EÖTVÖS LORÁND
UNIVERSITY | BUDAPEST

DATA MINING AND MACHINE LEARNING - PROJECT 1

Analyzing Environmental Air Pollution in Cities

Zsófia Jólesz

Physics MSc, 3rd semester

Contents

1 Introduction

One of the most concerning problems of today's civilization is air pollution. It poses a serious threat to not only our health condition, but to the overall ecosystem and environment as well. There are air pollutants, such as PM2.5 and NO_2 that are known to cause respiratory diseases (e.g. asthma), cardiovascular diseases, diabetes and cancer. What is more, these pollutants also influence mortality in infants [1].

The major sources of these pollutants are automobiles, power plants and other heavy industries. The pollution level is therefore especially high in highly populated cities and industrial zones, causing very poor living conditions in these areas. Beside traffic and industry, the meteorological factors also play a role in the changing of the distribution and levels of the air pollutants, hence these should be investigated as well.

The aim of my work is to show statistics of air pollutants in different cities and analyze the data. I also want to develop a model which predicts the levels of the certain air pollutants based on meteorological factors, traffic levels and presence of power plants.

For my work I have used the given dataset which includes daily levels of different pollutants and the major casual agents. This dataset encompasses 2 years of data from more than 50 cities of the United States. I have also used a dataset which includes data of the emission of different power plants I have relied on and taken a significant amount of inspiration from the article cited.

2 Data exploration

2.1 Valid datapoints and distributions

The dataset I have used is the largest dataset of this topic, regarding the number of locations and days involved. The dataset contains a total of 35,596 unique sample points, 54 cities and 24 months with each sample point representing a unique (date, city) combination. Since some cities and dates have some data missing for each pollutant, first I have calculated the number of valid samples and valid cities for each pollutant. A city is considered valid if it has at least 2 months data of the pollutant levels. The result is summarized in Table 1.

Pollutant	Valid Samples	Valid Cities
O3	33950	54
PM2.5	35134	54
NO2	23558	43
CO	24538	42
SO2	14676	43
PM10	16965	31

Table 1: Valid samples and cities for each pollutant.

I also found useful and necessary to show the monthly distribution of each pollutant. From that plot (see Figure 1) we can easily see that there are some outlier points which would modify our results if not treated correctly. I have calculated the number of outliers, for which I set different thresholds for each pollutant. The thresholds, the number of pollutants above the thresholds and the percentage of the number of outliers in the whole data is summarized in Table 2.

Pollutant	Threshold	Data Above Threshold	Percentage
O3	50	123	0.346
PM2.5	125	280	0.787
NO2	50	1	0.003
CO	10	99	0.278
SO2	3	1	0.003
PM10	80	48	0.135

Table 2: The number and percentage of outliers for each pollutant.

