UDACITY

PROJECT

Investigate a Dataset

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

5 SPECIFICATIONS REQUIRE CHANGES

Você está no caminho certo, seu trabalho está ficando bom. Achei o seu relatório interessante e fiz questão de revisar com cuidado ponto a ponto. Peço que leia os comentários, faça as correções e submeta novamente seu projeto. Espero que possa contribuir no seu aprendizado. Continue o bom trabalho, e até a próxima submissão!

P.s.: O seu projeto caiu para os revisores da língua inglesa. Como você tinha escrito comentários em português, a revisão foi enviada para mim. Por favor, antes de submeter, verifique se a sua configuração está em português para que da próxima vez já caia para os revisores brasileiros e acelere o processo de revisão.

Funcionalidade do Código

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

Excelente trabalho! Código executa sem erros.

O projeto utiliza vetores Numpy e Series Pandas e DataFrames quando apropriado ao invés de listas e dicionários do Python. Sempre que possível, as operações vetorizadas e funções padrões são usados em vez de loops.

Foi utilizada a biblioteca `pandas` e funções padrões da biblioteca, evitando a utilização de listas, dicionários e loops desnecessários.

Gostaria de deixar como sugestão a utilização da função corr da biblioteca pandas. A função calcula a correlação entre todas as variáveis do DataFrame, não sendo necessário definir uma nova função para calcular a correlação.

Veja como ficaria simples o código utilizando essa função:

```
correlation = titanic_data.corr()
print "Correlation between age/gender and survival: ",correlation['Survived']['GA
class']
print "Correlation between cabin class and survival: ",correlation['Survived']['P
class']
```

O código faz uso das funções para evitar código repetitivo. O código contém bons comentários e nomes de variáveis, tornando-o fácil de ler.

Foram criadas funções definidas pelo usuário para evitar o código repetitivo como a função `classify` e `correlation`.

Porém não foi utilizado nenhum comentário no código. Comentários são importantes para tornar o código mais fácil de ler e facilitar a manutenção futura. O único comentário na função `correlation` parece ter sido copiado e não criado pelo usuário. 😄

Veja esse guia de boas práticas do google para python dicas de como inserir comentários em seu código.

Qualidade das Análises

O projeto estabelece claramente uma ou mais perguntas que atende à essas perguntas no resto da análise.

Excelente! A pergunta foi formulada logo no começo, junto com uma ótima introdução que contextualiza o leitor.

Fase de Data Wrangling

O projeto documenta todas as alterações que foram feitas para limpar os dados, tais como união de vários arquivos, manipulação dos valores ausentes, etc.

Aqui era esperado que fosse feita uma checagem e limpeza dos dados.

Por exemplo, existem valores nulos em algumas das variáveis? Existem valores diferentes do esperado para cada variável (como um valor para `Pclass` diferente de 1 a 3)?

Qual técnica utilizar para tratar os valores ausentes?

Esses são alguns pontos que você pode explorar. Não esqueça de documentar todas as alterações realizadas.

Fase de Exploração

O projeto investiga a(s) questão(ões) indicada(s) a partir de vários ângulos. Pelo menos três variáveis são investigados usando tanto variável simples (1d) e explorações (2d) de múltipla variáveis.

Muito bom! O relatório inclui análises univariadas, como o histograma de Age, bivariadas como a análise entre Age e Survived e Pclass e Survived, e também com três variáveis como GAclass, Survived e Pclass.

Criar novas variáveis a partir de variáveis existentes também é uma estratégia interessante, e muitas as vezes expõe padrões ocultos nos dados, como você fez criando a variável GAclass. Outra variáveis também poderiam ter sido criadas, como por exemplo, a partir do nome, que é uma variável textual, podemos extrair a família (pelo sobrenome) e o status social (pelo prefixo Mr., Ms., Miss, etc). Essas duas variáveis também correlação significativa com a variável Survived.

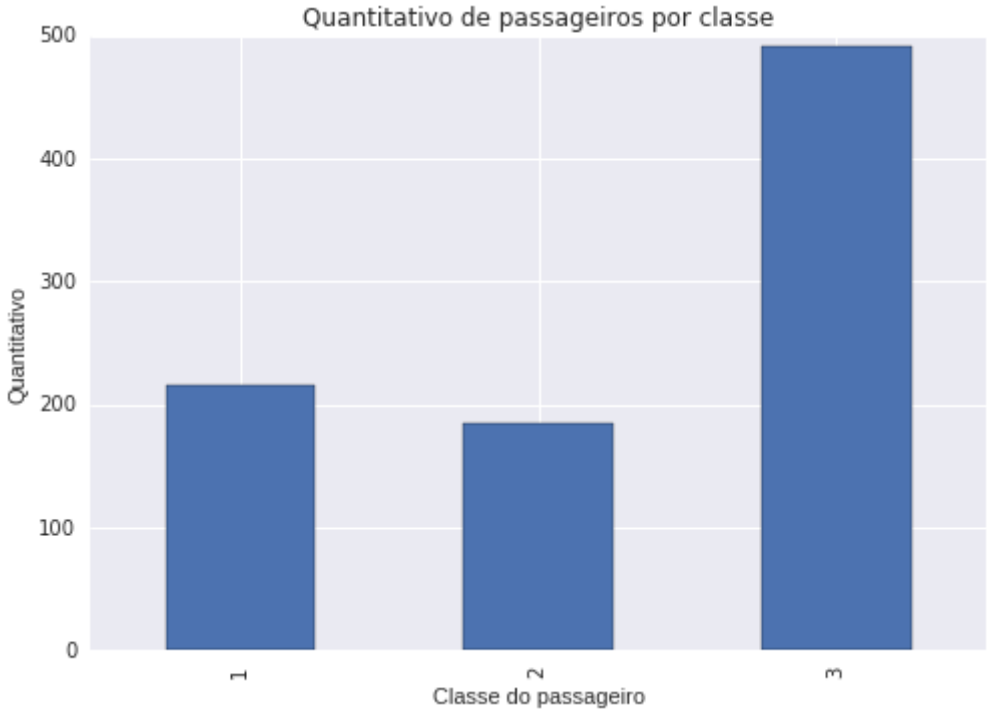
As visualizações do projeto são variadas e mostram comparações e tendências. Estatísticas relevantes são computadas ao longo da análise, quando uma inferência é feita sobre os dados.

