Limpieza de datos para ENAHO

Dereck Amesquita - DAEC Consultoría

Julio - 2022

El fin de este documento es proporcionar una forma amigable de procesar los datos de la Encuesta Nacional de Hogares (Enaho).

Librerías necesarias :

```
library(dplyr)
library(readxl)
library(stargazer)
library(ggplot2)
```

Sobre la base de datos

ENAHO maneja actualmente 8 módulos (http://iinei.inei.gob.pe/microdatos/), nos concentraremos en 3 de ellos, "Educación", "Salud" y "Empleo e Ingresos". Desde el portal de microdatos podemos acceder a su ficha técnica y un archivo CSV el cual previamente ya ha sido descargado y alojado en mi cuenta Github (Esto no es necesario, se subió a GitHub con el fin de que cualquier persona pueda ejecutarlo).

Aclaraciones

R es un software de código abierto, al mismo tiempo este RMarkdown puede ser usado de forma libre.

Resumen

Se iniciaría con el cargado de los archivos necesarios para ser cruzados entre sí, generando una sola base de datos, se procederá a limpiar los valores ausentes y filtrar la data. Posteriormente, se le dará la estructura necesaria con la ayuda de la ficha técnica de la encuesta. Finalmente, se crearán unos gráficos simples juntos a modelos de regresión lineal.

Descarga de Data

```
edu <-"https://raw.githubusercontent.com/dereckamesquita/Learning-Before-Estadistica/main/Enaho/Data_Ena
sal<-"https://raw.githubusercontent.com/dereckamesquita/Learning-Before-Estadistica/main/Enaho/Data_Enal
iyl<-"https://raw.githubusercontent.com/dereckamesquita/Learning-Before-Estadistica/main/Enaho/Data_Enal
dedu <- read.csv(edu)
dsal<- read.csv(sal)
diyl<- read.csv(iyl)
```

Unión de las bases de datos

Especificamos que la unión se hace en la columna llamada "MES" la cual contiene un id único. También eliminamos los valores ausentes.

```
datafull<-full_join(dedu,diyl,by="MES")
data<-na.omit(datafull)
```

En este caso, con fines específicos limpiamos la data de acuerdo a lo que necesitamos. Asignamos valores NA y luego omitimos todos los NA (incluyendo los que vienen por defecto en la base)

```
datafull1<-full_join(data,dsal,by="MES")
data1<-na.omit(datafull1)
data1$P524A1[data1$P524A1=="999999"]=NA
data1$P301A[data1$P301A=="Básica especial"]=NA
data1 <- na.omit(data1)
```

Exportación data

Este cruce de datos, finalmente será guardado.

```
write.csv(data1, file="data.csv")
```

Importación de la nueva data

El fin de hacer esto es para consolidar un único CSV con los cruces correspondientes, una vez hecho esto se podrá comenzar a trabajar desde este punto.

```
data1<-read.csv(file ="data.csv", header=T )
```

Procesando la data

El proceso de la data requiere entender los items de cada variable los cuales se encuentra en la ficha técnica que se encuentra en microdatos. Por ejemplo, la P207 recoge la información del sexo donde 1 es Hombre y 2 es Mujer.

Asignando niveles a las variables

Usaremos un bucle for, que recorra los números 1 y 2 para cambiar los valores por un string denominado según el sexo. Ahora le daremos los niveles necesarios usando "factor". Usamos "model.matrix" para tener el formato necesario que requieren las regresiones categóricas (opcional).

```
P207 <- c("Hombre","Mujer")
#Forma 1
for (a in (1:2)){
  data1$P207[data1$P207==a]=P207[a]
}

data1$P207<- factor(data1$P207,levels =P207)
modelP207<- model.matrix(data1$P524A1~data1$P207+data1$P513T)
```

De esto modo ahora hemos modificado la data y P207 ahora tiene la estructura de variable categórica nominal, es decir que no tiene un orden o importancia y que es mutuamente excluyente de otra.

