Project 2.1: Data Cleanup

## **Entendimento do Negócio e dos Dados**

**Que decisões devem ser tomadas?**

Pawdacity é a principal cadeia de pet shops no estado do Wyoming com 13 lojas. Este ano, gostaria de expandir sua atuação com a 14º loja. O objetivo deste projeto é realizar uma análise para recomendar qual a melhor cidade para a mais nova loja de Pawdacity, com base nas vendas anuais previstas.

**Que dados são necessários para subsidiar essas decisões?**

O primeiro passo é formatar e agregar dados vindos de diferentes bases de dados e lidar com dados anômalos (outliers). E para isso temos 3 data set disponíveis:

*p2-2010-pawdacity-monthly-sales.csv*,

*p2-partially-parsed-wy-web-scrape.csv*,

*p2-wy-453910-naics-data.csv*.

Precisamos descobrir quais dados dos arquivos acima serão necessários para prever onde deve ser sua próxima loja.

## **Construindo o Conjunto de Treinamento**

Precisamos extrair as seguintes colunas de dados dos arquivos acima:

|  |
| --- |
| City |
| 2010 Census Population |
| Total Pawdacity Sales |
| Households with under 18 |
| Land Area |
| Population Density |
| Total Families |

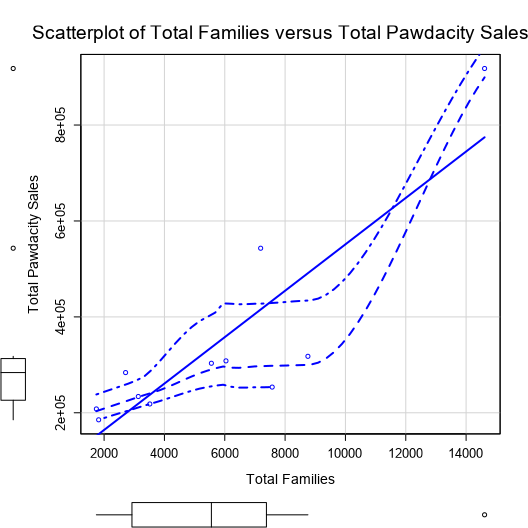
Os dados dos campos acima serão usados posteriormente para criar um modelo de previsão para o novo local da loja (Projeto 2.2).

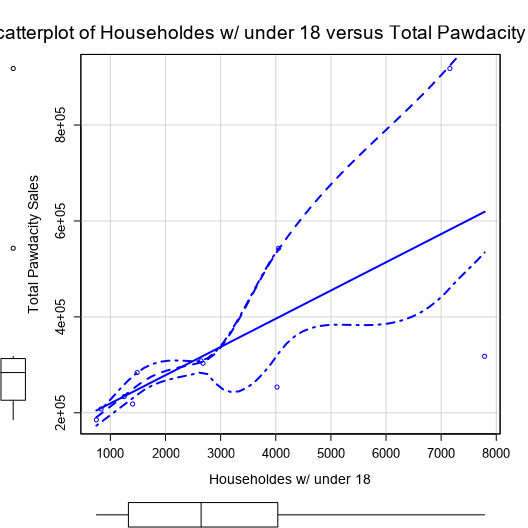
Abaixo é apresentado um resumo os dados.

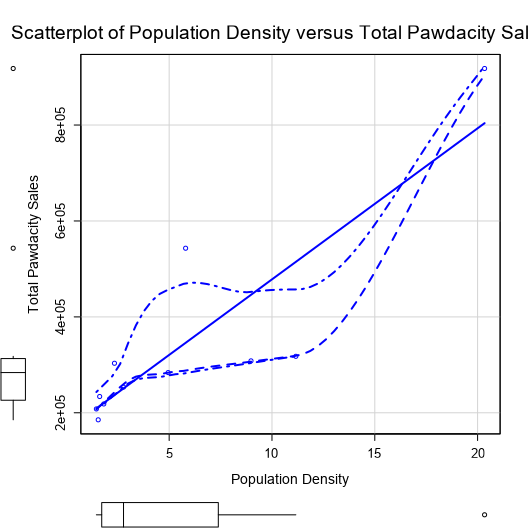
|  |  |  |
| --- | --- | --- |
| Name | Sum | Avg |
| Total Pawdacity Sales | 3773304 | 343027.6364 |
| Land Area | 33071.38039 | 3006.489126 |
| Households with Under 18 | 34064 | 3096.727273 |
| Population Density | 62.8 | 5.709090909 |
| Total Families | 62652.79 | 5695.708182 |

