High Level Documentation

Our codebase consists primarily of 2 parts: the R file which performs the data cleaning and merging, and then the python notebook which performs the ridge and lasso regression.

R File: Data Cleaning and Merging

In order to process the varying challenges presented in the data we are using dplyr and tidyverse to efficiently process the dataframes.

One of the important things that this performs is it unifies the datetime variables which makes it possible to merge them together based off of the time.

The csv file that this produces is also provided in the code base to simplify testing it out.

Python Notebook: Ridge and Lasso Regression

After importing the data, it checks for any NA values within the data and drops them.

We then move on to splitting the data, which we perform on 3 different sets of predictors so that we are able to compare ridge and lasso regression across different dimensionalities of datasets.

After fitting the models, we use the data we gathered to compare them.