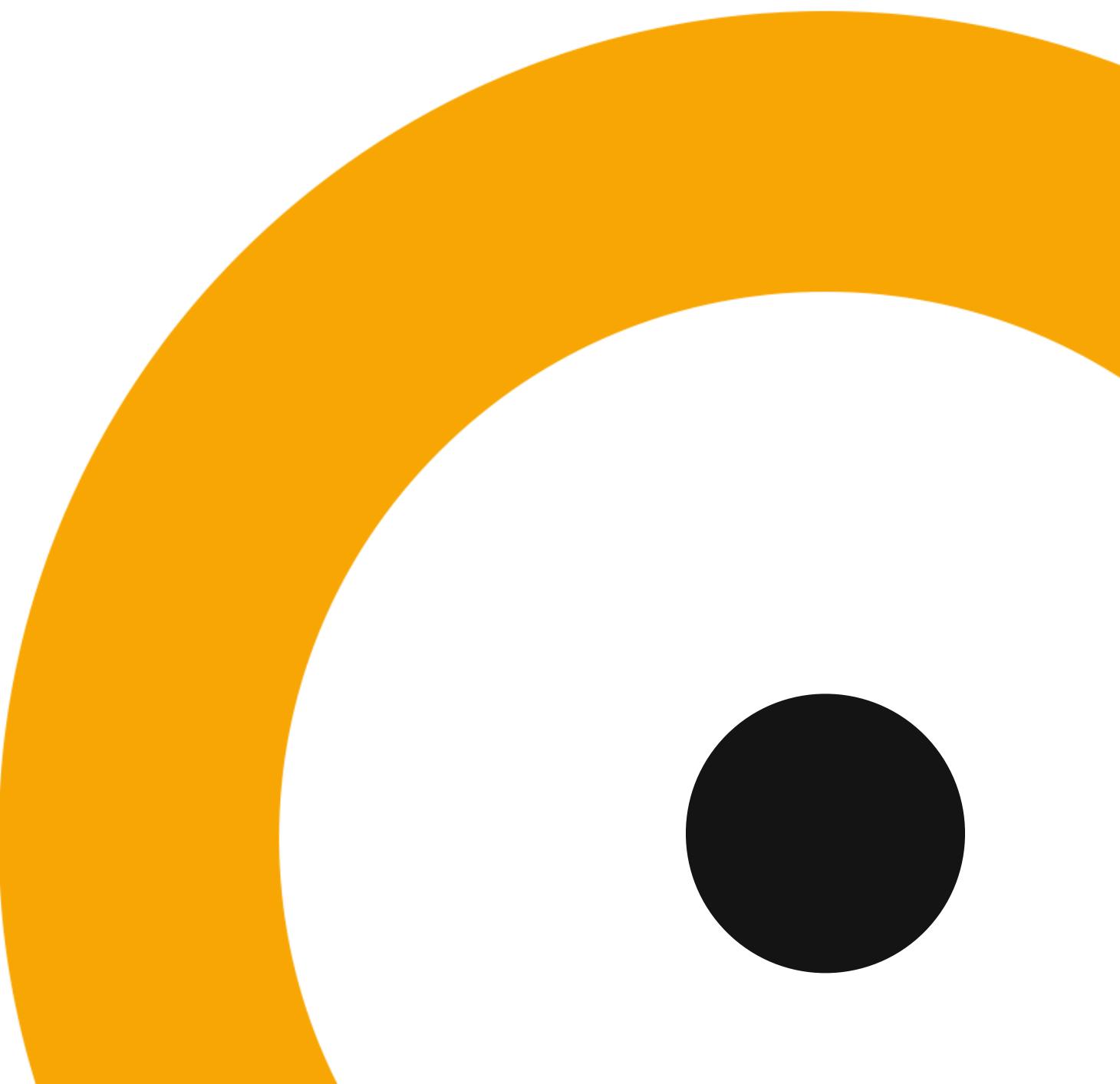


Mestrado Data Science

# Data Cleaning

**Módulo: Data Cleaning and Transformation in Python**





# Fábio Ferreira

Data Science Team Leader

---

**Formação:**

Mestrado em Engenharia Matemática – FCUP  
Pós-Graduação em BI & Analytics – PBS

**Experiência Profissional:**

2016-2017: Investigador em Data Science – FMUP  
2017-2018: Consultor em Implementação de Seguros – i2S  
2018-2022: Data Scientist – BNP Paribas PF  
2022-momento: Team Leader – BNP Paribas PF

# O nosso módulo



# Conteúdo

Sessão nº 1

01

Motivação

---

02

Tipos de dados

---

03

Outliers

---

04

Dados em falta

---

05

Correlações

---

06

Possíveis perguntas numa entrevista

# 01

## Motivação

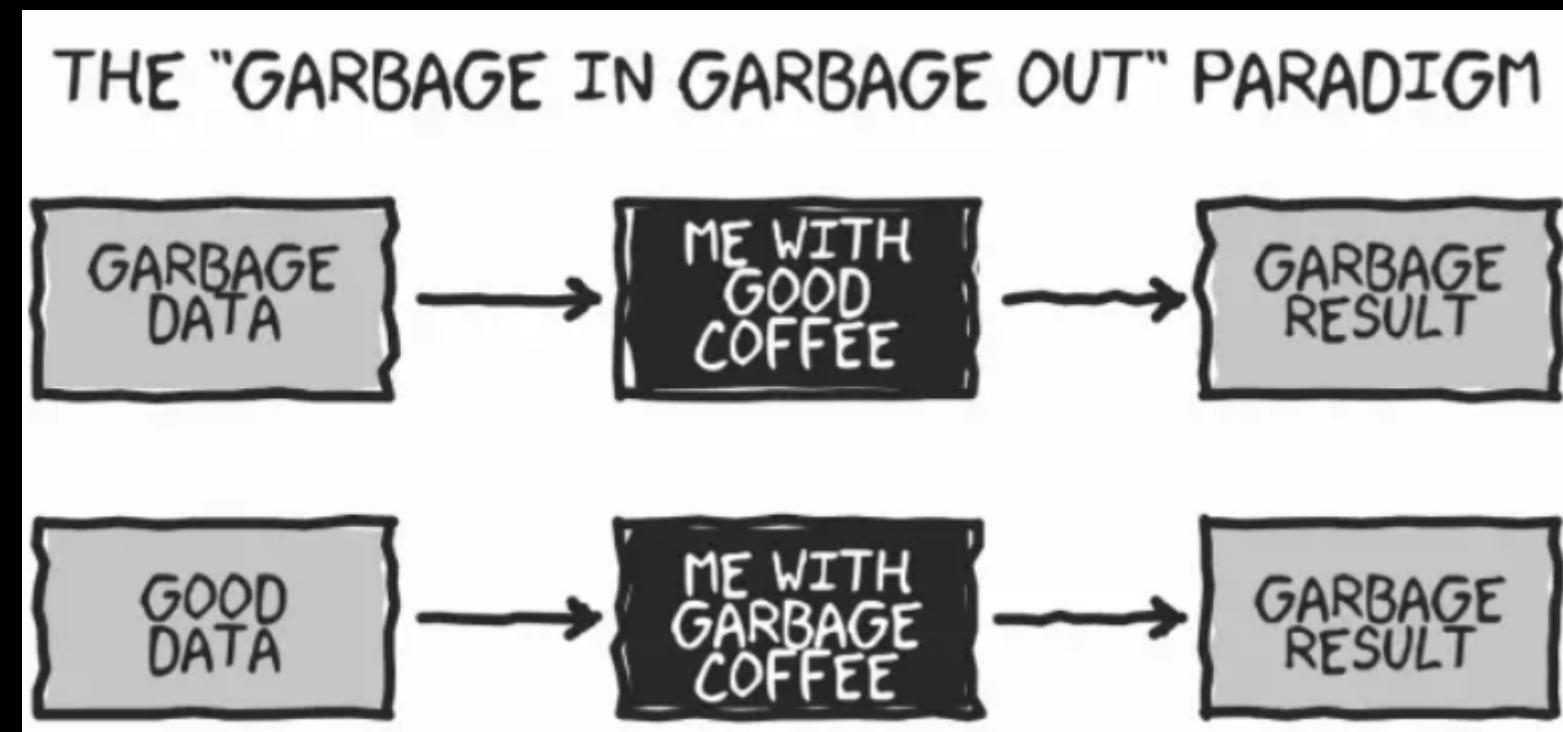
Porque é importante limpamos os dados?



