Data 100 Final Project

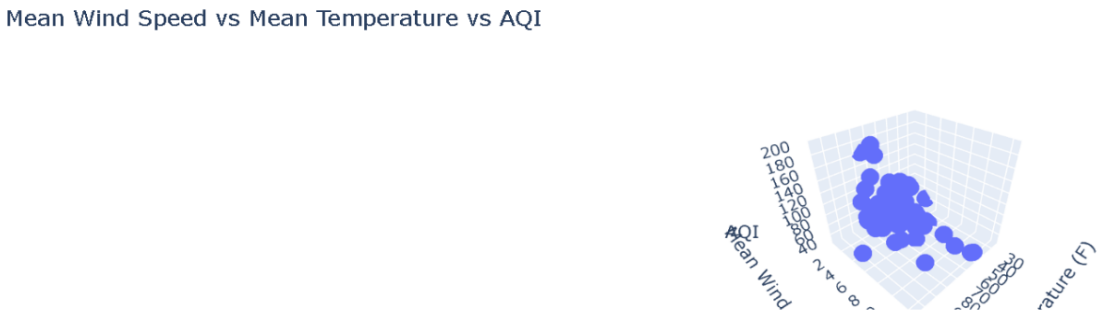
AQI Dataset

Section 107

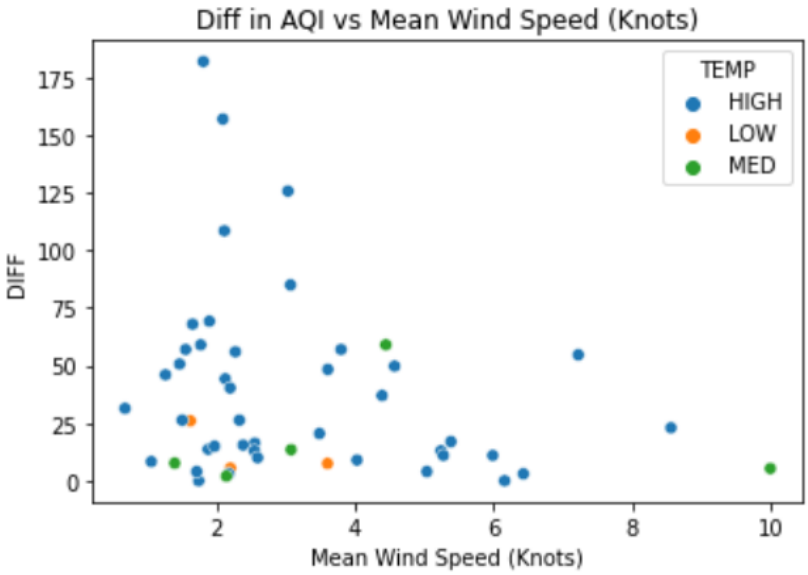
Momo Siu, Nathan McKay, Nathan Kelsh, Emre Kusakci

Open-Ended EDA

For our open-ended EDA, we speculated that there was some relationship between windspeed, temperature, and AQI.



We plotted a 3D graph of the relationship between mean wind speed, mean temperature, and AQI. We found that there might be some correlation between the variables because higher AQI levels appeared to be correlated with higher temperature and wind speed.



We also plotted the difference in AQI max and mean values by the mean wind speed and color-coded high, medium, and low temperature points to observe if there is any correlation between wind speed and AQI. This plot appears to have values that are clustered near the origin, which made us curious to further investigate the relationship between wind speed and AQI.

Questions about Data:

Are there other variables besides temperature and wind speed that can help predict air quality?
Does geography or altitude affect AQI?
Do the amount and type of air particles change throughout the year (across different months)?
Is this data generalizable to other states, or even other areas of the world?
Are any measurements related to natural occurrences/disasters?

Problem

The hypothesis we wanted to evaluate was that temperature and observed AQI data for counties could be used to predict the AQI levels of neighboring counties. Neighboring counties are defined as counties within two degrees of latitude or one degree of longitude from the observed county. To prove or disprove this hypothesis, we first attempted to establish a correlation between temperature and AQI in our dataset.

Answer

Reject: Temperature and observed county AQI data cannot be used to accurately predict AQI for neighboring counties. Our linear model that was trained first on local data and second on neighboring county AQI data generally had predictions that were off by an amount that could be considered significant on the AQI measurement scale. Even when attempting to adjust for certain parameters (wind, latitude, longitude), it seemed that our model was not accurate enough to predict AQI measurements for neighboring counties.

