Final Report - Air Quality Index

Luis Navarrete Rios, Ishika Prashar, and Lynda Solis Chavez

Akshara Majjiga

Data 100 - Section 114

13 December 2021

(Optional Open Ended EDA)

We performed a principal component analysis (PCA) utilizing chemical concentrations as our components split amongst monthly data. Specifically, we focused on the chemical concentrations of ozone, SO_2, CO, and NO_2. Our resulting PCA clustering:

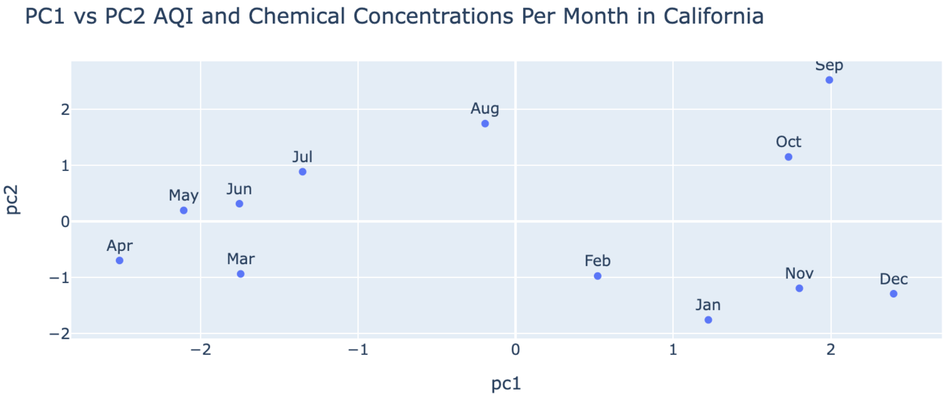


Figure 1. PCA clustered data

Looking through the clusters, we could clearly see that large values for PC1 represent the colder climate months and small values for PC1 represent the warmer temperatures. We also noticed that the colder months of fall seem to have larger SO_2, CO, and NO_2 values. PC2 was slightly more difficult to interpret in this case as the clusters were not very clearly defined. Our guess is that larger PC2 values are associated with larger AQIs and respectively smaller PC2 values with lower AQI levels. The scree plot was used as a form of visualization for the magnitude of variability associated with the individual components obtained from performing PCA. From this plot, we were able to determine the results from our analysis were useful because the scree plot below shows that PC1 and PC2 capture more than 80 % of all variance. This allows us to ignore the rest of the PC's without losing any valuable information.

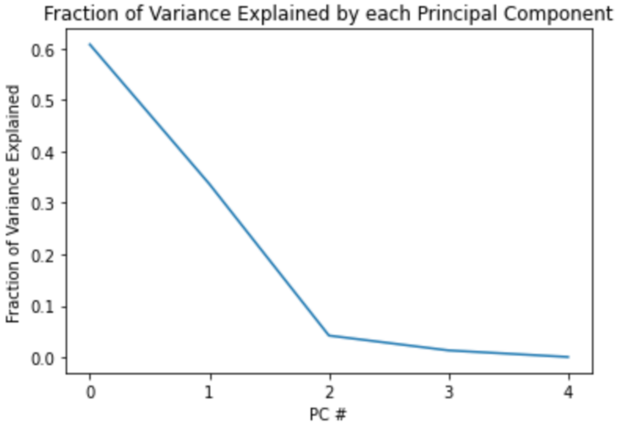


Figure 2. Scree plot representing fraction of variance amongst each principle component

To further study the pattern noticed of rising chemical concentrations and AQI during colder months we plotted the following graph using the normalized dataset from above to demonstrate that there does indeed seem to be an increase of AQI and and chemical concentrations in colder climate months, specifically the months in the fall season.

