Resolución PRA 2

Alexiss Mejía Borja - Julio Morales

9 de enero, 2023

Contents

1. Descripción

En esta actividad se elabora un caso práctico, consistente en el tratamiento de un conjunto de datos (en inglés, dataset), 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.

1.2. Objetivos

2. Resolución

2.1 Descripción del dataset

2.3 Limpieza de los datos

Para comenzar con el proceso de limpieza de los datos, procedemos a realizar la lectura del fichero en formato CSV. Esto tiene con finalidad cargar el archivo de datos y examinar el tipo de datos con los que R ha interpretado cada variable. Adocionalmente se va a examinar los valores resumen de cada tipo de variable.

Cargar el Archivo

```
# Lectura del archivo seleccionado en  formato csv
healthcare<- read_delim("C:/MAESTRIA/TERCER_SEMESTRE/TIPOLOGIA_CICLO_VIDA_DATOS/PRA2/healthcare-dataset-
                        delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 5110 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (5): gender, ever_married, work_type, Residence_type, smoking_status
## dbl (7): id, age, hypertension, heart_disease, avg_glucose_level, bmi, stroke
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(healthcare,n=4L)
```

```
## # A tibble: 4 x 12
##      id gender   age hypertension heart_disease ever_married work_type
##   <dbl> <chr>  <dbl>        <dbl>         <dbl> <chr>        <chr>
## 1  9046 Male      67            0             1 Yes          Private
## 2 51676 Female    61            0             0 Yes          Self-employed
## 3 31112 Male      80            0             1 Yes          Private
## 4 60182 Female    49            0             0 Yes          Private
## # ... with 5 more variables: Residence_type <chr>, avg_glucose_level <dbl>,
## #   bmi <dbl>, smoking_status <chr>, stroke <dbl>
```

Una vez cargado el dataset es necesario verificar la tipología de los atributos, asi como la verificación de la existencia de posibles valores perdidos y de valores atípicos

Tipos de datos

```
# Tipo de datos asigando para cada atributo
str(healthcare)
```

```
## spec_tbl_df [5,110 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ id               : num [1:5110] 9046 51676 31112 60182 1665 ...
##  $ gender           : chr [1:5110] "Male" "Female" "Male" "Female" ...
##  $ age              : num [1:5110] 67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : num [1:5110] 0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : num [1:5110] 1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : chr [1:5110] "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr [1:5110] "Private" "Self-employed" "Private" "Private" ...
##  $ Residence_type   : chr [1:5110] "Urban" "Rural" "Rural" "Urban" ...
##  $ avg_glucose_level: num [1:5110] 229 202 106 171 174 ...
##  $ bmi              : num [1:5110] 36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...
##  $ smoking_status   : chr [1:5110] "formerly smoked" "never smoked" "never smoked" "smokes" ...
##  $ stroke           : num [1:5110] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
```

```
##   .. cols(
##   ..   id = col_double(),
##   ..   gender = col_character(),
##   ..   age = col_double(),
##   ..   hypertension = col_double(),
##   ..   heart_disease = col_double(),
##   ..   ever_married = col_character(),
##   ..   work_type = col_character(),
##   ..   Residence_type = col_character(),
##   ..   avg_glucose_level = col_double(),
##   ..   bmi = col_double(),
##   ..   smoking_status = col_character(),
##   ..   stroke = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Se puede notar la existencia de variables de tipo numericas como de tipo categóricas, adicionalmente se puede notar la existencia de valores atípicos en el atriburto denominado bmi (índice de masa coporal), para verificar la existencia de más valores perdidos se ejecuta el siguiente proceso:

2.3.1. Normalización de los datos cualitativos

```
# Para el caso del genero
table(healthcare$gender)
```

```
##
## Female   Male  Other
##   2994   2115      1
```

```
# Para el caso de si alguna vez estuvo casado
table(healthcare$ever_married)
```

```
##
##   No  Yes
## 1757 3353
```

```
# Para el caso del tipo de trabajo
table(healthcare$work_type)
```

```
##
##      children      Govt_job  Never_worked       Private Self-employed
##           687           657            22          2925           819
```

```
# Para el caso del tipo de residencia
table(healthcare$Residence_type)
```

```
##
## Rural Urban
##  2514  2596
```

```r
# Para el caso de si alguna vez a fumado
table(healthcare$smoking_status)
```

