Exploring the Effects of COVID-19 on the Air Quality of Hong Kong

Milind Sharma
City University of Hong Kong

Abstract

In this report I display the effects of the COVID-19 pandemic on the air pollution levels in Hong Kong in 2020. The basis for this analysis is a comparison of the concentrations of CO, NO_2, O_3, $PM_{2.5}$ (FSP) and NO_X (nitrogen oxides) in 2020 to their concentration levels in 2019. I fit the pollutant distributions to standard probability distributions, in order to perform further probability calculations on the pollutants. T-tests I conducted showed that mean levels for all pollutants except O_3 fell from 2019 to 2020. I also conducted correlation analysis using simple linear regression to explore the relationship between different pollutants. The use of daily data showed inconclusive, weak correlations between NO_X and O_3, and NO_2 and O_3. A more in-depth analysis using hourly data revealed significant negative correlations between NO_X and O_3, and NO_2 and O_3. This analysis showed that in order to effectively manipulate and lower air pollutant levels, it is imperative to understand what effects the air quality. Relationships between air pollutant concentrations and things like COVID-19 should be further studied.

Introduction

Hong Kong's position as a bustling urban metropolis ensures that there are significant levels of aerial pollutants in the city's air throughout the year, despite the Hong Kong government's constant efforts to curb air pollution. Most of these pollutants are due to emissions from vehicles, marine vehicles and power plants. However, in the past year (2020), with the COVID-19 pandemic affecting large parts of the world, significant deviations from the normal way of life have taken place in many cities including Hong Kong, to try and curb the spread of the disease. Much of this may also indirectly impact air quality in the city.

For example, a ban on non-essential travel to the city and a mandate that all international arrivals must complete a two-week quarantine period would have reduced footfall in the city dramatically and stopped tourism almost entirely. Both these things may impact pollution. Within the city, frequent enforcement of COVID restrictions, including bans on gathering in groups, and restrictions on eating in restaurants, have resulted in even less hustle and bustle in the city. Lastly, working from home instead of commuting to office resulted in less cars on the road, which may have also affected the air quality.
Any conclusions reached at the end of this report may help us better understand what factors affect air quality, specifically in Hong Kong. This would be helpful in learning how to change the concentrations of ground level atmospheric pollutants, something that could aid the population of Hong Kong breathe cleaner air and be healthier.

Plenty of research has already been done into investigating how the COVID 19 pandemic has affected the air quality in other places. In many regions of mainland China, a 12%-58% reduction in aerial pollutants was recorded once the compulsory virus restrictions were put in place (Chen *et al.*, 2020; Zhang *et al.*). Bedi *et al.*, 2020, observed reductions in particulate matter and NO2 concentrations in 4 major Indian metropolises by up to 50% at times. This report will attempt to investigate if any similar conclusions can be reached regarding the air in the city of Hong Kong.

In this report I aimed to:

1. Investigate the distribution of certain air pollutant data in Hong Kong in 2020 compared to the previous year, 2019 (when the effects of the pandemic hadn't been felt yet), to observe the shape of the distribution.
2. To fit a probability distribution to the data, in order to make probability calculations related to it.

3. Formulate hypotheses about how the air pollutants concentrations in 2020 would compare to those in 2019. Validate/reject these hypotheses.
4. Check for correlations between the pollutants and attempt to explain any correlations found.

Methods

Study Area

The area I studied in this report is the city of Hong Kong, a special administrative region of China, located in the Guangdong-Hong Kong-Macao Greater Bay Area. The city has a population of around 7.5 million people.

The specific AQHI measuring station used to collect the data that I analysed, is the Causeway Bay Roadside Station operated by the Environmental Protection Department (EPD) of Hong Kong. The reasons for selecting this station were:

1. Causeway Bay is in a central and busy location on the island, through which large numbers of people and vehicles pass through.
2. It is positioned directly adjacent to Victoria Harbour, a busy port.
3. Both the above factors, combined with the fact that it is a roadside station, would make it sensitive to changes in the concentrations of air pollutants.

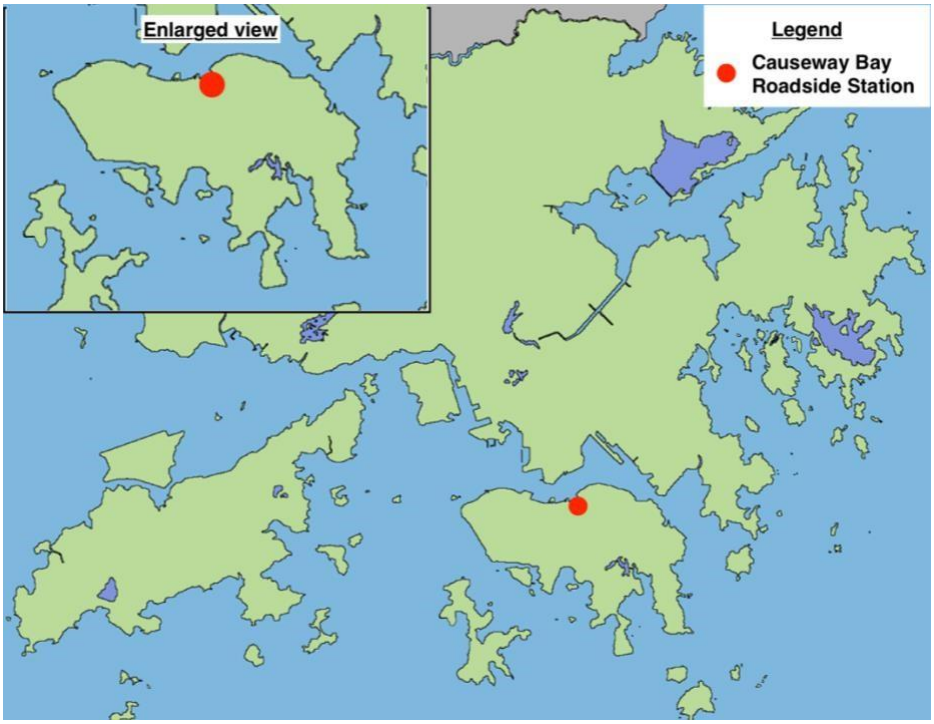


Figure 1: Map to show location of measuring station.

Data

The data I used in this analysis consists of daily averages of pollutant concentrations, from the aforementioned atmospheric pollutant measuring station in Causeway Bay. I also used a small amount of hourly data to further investigate a possible correlation that I discovered. I randomly chose the dates for the hourly data. (5 February 2019 and 5 September 2020)

Data collected during the period of January 23 to September 30 for both 2019 and 2020 has been utilised in this analysis. January 23, 2020 was the date that the first COVID-19 case was confirmed in Hong Kong, and September 30, 2020 is the latest date till where daily data was available at the time of this analysis. Thus, in this study, for the purpose of standardisation, I took the same date range for both years. The data has also been cleaned so that all measurements for a particular date are removed even if just one is missing (this has been done to preserve correlations). As the data from 2019 is independent of the data from 2020, a small difference in total number of data points will have negligible effect on the results.

For the purpose of this study, I analysed the concentrations of CO, NO_2, O_3, $PM_{2.5}$ (FSP) and NO_X (nitrogen oxides). I divided the data into two parts with data from 2019 regarded as historical data, as it was recorded prior to the COVID-19 pandemic. I regarded the 2020 data, which was recorded during the ongoing pandemic, as current/new data.

Methods of analysis

Computational aid

I used Python 3.7 to analyse the data, to generate graphs and figures, and plot data for me. I also used it to conduct the t test with a high level of accuracy, and to create linear regression models.

Visualisation

I generated boxplots to visualise the data and to gain an initial understanding of what trends and correlations may be discovered later.

Fitting to distribution

I then used histograms observe the general shape of the distributions of various pollutants. For the histograms, the bins were mostly generated using the Freedman-Diaconis estimator, which is resilient to outliers, and takes into account data variability and data size. For a few histograms I set the bin size myself in order to better highlight the shape. To smooth out the distribution and make it easier to observe the underlying distribution, I plotted kernel density estimates over the histograms.

