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

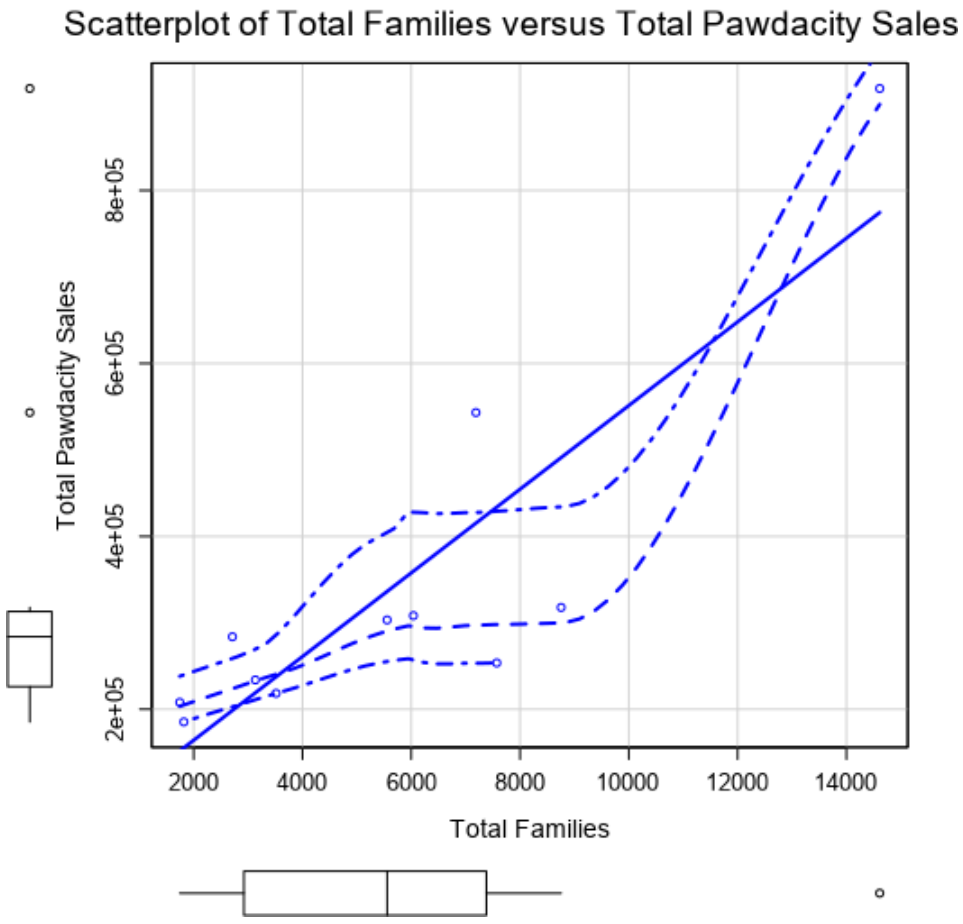
Os dados dos campos acima serão usados posteriormente para criar um modelo de previsão para o novo local da loja (Projeto 2.2).

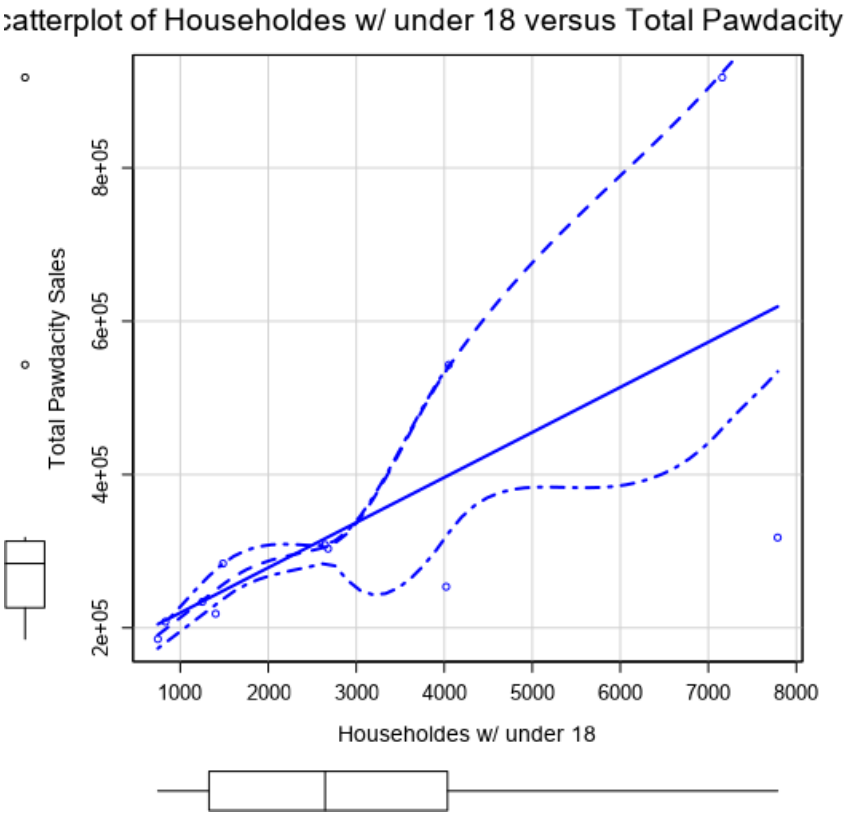
Abaixo é apresentado um resumo os dados.

