# Project 2.2: Recommend a City

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/3d606c26-cb8e-43af-9199-7e3577aa3392/project#

**Note that this project is a continuation from Project 2.1: Data Cleanup. You must meet specifications for Project 2.1 before you can continue on with this Project 2.2**

## Step 1: Linear Regression

*Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section. (750 word limit)*

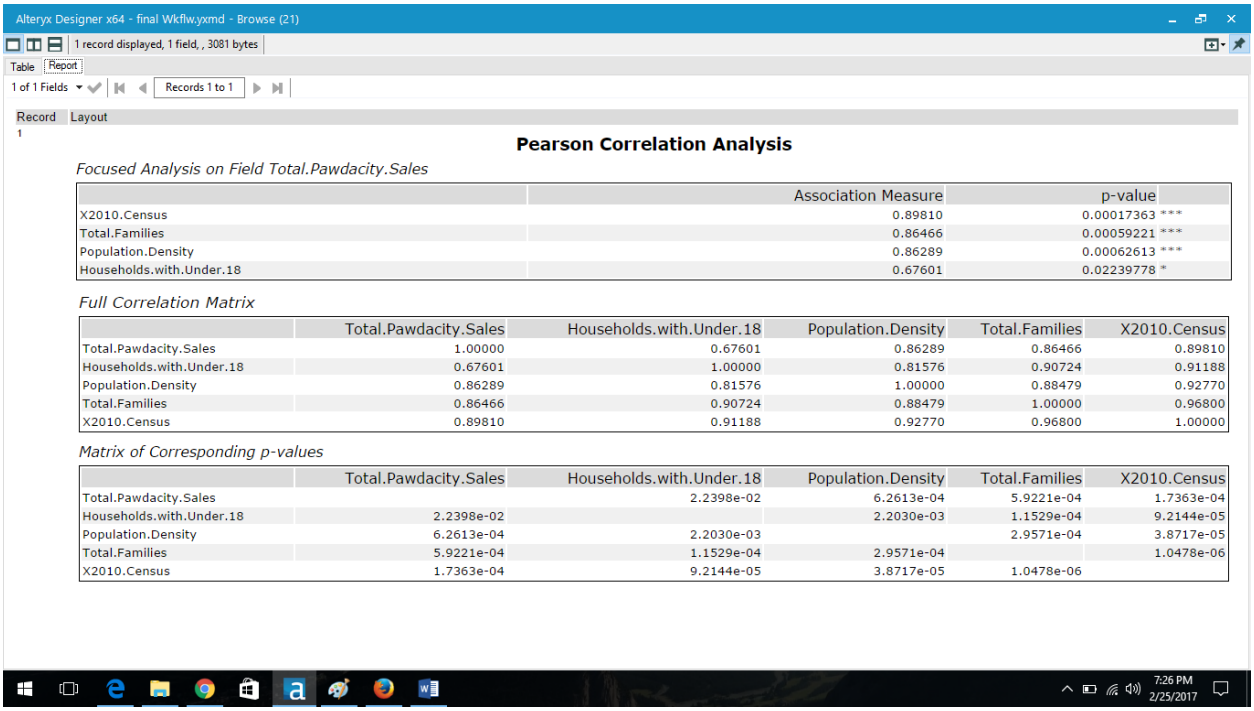
***Important:** Make sure you have dealt with outliers and removed one city from your training set. You should have **10 rows** of data before you begin modeling the dataset.*