“

**Data scientists , (...) spend from 50 percent to 80 percent of their time (...) collecting and preparing unruly digital data, before it can be explored for useful nuggets.** ”

New York Times, 80-20 rule, 2014



# Diferentes problemas de qualidade de dados

## 1. Precisão - representam a verdade?

### Exemplos:

- Age = 160
- Salary = -10

## 2. Completude – foram todos registados?

### Exemplos:

- Colaborador não preencheu um campo
- Equipamento parou

## 3. Consistencia – são coerentes?

### Exemplos:

- Age = 50; Birthday = “03/10/2001”
- Rating na tabela 1: {1, 2, 3}; Rating na tabela 2: {A, B, C}

## 4. Tempo – estão disponíveis?

### Exemplos:

- Informação do dia anterior indisponível

## 5. Únicos – registos duplicados?

### Exemplos:

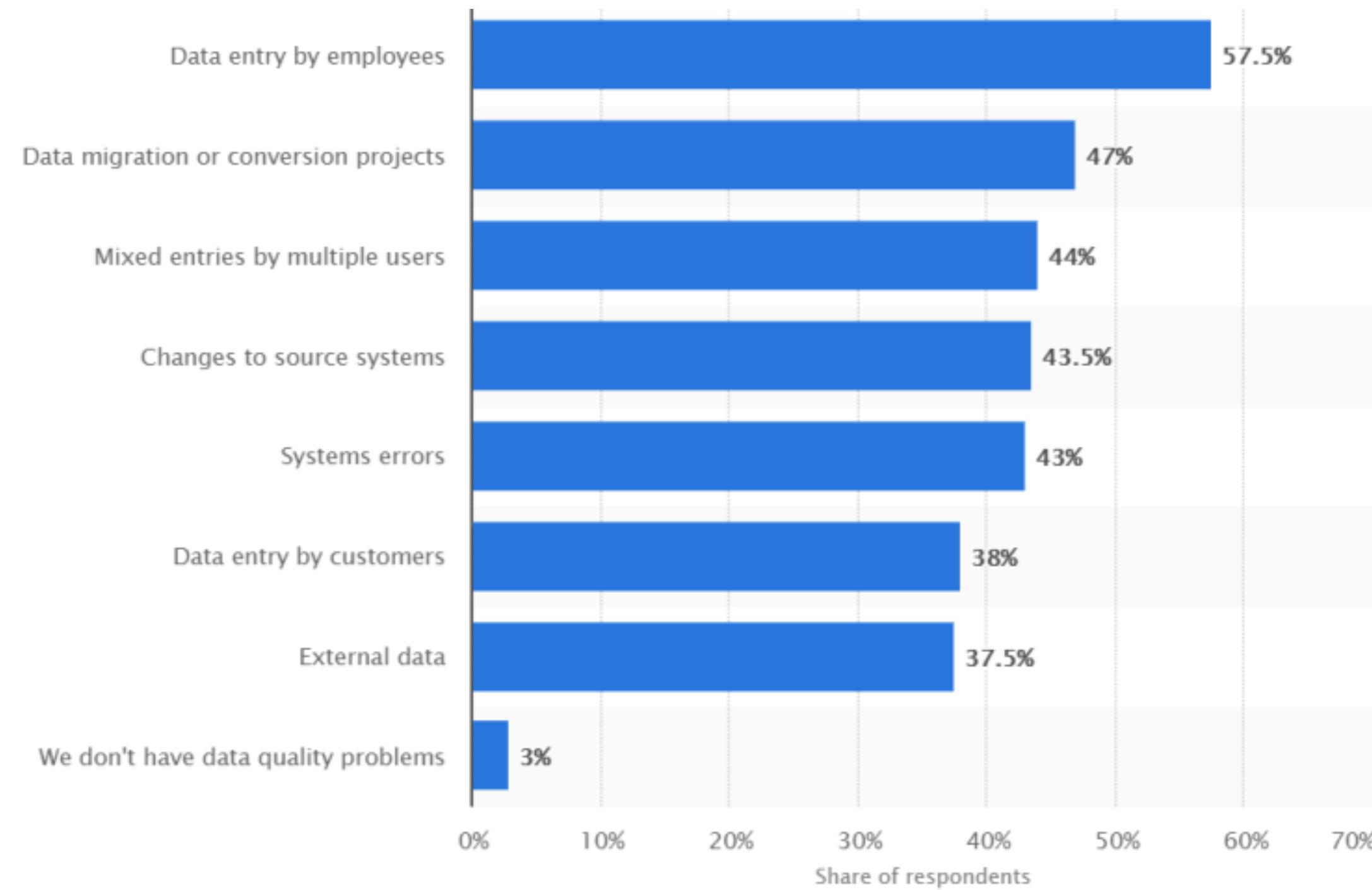
- Cliente A duplicado e com diferentes salários

## 6. Válidos – obedecem as regras?

### Exemplos:

- Obedece ao formato, tipo e amplitude esperados?

