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:

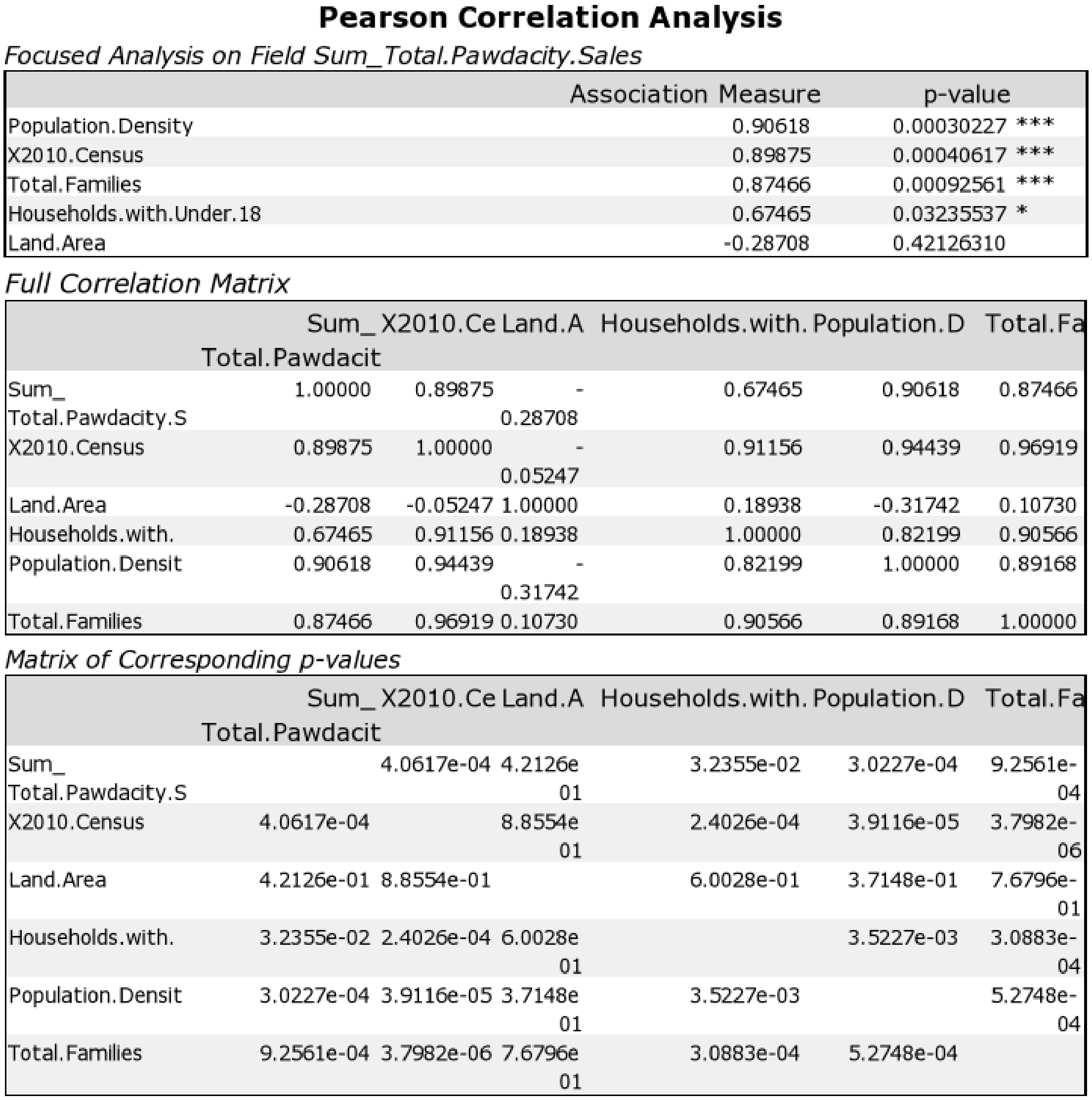
1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