Modeling

We trained a linear regression model using temperature, wind speed, AADT (a traffic measurement used earlier in the project), the day of the year, and mean NO_2 as the inputs. Initially, we found that the actual AQI values and our predicted AQI values seemed to follow a concave path. When we applied the natural log function to each of our features, however, we achieved a smaller RMSE and cross-validation error, and a more linear relationship between temperature and AQI. The outputs of this model were the predicted AQIs for a county, given these values. We created a range of typical temperatures, and for each individual temperature, we selected the other values uniformly at random from ranges of typical values for them, respectively. The motivation for this was to determine if there was a strong linear relationship between temperature and predicted AQI, which might tell us if temperature had an impact on the AQI of neighboring counties. If there was a strong linear relationship between temperature and AQI, then the scatter plot of temperature against predicted AQI should have had a clear linear relationship, despite the randomization of the other features. Regression seemed an appropriate model for predicting AQI, as earlier visualizations involving wind speed and

temperature during exploratory data analysis seemed to be promising. Once we established a relationship between temperature and AQI for observed counties, we then used latitude and longitude measurements for each county to identify their neighboring counties. Then, we would use temperature and observe AQI measurements to predict AQI levels for the neighboring counties. To achieve this, we also used linear regression to validate our model by comparing our predicted AQI values of neighboring counties to their actual values.

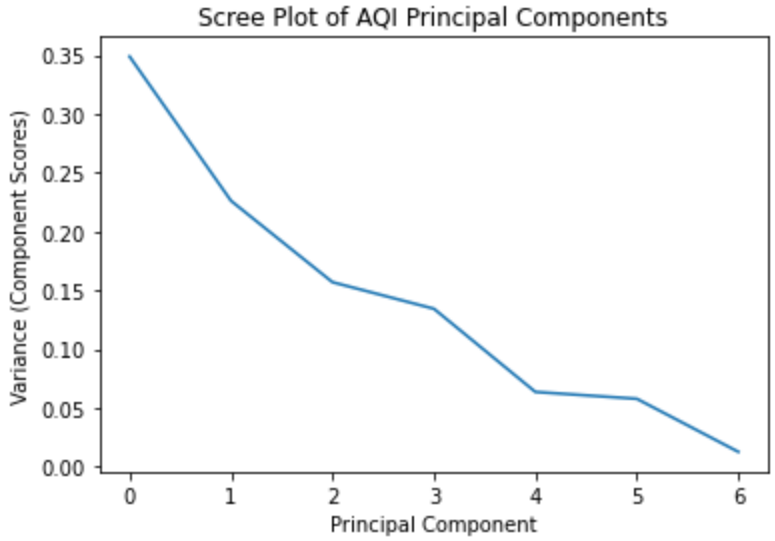
Model Evaluation and Analysis

In our first attempt at creating our model, we considered 3 features to predict AQI: wind speed, temperature, and day of the year. We used linear regression to try to predict AQI values based on these features. This model did not work too well, as we ended up with a cross-validation error of 33.93364372280335 and an RMSE of 39.83816189980789, which is a relatively high amount of error for our linear prediction model.

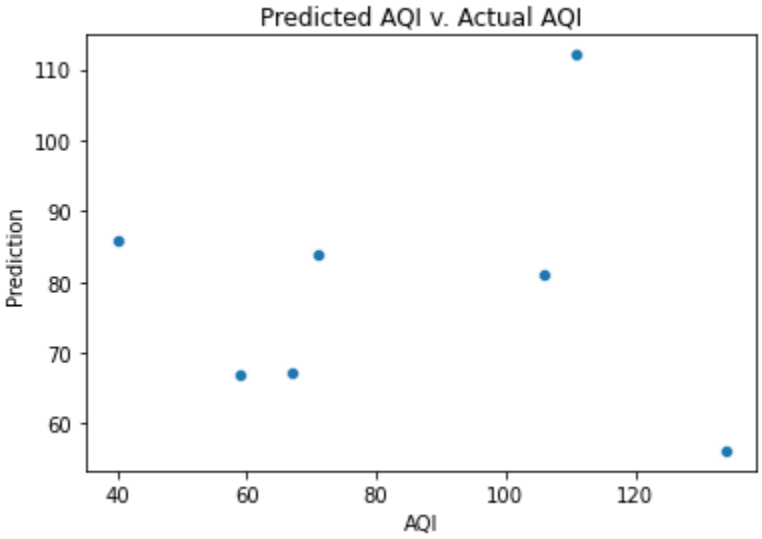
Taking information from our first model into account, we attempted to add more features to try to improve our model. Specifically, we chose to add data about NO_2 measurements and, AADT (traffic data). latitude, and longitude. The original model only considered temperature, wind speed, and day of the year. However, this caused both our cross-validation error and our RMSE to increase, and it more than doubled the number of features in our model. The results of our model are shown below:

```
Cross-Validation Error:  40.9781070552291
RMSE Error:  35.87185313075339
Sigma Values:  [12.30571786  9.90642808  8.25185372  7.63418849  5.25055958  5.00367942
  2.33514098]
```

