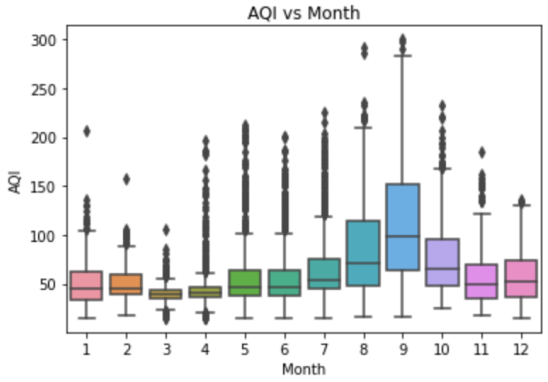
Authors: David Long, Brian Zheng, Winnie Xiao

Open Ended Modelling Report

1. Open Ended EDA

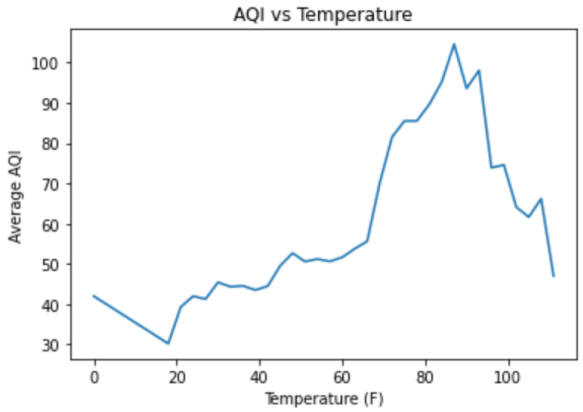
In the first section, we will present a brief overview of the open-ended EDA we conducted in part 1.

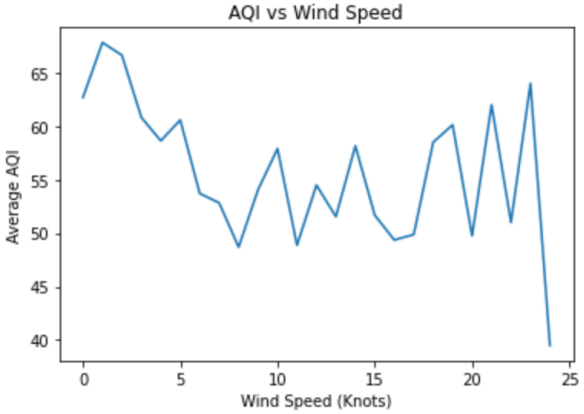
1. According to our box plot below, the median AQI is generally low from March to June--about 35, though there are occasionally some outliers. For the rest of the year, the median AQI is relatively high, ranging from 50 to over 100. Then we attempt to visualise the relationship between wind and AQI. We used linear regression to model wind and AQI and found that there is a weak negative relationship between wind and AQI. The slope and intercept of our linear regression line is -0.00354 and 68.64432747071004.



2. Next, we used both wind and temperature to model AQI. The coefficients are 1.27507759, -0.00799342, and the intercept is -9.754801085428369. We discovered a positive correlation between temperature and AQI and a negative correlation between wind and AQI. Therefore, our model suggests that the higher the temperature is, the higher AQI is, and the lower the strength of wind, the higher AQI is (the air is more polluted), which is helpful when predicting AQI. After modelling the data with linear regression, we also used line plots to visualise the correlation found utilising linear regression.

3. The first line plot visualises the relationship between temperature and AQI. The x-axis is temperature, which ranges from 15 Fahrenheit to 120 Fahrenheit and increases by 3 Fahrenheit each time; the y-axis is the mean AQI of every specific temperature. Although the mean AQI fluctuates as temperature increases, we can see that the mean AQI increases as temperature increases. The second line plot visualises the relationship between wind and AQI. The x-axis is the strength of wind, which ranges from 0 to 25 Knots; the y-axis is the mean AQI of every specific wind strength.





There are, however, a few other questions that still need to be addressed:

1. Are there other confounding factors that would mislead the experiment result?
2. How do we address the problem of other geographical variables that might be associated?

2. Problem

Hypothesis:
Population density is positively correlated with AQI level and the addition of population density as one of the features can improve the prediction model for categorical AQI levels.

