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)*

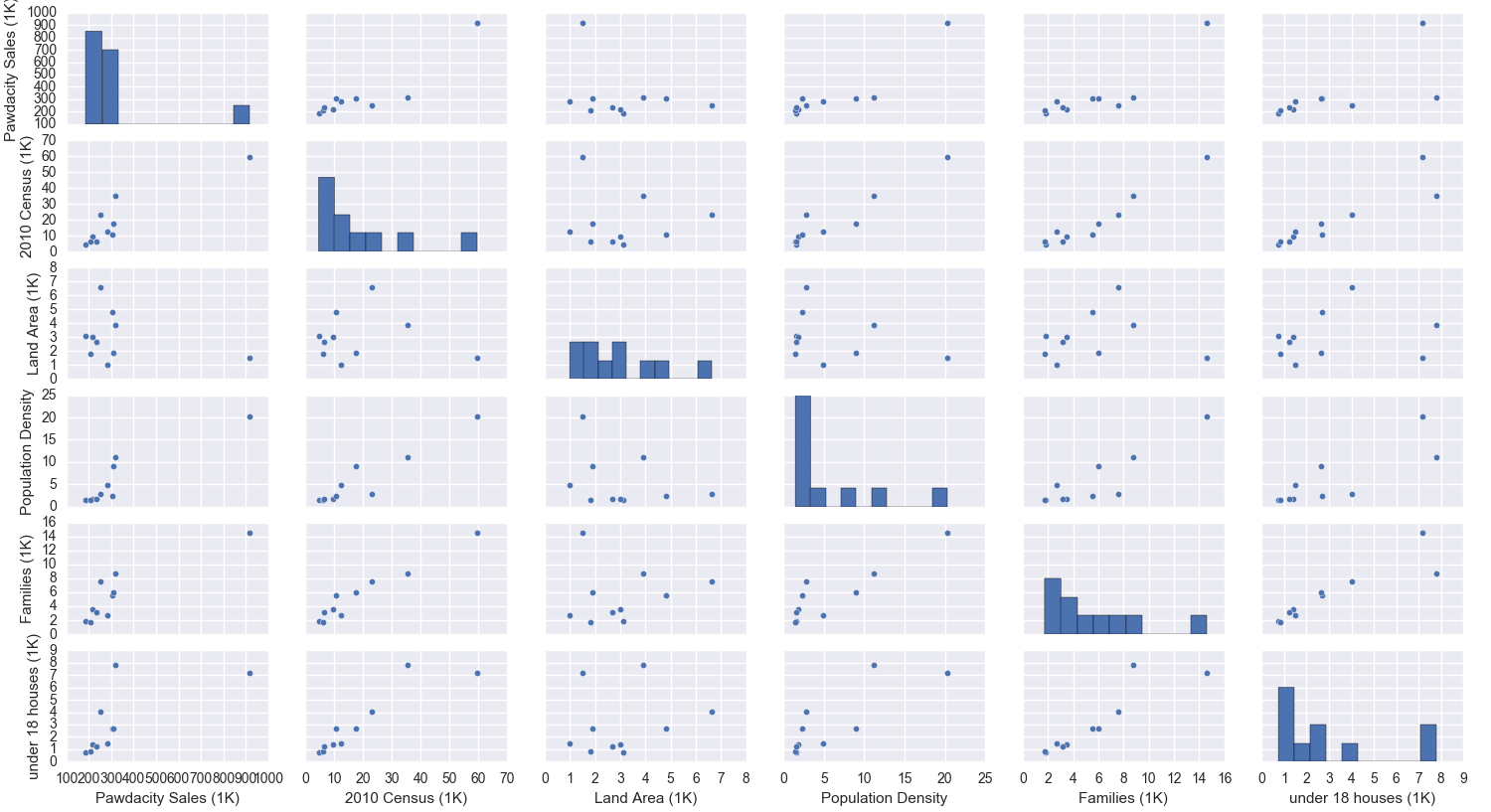
***Important:*** *Make sure you have dealt with outliers and removed one city from your training set. You should have* ***10 rows*** *of data before you begin modeling the dataset.*

