Maria Eduarda Rodrigues Garcia - 11796621
Mirela Mei - 11208392

Roteiro - Novo dataset

1. **Identificação do dataset escolhido: Efeitos do álcool no estudo (Alcohol Effects On Study)**
    - Fonte: https://www.kaggle.com/datasets/whenamancodes/alcohol-effects-on-study?resource=download

- Quantas linhas e quantas colunas tem o dataset?
    - 396 linhas e 33 colunas

2. **Procedimentos de amostragem**

   Estes dados abordam o desempenho dos alunos no ensino secundário de duas escolas portuguesas. Os atributos de dados incluem notas dos alunos, características demográficas, sociais e relacionadas à escola e foram coletados por meio de relatórios e questionários escolares.

3. **Variáveis**
    - school - qualitativa nominal
    - sex - qualitativa nominal
    - **age** - quantitativa discreta
    - address - qualitativa nominal
    - famsize - qualitativa nominal
    - Pstatus - qualitativa nominal
    - Medu - qualitativa ordinal
    - Fedu - qualitativa ordinal
    - Mjob - qualitativa nominal
    - Fjob - qualitativa nominal
    - reason - qualitativa nominal
    - guardian - qualitativa nominal
    - traveltime - qualitativa ordinal
    - studytime - qualitativa ordinal
    - failures - qualitativa ordinal
    - schoolsup - qualitativa nominal
    - famsup - qualitativa nominal
    - paid - qualitativa nominal
    - activities - qualitativa nominal
    - nursery - qualitativa nominal
    - higher - qualitativa nominal

    - internet - qualitativa nominal
    - romantic - qualitativa nominal
    - famrel - qualitativa ordinal
    - freetime - qualitativa ordinal
    - goout - qualitativa ordinal
    - Dalc - qualitativa ordinal
    - Walc - qualitativa ordinal
    - health - qualitativa ordinal
    - **absences** - quantitativa discreta
    - **G1** - quantitativa discreta
    - **G2** - quantitativa discreta
    - **G3** - quantitativa discreta

- Há missing data? Em quais variáveis? Indique como vocês trataram os missing data.
    - Não há missing data

4. **Observações, casos ou instâncias**
    - População observada em Portugal

5. **Estatística descritiva**

Medidas de tendência central e dispersão:

- Summary: valor mínimo, first quarter, mediana, third quarter, valor máximo
- Median: mediana
- Mode: moda
- Describe: número de registros, média, desvio padrão, mediana, valores mínimo e máximo, skew, kurtosis, amplitude, erro padrão

```
> summary(dados$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   15.0    16.0    17.0    16.7    18.0    22.0
> summary(dados$absences)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   4.000   5.709   8.000  75.000
> summary(dados$G1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00    8.00   11.00   10.91   13.00   19.00
> summary(dados$G2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.00   11.00   10.71   13.00   19.00
> summary(dados$G3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   11.00   10.42   14.00   20.00
```

```
> medianAge <- median(dados$age)
> medianAbsences <- median(dados$absences)
> medianG1 <- median(dados$G1)
> medianG2 <- median(dados$G2)
> medianG3 <- median(dados$G3)
>
> print(medianAge)
[1] 17
> print(medianAbsences)
[1] 4
> print(medianG1)
[1] 11
> print(medianG2)
[1] 11
> print(medianG3)
[1] 11
```

```
> getmode <- function(dados) {
+   uniqv <- unique(dados)
+   uniqv[which.max(tabulate(match(dados, uniqv)))]
+ }
> resultAge <- getmode(dados$age)
> resultAbsences <- getmode(dados$absences)
> resultG1 <- getmode(dados$G1)
> resultG2 <- getmode(dados$G2)
> resultG3 <- getmode(dados$G3)
>
> print(resultAge)
[1] 16
> print(resultAbsences)
[1] 0
> print(resultG1)
[1] 10
> print(resultG2)
[1] 9
> print(resultG3)
[1] 10
```

```
> describe(dados$age)
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 395 16.7 1.28     17   16.63 1.48  15  22     7 0.46    -0.03 0.06
> describe(dados$absences)
   vars   n mean sd median trimmed  mad min max range skew kurtosis  se
X1    1 395 5.71  8      4    4.24 5.93   0  75    75 3.64    21.31 0.4
> describe(dados$G1)
   vars   n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 395 10.91 3.32     11    10.8 4.45   3  19    16 0.24    -0.71 0.17
> describe(dados$G2)
   vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 395 10.71 3.76     11   10.84 2.97   0  19    19 -0.43     0.59 0.19
> describe(dados$G3)
   vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 395 10.42 4.58     11   10.84 4.45   0  20    20 -0.73     0.37 0.23
```

