# Project 2.1: Data Cleanup

## 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?
   Em que cidade abrir uma nova loja.

2. Que dados são necessários para subsidiar essas decisões?
   Receita potencial em cada cidade.

## 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* |
| *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.

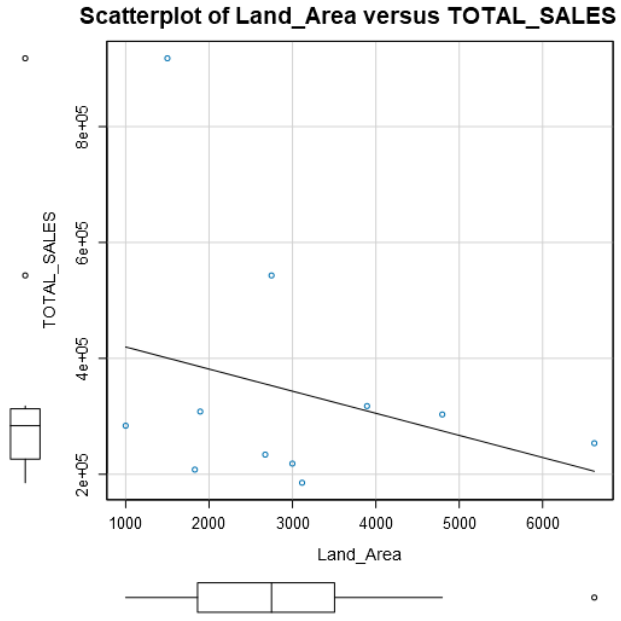
Existem 3 registros que podem ser considerados outliers em diferentes variáveis utilizando o método dos intervalos interquartis:

- ***Cheyenne - TOTAL_SALES; 2010_Census; Population_Density; Total_Families***
- ***Gillette - TOTAL_SALES***
- ***Rock Springs – Land_Area***

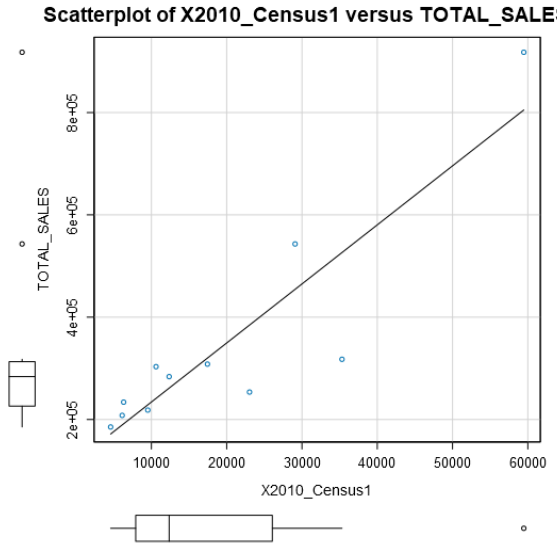
| CITY | TOTAL SALES | 2010 Census | Land Area | HH UNDER 18 | Pop Density | Total Families |
|---|---|---|---|---|---|---|
| Buffalo | 185,328 | 4,585 | 3,115.51 | 746 | 1.55 | 1,819.50 |
| Casper | 317,736 | 35,316 | 3,894.31 | 7,788 | 11.16 | 8,756.32 |
| Cheyenne | 917,892 | 59,466 | 1,500.18 | 7,158 | 20.34 | 14,612.64 |
| Cody | 218,376 | 9,520 | 2,998.96 | 1,403 | 1.82 | 3,515.62 |
| Douglas | 208,008 | 6,120 | 1,829.47 | 832 | 1.46 | 1,744.08 |
| Evanston | 283,824 | 12,359 | 999.50 | 1,486 | 4.95 | 2,712.64 |
| Gillette | 543,132 | 29,087 | 2,748.85 | 4,052 | 5.80 | 7,189.43 |
| Powell | 233,928 | 6,314 | 2,673.57 | 1,251 | 1.62 | 3,134.18 |
| Riverton | 303,264 | 10,615 | 4,796.86 | 2,680 | 2.34 | 5,556.49 |
| Rock Springs | 253,584 | 23,036 | 6,620.20 | 4,022 | 2.78 | 7,572.18 |
| Sheridan | 308,232 | 17,444 | 1,893.98 | 2,646 | 8.98 | 6,039.71 |
| **SUM** | **3,773,304** | **213,862** | **33,071** | **34,064** | **63** | **62,653** |
| **AVG** | **343,028** | **19,442** | **3,006** | **3,097** | **6** | **5,696** |

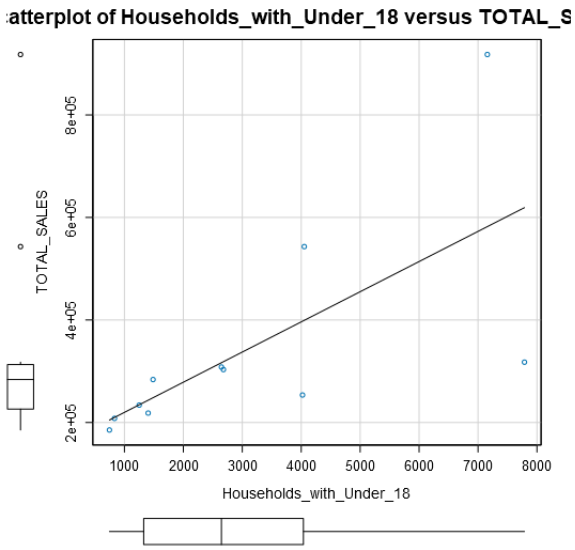
| | | | | | | |
|---|---|---|---|---|---|---|
| Q1 | 226152.0 | 7917.0 | 1861.7 | 1327.0 | 1.7 | 2923.4 |
| Q3 | 312984.0 | 26061.5 | 3504.9 | 4037.0 | 7.4 | 7380.8 |
| AIQ | 86832.0 | 18144.5 | 1643.2 | 2710.0 | 5.7 | 4457.4 |
| LIM INF | 95904.0 | -19299.8 | -603.1 | -2738.0 | -6.8 | -3762.7 |
| LIM SUP | 443232.0 | 53278.3 | 5969.7 | 8102.0 | 15.9 | 14066.9 |
| $R^2$ | - | 80.7% | 8.3% | 45.7% | 74.5% | 74.8% |

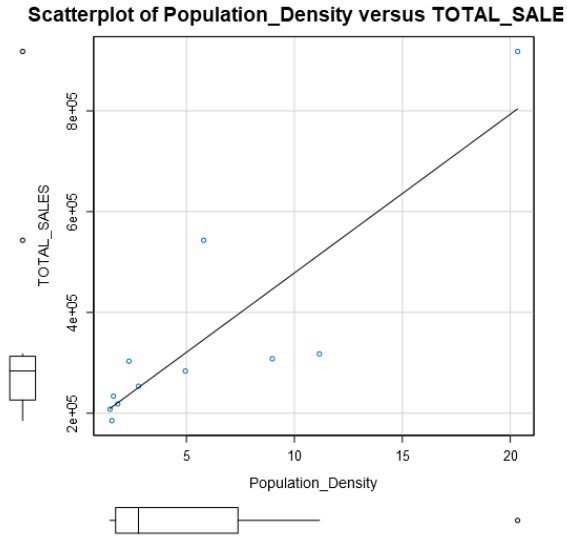
Explorando a força da relação de linearidade entre as variáveis, podemos ver que a variável Land_Area não se mostra relevante para o modelo e, portanto, podemos manter *Rock Springs* no modelo.

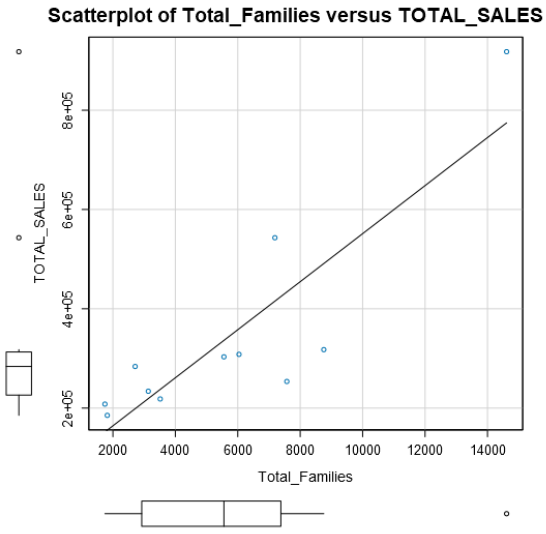


Já ao analisarmos a cidade de *Cheyenne*, percebemos que a mesma se mostra como outlier em diversas categorias, porém, sem afetar o declive da reta e, portanto, também será mantida.




Assim sendo, foi optado pela exclusão da o registro *Gillette*, que apresentava dados de vendas discrepantes quando comparado com os outros registros em relação às outras variáveis.