UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

ANÁLISIS DEL COVID-19 Y SUS CORRELACIONES A NIVEL INTERNACIONAL

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL ELÉCTRICA

TAMARA PAULINA NOVOA RODRÍGUEZ

PROFESOR GUÍA:
AIDAN HOGAN

MIEMBROS DE LA COMISIÓN:
FRANCISCO RIVERA SERRANO
EDUARDO VERA SOBRINO

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL ELÉCTRICA
POR: TAMARA PAULINA NOVOA RODRÍGUEZ
FECHA: 2021
PROF. GUÍA: AIDAN HOGAN

ANÁLISIS DEL COVID-19 Y SUS CORRELACIONES A NIVEL INTERNACIONAL

El COVID-19 (*Coronavirus disease 2019*) es una enfermedad infecciosa causada por el SARS-CoV-2. En diciembre de 2019 fue detectado por primera vez un brote en Wuhan, China y en cuestión de semanas, el virus se expandió a nivel mundial causando la pandemia que hoy se conoce. Debido a que esta es una situación sin precedentes, no hay un escenario claro sobre la relación entre esta enfermedad y otros factores. Por ejemplo, qué enfermedades de base influencian la propagación o si los países más populosos son aquellos más vulnerables frente al virus.

En este trabajo se propuso estudiar las correlaciones estadísticas que pueden tener variables referentes y no referentes al COVID-19. Para ello primero se hizo una investigación sobre las hipótesis que ya estaban planteadas en la Academia; luego se reunieron *datasets* que buscaran responder estas hipótesis, además de otras que se consideraron interesantes de estudiar. Se limpiaron y se pre-procesaron los datos para que fueran legibles para ser modelados como "*RDF Data Cube Vocabulary*" y fueron subidos al servidor Apache Jena Fuseki. Ya con los datos integrados, se calcularon sus correlaciones bajo dos coeficientes: Pearson y Spearman, obteniendo tanto su coeficiente como su valor p. Finalmente, se implementó una plataforma de visualización de estas correlaciones a través de *Flask*, con un *Heat Map* que mostraba lo fuerte que eran las correlaciones según la intensidad del color. Esta interfaz quedó disponible a través de un dominio web, pudiendo ser accesible desde cualquier lugar con internet.

Los resultados más destacados fueron 168 variables modeladas y 250 correlaciones significativas (valor $p \leq 0{,}05$). El top 10 de correlaciones en valor absoluto ordenadas según Spearman y con un coeficiente mayor a 0,5, contiene 5 determinantes de salud: *obesidad*, *cáncer*, *edad*, *tipo de sangre* y *enfermedades respiratorias*. Otras variables del top son *redes sociales* y el *ranking FIFA*, las cuales se cree que están influenciadas por una variable confusa. La única variable relacionada al clima es la *contaminación del aire*, particularmente PM2.5, que puede ser transportador del virus en el aire, favoreciendo una transmisión indirecta. Si bien se encontraron varias correlaciones interesantes, es importante resaltar que estas relaciones no implican necesariamente causalidad.

A mis héroes,
mis padres.

Tamara

Agradecimientos

Al profesor Aidan Hogan por haberme aceptado como memorista sin ser parte de su departamento, por los comentarios constructivos sin importar la hora o el día, por darse el tiempo de reunirse conmigo semana a semana y por ayudarme con las infinitas dudas y problemas que tuve a lo largo de la Memoria.

A mis amigos, por las palabras de aliento, ayuda y consejos independiente de la hora, por las desveladas para poder terminar este trabajo y por todos los maravillosos momentos que hemos compartido.

A Maite y Hiro, quienes son mi dúo favorito.

A mi familia, quienes son el mejor apoyo que pude haber pedido, por todos los sacrificios y esfuerzos a lo largo de estos años, por todo la paciencia, guía, risas y amor que me han brindado.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Los coronavirus son una gran familia de virus que pueden causar enfermedades en animales o humanos. En los humanos, se sabe que varios coronavirus causan infecciones respiratorias que van desde el resfriado común hasta enfermedades más graves, como el síndrome respiratorio del Medio Oriente (MERS) y el síndrome respiratorio agudo severo (SARS). El coronavirus más recientemente descubierto es el COVID-19, nombrado por sus siglas en inglés *Coronavirus disease 2019*.

El COVID-19 es una enfermedad infecciosa causada por SARS-CoV-2 [1]. El virus puede causar síntomas leves como rinorrea, odinofagia, cansancio, tos y fiebre, o síntomas más graves como dificultades respiratorias o neumonía. En algunas ocasiones, la enfermedad puede ser mortal. Las personas mayores y aquellos con problemas médicos subyacentes, como hipertensión, problemas cardíacos y pulmonares, diabetes o cáncer, tienen un mayor riesgo de desarrollar enfermedades graves [2]. El virus se propaga principalmente de persona a persona a través de pequeñas gotas de la nariz o la boca, que se expulsan cuando una persona con COVID-19 tose, estornuda o habla. Las estimaciones actuales apuntan a que el periodo de incubación varía entre 1 y 12,5 días, con una media estimada de 5-6 días. La OMS recomienda que el seguimiento de contactos de casos confirmados sea de 14 días [2].

Este nuevo virus y enfermedad eran desconocidos antes de que comenzara el brote en Wuhan, China, en diciembre de 2019. COVID-19 es ahora una pandemia que afecta a muchos países a nivel mundial. Debido a la situación emergente y la falta de protocolos para combatir el virus, los países han tomado diversas medidas: Japón, con 186 casos confirmados, decidió cerrar escuelas e instituciones académicas en febrero del 2020 hasta el fin del año escolar en abril del mismo año [3]; en Austria, a 13 días del primer contagiado, el ministro del interior Karl Nehammer decretó la prohibición de reuniones de más de 100 personas en recintos cerrados [4] y, el caso de Filipinas, donde el presidente Rodrigo Duterte ordenó "disparar a matar" a aquellos que violen la cuarentena [5].

No solo la respuesta frente al COVID-19 varía en demasía, sino que sus estadísticas también son dispares. Dado que el virus se propaga a través del contacto humano, se podría suponer que los contagios aumentan según la locación, sin embargo, Taiwán, país limítrofe a China, posee una de las menores tasas de contagios a nivel mundial [6, 7, 8].

Este ambiente poco claro y desconocido favorece la desinformación entre la ciudadanía, siendo aún más visible en las redes sociales. El trabajo de Ramez et al. [9] analiza la magnitud de la "desinformación" en Twitter, buscando *trending hashtags* y *keywords* asociados al COVID-19 para luego evaluar la veracidad de la información compartida. Las estadísticas mostraron que el 24,8 % contenía información errónea y un 17,4 % reportaba información no

verificable sobre la epidemia del COVID-19.

Al momento de explicar las variables influyentes en el comportamiento del virus, no existe un consenso a nivel mundial; esto ha generado varias hipótesis con diferentes grados de justificación, pero que generalmente, no están avalados por un trasfondo científico de datos. De este modo, es imperioso estudiar autores que planteen factores de propagación del COVID-19 con un sustento en datos, como es el caso de Lipsitch [10] y Wang et al. [11] que obtienen correlaciones ambientales como temperatura y humedad, la vitamina D que analiza Rhodes et al. [12], o la densidad poblacional vista por Rocklöv y Sjödin [13].

1.1. Identificación y Formulación del Problema

Debido a la contingencia actual, numerosos medios y fuentes han reportado información sobre el COVID-19, sin embargo, encontrar estadísticas sobre el virus y datos propios del país, es una tarea difícil. Es más, si se quiere dar explicación a la diferencia del virus entre las naciones, se necesita hacer un ejercicio manual de cruce de datos relevantes.

¿Qué enfermedades de base causan que una persona sea de riesgo? ¿existe relación entre la propagación del virus y la edad de la persona? ¿hay un efecto indirecto del COVID-19 en el calentamiento global? Contestar estas preguntas no es una tarea trivial, puesto que no existe un método directo que las responda.

1.2. Objetivos del Trabajo de Título

Como propuesta de solución al problema detallado en la sección anterior, se integrarán diferentes bases de datos, con variables referentes y no referentes al COVID-19, a los que se aplicarán métodos y análisis matemáticos con el fin de obtener relaciones. Es en este sentido que, como herramienta útil, se usarán medidas de correlación, ya que estas evalúan de manera cuantitativa la relación entre los datos. Por ejemplo, se puede relacionar el número de muertes en cada país, con variables que no están directamente asociadas al virus, como la humedad, la obesidad, envejecimiento de la población, entre otros.

De esta manera, el objetivo principal corresponde a:

Integrar y generar una base de datos con información por países dividida en dos categorías: datos asociados al COVID-19 y datos de fondo, para luego explorar la calidad de la base generada, buscando correlaciones entre ambas categorías. Finalmente, implementar una plataforma que permita visualizar las variables correlacionadas.

Un desafío importante a mencionar es poder integrar varias fuentes de información y luego, evaluar la calidad de las mismas. Este puede ser abordado a través de herramientas matemáticas para estimar las distribuciones de los datos, con el fin de observar la tendencia que presentan. Otro desafío que se prevé es la temporalidad de los datos, debido a su estructura como series de tiempo, pudiendo cambiar en el tiempo.

Así, se desprenden los siguientes objetivos específicos:

1. Comprender las variables que puedan afectar el fenómeno de la propagación del virus según la literatura.

2. Recopilar fuentes de información que incluyan variables que respondan hipótesis científicas del COVID-19.

3. Integrar una base de datos que incorpore variables relacionadas y no relacionadas al COVID-19.

4. Crear un modelo de correlaciones estadísticas sobre la base de datos integrada.

5. Implementar una plataforma de visualización de correlaciones.

Capítulo 2

Marco Teórico y Estado del Arte

2.1. Hipótesis relacionadas al COVID-19

Uno de los objetivos de este trabajo es "*Comprender las variables que afectan el fenómeno del contagio*", es por esto que la búsqueda de hipótesis que explican la propagación del virus es fundamental. La revisión de las explicaciones del comportamiento del COVID-19 que dan las publicaciones científicas, es un buen punto de partida en las correlaciones esperadas.

Una de las hipótesis encontradas fue la influencia que tienen agentes climáticos, como los ya mencionados por Lipsitch [10] y Wang et al. [11] con los parámetros de temperatura y humedad. En diferentes países alrededor del mundo, se ha analizado el efecto que tienen estas variables en la propagación del virus. El caso más estudiado es China, con Ma et al. [14], Xie y Zhu [15], Shi et al. [16] y Qi et al. [17] que mencionan la influencia de estos factores climáticos en las muertes del país. En Jakarta, Indonesia, Tosepu et al. [18] examinan el promedio de la temperatura como influencia en la transmisión del COVID-19. Briz-Redón y Serrano-Aroca [19] analizan el efecto de la temperatura en España, no mostrando relaciones con la propagación del virus. Se puede ver una relación en el trabajo de Prata et al. [20], entre las temperaturas promedios de Brasil y el número de casos confirmados. En Irán, Jahangiri et al. [21] observan que las variaciones de temperatura tienen una sensibilidad baja ante la transmisión del COVID-19. Otras investigaciones mencionan la influencia de otras variables climáticas como el viento [22], precipitaciones [23] y la calidad del aire [24, 25, 26].

Otra de las hipótesis encontradas es la relación con algunos determinantes en salud como lo son la obesidad y la edad. Grant et al. [27] indican que los niveles más bajos de vitamina D afecta la propagación del virus, además de mencionar que la ingesta de vitamina D puede disminuir el riesgo de infección. No solo parece tener una mitigación en el riesgo de contagio, sino también, parece tener un efecto en la gravedad de la enfermedad COVID-19 [12, 28, 29]. La obesidad también está dentro de las variables relacionadas al COVID-19, debido a los riesgos desarrollados al momento de enfrentar la enfermedad [30, 31, 32].

Este estudio contempla no sólo las hipótesis publicadas en revistas y conferencia científicas, sino también hipótesis más informales, por ejemplo, hipótesis compartidas en redes sociales, u otros medios. Para reducir el sesgo de la búsqueda, se publicó una encuesta en la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. En ella se le preguntaba a los usuarios de U-Cursos sobre posibles hipótesis relacionadas al COVID-19 y su respectivo respaldo a través de una publicación científica. De este modo, se obtuvieron 23 respuestas que en conjunto con la búsqueda recién descrita, resultaron en 33 hipótesis disponibles en el anexo A.

2.2. Datos crudos disponibles

Es importante destacar las fuentes de datos crudos que se usaron en esta memoria, ya que se quiere ver las relaciones entre diferentes indicadores y el desarrollo del virus por país. Se dividieron en dos categorías: los Referentes al COVID-19 y los No referentes al COVID-19.

2.2.1. Referentes al COVID-19

Los datos disponibles sobre el COVID-19 son, en su mayoría, series de tiempo de la cantidad de casos confirmados, muertos y recuperados diarios y/o acumulados, también, índices como la rigurosidad de respuesta del gobierno. Algunas de las fuentes de datos libres sobre el COVID-19 son:

1. "WHO Coronavirus Disease (COVID-19)" por World Health Organization [33].
2. "Excess mortality during the Coronavirus pandemic (COVID-19)" por Our World in Data [34].
3. "Coronavirus Pandemic (COVID-19) – the data" por Our World in Data [35].
4. "COVID-19 Data Repository" por Center for Systems Science and Engineering (CSSE) de la Johns Hopkins University [36].
5. "COVID-19 situation update worldwide" por European Centre for Disease Prevention and Control [37].
6. "Excess Deaths Associated with COVID-19" por Centers for Disease Control and Prevention (CDC) [38].
7. "Base de Datos COVID-19" por el Ministerio de Ciencia, Tecnología, Conocimiento e Innovación Chileno [39].

2.2.2. No referentes al COVID-19

Además de los datos propios del COVID-19, se buscaron bases de datos que otorgan información sobre los países en diferentes ámbitos, tales como indicadores económicos, políticos, de salud, entre otros.

1. "World Bank Open Data" por The World Bank [40].
2. "Our World in Data" [41].
3. "Wikipedia" [42].
4. "The World Factbook" [43].
5. "European mortality database" por European Health Information Gateway [37].
6. "The Global Health Observatory" por World Health Organization [44].
7. "Country Product Complexity Rankings" por Atlas of Economic Complexity del Growth Lab de Harvard University [45].

8. “FIBA World Ranking” [46].
9. “FIFA Ranking” [47].
10. “FIVB Senior World Ranking” [48].
11. “UNdata” [49].
12. “Dataset Publishing Language” de Google Developers [50].

2.3. Web Semántica

Uno de los desafíos centrales en este trabajo es poder integrar datos diversos e incompletos desde varias fuentes. Aunque las bases de datos relacionales constituyen la solución más popular para manejar datos, su modelo de datos, basado en tablas, no está bien adaptado para trabajar con datos incompletos, esquemas complejos, ni para integrar datos de varias fuentes.

Para esta memoria, se eligió usar un modelo de datos alternativo basado en grafos, que provee una estructura más flexible para poder integrar datos incompletos de varias fuentes. Usando el mismo modelo para representar todos los datos permite extraer la información que se necesita para medir alguna correlación en particular usando un lenguaje de consulta declarativa. En particular, el modelo de datos que se ha elegido ha sido propuesto en el contexto de la Web Semántica para representar datos diversos en la Web.

El término «*Web Semántica*» hace referencia a la visión de W3C (*World Wide Web Consortium*) sobre la *Web of Linked Data*, que prevee programas de *software* con metadatos legibles por máquina de los datos publicados en la red; en este sentido, se espera que los computadores sean capaces de hacer interpretaciones significativas similares a la forma en que los humanos procesan la información y que desarrollen sistemas que puedan soportar transacciones confiables a través de la red. La Web Semántica permite a los usuarios crear *data stores*, vocabularios y reglas para manejar los datos en la Web. Por ejemplo, una ontología puede describir conceptos, relaciones entre entidades y categorías de cosas. Para lograr esto, *Linked Data* está potenciado por tecnologías especiales para representar metadatos como *Resource Description Framework* (RDF), SPARQL, *Web Ontology Language* (OWL) y SKOS.

2.3.1. RDF

Resource Description Framework (RDF) es un modelo estándar de intercambio de datos y otras informaciones estructuradas en la Web Semántica. Fue uno de los primeros hitos de esta, siendo recomendada por primera vez por W3C en febrero de 1999. RDF extiende la estructura de enlaces de la Web usando URIs, literales y nodos blancos para nombrar la relación entre objetos con expresiones «*sujeto-predicado-objeto*» (referido usualmente como “tripleta”) [51]. Este tipo de estructura forma un grafo etiquetado en que el enlace entre dos nodos es la unión de ambos recursos, convirtiéndose en una explicación visual fácil de entender.

A modo de ejemplo, la Figura 2.2 representa un grafo de RDF que describe las relaciones indicadas en la 2.1, usando un vocabulario más formal.

