

## Data Science – Princípios e Técnicas

Agosto  
2024



# Data Science



Onde me encontrar:

<https://www.linkedin.com/in/marco-mialaret-junior/>

e

<https://github.com/MatmJr>

# Ciência de Dados, finalmente!

# Data Science

---

**“Um cientista de dados é alguém que sabe mais sobre estatística do que um cientista da computação e mais sobre ciência da computação do que um estatístico”**

# Data Science

---

A ciência de dados é uma área extremamente variada. Ela engloba desde profissionais que se assemelham à estatísticos até aqueles que são praticamente engenheiros de software. Alguns dominam o aprendizado de máquina, enquanto outros possuem conhecimento limitado sobre o tema.

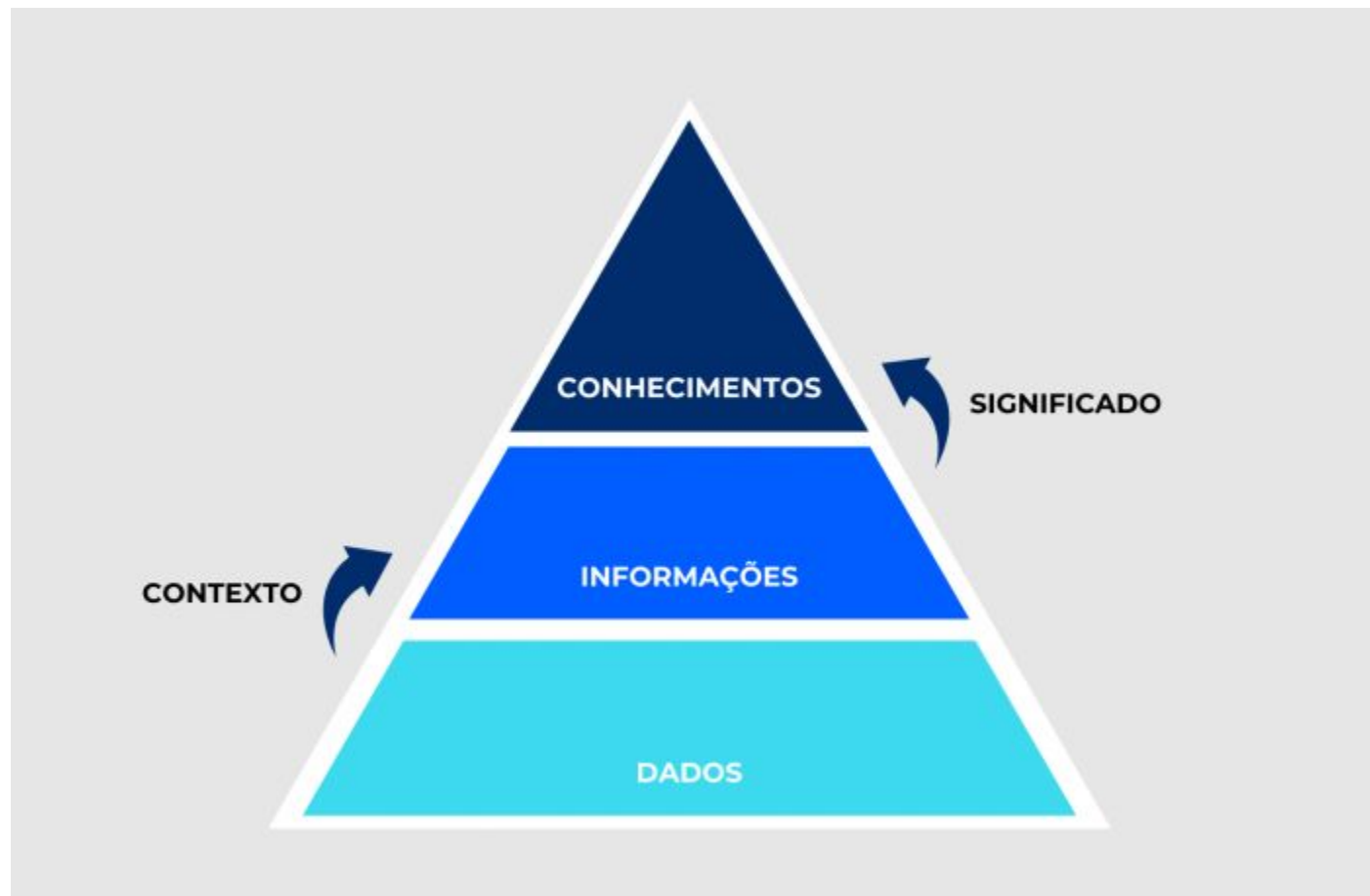
# Data Science

---

Apesar da diversidade na ciência de dados, podemos tentar definir um cientista de dados como alguém que extrai conhecimento a partir de dados desorganizados.

