Assignment 3 - Module 5: Linear Regression

Your Name Here

due date here

These exercises will require you to load the area_data.Rds file. For this problem set we're interested in predicting the percent of population in the labor force (dependent variable).

1. Plot the distribution of your outcome and provide 1-2 sentences interpreting what the plot tells you (please put sentences outside of code chunks).

2. Let's examine our first independent variable: Percent of population with commute to work of 30 minutes or more. Create a scatterplot of the relationship and provide 1-2 sentences interpreting what the plot tells you (please put sentences outside of code chunks).

3. Split your data into training and testing data. Please use `set.seed(26151)` so that our results are the same.

4. Run a model (on the *training* subset of the data) that predicts the percent of the population in the labor force (dependent variable) as a function of the percent of the population that has a commute >30 min (independent variable). Make sure you use the workflow process and present the regression table of results.

5. Provide a sentence interpreting the slope of your independent variable in your regression model.

6. Provide a sentence interpreting the intercept of your regression model.

7. Add your predictions to your testing dataset and calculate the RMSE. Provide 1-2 sentences interpreting the RMSE (please put sentences outside of code chunks).

8. Expand your regression model to a multiple linear regression model by adding two more independent variables: percent of the population that is college educated and percent who moved in to the locale. Provide a sentence interpreting the slope of each independent variable in this multiple linear regression model.

9. Add predictions from your multiple model to your testing dataset and calculate the RMSE. Provide 1-2 sentences on what the RMSE means and how it compares to the RMSE from the simple (one predictor) model.

10. Overall takeaways: Which predictors appear to be related to your outcome? How do you know? What were your overall model findings/takeaways?

BONUS: Run a third model ADDING a categorical variable (hint: you have to use as.factor in your code). Present the summary table. Did your model fit improve? Was your added variable a good predictor? Why/why not?