

Aprendizagem Supervisionada Aula 2

Análise Exploratória de
Dados (EDA)

Douglas Macedo Sgrott
Orientador: Rafael Parpinelli
02/06/2021

 **O que foi visto na aula passada...**

O que foi visto na aula passada...

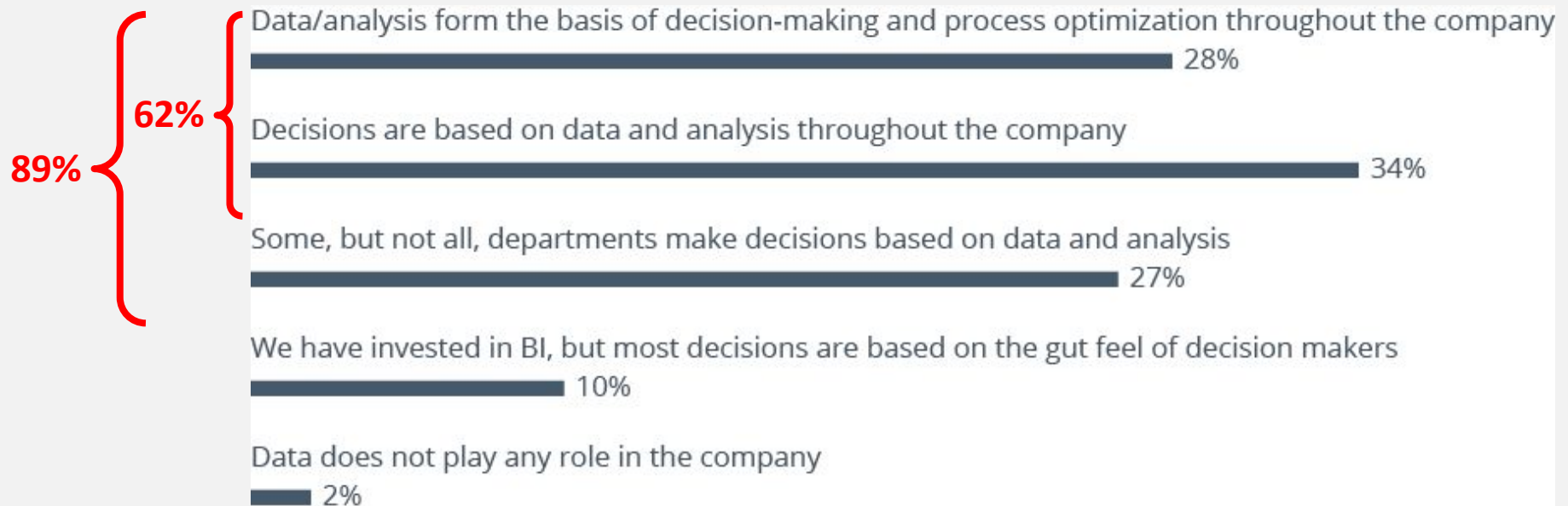
- **Definições sobre Aprendizagem de Máquina**
- **Aprendizagem Supervisionada: Regressão x Classificação**
- **Regressão:**
 - **Regressão Linear + RIDGE**
 - **Regressão Polinomial**
 - **Árvores de Decisão + Floresta Aleatória**
- **Classificação**
 - **Regressão Logística (apesar do nome, ainda é classificação)**
- **A escolha de um modelo adequado é IMPORTANTÍSSIMO**
 - **Isso depende das prioridades de modelagem ou de negócio**
 - **Depende dos DADOS.**

Objetivo dessa aula

- **Ter uma noção de como fazer uma Análise Exploratória de Dados (EDA) e pré-processar dados**

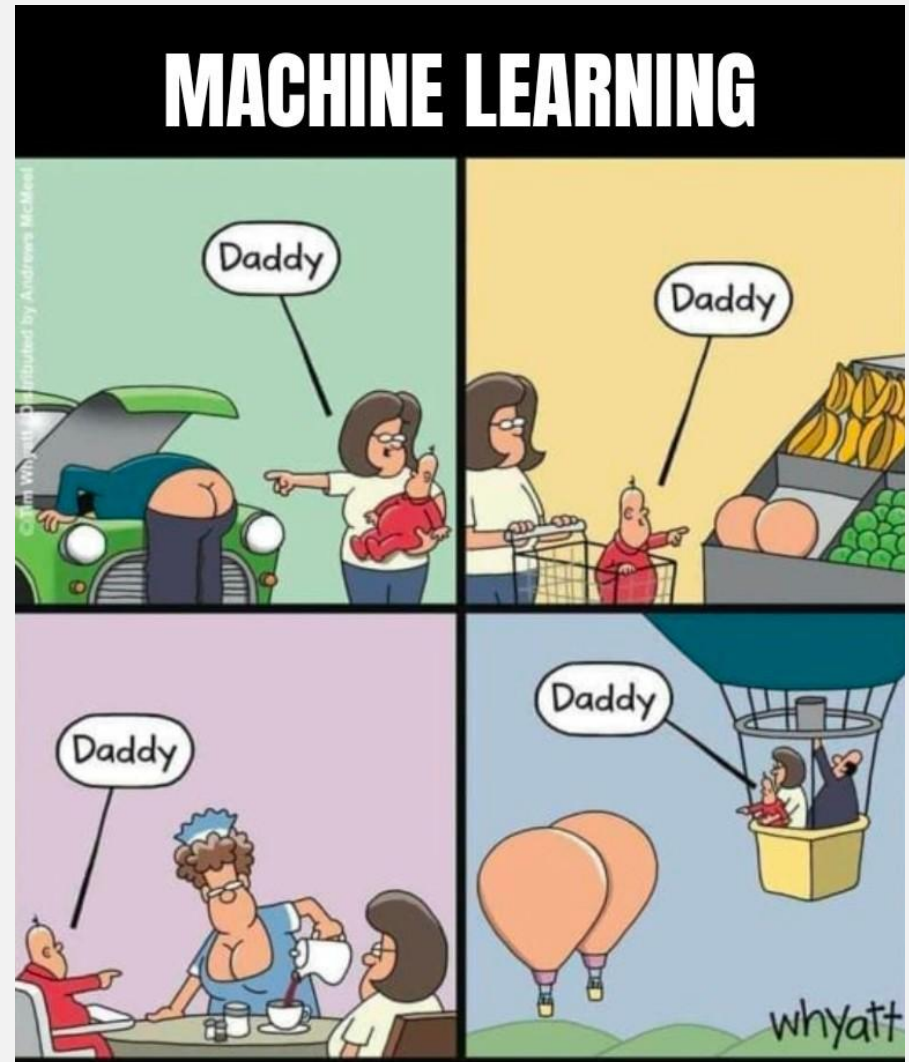
A importância dos dados

- Resultados de um estudo de 2014 sobre a utilização de dados em tomada de decisão em empresas da Alemanha, Áustria e Suíça



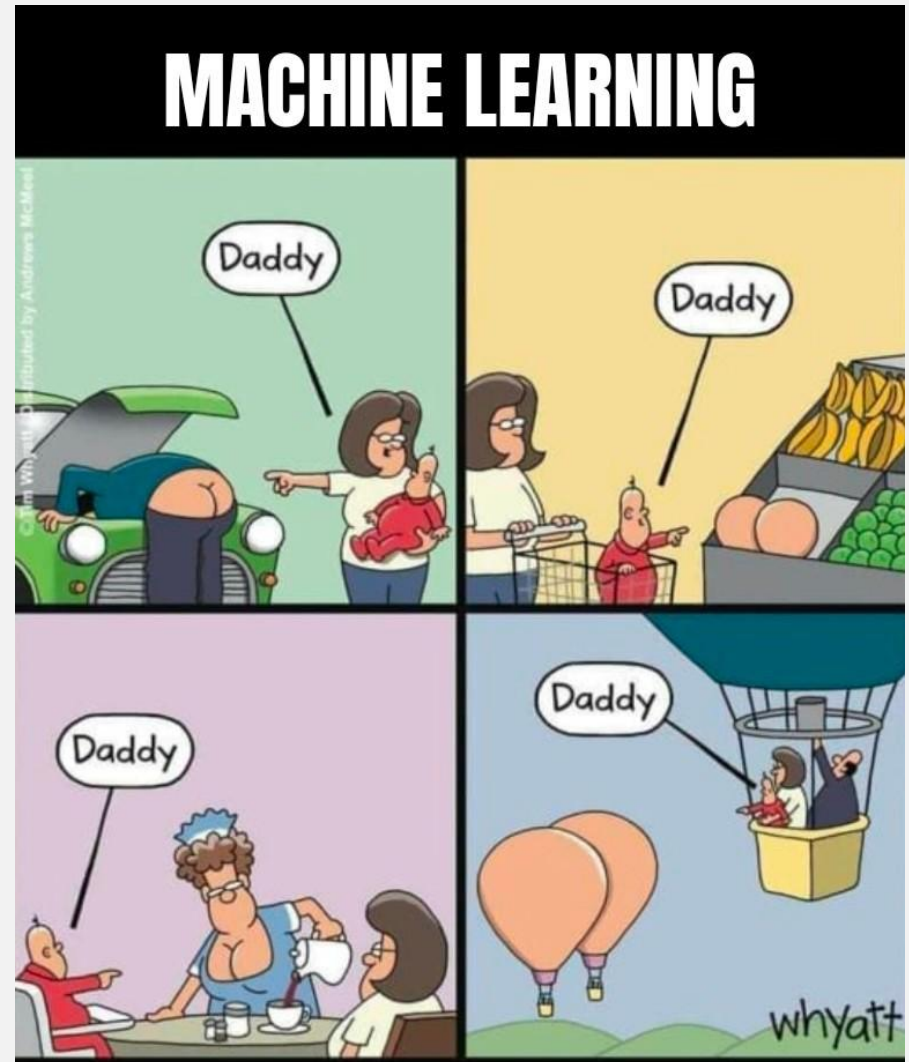
A importância de dados em aprendizagem de máquina

- Um modelo de Machine Learning é tão bom quanto os dados usados na sua modelagem.



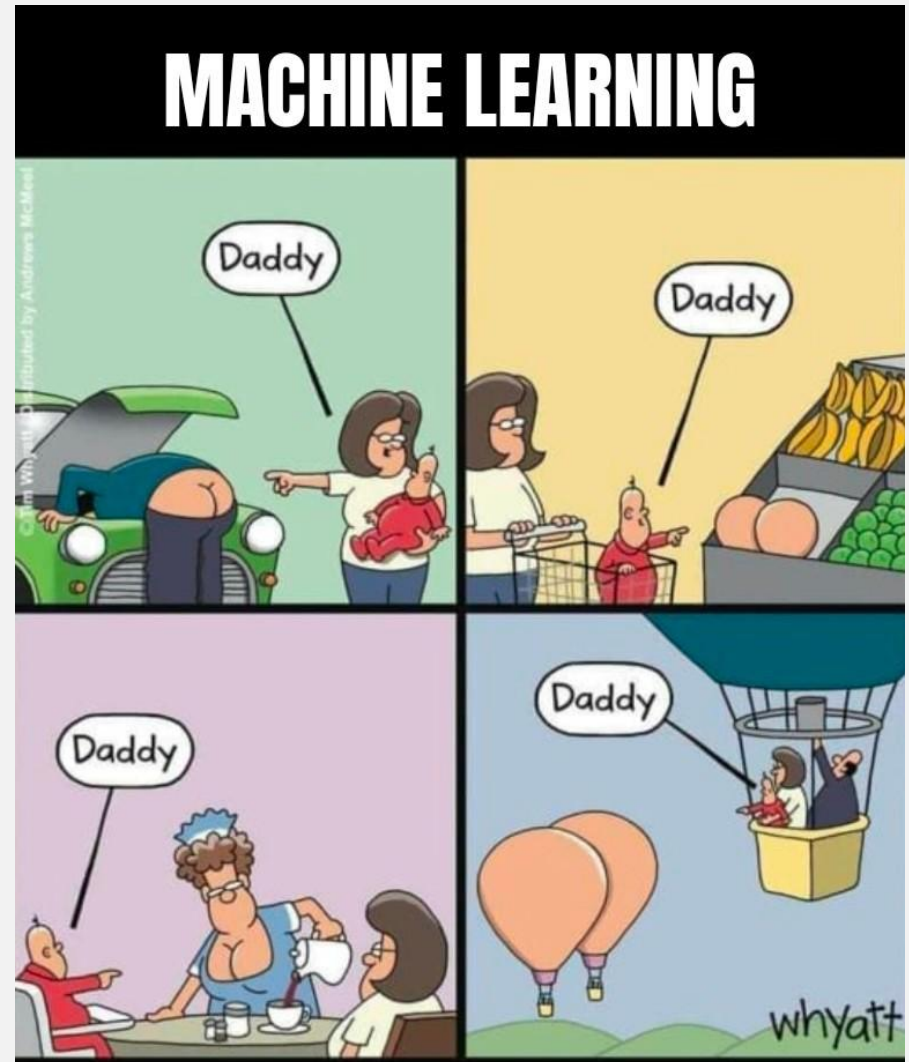
A importância de dados em aprendizagem de máquina

