Análise estatística descritiva

Parte 1 - Visualização do dataset AVC

Importando o dataset

In [ ]:
```
location <- 'stroke-data.csv'
data0 <- read.csv(location)
```

Visualizando a estrutura interna do dataset

In [ ]:
```
str(data0)
```

```
'data.frame':   5110 obs. of  12 variables:
 $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27
419 60491 ...
 $ gender           : chr  "Male" "Female" "Male" "Female" ...
 $ 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     : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ work_type        : chr  "Private" "Self-employed" "Private" "Private"
...
 $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
 $ avg_glucose_level: num  229 202 106 171 174 ...
 $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
 $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked"
"smokes" ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

Descrição das colunas

- id: identificador único
- gender: "Male", "Female" ou "Other" (Masculino, Feminino ou Outro)
- age: idade do paciente
- hypertension: 0 se o paciente não apresenta hipertensão, 1 se o paciente apresenta hipertensão
- heart_disease: 0 se o paciente não apresenta nenhuma doença cárdiaca, 1 se o paciente apresenta alguma doença cárdiaca
- ever_married: Paciente já foi casado? "No" ou "Yes" (Não ou Sim)

- work_type: tipo de trabalho do paciente "children", "Govt_job", "Never_worked", "Private" or "Self-employed" (Criança, Funcionário público, nunca trabalhou, iniciativa privada ou autônomo)
- Residence_type: tipo de residência: "Rural" ou "Urban" (Zona rural ou Urbana)
- avg_glucose_level: Nível de glicose médio no sangue
- bmi: Índice de Massa Corporal
- smoking_status: representa se o paciente é fumante. "formerly smoked", "never smoked", "smokes" or "Unknown" * (Ex-fumante, nunca fumou, fuma, desconhecido)
- stroke: 1 se o paciente sofreu AVC ou 0 se o paciente não sofreu AVC.
- Obs: "Unknown" na coluna smoking_status significa que a informação está indisponível para este paciente.

Estatísticas de sumário do dataset

In [ ]:
```
summary(data0)
```

```
       id           gender                age         hypertension    
 Min.   :   67   Length:5110        Min.   : 0.08   Min.   :0.00000  
 1st Qu.:17741   Class :character   1st Qu.:25.00   1st Qu.:0.00000  
 Median :36932   Mode  :character   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   Length:5110        Length:5110        Length:5110       
 1st Qu.:0.00000   Class :character   Class :character   Class :character  
 Median :0.00000   Mode  :character   Mode  :character   Mode  :character  
 Mean   :0.05401                                                          
 3rd Qu.:0.00000                                                          
 Max.   :1.00000                                                          
 avg_glucose_level     bmi            smoking_status         stroke        
 Min.   : 55.12    Length:5110        Length:5110        Min.   :0.00000  
 1st Qu.: 77.25    Class :character   Class :character   1st Qu.:0.00000  
 Median : 91.89    Mode  :character   Mode  :character   Median :0.00000  
 Mean   :106.15                                          Mean   :0.04873  
 3rd Qu.:114.09                                          3rd Qu.:0.00000  
 Max.   :271.74                                          Max.   :1.00000  
```

In [ ]:
```
# duplica o dataset
data1 <- data0
```

In [ ]:
```
# Visualizando todas as variáveis categóricas(qualitativas)

lapply(subset(data1, select = c(gender, ever_married, work_type, Residence_t
```

$gender 'Male' · 'Female' · 'Other'

$ever_married 'Yes' · 'No'

$work_type 'Private' · 'Self-employed' · 'Govt_job' · 'children' · 'Never_worked'

$Residence_type 'Urban' · 'Rural'

$bmi '36.6' · 'N/A' · '32.5' · '34.4' · '24' · '29' · '27.4' · '22.8' · '24.2' · '29.7' · '36.8' · '27.3' · '28.2' · '30.9' · '37.5' · '25.8' · '37.8' · '22.4' · '48.9' · '26.6' · '27.2' · '23.5' · '28.3' · '44.2' · '25.4' · '22.2' · '30.5' · '26.5' · '33.7' · '23.1' · '32' · '29.9' · '23.9' · '28.5' · '26.4' · '20.2' · '33.6' · '38.6' · '39.2' · '27.7' · '31.4' · '36.5' · '33.2' · '32.8' · '40.4' · '25.3' · '30.2' · '47.5' · '20.3' · '30' · '28.9' · '28.1' · '31.1' · '21.7' · '27' · '24.1' · '45.9' · '44.1' · '22.9' · '29.1' · '32.3' · '41.1' · '25.6' · '29.8' · '26.3' · '26.2' · '29.4' · '24.4' · '28' · '28.8' · '34.6' · '19.4' · '30.3' · '41.5' · '22.6' · '56.6' · '27.1' · '31.3' · '31' · '31.7' · '35.8' · '28.4' · '20.1' · '26.7' · '38.7' · '34.9' · '25' · '23.8' · '21.8' · '27.5' · '24.6' · '32.9' · '26.1' · '31.9' · '34.1' · '36.9' · '37.3' · '45.7' · '34.2' · '23.6' · '22.3' · '37.1' · '45' · '25.5' · '30.8' · '37.4' · '34.5' · '27.9' · '29.5' · '46' · '42.5' · '35.5' · '26.9' · '45.5' · '31.5' · '33' · '23.4' · '30.7' · '20.5' · '21.5' · '40' · '28.6' · '42.2' · '29.6' · '35.4' · '16.9' · '26.8' · '39.3' · '32.6' · '35.9' · '21.2' · '42.4' · '40.5' · '36.7' · '29.3' · '19.6' · '18' · '17.6' · '19.1' · '50.1' · '17.7' · '54.6' · '35' · '22' · '39.4' · '19.7' · '22.5' · '25.2' · '41.8' · '60.9' · '23.7' · '24.5' · '31.2' · '16' · '31.6' · '25.1' · '24.8' · '18.3' · '20' · '19.5' · '36' · '35.3' · '40.1' · '43.1' · '21.4' · '34.3' · '27.6' · '16.5' · '24.3' · '25.7' · '21.9' · '38.4' · '25.9' · '54.7' · '18.6' · '24.9' · '48.2' · '20.7' · '39.5' · '23.3' · '64.8' · '35.1' · '43.6' · '21' · '47.3' · '16.6' · '21.6' · '15.5' · '35.6' · '16.7' · '41.9' · '16.4' · '17.1' · '29.2' · '37.9' · '44.6' · '39.6' · '40.3' · '41.6' · '39' · ⋯ · '22.1' · '26' · '44.3' · '51' · '39.7' · '34.7' · '21.3' · '41.2' · '34.8' · '19.2' · '35.7' · '40.8' · '24.7' · '19' · '32.4' · '34' · '28.7' · '32.1' · '51.5' · '20.4' · '30.6' · '71.9' · '19.3' · '40.9' · '17.2' · '16.1' · '16.2' · '40.6' · '18.4' · '21.1' · '42.3' · '32.2' · '50.2' · '17.5' · '18.7' · '42.1' · '47.8' · '20.8' · '30.1' · '17.3' · '36.4' · '12' · '36.2' · '55.7' · '14.4' · '43' · '41.7' · '33.8' · '43.9' · '22.7' · '57.5' · '37' · '38.5' · '16.3' · '44' · '32.7' · '54.2' · '40.2' · '33.3' · '17.4' · '41.3' · '52.3' · '14.6' · '17.8' · '46.1' · '33.1' · '18.1' · '43.8' · '50.3' · '38.9' · '43.7' · '39.9' · '15.9' · '19.8' · '12.3' · '78' · '38.3' · '41' · '42.6' · '43.4' · '15.1' · '20.6' · '33.5' · '43.2' · '30.4' · '38' · '33.4' · '44.9' · '44.7' · '37.6' · '39.8' · '53.4' · '55.2' · '42' · '37.2' · '42.8' · '18.8' · '42.9' · '14.3' · '37.7' · '48.4' · '50.6' · '46.2' · '49.5' · '43.3' · '33.9' · '18.5' · '44.5' · '45.4' · '55' · '54.8' · '19.9' · '17.9' · '15.6' · '52.8' · '15.2' · '66.8' · '55.1' · '18.2' · '48.5' · '55.9' · '57.3' · '10.3' · '14.1' · '15.7' · '56' · '44.8' · '13.4' · '51.8' · '38.1' · '57.7' · '44.4' · '38.8' · '49.3' · '39.1' · '54' · '56.1' · '97.6' · '53.9' · '13.7' · '11.5' · '41.4' · '14.2' · '49.4' · '15.4' · '45.1' · '49.2' · '48.7' · '53.8' · '42.7' · '48.8' · '52.7' · '53.5' · '50.5' · '15.8' · '45.3' · '14.8' · '51.9' · '63.3' · '40.7' · '61.2' · '48' ·

'46.8' · '48.3' · '58.1' · '50.4' · '11.3' · '12.8' · '13.5' · '14.5' · '15' · '59.7' · '47.4' · '52.5' · '13.2' · '52.9' · '61.6' · '49.9' · '54.3' · '47.9' · '13' · '13.9' · '50.9' · '57.2' · '64.4' · '92' · '50.8' · '57.9' · '45.8' · '47.6' · '14' · '46.4' · '46.9' · '47.1' · '13.3' · '48.1' · '51.7' · '46.3' · '54.1' · '14.9'

$smoking_status 'formerly smoked' · 'never smoked' · 'smokes' · 'Unknown'

In [ ]:
```
#Visualizando o sumário da coluna IMC
summary(data1$bmi)
```

```
   Length     Class      Mode
     5110 character character
```

In [ ]:
```
# Contagem das variáveis únicas na coluna gênero
table(data1$gender)
```

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

In [ ]:
```
# Contagem das variáveis únicas na coluna Fumante
table(data1$smoking_status)
```

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

In [ ]:
```
# Contagem das variáveis únicas na coluna AVC
table(data1$stroke)
```

```
   0    1
4861  249
```

In [ ]:
```
# Contagem das variáveis únicas na coluna Hipertensão
table(data1$hypertension)
```

```
   0    1
4612  498
```

In [ ]:
```
# Contagem das variáveis únicas na coluna Ataque cardiáco
table(data1$heart_disease)
```

