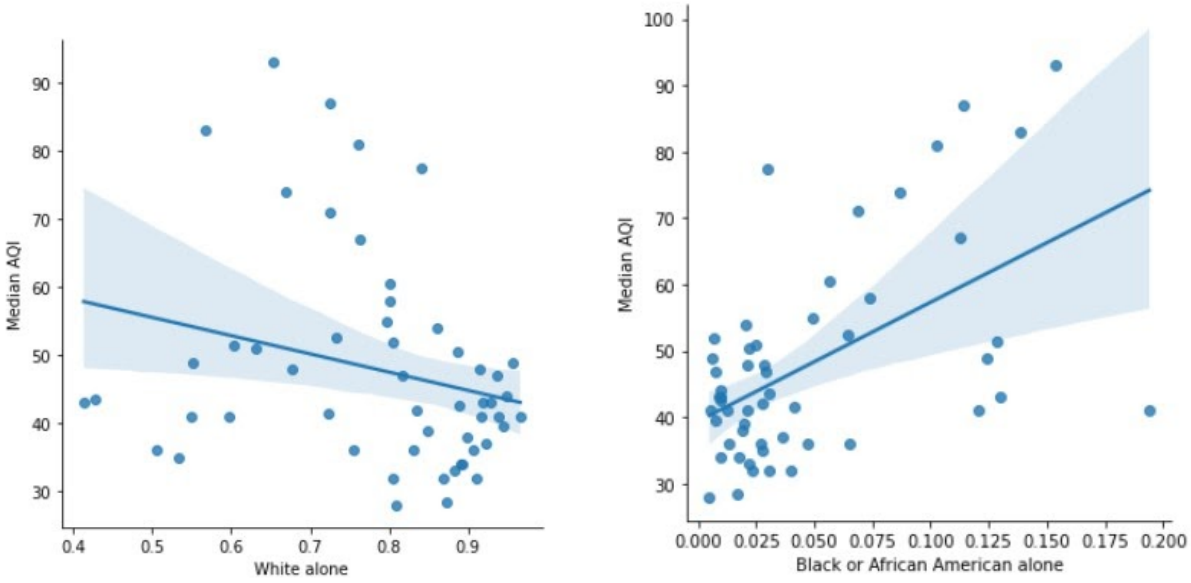
Data 100 Final Report (AQI)

Contributors: Andrew Chen, Casey Lei, Young Woong Min, Shoumik Chaudhuri

Open Ended EDA

In terms of exploratory data analysis, we tried to visualize how the median API and other variables were related with county racial makeup. As such, we utilized Python's Seaborn library and Matplotlib plot to visualize this relationship.

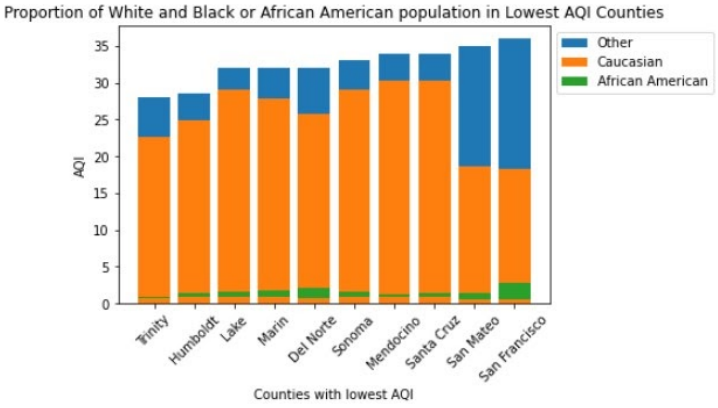
For our first plot, we are displaying the relationships between the White population density and the median AQI of counties in California, then for African-American population density and median AQI. We observe a **relatively linear relationship between the racial population densities and median AQI**, as well as a tight **cluster of data points in the range of 0 to 0.05 for the proportion of each county that is Black/African American** and in the range of 30 to 55 for Median AQI. As such, we can observe that our plot has large variance as our data fans out as we increase Median AQI or the Proportion of Black/African Americans. However, we still can observe an overall positive correlation between Median AQI and Black or African American proportion.

