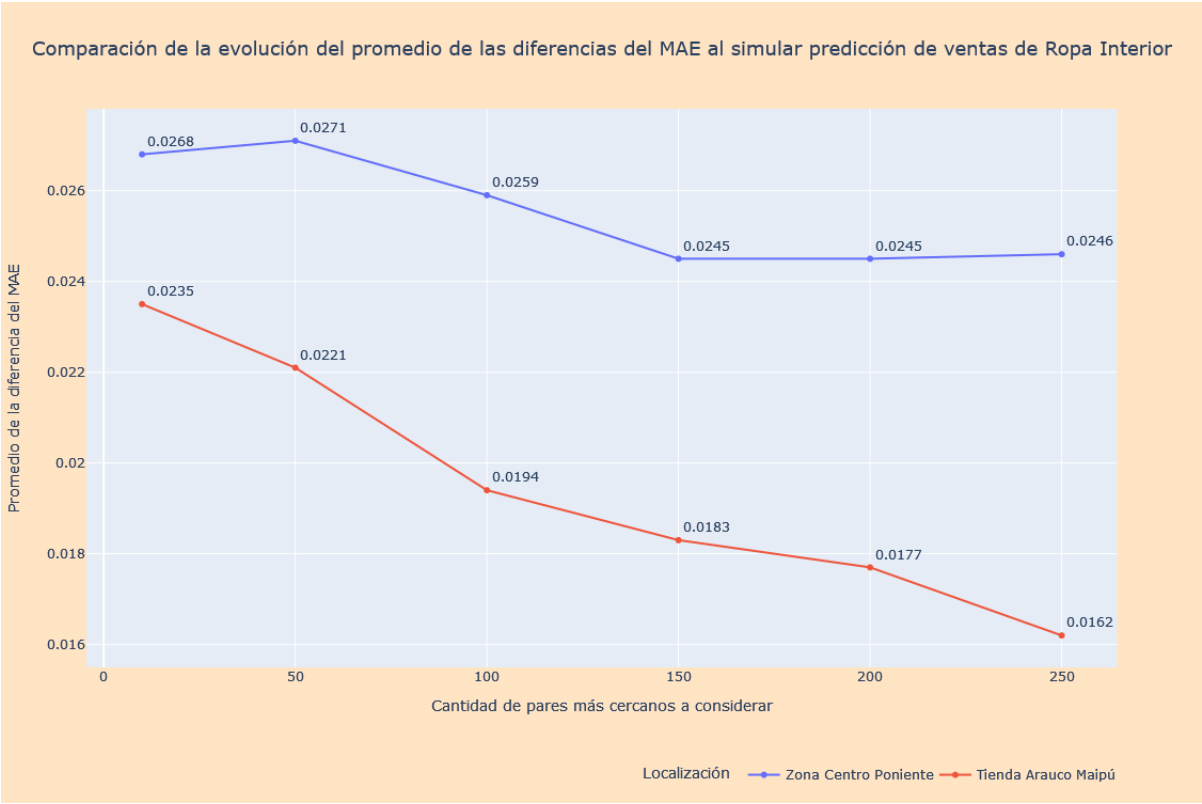


Figura 6.9: Evolución de la diferencia del MAE en las predicciones para el *dataset* de Ropa Interior en la zona Centro Poniente.

En la figura 6.9 podemos apreciar que en la evolución de las diferencias del MAE de la zona ocurre un fenómeno similar al comentado sobre la figura 6.5. Además, podemos observar que en todo momento, la diferencia de MAEs fue mayor a nivel de agrupación de zona con respecto al nivel de la tienda, implicando que las simulaciones de predicciones de ventas son mejores a nivel de zona. Esto permite concluir que es este modelo predictivo podría ser más útil a nivel de zona que a nivel de tiendas. Esta conclusión aporta gran valor a "La Empresa".

6.1.3. Pronóstico de ventas en una tienda específica del *dataset* de Lavadoras

La tienda seleccionada corresponde a Plaza Oeste y, luego de aplicar los diversos filtros de ventas y transformaciones, su conjunto de datos contempla 83 estilos de productos. A continuación de presentan los resultados más relevantes.

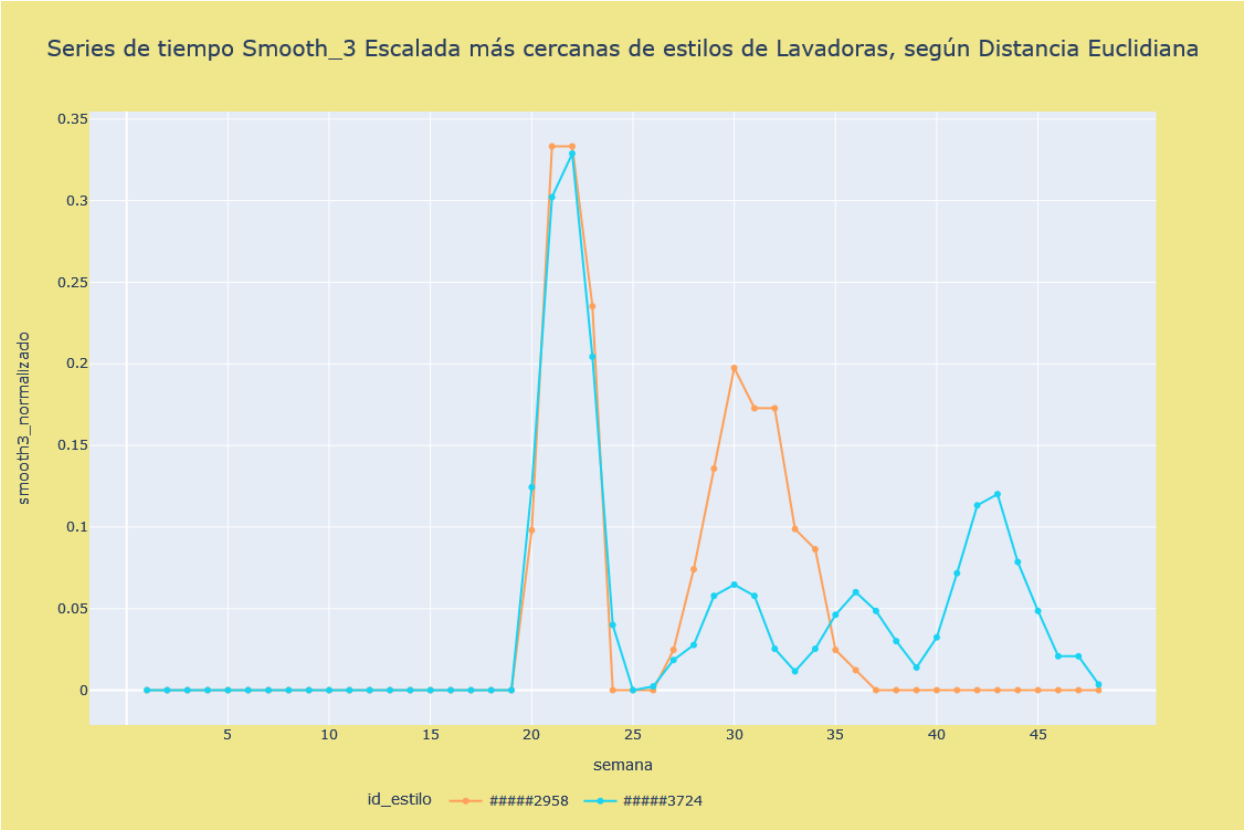


Figura 6.10: Primer par de series temporales más cercanas en el *dataset* de Lavadoras en la tienda Plaza Oeste.

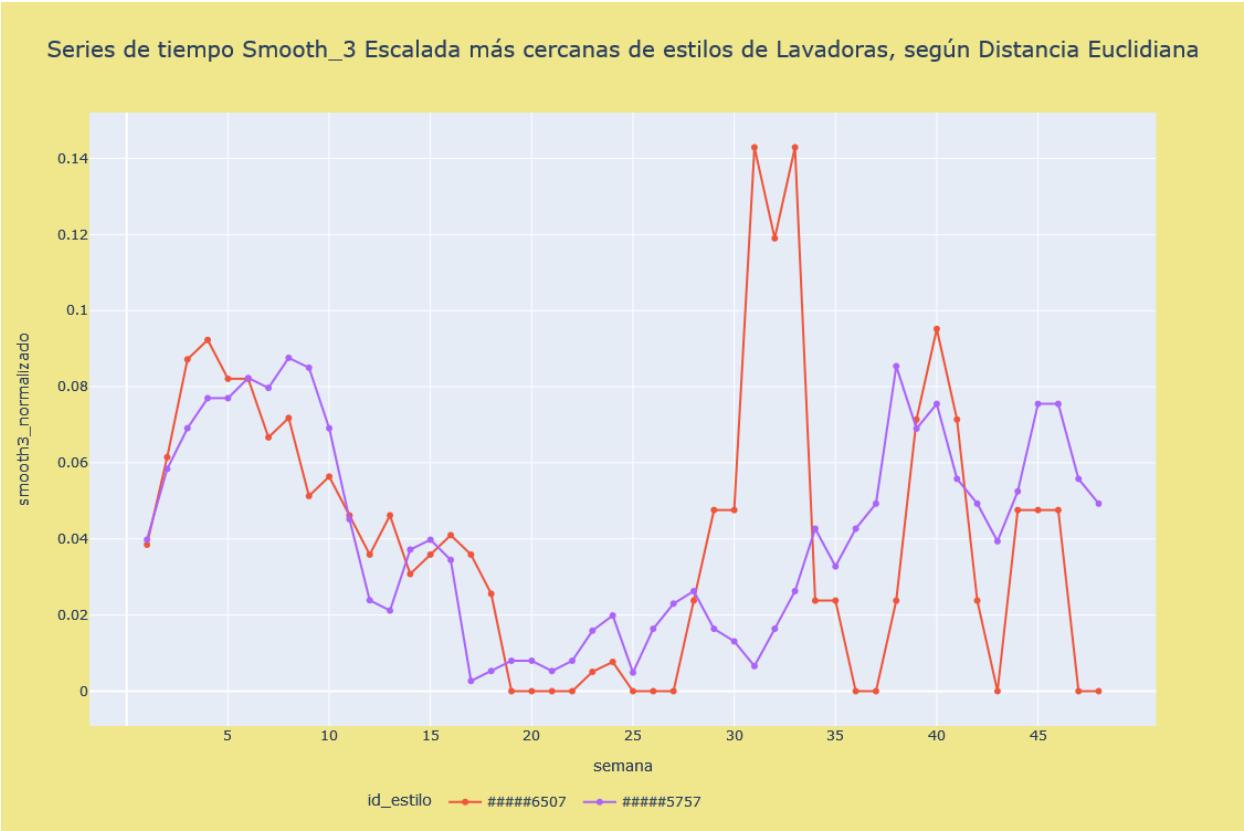


Figura 6.11: Segundo par de series temporales más cercanas en el *dataset* de Lavadoras en la tienda Plaza Oeste.

