Practica2

Sandra Milena Patiño

3/6/2020

#Práctica 2 (35% nota final)

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis. Objetivos Los objetivos concretos de esta práctica son:
- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

La Organización Mundial de la Salud ha estimado que ocurren 12 millones de muertes en todo el mundo, cada año debido a enfermedades del corazón. La mitad de las muertes en los Estados Unidos y otros países desarrollados se deben a enfermedades cardiovasculares. El pronóstico temprano de las enfermedades cardiovasculares puede ayudar a tomar decisiones sobre los cambios en el estilo de vida en pacientes de alto riesgo y, a su vez, reducir las complicaciones. Esta investigación tiene la intención de identificar los factores más relevantes de riesgo de enfermedad cardíaca, así como predecir el riesgo general mediante regresión logística

El conjunto de datos está disponible públicamente en el sitio web de Kaggle, y proviene de un estudio cardiovascular en curso en residentes de la ciudad de Framingham, Massachusetts. El objetivo de la clasificación es predecir si el paciente tiene 10 años de riesgo de enfermedad coronaria (CHD) en el futuro. El conjunto de datos proporciona la información del paciente. Incluye más de 4,000 registros y 15 atributos.

1. Carga de Datos

Cargamos los datos del archivo csv descargado de www.kaggle.com

```
dataset <- read.csv("datasets_framingham.csv")
head(dataset)
```

```
##   male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1  39         4             0          0      0               0
## 2    0  46         2             0          0      0               0
## 3    1  48         1             1         20      0               0
## 4    0  61         3             1         30      0               0
## 5    0  46         3             1         23      0               0
## 6    0  43         2             0          0      0               0
##   prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
## 1            0        0     195 106.0    70 26.97        80      77          0
## 2            0        0     250 121.0    81 28.73        95      76          0
## 3            0        0     245 127.5    80 25.34        75      70          0
## 4            1        0     225 150.0    95 28.58        65     103          1
## 5            0        0     285 130.0    84 23.10        85      85          0
## 6            1        0     228 180.0   110 30.30        77      99          0
```

Revisamos la estructura del dataset

```
str(dataset)
```

```
## 'data.frame':    4240 obs. of  16 variables:
##  $ male           : int  1 0 1 0 0 0 0 0 1 1 ...
##  $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
##  $ education      : int  4 2 1 3 3 2 1 2 1 1 ...
##  $ currentSmoker  : int  0 0 1 1 1 0 0 1 0 1 ...
##  $ cigsPerDay     : int  0 0 20 30 23 0 0 20 0 30 ...
##  $ BPMeds         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentHyp   : int  0 0 0 1 0 1 0 0 1 1 ...
##  $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ totChol        : int  195 250 245 225 285 228 205 313 260 225 ...
##  $ sysBP          : num  106 121 128 150 130 ...
##  $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
##  $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
##  $ heartRate      : int  80 95 75 65 85 77 60 79 76 93 ...
##  $ glucose        : int  77 76 70 103 85 99 85 78 79 88 ...
##  $ TenYearCHD     : int  0 0 0 1 0 0 1 0 0 0 ...
```

Diccionario de datos - Descripción de los atributos

Analisis de datos

No	Nombre	Tipo de dato	Descripción de atributo	Tipo de atributo
1.	male	Integer	Sexo	Independente
			0 : Femenino	
			1: Masculino	
2.	age	Integer	Edad del paciente	Independente
3.	education	Integer	Nivel de educación del pacente	Independente
			4 : Universidad	
			3 : Alguna universidad o escuela vocacional	
			2 : Escuela secundaria o GED	
			1: Alguna escuela secundaria	
4.	currentSmoker	Integer	Si el paciente es o no fumador actual	Independente
			0 : No	
			1: Si	
5.	cigsPerday	Integer	Cantidad de cigarrillos que la persona fumaba en promedio en un día	Independente
6.	BPMeds	Integer	Si el paciente estaba tomando medicamentos para la presión arterial	Independente
			0 : No	
			1: Si	
7.	prevalentStroke	Integer	Si el paciente había tenido previamente un accidente cerebrovascular	Independente
			0 : No	
			1: Si	
8.	prevalentHyp	Integer	Si el paciente era o no hipertenso	Independente
			0 : No	
			1: Si	
9.	diabetes	Integer	Si el paciente tenía o no diabetes	Independente
			0 : No	
			1: Si	
10.	totChol	Integer	Nivel de colesterol total mg/dL	Independente
11.	*sys*BP	Double	Presión arterial sistólica mmHg	Independente
12.	diaBP	Double	Presión arterial diastólica mmHg	Independente
13.	BMI	Double	Indice de masa corporal	Independente
14.	heartRate	Integer	Frecuencia cardíaca pulsaciones/min	Independente
15.	glucose	Integer	Nivel de glucosa mg/dL	Independente
16.	TenYearCHD	Integer	10 años de riesgo de enfermedad coronaria CHD	Objetivo
			0 : No	
			1: Si	

Figure 1: Tipos de Datos

```r
summary(dataset)
```

```
##       male              age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                   NA's   :105
##    cigsPerDay        BPMeds       prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.006   Mean   :0.02962   Mean   :0.005896   Mean   :0.3106
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##  NA's   :29       NA's   :53
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.0
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.0
##  Mean   :0.02571   Mean   :236.7   Mean   :132.4   Mean   : 82.9
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.0
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.5
##                    NA's   :50
##       BMI          heartRate         glucose         TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.0000
##  1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000
##  Median :25.40   Median : 75.00   Median : 78.00   Median :0.0000
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.96   Mean   :0.1519
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.0000
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.0000
##  NA's   :19      NA's   :1        NA's   :388
```

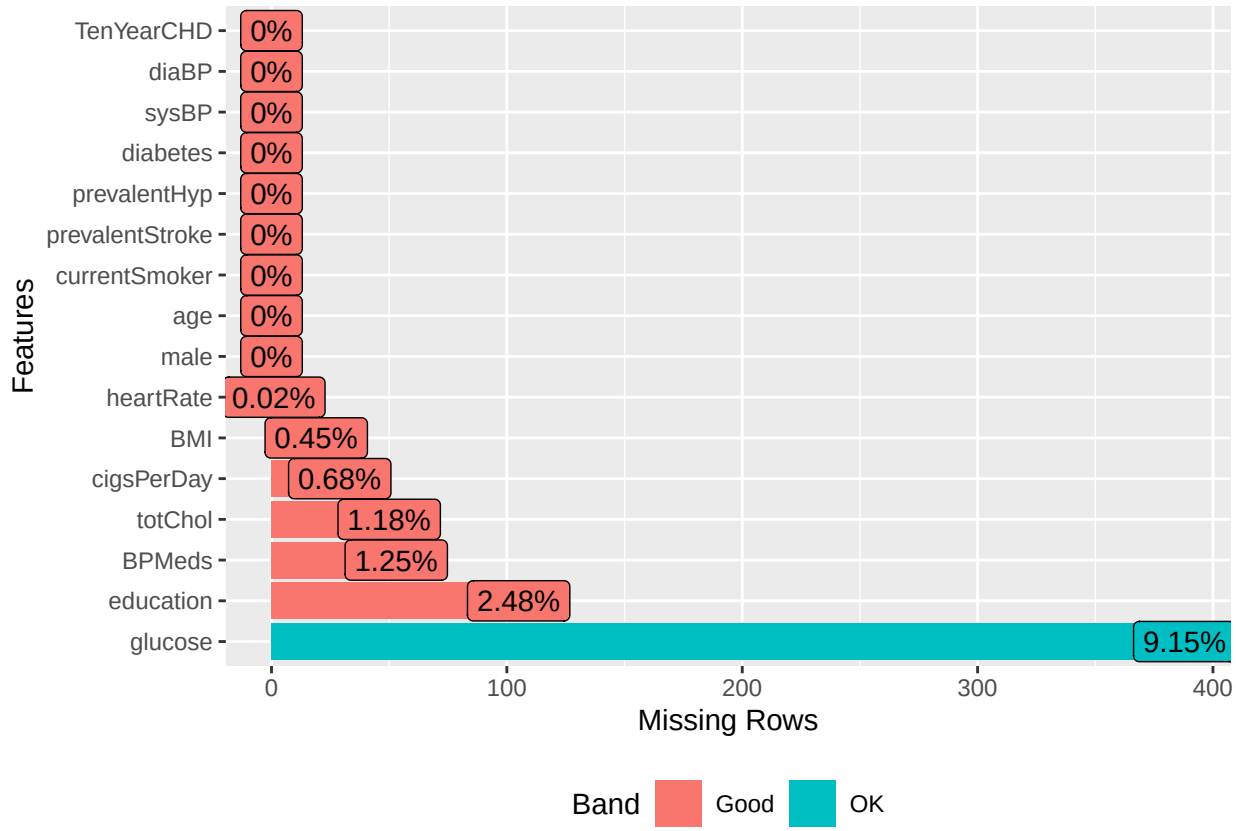
- Hay mayor cantidad de pacientes mujeres que hombres
- Edad promedio de los pacentes 49 años
- El mayor grupo de personas no terminaron la secundaria
- La cantidad de fumadores es aprox. igual a la de no fumadores
- La mayoria de pacientes no toma medicamentos para la presion arterial, no ha tenido accidentes cerebrovasculares ni sufre de diabetes.
- En el indice de masa corporal observamos una media de 25.8 esto indica sobrepeso
- Existen una gran cantidad de valores ausentes (NA) que deben ser imputados

```r
colSums(is.na(dataset))
```

```
##           male            age      education   currentSmoker      cigsPerDay
##              0              0            105              0             29
##         BPMeds prevalentStroke   prevalentHyp       diabetes        totChol
##             53              0              0              0             50
##          sysBP          diaBP            BMI      heartRate        glucose
##              0              0             19              1            388
```

```
##       TenYearCHD
##                0
```

```
options(repr.plot.width=14, repr.plot.height=6)
plot_missing(dataset)
```



Porcentaje de datos perdidos

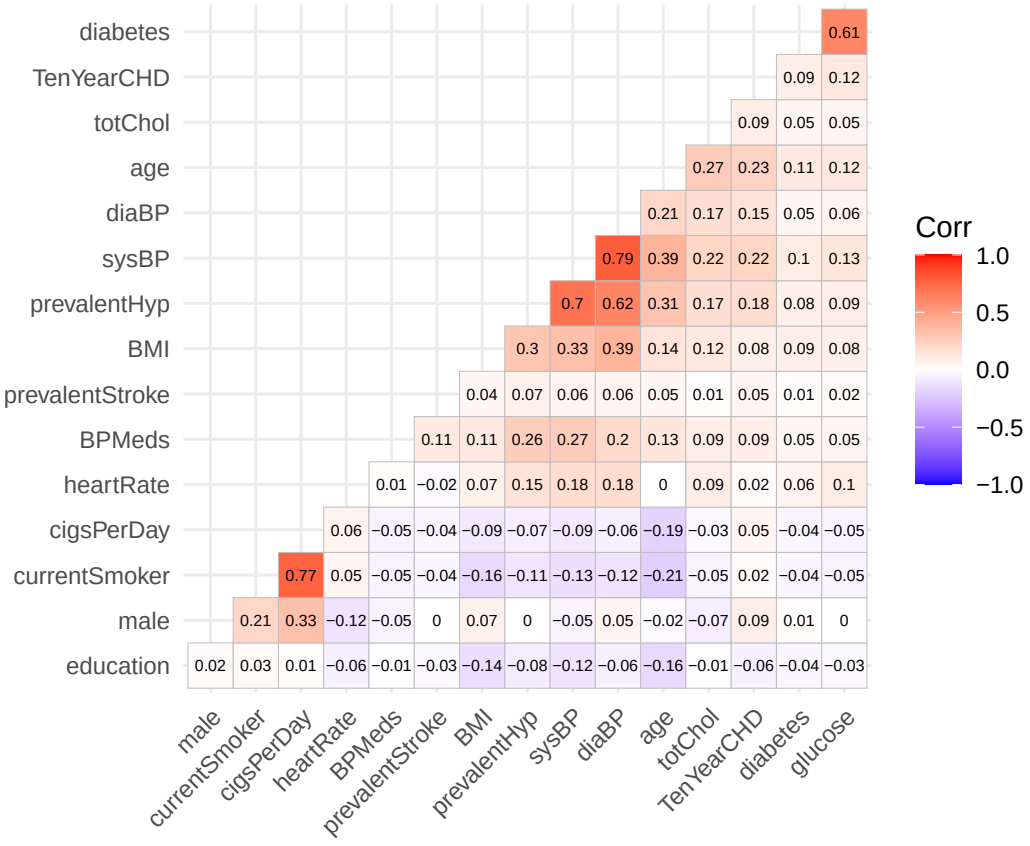
Correlación

```
corr <- round(cor(dataset, use="complete.obs"), 2)
corr
```

```
##                    male   age education currentSmoker cigsPerDay BPMeds
## male               1.00 -0.02      0.02          0.21       0.33  -0.05
## age               -0.02  1.00     -0.16         -0.21      -0.19   0.13
## education          0.02 -0.16      1.00          0.03       0.01  -0.01
## currentSmoker      0.21 -0.21      0.03          1.00       0.77  -0.05
## cigsPerDay         0.33 -0.19      0.01          0.77       1.00  -0.05
## BPMeds            -0.05  0.13     -0.01         -0.05      -0.05   1.00
## prevalentStroke    0.00  0.05     -0.03         -0.04      -0.04   0.11
## prevalentHyp       0.00  0.31     -0.08         -0.11      -0.07   0.26
## diabetes           0.01  0.11     -0.04         -0.04      -0.04   0.05
## totChol           -0.07  0.27     -0.01         -0.05      -0.03   0.09
## sysBP             -0.05  0.39     -0.12         -0.13      -0.09   0.27
## diaBP              0.05  0.21     -0.06         -0.12      -0.06   0.20
```

```
## BMI              0.07  0.14     -0.14         -0.16      -0.09   0.11
## heartRate       -0.12  0.00     -0.06          0.05       0.06   0.01
## glucose          0.00  0.12     -0.03         -0.05      -0.05   0.05
## TenYearCHD       0.09  0.23     -0.06          0.02       0.05   0.09
##                 prevalentStroke prevalentHyp diabetes totChol sysBP diaBP   BMI
## male                       0.00         0.00     0.01   -0.07 -0.05  0.05  0.07
## age                        0.05         0.31     0.11    0.27  0.39  0.21  0.14
## education                 -0.03        -0.08    -0.04   -0.01 -0.12 -0.06 -0.14
## currentSmoker             -0.04        -0.11    -0.04   -0.05 -0.13 -0.12 -0.16
## cigsPerDay                -0.04        -0.07    -0.04   -0.03 -0.09 -0.06 -0.09
## BPMeds                     0.11         0.26     0.05    0.09  0.27  0.20  0.11
## prevalentStroke            1.00         0.07     0.01    0.01  0.06  0.06  0.04
## prevalentHyp               0.07         1.00     0.08    0.17  0.70  0.62  0.30
## diabetes                   0.01         0.08     1.00    0.05  0.10  0.05  0.09
## totChol                    0.01         0.17     0.05    1.00  0.22  0.17  0.12
## sysBP                      0.06         0.70     0.10    0.22  1.00  0.79  0.33
## diaBP                      0.06         0.62     0.05    0.17  0.79  1.00  0.39
## BMI                        0.04         0.30     0.09    0.12  0.33  0.39  1.00
## heartRate                 -0.02         0.15     0.06    0.09  0.18  0.18  0.07
## glucose                    0.02         0.09     0.61    0.05  0.13  0.06  0.08
## TenYearCHD                 0.05         0.18     0.09    0.09  0.22  0.15  0.08
##                 heartRate glucose TenYearCHD
## male                -0.12    0.00       0.09
## age                  0.00    0.12       0.23
## education           -0.06   -0.03      -0.06
## currentSmoker        0.05   -0.05       0.02
## cigsPerDay           0.06   -0.05       0.05
## BPMeds               0.01    0.05       0.09
## prevalentStroke     -0.02    0.02       0.05
## prevalentHyp         0.15    0.09       0.18
## diabetes             0.06    0.61       0.09
## totChol              0.09    0.05       0.09
## sysBP                0.18    0.13       0.22
## diaBP                0.18    0.06       0.15
## BMI                  0.07    0.08       0.08
## heartRate            1.00    0.10       0.02
## glucose              0.10    1.00       0.12
## TenYearCHD           0.02    0.12       1.00
```

```
ggcorrplot(corr,hc.order = TRUE, type = "lower",lab = TRUE, tl.cex = 9, lab_size = 2, sig.level = .2) +
```



