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.

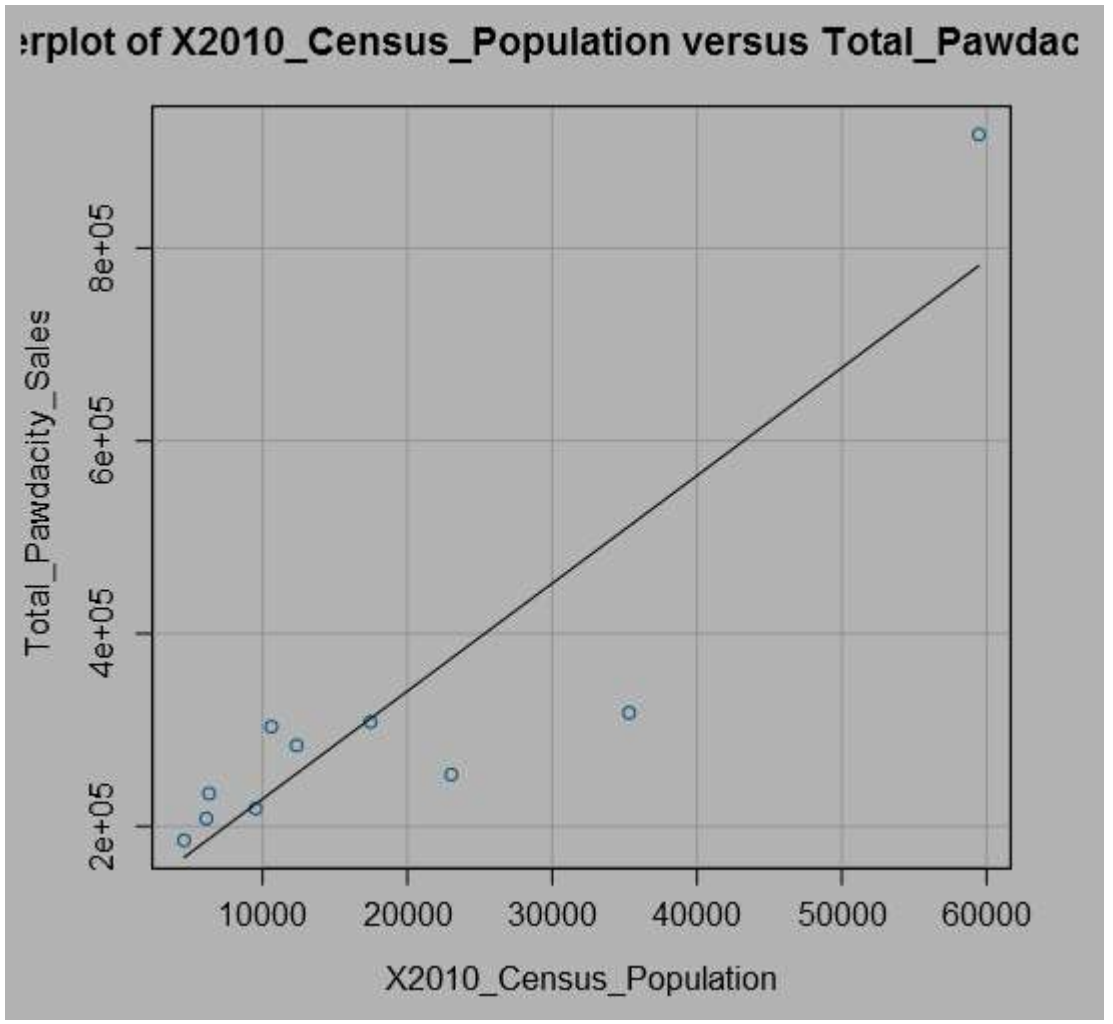
At the minimum, answer these questions:

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 using the Association Analysis tool. My target variable is the Total_Pawdacity_Sales and I selected all the other metrics as predictor variables. Although Land_Area has a low correlation and high p value, I want to do further testing before I decide to keep or remove it. 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. 2010_Census, with its Pearson correclation of 0.90618, Spearman correlation of 0.90303, and p value of 0.00037, has strong explanatory power of Total_Pawdacity_Sales. I chose to leave this variable in and remove the related variables. Land_Area has a low negative correlation to sales, -0.28708, and a large p value of 0.12668. A large p value and a weak correlation mean that Land_Area is not a good predictor of sales by itself.

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.
   2010_census has a p value of 0.00037 and Land_Area has a p value of 0.12668. The p

: Required: Indeed, we need to use Land Area as one of the predictor variables, but we need to combine Land Area with one of the highly correlated variables which will give us a model where both Land Area and the other predictor variable are BOTH statistically significant (p-values less than 0.05) . Hence, "2010_census" and the "land area" together do not give us the best performing model.

value of Land_Area does not look statistically significant. However, I created two linear regression models, one with Land_Area and one without. The model with Land_Area has an adjusted R squared value of 0.827 against the smaller 0.784 of the model without Land_Area. To further confirm these results, I used the Nested Test tool to help me select which model performs better. The f value of removing the Land_Area value was 0.12668, which is higher than 0.05. This means that the Land_Area variable helps make the model more accurate. The high correlations mean that as Total_Pawdacity_Sales increase or decrease, 2010_census can explain roughly 90% of this change. The model itself has an adjusted R squared value of 0.827. This means that roughly 82% of the change in the target variable, Predicted_Total_Pawdacity_Sales, can be accounted for by the predictor variables, 2010_census and Land_Area..



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 = 210827.04 + 11.03 * 2010_census – 30.32 * Land_Area

: Required: Please correct the model and include:

- a scatterplot that shows that a linear relationship between Land Area and the target variable Predicted sales exists

- a scatterplot that shows that a linear relationship between the other correct predictor variable and the target variable Predicted sales exists

: Required: The 2010 census variable should not be included in the model. There is another variable that will give us a higher adjusted R squared value (more than 0.82). Please create and compare these 4 models:

- Land Area + Total Families
- Land Area + 2010 census
- Land Area + Household under 18
- Land Area + Population density

Compare the model. We should choose the one that has the highest adjusted R-squared value, considering the fact that all of its predictor variables have p - values less than 0.05 (are statistically significant).

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 $474 790.55. 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.

: Required: this number will change after the correct model is used for calculating it.

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.