Data Science & Business Analytics

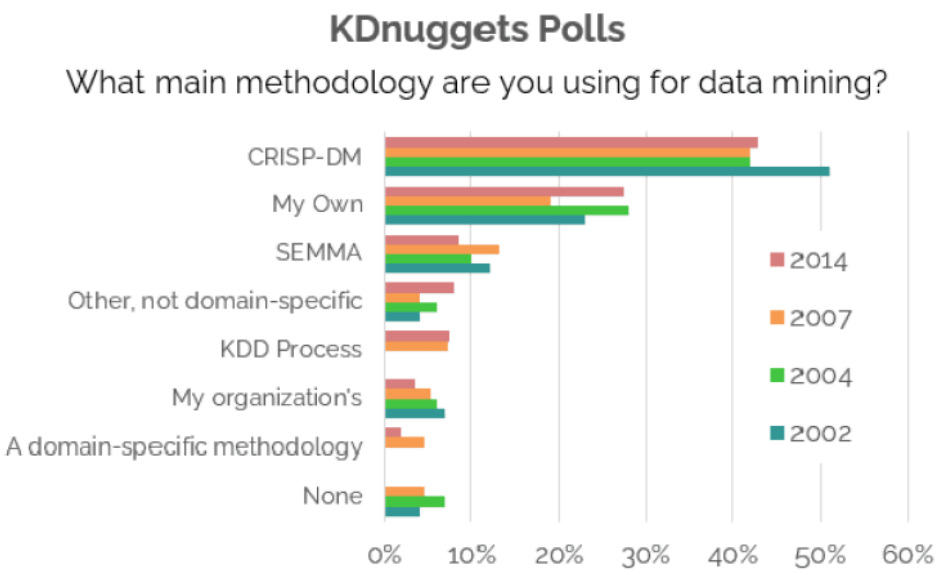
Machine Learning Models

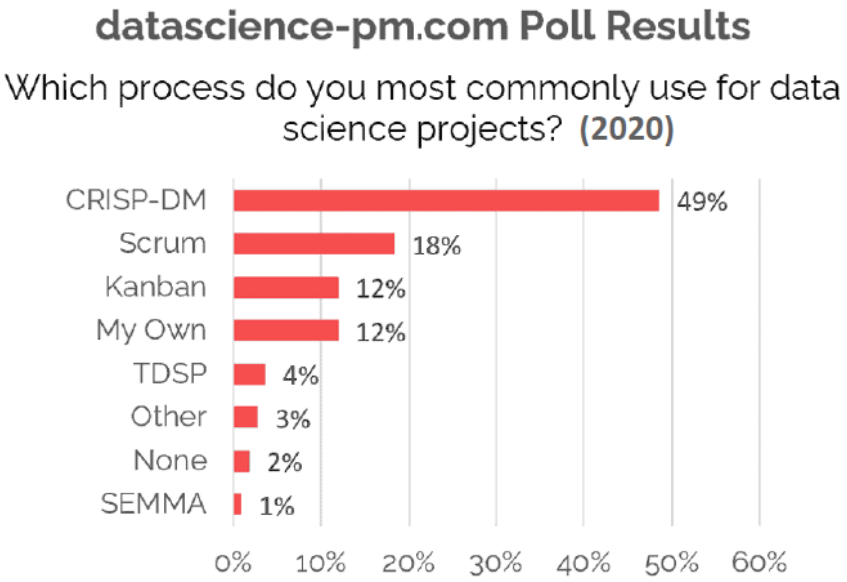
David Issá
davidribeiro.issa@gmail.com

1. Metodologias de Data Science

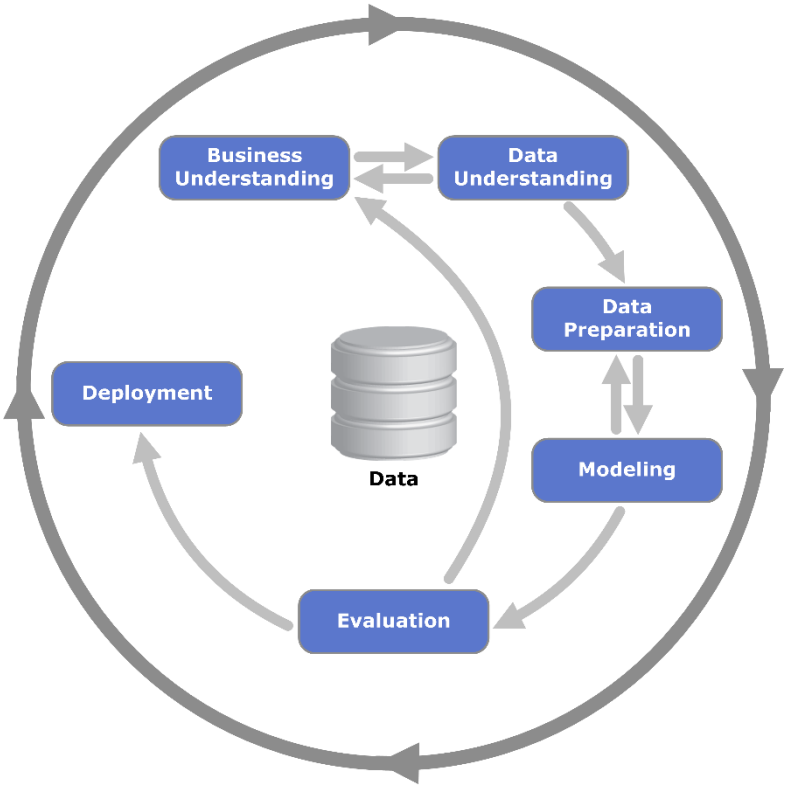
1. Metodologias de Data Science

Em Data Science, é importanto usar uma determinada metodologia em cada projeto, facilitanto a sua gestão e planeamento.



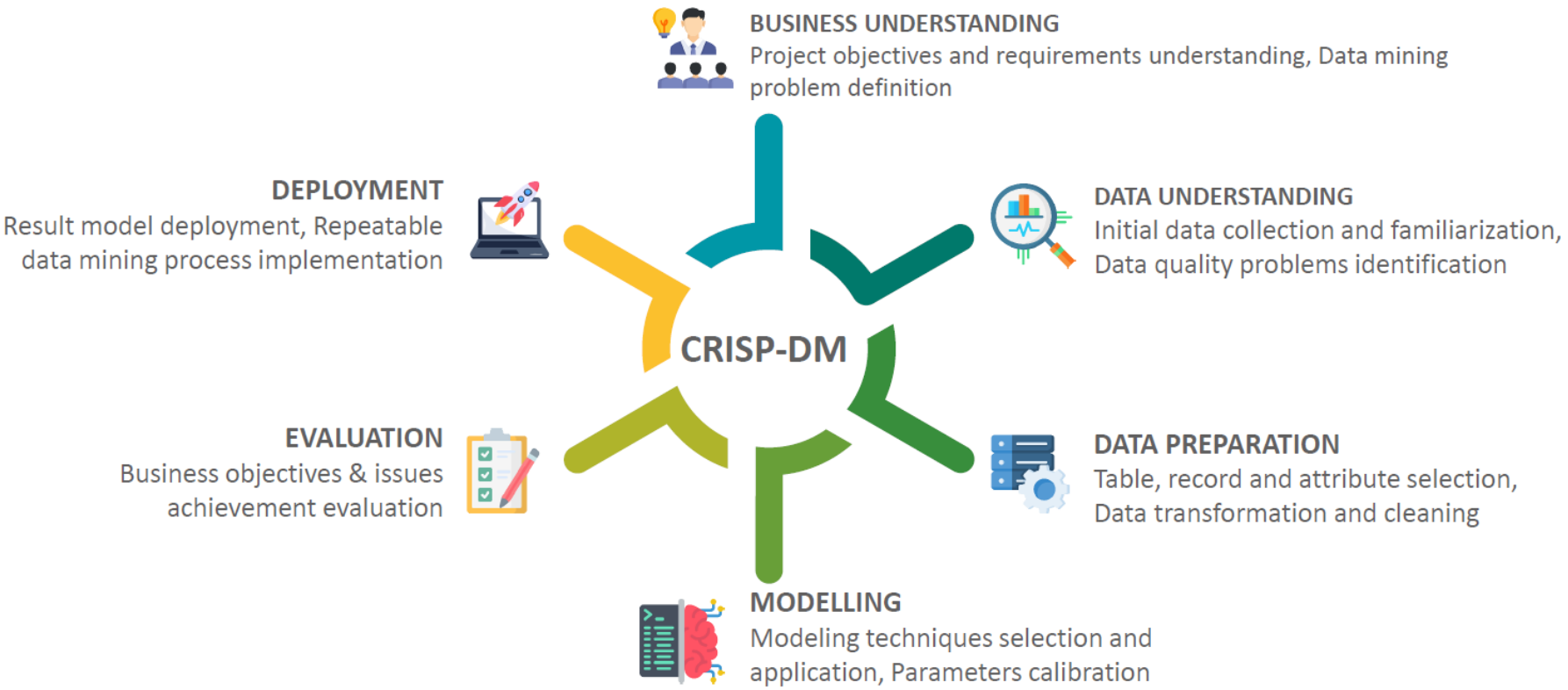


1. Metodologias de Data Science – CRISP-DM



- CRISP-DM (Cross Industry Standard Process for Data Mining), originado em 1996;
- Framework para orientação;
- Ciclo de vida: 6 fases;
- Neutro em termos de aplicação/indústria;
- Foco nas questões business, bem como na análise técnica.

