**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.

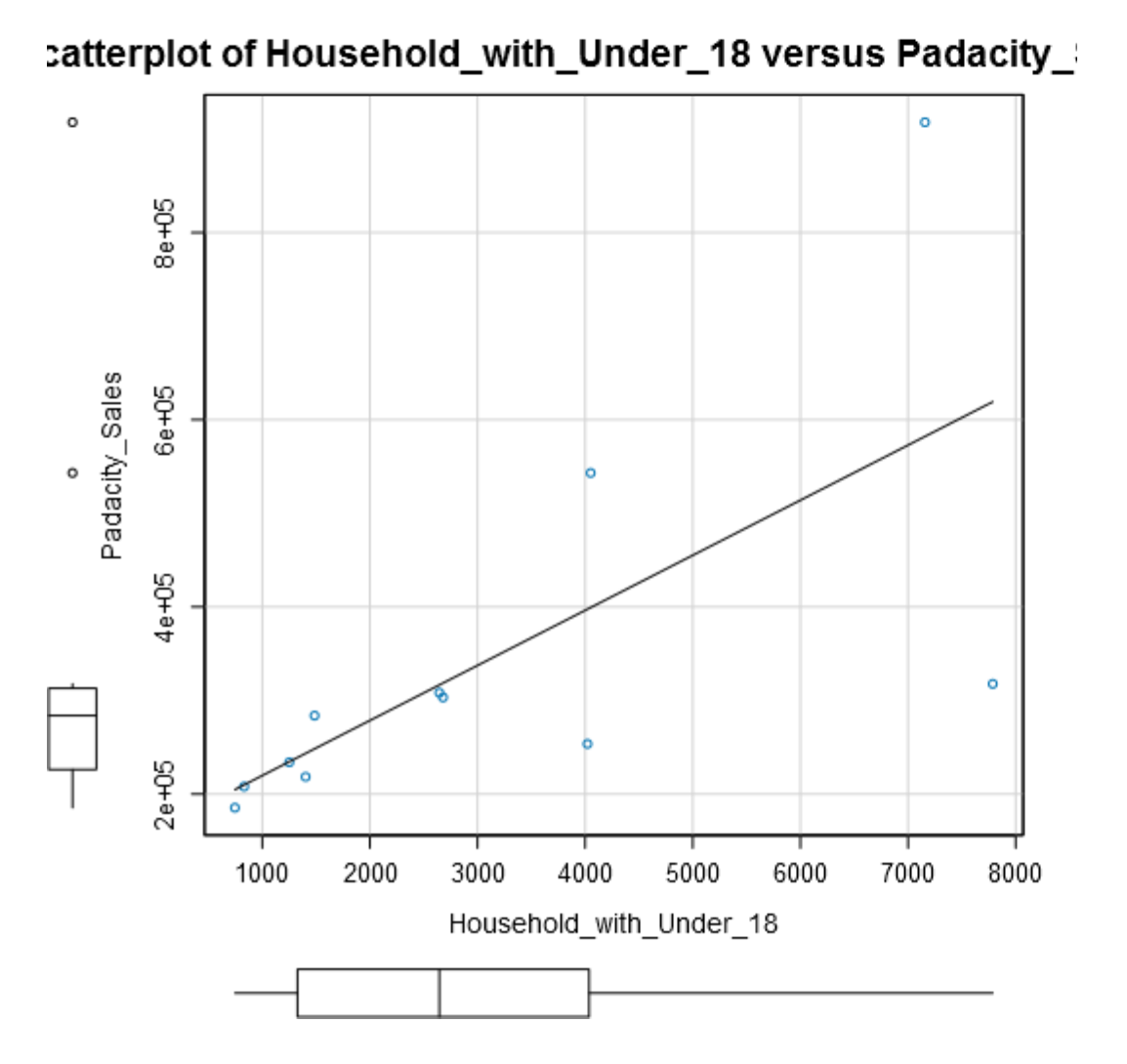
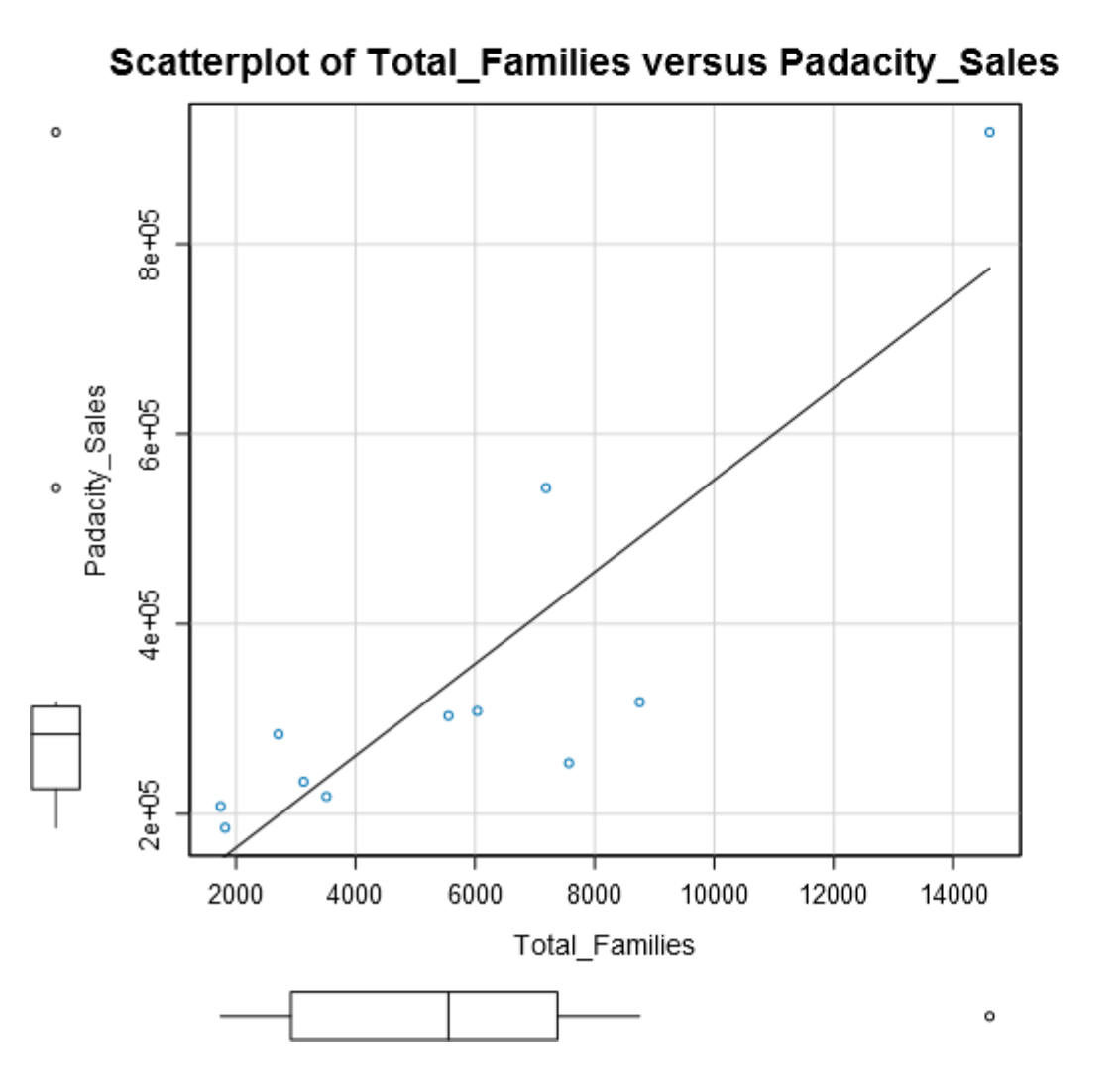
**The Dataset.**

