Fundamentos de Análise Exploratória

Exemplos em R e Python

Autor: Rafael Barbosa da Silva

Data Science Belém/PA

Fundamentos de Análise Exploratória

O que é Análise Exploratória?

- Segundo **Magalhães** (2004), a grosso modo, é umas das 3 principais áreas da estatística
- Consiste em resumir e organizar as informações (dados)
- Exemplos mais comuns
- Utilizada no início da análise para retirar padrões e informações úteis dos dados

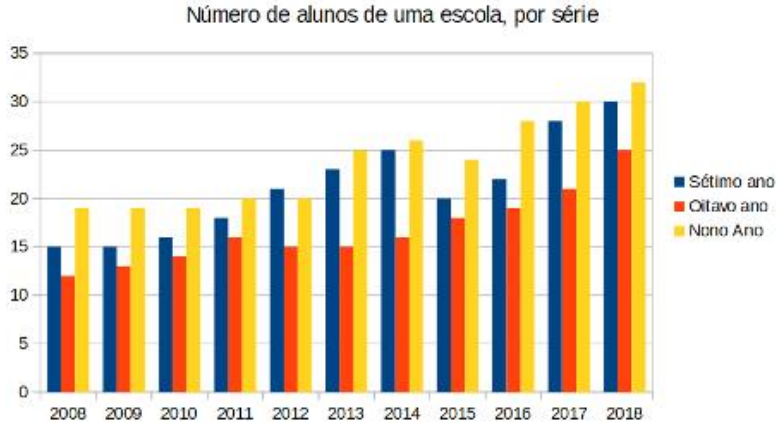


Tabela de contribuição dos segurados empregado, empregado doméstico e trabalhador avulso, a partir de 1º de janeiro de 2019

Salário de contribuição (R$)	Alíquota para fins de recolhimento ao INSS	Retenção
até R$ 1.751,81	8%	
de R$ 1.751,82 até R$ 2.919,72	9%	
de R$ 2.919,73 até R$ 5.839,45	11%	R$ 642,34
Desconto em cooperativa	20%	R$ 1.167,89

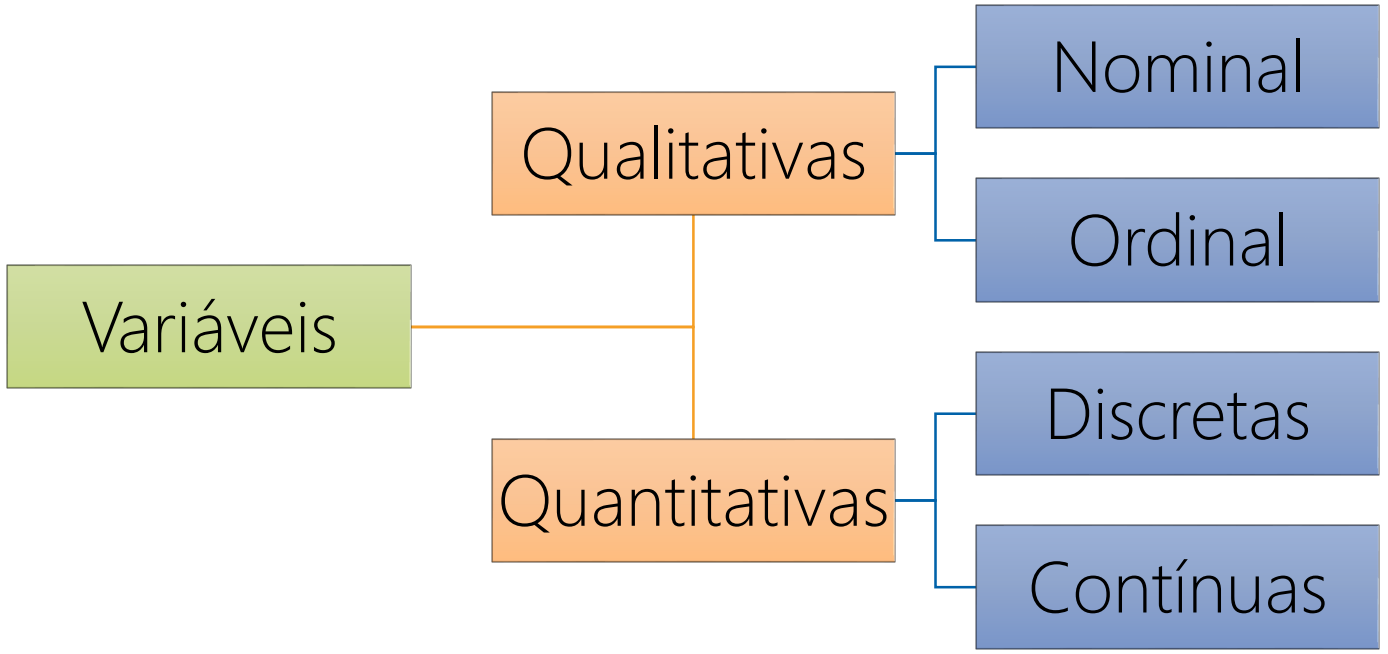
- Média
- Mediana
- Moda
- Variância
- Desvio Padrão
- Entre outras

Por quê a EDA é importante?

- Familiarização e conhecimento dos dados
- Melhor eficiência na aplicação dos modelos
- Respostas para as hipóteses iniciais da problema de negócio

Variáveis e seus tipos

- Informação da características de interesse
- Exemplo: Peso, altura, etc.
- Cada uma destas características podem ser divididas em categorias



T. E.: Arredondamento numérico

- Utilizado quando for conveniente eliminar unidades
- Regras propostas pelo IBGE em 1993
- Notas:
  - Nada de arredondamentos sucessivos
  - Em casos de porcentagens, quando queremos somar 100%

Regra 1 - Quando o primeiro algarismo a ser abandonado for: 0, 1, 2, 3 e 4. Fica inalterado o último algarismo a permanecer. Por exemplo, arredondar para uma casa decimal os números:

$53{,}24 \rightarrow 53{,}2$; $88{,}01 \rightarrow 88{,}0$; $10{,}43 \rightarrow 10{,}4$

Regra 2 - Quando o primeiro algarismo a ser abandonado for: 5, 6, 7, 8 ou 9. Aumenta-se uma unidade no algarismo a permanecer. Por exemplo, arredondar para uma casa decimal os números:

$53{,}25 \rightarrow 53{,}3$; $88{,}09 \rightarrow 88{,}1$; $10{,}47 \rightarrow 10{,}5$

Fundamentos de Análise Exploratória

Arredondamentos sucessivos

Objetivo: Arredondar o número 17,3452 para 1 casa decimal.

