Portland's Air Quality and the Drivers of Clean Air

Jake Warmflash

Willamette University MSDS

August 1, 2024

Introduction

In their journal article published in *The American Journal of Public Health*, Jacobs, Burgess, and Abbott tell the story of the Donora Pennsylvania Smog crisis. On October 30, 1948, the Donora High School Football team played through a dense smog to complete the game with hundreds of fans in the audience, despite very poor visibility. The team, fueled by resilience, took pride in playing through poor conditions, a testament to the high spirit of the small town. Soon after, calls to the town's medical offices began flooding in, complaining of difficulty breathing and respiratory issues. Donora, Pennsylvania, was a town of metalworks, built by the American Steel and Wire company and the Donora Zinc Works company, which made up major parts of the town's economy. The heavy smog and pollution clouds that covered the sky had been viewed as a sign of prosperity, owing to the industrial might that powered their economy. Within just twelve hours, seventeen people would be dead, 1440 seriously affected, and 4470 with mild to moderate conditions—almost half the town's working population (Jacobs, Burgess, Abbott).

This event, known as the Donora Smog of 1948, prompted the country to take a closer look at the negative impacts of air pollution. Widespread debate surrounding the event led to the first legislation aimed at regulating the air quality within the United States, ushering in a new era of tracking, combatting, and reversing the ill effects of poor air quality. The quality of air we breathe has direct impacts on our health. We must understand the factors that contribute to poor air quality and how we individually and collectively contribute to these changes. Until we can visualize the impact we have on our atmosphere, we will continue behavior that negatively impacts the air around us.

In this project, I will focus on the key factors influencing air quality in Portland, Oregon. I aim to understand the complex interplay between various environmental and human-made factors that contribute to high air pollution levels. Initially, I was surprised to discover that Portland, the city I reside in, has some of the best air quality for a city of its size in the United States. This led me to narrow my focus to understanding the factors that contribute to these favorable outcomes in this city.

This project aims to develop and validate machine learning models to analyze the various factors influencing air quality. By focusing on Portland in comparison to other large US cities, I hope to find things Portland does that lead to the greater AQI outcomes. This approach will consider a range of variables, including meteorological conditions, pollution sources, transit systems, park land, and livestock. Through this analysis, I aim to provide actionable insights and recommendations for sustaining and

improving air quality. By examining both the contributors to clean air and the sources of pollution, we can understand the factors affecting air quality and develop comprehensive strategies for enhancement.

Background

Federal regulation of air quality in the United States began in 1955 with the Air Pollution Control Act. This new piece of legislation provided funding for initial research into air quality and pollution in the US. Building off this and privately funded research, Congress passed the Clean Air Act of 1963, establishing the first federal regulation for controlling air pollution. This act established a new federal program within the US Public Health Service, dedicated to the monitoring and control of air quality. In 1967, Congress passed the Air Quality Act, which introduced more federal oversight and enforcement policies, allowing extensive monitoring of interstate air pollution. This all led to the passage of the 1970 Clean Air Act, aimed at restricting and regulating emissions, measuring and reducing pollutant particles, and addressing upcoming pollution threats (Environmental Protection Agency).

Also established in 1970, the Environmental Protection Agency (EPA) implemented and monitored the requirements established by these rulings. The EPA's authority extended beyond federal lands and roads to include all companies operating within the United States. Enforcement authority was expanded to allow upholding these established standards and prevent companies from circumventing the law. Much of the improvement in the quality of air in the US over the past fifty years can be attributed to these regulations. In 1990, when deaths due to air quality were first measured, an estimated 135,000 Americans died. By 2010, that number had dropped to 71,000 (Zhang et al.). Despite the significant improvements led by the federal guidelines of the late 70s, nearly four in ten Americans still live in places where they are exposed to unhealthy air (American Lung Association).

In 1999, the EPA developed the Air Quality Index (AQI), creating an easily understood measurement of air quality. The AQI measures air pollution levels on a scale from 0 to 500, divided into six categories. A score of 0 to 50 represents good air quality which poses little or no risk to those breathing it in, while a score above 300 signifies emergency conditions, an extremely high risk which impacts everyone. This measurement is mainly derived from five major pollutants: ozone (O3), particulate matter (2.5μm and 10μm), carbon monoxide (CO), nitrogen dioxide (NO2), and sulfur dioxide (SO2) (Airnow.gov). Poor air quality has been linked to a variety of diseases including respiratory infections, stroke, heart disease, lung cancer, and chronic obstructive pulmonary disease, among others (World Health Organization). An estimated seven million premature deaths annually can be attributed to air pollution, which equates to a global mean loss of life expectancy of 2.9 years, making it the largest

environmental risk factor for disease and premature death (Fuller, Landrigan, Balakrishnan, et al.). Thus, it is important to understand factors that contribute to poor air quality, and outcomes that can be attributed to the state of the AQI.

These harmful factors can originate from a variety of sources. Anything that releases a foreign substance into the air can lower the AQI. This includes smoking, vehicle exhaust, combustion processes for production and manufacturing, household cleaning products, appliances, central air and heating systems, agriculture pesticides, livestock, shipping and transportation, and much more. Individually, we can reduce our individual contributions by lowering our reliance on personal vehicles, watching our power usage, supporting companies that monitor and address their emissions, and more. However, there are many factors beyond our control. Larger pollutant sources, such as manufacturing and transportation, are often regulated to some extent but may still release significant amounts of pollutants into the atmosphere which we as individuals have no say over (Manisalidis, Stavropoulou, Stavropoulos, Bezirtzoglou). It is challenging to restrict and watch our personal contributions to the polluting of the environment without worrying about what others are doing. Measuring and analyzing the impact these pollutants have on air quality is a crucial step towards addressing these issues.

Data Ethics

The ethical implications of conducting this analysis should always be considered and addressed. Air quality has historically impacted lower income communities significantly more. Marginalized communities may be built near major pollution sources such as factories or highways. In Portland, for example, Interstate 5 completed construction in 1966. It was part of an Oregon State Highway Department (OSHD) development, a project run by the state as a part of the Federal Aid Highway Act. The highway was built on the existing Minnesota Avenue, cutting North Portland, the heart of Portland's African American community in half. It was decided to avoid affecting the higher income downtown properties. In fact, the OSHD contracted work on the Minnesota Ave portion of the highway to private contractors, in an effort to avoid a political battle (Oregon Encyclopedia). Today, the neighborhoods surrounding the large highway have worse AQI outcomes than those further away, all other factors equal (Shandas and George). Any recommendations given by this analysis should properly consider the ramifications of those impacted.

Another important ethical issue to be aware of is data transparency. Data can be used to push a specific narrative without faking or editing the specific data points. By leaving

out important information, pushing correlated data as causation to push an agenda, or by purposely misconstruing what the data means, one can create a narrative that can misinform or deceive readers. All analysis methodology must be openly discussed to ensure full understanding by audiences. This leads to credibility, reproducibility, and trust in the methods and conclusions of any piece of data analysis. All data is sourced from credible online sources that are open about their collection methods. Data privacy is addressed, and no personalized information is given. Limitations are also addressed allowing readers to understand where any uncertainty may be.

Methods

Python will be the primary programming language used to conduct this analysis. I will also use the R language in statistical applications. To perform this analysis, I will use the NumPy and Pandas libraries for data manipulation, Matplotlib for visualization, and the time series forecasting algorithm Prophet by Meta. I will address data inconsistencies, missing values and ensure that data is in a tidy format. I may need to normalize or standardize data if necessary and create new features through aggregation to enhance the model's performance.

Prophet

Prophet is an open-source forecasting tool developed by Meta, designed for forecasting time series data. It is suited for datasets with strong seasonal, monthly, weekly, or daily patterns, and it handles missing data and outliers well. I utilized prophet to gain a quick understanding of the AQI patterns, seeking to understand basic trends before conducting a more thorough analysis.

Key features of Prophet include seasonality detection and holiday incorporation, while providing easy use and understanding for users. I use this software to get complex understanding from simple applications.

To conduct this analysis, I prepare data into a two-column table, with columns date and AQI. Prophet uses the trends of past data to highlight similarities over days of the year, weeks, months, and seasons. From this, prophet is able to generate its predictions, cross validate, and give performance metrics such as mean absolute percentage error to quantify the accuracy of the results

Hypothesis Testing

Hypothesis tests for significance are conducted to understand the significance of differences in datasets. Statistics are done in R using a variety of statistical methods and measurements:

- *Null Hypothesis*: This serves as the baseline of a hypothesis test. It represents the idea that there is no effect, difference, or relationship between tested variables. Any observed differences are due to random chance.
- *Alternate Hypothesis*: On the other hand, the alternative hypothesis suggests a potential effect, difference, or relationship between tested variables exists. It reflects what is hoped to be found by the data.

- *Independence*: The idea that data in one sample or population has no impact on the data in another sample or population.
- *Normal Distribution*: A continuous probability distribution symmetric around the mean. It is characterized by mean and standard deviation.
- *Normal QQ Plot*: A plot to measure the normality of a distribution. It compares the quantiles of a dataset against the quantiles of a theoretical normal distribution. Data is mapped along a y=x trendline. Deviations of the data from the trendline indicate non normality. An "S" shaped bend suggests data with heavy or light tails. A convex or concave bend suggests skewness in the distribution.
- *Histogram*: Plot that displayed data frequencies of values within specified intervals known as bins. Histograms are used to visualize the shape and spread of a distribution.
- *Welch Two Sample t-test*: This type of hypothesis test measures the statistical significance of the difference in mean of the two samples. Unlike the Student's two sample t-test, the Welch two-sample t-test is used in cases where the groups being compared have unequal variances. A Welch two sample t-test includes the following:
    - *t-Statistic (t)*: Difference between sample means relative to variance of samples. Larger t-statistics indicate larger differences between means.
    - *Degrees of Freedom (df)*: Amount of information available to estimate variance of sample means.
    - *p-Value*: The probability of observing the sample data or something more extreme if the null hypothesis is true. A small p-value (< 0.05) suggests the observed difference is statistically significant.
    - *95 Percent Confidence Interval*: Range in which the true difference in means is 95% likely to lie within.