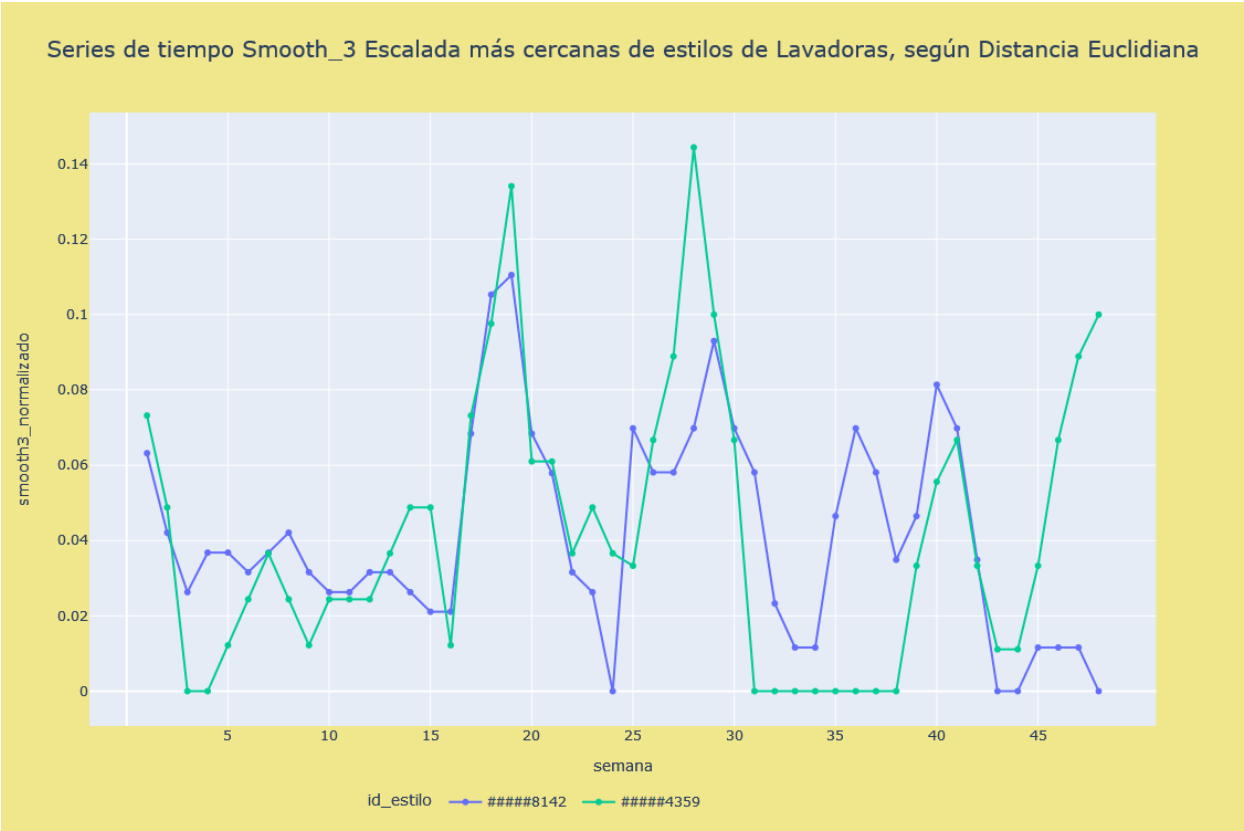


Figura 6.12: Tercer par de series temporales más cercanas en el *dataset* de Lavadoras en la tienda Plaza Oeste.

En los gráficos 6.10, 6.11 y 6.12, vemos que los 3 pares de series más cercanas, verde con azul, celeste con naranja, y roja con morado, son notoriamente similares entre sí, sin embargo, en esta ocasión, se observa que son más semejantes en su primera mitad (semanas 1 a 24), y no tan semejantes en su segunda mitad (25 a 48), al menos como sucedía en el caso de la Ropa Interior. Esto puede implicar que el pronóstico de ventas en base a la similitud de series temporales no sea tan viable para productos electrodomésticos, como lo es para productos de vestuario.

Además, nuevamente se observa que los pares más cercanos no son similares entre otros pares.

En cuanto a los valores de distancia Euclidiana obtenidos entre los pares, se tienen los siguientes resultados.

Estilo 1	**Estilo 2**	**Distancia**
#####2958	#####3724	0.065103
#####6507	#####5757	0.071993
#####8142	#####4359	0.086884

Tabla 6.6: Euclidianas de los 3 pares de estilos de Lavadoras más cercanos en la tienda

En la tabla 6.6 notamos que los pares se encuentran a distancias relativamente parecidas entre sí, pero ligeramente en menor medida que lo visto en Ropa Interior. Para investigar con

respecto a las características físicas de los productos, se consultó su información relevante, como se muestra a continuación.

Estilo	Clase	Sub-Clase	Descripción	Marca	Temporada	Precio Venta Promedio
#####8142	LAVADORAS CARGA SUPERIOR	DE 12KG/ 26 LBS A MAS	LAVADORA WA15J5730LS/ZS(D)	MARCA-A	S/T	###674.68
#####6507	LAVADORAS CARGA SUPERIOR	DE 12KG/ 26 LBS A MAS	LAVADORA S WA19F7L6DDB/ZS	MARCA-A	S/T	###732.95
#####4359	CENTROS DE LAVADO	LAVASECA	LAVASECA DWC-K963(D)	MARCA-B	S/T	###412.95
#####5757	LAVADORAS CARGA SUPERIOR	DE 12KG/ 26 LBS A MAS	LAVADORA S BRILLIANT 15 SG	MARCA-C	S/T	###420
#####2958	SECADORAS	ELECTRICA	SECADORA 8 KG WKR-08K1W3	MARCA-D	S/T	###718.86
#####3724	LAVADORAS CARGA SUPERIOR	DE 12KG/ 26 LBS A MAS	LAVADORA PREMIUM CARE PRO 21X	MARCA-C	S/T	###134.1

Tabla 6.7: Comparación cualitativa de los 3 pares de estilos de Lavadoras más cercanos en la tienda

En la tabla 6.7 podemos notar que en esta ocasión, los productos de los pares de series más cercanas también presentan cierto parecido entre sí, ya que presentan en su mayoría la misma clase y subclase, sin embargo hay una mayor variedad de marcas y una alta variación de precios. Nuevamente vemos que el primer par más cercano está conformado por productos que son distintos, ya que uno es una lavadora de carga superior y el otro es una secadora.

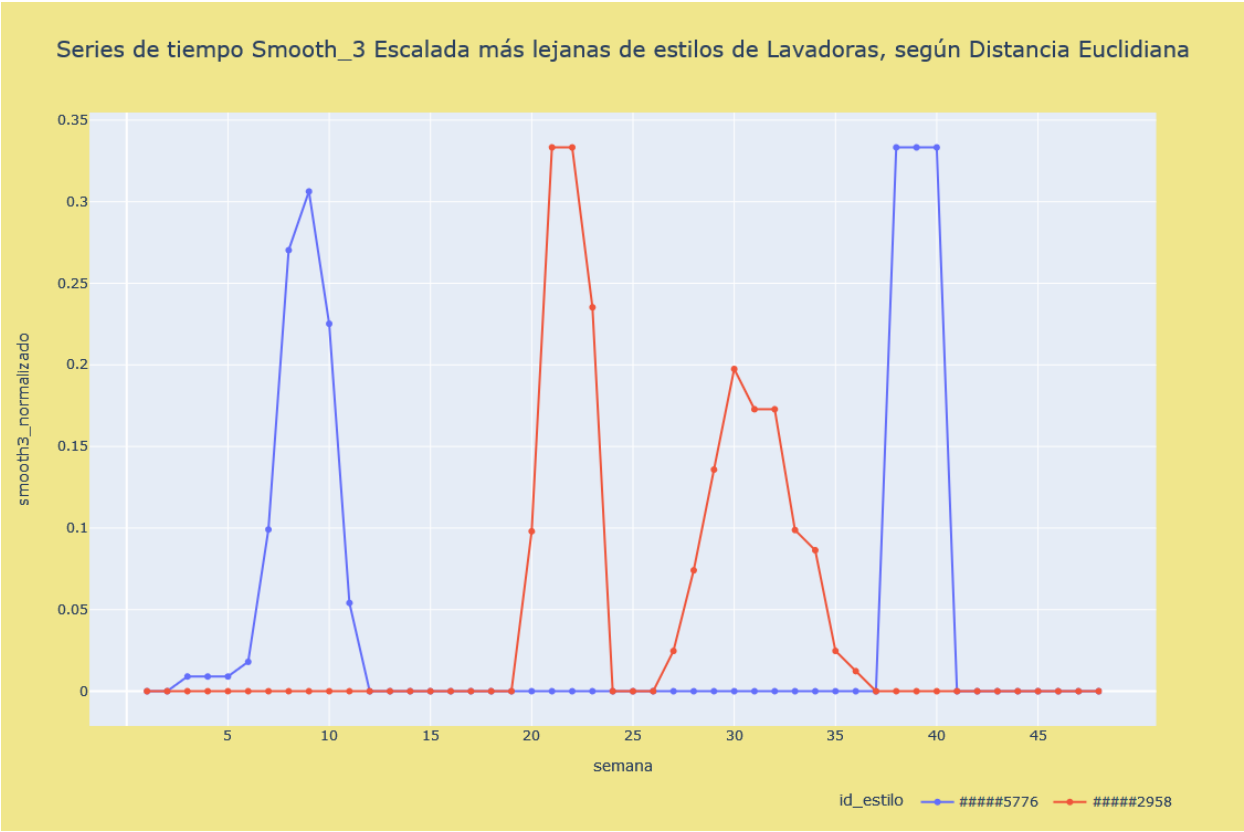


Figura 6.13: Par de series temporales más lejanas en el *dataset* de Lavadoras en la tienda Plaza Oeste.

En cuanto a las series menos similares, i.e. con mayor distancia Euclidiana entre ellas, notamos que están bastante alejadas entre sí, en cuanto a sus periodos de ventas, como se puede apreciar en la figura 6.13. Además, si bien ambas presentan dos *peaks* de ventas muy marcados, el comportamiento de ventas de las series es distinto. En la siguiente tabla se verá esto de manera cuantitativa.

Estilo 1	Estilo 2	Distancia
#####5776	#####2958	0.719763

Tabla 6.8: Distancia Euclidiana del par de estilos de Lavadoras más lejano en la tienda

A partir de la tabla 6.8, si comparamos el valor obtenido, 0.719, y lo comparamos con el de la distancia entre las series más cercanas, 0.065, notamos nuevamente una gran diferencia, considerando la escala porcentual. Con la siguiente tabla podemos comparar cualitativamente estos productos.

Estilo	Clase	Sub-Clase	Descripción	Marca	Temporada	Precio Venta Promedio
#####5776	LAVADORAS CARGA SUPERIOR	DE 12KG/ 26 LBS A MAS	LAVADORA S IMPRESSIVE 18SZ	MARCA-C	S/T	###490.58
#####2958	SECADORAS	ELECTRICA	SECADORA 8 KG WKR-08K1W3	MARCA-D	S/T	###718.86

Tabla 6.9: Comparación cualitativa del par de estilos de Lavadoras más lejano en la tienda

En la tabla 6.9 vemos que los productos más lejanos, son productos muy distintos entre sí, ya que pertenecen a distintas clases, subclases y marcas, sin embargo, es interesante notar que están dentro del mismo rango de precios.

Luego de observar los pares de series más cercanos y los más lejanos, se procede con la heurística de simulación de pronóstico de ventas descrita en el capítulo anterior. Los resultados de esta simulación, para los $N = 10$ pares más cercanos son los siguientes, evaluados por la métrica del MAE.

