

Fecomércio Sesc

Data Science – Princípios e Técnicas

Maio

2024



Onde me encontrar:

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

e

https://github.com/MatmJr





Vimos nas aulas passadas



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:



Montamos um script que acessa o arquivo e faz o processamento dos dados.



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



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.



Vimos na aula passada:

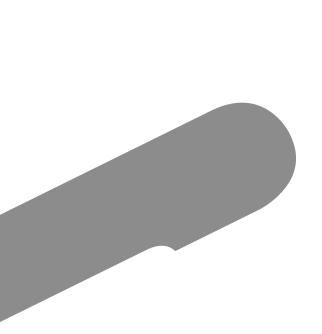
- A tendência central informa sobre os centros dos dados. Medidas úteis incluem a média, mediana e moda.
- A variabilidade informa sobre a dispersão dos dados. Medidas úteis incluem variância e desvio padrão.



Veremos na aula de hoje:

- A correlação (ou variabilidade conjunta) informa sobre a relação entre um par de variáveis em um conjunto de dados.





A Correlação



A análise de correlação mede o grau de dependência entre duas ou mais variáveis, ou seja, como uma variável influencia outra. Esta relação pode ser não-causal e é quantificada por coeficientes, como o coeficiente de Pearson, que será o foco desta aula.



O coeficiente de Pearson, também chamado de "coeficiente de correlação produto-momento" ou chamado de "p de Pearson", mede o grau de correlação através do cálculo de direção positiva ou negativa. Este coeficiente, normalmente representado por p assume apenas valores entre -1 e 1.



A análise de correlação vai retornar três possíveis cenários:

- correlação positiva;
- correlação negativa;
- não há correlação.



 Correlação positiva: quando duas variáveis que possuem correlação crescem ou decrescem juntas, ou seja, que possuem uma relação direta;



 Correlação negativa: quando duas variáveis que possuem correlação mas quando uma variável cresce a outra decresce, ou vice-versa;

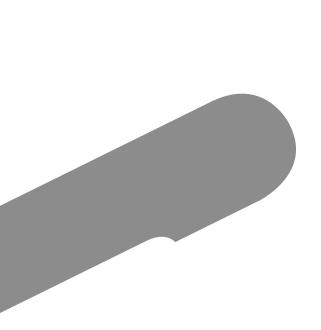


- Não ter correlação: quando o crescimento ou decrescimento de uma variável não tem efeito sobre outra variável.



```
\rho = 0,9 a 1 (positivo ou negativo): correlação muito forte; \rho = 0,7 a 09 (positivo ou negativo): correlação forte; \rho = 0,5 a 0,7 (positivo ou negativo): correlação moderada; \rho = 0,3 a 0,5 (positivo ou negativo): correlação fraca; \rho = 0 a 0,3 (positivo ou negativo): não possui correlação.
```









O primeiro passo para usar esse conceito no python é importar as bibliotecas necessárias: `pandas` e `seaborn`. `pandas` é essencial para manipular dados, tabelas e dataframes, enquanto `seaborn` é ideal para criar visualizações gráficas, especialmente para mapeamentos estatísticos.

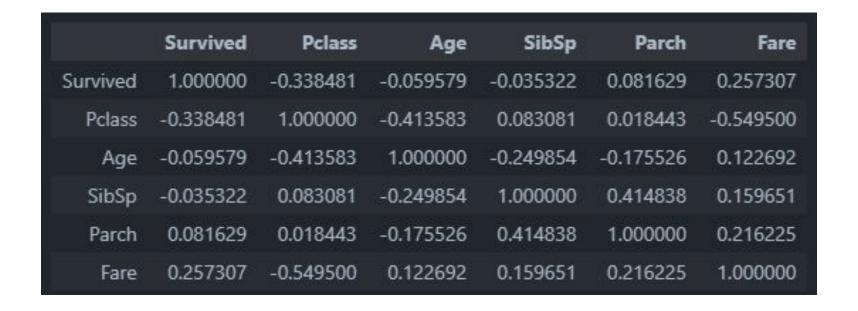


import pandas as pd import seaborn as sns

nums = ["Survived","Pclass","Age","SibSp","Parch","Fare"]

data[nums].corr()







Para transformar as colunas categóricas 'Sex', 'Cabin', e 'Title' em números usando LabelEncoder do pacote sklearn, você pode seguir os passos a seguir:



from sklearn.preprocessing import LabelEncoder

```
# Inicializando o LabelEncoder
label_encoder = LabelEncoder()
```

Colunas a serem transformadas columns_to_encode = ['Sex', 'Cabin', 'Title']



```
# Aplicando LabelEncoder às colunas categóricas
for column in columns_to_encode:
   data[column] = label_encoder.fit_transform(data[column])
```

Mostrando o DataFrame transformado data.head()



Para visualizarmos a matriz de correlação, vamos utilizar a função .heatmap() do pacote seaborn, essa função vai nos retornar uma forma gráfica da matriz com uma escala de cor em conjunto com uma escala numérica, as quais vão indicar o grau medido entre as variáveis.



correlation = data[vars].corr()

plot da matriz de correlação

sns.heatmap(correlation, annot = True, fmt=".1f", linewidths=.6)



Dúvidas?









Marco Mialaret, MSc

Telefone:

81 98160 7018

E-mail:

marco.junior@pe.senac.br