Machine Learning

Machine learning is done in scikit-learn, an open-source ML library built in Python. It is built on top of other common Python libraries such as NumPy, Pandas, and Matplotlib. The following tools are used:

- *Train Test Split:* Splits datasets randomly into a training and testing dataset, split up by a specified ratio.
- *One Hot Encoder*: Transforms categorical data into a binary matrix. Each category is represented as a vector where one element is tagged with a 1 and all

others are tagged as 0. This is done for certain machine learning algorithms that require all discrete data.

- *Standard Scaler*: Used to normalize or standardize numerical data. Certain machine learning algorithms are affected by unscaled data and therefore require normalized data.
- *Transformer*: Can be used to change multiple variables in one step in methods such as One Hot Encoder and Standard Scaler.
- *Pipeline*: Allows multiple functions in a series of steps that can include transformers and models. Incorporates chaining to supply the output of one step as the input of the next step.
- *Feature Selection*: Chooses a subset of most relevant features, with the ability to specify how many features.
- *Hyperparameter Optimization*: Selects the best set of hyperparameters in a model. These hyperparameters are set before training and control various aspects of the training process and model behavior. The optimization method used is a Randomized search. This evaluates a fixed number of random combinations from a specified distribution. This is far more efficient than a Grid search which systematically checks every single combination of the specified hyperparameter distribution.
- *Cohen Kappa Score*: A statistical measure used to assess agreement between two classifiers when they categorize items into classes. In ML specifically, the Cohen Kappa Score is used to evaluate the agreement between predicted and actual labels in classification algorithms. It is measured on a scale from -1 to 1, where 1 represents perfect agreement in classifiers, 0 represents agreement being no better or worse than random chance, and -1 represents perfect disagreement.
- *Random Forest Model*: An ensemble model used in prediction tasks. It combines multiple decision trees and aggregates their predictions to improve performance and reduce overfitting.
- *Confusion Matrix*: Measures and visualizes the accuracy of a classification model. It breaks down raw numbers of correctly and incorrectly classified values.
- *Classification Report*: A detailed evaluation of the prediction model. Includes:
    - *Precision*: The ratio of correctly predicted positive observations to the total predicted positives
    - *Recall*: The ratio of correctly predicted positive observations to all observations
    - *F1 score*: Mean of precision and recall
    - *Support*: Number of observations

Data

The data for this project was initially scattered across multiple sources and required significant organization and compilation. The focus of this project is on the air quality in Portland, Oregon, so various data sources were aggregated and processed to compare a variety of air quality indicators. There is a significant amount of missing data due to incomplete collection which will be addressed via filling in and row dropping.

Air Quality Data

Air quality data, specifically AQI values, were obtained from the United States Environmental Protection Agency (EPA) pre generated data files. The AQI values are calculated daily, based on a variety of factors including criteria gasses and measured pollutant concentrations, and measures how harmful breathing the air is. AQI is classified into one of six categories from 'good' to 'hazardous', each having long term health effects associated with it. The files were given daily on a county wide basis, separated into different files by year. After stacking year outputs together in R, columns regarding date, location (state, county), AQI, and AQI category were brought into the database.

Meteorological Data

Historical weather data was also sourced from the EPA database, measured by thousands of weather stations across the country. Measurements tracked include temperature, wind speed, air pressure, and humidity. Temperature is measured in degrees Fahrenheit. Wind speed is measured in knots, which are defined as one nautical mile per hour (equivalent to approximately 1.15mph). Wind speed is important in air quality as winds can blow different pollutants around and move and spread wildfires. Pressure is measured in millibars, where 1013.25 millibars is the standard atmospheric pressure (Earth's pressure at mean sea level). Finally, humidity is measured in percent relative humidity. This is the amount of water vapor in the air as a percentage of the maximum amount of water vapor possible at a given temperature. Humidity can make it more difficult to breathe and sweat, make the air feel hotter than it is, and prevent air pollutants from dispersing as easily. This data was given daily by city, separated into different files by year. Measurements were taken hourly, but pre-aggregated in the source database, giving an average value over the twenty-four hours and a maximum value. After stacking years in R, columns regarding date, location (state, city), weather observation average and maximum, and observation unit were brought into the database.

Pollution Source Data

Pollution data was again sourced from the EPA database, separated by criteria gasses (CO, NO2, O3, SO2), Toxins (lead), and particulate matter (PM2.5 and PM10). Criteria gas Carbon Monoxide (CO) is measured in parts per million and is especially dangerous as it is both colorless and odorless. CO binds to hemoglobin in the blood, making the transportation of oxygen around the body more difficult. Nitrogen Dioxide (NO2) is dangerous to breathe in at high levels. It can cause swelling in the throat, burning, reduced oxygenation of body tissues, and fluid buildup in the lungs. It is released in many common combustion reactions including in cars, coal plants, and cigarettes. It is measured in parts per billion. Ozone (O3) can harm our ability to breathe, especially in older people, children, and people with asthma. It is measured in parts per million. Sulfur Dioxide (SO2), measured in parts per billion, can irritate the eyes, mucous membranes, skin, and the respiratory tract. Lead is a toxin which can increase the risk of high blood pressure, cardiovascular problems, and complications during pregnancy. While exposure has gone down significantly in recent decades after use in gasoline, it still remains a dangerous toxin to breathe in. It is measured in micrograms per cubic meter. PM2.5 and PM10 are particulate matter, small inhalable particles with diameters of 2.5 microns or smaller, and 10 microns or smaller respectively. PM2.5 includes all sorts of common particles, metals, and organic compounds. PM10 includes dust, pollen, molds, and other larger (but still very small) particles. Due to the variability of particles included in the PM classification, there are a wide range of negative health impacts that come from breathing in these particles. PM2.5 and PM10 are measured in micrograms per cubic meter.

This data was given daily by city, separated into different files by year. They are sourced from thousands of individual sources, which measure various selections of these pollution sources. Because of the variety of different pollutants being measured, there was a significant amount of missing data, especially from small towns. Measurements were taken hourly, pre-compiled into a daily average and maximum. After stacking years in R, columns regarding date, location (state, city), pollutant observation average and maximum, and observation unit were brought into the database.

Public Transit Data

Information on motor buses was taken from the National Transit Database, produced by the Federal Transit Administration. It includes information about bus systems and ridership by city, separated by year. Data is recorded yearly, encompassing annual

totals for information such as number of buses, total revenue, passengers, and miles driven for the respective city transit systems. Information was given in yearly CSVs, separated by the city transit system. For cities with multiple systems, data was combined. Only motorbus data was compiled, which may not be reflective of cities with other large methods of public transportation, such as the New York subway system. Only public transit information was able to be collected. Accessible data on cars and trucks could not be found. This will be further discussed in the limitations section of the conclusion. After stacking years in R, columns regarding date, location (state, city), transport mode, number of buses, passenger trips in hours and miles, and bus operations in hours and miles were brought into the database.

Land Area Data

City and county area data was collected from the 2020 U.S. Census, collected by the U.S. Census Bureau. It is measured in km^2. City, county, state, city area, and county area columns were brought into the database.

Park Area Data

Park area data is sourced from the Trust for Public Land's 2024 annual City Park Survey. This information is collected via survey from a variety of park agencies including individual parks and recreation agencies, the National Park Service, the Bureau of Land Management, regional park authorities, private conservancies, watershed management agencies, private park services, and more. Totals are compiled by the Trust for Public Land and measured in acres. Only location (city, state) and park area columns are brought into the database.

Livestock Data

Information on livestock numbers was collected on a county wide basis from the 2017 United States Census of Agriculture by the National Agricultural Statistics Service within the United States Department of Agriculture. Cattle include cows and bulls raised for meat consumption, cows for milk production, and calves. Hogs includes all pigs, adult and adolescent. Although there are colloquial differences for the animals called hogs and pigs, they refer to the same species, though 'hogs' generally refers to the larger members of the species. Sheep includes sheep raised for wool production and lambs. Numbers may have slight inaccuracies for areas with low numbers to prevent disclosing identifying information and are marked with a 'D'. All of these data points

will be replaced with zero. Location (county, state) and livestock varieties (cattle and calves, hogs, sheep and lambs) columns are collected and added to the database.

Population Data

Data on population and population density was sourced from the Simplemaps United States Cities database, which is built from multiple sources including the U.S. Geological Survey and the U.S. Census Bureau. Data was last updated on May 6, 2024, reflecting very up to date information.

Data Organization

Given the raw data available, the table structure was simplified compared to the original data sources. Data was organized around the air_quality table. This table tracks AQI, pollutant, weather and toxin data daily for each location. It includes all 1438 locations with a line for each of the 2922 days in the eight-year time period, reaching a total of over 4.2 million rows. Location is split into a separate table to reduce repeated data. This table lists cities, labeled with their city name, county, and state. It includes metropolitan area population data given in raw numbers and as a density. Additionally, this table includes the city area in square kilometers and park area in acres. The aqi_category table is a short list of AQI value categories (Good, Unhealthy, Hazardous, etc.) with their respective AQI value range as minimum and maximum values. The yearly_transit table is connected to the location table and gives the information for the transit system of the respective city during the specified year. Finally, the livestock table, also connected via locations, includes counts for cattle, hogs, and sheep.

The central table has a compound primary key composed of location_id and date. Each other table has a serialized primary key, which are used to connect to each other. Several additional indexes are included on columns that will be queried often. Finally, constraints have been added to limit unusual or impossible data. For example, an AQI value less than 0 or greater than 500 would be impossible and thus would be caught by the constraint.