Observamos alta correlación entre diatebes y glucose, currentSmoker y cigsPerDay, prevalentHyp y sys BP, sysBP y diaBP

2. Normalización de las variables cualitativas

Codificamos las variables categoricas cuyos valores sean (1 y 0) a valores representativos tipo char para cada variable

2.1 Male

```
dataset$male <- ifelse( dataset$male=="0" , "F", "M" )
table(dataset$male)
```

```
##
##    F    M
## 2420 1820
```

2.2 currentSmoker

```
dataset$currentSmoker <- ifelse( dataset$currentSmoker=="0" , "N", "S" )
table(dataset$currentSmoker)
```

```
##
##    N    S
## 2145 2095
```

2.3 BPMeds

```
dataset$BPMeds <- ifelse( dataset$BPMeds=="0" , "N", "S" )
table(dataset$BPMeds)
```

```
##
##    N    S
## 4063  124
```

2.4 prevalentStroke

```
dataset$prevalentStroke <- ifelse( dataset$prevalentStroke=="0" , "N", "S" )
table(dataset$prevalentStroke)
```

```
##
##    N    S
## 4215   25
```

2.5 prevalentHyp

```
dataset$prevalentHyp <- ifelse( dataset$prevalentHyp=="0" , "N", "S" )
table(dataset$prevalentHyp)
```

```
##
##    N    S
## 2923 1317
```

2.6 diabetes

```
dataset$diabetes <- ifelse( dataset$diabetes=="0" , "N", "S" )
table(dataset$diabetes)
```

```
##
##    N    S
## 4131  109
```

2.7 TenYearCHD

```
dataset$TenYearCHD <- ifelse( dataset$TenYearCHD=="0" , "N", "S" )
table(dataset$TenYearCHD)
```

```
##
##    N    S
## 3596  644
```

Convertimos las variables categoricas de INT a Factor

```r
dataset$male <- as.factor(dataset$male)
dataset$education <- as.factor(dataset$education)
dataset$currentSmoker <- as.factor(dataset$currentSmoker)
dataset$BPMeds <- as.factor(dataset$BPMeds)
dataset$prevalentStroke <- as.factor(dataset$prevalentStroke)
dataset$prevalentHyp <- as.factor(dataset$prevalentHyp)
dataset$diabetes <- as.factor(dataset$diabetes)
dataset$TenYearCHD <- as.factor(dataset$TenYearCHD)
str(dataset)
```

```
## 'data.frame':    4240 obs. of  16 variables:
##  $ male           : Factor w/ 2 levels "F","M": 2 1 2 1 1 1 1 1 2 2 ...
##  $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
##  $ education      : Factor w/ 4 levels "1","2","3","4": 4 2 1 3 3 2 1 2 1 1 ...
##  $ currentSmoker  : Factor w/ 2 levels "N","S": 1 1 2 2 2 1 1 2 1 2 ...
##  $ cigsPerDay     : int  0 0 20 30 23 0 0 20 0 30 ...
##  $ BPMeds         : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ prevalentStroke: Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ prevalentHyp   : Factor w/ 2 levels "N","S": 1 1 1 2 1 2 1 1 2 2 ...
##  $ diabetes       : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
##  $ totChol        : int  195 250 245 225 285 228 205 313 260 225 ...
##  $ sysBP          : num  106 121 128 150 130 ...
##  $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
##  $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
##  $ heartRate      : int  80 95 75 65 85 77 60 79 76 93 ...
##  $ glucose        : int  77 76 70 103 85 99 85 78 79 88 ...
##  $ TenYearCHD     : Factor w/ 2 levels "N","S": 1 1 1 2 1 1 2 1 1 1 ...
```

No se encuentra errores o inconsistencias en los datos

3. Valores perdidos

Analizar la presencia de valores perdidos. En el caso de detectar algún valor perdido en las variables cuantitativas realizar una imputación de valores en estas variables. La imputación debe hacerse con los 5 vecinos más cercanos usando la distancia de Gower, usando sólo la información de las variables cuantitativas y dentro de éstas, aquellas que tengan sentido en la imputación de la variable. Después de realizar la imputación es necesario verificar que los valores asignados se han copiado sobre el conjunto de datos originales

```r
#Imputación
output <- kNN( dataset, variable=c("education","cigsPerDay","BPMeds","totChol","BMI","heartRate","glucos

dataset[,c("education","cigsPerDay","BPMeds","totChol","BMI","heartRate","glucose")] <- output[,c("educa

# Registros imputados
filas_imp <- dataset[ output$education_imp==TRUE | output$cigsPerDay_imp==TRUE | output$BPMeds_imp==TRU
head(filas_imp)
```

```
##    male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 15    F  39         2             S          9      N               N
## 22    F  43         1             N          0      N               N
## 27    F  60         1             N          0      N               N
## 34    M  61         2             S          5      N               N
## 37    M  56         1             N          0      N               N
```

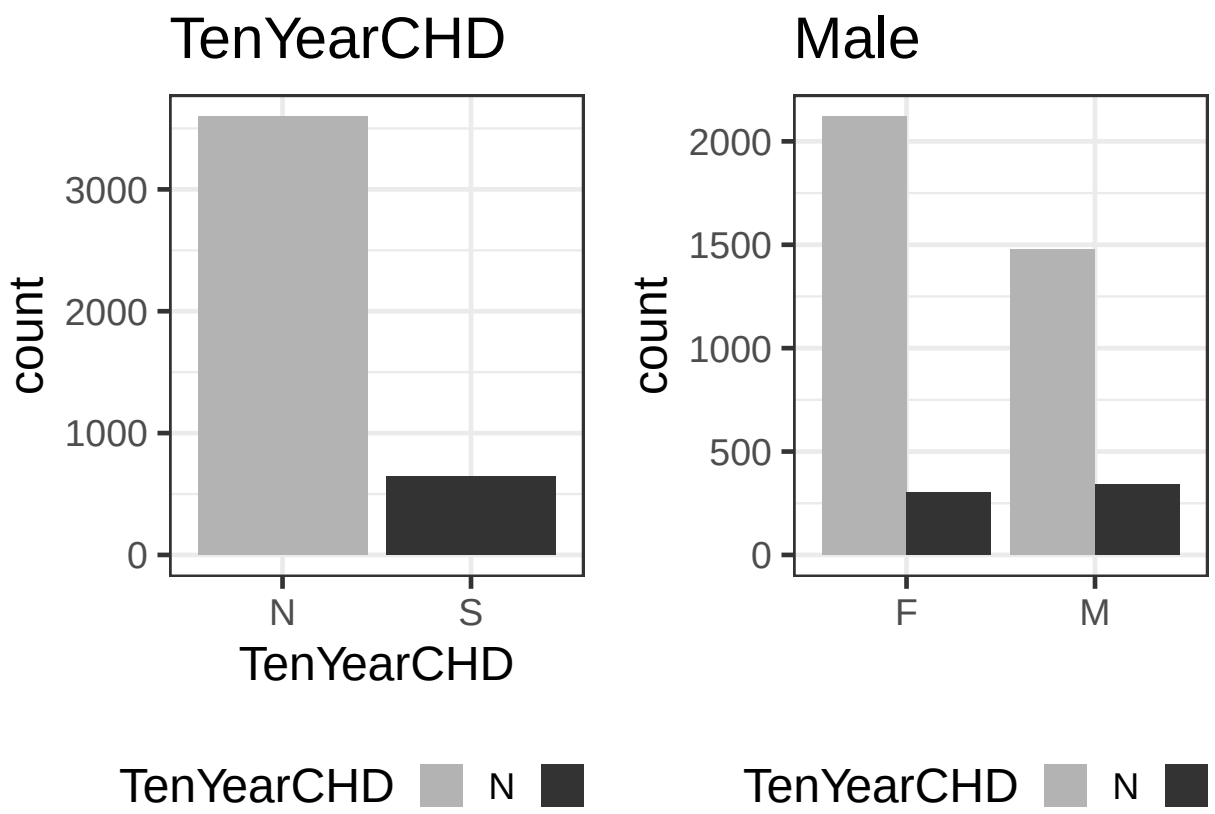
```
## 43    F  52         1             N          0      S               N
##    prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
## 15            N        N     226 114.0  64.0 22.35        85      77          N
## 22            N        N     185 123.5  77.5 29.89        70      77          N
## 27            N        N     260 110.0  72.5 26.59        65      90          N
## 34            N        N     175 134.0  82.5 18.59        72      75          S
## 37            N        N     257 153.5 102.0 28.09        72      75          N
## 43            S        N     233 148.0  92.0 25.09        70      75          S
```

```
colSums(is.na(dataset))
```

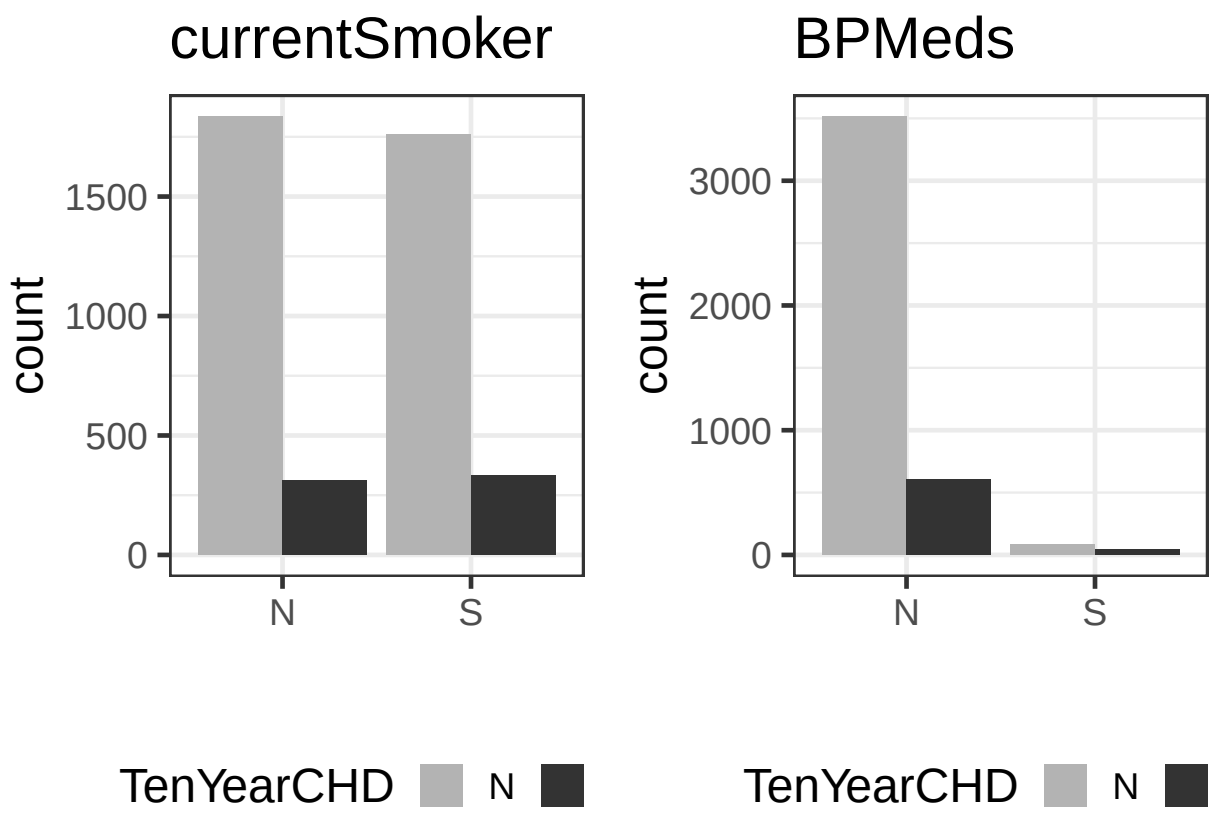
```
##            male             age       education   currentSmoker      cigsPerDay
##               0               0               0               0               0
##          BPMeds prevalentStroke    prevalentHyp        diabetes         totChol
##               0               0               0               0               0
##           sysBP           diaBP             BMI       heartRate         glucose
##               0               0               0               0               0
##      TenYearCHD
##               0
```