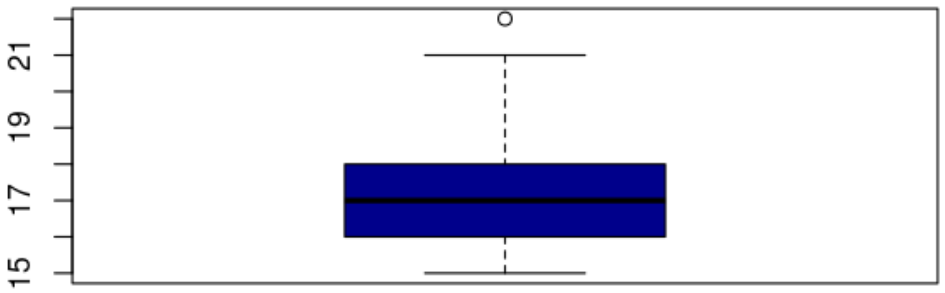
```
> # Erro
> std_mean <- function(x) sd(x)/sqrt(length(x))
> std_mean(dados$age)
[1] 0.06420468
> std_mean(dados$absences)
[1] 0.4026794
> std_mean(dados$G1)
[1] 0.1670068
> std_mean(dados$G2)
[1] 0.1892618
> std_mean(dados$G3)
[1] 0.2305174
```

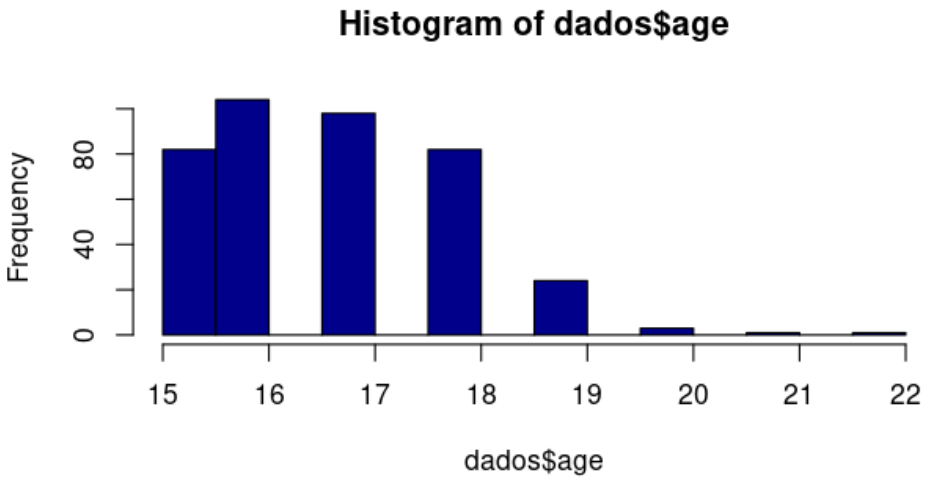
```
> # Variância
> var(dados$age)
[1] 1.628285
> var(dados$absences)
[1] 64.04954
> var(dados$G1)
[1] 11.01705
> var(dados$G2)
[1] 14.14892
> var(dados$G3)
[1] 20.98962
```

```
> # Amplitudes
> range(dados$age)
[1] 15 22
> range(dados$absences)
[1]  0 75
> range(dados$G1)
[1]  3 19
> range(dados$G2)
[1]  0 19
> range(dados$G3)
[1]  0 20
```