- Um modelo de Machine Learning é tão bom quanto os dados usados na sua modelagem.
- Dados não balanceados podem gerar modelos não balanceados



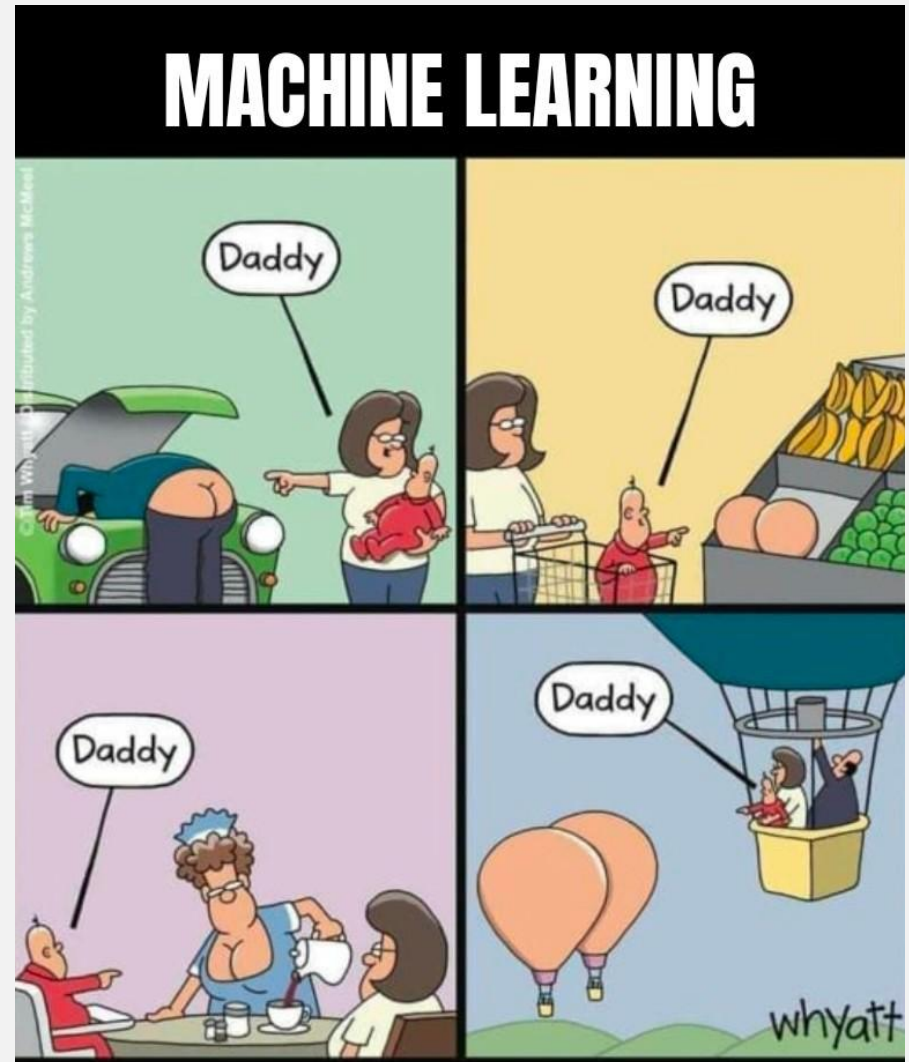
A importância de dados em aprendizagem de máquina

- Um modelo de Machine Learning é tão bom quanto os dados usados na sua modelagem.
- Dados não balanceados podem gerar modelos não balanceados
- Dados com tendência podem gerar modelos tendenciosos



A importância de dados em aprendizagem de máquina

- Um modelo de Machine Learning é tão bom quanto os dados usados na sua modelagem.
- Dados não balanceados podem gerar modelos não balanceados
- Dados com tendência podem gerar modelos tendenciosos
- Por que isso é perigoso?



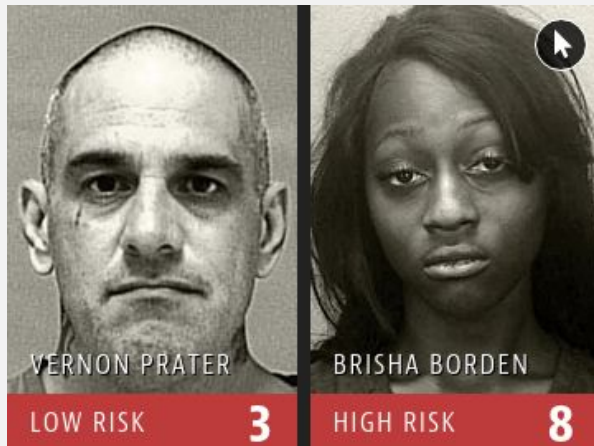
A importância de dados em aprendizagem de máquina

- Às vezes pode parecer cômico...



A importância de dados em aprendizagem de máquina

- Um modelo de ML em 2016 que calculava o “risco” de um “criminoso” pode inferir que pessoas negras têm maior risco que pessoas brancas.



A importância de dados em aprendizagem de máquina

- Modelos de ML podem inferir presença/ausência de doenças com base em fatores insignificantes.



A importância de dados em aprendizagem de máquina

Analysis | [Open Access](#) | Published: 15 March 2021

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts , Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Nature Machine Intelligence **3**, 199–217 (2021) | [Cite this article](#)

30k Accesses | **5** Citations | **874** Altmetric | [Metrics](#)

A importância de dados em aprendizagem de máquina

Abstract

Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, **62 studies were included in this systematic review**. Our review finds that **none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases**. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

Mas afinal... o que são dados?

- Possíveis definições:

Mas afinal... o que são dados?

- **Possíveis definições:**
 - **Dado é o registro do atributo de um ente, objeto ou fenômeno.**
 - **Unidades de informação, coletados através de observação.**

Tipo de dados

- **Dados numéricos vs Dados categóricos**
 - O que são, e qual a diferença?

Tipo de dados

- Dados numéricos ou quantitativos
- Normalmente são frutos de alguma medição

Tipo de dados

- **Dados numéricos ou quantitativos**
- **Normalmente são frutos de alguma medição**
 - **Contínuos**
 - **Admitem qualquer valor numérico**
 - [..., -0.2, -0.1, 0.0, 0.1, 0.2, 0.3, ... , 1000000.0, ...]
 - **Discretos**
 - **Admitem apenas valores inteiros**
 - [..., -2, -1, 0, 1, 2, 3, ...]

Tipo de dados

- **Dados numéricos ou quantitativos**
- **Normalmente são frutos de alguma medição**
 - **Contínuos**
 - **Admitem qualquer valor numérico**
 - [..., -0.2, -0.1, 0.0, 0.1, 0.2, 0.3, ... , 1000000.0, ...]
 - **Discretos**
 - **Admitem apenas valores inteiros**
 - [..., -2, -1, 0, 1, 2, 3, ...]
- **Dados categóricos ou qualitativos**
- **Normalmente servem como uma representação ou classificação**

Tipo de dados

- **Dados numéricos ou quantitativos**
- **Normalmente são frutos de alguma medição**
 - **Contínuos**
 - **Admitem qualquer valor numérico**
 - [..., -0.2, -0.1, 0.0, 0.1, 0.2, 0.3, ... , 1000000.0, ...]
 - **Discretos**
 - **Admitem apenas valores inteiros**
 - [..., -2, -1, 0, 1, 2, 3, ...]
- **Dados categóricos ou qualitativos**
- **Normalmente servem como uma representação ou classificação**
 - **Nominais**
 - **Não existe ordenação entre as categorias**
 - [Masculino, Feminino], [Sim, Não], [Alto, Médio, Baixo]

Tipo de dados

- **Dados numéricos ou quantitativos**
- **Normalmente são frutos de alguma medição**
 - **Contínuos**
 - **Admitem qualquer valor numérico**
 - [..., -0.2, -0.1, 0.0, 0.1, 0.2, 0.3, ... , 1000000.0, ...]
 - **Discretos**
 - **Admitem apenas valores inteiros**
 - [..., -2, -1, 0, 1, 2, 3, ...]
- **Dados categóricos ou qualitativos**
- **Normalmente servem como uma representação ou classificação**
 - **Nominais**
 - **Não existe ordenação entre as categorias**
 - [Masculino, Feminino], [Sim, Não], [Alto, Médio, Baixo]
 - **Ordinais**
 - **Existe ordenação entre as categorias**
 - Mês/ano/dia de observação, Nível de escolaridade, Estágio de doenças, etc.

Tipo de dados

- Existem exceções!

Tipo de dados

- **Existem exceções!**
- **Exemplos:**
 - **Número de telefone, CEP, CPF ...**
 - **Dados categóricos podem ser representados por números:**
 - **Ex: Verdadeiro = 1, Falso = 0**
 - **Ex: Ruim = 0, Indiferente = 1, Bom = 2**

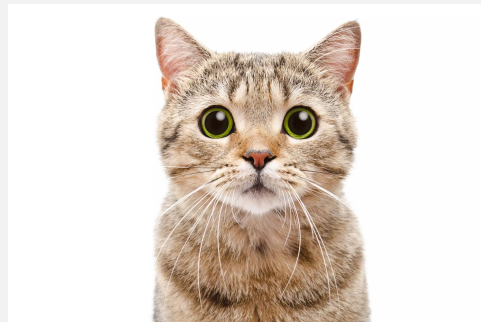
Tipo de dados

- Existem exceções!
- Exemplos:
 - Número de telefone, CEP, CPF ...
 - Dados categóricos podem ser representados por números:
 - Ex: Verdadeiro = 1, Falso = 0
 - Ex: Ruim = 0, Indiferente = 1, Bom = 2
- Existem outros tipos de dados:
 - Áudio



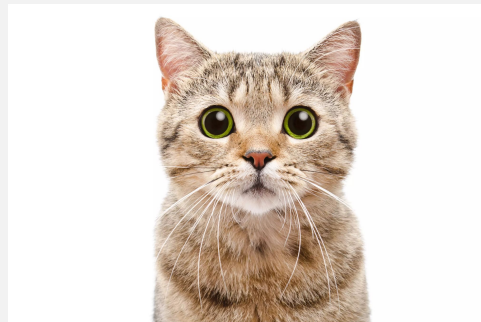
Tipo de dados

- Existem exceções!
- Exemplos:
 - Número de telefone, CEP, CPF ...
 - Dados categóricos podem ser representados por números:
 - Ex: Verdadeiro = 1, Falso = 0
 - Ex: Ruim = 0, Indiferente = 1, Bom = 2
- Existem outros tipos de dados:
 - Áudio
 - Imagens



Tipo de dados

- Existem exceções!
- Exemplos:
 - Número de telefone, CEP, CPF ...
 - Dados categóricos podem ser representados por números:
 - Ex: Verdadeiro = 1, Falso = 0
 - Ex: Ruim = 0, Indiferente = 1, Bom = 2
- Existem outros tipos de dados:
 - Áudio
 - Imagens
 - Texto