4. Graficas de Datos Categoricos

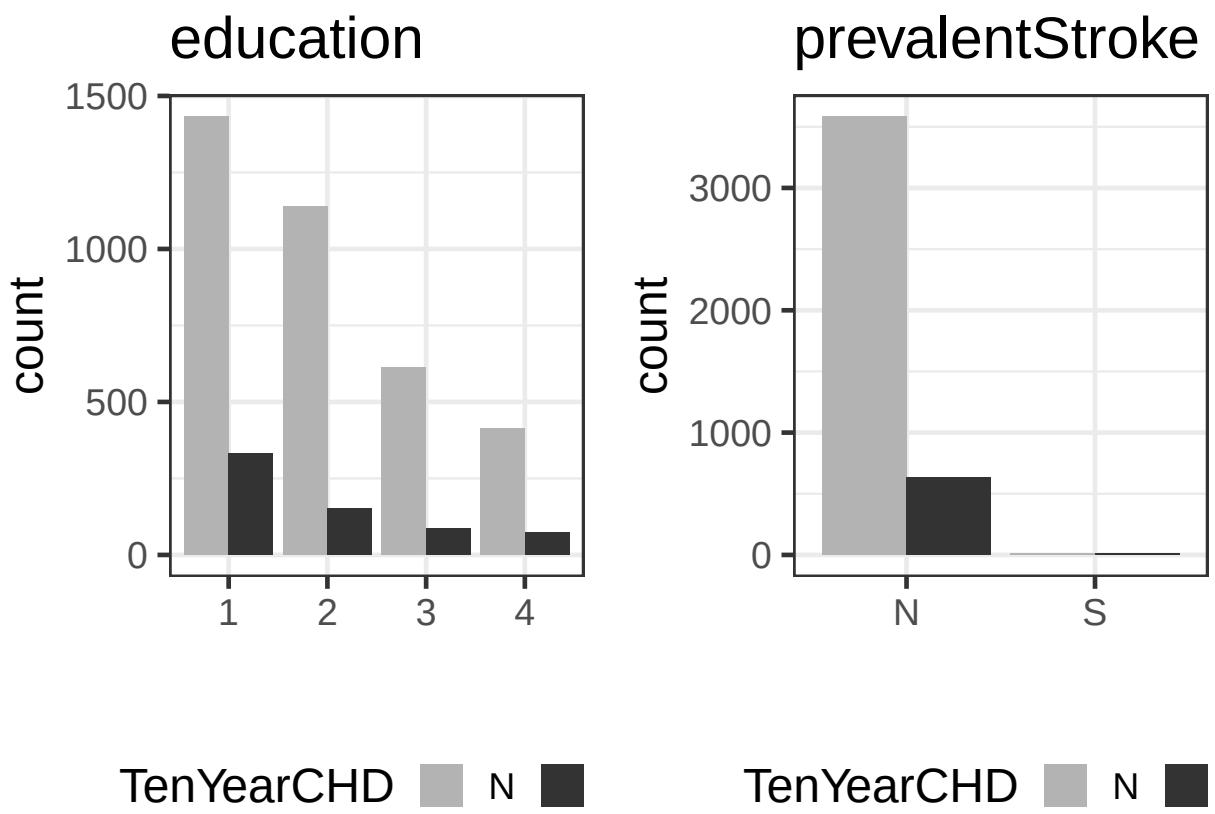
```
a = ggplot(dataset, aes(TenYearCHD, fill = TenYearCHD)) +
  geom_bar(stat = "count") + scale_fill_manual(values=c('grey70', 'grey20')) +
  labs(title = "TenYearCHD") +  theme_bw(base_size = 18) +
  theme(legend.position="bottom")

b = ggplot(dataset, aes(male, fill = TenYearCHD)) +
  geom_bar(stat = "count", position = "dodge") +
  scale_fill_manual(values=c('grey70', 'grey20')) +
  labs(title = "Male", x = "") +  theme_bw(base_size = 18) +
  theme(legend.position="bottom")

options(repr.plot.width=16, repr.plot.height=8)
plot_grid(a,b, ncol = 2, nrow = 1)
```



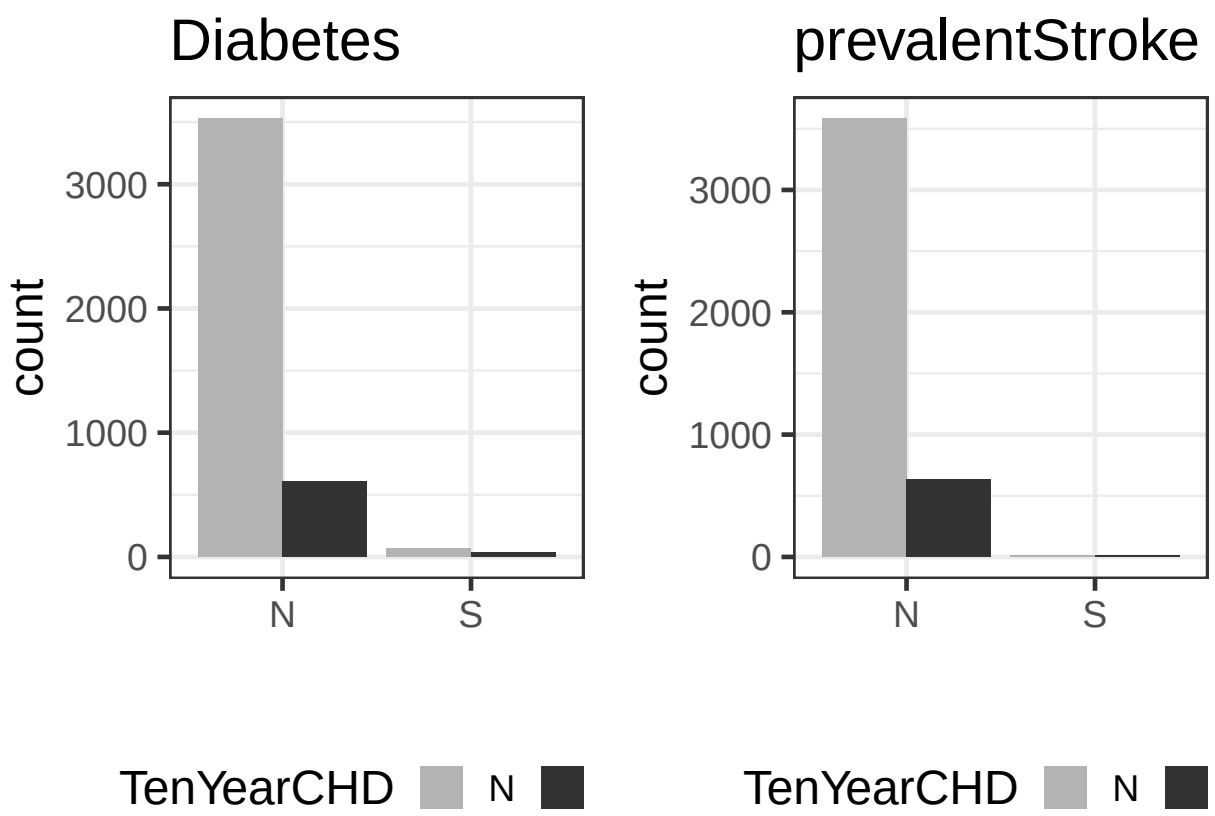

```
a = ggplot(dataset, aes(currentSmoker , fill = TenYearCHD)) +
  geom_bar(stat = "count", position = "dodge") +
  scale_fill_manual(values=c('grey70', 'grey20')) +
  labs(title = "currentSmoker", x = "") +
  theme_bw(base_size = 18) + theme(legend.position="bottom")

b = ggplot(dataset, aes(BPMeds, fill = TenYearCHD)) +
  geom_bar(stat = "count", position = "dodge") +
  scale_fill_manual(values=c('grey70', 'grey20')) +
  labs(title = "BPMeds", x = "") +
  theme_bw(base_size = 18) + theme(legend.position="bottom")

plot_grid(a,b, ncol = 2, nrow = 1)
```




```
a = ggplot(dataset,aes(education, fill = TenYearCHD)) +
  geom_bar(stat = "count", position = "dodge") +
  scale_fill_manual(values=c('grey70', 'grey20')) +
  labs(title = "education", x = "") +
  theme_bw(base_size = 18) + theme(legend.position="bottom")


b = ggplot(dataset, aes(prevalentStroke, fill = TenYearCHD)) +
  geom_bar(stat = "count", position = "dodge") +
  scale_fill_manual(values=c('grey70', 'grey20')) +
  labs(title = "prevalentStroke", x = "") +
  theme_bw(base_size = 18) + theme(legend.position="bottom")

plot_grid(a,b, ncol = 2, nrow = 1)
```



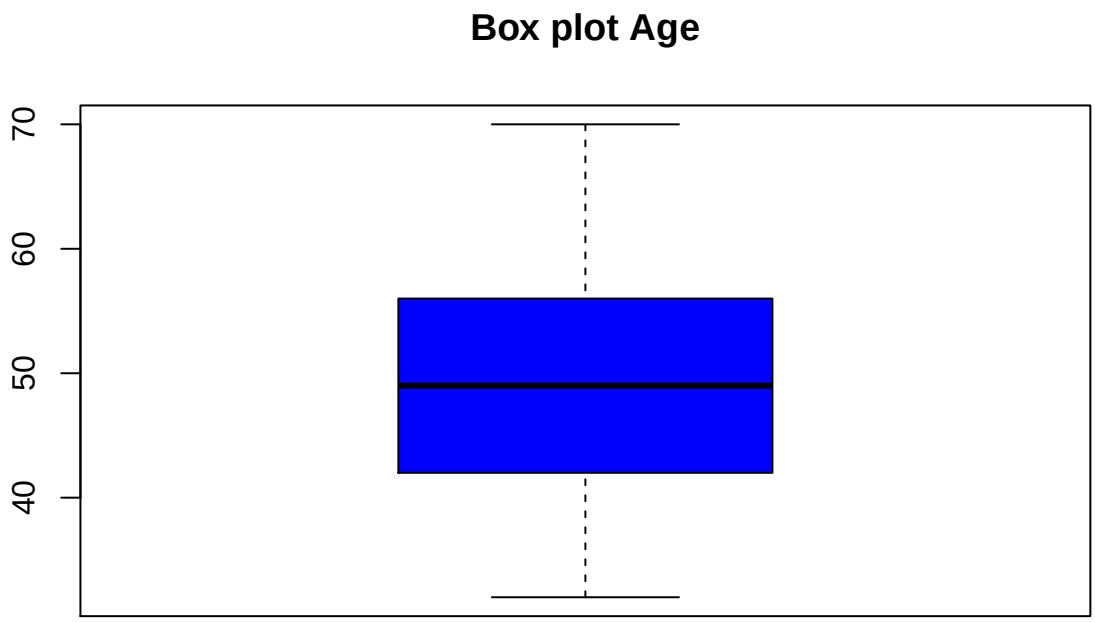

```
a = ggplot(dataset, aes(prevalentHyp, fill = TenYearCHD)) +
  geom_bar(stat = "count", position = "dodge") +
  scale_fill_manual(values=c('grey70', 'grey20')) +
  labs(title = "prevalentHyp", x = "") +
  theme_bw(base_size = 18) + theme(legend.position="bottom")

a = ggplot(dataset, aes(diabetes, fill = TenYearCHD)) +
  geom_bar(stat = "count", position = "dodge") +
  scale_fill_manual(values=c('grey70', 'grey20')) +
  labs(title = "Diabetes", x = "") +
  theme_bw(base_size = 18) + theme(legend.position="bottom")

plot_grid(a,b, ncol = 2, nrow = 1)
```