Con la pregunta P300A, sucede lo mismo, la ficha técnica nos informa que 1 significa Quechua, 2 significa Aimara, 3 significa otra lengua nativa, 4 significa Castellano y así sucesivamente.

```
P300A <- c("Quechua","Aimara","Otra lengua nativa","Castellano",
           "Portugués","Otra lengua extranjera","No escucha/no habla",
           "Lengua de señas peruanas","Ashaninka","Awajún",
           "Shipibo","Shawi","Matsigenka","Achuar")

for (a in (1:15)){
  data1$P300A[data1$P300A==a]=P300A[a]
}
data1$P300A<- factor(data1$P300A,levels =P300A)
modelP300A<- model.matrix(data1$P524A1~data1$P207+data1$P300A)
```

Respecto a la pregunta que recoge el nivel de educación sucede lo mismo.

```
P301A <- c("Sin nivel","Educación inicial","Primaria incompleta",
           "Primaria completa","Secundaria incompleta","Secundaria completa",
           "Superior no universitaria Incompleta","Superior no universitaria completa",
           "Superior universitaria completa","Maestria/Doctorado","Básica especial")

#Forma 1
for (a in (1:12)){
  data1$P301A[data1$P301A==a]=P301A[a]
}

#Limpieza de Educacion Basica y sin Nivel
data1$P301A[data1$P301A=="Básica especial"]=NA
data1$P301A[data1$P301A=="Sin nivel"]=NA
data1 <- na.omit(data1)

#Niveles
data1$P301A<- factor(data1$P301A,levels =P301A)
modelP301A<- model.matrix(data1$P524A1~data1$P301A+data1$P207)

#head(modelP301A[,-1])
```

En el caso de salud, tomamos las preguntas que buscan conocer las limitaciones en 6 campos visual, habla, olfato etc. Donde 1 afirma que posee dicha limitación. En este caso lo que haremos será sumar las 6 variables y dejar su resultado en una nueva. Ahora le daremos los niveles necesarios con factor.

```
#Cambiamos todos los 2 por el numero 0.
data1$P401H1[data1$P401H1==2]=0
data1$P401H2[data1$P401H2==2]=0
data1$P401H3[data1$P401H3==2]=0
data1$P401H4[data1$P401H4==2]=0
data1$P401H5[data1$P401H5==2]=0
data1$P401H6[data1$P401H6==2]=0
#Creamos una columna con el total de los que tienen limitaciones

data1$P401HT <- data1$P401H1+data1$P401H2+data1$P401H3+data1$P401H4+data1$P401H5+data1$P401H6

data1$P401HT <- data1$P401HT+1

#Categorizamos

P401HT <- c("Sano","Una_limitacion","Dos_limitacion",
            "Tres_limitacion","Cuatro_limitacion")

for (a in (1:5)){
  data1$P401HT[data1$P401HT==a]=P401HT [a]
}

data1$P401HT <- factor(data1$P401HT ,levels =P401HT )
```

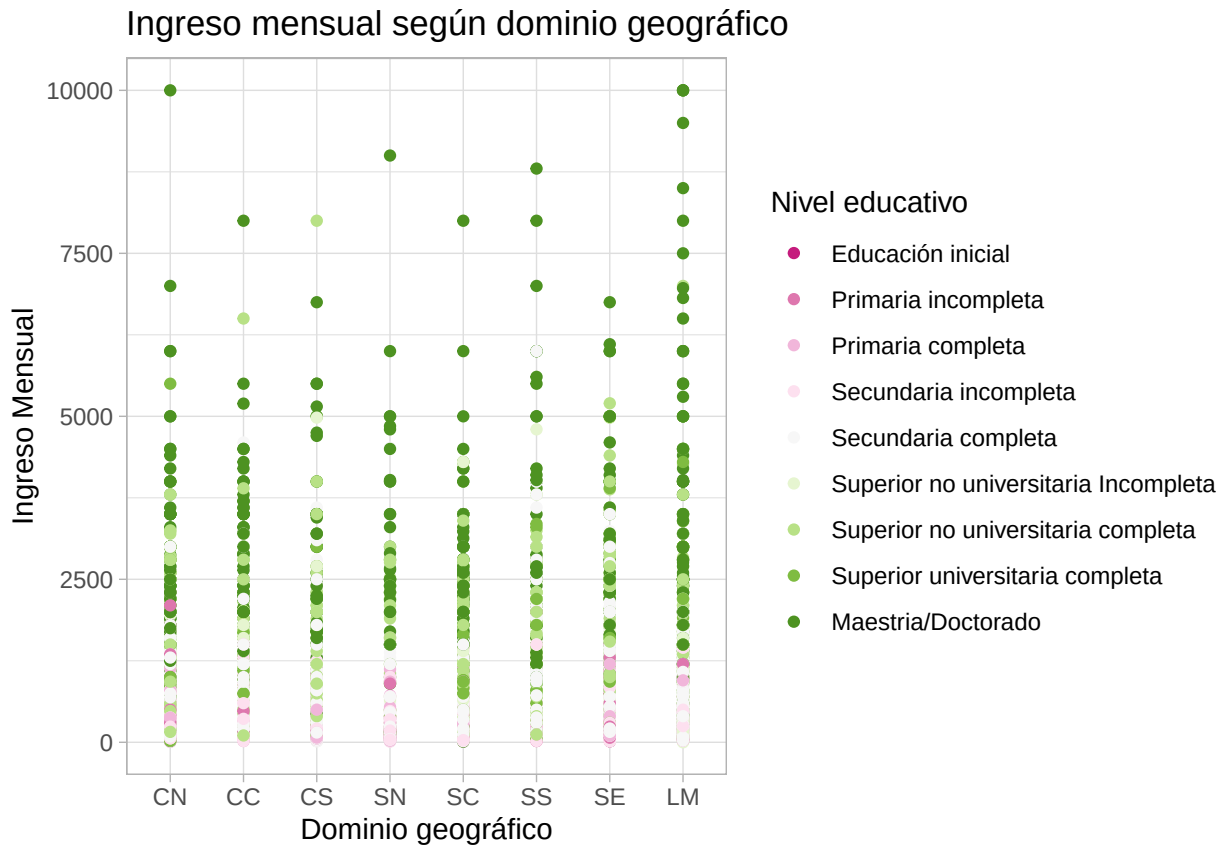
Para trabajar el dominio geográfico, repetimos el mismo proceso ayudándonos de la ficha técnica. Para no extender los nombres he nombrado cada dominio con las primeras letras. Como: Costa Norta (CN), Lima Metropolitana (LM), etc.

```
#Renombrando DOMINIO
names(data1)[names(data1) == 'DOMINIO.x'] <- 'DOMINIO'
data1$DOMINIO <- as.numeric(data1$DOMINIO)
Dominios1 <- c("CostaNorte", "CostaCentro", "CostaSur","SierraNorte","SierraCentro","SierraSur","Selva"

Dominios <- c("CN", "CC", "CS","SN","SC","SS","SE", "LM")
#Cambiamos numeros
for (a in (1:8)){
  data1$DOMINIO[data1$DOMINIO==a]=Dominios[a]
}

data1$DOMINIO<- factor(data1$DOMINIO,levels =Dominios)
```

