

Fecomércio Sesc

Data Science – Princípios e Técnicas

Março

2024



Onde me encontrar:

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

e

https://github.com/MatmJr









O Ciclo de Vida da Ciência de Dados envolve várias etapas, desde a limpeza de dados até a modelagem e avaliação, para extrair insights e previsões úteis para objetivos comerciais.

As principais etapas são:







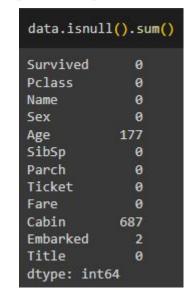
Estamos trabalhando com 891 observações de 12 variáveis. Para tornar as coisas um pouco mais explícitas, já que alguns dos nomes das variáveis não são totalmente claros, aqui está o que temos que lidar:

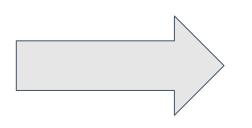


Nome da Variável	Descrição
Survived	Sobreviveu (1) ou morreu (0)
Pclass	Classe do passageiro (1 = Upper Class, 2 = Middle Class, 3 = Lower Class)
Name	Nome do passageiro
Sex	Sexo do passageiro
Age	Idade do passageiro
SibSp	Número de irmãos/cônjuges a bordo
Parch	Número de pais/filhos a bordo
Ticket	Número do bilhete
Fare	Tarifa
Cabin	Cabine
Embarked	Porto de embarque C = Cherbourg, Q = Queenstown, S = Southampton









data.isna().sum()			
Survived	0		
Pclass	0		
Name	0		
Sex	0		
Age	0		
SibSp	0		
Parch	ø		
Ticket	ө		
Fare	0		
Cabin	0		
Embarked	0		
Title	0		
dtype: int	64		





Transformando os passos da aula passada em um Script



```
# Criação do dataFrame dos alunos
# ID do arquivo no Google Drive
file_id = '1S5Nl793vcL5ZPTGjzKaIEbwbLaDplvIP'

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





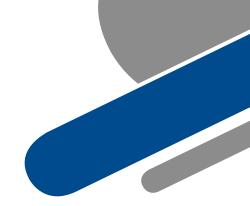
```
# Tentando obter o arquivo com requests
try:
    response = requests.get(url)
    response.raise_for_status() # Lança um erro para respostas não-sucedidas
    # Usando StringIO para converter o texto em um arquivo em memória e, então, lendo com o Pandas
    csv_raw = StringIO(response.text)
    data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



```
#Transformando o PassengerId no indice
data = data.set_index("PassengerId")

#Criando uma função que encontra os pronomes de tratamentos dos passageiros
def extract_title(name):
    title_search = re.search(' ([A-Za-z]+)\.', name)
    if title_search:
        return title_search.group(1)
    return ""
```





```
#Aplicando a função na coluna de nomes, isto é, criando uma coluna só com os pronomes de tratamentos
data['Title'] = data['Name'].apply(extract_title)

#Substituindo os valores ausentes das idades pela mediana agrupada por sexo e classe de passageiro
data['Age'] = data.groupby(['Sex', 'Pclass'])['Age'].transform(lambda x: x.fillna(x.median()))

#Substituindo os valores ausentes do porto de embarque pelo valor que mais apareceu
data['Embarked'] = data['Embarked'].fillna('S')
```





```
#Substituindo os valores ausentes da cabine com base na Classe e no mapa do navio
for num in [1, 2, 3]:
    if num == 1:
        data.loc[data['Pclass'] == 1, 'Cabin'] = data.loc[data['Pclass'] == 1, 'Cabin'].fillna('ABC')
    elif num == 2:
        data.loc[data['Pclass'] == 2, 'Cabin'] = data.loc[data['Pclass'] == 2, 'Cabin'].fillna('DE')
    elif num == 3:
        data.loc[data['Pclass'] == 3, 'Cabin'] = data.loc[data['Pclass'] == 3, 'Cabin'].fillna('FG')
```



Agora temos um script que baixa o arquivo e faz o processamento dos dados.



Nesta aula, você aprenderá:

- Métricas usadas para descrever e resumir dados, as estatísticas descritivas.
- As principais bibliotecas Python usadas no estudo das estatísticas descritivas.









Nos últimos anos, houve um crescimento exponencial no volume de dados gerados pela humanidade, o que gerou uma demanda crescente por profissionais capazes de extrair informações e tomar decisões fundamentadas com base nesses dados. Para atender a essa demanda, os profissionais da área de dados precisam dominar campos essenciais como big data, inteligência artificial, ciência de dados, aprendizado de máquina, entre outros.



Um aspecto crucial ao trabalhar com dados é a habilidade de descrevê-los, resumi-los e representá-los visualmente. A estatística descritiva é uma ferramenta essencial nesse processo, utilizando duas abordagens principais:



- A abordagem quantitativa, que descreve e resume os dados numericamente.
- A abordagem visual, que ilustra os dados por meio de gráficos e visualizações.



Na análise quantitativa, destacamos:

- A tendência central informa sobre os centros dos dados. Medidas úteis incluem a média, mediana e moda.
- A variabilidade informa sobre a dispersão dos dados. Medidas úteis incluem variância e desvio padrão.



