Aprendizagem Automática em Sistemas Empresariais

PEDRO PEREIRA

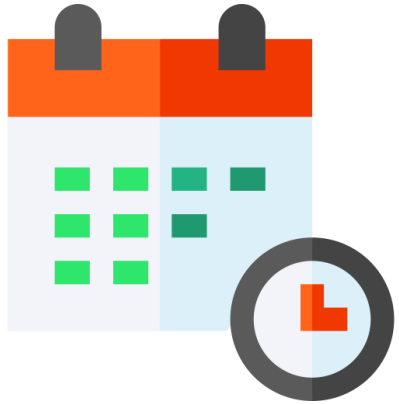
AULA 3

Agenda

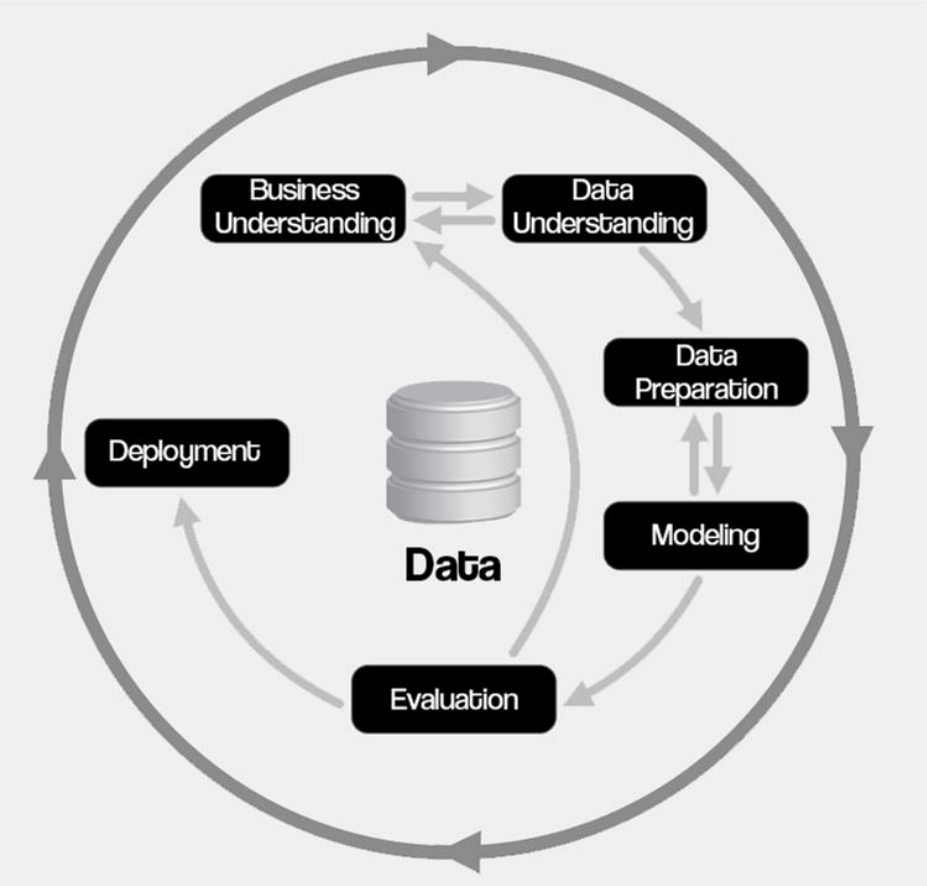