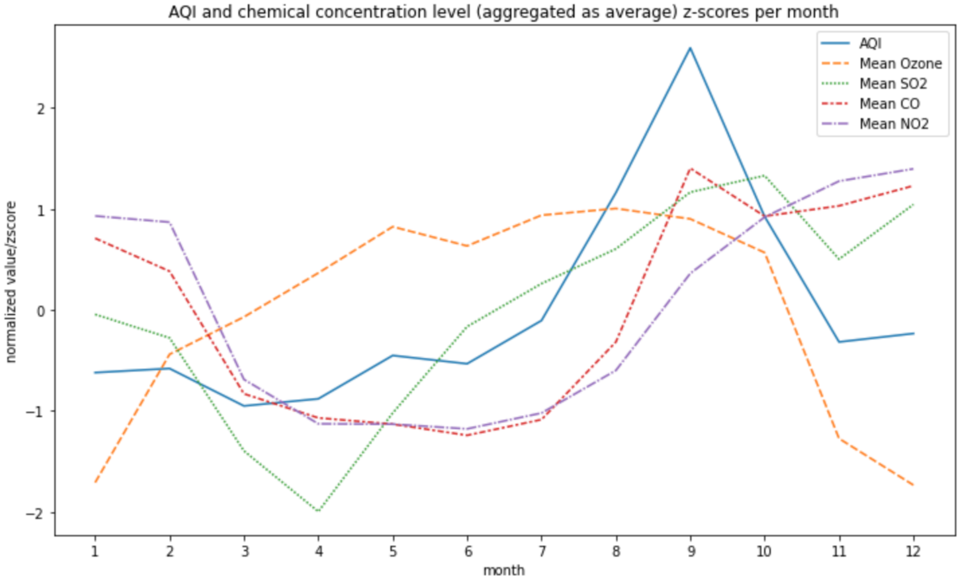


Figure 3: Line plot representing the AQI and chemical concentration z-scores per month

Figure 4 demonstrates that and confirms our findings that the fall months seem to have much greater AQI levels overall.

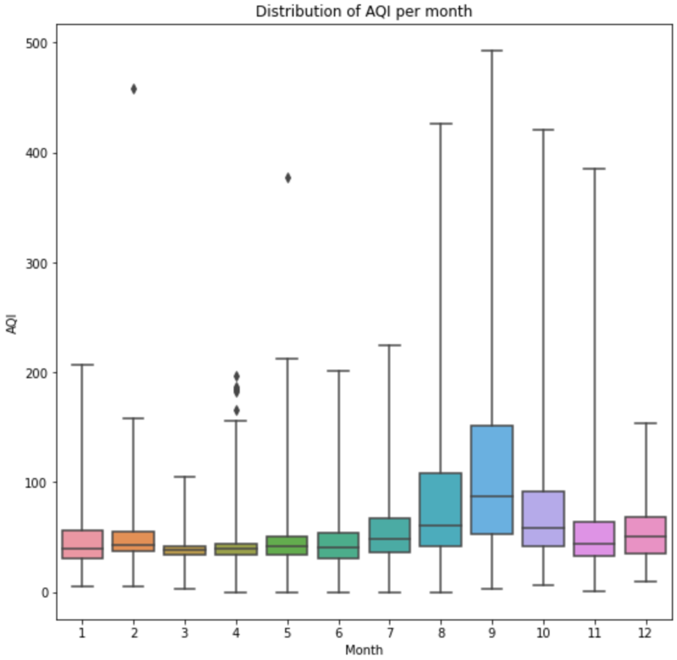


Figure 4: Whisker plots representing distribution of AQI (capped at 500) per month

Problem/Hypothesis

From our open ended EDA results (as well as the association found between temperature and AQI as shown in Figure 6) we propose the following hypothesis: AQI levels are positively correlated with increasing chemical concentrations of SO_2, CO, and NO_2 and increasing temperature. However, we have learned that chemical concentrations are a direct predictor of AQI and thus not a valuable set of features to utilize when it comes to predicting AQI. We still believe that these chemical concentrations are highly predictive and important, so we plan to use sulfur dioxide as well as carbon monoxide in our improved model below. Additionally, we find that traffic may be an important predictor of AQI as increased traffic levels cause increased pollution. Considering that temperature seems to have a linear association with AQI, we believe that other weather variables such as precipitation may also be highly predictive and linearly correlated with AQI. Thus we will be utilizing two external datasets of traffic and precipitation to generate two alternative features for our model. To summarize, our improved and final hypothesis (utilized in the improved model section of this report) is the following: AQI levels are positively correlated with carbon monoxide levels, sulfur dioxide levels, traffic, temperature, and precipitation. We expect that running a linear regression with features defined above will be highly predictive in finding AQI values. We also suspect to see a R^2value that is positive and close to 1 to represent the positive linear relationship we suspect is present.

