Análise de Dados e Inferência Estatística com Python



Aula de hoje!

- X Como serão as aulas?
- Pandas Leitura de DataFrames
- Manipulação de Dados com Python



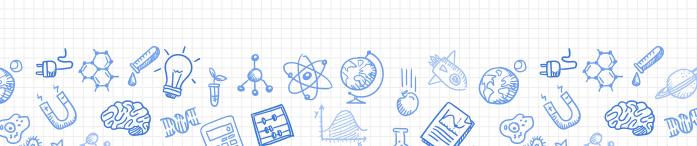
Como serão as aulas?

Aulas

- Apresentar, de forma detalhada, as técnicas de estatística utilizando Python e ciência de dados:
 - o Teoria;
 - Trabalhos práticos;
 - Frameworks;
 - Estatística.

- Unidade I Apresentação do que é ciência de dados.
- Unidade II Frameworks.
- Unidade III Modelos estatísticos.
- Unidade IV Solução de modelos estatísticos.





Os DataFrames são estruturas que comportam dados de forma tabular. Os DataFrames são compostos de linhas e colunas, sendo cada coluna um campo da tabela e cada linha um registro. Cada coluna possui dados de um mesmo tipo. Você pode imaginar uma tabela de Excel, mas nesse caso, cada coluna é limitada a um tipo de dado.



Primeiro, uma forma bem simples de criar um DataFrame, através do Dictionary do Python:

Link: https://github.com/paulovpcotta/aula_l_python_estatis



Vamos utilizar o famoso exemplo do Titanic, do Kaggle, para exemplificar. Para quem não sabe o que é, o Kaggle é um site de competições de Ciência de Dados.



Manipulação de Dados com Python

Agora do básico de manipulação de dados com Python, alteração de valores, remoção de colunas, etc. Vamos novamente usar o Dataset do Titanic para a maior parte de nossos exemplos.



Removendo info. do DataFrame

Vamos começar falando sobre a remoção de informações de um DataFrame. Informações que não sejam úteis em um dado momento sempre podem ser removidas para facilitar seu trabalho.

Para remover informações de um DataFrame, use o drop. Para o drop, você deve incluir como argumento o índice a ser retirado, seja ele nome ou posição. Se você não definir o argumento axis, ele remove a linha. Se você definir axis=1 como argumento, ele remove a coluna. Você também pode passar uma lista como argumento, e todas as linhas ou colunas dentro da lista serão retiradas. É importante notar que ele não remove a informação diretamente do DataFrame em que for usado. Caso você queira manter o drop, deve criar um novo objeto junto com o comando.



Removendo info. do DataFrame

Vamos começar falando sobre a remoção de informações de um DataFrame. Informações que não sejam úteis em um dado momento sempre podem ser removidas para facilitar seu trabalho.

Para remover informações de um DataFrame, use o drop. Para o drop, você deve incluir como argumento o índice a ser retirado, seja ele nome ou posição. Se você não definir o argumento axis, ele remove a linha. Se você definir axis=1 como argumento, ele remove a coluna. Você também pode passar uma lista como argumento, e todas as linhas ou colunas dentro da lista serão retiradas. É importante notar que ele não remove a informação diretamente do DataFrame em que for usado. Caso você queira manter o drop, deve criar um novo objeto junto com o comando.



NaN

Caso o Dataset em que você esteja trabalhando esteja com valores inexistentes demais, o Pandas possui funções para preencher estes valores inexistentes da melhor forma possível.

Primeiramente, você pode verificar onde no seu DataFrame existem valores não existentes, com a função isnull①. No caso do Dataset do Titanic especificamente, a coluna "Cabin" possui muitos valores não existentes, então você pode conferir o comando filtrando esta coluna.



Algumas Estatíst. em DataFrames

O Pandas também possui uma série de comandos para cálculo de estatísticas descritivas ao longo de um DataFrame, podendo calcular ao longo de linhas ou colunas. Entre estas estatísticas estão média, moda, mediana, contagem, soma, entre muitos outros. Por padrão, o cálculo é feito por coluna, mas caso você passe o argumento axis=1, o cálculo é feito ao longo das linhas.



Ordenando DataFrames (sorting)

Para leitura dos dados, também pode ser importante ordenar um DataFrame, ou seja, fazer um Sort. Para ordenar pelos índices das linhas, vamos primeiro reindexar o DataFrame para que os índices fiquem revertidos. Depois vamos usar a função sort_index() para ordena-lo. Você também pode passar o argumento ascending=False para ordena-lo de forma descendente.