```
   0    1
4834  276
```

In [ ]:
```
# Contagem das variáveis únicas na coluna Tipo de Trabalho
table(data1$work_type)
```

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

Parte 2 - Limpeza dos dados

In [ ]:
```
# Converter BMI para numérico

data1$bmi <- as.numeric(data1$bmi)
```

```
Warning message in eval(expr, envir, enclos):
“NAs introduced by coercion”
```

In [ ]:
```r
# Substituir N/As na coluna BMI pela média

data1$bmi[is.na(data1$bmi)] <- mean(data1$bmi,na.rm=TRUE)

# Visualizando o novo sumário da coluna BMI
summary(data1$bmi)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  10.30   23.80   28.40   28.89   32.80   97.60 
```

In [ ]:
```r
# como o gênero feminino é maioria, substituimos 'Other' para 'Female'
data1$gender <- ifelse(data1$gender == "Other", "Female", data1$gender)

# Contagem das variáveis únicas na coluna gênero revisada
table(data1$gender)
```

```
Female   Male 
  2995   2115 
```

In [ ]:
```r
# Calculamos a probabilidade de cada categoria de fumante, dado que só temos
prob.FS <- 885 / (885 + 1892 + 789)
prob.NS <- 1892 / (885 + 1892 + 789)
prob.S <- 789 / (885 + 1892 + 789)

# Duplica o dataset
data2 <- data1

# Subistituindo a categoria "Unknown" na coluna Fumante pelas outras 3 categ
library(tidyverse)

data2$rand <- runif(nrow(data2))
data2 <- data2%>%mutate(Probability = ifelse(rand <= prob.FS, "formerly smok
data2 <- data2%>%mutate(smoking.status = ifelse(smoking_status == "Unknown",

# Contagem das variáveis únicas na coluna gênero revisada
table(data2$smoking.status)
```

```
formerly smoked    never smoked          smokes 
           1258            2735            1117 
```

In [ ]:
```r
# Removendo as colunas que não são necessárias
data2 <- subset(data2, select = -c(rand,Probability,smoking_status,id))
```

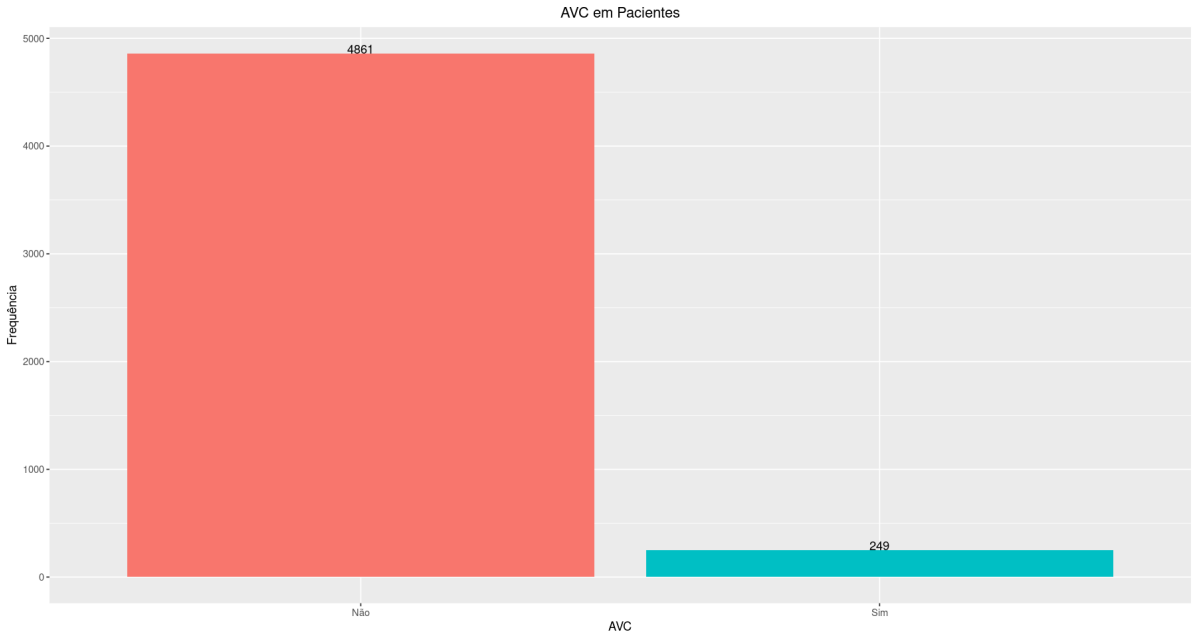
Parte 3 - Análise exploratória dos dados

Parte 3.1 - Gráfico de barras por coluna

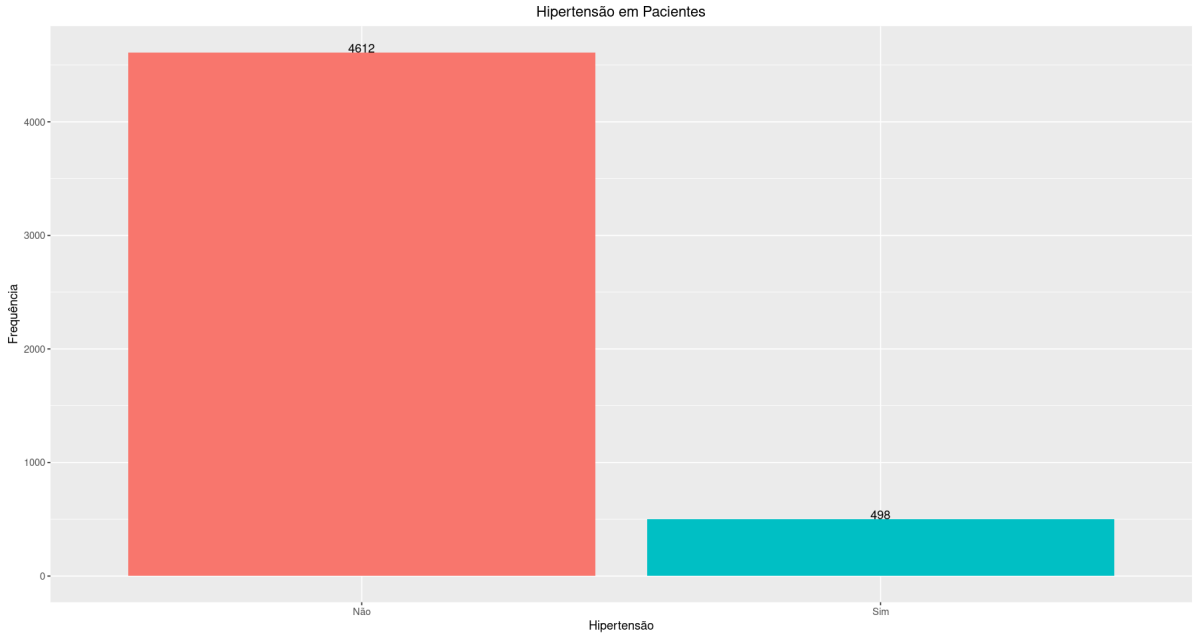
In [ ]:
```r
library("ggplot2")
```

In [ ]:
```r
# Cria a tabela de contagem de AVC
strokecounts <- as.data.frame(table(data2$stroke))
```

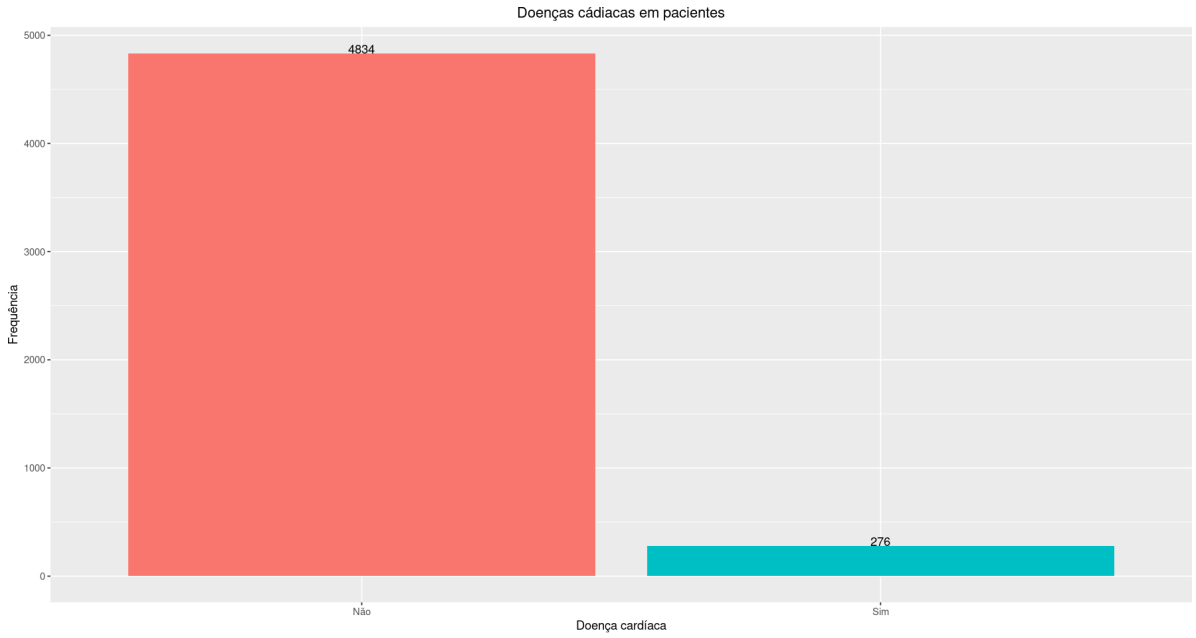
```
# Troca os valores 1 e 0 por Sim e Não, respectivamente
strokecounts$Var1 <- ifelse(strokecounts$Var1 == 1, "Sim", 'Não')

# Gráfico de Barras dos Pacientes que sofreram e não sofreram AVC
ggplot(strokecounts, aes(x = Var1, y = Freq, fill = Var1)) +
        geom_bar(stat = "identity") + theme(legend.position="none") +
        geom_text(aes(label = Freq), vjust = 0) +
        labs(title="AVC em Pacientes",x ="AVC", y = "Frequência") +
        theme(plot.title = element_text(hjust = 0.5))
```



In [ ]:
```
# Cria a tabela de contagem de Hipertensão
hypercounts <- as.data.frame(table(data2$hypertension))