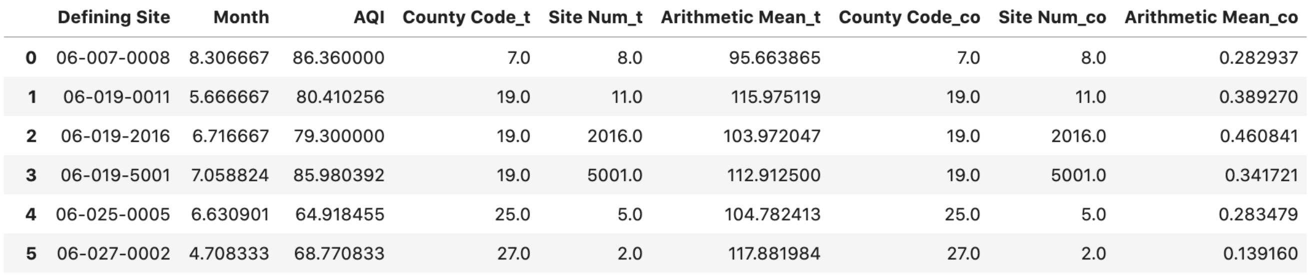
3 Feature Model

Initial Workflow

From our hypothesis and initial EDA we decided to use the following three features to discover the type of relationship they have for our initial model analysis: average carbon monoxide concentration by defining site, average temperature by defining site, and the month in which the data was collected. We decided to use only one chemical concentration as it was shared in question 10 of the project that AQI is directly based on chemical concentrations so there would be too high of a correlation. Instead of multiple chemical concentrations, we determined month would be a significant feature given initial EDA showed that the months of fall seemed to show an increase in AQI. We believe that the temporality of the data collected holds great significance in determining AQI.

We began by cleaning the data and focusing on columns that were of importance for our analysis. We included the county code and site code to use as keys for merging and grouping DataFrames. Additionally, we used the columns to create a defining site column for those tables that did not include it. We noticed that some rows did not have values for latitude/longitude or site code so we removed those rows. In order to keep our data consistent, we decided to get rid of outlier AQI values. Our team chose any value greater than 500 to be unrealistic in California by AQI standards.

To summarize, our model inputs/features initially included CO concentration, temperature, and month. Our model output/target is the AQI value.



	Defining Site	Month	AQI	County Code_t	Site Num_t	Arithmetic Mean_t	County Code_co	Site Num_co	Arithmetic Mean_co
0	06-007-0008	8.306667	86.360000	7.0	8.0	95.663865	7.0	8.0	0.282937
1	06-019-0011	5.666667	80.410256	19.0	11.0	115.975119	19.0	11.0	0.389270
2	06-019-2016	6.716667	79.300000	19.0	2016.0	103.972047	19.0	2016.0	0.460841
3	06-019-5001	7.058824	85.980392	19.0	5001.0	112.912500	19.0	5001.0	0.341721
4	06-025-0005	6.630901	64.918455	25.0	5.0	104.782413	25.0	5.0	0.283479
5	06-027-0002	4.708333	68.770833	27.0	2.0	117.881984	27.0	2.0	0.139160

Figure 5: Initial feature table

Model

With the three selected features, we decided to use a linear regression model to predict the level of AQI. We justify the usage of a linear regression model throughout initial EDA, as there was a positive linear correlation between the temperature arithmetic mean and AQI.

