Documento2do Corte

Raul Pinilla

2023-10-10

Summary of the dataset necessary to understand the exercise

People with diabetes no diabetes=0; prediabetes=1; diabetes= 2

People with high Blood Pressure (BP) No high BP= 0; High BP = 1

People with high cholesterol No high cholesterol= 0; high cholesterol = 1

Cholesterol control in the last 5 years (CholCheck) No CholCheck= 0; Yes cholesterol control in 5 years = 1

BMI: body mass index

Smoker: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] No= 0; Yes= 1

Stroke: (Were you ever told) that you had a stroke. No= 0; Yes= 1

HeartDiseaseorAttack: Coronary heart disease (CHD) or myocardial infarction (MI) No=0; Yes= 1

PhysActivity: physical activity in the last 30 days - not including work No=0; Yes= 1

Fruits: Consume fruit 1 or more times a day No=0; Yes= 1

Vegetables: Eat Vegetables 1 or more times a day No=0; Yes= 1

HvyAlcoholConsump: (adult men >=14 drinks per week and adult women >=7 drinks per week) No=0; Yes= 1

AnyHealthcare: Have any type of health care coverage, including health insurance, prepaid plans such as HMOs,etc. No=0; Yes= 1

NoDocbcCost: Was there a time in the last 12 months when you needed to see a doctor but couldn't because cost? No=0; Yes= 1

GenHlth: Would you say that in general your health is: scale 1-5 1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 =poor

MentHlth: days of poor mental health scale 1-30 days

PhysHlth: days of illness or physical injury in the last 30 days scale 1-30

DiffWalk: Do you have serious difficulties walking or climbing stairs? No=0; Yes= 1

Sex: Female=0; Male=1

Age: 13-level age category (_AGEG5YR see codebook) 18-24=1; 60-64=9; 80 or more=13

Education: Educational level (EDUCA see code book) Scale 1-6 1 = Never attended school or only kindergarten 2 = elementary etc.

Income: Income scale (INCOME2 see codebook) Less than $10,000= scale 1-8 1; Less than $35,000= 5; $75,000 or more= 8

Summary of the variables in the dataset 0 = no diabetes 1 = prediabetes 2 = diabetes

Data exploration and data wrangling

Initially, the database "diabetes_012_health_indicators_BRFSS2015.csv" provided by the teacher is loaded, for this the following function was used:

{r cars, include=FALSE} data <- read.delim("clipboard") data

Where:

"data" loads specified data sets or lists available data sets.

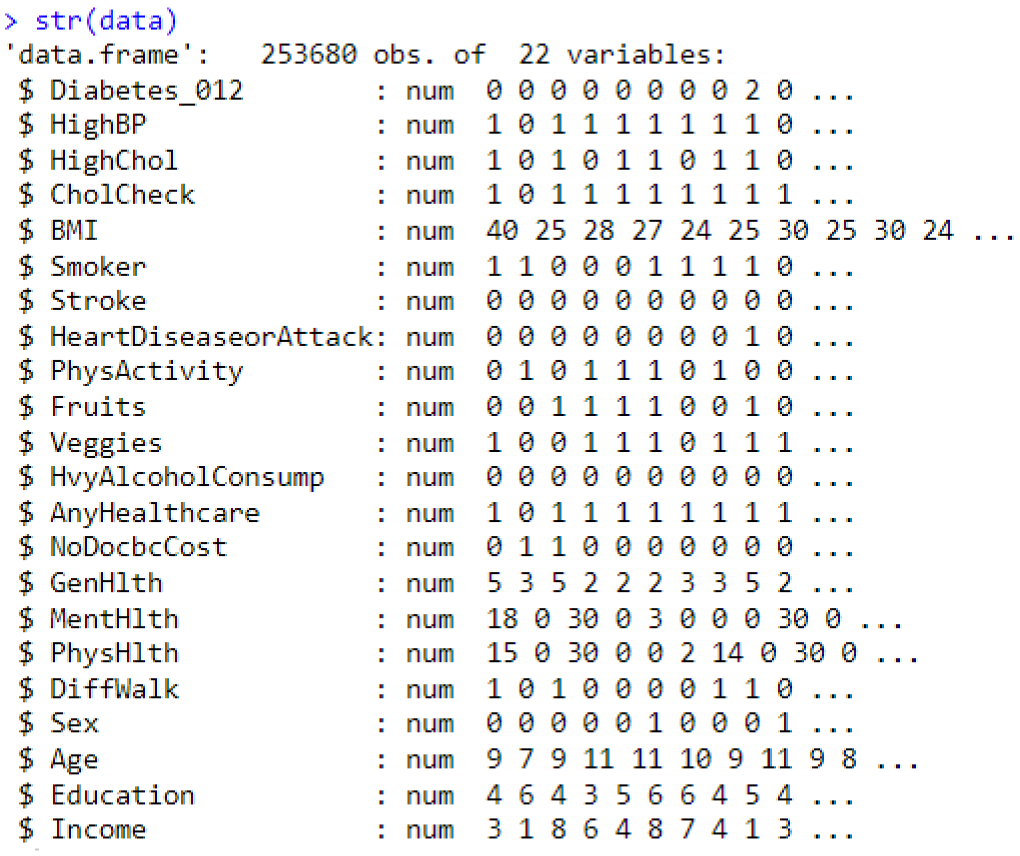
"read.delim" = Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

I used "clipboard" to paste the data from diabetes_012_health_indicators_BRFSS2015.csv since I had it in an excel file and I clicked run on the code to save it.

"Data" We use it to review the data in general, where it tells us that we have 22 variables and 253680 observations.

