Trabalho Caso 12 - AED

Catarina Castanheira (92478), João Martins (93259), Joel Paula (93392)

Contents

Introdução

No seguimento da análise exploratória dos dados do caso prático 12, é interessante também aproveitar as funcionalidades que o R nos pode dar neste âmbito. Sendo assim, utilizando já o ficheiro xlsx respeitante ao *dataset* em questão e que tinha sido previamente tratado nas tarefas anteriores, nesta parte teremos como objetivo: atribuir nomes a variáveis que estavam com códigos, identificar valores omissos, imputar novos valores a estes omissos, e analisar descritivamente alguns dados.

Começa-se então pela importação do ficheiro Excel e dos *packages* que vão ser utilizados (tarefa II - 1).

```
library(openxlsx)
bd <- read.xlsx("CP12_Ecologia_final.xlsx", sheet = "Dados_Eco1")


library(flextable)
library(ggplot2)

print(c("Número de registos da amostra:", nrow(bd))) ### MODIFICAR ###
```

```
## [1] "Número de registos da amostra:" "354"
```

Depois de importado o ficheiro de *Excel* podemos atribuir novos nomes às variáveis que estavam identificadas com código (exemplo p8.1 - Incêndios florestais), da seguinte forma:

```
colnames(bd)[2] <- c("incendios_florestais")
colnames(bd)[3] <- c("Acidentes_trabalho")
colnames(bd)[4] <- c("Poluicao_rios")
colnames(bd)[5] <- c("Poluicao_praias")
colnames(bd)[6] <- c("Ma_qualidade_ar")
colnames(bd)[7] <- c("Utilizacao_pesticidas")
colnames(bd)[8] <- c("Efeito_estufa")
colnames(bd)[9] <- c("Reciclagem_lixos")
colnames(bd)[10] <- c("Destruicao_florestas")
```

Valores omissos

Depois disto procedemos à substituição de valores omissos (identificados por 99), para *NA* para facilitar a análise posterior.

```
bd$incendios_florestais[which(bd$incendios_florestais == 99)] <- c(NA)
```

Procurámos os valores omissos da idade e verificámos que existiam 3 valores omissos na idade usando a seguinte instrução (Tarefa II - 3):

```
agedf <- as.data.frame(bd$ID[which(is.na(bd$idade))])
ageft <- flextable(agedf)
ageft <- set_header_labels(ageft,`bd$ID[which(is.na(bd$idade))]`="Id")
ageft
```

Verificámos que existiam 3 valores omissos e substituímos pela mediana das idades (Tarefa II - 4):

```
bd$idade[which(is.na(bd$idade))] <- c(median(bd$idade, na.rm=TRUE))
```

Sendo a mediana das idades:

```
median(bd$idade, na.rm=TRUE)
```

```
## [1] 34
```

Da mesma forma, procurámos os valores omissos para a variável "Preocupação ambiental" (Tarefa II - 3).

```
agedf <- as.data.frame(bd$ID[which(is.na(bd$preoc_ambiente))])
ageft <- flextable(agedf)
ageft <- set_header_labels(ageft,`bd$ID[which(is.na(bd$preoc_ambiente))]`="Id")
ageft
```

Neste caso encontrámos 2 valores omissos aos quais imputámos também o valor da mediana da variável (Tarefa II - 4):

```
bd$preoc_ambiente[which(is.na(bd$preoc_ambiente))] <- c(median(bd$preoc_ambiente, na.rm=TRUE))
```

O valor da mediana desta variável é dado por:

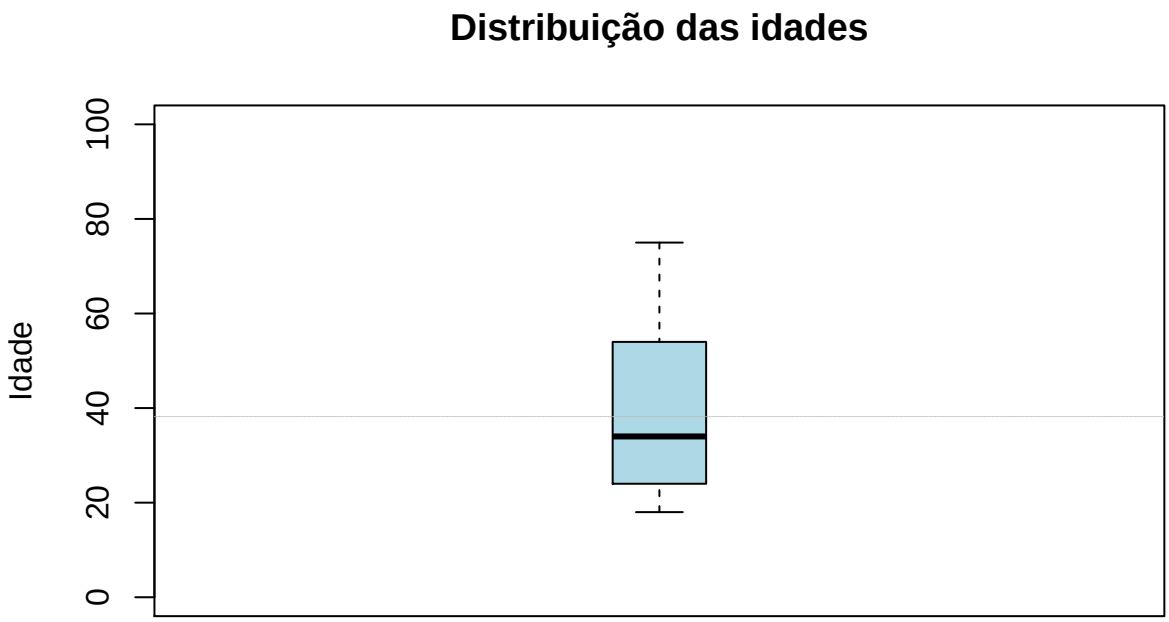
```
median(bd$preoc_ambiente, na.rm=TRUE)
```

```
## [1] 80
```

Caracterização da População

Depois de limpos os dados, podemos começar a detalhar alguma informação relativamente às características amostra e a concretizar uma breve análise das respostas ao questionário (Tarefa II - 5). Para a caracterização da amostra da população é necessário analisar as variáveis: idade, sexo, tem/não tem filhos, rendimento e tem/não tem automóvel. Socorrendo-nos do *boxplot* obtemos o seguinte, em relação à variável idade:

```
boxplot(bd$idade, ylab = "Idade", main = "Distribuição das idades", col="light blue", ylim = c(0,100), l
abline(h=mean(bd$idade, na.rm = TRUE), lty="dashed", lwd="0.1", col="grey")
```




```
summary(bd$idade)
```