- A correlação (ou variabilidade conjunta) informa sobre a relação entre um par de variáveis em um conjunto de dados. Medidas úteis incluem a covariância e o coeficiente de correlação.





As medidas de tendência central



Média:

A média aritmética, ou simplesmente média, de um conjunto de valores é a medida de centro encontrada somando todos os valores do conjunto e dividindo pelo número de valores. Assim:

$$M\'{e}dia = \frac{soma~dos~valores}{total~de~observa\~{c}\~{o}es}$$



Exemplo 1: Determine a idade média dos passageiros do Titanic

data["Age"].mean()

round(data["Age"].mean(),2)



O Exemplo 1 nos mostrou que a idade média das pessoas que estavam no Titanic foi de aproximadamente 29 anos, mas será que esse número representa bem todos os passageiros da embarcação?



Para tentar responder essa pergunta vamos dividir o nosso conjunto de dados em subconjuntos menores, buscaremos características que separam o conjunto em subconjunto complementares, por exemplo: Survived, Pclass, Sex, Age, Siblings/Spouses Aboard...



Essa ideia de dividir o conjunto original em subconjuntos com características determinadas é conhecida como **Estratificação**.



Exemplo 2:



Vimos na aula passada:

data.groupby(["Sex","Pclass"])["Age"].mean()



Obs: Existem outras médias, porém cada uma delas é usada em situações específicas. A saber:

- Média Ponderada: Você deve usar uma média ponderada quando deseja atribuir mais importância a alguns números em um conjunto de dados do que a outros. Isso é útil em cenários onde um evento pode ter vários resultados positivos ou negativos, e a magnitude desses resultados varia.



- Média Harmônica: A média harmônica é calculada como o número de valores dividido pela soma do inverso de cada valor. É apropriada quando os dados representam grandezas que são inversamente proporcionais, como taxas.



- Média Geométrica: A média geométrica é calculada como a raiz N-ésima do produto de todos os valores, onde N é o número de valores. É útil quando os dados estão em uma escala multiplicativa, como em situações envolvendo crescimento ou taxa de variação entre diferentes unidades de medida.



Mediana

A **mediana** da amostra é o elemento central de um conjunto de dados ordenado (crescente ou decrescente). Se o número de elementos n do conjunto de dados for ímpar, então a mediana é o valor na posição do meio. Se n for par, então a mediana é a média aritmética dos dois valores no meio



Exemplo 3: Se tivermos os pontos de dados 3, 5, 1, 2 e 8, o valor mediano será 3, pois 3 é o elemento central após a ordenação do conjunto (1, 2, 3, 5, 8). Se os pontos de dados forem 3, 5, 1 e 8, então a mediana será 4, que é a média dos dois elementos centrais da sequência ordenada (1, 3, 5, 8).



Importante: A principal diferença entre o comportamento da média e da mediana está relacionada aos valores extremos (outliers) do conjunto de dados. De uma maneira geral:



- Se você colocar um valor discrepante em um conjunto de dados, a média aumentará, mas o valor da mediana permanecerá inalterado.
- Se você remover um valor discrepante de um conjunto de dados, a média diminuirá, mas a mediana continuará a mesma.



Exemplo 4: Vamos determinar a mediana das Idades:

data["Age"].median()

data.groupby(["Sex","Pclass"])["Age"].median()



Moda

A moda da amostra é o valor no conjunto de dados que ocorre com mais frequência. Se não houver um único valor desse tipo, o conjunto será multimodal, pois possui vários valores modais.



Exemplo 5: Determine a moda dos conjuntos a seguir:

data["Age"].mode()

data.groupby(["Sex", "Pclass"])["Age"].apply(lambda x: x.mode().iloc[0] if not x.empty else None)



Medidas de Localização

O percentil p da amostra é o elemento no conjunto de dados tal que p% dos elementos no conjunto de dados são menores ou iguais a esse valor. Além disso, (100 - p)% dos elementos são maiores ou iguais a esse valor. Se houver dois desses elementos no conjunto de dados, o percentil p da amostra é a média aritmética deles.



- O primeiro quartil Q1 é o percentil 25 da amostra. Ele divide aproximadamente 25% dos menores itens do restante do conjunto de dados.
- O segundo quartil Q2 é o percentil 50 da amostra, também conhecido como a mediana. Aproximadamente 25% dos itens situam-se entre o primeiro e o segundo quartis, e outros 25% entre o segundo e o terceiro quartis.



- O terceiro quartil Q3 é o percentil 75 da amostra. Ele divide aproximadamente 25% dos maiores itens do restante do conjunto de dados.



Exemplo 6: Determine os quartis das idades dos passageiros do Titanic.

from statistics import quantiles

quantiles(data.Age, n=4, method='inclusive')

data.groupby(["Sex", "Pclass"])["Age"].quantile([0.25, 0.50, 0.75])



Dúvidas?







Marco Mialaret, MSc

Telefone:

81 98160 7018

E-mail:

marco.junior@pe.senac.br