Num Par Cercano	ID TS Orig	ID TS Pred	MAE Pred	MAE Rand Prom	MAE Dif
1	#####2958	#####3724	0.0562	0.0548	-0.0014
2	#####6507	#####5757	0.0382	0.0467	0.0085
3	#####8142	#####4359	0.0331	0.0338	0.0007
4	#####1360	#####7473	0.0731	0.0588	-0.0143
5	#####8105	#####7547	0.037	0.0409	0.0039
6	#####6747	#####2956	0.0531	0.0423	-0.0108
7	#####2535	#####2957	0.048	0.0409	-0.0071
8	#####4200	#####7547	0.0371	0.0495	0.0124
9	#####0873	#####2535	0.0327	0.0377	0.005
10	#####0380	#####5757	0.0296	0.0346	0.005

Tabla 6.10: Cálculo del MAE en la simulación de pronóstico de ventas de Lavadoras en la tienda

En la tabla 6.10, observamos que las predicciones fueron mucho menos favorables que en el estudio sobre Ropa Interior, ya que en varias ocasiones se tienen diferencias de MAE negativas, es decir, donde simular el pronóstico de venta con respecto a series aleatorias resultó mejor que con respecto al pronóstico realizado con su serie más cercana.

Este fenómeno puede contribuir a la conclusión de que esta heurística de pronóstico de ventas no es tan relevante para los productos de electrodomésticos como lo es para productos de vestuario. Esto queda más en evidencia en el siguiente gráfico.

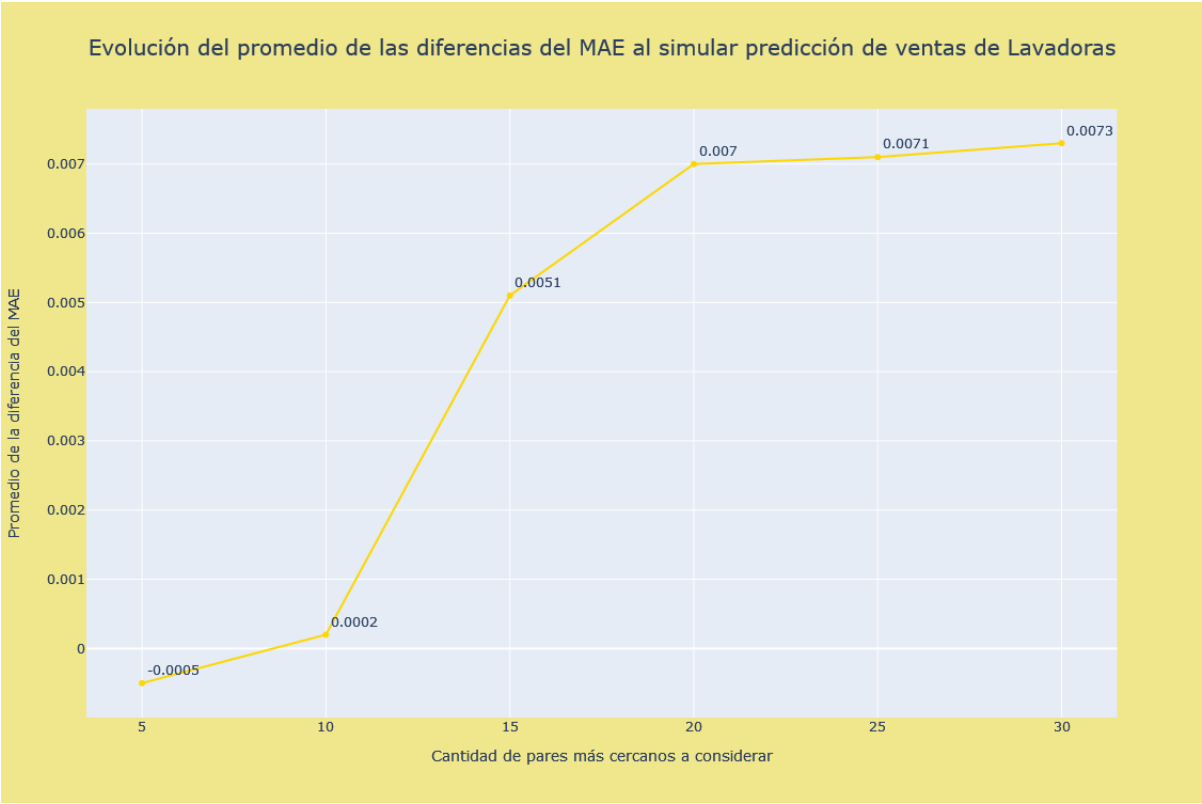


Figura 6.14: Evolución de la diferencia del MAE en las predicciones para el *dataset* de Lavadoras en la tienda Plaza Oeste.

En la figura 6.14 ocurre totalmente lo opuesto a lo observado en la evolución de las diferencias de MAE vista en Ropa Interior (figura 6.5). Vemos que a medida que aumentamos la cantidad de series a considerar, N, el promedio de las diferencias entre el MAE de la predicción aleatoria y la predicción con la serie más cercana va aumentando, lo que quiere decir que, en promedio, la calidad de las predicciones va mejorando, contrario a lo esperado.

6.1.4. Pronóstico de ventas en una zona geográfica del *dataset* de Lavadoras

La zona de mayor suma de ventas corresponde a la Centro Poniente. Luego de agrupar las ventas de las tiendas de esta zona y aplicar los diversos filtros de ventas y transformaciones, su conjunto de datos contempla 144 estilos de productos.

Nuevamente notamos que al pasar a replicar el estudio a nivel de agrupación de tiendas de una misma zona, la cantidad de series temporales aumentó considerablemente, implicando nuevamente una mejora con respecto al estudio realizado para una tienda en específico, tal como se dio en el estudio realizado para el *dataset* de Ropa Interior. Los resultados más relevantes, en comparación con el mismo estudio aplicado a nivel de tiendas, son los siguientes.

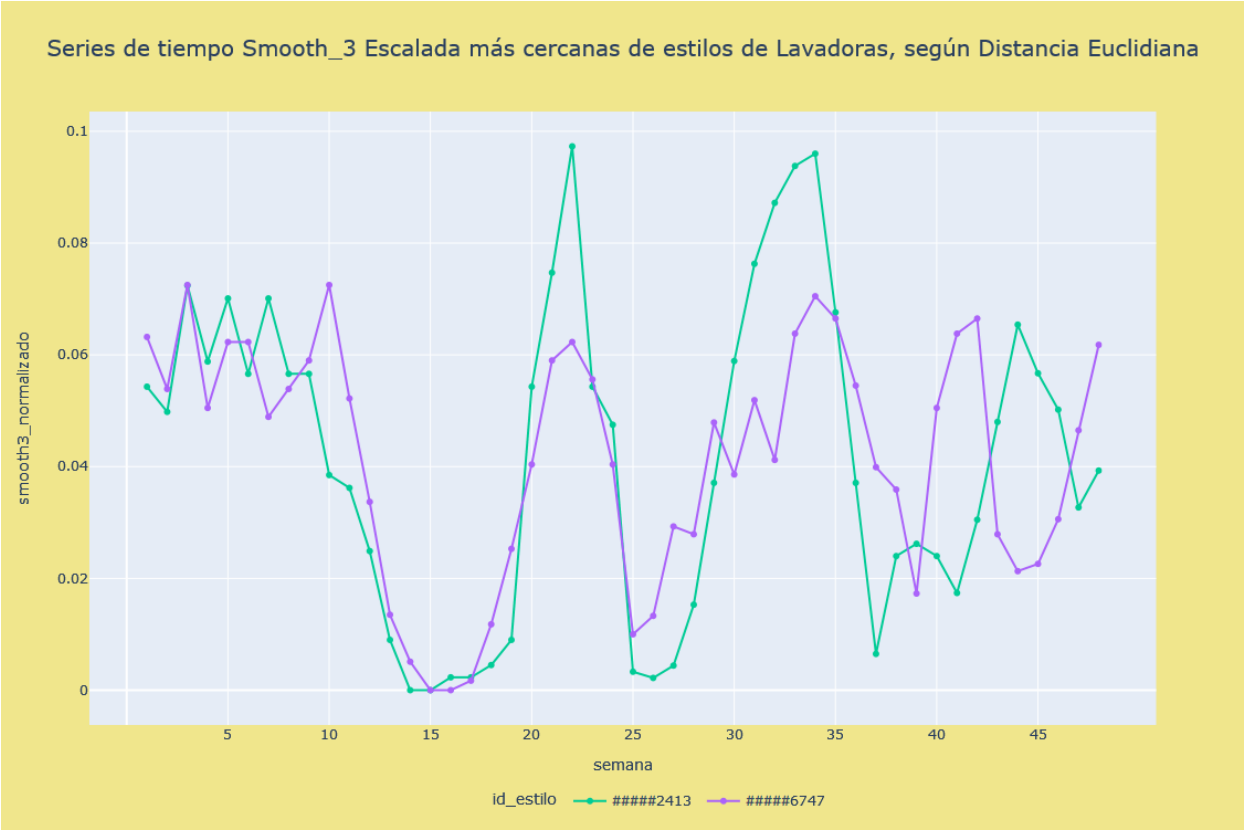


Figura 6.15: Primer par de series temporales más cercanas en el *dataset* de Lavadoras en la zona Centro Poniente.

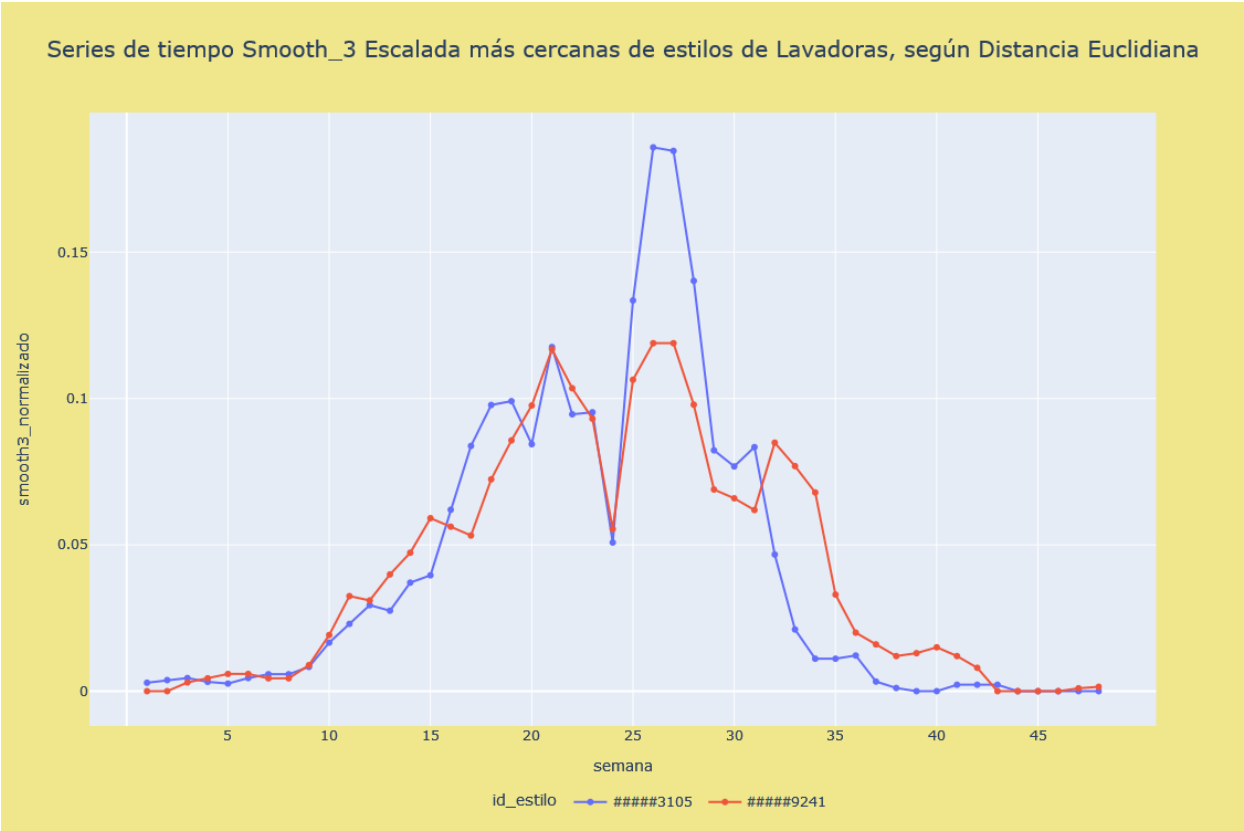


Figura 6.16: Segundo par de series temporales más cercanas en el *dataset* de Lavadoras en la zona Centro Poniente.