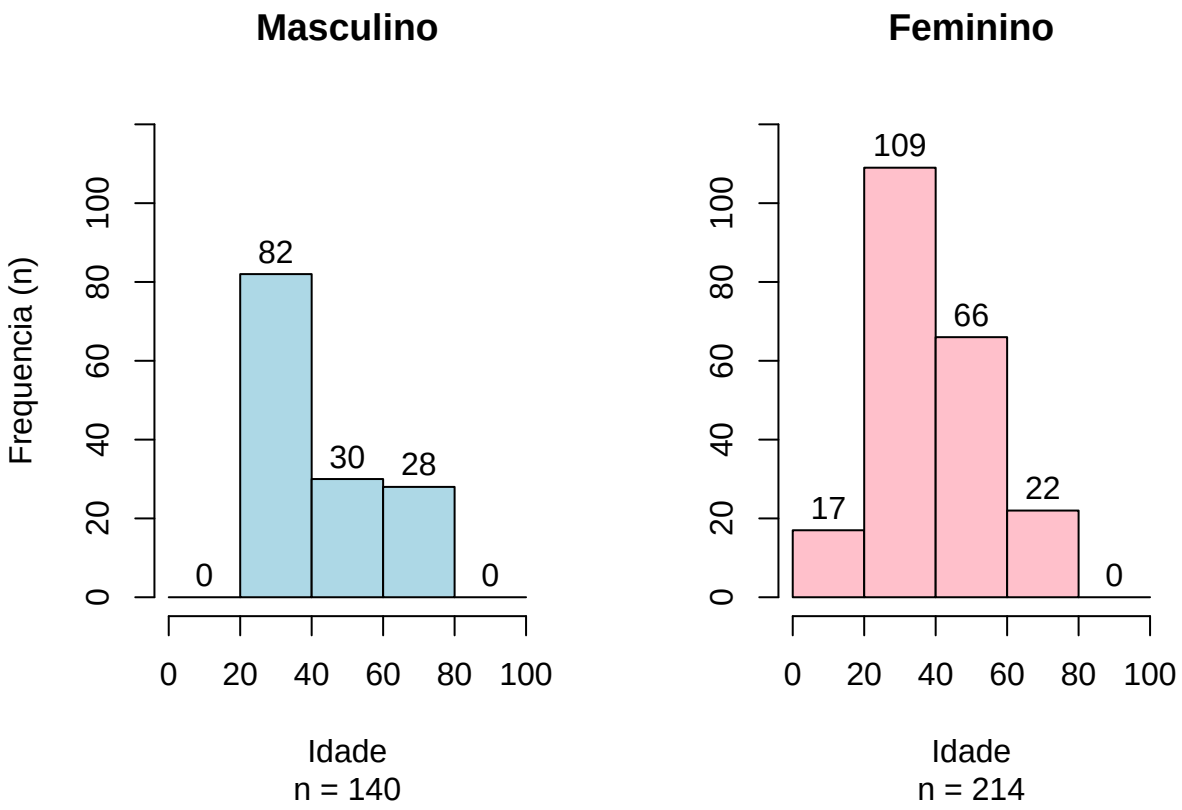
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0    24.0    34.0    38.2    54.0    75.0
```

Por observação do *boxplot* e da tabela *summary*, observámos que o mínimo de idades situa-se nos 18 anos e o máximo situa-se nos 75 anos. Observamos também que 25% da população tem idades até 24 anos (1º Quartil), 50% tem idades até 34 anos (Mediana) e 75% até 74 anos. A média de idades situa-se nos 38,2 anos.

Ainda relativamente à idade conseguimos fazer uma análise pormenorizada considerando a distribuição respetiva pelos sexos. Desta vez torna-se mais interessante o recurso a histogramas. Obtemos então a seguinte distribuição das idades pelo sexo:

```
par(mfrow=c(1,2))
hist(bd$idade[which(bd$sexotexto=="Masculino")], ylim = c(0, 120), xlim = c(0,100),breaks = c(0,20,40,6

hist(bd$idade[which(bd$sexotexto=="Feminino")], ylim = c(0, 120), xlim = c(0,100),breaks = c(0,20,40,60
```




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

De acordo com os gráficos facilmente se percebe que temos uma amostra do sexo masculino composta por 140 indivíduos e de 214 indivíduos do sexo feminino. Em ambos os casos a classe de idades com maior registo de observações é a que vai dos 20 aos 40 anos, com 82 casos na amostra masculina e 109 na amostra feminina. Observámos também que existem 17 indivíduos do sexo feminino com menos de 20 anos. Já no caso masculino não observamos qualquer registo nesta classe de idades.

Estatística descritiva relativa às idades, considerando a amostra masculina:

```
summary(bd$idade[which(bd$sexotexto=="Masculino")])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.00   24.00   28.50   38.61   60.00   72.00
```

Estatística descritiva relativa às idades, considerando a amostra feminina:

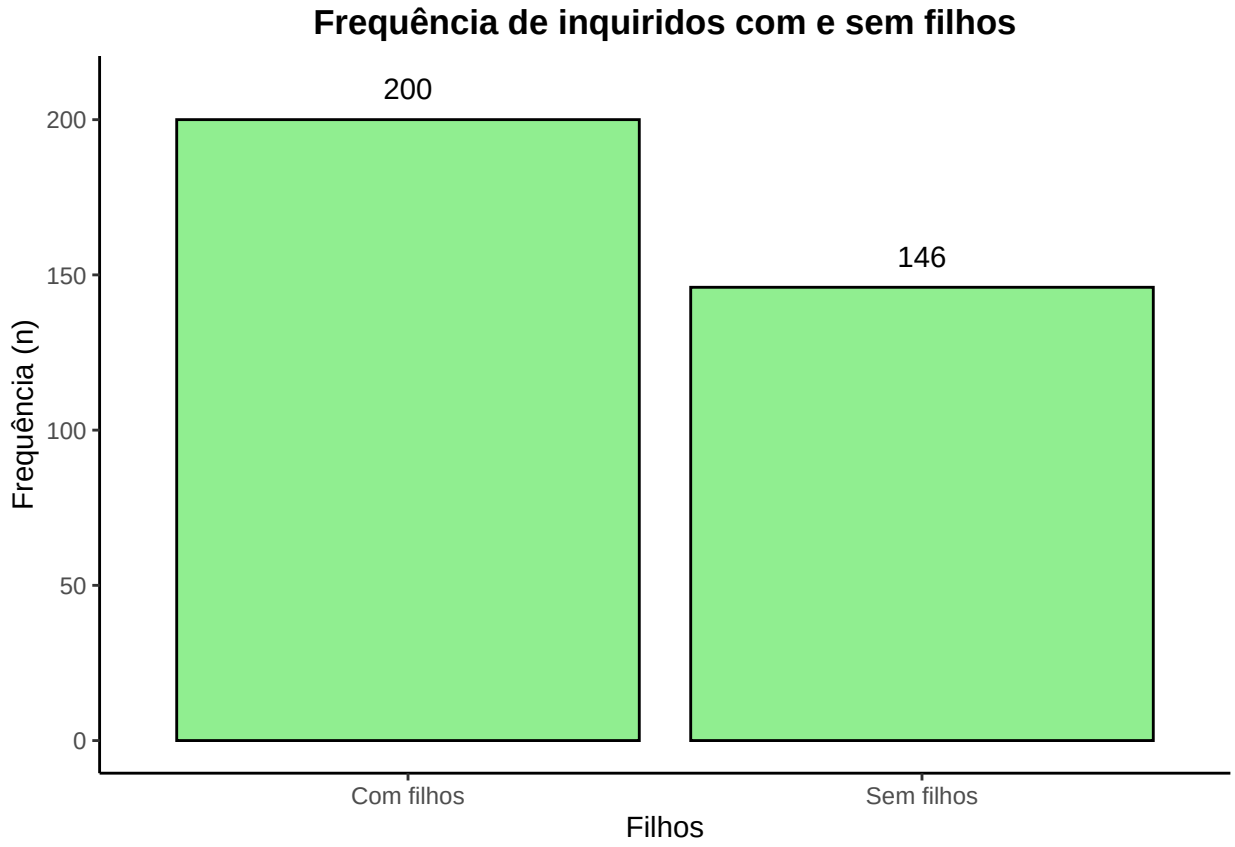
```
summary(bd$idade[which(bd$sexotexto=="Feminino")])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   24.00   36.00   37.93   50.00   75.00
```

Complementando os gráficos acima com as estatísticas descritivas dadas pelas tabelas para cada um dos casos, percebemos que o mínimo de idade é de 21 anos no caso masculino e 18 anos no caso feminino e que o máximo de idades é 72 e 75 anos, respetivamente. A média de idades é 38,61 no sexo masculino e 37,93 no sexo feminino.