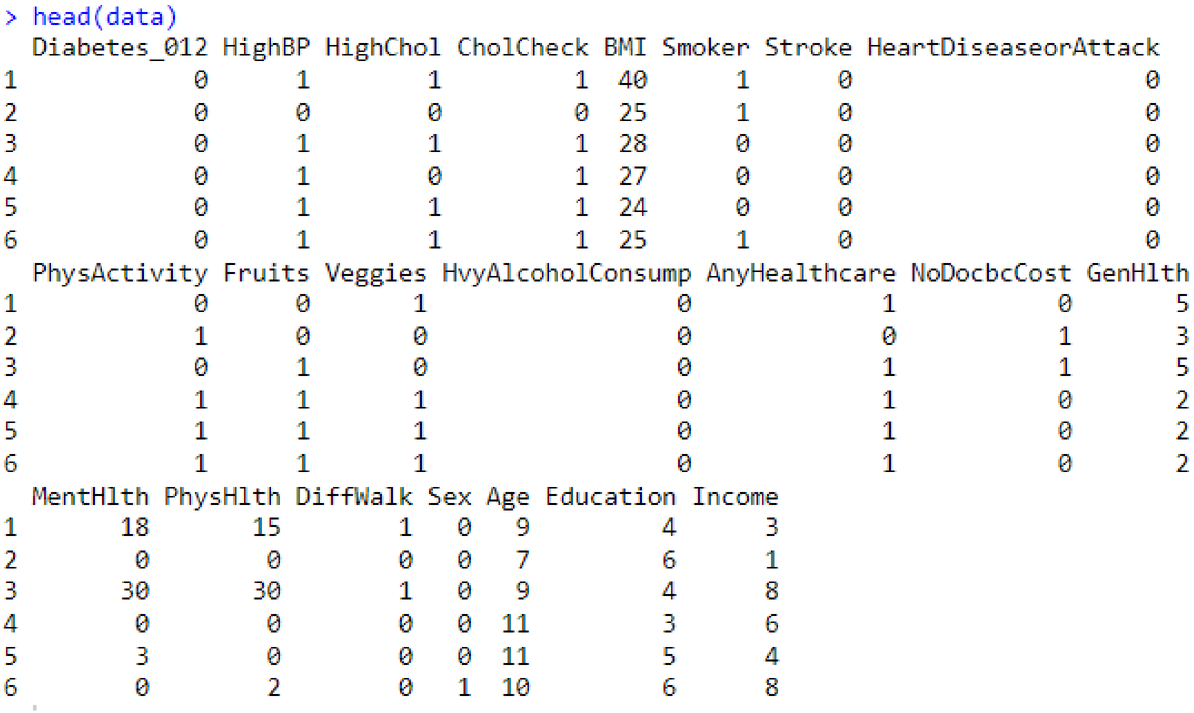
Variables present in the database

"Str" is used to view each of the variables contained in the database.



```
> str(data)
'data.frame':	253680 obs. of  22 variables:
 $ Diabetes_012        : num  0 0 0 0 0 0 0 0 2 0 ...
 $ HighBP              : num  1 0 1 1 1 1 1 1 1 0 ...
 $ HighChol            : num  1 0 1 0 1 1 0 1 1 0 ...
 $ CholCheck           : num  1 0 1 1 1 1 1 1 1 1 ...
 $ BMI                 : num  40 25 28 27 24 25 30 25 30 24 ...
 $ Smoker              : num  1 1 0 0 0 1 1 1 1 0 ...
 $ Stroke              : num  0 0 0 0 0 0 0 0 0 0 ...
 $ HeartDiseaseorAttack: num  0 0 0 0 0 0 0 0 1 0 ...
 $ PhysActivity        : num  0 1 0 1 1 1 0 1 0 0 ...
 $ Fruits              : num  0 0 1 1 1 1 0 0 1 0 ...
 $ Veggies             : num  1 0 0 1 1 1 0 1 1 1 ...
 $ HvyAlcoholConsump   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AnyHealthcare       : num  1 0 1 1 1 1 1 1 1 1 ...
 $ NoDocbcCost         : num  0 1 1 0 0 0 0 0 0 0 ...
 $ GenHlth             : num  5 3 5 2 2 2 3 3 5 2 ...
 $ MentHlth            : num  18 0 30 0 3 0 0 0 30 0 ...
 $ PhysHlth            : num  15 0 30 0 0 2 14 0 30 0 ...
 $ DiffWalk            : num  1 0 1 0 0 0 0 1 1 0 ...
 $ Sex                 : num  0 0 0 0 0 1 0 0 0 1 ...
 $ Age                 : num  9 7 9 11 11 10 9 11 9 8 ...
 $ Education           : num  4 6 4 3 5 6 6 4 5 4 ...
 $ Income              : num  3 1 8 6 4 8 7 4 1 3 ...
```

Figure 1: Variables



```
> head(data)
  Diabetes_012 HighBP HighChol CholCheck BMI Smoker Stroke HeartDiseaseorAttack
1            0      1        1         1  40      1      0                    0
2            0      0        0         0  25      1      0                    0
3            0      1        1         1  28      0      0                    0
4            0      1        0         1  27      0      0                    0
5            0      1        1         1  24      0      0                    0
6            0      1        1         1  25      1      0                    0
  PhysActivity Fruits Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost GenHlth
1            0      0       1                 0             1           0       5
2            1      0       0                 0             0           1       3
3            0      1       0                 0             1           1       5
4            1      1       1                 0             1           0       2
5            1      1       1                 0             1           0       2
6            1      1       1                 0             1           0       2
  MentHlth PhysHlth DiffWalk Sex Age Education Income
1       18       15        1   0   9         4      3
2        0        0        0   0   7         6      1
3       30       30        1   0   9         4      8
4        0        0        0   0  11         3      6
5        3        0        0   0  11         5      4
6        0        2        0   1  10         6      8
```

