Problem Set 3: LMS Modules 6 & 7

YOUR NAME HERE

due INSERT DATE HERE

Structural stuff:
1. Be sure to change the "author" above to your name. Also insert the due date for this term/assignment.
2. Save your .Rmd file as LastName_FirstName.Rmd (do this before you knit).
3. You need to submit your .Rmd code file AND a knit file (upload both simultaneously to the course webpage; you can't upload them one-by-one). You will only receive full credit if you upload both files (note: you get 20 points JUST for uploading both). 4. Below I have set up the file for you with the libraries you'll need. I have also inserted code chunks for you (note, I won't do this every time).
5. I expect that the .Rmd file you submit will run cleanly, and that the knit file won't contain any errors (LOOK at the knit file after you create it - if questions/text are running into each other, if you see error messages, etc., you're not done).
6. You can use comments to tell me what you are doing either in text or in code chunks, but remove "old" code that didn't run/work.

This assignment includes questions from Modules 6 & 7 (regression and scatterplots). These exercises will require you to load the area_data.Rds file.

Area Dataset Codebook
* name - Name of Micro/Metro Area
* college_educ - Percent of population with at least a bachelor's degree
* perc_commute_30p - Percent of population with commute to work of 30 minutes or more * perc_insured - Percent of population with health insurance
* perc_homeown - Percent of housing units owned by occupier
* geoid - Geographic FIPS Code (id) * income_75 - Percent of population with income over 75,000
* perc_moved_in - Percent of population that moved from another state in last year
* perc_in_labor_force - Percent of population in labor force * metro - Metropolitan Area? Yes/No * state - State Abbreviation * region - Census Region * division - Census Division

Research Topic: Using this data, we're trying to help advise local politicians and city planners in our area to see if we can help predict factors that contribute to population growth. Our dependent variable for this analysis will be the percent of our population that has moved to our city from another state in the past year (perc_moved_in). Make sure you keep this goal in mind as you choose potential predictors for your analysis.

1. Import/load the "area_data.Rds" dataset here. (4pts)

2. Examine the 5-number summary of the outcome variable (`perc_moved_in`) and plot its distribution. (6pts)

2a. What did you learn about the outcome from the summary and plot? (3pts)

3. The first predictor we think might relate to growth is the percent of population in the labor force (i.e., employment rate). Plot a scatterplot (and include a linear regression line) to examine the possible association between this predictor and our outcome. (5pts)

3a. What does the scatterplot indicate about the potential relationship between these two variables? (2pts)

4. Split your data into training and testing data. Please use `set.seed(1540)` so that our results are the same. (5pts)

5. Run a model (on the *training* subset of the data) that predicts the percent of the population who move in (dependent variable) as a function of the percent of the population in the labor force (independent variable). Make sure you us present the regression table of results. (8pts)

5a. Provide a sentence interpreting the slope of your independent variable in your regression model. (2pts)

5b. Provide a sentence interpreting the intercept of your regression model. (2pts)

5c. Provide a sentence interpreting the overall model (hint: Rsquared). (2pts)

6. Add predictions from your simple linear regression above to your testing dataset and calculate the RMSE. (8pts)

6a. Provide a sentence interpreting your RMSE. (2pts)

7. Expand your regression model to a multiple linear regression model by adding two more independent variables: the percent of the population who own their home and the percent of the population who are college educated. (5pts)

7a. Provide a sentence interpreting the intercept in this multiple linear regression model. (2pts)

7b. Provide a sentence interpreting the slope of each independent variable in this multiple linear regression model (one sentence per slope). (6pts)

7c. Provide a sentence interpreting the overall model (hint: Rsquared). Did this improve from your simple model? (2pts)

8. Add predictions from your multiple linear regression to your testing dataset and calculate the RMSE. (8pts)

8a. Did your RMSE improve from the simple model? If so, how much? (4pts)

8b. What are your final thoughts on this model? Were you able to adequately predict the percent moved in based on these variables? If you had more time, what other variables might you explore (and why)? (4pts)