Caracterização da amostra em relação à variável filhos:

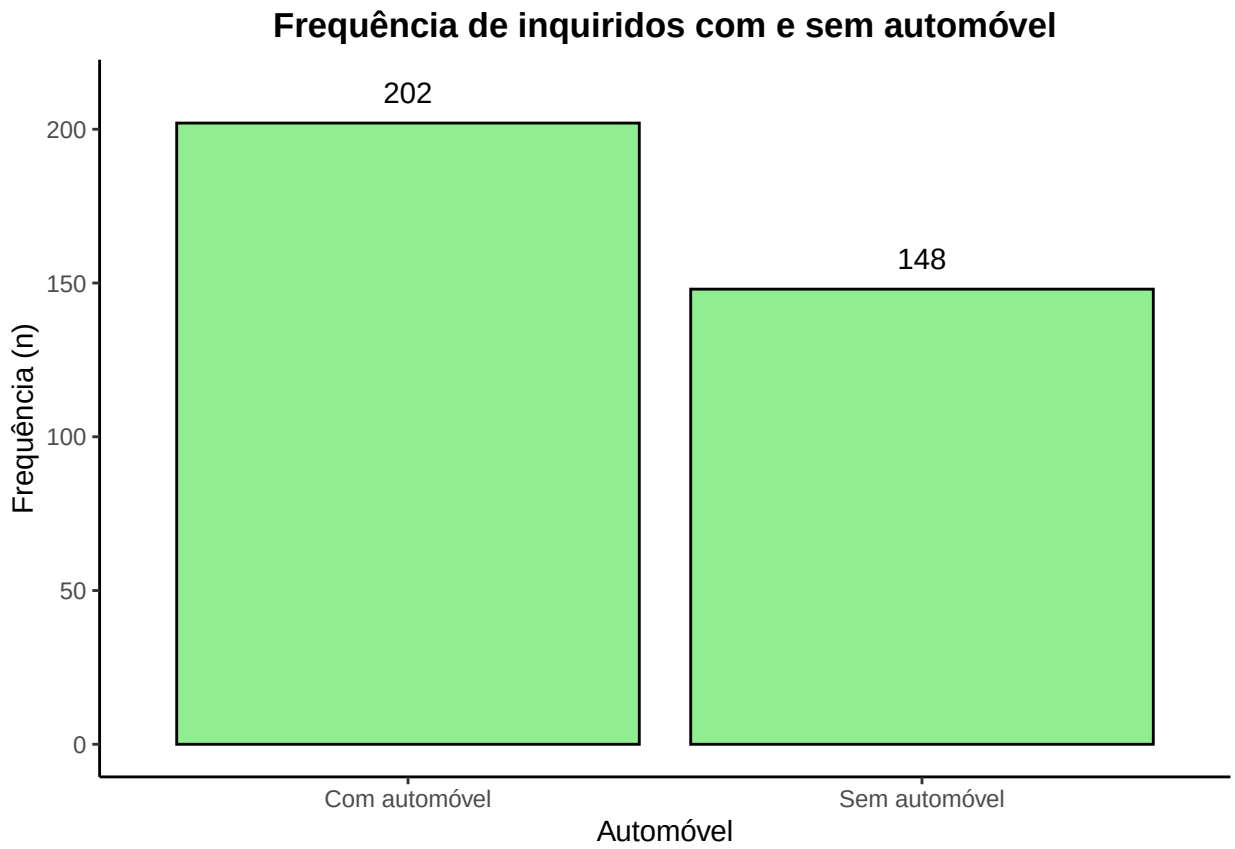
```
tab.ft <- table(bd$filhostexto)
bp <- ggplot(bd[which(is.na(bd$filhostexto)==0),], aes(filhostexto, na.rm = TRUE)) + geom_histogram(sta
bp
```



Por observação do gráfico percebemos que 200 dos inquiridos têm filhos e 146 não têm.

Caracterização da amostra em relação à variável automóvel:

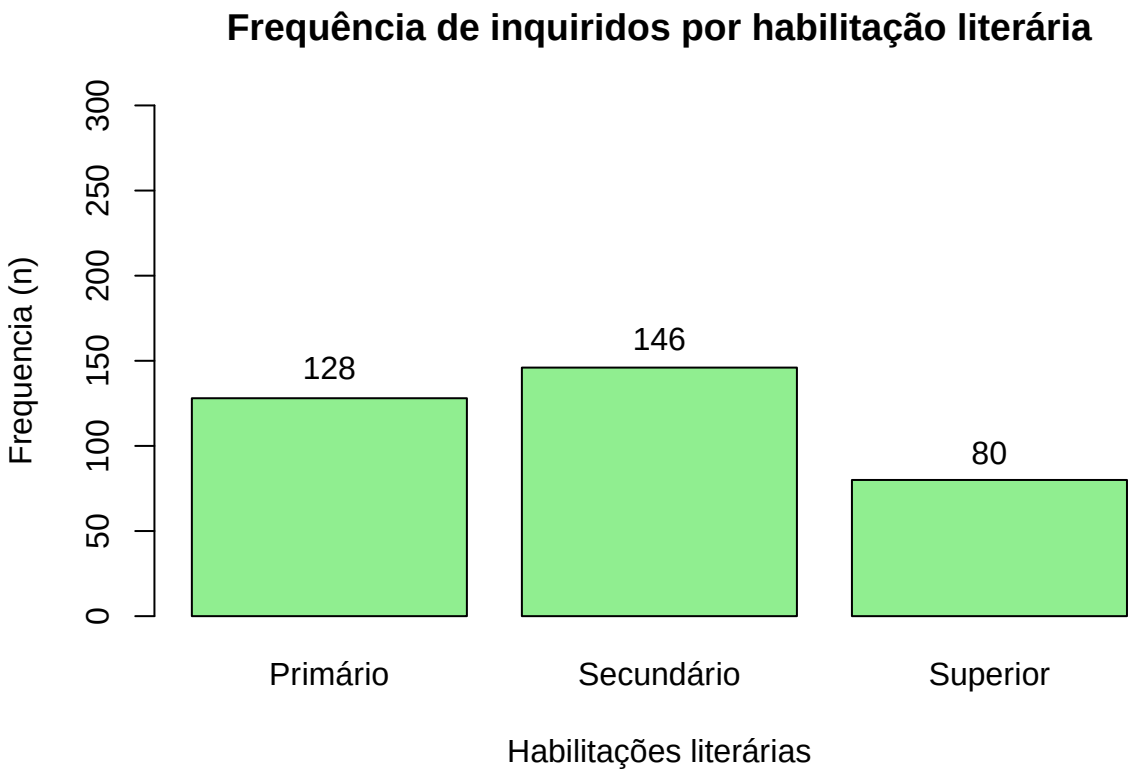
```
tab.ft <- table(bd$automovel)
bp <- ggplot(bd[which(is.na(bd$automovel)==0),], aes(automovel)) + geom_histogram(stat="count", fill="#
bp
```



