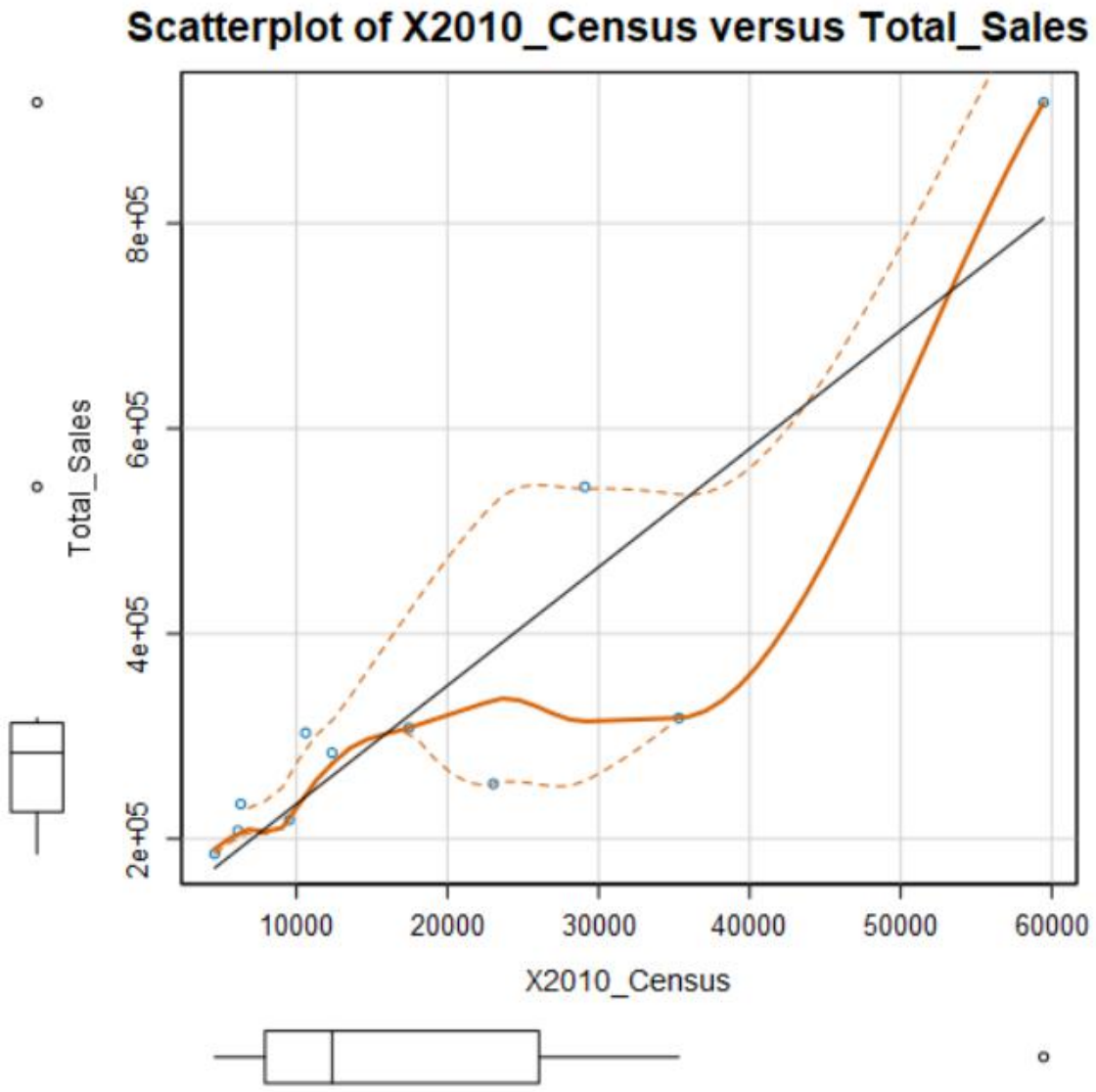
Project 2.2: Recommend a City of a New Pet Store

