UNIVERSIDAD
NACIONAL DEL OESTE

Explotación de Datos

Actividad N^{o} 1

Correlación de Tasa de Movilidad con Casos

Profesores:

Dejean, Gustavo
Españadero, Juan
Mendoza, Dante

Integrantes Grupo B:

Benitez, Nicolas
Garcia Ravlic, Ignacio Agustin
Rechimon, Pablo Hernan
Rodriguez, Miguel Angel

Fecha de entrega:

29 de Agosto de 2020

Índice

Graficos

1 Resumen

En base al contexto de pandemia decidimos estudiar la correlación entre la tasa de movilidad diaria y los casos diarios de cada 10000 habitantes para afirmar o negar que *"a mayor movilidad aumentan los casos diarios"*, para esto contamos con 3 datasets, los cuales manipulamos a través del lenguaje R, dando como resultado que podamos afirmar que el aumento de la movilidad provoque, en alta probabilidad, un aumento de casos.

Palabras Clave: *Analisis de Datos - Programacion - Estudio de Correlacion - COVID-19 - Movilidad Peatonal - Diagramas de Dispersion - Time Series*

---

2 Introducción

2.1 Problemática

Dado el contexto de pandemia, decidimos investigar y analizar la correlacíon entre la tasa de movilidad y los casos diarios de Covid-19 por cada 10000 habitantes en la región sudamericana, específicamente en Chile, Brasil y Argentina.

2.2 Datos a utilizar

Para esto contamos con tres datasets:

- Tasa de Movilidad Mundial en la Pandemia, provisto por Apple
- Confirmados Globales de COVID-19, provisto por Hopkings University
- Habitantes del mundo, provisto por las Organización de las Naciones Unidas

2.3 Objetivo

Confirmar que a mayor tasa de movilidad, mayor cantidad de contagios diarios de Covid-19.

3 Desarrollo

3.1 Análisis de los datos

Analizamos los distintos dataset para verificar la correcta carga de datos y poder empezar a prepararlos. Para esto utilizamos las siguientes funciones:

- str()
- colnames()
- View()
- head()
- tail()

A partir de esto pudimos observar que algunos se encontraban en formato "horizontal", que tenían columnas que no nos interesaban y que los tipos de datos no eran los apropiados, tales observaciones las tuvimos en cuenta a la hora de preparar los datos.

3.2 Preparación de los datos

Preparamos los datos en base a nuestros requerimientos. Para ello, obtenemos datos estadísticos, modificamos el formato de la tabla de "horizontal" a "vertical", los tipos de datos, eliminamos y renombramos columnas, filtramos los datos creando subconjuntos de interés. Para esto se utilizaron las funciones:

- summary()
- sd()
- gather()
- setnames()
- as.Date(), as.factor(), as.character()
- select()
- subset()
- sqldf()

Es importante aclarar que para esto hubo que importar las bibliotecas *RSQLite - sqldf - lubridate - tidyr - dpylr*. También tuvimos que calcular los Casos Diarios por cada 10.000 Habitantes, para esto usamos la estructura iterativa while para calcular esa diferencia entre dia y dia.

Una vez terminada la preparación nos encontrabámos con un *data frame* final, el cual contenida la fecha y de los países de interés sus casos nuevos cada dia y su tasa de movilidad, por ende nos encontrábamos en posición de hacer un análisis a través de las visualizaciones.

3.3 Visualización de datos

Para la visualización de los datos, hemos utilizado las siguientes bibliotecas:

- ggplot2 para graficar scatterplots.
- plotly para gráficos dinámicos.

3.4 Análisis de las visualizaciones

- En la Fig. 1 se puede ver la evolución de la Tasa de Movilidad en el Tiempo, precisamente de los países Argentina, Brasil y Chile. Podemos observar que en un primer segmento la tasa se encuentra arriba del 60% y luego tiene una notable caída de la cual aún se está en recuperación, esto producto de las medidas de prevención tomadas por los diferentes países. A partir de esto podemos decir que las medidas fueron más duras en Argentina y más leves en Brasil.

Fig. 1: Comparación de Time Series de Movilidad 1

- En la Fig. 2 se puede observar lo mismo, pero realizado a través de plotly, lo cual permite interactuar.

Fig. 2: Comparación de Time Series de Movilidad 2

- En la Fig. 3 podemos observar que hay 3 dimensiones a partir de los datos de Argentina únicamente, tenemos la evolución de la movilidad y los casos en el tiempo. Podemos comprobar que a mayor fecha, mayor movilidad y junto a esto, mayor cantidad de casos.

