Análisis Exploratorio_v2

Jorge Luis Moreno Monreal

04/8/2020

En este markdown se exploraran las varaibles del DataFrame_Variables_brutas.csv y se generaran estadísticas desriptivas y visualizaciones que nos permitirán crear nuestro modelo de regresión lineal. Además, dichas visualizaciones nos permitiran redactar los primeros comentarios en el capítulo de resultados.

- Cargamos paquetería

```
if (!require(pacman)) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
library(pacman)
p_load("readstata13" , #Lector STATA .DTA
       "tidyverse",    #ManipulaciÃ³n de datos
       "survey",       #Marco Estadistico
       "stargazer",    #PresentaciÃ³n de tablas
       "readxl",       #Excel
       "corrplot",     #Visualización de Correlaciones
       "knitr","tinytex") #Knit pdf
```

- Cargamos el .csv

```
df <- read.csv("DataFrame_Variables_brutas.csv")
```

Descriptor de Archivos

Tenemos una base de datos con 62 variables:

- Sociodemográficas - 8
- Salud (dependientes) - 3
- Construcción de la salud - 13 (No forman parte del modelo)
- Salud (Enfermedades) - 6
- Económicas - 25
- Variables de identificacion - 5
- Factor de Expansión - 1

n=14,779

Sociodemográficas

- sexo_mujer
- edad
- edad2
- escolaridad
- escolaridad2
- rural
- altamigración
- n_hijos (Número de hijos)

```
stargazer(df[c("sexo_mujer","edad","edad2","escolaridad","escolaridad2","rural",
               "altamigración","n_hijos")])
```

Table 1: Resumen de variables dependientes de sociodemográficas

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
sexo_mujer	14,779	0.584	0.493	0	0	1	1
edad	14,775	66.131	10.662	22.000	58.000	73.000	113.000
edad2	14,775	4,486.955	1,440.643	484.000	3,364.000	5,329.000	12,769.000
escolaridad	14,604	5.767	4.748	0.000	2.000	9.000	22.000
escolaridad2	14,604	55.808	78.538	0.000	4.000	81.000	484.000
rural	14,779	0.428	0.495	0	0	1	1
altamigración	14,779	0.380	0.485	0	0	1	1
n_hijos	14,127	5.085	3.208	0.000	3.000	7.000	23.000

Variables de Salud Dependientes

Este conjunto de variables serán las que se pretenden utilizar en el modelo como variables dependientes (Y) para evaluar la salud de los individuos.

- aes = Auto Evaluación de la Salud
- laf = Limitación en Actividades Fisicas
- lavd = Limitaciones en Actividades de la Vida Diaria

```
stargazer(df[c("aes","laf","lavd")])
```

Table 2: Resumen de variables dependientes de Salud

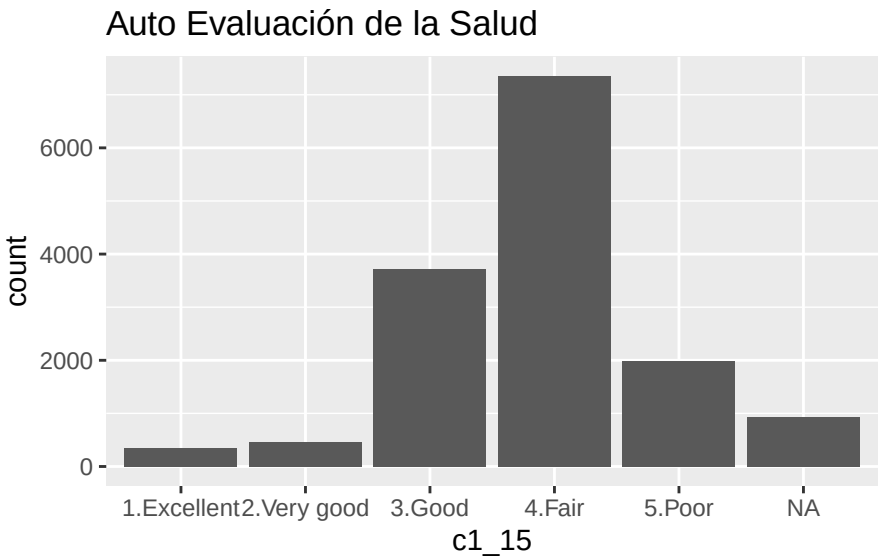
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
aes	13,847	0.674	0.469	0.000	0.000	1.000	1.000
laf	13,810	2.124	2.079	0.000	0.000	4.000	7.000
lavd	9,807	0.249	0.433	0.000	0.000	0.000	1.000

En el primer indicador de salud Auto Evaluación de la Salud AES, resalta que, en 2015, el 67.4% de la muestra reportó tener una salud regular o mala, es decir, 32.6% manifestó tener una buena, muy buena o excelente salud. De las siete Limitaciones a Actividades Físicas (LAF), los entrevistados respondieron en promedio tener 2.12 limitaciones. Destaca que la variable de Limitaciones con Actividades de la Vida Diaria(LAVD) fue la que reportó mayores casos de no respuesta, sin embargo, el 24.9% de los que respondieron aseguraron tener alguna dificultad para hacer alguna de ellas.