Análise Univariada e Bivariada

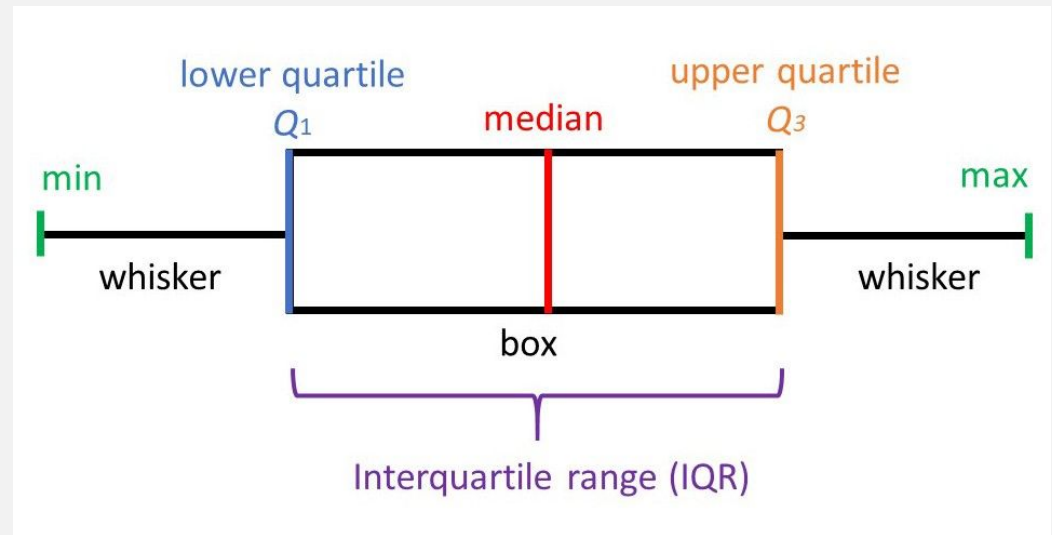
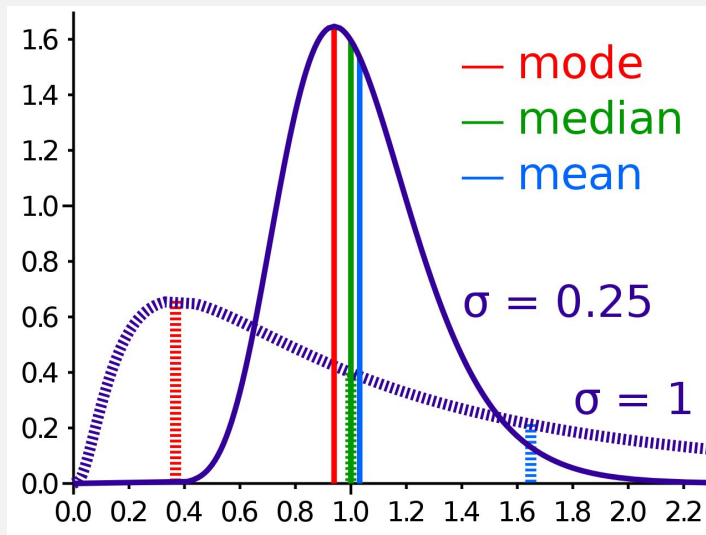
- **Análise Univariada**
 - **Análise mais simples de se fazer**
 - **Objetivos:**
 - **Entender melhor a distribuição dos dados**
 - **Facilitar a interpretação dos dados**

Análise Univariada e Bivariada

- **Análise Univariada**
 - **Análise mais simples de se fazer**
 - **Objetivos:**
 - **Entender melhor a distribuição dos dados**
 - **Facilitar a interpretação dos dados**
- **Análise Bivariada:**
 - **Objetivos:**
 - **Encontrar a RELAÇÃO ou causa entre duas (ou mais) variáveis**

Análise Univariada

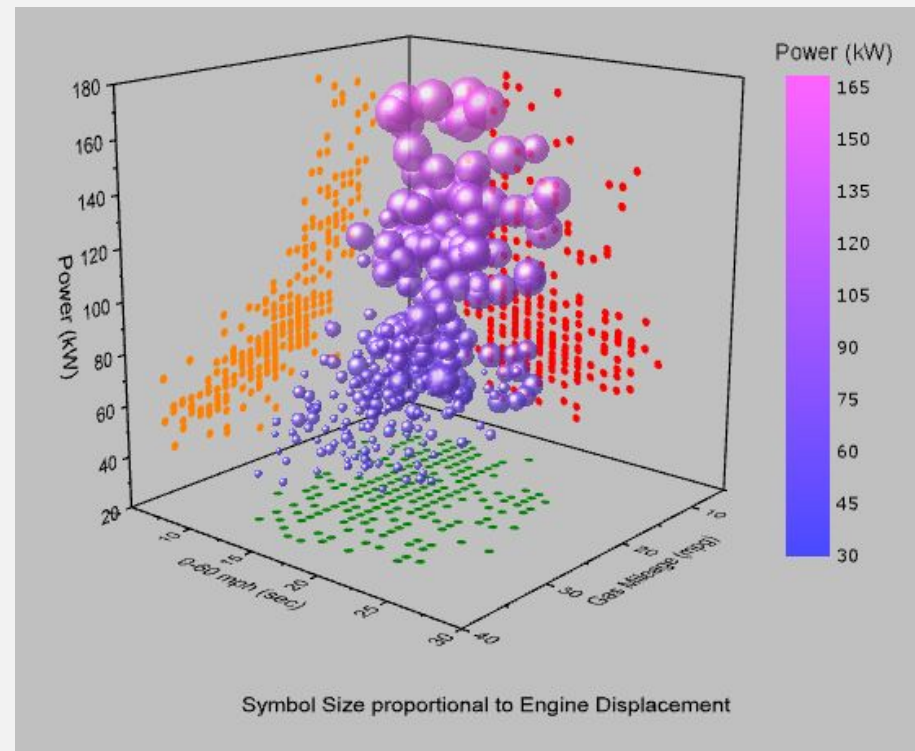
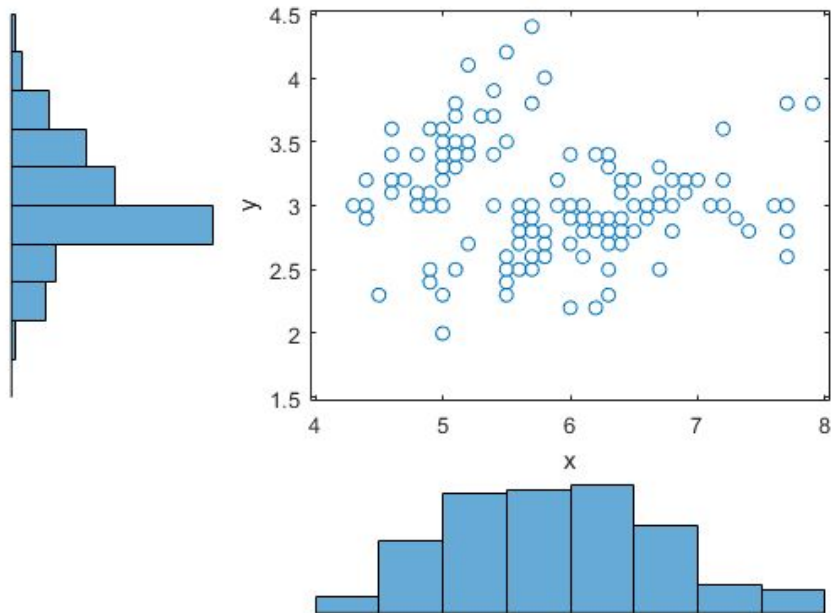
- Para dados quantitativos
 - Média, Mediana, Moda, Mínimo, Máximo, Intervalo, Percentil, Amplitude Interquartil, Assimetria, Histograma, Box Plot / Diagrama de Caixa



- Para dados qualitativos
 - Moda, Frequência, Histograma

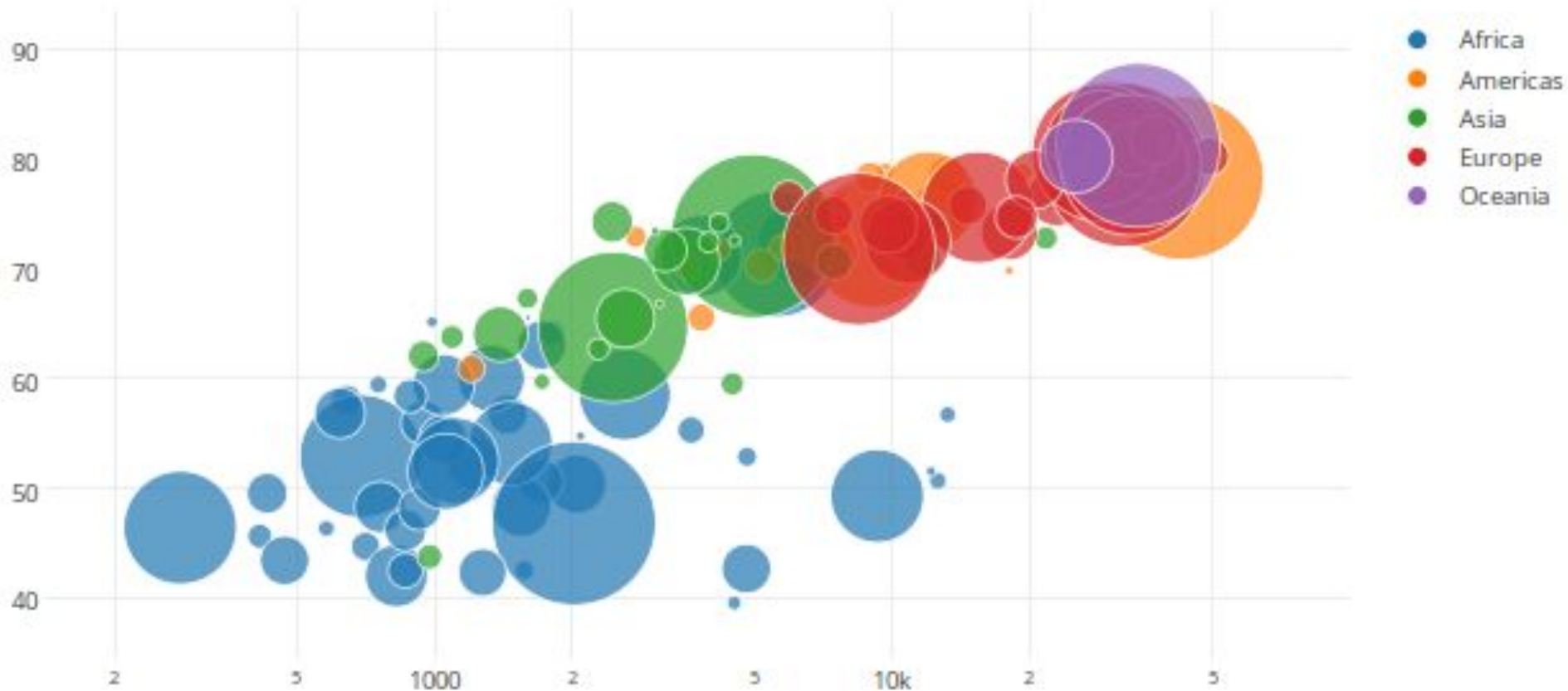
Análise Bivariada

- Gráfico de Dispersão (Scatter plot)



