Práctica 2: Limpieza y Análisis de Datos

Richard Jácome - Andrea Martínez

Junio 2021

Contents

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

El dataset a ser utilizado en esta práctica ha sido obtenido de la página de Kaggle y se puede acceder en el siguiente link: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Este dataset contiene información para predecir si un paciente tiene la probabilidad de sufrir un accidente cerebral vascular (AVC) basado en ciertas características como género, edad, ciertas enfermedades, etc.

De acuerdo a la Organización Mundial de la Salud, las enfermedades cardiovasculares son la principal causa de muerte en todo el mundo, dentro de las cuales se encuentran los ataques al corazón y los AVC. https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Es importante este dataset ya que permite analizar si cierto grupo de pacientes tiene mayor incidencia a sufrir AVC en comparación con otro y en base a esta predisposición se pueden definir políticas de medicina preventiva para evitar la ocurrencia de un derrame.

. Integración y selección de los datos de interés a analizar.

Los datos están contenidos en un solo dataset en formato csv con la siguiente estructura:

```
#Cargamos el archivo  respectivo
df_stroke <- read.csv("healthcare-dataset-stroke-data.csv", header=T, sep=",", stringsAsFactors = TRUE,
#Verificamos la estructura del archivo
str(df_stroke)
```

```
## 'data.frame':    5110 obs. of  12 variables:
##  $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
##  $ gender           : Factor w/ 3 levels "Female","Male",..: 2 1 2 1 1 2 2 1 1 1 ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 1 2 2 ...
##  $ work_type        : Factor w/ 5 levels "children","Govt_job",..: 4 5 4 4 5 4 4 4 4 4 ...
##  $ Residence_type   : Factor w/ 2 levels "Rural","Urban": 2 1 1 2 1 2 1 2 1 2 ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : Factor w/ 419 levels "10.3","11.3",..: 240 419 199 218 114 164 148 102 419 116
##  $ smoking_status   : Factor w/ 4 levels "formerly smoked",..: 1 2 2 3 2 1 2 2 4 4 ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

Podemos observar que el dataset contiene 5.110 observaciones con 12 variables de las cuales 6 variables son numéricas y 6 categóricas. La variable objetivo es “stroke” que puede tomar valores 0 o 1

Para efectos de este análisis de van a utilizar todas las variables proporcionadas en el dataset.

. Limpieza de Datos.

. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Primero verificamos el resumen de los datos:

```
# Estadísticas de valores NA
summary(df_stroke)
```

```
##        id           gender          age         hypertension
##  Min.   :   67   Female:2994   Min.   : 0.08   Min.   :0.00000
##  1st Qu.:17741   Male  :2115   1st Qu.:25.00   1st Qu.:0.00000
##  Median :36932   Other :   1   Median :45.00   Median :0.00000
```

```
##  Mean   :36518                 Mean   :43.23   Mean   :0.09746
##  3rd Qu.:54682                 3rd Qu.:61.00   3rd Qu.:0.00000
##  Max.   :72940                 Max.   :82.00   Max.   :1.00000
##
##  heart_disease     ever_married         work_type    Residence_type
##  Min.   :0.00000   No :1757     children     : 687   Rural:2514
##  1st Qu.:0.00000   Yes:3353     Govt_job     : 657   Urban:2596
##  Median :0.00000                Never_worked :  22
##  Mean   :0.05401                Private      :2925
##  3rd Qu.:0.00000                Self-employed: 819
##  Max.   :1.00000
##
##  avg_glucose_level      bmi               smoking_status     stroke
##  Min.   : 55.12    N/A    : 201   formerly smoked: 885   Min.   :0.00000
##  1st Qu.: 77.25    28.7   :  41   never smoked   :1892   1st Qu.:0.00000
##  Median : 91.89    28.4   :  38   smokes         : 789   Median :0.00000
##  Mean   :106.15    26.1   :  37   Unknown        :1544   Mean   :0.04873
##  3rd Qu.:114.09    26.7   :  37                          3rd Qu.:0.00000
##  Max.   :271.74    27.6   :  37                          Max.   :1.00000
##                    (Other):4719
```

Podemos apreciar que la variable bmi tiene valores N/A que deben ser saneados.

```
# Estadísticas de valores NA
colSums(is.na(df_stroke))
```

```
##                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                 0                 0                 0
```

Se puede apreciar que no existen valores nulos (NA).

```
# Estadísticas de valores vacíos
colSums(df_stroke=="")
```

```
##                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                 0                 0                 0
```

También se comprueba que no hay datos vacíos.

Sabemos que bmi contiene valores NA, si embargo no se visualiza con los procesos ejecutados anteriormente, esto se debe que está en tipo factor, por lo cual se lo debe pasar atributo numérico para poder imputar valores.

```
library(varhandle)
df_stroke$bmi <- unfactor(df_stroke$bmi)
df_stroke$bmi <- as.double(df_stroke$bmi)
```

```
# Estadísticas de valores NA
colSums(is.na(df_stroke))
```

```
##                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               201                 0                 0
```

Para reemplazar los valores perdidos utilizaremos el método kNN, que se basa en los k vecinos más próximos de acuerdo con los valores de los registros.

```
# Imputación de valores mediante la función kNN() del paquete VIM
suppressWarnings(suppressMessages(library(VIM)))
df_stroke$bmi <- kNN(df_stroke)$bmi
```

```
# Estadísticas de valores NA
colSums(is.na(df_stroke))
```

```
##                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                 0                 0                 0
```

Volvemos a analizar los valores NA y ya no están presentes.

. Identificación y tratamiento de valores extremos.

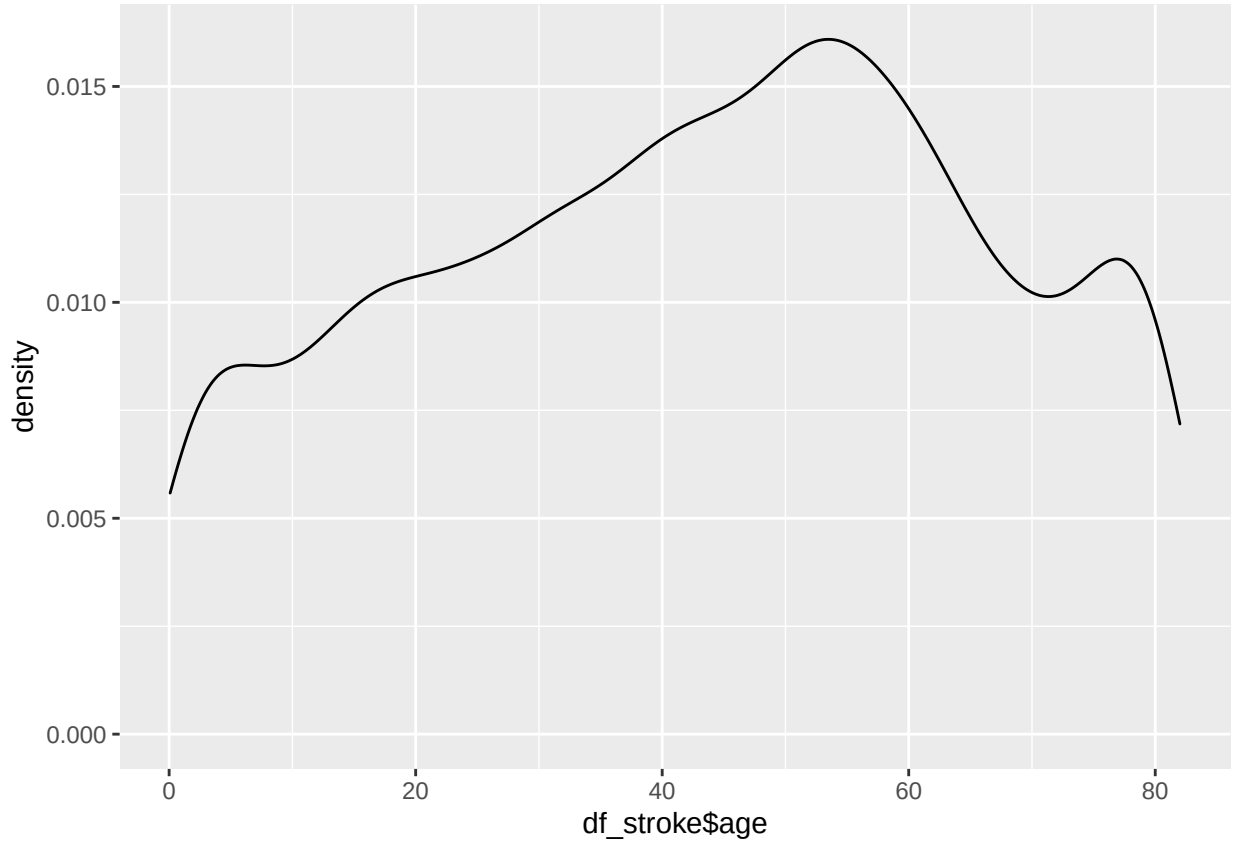
```
# número de variantes por variable del data frame
apply(df_stroke,2, function(x) length(unique(x)))
```

```
##                id            gender               age      hypertension
##              5110                 3               104                 2
##     heart_disease      ever_married         work_type    Residence_type
##                 2                 2                 5                 2
## avg_glucose_level               bmi    smoking_status            stroke
##              3979               418                 4                 2
```

Primero verificamos cuantas variantes tenemos en los atributos para buscar outliers, donde se tenga una cantidad alta de variantes, por lo cual analizaremos los atributos: age, avg_glucose_level y bmi, id no se toma en cuenta, ya que solo es un identificador.

Vamos a representar la distribución de los valores de las variables para visualizar picos atípicos que se tomarán como inconsistencias:

```
library(ggplot2)
ggplot(mapping= aes(x=df_stroke$age))+ geom_density()
```



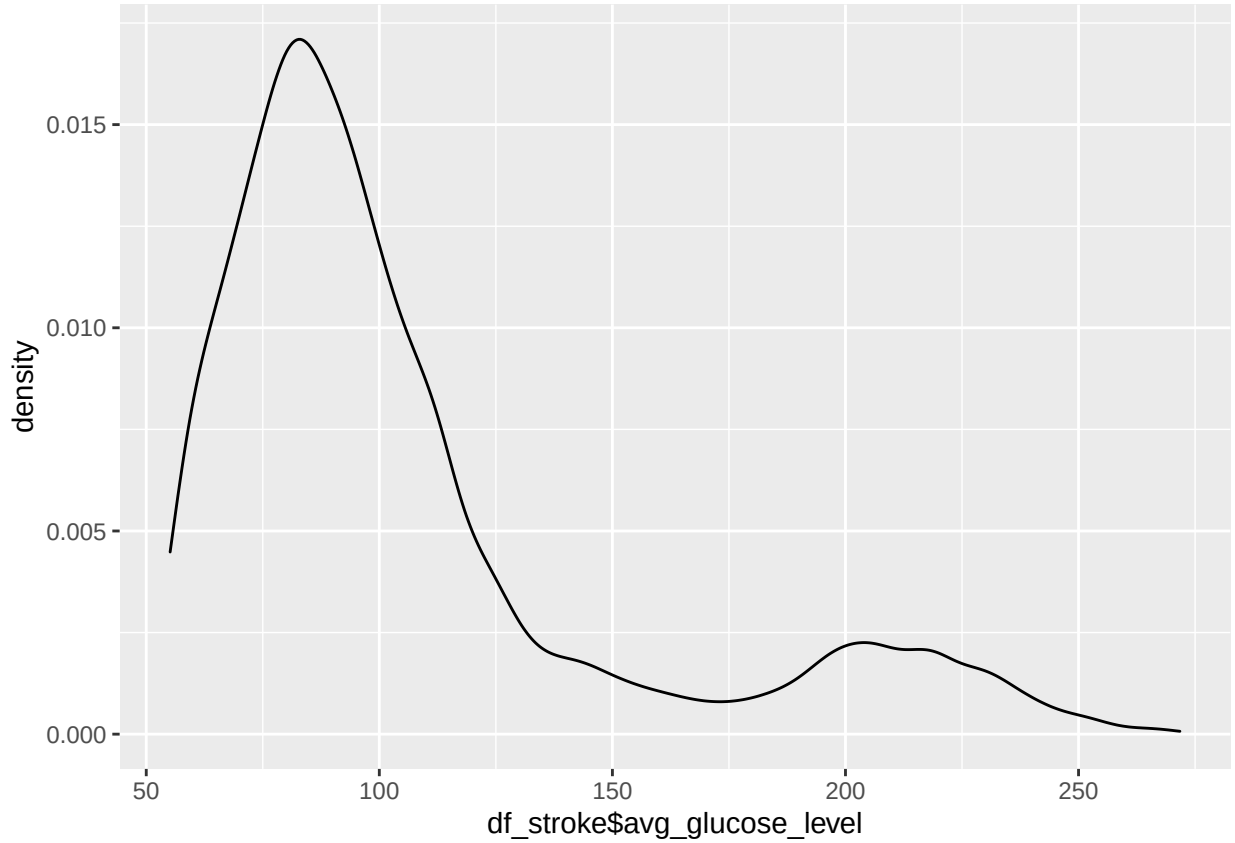
En el atributo age podemos evidenciar que se tiene una tendencia a la normalidad y visualizan variaciones diferentes en los extremos que no necesariamente son valores atípicos por lo cual verificaremos con boxplot.

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

```
## numeric(0)
```

Se confirma que la variable age no contiene valores atípicos.

```
library(ggplot2)
ggplot(mapping= aes(x=df_stroke$avg_glucose_level))+ geom_density()
```



Se puede evidenciar en la gráfica que los datos tienden a la normalidad, en el extremos derecho se tiene una curva con valores que no siguen la tendencia, sin embargo no justican ser atípicos.

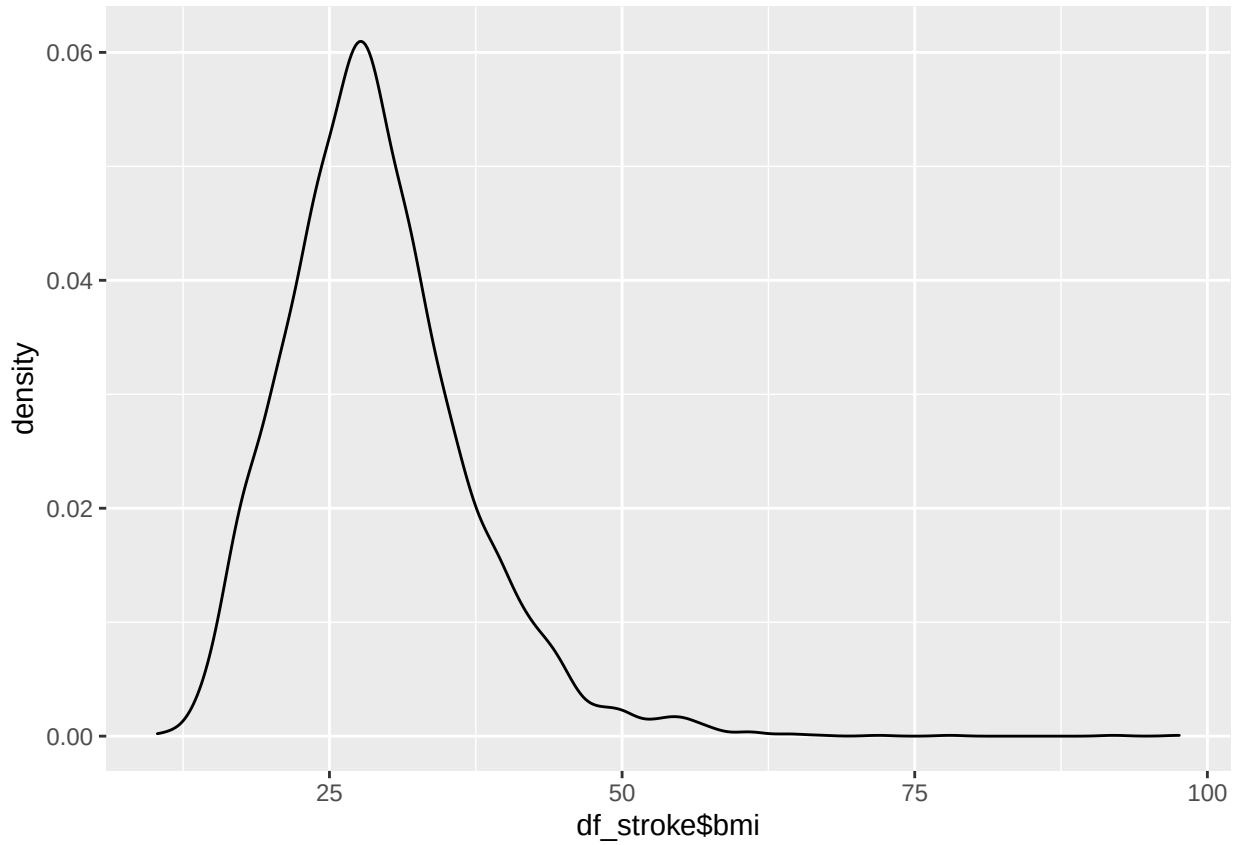
```
(boxplot.stats(df_stroke$avg_glucose_level)$out)
```

```
##   [1] 228.69 202.21 171.23 174.12 186.21 219.84 214.09 191.61 221.29 217.08
##  [11] 193.94 233.29 228.70 208.30 189.84 195.23 211.78 212.08 196.92 252.72
##  [21] 219.72 213.03 243.58 197.54 196.71 237.75 194.99 180.93 185.17 221.58
##  [31] 179.12 228.56 240.09 226.98 235.63 240.59 190.32 231.61 191.82 224.10
##  [41] 216.94 259.63 249.31 219.91 200.59 190.14 182.99 206.09 263.32 207.28
##  [51] 194.37 199.20 221.79 239.07 169.67 223.83 231.56 221.89 195.71 203.87
##  [61] 185.49 213.22 215.94 209.86 205.77 271.74 200.62 242.52 175.29 208.65
##  [71] 205.33 210.40 199.86 219.73 250.89 205.35 216.58 184.40 199.84 218.46
##  [81] 211.06 197.28 233.94 247.51 210.95 243.53 205.84 198.21 206.72 214.45
##  [91] 190.70 203.04 242.30 220.49 218.46 216.70 234.58 235.85 243.50 182.20
## [101] 229.92 215.60 239.64 200.28 205.23 209.58 210.78 251.60 213.37 223.36
## [111] 178.29 203.81 205.26 211.03 225.47 180.63 227.10 201.76 170.05 217.30
## [121] 196.01 184.15 198.69 186.17 183.45 210.48 193.83 183.34 247.69 191.47
## [131] 239.82 189.57 207.58 182.86 215.64 196.36 188.11 205.50 204.86 228.08
## [141] 219.53 219.97 214.05 200.49 240.71 197.10 194.62 222.21 250.20 173.43
## [151] 184.25 254.60 212.01 186.45 189.49 186.32 226.70 183.10 194.04 237.15
## [161] 231.19 207.32 207.64 236.84 204.63 232.89 195.03 170.95 227.91 204.50
## [171] 206.25 254.63 246.34 195.16 223.68 229.20 193.22 204.57 251.46 220.52
## [181] 195.04 218.65 211.49 224.71 226.11 210.94 230.68 198.02 204.17 267.76
## [191] 217.71 180.76 239.52 229.86 210.96 195.25 217.39 201.25 197.79 214.77
## [201] 181.23 189.45 206.40 178.76 197.58 199.96 205.77 237.21 246.53 206.33
## [211] 206.98 227.28 228.70 169.97 244.28 251.99 191.79 216.88 222.29 213.11
```