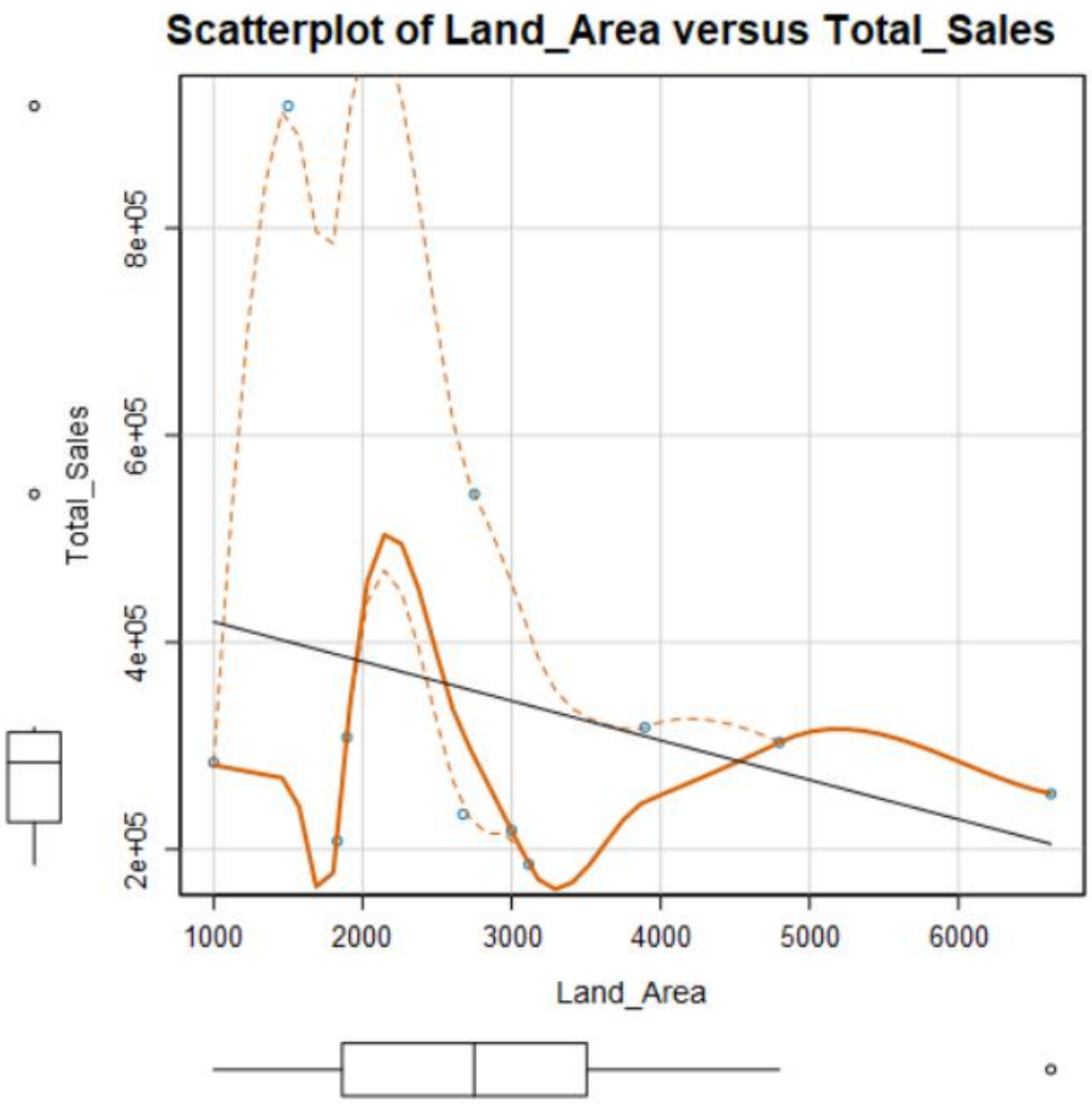
Note that this project is a continuation of the Data Cleanup project.

