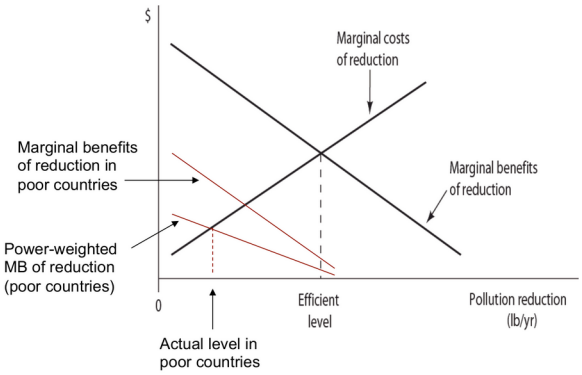
Will Chan

Final Report
Air Pollution and Socio-Economic Factors

Context: The power weighted social decision rule is a economic concept that explains how more powerful sociopolitical groups are less impacted by negative externalities such as environment and health, than less powerful groups; which points to the conclusion that socio-political factors can dictate whether groups of individuals will have less environmental protections or increased levels of environmental degradation such as air pollution or poor water quality.

Valuing human health risks

Purchasing power + political power combined



Above is a graph that uses the power weighted social decision rule to demonstrate how more powerful groups perceive efficiency in polluting to less powerful groups.
Motivated by this concept, this project focuses on air pollution, particularly pollutant, PM10.
PM10 is an airborne particulate matter that can cause adverse health effects such as respiratory heart and lung damage, and it can also have adverse environmental impacts, such as altering plant growth and yield, soiling of materials, and impacting water quality and clarity. PM10 is not only emitted by natural sources such as trees and vegetation but it is also created from anthropogenic sources such as industrial processes and motor vehicle exhaust. Following these concerns this project focuses on PM10 pollution levels of US counties.

Objective: To predict the 2019 annual air pollution, PM10, of a county using that counties social-economic factors as well as that counties past pollution readings from 2017-2018. The social-economic factors being observed include county demographics, poverty and income levels, unemployment and employment rate, education attainment levels, and vehicle totals. Last, the observations are inclusive of 208 counties throughout the US.

Data Collection: Data for this project was collected from various sources and then aggregated into one dataset.

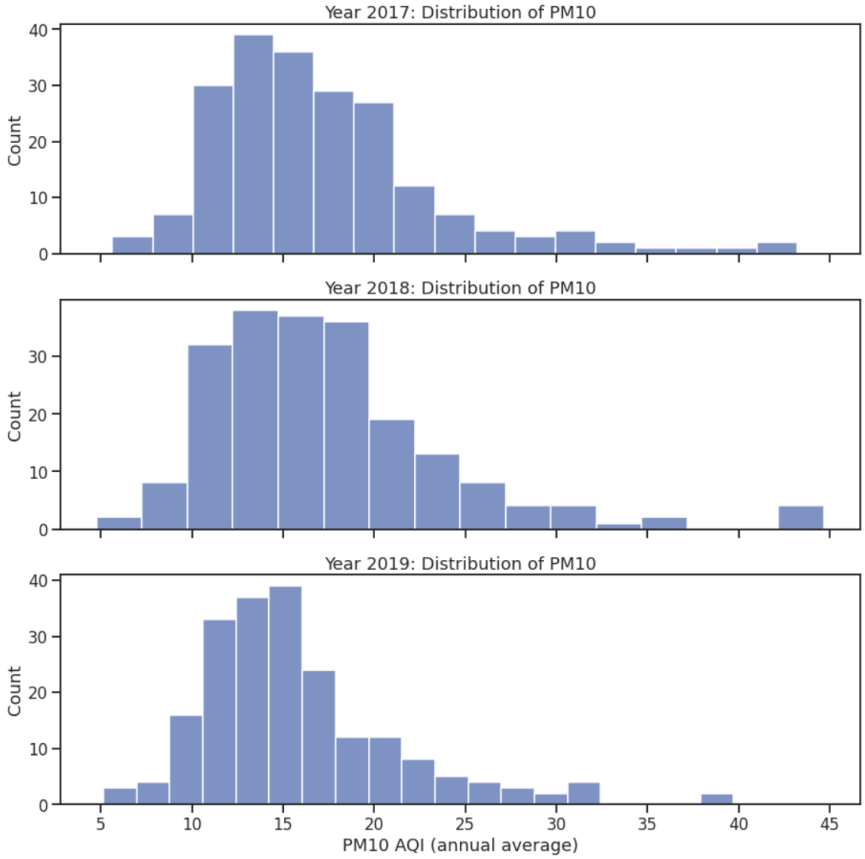
- Epa.gov
  - Daily PM10 AQI

        - Air Quality System (AQS) monitors
- Data.census.gov
    - County Characteristics Resident Population Estimates
        - Age, Sex, and Race
    - SAIPE (Small Area Income and Poverty Estimates Program)
        - Poverty, Median Household Incomes
    - ACS1 American Community Survey)
        - Education attainment levels
        - Total Vehicles
- Data.ers.usda.gov
    - Economic Research Service
    - County Unemployment and Median Household Incomes
        - Civilian Labor Force, Employed, Unemployed, Unemployment rate
- Bea.gov
    - Bureau of Economics Analysis
    - County Annual Income (CAINC1):
        - County Personal Income, Population, Per Capita Personal Income

Data Wrangling: Starting by looking at the PM10 dataset from Epa.gov. This dataset had a time frequency of incurring daily observations and this would become a problem in merging with the other datasets collected which were of annual frequency. Because of this, the PM10 Daily Data was converted into PM10 annual data; so that the dataset consisted of each counties annual PM10 AQI levels for that year (note usage of mean average and mode average did not signify drastic differences); the mean average was used for this transformation. This transformation led observation count to go from around 16000 observations to 340 observations. This transformation was PM10 daily data from 2017 to 2019 and afterwards these transformed datasets were merged together by their GeoFIPS.