How to test for test results:
We would first develop a prediction model using only environmental features like temperature mean, wind mean, and so2 level mean. Then we would use the same model but include the feature of population density in. To test our model, we would split the dataset into training and testing sets to be predicted by both models. We would run predictions with the two models for 100 times each and calculate the prediction accuracy (binary error) each time. To test for our hypothesis, we would use the t-test to see if the difference between the two model accuracies are statistically significant with the p-values calculated. For this report, we would use the significance level of 95%. Therefore, if our resulting p-value is less than 0.05, the difference is statistically significant for use to reject the null hypothesis in favor of the alternative hypothesis that population density is indeed a good feature to improve prediction accuracy for AQI level.

Result from test:
After testing our hypothesis, we obtained a t-score of -13 and the corresponding p-value of 1.4 * 10^-39. This value is significantly smaller than 0.05, so we would reject our null hypothesis (no difference) in favor of the alternative. Thus, the answer to our hypothesis is that the addition of

population density as one of the features can indeed improve the prediction model for categorical AQI levels.

3. Modeling

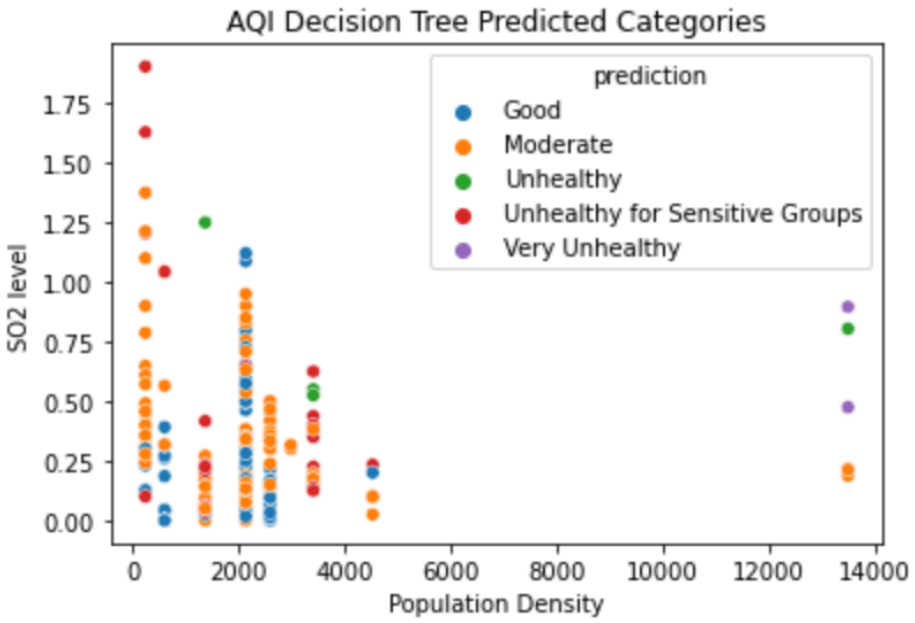
Since we are trying to predict the categorical AQI level that are standardized by AirNow.gov (Home of the US Air Quality Index), we decided to use a **decision tree** model.

The **input** to our model is the features of average temperature, average wind, average SO2 level, and population density by county in California. Our input data is based on the daily data entries from EPA.gov on the environmental statistics periodically, and the population density data collected in 2021 by the US Census Bureau. (https://www.california-demographics.com/counties_by_population)

The **output** to our model is AQI levels in 6 categories: Good, Moderate, Unhealthy for sensitive groups, Unhealthy, Very Unhealthy, and Hazardous.