```
##
## formerly smoked    never smoked          smokes         Unknown
##             885            1892             789            1544
```

No se puede ver valores que ameriten una normalización de los datos cualitativos del dataset por lo que ha optado por conservar la estructura inicial del mismo.

2.3.2. Normalización de los datos cuantitativos

```r
# Verificación de la edad

summary(healthcare$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.08   25.00   45.00   43.23   61.00   82.00
```

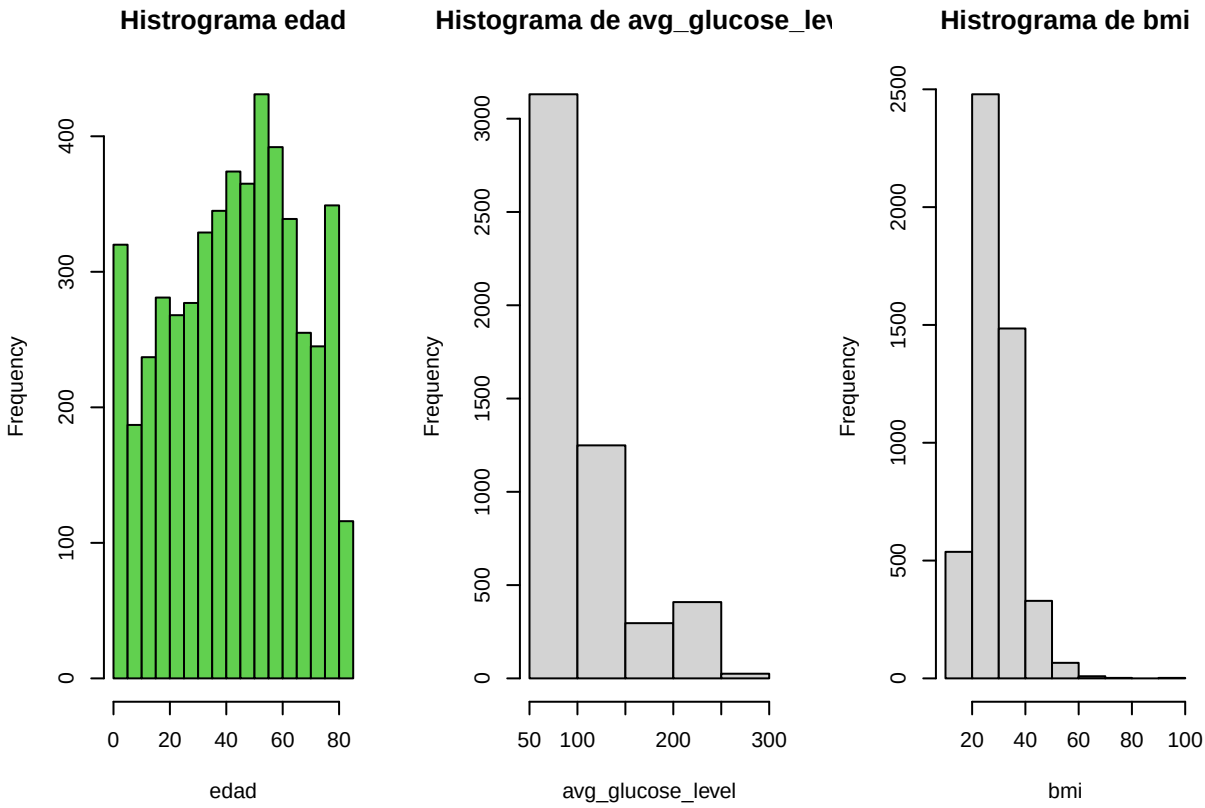
```r
# Verificación de la avg_glucose_level

summary(healthcare$avg_glucose_level)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.12   77.25   91.89  106.15  114.09  271.74
```

```r
# Verificación del bmi

summary(healthcare$bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.30   23.50   28.10   28.89   33.10   97.60     201
```

```r
# Histogramas de la edad, glucosa y bmi

par(mfrow=c(1,3))