```
## [221] 227.51 201.01 210.00 237.58 207.45 226.93 253.16 238.53 207.79 196.20
## [231] 231.76 216.92 194.98 218.54 183.00 237.17 178.33 197.09 242.84 202.66
## [241] 216.90 210.00 208.05 222.60 199.14 191.48 200.16 190.40 215.90 233.52
## [251] 213.54 188.69 219.50 217.66 227.16 209.90 176.48 211.88 225.60 210.23
## [261] 234.82 230.59 224.63 185.71 208.17 185.31 203.04 187.87 213.87 222.85
## [271] 198.36 196.25 194.53 204.05 199.18 209.26 217.11 222.46 187.52 237.74
## [281] 223.35 201.07 208.06 186.95 198.24 229.21 209.06 228.42 212.97 202.05
## [291] 206.25 231.69 219.96 197.69 199.88 170.22 208.78 222.29 220.36 187.88
## [301] 191.66 217.75 226.88 186.40 169.49 203.81 170.76 189.44 249.29 211.35
## [311] 206.59 196.33 242.94 226.75 185.00 199.83 227.81 240.81 239.28 231.50
## [321] 192.37 220.47 196.91 180.80 247.48 216.00 219.39 220.47 173.96 198.33
## [331] 191.33 206.52 216.96 170.93 232.81 207.95 229.58 187.22 227.04 214.42
## [341] 233.71 216.40 266.59 227.94 205.00 203.44 243.73 176.25 200.28 221.43
## [351] 213.38 192.16 215.72 173.14 202.57 209.50 203.16 201.45 206.15 196.61
## [361] 219.92 231.95 216.38 213.33 172.33 243.59 169.43 183.87 227.98 208.20
## [371] 199.42 190.13 235.54 178.89 227.74 213.80 250.80 217.84 217.00 217.40
## [381] 190.92 182.90 255.17 217.55 227.96 231.71 196.81 222.66 223.58 198.79
## [391] 192.39 233.30 201.38 236.14 193.81 239.95 170.88 202.21 181.30 198.79
## [401] 202.55 232.12 203.57 230.78 204.98 227.89 216.71 202.67 221.80 202.38
## [411] 215.81 220.24 195.61 267.61 176.71 207.62 201.58 231.43 220.26 211.12
## [421] 177.91 215.33 212.02 228.20 260.85 223.90 169.74 207.96 176.78 205.01
## [431] 191.78 214.43 220.64 204.77 248.37 194.53 228.92 227.68 226.73 219.17
## [441] 215.92 198.12 240.86 263.56 200.14 235.45 207.71 228.05 223.14 174.43
## [451] 214.51 231.31 238.78 233.59 188.13 205.97 190.89 193.87 214.77 189.88
## [461] 197.11 192.47 199.38 202.98 198.32 226.38 236.79 219.82 239.19 206.62
## [471] 216.88 204.92 226.84 234.35 200.73 202.51 218.00 209.15 202.66 196.50
## [481] 209.50 219.81 205.23 234.27 239.21 196.08 176.38 175.74 193.45 180.45
## [491] 219.38 173.90 217.94 216.64 173.97 208.85 219.70 208.05 185.28 198.30
## [501] 206.66 200.68 218.60 223.26 172.27 221.83 218.10 200.46 217.79 233.47
## [511] 181.23 200.98 219.67 207.60 247.97 231.15 186.54 221.06 212.62 217.74
## [521] 208.99 197.36 222.52 232.64 207.37 201.96 213.43 248.24 229.94 202.06
## [531] 253.93 194.75 207.84 228.26 203.76 205.78 179.67 230.74 216.19 200.66
## [541] 228.50 232.29 200.91 236.04 254.95 196.58 189.82 193.61 195.74 221.24
## [551] 192.50 212.92 191.94 247.87 229.73 261.67 256.74 221.08 208.39 227.23
## [561] 203.27 234.50 190.67 197.06 216.07 179.14 203.87 235.06 195.43 200.25
## [571] 223.64 199.78 176.42 244.30 223.16 226.28 172.86 213.92 212.19 200.80
## [581] 222.58 206.53 232.78 187.47 234.06 242.62 174.54 231.54 219.80 187.99
## [591] 234.45 240.69 217.57 234.51 182.22 214.73 208.69 231.72 206.53 193.80
## [601] 203.01 177.56 198.84 243.52 238.27 208.31 176.34 211.83 215.69 267.60
## [611] 215.07 225.35 196.26 182.52 212.87 183.43 185.27 206.49 253.86 203.36
## [621] 175.92 191.15 223.78 211.58 179.38 193.88 174.37
```

Al comprobar con boxplot nos arroja todos lo valores altos de los datos, ya que no siguen el patrón de la normal, sin embargo, los datos no son erróneos, por lo cual no se los debe modificar o quitar.

```
library(ggplot2)
ggplot(mapping= aes(x=df_stroke$bmi))+ geom_density()
```



Al igual que los atributos analizados anteriormente, se tiene una normal y cuando llega a valores altos queda fuera de tendencia, por lo cual se comprueba con boxplot:

```
boxplot.stats(df_stroke$bmi)$out
```

```
##   [1] 48.9 47.5 56.6 50.1 54.6 60.9 54.7 48.2 64.8 47.3 54.7 49.8 60.2 51.0 51.5
##  [16] 71.9 50.2 47.8 54.6 55.7 55.7 57.5 54.2 52.3 50.3 78.0 50.2 53.4 55.2 48.4
##  [31] 50.6 49.5 55.0 54.8 50.2 47.5 52.8 66.8 55.1 48.5 55.9 57.3 49.8 56.0 51.8
##  [46] 57.7 48.9 49.3 49.8 54.0 56.1 97.6 53.9 49.4 48.5 49.2 48.7 48.9 53.8 48.8
##  [61] 52.7 52.8 55.7 53.5 50.5 51.9 63.3 52.8 61.2 48.0 50.1 48.3 58.1 49.3 50.4
##  [76] 52.7 48.3 49.3 51.9 53.4 50.3 59.7 47.4 52.5 52.9 54.7 61.6 49.9 53.8 47.3
##  [91] 54.3 47.9 55.0 50.9 50.6 57.2 64.4 92.0 50.8 55.9 57.9 47.6 55.7 48.8 57.2
## [106] 47.5 50.2 47.1 48.1 51.7 60.9 47.8 47.6 54.1 56.6 49.5 47.6
```

Una vez comprobamos que con valores altos los datos no siguen con la tendencia normal, sin embargo no indica que los valores son atípicos o frutos del error, sino que pertenecen a personas con características diferentes al promedio.

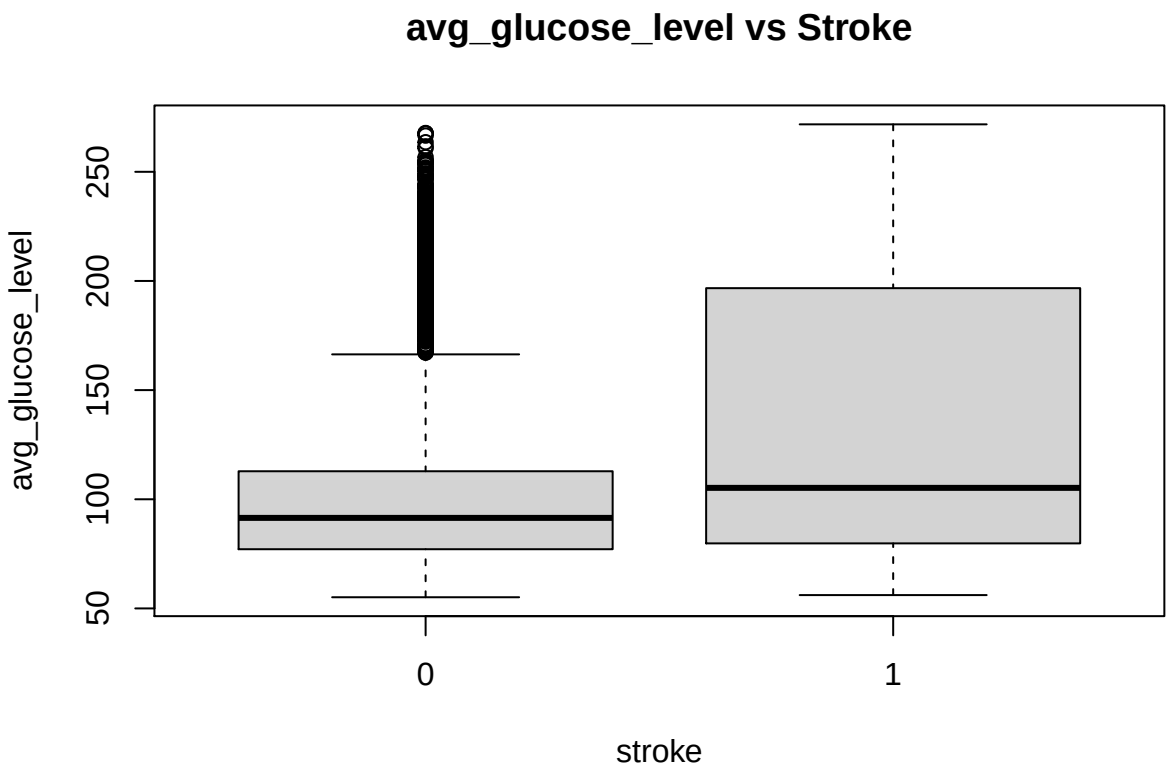
. Análisis de los Datos.

Creamos diversos diagramas de caja para observar la distribución de las variables 'Age', 'avg_glucose_level' y 'bmi' respecto de la variable 'stroke'

