UNIVERSIDADE DE SÃO PAULO - ICMC
Departamento de Ciências de Computação
SCC-0275 – Introdução à ciência de dados – Graduação - 2° Sem /2018
PROFa: Roseli Aparecida Francelin Romero
Estagiário PAE: Daniel Moreira Cestari

Aula 03 - Exploração de dados - Parte I

Para as atividades da prática vocês devem escolher 1 dataset dentre:

- Titanic
- Diabetes
- Breast Cancer
- Wine
- Boston House Pricing

E escolher um conjunto de dados dentre os seguintes:

- Agro
- Instituições de ensino básico
- Algum dos microdados do censo superior 2014.

Todos os conjuntos de dados estão no arquivo datasets.zip disponível no tidia.

Os conjuntos de dados Diabetes, Breast Cancer, Wine e Boston House Pricing estão disponíveis na biblioteca scikit-learn, vide exemplo de como carregar pela biblioteca no notebook auxiliar da prática (Aula03_EDA_pratica_I.ipynb).

Para cada conjunto escolhido, realizar as seguintes atividades:

1- Baseado na descrição do conjunto de dados formule algumas hipótese, perguntas que você acha que podem ser respondidas/entendidas com este conjunto de dados. Tente formular até 3 questões. Numere cada questão.

2- Escolha 5 atributos que você acha representar bem o problema e faça uma análise dos tipos de atributos e quais operações fazem sentido para cada um. Justifique a escolha desses atributos no contexto do conjunto de dados e com relação às hipótese/questões levantadas na pergunta 1.

3- Faça uma análise exploratória com as medidas vistas em aula (média, mediana, moda, ...) nos atributos anteriormente selecionados. Explique o porque você usou tal medida e como ela ajuda no entendimento do conjunto de dados. Sinta-se livre para utilizar outras medidas, mas explique porque você está usando elas. Se achar necessário, pode utilizar mais atributos na análise.

4- Após a análise exploratória reanalise as hipóteses/questões e comente se elas já podem ser respondidas, se elas não fazem mais sentido, eventualmente se elas podem ser reformuladas para questões

UNIVERSIDADE DE SÃO PAULO - ICMC
Departamento de Ciências de Computação
SCC-0275 – Introdução à ciência de dados – Graduação - 2° Sem /2018
PROFa: Roseli Aparecida Francelin Romero
Estagiário PAE: Daniel Moreira Cestari

5- Quais "problema(s)" o conjunto de dados apresenta e como isso impactou formulação e entendimento do conjunto de dados e o problema que ele representa. Argumente. Por problemas entenda valores ausentes, inconsistências e etc.

Entrega

A entrega da prática deve ser feita em formato de notebook do jupyter. Crie um notebook na mesma pasta que você descompactar o arquivo datasets.zip, nas primeiras linhas identifique os membros da dupla, salve o notebook com o nome pratica_1_nome1_nome2.

No corpo da prática primeiro identifique qual o conjunto de dados escolhido e comeÇe a responder as perguntas no corpo do notebook. Procure utilizar as caixas de texto para a discussão da discussão (o notebook tem caixas de texto e de código).

Quando terminarem submetam a prática no seu escaninho, os dois alunos precisam subir o arquivo do notebook.

O prazo de entrega é até sexta-feira às 18:00.

Descrição dos datasets utilizados na prática

Iris

https://archive.ics.uci.edu/ml/datasets/iris

https://en.wikipedia.org/wiki/Iris_flower_data_set

Diabetes

https://archive.ics.uci.edu/ml/datasets/diabetes

https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

UNIVERSIDADE DE SÃO PAULO - ICMC
Departamento de Ciências de Computação
SCC-0275 – Introdução à ciência de dados – Graduação - 2° Sem /2018
PROFa: Roseli Aparecida Francelin Romero
Estagiário PAE: Daniel Moreira Cestari

Breast Cancer

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Wine

https://archive.ics.uci.edu/ml/datasets/wine

Boston House Princing

https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

Titanic

https://www.kaggle.com/c/titanic/data

Enron Email Dataset

https://www.cs.cmu.edu/~enron/

https://en.wikipedia.org/wiki/Enron_Corpus

Reuters-21578 text categorization test collection

https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

Microdados Enem 2014

http://dados.gov.br/dataset/microdados-do-exame-nacional-do-ensino-medio-enem

Instituições Ensino Superior (SP)

http://dados.gov.br/dataset/instituicoes-de-ensino-basico

UNIVERSIDADE DE SÃO PAULO - ICMC
Departamento de Ciências de Computação
SCC-0275 – Introdução à ciência de dados – Graduação - 2° Sem /2018
PROFa: Roseli Aparecida Francelin Romero
Estagiário PAE: Daniel Moreira Cestari

Domínios .gov

http://dados.gov.br/dataset/dominios-gov-br

Agro

https://drive.google.com/file/d/14jcPwUUS-ldXdwYBgo1CHpJTEuXD-EDY/view?usp=sharing

Funções do pandas que podem ser utilizadas na prática

- mean(axis=None, skipna=None, level=None, numeric_only=None)
- median(axis=None, skipna=None, level=None, numeric_only=None)
- mode
- quantile(q=0.5, interpolation='linear')
- unique
- describe(percentiles=None, include=None, exclude=None)

No notebook auxiliar da prática (Aula03_EDA_pratica_I.ipynb) há vários exemplos de utilização de funções para carregar os datasets e como calcular algumas medidas. Há ainda um exemplo de resolução da prática no final o notebook.

Bibliografia para práticas de EDA:

- Charu C. Aggarwal. (2015). Data Mining: The Textbook. Capítulos 1, 2.
- Verónica Bolón-Canedo, Noelia Sánchez-Maroño, Amparo Alonso-Betanzos. (2015). Feature Selection For High-Dimensional Data. Capítulos 1, 2.
- Wilfried Grossmann, Stefanie Rinderle-Ma. (2015). Fundamentals of Bussines Intelligent. Capítulos 1, 2.1, 2.5, 2.6, 4.1, 4.3, 4.4, 4.5.
- Thomas A. Runkler. (2016). Data Analytics. Capítulos 1, 2.1, 2.2.

- Rajendra Akerkar, Priti Srinivas Sajja. (2016). Intelligent Techniques for Data Science. Capítulos 1, 2.
- **Doug Rose. (2016). Data Science. Capítulos 1, 3, 4, 8, 9, 11, 12, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24.**
- Max Bramer. (2016). Principles for Data Mining. Capítulos 1, 2.
- Steven S. Skiena. (2017). The Data Science Design Manual. Capítulo 1.

Recomendo a leitura de alguns desses capítulos para as próximas práticas, especialmente do livro Data Science do Doug Rose. Este livro é mais voltado para a comunicação das análises e como trabalhar com ciência de dados. Vários desses livros falam da mesma coisa mas de forma diferente, então é interessante ler mais de uma referência. Esses capítulos são sugestões dos conteúdos mais relacionados às aulas e práticas, sinta-se a vontade para os outros capítulos se tiver curiosidade.
Esses livros estão disponíveis no tidia numa pasta chamada bibliografia.