1. Metodologias de Data Science – CRISP-DM



1. Metodologias de Data Science – CRISP-DM

BUSINESS UNDERSTANDING

- Determinar objetivos de negócio: quais as perguntas a responder?
- Determinar objetivos de data mining: o que pretendo prever?
- Produzir plano do projeto: step by step

DATA UNDERSTANDING

- Adquirir dados inicias.
- Descrever os dados.
- Explorar os dados, com os objetivos definidos anteriormente em mente: estatísticas, visualizações, etc.
- Qualidade de dados: missing values, incosistências, etc.

1. Metodologias de Data Science – CRISP-DM

DATA PREPARATION

- Selecionar e limpar os dados, resolvendo as issues detetadas.
- Constuir novos dados com base nos dados iniciais.
- Integrar e combinar dados de diferentes fontes.

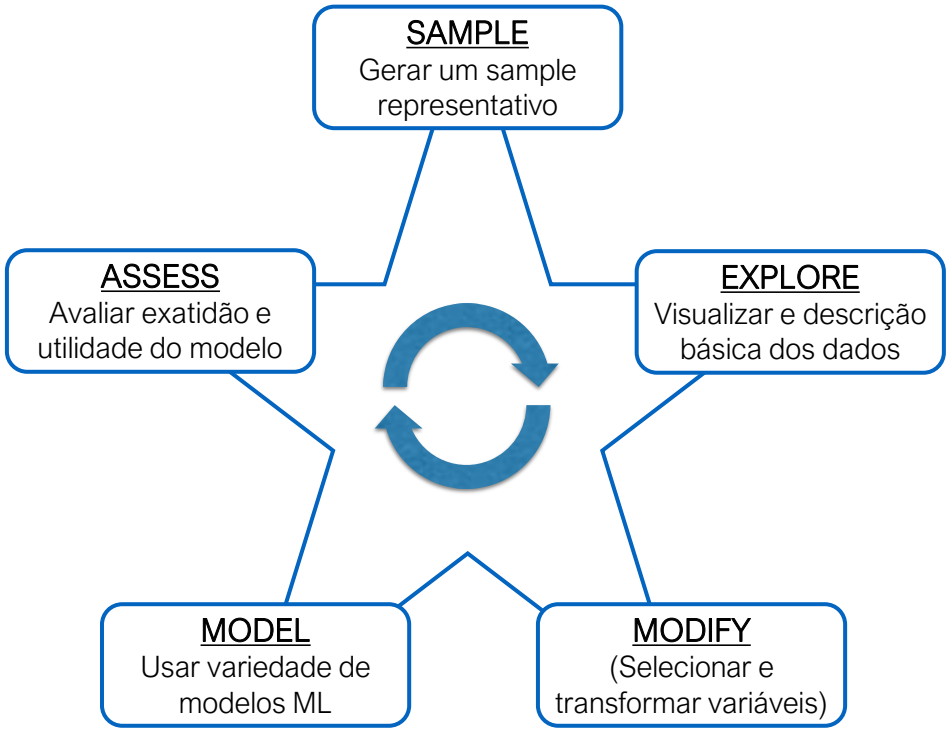
MODELLING

- Selecionar modelos a testar.
- Definir o processo de teste.
- Construir o modelo e interpretar os resultados.

EVALUATION

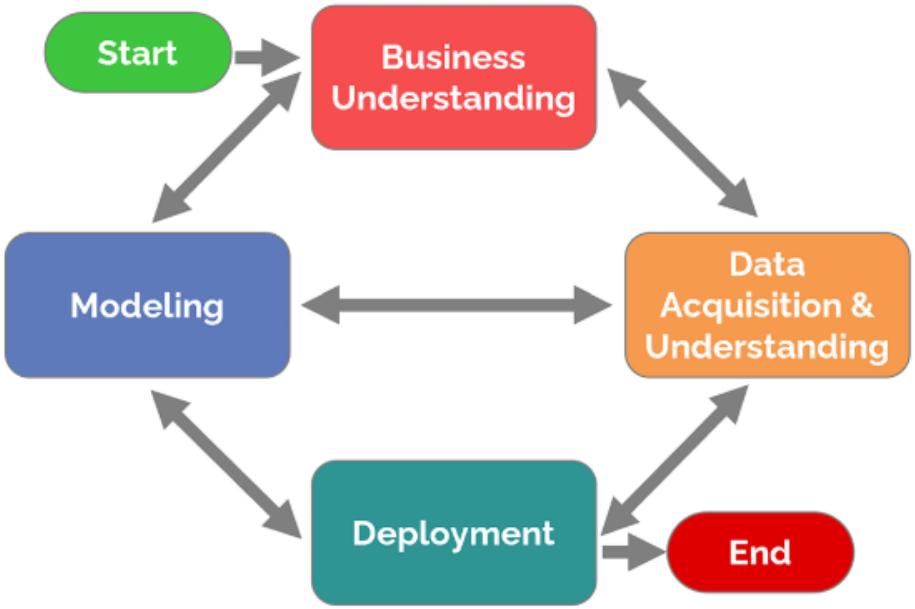
- Avaliar os resultados do modelo.
- Rever todo o processo de data mining.
- Determinar os próximos passos.

1. Metodologias de Data Science – SEMMA



- SEMMA (Explore, Modify, Model, Assess) é uma metodologia utilizada pelo SAS Institute para projectos de data mining, com ênfase na análise estatística.
- Começa com a amostra inicial de dados e passa por uma série de etapas que envolvem a exploração e modificação de dados, o desenvolvimento de modelos e a sua avaliação.

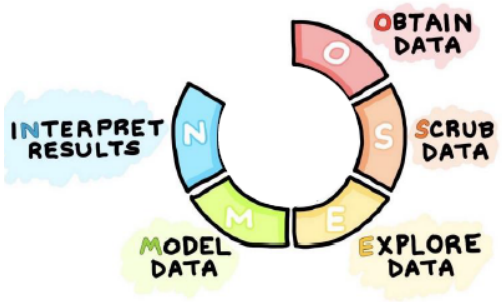
1. Metodologias de Data Science – TDSP



- TDSP (Team Data Science Process) é uma metodologia desenvolvida pela Microsoft.
- Enfatiza a colaboração entre os membros da equipa, a escalabilidade e a reprodutibilidade no processo de análise de dados.
- Processo moderno que combina elementos do ciclo de vida da Data Science, sofware engineering e processos Agile.

1. Metodologias de Data Science – OSEMN

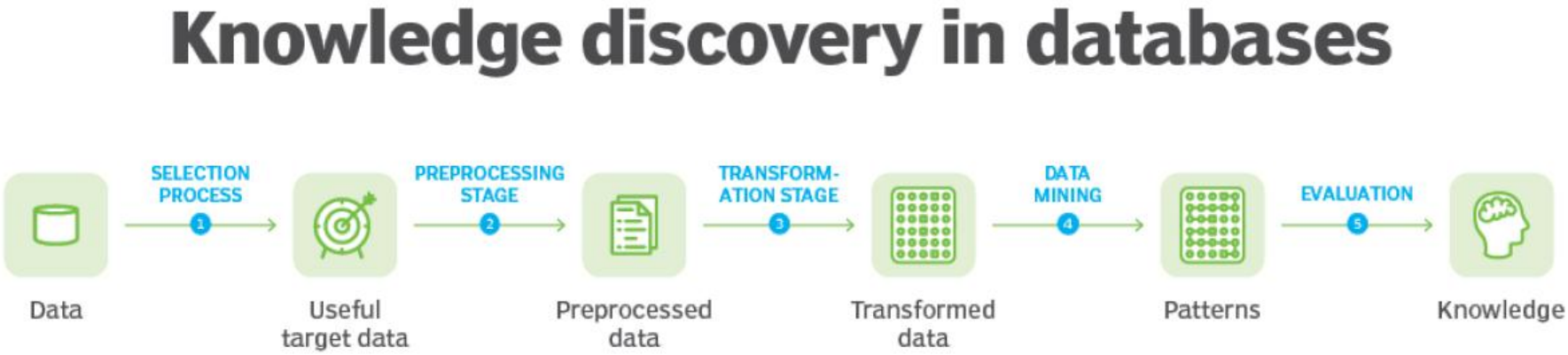
- OSEMN (Obtain, Scrub, Explore, Model, Interpret) é uma metodologia, nascida em 2010, que foi popularizada pela comunidade de data science;
- Enfatiza a importância da qualidade e exploração dos dados no processo de análise de dados.



