Taller 4 - Análisis descriptivo para Proyecto Final

Valentina Moreno y Laura Barragan

16/5/2021

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.5 v dplyr 1.0.3  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(haven)  
library(knitr)  
library(naniar)

## Warning: package 'naniar' was built under R version 4.0.5

library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

library(GGally)

## Warning: package 'GGally' was built under R version 4.0.5

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

## Cargar las bases de datos

rm(list = ls())  
Base\_ninos <- read\_sav("Datos/Enti\_ninos.sav")  
Base\_adultos <- read\_sav("Datos/Enti\_adultos.sav")

## Adecuación de la base de datos para el análisis

Base\_ninos <- Base\_ninos %>%  
rename(Ninos\_tiempo = "P6237")%>%  
rename(Nino\_sexo = "P220")%>%  
rename(Nino\_edad = "P6037")%>%  
rename(Nino\_personas\_hogar = "P6008")%>%  
rename(Nino\_educ = "P6170")%>%  
rename(Nino\_factor\_exp = Fex\_c)  
  
Ninos\_interes <- Base\_ninos %>%  
 select(starts\_with("Nino"), Directorio, Secuencia\_p, Orden)  
  
Base\_adultos <- Base\_adultos %>%  
rename(Adult\_disminucion\_ing = "P6047") %>%  
rename(Adult\_factor\_exp = Fex\_c)  
  
Adult\_interes <- Base\_adultos %>%  
 select(starts\_with("Adult"), Directorio, Secuencia\_p, Orden)  
  
Base\_completa <- full\_join(Ninos\_interes, Adult\_interes, by = c("Directorio", "Secuencia\_p", "Orden"))  
  
Base\_completa <- Base\_completa %>%  
 relocate("Directorio", "Secuencia\_p", "Orden","Ninos\_tiempo", "Adult\_disminucion\_ing", "Nino\_personas\_hogar", "Nino\_sexo", "Nino\_edad", "Nino\_educ", "Adult\_factor\_exp", "Nino\_factor\_exp")

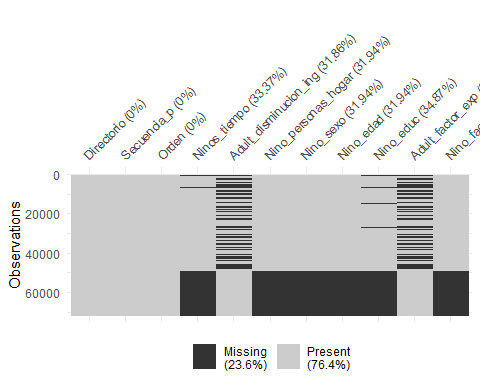
## Manejo de valores faltantes

En cuanto al manejo de datos faltantes para la variable llamada *Adult\_disminucion\_ing* que da cuenta de la disminucion de los ingresos en el hogar durante los ultimos 12 meses, en principio tenia dos categorias de respuesta, 1 para si y 2 para no; esta variable tenia un porcentaje de 31,86% de datos faltantes, sin embargo, luego de el analisis realizado, decidimos reemplazar esos valores faltantes por una tercera categoria de respuesta representada por el numero 0, esta nueva categoria de respuesta representara aquellas familias cuyos ingresos se mantuvieron constantes o aumentaron durante los ultimos 12 meses.

Por otro lado, en cuanto al factor de expansion de la base de datos de los adultos, que debe entenderse como la cantidad de personas en la población que representa una persona en la muestra tenia un porcentaje de datos faltantes del 31,86%. Dado el comportamiento de las observaciones de esta variable, se decidio rellenar esos valores faltantes con valores completos que estuvieran antes o despúes de los faltantes.

Finalmente, todas las variables obtenidas de la base de datos de los niños tenian un porcentaje de valores faltantes entre el 30-35% y segun la grafica pudimos evidenciar que estos valores faltantes aparecian a partir de la observacion 50.000 y teniendo en cuenta que las variables de mayor interes para este estudio son las obtenidas de la base de datos que da informacion de los niños, decidimos eliminar los valores faltantes de todas las variables, teniendo en cuenta que el procentaje de datos que quedaba disponible para realizar los analisis era del 76,4%.