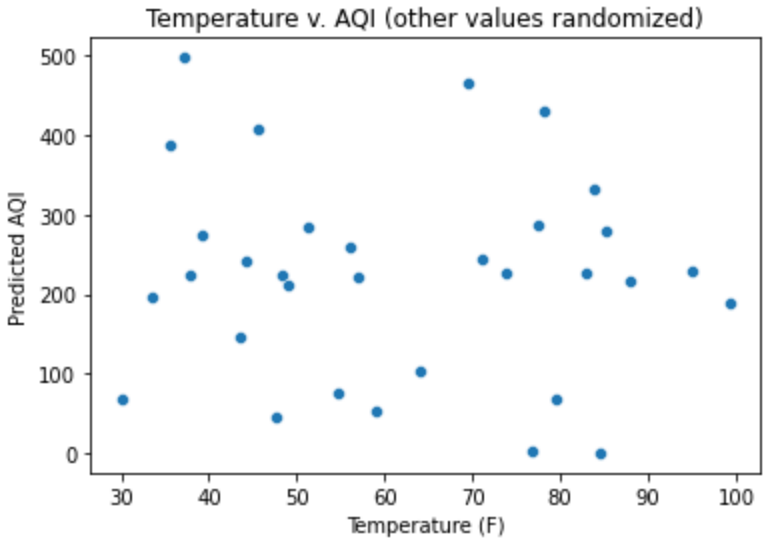
This result is undesirable because the RMSE and cross-validation errors were even higher than in our initial model. Because the predictions are off by this much, on average, it would not be uncommon for the predicted AQI category to be wrong. The singular values, however, looked promising, as their gradual decline indicated that our model was dropping in variance of our AQI predictions.



To try to improve our model even further, we attempted to perform PCA to single out the features most important to the model. Instead, however, we discovered that all features were fairly important for predictions, as there was no significant dropoff in the variance accounted for by each component, so this improvement failed.



This plot shows that there is no pattern between predicted and actual AQI values, meaning that the model is not predicting correctly. Ideally, the relationship between them would be roughly linear.



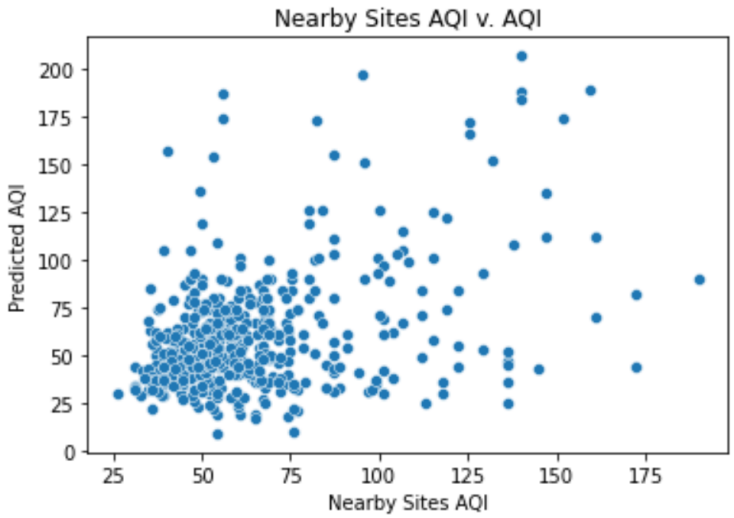
On a scatter plot of temperature against predicted AQI, we hoped to identify some linear pattern. However, it turns out that temperature alone was not enough to predict AQI, which is why when the other values were randomized there appears to be no pattern for our model.

Model Improvement

To create an improved model, we took two different approaches. In the first approach we used the same features as the original, but added a new feature nearby sites AQI. The function within_2_degrees takes a site number and returns an array of all the nearby sites within 2 degrees of longitude or latitude of the site. We used this function to create a new column avg_aqi_of_nearby in the merged df. The avg_aqi_of_nearby is the average AQI of all the nearby sites on the day of the specified row.