O	S	E	M	N
OBTAIN	SCRUB	EXPLORE	MODEL	INTEGRATE
Gather data from relevant sources	Data cleaning to correct formats	Exploratory data analysis and find patterns	Build models based to predict and forecast	Deploy models and turn insights into actions

1. Metodologias de Data Science – KDD

- KDD (Knowledge Discovery in Databases) tem como principal objetivo extrair conhecimento de grandes conjuntos de dados, utilizando técnicas estísticas e de Machine Learning.
- É um processo iterativo que envolve várias técnicas, tais como clustering, análise de associação, análise de classificação e regressão, entre outras.



1. Metodologias de Data Science – Resumo

Cada metodologia fornece uma abordagem estruturada para a análise de dados, mas têm fases e abordagens diferentes, enfatizando diferentes aspectos do processo de análise de dados:

- CRISP-DM enfatiza a compreensão do problema de negócio
- SEMMA enfatiza a análise estatística
- TDSP privilegia a colaboração e a escalabilidade
- OSEMN dá ênfase à qualidade e exploração dos dados
- KDD enfatiza a extração de conhecimentos a partir de grandes conjuntos de dados utilizando técnicas estatísticas e de Machine Learning.

2. Data Understanding

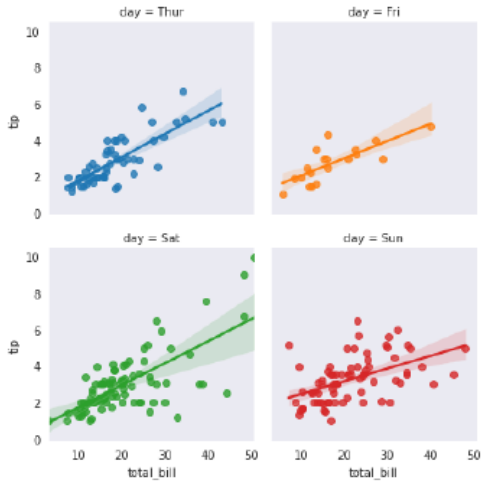
2. Data Understanding

Como compreender os dados?

	id	iv2	rt
count	120.000000	120.00000	120.000000
mean	9.500000	2.00000	877.587425
std	5.790459	0.81992	309.293048
min	0.000000	1.00000	283.240752
25%	4.750000	1.00000	582.630955
50%	9.500000	2.00000	902.719888
75%	14.250000	3.00000	1114.050194
max	19.000000	3.00000	1472.688933

Summary statistics

+



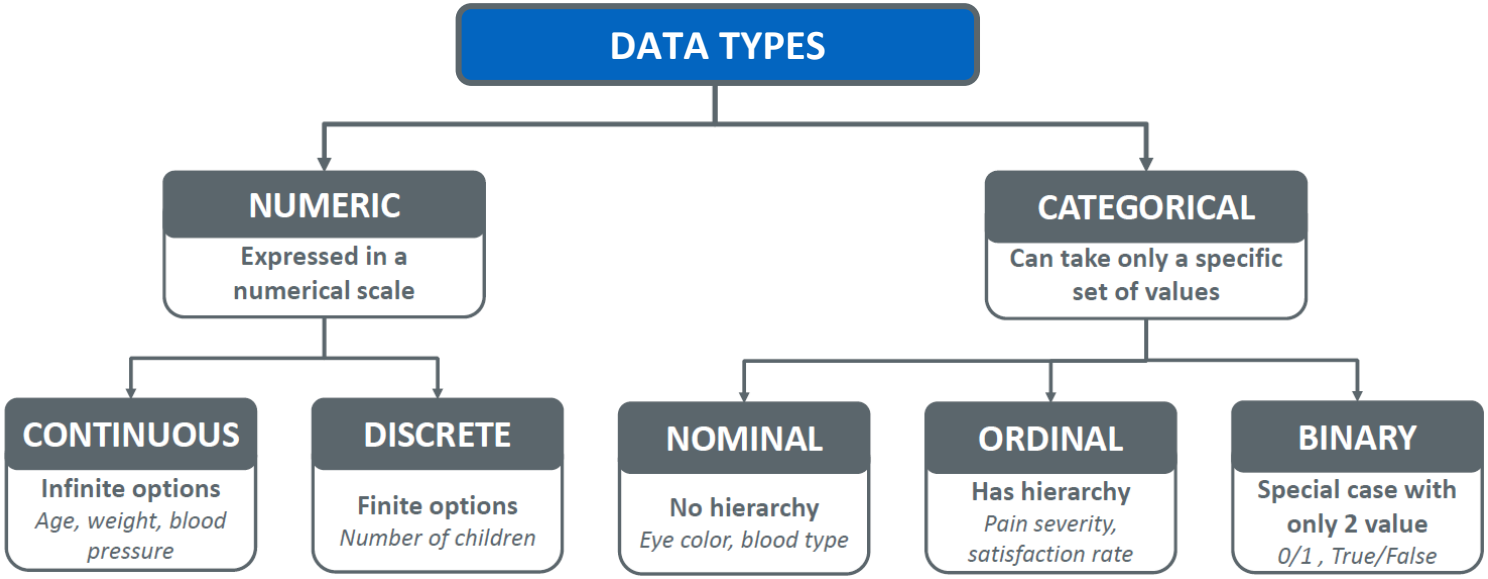
Visualizações

=

Exploratory Data Analysis (EDA)

2.1 Data Understanding – Summary Statistics

As summary statistics a analisar dependem do tipo de dados das nossas variáveis:



2.1 Data Understanding – Summary Statistics

Esta etapa pretende responder às seguintes questões:

- Os dados são bons? Estão limpos? São representativos do que é suposto medirem? Estão preenchidos? Estão distribuídos como esperado? Serão úteis para a construção de modelos?

As summary statistics fornecem um resumo rápido dos dados, descrevendo assim as principais características do nosso dataset. Existem 2 tipos de summary statistics:

- Medidas de tendência central: média, mediana, moda, etc.
- Medidas de dispersão: variância, desvio padrão, range, range interquartil, etc.

2.1 Data Understanding – Summary Statistics

Média (μ) - a média de todos os valores de uma variável numérica.

- Soma de todos os valores da variável dividida pela sua contagem de valores.
- Representa a tendência central se a distribuição da variável for normal ou uniforme. Para dados enviesados, a média torna-se menos representativa da tendência central.
- A média é sensível a outliers.

$$Mean = \bar{x} = \frac{\sum_{i=1}^{n} \bar{x_i}}{n}$$

Mediana - o valor “do meio” numa variável numérica quando os valores são ordenados.

- Menos sensível a outliers ou dados enviasados do que a média.

2.1 Data Understanding – Summary Statistics

Range - a diferença entre os valores mais altos e mais baixos de uma variável.

$$Range = max - min$$

Variância – medida do grau de dispersão de uma variável.

$$Variance = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Desvio-Padrão – raiz quadrada da variância e indica o grau de desvio em relação à média.