CRISP-DM: Compreensão dos Dados

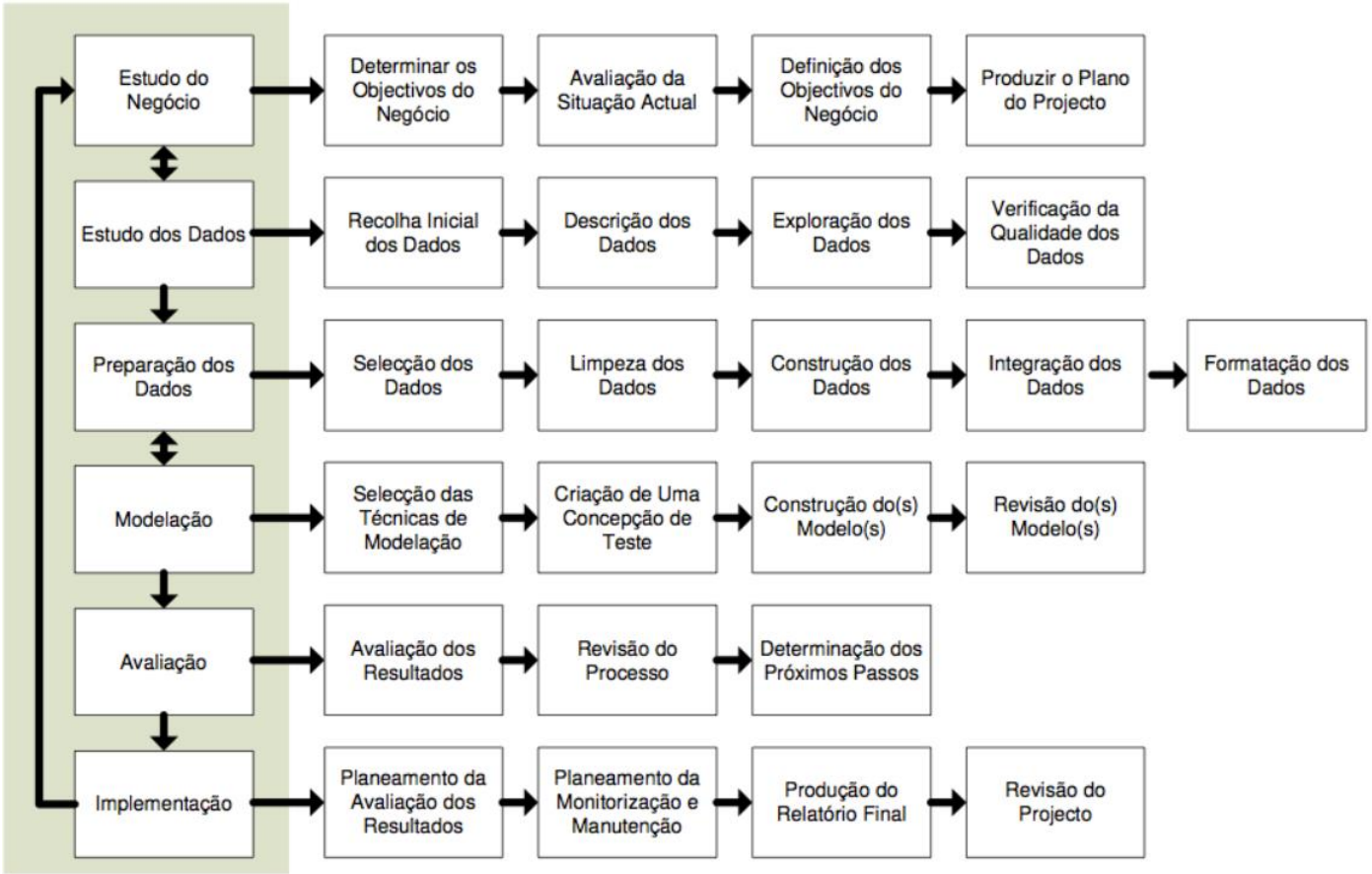
- Análise Exploratória dos Dados (EDA)
- Exemplo