# Troca os valores 1 e 0 por Sim e Não, respectivamente
hypercounts$Var1 <- ifelse(hypercounts$Var1 == 0, "Não", 'Sim')

# Gráfico de Barras para a coluna hipertensão : Não / Sim
ggplot(hypercounts, aes(x = Var1, y = Freq, fill = Var1)) +
        geom_bar(stat = "identity") + theme(legend.position="none") +
        geom_text(aes(label = Freq), vjust = 0) +
        labs(title="Hipertensão em Pacientes",x ="Hipertensão", y = "Frequêr
        theme(plot.title = element_text(hjust = 0.5))
```



In [ ]:
```
# Cria a tabela de contagem de Doenças Cardíacas

heartcounts <- as.data.frame(table(data2$heart_disease))


# Troca os valores 1 e 0 por Sim e Não, respectivamente
heartcounts$Var1 <- ifelse(heartcounts$Var1 == 0, "Não", 'Sim')

# Gráfico de Barras para a coluna Doenças Cárdiacas: Não / Sim
ggplot(heartcounts, aes(x = Var1, y = Freq, fill = Var1)) +
        geom_bar(stat = "identity") + theme(legend.position="none") +
        geom_text(aes(label = Freq), vjust = 0) +
        labs(title="Doenças cádiacas em pacientes",x ="Doença cardíaca", y =
        theme(plot.title = element_text(hjust = 0.5))
```

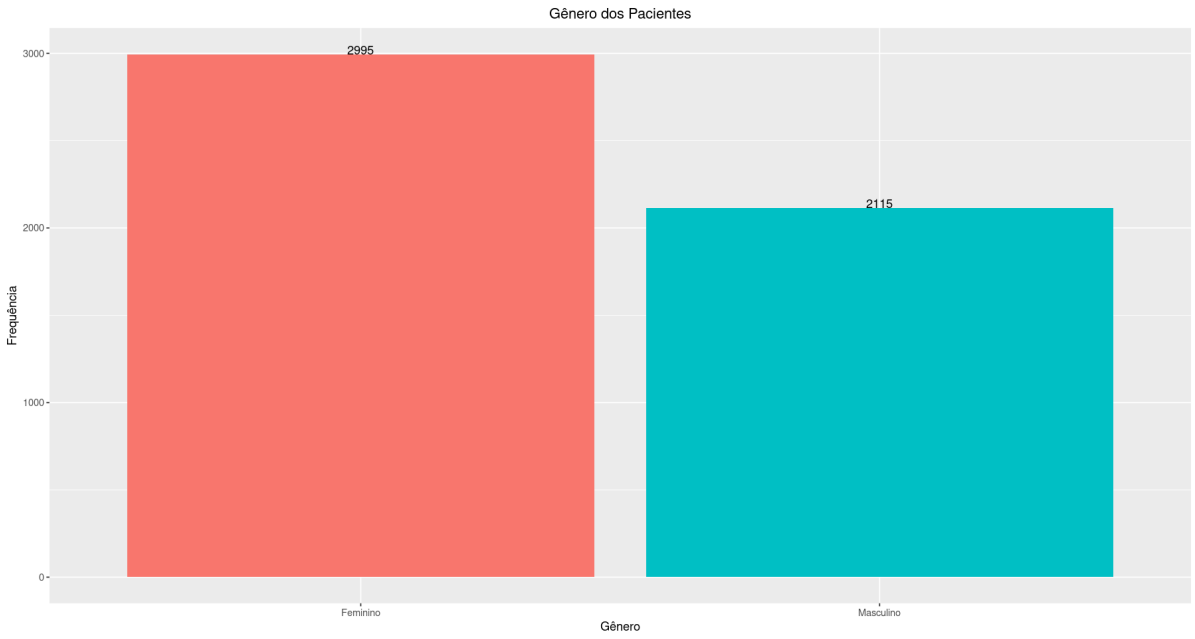


- A quantidade de pacientes que não sofreram AVC é alta em relação aos que sofreram AVC.
- A quantidade de pacientes que não sofrem de Hipertensão é alta em relação aos que sofrem Hipertensão,

mas a diferença é relativamente menor que a diferença vista em vítimas de AVC.

- A diferença entre pacientes com e sem doênças cardíacas se aproxima da diferença entre os pacientes que tiveram e não tiveram AVC.

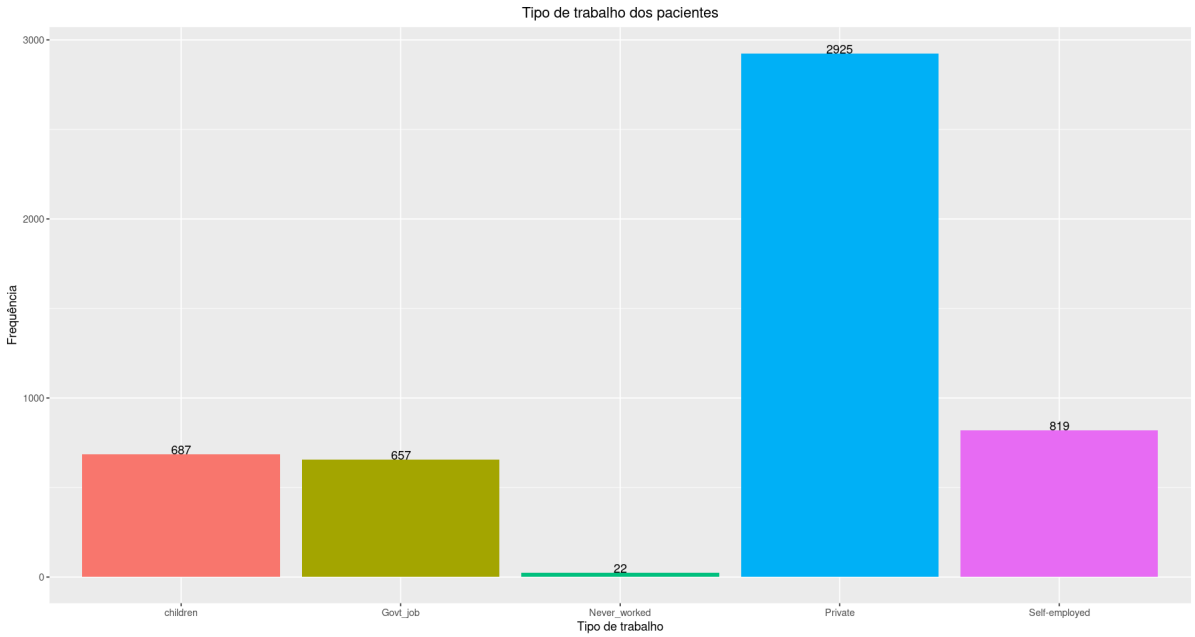
In [ ]:
```
# Cria a tabela de contagem de Gênero

gendercounts <- as.data.frame(table(data2$gender))

# Troca os valores "Male" e "Female" por Masculino e Feminino, respectivamer
gendercounts$Var1 <- ifelse(gendercounts$Var1 == "Male", "Masculino", 'Femir

# Gráfico de Barras para a coluna Gênero

ggplot(gendercounts, aes(x = Var1, y = Freq, fill = Var1)) +
        geom_bar(stat = "identity") + theme(legend.position="none") +
        geom_text(aes(label = Freq), vjust = 0) +
        labs(title="Gênero dos Pacientes",x ="Gênero", y = "Frequência") +
        theme(plot.title = element_text(hjust = 0.5))
```



- Há mais pacientes do gênero feminino que pacientes do gênero masculino.
- O único paciente que estava listado como "Other" foi adicionado na categoria "Female", dado que há mais pacientes do gênero feminino.

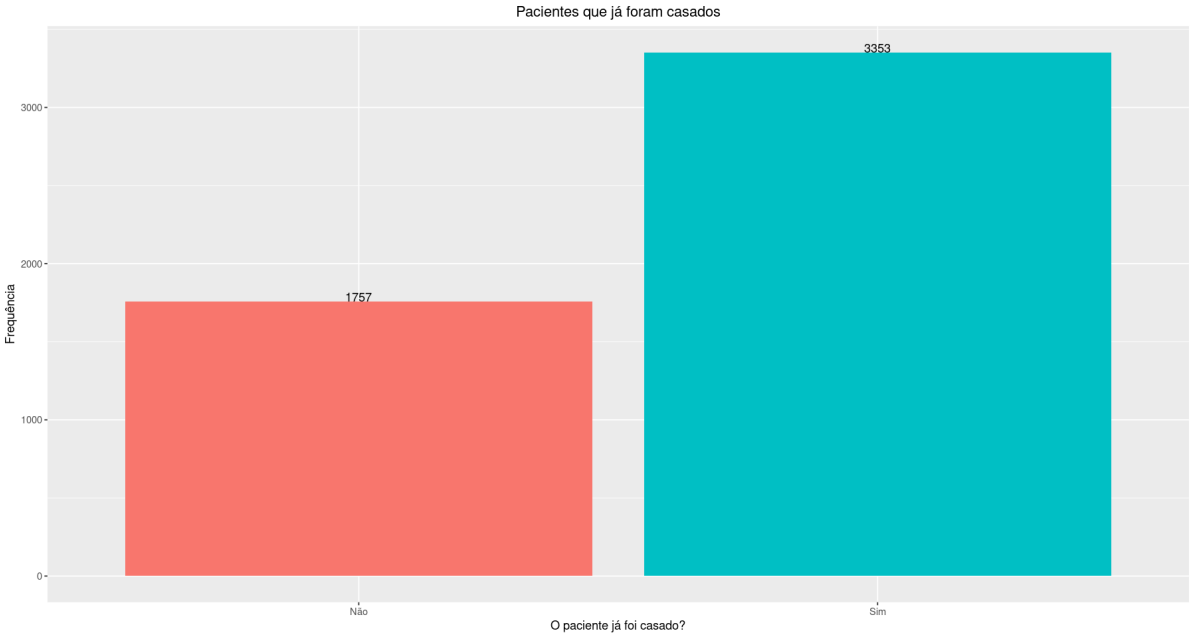
In [ ]:
```
# Cria a tabela de contagem da coluna Tipo de trabalho