Fig. 3: Time Series de Movilidad con aumento de Casos

- En la Fig. 4 tenemos la dispersión entre la tasa de movilidad y los casos cada 10.000 habitantes, podemos ver que en cada país la correlación es diferente, el caso de Chile se ve un tanto especial, ya que la movilidad se ve casi constante

pero aun asi tiene altos niveles de casos diarios, en cambio entre Brasil y Argentina vemos que la "curva" es bastante similar, donde vemos que el aumento de la movilidad va acompañado del aumento de casos.

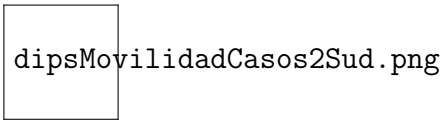

Fig. 4: Dispersion entre Movilidad y Casos

- En la Fig. 5 tenemos una extensión de la figura anterior donde podemos observar unos outliers pertenecientes a los datos de Chile:

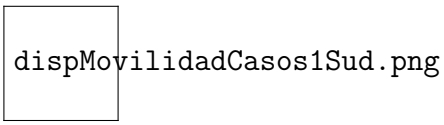

Fig. 5: Dispersion entre Movilidad y Casos completo

- Por último, en la Fig. 6 tenemos la evolución en el tiempo de cantidad de casos confirmados cada 10.000 habitantes desde el 12/03/2020 en Argentina, Chile y Brasil, aquí también presentes los ya nombrados outliers de Chile, pero además podemos ver más picos bruscos tanto en Chile como en Brasil.

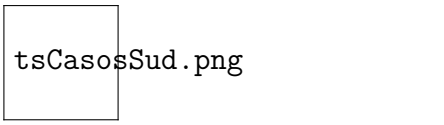

Fig. 6: Comparacion de confirmados c/10.000 habitantes

4 Conclusión

Se puede observar en las figuras, sobre todo en los gráficos de dispersión de los tres paises que la relación entre la tasa de movilidad, los casos diarios por 10000 habitantes y su evolucion en el tiempo va siguiendo la misma tendencia en los tres paises, diagonal hacia arriba, osea a medida que aumenta la tasa, aumentan los casos. Por lo cual podemos afirmar que el aumento de la movilidad provoque, en alta probabilidad, un aumento de casos.

5 Anexo

5.1 Código en R

```r
# Corrleacion entre Casos Diarios Relativos y Movilidad urbana en Argentina vs Chile vs Brasil

# creado: 2020-08-23   v.  2020-8-28
# ultima modificacion: se hicieron los resumenes y se terminaron las vistas

# Autor: GAD, Ignacio Garcia, Pablo Rechimon, Miguel Rodriguez, Nicolas Benitez

# archivo MOVILIDAD:  https://covid19.apple.com/mobility
# archivo HABITANTES:  https://drive.google.com/file/d/1wi9LrbbJXqwmNhnTSvJT1pcSEE_z-nd5/view?usp=sharing

################################################


###################### IMPORTAR BIBLIOTECAS
check_packages <- function(packages) {
  if (all(packages %in% rownames(installed.packages()))) {
    TRUE
  } else{
    cat(
      "Instalar los siguientes packages antes de ejecutar el presente script\n",
      packages[!(packages %in% rownames(installed.packages()))],
      "\n"
    )
  }
}
packages_needed <- c("ggplot2", "plotly", "sqldf", "lubridate",
                     "tidyr", "data.table", "readr", "dplyr", "RColorBrewer" )
check_packages(packages_needed)

library(gsubfn)
library(proto)
library(RSQLite)
library(sqldf)
library(data.table)
library(lubridate)
library(tidyr)
library(readr)
library(ggplot2)
library(plotly)
library(dplyr)
library(RColorBrewer)

options(scipen = 6) #para evitar notacion cientifica

