**CMPE 255: Group 9: Project Report**

**Team Members:**

Varsha Suresh (015245347)

Lasya Bheemendra Nalini (014618695)

Auni Bagchi (011256895)

**Project Title: Predicting Air Quality and Pollutant Levels in the US**

**Link to Shared Google Folder Containing Source Files:** <https://drive.google.com/file/d/18IsBtl4SbWQEzYJawsVS_l3qKmI67kj-/view?usp=sharing>

# Section 1: Introduction

• Motivation

Air quality is a critical measure to help us understand environmental conditions and make informed public health decisions. Living in an area with poor air quality can lead to bad long-term health outcomes. Currently, the Environmental Protection Agency, or EPA, predicts Air Quality only days ahead, which does not give much time for the public to prepare. Thus, being able to predict air quality much ahead of time is very important.

• Objective

Our objective was to use EPA’s Air Quality Index, or AQI, data for past years in order to predict AQI in upcoming years, and check how accurate such predictions can be. As a next step, we aimed to find out if there was an association/correlation between the AQI in a certain location with the population density of the place. We also explored the possible association between AQI and the location setting, and AQI and land use of the site in question which measures air quality.

The air quality, land use, and location setting data from the EPA is collated for a period of 21 years in our analysis, from 2000 to 2020. The population data and estimates are sourced from the US Census Bureau. A description of land use and location setting can be found in the following table:

| Land Use | A category describing the predominant land use within a 1/4 mile radius of the site (ex: residential, commercial, agricultural, forest, desert, etc.) |
| --- | --- |
| Location Setting | A description of the setting within which the monitoring site is located (ex: rural, urban and city center, suburban, etc.) |

• Literature Review

We looked at different papers to understand approaches that have been taken previously using pollutant data.

To understand the role of the measurement method, we read through [Mou et al (2020)](https://aaqr.org/articles/aaqr-20-05-oa-0217). We learned that though there are various factors that determine variability for a measurement like PM 2.5, many of them can be explained by meteorological factors, as well as the levels of other pollutants. Hence, we have used data of all five main pollutants and their levels in our analysis.

To understand approaches taken to do regression analysis, we reviewed [Castelli et al (2020)](https://www.hindawi.com/journals/complexity/2020/8049504/). From this, we took away the importance of having a large dataset to improve predictive modeling accuracy, as suggested in the paper. By looking at national level data (as opposed to data for only one state as is done in the paper) and compiling 21 years of daily summary data, we have made an effort to improve our ability to successfully predict AQI.

# Section 2: System Design & Implementation Details

• Algorithms Considered/Selected

We considered various regression algorithms for our project, including Gradient Boosting Regression, Random Forest Regression, Linear Regression, and Bayesian Ridge Regression. We selected the algorithm that gave us the best combination of accuracy and execution time for training the model for each of the 5 pollutants that are the focus of our project.

For example, in the case of PM 2.5, Linear Regression yielded an accuracy of 73.98% on the test data. Random Forest resulted in a much better accuracy of 99.92%, but took 11 minutes to train. Gradient Boosting led to 99.98% accuracy, and only took 7 min to train, so that was our choice of algorithm for that pollutant.

A similar process was followed for the remaining pollutants and the most accurate models were chosen and combined for further AQI analysis with county-wise population and location data.

For association analysis, we chose the chi-square test to check the correlation between Air Quality, Location Setting, and Land Use. The chi-square test best fits our needs as all features mentioned earlier are categorical, and we intended to analyze cross-tabulated data. To find the correlation between AQI and county population, Pearson, Kendall, and Spearman methods were considered.

• Technologies & Tools Used

We used Google Colab, Numpy, Pandas, Scikit-learn, and Tableau. Google Colab allowed us to easily centralize our large amount of data and run our notebooks using the same files. The various Python libraries were ones that we had used previously in this course, though we did utilize different parts of those libraries than we had before in order to create our prototype.

To generate the visualization for predicted AQI values, we created a Tableau dashboard with different sheets. The first one contains all the counties that our dataset predicts AQI values for. The second sheet is used to drill down into the counts of the number of good, moderate, and unhealthy AQI days present in each county across sites per month. Using Tableau enabled us to abstract away details on how the graphs were generated while having a template which can be refreshed with newer data at will.

• System Design / Data Flow

After initial preprocessing and evaluation of the regression algorithms for each of the five pollutants’ data, we applied the appropriate regression and saved the predicted AQI values to csv files.

Those csv files were then combined and merged with Census data, specifically Population, and Site Description data, specifically Land Use and Location Setting, in order to allow us to have everything we needed to create our data visualization.

