Step 1: Business and Data Understanding

The Business Problem

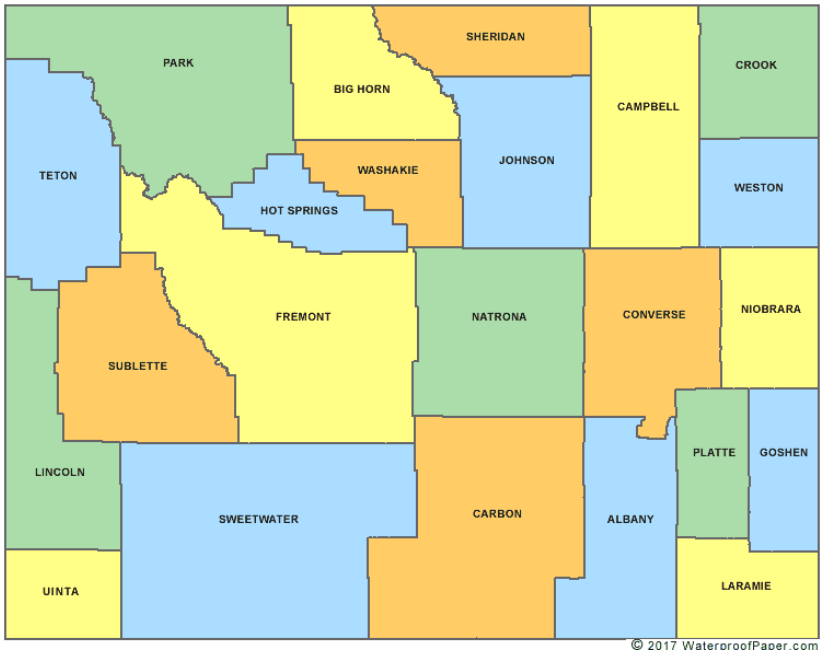
Pawdacity is a leading pet store chain in Wyoming with 13 stores throughout the state. This year, Pawdacity would like to expand and open a 14th store. Your manager has asked you to perform an analysis to recommend the city for Pawdacity's newest store, based on predicted yearly sales.

Your first step in predicting yearly sales is to first format and blend together data from different datasets and deal with outliers.

Your manager has given you the following information to work with:

1. The monthly sales data for all of the Pawdacity stores for the year 2010.
2. NAICS data on the most current sales of all competitor stores where total sales is equal to 12 months of sales.
3. A partially parsed data file that can be used for population numbers.
4. Demographic data (Households with individuals under 18, Land Area, Population Density, and Total Families) for each city and county in the state of Wyoming. For people who are unfamiliar with the US city system, a state contains counties and counties contains one or more cities.

Map of Wyoming Counties



What decisions need to be made?

There are 4 sets of data:

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

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

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

p2-wy-demographic-data.csv

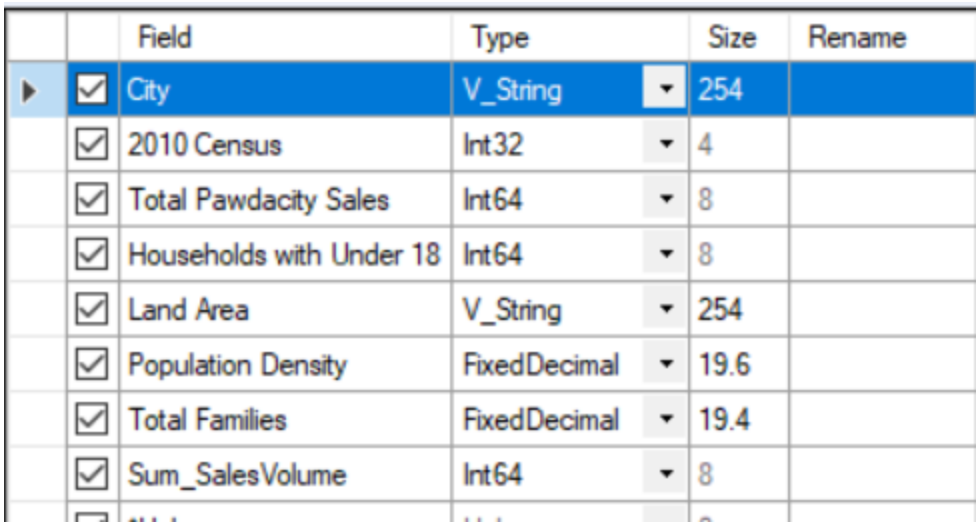
We need to work out what data from the above files will be necessary to predict where our next store should be.

