INSTITUTO INFNET
MIT BIG DATA
Analytics com R
Prof.: Cassius Figueiredo

AVALIAÇÃO

INSTRUÇÕES

- Cada grupo de alunos deve apresentar um relatório com as respostas para os cinco problemas propostos.
- Todas as informações disponíveis estão abaixo e, caso seja necessário, os grupos deverão complementar as informações por sua conta, deixando claramente indicadas as referências.
- A abordagem de solução dos problemas é livre, porém apenas o uso de linguagem R será permitido.
- O relatório final deverá ser acompanhado dos códigos fartamente comentados para documentação das soluções apresentadas.

1) Titanic

Leia o arquivo **titanic_train.csv,** com os dados de alguns passageiros do acidente do Titanic. A base de dados possui a seguinte estrutura:

Variable Definition Key:

- survival - Survival 0 = No, 1 = Yes
- pclass - Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd
- sex - Sex
- Age - Age in years
- sibsp - # of siblings / spouses aboard the Titanic
- parch - # of parents / children aboard the Titanic
- ticket - Ticket number
- fare - Passenger fare
- cabin - Cabin number
- embarked - Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes:

- pclass: A proxy for socio-economic status (SES)
    - 1st = Upper
    - 2nd = Middle
    - 3rd = Lower
- age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- sibsp: The dataset defines family relations in this way...
    - Sibling = brother, sister, stepbrother, stepsister
    - Spouse = husband, wife (mistresses and fiancés were ignored)
- parch: The dataset defines family relations in this way...
    - Parent = mother, father
    - Child = daughter, son, stepdaughter, stepson
    - Some children travelled only with a nanny, therefore parch=0 for them.

Agora, execute o que é pedido abaixo:

- Quantas variáveis e observações possui o arquivo?
- Quais são as classes das variáveis?
- Qual é a média das dos preços dos tickets?
- Faça um filtro na tabela e crie dois outros data frames. Um para o gênero masculino e o outro para o gênero feminino.
- Crie duas listas, uma para informações do dataframe do gênero feminino e outra para o gênero masculino. Cada lista deve ser composta de:
    - Número total de passageiros
    - Número de sobreviventes
    - Número de passageiros na primeira classe
    - Preço do ticket
    - Número de parentes\filhos
- Com base nas listas criadas, responda:
    - Qual gênero teve o maior número de pessoas embarcadas?
    - Qual gênero sobreviveu mais?
    - Qual gênero teve a maior média do número de parentes?

2) Human Development Index (HDI)

Leia a base **Human_development_index_HDI.csv** com os dados da evolução do IDH (Índice de Desenvolvimento Humano) dos países. Use a tabela abaixo como referência para avaliação do IDH:

Valor IDH	Classificação
IDH ≤ 0.534	Baixo
$0.534 <$IDH ≤ 0.710	Médio
$0.710<$IDH ≤ 0.796	Alto
IDH >0.796	Muito Alto

Agora, execute o que está sendo pedido abaixo:

- Crie uma função que classifique os países (em uma coluna extra) em 2014 de acordo com a tabela acima.
- Qual país cresceu mais em relação à 2013?
- Qual país caiu mais em relação à 2013?
- Quantos países estão com classificação baixa?
- Qual é a posição do Brasil?

3) Dados ANP

Crie um objeto chamado anp que receba a leitura de dados **dados_anp2.csv**.

Dica 1: na leitura use o argumento stringsAsFactors = FALSE.
Dica 2: transforme a coluna PRECO_COMPRA em númerica, com a função as.numeric().
Dica 3: faça os valores nulos de PRECO_COMPRA receberem NA.

- Use as funções **summary()** e **str()** para entender a base. Faça o que se pede:
- Quantos preços foram coletados?
- Crie uma tabela com a frequência de postos por combustível, atribua essa tabela à variável “quantidade_postos”.
- Qual combustível teve menos preços coletados? (Interpretar este resultado!)
- Qual é o posto com menor preço de venda? É confiável essa fonte (Dica: olhe para o fornecedor e a bandeira.)
- Crie o dataframe “dados_etanol”, como um filtro do dataframe ***anp***. Apresente “dados_etanol” por UF e média dos preços de venda do etanol.
- Qual é o estado com a menor média de preços de venda do etanol. Isso faz sentido?
- Exporte este último dataframe no formato CSV.

4) Dietas de galinhas

Essa base contém dados de galinhas divididas em quatro dietas distintas ao longo do tempo. Ela pode ser carregada diretamente do R com a instrução **chickwts**. Mais detalhes sobre a base abaixo:

Description
An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens.

Usage
chickwts

Format
A data frame with 71 observations on the following 2 variables.

- **Weight**: a numeric variable giving the chick weight.
- **Feed**: a factor giving the feed type.

Details
Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights in grams after six weeks are given along with feed types.

Source
Anonymous (1948) Biometrika, 35, 214.

References
McNeil, D. R. (1977) Interactive Data Analysis. New York: Wiley.

Diante dessa base, faça:

- Qual dieta apresentou maior média de peso?
- Qual dieta apresentou maior homogeneidade dos pesos?
- Qual dieta seria escolhida para aumento de peso?

5) Um estudo sobre salários de jogadores de Baseball

Para executar esse exercício você precisa:

- Instalar o pacote ISLR.
- chamar a base **“Hitters”**.
- Descrição da base.

A partir desta base, suponha que você precise fazer um rápido estudo sobre o salário dos jogadores de baseball. A intenção é entender quais seriam os fatores que mais influenciam no salário dos jogadores. Atenção! É um relatório para a diretoria, portanto tem que ser visual.

Crie um relatório simples com as seguintes informações:

- Histograma e boxplot dos salários.
- Analise os valores faltantes
- Qual liga apresenta os maiores salários?
- Qual divisão apresenta os maiores salários?
- Quais variáveis quantitativas apresentam maior correlação com o salário?
- Apresente suas conclusões, relacionadas à questão da influência dos fatores apresentados no salário.