Primeros gráficos

Creamos un gráfico a partir de la estructura de datos realizada.

```
data1$P524A1=as.numeric(data1$P524A1)
ggplot(data1, aes(x=DOMINIO, y=P524A1, color=P301A )) +
  geom_point() + theme_light()  + labs(x="Dominio geográfico", y="Ingreso Mensual", colour= "Nivel educa
```




```
mod1 <- lm(P524A1 ~ DOMINIO ,data=data1)
```

Del mismo modo podemos crear nuestra primera regresión. Utilizamos Stargazer para darle formato a los resultados.

```
stargazer(mod1, type="text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                P524A1
## ------------------------------------------------
## DOMINIOCC                     -69.745
##                               (59.351)
##
## DOMINIOCS                    285.008***
##                               (71.041)
##
## DOMINIOSN                      84.846
##                               (89.518)
##
## DOMINIOSC                      25.373
##                               (60.715)
##
```

```
## DOMINIOSS                     137.727**
##                               (61.706)
##
## DOMINIOSE                      -67.741
##                               (52.673)
##
## DOMINIOLM                    325.198***
##                               (57.140)
##
## Constant                     955.407***
##                               (37.855)
##
## ------------------------------------------------
## Observations                    4,975
## R2                              0.015
## Adjusted R2                     0.014
## Residual Std. Error    1,150.077 (df = 4967)
## F Statistic          11.154*** (df = 7; 4967)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

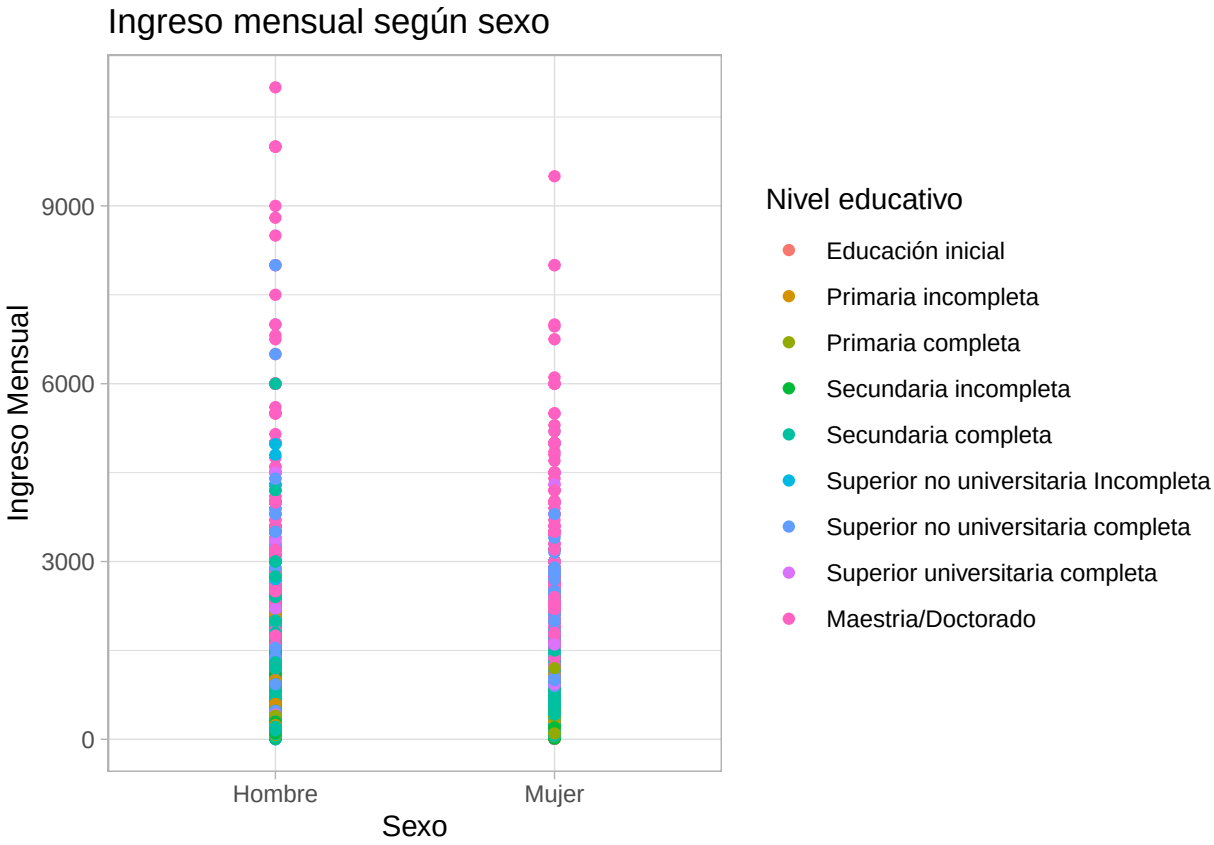
Tambien podemos platear otras relaciónes, en este caso no utilizaremos Stargazer para notar la diferencia.

```
mod <- lm(P524A1 ~ P301A
          , data=data1)

summary(mod)
```

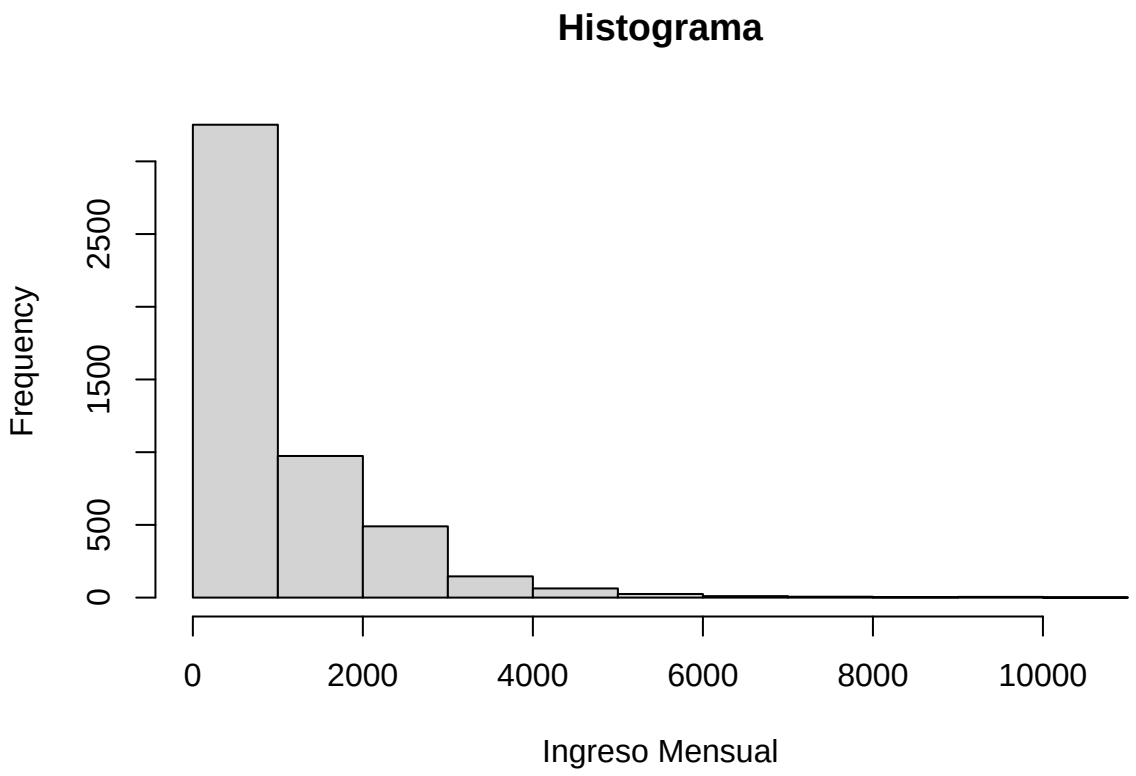
```
##
## Call:
## lm(formula = P524A1 ~ P301A, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2462.6  -497.3  -189.5   327.4  8527.4
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                260.00     534.46   0.486   0.6267
## P301APrimaria incompleta                    90.58     536.98   0.169   0.8661
## P301APrimaria completa                     170.95     536.69   0.319   0.7501
## P301ASecundaria incompleta                 129.50     535.60   0.242   0.8090
## P301ASecundaria completa                   460.02     534.98   0.860   0.3899
## P301ASuperior no universitaria Incompleta  537.25     537.30   1.000   0.3174
## P301ASuperior no universitaria completa   1255.17     535.58   2.344   0.0191
## P301ASuperior universitaria completa       740.88     536.96   1.380   0.1677
## P301AMaestria/Doctorado                   2212.57     535.55   4.131 3.67e-05
##
## (Intercept)
## P301APrimaria incompleta
## P301APrimaria completa
## P301ASecundaria incompleta
## P301ASecundaria completa
```