What data is needed to inform those decisions?

We will need to extract the following columns of data from the above files:

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

The data from the above fields will later be used to create a prediction model for the new store location.



	Field	Type	Size	Rename
▶	☑	City	V_String	254	
	☑	2010 Census	Int32	4	
	☑	Total Pawdacity Sales	Int64	8	
	☑	Households with Under 18	Int64	8	
	☑	Land Area	V_String	254	
	☑	Population Density	FixedDecimal	19.6	
	☑	Total Families	FixedDecimal	19.4	
	☑	Sum_SalesVolume	Int64	8	

Step 2: Building the Training Set

Results:

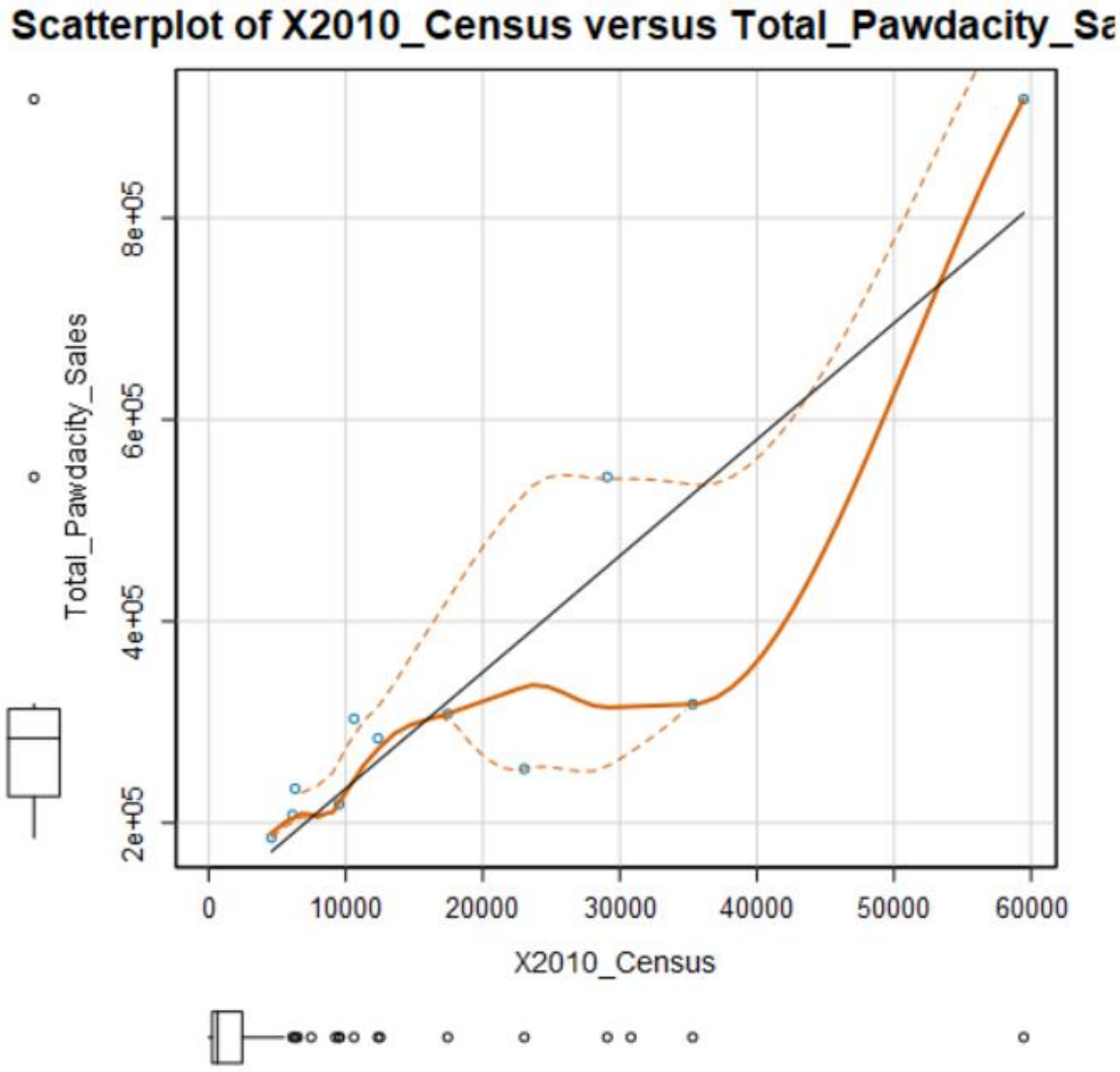
Record #	City	2010 Census	Total Pawdacity Sales	Households with Under 18	Land Area	Population Density	Total Families
1	Buffalo	4585	185328	746	3115.5075	1.550000	1819.5000
2	Casper	35316	317736	7788	3894.3091	11.160000	8756.3200
3	Cheyenne	59466	917892	7158	1500.1784	20.340000	14612.6400
4	Cody	9520	218376	1403	2998.95696	1.820000	3515.6200
5	Douglas	6120	208008	832	1829.4651	1.460000	1744.0800
6	Evanston	12359	283824	1486	999.4971	4.950000	2712.6400
7	Gillette	29087	543132	4052	2748.8529	5.800000	7189.4300
8	Powell	6314	233928	1251	2673.57455	1.620000	3134.1800
9	Riverton	10615	303264	2680	4796.859815	2.340000	5556.4900
10	Rock Springs	23036	253584	4022	6620.201916	2.780000	7572.1800
11	Sheridan	17444	308232	2646	1893.977048	8.980000	6039.7100