workcounts <- as.data.frame(table(data2$work_type))
```

```
# Gráfico de Barras para a coluna tipo de trabalho

ggplot(workcounts, aes(x = Var1, y = Freq, fill = Var1)) +
        geom_bar(stat = "identity") + theme(legend.position="none") +
        geom_text(aes(label = Freq), vjust = 0) +
        labs(title="Tipo de trabalho dos pacientes",x ="Tipo de trabalho", y
        theme(plot.title = element_text(hjust = 0.5))
```



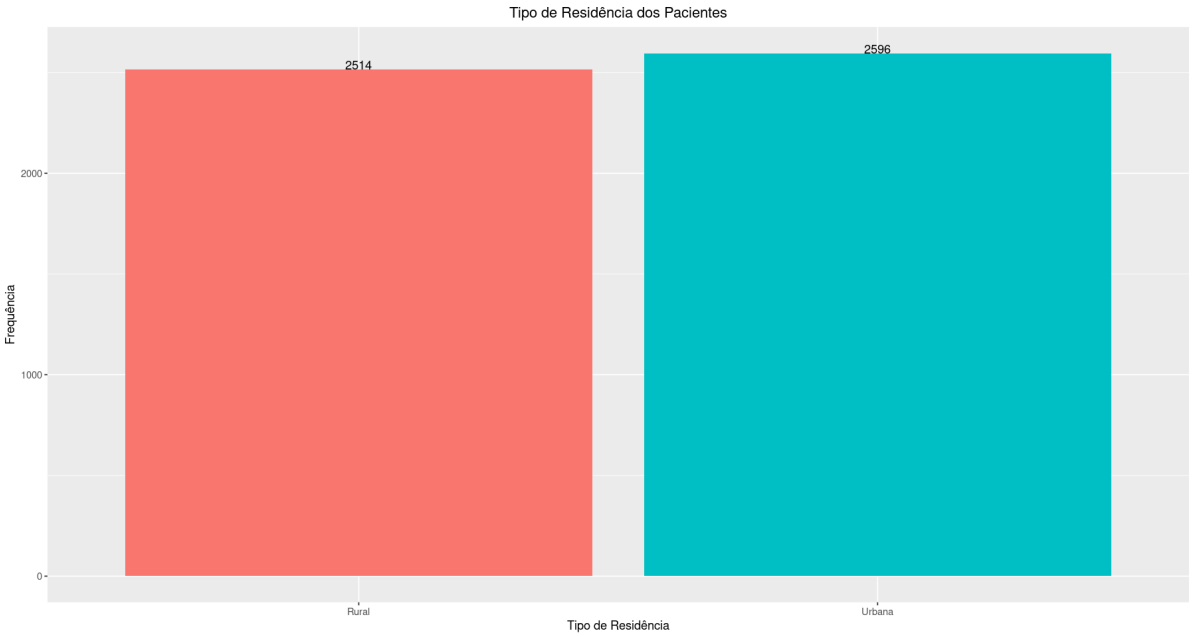
- A quantidade de pacientes que trabalham em cargos públicos, autônomos e crianças se aproximam.
- A maioria dos pacientes trabalham na iniciativa privada e a minoria nunca trabalhou.

In [ ]:
```
# Cria a tabela de contagem da coluna de pacientes que já se casaram
marriedcounts <- as.data.frame(table(data2$ever_married))


# Troca os valores "Yes" e "No" por Sim e Não, respectivamente
marriedcounts$Var1 <- ifelse(marriedcounts$Var1 == "No", "Não", 'Sim')

# Gráfico de Barras para a coluna de pacientes que já foram casados
ggplot(marriedcounts, aes(x = Var1, y = Freq, fill = Var1)) +
        geom_bar(stat = "identity") + theme(legend.position="none") +
        geom_text(aes(label = Freq), vjust = 0) +
        labs(title="Pacientes que já foram casados",x ="O paciente já foi ca
        theme(plot.title = element_text(hjust = 0.5))
```



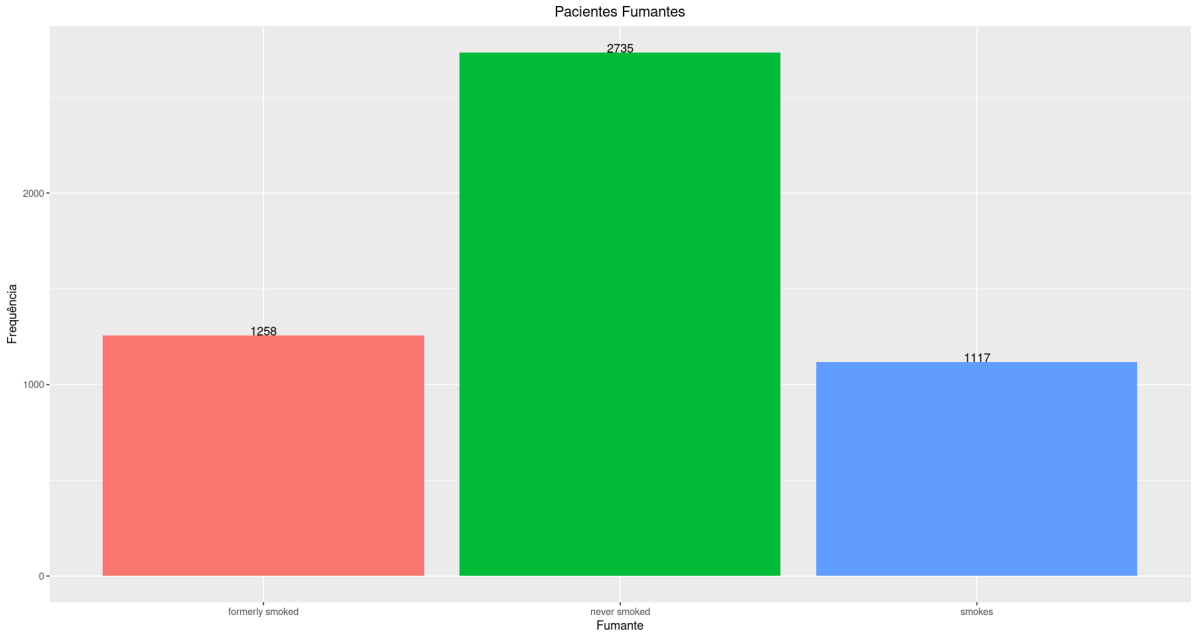
- Há aproximadamente o dobro de paciências que foram casados em relação aos que não foram.

In [ ]:
```
# Cria a tabela de contagem da coluna Tipo de Residência
rescounts <- as.data.frame(table(data2$Residence_type))

# Troca os valores "Urban" por Urbana
rescounts$Var1 <- ifelse(rescounts$Var1 == "Urban", "Urbana", 'Rural')

# Gráfico de Barras para a coluna Tipo de Residência
ggplot(rescounts, aes(x = Var1, y = Freq, fill = Var1)) +
        geom_bar(stat = "identity") + theme(legend.position="none") +
        geom_text(aes(label = Freq), vjust = 0) +
        labs(title="Tipo de Residência dos Pacientes",x ="Tipo de Residência
        theme(plot.title = element_text(hjust = 0.5))
```



- Os pacientes estão aproxidamente divididos por igual em residências rurais e urbanas.

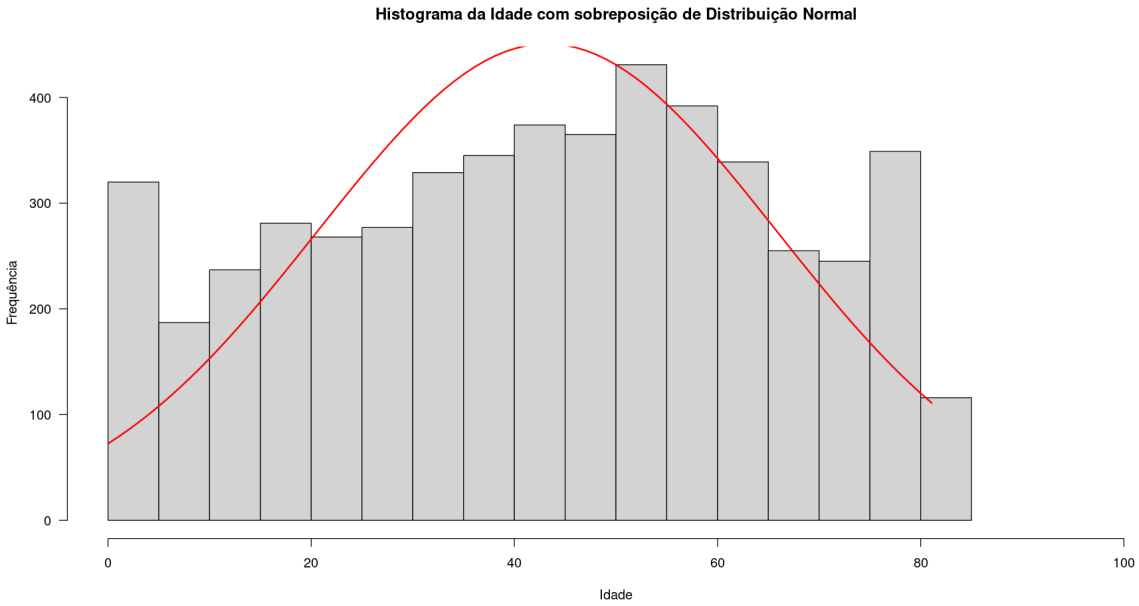
In [ ]:
```r
# Cria a tabela de contagem da coluna Fumante
smokecounts <- as.data.frame(table(data2$smoking.status))

