Escuela de Gobierno y Transformación Pública
Tecnológico de Monterrey

TAREA # 1, PARTE 2

Ciencia de datos

Entrega: 22 de febrero del 2021

Ejercicios resueltos en colaboración con Nagib Gobera Mac Farland, Elva Deyanira Martínez González, Álvaro Isaac Vázquez Aguilar y Elvia Daniela Flores Resendez

Guillermo Alberto García Candanosa
A01034958@itesm.mx

Generales

Corre el siguiente código en R para descargar los resultados del Censo 2020 de INEGI. En esta tarea trabajaremos con estos datos.

```
#### Generales ####

rm(list =ls())


# Descarga el archivo del censo 2020 por AGEB para # Nuevo Leon

download.file (url =

'https://www.inegi.org.mx/contenidos/programas/ccpv/2020/microdatos/ageb_manzana/RESAGEBURB_19_2020_csv.zip',

destfile ='censo20_NL.zip', method ='curl')


# Descomprime el archivo . zip

zipF <- "censo20_NL.zip"

outDir <-"./unzipfolder20"

unzip (zipF , exdir = outDir )


# Lee el archivo que descomprimes .

d20 <- read.csv(

  paste(outDir, "/", "RESAGEBURB_19CSV20.csv", sep =""),

  encoding ="UTF-8", na.strings ="*",

  stringsAsFactors = FALSE )

head ( d20 )
```

El resultado del código anterior debe de ser un objeto en memoria llamado "d20" con la base de datos de manzanas y Áreas Geoestadísticas Básicas (AGEBs) del Censo del 2020 de INEGI. Consulta el pdf adjunto en el Canvas de esta tarea que contiene el diccionario de datos de la base que acabas de descargar. Esta base de datos contiene los resultados por manzana, AGEB, localidad, municipio y estado para Nuevo León. Explora el contenido de la base y sus variables.

1. Tidyverse y SQL

Utiliza Tidyverse o SQL en R para resolver los siguientes ejercicios. El objetivo es que generes una base de datos con los resultados de las manzanas y generes una serie de estadísticos.

(a) Filtra tu base de datos de tal forma que solamente despliegue los resultados de las manzanas, elimina los totales por AGEB, localidad, municipio y estado. Quédate solamente con las manzanas.

```
# Función nueva
`%!in%` = Negate("%in%")

# Ejercicio 1a
d20_1a <- d20 %>%
  filter(NOM_LOC %!in% c("Total de la entidad", "Total del municipio",
                         "Total de la localidad urbana", "Total AGEB urbana"))

# otra forma de llegar al mismo resultado
d20_1a2 <- d20 %>%
  filter(MZA != 0)
```

(b) Filtra tu base de datos del punto anterior, de tal forma que contenga solamente las manzanas de los municipios de la ZMM (once o dieciocho municipios, según la definición que utilices) y guárdalo como un nuevo objeto. En este objeto selecciona solamente las columnas siguientes: grado promedio de escolaridad, población total, viviendas particulares habitadas, viviendas con internet, promedio de ocupantes por cuarto, población sin afiliación a servicios de salud.

```
# Ejercicio 1b
d20_1b <- d20_1a %>%
  filter(NOM_MUN %in% c("Monterrey", "San Nicolás de los Garza", "San Pedro Garza
García", "Santa Catarina", "Guadalupe", "General Escobedo", "Apodaca",
"Juárez", "García", "Cadereyta Jiménez", "Pesquería", "El Carmen",
"General Zuazua", "Salinas Victoria", "Ciénega de Flores",
"Santiago", "Hidalgo", "Abasolo"))

d20_1b2 <- d20_1b %>%
  select(GRAPROES,POBTOT,TVIVPARHAB,VPH_INTER,PRO_OCUP_C,PSINDER)
```

(c) En la base de datos del inciso anterior, elimina las manzanas que contengan valores nulos en alguna de las variables que seleccionaste.

```
# Ejercicio 1c

d20_1c <- d20_1b2 %>%
  na.exclude() %>%
  filter(POBTOT != 0)
```

