Primera entrega
TFM – COVID-19

INESDI

Programa: Máster en Business Intelligence &
Data Management (Online)

Autores: Grupo 6 COVID - Grupo I (Proyecto Inesdi)

Tutor: Pier Paolo Rossi

- Amaia Miranda Ulloa
- Fabián Ascheri Aguerre
- José Chavarría Montero
- Juan Carlos Valcuende Aláez
- Patricia Peña Torres

05 de abril de 2023

INTRODUCCIÓN

El presente documento corresponde a la primera entrega de nuestro proyecto final de máster relativo al Covid-19

Está estructurado en esta breve introducción seguida de la definición del proyecto del proyecto y análisis de viabilidad del mismo, aquí estará por una parte el análisis interno en el que hablamos de los conocimientos y/o familiaridad de los integrantes con la temática así como un repaso de la petición del cliente, aquí podrán encontrarse técnicas como el mapa de empatía, post-motorola, team alignment map y los respectivos insights; así mismo cerramos este apartado con el análisis externo en el cual hacemos una breve contextualización de la problemática del COVID-19. Por último, en esta sección habrá la definición detallada del proyecto que también contendrá los objetivos tanto generales como específicos del proyecto a desarrollar.

Seguidamente tenemos lo relativo al perfilado y limpieza de las bases de datos preseleccionadas, así como la estructura de las bases de datos.

A continuación, nos encontramos con la planificación en la cual entraremos en temas como son los recursos económicos, materiales, humanos, tiempos estimados, cronograma y alcance esperado del proyecto.

Para finalizar tenemos un apartado de conclusiones en el que reflexionamos sobre el trabajo desarrollado por el equipo en este entregable, seguido de las referencias bibliográficas en formato APA y los anexos donde hacemos una breve aproximación a todas las bases de datos facilitadas por INESDI.

Para la gestión y trabajo interno nos estamos basando en metodologías AGILE y de equipos de alto rendimiento con relaciones horizontales entre los miembros del equipo.

Sin nada más que decir damos inicio al informe del primer entregable.

ÍNDICE

2. DEFINICIÓN DEL PROYECTO Y ANÁLISIS DE LA VIABILIDAD.

Se nos hizo la petición de desarrollar un proyecto final de máster utilizando los diferentes datasets y fuentes de datos proporcionadas sobre el COVID ser capaz de realizar un cuadro de mando que nos permita encontrar insights en los datos relacionados con el COVID.

Se nos dan también tres objetivos a cumplir:

1. Integrar las diferentes fuentes en un repositorio (base de datos), realizar perfilado de los datos, reglas de calidad y realizar el modelo de datos. Análisis exploratorio de todos los datos de COVID escogidos de manera que se puedan encontrar insights en los datos

2. Analizar y comparar los datos de diferentes zonas y épocas del año para encontrar diferencias y explicar a qué se deben estas diferencias.

3. Opcional: construir un modelo predictivo para predecir la siguiente ola de COVID o en su defecto explicar qué variables son las más representativas para explicar la mortalidad o la inafectabilidad.

2.1. DESCRIPCIÓN DETALLADA DE LA EMPRESA.

Este proyecto fue propuesto por INESDI por lo cual no tiene vinculación con ninguna empresa.

Acotamos que nuestro target serán profesionales del mundo de la salud.

2.1.1. ANÁLISIS INTERNO

Dentro del equipo hay varios miembros que ya tienen experiencia previa en ámbitos de la salud con Covid-19.

Un miembro del equipo es profesional de salud en activo por tanto conoce de primera mano las necesidades que tiene su colectivo al respecto de recibir información sobre el Covid-19 y tiene un punto de vista desde dentro de nuestro potencial cliente

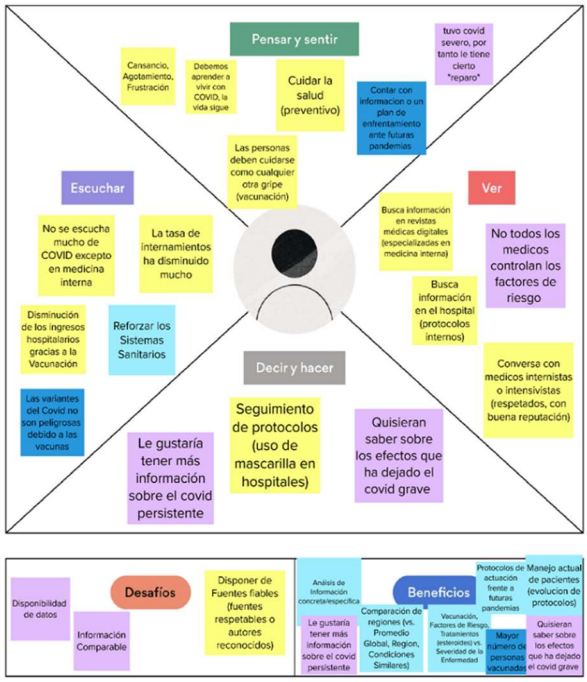
Otro tiene experiencia en el ámbito administrativo hospitalario concretamente en el área de sistemas de información así que creemos que puede aportar una visión muy clara sobre cómo hemos de presentar esta información para que sea accesible e inteligible.

Una miembro del equipo tiene conocimiento de enfermedades respiratorias debido a su experiencia laboral en salud digital enfocada en el diagnóstico y monitorización de patologías respiratorias crónicas y agudas, acostumbra a trabajar de forma frecuente con información y estadísticas de patologías respiratorias incluyendo el Covid-19.

2.1.1.1. MAPA DE EMPATÍA

Un mapa de empatía es una herramienta utilizada en el contexto empresarial para comprender mejor las necesidades y motivaciones de los clientes. Consiste en un diagrama que muestra los

pensamientos, sentimientos, comportamientos y necesidades de los clientes en relación con un producto o servicio en particular. Al mapear estos elementos, las empresas pueden crear una imagen más clara y detallada de su cliente ideal, lo que les permite desarrollar productos y servicios que satisfagan mejor sus necesidades y deseos.



2.1.1.1.2. INSIGHTS

Respecto a la <u>conformación del equipo de trabajo</u>

- Retos
  - Los distintos husos horarios hacen difícil que todos los integrantes estemos disponibles al mismo tiempo
  - El uso de tecnología es imprescindible para el éxito colectivo, lo que requiere que todos aprendamos a utilizar nuevas herramientas, y rápido
  - No logramos conectar con una de nuestras integrantes (Maureen Fleming) - seguimos en ello
- Acciones
  - Generación de confianza: el primer día cada integrante realizó una introducción de sí mismo, mediante la cual compartió información personal para conectar con el resto
  - Coordinación:
    - Se utilizó Doodle para identificar los mejores espacios de tiempo para conectar como equipo (se han realizado dos sesiones de grupo)
    - Se creó un grupo de WhatsApp para mantenernos en comunicación constante
  - Herramientas: se definieron las principales herramientas a utilizar
    - Comunicación: MS Teams y WhatsApp
    - Conocimiento: MS Teams (One Drive)
    - Interactivas: Mural
    - Tiempo: Doodle

- Planificación de Tareas: MS Teams

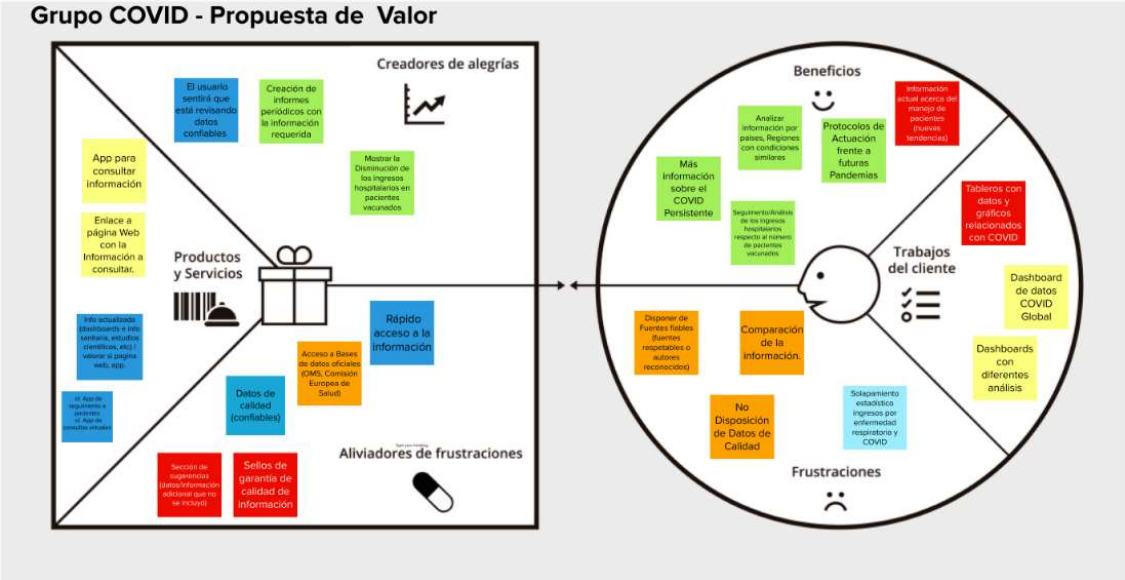
- Entregables de la Semana
  - Se definió un potencial usuario, esto para saber más o menos a quién entrevistar (en grupo)
  - Se realizaron las entrevistas (de manera individual)
  - Se consolidó toda la información mediante el mapa de empatía (en grupo)
  - Se asignaron el resto de las tareas para preparar el entregable final: portada, mapa de empatía y perspectivas

Respecto al usuario

- El usuario potencial del producto final del proyecto es el profesional del área de salud (médico, farmacéutico, enfermero, entre otros)
- El usuario asigna un alto valor a la calidad de la información (fuentes fidedignas, respetables, confiables), por lo que los datos deben provenir de entidades serias, especializadas, con reputación dentro del sector salud
  - También se puede considerar contar con sponsors del producto (por ejemplo, colegios de profesionales de salud, ministerios de salud, etc.)
- El usuario busca información actualizada (la ciencia avanza a un ritmo que hace que los tratamientos o protocolos queden obsoletos)
- El usuario busca correlacionar distintas variables a resultados observados:
  - Características de Paciente vs. Severidad
  - Factores de Riesgo vs. Severidad
  - Tratamientos vs. Severidad
  - % Vacunación vs. Mortalidad
- El usuario muestra un interés especial por los efectos a largo plazo del COVID

2.1.1.2. PROPUESTA DE VALOR

A continuación, se presenta el canvas de la propuesta de valor realizado por el equipo de trabajo. El mismo se encuentra disponible en el siguiente mural en caso de que se requiera una revisión detallada del mismo (la calidad de la imagen no es la mejor).

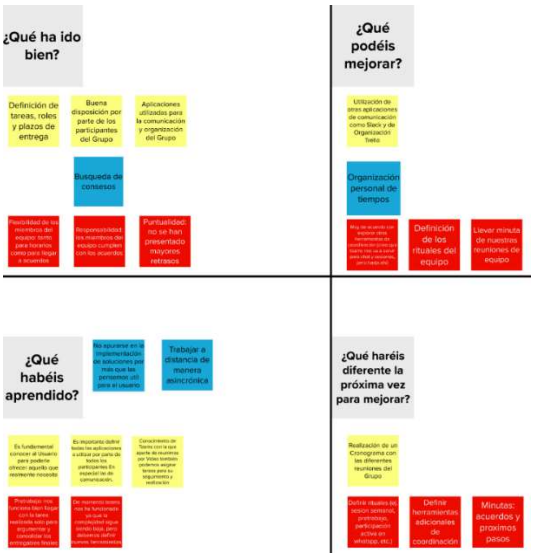


2.1.1.2.1. INSIGHTS

- Se realizó la propuesta de valor mediante la utilización del canvas, alineando los resultados del mapa de empatía realizado la semana pasada.
- El contexto, alineamiento y valor que brinda conocer los beneficios y las frustraciones de los potenciales usuarios antes de trabajar en una propuesta de valor es algo que nos impresionó. Especialmente al identificar los creadores de alegrías y los aliviadores de frustraciones como parte de la propuesta.
- A partir del entendimiento de lo que nuestro potencial cliente hace (trabajos del cliente), los beneficios que espera y las frustraciones que tiene, se propone un conjunto de dashboards que brinde información precisa (confiable), relevante, actual y de fácil y rápido acceso.
  - Confiable
    - Fuentes fidedignas y respetables, que gocen de buena reputación
    - Quizá "sponsors" respetables que promuevan la solución
  - Relevante
    - Alineado con los intereses de nuestros usuarios
    - Estadísticas de relación entre factores vs. efectos (por ejemplo, severidad de la enfermedad vs. % vacunación en el país)
    - Benchmark (por ejemplo, resultados país vs. promedio global, regional o países con condiciones similares
  - Actual
    - Fuentes e información actual (por ejemplo, nuevas tendencias globales o datos relacionados a nuevos tratamientos)
  - Fácil y Rápido Acceso
    - Toda la información consolidada en formatos que faciliten el análisis, la comparación y la toma de decisiones por parte de los usuarios

2.1.1.3. POST-MOTOROLA

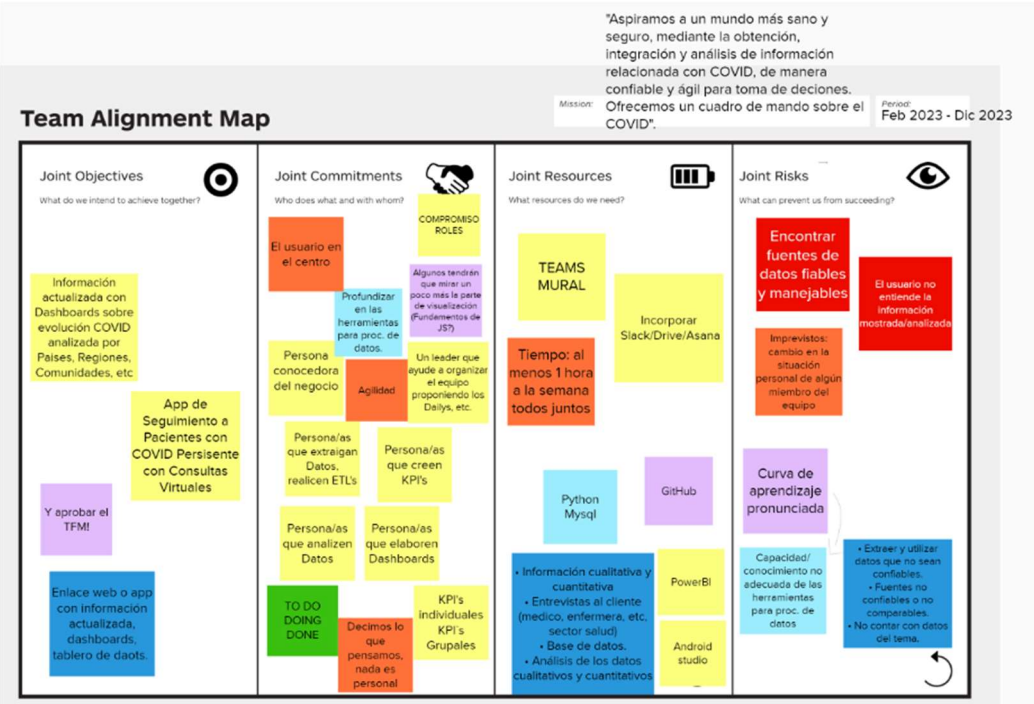
De la misma manera, se presenta el ejercicio post-motorola realizado por el equipo de trabajo. Al igual que el ejercicio anterior, el mismo se encuentra disponible en el siguiente mural en caso de que se requiera una revisión detallada del mismo.



2.1.1.3.1. INSIGHTS

- Se concluye que el equipo ha logrado coordinar de manera exitosa todas las tareas y esfuerzos que se debían ejecutar a la fecha, debido principalmente a la apertura, al compromiso y a la buena voluntad de todos por colaborar.
- El equipo debe trabajar en una mejor definición de "rutinas" y herramientas de colaboración (hasta el momento se ha utilizado MS Teams). Esto debido a que, con el paso del tiempo, la complejidad de las tareas incrementará, por lo que será vital contar con una excelente coordinación y visibilidad de acuerdos, tareas y compromisos.
    - Se agendó una sesión semanal, todos los Lunes a las 19h España para tratar todos los temas relacionados a entregables y avance del TFM
    - Se ha trabajado muy bien mediante un esquema de pre-trabajo, en el que los miembros del equipo realizan los ejercicios de manera individual previo a la sesión de equipo, en la cual finalmente debatimos, tomamos decisiones y consolidamos.
    - Se definirán las nuevas herramientas a utilizar para mejorar la coordinación de las tareas (por ejemplo, Trello, Slack, entre otras)
    - Se llevará una minuta de nuestras sesiones en un formato sencillo y práctico (acuerdos, próximos pasos y responsables)
- El equipo ha aprendido que no debe apresurarse a ejecutar tareas sin pensar de manera estratégica primero, asegurándose de que todos los miembros entienden el resultado esperado y existe un alineamiento.

2.1.1.4. TEAM ALIGNMENT MAP



El mismo puede seleccionando el siguiente [enlace](#).

2.1.1.4.1. INSIGHTS

Misión

- "Aspiramos a un mundo más sano y seguro, mediante la obtención, integración y análisis de información relacionada con COVID, de manera confiable y ágil para toma de decisiones. Ofrecemos un cuadro de mando sobre el COVID".

Línea de Tiempo

- Febrero 2023 - Diciembre 2023

Objetivos

- Aprobar el Trabajo Final del Máster
- Crear un cuadro de mando con información relevante, confiable y actual relacionada con COVID

Compromisos

- Agilidad: utilización de la metodología SCRUM
- Mantendremos al usuario en el centro (comprobación luego de cada sprint)
- Reuniones semanales, al menos una hora para conectar, siendo conscientes de que debemos ser flexibles (en algunas ocasiones no todos los miembros del equipo podrán participar)
- Comunicación: nuestro equipo contará con un ambiente seguro, en donde cada integrante podrá decir lo que piensa y como se siente, siempre dentro de un marco de respeto
- Hemos conversado sobre la figura del tutor y hemos hecho el primer contacto con el mismo (también por LinkedIn)

Recursos

- Herramientas del equipo:
    - Comunicación: Teams (quizá lo reemplazaremos por Slack), Whatsapp
    - Repositorio: Teams (lo reemplazaremos por OneDrive)
    - Coordinación de Tareas: adoptaremos un gestor de tareas más visual, esto debido a que la versión gratuita de Teams cuenta con una herramienta muy básica (lo reemplazaremos por Asana o Trello)
    - Colaboración: Mural
- Herramientas para el desarrollo del proyecto:
    - Python
    - MySQL
    - PowerBI
- Tiempo de cada integrante del grupo

Riesgos

- Curva de aprendizaje respecto a herramientas tecnológicas para el desarrollo del proyecto
- Alcance del proyecto: hay una gran cantidad de datos alrededor de esta enfermedad, por lo que debemos definir muy bien el alcance del proyecto (acotarlo)
- Calidad y confiabilidad de la información (fuentes respetables, información comparable)
- Imprevistos relacionados con la situación personal de cada integrante del equipo (nuevo trabajo, muerte de algún familiar, vacaciones, etc.)

2.1.2. ANÁLISIS EXTERNO

Sólo en España, las patologías respiratorias han producido 260.000 ingresos hospitalarios en un año hasta 2018, los datos de 2020 indican que se ingresaron 528.554 mil personas por enfermedad respiratoria y 137.623 específicamente por Covid-19, esto sugiere que se han infraregistrado los ingresos por covid-19. En el caso de Cataluña hablaríamos de 104.198 ingresos por enfermedades respiratorias en general y 31.114 por Covid-19 (INE, 2023).

La pandemia de COVID-19 es una enfermedad infecciosa causada por el virus SARS-CoV-2. Se cree que el virus se originó en Wuhan, China, a fines de 2019 y se propagó rápidamente por todo el mundo, llegando a ser declarada una pandemia por la Organización Mundial de la Salud en marzo de 2020 (OMS, 2023). Según la OMS la pandemia por Covid-19 ha dejado 674 millones de afectados en todo el mundo con el consiguiente colapso de los sistemas de emergencia y limitaciones en la realización de visitas presenciales, así mismo se contabilizan más de 6 millones de fallecidos. (Center for Systems Science and Engineering, 2023)

Los estudios señalan la necesidad de que los profesionales de la salud se mantengan actualizados al respecto de temas relativos al Covid-19 (Steinbach et al, 2020), así mismo se destaca la importancia de hacer estudios colaborativos de perspectiva internacional para un mejor abordaje de la pandemia. (Dolan & Mackey, 2020); otros artículos destacarán la importancia de que la información sea accesible, precisa y actualizada.

2.1.3. EXPLICACIÓN DETALLADA DEL PROYECTO

Nos encontramos con que desde INESDI se nos facilitan múltiples carpetas, relativas al Covid-19, serán relativas a la situación mundial, en Estados Unidos, Corea del Sur y España con especial interés en la Comunidad de Madrid.

Después de realizar una revisión preliminar de las bases de datos[1] seleccionamos cuáles veíamos más adecuadas en función de las necesidades del cliente.

Hasta el día de hoy, los avances actuales llegan a la selección de las bases de datos, así como limpieza y perfilado de las mismas.

Se prevé más adelante que con los resultados finales se presenten una serie de dashboards que plasmen la información.

A continuación, el listado completo de las mismas

Carpeta COVID ESPAÑA

- ✓ casos_diag_ccaadecl.csv
- ✓ casos_hosp_uci_def_sexo_edad_provres.csv
- ✓ casos_tecnica_ccaa.csv
- ✓ casos_tecnica_provincia.csv

[1] Presente en el apartado de anexos

- ✓ metadata_diag_ccaa_decl_prov_edad_sexo.pdf
- ✓ metadata_tecnica_ccaa_prov_res.pdf

Carpeta COVID MADRID

- ✓ covid19_tia_muni_y_distritos.csv
- ✓ covid19_tia_muni_y_distritos_s.csv
- ✓ municipios_y_distritos_madrid.zip
- ✓ municipios_y_distritos_madrid_20.zip

Carpeta datos COVID Corea del Sur

- ✓ Case.csv
- ✓ PatientInfo.csv
- ✓ Policy.csv
- ✓ Region.csv
- ✓ SearchTrend.csv
- ✓ SeoulFloating.csv
- ✓ Time.csv
- ✓ TimeAge.csv
- ✓ TimeGender.csv
- ✓ TimeProvince.csv
- ✓ Weather.csv

Carpeta datos COVID España

- ✓ ccaa_camas_uci_2017.csv
- ✓ ccaa_covid19_altas_long.csv
- ✓ ccaa_covid19_casos_long.csv
- ✓ ccaa_covid19_fallecidos_long.csv
- ✓ ccaa_covid19_hospitalizados_long.csv
- ✓ ccaa_covid19_mascarillas.csv
- ✓ ccaa_covid19_uci_long.csv
- ✓ nacional_covid19.csv
- ✓ nacional_covid19_rango_edad.csv

Carpeta datos COVID Mundial

- ✓ covid_19_data.csv
- ✓ time_series_covid_19_confirmed.csv
- ✓ time_series_covid_19_confirmed_US.csv
- ✓ time_series_covid_19_deaths.csv
- ✓ time_series_covid_19_deaths_US.csv
- ✓ time_series_covid_19_recovered.csv

Carpeta datos COVID USA

- ✓ us_counties_covid19_daily.csv
- ✓ us_covid19_daily.csv
- ✓ us_states_covid19_daily.csv

Así mismo se nos facilitará acceso a otros datos covid-19 entre los cuales habrá repositorios de WHO y del Centro de Investigación John Hopkins a través de GitHub.

2.3.1. DEFINICIÓN DE OBJETIVOS

Una vez analizada la petición del cliente podemos hablar de los siguientes objetivos generales

1. Análisis y explicación de variación en función de zona y época.
2. Realización de modelos predictivos
3. Desarrollo del storytelling mediante iteración y/o pivotación constante.

A continuación, una repetición de los objetivos generales esta vez con sus respectivos objetivos específicos

1. Análisis y explicación de variación en función de zona y época.
   1.1. Integración, perfilación y modelado de base de datos.
   1.2. Análisis exploratorio de datos de Covid-19
2. Realización de modelos predictivos
   2.1. Regresión temporal para predecir siguientes olas de Covid-19.
   2.2. Análisis de regresión logística múltiple para identificar las variables más representativas en cuanto a la mortalidad.
   2.3. Análisis de regresión lineal múltiple para identificar las variables más representativas de las tasas de contagio.
3. Desarrollo del storytelling mediante iteración y/o pivotación constante.
   3.1. Traducción estadística: adaptar los análisis a información inteligible por el usuario.
   3.2. Visualización de datos en dashboards.
   3.3. Pruebas piloto.
   3.4. Pruebas de concepto.

3. BASES DE DATOS PRESELECCIONADAS

Finalmente hemos decidido preliminarmente que se utilizarán las siguientes bases de datos y se ha hecho así mismo la limpieza y perfilado de las mismas.

Nombre del Fichero	Informacion General	Variables	Rango de Fechas	Enlace	KPI
Data on COVID-19 vaccination in the EU/EEA	Contiene información de vacunacion para paises europeos.	YearWeekISO; ReportingCountry; Denominator; NumberDosesReceived; NumberDosesExported; FirstDose; FirstDoseRefused; SecondDose; DoseAdditional1; UnknownDose; Region; Population; Vaccine; Population;	2021 - today	https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea	Vacunacion

		DoseAdditional2; DoseAdditional3			
Data on SARS-CoV-2 variants in the EU/EEA	Contiene informacion de las secuencias y variantes del covid.	country; country_code; year_week; source; new_cases; number_sequenced; percent_cases_sequenced; valid_denominator; variant; number_detections_variant; number_sequenced_known_variant; percent_variant	2020 - today	https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea	Casos Notificados por Variantes de Covid
Data on 14-day notification rate of new COVID-19 cases and deaths	Ratio de casos (1 en 100k) o de muertes (1 en 1MM), asi como acumulativos en paises europeos.	country; country_code; continent; population; indicator; weekly_count; year_week; rate_14_day; cumulative_count; source; note	2020 - today	https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19	Casos Notificados Muertes
Data on hospital and ICU admission rates and current occupancy for COVID-19	Contiene informacion de ingresos en hospital y UCI por dia o semana para paises europeos.	country; indicator; date; year_week; value; source; url	2020 - today	https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-covid-19	Hospitalizaciones UCI
Data on testing for COVID-19 by week and country	Informacion acerca de tests realizados en paises europeos.	country; country_code; year_week; level; region; region_name; new_cases; tests_done; population; testing_rate; positivity_rate; testing_data_source	2020 - today	https://www.ecdc.europa.eu/en/publications-data/covid-19-testing	Pruebas Realizadas
Data on the 14-day age-specific notification rate of new COVID-19 cases	Ratio de casos (1 en 100k) en paises europeos.	country; country_code; year_week; age_group; new_cases; population; rate_14_day_per_100k; source	2020 - today	https://www.ecdc.europa.eu/en/publications-data/covid-19-data-14-day-age-notification-rate-new-cases	Casos Notificados (ratio)
csse_covid_19_daily_reports	This folder contains daily case reports. All timestamps are in UTC (GMT+0).	FIPS: US only. Federal Information Processing Standards code that uniquely identifies counties within the USA. Admin2: County name. US only. Province_State: Province, state or dependency name. Country_Region: Country, region or sovereignty name. The names of locations included on the Website correspond with the official designations used by the U.S. Department of State. Last Update: MM/DD/YYYY HH:mm:ss (24 hour format, in UTC). Lat and Long_: Dot	2021 - today	https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports	Casos Notificados US

		locations on the dashboard. All points (except for Australia) shown on the map are based on geographic centroids, and are not representative of a specific address, building or any location at a spatial scale finer than a province/state. Australian dots are located at the centroid of the largest city in each state. Confirmed: Counts include confirmed and probable (where reported). Deaths: Counts include confirmed and probable (where reported). Recovered: Recovered cases are estimates based on local media reports, and state and local reporting when available, and therefore may be substantially lower than the true number. US state-level recovered cases are from COVID Tracking Project. We stopped to maintain the recovered cases (see Issue #3464 and Issue #4465). Active: Active cases = total cases - total recovered - total deaths. This value is for reference only after we stopped to report the recovered cases (see Issue #4465) Incident_Rate: Incidence Rate = cases per 100,000 persons. Case_Fatality_Ratio (%): Case-Fatality Ratio (%) = Number recorded deaths / Number cases. All cases, deaths, and recoveries reported are based on the date of initial report. Exceptions to this are noted in the "Data Modification" and "Retrospective reporting of (probable) cases and deaths" subsections below.			
csse_covid_19_daily_reports_us	This table contains an aggregation of each USA State level data.	Province_State - The name of the State within the USA. Country_Region - The name of the Country (US). Last_Update - The most recent date the file was pushed. Lat - Latitude. Long_ - Longitude. Confirmed - Aggregated case count for the state. Deaths - Aggregated death toll for the state.	2021 - today	https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports_us	Casos Notificados US by State

		Recovered - Aggregated Recovered case count for the state. Active - Aggregated confirmed cases that have not been resolved (Active cases = total cases - total recovered - total deaths). FIPS - Federal Information Processing Standards code that uniquely identifies counties within the USA. Incident_Rate - cases per 100,000 persons. Total_Test_Results - Total number of people who have been tested. People_Hospitalized - Total number of people hospitalized. (Nullified on Aug 31, see Issue #3083) Case_Fatality_Ratio - Number recorded deaths * 100/ Number confirmed cases. UID - Unique Identifier for each row entry. ISO3 - Officially assigned country code identifiers. Testing_Rate - Total test results per 100,000 persons. The "total test results" are equal to "Total test results (Positive + Negative)" from COVID Tracking Project. Hospitalization_Rate - US Hospitalization Rate (%): = Total number hospitalized / Number cases. The "Total number hospitalized" is the "Hospitalized – Cumulative" count from COVID Tracking Project. The "hospitalization rate" and "Total number hospitalized" is only presented for those states which provide cumulative hospital data. (Nullified on Aug 31, see Issue #3083)			
Daily cases and deaths by date reported to WHO	Casos y muertes diarias por país.	Date_reported, Country_code, Country, WHO_region, New_cases, Cumulative_cases, New_deaths, Cumulative_deaths	2020 - Today	https://covid19.who.int/WHO-COVID-19-global-data.csv	Casos Notificados Muertes
Latest reported counts of cases and deaths	Ultimo reporte de casos y muertes notificadas a la OMS.	Name, WHO_region, Cases - cumulative total, Cases - cumulative total per 100000 population, Cases - newly reported in last 7 days, Cases - newly reported in last 7 days per 100000 population, Cases - newly reported in last 24 hours, Deaths - cumulative	2020 - Today	https://covid19.who.int/WHO-COVID-19-global-table-data.csv	Casos Notificados Muertes

		total, Deaths - cumulative total per 100000 population, Deaths - newly reported in last 7 days, Deaths - newly reported in last 7 days per 100000 population, Deaths - newly reported in last 24 hours			
Vaccination data	Inforación de vacunación por país.	COUNTRY, ISO3, WHO_REGION, DATA_SOURCE, DATE_UPDATED, TOTAL_VACCINATIONS, PERSONS_VACCINATED_1PLUS_DOSE, TOTAL_VACCINATIONS_PER100, PERSONS_VACCINATED_1PLUS_DOSE_PER100, PERSONS_FULLY_VACCINATED, PERSONS_FULLY_VACCINATED_PER100, VACCINES_USED, FIRST_VACCINE_DATE, NUMBER_VACCINES_TYPES_USED, PERSONS_BOOSTER_ADD_DOSE, PERSONS_BOOSTER_ADD_DOSE_PER100	May 2021 - Today	https://covid19.who.int/who-data/vaccination-data.csv	Vacunacion
Vaccination metadata	Informacion general de vacunas utilizadas en cada país.	ISO3, VACCINE_NAME, PRODUCT_NAME, COMPANY_NAME, FIRST_VACCINE_DATE, AUTHORIZATION_DATE, START_DATE, END_DATE, COMMENT, DATA_SOURCE	May 2021 - Today	https://covid19.who.int/who-data/vaccination-metadata.csv	N/A

3.1. LIMPIEZA DE LAS BASES DE DATOS

Data on COVID-19 vaccination in the EU/EEA

```
import pandas as pd
import numpy as np
import datetime as dt

df_datos = pd.read_csv('data.csv')
df_datos.info()
df_datos.dtypes
df_datos.duplicated()
```

```
df_datos['YearWeekISO'] = pd.to_datetime(df_datos['YearWeekISO'] + '-1', format='%Y-
W%W-%w')
df_datos
df_datos['ReportingCountry'] = df_datos['ReportingCountry'].astype('category'); ; ; ;
assert df_datos['ReportingCountry'].dtype == 'category'
df_datos['Region']   =   df_datos['Region'].astype('category')   ;   ;   ;   ;   assert
df_datos['Region'].dtype == 'category'
df_datos['TargetGroup'] = df_datos['TargetGroup'].astype('category') ; ; ; ; assert
df_datos['TargetGroup'].dtype == 'category'
df_datos['Vaccine']   =   df_datos['Vaccine'].astype('category')   ;   ;   ;   ;   assert
df_datos['Vaccine'].dtype == 'category'
df_datos['FirstDoseRefused'].fillna(0, inplace=True)

df_datos.info()
df_datos.dtypes
df_datos.duplicated()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 659339 entries, 0 to 659338
Data columns (total 16 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   YearWeekISO          659339 non-null  datetime64[ns]
 1   ReportingCountry     659339 non-null  category
 2   Denominator          348937 non-null  float64
 3   NumberDosesReceived  82880 non-null   float64
 4   NumberDosesExported  88208 non-null   float64
 5   FirstDose            659339 non-null  int64
 6   FirstDoseRefused     659339 non-null  float64
 7   SecondDose           659339 non-null  int64
 8   DoseAdditional1      659339 non-null  int64
 9   DoseAdditional2      659339 non-null  int64
 10  DoseAdditional3      659339 non-null  int64
 11  UnknownDose          659339 non-null  int64
 12  Region               659339 non-null  category
 13  TargetGroup          659339 non-null  category
 14  Vaccine              659339 non-null  category
 15  Population           659339 non-null  int64
dtypes: category(4), datetime64[ns](1), float64(4), int64(7)
memory usage: 62.9 MB
```

Out[31]:

```
0         False
1         False
2         False
3         False
4         False
          ...
659334    False
659335    False
659336    False
659337    False
659338    False
Length: 659339, dtype: bool
```

Data on SARS-CoV-2 variants in the EU/EEA

```
import pandas as pd
import numpy as np
import datetime as dt

df_datos1 = pd.read_csv('data1.csv')
df_datos1.info()
df_datos1.dtypes
df_datos1.duplicated()
df_datos1

df_datos1['country'] = df_datos1['country'].astype('category') ; ; ; ; assert df_datos1['count
ry'].dtype == 'category'
df_datos1['country_code'] = df_datos1['country_code'].astype('category') ; ; ; ; assert df_dat
os1['country_code'].dtype == 'category'
df_datos1['year_week'] = pd.to_datetime(df_datos1['year_week'], format='%Y-W%U', errors='coerc
e') df_datos1['year_week'] = pd.to_datetime(df_datos1['year_week'], format='%Y-%m-%d', errors=
'coerce')
```

```
df_datos1['source'] = df_datos1['source'].astype('category') ; ; ; ; assert df_datos1['source'
].dtype == 'category'
df_datos1['new_cases'].fillna(0, inplace=True)
df_datos1['variant'] = df_datos1['variant'].astype('category') ; ; ; ; assert df_datos1['varia
nt'].dtype == 'category'
df_datos1['percent_variant'].fillna(0, inplace=True)
df_datos1['percent_variant'] = df_datos1['percent_variant'].astype('string') ; ; ; ; assert df
_datos1['percent_variant'].dtype == 'string'

df_datos1.info()
df_datos1.dtypes
df_datos1.duplicated()
df_datos1
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 120884 entries, 0 to 120883
Data columns (total 12 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   country                         120884 non-null  category
 1   country_code                    120884 non-null  category
 2   year_week                       0 non-null       datetime64[ns]
 3   source                          120884 non-null  category
 4   new_cases                       120884 non-null  float64
 5   number_sequenced                120884 non-null  int64
 6   percent_cases_sequenced         119314 non-null  float64
 7   valid_denominator               120884 non-null  bool
 8   variant                         120884 non-null  category
 9   number_detections_variant       120884 non-null  int64
 10  number_sequenced_known_variant  120884 non-null  int64
 11  percent_variant                 120884 non-null  string
dtypes: bool(1), category(4), datetime64[ns](1), float64(2), int64(3), string(1)
memory usage: 7.0 MB
```

Data on 14-day notification rate of new COVID-19 cases and deaths

```
import pandas as pd
import numpy as np
import datetime as dt

df_datos3 = pd.read_csv('data3.csv')
df_datos3.info()
df_datos3.dtypes
df_datos3.duplicated()

df_datos3['country'] = df_datos3['country'].astype('category'); ; ; ; assert df_datos3['countr
y'].dtype == 'category'
df_datos3['country_code'] = df_datos3['country_code'].astype('category'); ; ; ; assert df_dato
s3['country_code'].dtype == 'category'
df_datos3['continent'] = df_datos3['continent'].astype('category') ; ; ; ; assert df_datos3['c
ontinent'].dtype == 'category'
df_datos3['indicator'] = df_datos3['indicator'].astype('category'); ; ; ; assert df_datos3['in
dicator'].dtype == 'category'
df_datos3.dropna(subset=['weekly_count'], inplace=True)
df_datos3['year_week'] = pd.to_datetime(df_datos3['year_week'], format='%Y-W%U', errors='coerc
e')
df_datos3['year_week'] = pd.to_datetime(df_datos3['year_week'], format='%Y-%m-%d', errors='coe
rce')
df_datos3.dropna(subset=['rate_14_day'], inplace=True)
df_datos3.dropna(subset=['cumulative_count'], inplace=True)
df_datos3['source'] = df_datos3['source'].astype('category') ; ; ; ; assert df_datos3['source'
].dtype == 'category'

df_datos3.dropna(subset=['note'], inplace=True)
df_datos3.info()
df_datos3.dtypes
df_datos3.duplicated()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 0 entries
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   country           0 non-null      category
 1   country_code      0 non-null      category
 2   continent         0 non-null      category
```

```
 3   population        0 non-null      int64
 4   indicator         0 non-null      category
 5   weekly_count      0 non-null      float64
 6   year_week         0 non-null      datetime64[ns]
 7   rate_14_day       0 non-null      float64
 8   cumulative_count  0 non-null      float64
 9   source            0 non-null      category
 10  note              0 non-null      float64
dtypes: category(5), datetime64[ns](1), float64(4), int64(1)
memory usage: 2.9 KB
Out[18]:
Series([], dtype: bool)
```

Data on hospital and ICU admission rates and current occupancy for COVID-19

```
import pandas as pd
import numpy as np
import datetime as dt

df_datos2 = pd.read_csv('data2.csv')
df_datos2.info()
df_datos2.dtypes
df_datos2.duplicated()

df_datos2['country'] = df_datos2['country'].astype('category'); ; ; ; assert df_datos2['countr
y'].dtype == 'category'
df_datos2['indicator'] = df_datos2['indicator'].astype('category'); ; ; ; assert df_datos2['in
dicator'].dtype == 'category'
df_datos2['date'] = pd.to_datetime(df_datos2['date'])
df_datos2['year_week'] = pd.to_datetime(df_datos2['year_week'], format='%Y-W%U', errors='coerc
e')
df_datos2['year_week'] = pd.to_datetime(df_datos2['year_week'], format='%Y-%m-%d', errors='coe
rce')
df_datos2['source'] = df_datos2['source'].astype('category'); ; ; ; assert df_datos2['source']
.dtype == 'category'
df_datos2['url'] = df_datos2['url'].astype('category'); ; ; ; assert df_datos2['url'].dtype ==
'category'
df_datos2.dropna(subset=['url'], inplace=True)

df_datos2.info()
df_datos2.dtypes
df_datos2.duplicated()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 0 entries
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   country    0 non-null      category
 1   indicator  0 non-null      category
 2   date       0 non-null      datetime64[ns]
 3   year_week  0 non-null      datetime64[ns]
 4   value      0 non-null      float64
 5   source     0 non-null      category
 6   url        0 non-null      category
dtypes: category(4), datetime64[ns](2), float64(1)
memory usage: 1.4 KB
```

