Covid ETL

2024-03-01

ETL (Extraer, Transformar, Cargar)

El siguiente documento tiene como objetivo explicar el proceso de ETL realizado en base a mis conocimientos actuales, el resultado obtenido puede no ser muy bueno en su totalidad, en caso de encontrar fallas en el analisis la documentacion queda adjuntada con el codigo realizado en RStudio para su correccion de parte de cualquier persona que posea los conocimientos adecuados.

Primero hacemos el cargado de las librerias/bibliotecas y el set de datos.

```
library(tidyquant)
library(readr)
library(plotly)
library(webshot)
library(orca)
library(corrplot)
library(dplyr)

getwd()
```

```
## [1] "C:/Users/Facundo/Desktop/UNO/DS/ETL/Covid/DataSet"
```

```
datos <- COVID_19_Global_Statistics_Dataset <- read_csv("COVID-19 Global Statistics Dataset.csv")
attach(datos)
```

El frame de datos a trabajar almacena diversas categorias de casos de covid agrupadas por cada Pais. Entre las variables encontramos las siguientes: Pais, Casos Totales, Casos Nuevos, Muertes Totales, Muertes Nuevas, Recuperados Totales, Nuevos Recuperados, Casos Activos, Casos Serios/Criticos, Casos Infantiles por millon, Muertes por cada millon, Test Totales, Test por cada millon y Poblacion. Aclaracion no se encontro un diccionario de datos que explique de forma detallada lo que significa cada variable.

```
head(datos)
```

```
## # A tibble: 6 x 14
##   Country  `Total Cases` `New Cases` `Total Deaths` `New Deaths`
##   <chr>            <dbl>       <dbl>          <dbl>        <dbl>
## 1 USA          111367209          NA        1199031           NA
## 2 India         45028429         161         533475            2
## 3 France        40138560          NA         167642           NA
```

```
## 4 Germany       38819284         574         182439           28
## 5 Brazil        38407327          NA         709765           NA
## 6 S. Korea      34571873          NA          35934           NA
## # i 9 more variables: `Total Recovered` <chr>, `New Recovered` <chr>,
## #   `Active Cases` <chr>, `Serious, Critical` <chr>, `Tot Cases/1M pop` <dbl>,
## #   `Deaths/1M pop` <dbl>, `Total Tests` <dbl>, `Tests/1M pop` <dbl>,
## #   Population <dbl>
```