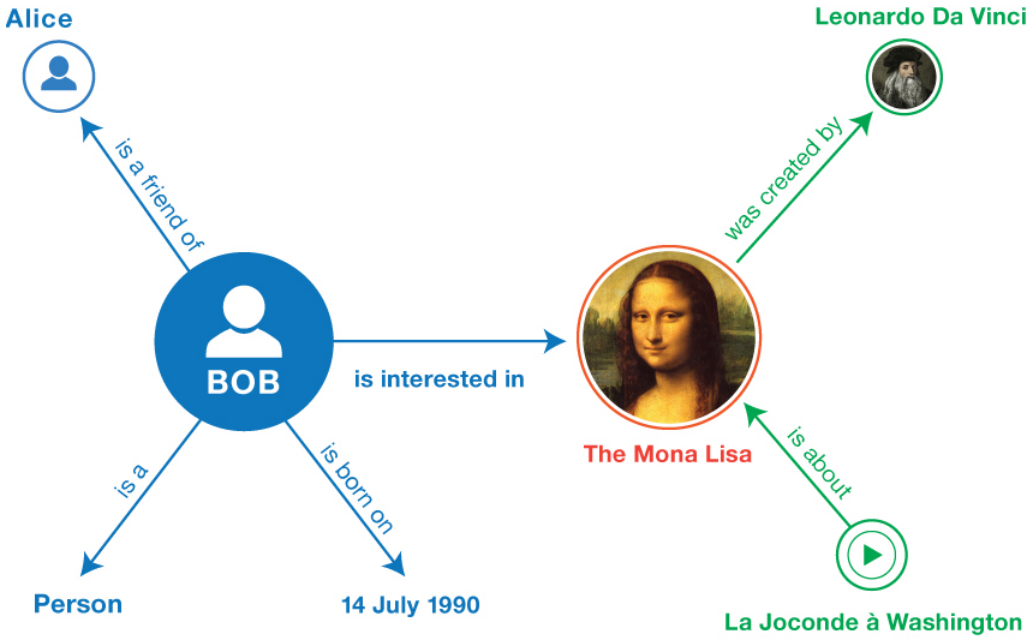


Figura 2.1: Grafo informal de una muestra de tripletas [52].

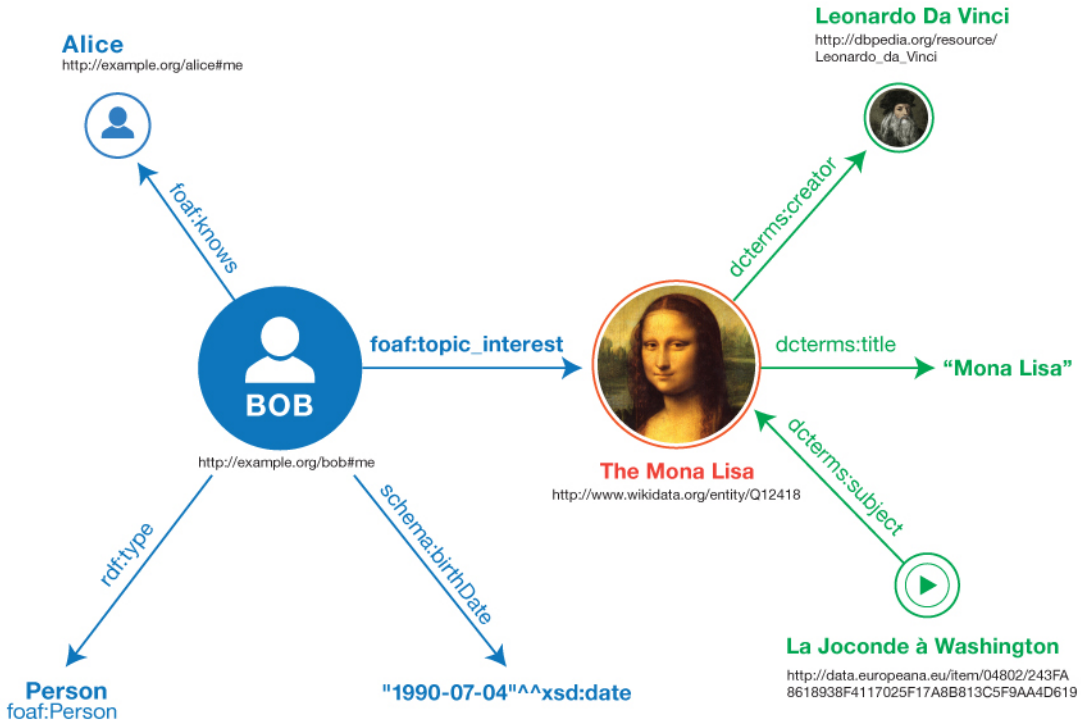


Figura 2.2: Grafo RDF resultante del ejemplo de N-Tripletas [52].

2.3.2. RDF Data Cube

Data Cube Vocabulary nace de la necesidad de agregar un contexto estadístico a través de *datasets*, un conjunto de observaciones organizadas a través de diferentes dimensiones. Este vocabulario permite representar estos datos multi-dimensionales en la Web a través de RDF y publicar la información según los principios de *Linked Data*.

El espacio donde residen los *datasets* es llamado *cubo*; éste está formado por *componentes* que son un conjunto de *dimensiones*, *atributos* y *medidas*. Una *dimensión* es aquella que identifica a la observación, como por ejemplo la ubicación geográfica o el período al que corresponde. Por otra parte, los *atributos* tienen el objetivo de posibilitar la interpretación cualitativa de las observaciones como es el número de decimales o la unidad multiplicadora. Finalmente, las *medidas* representan el fenómeno observado.

La idea de poder utilizar este tipo de vocabulario es que permite a los usuarios acceder de forma fácil y rápida a las referencias que unen a los *datasets*, llevándose a cabo a través de los URI enlazados a las entidades o conceptos. RDF provee de un estándar para la representación de la información que describe entidades y conceptos, para retornar el URI enlazado [53].

Los datos contenidos en un *dataset* pueden pertenecer a uno de los siguientes grupos:

1. **Observaciones:** se refiere a los valores medidos, por ejemplo, en un tabla estadística corresponden a los valores de cada celda.

2. **Estructura organizacional:** para ubicarse en el cubo, es necesario conocer los valores de las dimensiones sobre las cuales se está haciendo referencia, teniendo que estar explícitamente para cada observación. Otro tipo de estructuras organizacionales son las *slices*, que representan selecciones de regiones del cubo basado en restringir una o más dimensiones a un valor o varios valores, mas no serán tratadas en profundidad debido a su irrelevancia dentro de este trabajo.

3. **Metadatos estructurales:** ya ubicados los datos, se necesita de ciertos metadatos que den una interpretación a las observaciones, tales como unidades de medidas, multiplicadores de unidades, entre otros. Los metadatos son considerados como "atributos" y pueden estar adjuntos tanto a nivel de observación como a niveles superiores.

4. **Metadatos de referencia:** son los metadatos que ayudan a describir el *dataset* como un todo. Aquí se incluyen datos como el autor, fuente, caracterización del *dataset* y un *SPARQL endpoint* al que se puede acceder.

Todos los conceptos recién descritos están representados en el gráfico de la Figura 2.3.

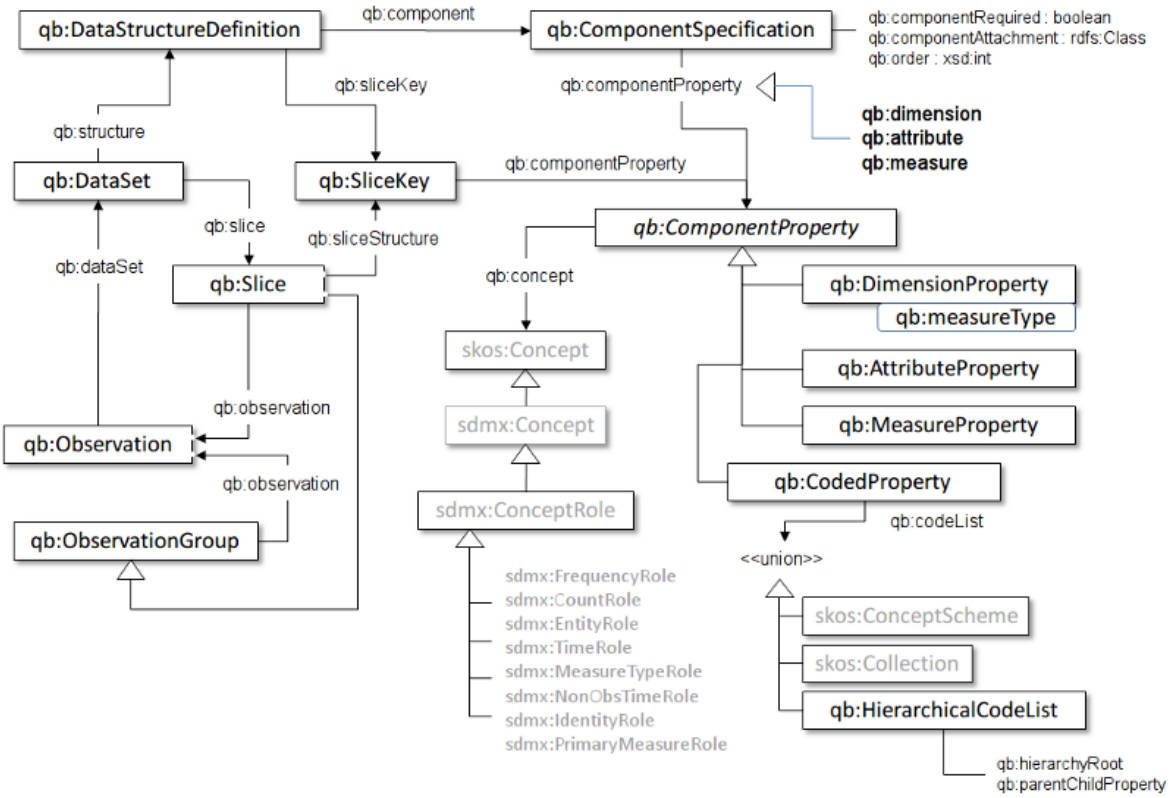


Figura 2.3: Resumen ilustrado de términos clave y su relación [53].

2.3.3. Wikidata

Wikidata es un repositorio central de almacenamiento, libre y de código abierto, que puede ser leído y editado tanto por humanos como máquinas. Su fin es tener una base común que puede ser accedida desde cualquier *wiki* de la Fundación Wikimedia [54], como son los proyectos de Wikipedia, Wikivoyage, Wiktionary, Wikisource y de cualquier persona siempre que se use una licencia de dominio público. Puede ser exportada en formatos estándar y ser enlazada a cualquier *dataset* abierta en la Web.

Opera con el software de Wikibase [55], consistente, mayoritariamente, en elementos donde cada uno contiene etiquetas, descripciones y alias para cada lenguaje enlazado. Debido a su característica de base de datos es que busca ser insesgada al idioma, conteniendo un identificador único para cada elemento, que parte con el prefijo Q y un sufijo numérico único; por ejemplo, Q42 es el elemento "Douglas Adams" detallado en la Figura 2.4. En ésta se muestran las *declaraciones*, información que se agrega a los elementos para describir características en forma de *propiedad* y *valor*. En el ejemplo se agregó la declaración de *dónde estudió* Douglas Adams, con la propiedad "alma mater" (P69) y dos valores: St John's College (Q691283) y Brentwood School (Q4961791); también se pueden incluir *declaraciones* como el lugar/fecha de nacimiento/muerte, padres, ocupación, trabajos destacados, entre otros.

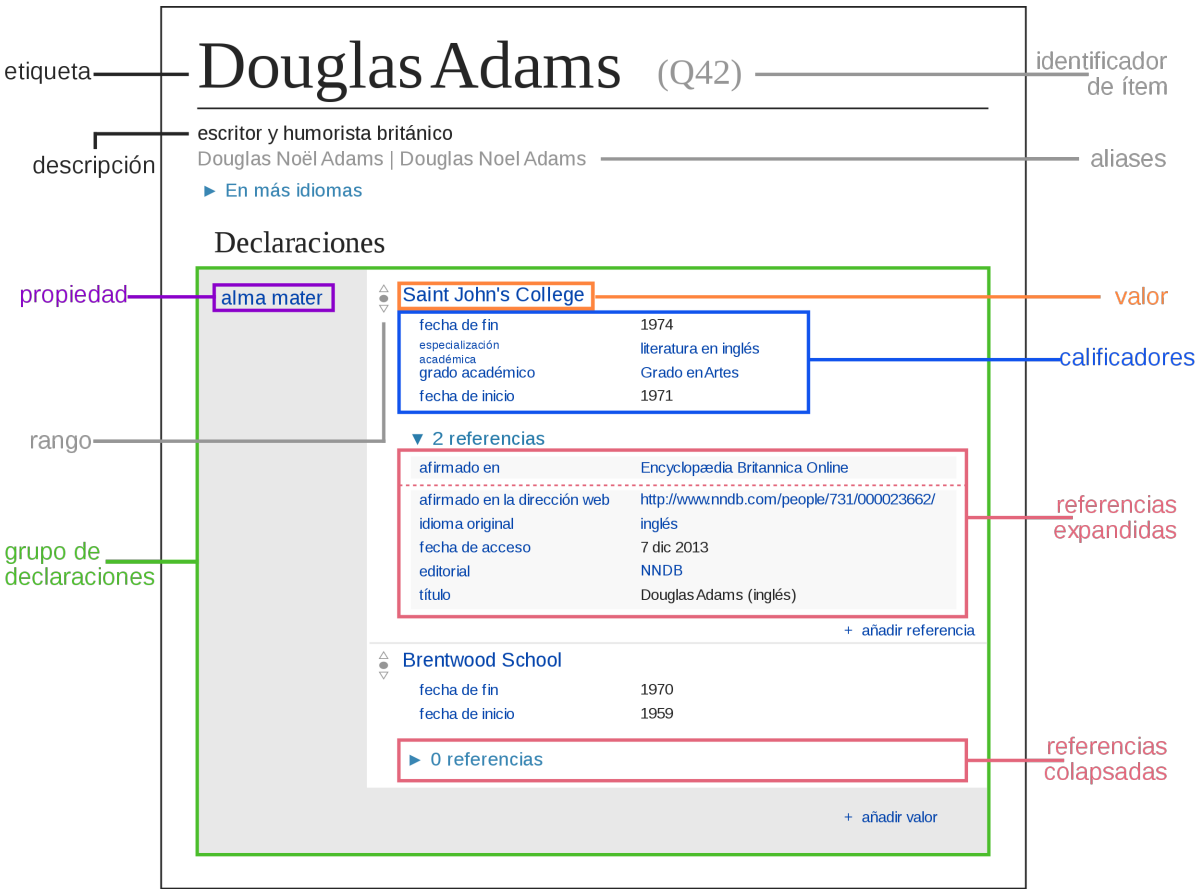


Figura 2.4: Ejemplo de elemento de Wikidata [56]

2.3.4. Turtle

El lenguaje Turtle [57] (del inglés *Terse RDF Triple Language*) es una sintaxis concreta y formato de archivos definida en RDF Concepts and Abstract Syntax [58], además de ser una extensión de N-Triples [59]. Así, Turtle permite escribir un grafo RDF de forma compacta a través de texto.

A modo de ejemplo, se muestra una parte de un archivo sobre la extracción de la demografía de Islandia (Código 2.1). Los datos corresponden a una observación del *dataset* de poblaciones, incluye, además, información como el período de la toma de la muestra y cantidad de habitantes; cabe destacar que Islandia tiene el código Q189 [60] en *Wikidata* [56].

Código 2.1: Ejemplo de archivo *.ttl* sobre la población de Islandia [40]

```
c19:population-o086
    rdf:type                        qb:Observation ;
    qb:dataSet                      c19:dataset-population ;
    c19-dimension:refPeriod         c19-interval:Year2019 ;
    c19-measure:population          0.4 ;
    sdmx-attribute:unitMultiplier   6 ;
    sdmx-attribute:decimals         1 ;
    c19-dimension:refArea           wd:Q189 .
```

2.3.5. SPARQL

SPARQL Protocol and RDF Query Language (SPARQL) es un lenguaje de consultas y protocolo estándar para RDF, donde su sintaxis y semántica viene definida por la estructura de grafos. SPARQL permite a los usuarios enfocarse en el *qué* quieren consultar, en vez del *cómo* se debería evaluar la consulta. Las consultas incluyen agregación, subconsultas, negaciones, crear valores por expresiones, *value testing* y consultas de restricción. Además, SPARQL contiene capacidades para consultar patrones de grafos obligatorios y opcionales junto con sus conjunciones y disyunciones [61].

La mayoría de las consultas de SPARQL contiene un set de patrones de tripletas, similares a las tripletas de RDF, pero donde el sujeto, predicado u objeto pueden ser las variables. Una consulta básica consiste en dos partes: «`SELECT`», una cláusula que identifica las variables a aparecer en los resultados y «`WHERE`», cláusula que provee al patrón de grafo básico para que coincida con el grafo de los datos.

Para obtener los habitantes del Código 2.1 se utiliza la consulta detallada en el Código 2.2; esta consiste en la declaración de los prefijos, un *select* que proyecta la población (?population) en la posición del objeto y un *where* que compara el patrón básico, seleccionando la medida de la observación o086.

Código 2.2: Ejemplo de *query SPARQL* sobre la extracción de la población de Islandia (ver Código 2.1).

```
PREFIX c19:  <http://example.org/ns#>
PREFIX c19-measure:  <http://example.org/measure#>
SELECT ?population
WHERE {
   c19:obs-o086 c19-measure:population ?population .
}
```

El resultado puede ser retornado en formato tabla o datos crudos en formato *json*. Para el ejemplo del Código 2.2 su *output* como tabla es:

population
"357050.0"^^xsd:decimal

2.4. Coeficientes de Correlación

Para comprobar las hipótesis planteadas con los datos disponibles, se hace uso de las herramientas matemáticas: el coeficiente de correlación de Pearson y el Coeficiente de Spearman. Si bien existen más coeficientes de correlación, en este trabajo solo se hará uso de los dos ya mencionados. También, en esta sección se hablará, sobre las restricciones y consideraciones importantes de utilizar estos coeficientes.

En estadística, un coeficiente de correlación es una evaluación cuantitativa que mide, tanto la dirección como la fuerza y la dirección a variar juntas, es decir, la relación entre dos conjuntos de datos. Los coeficientes poseen un rango entre $-1{,}0$ y $1{,}0$. Los valores cercanos a $1{,}0$ indican un fuerte acuerdo, mientras que los valores cercanos a $-1{,}0$ indican un fuerte desacuerdo. Una correlación de $0{,}0$ muestra que no hay relación entre las dos variables.