```r
#Gráficos
boxplot(formula = age ~ stroke, data =  df_stroke,  main = "Age vs Stroke")
```

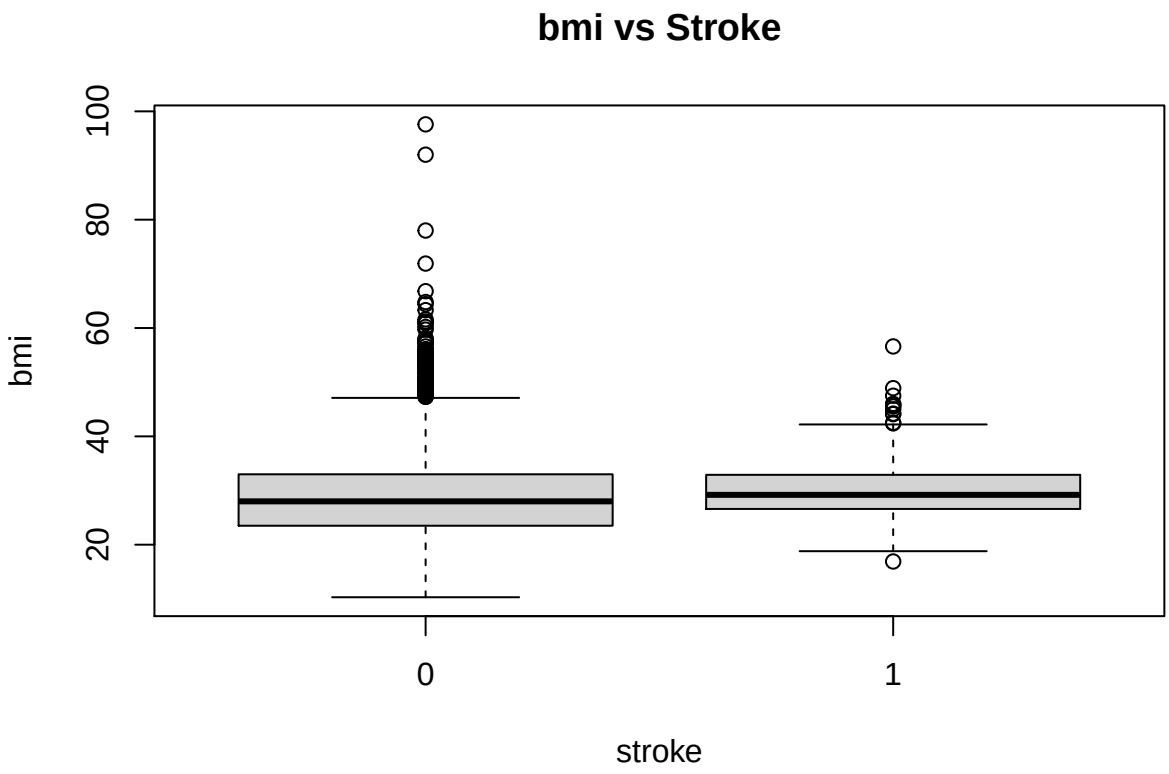
Podemos observar que la mediana de la edad cuando se ha tenido un ACV es mayor que cuando no hay incidencia de AVC.

```r
#Gráficos
boxplot(formula = avg_glucose_level ~ stroke, data =  df_stroke,  main = "avg_glucose_level vs Stroke")
```



Podemos observar que la mediana de los niveles de glucosa cuando se ha tenido un AVC es mayor que cuando no hay incidencia de ACV.

```
#Gráficos
boxplot(formula = bmi ~ stroke, data =  df_stroke,  main = "bmi vs Stroke")
```



Podemos observar que se tiene mayor cantidad de valores outliers en el bmi de las personas que no han tenido un AVC.

. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

En el análisis no se tomará en cuenta la variable: id, porque no genera información adicional que el conteo de registros e indentificación de los mismos.

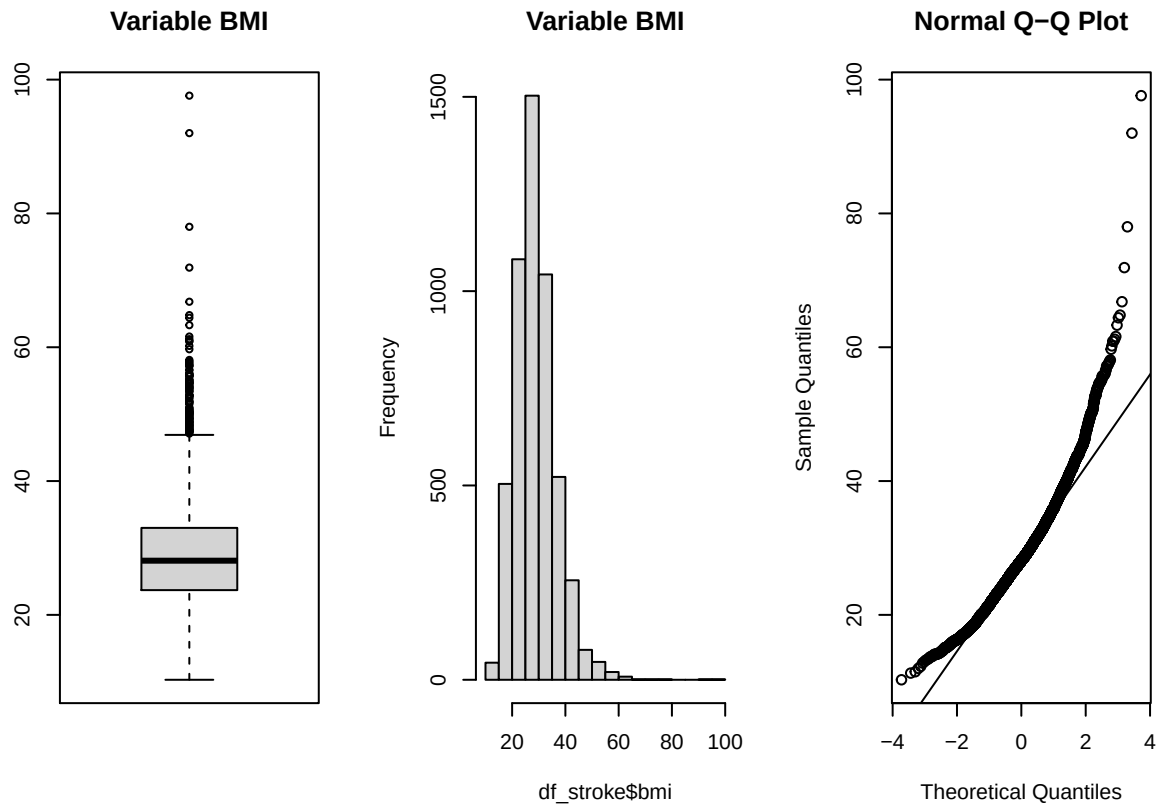
Para las gráficas se tomará en cuenta el número de variantes por atributo, ya que en el caso de atributos como age, avg_glucose_level y bmi, tienen múltiples valores, por lo tanto se los agrupará en conjuntos para tener una mejor visión en gráficas.

. Comprobación de la normalidad y homogeneidad de la varianza.

Vamos a comprobar la normalidad de la variable bmi

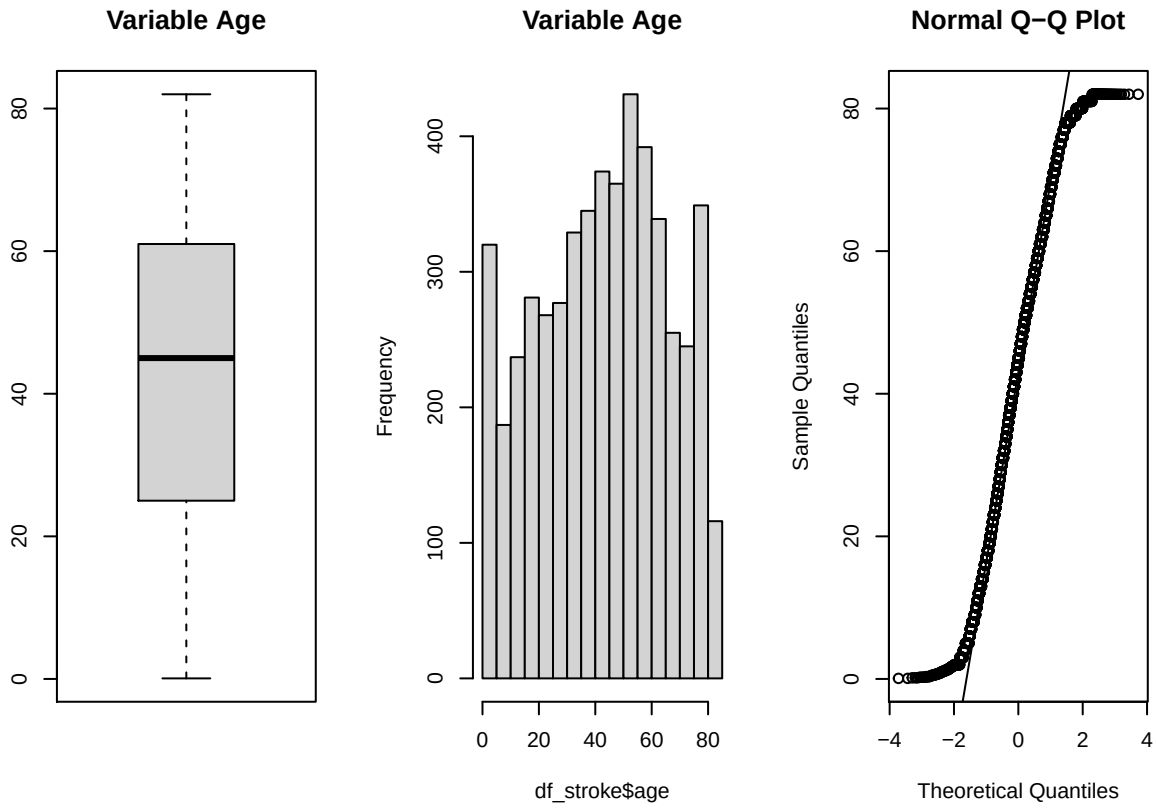
```
#Gráficos Boxplot e histograma
par(mfrow=c(1,3))
b_bmi <- boxplot(df_stroke$bmi,  main = "Variable BMI")
hist(df_stroke$bmi, main = "Variable BMI")
```