# Gráfico de Barras para a coluna Fumante
ggplot(smokecounts, aes(x = Var1, y = Freq, fill = Var1)) +
        geom_bar(stat = "identity") + theme(legend.position="none") +
        geom_text(aes(label = Freq), vjust = 0) +
        labs(title="Pacientes Fumantes",x ="Fumante", y = "Frequência") +
        theme(plot.title = element_text(hjust = 0.5))
```



- Os dados desconhecidos foram aleatoriamente adicionados nas três categorias baseado em suas probabilidades.
- A maioria dos pacientes nunca fumaram
- A quantidade de pacientes ex-fumante e pacientes que fumam estão próximas.

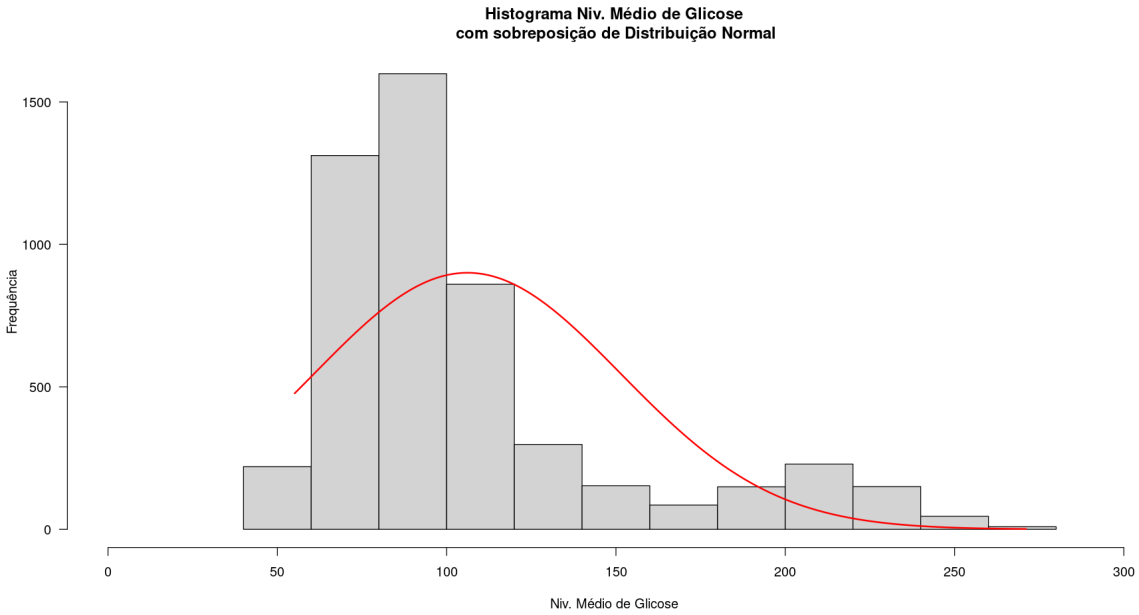
Parte 3.2 - Histogramas

In [ ]:
```r
# Histograma da coluna Idade com sobreposição de Distribuição Normal (Gaussi

histage <- hist(data2$age,xlim=c(0,100),
                main="Histograma da Idade com sobreposição de Distribuição N
                xlab="Idade",las=1,
                ylab="Frequência", las =1)
xfit <- seq(min(data2$age),max(data2$age))
yfit <- dnorm(xfit,mean=mean(data2$age),sd=sd(data2$age))
yfit <- yfit*diff(histage$mids[2:4])*length(data2$age)
lines(xfit,yfit,col="red",lwd=2)
```



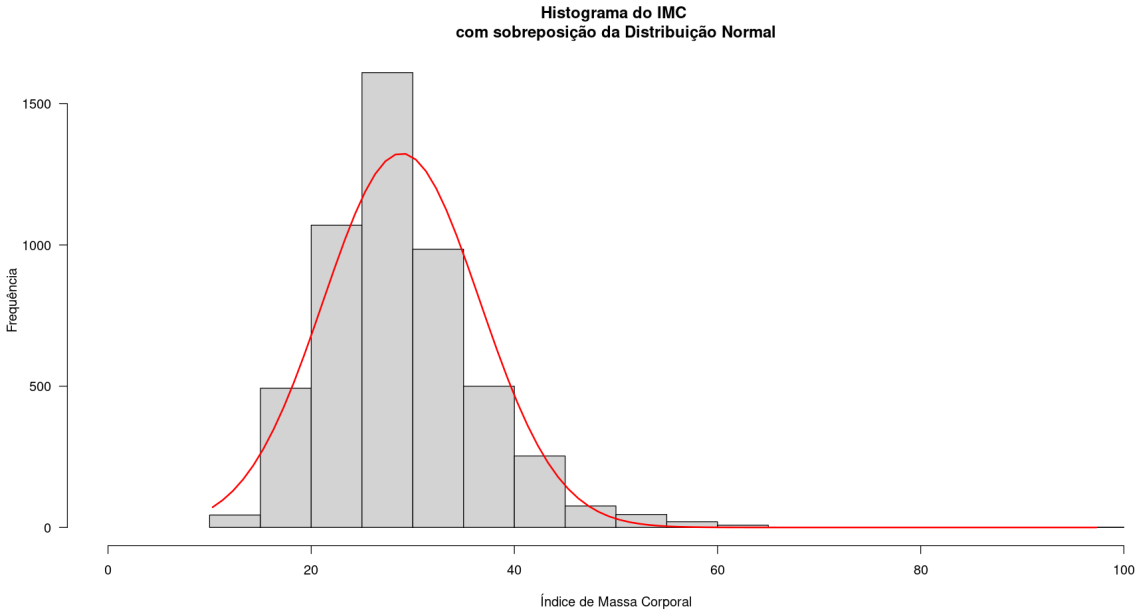
- As idades dos pacientes no estudo estão próximas de uma distribuição normal, com uma média de idade de 43,23 retirado da função summary().
- Baseado na informação do sumário anteriormente e no gráfico acima, a maioria dos pacientes está em torno dos seus 40 anos.

In [ ]:
```
# Histograma da coluna Nível Médio de Glicose com sobreposição de Distribuiç

histglucose <- hist(data2$avg_glucose_level,xlim=c(0,300),
                    main="Histograma Niv. Médio de Glicose \ncom sobreposição de
                    xlab="Niv. Médio de Glicose ",las=1,
                    ylab="Frequência", las =1)
xfit <- seq(min(data2$avg_glucose_level),max(data2$avg_glucose_level))
yfit <- dnorm(xfit,mean=mean(data2$avg_glucose_level),sd=sd(data2$avg_glucos
yfit <- yfit*diff(histglucose$mids[1:2])*length(data2$avg_glucose_level)
lines(xfit,yfit,col="red",lwd=2)
```



- Os níveis médios de glicose dos pacientes no estudo estão enviesados para a direita, com média de 106,15 retirado da função summary() anteriormente.

In [ ]:
```
# Histograma do IMC com sobreposição de Distribuição Normal.
histbmi <- hist(data2$bmi,xlim=c(0,100),
                main="Histograma do IMC \ncom sobreposição da Distribuição N
                xlab="Índice de Massa Corporal",las=1,
                ylab="Frequência", las=1)
xfit <- seq(min(data2$bmi),max(data2$bmi))
yfit <- dnorm(xfit,mean=mean(data2$bmi),sd=sd(data2$bmi))
yfit <- yfit*diff(histbmi$mids[1:2])*length(data2$bmi)
lines(xfit,yfit,col="red",lwd=2)
```



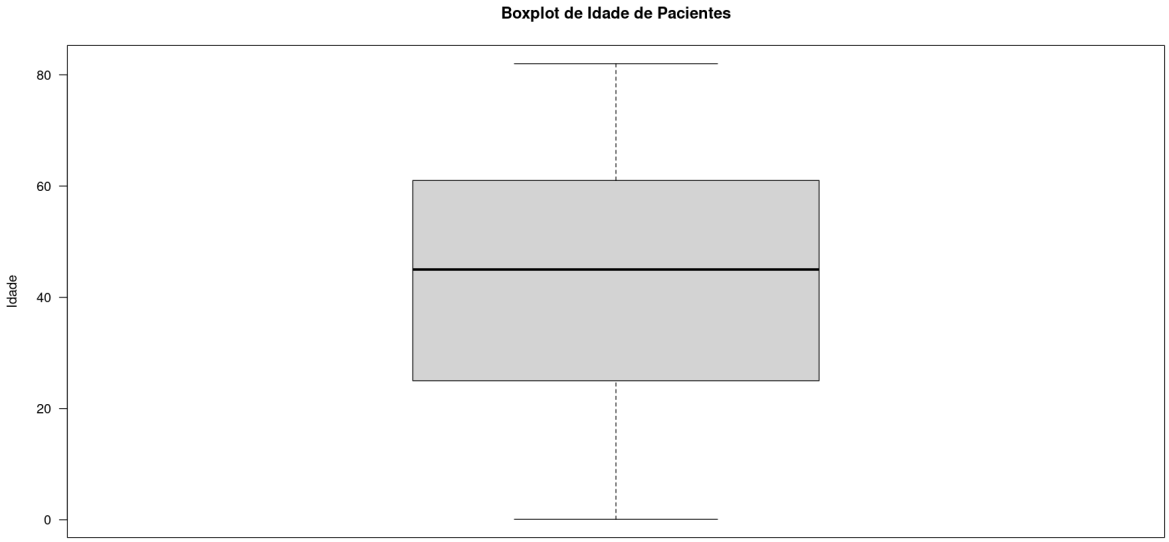
- Os dados de IMC dos pacientes estão enviesados para a direita, com uma média de

28.89 retirado da função summary(), após a modificação.

- Todos os registros "NA" foram atualizados para a média na etapa de Limpeza dos Dados, Parte 2 do estudo.

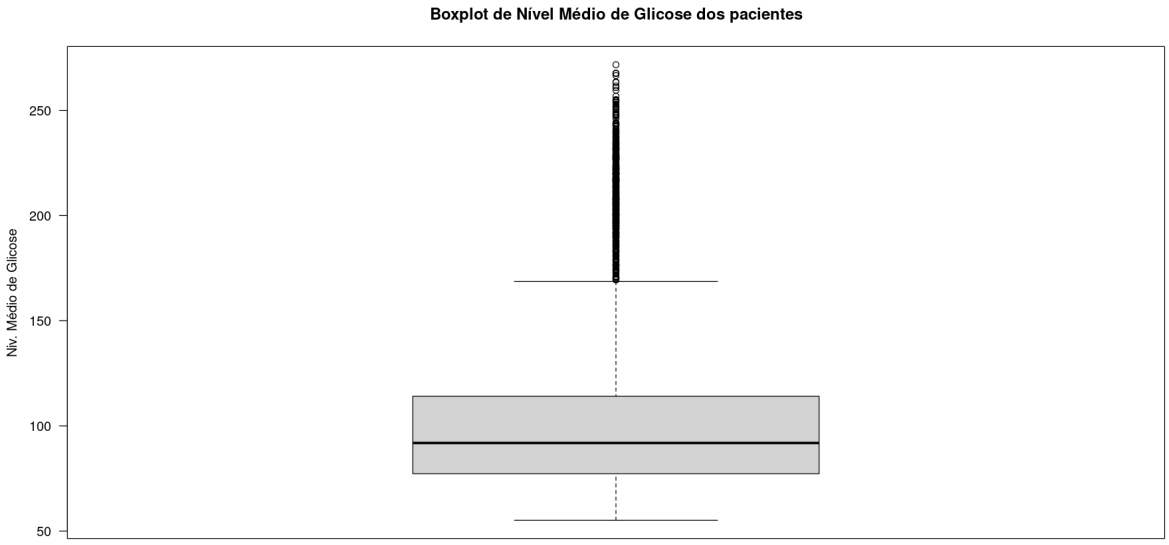
Parte 3.3 - Gráficos de Caixa (Boxplot)

In [ ]:
```
# Boxplot da Idade de Pacientes
boxplot(data2$age,main="Boxplot de Idade de Pacientes",ylab="Idade",las=1)
```



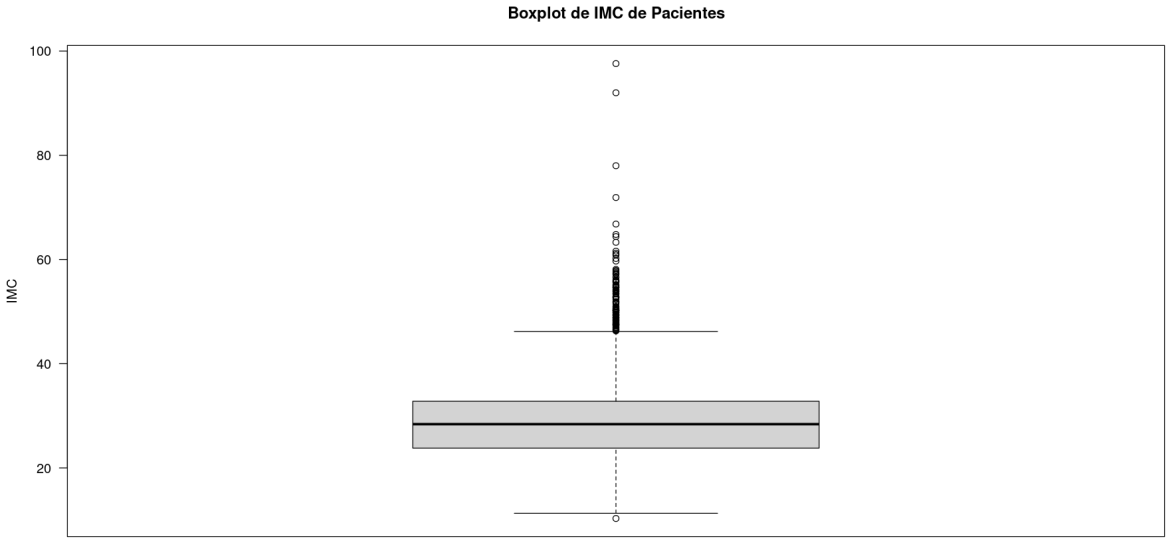
- O boxplot da Idade dos Pacientes não mostra dados potencialmente discrepantes (outliers)

In [ ]:
```
# Boxplot de Nível Médio de Glicose dos pacientes.