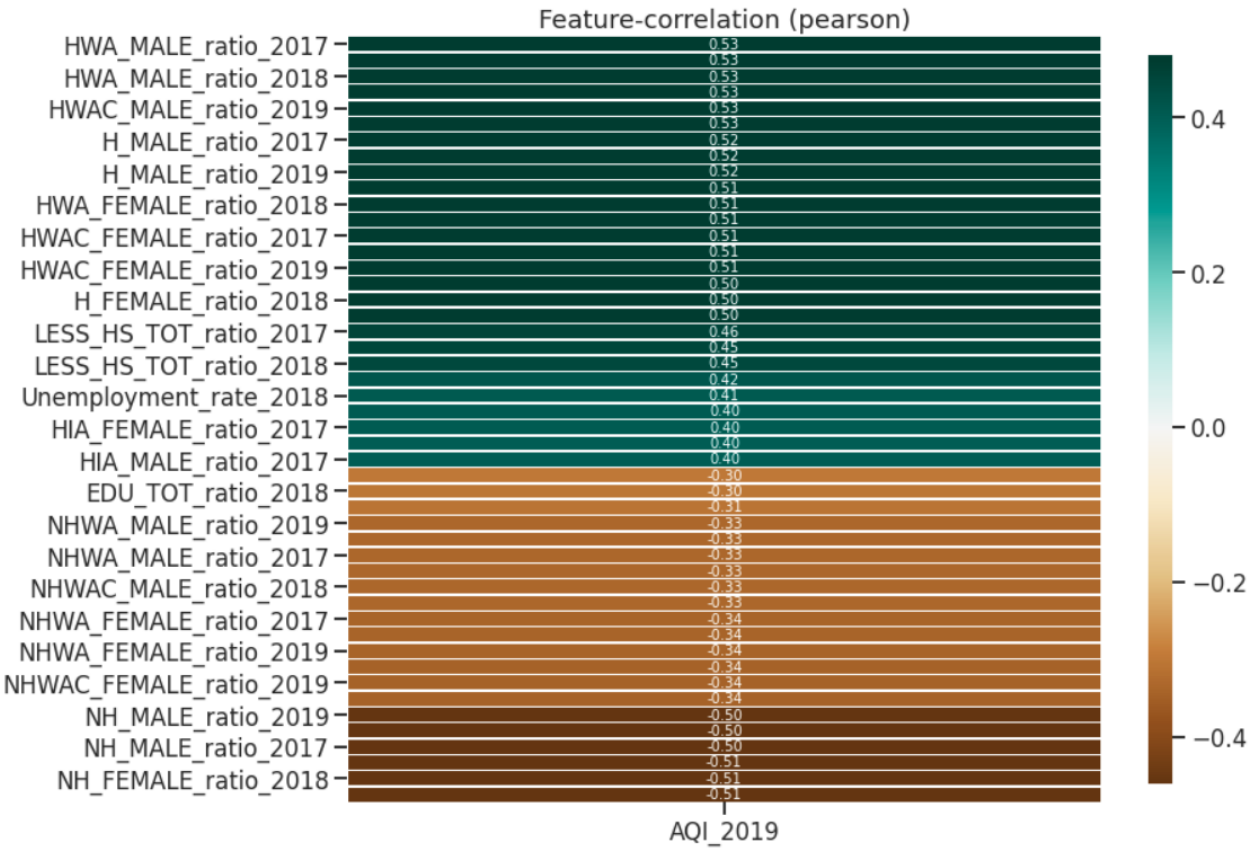
Afterwards the remaining datasets that are listed above are merged to the annual PM10 dataset. This is done by creating a GeoFIPS attribute for each social political dataset and then merging all datasets bythis GeoFIPS. Because some of the dataset do not have the same range of county data, the merging of datasets led to loss of some observations. The final dataset resulted with 208 observations with each year having 567 columns (feature engineering was also performed to obtain more columns to analyze).

Exploratory Data Analysis: Graphs and analysis will be provided below.



Above is graphing that displays the counts of annual PM10 levels from years 2017 to 2019. Graphing the distribution of annual PM10 datasets from years 2017 to 2019 assisted in determining whether there were any major fluctuations throughout the years. From above it appears that were no major fluctuations. The majority of PM10 levels appear to range from 10-20 annually with a right skew in distribution; the max is often around 40 and min is often around 5.

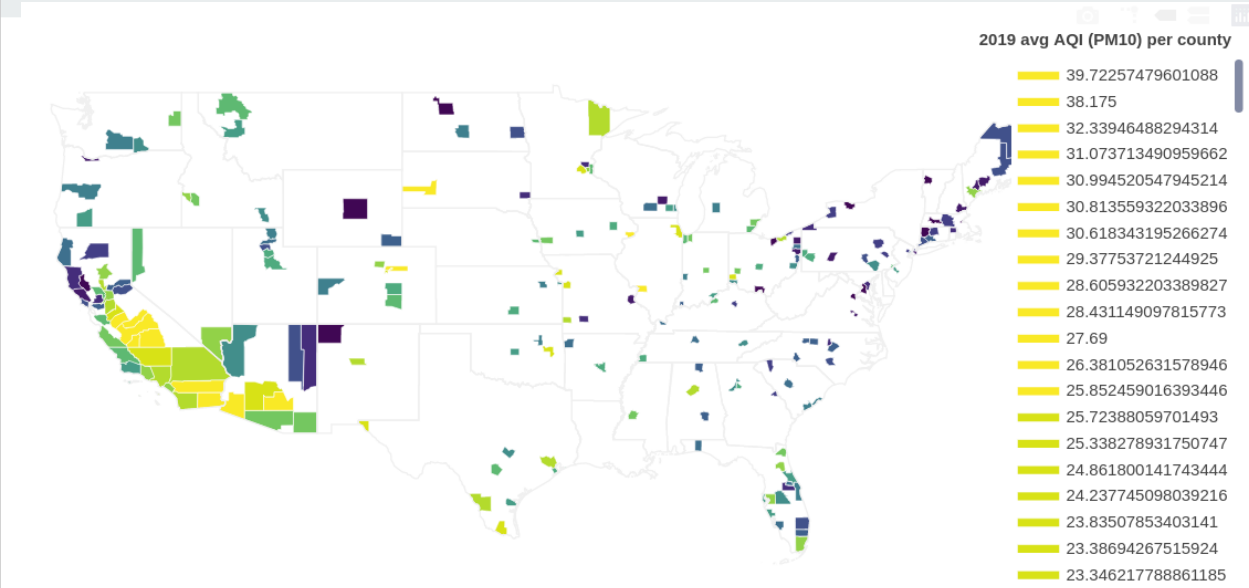
Another analysis done was a Two Sided T Test, to determine whether the top correlated attribute with PM10 levels of year 2017, which was HWAC_MALE_rate had a significant impact on 2017 PM10 levels.
HO = On average, counties that have above average HWAC ratio will have about the same pm10 aqi as other counties with below ratios
HA = On average, counties with below or above average HWAC ratios will have significantly different levels of pm10 aqi
Results of Two Sided T Test:
T-Statistic = 5.344351656283399
p-value = 6.201321040496359e-07
Reject the null hypothesis and assume that on average there is a significant impact on county PM10 aqi levels when the ratio of HWAC individuals is above average, which is above 8.50% in our situation

