Predicting New Store Location

**Part 1 – Cleaning the Data.**

**Business decisions.**

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. The aim of this project is to perform analysis to recommend the city for Pawdacity’s newest store, based on predicted yearly sales.

**What decisions need to be made?**

There are three sets of data:

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

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

*p2-wy-453910-naics-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.

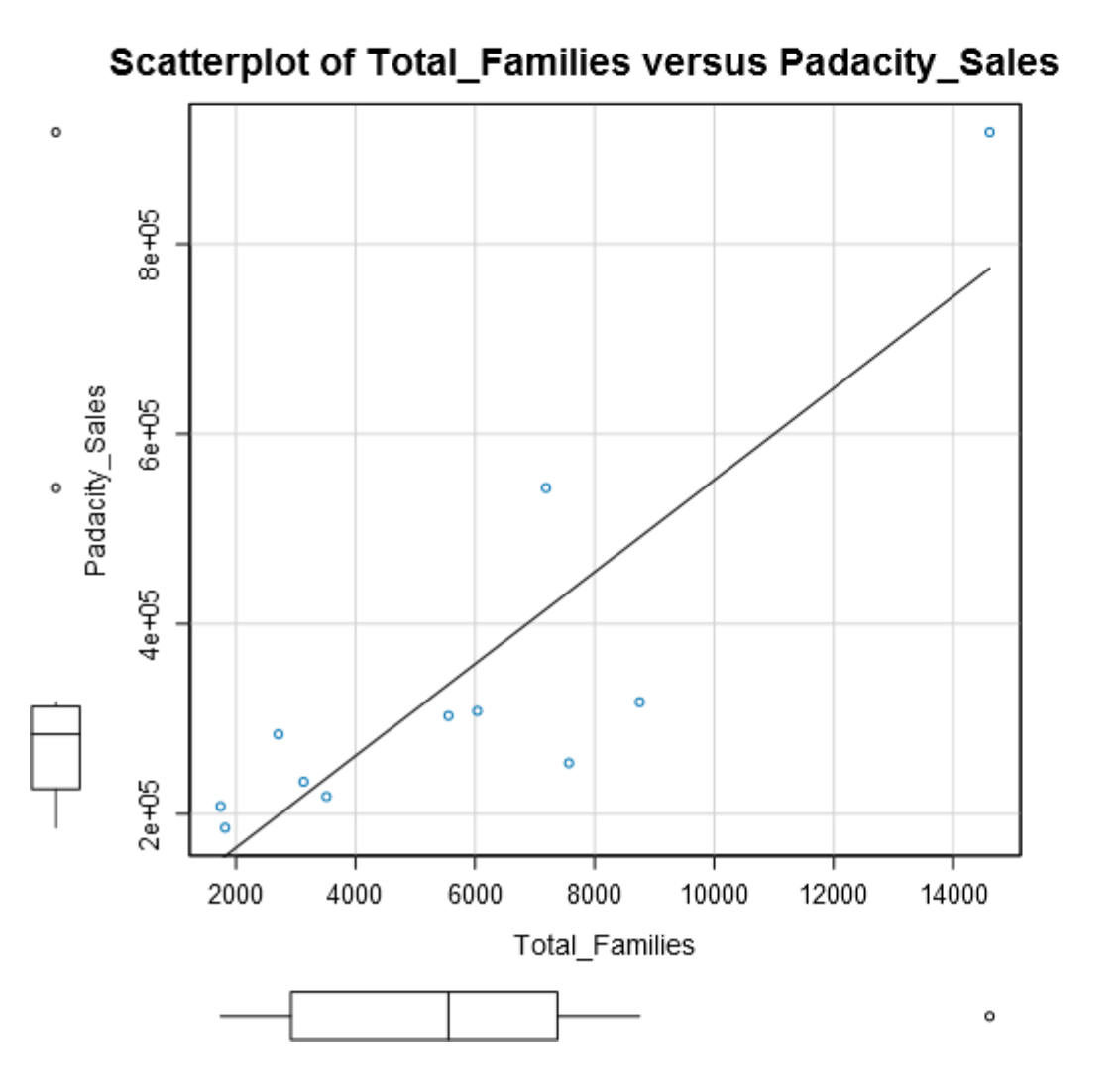
**The Dataset.**

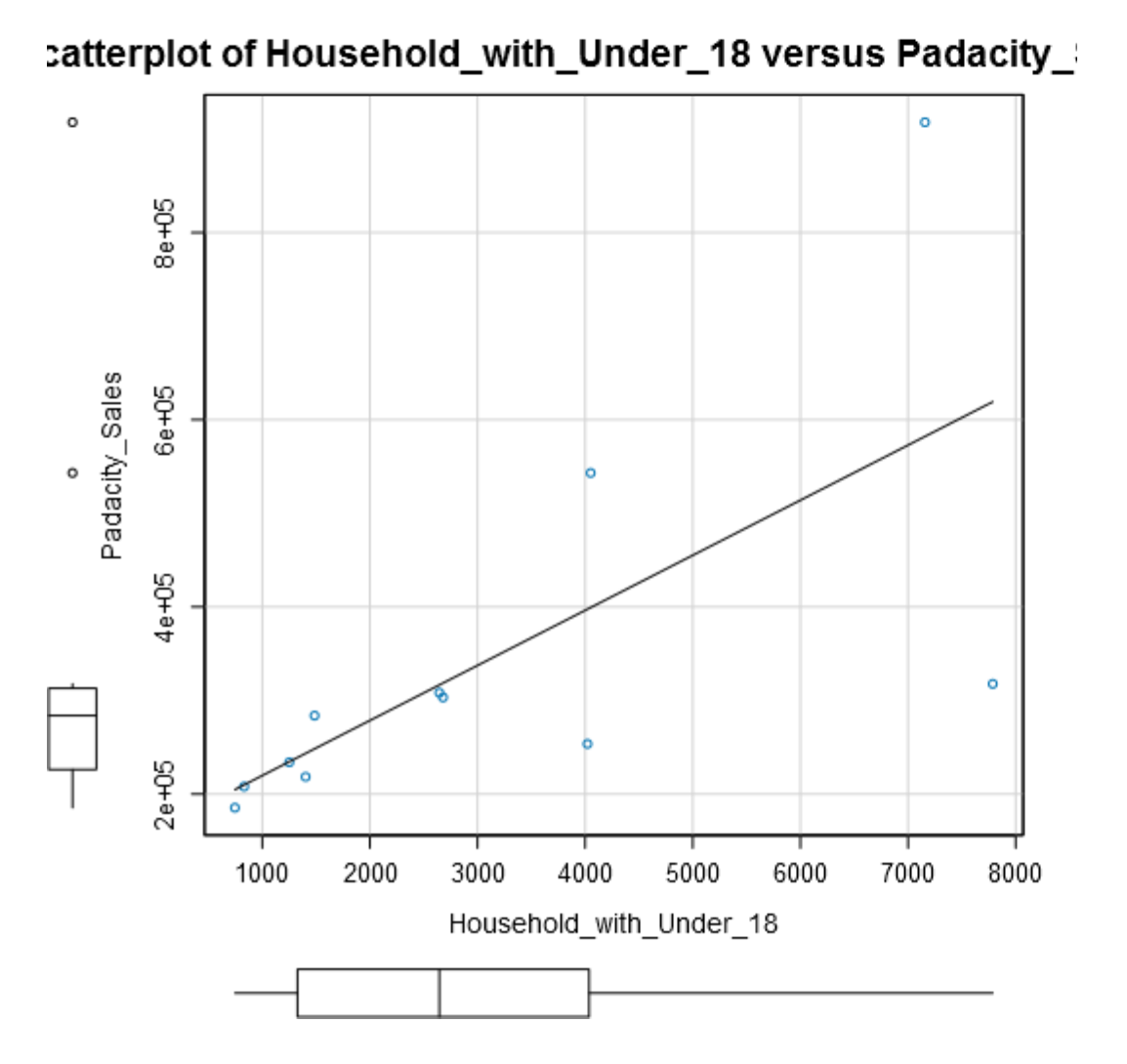
The below is a summary of the dataset.