Para diferenciar entre coeficientes similares se utiliza el valor p (*p-value* en inglés). El valor p se define como la probabilidad de obtener resultados dado que la hipótesis nula (H_0) es cierta [62]. En este caso, los resultados son producto del azar del muestreo y no son significantes en términos de soportar la correlación investigada. Se ha establecido por convención que valores p asociados a un nivel de significación α iguales o menores a 0,05 pueden rechazar H_0. De esta manera, este valor ayuda a determinar la significancia de los resultados, en relación a la hipótesis nula.

En este trabajo se utilizarán dos tipos de coeficientes de correlación: coeficiente de Pearson (r) y el coeficiente de Spearman (ρ). Estos pueden ser implementados a través de la librería *SciPy* de *Python*, particularmente *Scipy Stats*.

2.4.1. Coeficiente de Pearson

El coeficiente de correlación de Pearson mide la relación lineal entre dos variables aleatorias cuantitativas y es independiente de la escala de medida de las variables. Las correlaciones positivas implican que si una variable incrementa, la otra también lo hará, de igual modo, si corresponde a una correlación negativa, ambas variables tienden a disminuir.

El valor r es calculado según la Fórmula 2.1, donde m_x es el promedio del vector x y m_y es el promedio del vector y, además de asumir que ambos vectores tienen distribuciones normales independientes.

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2(y - m_y)^2}} \tag{2.1}$$

El valor p está calculado bajo la suposición de que cada variable está distribuida normalmente; está computado como la probabilidad que el valor absoluto de r de una muestra al azar de x e y, extraído de la población con correlación cero, será más grande o igual al valor absoluto r [63].

2.4.2. Coeficiente de Spearman

El coeficiente de correlación de orden de rango de Spearman, es una medida no paramétrica de la monotonicidad de la relación entre dos conjuntos de datos. A diferencia de la correlación de Pearson, la correlación de Spearman no supone que ambos conjuntos de datos se distribuyan normalmente.

Así, el coeficiente de Spearman está definido como el coeficiente de Pearson entre rankings de variables, transformando las variables X_i, Y_i a los rankings rg_{Xi}, rg_{Yi}. La Fórmula 2.2 muestra el cómputo de ρ, donde $cov(rg_X, rg_Y)$ es la covarianza y σ_{rg_X}, σ_{rg_Y} las desviaciones estándar de los rankings de variables.

$$\rho = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{2.2}$$

Si los datos corresponden a enteros, se puede computar el coeficiente de acuerdo a la siguiente fórmula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{2.3}$$

En la Fórmula 2.3, d es la diferencia entre los correspondientes estadísticos de orden $x - y$ y n es el número de parejas de datos [64].

2.4.3. Limitaciones

Existen muchas consideraciones importantes al momento de trabajar con correlaciones, la principal es que una correlación no implica una causalidad, conocida en latín como *cum hoc ergo propter hoc* (con esto, por tanto a causa de esto) [65]. Un ejemplo de esta falacia pueden ser variables confusas, como lo son el número de universidades de un país y su número de contagios por COVID-19. Se puede observar en los rankings mundiales que las naciones que poseen un mayor número de infectados (ver Tabla 2.1), son también aquellas con gran cantidad de universidades; sin embargo, esta correlación se ve afectada por la variable confusa de la población.

Tabla 2.1: Ejemplo de variables confusas, extracción de datos del número de infectados de COVID-19 [66], el número de universidades del mundo [67] y su población en millones respectiva [68].

País	**Nº contagios COVID-19**	**Nº universidades**	**Población [millones]**
Estados Unidos	30.705.525	3.254	328,2
Brasil	12.984.956	1.349	211,0
India	12.485.509	4.381	1.366,4

Otro punto relevante a mencionar es la gran cantidad de correlaciones que puede analizar un sistema, en particular, hay que tener especial cuidado con los valores p. Si se considera un valor p de 0,05 para tener resultados significativos, este 0,05 implica que la probabilidad de rechazar la hipótesis nula es de 1/20; así, al momento de tomar una muestra aleatoria de correlaciones para 20 pares de variables, es probable que se vaya a encontrar una correlación significativa aunque los datos sean aleatorios.

Si bien las limitaciones existentes respecto a las correlaciones son claves al momento de realizar un análisis estadístico, especialmente la interpretación de los resultados, igualmente pueden informar sobre el comportamiento del COVID-19. Por ejemplo, de encontrarse una correlación fuerte entre dos variables, sin que contenga variables confusas u otra explicación obvia, puede incitar a generar una nueva hipótesis y conllevar a una investigación más profunda; o por el contrario, si no existe la correlación esperada, se puede rechazar la hipótesis existente y sugerir que no existe dicha relación causal.

2.5. Sistemas similares

Se buscaron plataformas que intenten responder a la problemática planteada, sin embargo, no se encontraron aplicaciones que permitan buscar correlaciones a nivel de país con respecto al COVID-19, en un contexto económico, político, geográfico, social, entre otros.

Las aplicaciones encontradas corresponden a visualizaciones de los datos del COVID-19, como el número de casos confirmados, casos de muertos, casos de recuperados, entre otros. En la mayoría de estas plataformas es posible filtrar la información por locación geográfica, como un país o región. Ejemplos de plataformas internacionales son *"IBM Global COVID-19 Statistics"* [69] (Figura 2.5), *"COVID-19 Dashboard by Johns Hopkins University"* [66] (Figura 2.6), ‘*'Coronavirus Pandemic (COVID-19)"* [70] (Figura 2.7) y, como ejemplo nacional, *"COVID-19 en tu comuna"* [71] (Figura 2.8).

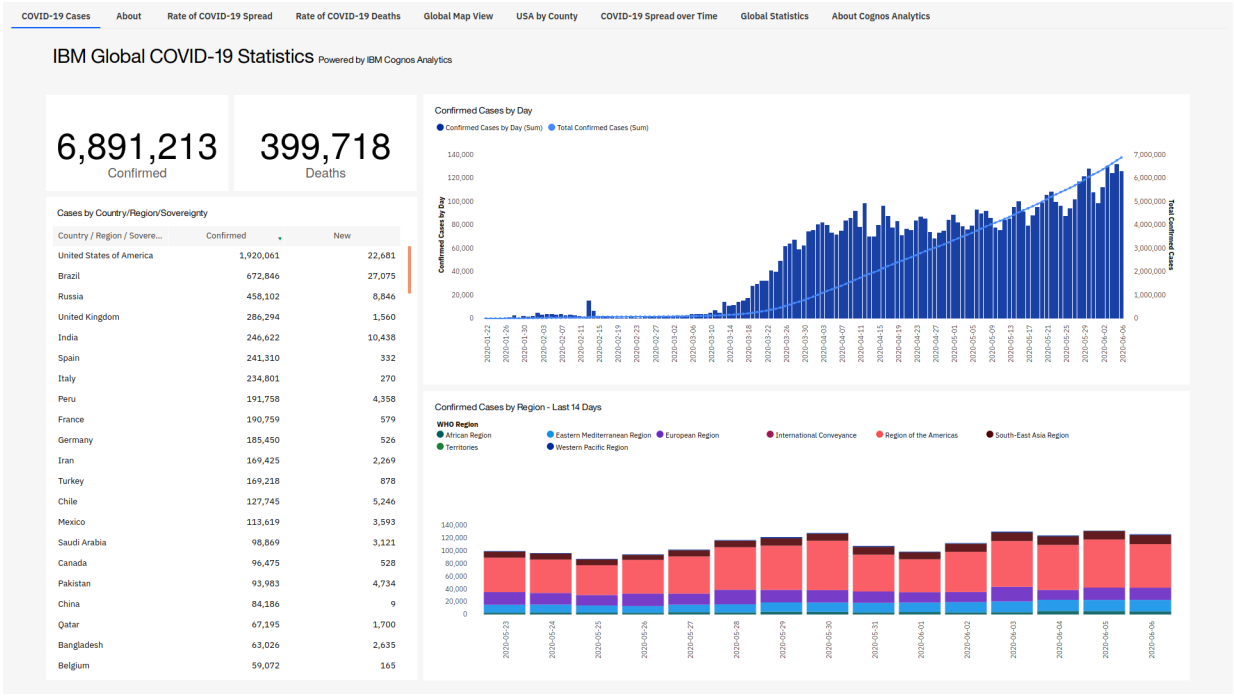


Figura 2.5: IBM Global COVID-19 Statistics.

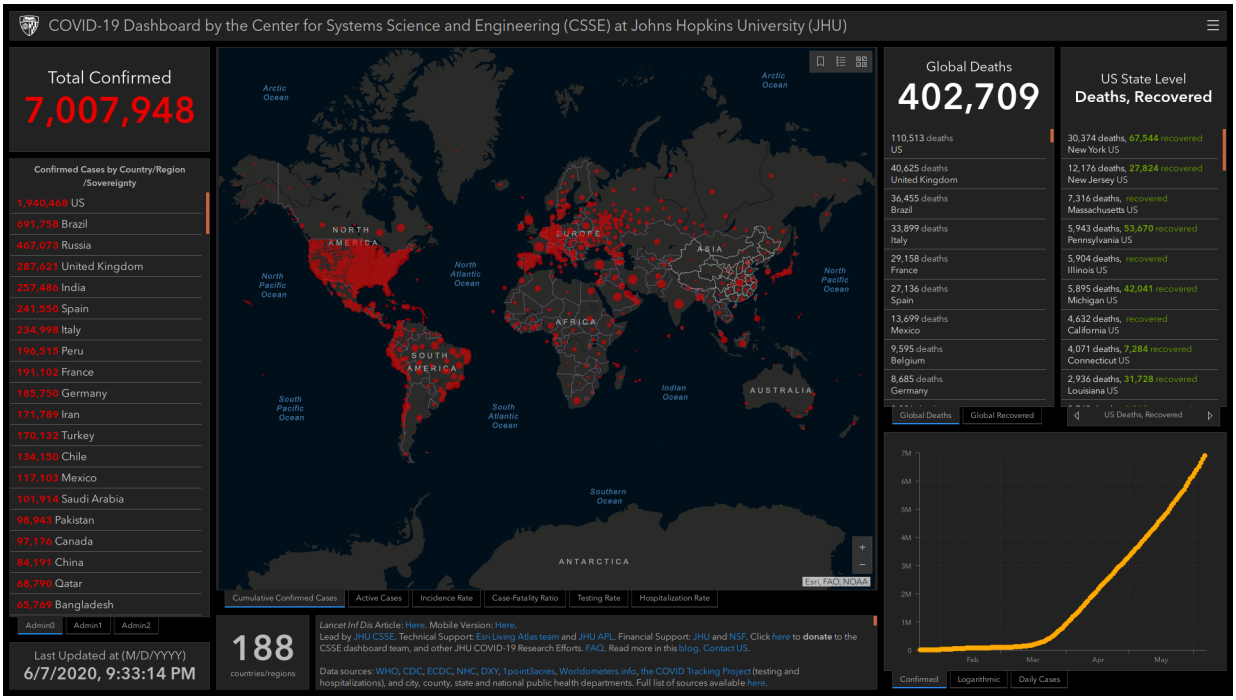


Figura 2.6: COVID-19 Dashboard by CSSE at Johns Hopkins University (JHU).

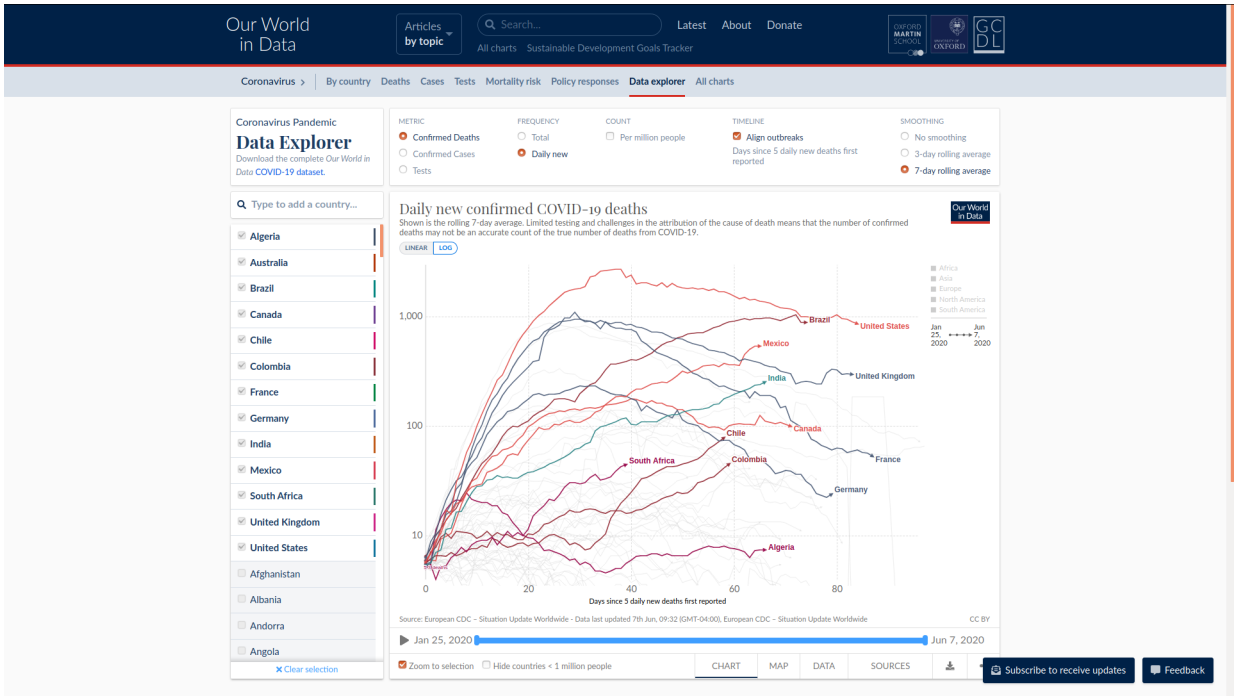


Figura 2.7: Coronavirus Pandemic (COVID-19).

Figura 2.8: COVID-19 en tu comuna.

Capítulo 3

Manejo de los Datos

3.1. Recolección de datos

En un principio, la recolección de datos se centraría solamente sobre aquellas variables enfocadas en responder las hipótesis planteadas (ver Anexo A), sin embargo, a medida que avanzaba la búsqueda, se optó por una colección más extensa, dado que en particular, algunas hipótesis tenían diversas variables asociadas. Así se logró armar un *database* con 525 variables, de las cuales, 426 variables (de diferentes ámbitos) provenían de Our World in Data [70]. No obstante, estas no eran suficientes para cumplir con el objetivo de "*Recopilar fuentes de información que incluyan variables que respondan hipótesis científicas del COVID-19*". De esta manera, se buscaron fuentes de información que completaran y complementaran los datos faltantes como Wikipedia [42], The World Bank [40], World Health Organization [44], entre otros. Con respecto a información relativa al COVID-19, se recopilaron 14 *datasets* con reportes diarios de los casos asociados al virus, excesos de muertes, casos acumulados e índices de respuesta; estos contienen información desde el inicio de la pandemia hasta el 27 de octubre.

Ya seleccionados los datos, se guardaron como formato *.tsv* (*tab-separated values*) para evitar confusiones de lectura. Adicionalmente, en aquellas bases de datos que contenían más de una variable, se almacenaron de forma separada con el fin de optimizar el modelamiento. Finalmente, los archivos se respaldaron en *GitHub*.

Al tener datos de fuentes de información tan variadas, se debe considerar que toda la muestra tenga el mismo tipo de número, es decir, enteros, racionales o reales, de lo contrario no se pueden procesar. En este caso se utilizó Python, donde se armó un *script* que codificaba los números como *floats*; por ejemplo 123.345,67 se convirtió en 123345.67, también se transformaron los valores vacíos o nulos como valores aptos `NaNs` para poder ser leídos por la máquina.

A continuación se muestra un ejemplo con 6 territorios de la base de datos "Number of deaths by risk factor: Smoking" de Our World in Data [70] que muestra el número total de muertes al año por factor de riesgo de fumadores, medido en todos los grupos etarios y ambos sexos. Los datos crudos se encuentran en Tabla 3.1, en ella destacan que se utiliza la coma como separador de millares y la utilización de la palabra "*million*" para representar el multiplicador 10^6.

Tabla 3.1: Datos crudos extraídos de la base de datos *Number of deaths by risk factor: Smoking* de Our World in Data [70].

Territory	**Number of associated deaths**
Central Latin America	102,12
Central Sub-Saharan Africa	30,294
Chad	1,658
Chile	10,405
China	2.20 million
Colombia	19,434

Siguiendo con la unificación de datos, se cambiaron los nombres de los territorios a un identificador único, utilizando el ID que posee *Wikidata* (ver Sección 2.3.3). Una de las ventajas de utilizar este identificador es que permite extraer información extra en caso de ser necesario, tales como la expectativa de vida, ubicación geográfica en coordenadas, población, entre otros. Se construyó un diccionario de forma manual tomando como base los países más comunes y agregando casos bordes según el *dataset* como: territorios válidos, nombres escritos en otros idiomas que no fueran inglés o con caracteres especiales.

Uno de los países que tuvo más formas de escribir fue la República Democrática del Congo, el cual no se debe confundir con su nación vecina la República del Congo, con 12 casos como se muestra en el Código 3.1. Gracias a mantener un ID único es que se pueden considerar todas estas acepciones como un mismo país, independiente de la fuente de origen.

