Assignment for Week 9 / Modules 5 and 6 (Merged and consolidated)

In this assignment, you'll be asked to predict reading scores using the ELS data.

Scatterplots.

1. Choose a dependent variable and plot that dependent variable as a function of a continuous independent variable.

Use ggplot ☺

2. If you need to, adjust the scale of either the dependent variable or independent variable to make more sense.

Yes – try this and show the results (even if it makes the plot "worse").

3. Add a line of best fit to your graphic.

Show **both** the loess and lm lines. Hint: Use geom smooth.

4. Clean up the labeling and description of the graphic so that it could be presented to a broader audience.

Make it pretty. Hint: Investigate the resources for ggplot and find which parameters can be altered for aesthetics.

5. In your .Rmd file, include a few sentences describing what the graphic above shows.

Also, clearly explain the difference between the loess and the lm estimates**. Explain which would be "better" for prediction purposes and why.**

Regression.

6. Create a regression that predicts reading scores as a function of SES plus one other covariate, using the training data.

First load the training data and proceed with "learning" a regression model. This is synonmous with learning the parameters to the linear equation, such that the linear equation will "best" model the data. Hint: use the lm(…) function; you will need to clearly identify the independent and dependent variable(s)

Hint: I encourage you to plot the results so that you can visually confirm the result.

7. Report the RMSE from a validation of your model using the testing data.

First you will need to load the testing data. Then apply your learned model (from #1) to the test data. The prediction will likely not be perfect, and thus there will be some error … that's what we are computing here !! Hint: use the modelr::rmse(…) function. This function will apply the regression model and measure the error!

Hint: (As you have done previously) You may wish to store intermediate computational results of the error back into the data frame using mutate.

Hint: I encourage you to plot the results so that you can visually confirm the result.