#hist(healthcare$age, main = "Histograma de la edad",
 #    xlab = "Edad",breaks = 30)

hist(healthcare$age,main="Histrograma edad",xlab="edad",col=3,breaks="Sturges")

hist(healthcare$avg_glucose_level, main = "Histograma de avg_glucose_level",
     xlab = "avg_glucose_level", breaks = 7)

hist(healthcare$bmi, main = "Histrograma de bmi",
     xlab = "bmi", breaks = 10)
```



Un hecho particular sucede con la edad ya para ser un dataset relacionado los ataques cardiacos la distribución de la edad refleja un grupo considerable de individuos que se encuentran por debajo de los 10 años de edad incluso infantes tal y como se puede ver el summary de este atributo, cabe aquí la cuestión de si es necesario normalizar la edad de los individuos o simplemente realizar una selección de datos que son de mayor interés y lógica para el presente análisis.

En el caso de la vaiable relacionada con el promedio del nivel de la glucosa tanto el resúmen como los graficos no muestra indicios de que requieran un tratamiento de normalización, sin embargo como se advirtio en la anterior sección el bmi presenta valores perdidos, hecho que lo tratara en la siguiente sección.

2.3.3. Selección de datos de interés (reducción de cantidad)

Como se vio en la sección anterior existen registros con edades inferiores a 18 años incluso existen registros de personas que no llegan a cumplir el año de vida, este es un hecho que llama la atención, pues sería necesatio verificar la edad con otra variable de control, por ejemplo el tipo de trabajo, ya que la lógica nos indica que las personas menores de edad no debería encontrarse trabajando.

```
# se selecciona de acuerdo a una edad <18 años y que en la variable work_type  sea distinto
# a children  y Never_worked
seleccio_edad<- subset(healthcare, age < 18 & work_type !="children" &
                         work_type !="Never_worked")

# se verifica  las caracteristcias de menores de edad
seleccio_edad%>% count (work_type, age)
```

```
## # A tibble: 16 x 3
```

```
##    work_type       age     n
##    <chr>         <dbl> <int>
##  1 Govt_job         14     1
##  2 Govt_job         15     1
##  3 Govt_job         16     1
##  4 Govt_job         17     3
##  5 Private           8     1
##  6 Private          13     9
##  7 Private          14    19
##  8 Private          15    21
##  9 Private          16    35
## 10 Private          17    49
## 11 Self-employed     7     1
## 12 Self-employed    13     1
## 13 Self-employed    14     4
## 14 Self-employed    15     2
## 15 Self-employed    16     1
## 16 Self-employed    17     3
```

Se puede notar como existen incluso niños con 8 y 7 años de edad que trabajan el en sector privado o que son auto empleados, hecho que no tiene mucha coherencia, en el mayor de los casos existen 49 adolescentes de 17 años que trabajan en el sector privado; con estos antecedentes se ha decidido excluir los casos con una edad menor a los 15 años y que el tipo de trabajo sea distinto y a children y Never_worked

```
# se excluyen a los casos  de acuerdo a una edad menor a 15 años y que en la variable work_type
#sea distinto a children  y Never_worked
healthcare$filtro<-ifelse(healthcare$age<15 & (healthcare$work_type !="children"
                                               &  healthcare$work_type !="Never_worked" ),1,0)