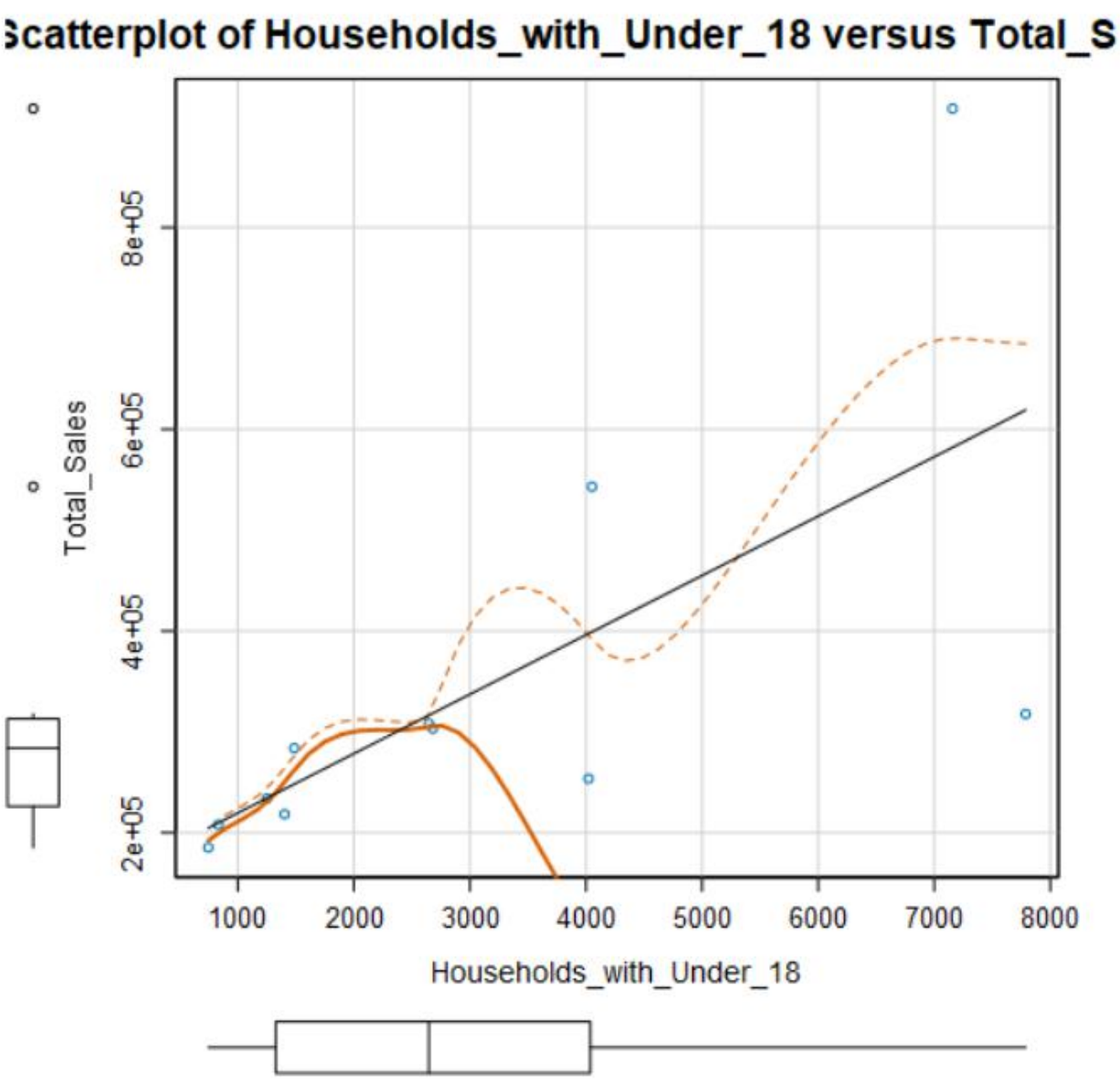
Step 1: Linear Regression