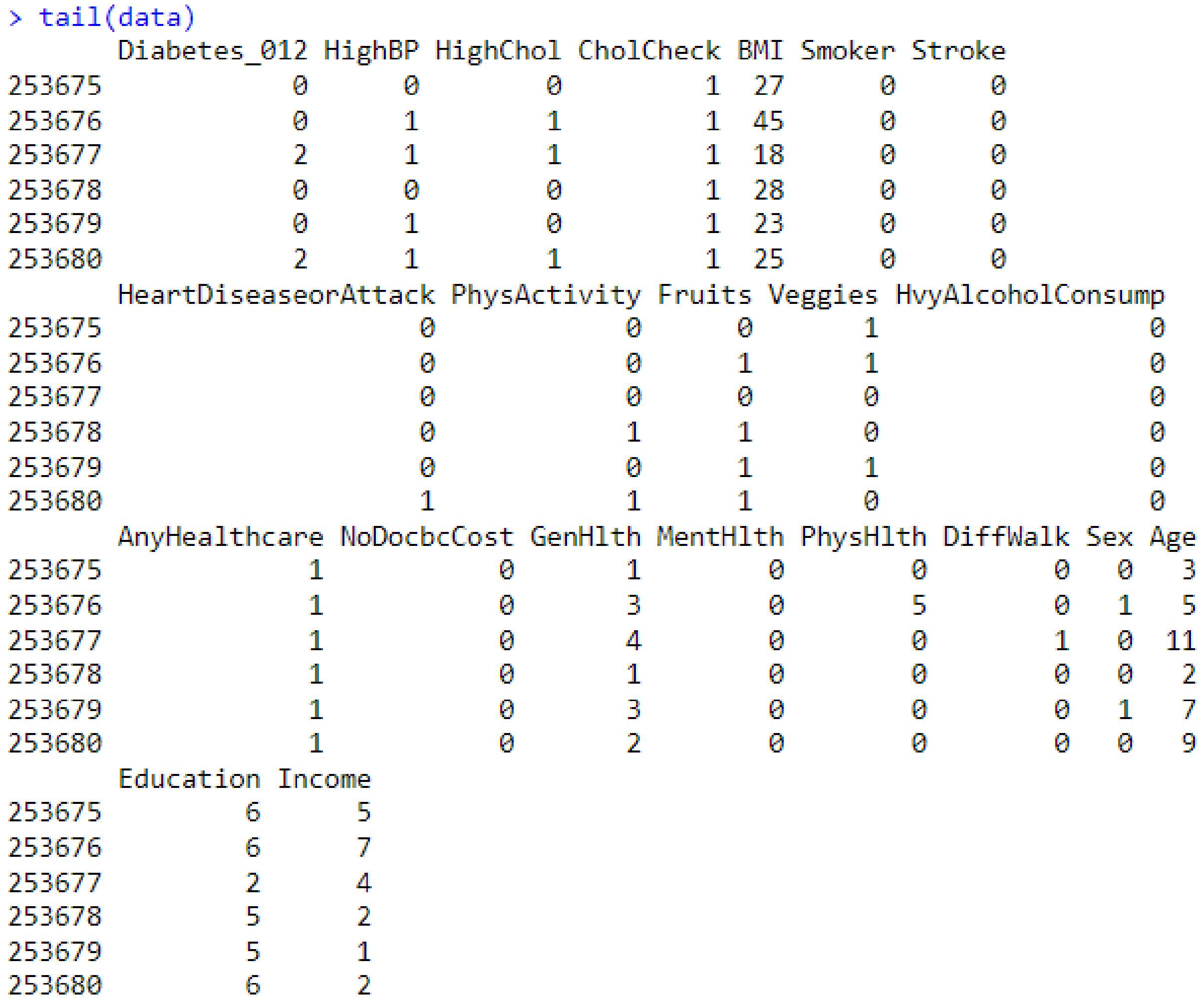
Figure 2: Variables

First Observations

With the “head” function it shows me the first observations of the ENTIRE database.

Ultimas Observaciones

With the “tail” function it shows me the latest observations of the ENTIRE database.

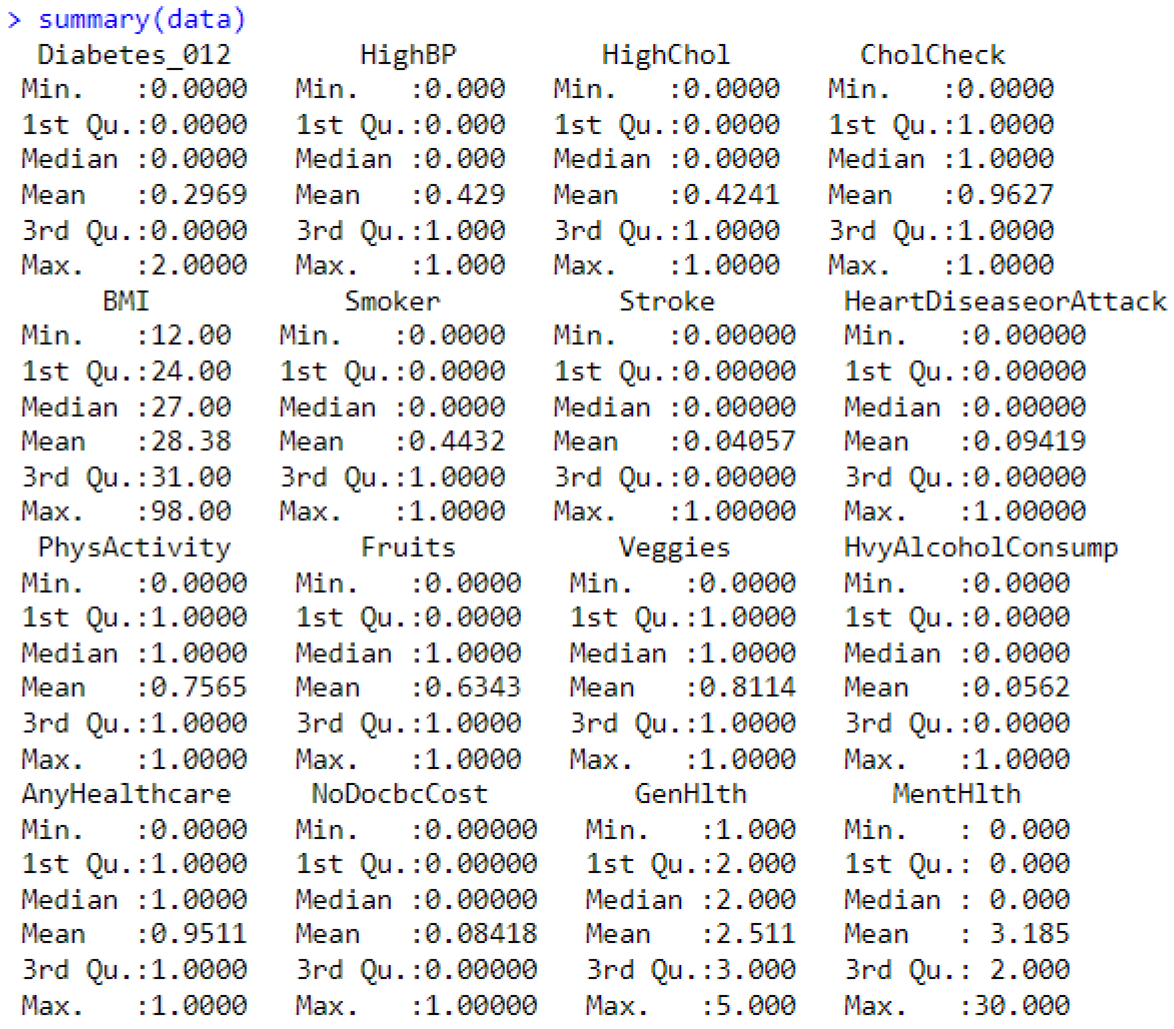


```
> tail(data)
       Diabetes_012 HighBP HighChol CholCheck BMI Smoker Stroke
253675            0      0        0         1  27      0      0
253676            0      1        1         1  45      0      0
253677            2      1        1         1  18      0      0
253678            0      0        0         1  28      0      0
253679            0      1        0         1  23      0      0
253680            2      1        1         1  25      0      0
       HeartDiseaseorAttack PhysActivity Fruits Veggies HvyAlcoholConsump
253675                    0            0      0       1                 0
253676                    0            0      1       1                 0
253677                    0            0      0       0                 0
253678                    0            1      1       0                 0
253679                    0            0      1       1                 0
253680                    1            1      1       0                 0
       AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
253675             1           0       1        0        0        0   0   3
253676             1           0       3        0        5        0   1   5
253677             1           0       4        0        0        1   0  11
253678             1           0       1        0        0        0   0   2
253679             1           0       3        0        0        0   1   7
253680             1           0       2        0        0        0   0   9
       Education Income
253675         6      5
253676         6      7
253677         2      4
253678         5      2
253679         5      1
253680         6      2
```

Figure 3: Variables

General Summary

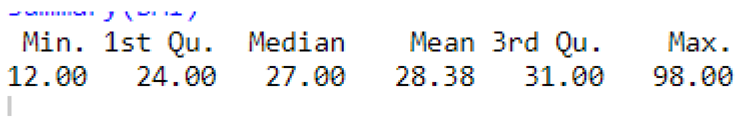
With the “summary” function it shows me a summary of the ENTIRE database, discriminating for each variable, the mean, median, minimum, maximum and others.



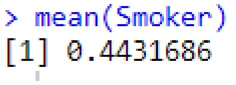
```
> summary(data)
  Diabetes_012        HighBP         HighChol        CholCheck
 Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.0000
 Median :0.0000   Median :0.000   Median :0.0000   Median :1.0000
 Mean   :0.2969   Mean   :0.429   Mean   :0.4241   Mean   :0.9627
 3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :2.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
      BMI            Smoker           Stroke        HeartDiseaseorAttack
 Min.   :12.00   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
 1st Qu.:24.00   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
 Median :27.00   Median :0.0000   Median :0.00000   Median :0.00000
 Mean   :28.38   Mean   :0.4432   Mean   :0.04057   Mean   :0.09419
 3rd Qu.:31.00   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :98.00   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
  PhysActivity        Fruits          Veggies        HvyAlcoholConsump
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
 Median :1.0000   Median :1.0000   Median :1.0000   Median :0.0000
 Mean   :0.7565   Mean   :0.6343   Mean   :0.8114   Mean   :0.0562
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
 AnyHealthcare      NoDocbcCost         GenHlth         MentHlth
 Min.   :0.0000   Min.   :0.00000   Min.   :1.000   Min.   : 0.000
 1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 0.000
 Median :1.0000   Median :0.00000   Median :2.000   Median : 0.000
 Mean   :0.9511   Mean   :0.08418   Mean   :2.511   Mean   : 3.185
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:3.000   3rd Qu.: 2.000
 Max.   :1.0000   Max.   :1.00000   Max.   :5.000   Max.   :30.000
```

To see the summary of each variable you must use the "attach" function

Now it is possible to request the summary or the mean, or median among others for each variable