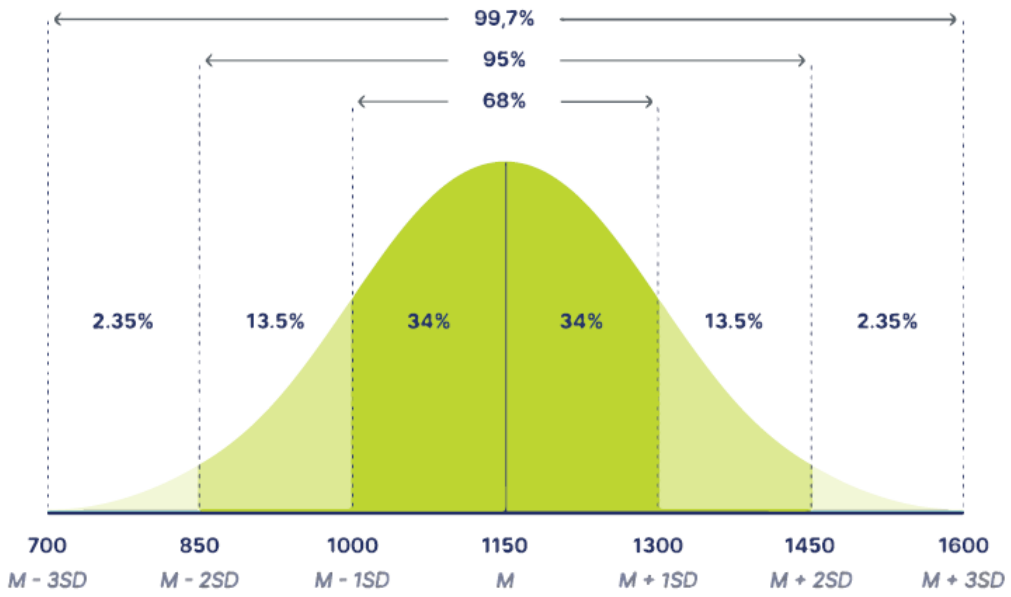
- A interpretabilidade é maior versus a variância, visto que ao tirar a raiz quadrada, a medida regressa às unidades originais da variável em questão.
- Quanto maior, maior é a amplitude aa distribuição dos valores da variável.

$$Standard\ Deviation = \sigma = \sqrt{Variance}$$

2.1 Data Understanding – Summary Statistics

Distribuição Normal

- Muitos algoritmos assumem distribuições normais;
- A distribuição é simétrica;
- O valor médio é o valor mais provável de ocorrer na distribuição;
- A média, a mediana e a moda têm todos o mesmo valor;
- Aproximadamente 95% dos dados situar-se-ão entre amédia e ± 2 desvios-padrão da média.

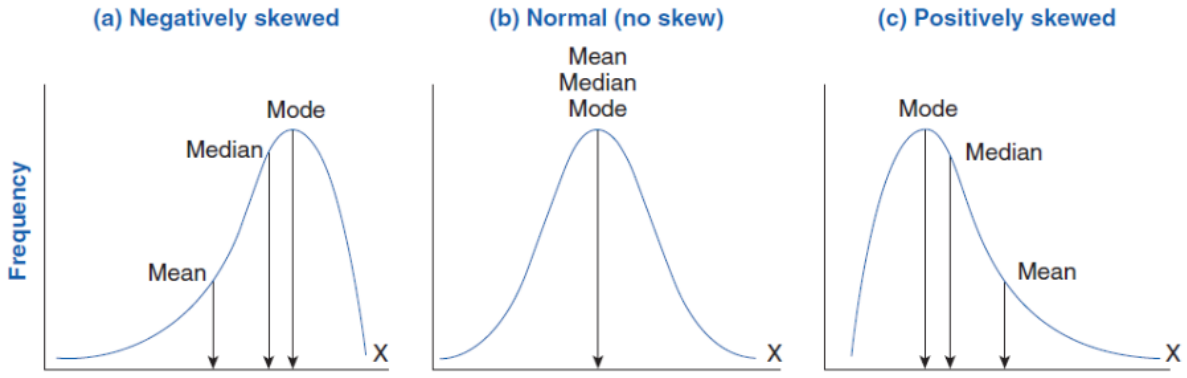


2.1 Data Understanding – Summary Statistics

Skeweness (Enviazamento): uma medida da assimetria da distribuição dos dados.

- Uma distribuição normal tem uma skweness = 0, pelo que a distribuição é simétrica;
- Skweness negativa: a distribuição tem uma cauda à esquerda do corpo principal da distribuição.
- Skweness positiva: a distribuição tem uma cauda à direita do corpo principal da distribuição;

$$Skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\sigma^3}$$

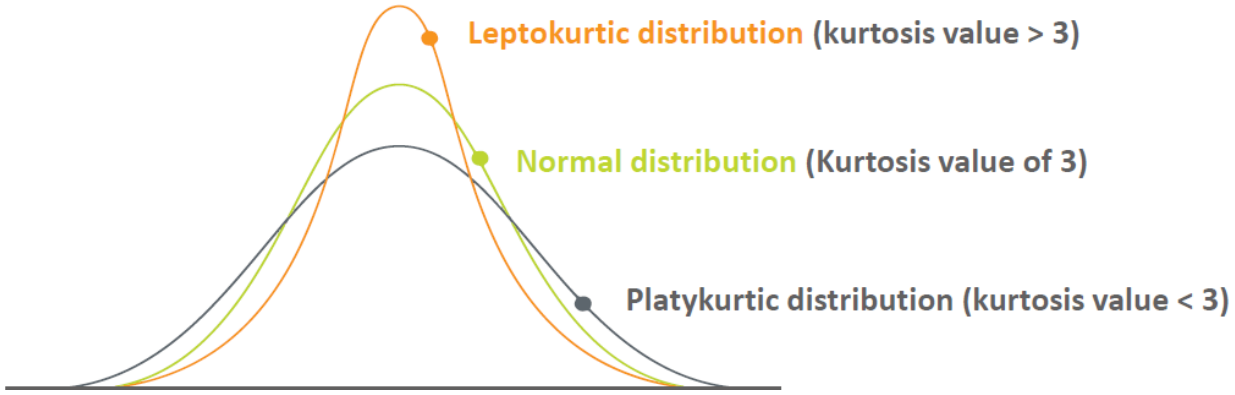


2.1 Data Understanding – Summary Statistics

Kurtosis: medida do grau de pico da distribuição dos dados.

- mede o quanto a distribuição é mais fina ou mais gorda em comparação com as distribuições normais.

$$Kurtosis = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4}$$



2.1 Data Understanding – Summary Statistics

Para variáveis categóricas.

Contagem do número de ocorrências para cada nível. Questões abordadas:

- Os valores fazem sentido?
- Há valores em falta? Quantos?
- Como é que estão codificados?
- Quantos níveis existem? Existe apenas um valor? Existem mais de 50 ou 100 níveis?
- Qual é a moda dos níveis?

2.1 Data Understanding – Summary Statistics

Porque é que as summary statistics são importantes?