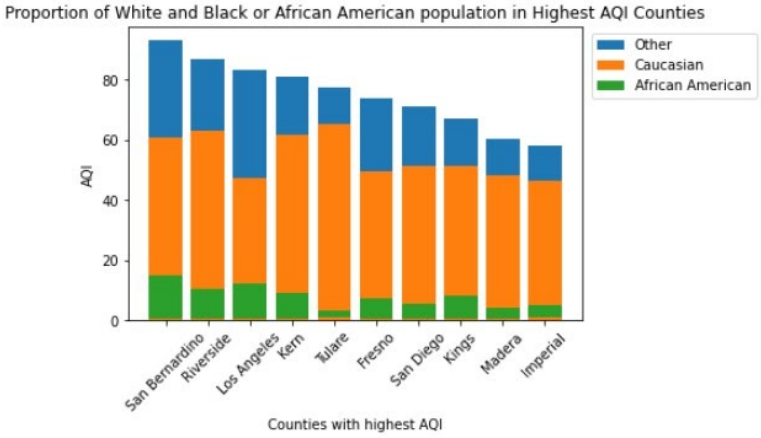


When visualizing the relationship between White Proportion and Median AQI, we find **key insights** regarding the distribution of races in California. First, the range of values for the Proportion of Whites (0.4 to 1.0) is drastically different from most other races (i.e. 0.0 to 0.2 for the Proportion of Blacks/African Americans). This tells us that not only are most counties in California actually predominantly white, but **the African-American population is proportionally negligible in some areas**. In other words, when performing analyses, the size of our regions may render the predictor of African-American population density so small that there

is little variation in this feature among counties and may make it more difficult to confirm our hypothesis. In this visualization, we can observe a clustering of values towards the larger proportions. Most importantly, there appears to be a negative correlation between Proportion of Whites and Median AQI. This seems to indicate that there is some additional analysis that could be done on how racial makeup is related with Air Quality, which is related to the overall theme of the racial environmental justice movement.

For our second plot, we used a stacked bar chart to compare the racial makeup of the counties with the Top 10 Median AQI and Bottom 10 Median AQI. The bar chart has the Counties with Lowest/Highest AQI on the x-axis and distribution of AQI on the y-axis by proportion of race. Looking at the top 10 highest Median AQI counties, we observe that there is a much larger proportion of African Americans relative to the figure with the top 10 lowest Median AQI counties. The range of the proportion of African-Americans in the counties with the Top 10 Highest Median AQI are greater than the proportion of African-Americans in the counties with the Bottom 10 Median AQI Counties. We can observe this just by looking at what proportion of each bar is green. As such, there appears to be some relationship between African American proportion and the AQI. As such, we hoped to explore whether some quantifiable relationship exists between African American proportion and AQI and whether we can create an accurate model that utilizes racial demographics as a feature in our project.





In performing our initial analysis, we felt that **further open-ended EDA** could be conducted by reexamining the relationship between POC population density and various indicators of pollution and environmental wellness, such as median temperature, AQI, and even proximity to hazardous materials factories by separating geographic locations by whether they were historically redlined or not. This would target the question of whether the legacy of redlining has continued to subject people of color to unequal living situations even in modern day.

Problem

Our hypothesis is that there is a strong positive correlation between the proportion of African Americans and the median AQI of a region, and we focused our analysis on data from California. In other words, we believe that incorporating the proportion of African-Americans as a feature of our multiple linear regression model that predicts AQI will improve the RMSE and R2 value of that model.

We can explicitly confirm the first part of our hypothesis by calculating the Pearson correlation (correlation coefficient) to confirm that there exists a positive correlation between the two variables. For Median AQI and African-American proportion, there is a correlation coefficient of 0.537, confirming the first part of our hypothesis. For the second part of our hypothesis, we can confirm the predictive capability of African Americans as a feature in our model through k-fold cross-validation and also comparing metrics like RMSE and R2 across different models.

