A primeira coisa que a gente pode fazer para entender os dados é ordenar as ocorrências.

Depois a gente pode usar algumas técnicas interessantes para visualizar, como por exemplo:

Diagrama de Folhas e Ramos, que é um **diagrama de ramos e folhas** é um dispositivo para apresentação de [dados quantitativos](https://pt.wikipedia.org/wiki/Escala_(estat%C3%ADstica)) em um formato [gráfico](https://pt.wikipedia.org/wiki/Infografia), semelhante a um [histograma](https://pt.wikipedia.org/wiki/Histograma), que ajuda a visualizar a forma de uma [distribuição](https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_de_probabilidade).

Sua popularidade durante estes anos pode ser atribuída ao uso de fontes [mono-espaçadas](https://pt.wikipedia.org/wiki/Monospace) que permitiam à tecnologia computacional da época produzir facilmente o gráfico.

#Obter a Idade

install.packages("eeptools")

library("eeptools")

age\_calc(as.character.Date(mydata$D\_Nasc))

floor(age\_calc(as.Date(mydata$D\_Nasc, format="%d/%m/%Y"),units = "years"))

tbl\_idade <- cbind(as.data.frame(mydata$Nome), as.data.frame(floor(age\_calc(as.Date(mydata$D\_Nasc, format="%d/%m/%Y"), units = "years"))))

colnames(tbl\_idade) <-c("NOME","IDADE")

# Diagrama de Folhas e Dados

stem(tbl\_idade$IDADE)

The decimal point is 1 digit(s) to the right of the |

1 | 99

2 | 0000111223334445778

3 |

4 | 5

Ainda melhor que o diagrama de ramos e folhas é o histograma:

O **histograma**, também conhecido como **distribuição de frequências**, é a representação gráfica em colunas ou em barras (retângulos) de um conjunto de dados previamente tabulado e dividido em classes uniformes ou não uniformes. A base de cada retângulo representa uma classe. A altura de cada retângulo representa a quantidade ou a frequência absoluta com que o valor da classe ocorre no conjunto de dados para classes uniformes ou a densidade de frequência para classes não uniformes

*Nota: A palavra "histograma" se origina dos termos gregos: istos (mastro) e gramma (escrita) e pode ser interpretada como uma escrita em longas formas verticais ou algo semelhante.*

**

Não há um número ideal de intervalos de classe. Diferentes tamanhos de intervalos de classe podem revelar diferentes características dos dados. Usando intervalos de classe mais largos, a baixa densidade reduz o ruído devido à aleatoriedade da amostragem. Usando intervalos de classe mais estreitos, a densidade elevada garante maior precisão à estimativa da densidade. Então, variar a largura do intervalo de classe em um histograma pode ser benéfico. Entretanto, intervalos de classe com larguras iguais são amplamente utilizados.

Como Construir:

1. Organizar os dados em ordem crescente.
2. Definir os intervalos e o número de classes.
3. Construir o gráfico.

Um histograma cumulativo é um mapeamento que conta o número cumulativo de observações em todos os intervalos de classe até o intervalo de classe especificado.

Histograma é amplamente usado para processamento de imagem:

#HISTOGRAMA DE UMA IMAGEM

install.packages("imager")

library(imager)

#tem uma imagem no pacote chamada boats

plot(boats)

class(boats)

library(ggplot2)

bdf <- as.data.frame(boats)

head(bdf,3)

bdf <- plyr::mutate(bdf,channel=factor(cc,labels=c('R','G','B')))

ggplot(bdf,aes(value,col=channel))+geom\_histogram(bins=30)+facet\_wrap(~ channel)



Histograma Distorcido a Esquerda

Histograma Distorcido a Direita

Histograma Bimodal, Multimodal, Platô

Exercícios:

1. Fazer o histograma da idade com 6 classes e com 12 classes
2. Fazer um histograma acumulado.

A **análise exploratória de dados (AED)**

é uma abordagem à [análise](https://pt.wikipedia.org/wiki/An%C3%A1lise_de_dados) de [conjuntos de dados](https://pt.wikipedia.org/wiki/Conjunto_de_dados) de modo a resumir suas características principais, frequentemente com métodos visuais.