boxplot(data2$avg_glucose_level,main="Boxplot de Nível Médio de Glicose dos
        ylab="Niv. Médio de Glicose",las=1)
```



- O Boxplot mostra muitos dados discrepantes em potencial na parte superior do Nível Médio de Glicose dos pacientes.

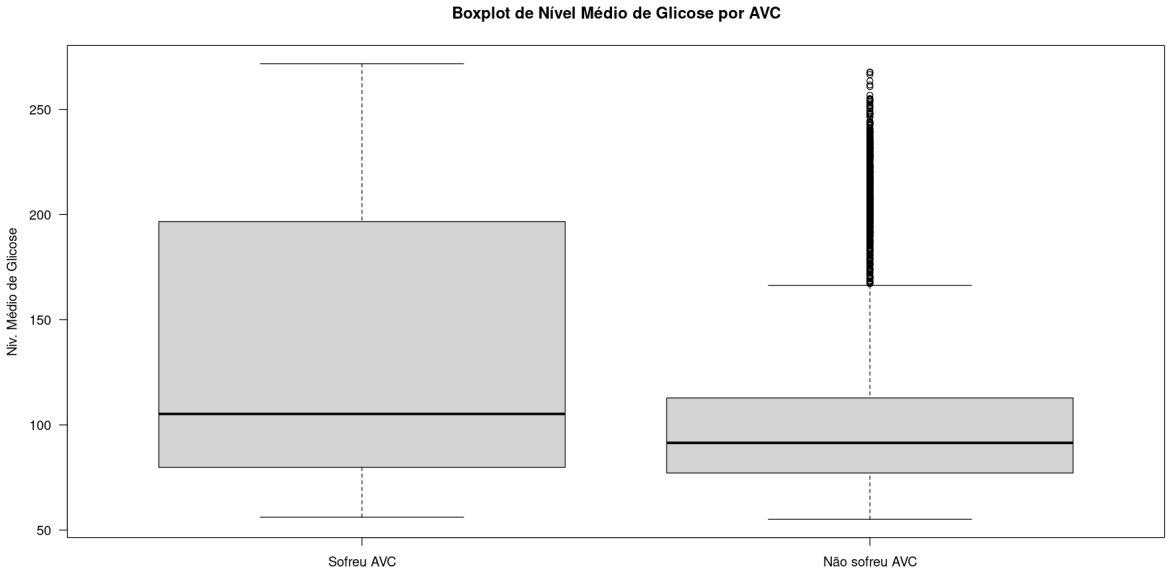
In [ ]:
```
# Boxplot de Índice de Massa Corporal dos pacientes.

boxplot(data2$bmi,main="Boxplot de IMC de Pacientes",ylab="IMC",las=1)
```



- O Boxplot mostra muitos dados discrepantes em potencial na parte superior do Índice de Massa Corporal dos pacientes.

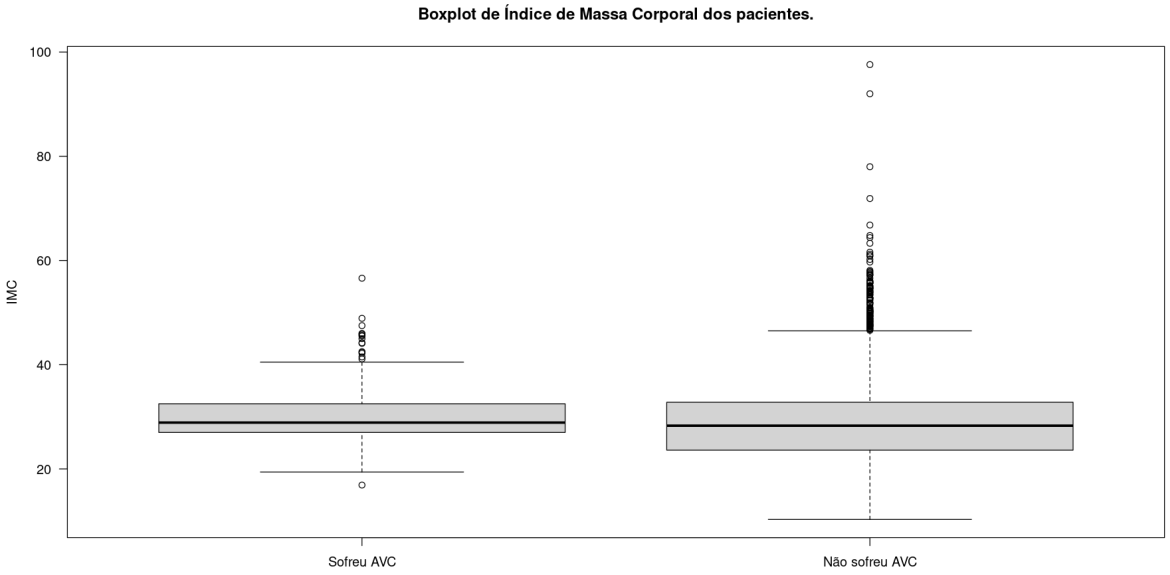
In [ ]:
```
# Subconjunto de Pacientes que sofreram e que não sofreram AVC em "Yes" e "N
Yes <- subset(data2, stroke == '1')
No <- subset(data2, stroke == '0')
```

In [ ]:
```
# Boxplot de Nível Médio de Glicose em pacientes que sofreram e não sofreram
boxplot(Yes$avg_glucose_level,No$avg_glucose_level,
        main="Boxplot de Nível Médio de Glicose por AVC",
        ylab="Niv. Médio de Glicose",las=1,names=c("Sofreu AVC","Não sofreu
```



- O Boxplot mostra a mediana e primeiro quartil relativamente semelhante para nível médio de glicose em pacientes que sofreram e que não sofreram AVC.
- Há muitos outliers entre os pacientes que não sofreram AVC.

In [ ]:
```
# Boxplot de Índice de Massa Corporal em pacientes que sofreram e não sofrer