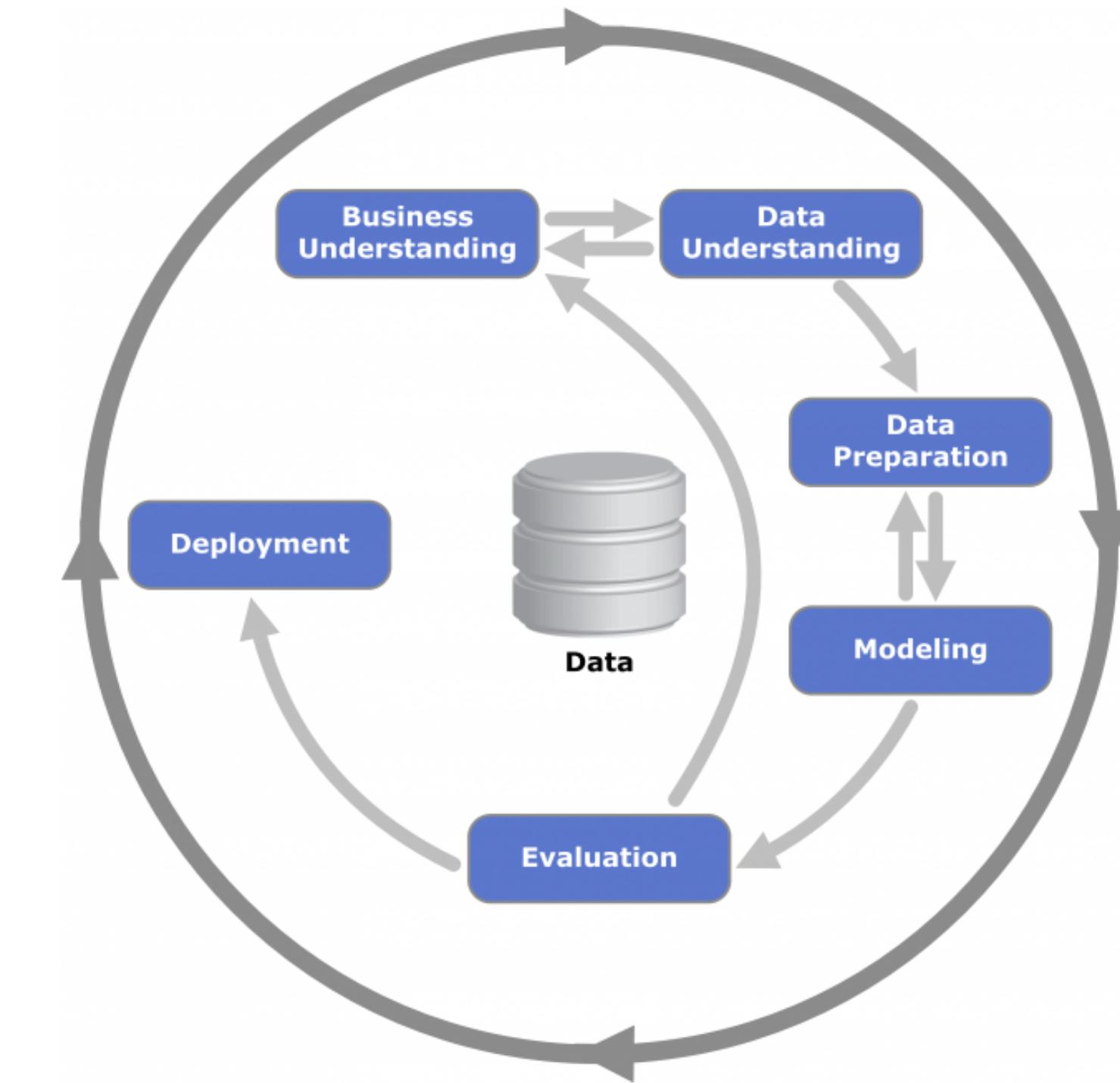
# Diferentes problemas de qualidade de dados



Survey of North American IT executives conducted by 451 Research in 2015

# Metodología CRISP-DM

## Cross-Industry Standard Process for Data Mining



CRISP-DM: Step-by-step data mining guide: <https://www.the-modeling-agency.com/crisp-dm.pdf>

02

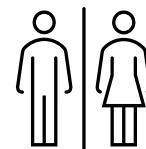
## Tipos de dados

Quais os diferentes tipos de dados que conhecem?



# Variáveis qualitativas

Representam uma classificação, podem ser nominais ou ordinais.



## Nominais

São variáveis que **não têm uma ordenação entre si**. Se tiverem apenas duas categorias são consideradas variáveis binárias.

Exemplo: género (homem/mulher), tipos de empréstimo (hipotecário, pessoal, automóvel, ...), etc.

---



## Ordinais

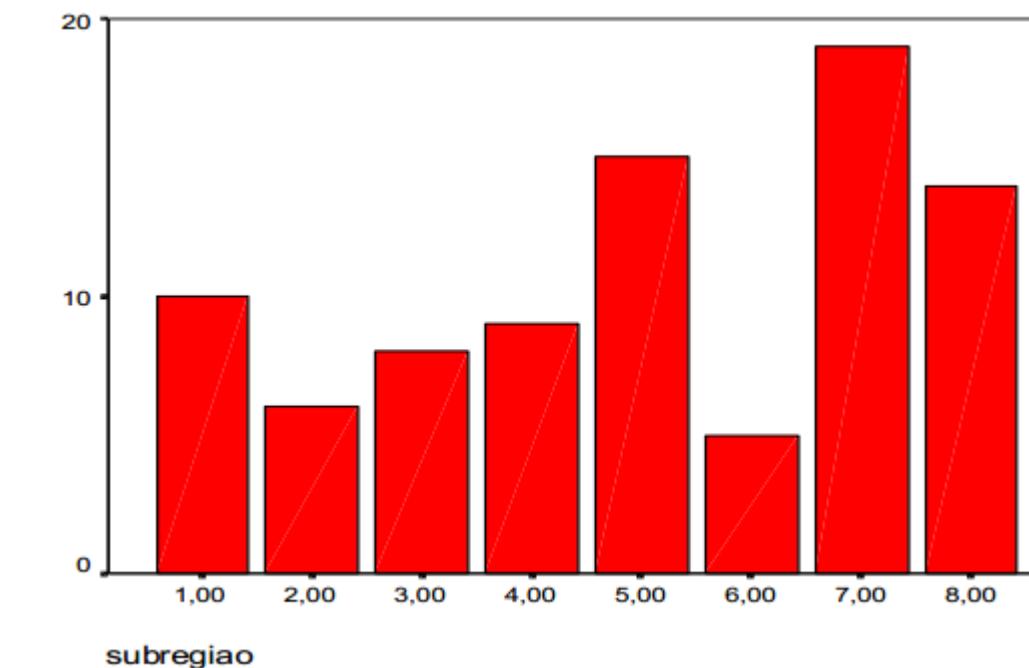
Se conseguirmos atribuir uma **ordem** trata-se de uma variável ordinal.

Exemplo: qualidade de um produto (má, média, boa), tamanho da camisola (S, M, L, XL), etc.

# Estudo de variáveis qualitativas

Algumas ferramentas que ajudam no estudo destas variáveis são: **distribuição de frequências** e a **moda**.

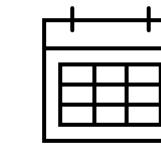
subregiao		
	Frequency	Percent
Valid		
1,00	10	11,6
2,00	6	7,0
3,00	8	9,3
4,00	9	10,5
5,00	15	17,4
6,00	5	5,8
7,00	19	22,1
8,00	14	16,3
Total	86	100,0



Frequências absolutas; Frequências relativas

# Variáveis quantitativas

Representam um valor numérico com uma ordem entre si, podem ser discretas ou contínuas.