Normalizamos la informacion estandarizando el idioma de las variables junto a una conversion de tipo de dato y removemos los caracteres especiales como lo son las “,” .

```
datos_normalizados <- c("Pais","Casos Totales","Casos Nuevos","Muertes Totales"
                        ,"Muertes Nuevas","Recuperados Totales","Nuevos Recuperados"
                        ,"Casos Activos","Casos Serios/Criticos"," Casos Infantiles c/1Millon"
                        ,"Muertes c/1Millon","Test Totales","Test c/1Millon","Poblacion")

names(datos) <- datos_normalizados

#datos <- datos[,-1]
datos$Pais <- as.character(gsub(",","",datos$Pais))
datos$`Casos Totales` <- as.numeric(gsub(",","",datos$`Casos Totales`))
datos$`Casos Nuevos` <- as.numeric(gsub(",","",datos$`Casos Nuevos` ))
datos$`Muertes Totales` <- as.numeric(gsub(",","",datos$`Muertes Totales`))
datos$`Muertes Nuevas` <- as.numeric(gsub(",","",datos$`Muertes Nuevas`))
datos$`Recuperados Totales` <- as.numeric(gsub(",","",datos$`Recuperados Totales` ))
datos$`Nuevos Recuperados` <- as.numeric(gsub(",","",datos$`Nuevos Recuperados`))
datos$`Casos Activos` <- as.numeric(gsub(",","",datos$`Casos Activos`))
datos$`Casos Serios/Criticos` <- as.numeric(gsub(",","",datos$`Casos Serios/Criticos`))
datos$` Casos Infantiles c/1Millon` <- as.numeric(gsub(",","",datos$` Casos Infantiles c/1Millon`))
datos$`Muertes c/1Millon` <- as.numeric(gsub(",","",datos$`Muertes c/1Millon`))
datos$`Test Totales` <- as.numeric(gsub(",","",datos$`Test Totales`))
datos$`Test c/1Millon` <- as.numeric(gsub(",","",datos$`Test c/1Millon`))
datos$Poblacion <- as.numeric(gsub(",","",datos$Poblacion))

str(datos)
```

```
## spc_tbl_ [239 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Pais                        : chr [1:239] "USA" "India" "France" "Germany" ...
##  $ Casos Totales               : num [1:239] 1.11e+08 4.50e+07 4.01e+07 3.88e+07 3.84e+07 ...
##  $ Casos Nuevos                : num [1:239] NA 161 NA 574 NA NA NA NA NA NA ...
##  $ Muertes Totales             : num [1:239] 1199031 533475 167642 182439 709765 ...
##  $ Muertes Nuevas              : num [1:239] NA 2 NA 28 NA NA NA NA NA NA ...
##  $ Recuperados Totales         : num [1:239] 1.09e+08 NA 4.00e+07 3.82e+07 3.62e+07 ...
##  $ Nuevos Recuperados          : num [1:239] NA NA NA NA NA NA NA NA NA NA ...
##  $ Casos Activos               : num [1:239] 1114929 NA 0 396245 1448401 ...
##  $ Casos Serios/Criticos       : num [1:239] 1771 NA NA NA NA ...
##  $  Casos Infantiles c/1Millon: num [1:239] 332633 32012 612013 462776 178345 ...
##  $ Muertes c/1Millon           : num [1:239] 3581 379 2556 2175 3296 ...
##  $ Test Totales                : num [1:239] 1.19e+09 9.36e+08 2.71e+08 1.22e+08 6.38e+07 ...
##  $ Test c/1Millon              : num [1:239] 3544577 665334 4139547 1458359 296146 ...
##  $ Poblacion                   : num [1:239] 3.35e+08 1.41e+09 6.56e+07 8.39e+07 2.15e+08 ...
##  - attr(*, "spec")=
```

```
##   .. cols(
##   ..   Country = col_character(),
##   ..   `Total Cases` = col_number(),
##   ..   `New Cases` = col_number(),
##   ..   `Total Deaths` = col_number(),
##   ..   `New Deaths` = col_double(),
##   ..   `Total Recovered` = col_character(),
##   ..   `New Recovered` = col_character(),
##   ..   `Active Cases` = col_character(),
##   ..   `Serious, Critical` = col_character(),
##   ..   `Tot Cases/1M pop` = col_number(),
##   ..   `Deaths/1M pop` = col_number(),
##   ..   `Total Tests` = col_number(),
##   ..   `Tests/1M pop` = col_number(),
##   ..   Population = col_number()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Indagando por el set de datos se puede observar un valor anomalo en la variable "Casos Activos" siendo este -1. Es un valor anomalo porque estamos trabajando con cantidades reales nosotros no podemos tener casos de covid negativos si nos referimos a totales por eso debe ser eliminado.

```
summary(datos['Casos Activos'])
```

```
##  Casos Activos
##  Min.   :      -1
##  1st Qu.:      52
##  Median :     971
##  Mean   :  264825
##  3rd Qu.:    8115
##  Max.   :22242287
##  NA's   :48
```

```
datos <- datos[-82,]
```

Verificamos si existen valores NA en nuestro set de datos, en este caso podemos ver que los hay en gran cantidad en casi todas las variables.