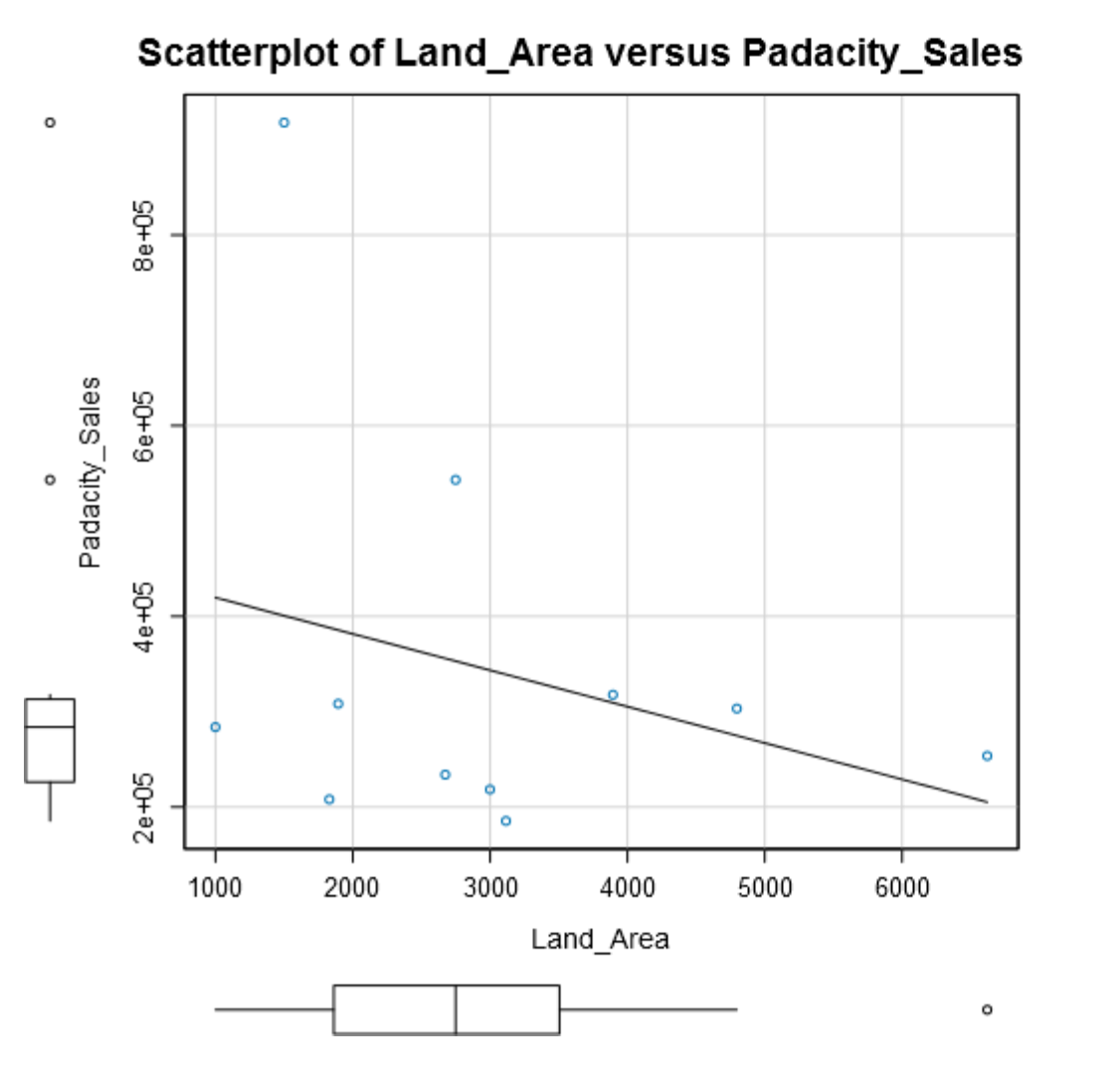
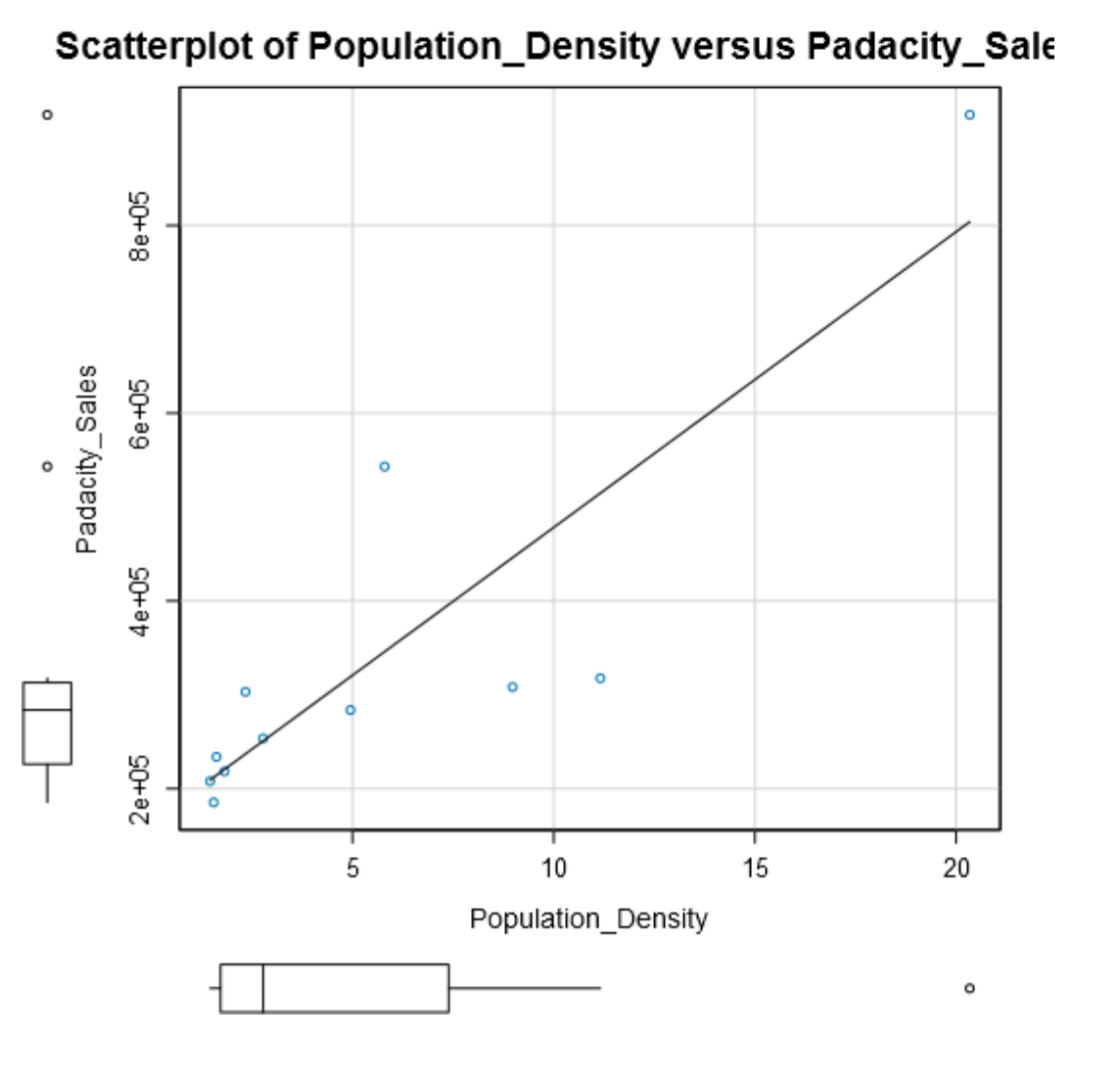
The below is a summary of the dataset.

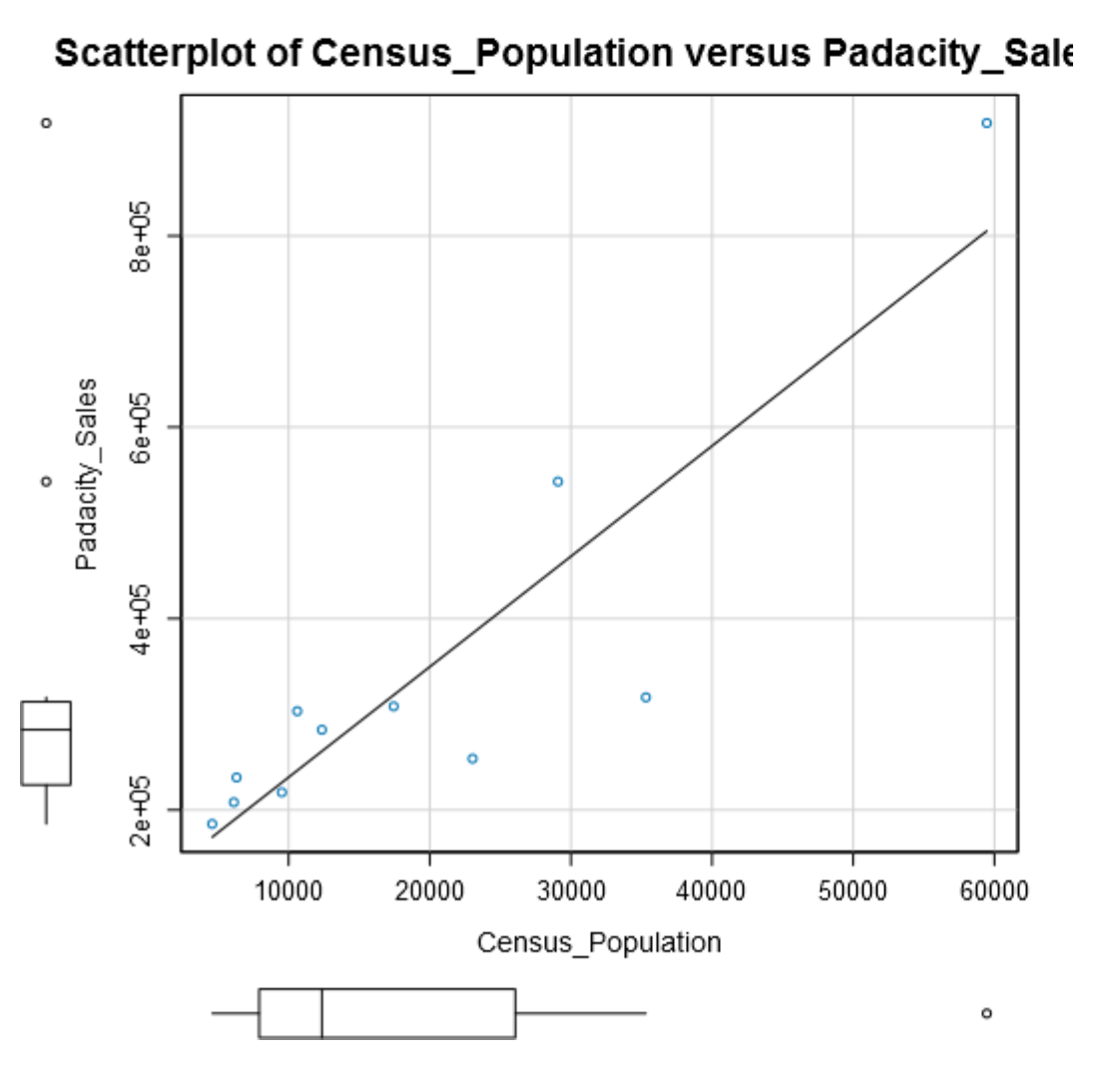
|  |  |  |
| --- | --- | --- |
| Column | Sum | Average |
| Census Population | 213862 | 19442 |
| Total Pawdacity Sales | 3773304 | 343027.64 |
| Households with Under 18 | 34064 | 3096.73 |
| Land Area | 33071 | 3006.49 |
| Population Density | 63 | 5.71 |
| Total Families | 62653 | 5695.71 |