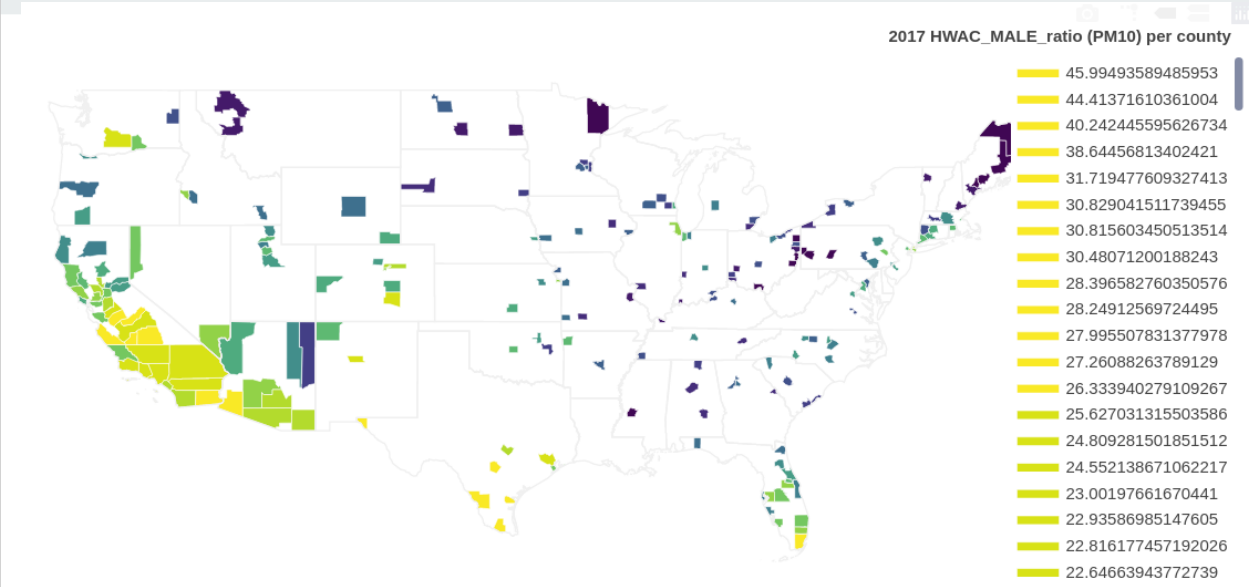


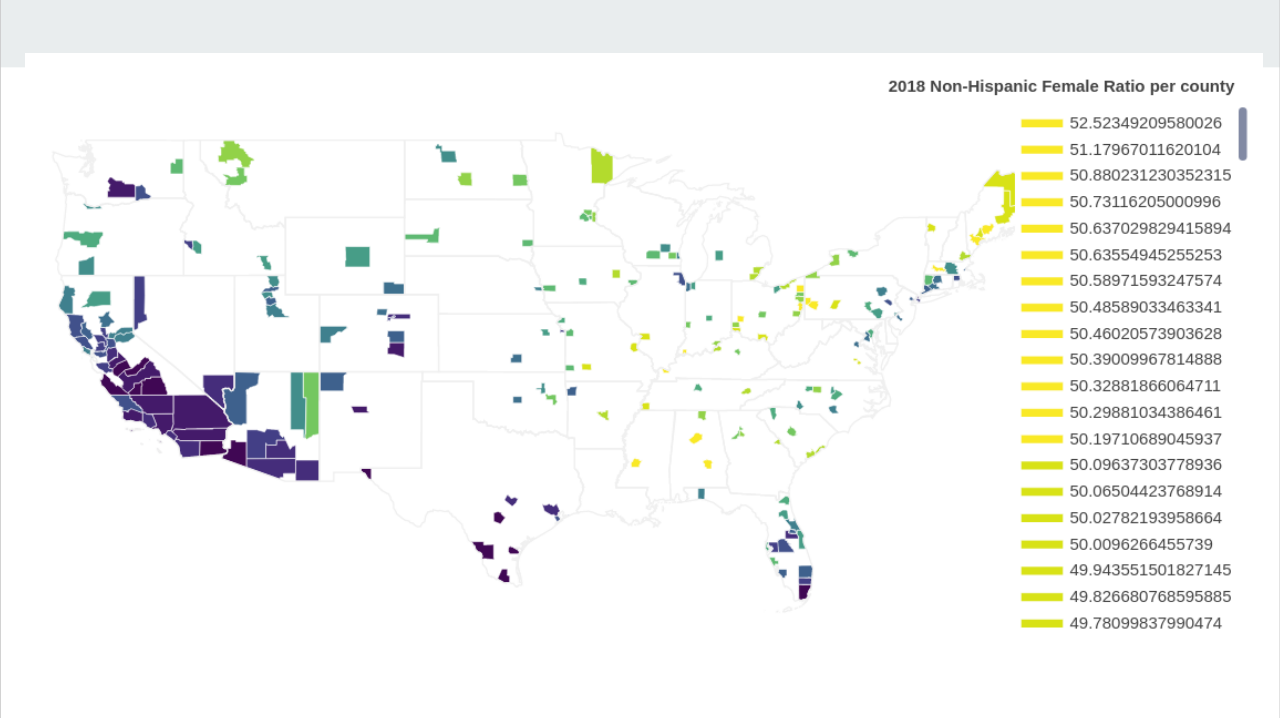
Because the objective is to predict PM10 levels for the year 2019, it would be of interest to investigate features that correlated with PM10 levels from our aggregated datasets. From the Pearson Correlation Table, notable points that can be summarized are that counties with large ratio of Hispanic communities are strong positively associated with PM10 levels of 2019; meaning counties with higher levels of PM10 pollution level are often associated with counties with high ratio of Hispanic demographics (note this association and not causation).