4. Model Analysis and Evaluation

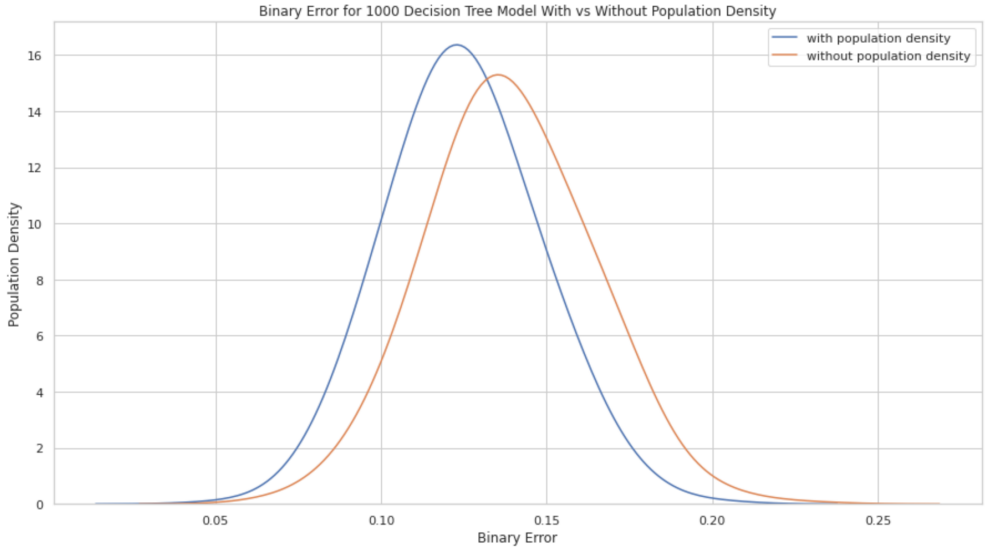
1) Basic Decision Tree



After feeding the input into our and generating the predictions, the statistics favors the hypothesis. The binary error is around 0.11, which is rather small. It indicates that the model correctly predicts around 90% of the AQI categories, a relatively high success rate. The cross validation error is 1.5, which is also relatively low. Then, we visualized the result of the decision tree predictions using a scatter plot.

As seen from the graph above, the relationship between population density and AQI category is not very well-reflected by the visualization. A vague positive relationship exists between these two variables, but much stronger and undeniable evidence is needed to validate our hypothesis. Meanwhile, it's unclear whether it's the population density that leads to the relationship observed or it's due to other variables. Thus, we need to isolate other variables and provide clearer evidence. This motivates us to find a better visualization tool and generate something easier to visualise, which eventually leads us to use t-test.

2) <u>One Sided t-test</u>



We first used the decision tree model with all the other variables present except for population density, to predict for one thousand times and keep track of the binary error each time. We did the same for our original model with population density present, and recorded the binary errors in an array. Then, we used the t-test tool in numpy to conduct a one-sided t-test at a significance level of 0.05. Our null hypothesis for the t-test is that including population density in our model results in a larger or equal binary error, and the alternative hypothesis is that including population density contributes to a lower binary error. The result shows extreme strong evidence favoring the hypotheses, with a t-score of -13 and the corresponding p-value being 1.4 * 10^-39, which is significantly lower than the 0.05 significance level. As a result, we have proved that population density is correlated with AQI level. We have provided the visualization below for population density vs binary error for the 2 models, one without and one with population density. The graph also

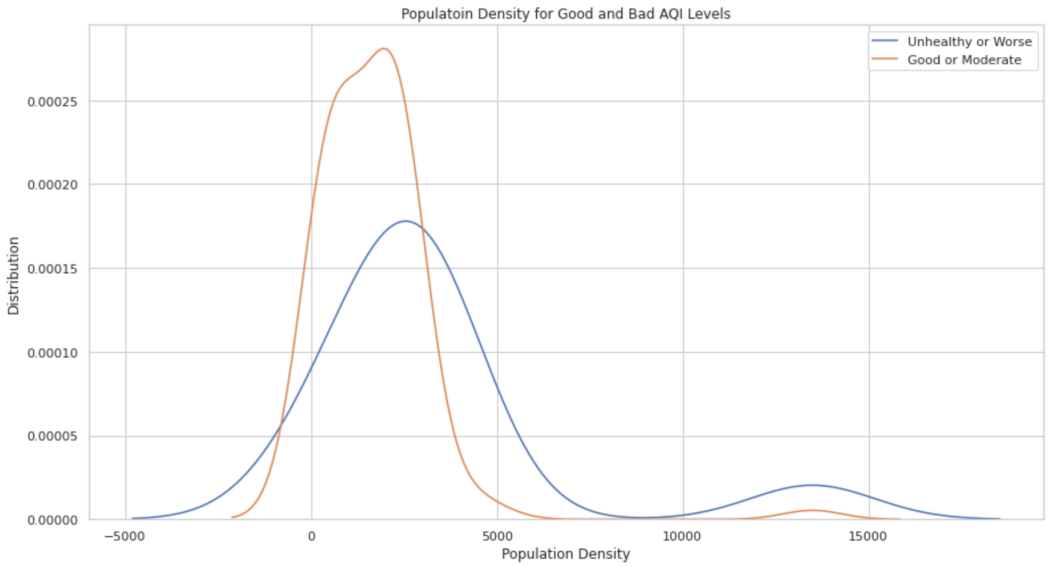
clearly shows that by including population density, binary error decreases noticeably for the decision tree model, from around 0.14 to 0.12.