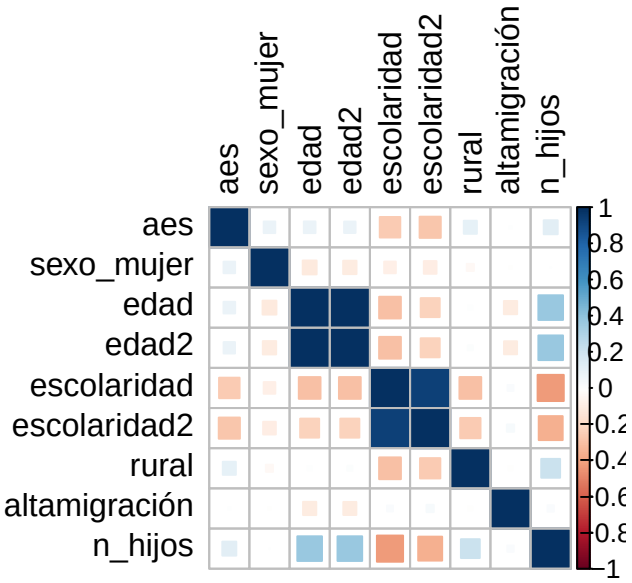
Auto Evaluación de la Salud

Recordemos que la variable AES couando toma el valor de 1 indica peor estado de salud en comparación con cuando toma el valor de 0. Excellent = 0, Very good = 0, Good = 0, Fair = 1, Poor = 1

```
ggplot(data = df)+geom_bar(mapping = aes(x=c1_15))+
  ggtitle("Auto Evaluación de la Salud")
```




```
m <- na.omit(df[c("aes","sexo_mujer","edad","edad2","escolaridad",
                  "escolaridad2","rural","altamigración","n_hijos")])
m <- cor(m)
corrplot(m, method = "square", tl.col = "black")
```



$N = 13,188$ en Corrplot.

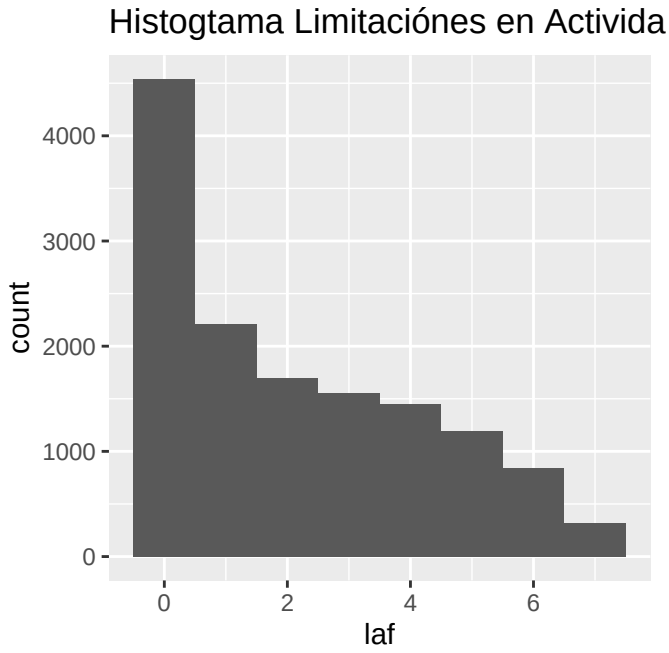
Del la figura 1 resalta la falta de correlación entre la salud y la pertenencia a estados de alta migración a Estados Unidos, así mismo, destaca la correlación entre el número de hijos y las variables socio demográficas de edad y escolaridad.

