Senac

Fecomércio Sesc

Big Data

Prof. Marco Mialaret

Março

2024

Big Data

Onde me encontrar:

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

e

https://github.com/MatmJr

Na aula passada...

Eita, já esqueci ...

- Vimos os conceitos introdutório de Big Data
- Trabalhamos com um conjunto de dados com mais de 6 milhões de entradas.

Analisando dados

Big Data

A fase de análise em Big Data é destacada como a mais empolgante, permitindo a transformação de dados brutos em conhecimento útil. Essa prática de análise de dados, não sendo nova, ganhou ampla adesão recentemente como meio essencial para impulsionar negócios, reduzir custos e melhorar produtos e serviços.

Big Data

Para compreender os objetivos em uma análise de dados, é essencial questionar a origem, posse, contexto, benefícios e valor agregado dos dados. Essas perguntas ajudam a avaliar a viabilidade de investimento na análise.

Casos onde os dados não atendem às expectativas podem levar à descontinuação do uso, enquanto dados valiosos, porém imprecisos, podem ser ajustados na fase de preparação para alinhá-los aos objetivos desejados.

Big Data

Após a fase inicial de compreensão e ajuste dos dados, o processo de análise aprofunda-se, potencializando a descoberta de novos insights. É crucial agir corretamente, pois erros podem impedir a obtenção de resultados positivos.

Big Data

O sucesso na análise depende da capacidade de aplicar tanto uma abordagem científica quanto criativa, utilizando as ferramentas e técnicas adequadas. Esta aula discutirá como essas duas abordagens se complementam na análise de dados.

Características da análise de dados

Big Data

Os dados utilizados estão normalmente “sujos”

Muitas vezes, as bases contêm dados incompletos, inconsistentes, corrompidos, duplicados ou em formatos inadequados, entre outros problemas. Portanto, é essencial a intervenção de um profissional capacitado para tratar esses dados antes de iniciar a análise propriamente dita.

Big Data

Gasta-se mais tempo preparando do que analisando os Dados

Devido às peculiaridades de cada base de dados, o tratamento necessário muitas vezes requer avaliação e definição manual, limitando a automação do processo. Por isso, a etapa de tratamento dos dados pode ser demorada. Apesar de onerosa, essa fase é crucial para prevenir inconsistências nos resultados das análises.

Big Data

Procura de uma agulha em um palheiro

Analisar grandes volumes de dados para descobrir padrões é comparável a procurar uma agulha em um palheiro, destacando a complexidade e o tempo exigido para essa tarefa. No entanto, no âmbito do Big Data, o desafio se amplia: não se trata apenas de encontrar a agulha, mas de definir o que constitui uma "agulha" — ou seja, determinar qual pergunta os dados podem responder.

Big Data

Correlação não implica causalidade

Um princípio fundamental da estatística é que correlação não implica causalidade. Correlação indica que dois eventos, "A" e "B", ocorrem juntos com frequência, mas isso não significa que um cause o outro. Às vezes, a correlação pode ser mera coincidência. Para estabelecer causalidade, são necessários testes estatísticos e experimentos controlados.

Big Data

Interpretar uma correlação como causalidade sem essas verificações pode levar a conclusões errôneas, como associar erroneamente o aumento das vendas de sorvete ao aumento de afogamentos. Portanto, é crucial ter cautela ao interpretar dados.

Big Data

É fácil fazer a análise de dados de forma errada

Pesquisadores alertam para o risco de que as atuais ferramentas de análise de dados, ao facilitarem a criação de algoritmos com variados conjuntos de dados, possam também levar a erros ou interpretações equivocadas.

Big Data

Esses equívocos podem produzir resultados aparentemente promissores que, na realidade, não refletem a verdade. Por isso, é crucial validar as conclusões obtidas, especialmente ao trabalhar com grandes volumes de dados, onde inconsistências podem não ser imediatamente óbvias.

O Processo de Análise de Dados

Big Data

Quando falamos em Big Data e em análise de dados, é comum ouvirmos palavras como identificação de padrões, modelagem dos dados, detecção de grupos, classificação de dados. Essas atividades são possíveis por meio da utilização de técnicas há muito tempo desenvolvidas, como técnicas estatísticas, matemáticas, de aprendizado de máquina e de mineração de dados.