# Data Science

---



# Cientista x Analista x Eng. de Dados



# Data Science



---

Embora possam parecer similares, essas Três áreas têm sim uma diferença entre elas.

A **Ciência de Dados** abrange muitos modelos e métodos científicos, matemáticos e estatísticos, além de ferramentas para analisar e manipular dados.

# Data Science



---

A **Análise de Dados** é mais direcionada e focada em objetivos específicos. Em vez de buscar conexões gerais entre os dados, essa área trabalha com metas diretas e definidas, organizando os dados de maneira que permita extrair informações relevantes para alcançar esses objetivos.

# Data Science

---

Além disso, a análise de dados frequentemente envolve a criação de visualizações que ajudam a comunicar os insights de forma clara e impactante, facilitando a tomada de decisões estratégicas para o sucesso da empresa. Pode ser uma boa porta de entrada na área.

# Data Science



---

A **Engenharia de Dados** é responsável por projetar, construir e manter a infraestrutura necessária para a coleta, preparação e organização dos dados. Esse trabalho cria a base sólida sobre a qual a Ciência de Dados pode atuar, permitindo que os dados sejam analisados de maneira eficiente e precisa para gerar insights valiosos.

# E o mercado de trabalho?

# Data Science

---

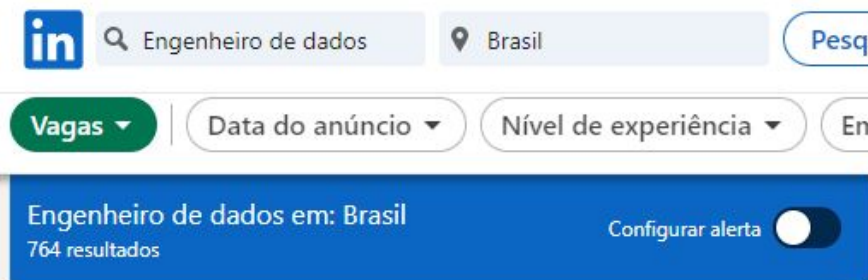
Segundo a IDC, cerca de 2,5 quintilhões de dados são gerados diariamente. O grande desafio para as empresas é converter essa imensa quantidade de dados brutos em informações valiosas e insights úteis para alcançar objetivos específicos, papel fundamental desempenhado pela Ciência de Dados.