healthcare<- subset(healthcare, filtro==0)
```

El dataset final consta de 5074 observaciones sobre las cuales se ejecutaran el resto de proceso y análisis.

2.3.4. Detección de Valores perdidos

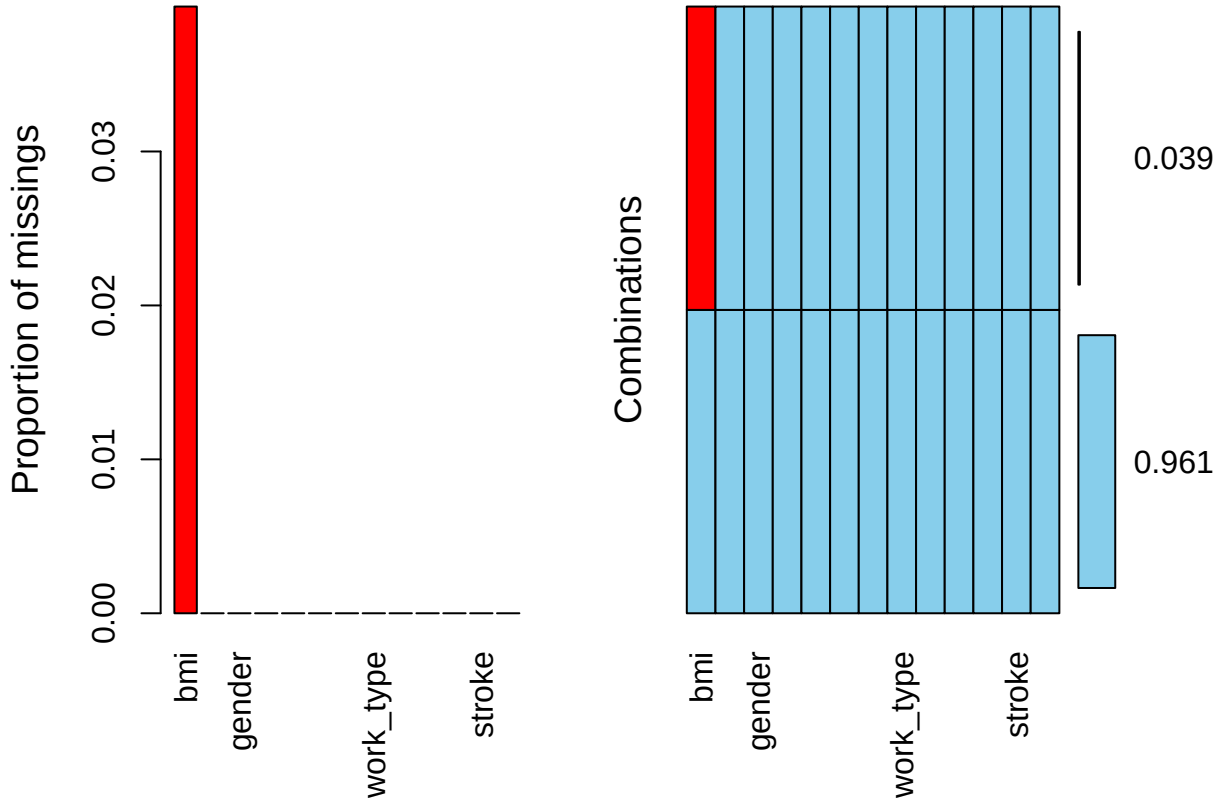
```
# Valores perdidos del dataset  healthcare
apply(is.na(healthcare), 2, sum)
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    Residence_type
##                 0                 0                 0                 0
## avg_glucose_level               bmi    smoking_status            stroke
##                 0               200                 0                 0
##            filtro
##                 0
```

Tal y como se sospechaba existen datos perdidos en la variable relacionada con el índice de masa corporal (200 casos perdiddos), para este caso y al considerar un atributo importante se ha optado por realizar imputaciones mediante la técnica knn con el fin de recuperar la información perdida.

De igual forma de manera grafica podemos ver la distribución de los valores perdidos en el dataset del ejercicio

```