Código 3.1: Ejemplo en entrada del diccionario

```
{
"Q974": ["Congo, Democratic Republic",
        "The Democratic Republic of the Congo",
        "Congo, Dem. Rep.", "Democratic Republic of Congo",
        "Democratic Republic of the Congo", "DR Congo",
        "Congo Democratic Republic", "Congo DR",
        "Congo, Democratic Republic of the", "Congo [DRC]",
        "Congo, Democratic Republic of", "Democratic_Republic_of_the_Congo"]
}
```

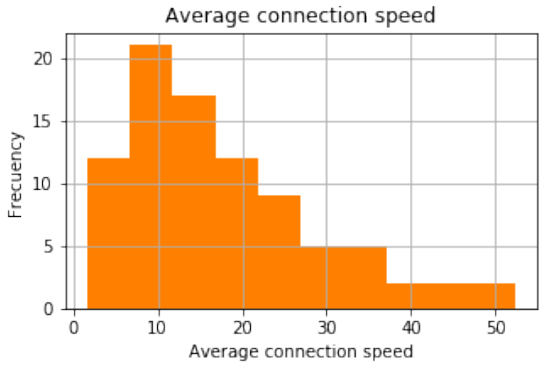
Continuando el ejemplo anterior, la Tabla 3.2 contiene los datos limpios y preprocesados del factor de riesgo de fumadores (Tabla 3.1), donde solo se mantuvieron los datos de aquellos territorios que correspondían a países y los datos numéricos como `floats`, además del ID de Wikidata correspondiente.

Tabla 3.2: Datos limpios y preprocesados extraídos de la base de datos *Number of deaths by risk factor: Smoking* de Our World in Data [70].

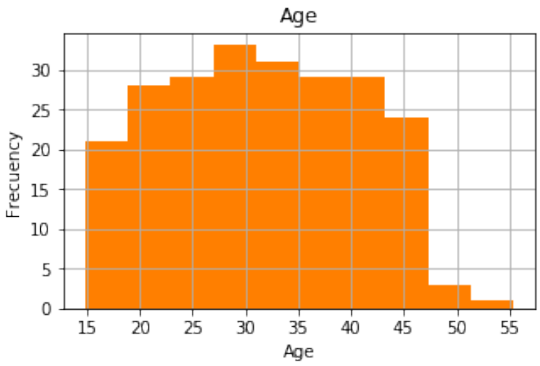
Territory	**Number of associated deaths**	**ID wikidata**
Central African Republic	2509.0	Q929
Chad	1658.0	Q657
Chile	10405.0	Q298
China	2200000.0	Q148
Colombia	19434.0	Q739

3.2. Análisis exploratorio

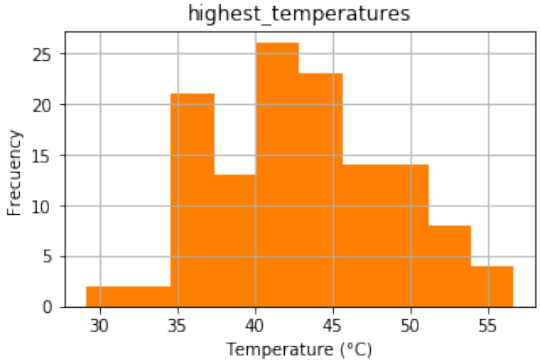
Como primer paso para el modelamiento, es necesario comprender el comportamiento de los datos; para ello se estudiaron las distribuciones de una muestra estocástica de 50 variables. En este grupo se vieron variadas distribuciones como uniformes, gaussianas o fisher. Algunos ejemplos son mostrados en la Figura 3.2.



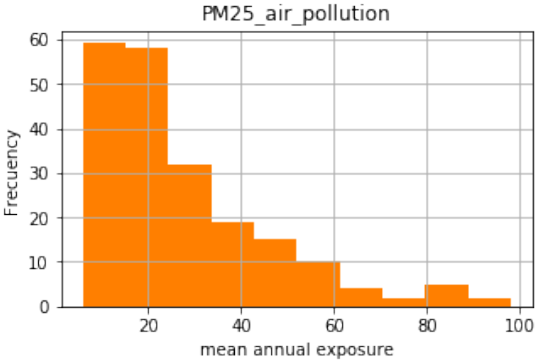
(a) Velocidad de conexión a Internet móvil.



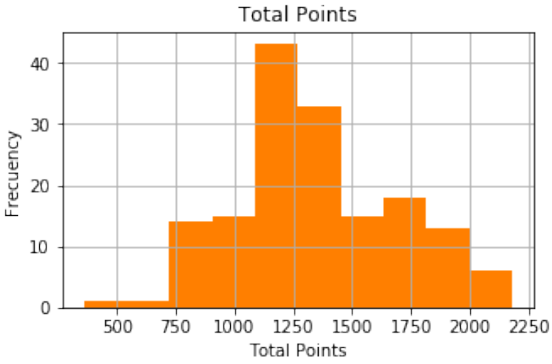
(b) Edad Media.



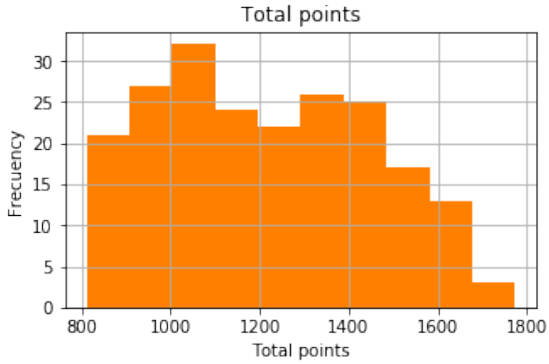
(c) Temperaturas más altas registradas.



(d) Contaminación del aire PM25.



(e) Ranking FIFA mujeres.



(f) Ranking FIFA hombres.

Figura 3.1: Distribuciones de diferentes variables.

3.3. Modelamiento

Ya que los *datasets* no poseían las mismas dimensiones (no contaban con la misma cantidad de territorios o columnas de variables), no se consideran datos estructurados, por lo que una base de datos relacional clásica complicaría la búsqueda de información. Por este motivo se decidió utilizar RDF, un modelo de datos basados en grafos detallado en el capítulo 2.3.1, que con su estructuración a través de sujetos, predicados y objetos, permite la flexibilidad necesaria.

Ya decidido el modelo de datos, se procedió a investigar la estructura más adecuada para *datasets*, escogiendo un vocabulario creado especialmente para este tipo de datos: RDF Data Cube (ver capítulo 2.3.2). Para convertir los datos crudos a RDF, se hizo uso de la herramienta *Tarql* [72], que permite a través de una consulta SPARQL, una transformación de los archivos de *.tsv* a *.ttl* (ver Capítulo 2.3.4).

A modo de ejemplo, se convertirán los datos sobre la población (Tabla 3.3) para dejarlos en el formato del Código 2.1; a través de la *query* Tarql mostrada en el Código 3.2. En ella, en la sección de `WHERE`, se eligen columnas de los datos crudos (en formato CSV, TSV, etc.), y se generan términos en RDF para cada valor elegido. Luego, en `CONSTRUCT` se define cómo se debe estructurar los datos elegidos de cada fila como un grafo RDF. En este caso, se usa la estructura de RDF Data Cube. La salida del proceso es la unión de los grafos generados por cada fila del archivo original. Es necesario definir una consulta Tarql para cada archivo CSV o TSV de entrada.

Tabla 3.3: Extracción de los datos pre-modelados de la población de "The World Bank" [68].

Country	Population	ID	id_row
Hungary	9.8	Q28	85
Iceland	0.4	Q189	86
India	1366.4	Q668	87
Indonesia	270.6	Q252	88
Iran, Islamic Rep.	82.9	Q794	89
Iraq	39.3	Q796	90
Ireland	4.9	Q27	91

Código 3.2: Ejemplo de *query* SPARQL utilizada en Tarql [72].

```
PREFIX ex: <http://ex.org/a#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX qb: <http://purl.org/linked-data/cube#>
PREFIX c19: <http://example.org/ns#>
PREFIX c19-measure: <http://example.org/measure#>
PREFIX c19-interval: <http://example.org/interval#>
PREFIX c19-dimension: <http://example.org/dimension#>
PREFIX sdmx-attribute:  <http://purl.org/linked-data/sdmx/2009/attribute#>
PREFIX wd: <http://www.wikidata.org/entity/>
CONSTRUCT {
  ?obs a qb:Observation;
    qb:dataSet  c19:dataset-population;
    c19-dimension:refPeriod c19-interval:Year2019;
```

```
    c19-measure:population ?data ;
    sdmx-attribute:unitMultiplier  6;
    c19-dimension:refArea ?wdID;
}
FROM <file:population.tsv>
WHERE {
  BIND (URI(CONCAT('c19:population-o0', ?id_row)) AS ?obs)
  BIND (URI(CONCAT('wd:',?ID)) AS ?wdID)
  BIND (xsd:decimal(?Population) AS ?data)
}
```

Esta herramienta es muy útil para modelar los datos que ya estaban en formato *dataset*. Sin embargo, no es muy eficiente cuando se tienen metadatos como el autor, la fecha de extracción o el enlace a los datos que se modelan como `qb:Dataset` de RDF Data Cube. Es por esto que se utilizó un *script* de Python que convertía la *metadata* en archivos *.ttl* de forma manual, finalmente, anexaba los datos generados por *Tarql* para cumplir con las normas de Data Cube Vocabulary.

Otros datos importantes que se tuvieron en cuenta son los atributos y medidas, donde no todas las variables estaban modeladas con la misma magnitud de medida o donde la misma variable estaban divididas en columnas diferentes como es el caso de la separación de género. Asimismo, se tuvieron que definir rangos etarios como *c19-interval:Age0-5* para el *dataset* de "Under-five mortality rate" [40].

Se debe señalar que este minucioso proceso ocupó gran parte del tiempo destinado a la Memoria, y debido a las idas y vueltas de la redefinición de atributos y medidas, se solicitó ayuda a un grupo de estudiantes del ramo "Web de Datos" del semestre de Primavera 2020 de la Universidad, para que tomaran los datos ya limpios y preprocesados para que los transformaran en archivos *.ttl*. Sin embargo, debido a la gran cantidad de variables recopiladas y el tiempo limitado de esta Memoria, es que no se alcanzó a convertir todas las variables, dando prioridad a las que estaban asociadas a una hipótesis (ver el listado en el Anexo A). En total se lograron modelar 79 *datasets* (grafos) ligadas a una hipótesis y 39 *datasets* independientes (ver Anexo B), resultando en un total de 118 grafos con información no referente al COVID-19.

Ya modelados los datos, se decidió almacenar los archivos en un servidor SPARQL llamado Apache Jena Fuseki [73]. Este funciona como "*webapp*", donde se pueden subir los datos y posteriormente consultarlos. Una gran ventaja de este software es que detecta automáticamente los errores de sintaxis o la falta de definiciones, además de poder utilizarse de manera local. Una *query* de esta plataforma se muestra en la Figura 3.2, donde el resultado está en formato tabla.

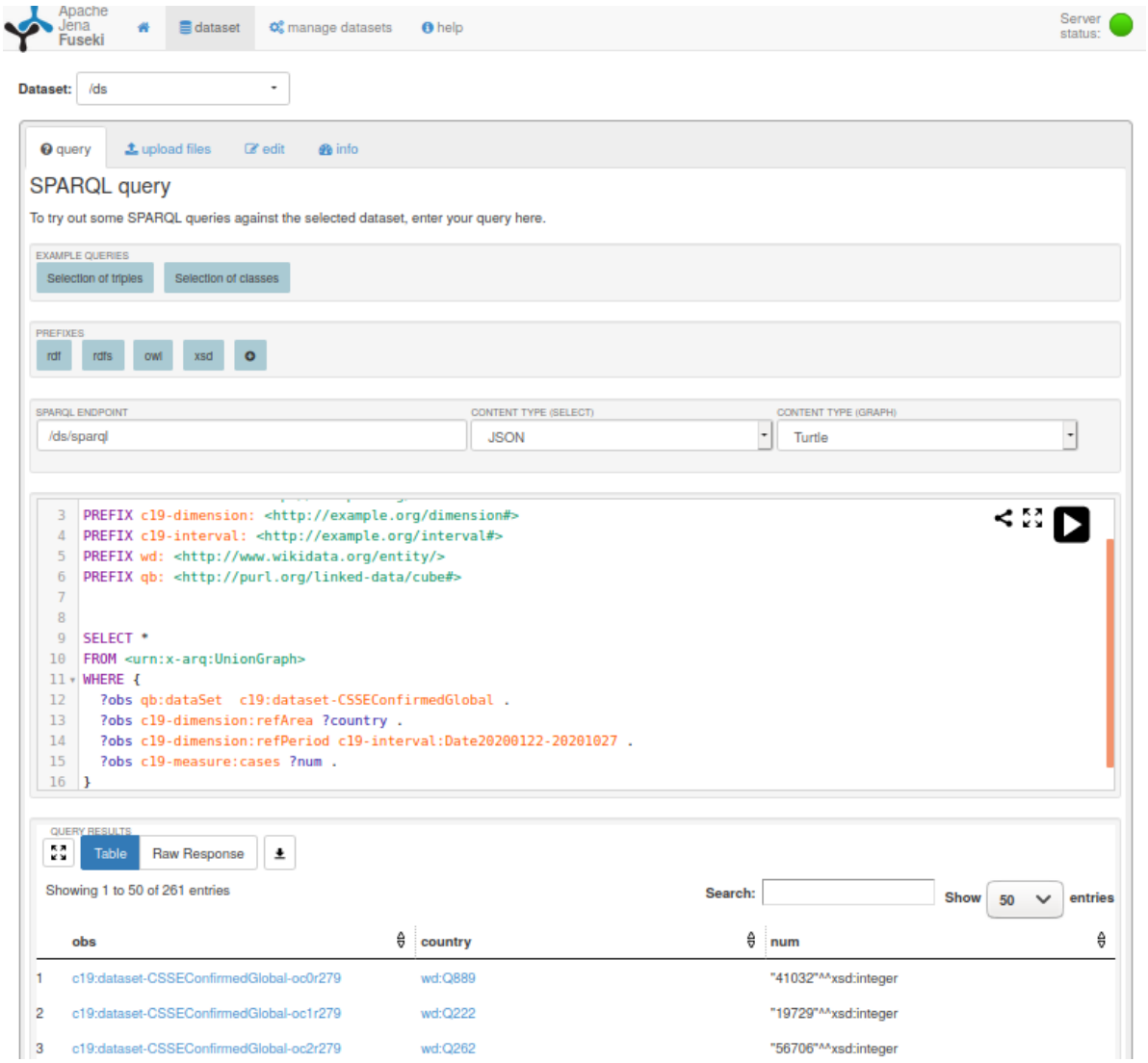


Figura 3.2: Visualización de consulta de la plataforma Apache Jena Fuseki

Sin lugar a dudas, se obtienen muchos beneficios de tener esta nueva base de datos integrada, como es el poder cruzar información de variadas fuentes; agregar, editar y eliminar grafos; seleccionar información específica; entre otros. A continuación se muestran tres ejemplos sobre las ventajas de la integración:

1. En la Tabla 3.4 se seleccionaron los *datasets* que contienen a Chile (Q298), donde en la columna izquierda está la consulta y en la derecha, parte de los títulos resultantes.

2. La Figura 3.3 muestra un extracto del listado de los grafos presentes en el servidor, además de la opción de editar el grafo "http://example.org/graph/ncd_diabetes_mellitus", que incluyen cambiar, agregar o eliminar sus datos.

3. Una consulta a Apache Jena Fuseki por el número de fuentes de información diferentes asociadas para cada país. La Figura 3.4 expone los 3 primeros resultados.