```
qqnorm(df_stroke$bmi)
qqline(df_stroke$bmi)
```



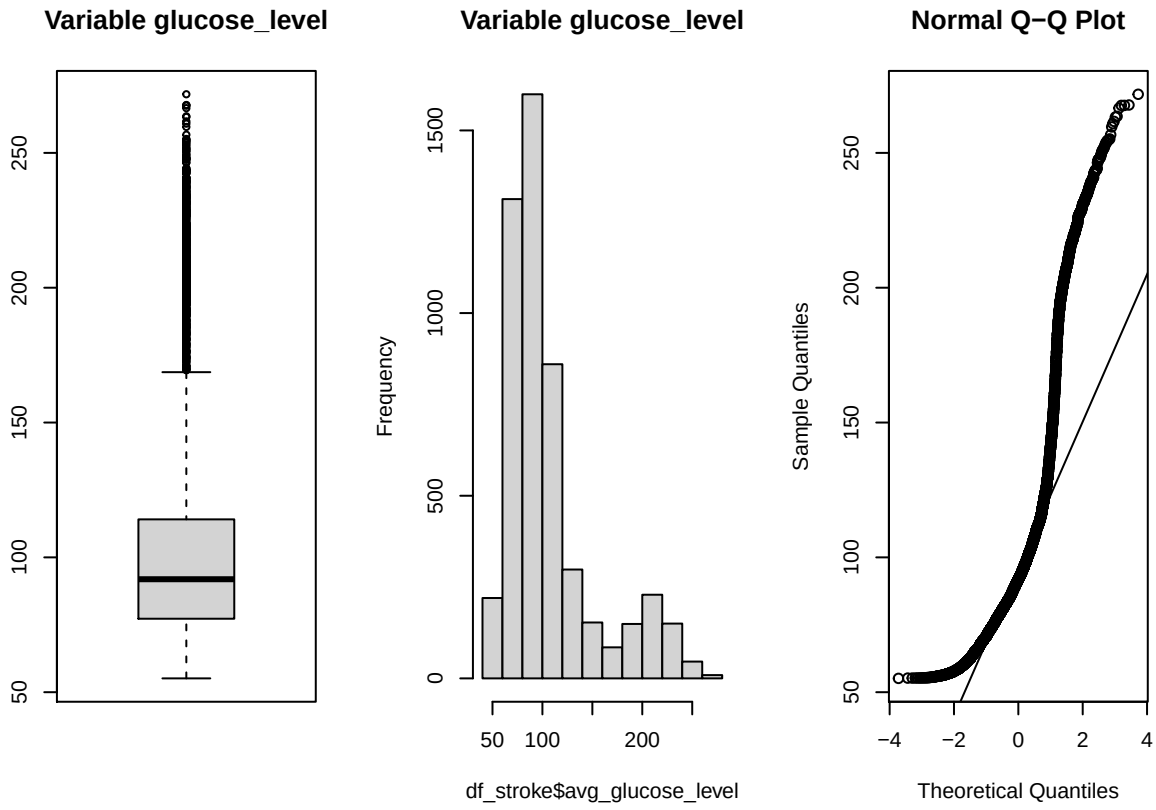
Podemos observar en los gráficos que la variable bmi no tiene una distribución normal

```
#Gráficos Boxplot e histograma
par(mfrow=c(1,3))
b_age <- boxplot(df_stroke$age,  main = "Variable Age")
hist(df_stroke$age, main = "Variable Age")
qqnorm(df_stroke$age)
qqline(df_stroke$age)
```



Podemos observar en los gráficos que la variable age no tiene una distribución normal

```
#Gráficos Boxplot e histograma
par(mfrow=c(1,3))
b_age <- boxplot(df_stroke$avg_glucose_level,  main = "Variable glucose_level")
hist(df_stroke$avg_glucose_level, main = "Variable glucose_level")
qqnorm(df_stroke$avg_glucose_level)
qqline(df_stroke$avg_glucose_level)
```



Podemos observar en los gráficos que la variable avg_glucose_level no tiene una distribución normal

A pesar que estas tres variables no son normales en la distribución total de los datos, al tener una muestra mayor a 400, se puede asumir normalidad por el teorema del límite central (distribución de la media puede ser aproximadamente normal).

Procedemos a comprobar la homogeneidad de la varianza

```
#Test de varianzas iguales
var.test(df_stroke$age[df_stroke$stroke=="0"], df_stroke$age[df_stroke$stroke=="1"])
```

```
##
##  F test to compare two variances
##
## data:  df_stroke$age[df_stroke$stroke == "0"] and df_stroke$age[df_stroke$stroke == "1"]
## F = 3.0677, num df = 4860, denom df = 248, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.541299 3.648292
## sample estimates:
## ratio of variances
##           3.067715
```

El valor p-value es menor alfa, por lo tanto se descarta la hipótesis nula, es decir se descarta la igualdad de varianzas en la edad cuando el paciente no ha tenido un AVC que cuando si.

```
#Test de varianzas iguales
var.test(df_stroke$bmi[df_stroke$stroke=="0"], df_stroke$bmi[df_stroke$stroke=="1"])
```

```
##
##  F test to compare two variances
##
## data:  df_stroke$bmi[df_stroke$stroke == "0"] and df_stroke$bmi[df_stroke$stroke == "1"]
## F = 1.6986, num df = 4860, denom df = 248, p-value = 1.18e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.407155 2.020114
## sample estimates:
## ratio of variances
##            1.69864
```

El valor p-value es menor alfa, por lo tanto se descarta la hipótesis nula, es decir se descarta la igualdad de varianzas en el bmi cuando el paciente no ha tenido un AVC que cuando si.

```
#Test de varianzas iguales
var.test(df_stroke$avg_glucose_level[df_stroke$stroke=="0"], df_stroke$avg_glucose_level[df_stroke$strol
```

```
##
##  F test to compare two variances
##
## data:  df_stroke$avg_glucose_level[df_stroke$stroke == "0"] and df_stroke$avg_glucose_level[df_stroke
## F = 0.5014, num df = 4860, denom df = 248, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4153606 0.5962922
## sample estimates:
## ratio of variances
##          0.5014003
```

El valor p-value es menor alfa, por lo tanto se descarta la hipótesis nula, es decir se descarta la igualdad de varianzas en el bmi cuando el paciente no ha tenido un AVC que cuando si.

. Aplicación de pruebas estadísticas para comparar los grupos de datos.

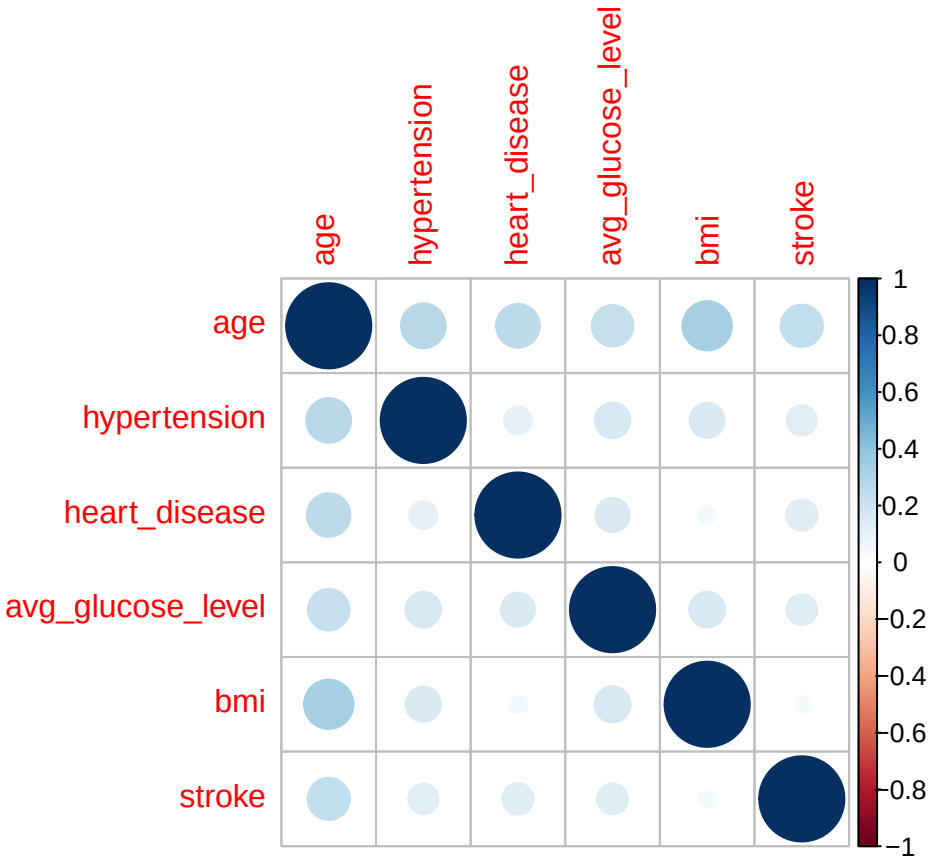
En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

. Analisis de correlacion

```
library(corrplot)
res <- cor(df_stroke[,c('age','hypertension','heart_disease','avg_glucose_level','bmi', 'stroke')])
round(res, 2)
```

```
##                    age hypertension heart_disease avg_glucose_level  bmi stroke
## age               1.00         0.28          0.26              0.24 0.34   0.25
## hypertension      0.28         1.00          0.11              0.17 0.17   0.13
## heart_disease     0.26         0.11          1.00              0.16 0.04   0.13
## avg_glucose_level 0.24         0.17          0.16              1.00 0.18   0.13
## bmi               0.34         0.17          0.04              0.18 1.00   0.04
## stroke            0.25         0.13          0.13              0.13 0.04   1.00
```

```
corrplot(res)
```



Podemos observar que las seis variables numericas no tiene correlacion entre si.

. Contraste de hipótesis

Podemos analizar que las personas de mas edad tienen mayor incidencia de presentar un AVC con un nivel de confianza del 95%.

Vamos a plantear una prueba de contraste de hipotesis:

La hipótesis nula es que las medias de la edad son iguales en pacientes con AVC que sin AVC

H0: media_edad_stroke = media_edad_sin_stroke

La hipótesis alternativa es

H1: media_edad_stroke > media_edad_sin_stroke

Debemos aplicar un test de dos muestras independientes sobre la media con varianza desconocida y diferente. Es un test unilateral por la derecha.

