Mineração de Dados

Profa. Fernanda Maria da Cunha Santos

Universidade Federal de Uberlânida

Sistema de Informação – Campus Monte Carmelo

Site de referência: https://www.kaggle.com/code/edergillian/2-pre-processamento-de-dados

Panda

- Nesta aula vamos aprender a manipular e formatar os dados com Python usando o Pandas. Pandas é uma biblioteca para a análise de dados e manipulação da estrutura dos dados, usado para limpar, formatar e padronizar os dados.
- Antes de qualquer coisa, vamos importar a biblioteca Pandas
  - import pandas as pd

1. Como criar um dataset

In [2]:
```
# Vamos criar um DataFrame do resultado de uma votação.
pd.DataFrame({'Sim':[10,15], 'Nao':[21, 4]})
```

Show hidden output

In [3]:
```
# Uma Serie tem apenas uma lista. Vamos criar uma Serie:
pd.Series([10,15])
```

1.1) Crie um DataFrame que tenha 3 colunas: Produto | Quantidade | Preço. Como valores, tenha duas linhas com os seguintes valores na primeira linha: Chocolate | 200 | 3,00. Na segunda linha os valores: Banana | 80 | 0,50

1.2) Crie uma Serie que contenha o total dos produtos. Com os seguintes valores 600, 40

2. Como salvar um DataFrame como arquivo CSV

Para salvar como CSV um dataframe, basta usar a sintaxe

```
nomedodataframe.to_csv('nomearquivo.csv')
```

2.1) Salve o dataframe criado no exercício 1.1 num arquivo csv chamado vendas.csv. Descubra... onde o arquivo CSV foi salvo? Você consegue fazer o download deste arquivo?

3. Como abrir dados CSV já existentes

```
# Importa arquivo CSV
admissoes = pd.read_csv('../input/ipeadata/ipea_admissoes.csv')

# Visualiza as primeiras linhas do dataset
admissoes.head()
```

3.1) Faça a importação do arquivo CSV de demissões num novo DataFrame chamado demissoes e visualize suas primeiras linhas.

4. Como visualizar a estrutura dos dados

- df.shape: (quantidade de linhas, quantidade de colunas)
- df.index: descreve os indices
- df.columns: descreve as colunas
- df.info(): descreve o dataframe
- df.count(): conta o número de linhas que não tem valores NA

4.1) Mostre a estrutura de dados do dataset de demissoes. Ou seja, mostre sua quantidade de linhas e colunas, seus índices, descreva suas colunas, descreva o dataframe e mostre a quantidade de linhas de demissoes.

5. Como gerar o Sumário dos dados

- df.sum(): mostra a soma de valores
- df.cumsum(): mostra a soma acumulada dos valores
- df.min(): mostra o mínimo valor
- df.max(): mostra o valor máximo
- df.mean(): mostra a média
- df.median(): mostra a mediana
- df.describe(): mostra um sumário estatístico com a quantidade, média, desvio padrão (std), mínimo, primeiro quartil (25%), segundo quartil (50%), terceiro quartil (75%), valor máximo.

5.1) Para o dataset de admissoes, gere seus sumários e responda às seguintes perguntas:

- Qual a soma de valores de admissoes?
- Qual a diferença entre o resultado de sum() e cumsum() ?
- Qual o valor mínimo de admissões?
- Qual a maior data?
- Qual a média de admissoes?
- Qual a mediana de admissoes?
- Qual o desvio padrão de admissoes?

6. Como selecionar subconjuntos de dados

- .loc: retorna pelo nome (textual)
- .iloc: retorna pela posição (numérico)
- [ ]: retorna um elemento ou o range de elementos. Sendo: [linhas/indices, colunas]

6.1)Crie um dataset apenas com as colunas data e admissoes das admissoes do ano de 2017: a2017