```
## P301ASuperior no universitaria Incompleta
## P301ASuperior no universitaria completa   *
## P301ASuperior universitaria completa
## P301AMaestria/Doctorado                   ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 925.7 on 4966 degrees of freedom
## Multiple R-squared:  0.3623, Adjusted R-squared:  0.3612
## F-statistic: 352.6 on 8 and 4966 DF,  p-value: < 2.2e-16
```

```
ggplot(data1, aes(x=P207, y=P524A1, color=P301A)) +
  geom_point() + theme_light()  + labs(x="Sexo", y="Ingreso Mensual", colour= "Nivel educativo", title =
```



```
hist(data1$P524A1, main="Histograma", xlab  = "Ingreso Mensual")
```

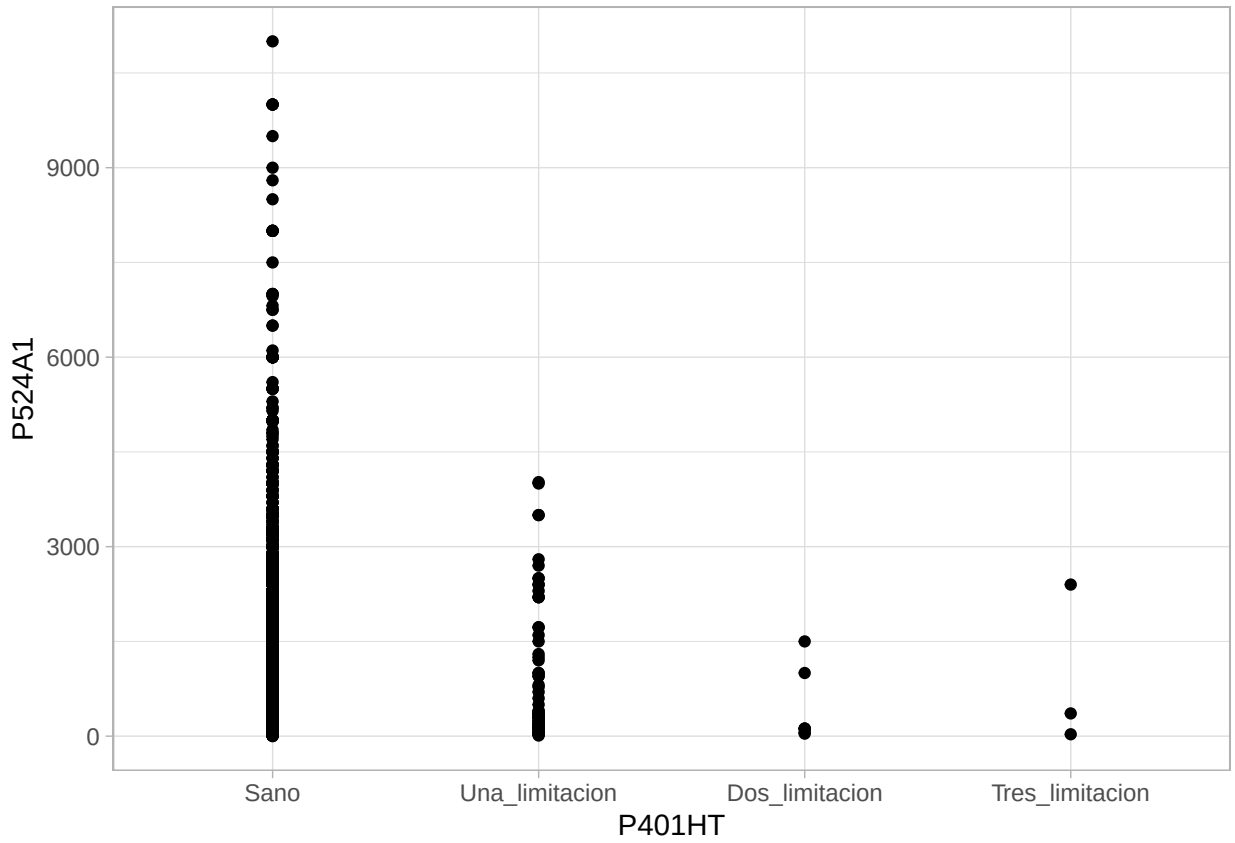