Acompanhamento ao projeto

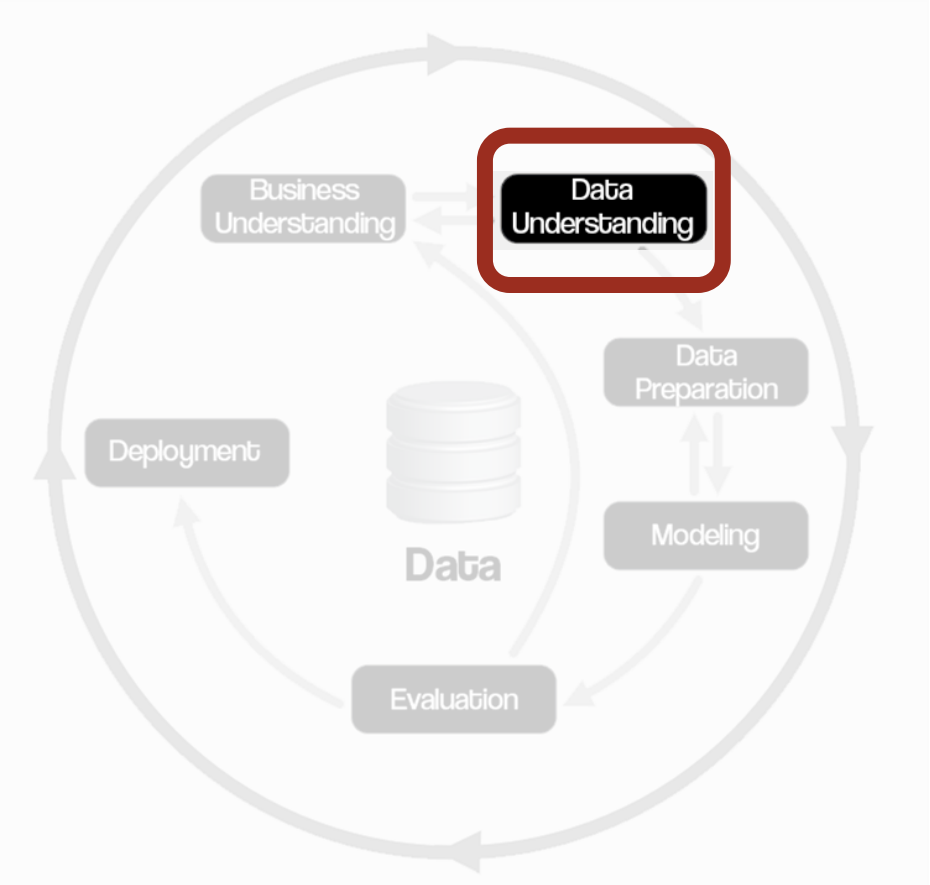