A análise exploratória de dados emprega grande variedade de técnicas gráficas e quantitativas, visando maximizar a obtenção de informações ocultas na sua estrutura, descobrir variáveis importantes em suas tendências, detectar comportamentos anômalos do fenômeno, testar se são válidas as hipóteses assumidas, escolher modelos e determinar o número ótimo de variáveis.

análise inicial de dados (AID)

Em [estatística](https://pt.wikipedia.org/wiki/Estat%C3%ADstica), uma **tendência central** (ou, normalmente, uma **medida de tendência central**) é um valor central ou valor típico para uma [distribuição de probabilidade](https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_de_probabilidade).[[1]](https://pt.wikipedia.org/wiki/Tend%C3%AAncia_central#cite_note-Weisberg-1) É chamada ocasionalmente como [média](https://pt.wikipedia.org/wiki/M%C3%A9dia) ou apenas **centro** da distribuição. As medidas de tendência central mais comuns são a [média aritmética](https://pt.wikipedia.org/wiki/M%C3%A9dia_aritm%C3%A9tica), a [mediana](https://pt.wikipedia.org/wiki/Mediana_(estat%C3%ADstica)) e [moda](https://pt.wikipedia.org/wiki/Moda_(estat%C3%ADstica)).

(ou **centralidade**)

Tendências centrais podem ser calculadas tanto para um número finito de valores quanto para uma distribuição teórica, a exemplo da [distribuição normal](https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_normal).

[Média aritmética](https://pt.wikipedia.org/wiki/M%C3%A9dia_aritm%C3%A9tica) (ou simplesmente, média) (Mean) - a soma de todas as medições divididas pelo número de observações no conjunto de dados.

Exercício: a) Calcular a Média em minutos do tempo de viagem.

Mediana (median).: o valor do meio que separa a metade maior da metade menor no conjunto de dados. A mediana e a moda são as únicas medidas de tendência central que podem ser usadas para [nível de medição](https://pt.wikipedia.org/w/index.php?title=N%C3%ADvel_de_medi%C3%A7%C3%A3o&action=edit&redlink=1), onde valores recebem ranks relativos aos outros mas não são medidos absolutamente.

Exercícios:

1. criar uma coluna em mydata em R que faça o rank das viagens mais demoradas.
2. Calcular a mediana.

[Moda](https://pt.wikipedia.org/wiki/Moda_(estat%C3%ADstica)) - O valor que aparece com mais frequência no conjunto de dados. Essa é a única medida de tendência central que pode ser usada com [dados nominais](https://pt.wikipedia.org/w/index.php?title=Dados_nominais&action=edit&redlink=1), os quais tem atribuições de categoria puramente qualitativa.

Medidas de Espalhamento

A primeira coisa que podemos obter é obter o alcance e a amplitude dos dados, mais uma vez facilita ordenar os dados antes.

min(tbl\_idade$IDADE)

max(tbl\_idade$IDADE)

range(tbl\_idade$IDADE)

Em [Estatística](https://pt.wikipedia.org/wiki/Estat%C3%ADstica), **dispersão** (também chamada de **variabilidade** ou **espalhamento**) mostra o quão esticada ou espremida[[1]](https://pt.wikipedia.org/wiki/Dispers%C3%A3o_estat%C3%ADstica#cite_note-1) uma [distribuição](https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_de_probabilidade) (teórica ou que define uma [amostra](https://pt.wikipedia.org/wiki/Amostragem_(estat%C3%ADstica))).

Definir Distribuição teórica ou de amostra.

se as medidas das amostras são feitas em metros ou segundos, a medida de dispersão também serão.

Variância: é uma medida da sua [dispersão estatística](https://pt.wikipedia.org/wiki/Dispers%C3%A3o_estat%C3%ADstica), indicando "o quão longe" em geral os seus valores se encontram do [valor esperado](https://pt.wikipedia.org/wiki/Valor_esperado)

Visualização de Dados

Em [estatística descritiva](https://pt.wikipedia.org/wiki/Estat%C3%ADstica_descritiva), **diagrama de caixa**, **diagrama de extremos e quartis,** ***boxplot*** ou ***box plot*** é uma ferramenta gráfica para representar a variação de dados observados de uma variável numérica por meio de [quartis](https://pt.wikipedia.org/wiki/Quartil) (ver figura 1, onde o eixo horizontal representa a variável). O boxplot tem uma reta (*whisker* ou fio de bigode) que estende–se verticalmente ou horizontalmente a partir da caixa, indicando a variabilidade fora do quartil superior e do quartil inferior.[[1]](https://pt.wikipedia.org/wiki/Diagrama_de_caixa#cite_note-1) Os valores atípicos ou [*outliers*](https://pt.wikipedia.org/wiki/Outlier)(valores discrepantes) podem ser plotados como pontos individuais

<https://pt.wikipedia.org/wiki/An%C3%A1lise_explorat%C3%B3ria_de_dados>

https://cran.r-project.org/web/packages/imager/vignettes/gettingstarted.html