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 05 - Exploração de dados - Parte III

Caso os atributos que você escolheu nas práticas anteriores não é numérico, escolha outros para a realização desta prática.

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- Histograma

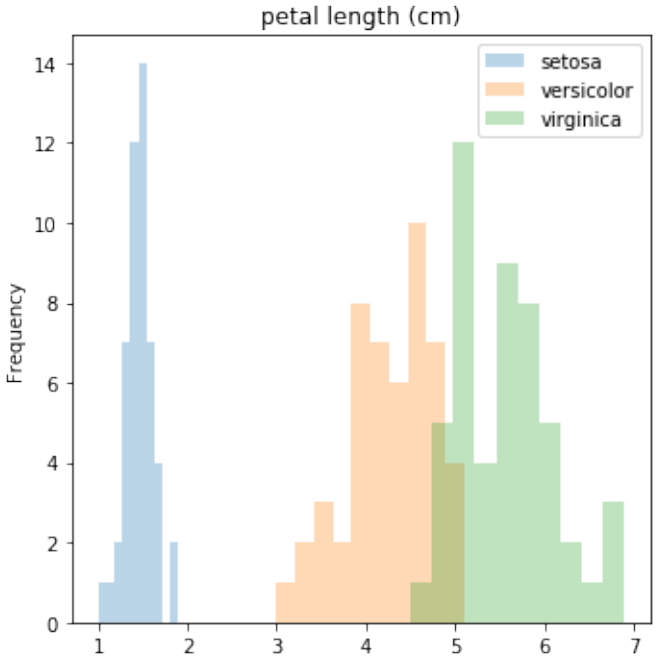
Plote os histograma dos atributos selecionados.

Como você fez a escolha do parâmetro de intervalo (*bin*)? Por faixa (para altura, por exemplo, intervalo de 5cm), por quantidade (cada intervalo com tamanho 5, 0-5, 5-10, ...) ... ? Por que?

O parâmetro *bin* pode ser complicado para ajustar corretamente e pode distorcer bastante o histograma. Uma regra prática é tentar garantir que cada intervalo tenha um número mínimo de exemplos, uns 30 em cada *bin* deve ser suficiente. Mesmo assim nos *bins* das pontas é difícil cumprir essa regra.

Se achar interessante para sua análise, divida os dados de algum atributo e plote o histograma para cada parte sobreposto.

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



Explique a relação do histograma com o boxplot, o que um mostra que o outro não? Como é possível inferir um do outro? Sugestão: plotar um boxplot e um histograma de um mesmo atributo pode ajudar a entender e responder a pergunta.

Análise multivariada.

Muitas vezes a análise univariada, análise de cada atributo individualmente, não é muito informativa. Por isso também utilizamos a análise multivariada, nesta prática utilizaremos uma análise bivariada, analisando 2 atributos ao mesmo tempo.

3- Execute a covariância e a correlação entre os atributos selecionados.

No seu caso, qual das duas é mais informativa? Explique.

Qual a diferença entre as duas e quando é mais indicado utilizar uma ao invés da outra?

De que forma elas ajudam, se é que ajudam, a responder suas perguntas?

4- Gráfico de dispersão (*scatter plot*)

Gere em um único gráfico os *scatter plots* dos atributos selecionados, como no slide. Mude a cor dos exemplos de maneira a ajudar na interpretação do problema e a responder suas perguntas.

Por exemplo, no iris se colorir os exemplos por classe é possível ver que a classe setosa não tem sobreposição com as outras duas para dois dos scatter plots.

Além da separação entre classes, o scatter plot permite uma inspeção visual rápida de grupos (*clusters*).

5- Explique como os gráficos permitem uma inspeção mais fácil e intuitiva de tendências, padrões e outliers? Se achar mais fácil use exemplos já explorados mas refaça os gráficos aqui para ilustrar seu ponto.

6- 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.

Tenha em mente que futuramente utilizaremos alguns algoritmos simples de classificação e suas questões poderão ser respondidas por eles. Isto não influenciará sua nota, é apenas para motivar que a fazer perguntas que poderão ser efetivamente respondidas. Até agora o que foi feita foi uma análise descritiva, vocês tentaram entender um pouco dos dados e do problema que os dados representa.

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. Submeta no tidia apenas o arquivo do notebook .ipynb, não crie pastas para separar as práticas, apenas faça o upload do arquivo .ipynb Não precisa fazer o upload dos datasets utilizados. 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 (o notebook tem caixas de texto e de código).

Funções utilizadas na prática

Pandas.DataFrame

- plot.hist(self, by=None, bins=10)
- cov(min_periods=None)
- corr(method='pearson', min_periods=1)

Matplotlib

- tight_layout(pad=1.08, h_pad=None, w_pad=None, rect=None)
- title(s, *args)
- legend(*args)

seaborn

- heatmap(data, vmin=None, vmax=None, cmap=None, center=None, robust=False, annot=None, fmt='.2g', annot_kws=None, linewidths=0, linecolor='white', cbar=True, cbar_kws=None, cbar_ax=None, square=False, xticklabels='auto', yticklabels='auto', mask=None, ax=None)
- pairplot(data, hue=None, hue_order=None, palette=None, vars=None, x_vars=None, y_vars=None, kind='scatter', diag_kind='hist', markers=None, size=2.5, aspect=1, dropna=True, plot_kws=None, diag_kws=None, grid_kws=None)

Nos exemplos eu utilizei heatmap e pairplot da seaborn mas a matplotlib também possui funções para esses tipos de plots.

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.