Project 2: Data Cleanup And Building Linear Regression model

Analyst: Shahrooz Govahi

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?
   Answer: A pet store chain in Wyoming has 13 stores in that state. They are planning to open a new store in Wyoming. Based on yearly sales of current stores they want to choose the best city for the 14th branch and it should be the most profitable store between available option.

2. What data is needed to inform those decisions?
   Answer: We need current stores locations and yearly sales amount for each of them. More over for predicting a new location sales we need some other predictors that will describe potential customers of a city. It means demographic data for all cities in the state. Variables like total population, number of households with young children and land area. On the other hand we need to know how many pet stores as competitors are in each city and how much yearly sales they acquire. Based on these information we can build a linear regression to predict expected sales amount for potential cities and after that we can choose the city with the most predicted yearly sales amount for the new store.

Step 2: Building the Training Set

Build your training set given the data provided to you. Your column sums of your dataset should match the sums in the table below.

In addition provide the averages on your data set here to help reviewers check your work. You should round up to two decimal places, ex: 1.24

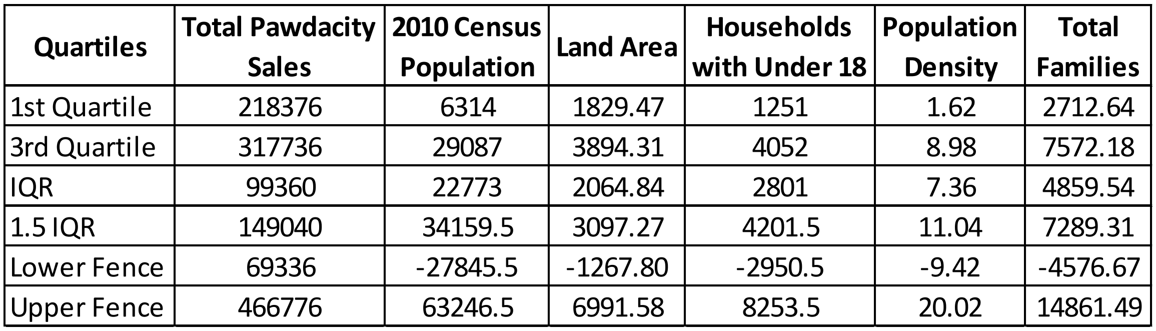
Column	**Sum**	**Average**
Census Population	*213,862*	*19,442*
Total Pawdacity Sales	*3,773,304*	*343,027.64*
Households with Under 18	*34,064*	*3,096.73*
Land Area	*33,071*	*3,006.45*
Population Density	*63*	*5.73*
Total Families	*62,653*	*5,695.73*

Step 3: Dealing with Outliers

Answer these questions

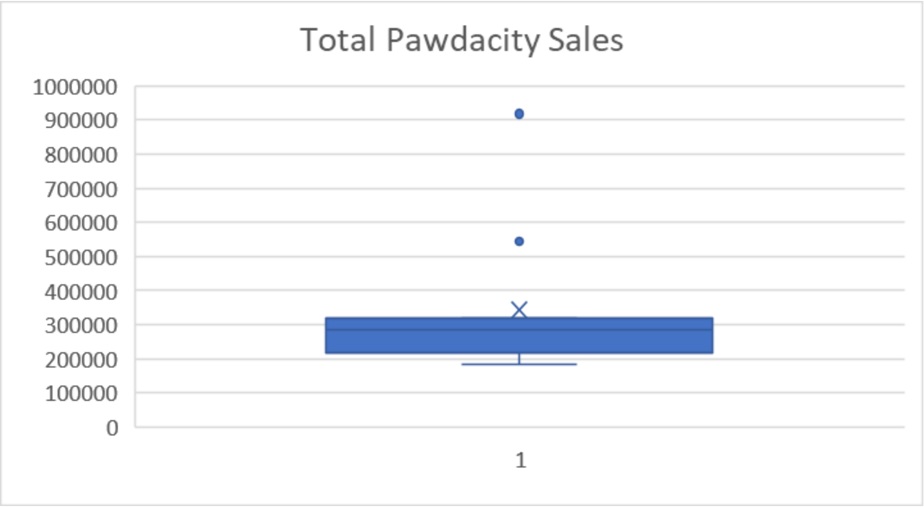
Are there any cities that are outliers in the training set? Which outlier have you chosen to remove or impute? Because this dataset is a small data set (11 cities), **you should only remove or impute one outlier**. Please explain your reasoning.

