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:

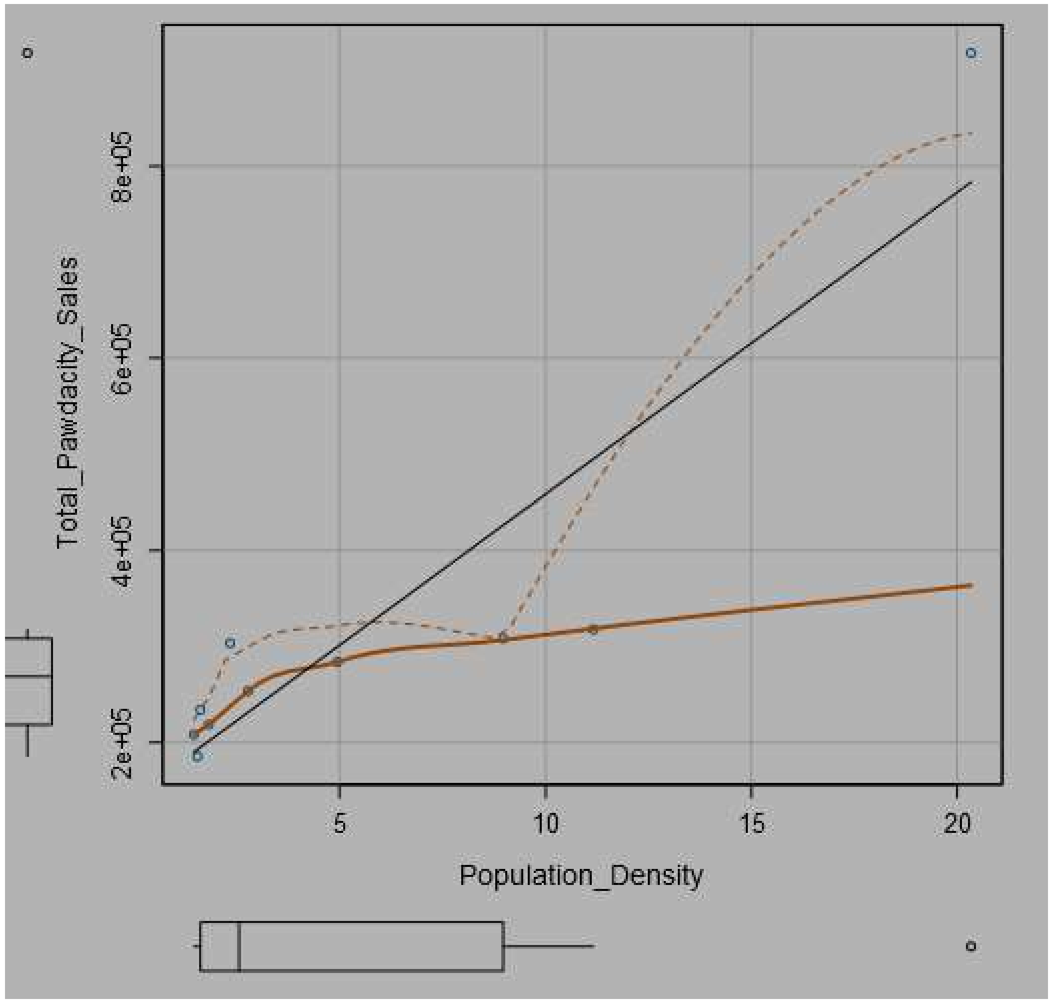
1. How and why did you select the predictor variables (see supplementary text) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.
   I began my analysis by using the Association Analysis tool. My target variable is the Total_Pawdacity_Sales and I selected all of the other metrics as predictor variables. Land_Area has a low negative correlation to sales, -0.28708, and a large p value of 0.42126310. A large p value and a weak correlation mean that Land_Area is not a good predictor of sales. As Land_Area moves up and down, Total_Pawdacity_Sales will not change in a predictable or statistically significant way. The other variables have a higher measure of inner-correlation. 2010_Census, Population_Density, Total_Families, and Households_with_Under_18 are all logically related and the data bear this out with a correlation of at least 0.8. This means that they all move in the same way and will tend to make my model less accurate if they are all used as predictor variables. Population_Density, with its Pearson correclation of 0.90618, Spearman correlation of 0.939393, and p value of 0.00030227, has the strongest explanatory power of Total_Pawdacity_Sales. I chose to leave this variable in and remove the related variables. I didn't choose 2010_Census because, ultimately, the chosen city will be restricted to absolute numbers above 4000 people based on 2014_Census data. I also chose Population_Density because it will be a crucial predictor of Total_Pawdacity_Sales because the higher the Population_Density, the more people are located within a smaller area, and therefore, there will be more people closer to the new store.

: Required: The variable Land Area should not be discarded so fast even though its p-value here is above 0.05. Take a look at this discussion for more detailed explanation why - (https://discussions.udacity.com/t/selecting-predictor-variables-is-there-any-rule-of-thumb/206999)

: Awesome: You are correct here - we need to take multicollinearity into consideration. Using the Association Analysis tool we can verify how associated each variable is with each other. We can see that four variables are highly correlated with each other, meaning that using more than one of them in our model can create a lot of problems when constructing predictive models. One such problem is [overfitting](http://blog.minitab.com/blog/adventures-in-statistics-2/the-danger-of-overfitting-regression-models).

: Required: Using Population Density as a predictor variable won't give us the best performing model.

Could you please create/ look at the correlation matrix heatmap where we should be able to see that HHU18, 2010 Census Population, Families, and Population Density have strong correlations which each other as you noted. Land Area, however, is not as highly



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.
Population_Density, with its Pearson correclation of 0.90618, Spearman correlation of 0.939393, and p value of 0.00030227, is the only variable I selected. The high correlations mean that as Total_Pawdacity_Sales increase or decrease, Population_Density can explain roughly 90% of this change. The model itself has an adjusted R squared value of 0.821. This means that roughly 82% of the change in the target variable, Predicted_Total_Pawdacity_Sales, can be accounted for by the predicted variable, Population_Density. The p-value of the predictor variable is statistically significant by being lower than 0.05.

correlated. So we should start by using Land Area as one predictor and then test the four variables that are correlated. By doing this, we will be able to find the best optimal model. Following this, we can create 4 models with 2 predictor variables each and compare them with each other to see which one is the optimal model. The optimal model will be the one that has the highest adjusted R-squared value, and all of its predictor variables will have p values less than 0.05.

: Required: Nice work with the plot. However, since the predictor variables will change once the correct model is created - please include scatter plots that show that a linear relationship between the correct predictor variables and the target variables exists.

: Required: Very thorough explanation here - nice work! However, the correct model will be able to achieve a higher R-squared values.

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_Total_Pawdacity_Sales = 143800 + 31442 * Population_Density

: Required: This is not the best performing model. Please follow my suggestions above for creating the best performing model.

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?
   I recommend Pawdacity opens a store in Laramie, Wyoming. Laramie does not currently have a Pawdacity store. Competitor sales in Laramie are $76 000, less than $500 000. The 2014 census for Laramie is 32 081, which is greater than 4 000. Predicted sales for Laramie are $306 981.94. Finally, the predicted sales for Laramie are the largest predicted sales for all cities which meet the previous four conditions. Pawdacity should open its 14th store in Laramie.

: Awesome: Indeed, Laramie is the recommended city - well done!

: Required: The predicted sales will change slightly when the correct model is used for calculating them.

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.