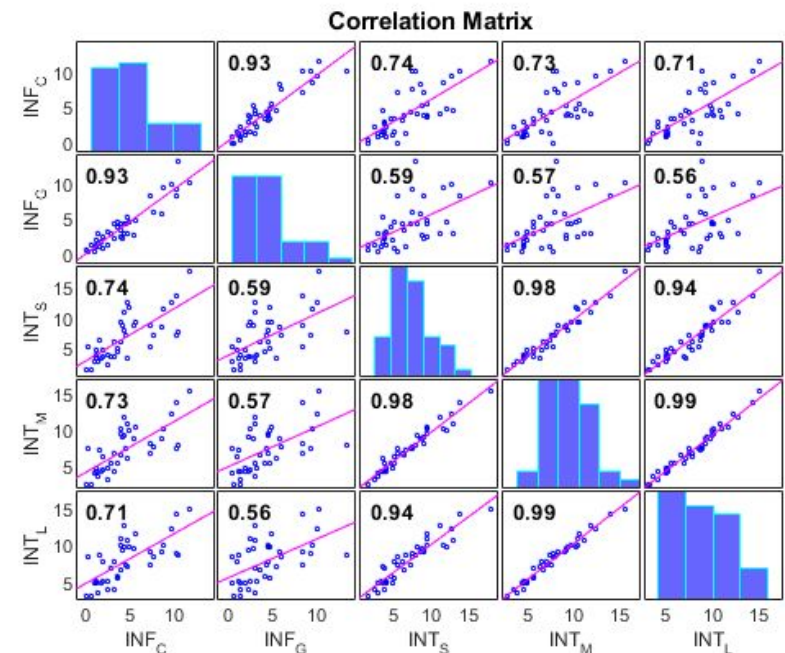
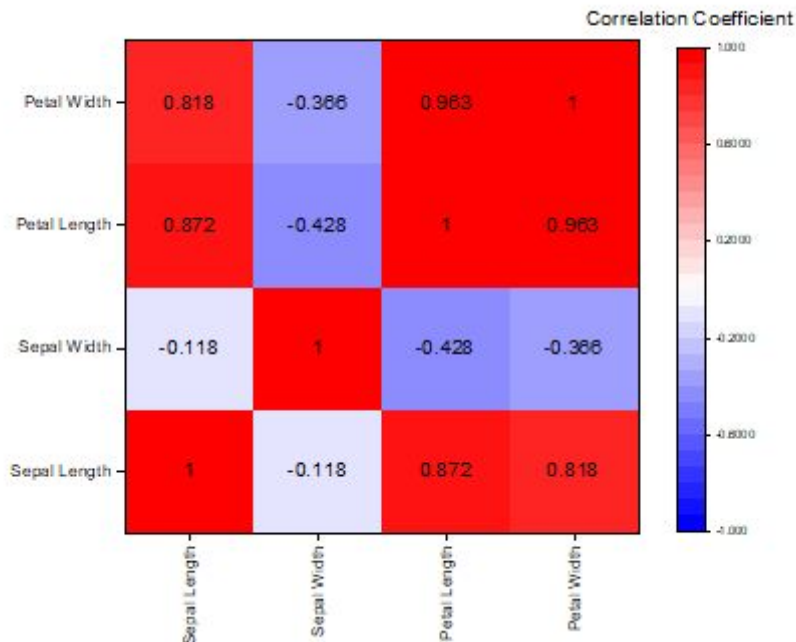
Análise Bivariada

- **Gráfico de Bolhas (Bubble plot)**
 - É uma variação do Gráfico de Dispersão onde cores, símbolos e tamanho também representam variáveis



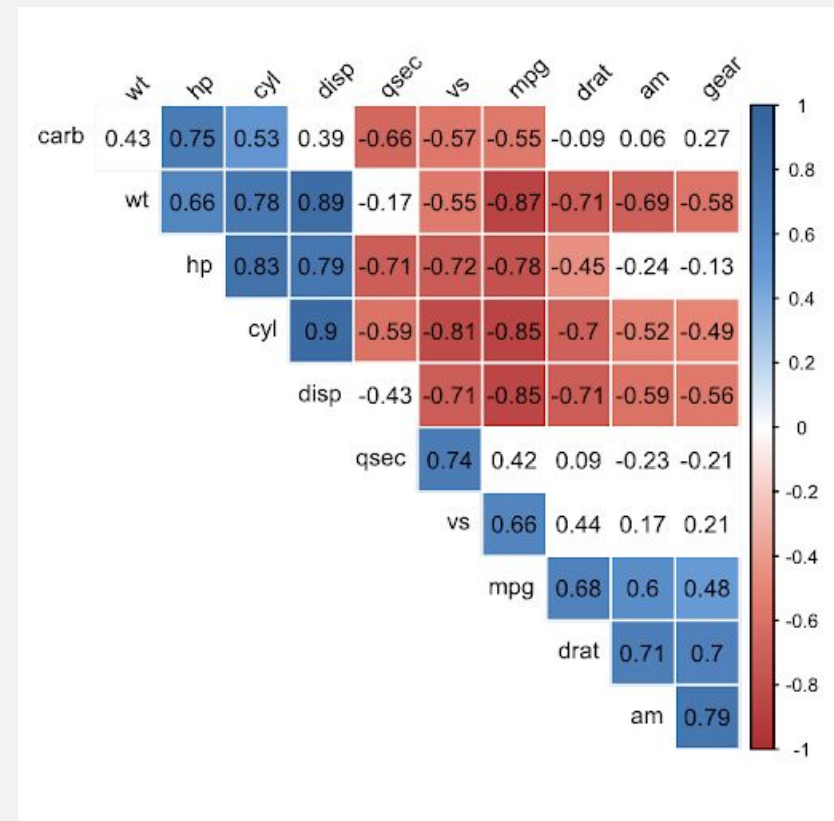
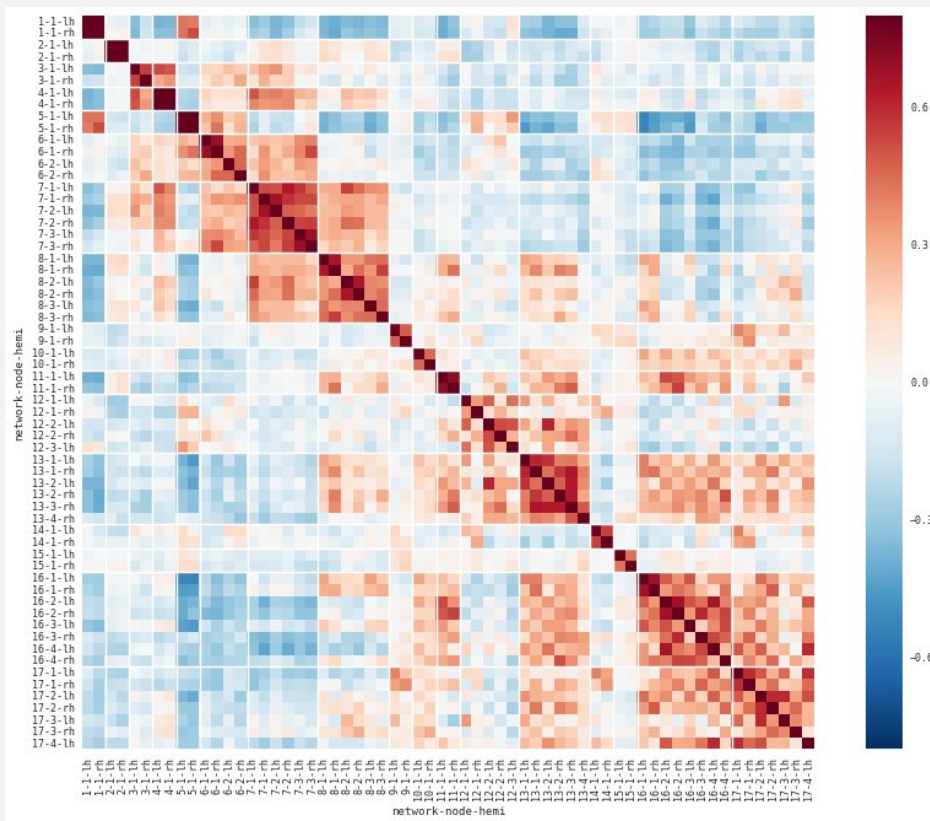
Análise Bivariada

- Gráfico de correlação - Heatmap



Análise Bivariada

- Gráfico de correlação - Heatmap



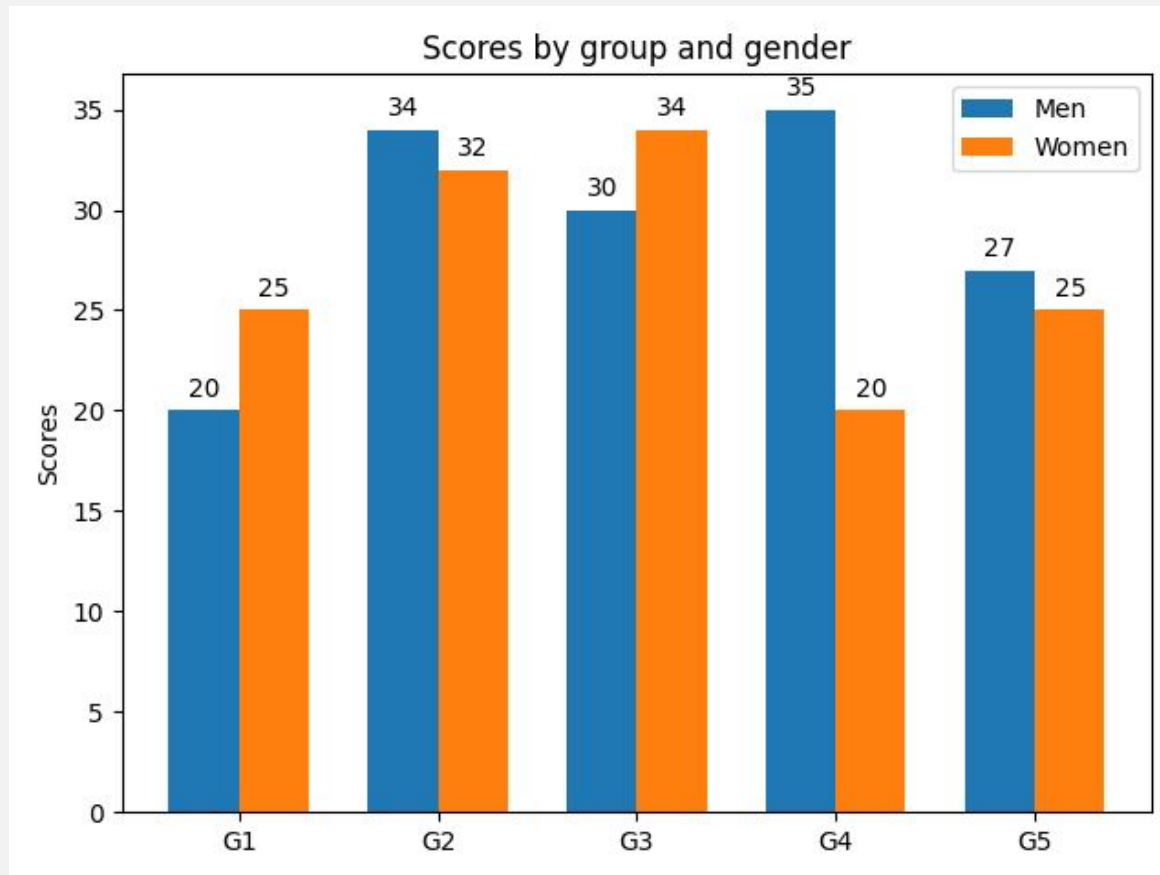
Análise Bivariada

- Tabela Two-way
 - Mostra a frequência de duas variáveis categóricas em linha e coluna.

	Baseball	Basketball	Football	Total
Male	13	15	20	48
Female	23	16	13	52
Total	36	31	33	100