However, this only proves half of our hypothesis. We have shown that correlation exists between population density and AQI level, but we have not provided enough evidence to show that a positive relationship exists. Thus, we decided to investigate further upon the direction of this relationship.

3) Further one-sided t-test

Still the same decision tree model with average temperature, average wind speed, SO2 level, and population density as input, but this time split the predicted AQI categories into 2 groups: one for good and moderate, and one for the rest. For each group, we put their corresponding population densities in an array and conducted another t-test. Our null hypothesis for t-test is that the population density for the unhealthy or worse group is less or equal than the population density for the good or moderate group. The resulting t-score is 4.8, and the corresponding p-value is 1.2 * 10^-6, which is statistically significant at 0.05 level. Thus, we have proved that the higher the population density, the higher the AQI level.

In the line plot below, we have provided the distribution of population density for good (good and moderate) and bad (unhealthy to sensitive people or worse). We can easily see that the population density tends to be less when the weather is good, and vice versa. Therefore, we have successfully proved our hypothesis that population density is positively correlated with AQI level.



5. Model Improvements

Improvement #1: Change from Linear Regression to Decision Tree Model
Problem: Originally, we were planning to use a linear regression model for prediction purposes. We would first predict the numerical AQI levels with the specified features of temperature, wind, so2 concentration, and population density. With the predicted numerical values, we would bin them into the categories according to the categorical standards issued by AirNow.gov (Home of the U.S. Air Quality Index). But we recognized how the binary error of this linear model is relatively high, being around 0.3, which means it only has a prediction accuracy of 70%.

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

https://www.airnow.gov/aqi/aqi-basics/

Solution: Since the targeted data is categorical, we decided to try decision tree models instead to better accommodate the characteristics of the data we are working with.

Result: With the same set of data and features inputted, we found out that a decision tree model can reduce the binary to around 0.15, which is much less than that using a linear regression model.

Improvement #2: Testing Model Results
Problem: Initially, we were only planning to run a decision tree with population density as one of its features for once to see if the model can have a high prediction accuracy. However, we realized that this would not be a legitimate statistical test.

Solution: Since we are trying to test whether the addition of population density as one of the features can improve prediction models, we decided to model two different decision tree models: one with the features of temperature, wind, and so2 concentration, and the other with the addition of population density.

To legitimize our hypothesis testing, we decided to run the prediction models for 1000 times each and obtain the binary error of each prediction. With the two sets of experiment data, we would choose to use an one sided t-test to see if the difference between the two sets of binary error is actually statistically significant. With the t-stat, we can then obtain a p-value to actually evaluate our hypothesis test with confidence intervals instead of a manually determined level of prediction accuracy to be considered a “good model”.

Result: This method indeed worked giving us a logical path to test out our hypothesis. The result is also clearly depicted in the two one-sided test graphs that we included above.

6. Future Works:

Admittedly, the decision tree model that we used has a high accuracy of predicting AQI from So2 concentration, wind, temperature, and population density. However, since we used a decision tree as our model, we could not know the exact relationship between each of the explanatory variables (So2 concentration, wind, temperature, and population density) and the response variable (AQI). In other words, we want to see the strength of the correlation between each of our explanatory variables and AQI, which can be useful to some real world applications. Therefore, one aspect of future work could be deriving the strength of these correlations. Additionally, the accuracy of the model used in predicting AQI should remain roughly the same as our current model. Otherwise, it kind of defeats the purpose of switching to another model to find these correlations. For example, a naive implementation of a model involves using linear regression. Although now the correlations can be derived, the accuracy of the model decreases significantly, which means these correlations are not very useful. If we can get the correlations between each of our explanatory variables and AQI, we are able to know which variables have strong impacts on the AQI level, making it possible to reduce AQI level by manipulating and controlling these variables.