These kernel density estimation plots were then compared to whichever probability distribution appeared suitable (normal, chi square). The parameters of the probability distribution were set to that of the pollutant data, to make it fit the distribution appropriately.

Determining differences

To determine whether there was any significant difference between the mean concentration of pollutants in 2020 vs 2019, I conducted hypothesis T-tests for two independent populations (2019 and 2020). I created null and alternate hypotheses for each pollutant based on whether their mean values were higher or lower in 2020. For example, if the mean of a parameter was higher in 2020 than in 2019, I would formulate the following pair of hypotheses:

- $H_0: \mu_{2020} = \mu_{2019}$
- $H_1: \mu_{2020} > \mu_{2019}$

I then conducted one sided hypothesis tests for each pollutant to accept or reject these hypotheses. A significance level of $\alpha=0.05$ was decided on, as a good balance between ensuring a low probability of type I error, and enough probability of detecting a difference. I also directly compared the means of each pollutant in 2019 and in 2020, to determine how much of a percentage increase/decrease there was from one year to the next.

By default, the t-test used in SciPy's statistics module is a two tailed hypothesis test. Since I was conducting one tailed test, I halved the resulting p value to get the appropriate probability for the test before comparing it to the value.

I also did not assume the population variances to be equal in the t tests, due to differences in the variability of the samples I had.

Correlation testing

I chose to search for correlations between ozone and other pollutants with ozone as the dependent variable. This is because ozone is the only pollutant examined here that is not directly emitted from vehicles/power stations etc, but forms due to interactions between other pollutants, organic compounds (VOC) and sunlight.

To determine these correlations, I used simple linear regression. I also utilised scatter plots and simple line plots to visually identify any possible correlations. Eventually I also analysed some hourly data, chosen at random, for certain pollutants, to further investigate the extent of correlation.

Since I was only carrying out simple linear regression (one independent variable, one dependent variable), I use the r squared value and not the adjusted r squared value, which would be slightly lower for little reason.

Results

Data Visualisation

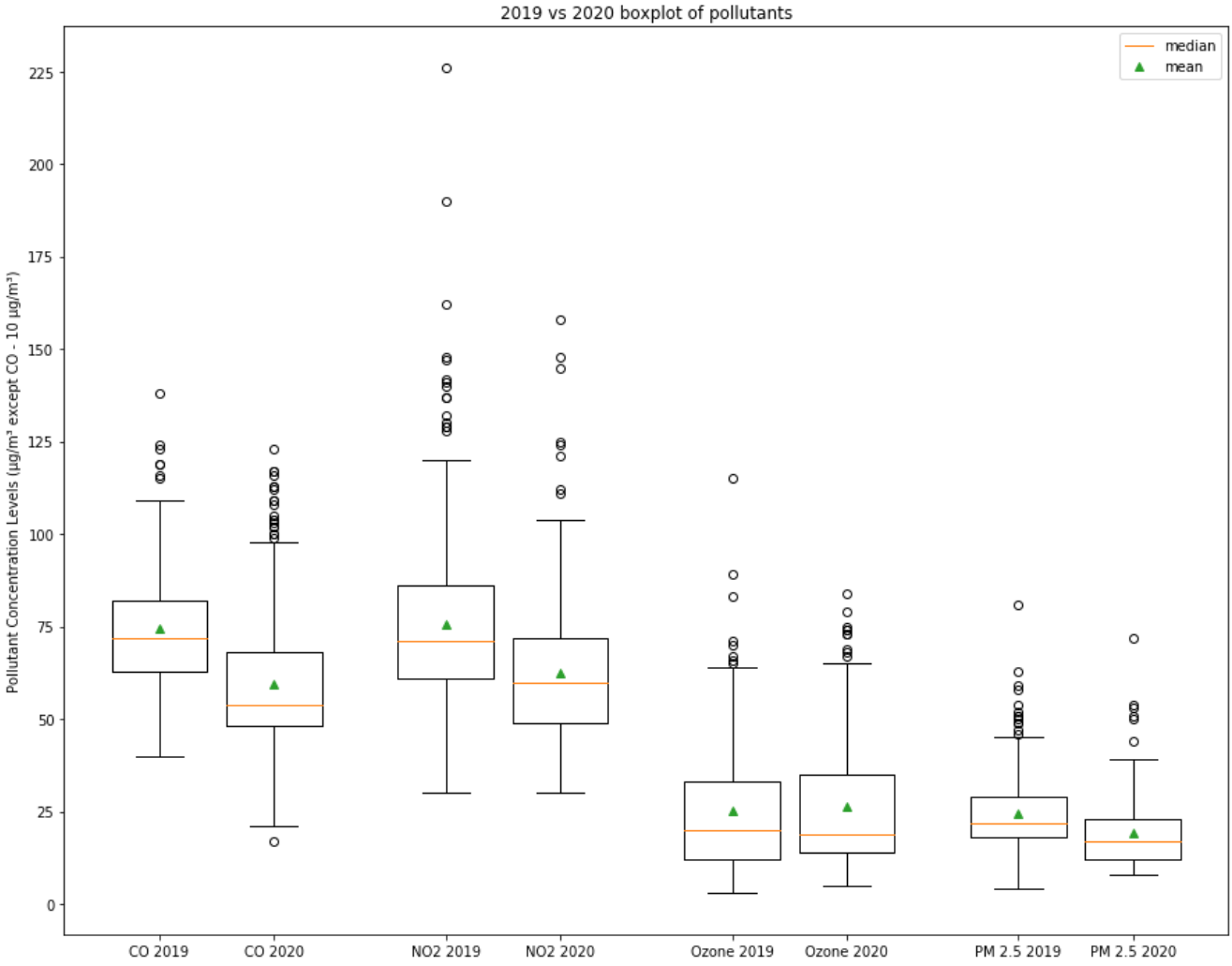


Figure 2: Box plot of CO, NO_2, O_3, $PM_{2.5}$ pollutant concentrations.

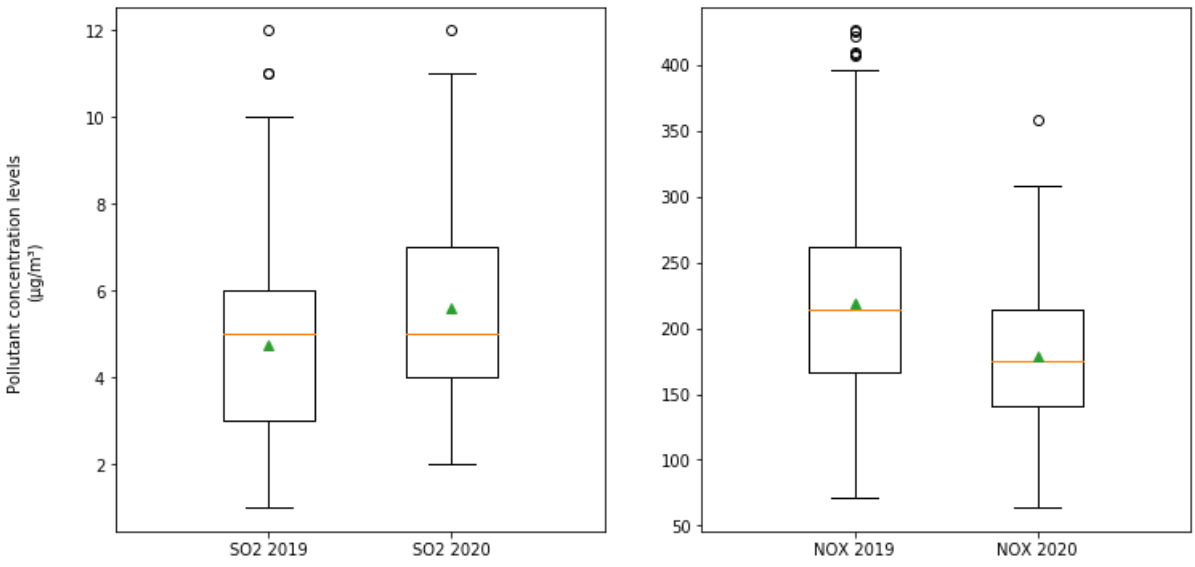


Figure 3: Box plot of SO_2, NO_X pollutant concentrations

From the boxplots of the pollutant concentrations, it is clear that the means of the concentration of CO_2, NO_2, NO_X and $PM_{2.5}$ are lower in 2020 than in 2019. For SO_2 and O_3 the mean values are slightly higher in 2020.