Data on testing for COVID-19 by week and country

```
import pandas as pd
import numpy as np
import datetime as dt

df_datos5 = pd.read_csv('data5.csv')
df_datos5.info()
df_datos5.dtypes

df_datos5['country'] = df_datos5['country'].astype('category'); assert
df_datos5['country'].dtype == 'category'
df_datos5['country_code'] = df_datos5['country_code'].astype('category'); assert
df_datos5['country_code'].dtype == 'category'
df_datos5['year_week'] = pd.to_datetime(df_datos5['year_week'], format='%Y-W%U',
errors='coerce')
```

```
df_datos5['year_week'] = pd.to_datetime(df_datos5['year_week'], format='%Y-%m-%d',
errors='coerce')
df_datos5['age_group'] = df_datos5['age_group'].astype('category'); assert
df_datos5['age_group'].dtype == 'category'
df_datos5.dropna(subset=['new_cases'], inplace=True)
df_datos5.dropna(subset=['new_cases'], inplace=True)
df_datos5.dropna(subset=['rate_14_day_per_100k'], inplace=True)
df_datos5.dropna(subset=['source'], inplace=True)
df_datos5['source'] = df_datos5['source'].astype('category'); assert df_datos5['source'].dtype
== 'category'

df_datos5.info()
df_datos5.dtypes
       <class 'pandas.core.frame.DataFrame'>
       Int64Index: 27806 entries, 9 to 29405
       Data columns (total 8 columns):
        #   Column                Non-Null Count  Dtype
       ---  ------                --------------  -----
        0   country               27806 non-null  category
        1   country code          27806 non-null  category
        2   year week             0 non-null      datetime64[ns]
        3   age group             27806 non-null  category
        4   new_cases             27806 non-null  float64
        5   population            27806 non-null  int64
        6   rate 14 day per 100k  27806 non-null  float64
        7   source                27806 non-null  category
       dtypes: category(4), datetime64[ns](1), float64(2), int64(1)
       memory usage: 1.2 MB
```

csse_covid_19_daily_reports

```
import pandas as pd
import os
import zipfile
ruta = 'C:/Users/Patricia/Downloads/COVID-19-master/csse_covid_19_data/csse_covid_19_daily_rep
orts'
archivos_csv = []
if ruta.endswith('.zip'):
    with zipfile.ZipFile(ruta) as archivo_zip:
        for archivo in archivo_zip.namelist():
            if archivo.endswith('.csv'):
                with archivo_zip.open(archivo) as archivo_csv:
                    archivos_csv.append(pd.read_csv(archivo_csv))
else:
    for archivo in os.listdir(ruta):
        if archivo.endswith('.csv'):
            archivos_csv.append(pd.read_csv(os.path.join(ruta, archivo)))
datos_combinados = pd.concat(archivos_csv, ignore_index=True)
datos_combinados.to_csv('datos_combinados_csse_covid_19_daily_reports.csv', index=False)


df_datos6 = df_datos6.astype({'Admin2':'category', 'Province_State':'category', 'Country_Regio
n':'category','Combined_Key':'category', 'Province/State':'category', 'Country/Region':'catego
ry', 'Last Update':'category'})

df_datos6.dropna(inplace=True)

df_datos6['Last_Update'] = df_datos6['Last_Update'].astype('object')

df_datos6.info()
df_datos6.dtypes
       <class 'pandas.core.frame.DataFrame'>
       Int64Index: 0 entries
       Data columns (total 20 columns):
        #   Column               Non-Null Count  Dtype
       ---  ------               --------------  -----
        0   FIPS                 0 non-null      float64
        1   Province_State       0 non-null      category
        2   Country_Region       0 non-null      category
        3   Last_Update          0 non-null      object
        4   Lat                  0 non-null      float64
        5   Long                 0 non-null      float64
        6   Confirmed            0 non-null      float64
        7   Deaths               0 non-null      float64
```

```
 8   Recovered            0 non-null      float64
 9   Active               0 non-null      float64
 10  Combined Key         0 non-null      category
 11  Incident Rate        0 non-null      float64
 12  Case_Fatality_Ratio  0 non-null      float64
 13  Province/State       0 non-null      category
 14  Country/Region       0 non-null      category
 15  Last Update          0 non-null      category
 16  Latitude             0 non-null      float64
 17  Longitude            0 non-null      float64
 18  Incidence_Rate       0 non-null      float64
 19  Case-Fatality_Ratio  0 non-null      float64
dtypes: category(6), float64(13), object(1)
memory usage: 291.0+ KB
```

csse_covid_19_daily_reports_us

```
import pandas as pd
import os
import zipfile
ruta = 'C:/Users/Patricia/Downloads/COVID-19-master/csse_covid_19_data/csse_covid_19_daily_rep
orts_us'
archivos_csv = []
if ruta.endswith('.zip'):
    with zipfile.ZipFile(ruta) as archivo_zip:
        for archivo in archivo_zip.namelist():
            if archivo.endswith('.csv'):
                with archivo_zip.open(archivo) as archivo_csv:
                    archivos_csv.append(pd.read_csv(archivo_csv))
else:
    for archivo in os.listdir(ruta):
        if archivo.endswith('.csv'):
            archivos_csv.append(pd.read_csv(os.path.join(ruta, archivo)))
datos_combinados = pd.concat(archivos_csv, ignore_index=True)
datos_combinados.to_csv('datos_combinados_csse_covid_19_daily_reports_us.csv', index=False)

import pandas as pd
import numpy as np
import datetime as dt

df_datos7 = pd.read_csv('datos_combinados_csse_covid_19_daily_reports_us.csv')
df_datos7.info()
df_datos7.dtypes

for col in df_datos7.select_dtypes(include=['object']):
    if 'Rate' in col:
        df_datos7[col] = df_datos7[col].astype('Int64').fillna(0)
    else:
        df_datos7[col] = df_datos7[col].astype('category')

df_datos7.info()
df_datos7.dtypes
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 61614 entries, 0 to 61613
       Data columns (total 21 columns):
        #   Column               Non-Null Count  Dtype
       ---  ------               --------------  -----
        0   Province_State       61614 non-null  category
        1   Country_Region       61614 non-null  category
        2   Last Update          61595 non-null  category
        3   Lat                  59472 non-null  float64
        4   Long                 59472 non-null  float64
        5   Confirmed            61614 non-null  int64
        6   Deaths               61614 non-null  int64
        7   Recovered            15122 non-null  float64
        8   Active               15122 non-null  float64
        9   FIPS                 61595 non-null  float64
        10  Incident_Rate        59472 non-null  float64
        11  Total_Test_Results   36637 non-null  float64
        12  People Hospitalized  5129 non-null   float64
        13  Case Fatality Ratio  49027 non-null  float64
        14  UID                  61614 non-null  float64
        15  ISO3                 61614 non-null  category
        16  Testing_Rate         45921 non-null  float64
        17  Hospitalization_Rate 5129 non-null   float64
```

```
 18  Date                  51754 non-null  category
 19  People Tested         11816 non-null  float64
 20  Mortality Rate        12027 non-null  float64
dtypes: category(5), float64(14), int64(2)
memory usage: 8.0 MB
```

WHO_Global_Data_1

```
import pandas as pd
import numpy as np
import datetime as dt

df_who1 = pd.read_csv('WHO_Global_Data_1.csv')
df_who1.info()
df_who1.dtypes

df_who1['Date_reported'] = pd.to_datetime(df_who1['Date_reported'], format='%Y-%m-%d')
df_who1 = df_who1.astype({'Date_reported': 'datetime64[ns]', 'Country_code': 'category', 'Coun
try': 'category', 'WHO_region': 'category'})

df_who1.info()
df_who1.dtypes
df_who1.duplicated()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 280134 entries, 0 to 280133
Data columns (total 8 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Date_reported      280134 non-null  datetime64[ns]
 1   Country_code       278952 non-null  category
 2   Country            280134 non-null  category
 3   WHO_region         280134 non-null  category
 4   New_cases          280134 non-null  int64
 5   Cumulative_cases   280134 non-null  int64
 6   New_deaths         280134 non-null  int64
 7   Cumulative_deaths  280134 non-null  int64
dtypes: category(3), datetime64[ns](1), int64(4)
memory usage: 12.0 MB
```

WHO_Global_Data_2

```
df_who2 = pd.read_csv('WHO_Global_Data_2.csv')
df_who2.info()
df_who2.dtypes

df_who2['WHO Region'] = df_who2['WHO Region'].astype('category'); assert df_who2['WHO Region']
.dtype == 'category'
df_who2['Name'] = df_who2['Name'].astype('category'); assert df_who2['Name'].dtype == 'categor
y'
df_who2 = df_who2.dropna()

df_who2.info()
df_who2.dtypes
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 0 entries
Data columns (total 12 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   Name                                                  0 non-null      category
 1   WHO Region                                            0 non-null      category
 2   Cases - cumulative total                              0 non-null      float64
 3   Cases - cumulative total per 100000 population        0 non-null      int64
 4   Cases - newly reported in last 7 days                 0 non-null      float64
 5   Cases - newly reported in last 7 days per 100000 population  0 non-null  int64
 6   Cases - newly reported in last 24 hours               0 non-null      int64
 7   Deaths - cumulative total                             0 non-null      float64
 8   Deaths - cumulative total per 100000 population       0 non-null      int64
 9   Deaths - newly reported in last 7 days                0 non-null      float64
 10  Deaths - newly reported in last 7 days per 100000 population  0 non-null  int64
 11  Deaths - newly reported in last 24 hours              0 non-null      float64
dtypes: category(2), float64(5), int64(5)
memory usage: 464.0+ bytes
```

WHO_Vaccination_Data_1.csv

```
df_who3 = pd.read_csv('WHO_Vaccination_Data_1.csv')
df_who3.info()
df_who3.dtypes

df_who3[['COUNTRY', 'ISO3', 'WHO_REGION', 'DATA_SOURCE', 'VACCINES_USED']] = df_who3[['COUNTRY
', 'ISO3', 'WHO_REGION', 'DATA_SOURCE', 'VACCINES_USED']].astype('category')
df_who3[['DATE_UPDATED', 'FIRST_VACCINE_DATE']] = df_who3[['DATE_UPDATED', 'FIRST_VACCINE_DATE
']].apply(pd.to_datetime)

df_who3.info()
df_who3.dtypes
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 229 entries, 0 to 228
       Data columns (total 16 columns):
        #   Column                                Non-Null Count  Dtype
       ---  ------                                --------------  -----
        0   COUNTRY                               229 non-null    category
        1   ISO3                                  229 non-null    category
        2   WHO_REGION                            229 non-null    category
        3   DATA_SOURCE                           229 non-null    category
        4   DATE_UPDATED                          229 non-null    datetime64[ns]
        5   TOTAL_VACCINATIONS                    228 non-null    float64
        6   PERSONS_VACCINATED_1PLUS_DOSE         229 non-null    int64
        7   TOTAL_VACCINATIONS_PER100             227 non-null    float64
        8   PERSONS_VACCINATED_1PLUS_DOSE_PER100  229 non-null    float64
        9   PERSONS_FULLY_VACCINATED              229 non-null    int64
        10  PERSONS_FULLY_VACCINATED_PER100       229 non-null    float64
        11  VACCINES_USED                         225 non-null    category
        12  FIRST_VACCINE_DATE                    207 non-null    datetime64[ns]
        13  NUMBER_VACCINES_TYPES_USED            225 non-null    float64
        14  PERSONS_BOOSTER_ADD_DOSE              210 non-null    float64
        15  PERSONS_BOOSTER_ADD_DOSE_PER100       210 non-null    float64
       dtypes: category(5), datetime64[ns](2), float64(7), int64(2)
       memory usage: 46.7 KB
```

WHO_Vaccination_Data_1.csv

```
df_who4 = pd.read_csv('WHO_Vaccination_Data_2.csv')
df_who4.info()
df_who4.dtypes

df_who4['AUTHORIZATION_DATE'] = pd.to_datetime(df_who4['AUTHORIZATION_DATE'])
df_who4['START_DATE'] = pd.to_datetime(df_who4['START_DATE'])
df_who4['END_DATE'] = pd.to_datetime(df_who4['END_DATE'])
df_who4[['ISO3', 'VACCINE_NAME', 'PRODUCT_NAME', 'COMPANY_NAME', 'DATA_SOURCE']] = df_who4[['I
SO3', 'VACCINE_NAME', 'PRODUCT_NAME', 'COMPANY_NAME', 'DATA_SOURCE']].astype('category')
df_who4.dropna(inplace=True)

df_who4.info()
df_who4.dtypes
       <class 'pandas.core.frame.DataFrame'>
       Int64Index: 0 entries
       Data columns (total 9 columns):
        #   Column              Non-Null Count  Dtype
       ---  ------              --------------  -----
        0   ISO3                0 non-null      category
        1   VACCINE_NAME        0 non-null      category
        2   PRODUCT_NAME        0 non-null      category
        3   COMPANY_NAME        0 non-null      category
        4   AUTHORIZATION_DATE  0 non-null      datetime64[ns]
        5   START_DATE          0 non-null      datetime64[ns]
        6   END_DATE            0 non-null      datetime64[ns]
        7   COMMENT             0 non-null      float64
        8   DATA_SOURCE         0 non-null      category
       dtypes: category(5), datetime64[ns](3), float64(1)
       memory usage: 13.4 KB
```

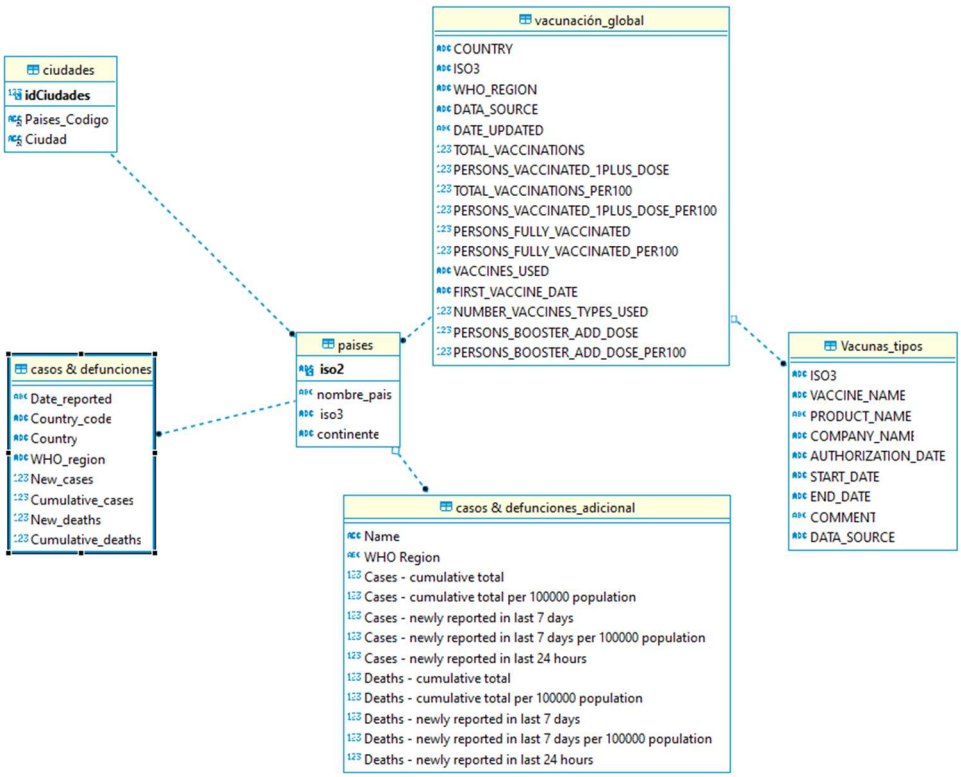
3.2. ESTRUCTURA DE LAS BASES DE DATOS

Una vez realizada la exploración inicial de los datos de COVID-19, y entendiendo la cantidad y calidad de la información de la que disponemos, se tomó la decisión de iniciar con un modelo de base de datos simple, que facilite la normalización y consolidación de los datos más generales, eso sí, con un alcance global.

La intención del equipo es seguir construyendo alrededor de esta base de datos inicial, una estructura cada vez más robusta y compleja, dependiendo de la información que se decida incluir en los siguientes "sprints".

De momento, la estructura cuenta con 3 tablas de hechos y 3 tablas de dimensiones:

- Tablas de Hechos
    - Casos y Defunciones Históricos (casos & defunciones)
    - Casos y Defunciones Actual (casos & defunciones_adicional)
    - Vacunación (vacunacion_global)
- Tablas de Dimensiones
    - Países
    - Ciudades
    - Vacunas (vacunas_tipo)



4. PLANIFICACIÓN

4.1. ESTIMACIÓN DE LOS RECURSOS ECONÓMICOS

Debido a que es un proyecto de TFM no disponemos de recursos económicos para llevar a cabo el mismo y se usarían herramientas gratuitas o de las que dispusiéramos con anterioridad.

4.2. ESTIMACIÓN DE LOS RECURSOS MATERIALES.

Entre los recursos materiales contaríamos con que todos los integrantes disponemos de ordenadores y conexión a internet, así mismo se prevé la utilización de Python mediante Jupyter Notebooks y para SQL utilizar DBeaver, DB Browser o PhpMyAdmin.

4.3. ESTIMACIÓN DE LOS RECURSOS HUMANOS.

Contamos con cinco integrantes, que serían:

Amaia Miranda, economista, con experiencia en seguimiento y monitoreo de proyectos.

Fabián Ascheri, enfermero con experiencia principalmente en asistencia al paciente.

Jose Chavarria, ingeniero industrial que actualmente se desempeña como director de cuentas por pagar.

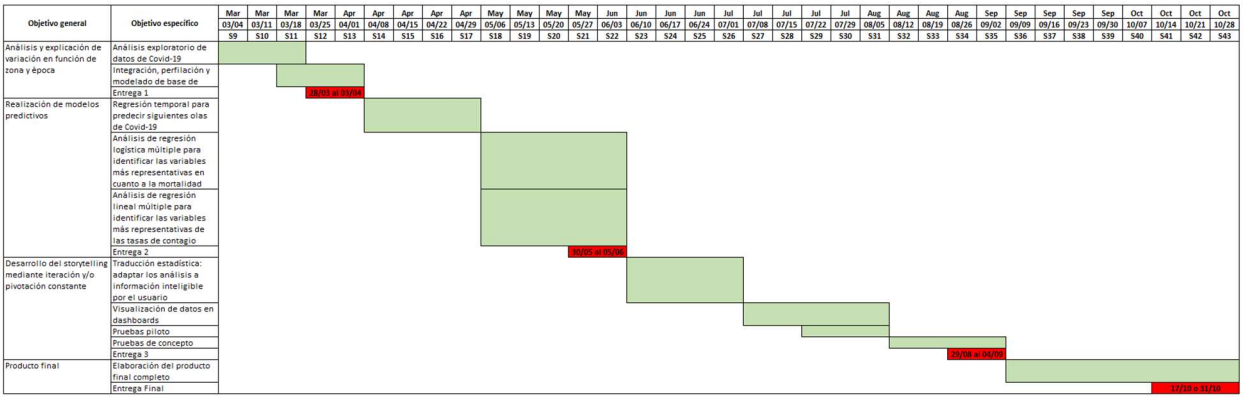
Juan Carlos Valcuende, administrativo de sistemas de la información en el Hospital de Bellvitge con amplia experiencia en el ámbito de datos hospitalarios.

Patricia Peña, socióloga con experiencia en emprendimiento social en salud digital en patologías respiratorias y análisis de datos en el ámbito de la investigación.

4.4. ESTIMACIÓN DE LOS RECURSOS TIEMPO.

Se ha propuesto realizar dos reuniones semanalmente, también habrá trabajo autónomo y asincrónico que se distribuirá cada integrante en función de su disponibilidad.

4.5 CRONOGRAMA DEL PROYECTO



4.6. DEFINICIÓN DEL ALCANCE DEL PROYECTO.

Entre los entregables estarían una serie de dashboards y sus respectivos resúmenes en los que se visualice de forma sencilla, atractiva y amigable la información recabada; para esto será importante también un trabajo de storytelling así como de conversar frecuentemente con el cliente para no perder el enfoque.

Así mismo el ideal sería que estos fuesen utilizados así que nos gustaría poder llevar a cabo las pruebas piloto y de concepto en alguna institución de salud.

5. CONCLUSIONES

En general hemos tenido diversas dificultades como por ejemplo,

Problemas con archivos de datos que se encontraban desagrupados

Archivos demasiado grandes y que nuestros procesadores no podían afrontar

Problemas para descargar algunos repositorios

Dificultad para hacer nuestro cronograma de trabajo

Finalmente, los problemas los fuimos resolviendo entre todos repartiéndonos las tareas y siempre alguien se ponía a investigar, lo que nos causó mayores problemas lo conversamos con nuestro tutor y nos dió ideas de cómo abordarlo.

En general este entregable nos ha servido para poder aclarar mejor nuestras ideas, aclarar más el rumbo del trabajo, así como empezar a tener un contacto más directo con el mundo de los datos, también ha sido la primera vez de tener que preparar un entregable grande juntos como equipo debido a que en otras asignaturas hasta ahora siempre eran trabajos relativamente superficiales que en una o dos sesiones ya teníamos cerrados

6. BIBLIOGRAFÍA

Center for Systems Science and Engineering. (n.d.). COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. GitHub. https://github.com/CSSEGISandData/COVID-19

Dolan, B., & Mackey, T. K. (2020). COVID-19 and the Need for Action on Health Information. The Lancet Digital Health, 2, e188-e189. https://doi.org/10.1016/S2589-7500(20)30084-4

Gupta, A. K. (2020, 20 de mayo). The Importance of Data in the Fight Against COVID-19. Forbes. Recuperado de https://www.forbes.com/sites/forbestechcouncil/2020/05/20/the-importance-of-data-in-the-fight-against-covid-19/?sh=497ccde22a9f

Instituto Nacional de Estadística. (2023). Altas hospitalarias según el sexo, el diagnóstico principal, la provincia, Comunidad y Ciudad autónoma de hospitalización. Recuperado el 24 de febrero de 2023

Organización Mundial de la Salud (OMS). (2020, 11 de marzo). Rolling updates on coronavirus disease (COVID-19). Recuperado el 21 de marzo de 2023 de https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229

Siu Hing Lo, C. P. Choy, & T. H. Lam. (2020). COVID-19 and the Need for Action on Health Information. Journal of Racial and Ethnic Health Disparities, 7, 937-941. https://doi.org/10.1007/s40615-020-00849-y

Steinbach, W. J., Scheetz, M. H., Postelnick, T. C., Reed, D. R., Pickering, K. S., Smith, A. M., & Landon, E. (2020). COVID-19: Knowledge is Power for Frontline Healthcare Workers. The Lancet, 8, e484-e485. https://doi.org/10.1016/S2213-2600(20)30165-1

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Niu, P. (2020). A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine, 382(8), 727-733.

7. ANEXOS

Perfilado de Datos

ccaa_covid19_mascarillas

```
In [1]: import pandas as pd
        import numpy as np
        import datetime
        import pandas_profiling
```

```
In [2]: import chardet

        with open('ccaa_covid19_mascarillas.csv', 'rb') as f:
            result = chardet.detect(f.read())

        print(result['encoding'])
```

```
ISO-8859-1
```

```
In [3]: df1 = pd.read_csv('ccaa_covid19_mascarillas.csv', encoding='ISO-8859-1')
```

```
In [4]: df1.head()
```

Out[4]:

	fecha	cod_ine	CCAA	mascarillas_acumulado_desde_2020-03-10
0	2020-03-22	1.0	ANDALUCÍA	444083
1	2020-03-22	2.0	ARAGÓN	72455
2	2020-03-22	3.0	ASTURIAS	60229
3	2020-03-22	4.0	BALEARES	49476
4	2020-03-22	5.0	CANARIAS	124165

Alerts

`cod_ine` is highly overall correlated with `CCAA`	High correlation
`CCAA` is highly overall correlated with `cod_ine`	High correlation
`cod_ine` has 1 (1.7%) missing values	Missing
`CCAA` is uniformly distributed	Uniform
`mascarillas_acumulado_desde_2020-03-10` has unique values	Unique

Este fichero contiene número de mascarillas acumuladas diarias por Comunidad Autónoma.

Presenta las variables: fecha , cod_ine , CCAA , total

ccaa_covid19_uci_long

```
import chardet

with open('ccaa_covid19_uci_long.csv', 'rb') as f:
    result = chardet.detect(f.read())

print(result['encoding'])
```

```
ISO-8859-1
```

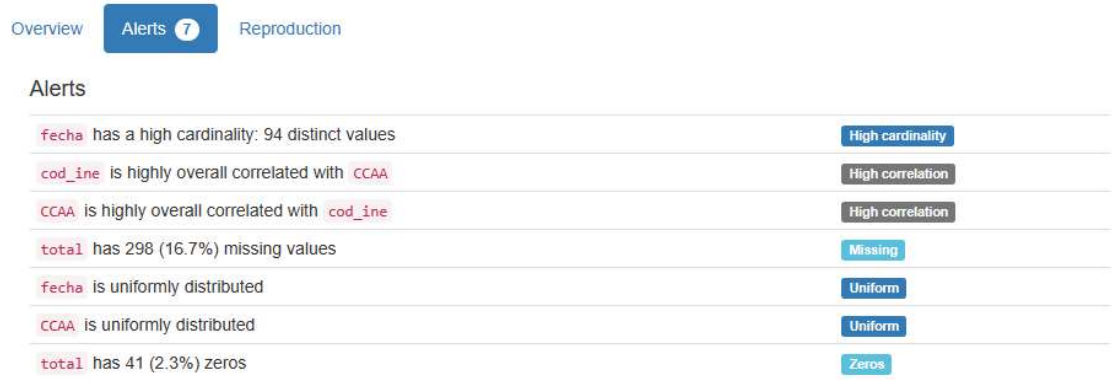
```
df2 = pd.read_csv('ccaa_covid19_uci_long.csv', encoding='ISO-8859-1')
```

```
df2.head()
```

	fecha	cod_ine	CCAA	total
0	2020-02-21	1	Andalucía	NaN
1	2020-02-22	1	Andalucía	NaN
2	2020-02-23	1	Andalucía	NaN
3	2020-02-24	1	Andalucía	NaN
4	2020-02-25	1	Andalucía	NaN

```
df2.profile_report()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```



Este fichero contiene número de ingresados en UCI diarios por Comunidad Autónoma.

Presenta las variables: cod_ine , CCAA , Públicos , Privados , Total

ccaa_covid19_hospitalizados_long

```
df3 = pd.read_csv('ccaa_covid19_hospitalizados_long.csv', encoding='ISO-8859-1')
```

```
df3.head()
```

	fecha	cod_ine	CCAA	total
0	2020-02-21	1	Andalucía	NaN
1	2020-02-22	1	Andalucía	NaN
2	2020-02-23	1	Andalucía	NaN
3	2020-02-24	1	Andalucía	NaN
4	2020-02-25	1	Andalucía	NaN

```
df3.profile_report()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

Este fichero contiene número de hospitalizados diarios por Comunidad Autónoma.

Presenta las variables: fecha , cod_ine , CCAA , total

ccaa_covid19_fallecidos_long.csv

```
In [16]: df4 = pd.read_csv('ccaa_covid19_fallecidos_long.csv', encoding='ISO-8859-1')
```

```
In [17]: df4.head()
```

Out[17]:

	fecha	cod_ine	CCAA	total
0	2020-03-04	1	Andalucía	0
1	2020-03-05	1	Andalucía	0
2	2020-03-06	1	Andalucía	0
3	2020-03-07	1	Andalucía	0
4	2020-03-08	1	Andalucía	0

```
In [18]: df4.profile_report()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

Alerts

`fecha` has a high cardinality: 82 distinct values	High cardinality
`cod_ine` is highly overall correlated with `CCAA`	High correlation
`CCAA` is highly overall correlated with `cod_ine`	High correlation
`fecha` is uniformly distributed	Uniform
`CCAA` is uniformly distributed	Uniform
`total` has 173 (11.1%) zeros	Zeros

Este fichero contiene número de fallecidos diarios por Comundiad Autónoma.

Presenta las variables: fecha , cod_ine , CCAA , total

ccaa_covid19_casos_long

```
In [20]: df5 = pd.read_csv('ccaa_covid19_casos_long.csv', encoding='ISO-8859-1')
```

```
In [21]: df5.head()
```

Out[21]:

	fecha	cod_ine	CCAA	total
0	2020-02-21	1	AndalucÃa	0
1	2020-02-22	1	AndalucÃa	0
2	2020-02-23	1	AndalucÃa	0
3	2020-02-24	1	AndalucÃa	0
4	2020-02-25	1	AndalucÃa	0

```
In [22]: df5.profile_report()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

Alerts

`fecha` has a high cardinality: 91 distinct values	High cardinality
`cod_ine` is highly overall correlated with `CCAA`	High correlation
`CCAA` is highly overall correlated with `cod_ine`	High correlation
`fecha` is uniformly distributed	Uniform
`CCAA` is uniformly distributed	Uniform
`total` has 171 (9.9%) zeros	Zeros

Este fichero contiene número de positivos diarios por Comunidad Autónoma.

Presenta las variables: fecha , cod_ine , CCAA , total

ccaa_covid19_altas_long

```
In [24]: df6 = pd.read_csv('ccaa_covid19_altas_long.csv', encoding='ISO-8859-1')
```

```
In [25]: df6.head()
```

Out[25]:

	fecha	cod_ine	CCAA	total
0	2020-03-01	1	Andalucía	NaN
1	2020-03-09	1	Andalucía	11.0
2	2020-03-10	1	Andalucía	11.0
3	2020-03-11	1	Andalucía	11.0
4	2020-03-12	1	Andalucía	11.0

```
In [26]: df6.profile_report()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

`fecha` has a high cardinality: 72 distinct values	High cardinality
`cod_ine` is highly overall correlated with `CCAA`	High correlation
`CCAA` is highly overall correlated with `cod_ine`	High correlation
`total` has 24 (1.8%) missing values	Missing
`fecha` is uniformly distributed	Uniform
`CCAA` is uniformly distributed	Uniform
`total` has 92 (6.7%) zeros	Zeros

Este fichero contiene número de altas hospitalarias diarias por Comunidad Autónoma.

Presenta las variables: fecha , cod_ine , CCAA , total

ccaa_camas_uci_2017

```
In [28]: df7 = pd.read_csv('ccaa_camas_uci_2017.csv', encoding='ISO-8859-1')
```

```
In [29]: df7.head()
```

Out[29]:

	ï»¿cod_ine	CCAA	PÃºblicos	Privados	Total
0	1	ANDALUCÃDA	572	162	734
1	2	ARAGÃ□N	113	22	135
2	3	ASTURIAS	86	7	93
3	15	C. FORAL DE NAVARRA	46	19	65
4	5	CANARIAS	187	50	237

```
In [30]: df7.profile_report()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

`ï»¿cod_ine` is highly overall correlated with `CCAA`	High correlation
`PÃºblicos` is highly overall correlated with `Privados` and 2 other fields	High correlation
`Privados` is highly overall correlated with `PÃºblicos` and 2 other fields	High correlation
`Total` is highly overall correlated with `PÃºblicos` and 2 other fields	High correlation
`CCAA` is highly overall correlated with `ï»¿cod_ine` and 3 other fields	High correlation
`CCAA` is uniformly distributed	Uniform
`ï»¿cod_ine` has unique values	Unique
`CCAA` has unique values	Unique
`PÃºblicos` has unique values	Unique
`Total` has unique values	Unique
`Privados` has 1 (5.9%) zeros	Zeros

Este fichero contiene número de camas hospitalarias por Comunidad Autónoma diferenciando por públicas y privadas.

Presenta las variables: fecha , cod_ine , CCAA , mascarillas_acumulado_desde_2020-03-10

nacional_covid19

```
In [32]: df8 = pd.read_csv('nacional_covid19.csv', encoding='ISO-8859-1')
```

```
In [33]: df8.head()
```

Out[33]:

	fecha	casos_total	casos_pcr	casos_test_ac	altas	fallecimientos	ingresos_uci	hospitalizados
0	2020-02-21	3.0	3	NaN	NaN	NaN	NaN	NaN
1	2020-02-22	3.0	3	NaN	NaN	NaN	NaN	NaN
2	2020-02-23	3.0	3	NaN	NaN	NaN	NaN	NaN
3	2020-02-24	3.0	3	NaN	NaN	NaN	NaN	NaN
4	2020-02-25	4.0	4	NaN	NaN	NaN	NaN	NaN

`fecha` has a high cardinality: 263 distinct values	High cardinality
`casos_total` is highly overall correlated with `casos_pcr` and 5 other fields	High correlation
`casos_pcr` is highly overall correlated with `casos_total` and 5 other fields	High correlation
`casos_test_ac` is highly overall correlated with `casos_total` and 5 other fields	High correlation
`altas` is highly overall correlated with `casos_total` and 5 other fields	High correlation
`fallecimientos` is highly overall correlated with `casos_total` and 5 other fields	High correlation
`ingresos_uci` is highly overall correlated with `casos_total` and 5 other fields	High correlation
`hospitalizados` is highly overall correlated with `casos_total` and 5 other fields	High correlation
`casos_total` has 172 (65.4%) missing values	Missing
`casos_test_ac` has 215 (81.7%) missing values	Missing
`altas` has 192 (73.0%) missing values	Missing
`fallecimientos` has 12 (4.6%) missing values	Missing
`ingresos_uci` has 70 (26.6%) missing values	Missing
`hospitalizados` has 68 (25.9%) missing values	Missing
`fecha` is uniformly distributed	Uniform
`fecha` has unique values	Unique

Este fichero contiene número de casos, pcr, test_ac, altas, fallecimientos ingresos uci y hospitalizados diarios a nivel nacional.

Presenta las variables: fecha , casos_total , casos_pcr , casos_test_ac , altas , fallecimientos , ingresos_uci , hospitalizados

nacional_covid19_rango_edad

```
In [36]: df9 = pd.read_csv('nacional_covid19_rango_edad.csv', encoding='ISO-8859-1')
```

```
In [37]: df9.head()
```

Out[37]:

	ï»¿fecha	rango_edad	sexo	casos_confirmados	hospitalizados	ingresos_uci	fallecidos
0	2020-03-23	0-9	ambos	129	34	1	0
1	2020-03-23	10-19	ambos	221	15	0	1
2	2020-03-23	20-29	ambos	1285	183	8	4
3	2020-03-23	30-39	ambos	2208	365	15	3
4	2020-03-23	40-49	ambos	2919	663	40	9

```
In [38]: df9.profile_report()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

`ï»¿fecha` has a high cardinality: 57 distinct values	High cardinality
`casos_confirmados` is highly overall correlated with `hospitalizados` and 2 other fields	High correlation
`hospitalizados` is highly overall correlated with `casos_confirmados` and 2 other fields	High correlation
`ingresos_uci` is highly overall correlated with `casos_confirmados` and 2 other fields	High correlation
`fallecidos` is highly overall correlated with `casos_confirmados` and 2 other fields	High correlation
`ï»¿fecha` is uniformly distributed	Uniform
`sexo` is uniformly distributed	Uniform
`ingresos_uci` has 26 (1.4%) zeros	Zeros
`fallecidos` has 75 (4.0%) zeros	Zeros

Este fichero contiene número de casos confirmados, hospitalizados, ingresos en uci y fallecidos diarios por sexo y rango de edad.

Presenta las variables: fecha , rango_edad , sexo , casos_confirmados , hospitalizados , ingresos_uci , fallecidos

Fabian

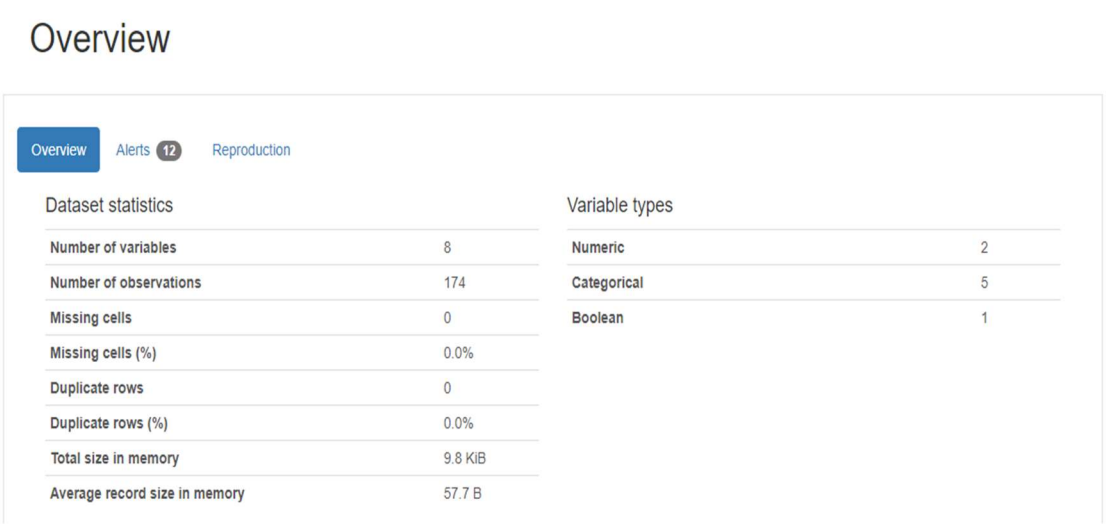
- **Nombre del Archivo: Case.csv**
- Ejemplo de Datos (head)

```
In [3]: df_ride.head()
```

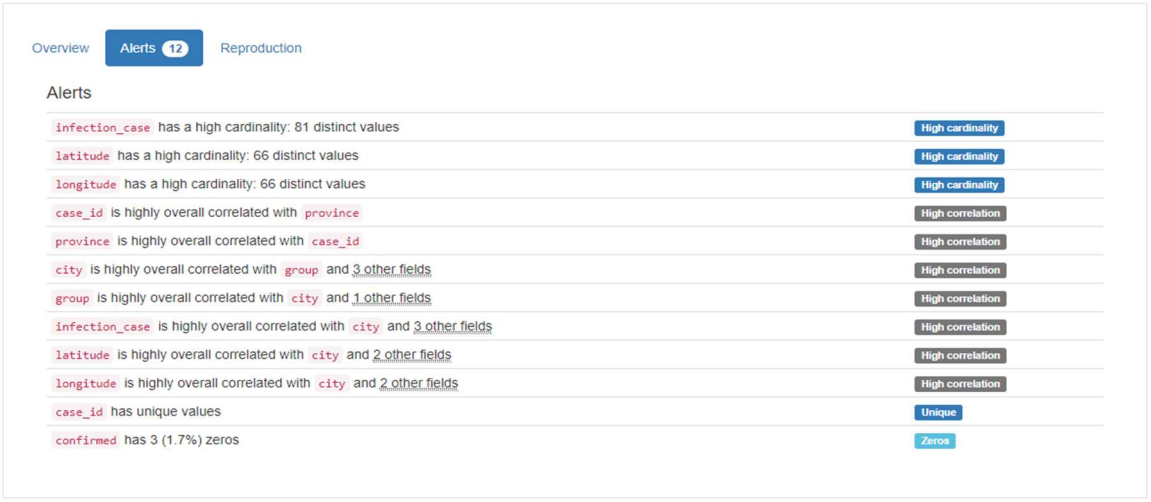
Out[3]:

	case_id	province	city	group	infection_case	confirmed	latitude	longitude
0	1000001	Seoul	Yongsan-gu	True	Itaewon Clubs	139	37.538621	126.992652
1	1000002	Seoul	Gwanak-gu	True	Richway	119	37.48208	126.901384
2	1000003	Seoul	Guro-gu	True	Guro-gu Call Center	95	37.508163	126.884387
3	1000004	Seoul	Yangcheon-gu	True	Yangcheon Table Tennis Club	43	37.546061	126.874209
4	1000005	Seoul	Dobong-gu	True	Day Care Center	43	37.679422	127.044374

- Resumen General (overview)



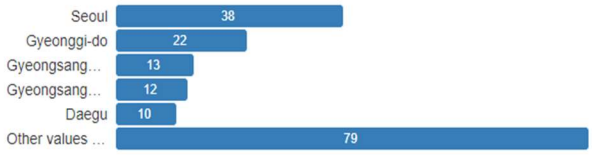
- Alertas (Alerts)



- Variables

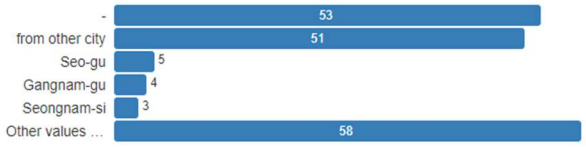
province
Categorical

Distinct	17
Distinct (%)	9.8%
Missing	0
Missing (%)	0.0%
Memory size	1.5 KiB



city
Categorical

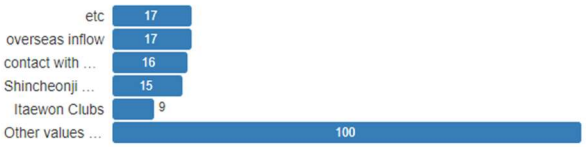
Distinct	48
Distinct (%)	27.6%
Missing	0
Missing (%)	0.0%
Memory size	1.5 KiB



infection_case
Categorical

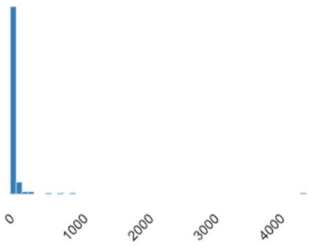
HIGH CARDINALITY HIGH CORRELATION

Distinct	81
Distinct (%)	46.6%
Missing	0
Missing (%)	0.0%
Memory size	1.5 KiB



confirmed
Real number (ℝ)

Distinct	62	Minimum	0
Distinct (%)	35.6%	Maximum	4511
Missing	0	Zeros	3
Missing (%)	0.0%	Zeros (%)	1.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	65.488506	Memory size	1.5 KiB



Explicación del resultado

Este archivo brinda datos sobre los focos de contagio de covid en Corea del Sur. Con 174 observaciones ofrece información sobre la localización, el número de contagiados, la ciudad y el nombre del caso. También identifica si ha sido un grupo el contagiado o no.

Observamos que el fichero consta de 8 variables, de las cuales 5 son categóricas, 2 numéricas y 1 boolena. No existen valores duplicados ni nulos. Hay variables que están altamente correlacionadas como son los case_id, province, city, group, infection_case, latitude y lingitude. Algunas de estas correlaciones son esperadas dado que los valores de latitude y longitude para city se mantiene constantes.

- **Nombre del Archivo: PatientInfo.csv**

- Ejemplo de Datos (head)

