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 04 - Exploração de dados - Parte II

Para cada conjunto escolhido na prática anterior, realizar as seguintes atividades:

1- Listar novamente as perguntas/hipóteses em sua versão final da prática anterior, listar quais colunas foram selecionadas. Se alguma pergunta já foi respondida com a análise anterior refine a questão de modo que você ache que as análises realizadas hoje ajudarão a responder.

2- Executar as medidas em todos os atributos selecionados e responder as seguintes perguntas: Lembre-se que as análises devem ter como foco entender o dataset e responder as perguntas/hipóteses levantadas.

- Boxplot;

  O boxplot ajuda numa melhor compreensão dos dados?

  Porque e quando você usaria o boxplot?

  De acordo com o boxplot, há outliers nos seus atributos? se sim esses pontos marcados como outliers parecem mesmo outliers? explique.

  Considere outlier algum tipo de erro, de medição, ou digitação, ou inconsistência nos dados. Por exemplo, se um atributo estiver medindo altura de pessoas e um valor marcado como outlier é 19.5m, provavelmente é um erro de digitação a altura correta deveria ser 1.95m

- Intervalo máximo;

  O intervalo máximo pode induzir ao erro ao interpretá-lo? Em qual cenário? Mostrar no seu dataset se possível, se não for o caso para o dataset explicar em um cenário que ocorre.

- Momentos estatísticos centrados e originais, até o quarto.

  Quando é indicado utilizar os momentos estatísticos para EDA?

  Pelos valores dos momentos centrados, os atributos parecem seguir uma distribuição normal (nos slides tem uma indicação de quando um atributo parece seguir uma normal)? Explique.

  Se já existe uma função para alguma medida pode utilizar, caso contrário você precisa implementar tal método.

Responder da seguinte maneira:

Medida: boxplot

codigo

resposta das perguntas

3- 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

Entrega

A entrega da prática deve ser feita em formato de notebook do jupyter. Crie um notebook seu, nas primeiras linhas identifique os membros da dupla, salve o notebook com o nome pratica_2_nUSP1_nUSP2. A resolução desta prática deve seguir a mesma maneira da anterior. 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).

Funções utilizadas na prática

- boxplot(column=None, by=None, ax=None, fontsize=None, rot=0, grid=True, figsize=None, layout=None, return_type=None, **kwds)
- min(axis=None, skipna=None, level=None, numeric_only=None, **kwargs)
- max(axis=None, skipna=None, level=None, numeric_only=None, **kwargs)
- mean(axis=None, skipna=None, level=None, numeric_only=None, **kwargs)
- var(axis=None, skipna=None, level=None, ddof=1, numeric_only=None, **kwargs)
- skew(axis=None, skipna=None, level=None, numeric_only=None, **kwargs)
- kurt(axis=None, skipna=None, level=None, numeric_only=None, **kwargs)

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

No notebook auxiliar da prática (Aula04_EDA_pratica_II.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.