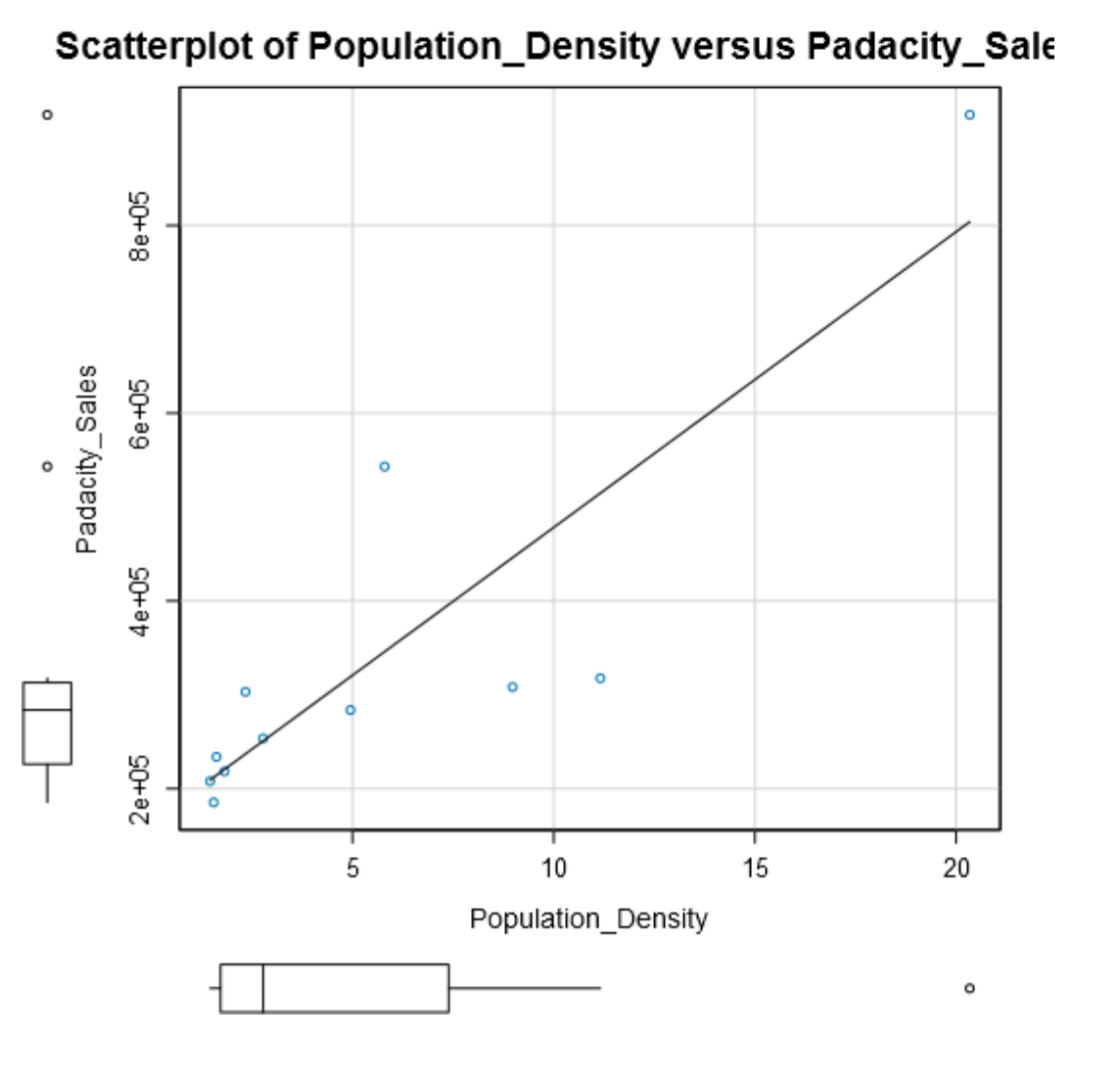
|  |  |  |
| --- | --- | --- |
| Column | Sum | Average |
| Census Population | 213862 | 19442 |
| Total Pawdacity Sales | 3773304 | 343027.64 |
| Households with Under 18 | 34064 | 3096.73 |
| Land Area | 33071 | 3006.49 |
| Population Density | 63 | 5.71 |
| Total Families | 62653 | 5695.71 |