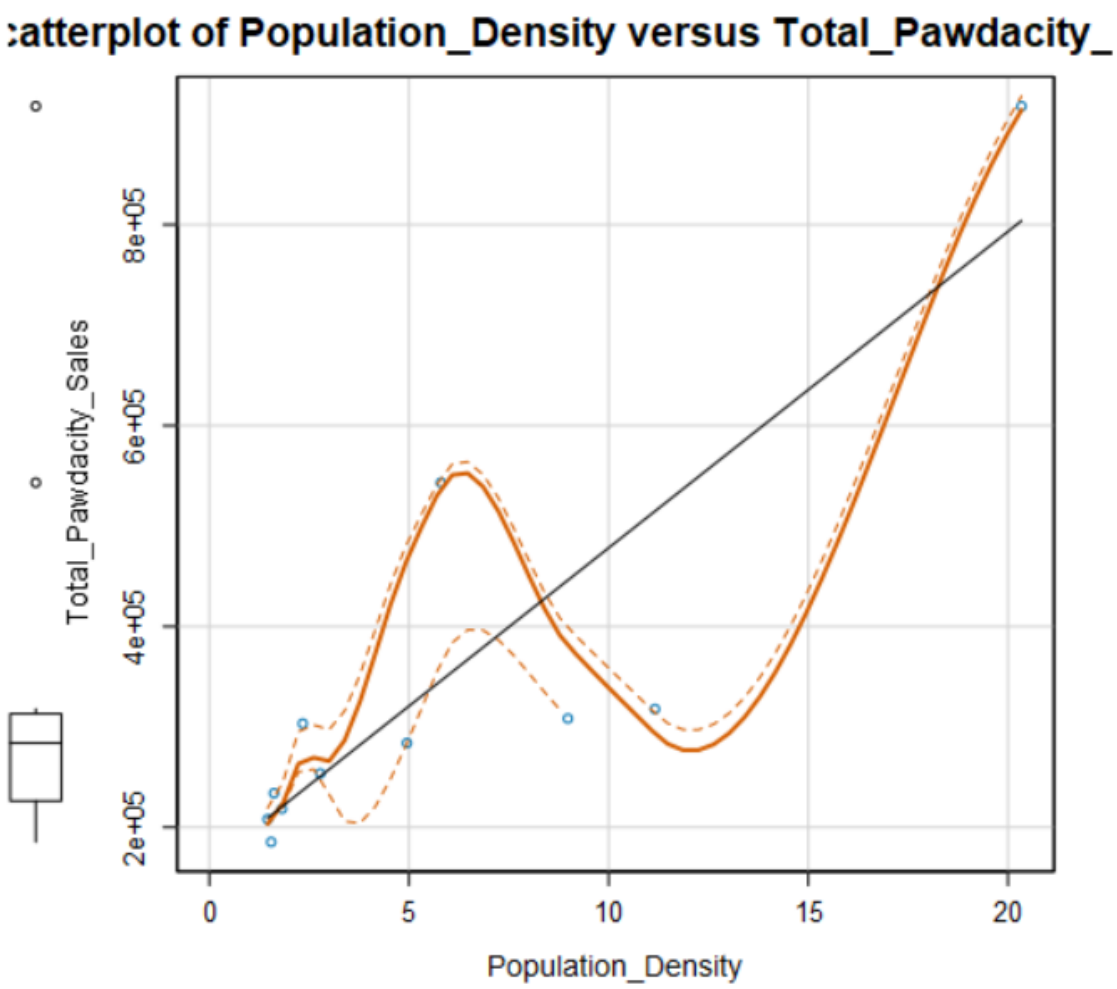
Column	Sum	Average
Census Population	213862	19442
Total Pawdacity Sales	3773304	343027.64
Households with Under 18	34064	3096.73
Land Area	33071	3006.49
Population Density	63	5.71
Total Families	62653	5695.71