(d) Con esta base de datos final, genera una tabla que despliegue el mínimo, máximo y promedio de grado promedio escolaridad por municipio a partir de las manzanas. Es decir, agrega las manzanas por municipio y genera el mínimo, máximo y promedio. Presenta tu tabla en los resultados, así como el código de R.

```
# Ejercicio 1d

d20_1d <- d20_1b %>%

  select(NOM_MUN,
GRAPROES,POBTOT,TVIVPARHAB,VPH_INTER,PRO_OCUP_C,PSINDER) %>%

  na.exclude() %>% filter(POBTOT != 0) %>%

  group_by(NOM_MUN) %>%

  transmute(GRAPROES_NUM = as.numeric(GRAPROES)) %>%

  filter(!is.na(GRAPROES_NUM)) %>%

  summarise(MIN = min(GRAPROES_NUM), PROM = mean(GRAPROES_NUM), MAX =
max(GRAPROES_NUM)) %>%

  format.data.frame(digits=3)
```

Tabla 1 Mínimo, máximo y promedio de grado promedio de escolaridad por municipio a partir de las manzanas.

NOM_MUN	MIN	PROM	MAX
Apodaca	4.88	11.55	17.2
Cadereyta Jiménez	3.75	9.76	16.7
Ciénega de Flores	5.82	9.4	14.1
El Carmen	6.56	9.5	15.8
García	4.64	10.62	18.3
General Escobedo	3.83	10.61	17.6
General Zuazua	5.33	9.68	14.9
Guadalupe	4.1	11.11	18
Hidalgo	4.33	9.92	15.2
Juárez	4	9.95	17.8
Monterrey	0	11.48	19
Pesquería	4.11	9.55	16.1
Salinas Victoria	5	9.51	13.6
San Nicolás de los Garza	3.43	11.94	17.2
San Pedro Garza García	5.23	13.23	18.4
Santa Catarina	0	10.73	18.6
Santiago	4.71	10.8	16.8

Regresión Lineal

(a) Con la base de datos del ejercicio anterior, calcula las siguientes dos nuevas variables: a) promedio de viviendas particulares habitadas con internet; b) porcentaje de la población sin afiliación a servicios de salud.

```
# Ejercicio 2a
```

```
#install.packages("scales")
library(scales)

d20_2a <- d20_1b %>%

select(NOM_MUN,GRAPROES,POBTOT,TVIVPARHAB,VPH_INTER,PRO_OCUP_C,PSINDER
) %>%
  na.exclude() %>%
  filter(POBTOT != 0) %>%
  group_by(NOM_MUN) %>%
  mutate(GRAPROES_NUM = as.numeric(GRAPROES),VPH_INTER_NUM =
as.numeric(VPH_INTER),PSINDER_NUM = as.numeric(PSINDER),
         TVIVPARHAB_NUM = as.numeric(TVIVPARHAB)) %>%
  transmute(GRAPROES_NUM, VPH_INTER_NUM, POBTOT, PSINDER_NUM,
TVIVPARHAB_NUM) %>%
  na.exclude() %>%
  summarise(MIN_GRAPROES = min(GRAPROES_NUM),
            PROM_GRAPROES = mean(GRAPROES_NUM),
            MAX_GRAPROES = max(GRAPROES_NUM),
            PER_VPH_INTER =
label_percent()(sum(VPH_INTER_NUM)/sum(TVIVPARHAB_NUM)),
            PER_PSINDER = label_percent()(sum(PSINDER_NUM)/sum(POBTOT))) %>%
  format.data.frame(digits=3)
```

Tabla 2 Mínimo, máximo y promedio de grado promedio de escolaridad, porcentaje de viviendas particulares habitadas con internet y porcentaje de personas sin afiliación a servicios de salud por municipio a partir de las manzanas.

NOM_MUN	MIN_GRAPROES	PROM_GRAPROES	MAX_GRAPROES	PER_VPH_INTER	PER_PSINDER
Apodaca	4.88	11.55	17.2	81%	16%
Cadereyta Jiménez	3.75	9.76	16.7	59%	20%
Ciénega de Flores	5.82	9.4	14.1	52%	21%
El Carmen	6.56	9.5	15.8	52%	22%
García	4.64	10.62	18.3	62%	21%
General Escobedo	3.83	10.61	17.6	71%	22%
General Zuazua	5.33	9.68	14.9	61%	18%
Guadalupe	4.1	11.11	18	78%	20%
Hidalgo	4.33	9.92	15.2	50%	13%
Juárez	4	9.95	17.8	62%	22%
Monterrey	0	11.48	19	78%	20%
Pesquería	4.11	9.55	16.1	50%	19%
Salinas Victoria	5	9.51	13.6	50%	21%

San Nicolás de los Garza	3.43	11.94	17.2	83%	16%
San Pedro Garza García	5.23	13.23	18.4	87%	13%
Santa Catarina	0	10.73	18.6	77%	18%
Santiago	4.71	10.8	16.8	68%	15%

