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

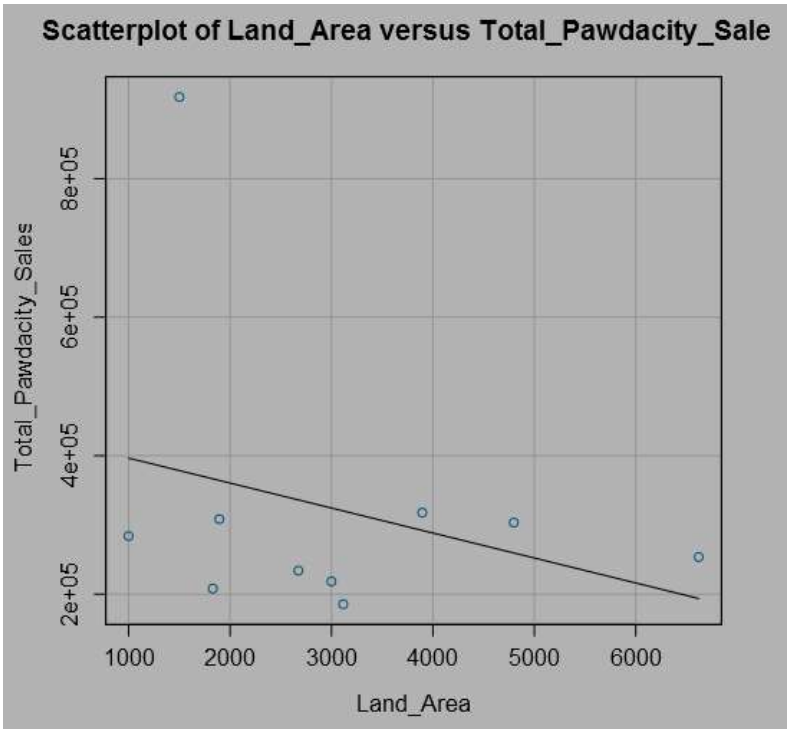
Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section. (750 word limit)

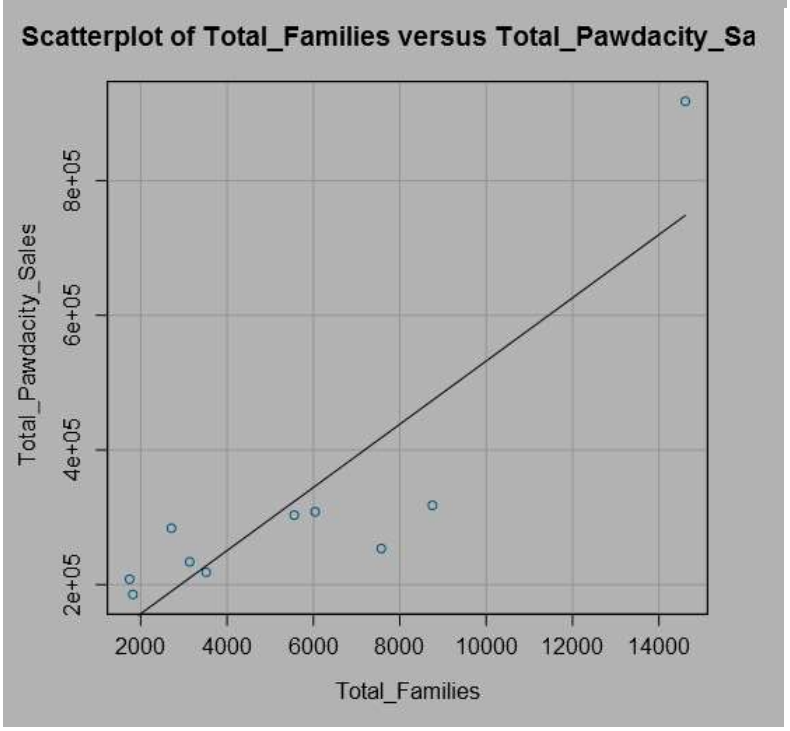
Important: *Make sure you have dealt with outliers and removed one city from your training set. You should have* ***10 rows*** *of data before you begin modeling the dataset.*

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 linear relationship with Total_Pawdacity_Sales as shown by the scatterplot, it has low correlation. 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. Total_Families, with its Pearson correclation of 0.874663 and Spearman correlation of 0.842424, has strong explanatory power of Total_Pawdacity_Sales. I chose to leave this variable in and remove the related 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.
   Total_Families has a p value of $8e^{-5}$ and Land_Area has a p value of 0.01123. The model has an R squared value of 0.912 adjusted R squared value of 0.887. The high

correlation means that as Total_Pawdacity_Sales increase or decrease, the model can explain roughly 91% of this change.

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 = 197330.41 + 49.14 * Total_Families – 48.42 * Land_Area

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 $305 013.88. 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.

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.