Relatório Atividade A2

Liz Alexandrita de Souza Barreto

21/10/2021

Universidade de Franca
RGM: 21125066

Preparação do Ambiente

Este relatório foi feito em Rmarkdown!

```
library('readtext')
library('tidyverse')
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Leitura do Arquivo e entendimento dos campos

```
idades_text <- readtext(file = 'idades.docx')
glimpse(idades_text)
```

```
## Rows: 1
## Columns: 2
## $ doc_id <chr> "idades.docx"
## $ text   <chr> "7, 7, 7, 89, 89, 47, 47, 47, 4, 4, 4, 39, 73, 73, 73, 73, 70, ~
```

Preparação do texto para dataframe numérico para análise

```
tmp <- strsplit(x=idades_text$text, split=", ")
glimpse(tmp)
```

```
## List of 1
##  $ : chr [1:733] "7" "7" "7" "89" ...
```

```
tmp <- as.data.frame(tmp)
glimpse(tmp)
```

```
## Rows: 733
```

```
## Columns: 1
## $ c..7....7....7....89....89....47....47....47....4....4....4... <chr> "7", "7~
```

```
names(tmp) <- 'idades'
tmp$idades <- as.numeric(tmp$idades)
glimpse(tmp)
```

```
## Rows: 733
## Columns: 1
## $ idades <dbl> 7, 7, 7, 89, 89, 47, 47, 47, 4, 4, 4, 39, 73, 73, 73, 73, 70, 3~
```

Análise - Estatística Descritiva

Entendimento do Boxplot

```
summary(tmp)
```

```
##      idades
##  Min.   : 1.00
##  1st Qu.:38.00
##  Median :55.00
##  Mean   :53.55
##  3rd Qu.:71.00
##  Max.   :96.00
```

Entendimento da distribuição

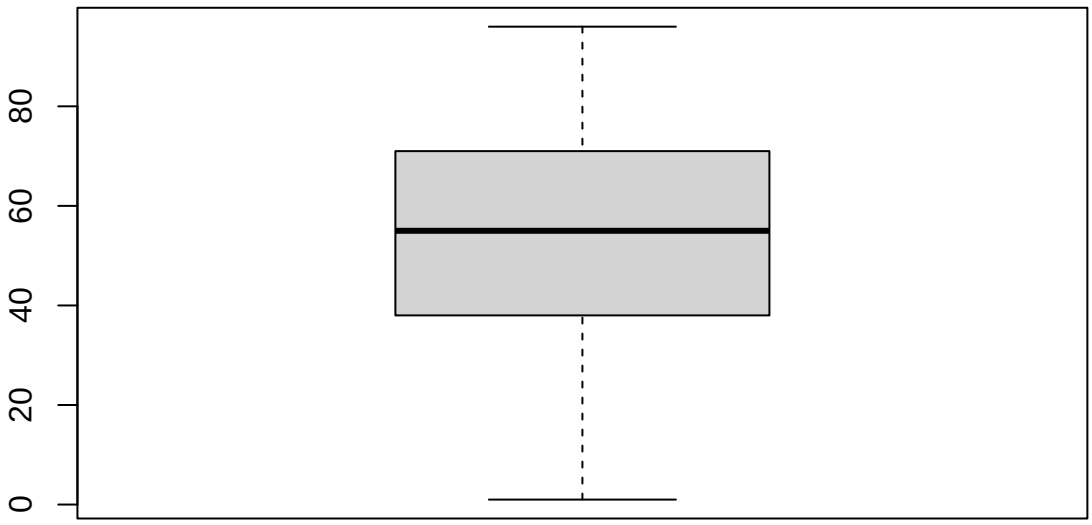
```
tmp %>% summarize(
  min = min(idades),
  max = max(idades),
  mean = mean(idades),
  median = median(idades),
  var = var(idades),
  sd = var(idades) ^ (1/2))
```

```
##   min max     mean median      var       sd
## 1   1  96 53.54707     55 476.0159 21.81779
```

Análise - Gráfica

Boxplot pedido

