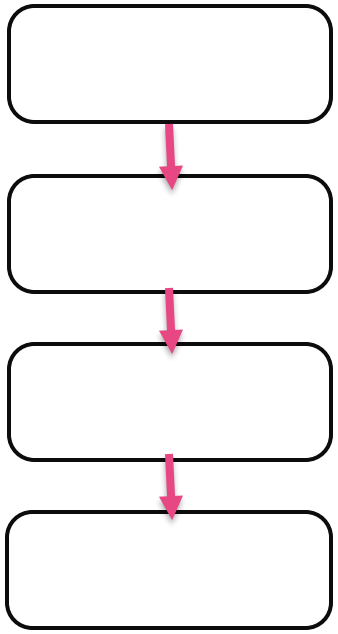
Python para dados

Data Wrangling com Python

WOMAKERS CODE

O que é?



- Exploração
- Limpeza
- Transformação
- Análise (modelos de ML, dashboard)

É um componente chave de qualquer projeto de análise de dados. Wrangling é um processo no qual se transforma dados "brutos" para torná-los mais adequados para análise, melhorando assim a qualidade dos seus dados.

Em português, pode ser traduzido como "**manipulação de dados**" ou "**tratamento de dados**". Trata-se de preparar os dados para que possam ser facilmente explorados e analisados.

Exploração e coleta de dados

Exploração e coleta de dados

Coletamos dados de diferentes fontes e é necessário checá-las antes. Normalmente em um projeto nós temos a fonte de dados disponível, mas em outros ainda temos que buscar essa informação e checar a qualidade desse dado, como a fonte.

Exploração de Dados: Verificação dos tipos de dados das características, valores únicos e descrição dos dados.

Limpeza

Por que precisamos limpar dados?

- Qualidade
- Consistência
- Integridade
- Padronização
- Melhor precisão
- Melhor eficiência
- Prevenção de viés
- Conformidade com regulamentação

Manipulando o dataframe para limpeza

- **rfewr**
- werwerwe

- **werwer**
- werwerw

Manipulando o dataframe para limpeza

- ## Removendo colunas e linhas

  O pandas oferece uma maneira prática de remover colunas ou linhas indesejadas de um DataFrame com a função drop().

- ## Mudando index

  Em muitos casos, é útil utilizar um campo de identificação com valores únicos nos dados como seu índice.

- ## Removendo duplicados

Lidando com dados faltosos (missing data)

Lidar com dados ausentes em um DataFrame é uma parte essencial do processo de análise de dados. A escolha da estratégia de lidar com dados ausentes depende do contexto específico do seu conjunto de dados e do impacto potencial nas análises.

- **Identificar Dados Ausentes:** antes de lidar com dados ausentes, é importante identificá-los. Utilize o método **df.isna()** para identificar células com valores nulos.
- **Remover Dados Ausentes:** se a quantidade de dados ausentes for pequena e não comprometer a análise, você pode optar por remover as linhas ou colunas correspondentes.
- **Preencher Dados Ausentes:** preencher valores ausentes pode ser feito utilizando o método **fillna()**, onde você pode especificar um valor constante ou usar métodos de preenchimento mais avançados, como preenchimento pela média ou mediana.
- **Imputação Estatística:** outra abordagem é a imputação estatística, onde você preenche valores ausentes com base em modelos estatísticos.

Transformação

Várias transformações podem ser aplicadas aos dados

1. **Tratamento de Valores Ausentes:** Preencher ou remover valores ausentes para evitar problemas durante a análise.
2. **Codificação de Variáveis Categóricas:** Converter variáveis categóricas em uma forma numérica, como one-hot encoding, para que possam ser utilizadas em algoritmos de machine learning.
3. **Normalização ou Padronização:** Escalar valores numéricos para uma escala comum, como normalização (escala de 0 a 1) ou padronização (média 0, desvio padrão 1).
4. **Conversão de Tipos de Dados:** Garantir que os tipos de dados das colunas estão corretos para a análise, convertendo strings em datas, números inteiros, etc.
5. **Remoção de Dados Duplicados:** Identificar e remover entradas duplicadas que podem prejudicar a precisão da análise.

Várias transformações podem ser aplicadas aos dados

6. **Agregação de Dados:** Agrupar dados com base em determinadas características e calcular estatísticas agregadas, como média, soma, mínimo ou máximo.
7. **Criação de Novas Variáveis (Feature Engineering):** Criar novas variáveis que podem ser mais informativas para a análise, como calcular a idade a partir da data de nascimento.
8. **Divisão de Dados:** Separar conjuntos de dados em subconjuntos com base em condições específicas.
9. **Renomeação de Colunas:** Modificar os nomes das colunas para torná-los mais descritivos e compreensíveis.
10. **Mudança do Formato do DataFrame:** Reorganizar ou pivotar o DataFrame para facilitar a análise ou a criação de visualizações.

HORA DE PRATICAR

WOMAKERS CODE

Cenário 1

A empresa Carri Construtora contratou a empresa que você trabalha para encontrar possíveis compradores para o seus novos empreendimentos.

A empresa quer entender as necessidades dos clientes, e quer informações com:

- Quais cliente devemos abordar;
- Qual empreendimento nós devemos mostrá-los;
- Esse cliente está em busca em investir em um imóvel ou comprar para moradia?

Cenário 2

Uma empresa chamada Hillo, é uma empresa que está estudando o mercado e quer encontrar uma parceria com uma empresa de Streeming (netflix, Disney plus etc), mas gostaria de saber quais as empresas dariam o maior retorno de investimento.

Eles podem fazer análise de machine learning também com esses dados.

Cenário 3

Uma influenciadora digital de bem estar gostaria de analisar possíveis empreendimentos dentro de diferentes propostas que recebe. Essas propostas podem ser excludentes ou somatórias.