Tabla 3.4: Consulta sobre los *dataset* asociados a Chile (Q298) y extracto de sus resultados

Query	**Título *dataset***

`PREFIX c19-dimension: <http://example.org/dimension#>`
`PREFIX wd: <http://www.wikidata.org/entity/>`
`PREFIX qb: <http://purl.org/linked-data/cube#>`
`PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>`
`PREFIX dct: <http://purl.org/dc/terms/>`
`SELECT distinct ?title`
`FROM <urn:x-arq:UnionGraph>`
`WHERE {`
`?obs rdf:type qb:Observation ;`
`qb:dataSet ?dataset;`
`c19-dimension:refArea wd:Q298 .`
`?dataset a qb:Dataset;`
`dct:title ?title.`
`}`
	Under-five mortality rate
	Wage and salary workers
	Surface area
	Renewable energy consumption
	Purchasing power parity gross national income
	Primary completion rate
	Prevalence of child malnutrition
	Access to electricity
	Account ownership at a financial institution
	Adjusted net savings
	Carbon dioxide emissions
	Contributing family workers
	Expenditures for R&D
	Gross domestic product growth rate
	Incidence of HIV
	Incidence of tuberculosis

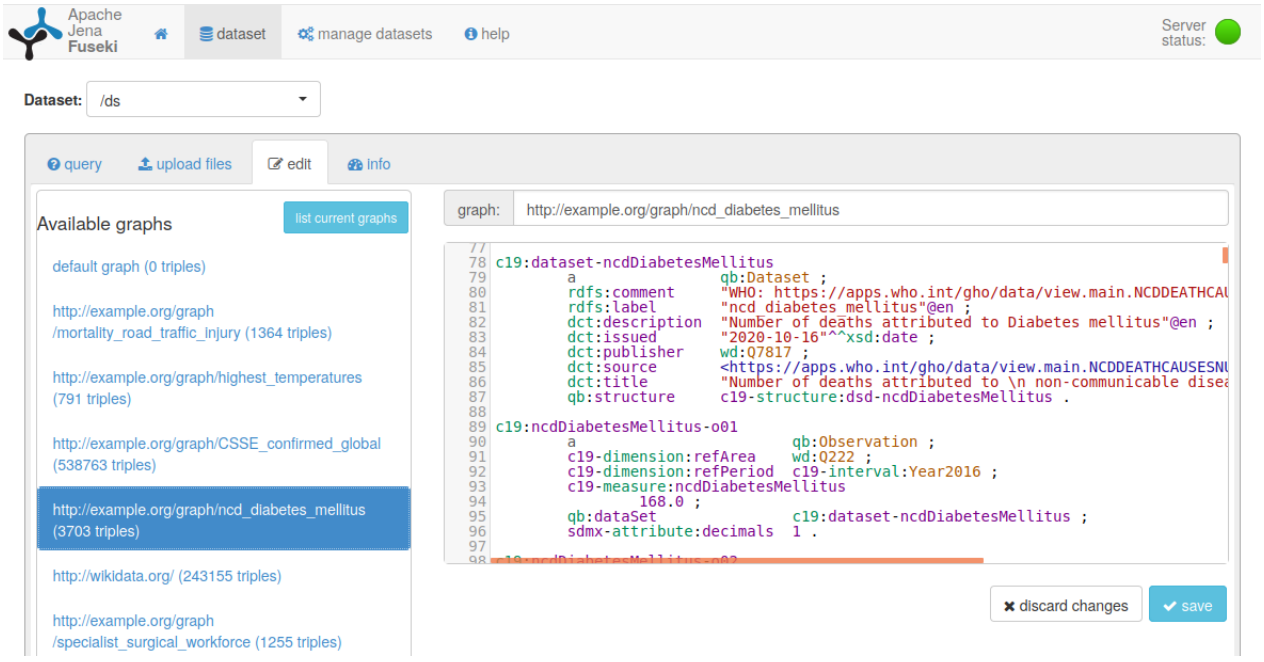


Figura 3.3: Listado y edición de *datasets* en Apache Jena Fuseki

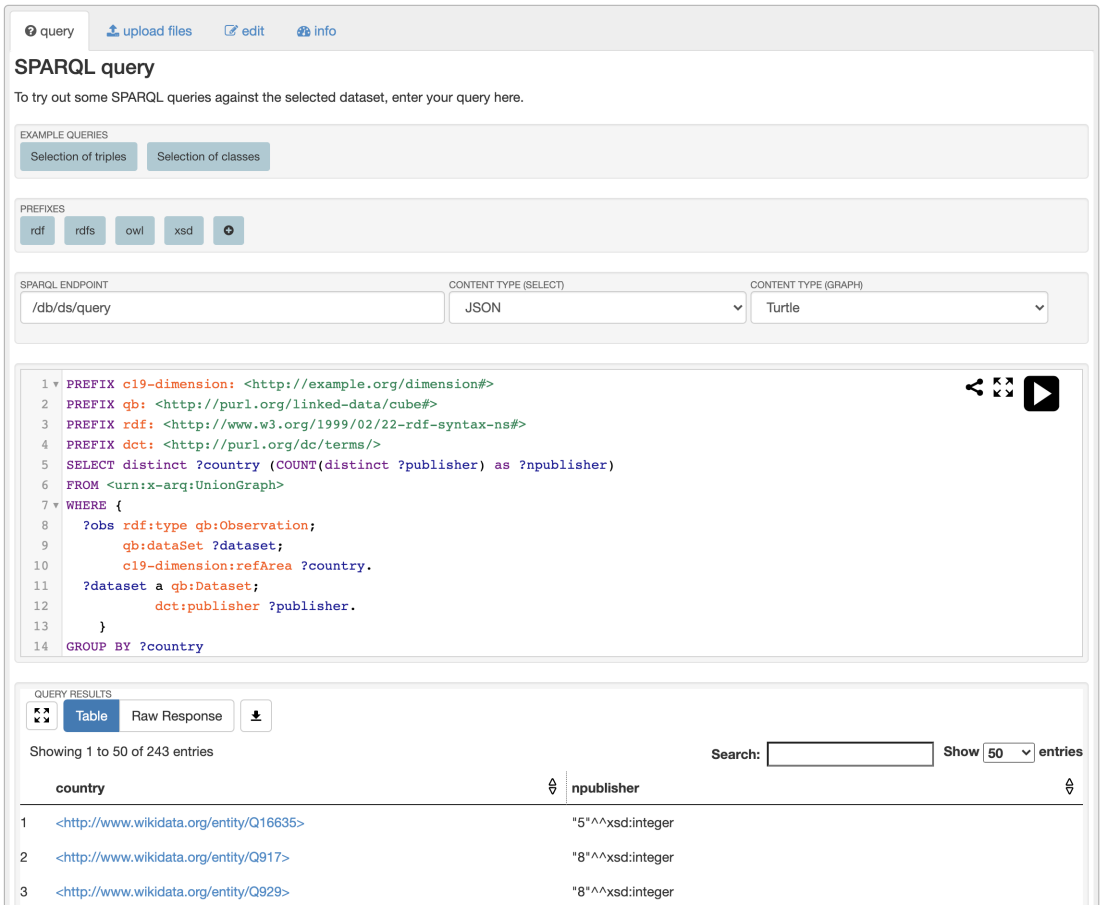


Figura 3.4: Ejemplo de consulta en Apache Jena Fuseki, se muestran el número de fuentes diferentes para los 3 primeros países del resultado.

3.4. Resultado de la integración

Este subcapítulo incluye algunas estadísticas sobre el resultado de la integración de los datos, resumidas en la Tabla 3.5, donde se muestra la cantidad de grafos, tripletas, estructuras y observaciones que posee la base, entre otros.

Tabla 3.5: Estadísticas del resultado de la base integrada.

Medida	Total
Grafos	123
Tripletas	3.331
Variables	168
Estructuras	126
Observaciones	442.420
Territorios	251
Fuentes de información	37
Publicadores	8

Capítulo 4

Desarrollo de la Interfaz de Usuario

Ya teniendo los datos modelados en *.ttl* y subidos a Fuseki, se necesitaba de un sistema para computar las correlaciones y una interfaz que muestra las correlaciones. Según lo planteado en los objetivos de la Memoria (ver Capítulo 1.2), uno de ellos busca implementar una plataforma de visualización de los resultados, por lo que se optó por una página web que exponga las correlaciones a través de una matriz con los valores asociados.

4.1. Requisitos y objetivos

De esta forma, se idearon requerimientos propios para satisfacer el objetivo específico recién mencionado:

- Sistema utilizable por cualquier tipo de usuarios (no restringido a expertos).
- Desplegar los resultados de forma gráfica a través de un mapa de calor, con colores intuitivos para demostrar la fuerza de la correlación.
- Exhibir tanto el valor de la correlación como su valor p.
- Plataforma accesible desde un sitio web.

4.2. Arquitectura

Desde un principio se planeó trabajar con un sistema que fuera utilizable no solo de manera local, sino también que se pudiera conectar a través de una dirección web. Es por ello que se habilitó un servidor en el dominio del Departamento de Ciencias de la Computación de la Universidad, tanto por recursos disponibles como una forma de seguridad y protección de los datos.

La arquitectura del sistema utiliza como base el *micro web framework* Flask [74], escrito en Python. Este funciona como *back end* de la aplicación, extrayendo datos desde el servidor, procesándolos a través de *script* de Python, para finalmente, entregando la información limpia y tratada al *front end*.

Como bien ya se mencionó el Capítulo 3, los datos fueron almacenados en el servidor Apache Fuseki Jena, que corre en segundo plano un proceso continuo en la máquina. Ya conseguidos los datos crudos del servidor, se comienza el proceso de cálculo de las correlacio-

nes, a través de un *script* de Python; con la librería `spicy.stats` se procesan las variables, obteniéndose la información del coeficiente y su valor p respectivo.

Con los datos listos para su visualización, estos son enviados al *front end*, donde se trabaja con el lenguaje HTML5 [75] y JavaScript [76]. Para una mejor representación gráfica, se hizo uso de una matriz de calor que muestra los valores numéricos a través de colores, en particular se empleó un *template* de "*Heat Map Chart*" de AnyChart [77].

La Figura 4.1 es un resumen gráfico de cómo funciona el *back end* y sus conexiones, para obtener un *front end* como sitio web.

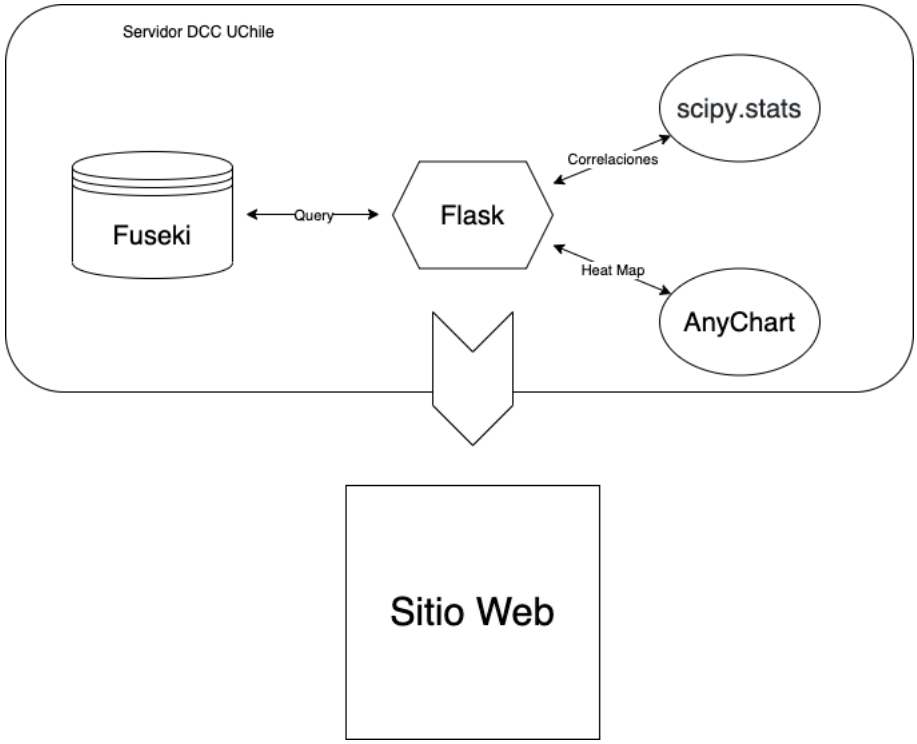


Figura 4.1: Arquitectura del Sistema

4.3. Interfaz

La interfaz está compuesta por dos partes: *back end* y *front end*. El *front end* corresponde a la capa de presentación del Sistema, mientras que el *back end* es la capa de acceso a datos y procesamiento de los mismos. La interfaz está disponible a través de la página web https://c19.dcc.uchile.cl/.

4.3.1. *Back end*

Flask actúa como *back end* del Sistema. La elección de este *framework* sobre otros se debe a su no requerimiento de librerías o herramientas particulares, posibilitando la creación de aplicaciones web en unas pocas líneas de código, es decir, una opción fácil y útil a la hora de utilizar la máquina remota. Cada vez que un usuario ingresa a la página web, Flask trabaja conectándose con Apache Jena Fuseki a través de un *script* de Python, extrayendo todos los grafos existentes y filtrando la información relevante para el *front end*.

Ya con estos datos, se separan en dos grupos: variables referentes y no referentes al COVID-19. Las variables referentes al virus, que finalmente fueron incorporadas en el sistema, son las siguientes:

- **CSSE: Confirmed Global**: Casos confirmados diarios acumulados, información como serie de tiempo.
- **CSSE: Deaths Global**: Muertes diarias acumuladas, información como serie de tiempo.
- **CSSE: Recovered Global**: Casos de recuperados diarios acumulados, información como serie de tiempo.
- **OWiD: Stringency Index**: Medida compuesta basada en 9 indicadores de respuesta que incluyen cierres de escuelas, cierres de lugares de trabajo y prohibiciones de viaje, reescalado a un valor de 0 a 100 (100 = respuesta más estricta)

Con `spicy.stats` se calculan los valores del coeficiente de correlación y su valor p entre los grupos. Algunos de los datos tenían dependencia temporal, especialmente aquellas referentes al COVID-19, de modo que, por simplificación, se tomaron solamente las observaciones con la fecha más reciente como válidas, o en el caso del número de confirmados, muertes y recuperados, se tomaron la suma de los datos sobre todo el periodo disponible. Esta información, complementada con metadatos como el título y dimensiones como el sexo, son enviados al *front end* como formato *json*.

4.3.2. *Front end*

El *front end* consiste en dos vistas principales, una con los resultados según el coeficiente de Pearson (Capítulo 2.4.1) y la segunda con los resultados calculados con el coeficiente de Spearman (Capítulo 2.4.2). Ambas vistas están presentes en las Figuras 4.2 y 4.3, siendo totalmente análogas pero con la diferencia del coeficiente utilizado.

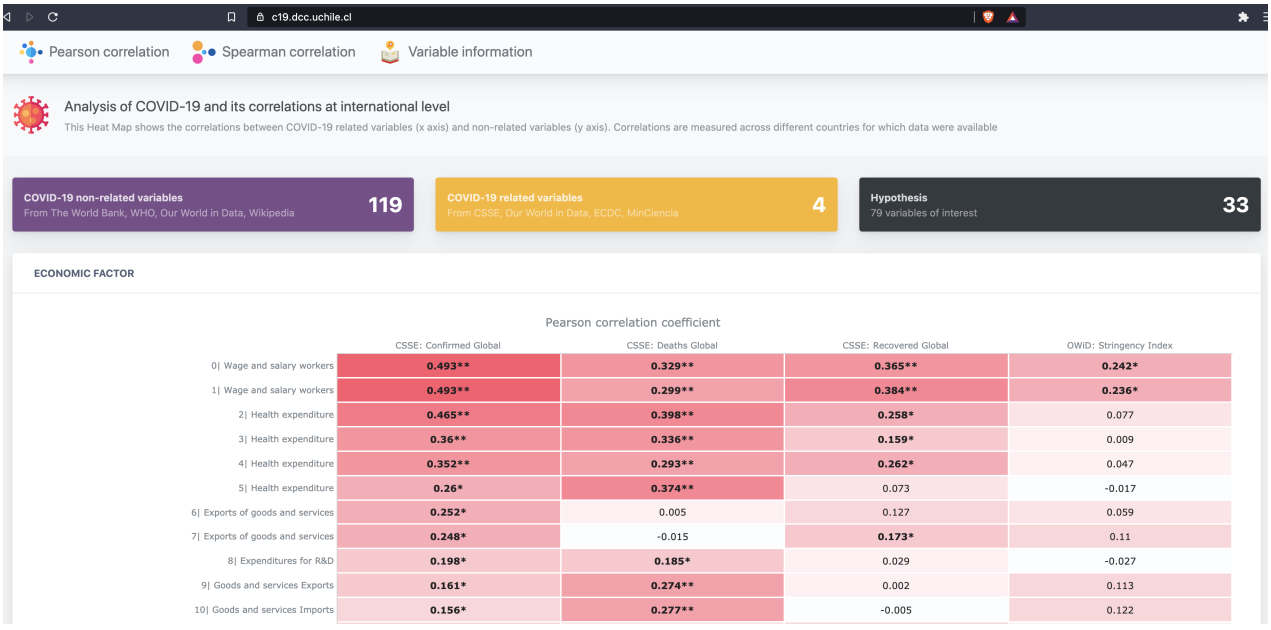


Figura 4.2: Vista del coeficiente de Pearson

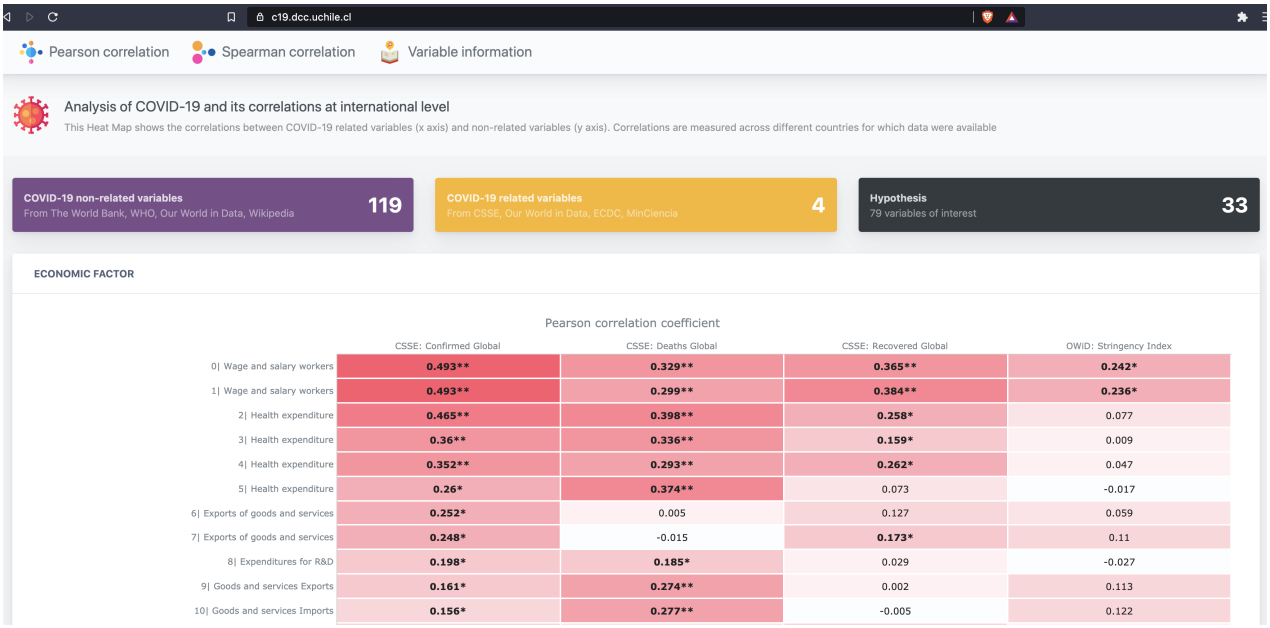


Figura 4.3: Vista del coeficiente de Spearman

Para tener un orden y una mejor visualización de los datos, las variables fueron agrupadas según las categorías descritas en el Anexo A, dividiéndose no solo a través de *cards*, sino también con una matriz por factor. De este modo, se observan las relaciones según el tipo de hipótesis a responder, en vez de una única matriz con todas las variables integradas. Esta separación se muestra en la Figura 4.4.