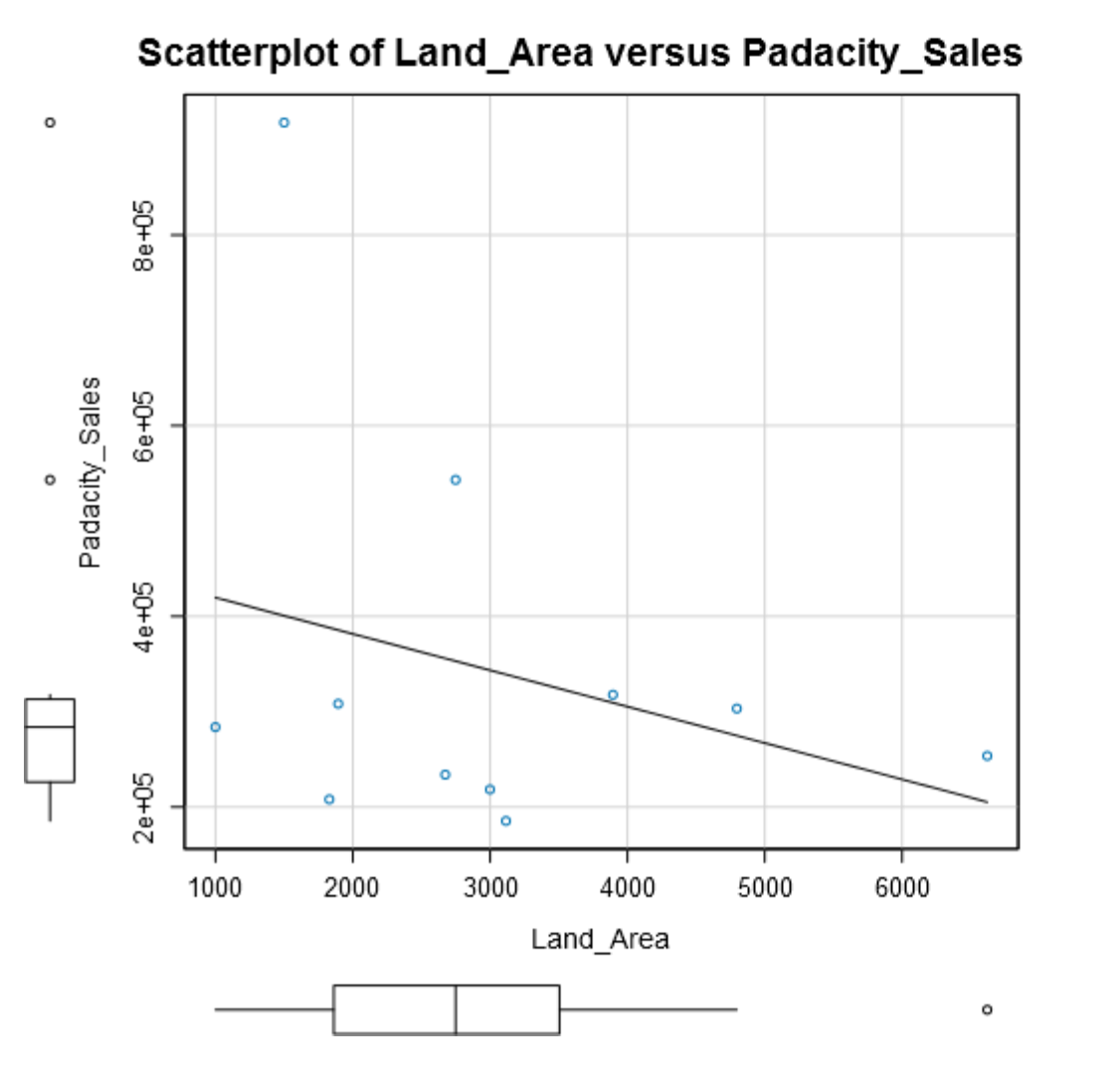
**Outliers in the dataset.**

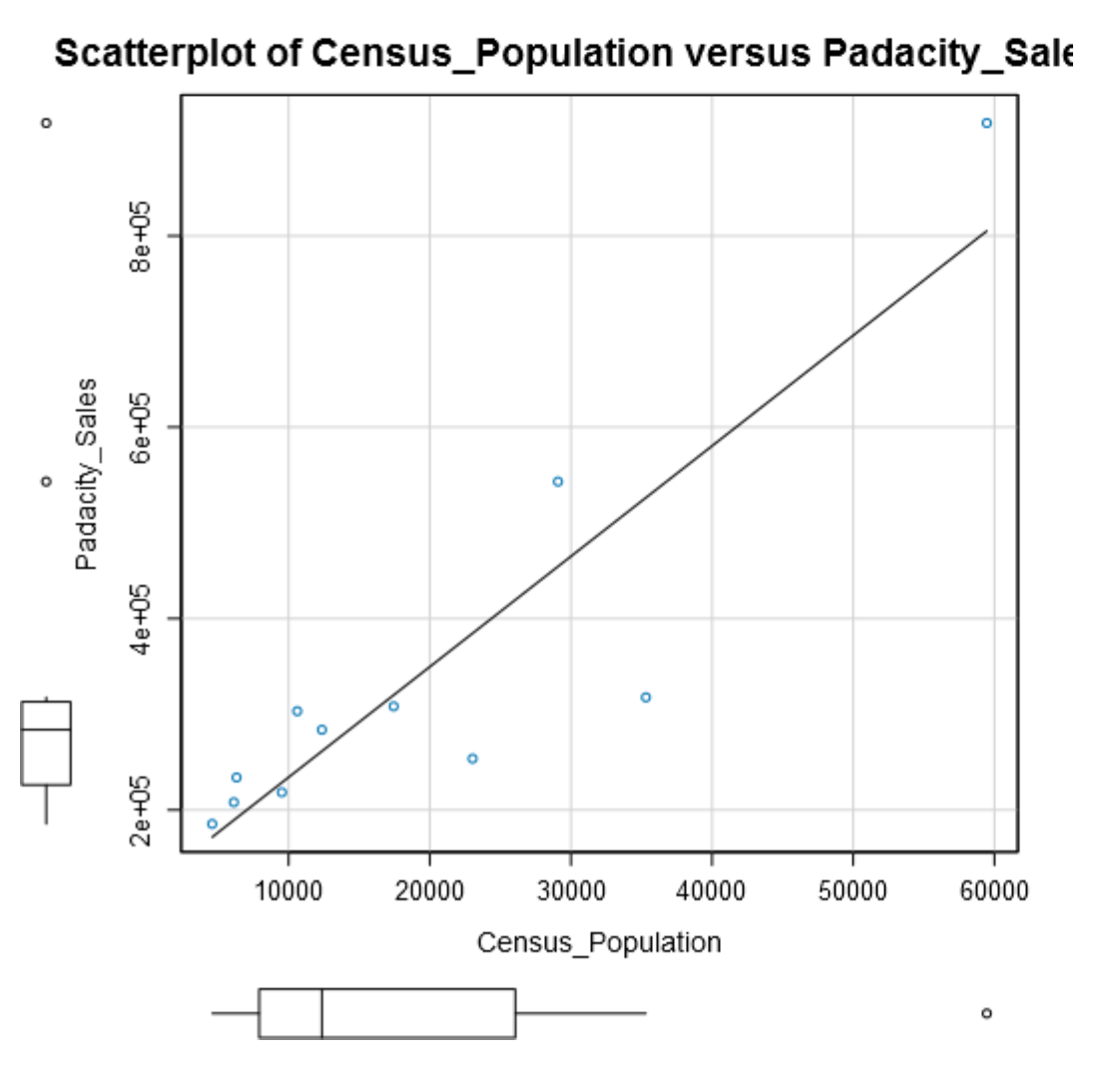
Below are scatter plots and boxplots of the dataset, with each potential predictor variable plotted against the Pawdacity Sales for that city.











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.