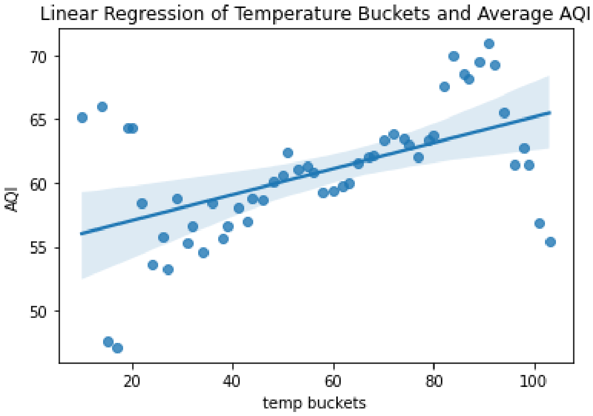


Figure 6: Scatter plot comparing temperature (aggregated to buckets) and AQI value

We also performed a PCA which provided insight on chemical concentrations and the possibility that the months of fall seemed to be associated with higher chemical concentration amounts and higher AQI values, suggesting some sort of positive correlation.

We split our data into a testing set that contained 70% of our original data and a validation set that contained 30% of the remaining original data. Using the SKLearn library, we fit our training data to a linear regression model and obtain the accuracy as well as a validation error through the use of a five-fold cross-validation error. Additionally, we tested the training and test error through root-mean-square error (RMSE).

Model Analysis

Our training accuracy was found to be approximately 14.92% and RMSE approximately 39.05. The test accuracy was found to be approximately 10.26% and RMSE approximately 37.81. Finally, we performed five-fold cross-validation of the training data set and found the misclassification error rate of approximately 39.09. This score implied that our first model did not have an accurate prediction of AQI with the three features used. The accuracy is very low, at ten percent it seems that what the model did end up predicting correctly may have just been by random chance. The RMSE value also seems high considering that AQI levels were capped at 500 and the average square root difference between the residuals hints that there was great variability between actual and predicted RMSE values. It follows that there does not seem to be a linear relationship between our selected features and AQI values. Our extremely high RMSE values further tell us that our model is not a good predictor of AQI and our features may not necessarily have a linear association with AQI.

As shown in Figure 7, our cross-validation predicted values were not quite accurate with the actual AQI values. We find from this plot that the current model vastly underestimated the true AQI values.

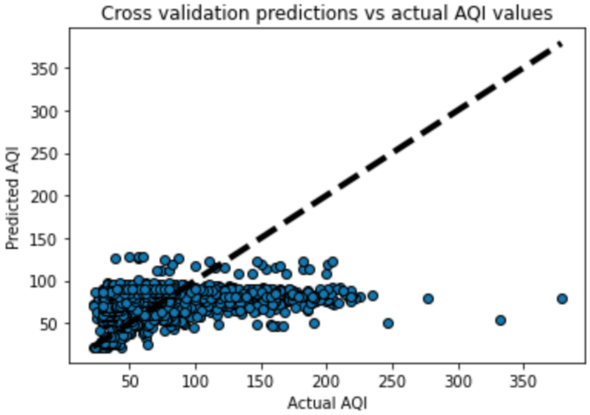


Figure 7: Relationship between actual and CV predicted AQI

From this analysis, it seems we would have to reject our null hypothesis that AQI values are positively correlated with the increase of chemical concentration of carbon monoxide, month, and the temperature.

We also find that our R^2 value on the training set is very close to 0, at approximately 0.15. Since our model did not do a good job of accounting for variance on AQI, our regression did a very poor job of fitting the model. There is a very weak linear association between our features and the AQI target.

We also plotted residuals (Figure 8) and found that the values are not randomly dispersed across the horizontal $y = 0$line, meaning that a linear model was not the optimal selection of model for our described features. A nonlinear model would be needed to find an association between the three features and AQI.