**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.

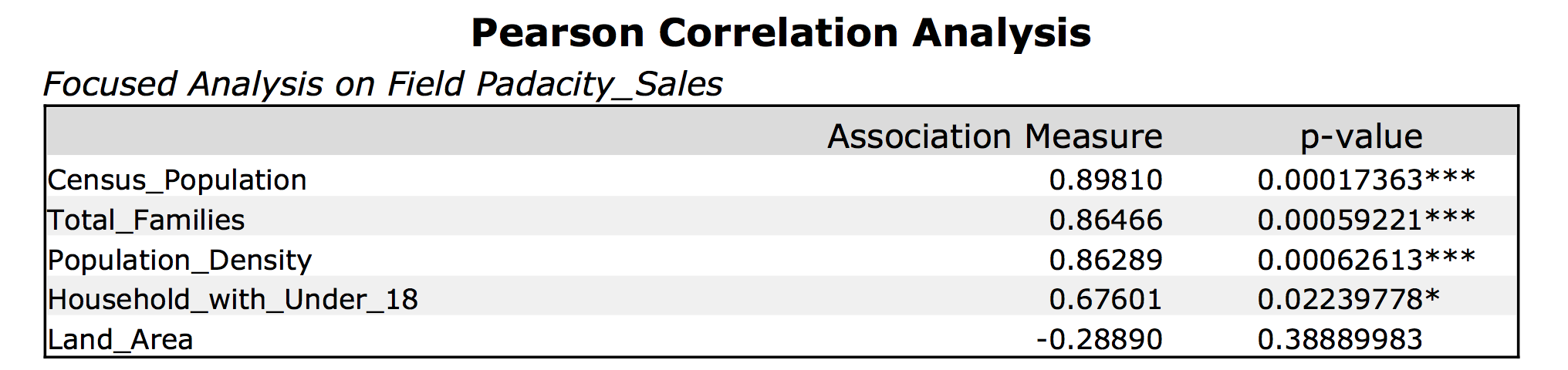
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Min** | **Max** | **Median** | **Mean** | **Std. Dev.** |
| **Census\_Population** | 4585.00 | 59466.00 | 12359.00 | 19442.00 | 16616.02 |
| **Household\_with\_Under\_18** | 746.00 | 7788.00 | 2646.00 | 3096.73 | 2453.00 |
| **Land\_Area** | 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.

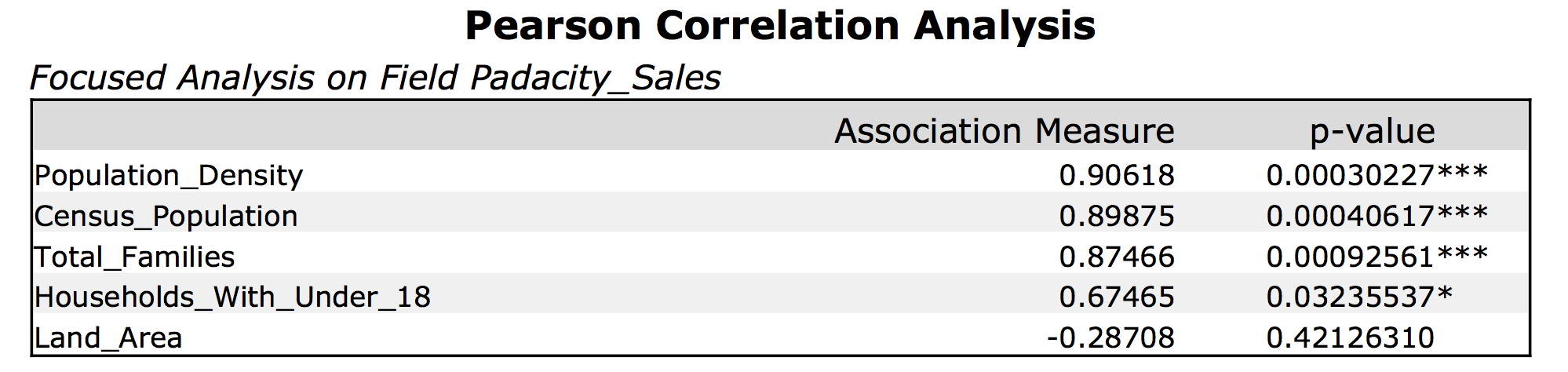
**Creating the model.**

Below is the final dataset used for the regression model.

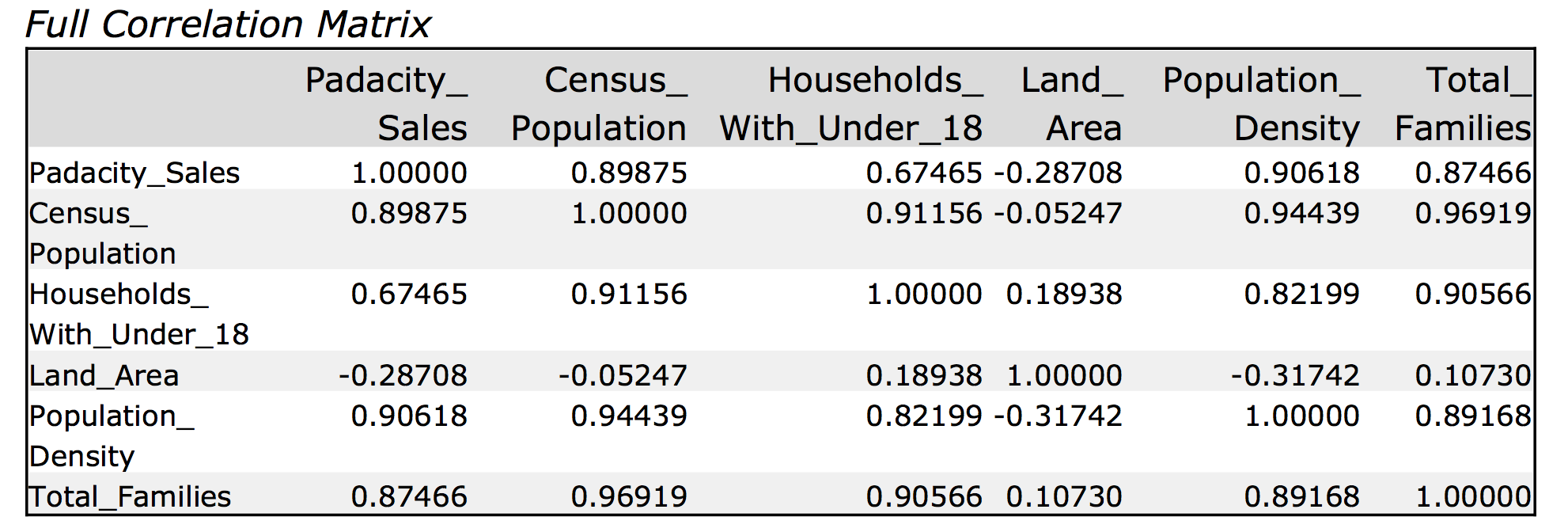
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **City** | **Census\_Population** | **Household\_with\_Under\_18** | **Land\_Area** | **Padacity\_Sales** | **Population\_Density** | **Total\_Families** |
| **Buffalo** | 4585 | 746 | 3115.5075 | 185328 | 1.55 | 1819.5 |
| **Casper** | 35316 | 7788 | 3894.3091 | 317736 | 11.16 | 8756.32 |
| **Cheyenne** | 59466 | 7158 | 1500.1784 | 917892 | 20.34 | 14612.64 |
| **Cody** | 9520 | 1403 | 2998.95696 | 218376 | 1.82 | 3515.62 |
| **Douglas** | 6120 | 832 | 1829.4651 | 208008 | 1.46 | 1744.08 |
| **Evanston** | 12359 | 1486 | 999.4971 | 283824 | 4.95 | 2712.64 |
| **Powell** | 6314 | 1251 | 2673.57455 | 233928 | 1.62 | 3134.18 |
| **Riverton** | 10615 | 2680 | 4796.859815 | 303264 | 2.34 | 5556.49 |
| **Rock Springs** | 23036 | 4022 | 6620.201916 | 253584 | 2.78 | 7572.18 |
| **Sheridan** | 17444 | 2646 | 1893.977048 | 308232 | 8.98 | 6039.71 |

**Selecting the predictor variables.**

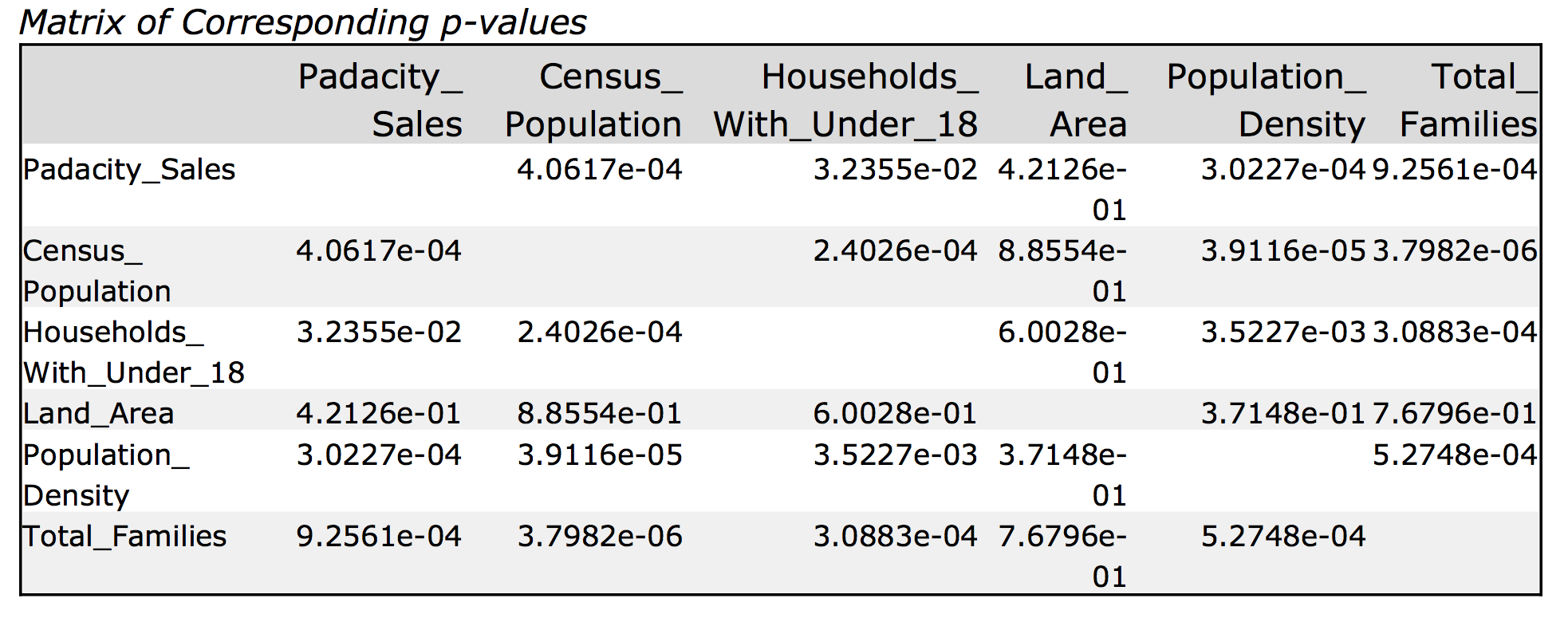
Below is a table of all the variables and their Pearson correlation.