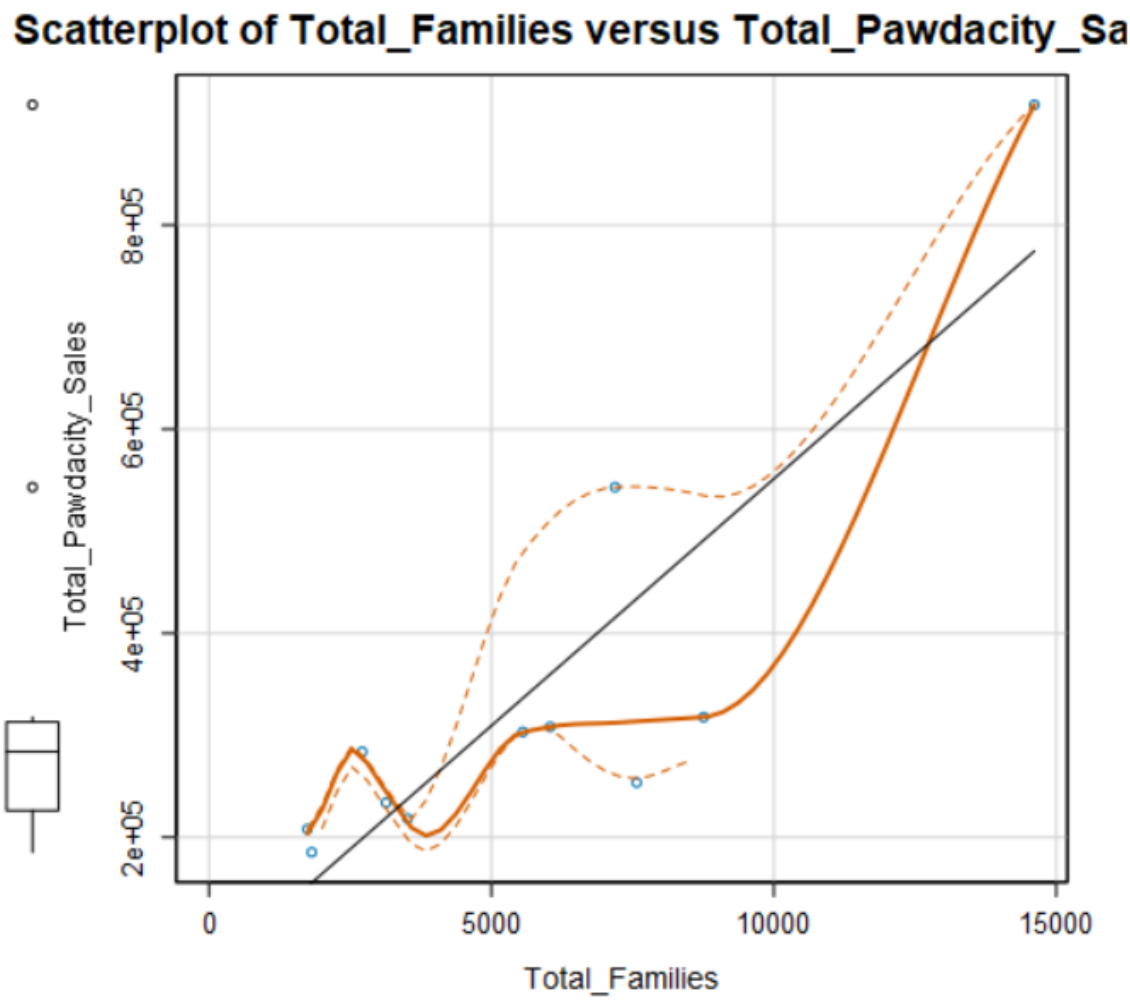
Step 3: Dealing with Outliers