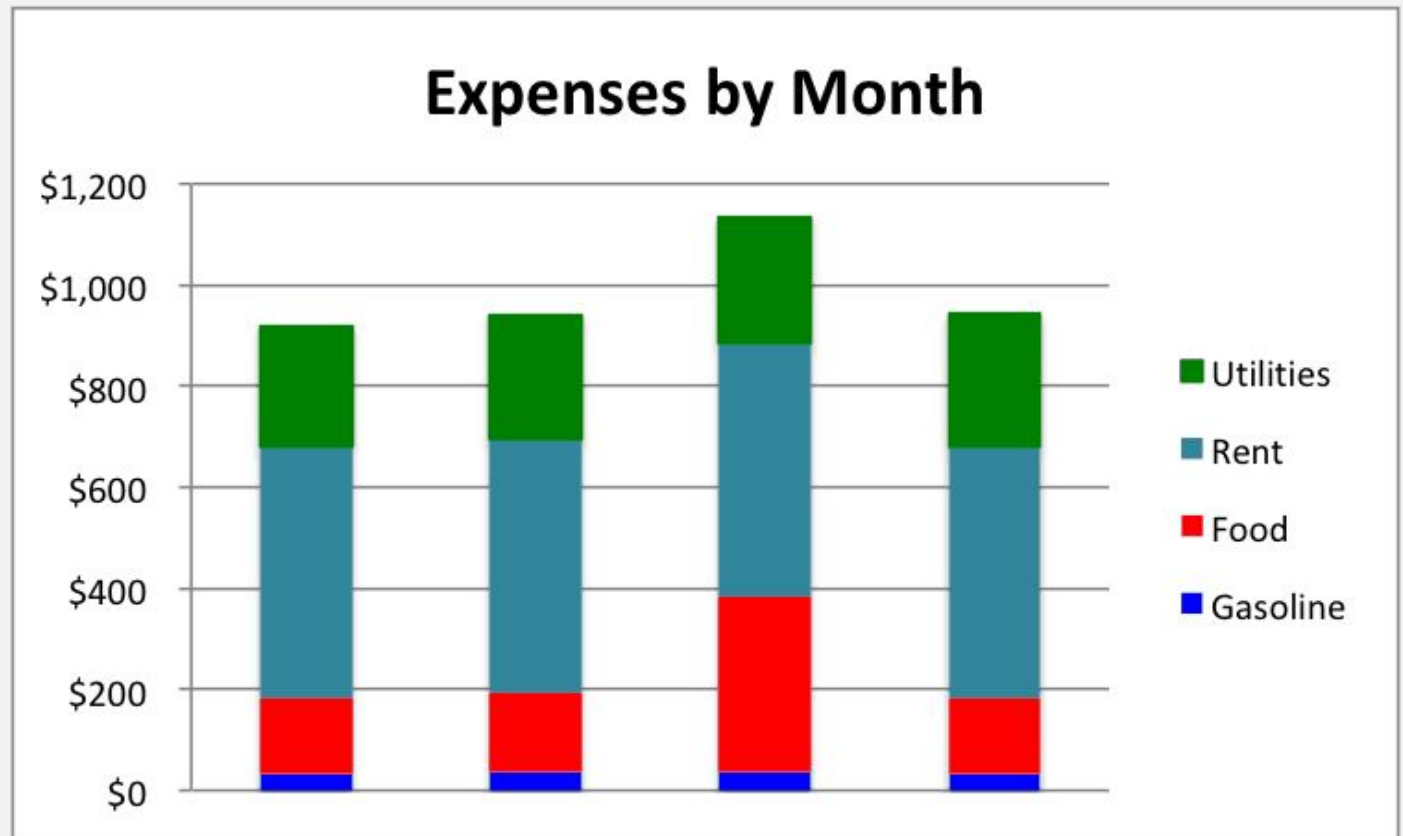
Análise Bivariada

- Gráfico de barras e barras empilhadas



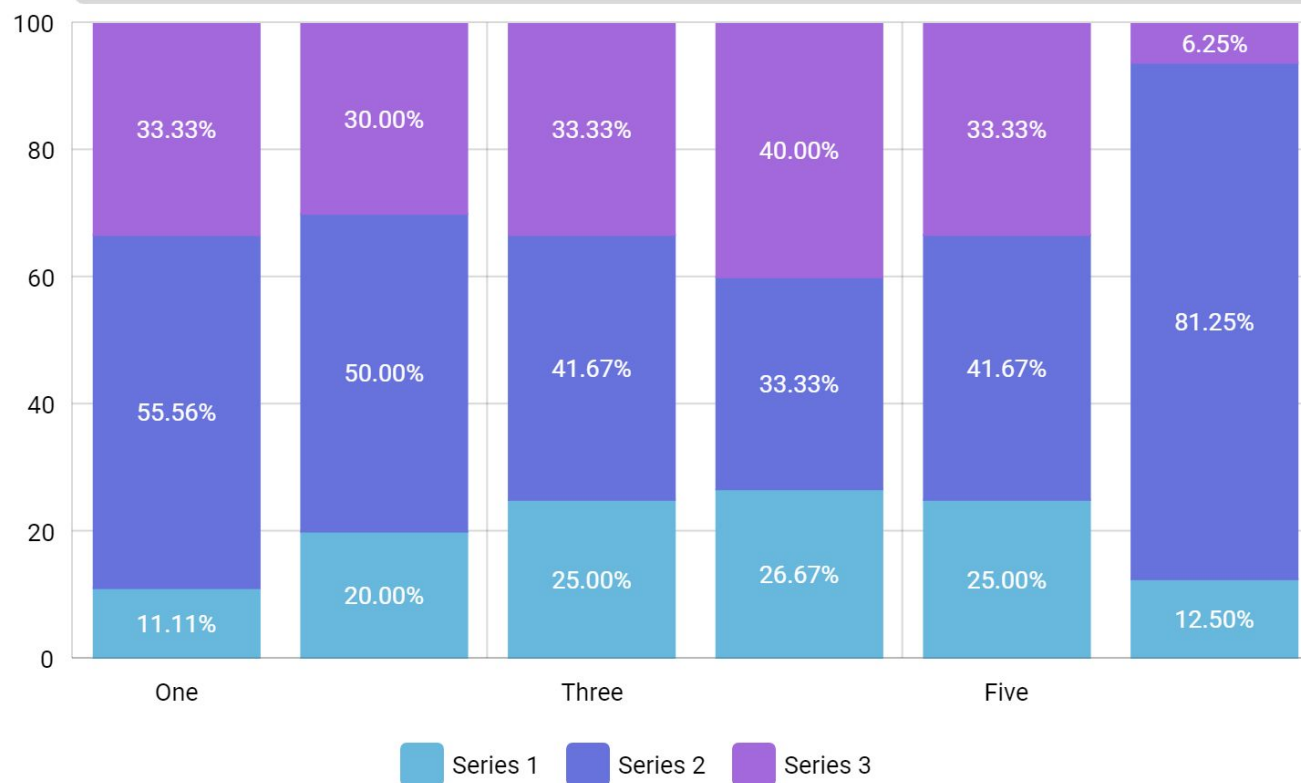
Análise Bivariada

- Gráfico de barras e barras empilhadas
 - Atenção com o contexto!



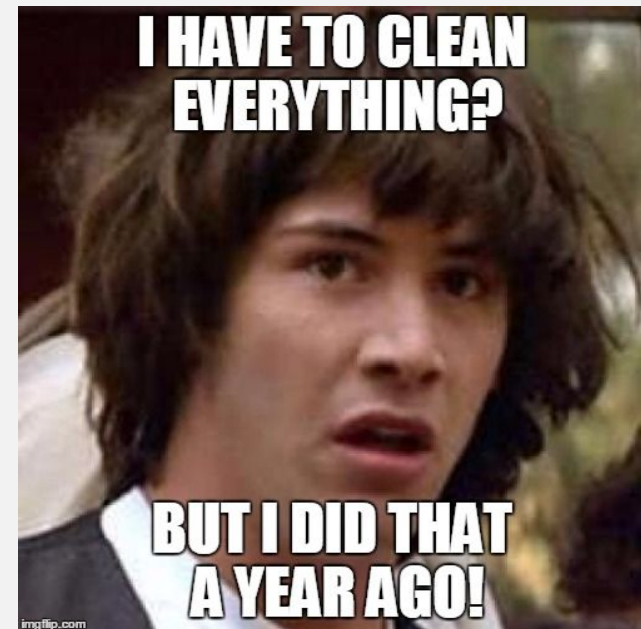
Análise Bivariada

- Gráfico de barras e barras empilhadas
 - Atenção com o contexto!



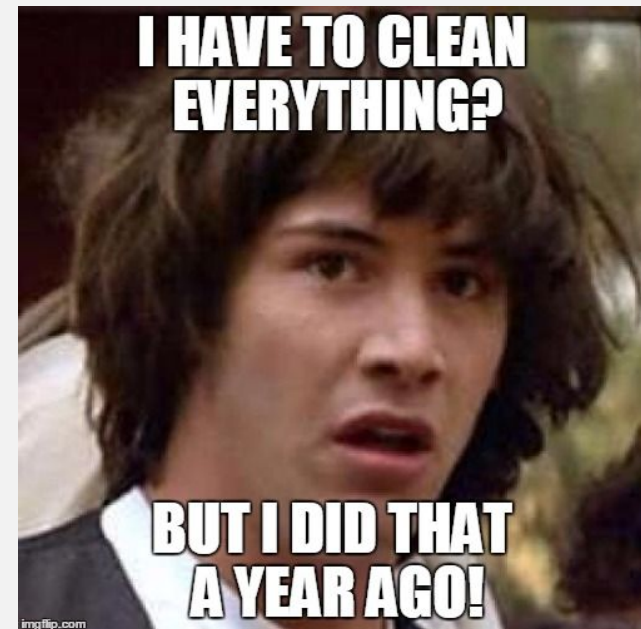
■ Limpeza de dados

- O que significa “limpar dados” ?



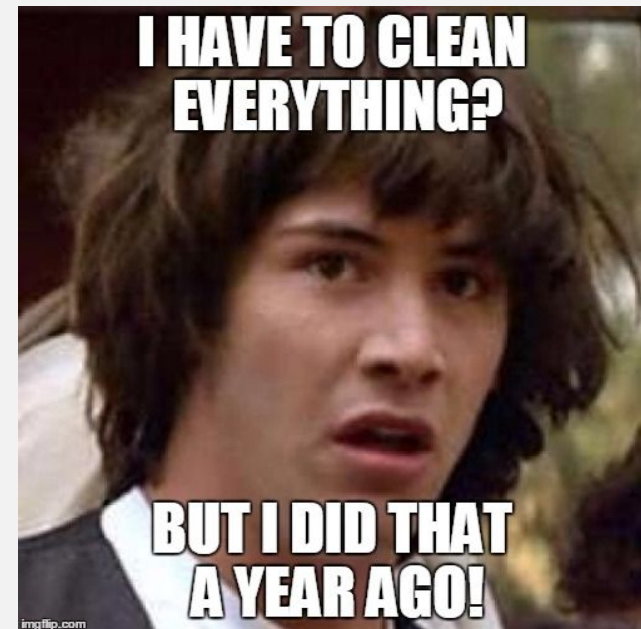
■ Limpeza de dados

- O que significa “limpar dados” ?
- Limpeza de dados (ou Data cleaning) consiste em detectar, remover OU corrigir dados corruptos, incorretos, ausentes ou irrelevantes.



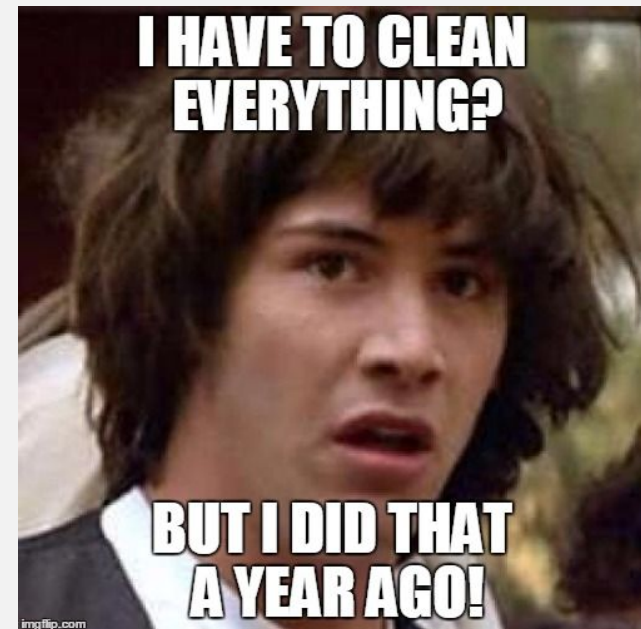
■ Limpeza de dados

- O que significa “limpar dados” ?
- Limpeza de dados (ou Data cleaning) consiste em detectar, remover OU corrigir dados corromptos, incorretos, ausentes ou irrelevantes.
- Limpeza de dados (Data cleaning) é muitas vezes confundido com Pré processamento de dados (Data preprocessing)


















Limpeza de dados

- O que significa “limpar dados” ?
- Limpeza de dados (ou Data cleaning) consiste em detectar, remover OU corrigir dados corruptos, incorretos, ausentes ou irrelevantes.
- Limpeza de dados (Data cleaning) é muitas vezes confundido com Pré processamento de dados (Data preprocessing)
- Principais causas de sujeira:
 - Valores ausentes
 - Caracteres especiais
 - Inconsistências
 - Outliers



Limpeza de dados




- Valores ausentes
 - Em Python normalmente são variáveis do tipo **None**.
 - Dependendo do módulo/biblioteca que estiver usando, pode admitir outros valores, como **NaN** ou **np.nan** (numpy)

entree	pets	emergency_contact
		
shrimp		Pepper
beef		Jane
chicken	62	Janet
beef		Henry
		
veggie		NA
chicken		n/a
shrimp	3	None
shrimp		empty
		-
veggie	1
chicken		null

Limpeza de dados

- Valores ausentes

- Em Python normalmente são variáveis do tipo **None**.
- Dependendo do módulo/biblioteca que estiver usando, pode admitir outros valores, como **NaN** ou **np.nan** (numpy)
- Mas dependendo do caso pode ser valores “especiais”, exemplo:
 - “-” (string)
 - “empty” (string)

entree	pets	emergency_contact
		
shrimp		Pepper
beef		Jane
chicken	62	Janet
beef		Henry
		NA
veggie		n/a
chicken		None
shrimp	3	empty
shrimp		-
	
veggie	1	
chicken		null

Limpeza de dados

O que fazer com valores ausentes?