```
sapply(datos , function(x) sum(is.na(x)))
```

```
##                        Pais              Casos Totales
##                           0                          0
##               Casos Nuevos            Muertes Totales
##                         226                          5
##             Muertes Nuevas        Recuperados Totales
##                         231                         49
##          Nuevos Recuperados              Casos Activos
##                         221                         48
```

```
##       Casos Serios/Criticos  Casos Infantiles c/1Millon
##                         178                           9
##           Muertes c/1Millon               Test Totales
##                          14                          26
##              Test c/1Millon                   Poblacion
##                          26                          10
```

La presencia de estos valores faltantes en gran medida llega a ser un problema por la siguiente razon:

(Observaciones del set de datos original)

```
# Observaciones del set de datos original
nrow(datos)
```

```
## [1] 238
```

Al momento de realizar la limpieza vemos que pasamos de tener 238 observaciones a solamente 1 siendo una perdida del 99% de la informacion total.

```
# Observaciones del set de datos limpio de NA's
clean_datos <- na.omit(datos)
nrow(clean_datos)
```

```
## [1] 1
```

Si no existe un mejor set de datos en el cual trabajar lo que podemos hacer para resolver este problema es lo siguiente:

Podemos sacar el porcentaje que representa los valores perdidos existentes en cada variable para ello primero debemos conseguir la cantidad de valores nulos en cada variable individualmente para luego dividirlo por la cantidad de observaciones en el set de datos.

Aqui el resultado:

```
#VEMOS PORCENTAJE DE NA'S POR VARIABLE

# sum(is.na(datos$Pais))/nrow(datos)                          0.0
# sum(is.na(datos$`Casos Totales`))/nrow(datos)               0.0
# sum(is.na(datos$`Casos Nuevos`))/nrow(datos)                0.94
# sum(is.na(datos$`Muertes Totales`))/nrow(datos)             0.02
# sum(is.na(datos$`Muertes Nuevas`))/nrow(datos)              0.97
# sum(is.na(datos$`Recuperados Totales`))/nrow(datos)         0.20
# sum(is.na(datos$`Nuevos Recuperados`))/nrow(datos)          0.92
# sum(is.na(datos$`Casos Activos`))/nrow(datos)               0.20
# sum(is.na(datos$`Casos Serios/Criticos`))/nrow(datos)       0.74
```

```
# sum(is.na(datos$` Casos Infantiles c/1Millon`))/nrow(datos)      0.03
# sum(is.na(datos$`Muertes c/1Millon`))/nrow(datos)                0.05
# sum(is.na(datos$`Test Totales`))/nrow(datos)                     0.10
# sum(is.na(datos$`Test c/1Millon`))/nrow(datos)                   0.10
```

Ahora que sabemos el porcentaje de cada variable podemos definir un criterio para eliminar algunas, el cual sera remover aquellas que posean un porcentaje igual o superior al 50%.

Aqui tenemos las variables que cumplen ese criterio:

```
sum(is.na(datos$`Casos Nuevos`))/nrow(datos)
```

```
## [1] 0.9495798
```

```
sum(is.na(datos$`Muertes Nuevas`))/nrow(datos)
```

```
## [1] 0.9705882
```

```
sum(is.na(datos$`Nuevos Recuperados`))/nrow(datos)
```

```
## [1] 0.9285714
```

```
sum(is.na(datos$`Casos Serios/Criticos`))/nrow(datos)
```

```
## [1] 0.7478992
```

Eliminamos aquellas variables que cumplan con el criterio y comparamos resultados.