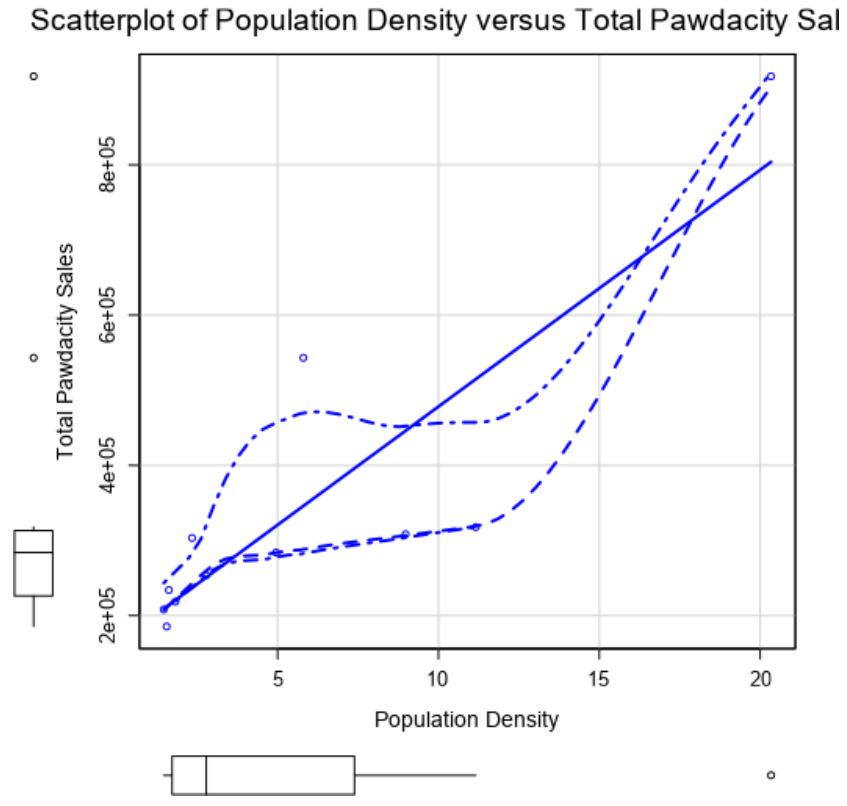
Name	Sum	Avg
Total Pawdacity Sales	3773304	343027.6364
Land Area	33071.38039	3006.489126
Households with Under 18	34064	3096.727273
Population Density	62.8	5.709090909
Total Families	62652.79	5695.708182

