Project 2.2: Recommend a City

Jean Paul Barddal

Github repo: https://github.com/jaycwb/udacity-select-pet-store-location

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

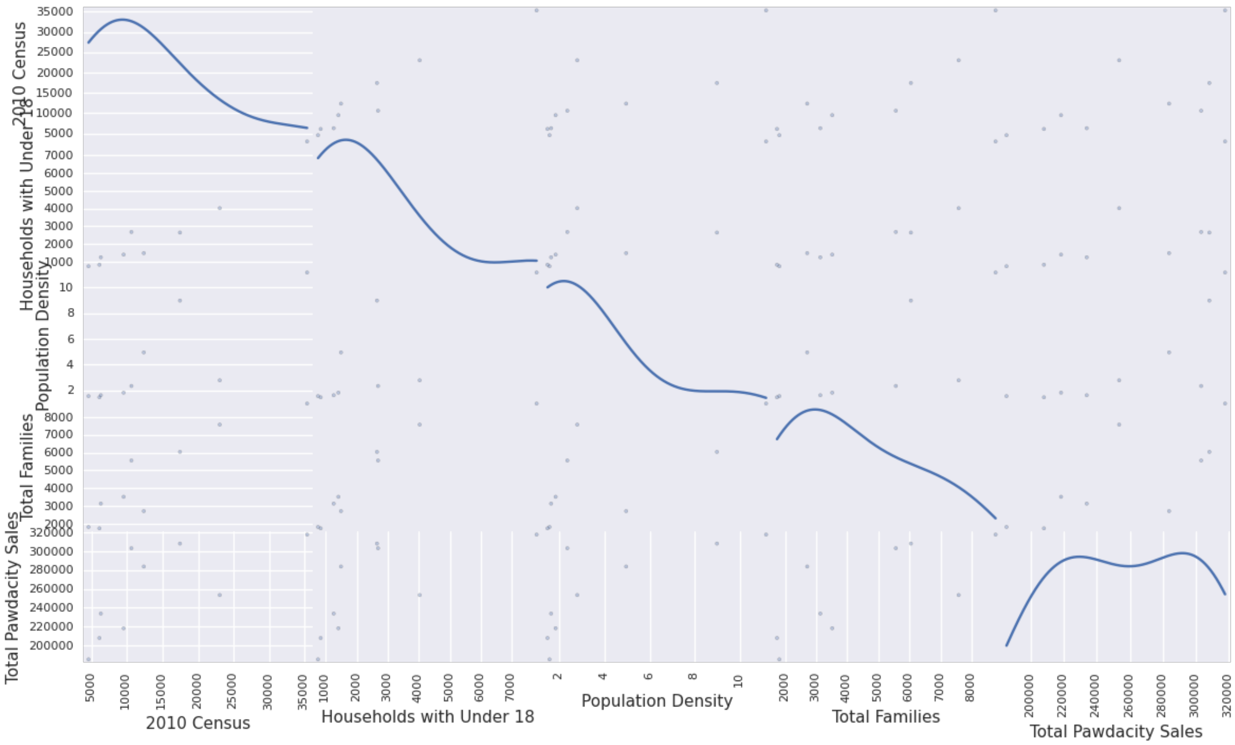
Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section. (750 word limit)

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.

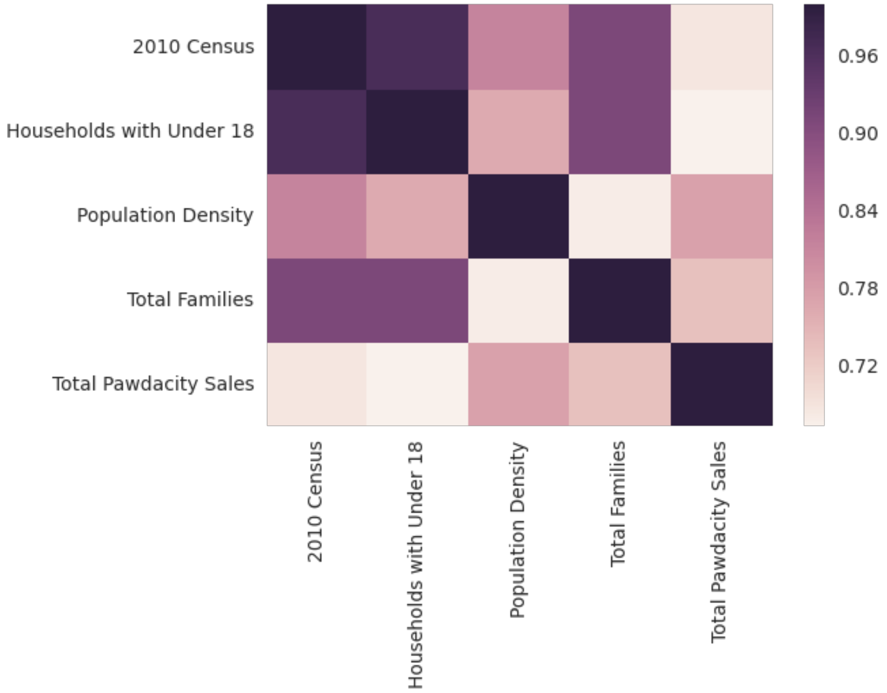
At the minimum, answer these questions:

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. During the P2.1 project, I analyzed the variables, and my rationale to build a regression model is very simple: use all but the city variable in a wrapper feature selection procedure that will maximize the R2 score in the training dataset. Working with the city doesn't make any sense because one of the restrictions that we have in this project is that we should recommend a city where we don't have a store already. Also, optimizing the R2 scores in the full training dataset may lead us to finding optimistic results. Anyway, with the aid of scatter plots, I have verified the data distribution, which was the following:



Here, we can see that the variables seem somewhat disperse and don't show linear trends. But again, we only have 10 instances in the dataset, so any judgement here would be inconclusive.

I've also checked for redundancy using correlations, and the result is the following:



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. To be honest, I don't believe the model I built is robust. The first reason is that we have so little data to work on (10 instances), which I doubt is a sufficient number to correctly identify real-world phenomena. The second is that we don't have a way to be sure that our r2 scores and p-values are reliable. The values I obtained are the following:
    - 2010 Census has a p-value of 0.0404993770689
    - Land Area has a p-value of 0.621327061156
    - Households with Under 18 has a p-value of 0.0466060033703
    - Population Density has a p-value of 0.0141989821257
    - Total Families has a p-value of 0.0240689813987

    - The r2 score obtained during training is 0.70156

    First of all, a widely used strategy is to remove the variables with p-values above a threshold (usually 0.05), and even though I tried, the r2 score decreased to 0.67. Next, our r2 score may be overfitted, since it was obtained by predicting the values for the training data.

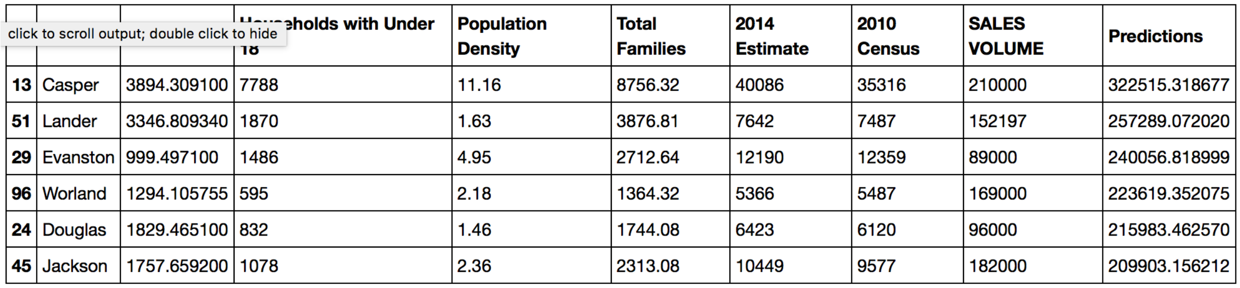
3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)
   Y = -10.74 * 2010 Census + -17.50 * Land Area + 26.26 * Households with Under 18 + 10043.76 * Population Density + 25.12 * Total Families + 233407.99

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? After learning a linear regression model, I've used it to predict the sales for all cities that we had available in our study. After that, I've gathered some information that was available in other datasets (census data and so on), to be able to correctly guess the best city given the project constraints (should be in a new city, total sales should be above 500,000 and so on). The final result is presented in the following table, where Casper is the chosen city, because it has a predict sale value of 322515.32, which is the biggest among all cities tested.



click to scroll output; double click to hide

			Households with Under 18	Population Density	Total Families	2014 Estimate	2010 Census	SALES VOLUME	Predictions
13	Casper	3894.309100	7788	11.16	8756.32	40086	35316	210000	322515.318677
51	Lander	3346.809340	1870	1.63	3876.81	7642	7487	152197	257289.072020
29	Evanston	999.497100	1486	4.95	2712.64	12190	12359	89000	240056.818999
96	Worland	1294.105755	595	2.18	1364.32	5366	5487	169000	223619.352075
24	Douglas	1829.465100	832	1.46	1744.08	6423	6120	96000	215983.462570
45	Jackson	1757.659200	1078	2.36	2313.08	10449	9577	182000	209903.156212

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.