Tracking these identifiers independently allows for accurate analysis of changes over time and across different areas and allows adding new information should I need to update the database. Figure 1 illustrates the resulting ERD diagram using drawSQL.

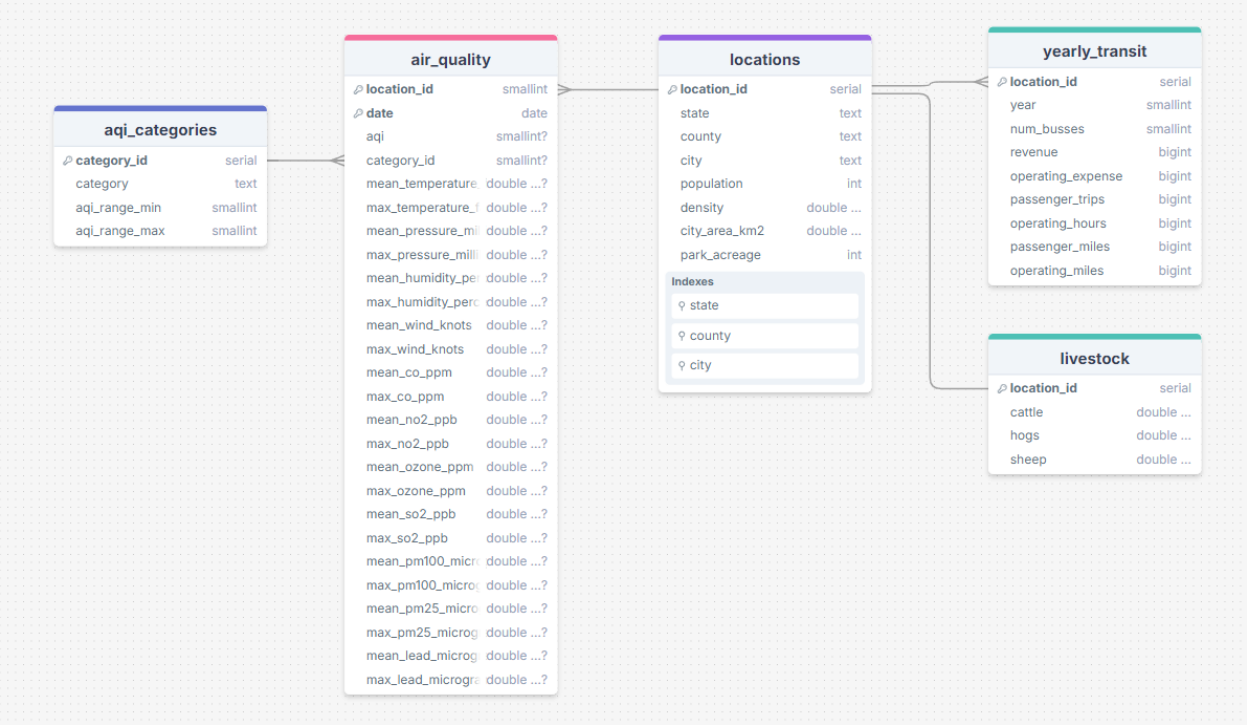

Figure 1: ERD diagram

Prophet AQI Trend Forecasting and Statistical Testing

To gain an understanding of how Portland's AQI differs from other large cities, we can start by running a hypothesis test to determine the significance.

Initially we can run a Two-samples t-Test to show that Portland's AQI average is statistically greater than the average AQI of all large metropolitan areas within the US. I compare the sample of Portland AQI datapoints in the time period with the sample of all metropolitan areas with a population greater than one million.

Null Hypothesis: The Portland AQI is greater than or equal to the AQI of all large metro cities in the US.

Alternate Hypothesis: The Portland AQI is less than the AQI of all large metro cities in the US.

Assumptions:

1. Simple Random Sample: Data is a simple random sample. I have selected 10% of values from each population set.
2. Independence: AQI in Portland does not affect AQI in the rest of the country. However, the dataset for the large metro areas does include Portland, so Portland AQI will be repeated within both population groups. Both groups are largely independent, though this should be noted.
3. Normal Distribution:

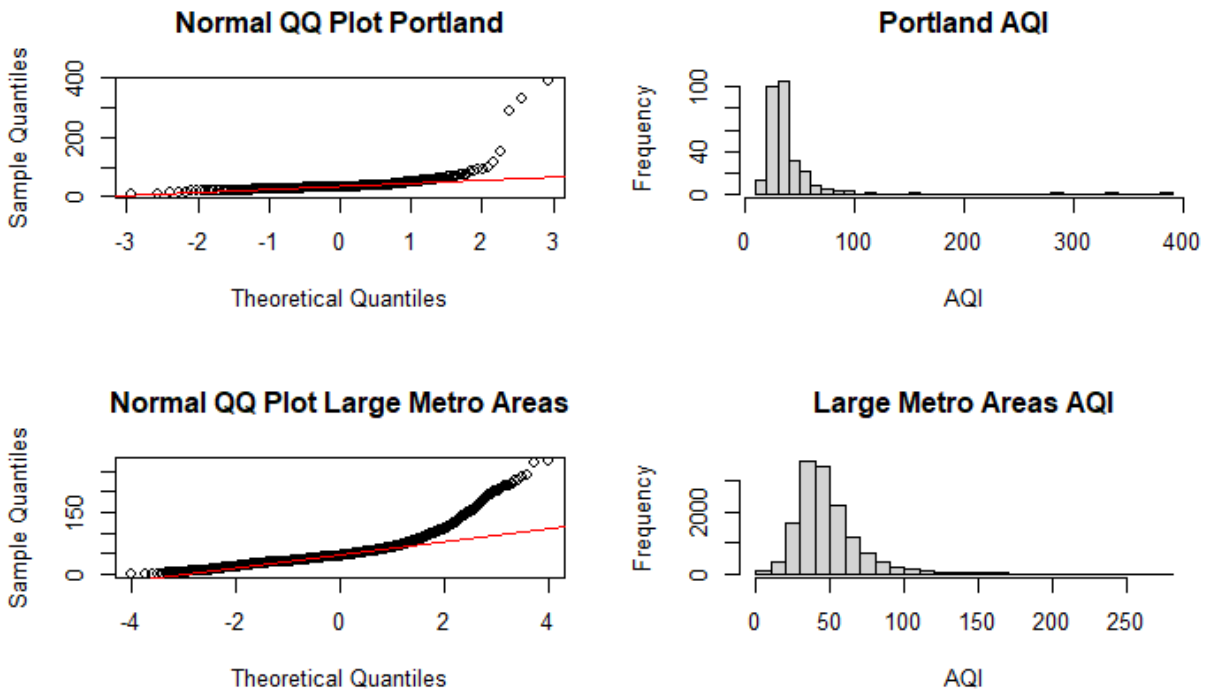

Figure 2: PDX Vs Large Metro Areas Distribution

From the QQ plots and histograms, we can see both datasets are clearly not normally distributed. They have long tails to the right. However, since the tails have such low frequencies and the samples are very large, this likely will not impact the results.

The partial violations of the assumptions in the t-test in the analysis suggest that the conclusions should be considered with a degree of caution.

Two-Sample t-Test Portland AQI vs Large Metro Areas AQI

```
    Welch Two Sample t-test

data:  pdx_aqi_sample and lmaqi_sample
t = -3.2311, df = 295.69, p-value = 0.0006862
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.378713
sample estimates:
mean of x mean of y
 41.36301  48.26777
```

Based on the low P-value of $6.862*10^{-4}$, we can safely reject the null hypothesis. I conclude that Portland's AQI mean is not greater than or equal to the AQI of all large metropolitan areas. This agrees with my initial observations on the greater AQI outcomes of Portland and specifies the significance of this statistically.

Data Forecasting with Prophet

Prophet by Meta is used as time series prediction to estimate and map trends based on the data. I use this to see how AQI trends vary by day, month, and year. It is also able to give us a forecast for a given period after the end of the data which I can analyze and use to anticipate future AQI values.

To use the package, data must be in the format of a two-column graph, with the first column being the date data, and the second being the variable being mapped and predicted. In this case, this predicted variable is AQI.

The package allows a future period to be generated, which can be specified and added to the end of the time data. It can then predict future AQI values for the new data period.

The graph below (figure 3), shows the supplied data with the forecast over the future period. The actual values are shown with the black datapoints, and the blue line represents the prediction. The upper and lower bounds of the error are represented with the transparent blue area. Each year, the data spikes during the late summer to early fall. Even more significant is the large spike in September 2020. What caused it? How significant was it?