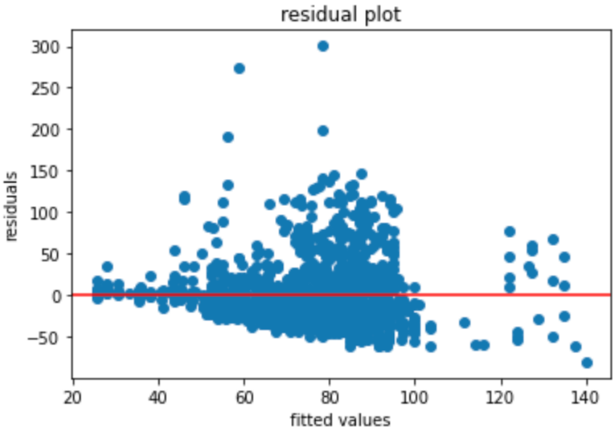


Figure 8: Residual plot demonstrating a non-random relationship for 3-feature model

Problems addressed with the Model & Improvements

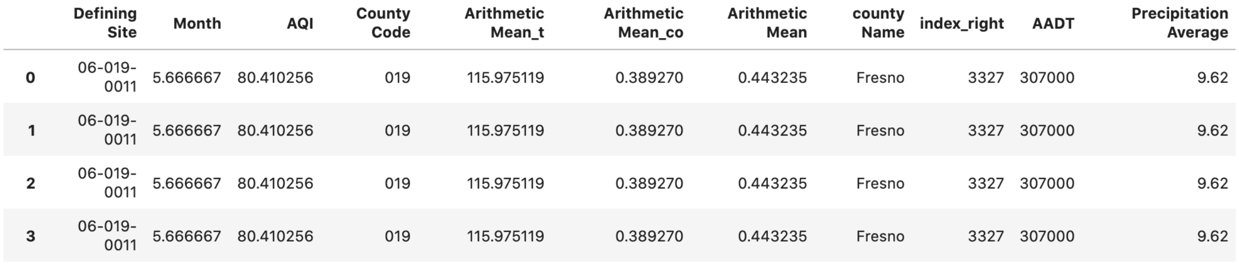
To reiterate, we obtained a training accuracy of 14.92% on our initial model that predicted AQI using linear regression with the month, carbon monoxide, and temperature as the features. Additionally, we obtained high cross-validation and training errors. Being that the accuracy was low, we could see that our model did not accurately represent AQI alone, resulting in an underfitted model. The model contained too few features so it is important to note our model had high bias and low variance because the accuracy was low and could not accurately predict the target variable. To improve our model accuracy, we decided to increase our feature selection.

To increase our feature selection, we researched what variables could potentially have the largest impact on AQI. Our prediction that month, carbon monoxide, and temperature were important features but alone, did not represent AQI enough. Although we could not use all chemical concentrations, we determined that it would be beneficial to use one more chemical concentration. In our hypothesis, we analyzed that a combination of all chemical concentrations seemed to be greatest correlated with AQI. Thus, we chose to add sulfur dioxide to our model as we had determined from our initial EDA and PCA that it seemed to change drastically by month. We also believed that traffic could be a big indicator of AQI without directly using chemical concentrations due to pollutants released by cars. Through research, we found that carbon monoxide and sulfur oxides are released through the burning of fuels, so we believed traffic data would be a feature that would be an asset to our already selected chemical concentrations. We used the traffic data introduced in part 1 of this project to add a traffic feature of Annual Average Daily Traffic (AADT). Additionally, we decided to add precipitation data into our model to further improve accuracy as precipitation decreases the pollutants in the air. Again, we believed this feature would fit nicely in our model as it is a feature that still has a relation with chemical concentrations and seems to fit the linear association we are looking for.

Improved Model

Workflow

Once we found features that could potentially increase our accuracy, we moved on to further clean our data and add the additional features. We merged the data with our previous three feature dataframe, using the Defining Site as a key. Then, we proceeded with an external dataset, California traffic data for the year 2020. We cleaned the data to find the AADT per longitudes and latitudes. Since our traffic data contained longitudes and latitudes that differed by some decimal points from our feature dataset, we used the GeoDataFrame library in order to join the feature DataFrame with the nearest latitude/longitude location within the traffic dataset. We continued by grouping the merged features by the Defining Site. Additionally, we used a second external dataset, the yearly average precipitation in California per county. From our research, we found that precipitation positively affects AQI. After cleaning the precipitation dataset and merging on county code, we created our feature matrix “multiple_features" that contained the following features: month, arithmetic mean for temperature, arithmetic mean for carbon monoxide per defining site, arithmetic mean for sulfur dioxide per defining site, average annual daily traffic, and the average annual precipitation per county.