```
In [3]: df_ride.head()
```

Out[3]:

	patient_id	sex	age	country	province	city	infection_case	infected_by	contact_number	symptom_onset_date	confirmed_date	released_date	deceased_date	state
0	1000000001	male	50s	Korea	Seoul	Gangseo-gu	overseas inflow	NaN	75	2020-01-22	2020-01-23	2020-02-05	NaN	released
1	1000000002	male	30s	Korea	Seoul	Jungnang-gu	overseas inflow	NaN	31	NaN	2020-01-30	2020-03-02	NaN	released
2	1000000003	male	50s	Korea	Seoul	Jongno-gu	contact with patient	2002000001	17	NaN	2020-01-30	2020-02-19	NaN	released
3	1000000004	male	20s	Korea	Seoul	Mapo-gu	overseas inflow	NaN	9	2020-01-26	2020-01-30	2020-02-15	NaN	released
4	1000000005	female	20s	Korea	Seoul	Seongbuk-gu	contact with patient	1000000002	2	NaN	2020-01-31	2020-02-24	NaN	released

- Resumen General (overview)

Overview

Overview | Alerts 22 | Reproduction

Dataset statistics

Number of variables	14
Number of observations	5165
Missing cells	24863
Missing cells (%)	34.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	565.0 KiB
Average record size in memory	112.0 B

Variable types

Numeric	1
Categorical	13

- Alertas (Alerts)

Overview | Alerts 22 | Reproduction

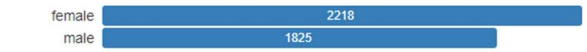
Alerts

`city` has a high cardinality: 163 distinct values	High cardinality
`infection_case` has a high cardinality: 51 distinct values	High cardinality
`infected_by` has a high cardinality: 606 distinct values	High cardinality
`contact_number` has a high cardinality: 88 distinct values	High cardinality
`symptom_onset_date` has a high cardinality: 129 distinct values	High cardinality
`confirmed_date` has a high cardinality: 148 distinct values	High cardinality
`released_date` has a high cardinality: 130 distinct values	High cardinality
`patient_id` is highly overall correlated with `province`	High correlation
`province` is highly overall correlated with `patient_id` and 1 other fields	High correlation
`contact_number` is highly overall correlated with `deceased_date`	High correlation
`deceased_date` is highly overall correlated with `contact_number` and 1 other fields	High correlation
`state` is highly overall correlated with `province` and 1 other fields	High correlation
`country` is highly imbalanced (97.6%)	Imbalance
`sex` has 1122 (21.7%) missing values	Missing
`age` has 1380 (26.7%) missing values	Missing
`city` has 94 (1.8%) missing values	Missing
`infection_case` has 919 (17.8%) missing values	Missing
`infected_by` has 3819 (73.9%) missing values	Missing
`contact_number` has 4374 (84.7%) missing values	Missing
`symptom_onset_date` has 4475 (86.6%) missing values	Missing
`released_date` has 3578 (69.3%) missing values	Missing
`deceased_date` has 5099 (98.7%) missing values	Missing

- Variables

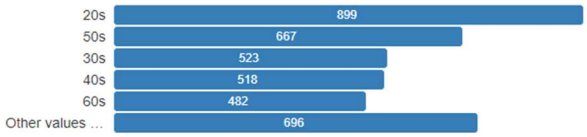
sex
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	1122
Missing (%)	21.7%
Memory size	40.5 KiB



age
Categorical

Distinct	11
Distinct (%)	0.3%
Missing	1380
Missing (%)	26.7%
Memory size	40.5 KiB

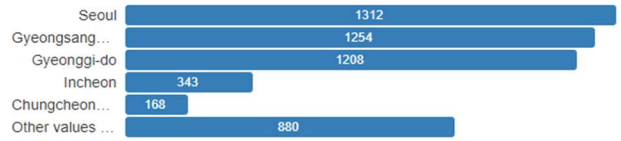


country
Categorical

Distinct	16
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	40.5 KiB

province
Categorical

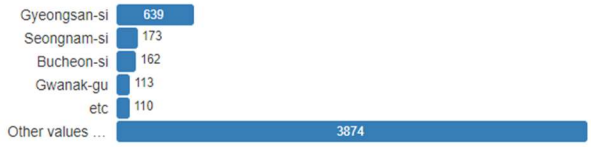
Distinct	17
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	40.5 KiB



city
Categorical

HIGH CARDINALITY MISSING

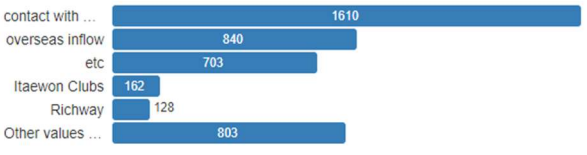
Distinct	163
Distinct (%)	3.2%
Missing	94
Missing (%)	1.8%
Memory size	40.5 KiB



infection_case
Categorical

HIGH CARDINALITY MISSING

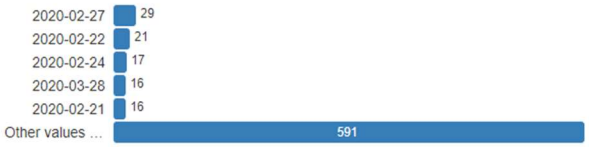
Distinct	51
Distinct (%)	1.2%
Missing	919
Missing (%)	17.8%
Memory size	40.5 KiB



symptom_onset_date
Categorical

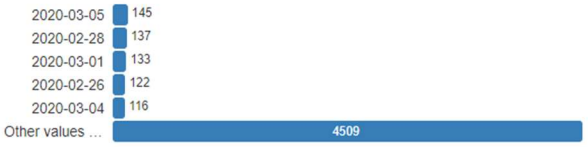
HIGH CARDINALITY MISSING

Distinct	129
Distinct (%)	18.7%
Missing	4475
Missing (%)	86.6%
Memory size	40.5 KiB



confirmed_date
Categorical

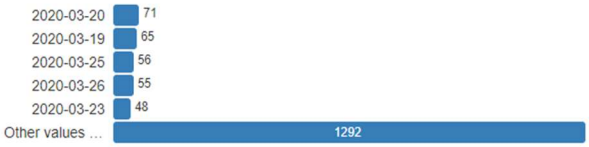
Distinct	148
Distinct (%)	2.9%
Missing	3
Missing (%)	0.1%
Memory size	40.5 KiB



released_date
Categorical

HIGH CARDINALITY MISSING

Distinct	130
Distinct (%)	8.2%
Missing	3578
Missing (%)	69.3%
Memory size	40.5 KiB



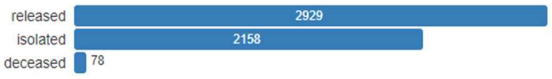
deceased_date
Categorical

HIGH CORRELATION MISSING

Distinct	39
Distinct (%)	59.1%
Missing	5099
Missing (%)	98.7%
Memory size	40.5 KiB

state
Categorical

Distinct	3
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	40.5 KiB



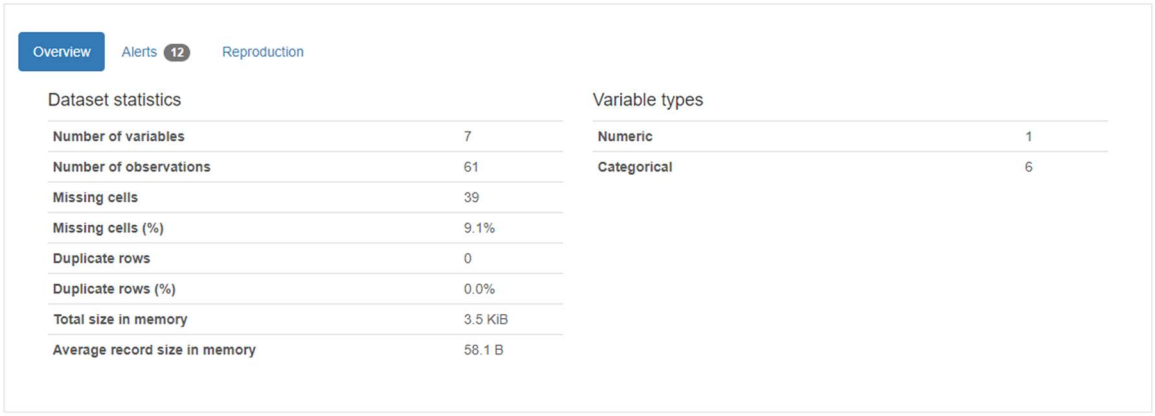
Explicación del resultado

Este archivo ofrece información sobre pacientes infectados a través de 5165 observaciones. Podemos ver la edad, el sexo, la dirección si está relacionado con un caso del fichero anterior, fechas y si el paciente murió. Presenta 13 variables categóricas y 1 numéricas, no existiendo valores duplicados, pero si un 34.4% de valores nulos. Existe una alta correlación entre varias variables como ser patient_id, province, state, contact number y deceased. La correlación entre state y province es esperable debido a que uno pertenece a otro.

- **Nombre del Archivo: Policy.csv**
- Ejemplo de Datos (head)

	policy_id	country	type	gov_policy	detail	start_date	end_date
0	1	Korea	Alert	Infectious Disease Alert Level	Level 1 (Blue)	2020-01-03	2020-01-19
1	2	Korea	Alert	Infectious Disease Alert Level	Level 2 (Yellow)	2020-01-20	2020-01-27
2	3	Korea	Alert	Infectious Disease Alert Level	Level 3 (Orange)	2020-01-28	2020-02-22
3	4	Korea	Alert	Infectious Disease Alert Level	Level 4 (Red)	2020-02-23	NaN
4	5	Korea	Immigration	Special Immigration Procedure	from China	2020-02-04	NaN

- Resumen General (overview)



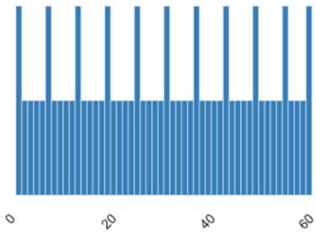
- Alertas (Alerts)

- Variables

policy_id
Real number (ℝ)

HIGH CORRELATION UNIFORM UNIQUE

Distinct	61	**Minimum**	1
Distinct (%)	100.0%	**Maximum**	61
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	31	**Memory size**	616.0 B

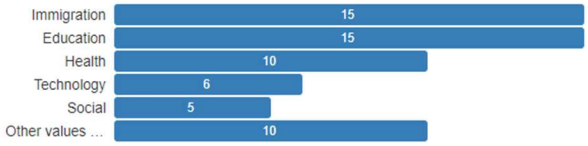


country
Categorical

Distinct	1
Distinct (%)	1.6%
Missing	0
Missing (%)	0.0%
Memory size	616.0 B

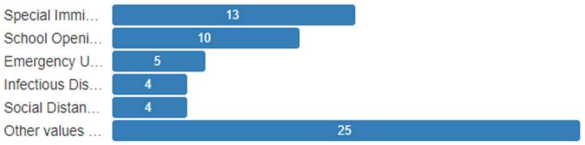
type
Categorical

Distinct	8
Distinct (%)	13.1%
Missing	0
Missing (%)	0.0%
Memory size	616.0 B



gov_policy
Categorical

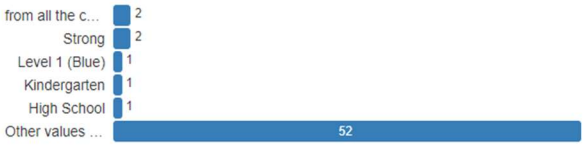
Distinct	24
Distinct (%)	39.3%
Missing	0
Missing (%)	0.0%
Memory size	616.0 B



detail
Categorical

HIGH CARDINALITY MISSING UNIFORM

Distinct	57
Distinct (%)	96.6%
Missing	2
Missing (%)	3.3%
Memory size	616.0 B



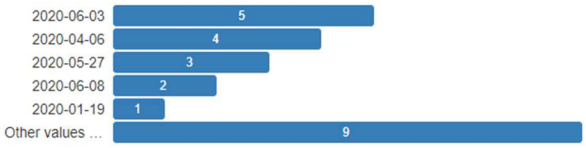
start_date
Categorical

Distinct	35
Distinct (%)	57.4%
Missing	0
Missing (%)	0.0%
Memory size	616.0 B

end_date
Categorical

HIGH CORRELATION MISSING

Distinct	14
Distinct (%)	58.3%
Missing	37
Missing (%)	60.7%
Memory size	616.0 B



Explicación del resultado

Este archivo tiene como datos las diferentes políticas que sacó el gobierno coreano, diferenciando el tipo y con fecha inicio y fecha fin, siendo 61 observaciones en total. Esta compuesto por una variable numérica y 6 categoricas, sin presencia de duplicados pero con un 9.1% de datos nulos.

Existe una alta correlación entre las variables policy_id, type, gov_policy, end_date, start_date

- **Nombre del Archivo: Region.csv**

- Ejemplo de Datos (head)

	code	province	city	latitude	longitude	elementary_school_count	kindergarten_count	university_count	academy_ratio	elderly_population_ratio	elderly_alone_ratio	nursing_home_count
0	10000	Seoul	Seoul	37.566953	126.977977	607	830	48	1.44	15.38	5.8	22739
1	10010	Seoul	Gangnam-gu	37.518421	127.047222	33	38	0	4.18	13.17	4.3	3088
2	10020	Seoul	Gangdong-gu	37.530492	127.123837	27	32	0	1.54	14.55	5.4	1023
3	10030	Seoul	Gangbuk-gu	37.639938	127.025508	14	21	0	0.67	19.49	8.5	628
4	10040	Seoul	Gangseo-gu	37.551166	126.849506	36	56	1	1.17	14.39	5.7	1080

- Resumen General (overview)

Overview Alerts 15 Reproduction

Dataset statistics

Number of variables	12
Number of observations	244
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	23.0 KiB
Average record size in memory	96.5 B

Variable types

Numeric	10
Categorical	2

- Alertas (Alerts)

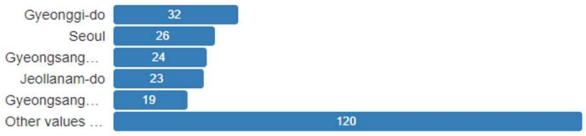
Alerts

`city` has a high cardinality: 222 distinct values	High cardinality
`code` is highly overall correlated with `elderly_alone_ratio` and 1 other fields	High correlation
`latitude` is highly overall correlated with `province`	High correlation
`longitude` is highly overall correlated with `province`	High correlation
`elementary_school_count` is highly overall correlated with `kindergarten_count` and 6 other fields	High correlation
`kindergarten_count` is highly overall correlated with `elementary_school_count` and 6 other fields	High correlation
`university_count` is highly overall correlated with `elementary_school_count` and 3 other fields	High correlation
`academy_ratio` is highly overall correlated with `elementary_school_count` and 4 other fields	High correlation
`elderly_population_ratio` is highly overall correlated with `elementary_school_count` and 4 other fields	High correlation
`elderly_alone_ratio` is highly overall correlated with `code` and 5 other fields	High correlation
`nursing_home_count` is highly overall correlated with `elementary_school_count` and 6 other fields	High correlation
`province` is highly overall correlated with `code` and 6 other fields	High correlation
`city` is uniformly distributed	Uniform
`code` has unique values	Unique
`university_count` has 92 (37.7%) zeros	Zeros

- Variables

province
Categorical

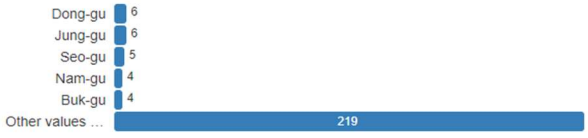
Distinct	18
Distinct (%)	7.4%
Missing	0
Missing (%)	0.0%
Memory size	2.0 KiB



city
Categorical

HIGH CARDINALITY UNIFORM

Distinct	222
Distinct (%)	91.0%
Missing	0
Missing (%)	0.0%
Memory size	2.0 KiB

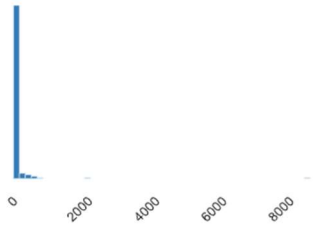


elementary_school_count
Real number (ℝ)

Distinct	78	**Minimum**	4
Distinct (%)	32.0%	**Maximum**	6087
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	74.180328	**Memory size**	2.0 KiB

kindergarten_count
Real number (ℝ)

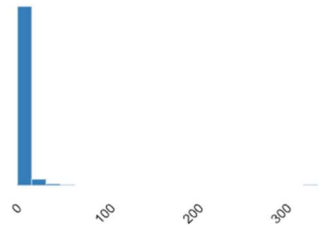
Distinct	101	**Minimum**	4
Distinct (%)	41.4%	**Maximum**	8837
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	107.90164	**Memory size**	2.0 KiB



university_count
Real number (ℝ)

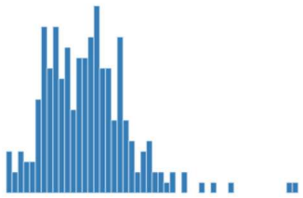
HIGH CORRELATION ZEROS

Distinct	21	**Minimum**	0
Distinct (%)	8.6%	**Maximum**	340
Missing	0	**Zeros**	92
Missing (%)	0.0%	**Zeros (%)**	37.7%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4.1516393	**Memory size**	2.0 KiB



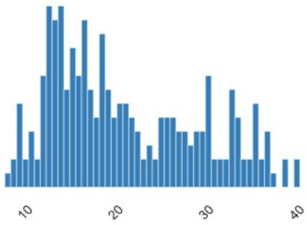
academy_ratio
Real number (ℝ)

Distinct	144	**Minimum**	0.19
Distinct (%)	59.0%	**Maximum**	4.18
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.2947541	**Memory size**	2.0 KiB



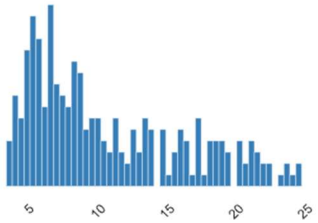
elderly_population_ratio
Real number (ℝ)

Distinct	229	**Minimum**	7.69
Distinct (%)	93.9%	**Maximum**	40.26
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	20.92373	**Memory size**	2.0 KiB



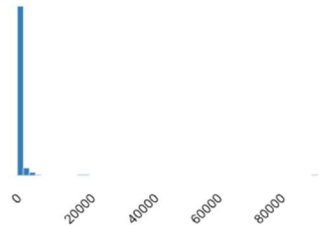
elderly_alone_ratio
Real number (ℝ)

Distinct	130	**Minimum**	3.3
Distinct (%)	53.3%	**Maximum**	24.7
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	10.644672	**Memory size**	2.0 KiB



nursing_home_count
Real number (ℝ)

Distinct	208	**Minimum**	11
Distinct (%)	85.2%	**Maximum**	94865
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1159.2582	**Memory size**	2.0 KiB



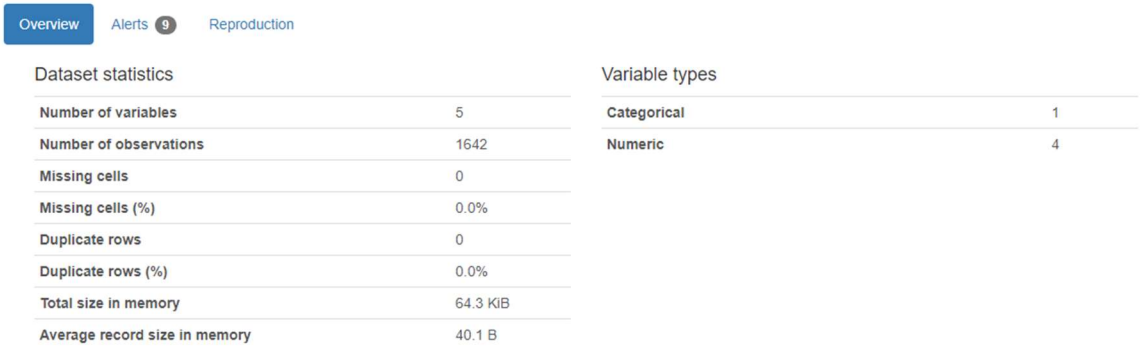
Explicación del resultado

Este archivo brinda datos poblacionales de las diferentes regiones de Corea, siendo en total 244 observaciones, presentando 10 variable numéricas y una categórica, no existiendo valores duplicados ni nulos. Existe una alta correlación entre la mayoría de ellas, siendo esperada la correlación entre latitude y longitude con province particularmente.

- **Nombre del Archivo: SerachTrend.csv**
- Ejemplo de Datos (head)

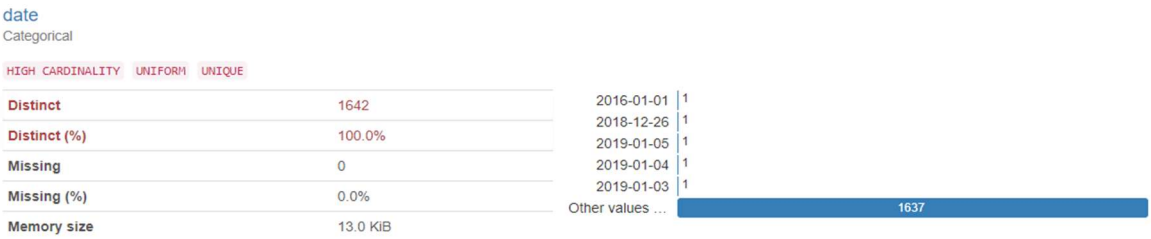
	date	cold	flu	pneumonia	coronavirus
0	2016-01-01	0.11663	0.05590	0.15726	0.00736
1	2016-01-02	0.13372	0.17135	0.20826	0.00890
2	2016-01-03	0.14917	0.22317	0.19326	0.00845
3	2016-01-04	0.17463	0.18626	0.29008	0.01145
4	2016-01-05	0.17226	0.15072	0.24562	0.01381

- Resumen General (overview)



- Alertas (Alerts)

- Variables



cold
Real number (ℝ)

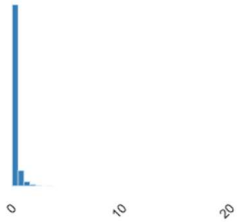
HIGH CORRELATION SKEWED

Distinct	1085	**Minimum**	0.05163
Distinct (%)	66.1%	**Maximum**	15.72071
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.19050643	**Memory size**	13.0 KiB

flu
Real number (ℝ)

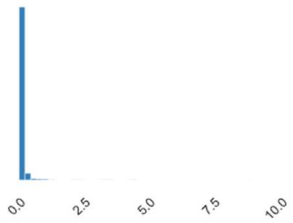
HIGH CORRELATION SKEWED

Distinct	1312	**Minimum**	0.00981
Distinct (%)	79.9%	**Maximum**	27.32727
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.24494582	**Memory size**	13.0 KiB



pneumonia
Real number (ℝ)

Distinct	1184	**Minimum**	0.06881
Distinct (%)	72.1%	**Maximum**	11.3932
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.22142959	**Memory size**	13.0 KiB



coronavirus
Real number (ℝ)

Distinct	417	**Minimum**	0.00154
Distinct (%)	25.4%	**Maximum**	100
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.8625222	**Memory size**	13.0 KiB

Explicación del resultado

Este archivo aporta la tasa de incidencia de diferentes enfermedades desde 2016. Se puede ver la tasa de incidencia de resfriado, gripe, neumonía y coronavirus. Con 1642 observaciones, contando con 1 variable categórica y 4 numéricas. No presenta valores duplicados ni nulos, presentando un alta correlación entre cold, flu, pneumonia y coronavirus

- **Nombre del Archivo: SeoulFloating**

- Ejemplo de Datos (head)

	date	hour	birth_year	sex	province	city	fp_num
0	2020-01-01	0	20	female	Seoul	Dobong-gu	19140
1	2020-01-01	0	20	male	Seoul	Dobong-gu	19950
2	2020-01-01	0	20	female	Seoul	Dongdaemun-gu	25450
3	2020-01-01	0	20	male	Seoul	Dongdaemun-gu	27050
4	2020-01-01	0	20	female	Seoul	Dongjag-gu	28880

- Resumen General (overview)

Overview | Alerts 7 | Reproduction

Dataset statistics

Number of variables	7
Number of observations	1084800
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	57.9 MiB
Average record size in memory	56.0 B

Variable types

Categorical	4
Numeric	3

- Alertas (Alerts)

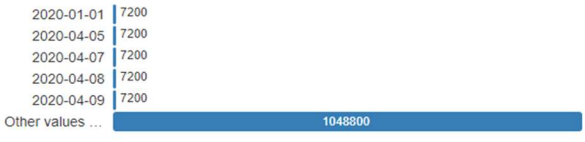
Alerts

`province` has constant value "Seoul"	Constant
`date` has a high cardinality: 151 distinct values	High cardinality
`birth_year` is highly overall correlated with `fp_num`	High correlation
`fp_num` is highly overall correlated with `birth_year`	High correlation
`sex` is uniformly distributed	Uniform
`city` is uniformly distributed	Uniform
`hour` has 45300 (4.2%) zeros	Zeros

- Variables

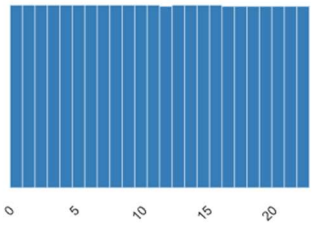
date
Categorical

Distinct	151
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	8.3 MiB



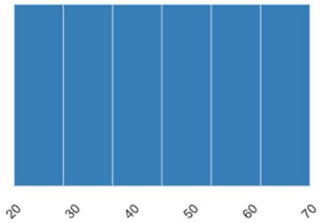
hour
Real number (ℝ)

Distinct	24	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	23
Missing	0	**Zeros**	45300
Missing (%)	0.0%	**Zeros (%)**	4.2%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	11.483407	**Memory size**	8.3 MiB



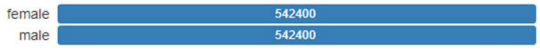
birth_year
Real number (ℝ)

Distinct	6	**Minimum**	20
Distinct (%)	< 0.1%	**Maximum**	70
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	45	**Memory size**	8.3 MiB



sex
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	8.3 MiB

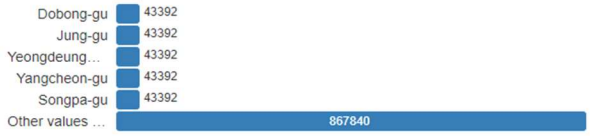


province
Categorical

Distinct	1
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	8.3 MiB

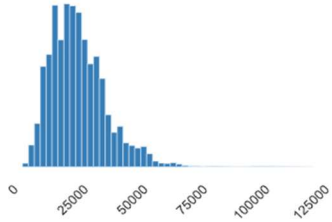
city
Categorical

Distinct	25
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	8.3 MiB



fp_num
Real number (ℝ)

Distinct	10669	**Minimum**	3630
Distinct (%)	1.0%	**Maximum**	127640
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	27426.965	**Memory size**	8.3 MiB



Explicación del resultado

Este archivo brinda datos de la población flotante en Seul por distritos. Nos marca la cantidad de gente que hay en cada distrito por sexo, edad y hora del día contando con un total de 1084800 observaciones. Compuesto por 4 variables categóricas y 3 numéricas, no presenta valores nulos ni duplicados. Presenta una alta correlación entre las variables fp_num y birth_year

- **Nombre del Archivo: Time.csv**
- Ejemplo de Datos (head)

	date	time	test	negative	confirmed	released	deceased
0	2020-01-20	16	1	0	1	0	0
1	2020-01-21	16	1	0	1	0	0
2	2020-01-22	16	4	3	1	0	0
3	2020-01-23	16	22	21	1	0	0
4	2020-01-24	16	27	25	2	0	0

- Resumen General (overview)

Overview | Alerts 12 | Reproduction

Dataset statistics

Number of variables	7
Number of observations	163
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	9.0 KiB
Average record size in memory	56.8 B

Variable types

Categorical	2
Numeric	5

- Alertas (Alerts)

Alerts

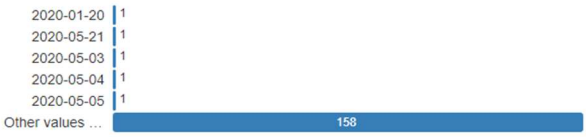
`date` has a high cardinality: 163 distinct values	High cardinality
`test` is highly overall correlated with `negative` and 4 other fields	High correlation
`negative` is highly overall correlated with `test` and 4 other fields	High correlation
`confirmed` is highly overall correlated with `test` and 4 other fields	High correlation
`released` is highly overall correlated with `test` and 4 other fields	High correlation
`deceased` is highly overall correlated with `test` and 4 other fields	High correlation
`time` is highly overall correlated with `test` and 4 other fields	High correlation
`date` is uniformly distributed	Uniform
`date` has unique values	Unique
`negative` has 2 (1.2%) zeros	Zeros
`released` has 16 (9.8%) zeros	Zeros
`deceased` has 31 (19.0%) zeros	Zeros

- Variables

date
Categorical

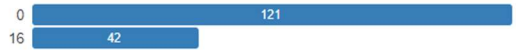
HIGH CARDINALITY UNIFORM UNIQUE

Distinct	163
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	1.4 KiB



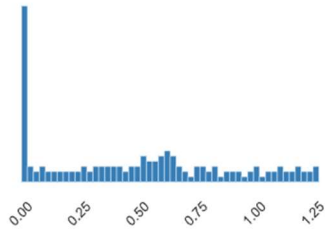
time
Categorical

Distinct	2
Distinct (%)	1.2%
Missing	0
Missing (%)	0.0%
Memory size	1.4 KiB



test
Real number (ℝ)

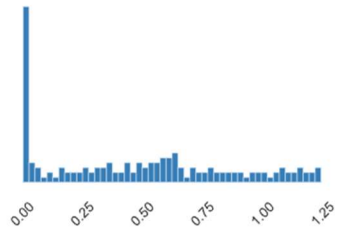
Distinct	161	**Minimum**	1
Distinct (%)	98.8%	**Maximum**	1273766
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	497779.72	**Memory size**	1.4 KiB



negative
Real number (ℝ)

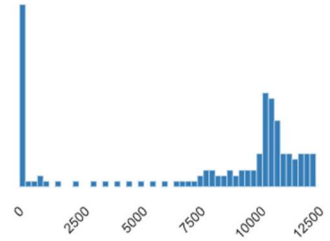
HIGH CORRELATION ZEROS

Distinct	161	**Minimum**	0
Distinct (%)	98.8%	**Maximum**	1240157
Missing	0	**Zeros**	2
Missing (%)	0.0%	**Zeros (%)**	1.2%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	475483.9	**Memory size**	1.4 KiB



confirmed
Real number (ℝ)

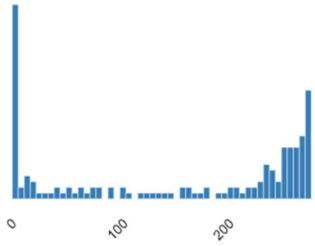
Distinct	150	**Minimum**	1
Distinct (%)	92.0%	**Maximum**	12800
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	7834.865	**Memory size**	1.4 KiB



deceased
Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	93	Minimum	0
Distinct (%)	57.1%	Maximum	282
Missing	0	Zeros	31
Missing (%)	0.0%	Zeros (%)	19.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	157.10429	Memory size	1.4 KiB



Explicación del resultado

Este archivo aporta los datos diarios del número de test realizados y sus resultados a través de 163 observaciones compuestas por 5 variables numéricas y 2 categóricas sin presencia de valores nulos o duplicados. Presenta una alta correlación entre test y negative, confirmed, deceased, released y time

Nombre del Archivo: TimeAge.csv

- Ejemplo de Datos (head)

	date	time	age	confirmed	deceased
0	2020-03-02	0	0s	32	0
1	2020-03-02	0	10s	169	0
2	2020-03-02	0	20s	1235	0
3	2020-03-02	0	30s	506	1
4	2020-03-02	0	40s	633	1

- Resumen General (overview)

Overview Alerts 8 Reproduction

Dataset statistics

Number of variables	5
Number of observations	1089
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	42.7 KiB
Average record size in memory	40.1 B

Variable types

Categorical	3
Numeric	2

- Alertas (Alerts)

Alerts

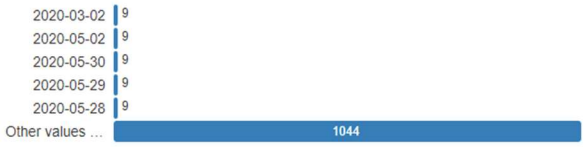
`time` has constant value "0"	Constant
`date` has a high cardinality: 121 distinct values	High cardinality
`confirmed` is highly overall correlated with `age`	High correlation
`deceased` is highly overall correlated with `age`	High correlation
`age` is highly overall correlated with `confirmed` and 1 other fields	High correlation
`date` is uniformly distributed	Uniform
`age` is uniformly distributed	Uniform
`deceased` has 363 (33.3%) zeros	Zeros

- Variables

date
Categorical

`HIGH CARDINALITY` `UNIFORM`

Distinct	121
Distinct (%)	11.1%
Missing	0
Missing (%)	0.0%
Memory size	8.6 KiB



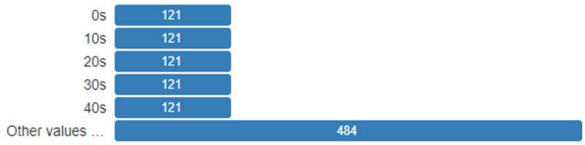
time
Categorical

Distinct	1
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	8.6 KiB

age
Categorical

`HIGH CORRELATION` `UNIFORM`

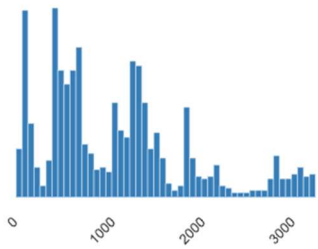
Distinct	9
Distinct (%)	0.8%
Missing	0
Missing (%)	0.0%
Memory size	8.6 KiB



confirmed
Real number (ℝ)

Distinct	827
Distinct (%)	75.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1158.1295

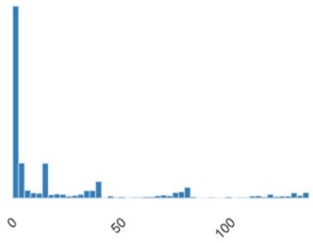
Minimum	32
Maximum	3362
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	8.6 KiB



deceased
Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	107	Minimum	0
Distinct (%)	9.8%	Maximum	139
Missing	0	Zeros	363
Missing (%)	0.0%	Zeros (%)	33.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	23.419651	Memory size	8.6 KiB



Explicación del resultado

Este archivo presenta los datos diarios por franja de edad siendo 1089 observaciones compuestas por 3 variables categóricas y 2 numéricas que no presentan valores nulos ni duplicados, existiendo una alta correlación entre age, confirmed y deceased.

Nombre del Archivo: TimeGender.csv

- Ejemplo de Datos (head)

	date	time	sex	confirmed	deceased
0	2020-03-02	0	male	1591	13
1	2020-03-02	0	female	2621	9
2	2020-03-03	0	male	1810	16
3	2020-03-03	0	female	3002	12
4	2020-03-04	0	male	1996	20

- Resumen General (overview)

Overview Alerts 6 Reproduction

Dataset statistics

Number of variables	5
Number of observations	242
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	9.6 KiB
Average record size in memory	40.5 B

Variable types

Categorical	3
Numeric	2

- Alertas (Alerts)

Alerts

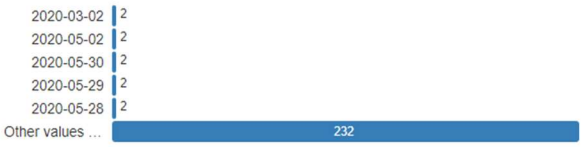
`time` has constant value "0"	Constant
`date` has a high cardinality: 121 distinct values	High cardinality
`confirmed` is highly overall correlated with `sex`	High correlation
`sex` is highly overall correlated with `confirmed`	High correlation
`date` is uniformly distributed	Uniform
`sex` is uniformly distributed	Uniform

- Variables

date
Categorical

HIGH CARDINALITY UNIFORM

Distinct	121
Distinct (%)	50.0%
Missing	0
Missing (%)	0.0%
Memory size	2.0 KiB



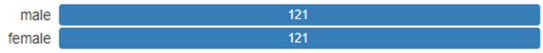
time
Categorical

Distinct	1
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	2.0 KiB

sex
Categorical

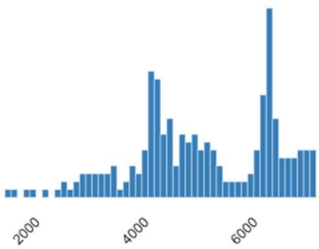
HIGH CORRELATION UNIFORM

Distinct	2
Distinct (%)	0.8%
Missing	0
Missing (%)	0.0%
Memory size	2.0 KiB



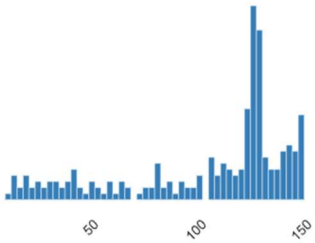
confirmed
Real number (ℝ)

Distinct	238	**Minimum**	1591
Distinct (%)	98.3%	**Maximum**	7305
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	5211.5455	**Memory size**	2.0 KiB



deceased
Real number (ℝ)

Distinct	102	**Minimum**	9
Distinct (%)	42.1%	**Maximum**	151
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	105.3843	**Memory size**	2.0 KiB



Explicación del resultado

Este archivo aporta los datos diarios por sexo, siendo 242 observaciones, compuestas por 3 variables categóricas y 2 numéricas, sin presencia de valores nulos ni duplicados. Existe una alta correlación entre sex y confirmed.

Nombre del Archivo: TimeProvince.csv

- Ejemplo de Datos (head)

	date	time	province	confirmed	released	deceased
0	2020-01-20	16	Seoul	0	0	0
1	2020-01-20	16	Busan	0	0	0
2	2020-01-20	16	Daegu	0	0	0
3	2020-01-20	16	Incheon	1	0	0
4	2020-01-20	16	Gwangju	0	0	0

- Resumen General (overview)

Overview | Alerts 9 | Reproduction

Dataset statistics

Number of variables	6
Number of observations	2771
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	130.0 KiB
Average record size in memory	48.0 B

Variable types

Categorical	3
Numeric	3

- Alertas (Alerts)

Alerts

`date` has a high cardinality: 163 distinct values	High cardinality
`confirmed` is highly overall correlated with `released` and 1 other fields	High correlation
`released` is highly overall correlated with `confirmed` and 1 other fields	High correlation
`deceased` is highly overall correlated with `confirmed` and 1 other fields	High correlation
`date` is uniformly distributed	Uniform
`province` is uniformly distributed	Uniform
`confirmed` has 418 (15.1%) zeros	Zeros
`released` has 638 (23.0%) zeros	Zeros
`deceased` has 1912 (69.0%) zeros	Zeros

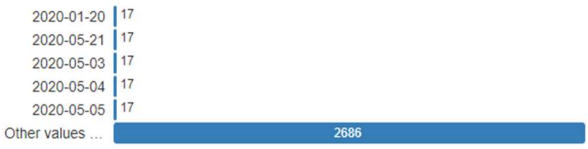
Variables

date

Categorical

`HIGH CARDINALITY` `UNIFORM`

Distinct	163
Distinct (%)	5.9%
Missing	0
Missing (%)	0.0%
Memory size	21.8 KiB



time

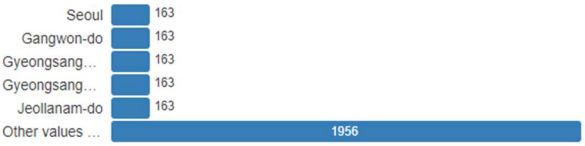
Categorical

Distinct	2
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	21.8 KiB

province

Categorical

Distinct	17
Distinct (%)	0.6%
Missing	0
Missing (%)	0.0%
Memory size	21.8 KiB

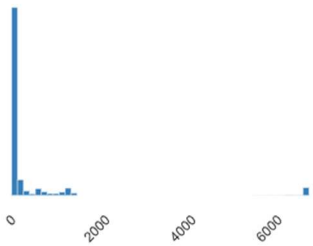


confirmed

Real number (ℝ)

`HIGH CORRELATION` `ZEROS`

Distinct	544	**Minimum**	0
Distinct (%)	19.6%	**Maximum**	6906
Missing	0	**Zeros**	418
Missing (%)	0.0%	**Zeros (%)**	15.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	444.30819	**Memory size**	21.8 KiB

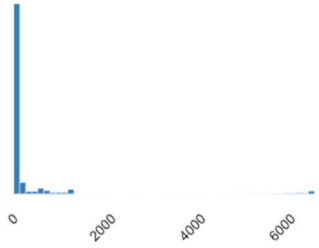


released

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	507	Minimum	0
Distinct (%)	18.3%	Maximum	6700
Missing	0	Zeros	638
Missing (%)	0.0%	Zeros (%)	23.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	320.72645	Memory size	21.8 KiB

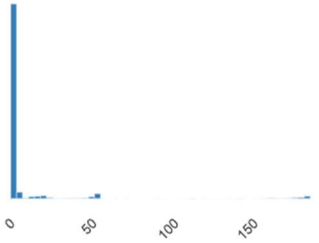


deceased

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	104	Minimum	0
Distinct (%)	3.8%	Maximum	189
Missing	0	Zeros	1912
Missing (%)	0.0%	Zeros (%)	69.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	9.2399856	Memory size	21.8 KiB



Explicación del resultado

Este archivo aporta los datos diarios por provincia con 2771 observaciones, compuesto por 3 variables categóricas y 3 numéricas, sin presencia de valores nulos ni duplicados, estando confirmed altamente correlacionado con realised y deceased

Nombre del Archivo: Weather.csv

- Ejemplo de Datos (head)

	code	province	date	avg_temp	min_temp	max_temp	precipitation	max_wind_speed	most_wind_direction	avg_relative_humidity
0	10000	Seoul	2016-01-01	1.2	-3.3	4.0	0.0	3.5	90.0	73.0
1	11000	Busan	2016-01-01	5.3	1.1	10.9	0.0	7.4	340.0	52.1
2	12000	Daegu	2016-01-01	1.7	-4.0	8.0	0.0	3.7	270.0	70.5
3	13000	Gwangju	2016-01-01	3.2	-1.5	8.1	0.0	2.7	230.0	73.1
4	14000	Incheon	2016-01-01	3.1	-0.4	5.7	0.0	5.3	180.0	83.9

- Resumen General (overview)

Overview | Alerts 9 | Reproduction

Dataset statistics

Number of variables	10
Number of observations	26271
Missing cells	81
Missing cells (%)	< 0.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	2.0 MiB
Average record size in memory	80.0 B

Variable types

Numeric	8
Categorical	2

- Alertas (Alerts)

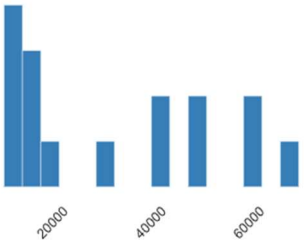
Alerts

`date` has a high cardinality: 1642 distinct values	High cardinality
`code` is highly overall correlated with `province`	High correlation
`avg_temp` is highly overall correlated with `min_temp` and 1 other fields	High correlation
`min_temp` is highly overall correlated with `avg_temp` and 2 other fields	High correlation
`max_temp` is highly overall correlated with `avg_temp` and 1 other fields	High correlation
`avg_relative_humidity` is highly overall correlated with `min_temp`	High correlation
`province` is highly overall correlated with `code`	High correlation
`date` is uniformly distributed	Uniform
`precipitation` has 22778 (86.7%) zeros	Zeros

- Variables

code
Real number (ℝ)

Distinct	16	Minimum	10000
Distinct (%)	0.1%	Maximum	70000
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	32124.662	Memory size	205.4 KiB



Explicación del resultado

Este archivo aporta datos meterológicos por provincia y diarios. Esta compuesto por 26271 observaciones, en 10 variables, 8 numéricas y 10 categóricas, no presentando valores duplicados y menos de 0.1% de datos nulos. Existe una alta correlación entre avg_temp y max_temp y min_temp la cual es esperable dado que una se calcula en función de las segundas y también entre average_relative_humidity y min_temp. También hay una alta correlación entre code y province, la cual es esperable dado que a una misma province se le asigna el mismo code.

Amaia

Nombre del Archivo: us_counties_covid19_daily - Carpeta COVID USA

- Ejemplo de Datos (head)

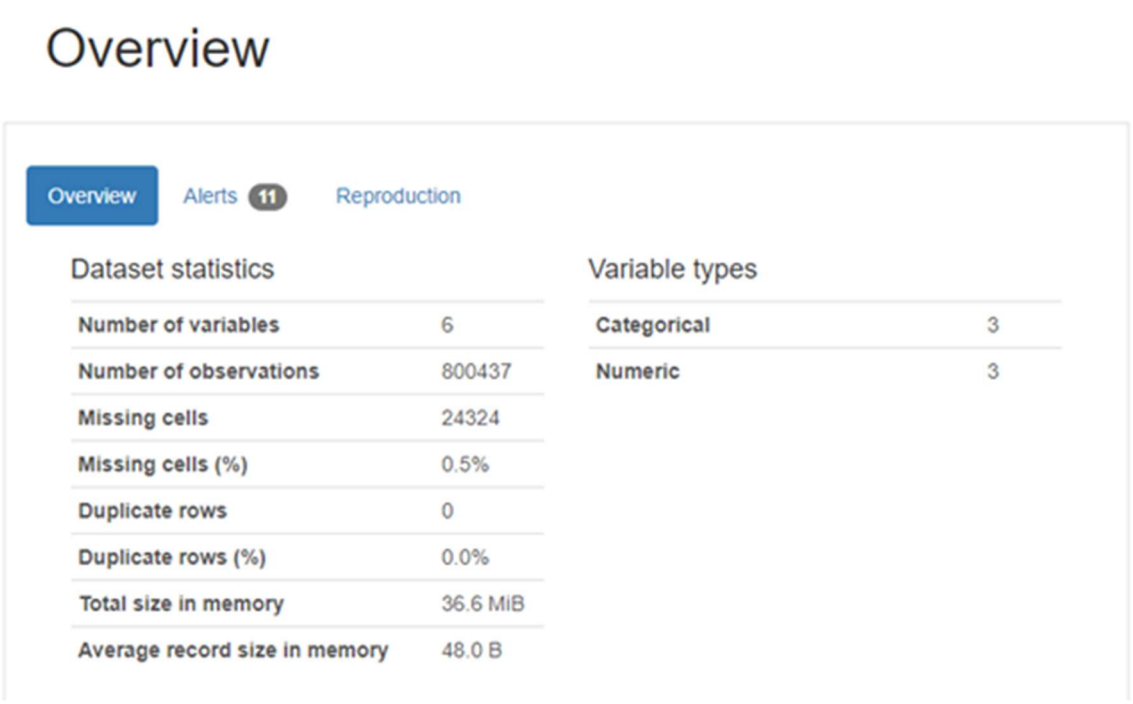
```
In [7]: df_ride.head()
```

Out[7]:

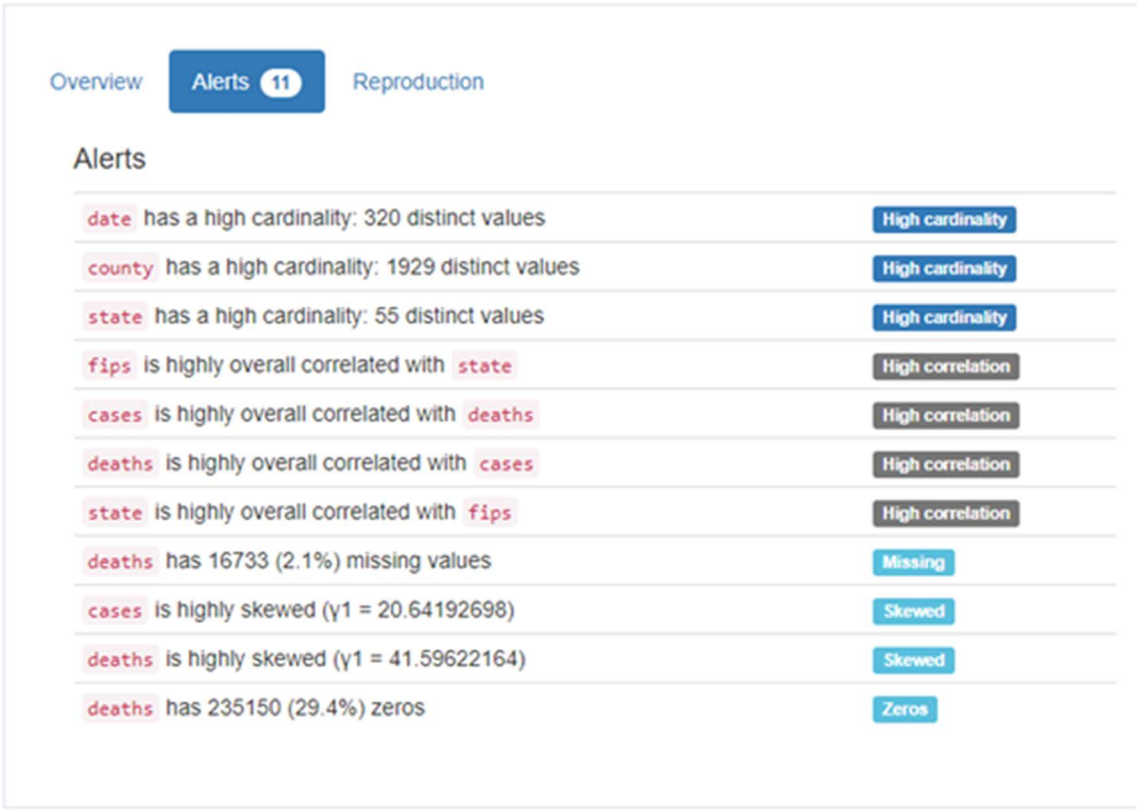
	date	county	state	fips	cases	deaths
0	2020-01-21	Snohomish	Washington	53061.0	1	0.0
1	2020-01-22	Snohomish	Washington	53061.0	1	0.0
2	2020-01-23	Snohomish	Washington	53061.0	1	0.0
3	2020-01-24	Cook	Illinois	17031.0	1	0.0
4	2020-01-24	Snohomish	Washington	53061.0	1	0.0