```
tmp %>% boxplot()
```

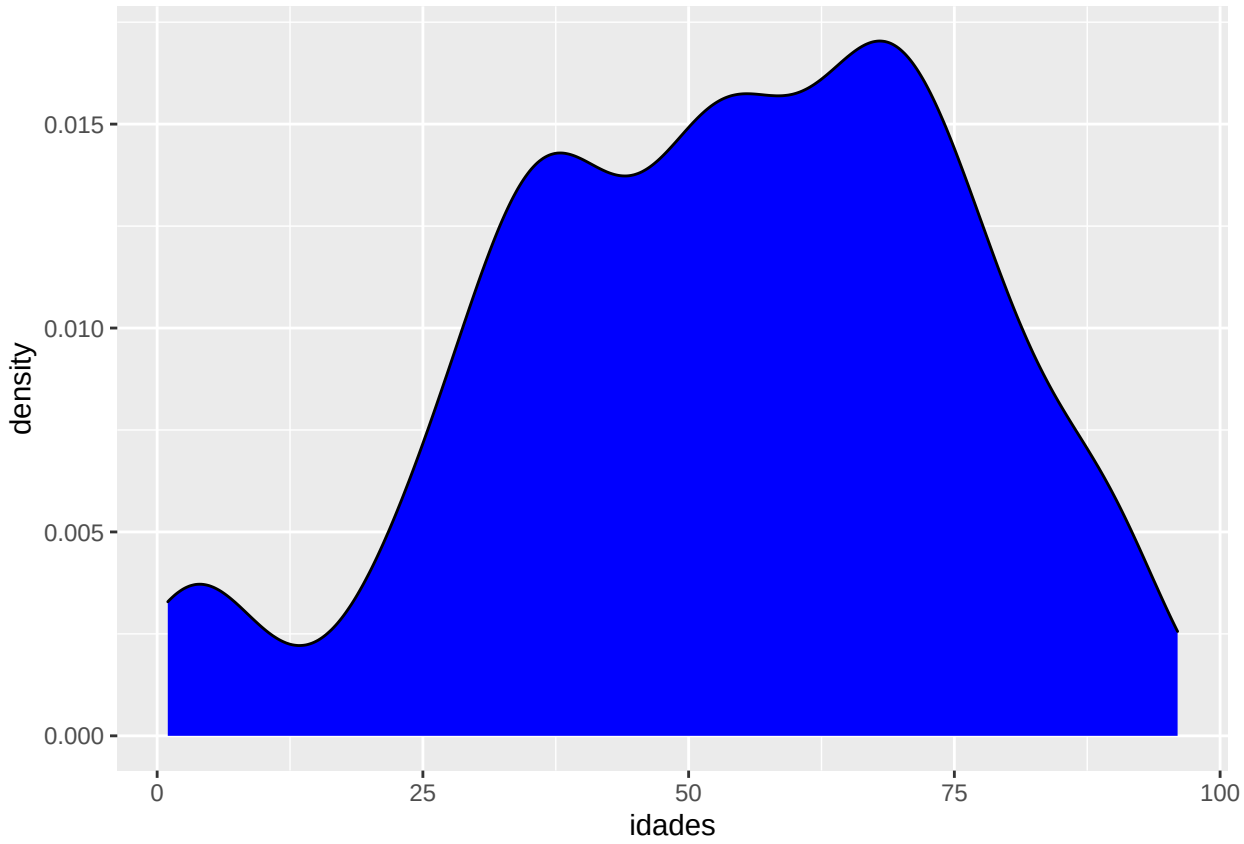


Aprofundamento da compreensão dos dados

Para esse aprofundamento eu preferi olhar tanto para o histograma quanto para a densidade para compreender a distribuição dos dados visualmente. Consigo assim observar se a disposição dos dados está próxima de numa normal ou não, quais as maiores frequências de dados de idade e etc que não calculei no *summary*. Diga-se de passagem não existe um cálculo direto de moda em R.

Densidade

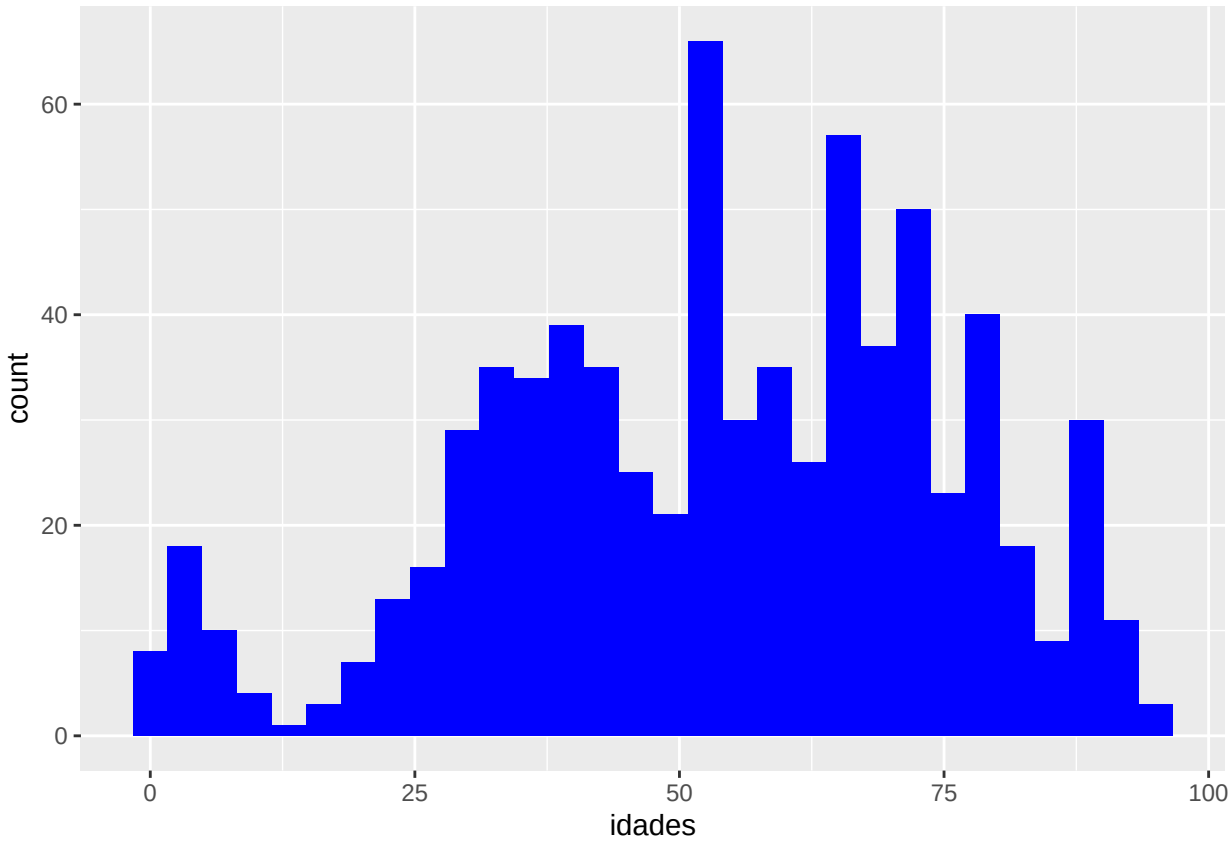
```
tmp %>% ggplot(aes(idades)) +
  geom_density(fill = "blue")
```



Histograma