5. Valores extremos

Analizar la presencia de posibles valores extremos (outliers) en las variables

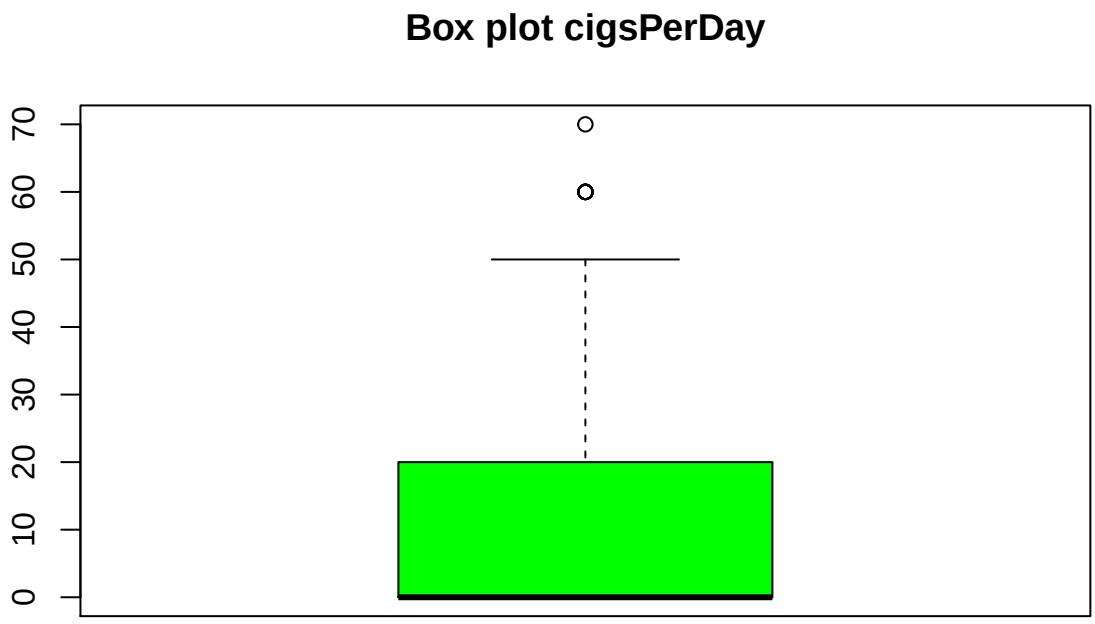
```
boxplot(dataset$age,main="Box plot Age", col="blue")
```




```
boxplot.stats(dataset$age)$out
```

```
## integer(0)
```

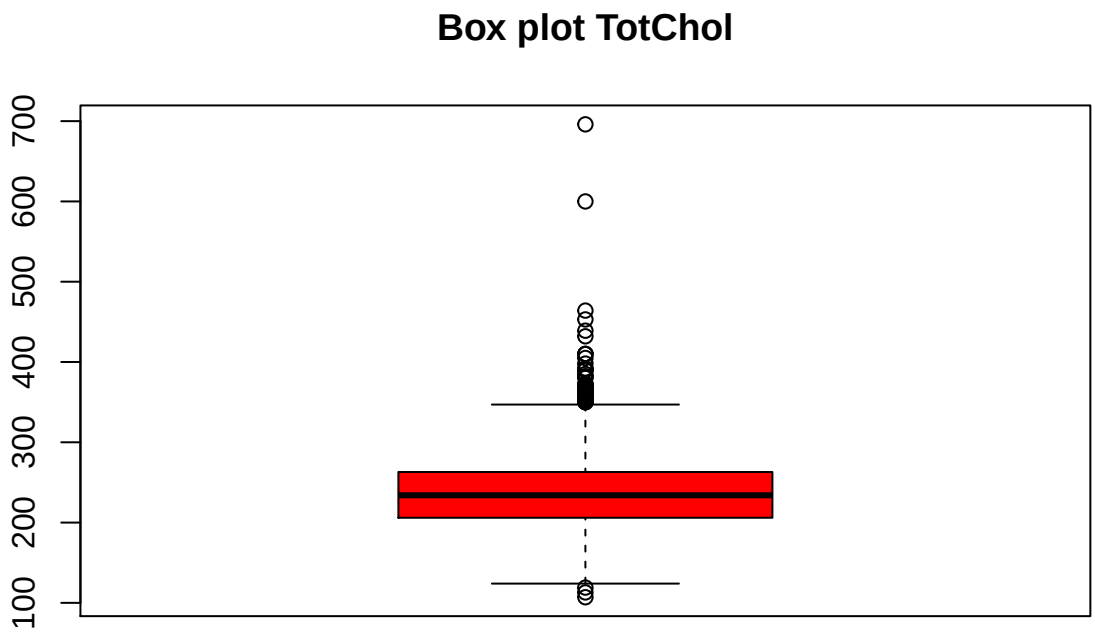
```
boxplot(dataset$cigsPerDay, main="Box plot cigsPerDay", col="green")
```




```
boxplot.stats(dataset$cigsPerDay)$out
```

```
##  [1] 60 60 60 60 60 60 60 60 60 70 60 60
```

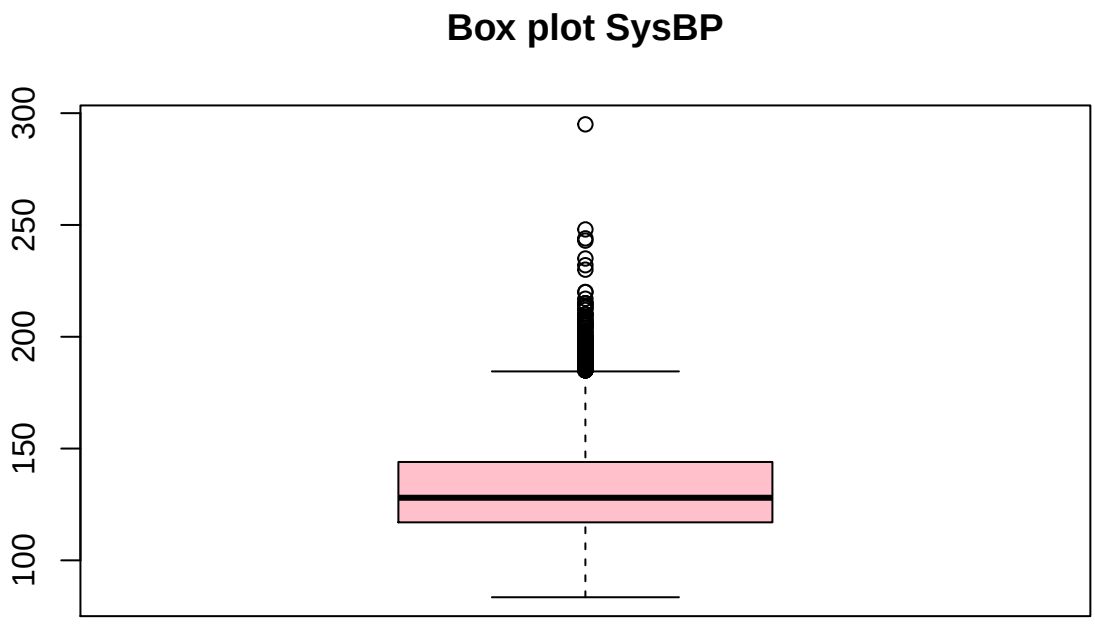
```
boxplot(dataset$totChol,main="Box plot TotChol", col="red")
```




```
boxplot.stats(dataset$totChol)$out
```

```
##  [1] 464 352 368 370 439 398 355 353 360 372 352 600 392 358 391 410 356 107 372
## [20] 366 365 362 410 351 390 405 359 350 380 355 390 371 113 350 354 382 364 367
## [39] 352 432 351 696 363 382 361 453 352 366 410 350 391 358 373 385 366 119
```

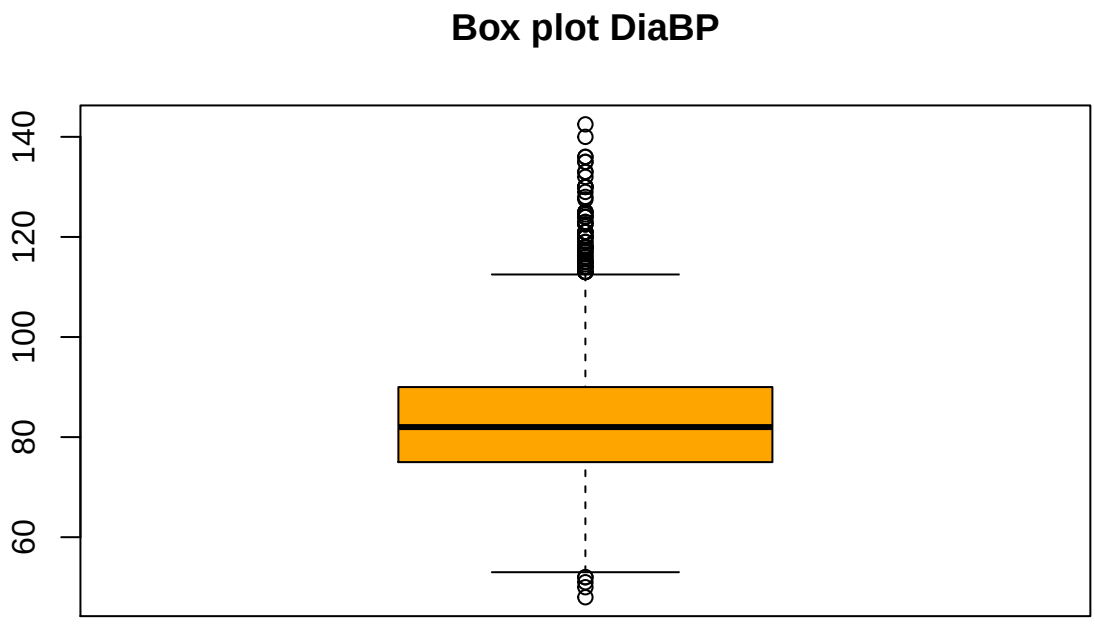
```
boxplot(dataset$sysBP,main="Box plot SysBP", col="pink")
```




```
boxplot.stats(dataset$sysBP)$out
```

```
##   [1] 206.0 190.0 200.0 187.0 212.0 191.0 200.0 189.0 197.5 195.0 189.0 204.0
##  [13] 215.0 188.0 197.0 209.0 295.0 189.0 188.0 185.0 220.0 205.5 186.0 192.0
##  [25] 185.0 195.0 200.0 244.0 213.0 206.0 199.0 198.0 206.0 201.0 189.0 243.0
##  [37] 187.5 185.5 195.0 199.0 186.5 186.0 204.0 217.0 196.0 193.0 187.0 196.0
##  [49] 189.0 196.0 190.0 185.0 202.0 195.0 200.0 232.0 191.0 235.0 188.0 205.0
##  [61] 185.0 220.0 210.0 193.0 188.5 190.0 185.0 192.0 199.0 197.5 190.0 195.0
##  [73] 210.0 202.5 191.5 208.0 191.0 205.0 190.0 210.0 190.0 197.0 198.0 190.0
##  [85] 185.0 204.0 207.5 191.0 195.0 198.0 197.0 186.5 193.0 215.0 196.0 199.5
##  [97] 193.0 195.0 248.0 196.0 202.0 185.0 185.0 230.0 197.0 189.0 214.0 196.0
## [109] 215.0 192.5 188.0 187.0 194.0 207.0 185.5 213.0 192.5 192.5 200.0 187.0
## [121] 190.0 206.0 210.0 195.0 188.0 190.0
```

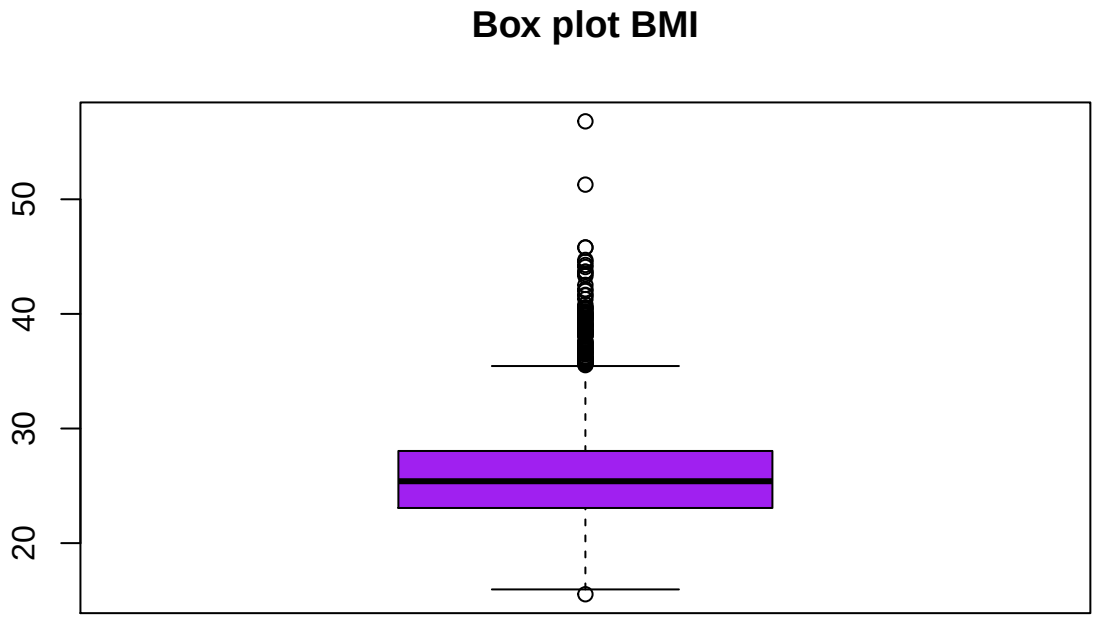
```
boxplot(dataset$diaBP,main="Box plot DiaBP", col="orange")
```




```
boxplot.stats(dataset$diaBP)$out
```

```
##  [1] 121.0 114.0 124.5 122.5 123.0 120.0 118.0 120.0 133.0 135.0 117.0 121.0
## [13] 114.0 118.0 114.5 118.0 140.0 124.0 115.0 115.0 142.5 117.5 116.5 118.0
## [25] 116.0 120.0 119.0 118.0 132.0 124.0 123.0 120.0 114.0  50.0 136.0  51.0
## [37] 120.0 128.0 120.0 115.0 114.0 125.0 130.0 113.0 117.0 136.0 130.0 135.0
## [49] 115.0 114.0 118.0 113.0 121.0 118.0 113.0 129.0 120.0  52.0 130.0 119.0
## [61] 124.0 121.0  52.0  48.0 118.0 113.0 113.0 130.0 122.5 115.5 133.0 115.0
## [73] 125.0 125.0 116.0 127.5 130.0
```