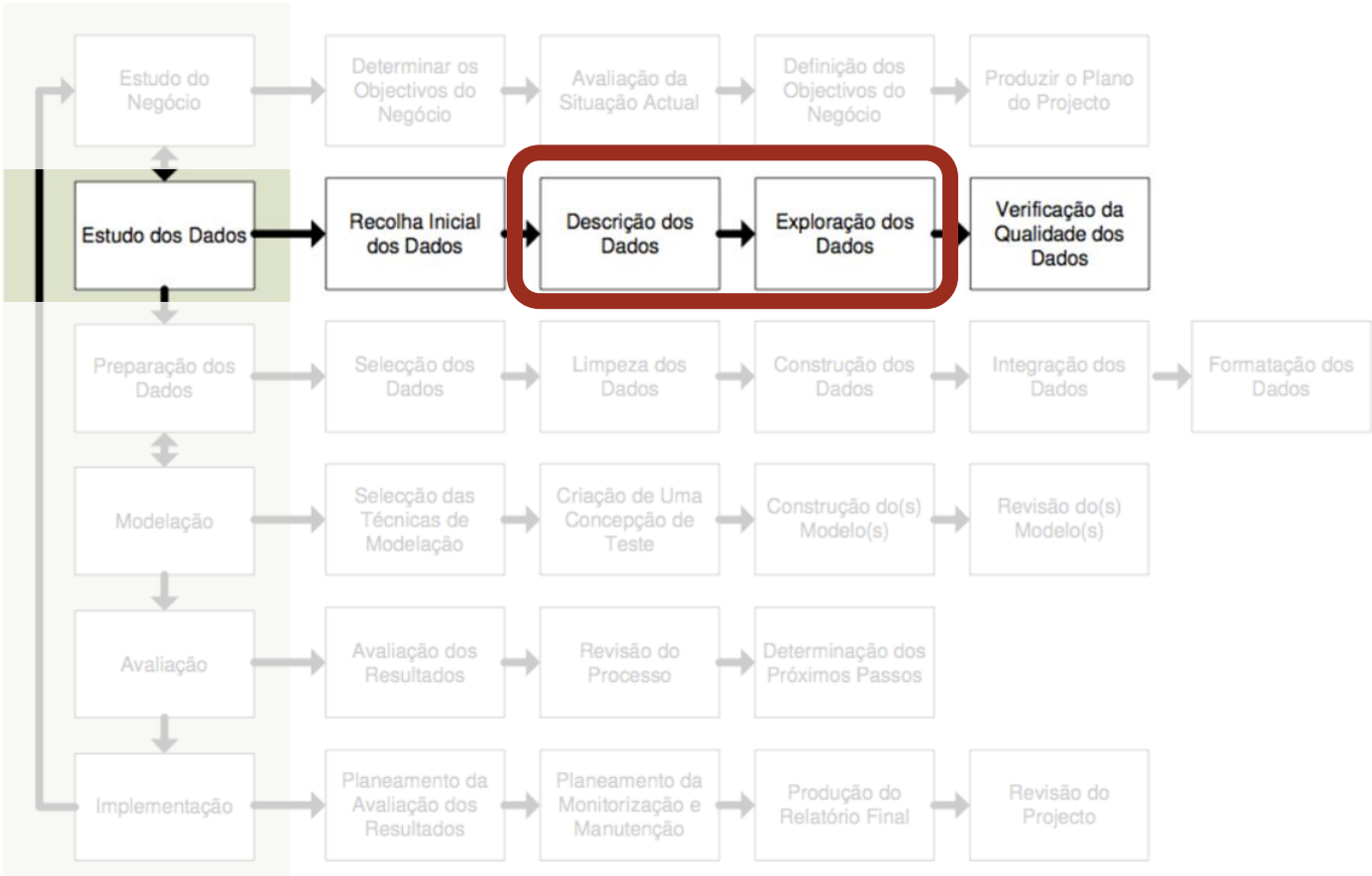
Cross Industry Process for Data Mining (CRISP-DM)



