



Big data e ciência de dados

aula 03

Luisa Mariele Strauss

The background of the slide features a dark blue, abstract graphic. It includes a line graph with several data points connected by lines. One data point is highlighted with a yellow circle and labeled with the number '289.33'. The overall aesthetic is technical and data-oriented.

Agenda

- Análise descritiva e exploratória
- Tipos de dados
- Ferramentas: Power BI, Excel, Python
- Cuidados e análises para “entender” a base de dados



Exploratória

- Explorar o conjunto de dados – compreensão do próprio analista
- Apresentar a análise exploratória/descritiva para audiência
- Encontrar padrões e explicações interessantes – explanatória

Knafllic (2018)

Descritiva

- O que está acontecendo/aconteceu?
- Análise do passado
- Informações instantâneas
- Uso de estatísticas “básicas”
- Gráficos, tabelas, infográficos etc
- Exemplos?



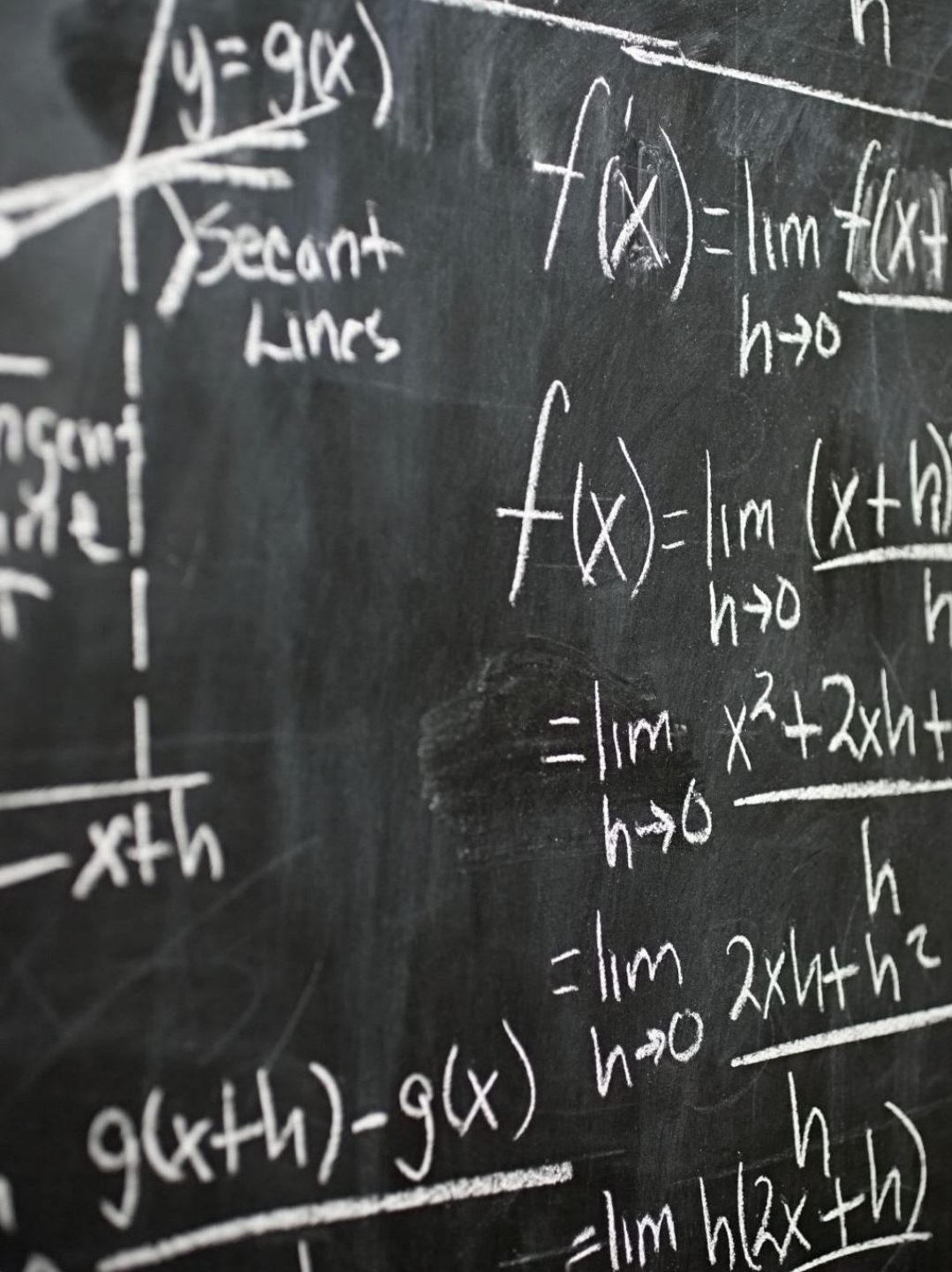
Tipos de dados/variáveis

Quantitativas

- Contínuas – números reais
- Discretas – números inteiros

Qualitativas

- Nominais ou Categóricas – caracterizam o dado
- Ordinais – caracterizam e categorizam com uma ordem