We used ***two external datasets*** for our inputs: one gives the proportion of races (specifically African-American) in each county in California, the second gives counties divided into Census regions of California, which we used in determining our data granularity. Looking at our models, we can confirm our hypothesis in principle. Our sample size consists of 58 California counties (some dropped due to unavailability of racial demographic/ozone data). Some may argue that our sample size is too small for determining the predictive power of African American proportion on region AQI. However, with access to data from all 3,006 United States counties, we could create a model from that larger dataset and confidently confirm our hypothesis. This larger dataset is available through the U.S. Census. Another potential solution to our lack of data would be to bootstrap the sample of 53 data points. Since the bootstrapping technique is independent of distribution, we could conduct hypothesis testing and assess the underlying distribution characteristics of our California data.

Our hypothesis looks at the relationship between race and AQI levels by geographic regions, and we were able to use a U.S. Census dataset containing racial demographics for every county in California. If we expand our scope to include counties throughout the United States, we would be able to find relevant racial demographic data off the Census website. We believe race and AQI are correlated; specifically, we believe that the proportion of African-Americans in a geographic area is positively correlated with the region's median AQI. In order to confirm this, we can look

at the correlation coefficient R2 of our baseline model, which does not include any racial demographics data as a feature, then compare it to the R2 once African-American population density is included. We can also look at RMSE and k-fold cross-validation R2. If R2 increases in either case and loss decreases, we can say that there is a positive correlation between African-American population density and median AQI in a region, confirming our hypothesis.

Answer

After our modeling and hypothesis testing, we confirm our hypothesis. For Median AQI and African-American proportion, there is a correlation coefficient of 0.537, meaning that there is a moderate, positive relationship between these variables. Upon examining our *optimal model*, which included all baseline model features in addition to using African-American population density as a predictor of median region AQI, we found that the R2 of our *optimal model* was slightly higher than the *improved model* (which included the baseline model features with Days of Unhealthy for Sensitive Groups replacing Very Unhealthy Days), increasing from 0.8839 to 0.8872, and our RMSE was slightly lower (5.6810 vs 5.7627). However, our cross-validation R2 nearly doubled from 0.2933 in our *improved model* to 0.5128 and the Negative RMSE increased from -8.96 to -7.7 in our *optimal model*. Due to these improvements, we confirm that African-American population density in a region is positively correlated with the area's median AQI.

Modeling

We used the supervised method of several multiple linear regressions to featurize a variety of factors affecting median AQI and to specifically examine the predictive ability of race demographics. With regard to our hypothesis, multiple linear regression modeling makes the most sense to utilize since we want to model the impact of a variety of factors like the percentage of BIPOC individuals on AQI. Utilizing scipy-learn, we created a model to featurize our factors. For categorical variables like region and median household income, we would utilize One-Hot Encoding. In the case of median household income, this would mean that we would need to merge the table with additional information on median county income by introducing columns which represent each mutually exclusive bin and assign 0s and 1s for each data point (0 if the data point does not fall into that bin, 1 if the data point does fall into that category).

Our *baseline model* utilizes the features Region, Median Ozone PPM per County, and Number of Very Unhealthy Days to predict AQI. Region comes from an external dataset of California counties classified by the U.S. Census. The Number of Very Unhealthy Days refers to the annual number of days described as having extremely unhealthy air quality for each county. In our *optimal model* we utilize Region, Median Ozone PPM per County, Number of Days Unhealthy for Sensitive Groups, and African-American population density in each county in California.

As for why we decided to utilize these features, in Part 1, we noticed a linear relationship in our open-ended EDA between median AQI and White and African-American population densities. Thus, we decided to look into this relationship between race and AQI, comparing between several multiple linear regression models with different racial demographic inputs (African-American, White, and American Indian).

Since we are looking at counties as data points, we wanted to include a broader categorical geographic input and chose California regions. Additionally, we utilized medians throughout the project when looking at ozone levels and AQI because it takes into account extreme outliers that may be due to California wildfires. We wanted to exclude these outliers in order to focus on our factor of interest, which was race and its influence on a region's normal day-to-day AQI.