```
tmp %>% ggplot(aes(idades)) +
  geom_histogram(fill = "blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Conclusões

O boxplot pode ser interpretado da seguinte forma: A linha em negrito no meio se refere à posição da mediana (55 anos) e a dispersão dos dados é o tamanho geral da caixinha que no caso é o intervalo interquartílico 3º-1º que seria 71-38 = 33 anos (eu também sumarizei calculando variância e desvio padrão e vi que a maior parte dos dados fica entre 31 e 75 anos, 53 anos da média +- 22 anos de desvio padrão), os valores máximo e mínimo são as retas mais alta e mais baixa respectivamente e a simetria, ou *skewness* (que eu chamei na primeira atividade) está próxima à simétrica, ligeiramente negativa, porém só dá pra perceber isso olhando nos gráficos que eu usei no aprofundamento, que são de densidade e histograma. A cauda é curta para ambos os lados. porém é maior pra baixo (*negative skewness*) e vemos que não há outliers. (Inicialmente eu até construí o boxplot com o ggplot, mas como vi que não tem qse informação relevante alguma deixei com o código mais limpo e simples possível).

Código em R na íntegra

```
## Preparação do Ambiente
library('readtext')
library('tidyverse')

## Leitura do Arquivo
idades_text <- readtext(file = 'idades.docx')
glimpse(idades_text)

## Preparação do texto para dataframe numérico para análise
tmp <- strsplit(x=idades_text$text, split=", ")
glimpse(tmp)
```

```r
tmp <- as.data.frame(tmp)
glimpse(tmp)
names(tmp) <- 'idades'
tmp$idades <- as.numeric(tmp$idades)
glimpse(tmp)

## Análise - Estatística Descritiva
summary(tmp)
tmp %>% summarize(
  min = min(idades),
  max = max(idades),
  mean = mean(idades),
  median = median(idades),
  var = var(idades),
  sd = var(idades) ^ (1/2))

## Análise - Gráfica
tmp %>% boxplot()

tmp %>% ggplot(aes(idades)) +
  geom_density(fill = "blue")

tmp %>% ggplot(aes(idades)) +
  geom_histogram(fill = "blue")
```