*Build a linear regression model to help you predict total sales.*

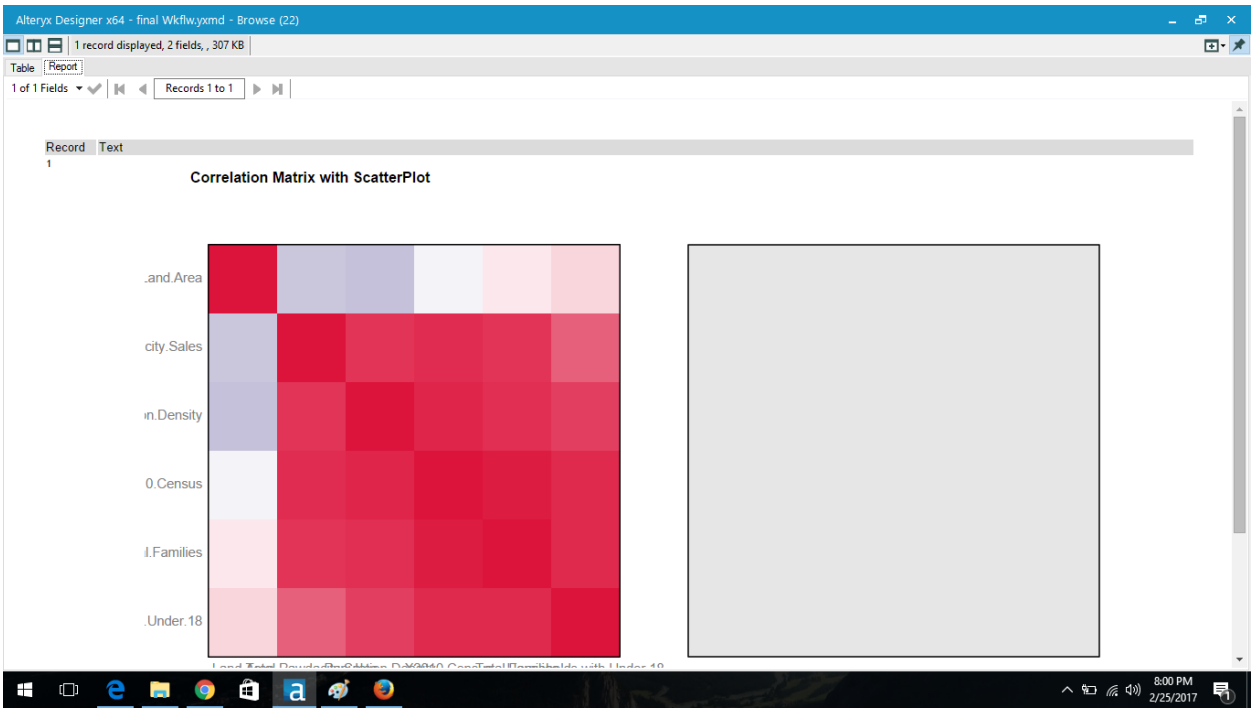
*At the minimum, answer these questions:*

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.

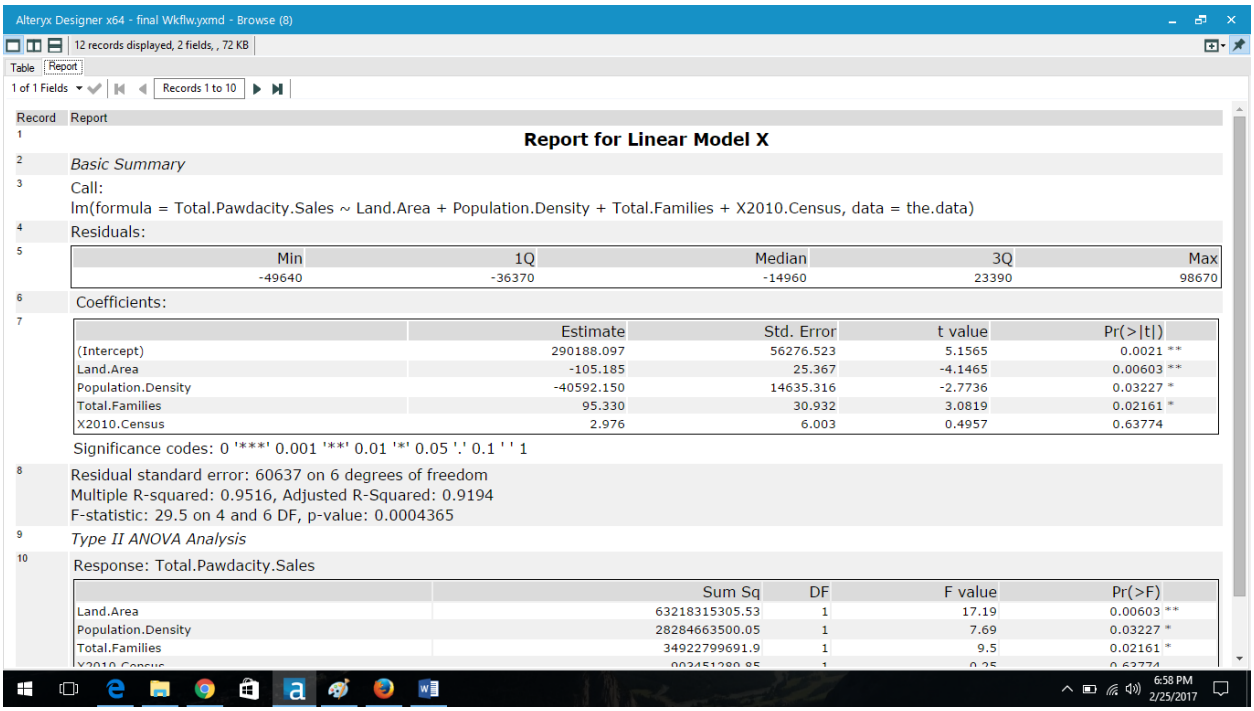
Answer : I ran the Association analysis tool and opened the report option of the association analysis tool to find significance of variables to “Total Pawdacity Sales”.