```
In [13]: df_ride.profile_report()
```

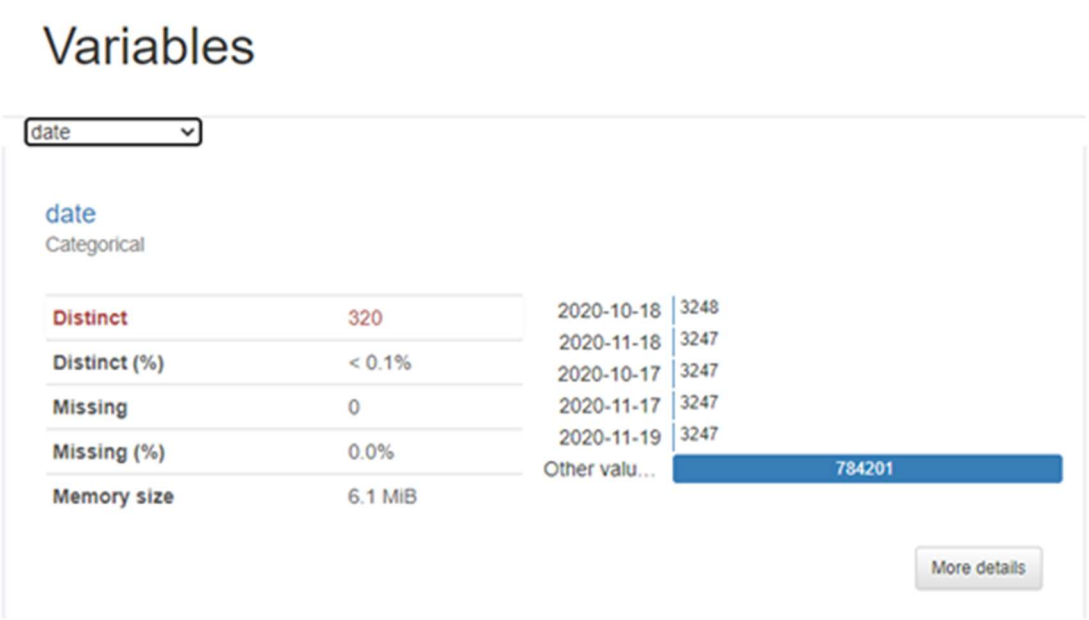
- Resumen General (overview)

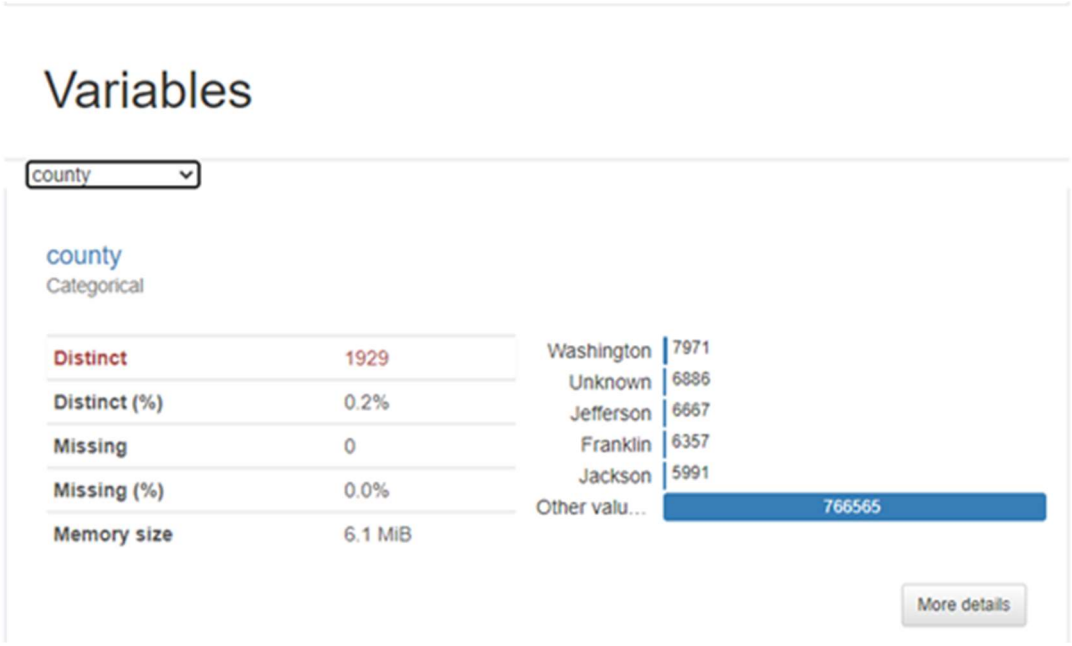


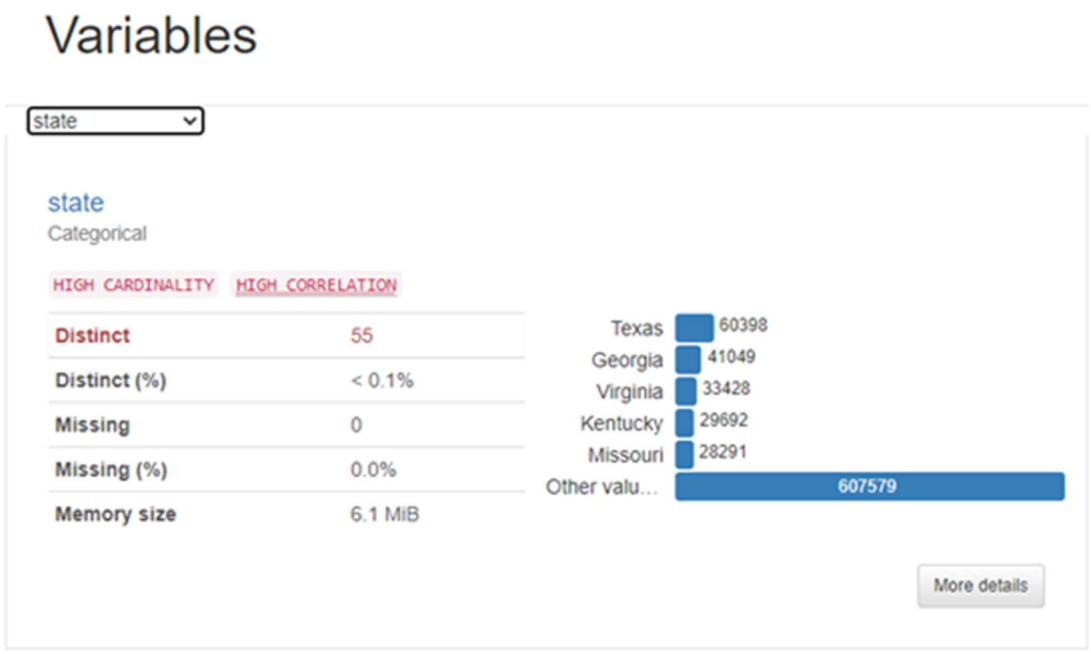
- Alertas (Alerts)



- Variables (de todas las variables que sean relevantes, no más de 10)




Variables
county
county
Categorical
Distinct 1929
Distinct (%) 0.2%
Missing 0
Missing (%) 0.0%
Memory size 6.1 MiB
Washington 7971
Unknown 6886
Jefferson 6667
Franklin 6357
Jackson 5991
Other valu... 766565
More details


Variables
state
state
Categorical
HIGH CARDINALITY
HIGH CORRELATION
Distinct 55
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 6.1 MiB
Texas 60398
Georgia 41049
Virginia 33428
Kentucky 29692
Missouri 28291
Other valu... 607579
More details

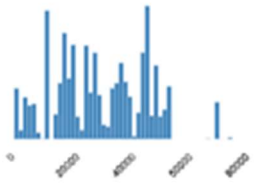
Variables

fips

fips

Real number (ℝ)

Distinct	3217	**Minimum**	1001
Distinct (%)	0.4%	**Maximum**	78030
Missing	7591	**Zeros**	0
Missing (%)	0.9%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	31238.199	**Memory size**	6.1 MiB



More details

Variables

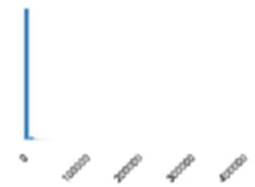
cases

cases

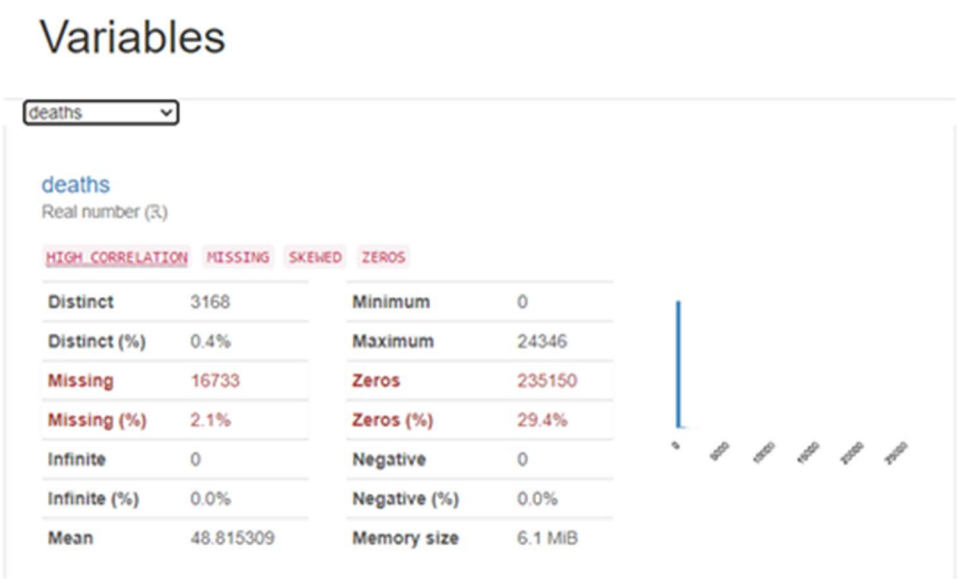
Real number (ℝ)

HIGH CORRELATION SKEWED

Distinct	25785	**Minimum**	0
Distinct (%)	3.2%	**Maximum**	439538
Missing	0	**Zeros**	664
Missing (%)	0.0%	**Zeros (%)**	0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1589.6674	**Memory size**	6.1 MiB



More details

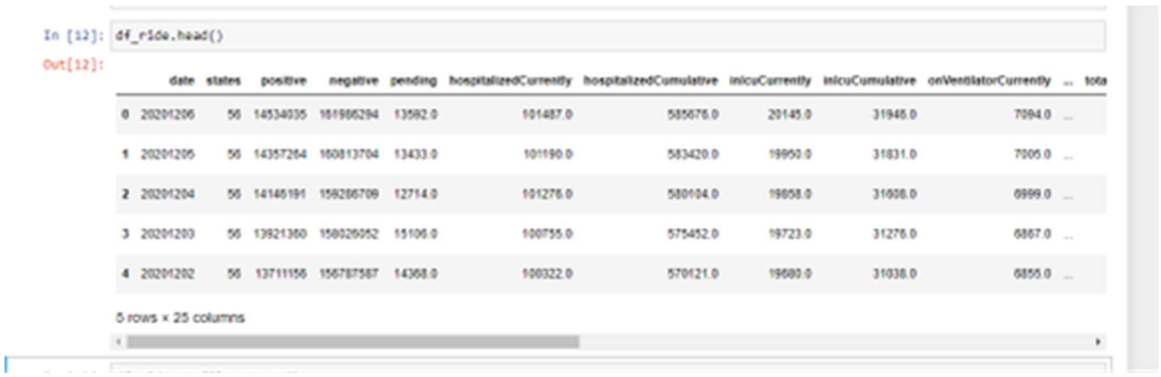


Explicación del resultado

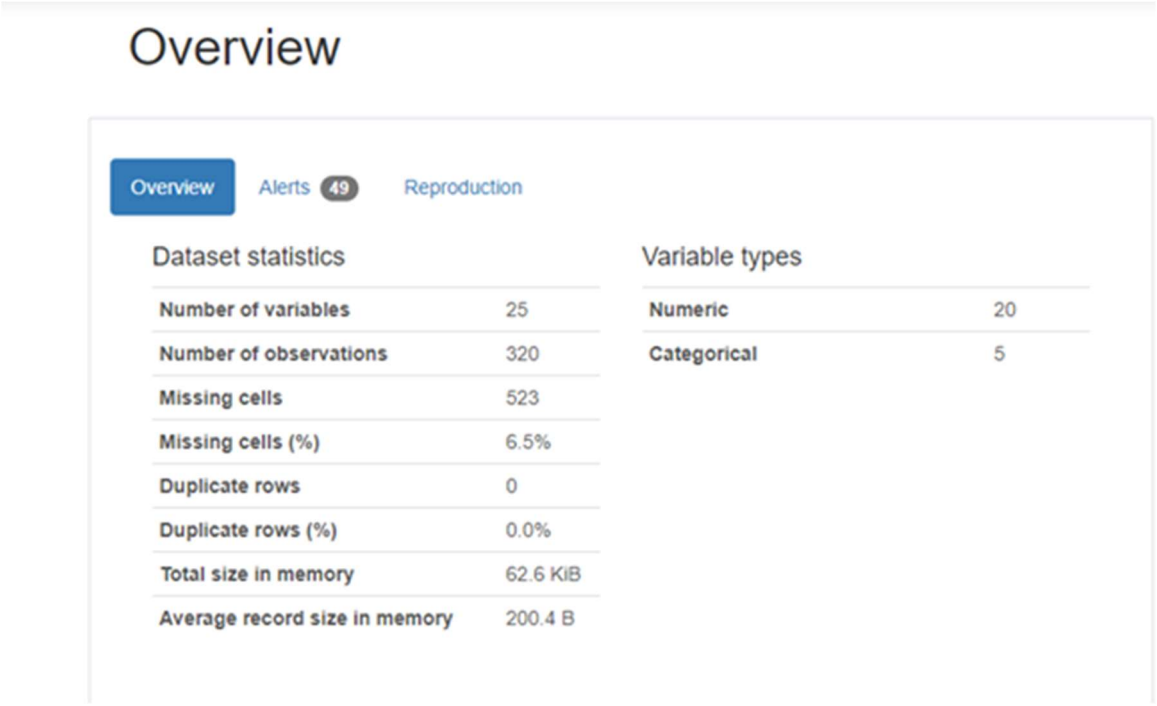
En el presente fichero se cuenta con 6 variables, de las cuales 3 son categóricas y 3 numéricas. No se cuenta con duplicados y 0.5% missing cells. Así mismo, hay 4 variables que están altamente correlacionadas como son los confirmados, muertes, Estado y Fips (Código del estado).

Nombre del Archivo: us_covid19_daily - Carpeta COVID USA

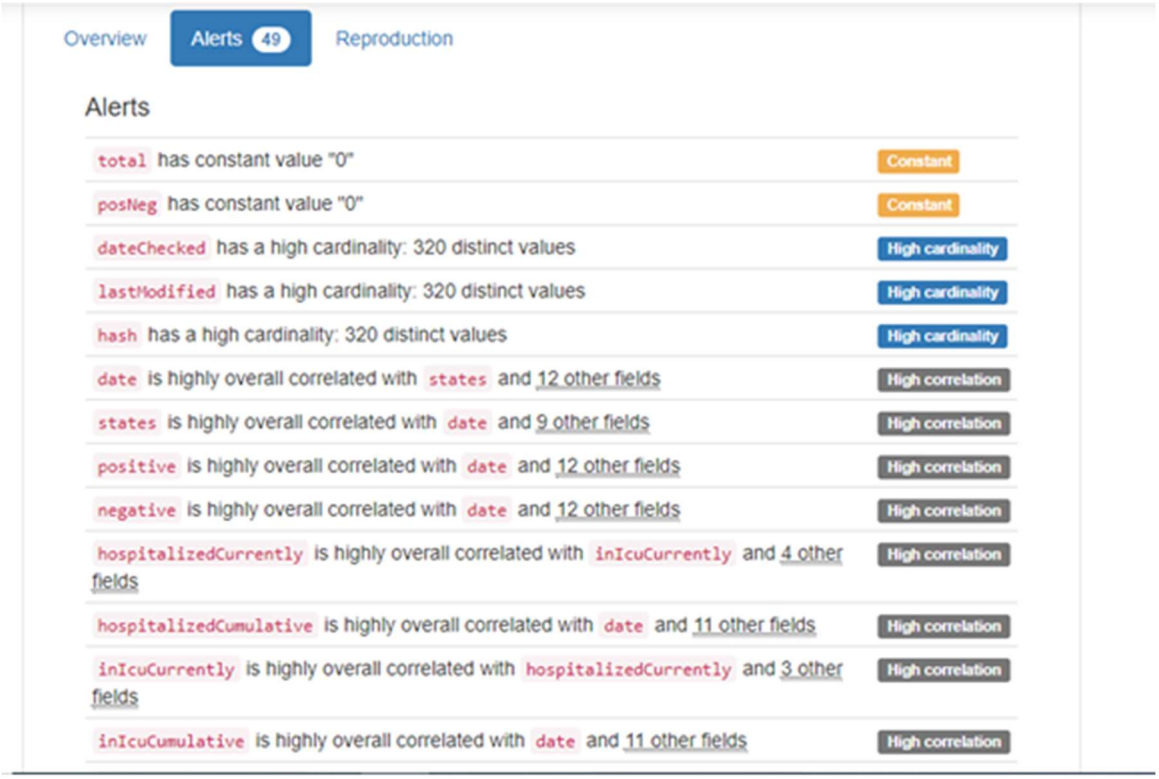
- Ejemplo de Datos (head)



- Resumen General (overview)



- Alertas (Alerts)



`hospitalizedCumulative` is highly overall correlated with `date` and 11 other fields	High correlation
`inIcuCurrently` is highly overall correlated with `hospitalizedCurrently` and 3 other fields	High correlation
`inIcuCumulative` is highly overall correlated with `date` and 11 other fields	High correlation
`onVentilatorCurrently` is highly overall correlated with `hospitalizedCurrently` and 2 other fields	High correlation
`onVentilatorCumulative` is highly overall correlated with `date` and 11 other fields	High correlation
`recovered` is highly overall correlated with `date` and 11 other fields	High correlation
`death` is highly overall correlated with `date` and 12 other fields	High correlation
`hospitalized` is highly overall correlated with `date` and 11 other fields	High correlation
`totalTestResults` is highly overall correlated with `date` and 12 other fields	High correlation
`deathIncrease` is highly overall correlated with `states` and 5 other fields	High correlation
`hospitalizedIncrease` is highly overall correlated with `states` and 4 other fields	High correlation
`negativeIncrease` is highly overall correlated with `date` and 12 other fields	High correlation
`positiveIncrease` is highly overall correlated with `date` and 15 other fields	High correlation
`totalTestResultsIncrease` is highly overall correlated with `date` and 12 other fields	High correlation
`pending` has 42 (13.1%) missing values	Missing
`hospitalizedCurrently` has 55 (17.2%) missing values	Missing

`hospitalizedCurrently` has 55 (17.2%) missing values	Missing
`hospitalizedCumulative` has 42 (13.1%) missing values	Missing
`inIcuCurrently` has 64 (20.0%) missing values	Missing
`inIcuCumulative` has 63 (19.7%) missing values	Missing
`onVentilatorCurrently` has 63 (19.7%) missing values	Missing
`onVentilatorCumulative` has 70 (21.9%) missing values	Missing
`recovered` has 63 (19.7%) missing values	Missing
`death` has 19 (5.9%) missing values	Missing
`hospitalized` has 42 (13.1%) missing values	Missing
`dateChecked` is uniformly distributed	Uniform
`lastModified` is uniformly distributed	Uniform
`hash` is uniformly distributed	Uniform
`date` has unique values	Unique
`dateChecked` has unique values	Unique
`lastModified` has unique values	Unique
`hash` has unique values	Unique
`positive` has 38 (11.9%) zeros	Zeros

`lastModified` is uniformly distributed	Uniform
`hash` is uniformly distributed	Uniform
`date` has unique values	Unique
`dateChecked` has unique values	Unique
`lastModified` has unique values	Unique
`hash` has unique values	Unique
`positive` has 38 (11.9%) zeros	Zeros
`negative` has 38 (11.9%) zeros	Zeros
`death` has 16 (5.0%) zeros	Zeros
`deathIncrease` has 36 (11.2%) zeros	Zeros
`hospitalizedIncrease` has 45 (14.1%) zeros	Zeros
`negativeIncrease` has 38 (11.9%) zeros	Zeros
`positiveIncrease` has 38 (11.9%) zeros	Zeros
`totalTestResultsIncrease` has 15 (4.7%) zeros	Zeros

- Variables (de todas las variables que sean relevantes, no más de 10)

Variables

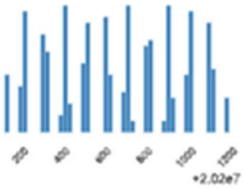
date

date

Real number (ℝ)

HIGH CORRELATION UNIQUE

Distinct	320	Minimum	20200122
Distinct (%)	100.0%	Maximum	20201206
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	20200661	Memory size	2.6 KiB



More details

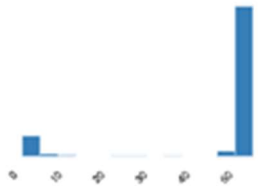
Variables

states

states

Real number (ℝ)

Distinct	13	**Minimum**	2
Distinct (%)	4.1%	**Maximum**	56
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	48.853125	**Memory size**	2.6 KiB



More details

Variables

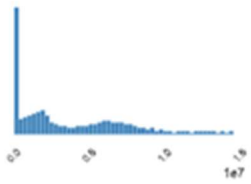
positive

positive

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	283	**Minimum**	0
Distinct (%)	88.4%	**Maximum**	14534035
Missing	0	**Zeros**	38
Missing (%)	0.0%	**Zeros (%)**	11.9%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	3963366.1	**Memory size**	2.6 KiB



More details

Variables

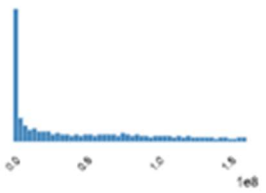
negative

negative

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	283	**Minimum**	0
Distinct (%)	88.4%	**Maximum**	1.6198629×10^8
Missing	0	**Zeros**	38
Missing (%)	0.0%	**Zeros (%)**	11.9%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	47071503	**Memory size**	2.6 KiB



Variables

hospitalizedCumulative

hospitalizedCumulative

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	275	**Minimum**	4
Distinct (%)	98.9%	**Maximum**	585676
Missing	42	**Zeros**	0
Missing (%)	13.1%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	276808.1	**Memory size**	2.6 KiB

More details

Variables

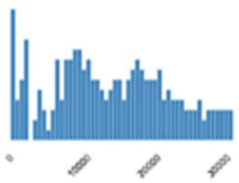
inIcuCumulative

inIcuCumulative

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	257	**Minimum**	74
Distinct (%)	100.0%	**Maximum**	31946
Missing	63	**Zeros**	0
Missing (%)	19.7%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	14374.735	**Memory size**	2.6 KiB



More details

Variables

onVentilatorCumulative

onVentilatorCumulative

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	234	**Minimum**	32
Distinct (%)	93.6%	**Maximum**	3322
Missing	70	**Zeros**	0
Missing (%)	21.9%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1553.636	**Memory size**	2.6 KiB

More details

Variables

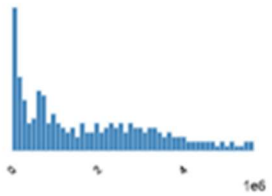
recovered

recovered

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	257	**Minimum**	97
Distinct (%)	100.0%	**Maximum**	5624444
Missing	63	**Zeros**	0
Missing (%)	19.7%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1782062.5	**Memory size**	2.6 KiB



More details

Variables

death

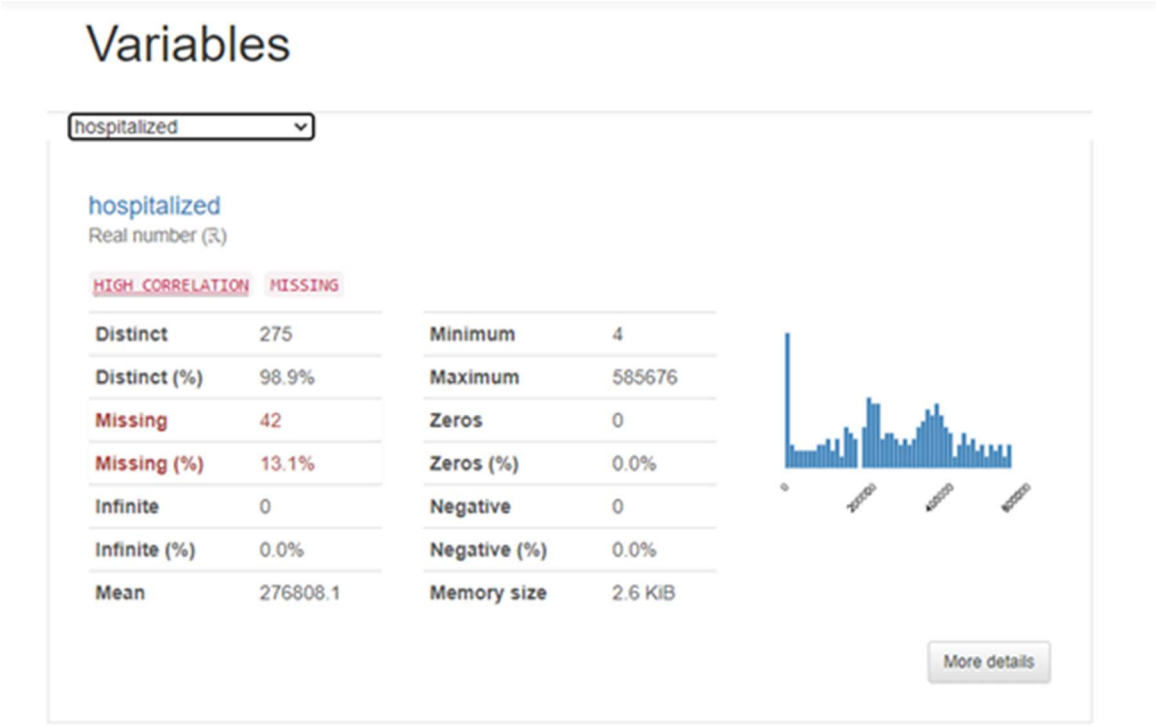
death

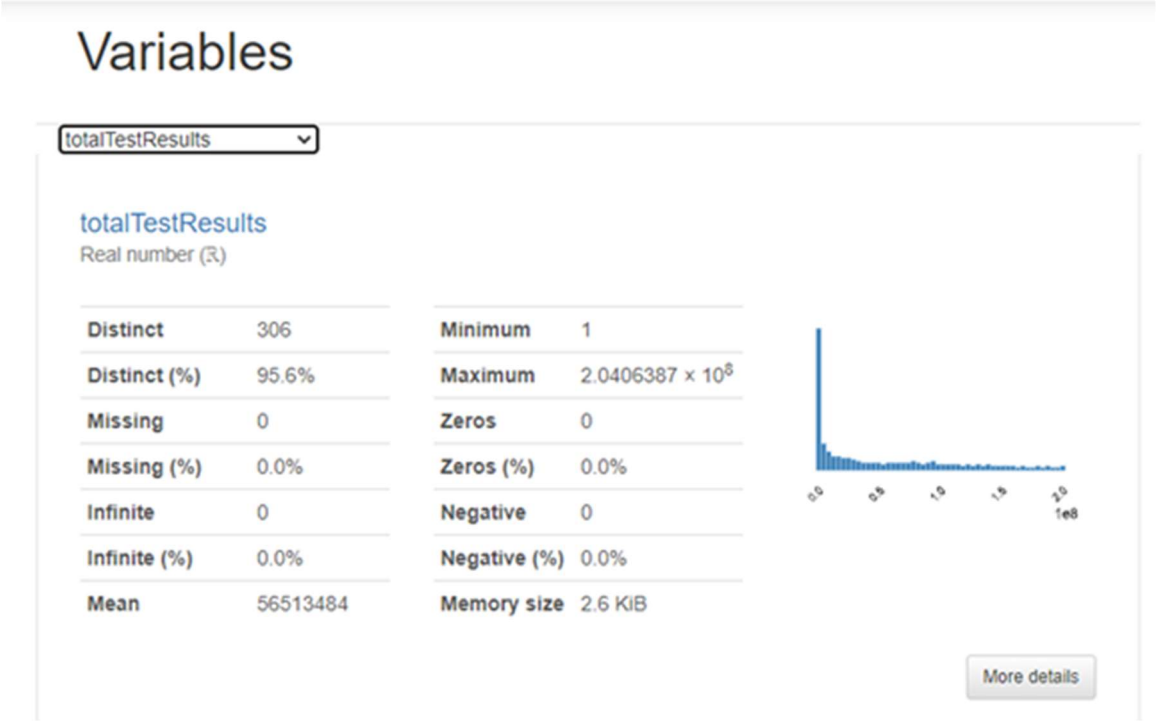
Real number (ℝ)

HIGH CORRELATION MISSING ZEROS

Distinct	285	**Minimum**	0
Distinct (%)	94.7%	**Maximum**	273374
Missing	19	**Zeros**	16
Missing (%)	5.9%	**Zeros (%)**	5.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	122746.14	**Memory size**	2.6 KiB

More details





Explicación del resultado

En el presente fichero se cuenta con 25 variables, de las cuales 5 son categóricas y 20 numéricas. No se cuenta con duplicados y 6.5% missing cells. Así mismo, hay 15 variables que están altamente correlacionadas.

Nombre del Archivo: us_states_covid19_daily - Carpeta COVID USA

- Ejemplo de Datos (head)