# Valores perdidos del dataset  healthcare
aggr(healthcare,numbers=T,sortVar=T)
```




```
##
##  Variables sorted by number of missings:
##           Variable      Count
##                bmi 0.03941663
##                 id 0.00000000
##             gender 0.00000000
##                age 0.00000000
##       hypertension 0.00000000
##      heart_disease 0.00000000
##       ever_married 0.00000000
##          work_type 0.00000000
##     Residence_type 0.00000000
##  avg_glucose_level 0.00000000
##     smoking_status 0.00000000
##             stroke 0.00000000
##             filtro 0.00000000
```

Aunque tanto los graficos como la estructura del dataset muestra que la variable bmi presenta valores perdidos el atributo smoking_status presenta una categoria denominada como Unknown que tambien podria ser considerada como valor perdido.

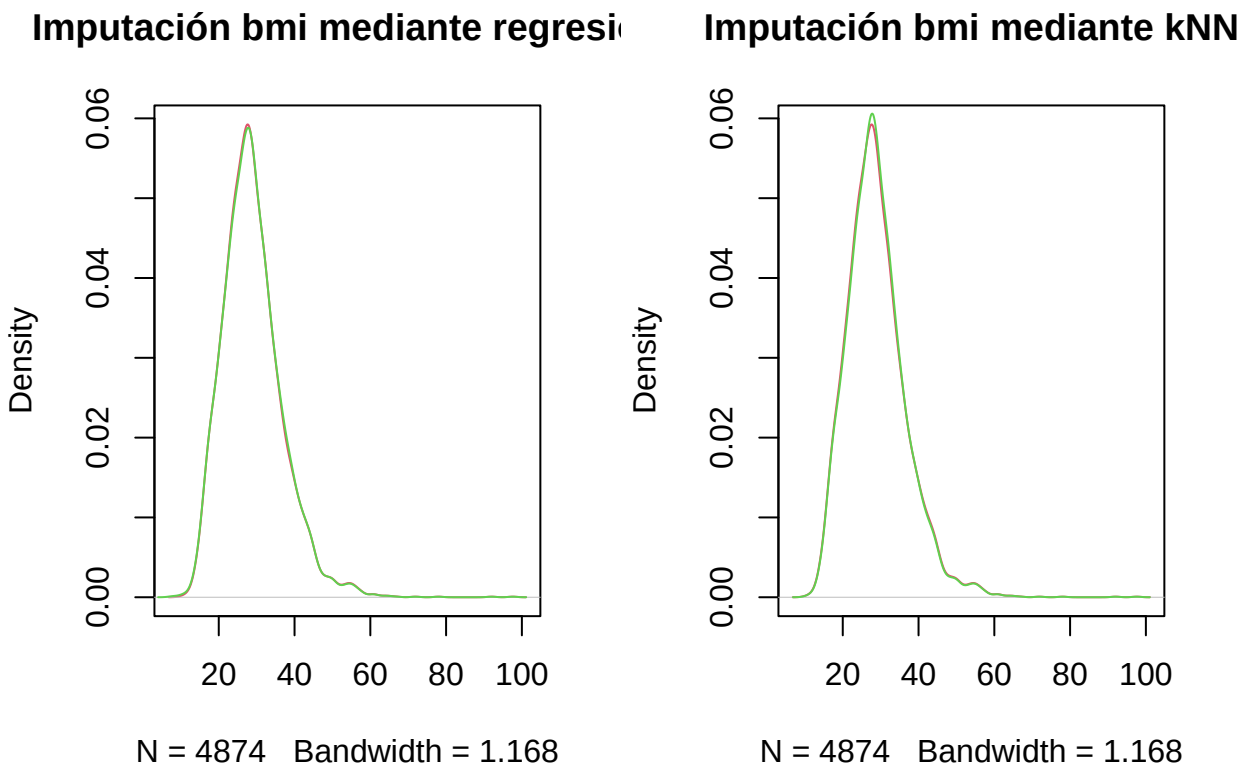
2.3.5. Imputación de los Valores perdidos

Para la imputación de los valores perdidos del índice de masa corporal se va a usar dos métodos y elegir el mejor. El primero de elloses la imputación por kNN, y el segundo mediante una Regresión Estocastica

