For this assignment, you are going to build a model to predict the trip time in seconds of a taxi trip. Using the previously provided sample of the NYC taxi trip database.

1) Split the data into three sets: a training set, a validation set, and a test set. The test set should be about 10% of the data, the validation set should be another 10% of the data, and the training set should be the remaining 80% of the data.

2) The most basic prediction one can make for trip\_time\_in\_secs is the average trip time in seconds. Calculate the mean\_trip\_time\_in\_secs on the *training*dataset, then use this value as your prediction for every observation in the *validation* dataset. Calculate the mean squared error on the validation set using these predictions. This is the baseline score that we want to beat.

3) Build a linear model using one predictor variable. Train the linear model on the *training* set, then use the model to make predictions on the *validation* set. Calculate the mean squared error and compare to the baseline score (it should be considerably better).

4) Build several more linear models using different combinations of predictor variables (minimum of two additional models total). As in step 3, for each model, make predictions on the *validation* set, and calculate the mean squared error.

5) Select your best performing model from step four. Use this model to make predictions on the *tes*t set. This is your final assessment of model performance. Calculate the test set mean squared error.

In steps 3 - 4, limit your choice of predictors to only features that would be available at the start of a taxi trip. For instance, dropoff\_datetime is not a reasonable choice of predictor, nor is fare\_amount, but pickup\_longitude is a reasonable choice, as is medallion, or hack\_license.

Do your work in a Jupyter notebook; when complete, export to html and submit here by the due date.