Python - Pandas

Pandas

- (http://pandas.pydata.org/)
- Desde o seu surgimento em 2010, tem ajudado a viabilizar o Python como um ambiente eficaz e produtivo para análise de dados
- Oferece estruturas de dados de alto nível e funções, projetadas para fazer com que trabalhar com dados estruturados ou tabulares seja rápido, fácil e expressivo.

Pandas

Principais objetos:

DataFrame:

– uma estrutura de dados tabular, orientada a colunas, com rótulos (labels) tanto para linhas quanto para colunas

Series:

– um objeto array unidimensional, com rótulo

- Combina as ideias de processamento de alto desempenho de arrays da NumPy com os recursos flexíveis de manipulação de dados das planilhas e dos bancos de dados relacionais (como o SQL)

Pandas

- disponibiliza uma funcionalidade sofisticada de indexação para facilitar a reformatação, a manipulação, as agregações e a seleção de subconjuntos de dados.

- manipulação, preparação e a limpeza de dados são habilidades importantes na análise de dados, justificando a importância da biblioteca Pandas.

Pandas

Outras características:

- funcionalidade para séries temporais integradas;
- mesmas estruturas de dados para lidar com séries de dados tanto temporais quanto não temporais;
- operações aritméticas e reduções que preservem metadados;
- tratamento flexível para dados ausentes;
- combinações (merge) e outras operações relacionais que se encontram em bancos de dados populares (baseados em SQL, por exemplo).

Limpeza e Tratamento de Dados

Cientista de Dados: gasta 80% do tempo com tratamento de dados

Produção de Dados x Análise de Dados

Limpeza e Tratamento de Dados

Por que Dados tem problemas?

Sistemas de Operações e bancos de dados sem restrições de entrada

Atualizações diretas em bancos de dados

Sistemas antigos, codificações diferentes

Inconsistência nos processos de carga:

1. Origem da informação é diversa
2. Mudanças no processo
3. Erros no processo

Limpeza e Tratamento de Dados

- Operação x Analítico
- Na operação o dado em seu formato individual não pode ser alterado para um valor padrão
- No analítico, o dado não tem valor individual , mas coletivo. Ele pode ser “corrigido” pelo bem do modelo.
- Ex: Cliente do plano de saúde tem data de nascimento vazia
- => Não podemos preencher com a mediana, pois isso influencia o valor do plano!