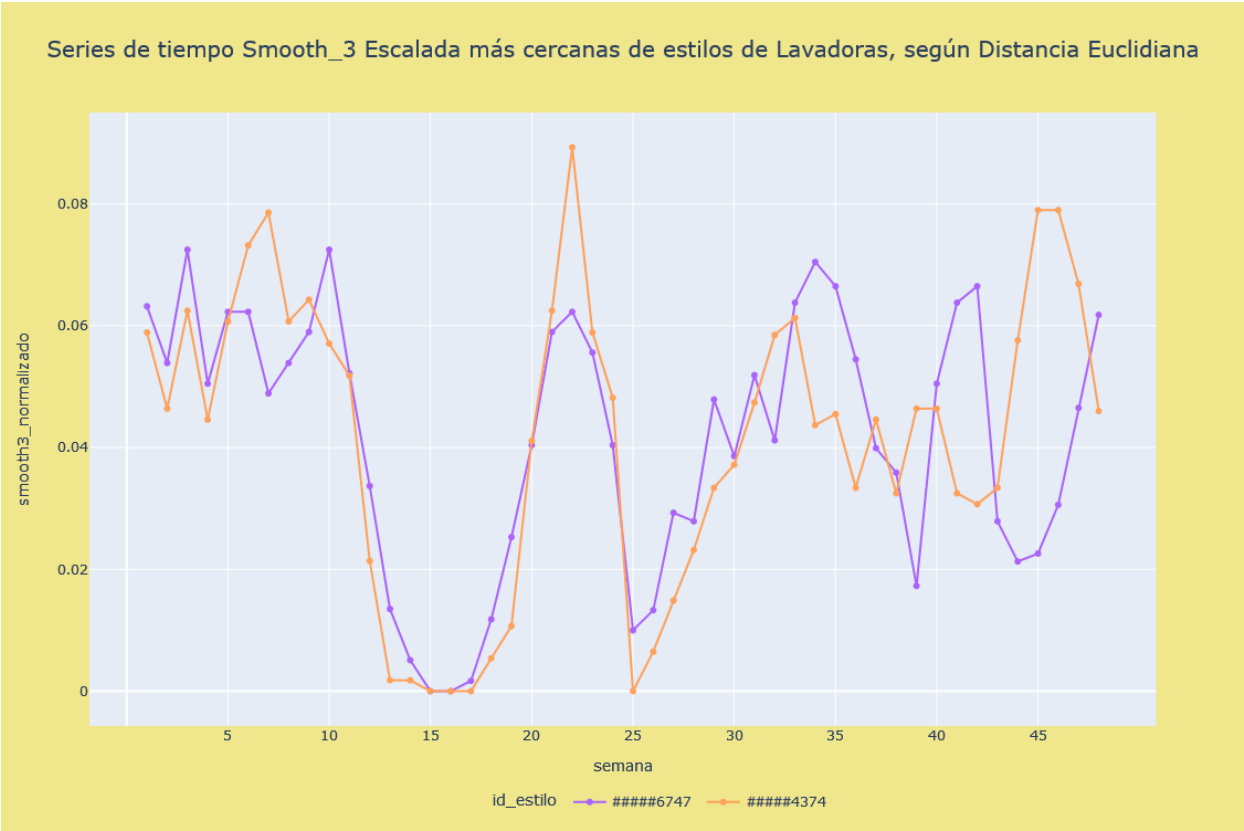


Figura 6.17: Tercer par de series temporales más cercanas en el *dataset* de Lavadoras en la zona Centro Poniente.

En los gráficos 6.15, 6.16 y 6.17, a diferencia de los demás vistos en el estudio aplicado a Lavadoras, sí muestran la existencia de pares de series más cercanos que se parecen tanto en su primera mitad, como en su segunda mitad. Lo cual puede deberse a que el conjunto de productos en este caso era mayor al del usado al aplicar este estudio a nivel de una tienda específica.

Por último, veremos a continuación el gráfico que compara las predicciones realizadas a nivel de tienda y de zona y los efectos de variar el número N de pares a considerar en la simulación de pronóstico de ventas.

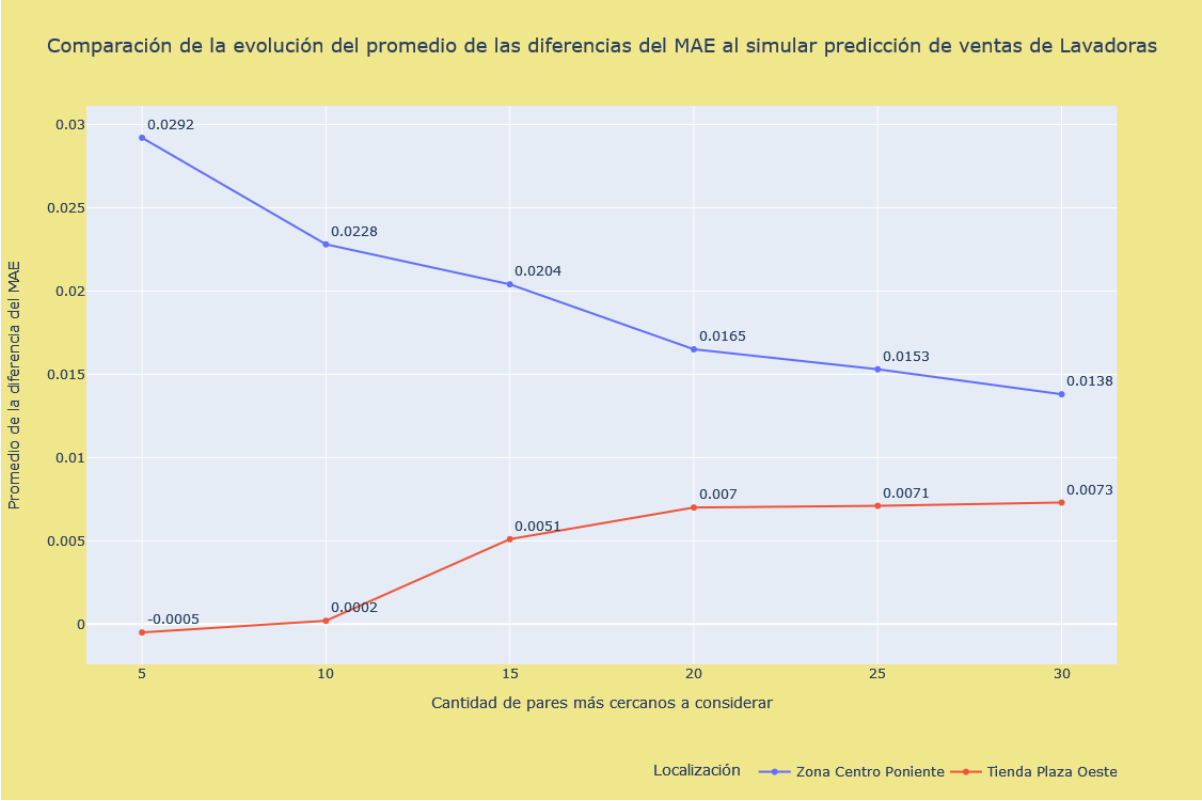


Figura 6.18: Evolución de la diferencia del MAE en las predicciones para el *dataset* de Lavadoras en la zona Centro Poniente.

En la figura 6.18 podemos observar que, a diferencia de lo visto en la figura 6.14, la evolución de las diferencias del MAE de la zona presenta un fenómeno similar al comentado sobre la figura 6.5. Además, nuevamente se aprecia que en todo momento, la diferencia de MAEs fue mayor a nivel de agrupación de zona con respecto al nivel de la tienda, corroborando el supuesto de que las simulaciones de predicciones de ventas son mejores a nivel de zona.

6.2. Discusión

Como se observó en el trabajo reportado, los resultados de las simulaciones de pronóstico de ventas en base a la similitud de series temporales no se ven tan favorables para los productos de electrodomésticos como para los de vestuario, pero esto puede deberse a que la cantidad de productos con los que se trabajó en los *datasets* de electrodomésticos fue considerablemente menor con respecto al de vestuario. Por ejemplo, para el estudio realizado en la tienda específica para Ropa Interior, se contó con 626 estilos, mientras que para la tienda específica para Lavadoras solo se contó con 83 estilos, mientras que para el estudio realizado en la zona geográfica para Ropa Interior se contó con 1,088 estilos, mientras que en la zona geográfica para Lavadoras nada más se contó con 144 estilos. Esto sugiere que sería útil volver a replicar este trabajo, pero comparando entre *datasets* con un tamaño más parecido, de tal

manera de poder realizar observaciones más concluyentes.

Por otro lado, dado que fue posible encontrar series temporales similares en su primera mitad, que permitieron predecir de buena manera las ventas de la segunda mitad, el presente estudio sugiere que es factible la aplicación de técnicas de *clustering*[1] para encontrar los pares de series temporales más cercanos y repetir la heurística presentada para el pronóstico de ventas, pero aplicado sobre los *centroides* de los grupos encontrados al realizar *clustering*.

Otra observación relevante es que el estudio realizado apunta a la evaluación del modelo predictivo en relación a sus propias predicciones realizadas sobre productos aleatorios, sin embargo, para agregar un mayor grado de interpretabilidad del modelo, sería necesario evaluar concretamente las predicciones de ventas pudiendo concluir si estas son aceptables o no.

Por último, surgió la idea de mejorar la generalización de las observaciones y análisis realizados, ya que la heurística para el pronóstico de ventas presentada contempla la predicción en base a la serie temporal más cercana, mientras que una heurística que permita generalizar aún más las conclusiones podría abarcar además la predicción en base a varias series temporales más cercanas, es decir, como un enfoque parecido al del algoritmo de *k-nearest neighbors*[2].

6.3. Validación

De acuerdo a las características deseadas de la solución a desarrollar que fueron indicadas en la sección 3.1.4, se observa que la serie de gráficos, métricas y otras visualizaciones elaboradas en el presente trabajo sirvieron, de manera apropiada y suficiente, en la compresión de la relación entre la similitud de las series temporales con el pronóstico de ventas, ya que, como vimos en los resultados y análisis anteriormente mencionados, fue posible determinar que existen productos cuya similitud de series temporales en la primera mitad del año permitió llevar a cabo una predicción de ventas para la segunda mitad del año. Esto se vio de manera más evidente para el *dataset* de Ropa Interior.