```
mod <- lm(P524A1 ~ P301A + P207+ P401H1+ P401H2 + P401H3 + P401H4 + P401H5 + P401H6
          , data=data1)
summary(mod)
```

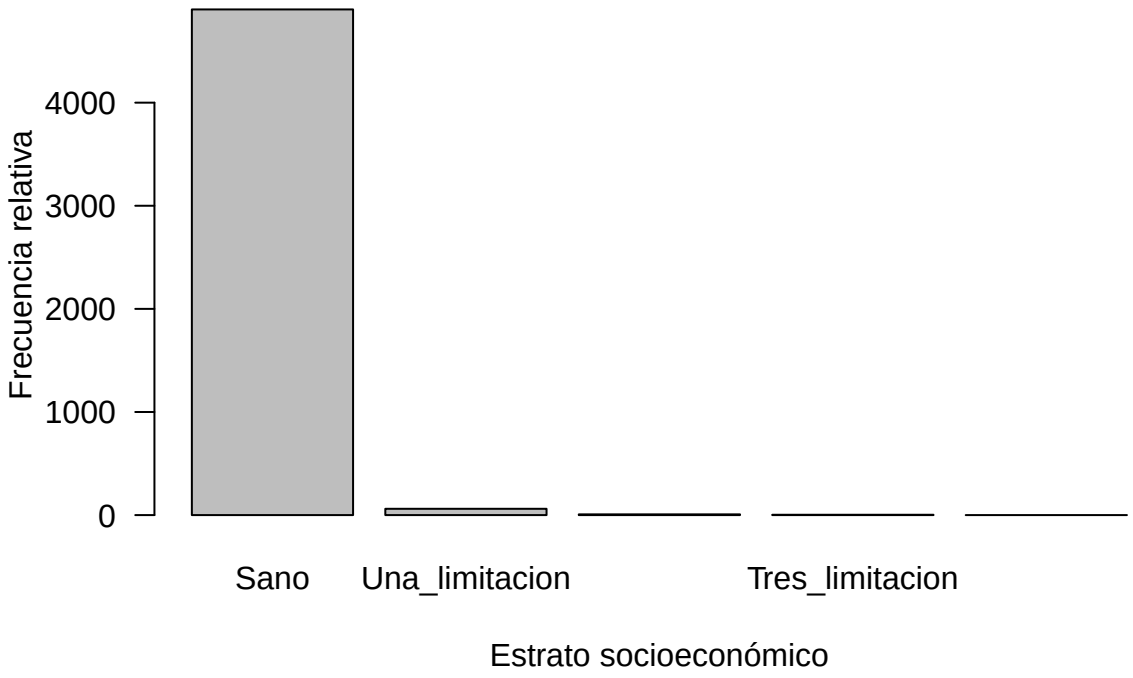
```
##
## Call:
## lm(formula = P524A1 ~ P301A + P207 + P401H1 + P401H2 + P401H3 +
##     P401H4 + P401H5 + P401H6, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2552.5  -477.9  -175.8   348.8  8427.5
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                391.11     532.16   0.735   0.4624
## P301APrimaria incompleta                    40.24     534.41   0.075   0.9400
## P301APrimaria completa                     107.87     534.17   0.202   0.8400
## P301ASecundaria incompleta                  64.72     533.07   0.121   0.9034
## P301ASecundaria completa                   386.80     532.46   0.726   0.4676
## P301ASuperior no universitaria Incompleta  480.63     534.74   0.899   0.3688
## P301ASuperior no universitaria completa   1209.25     533.00   2.269   0.0233
## P301ASuperior universitaria completa       689.99     534.39   1.291   0.1967
```