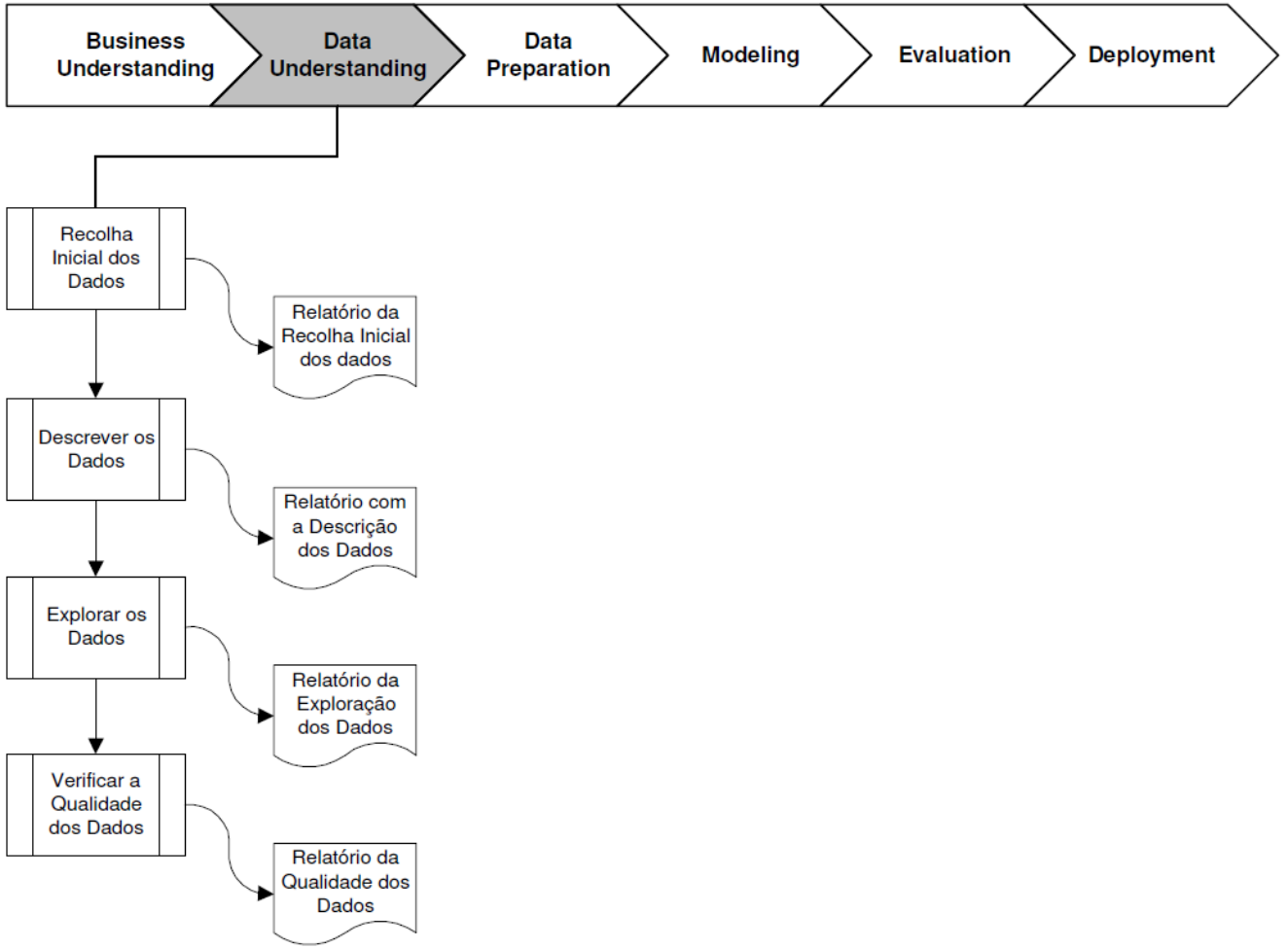


CRISP-DM – Compreensão dos Dados





CRISP-DM – Atividades da Compreensão de Dados



CRISP-DM – Compreensão de Dados

Familiarização com os dados e perceção da sua utilidade.

Processo de coleta de dados (ex.: via *query* a DW, *dowlnoad* .csv, etc.).

Descrição dos dados (ex.: atributos e significado, tipos de dados, nº de linhas, etc.).

Identificação de problemas na qualidade dos dados e possíveis soluções a implementar ao longo do projeto.

Tipos de Atributos

Data Mining é frequentemente aplicado a simples conjuntos de dados em formato **tabular**.