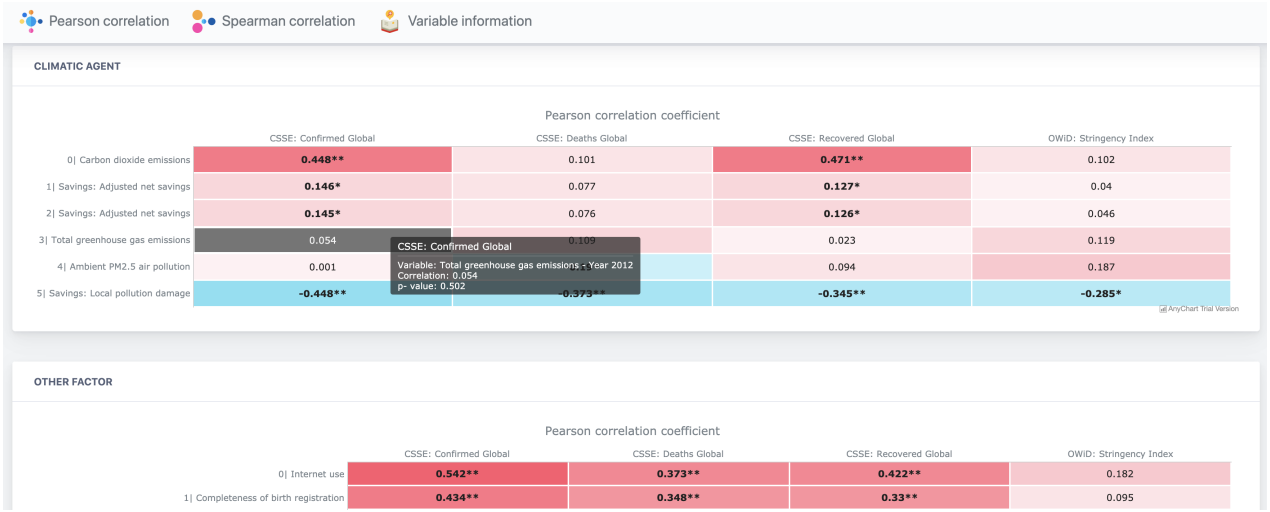


Figura 4.4: Detalle del *Heat Map*.

Un *tooltip* es un cuadro de texto que está oculto de forma predeterminada y se puede mostrar cuando se coloca el cursor sobre un punto de la gráfica, conteniendo información de cada punto. Tal cual se observa en la Figura 4.4, el *tooltip* está configurado para mostrar tanto el título de las variables que están siendo consideradas, como el valor del coeficiente con su valor p.

Otra característica de este *Map Heat* es su gama de colores, donde el rango va desde un celeste claro para los menores coeficientes, hasta un rojo intenso para los mayores valores. Los coeficientes intermedios se asocian con una combinación entre estos dos colores, dependiendo de qué tan grande o pequeño sea el número. Cabe destacar que las variables están ordenadas de mayor a menor, según sus valores en la columna "CSSE: Confirmed Global".

Si bien hay 2 vistas principales, existe una tercera vista que es un diccionario sobre todas las variables que están en los *Heat Maps*. Esta vista se muestra en la Figura 4.5 y consiste en una tabla con tres columnas: el nombre de la variables, las descripción de la misma y si está o no normalizada manualmente.

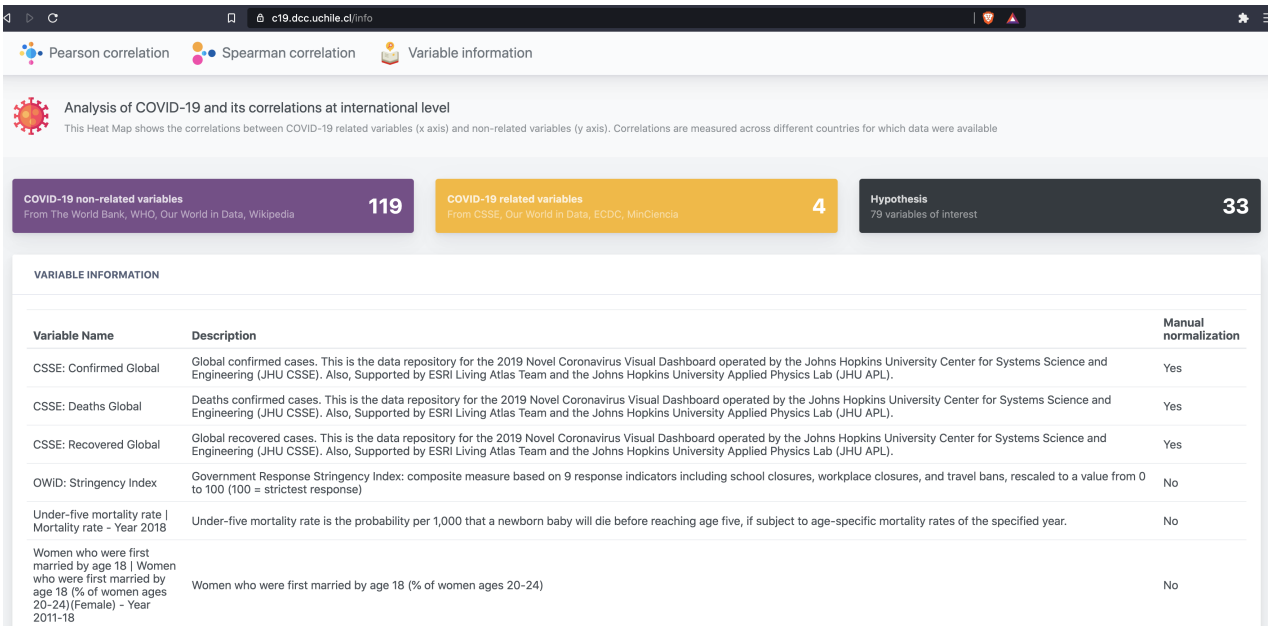


Figura 4.5: Tercera vista de la interfaz: Diccionario de las variables.

4.4. *Feedback*

Ya teniendo toda la plataforma implementada, es necesario incluir un aspecto de evaluación del sistema final. Se realizó una encuesta para determinar la usabilidad de la interfaz, al igual que la encuesta de búsqueda de hipótesis. Esta fue publicada en el Foro de *U-cursos* y permaneció abierta durante tres días.

La encuesta consiguió 52 respuestas y estuvo compuesta por 9 preguntas: 6 con evaluación numérica y 3 como respuestas abiertas. La evaluación numérica tuvo una escala del 1 al 5, donde 5 es el mejor puntaje a alcanzar. Los resultados se encuentran en la Tabla 4.1, donde lo más valorado por los usuarios es el funcionamiento-integración y la utilidad de la plataforma con una media sobre los 4 puntos. Por el contrario, el peor resultado fue el del entendimiento de la interfaz.

Tabla 4.1: Resultados de la encuesta.

Pregunta	**Promedio**
¿Encuentras útil la plataforma?	4,12
¿Qué tan fácil fue entender la información de la plataforma?	2,98
¿Crees que necesitas de un experto para entender la información de la plataforma?	3,42
¿Cómo es la visual de la plataforma?	3,77
¿Qué tanto confías en la información de la plataforma?	3,96
¿Encuentras que la plataforma funciona sin problemas y está bien integrada?	4,25

Las preguntas de respuesta abierta estaban enfocadas en obtener la opinión libre sin las restricciones que tienen los puntajes. Las preguntas en cuestión fueron las siguientes:

1. ¿Qué es lo que más te gustó de la plataforma?
2. ¿Qué es lo que menos te gustó de la plataforma?
3. ¿Tienes algún comentario que ayude a mejorar la plataforma?

A modo general, lo que más gustó a los usuarios fue la estética que tenía la plataforma y la gran cantidad de correlaciones disponibles. Dentro de las posibles mejoras populares está el dar más información sobre el propósito de la página para estar autocontenida y una visualización directa sobre el significado de cada variable.

Capítulo 5

Análisis Estadístico de los Datos

Ya teniendo implementado el sistema completo, con los datos integrados y sus correlaciones respectivas, hace falta ver cuáles son las correlaciones fuertes y en particular, como se comportan las hipótesis planteadas en la Anexo A. Cabe destacar que para evitar que la población fuera una variable confusa (ver Capítulo 2.4.3), se normalizaron manualmente por población aquellas variables que no estaban normalizadas en su fuente de origen.

5.1. Variables con mayor coeficiente

5.1.1. Agentes Climáticos

Dentro de las variables relacionadas con el clima, el coeficiente de Spearman tiene 17 correlaciones significativas (valor $p \leq 0{,}05$), mientras que el coeficiente de Pearson posee 11. Sin embargo, las correlaciones más fuertes se observan en en las mismas variables: "Carbon dioxide emissions" y "Particulate emission damage", para las variables COVID de casos confirmados, muertes y recuperados. Dentro de estas, Pearson cuenta con 5 correlaciones que van de 0,345 a 0,471 en valor absoluto y en el caso de Spearman, son 6 correlaciones con un intervalo de 0,445 a 0,613. La correlación en que difieren es la que está dada por "Carbon dioxide emissions" y "CSSE: Deaths Global", donde posee un valor de 0,101 según Pearson pero con una valor p de 0,18, considerándose no significativa. Su máximo se alcanza en el cruce de "Carbon dioxide emissions" y "CSSE: Recovered Global" con un r de 0,471.

En general, se observa una tendencia a un fuerte acuerdo en la variable "Carbon dioxide emissions" y un fuerte desacuerdo para la variable "Particulate emission damage", en los casos de COVID-19. El detalle de las correlaciones más fuertes se encuentran en la Tabla 5.1 y la Tabla 5.2 para los coeficientes de Pearson y Spearman respectivamente.

Tabla 5.1: Agentes Climáticos: Estadísticas más importantes para el coeficiente de Pearson

Variable no-referente COVID-19	**Variable referente COVID-19**	r	**Valor** p
Carbon dioxide emissions	CSSE: Recovered Global	0,471	0
Carbon dioxide emissions	CSSE: Confirmed Global	0,448	0
Particulate emission damage	CSSE: Confirmed Global	-0,448	0
Particulate emission damage	CSSE: Deaths Global	-0,373	0
Particulate emission damage	CSSE: Recovered Global	-0,345	0

Tabla 5.2: Agenties Climáticos: Estadísticas más importantes para el coeficiente de Spearman

Variable no-referente COVID-19	**Variable referente COVID-19**	ρ	**Valor** p
Particulate emission damage	CSSE: Confirmed Global	-0,613	0
Particulate emission damage	CSSE: Deaths Global	-0,522	0
Carbon dioxide emissions	CSSE: Confirmed Global	0,515	0
Particulate emission damage	CSSE: Recovered Global	-0,471	0
Carbon dioxide emissions	CSSE: Recovered Global	0,458	0
Carbon dioxide emissions	CSSE: Deaths Global	0,445	0

5.1.2. Determinantes de Salud

De las 208 correlaciones que cuenta esta categoría, 110 correlaciones de Pearson y 137 correlaciones de Spearman permiten rechazar la hipótesis nula. Existen 63 correlaciones significativas de Pearson que poseen un rango entre 0,301 y 0,529 en valor absoluto; de estas mismas, las más que poseen un correlación más fuerte son "Population by gender: Female" y "Population by gender: Male", con $-0{,}529$ y 0,529 respectivamente, ambos para el cruce de con "CSSE: Recovered Global". En el caso de Spearman, hay 94 correlaciones significativas dentro del rango 0,304-0,610 en valor absoluto; sus correlaciones más fuertes fueron las variables no referentes al virus "No access to handwashing facility" y "Under-five mortality rate" y la variable referente al COVID-19 "CSSE: Confirmed Global", con un valor ρ de $-0{,}610$ y $-0{,}605$ correspondientemente.

Es notable señalar que la única variable no referente al COVID-19 que posee una correlación Pearson fuerte entre todas las variables referentes al COVID-19 fue "Prevalence of child malnutrition", obteniendo un promedio de $-0{,}429$. Esta misma variables también tiene una correlación fuerte con todas las variables referentes al COVID-19 en las correlaciones de Spearman (promedio de $-0{,}472$), pero además se le suma la variable "Body Mass Index: Female" con un promedio de 0,381. Las Tablas 5.3 y 5.4 contienen los detallles para las 6 variables más fuertes de Pearson y Spearman respectivamente.

Tabla 5.3: Determinantes de Salud: Estadísticas más importantes para el coeficiente de Pearson

Variable no-referente COVID-19	**Variable referente COVID-19**	r	**Valor** p
Population by gender: Female	CSSE: Recovered Global	-0,529	0
Population by gender: Male	CSSE: Recovered Global	0,529	0
Prevalence of overweight children	CSSE: Confirmed Global	0,512	0
Body Mass Index: Male	CSSE: Confirmed Global	0,494	0
Prevalence of child malnutrition	CSSE: Confirmed Global	-0,489	0
Life expectancy birth total	CSSE: Confirmed Global	0,479	0

Tabla 5.4: Determinantes de Salud: Estadísticas más importantes para el coeficiente de Spearman

Variable no-referente COVID-19	**Variable referente COVID-19**	ρ	**Valor** p
No access to handwashing facility	CSSE: Confirmed Global	-0,610	0
Under-five mortality rate	CSSE: Confirmed Global	-0,605	0
Infant mortality rate	CSSE: Confirmed Global	-0,597	0
Adult mortality rate	CSSE: Confirmed Global	-0,592	0
Life expectancy at birth: Female	CSSE: Confirmed Global	0,586	0
Life expectancy birth total	CSSE: Confirmed Global	0.583	0

5.1.3. Factores Económicos

En el servidor se encuentran 156 correlaciones según el coeficiente de Pearson, de las cuales 51 se consideran significativas y de estas hay 19 correlaciones fuertes con un rango de 0,300 a 0,493. El valor más alto es de 0,493, siendo un empate entre las variables "Wage and salary workerd: Female" en conjunto con "CSSE: Confirmed Global" y "Wage and salary workers: Male" con la misma variable referente al COVID-19. Cabe destacar que la única variable que tiene correlaciones fuertes para todas las variables referentes al virus es "Urban population living in slums" que posee un promedio de 0,411. La Tabla 5.5 es una especificación de las 6 correlaciones más fuertes de Pearson.

Tabla 5.5: Factores Económicos: Estadísticas más importantes para el coeficiente de Pearson

Variable no-referente COVID-19	**Variable referente COVID-19**	r	**Valor** p
Wage and salary workers: Female	CSSE: Confirmed Global	0,493	0
Wage and salary workers: Male	CSSE: Confirmed Global	0,493	0
Current health expenditure per capita, PPP	CSSE: Confirmed Global	0,465	0
Urban population living in slums	OWiD: Stringency Index	-0,448	0
Urban population living in slums	CSSE: Confirmed Global	-0,427	0
Urban population living in slums	CSSE: Recovered Global	-0,411	0

Dentro de las correlaciones de Spearman, existen 74 correlaciones que rechazan la hipótesis nula, de ella hay 36 correlaiones que pueden ser consideradas como fuertes por su rango entre 0,300 a 0,634, en valor absoluto. Al igual que en el caso de Pearson, "Urban population living in slums" también es la única variable que posee correlaciones fuertes en las cuatro variables referentes al virus, con un promedio de $-0{,}441$. La correlación más fuerte de 0,634 está en el cálculo entre "Current health expenditure per capita, PPP" y "CSSE: Confirmed Global", más correlaciones fuertes están detalladas en la Tabla 5.6.

Tabla 5.6: Factores Económicos: Estadísticas más importantes para el coeficiente de Spearman

Variable no-referente COVID-19	**Variable referente COVID-19**	ρ	**Valor** p
Current health expenditure per capita, PPP	CSSE: Confirmed Global	0,634	0
Wage and salary workers: Female	CSSE: Confirmed Global	0,623	0
Current health expenditure per capita	CSSE: Confirmed Global	0,621	0
Wage and salaried workers: Male	CSSE: Confirmed Global	0,611	0
External health expenditure	CSSE: Confirmed Global	-0,567	0
Current health expenditure per capita, PPP	CSSE: Deaths Global	0,553	0

5.1.4. Otros Factores

Las correlaciones que pueden rechazar la hipótesis nula de Otros Factores, calculadas según Pearson, son 94 correlaciones de un total de 292. De estas 94, 42 son correlaciones fuertes que están entre 0,300 y 0,542 en valor absoluto. Adicionalmente, se distingue una sola variable no relacionada al virus que cuenta con correlaciones fuertes en todas las variables COVID-19 disponibles: "Contributing family workers: Male" que tiene un fuerte desacuerdo en promedio de $-0{,}359$.

En Otros Factores los temas abarcados son muy diversos, en la Tabla 5.7 los tópicos más populares comprenden desde el uso de Internet, el porcentaje de registros de nacimiento completos, los sostenedores del hogar hasta los gastos en subsidios del gobierno.