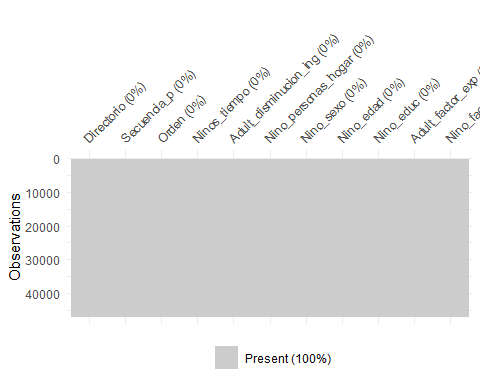
faltantes\_antes <- vis\_miss(Base\_completa)  
faltantes\_antes



ggsave( "Faltantes\_antes.pdf", faltantes\_antes)

## Saving 5 x 4 in image

Base\_completa <- replace\_na(data = Base\_completa, replace = list(Adult\_disminucion\_ing = 0))  
Base\_completa <- fill(Base\_completa, Adult\_factor\_exp)  
Base\_completa <-na.omit(Base\_completa)  
  
faltantes\_final <- vis\_miss(Base\_completa)  
faltantes\_final



ggsave( "Faltantes\_final.pdf", faltantes\_final)

## Saving 5 x 4 in image

## Reconocimiento de las variables como numero o factor en R

class(Base\_completa$Directorio)

## [1] "numeric"

class(Base\_completa$Secuencia\_p)

## [1] "numeric"

class(Base\_completa$Orden)

## [1] "numeric"

class(Base\_completa$Ninos\_tiempo)

## [1] "haven\_labelled" "vctrs\_vctr" "double"

class(Base\_completa$Nino\_sexo)

## [1] "haven\_labelled" "vctrs\_vctr" "double"

class(Base\_completa$Nino\_edad)

## [1] "numeric"

class(Base\_completa$Nino\_personas\_hogar)

## [1] "numeric"

class(Base\_completa$Nino\_educ)

## [1] "haven\_labelled" "vctrs\_vctr" "double"

class(Base\_completa$Nino\_factor\_exp)

## [1] "numeric"

class(Base\_completa$Adult\_factor\_exp)

## [1] "numeric"

class(Base\_completa$Adult\_disminucion\_ing)

## [1] "haven\_labelled" "vctrs\_vctr" "double"

Base\_completa <- Base\_completa %>%   
 mutate(Ninos\_tiempo = as.factor(Ninos\_tiempo)) %>%   
 mutate(Nino\_sexo = as.factor(Nino\_sexo)) %>%   
 mutate(Nino\_edad = as.numeric(Nino\_edad)) %>%   
 mutate(Nino\_personas\_hogar = as.numeric(Nino\_personas\_hogar)) %>%   
 mutate(Nino\_educ = as.factor(Nino\_educ)) %>%   
 mutate(Adult\_disminucion\_ing = as.factor(Adult\_disminucion\_ing))

Base\_completa <- Base\_completa %>%   
 mutate(Mujer = recode(Nino\_sexo, "1" = 0, "2" = 1))%>%   
 mutate(Asistencia = recode(Nino\_educ, "2" = 0, "1" = 1))  
  
class(Base\_completa$Mujer)

## [1] "numeric"

class(Base\_completa$Asistencia)

## [1] "numeric"

Base\_completa <- Base\_completa %>%   
 mutate(Mujer = as.factor(Mujer)) %>%   
 mutate(Asistencia = as.factor(Asistencia))

## Estadísticas descriptivas de las variables de interes

f\_abs\_dep <- table(Base\_completa$Ninos\_tiempo)  
f\_abs\_dep

##   
## 1 2 3 4 5 6 7   
## 1541 90 25411 15878 1847 43 1978

f\_rel\_dep <- prop.table(table(Base\_completa$Ninos\_tiempo))\*100  
f\_rel\_dep

##   
## 1 2 3 4 5 6   
## 3.29357955 0.19235701 54.31093443 33.93605198 3.94759340 0.09190391   
## 7   
## 4.22757972

distribucion\_dep <- cbind(Absoluta = f\_abs\_dep, Relativa = f\_rel\_dep)  
distribucion\_dep

## Absoluta Relativa  
## 1 1541 3.29357955  
## 2 90 0.19235701  
## 3 25411 54.31093443  
## 4 15878 33.93605198  
## 5 1847 3.94759340  
## 6 43 0.09190391  
## 7 1978 4.22757972

f\_abs\_indep <- table(Base\_completa$Adult\_disminucion\_ing)  
f\_abs\_indep

##   
## 0 1 2   
## 22082 7947 16759

f\_rel\_indep <- prop.table(table(Base\_completa$Adult\_disminucion\_ing))\*100  
f\_rel\_indep

##   
## 0 1 2   
## 47.19586 16.98512 35.81901

distribucion\_indep <- cbind(Absoluta = f\_abs\_indep, Relativa = f\_rel\_indep)  
distribucion\_indep