```
# Imputación de los valores perdidos del bmi mediante kNN

#Se extrae las variables de horas, genero y ClaimNumber en otro dataframe
gender<-(healthcare[, "gender"])
age<-(healthcare[,"age"])
hypertension<-(healthcare[,"hypertension"])
bmi<-(healthcare[,"bmi"])
id<-(healthcare[,"id"])

df<- data.frame(id,bmi,hypertension,age,gender)

df_imput<-kNN(df,k=4)
```

```
# Imputación de los valores perdidos del bmi mediante la regresión


columns <- c("bmi","id")
imputed_data1 <- mice(healthcare[,names(healthcare) %in% columns],m = 1,
                      maxit = 1, method = "norm.nob",seed = 2018,print=F)
complete.data1 <- mice::complete(imputed_data1)
```

```
# Gráficos de densidad de la variable bmi original vs los dos metodos de imputación


par(mfrow=c(1,2))
plot(density(healthcare$bmi,na.rm = T),col=2,main="Imputación bmi mediante regresión")
lines(density(complete.data1$bmi),col=3)

plot(density(healthcare$bmi,na.rm = T),col=2,main="Imputación bmi mediante kNN")
lines(density(df_imput$bmi),col=3)
```



Como se puede ver en la comparación de los dos métodos, no existe una variación muy significativa, sin embargo la Imputación mediante regresión presenta un mejor ajuste entre la distribución de la variable original y el bmi imputado, por lo que se ha optado por usa este método estocástico como técnica de imputación.

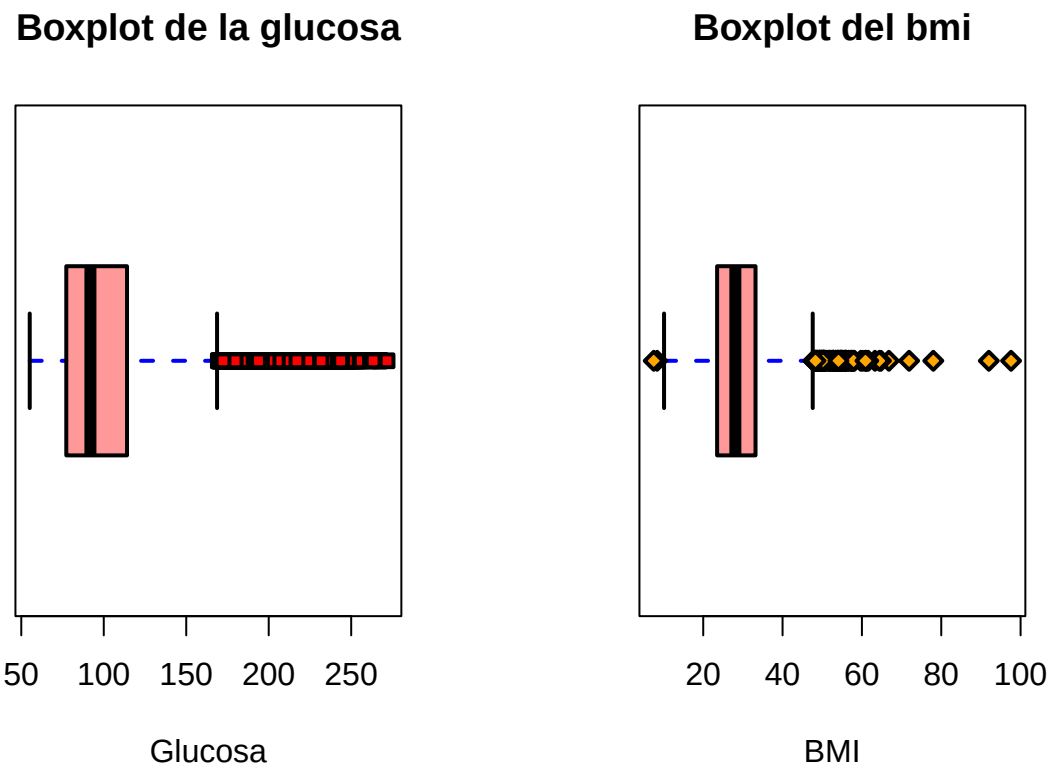
```
# Se genera una verificación de la imputación hecha en bmi


complete.data1 <-dplyr::rename(complete.data1, c( bmi_imput  = "bmi"))

healthcare<-merge(healthcare, complete.data1, by ="id", all = TRUE)

borrar <- c("bmi", "filtro")
healthcare <- healthcare[ , !(names(healthcare) %in% borrar)]

apply(is.na(healthcare), 2, sum)
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    Residence_type
##                 0                 0                 0                 0
## avg_glucose_level    smoking_status            stroke        bmi_imput
##                 0                 0                 0                 0
```

2.3.4 Verificación de datos extremos

Se realiza la verificación de la existencia o no de datos atípicos de las variables cuantitativas

```
# Se genera una verificación de los datos atípicos del bmi y del nivel de la glucosa

par(mfrow=c(1,2))