Junto con lo anterior, la solución desarrollada permitió determinar que este estudio resulta más factible para ser realizado sobre *datasets* de productos de la categoría de vestuario y, aparentemente, también para los productos de la categoría de electrodomésticos, aunque en menor medida. Este trabajo además permitió determinar que la mejor forma de replicar la heurística de pronóstico de ventas planteada es teniendo un enfoque a nivel de estilos, más que a nivel de SKUs, y para una zona geográfica, más que para una tienda en específico.

Además de esto, vemos que cada uno de los objetivos específicos señalados en la sección 1.2.2 fueron correctamente logrados, tal como se verá en el recuento de objetivos cumplidos de la sección 7.2. Por este motivo, se da por aceptada la solución propuesta, en el contexto del estudio del pronóstico de ventas en base a la similitud de series temporales de productos en la industria del *retail*.

[1] También conocido como agrupamiento, es una técnica de Aprendizaje No Supervisado cuyo propósito es encontrar en los datos patrones no triviales permitan agruparlos.

[2] También conocido como k-NN, es una técnica de Aprendizaje Supervisado utilizada en los problemas de clasificación y regresión, que utiliza la información de los k datos más cercanos.

Capítulo 7

Conclusiones y Trabajo Futuro

En este capítulo se resumirá brevemente el trabajo llevado a cabo, junto con un repaso de los objetivos planteados inicialmente que fueron logrados. También se mencionan las principales conclusiones obtenidas, además de reflexiones y observaciones sobre el impacto y relevancia del trabajo realizado. Por último, se comentarán las posibles soluciones adicionales a probar como parte del trabajo futuro que surge a partir del presente trabajo.

7.1. Resumen del trabajo realizado

En el presente trabajo se llevó a cabo una exploración inicial de los todos los datos presentes en los 4 *datasets* de productos disponibles. Esto tuvo la finalidad de comprender mejor los datos y la realidad de "La Empresa" en el contexto de la industria del *retail*, y sirvió para determinar filtros útiles para aplicar a los datos de manera preliminar. Luego se realizó una prueba de concepto que tuvo por finalidad encontrar el formato más adecuado, con los filtros y transformaciones adicionales necesarias, para poder realizar análisis de series de tiempo con los datos de ventas históricos disponibles. Además, esto fue útil para determinar la existencia de series temporales similares entre sí, así como encontrar una métrica para evaluar cuantitativamente la similitud entre las series.

Por último se desarrolló el trabajo final relacionado con la simulación del pronóstico de ventas en base a la similitud entre las series temporales, donde se calcularon matrices de distancia para cada posible par de series, se obtuvieron los N pares de series más cercanas en sus datos de ventas de las semanas 1 a 24 y se simuló la predicción de ventas con los datos de las semanas 25 a 48, dejando afuera las siguientes semanas debido a que todas las series presentaban el mismo patrón de comportamiento. Se evaluaron estas simulaciones de predicciones, comparándolas con el escenario de predicción con respecto a series aleatorias y se concluyó sobre la capacidad predictiva del modelo propuesto, mediante gráficos y métricas.

El trabajo realizado permitió evaluar la factibilidad de realizar pronósticos de ventas a partir de la demanda histórica de productos de una misma categoría con comportamientos similares en el tiempo, concluyendo que sí es factible de realizar.

7.2. Revisión de objetivos y conclusiones

A continuación se realizará un recuento del cumplimiento de los objetivos indicados en la sección 1.2.

- **Primer objetivo**. Logrado. La fase de análisis exploratorio reportada en el capítulo 4 permitió acotar la cantidad de datos inicialmente disponibles, mediante filtros de tiendas y precios.

- **Segundo objetivo**. Logrado. La prueba de concepto descrita en la sección 5.2 permitió descubrir transformaciones útiles sobre los *datasets* originales, tales como agrupaciones de venta por semana, adición de los días sin ventas, recorte de las semanas 49 en adelante, agrupación de SKUs de un mismo estilo, entre otras. Con esto se logró tener un conjunto de datos apropiado para la realización del resto del trabajo.

- **Tercer objetivo**. Logrado. La prueba de concepto permitió establecer la distancia Euclidiana como la métrica más adecuada para el presente análisis de series temporales.

- **Cuarto objetivo**. Logrado. Como se observó en los análisis relacionados con la figura 6.6 y la tabla 6.7, se comprobó la existencia de productos cualitativamente distintos que tuvieron un comportamiento de ventas similar.

- **Quinto objetivo**. Logrado. Tal como fue detallado en la sección 5.1.2, fue posible definir una heurística para la simulación del pronóstico de ventas para los pares de series más cercanas.

- **Sexto objetivo**. Logrado. En la sección 6.1 se reportaron los resultados de los experimentos realizados, donde se utilizó una gran variedad de visualizaciones, gráficos de series temporales, tablas comparativas y cálculo de métricas de predicción que permitieron evaluar el modelo propuesto.

- **Séptimo objetivo**. Logrado. El estudio realizado fue replicado en cada uno de los *datasets* disponibles: Ropa Interior (secciones 6.1.1 y 6.1.2), Lavadoras (secciones 6.1.3 y 6.1.4), Pantalones (apéndice A) y Microondas (apéndice B). Además, se pudo concluir que la heurística de pronóstico de ventas presentada es más apropiada para vestuario que para electrodomésticos.

Dicho esto, al haber cumplido con la totalidad de los objetivos específicos, junto con lograr satisfacer los criterios de aceptación de la solución, como se mencionó en la sección 6.3, es que se da por cumplido el objetivo general señalado en la sección 1.2.1.

Como conclusiones adicionales a las ya vistas en las secciones de interpretación de los experimentos llevados a cabo (6.1) y de discusión sobre el trabajo realizado (6.2), se pueden concluir con respecto a la utilidad que representa el trabajo bajo el punto de vista de negocio de "La Empresa" es grande, ya que, finalmente, este estudio implica la posibilidad de predecir de manera porcentual la distribución de ventas en el segundo semestre de un producto nuevo, es decir, sin datos históricos de venta, más allá de la semana 24. Esto debido a que se observó que existen productos para los cuales es posible predecir el porcentaje de ventas de cada una de las semanas del segundo semestre en base a los porcentajes de ventas del primer semestre.

7.3. Trabajo futuro

En el trabajo realizado, la heurística para la simulación del pronóstico de ventas fue realizada sin mezclar los productos de distintos *datasets*: siempre se trabajó exclusivamente con Pantalones o con Ropa Interior. Sin embargo, resulta interesante la idea de replicar este mismo estudio sobre series temporales de productos de una misma categoría, e incluso de categorías distintas, para ver si es posible obtener conclusiones adicionales sobre la posibilidad de realizar pronósticos de ventas en base a la similitud de series de productos heterogéneos.

Tal como se mencionó en la sección 6.2, se propone adaptar el enfoque del algoritmo de *k-nearest neighbors* para aplicarlo en la simulación de predicciones de ventas, seleccionando no solo la serie temporal más similar para un producto, sino más bien las k series temporales más similares a ella, y replicar las predicciones sobre dicho conjunto, para obtener métricas promedio que permitan generalizar de mejor forma la capacidad predictiva del modelo propuesto.

En cuanto al margen de mejora del presente estudio realizado, se propone la idea de conseguir una métrica más apropiada para la interpretación de la calidad general de los pronósticos de ventas realizados. Si bien fueron utilizadas las métricas MAE, MAPE y RMSE, de momento, estas métricas solo sirvieron al propósito de comparar los rendimientos entre las predicciones realizadas, sin embargo, se hace necesaria alguna métrica adicional que permita determinar si la predicción se acepta como buena o no. Paralelamente, se podría estudiar a más a fondo el modelo predictivo actual de "La Empresa" y comparar sus predicciones de ventas, en contraste con la heurística propuesta en este trabajo, con el fin de obtener un umbral de aceptación para los valores del MAE calculado.

Por último, se sugiere aprovechar la conclusión obtenida en este trabajo, de que es posible realizar pronósticos de ventas en base a la similitud de series temporales, para aplicar y evaluar algoritmos de *clustering* para encontrar grupos de series de tiempo cercanas y verificar si es que dichos grupos siguen alguna tendencia marcada, o si los *centroides* de dichos *clusters* permiten mejorar las predicciones de ventas de la heurística propuesta. Dicho esto, el trabajo a realizar por "La Empresa" es continuar el presente estudio y elaborar un modelo predictivo funcional basado en la similitud de las series temporales de ventas, con la finalidad de mejorar su modelo actual.

Bibliografía

[1] Alexandra Amidon. How to apply k-means clustering to time series data. URL: https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3, Jul. 2020. [Online; visitado el 13-04-2021].

[2] José Manuel Arreaza. Calendario de campañas comerciales en Chile 2022: los mejores momentos de venta del año. URL: https://marketing4ecommerce.cl/campanas-comerciales-en-chile/, Ene. 2022. [Online; visitado el 24-09-2022].

[3] Jan Belke. Data Science terms explained, part I. URL: https://medium.com/@datadrivenscience/data-science-terms-explained-part-i-9eaf2cae6b16, Feb. 2020. [Online; visitado el 21-01-2021].

[4] Carmelo Cassisi, Placido Montalto, Marco Aliotta, Andrea Cannata, and Alfredo Pulvirenti. *Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining*, chapter 3, pages 70–82. IntechOpen, Sep. 2012. doi:10.5772/49941.

[5] ICR Chile. Análisis de Industria Retail - Reporte sectorial. URL: https://www.icrchile.cl/index.php/destacados/4287-proyecciones-sobre-la-industria-del-retail-continuan-desfavorables-con-bajas-expectativas-de-un-fortalecimiento-operacional-durante-2021/file, Jul. 2020. [Online; visitado el 13-01-2022].

[6] Wikimedia Commons. Euclidean matching vs. Dynamic Time Warping matching. URL: https://commons.wikimedia.org/wiki/File:Euclidean_vs_DTW.jpg, Sep. 2011. [Online; visitado el 16-04-2021].

[7] Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 894–903. PMLR, Ago. 2017. URL: http://proceedings.mlr.press/v70/cuturi17a.html.

[8] Xie dairu and Zhang Shilong. Machine Learning Model for Sales Forecasting by using XGBoost. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 480–483, 2021. doi:10.1109/ICCECE51280.2021.9342304.

[9] Sourav Dash. Smoothing techniques for Time Series data. URL: https://medium.com

/@srv96/smoothing-techniques-for-time-series-data-91cccfd008a2, May. 2020. [Online; visitado el 17-06-2022].

[10] Consuelo de la Jara. Cuarentena podría afectar al retail más que el estallido social. URL: https://www.dii.uchile.cl/wp-content/uploads/2020/03/24-03-202-El-Mercurio-Inversiones-Cuarentena-podr%C3%ADa-afectar-al-retail-m%C3%A1s-que-el-estallido-social.pdf, Mar. 2020. [Online; visitado el 24-09-2022].

[11] Departamento de Estudios de la Cámara Nacional de Comercio. Pese a enfrentar una alta base de comparación las ventas del retail marcan un alza promedio de 19.2 % anual en las semanas de diciembre. URL: https://www.cnc.cl/pese-a-enfrentar-una-alta-base-de-comparacion-las-ventas-del-retail-marcan-un-alza-promedio-de-192-anual-en-las-semanas-de-diciembre/, Ene. 2022. [Online; visitado el 21-01-2022].