```

```r


###################### LEER DATOS DEL COVID
# URL con datos del COVID-19
URL           <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/"

# time_series_covid19_confirmed_global.csv    este es el archivo que a leer
url_archivo  <- paste(URL,"time_series_covid19_confirmed_global.csv", sep = "")

COVID_19_h    <- read.csv(url_archivo, sep = ",", header = T)

#Analizamos lo leido y corroboramos que se leyo bien
str(COVID_19_h)


###################### PREPARAR SUBSET DE DATOS PARA ARGENTINA, BRASIL y CHILE
colnames(COVID_19_h)
#Eliminamos columnas que no nos interesan
COVID_19_h <- select(COVID_19_h, -c(Lat, Long, Province.State))

#Cambiamos el nombre de una columna a algo mas entendible
setnames(COVID_19_h, "Country.Region", "pais")

#Pasamos de formato horizontal a vertical
COVID_19 <- COVID_19_h %>% gather(fecha, acumCasos, 2:ncol(COVID_19_h))


str(COVID_19)
#Cambiamos el formato de la fecha
COVID_19$fecha <- as.Date(COVID_19$fecha, format = "X%m.%d.%y")


#Quedarse solo con datos de ARG
ARG_target = "Argentina"
C19_ARG <- subset(COVID_19, pais == ARG_target)

#Quedarse solo con datos de BR
BR_target = "Brazil"
C19_BR <- subset(COVID_19, pais == BR_target)

#Quedarse solo con datos de CHL
CHL_target = "Chile"
C19_CHL <- subset(COVID_19, pais == CHL_target)

C19_Sudamerica <- sqldf("
   SELECT arg.fecha fecha, arg.acumCasos acumCasosARG,
      br.acumCasos acumCasosBR, chl.acumCasos acumCasosCHL
   FROM C19_ARG arg, C19_BR br, C19_CHL chl
   WHERE arg.fecha==br.fecha AND arg.fecha==chl.fecha")

#Analizamos la salida del join entre los casos de cada pais
```

```r
str(C19_Sudamerica)
C19_Sudamerica


##################### LEER DATOS DE HABITANTES DEL MUNDO
data_habitantes <- read.csv("C:/work/Explotacion/Actividad1/habitantesMundo.csv",
                            sep=";", stringsAsFactors=TRUE)

#Analizamos lo leido y corroboramos que se leyo bien
str(data_habitantes)



##################### PREPARAR DATOS DE HABITANTES
#Quedarse solo con datos de ARG, BR y CHL
SUD_habitantes <- subset(data_habitantes, pais == ARG_target |
                         pais == BR_target |
                         pais == CHL_target)

#Eliminar columnas extras
colnames(SUD_habitantes)
SUD_habitantes <- select(SUD_habitantes, -c(Variant, Index, Notes, codigo, Type))

#Analizamos la salida
str(SUD_habitantes)
SUD_habitantes


##################### CALCULAR CASOS DIARIOS EN RELACION A LOS HABITANTES
#Agregar columnas para los casos
C19_Sudamerica$casosARG <- 0
C19_Sudamerica$casosBR <- 0
C19_Sudamerica$casosCHL <- 0

C19_Sudamerica

#Contador y tope
i <- 1
max <- nrow(C19_Sudamerica)

print(max)#Visualizar las filas

#Recorrer y calcular casos diarios de cada 10.000 habitantes
while (i<=max) {
  if(C19_Sudamerica$acumCasosARG[i] > 0){
    C19_Sudamerica$casosARG[i] <- (C19_Sudamerica$acumCasosARG[i] - C19_Sudamerica$acumCasosARG[i-1]) * 10000 / SUD_habitantes$cantidad[1]
  }
  if(C19_Sudamerica$acumCasosBR[i] > 0){
    C19_Sudamerica$casosBR[i] <- (C19_Sudamerica$acumCasosBR[i] - C19_Sudamerica$acumCasosBR[i-1]) * 10000 / SUD_habitantes$cantidad[2]
```

```
  }
  if(C19_Sudamerica$acumCasosCHL[i] > 0){
    C19_Sudamerica$casosCHL[i] <- (C19_Sudamerica$acumCasosCHL[i] -
    C19_Sudamerica$acumCasosCHL[i-1]) * 10000 / SUD_habitantes$
   cantidad[3]
  }
  i <- i+1
}

C19_Sudamerica



###################### LEER DATOS DE MOVILIDAD
data_movilidad_h <- read_csv("applemobilitytrends-2020-08-26.csv",
   locale = locale(grouping_mark = ""))

#Analizamos lo leido y corroboramos que se leyo bien
str(data_movilidad_h)


###################### PREPARAR DATOS
colnames(data_movilidad_h)
setnames(data_movilidad_h, "region", "pais")
setnames(data_movilidad_h, "transportation_type", "transporte")