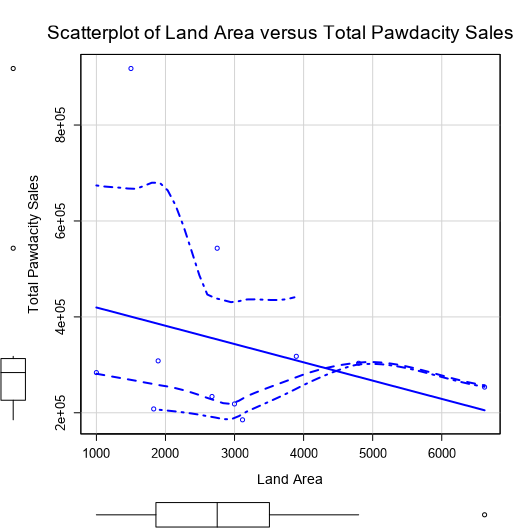
## **Tratando os Outliers**

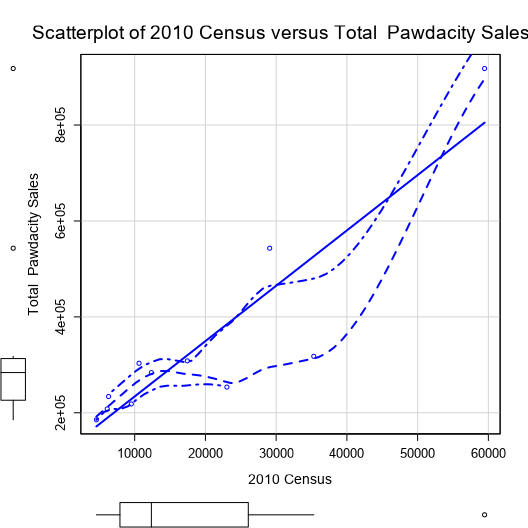
Abaixo estão os gráficos de dispersão e os gráficos de caixa do conjunto de dados, com cada variável preditora em potencial plotada em relação às vendas de Pawdacity para essa cidade.



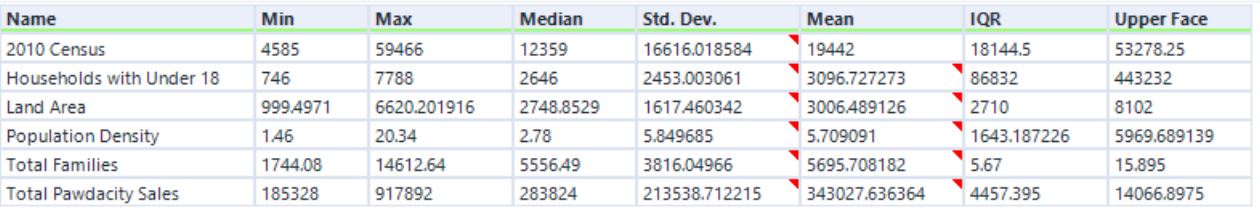








Abaixo está um resumo do conjunto de dados, com uma análise adicional dos intervalos interquartis para as variáveis e seu limite superior subsequente, que para este projeto será [1,5 \* Intervalo Interquartil] + 3º Quartil.

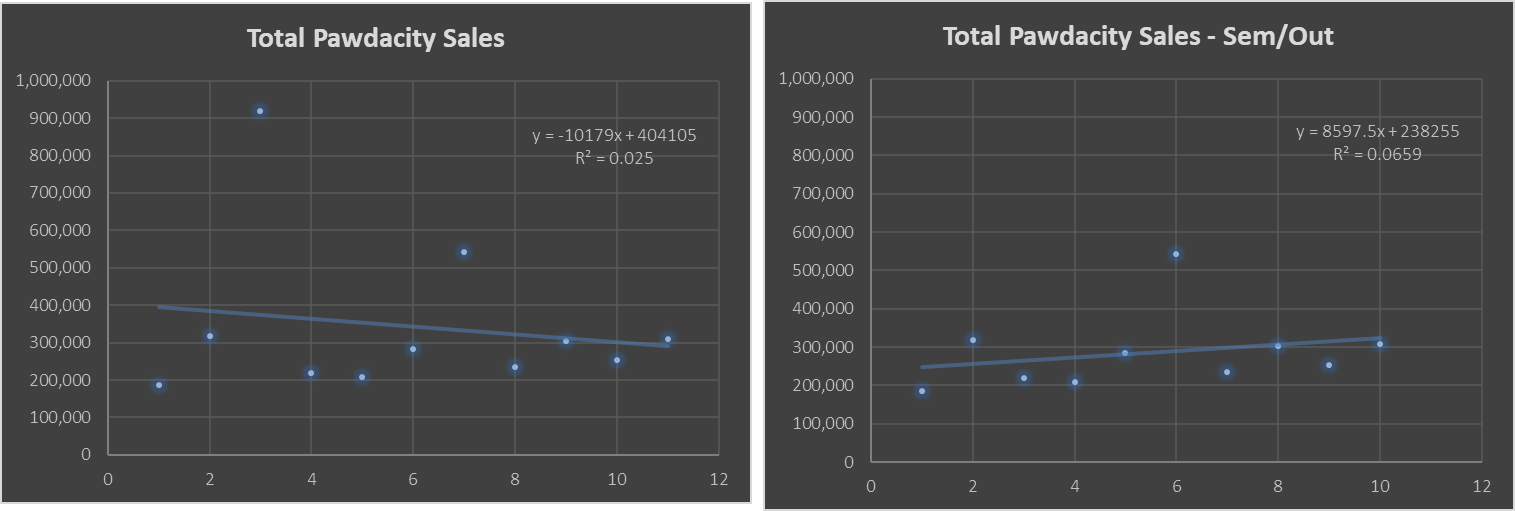


A tabela abaixo mostra os máximos pontos acima do “Upper Face”

|  |  |  |
| --- | --- | --- |
| 2010 Census Population | Cheyenne | 59466 |
| Total Pawdacity Sales | Cheyenne e Gillette | 917892 |
| Land Area | Rock Spring | 6620.201916 |
| Population Density | Cheyenne | 20.34 |
| Total Families | Cheyenne | 14612.64 |

