Predicting Shared Bike Usage in San Francisco

Bike Share: Data Wrangling

Data Wrangling

The three datasets for this project are fairly clean. The weather dataset had the most missing data, primarily in the 'events' column (~85% missing). This was expected as this column recorded significant weather events such as storms. Missing data was filled with NAN. Max gust speeds also had a significant amount of missing data (~25%). As this column records gusts and regular wind patterns were recorded elsewhere, missing data was filled with NAN.

The trip dataset is very complete. The data was subsetted to just trips in San Francisco (a majority of the data). The main cleaning in this set was in removing outliers, specifically trips which seemed to be errors. Plotting the trip durations revealed many implausible trip times. The mean duration of 17.117811 was greater than the 75% of the data, and the std was 380.557814 (over 6 hours). The max duration is 4797 hours. Clearly, outliers are strongly affecting the data. Rather than replacing the values with the median value for duration, these trips will not be included in the clean dataset as these trips are most likely errors in usage. Trips falling outside of the 98 percentile were discarded. There were many trips with durations of less than five minutes. Further exploration revealed that many of their start and end stations were the same and presumably due to user error in checking out the bikes. Trips with the same start and end stations were removed. After the removal of the outliers, the mean is now 10.652207 and the std is 9.414143.