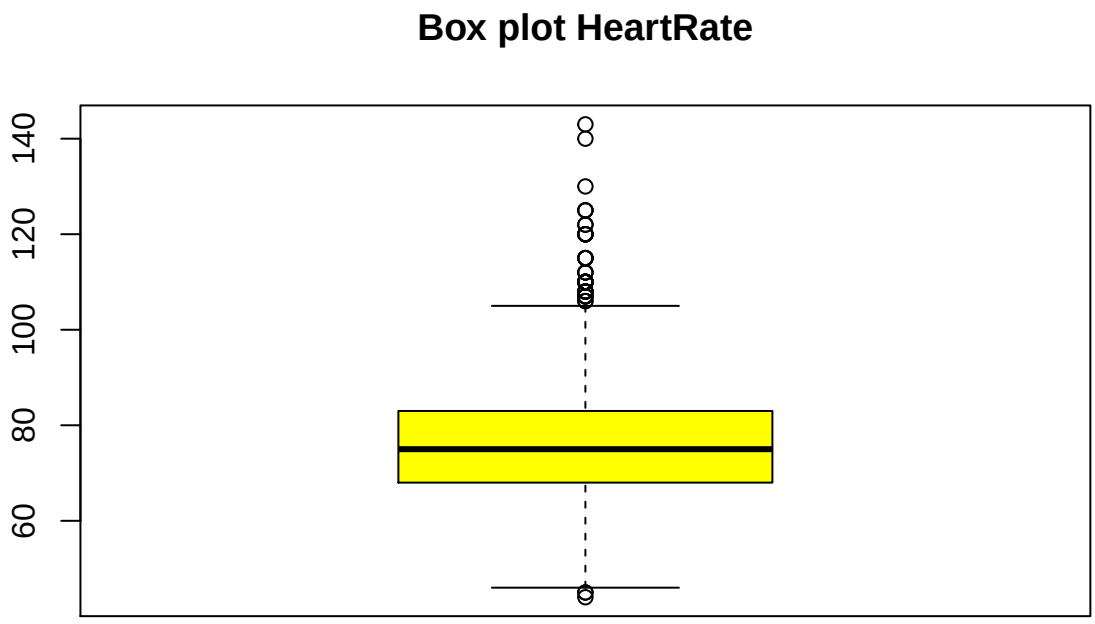
```
boxplot(dataset$BMI,main="Box plot BMI", col="purple")
```




```
boxplot.stats(dataset$BMI)$out
```

```
##  [1] 38.53 40.11 45.80 38.46 40.52 42.15 36.81 38.39 42.00 44.27 36.29 38.14
## [13] 15.54 39.88 36.12 35.58 36.11 45.79 38.82 37.41 36.62 37.48 39.60 44.09
## [25] 40.58 43.30 43.69 42.53 36.21 36.52 38.88 38.38 35.99 35.85 38.75 44.55
## [37] 39.64 36.46 38.06 35.78 39.91 35.85 38.43 36.04 37.04 44.71 38.54 39.04
## [49] 38.42 39.53 39.54 35.62 43.48 36.91 39.08 39.69 36.65 39.82 36.79 37.02
## [61] 35.96 35.53 37.38 36.54 56.80 38.11 40.21 37.15 39.40 40.81 38.61 39.21
## [73] 36.01 40.51 38.31 35.68 37.10 38.96 39.94 39.94 39.22 41.29 40.08 36.12
## [85] 36.18 41.61 37.62 40.38 37.58 51.28 38.94 37.30 41.66 38.17 36.07 39.17
## [97] 43.67
```

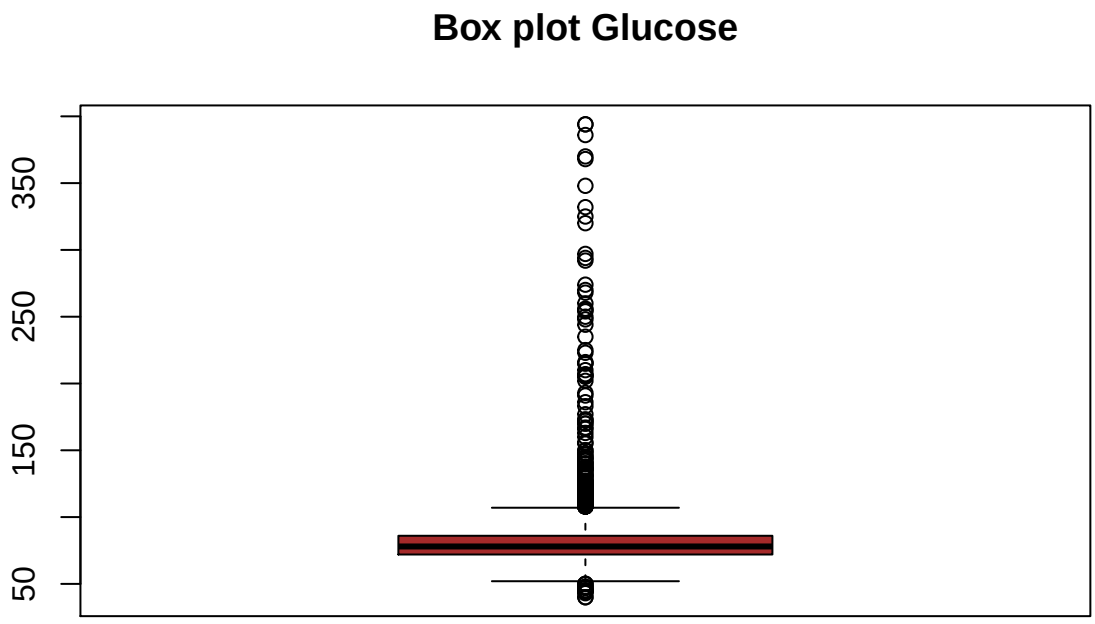
```
boxplot(dataset$heartRate,main="Box plot HeartRate", col="yellow")
```




```
boxplot.stats(dataset$heartRate)$out
```

```
##  [1] 110 110 140 130 108 110 110 108 110 106 110 110 107 108 110 108 110 110 112
## [20] 125 110  44 112 108 110 110 110 110 110  45 110 110 110 110 122 110 110 106
## [39] 110 110 107 107 120 120 108 120 115 110 120 110 143 110 110 120 110 115 115
## [58] 110 107 110 115  45 120 108 110 115 110 122 110 120 108 110 110 125 125 112
```

```
boxplot(dataset$glucose,main="Box plot Glucose", col="brown")
```




```
boxplot.stats(dataset$glucose)$out
```

```
##   [1] 113 225 215  45 202 126 120 117 109 132 150 120 113 135 115 117 113 140
##  [19] 112 118 143 113 114 160 110 117 115 123 108 145 126 118 108 108 117 108
##  [37] 120 122 137 127 205 114 115 113  45 118 130 113 118 112 120  47 108 115
##  [55] 112 216 163 113 113 112 144 116 145 121 172 140 124 112 111  40 126 113
##  [73] 110 186 223 117 325 114 108  44 156 268 120 122 117  50 274 292 111 118
##  [91] 114 112 116 108 114 127 120 115 115 118 255 110 123 136 123 206 127  47
## [109] 131 148 297  50 120 118 132  43 113 173  48 118 126 115 206 140 108 386
## [127] 127 121 155 215 150 112 112 108 147 117 123 110 118 170 115 112 112 108
## [145] 320 132 140 109 108  44 170 137 254 394 394 124 270 244 130 183 115 142
## [163] 108 137  45 117 119 135 167 113  47 135 207 110  45 110 115 129 115 110
## [181] 112 137 115 177 119 108 250 136 113 117 116 294 115 166 123 125 108 332
## [199] 115 109 368 348 122 248 116 110 370 173  40 120 110 117  50 193 191 256
## [217] 235 115 210 118 113 120 116 260
```

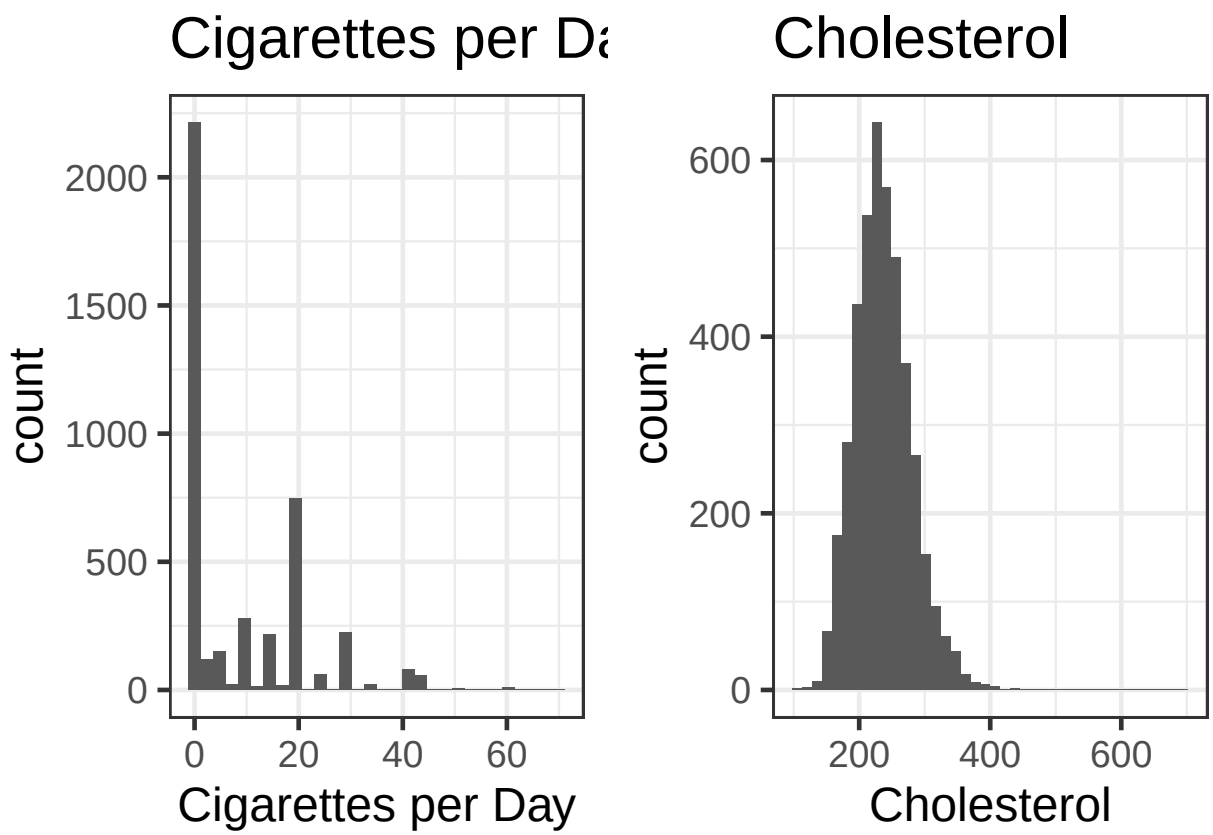
Revisando los outliers de cada variable podemos observar que corresponden a datos validos

6.Comprobación de la normalidad y homogeneidad de la varianza

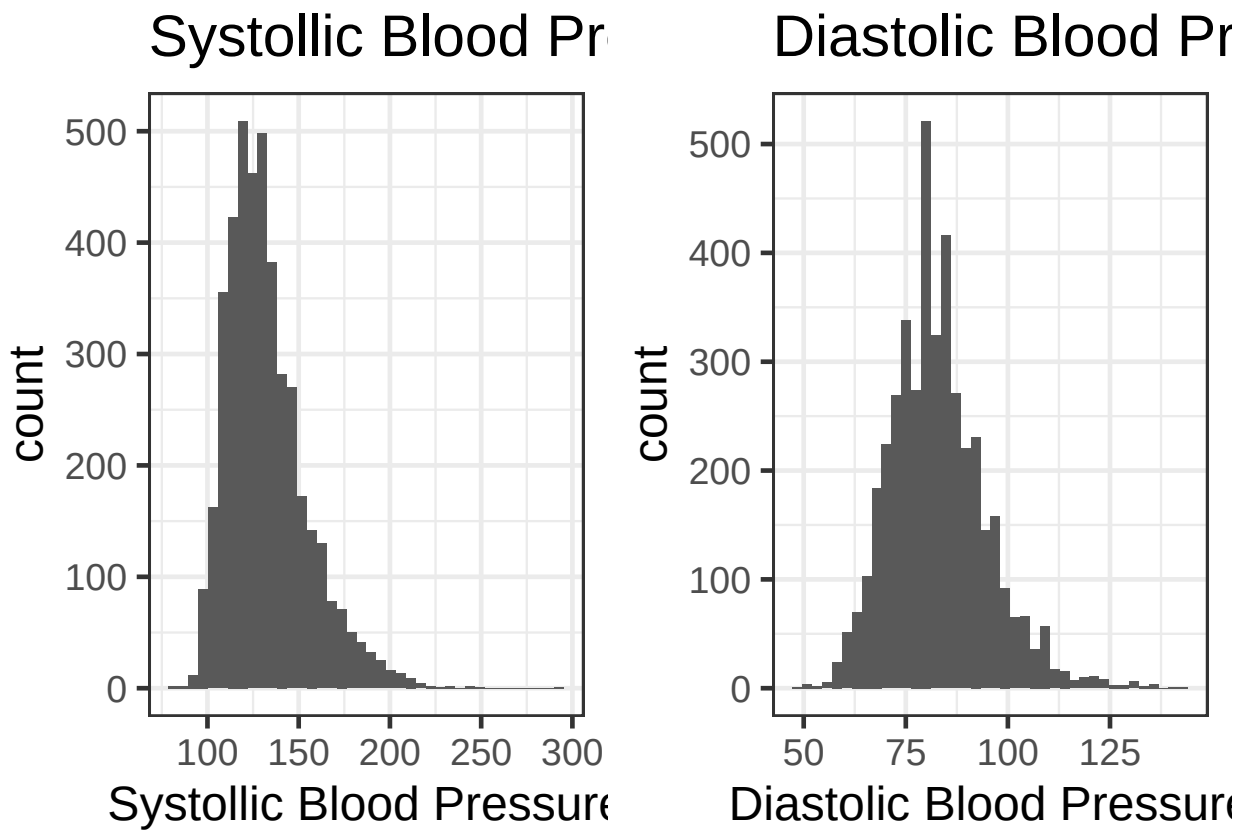
```
a = ggplot(dataset, aes(cigsPerDay)) + geom_histogram() +
  labs(title = "Cigarettes per Day", x = "Cigarettes per Day") +
  theme_bw(base_size = 18)

b = ggplot(dataset, aes(totChol)) + geom_histogram(bins = 40) +
```

```
  labs(title = "Cholesterol", x = "Cholesterol") +
  theme_bw(base_size = 18)


plot_grid(a,b, ncol = 2, nrow = 1)
```

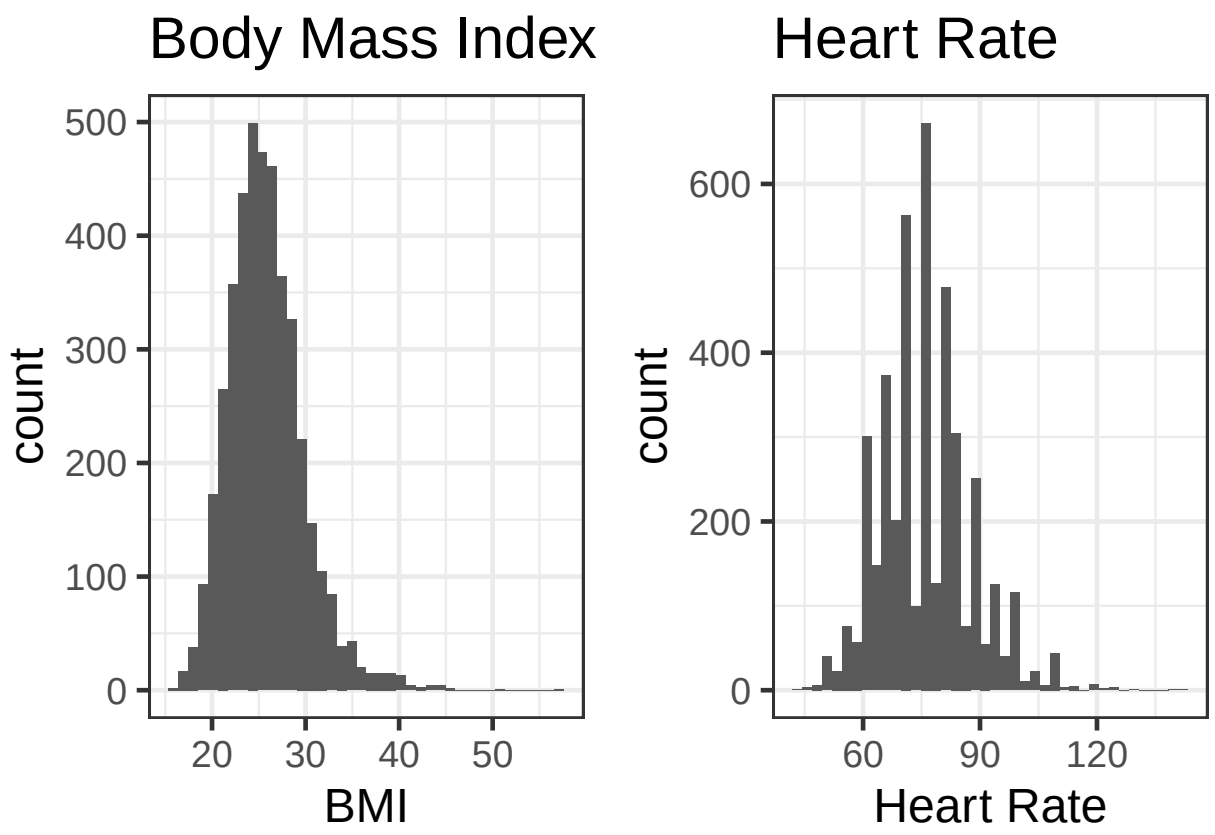
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



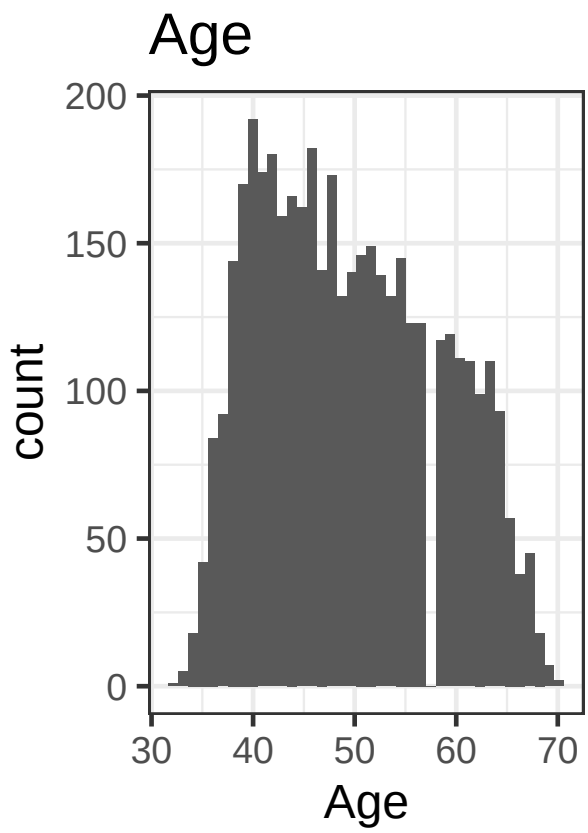

```
a = ggplot(dataset, aes(sysBP)) + geom_histogram(bins = 40) +
  labs(title = "Systollic Blood Pressure", x = "Systollic Blood Pressure") +
  theme_bw(base_size = 18)

b = ggplot(dataset, aes(diaBP)) + geom_histogram(bins = 40) +
  labs(title = "Diastolic Blood Pressure", x = "Diastolic Blood Pressure") +
  theme_bw(base_size = 18)


plot_grid(a,b, ncol = 2, nrow = 1)
```

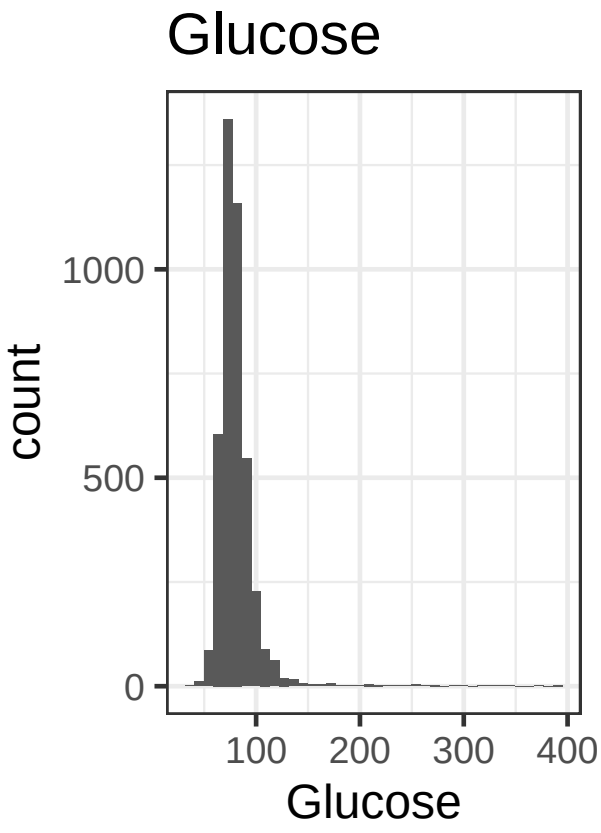



```
a = ggplot(dataset, aes(BMI)) + geom_histogram(bins = 40) +
  labs(title = "Body Mass Index", x = "BMI") +
  theme_bw(base_size = 18)

b = ggplot(dataset, aes(heartRate)) + geom_histogram(bins = 40) +
  labs(title = "Heart Rate", x = "Heart Rate") +
  theme_bw(base_size = 18)

plot_grid(a,b, ncol = 2, nrow = 1)
```




```
a = ggplot(dataset, aes(age)) + geom_histogram(bins = 40) +
  labs(title = "Age", x = "Age") +
  theme_bw(base_size = 18)

b = ggplot(dataset, aes(glucose)) + geom_histogram(bins = 40) +
  labs(title = "Glucose", x = "Glucose") +
  theme_bw(base_size = 18)

plot_grid(a,b, ncol = 2, nrow = 1)
```





Sesgo de la distribución

```
cat("\ncigsPerDay = ", skewness(dataset$cigsPerDay))
```

```
##
## cigsPerDay =  1.23032
```

```
cat("\ntotChol = ", skewness(dataset$totChol))
```

```
##
## totChol =  0.8711289
```

```
cat("\nsysBP = ", skewness(dataset$sysBP))
```

```
##
## sysBP =  1.144475
```

```
cat("\ndiaBP = ", skewness(dataset$diaBP))
```

```
##
## diaBP =  0.7127456
```

```r
cat("\nBMI = ", skewness(dataset$BMI))
```

```
##
## BMI =  0.9828367
```

```r
cat("\nheartRate = ", skewness(dataset$heartRate))
```

```
##
## heartRate =  0.6439241
```

```r
cat("\nage = ", skewness(dataset$age))
```

```
##
## age =  0.2287051
```

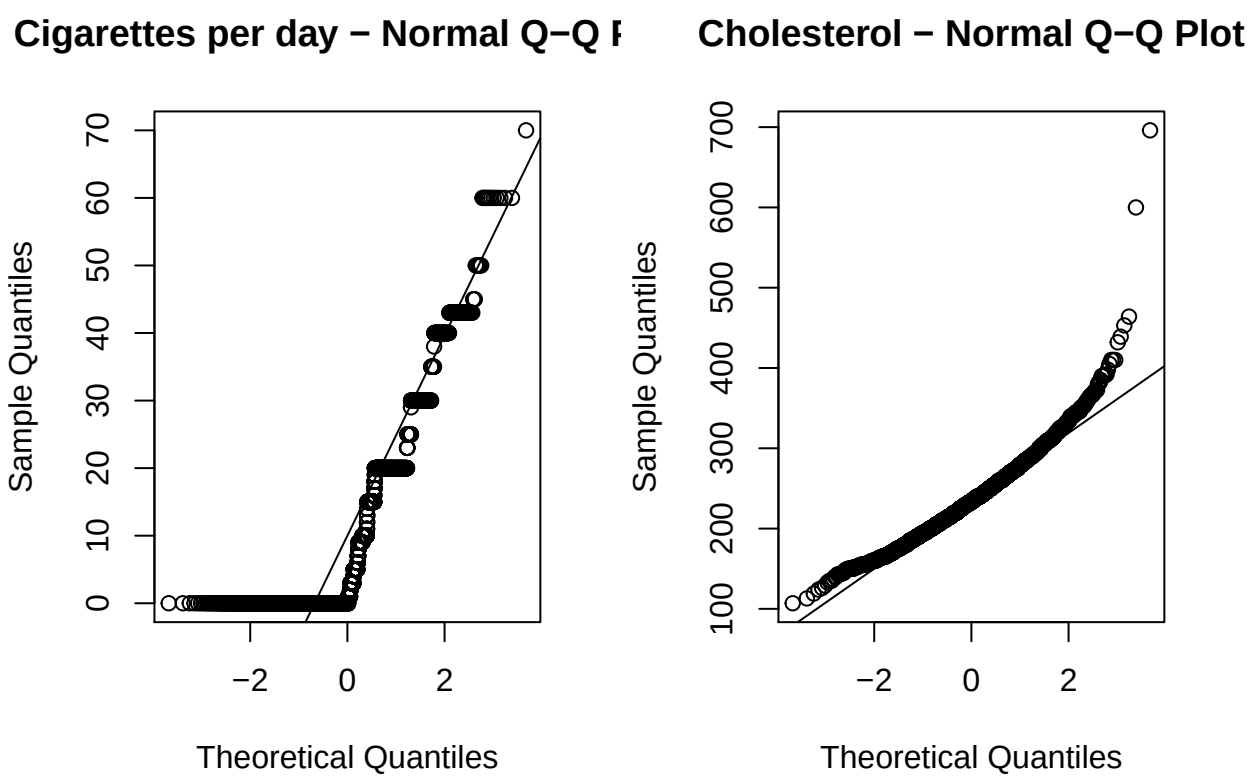
```r
cat("\nglucose = ", skewness(dataset$glucose))
```

```
##
## glucose =  6.459213
```