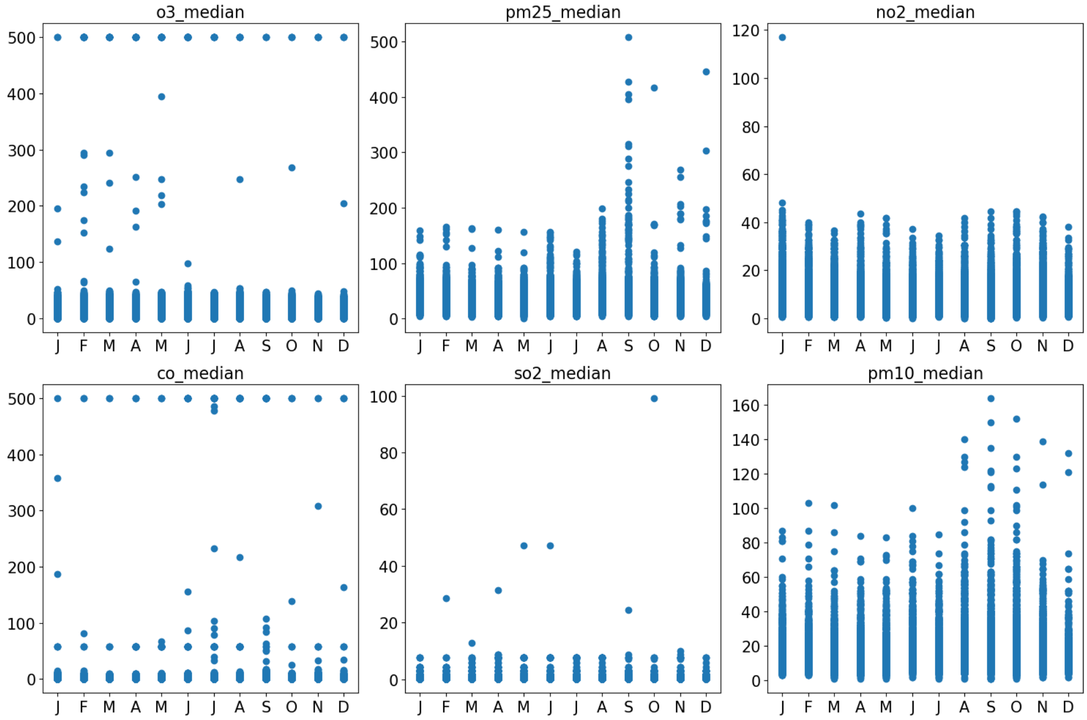

Figure 1: Monthly distribution of the air pollutants.

It can be seen that only a negligible amount of data falls above the given thresholds, therefore I have decided to throw away these outliers and I have used the dataset without these points in the following.

With the truncated dataset I have created the violin plots of the air pollutants' monthly distribution and variation, which gives a more spectacular visualization of the data. This can be seen on Figure 2.

I was interested in the cities that have the largest and smallest amount of air pollutants, since it is an informative feature of the dataset. The boxplot of the five cities with the largest and smallest air pollutant concentration (for each pollutant) can be seen on Figure 3 and 4. The data needed to be normalized first, since the amount of data is not necessarily the same for every city. This data is later used in section Hypothesis.

2.2 Meteorological factors

Another characteristic feature of the dataset was the different meteorological factors that have a possible impact on the pollution level as well. The preprocessing was also required for this part of the data, therefore I checked the number of NaN values first. I have found that except

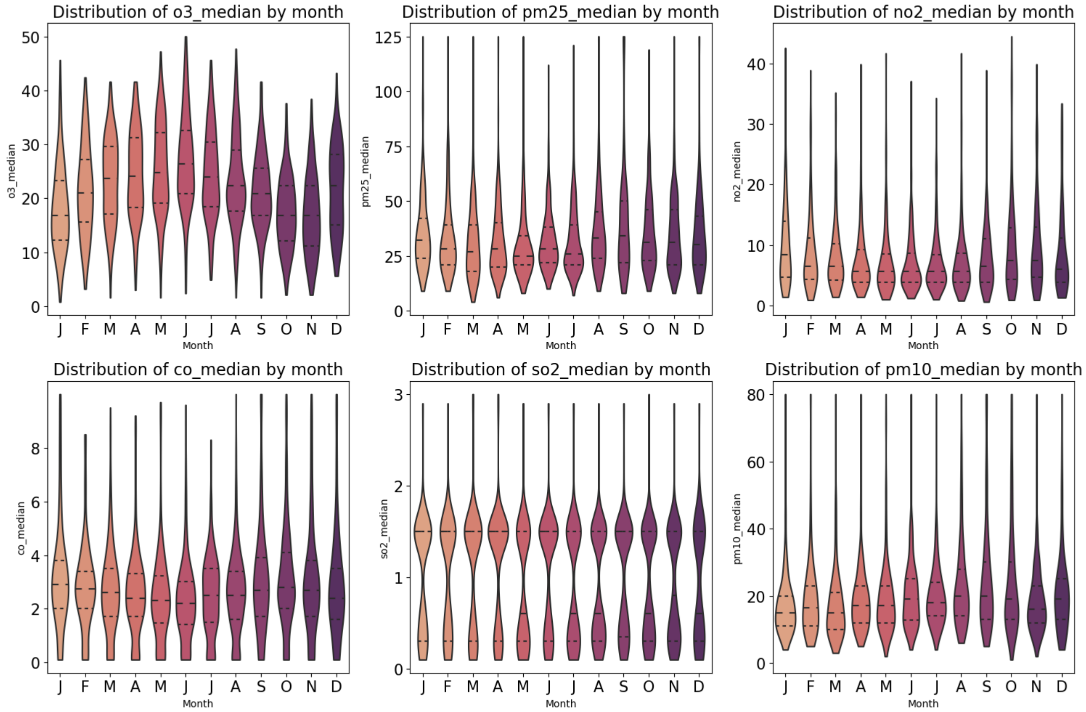

Figure 2: Violin plot of the monthly distribution of the air pollutants.