I then tried finding similar predictor variables if any



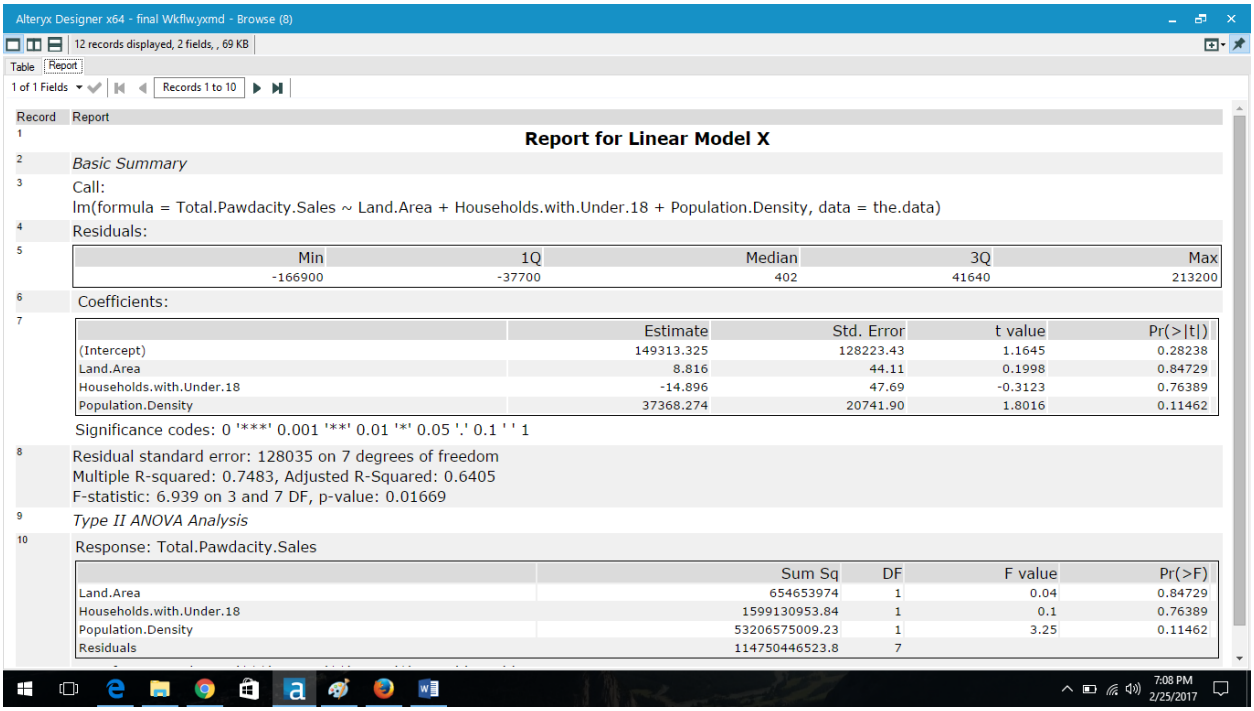

As you can see all variables are very highly co-related( except Land area), thus there may be a presence of similar variables which would need to be excluded. Before start analyzing duplicate variables, I ran the linear regression tool, to check for any un-significant variable



From the above report , we can safely remove 2010 census from further analysis as it is not significant because its p-value is above 0.05

Now to check for the best model, factoring the high co-relation between some variables

MODEL 1 : without Total Families



Alteryx Designer x64 - final Wkflw.yxmd - Browse (8)

12 records displayed, 2 fields, , 69 KB

Table | Report

1 of 1 Fields | Records 1 to 10

Record | Report

**Report for Linear Model X**

*Basic Summary*

Call:
lm(formula = Total.Pawdacity.Sales ~ Land.Area + Households.with.Under.18 + Population.Density, data = the.data)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -166900 | -37700 | 402 | 41640 | 213200 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 149313.325 | 128223.43 | 1.1645 | 0.28238 |
| Land.Area | 8.816 | 44.11 | 0.1998 | 0.84729 |
| Households.with.Under.18 | -14.896 | 47.69 | -0.3123 | 0.76389 |
| Population.Density | 37368.274 | 20741.90 | 1.8016 | 0.11462 |

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

Residual standard error: 128035 on 7 degrees of freedom
Multiple R-squared: 0.7483, Adjusted R-Squared: 0.6405
F-statistic: 6.939 on 3 and 7 DF, p-value: 0.01669