We had to filter out a lot of the rows however since many of the sites had no nearby sites. Even with the smaller dataframe however, the new model performed much better. This makes sense since the nearby sites AQI are very likely to be related to a site's AQI as seen by the decreased validation errors in the new model.

```
Cross-Validation Error:  23.734351782045188
RMSE Error:  21.130799204301393
Sigma Values:  [28.86391898 21.36399049 18.2519014  12.70126712]
```

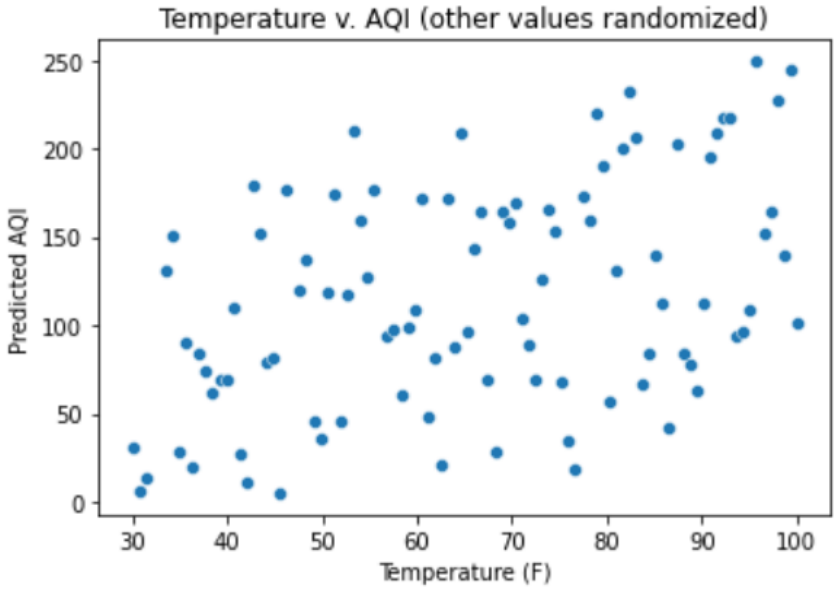


In the graph of Nearby Sites AQI vs Predicted AQI, we can see that there is a moderately strong linear relationship between the two. The points are mostly clustered around the line y = x, which makes sense since the AQI of a site is probably close to the average of its nearby sites.

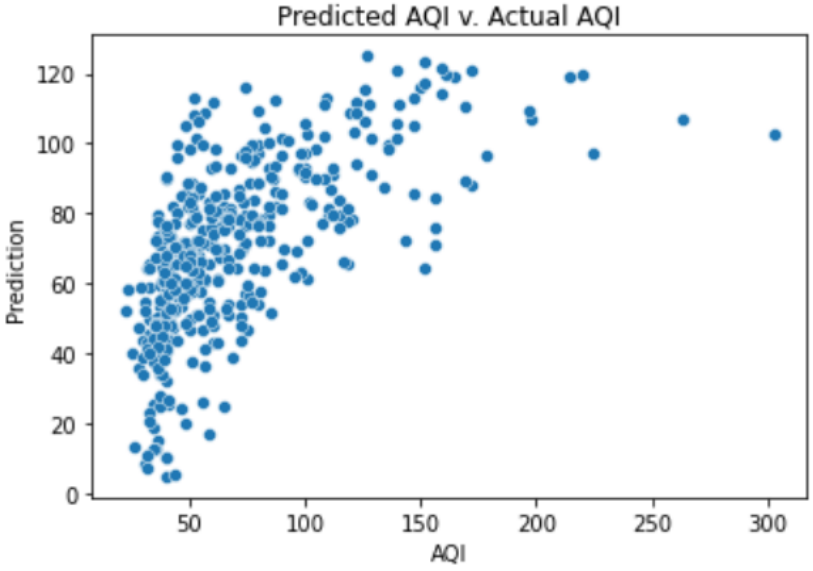
In our second approach, we took the natural log of each of our features (except longitude and latitude) in our dataset in order to try to achieve some linear relationship between temperature and AQI within our dataset. Upon doing this, we successfully dropped both our cross-validation and RMSE errors significantly.

```
Cross-Validation Error:  30.57010616418818
RMSE Error:  30.939762258802965
Sigma Values:  [87.04047632 81.28850544 64.32444697 57.27014929 43.34194596 37.6053499
 13.6727199 ]
```

However, this result is still somewhat undesirable because the cross-validation and RMSE error magnitudes are still relatively high. However, the declining sigma values in the array show us that as we add more features to our model, we are successfully dropping off some of the variance from our AQI predictions.