For the visualization, we grouped our combined data by County and Year to make the visualization easier to follow.

Using the training data, with years 2000-2017, we analyzed the association/correlation between Air Quality and Population, Air Quality and Location Setting, and Air Quality and Land Use.

• Screenshots

Figure 1, Data for Santa Clara County, year 2018:

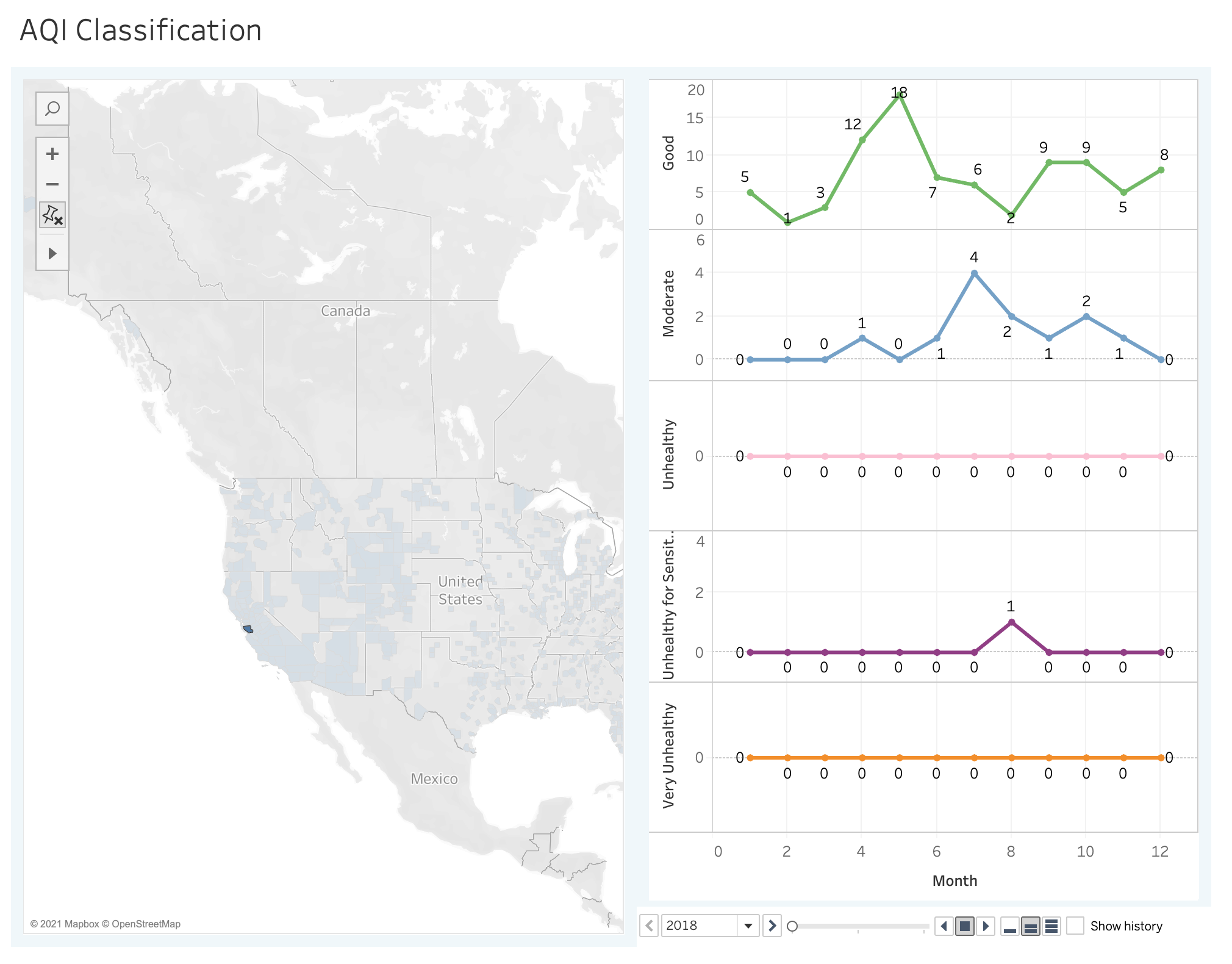
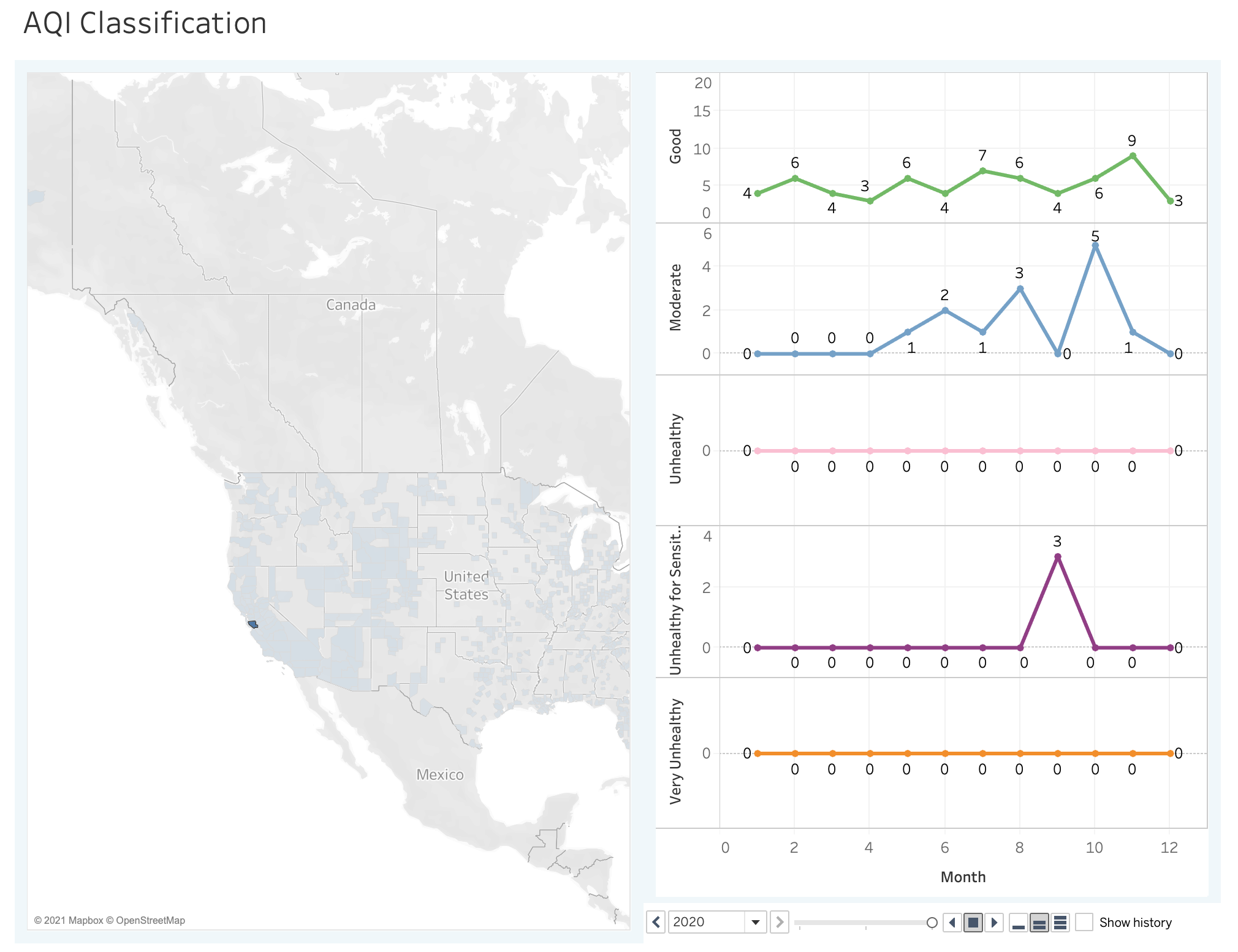