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

   a. Census 2010
   b. Land Area
   c. Household with under 18
   d. Population Density
   e. 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”



Pearson Correlation Analysis

Focused Analysis on Field Sum_Total.Pawdacity.Sales

	Association Measure	p-value
Population.Density	0.90618	0.00030227 ***
X2010.Census	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

	Sum_ Total.Pawdacit	X2010.Ce	Land.A	Households.with.	Population.D	Total.Fa
Sum_ Total.Pawdacity.S	1.00000	0.89875	- 0.28708	0.67465	0.90618	0.87466
X2010.Census	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.	0.67465	0.91156	0.18938	1.00000	0.82199	0.90566
Population.Densit	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

Matrix of Corresponding p-values

	Sum_ Total.Pawdacit	X2010.Ce	Land.A	Households.with.	Population.D	Total.Fa
Sum_ Total.Pawdacity.S		4.0617e-04	4.2126e 01	3.2355e-02	3.0227e-04	9.2561e-04
X2010.Census	4.0617e-04		8.8554e 01	2.4026e-04	3.9116e-05	3.7982e-06
Land.Area	4.2126e-01	8.8554e-01		6.0028e-01	3.7148e-01	7.6796e-01
Households.with.	3.2355e-02	2.4026e-04	6.0028e 01		3.5227e-03	3.0883e-04
Population.Densit	3.0227e-04	3.9116e-05	3.7148e 01	3.5227e-03		5.2748e-04
Total.Families	9.2561e-04	3.7982e-06	7.6796e 01	3.0883e-04	5.2748e-04	

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

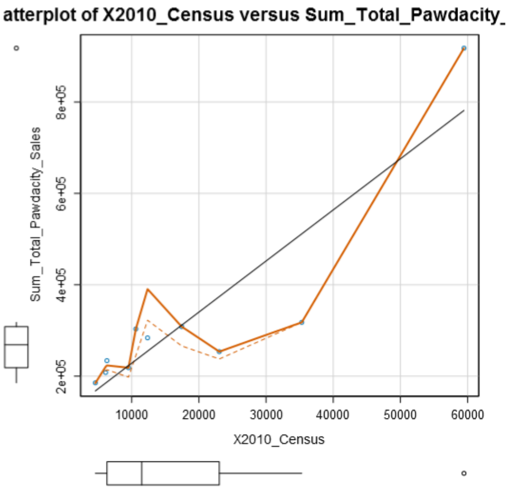
If we take a look on the first three potential predictor variable "Population Density", "Census 2010" and "Total Family" they are all very similar in regards of counting number of human in an area, that also can explains why 3 of them has pretty similar association result, (0.87, 0.89, 0.90). I personally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

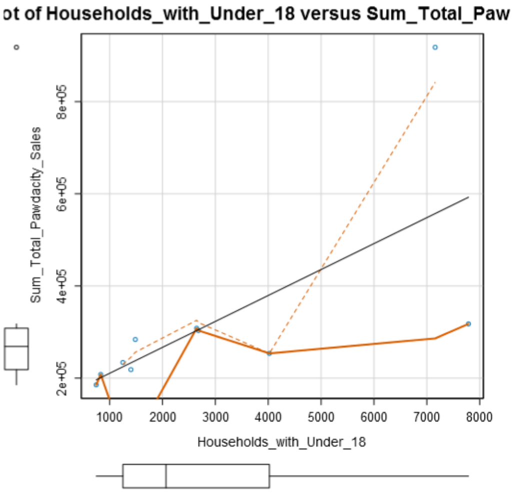
While “Population Density” has the strongest association with sales, I will use Census 2010 as a measure so we can compare with Census 2014 estimate as sales predictor variable. Because I will use “Census 2010”, I will drop both “Density” and “Total Families”

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”

