

## Data Science – Princípios e Técnicas

Setembro  
2024



# Data Science



Onde me encontrar:

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

e

<https://github.com/MatmJr>

# Combinando datasets

# 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

---

Link do dataset:

[https://drive.google.com/uc?id=15aOJIGAyLMSY1gecjiCgu2ko\\_rilcKQy](https://drive.google.com/uc?id=15aOJIGAyLMSY1gecjiCgu2ko_rilcKQy)

# Data Science

escolas

	ID_Escola	Nome_Escola	Tipo_Escola	Numero_Alunos	Orcamento_Anual
0	0	Escola A	Publica	2917	1910635
1	1	Escola B	Publica	2949	1884411
2	2	Escola C	Particular	1761	1056600
3	3	Escola D	Publica	4635	3022020
4	4	Escola E	Particular	1468	917500
5	5	Escola F	Particular	2283	1319574
6	6	Escola G	Particular	1858	1081356

# Data Science

estudantes

	ID_Estudante	Nome_Estudante	Genero	Serie	Nome_Escola	Nota_Redacao	Nota_Matematica
0	0	Kevin Bradley	M	6	Escola A	66	79
1	1	Paul Smith	M	9	Escola A	94	61
2	2	John Rodriguez	M	9	Escola A	90	60
3	3	Oliver Scott	M	9	Escola A	67	58
4	4	William Ray	F	6	Escola A	97	84

# Data Science

---

Perguntas:

Existem dados faltantes nas tabelas?

Existe algo em comum nas duas tabelas?



# Data Science

---

Verificação de dados faltantes:

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

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

Repetir o processo para escolas.

# Data Science

---

Combinando os datasets:

```
data = pd.merge(estudantes,  
                 escolas,  
                 how='left',  
                 on=["Nome_Escola", "Nome_Escola"])
```

✓ 0.0s

# Data Science

---

Entendendo as variáveis categóricas:

```
data["Genero"].unique()
```

```
data["Serie"].unique()
```

```
data["Tipo_Escola"].unique()
```

# Data Science

---

Algumas perguntas:

- Qual o orçamento total das escolas?
- Qual a nota média dos alunos nas disciplinas analisadas?

# Data Science



Algumas perguntas:

- Qual o orçamento total das escolas?

```
escolas["Orçamento_Anual"].sum()
```

- Qual a nota média dos alunos nas disciplinas analisadas?

```
mediaRedacao = data["Nota_Redacao"].mean()  
mediaMatematica = data["Nota_Matematica"].mean()
```

# Data Science

Quantos alunos ficaram com nota superior a 90 em redação?

```
highRed = data[data["Nota_Redacao"]>90]  
len(highRed)
```

E qual o percentual?

```
len(highRed)/len(estudantes) * 100
```

# Data Science

Quantos alunos ficaram com nota superior a 90 em matemática?

```
highMat = data[data["Nota_Matematica"]>90]  
len(highMat)
```

Qual o percentual?

```
len(highMat)/len(estudantes) * 100
```

# Data Science

Quantos Alunos tiraram nota maior do que 90 nas duas disciplinas?

```
highBoth = data[(data["Nota_Redacao"]>90) & (data["Nota_Matematica"]>90)]  
len(highBoth)
```

Qual o percentual?

```
len(data[(data["Nota_Redacao"]>90) & (data["Nota_Matematica"]>90)]) / len(estudantes) * 100
```



# Data Science

Quantos alunos que obtiveram alto desempenho em ambas as disciplinas são de escolas públicas?

```
highBoth["Tipo_Escola"].value_counts()
```

✓ 0.0s

```
Tipo_Escola
```

```
Publica      1321
```

```
Particular   1002
```

```
Name: count, dtype: int64
```

# Data Science

Quantas alunas obtiveram alto desempenho em ambas as disciplinas?

```
highBoth["Genero"].value_counts()
```

✓ 0.0s

Genero

M 1167

F 1156

Name: count, dtype: int64

# Data Science

Como ficou a distribuição dos alunos de alto desempenho pelas série?

```
highBoth["Serie"].value_counts()
```

✓ 0.0s

Serie

6 663

8 604

7 599

9 457

Name: count, dtype: int64

# Data Science

---

Qual o total por escolas?

```
highBoth["Nome_Escola"].value_counts()
```

✓ 0.0s

# Data Science



Qual o total por escolas?

```
# Agrupando por 'Nome_Escola' e 'Tipo_Escola', e contando o número de registros em cada grupo
sumario = highBoth.groupby(['Nome_Escola', 'Tipo_Escola']).size().sort_values(ascending=False)

# Exibindo o sumário
print(sumario)
```

# Data Science

---

Qual o orçamento per capita de cada escola?

```
perCapita = escolas["Orçamento_Anual"]/escolas["Numero_Alunos"]  
escolas["Per_Capita"] = perCapita
```

✓ 0.0s

# Exercício da Semana

# Data Science

---

Quais escolas demonstraram maior eficiência?

Instruções:

1. Antes de analisar os dados, explique o que significa "eficiência" no contexto de escolas. Considere aspectos como o uso de recursos, desempenho acadêmico, custo-benefício, entre outros critérios relevantes.



# Data Science

---

2. Utilize os dados fornecidos neste documento para identificar quais escolas se destacam em termos de eficiência, de acordo com a definição estabelecida.
3. Quais diferentes formas de análise, novos resultados ou novas métricas podem ajudar a avaliar melhor a eficiência das escolas?

# Dúvidas?

---



81 98160 7018

marcomialaret@gmail.com