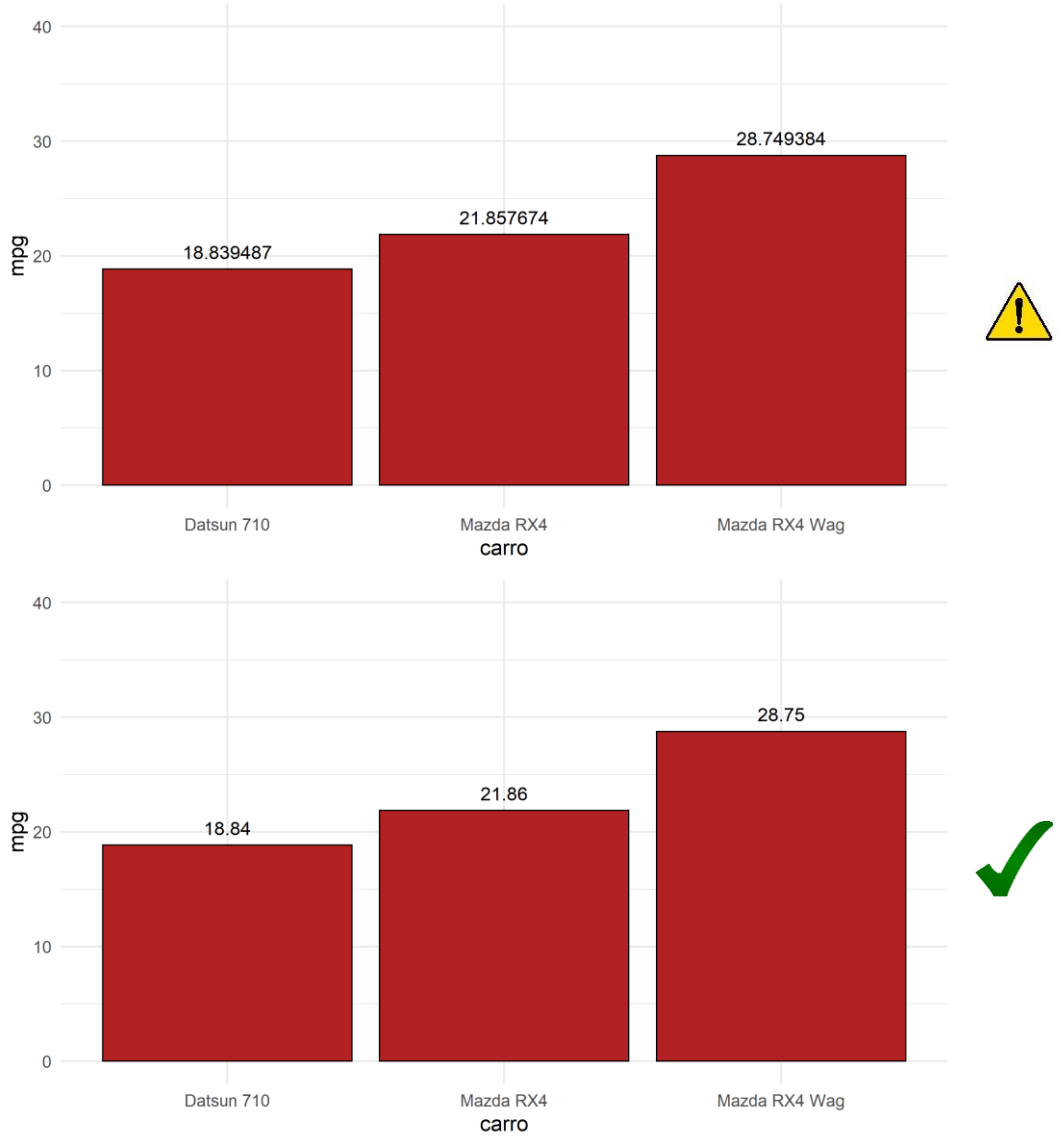
Errado:

17,3452 → 17,35 → 17,4

Correto:

17,3452 → 17,3

Arredondamentos em gráficos



T. E.: Missing data

- Informações que não estão disponíveis nos dados

- Principais motivos:

  - Não gostaria de responder determinada pergunta e deixou em branco
  - Alguém se esqueceu de registrar a informação

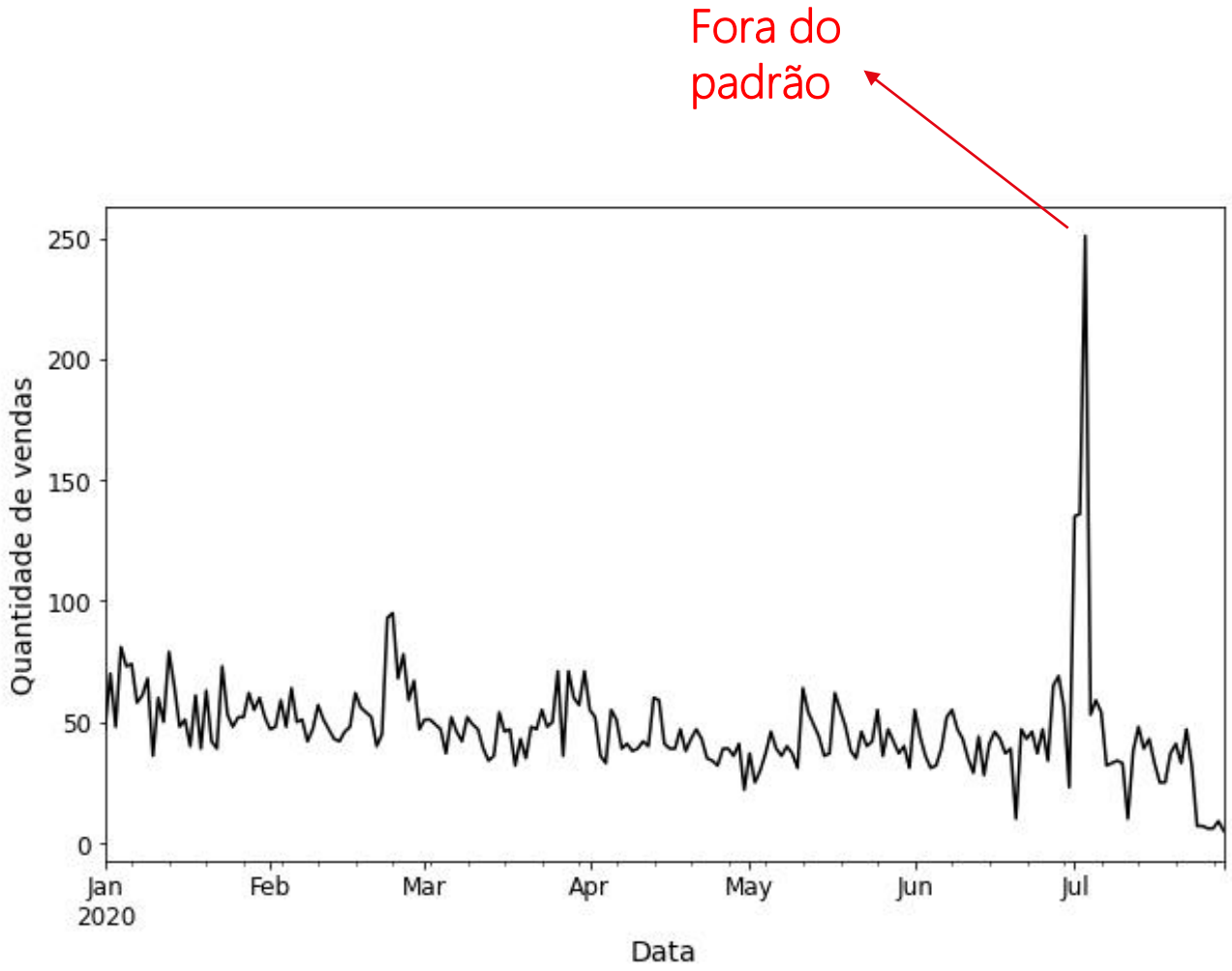
- Soluções:

  - Eliminação completa do registro
  - Preenchimento com a média, mediana, moda, interpolação, entre outros
  - Criação de uma categoria "SI"

F1	F2	F3	F4	F5	Class
Good	20	5	7	Old	Normal
Good	Missing	8	8	Old	Normal
Good	15	10	10	Old	Normal
Good	50	10	10	Old	Normal
Good	70	10	10	Old	Abnormal
Bad	20	5	7	Old	Abnormal
Good	20	5	80	Old	Abnormal
Good	85	100	100	Old	Abnormal
Good	20	100	Missing	Old	Abnormal
Good	24	6	8.4	Old	Normal
Good	12	9.6	9.6	Old	Normal
Good	18	12	12	Old	Normal
Good	60	12	12	Old	Normal
Good	84	Missing	12	Old	Abnormal
Bad	24	6	8.4	Old	Abnormal
Good	24	6	96	Old	Abnormal
Good	102	120	120	Old	Abnormal
Good	24	120	72	Old	Abnormal

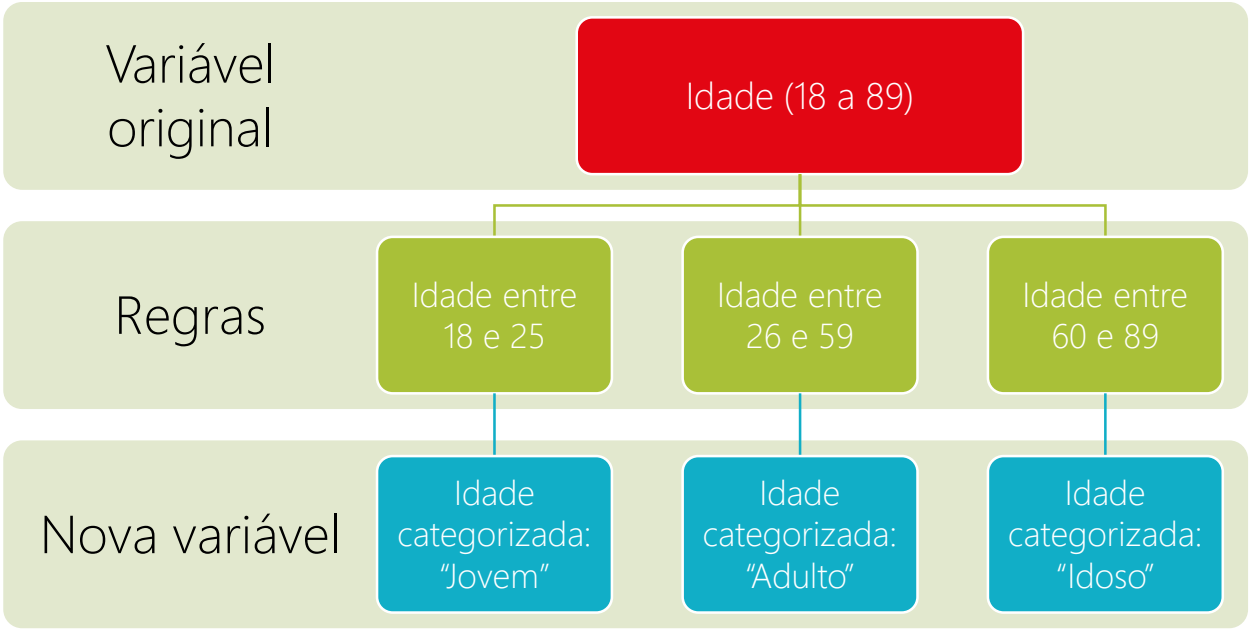
T. E.: Outliers

- Especificamente para variáveis quantitativas
- Valores discrepantes de uma variável que possui um comportamento "padrão"
- Motivos:
  - Erros humanos
  - Erros de mensuração
- Detecção de outliers:
  - Gráficos exploratórios
  - Z-score
  - Algoritmos sofisticados (Dbscan, Isolation Forests, entre outros)



T. E.: Recodificação

- Criação de novas variáveis a partir de uma já existente
- Quando utilizar:
  - Variável Qualitativa com muitas categorias
  - Variável Quantitativa será transformada para **qualitativa** ⚠