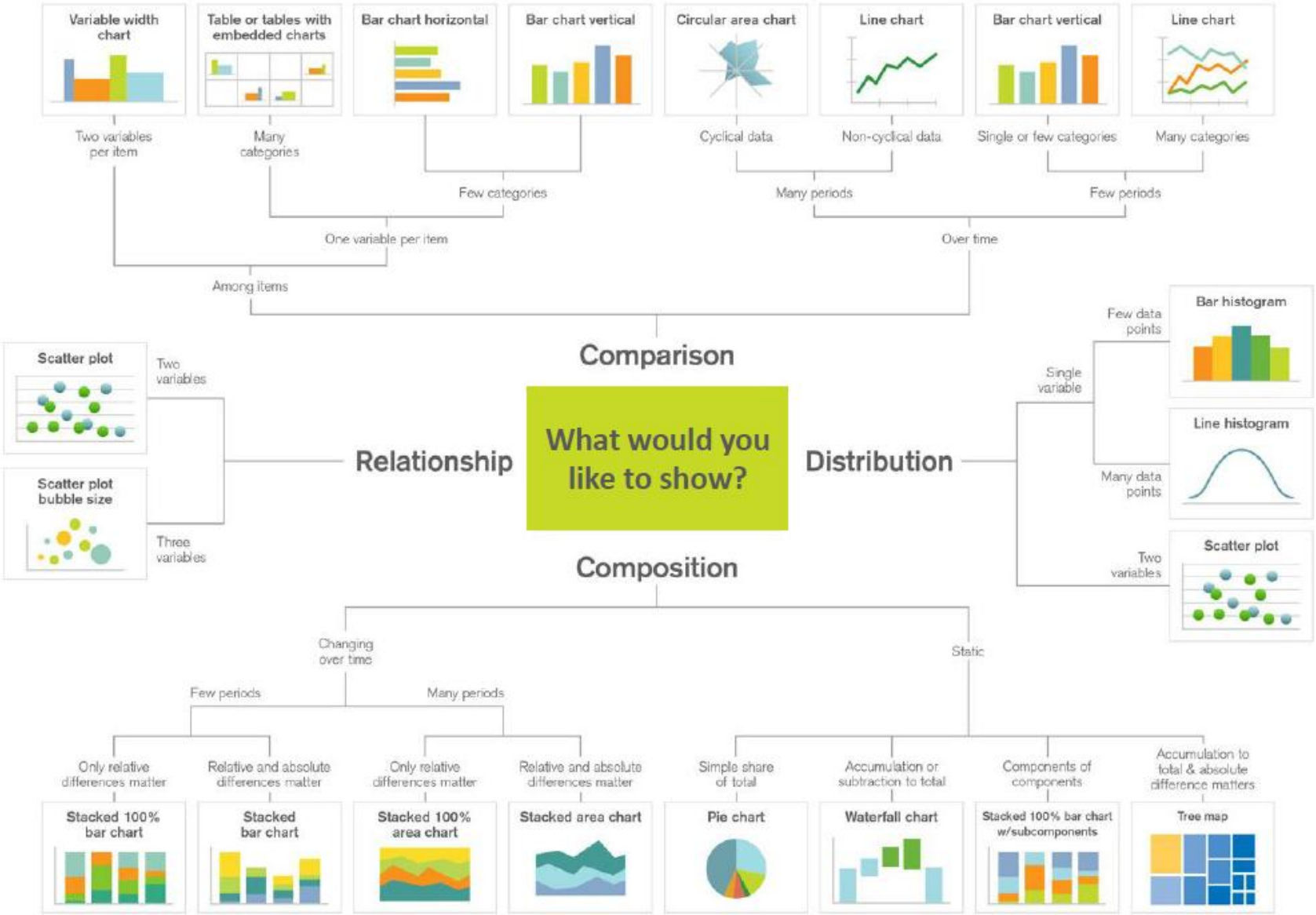
- Deteção de outliers: os outliers podem distorcer as estatísticas e afetar a precisão dos modelos. Examinando a distribuição de cada variável e a sua kurtosis, é mais fácil identificar outliers e determinar a forma de os tratar.
- Seleção do modelo: Diferentes tipos de modelos são mais adequados a diferentes distribuições de variáveis. Por exemplo: modelos de árvore de decisão podem ser mais adequados a distribuições não normais.
- Feature engineering: Explorar as variáveis pode ajudar a determinar como transformar as variáveis para melhorar o desempenho dos modelos. Por exemplo: uma variável enviesada, pode ser transformada utilizando uma transformação logarítmica.
- Processamento de dados: A verificação da distribuição das variáveis pode ajudar a identificar potenciais problemas de qualidade dos dados, tais como dados em falta, dados inconsistentes, etc.

2.2 Data Understanding – Visualizações

O processo de apresentação de dados em formato visual fornece insights que permitem uma mais rápida e fácil tomada de decisões.

Os seres humanos são muito eficientes a detetarpadrões visuais. Com componentes visuais, podemos:

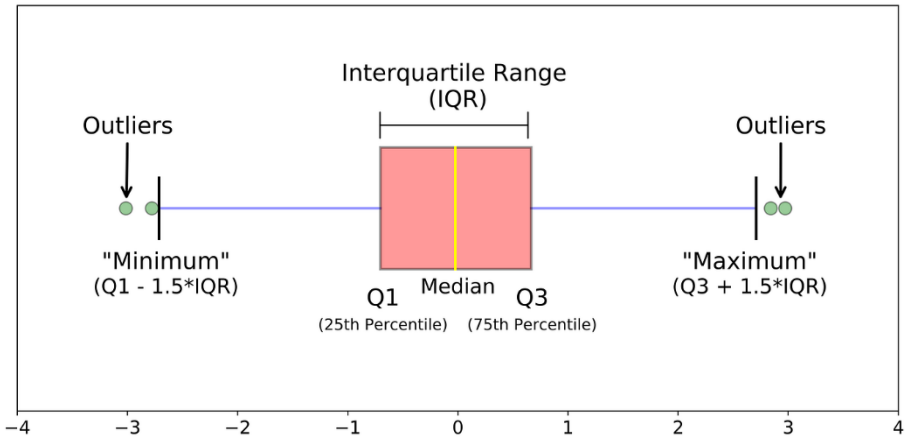
- Interpretar grandes quantidades de dados;
- Descobrir padrões interessantes;
- Explicar temas complexos;
- Transmitir informações rapidamente.


What would you like to show?
Comparison
Relationship
Distribution
Composition
Variable width chart
Table or tables with embedded charts
Bar chart horizontal
Bar chart vertical
Circular area chart
Line chart
Bar chart vertical
Line chart
Two variables per item
Many categories
Few categories
Cyclical data
Non-cyclical data
Single or few categories
Many categories
Many periods
Few periods
One variable per item
Over time
Among items
Scatter plot
Two variables
Scatter plot bubble size
Three variables
Few data points
Single variable
Many data points
Two variables
Bar histogram
Line histogram
Scatter plot
Changing over time
Static
Few periods
Many periods
Only relative differences matter
Relative and absolute differences matter
Only relative differences matter
Relative and absolute differences matter
Simple share of total
Accumulation or subtraction to total
Components of components
Accumulation to total & absolute difference matters
Stacked 100% bar chart
Stacked bar chart
Stacked 100% area chart
Stacked area chart
Pie chart
Waterfall chart
Stacked 100% bar chart w/subcomponents
Tree map

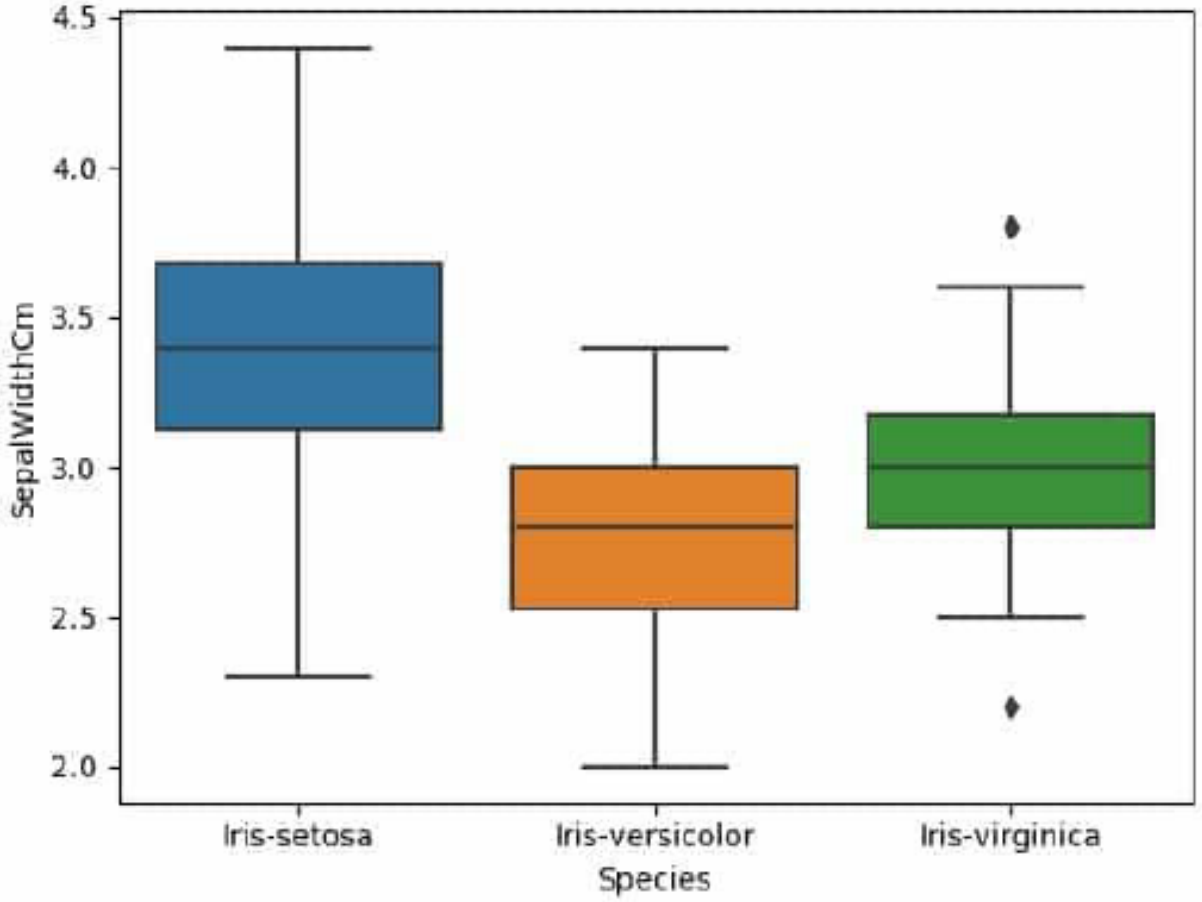
2.2 Data Understanding – Visualizações

Boxplots: Os boxplots visam responder às seguintes questões:

- Quais são os valores-chave, como a mediana, o percentil 25 e outros?
- Existem outliers? Quais são os seus valores?
- Os dados são simétricos?
- Qual é o grau de concentração dos dados?
- Os dados são enviesados? Em caso afirmativo, em que direção?