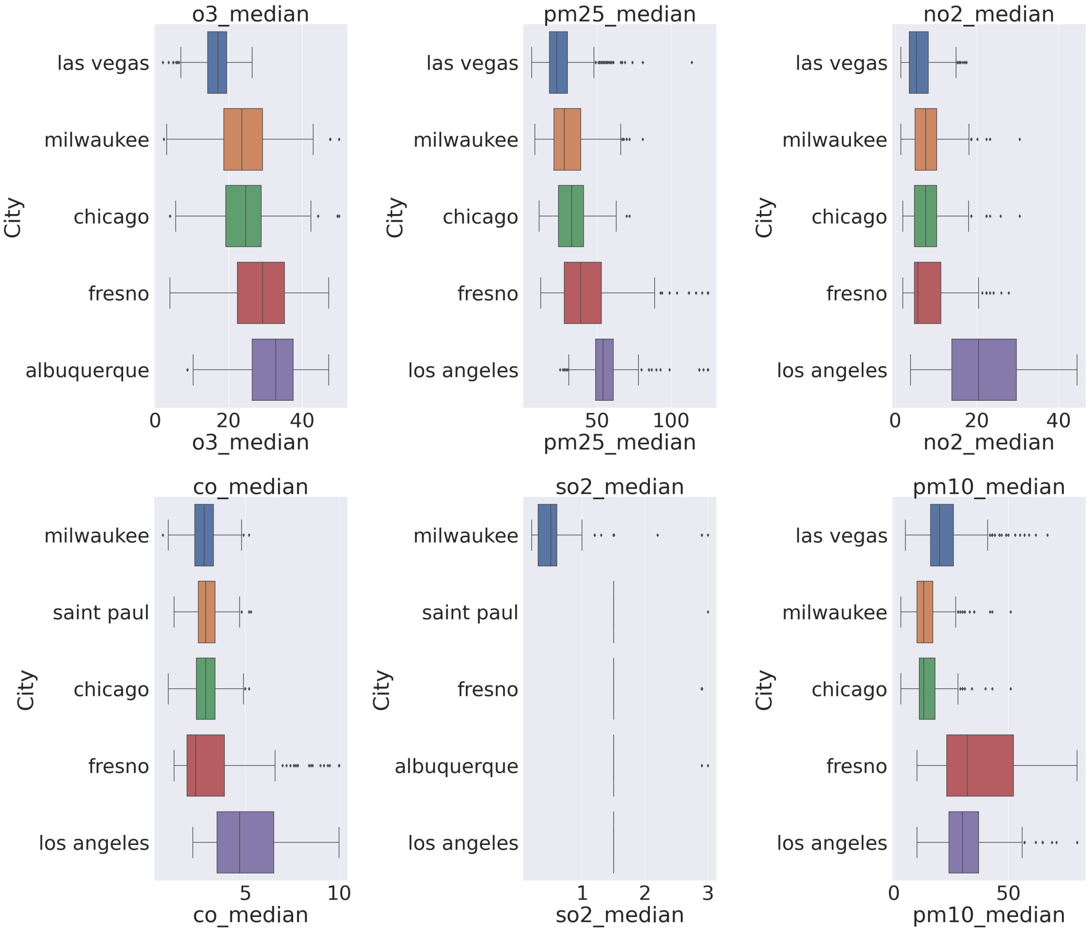

Figure 3: Cities with the highest pollutant concentrate.