Full correlation matrix.

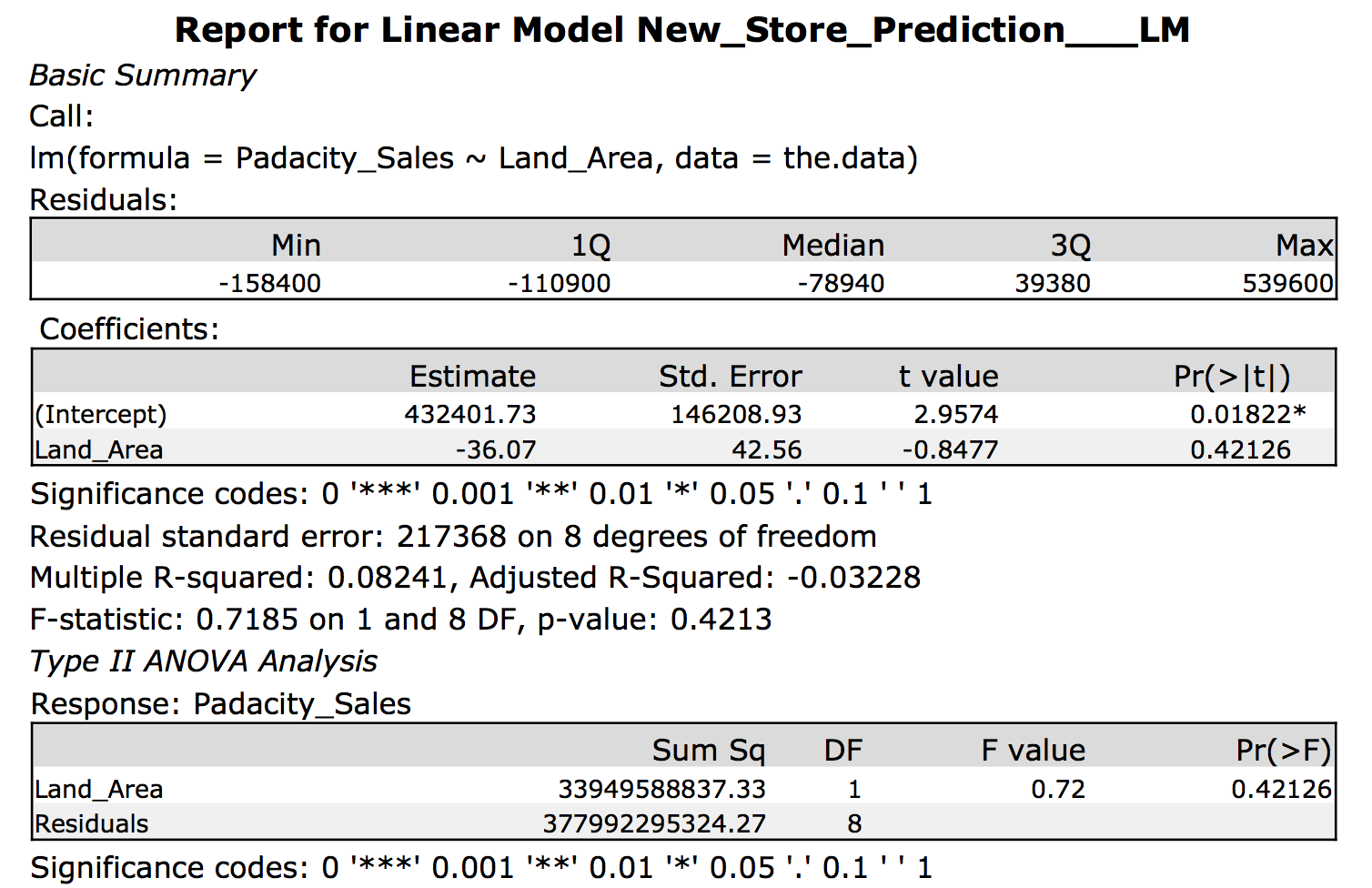


Matrix of p-values for Predictor Variables.

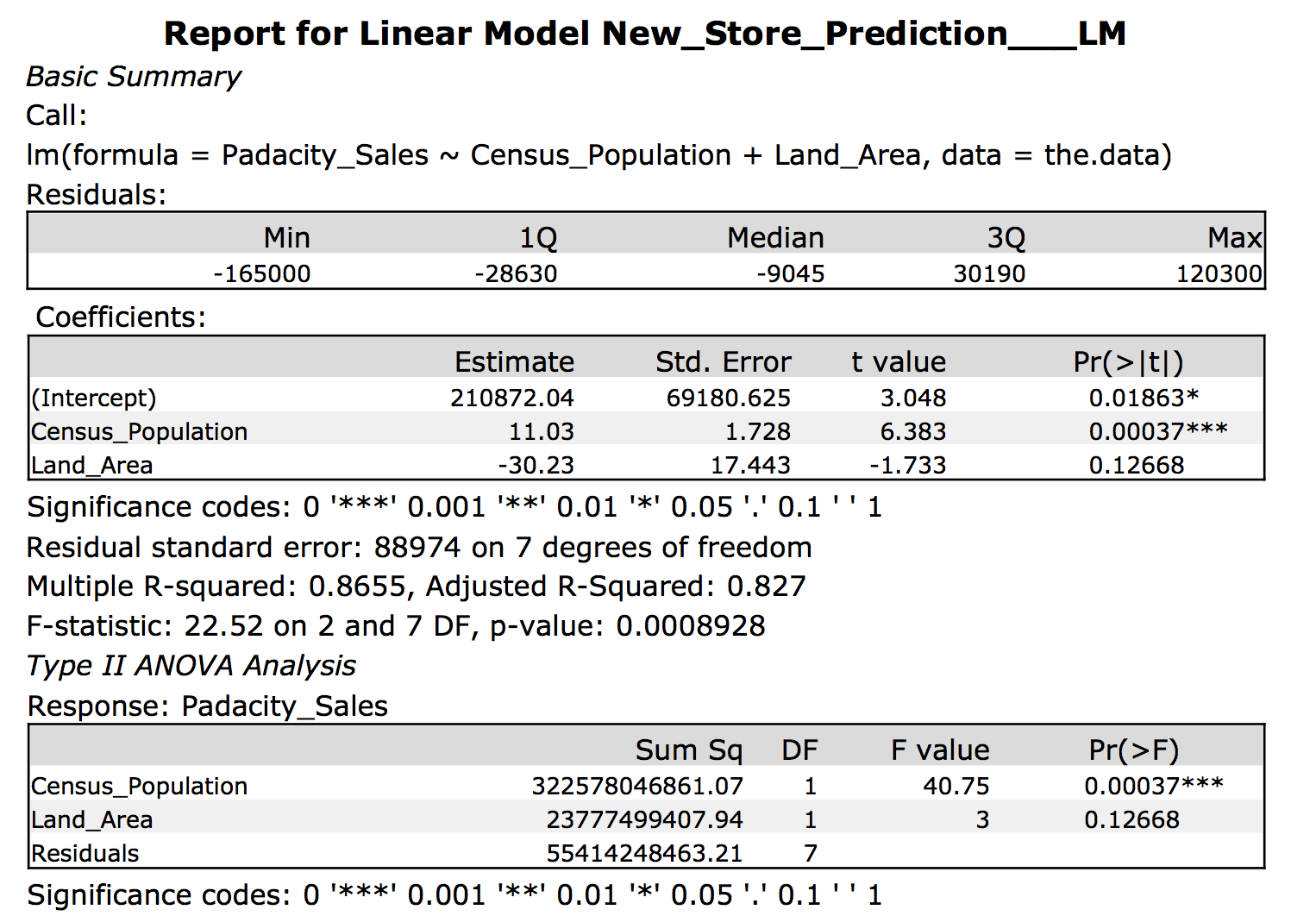


The full correlation matrix shows good correlation between predictor variables, Census\_Population, Households\_with\_Under\_18, Population\_Density and Total\_Families. There may be some multicollinearity here.