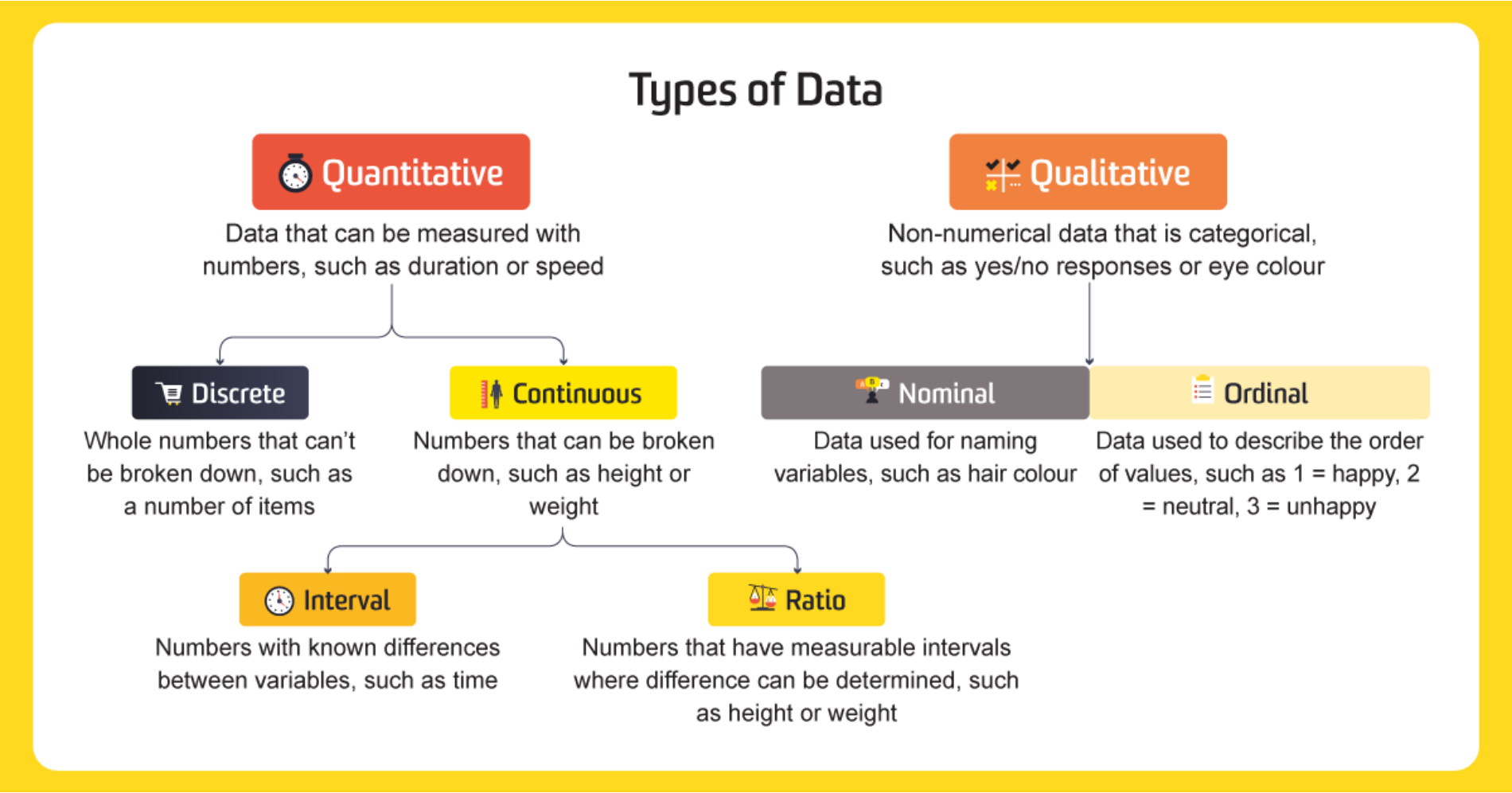
As tabelas contêm um conjunto de **instâncias** (exemplos, linhas) e **atributos** (colunas).

Em dados tabulares, os atributos são maioritariamente:

- **Categóricos:** valores qualitativos; discretos ou enumerados; “classes”. Podem ser:
  - **Nominais:** sem ordem nos valores possíveis (ex.: “cão”, “gato”, “pássaro”);
  - **Ordenados:** valores possíveis seguem uma determinada ordem (ex.: “Janeiro”, “Fevereiro”, “Março”, …).
- **Numéricos:** valores quantitativos; inteiros ou reais, com ou sem limites. Podem ser:
  - **Discretos:** opções finitas (ex.: dias do mês, tamanho do calçado, …);
  - **Contínuos:** opções infinitas (ex.: altura, peso, …).

O tipo de atributo vai estabelecer o modo como este é analisado e manipulado.