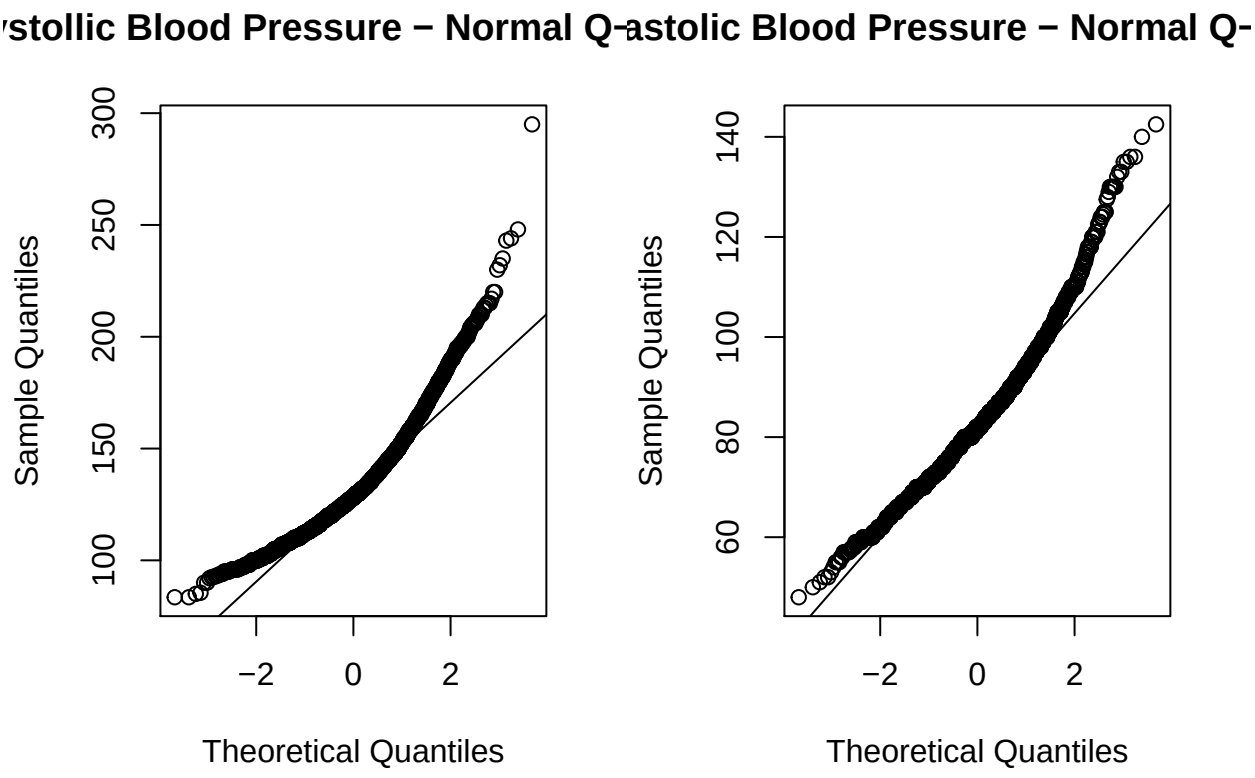
Ligeramente sesgado hacia la derecha (>1): cigsPerDay, sysBP Muy sesgado hacia la derecha: glucose

QQ-Plots de la distribución

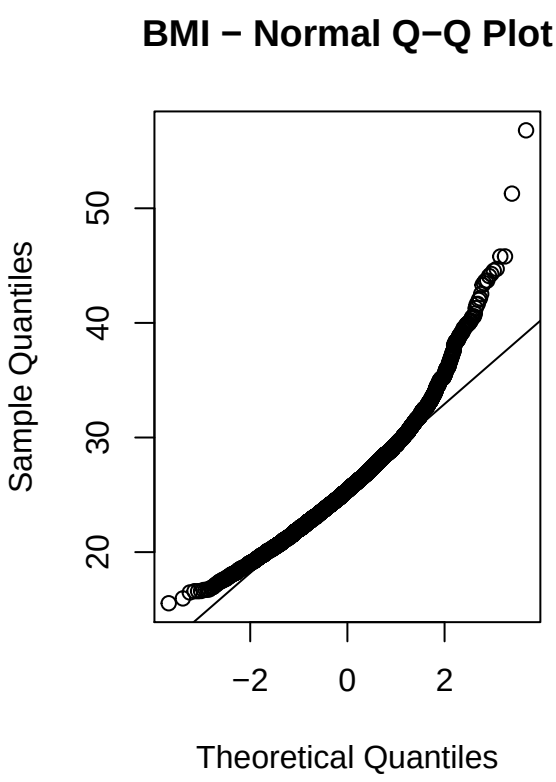
```r
par(mfrow=c(1,2))
qqnorm(dataset$cigsPerDay, main = "Cigarettes per day - Normal Q-Q Plot");
qqline(dataset$cigsPerDay)
qqnorm(dataset$totChol, main = "Cholesterol - Normal Q-Q Plot");
qqline(dataset$totChol)
```

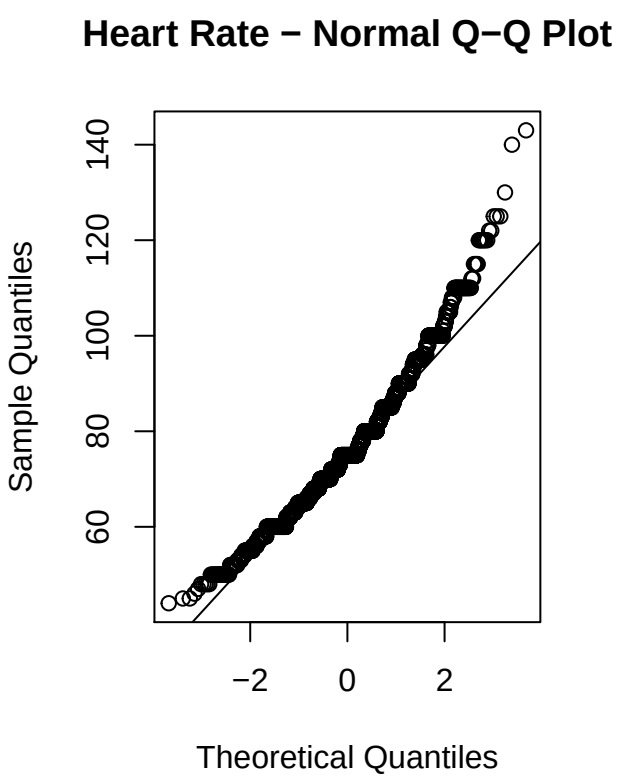



```
par(mfrow=c(1,2))
qqnorm(dataset$sysBP, main = "Systollic Blood Pressure - Normal Q-Q Plot");
qqline(dataset$sysBP)
qqnorm(dataset$diaBP,  main = "Diastolic Blood Pressure - Normal Q-Q Plot");
qqline(dataset$diaBP)
```

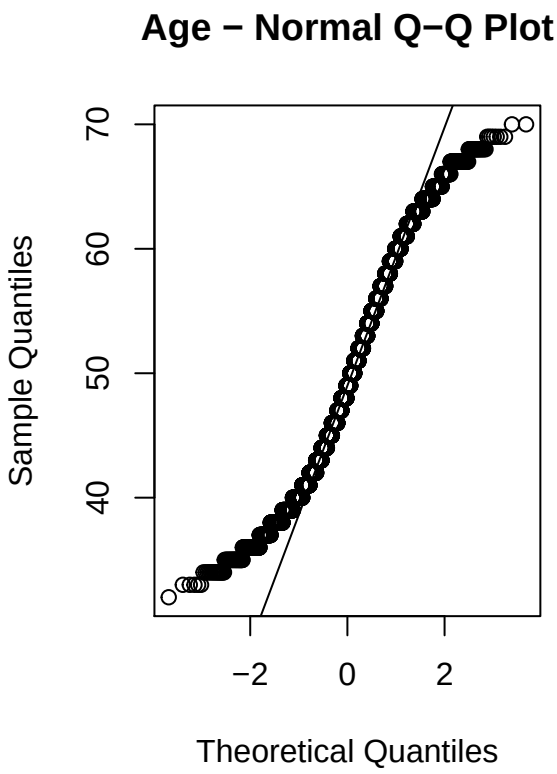


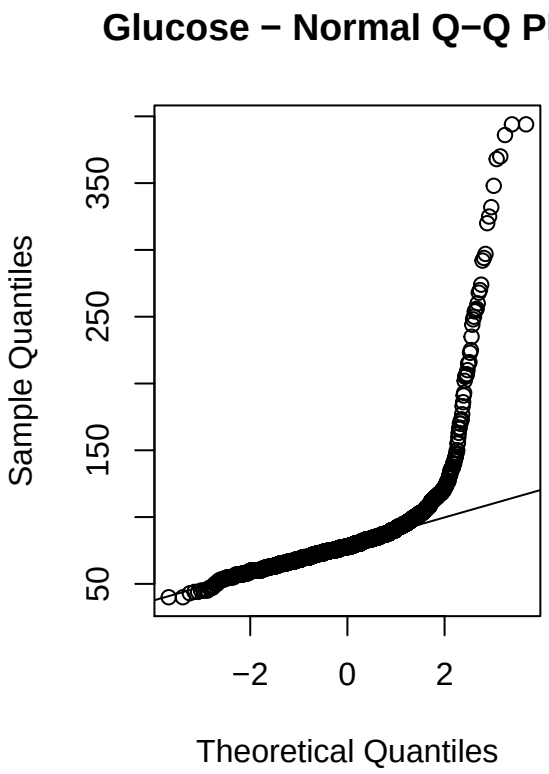

```
par(mfrow=c(1,2))
qqnorm(dataset$BMI, main = "BMI - Normal Q-Q Plot");
qqline(dataset$BMI)
qqnorm(dataset$heartRate, main = "Heart Rate - Normal Q-Q Plot");
qqline(dataset$heartRate)
```






```r
par(mfrow=c(1,2))
qqnorm(dataset$age, main = "Age - Normal Q-Q Plot");
qqline(dataset$age)
qqnorm(dataset$glucose, main = "Glucose - Normal Q-Q Plot");
qqline(dataset$glucose)
```





Prueba de normalidad Shapiro Test

Ho: Los datos estan normalmente distribuidos H1: Los datos no estan normalmente distribuidos

```
cat("\ncigsPerDay p-value =  ", as.numeric(shapiro.test(dataset$cigsPerDay)[2]))
```

```
##
## cigsPerDay p-value =   3.529751e-61
```

```
cat("\ntotChol p-value =  ", as.numeric(shapiro.test(dataset$totChol)[2]))
```

```
##
## totChol p-value =   1.763243e-29
```

```
cat("\nsysBP p-value =  ", as.numeric(shapiro.test(dataset$sysBP)[2]))
```

```
##
## sysBP p-value =   1.580916e-39
```

```
cat("\ndiaBP p-value =  ", as.numeric(shapiro.test(dataset$diaBP)[2]))
```

```
##
## diaBP p-value =   2.996581e-27
```

```r
cat("\nBMI p-value =  ", as.numeric(shapiro.test(dataset$BMI)[2]))
```

```
##
## BMI p-value =   1.663975e-33
```

```r
cat("\nheartRate p-value =  ", as.numeric(shapiro.test(dataset$heartRate)[2]))
```

```
##
## heartRate p-value =   7.913998e-27
```

```r
cat("\nage p-value =  ", as.numeric(shapiro.test(dataset$age)[2]))
```

```
##
## age p-value =   3.633442e-30
```

```r
cat("\nglucose p-value =  ", as.numeric(shapiro.test(dataset$glucose)[2]))
```

```
##
## glucose p-value =   9.114873e-74
```

Como los valores de p-value son menores a alfa (0.05) se rechaza la hipotesis nula (H0) por tanto, Los datos no estan normalmente distribuidos, lo mismo podemos concluir observando las graficas QQplot

7. Modelo de Regresión Logística

```r
logistic_model <- glm(dataset$TenYearCHD~.,family = "binomial", data = dataset)
summary(logistic_model)
```

```
##
## Call:
## glm(formula = dataset$TenYearCHD ~ ., family = "binomial", data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9263  -0.5927  -0.4308  -0.2919   2.8328
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -7.960946   0.658397 -12.091  < 2e-16 ***
## maleM             0.485255   0.101465   4.783 1.73e-06 ***
## age               0.060686   0.006298   9.636  < 2e-16 ***
## education2       -0.170968   0.114512  -1.493 0.135434
## education3       -0.110582   0.139200  -0.794 0.426955
## education4        0.033770   0.151681   0.223 0.823818
## currentSmokerS    0.016817   0.144979   0.116 0.907653
## cigsPerDay        0.021051   0.005734   3.671 0.000241 ***
## BPMedsS           0.252310   0.220372   1.145 0.252239
## prevalentStrokeS  0.971338   0.443501   2.190 0.028513 *
## prevalentHypS     0.231753   0.128655   1.801 0.071646 .
## diabetesS         0.197107   0.295133   0.668 0.504223
```

```
## totChol           0.001845   0.001028   1.795 0.072716 .
## sysBP             0.014153   0.003546   3.991 6.57e-05 ***
## diaBP            -0.002849   0.005984  -0.476 0.634002
## BMI               0.001067   0.011849   0.090 0.928222
## heartRate        -0.001224   0.003885  -0.315 0.752659
## glucose           0.006541   0.002141   3.056 0.002246 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3612.2  on 4239  degrees of freedom
## Residual deviance: 3208.1  on 4222  degrees of freedom
## AIC: 3244.1
##
## Number of Fisher Scoring iterations: 5
```

```
varImp(logistic_model)
```

```
##                     Overall
## maleM            4.78250210
## age              9.63636442
## education2       1.49301139
## education3       0.79441294
## education4       0.22263664
## currentSmokerS   0.11599923
## cigsPerDay       3.67111500
## BPMedsS          1.14492834
## prevalentStrokeS 2.19015818
## prevalentHypS    1.80136055
## diabetesS        0.66785947
## totChol          1.79460714
## sysBP            3.99134068
## diaBP            0.47610096
## BMI              0.09008184
## heartRate        0.31513535
## glucose          3.05558013
```