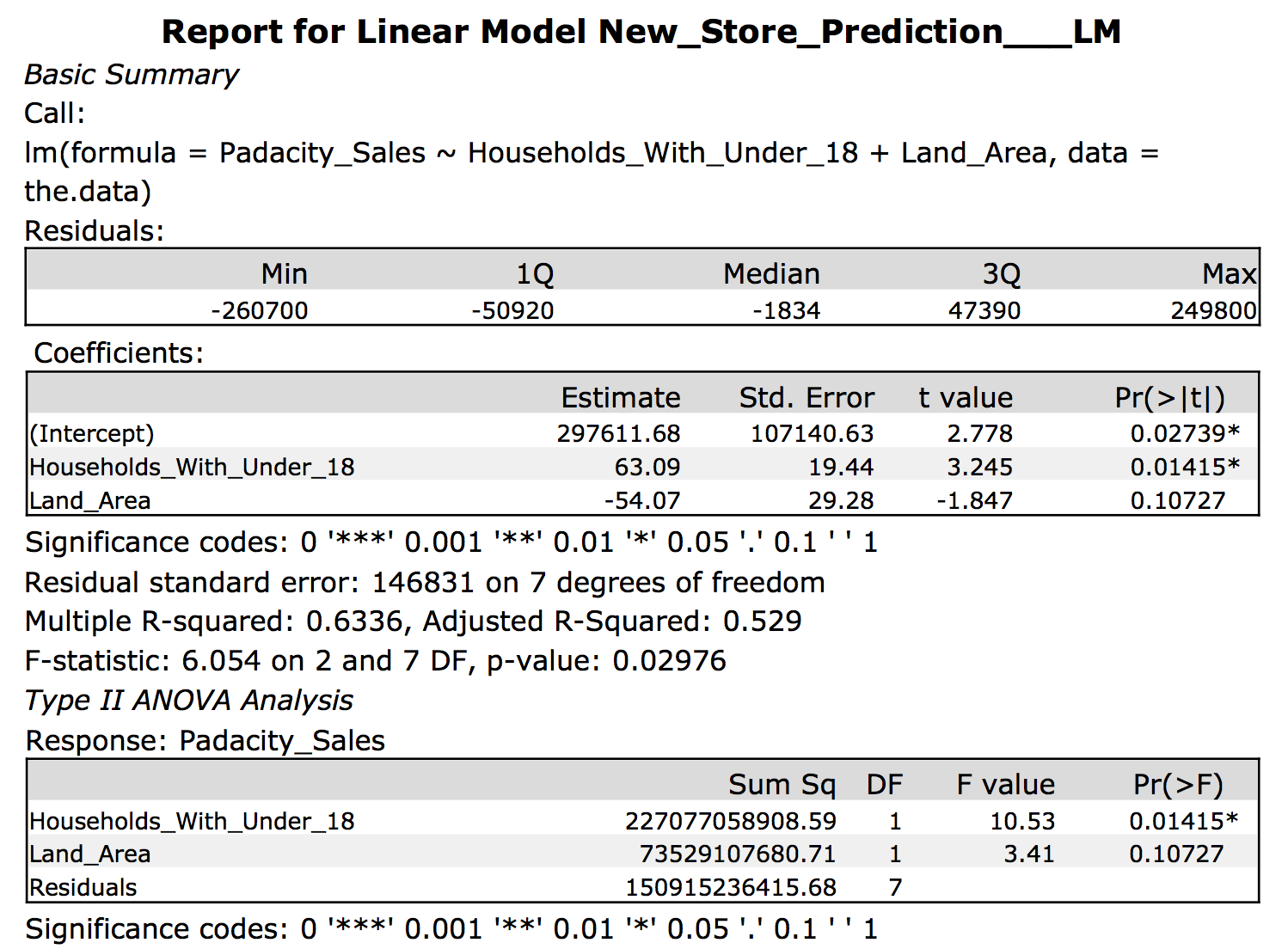
Land\_Area does not show great correlation with the other predictor variables so I will start by running a regression with Land\_Area and add other predictor variables to the regression.



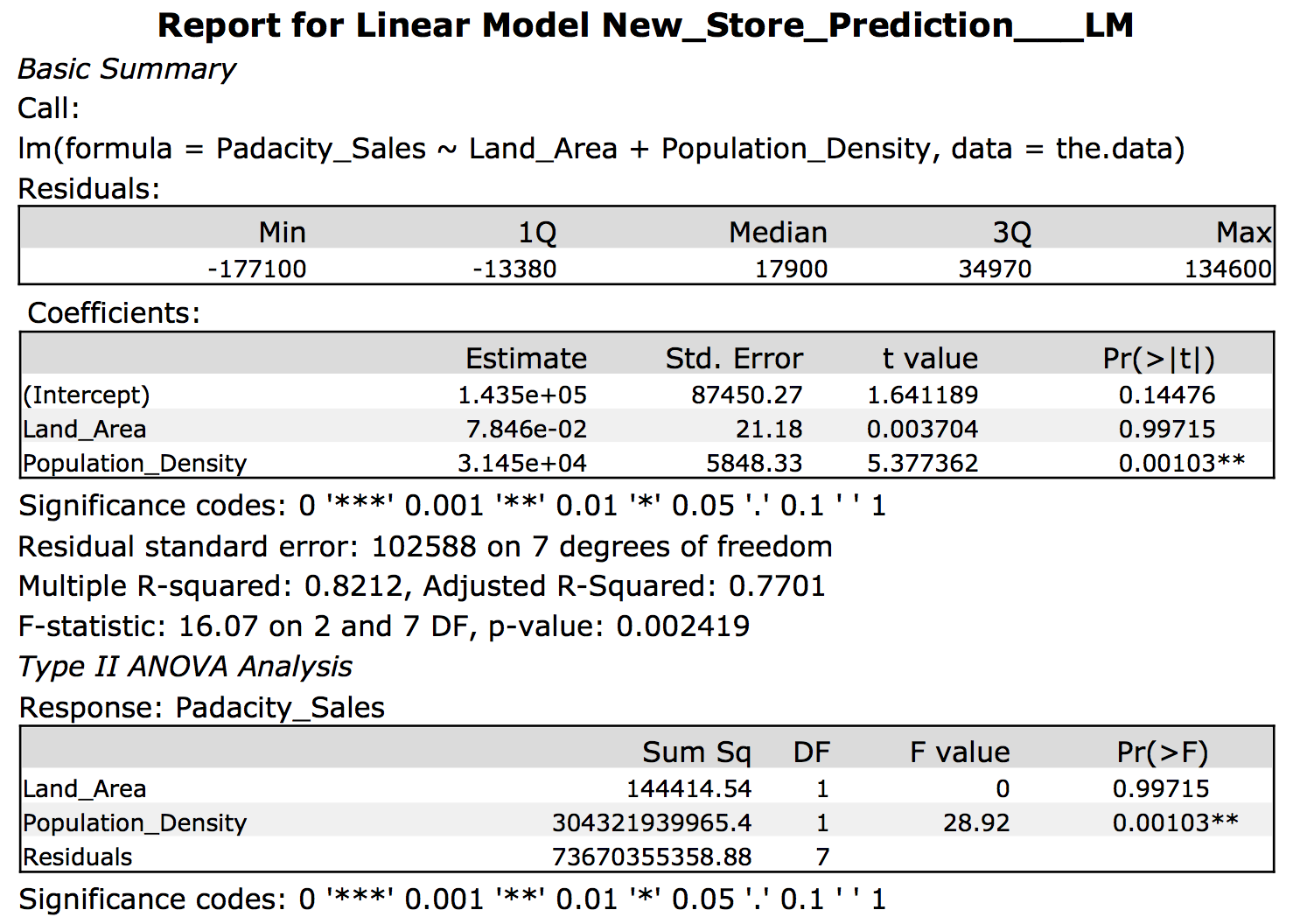
R-Squared for linear model between Sales vs Land\_Area = **0.08241**



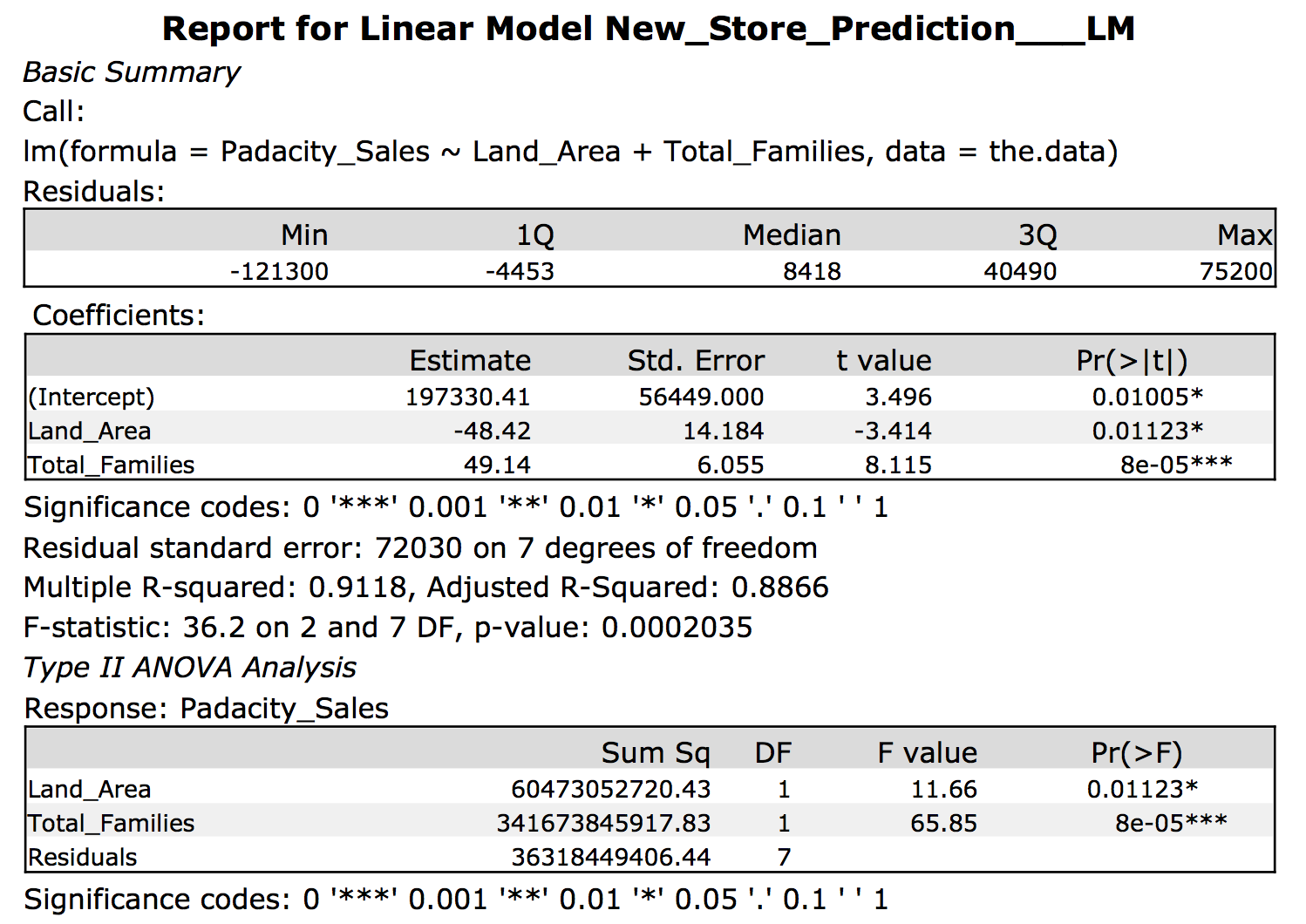
Adjusted R-Squared for linear model between Sales vs Land\_Area vs Census\_Population = **0.827**