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

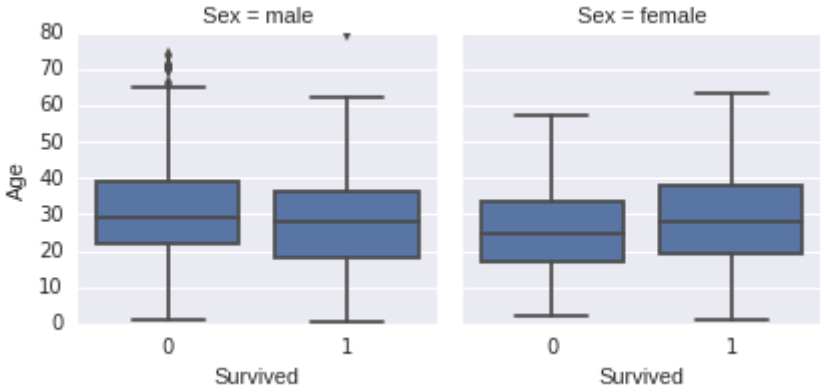
Muito bom! Os gráficos são interessantes. Vamos ter a oportunidade de trabalhar mais a fundo com visualização em vários outros projetos a frente no curso, mas tem alguns pontos, como a escolha do gráfico, que é interessante começar a trabalhar a partir de agora.

- Histogramas são recomendados para variáveis com valores contínuos. Observe no gráfico da célula 49 como o eixo x ficou estranho. O gráfico dá a entender que a variável Pclass é contínua, veja os valores 1.5 e 2.5. Para variáveis categóricas um gráfico de barras é mais adequado. Veja como ficaria

um gráfico de barras para o caso:



- O mesmo vale para os gráficos da célula 106, repare no eixo da variável Survived. Além disso o gráfico está muito confuso e sem uma explicação para seu uso. Porque foi plotada a regressão linear (regplot)? Para a análise de Age, Survived e Sex poderia ser utilizado um boxplot por exemplo. Veja como ficaria:



- Todos os gráficos devem conter título e descrição dos eixos para facilitar a interpretação.
- Inclua também estatísticas que corroborem com as interpretações dos gráficos. Os gráficos são importantes ferramentas visuais mas é importante que sejam acompanhados de estatísticas relevantes.

Fase de Conclusão

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

Você fez conclusões corretas porém é preciso elaborar mais a justificativa. Alguns pontos precisam de correção.

Por exemplo, o coeficiente de correlação entre GAclass e Survived mede a correlação entre as duas variáveis mas não explica que mulheres tem mais chance de sobreviver que homens. O mesmo vale a classe

econômica. Talvez seja mais interessante calcular o percentual de cada um dos valores e compará-los (por exemplo o percentual de mulheres e crianças que sobreviveram em relação aos homens).

Importante deixar claro as limitações da análise. Alguns pontos que devem ser ressaltados:

- Deixar claro que correlação não significa causalidade. O fato da variável classe estar correlacionado com a variável Survived pode ser uma correlação sem nenhum valor causal - outros fatores podem ter influenciado na chance do passageiro sobreviver, e a correlação de Pclass e Survived ser um fato casual e sem relevância para análise. Veja alguns exemplos de como confundir correlação com causalidade pode dar errado: http://www.tylervigen.com/spurious-correlations
- Informar que é apenas uma teoria com base nas observações, com nenhum valor estatístico. Nenhum teste foi aplicado para saber se as diferenças observadas poderiam ser causadas por mera chance. Se eu jogo dois dados comuns por 5 jogadas, no primeiro tenho uma média de 2, e no segundo de 4, com qual grau de certeza posso afirmar que os dados são diferentes? Esse trabalho não exige que sejam aplicados testes estatísticos, apenas que seja reconhecido a limitação da análise de dados sem uma análise estatística rigorosa.

Para ficar ainda melhor a conclusão, sugiro ainda discutir a possibilidade de trabalhos futuros. Que outras variáveis podem ser analisadas? Como contornar as limitações das conclusões acima ? É possível construir um modelo preditivo de quem sobreviveu ou não ao acidente do Titanic?

Fase de Comunicação

Raciocínio é fornecido para cada decisão analítica, gráfico e resumo estatístico.

Inclua uma interpretação de cada gráfico, estatística ou decisão que fizer durante a análise. É importante deixar claro qual a contribuição de cada etapa para a análise e o que é possível inferir a partir de cada uma delas.

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Student FAQ