Tabla 5.7: Otros Factores: Estadísticas más importantes para el coeficiente de Pearson

Variable no-referente COVID-19	**Variable referente COVID-19**	r	**Valor** p
Individuals using the Internet	CSSE: Confirmed Global	0,542	0
Completeness of birth registration	CSSE: Confirmed Global	0,434	0
Individuals using the Internet	CSSE: Recovered Global	0,422	0
Contributing family workers: Male	CSSE: Confirmed Global	-0,415	0
Contributing family workers: Female	CSSE: Confirmed Global	-0,408	0
Government Expenditure: Subsidies	CSSE: Deaths Global	0,406	0

Para el coeficiente de Spearman se encontraron 134 correlaciones significantes, de las cuales 89 son correlaciones fuertes con un rango de $-0{,}301$ a 0,637 según su valor abosoluto de ρ. Si bien no hay ninguna correlación que destaque por ser fuerte en las cuatro variables COVID-19, sí resalta en la Tabla 5.8, el hecho que aparezcan correlaciones fuertes relacionadas al número de trabajadores de la salud ("Specialist surgical workforce" y "Health workers: Physicians") respecto a la Tabla 5.7.

Tabla 5.8: Otros Factores: Estadísticas más importantes para el coeficiente de Spearman

Variable no-referente COVID-19	**Variable referente COVID-19**	ρ	**Valor** p
Specialist surgical workforce	CSSE: Confirmed Global	0,637	0
Individuals using the Internet	CSSE: Confirmed Global	0,634	0
Contributing family workers: Male	CSSE: Confirmed Global	-0,598	0
Completeness of birth registration	CSSE: Confirmed Global	0,594	0
Health workers: Physicians	CSSE: Confirmed Global	0,586	0
Specialist surgical workforce	CSSE: Deaths Global	0,566	0

5.2. Estadísticas asociadas

Para comprender mejor los resultados mostrados en la página web, es importante tener información complementaria al resultado visual. Es por ello que se calcularon estadísticas propias de los coeficientes con el fin de entender su comportamiento. El cómputo considera como válidas a aquellas relaciones cuyo *p-value* fuera menor o igual a 0,05. Estas estadísticas corresponden a la cantidad de variables aceptadas, el promedio, la desviación estándar, el mínimo, los cuartiles (25 %, 50 % y 75 %) y el máximo para cada variable referente al COVID-19. Es relevante mencionar que la base contiene 168 variables disponibles, resultando en 672 correlaciones.

Las estadísticas están detalladas en la Tabla 5.10 para el coeficiente de Pearson y la Tabla 5.9 para el coeficiente de Spearman. Ambas tablas consideran el cálculo entre todas las variables no referentes al COVID-19 versus cada una de las variables referentes al virus.

Tabla 5.9: Estadísticas más importantes para el coeficiente de Pearson

	CSSE: Confirmed Global	**CSSE: Deaths Global**	**CSSE: Recovered Global**	**OWiD: Stringency Index**
count	87	90	59	26
mean	0,0487	0,0409	-0,0048	-0,1184
std	0,3549	0,3103	0,3119	0,2708
min	-0,4890	-0,4390	-0,5290	-0,4480
25 %	-0,3165	-0,2913	-0,2905	-0,3000
50 %	0,1980	0,2125	-0,1660	-0,2610
75 %	0,3490	0,3110	0,3090	0,2360
max	0,5420	0,4110	0,5290	0,3700

Tabla 5.10: Estadísticas más importantes para el coeficiente de Spearman

	CSSE: Confirmed Global	**CSSE: Deaths Global**	**CSSE: Recovered Global**	**OWiD: Stringency Index**
count	118	119	87	32
mean	0,1115	0,1072	0,0812	0,0981
std	0,4091	0,3858	0,3394	0,2663
min	-0,6130	-0,5460	-0,4840	-0,4360
25 %	-0,3185	-0,3200	-0,3000	-0,2322
50 %	0,2875	0,2720	0,2210	0,2375
75 %	0,4122	0,4280	0,3545	0,2820
max	0,6370	0,5660	0,5100	0,3950

5.3. Verificación de hipótesis

Como se mencionó antes, se trató de dar prioridad a las variables que estaban ligadas a una hipótesis, por lo que la verificación de hipótesis es solo sobre aquellas variables que fueron subidas a Apache Jena Fuseki. La información de la Tabla 5.11 muestra para cada hipótesis el número de variables asociadas originalmente y el valor absoluto máximo para las correlaciones con un valor p menor o igual a 0,05, ordenadas de mayor a menor por la última columna. Como no se tienen las distribuciones de todas las variables, los datos detallados son sobre las correlaciones del coeficiente de Spearman ya que ésta es más confiable al no asumir una distribución normal.

Al tomar las 10 primeras correlaciones máximas absolutas, todas poseen un máximo superior a 0,5. Además, 5 hipótesis corresponden a la categoría de Determinantes de Salud, resultado muy esperable dado que el desarrollo de cualquier enfermedad depende de las enfermedades de base o las características particulares de la persona como pueden ser la edad y el tipo de sangre.

Un dato no esperable a simple vista es el puesto número 1 que obtenido por "*Social Media*", en particular, su cruce de "*Internet use*" con "*CSSE: Confirmed Global*" con una correlación con un fuerte acuerdo de 0,634; habría que buscar más información y modelar otro tipo de variables para descartar que esta correlación no sea producto de una variable confusa, como puede ser la cantidad de recursos o el nivel de desarrollo del país. Otro dato que podría contener una variable confusa como el "Producto Interno Bruto" es el ranking FIFA, debido a que en su mayoría, los primeros lugares del ranking lo ocupan países con un alto PIB.

La variable "*Family members who work*" fue propuesta para evaluar la necesidad que tienen las personas de salir a trabajar en cuarentena. En general, este valor puede indicar una economía poco desarrollada (bajas tasas de empleabilidad, economía rural, etc). Un país con estas características puede que tenga una bajo nivel de testeo del virus, por lo que los números asociados al COVID-19 pueden que sean más bajos que la realidad. Para aceptar o rechazar estas teorías habría que hacer un estudio más profundo.

Finalmente, la única hipótesis relativa a los Agentes Climáticos es la contaminación del aire; esta correlación puede estar dada por el hecho de que siendo el COVID-19 una enfermedad respiratoria infecciosa, la contaminación reduce la capacidad respiratoria; también puede estar dado por la transmisión indirecta que favorece contaminantes como el PM2.5, ya que actúa como transportador del virus, plateado Nor et.al [78]

Tabla 5.11: Correlaciones absolutas máximas para cada hipótesis según el coeficiente de Spearman

Hipótesis	**Var. Originales**	**Var. Incluidas**	**Abs. Máx**
Social Media	3	3	0,634
Family members who work	3	1	0,598
Obesity	4	4	0,581
Cancer	2	1	0,558
Age	2	1	0,525
Air Pollution	10	3	0,515
Blood Type	1	1	0,512
Respiratory diseases	5	2	0,511
Government	19	19	0,502
FIFA ranking	2	2	0,501
Depression	2	1	0,480
Imports/Exports	15	15	0,467
Poverty	4	1	0,461
International flights	2	2	0,441
Income	5	3	0,432
Mobility	3	3	0,391
Women leadership	2	2	0,346
Cardiovascular Disease	3	3	0,344
Smoking	3	1	0,325
Diabetes	4	4	0,276
Dementia	2	1	0,259
Hypertension	4	1	0,228
Gender	1	1	0,163
Pollution	3	2	0,160
Domestic violence	2	1	0,143
Density Poblation	3	1	0,087
Winter	1	0	N/A
Humidity	1	0	N/A
Weather	1	0	N/A
Temperature	4	0	N/A
Precipitation	1	0	N/A
Vitamin D	1	0	N/A
Ethnic group	1	0	N/A

Capítulo 6

Conclusiones

El COVID-19, una enfermedad viral que comenzó como un brote en Wuhan, China y rápidamente escaló a nivel mundial, ha marcado un hito en la historia del mundo. Frente a la falta de información y protocolos adecuados para manejar una situación de esta escala, la pandemia llevó numerosos países a imponer cuarentenas, sin tener una clara idea sobre qué factores son los que afectan a la propagación de este virus sin precedentes. Es por esto, que en este trabajo se planteó la idea de explorar cuáles son los fenómenos que están relacionados al COVID-19.

Para ello, se propuso correlacionar estadísticamente variables referentes al virus y variables no referentes al COVID-19, según hipótesis ya establecidas en el mundo científico. Para la búsqueda de estas hipótesis, se realizó una investigación a través de las publicaciones que ha proporcionado la Academia. Estas abordaban relaciones , por ejemplo, el efecto de la obesidad en la gravedad de la enfermedad. Sin embargo, este tipo de exploración contiene el sesgo de la información recopilada por un grupo reducido de personas, por consiguiente, se publicó una encuesta en la plataforma "*U-cursos*", específicamente en el Foro de la Facultad de Ciencias Físicas y Matemáticas, pidiendo a los usuarios dar hipótesis que conocieran al respecto. De este modo, se logró reunir 33 hipótesis.

Al contar con una noción de la dirección que debía tomar la búsqueda de variables, se exploraron diferentes fuentes de información que contuvieran datos no solamente sobre el virus, sino que tuvieran contenido que, *a priori*, no estuviera relacionado. Se examinaron *datasets* que no solo incluyeran tópicos que tradicionalmente afectan en una pandemia, como lo son la densidad poblacional o enfermedades de base respiratorias; sino también temas atípicos como son el ranking FIFA o el uso de internet. Adicionalmente, se decidió profundizar la recopilación, tomando todas las posibles variables encontradas. De esta forma, se recolectaron 525 variables no referentes al virus de fuentes como The World Bank [40], Our World in Data [41], entre otros; además de 14 variables relacionadas al COVID-19 de sitios como el de Johns Hopkins University [36]. Con estos dos pasos mencionados, se cumplieron los dos primeros objetivos específicos planteados en el Capítulo 1.2.

Para lograr el tercer objetivo específico sobre la integración de las variables, se hizo un pre-procesamiento en el que se eliminaron datos no válidos como lo era, por ejemplo, información por continente, se cambiaron los formatos para que fueran compatibles como *floats* en Python y se les agregó un *id* para cada territorio según su codificación en Wikidata [56] con el fin de asociar los países con un identificador único. Ya con esto, se realizó un modelamiento de los datos basado en el vocabulario *RDF Data Cube* debido a su compatibilidad con estructuras tipo *datasets*. Para transformar los archivos a RDF, se ocupó la herramienta

Tarql que permitía dejarlos en formato *turtle* mediante un consulta SPARQL; en conjunto, se utilizó un *script* de Python para modelar metadatos. Una vez convertidos, estos se subieron al servidor Apache Fuseki Jena; las mayores ventajas de tener los datos en el servidor son el poder cruzar diferentes variables, editar grafos y poder seleccionar información específica. Esta etapa del trabajo fue la más extensa y compleja de realizar, debido a la inversión de tiempo que llevó modelar los datos y el manejo de nuevos *software*.

Ya teniendo integrados los datos, se procedió a calcular las correlaciones. Se emplearon los coeficientes de correlación de Pearson y Spearman a través de las librerías `scipy.stats` de Python `pearsonr` y `spearmanr`. Con un *script* se extrajeron los datos desde el servidor, donde se asoció cada variable no referente al virus con todas las relacionadas a este. Se computaron dos valores para cada correlación, el de su coeficiente que exhibe qué tan fuerte es esta correlación, y el valor p que habla sobre la capacidad de rechazar la hipótesis nula.

Finalmente, se implementó una plataforma que permite la visualización de las correlaciones. La arquitectura se llevó a cabo en un servidor del Departamento de Computación de la Universidad. En él se montó el *framework* Flask que funcionó como eje central, trayendo los datos desde Fuseki, luego calculando las correlaciones a través `scipy.stats` para enviarlas al *front end*. La visualización se realizó mediante un *Heat Map*, donde el color se hacía más intenso a medida que la correlación fuera más fuerte. La interfaz quedó completamente integrada y ejecutada bajo un dominio web, pudiendo ser accesible desde cualquier lugar.

Asimismo se realizó una encuesta para evaluar la plataforma, recibiendo un total de 52 respuestas. En ella se destacaba la utilidad y la funcionalidad de la plataforma, además, comentar posibles mejoras como lo fueron un contexto explícito para que la interfaz fuera autocontenida y una visualización directa del significado de cada variable, en vez de la vista del diccionario.

Como resultado de este trabajo se obtuvieron 79 variables asociadas a alguna hipótesis en el servidor, con 250 correlaciones significativas (valor $p \leq 0{,}05$). También se observó que las correlaciones computadas según el coeficiente de Pearson eran más estrictas. De ellas, las variables que más destacan son "*Individuales using the Internet*", "*Population by gender*", "*Prevalence of overweight in children*", "*Body Mass Index*" que poseen valores r mayores a 0,5. De las 33 hipótesis planteadas al principio de la Memoria, se tomó el valor absoluto máximo de cada variable asociada a estas, con el fin de tener una idea sobre el comportamiento de la hipótesis. Para crear un ranking con las mayores correlaciones absolutas, se utilizó el coeficiente de Spearman debido que no asume una distribución normal. De los 10 primeros puestos, 5 son ocupados por determinantes de la salud que abarcan desde enfermedades pre-existentes como la obesidad, cáncer y enfermedades respiratorias hasta características individuales como la edad y el tipo de sangre. El único Agente Climático del top 10 es "*Air Pollution*" que puede estar relacionado con la naturaleza del COVID-19, una enfermedad infecciosa respiratoria y con la capacidad del PM2.5 de ser un gente transportador del virus [78]. Es importante resaltar que si bien se consideran como aceptadas las hipótesis, estas en ningún momento expresan una relación de causalidad.

A modo de trabajo a futuro se pueden modelar las variables que quedaron pendientes, así el sistema quedaría más robusto, pudiendo llegar a resultados más representativos o correlaciones inesperadas. También, queda pendiente el modelamiento a través del tiempo, donde se aplicaría *Heat Maps* con dimensiones temporales como pueden ser meses o semanas. Otro punto es el ahondamiento de las correlaciones ya encontradas, donde se pueden plantear nuevas hipótesis o investigaciones más profundas que ayuden a descartar variables confusas. Además, la plataforma puede ser mejorada en base a los comentarios de la encuesta y se

puede agregar una opción que permita realizar consultas para que cualquier usuario pueda obtener información de esta base integrada.

Bibliografía

[1] A. E. Gorbalenya, S. C. Baker, R. S. Baric, R. J. de Groot, C. Drosten, A. A. Gulyaeva, B. L. Haagmans, C. Lauber, A. M. Leontovich, B. W. Neuman, D. Penzar, S. Perlman, L. L. Poon, D. V. Samborskiy, I. A. Sidorov, I. Sola, and J. Ziebuhr, "The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, volume = 5, year = 2020," *Nature Microbiology*, no. 4, pp. 536–544.

[2] W. H. Organization, "QA on coronaviruses (COVID-19), howpublished = https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses, note = Accessed: 2020-08-01."

[3] T. J. Times, "All schools in Japan told to close until April over virus outbreak." https://www.japantimes.co.jp/news/2020/02/27/national/hokkaido-coronavirus-school/#.XwvWaBEpA5k. Accessed: 2020-08-01.

[4] Reuters, "Austria bans indoor events of more than 100 people, minister says." https://www.reuters.com/article/us-health-coronavirus-austria-events-idUSKBN20X1EV. Accessed: 2020-08-01.

[5] CNA, "Shoot them dead': Philippine President Duterte warns COVID-19 lockdown violators." https://www.channelnewsasia.com/news/asia/covid-19-philippines-lockdown-duterte-coronavirus-12601384. Accessed: 2020-08-01.

[6] O. W. in Data, "Total and daily confirmed COVID-19 cases, Taiwan." https://ourworldindata.org/grapher/total-and-daily-cases-covid-19?country=~TWN. Accessed: 2020-08-01.

[7] O. W. in Data, "Total confirmed COVID-19 deaths: how rapidly are they increasing? Taiwan." https://ourworldindata.org/grapher/covid-confirmed-deaths-since-5th-death?country=~TWN. Accessed: 2020-08-01.

[8] S. Mike Moffitt, "Why Taiwan's COVID-19 death rate is shockingly low." https://www.sfgate.com/bayarea/article/Why-Taiwan-s-COVID-19-death-rate-is-shockingly-low-15130341.php. Accessed: 2020-08-01.

[9] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour, "Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter," *Cureus*, vol. 12, no. 3, 2020.

[10] M. Lipsitch, "Seasonality of SARS-CoV-2: Will COVID-19 go away on its own in warmer weather," *Center for Communicable Disease Dynamics, Harvard University (https://ccdd. hsph. harvard. edu/will-covid-19-go-away-on-its-own-in-warmer-weather/, accessed 19 April 2020)*, 2020.

[11] J. Wang, K. Tang, K. Feng, and W. Lv, "High temperature and high humidity reduce

the transmission of COVID-19," *Available at SSRN 3551767*, 2020.

[12] J. M. Rhodes, S. Subramanian, E. Laird, and R. A. Kenny, "low population mortality from COVID-19 in countries south of latitude 35 degrees North supports vitamin D as a factor determining severity," *Alimentary pharmacology & therapeutics*, vol. 51, no. 12, pp. 1434–1437, 2020.

[13] J. Rocklöv and H. Sjödin, "High population densities catalyse the spread of COVID-19," *Journal of travel medicine*, vol. 27, no. 3, p. taaa038, 2020.

[14] Y. Ma, Y. Zhao, J. Liu, X. He, B. Wang, S. Fu, J. Yan, J. Niu, J. Zhou, and B. Luo, "Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China," *Science of The Total Environment*, p. 138226, 2020.

[15] J. Xie and Y. Zhu, "Association between ambient temperature and COVID-19 infection in 122 cities from China," *Science of the Total Environment*, vol. 724, p. 138201, 2020.

[16] P. Shi, Y. Dong, H. Yan, C. Zhao, X. Li, W. Liu, M. He, S. Tang, and S. Xi, "Impact of temperature on the dynamics of the COVID-19 outbreak in China," *Science of The Total Environment*, p. 138890, 2020.