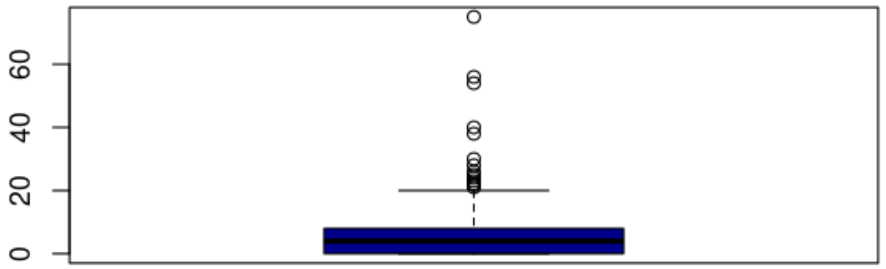
Gráficos

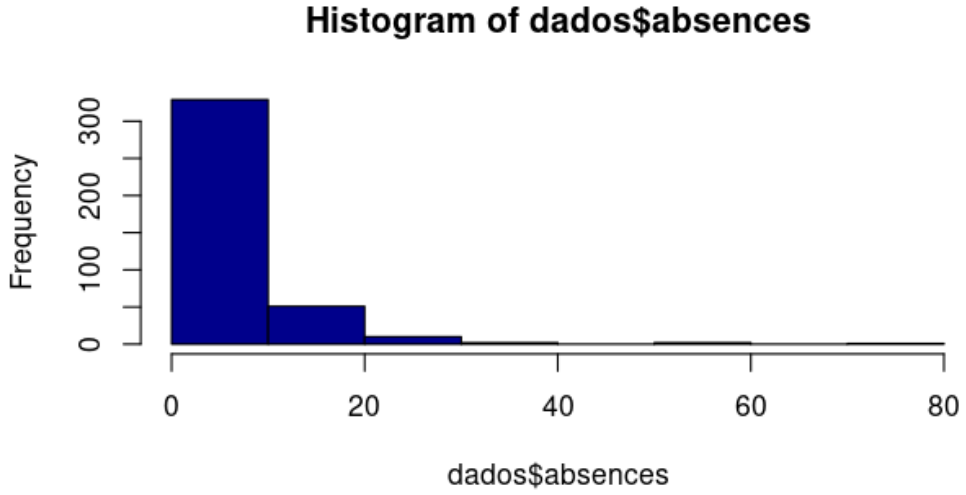
- Idade



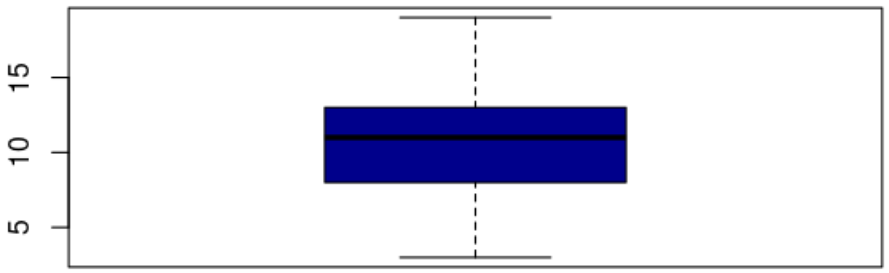


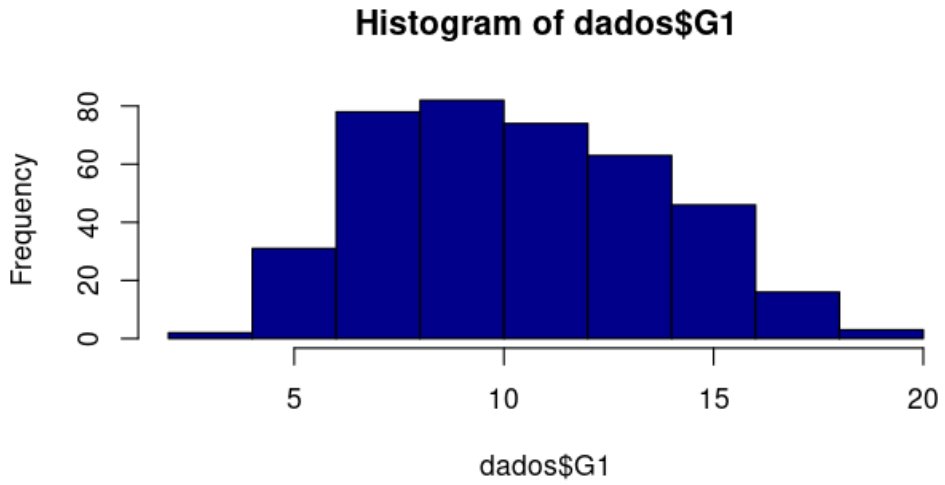
- Faltas



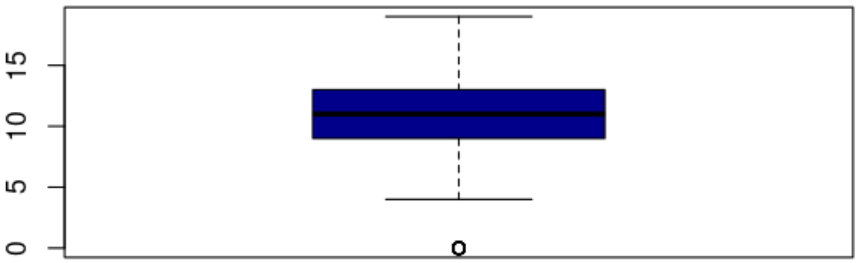


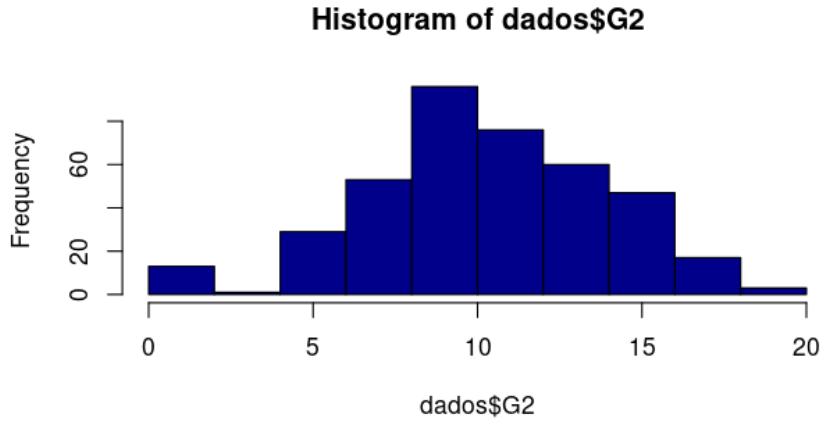
- Nota do primeiro período



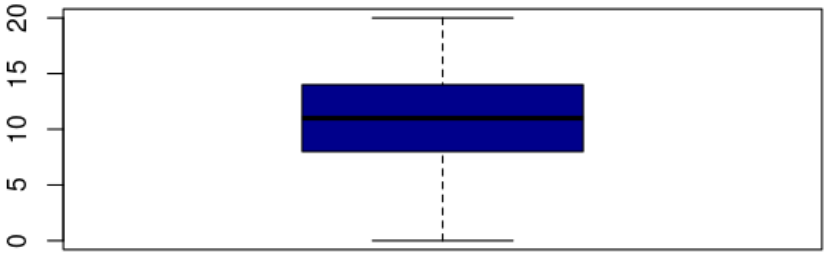


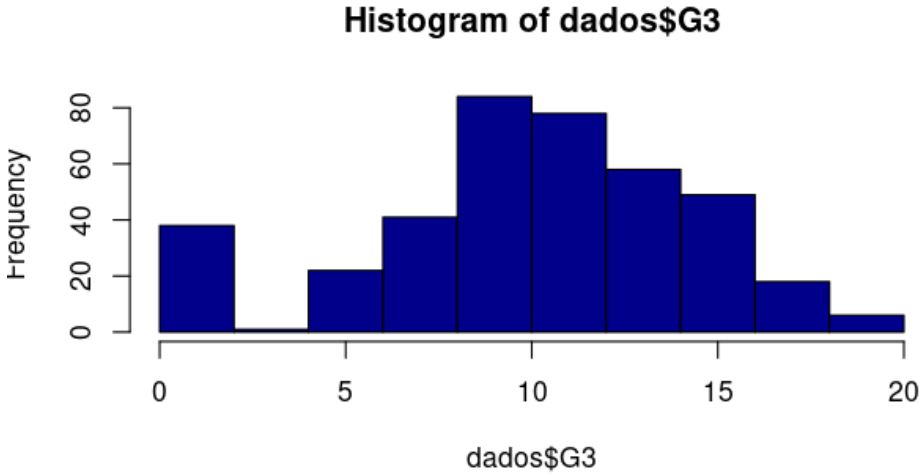
- Notas do segundo período





- Notas do terceiro período





6. **Que tipo de pesquisa/pergunta você pretende fazer com este dataset?**
    - Como o consumo de álcool interfere nas notas dos alunos?
    - A idade faz diferença em relação ao consumo de álcool?
    - A idade interfere nas notas?
    - O consumo de álcool interfere na quantidade de faltas?

- Qual seria um título adequado para o seu trabalho?
    - Consumo de álcool na adolescência: quais as consequências?