DATA MINING

LABDATA

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Análise de Dados e Data Mining

Tema da Aula: **Aula 3 - Análise Exploratória de Dados com Python**

Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

Prof.: **Dino Magri**

- Contatos:
  - E-mail: professor.dinomagri@gmail.com
  - Twitter: https://twitter.com/prof_dinomagri
  - LinkedIn:http://www.linkedin.com/in/dinomagri
  - Site: http://www.dinomagri.com

Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

Currículo

- **(2014-Presente)** – Professor no curso de Extensão, Pós e MBA na Fundação Instituto de Administração (FIA) – www.fia.com.br
- **(2018-Presente)** – Pesquisa e Desenvolvimento de Big Data e Machine Learning na Beholder (http://beholder.tech)
- **(2013-2018)** – Pesquisa e Desenvolvimento no Laboratório de Arquitetura e Redes de Computadores (LARC) na Universidade de São Paulo – www.larc.usp.br
- **(2012)** – Bacharel em Ciência da Computação pela Universidade do Estado de Santa Catarina (UDESC) – www.cct.udesc.br
- **(2009/2010)** – Pesquisador e Desenvolvedor no Centro de Computação Gráfica – Guimarães – Portugal – www.ccg.pt
- **Lattes:** http://lattes.cnpq.br/5673884504184733

Material das aulas

- Caso esteja utilizando seu próprio computador, realize o download de todos os arquivos e salve na **Área de Trabalho** para facilitar o acesso.
  - Lembre-se de instalar os softwares necessários conforme descrito no documento de Instalação (**`InstalaçãoPython3v1.2.pdf`**).

- Nos computadores da FIA os arquivos já estão disponíveis, bem como a instalação dos softwares necessários.

Conteúdo da Aula

- Objetivo
- Análise Exploratória de Dados
- Referências Bibliográficas

Conteúdo da Aula

- **Objetivo**
- Análise Exploratória de Dados
- Referências Bibliográficas

Objetivo

- O objetivo dessa aula é aplicar os conceitos sobre a biblioteca Pandas para realizar a **análise exploratória de dados**.

Conteúdo da Aula

- Objetivo
- **Análise Exploratória de Dados**
- Referências Bibliográficas

Análise Exploratória de Dados

- A finalidade da Análise Exploratória de Dados (AED) é examinar os dados previamente à aplicação de qualquer técnica estatística ou de aprendizagem de máquina.
- **Por que é importante realizar a análise exploratória dos dados?**

Análise Exploratória de Dados

- Para aplicar os conceitos envolvidos na análise exploratória iremos utilizar o conjunto de dados de Preços de casas no Reino unido.

- É um desafio do Kaggle - https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Análise Exploratória de Dados

- Essa base, tem muitas informações relacionadas as casas que estão sendo analisadas.
- Para facilitar o estudo esse conjunto foi modificado para conter **23 atributos** (colunas).
- Essas colunas representam as características das casas.

Análise Exploratória de Dados

- O conjunto de dados contem 3 arquivos:
  - `preco_casas.csv` – Conjunto de dados com as características das casas, bem como o preço.
  - `descricao_conjunto_dados.txt` – Descrição das variáveis do conjunto de dados.
  - `descricao_codigos.csv` – Respectivos códigos e suas descrições para facilitar o entendimento.

</> Abra o arquivo **"aula3-parte1-dados.ipynb"**

Análise Exploratória de Dados

- Variável Qualitativa
  - Nominal: sexo, estados brasileiros
  - Ordinal: tamanho (pequeno, médio, grande), nível de escolaridade
- Variável Quantitativa
  - Contínua: salário, cotação do dólar
  - Discreta: número de clientes, número de carros, número de unidades vendidas

Análise Exploratória de Dados

- Medidas de posição: valor ao redor do qual os dados estão distribuídos
  - Máximo
  - Mínimo
  - Moda
  - Média
  - Mediana
  - Quartis (Q1 e Q3)

</> Abra o arquivo **"aula3-parte2-aed.ipynb"**

Análise Exploratória de Dados

- Medidas de dispersão: A finalidade é encontrar um valor que resuma a **variabilidade** de um conjunto de dados
  - Amplitude: diferença entre o valor máximo e o valor mínimo
  - Intervalo-Interquartil – É a diferença entre o terceiro e o primeiro quartil (Q3 – Q1)
  - Variância: média dos quadrados dos desvios em relação à média aritmética
  - Desvio Padrão: mede a variabilidade independente do número de observações e com a mesma unidade de medida da média.
  - Coeficiente de Variação: mede a variabilidade numa escala percentual independente da unidade de medida ou da ordem de grandeza da variável.