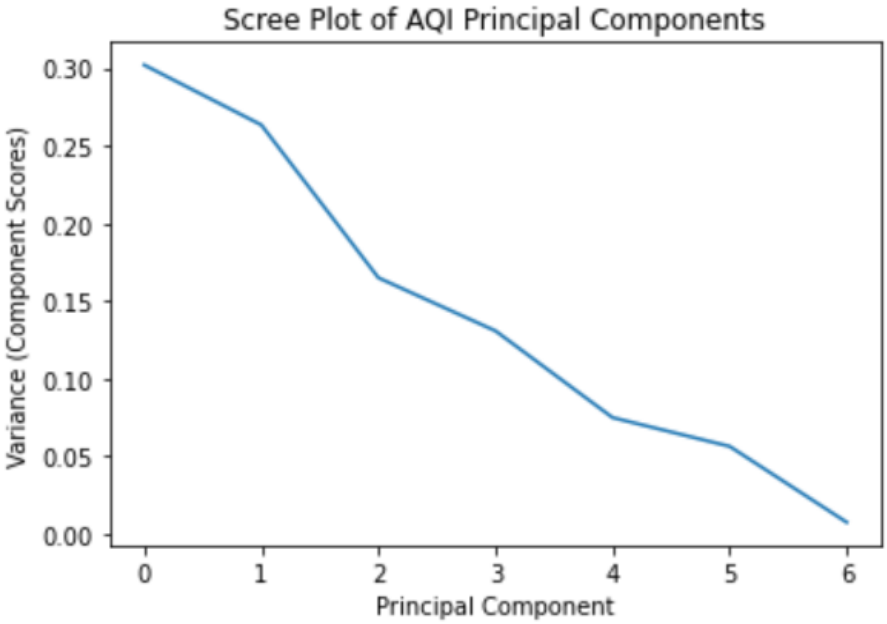


The graph above shows the relationship between temperature (in fahrenheit) and our predicted AQI points after we transformed our features by taking the natural log of them. There appears to be a faintly detectable linear relationship between temperature and AQI, which is an improvement from our previous modeling but not enough to conclude that temperature has any direct correlation to the AQI level.

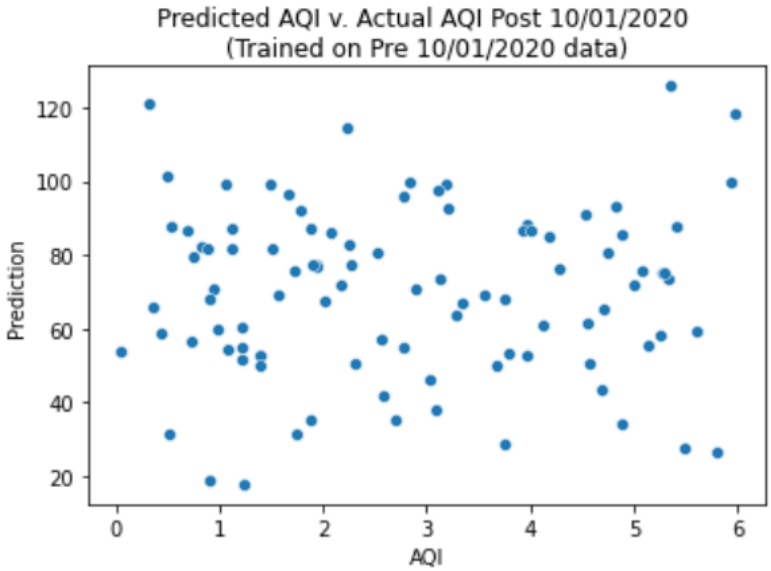


This plot shows that there is somewhat of a relationship between predicted AQI values and actual AQI values, especially around the 0-120 unit range. Within this range, datapoints appear to be highly clustered, and as actual AQI values increase, predicted AQI values appear to increase as well. However, beyond the 150 unit threshold, there appears to be little to no correlation between actual AQI and predicted AQI values. As actual AQI values increase

beyond the 150 unit mark, predicted AQI values do not move past the 120 unit mark. We, therefore, cannot conclude that our model is highly accurate, but there appears to be some faint correlation between actual and predicted AQI.



Then, we performed PCA again on our features to determine if there was any improvement in the variance by transforming our features. We found that our model performed pretty similar to our previous model, as there is still no significant dropoff when we added new features, which tells us that each of our features are equally important in reducing the variance of our model.



Future Work

It may be better to consider a different model, such as a decision tree, or perhaps consider more than temperature as a way to predict AQI. As demonstrated by our PCA, the number of variables that have a role in determining AQI is large, so our attempt to use a single variable to capture the whole relationship was a failure. This could reduce error in the model, and allow for accurate predictions of the AQI of neighboring counties. On another note, issues with table merges only allowed for a relatively small number of usable data points for our model (this can be seen by the very sparse scatter plot above). Improving our data would allow for more training, and possibly lead to more accuracy as well.