```
t.test(df_stroke$age[df_stroke$stroke=="1"], df_stroke$age[df_stroke$stroke=="0"], var.equal=FALSE, alt
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  df_stroke$age[df_stroke$stroke == "1"] and df_stroke$age[df_stroke$stroke == "0"]
## t = 29.686, df = 331.65, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  24.32553      Inf
## sample estimates:
## mean of x mean of y
##  67.72819  41.97154
```

Podemos observar que el p_value es significativamente menor que alfa (0.05), por lo tanto tenemos evidencia estadistica para rechazar la hipótesis nula, esto es que la media de la edad de los pacientes con AVC es mayor que sin AVC.

. Modelo de regresión logistica

Vamos a aplicar un modelo de regresión logística para predecir la probabilidad de tener un AVC en función de las demás variables

```
mod.log.1<-glm(stroke~gender+age+hypertension+heart_disease+ever_married+work_type+Residence_type+avg_g
summary(mod.log.1)
```

```
##
## Call:
## glm(formula = stroke ~ gender + age + hypertension + heart_disease +
##     ever_married + work_type + Residence_type + avg_glucose_level +
##     bmi + smoking_status, family = binomial(), data = df_stroke)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1274  -0.3196  -0.1637  -0.0871   3.5691
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -6.780e+00  7.845e-01  -8.643  < 2e-16 ***
## genderMale                   1.285e-02  1.419e-01   0.091 0.927857
## genderOther                 -1.053e+01  1.455e+03  -0.007 0.994229
## age                          7.474e-02  5.835e-03  12.809  < 2e-16 ***
## hypertension                 4.035e-01  1.652e-01   2.442 0.014586 *
## heart_disease                2.796e-01  1.911e-01   1.463 0.143423
## ever_marriedYes             -1.839e-01  2.254e-01  -0.816 0.414696
## work_typeGovt_job           -9.450e-01  8.366e-01  -1.130 0.258633
## work_typeNever_worked       -1.033e+01  3.092e+02  -0.033 0.973355
## work_typePrivate            -8.026e-01  8.205e-01  -0.978 0.327985
## work_typeSelf-employed      -1.180e+00  8.411e-01  -1.403 0.160761
## Residence_typeUrban          8.336e-02  1.383e-01   0.603 0.546782
## avg_glucose_level            4.027e-03  1.204e-03   3.346 0.000821 ***
## bmi                          1.072e-03  1.128e-02   0.095 0.924275
## smoking_statusnever smoked -2.067e-01  1.759e-01  -1.175 0.240122
## smoking_statussmokes         1.126e-01  2.154e-01   0.523 0.600989
## smoking_statusUnknown       -7.278e-02  2.084e-01  -0.349 0.726880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1990.4  on 5109  degrees of freedom
## Residual deviance: 1581.2  on 5093  degrees of freedom
## AIC: 1615.2
##
## Number of Fisher Scoring iterations: 14
```

Podemos observar que con un valor de alfa (0.05), las variables estadisticamente mas significativas para predecir la probabilidad de un accidente AVC son la edad, si el paciente tiene hipertension y los niveles de glucosa.

. Representación de los resultados a partir de tablas y gráficas.

Para la representación gráfica de los atributos, se toma en cuenta el número de variables que estos tienen, por lo cual se vuelve a analizar el número de variantes:

```
#conteo de variables por campo
apply(df_stroke,2, function(x) length(unique(x)))
```

```
##                id            gender               age      hypertension
##              5110                 3               104                 2
##     heart_disease      ever_married         work_type    Residence_type
##                 2                 2                 5                 2
## avg_glucose_level               bmi    smoking_status            stroke
##              3979               418                 4                 2
```

Como se detalló al principio de la sección no se tomará en cuenta la variable id, además, se discretizarán los valores de los atributos age, avg_glucose_level y bmi. Para esto se crea un nuevo dataset el cual excluye el atributo id:

```
nuevo_dataset <- df_stroke
nuevo_dataset$id <- NULL
filas=dim(nuevo_dataset)[1]
```

Las edades separamos en 5 grupos entre niños, adolecentes, adultos jóvenes, adultos maduros y adultos mayores.

```
nuevo_dataset[["age"]] <- cut(nuevo_dataset[["age"]],
                              c(min(nuevo_dataset$age)-0.1,
                                min(nuevo_dataset$age)+12,
                                min(nuevo_dataset$age)+19,
                                min(nuevo_dataset$age)+35,
                                min(nuevo_dataset$age)+55,
                                max(nuevo_dataset$age)),
                              labels = c("Niños", "Adolecentes",
                                         "Adultos Jóvenes",
                                         "Adultos Maduros",
                                         "Adultos Mayores"))
nuevo_dataset$age <- as.factor(nuevo_dataset$age)
```

El nivel de glucosa lo categorizaremos en 3 niveles Alto, Medio y Bajo.

El primer intervalo debe tomar en cuenta el mínimo por lo cual se debe poner un número menor a este, para que sea incluido, ya que se dividirá en 3 intervalos, se necesitarán 4 valores, para lo cual primero se calcula el valor del incremento entre valores, este valor es la diferencia entre el máximo y mínimo dividio para 3, ya con esto podemos formar 3 intervalos con 4 valores.

```
incr_fixed <- (max(nuevo_dataset$avg_glucose_level) - min(nuevo_dataset$avg_glucose_level))/3
nuevo_dataset[["avg_glucose_level"]] <- cut(nuevo_dataset[["avg_glucose_level"]],
                                    c(min(nuevo_dataset$avg_glucose_level) - 0.01,
                                      min(nuevo_dataset$avg_glucose_level) + incr_fixed,
                                      min(nuevo_dataset$avg_glucose_level) + 2 * incr_fixed,
                                      max(nuevo_dataset$avg_glucose_level)),
                                    labels = c("Bajo", "Medio","Alto"))
nuevo_dataset$avg_glucose_level <- as.factor(nuevo_dataset$avg_glucose_level)
```

La variable bmi se la categoriza de acuerdo con su significado en el sentido de las medidas de índice corporal, es decir, Bajo Peso, Normal, Sobrepeso y Obesidad.

```
nuevo_dataset[["bmi"]] <- cut(nuevo_dataset[["bmi"]],
                                    c(min(nuevo_dataset$bmi)-0.1,
                                      min(nuevo_dataset$bmi)+18.5,
                                      min(nuevo_dataset$bmi)+24.9,
                                      min(nuevo_dataset$bmi)+29.9,
                                      max(nuevo_dataset$bmi)),
                                    labels = c("Bajo Peso", "Normal",
                                               "Sobrepeso",
                                               "Obesidad"))
nuevo_dataset$bmi <- as.factor(nuevo_dataset$bmi)
```

Finalmente transformamos los atributos faltantes a tipo factor para poder graficarlos:

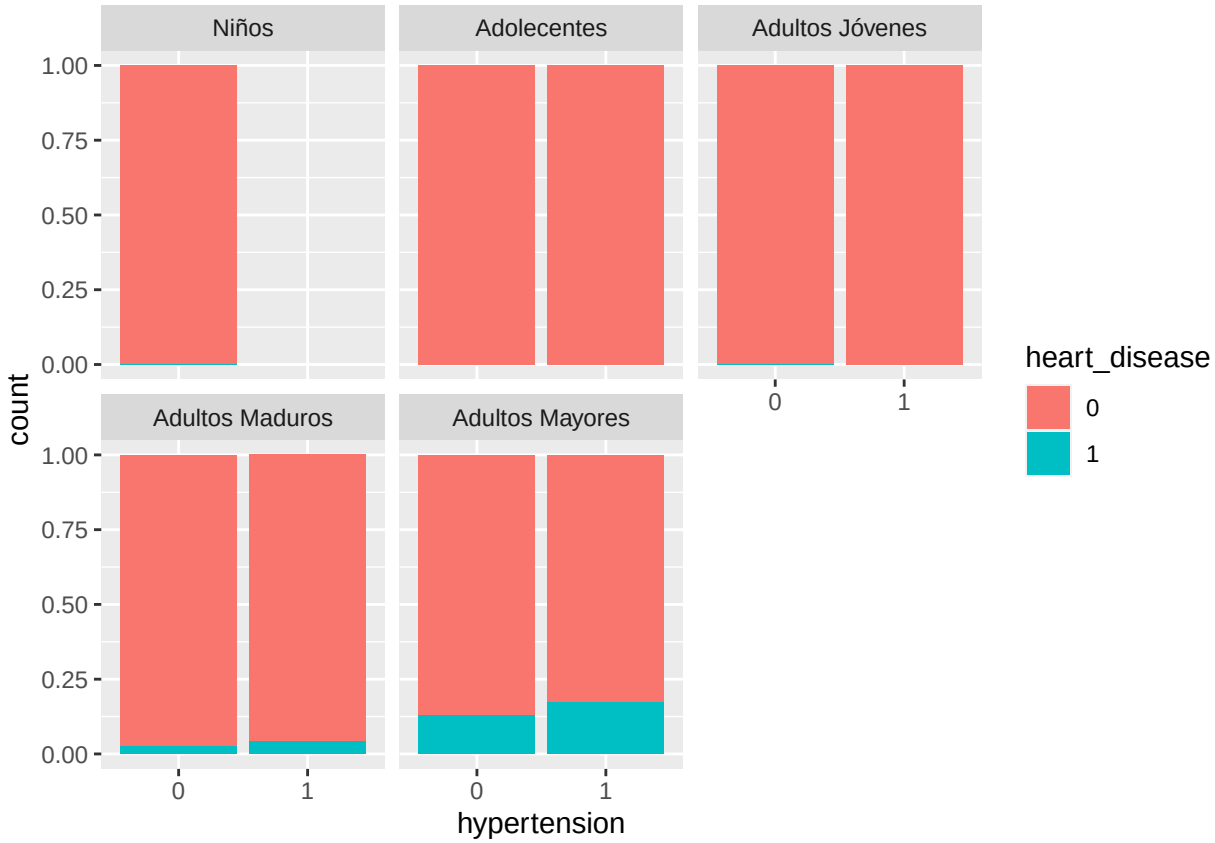
```
cols<-c("hypertension","heart_disease","stroke")
for (i in cols){
  nuevo_dataset[,i] <- as.factor(nuevo_dataset[,i])
}
```

Gráficas