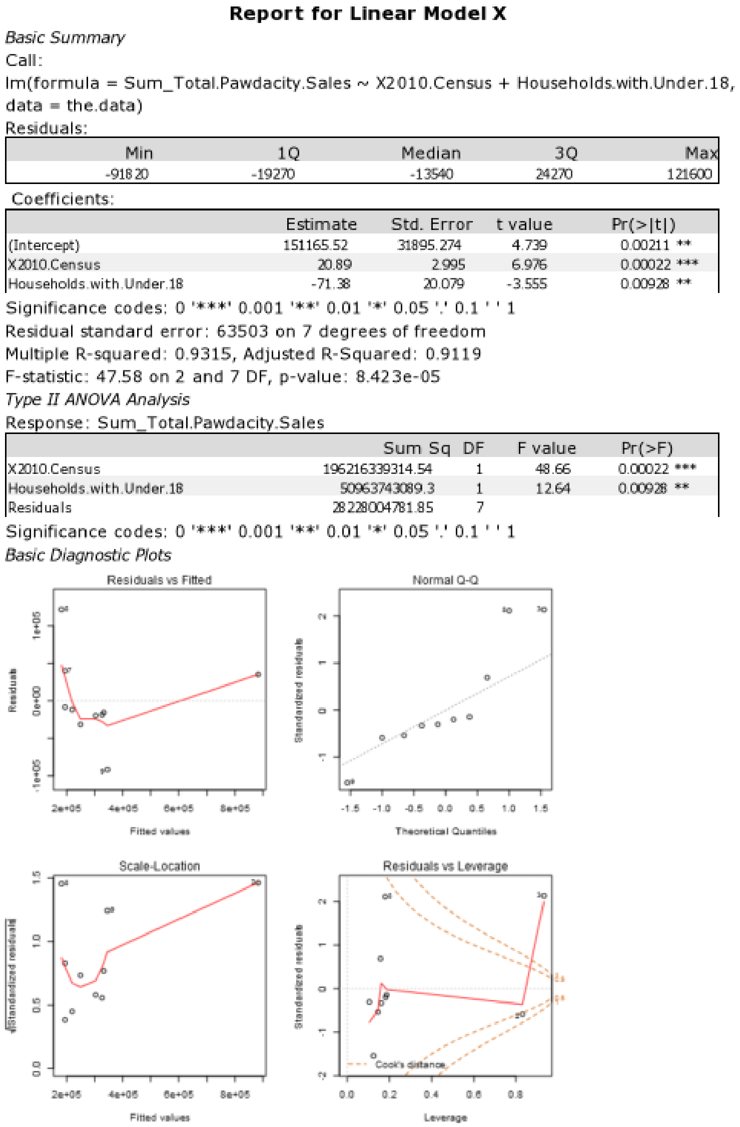
We can see both give correlation , where as “Household Under 18” give more moderate correlation, not as steep as”Census 2010”





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.

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



Report for Linear Model X

Basic Summary

Call:

```
lm(formula = Sum_Total.Pawdacity.Sales ~ X2010.Census + Households.with.Under.18,
data = the.data)
```

Residuals:

Min	1Q	Median	3Q	Max
-91820	-19270	-13540	24270	121600

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	151165.52	31895.274	4.739	0.00211 **
X2010.Census	20.89	2.995	6.976	0.00022 ***
Households.with.Under.18	-71.38	20.079	-3.555	0.00928 **

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

Residual standard error: 63503 on 7 degrees of freedom

Multiple R-squared: 0.9315, Adjusted R-Squared: 0.9119

F-statistic: 47.58 on 2 and 7 DF, p-value: 8.423e-05

Type II ANOVA Analysis

Response: Sum_Total.Pawdacity.Sales

	Sum Sq	DF	F value	Pr(>F)
X2010.Census	196216339314.54	1	48.66	0.00022 ***
Households.with.Under.18	50963743089.3	1	12.64	0.00928 **
Residuals	28228004781.85	7		

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

Basic Diagnostic Plots

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

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)

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 here. Reviewers will use this rubric to grade your project.