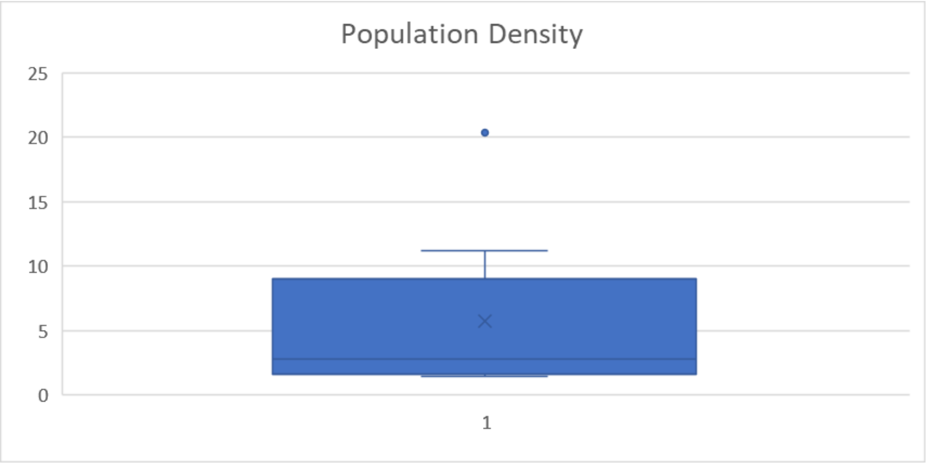
Answer: I created a clean data set with 6 variables and 11 cities. Based on IQR calculations for each variable in the dataset I concluded that there are 2 potentially outliers between these 11 cities. Here is the IQR details:



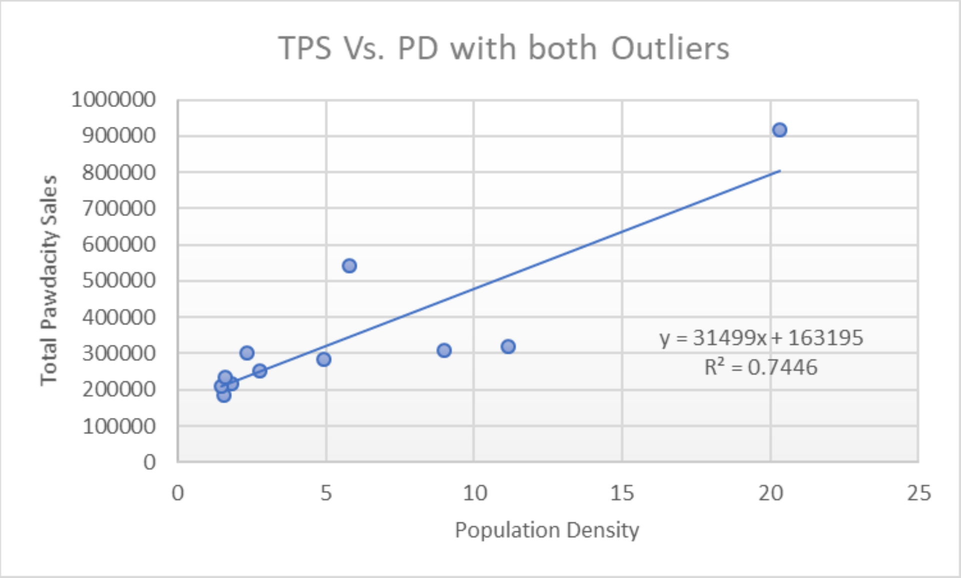
Quartiles	Total Pawdacity Sales	2010 Census Population	Land Area	Households with Under 18	Population Density	Total Families
1st Quartile	218376	6314	1829.47	1251	1.62	2712.64
3rd Quartile	317736	29087	3894.31	4052	8.98	7572.18
IQR	99360	22773	2064.84	2801	7.36	4859.54
1.5 IQR	149040	34159.5	3097.27	4201.5	11.04	7289.31
Lower Fence	69336	-27845.5	-1267.80	-2950.5	-9.42	-4576.67
Upper Fence	466776	63246.5	6991.58	8253.5	20.02	14861.49