```
In [7]: df_ride.head()
Out[7]:
```

	date	state	positive	probableCases	negative	pending	totalTestResultsSource	totalTestResults	hospitalizedCurrently	hospitalizedCumulative	...	po
0	20201206	AK	35720.0	NaN	1042056.0	NaN	totalTestsViral	1077776.0	164.0	799.0	...	107
1	20201206	AL	268877.0	45962.0	1421126.0	NaN	totalTestsPeopleViral	1645041.0	1927.0	26331.0	...	169
2	20201206	AR	170924.0	22753.0	1614979.0	NaN	totalTestsViral	1763150.0	1076.0	9401.0	...	178
3	20201206	AS	0.0	NaN	2140.0	NaN	totalTestsViral	2140.0	NaN	NaN	...	
4	20201206	AZ	364276.0	12590.0	2018813.0	NaN	totalTestsPeopleViral	2370489.0	2977.0	28248.0	...	238

5 rows × 55 columns

- Resumen General (overview)
-

Overview | Alerts 107 | Reproduction

Dataset statistics

Number of variables	55
Number of observations	15633
Missing cells	326265
Missing cells (%)	37.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	6.6 MiB
Average record size in memory	440.0 B

Variable types

Numeric	41
Categorical	13
Unsupported	1

- Variables (de todas las variables que sean relevantes, no más de 10)

date

Real number (ℝ)

Distinct	320
Distinct (%)	2.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	20200727

state

Categorical

HIGH CARDINALITY HIGH CORRELATION UNIFORM

Distinct	56
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	122.3 KiB

positive

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	12299
Distinct (%)	79.4%
Missing	152
Missing (%)	1.0%
Infinite	0
Infinite (%)	0.0%
Mean	81924.756

probableCases

Real number (ℝ)

HIGH CORRELATION MISSING ZEROS

Distinct	3432
Distinct (%)	63.0%
Missing	10184
Missing (%)	65.1%
Infinite	0
Infinite (%)	0.0%
Mean	5132.6115

negative

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	13027
Distinct (%)	85.0%
Missing	310
Missing (%)	2.0%
Infinite	0
Infinite (%)	0.0%
Mean	983024.27

pending

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	769
Distinct (%)	45.7%
Missing	13949
Missing (%)	89.2%
Infinite	0
Infinite (%)	0.0%
Mean	1492.7185

totalTestResults

Real number (ℝ)

Distinct	13975
Distinct (%)	89.6%
Missing	35
Missing (%)	0.2%
Infinite	0
Infinite (%)	0.0%
Mean	1159399.6
Minimum	0

hospitalizedCurrently

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	2922
Distinct (%)	23.3%
Missing	3117
Missing (%)	19.9%
Infinite	0
Infinite (%)	0.0%
Mean	942.66771

hospitalizedCumulative

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	5457
Distinct (%)	57.8%
Missing	6199
Missing (%)	39.7%
Infinite	0
Infinite (%)	0.0%
Mean	8156.9486

onVentilatorCurrently

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	645
Distinct (%)	10.4%
Missing	9422
Missing (%)	60.3%
Infinite	0
Infinite (%)	0.0%
Mean	130.17872

recovered

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	7329
Distinct (%)	66.0%
Missing	4522
Missing (%)	28.9%
Infinite	0
Infinite (%)	0.0%
Mean	41219.517

death

Real number (ℝ)

HIGH CORRELATION MISSING ZEROS

Distinct	5194
Distinct (%)	35.1%
Missing	826
Missing (%)	5.3%
Infinite	0
Infinite (%)	0.0%
Mean	2495.2109

hospitalized

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	5457
Distinct (%)	57.8%
Missing	6199
Missing (%)	39.7%
Infinite	0
Infinite (%)	0.0%

Explicación del resultado

En el presente fichero se cuenta con 55 variables, de las cuales 13 son categóricas, 41 numéricas y 1 unsupported. No se cuenta con duplicados y 37.9% missing cells.

Nombre del Archivo : covid_19_data - Carpeta MUNDIAL

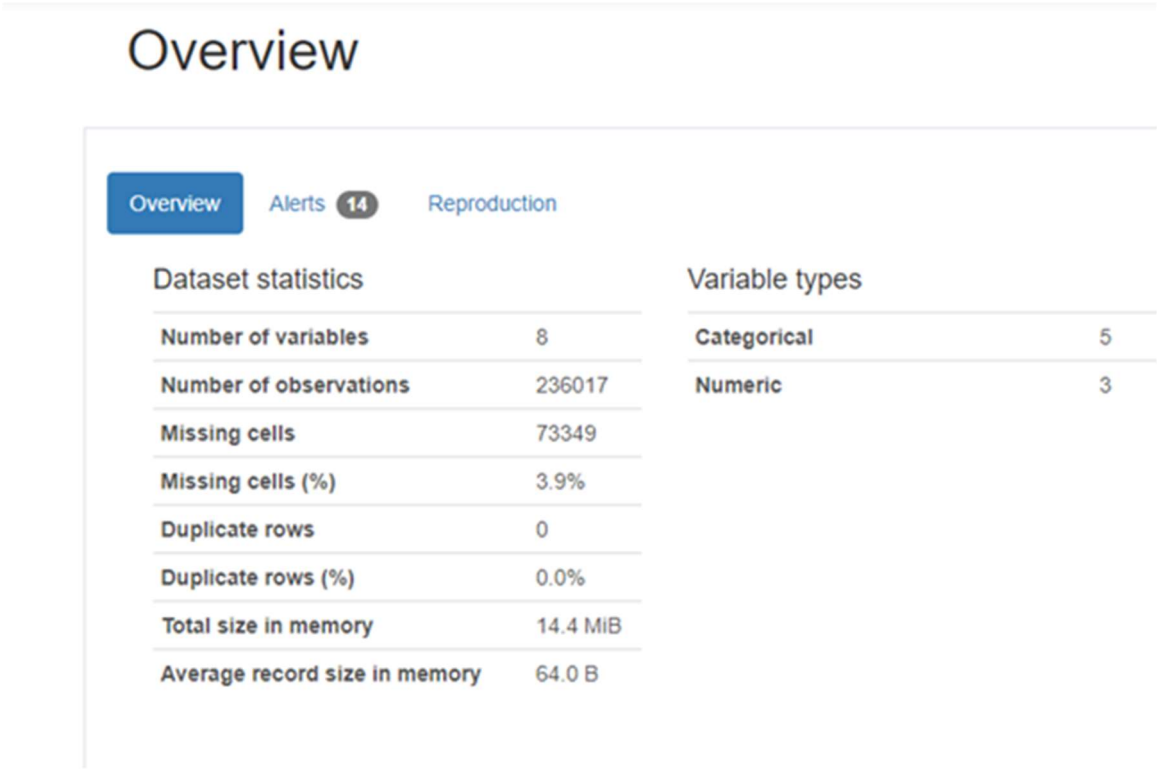
Ejemplo de Datos (head)