## Absoluta Relativa  
## 0 22082 47.19586  
## 1 7947 16.98512  
## 2 16759 35.81901

f\_absoluta\_con <- table(Base\_completa$Nino\_sexo,  
 Base\_completa$Asistencia)  
f\_absoluta\_con

##   
## 0 1  
## 1 1958 21830  
## 2 1485 21515

f\_relativa\_total <- prop.table(f\_absoluta\_con)  
round(f\_relativa\_total, 4) \* 100

##   
## 0 1  
## 1 4.18 46.66  
## 2 3.17 45.98

stargazer(as.data.frame(Base\_completa), type = "text",   
 out = "Tabla\_1",   
 title = "Tabla 1 - Estadística Descriptiva")

##   
## Tabla 1 - Estadística Descriptiva  
## ==================================================================================  
## Statistic N Mean St. Dev. Min Pctl(25) Pctl(75) Max   
## ----------------------------------------------------------------------------------  
## Directorio 46,788 23,904.510 9,105.695 8,351 15,624.8 31,808.2 39,324   
## Secuencia\_p 46,788 1.018 0.150 1 1 1 4   
## Orden 46,788 1.607 0.884 1 1 2 9   
## Nino\_personas\_hogar 46,788 5.117 2.011 1 4 6 20   
## Nino\_edad 46,788 11.195 3.656 4 8 14 17   
## Adult\_factor\_exp 46,788 697.848 1,061.992 8.588 123.188 920.972 19,961.130  
## Nino\_factor\_exp 46,788 687.119 973.981 8.588 124.182 951.415 19,961.130  
## ----------------------------------------------------------------------------------

#Coeficientes de variacion  
  
coeficiente\_personas <- sd(Base\_completa$Nino\_personas\_hogar, na.rm=TRUE)/mean(Base\_completa$Nino\_personas\_hogar, na.rm=TRUE) \* 100  
coeficiente\_personas

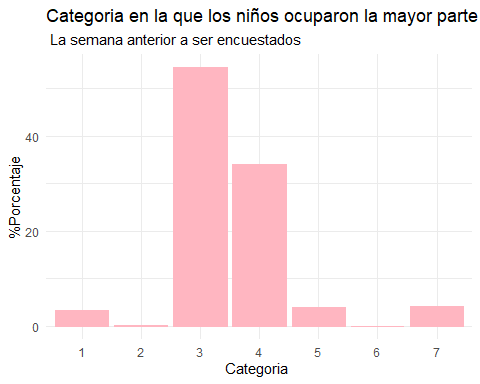
## [1] 39.31017

coeficiente\_edad <- sd(Base\_completa$Nino\_edad, na.rm=TRUE)/mean(Base\_completa$Nino\_edad, na.rm=TRUE) \* 100  
coeficiente\_edad

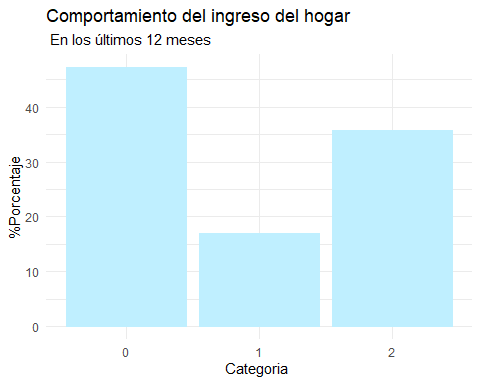
## [1] 32.66027

## Gráﬁcos de barras para la variable dependiente y para la principal variable independiente.

Variable\_depen <- Base\_completa %>% count(Ninos\_tiempo)%>%  
 mutate(prop = n/sum(n)\*100)  
  
ggplot(Variable\_depen, aes(Ninos\_tiempo, prop)) +  
 geom\_bar(stat = "identity", fill = "lightpink", color = c ("lightpink")) +   
 labs(x= "Categoria ", y= "%Porcentaje", title = "Categoria en la que los niños ocuparon la mayor parte de su tiempo ", subtitle = " La semana anterior a ser encuestados") +   
 theme\_minimal()



Variable\_indep <- Base\_completa %>% count(Adult\_disminucion\_ing)%>%  
 mutate(prop1 = n/sum(n)\*100)  
  
ggplot(Variable\_indep, aes(Adult\_disminucion\_ing, prop1)) +  
 geom\_bar(stat = "identity", fill = "lightblue1", color = ("lightblue1")) +   
 labs(x= "Categoria ", y= "%Porcentaje", title = "Comportamiento del ingreso del hogar ", subtitle = " En los últimos 12 meses") +   
 theme\_minimal()