Fitting data to distributions

1. CO

The distribution of carbon monoxide appears to be roughly normal. The kernel density estimate also indicated that the distribution was normal.

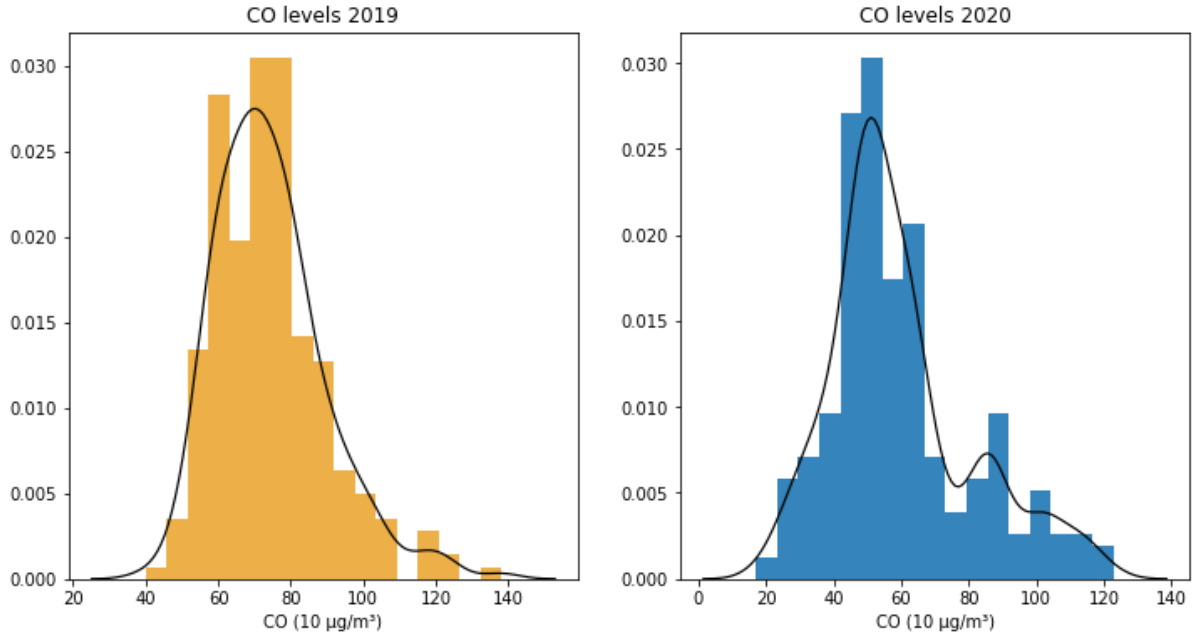


Figure 4: Distributions of CO in 2019 and 2020, with kernel density estimate overlayed.

A comparison of the kernel density estimate to the appropriate normal distribution shows a good match for 2019, but a poor one for 2020.

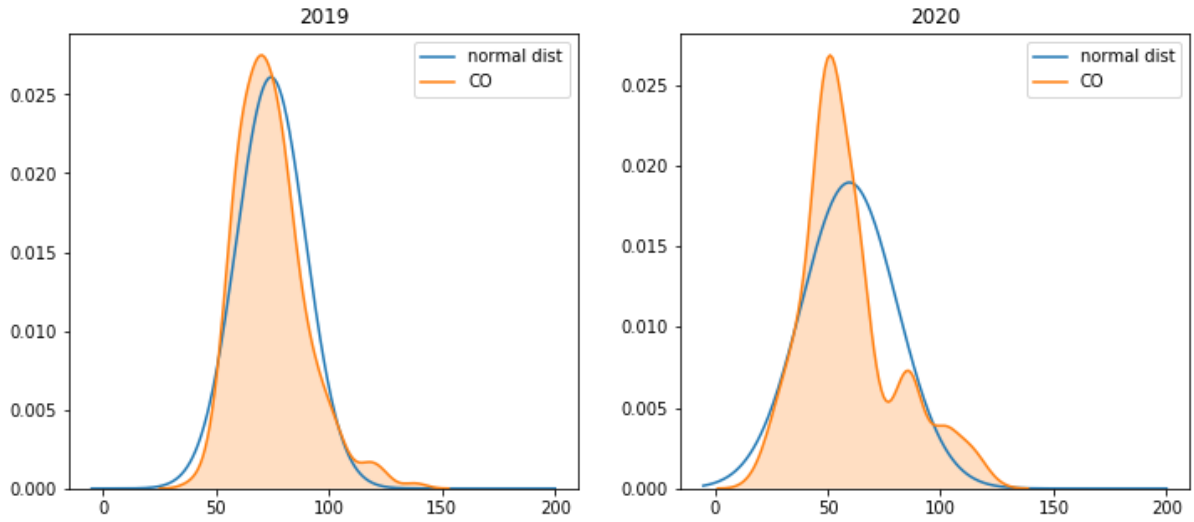


Figure 5: comparison of CO KDE plot to normal distribution

2. NO_2

The distribution of NO_2 appears to be normal in nature, as made clearer by its kernel density estimate plot.

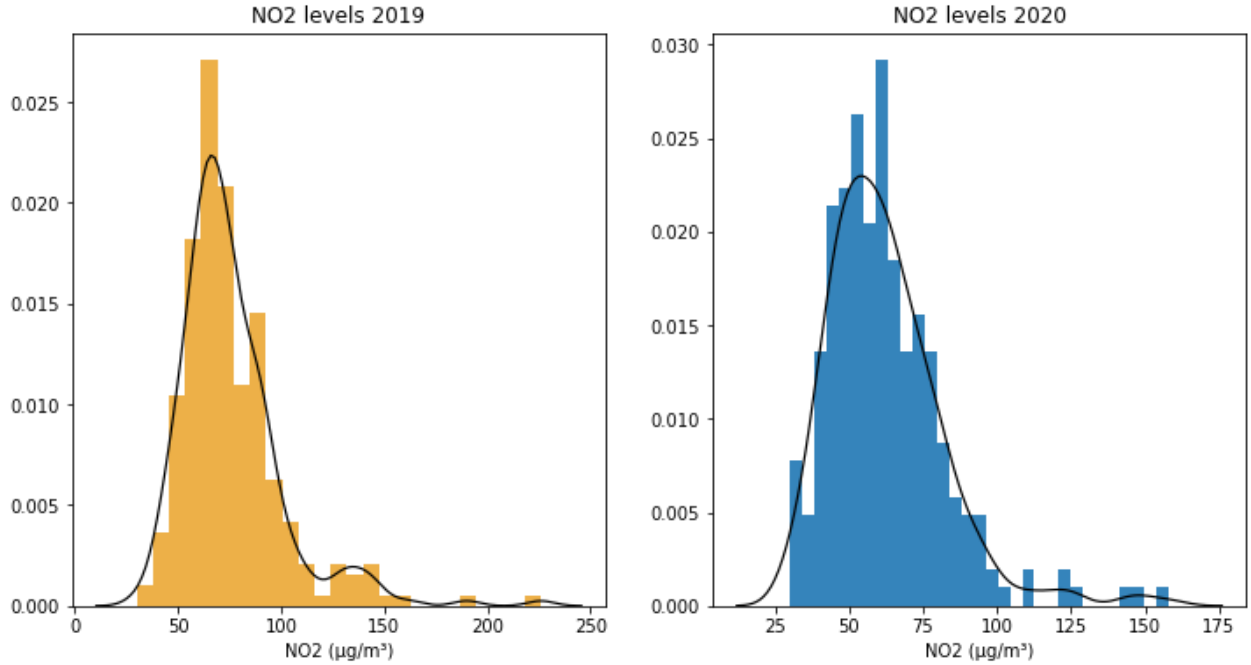


Figure 6: Distributions of NO_2 in 2019 and 2020, with kernel density estimate overlayed.

