Project 2.1: Limpeza de dados

Passo 1: Entendimento do Negócio e dos Dados

Forneça uma explicação das principais decisões que precisam ser feitas. (Limite de 250 palavras)
A rede de petshop Pawdacity possui 13 lojas e quer expandir sua atuação, para isso pretende abrir a 14ª loja no estado de Wyoming. Para isso temos alguns dados que será analisado com o objetivo de definir a cidade mais adequada para a nova loja da rede de pet shop, baseando-se nas previsões de vendas anuais, populações e densidade demográfica.

Decisões Chave:

Responda estas perguntas

1. Que decisões devem ser tomadas?
   Definir qual cidade irá sediar a nova loja.

2. Que dados são necessários para subsidiar essas decisões?
   Dados de cidade, população no ano de 2010, total de vendas, tamanho das localizações, densidade populacional e o total de famílias. Com isso saberemos a quantidade de clientes que temos em cada local.

Passo 2: Construindo o Conjunto de Treinamento

Construa seu conjunto de treinamento dado os dados fornecidos a você. As somas de coluna do seu conjunto de dados devem corresponder às somas na tabela abaixo.

Além disso, forneça as médias do seu conjunto de dados aqui para ajudar os revisores a verificar o seu trabalho. Você deve arredondar até duas casas decimais, ex: 1.24

Column	**Sum**	**Average**
Census Population	*213,862*	*19,442.00*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

Passo 3: Tratando os Outliers

Responda estas perguntas

Existem cidades que são outliers no conjunto de treinamento? R: Sim, existem. Cheyenne, Gillette e Rock Springs.
Qual outlier você escolheu para remover ou imputar? R: Removi a cidade de Cheyenne.

Como esse conjunto de dados é um conjunto de dados pequeno (11 cidades), **você deve apenas remover ou imputar um outlier**. Explique o seu raciocínio.
R: Cheyenne é uma cidade grande e seus números estão acima de todas as outras cidades no conjunto de treinamento nos campos Total Sales, Population Density, Total Families e 2010 Census. Isso significa que devemos remove-la, porque essa cidade é diferente de outras cidades para a maioria dos campos.

