Project 2.1: Data Cleanup

Passo 1: Entendimento do Negócio e dos Dados

Decisões Chave:

Responda estas perguntas

1. Que decisões devem ser tomadas?

Pawdacity é a principal cadeia de pet shops no estado do Wyoming com 13 lojas. Este ano, Pawdacity gostaria de expandir sua atuação com a 14ª loja. Objetivo deste projeto é realizar análise e determinar a melhor localização para abrir a mais nova loja da Pawdacity, se a decisão de expandir será lucrativa ou não. Tomar essas decisões exige processar as informações do conjunto de dados das projeções da receita da nova loja para diferentes cidades no estado de Wyoming.

2. Que dados são necessários para subsidiar essas decisões?

Essas decisões serão conduzidas por um conjunto de dados contendo as seguintes informações para cidade do estado de Wyoming em que cada uma das lojas de Pawdacity está atualmente:

- Total de vendas em 2010
- Quantidade da população com base nos dados do censo de 2010
- Área
- Densidade demográfica
- Quantidade de famílias com menores de 18 anos
- Quantidade total de famílias

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

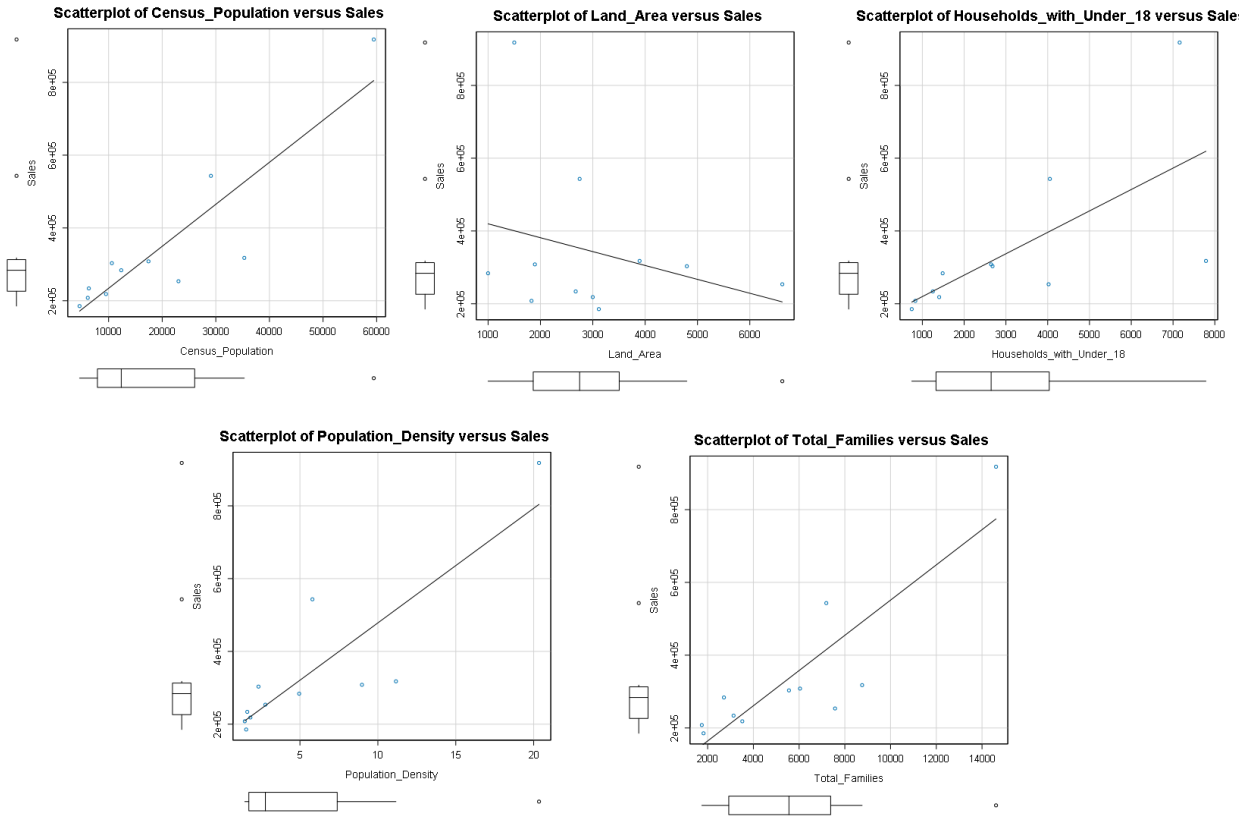
Column	**Sum**	**Average**
Census Population	*213,862*	*19,442.00*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.49*
Population Density	*63*	*5.71*
Total Families	*62,653*	*5,695.71*

Passo 3: Tratando os Outliers

Responda estas perguntas

Existem cidades que são outliers no conjunto de treinamento? 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.