 ## Correlación entre cada pareja de variables cuantitativas y tablas cruzadas para cada par de variables categoricas.

correlaciones\_num <- cor(Base\_completa$Nino\_personas\_hogar, Base\_completa$Nino\_edad)  
correlaciones\_num\*100

## [1] -2.705841

table(Base\_completa$Ninos\_tiempo, Base\_completa$Adult\_disminucion\_ing)

##   
## 0 1 2  
## 1 829 247 465  
## 2 54 13 23  
## 3 13627 4108 7676  
## 4 5686 2990 7202  
## 5 946 257 644  
## 6 21 4 18  
## 7 919 328 731

table(Base\_completa$Ninos\_tiempo, Base\_completa$Nino\_sexo)

##   
## 1 2  
## 1 1163 378  
## 2 69 21  
## 3 12730 12681  
## 4 7948 7930  
## 5 603 1244  
## 6 28 15  
## 7 1247 731

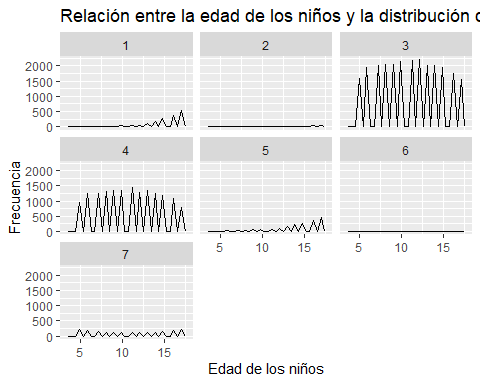
table(Base\_completa$Ninos\_tiempo, Base\_completa$Asistencia)

##   
## 0 1  
## 1 932 609  
## 2 72 18  
## 3 65 25346  
## 4 96 15782  
## 5 1289 558  
## 6 37 6  
## 7 952 1026

## Gráﬁcas de relación entre cada variable independiente y la variable dependiente.

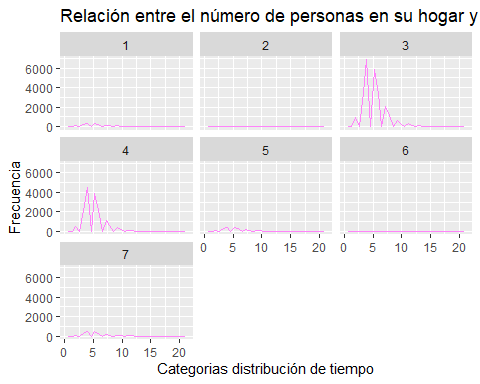
ggplot(data = Base\_completa, mapping = aes(x = Nino\_edad )) +   
 geom\_freqpoly() +   
 facet\_wrap(vars(Ninos\_tiempo))+  
 labs(x= "Edad de los niños", y= "Frecuencia", title = "Relación entre la edad de los niños y la distribución de su tiempo")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

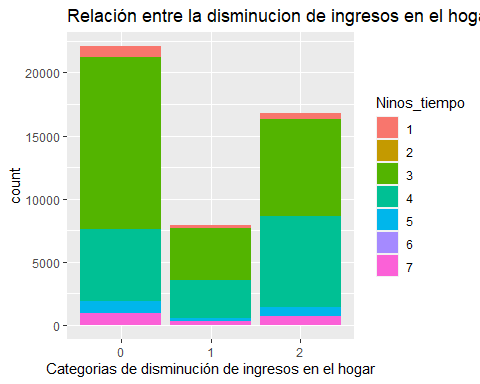


ggplot(data = Base\_completa, mapping = aes(x = Nino\_personas\_hogar)) +   
 geom\_freqpoly(color= "orchid1") +   
 facet\_wrap(vars(Ninos\_tiempo ))+  
 labs(x= "Categorias distribución de tiempo", y= "Frecuencia", title = "Relación entre el número de personas en su hogar y la distribución del tiempo de los niños ")

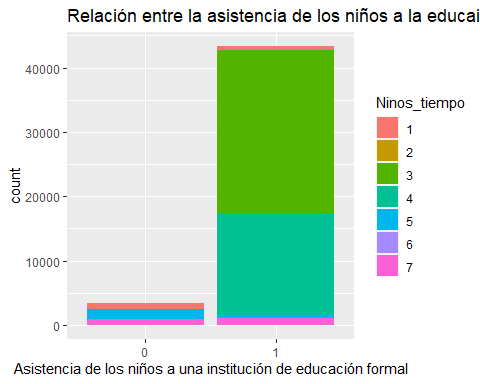
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