T. E.: Transformação

$$\frac{TC}{5} = \frac{(TF - 32)}{9}$$

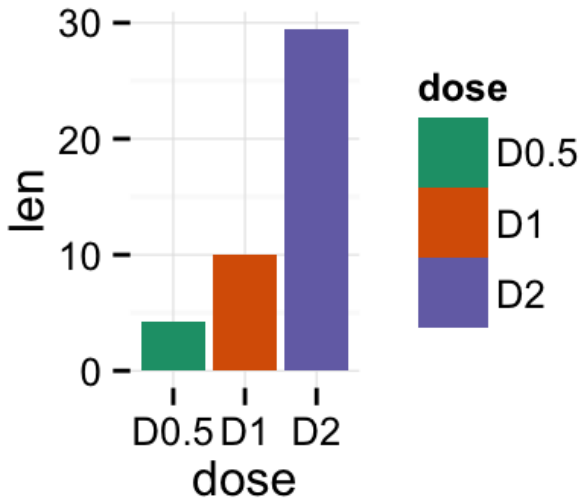
TC = Temperatura em Celsius

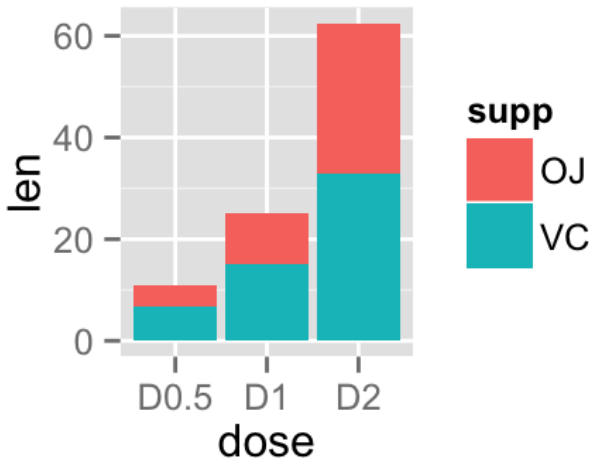
TF = Temperatura em Fahrenheit

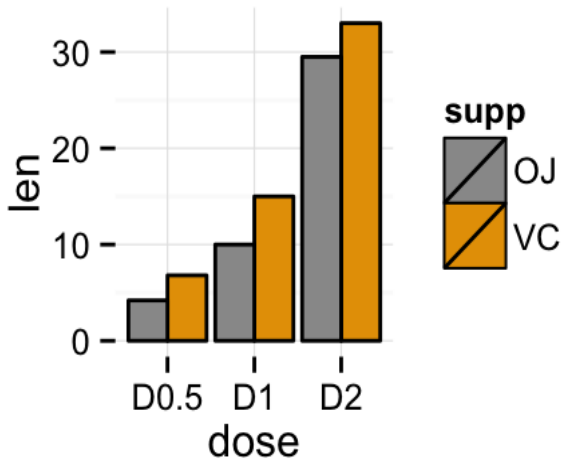
- Comumente utilizada em Variáveis Quantitativas
- Novas variáveis por meio de expressões matemáticas
- Exemplo comum: Celsius → Fahrenheit

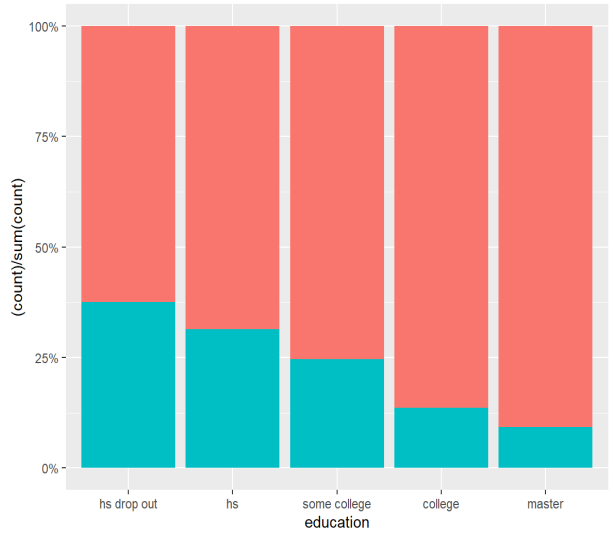
	A	B	C	D
1				
2				
3				
4		**Conversão de temperatura**		
5				
6		Grau Celsius	Grau Fahrenheit	
7		-10	14	
8		-2		
9		4		
10		15		
11		0		
12		22		
13		29		
14		35		
15		42		
16				
17				
18				
19				
20				