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   24.00   27.00   28.38   31.00   98.00
```



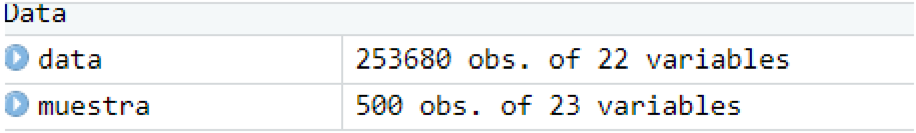
```
> mean(Smoker)
[1] 0.4431686
```

We can request the variance per variable and/or the standard deviation

Database Sampling

I will start using the following function, which allows me to choose a specific sample. If I don't have it, every time I compile the code it will show me a different sample.

We will call a new variable for the random sample and select the number of observations to show, I chose 500



Data	
data	253680 obs. of 22 variables
muestra	500 obs. of 23 variables

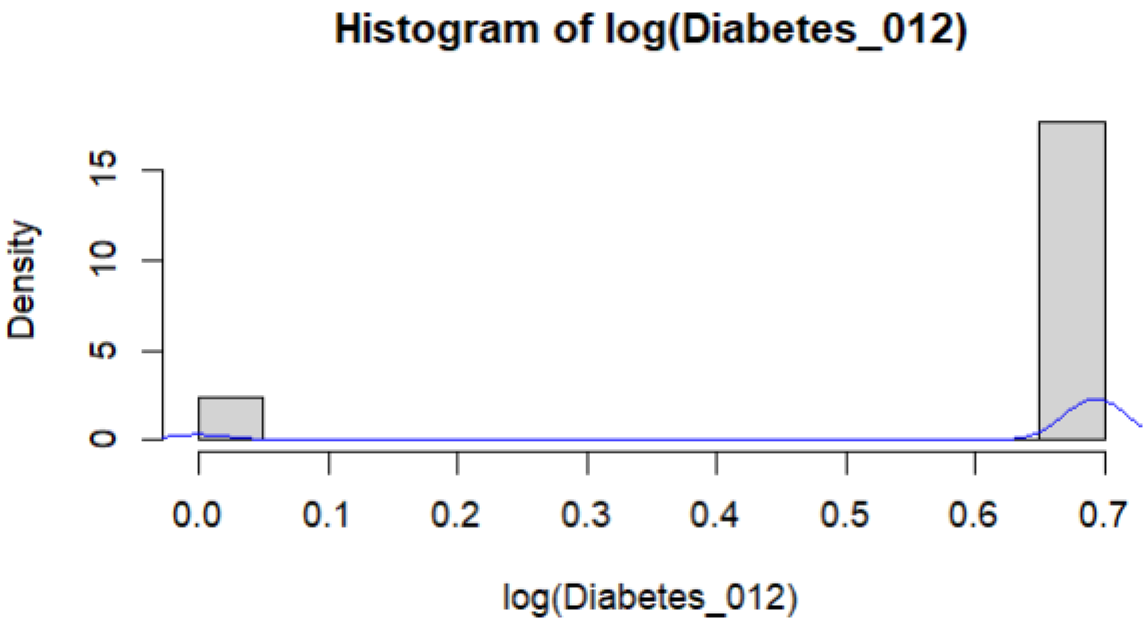
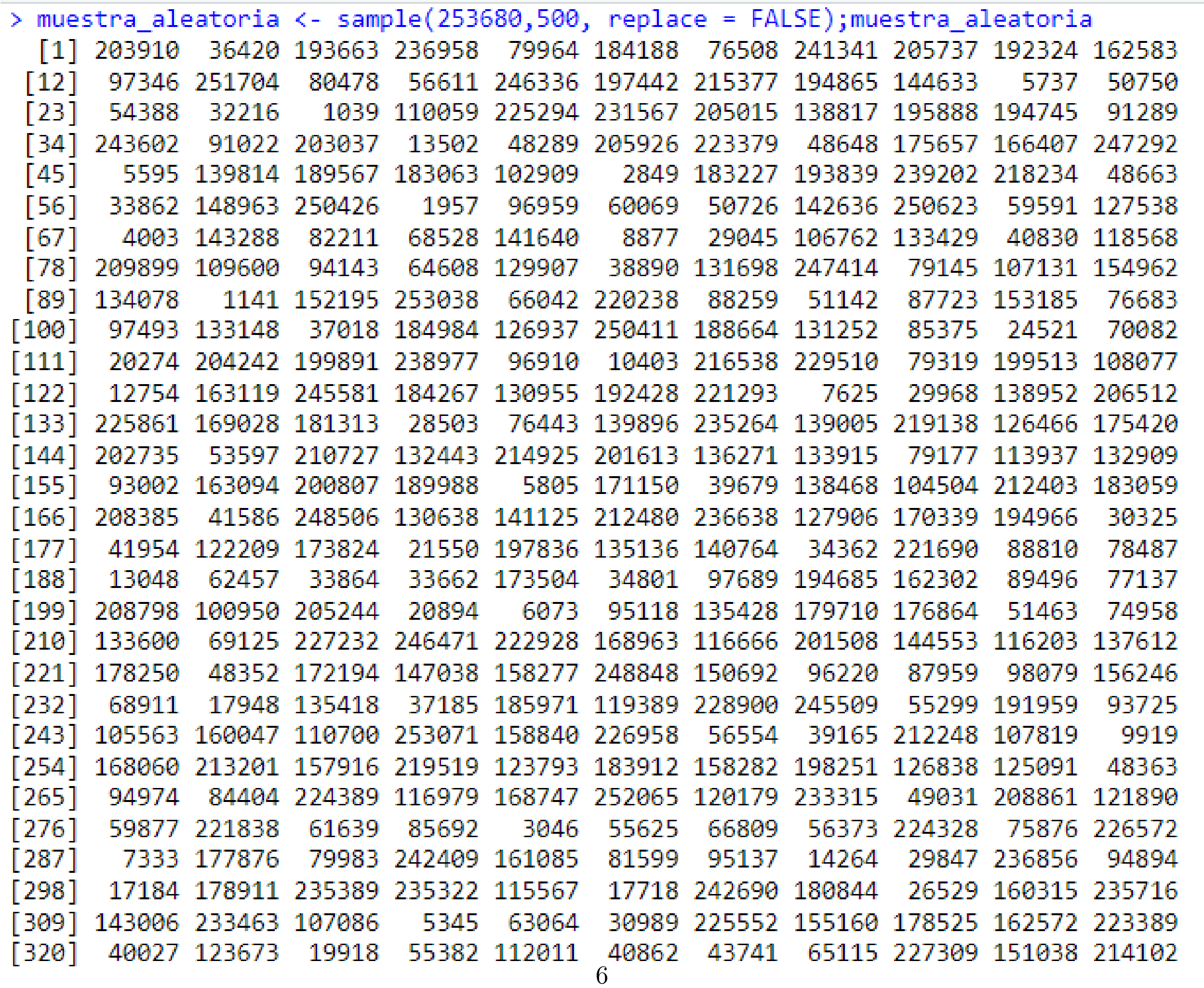


Figure 4: Variables