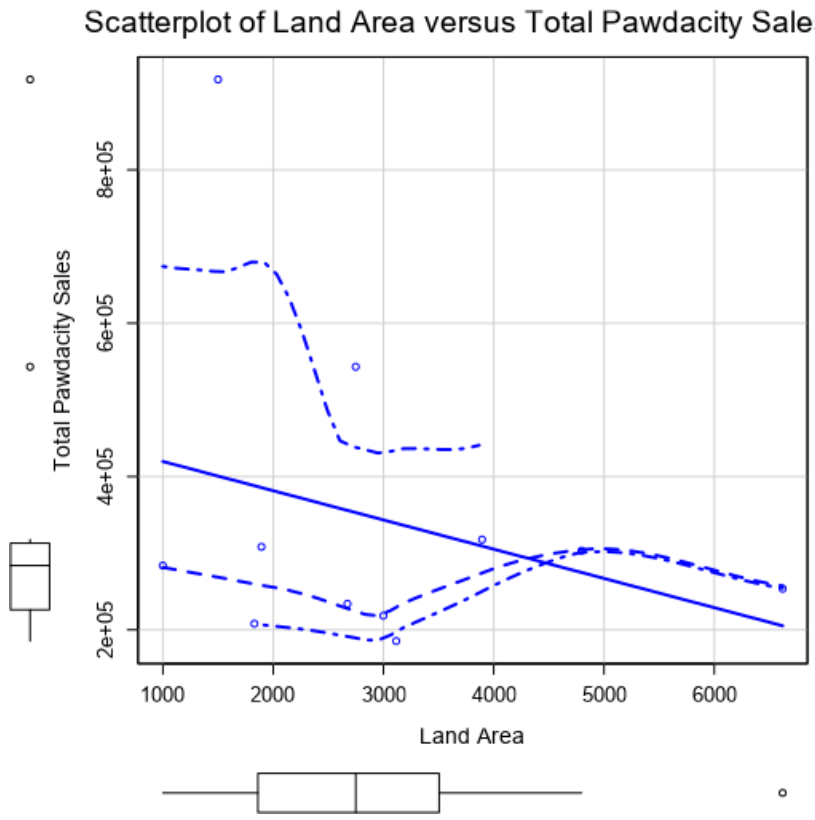
Tratando os Outliers

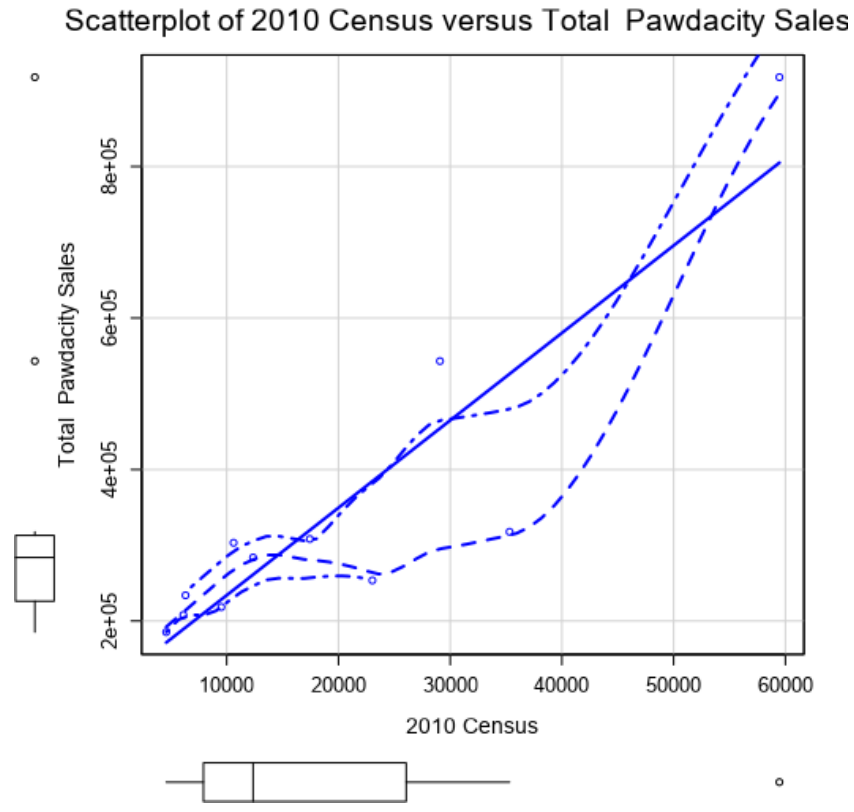
Abaixo estão os gráficos de dispersão e os gráficos de caixa do conjunto de dados, com cada variável preditora em potencial plotada em relação às vendas de Pawdacity para essa cidade.




catterplot of Householdes w/ under 18 versus Total Pawdacity
Total Pawdacity Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
7000
8000
Householdes w/ under 18


Scatterplot of Population Density versus Total Pawdacity Sal
Total Pawdacity Sales
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Population Density


Scatterplot of Land Area versus Total Pawdacity Sales
Total Pawdacity Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Land Area


Scatterplot of 2010 Census versus Total Pawdacity Sales
Total Pawdacity Sales
2e+05
4e+05
6e+05
8e+05
10000
20000
30000
40000
50000
60000
2010 Census

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.

Name	Min	Max	Median	Std. Dev.	Mean	IQR	Upper Face
2010 Census	4585	59466	12359	16616.018584	19442	18144.5	53278.25
Households with Under 18	746	7788	2646	2453.003061	3096.727273	86832	443232
Land Area	999.4971	6620.201916	2748.8529	1617.460342	3006.489126	2710	8102
Population Density	1.46	20.34	2.78	5.849685	5.709091	1643.187226	5969.689139
Total Families	1744.08	14612.64	5556.49	3816.04966	5695.708182	5.67	15.895
Total Pawdacity Sales	185328	917892	283824	213538.712215	343027.636364	4457.395	14066.8975

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.