A comparison of the kernel density estimate to the a fitted normal distribution shows that a normal distribution appears to be a good fit for 2020,

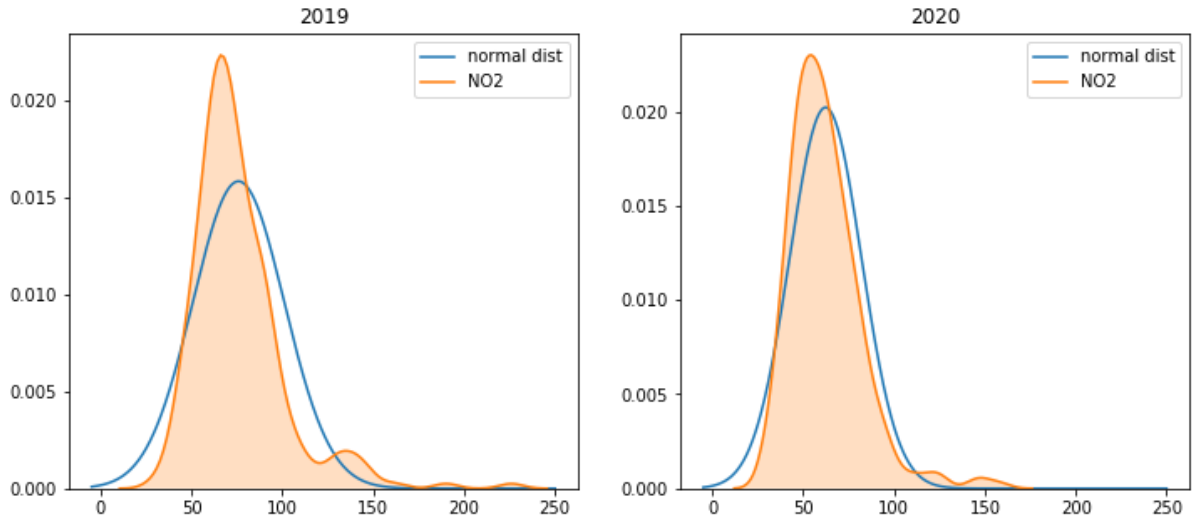


Figure 7: comparison of NO_2 KDE plot to normal distribution.

3. O_3

The distribution of ozone in both 2019 and 2020 is clearly right skewed, and certainly not a good fit for a normal/Gaussian distribution. Here I fit a Chi Square distribution to the data.

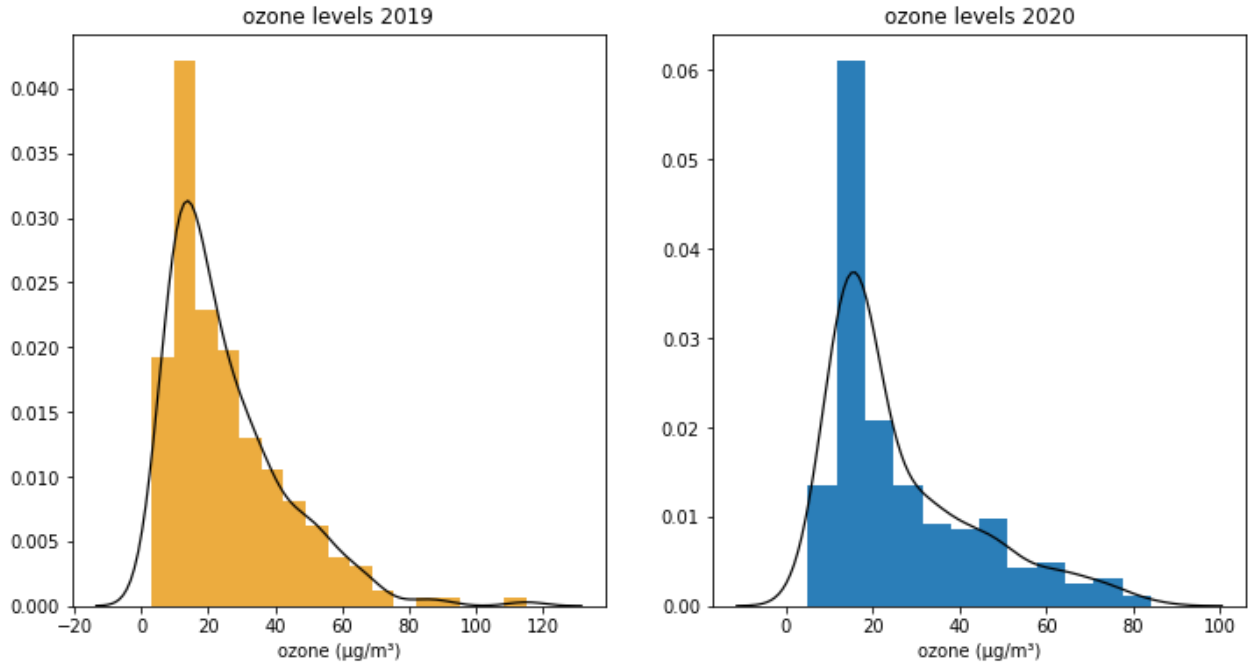


Figure 8: Distributions of NO_2 in 2019 and 2020, with kernel density estimate overlayed.

As shown below, the chi square distribution appears to be a good fit for the distribution of ozone.

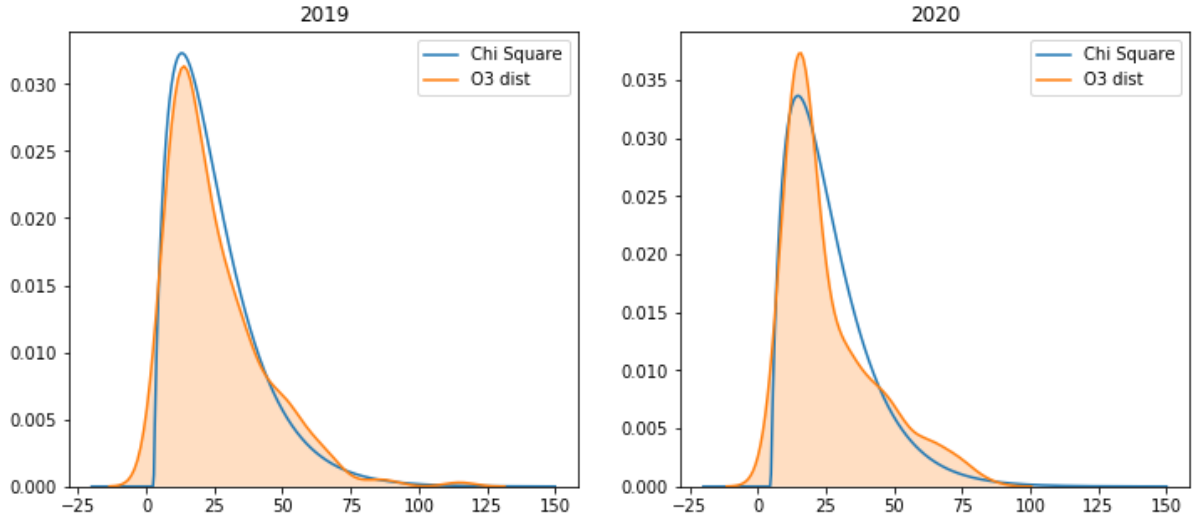


Figure 9: comparison of O_3 KDE plot to chi square distribution.

4. $PM_{2.5}$

The distribution of $PM_{2.5}$ also appears to be right skewed, and not normally distributed. Again, I fit a chi square distribution to the data.

