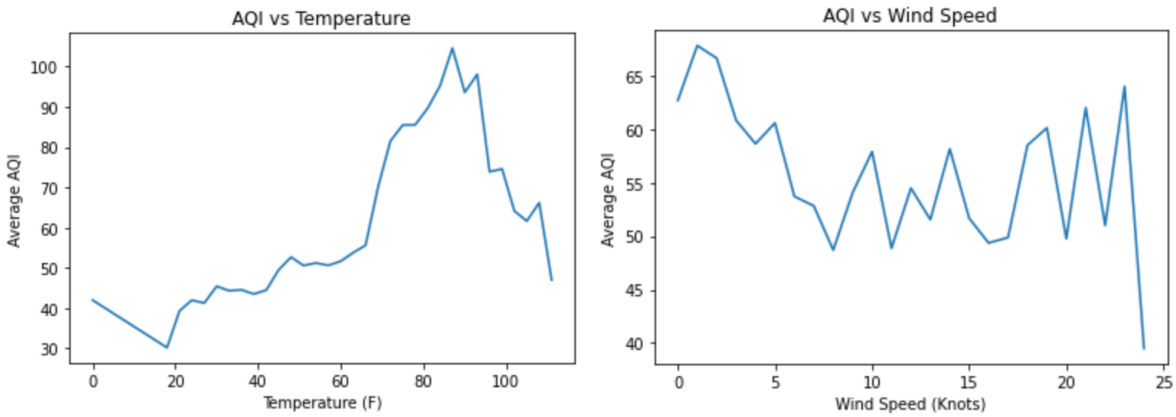
AQI Project Design Document

Author: Winnie Xiao, David Long, Brian Zheng

1. Overview:

Air quality is a great concern in our lives. It is deeply connected with our health and daily wellbeing. Polluted air hinders our outdoor activities, and long-term exposure to it can lead to various respiratory conditions and endanger our health. Thus, studying the air quality and factors that impact it is of great significance. In this project, we will dive deep into what influences AQI(Air Quality Index) level, and in specific, we will investigate the relationship between population density and AQI, using data obtained from the USA EPA in the state of California in 2020.

2. Part1 EDA Summary

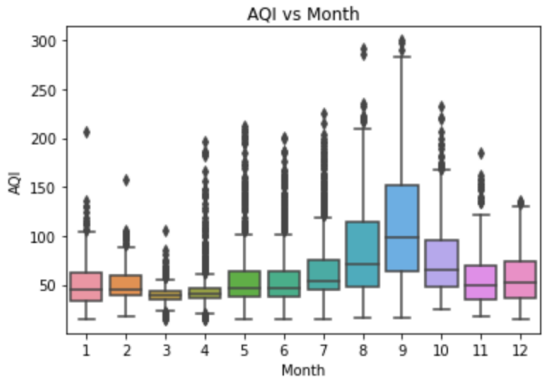


1. According to our box plot. The median AQI is generally low from March to June--about 35, though there are occasionally some outliers. For the rest of the year, the median AQI is relatively high, ranging from 50 to over 100. Then we attempt to visualize the relationship between wind and AQI. We used 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 modeling the data with linear regression, we also used line plots to visualize the correlation found utilizing linear regression.

3. The first line plot visualizes 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 visualizes 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.



3. Hypothesis: ***Population density is positively correlated with AQI level.***

4. Motivation

After some exploration with the EPS dataset, we found that the heatmap(fig.1) is a very straightforward illustration for the relationship between geo-location and AQI, and is worthy to be investigated further.

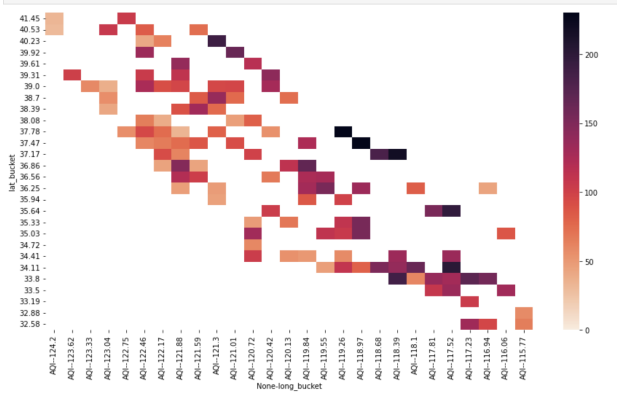


Fig. 1: Heatmap for California's Median AQI in September 2020

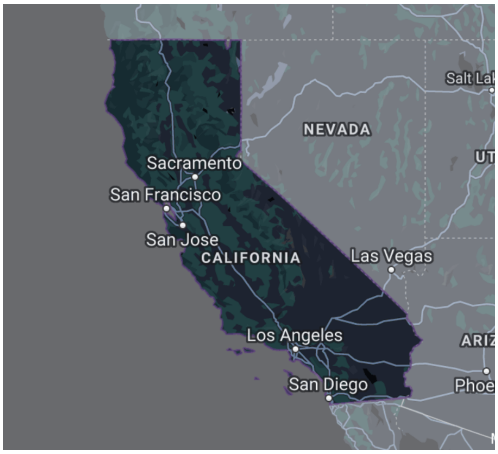


Fig. 2: Map of California

The heat map visualizes the geographical locations in California in relation to the AQI level that were recorded in September of 2020. By comparing the heatmap (Fig.1) of California and AQI levels with the map of California from Google Maps (Fig.2), we found that the AQI levels tend to be higher in locations of the major city areas: Los Angeles, San Diego, and the Bay Area. However, we do recognize that there are higher AQI levels around the western border of California where cities are not located. After some background research, we found out that it's due to the large amount of wildfires that occurred in California in September 2020. Thus, we still remain curious on our intuitive hypothesis on the relationship between population density and AQI levels.

5. Methodology

5.1 Data:

This project is primarily based on the data samples from the US EPA data. Files that are included in this dataset are: annual_county_aqi, aqs_sites, daily_co, daily_county_aqi, daily_no2, daily_ozone, daily_so2, daily_temp, and daily_wind. For the purpose of this experiment, we will also need the data on population density levels for each US county, which can be obtained from https://www.california-demographics.com/counties_by_population.

How do we scientifically split the data so that we can correctly examine the relationship between a region's AQI level and its population? We decided to use the measuring unit of counties and the variable of population density to best validate the comparison. This decision is made after recognizing how data about population density for each county can be easily found to avoid the issue of missing values. Also by using population density, we can best standardize all of our measurements within a geographical area.

5.2 Experiment:

To test our hypothesis, we will conduct an experiment to discover the relationship between AQI levels and the population density of the site. To be more specific, we will collect data on counties of the United States as the unit of geographical measurement, and narrow down the sample frame to the year of 2020. With the data, we will establish a linear regression model that best predicts the AQI with the given population density. As a result, the regression would give the best estimated values. In addition, we will conduct a multivariate regression by adding more variables such as wind and temperature to see if it produces a better regression result, and to check which variable has the most effect on AQI.

5.3 Evaluation:

We will plot a scatterplot to see if there is a clear association between AQI and population density. Additionally, we will divide our data into a training set and a test set. We train our model using data from the training set and do cross-validation to improve the accuracy of our regression model. Then, we compute the estimated values given by the model on the test set. More rigorously, we will validate or reject our hypothesis by evaluating the estimator with the significance level of 95% .

6. Further Questions:

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?