3ª Lista de exercícios utilizando o R

Na área de Ciência dos Dados e Machine Learning, existem várias bases de dados **clássicas**, ou seja, bases de dados que têm sido amplamente utilizadas para validar modelos estatísticos e computacionais. A base de dados disponível no arquivo "boston_corrected.txt", e conhecida como *Boston Housing data set*, foi publicada originalmente por Harrison and Rubinfeld (1978).

Neste exercício, o objetivo é avaliar potenciais variáveis preditoras para a estimação do preço de valores de imóveis na região de Boston/USA. Uma descrição sucinta de cada uma das variáveis é apresentada a seguir:

Variável	**Descrição**
CRIM	Taxa de crime per capita
ZN	Proporção de terrenos residenciais destinados a lotes com mais de 25.000 pés quadrados.
INDUS	proporção de acres de negócios não varejistas por cidade.
CHAS	Variável binária indicando a proximidade ao Rio Charles (1, se o setor contiver o rio; 0, caso contrário).
NOX	Concentração de óxidos nítricos (partes por 10 milhões)
RM	Número médio de quartos por habitação.
AGE	Proporção de unidades construídas antes de 1940 e ocupadas pelo proprietário.
DIS	Distância ponderadas com relação a cinco centros de emprego de Boston.
RAD	Índice de acessibilidade a rodovias radiais
TAX	Taxa de imposto predial de valor integral por US $ 10.000.
PTRATIO	Taxa aluno-professor por cidade.
B	1000 (Bk - 0,63) ^ 2 onde Bk é a proporção de negros por cidade
LSTAT	% população de baixa renda
MEDV (**y**)	Valor mediano (preço), em US $ 1000, das residências ocupadas pelo proprietário.

Utilizando a base de dados indicada, faça uma análise dos dados e procure encontrar as variáveis mais preditivas para o preço dos imóveis (MEDV). As seguintes análises são indicadas:

a) Faça uma análise descritiva da variável resposta (MEDV), incluindo estatísticas descritivas, histograma e boxplot. Faça uma análise crítica dos resultados obtidos.

b) Utilizando o pacote **corrplot**, faça um gráfico de correlação linear considerando todas as variáveis da base de dados. Procure identificar a partir do gráfico, a variável preditora que apresenta a maior correlação linear para com a variável resposta (MEDV).

c) Ajuste modelos de regressão linear simples considerando cada uma das 12 variáveis preditoras. Para isso, o pacote **exploreR** é indicado. Faça uma análise dos resultados encontrados.

d) Ajuste um modelo de regressão linear múltipla e faça uma análise dos resíduos. É possível afirmar que os resíduos seguem uma distribuição normal? É possível afirmar que os resíduos são homocedásticos? Inclua os gráficos de análise dos resíduos no seu relatório. Faça uma análise do modelo encontrado.

Elabore o seu relatório utilizando o **Rmarkdown** e envie o documento em formato **PDF** para avaliação pelo sistema minha.ufmg.

Referência bibliográfica

Harrison, David, and Rubinfeld, Daniel L., Hedonic Housing Prices and the Demand for Clean Air, Journal of Environmental Economics and Management, Volume 5, (1978), 81-102.