Figure 2, Data for Santa Clara County, year 2020:



# Section 3: Experiments / Proof of Concept Evaluation

• Dataset Information

Datasets:

We used the EPA Daily Summary Dataset for 5 pollutants for our AQI analysis.

The US Census data was used to get the population sizes over time across counties.

The EPA Site Description data was used to get land use and location setting for each monitoring site.

Type of Data:

We combined three different datasets for our analysis and chose the relevant subset of features which made our dataset a heterogeneous one. This dataset includes numeric and categorical data, such as population and land use, respectively.

Size of Data:

Our training data for AQI consisted of daily summary data for a period of 18 years. For each of the 5 pollutants, this size is of the order of ~1 million records. The test data used was for a period of 3 years. The size of this data was in the order of ~200k records for each pollutant.

Preprocessing:

Each of the pollutant daily summary data files underwent a similar series of preprocessing steps.

1. The day, month, year were extracted from the Data Local column
2. An appropriate Sample Duration was chosen. (This varies across pollutants)
3. Other irrelevant features and those with categorical values were discarded as they were found to not be salient for the purpose of predicting AQI.
4. Label encoder was used wherever necessary to convert categorical values to numeric data for use in the prediction models.
5. Duplicate Records and Records having nan values were dropped.

EPA site data was preprocessed by removing an irrelevant site number value.

Data Merge:

Once the predicted AQI values were generated for the different pollutants, the predicted values were merged into a single dataset. This dataset was then combined with the US census data and site information to perform an association analysis.

1. To merge the census data, the State and County Codes were used
2. To merge the EPA site data, the State and County codes were used in addition to the site number which identifies the site where the pollutant level was captured

• Methodology Followed

Regression Analysis:

Post data preprocessing we choose data of years 2000-2017 as our training data, and data of years 2018-2020 as our testing data, following a roughly 85-15 training-test split. This has been followed in regression analysis of all 5 pollutants.

We trained a Gradient Boosting Regressor model for pollutants PM 2.5, CO, Ozone, and SO2. For NO2, a Linear Regression model was trained to predict AQI. We then merged the AQI predicted for all pollutants and extracted AQI for each day.

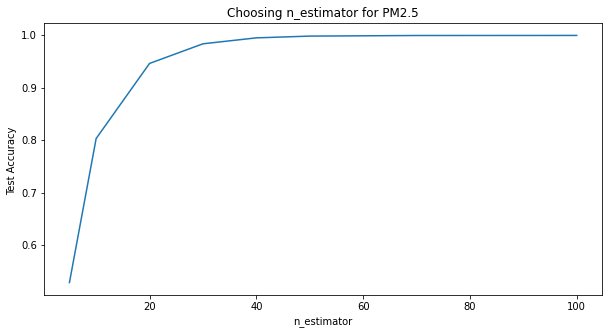
Association Analysis:

To find association/correlation between Air Quality and Location Setting, first we created a cross tabulation between Air Quality and Location Setting to capture the frequency of Air Quality against Location Setting (Rural, Urban and Suburban).

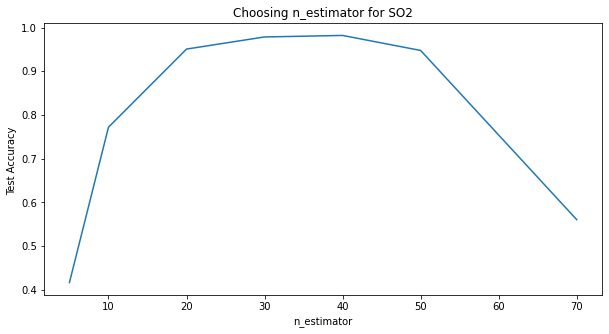
To evaluate the test we chose a significance value of 0.05 and had the null hypothesis: The features Quality and Location Setting are NOT dependent (which means they are not associated). After applying the chi-squared test, based on our p-value, which was greater than the significance value, we decided to reject our null hypothesis. The same methodology was applied to find association/correlation between Air Quality and Land Use.

• Comparison of Different Parameters/Algorithms Evaluated

For PM2.5, we were able to achieve accuracy of 99.98% using gradient boosting regressor with default n\_estimator 100. We plotted the graph to determine the n\_estimator value for which accuracy is stabilized. We chose n\_estimator as 40 based on the below graph.



For SO2, we were able to improve accuracy from 35.46% to 98.20% by choosing n\_estimator as 40 instead of default n\_estimator of 100 for gradient boosting regressor. We plotted the graph to get the n\_estimator value for which maximum accuracy is achieved. We choose n\_estimator as 40 based on the below graph.



We tried to further hypertune parameters of the gradient booster using GridSearchCV. Training data was split into a training and validation set and passed as input to GridSearchCV, which returned best parameters {'max\_depth': 50, 'n\_estimators': 100}. However, accuracy dropped with these set of parameters.