data_movilidad_h <- select(data_movilidad_h, -c(geo_type,
   alternative_name, country, 'sub-region'))

data_movilidad_h$pais <- as.factor(data_movilidad_h$pais)
data_movilidad_h$transporte <- as.factor(data_movilidad_h$
   transporte)

#Pasar a formato vertical
movilidad <- data_movilidad_h     %>% gather(fecha, tasa, 3:ncol(
   data_movilidad_h))

#Formatear fecha
movilidad$fecha <- as.Date(as.character(movilidad$fecha))

#Corroboramos que los cambios se efectuaron bien
str(movilidad)

#Quedarse solo con datos de ARG, BR y CHL
MOV_ARG <- subset(movilidad, transporte == "walking" & pais == ARG_
   target)
MOV_BR <- subset(movilidad, transporte == "walking" & pais == BR_
   target)
MOV_CHL <- subset(movilidad, transporte == "walking" & pais == CHL_
   target)


#Juntar los subsets de movilidad
MOV_Sudamerica <- sqldf("
  SELECT arg.fecha fecha, arg.tasa tasaARG, br.tasa tasaBR, chl.
   tasa tasaCHL
```

```
  FROM MOV_ARG arg, MOV_BR br, MOV_CHL chl
  WHERE arg.fecha == br.fecha AND arg.fecha == chl.fecha")

#Analizamos que el join no afecto y salio bien
str(MOV_Sudamerica)
MOV_Sudamerica



###################### JOIN ENTRE MOVILIDAD Y CASOS
SUD_MovC19 <- sqldf("
  SELECT mov.fecha fecha, c19.casosARG, mov.tasaARG, c19.casosBR,
   mov.tasaBR, c19.casosCHL, mov.tasaCHL
  FROM C19_Sudamerica c19, MOV_Sudamerica mov
  WHERE mov.fecha == c19.fecha")

#Analizamos que el join no afecto y salio bien
str(SUD_MovC19)
SUD_MovC19


#Calculamos y analizamos medidas estadisticas
desviaciones <- c( #Desviaciones tipicas
sd(SUD_MovC19$casosARG, na.rm = T),
sd(SUD_MovC19$tasaARG, na.rm = T),
sd(SUD_MovC19$casosBR, na.rm = T),
sd(SUD_MovC19$tasaBR, na.rm = T),
sd(SUD_MovC19$casosCHL, na.rm = T),
sd(SUD_MovC19$tasaCHL, na.rm = T))

desviaciones


resumen <- summary(SUD_MovC19) #Minimos, Maximos, Quartiles y
   Promedio

resumen



###################### GUARDAR DATOS
write.csv2(C19_Sudamerica, "datos_casosXDia.csv",  row.names =
   FALSE, fileEncoding = "UTF-8")
write.csv2(MOV_Sudamerica, "datos_movilidadXDia.csv",  row.names =
   FALSE, fileEncoding = "UTF-8")
write.csv2(SUD_MovC19, "datos_correlacion.csv",  row.names = FALSE,
    fileEncoding = "UTF-8")



###################### GENERAR FIGURA
colores <- c("blue", "lightblue", "yellow", "orange", "red", "
   darkred")

#Time Series de Movilidad en Arg, Br y Chl con qplot
```

```
fig1 <- ggplot(SUD_MovC19, aes(x = fecha, y = tasaARG)) +
  geom_line(aes(y = tasaBR), color="darkgreen") +
  geom_line(aes(y = tasaCHL), color="darkred") +
  geom_line(size = 0.6) +
  ggtitle(paste("COVID-19 - Movilidad",sep = "")) +
  scale_x_date(date_breaks = "7 day", date_labels =  "%d %b") +
  theme(plot.title = element_text(lineheight = 1,face ='bold'))   +
  ylab("Tasa de Movilidad") +
  xlab("") +
  labs(caption = "Fuente de los datos: apple.com/covid19/mobility")
    +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 7)
   )
fig1


#Time Series de Movilidad en Arg, Br y Chl con plotly
fig2 <- plot_ly(SUD_MovC19,  x = ~fecha, y = ~tasaARG,
     name = 'Argentina', type = 'scatter', mode = 'lines') %>%
  add_trace(y = ~tasaBR, name = 'Brasil', mode = 'lines') %>%
  add_trace(y = ~tasaCHL, name = 'Chile', mode = 'lines') %>%
  layout(title = "Time Series de Movilidad",
         xaxis = list(title = "Fecha", type = "date",
                      tickmode = "linear", tick0 = min(SUD_MovC19$
   fecha),
                      tickformat = "%d/%m", dtick = 86400*10000,
   tickangle = 75),
         yaxis = list (title = "Tasa de Movilidad", tickangle =
   -45))

fig2


#Time Series de Movilidad en Arg, con color en relacion a cantidad
   de casos con plotly
fig3 <- qplot(data = SUD_MovC19, x = fecha, y = tasaARG, colour =
   casosARG, geom = "line") +
  geom_line(size = 1) +