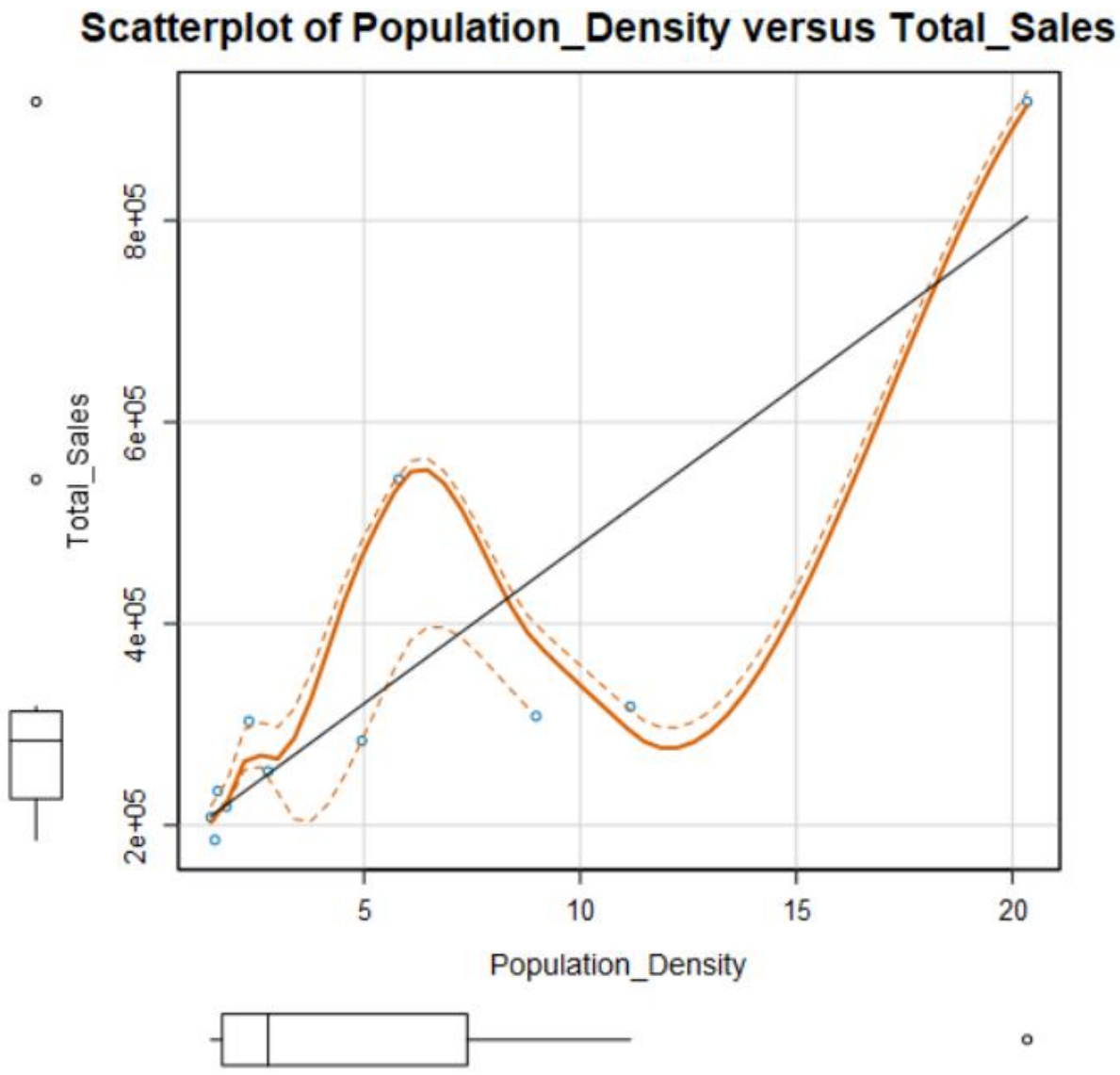
1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