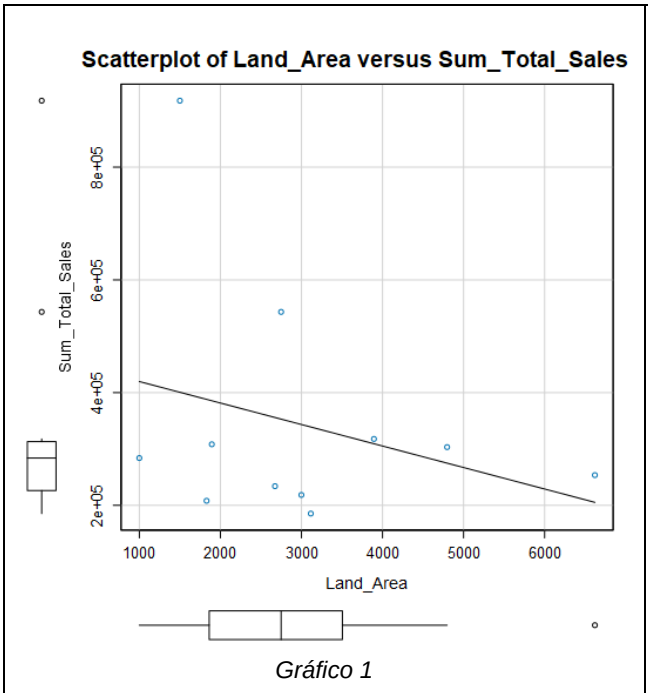

Gráfico 1

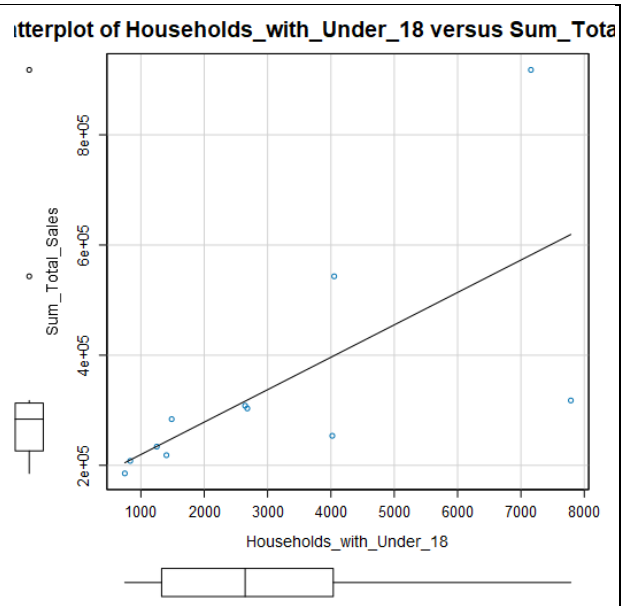

Gráfico 2

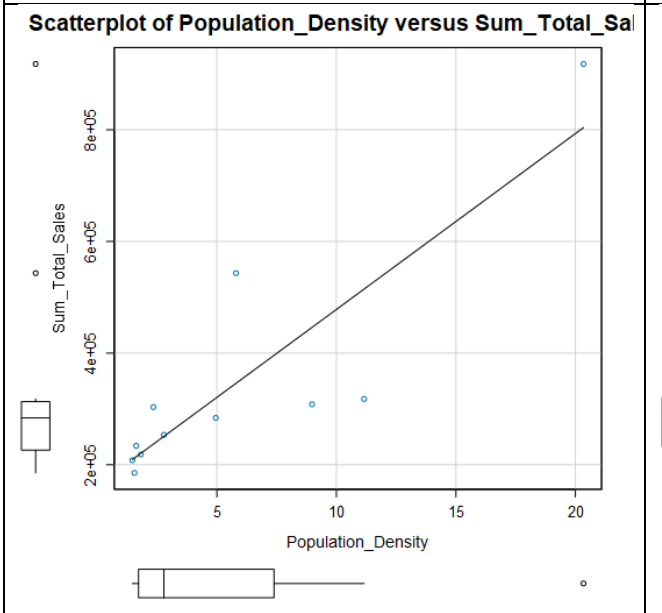

Gráfico 3

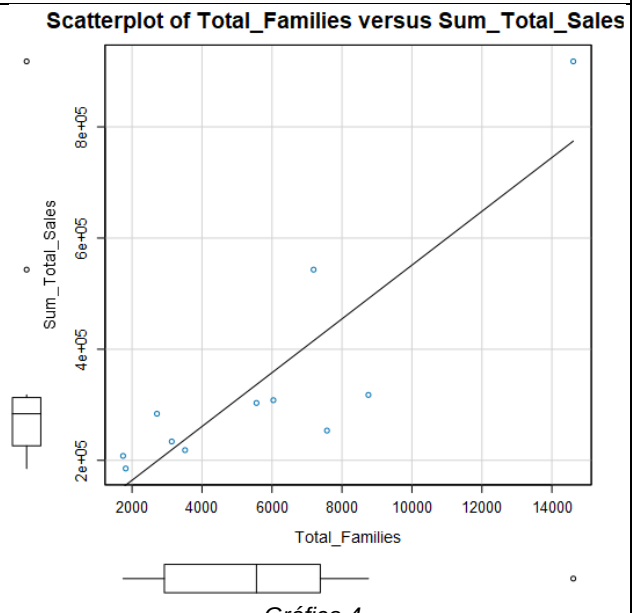

Gráfico 4

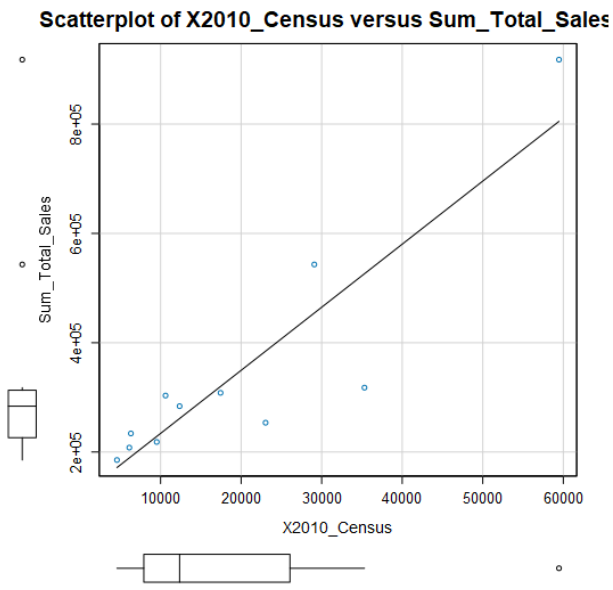
Scatterplot of X2010_Census versus Sum_Total_Sales
Sum_Total_Sales
2e+05
4e+05
6e+05
8e+05
10000
20000
30000
40000
50000
60000
X2010_Census

Gráfico 5