UDACITY

Back to Data Analyst Nanodegree

Investigate a Dataset

REVISÃO | HISTORY

Requires Changes

3 ESPECIFICAÇÕES NECESSITAM DE MUDANÇAS

Muito impressionante. Dá pra perceber seu trabalho duro refletido no seu projeto 😄 Seu código é muito sólido também, você só precisa de algumas modificações para continuar. Boa sorte na sua próxima submissão.

Não hesite em entrar em contato com seu mentor ou usar o canal slack para obter ajuda. Estamos aqui para ajudá-lo a ter sucesso 🏆

Funcionalidade do código

Todo o código é funcional e não produz erros quando executado. O código dado é suficiente para reproduzir os resultados descritos.

Uma coisa importante que você precisa fazer é usar comentários exclusivamente para documentar seu código, qualquer tipo de análise deve ser incluído em uma markdown cell (basta selecionar uma célula normal, clicar no tipo "Cell", depois "Cell Type" e finalmente selecione "Markdown")

```
In [12]: # Verificando se os nulos foram preenchidos
         df_titanic['Age'].isnull().sum()
Out[12]: 0

In [13]: # Estatísticas do Dataframe: Média de idade
         df_titanic['Age'].mean()
Out[13]: 29.699117647058763

In [15]: # Estatísticas do Dataframe: Média de idade dos sobreviventes
         df_titanic[df_titanic['Survived']== 1].mean()['Age']
Out[15]: 28.54977812177503

In [16]: # Estatísticas do Dataframe: Média de idade dos mortos
         df_titanic[df_titanic['Survived']== 0].mean()['Age']
Out[16]: 30.415099646415896

In [17]: # Estatísticas do Dataframe: Média de idade dos homens
         df_titanic[df_titanic['Sex']== 'male'].mean()['Age']
Out[17]: 30.50582424304206

In [500]: # Estatísticas do Dataframe: Média de idade das mulheres
          df_titanic[df_titanic_cleaned['Sex']== 'female'].mean()['Age']
Out[500]: 28.216730048707397
```

O projeto utiliza vetores Numpy, séries Pandas e DataFrames quando apropriado, em vez de listas e dicionários de Python. Sempre que possível, as operações vetorizadas e funções padrão são usadas em vez de loops.

Bom trabalho 👍
Sugestão

Aqui estão alguns métodos internos do Pandas que são muito úteis para explorar variáveis neste projeto:

1. Boolean-Indexing
2. Group-by
3. Value-Counts
4. Series.map

5. Working-with-text-data

O código utiliza funções para evitar repetição. O código contém bons comentários e nomes de variáveis, tornando-o fácil de ler.

Nós encorajamos você a fazer uso de funções para evitar códigos repetitivos, você pode usá-los, por exemplo, na data cleaning ou plotting de dados 😉

Qualidade das análises

O projeto estabelece claramente uma ou mais perguntas que são respondidas pela análise.

Perguntas muito pertinentes e importantes 👏🏽 Estas são algumas perguntas que eu normalmente sugiro:

1. Quantos passageiros havia em cada classe do navio?
2. Quantas mulheres havia a bordo do navio?
3. Quantos homens havia a bordo do navio?
4. Qual a média das idades das mulheres?
5. Qual a média das idades dos homens?
6. Qual a idade da mulher mais idosa?
7. Qual a idade do homem mais idoso?
8. Quantas mulheres havia em cada classe do navio?
9. Quantos homens havia em cada classe do navio?
10. Qual era a porcentagem de homens em cada classe do navio?
11. Qual a média de idade das mulheres em cada classe do navio?
12. Qual a média de idade dos homens em cada classe do navio?
13. Quantos passageiros viajavam acompanhados de membros da família?
14. Quantos passageiros viajavam desacompanhados de membros da família?
15. Quantos passageiros embarcaram em cada porto?
16. Quantos passageiros embarcaram em cada classe em cada porto?
17. Qual o valor médio do tíquete para cada classe?
18. Quantas vezes o tíquete da primeira classe era mais caro que o da segunda?
19. Quantas vezes o tíquete da segunda classe era mais caro que o da terceira?
20. Quantas vezes o tíquete da primeira classe era mais caro que o da terceira?
21. Qual o valor médio do tíquete em cada classe para aqueles que viajavam com membros da família?
22. Qual o valor médio do tíquete em cada classe para aqueles que viajavam desacompanhados da família?
23. Qual o valor médio do tíquete para crianças em cada classe?
24. Qual o valor médio do tíquete para mulheres em cada classe?
25. Qual o valor médio do tíquete para homens em cada classe?
26. Qual o valor médio do tíquete para idosos em cada classe?
27. Qual a idade média dos sobreviventes do naufrágio?
28. Qual a idade média das vítimas do naufrágio?
29. Qual a idade do sobrevivente mais novo?
30. Qual a idade da vítima mais nova?
31. Qual a idade do sobrevivente mais idoso?
32. Qual a idade da vítima mais idosa?
33. Qual a porcentagem de sobreviventes em cada classe do navio?
34. Qual a porcentagem de sobreviventes entre as crianças?
35. Qual a porcentagem de sobreviventes entre as mulheres?
36. Qual a porcentagem de sobreviventes entre os homens?
37. Qual a porcentagem de sobreviventes entre os idosos?
38. Qual a porcentagem de sobreviventes entre os que viajavam com a família?
39. Qual a porcentagem de sobreviventes entre os que viajavam desacompanhados?
40. Qual o valor médio dos tíquetes de sobreviventes em cada classe do navio?
41. Qual o valor médio dos tíquetes das vítimas em cada classe do navio?