- Deleta toda a linha ou faz uma imputação
- Tipos de imputação
 - Hot-Deck
 - Substitui o dado ausente por algum outro dado do banco de dados

Limpeza de dados

O que fazer com valores ausentes?

- Deleta toda a linha ou faz uma imputação
- Tipos de imputação
 - Hot-Deck
 - Substitui o dado ausente por algum outro dado do banco de dados
 - Cold-Deck
 - Substitui o dado ausente por um dado de OUTRO banco de dados

Limpeza de dados

O que fazer com valores ausentes?

- Deleta toda a linha ou faz uma imputação
- Tipos de imputação
 - Hot-Deck
 - Substitui o dado ausente por algum outro dado do banco de dados
 - Cold-Deck
 - Substitui o dado ausente por um dado de OUTRO banco de dados
 - Imputação através da média
 - Calcula a média dos dados presentes e substitui nos dados ausentes

Limpeza de dados

O que fazer com valores ausentes?

- Deleta toda a linha ou faz uma imputação
- Tipos de imputação
 - Hot-Deck
 - Substitui o dado ausente por algum outro dado do banco de dados
 - Cold-Deck
 - Substitui o dado ausente por um dado de OUTRO banco de dados
 - Imputação através da média
 - Calcula a média dos dados presentes e substitui nos dados ausentes
 - Imputação através da regressão
 - Cria uma regressão com base nos dados presentes para prever os dados ausentes

Limpeza de dados

Caracteres especiais

- Dados nulos mascarados
 - Ex: “nulo”, “-”, “*”, “vazio” <- String

Limpeza de dados

Caracteres especiais

- Dados nulos mascarados
 - Ex: “nulo”, “-”, “*”, “vazio” <- String
- “Números” especiais
 - Ex: infinito, np.inf (numpy)

Limpeza de dados

Caracteres especiais

- **Dados nulos mascarados**
 - Ex: “nulo”, “-”, “*”, “vazio” <- String
- **“Números” especiais**
 - Ex: infinito, np.inf (numpy)
- **Dados numéricos mascarados**
 - Ex: 42,0 → String
42.0 → Float

Limpeza de dados

Caracteres especiais

- **Dados nulos mascarados**
 - Ex: “nulo”, “-”, “*”, “vazio” <- String
- **“Números” especiais**
 - Ex: infinito, np.inf (numpy)
- **Dados numéricos mascarados**
 - Ex: 42,0 → String
42.0 → Float

Observação: Caracteres especiais podem “poluir” resumos estatísticos do módulo Pandas ou operações matemáticas

Limpeza de dados

Inconsistências

- **Valores que são claramente incompatíveis com o resto dos dados ou que fogem do bom senso.**
 - **Ex: Altura de uma pessoa ser 10 metros**

Limpeza de dados

Inconsistências

- **Valores que são claramente incompatíveis com o resto dos dados ou que fogem do bom senso.**
 - **Ex: Altura de uma pessoa ser 10 metros**
 - **Ex: -10°C em Fortaleza em pleno verão**

Limpeza de dados

Inconsistências

- **Valores que são claramente incompatíveis com o resto dos dados ou que fogem do bom senso.**
 - **Ex: Altura de uma pessoa ser 10 metros**
 - **Ex: -10°C em Fortaleza em pleno verão**
 - **Ex: 0°C em um forno que opera na faixa de 500°C**

Limpeza de dados

Inconsistências

- **Valores que são claramente incompatíveis com o resto dos dados ou que fogem do bom senso.**
 - **Ex: Altura de uma pessoa ser 10 metros**
 - **Ex: -10°C em Fortaleza em pleno verão**
 - **Ex: 0°C em um forno que opera na faixa de 500°C**
 - **Ex: \$1.000.000,00 na minha conta bancária**

Limpeza de dados

Outliers

- **Valores que são incompatíveis com o resto dos dados, mas que não necessariamente estão incorretos.**
 - **Ex: Altura de uma pessoa ser 2.3 metros**

Limpeza de dados

Outliers

- **Valores que são incompatíveis com o resto dos dados, mas que não necessariamente estão incorretos.**
 - **Ex: Altura de uma pessoa ser 2.3 metros**
 - **Ex: 15°C em Fortaleza em pleno verão**

Limpeza de dados

Outliers

- **Valores que são incompatíveis com o resto dos dados, mas que não necessariamente estão incorretos.**
 - **Ex: Altura de uma pessoa ser 2.3 metros**
 - **Ex: 15°C em Fortaleza em pleno verão**
 - **Ex: 300°C em um forno que opera na faixa de 500°C**

Limpeza de dados

Outliers

- **Valores que são incompatíveis com o resto dos dados, mas que não necessariamente estão incorretos.**
 - **Ex: Altura de uma pessoa ser 2.3 metros**
 - **Ex: 15°C em Fortaleza em pleno verão**
 - **Ex: 300°C em um forno que opera na faixa de 500°C**
 - **Ex: A conta bancária do Bill Gates**

Limpeza de dados

Como detectar Outliers?

Como detectar outliers?

Limpeza de dados

Como detectar Outliers?

Height M
1.5895
1.6508
1.7131
1.7136
1.7212
1.7296
1.7343
1.7663
1.8018
1.8394
1.8869
1.9357
1.9482
2.1038
10.8135

Ordenando dados

Limpeza de dados

Como detectar Outliers?

Height M
1.5895
1.6508
1.7131
1.7136
1.7212
1.7296
1.7343
1.7663
1.8018
1.8394
1.8869
1.9357
1.9482
2.1038
10.8135

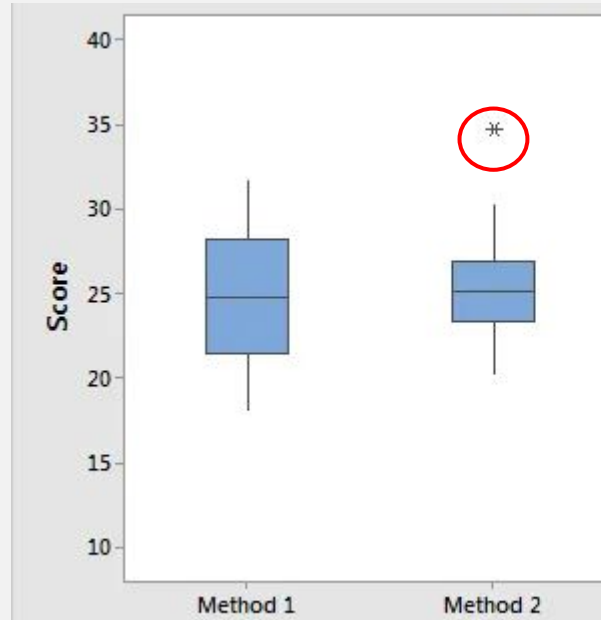


Diagrama de caixas

Ordenando dados

Limpeza de dados

Como detectar Outliers?

Height M
1.5895
1.6508
1.7131
1.7136
1.7212
1.7296
1.7343
1.7663
1.8018
1.8394
1.8869
1.9357
1.9482
2.1038
10.8135

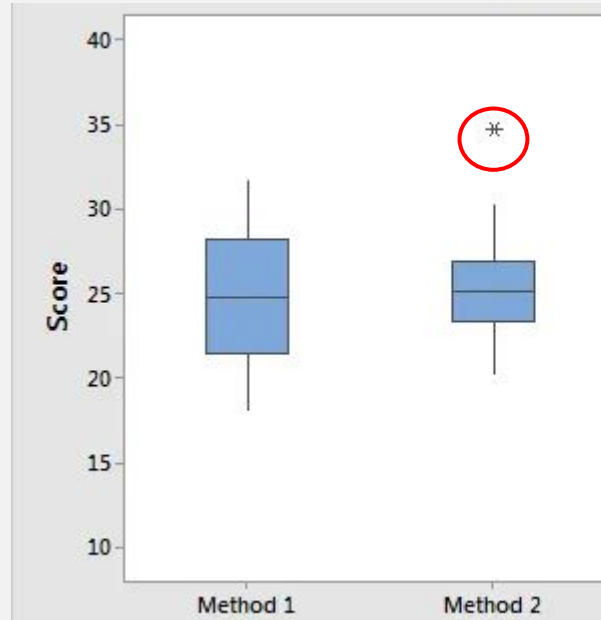
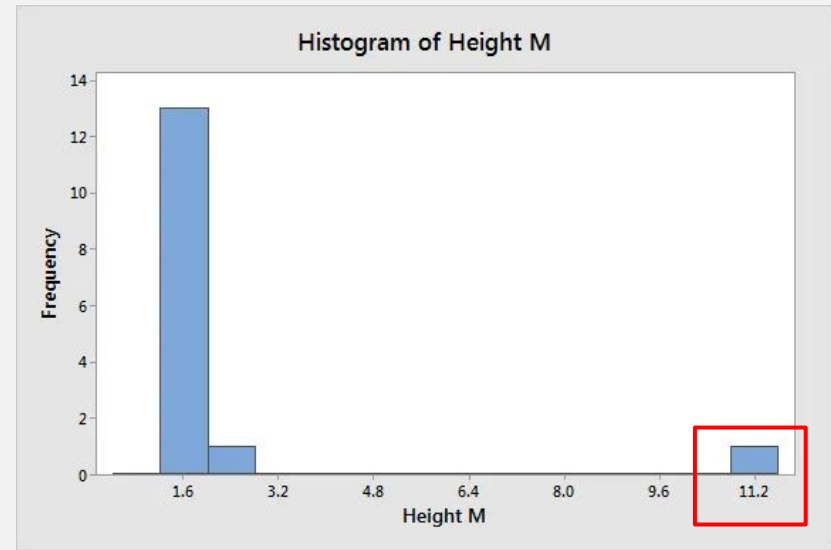


Diagrama de caixas



Histogramas

Ordenando dados

Limpeza de dados

Como detectar Outliers?

Height M
1.5895
1.6508
1.7131
1.7136
1.7212
1.7296
1.7343
1.7663
1.8018
1.8394
1.8869
1.9357
1.9482
2.1038
10.8135

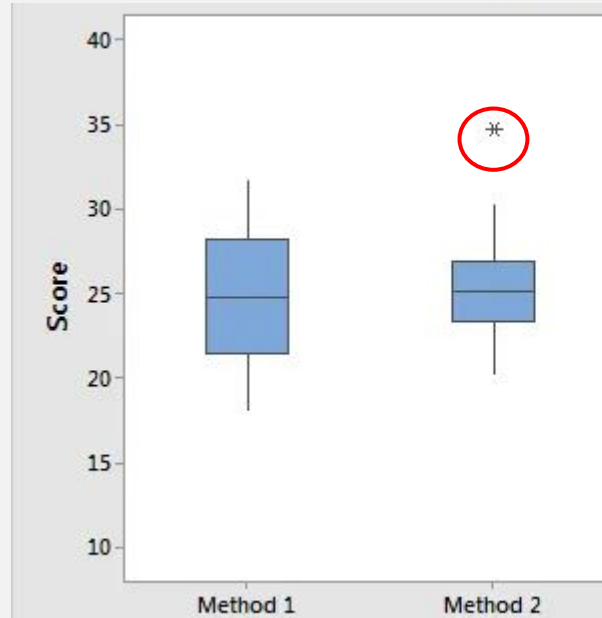
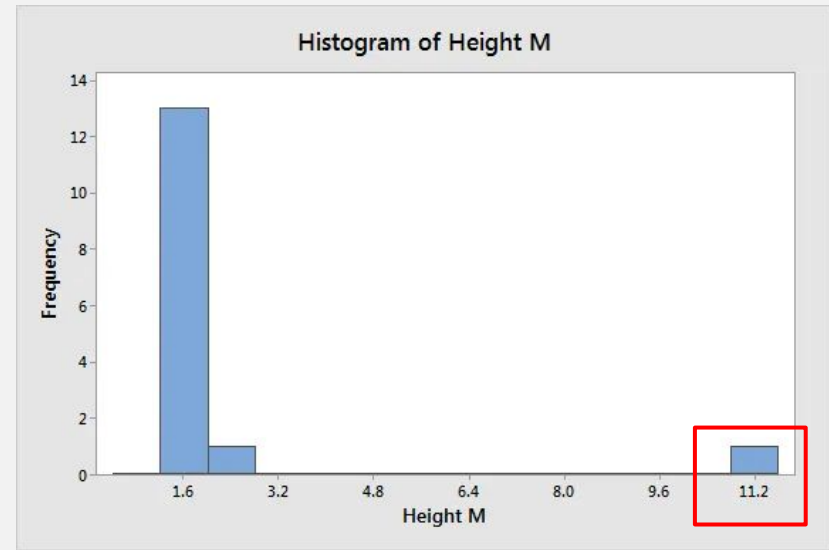