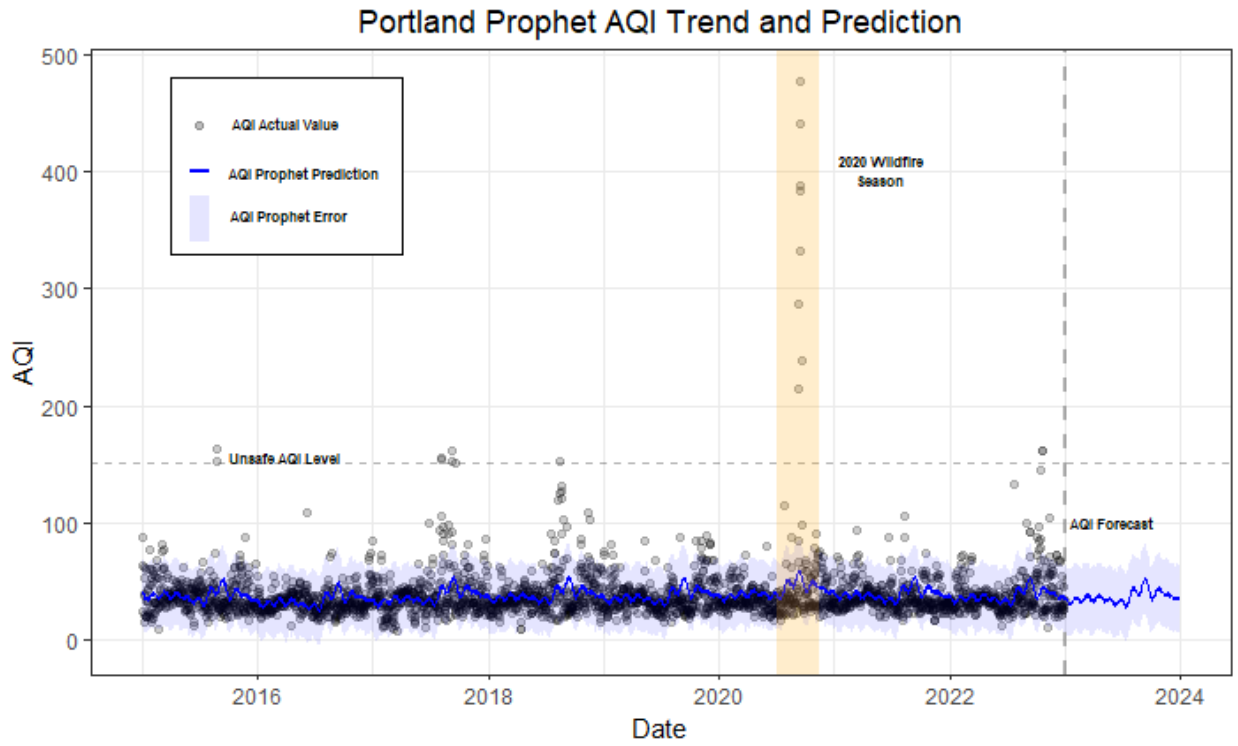

Figure 3: Prophet Trend and Prediction with 2020 Wildfire Season Highlight

2020 Wildfires

In September 2020, a wildfire ravaged the state of Oregon, as well as many other areas of the United States and Canada. The fires burned more than one million acres of land, destroying thousands of homes, and killing 11 people. 500,000 Oregonians were on evacuation alert, and 40,000 were actually forced to leave (Oregon Department of Emergency Management). Anyone around during that time will recall the orange skies, thick atmosphere, and strong smoke smell, but how unusual was this period actually?

To understand, I will conduct a Two-samples t-Test to see whether or not this month had greater than usual AQI.

Null Hypothesis: September 2020 AQI is less than or equal to AQI of entire period in Portland

Alternate Hypothesis: September 2020 AQI is greater AQI of entire period in Portland

Assumptions:

1. Simple Random Sample: Data is not a simple random sample. Since there are so few datapoints for the September 2020 population (30 datapoints), taking a 10%

sample may not be suitable to accurately capture the variance of this month. Thus, it was decided to use the entire population as the sample. A 10% sample will be taken of the entire Portland data population.

2. Independence: AQI in September 2020 does not affect AQI in the rest of the timeframe.

3. Normal Distribution:

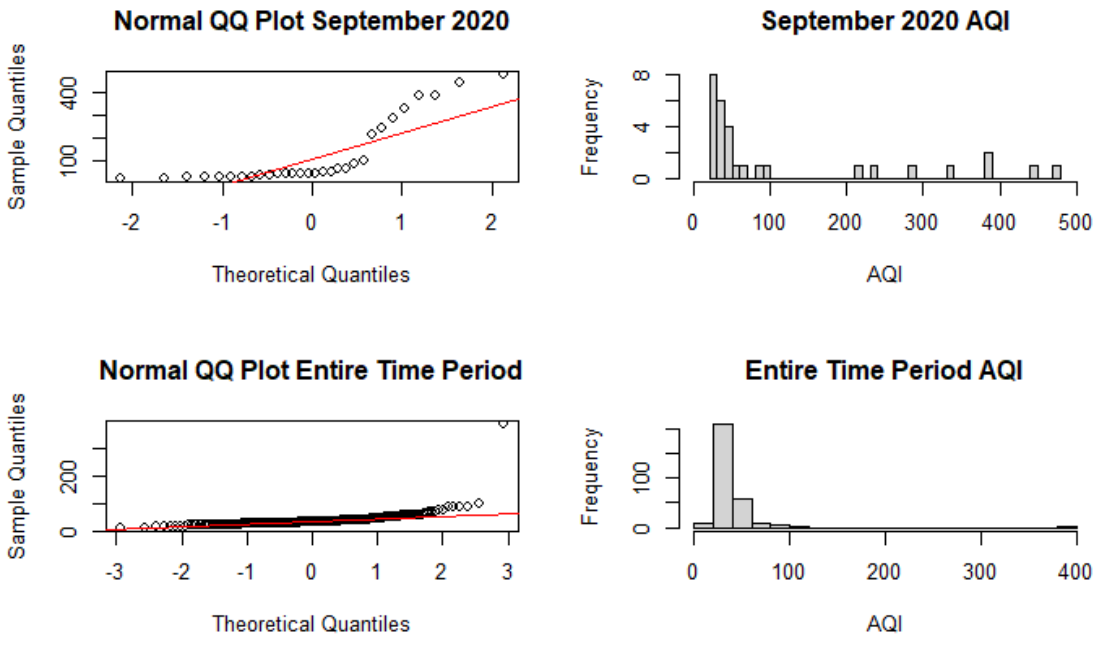

Figure 4: Wildfire Distribution

From the QQ plots and histograms, we can see both datasets are clearly not normally distributed. They have long tails to the right, and the September 2020 dataset is oddly shaped due to lack of data.

The partial violations of the assumptions in the t test in our analysis suggest that the conclusions should be considered with a degree of caution.

Two-Sample t-Test September 2020 vs Full Period of Portland AQI

```
    Welch Two Sample t-test

data:  pdx_sep_20_aqi and pdx_data_sample
t = 3.1563, df = 29.109, p-value = 0.00185
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 38.656    Inf
sample estimates:
mean of x mean of y
122.66667  38.94863
```

Based on the low p value of $1.85*10^{-3}$, we reject the null. I conclude that the September AQI in Portland is not less than or equal to the AQI for the entire period. The wildfire had a significant effect on the air quality, making it much more difficult to breathe. This aligns with the observation of the large spike during this period. We must do more to address and combat the wildfires that not only harm the air we breathe but cause long lasting damage to the local environment.

AQI Trends

Prophet's plot component function allows me to see specific trends including the total (entire eight years), weekly, and yearly trends in figure 5.

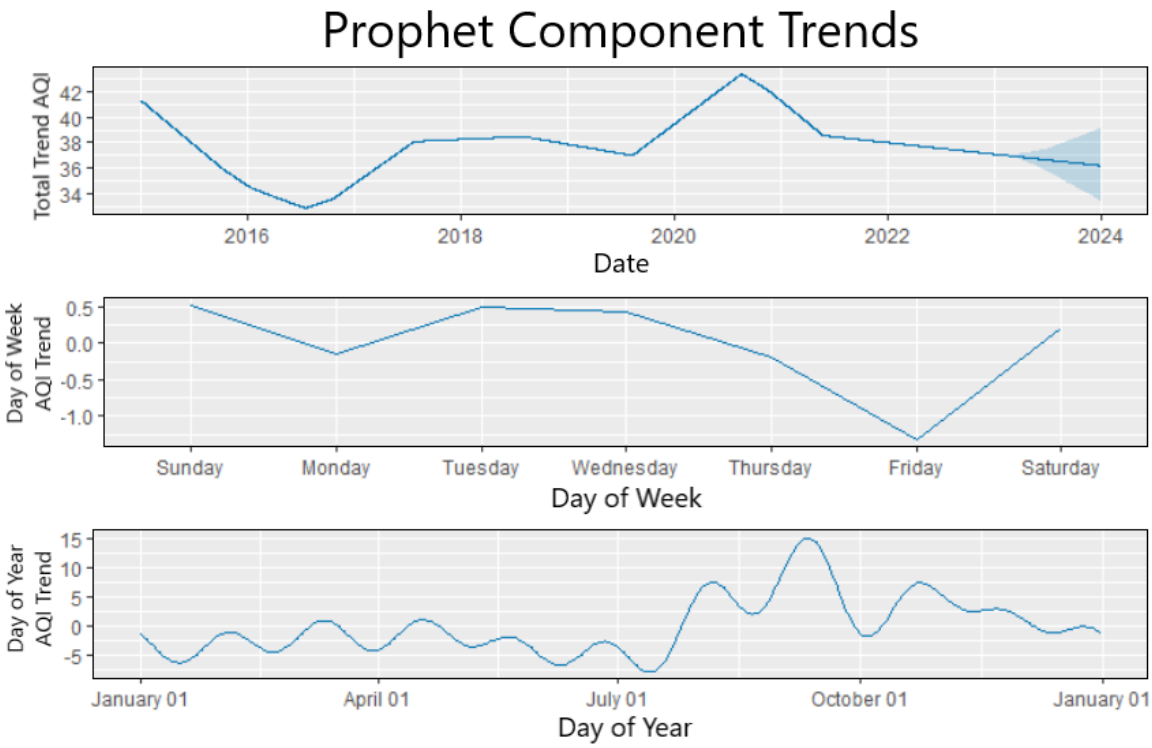

Figure 5: Prophet Component Trend Aggregations

This allows us to see how AQI changes by day. In the top full time span graph, it shows the potential spread of data for the predicted period with the transparent blue area. We can also see that day of the week tends to have very little impact on the trend (it may look significant, but it is only moving up and down less than 1.5 AQI between days). Finally, from the yearly graph we can see the consistent increase during the late summer to mid fall each year.

Cross Validation

How accurate are these predictions? Cross validation predicts over the period from the cutoff date up to specified date.

Cross validating the predictions allows us to create many estimates from a starting date up to an end date within the timeframe. In this instance, the tool predicts using start dates every half a year. It will predict AQI for that date to everyday up to a year after the

start date, giving us 365 estimates per start date. I then compare all estimations, seeing how accurate the prediction is for a number of days after the start date.