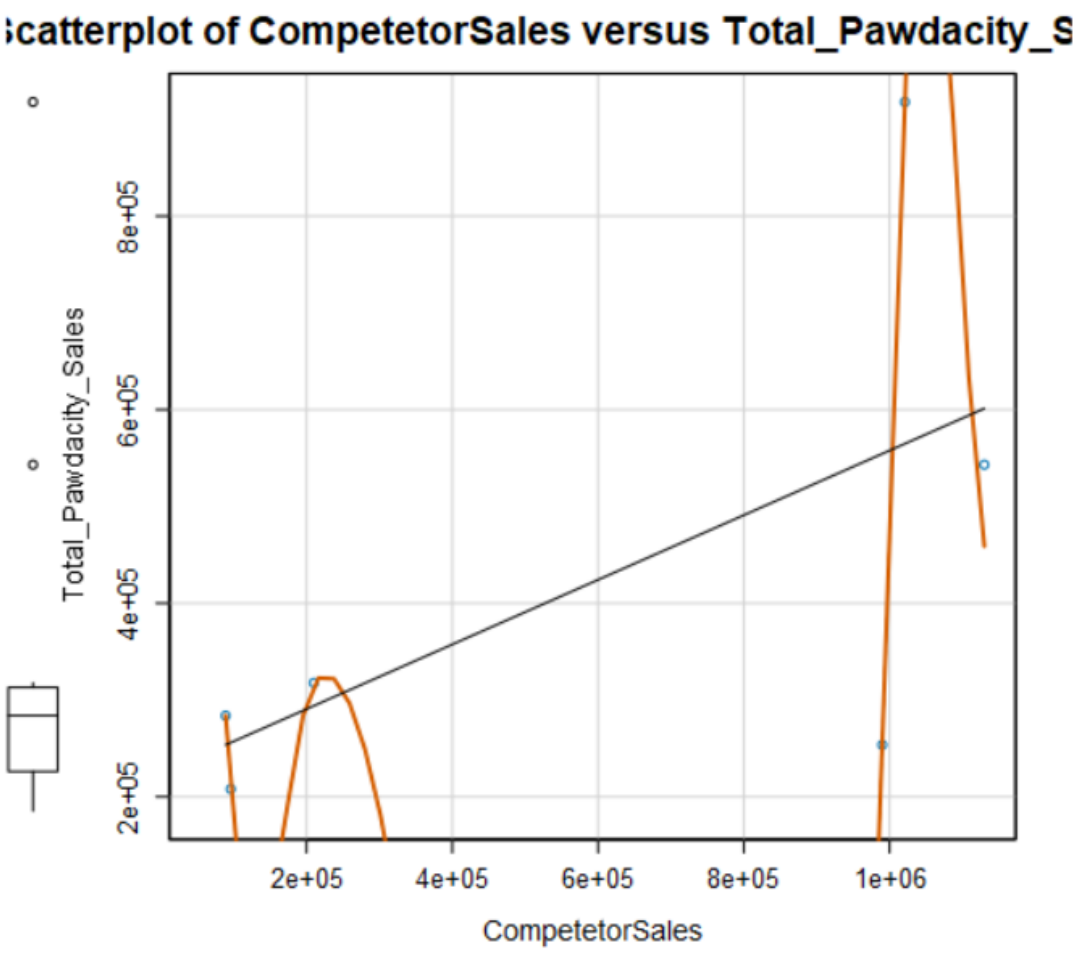
Outliers in the dataset.

