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

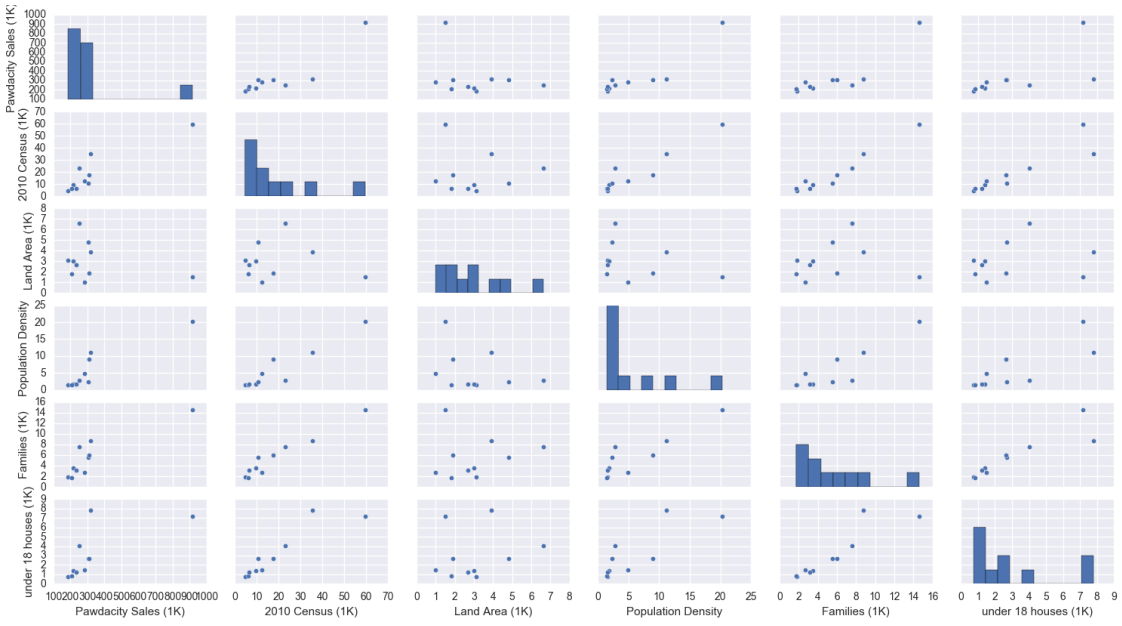
Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section. (750 word limit)

***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.

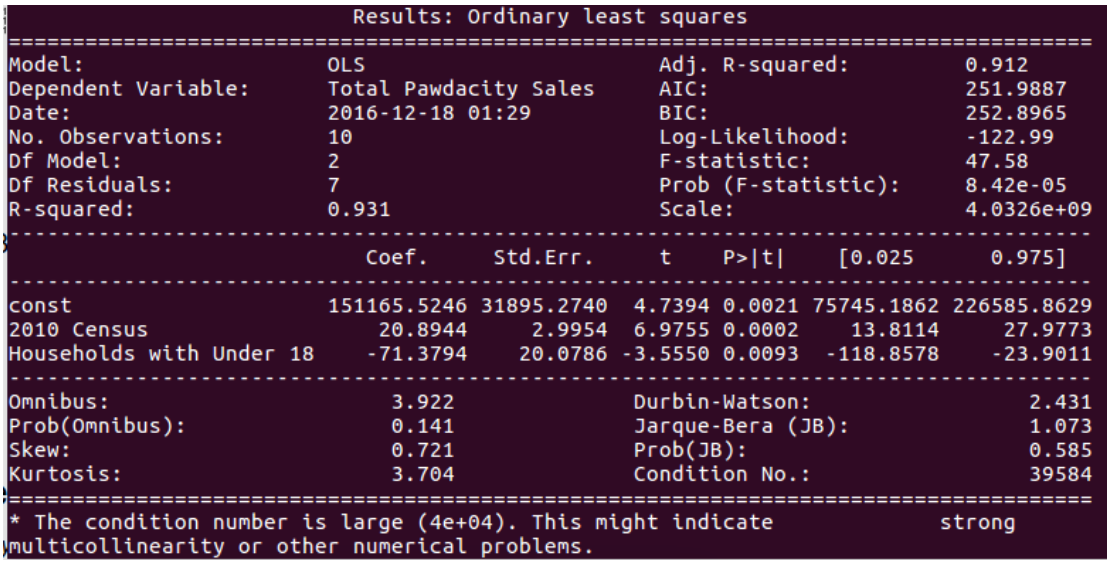


I looked at plots, and eyeballed which ones looked somewhat linear. I also looked at the correlation matrix for which independent variables had high correlation to sales. I chose everything but 'Land Area' as predictor variables, because that one looked like sales were flat with changing land area. Everything else had some linearity to it. Population measures (census, pop density, and total families) were highly correlated, and so I only

used the 2010 census as a population measure. So in the end, I used 2010 Census and Households under 18 as predictor variables.

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.

   It's a good model because the adjusted r-squared is high (0.912) and the p-values of the coefficients are below 0.05, meaning they have a meaningful relationship to the target variable.



Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.912
Dependent Variable:	Total Pawdacity Sales	AIC:	251.9887
Date:	2016-12-18 01:29	BIC:	252.8965
No. Observations:	10	Log-Likelihood:	-122.99
Df Model:	2	F-statistic:	47.58
Df Residuals:	7	Prob (F-statistic):	8.42e-05
R-squared:	0.931	Scale:	4.0326e+09

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	151165.5246	31895.2740	4.7394	0.0021	75745.1862	226585.8629
2010 Census	20.8944	2.9954	6.9755	0.0002	13.8114	27.9773
Households with Under 18	-71.3794	20.0786	-3.5550	0.0093	-118.8578	-23.9011

Omnibus:	3.922	Durbin-Watson:	2.431
Prob(Omnibus):	0.141	Jarque-Bera (JB):	1.073
Skew:	0.721	Prob(JB):	0.585
Kurtosis:	3.704	Condition No.:	39584

* The condition number is large (4e+04). This might indicate strong multicollinearity or other numerical problems.

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)
   sales =
   2010 Census * 20.89
   + Households with under 18 * -71.38
   + 151,165.52

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?

   Laramie looks to be the best City for a new store. It has the highest predicted sales based on the available data, and meets all the criteria (more than 4000 people, etc). Makes sense because it is the most populated City with low competition.

	City	Land Area	Households with Under 18	Population Density	\
0	Laramie	2513.745235	2075	5.19	
30	Jackson	1757.659200	1078	2.36	
77	Green River	3477.361206	2113	1.46	
17	Rawlins	5322.661628	1307	1.32	
35	Worland	1294.105755	595	2.18	
30	Lander	3346.809340	1870	1.63	

	Total Families	2014 Estimate	2010 Census	SALES VOLUME	predicted sales
0	4668.93	32081	30816	76000.0	646933.905708
30	2313.08	10449	9577	182000.0	274323.821821
77	3977.40	12630	12515	0.0	261833.762245
17	2722.43	9227	9259	0.0	251333.528183
35	1364.32	5366	5487	169000.0	223342.137978
30	3876.81	7642	7487	152197.0	174122.106515

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.