Neste caso, observamos 202 inquiridos com automóvel e 148 sem automóvel.

Caracterização da amostra em relação à variável habilitações literárias:

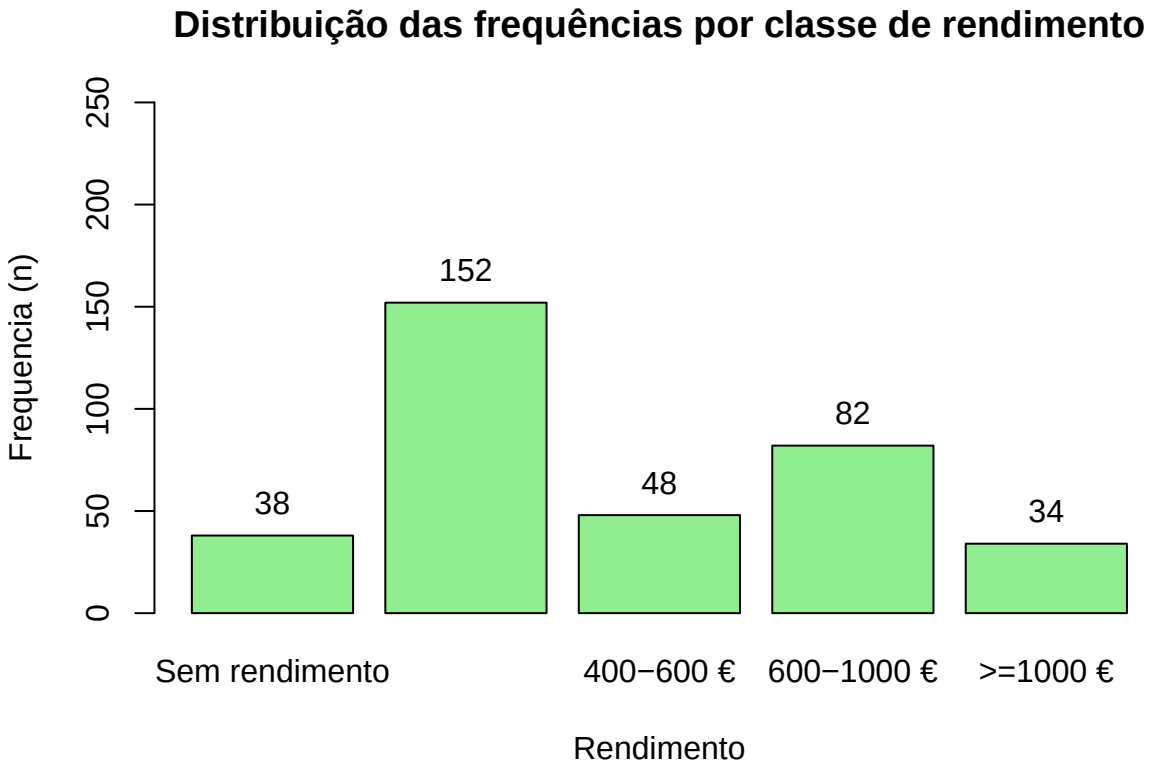
```
bp2 <- barplot(table(bd$habilitacoestexto),ylim = c(0, 300), col="lightgreen", xlab = "Habilitações lit
text(bp2, c(128, 145, 78), table(bd$habilitacoestexto), cex=1, pos = 3)
```



Quanto a este gráfico, percebemos que 128 inquiridos têm habilitações literárias até ao ensino primário, 146 até ao ensino secundário e outros 80 até ao ensino superior.

Caracterização da amostra em relação à variável rendimento:

```
bp3 <- barplot(table(bd$rendimento)[c(5,1,3,4,2)], ylim = c(0, 250), col="lightgreen", xlab = "Rendimen
text(bp3, table(bd$rendimento)[c(5,1,3,4,2)] + 1, table(bd$rendimento)[c(5,1,3,4,2)], cex=1, pos = 3)
```



Neste caso, o maior número de registos é observado em rendimentos superiores a 0€ e inferiores a 400€, com 152 casos. A classe de rendimentos com menor número de observações é aquela com valor igual ou superior a 1000€.

Em todos estes gráficos, a diferença entre o total de observações e a frequência de cada tipo de resposta dá-nos os casos de não resposta.

Desta forma conclui-se a análise exploratória de dados com recurso ao R e procedemos para a análise das variáveis Preocupação ambiental, Efeito de estufa e Incêndios florestais com recurso ao *Jamovi*.