Kelsey Hoekstra
Final Project

This project was interesting because there are many variables but none are named, so I couldn't common-sense my way through this. There also were too many variables to just run a regression and see what happens (I tried and R crashed).

The first thing I did was visualize as much as I could. I could graph the continuous variables and kept randomly selecting categorical ones to create boxplots. After that, I ran a simple linear model using only the continuous variables. As you'd expect, it didn't score that well. I also did a random categorical submission. Also not that successful. All this was done on the file final!.r

Since I couldn't pick variables based on common sense, I had to do a more exhaustive search. I worked in groups of 20 variables to run a lasso regression, including cross validating for lambda and final scores. I used the lasso graphs to determine which variables contributed most in each chunk. This was done in final part 2.r. There were some variables that weren't working with the models, so I ran troubleshooting.r to try to identify the problem ones and remove them. Since I had no idea what variable meant what and what the errors were saying I just moved on. I then took all my impactful variables from the exhaustive lasso tests to create the super combination. It did much better.

After that, I moved on to final part 3.r because I don't like it when code gets too long. I wanted to see if I could make the model simpler, so I removed the variables one at a time and saw the impact on the score. Those who barely impacted the score were removed. I also used that r file to compare lasso and ridge regression, and it turned out that ridge was more accurate than lasso. I changed my model for the final submission.

In the end, in my final submission I went with the ridge simplified model. It didn't score as well as the lasso super model, but it worked much faster.

For my final project submission I was unable to upload train or test data sets. They can be found here: https://www.kaggle.com/c/allstate-claims-severity/data