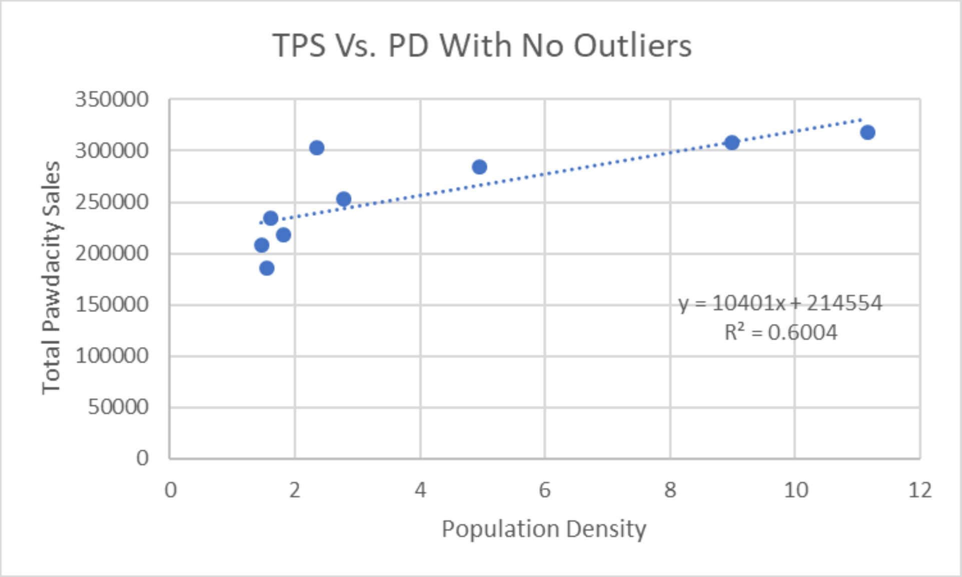
And here are Box and Whisker plots for two variables that has outliers:

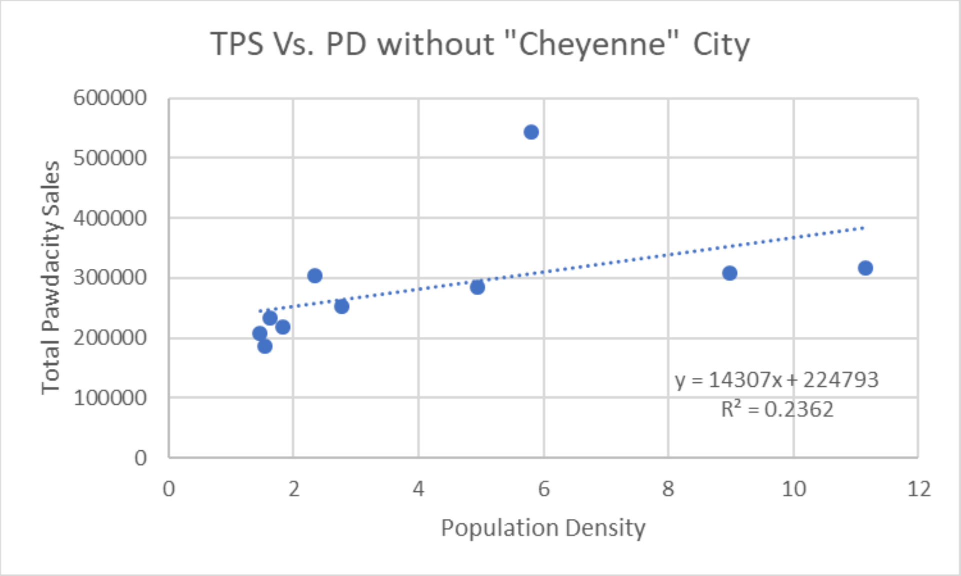


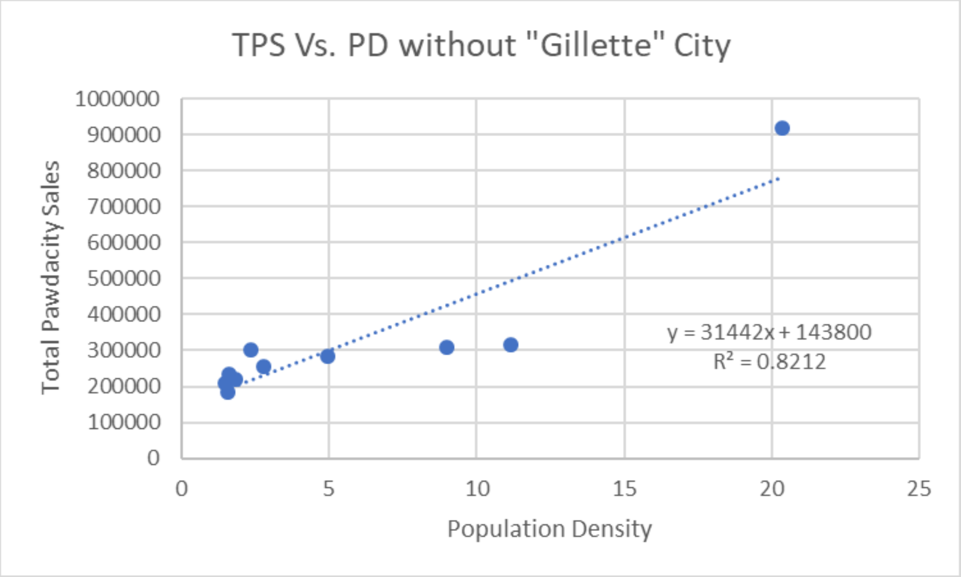


Based on these analysis I understand there are 2 cities that play as outliers in our data set: “Cheyenne” and “Gillette”. Now I draw scatterplots for different conditions. First I draw it with both outliers and without none of them and next in each chart I delete one of them and look at the slope of trend line and R-Squared:









Based on these scatter plots, I think the last choice is the best one. Because more than 82% of the variance in the target variable(TPS) is explained with the linear model. “Cheyenne” is the most populated and the most dense city in the dataset and it make sense it has the most yearly sales amount. But “Gillette” population density and yearly sales amount are not both in the outlier part. I mean the sales amount for this city is an outlier, but the population density is less than 3rd quartile. This city will create problem for our model and will decrease the power of “population density” as a predictor.

So I decide to delete the “Gillette” city from the data set and keep the “Cheyenne” in it. Now we have 10 cities with 6 variables in the final version of our cleaned data set.