```
## P301AMaestria/Doctorado                    2181.36     532.96   4.093 4.33e-05
## P207Mujer                                  -196.66      27.32  -7.198 7.01e-13
## P401H1                                       16.95     154.13   0.110   0.9124
## P401H2                                      343.41     250.74   1.370   0.1709
## P401H3                                     -214.84     392.46  -0.547   0.5841
## P401H4                                      -96.91     224.67  -0.431   0.6662
## P401H5                                     -513.85     464.45  -1.106   0.2686
## P401H6                                      271.55     522.16   0.520   0.6031
##
## (Intercept)
## P301APrimaria incompleta
## P301APrimaria completa
## P301ASecundaria incompleta
## P301ASecundaria completa
## P301ASuperior no universitaria Incompleta
## P301ASuperior no universitaria completa   *
## P301ASuperior universitaria completa
## P301AMaestria/Doctorado                  ***
## P207Mujer                                ***
## P401H1
## P401H2
## P401H3
## P401H4
## P401H5
## P401H6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 921.2 on 4959 degrees of freedom
## Multiple R-squared:  0.3694, Adjusted R-squared:  0.3675
## F-statistic: 193.6 on 15 and 4959 DF,  p-value: < 2.2e-16
```

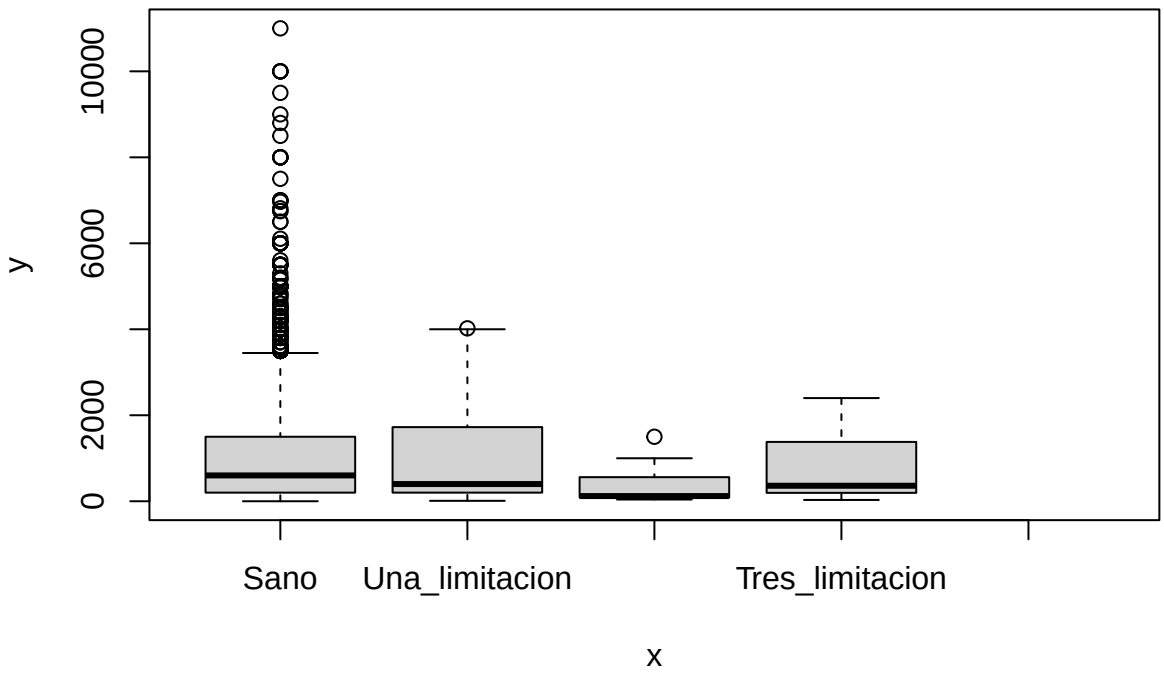
```
ggplot(data1, aes(x=P401HT, y=P524A1)) +
  geom_point() + theme_light()
```




```
tablasalud <- table(data1$P401HT)
tablasalud_ <- prop.table(tablasalud)
barplot(tablasalud, xlab='Estrato socioeconómico',
        ylab='Frecuencia relativa', las=1)
```




```
plot(data1$P401HT,data1$P524A1)
```




```
log_sal <- log(data1$P524A1)
mod <- lm(log(log_sal) ~ data1$P401HT)
```

```
## Warning in log(log_sal): NaNs produced
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = log(log_sal) ~ data1$P401HT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33773 -0.14622  0.04221  0.17607  0.41701
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.813610   0.003166 572.816   <2e-16 ***
## data1$P401HTUna_limitacion   -0.020377   0.028561  -0.713   0.4756
## data1$P401HTDos_limitacion   -0.195042   0.083853  -2.326   0.0201 *
## data1$P401HTTres_limitacion -0.130713   0.128036  -1.021   0.3073
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2217 on 4970 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.001394,   Adjusted R-squared:  0.0007915
## F-statistic: 2.313 on 3 and 4970 DF,  p-value: 0.07399
```

Division por dominio

Dominio Geográfico 1. Costa Norte 2. Costa Centro 3. Costa Sur 4. Sierra Norte 5. Sierra Centro 6. Sierra Sur 7. Selva 8. Lima Metropolitana

```
dataCostaNorte <- filter(data1, DOMINIO=="CN")
dataCostaCentro <- filter(data1, DOMINIO=="CC")
dataCostaSur <- filter(data1, DOMINIO=="CS")
dataSierraNorte <- filter(data1, DOMINIO=="SN")
dataSierraCentro <- filter(data1, DOMINIO=="SC")
dataSierraSur <- filter(data1, DOMINIO=="SS")
dataSelva <- filter(data1, DOMINIO=="SE")
dataLimaMetropolitana <- filter(data1, DOMINIO=="LM")
```

```
model_CN<-lm(P524A1 ~ P301A, data=dataCostaNorte)
model_CC<-lm(P524A1 ~ P301A, data=dataCostaCentro)
model_LM<-lm(P524A1 ~ P301A, data=dataLimaMetropolitana)

beta_CN <- model_CN$coefficients[2:8]
beta_CC <- model_CC$coefficients[2:8]
beta_LM <- model_LM$coefficients[2:8]

unidos<-data.frame(beta_CN,beta_CC,beta_LM)


summary(model_CN)
```

```
##
## Call:
## lm(formula = P524A1 ~ P301A, data = dataCostaNorte)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2310.2  -461.6  -180.1   316.4  7669.8
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                454.44     104.15   4.363 1.43e-05
## P301APrimaria completa                      25.69     137.89   0.186  0.85223
## P301ASecundaria incompleta                 -65.21     128.41  -0.508  0.61171
## P301ASecundaria completa                   217.19     115.29   1.884  0.05989
## P301ASuperior no universitaria Incompleta  318.04     151.74   2.096  0.03636
## P301ASuperior no universitaria completa    978.94     126.45   7.742 2.59e-14
## P301ASuperior universitaria completa       436.02     157.26   2.773  0.00567
## P301AMaestria/Doctorado                   1875.71     128.06  14.647  < 2e-16
##
## (Intercept)                               ***
```

```
## P301APrimaria completa
## P301ASecundaria incompleta
## P301ASecundaria completa                   .
## P301ASuperior no universitaria Incompleta *
## P301ASuperior no universitaria completa   ***
## P301ASuperior universitaria completa      **
## P301AMaestria/Doctorado                   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 833.2 on 915 degrees of freedom
## Multiple R-squared:  0.3661, Adjusted R-squared:  0.3612
## F-statistic: 75.48 on 7 and 915 DF,  p-value: < 2.2e-16
```

for(i in 1:ncol(portfolios)){

linear regression mod <- lm(portfolios[,i] ~ data$Mkt.RF)

summary mod.s <- summary(mod)

store residuals eps[,i] <- residuals(mod)

extract coefficients alpha <- mod.s$coefficients[1,'Estimate']beta <- mod.s$coefficients[2,'Estimate']

extract standard errors of the estimates sd.alpha <- mod.s$coefficients[1,'Std.Error']sd.beta <- mod.s$coefficients[2,'Std. Error']

compute the average excess return excess <- mean(portfolios[,i])

store everything into the capm dataframe row <- c(excess, alpha, sd.alpha, beta, sd.beta) capm <- rbind(capm, row) }