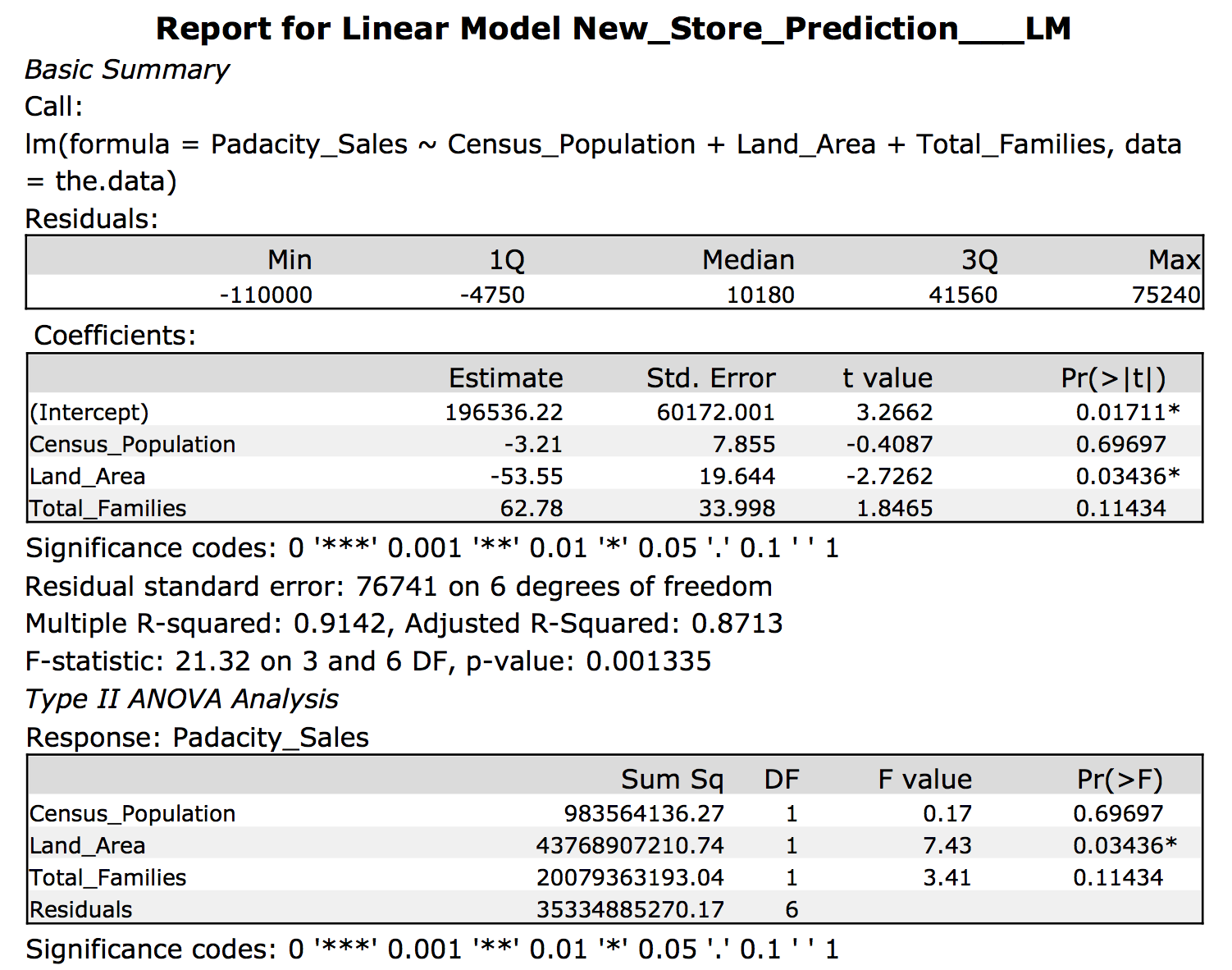
Adjusted R-Squared for linear model between Sales vs Land\_Area vs Households\_With\_Under\_18 = **0.529**



Adjusted R-Squared for linear model between Sales vs Land\_Area vs Population\_Density = **0.7701**

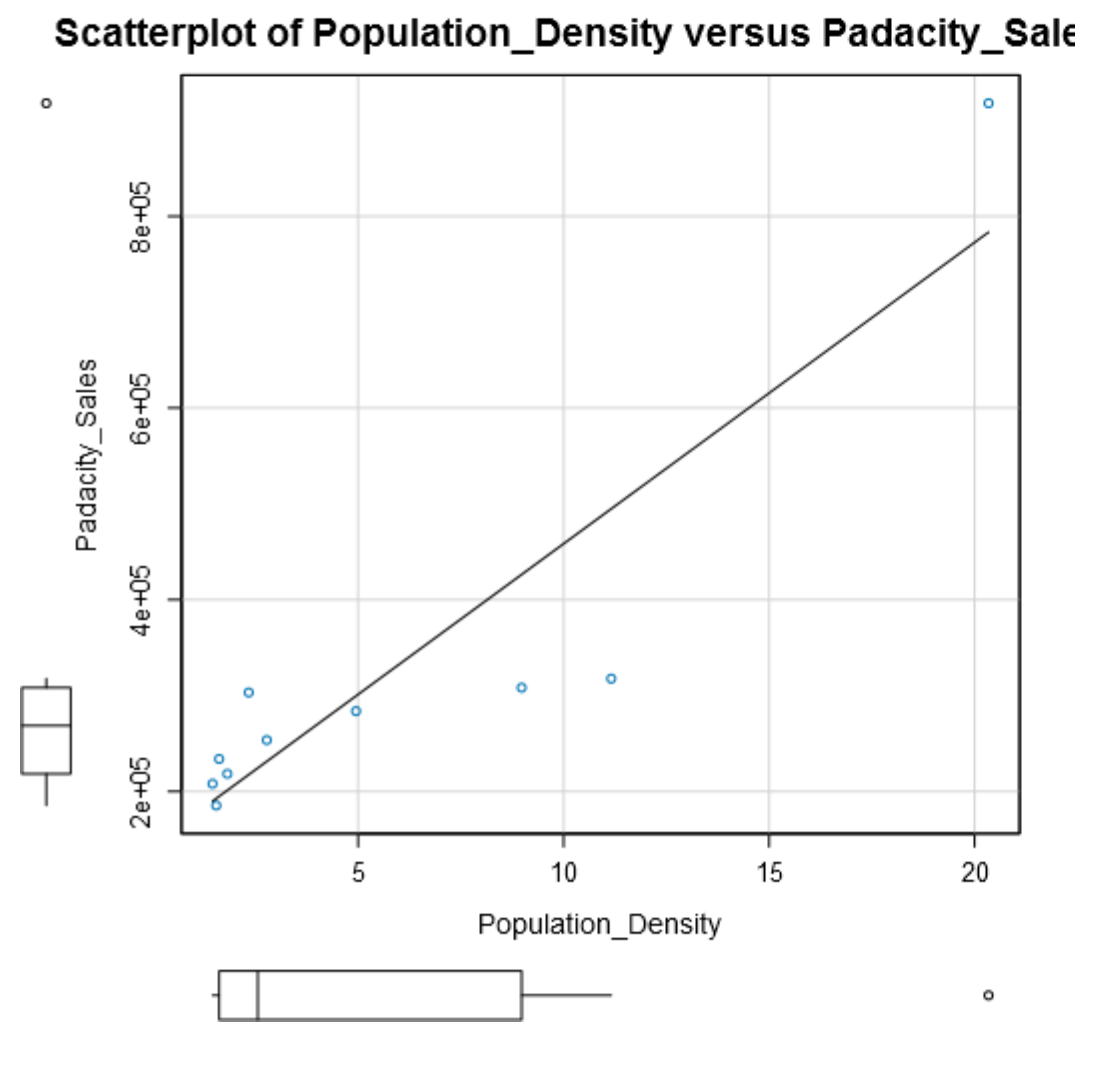
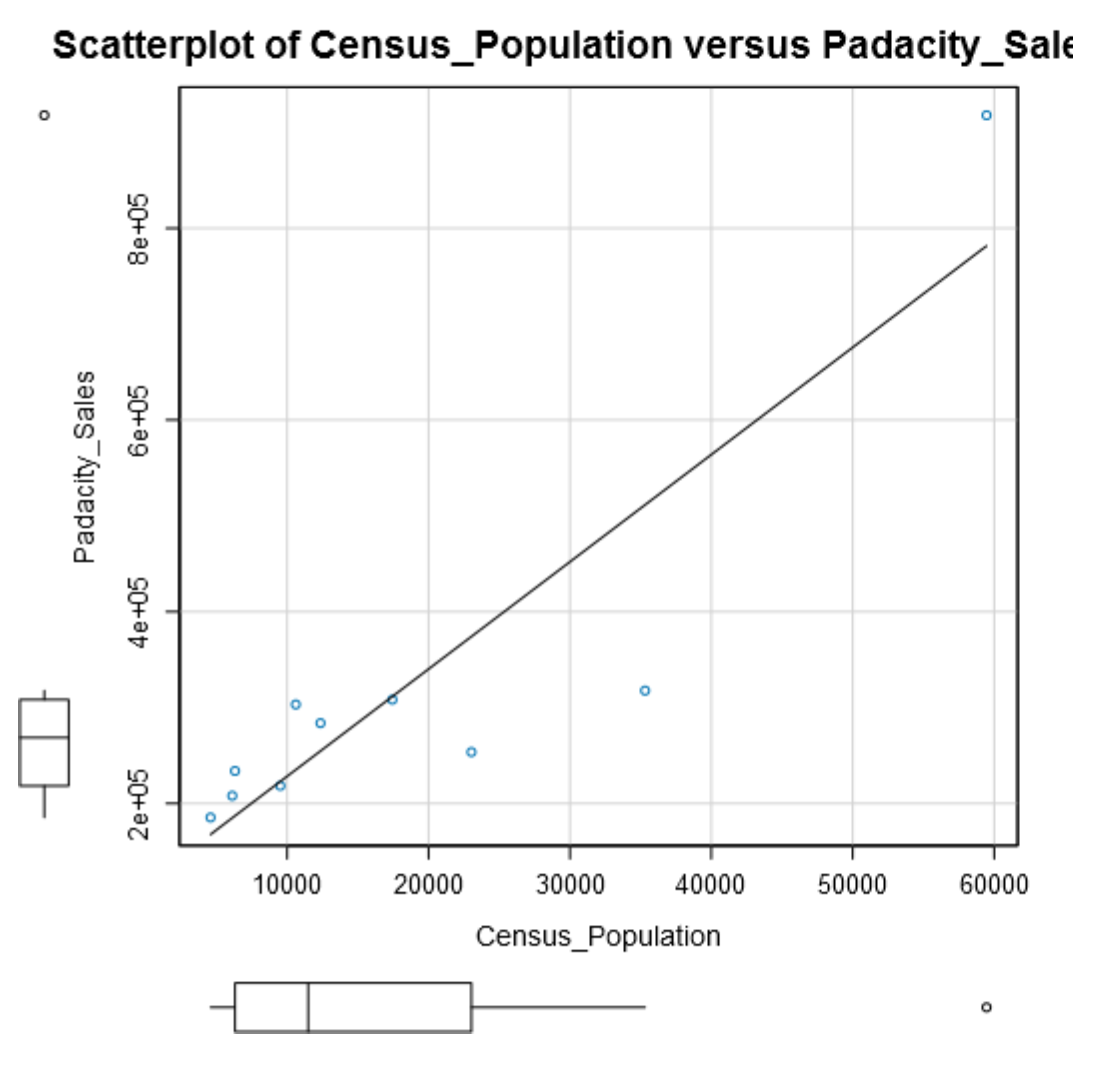


