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#](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:*

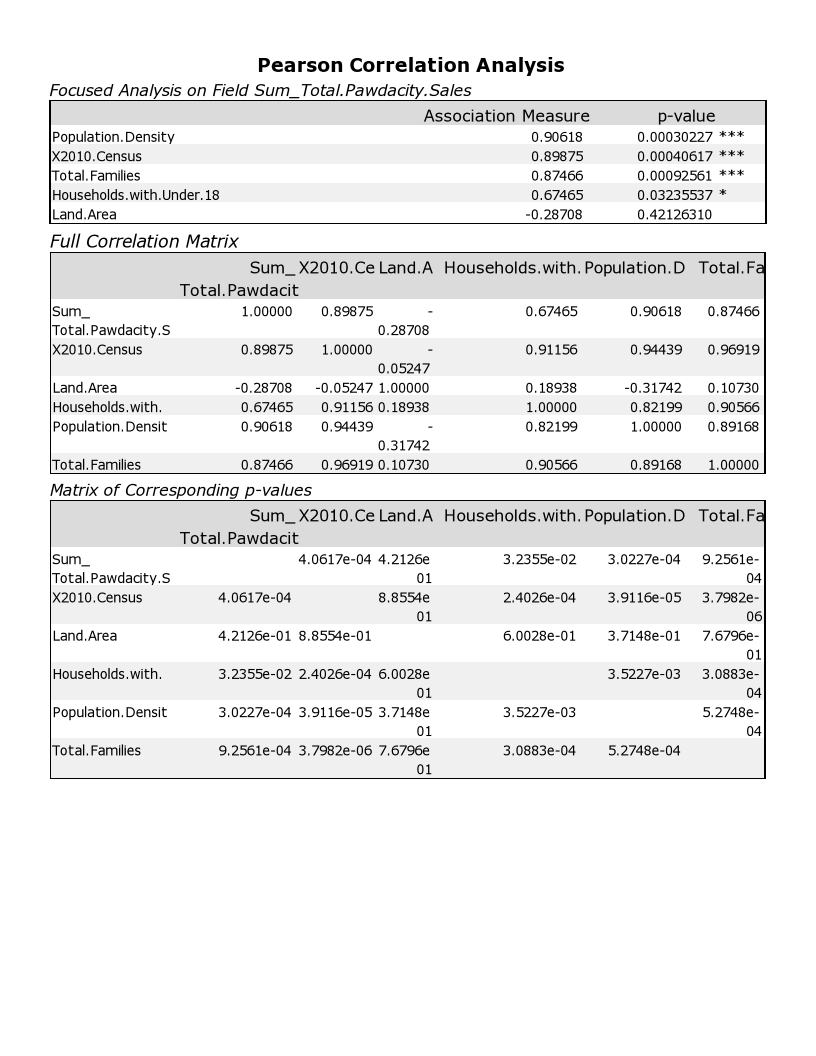
1. How and why did you select the [predictor variables (see supplementary text)](https://classroom.udacity.com/courses/ud976/lessons/4e33b70a-72a4-47cb-959a-28632ae6aaff/concepts/631d190c-8626-4dd7-92df-f5bd96913c48) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

Initially according to the data set we have, there are 5 potential predictor variables to predict sales:

* 1. Census 2010
  2. Land Area
  3. Household with under 18
  4. Population Density
  5. Total Families

We can see immediately that some of these variables are explaining each other, and most probably correlates with each other. For example “Density”, was logically made from land area divided by total population of that area, which can correlates with variables “total families”, “Household under 18” or “Cencus 2010”. I assume I have to select only few of these variables.

First I conduct Association Analysis to find variables to explain “Pawdacity Sales”



From the result above we can see that the significant variable that associated with sales were :

1. Population Density

2. Census 2010

3. Total Families

4. Household under 18

If we take a look on the first three potential predictor variable “Population Density”, “Census 2010” and “Total Family” they are all very similar in regards of counting number of human in an area, that also can explains why 3 of them has pretty similar association result, (0.87, 0.89, 0.90). I personally would take only one predictor variable out of the three to put into the regression model.