## Discretas

São variáveis com valores finitos distribuídos numa escala de igual distância entre cada valor.

Exemplo: idade, número de cafés por dia, etc

---



## Contínuas

Podem assumir valores com casas decimais

Exemplo: rácios, montantes de pagamento, pesos, etc

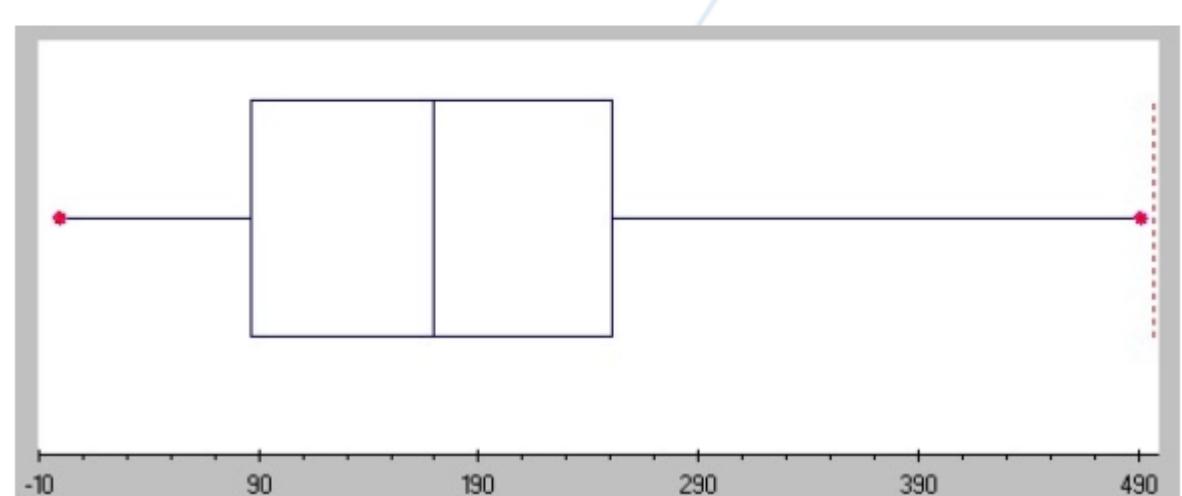
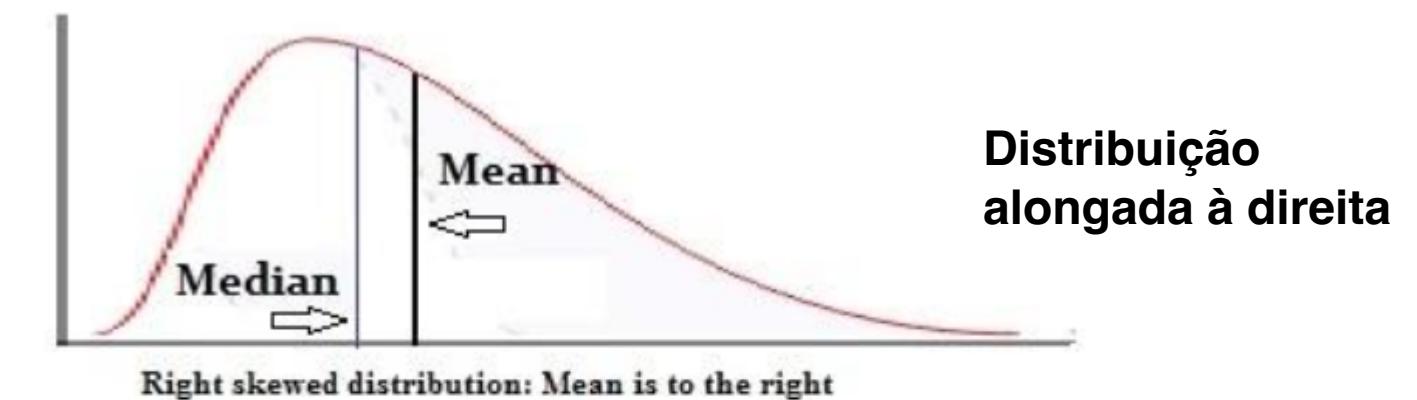
# Estudo de variáveis quantitativas

Algumas ferramentas que ajudam no estudo destas variáveis são: **medidas de localização (1)** e **medidas de dispersão (2)**.

- (1) localizam o centro da distribuição – média, moda e mediana;
- (2) avaliam a variação dos dados em relação ao centro da distribuição – variância e desvio padrão.

## Simetria

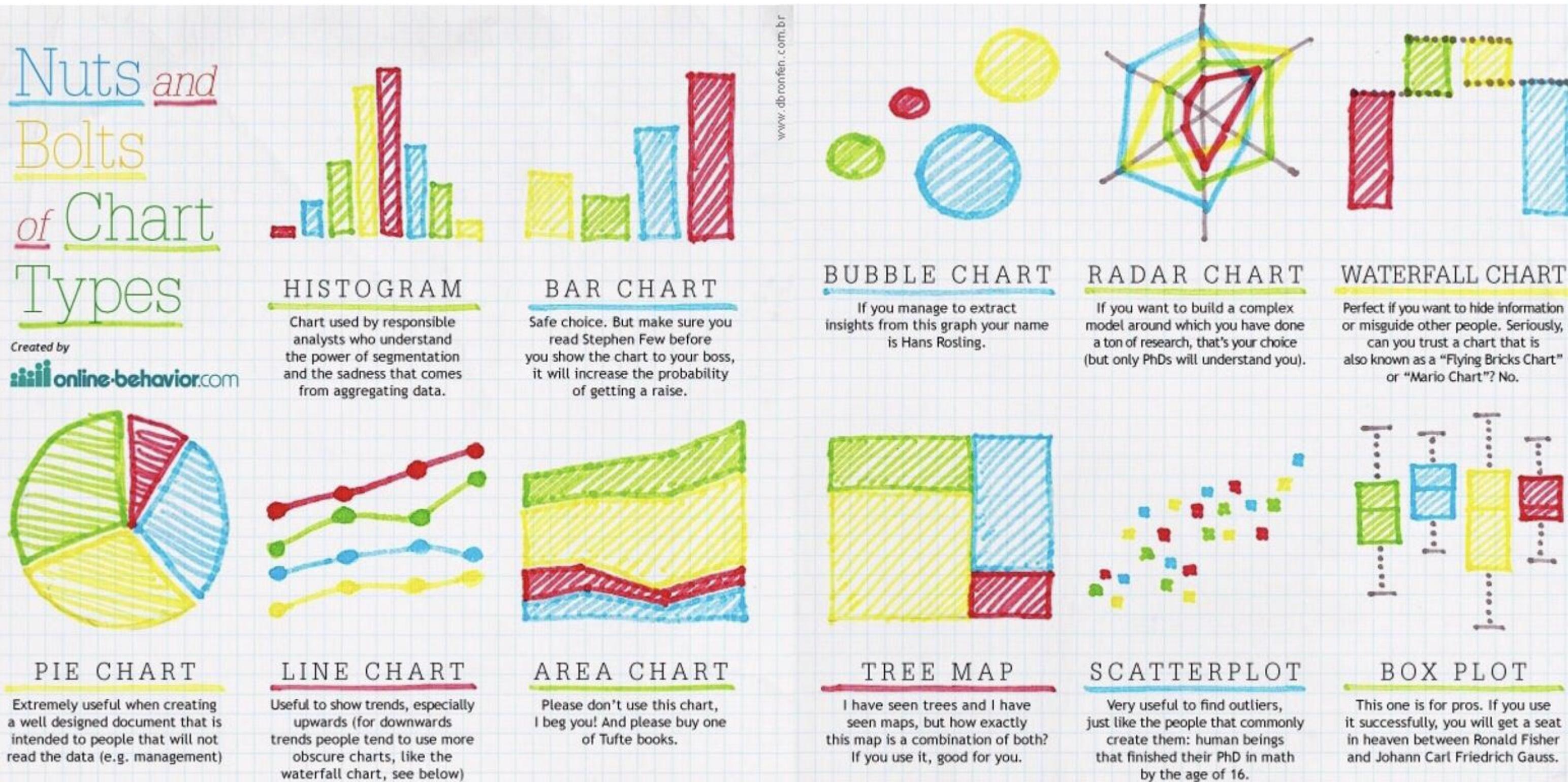
A maioria das estatísticas descritivas referentes a variáveis numéricas, como a média, a variância e o desvio padrão devem ser analisadas com cuidado, excepto se a distribuição não for muito assimétrica, nem contiver observações aberrantes.