Limitaciones en Actividades Físicas

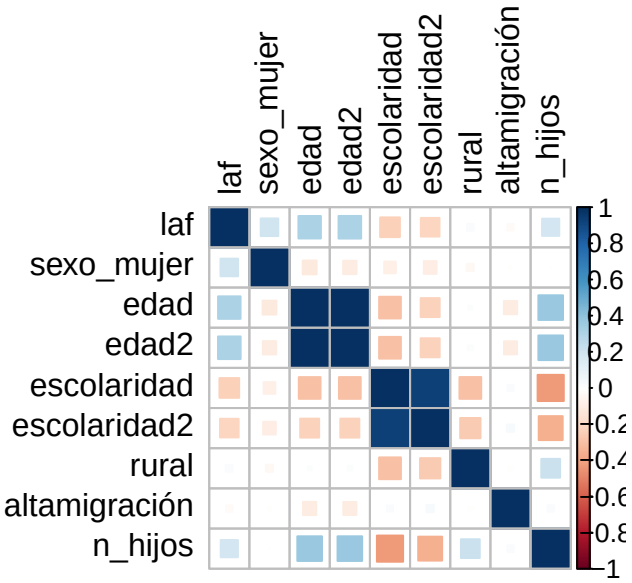
La variable LAF toma el valor máximo de 7 y el valor mínimo de 0 dependiendo de la cantidad de limitaciones a las que tiene en las siguientes actividades:

- Caminar varias cuadras;
- Subir varios pisos de escaleras sin descansar;
- Levantarse de una silla después de haber estado sentado durante un largo tiempo;
- Arrodillarse, agacharse o ponerse en cuclillas;
- Subir o extender los brazos más arriba de los hombros;
- Levantar o transportar objetos que pesan más de 5 kilos, como una bolsa pesada de alimentos;
- Recoger una moneda de 1 peso de la mesa.

```
ggplot(data = df)+geom_histogram(mapping = aes(x=laf),binwidth = 1)+
  ggtitle("Histogtama Limitaciónes en Actividades Físicas")
```




```
m <- na.omit(df[c("laf","sexo_mujer","edad","edad2","escolaridad",
                  "escolaridad2","rural","altamigración","n_hijos")])
m <- cor(m)
corrplot(m, method = "square", tl.col = "black")
```



$N = 13,158$ en Corrplot

Resalta que la variable LAF tiene correlaciones con diversas variables socio demográgicas tengo la expectastiva de que el modelo de regrasión lineal que se proponga será más explicativo para la variable de limitaciones(laf) que para la variable de auto evaluación de la salud (aes).

Limitaciones en Actividades de la Vida Diaria

La variable LAVD toma el valor de 1 si el individuo reporta dificultades para efectuar cualquiera de las siguientes actividades y 0 no tiene limitaciones:

- Caminar de un lado a otro de un cuarto;
- Bañarse;
- Comer, cortar su comida;
- Acostarse y levantarse de la cama y;
- Usar el excusado

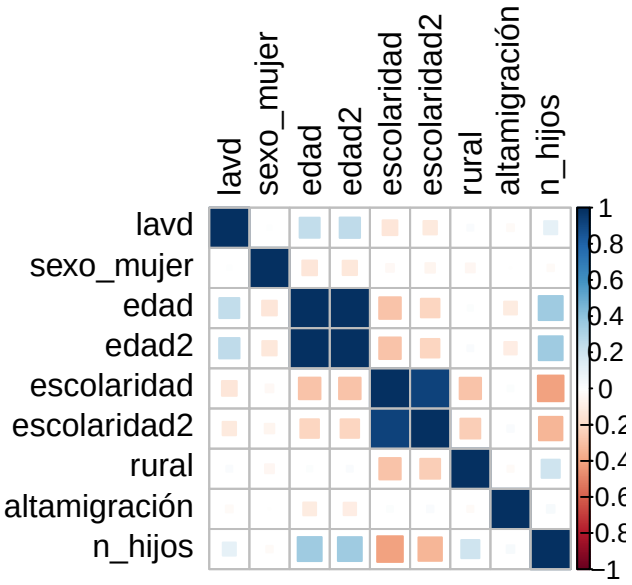
```
tabla<-df_2 %>%
  group_by(lavd) %>%
summarise(
  n=n(),
  Caminar=sum(Caminar,na.rm=T),
  Bañarse=sum(Bañarse,na.rm=T),
  Comer=sum(Comer,na.rm=T),
  Ir_a_la_cama=sum(Ir_a_la_cama,na.rm=T),
  Usar_el_excusado=sum(Usar_el_excusado,na.rm=T)
)
```

Destaca que el 24.9% de la muestra presenta limitaciónes con alguna actividad. Camiar e ir a la cama son los que se encuentran más presentes. En caso de tener limitaciones en cualquiera de als actividades la variable LAVD toma el valor de uno.