Gráficos: Barras

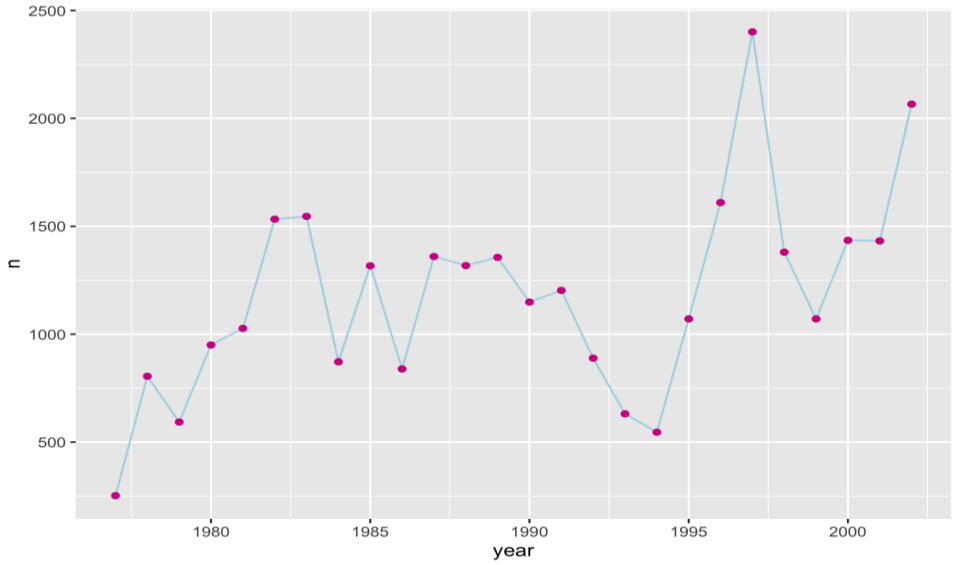


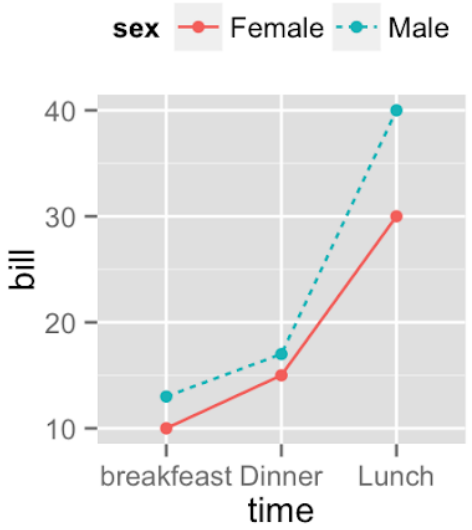


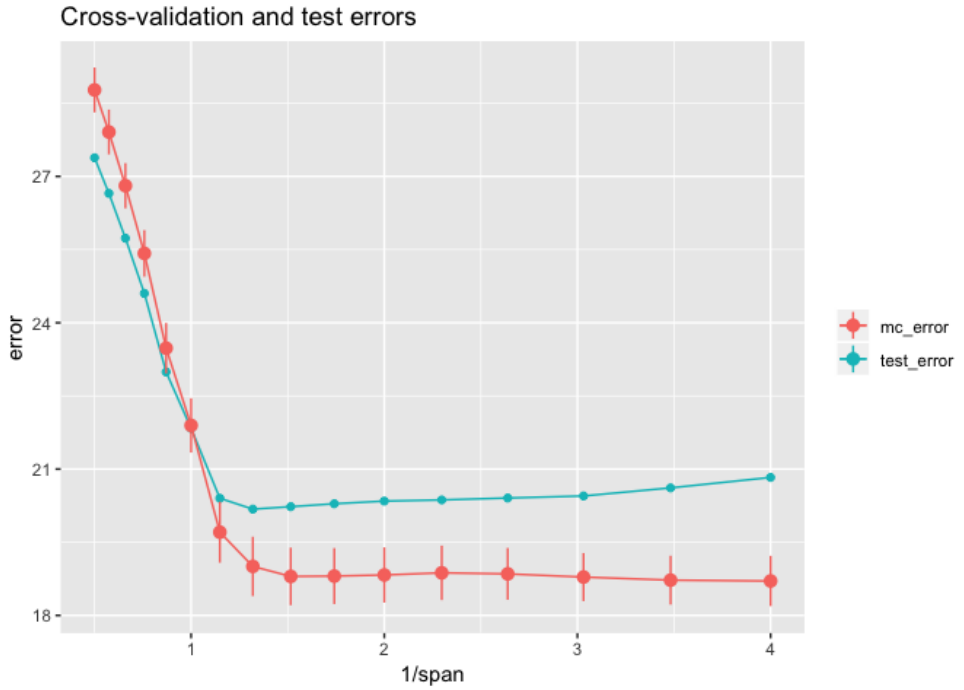




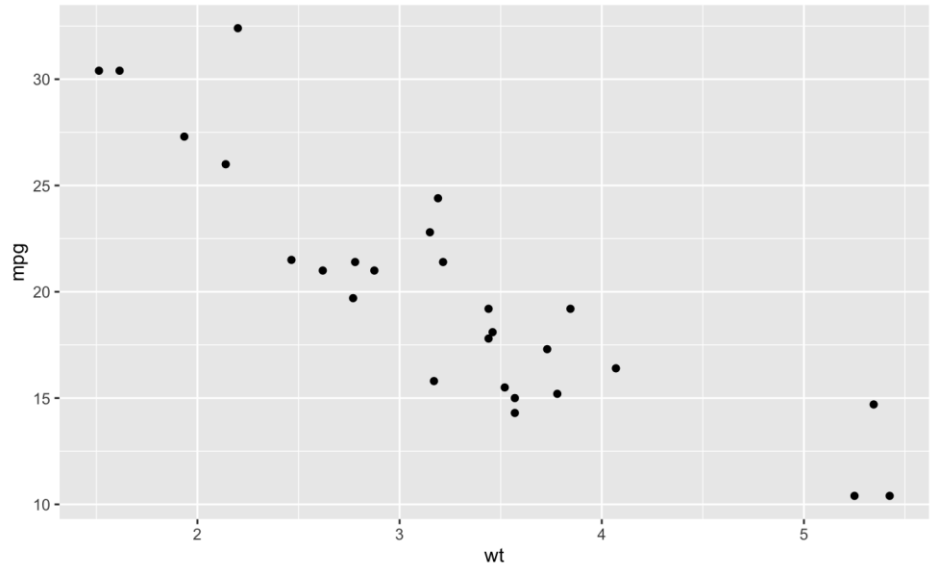
Gráficos: Linhas

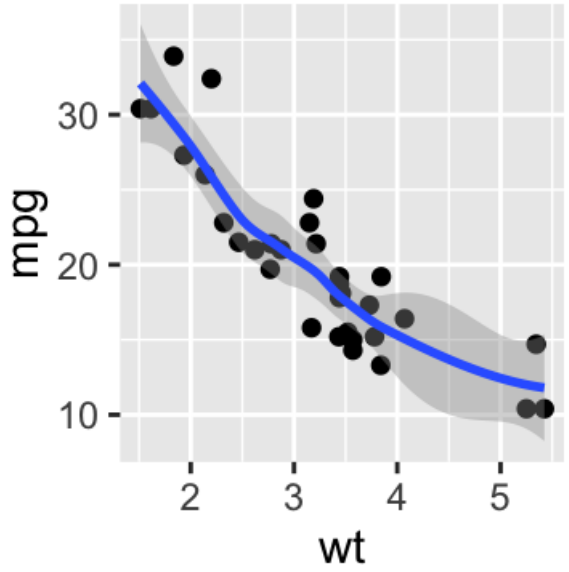


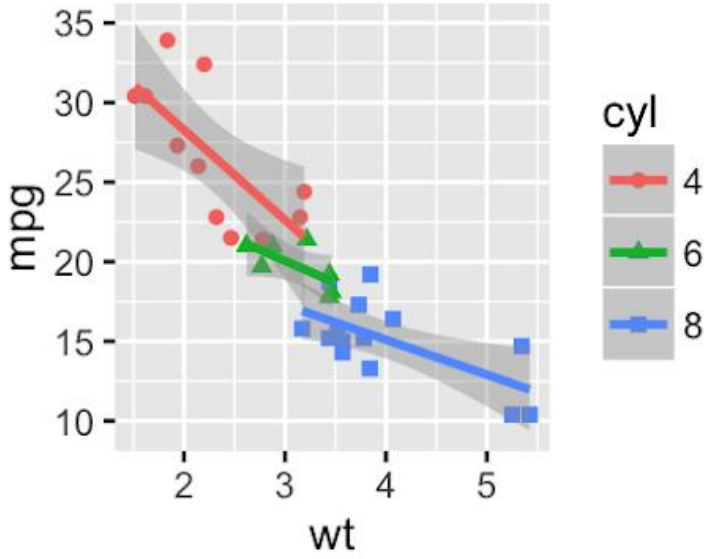




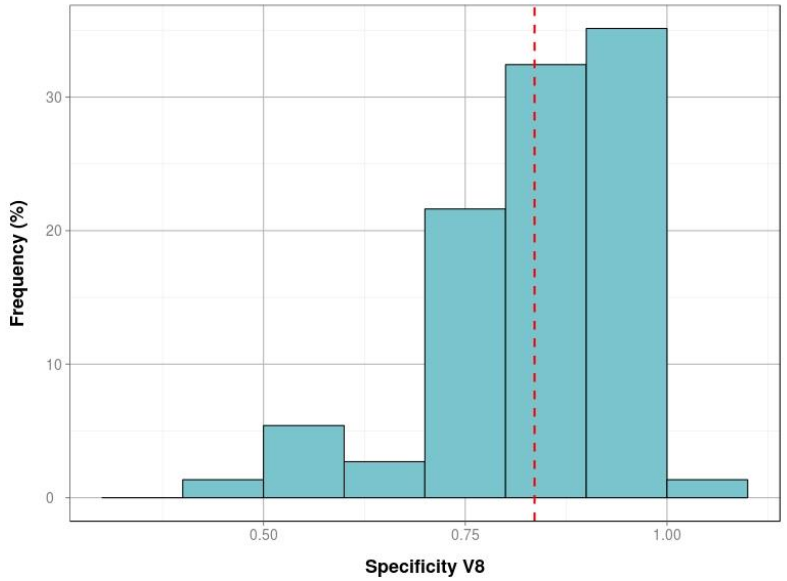
Gráficos: Pontos (Scatterplot)

