MAKE A COPY: Practice Project: Recommend a City

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

## 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)](https://classroom.udacity.com/courses/ud976/lessons/4e33b70a-72a4-47cb-959a-28632ae6aaff/concepts/631d190c-8626-4dd7-92df-f5bd96913c48) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

The next variables have been selected as predictor variables

* Population Density
* 2010 Census Population
* Total Families
* Households with under 18

Because these variables have a p values less than 0.05 and between three and 1 stars (\*\*\*) per the Pearson correlation analysis, which means that are statistically significant to predict Total Sales values with good level of confidence. See table below:

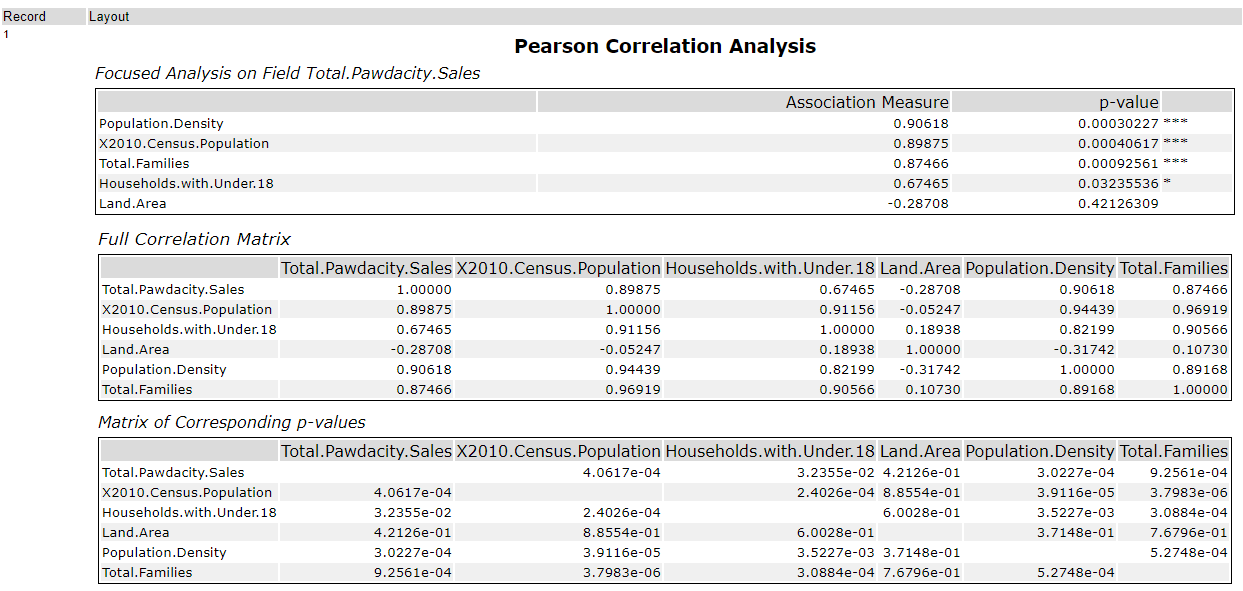


Table Pearson correlation for predictor variables

Also has been generated the scatter plots, to verify there is a linear relationship between the each predictor variables and the target variable, in this case the Total Pawdacity Sales.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Total Pawdacity Sales** | **p-value** | **~~chosen~~** |
| **Census population** |  | 0.0004 | ~~yes~~ |
| ***Households with Under 18*** |  | .03 | ~~yes~~ |
| ***Land Area*** |  | 0.42 | ~~no~~ |
| ***Population Density*** |  | 0.0003 | ~~yes~~ |
| ***Total Families*** |  | 0.0009 | ~~yes~~ |

Observation after review solution: the predictor variables were not analyzed to avoid potential duplicity, mainly the variables about population, are highly correlated each other, which may cause the prediction get skewed.