Table 3: Construcción de la variable LAVD

Variables	LAVD	Caminar	Bañarse	Comer	Ir a la cama	Usar el excusado
0	7364	8521	8834	9248	8363	8851
1	2443	1286	973	559	1444	956
Promedio	0.249	0.131	0.099	0.057	0.147	0.097
NA	4972	4972	4972	4972	4972	4972

```
m <- na.omit(df[c("lavd","sexo_mujer","edad","edad2","escolaridad",
                  "escolaridad2","rural","altamigración","n_hijos")])
m <- cor(m)
corrplot(m, method = "square", tl.col = "black")
```



$N = 9,449$ en Corrplot.

Salud (Enfermedades)

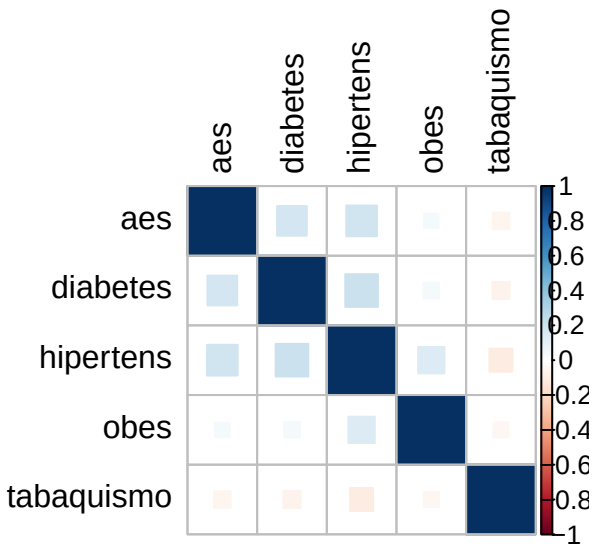
- Diabetes
- Hipertensión
- Obesidad (IMC >= 30)
- Tabaquismo

```
stargazer(df[c("diabetes","hipertens","obes","tabaquismo")])
```

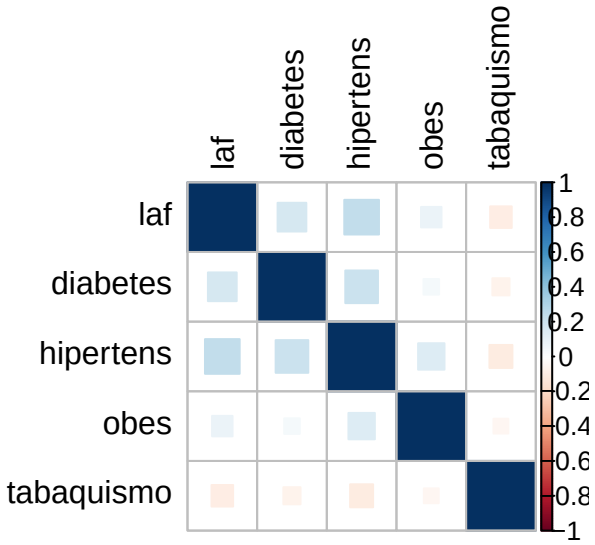
```
corrplot(a, method = "square", tl.col = "black")
```

Table 4: Resumen enfremedades

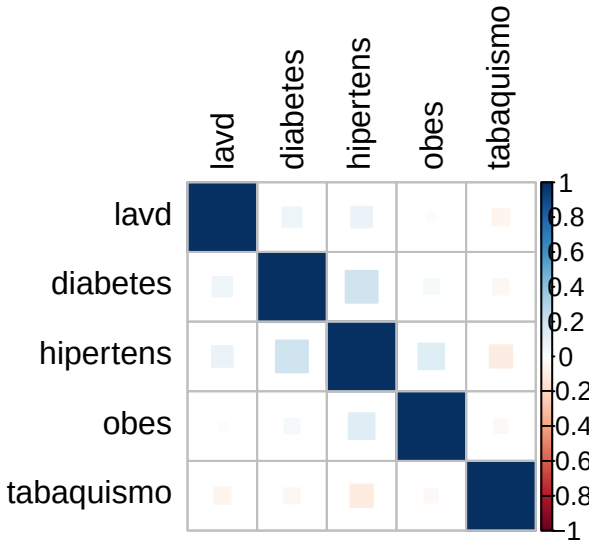
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
diabetes	14,759	0.247	0.431	0.000	0.000	0.000	1.000
hipertens	14,756	0.474	0.499	0.000	0.000	1.000	1.000
obes	14,077	0.244	0.430	0.000	0.000	0.000	1.000
tabaquismo	14,779	0.115	0.320	0	0	0	1




```
corrplot(b, method = "square", tl.col = "black")
```




```
corrplot(c, method = "square", tl.col = "black")
```



En cuanto a las enfermedaes preocupa la alta correlacion entre las variables y destaca el caso curioso del tabaquiso (correlacion nefativa), que se contruye con la pregunta, “Fuma actualmente” entonces es porible que las personas que con mala salud dejan de fumar para evitar agravar dichas condiciones de salud.

Económicas - 25 Variables

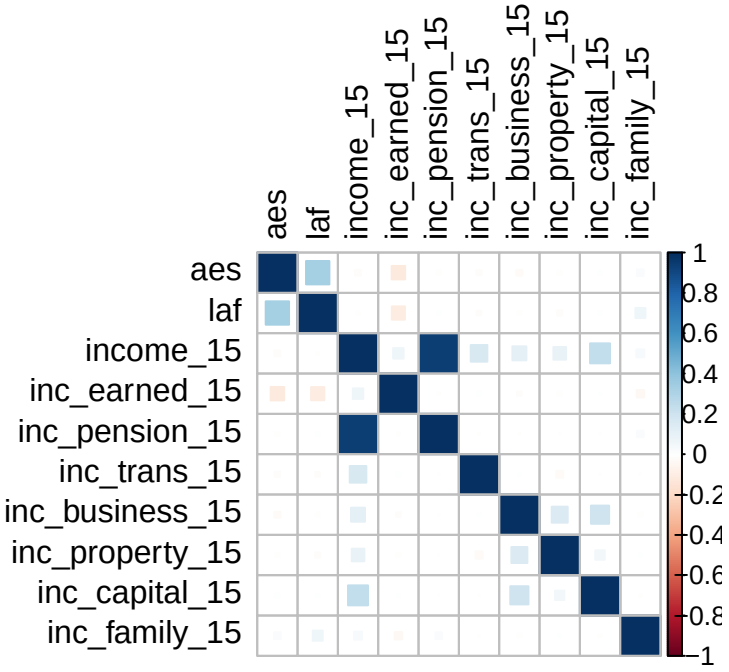
```
stargazer(df[c("income_15","inc_earned_15","inc_pension_15","inc_trans_15",
               "inc_business_15","inc_property_15","inc_capital_15","inc_family_15")])
```

Table 5: Resumen fuentes de ingreso (Pesos mexicanos)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
income_15	14,157	7,182.007	104,287.500	−48,366.670	500.000	5,000.000	9,005,000.000
inc_earned_15	14,745	1,399.114	7,531.631	0.000	0.000	0.000	508,616.900
inc_pension_15	14,745	2,802.862	96,632.370	0.000	0.000	630.000	9,000,000.000
inc_trans_15	14,745	875.007	17,755.900	0.000	0.000	0.000	1,200,000.000
inc_business_15	14,745	846.899	6,365.880	0.000	0.000	0.000	250,000.000
inc_property_15	14,745	296.268	8,365.293	−50,000.000	0.000	0.000	497,500.000
inc_capital_15	14,745	329.642	23,331.310	0.000	0.000	0.000	2,001,353.000
inc_family_15	14,157	528.340	1,840.974	0.000	0.000	500.000	116,360.200

```
d <- na.omit(df[c("aes","laf","income_15","inc_earned_15","inc_pension_15","inc_trans_15",
               "inc_business_15","inc_property_15","inc_capital_15","inc_family_15")])
d <- cor(d)

corrplot(d, method = "square", tl.col = "black")
```



Destaca de la tabla 5 la desigualdad en el ingreso y en el mapa de correlaciones destaca que el ingreso percibido o ganado es el que tiene correlación negativa con la salud.

Gastos en hospitalización y servicios médicos

```
stargazer(df[c("imamd6_15","imamd9_1_15","imamd9_2_15","imamd9_3_15",
               "imamd9_4_15","imamd12a_15")])
```

Table 6: Resumen gastos en salud (Pesos mexicanos)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Total hospitalization costs	14,763	1,301.861	12,628.880	0.000	0.000	0.000	500,000.000
Total curandero homeopath costs	14,763	66.014	656.870	0.000	0.000	0.000	40,000.000
Total dentist costs	14,763	780.452	3,084.678	0.000	0.000	50.000	100,000.000
Total outpatient procedure costs	14,763	207.572	2,623.970	0.000	0.000	0.000	120,000.000
Total medical visits costs	14,763	621.358	3,113.482	0.000	0.000	60.000	100,000.000
Medications costs	14,763	476.233	2,098.566	0.000	0.000	300.000	120,000.000

En la tabla 6 destacan las variables gasto en medicamentos, gasto en visitas medicas y gasto en dentista como potenciales variables explicativas en el modelo.