El modelo es bueno ya que la devianza residual es menor que la devianza nula

las variables de mayor signifcancia en el modelo son: age, maleM, cigsPerDay, prevalentStrokeS, sysBP, glucose

Odds Ratio e intervalo de confianza

```
exp(logistic_model$coefficients)
```

```
##      (Intercept)            maleM              age        education2
##     0.0003488229     1.6245892351     1.0625655864     0.8428483166
##        education3       education4   currentSmokerS       cigsPerDay
##     0.8953129162     1.0343464640     1.0169596924     1.0212738070
##          BPMedsS prevalentStrokeS    prevalentHypS        diabetesS
##     1.2869944388     2.6414756011     1.2608085262     1.2178744573
```

```
##          totChol            sysBP            diaBP              BMI
##     1.0018462850     1.0142540407     0.9971548341     1.0010679791
##        heartRate          glucose
##     0.9987764570     1.0065620520
```

```
exp(confint(logistic_model))
```

```
## Waiting for profiling to be done...
```

```
##                         2.5 %    97.5 %
## (Intercept)      9.503102e-05 0.0012563
## maleM            1.331877e+00 1.9827271
## age              1.049584e+00 1.0758264
## education2       6.723973e-01 1.0536301
## education3       6.787760e-01 1.1720506
## education4       7.644256e-01 1.3863444
## currentSmokerS   7.639379e-01 1.3489209
## cigsPerDay       1.009826e+00 1.0327992
## BPMedsS          8.288179e-01 1.9698023
## prevalentStrokeS 1.090042e+00 6.3062783
## prevalentHypS    9.790520e-01 1.6214673
## diabetesS        6.728392e-01 2.1464468
## totChol          9.998175e-01 1.0038597
## sysBP            1.007230e+00 1.0213371
## diaBP            9.855462e-01 1.0089510
## BMI              9.779545e-01 1.0244744
## heartRate        9.911559e-01 1.0063710
## glucose          1.002397e+00 1.0108633
```

```
confint(logistic_model)
```

```
## Waiting for profiling to be done...
```

```
##                          2.5 %       97.5 %
## (Intercept)      -9.2613071884 -6.679584765
## maleM             0.2865892521  0.684473232
## age               0.0483943059  0.073089102
## education2       -0.3969059567  0.052241474
## education3       -0.3874641687  0.158754887
## education4       -0.2686306264  0.326670346
## currentSmokerS   -0.2692687227  0.299304919
## cigsPerDay        0.0097778112  0.032272739
## BPMedsS          -0.1877548553  0.677933166
## prevalentStrokeS  0.0862163131  1.841545698
## prevalentHypS    -0.0211705031  0.483331500
## diabetesS        -0.3962489061  0.763813845
## totChol          -0.0001824672  0.003852262
## sysBP             0.0072040074  0.021112684
## diaBP            -0.0145593044  0.008911202
## BMI              -0.0222921303  0.024179676
## heartRate        -0.0088834215  0.006350762
## glucose           0.0023937094  0.010804690
```

Segun los Odds ratio la probabilidad de sufrir una enfermedad coronaria en los proximos 10 años aumenta si se es hombre,si se toma medicamentos para la tensión, si el paciente había tenido previamente un accidente cerebrovascular o ha sido o es hipertenso y Si el paciente tenía diabetes

Bondad del ajuste

Usa el test de Hosman-Lemeshow para ver la bondad de ajuste del modelo final

```
hl <- hoslem.test(logistic_model$y, fitted(logistic_model), g=10)
hl
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  logistic_model$y, fitted(logistic_model)
## X-squared = 8.1845, df = 8, p-value = 0.4157
```

```
cbind(hl$expected,hl$observed)
```

```
##                     yhat0     yhat1  y0  y1
## [0.0169,0.0417] 410.2043  13.79571 415   9
## (0.0417,0.0574] 402.9422  21.05776 400  24
## (0.0574,0.0756] 395.7716  28.22843 392  32
## (0.0756,0.0947] 387.8565  36.14353 390  34
## (0.0947,0.119]  379.0970  44.90305 379  45
## (0.119,0.146]   368.3886  55.61141 370  54
## (0.146,0.182]   355.3203  68.67974 361  63
## (0.182,0.23]    338.0840  85.91605 324 100
## (0.23,0.31]     310.6052 113.39477 305 119
## (0.31,0.925]    247.7304 176.26956 260 164
```

De acuerdo al p-value el cual es mayor a alfa (0.05),indica que no hay evidencia de mal ajuste. El modelo está correctamente especificado. Implica que lo que observamos se ajusta suficientemente a lo que esperado bajo el modelo. Hay mucha proximidad entre estos valores reales y teóricos. Esto es lo que permite pensar que usar este modelo y calcular predicciones con él es suficientemente correcto

Evaluación del modelo

Likelihood ratio

```
# Diferencia de residuos
dif_residuos <- logistic_model$null.deviance - logistic_model$deviance

# Grados libertad
df <- logistic_model$df.null - logistic_model$df.residual

# p-value
p_value <- pchisq(q = dif_residuos,df = df, lower.tail = FALSE)

paste("Diferencia de residuos:", round(dif_residuos, 4))
```

```
## [1] "Diferencia de residuos: 404.1422"
```

```
paste("Grados de libertad:", df)
```

```
## [1] "Grados de libertad: 17"
```

```
paste("p-value:", p_value)
```

```
## [1] "p-value: 2.52388365869113e-75"
```

El modelo en conjunto sí es significativo

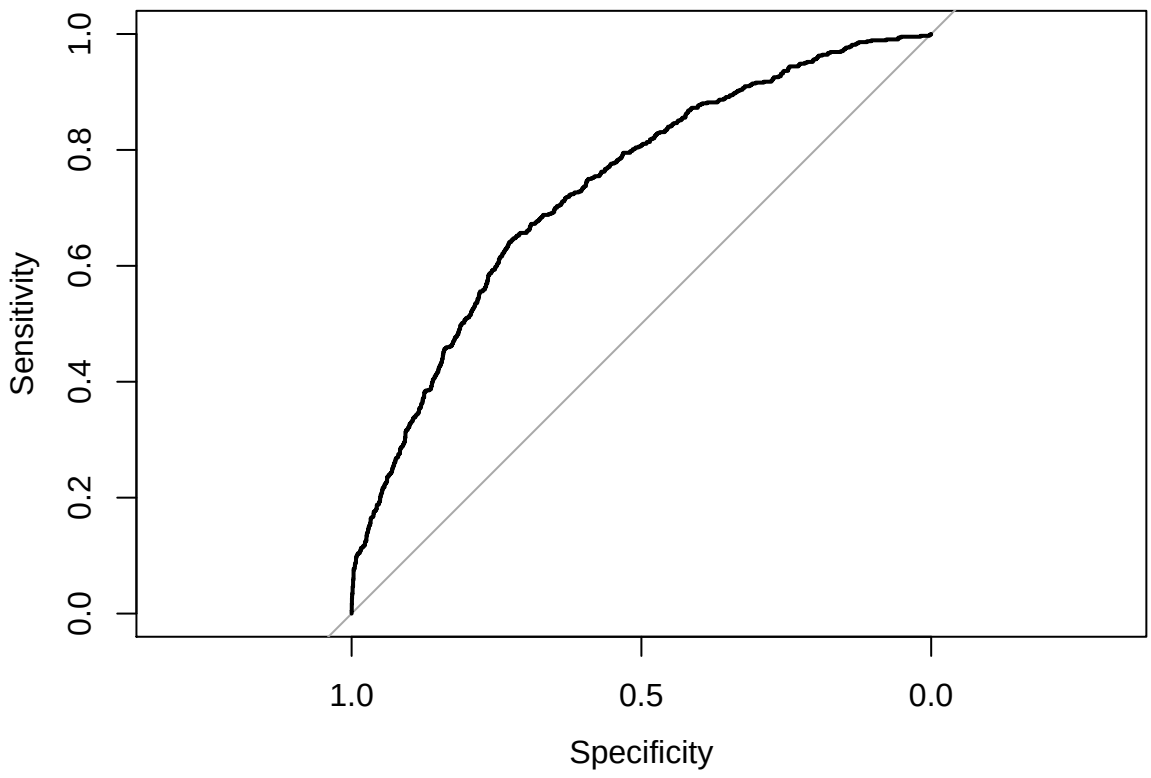
Curva ROC

```
predpr <- predict(logistic_model,type=c("response"))
roccurve <- roc(dataset$TenYearCHD ~ predpr)
```

```
## Setting levels: control = N, case = S
```

```
## Setting direction: controls < cases
```

```
plot(roccurve)
```



```r
auc(roccurve)
```

```
## Area under the curve: 0.7327
```

El área por debajo de esa curva toma el valor de 0.73, por lo que la habilidad del modelo para discriminar es relativamente buena

Matriz de confusión

Para este estudio se va a emplear un threshold de 0.5

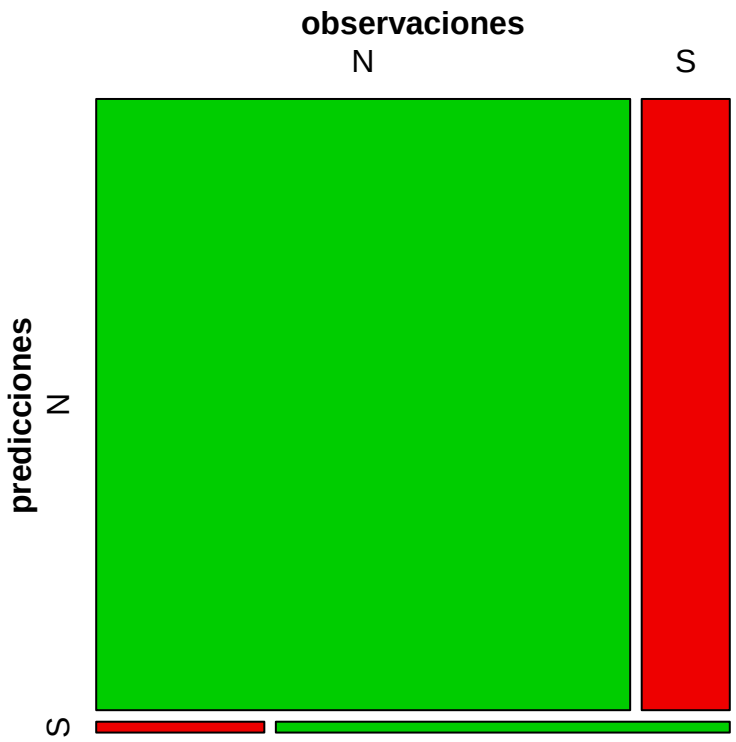
```r
pr1 <- ifelse(predict(logistic_model, type = "response") > 0.5, "S", "N")
t = table(predicciones = pr1, observaciones = dataset$TenYearCHD)
t
```

```
##             observaciones
## predicciones    N    S
##            N 3576  590
##            S   20   54
```

```r
cm = confusionMatrix(t, positive = "S")
cm
```

```
## Confusion Matrix and Statistics
##
##             observaciones
## predicciones    N    S
##            N 3576  590
##            S   20   54
##
##                Accuracy : 0.8561
##                  95% CI : (0.8452, 0.8666)
##     No Information Rate : 0.8481
##     P-Value [Acc > NIR] : 0.0751
##
##                   Kappa : 0.123
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.08385
##             Specificity : 0.99444
##          Pos Pred Value : 0.72973
##          Neg Pred Value : 0.85838
##              Prevalence : 0.15189
##          Detection Rate : 0.01274
##    Detection Prevalence : 0.01745
##       Balanced Accuracy : 0.53914
##
##        'Positive' Class : S
##
```

```
mosaic(t, shade = T, colorize = T,gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
```




```
cat("\nAccuracy del modelo =  ", round(cm$overall["Accuracy"],3));
```

```
##
## Accuracy del modelo =   0.856
```

```
cat("\nSensitivity del modelo = ", round(cm$byClass["Sensitivity"], 3));
```

```
##
## Sensitivity del modelo =  0.084
```

```
cat("\nSpecificity del modelo = ", round(cm$byClass["Specificity"],3))
```

```
##
## Specificity del modelo =  0.994
```

tasa de error de clasificación

```
class_err = function(actual, predicted) {
  mean(actual != predicted)}
```

```r
ac1 = mean(pr1 == dataset$TenYearCHD)
err1 <- class_err(actual = dataset$TenYearCHD, predicted = pr1)

cat("Accuracy del Modelo = ", ac1)
```

```
## Accuracy del Modelo =  0.8561321
```

```r
cat("\nError del Modelo = ", err1)
```

```
##
## Error del Modelo =  0.1438679
```

```r
cat("\nAccuracy + Error = ", err1 + ac1)
```

```
##
## Accuracy + Error =  1
```

8.Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Elegimos regresión logística ya que el atributo objetivo es categóricos.

El modelo puede predecir valores correctos en un 85.6% utilizando regresión logística binomial.

La Sensibilidad del modelo es de 0.084, Tasa de Verdaderos Positivos (True Positive Rate) (TP). Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

La Especificidad del modelo es de 0.994, Tasa de Verdaderos Negativos, ("true negative rate" o TN). Se trata de los casos negativos que el algoritmo ha clasificado correctamente.

El modelo es bueno al predecir a las personas que no sufriran enfermedades coronarias per no es muy bueno al predecir a las personas que si las sufriran.

El código en R esta incluido en este fichero con extensión rmd y tambien se puede descargar en GitHub

9. Archivo CSV con datos finales analizados

```r
write_excel_csv2(dataset, "datasets_framingham_Processed.csv")
```