Abaixo estão os gráficos de dispersão e os boxplots para cada possível variável preditor no conjunto de dados:



Segue um resumo do conjunto de dados:

Name	Field Category	Min	Max	Median	Std. Dev.
Census Population	Numeric	4585	59466	12359	16616.018584
Households with Under 18	Numeric	746	7788	2646	2453.003061
Land Area	Numeric	999.4971	6620.201916	2748.8529	1617.460342
Population Density	Numeric	1.46	20.34	2.78	5.849685
Sales	Numeric	185328	917892	283824	213538.712215
Total Families	Numeric	1744.08	14612.64	5556.49	3816.04966

Abaixo planilha com valores da faixa interquartil seguido da faixa superior de cada variável:

Sales IQR	Census Population IQR	Land Area IQR	Households with Under 18 IQR	Population Density IQR	Total Families IQR
86832	18144.5	1643.187226	2710	5.67	4457.395
Sales Upper	Census Population Upper	Land Area Upper	Households with Under 18 Upper	Population Density Upper	Total Families Upper
443232	53278.25	5969.689139	8102	15.895	14066.8975

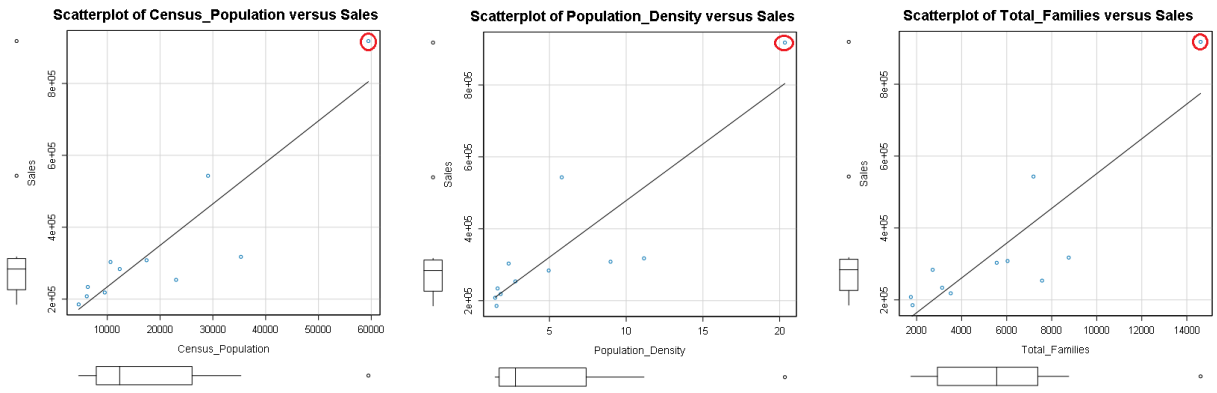
Informações que possuo das variáveis para cada cidade:

Record #	City	Sales	Census Population	Land Area	Households with Under 18	Population Density	Total Families
1	Buffalo	185328	4585	3115.5075	746	1.55	1819.5
2	Casper	317736	35316	3894.3091	7788	11.16	8756.32
3	Cheyenne	917892	59466	1500.1784	7158	20.34	14612.64
4	Cody	218376	9520	2998.95696	1403	1.82	3515.62
5	Douglas	208008	6120	1829.4651	832	1.46	1744.08
6	Evanston	283824	12359	999.4971	1486	4.95	2712.64
7	Gillette	543132	29087	2748.8529	4052	5.8	7189.43
8	Powell	233928	6314	2673.57455	1251	1.62	3134.18
9	Riverton	303264	10615	4796.859815	2680	2.34	5556.49
10	Rock Springs	253584	23036	6620.201916	4022	2.78	7572.18
11	Sheridan	308232	17444	1893.977048	2646	8.98	6039.71

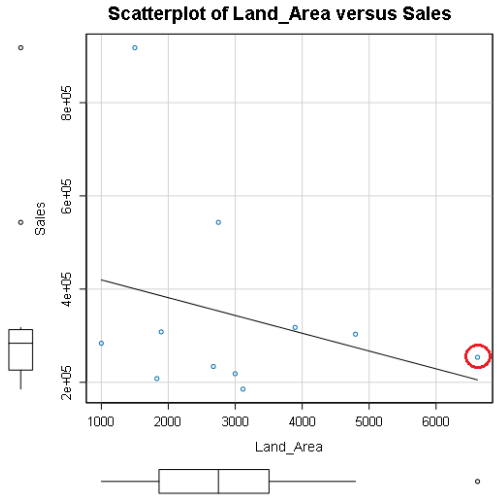
Analisando cada variável com a faixa superior calculado, verifiquei que seguintes pontos estão acima da faixa superior:

- Sales para Gillette e Cheyenne
- Census Population para Cheyenne
- Land Area para Rock Springs
- Population Density para Cheyenne
- Total Families para Cheyenne

Primeiro outlier é a cidade de Cheyenne que excede a faixa de quatro variáveis (Sales, Census Population, Population Density e Total Families). No entanto é esperado que a capital do estado seja uma cidade mais urbana em comparação com as demais cidades, de modo que a alta densidade demográfica, quantidade de habitantes e total de famílias se correlacionam entre si. Gráfico de dispersão do total de vendas entre as 3 variáveis também se correlacionam de forma linear, apesar de ser dados anômalos.



Rock Springs é a segunda cidade onde consideramos com um outlier baseado na variável Land Area. Apesar de ter uma área muito maior em comparação com as outras cidades no conjunto de dados, ela ainda se correlaciona linearmente entre total de vendas e área territorial. Como não tem nenhuma distorção nesse requisito, essa cidade também não será removida do conjunto de dados.



Gillette é a última cidade que foi definida como outlier baseado no total de vendas, apesar de estar dentro da média de todas as outras variáveis exceto a variável Sales. Como observamos que as vendas para essa cidade é um outlier e não conseguimos explicar por nenhum outro outlier encontrado na população ou métricas demográficas, deixar esta entrada no conjunto de dados tem potencial de distorcer qualquer modelo de treinamento nesses dados. Por tanto esta cidade será removida do conjunto de dados.