```
library(ggplot2)
ggplot(data = nuevo_dataset[1:filas,],aes(x=gender,fill=hypertension))+geom_bar(position="fill")+facet_
```

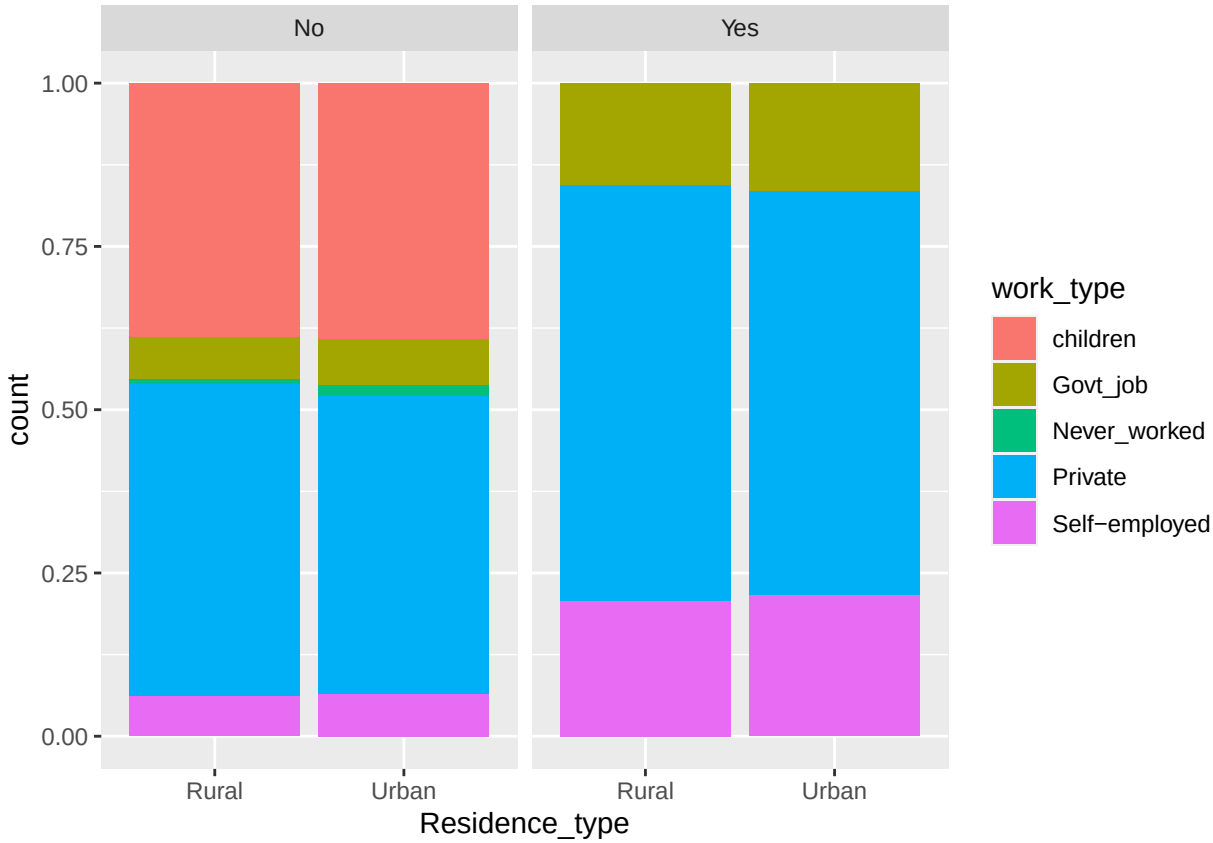
- En esta primera gráfica podemos identificar que de las personas que sufren de hipertensión la mayoría son hombres.
- También se puede distinguir que la hipertensión comienza a aparecer desde los adolescentes.
- La hipertensión es más probable encontrarla en adultos mayores.
- Además, se nota que solo Adultos Jóvenes se autoidentifican como de otro género.

```
ggplot(data = nuevo_dataset[1:filas,],aes(x=hypertension,fill=heart_disease))+geom_bar(position="fill")+
```



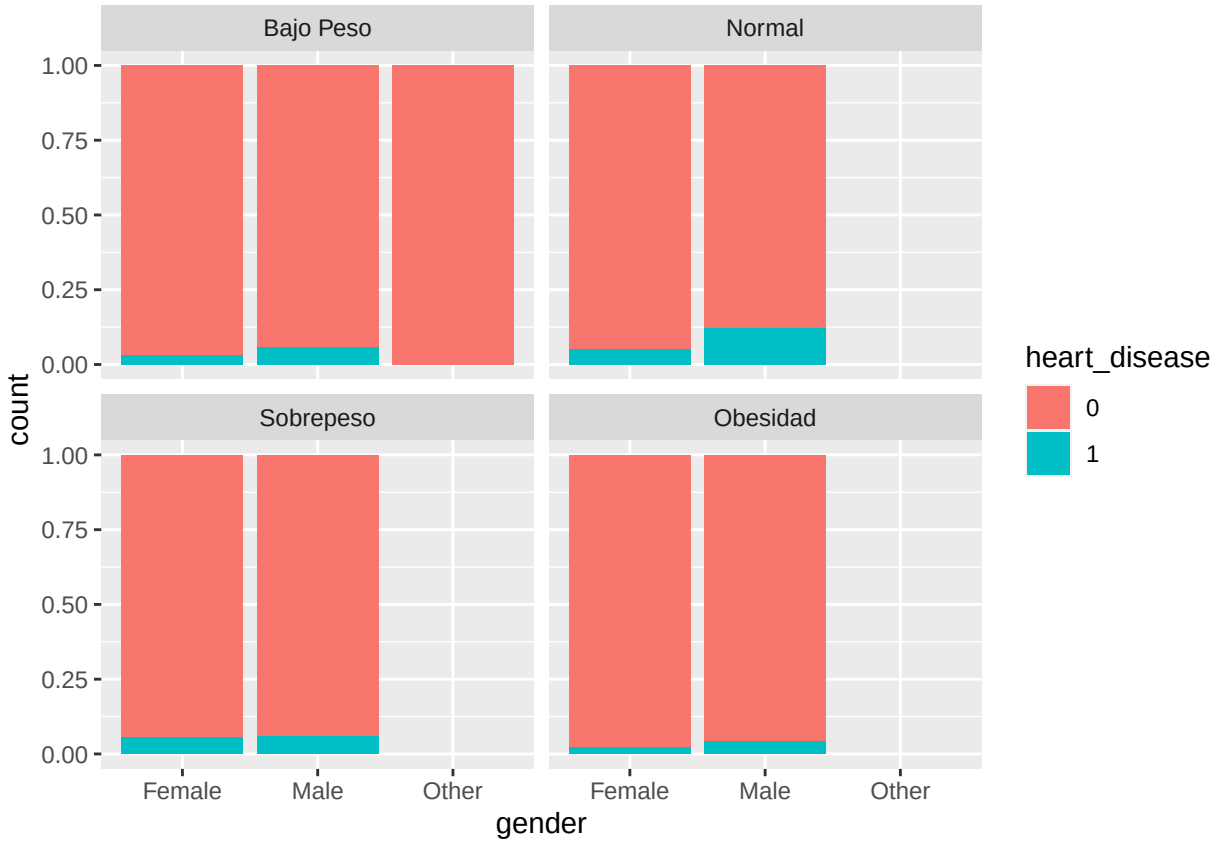
- En los datos obtenidos los niños no presentan enfermedades del corazón ni hipertensión.
- Adolecentes y Adultos Jóvenes no exiten datos de enfermedades del corazón, sin embargo si presentan hipertensión.
- En Adultos Madutos y Mayores aparecen las enfermedades del corazón, según los datos, es probable que una persona con hipertensión también padezca de enfermedades del corazón.