```
newdatos <- datos[,-c(3,5,7,9)]
clean_datos_wtv <- na.omit(newdatos)
```

Observaciones del set de datos sin limpiar:

```
nrow(datos)
```

```
## [1] 238
```

Observaciones del set de datos limpio

```
nrow(clean_datos)
```

```
## [1] 1
```

Observaciones del set de datos limpio con el criterio:

```
nrow(clean_datos_wtv)
```

```
## [1] 167
```

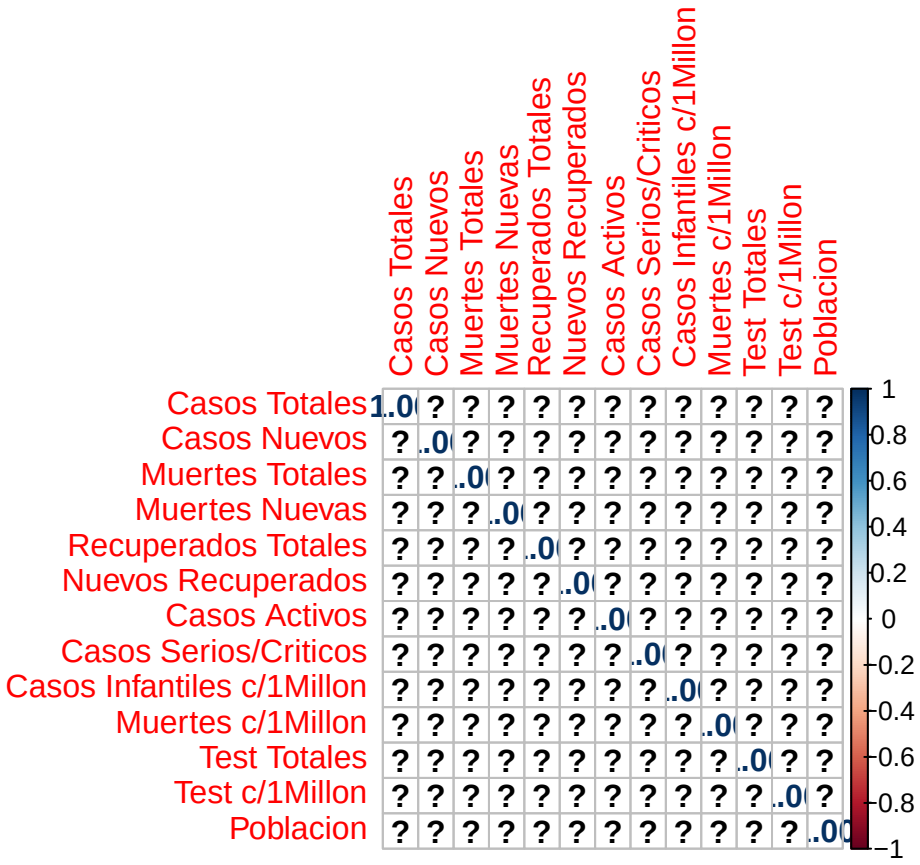
Podemos notar que ahora solamente perdemos un 29.83% de los datos sacando cuentas salvamos un 70.17% de la informacion total por ende el criterio que aplicamos fue bastante efectivo.

En conclusion, esta es la tecnica que utilizaria para salvar la informacion en caso de una perdida masiva en el proceso de ETL, sin embargo, no es la unica forma de hacerlo ya que podriamos contar con otro set de datos mas estable en cuanto a valores NULOS o se podria utilizar tecnicas de promedio si se trabajara con datos numericos como es en este caso.

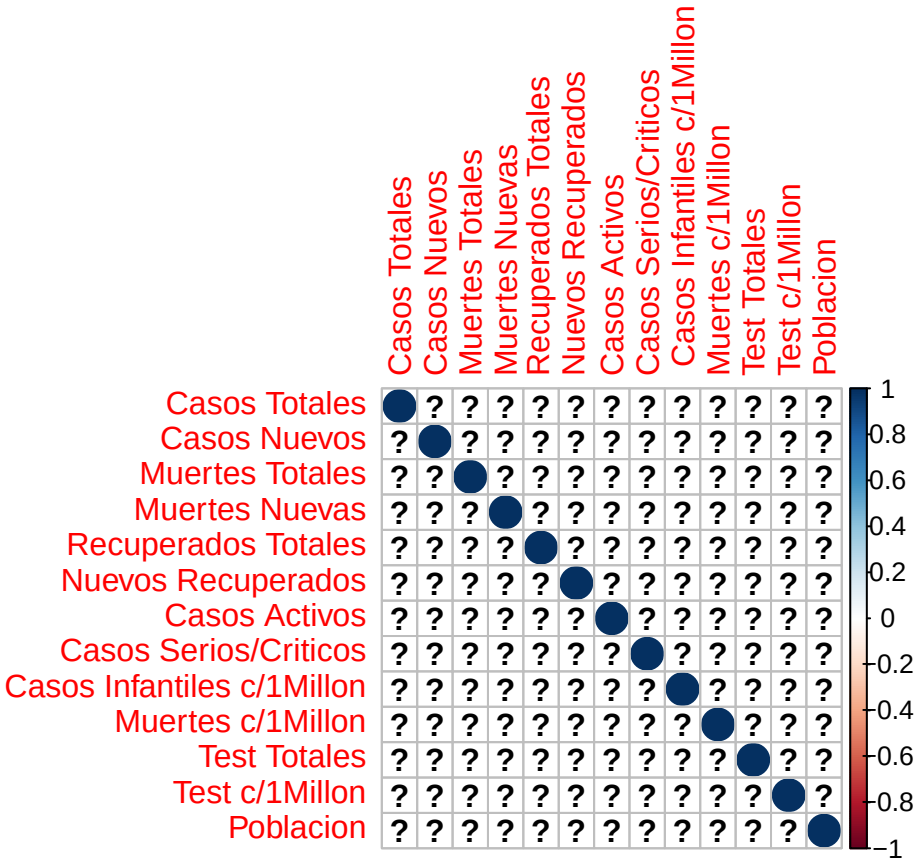
Para finalizar el analisis como apartado extra dejo una matriz de correlacion del data frame original y el que hicimos con el criterio.

La matriz de correlacion no responde bien al data frame original debido a la inmensa cantidad de valores nulos.