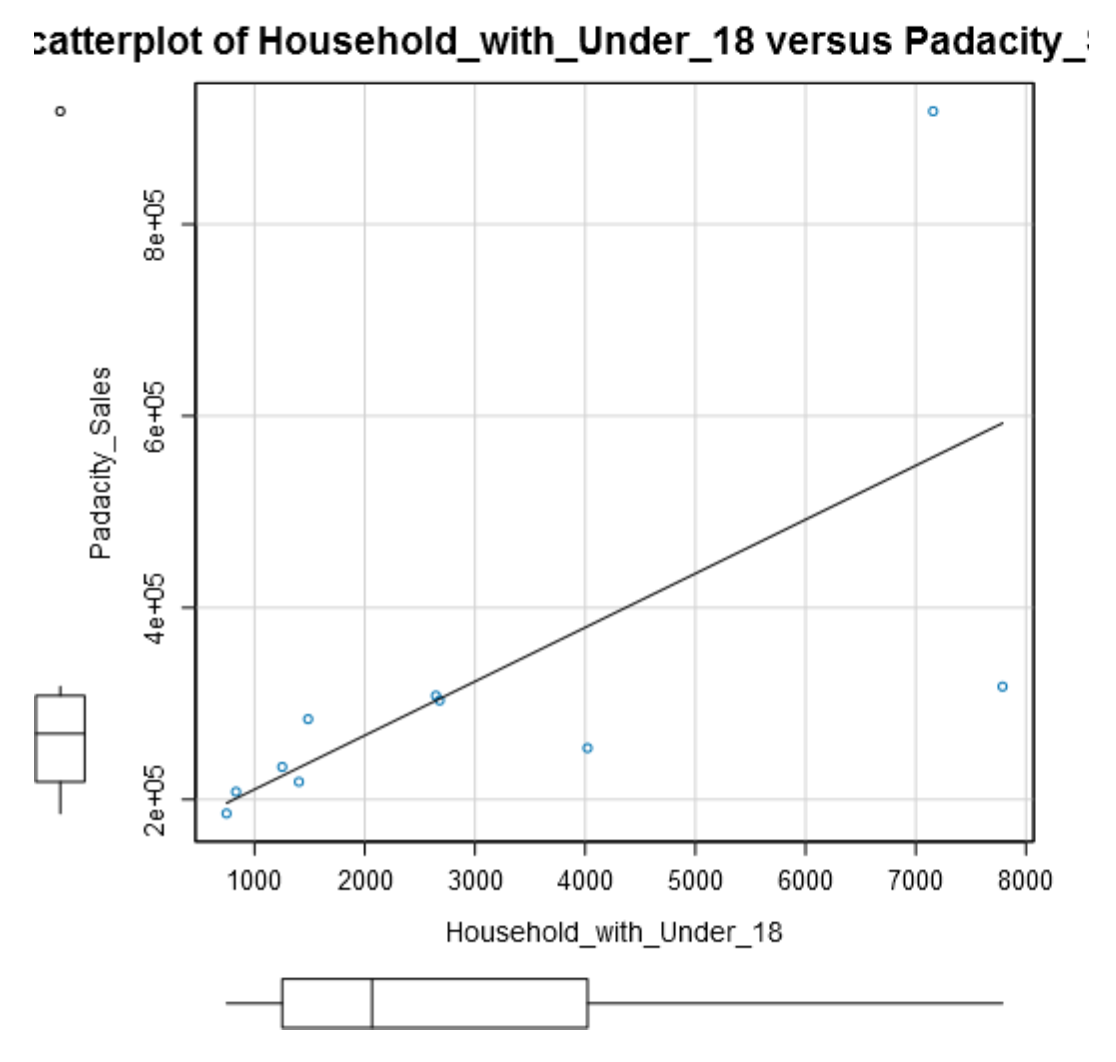
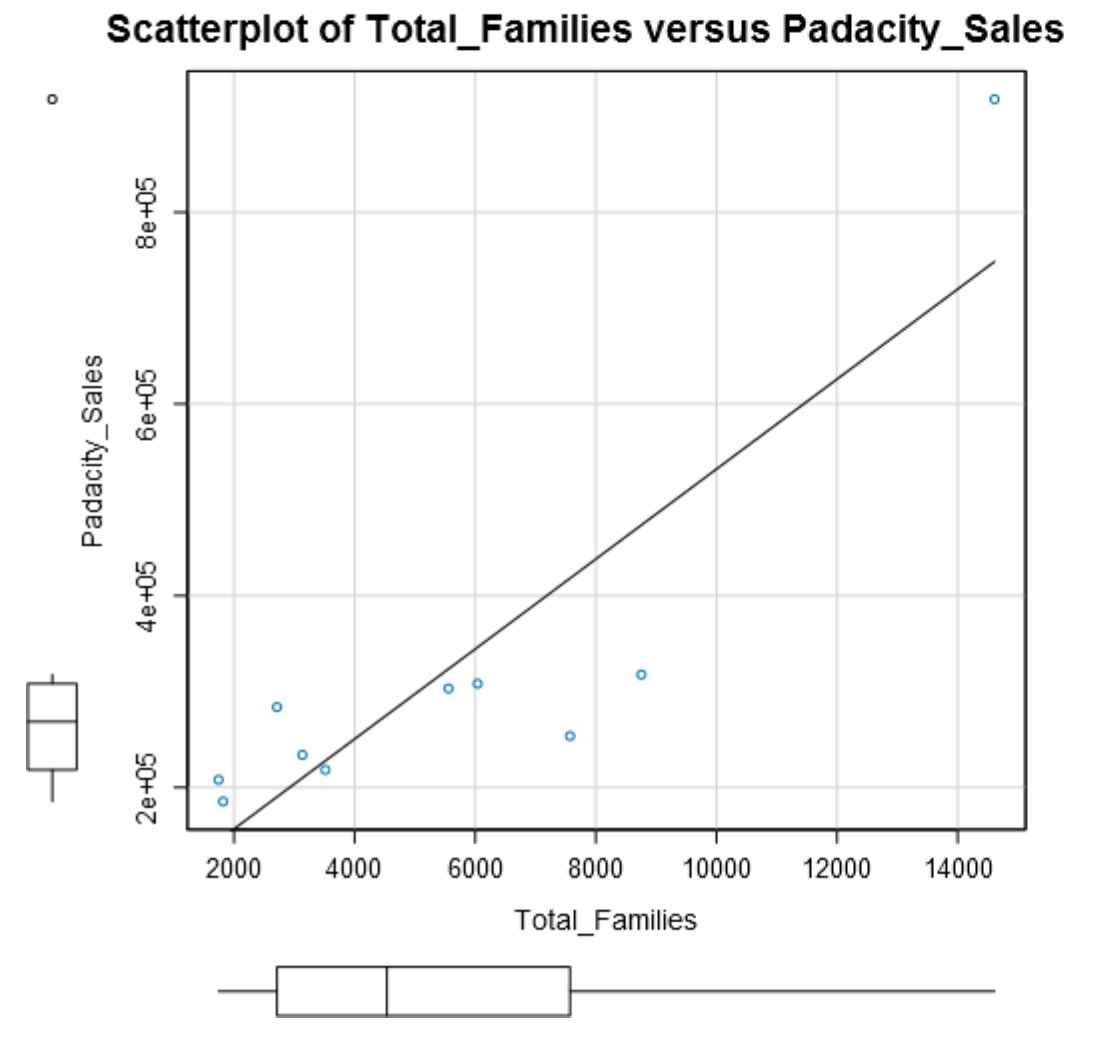
Adjusted R-Squared for linear model between Sales vs Land\_Area vs Total\_Families = **0.8866**