Another jutting point is that counties with larger ratio of individuals graduating with less than a high school diploma and also counties with larger unemployment rate are associated with higher annual PM10 readings.

Whereas, Non-Hispanic ratio of counties and ratio of total education attainments, had a negative association with PM10 levels. Meaning an increase in these social economic factors associated with a decrease in annual PM10 levels.


2019 avg AQI (PM10) per county
39.72257479601088
38.175
32.33946488294314
31.073713490959662
30.994520547945214
30.813559322033896
30.618343195266274
29.37753721244925
28.605932203389827
28.431149097815773
27.69
26.381052631578946
25.852459016393446
25.72388059701493
25.338278931750747
24.861800141743444
24.237745098039216
23.83507853403141
23.38694267515924
23.346217788861185


2017 HWAC_MALE_ratio (PM10) per county
45.99493589485953
44.41371610361004
40.242445595626734
38.64456813402421
31.719477609327413
30.829041511739455
30.815603450513514
30.48071200188243
28.396582760350576
28.24912569724495
27.995507831377978
27.26088263789129
26.333940279109267
25.627031315503586
24.809281501851512
24.552138671062217
23.00197661670441
22.93586985147605
22.816177457192026
22.64663943772739



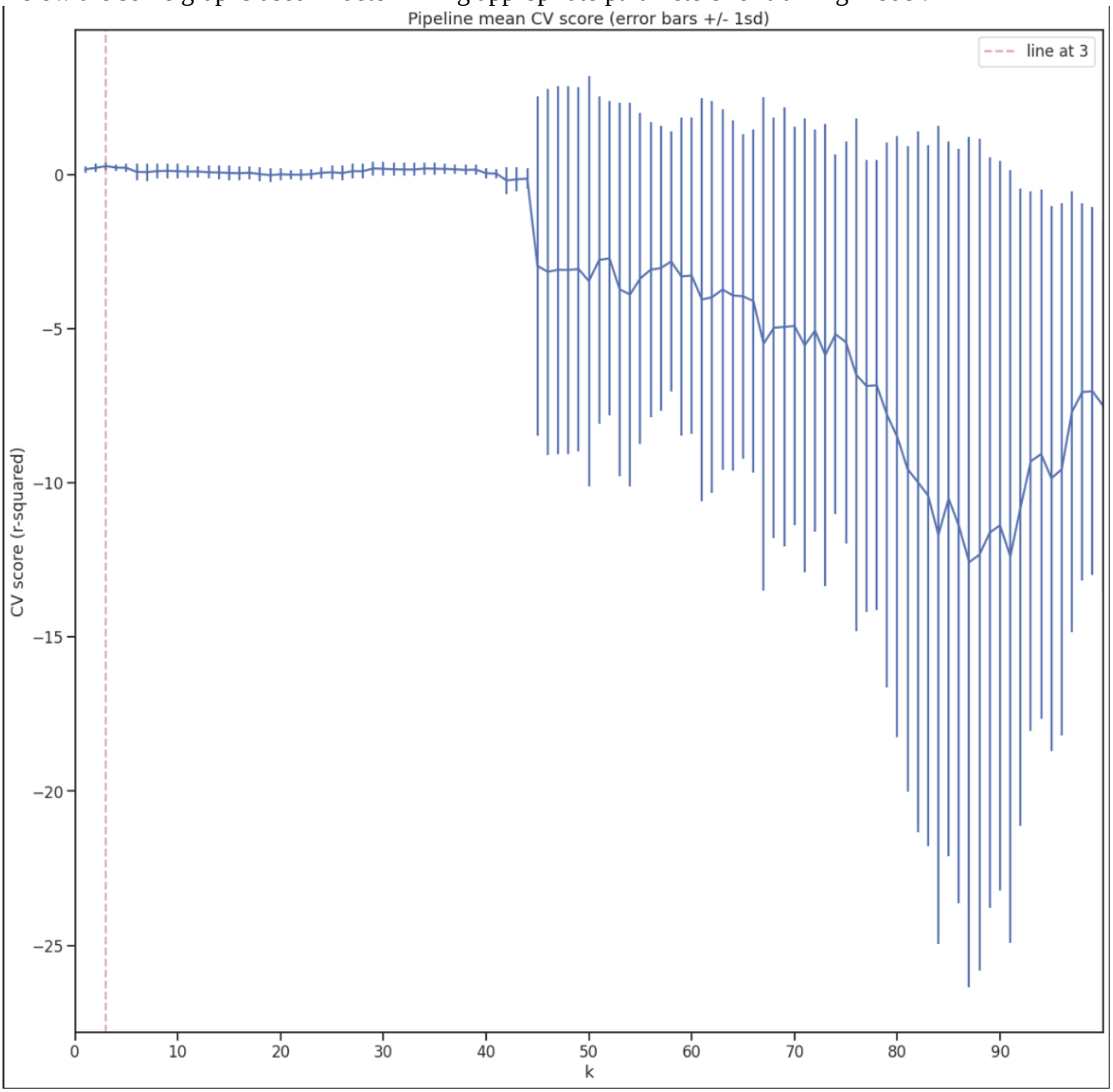
The above represents choropleth map of attributes of 2019 PM10 levels, 2017 Hispanic-White-Alone-Combination ratio per county, and 2018 Non-Hispanic Female ratio per county. Choropleth maps were used to display the counties being represented in this analysis as well as for further data exploration.