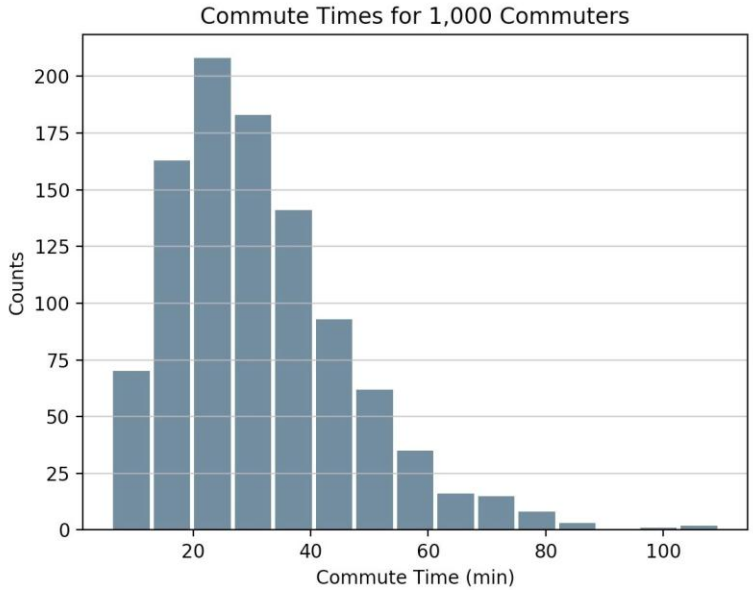
2.2 Data Understanding – Visualizações



2.2 Data Understanding – Visualizações

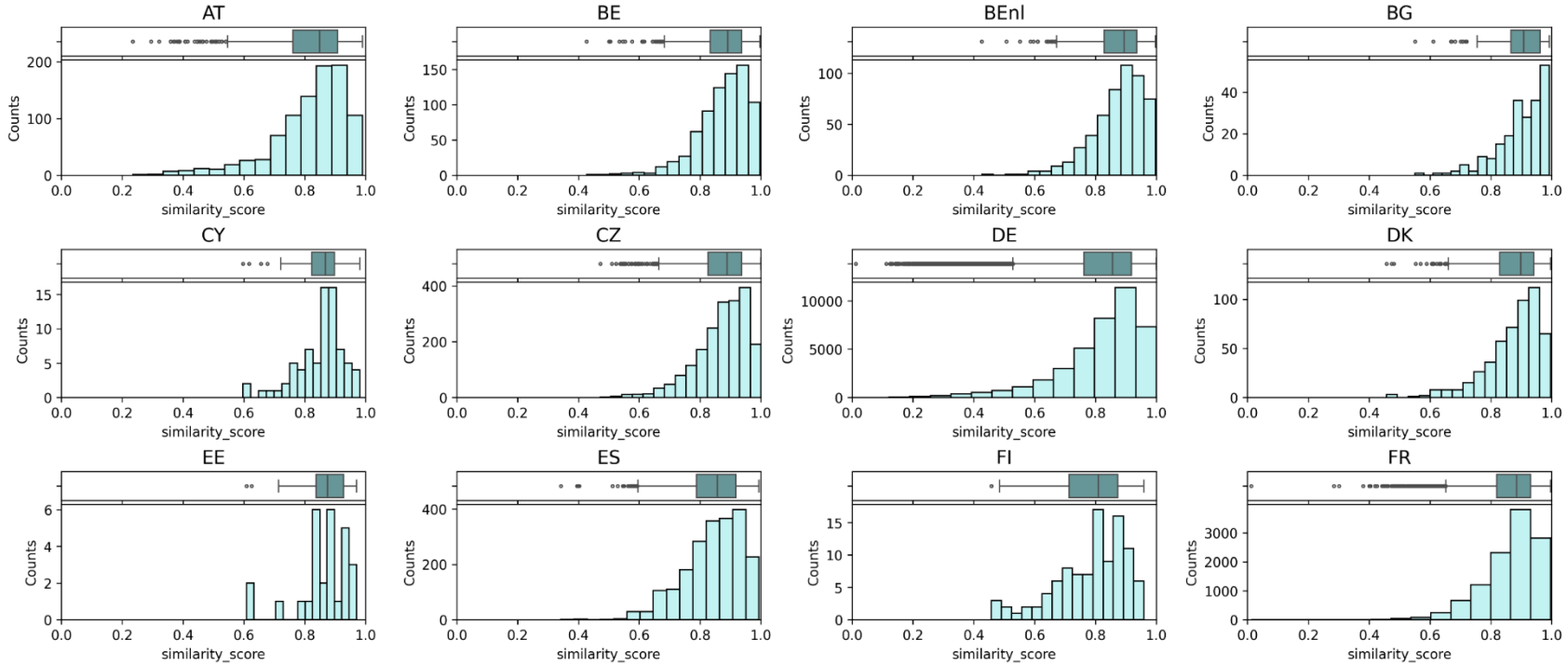
Histogramas: Dão uma estimativa da concentração de valores.

- Quais são os extremos e se existem lacunas ou outliers.
- Quanto maior o número de barras a apresentar, mais o gráico se aproxima de uma distrbuição.



2.2 Data Understanding – Visualizações

Boxplots + Histogramas

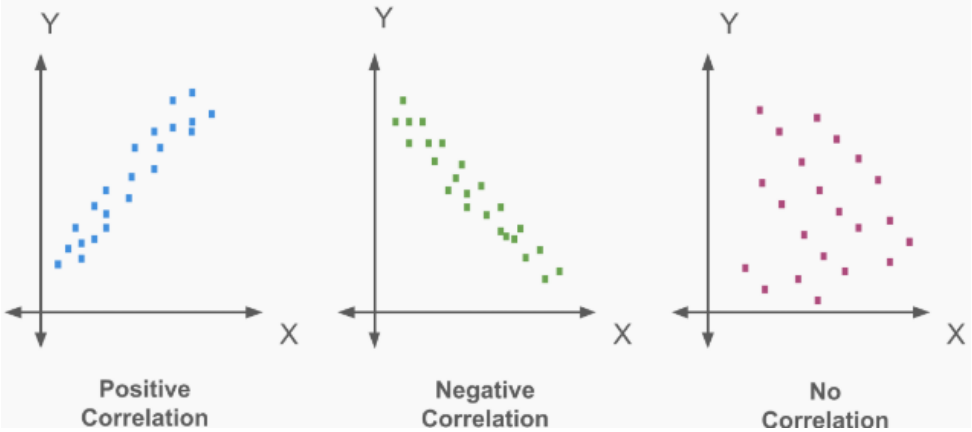


2.2 Data Understanding – Visualizações

Scatter Plots: ao apresentar uma variável em cada eixo, é possível detetar se existe uma relação entre as duas variáveis.