Fase de preparação dos dados

O projeto documenta todas as alterações que foram feitas para limpar os dados, como manipulação dos valores ausentes.

Bom trabalho na implementação de uma fase de data wrangling
Sugestão

O aspecto mais importante da disputa de dados é limpar ou transformar os dados preparando-os para análise.

Uma questão principal é a falta de dados durante a análise, o que pode fornecer resultados de desvio / polarização. Felizmente, existem alguns métodos que o Pandas oferece para lidar com esses problemas:

- A primeira coisa a fazer é sempre Identificar os valores ausentes dentro do conjunto de dados. Os poucos passos depois disso explicam como lidar com os dados perdidos
- Se houver colunas com algumas linhas de dados ausentes, o Método Dropna poderá ser usado para descartar as linhas ausentes.
- Se houver linhas com dados ausentes, o método Fillna-method poderá ser usado de largá-los completamente (Esse método pode variar de acordo com os dados e o projeto)
- A opção final é se houver muitos valores ausentes dentro de uma coluna, é melhor soltar a coluna completamente usando o método Drop-column. /0.17.0/generated/pandas.DataFrame.drop.html)

A Disputagem de Dados não envolve apenas a identificação e o tratamento de valores ausentes, mas também envolve a transformação dos dados em um estado mais efetivo para direcionar a análise. Aqui estão outros métodos de manipulação:

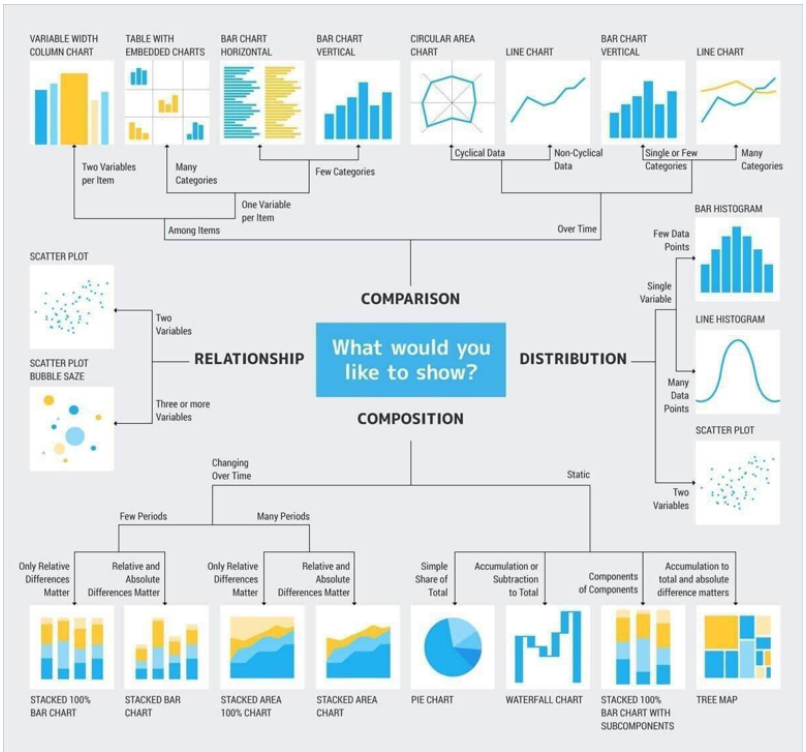
- Binning ou Cutting Agrupa valores contínuos ou numéricos em grupos menores ou "bins"
- Pandas-Dummies Transforma dados categóricos em variáveis dummy / indicadoras

Fase de exploração

O projeto investiga a(s) questão(ões) indicada(s) a partir de vários ângulos. Pelo menos três variáveis são investigadas.