```
corr_datos <- cor(datos[,-1])
corrplot(corr_datos, method = 'number')
```

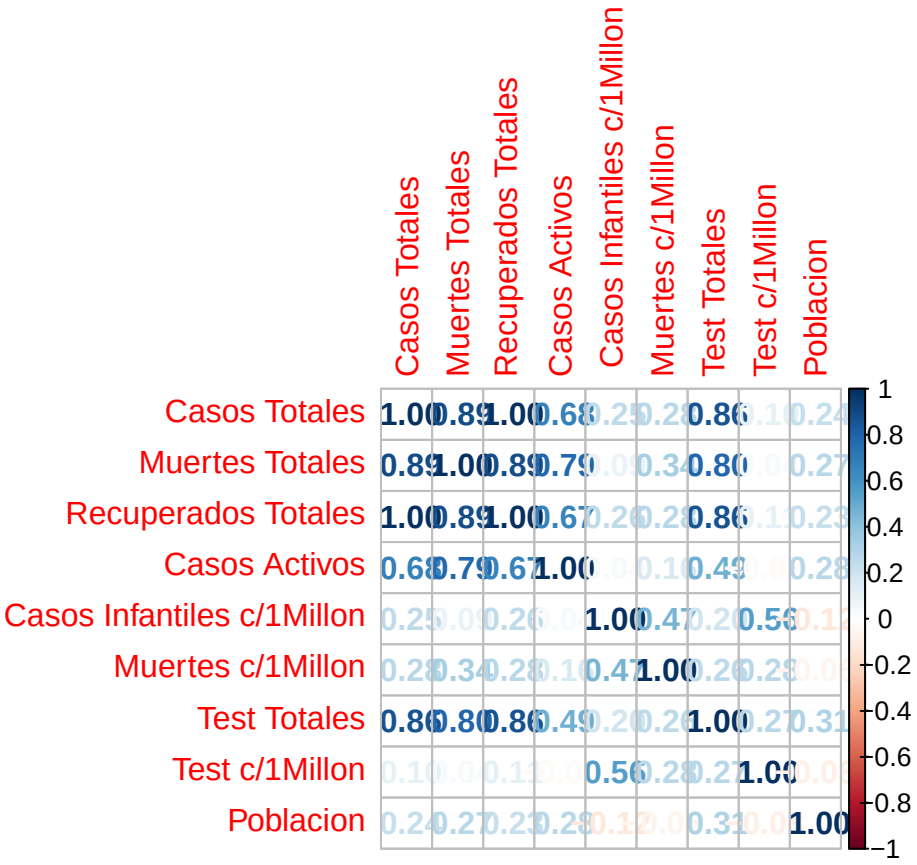


