Teo Speece

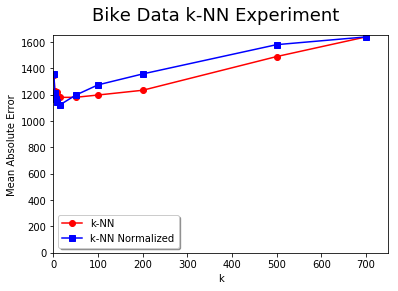
# Problems and Data

# In this assignment, the objective is to implement the k-Nearest-Neighbor algorithm on the bikeshare\_projectA.csv data set to predict the demand for a bike sharing service based on information about the date and the weather. To obtain this prediction we used the predictors: season, month, weekday, weathersit, tempc, and windspeed. The only transformation made to the variables was normalization by the Z-score method. This method makes the distribution more normally distributed which helps the values be centered more around the mean

# Functionality

# A completed run of the predictions can be found in output lines 105

# Expirements and Results

# copy of visualization 

Insights

# at around k = 50 the normalized model the performance grew worse than original data. At around 15 for k-NN normalized and 50 for the k-NN the k grew too large and the performance dropped. Based on the plot I learned that k nearest neighbor by itself is not effective at large values of k. As k approaches the observations more values are used to predict the one value. A good way to think of it is if 600 nearest neighbors were used, that is almost the size of all the observations and all variables are taken into account making it difficult to predict. 600 does not explain much better than using all the observations. To summarize how different k’s work- the lower the k the less stable the predictions are. Increasing k increases how close predictions are until k becomes too large and error increases again.

# Interpertation

# k-nearest neighbor works best with lower k’s because it is easier to predict values when points are more compact. At high k-values the range of values it could be is higher so it becomes more difficult to predict the value. The more observations there are the higher the k can be and still perform well. To summarize how different k’s work- the lower the k the less stable the predictions are. Increasing k increases how close predictions are until k becomes too large and error increases again. Comparing the plots the normalized was better than the original data at low ks.

When comparing the non-normalized plot to the normalized plot, using Z-score method, at low k’s the normalized method is better, (turning point is around 50) but as k increases the error for the normalized plot surpasses the error for the non-normalized plot. This higher error is because the normalization has smaller values therefore overfitting occurs earlier than the non-normalized.