Tipos de Atributos



Tipos de Atributos

		Categóricos		Numéricos	
		Nominais	Ordenados	Intervalos	*Ratio*
Operações Lógicas / Matemáticas	÷ ×	✗	✗	✗	✓
	+ -	✗	✗	✓	✓
	> <	✗	✓	✓	✓
	= ≠	✓	✓	✓	✓
Exemplos (valores)		Género (M, F, Outro)	Opinião (concordo, neutro, discordo)	Latitude (de +90 a -90)	Idade (0 a 100)
Medida de tendência central		Moda	Mediana	Média aritmética	Média geométrica

Análise Exploratória de Dados (EDA)

"EDA está maioritariamente relacionada com a visualização e sumarização dos dados antes de se iniciar o processo de modelação.

Razões para a criação de gráficos na exploração de dados:

- **Compreender** as propriedades dos dados;
- **Inspecionar** atributos qualitativos ao invés de olhar para grandes tabelas de dados em bruto;
- **Descobrir** novos padrões ou associações;
- Entre outras…

Análise interativa é a melhor forma de explorar os dados." (Jeff Leck, 2015)

Análises de dados podem ser:

- **Univariadas:** apenas um atributo;
- **Multivariadas:** 2 ou mais atributos (mais comum).

Análise Exploratória de Dados (EDA)

EDA pode ser feito em diversas ferramentas.

Python: pandas-profiling, sweetviz, AutoViz,dtale, …

R: datavis, plotly, ggplot, …

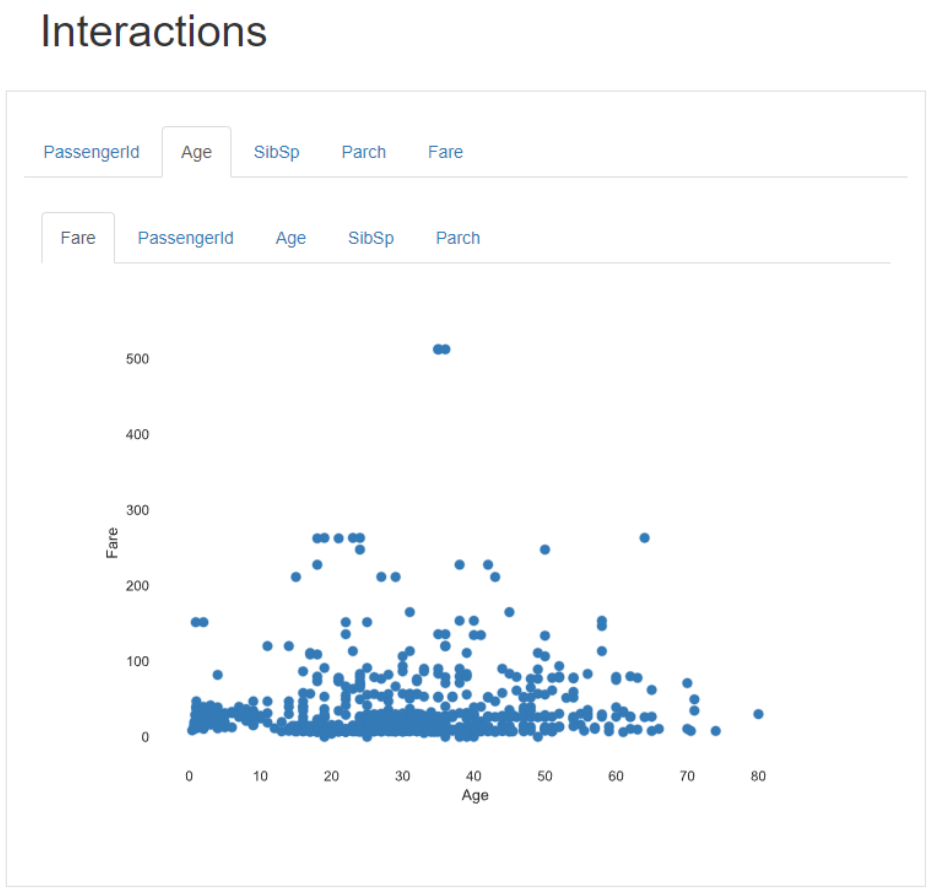
Weka.

Tableau.