Base Student Data

Student Performance Data Set

Contém dados sobre disciplina de matemática numa escola secundária

- school ID
- gender
- age
- Address (rural/urban)
- Mother education (Medu)
- Study time
- Father education (Fedu)
- Tempo dedicado aos estudos
- Paid
- Internet
- Absences
- Grades (G1, G2 e G3)

Power BI desktop

- Gratuito
- Plataforma líder segundo o Gartner
- Para baixar: [Fazer o download do Power BI | Microsoft Power Platform](#) (link também no Moodle)
- Baixar o executável e seguir os passos
- Assim que instalar, vai abrir uma página de cadastro na Microsoft – não é obrigatório, basta fechar e estará pronto para usar



Fonte da imagem:

<https://exceleratorbi.com.au/extract-numerical-data-points-from-an-image/>

Python

- Google Colab
 - Serviço do Google que hospeda o Jupyter Notebook
 - Basta ter uma conta no Google
 - colab.google
- Jupyter Notebook



Instalando Jupyter Notebook a partir da distribuição Anaconda

- Acessar site <https://www.anaconda.com/> e a opção Free Download
- Fazer o login de acordo com sua preferência
- Escolher o pacote. Sugerido: Distribution
- Baixar o arquivo, executar, seguir as orientações de instalação e as seguintes opções de instalação:



- No prompt de comando, chamar “jupyter notebook”. Isso vai instalar os pacotes e abrir o navegador, com todas as pastas do seu computador OU Abrir no Anaconda Navigator e clicar na opção Jupyter Notebook Launch
- Pronto, é navegar nos arquivos e, para criar um novo notebook, clicar em New > Python

Python Bibliotecas

NumPy: para estruturas de dados básicos

Pandas: oferece estruturas de dados de alto

Seaborn: visualização de dados (gráficos) Python baseada no matplotlib

Matplotlib: biblioteca de plotagem 2D do Python que produz números de qualidade de publicação em vários formatos de cópia impressa e ambientes interativos entre plataformas

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Python - estruturas


- Série: array unidimensional, semelhante a uma lista em Python, no entanto criado sobre o numpy. Além da velocidade de processamento, a principal característica que o difere de uma lista comum é que seus índices podem ser mutáveis.

```
my_series = pd.Series([10,20,30,40,50])  
print(my_series)  
print(my_series[2])  
my_series.index = ['A', 'B', 'C', 'D', 'E']  
print(my_series)
```

```
0    10  
1    20  
2    30  
3    40  
4    50  
dtype: int64  
30  
A     10  
B     20  
C     30  
D     40  
E     50  
dtype: int64
```


Python - estruturas

- Dataframe é uma estrutura de dados tabular bidimensional e mutável em tamanho, potencialmente heterogênea, com eixos rotulados (linhas e colunas).



The screenshot shows a Jupyter Notebook interface. The top bar contains navigation icons: up, down, link, chat, settings, and a menu. Below the bar, a code cell contains the following Python code:

```
data = pd.read_csv("nba.csv")  
data.head()
```

The output of the code is a DataFrame with 10 columns: Name, Team, Number, Position, Age, Height, Weight, College, and Salary. The first five rows of data are displayed, showing players from the Boston Celtics.

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
2	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0

Mão na massa!

Cada aluno (ou dupla) vai escolher uma base de dados (sugestão: Kaggle), com tipos de dados heterogêneos (numérico, ordinal, categórico).

Usando a ferramenta de sua escolha, Power BI ou Python, realizar a análise descritiva, com as principais estatísticas (contagem, distinct, min, max, média, desvio-padrão, mediana etc).

Deve ser entregue um documento em PDF contendo:

- Breve descrição sobre a base escolhida
- Identificação do tipo de cada variável (contínua, discreta, ordinal, categórica)
- Ajuste do tipo da variável se necessário
- Uma interpretação sobre as estatísticas descritivas
- Identificação de dados com problemas, faltantes e alternativas de solução (basta citar as alternativas, não precisa implementar)
- Entrega até as 19h30 da próxima semana – aproveitem o restante da aula para exercitar!

Principais referências

- Tutoriais do Power BI
- Documentação Python
- Conhecimento e experiência em estatística (faculdade + trabalho)
- Material de aula prof. Renato Carlson
- Material de aula prof. Felipe de Moraes
- LOPES, G.R. et al. Introdução à Análise Exploratória de Dados com Python. Conference Paper, 2019.