```
In [12]: df_ride.head()
```

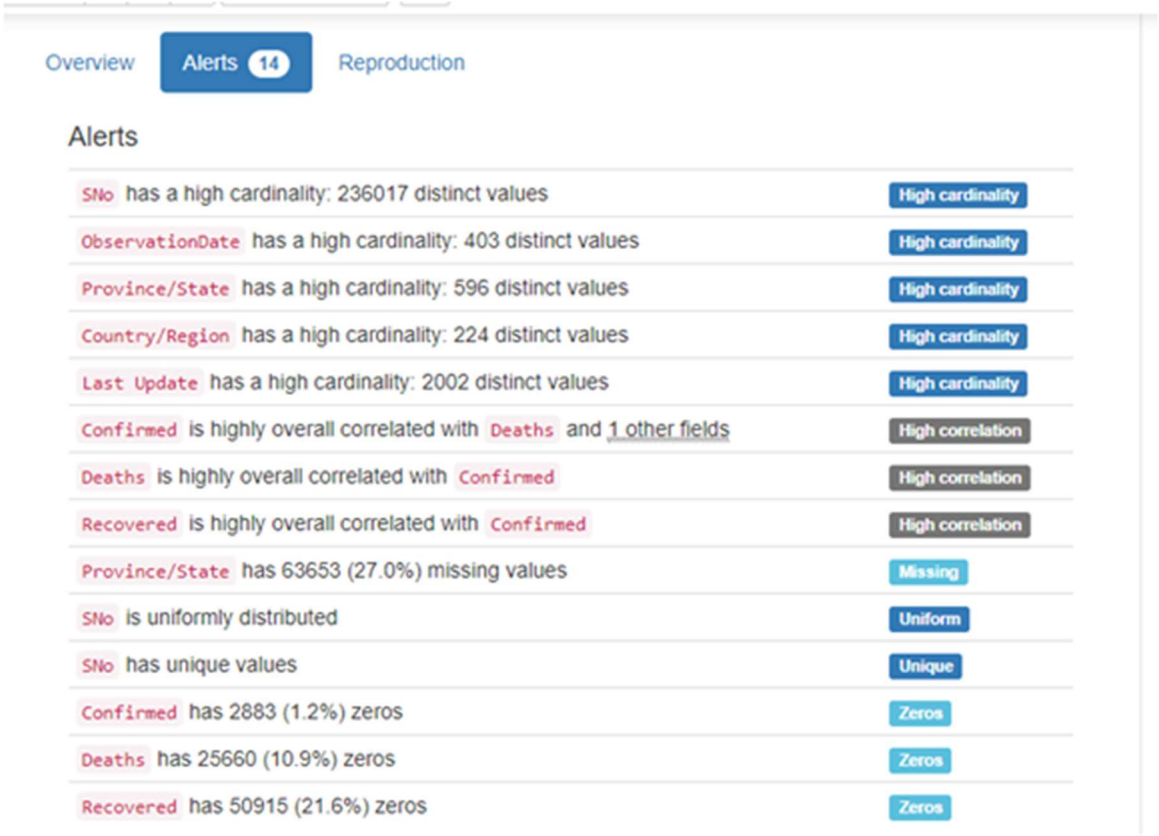
Out[12]:

	SNo	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered
0	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
1	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0
2	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0
3	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0
4	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0

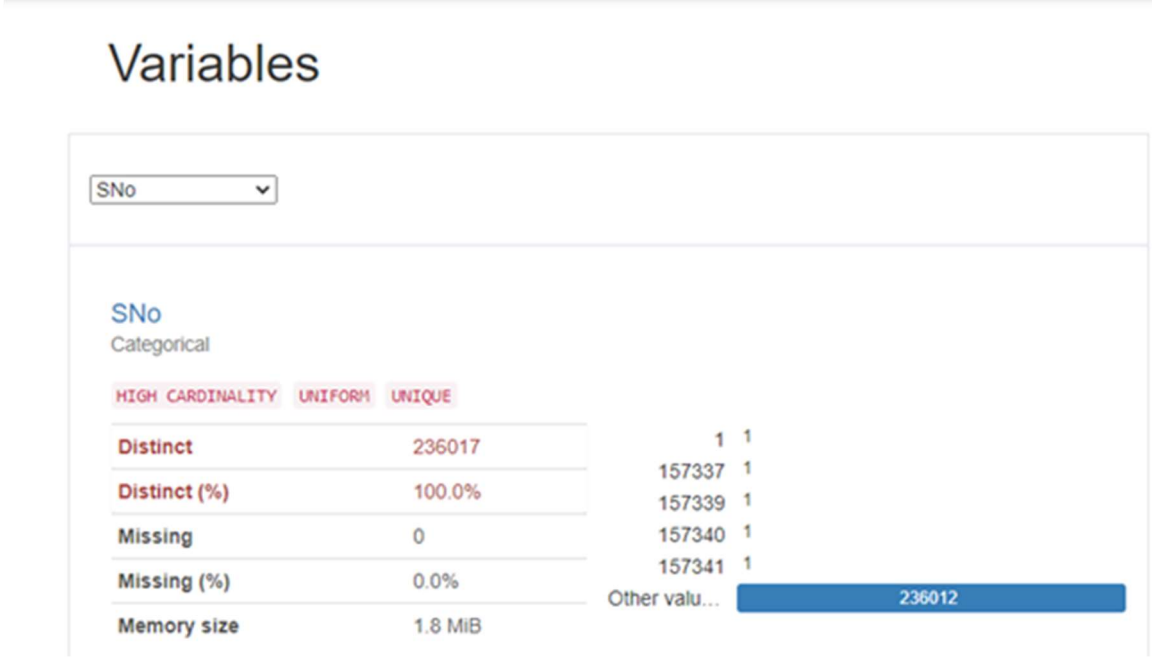
Resumen General (overview)

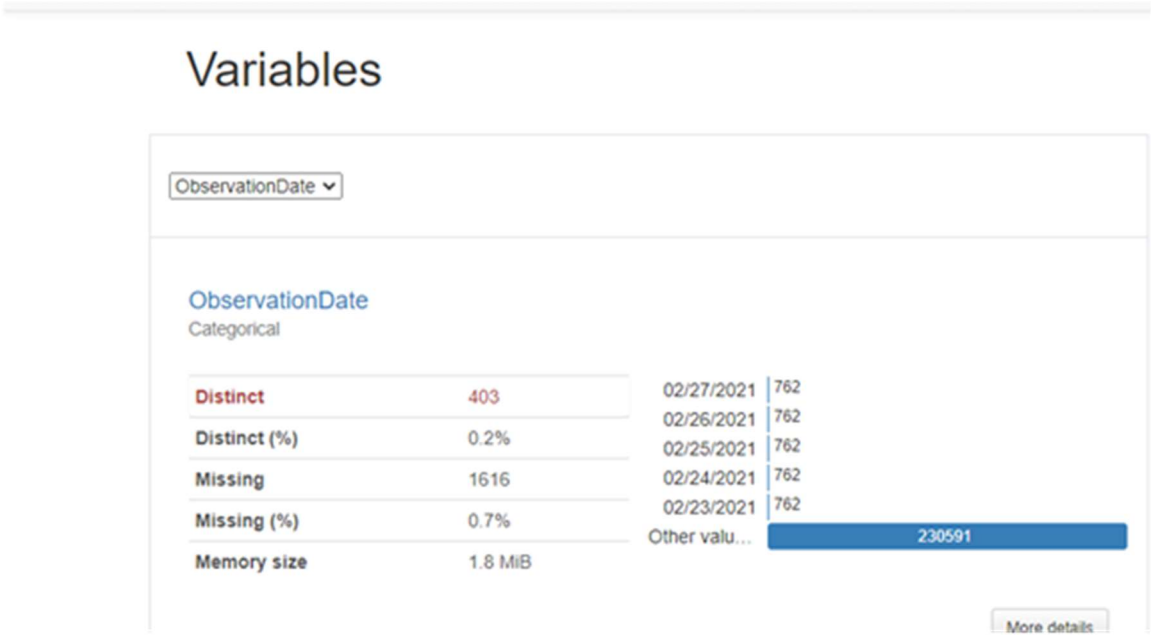


Alertas (Alerts)

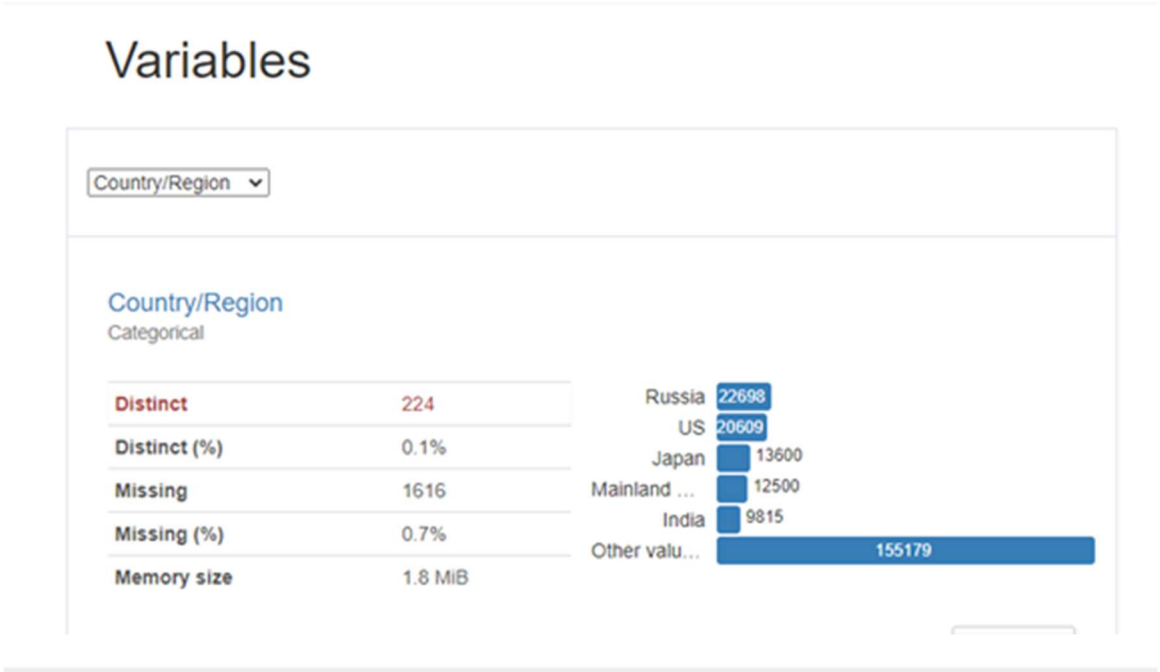


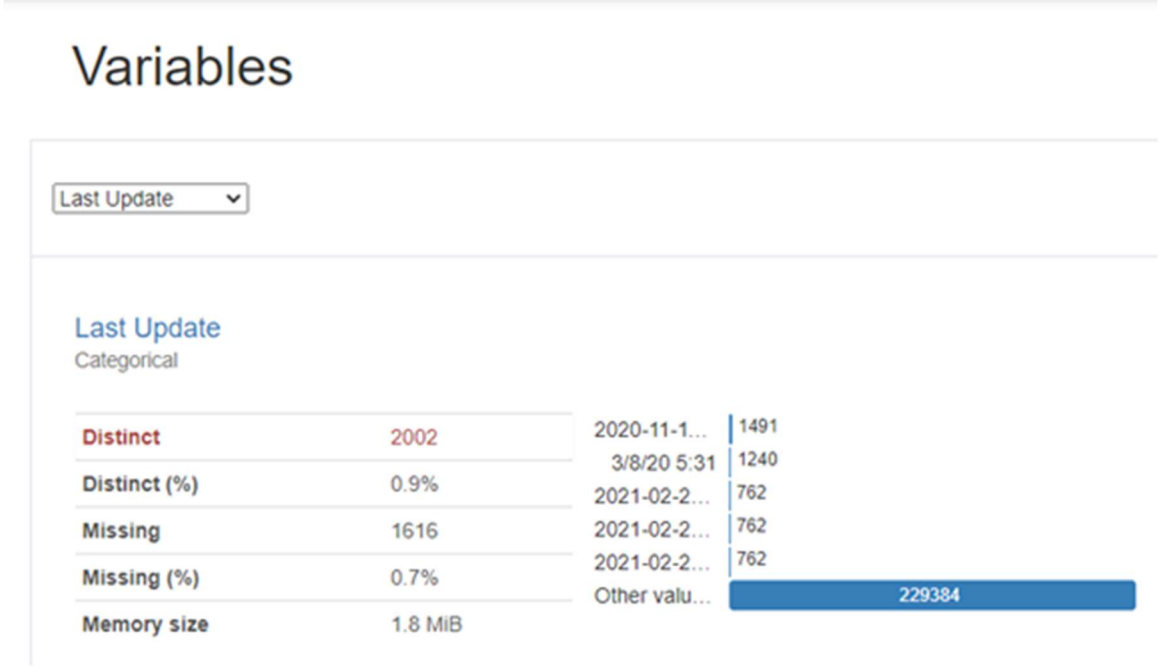
Variables (de todas las variables que sean relevantes, no más de 10)




Variables
ObservationDate
ObservationDate
Categorical
Distinct 403
Distinct (%) 0.2%
Missing 1616
Missing (%) 0.7%
Memory size 1.8 MiB
02/27/2021 762
02/26/2021 762
02/25/2021 762
02/24/2021 762
02/23/2021 762
Other valu... 230591
More details


Variables
Province/State
Province/State
Categorical
HIGH CARDINALITY
MISSING
Distinct 596
Distinct (%) 0.3%
Missing 63653
Missing (%) 27.0%
Memory size 1.8 MiB
Unknown 2637
Amazonas 836
Diamond ... 742
Grand Pri... 700
Punjab 526
Other valu... 166923


Variables
Country/Region
Country/Region
Categorical
Distinct 224
Distinct (%) 0.1%
Missing 1616
Missing (%) 0.7%
Memory size 1.8 MiB
Russia 22698
US 20609
Japan 13600
Mainland ... 12500
India 9815
Other valu... 155179


Variables
Last Update
Last Update
Categorical
Distinct 2002
Distinct (%) 0.9%
Missing 1616
Missing (%) 0.7%
Memory size 1.8 MiB
2020-11-1... 1491
3/8/20 5:31 1240
2021-02-2... 762
2021-02-2... 762
2021-02-2... 762
Other valu... 229384

Variables

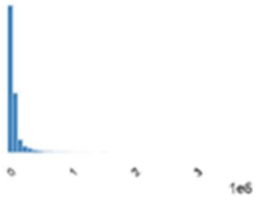
Confirmed

Confirmed

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	78146	**Minimum**	-302844
Distinct (%)	33.3%	**Maximum**	3664050
Missing	1616	**Zeros**	2883
Missing (%)	0.7%	**Zeros (%)**	1.2%
Infinite	0	**Negative**	1
Infinite (%)	0.0%	**Negative (%)**	< 0.1%
Mean	57551.918	**Memory size**	1.8 MiB



Variables

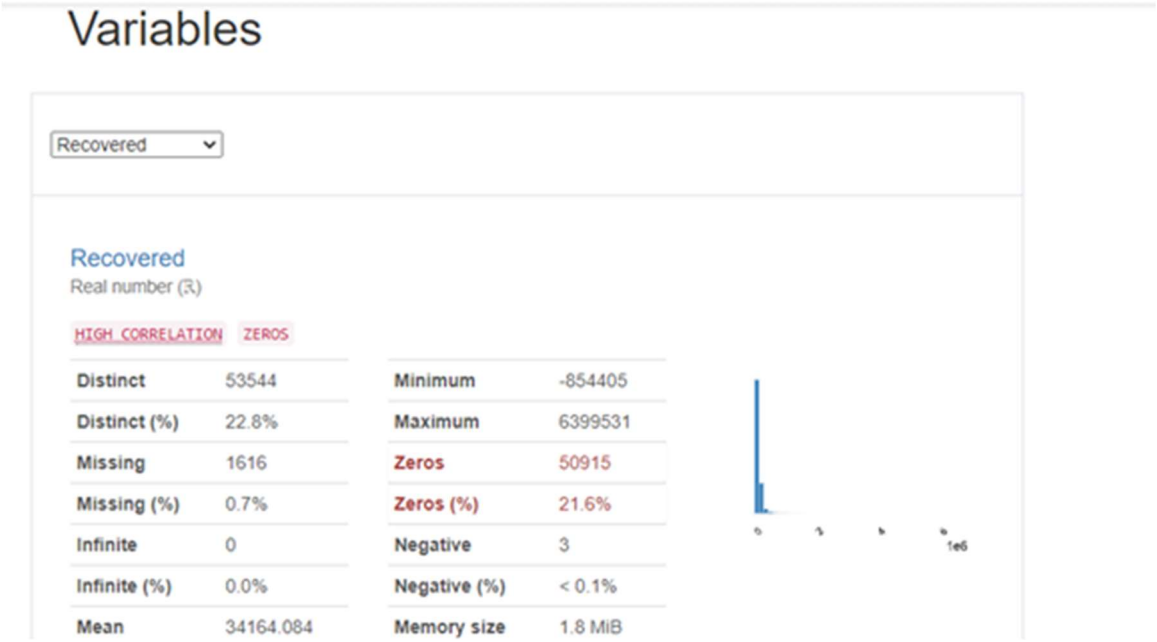
Deaths

Deaths

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	15099	**Minimum**	-178
Distinct (%)	6.4%	**Maximum**	108208
Missing	1616	**Zeros**	25660
Missing (%)	0.7%	**Zeros (%)**	10.9%
Infinite	0	**Negative**	2
Infinite (%)	0.0%	**Negative (%)**	< 0.1%
Mean	1497.9736	**Memory size**	1.8 MiB



Explicación del resultado

En el presente fichero se cuenta con 8 variables, de las cuales 5 son categóricas y 3 numéricas. No se cuenta con duplicados y 3.9% missing cells. Así mismo, hay 3 variables que están altamente correlacionadas como son los confirmados, muertes y recuperados.

Jose

- **EU_Cases**
- Ejemplo de Datos

	country	country_code	year_week	age_group	new_cases	population	rate_14_day_per_100k	source
0	Austria	AT	2020-01	<15yr	NaN	1285488	NaN	TESSy COVID-19
1	Austria	AT	2020-02	<15yr	NaN	1285488	NaN	TESSy COVID-19
2	Austria	AT	2020-03	<15yr	NaN	1285488	NaN	TESSy COVID-19
3	Austria	AT	2020-04	<15yr	NaN	1285488	NaN	TESSy COVID-19
4	Austria	AT	2020-05	<15yr	NaN	1285488	NaN	TESSy COVID-19

- Resumen General

Overview

Overview | Alerts 15 | Reproduction

Dataset statistics

Number of variables	8
Number of observations	29058
Missing cells	2964
Missing cells (%)	1.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.8 MiB
Average record size in memory	64.0 B

Variable types

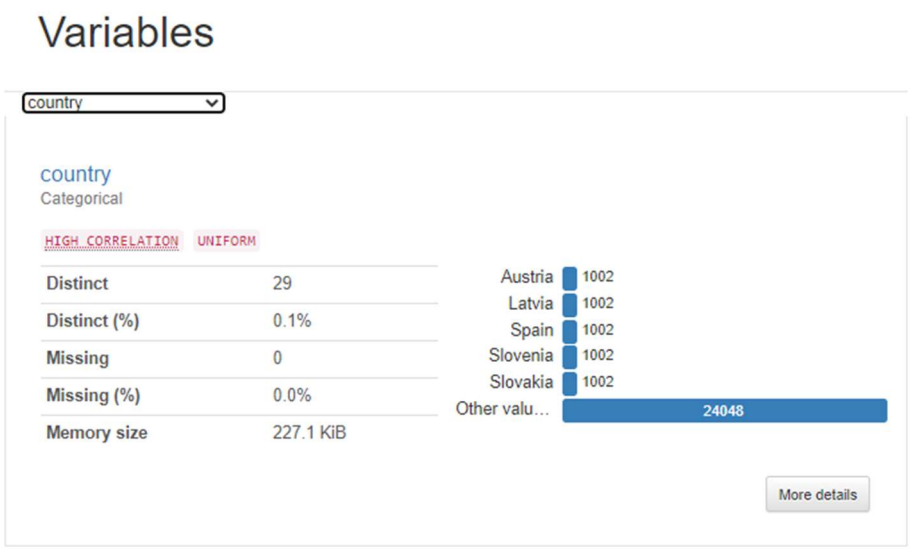
Categorical	5
Numeric	3

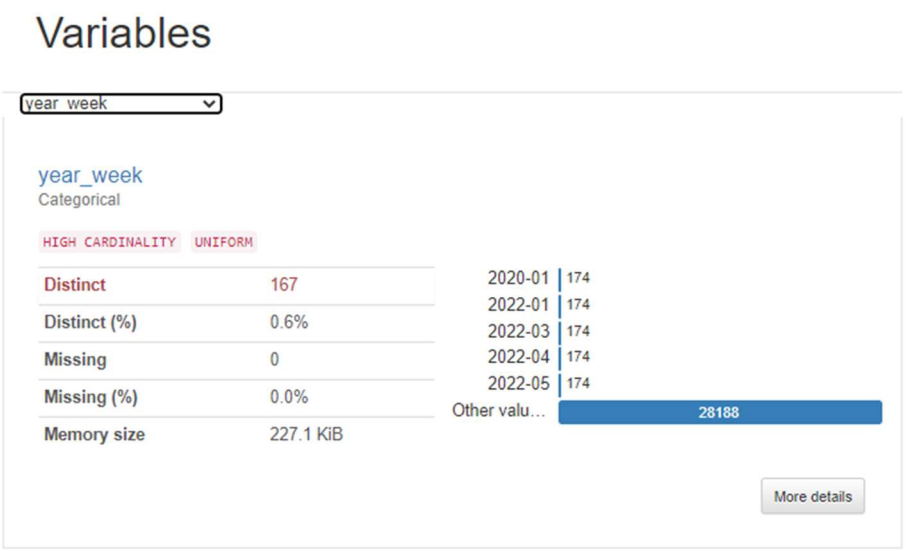
- Alertas

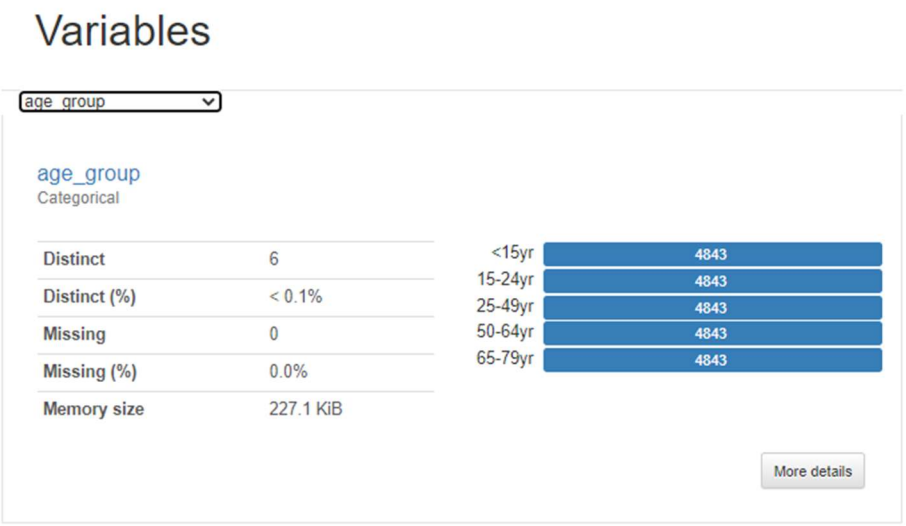
Alerts

`source` has constant value "TESSy COVID-19"	Constant
`year_week` has a high cardinality: 167 distinct values	High cardinality
`new_cases` is highly overall correlated with `population` and 1 other fields	High correlation
`population` is highly overall correlated with `new_cases`	High correlation
`rate_14_day_per_100k` is highly overall correlated with `new_cases`	High correlation
`country` is highly overall correlated with `country_code`	High correlation
`country_code` is highly overall correlated with `country`	High correlation
`new_cases` has 1358 (4.7%) missing values	Missing
`rate_14_day_per_100k` has 1606 (5.5%) missing values	Missing
`country` is uniformly distributed	Uniform
`country_code` is uniformly distributed	Uniform
`year_week` is uniformly distributed	Uniform
`age_group` is uniformly distributed	Uniform
`new_cases` has 881 (3.0%) zeros	Zeros
`rate_14_day_per_100k` has 526 (1.8%) zeros	Zeros

- Variables


Variables
country
country
Categorical
HIGH CORRELATION
UNIFORM
Distinct 29
Distinct (%) 0.1%
Missing 0
Missing (%) 0.0%
Memory size 227.1 KiB
Austria 1002
Latvia 1002
Spain 1002
Slovenia 1002
Slovakia 1002
Other valu… 24048
More details


Variables
year_week
year_week
Categorical
HIGH CARDINALITY
UNIFORM
Distinct 167
Distinct (%) 0.6%
Missing 0
Missing (%) 0.0%
Memory size 227.1 KiB
2020-01 174
2022-01 174
2022-03 174
2022-04 174
2022-05 174
Other valu… 28188
More details


Variables
age_group
age_group
Categorical
Distinct 6
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 227.1 KiB
<15yr 4843
15-24yr 4843
25-49yr 4843
50-64yr 4843
65-79yr 4843
More details

Variables

new_cases

new_cases

Real number (ℝ)

HIGH CORRELATION MISSING ZEROS

Distinct	8814	Minimum	0
Distinct (%)	31.8%	Maximum	967879
Missing	1358	Zeros	881
Missing (%)	4.7%	Zeros (%)	3.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	6696.7175	Memory size	227.1 KiB

More details

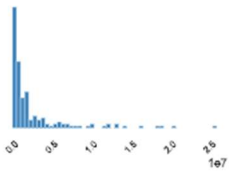
Variables

population

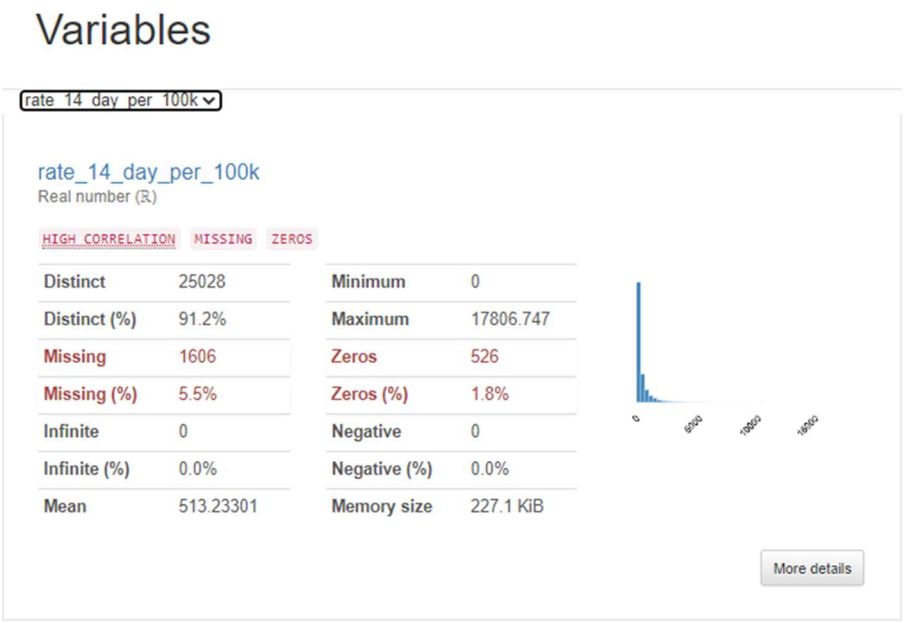
population

Real number (ℝ)

Distinct	174	Minimum	1609
Distinct (%)	0.6%	Maximum	25855651
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2563736.5	Memory size	227.1 KiB



More details



Explicación del resultado

Observamos que el fichero consta de 8 variables, de las cuales 5 son categóricas y 3 numéricas. No se observan duplicados, si 2964 casos nulos. Así mismo, hay 5 variables que están altamente correlacionadas

- **EU_Cases_And_Deaths_14d**
- Ejemplo de Datos

	country	country_code	continent	population	indicator	weekly_count	year_week	rate_14_day	cumulative_count	source	note
0	Austria	AUT	Europe	8932664	cases	NaN	2020-01	NaN	NaN	TESSy COVID-19	NaN
1	Austria	AUT	Europe	8932664	cases	NaN	2020-02	NaN	NaN	TESSy COVID-19	NaN
2	Austria	AUT	Europe	8932664	cases	NaN	2020-03	NaN	NaN	TESSy COVID-19	NaN
3	Austria	AUT	Europe	8932664	cases	NaN	2020-04	NaN	NaN	TESSy COVID-19	NaN
4	Austria	AUT	Europe	8932664	cases	NaN	2020-05	NaN	NaN	TESSy COVID-19	NaN

- Resumen General

Overview

Overview | Alerts 22 | Reproduction

Dataset statistics

Number of variables	11
Number of observations	10354
Missing cells	12155
Missing cells (%)	10.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	889.9 KiB
Average record size in memory	88.0 B

Variable types

Categorical	6
Numeric	4
Unsupported	1

- Alertas

Alerts

`continent` has constant value "Europe"	Constant
`source` has constant value "TESSy COVID-19"	Constant
`year_week` has a high cardinality: 167 distinct values	High cardinality
`population` is highly overall correlated with `country_code`	High correlation
`weekly_count` is highly overall correlated with `rate_14_day` and 1 other fields	High correlation
`rate_14_day` is highly overall correlated with `weekly_count` and 1 other fields	High correlation
`cumulative_count` is highly overall correlated with `weekly_count` and 1 other fields	High correlation
`country` is highly overall correlated with `country_code`	High correlation
`country_code` is highly overall correlated with `population` and 1 other fields	High correlation
`country_code` has 334 (3.2%) missing values	Missing
`weekly_count` has 460 (4.4%) missing values	Missing
`rate_14_day` has 547 (5.3%) missing values	Missing
`cumulative_count` has 460 (4.4%) missing values	Missing
`note` has 10354 (100.0%) missing values	Missing
`country` is uniformly distributed	Uniform
`country_code` is uniformly distributed	Uniform
`indicator` is uniformly distributed	Uniform
`year_week` is uniformly distributed	Uniform
`note` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`weekly_count` has 561 (5.4%) zeros	Zeros
`rate_14_day` has 396 (3.8%) zeros	Zeros
`cumulative_count` has 151 (1.5%) zeros	Zeros

- Variables

Variables

country

country

Categorical

HIGH CORRELATION UNIFORM

Distinct	31
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	81.0 KiB

More details

Variables

population

population

Real number (ℝ)

Distinct	60	**Minimum**	39055
Distinct (%)	0.6%	**Maximum**	4.530067×10^8
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	28742687	**Memory size**	81.0 KiB

More details

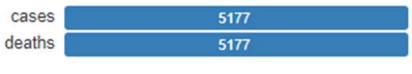
Variables

indicator

indicator

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	81.0 KiB



More details

Variables

weekly_count

weekly_count

Real number (ℝ)

HIGH CORRELATION MISSING ZEROS

Distinct	4657	Minimum	0
Distinct (%)	47.1%	Maximum	9066537
Missing	460	Zeros	561
Missing (%)	4.4%	Zeros (%)	5.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	37797.388	Memory size	81.0 KiB

More details

Variables

year_week

year_week

Categorical

HIGH CARDINALITY UNIFORM

Distinct	167
Distinct (%)	1.6%
Missing	0
Missing (%)	0.0%
Memory size	81.0 KiB

More details

Variables

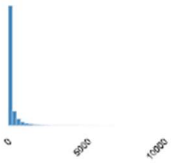
rate_14_day

rate_14_day

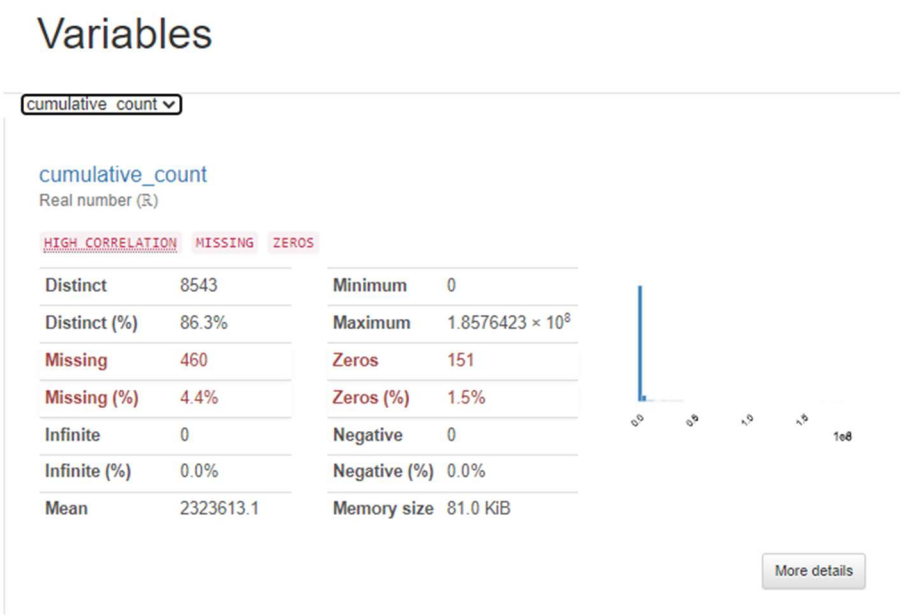
Real number (ℝ)

HIGH CORRELATION MISSING ZEROS

Distinct	8168	Minimum	0
Distinct (%)	83.3%	Maximum	13728.646
Missing	547	Zeros	396
Missing (%)	5.3%	Zeros (%)	3.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	286.60006	Memory size	81.0 KiB



More details



Explicación del resultado

Observamos que el fichero consta de 11 variables, de las cuales 6 son categóricas, 4 numéricas y 1 no reconocible. No se observan duplicados, si 12155 casos nulos. Así mismo, hay 6 variables que están altamente correlacionadas

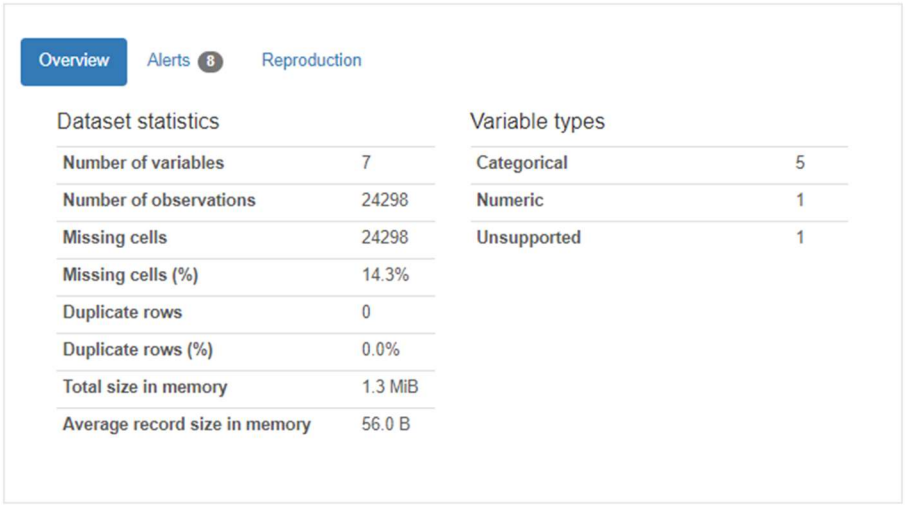
- **EU_Hospital_And_ICU**

- Ejemplo de Datos

	country	indicator	date	year_week	value	source	url
0	Austria	Daily hospital occupancy	2020-04-01	2020-W14	856.0	Country_Website	NaN
1	Austria	Daily hospital occupancy	2020-04-02	2020-W14	823.0	Country_Website	NaN
2	Austria	Daily hospital occupancy	2020-04-03	2020-W14	829.0	Country_Website	NaN
3	Austria	Daily hospital occupancy	2020-04-04	2020-W14	826.0	Country_Website	NaN
4	Austria	Daily hospital occupancy	2020-04-05	2020-W14	712.0	Country_Website	NaN

- Resumen General

Overview



Overview | Alerts 8 | Reproduction

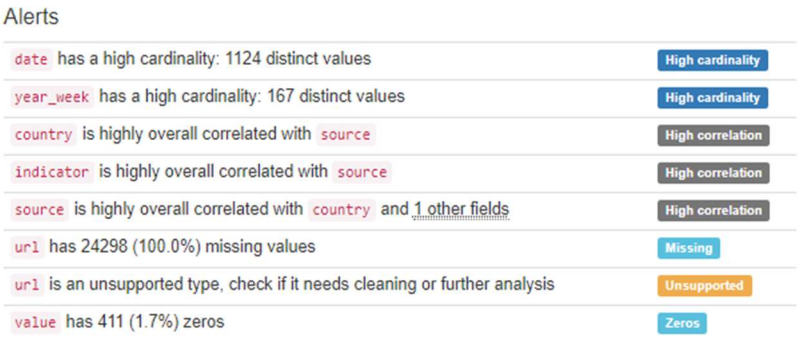
Dataset statistics

Number of variables	7
Number of observations	24298
Missing cells	24298
Missing cells (%)	14.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.3 MiB
Average record size in memory	56.0 B

Variable types

Categorical	5
Numeric	1
Unsupported	1

- Alertas

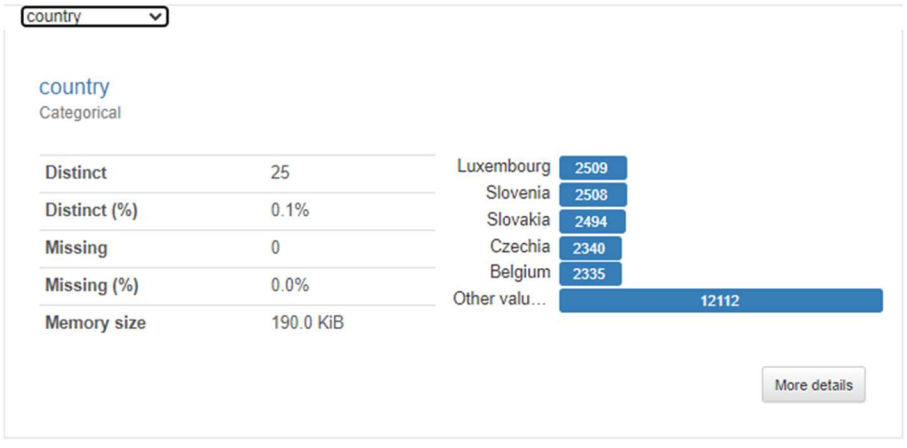


Alerts

`date` has a high cardinality: 1124 distinct values	High cardinality
`year_week` has a high cardinality: 167 distinct values	High cardinality
`country` is highly overall correlated with `source`	High correlation
`indicator` is highly overall correlated with `source`	High correlation
`source` is highly overall correlated with `country` and 1 other fields	High correlation
`url` has 24298 (100.0%) missing values	Missing
`url` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`value` has 411 (1.7%) zeros	Zeros

- Variables

Variables



country

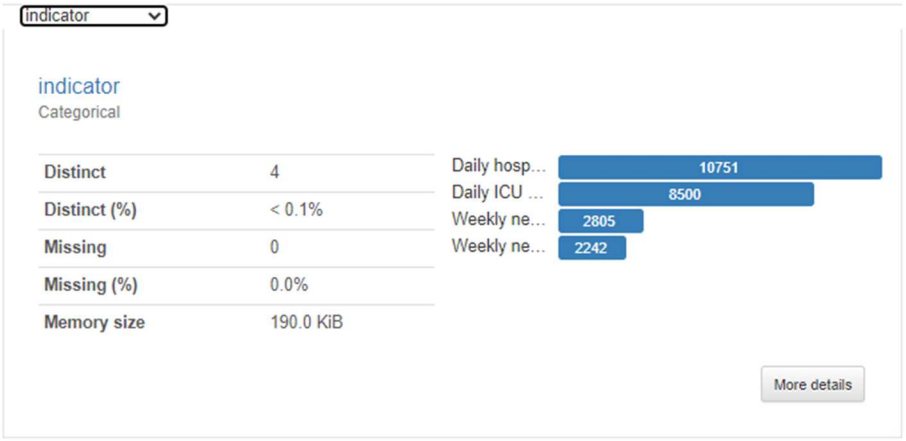
country
Categorical

Distinct	25
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	190.0 KiB

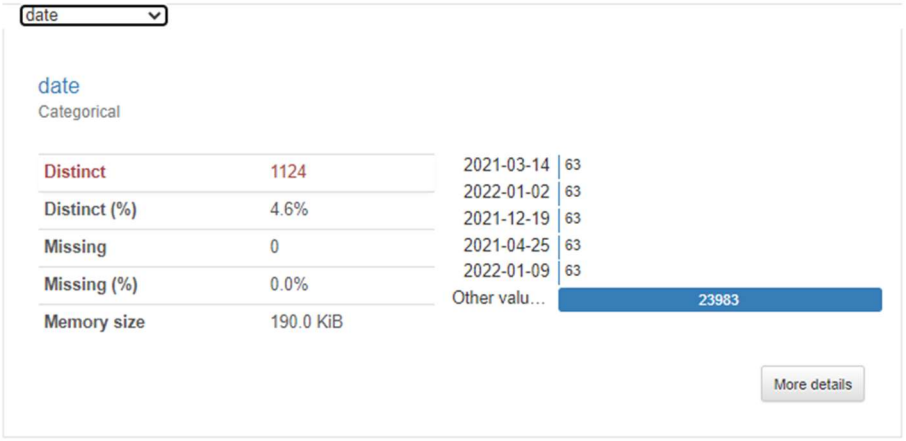
Luxembourg 2509
Slovenia 2508
Slovakia 2494
Czechia 2340
Belgium 2335
Other valu… 12112

More details

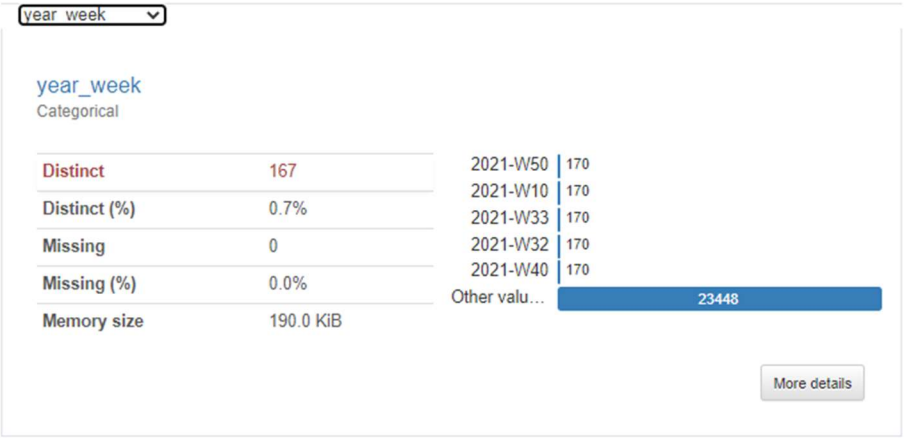
Variables

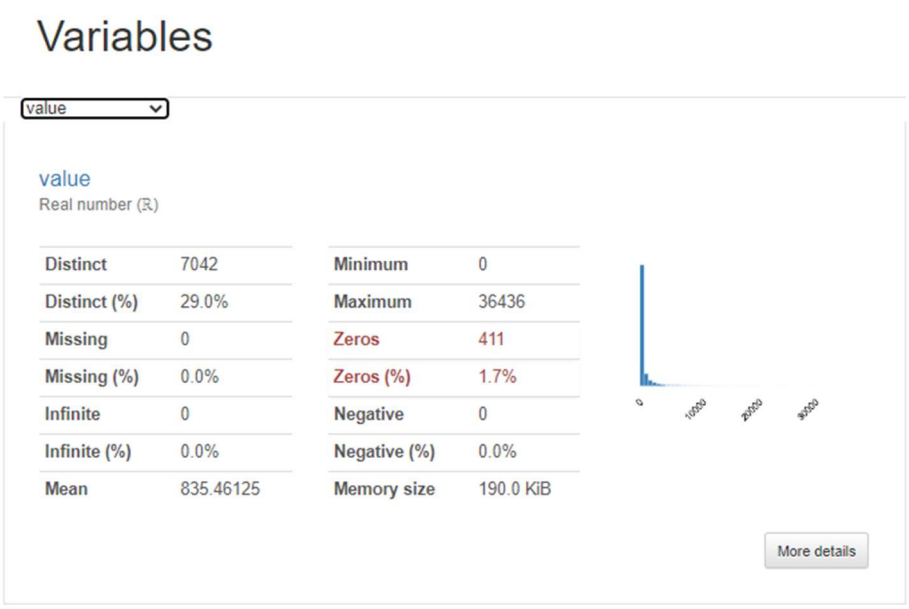


Variables



Variables





Explicación del resultado

Observamos que el fichero consta de 7 variables, de las cuales 5 son categóricas, 1 numérica y 1 no reconocible. No se observan duplicados, si 24296 casos nulos. Así mismo, hay 3 variables que están altamente correlacionadas

- **EU_Testing**

- Ejemplo de Datos

	country	country_code	year_week	level	region	region_name	new_cases	tests_done	population	testing_rate	positivity_rate	testing_data_source
0	Austria	AT	2020-W01	national	AT	Austria	NaN	NaN	8932664	NaN	NaN	NaN
1	Austria	AT	2020-W02	national	AT	Austria	NaN	NaN	8932664	NaN	NaN	NaN
2	Austria	AT	2020-W03	national	AT	Austria	NaN	NaN	8932664	NaN	NaN	NaN
3	Austria	AT	2020-W04	national	AT	Austria	NaN	NaN	8932664	NaN	NaN	NaN
4	Austria	AT	2020-W05	national	AT	Austria	NaN	NaN	8932664	NaN	NaN	NaN

- Resumen General

Overview

Overview | Alerts 21 | Reproduction

Dataset statistics

Number of variables	12
Number of observations	5011
Missing cells	4026
Missing cells (%)	6.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	469.9 KiB
Average record size in memory	96.0 B

Variable types

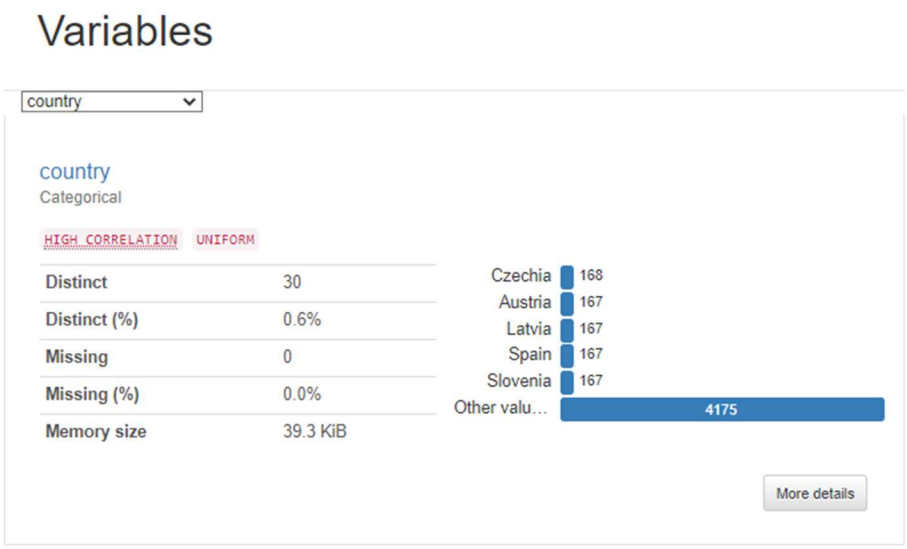
Categorical	7
Numeric	5

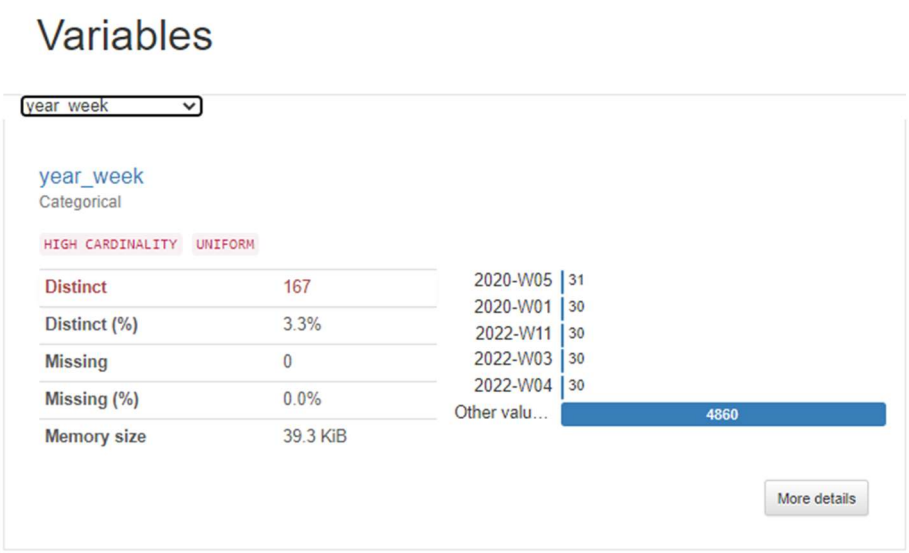
- Alertas

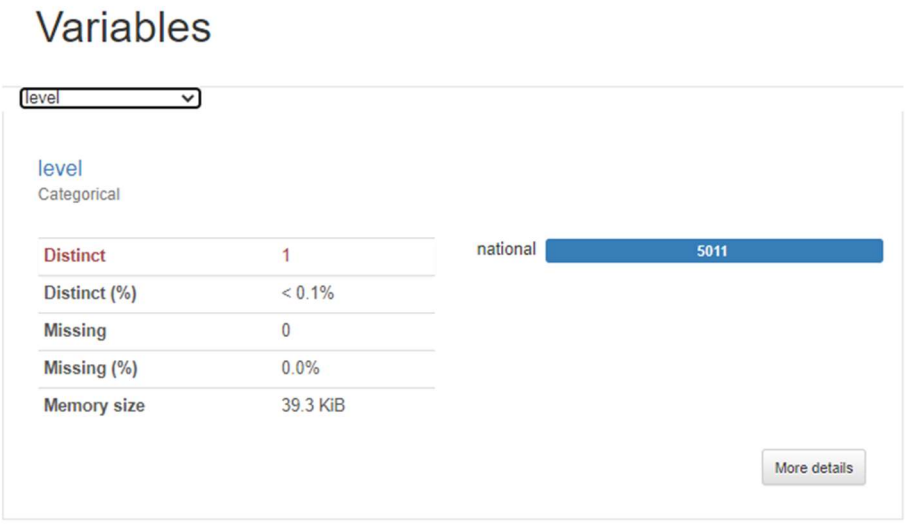
Alerts

`level` has constant value "national"	Constant
`year_week` has a high cardinality: 167 distinct values	High cardinality
`new_cases` is highly overall correlated with `tests_done` and 1 other fields	High correlation
`tests_done` is highly overall correlated with `new_cases` and 2 other fields	High correlation
`population` is highly overall correlated with `new_cases` and 5 other fields	High correlation
`testing_rate` is highly overall correlated with `tests_done`	High correlation
`country` is highly overall correlated with `population` and 3 other fields	High correlation
`country_code` is highly overall correlated with `population` and 3 other fields	High correlation
`region` is highly overall correlated with `population` and 3 other fields	High correlation
`region_name` is highly overall correlated with `population` and 3 other fields	High correlation
`testing_data_source` is highly imbalanced (94.0%)	Imbalance
`new_cases` has 821 (16.4%) missing values	Missing
`tests_done` has 790 (15.8%) missing values	Missing
`testing_rate` has 790 (15.8%) missing values	Missing
`positivity_rate` has 835 (16.7%) missing values	Missing
`testing_data_source` has 790 (15.8%) missing values	Missing
`country` is uniformly distributed	Uniform
`country_code` is uniformly distributed	Uniform
`year_week` is uniformly distributed	Uniform
`region` is uniformly distributed	Uniform
`region_name` is uniformly distributed	Uniform

- Variables


Variables
country
country
Categorical
HIGH CORRELATION
UNIFORM
Distinct 30
Distinct (%) 0.6%
Missing 0
Missing (%) 0.0%
Memory size 39.3 KiB
Czechia 168
Austria 167
Latvia 167
Spain 167
Slovenia 167
Other valu… 4175
More details


Variables
year_week
year_week
Categorical
HIGH CARDINALITY
UNIFORM
Distinct 167
Distinct (%) 3.3%
Missing 0
Missing (%) 0.0%
Memory size 39.3 KiB
2020-W05 31
2020-W01 30
2022-W11 30
2022-W03 30
2022-W04 30
Other valu… 4860
More details


Variables
level
level
Categorical
Distinct 1
Distinct (%) < 0.1%
Missing 0
Missing (%) 0.0%
Memory size 39.3 KiB
national 5011
More details

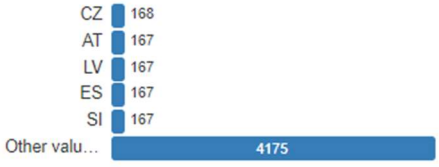
Variables

region

region
Categorical

HIGH CORRELATION UNIFORM

Distinct	30
Distinct (%)	0.6%
Missing	0
Missing (%)	0.0%
Memory size	39.3 KiB



More details

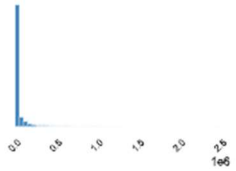
Variables

new_cases

new_cases
Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	3583	Minimum	0
Distinct (%)	85.5%	Maximum	2486551
Missing	821	Zeros	50
Missing (%)	16.4%	Zeros (%)	1.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	41890.189	Memory size	39.3 KiB



More details

Variables

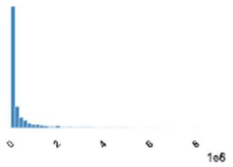
tests_done

tests_done

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	4132	Minimum	0
Distinct (%)	97.9%	Maximum	8832046
Missing	790	Zeros	14
Missing (%)	15.8%	Zeros (%)	0.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	410616.66	Memory size	39.3 KiB



More details

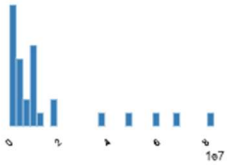
Variables

population

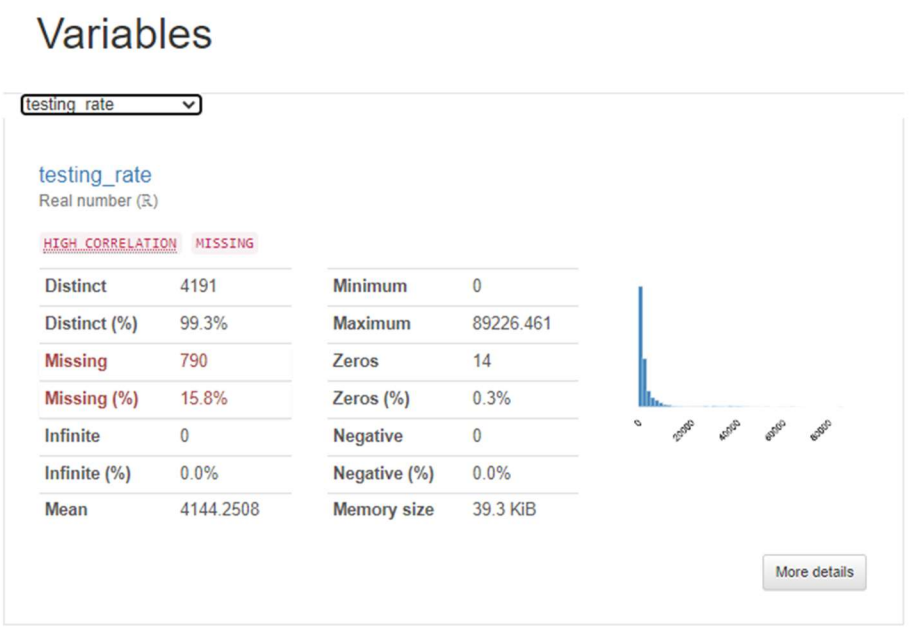
population

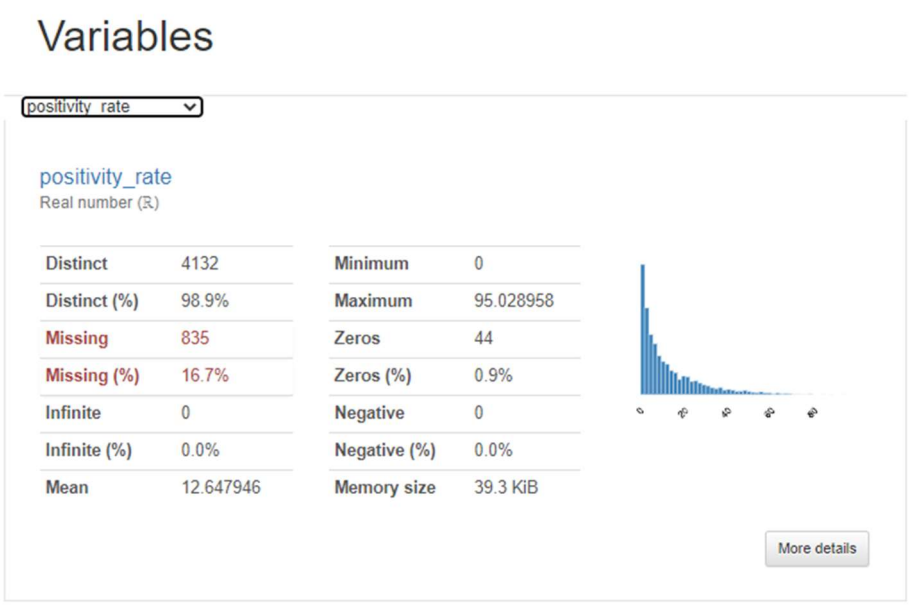
Real number (ℝ)

Distinct	30	Minimum	39055
Distinct (%)	0.6%	Maximum	83155031
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	15099346	Memory size	39.3 KiB



More details





Explicación del resultado

Observamos que el fichero consta de 12 variables, de las cuales 7 son categóricas y 5 numéricas. No se observan duplicados, si 40026 casos nulos. Así mismo, hay 8 variables que están altamente correlacionadas

- **EU_Vaccination**
- Ejemplo de Datos

	YearWeekISO	ReportingCountry	Denominator	NumberDosesReceived	NumberDosesExported	FirstDose	FirstDoseRefused	SecondDose	DoseAdditional1	U
0	2020-W53	AT	7388778.0	0.0	0.0	0	NaN	0	0	
1	2020-W53	AT	7388778.0	0.0	0.0	0	NaN	0	0	
2	2020-W53	AT	7388778.0	0.0	0.0	2	NaN	0	0	
3	2020-W53	AT	7388778.0	61425.0	0.0	5344	NaN	0	0	
4	2020-W53	AT	7388778.0	0.0	0.0	0	NaN	0	0	

UnknownDose	Region	TargetGroup	Vaccine	Population	DoseAdditional2	DoseAdditional3
0	AT	ALL	AZ	8901064	0	0
0	AT	ALL	JANSS	8901064	0	0
0	AT	ALL	COMBA.1	8901064	0	0
0	AT	ALL	COM	8901064	0	0
0	AT	ALL	NVXD	8901064	0	0

- Resumen General

Overview

Overview | Alerts 30 | Reproduction

Dataset statistics

Number of variables	16
Number of observations	659339
Missing cells	2115884
Missing cells (%)	20.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	80.5 MiB
Average record size in memory	128.0 B

Variable types

Categorical	5
Numeric	11

- Alertas

Alerts

`YearWeekISO` has a high cardinality: 120 distinct values	High cardinality
`Region` has a high cardinality: 126 distinct values	High cardinality
`FirstDose` is highly overall correlated with `SecondDose`	High correlation
`FirstDoseRefused` is highly overall correlated with `Population`	High correlation
`SecondDose` is highly overall correlated with `FirstDose` and 1 other fields	High correlation
`DoseAdditional1` is highly overall correlated with `SecondDose` and 1 other fields	High correlation
`Population` is highly overall correlated with `FirstDoseRefused` and 1 other fields	High correlation
`DoseAdditional2` is highly overall correlated with `DoseAdditional1` and 1 other fields	High correlation
`DoseAdditional3` is highly overall correlated with `DoseAdditional2`	High correlation
`ReportingCountry` is highly overall correlated with `Population`	High correlation
`Denominator` has 310402 (47.1%) missing values	Missing
`NumberDosesReceived` has 576459 (87.4%) missing values	Missing
`NumberDosesExported` has 571131 (86.6%) missing values	Missing
`FirstDoseRefused` has 657892 (99.8%) missing values	Missing
`NumberDosesReceived` is highly skewed (γ1 = 31.14776629)	Skewed
`NumberDosesExported` is highly skewed (γ1 = 77.00355641)	Skewed
`FirstDose` is highly skewed (γ1 = 63.86394664)	Skewed
`SecondDose` is highly skewed (γ1 = 68.6765303)	Skewed
`DoseAdditional1` is highly skewed (γ1 = 100.401705)	Skewed
`UnknownDose` is highly skewed (γ1 = 229.6305681)	Skewed
`DoseAdditional2` is highly skewed (γ1 = 74.20472884)	Skewed
`DoseAdditional3` is highly skewed (γ1 = 111.1857112)	Skewed

`NumberDosesReceived` has 75964 (11.5%) zeros	Zeros
`NumberDosesExported` has 88018 (13.3%) zeros	Zeros
`FirstDose` has 459175 (69.6%) zeros	Zeros
`SecondDose` has 486576 (73.8%) zeros	Zeros
`DoseAdditional1` has 521546 (79.1%) zeros	Zeros
`UnknownDose` has 650845 (98.7%) zeros	Zeros
`DoseAdditional2` has 573218 (86.9%) zeros	Zeros
`DoseAdditional3` has 631090 (95.7%) zeros	Zeros

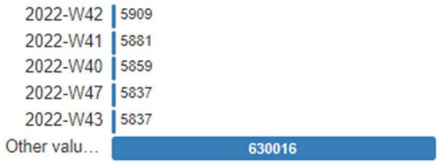
- Variables

Variables

YearWeekISO

YearWeekISO
Categorical

Distinct	120
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.0 MiB



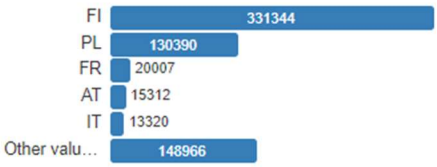
More details

Variables

ReportingCountry

ReportingCountry
Categorical

Distinct	30
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.0 MiB



More details

Variables

Denominator

Denominator
Real number (ℝ)

Distinct	626	Minimum	981
Distinct (%)	0.2%	Maximum	69411087
Missing	310402	Zeros	0
Missing (%)	47.1%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1491709.9	Memory size	5.0 MiB

More details

Variables

NumberDosesReceived

NumberDosesReceived

Real number (ℝ)

MISSING SKEWED ZEROS

Distinct	2261	Minimum	0
Distinct (%)	2.7%	Maximum	18892800
Missing	576459	Zeros	75964
Missing (%)	87.4%	Zeros (%)	11.5%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	17603.202	Memory size	5.0 MiB

More details

Variables

NumberDosesExported

NumberDosesExported

Real number (ℝ)

MISSING SKEWED ZEROS

Distinct	139	Minimum	0
Distinct (%)	0.2%	Maximum	6488820
Missing	571131	Zeros	88018
Missing (%)	86.6%	Zeros (%)	13.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1019.996	Memory size	5.0 MiB

More details

Variables

FirstDose

FirstDose

Real number (ℝ)

HIGH CORRELATION SKEWED ZEROS

Distinct	18528	Minimum	0
Distinct (%)	2.8%	Maximum	4021725
Missing	0	Zeros	459175
Missing (%)	0.0%	Zeros (%)	69.6%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1415.3767	Memory size	5.0 MiB

More details

Variables

FirstDoseRefused

FirstDoseRefused

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	23	Minimum	0
Distinct (%)	1.6%	Maximum	73
Missing	657892	Zeros	1377
Missing (%)	99.8%	Zeros (%)	0.2%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.40152039	Memory size	5.0 MiB

More details

Variables

SecondDose

SecondDose

Real number (ℝ)

HIGH CORRELATION SKEWED ZEROS

Distinct	17429	Minimum	0
Distinct (%)	2.6%	Maximum	4149195
Missing	0	Zeros	486576
Missing (%)	0.0%	Zeros (%)	73.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1288.261	Memory size	5.0 MiB

More details

Variables

DoseAdditional1

DoseAdditional1

Real number (ℝ)

HIGH CORRELATION SKEWED ZEROS

Distinct	12886	Minimum	0
Distinct (%)	2.0%	Maximum	6647762
Missing	0	Zeros	521546
Missing (%)	0.0%	Zeros (%)	79.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1020.7142	Memory size	5.0 MiB

More details

Variables

UnknownDose

UnknownDose

Real number (ℝ)

SKEWED ZEROS

Distinct	816	Minimum	0
Distinct (%)	0.1%	Maximum	429061
Missing	0	Zeros	650845
Missing (%)	0.0%	Zeros (%)	98.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	11.498014	Memory size	5.0 MiB

More details

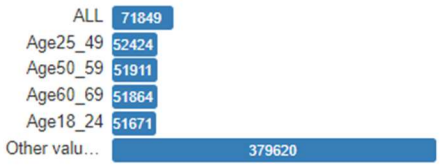
Variables

TargetGroup

TargetGroup

Categorical

Distinct	17
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.0 MiB



More details

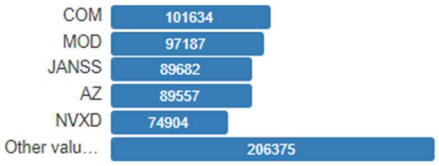
Variables

Vaccine

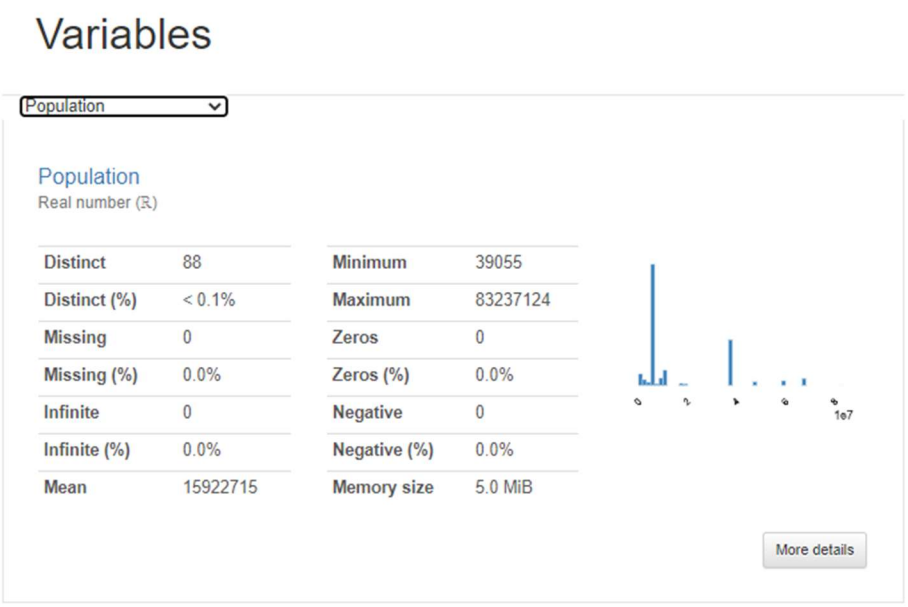
Vaccine

Categorical

Distinct	18
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.0 MiB



More details



Explicación del resultado

Observamos que el fichero consta de 16 variables, de las cuales 5 son categóricas y 11 numéricas. No se observan duplicados, si 2115884 casos nulos, un 20.1%. Así mismo, hay 8 variables que están altamente correlacionadas

- **EU_Variants**
- Ejemplo de Datos

	country	country_code	year_week	source	new_cases	number_sequenced	percent_cases_sequenced	valid_denominator	variant	number_detections_variant
0	Austria	AT	2020-01	GISAID	NaN	0	0.0	True	Other	0
1	Austria	AT	2020-01	GISAID	NaN	0	0.0	True	P.1	0
2	Austria	AT	2020-01	GISAID	NaN	0	0.0	True	P.3	0
3	Austria	AT	2020-01	GISAID	NaN	0	0.0	True	B.1.1.7	0
4	Austria	AT	2020-01	GISAID	NaN	0	0.0	True	XBB	0

number_sequenced_known_variant	percent_variant
0	NaN
0	NaN
0	NaN
0	NaN
0	NaN

- Resumen General

Overview

Overview | Alerts 20 | Reproduction

Dataset statistics

Number of variables	12
Number of observations	120066
Missing cells	23979
Missing cells (%)	1.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	10.2 MiB
Average record size in memory	89.0 B

Variable types

Categorical	6
Numeric	5
Boolean	1

- Alertas

Alerts

`valid_denominator` has constant value "True"	Constant
`year_week` has a high cardinality: 167 distinct values	High cardinality
`percent_variant` has a high cardinality: 1085 distinct values	High cardinality
`new_cases` is highly overall correlated with `number_sequenced` and 1 other fields	High correlation