(b) Estima una regresión lineal de grado promedio de escolaridad por manzana sobre porcentaje de viviendas con internet. Presenta el resultado e interpreta el coeficiente.

```
# Ejercicio 2b

d20_2b <- d20_1a %>%
  select(GRAPROES, TVIVPARHAB, VPH_INTER, POBTOT) %>%
  na.exclude() %>%
  filter(POBTOT != 0) %>%
  select(GRAPROES, TVIVPARHAB, VPH_INTER) %>%
  transmute(GRAPROES_NUM = as.numeric(GRAPROES), TVIVPARHAB_NUM =
as.numeric(TVIVPARHAB),
            VPH_INTER_NUM = as.numeric(VPH_INTER)) %>%
  na.exclude() %>%
  transmute(GRAPROES_NUM, PER_VPH_INTER =
100*(VPH_INTER_NUM/TVIVPARHAB_NUM))

ols_2b <- lm(GRAPROES_NUM ~ PER_VPH_INTER, data = d20_2b)
cor(d20_2b$PER_VPH_INTER,d20_2b$GRAPROES_NUM) #grado de asociación entre
variables

summary(ols_2b)
```

Resultado:

```
Call:
lm(formula = GRAPROES_NUM ~ PER_VPH_INTER, data = d20_2b)

Residuals:
    Min      1Q  Median      3Q     Max
-13.1472  -1.1631  -0.1527   1.1560  11.6560

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3440326  0.0242729   220.2   <2e-16 ***
PER_VPH_INTER 0.0780313  0.0003228   241.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 1.677 on 57323 degrees of freedom
Multiple R-squared: 0.5048, Adjusted R-squared: 0.5048
F-statistic: 5.843e+04 on 1 and 57323 DF, p-value: < 2.2e-16

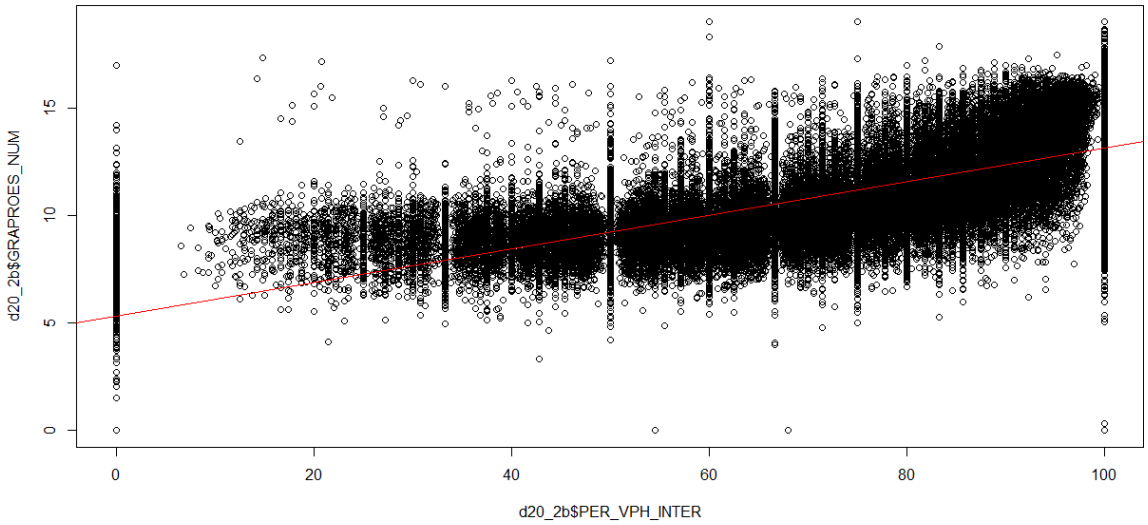
Interpretación:

Por cada unidad adicional de porcentaje de viviendas particulares habitadas con internet, el grado promedio de escolaridad de la manzana evaluada aumenta en promedio 0.0780.

(c) Dibuja un plot donde grafiques grado promedio de escolaridad contra porcentaje de viviendas, agregando al plot de puntos una línea en color rojo que represente tu línea de regresión. ¿Qué te dice este plot?