Pelo menos dois tipos de gráficos são criados como parte das explorações.

Muito bom ! para projetos futuros, deixe-me recomendar-lhe estas ferramentas para escolher suas visualizações



Fase de conclusão

Avalie esta revisão

Os resultados da análise são apresentados de tal forma que quaisquer limitações são claras. A análise não indica nem sugere que uma alteração causa outra baseada unicamente em uma correlação.

Deveria haver uma subseção separada dentro da seção de conclusão chamada 'Limitações', na qual você teria que discutir as limitações desse conjunto de dados que poderiam ter afetado negativamente sua análise. Exemplos seriam valores nulos ou ausentes, se essa amostra for uma representação efetiva da população ou talvez você possa mergulhar mais fundo em sua análise com informações específicas adicionais.

As conclusões e limitações devem ter a seguinte estrutura:

Conclusões

Analisando os dados do acidente do Titanic, fica claro que os passageiros das classes superiores foram previlegiados no momento de embarcarem nos botes salva vidas.
Apenas 24% dos passageiros da 3ª classe sobreviveram, enquanto 47% se salvaram na 2ª classe e 63% da 1ª.
A idade média dos sobreviventes era de 29 anos.
Apenas 38% dos passageiros sobreviveram ao acidente, sendo a maior deles mulheres (68%).

Limitações

Numa primeira visão encontrou-se como fator limitante para a análise que algumas propriedades não possuiam valores para alguns do passageiros. Estas características são: Age, Cabin e Embarked.

- Medida tomada:

```
        * Age: Os valores faltantes desta característica foram preenchidos com base na idade média da me
sma classe de embarque do passageiro em questão.
        Cabin: Como esta característica não foi considerada na fase de análise optou-se por não alterá-l
a.
        * Embarked: O local de embarque foi preenchido com base na probabilidade de selecionar um dos lo
cais de embarque da mesma classe do passageiro em questão.
```

- Após ajustou-se o tipo de representação de algumas característica para facilitar o processo de análise. Alterou-se o tipo do sexo e local de embarque para uma representação numérica inteira.

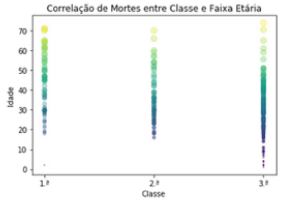
Fase de comunicação

É fornecido um raciocínio para cada decisão analítica, gráfico e resumo estatístico.

Nem toda análise é seguida de explicação. Por exemplo: Os gráficos.

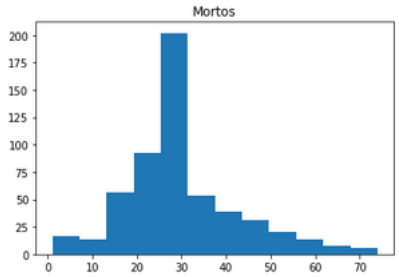
Correlação de Mortos entre Classe e Faixa Etária

```
In [312]: # Gráfico de correlação de morte entre Classe e Faixa Etária
plt.scatter(df_titanic_m['Pclass'], df_titanic_m['Age'], alpha=0.3, c=df_titanic_m['Age'], s=df_titanic_m['Age'])
plt.title('Correlação de Mortes entre Classe e Faixa Etária')
plt.xticks([1,2,3], ['1.ª','2.ª','3.ª'])
plt.xlabel('Classe')
plt.ylabel('Idade')
plt.show()
```



Relações envolvendo o Valor da Tarifa (Fare)

Por favor inclua uma ou duas sentenças falando sobre o que informação você pode retirar dos dados. O mesmo se aplica para as tabelas que você gerou. O que podemos concluir ao analisar estes valores?



Por favor, certifique-se de que cada grafico tenha as seguintes três características:

1. Um título
2. Nomes dos eixos
3. Etiquetas

REENVIAR PROJETO

BAIXAR PROJETO

Melhores práticas para sua resubmissão do projeto

Ben compartilha 5 dicas úteis para a revisão ressubmissão do seu projeto.

Assistir Vídeo (3:01)

RETORNAR

FAQ do Estudante