```
> muestra_aleatoria <- sample(253680,500, replace = FALSE);muestra_aleatoria
  [1] 203910  36420 193663 236958  79964 184188  76508 241341 205737 192324 162583
 [12]  97346 251704  80478  56611 246336 197442 215377 194865 144633   5737  50750
 [23]  54388  32216   1039 110059 225294 231567 205015 138817 195888 194745  91289
 [34] 243602  91022 203037  13502  48289 205926 223379  48648 175657 166407 247292
 [45]   5595 139814 189567 183063 102909   2849 183227 193839 239202 218234  48663
 [56]  33862 148963 250426   1957  96959  60069  50726 142636 250623  59591 127538
 [67]   4003 143288  82211  68528 141640   8877  29045 106762 133429  40830 118568
 [78] 209899 109600  94143  64608 129907  38890 131698 247414  79145 107131 154962
 [89] 134078   1141 152195 253038  66042 220238  88259  51142  87723 153185  76683
[100]  97493 133148  37018 184984 126937 250411 188664 131252  85375  24521  70082
[111]  20274 204242 199891 238977  96910  10403 216538 229510  79319 199513 108077
[122]  12754 163119 245581 184267 130955 192428 221293   7625  29968 138952 206512
[133] 225861 169028 181313  28503  76443 139896 235264 139005 219138 126466 175420
[144] 202735  53597 210727 132443 214925 201613 136271 133915  79177 113937 132909
[155]  93002 163094 200807 189988   5805 171150  39679 138468 104504 212403 183059
[166] 208385  41586 248506 130638 141125 212480 236638 127906 170339 194966  30325
[177]  41954 122209 173824  21550 197836 135136 140764  34362 221690  88810  78487
[188]  13048  62457  33864  33662 173504  34801  97689 194685 162302  89496  77137
[199] 208798 100950 205244  20894   6073  95118 135428 179710 176864  51463  74958
[210] 133600  69125 227232 246471 222928 168963 116666 201508 144553 116203 137612
[221] 178250  48352 172194 147038 158277 248848 150692  96220  87959  98079 156246
[232]  68911  17948 135418  37185 185971 119389 228900 245509  55299 191959  93725
[243] 105563 160047 110700 253071 158840 226958  56554  39165 212248 107819   9919
[254] 168060 213201 157916 219519 123793 183912 158282 198251 126838 125091  48363
[265]  94974  84404 224389 116979 168747 252065 120179 233315  49031 208861 121890
[276]  59877 221838  61639  85692   3046  55625  66809  56373 224328  75876 226572
[287]   7333 177876  79983 242409 161085  81599  95137  14264  29847 236856  94894
[298]  17184 178911 235389 235322 115567  17718 242690 180844  26529 160315 235716
[309] 143006 233463 107086   5345  63064  30989 225552 155160 178525 162572 223389
[320]  40027 123673  19918  55382 112011  40862  43741  65115 227309 151038 214102
```

We select the sort function to organize our sample and call a new variable

To view it in a table we select a name for the sample, call the main database, and write "view"

Now we can do the same as we did before, but with the sample, select a summary, the mean of a variable, the median, the maximum, minimum or others.

I can show the general data of the sample

I can make a sample table

Figure 5: Variables

I will show a sample bar graph

You can also show a pie chart but it is more useful in categorical variables, in this case select only 50 observations so that it looks better

Graph correlation, the closest to diabetes is the body mass index "BMI"

To check how many people have diabetes, prediabetes or do not have in numerical values

We call a new variable new to change the variable from numeric to categorical. The function "as.factor" is used to encode a vector as a factor (the terms "category" and "enumerated type" are also used for factors).

Taking into account what was provided by the teacher, it is taken into account that patients without diabetes are represented as 0, prediabetes 1 and diabetes 2

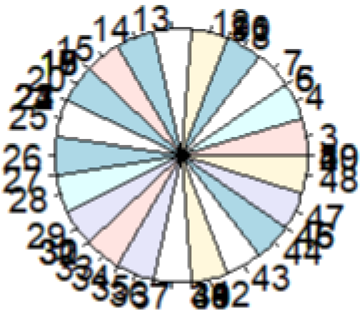


Figure 6: Variables

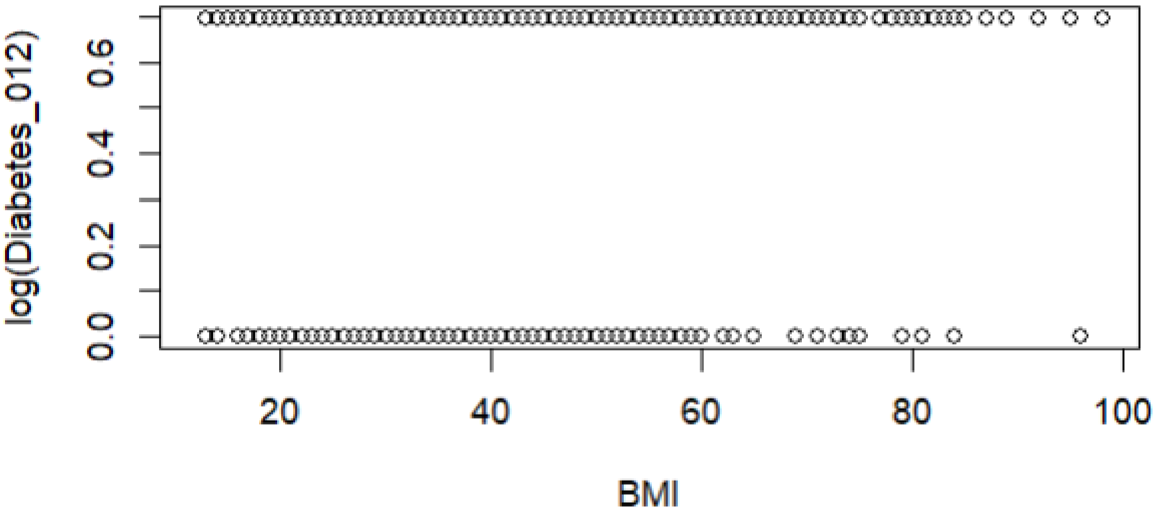
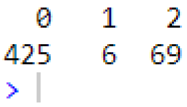
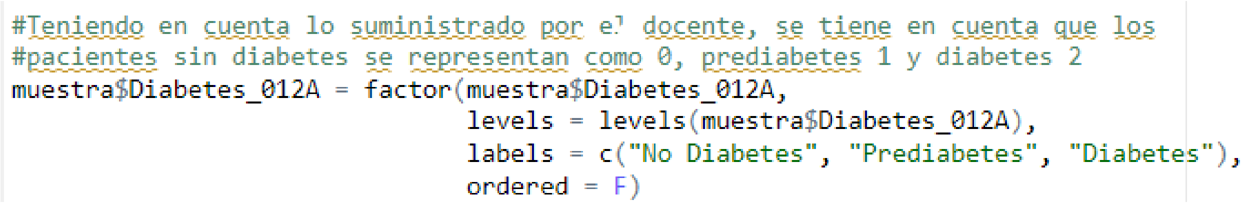


Figure 7: Variables