[12] Paul Dix. What is time series data? URL: https://www.influxdata.com/what-is-time-series-data/, Jun. 2020. [Online; visitado el 17-03-2021].

[13] Efficient Costumer Response Community. Optimal Shelf Availability. URL: https://ecr-community.org/wp-content/uploads/2016/10/ecr-europe-osa-optimal-shelf-availability.pdf, 2003. [Online; visitado el 22-09-2022].

[14] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3):37, Mar. 1996. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/1230, doi:10.1609/aimag.v17i3.1230.

[15] Maarten Grootendorst. 9 Distance Measures in Data Science. URL: https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa, Feb. 2021. [Online; visitado el 05-04-2021].

[16] Jiawei Han, Micheline Kamber, and Jian Pei. Getting to Know Your Data. In Jiawei Han, Micheline Kamber, and Jian Pei, editors, *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 39–82. Morgan Kaufmann, Boston, third edition edition, 2012. URL: https://www.sciencedirect.com/science/article/pii/B9780123814791000022, doi:https://doi.org/10.1016/B978-0-12-381479-1.00002-2.

[17] Jose Martinez Heras. La Maldición de la Dimensión en Machine Learning. URL: https://www.iartificial.net/la-maldicion-de-la-dimension-en-machine-learning/, Sep. 2020. [Online; visitado el 07-04-2021].

[18] Gaoxia Jiang, Wenjian Wang, and Wenkai Zhang. A novel distance measure for time series: Maximum shifting correlation distance. *Pattern Recognition Letters*, 117:58–65, 2019. URL: https://www.sciencedirect.com/science/article/pii/S0167865518308985, doi:https://doi.org/10.1016/j.patrec.2018.11.013.

[19] A. Kianimajd, M.G. Ruano, P. Carvalho, J. Henriques, T. Rocha, S. Paredes, and A.E. Ruano. Comparison of different methods of measuring similarity in physiologic time

series. *IFAC-PapersOnLine*, 50(1):11005–11010, 2017. 20th IFAC World Congress. URL: https://www.sciencedirect.com/science/article/pii/S2405896317333967, doi:https://doi.org/10.1016/j.ifacol.2017.08.2479.

[20] Clifford Lam. Challenges to Time Series Analysis in the computer age. In *Statistics: Discovering your future power*. Qian Meng (ed.), China Statistics Press, 2012. URL: http://stats.lse.ac.uk/lam/bookarticle1. [Online; visitado el 27-03-2021].

[21] Vinh-Trung Luu, Germain Forestier, Jonathan Weber, Paul Bourgeois, Fahima Djelil, and Pierre-Alain Muller. A review of alignment based similarity measures for web usage mining. *Artificial Intelligence Review*, 53(3):1529–1551, Mar. 2020. doi:10.1007/s10462-019-09712-9.

[22] Nehal Magdy, Mahmoud Sakr, Tamer Abdelkader, and Khaled Elbahnasy. Review on trajectory similarity measures. In *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 613–619, 2015. doi:10.1109/IntelCIS.2015.7397286.

[23] Tetsuya Nakamura, Keishi Taki, Hiroki Nomiya, Kazuhiro Seki, and Kuniaki Uehara. A Shape-based Similarity Measure for Time Series Data with Ensemble Learning. *Pattern Analysis and Applications*, 16, 11 2012. doi:10.1007/s10044-011-0262-6.

[24] Emiliano Carrizo Ortiz. Las ventas del comercio sufrieron en 2019 la mayor caída de su historia producto del estallido social. URL: https://www.latercera.com/pulso/noticia/las-ventas-del-comercio-sufrieron-en-2019-la-mayor-caida-de-la-historia-producto-del-estallido-social/991091/, Ene. 2020. [Online; visitado el 24-09-2022].

[25] Nicola Petty. Understanding time series analysis. URL: https://creativemaths.net/videos/video-time-series/, Feb. 2014. [Online; visitado el 22-03-2021].

[26] Selva Prabhakaran. Cosine Similarity – Understanding the math and how it works. URL: https://www.machinelearningplus.com/nlp/cosine-similarity/, Oct. 2018. [Online; visitado el 03-05-2021].

[27] Vinit Saini. Overview of the Data Science Pipeline. URL: https://dzone.com/articles/overview-of-the-data-science-pipeline, Nov. 2018. [Online; visitado el 22-01-2021].

[28] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi:10.1109/TASSP.1978.1163055.

[29] Ana Isabel Sordo. Los 10 factores que afectan el comportamiento de tus consumidores. URL: https://blog.hubspot.es/marketing/factores-comportamiento-del-consumidor, May. 2022. [Online; visitado el 24-09-2022].

[30] Levente Szász, Csaba Bálint, Ottó Csíki, Bálint Zsolt Nagy, Béla-Gergely Rácz, Dénes Csala, and Lloyd C. Harris. The impact of COVID-19 on the evolution of online retail:

The pandemic as a window of opportunity. *Journal of Retailing and Consumer Services*, 69:103089, 2022. URL: https://www.sciencedirect.com/science/article/pii/S0969698922001825, doi:https://doi.org/10.1016/j.jretconser.2022.103089.

[31] Romain Tavenard. Dynamic Time Warping - tslearn documentation. URL: https://tslearn.readthedocs.io/en/stable/user_guide/dtw.html, Jun. 2020. [Online; visitado el 15-04-2021].

[32] Xiao Wang, Fusheng Yu, and Witold Pedrycz. An area-based shape distance measure of time series. *Applied Soft Computing*, 48:650–659, 2016. URL: https://www.sciencedirect.com/science/article/pii/S1568494616303131, doi:https://doi.org/10.1016/j.asoc.2016.06.033.

[33] Hang Xu, Wenhua Zeng, Xiangxiang Zeng, and Gary Yen. An evolutionary algorithm based on Minkowski Distance for many-objective optimization. *IEEE transactions on cybernetics*, 49, Jul. 2018. doi:10.1109/TCYB.2018.2856208.

Anexos

Anexo A

Resultados de la simulación de pronóstico de ventas con el *dataset* de Pantalones

En esta sección se incluyen los resultados de los experimentos llevados a cabo en el *dataset* de Pantalones, tanto a nivel de tienda (figuras A.1 y A.2), como a nivel de zona (figuras A.3 y A.4).

Estos resultados complementan las observaciones, análisis y conclusiones obtenidas para el *dataset* de Ropa Interior y sirven como sustento para generalizar lo mencionado con respecto a los *datasets* de la categoría vestuario.

A.1. Experimentos realizados para una tienda específica

A.1.1. Resultados obtenidos

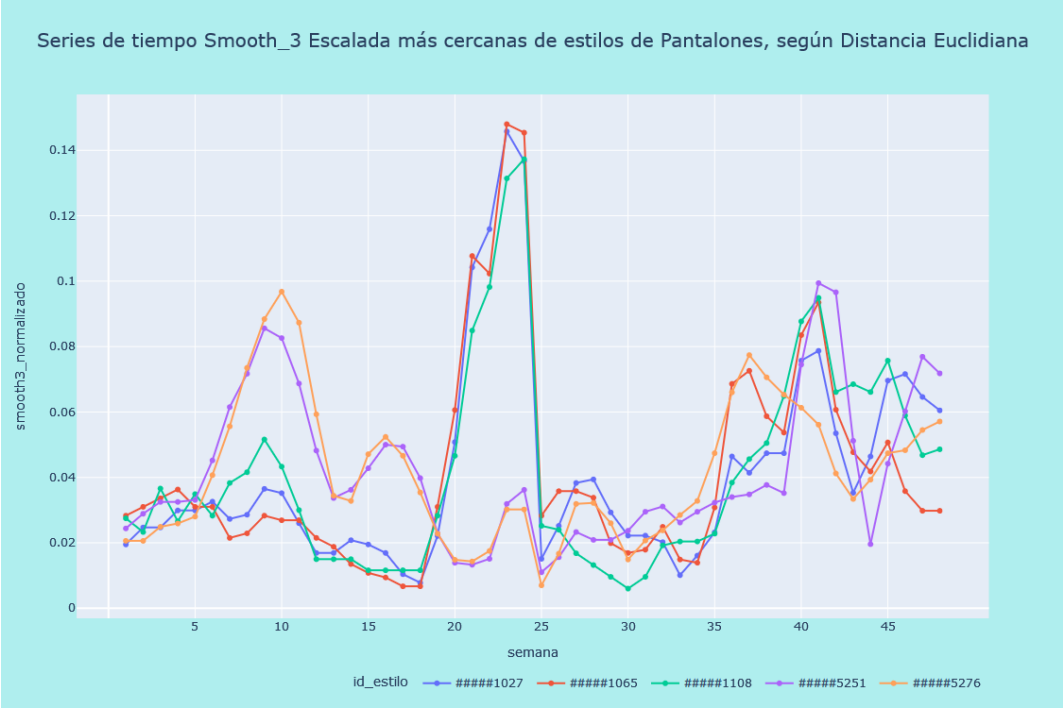


Figura A.1: Pares de series temporales más cercanas en el *dataset* de Pantalones en la tienda Plaza Vespucio.

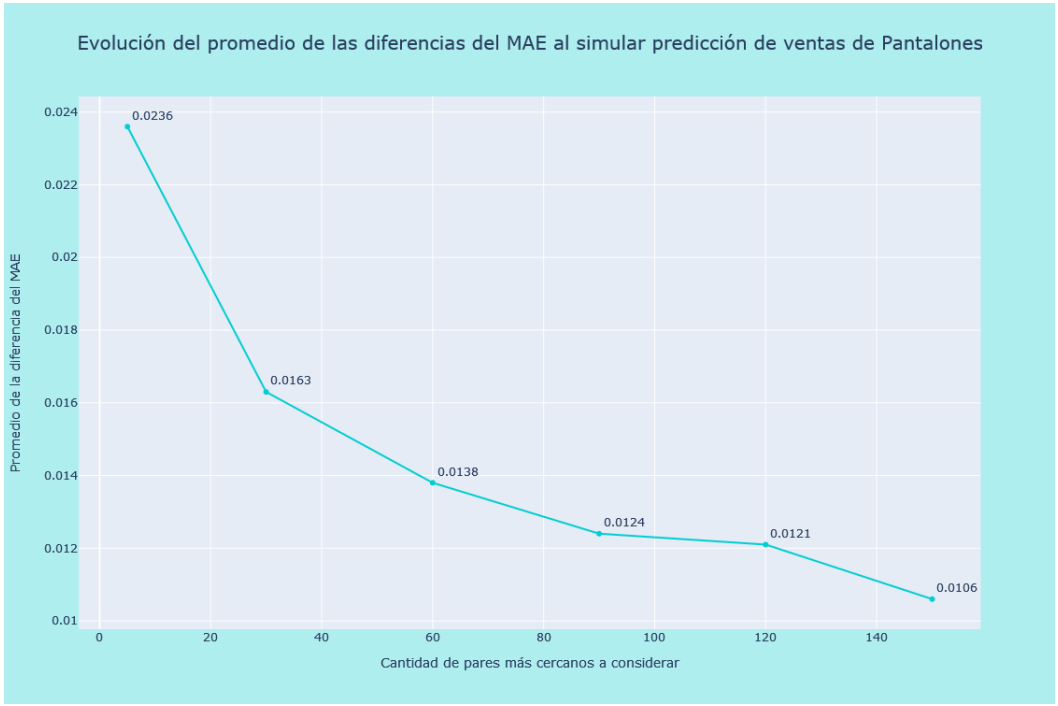


Figura A.2: Evolución de la diferencia del MAE en las predicciones para el *dataset* de Pantalones en la tienda Plaza Vespucio.