Here are some of the pertinent regression algorithm comparisons:

| Pollutant | Model | Accuracy |
| --- | --- | --- |
| CO | RandomForestRegressor | 0.7567426500776686 |
| CO | LinearRegression | 0.9904037676772358 |
| CO | **GradientBoostingRegressor (n\_estimators = 30)** | 0.9940943902841846 |
| Ozone | RandomForestRegressor | 0.8286162428211814 |
| Ozone | LinearRegression | 0.8474767432110633 |
| Ozone | **GradientBoostingRegressor (n\_estimators = 30)** | 0.9928378665232495 |
| NO2 | RandomForestRegressor | 0.8896325733253335 |
| NO2 | **LinearRegression** | 0.997654529231295 |
| NO2 | GradientBoostingRegressor (n\_estimators = 30) | 0.9964066571611149 |
| PM2.5 | LinearRegression | 0.7398197103884219 |
| PM2.5 | RandomForestRegressor | 0.992764747099709 |
| PM2.5 | **GradientBoostingRegressor (n\_estimators = 40)** | 0.995248 |
| SO2 | LinearRegression | 0.49642129625292253 |
| SO2 | RandomForestRegressor | 0.2957474436992166 |
| SO2 | **GradientBoostingRegressor (n\_estimators = 40)** | 0.9820155519901449 |

• Analysis of Results

1. AQI prediction for each pollutant is above 0.98.
2. The visualization in Figures 1 and 2 show the variation of AQI ranges for Santa Clara county in 2018 and 2020 respectively. The site with the maximum AQI value has been chosen to represent the county as per EPA guidelines. From Figure 2, we can observe that the AQI is relatively high for the months of September through November. Historically, these are the months when the county is most prone to have wildfires. The predicted AQI values therefore corroborate real life findings.
3. Association analysis shows that Air Quality is dependent on Location setting and Land Use.
4. Association analysis also shows that Air quality is positively correlated with County population. As correlation coefficient is 0.07, we can say that the strength of association is small.

# Section 4: Discussion & Conclusions

• Decisions made

1. Each pollutant had to be preprocessed and an appropriate sample duration had to be chosen for each.
2. The model for predicting AQI had to be implemented separately for each pollutant.
3. Train-test split had to be decided so that the overall accuracy across pollutants would be high.
4. Decisions on which features of the source dataset had to be retained and which ones discarded.

• Difficulties faced

1. Choice of Monthly summary data vs Daily summary data for analysis.
2. Deciding what data/which association to display in the visualization step and how to group the data.
3. Deciding how to approach association analysis for the dataset.

• Things that worked

1. Leaving out a lot of categorical values in EPA dataset helped improve the accuracy of the model as they were not influencing the AQI.
2. Discussions on pre-processing helped each member of the team to follow similar strategies to achieve high accuracy. Without proper preprocessing the accuracy was very low in the first few attempts to predict AQI.

• Things that didn’t work well

1. Monthly summary was chosen first but did not give accurate results and accuracy was low, therefore we decided to go with daily summary.
2. For generating visualization, averaging AQI over a month did not help as it seemed to smooth out days where the AQI was very high.
3. Simple association analysis like Apriori did not work well since that was not suitable for the data at hand. Frequent itemsets could not be extracted meaningfully from the dataset. An extremely high number of nonsensical rules were generated.

• Conclusion

1. Accuracy for each pollutant:
   1. CO data has been predicted with Gradient Boosting Regressor with 99.4% accuracy.
   2. Ozone data has been predicted with Gradient Boosting Regressor with 99.28% accuracy.
   3. NO2 data has been predicted using Linear Regression with 99.76% accuracy.
   4. PM2.5 data has been predicted with Gradient Boosting Regressor with 99.52% accuracy.
   5. SO2 data has been predicted with Gradient Boosting Regressor with 98.2% accuracy.
2. Association Analysis:
   1. A positive but small relation has been found between AQI values and population across all states.
   2. A positive but small relation has been found between AQI values and land use in the site location across all states.
   3. A positive but small relation has been found between AQI values and location setting in the site location across all states.
3. Visualization:  
   Visualization graphs have been generated to show the number of good/moderate/very unhealthy etc ranges of AQI for a county over different years. This data is shown across the multiple sites that can be present within a county.

# Section 5: Project Plan / Task Distribution

o Who was assigned to what task

1. Preprocessing and predictive models for AQI
   1. PM2.5 and SO2: Lasya
   2. Ozone and NO2: Auni
   3. CO: Varsha
2. Merge predicted AQI data with population and site info and generate Tableau graphs: Varsha
3. Association Analysis:
   1. Correlation, Chi-test, nominal associations: Lasya
   2. Apriori unsuitability: Varsha
4. Debug issues and brainstorm alternate possibilities: Team Effort
   * 1. Had multiple meetings to ensure progress and help each other with roadblocks
5. Project Report: Team Effort
6. Project Slide Deck for Presentation: Team Effort
7. Initial project proposal and exploratory analysis: Team Effort

o Who ended up doing what task

As detailed above, with the following addition:

1. Analysis to update visualization without monthly mean AQI: Lasya