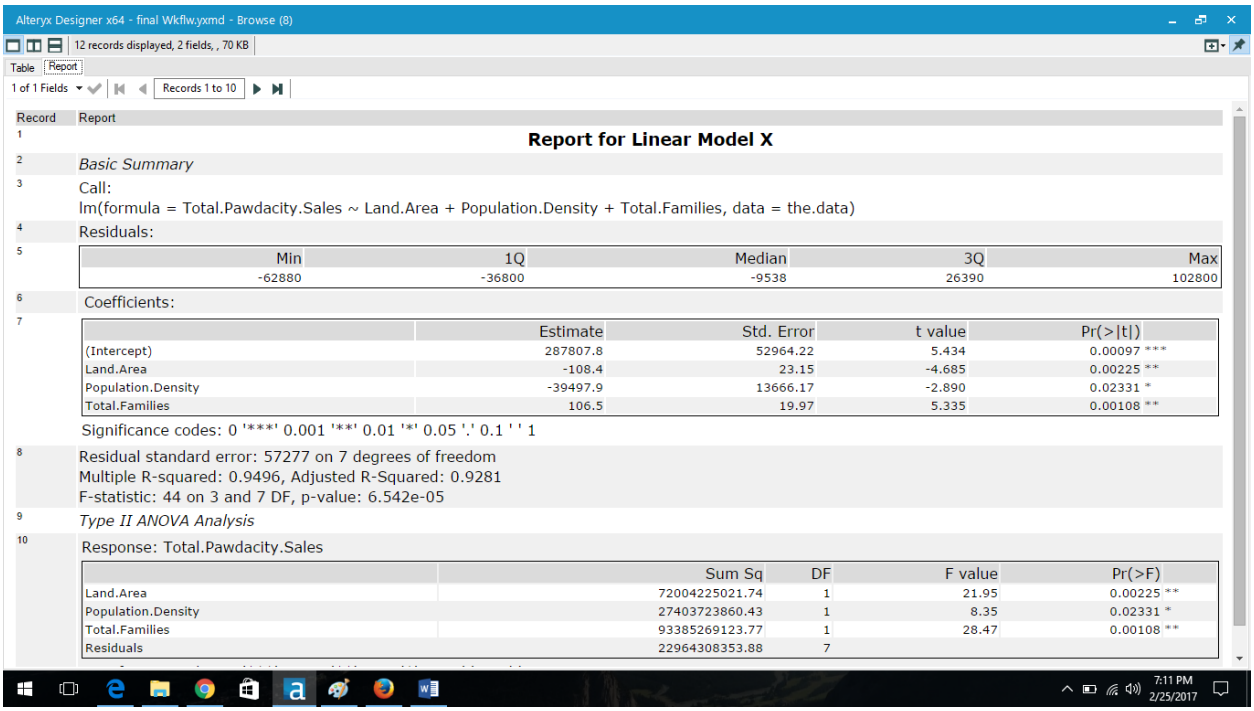
*Type II ANOVA Analysis*

Response: Total.Pawdacity.Sales

| | Sum Sq | DF | F value | Pr(>F) |
|---|---|---|---|---|
| Land.Area | 654653974 | 1 | 0.04 | 0.84729 |
| Households.with.Under.18 | 1599130953.84 | 1 | 0.1 | 0.76389 |
| Population.Density | 53206575009.23 | 1 | 3.25 | 0.11462 |
| Residuals | 114750446523.8 | 7 | | |

7:08 PM 2/25/2017

MODEL 2 : without Households under 18



Alteryx Designer x64 - final Wkflw.yxmd - Browse (8)

12 records displayed, 2 fields, , 70 KB

Table | Report

1 of 1 Fields | Records 1 to 10

Record | Report

**Report for Linear Model X**

*Basic Summary*

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

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -62880 | -36800 | -9538 | 26390 | 102800 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 287807.8 | 52964.22 | 5.434 | 0.00097 *** |
| Land.Area | -108.4 | 23.15 | -4.685 | 0.00225 ** |
| Population.Density | -39497.9 | 13666.17 | -2.890 | 0.02331 * |
| Total.Families | 106.5 | 19.97 | 5.335 | 0.00108 ** |

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

Residual standard error: 57277 on 7 degrees of freedom
Multiple R-squared: 0.9496, Adjusted R-Squared: 0.9281
F-statistic: 44 on 3 and 7 DF, p-value: 6.542e-05