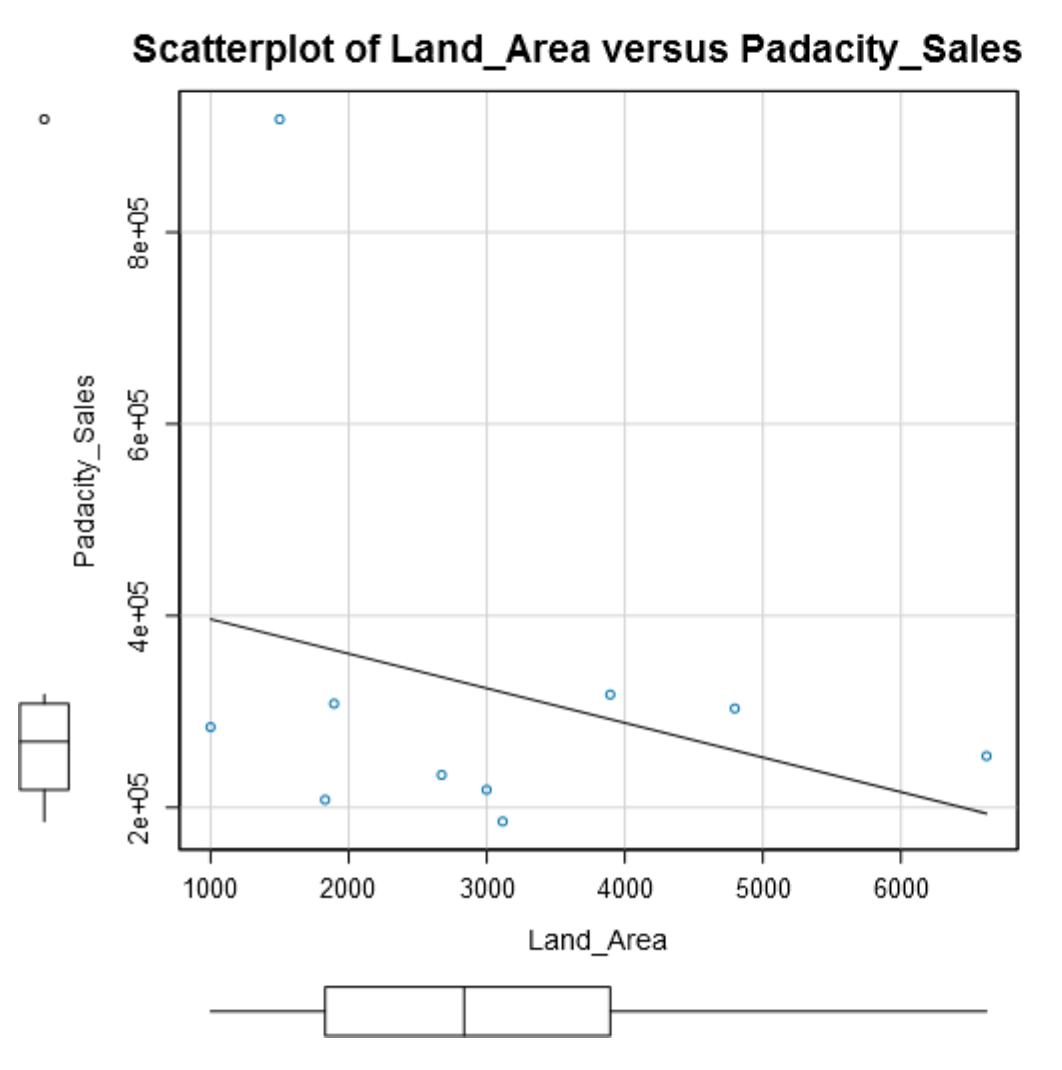


Adjusted R-Squared for linear model between Sales vs Land\_Area vs Total\_Families vs Census\_Population = **0.8713**

Below are the scatter plots of the predictor variables vs our target variable (Pawdacity\_Sales).





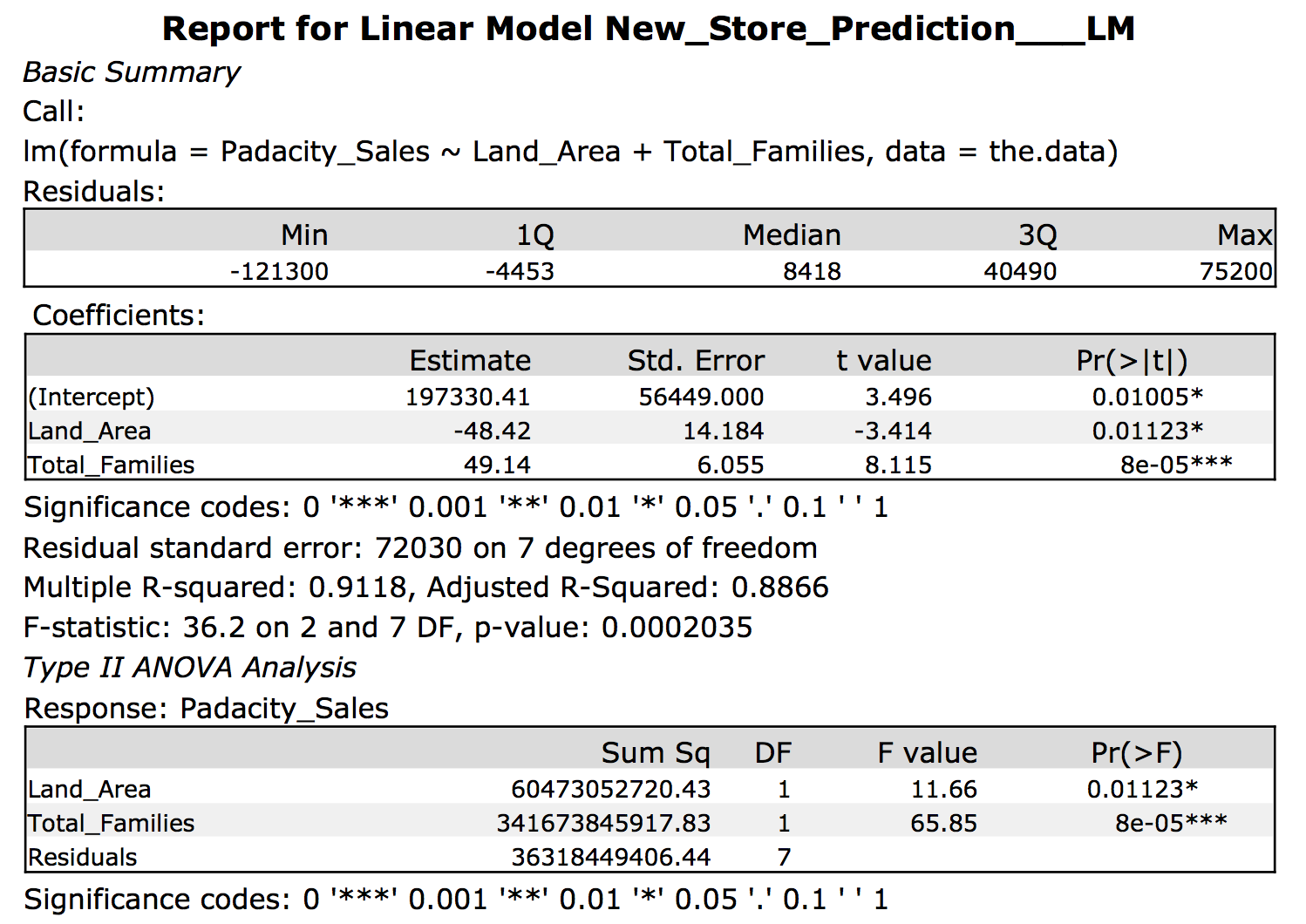


The scatter plots above give a good representation of the linearity between the target variable and its respective predictor variable.

Starting with Land\_Area as a predictor variable (R-Squared = 0.08241) and adding the other variables, I can see that the largest jump in R-Squared comes from Land\_Area and Total\_Families (adjusted r-squared = 0.8866)

I will use Land\_Area and Total\_Families as my predictor variables for my linear model.

**Below is the summary of the multilinear regression model.**



From the summary, the equation for the linear regression model is:

*Y (Pawdacity\_Sales) = 197330.41 – 48.42(Land\_Area) + 49.14(Total\_Families)*

**Final recommendation.**

Here are the criteria’s given to you in choosing the right city:

1. The new store should be located in a new city. That means there should be no existing stores in the new city.
2. The total sales for the entire competition in the new city should be less than $500,000
3. The new city where you want to build your new store must have a population over 4,000 people (based upon the 2014 US Census estimate).
4. The predicted yearly sales must be over $200,000.
5. The city chosen has the highest predicted sales from the predicted set.

With the required criteria, I would recommend Laramie City. Laramie City does not currently contain a store, has an estimated census population for 2014 of 32,081 and predicted sales of **$305,013.88**.

Below is a summary of the final possibilities for a new store with the highlighted row as the recommendation.

|  |  |  |  |
| --- | --- | --- | --- |
| **City** | **2014\_Census\_Pop\_Est** | **Total\_Families** | **Score** |
| **Laramie** | 32081.00 | 4668.93 | 305013.88 |
| **Torrington** | 6736.00 | 2548.50 | 245081.79 |
| **Jackson** | 10449.00 | 2313.08 | 225870.82 |
| **Lander** | 7642.00 | 3876.81 | 225751.40 |
| **Green River** | 12630.00 | 3977.40 | 224372.00 |
| **Worland** | 5366.00 | 1364.32 | 201700.33 |