	Defining Site	Month	AQI	County Code	Arithmetic Mean_t	Arithmetic Mean_co	Arithmetic Mean	county Name	index_right	AADT	Precipitation Average
0	06-019-0011	5.666667	80.410256	019	115.975119	0.389270	0.443235	Fresno	3327	307000	9.62
1	06-019-0011	5.666667	80.410256	019	115.975119	0.389270	0.443235	Fresno	3327	307000	9.62
2	06-019-0011	5.666667	80.410256	019	115.975119	0.389270	0.443235	Fresno	3327	307000	9.62
3	06-019-0011	5.666667	80.410256	019	115.975119	0.389270	0.443235	Fresno	3327	307000	9.62

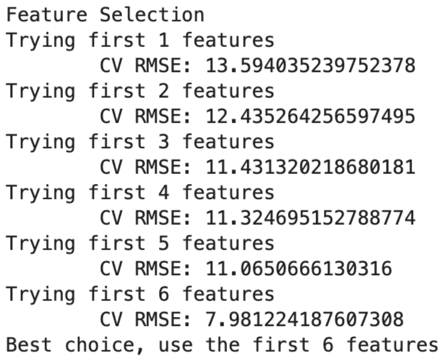
Figure 9: Improved feature table

Again, we fit our linear regression model with the previously mentioned features and used AQI as our targeted output.

Model Analysis & Evaluation

Our training accuracy was found to be approximately 66.99% and RMSE approximately 7.85. The test accuracy was found to be approximately 71.1% and RMSE approximately 7.02. Finally, we performed five-fold cross-validation of the training data set and found the misclassification error rate of approximately 7.98 a big improvement from our CV error from our model that used only three features. We find that our RMSE has decreased significantly from our baseline model suggesting that the difference between actual and predicted values has decreased, and our accuracy also suggests the same. Given that AQI values are capped at 500, the RMSE values all ranging around 7 is quite good as the residuals, or difference in predicted versus actual AQI values is not significantly high. Although we would have wanted to see higher accuracy than just around 66-71%, we believe that this is a stronger linear correlation compared to the baseline and definitely not just due to random chance.

We find that our model was much more accurate with the additional features and we were also able to reduce our errors. Since we had a total of six features, we thought it would be beneficial to compute the cross validation error to determine how many of the first N (up to six) features would give us the lowest misclassification error rate. We find from our analysis that all six features when used together provide the lowest RMSE as shown in Figure 10 so no further analysis was necessary.



```
Feature Selection
Trying first 1 features
        CV RMSE: 13.594035239752378
Trying first 2 features
        CV RMSE: 12.435264256597495
Trying first 3 features
        CV RMSE: 11.431320218680181
Trying first 4 features
        CV RMSE: 11.324695152788774
Trying first 5 features
        CV RMSE: 11.0650666130316
Trying first 6 features
        CV RMSE: 7.981224187607308
Best choice, use the first 6 features
```

Figure 10: Cross Validation Feature Selection Analysis

 We also find, as shown in Figure 11, that our cross-validation predictions are much more accurate compared to our three feature model. This further shows that even though our accuracy is not extremely close to 100%, we have improved our model significantly and find that it does a fair job at predicting accurate AQI values.