A.2. Experimentos realizados para una zona específica

A.2.1. Resultados obtenidos

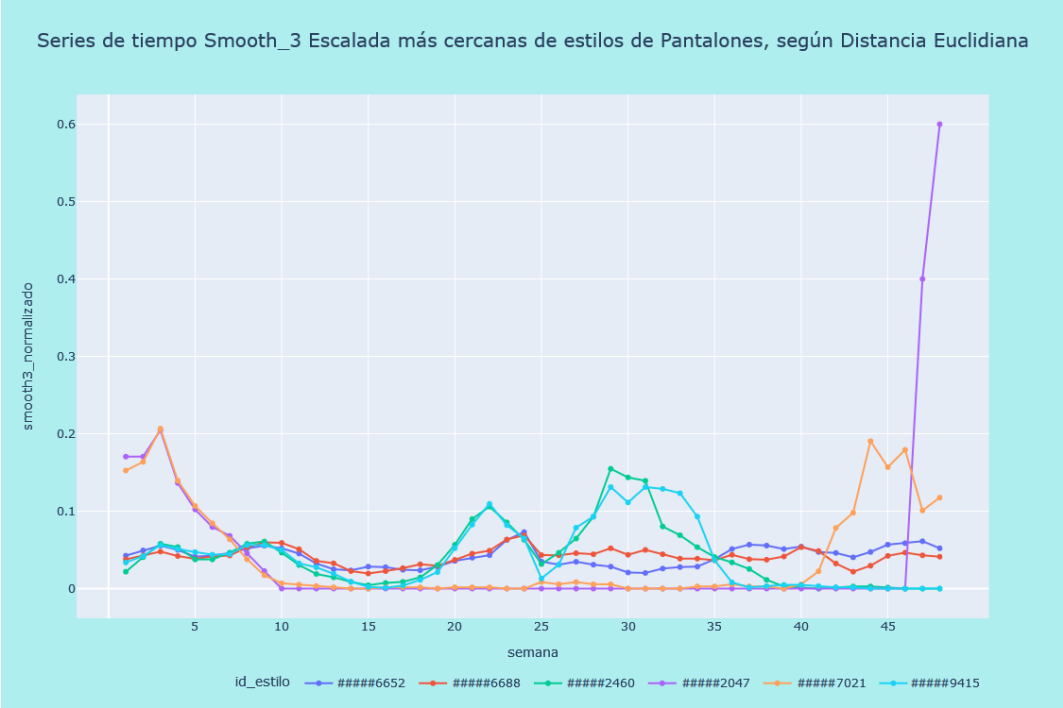


Figura A.3: Pares de series temporales más cercanas en el *dataset* de Pantalones en la zona Centro Poniente.

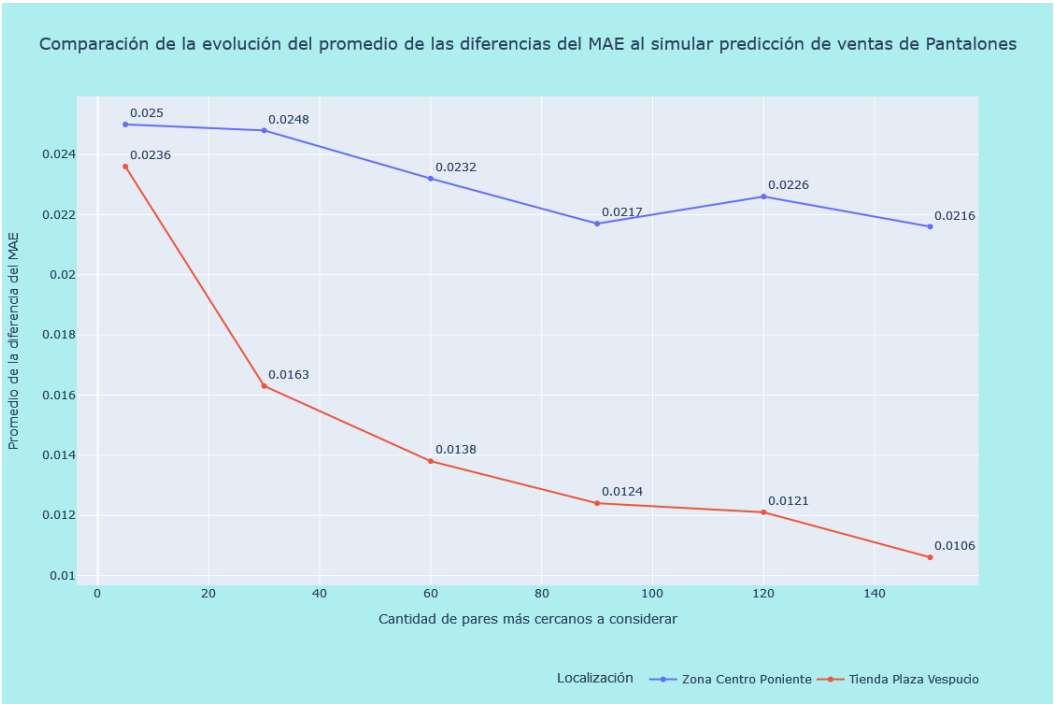


Figura A.4: Evolución de la diferencia del MAE en las predicciones para el *dataset* de Pantalones en la zona Centro Poniente.

Anexo B

Resultados de la simulación de pronóstico de ventas con el *dataset* de Microondas

En esta sección se incluyen los resultados de los experimentos llevados a cabo en el *dataset* de Microondas, tanto a nivel de tienda (figuras B.1 y B.2), como a nivel de zona (figuras B.3 y B.4).

Estos resultados complementan las observaciones, análisis y conclusiones obtenidas para el *dataset* de Lavadoras y sirven como sustento para generalizar lo mencionado con respecto a los *datasets* de la categoría de electrodomésticos.

B.1. Experimentos realizados para una tienda específica

B.1.1. Resultados obtenidos

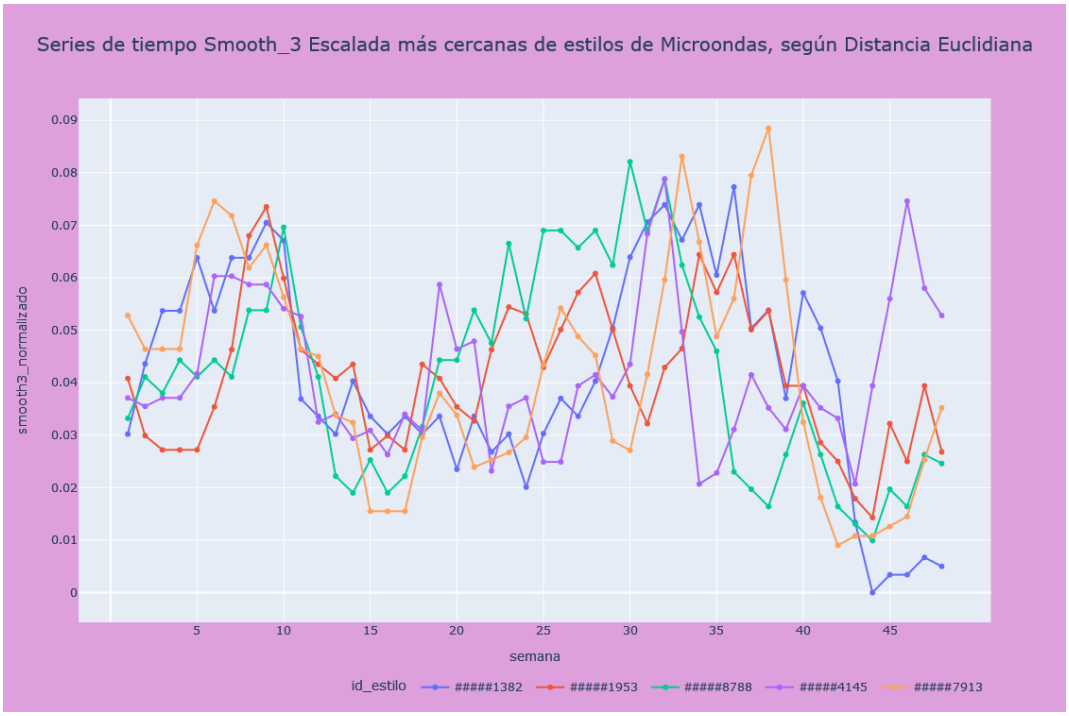


Figura B.1: Pares de series temporales más cercanas en el *dataset* de Microondas en la tienda Plaza Vespucio.

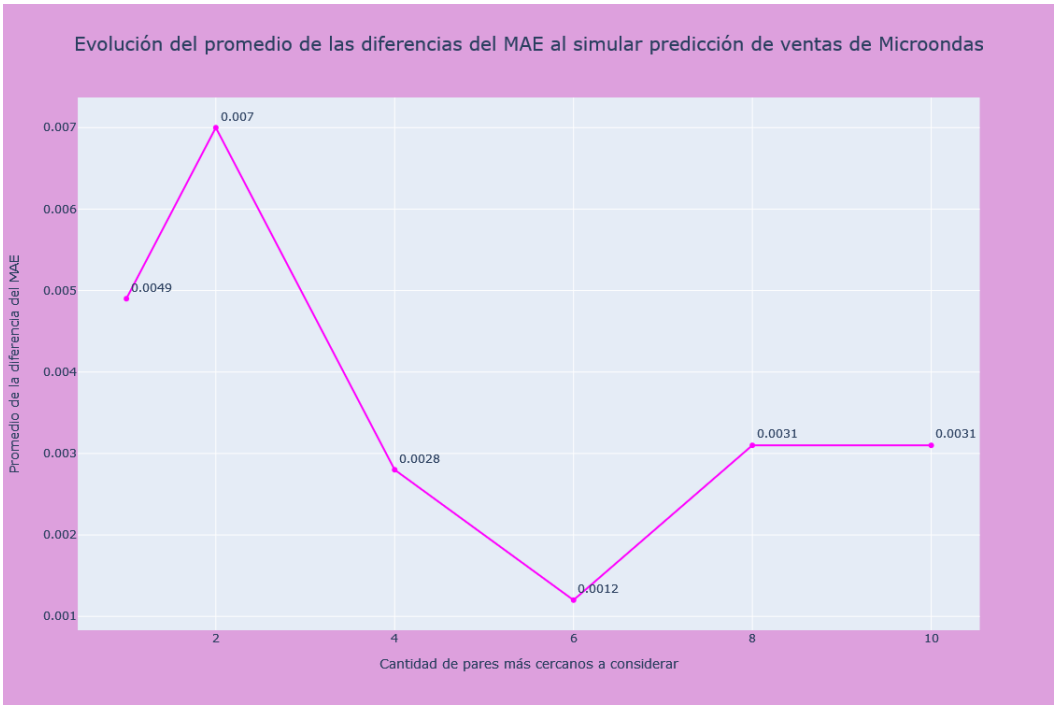


Figura B.2: Evolución de la diferencia del MAE en las predicciones para el *dataset* de Microondas en la tienda Plaza Vespucio.