```
  0   1   2
425   6  69
>
```

Figure 8: Variables



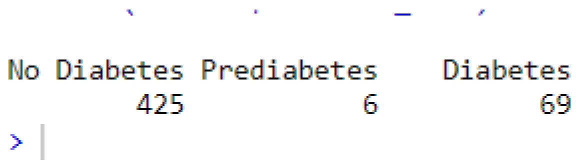
```
#Teniendo en cuenta lo suministrado por e⁷ docente, se tiene en cuenta que los
#pacientes sin diabetes se representan como 0, prediabetes 1 y diabetes 2
muestra$Diabetes_012A = factor(muestra$Diabetes_012A,
                               levels = levels(muestra$Diabetes_012A),
                               labels = c("No Diabetes", "Prediabetes", "Diabetes"),
                               ordered = F)
```

Figure 9: Variables

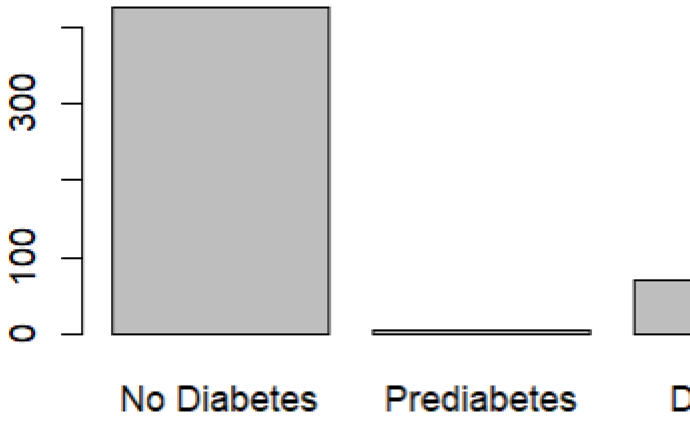
With the levels function, it links 0, 1 and 2 of the variable Diabetes_012 and then with labels I place the names that I will change, in this case, no diabetes, prediabetes and diabetes.

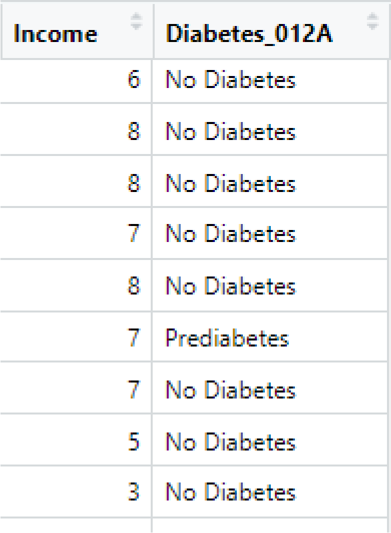
“Str” again to see if the change was made to these categorical variables

#Create a table that shows me how many patients have diabetes, how many do not, and how many have prediabetes.