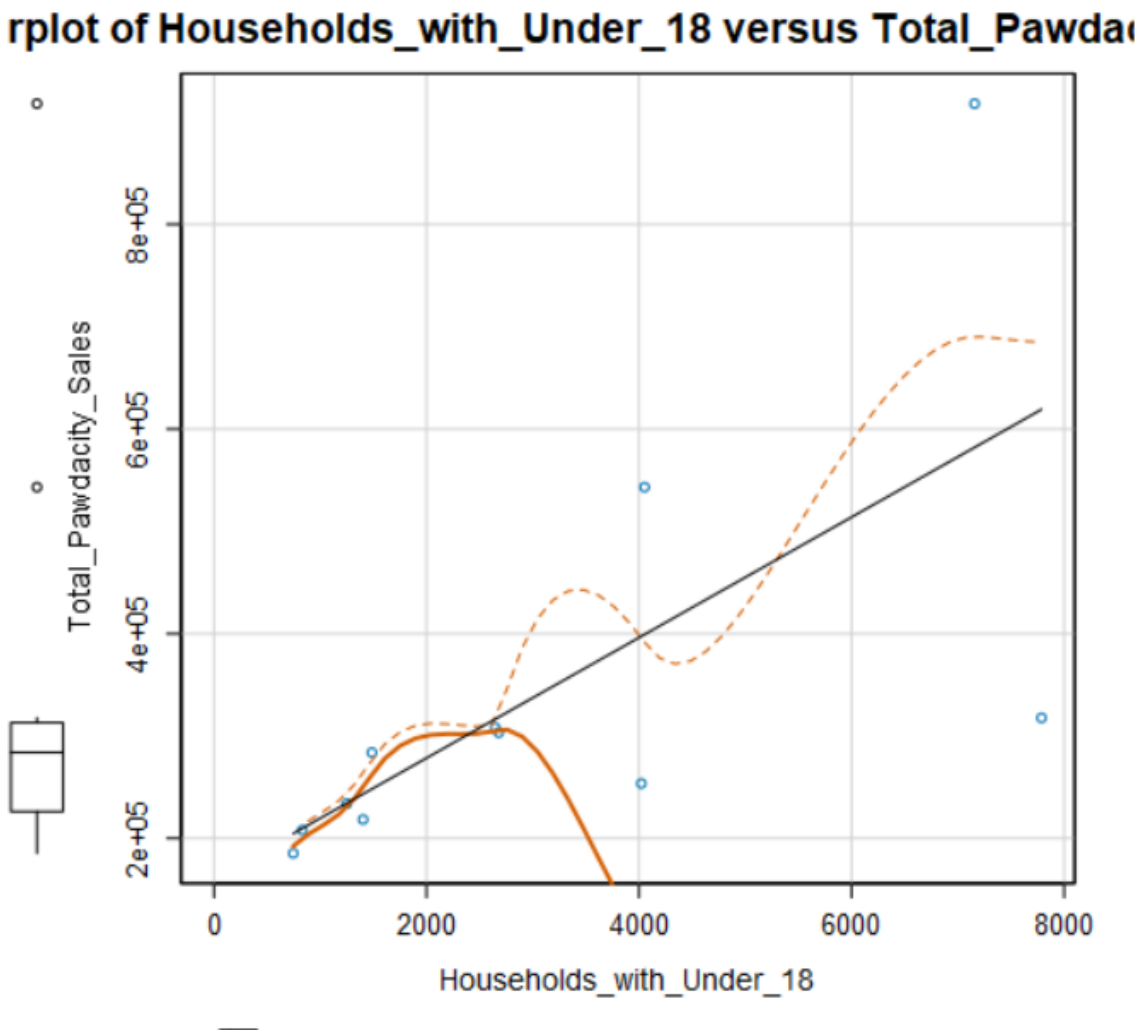
Below are scatter plots and boxplots of the dataset, with each potential predictor variable plotted against the Pawdacity Sales for that city.