B.2. Experimentos realizados para una zona específica

B.2.1. Resultados obtenidos

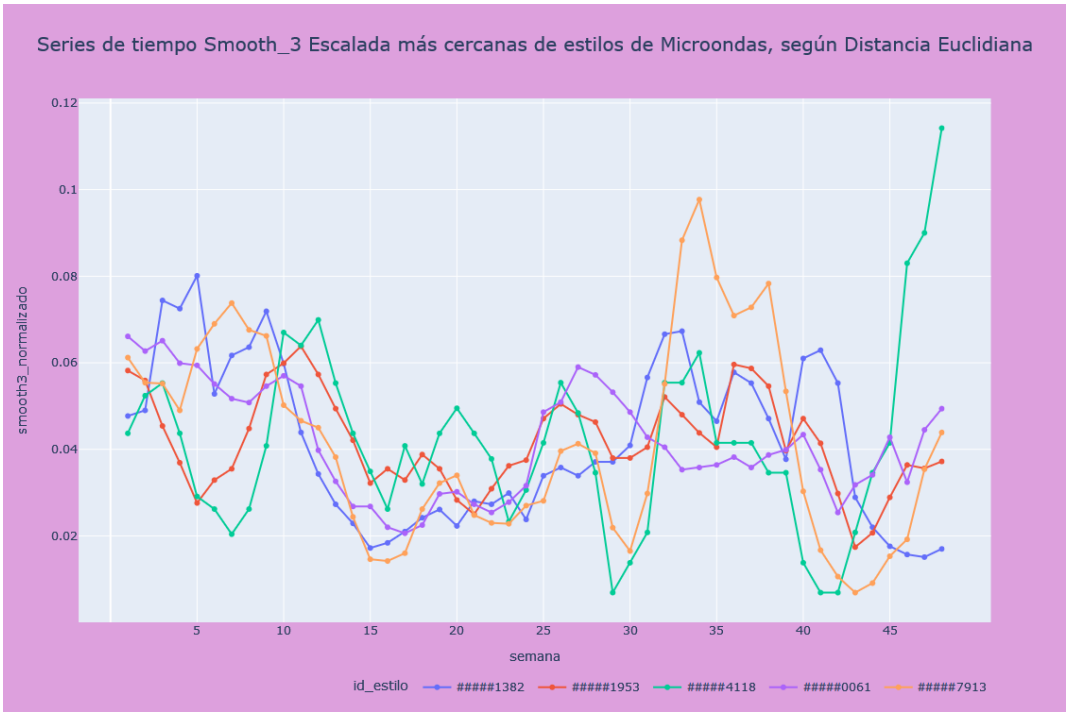


Figura B.3: Pares de series temporales más cercanas en el *dataset* de Microondas en la zona Centro Poniente.

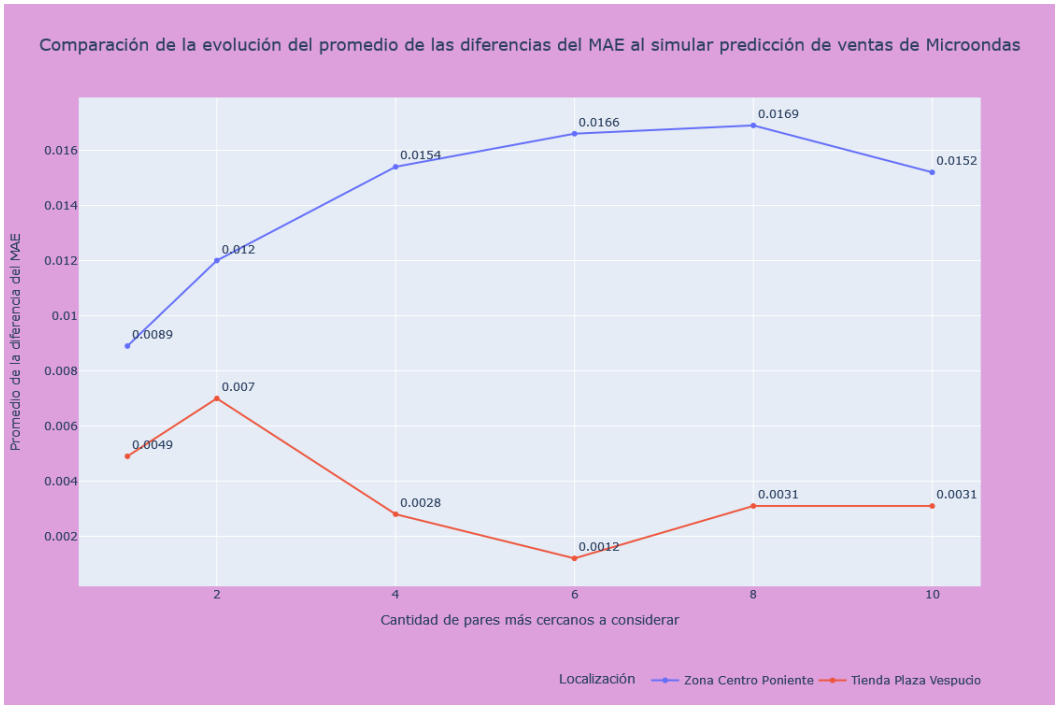


Figura B.4: Evolución de la diferencia del MAE en las predicciones para el *dataset* de Microondas en la zona Centro Poniente.

Anexo C

Nomenclaturas e información relevante

.

C.1. Equivalencia entre semanas comerciales y rangos de fechas

La tabla C.1 se elaboró para facilitar la comparación entre los intervalos de fechas correspondientes al número de semana de cada año, para los años que abarcaron los *datasets*: 2019, 2020 y 2021.

	Año 2019		Año 2020		Año 2021	
Semana	**Fecha inicial**	**Fecha final**	**Fecha inicial**	**Fecha final**	**Fecha inicial**	**Fecha final**
1	2018-12-31	2019-01-06	2019-12-30	2020-01-05	2020-12-28	2021-01-03
2	2019-01-07	2019-01-13	2020-01-06	2020-01-12	2021-01-04	2021-01-10
3	2019-01-14	2019-01-20	2020-01-13	2020-01-19	2021-01-11	2021-01-17
4	2019-01-21	2019-01-27	2020-01-20	2020-01-26	2021-01-18	2021-01-24
5	2019-01-28	2019-02-03	2020-01-27	2020-02-02	2021-01-25	2021-01-31
6	2019-02-04	2019-02-10	2020-02-03	2020-02-09	2021-02-01	2021-02-07
7	2019-02-11	2019-02-17	2020-02-10	2020-02-16	2021-02-08	2021-02-14
8	2019-02-18	2019-02-24	2020-02-17	2020-02-23	2021-02-15	2021-02-21
9	2019-02-25	2019-03-03	2020-02-24	2020-03-01	2021-02-22	2021-02-28
10	2019-03-04	2019-03-10	2020-03-02	2020-03-08	2021-03-01	2021-03-07
11	2019-03-11	2019-03-17	2020-03-09	2020-03-15	2021-03-08	2021-03-14
12	2019-03-18	2019-03-24	2020-03-16	2020-03-22	2021-03-15	2021-03-21
13	2019-03-25	2019-03-31	2020-03-23	2020-03-29	2021-03-22	2021-03-28
14	2019-04-01	2019-04-07	2020-03-30	2020-04-05	2021-03-29	2021-04-04
15	2019-04-08	2019-04-14	2020-04-06	2020-04-12	2021-04-05	2021-04-11
16	2019-04-15	2019-04-21	2020-04-13	2020-04-19	2021-04-12	2021-04-18
17	2019-04-22	2019-04-28	2020-04-20	2020-04-26	2021-04-19	2021-04-25
18	2019-04-29	2019-05-05	2020-04-27	2020-05-03	2021-04-26	2021-05-02
19	2019-05-06	2019-05-12	2020-05-04	2020-05-10	2021-05-03	2021-05-09
20	2019-05-13	2019-05-19	2020-05-11	2020-05-17	2021-05-10	2021-05-16
21	2019-05-20	2019-05-26	2020-05-18	2020-05-24	2021-05-17	2021-05-23
22	2019-05-27	2019-06-02	2020-05-25	2020-05-31	2021-05-24	2021-05-30
23	2019-06-03	2019-06-09	2020-06-01	2020-06-07	2021-05-31	2021-06-06
24	2019-06-10	2019-06-16	2020-06-08	2020-06-14	2021-06-07	2021-06-13
25	2019-06-17	2019-06-23	2020-06-15	2020-06-21	2021-06-14	2021-06-20
26	2019-06-24	2019-06-30	2020-06-22	2020-06-28	2021-06-21	2021-06-27
27	2019-07-01	2019-07-07	2020-06-29	2020-07-05	2021-06-28	2021-07-04
28	2019-07-08	2019-07-14	2020-07-06	2020-07-12	2021-07-05	2021-07-11
29	2019-07-15	2019-07-21	2020-07-13	2020-07-19	2021-07-12	2021-07-18
30	2019-07-22	2019-07-28	2020-07-20	2020-07-26	2021-07-19	2021-07-25
31	2019-07-29	2019-08-04	2020-07-27	2020-08-02	2021-07-26	2021-08-01
32	2019-08-05	2019-08-11	2020-08-03	2020-08-09	2021-08-02	2021-08-08
33	2019-08-12	2019-08-18	2020-08-10	2020-08-16	2021-08-09	2021-08-15
34	2019-08-19	2019-08-25	2020-08-17	2020-08-23	2021-08-16	2021-08-22
35	2019-08-26	2019-09-01	2020-08-24	2020-08-30	2021-08-23	2021-08-29
36	2019-09-02	2019-09-08	2020-08-31	2020-09-06	2021-08-30	2021-09-05
37	2019-09-09	2019-09-15	2020-09-07	2020-09-13	2021-09-06	2021-09-12
38	2019-09-16	2019-09-22	2020-09-14	2020-09-20	2021-09-13	2021-09-19
39	2019-09-23	2019-09-29	2020-09-21	2020-09-27	2021-09-20	2021-09-26
40	2019-09-30	2019-10-06	2020-09-28	2020-10-04	2021-09-27	2021-10-03
41	2019-10-07	2019-10-13	2020-10-05	2020-10-11	2021-10-04	2021-10-10
42	2019-10-14	2019-10-20	2020-10-12	2020-10-18	2021-10-11	2021-10-17
43	2019-10-21	2019-10-27	2020-10-19	2020-10-25	2021-10-18	2021-10-24
44	2019-10-28	2019-11-03	2020-10-26	2020-11-01	2021-10-25	2021-10-31
45	2019-11-04	2019-11-10	2020-11-02	2020-11-08	2021-11-01	2021-11-07
46	2019-11-11	2019-11-17	2020-11-09	2020-11-15	2021-11-08	2021-11-14
47	2019-11-18	2019-11-24	2020-11-16	2020-11-22	2021-11-15	2021-11-21
48	2019-11-25	2019-12-01	2020-11-23	2020-11-29	2021-11-22	2021-11-28
49	2019-12-02	2019-12-08	2020-11-30	2020-12-06	2021-11-29	2021-12-05
50	2019-12-09	2019-12-15	2020-12-07	2020-12-13	2021-12-06	2021-12-12
51	2019-12-16	2019-12-22	2020-12-14	2020-12-20	2021-12-13	2021-12-19
52	2019-12-23	2019-12-29	2020-12-21	2020-12-27	2021-12-20	2021-12-26
53	2019-12-30	2020-01-05	2020-12-28	2021-01-03	2021-12-27	2022-01-02

Tabla C.1: Intervalos de fechas de cada semana de los años 2019, 2020 y 2021.