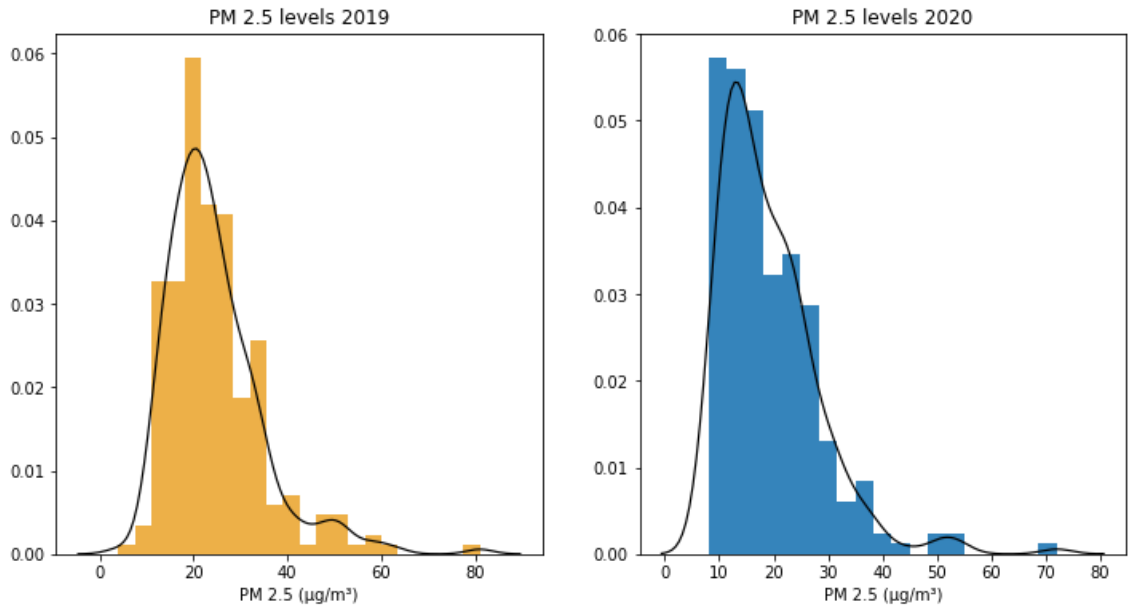


Figure 10: Distributions of $PM_{2.5}$ in 2019 and 2020, with kernel density estimate overlayed.

Shown below, the chi square distribution seems like a good fit for the distribution of fine suspended particulate matter ($PM_{2.5}$) in 2019 and 2020.

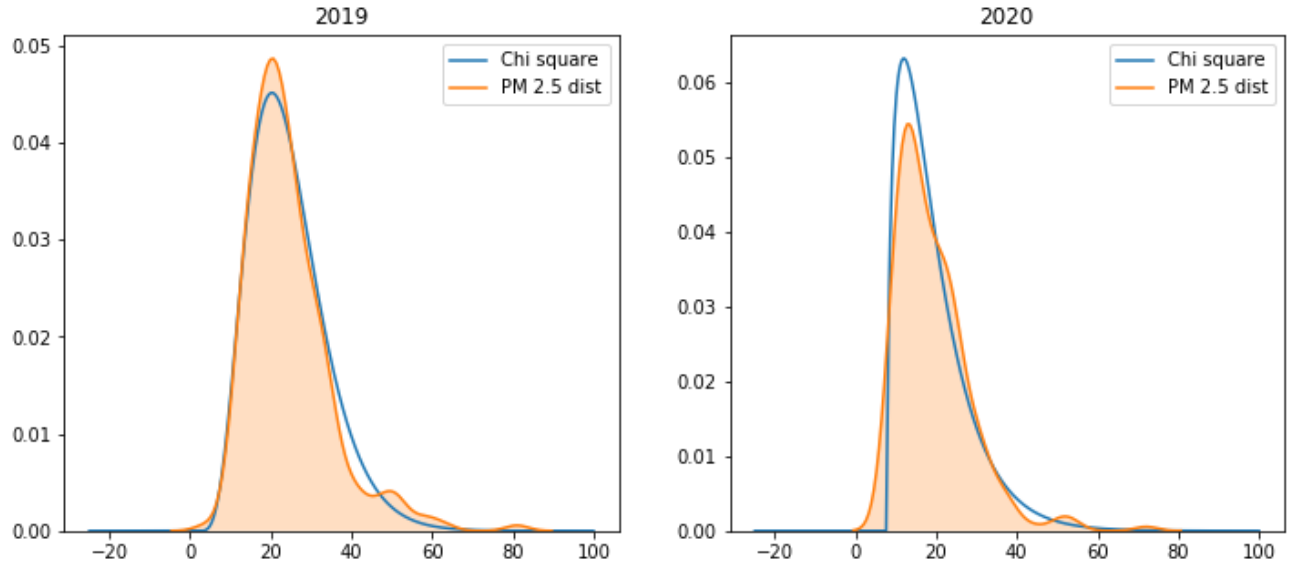


Figure 11: comparison of $PM_{2.5}$ KDE plot to chi square distribution

5. NO_X

From the histograms and from the shape of the kernel density estimate plot, NO_X in both 2019 and 2020 look like they are approximately normally distributed.

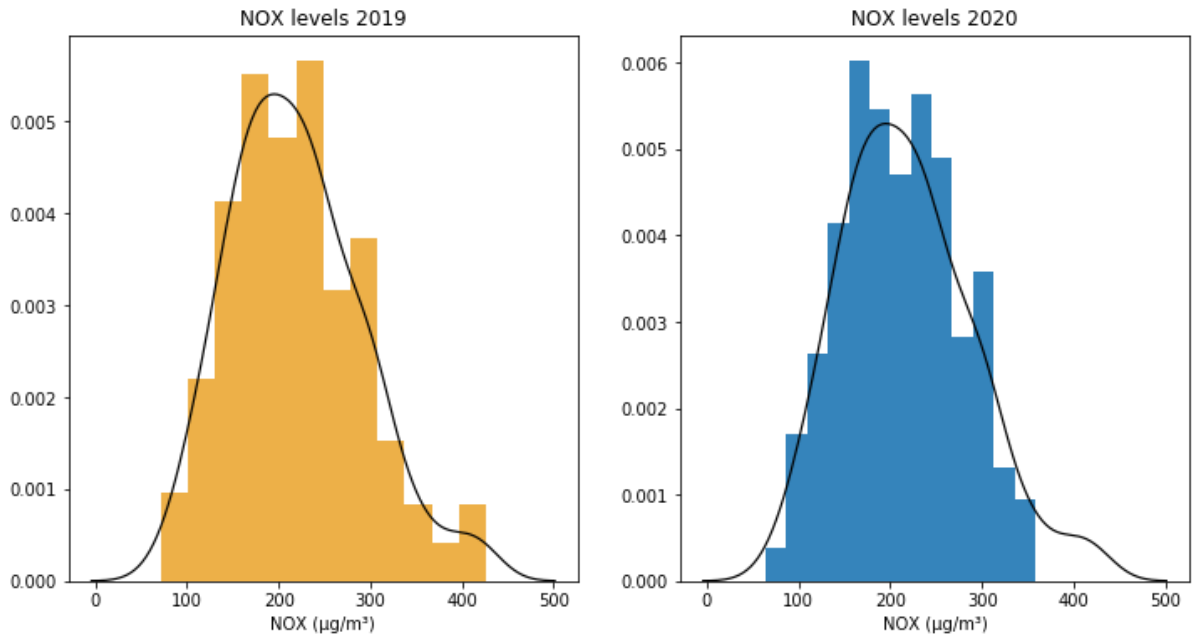


Figure 12: Distributions of NO_X in 2019 and 2020, with kernel density estimate overlayed.

The overlayed normal distributions shown in the following figure fit the data very well, so it can be concluded that the data is normally distributed.