CITY	Country	Total Pawdacity Sales	2010 Census	vendas /pop	Land Area	Households with Under 18	vendas />18	Population Density	Vendas /popDen	Total Families
Buffalo	Johnson	$ 185,328.00	4585.00	40	3,115.5075	746	248	1.55	119566.45	1819.5
Casper	Natrona	$ 317,736.00	35316.00	9	3,894.3091	7,788	41	11.16	28470.97	8756.32
Cheyenne	Laramie	$ 917,892.00	59466.00	15	1,500.1784	7,158	128	20.34	45127.43	14612.64
Cody	Park	$ 218,376.00	9520.00	23	2,998.9570	1,403	156	1.82	119986.81	3515.62
Douglas	Converse	$ 208,008.00	6120.00	34	1,829.4651	832	250	1.46	142471.23	1744.08
Evanston	Uinta	$ 283,824.00	12359.00	23	999.4971	1,486	191	4.95	57338.18	2712.64
Gillette	Campbell	$ 543,132.00	29087.00	19	2,748.8529	4,052	134	5.8	93643.45	7189.43
Powell	Park	$ 233,928.00	6314.00	37	2,673.5746	1,251	187	1.62	144400.00	3134.18
Riverton	Fremont	$ 303,264.00	10615.00	29	4,796.8598	2,680	113	2.34	129600.00	5556.49
Rock Springs	Sweetwater	$ 253,584.00	23036.00	11	6,620.2019	4,022	63	2.78	91217.27	7572.18
Sheridan	Sheridan	$ 308,232.00	17444.00	18	1,893.9770	2,646	116	8.98	34324.28	6039.71
Q1		**$ 226,152.00**	**$ 7,917.00**	**17**	**1,861.7211**	**1,327**	**115**	**1.72**	**51,233**	**2,923.41**
Q3		**$ 312,984.00**	**$ 26,061.50**	**31**	**3,504.9083**	**4,037**	**189**	**7.39**	**124,793**	**7,380.81**
IQR		**$ 86,832.00**	**$ 18,144.50**	**15**	**1,643.1872**	**2,710**	**74**	**5.67**	**73,561**	**4,457.40**
LinfUP		**$ 473,275.64**	**$ 46,658.75**	**46**	**5,471.2700**	**7,162**	**259**	**14.21**	**201,809**	**12,381.80**
LinfLO		**$ 212,779.64**	**$ (7,774.75)**	**1**	**541.7083**	**-968**	**37**	**-2.80**	**-18,873**	**-990.38**
Média		**$ 343,027.64**	**$ 19,442.00**	**23**	**3,006.4891**	**3,097**	**148**	**5.71**	**91,468**	**5,695.71**

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.

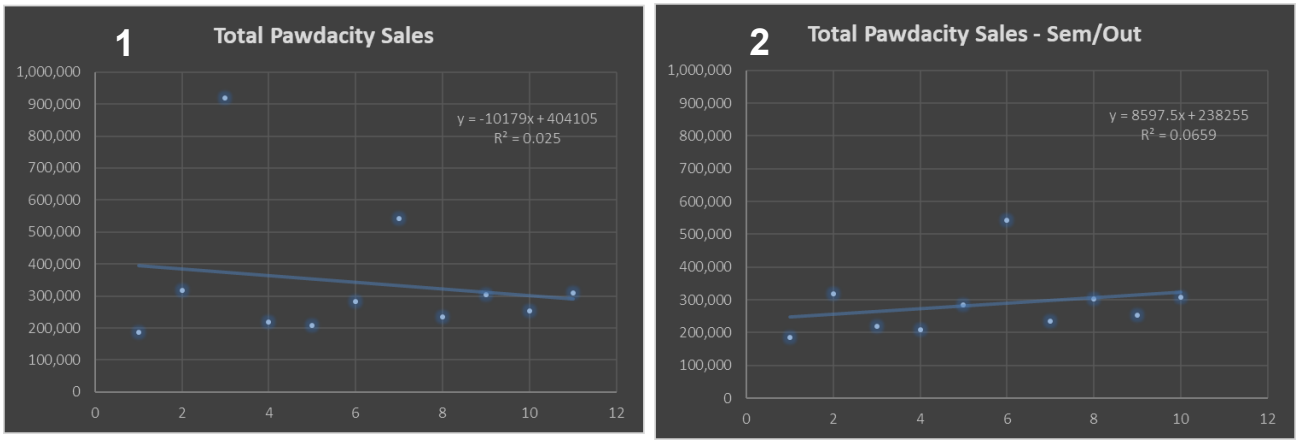


Figure 1 - Gráficos com e sem outliers

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
$\frac{300-390}{10-2} = \frac{-90}{8} = -11.25$	$\frac{320-250}{10-2} = \frac{70}{8} = 8.75$

E podemos notar também uma diferencia na equação e no R^2.

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.

Pearson Correlation Analysis

Focused Analysis on Field Total.Pawdacity.Sales

	Association Measure	p-value
X2010.Census	0.89810	0.00017363 ***
Total.Families	0.86466	0.00059221 ***
Population.Density	0.86289	0.00062613 ***
Households.with.Under.18	0.67601	0.02239778 *
Land.Area	-0.28890	0.38889985

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