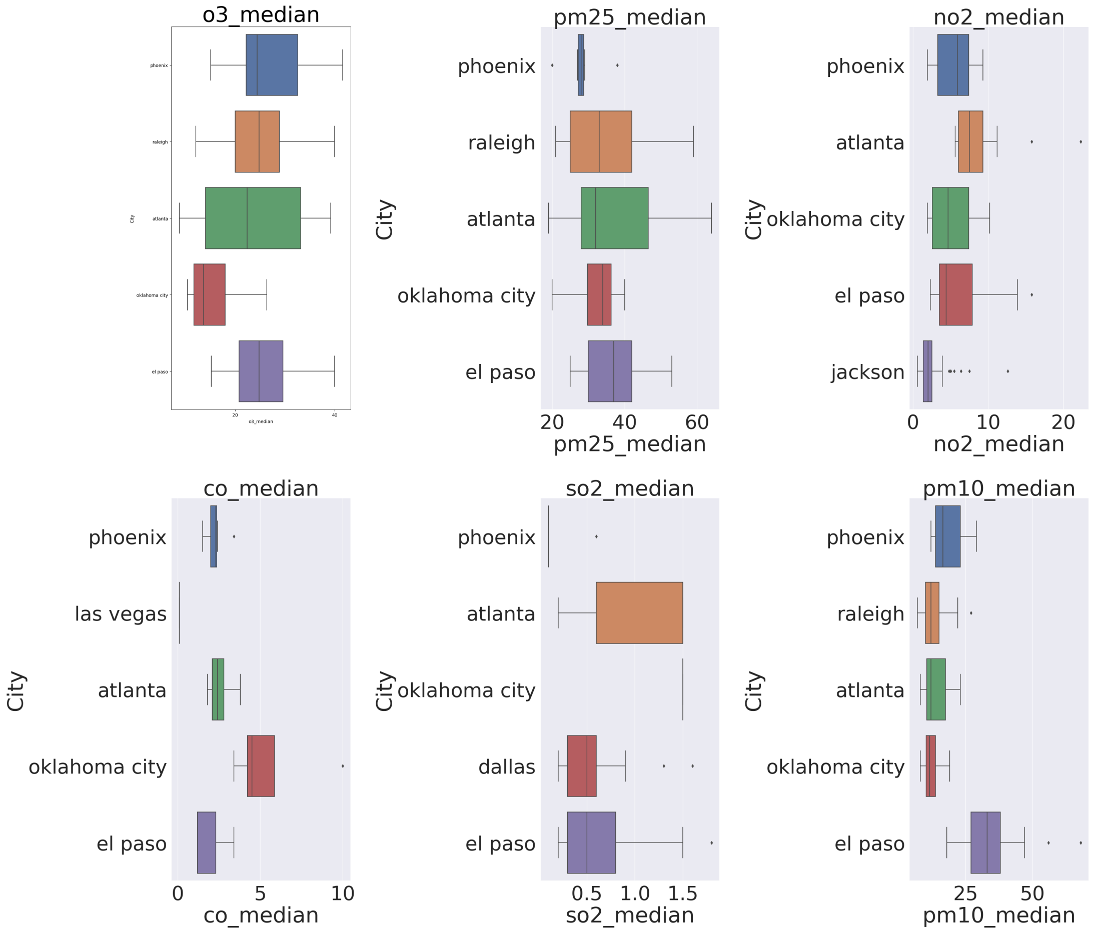

Figure 4: Cities with the lowest pollutant concentrate.

for meteorological factor 'dew' every parameter had a negligible amount of NaN values, hence they were thrown away. In the case of parameter 'dew' which had almost its 50% missing, I have inputed the NaN values with -1. After that truncation the number of valid samples remained 3733, which I considered a feasible amount of data to work with.

Since the impact of the certain meteorological factors can be significant, I have shown the correlation between the factors and the pollutant levels. The correlation is shown as a heatmap on Figure 5. It can be seen that O3 does not correlate strongly with any of the meteorological factors, while SO2 has a strong correlation with each one of them. The meteorological factors that correlate the strongest with the air pollutants are wind speed and dew.

2.3 Power plants

I have read in the dataset that contains the power plant types, the latitudinal and longitudinal coordinates of them, the dates and the net generation of the fuel for the dates. My goal with this dataset was to calculate the feature given in the article, which represents the effects of the power plants for a certain (city,date) pair. The equation used for determining this factor was the following:

$$I_{pp,c,t} = \sum_{p} G_p / r_{cp}^2, \tag{1}$$

for $r_{cp} < R_{\text{limit}}$, where I_{pp} is the feature obtained from power plants for a city c on a date t. G_p is the average daily generating capacity for the plant for that month and r_{cp} is the linear