boxplot(Yes$bmi,No$bmi,main="Boxplot de Índice de Massa Corporal dos pacient
        ylab="IMC",las=1,names=c("Sofreu AVC","Não sofreu AVC"))
```



- O Boxplot mostra a mediana e terceiro quartil relativamente similares para o índice de massa corporal de pacientes que sofreram e que não sofreram AVC.
- Há poucos outliers entre pacientes que sofreram AVC
- Há muitos outliers entre os pacientes que não sofreram AVC.

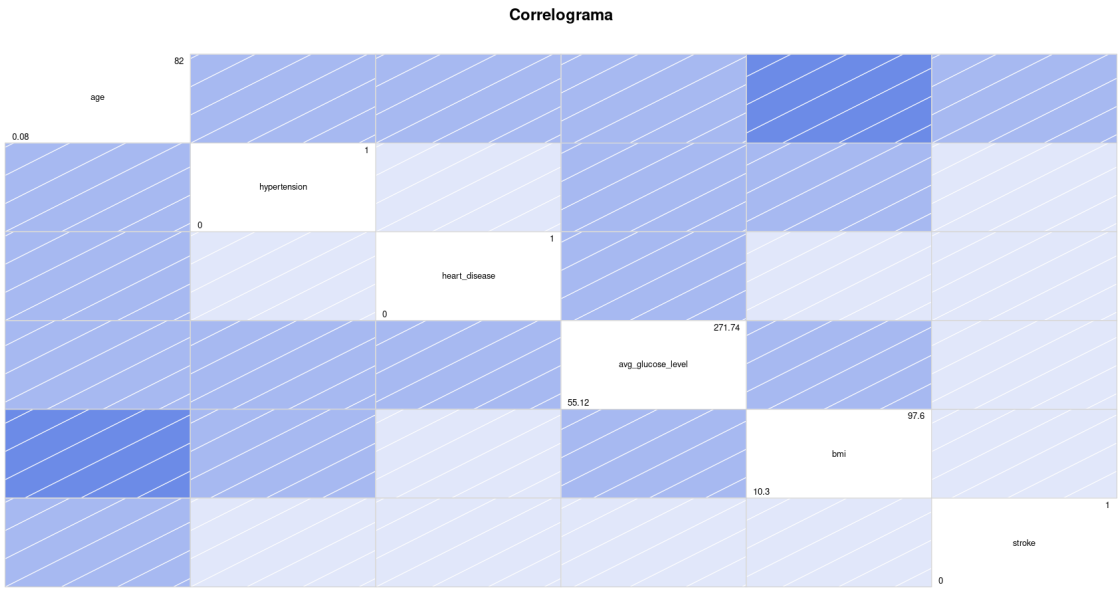
Parte 3.4 - Correlogramas

Correlograma das variáveis numéricas

- O gráfico abaixo, mostra a correlação entre todas as variáveis numéricas nos dados limpos(Parte 2).
- Os valores nas células da diagonal representam os valores de mínimo e máximo. Por exemplo, o menor IMC é 10,3, enquanto que o maior IMC é 97,6.
- A partir do correlograma e da tabela de correação, todas as variáveis numéricas estão positivamente correlacionadas com a variável preditora (AVC).
- A idade tem o maior índice de correlação com o AVC.

In [ ]:
```
library(corrgram)
```

In [ ]:
```
# Cria o correlograma para variáveis numéricas.
corrgram(data2, order=NULL, panel=panel.shade, text.panel=panel.txt,
         diag.panel=panel.minmax, main="Correlograma")
```



In [ ]:
```
# Valores de correlação das variáveis numéricas exibidos em 2 casas decimais
round(cor(subset(data2, select=c(age,hypertension, heart_disease,avg_glucose
```

A matrix: 6 × 6 of type dbl

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
age	1.00	0.28	0.26	0.24	0.33	0.25
hypertension	0.28	1.00	0.11	0.17	0.16	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	0.17	0.13
bmi	0.33	0.16	0.04	0.17	1.00	0.04
stroke	0.25	0.13	0.13	0.13	0.04	1.00

- A partir da matriz de correlação das variáveis numéricas, há 4 fatores principais para determinar se um paciente será vítima de AVC. São elas, idade, hipertensão, doença cardíaca, e nível médio de glicose no sangue.
- Como não há uma correlação forte entre as variáveis, podemos ignorar o risco de multicolinearidade

Correlograma das variáveis numéricas e categóricas.

In [ ]:
```
library(caret)
library(corrplot)
library(rpart)
library(rpart.plot)
```

In [ ]:
```
# Converte as variáveis categóricas para variáveis numéricas
# o novo dataset se chama data3

dmy <- dummyVars(" ~ .", data = data2)
data3 <- data.frame(predict(dmy, newdata = data2))

# Mostra os cabeçalhos do novo dataset
names(data3)
```

'genderFemale' · 'genderMale' · 'age' · 'hypertension' · 'heart_disease' · 'ever_marriedNo' · 'ever_marriedYes' · 'work_typechildren' · 'work_typeGovt_job' · 'work_typeNever_worked' · 'work_typePrivate' · 'work_typeSelf.employed' · 'Residence_typeRural' · 'Residence_typeUrban' · 'avg_glucose_level' · 'bmi' · 'stroke' · 'smoking.statusformerly.smoked' · 'smoking.statusnever.smoked' · 'smoking.statussmokes'

In [ ]:
```
# Tabela de correlação
cor_data3 <- correlate(data3)
```

```
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
```

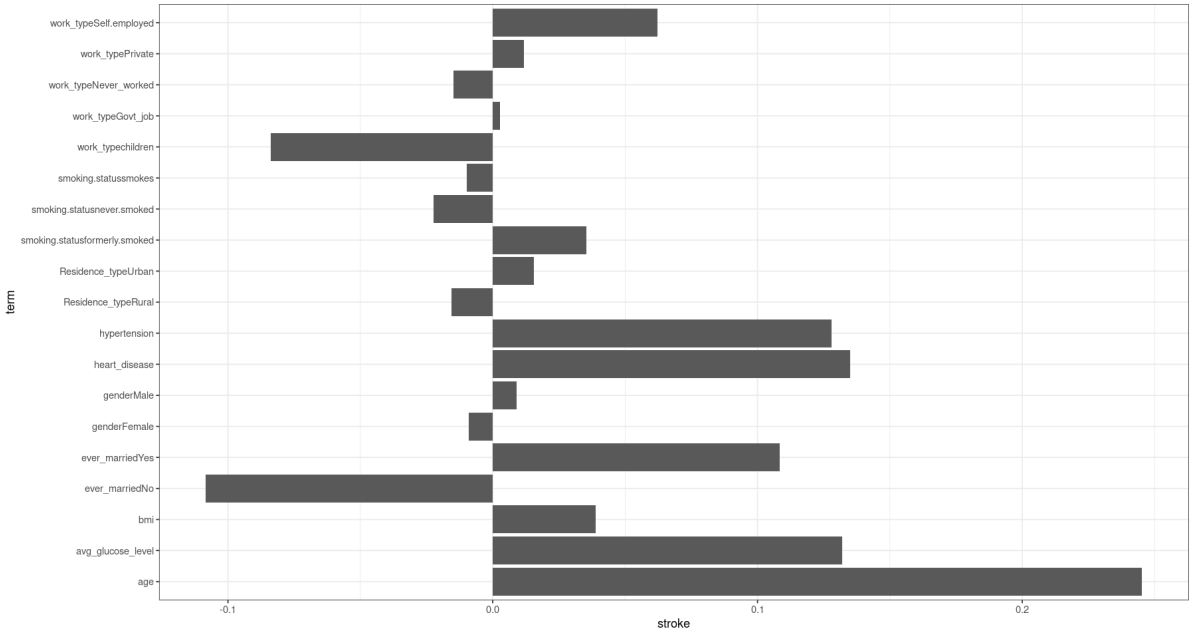
In [ ]:
```
# Extrai a correlação relacionada ao AVC
cor_data3%>% focus(stroke)
```

A tibble: 19 × 2

term	stroke
<chr>	<dbl>
genderFemale	-0.009117154
genderMale	0.009117154
age	0.245257346
hypertension	0.127903823
heart_disease	0.134913997
ever_marriedNo	-0.108339742
ever_marriedYes	0.108339742
work_typechildren	-0.083869266
work_typeGovt_job	0.002676705
work_typeNever_worked	-0.014882458
work_typePrivate	0.011888235
work_typeSelf.employed	0.062168257
Residence_typeRural	-0.015457965
Residence_typeUrban	0.015457965
avg_glucose_level	0.131945441
bmi	0.038946597
smoking.statusformerly.smoked	0.035236804
smoking.statusnever.smoked	-0.022362956
smoking.statussmokes	-0.009740997

- O gráfico abaixo inclui tanto variáveis númericas quanto categóricas. Os 4 fatores principais permanecem os mesmos.
- A variável que representa pacientes que já foram casados (Ever_married) tem a maior correlação com AVC dentre todas as variáveis categóricas.

In [ ]:
```
# Gráfico de correlação entre a variável AVC e todas as outras
cor_data3 %>%
  focus(stroke) %>%
  mutate(rowname = reorder(term, stroke)) %>%
  ggplot(aes(term, stroke)) +
    geom_col() + coord_flip() +
  theme_bw()
```



- Não há correlações fortes entre AVC e todas as outras variáveis
- A maior correlação com AVC pode ser observada na variável idade
- A correlação mais fraca entre AVC é com pacientes que nunca foram casados

Parte 3.5 - Gráfico de Barras agrupado por faixa etária

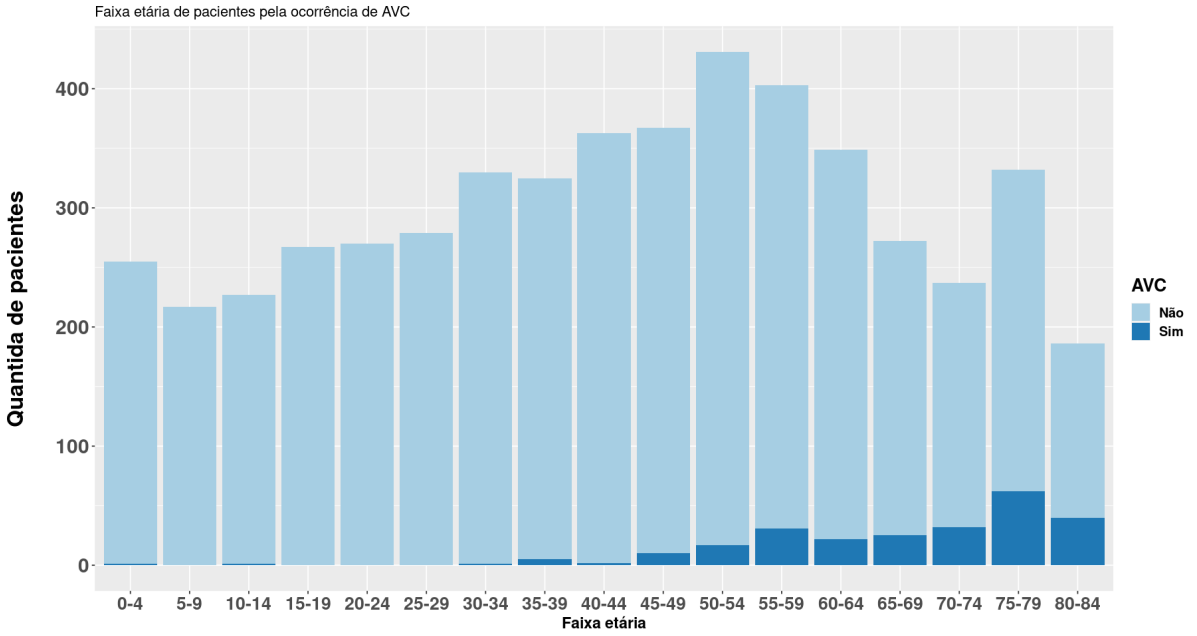
In [ ]:
```
# Divide a coluna idade em 17 faixas etárias
data3$age <- cut(data3$age,
                         breaks = c(-Inf
                                    ,5 ,10 ,15,20,25,30,35,40,45,50,55,60 ,6
                                    , Inf),