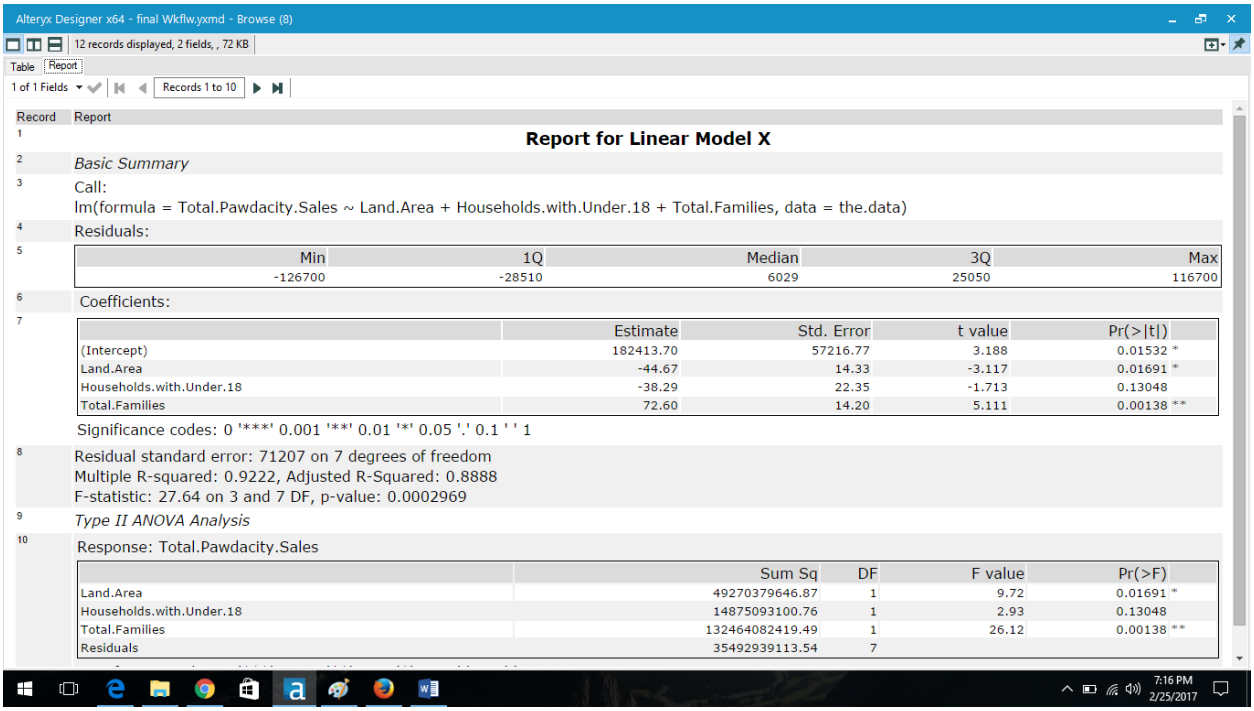
*Type II ANOVA Analysis*

Response: Total.Pawdacity.Sales

| | Sum Sq | DF | F value | Pr(>F) |
|---|---|---|---|---|
| Land.Area | 72004225021.74 | 1 | 21.95 | 0.00225 ** |
| Population.Density | 27403723860.43 | 1 | 8.35 | 0.02331 * |
| Total.Families | 93385269123.77 | 1 | 28.47 | 0.00108 ** |
| Residuals | 22964308353.88 | 7 | | |