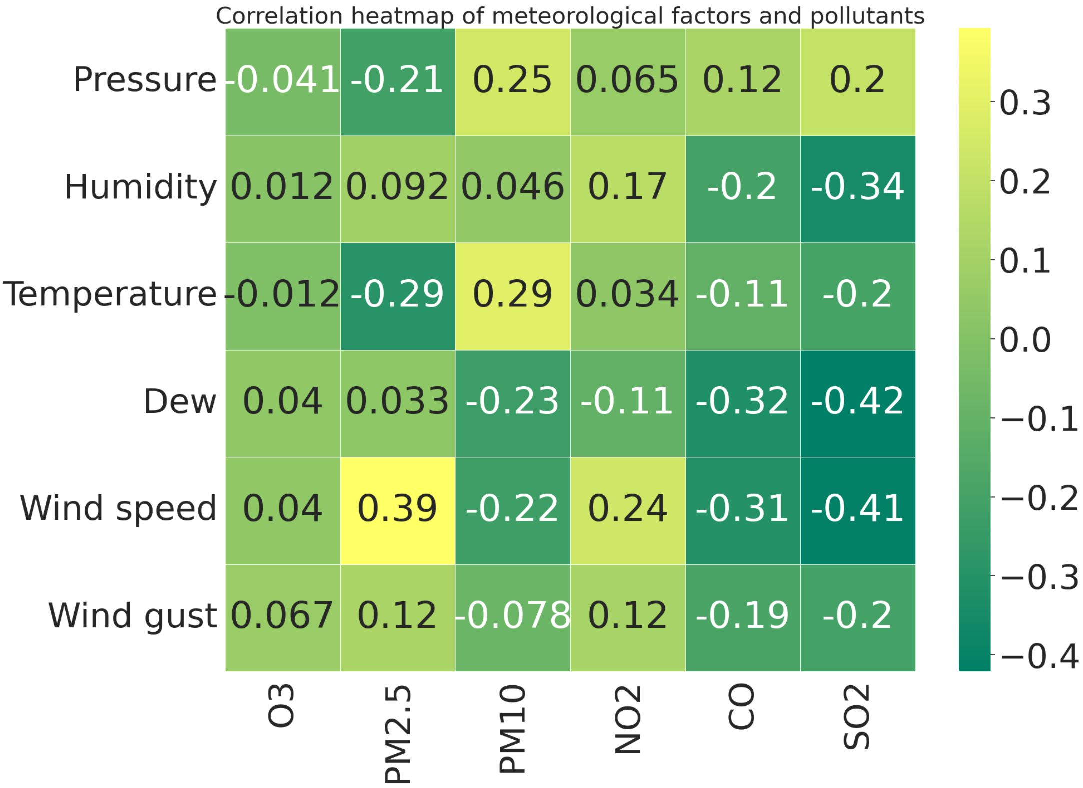

Figure 5: Heatmap of the correlation of the pollutants and the meterorological factors.

distance between the power plant and the center of the city. R_{limit} was taken as 30 km and the net generation of the power plants was averaged monthly to a daily level. Due to some difficulties with the code and lack of time, this part of my project could not be handled out. Including this dataset would have most likely increased the precision of the later used machine learning algorithms, this should be taken into account when evaluating the model.

2.4 Traffic

The last parameter I have found informative and necessary for EDA was the driven miles (given in million miles). I have implemented a correlation between the 'mil_miles' values and the different air pollutants and found that the highest correlation was with air pollutants PM10, CO and SO2. This correlates with the fact that these pollutants are usually the artifacts of vehicles. The correlation heatmap can be seen on Figure 6.

3 Hypothesis

In this part of my project I have decided to examine the relation between the volume of traffic and the cities that have the highest and lowest pollution concentration. Since the parameter 'mil_miles' had the highest correlation with pollutants PM10, CO and SO2, I have decided to choose the cities that have the highest and the lowest concentration of these pollutants and plot their driven miles in monthly distribution. The cities with the highest PM10, CO and SO2 concentrate were Fresno, Los Angeles and Fresno, respectively, but Los Angeles was also in the top 5 for PM10 and has the same level of SO2 as Fresno, thus I have taken Los Angeles as the city with the highest pollutant level. The cities with the lowest levels of the same pollutants were Oklahoma City, Las Vegas and El Paso, respectively. Of these cities I have chosen El

