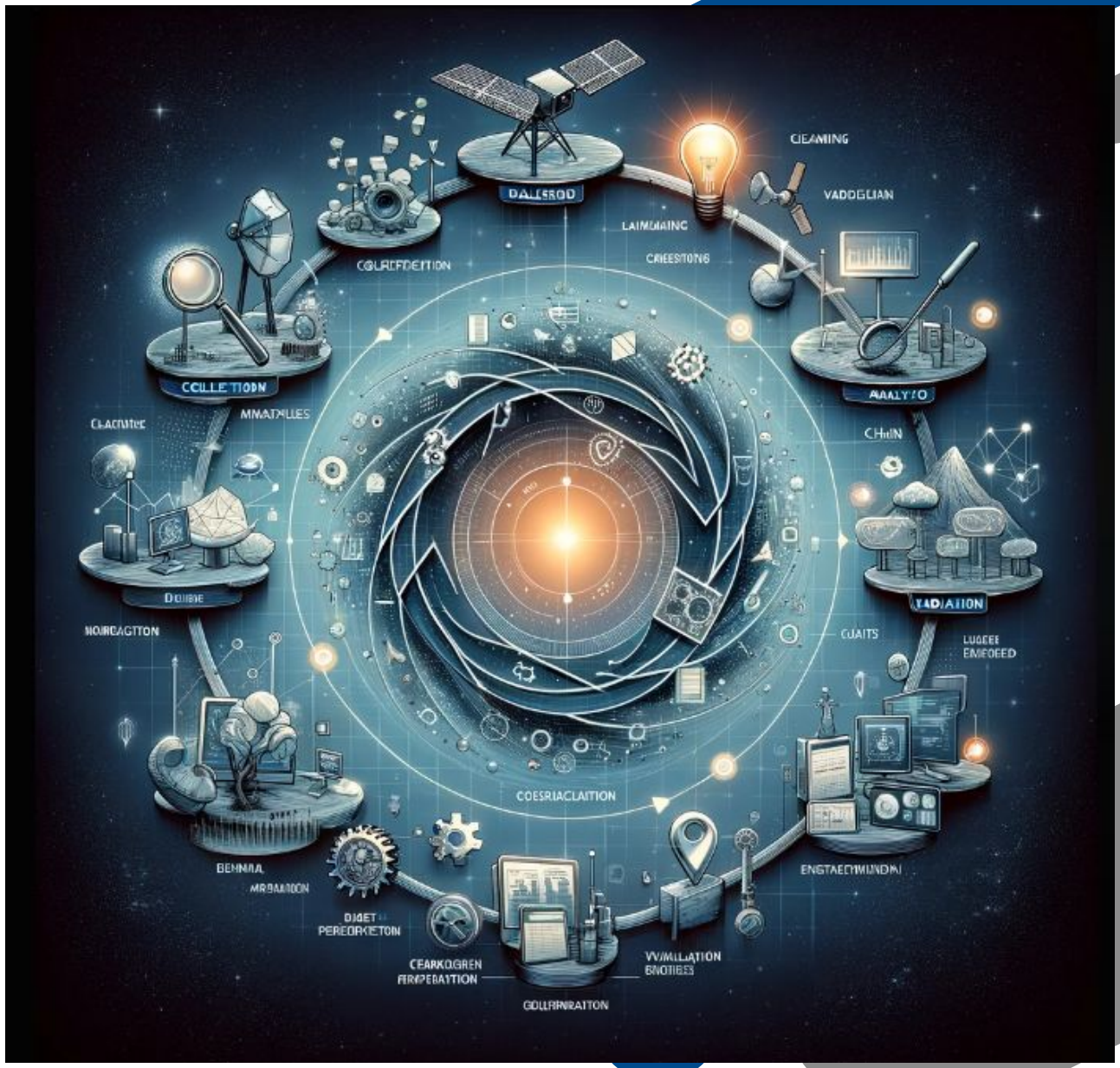
Senac

Fecomércio Sesc

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

Montando o ambiente

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)
```

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.

```
data['coluna_de_interesse'].sum()
```

Data Science

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

No pandas, se quisermos criar uma coluna que seja o resultado da operação de duas colunas já existentes, 'coluna_interesse1' e 'coluna_interesse2', podemos fazer isso diretamente.

```
data['Nova_coluna'] = data['coluna_interesse1'] / data['coluna_interesse2']
```

pandas permite realizar operações matemáticas de forma vetorizada, ou seja, operação por operação em cada linha de uma vez, sem a necessidade de loops.

Data Science

Para encontrar o índice do maior valor:

```
data['coluna_interesse'].idxmax()
```

Para encontrar o maior valor do atributo estudado:

```
data['coluna_interesse'].max()
```

Data Science

Fazer o mesmo para o mínimo.

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).

Quando utilizamos o método groupby no pandas, estamos agrupando os dados com base em uma variável categórica (ou coluna) e, em seguida, aplicando uma operação sobre uma variável numérica.

Data Science

```
data.groupby('variavel_categorica')['variavel_numerica'].metodo()
```

.metodo(): É o método que desejamos aplicar aos grupos da variável numérica. Incluem .sum(), .count(), .max(), .min(), entre outros.

Data Science

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.

```
data['coluna_interesse'] comparativo condição
```

Data Science

- comparativo pelo operador de comparação que deseja usar (por exemplo, >, <, >=, <=, ==, !=).

- condição pelo valor que define o critério de filtragem.

Dúvidas?

Marco Mialaret, MSc

Telefone:
81 98160 7018

E-mail:
marcomialaret@gmail.com