Exploratory Data Analysis

Zuil Pirola

UNIVERSIDADE
LUSÓFONA

Objetivo

Compreender o papel da EDA

Saber **explorar dados** antes da modelação

Identificar **problemas** comuns em **datasets** reais

Data Science e EDA

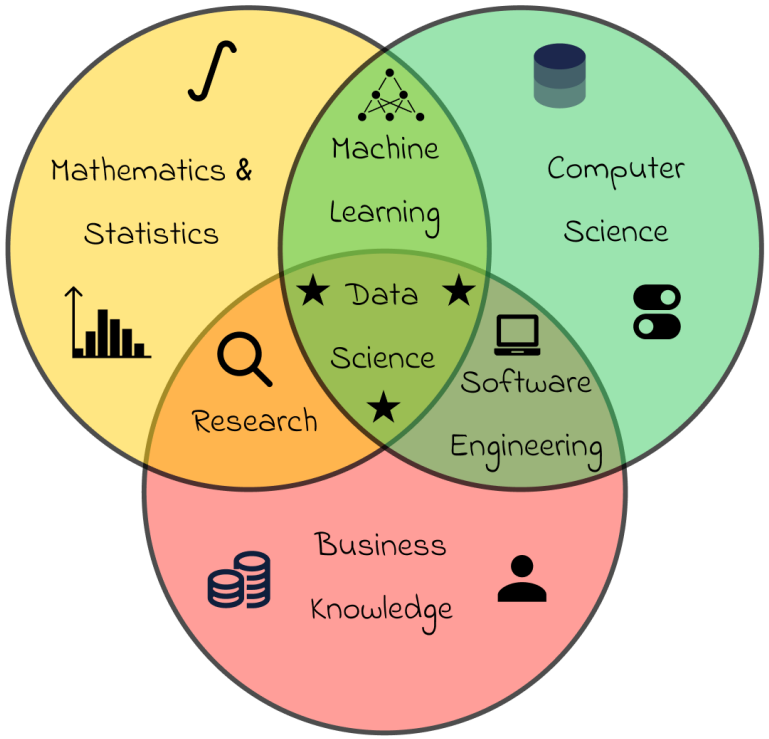
Area Interdisciplinar

- **Estatística**
- Computação
- Conhecimento do domínio

Objetivo:

- **Extrair conhecimento** dos dados
- Apoiar decisões

Data Science e EDA



https://towardsdatascience.com/roadmap-to-becoming-a-data-scientist-part-1-maths-2dc9beb69b27/

Data Science e EDA

EDA é uma das **primeiras etapas**.

Exemplo:
Netflix usa Data Science para:

- Recomendações
- Previsão churn
- Otimização conteúdo

O que é EDA?

Processo de:

- Resumir **estatisticamente**
- Explorar dados **sem hipóteses rígidas**
- **Visualizar** padrões
- Identificar **inconsistências**

John Tukey (1977):

> "Exploratory data analysis is **detective** work."

O que é EDA?

1. **Compreender** estrutura dos dados
2. **Identificar problemas** de qualidade
3. Detectar **padrões** iniciais
4. Gerar hipóteses analíticas

Não é ainda Machine Learning.

Exemplo:
Antes de prever vendas:

- Há sazonalidade?
- Existem valores errados?

Por que EDA?

Dados reais frequentemente têm:

- **Missing values**
- **Outliers**
- Erros de **recolha**
- Viés **amostral**

Sem EDA:

- Modelos aprendem **erros**
- Decisões podem **falhar**

"Garbage in, garbage out."

O que é CDA?

Processo analítico focado em **testar hipóteses** específicas com base em modelos estatísticos formais.

Características principais:

- Parte de uma **hipótese definida previamente**
- Utiliza testes estatísticos **rigorosos**
- Procura **validar** ou **rejeitar** teorias
- Resultados com maior formalidade inferencial

EDA x CDA

	EDA	**CDA**
Objetivo	Explorar dados e descobrir padrões	Testar hipóteses específicas
Métodos	Visualização, estatísticas descritivas	Testes estatísticos formais
Flexibilidade	Alta (abordagem aberta)	Mais estruturada
Sobre a Hipótese	Gera a Hipótese	Testa a hipótese
Exemplo típico	Explorar dataset clientes	Testar se campanha aumentou vendas

Passos para uma EDA

1. Definir **perguntas** de investigação

O que queremos **compreender** nos dados?

Exemplos:

- Clientes mais rentáveis?
- Fatores que influenciam desempenho académico?

Passos para uma EDA

2. **Preparação e reestruturação** dos dados

- **Criar** novas **variáveis** relevantes (feature engineering inicial)
- Converter unidades ou formatos

Exemplo:

Data → idade

Receita total → receita mensal

Passos para uma EDA

3. Criar **métricas** mais informativas

- Taxas, rácios ou percentagens podem ser mais **úteis**

Exemplo:

- Taxa conversão = vendas / visitas
- Taxa sucesso alunos = aprovados / inscritos

Frequentemente revelam **padrões escondidos**.

Passos para uma EDA

4. Codificação de variáveis categóricas

- Criar variáveis dummy (**one-hot encoding**)
- Facilita análise estatística e **modelação**

Exemplo:

Género → {Masculino=0, Feminino=1}

Passos para uma EDA

5. **Visualização** e estatísticas descritivas

- **Histogramas**, **boxplots**, **scatterplots**
- Média, mediana, variância, quartis
- Primeira visão dos dados

Objetivo: **compreender** antes de modelar.

Passos para uma EDA

6. **Compreender** estrutura e **relações** nos dados

- Identificar padrões e correlações
- Detectar **outliers** ou comportamentos inesperados
- Explorar dependências entre variáveis

7. Identificar fatores confundidores e relações complexas

- Variáveis que influenciam outras simultaneamente
- Interações entre variáveis
- Possível **multicolinearidade**

Importante para análises posteriores.

Passos para uma EDA

8. Tratar **dados em falta**

Possíveis abordagens:

- Remover observações
- **Imputar** valores (média, mediana, modelo)
- Analisar padrão de **missingness**

9. Avaliar necessidade de transformações

- Logaritmos, **normalização**, scaling
- Ajustes para **melhorar interpretação** ou modelação

Passos para uma EDA

10. Da **exploração** à **confirmação**

- Formular hipóteses com base na EAD
- Aplicar Análise Confirmatória (CAD):
  - Testes estatísticos
  - Modelação inferencial

Finalmente:
Comunicar conclusões de forma clara e visual.

EDA no pipeline

Pipeline típico Data Science

1. Definição problema
2. Recolha dados
3. **EDA (compreensão)**
4. Preparação dados
5. Modelação
6. Avaliação
7. Deploy

EDA é iterativa

- Exploras dados
- Descobres problemas
- Voltas atrás para limpar dados
- Exploras novamente

Processo **cíclico**.

Exemplo:
Detetar outliers → limpar → **reavaliar** distribuição.

Sobre avaliação

- 3 mini-Quizzes - 40%
- 1 Group Project - 60%

Or

- Exam - 100%

Quiz e exames podem conter questões práticas