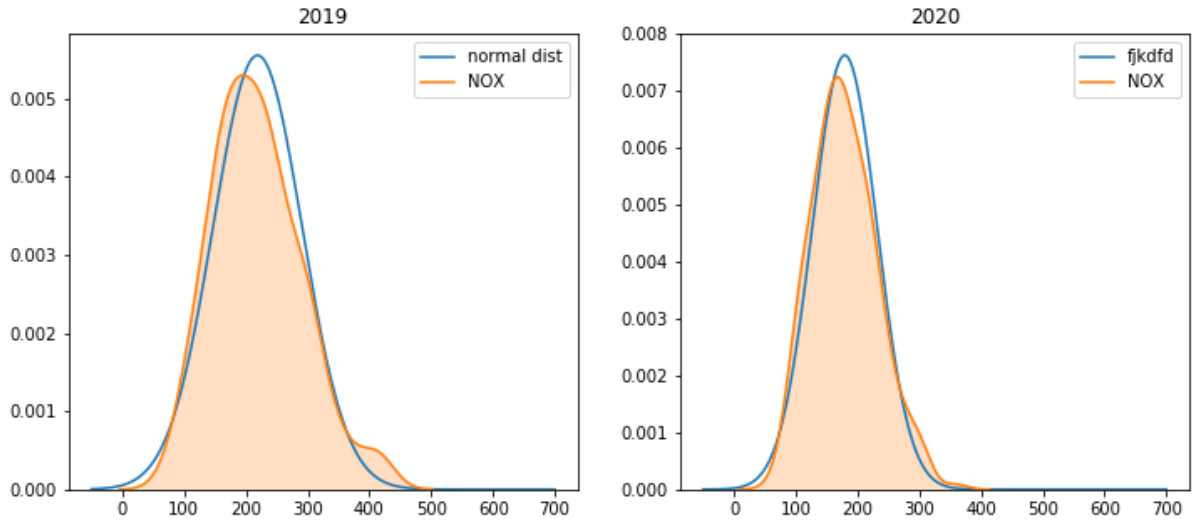


Figure 13: comparison of NO_X KDE plot to normal distribution.

Hypothesis Testing

The performed T tests showed that there is evidence to support the claims that the mean concentrations of CO, NO_2, NO_X, $PM_{2.5}$ were higher in 2019 than in 2020 at the 5% level of significance ($\alpha = 0.05$), for the date range that I selected (23 January - 30 September). The T test for O_3 showed that for this date range, there is insufficient evidence to say that there is a significant increase in the mean concentration from 2019 to 2020. I have summarised the individual t tests as follows:

1. CO

From 2019 to 2020, the mean concentration of CO decreased from 74.391837 (10µg/m^3) to 59.618474 (10µg/m^3). The one tailed hypothesis test I conducted had the following hypotheses:

H_0: $\mu_{2019} = \mu_{2020}$

H_1: $\mu_{2019} > \mu_{2020}$

The hypothesis test resulted in a p value = 5.330836×10^{-18}, which is far less than the alpha value, so we reject the null hypothesis. This is strong evidence to show that the mean concentration of CO in 2019 was significantly higher than that in 2020.

2. NO_2

From 2019 to 2020, the mean concentration of NO_2 decreased from 75.8122445 (µg/m^3) to 62.301205 (µg/m^3). The one tailed hypothesis test I conducted had the following hypotheses:

H_0: $\mu_{2019} = \mu_{2020}$

H_1: $\mu_{2019} > \mu_{2020}$

The hypothesis test resulted in a p value = 4.738092×10^{-11}, which is far lower than the alpha value, so we reject the null hypothesis. This is strong evidence to show that the mean concentration of NO_2 in 2019 was significantly higher than that in 2020.

3. O_3

From 2019 to 2020, the mean concentration of O_3 increased from 25.318367 (µg/m^3) to 26.433735 (µg/m^3). The one tailed hypothesis test I conducted had the following hypotheses:

H_0: $\mu_{2020} = \mu_{2019}$

H_1: $\mu_{2020} > \mu_{2019}$

The hypothesis test resulted in a p value= 0.235426, which is larger than the alpha value of 0.05, so we fail to reject the null hypothesis. There is insufficient evidence to support the claim that the mean concentration of ozone is significantly higher in 2020 than in 2019.

4. $PM_{2.5}$

From 2019 to 2020, the mean concentration of $PM_{2.5}$ decreased from 24.416327 (µg/m^3) to 24.416327 (µg/m^3). The one tailed hypothesis test I conducted had the following hypotheses:

H_0: $\mu_{2019} = \mu_{2020}$

H_1: $\mu_{2019} > \mu_{2020}$

The hypothesis test resulted in a p value= 2.228337×10^{-9}, which is much lower than the alpha level of 0.05, so we reject the null hypothesis. There is evidence to support the claim that the mean concentration of $PM_{2.5}$ was significantly higher in 2019 than in 2020.

5. NO_X

From 2019 to 2020, the mean concentration of NO_X decreased from 218.897959 ($\mu g/m^3$) to 179.317267 ($\mu g/m^3$). The one tailed hypothesis test I conducted had the following hypotheses:

H_0: $\mu_{2019} = \mu_{2020}$

H_1: $\mu_{2019} > \mu_{2020}$

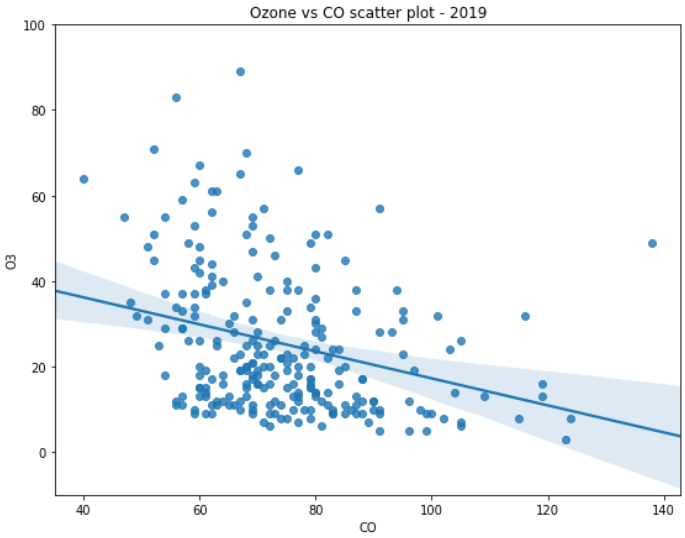
The hypothesis test resulted in a p value= 4.047722×10^{-12}, which is much smaller than the alpha level of 0.05, so we reject the null hypothesis. There is evidence to support the claim that the mean concentration of $PM_{2.5}$ was appreciably higher in 2019 than in 2020

Correlation Analysis

The correlation analysis of daily data showed poor negative correlation between CO (independent variable) and O_3 in 2019 and 2020 with their respective r^2 values = 0.078, 0.049. It showed similarly poor positive correlation between NO_2 (independent variable) and O_3, with r^2 values of 0.034 with 2019 data, and 0.060 with 2020 data.

The analysis revealed a slightly better positive correlation between $PM_{2.5}$ (independent variable) and O_3, with r^2 values of 0.109 in 2019, and 0.134 in 2020. Nitrogen oxides had the strongest correlation with O_3. NO_X displayed a positive correlation with r^2 values = 0.275 in 2019 and 0.331 in 2020.

