Project 2.1: Data Cleanup

Faça uma cópia deste documento. Complete cada seção. Quando estiver pronto, salve seu arquivo como um documento PDF e envie-o aqui: https://classroom.udacity.com/nanodegrees/nd008/parts/8d60a887-d4c1-4b0e-8873-b2f36435eb39/project

Passo 1: Entendimento do Negócio e dos Dados

Forneça uma explicação das principais decisões que precisam ser feitas. (Limite de 250 palavras)

Decisões Chave: *Responda estas perguntas*

1. Que decisões devem ser tomadas?

Uma rede líder de pet shops localizados no estado de Wyoming, chamada Pawcity, tem necessidade de um estudo sobre onde abrir a 14ª loja da rede.

2. Que dados são necessários para subsidiar essas decisões?

Para obtermos resultados satisfatórios no estudo, precisamos de informações para incluirmos na modelagem do projeto. Precisamos de informações como: população total das cidades do estador de Wyoming, as vendas das lojas Pawcity por cidade, vendas dos competidores, número de casas com pessoas com menos de 18 anos, tamanho das áreas, densidade populacional e total de famílias por cidades. Estas informações são cruciais para entendermos o tamanho das cidades, tamanhos das famílias, propensão a consumir produtos de petshop e entendermos as nossas vendas x competidores por cada uma das variáveis a serem analisadas.

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

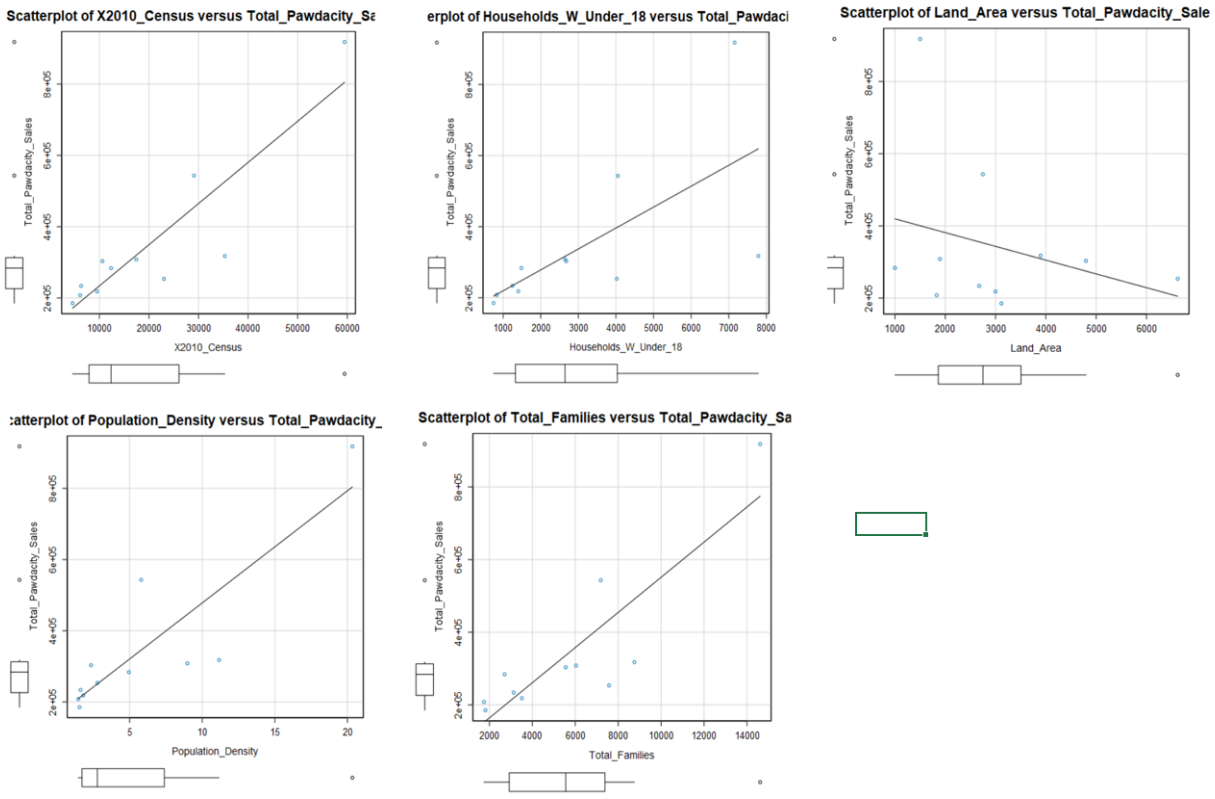
Indicator	Sum	Average
Census Population	213,862	19,442
Total Pawdacity Sales	3,773,304	343,027.64
Households with Under 18	34,064	3,096.73
Land Area	33,071	3,006.49
Population Density	63	5.71
Total Families	62,653	5,695.71

Passo 3: Tratando os Outliers

Responda estas perguntas

Existem cidades que são outliers no conjunto de treinamento? Qual outlier você escolheu para remover ou imputar? 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.

Analisando os resultados obtidos pelos scatterplots abaixo, vemos que as cidades de Cheyenne e Gillette são outliers, pois seus números de vendas são muito maiores em relação às outras cidades. Cheyenne podemos ignorá-la pois os dados são de uma cidade grande e não influenciam nos demais resultados. Enquanto isso, eu excluiria Gillette por conta que apenas a variável de vendas é outlier, com isso a cidade não tem característica de uma cidade grande, portanto influenciaria nos demais resultados, podemos considerá-la como anormal.



Antes de enviar

Por favor, verifique suas respostas contra os requisitos do projeto ditados por esta rubrica usada pelos revisores para classificar seu projeto.