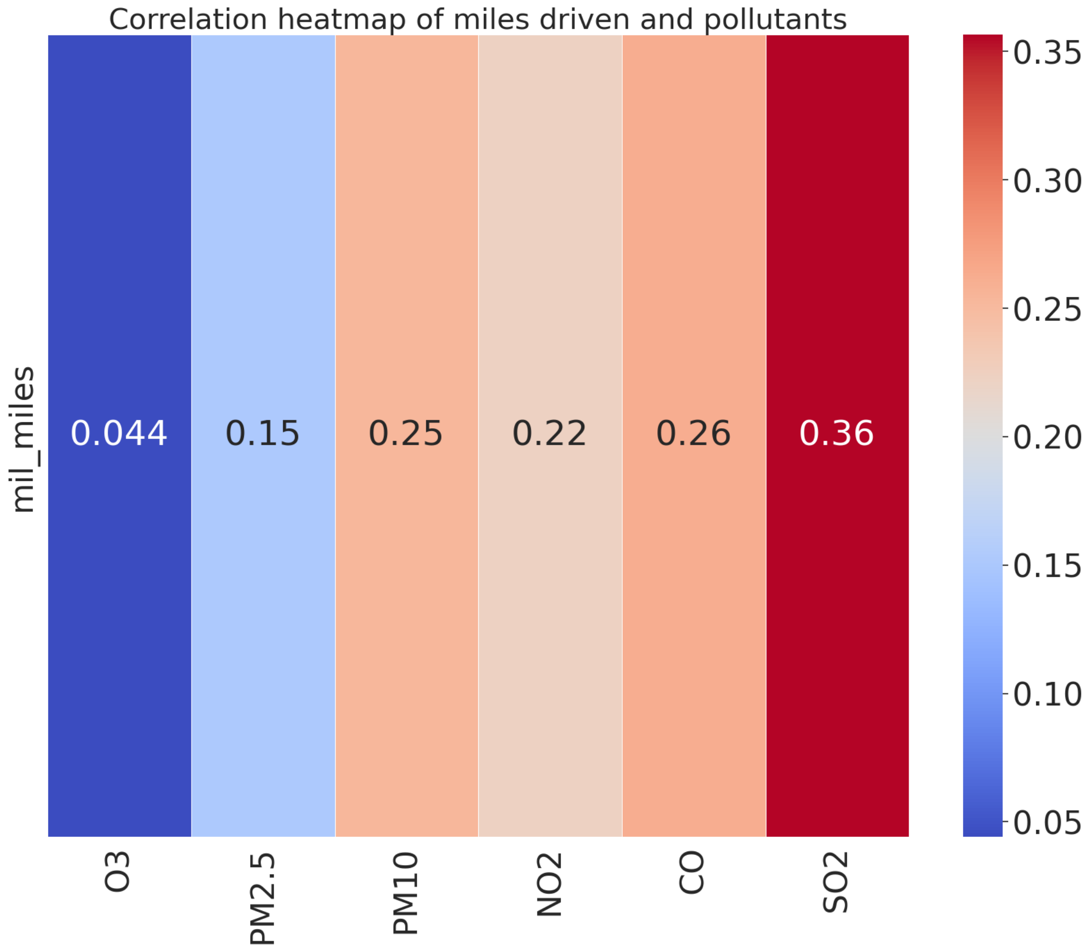

Figure 6: Heatmap of the correlation of air pollutants and traffic level.

Paso, due to the fact that this was the city that appeared in every category. On Figure 7 it can be seen that the driven miles in Los Angeles are significantly higher than in El Paso, hence the presumption was correct and the correlation between the driven miles and the aforementioned pollutants is indeed high.

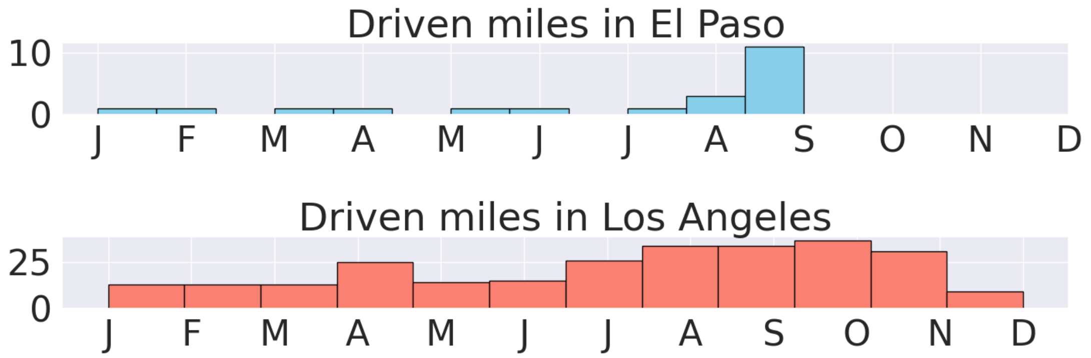

Figure 7: Cities with the highest and lowest pollutant levels.