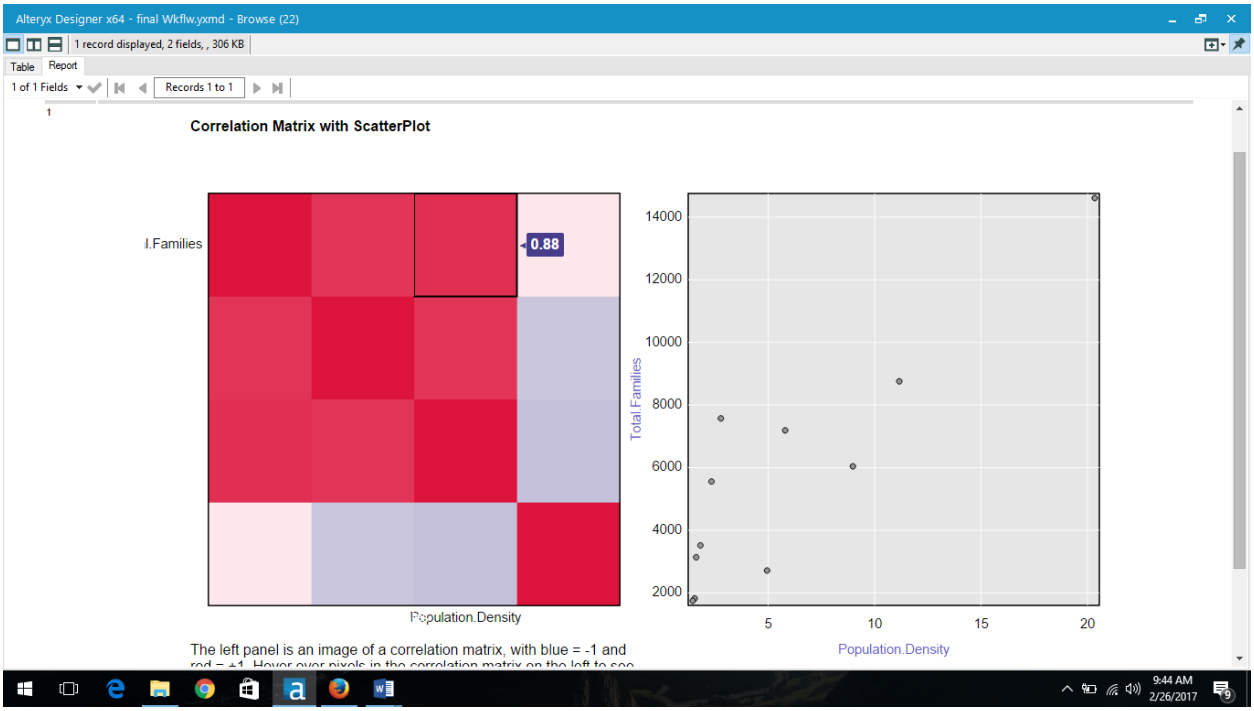
7:11 PM 2/25/2017

Model 3: Without Population density



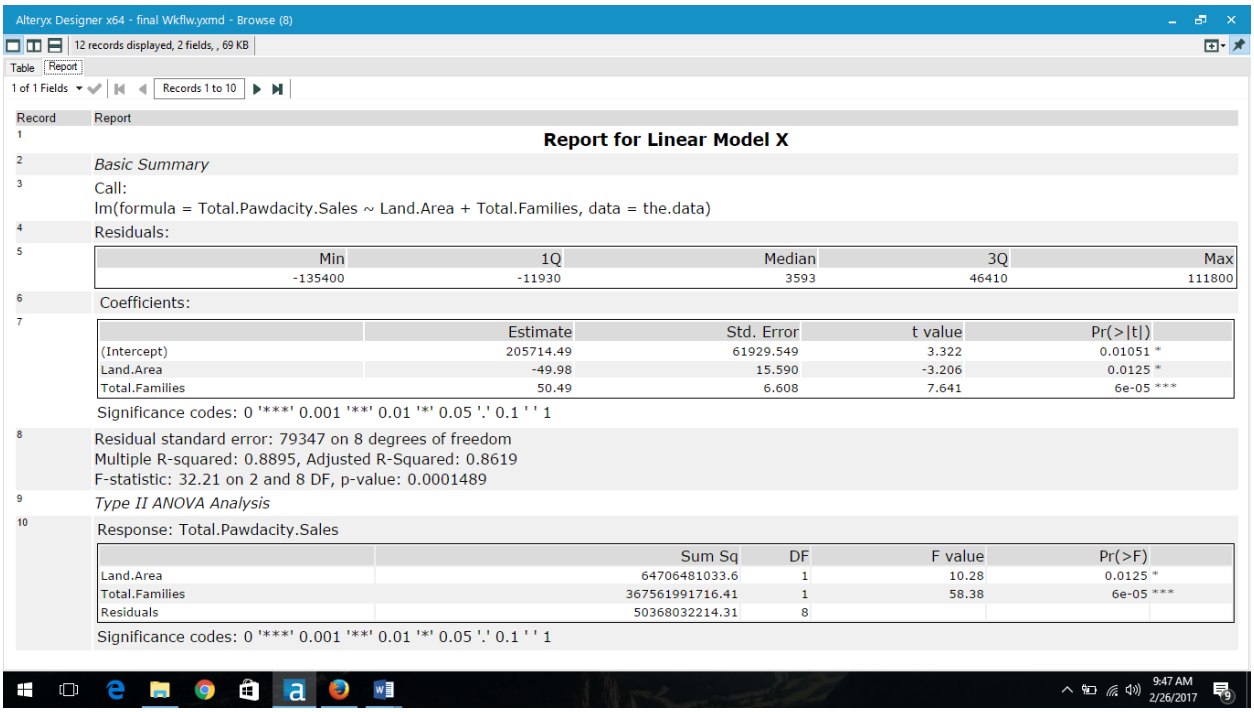

We conclude from the above the 3 models that MODEL 2 is the best fit (without Households under 18) as it has the highest Adjusted R-squared value and the lowest p-values amongst all the list of its variables.