While “Population Density” has the strongest association with sales, I will use Census 2010 as a measure so we can compare with Census 2014 estimate as sales predictor variable. Because I will use “Census 2010”, I will drop both “Density” and “Total Families”

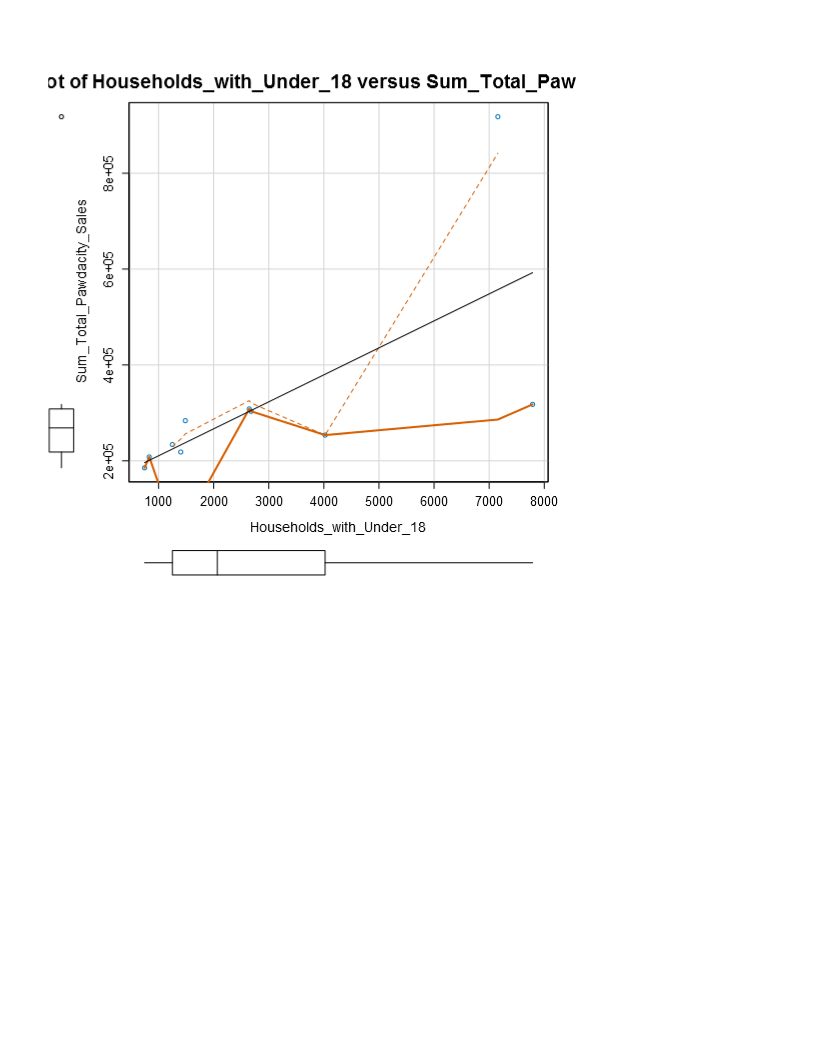
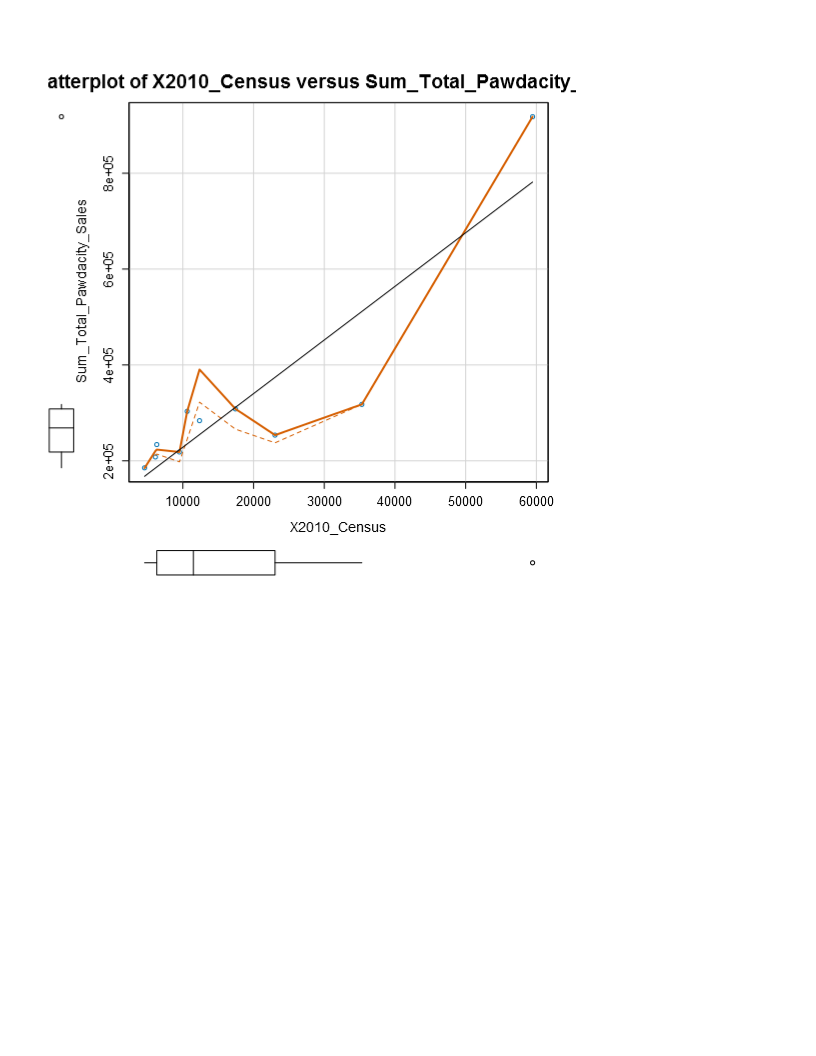
I will still include “Household Under 18” because I assume this variable will be important because it will be a somewhat negative correlation with the model of sum of individual in the city. “ the more household under 18 in a city, means more people are not potential consumer, means less predicted sales”