Big Data

Embora cada processo tenha definições distintas, em geral, eles envolvem as seguintes etapas:

1. **Entendimento do negócio:** aqui são definidas as perguntas, o objetivo da análise de dados e o plano a ser seguido;

2. **Compreensão dos dados:** etapa utilizada para coletar e explorar os dados, aumentando a compreensão sobre sua estrutura, atributos e contexto;

Big Data

3. **Preparação dos dados:** após a análise exploratória, inicia-se o processo de limpeza, filtragem, estruturação, redução e integração dos dados;

4. **Modelagem dos dados:** envolve as tarefas de seleção dos dados, definição e construção do modelo;

Big Data

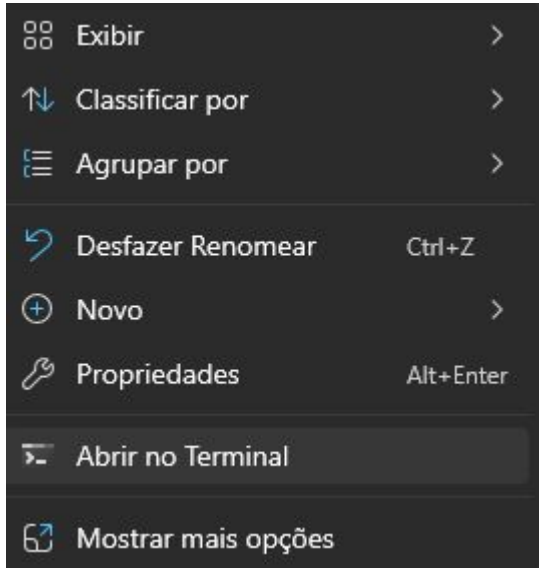
5. **Validação do modelo:** os resultados gerados pelo modelo são avaliados, para verificar se a precisão obtida está satisfatória e coesa;

6. **Utilização do modelo:** após serem validados, os resultados dos modelos são utilizados e monitorados.

Criação do Ambiente de Trabalho

Big Data

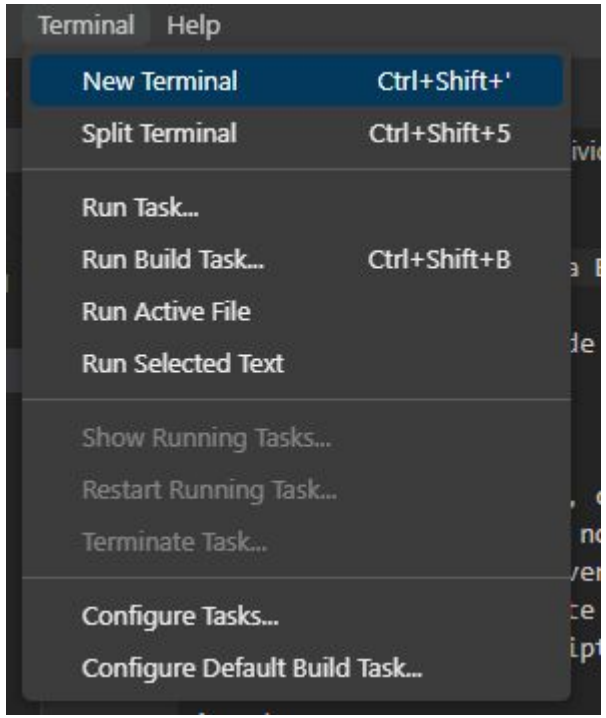
Após criar a pasta chamada BigData, clicar com o botão do lado direito e selecionar a opção abrir no terminal

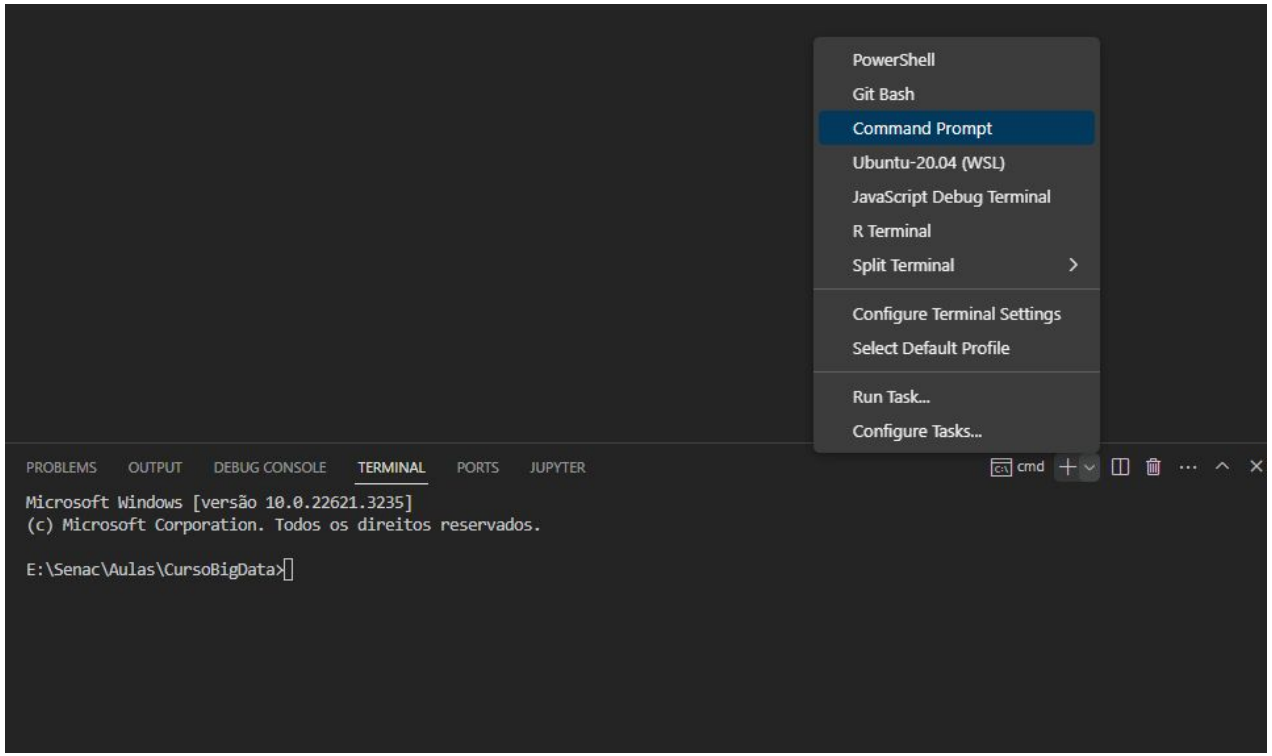



```
egamento de perfis pessoais e do sistema levou 729ms.
PS E:\Senac\Aulas\Slides\BigData\BigData> code .
```

Big Data

No VSCode:





Big Data

Criando o ambiente:

```
E:\Senac\Aulas\CursoBigData>python -m venv venv
```

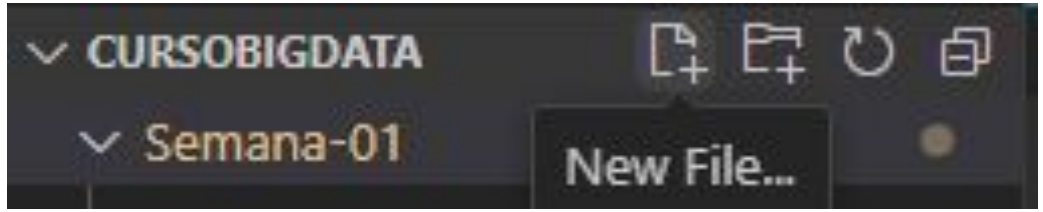
Ativando o ambiente:

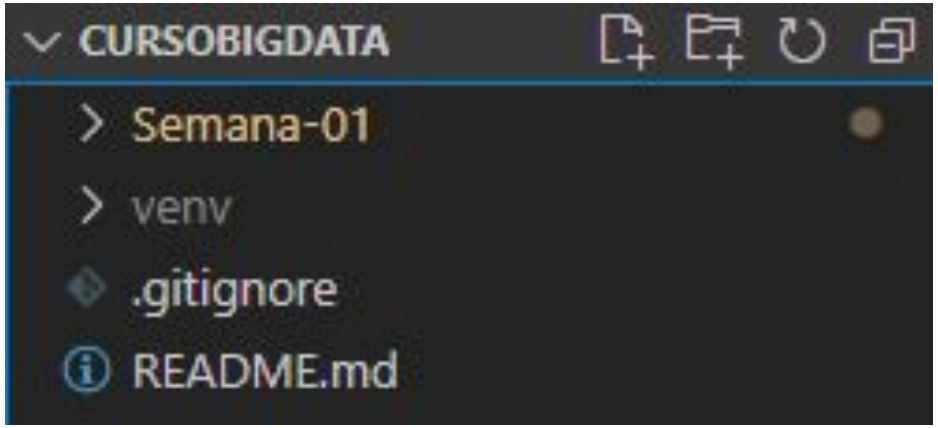
```
E:\Senac\Aulas\CursoBigData>.\venv\Scripts\activate

(venv) E:\Senac\Aulas\CursoBigData>
```

Big Data

Criação de arquivos no VSCode:





Big Data

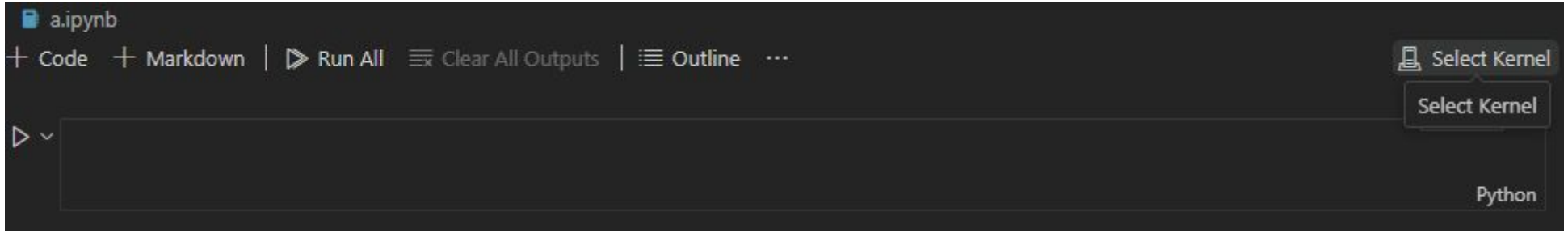
Crie um arquivo chamado aula01.ipynb.