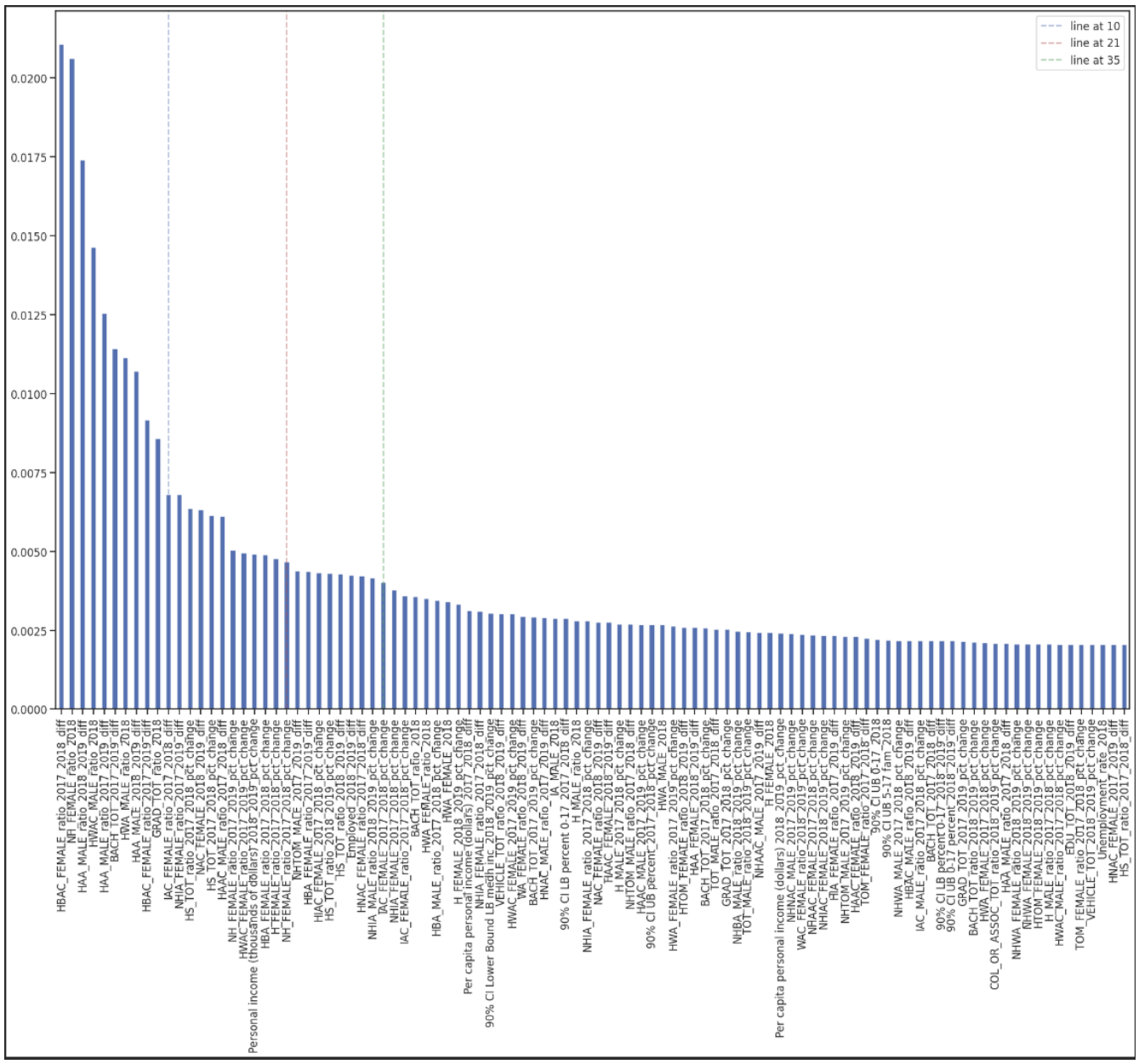
For data exploration, it can be noted that one of the top correlated attributes, 2017 HWAC Male Ratio, shows a near identical color mapping to the target attribute, 2019 PM10 levels. It can be depicted that there are higher concentration of high PM10 levels in lighter colored areas of the map which are predominantly in the western region of the US, encompassing California, Oregon, Washington, and Arizona. And it can be depicted that there are higher concentration of low PM10 levels in darker colored ares of the map which are predominantly the eastern region of the US, encompassing the New England states. These patterns can be shared between the positive correlated feature and the target feature; whereas a opposite contrast appears for a negative correlated feature with the target feature.

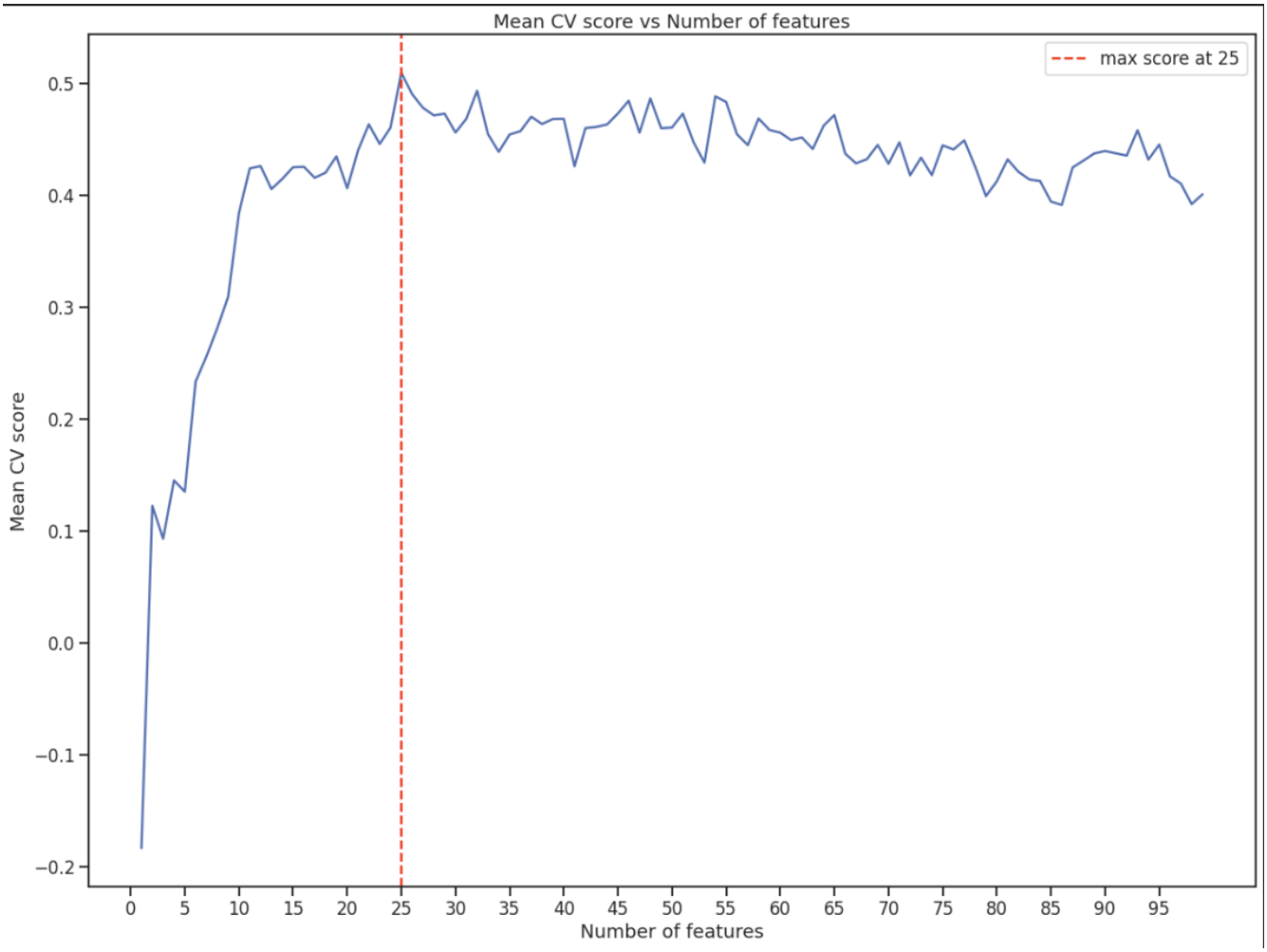
Pre-Processing: Because the dataset observations were annual recorded measures, the dataset can be considered as a time-series. And in dealing with a time-series dataset, stationarization of data is necessary. The data was stationarized to eliminate heteroscedasticity; this process was done via feature generating attributes that were the resulted difference of years 2017 to 2019. This created new features that contained resulted year differences of (2017-2018, 2017-2019, 2017-2018, 2018-2019), which contributed to having a dataset of 1311 features. These features were also standardized afterwards to allow data to be more palatable when training models.