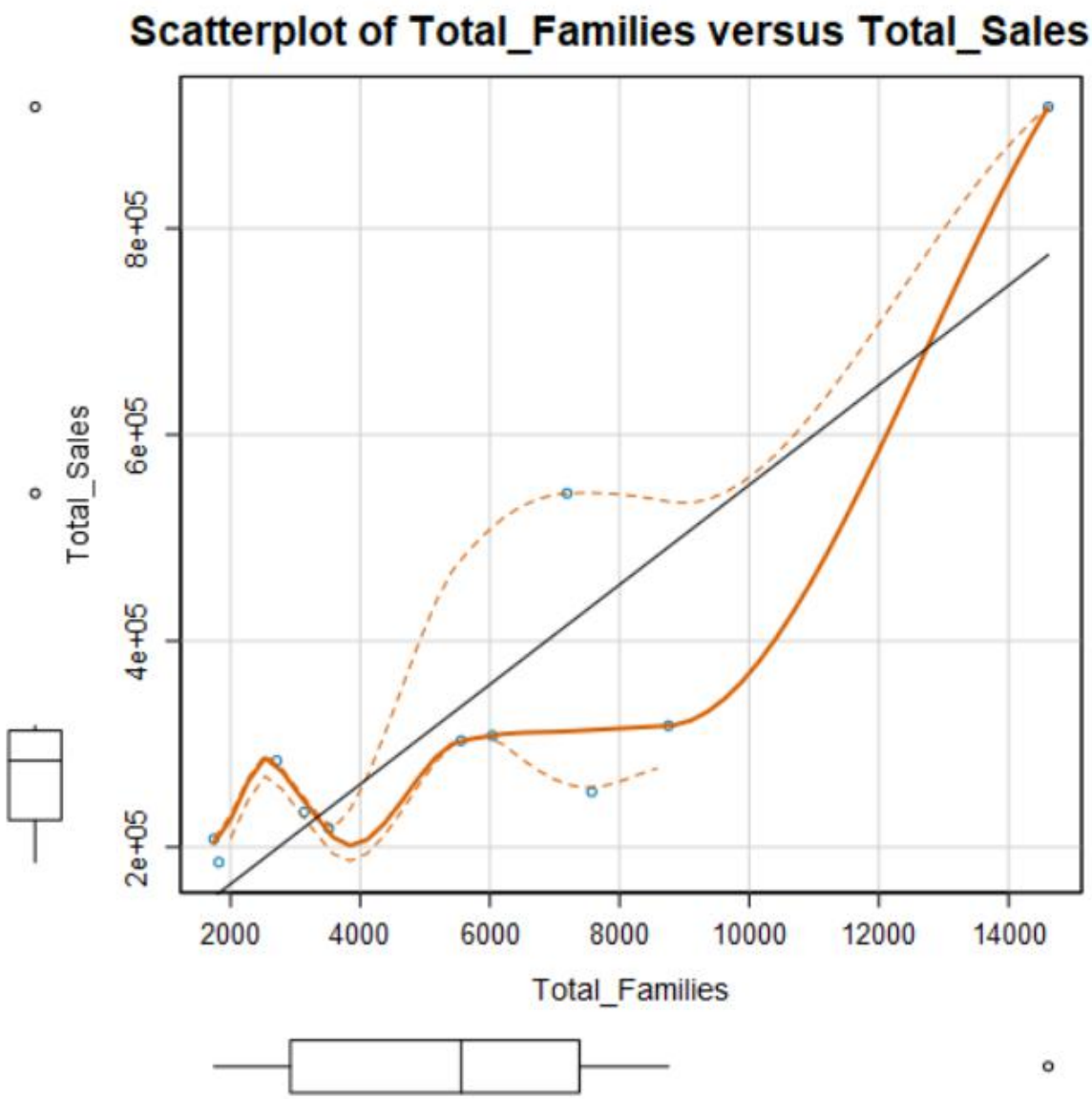
I first plotted each predictor variables against my target variable. I can conclude all predictor variables are good potential predictor variables since they show linearity with sales.




Scatterplot of Land_Area versus Total_Sales
Total_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
Land_Area


Scatterplot of Households_with_Under_18 versus Total_S
Total_Sales
2e+05
4e+05
6e+05
8e+05
1000
2000
3000
4000
5000
6000
7000
8000
Households_with_Under_18


Scatterplot of Population_Density versus Total_Sales
Total_Sales
2e+05
4e+05
6e+05
8e+05
5
10
15
20
Population_Density



Secondly, i checked for correlations between the predictor variables to see if there is any possiblity of multicolinearity in my dataset. Below is a table that shows the correlation between the different predictor variables:

Pearson Correlation Analysis

Focused Analysis on Field Total.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	