# Alguns tipos de gráficos

**Nuts and Bolts of Chart Types**

Created by  online-behavior.com



**HISTOGRAM**  
Chart used by responsible analysts who understand the power of segmentation and the sadness that comes from aggregating data.

**BAR CHART**  
Safe choice. But make sure you read Stephen Few before you show the chart to your boss, it will increase the probability of getting a raise.

**BUBBLE CHART**  
If you manage to extract insights from this graph your name is Hans Rosling.

**RADAR CHART**  
If you want to build a complex model around which you have done a ton of research, that's your choice (but only PhDs will understand you).

**WATERFALL CHART**  
Perfect if you want to hide information or misguide other people. Seriously, can you trust a chart that is also known as a "Flying Bricks Chart" or "Mario Chart"? No.

**PIE CHART**  
Extremely useful when creating a well designed document that is intended to people that will not read the data (e.g. management)

**LINE CHART**  
Useful to show trends, especially upwards (for downwards trends people tend to use more obscure charts, like the waterfall chart, see below)

**AREA CHART**  
Please don't use this chart, I beg you! And please buy one of Tufte books.

**TREE MAP**  
I have seen trees and I have seen maps, but how exactly this map is a combination of both? If you use it, good for you.

**SCATTERPLOT**  
Very useful to find outliers, just like the people that commonly create them: human beings that finished their PhD in math by the age of 16.

**BOX PLOT**  
This one is for pros. If you use it successfully, you will get a seat in heaven between Ronald Fisher and Johann Carl Friedrich Gauss.

## 03 **Outliers**

O que é um outlier?

Os outliers são observações aberrantes que podem existir numa distribuição de frequências e classificam-se como severos ou moderados consoante o seu afastamento em relação às outras observações seja mais ou menos pronunciado.



# Outliers

Qual a importância de se detetar os outliers?



## Ponto 1

Melhor compreensão da base de dados e da distribuição das variáveis.

---



## Ponto 2

Identificação de erros potenciais.

---



## Ponto 3

Deteção de anomalias fraudulentas.

# COMO DETETAR?

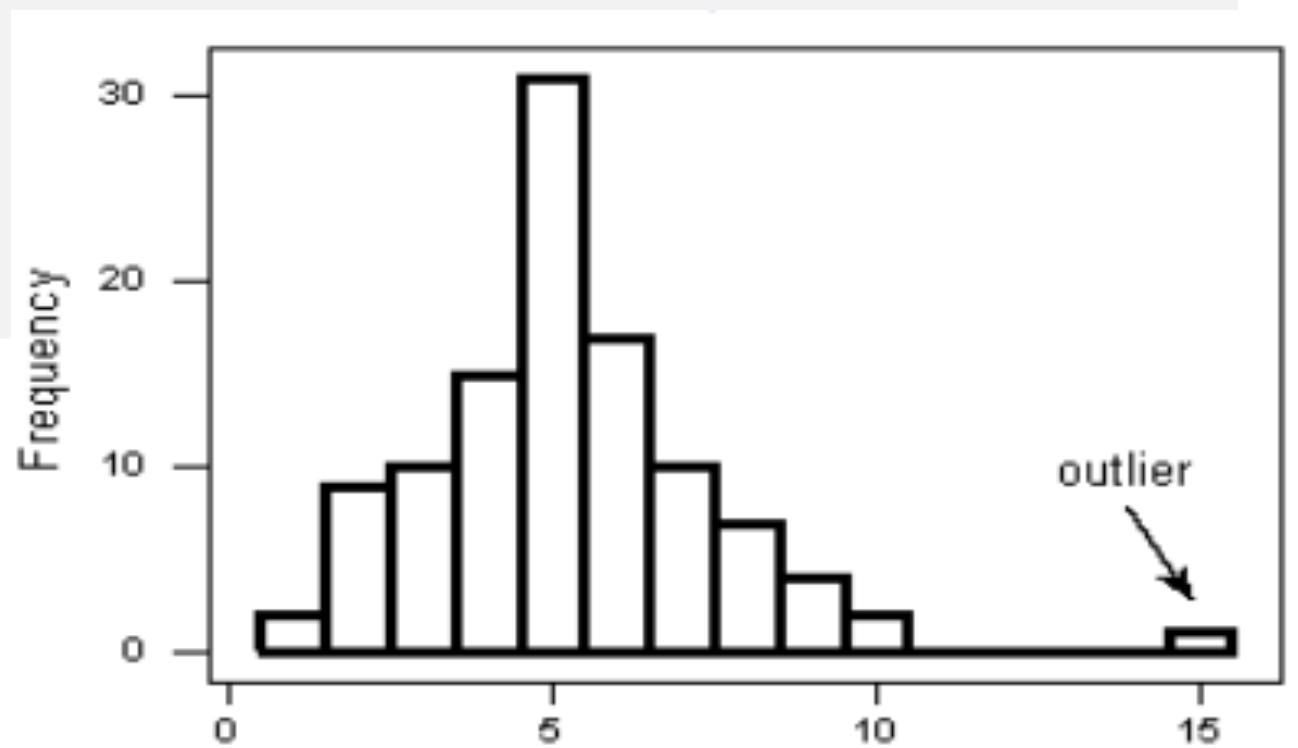
Abordagem baseada em distâncias

Calcular a média ( $\bar{x}$ ) e o desvio padrão ( $\sigma$ ).