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

*At the minimum, answer these questions:*

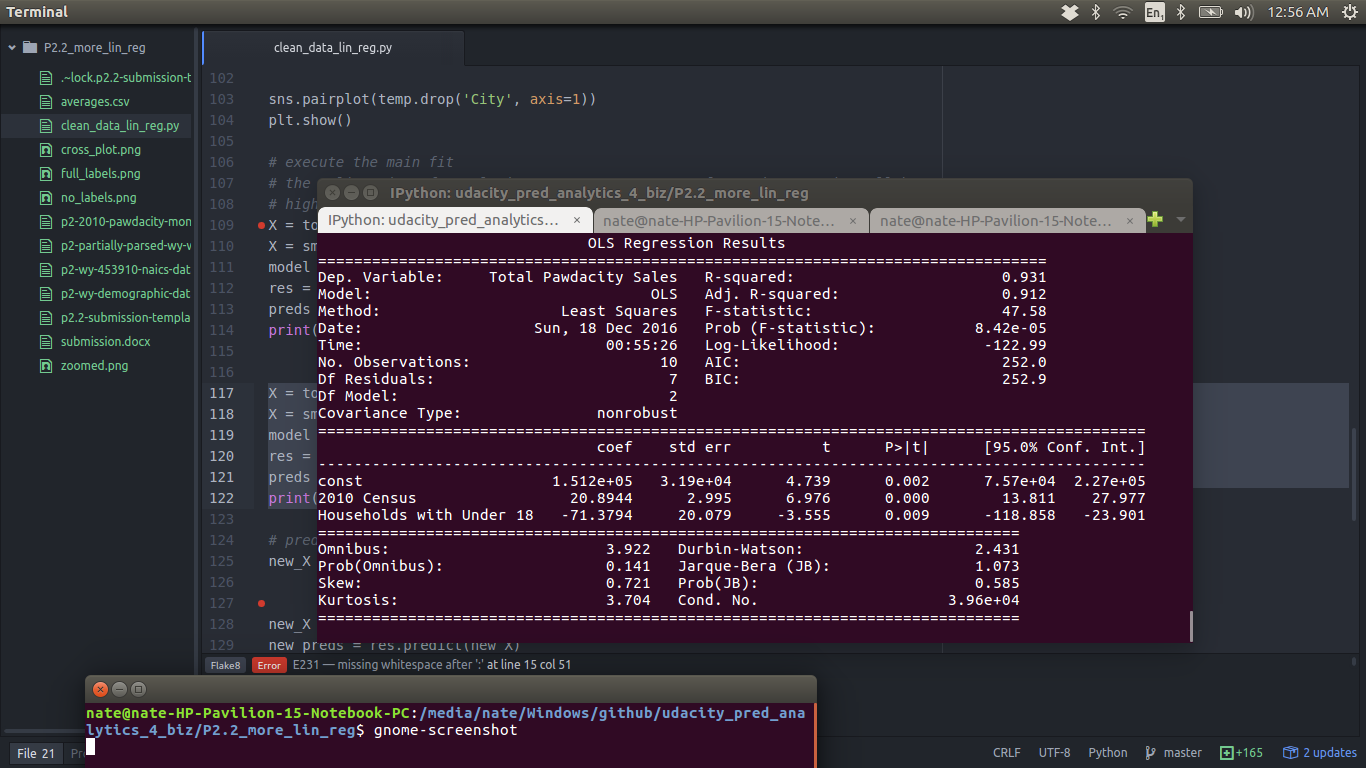
1. How and why did you select the [predictor variables (see supplementary text)](https://classroom.udacity.com/courses/ud976/lessons/4e33b70a-72a4-47cb-959a-28632ae6aaff/concepts/631d190c-8626-4dd7-92df-f5bd96913c48) in your model? You must show that each predictor variable has a linear relationship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

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.



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.

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.



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)

sales =

2010 Census \* 20.89

+ Households with under 18 \* -71.38

+ 15,165.52

## 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?

Before you Submit

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