```
corrplot(corr_datos, method = 'circle')
```

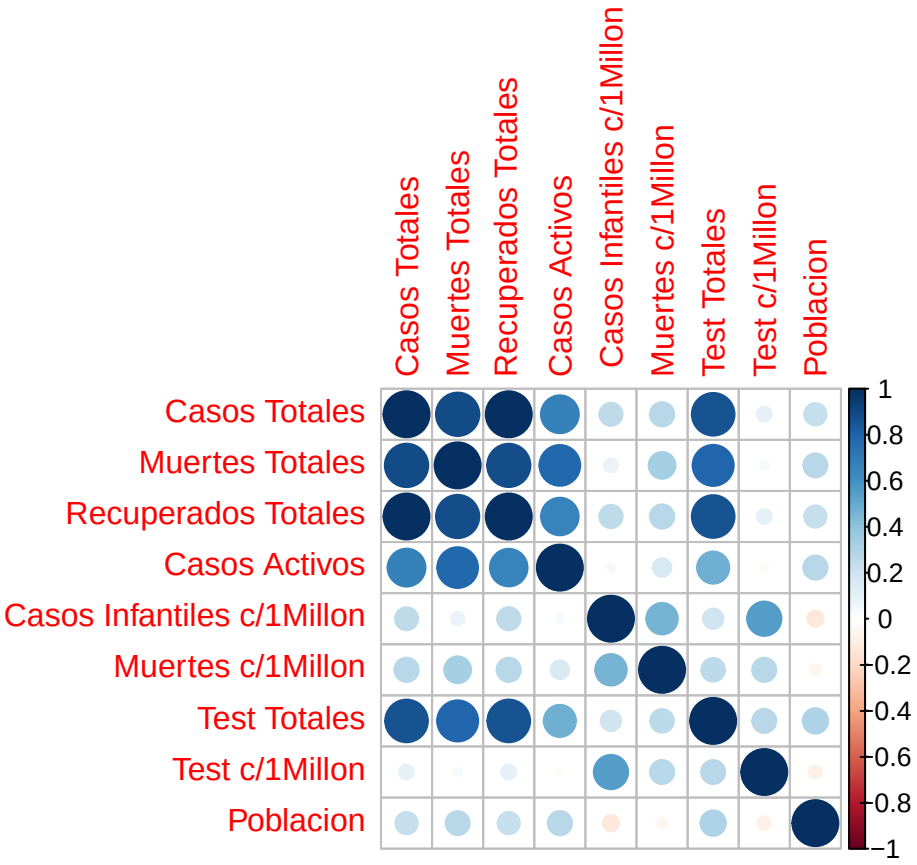


La matriz de correlacion sin aquellas variables con una gran de valores nulos se puede realizar de forma exitosa.

```
corr_datos_2 <- cor(clean_datos_wtv[,-1])
corrplot(corr_datos_2, method = 'number')
```




```
corrplot(corr_datos_2, method = 'circle')
```


Casos Totales
Muertes Totales
Recuperados Totales
Casos Activos
Casos Infantiles c/1Millon
Muertes c/1Millon
Test Totales
Test c/1Millon
Poblacion
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1