The tool allows us to measure the difference in predicted y and actual y with a variety of different measurements. In figure 6 below, I have selected mean absolute percent error, to show if the error on average increased as the prediction got further from the starting point.

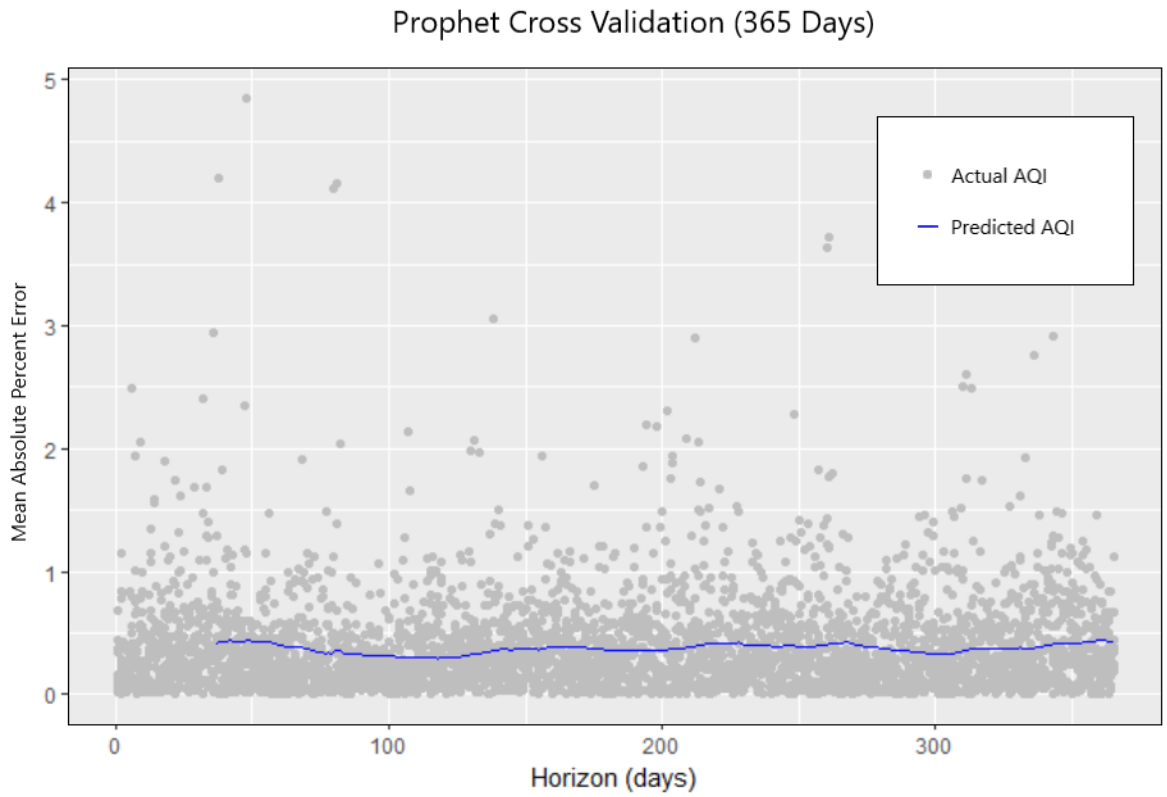

Figure 6: Prophet Error on Cross Validated Estimates

The light grey datapoints represent the mean absolute percent error. A datapoint with an x value of 200 and a y value of 1 means it was predicted for a period of 200 days after the cutoff date and was 1% off the actual value. We can see that most predictions are under 1%, making this a pretty decent estimation. The fact that they do not increase over time allows us to have a certain degree of confidence in the prediction even a year out. The blue line shows the mean of these predictions.

Portland's Transit System

As shown in the graphs and tests above, Portland has greater than normal AQI outcomes when compared with other cities of high populations. The AQI follows certain

yearly trends, with the notable exception in September 2020 due to the large wildfires. Using Prophet, I am able to predict AQI up to a year out of the final datapoint in the time period. What exactly is the reason for these outcomes in this city?

One hypothesis for Portland's better air quality outcomes is due to the increased focus on public transportation. Portland has placed high emphasis on utilizing public transit, with rates comparable to larger cities and higher than most other cities of its size. I will run a two-samples t-Test to see how Portland's rate of ridership compares with other large cities.

I have standardized rates by passenger miles ridden per person. This is the total number of miles ridden in a given year divided by the population. This allows us to properly compare cities with different population sizes.

Null Hypothesis: Transit ridership in Portland is less than or equal to transit ridership across the country (large metro areas only).

Alternate Hypothesis: Transit ridership in Portland is greater than transit ridership across the country (large metro areas only).

Assumptions:

1. Simple Random Sample: Data is not a simple random sample. Since there are so few datapoints for the Portland transit ridership population (8 datapoints), taking a 10% sample may not be suitable to accurately capture the variance of this month. Thus, it was decided to use the entire population as the sample. A 10% sample will be taken of the large cities transit population.
2. Independence: Transit ridership in Portland does not affect transit ridership in the rest of the country.
3. Normal Distribution:

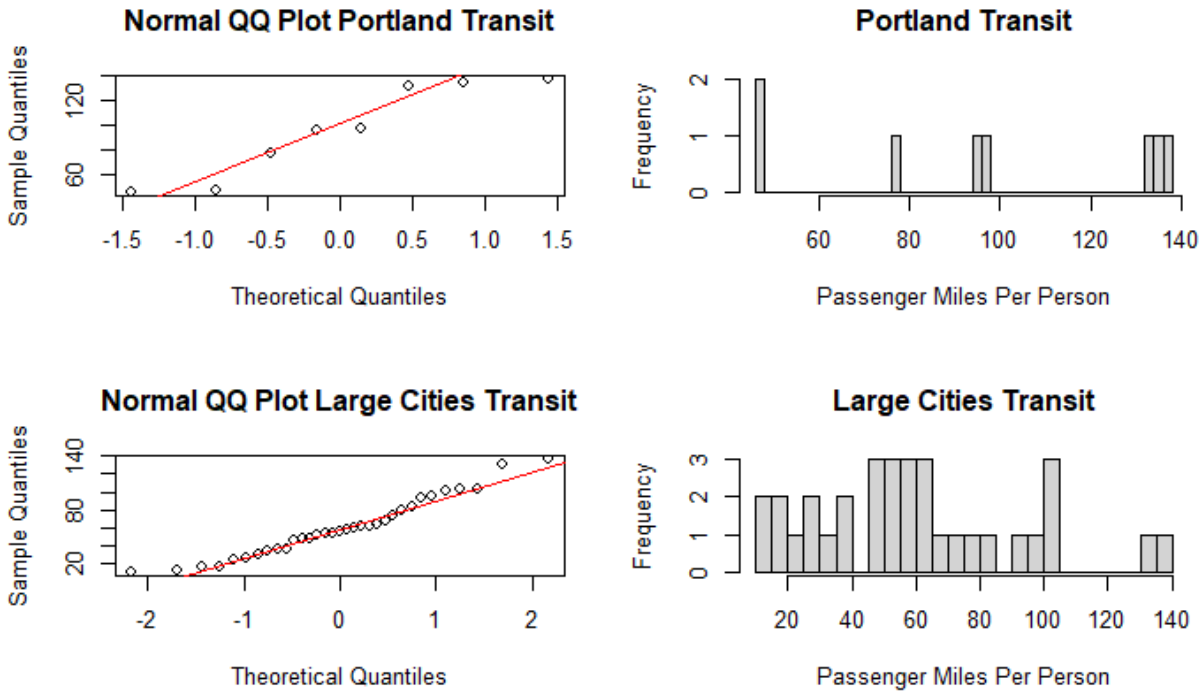

Figure 7: Transit Distribution

From the QQ plots and histograms, we can see both datasets are clearly not normally distributed. The Portland data has so few datapoints that it is hard to tell if it has a normal distribution or not. The large city transit data seems fairly normal, though it too does not have enough datapoints to visually approach a bell curve in the histogram.

The partial violations of the assumptions in the t test in the analysis suggest that the conclusions should be considered with a degree of caution.

Two-Sample t-Test Portland Transit Ridership vs Large Metro Area Ridership

```
    Welch Two Sample t-test

data:  pdx_transit$pass_miles_per_person and large_metro_transit_sample
t = 0.69167, df = 34.088, p-value = 0.2469
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -23.10491       Inf
sample estimates:
mean of x mean of y
 95.90601  79.91106
```

Based on the high p value of 0.2469, we fail to reject the null. We cannot conclude that Portland's transit ridership is different from the average transit ridership of other large metro areas across the country. Transit ridership may have some impact on AQI but not enough to make a statistical difference by itself. Likely, this is a situation of correlated variables. Cities that invest more in public infrastructure are more likely to make a concerted effort into being environmentally conscious.

It should also be noted that some cities have higher public transit numbers not reflected in the data. For example, in New York City, subway ridership is almost double that of its motor bus numbers. In this data, only motor bus numbers were included which may contribute to the lack of evidence for transit numbers influence on AQI. I will explore which features have a greater effect on AQI in the upcoming Machine Learning section.

Machine Learning with Scikit-Learn

Ultimately, I want to see which variables have the greatest impact on AQI. To do this I perform a machine learning analysis and create a prediction algorithm. As the AQI is defined by the four criteria gasses and particulate matter (PM10 and PM2.5), those should not be included in the algorithm. Thus, I start with all the other variables.

After data is cleaned, I am left with a total of eighteen cities across the country with a total metropolitan area population of greater than one million. To perform the ML prediction algorithm, AQI will be predicted. I will use existing AQI categories as classifiers. A feature selector is run on the set of all variables except for the six that define AQI (CO, NO2, O3, SO2, PM10, PM2.5). This chooses the best predictors of the dependent variable AQI.