boxplot((healthcare$avg_glucose_level),
        horizontal = T,
        lwd = 2,
        col = rgb(1, 0, 0, alpha = 0.4),
        xlab = "Glucosa",
        main = "Boxplot de la glucosa",
        border = "black",
        outpch = 22,
        outbg = "red",
        whiskcol = "blue")
boxplot((healthcare$bmi),
        horizontal = T,
        lwd = 2,
        col = rgb(1, 0, 0, alpha = 0.4),
        xlab = "BMI",
        main = "Boxplot del bmi",
        border = "black",
        outpch = 23,
        outbg = "orange",
        whiskcol = "blue")
```



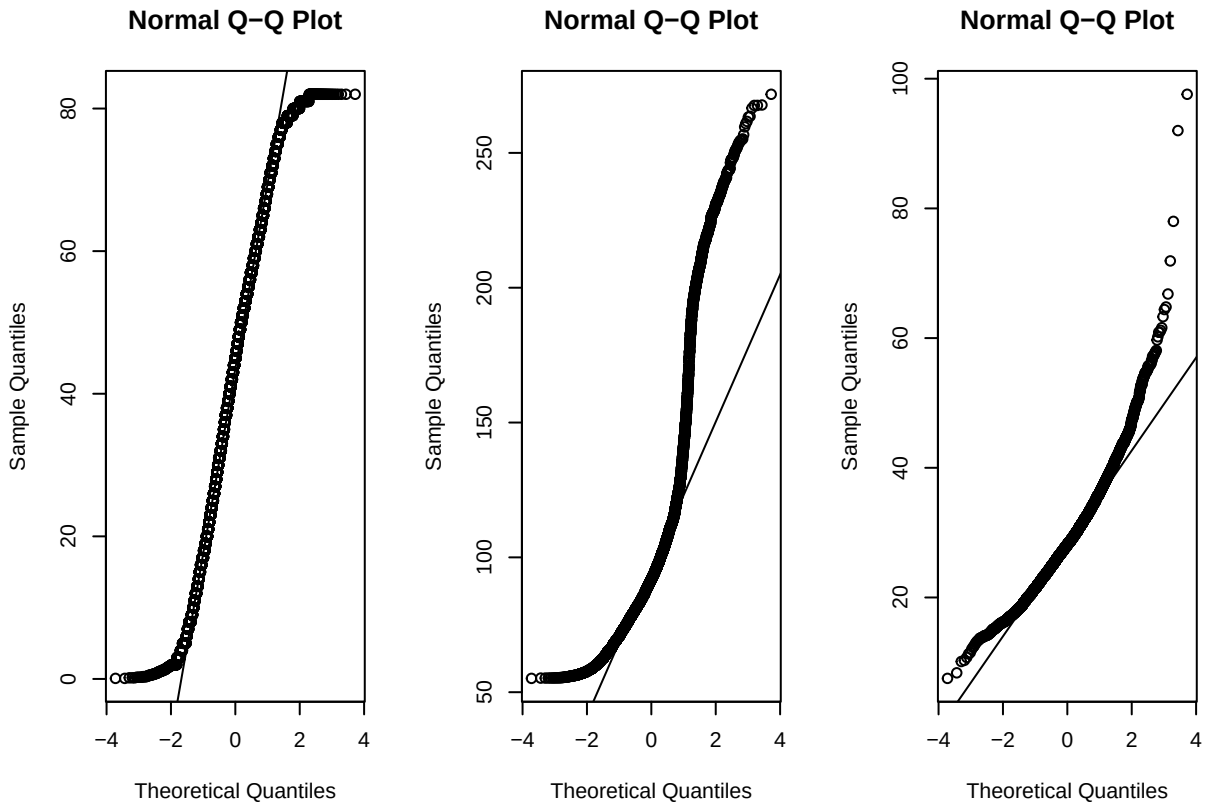
Según el diagrama de caja del bmi y de la glucosa, se podría concluir que existen datos atípicos, sin embargo al tratarse de un fenómeno medico donde la prevalencia o no de un ataque cardiaco puede deberse a la existencia de datos atípicos se ha optado por consérvalos y no aplicar ningún tratamiento sobre ellos.

2.4. Análisis de los datos

2.4.1 normalidad de las variables cuantitativas

Se verifica la normalidad de los atributos cuantitativos del dataset

```
par(mfrow=c(1,3))

#  Normalidad de la edad
qqnorm(healthcare$age)
qqline(healthcare$age)

#  Normalidad de la glucosa
qqnorm(healthcare$avg_glucose_level)
qqline(healthcare$avg_glucose_level)

#  Normalidad de la bmi
qqnorm(healthcare$bmi)
qqline(healthcare$bmi)
```




```
#  Se realiza la Prueba de Lilliefors

# edad
lillie.test(healthcare$age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  healthcare$age
## D = 0.050496, p-value < 2.2e-16
```

```
# glucosa
lillie.test(healthcare$avg_glucose_level)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  healthcare$avg_glucose_level
## D = 0.18318, p-value < 2.2e-16
```