Linear Regression

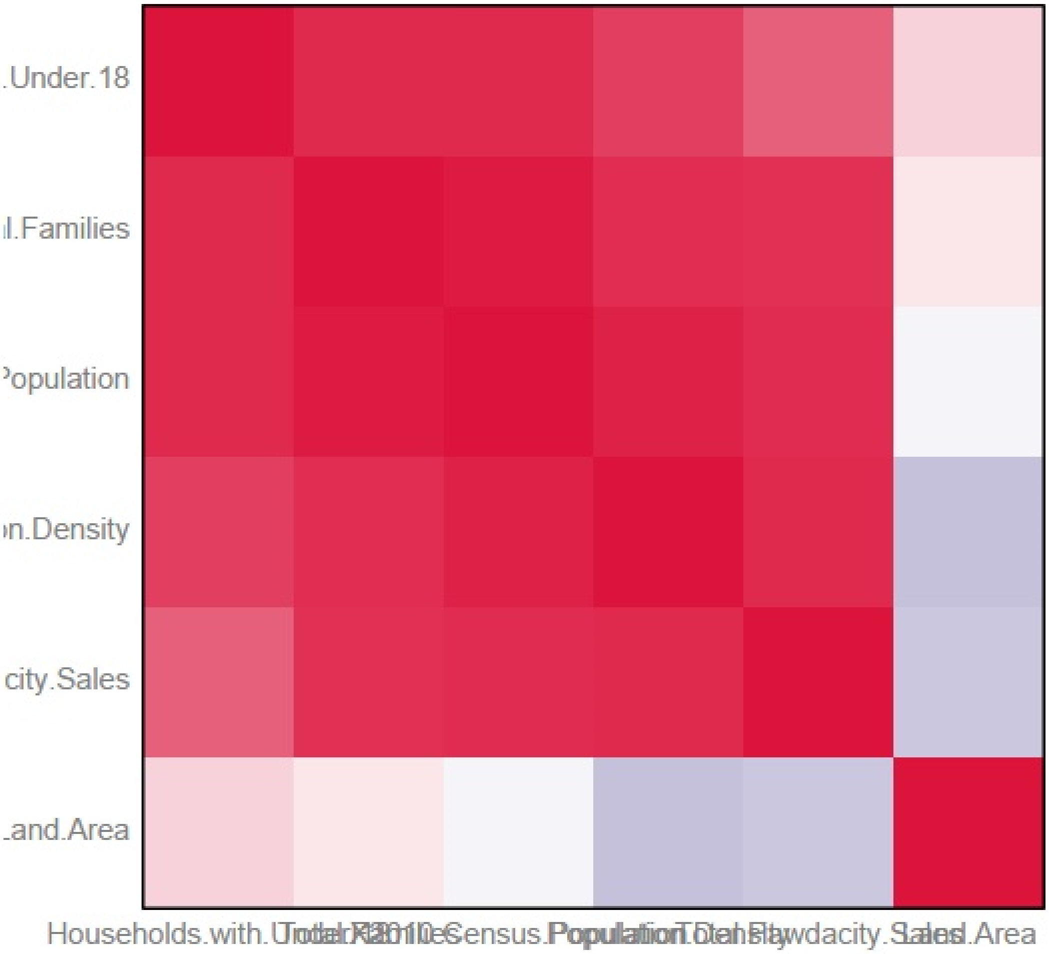
Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section.

Answer: Now that we have the data we were looking for, it’s time for consuming the data and creating a linear model for predicting sales in the cities that Pawdacity has no branch and choose the best city for opening the 14th store.

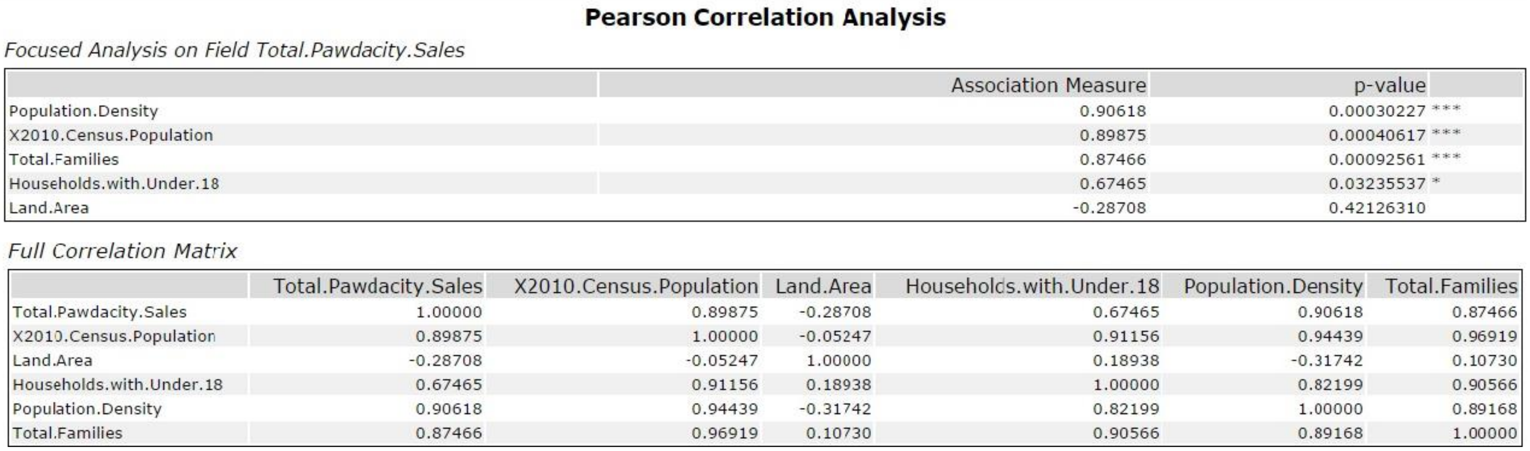
Out Target variable is “Total Pawdacity Sales” and we have these potential predictors:

- 2010 Census Population
- Land area
- Households with under 18
- Population Density
- Total Families

First I will run a correlation analysis between predictors:



These red regions means that there are strong correlation between 4 variables.



Pearson Correlation Analysis

Focused Analysis on Field Total.Pawdacity.Sales

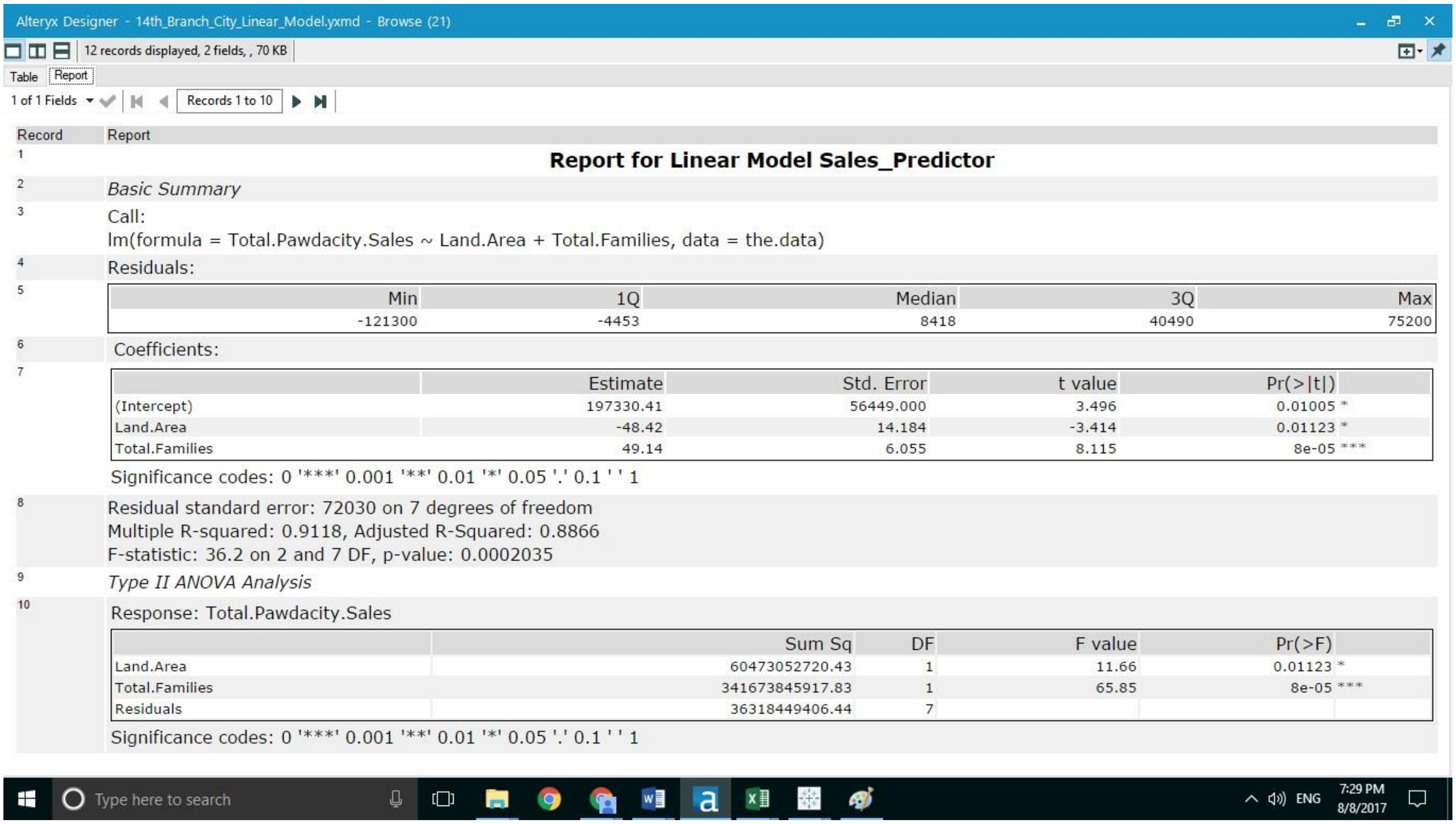
	Association Measure	p-value
Population.Density	0.90618	0.00030227 ***
X2010.Census.Population	0.89875	0.00040617 ***
Total.Families	0.87466	0.00092561 ***
Households.with.Under.18	0.67465	0.03235537 *
Land.Area	-0.28708	0.42126310