Full Correlation Matrix

	Total.Sales	X2010.Census	Land.Area	Households.with.Under.18	Population.Density	Total.Families
Total.Sales	1.000000	0.898099	-0.288898	0.676012	0.862894	0.864660
X2010.Census	0.898099	1.000000	-0.061587	0.911883	0.927702	0.968005
Land.Area	-0.288898	-0.061587	1.000000	0.180704	-0.317244	0.099389
Households.with.Under.18	0.676012	0.911883	0.180704	1.000000	0.815756	0.907242
Population.Density	0.862894	0.927702	-0.317244	0.815756	1.000000	0.884792
Total.Families	0.864660	0.968005	0.099389	0.907242	0.884792	1.000000

Matrix of Corresponding p-values

	Total.Sales	X2010.Census	Land.Area	Households.with.Under.18	Population.Density	Total.Families
Total.Sales		1.7363e-04	3.8890e-01	2.2398e-02	6.2613e-04	5.9221e-04
X2010.Census	1.7363e-04		8.5725e-01	9.2143e-05	3.8717e-05	1.0478e-06
Land.Area	3.8890e-01	8.5725e-01		5.9492e-01	3.4180e-01	7.7125e-01
Households.with.Under.18	2.2398e-02	9.2143e-05	5.9492e-01		2.2030e-03	1.1529e-04
Population.Density	6.2613e-04	3.8717e-05	3.4180e-01	2.2030e-03		2.9571e-04
Total.Families	5.9221e-04	1.0478e-06	7.7125e-01	1.1529e-04	2.9571e-04	

We can see that Census, Families, Population Density have strong correlations with each other. Land area however, is not as highly correlated. So i started by using Land Area as one predictor and then tested the four other variables that are correlated. I've found out that using Land Area and Total Families as the predictor variables produced the best model.

```
lm(formula = Total.Sales ~ Land.Area + Total.Families, data =
the.data)
```

Residuals:

Min	1Q	Median	3Q	Max
-121261	-4453	8418	40491	75205

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	197330.41	56449.000	3.496	0.01005	*
Land.Area	-48.42	14.184	-3.414	0.01123	*
Total.Families	49.14	6.055	8.115	8e-05	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72030 on 7 degrees of freedom
Multiple R-squared: 0.9118, Adjusted R-Squared: 0.8866
F-statistic: 36.2 on 2 and 7 degrees of freedom (DF), p-value 0.0002035

Type II ANOVA Analysis

Response: Total.Sales

	Sum Sq	DF	F value	Pr(>F)	
Land.Area	60473052720.43	1	11.66	0.01123	*
Total.Families	341673845917.83	1	65.85	8e-05	***
Residuals	36318449406.44	7			

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. . For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

The p-values for land area and total families are both below 0.05 and the Multiple R-squared value is at .91 which is close to 1. This is model is a decent model.

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

Y = 197,330 – 48.42 * [Land Area] + 49.14 * [Total Families]

Step 2: Analysis

1. **Which city would you recommend and why did you recommend this city?**

I started with the Web Scraped Data from the Wyoming Wikipedia page, and used text to

columns and select tools and the Data Cleansing to parse out the City, County, 2010 Census, and 2014 Estimate and remove all of the extra punctuation.

For the demographic data, I used the Auto-field tool to combine all of the numbers labeled as String fields.

Before each join, I summarized the amounts by city to ensure that there were no duplicate city names within the data.

For Pawdacity sales file, I transposed the data to get City, Month, and Amount, and then summarized by City to get the total amount for each city.

From there, I created my data set used to train my regression model.

Once the model was created, I applied the model to the cities that were not already in the Pawdacity Sales file by taking the left output from the join on the Pawdacity sales file.

I took the competitor data with an autofield tool and joined it, with a formula off of the left join to create a 0 in the Competitor Amount so I could union the cities that have no competitor back into the overall dataset. I don't want to exclude cities where no competitors are present.

I then applied the filters laid out in the project plan to come up with my list of possible cities, and sorted on the expected revenue to bring the best choice to the top.

2. **What were the sales prediction steps did you do?**

I filtered my cities according to the given the criteria in the project and calculated revenue off the population density information using my linear model.

3. **Which city would you recommend and why did you recommend this city?**

I would recommend the city of Laramie with a predicted sales of $305,014