```
No Diabetes Prediabetes    Diabetes 
        425           6          69 
> 
```



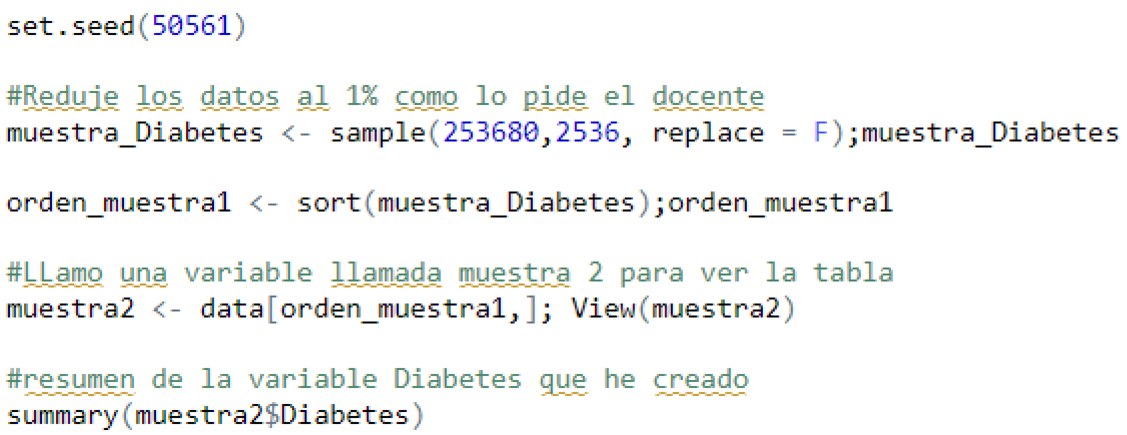


Income	Diabetes_012A
6	No Diabetes
8	No Diabetes
8	No Diabetes
7	No Diabetes
8	No Diabetes
7	Prediabetes
7	No Diabetes
5	No Diabetes
3	No Diabetes

# Segundo punto

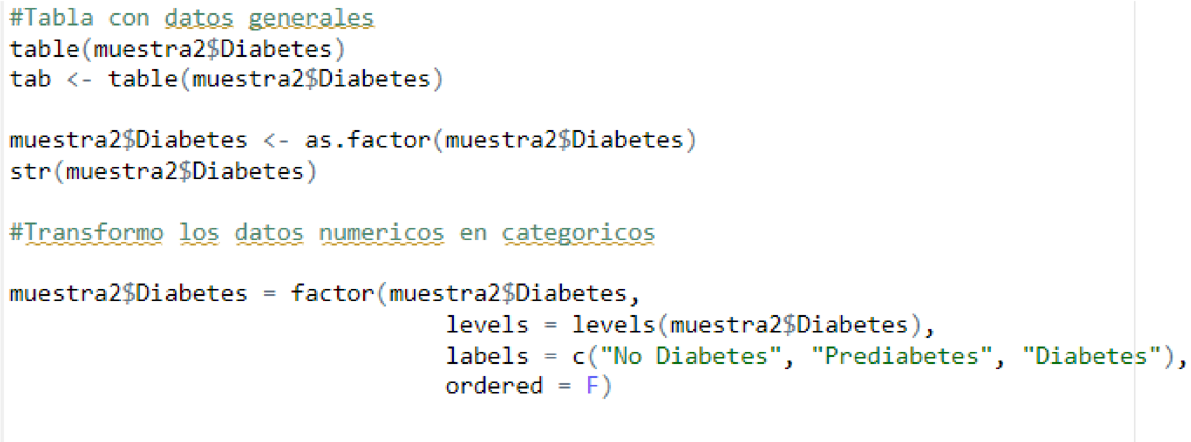
Se selecciona una muestra aleatoria en especifico y se reduce al 1% de la base total

Se creyo necesario pasar valores a categoricos para luego binarizarlos



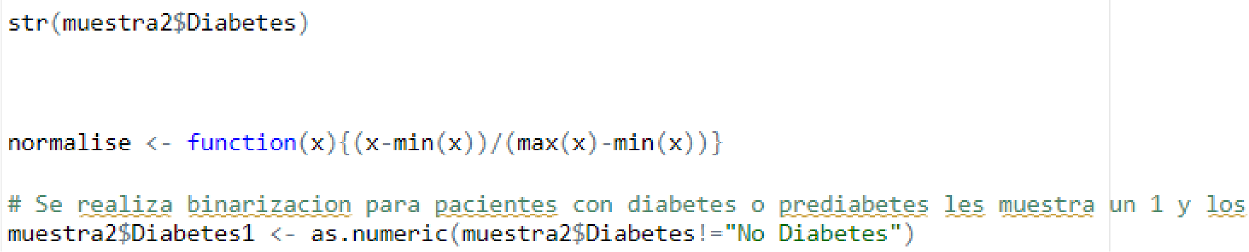
```
set.seed(50561)

#Reduje los datos al 1% como lo pide el docente
muestra_Diabetes <- sample(253680,2536, replace = F);muestra_Diabetes

orden_muestra1 <- sort(muestra_Diabetes);orden_muestra1

#LLamo una variable llamada muestra 2 para ver la tabla
muestra2 <- data[orden_muestra1,]; View(muestra2)

#resumen de la variable Diabetes que he creado
summary(muestra2$Diabetes)
```

Figure 10: Variables

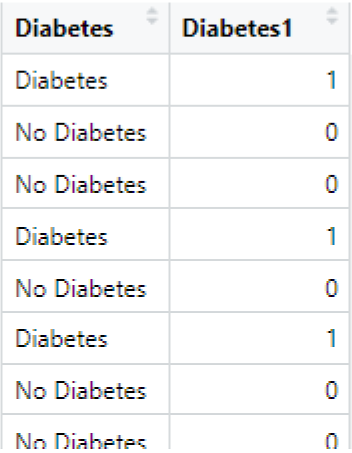


```
#Tabla con datos generales
table(muestra2$Diabetes)
tab <- table(muestra2$Diabetes)

muestra2$Diabetes <- as.factor(muestra2$Diabetes)
str(muestra2$Diabetes)

#Transformo los datos numericos en categoricos

muestra2$Diabetes = factor(muestra2$Diabetes,
                              levels = levels(muestra2$Diabetes),
                              labels = c("No Diabetes", "Prediabetes", "Diabetes"),
                              ordered = F)
```

Por ultimo segun lo explicado por el docente se binariza para que muestre 0 si es igual a No diabetes y a los valores diferentes muestre 1



```
str(muestra2$Diabetes)



normalise <- function(x){(x-min(x))/(max(x)-min(x))}

# Se realiza binarizacion para pacientes con diabetes o prediabetes les muestra un 1 y los
muestra2$Diabetes1 <- as.numeric(muestra2$Diabetes!="No Diabetes")
```



Diabetes	Diabetes1
Diabetes	1
No Diabetes	0
No Diabetes	0
Diabetes	1
No Diabetes	0
Diabetes	1
No Diabetes	0
No Diabetes	0