4 Data modeling

For this dataset I have found that the most fitting model would be a decision tree, more precisely for a regression task in order to predict certain pollutant levels, thus my goal was to implement this machine learning model on my dataset. Since I aimed for a prediction of the certain pollutant levels, I used them as the target variables and columns 'Population Staying at Home', 'Population Not Staying at Home', 'mil_miles' amd the meteorological factors as the predictor variables.

The first pollutant I used was the PM10. My first attempt was the usage of a DecisionTreeRegressor, which I evaluated by calculating the RMSE (root mean squared error). For this model, the RMSE was 10.42, which indicates that the DecisionTreeRegressor model does not fit the data well. The decision tree can be seen on Figure 8. Since this was not the best approach, I have decided to utilise a RandomForest model. Before this, I searched the best n-estimators with cross-validation, then applied the RandomForestRegressor. The RMSE was 7.61 at this point, which is still considered too high, yet an improving trend was visible. I have also tried the Gradient Boosting ensemble method, with whom I have reached an RMSE of 7.59.

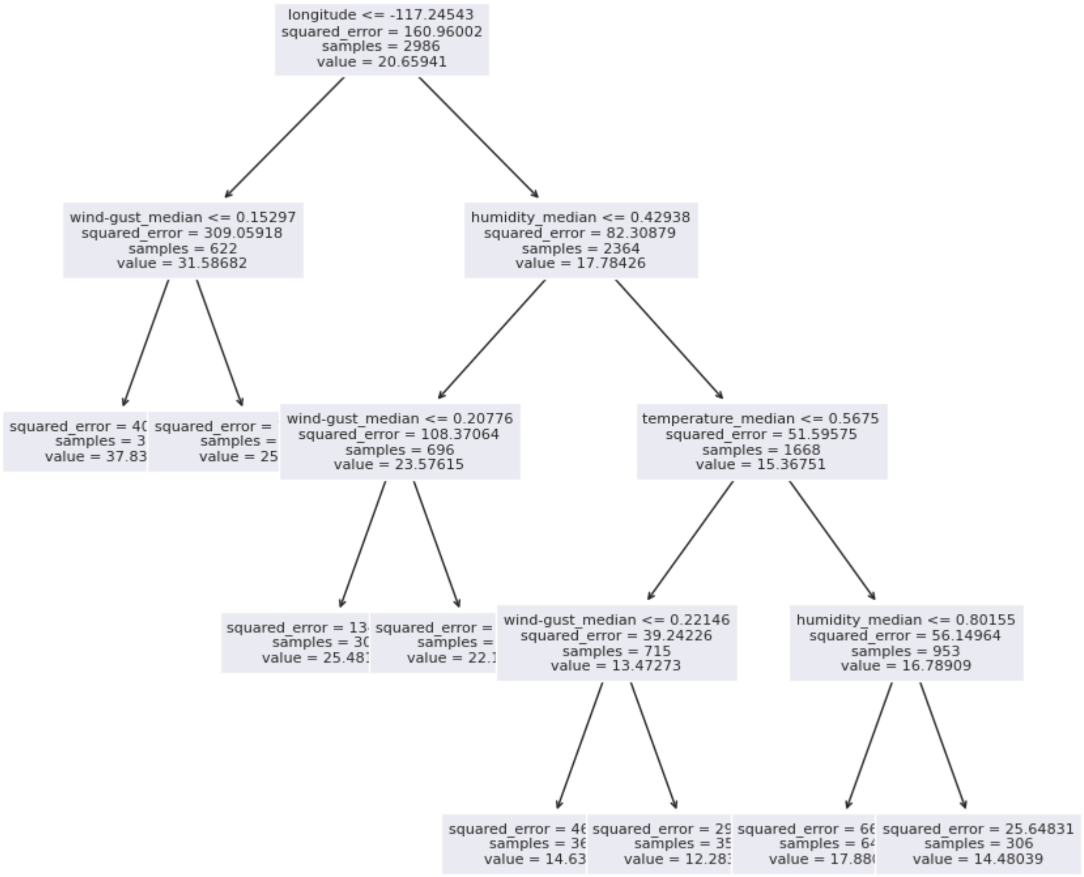

Figure 8: Decision tree of PM10.

I have also used the DecisionTreeRegressor on pollutant O3 and I got slightly better values, namely a 7.01 of RMSE. The decision tree can be seen on Figure 9. The Gradient Boosting method resulted in an RMSE of 4.91 for this pollutant.