Diagrama de caixas



Histogramas

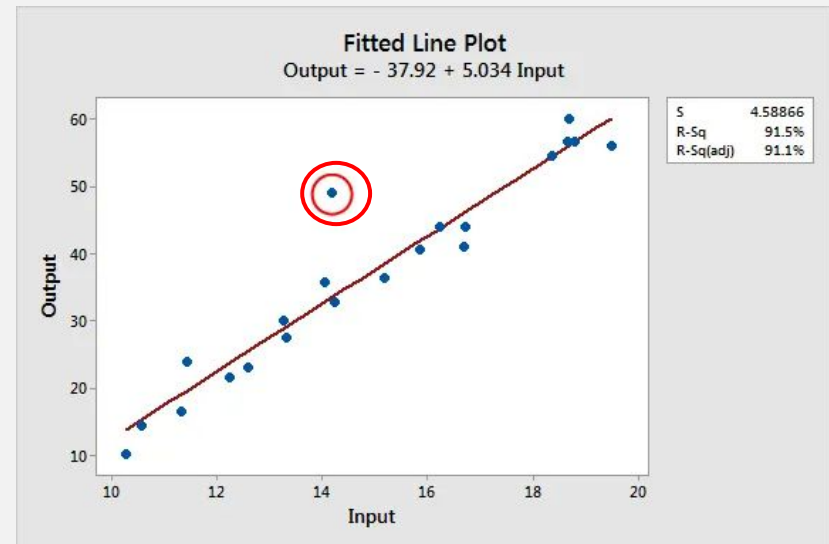
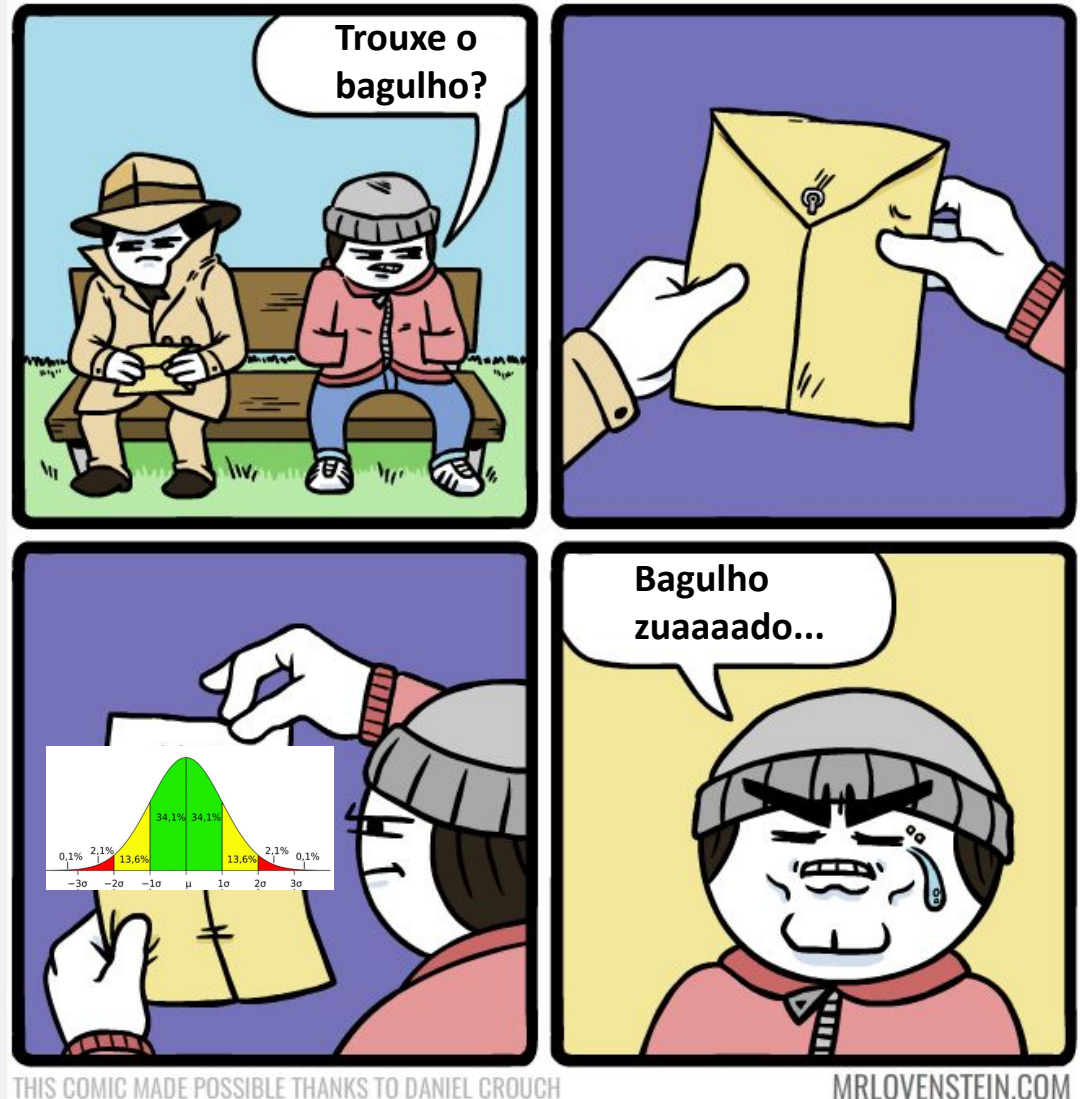


Gráfico de Dispersão

Limpeza de dados

Como detectar Outliers?

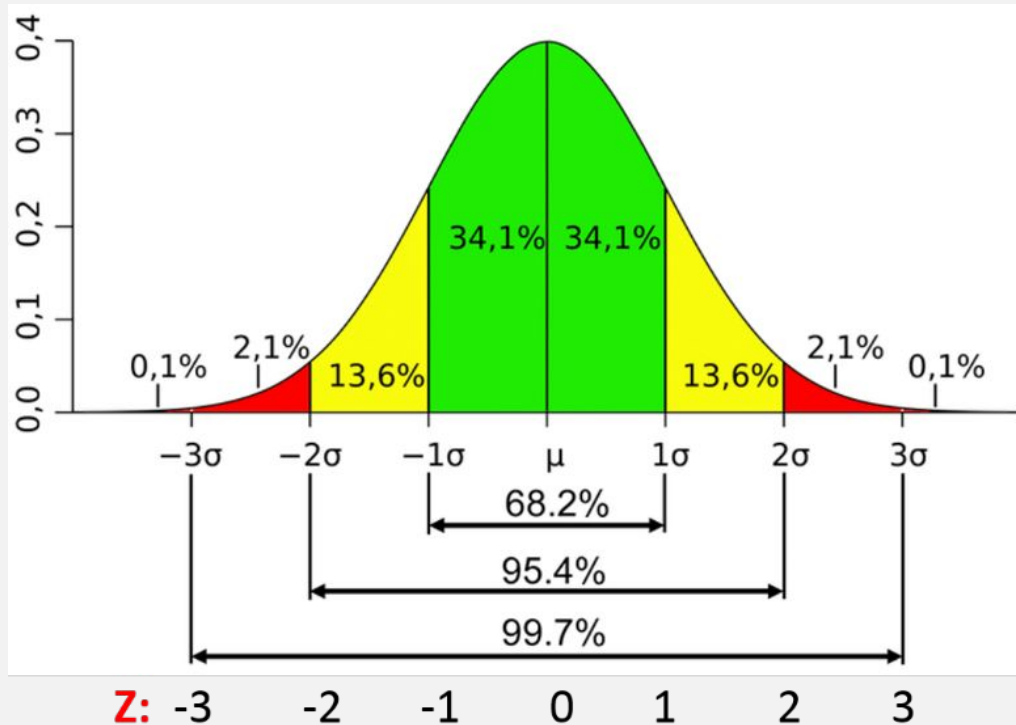
Se os dados seguirem uma distribuição normal, existe uma técnica que pode ser utilizada de forma QUANTITATIVA de forma robusta: Z-test ou T-test



Limpeza de dados

Como detectar Outliers?

Consiste em aplicar uma fórmula para calcular o quão “dentro do esperado” é um valor em uma distribuição



Limpeza de dados

Como detectar Outliers?

Consiste em aplicar uma fórmula para calcular o quão “dentro do esperado” é um valor em uma distribuição

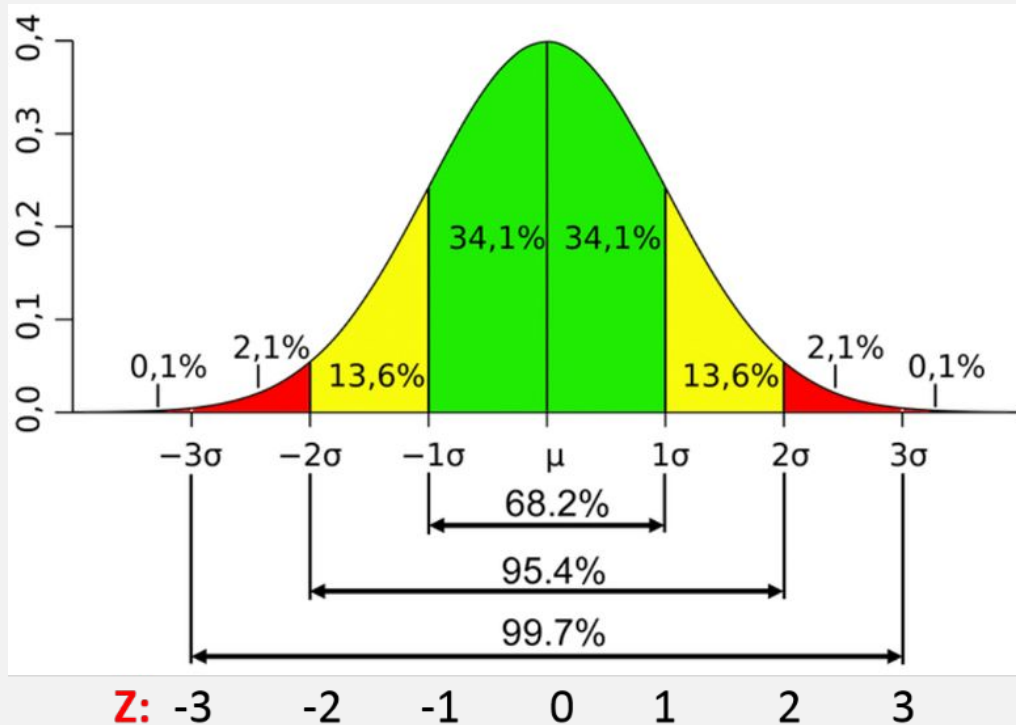
$$Z = \frac{X - \mu}{\sigma}$$

Onde:

X = valor

μ = média

σ = desvio padrão



Limpeza de dados

Como detectar Outliers?

