Projeto - Análise de Dados

IMD0601 - Bioestatística - Instituto Metrópole Digital - UFRN

Prof. Tetsu Sakamoto (IMD) e Profa. Beatriz Stransky (DEB)

Instruções:

- Este trabalho consiste no desenvolvimento de uma análise de dados completa, como ilustrado pelo fluxograma de Rolemund e Wickham no livro R para Data Science (https://r4ds.had.co.nz/) - importar, organizar, modelar, comunicar.
- Este trabalho avaliará o conhecimento e capacidade do aluno em utilizar métodos estatísticos para manipulação e análise de dados em ambiente R.
- Realize todos os procedimentos em ambiente R.
- Apresente um script com as etapas de análise e comandos em R utilizados;
- Utilize a biblioteca ggplot2 do R para gerar os gráficos.
- Não será permitido a utilização do mesmo conjunto de dados por mais de um grupo.
- **Será considerado plágio a utilização de análises e vinhetas já publicadas.**
- A 1a etapa deste trabalho será apresentada no dia 28/09/2020 e o código em R deverá ser submetido pelo SIGAA até o dia 27/09/2020 às 23:59h.
- A 2a etapa deste trabalho será apresentada no dia 11/11/2020 e submetida até o dia 10/11/2020 às 23:59h.
- O trabalho completo será apresentado nos dias 07 e 09/12/202, devendo ser submetido pelo SIGAA até dia 06/12/202 às 23:59h.

Após a seleção do conjunto de dados, o trabalho será desenvolvido em 3 etapas.

1a Etapa:

- Leia o artigo FiveThirtyEight original e quaisquer outros materiais complementares necessários para entender o contexto.
- Examine o arquivo de ajuda do conjunto de dados para saber a quais variáveis eles têm acesso.
- Realize uma análise de dados exploratória dos dados, inspecionando visualmente os valores brutos, computando estatísticas de resumo e criando visualizações.
- Elabore uma apresentação de cerca de 6-7 minutos, com a seguinte estrutura:
    - Capa

    - Sobre o dataset (introdução sobre o tema e objetivo pelo qual o dataset foi criado);
    - Estrutura do dataset;
    - Análises de estatística descritiva que podem auxiliar no alcance do objetivo;
    - Visualização de dados que pode auxiliar no alcance do objetivo.

2a Etapa:

- Elabore as hipóteses a serem testadas;
- Realize os testes adequados para verificar a plausibilidade das mesmas;
- Discuta sobre os seus resultados obtidos, procurando por outras fontes que dêem suporte ou não a sua descoberta.

3a Etapa:

- Execute regressões com uma das variáveis de resultado para investigar quais características estão associadas à variável dependente.
- Resuma suas descobertas por escrito, discutindo as implicações e limitações da sua análise. Caso haja outras fontes de dados ou informações, não deixe de citá-las corretamente.
- Elabore uma apresentação completa de cerca de 15 min.

Sobre o conjunto de dados

Os dados a serem utilizados serão os disponibilizados no pacote 'fivethirtyeight' (https://cran.r-project.org/web/packages/fivethirtyeight/). Este pacote apresenta 128 conjuntos de dados publicados pelo site 'FiveThirtyEight', um site de jornalismo baseado em dados fundado por Nate Silve, de propriedade da Disney/ESPN que reporta sobre política, economia, ciência, esportes e outros eventos atuais.

O pacote 'fivethirtyeight' foi desenvolvido com o objetivo de expor o aluno de disciplinas introdutórias de estatística/ciência de dados a todos os elementos do ciclo de análise, desde a importação de dados até a comunicação dos resultados. Desta forma, o pacote consegue harmonizar dois objetivos contrários, de 'minimizar os pré-requisitos de pesquisa' ao mesmo tempo que usa dados do mundo real. De acordo com os próprios autores, o conjunto de dados apresentados é:

- Rico o suficiente para responder a perguntas significativas;
- Real o suficiente para garantir que haja contexto;
- Realista o suficiente para transmitir aos alunos que os dados existentes "na natureza" geralmente precisam de preparação / pré-processamento.