Features Selected in Initial Model:

- City
- Month
- Population
- City Area
- Park Area
- Temperature
- Humidity
- Cattle

In pretesting, it was found that a Random Forest Model performed the best on this dataset. Therefore, this model type is used in all models.

An initial Random Forest Model run with the selected features returns the following Cohen Kappa score and accuracy:

```
Cohen Kappa Score:  0.4629814329018088
Accuracy:  0.7316276537833424
```

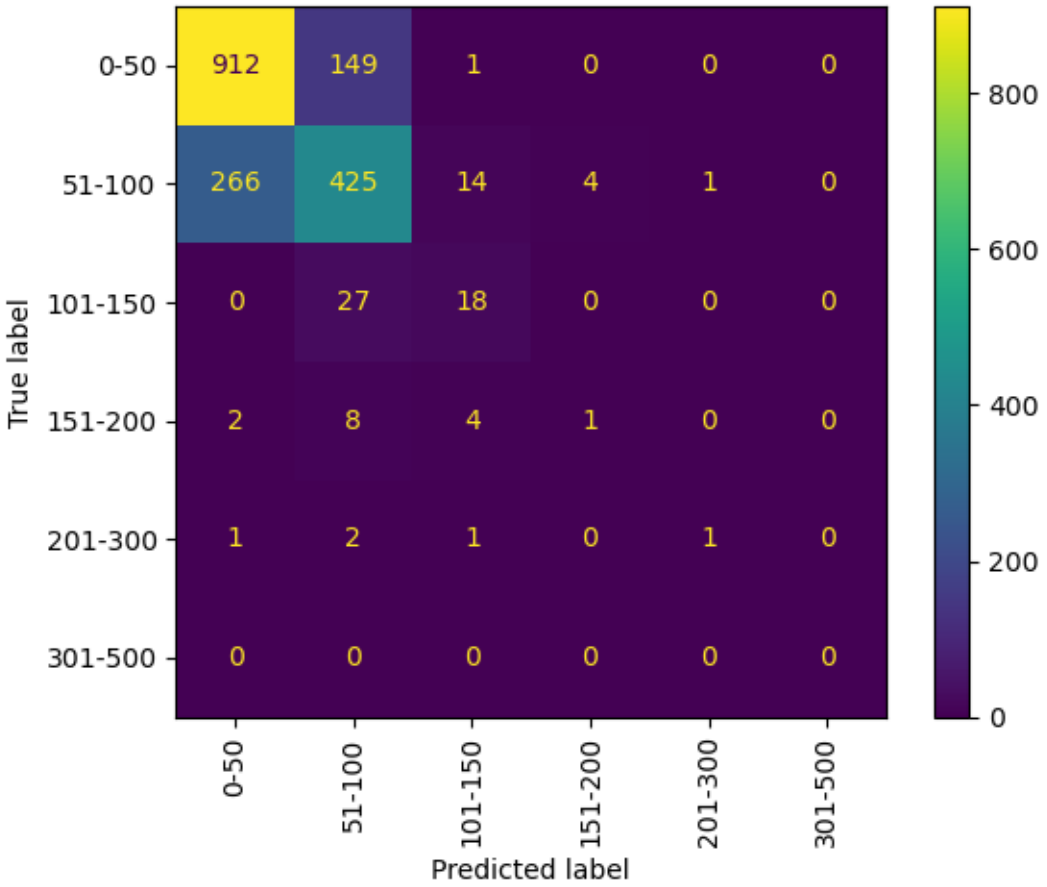

Figure 8: Confusion Matrix for Original AQI Categories

Looking only at Cohen Kappa score and accuracy, this seems like a good prediction model. However, the confusion matrix shows the data's skew, centered around the 0-100 range. Therefore, more bins and bin size combinations are tested in order to provide a more meaningful prediction.

After further testing the following bins were decided on:

- 0-25
- 26-50
- 51-75
- 76-100
- 101-200
- 201-500

As bins are made smaller, predictions become far less accurate. This bin size serves as a good balance between having enough bins to generate conclusions and maintaining decent accuracy. Different bin sizes also lead to different features being selected. The following features will be present in the final model:

- City
- Population
- City Area
- Park Area
- Temperature
- Humidity
- Cattle

Hyperparameter optimization will be done to further improve the model. A randomized search is run with 100 iterations. The following hyperparameters are optimized:

- Max Categories
- Min Frequency
- Max Depth
- Max Features
- Min Samples Leaf
- Min Samples Split
- Num Estimators
- Bootstrap

Table 1 below shows the selected hyperparameters.

Parameter	Value
RF_model__bootstrap	True
RF_model__max_depth	8

RF_model__max_features	log2
RF_model__min_samples_leaf	4
RF_model__min_samples_split	6
RF_model__n_estimators	147
aqi_transformer__categories__max_categories	10
aqi_transformer__categories__min_frequency	18

Table 1: Selected Hyperparameters following hyperparameter optimization

Using these hyperparameters, we reach an accuracy of over 70%.

```
Cohen Kappa Score:  0.4474900611032967
Accuracy:  0.7103973870440936
```

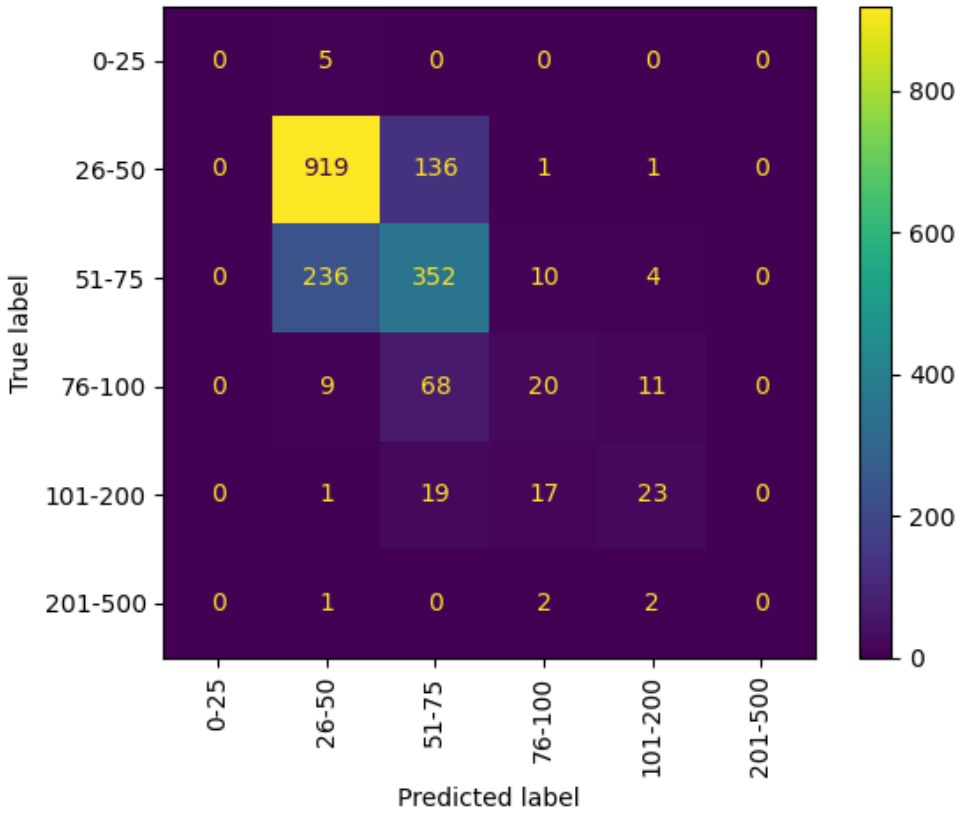

Figure 9: Final Model Confusion Matrix

We can use this model for the prediction of AQI, exploring the features that predict it, and use that to generate conclusions for what variables or systems should be addressed.

Below, table 2 shows a classification report of the model. Support represents the total number of data points in that bin. We can see that AQI groups with more data tend to be

more accurate (higher f1-score). This makes sense as the model has more of this data to train on. To make the model more accurate, more data from AQI values outside this 26-200 range will need to be collected.

	precision	recall	f1-score	support
0-25	0.000000	0.000000	0.000000	5.0
26-50	0.773899	0.880795	0.823894	1057.0
51-75	0.609982	0.548173	0.577428	602.0
76-100	0.393443	0.222222	0.284024	108.0
101-200	0.625000	0.333333	0.434783	60.0
201-500	0.000000	0.000000	0.000000	5.0

Table 2: Final Model Classification Report

We can also take a look at which features end up being most influential in the model. Below, figure 10 shows temperature and humidity being by far the most influential variables in the model. However, temperature and humidity are not variables we can directly impact easily, and therefore, the four other variables are where we should put our efforts.