Before we proceed with finalizing Model 2 variables, let us check for duplicate variables again with the association analysis tool:

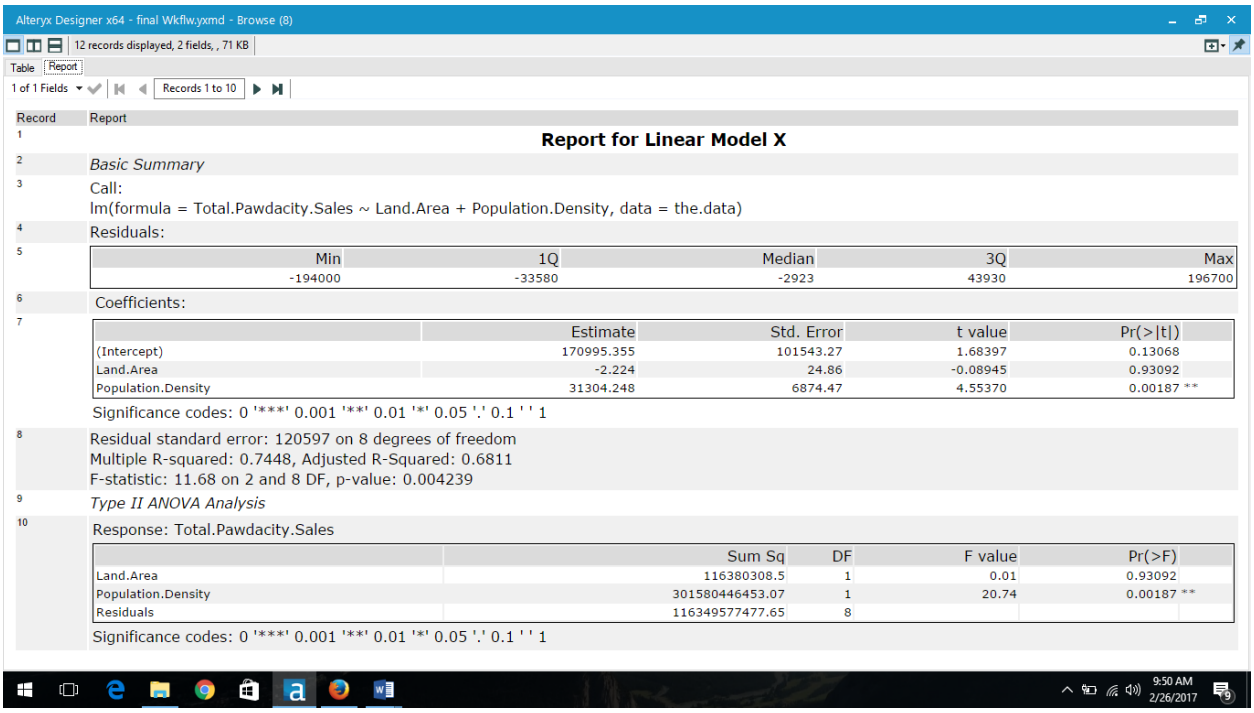



We can see the predictor variables are still highly co-related. Let's see which variable best fits our model and eliminate the duplicate variable.