                         labels = c("0-4"
                                    ,"5-9","10-14","15-19","20-24"
                                    ,"25-29","30-34","35-39","40-44"
                                    ,"45-49","50-54","55-59","60-64"
                                    ,"65-69","70-74","75-79","80-84"
                                    ),
                         right = FALSE)
```

In [ ]:
```
stroke <- data2$stroke
agetablestroke <- as.data.frame(table(data3$age, stroke))
agetablestroke
```

A data.frame: 34 × 3

Var1	stroke	Freq
<fct>	<fct>	<int>
0-4	0	254
5-9	0	217
10-14	0	226
15-19	0	267
20-24	0	270
25-29	0	279
30-34	0	329
35-39	0	320
40-44	0	361
45-49	0	357
50-54	0	414
55-59	0	372
60-64	0	327
65-69	0	247
70-74	0	205
75-79	0	270
80-84	0	146
0-4	1	1
5-9	1	0
10-14	1	1
15-19	1	0
20-24	1	0
25-29	1	0
30-34	1	1
35-39	1	5
40-44	1	2
45-49	1	10
50-54	1	17
55-59	1	31
60-64	1	22
65-69	1	25
70-74	1	32
75-79	1	62
80-84	1	40

In [ ]:
```
# Gráfico de Barra mostrando as faixas etários de pacientes separados pelos
options(repr.plot.width=15, repr.plot.height=8)

ggplot(agetablestroke, aes(x=Var1, y=Freq, fill=stroke)) + geom_bar(stat="ic
  theme_gray() + scale_x_discrete(name = "Faixa etária") +
  ggtitle("Faixa etária de pacientes pela ocorrência de AVC") + ylab("Quanti
  scale_fill_brewer(palette="Paired", labels=c("Não","Sim"), name = "AVC") +
  theme(axis.title.x = element_text(face="bold", size=14, hjust = 0.5),
        axis.title.y = element_text(face="bold", size=20, hjust=0.5),
        axis.text.x = element_text(face="bold", size=16),
        axis.text.y = element_text(face="bold", size=18),
        legend.text = element_text(face = "bold", size=12),
        legend.title = element_text(face = "bold", size=16))
```

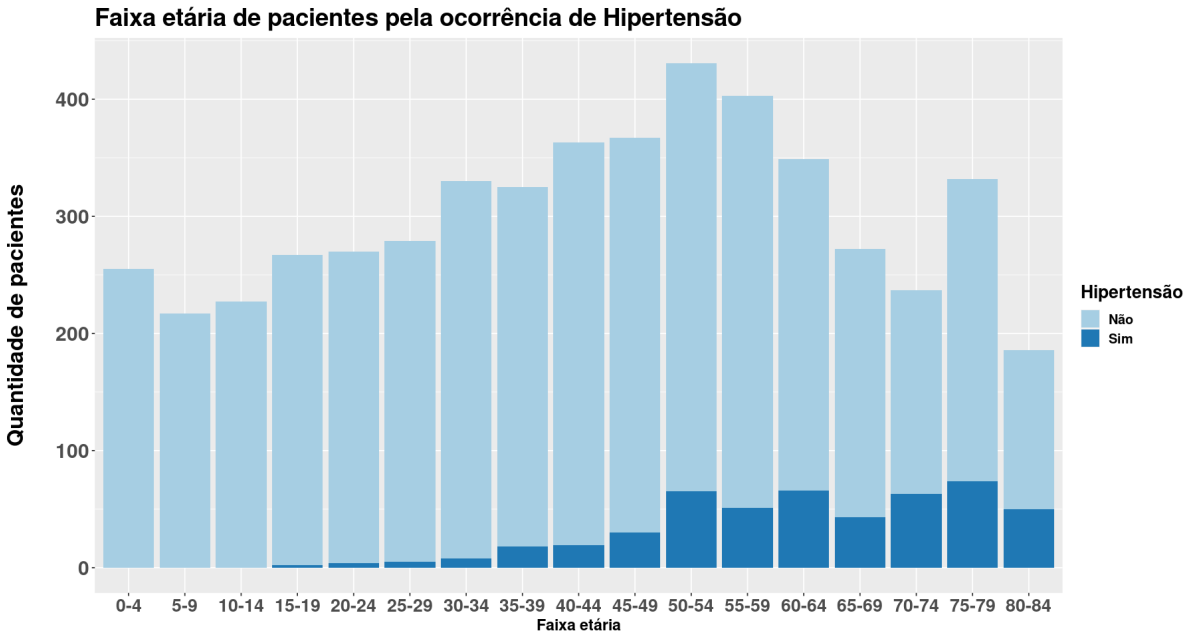


In [ ]:
```
hypertension <- data2$hypertension
agetablehypertension <- as.data.frame(table(data3$age, hypertension))
agetablehypertension
```

A data.frame: 34 × 3

Var1	hypertension	Freq
<fct>	<fct>	<int>
0-4	0	255
5-9	0	217
10-14	0	227
15-19	0	265
20-24	0	266
25-29	0	274
30-34	0	322
35-39	0	307
40-44	0	344
45-49	0	337
50-54	0	366
55-59	0	352
60-64	0	283
65-69	0	229
70-74	0	174
75-79	0	258
80-84	0	136
0-4	1	0
5-9	1	0
10-14	1	0
15-19	1	2
20-24	1	4
25-29	1	5
30-34	1	8
35-39	1	18
40-44	1	19
45-49	1	30
50-54	1	65
55-59	1	51
60-64	1	66
65-69	1	43
70-74	1	63
75-79	1	74
80-84	1	50

In [ ]:
```
# Gráfico de Barra mostrando as faixas etários de pacientes separados pelos
ggplot(agetablehypertension, aes(x=Var1, y=Freq, fill=hypertension)) + geom_
  theme_gray() + scale_x_discrete(name = "Faixa etária") +
  ggtitle("Faixa etária de pacientes pela ocorrência de Hipertensão") + ylab
  scale_fill_brewer(palette="Paired", labels=c("Não","Sim"), name = "Hiperte
  theme(axis.title.x = element_text(face="bold", size=14, hjust = 0.5),
        axis.title.y = element_text(face="bold", size=20, hjust=0.5),
        axis.text.x = element_text(face="bold", size=16),
        axis.text.y = element_text(face="bold", size=18),
        legend.text = element_text(face = "bold", size=12),
        legend.title = element_text(face = "bold", size=16),
        plot.title = element_text(face="bold", size = 22))
```



- Assim como o AVC, a hipertensão atinge com mais frequência pessoas de uma faixa etária mais elevada.

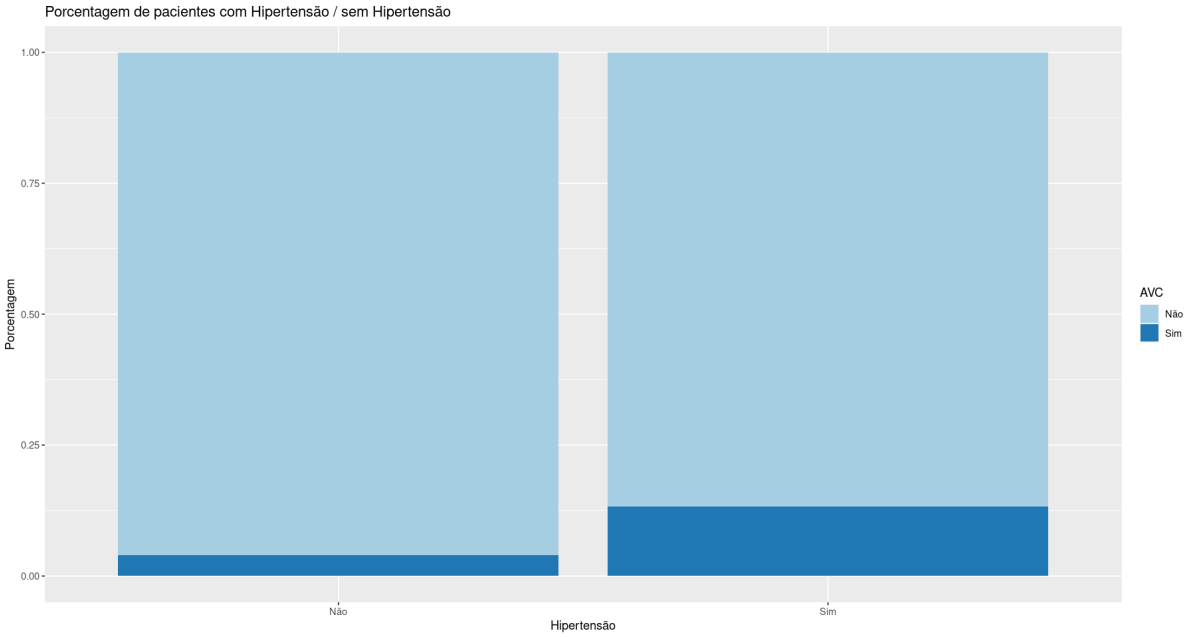
Parte 4 - Como a hipertensão influencia a probabilidade de sofrer um AVC?

In [ ]:
```
grouped_data <- mutated_data %>%
  group_by(hypertension, stroke) %>%
  count(stroke)
grouped_data
```

A grouped_df: 4 × 3

hypertension	stroke	n
<chr>	<chr>	<int>
Não	Não	4429
Não	Sim	183
Sim	Não	432
Sim	Sim	66

In [ ]:
```
grouped_data %>%
ggplot(mapping = aes(x = hypertension, y = n, fill = stroke) ) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Porcentagem de pacientes com Hipertensão / sem Hipertensão",
       x = "Hipertensão",
       y = "Porcentagem") +
  scale_fill_brewer(palette = "Paired", name = "AVC")
```



A quantidade de pacientes sem hipertensão é muito maior do que a de pacientes com hipertensão neste conjunto de dados. A percentagem de indivíduos que sofreram um AVC com hipertensão é maior do que a percentagem de indivíduos que tiveram um AVC sem hipertensão. Portanto, este conjunto de dados provou o fato de que a hipertensão é um fator de risco significativo para acidente vascular cerebral. Na verdade, é o fator de risco mais importante para o AVC. Hipertensão tende a ocorrer na família. Se um membro da família tem hipertensão, a pessoa tem um risco maior de tê-la. No entanto, a hipertensão pode ser prevenida ou controlada por medicamentos e um estilo de vida saudável.