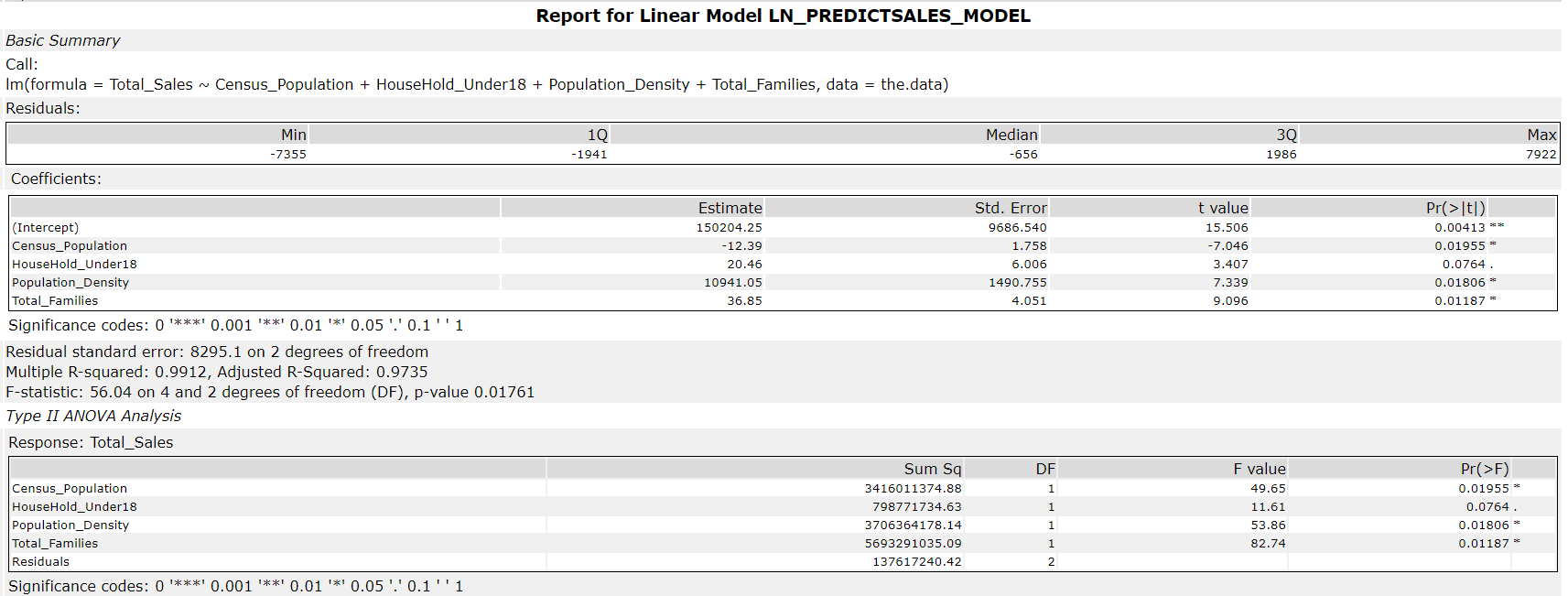
This can be avoided by reviewing correlation matrix and remove those variables highly correlated from the model, also called multicollinearity.

Total Families (0.97), Household Under 18 (0.91), Population Density (0.94) have a high correlation with Census Population.

Therefore, only Land Area and Total Families will be keep to use in the model, because the lower positive value for correlation of 0.11.

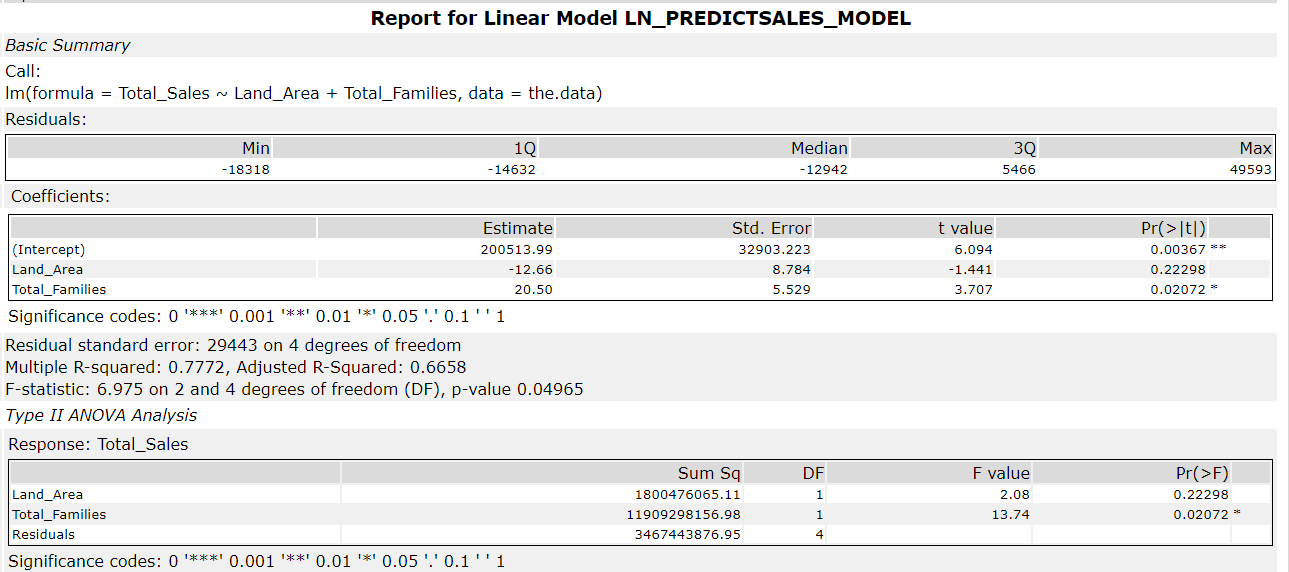
1. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that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

The Adjusted R-Squared value for the Linear Model is 0.97, which mean a good level of confidence to predict the Total Sales. And the P values are statistically significant, because its values are below 0.05.



After review:

The Multiple R-Squared is 0.77, and the Adjusted R-Squared value for the Linear Model is 0.66, which mean a enough level of confidence to predict the Total Sales. And the P values for Total\_Families is statistically significant, because its values are below 0.05. Although Land\_Area is not lower than 0.05 it will be implemented in the model.



1. 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)

Total\_Sales = (-12.39) \* Census\_Population + (20.46) \* HouseHold\_Under18 + (10941.05) \* Population\_Density + ( 36.85) Total\_Families + 150,204.25

After review:

Total\_Sales = -12.66 Land\_Area + ( 36.85) Total\_Families + 200,513.99

## 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?

Lander is the best city for the new store, per the output from the model for sales prediction with higher Predicted Sales of: $256,425.59



After review and choosing the most appropriate predictor variables, the best city for the new store is Laramie with the higher predicted sales of: $264,383. See table below:

