# 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)*

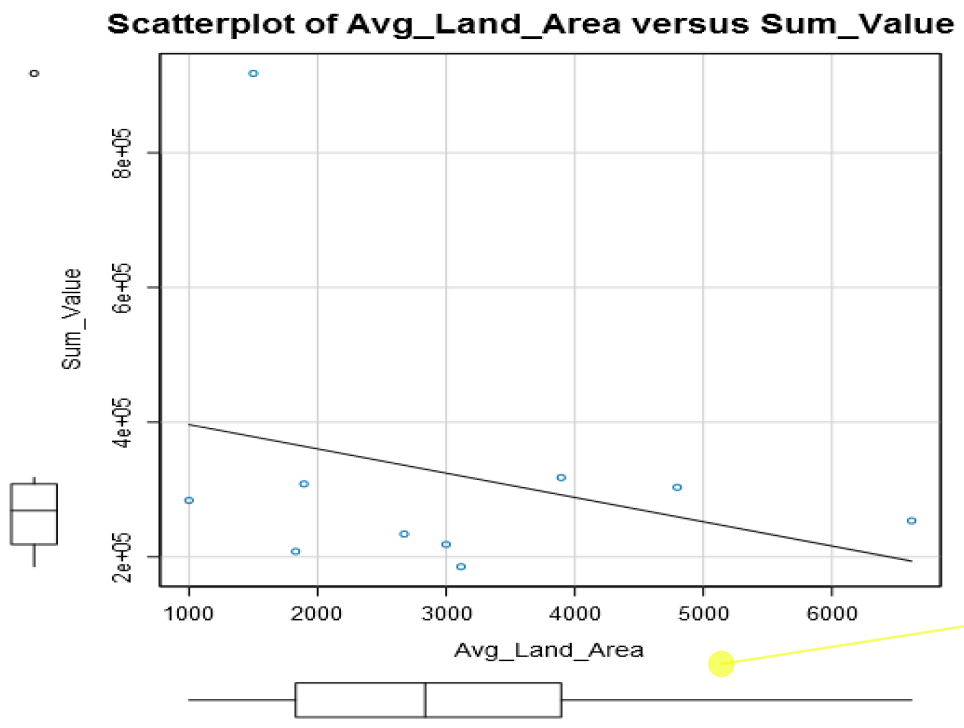
***Important:*** *Make sure you have dealt with outliers and removed one city from your training set. You should have* ***10 rows*** *of data before you begin modeling the dataset.*

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

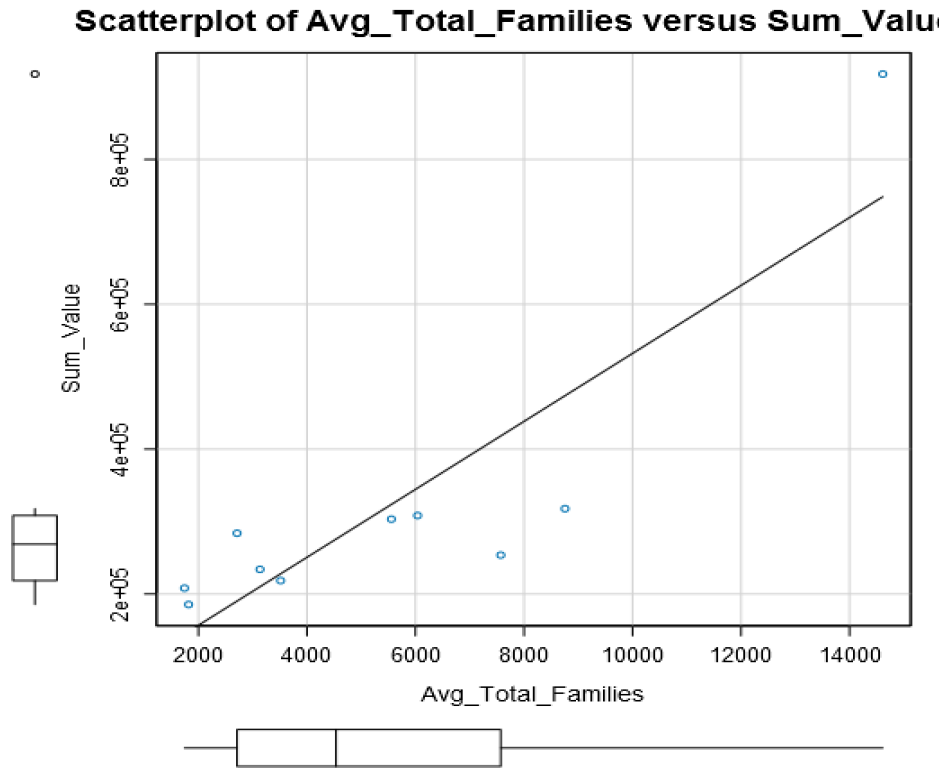
*At the minimum, answer these questions:*

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.

   First, I verified the correlation between each variable and the target variable (sales). The result shows that most variables present relatively high correlations, while one (land area) has a small one. I draw scatterplot to find the relationship among different variables. I find that land area should be considered as a predictor variable because the scatterplot shows a negative relationship between these two variables. Next, I verify how variables correlated each other. I see that many are redundant, and thus, I pick only one of them to be used in the model. I play around with the variable combinations, and I find the model that use land area and total family number as predictor variables presents a reasonably high R2 score and low p-values. The R2 score is 0.9118 and the p-value is 0.0002035.




: Suggestion: The two important scatterplots have been included. Please note that we don't need to look into the Averages of the predictor variables. We should use them as they are just Land Area not Avg_Land_Area. Same goes for Total Families.



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.

   The p-value of land area is 0.01123 and the p-value of total families is 8e-05. Both of them have statistics significance. The multiple R-squared value is 0.9118 and the adjusted R-squared value is 0.8866, which shows that this linear model is a good model.

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)

   Sum_Value = 49.14 * Total_Families – 48.42 * Land_Area + 197330.41

: Awesome: Correct!

## 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 would recommend Casper and the predict sum value of this city is $438997.17. There are no current Pawdacity stores in this city. The 2014 census shows that about 7788 people in this city and the predict yearly sales is over $200000.

: Required: Casper is not the recommended city. When I looked into your workflow it appears that you are just using the 10 cities to select from - that is not correct. Please use your regression model to calculate predicted sales for all of the cities and use the criteria given to you to make a recommendation. Hint: The name of the recommended city starts with "L".

## 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.