Vários tipos de correlação podem ser interpretados através dos padrões apresentados:

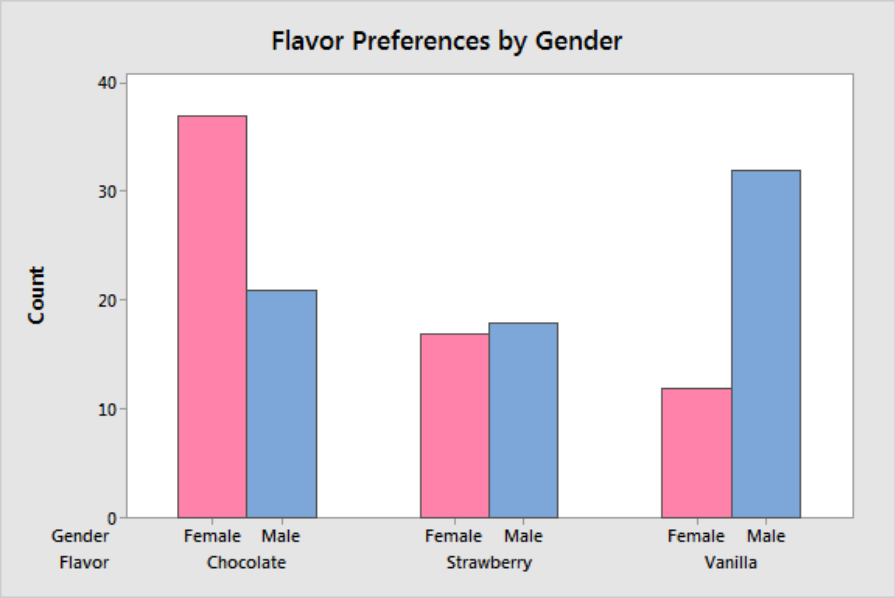
- Correlação positiva (os valores aumentam juntos)
- Correlação negativa (um valor diminui à medida que o outro aumenta)
- Correlação nula (sem correlação)



2.2 Data Understanding – Visualizações

Bar Plots: utiliza barras para representar categorias de dados, com o comprimento ou altura das barras proporcional aos seus valores.

- Compara categorias discretas, com um eixo para as categorias e o outro para os valores.



2.3 Data Understanding – Conclusão

Os principais objectivos da EDA são:

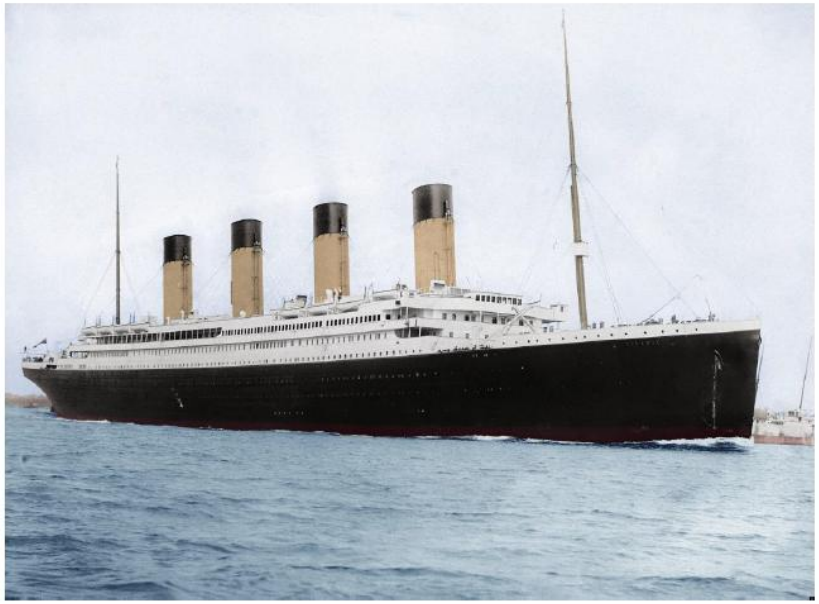
1. Compreender a distribuição e a variabilidade dos dados;
2. Identificar padrões, tendências e relações entre variáveis;
3. Verificar a existência de outliers, valores em falta e anomalias nos dados;
4. Identificar potenciais problemas com os dados, tais como erros ou enviesamentos;
5. Gerar ideias para uma análise mais aprofundada.

As técnicas comuns utilizadas na EDA incluem:

1. Histogramas e Box Plots para visualizar a distribuição e a variabilidade dos dados;
2. Scatter Plots e matrizes de correlação para examinar a relação entre variáveis;
3. Heatmaps e análise de clusters para identificar padrões e relações entre variáveis;
4. Estatísticas sumárias, como a média, a mediana e o desvio padrão, para descrever a tendência central e a variabilidade dos dados.

3. Exploratory Data Analysis - Exemplo

3. Exploratory Data Analysis - Exemplo



Variável	Definição
survived	Survival (Key : 0 = No, 1 = yes)
pclass	Ticket Class (Key : 1 = 1st , 2 = 2nd , 3 = 3rd)
name	Name of the Passenger
sex	Sex
age	Age in yeas
sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic
ticket	Ticket Number
fare	Passanger Fare
cabin	Cabin Number
embarked	Port of Embarkation (Key: C = Cherbourg, Q = Queenstown, S = Southampton)

3. Exploratory Data Analysis - Exemplo

- Neste caso, não estamos realmente a falar de "necessidades comerciais", mas temos um objetivo em mente:
- Prever quais passageiros sobreviveram ao Titanic e perceber quais os factores que levam a uma maior probabilidade de sobrevivência.

- Todos os nossos dados estão disponíveis num único ficheiro, pelo que não é necessária qualquer integração de dados.
- Precisamos agora de fazer a nossa Exploratory Data Analysis. Mas vamos primeiro olhar para as nossas variáveis.

3. Exploratory Data Analysis - Exemplo

pid	survived	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
1	0	3	BRAUND, MR. OWEN HARRIS	male	21	1	0	A/5 21171	7,25		S
2	1	1	CUMINGS, MRS. JOHN BRADLEY (FLORENCE BRIGGS THAYER)	female	38	1	0	PC 17599	71,2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	25	0	0	STON/O2. 3101282	7,925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53,1	C123	S
5	0	3	Allen, Mr. William Henry	male	34	0	0	373450	8,05		S

Target variable

Temos um conjunto de dados com 891 observações, e a coluna “survived” é a nossa variável dependente, que nos diz se um determinado passageiro sobreviveu ou não.

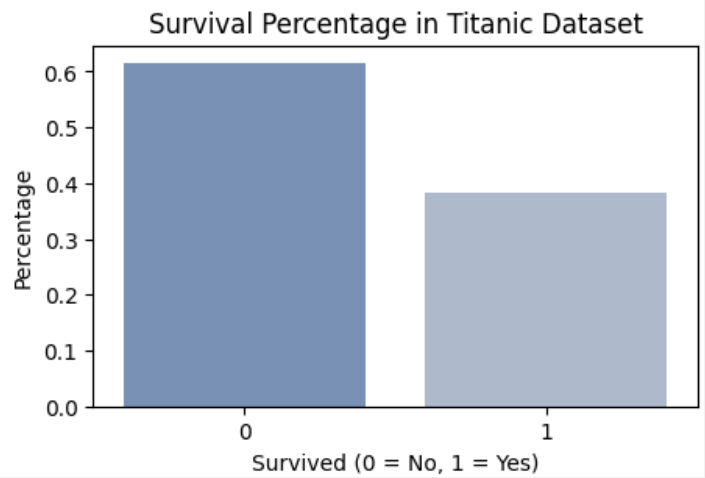
3. Exploratory Data Analysis - Exemplo

Qual o tipo de cada variável?

```
pid             int64
survived        int64
pclass          int64
name           object
sex            object
age             int32
sibsp           int64
parch           int64
ticket         object
fare          float64
cabin          object
embarked       object
dtype: object
```

Quantas observações temos para cada classe da variável target?

survived	
0	549
1	342



Nota: Unbalanced data

3. Exploratory Data Analysis - Exemplo

Explorar as variáveis numéricas:

	pid	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.745230	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	13.833487	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	21.000000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	79.000000	8.000000	6.000000	512.329200

Existem missing values?

As variáveis seguem uma distribuição normal?

Existem outliers?

Quantas pessoas sobreviveram?