`number_sequenced` is highly overall correlated with `new_cases` and 2 other fields	High correlation
`percent_cases_sequenced` is highly overall correlated with `number_sequenced` and 1 other fields	High correlation
`number_sequenced_known_variant` is highly overall correlated with `new_cases` and 2 other fields	High correlation
`country` is highly overall correlated with `country_code`	High correlation
`country_code` is highly overall correlated with `country`	High correlation
`source` is highly imbalanced (57.2%)	Imbalance
`percent_variant` is highly imbalanced (74.1%)	Imbalance
`new_cases` has 4084 (3.4%) missing values	Missing
`percent_cases_sequenced` has 1540 (1.3%) missing values	Missing
`percent_variant` has 18354 (15.3%) missing values	Missing
`number_detections_variant` is highly skewed (γ1 = 25.12523464)	Skewed
`new_cases` has 1497 (1.2%) zeros	Zeros
`number_sequenced` has 18162 (15.1%) zeros	Zeros
`percent_cases_sequenced` has 20844 (17.4%) zeros	Zeros
`number_detections_variant` has 95412 (79.5%) zeros	Zeros

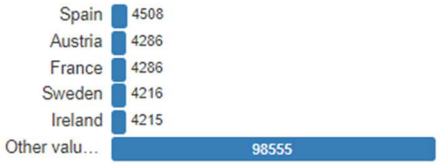
- Variables

Variables

country

country
Categorical

Distinct	30
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	938.1 KiB



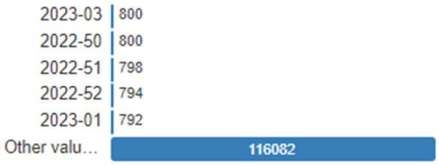
More details

Variables

year_week

year_week
Categorical

Distinct	167
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	938.1 KiB



More details

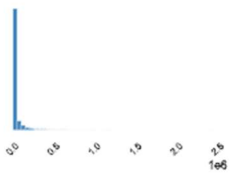
Variables

new_cases

new_cases
Real number (ℝ)

HIGH CORRELATION MISSING ZEROS

Distinct	3929	Minimum	0
Distinct (%)	3.4%	Maximum	2486551
Missing	4084	Zeros	1497
Missing (%)	3.4%	Zeros (%)	1.2%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	39706.83	Memory size	938.1 KiB



More details

Variables

number_sequenced

number_sequenced

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	2414	Minimum	0
Distinct (%)	2.0%	Maximum	60124
Missing	0	Zeros	18162
Missing (%)	0.0%	Zeros (%)	15.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	959.49531	Memory size	938.1 KiB

More details

Variables

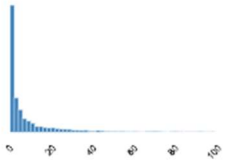
percent_cases_sequenced

percent_cases_sequenced

Real number (ℝ)

HIGH CORRELATION MISSING ZEROS

Distinct	737	Minimum	0
Distinct (%)	0.6%	Maximum	100
Missing	1540	Zeros	20844
Missing (%)	1.3%	Zeros (%)	17.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	9.8519515	Memory size	938.1 KiB



More details

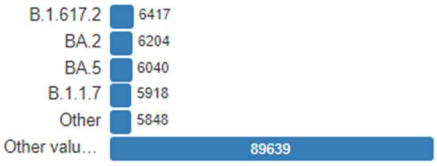
Variables

variant

variant

Categorical

Distinct	30
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	938.1 KiB



More details

Variables

number_detections_variant

number_detections_variant

Real number (ℝ)

SKEWED ZEROS

Distinct	2390	Minimum	0
Distinct (%)	2.0%	Maximum	50774
Missing	0	Zeros	95412
Missing (%)	0.0%	Zeros (%)	79.5%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	76.756884	Memory size	938.1 KiB

More details

Variables

number_sequenced_known_variant

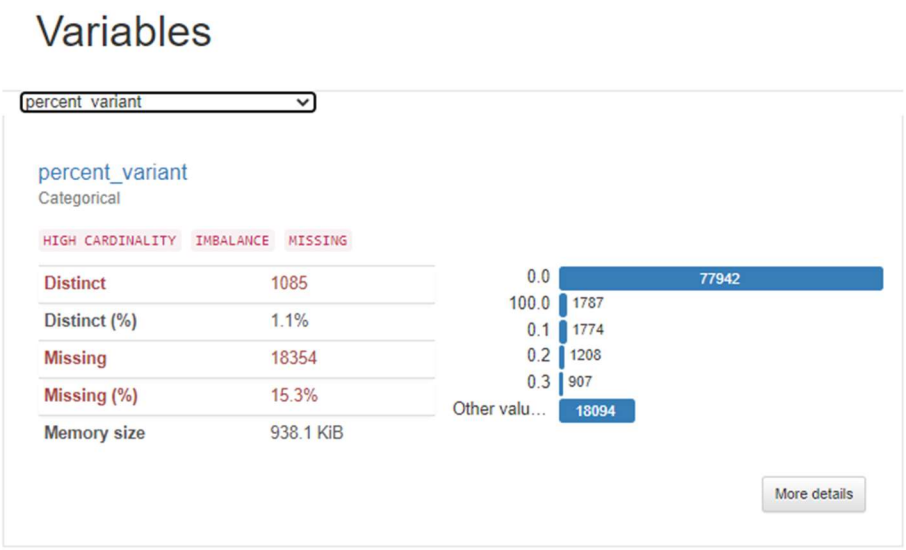
number_sequenced_known_variant

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	2366	Minimum	0
Distinct (%)	2.0%	Maximum	60124
Missing	0	Zeros	18360
Missing (%)	0.0%	Zeros (%)	15.3%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	927.7182	Memory size	938.1 KiB

More details



Explicación del resultado

Observamos que el fichero consta de 12 variables, de las cuales 6 son categóricas, 5 numéricas y 1 boolean. No se observan duplicados, si 23979 casos nulos. Así mismo, hay 6 variables que están altamente correlacionadas

- **WHO_Global_Data_1 (Número de Casos globales históricos)**
- Ejemplo de Datos

	Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
0	2020-01-03	AF	Afghanistan	EMRO	0	0	0	0
1	2020-01-04	AF	Afghanistan	EMRO	0	0	0	0
2	2020-01-05	AF	Afghanistan	EMRO	0	0	0	0
3	2020-01-06	AF	Afghanistan	EMRO	0	0	0	0
4	2020-01-07	AF	Afghanistan	EMRO	0	0	0	0

- Resumen General

Overview

Overview | Alerts 15 | Reproduction

Dataset statistics

Number of variables	8
Number of observations	280134
Missing cells	1182
Missing cells (%)	0.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	17.1 MiB
Average record size in memory	64.0 B

Variable types

Categorical	4
Numeric	4

- Alertas

Alerts

`Date_reported` has a high cardinality: 1182 distinct values	High cardinality
`Country_code` has a high cardinality: 236 distinct values	High cardinality
`Country` has a high cardinality: 237 distinct values	High cardinality
`New_cases` is highly overall correlated with `Cumulative_cases` and 2 other fields	High correlation
`Cumulative_cases` is highly overall correlated with `New_cases` and 2 other fields	High correlation
`New_deaths` is highly overall correlated with `New_cases` and 2 other fields	High correlation
`Cumulative_deaths` is highly overall correlated with `New_cases` and 2 other fields	High correlation
`New_cases` is highly skewed (γ1 = 110.3781991)	Skewed
`Date_reported` is uniformly distributed	Uniform
`Country_code` is uniformly distributed	Uniform
`Country` is uniformly distributed	Uniform
`New_cases` has 110629 (39.5%) zeros	Zeros
`Cumulative_cases` has 27440 (9.8%) zeros	Zeros
`New_deaths` has 175752 (62.7%) zeros	Zeros
`Cumulative_deaths` has 48731 (17.4%) zeros	Zeros

- Variables

Variables

Date_reported

Date_reported

Categorical

HIGH CARDINALITY UNIFORM

Distinct	1182
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	2.1 MiB

More details

Variables

Country_code

Country_code

Categorical

HIGH CARDINALITY UNIFORM

Distinct	236
Distinct (%)	0.1%
Missing	1182
Missing (%)	0.4%
Memory size	2.1 MiB

More details

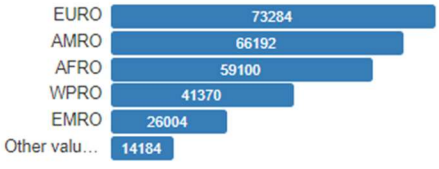
Variables

WHO_region

WHO_region

Categorical

Distinct	7
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.1 MiB



More details

Variables

New_cases

New_cases

Real number (ℝ)

HIGH CORRELATION SKEWED ZEROS

Distinct	19296	Minimum	-8261
Distinct (%)	6.9%	Maximum	6966046
Missing	0	Zeros	110629
Missing (%)	0.0%	Zeros (%)	39.5%
Infinite	0	Negative	59
Infinite (%)	0.0%	Negative (%)	< 0.1%
Mean	2717.9931	Memory size	2.1 MiB

More details

Variables

Cumulative_cases

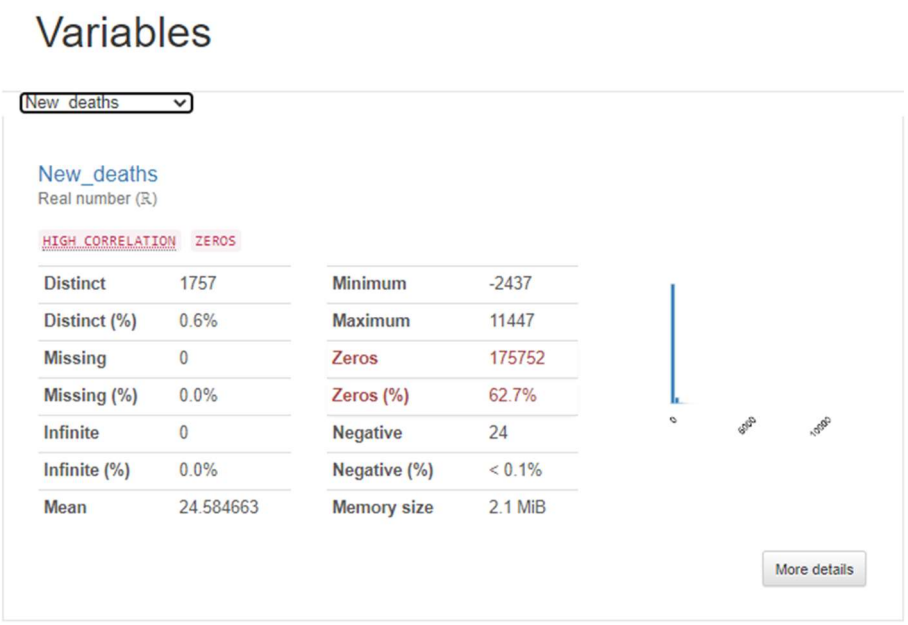
Cumulative_cases

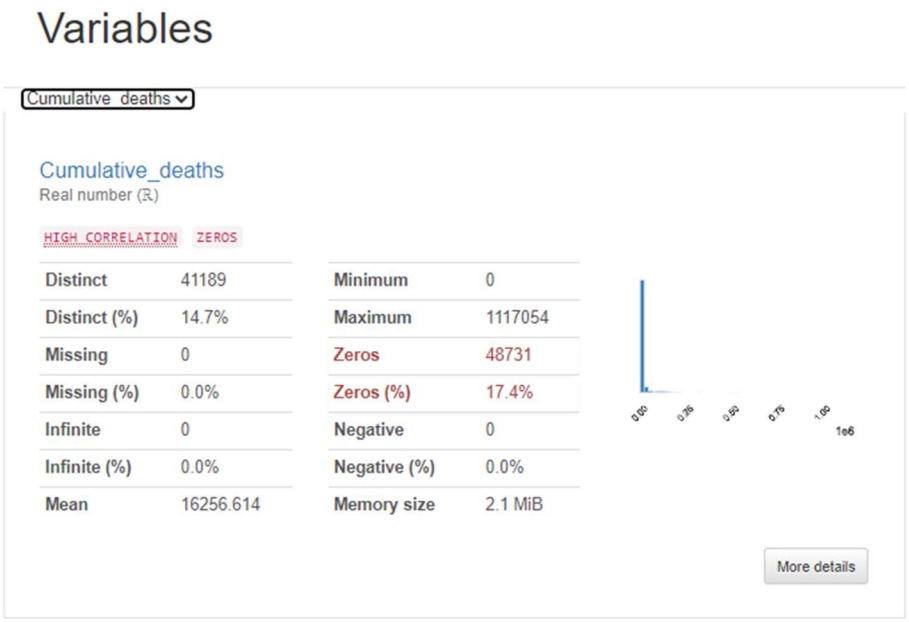
Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	125273	Minimum	0
Distinct (%)	44.7%	Maximum	1.0269757×10^8
Missing	0	Zeros	27440
Missing (%)	0.0%	Zeros (%)	9.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1202945.5	Memory size	2.1 MiB

More details





Explicación del resultado

Observamos que el fichero consta de 8 variables, de las cuales 4 son categóricas y 4 numéricas. No se observan duplicados, si 1182 casos nulos. Así mismo, hay 4 variables que están altamente correlacionadas

- **WHO_Global_Data_2 (Número de Casos globales actuales)**
- Ejemplo de Datos

	Name	WHO Region	Cases - cumulative total	Cases - cumulative total per 100000 population	Cases - newly reported in last 7 days	Cases - newly reported in last 7 days per 100000 population	Cases - newly reported in last 24 hours	Deaths - cumulative total	Deaths - cumulative total per 100000 population	Deaths - newly reported in last 7 days	Deaths - newly reported in last 7 days per 100000 population	Deaths - newly reported in last 24 hours
Global	NaN	761402282	9768.405152	534869	6.862098	2692	6887000	88.356717	4243	0.054436	25	NaN
United States of America	Americas	102697566	31026.206000	152968	46.214000	0	1117054	337.476000	2084	0.630000	0	NaN
China	Western Pacific	99238143	6744.989000	503	0.034000	0	120894	8.217000	70	0.005000	0	NaN
India	South-East Asia	44707525	3239.665000	9407	0.682000	0	530841	38.467000	28	0.002000	0	NaN
France	Europe	38677413	59467.703000	42503	65.350000	3	161857	248.860000	122	0.188000	0	NaN

- Resumen General

Overview

Overview | Alerts 20 | Reproduction

Dataset statistics

Number of variables	12
Number of observations	238
Missing cells	243
Missing cells (%)	8.5%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	32.3 KiB
Average record size in memory	138.9 B

Variable types

Categorical	2
Numeric	9
Unsupported	1

- Alertas

Alerts

`WHO Region` is highly overall correlated with `Cases - cumulative total per 100000 population` and 4 other fields	High correlation
`Cases - cumulative total` is highly overall correlated with `Deaths - cumulative total`	High correlation
`Cases - cumulative total per 100000 population` is highly overall correlated with `WHO Region` and 5 other fields	High correlation
`Cases - newly reported in last 7 days` is highly overall correlated with `Cases - cumulative total per 100000 population` and 2 other fields	High correlation
`Cases - newly reported in last 7 days per 100000 population` is highly overall correlated with `Deaths - newly reported in last 7 days per 100000 population`	High correlation
`Cases - newly reported in last 24 hours` is highly overall correlated with `WHO Region` and 4 other fields	High correlation
`Deaths - cumulative total` is highly overall correlated with `Cases - cumulative total`	High correlation
`Deaths - cumulative total per 100000 population` is highly overall correlated with `WHO Region` and 5 other fields	High correlation
`Deaths - newly reported in last 7 days` is highly overall correlated with `WHO Region` and 4 other fields	High correlation
`Deaths - newly reported in last 7 days per 100000 population` is highly overall correlated with `WHO Region` and 4 other fields	High correlation
`Deaths - newly reported in last 7 days per 100000 population` is highly imbalanced (93.3%)	Imbalance
`Deaths - newly reported in last 24 hours` has 238 (100.0%) missing values	Missing
`Deaths - newly reported in last 24 hours` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`Cases - cumulative total per 100000 population` has 122 (51.3%) zeros	Zeros
`Cases - newly reported in last 7 days` has 121 (50.8%) zeros	Zeros
`Cases - newly reported in last 7 days per 100000 population` has 227 (95.4%) zeros	Zeros
`Cases - newly reported in last 24 hours` has 9 (3.8%) zeros	Zeros
`Deaths - cumulative total` has 9 (3.8%) zeros	Zeros
`Deaths - cumulative total per 100000 population` has 176 (73.9%) zeros	Zeros
`Deaths - newly reported in last 7 days` has 175 (73.5%) zeros	Zeros

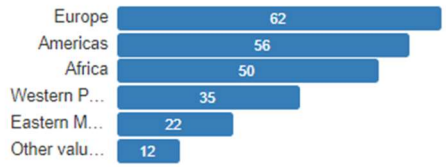
- Variables

Variables

Name

Name
Categorical

Distinct	7
Distinct (%)	3.0%
Missing	1
Missing (%)	0.4%
Memory size	3.7 KiB



More details

Variables

WHO Region

WHO Region

Real number (ℝ)

Distinct	237	**Minimum**	0
Distinct (%)	99.6%	**Maximum**	7.6140228×10^8
Missing	0	**Zeros**	2
Missing (%)	0.0%	**Zeros (%)**	0.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	6398338.5	**Memory size**	3.7 KiB

More details

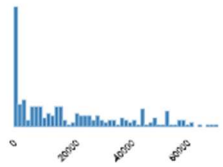
Variables

Cases - cumulative total

Cases - cumulative total

Real number (ℝ)

Distinct	236	**Minimum**	0
Distinct (%)	99.6%	**Maximum**	73728.639
Missing	1	**Zeros**	2
Missing (%)	0.4%	**Zeros (%)**	0.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	20186.978	**Memory size**	3.7 KiB



More details

Variables

Cases - cumulative total per 100000 population

Cases - cumulative total per 100000 population

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	98	**Minimum**	0
Distinct (%)	41.2%	**Maximum**	534869
Missing	0	**Zeros**	122
Missing (%)	0.0%	**Zeros (%)**	51.3%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4494.6975	**Memory size**	3.7 KiB

More details

Variables

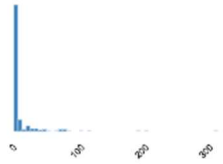
Cases - newly reported in last 7 days

Cases - newly reported in last 7 days

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	117	**Minimum**	0
Distinct (%)	49.4%	**Maximum**	317.501
Missing	1	**Zeros**	121
Missing (%)	0.4%	**Zeros (%)**	50.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	10.620916	**Memory size**	3.7 KiB



More details

Variables

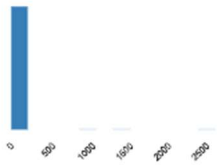
Cases - newly reported in last 7 days per 100000 population

Cases - newly reported in last 7 days per 100000 population

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	12	Minimum	0
Distinct (%)	5.0%	Maximum	2692
Missing	0	Zeros	227
Missing (%)	0.0%	Zeros (%)	95.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	22.621849	Memory size	3.7 KiB



More details

Variables

Cases - newly reported in last 24 hours

Cases - newly reported in last 24 hours

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	218	Minimum	0
Distinct (%)	91.6%	Maximum	6887000
Missing	0	Zeros	9
Missing (%)	0.0%	Zeros (%)	3.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	57873.95	Memory size	3.7 KiB

More details

Variables

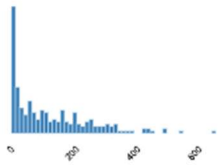
Deaths - cumulative total

Deaths - cumulative total

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	229	Minimum	0
Distinct (%)	96.6%	Maximum	666.45
Missing	1	Zeros	9
Missing (%)	0.4%	Zeros (%)	3.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	123.78705	Memory size	3.7 KiB



More details

Variables

Deaths - cumulative total per 100000 population

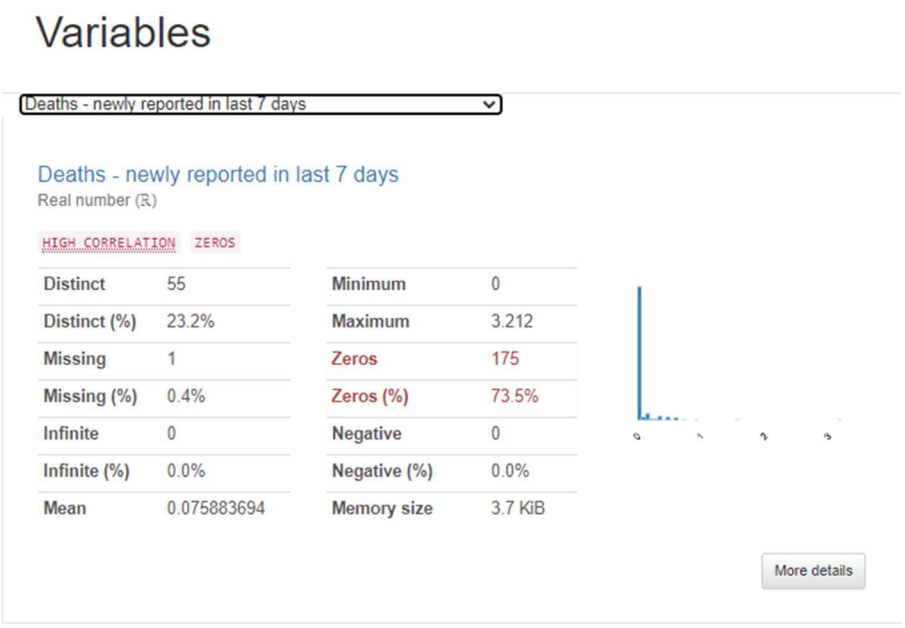
Deaths - cumulative total per 100000 population

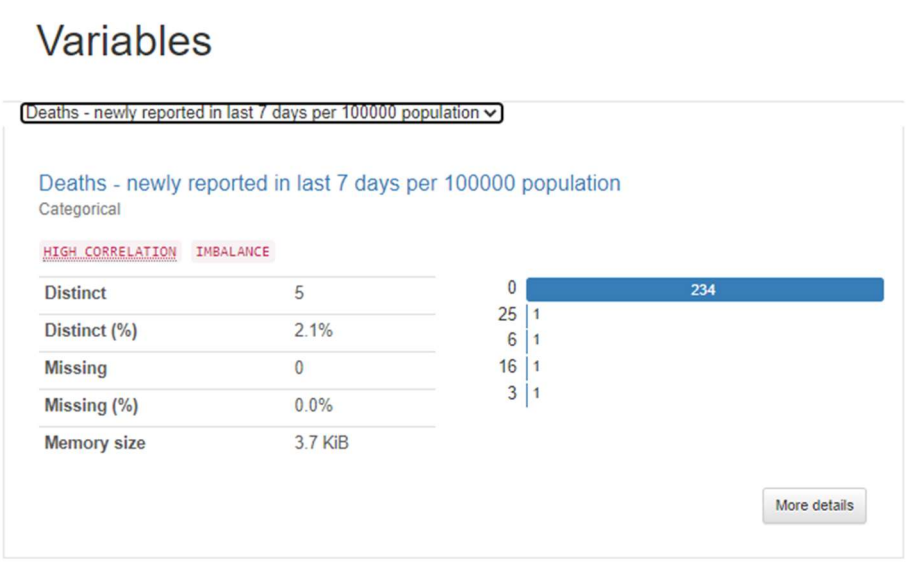
Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	41	Minimum	0
Distinct (%)	17.2%	Maximum	4243
Missing	0	Zeros	176
Missing (%)	0.0%	Zeros (%)	73.9%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	35.655462	Memory size	3.7 KiB

More details





Explicación del resultado

Observamos que el fichero consta de 12 variables, de las cuales 2 son categóricas, 9 numéricas y 1 no reconocible. No se observan duplicados, si 243 casos nulos. Así mismo, hay 10 variables que están altamente correlacionadas

- **WHO_Vaccination_Data_1 (Vacunaciones globales)**
- Ejemplo de Datos

	COUNTRY	ISO3	WHO_REGION	DATA_SOURCE	DATE_UPDATED	TOTAL_VACCINATIONS	PERSONS_VACCINATED_1PLUS_DOSE
0	Afghanistan	AFG	EMRO	REPORTING	2023-03-28	16499153.0	14457250
1	Albania	ALB	EURO	REPORTING	2023-03-19	3070468.0	1347054
2	Algeria	DZA	AFRO	REPORTING	2022-09-04	15267442.0	7840131
3	American Samoa	ASM	WPRO	REPORTING	2023-03-15	111316.0	46201
4	Andorra	AND	EURO	REPORTING	2023-02-26	156957.0	57904

TOTAL_VACCINATIONS_PER100	PERSONS_VACCINATED_1PLUS_DOSE_PER100	PERSONS_FULLY_VACCINATED	PERSONS_FULLY_VACCINATED_PER100
42.383	37.138	13743213	35.304
106.700	47.332	1276432	44.851
34.817	17.879	6481186	14.780
201.670	83.702	42473	76.948
203.100	76.012	53492	70.221

VACCINES_USED	FIRST_VACCINE_DATE	NUMBER_VACCINES_TYPES_USED	PERSONS_BOOSTER_ADD_DOSE	PERSONS_BOOSTER_ADD_DOSE_PER100
AstraZeneca - Vaxzevria,Beijing CNBG - BBIBP-C...	2021-02-22	11.0	1016263.0	2.611
AstraZeneca - Vaxzevria,Gamaleya - Gam-Covid-V...	2021-01-13	5.0	395384.0	13.893
Beijing CNBG - BBIBP-CorV,Gamaleya - Gam-Covid...	2021-01-30	4.0	575651.0	1.313
Janssen - Ad26.COV 2-S,Moderna - Spikevax,Pfiz...	2020-12-21	3.0	24160.0	43.770
AstraZeneca - Vaxzevria,Moderna - Spikevax,Pfi...	2021-01-20	3.0	43060.0	56.526

- Resumen General

Overview

Overview | Alerts 29 | Reproduction

Dataset statistics

Number of variables	16
Number of observations	229
Missing cells	71
Missing cells (%)	1.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	28.8 KiB
Average record size in memory	128.6 B

Variable types

Categorical	7
Numeric	9

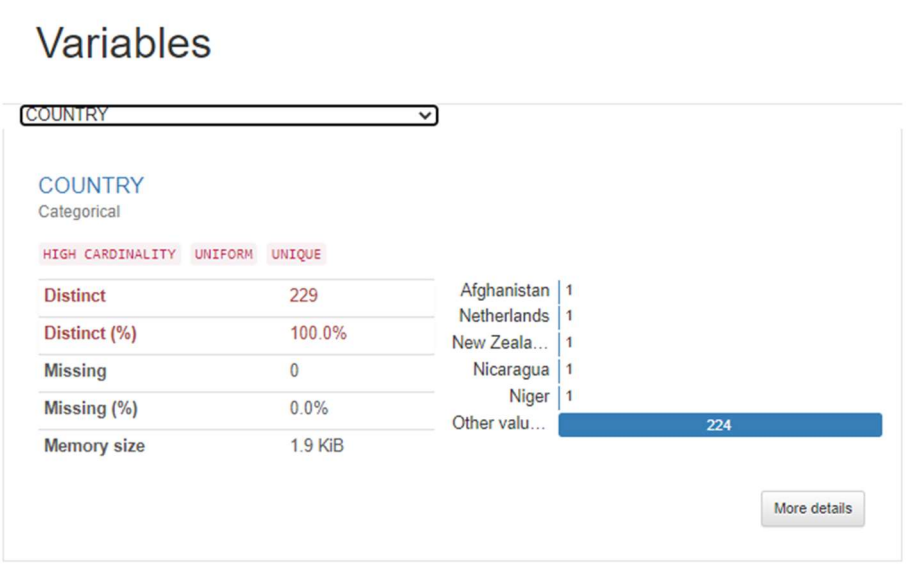
- Alertas

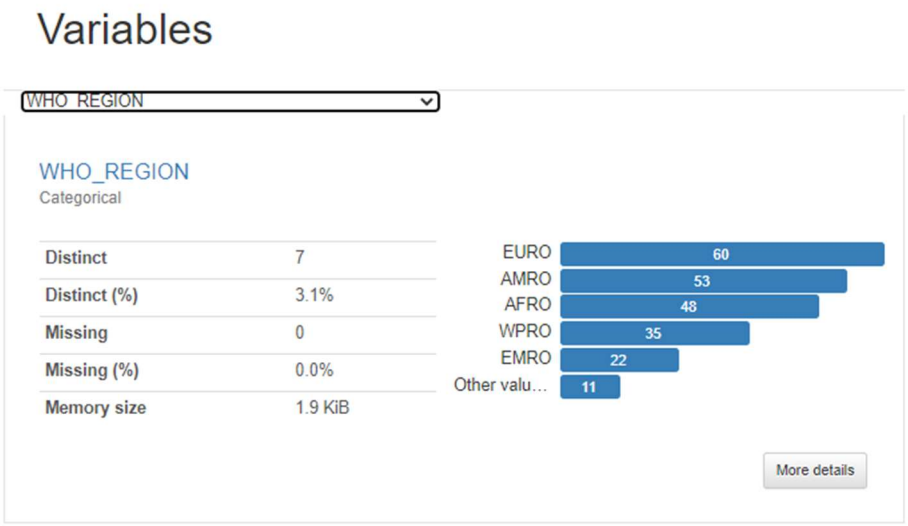
Alerts

`COUNTRY` has a high cardinality: 229 distinct values	High cardinality
`ISO3` has a high cardinality: 229 distinct values	High cardinality
`DATE_UPDATED` has a high cardinality: 77 distinct values	High cardinality
`VACCINES_USED` has a high cardinality: 123 distinct values	High cardinality
`FIRST_VACCINE_DATE` has a high cardinality: 102 distinct values	High cardinality
`TOTAL_VACCINATIONS` is highly overall correlated with `PERSONS_VACCINATED_1PLUS_DOSE` and 2 other fields	High correlation
`PERSONS_VACCINATED_1PLUS_DOSE` is highly overall correlated with `TOTAL_VACCINATIONS` and 2 other fields	High correlation
`TOTAL_VACCINATIONS_PER100` is highly overall correlated with `PERSONS_VACCINATED_1PLUS_DOSE_PER100` and 2 other fields	High correlation
`PERSONS_VACCINATED_1PLUS_DOSE_PER100` is highly overall correlated with `TOTAL_VACCINATIONS_PER100` and 2 other fields	High correlation
`PERSONS_FULLY_VACCINATED` is highly overall correlated with `TOTAL_VACCINATIONS` and 2 other fields	High correlation
`PERSONS_FULLY_VACCINATED_PER100` is highly overall correlated with `TOTAL_VACCINATIONS_PER100` and 2 other fields	High correlation
`PERSONS_BOOSTER_ADD_DOSE` is highly overall correlated with `TOTAL_VACCINATIONS` and 2 other fields	High correlation
`PERSONS_BOOSTER_ADD_DOSE_PER100` is highly overall correlated with `TOTAL_VACCINATIONS_PER100` and 3 other fields	High correlation
`WHO_REGION` is highly overall correlated with `DATE_UPDATED`	High correlation
`DATA_SOURCE` is highly overall correlated with `PERSONS_BOOSTER_ADD_DOSE_PER100` and 1 other fields	High correlation
`DATE_UPDATED` is highly overall correlated with `WHO_REGION` and 1 other fields	High correlation

DATA_SOURCE is highly imbalanced (76.1%)	Imbalance
VACCINES_USED has 4 (1.7%) missing values	Missing
FIRST_VACCINE_DATE has 22 (9.6%) missing values	Missing
NUMBER_VACCINES_TYPES_USED has 4 (1.7%) missing values	Missing
PERSONS_BOOSTER_ADD_DOSE has 19 (8.3%) missing values	Missing
PERSONS_BOOSTER_ADD_DOSE_PER100 has 19 (8.3%) missing values	Missing
COUNTRY is uniformly distributed	Uniform
ISO3 is uniformly distributed	Uniform
COUNTRY has unique values	Unique
ISO3 has unique values	Unique
PERSONS_VACCINATED_1PLUS_DOSE has unique values	Unique
PERSONS_VACCINATED_1PLUS_DOSE_PER100 has unique values	Unique
PERSONS_FULLY_VACCINATED has unique values	Unique

- Variables





Variables

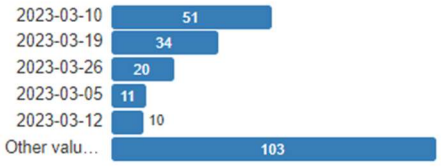
DATE_UPDATED

DATE_UPDATED

Categorical

HIGH CARDINALITY HIGH CORRELATION

Distinct	77
Distinct (%)	33.6%
Missing	0
Missing (%)	0.0%
Memory size	1.9 KiB



More details

Variables

TOTAL_VACCINATIONS

TOTAL_VACCINATIONS

Real number (ℝ)

Distinct	228	Minimum	138
Distinct (%)	100.0%	Maximum	3.5158728×10^9
Missing	1	Zeros	0
Missing (%)	0.4%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	58473576	Memory size	1.9 KiB

More details

Variables

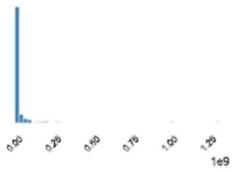
PERSONS_VACCINATED_1PLUS_DOSE

PERSONS_VACCINATED_1PLUS_DOSE

Real number (ℝ)

HIGH CORRELATION UNIQUE

Distinct	229	Minimum	0
Distinct (%)	100.0%	Maximum	1.3179949×10^9
Missing	0	Zeros	1
Missing (%)	0.0%	Zeros (%)	0.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	24169330	Memory size	1.9 KiB



More details

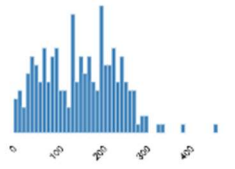
Variables

TOTAL_VACCINATIONS_PER100

TOTAL_VACCINATIONS_PER100

Real number (ℝ)

Distinct	227	Minimum	0.301
Distinct (%)	100.0%	Maximum	469.778
Missing	2	Zeros	0
Missing (%)	0.9%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	153.29389	Memory size	1.9 KiB



More details

Variables

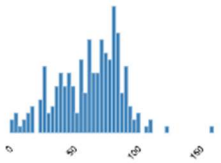
PERSONS_VACCINATED_1PLUS_DOSE_PER100

PERSONS_VACCINATED_1PLUS_DOSE_PER100

Real number (ℝ)

HIGH CORRELATION UNIQUE

Distinct	229	Minimum	0
Distinct (%)	100.0%	Maximum	163.185
Missing	0	Zeros	1
Missing (%)	0.0%	Zeros (%)	0.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	63.275991	Memory size	1.9 KiB



More details

Variables

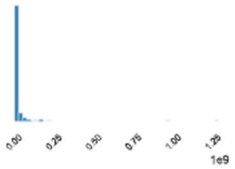
PERSONS_FULLY_VACCINATED

PERSONS_FULLY_VACCINATED

Real number (ℝ)

HIGH CORRELATION UNIQUE

Distinct	229	Minimum	0
Distinct (%)	100.0%	Maximum	1.2844137×10^9
Missing	0	Zeros	1
Missing (%)	0.0%	Zeros (%)	0.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	22310357	Memory size	1.9 KiB



More details

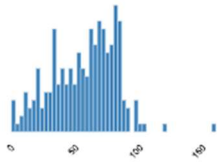
Variables

PERSONS_FULLY_VACCINATED_PER100

PERSONS_FULLY_VACCINATED_PER100

Real number (ℝ)

Distinct	227	**Minimum**	0
Distinct (%)	99.1%	**Maximum**	163.185
Missing	0	**Zeros**	1
Missing (%)	0.0%	**Zeros (%)**	0.4%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	58.147856	**Memory size**	1.9 KiB



More details

Variables

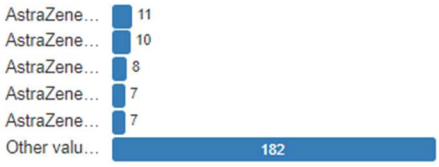
VACCINES_USED

VACCINES_USED

Categorical

HIGH CARDINALITY MISSING

Distinct	123
Distinct (%)	54.7%
Missing	4
Missing (%)	1.7%
Memory size	1.9 KiB



More details

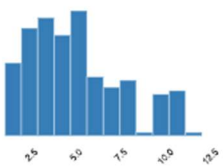
Variables

NUMBER_VACCINES_TYPES_USED

NUMBER_VACCINES_TYPES_USED

Real number (ℝ)

Distinct	12	**Minimum**	1
Distinct (%)	5.3%	**Maximum**	12
Missing	4	**Zeros**	0
Missing (%)	1.7%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4.8577778	**Memory size**	1.9 KiB



More details

Explicación del resultado

Observamos que el fichero consta de 16 variables, de las cuales 7 son categóricas y 9 numéricas. No se observan duplicados, si 71 casos nulos. Así mismo, hay 11 variables que están altamente correlacionadas

- **WHO_Vaccination_Data_2 (Información sobre el tipo de Vacunas)**
- Ejemplo de Datos

	ISO3	VACCINE_NAME	PRODUCT_NAME	COMPANY_NAME	AUTHORIZATION_DATE	START_DATE	END_DATE	COMMENT	DATA_SOURCE
0	GRL	Moderna - mRNA-1273	mRNA-1273	Moderna	NaN	NaN	NaN	NaN	OWID
1	FRO	Moderna - mRNA-1273	mRNA-1273	Moderna	NaN	NaN	NaN	NaN	OWID
2	FRO	Pfizer BioNTech - Comirnaty	Comirnaty	Pfizer BioNTech	NaN	NaN	NaN	NaN	OWID
3	JEY	Moderna - mRNA-1273	mRNA-1273	Moderna	NaN	NaN	NaN	NaN	OWID
4	JEY	AstraZeneca - AZD1222	AZD1222	AstraZeneca	NaN	NaN	NaN	NaN	OWID

- Resumen General

Overview

Overview | Alerts 15 | Reproduction

Dataset statistics

Number of variables	9
Number of observations	1074
Missing cells	3044
Missing cells (%)	31.5%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	75.6 KiB
Average record size in memory	72.1 B

Variable types

Categorical	7
Unsupported	2

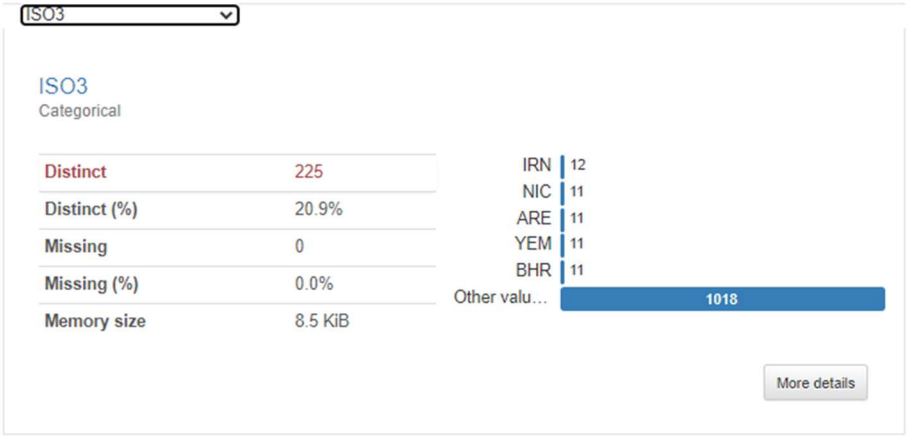
- Alertas

Alerts

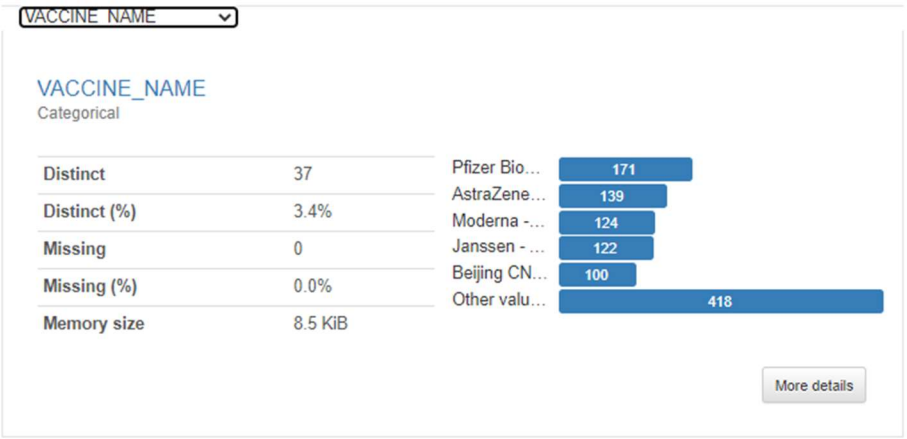
ISO3 has a high cardinality: 225 distinct values	High cardinality
AUTHORIZATION_DATE has a high cardinality: 179 distinct values	High cardinality
START_DATE has a high cardinality: 239 distinct values	High cardinality
VACCINE_NAME is highly overall correlated with PRODUCT_NAME and 2 other fields	High correlation
PRODUCT_NAME is highly overall correlated with VACCINE_NAME and 2 other fields	High correlation
COMPANY_NAME is highly overall correlated with VACCINE_NAME and 1 other fields	High correlation
DATA_SOURCE is highly overall correlated with VACCINE_NAME and 1 other fields	High correlation
DATA_SOURCE is highly imbalanced (85.1%)	Imbalance
COMPANY_NAME has 32 (3.0%) missing values	Missing
AUTHORIZATION_DATE has 584 (54.4%) missing values	Missing
START_DATE has 279 (26.0%) missing values	Missing
END_DATE has 1074 (100.0%) missing values	Missing
COMMENT has 1074 (100.0%) missing values	Missing
END_DATE is an unsupported type, check if it needs cleaning or further analysis	Unsupported
COMMENT is an unsupported type, check if it needs cleaning or further analysis	Unsupported

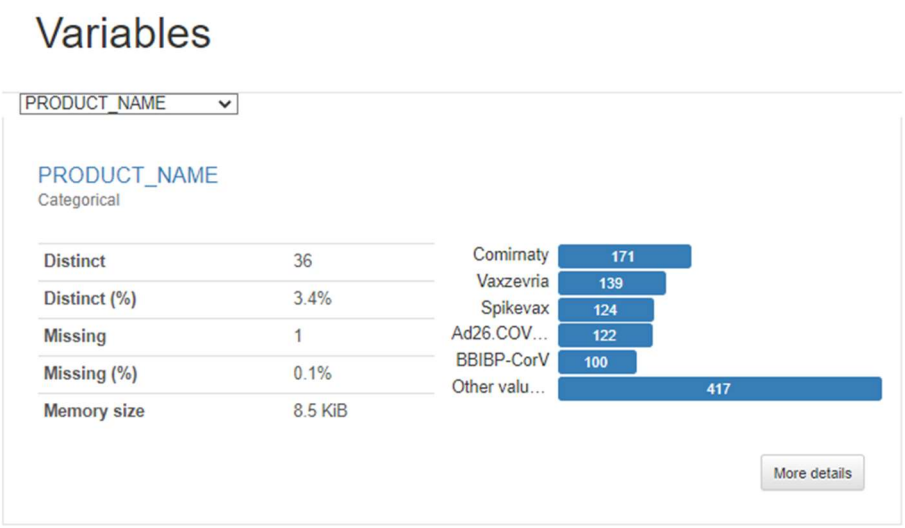
- Variables

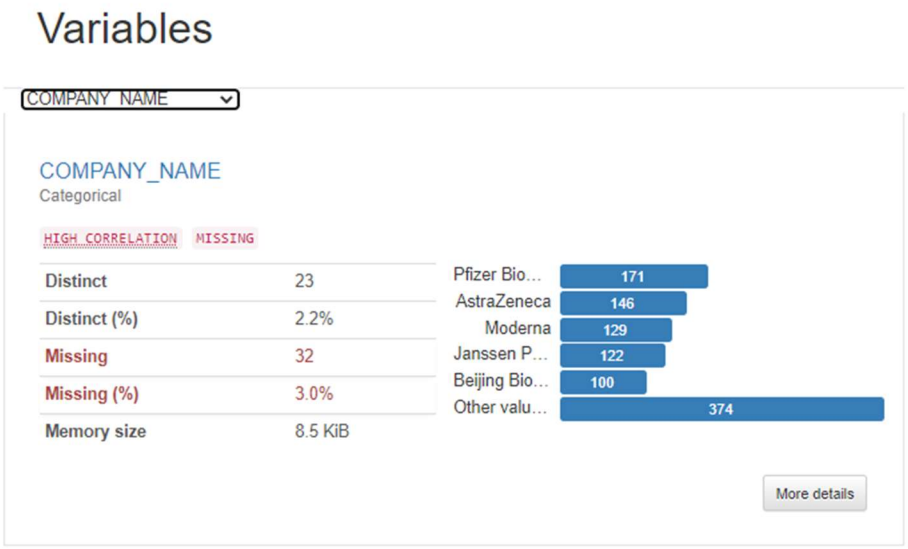
Variables

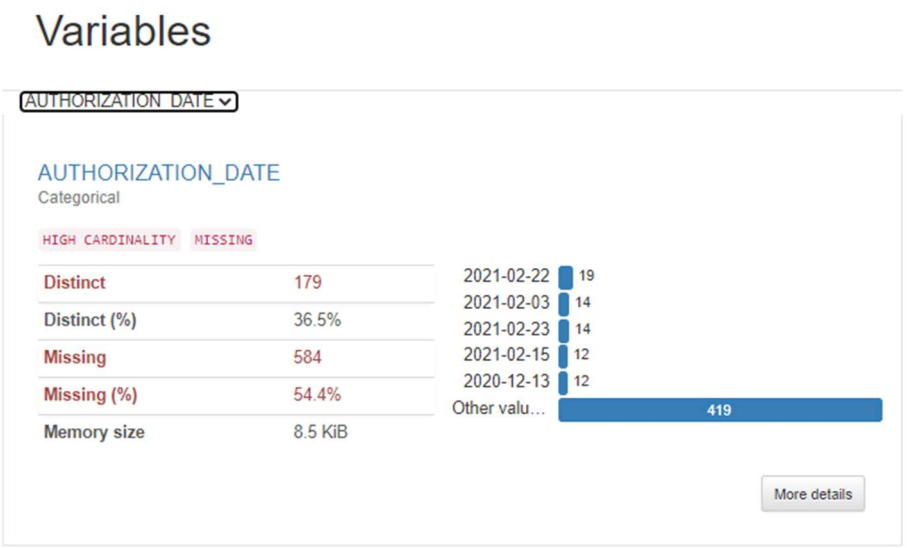


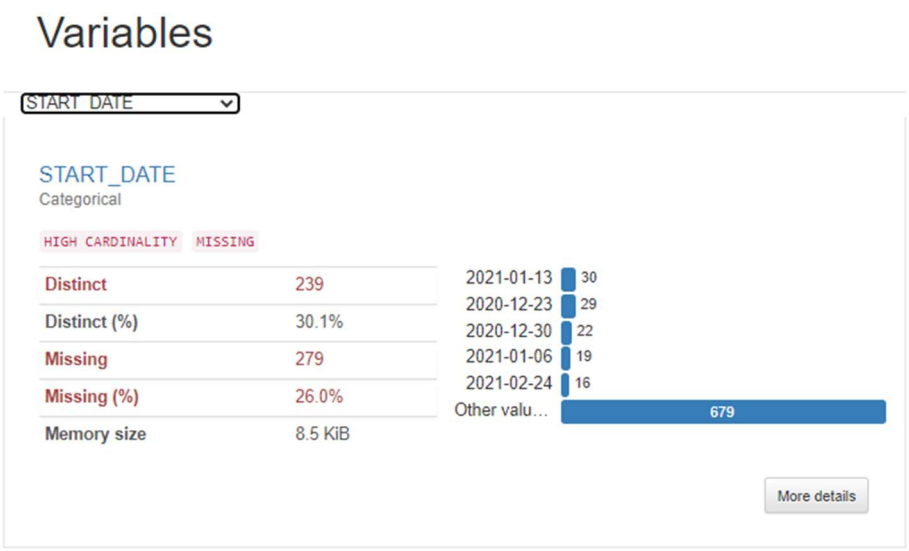
Variables




Variables
PRODUCT_NAME
PRODUCT_NAME
Categorical
Distinct 36
Distinct (%) 3.4%
Missing 1
Missing (%) 0.1%
Memory size 8.5 KiB
Comirnaty 171
Vaxzevria 139
Spikevax 124
Ad26.COV... 122
BBIBP-CorV 100
Other valu... 417
More details


Variables
COMPANY_NAME
COMPANY_NAME
Categorical
HIGH CORRELATION MISSING
Distinct 23
Distinct (%) 2.2%
Missing 32
Missing (%) 3.0%
Memory size 8.5 KiB
Pfizer Bio... 171
AstraZeneca 146
Moderna 129
Janssen P... 122
Beijing Bio... 100
Other valu... 374
More details


Variables
AUTHORIZATION_DATE
AUTHORIZATION_DATE
Categorical
HIGH CARDINALITY MISSING
Distinct 179
Distinct (%) 36.5%
Missing 584
Missing (%) 54.4%
Memory size 8.5 KiB
2021-02-22 19
2021-02-03 14
2021-02-23 14
2021-02-15 12
2020-12-13 12
Other valu... 419
More details



Explicación del resultado

Observamos que el fichero consta de 9 variables, de las cuales 7 son categóricas y 2 no reconocibles. No se observan duplicados, si 3044 casos nulos. Así mismo, hay 4 variables que están altamente correlacionadas

Juan.

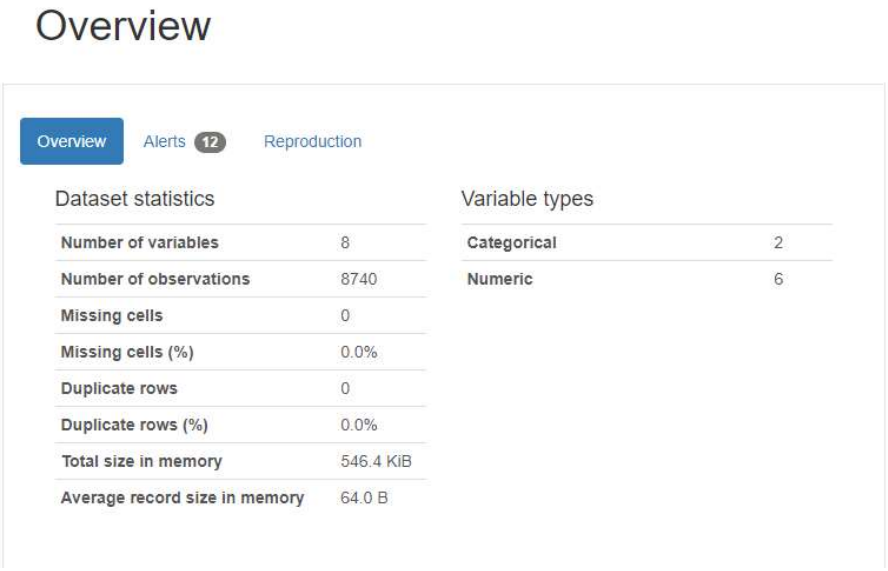
Perfilado de Datos // COVID ESPAÑA

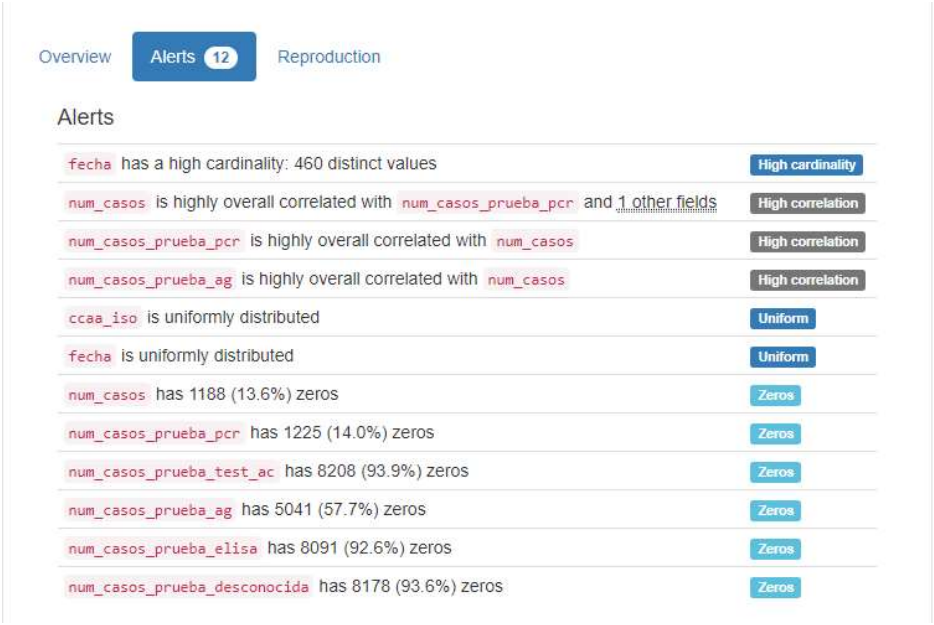
1. **casos_diag_ccaadecl (***Número de casos por técnica diagnóstica y CCAA (de declaración)*