In this case I will only take 2 variable as predictor:

1st “Census 2010”

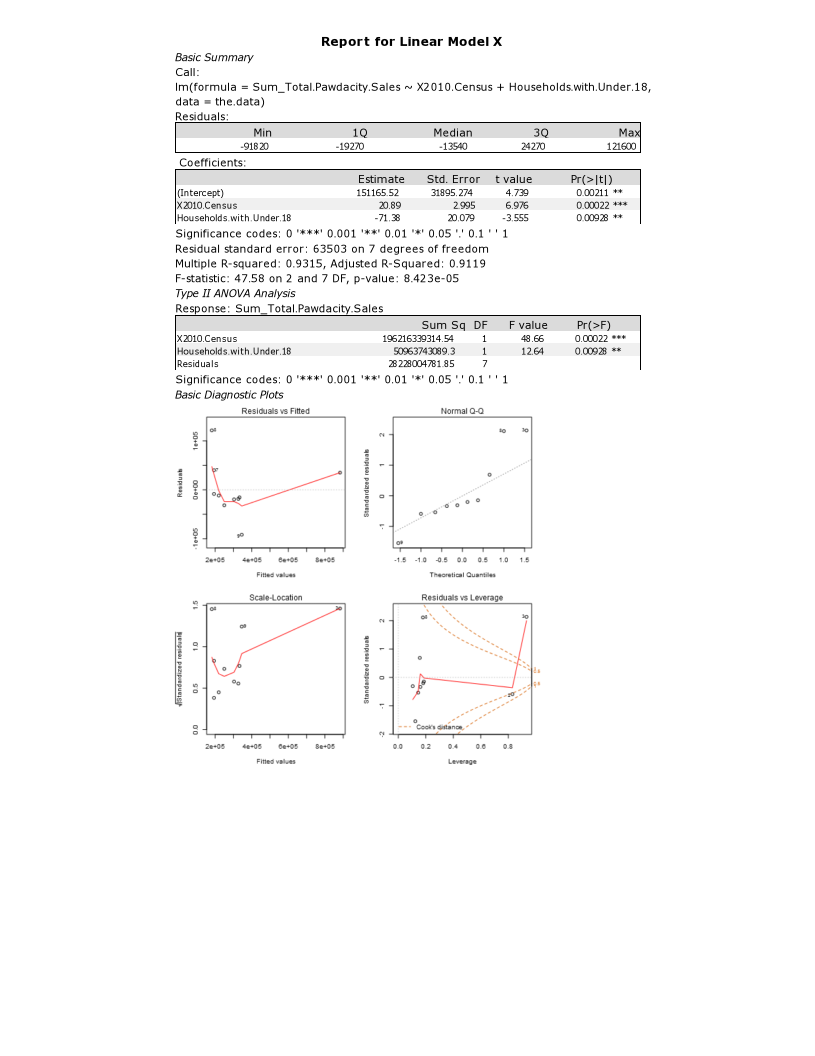
And 2nd “Household Under 18”

We can see both give correlation , where as “Household Under 18” give more moderate correlation, not as steep as”Census 2010”



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.

After I conduct regression of the two predictor variable against pawdacity sales



We can see the model was quite robust with R-squared 0.9315 (much more the minimum standard 0.70), and each of the predictor variable are statistically significant with p-value Census 2010 = 0.00022 ,and Household under 18 = 0.00928

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)

*Predicted Sales = 151165.52 + 20.89 (Census 2010) –* 71.38(Number of Household Under 18)

## 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?

Based on the required criteria :

* City Should be in new city (no pawdacity stores exist)
* Competition sales < $ 500,000.00
* 2014 US Census estimate > 4,000
* Predicted Sales > 200,000.00
* Highest Predicted sales from predicted set

The recommendation goes to Worland , with

Predicted sales at $ 227,987.35

Competition sales at $ 169,000.00

2014 Census estimate 5,366

And the pawdacity is non-existent in that city at the moment

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/381/view) here. Reviewers will use this rubric to grade your project.