Há 4 cidades que são consideradas outliers no conjunto de dados como mostra a imagem abaixo, contudo vamos focar nas 3 cidades que estão no “Upper Face”, ou seja que ultrapassam o limite superior. De uma forma mais detalhada para identificar os outliers no conjunto de dados, foi feita uma análise do boxplot, utilizando o 1° e 3° quartil (Q1, Q4) para achar o limite superior (LinfUP) e inferior (LinfLO), feito isso foi calculado os valores que ultrapassam o 1.5x ambos os limites. O resultado está mostrado na imagem abaixo, onde as células grivadas em vermelho mostra que passaram do limite superior, e as células grifadas em azul representam os valores que passaram o limite inferior, ambas são outliers.

Para uma análise, escolhi remover a cidade Cheyene, pois ela apresenta 4 outliers que ultrapassam o limite superior nas colunas: Total Pawdacity Sales, 2010 Census, Popolation Density e Total de Familie; e por conta da mudança da linha de tendência amostrada no gráfico abaixo.



**2**

**1**

Figure 1 - Gráficos com e sem outliers

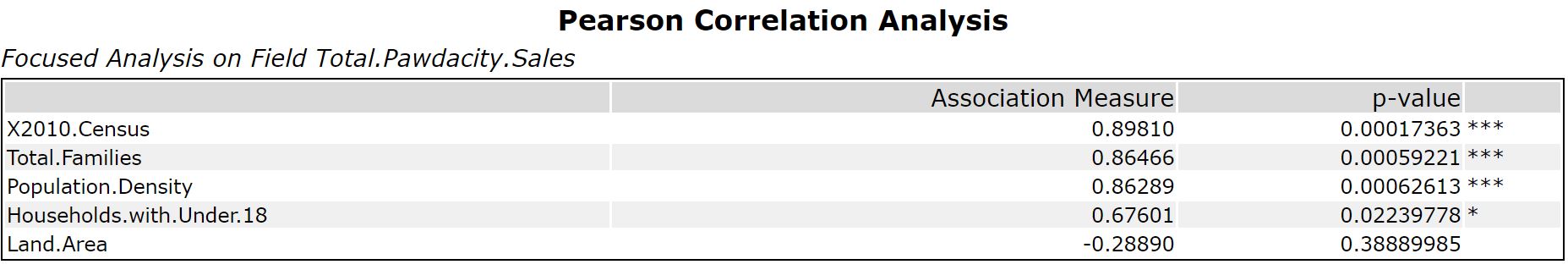
No gráfico (1) a esquerda é dos dados com os outliers (Cheyene), o da direita (2) já é com os outliers removidos, podemos notar uma grande diferença na linha de tendência. No primeiro caso o gráfico 1 mostra um decline na linha, já o gráfico 2 a linha de tendência apresenta uma ascensão, como podemos mostra a tabela abaixo.

Onde a inclinação é igual a diferença do eixo Y dividida pela diferença do eixo X.

|  |  |
| --- | --- |
| Gráfico 1 | Gráfico 2 |
|  |  |

E podemos notar também uma diferencia na equação e no R2.

Abaixo está um resumo da correlação de Pearson calculada a partir das variáveis preditivas e da variável de destino que, neste caso, é Total Pawdacity Sales.



O gráfico de dispersão para Land Area vs Total Pawdacity Sales, indicaria para mim que Rock Springs segue a direção descendente da linha de melhor ajuste para esse plot, com vendas aproximadamente alinhadas com outros valores de vendas nesse plot.

Cheyenne, por outro lado, possui duas lojas e seus dados são agregados nessa análise, o que pode causar uma discrepância, no entanto, como estamos procurando onde colocar a nova loja, devemos analisar esses dados no nível da cidade. Isso significaria que Cheyenne justificadamente é a cidade que produz mais vendas para manter as duas lojas.

A Gillette também tem duas lojas, no entanto, analisando as outras categorias, os dados da Gillette são relativamente parecidos em nossa linha mais externa “Upper Face”, exceto por suas vendas. Não parece haver uma boa razão para isso, com base na pequena quantidade de informações que eu conheço.

Minha recomendação aqui seria manter Cheyenne e Rock Springs, pois acredito que seus dados parecem adequados. Já Gillette, no entanto, é mais difícil de explicar e seria melhor remover totalmente essa cidade do nosso conjunto de dados.

**Alteryx Workflow**