```
ggplot(data = nuevo_dataset[1:filas,],aes(x=Residence_type ,fill=work_type))+geom_bar(position="fill")+
```



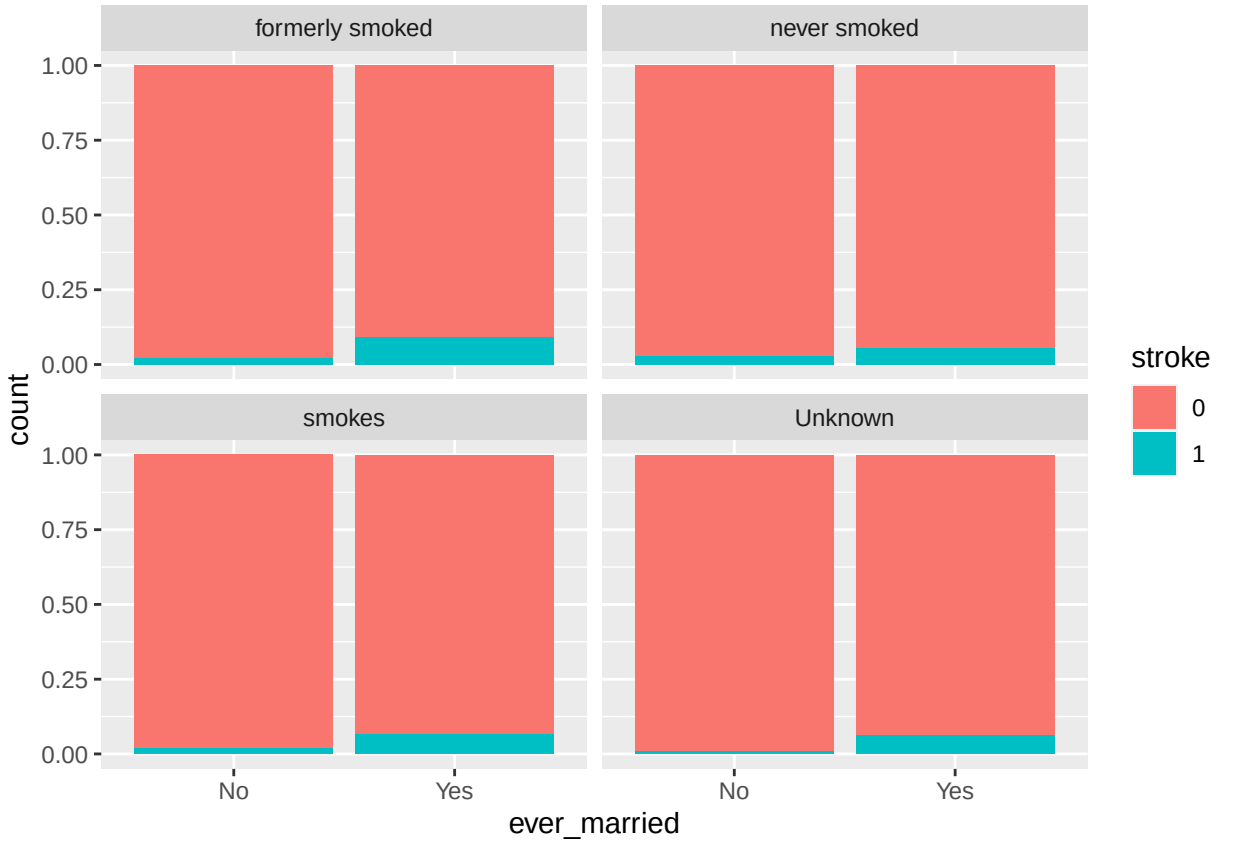
- El trabajo en el sector privado es el más frecuente, mientras que personas que nunca han trabajado aparece en menor frecuencia.
- Las personas que no se han casado es menos probable encontrarlas en el sector del gobierno.
- Las personas casadas tienden a acaparar más plazas de trabajo en el sector privado, además, hay más casados que son trabajadores independientes que los solteros.
- Los niños pueden vivir en el sector urbano o rural, esta variable no es distintiva.

```
ggplot(data = nuevo_dataset[1:filas,],aes(x=gender ,fill=heart_disease))+geom_bar(position="fill")+face
```



- Todos los datos correspondientes a otro género representan a personas de bajo peso.
- La obesidad en hombres aparece con mayor frecuencia con enfermedades del corazón.

```
ggplot(data = nuevo_dataset[1:filas,],aes(x=ever_married,fill=stroke))+geom_bar(position="fill")+facet_
```



- Lo más frecuente es encontrar personas fumadoras que son casadas, además están asociadas con stroke.
- De las peronas que nunca fuman, es más probable que estén marcadas con stroke

Tablas

```
t<-table(nuevo_dataset[1:filas,]$age,nuevo_dataset[1:filas,]$bmi)
for (i in 1:dim(t)[1]){
  t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

```
##
##                   Bajo Peso     Normal  Sobrepeso   Obesidad
##   Niños           96.7687075  2.8911565  0.3401361  0.0000000
##   Adolecentes     76.1904762 14.8148148  4.2328042  4.7619048
##   Adultos Jóvenes 58.8424437 21.9721329 10.3965702  8.7888532
##   Adultos Maduros 41.6501650 32.7392739 13.3993399 12.2112211
##   Adultos Mayores 44.7523585 37.3231132 11.6155660  6.3089623
```

- La mayoría de los niños tiene bajo peso y no presentan obesidad.
- De las personas con obesidad es más probable encontrarlas en adultos maduros (35-55 años).
- De las personas con peso normal es más probable encontrarlas en adultos mayores.

```
t<-table(nuevo_dataset[1:filas,]$smoking_status,nuevo_dataset[1:filas,]$hypertension)
for (i in 1:dim(t)[1]){
  t[i,]<-t[i,]/sum(t[i,])*100
}
```

```
t
```

```
##
##                            0         1
##   formerly smoked 86.440678 13.559322
##   never smoked    87.737844 12.262156
##   smokes          88.086185 11.913815
##   Unknown         96.632124  3.367876
```

- De las personas que presentan hipertensión lo más probable es que fumen.
- La mayoría de personas de los datos no sufren de hipertensión

```
t<-table(nuevo_dataset[1:filas,]$gender,nuevo_dataset[1:filas,]$avg_glucose_level)
for (i in 1:dim(t)[1]){
  t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

```
##
##                Bajo      Medio       Alto
##   Female  82.965932   9.218437   7.815631
##   Male    79.054374  11.205674   9.739953
##   Other    0.000000 100.000000   0.000000
```

- Las personas identificadas como género otro tienen un nivel medio de glucosa en la sangre.
- Es más probable encontrar hombres con alto nivel de glucosa que mujeres.
- Es más probable encontrar mujeres con bajo nivel de glucosa que hombres

Exportación de los datos limpios y procesados:

```
write.csv(nuevo_dataset, file="datos_preprocesados.csv",sep = ";",row.names=FALSE)
```

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

En funcion de la informacion del dataset y de acuerdo a los analisis realizados, se concluye que las variables mas importantes en la incidencia de padecer un AVC son la edad, la hipertension y los niveles de glucosa. En ese sentido este dataset nos ha ayudado a responder el problema y en base a lo cual se podra definir programas especificos de control de glucosa e hipertension en las pacientes para disminuir la incidencia de un AVC.

La utilidad de la información es evidente, ya que permite identificar grupos vulnerables, zonas, características sobres las cuales se debe prevenir o crear campañas contra enfermedades o condiciones tales como problemas del corazón, obesidad, hipertensión y sus derivados.

Se identifica que los hombres tienden a generar hipertensión con mayor frecuencia que las mujeres, según los datos estos problemas comienzan a aparecer desde la adolescencia, sin embargo, lo más probable es que encontremos adultos mayores con problemas de hipertención, además los datos indican que es más probable que una persona que tenga hipertensión también padezca de alguna enfermedad del corazón.

La condición física de las personas registradas nos indica que las personas que no se identifican como hombres o mujeres tienden a presentar bajo peso, además se obtiene que la obesidad en hombres aparece con mayor frecuencia con enfermedades del corazón. También se encuentra que los niños tienen bajo peso, con este conocimiento el entorno de la salud puede ejecutar campañas o programas de consciencia en esta sección. Con respecto a esta misma línes se identifica que lo más probable es encontrar mujeres con bajo nivel de glucosa, esto notándose que la relación es inversa en el género, es decir, que los hombres con más frecuencia que las mujeres concentran altos niveles de glucosa.

En las condiciones sociales se tiene que lo más probable es que una persona fumadora sea casada y esta presente hipertensión.

. Contribución de los Integrantes

```
library(kableExtra)
contribuciones = data.frame(stringsAsFactors=FALSE,
Contribuciones = c("Investigación previa ", "Redacción de las respuestas", "Desarrollo código"),
Firma = c("Andrea Martínez/Richard Jácome", "Andrea Martínez/Richard Jácome", "Andrea Martínez/Richard J
)
```

```
kbl(contribuciones) %>%
kable_paper("hover",
               full_width = F)
```

Contribuciones	Firma
Investigación previa	Andrea Martínez/Richard Jácome
Redacción de las respuestas	Andrea Martínez/Richard Jácome
Desarrollo código	Andrea Martínez/Richard Jácome

. Referencias

Organizacion Mundial de la Salud. Enfermedades cardiovasculares. [Fecha de consulta: 01 de junio del 2021]. https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Centro Nacional para la Prevención de Enfermedades Crónicas y Promoción de la Salud, División de Nutrición, Actividad Física, y Obesidad. IMC Índice de masa corporal. [Fecha de consulta: 05 de junio del 2021]. https://www.cdc.gov/healthyweight/spanish/assessing/index.html

Durán, Xavier. (2020, febrero). Limpieza del conjunto de datos de R. Catalunya: Universitat Oberta de Catalunya

Petry NM.A comparison of young, middle-aged, and older adult treatment-seeking pathological gamblers. [Fecha de consulta: 05 de junio del 2021]. https://www.ncbi.nlm.nih.gov/pubmed/11815703.

R Documentation. Uso de la función cut. Disponible en línea. [Fecha de consulta: 05 de junio del 2020]. https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cut.

R Pubs. Tablas en R Markdown. Disponible en línea. [Fecha de consulta: 05 de junio del 2020].https://rpubs.com/Juve_Campos/tablasRMarkdown