[17] H. Qi, S. Xiao, R. Shi, M. P. Ward, Y. Chen, W. Tu, Q. Su, W. Wang, X. Wang, and Z. Zhang, "COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis," *Science of the Total Environment*, p. 138778, 2020.

[18] R. Tosepu, J. Gunawan, D. S. Effendy, H. Lestari, H. Bahar, P. Asfian, *et al.*, "Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia," *Science of The Total Environment*, p. 138436, 2020.

[19] Á. Briz-Redón and Á. Serrano-Aroca, "A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain," *Science of the Total Environment*, p. 138811, 2020.

[20] D. N. Prata, W. Rodrigues, and P. H. Bermejo, "Temperature significantly changes COVID-19 transmission in (sub) tropical cities of Brazil," *Science of the Total Environment*, p. 138862, 2020.

[21] M. Jahangiri, M. Jahangiri, and M. Najafgholipour, "The sensitivity and specificity analyses of ambient temperature and population size on the transmission rate of the novel coronavirus (COVID-19) in different provinces of Iran," *Science of The Total Environment*, p. 138872, 2020.

[22] M. Şahin, "Impact of weather on COVID-19 pandemic in Turkey," *Science of The Total Environment*, p. 138810, 2020.

[23] M. F. F. Sobral, G. B. Duarte, A. I. G. da Penha Sobral, M. L. M. Marinho, and A. de Souza Melo, "Association between climate variables and global transmission oF SARS-CoV-2," *Science of The Total Environment*, vol. 729, p. 138997, 2020.

[24] M. F. Bashir, B. Ma, B. Komal, M. A. Bashir, D. Tan, M. Bashir, *et al.*, "Correlation between climate indicators and COVID-19 pandemic in New York, USA," *Science of The Total Environment*, p. 138835, 2020.

[25] Z. Yongjian, X. Jingu, H. Fengming, and C. Liqing, "Association between short-term

exposure to air pollution and COVID-19 infection: Evidence from China," *Science of the total environment*, p. 138704, 2020.

[26] Y. Ogen, "Assessing nitrogen dioxide (NO2) levels as a contributing factor to the coronavirus (COVID-19) fatality rate," *Science of The Total Environment*, p. 138605, 2020.

[27] W. B. Grant, H. Lahore, S. L. McDonnell, C. A. Baggerly, C. B. French, J. L. Aliano, and H. P. Bhattoa, "Evidence that vitamin D supplementation could reduce risk of influenza and COVID-19 infections and deaths," *Nutrients*, vol. 12, no. 4, p. 988, 2020.

[28] M. Alipio, "Vitamin D Supplementation Could Possibly Improve Clinical Outcomes of Patients Infected with Coronavirus-2019 (COVID-19)," *Available at SSRN 3571484*, 2020.

[29] A. Mendy, S. Apewokin, A. A. Wells, and A. L. Morrow, "Factors associated with hospitalization and disease severity in a racially and ethnically diverse population of COVID-19 patients," *MedRxiv*, 2020.

[30] W. Dietz and C. Santos-Burgoa, "Obesity and its Implications for COVID-19 Mortality," *Obesity*, vol. 28, no. 6, pp. 1005–1005, 2020.

[31] R. Kassir, "Risk of COVID-19 for patients with obesity," *Obesity Reviews*, vol. 21, no. 6, 2020.

[32] J. Lighter, M. Phillips, S. Hochman, S. Sterling, D. Johnson, F. Francois, and A. Stachel, "Obesity in patients younger than 60 years is a risk factor for Covid-19 hospital admission," *Clinical Infectious Diseases*, 2020.

[33] W. H. Organization, "WHO Coronavirus Disease (COVID-19) Dashboard." https://apps.who.int/gho/data/node.home. Accessed: 2020-10-16.

[34] O. W. in Data, "Excess mortality during the Coronavirus pandemic (COVID-19)." https://ourworldindata.org/excess-mortality-covid. Accessed: 2020-10-19.

[35] O. W. in Data, "Coronavirus Pandemic (COVID-19) – the data." https://ourworldindata.org/coronavirus-data. Accessed: 2020-11-04.

[36] C. for Systems Science and E. C. at Johns Hopkins University, "COVID-19 Data Repository." https://github.com/CSSEGISandData/COVID-19. Accessed: 2020-10-28.

[37] E. C. for Disease Prevention and Control, "COVID-19 situation update worldwide." https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases. Accessed: 2020-10-28.

[38] C. for Disease Control and Prevention, "Excess Deaths Associated with COVID-19." https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm. Accessed: 2020-11-04.

[39] C. e. I. d. C. Ministerio de Ciencia, Tecnología, "Datos-COVID19." https://github.com/MinCiencia/Datos-COVID19. Accessed: 2020-11-03.

[40] T. W. Bank, "World Development Indicators." http://wdi.worldbank.org/table. Accessed: 2020-09-22.

[41] O. world in Data, "Research and data to make progress against the world's largest problems." https://ourworldindata.org/. Accessed: 2020-10-19.

[42] Wikipedia, "The free encyclopedia that anyone can edit." https://en.wikipedia.org/wiki/Main_Page. Accessed: 2020-10-07.

[43] C. I. Agency, "The World Factbook." https://www.cia.gov/llibrary/publications/the-world-factbook/fields/343.html. Accessed: 2020-10-02.

[44] W. H. Organization, "Global Health Observatory data repository." https://apps.who.int/gho/data/node.home. Accessed: 2020-10-16.

[45] A. of Economic Complexity by GrowthLab at Harvard University, "Country Product Complexity Rankings." https://atlas.cid.harvard.edu/rankings/. Accessed: 2020-10-02.

[46] F. I. B. Federation, "FIBA World Ranking." http://www.fiba.basketball/rankingwomen, http://www.fiba.basketball/rankingmen. Accessed: 2020-10-07.

[47] F. I. de Football Association (FIFA), "Women's Ranking | Men's Ranking ." https://www.fifa.com/fifa-world-ranking/ranking-table/women/ | https://www.fifa.com/fifa-world-ranking/ranking-table/men/. Accessed: 2020-10-07.

[48] F. I. de Volleyball, "FIVB Senior World Ranking." https://www.fivb.com/en/volleyball/rankings/seniorworldrankingwomen | https://www.fivb.com/en/volleyball/rankings/seniorworldrankingmen. Accessed: 2020-10-07.

[49] UNdata, "Relative Humidity." http://data.un.org/Data.aspx?d=CLINO&f=ElementCode%3A11. Accessed: 2020-10-16.

[50] G. Developers, "Dataset Publishing Language." https://developers.google.com/public-data/docs/canonical/countries_csv. Accessed: 2020-10-20.

[51] W. W. W. C. (W3C), "Resource Description Framework (RDF)." https://www.w3.org/RDF/. Accessed: 2021-01-18.

[52] W. W. W. C. (W3C), "Resource Description Framework (RDF): Concepts and Abstract Syntax." https://www.w3.org/TR/rdf-concepts/. Accessed: 2021-01-18.

[53] W. W. W. C. (W3C), "The RDF Data Cube Vocabulary." https://www.w3.org/TR/vocab-data-cube/. Accessed: 2021-01-19.

[54] I. The Wikimedia Foundation, "Wikimedia Foundation." https://wikimediafoundation.org/. Accessed: 2021-03-07.

[55] Wikidata, "Wikibase." https://www.wikiba.se/. Accessed: 2021-03-07.

[56] W. Foundation, "Wikidata." https://www.wikidata.org. Accessed: 2021-02-06.

[57] W. W. W. C. (W3C), "Turtle - Terse RDF Triple Language." https://www.w3.org/TeamSubmission/turtle/. Accessed: 2021-02-21.

[58] W. W. W. C. (W3C), "Resource Description Framework (RDF): Concepts and Abstract Syntax." https://www.w3.org/TR/rdf-concepts/. Accessed: 2021-02-21.

[59] W. W. W. C. (W3C), "RDF Test Cases: N-Triples." https://www.w3.org/TR/rdf-testcases/#ntriples. Accessed: 2021-02-21.

[60] Wikidata, "Iceland ." https://www.wikidata.org/wiki/Q189. Accessed: 2021-01-19.

[61] W. W. W. C. (W3C), "SPARQL 1.1 Query Language." https://www.w3.org/TR/sparql11-query/. Accessed: 2021-01-20.

[62] C. Aschwanden, "Not Even Scientists Can Easily Explain P-values," *FiveThirtyEight*, 2015-11-24.

[63] SciPy, "scipy.stats.pearsonr." https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html. Accessed: 2021-02-03.

[64] D. Zwillinger and S. Kokoska, *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

[65] I. The Wikimedia Foundation, "Cum hoc ergo propter hoc." https://es.wikipedia.org/wiki/Cum_hoc_ergo_propter_hoc. Accessed: 2021-04-04.

[66] C. for Systems Science and J. H. U. Engineering, "COVID-19 Dash-board by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University(JHU)." https://www.arcgis.com/apps/opsdashboard/. Accessed: 2020-07-06.

[67] Statista, "Estimated number of universities worldwide as of July 2020, by country." https://www.statista.com/statistics/918403/number-of-universities-worldwide-by-country/. Accessed: 2021-04-04.

[68] T. W. Bank, " World Development Indicators: Size of the economy." ://w-di.worldbank.org/table/WV.1.

[69] IBM, "IBM Global COVID-19 Statistics." https://accelerator.weather.com/. Accessed: 2020-07-06.

[70] O. W. in Data, "Coronavirus Pandemic (COVID-19)." https://ourworldindata.org/coronavirus/. Accessed: 2020-07-06.

[71] J. L. A, "COVID-19 en tu comuna." https://covid19entucomuna.cl/. Accessed: 2020-07-10.

[72] G. Groups, "Tarql: SPARQL for Tables: Turn CSV into RDF using SPARQL syntax." http://tarql.github.io/. Accessed: 2021-02-10.

[73] T. A. S. Foundation, "Apache Jena Fuseki." https://jena.apache.org/documentation/fuseki2/. Accessed: 2021-02-10.

[74] T. Pallets, "Flask." https://flask.palletsprojects.com/en/1.1.x/. Accessed: 2021-03-24.

[75] A. Wood, "HTML5 Basics For Everyone Tired Of Reading About Deprecated Code." https://html.com/html5/. Accessed: 2021-03-24.

[76] Pluralsight, "JavaScript." https://www.javascript.com/. Accessed: 2021-03-24.

[77] AnyChart, "AnyChart - JS charts." https://www.anychart.com/. Accessed: 2021-03-24.

[78] N. Md, Y. Wai, N. Ibrahim, Z. Rashid, N. Mustafa, H. Hamid, M. Latif, S. Er, L. Yik, K. Alhasa, *et al.*, "Particulate matter (pm2. 5) as a potential sars-cov-2 carrier," 2020.

Anexo A

Hipótesis y sus posibles variables relacionadas

A.1. Agentes Climáticos

Tabla A.1: Agentes Climáticos y sus posibles variables relacionadas

Hypothesis	**Associated variable**
Winter	Latitude, Longitude
Humidity	Humidity: Mean Daily Maximum Value
Air Pollution	Ambient PM2.5 air pollution
	Carbon dioxide emissions
	Total greenhouse gas emissions
	Methane emissions
	Nitrous oxide emissions
	Other greenhouse gas emissions
	Number of deaths by risk factor: Indoor air pollution
	Number of deaths by risk factor: Air pollution (outdoor & indoor)
	Number of deaths by risk factor: Outdoor air pollution
	Carbon dioxide emissions
Weather	Precipitation, Mean Monthly Value
Temperature	Highest temperatures ever recorded
	Lowest temperatures ever recorded
	Average temperature, monthly
	Population affected by droughts, floods, and extreme temperatures
Pollution	Local pollution damage
	Vehicles per capita
Precipitation	Average annual precipitation

A.2. Determinantes de Salud

Tabla A.2: Determinantes de Salud y sus posibles variables relacionadas

Hypothesis	**Associated variable**
Vitamin D	Sunshine duration
Obesity	Obesity rate
	Prevalence of overweight children
	Number of deaths by risk factor: Obesity
	Body mass index
Hypertension	Mean systolic blood pressure trends, age-standardized (mmHg)
	Number of deaths by risk factor: High blood pressure
	Raised blood pressure (SBP 140 OR DBP 90), crude (%)
	Raised blood pressure (SBP 140 OR DBP 90), age-standardized (%)
Cardiovascular Disease	Number of deaths attributed to non-communicable diseases: Cardiovascular diseases
	Deaths - Cardiovascular diseases
	Mortality between age 30 and exact age 70 from cardiovascular diseases, cancer, diabetes or chronic respiratory diseases
Age	Median Age
	Population age composition
Cancer	Cancer frequency
	Mortality between age 30 and exact age 70 from cardiovascular diseases, cancer, diabetes or chronic respiratory diseases
Smoking	Prevalence of smoking
	Number of deaths by risk factor: Smoking
	Number of deaths by risk factor: Secondhand smoke
Gender	Population Gender
Blood Type	Blood type distribution
Ethnic group	Ethnic and cultural diversity level
Diabetes	Prevalence of diabetes
	Mortality between age 30 and exact age 70 from cardiovascular diseases, cancer, diabetes or chronic respiratory diseases
	Number of deaths by risk factor: High blood sugar
	Number of deaths attributed to non-communicable diseases: Diabetes mellitus
Respiratory diseases	Acute respiratory infection (ARI) prevalence
	Number of deaths attributed to non-communicable diseases: Chronic obstructive pulmonary disease
	Deaths - Chronic respiratory diseases
	Mortality between age 30 and exact age 70 from cardiovascular diseases, cancer, diabetes or chronic respiratory diseases
	Acute respiratory infection (ARI) treatment
Depression	Suicide rates
	Deaths - Mental and substance use disorders
Dementia	Deaths from dementia-related diseases, by age
	Deaths - Alzheimer disease and other dementias

A.3. Factores Económicos

Tabla A.3: Factores Económicos y sus posibles variables relacionadas

Hypothesis	**Associated variable**
Income	Gross national income
	Percentage share of income or consumption
	Annualized growth in mean consumption or income per capita
	Mean consumption or income per capita
	Adjusted net national income Growth
Poverty	Population below national poverty lines
	Urban population living in slums
	Wealth per adult
	International poverty lines
Imports/Exports	Growth of merchandise trade: Export volume
	Growth of merchandise trade: Import volume
	Growth of merchandise trade: Export value
	Growth of merchandise trade: Import value
	Growth of merchandise trade: Net barter terms of trade index
	Exports: To low- and middle-income economies, Within region
	Exports: To low- and middle-income economies, Outside region
	Exports: To high-income economies
	Imports: To low- and middle-income economies, Within region
	Imports: To low- and middle-income economies, Outside region
	Imports: To high-income economies
	Goods and services Exports
	Goods and services Imports
	Exports of goods and services
	Imports of goods and services

A.4. Otros Factores

Tabla A.4: Otros Factores y sus posibles variables relacionadas

Hypothesis	**Associated variable**
Density Poblation	Population density
	Population in urban agglomerations of more than 1 million
	Population in the largest city
Women leadership	Women in parliaments
	Women Business and the Law Index Score
International flights	International tourists: Inbound
	International tourists: Outbound
Domestic violence	Intentional homicides
	Divorce statistics
FIFA ranking	FIFA World Ranking: Men
	FIFA World Ranking: Women
Mobility	Railways
	Ports
	Air
Government	Government: Revenue
	Government: Expense
	Government: Net investment in nonfinancial assets
	Government: Net lending (+) / net borrowing (-)
	Government: Net acquisition of financial assets
	Government: Net incurrence of liabilities
	Government: Debt payments
	Government: Interest payments
	Government Expenditure: Goods and services
	Government Expenditure: Compensation of employees
	Government Expenditure: Interest payments
	Government Expenditure: Subsidies and other transfers
	Government Expenditure: Other expense
	Government Revenues: Taxes on income, profits and capital gains
	Government Revenues: Taxes on goods and services
	Government Revenues: Taxes on international trade
	Government Revenues: Other taxes
	Government Revenues: Social contributions
	Government Revenues: Grants and other revenue
Social Media	Internet use
	Internet application
	Mobile (Cellular only) download speed
Individuals per house/ family members who work	Contributing family workers
	Vulnerable employment
	Dependency ratio

Anexo B

Variables agregadas

Tabla B.1: Variables extras que fueron agregadas al servidor.

Category associated	**Variable**
Climate Factor	Savings: Local pollution damage
Economic Factor	Adjusted net savings
	Expenditures for RD
	External health expenditure
	Health expenditure
	Wage and salary workers
	Youth unemployment
Health Determinant	Adult mortality rate
	Infant mortality rate
	Life expectancy at birth
	Life expectancy birth total
	Maternal mortality ratio
	Mortality caused by road traffic injury
	Mortality: cardiovascular diseases, cancer, diabetes or chronic respiratory diseases
	Neonatal mortality rate
	Number of deaths attributed to non-communicable diseases: Cardiovascular diseases
	Number of deaths attributed to non-communicable diseases: Chronic obstructive pulmonary disease
	Number of deaths by risk factor: Alcohol use
	Number of deaths by risk factor: Drug use
	Number of deaths by risk factor: No access to handwashing facility
	People using safely managed drinking water services
	People using safely managed sanitation services
	Population by gender
	Prevalence of child malnutrition
	Under-five mortality rate
Other Factor	Access to electricity
	Account ownership at a financial institution
	Completeness of birth registration
	Completeness of death registration
	Health workers
	Nationally protected terrestrial and marine areas
	Population
	Renewable energy consumption
	Specialist surgical workforce
	Surface area
	Transport: Air
	Transport: Ports
	Transport: Railways
	Women who were first married by age 18