  scale_colour_gradient2(low = "lightblue", mid="orange", high = "
   red",
                         midpoint = max(SUD_MovC19$casosARG)/2) +
  ggtitle("Time Series de Movilidad con aumento de Casos") +
  scale_x_date(date_breaks = "7 day", date_labels =  "%d/%m") +
  theme(plot.title = element_text(lineheight = 1,face ='bold'))   +
  ylab("Tasa de Movilidad") +
  xlab("") +
  labs(caption = "Fuente de los datos: apple.com/covid19/mobility")
    +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 7)
   )
fig3


#Movilidad con calor de Arg con plotly
```

```
fig4 <- plot_ly(data = SUD_MovC19,  x = ~fecha, y = ~tasaARG, name = "Argentina",
                type = 'scatter', mode = 'markers', color = ~SUD_MovC19$casosARG "Casos Arg",
                size = 1, colors = colores) %>%
  layout(title = "Time Series de Movilidad con aumento de Casos",
         xaxis = list(title = "", type = "date",
                      tickmode = "linear", tick0 = min(SUD_MovC19$fecha),
                      tickformat = "%d/%m", dtick = 86400*10000, tickangle = 75,
                      range=c(min(SUD_MovC19$fecha), max(SUD_MovC19$fecha))),
         yaxis = list (title = "Tasa de Movilidad", tickangle = -45))
fig4



#Filtramos puntos antes de que llegue el primer caso
filtered_SUD_MovC19 <- sqldf("SELECT * FROM SUD_MovC19
                      WHERE casosARG > 0 AND casosBR > 0 AND casosCHL > 0")


#Correlacion entre Movilidad y Casos en Arg
fig5 <- plot_ly(data = filtered_SUD_MovC19,  x = ~casosARG, y = ~tasaARG, name = "Argentina",
                type = 'scatter', mode = 'markers') %>%
  layout(title = "Dispercion entre Tasa de Movilidad y Casos",
         xaxis = list(title = "Casos c/10.000 hab"),
         yaxis = list (title = "Tasa de Movilidad", tickangle = -45))
fig5


#Correlacion entre Movilidad y Casos en Arg, Br y Chl
fig6 <- plot_ly(data = filtered_SUD_MovC19,  x = ~casosARG, y = ~tasaARG, name = "Argentina",
                type = 'scatter', mode = 'markers') %>%
  add_trace(x = ~casosBR, y = ~tasaBR, name = 'Brasil', mode = 'markers') %>%
  add_trace(x = ~casosCHL, y = ~tasaCHL, name = 'Chile', mode = 'markers') %>%
  layout(title = "Dispercion entre Tasa de Movilidad y Casos",
         xaxis = list(title = "Casos c/10.000 hab"),
         yaxis = list (title = "Tasa de Movilidad", tickangle = -45))
fig6


#Correlacion entre Movilidad y Casos en Arg, Br y Chl filtrando tasa mayor a 100 y casos mayor a 4
fig7 <- plot_ly(data = filtered_SUD_MovC19,  x = ~casosARG, y = ~tasaARG, name = "Argentina",
                type = 'scatter', mode = 'markers') %>%
```

```
328   add_trace(x = ~casosBR, y = ~tasaBR, name = 'Brasil', mode = '
      markers') %>%
329   add_trace(x = ~casosCHL, y = ~tasaCHL, name = 'Chile', mode = '
      markers') %>%
330   layout(title = "Dispercion entre Tasa de Movilidad y Casos",
331          xaxis = list(title = "Casos c/10.000 hab", range= c(0, 4))
      ,
332          yaxis = list (title = "Tasa de Movilidad", tickangle =
      -45, range= c(0, 100)))
333 fig7
334
335
336 #Time Series de casos de Covid-19 en Arg, Br y Chl
337 fig8 <- plot_ly(data = filtered_SUD_MovC19,  x = ~fecha, y = ~
      casosARG, name = "Argentina",
338                 type = 'scatter', mode = 'lines') %>%
339   add_trace(y = ~casosBR, name = 'Brasil', mode = 'lines') %>%
340   add_trace(y = ~casosCHL, name = 'Chile', mode = 'lines') %>%
341   layout(title = "Time Series de Confirmados c/10.000 habitantes",
342          xaxis = list(title = "Fecha", type = "date",
343                       tickmode = "linear", tick0 = min(filtered_SUD
      _MovC19$fecha),
344                       tickformat = "%d/%m", dtick = 86400*10000,
      tickangle = 75),
345          yaxis = list (title = "Nro de Casos", tickangle = -45))
346 fig8
```