Model Evaluation and Analysis

In terms of evaluating our model's performance, we first split the dataset through a 70/30 train-test split with a random seed of 42 for future reproducibility. We utilized the metrics of R2 (which checks how well fit the data is to our regression) and root mean squared error (which checks the spread of the residuals and standard error of the residuals). After creating our initial model as a reference point, we compared our newer models based on these metrics. Specifically, we utilized 5-fold cross-validation on our training data and averaged the RMSE and R2 across these splits as a metric and also examined the RMSE and R2 of our test data. We experimented with a train-test split ranging from 80/20 to 50/50. However, we settled on 70/30 because we realized there was a trade-off for large train-test splits like 80/20 versus more even splits like 50/50 because our dataset was fairly small. For instance, with a large train-test split, the cross validation set was very large (80% of the total data set). However, the test-set was very small so the RMSE and R2 would have a lot of variance. As for more even train-test splits like 50/50, the test-set was larger so the RMSE and R2 had less variance but the cross validation set was very small (50% of the data set). As such, this resulted in large variance within the k-folds validation since each fold contained a very few data points. With a larger data set (the data set of all U.S. counties), our model evaluation and validation would be significantly stronger since it could help account for this variance that is a result of a small dataset.

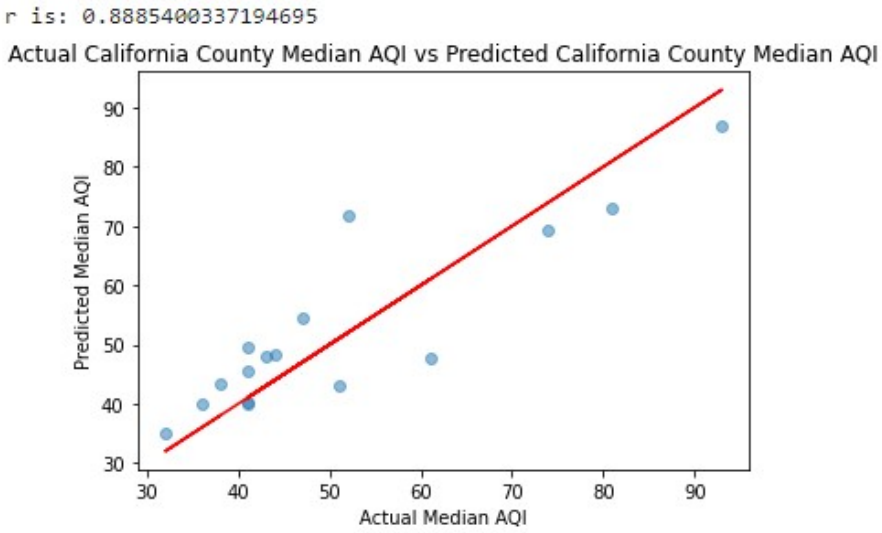
In terms of evaluating the efficacy of a model, we compared improved models to our *baseline model's* RMSE and R2 score (7.894 and 0.782 respectively). By using this "naive" approach, we could determine whether future models were effective relative to this baseline. Below is a figure of all of our models and their RMSE and R2 from the test-train split and cross-validation which we will repeatedly reference during our "Model Improvement" section:

```
R2, MSE, RMSE (random state = 42)

Features: Median Ozone PPM, Region, Very Unhealthy Days
R2 Score: 0.7821670890022685
Mean Squared Error: 62.327441659225926
Root Mean Squared Error: 7.8947730593872

1) Features: Median Ozone PPM, Region, Unhealthy for Sensitive Groups (Change 1)
Model 1 R2 Score: 0.8839337435746579
Mean Squared Error: 33.209457619701006
Root Mean Squared Error: 5.762764754846497

2) Features: Median Ozone PPM, Region, Unhealthy for Sensitive Groups, Black
Model 2 R2 Score: 0.8872024922209899
Mean Squared Error: 32.27418691326927
Root Mean Squared Error: 5.681037485641973

3) Features: Median Ozone PPM, Region, White, Black, Unhealthy for Sensitive Groups
Model 3 R2 Score: 0.8991358693338377
Mean Squared Error: 28.85974938685569
Root Mean Squared Error: 5.372127082158024

4) Features: Median Ozone PPM, Region, Black, White, American Indian, Unhealthy for Sensitive Groups
Model 4 R2 Score: 0.860821926365825
Mean Squared Error: 39.822326318578305
Root Mean Squared Error: 6.310493349856119

Cross Validation (random state = 42)

Initial Model
Cross Validation Negated Root Mean Squared Error Score: -11.109334655215342
Cross Validation r2 Score: 0.0786720720418694

1) Model 1 (Modification 1)
Cross Validation Negated Root Mean Squared Error Score: -8.961237807872227
Cross Validation r2 Score: 0.29332819961372275

Modification 2
2) Model 2
Cross Validation Negated Root Mean Squared Error Score: -7.703691632389868
Cross Validation r2 Score: 0.5128870698995398

3) Model 3
Cross Validation Negated Root Mean Squared Error Score: -7.964879409018662
Cross Validation r2 Score: 0.43685331701597774

4) Model 4
Cross Validation Negated Root Mean Squared Error Score: -8.235106763239681
Cross Validation r2 Score: 0.42742354015672934
```

In terms of visualizations, we evaluated the model via residual plots, looking for no pattern amongst the data and also an even spread across the x-axis and y-axis (residuals). If our residual plot follows the aforementioned requirements, we can conclude that the data points roughly have a linear relationship and that our model is not biased. Additionally, we also utilized a scatter plot of actual AQI v. predicted AQI. Because a perfect model would predict all values such that actual AQI equals predicted AQI, we also decided to plot a y = x line to determine whether our model was systemically overestimating or underestimating the AQI. Additionally, the scatterplot

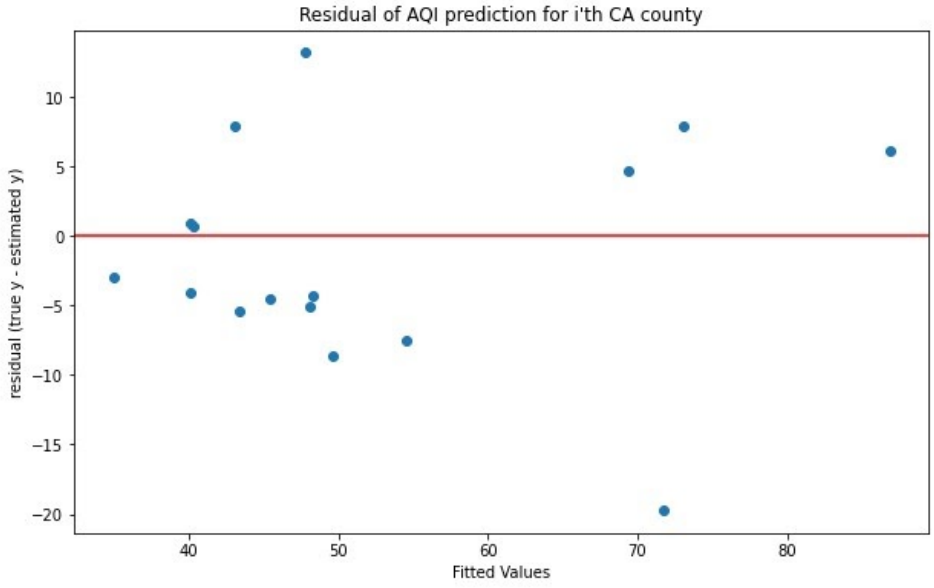
allows us to observe how well fit the data points to a perfect model, giving us a better understanding of the predictive capability of our model.

***Baseline Model* Figure 1:**



Our first plot of the *baseline model* is a scatter plot of Actual Median AQI v. Predicted Median AQI on the test dataset. Our dataset does have a higher density of values in lower median AQI, so the train-test split on this data may affect model accuracy in that higher median AQI values will not be predicted as accurately. Our *baseline model* makes no perfect predictions and tends to underestimate higher median AQI values and overestimate the lower ones. This will motivate our next improvement.

***Baseline Model* Figure 2:**



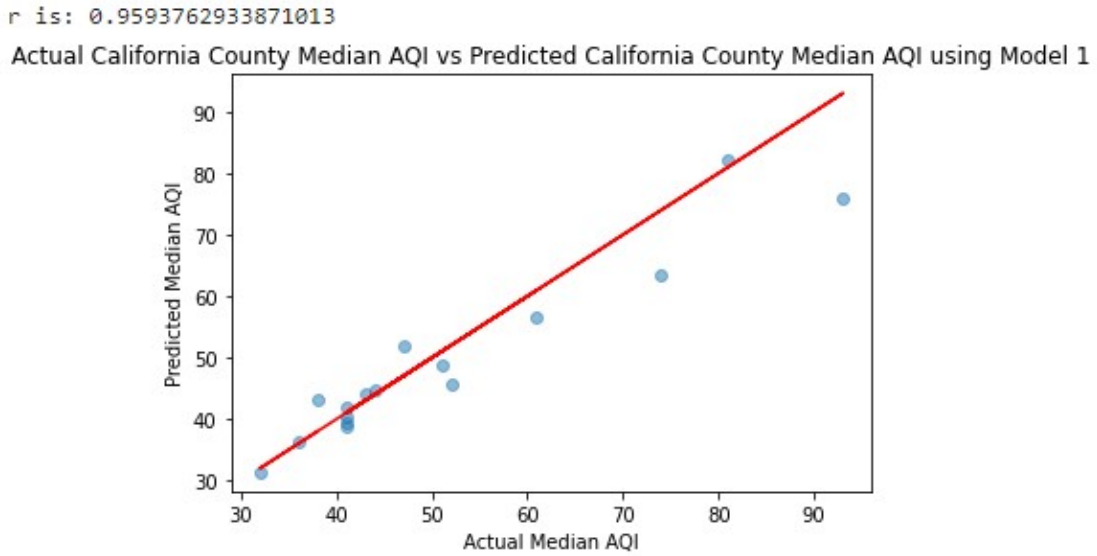
This residual plot shows no pattern in the residuals for our model, which tells us that the linear regression model has no glaring bias. We do see greater residuals for larger median AQI, once again likely because of the concentration of data in the lower half of the range of median AQI that allows the model to be trained to better predict lower median AQI's. In improving our model, our goal should be to make sure that the residual plot does not develop any new patterns and that the plot's points are not densely packed in any region.

Model Improvement

Improvement 1:

Initially, we used the annual number of very unhealthy air quality days as a predictor of median AQI as a feature in our *baseline model* because we believed it would be a strong indicator of worse air quality. However, "very unhealthy days" is an extreme variable that could be affected by factors other than institutional problems like racial demographic background. For instance, wildfires will heavily skew the number of very unhealthy days, muddling the predictive power of our model. Thus, our **first improvement** was switching out this feature for the more moderate "Number of Days Unhealthy for Sensitive Groups." This improved our model's performance, with RMSE improving from 7.8948 to 5.7628 and R2 improving from 0.7821 in the *baseline model* to 0.8839 in the *improved model.* Intuitively, we believe that this improvement in both RMSE and R2 was due to the less extreme feature of "Number of Days Unhealthy for Sensitive Groups" better accounted for the everyday AQI v. edge case AQI. Like the example mentioned before, a day of extreme wind or wildfires could potentially skew the predictive capability of our model. As such, with this improvement, our model better accounts for institutional impacts of AQI like the "Proportion of African Americans", which will be useful in determining the predictive power of racial demographics for AQI.