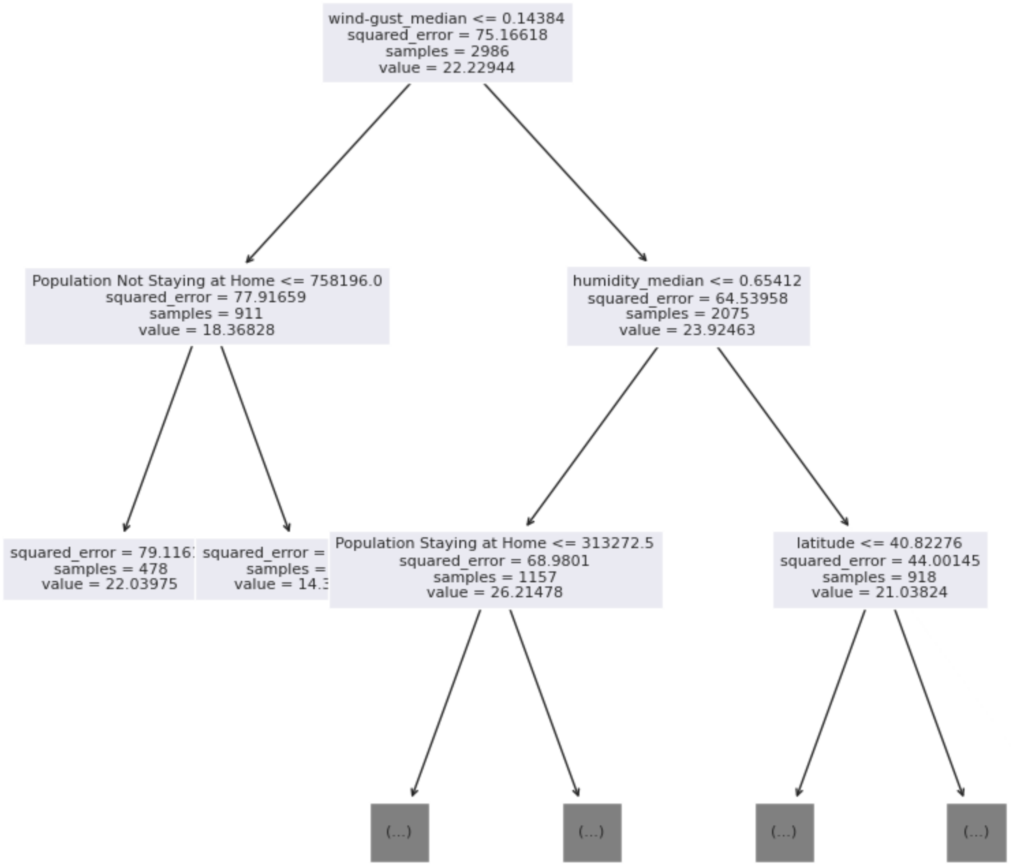

Figure 9: Decision tree of O3.

I have also performed Gradient Boosting on the other pollutants. The RMSE values of all pollutants are summarized in Table 3.

Pollutant type	RMSE
PM10	7.59
O3	4.91
PM2.5	13.18
NO2	3.26
CO	0.92
SO2	0.34

Table 3: The RMSE of each pollutant.

5 Summary

In my project I have examined the largest dataset of air pollution in the United States. I have performed exploratory data analysis on the given dataset and set up a hypothesis in connection with the traffic level, which was found to be true. I have also utilised different machine learning algorithms in order to predict pollutant levels, of which the Gradient Boosting method was the most suitable and precise for the problem. I have reached especially precise prediction for pollutants CO and SO2.

The explanation for the less accurate predictions could be the lack of usable and valid data, since I have truncated my dataset at the beginning, where I have thrown away the outliers. At this point, the biggest number of outliers was for PM2.5 and I find this a suitable explanation for the least precise prediction.

It also should be mentioned that due to unforeseen circumstances, I could not utilise the dataset of power plants, which could have further increased the punctuality of the predictions.

To conclude my work, I have learned a significant amount about machine learning algorithms and exploratory data analysis, through which I have successfully completed the tasks I wanted to, apart from the connection of power plants with the air pollutant levels.

References

[1] Mayukh Bhattacharyya, Sayan Nag, and Udita Ghosh. Deciphering environmental air pollution with large scale city data. *arXiv preprint arXiv:2109.04572*, 2021.