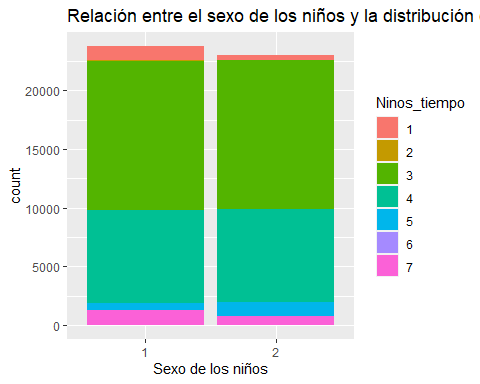
ggplot(data = Base\_completa,   
 mapping = aes(x = Adult\_disminucion\_ing ,  
 fill = Ninos\_tiempo )) +  
 geom\_bar()+  
 labs(x = "Categorias de disminución de ingresos en el hogar", title = "Relación entre la disminucion de ingresos en el hogar y la distribución del tiempo de los niños")



ggplot(data = Base\_completa,   
 mapping = aes(x = Asistencia ,  
 fill = Ninos\_tiempo )) +  
 geom\_bar()+  
 labs(x = "Asistencia de los niños a una institución de educación formal", title = "Relación entre la asistencia de los niños a la educaión formal y la distribución de su tiempo")



ggplot(data = Base\_completa,   
 mapping = aes(x = Nino\_sexo ,  
 fill = Ninos\_tiempo )) +  
 geom\_bar()+  
 labs(x = "Sexo de los niños", title = "Relación entre el sexo de los niños y la distribución de su tiempo")



## Conclusiones sobre la distribución de las variables de interés y sobre la relación entre estas variables

En primer lugar, es importante analizar el comportamiento de la variable dependiente, que corresponde al uso del tiempo de los niños; segun el diagrama de barras analizado, pudimos evidenciar que la mayoria de los niños ocupan su tiempo estudiando o en vacaciones escolares.

La variable independiente de mayor interes para esta investigación el nivel de ingresos de los hogares, utilizando como proxy la variable *Adult\_disminucion\_ing* que nos permite conocer si existio una disminución de los ingresos del hogar en los últimos doce meses resulto evidente que en mas del 40% de los hogares los ingresos se mantuvieron constantes o aumentaron y alrededor del 35% no tuvieron una disminucion de sus ingresos durante el ultimo año.

Por otra parte,el coeficiente de correlacion entre las unicas variables de control numericas, que son la edad de los niños y el numero de personas en el hogar fue de -2,7, lo cual demuestra una relacion inversa entre estas variables, es decir, a medida que la edad del niño aumenta, el numero de personas con las que vive en su hogar tiende a disminuir.

En cuanto a la relacion de la variable dependiente con la edad de los niños, es posible conluir que estudiar y estar en vacaciones presentan la mayor variabilidad a lo largo de todas las edades; por el contrario, en las actividades de buscar trabajo y estar incapacitado permanentemente para trabajar la frecuencia es practicamente nula. Además, los niños que se encontraban trabajando o realizando actividades en el hogar lo hacen a partir de los 10 años, en edades anteriores no se evidencia una participacion relevante en estas labores. Finalmente, en la categoria de otras actividades se observa una variación constante dentro de un rango pequeño a lo largo de todas las edades.

La relacion entre la distribucion del timpo de los niños y el numero de personas en su hogar nos muestra que los niños que gastan la mayor parte de su tiempo trabajando, realizando oficios en el hogar o haciendo otras actividades presentan una leve variacion en cuanto al numero de personas con las que viven en sus primeros 10 años de vida, de ahi hasta los 17 años, que es el periodo que se tiene en cuenta en la encuesta, la variacion es practicamente nula. Adicionalmente, los niños que dedican la mayor parte de su tiempo a estudiar y estar en vacaciones escolares tambien presentan una alta variacion durante sus primeros 10 años respecto al numero de personas con las que viven.

La relacion entre la disminucion de ingresos en el hogar y la distribución del tiempo de los niños muestra que para toda la muestra de hogares, independientemente de si se dio o no una disminucion de ingresos durante los ultimos 12 meses, los niños dedican la mayor parte de su tiempo a estudiar o a estar en vacaciones estudiantiles.

Por ultimo, con la relacion entre el sexo y la distribución del tiempo de los niños es posible deducir que los niños y niñas dedican la mayor parte de su tiempo a estudiar o a estar en vacaciones estudiantiles.