3. Exploratory Data Analysis - Exemplo

Explorar as variáveis numéricas:

fares por pclass

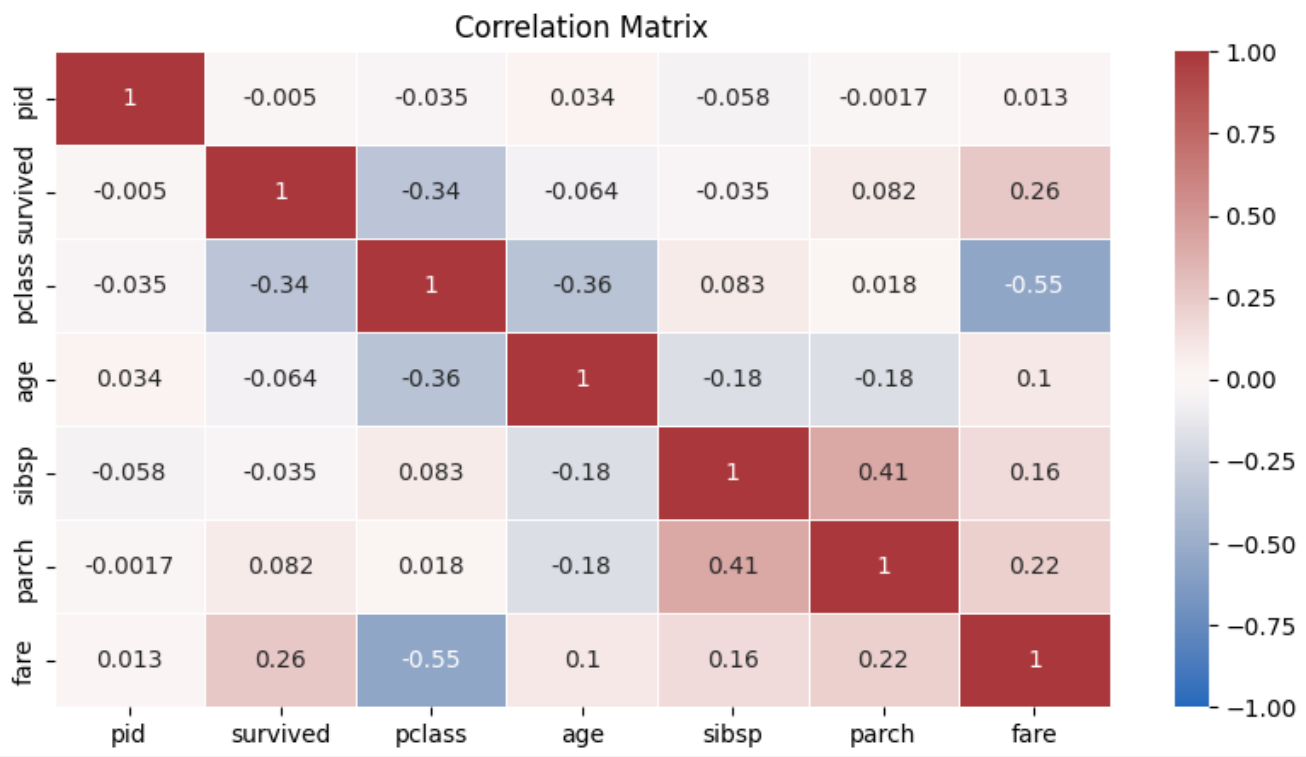
	count	mean	std	min	Q1	median	Q3	max
pclass								
1	216	84.154687	78.380373	0.0	30.92395	60.2875	93.5	512.3292
2	184	20.662183	13.417399	0.0	13.00000	14.2500	26.0	73.5000
3	491	13.675550	11.778142	0.0	7.75000	8.0500	15.5	69.5500

fares por survived

	count	mean	std	min	Q1	median	Q3	max
survived								
0	549	22.117887	31.388207	0.0	7.8542	10.5	26.0	263.0000
1	342	48.395408	66.596998	0.0	12.4750	26.0	57.0	512.3292

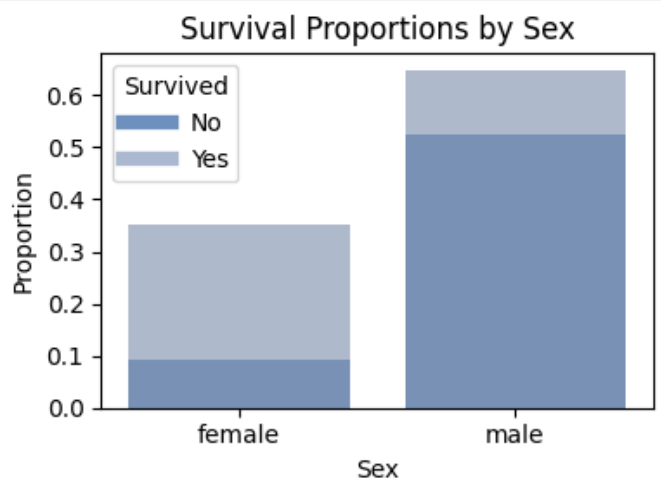
3. Exploratory Data Analysis - Exemplo

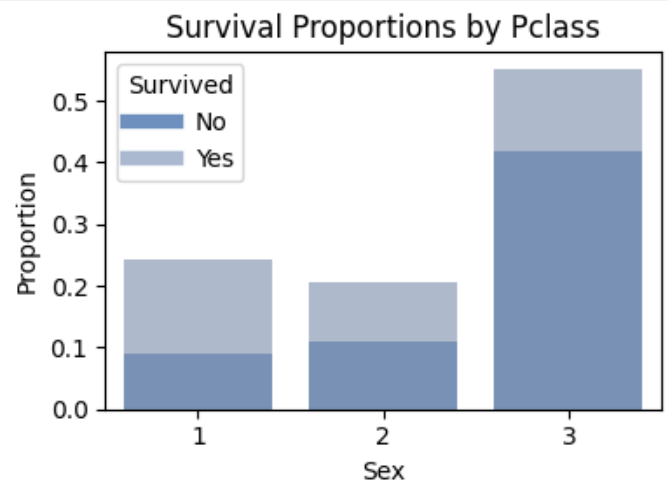
Explorar as variáveis numéricas:



3. Exploratory Data Analysis - Exemplo

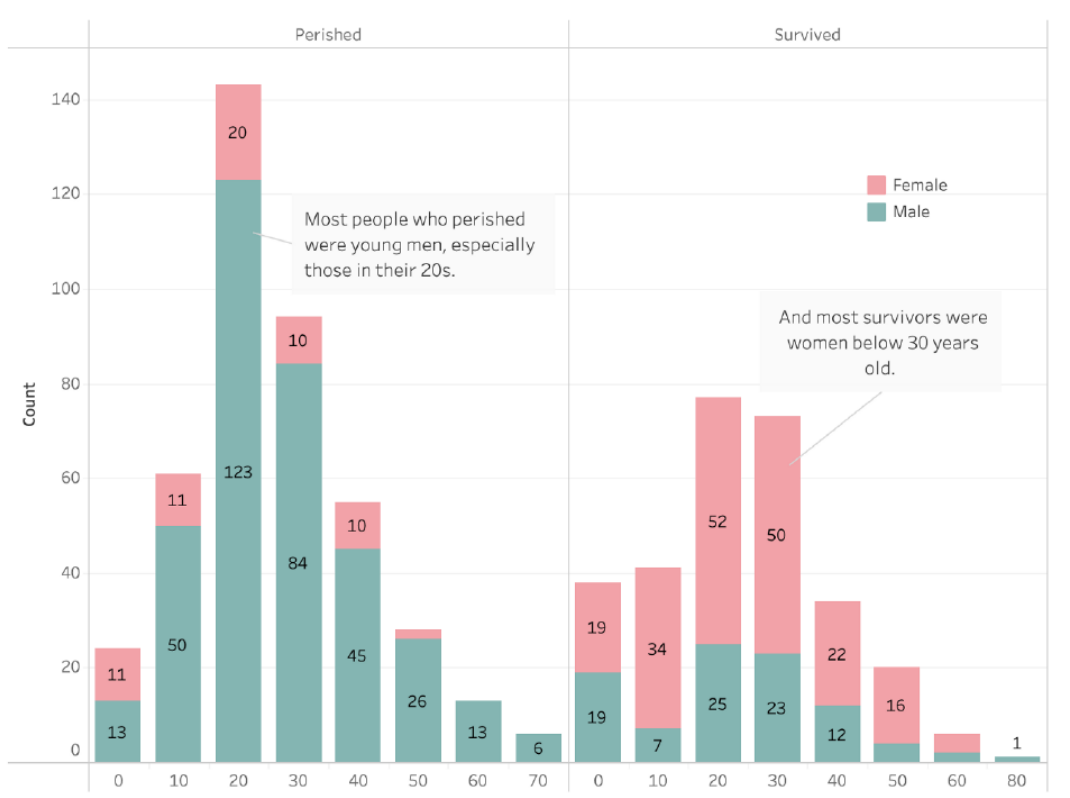
Explorar as variáveis categóricas:





3. Exploratory Data Analysis - Exemplo

Explorar as variáveis categóricas:



3. Exploratory Data Analysis - Exemplo

Insights obtidos através da EDA

- Dados a normalizar
- Outliers a remover
- Variáveis a criar
- Variáveis a remover
- Missing values
- …

Executar todos os passos necessários

DATA PREPARATION

- Selecionar e limpar os dados, resolvendo as issues detetadas.
- Constuir novos dados com base nos dados iniciais.
- Integrar e combinar dados de diferentes fontes.

Obrigado!