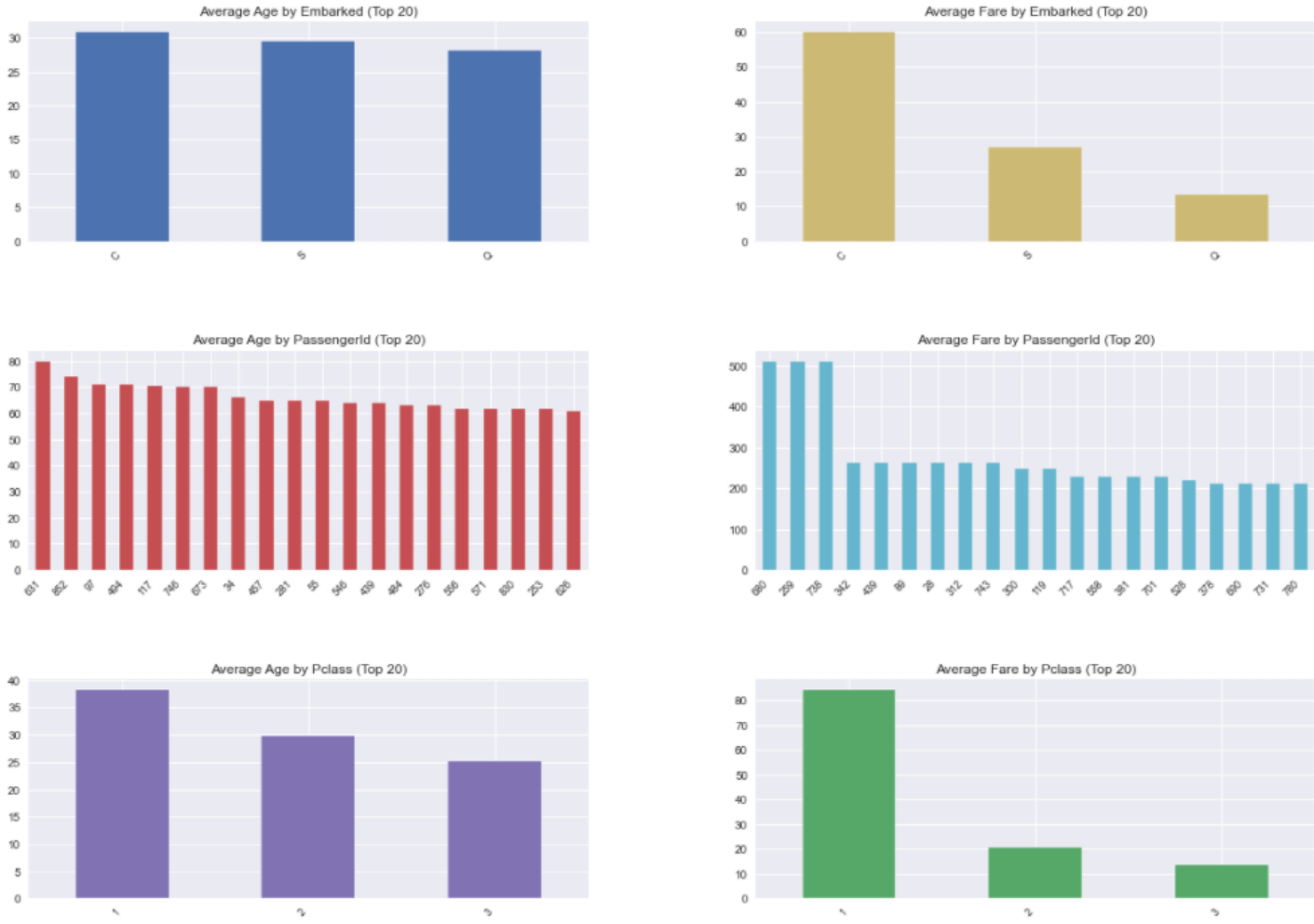
Excel.

…

EDA: Titanic Dataset



EDA com pandas-profiling



EDA com AutoViz

Aprendizagem Automática em Sistemas Empresariais

PEDRO PEREIRA

AULA 3