# Data Science

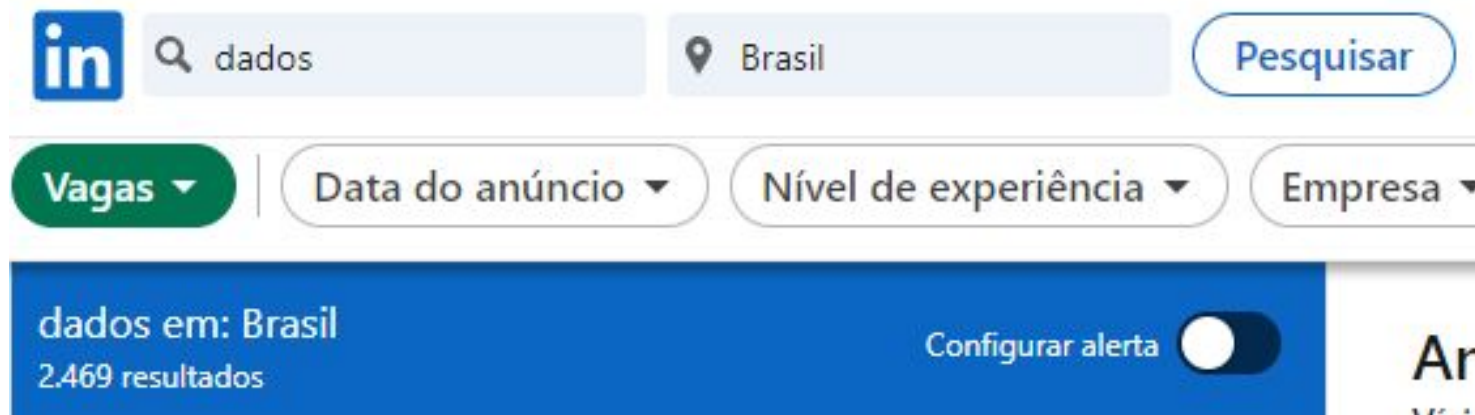
---

Este cenário suscita um interesse crescente em entender o mercado de Ciência de Dados e avaliar o potencial de investimento nessa carreira.

# Data Science



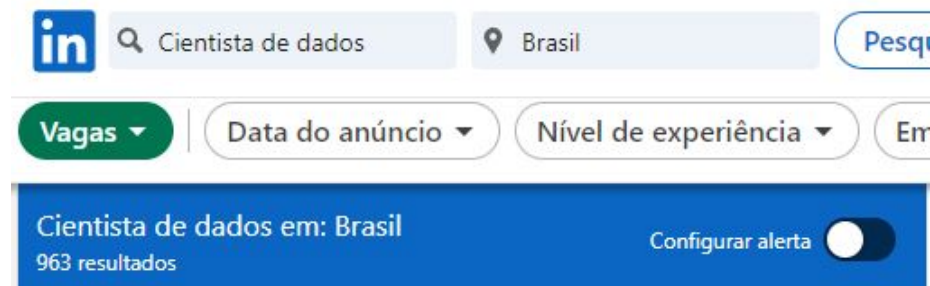
LinkedIn search results for "Engenheiro de dados" in Brazil. The search bar shows "Engenheiro de dados" and "Brasil". The results bar displays "Engenheiro de dados em: Brasil" with "764 resultados" and a "Configurar alerta" toggle switch.



LinkedIn search results for "dados" in Brazil. The search bar shows "dados" and "Brasil". The results bar displays "dados em: Brasil" with "2.469 resultados" and a "Configurar alerta" toggle switch.



LinkedIn search results for "Big Data" in Brazil. The results bar displays "Big Data em: Brasil" with "1.492 resultados" and a "Configurar alerta" toggle switch.



LinkedIn search results for "Cientista de dados" in Brazil. The search bar shows "Cientista de dados" and "Brasil". The results bar displays "Cientista de dados em: Brasil" with "963 resultados" and a "Configurar alerta" toggle switch.



# Definições importantes

# Data Science

---

## **Dados:**

Coleções de medições, características ou fatos sobre um grupo.

## **Ciência de dados:**

O processo de extrair significado dos dados.

# Data Science



---

## Registros:

são o local onde as informações individuais são armazenadas. Cada registro consiste em um ou mais campos. Em um banco de dados é uma linha de uma tabela que armazena informações individuais. Os campos correspondem às colunas da tabela.

## Variáveis:

Características particulares, medições ou fatos que compõem um registro.

# Data Science

---

## Dados quantitativos:

Dados numéricos. Dados quantitativos podem ser discretos (como o número de pessoas em uma família) ou contínuos (como a média de notas).

# Data Science

---

## **Exemplos de variáveis quantitativas:**

- altura
- peso
- renda anual
- GPA da faculdade

# Data Science



## Dados qualitativos:

Variáveis qualitativas são geralmente categorias não numéricas às quais os dados podem pertencer (como cor de cabelo). Algumas categorias podem ter uma ordem associada a elas, mas a ordem não implica uma natureza numérica para as categorias. Por exemplo, uma pergunta de pesquisa pode ter respostas que variam de Discordo Fortemente a Concordo Fortemente.

# Data Science

---

## **Exemplos de variáveis qualitativas:**

- cor do cabelo (loiro, castanho, preto, vermelho, cinza, ...)
- status atual da precipitação (sem precipitação, chovendo, nevando, granizando, ...)
- tipo de carro (sedan, coupé, SUV, minivan, ...)