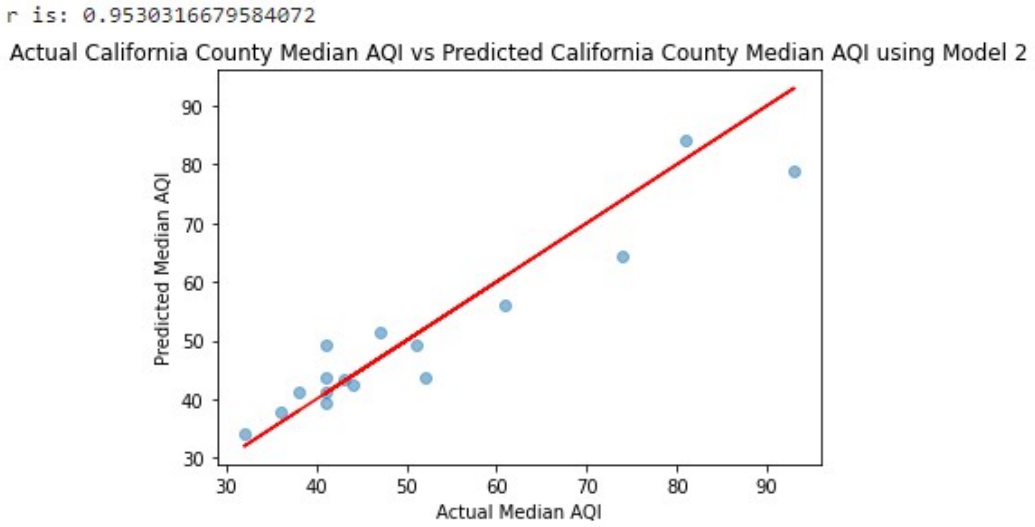
***Improved Model* Figure 1:**



Our improved model is another scatter plot of Actual Median AQI v. Predicted Median AQI, and we can see a much higher accuracy in median AQI predictions across all actual median AQI values. The model still predicts lower AQI's more accurately, but this is once again intrinsic to the dataset that we conducted the train-test split on. Were we to use a larger dataset, for example one containing data from throughout the country, this could likely be adjusted.

Improvement 2:

Optimal Model Figure 1 and 2:



In testing for an improved model, we built three candidate models, each with more racial demographic data as our **second improvement**: first with just African-American population density as a feature, then the addition of with White proportion, and finally with the addition of American Indian proportion. Loss varied and R2 gradually decreased with the addition of each racial group on top of African-American proportion. The introduction of just African-American population density greatly increased our predictive power, as seen in our k-fold cross validation scores, but the addition of more population densities decreased predictive power, which we concluded was due to overfitting. As such, there appears to be a trade-off between additional racial demographic features and AQI. Ultimately, we also looked for the highest k-fold cross validation score (0.5128 in the *optimal model* vs. 0.2933 in the first improved model), which also decreased with the addition of multiple racial features, to determine our *optimal model* was the *improved model* with the addition of African-American racial data.

Future Work

In the future, we would expand the scope of our data to include counties from all 50 states. Instead of using Census classification regions, we would sort regions by proportion of BIPOC to improve model accuracy. One possible flaw we found in our work is that the regions given by the U.S. Census are based upon population density. The African-American population in California

is very small, so the proportion of African-Americans in each region is rendered similarly negligible and therefore does not differentiate among the counties well enough for us to examine the relationship between race and AQI well. Stratifying by BIPOC population proportion would allow us to better look specifically at the legacy of redlining in the U.S, which was a major motivation behind our hypothesis. This could help identify target locations for environmental racial justice efforts. Environmental justice activists have called attention to the fact that marginalized populations (such as BIPOC peoples) tend to experience a disproportionate burden of the consequences of pollution and climate change. This is due to a number of factors, such as the lack of adequate pollutant disposals in urban areas, low land values in neighborhoods with significant Black and Hispanic populations, and the leniency shown toward corporations who violate environmental laws in Black communities. If we are able to confirm the relationships between poor environmental conditions and racial demographics, we will be able to better target policy change and advocacy in an organized and systematic way that maximizes impact on these marginalized communities.