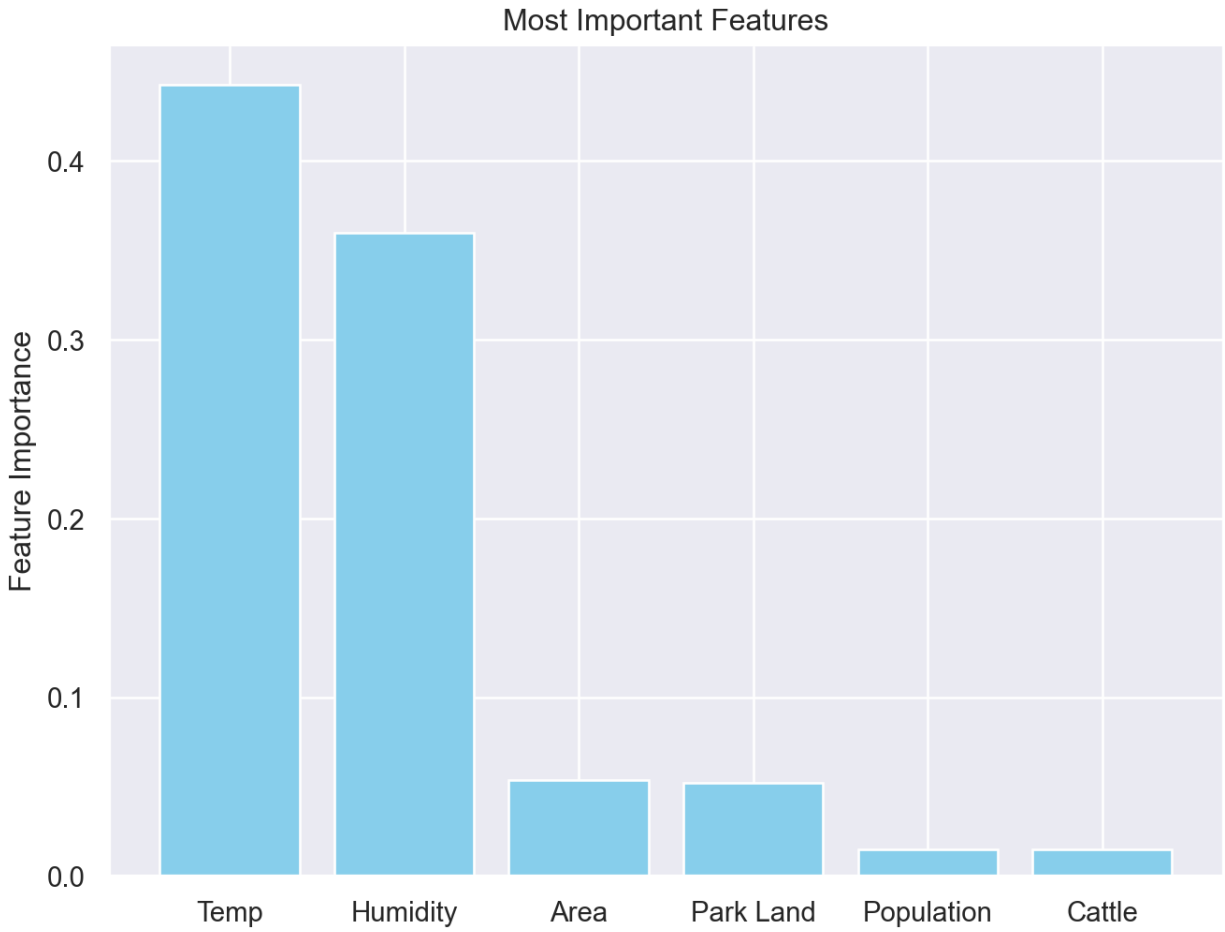

Figure 10: Feature Importance of Selected Variables in Final Model

Conclusion

After conducting my thorough research, I have landed on these specific recommendations.

AQI is impacted by six explanatory features:

- Temperature
- Humidity
- Park Land Area
- Number of Cattle
- City Area
- Population

Figure 11 shows how each of these features impacts AQI

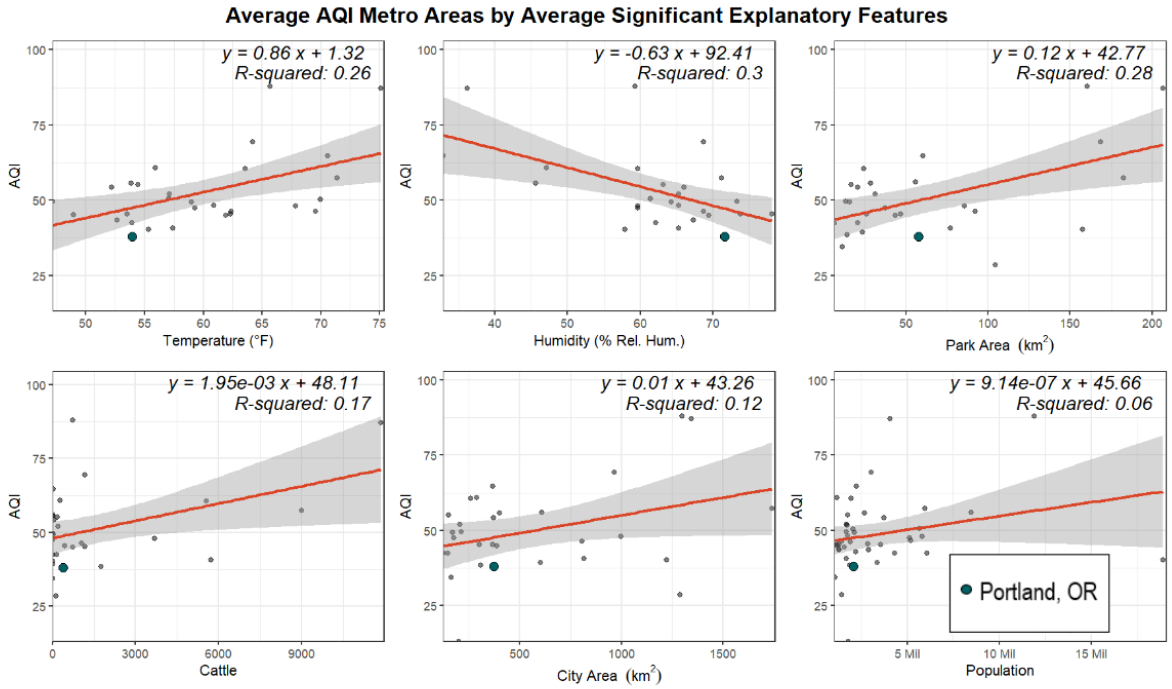

Figure 11: AQI explained by influential features, Portland highlighted

To improve AQI, a city should:

- Decrease temperatures
- Increase humidity
- Decrease park land area
- Decrease number of cattle
- Decrease city area
- Decrease population

Ialso found there are high seasonal trends that show up each year from the time series analysis models. AQI tends to be highest during late summer to early fall. We can attribute this increase to the increased temperatures and decreased humidity. As climate change raises temperatures and water sources dry up, the seasonal AQI increases will become more severe. We must be aware of the nature of air quality and how it changes at different parts of the year. We should understand how the AQI works and avoid being outside for too long when it reaches more dangerous levels.

We must focus on the variables that have the greatest impact on the AQI. Of these, we have little control over temperature and humidity, with the exception of fighting climate change. That leaves us with the city and park area, city population, and cattle numbers. City area and population are difficult to address. Population cannot be directly regulated; cities cannot put limits on the number of people moving in. Population, however, is impacted by economics. As the cost of living in certain areas increases, more people will be priced out and forced to move away. This is not a recommendation to raise living costs. That comes with its own set of issues that are outside the scope of this project such as a decrease in the physical and mental health outcomes of the population (Broadbent et al.), which would ultimately be counterproductive to the increase in air quality.

Some cities, such as Portland, have boundary restrictions to combat urban sprawl (Portland.gov). This increases the need to build vertically, increasing population density. I did not find evidence that population density has a significant impact on the AQI. Therefore, imposing urban growth boundaries can be useful for addressing land area increases and lead to lower AQI levels.

Next, cattle populations should be decreased. According to the EPA, 37% of annual methane emissions from human activities are a direct result of livestock. Cattle process food by fermentation within their stomachs. This process breaks food down and produces methane which is released into the atmosphere through flatulence or burps. A

single cow can produce up to 264 pounds of methane gas per year. In total, the 1.5 billion cattle raised specifically for meat production emit at least 231 billion pounds of methane annually (EPA). This figure doesn't include the millions of cattle raised for dairy production. Methane is a powerful short-lived greenhouse gas with a warming impact 86 times stronger than CO2 per unit of mass over a 20-year period. Methane is a precursor gas to Ozone, a criteria gas in the Air Quality Index. Additionally, methane traps heat, increasing the global temperature levels, which, as shown earlier, has negative impacts on the AQI.

Cattle reduction starts with less of a reliance on red meat and the dairy industry, mainstays of the American diet. This will be a huge shift, taking combined efforts of the citizens, government, and food industry. There will be significant pushback by the dairy and meat industries. Further investments should be made into lab-grown meat research or developing suitable plant-based alternatives. Many people are open to switching to traditional meat and dairy alternatives, but existing offerings still have ways to go with regards to taste and texture. As red meat and dairy consumption goes down, less livestock will need to be kept, and less feeding crops will need to be grown (Congressional Budget Office). It will be a difficult transition but a necessary one.

Finally, park land area was found to have a negative impact on AQI values. This surprised me as I had hypothesized that increased park land would increase the quality of air. This difference from expectations is likely due to the use of area as the variable measurement. Certain cities with poor AQI are significantly larger in size such as Phoenix and Los Angeles. Because of how large they are, the amount of park land will similarly be very large in raw square kilometers. Instead, I should have measured park land as a percentage of city area. This would have allowed us to see the proportion of park land within the city limits, which may have had a different result on AQI.

The 2020 wildfires had a significant impact on the air quality. Wildfires have progressively become more common over time. They not only increase the particulate matter in the air, but burn forests, causing long term damage to the soil and loss of plant life. Particulate matter in the air makes it more difficult to breathe, which is reflected by the increased AQI levels. Not all forest fires are started by man-made sources, but many are. Therefore, when in the woods, one should always obey fire restrictions, especially in the middle of the summer when it's most dry. If fires are allowed, they should always be watched and never left unattended. They must always be properly extinguished with all embers cool to the touch before leaving. Campsites should be properly cleaned, and all tools used correctly. One should also stay on

marked trails, avoiding trampling vegetation which can increase the risk of wildfire spreading (Oregon Wildfire Response and Recovery).

My other main focus was the reasons for Portland, Oregon's positive AQI outcomes in comparison to the other large cities. Below, figure 12 shows the rankings of the top and bottom ten average AQI cities from the dataset specified by the six explanatory variables. The top ten highest AQI cities are labeled in red and the top ten lowest AQI cities are shown in green. Additionally, Portland (ranked 4th best AQI) is noted with the cross-hatch lines to compare its measurements with other cities.