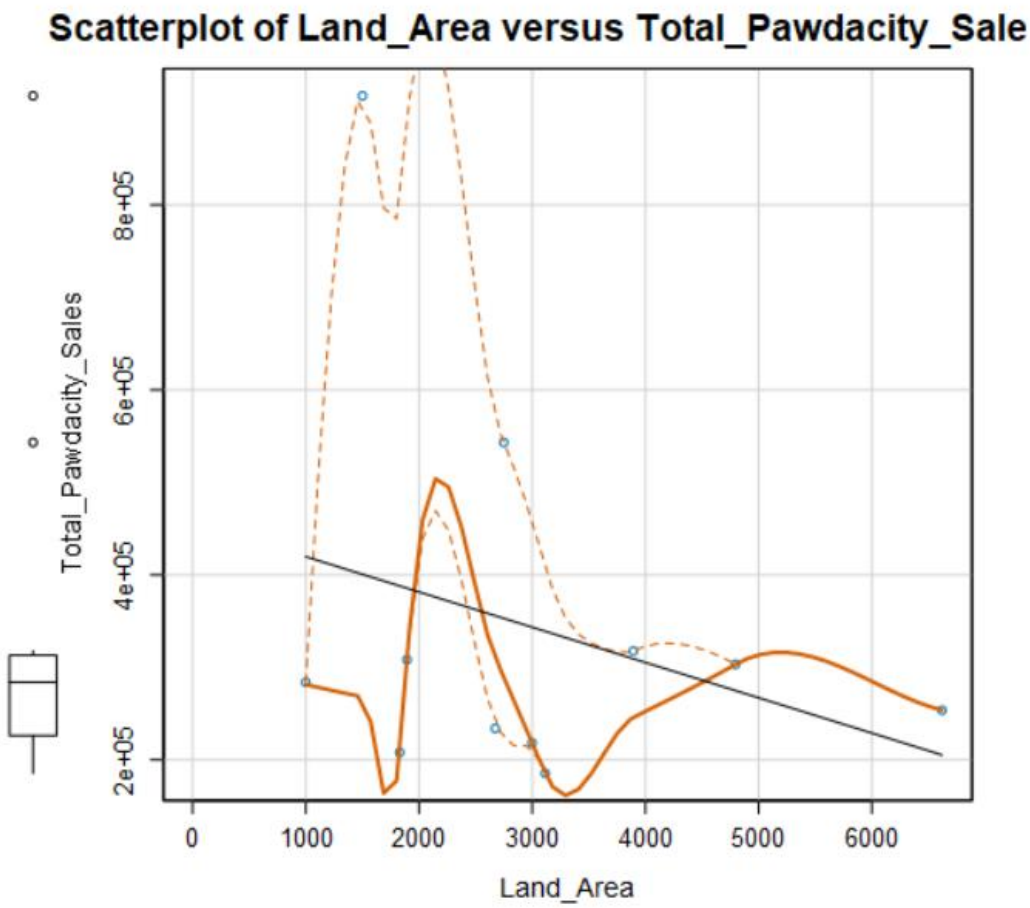



:atterplot of Population_Density versus Total_Pawdacity_
Total_Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
0
5
10
15
20
Population_Density


Scatterplot of Total_Families versus Total_Pawdacity_Sa
Total_Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
0
5000
10000
15000
Total_Families


scatterplot of CompetetorSales versus Total_Pawdacity_S
Total_Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
CompetetorSales
2e+05
4e+05
6e+05
8e+05
1e+06


rplot of Households_with_Under_18 versus Total_Pawda
Total_Pawdacity_Sales
2e+05
4e+05
6e+05
8e+05
0
2000
4000
6000
8000
Households_with_Under_18

Record	Graph
1	

Below is a summary of the dataset, with a further analysis of the interquartile ranges for the variables and their subsequent upper fence which for this project will be [1.5 * Interquartile Range] + 3rd Quartile.

Variables	25	50	75	Interquartile Range	1.5(Interquartile Range)	UpperFence	Lower Fence
Households with under 18	1327	2646	4037	2710	4065	8102	-2738
2010 Census	7917	12359	26061.5	18144.5	27216.75	53278.25	-19299.8
Land Area	1861.72	2748.853	3504.908	1643.1883	2464.78245	5969.690745	-603.062

Total Families	2923.41	5556.49	7380.805	4457.395	6686.0925	14066.8975	-3762.68
Population Density	1.72	2.78	7.39	5.67	8.505	15.895	-6.785

I will look into values that are above the “Upper Fence” for each variable.

