Disciplina: Minerção de Dados
Prof.: João Batista M. Pereira
e-mail: joao@dme.ufrj.br
Entrega: 24/02/2021

Atividade 1

O conjunto de dados a ser analisado corresponde a registros públicos de vendas de casas feitas de maio de 2014 a maio de 2015 no Condado de King, no estado de Washington, EUA.

O banco de dados está dividido em dois arquivos: `kc_housing_1.csv` e `kc_housing_2.csv`. No primeiro arquivo, as variáveis são:

- `id` - identificação única da casa vendida;
- `date` - data da venda;
- `price` - preço da venda;
- `bedrooms` - quantidade de quartos;
- `bathrooms` - quantidade de banheiros, em que .5 corresponde a banheiro sem chuveiro;
- `sqft_living` - pés quadrados do interior da casa;
- `sqft_lot` - pés quadrados do terreno;
- `floors` - quantidade de andares;
- `waterfront` - 1, se casa é à beira-mar e 0, caso contrário;
- `view` - um índice de 0 a 4 indicando a qualidade da visão da propriedade;
- `condition` - um índice de 0 a 5 indicando a condição do apartamento;
- `date2` - data da venda ("ano-mês").

No segundo arquivo as varáveis são:

- `id` - identificação única da casa vendida;
- `date` - data da venda;
- `grade` - um índice de 1 a 13, em que 1-3 corresponde a um baixo nível de construção e design, 7 corresponde a um nível médio de construção e design e 11-13 corresponde a um alto nível de construção e design;
- `sqft_above` - pés quadrados do interior da casa acima do nível do solo;

- `sqft_basement` - pés quadrados do interior da casa abaixo do nível do solo;
- `yr_built` - ano em que a casa começou a ser construída;
- `yr_renovated` - ano da última reforma da casa;
- `zipcode` - código postal da casa;
- `lat` - latitude;
- `long` - longitude;
- `sqft_living15` - pés quadrados do interior das 15 casas mais próximas;
- `sqft_lot15` - pés quadrados do terreno das 15 casas mais próximas.

1. Importe as bases de dados nos arquivos para o `R` e junte-as em um único *data frame* com todos os registros e todas as variáveis. Lembre-se que as linhas correspondem aos registros de vendas e uma casa pode ser vendida mais de uma vez.

   Escolha pelo menos três variáveis do banco de dados e faça uma análise exploratória. Utilize gráficos tais como histogramas, *box-plots*, gráfico de barras e/ou gráfico de setores. Para as variáveis quantitativas, pode-se também calcular algumas medidas-resumo tais como média, mediana, desvio padrão.

2. Para as variáveis quantitativas, calcule a matriz de correlação e faça um gráfico em que se possa facilmente visualizar as correlações entre as variáveis. Em seguida, faça diagramas de dispersão para os dois pares de variáveis mais correlacionados.

3. Verifique graficamente a associação entre o preço e pelo menos três outras variáveis do banco de dados. Investigue a melhor forma de visualização: gráfico de dispersão, *box-plots*.

4. Agrupe os dados por data (mês e ano) e calcule a média do preço para cada mês. Em seguida, faça um gráfico de linha com a variação do preço ao longo do tempo.

   A partir dos preços e pés quadrados do interior da casa, calcule, para cada registro, o preço médio do pé quadrado. Em seguida, faça um gráfico de linha com a variação do preço médio ao longo do tempo.

5. Calcule os quartis do preço (valores que dividem os preços em quatro faixas, cada um com 25% dos registros). Verifique em que faixa cada preço do banco de dados se encontra e associe a ele um cor diferente. Em seguida, faça um gráfico em três dimensões com a latitude, longitude e o preço (com as cores correspondentes à sua respectiva faixa).

A atividade deve ser feita no `R Markdown` *e entregue em HTML ou PDF com os códigos explicitados e as análises comentadas. Os gráficos devem ser explicativos, com nomes corretos nos eixos, por exemplo.*