MODEL 4: Without Population density



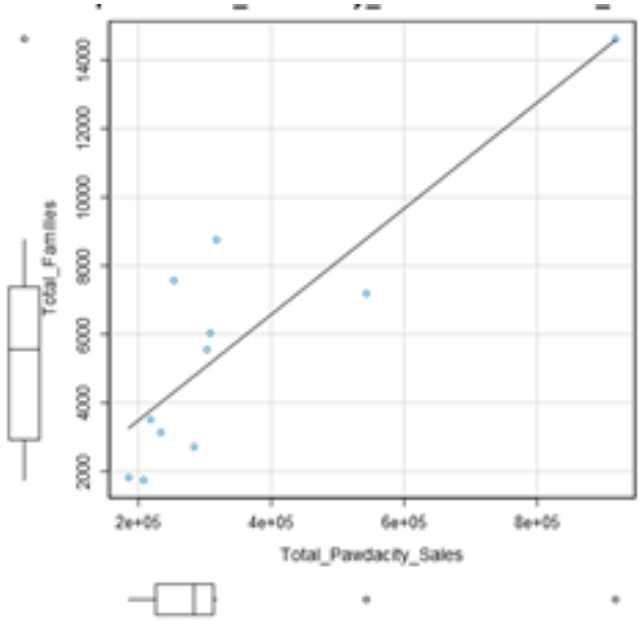
MODEL 5: without Total Families



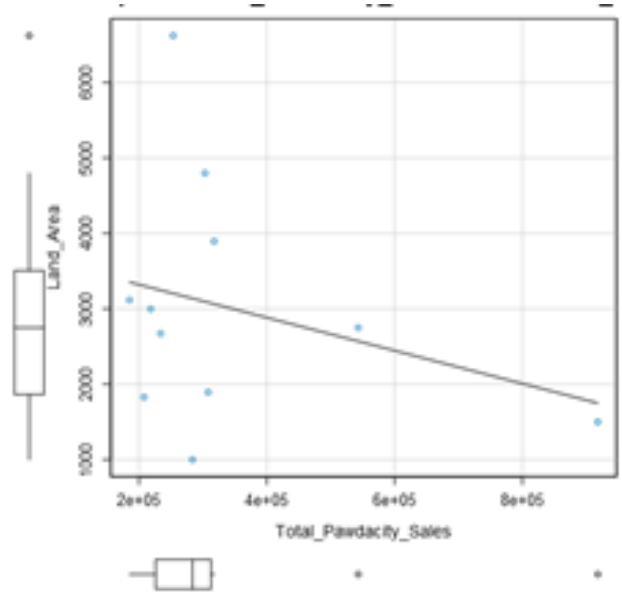

From the above two models, we can see Model 4 best fits our requirement as, the R-squared value is 0.8895 which is significantly higher and the p-values are more significant i.e. the values are less than 0.05.

And also, please find below, the scatterplots that establish linear relationship of target variable and the predictor variable (As per Model 4)

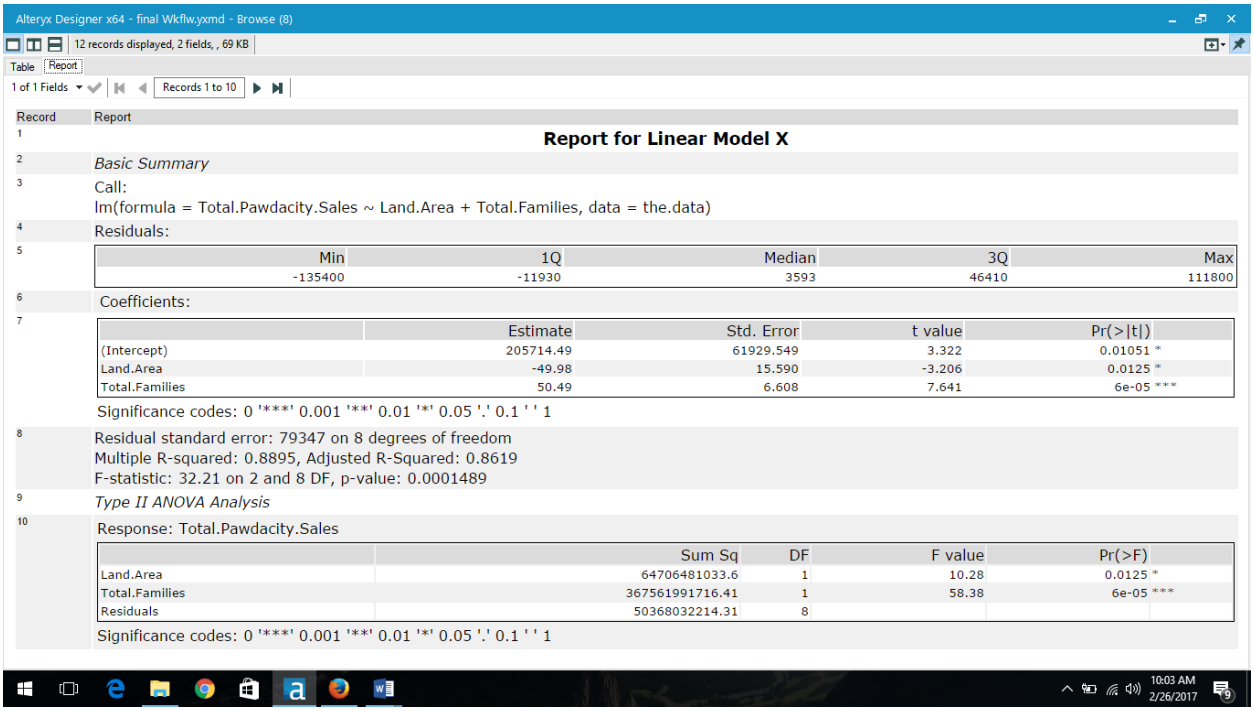
Scatterplot of Total Pawdacity sales and Total Families




Scatterplot of Total Pawdacity sales and Land area




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 linear model is a good model as all the predictor variables are highly significant ( P-values are less than 0.05)
And also, the adjusted R-squared value is 0.8619 which means nearly all variance is covered by the model and is well above the standard of 0.50 used in practice.

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)

*ANSWER : Y = m + b0x0 + b1x1 + b2x2*

*Total pawdacity sales = Intercept + (-49.98*Land area) + (50.49*Total families)*

*Total pawdacity sales =205714.49+ (-49.98*Land area) + (50.49*Total families)*

# Step 2: Analysis

*Use your model results to provide a recommendation. (500 word limit)*

*At the minimum, answer this question:*

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

I would recommend the 14th store to be opened on the city of “Laramie” as according to my calculations, it would generate the highest sales considering the given conditions/criteria of the project

The sales generated is estimated at $315,812.7370

## Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.