```
In [11]: df_ride.head()
```

Out[11]:

	ccaa_iso	fecha	num_casos	num_casos_prueba_pcr	num_casos_prueba_test_ac	num_casos_prueba_ag	num_casos_prueba_elisa	num_casos_prueba_desco
0	AN	2020-01-18	0	0	0	0	0	
1	AR	2020-01-18	0	0	0	0	0	
2	AS	2020-01-18	0	0	0	0	0	
3	CB	2020-01-18	0	0	0	0	0	
4	CE	2020-01-18	0	0	0	0	0	



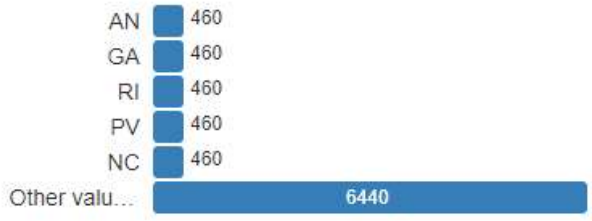


Variables

ccaa_iso

Categorical

Distinct	19
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	68.4 KiB



fecha

Categorical

HIGH CARDINALITY UNIFORM

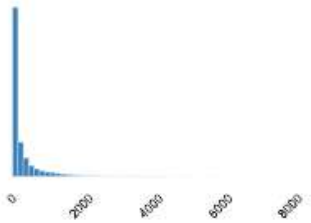
Distinct	460
Distinct (%)	5.3%
Missing	0
Missing (%)	0.0%
Memory size	68.4 KiB

num_casos_prueba_pcr

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	1476	**Minimum**	0
Distinct (%)	16.9%	**Maximum**	7721
Missing	0	**Zeros**	1225
Missing (%)	0.0%	**Zeros (%)**	14.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	310.22609	**Memory size**	68.4 KiB



num_casos_prueba_test_ac

Real number (ℝ)

Distinct	61	**Minimum**	0
Distinct (%)	0.7%	**Maximum**	157
Missing	0	**Zeros**	8208
Missing (%)	0.0%	**Zeros (%)**	93.9%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.54610984	**Memory size**	68.4 KiB

num_casos_prueba_ag

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	705	**Minimum**	0
Distinct (%)	8.1%	**Maximum**	3909
Missing	0	**Zeros**	5041
Missing (%)	0.0%	**Zeros (%)**	57.7%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	79.440961	**Memory size**	68.4 KiB

num_casos_prueba_desconocida
Real number (ℝ)

Distinct	104	**Minimum**	0
Distinct (%)	1.2%	**Maximum**	982
Missing	0	**Zeros**	8178
Missing (%)	0.0%	**Zeros (%)**	93.6%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	3.6185355	**Memory size**	68.4 KiB

Explicación del resultado

Observamos que el fichero consta de 8 variables, de las cuales 2 son categóricas y 6 numéricas. No se observan duplicados ni valores nulos. Así mismo, hay 3 variables que están altamente correlacionadas como son los *num_casos*, num_casos_prueba_pcr y num_casos_pruega_ag. Observamos un valor superior al 90% de celdas con valores 0 en las variables num_casos_prueba_test_ac, num_casos_prueba_elisa y num_casos prueba_elisa.

2. **casos_tecnica_provincia (***Número de casos por técnica diagnóstica y CCAA (de residencia)*

	provincia_iso	fecha	num_casos	num_casos_prueba_pcr	num_casos_prueba_test_ac	num_casos_prueba_ag	num_casos_prueba_elisa	num_casos_prueba_
0	A	2020-01-01	0	0	0	0	0	
1	AB	2020-01-01	0	0	0	0	0	
2	AL	2020-01-01	0	0	0	0	0	
3	AV	2020-01-01	0	0	0	0	0	
4	B	2020-01-01	0	0	0	0	0	

Overview

Overview | Alerts 15 | Reproduction

Dataset statistics

Number of variables	8
Number of observations	25281
Missing cells	477
Missing cells (%)	0.2%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.5 MiB
Average record size in memory	64.0 B

Variable types

Categorical	2
Numeric	6

Overview | Alerts 15 | Reproduction

Alerts

`provincia_iso` has a high cardinality: 52 distinct values	High cardinality
`fecha` has a high cardinality: 477 distinct values	High cardinality
`num_casos` is highly overall correlated with `num_casos_prueba_pcr` and 1 other fields	High correlation
`num_casos_prueba_pcr` is highly overall correlated with `num_casos` and 1 other fields	High correlation
`num_casos_prueba_ag` is highly overall correlated with `num_casos` and 1 other fields	High correlation
`provincia_iso` has 477 (1.9%) missing values	Missing
`num_casos_prueba_elisa` is highly skewed (γ1 = 24.7592003)	Skewed
`provincia_iso` is uniformly distributed	Uniform
`fecha` is uniformly distributed	Uniform
`num_casos` has 3999 (15.8%) zeros	Zeros
`num_casos_prueba_pcr` has 4142 (16.4%) zeros	Zeros
`num_casos_prueba_test_ac` has 23828 (94.3%) zeros	Zeros
`num_casos_prueba_ag` has 15145 (59.9%) zeros	Zeros
`num_casos_prueba_elisa` has 23978 (94.8%) zeros	Zeros
`num_casos_prueba_desconocida` has 24114 (95.4%) zeros	Zeros

Variables

Select Columns

provincia_iso

Categorical

HIGH CARDINALITY MISSING UNIFORM

Distinct	52
Distinct (%)	0.2%
Missing	477
Missing (%)	1.9%
Memory size	197.6 KiB

fecha

Categorical

HIGH CARDINALITY UNIFORM

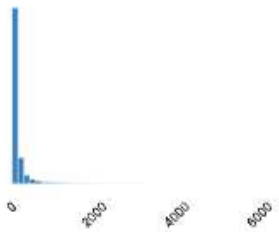
Distinct	477
Distinct (%)	1.9%
Missing	0
Missing (%)	0.0%
Memory size	197.6 KiB

num_casos

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	1398	**Minimum**	0
Distinct (%)	5.5%	**Maximum**	6974
Missing	0	**Zeros**	3999
Missing (%)	0.0%	**Zeros (%)**	15.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	136.31379	**Memory size**	197.6 KiB

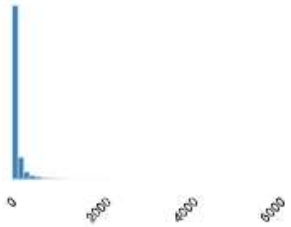


num_casos_prueba_pcr

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	1181	**Minimum**	0
Distinct (%)	4.7%	**Maximum**	6525
Missing	0	**Zeros**	4142
Missing (%)	0.0%	**Zeros (%)**	16.4%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	107.24956	**Memory size**	197.6 KiB



num_casos_prueba_test_ac

Real number (ℝ)

Distinct	31	**Minimum**	0
Distinct (%)	0.1%	**Maximum**	32
Missing	0	**Zeros**	23828
Missing (%)	0.0%	**Zeros (%)**	94.3%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.18879791	**Memory size**	197.6 KiB

num_casos_prueba_ag

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	644	**Minimum**	0
Distinct (%)	2.5%	**Maximum**	3262
Missing	0	**Zeros**	15145
Missing (%)	0.0%	**Zeros (%)**	59.9%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	27.463866	**Memory size**	197.6 KiB

num_casos_prueba_elisa
Real number (ℝ)

SKEWED ZEROS

Distinct	43	**Minimum**	0
Distinct (%)	0.2%	**Maximum**	70
Missing	0	**Zeros**	23978
Missing (%)	0.0%	**Zeros (%)**	94.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.16059491	**Memory size**	197.6 KiB

num_casos_prueba_desconocida
Real number (ℝ)

Distinct	139	**Minimum**	0
Distinct (%)	0.5%	**Maximum**	574
Missing	0	**Zeros**	24114
Missing (%)	0.0%	**Zeros (%)**	95.4%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.250979	**Memory size**	197.6 KiB

Explicación del resultado

Observamos que el fichero consta de 8 variables, de las cuales 2 son categóricas y 6 numéricas. No se observan duplicados y solamente el 0,2% de valores nulos. Así mismo, hay 3 variables que están altamente correlacionadas como son los *num_casos*, num_casos_prueba_pcr y num_casos_pruega_ag. Observamos un valor superior al 90% de celdas con valores 0 en las variables num_casos_prueba_test_ac, num_casos_prueba_elisa y num_casos prueba_elisa.

3. **casos_ hosp_uci_def_sexo_edad_prov_res:** (*Número de hospitalizaciones, número de ingresos en UCI, número de defunciones por sexo, edad y provincia de residencia)*

	provincia_iso	sexo	grupo_edad	fecha	num_casos	num_hosp	num_uci	num_def
0	A	H	0-9	2020-01-01	0	0	0	0
1	A	H	10-19	2020-01-01	0	0	0	0
2	A	H	20-29	2020-01-01	0	0	0	0
3	A	H	30-39	2020-01-01	0	0	0	0
4	A	H	40-49	2020-01-01	0	0	0	0

Overview

Overview | Alerts 16 | Reproduction

Dataset statistics

Number of variables	8
Number of observations	758430
Missing cells	14310
Missing cells (%)	0.2%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	46.3 MiB
Average record size in memory	64.0 B

Variable types

Categorical	4
Numeric	4

Alerts

`provincia_iso` has a high cardinality: 52 distinct values	High cardinality
`fecha` has a high cardinality: 477 distinct values	High cardinality
`num_casos` is highly overall correlated with `num_hosp`	High correlation
`num_hosp` is highly overall correlated with `num_casos`	High correlation
`provincia_iso` has 14310 (1.9%) missing values	Missing
`num_hosp` is highly skewed (γ1 = 30.98749075)	Skewed
`num_uci` is highly skewed (γ1 = 28.05787935)	Skewed
`num_def` is highly skewed (γ1 = 37.08136693)	Skewed
`provincia_iso` is uniformly distributed	Uniform
`sexo` is uniformly distributed	Uniform
`grupo_edad` is uniformly distributed	Uniform
`fecha` is uniformly distributed	Uniform
`num_casos` has 488735 (64.4%) zeros	Zeros
`num_hosp` has 656902 (86.6%) zeros	Zeros
`num_uci` has 737067 (97.2%) zeros	Zeros
`num_def` has 725644 (95.7%) zeros	Zeros

Variables

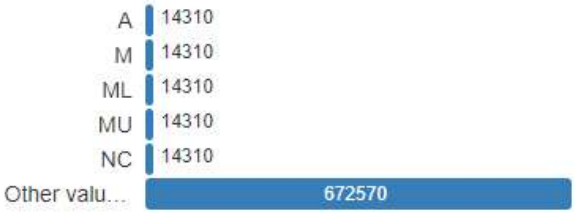
Select Columns

provincia_iso

Categorical

HIGH CARDINALITY MISSING UNIFORM

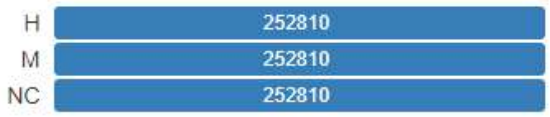
Distinct	52
Distinct (%)	< 0.1%
Missing	14310
Missing (%)	1.9%
Memory size	5.8 MiB



sexo

Categorical

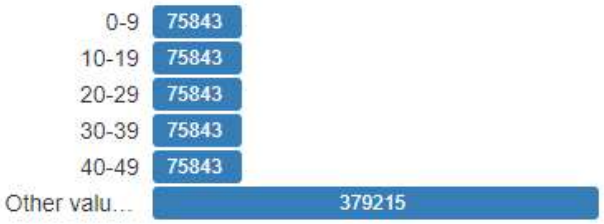
Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB



grupo_edad

Categorical

Distinct	10
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB

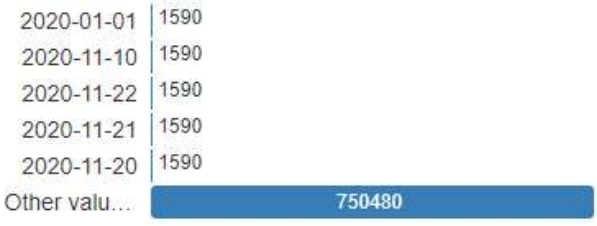


fecha

Categorical

HIGH CARDINALITY UNIFORM

Distinct	477
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB



num_casos

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	530	**Minimum**	0
Distinct (%)	0.1%	**Maximum**	772
Missing	0	**Zeros**	488735
Missing (%)	0.0%	**Zeros (%)**	64.4%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4.5437931	**Memory size**	5.8 MiB

num_hosp

Real number (ℝ)

HIGH CORRELATION SKEWED ZEROS

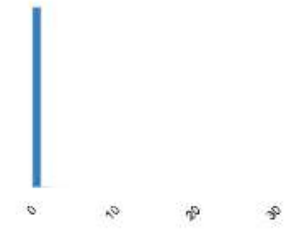
Distinct	178	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	269
Missing	0	**Zeros**	656902
Missing (%)	0.0%	**Zeros (%)**	86.6%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.45156705	**Memory size**	5.8 MiB

num_uci

Real number (ℝ)

SKEWED ZEROS

Distinct	32	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	35
Missing	0	**Zeros**	737067
Missing (%)	0.0%	**Zeros (%)**	97.2%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.041438234	**Memory size**	5.8 MiB



num_def
Real number (ℝ)

SKEWED ZEROS

Distinct	79	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	100
Missing	0	**Zeros**	725644
Missing (%)	0.0%	**Zeros (%)**	95.7%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.10200546	**Memory size**	5.8 MiB

Explicación del resultado

Observamos que el fichero consta de 8 variables, de las cuales 4 son categóricas y 4 numéricas. No se observan duplicados y solamente el 0,2% de valores nulos. Así mismo, hay 2 variables que están altamente correlacionadas como son los *num_casos y el num_hosp*. Observamos un valor superior al 90% de celdas con valores 0 en las variables num_uci y num_def.

4. **covid19_tia_muni_y_distritos:** (*Número de casos confirmados, tasa de incidencia acum en los últimos 14 días y global de la Provincia de Madrid y sus Municipios/Distritos)*

	municipio_distrito	fecha_informe	casos_confirmados_ultimos_14dias	tasa_incidencia_acumulada_ultimos_14dias	casos_confirmados_totales	tasa_incidencia_a
0	Madrid-Retiro	2020-07-01 09:00:00	28.0	23.47	1691.0	
1	Madrid-Salamanca	2020-07-01 09:00:00	23.0	15.74	1781.0	
2	Madrid-Centro	2020-07-01 09:00:00	18.0	13.35	1282.0	
3	Madrid-Arganzuela	2020-07-01 09:00:00	16.0	10.40	1769.0	
4	Madrid-Chamartín	2020-07-01 09:00:00	12.0	8.23	1800.0	

Overview

Overview | Alerts 8 | Reproduction

Dataset statistics

Number of variables	7
Number of observations	25273
Missing cells	25868
Missing cells (%)	14.6%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.3 MiB
Average record size in memory	56.0 B

Variable types

Categorical	1
DateTime	1
Numeric	5

Overview | Alerts 8 | Reproduction

Alerts

`municipio_distrito` has a high cardinality: 199 distinct values	High cardinality
`casos_confirmados_ultimos_14dias` is highly overall correlated with `casos_confirmados_totales`	High correlation
`casos_confirmados_totales` is highly overall correlated with `casos_confirmados_ultimos_14dias`	High correlation
`casos_confirmados_ultimos_14dias` has 16716 (66.1%) missing values	Missing
`casos_confirmados_totales` has 9152 (36.2%) missing values	Missing
`municipio_distrito` is uniformly distributed	Uniform
`tasa_incidencia_acumulada_ultimos_14dias` has 10122 (40.1%) zeros	Zeros
`tasa_incidencia_acumulada_total` has 4751 (18.8%) zeros	Zeros

Variables

Select Columns

municipio_distrito

Categorical

HIGH CARDINALITY UNIFORM

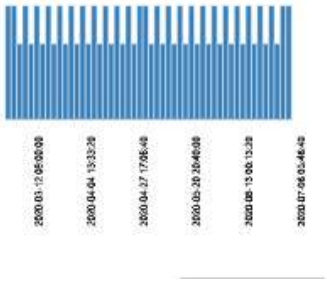
Distinct	199
Distinct (%)	0.8%
Missing	0
Missing (%)	0.0%
Memory size	197.6 KiB

fecha_informe

Date

Distinct	127
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	197.6 KiB

Minimum	2020-02-26 07:00:00
Maximum	2020-07-01 09:00:00



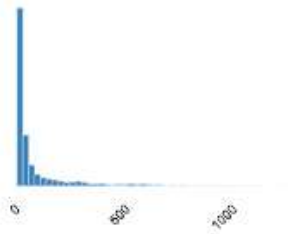
casos_confirmados_ultimos_14dias

Real number (ℝ)

HIGH CORRELATION MISSING

Distinct	733
Distinct (%)	8.6%
Missing	16716
Missing (%)	66.1%
Infinite	0
Infinite (%)	0.0%
Mean	96.318686

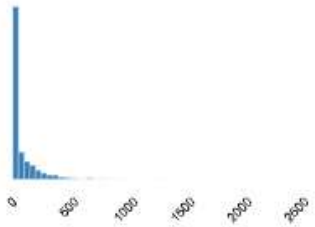
Minimum	6
Maximum	1403
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	197.6 KiB



tasa_incidencia_acumulada_ultimos_14dias

Real number (ℝ)

Distinct	5069	**Minimum**	0
Distinct (%)	20.1%	**Maximum**	2492.75
Missing	0	**Zeros**	10122
Missing (%)	0.0%	**Zeros (%)**	40.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	77.046346	**Memory size**	197.6 KiB

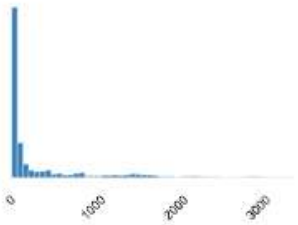


casos_confirmados_totales

Real number (ℝ)

HIGH CORRELATION MISSING

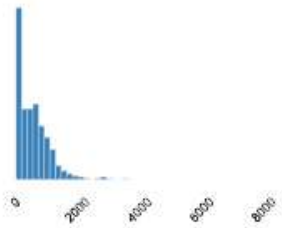
Distinct	2174	**Minimum**	6
Distinct (%)	13.5%	**Maximum**	3403
Missing	9152	**Zeros**	0
Missing (%)	36.2%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	364.62161	**Memory size**	197.6 KiB



tasa_incidencia_acumulada_total

Real number (ℝ)

Distinct	7821	**Minimum**	0
Distinct (%)	30.9%	**Maximum**	9090.91
Missing	0	**Zeros**	4751
Missing (%)	0.0%	**Zeros (%)**	18.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	536.94989	**Memory size**	197.6 KiB



codigo_geometria
Real number (ℝ)

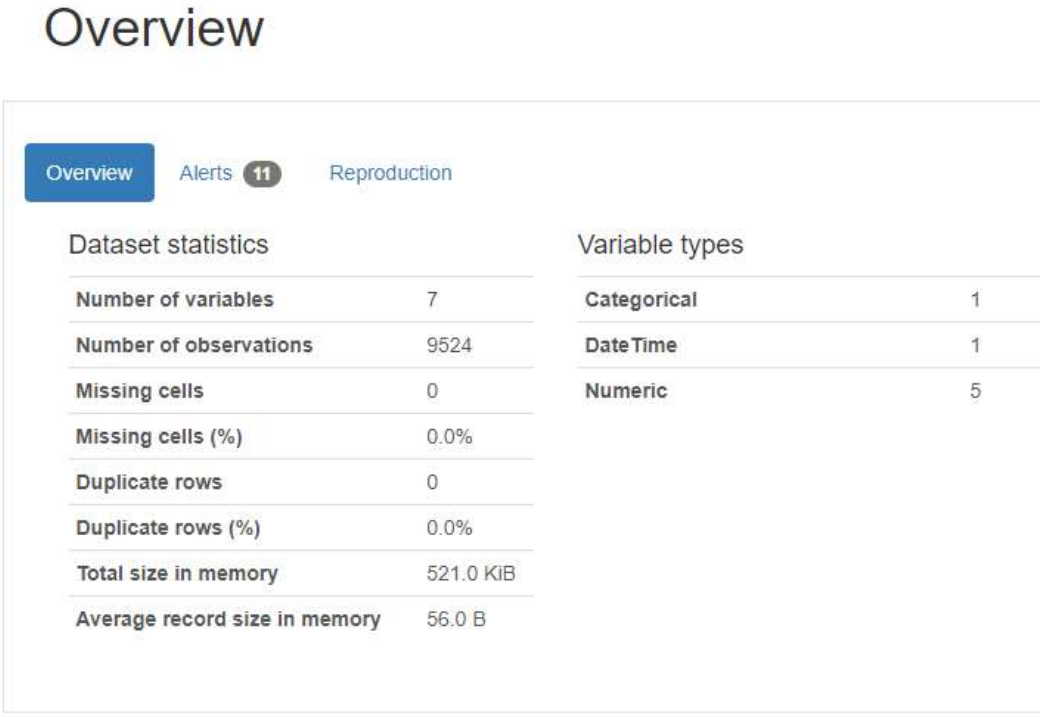
Distinct	199	**Minimum**	14
Distinct (%)	0.8%	**Maximum**	79621
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	9345.7739	**Memory size**	197.6 KiB

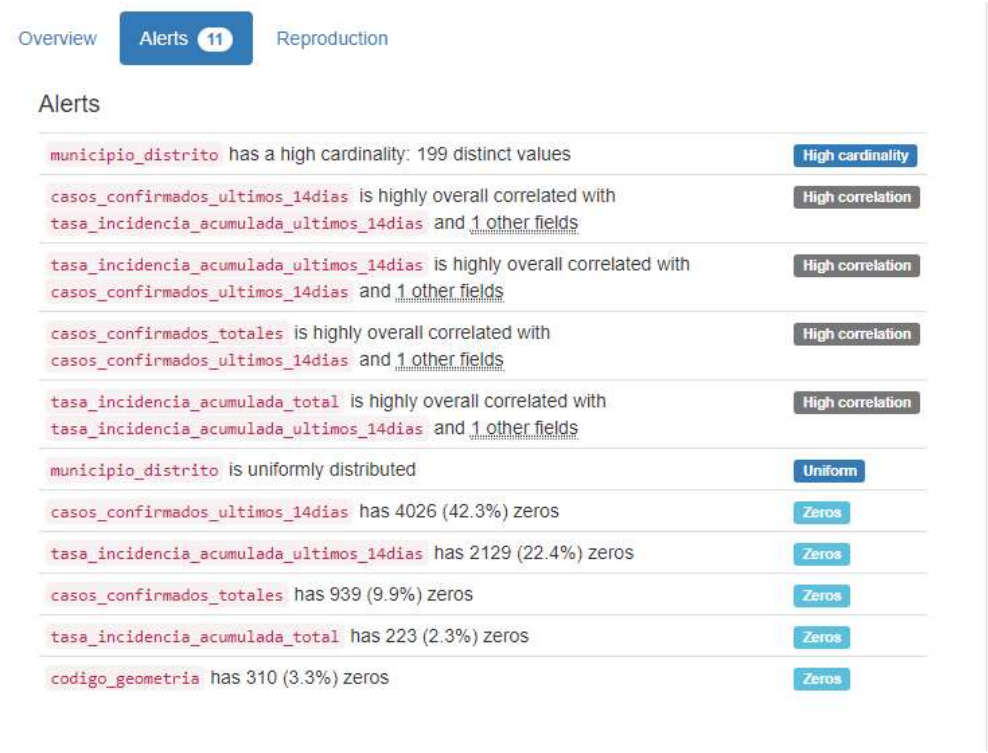
Explicación del resultado

Observamos que el fichero consta de 7 variables, de las cuales 5 són numéricas, 1 categórica y 1 formato fecha. No se observan duplicados y hay un 14,6% de valores nulos. Así mismo, hay 2 variables que están altamente correlacionadas como son los *el num_casos_confirmados_ultimos_14dias y el num_casos_confirmados_totales*.

5. **covid19_tia_muni_y_distritos_s:** (*Número de casos confirmados, tasa de incidencia acum en los últimos 14 días y global de la Provincia de Madrid y sus Municipios/Distritos) Continuación del fichero anterior.*

	municipio_distrito	fecha_informe	casos_confirmados_ultimos_14dias	tasa_incidencia_acumulada_ultimos_14dias	casos_confirmados_totales	tasa_incidencia_a
0	Madrid-Retiro	2021-04-20	498	413.76	11192	
1	Madrid-Salamanca	2021-04-20	741	501.21	14759	
2	Madrid-Centro	2021-04-20	739	525.89	14079	
3	Madrid-Arganzuela	2021-04-20	665	427.25	14580	
4	Madrid-Chamartín	2021-04-20	622	421.55	14191	




Overview
Alerts 11
Reproduction
Alerts
municipio_distrito has a high cardinality: 199 distinct values
High cardinality
casos_confirmados_ultimos_14dias is highly overall correlated with tasa_incidencia_acumulada_ultimos_14dias and 1 other fields
High correlation
tasa_incidencia_acumulada_ultimos_14dias is highly overall correlated with casos_confirmados_ultimos_14dias and 1 other fields
High correlation
casos_confirmados_totales is highly overall correlated with casos_confirmados_ultimos_14dias and 1 other fields
High correlation
tasa_incidencia_acumulada_total is highly overall correlated with tasa_incidencia_acumulada_ultimos_14dias and 1 other fields
High correlation
municipio_distrito is uniformly distributed
Uniform
casos_confirmados_ultimos_14dias has 4026 (42.3%) zeros
Zeros
tasa_incidencia_acumulada_ultimos_14dias has 2129 (22.4%) zeros
Zeros
casos_confirmados_totales has 939 (9.9%) zeros
Zeros
tasa_incidencia_acumulada_total has 223 (2.3%) zeros
Zeros
codigo_geometria has 310 (3.3%) zeros
Zeros

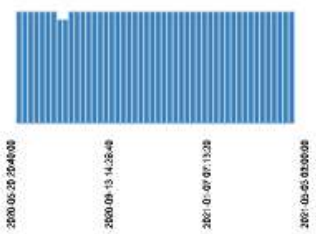

Variables
Select Columns
municipio_distrito
Categorical
HIGH CARDINALITY
UNIFORM
Distinct 199
Distinct (%) 2.1%
Missing 0
Missing (%) 0.0%
Memory size 74.5 KiB
Madrid-Re... 48
Casarrubu... 48
Corpa 48
Meco 48
Zarzalejo 48
Other valu... 9284

fecha_informe

Date

Distinct	48
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	74.5 KiB

Minimum	2020-05-26 00:00:00
Maximum	2021-04-20 00:00:00



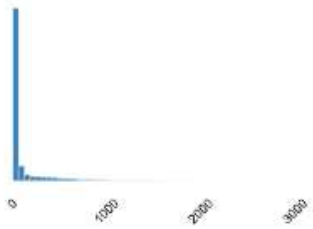
casos_confirmados_ultimos_14dias

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	939
Distinct (%)	9.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	109.16348

Minimum	0
Maximum	2988
Zeros	4026
Zeros (%)	42.3%
Negative	0
Negative (%)	0.0%
Memory size	74.5 KiB



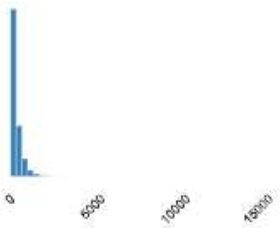
tasa_incidencia_acumulada_ultimos_14dias

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	4909
Distinct (%)	51.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	301.76676

Minimum	0
Maximum	17073.17
Zeros	2129
Zeros (%)	22.4%
Negative	0
Negative (%)	0.0%
Memory size	74.5 KiB

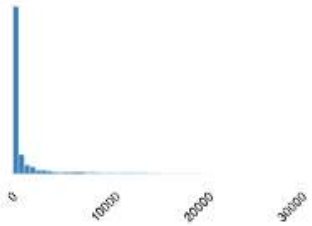


casos_confirmados_totales

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	2878	**Minimum**	0
Distinct (%)	30.2%	**Maximum**	29928
Missing	0	**Zeros**	939
Missing (%)	0.0%	**Zeros (%)**	9.9%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1571.2999	**Memory size**	74.5 KiB

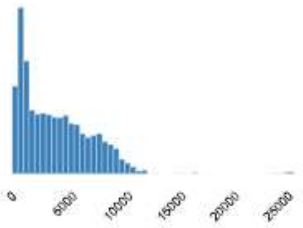


tasa_incidencia_acumulada_total

Real number (ℝ)

HIGH CORRELATION ZEROS

Distinct	6927	**Minimum**	0
Distinct (%)	72.7%	**Maximum**	26341.46
Missing	0	**Zeros**	223
Missing (%)	0.0%	**Zeros (%)**	2.3%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4083.0565	**Memory size**	74.5 KiB



codigo_geometria

Real number (ℝ)

Distinct	200	**Minimum**	0
Distinct (%)	2.1%	**Maximum**	79621
Missing	0	**Zeros**	310
Missing (%)	0.0%	**Zeros (%)**	3.3%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	9167.08	**Memory size**	74.5 KiB

Explicación del resultado

Observamos que el fichero consta de 7 variables, de las cuales 5 són numéricas, 1 categórica y 1 formato fecha. No se observan duplicados y no hay ningún valor nulo. Así mismo, hay 4 variables que están altamente correlacionadas como son los *el num_casos_confirmados_ultimos_14dias con tasa_incidencia_acumulada_ultimos_14dias, num_casos_confirmados_totales* y tasa_incidencia_acumulada_total.