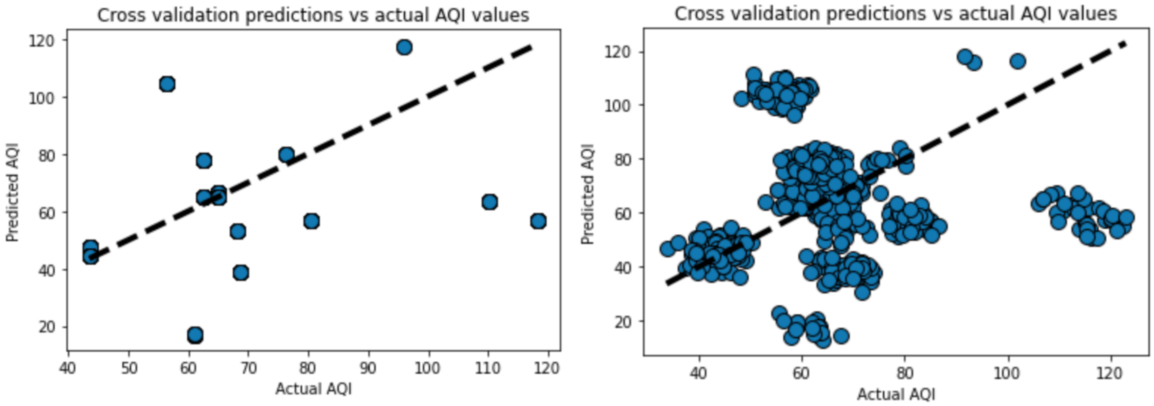


Figure 11: Relationships between actual and CV predicted AQI (without and with noise, respectively)

Note that there is significant overplotting occurring here as the total number of observations for this model is 862 values. Random noise was added to give a better estimate on the true number of predicted values.

We also find that our R^2 value on the training set is approximately 0.66 which is much closer to 1 and improved significantly from our three feature model. This value is also positive which means there is a positive linear association. This means that our regression does a much better job of fitting the model because our model does do a good job of accounting for variance on AQI. There is a good to a strong linear association between our features and the AQI target. Thus, we find that there does indeed seem to be some linear correlation between CO, SO_2, temperature, month, traffic, and precipitation. However, it is important to note that a large R^2 value isn't always good. There is a chance there could be covariates in our features or linearly dependent features that cause us to overfit our model.

We also plotted residuals (Figure 12) and found that the values are randomly dispersed across the y=0 line meaning that a linear model was a good fit for this regression. (Note that there is again overplotting in this Figure as stated with Figure 11 so random noise is added to get a better estimate on the true distribution.

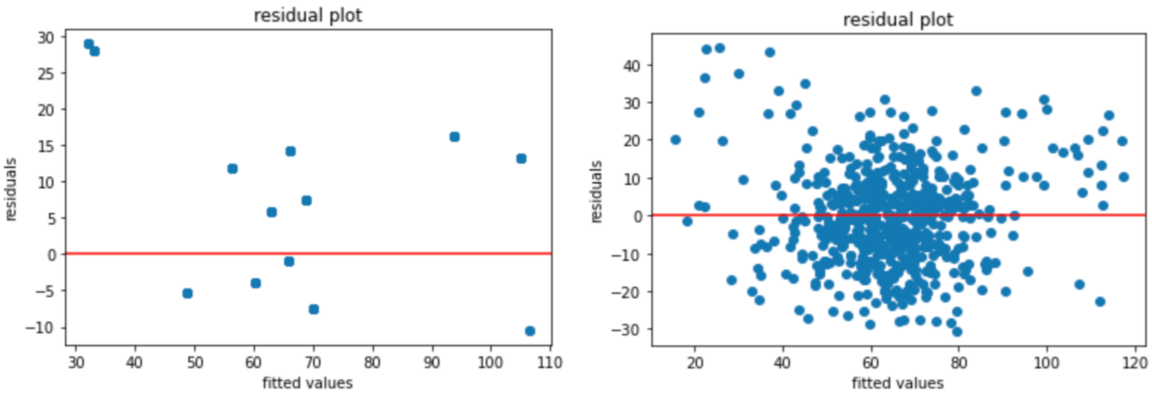


Figure 12: Residual plots demonstrating a non-random relationship for improved model (without and with noise, respectively)

Overall, we see a drastic improvement in our model after the addition of three features. Figure 13 shows the drastic decrease in RMSE with our improved model. We also find that our model is much more generalizable as the cross-validation error also decreased. Another interesting find across both the three feature model and the improved model is that the RMSE slightly decreased on the test set compared to the train or CV set, further explaining the generalizability.