X é um outlier se estiver fora dos limites.

$$(\bar{x} - k\sigma, \bar{x} + k\sigma)$$

**Nota:** Estamos a assumir uma distribuição normal (apropriado para variáveis simétricas).



# COMO DETETAR?

Abordagem baseada no boxplot

Uma observação é um **outlier moderado** se está fora destes limites:

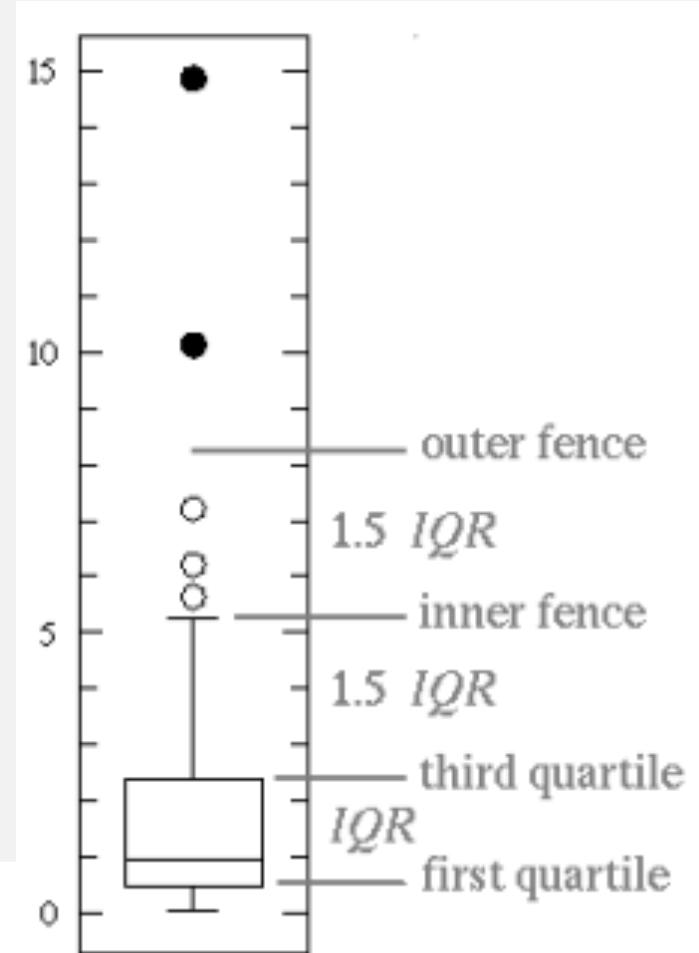
$$(Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR)$$

Uma observação é um **outlier severo** se está fora destes limites:

$$(Q1 - 3 \times IQR, Q3 + 3 \times IQR)$$

Onde  $IQR = Q3 - Q1$ .

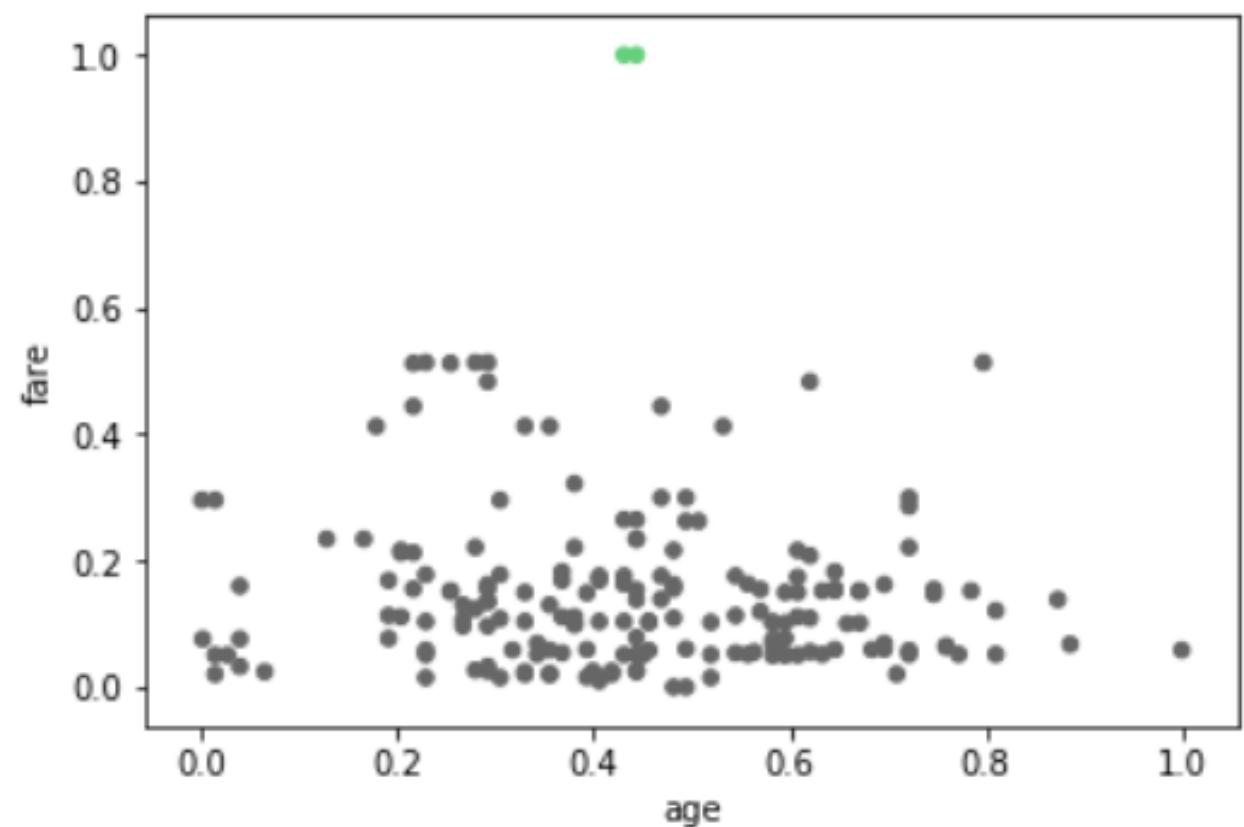
Apropriado para variáveis simétricas.



# COMO DETETAR?

Abordagem baseada em Clustering

Clusters com poucas observações são considerados outliers.