No terminal, instale o pandas no ambiente virtual:

```
(venv) E:\Senac\Aulas\CursoBigData>pip install pandas
Collecting pandas
  Using cached pandas-2.2.1-cp311-cp311-win_amd64.whl.metadata (19 kB)
Collecting numpy<2,>=1.23.2 (from pandas)
  Using cached numpy-1.26.4-cp311-cp311-win_amd64.whl.metadata (61 kB)
```

Big Data

Ativando o ambiente virtual no notebook Jupyter



O conjunto de Dados

Big Data

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.

Big Data

Acessando o Arquivo no drive

```
import pandas as pd
import requests
from io import StringIO

# Criação do dataFrame dos alunos
# ID do arquivo no Google Drive
file_id = '15aOJIGAyLMSY1gecjiCgu2ko_riIcKQy'

# URL modificada para forçar o download do arquivo
url = f"https://drive.google.com/uc?id={file_id}"
```

Big Data

```
# Tentando obter o arquivo com requests
try:
    response = requests.get(url)
    response.raise_for_status()  # Lança um erro para respostas não-sucedidas
    # Usando StringIO para converter o texto em um arquivo em memória e, então, lendo com o Pandas
    csv_raw = StringIO(response.text)
    estudantes = pd.read_csv(csv_raw)
except requests.RequestException as e:
    print(f"Erro ao acessar o arquivo: {e}")
```

Big Data

```
#Criação do dataFrame das escolas

# ID do arquivo no Google Drive
file_id = '1Jgto7psHaMRTAVzcFt7D6SgJiHMB7uGT'

# URL modificada para forçar o download do arquivo
url = f"https://drive.google.com/uc?id={file_id}"

# Tentando obter o arquivo com requests
try:
    response = requests.get(url)
    response.raise_for_status()  # Lança um erro para respostas não-sucedidas
    # Usando StringIO para converter o texto em um arquivo em memória e, então, lendo com o Pandas
    csv_raw = StringIO(response.text)
    escolas = pd.read_csv(csv_raw)
except requests.RequestException as e:
    print(f"Erro ao acessar o arquivo: {e}")
```

Big Data

```
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

Big Data

```
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

Big Data

Perguntas:

Existem dados faltantes nas tabelas?

Existe algo em comum nas duas tabelas?

Big Data

Verificação de dados faltantes:

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

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

Repetir o processo para escolas.

Big Data

Combinando os datasets:

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

✓ 0.0s

Big Data

Entendendo as variáveis categóricas:

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

data["Serie"].unique()

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

Big Data

Algumas perguntas:

- Qual o orçamento total das escolas?

- Qual a nota média dos alunos nas disciplinas analisadas?

Big Data

Algumas perguntas:

- Qual o orçamento total das escolas?

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

- Qual a nota média dos alunos nas disciplinas analisadas?

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

Big Data

Algumas perguntas:

- Qual o orçamento total das escolas?

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

- Qual a nota média dos alunos nas disciplinas analisadas?

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

Big Data

Personalizando os resultados obtidos:

```
# Criando um novo DataFrame com os resultados
resultados = pd.DataFrame({
    "Operação": ["Média Redação", "Média Matemática"],
    "Resultado": [mediaRedacao, mediaMatematica]
})

# Exibindo a tabela
display(resultados)
```

Big Data

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
```

Big Data

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
```

Big Data

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
```

Big Data

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
```

Big Data

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
```

Big Data

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
```

Big Data

Quantas alunas obtiveram alto desempenho em ambas as disciplinas?

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

✓ 0.0s

```
Genero
M    1167
F    1156
Name: count, dtype: int64
```

Big Data

Quantas alunas obtiveram alto desempenho em ambas as disciplinas?

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

✓ 0.0s

```
Genero
M    1167
F    1156
Name: count, dtype: int64
```

Big Data

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
```

Big Data

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
```

Big Data

Qual o total por escolas?

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

✓ 0.0s

Big Data

```
# 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)
```

Big Data

Qual o orçamento per capita de cada escola?

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

✓ 0.0s

Big Data

Qual o orçamento per capita de cada escola?

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

✓ 0.0s

Dúvidas?

Marco Mialaret, MSc

Telefone:
81 98160 7018

E-mail:
marcomialaret@gmail.com