Full Correlation Matrix

	Total.Pawdacity.Sales	X2010.Census.Population	Land.Area	Households.with.Under.18	Population.Density	Total.Families
Total.Pawdacity.Sales	1.00000	0.89875	-0.28708	0.67465	0.90618	0.87466
X2010.Census.Population	0.89875	1.00000	-0.05247	0.91156	0.94439	0.96919
Land.Area	-0.28708	-0.05247	1.00000	0.18938	-0.31742	0.10730
Households.with.Under.18	0.67465	0.91156	0.18938	1.00000	0.82199	0.90566
Population.Density	0.90618	0.94439	-0.31742	0.82199	1.00000	0.89168
Total.Families	0.87466	0.96919	0.10730	0.90566	0.89168	1.00000

It's clear that 4 variables has a strong correlation with each other and somehow they will create duplicate predictor variable problem. "Land Area" is the variable with the weakest correlation with others. So I choose that one as predictor and with trial and error I will choose the other variable between those 4 variables. The best combination will be the model with maximum R-Square. This best model contains "Land Area" and "Total Families" as predictors; The R-square in this model is 0.91 which is close to 1 and the linear model formula is:

Total Sales = 197330.41 – 48.42 * [Land Area] + 49.14 * [Total Families];


Alteryx Designer - 14th_Branch_City_Linear_Model.yxmd - Browse (21)
12 records displayed, 2 fields, , 70 KB
Table
Report
1 of 1 Fields
Records 1 to 10
Record
Report
Type here to search
ENG
7:29 PM
8/8/2017

1 **Report for Linear Model Sales_Predictor**

2 *Basic Summary*

3 Call:
lm(formula = Total.Pawdacity.Sales ~ Land.Area + Total.Families, data = the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Max
-121300	-4453	8418	40490	75200

6 Coefficients:

7

	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

8 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 DF, p-value: 0.0002035

9 *Type II ANOVA Analysis*

10 Response: Total.Pawdacity.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

The p-values for intercept and both variables are less than 0.05, so it is an acceptable model.

We want to choose a city with these criteria:

- The new store should be located in a new city. That means there should be no existing stores in the new city.
- The total sales for the entire competition in the new city should be less than $500,000
- 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).
- The predicted yearly sales must be over $200,000.
- The city chosen has the highest predicted sales from the predicted set.

I ran the model with respect to these guide lines and put appropriate filters to get the list of cities that meet these criteria. The 1st rank city in the result is: Laramie and the predicted sales is : **$305,014**;

My suggestion to the manager is opening the new store in **Laramie;**