The list below indicates max points above that of their respective “Upper Fence”:

Land Area for Rock Springs
Population Density for Cheyenne
Total Families for Cheyenne
Cheyenne city for 2010 Census

The scatterplot for Land Area vs Sales would indicate to me that Rock Springs have same sales roughly in line with other sales values of Cheyenne in that plot.

Cheyenne on the other hand has outlier with Population density, Total families and 2010 Census so I would recommend removing Cheyenne.

Creating the model.

Below is the final dataset used for the regression model.

Record #	City	2010 Census	Total Pawdacity Sales	Households with Under 18	Population Density	Total Families	CompetetorSales	Land Area
1	Buffalo	4585	185328	746	1.550000	1819.5000	[Null]	3115.507500
2	Casper	35316	317736	7788	11.160000	8756.3200	210000	3894.309100
3	Cody	9520	218376	1403	1.820000	3515.6200	[Null]	2998.956960
4	Douglas	6120	208008	832	1.460000	1744.0800	96000	1829.465100
5	Evanston	12359	283824	1486	4.950000	2712.6400	89000	999.497100
6	Gillette	29087	543132	4052	5.800000	7189.4300	1130000	2748.852900
7	Powell	6314	233928	1251	1.620000	3134.1800	[Null]	2673.574550
8	Riverton	10615	303264	2680	2.340000	5556.4900	[Null]	4796.859815
9	Rock Springs	23036	253584	4022	2.780000	7572.1800	990000	6620.201916
10	Sheridan	17444	308232	2646	8.980000	6039.7100	[Null]	1893.977048