```
# Ejercicio 2c

plot(d20_2b$PER_VPH_INTER,d20_2b$GRAPROES_NUM)
abline(lm(d20_2b$GRAPROES_NUM ~ d20_2b$PER_VPH_INTER),col="red")
```



Interpretación:

Existe una correlación positiva entre el porcentaje de viviendas con internet y el grado promedio de escolaridad en las manzanas.

(d) Estima una regresión lineal de grado promedio de escolaridad por manzana sobre porcentaje de viviendas con internet, promedio de ocupantes por cuarto y porcentaje de población sin afiliación a servicios de salud. Presenta los resultados e interpreta. ¿Cómo cambiaron los estimadores respecto al ejercicio anterior? ¿Por qué?

```
# Ejercicio 2d

d20_2d <- d20_1a %>%
  select(GRAPROES, TVIVPARHAB, VPH_INTER, POBTOT, PSINDER, PRO_OCUP_C) %>%
  na.exclude() %>%
```

```
  filter(POBTOT != 0) %>%
  transmute(POBTOT, GRAPROES_NUM = as.numeric(GRAPROES), TVIVPARHAB_NUM =
as.numeric(TVIVPARHAB),
            VPH_INTER_NUM = as.numeric(VPH_INTER), PSINDER_NUM =
as.numeric(PSINDER),
            PRO_OCUP_C_NUM = as.numeric(PRO_OCUP_C)) %>%
  na.exclude() %>%
  transmute(GRAPROES_NUM, PER_VPH_INTER =
100*(VPH_INTER_NUM/TVIVPARHAB_NUM),
            PRO_OCUP_C_NUM, PER_PSINDER = 100*(PSINDER_NUM/POBTOT))

ols_2d <- lm(GRAPROES_NUM ~
               PER_VPH_INTER + PRO_OCUP_C_NUM + PER_PSINDER, data = d20_2d)
summary(ols_2d)
```

Resultado:

```
Call:
lm(formula = GRAPROES_NUM ~ PER_VPH_INTER + PRO_OCUP_C_NUM +
    PER_PSINDER, data = d20_2d)

Residuals:
     Min       1Q   Median       3Q      Max
-14.3782  -0.9887  -0.0349   1.0110   9.9650

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.6649174  0.0451563  214.03   <2e-16 ***
PER_VPH_INTER   0.0576895  0.0003480  165.77   <2e-16 ***
PRO_OCUP_C_NUM -2.6390762  0.0297575  -88.69   <2e-16 ***
PER_PSINDER    -0.0319973  0.0005997  -53.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.462 on 52807 degrees of freedom
Multiple R-squared:  0.6047,       Adjusted R-squared:  0.6047
F-statistic: 2.693e+04 on 3 and 52807 DF,  p-value: < 2.2e-16
```

Interpretación:

- Por cada unidad adicional de porcentaje de viviendas particulares habitadas con internet, manteniendo constantes el promedio de ocupantes por cuarto y el porcentaje de población sin afiliación a servicios de salud, el grado promedio de escolaridad de la manzana evaluada aumenta en promedio 0.0577.
- Por cada unidad adicional de promedio de ocupantes por cuarto, manteniendo constantes el porcentaje de viviendas particulares habitadas con internet y el porcentaje de población sin afiliación a servicios de salud, el grado promedio de escolaridad de la manzana evaluada disminuye en promedio 2.6391.

- Por cada unidad adicional de porcentaje de población sin afiliación a servicios de salud, manteniendo constantes el porcentaje de viviendas particulares habitadas con internet y el promedio de ocupantes por cuarto, el grado promedio de escolaridad de la manzana evaluada disminuye en promedio 0.0320.
- El intercepto aumentó porque entraron en juego dos variables más en el modelo que tienen una correlación negativa con el grado promedio de escolaridad por manzana. La pendiente de la variable de porcentaje de viviendas particulares habitadas con internet disminuyó como reacción al aumento del intercepto.

(e) ¿Tu regresión tiene el sesgo de la variable omitida o no? ¿Por qué? ¿Son confiables tus estimadores? Explique.

Si hay sesgo de la variable omitida porque existen otras variables que se relacionan tanto con el Grado Promedio de Escolaridad como con el acceso a internet, los ocupantes por cuarto y el porcentaje de población sin afiliación a servicios de salud. Ejemplos serían los ingresos promedio per cápita de la población y la zona donde se encuentre ubicada esa manzana. Los estimadores son confiables en la medida en la que éstos se utilicen para llevar a cabo una predicción del grado promedio de escolaridad, dada la información de las tres variables independientes. Sin embargo, no se puede concluir que exista una relación causal entre las variables evaluadas.