MAKE A COPY: Practice Project: Recommend a City

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

Step 1: Linear Regression

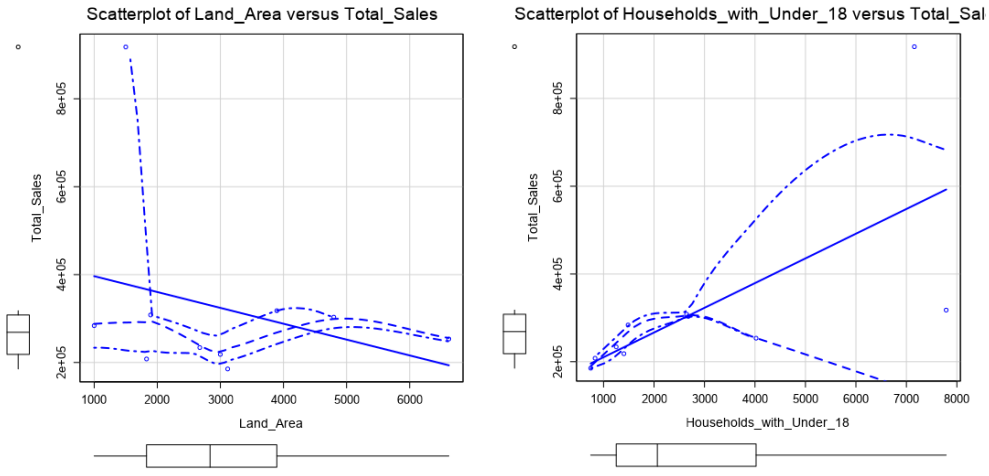
Create a linear regression model off your training set and present your model. Visualizations are highly encouraged in this section. (750 word limit)

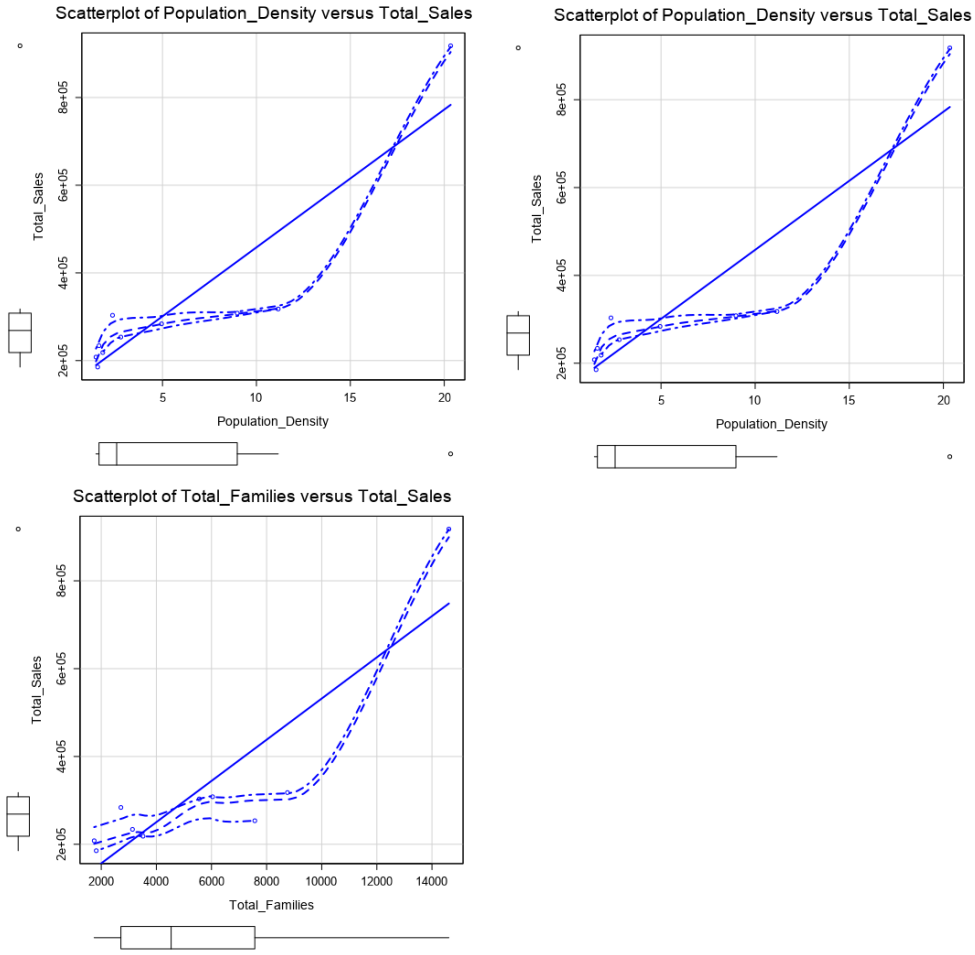
***Important:** Make sure you have dealt with outliers and removed one city from your training set. You should have **10 rows** of data before you begin modeling the dataset.*

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) in your model? You must show that each predictor variable has a linear relationship with your target variable with a scatterplot.
   First, I plotted each predictor variable against the target variable. All predictor variables show a linear relationship between sales.





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.

   When Linear regression is performed the lowest p-values were shown in Households.with.Under.18 and Total.Families. So I built a model with only those two predictors.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	209811.725	87568.456	2.3960	0.07468 .
Land.Area	-54.756	46.996	-1.1651	0.30872
Households.with.Under.18	-42.506	31.518	-1.3486	0.24875
Population.Density	-16308.444	26367.746	-0.6185	0.56973
Total.Families	64.293	39.043	1.6467	0.17496
X2010.Census	7.749	7.998	0.9688	0.38752

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61470 on 4 degrees of freedom
Multiple R-squared: 0.9633, Adjusted R-Squared: 0.9174
F-statistic: 21 on 5 and 4 degrees of freedom (DF), p-value 0.005675

   But in this case the value of the Households.with.Under.18 was greater then 0.05, so

with testing of the rest prediction variables the Land.Area showed good results
The p-value for Land.Area = 0.01123 (≤ 0.05)
The p-value for Total.Families = 8e-05 (≤ 0.05)
Multiple R-squared value = 0.9118 (>0.9)

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)

   Y = 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?

I would recommend the city of Laramie with a predicted sales of $305,014.