MAKE A COPY: Practice Project: Recommend a City

**Note that this project is a continuation of the Data Cleanup project.**

## 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)](https://classroom.udacity.com/courses/ud976/lessons/4e33b70a-72a4-47cb-959a-28632ae6aaff/concepts/631d190c-8626-4dd7-92df-f5bd96913c48) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.

Below are the plotted predictor variables against the target variable:

|  |  |
| --- | --- |
| A screenshot of a cell phone  Description automatically generated | A close up of a mans face  Description automatically generated |
| A close up of text on a white background  Description automatically generated | A close up of a map  Description automatically generated |
| A close up of a person  Description automatically generated |  |

I noticed that all the predictors could be considered as a good predictor variables because they are strongly related to the total sales .

Next, I tried to find out if there are correlations between the potential predictor variables, the martix below shows the correlations:

A screenshot of a cell phone

Description automatically generated

I highlighted the variables that shows a strong correlation:

Correlated variables: Census2010,HHU18,Population\_Density,Total\_Families

Uncorrelated variable: Land\_Area

Then a used a linear regression tool to test the generated models and select the best one, I found out that choosing land area with total families produce the best predicting model :

![A screenshot of a cell phone

Description automatically generated]()

1. 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.

Because the p-value for both predictors are less than 0.05, and the Multiple R-squared is 0.9 which is very good value as it is almost 1.

1. 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 = 197,330.41 - 48.42 \* [Land\_Area] + 49.14 \* [Total\_Families]

## 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?