Consiste em aplicar uma fórmula para calcular o quão “dentro do esperado” é um valor em uma distribuição

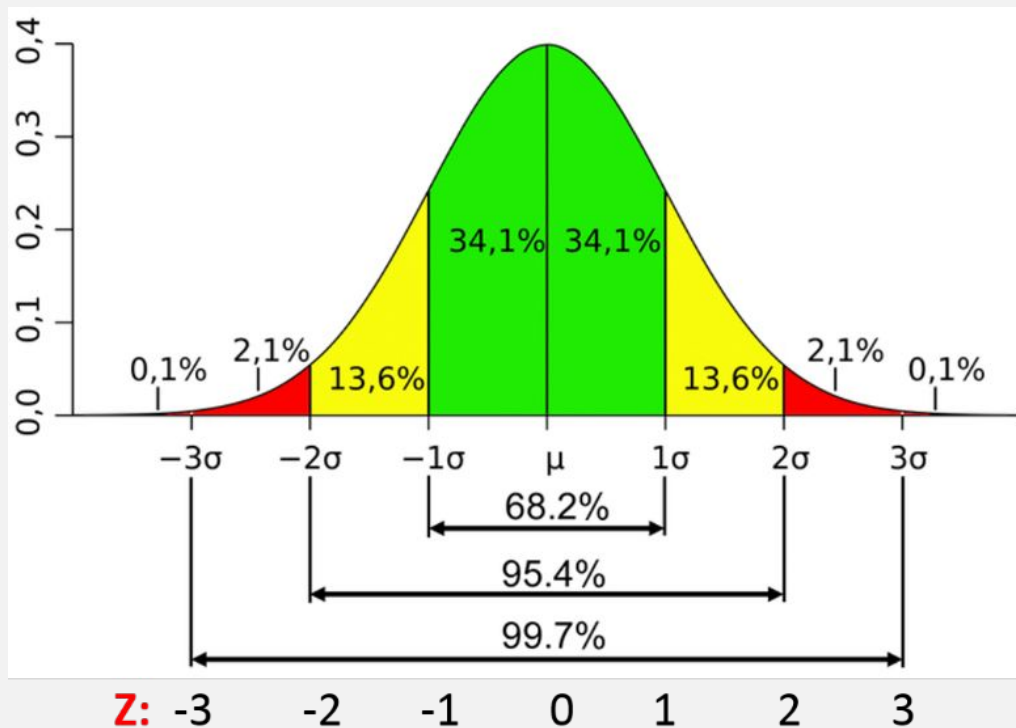
$$Z = \frac{X - \mu}{\sigma}$$

Onde:

X = valor

μ = média

σ = desvio padrão



Height M	Z-score
1.5895	-0.34603
1.6508	-0.31975
1.7131	-0.29301
1.7136	-0.29283
1.7212	-0.28954
1.7296	-0.28595
1.7343	-0.28394
1.7663	-0.27020
1.8018	-0.25501
1.8394	-0.23888
1.8869	-0.21852
1.9357	-0.19757
1.9482	-0.19223
2.1038	-0.12551
10.8135	3.60910

Feature Engineering

- **Consiste em CRIAR novas variáveis de entradas utilizando variáveis já presentes no banco de dados. São divididas em diferentes categorias:**
 - **Decomposição / Splitting**
 - **Crossing**
 - **Reframing / Reinterpretação**
 - **Discretização ou Binning**

Feature Engineering

- Decomposição / Splitting
 - Consiste em “separar” uma variável em múltiplas partes
 - Ex: Decompõe **Data** = 2014-09-20T20:45:40Z em **Ano** = 2014, **Mês** = 09, **Dia** = 20
 - **Peso** = 6.280 -> **Quilo** = 6, **Gramas** = 0.280

Feature Engineering

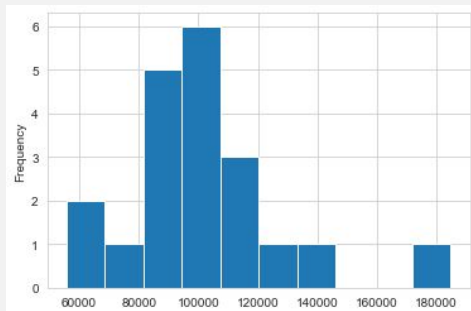
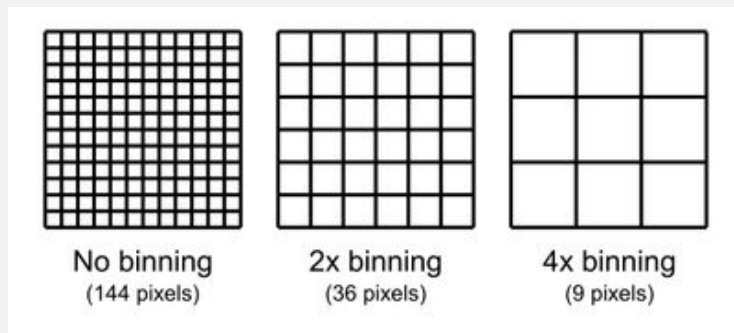
- Decomposição / Splitting
 - Consiste em “separar” uma variável em múltiplas partes
 - Ex: Decompõe **Data** = 2014-09-20T20:45:40Z em **Ano** = 2014, **Mês** = 09, **Dia** = 20
 - **Peso** = 6.280 -> **Quilo** = 6, **Gramas** = 0.280
- Crossing
 - Combina diferentes variáveis em uma nova variável
 - Ex: Multiplicação ou divisão de uma variável por outra variável

Feature Engineering

- Reframing
 - Consiste em “reinterpretar” um dado
 - Ex: trocar unidades de peso
 - Transformar kg em g, milhas em quilômetros, etc.

Feature Engineering

- Reframing
 - Consiste em “reinterpretar” um dado
 - Ex: trocar unidades de peso
 - Transformar kg em g, milhas em quilômetros, etc.
- Discretização ou Binning
 - Consiste em agrupar valores contínuos em categorias ou intervalos maiores



	id	age
0	30669	3.0
1	30468	58.0
2	16523	8.0
3	56543	70.0
4	46136	14.0
5	32257	47.0
6	52800	52.0
7	41413	75.0
8	15266	32.0
9	28674	74.0



	id	age	age_range
0	30669	3.0	minor
1	30468	58.0	old
2	16523	8.0	minor
3	56543	70.0	old
4	46136	14.0	minor
5	32257	47.0	young
6	52800	52.0	old
7	41413	75.0	very_old
8	15266	32.0	young
9	28674	74.0	very_old

Feature Selection e Redução de Dimensionalidade

• Seleção de Features e Redução de Dimensionalidade são o processo de REDUZIR o número de variáveis de entrada para uma modelagem com os objetivos de:

- ?
- ?
- ?

Feature Selection e Redução de Dimensionalidade

• Seleção de Features e Redução de Dimensionalidade são o processo de REDUZIR o número de variáveis de entrada para uma modelagem com os objetivos de:

- Remover variáveis de entrada redundantes
- Reduzir o custo computacional
- Reduzir a complexidade do modelo
- Aumentar a interpretabilidade do modelo
- Aumentar a acurácia do modelo, se possível

Feature Selection e Redução de Dimensionalidade

- Seleção de Features e Redução de Dimensionalidade são o processo de REDUZIR o número de variáveis de entrada para uma modelagem com os objetivos de:
 - Remover variáveis de entrada redundantes
 - Reduzir o custo computacional
 - Reduzir a complexidade do modelo
 - Aumentar a interpretabilidade do modelo
 - Aumentar a acurácia do modelo, se possível
- Seleção de Features → Supervisionado
- Redução de Dimensionalidade → Não supervisionado

Feature Selection e Redução de Dimensionalidade

- Seleção de Features e Redução de Dimensionalidade são o processo de REDUZIR o número de variáveis de entrada para uma modelagem com os objetivos de:
 - Remover variáveis de entrada redundantes
 - Reduzir o custo computacional
 - Reduzir a complexidade do modelo
 - Aumentar a interpretabilidade do modelo
 - Aumentar a acurácia do modelo, se possível
- Seleção de Features → Supervisionado
- Redução de Dimensionalidade → Não supervisionado
- Seleção de Features pode ser divididos em 2 tipos principais:
 - Abordagem Wrapper
 - Abordagem de Filtros

Feature Selection

- **Abordagem Wrapper**
 - **Treina vários modelos com diferentes subconjuntos de features e seleciona o melhor**
 - **Isso é bom? Isso é ruim?**

Feature Selection

- **Abordagem Wrapper**
 - **Treina vários modelos com diferentes subconjuntos de features e seleciona o melhor**
 - **Características:**
 - **Computacionalmente exigente**

Feature Selection

- **Abordagem Wrapper**
 - **Treina vários modelos com diferentes subconjuntos de features e seleciona o melhor**
 - **Características:**
 - **Computacionalmente exigente**
 - **O resultado é dependente do modelo (as melhores features de uma Rede Neural pode ser diferente de uma Random Forest)**

Feature Selection

- **Abordagem Wrapper**
 - **Treina vários modelos com diferentes subconjuntos de features e seleciona o melhor**
 - **Características:**
 - **Computacionalmente exigente**
 - **O resultado é dependente do modelo (as melhores features de uma Rede Neural pode ser diferente de uma Random Forest)**
 - **Se tiver poucos dados, aumenta o risco de overfit.**

Feature Selection

- **Abordagem Wrapper**
 - **Treina vários modelos com diferentes subconjuntos de features e seleciona o melhor**
 - **Características:**
 - **Computacionalmente exigente**
 - **O resultado é dependente do modelo (as melhores features de uma Rede Neural pode ser diferente de uma Random Forest)**
 - **Se tiver poucos dados, aumenta o risco de overfit.**
 - **Resultados tendem a ser melhores do que abordagem de filtro**

Feature Selection

- **Abordagem Wrapper**
 - **Treina vários modelos com diferentes subconjuntos de features e seleciona o melhor**
 - **Características:**
 - **Computacionalmente exigente**
 - **O resultado é dependente do modelo (as melhores features de uma Rede Neural pode ser diferente de uma Random Forest)**
 - **Se tiver poucos dados, aumenta o risco de overfit.**
 - **Resultados tendem a ser melhores do que abordagem de filtro**
 - **Técnicas:**
 - **Forward Selection, Recursive Feature Elimination (RFE), Algoritmos Genéticos, etc.**

Feature Selection

- Podem ser divididos em 2 tipos principais:
 - Abordagem de Filtros
 - Calcula métricas estatísticas das variáveis de entrada e elimina aquelas que forem menos interessantes

Feature Selection

- Podem ser divididos em 2 tipos principais:
 - Abordagem de Filtros
 - Calcula métricas estatísticas das variáveis de entrada e elimina aquelas que forem menos interessantes
 - Características:
 - Computacionalmente barato
 - Robusto contra overfit

Feature Selection

- Podem ser divididos em 2 tipos principais:
 - Abordagem de Filtros
 - Calcula métricas estatísticas das variáveis de entrada e elimina aquelas que forem menos interessantes
 - Características:
 - Computacionalmente barato
 - Robusto contra overfit
 - Técnicas:
 - Análise de correlação, Teste Chi-Quadrado, ANOVA: Análise de Variância, etc

Feature Encoding

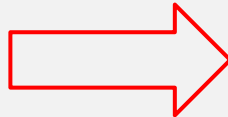
- Consiste em “traduzir” um dado para um formato que seja adequado para modelagem
 - Exemplo: Converter dados categóricos em numéricos



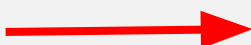
Feature Encoding

- **Label Encoding**
 - Forma mais simples de realizar Feature Encoding.
 - Consiste em substituir um dado categórico por um dado numérico.

Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000



Country	Age	Salary
0	44	72000
2	34	65000
1	46	98000
2	35	45000
1	23	34000

- **Problema:**
 - India = 0, Japan = 0, US = 2
 - $0 < 1 < 2$  India < Japan < US
- **Quando usar?**
 - Variáveis categóricas são ordinais
 - Número de categorias é muito grande

Feature Encoding

- One Hot Encoding
 - Substitui dados categóricos em uma LISTA de valores numéricos

Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000



India	Japan	US	Age	Salary
1	0	0	44	72000
0	0	1	34	65000
0	1	0	46	98000
0	0	1	35	45000
0	1	0	23	34000

- Problema:
 - Número de colunas adicionadas pode crescer exponencialmente
 - Pode surgir alta correlação entre as colunas
- Quando usar?
 - Variáveis categóricas não são ordinais
 - Número de categorias não é muito grande

Reescalonamento

- Altera TODO o intervalo de dados de modo que os dados fiquem mais adequados para o treinamento do modelo
- Existem escalonadores lineares e não lineares

Referências

- <https://bi-survey.com/role-of-data>
- <http://leg.ufpr.br/~silvia/CE055/node8.html>
- <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- <https://www.nature.com/articles/s42256-021-00307-0>
- <https://statisticsbyjim.com/basics/outliers/>
-