The regression scatter plots for the above analyses are shown below:



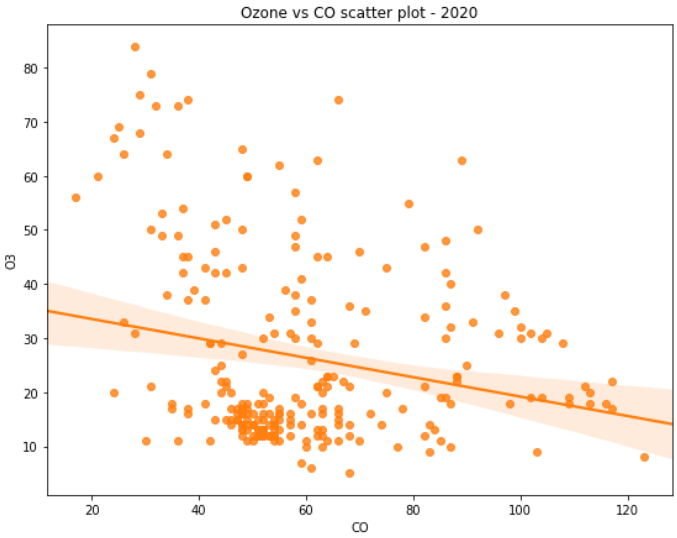


Figure 14: regression plots of O_3 vs CO - 2019 and 2020

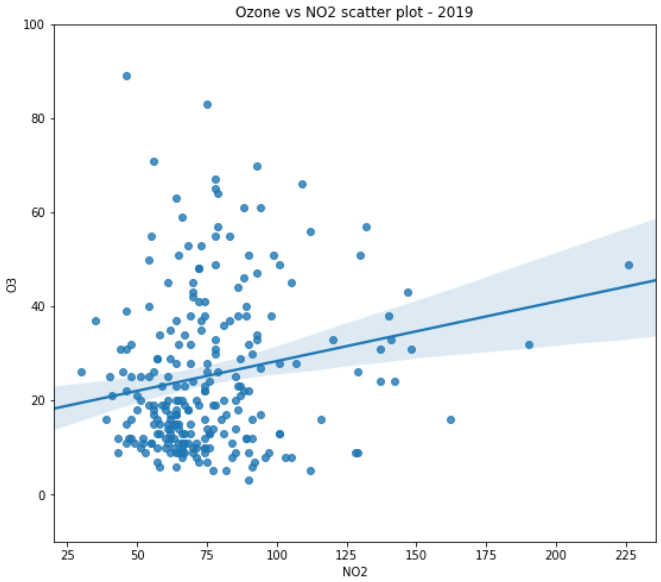
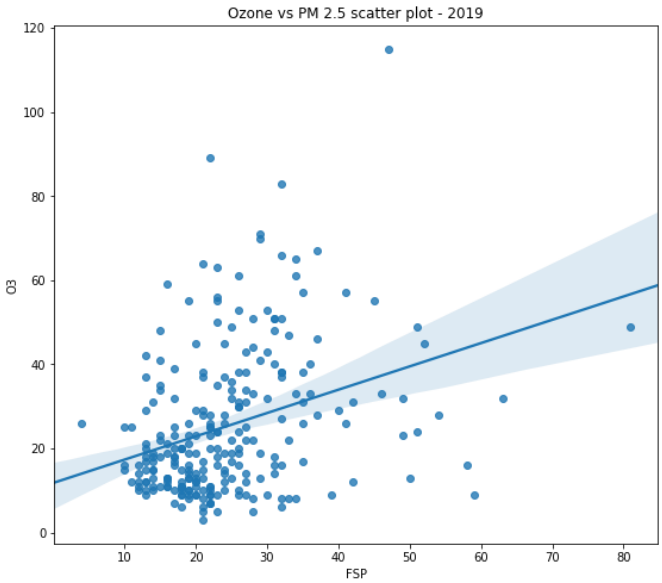


Figure 15: regression plots of O_3 vs NO_2 - 2019 and 2020



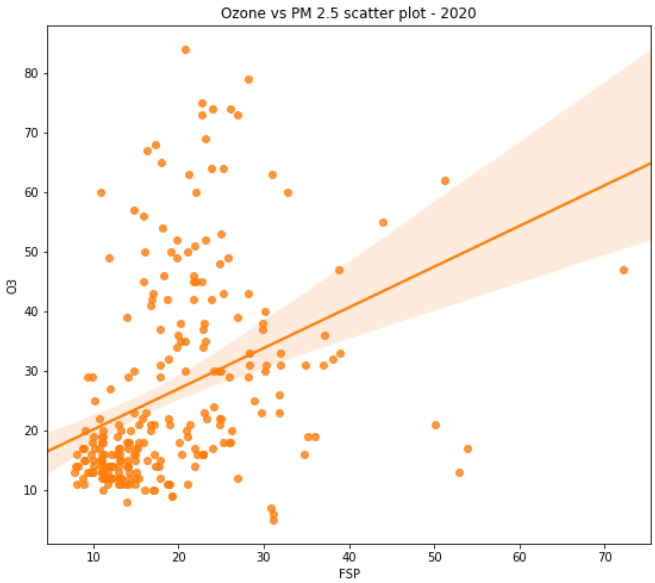


Figure 16: regression plots of O_3 vs $PM_{2.5}$ - 2019 and 2020

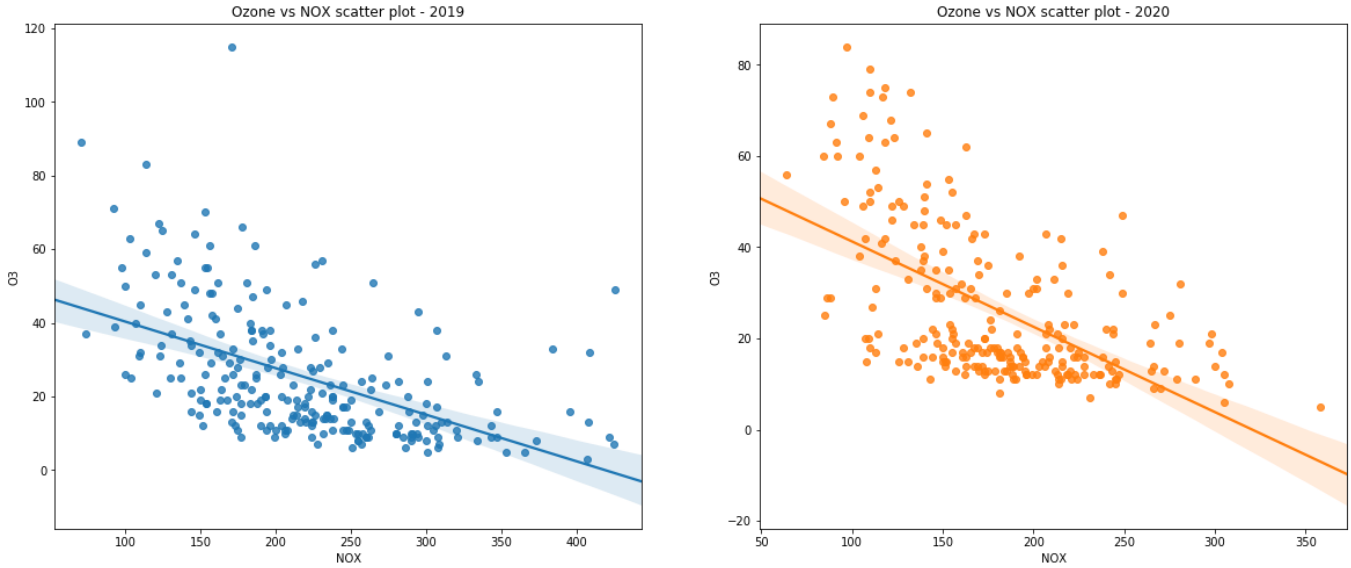


Figure 16: regression plots of O_3 vs NO_x - 2019 and 2020

Interestingly, with hourly data from randomly chosen dates the NO_2 and NO_X correlations with O_3 became much stronger. The r^2 values for NO_2 - ozone on the two dates are 0.816, 0.624. The r^2 values for NO_X - ozone on the two dates are 0.755, 0.619