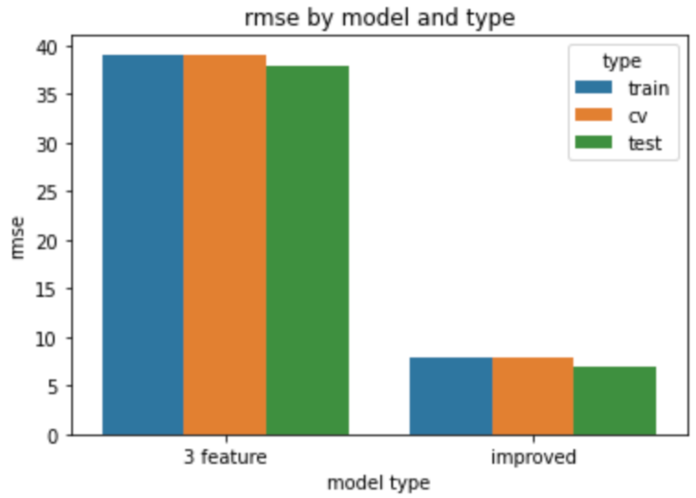


Figure 13: Bar plot demonstrating RMSE errors for baseline and improved model by dataset type

Limitations

We do acknowledge that both our models required a great deal of data loss to be utilizable. We found many NaN values, duplicates, and inconsistencies in the data that had to be removed. Daily analysis and county analysis were difficult to conduct as days reported for in the datasets were inconsistent as well as the counties reported for. A monthly analysis could not be done as well because that would create only twelve data points, which would not allow us to properly assess our model. Additionally, having to merge many different dataframes caused us to generalize our model to avoid running out of memory in the notebook. We also predict there is a possibility of collinearity in our features since traffic could be correlated with carbon monoxide levels or temperature correlated with precipitation, for example.

Discussion & Feedback

We believe that we did successfully answer our hypothesis given that our improved model accuracy improved drastically from the baseline. Though we could not use all chemical concentrations simultaneously, traffic and precipitation resulted in valuable indirect measures of chemical particulates to further help predict AQI levels. It seems that from this model we can draw the conclusion that AQI values have a positive linear correlation with weather, and pollutants. To provide further detail, temperature and precipitation were our weather values. Carbon monoxide, sulfur dioxide, and traffic as an indirect measure of pollution were our variables for pollutants. Our improved model increased accuracy, decreased RMSE, and had a positive R^2 value close to 1. Further analysis needs to be done on the impact of these features to further improve the model and aim for even higher accuracy.

To amend our hypothesis and add further complexity to it, it would be interesting to add a county-wide analysis to test whether county income and racial demographics of the county have an effect on the overall AQI. The amended hypothesis would be that AQI levels are linearly correlated with increasing chemical concentrations of sulfur dioxide, carbon monoxide, increasing temperature, income, and racial status/prevalence of racism. We would form some sort of numerical value for race and racism in order to run a linear regression. This would be an interesting approach due to socioeconomic disparities and environmental racism that is prevalent within the United States, specifically California. This approach would still be relevant to our current hypothesis that chemical concentrations have a positive correlation with AQI because we could test whether it is true that being in a less affluent, non-white neighborhood has an impact on the chemical concentrations in the air, and therefore the AQI of that county. We also find from research that many refineries and pollutant releases occur in cities and counties that are low income and minority based, so this could be an interesting approach to take for further work and analysis on this topic.

References

CDCR Population COVID-19 Tracking - CDCR Population COVID-19 Tracking. (2021, January 28). Retrieved December 4, 2021, from https://data.ca.gov/dataset/cdcr-population-covid-19-tracking/resource/5a3f496d-04be-4405-aea0-e83e26076efc

NOAA National Centers for Environmental information. (2021, December). Climate at a Glance: County Mapping. Retrieved December 4, 2021, from https://www.ncdc.noaa.gov/cag/county/mapping/4/pcp/202012/12/value