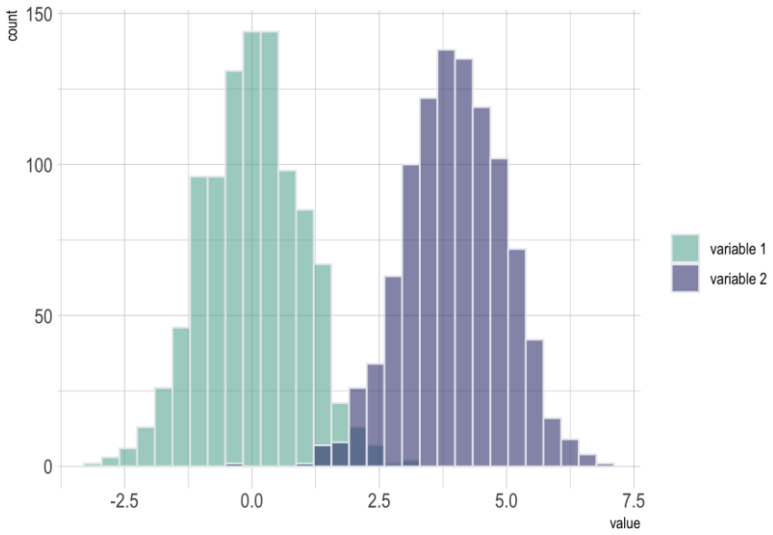


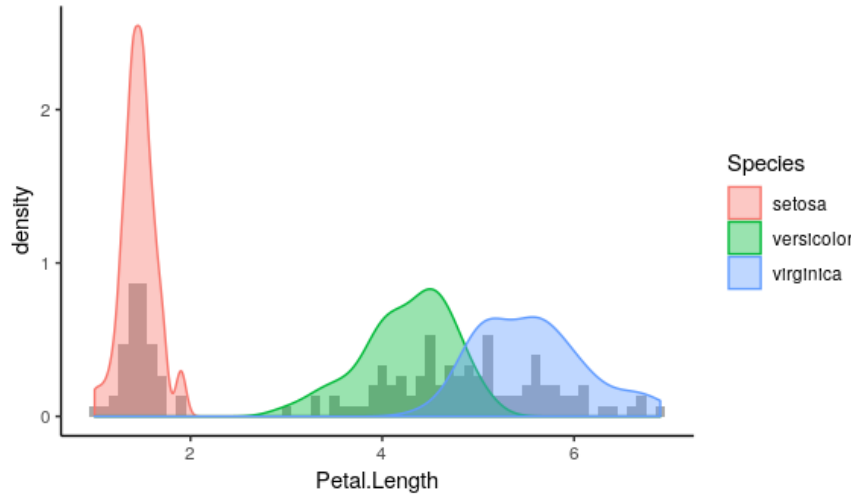




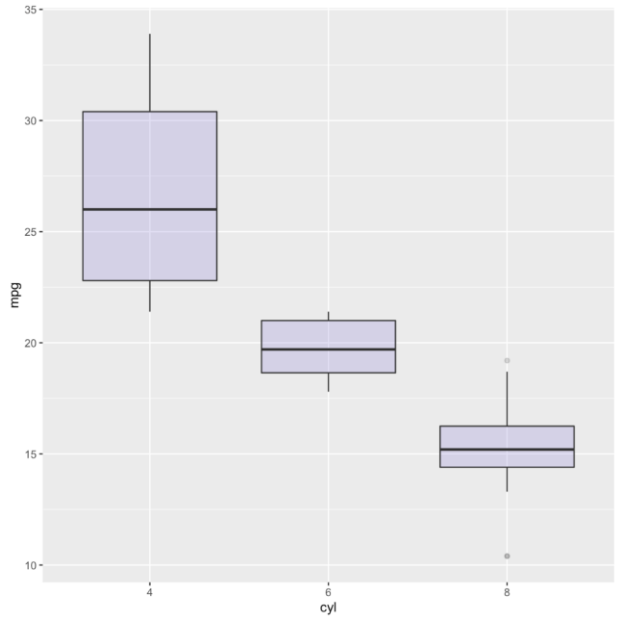
Gráficos: Histograma

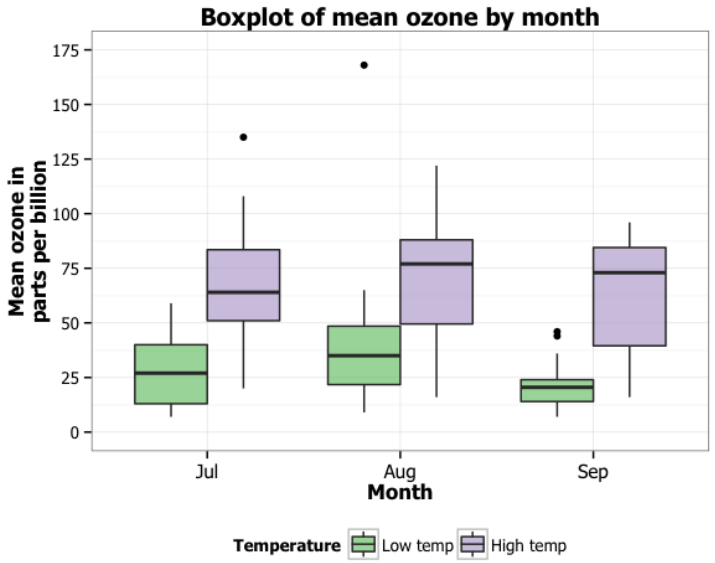


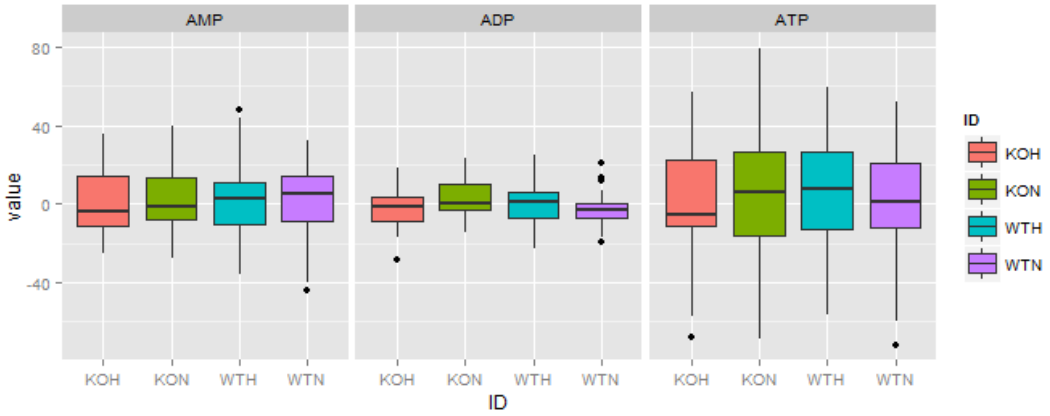




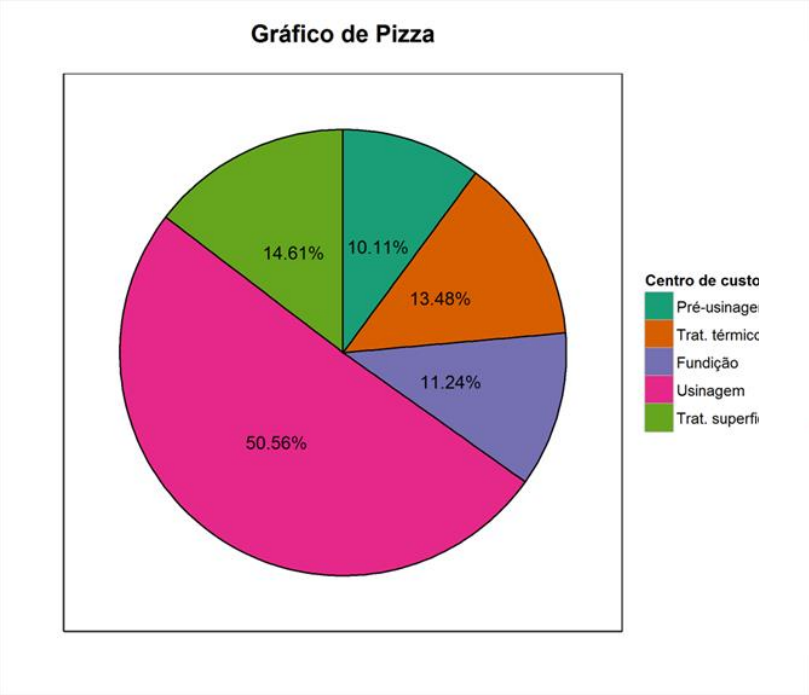
Gráficos: Boxplot

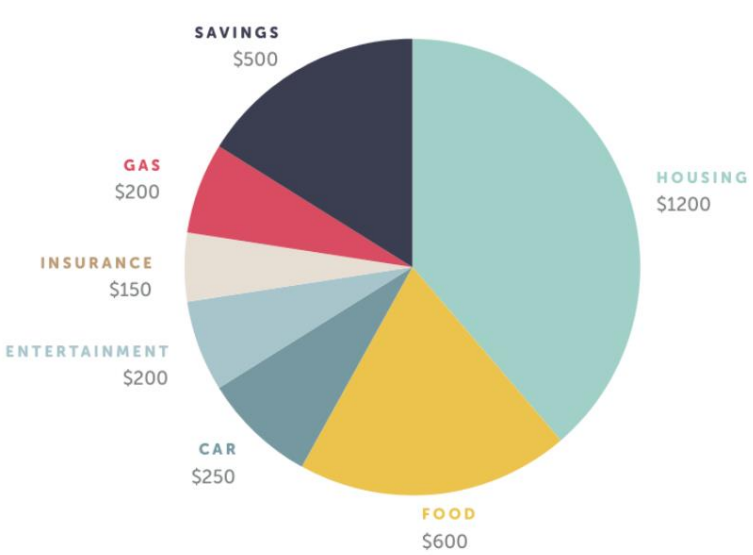


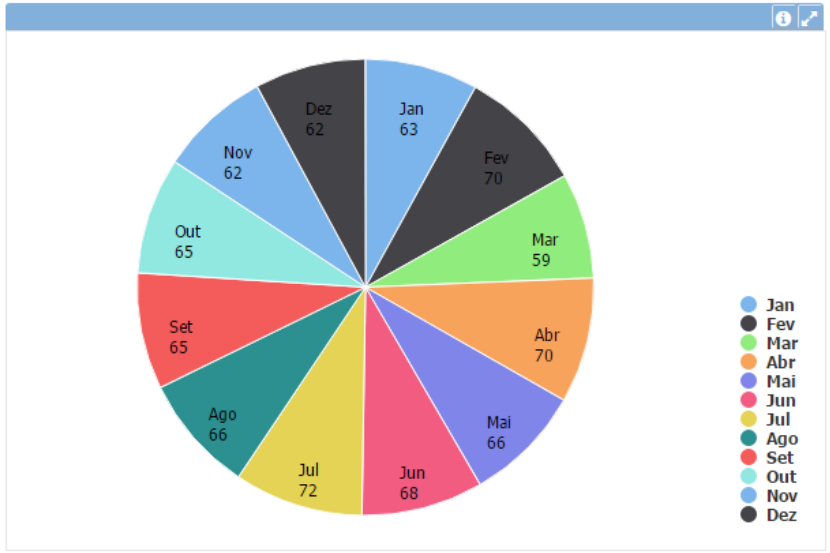




Gráficos: Setores (Pizza)







Tabela

- Maneira de apresentar de forma resumida um conjunto de observações
- Elementos:
  - Cabeçalho
  - Colunas
  - Corpo

Tabela 4: Número e porcentagem de causas de morte de residentes de Londrina, no período de 10 de agosto a 31 de dezembro de 2008

CAUSAS DA MORTE	Nº	%
Doenças do ap. circulatório	281	33,5
Neoplasias	115	13,7
Causas externas	92	11,0
Doenças do ap. respiratório	87	10,4
Doenças das glând. endóc./transt. Imunitários	56	6,7
Doenças do ap. digestivo	54	6,4
Doenças e infec. e parasitárias	46	5,5
Afecções do per. Perinatal	26	3,1
Demais grupos	82	9,8
TOTAL	**839**	**100,0**

FONTE: Núcleo de informação em mortalidade – PML

Tabela ou gráfico?