```
# glucosa
lillie.test(healthcare$bmi)
```

```
##
```

```
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  healthcare$bmi
## D = 0.057463, p-value < 2.2e-16
```

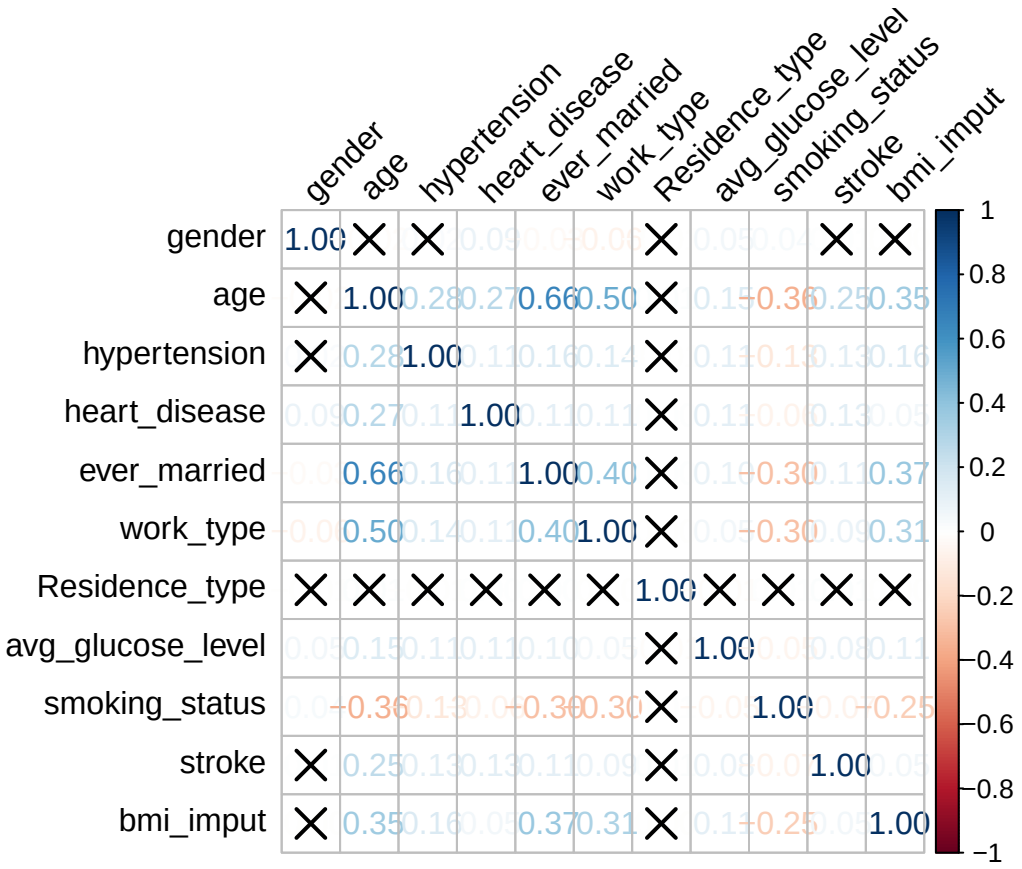
Mediante los QQ plot se puede ver que la variable de la glucosa como del bmi y la edad no se encuentran tan alineados respecto a la línea de referencia esto sobre todo en las partes más alejadas de la distribución, por otro lado en la prueba de Lilliefors se puede notar que el p calculado (2.2e-16) es mucho menor que el valor de contraste en este caso 0.05, por lo que se podría concluir que ninguna de las variables siguen una distribución normal.

2.4.2 Correlaciones

Como segundo punto dentro del análisis de los datos se va a proceder a calcular una matriz de correlaciones que nos permitirá determinar la medida de asociación entre los distintos atributos del dataset incluida la variable de interés para el presente estudio (stroke existencia o no de una ataque cardiaco), para este caso y al contar con variables de tipo cualitativo como cuantitativo y al haber verificado que las distribuciones de estas últimas no siguen una distribución normal se procede a usar el coeficiente de correlación de Spearman.

Para verificar la correlación de las variables categóricas, se les va aplicar un proceso de recodificación con el fin de que sean numéricas

```
# Se recodifica las variables categoricas en numericas con el fin de verificar la
#correlación existente, se trabaja en un nuevo dataset para no modificar al original

healthcare_corr<-healthcare
healthcare_corr$ever_married<-dplyr::recode(healthcare_corr$ever_married, No=0, Yes=1)
healthcare_corr$Residence_type<-dplyr::recode(healthcare_corr$Residence_type, Rural =0,
                                              Urban =1)
healthcare_corr$work_type<-dplyr::recode(healthcare_corr$work_type, children=1,
                                         Govt_job=2,Never_worked=3,Private =4,
                                         `Self-employed`=5)
healthcare_corr$smoking_status<-dplyr::recode(healthcare_corr$smoking_status,
                                         `formerly smoked`=1,
                                         `never smoked`=2, smokes=3, Unknown=4)

healthcare_corr$gender<-dplyr::recode(healthcare_corr$gender, Female=1, Male=2, Other=3)

columns2 <- c("id")
healthcare_corr <- healthcare_corr[ , !(names(healthcare_corr) %in% columns2)]


# se genera la matriz de correlaciones y se grafica para poder visualizar de mejor manera

matriz_correlacion<-cor_mat(healthcare_corr, method="spearman")

cor_plot(matriz_correlacion, method = "number")
```


gender
age
hypertension
heart_disease
ever_married
work_type
Residence_type
avg_glucose_level
smoking_status
stroke
bmi_imput
1.00
0.66
0.50
0.40
0.37
0.36
0.35
0.31
0.30
0.28
0.27
0.25
0.16
0.15
0.14
0.13
0.11
0.10
0.09
0.08
0.05
0.04
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1