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

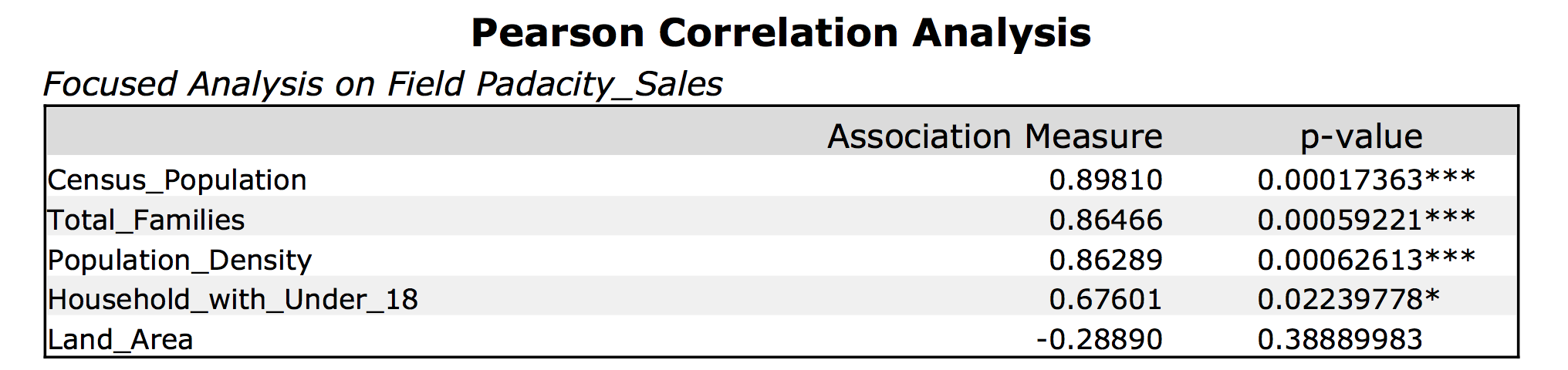
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Min** | **Max** | **Median** | **Mean** | **Std. Dev.** |
| **Census\_Population** | 4585.00 | 59466.00 | 12359.00 | 19442.00 | 16616.02 |
| **Household\_with\_Under\_18** | 746.00 | 7788.00 | 2646.00 | 3096.73 | 2453.00 |
| **Land\_Area** | 999.50 | 6620.20 | 2748.85 | 3006.49 | 1617.46 |
| **Padacity\_Sales** | 185328.00 | 917892.00 | 283824.00 | 343027.64 | 213538.71 |
| **Population\_Density** | 1.46 | 20.34 | 2.78 | 5.71 | 5.85 |
| **Total\_Families** | 1744.08 | 14612.64 | 5556.49 | 5695.71 | 3816.05 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Census\_Population\_IQR** | **Padacity\_Sales\_IQR** | **Household\_with\_Under\_18\_IQR** | **Land\_Area\_IQR** | **Population\_Density\_IQR** | **Total\_Families\_IQR** |
| 18144.50 | 86832.00 | 2710.00 | 1643.19 | 5.67 | 4457.40 |
| **Census\_Population\_Upper\_Fence** | **Padacity\_Sales\_Upper\_Fence** | **Household\_with\_Under\_18\_Upper\_Fence** | **Land\_Area\_Upper\_Fence** | **Population\_Density\_Upper\_Fence** | **Total\_Families\_Upper\_Fence** |
| 53278.25 | 443232.00 | 8102.00 | 5969.69 | 15.90 | 14066.90 |

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

|  |
| --- |
| Census Population for Cheyenne |
| Land Area for Rock Springs |
| Population Density for Cheyenne |
| Total Families for Cheyenne |
| Pawdacity Sales for Gillette and Cheyenne |

Below is a summary of the Pearson Correlation calculated from the predictor variables and the target variable which in this instance is Pawdacity Sales.



Currently, the outliers I need to investigate are Cheyenne City for Census Population, Land Area, Population Density, Rock Springs for Land Area and Pawdacity sales for Gillette.

The scatterplot for Land Area vs Sales would indicate to me that Rock Springs follows the downward direction of the line of best fit for that plot with sales roughly inline with other sales values in that plot.

Cheyenne on the other hand has two stores and their data is aggregated in this analysis which could cause it to be an outlier, however since we are looking at where to place the new store, we should look at this data at a city level. This would mean that Cheyenne justifiably is s city that produces higher sales to warrant two stores.

Gillette also has two stores, however looking through the other categories Gillette’s data looks relatively with in our outlier range except for its sales. There doesn’t seem to be a good reason for this based on the small amount of information that I know.

My recommendation here would be to keep Cheyenne and Rock Springs as I believe their data looks to be appropriate. Gillette however is harder to explain and it would be best to remove this city totally from our data set, however I am reluctantly removing Gillette due to the fact we already have a small amount of data.