- Grandeza, variação, tendência → Gráfico
- Detalhes, valores importantes → Tabela
- Dica: Se possível, priorize os gráficos

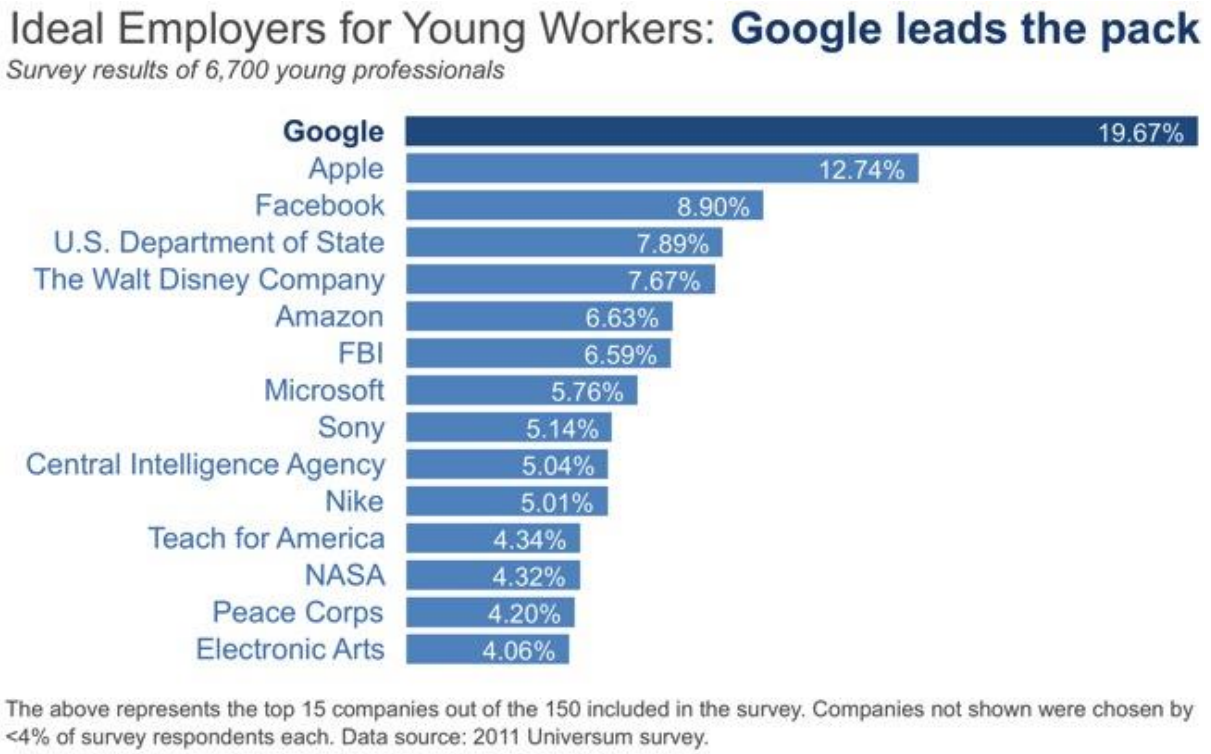
Tabela ou gráfico?

What's Your Dream Company?

Consulting firm Universum asked some 6,700 young professionals with one to eight years of work experience to choose their ideal employers from a list of firms. Here's the percentage of respondents that chose each company.

<< first < prev 1 2 3 next > last >>

Company	Rank	Percent
Google	1	19.67%
Apple	2	12.74%
Facebook	3	8.90%
U.S. Department of State	4	7.89%
The Walt Disney Company	5	7.67%
Amazon	6	6.63%
FBI	7	6.59%
Microsoft	8	5.76%
Sony	9	5.14%
Central Intelligence Agency	10	5.04%
Nike	11	5.01%
Teach for America	12	4.34%



Fundamentos de Análise Exploratória

Medidas de tendência central

- Média:
  - Soma de todos os valores da variável dividida pelo número de observações
  - Em todos os casos, existe somente uma média aritmética
  - Valores extremos influenciam
- Mediana:
  - Valor que ocupa a posição central de um conjunto de valores ordenados
  - Não é afetada por valores extremos
- Moda:
  - Valor que ocorre com mais frequência
  - Aplicada em Variáveis **Quantitativas** ou **Qualitativas**

Exemplo: Tempo de espera (em minutos) para consulta em determinado hospital.

- **Tempo medido**: 37, 21, 78, 64, 35, 123, 37, 54.

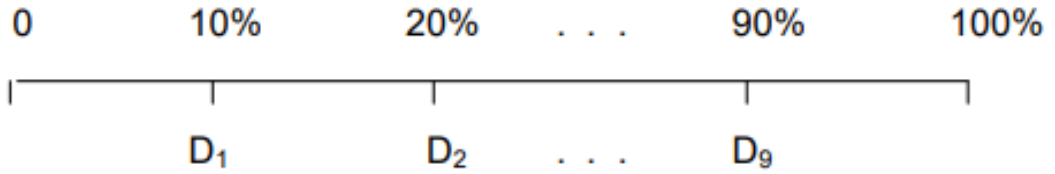
- O tempo médio de espera para consulta é de 56,13 minutos
- O tempo mediano de espera para consulta é de 45,5 minutos
- A maioria dos pacientes foram atendidos em 37 minutos

Medidas de dispersão

- Variância:
  - Mensura a dispersão dos dados em torno da média
  - Sua **unidade** é o quadrado da unidade dos dados
  - Quanto maior a variância, mais dispersos seus dados estão em torno da média

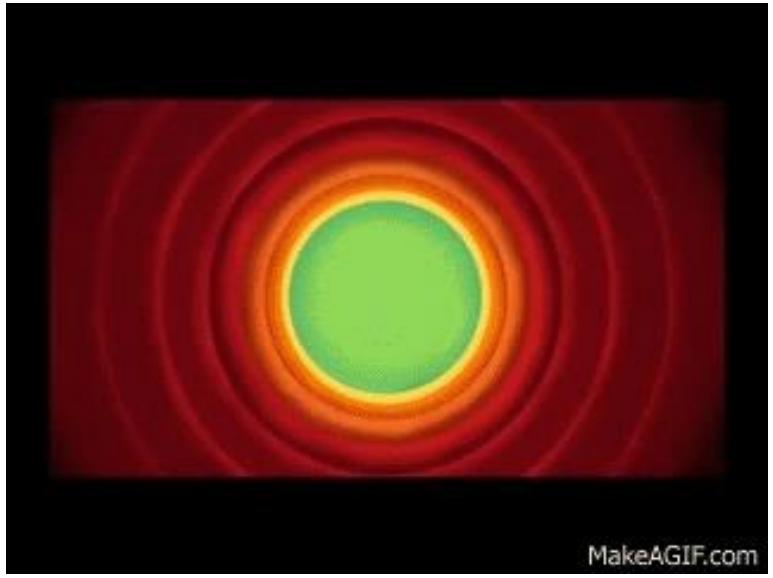
Medidas separatrizes

- Quartis, Decis e Percentis:
  - Dividem o conjunto de dados (em ordem) em *i* (4, 10 ou 100) partes iguais
  - Mediana sempre vai ser o quartil 2, 5 ou 50



Única consideração

- Cuidado com o excesso de Análise Exploratória de Dados



Obrigado