# Data Science

---

## Pergunta de pesquisa:

Uma pergunta que pode ser respondida usando pesquisa, incluindo coleta de dados e análise.



# Data Science

---

## Por exemplo:

- Mais educação se traduz em mais riqueza?
- O clima está mudando?
- Quão rápido o coronavírus COVID-19 estava se espalhando quando se tornou prevalente no Brasil dezembro de 2020?

# O conjunto de Dados

# Data Science

---

O engenheiro de dados da sua empresa forneceu acesso a dois conjuntos de dados pré-processados. Agora, cabe a você analisá-los cuidadosamente para identificar quais escolas utilizaram seus recursos de forma mais eficiente na preparação de estudantes que se destacaram nas Olimpíadas de Redação e de Matemática.

# Data Science

---

Primeiros passos:

- Criar um ambiente virtual;
- Instalar o pandas;
- Selecionar o ambiente virtual no Jupyter.

# Data Science

---

<https://drive.google.com/uc?id=1Jgto7psHaMRTAVzcFt7D6SgJiHMB7uGT>

```
import pandas as pd

url = "https://drive.google.com/uc?id=1Jgto7psHaMRTAVzcFt7D6SgJiHMB7uGT"

escolas = pd.read_csv(url)
```

# Data Science



---

Perguntas motivadoras:

- Quais são os tipos de variáveis presentes no conjunto de dados?
- Existem dados faltantes?
- Existem dados duplicados?

# Data Science

---

- Quais são os tipos de variáveis presentes no conjunto de dados?

```
escolas.info()
```

# Data Science

---

Verificação de dados faltantes:

```
escolas.isnull().sum()
```

```
escolas.isna().sum()
```



# Data Science

---

Verificação de dados duplicados:

```
escolas.duplicated().sum()
```

# Data Science

Para encontrar uma descrição dos dados usamos o describe:

```
escolas.describe()
```

✓ 0.0s

	ID_Escola	Numero_Alunos	Orcamento_Anual
count	15.000000	15.000000	1.500000e+01
mean	7.000000	2611.333333	1.643295e+06
std	4.472136	1420.915282	9.347763e+05
min	0.000000	427.000000	2.480870e+05
25%	3.500000	1698.000000	1.046265e+06
50%	7.000000	2283.000000	1.319574e+06
75%	10.500000	3474.000000	2.228999e+06
max	14.000000	4976.000000	3.124928e+06

# Data Science

---

Temos alguns problemas:

- Notação científica pode atrapalhar a interpretação;
- Onde estão as outras variáveis?

# Data Science

---

- Notação científica pode atrapalhar a interpretação;

```
# Ajustar a configuração global para evitar notação científica
pd.set_option('display.float_format', '{:.2f}'.format)

escolas.describe()
```

# Data Science

---

- Onde está o describe das outras variáveis?

```
quantitavas = ['Numero_Alunos', 'Orcamento_Anual']  
qualitativas = ['Nome_Escola', 'Tipo_Escola']
```

```
escolas[quantitavas].describe()
```

```
escolas[qualitativas].describe()
```

# Exercício da Semana

# Data Science

---

**Qual é o orçamento total de todas as escolas?**

Dica: Use a função `sum()` para calcular o total da coluna `Orcamento_Anual`.

**Qual escola tem o maior e o menor gasto per capita (gasto por aluno)?**

Dica: Crie uma nova coluna para calcular o gasto per capita, dividindo `Orcamento_Anual` pelo `Numero_Alunos`, e use as funções `idxmax()` e `idxmin()` para identificar as escolas com o maior e menor valor.

# Data Science

---

**Qual é a média do número de alunos por tipo de escola (Pública vs. Particular)?**

Dica: Use a função `groupby()` combinada com `mean()` para calcular a média do número de alunos agrupada por tipo de escola (`Tipo_Escola`).

**Quantas escolas têm um orçamento anual acima de 1,5 milhão?**

Dica: Use a função `sum()` para contar quantas escolas têm `Orcamento_Anual` maior que 1.5 milhão.



# Dúvidas?

---



**Marco Mialaret, MSc**

**Telefone:**

**81 98160 7018**

**E-mail:**

**marcomialaret@gmail.com**