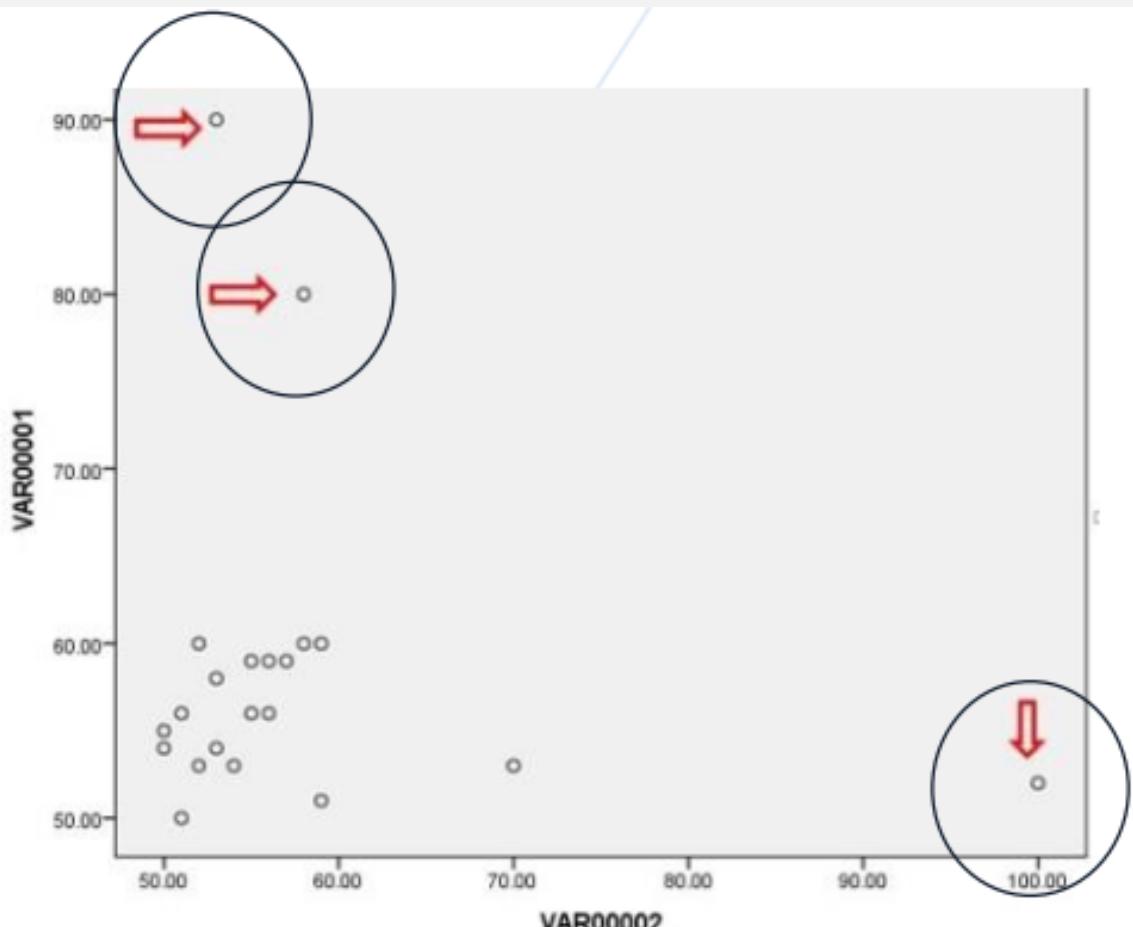
**Nota:** Este método usa conhecimento sobre Machine Learning não supervisionado  
(a verem em mais detalhe noutro módulo do curso)

# COMO DETETAR?

Abordagem baseada no “Vizinho mais próximo”

Método “Nearest neighbor”:

- Pontos onde existem menos de  $\rho$  vizinhos com uma distância  $D$ .
- Os outliers serão os top  $n$  pontos cuja distância para o  $k^{\text{th}}$  vizinho mais próximo é a maior.



# Outliers

Após identificar, o que fazer?



## **Não fazer nada**

Podem de facto não ser anomalias.



## **Remover as observações**

Considerar apenas as observações sem anomalias



## **Dar menos peso**

Atribuir um peso menor do que outras observações.



## **Substituir**

Podemos preencher com uma constante, uma métrica estatística (média, mediana, moda, etc), conhecimento de negócio ou inferir com recurso a Machine Learning (ML).

## 04

# Dados em falta

Numa base de dados real existem sempre dados em falta, muitas vezes com elevado impacto na construção do modelo.

**Os dados em falta podem ocorrer por diversos motivos:**

- Falhas no equipamento de registo
- Dados não considerados por parecerem irrelevantes no momento da recolha
- Ausência de informação em histórico
- Alteração da informação a recolher
- Dados eliminados por serem incoerentes com outras informações

# Dados em falta

O que fazer?

Idêntico aos outliers!



## Não fazer nada

Dependendo do objetivo pode ser relevante mantê-los.



## Remover as observações/variáveis

Considerar apenas as observações/variáveis com todos os valores preenchidos.



## Preencher os valores em falta

Podemos preencher com uma constante ou uma métrica estatística (média, mediana, moda, etc)



## Deduzir o valor mais provável

Com recurso a conhecimento de negócio ou inferir com recurso a Machine Learning (ML).

# 05

# Correlações

O que entendem por variáveis correlacionadas?

Medida de associação/semelhança (causal ou não) entre duas variáveis.



# Medidas de correlação mais conhecidas

**Entre variáveis  
Numéricas (ou ordinais)**

- Pearson;
- Spearman;
- Kendall

**Entre variáveis  
categóricas**

- Cramer's V

**Variável numérica vs  
categórica**

- Correlation ratio

# Pearson's r correlation

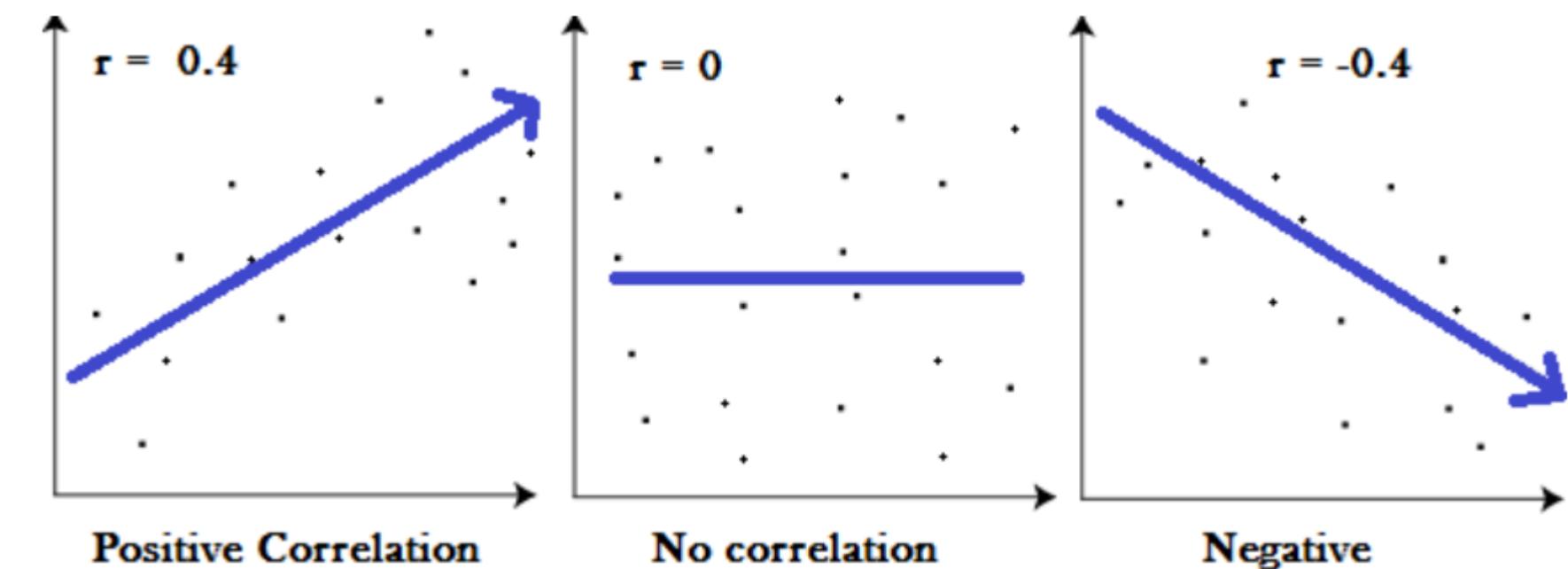
É uma medida de correlação linear entre duas variáveis numéricas:

$$r = \frac{COV(x, y)}{\sigma_x \times \sigma_y},$$

onde  $COV(x, y)$  é a covariância e  $\sigma_x$  é o desvio padrão.

O seu valor varia entre -1 e 1.

- 1 indica uma perfeita correlação linear negativa;
- 0 indica nenhuma correlação linear;
- 1 indica uma perfeita correlação linear positiva.



# Spearman's $\rho$ correlation

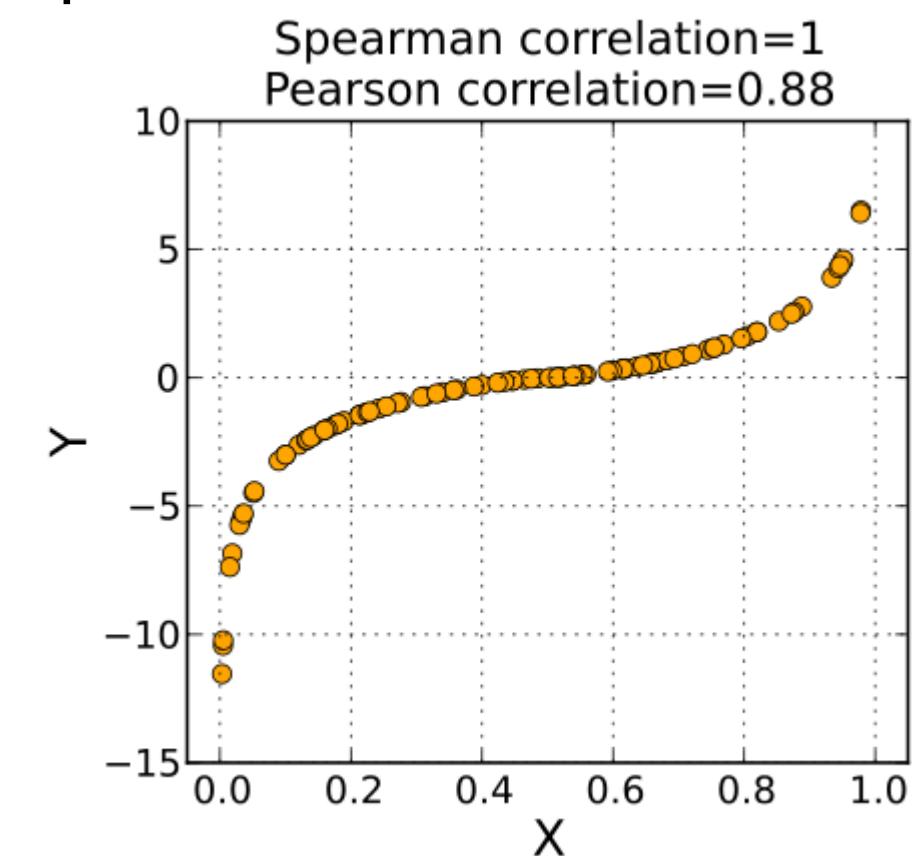
É uma medida de correlação monótona entre duas variáveis:

$$\rho = \frac{COV(R(x), R(y))}{\sigma_{R(x)} \times \sigma_{R(y)}},$$

onde  $COV(R(x), R(y))$  é a covariância dos ranks e  $\sigma_{R(x)}$  é o desvio padrão do rank.

O seu valor varia entre -1 e 1.

- 1 indica uma perfeita correlação monótona negativa;
- 0 indica nenhuma correlação monótona;
- 1 indica uma perfeita correlação monótona positiva.



# Kendal's $\tau$ correlation

É uma medida de correlação ordinal entre duas variáveis:

$$\tau = \frac{(n^o \text{ de pares concordantes}) - (n^o \text{ pares discordantes})}{\frac{n(n-1)}{2}}$$

O seu valor varia entre -1 e 1.

- 1 indica uma perfeita correlação negativa;
- 0 indica nenhuma correlação;
- 1 indica uma perfeita correlação positiva.

Idêntico ao Spearman!

# Cramér's V association

É uma medida de associação de variáveis aleatórias categóricas:

O seu valor varia entre 0 e 1.

- 0 indica independência;
- 1 indica uma perfeita associação.

*A sua forma original tem algumas limitações.*

*Para mais detalhes sobre uma forma otimizada ler: <http://stats.lse.ac.uk/bergsma/pdf/cramerV3.pdf>*

# Correlation Ratio

É uma medida de associação entre variáveis categóricas e numéricas:

$$\eta = \sqrt{\frac{\sum_x n_x (\bar{y}_x - \bar{y})^2}{\sum_{x,i} (y_{xi} - \bar{y})^2}}$$

Onde  $n_x$  é o número de observações na categoria x e:

$$\bar{y}_x = \frac{\sum_i y_{xi}}{n_x}, \bar{y} = \frac{\sum_i n_x \bar{y}_x}{\sum_x n_x}$$

O seu valor varia entre 0 e 1.

- 0 indica independência;
- 1 indica uma perfeita associação.

06

# Possíveis perguntas numa entrevista



# Exercício

**Enviar-me um notebook com outras abordagens de tratamento de valores em falta até dia 23/Dez;**

- 1. Eliminar nas duas bases as linhas com observações em falta em pelo menos uma dessas variáveis:  
3,5 pontos**
- 2. Eliminar 1 variável com dados em falta nas duas bases de dados: 3,5 pontos**
- 3. Preencher corretamente nas duas bases de dados os valores em falta com a média (para pelo menos 1 variável numérica): 5 pontos; e valor "missing" (para pelo menos 1 variável categoria) : 5 pontos**
- 4. Escolhe uma variável com dados em falta e preenche os valores com recurso a um algoritmo de ML nas duas bases de dados: 3 valores.**

**Nota: É importante que mantenham a abordagem das duas bases de dados (*df1* e *df2*);**

**Muito obrigado pela  
sua atenção.**

**Fábio Ferreira**

Endereço de e-mail:

[fabio7jnferreira@gmail.com](mailto:fabio7jnferreira@gmail.com)