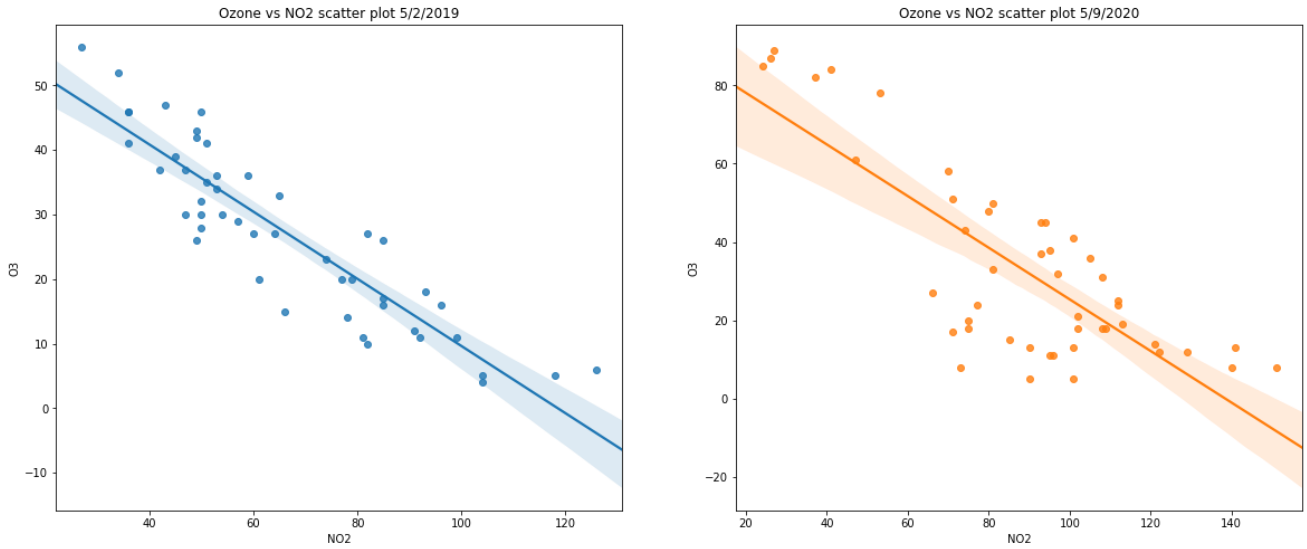


Figure 17: regression plots of O_3 vs NO_2 - hourly data

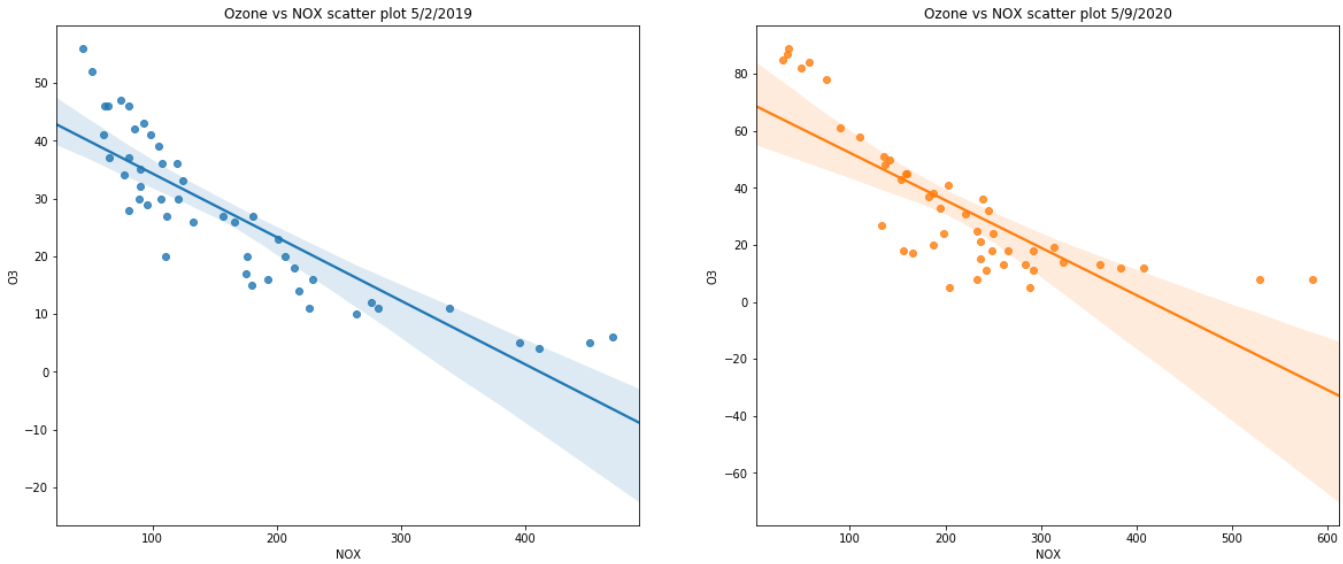


Figure 18: regression plots of O_3 vs NO_X - hourly data

Discussion

1. Distributions of Data

Most of the distributions fit normal distributions quite well, which is to be expected, as Gaussian/normal distributions occur naturally quite often in the world. However, the distribution of CO in 2020, and the distribution of NO_2 in 2019 did not fit their normal distributions perfectly. This appears to be due to a degree of randomness in the measurements, that cannot be accounted for. This tends to happen in cases such as this one, when raw data is collected directly, and outliers are not expelled from the analysis. Still, the fit of the data is still satisfactory to conduct reasonably accurate probability calculations using it.

2. Hypothesis Testing

The hypotheses for each parameter here, were formulated with regards to their mean concentrations in 2019 compared to 2020. For CO, NO_2, $PM_{2.5}$ (FSP) the t tests concluded that there were indeed significantly larger mean values for these parameters in 2019, as compared to 2020. This was as expected, as the mean values for each of these parameters differed significantly from 2019 to 2020. Thus, we can conclude that COVID 19 may have indeed led to reduced levels of pollution in Hong Kong as well.

However, one thing to remember is that correlation does not mean causation. It is indeed possible that in this analysis, the occurrence of COVID 19 and lowered pollution was simply a coincidence, and that COVID 19 did not actually affect the pollution levels.

For O_3 the t test showed that there wasn't any major increase in the mean value from 2019 to 2020 at the 5% alpha level. After taking a look at the mean value of O_3 in 2019, and its mean

value in 2020, this was also explainable. The mean value was higher by only ~ 3.16% in 2020, which is quite small. It is also possible that a larger value of alpha may have shown that there was a significant increase in the mean.

Another consideration to be taken into account is that O_3 is not directly emitted by vehicles/power plants, so while the reduction of the use of these may result in lowered levels for other pollutants, they may not necessarily affect the levels of O_3 as much.

It should also be understood that the complete data for 2019 and 2020 was not analysed. I deliberately selected a subset of the available data to better understand how the COVID 19 pandemic affected air pollution.

3. Correlation analysis

Weak correlations were discovered between CO and O_3, NO_2 and O_3, $PM_{2.5}$ and O_3, and NO_X and O_3, when daily data was considered. This was puzzling at first, but then I realised that certain correlations may occur only for certain time periods. The averaging of hourly data to obtain daily averages may thus obscure proper correlations.

This is likely the reason that NO_2 and NO_X showed much stronger correlations with O_3 with the hourly data. This occurs as the presence of NO_2 and NO_X in the air inhibits the formation of ozone beyond a point.

Conclusion

1. CO, NO_2 and NO_X fitted normal distributions.
2. O_3 and $PM_{2.5}$ fitted chi square distributions.
3. There were significantly higher mean concentration values in 2019 than in 2020 for CO, NO_2, $PM_{2.5}$, and NO_X at the 5% significance level.
4. There was no appreciably higher mean value of O_3 in 2020 at the 5% significance level.
5. There were weak correlations between Weak correlations were discovered between CO and O_3, NO_2 and O_3, $PM_{2.5}$ and O_3, and NO_X and O_3, with the use of only daily data.
6. Strong correlations between NO_2 and O_3 , NO_X and O_3 appeared with the use of the hourly data.

Recommendations

1. In order to lower air pollution levels on a regular basis, people should try and replicate the effects of COVID restrictions, maybe on a weekly or monthly basis. Less usage of private and public transport, would reduce pollution significantly.

2. The population should be more aware of the levels of pollution in the city, and cohesively attempt to lower the pollution levels to keep themselves safe.

References

1. https://aaqr.org/articles/aaqr-20-07-covid-0416
2. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjWxqzjx-7tAhXJDaYKHSthDAAQFjAAegQIAhAC&url=https%3A%2F%2Fwww.sciencedirect.com%2Fscience%2Farticle%2Fabs%2Fpii%2FS0269749120332887&usg=AOvVaw0jUGaXk7mhziNXVe683u6o

Appendix

All the code used in this project is available at my GitHub repository:
https://github.com/Milind220/SEE2003-project.git