Modeling:
Overall, three models were built for this project. The models were Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. Usage of feature selection via select-k-best and application of feature importance scores were used to reduce number of features and aid in improving the efficiency of model for future usage. Hyperparameter tuning was also utilized to improve results of the model as well.

Below are some graphs used in determining appropriate parameters for training model.




line at 10
line at 21
line at 35
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
HBAC_FEMALE_ratio_2017_2018_diff
NH_FEMALE_ratio_2018
HAA_MALE_ratio_2018_2019_diff
HWAC_MALE_ratio_2018
HAA_MALE_ratio_2017_2019_diff
BACH_TOT_ratio_2017_2018
HNA_MALE_ratio_2018
HAA_MALE_2018_2019_diff
HBAC_FEMALE_ratio_2017_2019_diff
GRAD_TOT_ratio_2018
IAC_FEMALE_ratio_2017_2018_diff
NHIA_FEMALE_ratio_2017_2019_diff
HS_TOT_ratio_2017_2018_pct_change
NAC_FEMALE_2018_2019_diff
HS_TOT_2017_2018_pct_change
HAAC_MALE_ratio_2017_2018_diff
NH_FEMALE_ratio_2017_2019_pct_change
HWAC_FEMALE_ratio_2017_2019_pct_change
Personal income (thousands of dollars) 2018 2019 pct change
HBA_FEMALE_ratio_2017_2018_pct_change
H_FEMALE_ratio_2017_2018_pct_change
NH_FEMALE_ratio_2017_2018_pct_change
NHTOM_MALE_2017_2019_diff
HBA_FEMALE_ratio_2017_2019_diff
HIAC_FEMALE_2017_2018_pct_change
HS_TOT_ratio_2018_2019_pct_change
HS_TOT_ratio_2018_2019_diff
Employed_2017_2018_diff
HNAC_FEMALE_ratio_2017_2018_diff
NHIA_MALE_ratio_2018_2019_pct_change
IAC_FEMALE_2017_2018_pct_change
NHIA_FEMALE_2017_2018_pct_change
IAC_FEMALE_ratio_2017_2018_pct_change
BACH_TOT_ratio_2018
HWA_FEMALE_ratio_2018
HBA_MALE_ratio_2017_2018_pct_change
H_FEMALE_2018_2019_pct_change
Per capita personal income (dollars) 2017 2018 diff
NHIA_FEMALE_ratio_2017_2018_diff
90% CI Lower Bound LB median inc_2018_2019_pct_change
VEHICLE_TOT_ratio_2018_2019_diff
HWAC_FEMALE_2017_2019_pct_change
WA_FEMALE_ratio_2018_2019_diff
BACH_TOT_2017_2019_pct_change
HNAC_MALE_ratio_2017_2019_diff
90% CI LB percent 0-17 2017 2018 diff
H_MALE_ratio_2018
NHIA_FEMALE_ratio_2017_2018_pct_change
NAC_FEMALE_ratio_2018_2019_diff
HAAC_FEMALE_2018_2019_diff
H_MALE_2017_2018_pct_change
NHTOM_MALE_ratio_2017_2018_diff
HAAC_MALE_2017_2019_pct_change
90% CI UB percent_2017_2018_pct_change
HWA_MALE_2018
HWA_FEMALE_ratio_2017_2018_diff
HTOM_FEMALE_ratio_2018_2019_diff
HAA_FEMALE_2018_2019_diff
BACH_TOT_2017_2018_pct_change
TOT_MALE_ratio_2017_2018_diff
GRAD_TOT_2017_2018_pct_change
NHBA_MALE_ratio_2018_2019_pct_change
TOT_MALE_ratio_2018_2019_pct_change
NHAAC_MALE_2017_2019_diff
HTE_MALE_2018
Per capita personal income (dollars) 2018 2019 pct change
NHNAC_MALE_2017_2019_pct_change
WAC_FEMALE_ratio_2018_2019_pct_change
NHAAC_FEMALE_2017_2018_pct_change
NHIAC_FEMALE_2018_2019_pct_change
HIA_FEMALE_ratio_2017_2019_diff
NHTOM_MALE_2017_2019_pct_change
HAAC_FEMALE_ratio_2018_2019_diff
TOM_FEMALE_ratio_2018_2019_diff
90% CI UB 5-17 Fam 2018
90% CI UB 5-17 Fam 2018
NHWA_MALE_2017_2018_pct_change
HBAC_MALE_ratio_2018_2019_diff
IAC_MALE_ratio_2017_2018_pct_change
BACH_TOT_2017_2018_diff
90% CI LB percent0-17 2018 2019 diff
90% CI UB 0-17 percent_2018_2019_diff
GRAD_TOT_2017_2019_pct_change
BACH_TOT_ratio_2018_2019_pct_change
HWA_FEMALE_2018_2019_pct_change
COL_OR_ASSOC_TOT_2018_diff
HAA_MALE_ratio_2017_2018_diff
NHWA_FEMALE_ratio_2018_2019_pct_change
NHWA_FEMALE_2018_2019_pct_change
HTOM_FEMALE_2018_2019_pct_change
H_MALE_ratio_2017_2018_pct_change
HWAC_MALE_ratio_2017_2018_pct_change
EDU_TOT_2018_2019_diff
TOM_FEMALE_ratio_2017_2018_pct_change
VEHICLE_TOT_2018_2019_pct_change
UNEMPLOYMENT_RATE_2018
HNAC_FEMALE_2017_2019_diff
HS_TOT_ratio_2017_2018_diff



The first graph was used in determining appropriate number of features for the Linear Regression model. From the graph it can be seen that after 40 (K Best Features) the train score dips and the variation starts to increase significantly whereas around 3 features the train score is at a local maxima on the graph and the train standard deviation is very small, which leads to 3 features leads to balance of bias and variance.

Second graph was used in determining appropriate number of features for Random Forest Regressor model. The graph shows the feature importance scores sorted from largest to smallest when running this model with all features. From looking at the graph it can seen that after 21 or 35 features the feature importance scores plateau, and it can be justified that the cumulative score of the 21 features represent a significant proportion of the features used in the model; thus 21 features were used in modeling with Random Forest Regressor.

Last graph was used in determining optimal number of features for Gradient Boosting Regressor model. The model was first run with all features to gather the feature importance scores of features used in Gradient Boosting Regressor model. Afterwards the model was run with a number of the top scores features; and from the graph above it can be depicted that at 25 top scored features, the model performed with best score.
(More graphs were used in determining parameters such as estimators used for models, and this can be further reviewed in notebook source code).

Model Evaluations: In comparing the results of the models, a baseline was first established using Dummy Regressors to determine how well our models had performed. Below are Dummy Regressor Evaluation metrics.

- Dummy Regressor w/ Mean Strategy
  - Train MAE mean: 4.002, Train Std 0.537
  - Test MAE mean:4.224
  - Train R2 : -0.0199, Train Std 0.016
  - Test R2 : -0.0147
- Dummy Regressor w/ Median Strategy
  - Train MAE mean: 3.919, Train Std 0.491
  - Test MAE mean:3.966
  - Train R2 : -0.0563, Train Std 0.077
  - Test R2 : -0.006

The Dummy Regressors used mean and median of the target feature, 2019 PM10 levels, as its strategy. And both regressors resulted with negative correlation of determination which interprets to abysmal results.
Below are Evaluation metrics for the Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor models.

- Linear Regression Model
  - Train MAE mean: 3.525, Train Std 0.479
  - Test MAE mean:3.182
  - Train R2 : 0.265, Train Std 0.118
  - Test R2 : 0.406

```
Number of features: 3
HWAC_MALE_ratio_2018    17.546961
HWA_MALE_ratio_2018     -3.238566
H_MALE_ratio_2018      -11.785481
dtype: float64
```

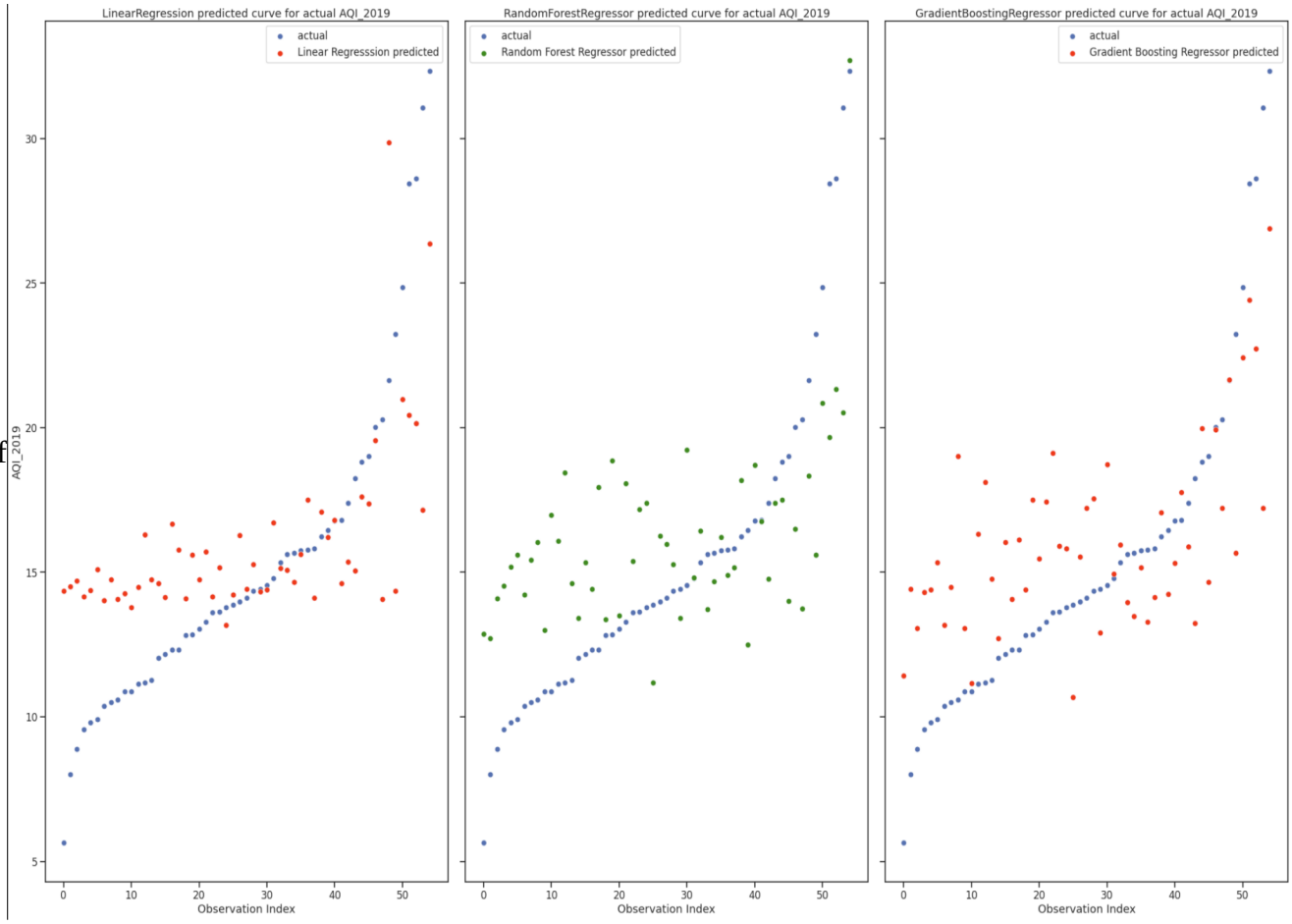
- Random Forest Regressor
  - Train MAE mean: 2.993, Train Std 0.466
  - Test MAE mean: 3.512
  - Train R2 : 0.407, Train Std 0.166
  - Test R2 : 0.404

```
Number of features: 21
BACH_TOT_2017_2019_diff                                          0.072646
HAA_MALE_ratio_2018_2019_diff                                    0.070659
HBAC_FEMALE_ratio_2017_2018_diff                                 0.066875
IAC_FEMALE_ratio_2017_2018_diff                                  0.059033
HBAC_FEMALE_ratio_2017_2019_diff                                 0.056135
HS_TOT_2017_2018_pct_change                                      0.055338
NHIA_FEMALE_ratio_2017_2019_diff                                 0.055242
Personal income (thousands of dollars) 2018_2019_pct_change      0.055009
HAA_MALE_ratio_2017_2019_diff                                    0.054966
HS_TOT_ratio_2017_2018_pct_change                                0.045067
NH_FEMALE_ratio_2018                                             0.044850
GRAD_TOT_ratio_2018                                              0.043479
HAA_MALE_2018_2019_diff                                          0.040214
HAAC_MALE_ratio_2017_2018_diff                                   0.039959
HWAC_FEMALE_ratio_2017_2019_pct_change                           0.039327
NH_FEMALE_ratio_2017_2019_pct_change                             0.036740
NAC_FEMALE_2018_2019_diff                                        0.036493
HWAC_MALE_ratio_2018                                             0.034532
HWA_MALE_ratio_2018                                              0.032657
H_FEMALE_ratio_2017_2018_pct_change                              0.031667
HBA_FEMALE_ratio_2017_2018_pct_change                            0.029112
dtype: float64
```

- Gradient Boosting Regressor
  - Train MAE mean: 2.957, Train Std 0.457
  - Test MAE mean: 3.329
  - Train R2 : 0.4437, Train Std 0.018
  - Test R2 : 0.4436

```
Number of features: 25
['HAA_MALE_ratio_2017_2019_diff',
 'HWA_MALE_ratio_2018',
 'HAA_MALE_ratio_2018_2019_diff',
 'HWAC_MALE_ratio_2018',
 'HBAC_FEMALE_ratio_2017_2018_diff',
 'HS_TOT_2017_2018_pct_change',
 'IAC_FEMALE_ratio_2017_2018_pct_change',
 'Employed_2018_2019_diff',
 'NHIA_FEMALE_ratio_2017_2018_pct_change',
 'AQI_2017_2018_diff',
 'HWA_FEMALE_2018',
 'NHNAC_MALE_2017_2018_pct_change',
 'HIA_FEMALE_ratio_2017_2019_diff',
 'NH_FEMALE_ratio_2018',
 'NHAA_MALE_2017_2019_pct_change',
 'NAC_FEMALE_2017_2019_pct_change',
 'NHNAC_MALE_ratio_2017_2018_diff',
 'BACH_TOT_2017_2019_diff',
 'NHBAC_FEMALE_ratio_2017_2018_diff',
 'HIA_FEMALE_ratio_2017_2019_pct_change',
 '90% CI LB percent 0-17 percent_2018',
 'Personal income (thousands of dollars) 2017_2018_diff',
 'TOM_FEMALE_ratio_2017_2019_diff',
 'TOT_MALE_ratio_2017_2018_pct_change',
 'NHWAC_MALE_ratio_2017_2019_diff']
```

ff



In comparing all three models, they all have a similar R^2 score of around .4 with the Gradient Boosting Regressor having the largest score of 0.44. The Linear Regression model uses the least number of features, 3, whereas Random Forest Regressor uses 21 and Gradient Boosting Regressor uses the most which is 25 features. From looking at the graphs above, which compare the actual to model predicted 2019 PM10 levels, it appears that although the Linear Regressor model uses least number of features, it is most variable, whereas the two other models are very similar in regards to bias and variance.
Although a R^2 of 0.4 is considered weak, these three models significantly outperform the Dummy Regressors, which were used as the baseline of this project and had yielded negative R^2 scores,

Limitation & Future Improvements:

- Low number of observations (Big P Little N)
  - 1311 Features vs 208 Observations
- Nested Hyperparameter Tuning was not utilized; however this would be suggested to improve generalization of results in future

Potential Insights:

- Information on trend of uneducated and large ratio of Hispanic communities being associated with higher pollution levels may be useful for property valuation where land may be less valuable in the future due to health harms such as air pollution

- Knowing areas where likely high levels of air pollution would be useful for environmental protection agencies in focusing where to measure and analyze pollution as well as focus to find solutions; also useful for areas less needed in focusing on.
- Highly polluted areas are most likely polluted due to reasons such as less environmental protection or lack of concern from residents of community, knowledge of less educated and large ratio Hispanic communities being associated with higher PM10 levels, could mean organizations could potentially take advantage of these communities with less friction.