</> Abra o arquivo **"aula3-parte2-aed.ipynb"**

Análise Exploratória de Dados

- Dentro da análise exploratória dos dados é muito importante **visualizar dados graficamente** para termos uma visão mais apropriada das variáveis.

Análise Exploratória de Dados

- Alguns gráficos que facilitam o entendimento das variáveis são:
  - Histograma (Distribuição)
  - Diagrama de dispersão (Relação entre as variáveis)
  - Box-plot (Diferenças entre grupos)
  - Gráficos de linhas
  - Gráfico de simetria
  - Entre outros

Análise Exploratória de Dados

- Para gerar alguns desses gráficos, iremos utilizar alguns bibliotecas que facilitam a exploração visual dos dados.
  - Matplotlib
  - Seaborn
  - Pandas-profiling

Matplotlib

- É uma das bibliotecas de visualização mais antiga do Python (2002), porém muito utilizada ainda.
- Funciona muito bem para realizarmos analises iniciais no dados, ter uma noção do que temos. Porém ela não é muito útil para a criação de gráficos com qualidade de publicação rápida e fácil.
- Ela é muito poderosa, porém complexa!
- Galeria de exemplos: http://matplotlib.org/examples/index.html

Seaborn

- É uma das bibliotecas de visualização baseada no matplotlib.
- Fornece uma interface de alto nível para criar gráficos de maneira simples e com informações estatísticas.
- Criado por Michael Waskom em meados de 2012.
- Galeria: https://seaborn.pydata.org/examples/index.htm
- Documentação: https://seaborn.pydata.org/tutorial.html
- Código fonte: https://github.com/mwaskom/seaborn/

Pandas Profiling

- É uma biblioteca que gera relatório de perfil de um DataFrame pandas.
- A função `df.describe()` do pandas tem bastante informação, porém são informações básicas para análise exploratória de dados.
- Para cada coluna as seguintes estatísticas (se relevante para o tipo da coluna) são apresentadas em um relatório HTML interativo:
  - Essenciais: tipos, valores únicos, valores ausentes.
  - Estatística como valor mínimo, Q1, mediana, Q3, máximo, entre outros.
  - Estatística descritivas como média, moda, desvio padrão, coeficiente de variação, entre outros.
  - Valores mais frequentes
  - Histograma
  - Correlações
- Documentação: https://github.com/pandas-profiling/

Análise Exploratória de Dados

</> Abra o arquivo **"aula3-parte3-visualizacao.ipynb"**

Conteúdo da Aula

- Objetivo
- Análise Exploratória de Dados
- Referências Bibliográficas

Conteúdo da Aula

- Objetivo
- Análise Exploratória de Dados
- **Referências Bibliográficas**

Referências Bibliográficas

- **Use a Cabeça! Python** – Paul Barry - Rio de Janeiro, RJ: Alta Books, 2012.

- **Use a Cabeça! Programação** – Paul Barry & David Griffiths – Rio de Janeiro RJ: Alta Books, 2010.

- **Aprendendo Python: Programação orientada a objetos** – Mark Lutz & David Ascher – Porto Alegre: Bookman, 2007

Referências Bibliográficas

- **Python Data Science Handbook** – Jake VanderPlas, USA: O'Reilly, 2016.

- **Python: Data Analytics and Visualization** – Phuong Vo.T.H, Martin Czygan, Ashish Kumar, Kirthi Raman, Packt, 2017.

- **Python Para Análise de Dados** – Wes McKinney, O'Reilly, 2018.

Referências Bibliográficas

- **Python for kids – A playful Introduction to programming** – Jason R. Briggs – San Francisco – CA: No Starch Press, 2013.

- **Python for Data Analysis** – Wes McKinney – USA: O'Reilly, 2013.

- **Python Cookbook** – David Beazley & Brian K. Jones – O'Reilly, 3th Edition, 2013.

- As referências de links utilizados podem ser visualizados em http://urls.dinomagri.com/refs