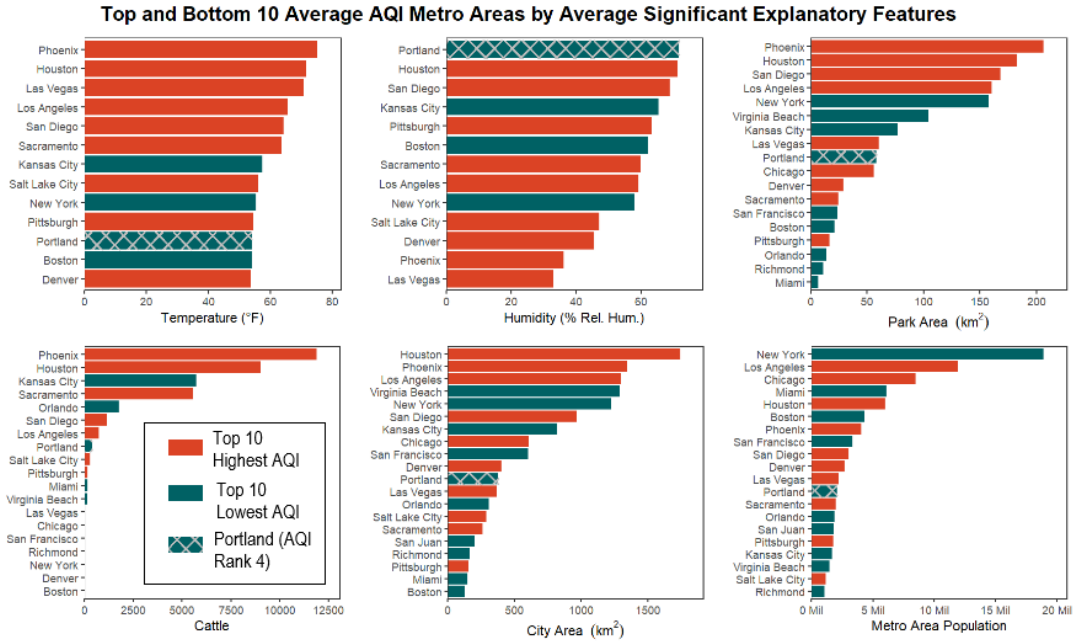

Figure 12: Top and Bottom ten average AQI cities explained by influential features ranked, Portland highlighted

It has been found that Portland has lower average temperatures and a higher average humidity. Portland's large amount of rainfall likely contributes to the higher amounts of water vapor in the air. Additionally, Portland has a smaller city boundary, due to the urban growth boundary. Both Portland's cattle and human population are relatively low. Finally, the size of park area ranks towards the middle of large cities in the top and bottom ten. However, as stated earlier, this is a flawed measurement so results from it are inconclusive.

Our main limitations arise from the lack of data. Importantly, data on the numbers of vehicles on the road could not be found for free which likely has an impact on the AQI. Though I collected public transit information, I didn't find much of an impact, which may have been more pronounced had I paired it with vehicle information. I could have also taken city walkability and metro/subway numbers into account. Other sources of

data that may impact AQI include road and highway location, geographical features, precipitation, industrial manufacturing, wildfire size and locations, agricultural practices, and more.

Even within this collected dataset, there were significant amounts of missing data. Having a more complete dataset may have given more accurate information and allowed me to run these models on more cities. By recognizing these limitations, I can better interpret and apply the results, while also identifying areas for future research. To expand my analysis in the future, I would look at finding more complete information, collecting data on other potentially influential factors described in the previous paragraph, and increasing the timespan of data collected.

Ultimately, we have a responsibility to take care of our planet and combat climate change. The worse climate conditions get, the more wildfires will spread, and the worse the air quality will become. As time goes on, with worsening air conditions, more people will catch and even die from preventable conditions sparked by poor air quality. Water, pollution, food, and financial problems will get worse. One should look at what they can do to make a difference, support those who vouch to make larger changes, and encourage people they know to do the same. While the situation may seem dire, there is hope for progress through concerted and informed efforts.

Bibliography

1. Airly. (n.d.). *How does humidity affect air quality? All you need to know*. airly.org/en/how-does-humidity-affect-air-quality-all-you-need-to-know/
2. AirNow. (n.d.). *Using air quality index*. www.airnow.gov/aqi/aqi-basics/using-air-quality-index
3. American Lung Association. (n.d.). *Key findings: State of the air*. www.lung.org/research/sota/key-findings
4. Broadbent, P., Grantz, D. A., & Leigh, A. (2023). Air quality in cities: Complexities and progress in mitigation. *Frontiers in Sustainable Cities*, *5*. www.ncbi.nlm.nih.gov/pmc/articles/PMC10068020/
5. California Air Resources Board. (n.d.). *Carbon monoxide & health*. ww2.arb.ca.gov/resources/carbon-monoxide-and-health
6. Centers for Disease Control and Prevention, Agency for Toxic Substances and Disease Registry. (2014, October 21). *Medical management guidelines for sulfur dioxide*. wwwn.cdc.gov/TSP/MMG/MMGDetails.aspx?mmgid=249&toxid=46
7. Centers for Disease Control and Prevention, Agency for Toxic Substances and Disease Registry. (2023, April 12). *Toxic substances portal - Nitrogen oxides*. wwwn.cdc.gov/TSP/ToxFAQs/ToxFAQsDetails.aspx?faqid=396&toxid=69
8. Climate & Clean Air Commission. (n.d.). Methane. Retrieved August 14, 2024, from https://www.ccacoalition.org/short-lived-climate-pollutants/methane
9. Congressional Budget Office. (2023, May). *Reducing emissions from transportation*. www.cbo.gov/publication/60030
10. Environmental Protection Agency. (n.d.-a). *Evolution of the Clean Air Act*. www.epa.gov/clean-air-act-overview/evolution-clean-air-act
11. Environmental Protection Agency. (n.d.). Agriculture and aquaculture: Food for thought. Retrieved August 14, 2024, from https://www.epa.gov/snep/agriculture-and-aquaculture-food-thought
12. Environmental Protection Agency. (n.d.-b). *Health effects of ozone pollution*. www.epa.gov/ground-level-ozone-pollution/health-effects-ozone-pollution
13. Environmental Protection Agency. (n.d.-c). *Particulate matter (PM) basics*. www.epa.gov/pm-pollution/particulate-matter-pm-basics
14. Federal Transit Administration. (n.d.). *NTD data tables*. American Public Transportation Association. www.apta.com/research-technical-resources/transit-statistics/ntd-data-tables/
15. Fuller, R., Landrigan, P. J., Balakrishnan, K., Bathan, G., Bose-O'Reilly, S., Brauer, M., Caravanos, J., Chiles, T., Cohen, A., Corra, L., Cropper, M., Ferraro, G., Hanna, J.,

Hanrahan, D., Hu, H., Hunter, D., Janata, G., Kupka, R., Lanphear, B., . . . Yadama, G. N. (2022). Pollution and health: A progress update. *The Lancet Planetary Health*, *6*(6), e535-e547. www.thelancet.com/journals/lanplh/article/PIIS2542-5196(22)00090-0/fulltext

16. Hog Wild Preserve. (n.d.). *Pig, boar, or hog: What's the difference?* www.hogwildok.com/blog/336-pig,-boar,-or-hog-what-s-the-difference.html
17. Jacobs, E. T., Burgess, J. L., & Abbott, M. B. (2018). The Donora smog revisited: 70 years after the event that inspired the clean air act. *American Journal of Public Health*, *108*(S2), S85-S88. www.ncbi.nlm.nih.gov/pmc/articles/PMC5922205/
18. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, *8*, 14. www.ncbi.nlm.nih.gov/pmc/articles/PMC7044178/
19. National Agricultural Statistics Service. (2017). *2017 Census of Agriculture*. United States Department of Agriculture. www.nass.usda.gov/Publications/AgCensus/2017/Full_Report/Volume_1,_Chapter_1_US/
20. National Oceanic and Atmospheric Administration. (n.d.). *What are a nautical mile and a knot?* oceanservice.noaa.gov/facts/nautical-mile-knot.html
21. National Weather Service. (n.d.). *Pressure and winds*. www.weather.gov/source/zhu/ZHU_Training_Page/winds/pressure_winds/Pressure.htm
22. Oregon Encyclopedia. (n.d.). *Interstate 5 in Oregon*. www.oregonencyclopedia.org/articles/interstate_5_in_oregon/
23. Oregon Wildfire Response and Recovery. (n.d.). *Wildfire prevention*. wildfire.oregon.gov/prevention
24. Portland.gov. (n.d.). *City of Portland Charter, Chapter 1*. www.portland.gov/charter/1/2
25. Shandas, V., & George, L. (2009). Neighborhood, neighborhood, neighborhood: Spatial patterns of air toxins and implications for metroscape residents and urban planners. *Metroscape*, Winter 2009. pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1033&context=usp_fac
26. Trust for Public Land. (n.d.). *Park data downloads*. www.tpl.org/park-data-downloads
27. United States Census Bureau. (n.d.). *Home page*. www.census.gov/en.html
28. World Health Organization. (2014, March 25). *7 million premature deaths annually linked to air pollution*. www.who.int/news/item/25-03-2014-7-million-premature-deaths-annually-linked-to-air-pollution

29. World Health Organization. (2022, October 31). *Lead poisoning and health*. www.who.int/news-room/fact-sheets/detail/lead-poisoning-and-health
30. Zhang, Y., Cooper, O. R., Gaudel, A., Thompson, A. M., Nédélec, P., Ogino, S. Y., & West, J. J. (2018). Tropospheric ozone change from 1980 to 2010 dominated by equatorward redistribution of emissions. *Nature Geoscience*, *11*, 637-644. acp.copernicus.org/articles/18/15003/2018/