For the foundation project I'm looking to do the Bike Sharing data on Kaggle.com. The data set contains hourly and daily bike usage total counts for casual and registered users with temperature, wind speed, and humidity for a two-year period from January 1, 2011 to December 31, 2012. The goal is to predict usage and detect anomalies.
I am planning to build a model using the first year as a training set and second year as a test set, and another model using randomly selected training and test sets from the entire two-year period perhaps using k-fold cross-validation. I am also going to learn and apply at least one anomaly detection method to automatically identify usage outliers.