STA9750TermProject

Shawn Meng, Rongnan He, Szufan Chen, Jiarui Guo

May 22 2021

Contents

Background

Air pollution is a leading environmental threat to the health of urban populations overall and specifically to New York City residents. Clean air laws and regulations have improved the air quality in New York and most other large cities, but several pollutants presented in the air are at levels that are harmful. In 2020, The emergence of a severe COVID-19 pandemic has posed a severe threat to human health and adversely affected all aspects of life, resulting in the implementation of lockdown in activities. While recession or economic slowdown will adversely affect countries' ongoing efforts towards climate mitigation, a significant improvement has been observed in air quality.

In this project, we aim to highlight the air pollution in New York City between 2010 to 2020. Our data is obtained from Wikipedia. It focuses on common air pollutants—fine particulate matter (PM2.5). Around the world, there are reports showed that the shutdown may or may not improved the air quality around the world.[1] The EPA(United States environmental protection agency) is tracking five common pollutants nationwide. In this project, we will focus on PM2.5 which are closely related to the traffic. It is Our data was access from EPA website. We will focus our analysis on PM2.5 because PM2.5 has a strong implication on human health.[2]

Objective

1. We want to see if there is a significant impact of the NY PAUSE executive order from the governor Cuomo on the air pollutant levels of PM2.5 in NYC.
2. We are also interested in the long term trend and seasonality influence of the air pollutants in NYC.

Exploratory data analysis

Introduction to the dataset

We obtained the data of PM2.5 from 2004 to 2020 from EPA website.

This dataset contains 193975 observations and 20 variables.

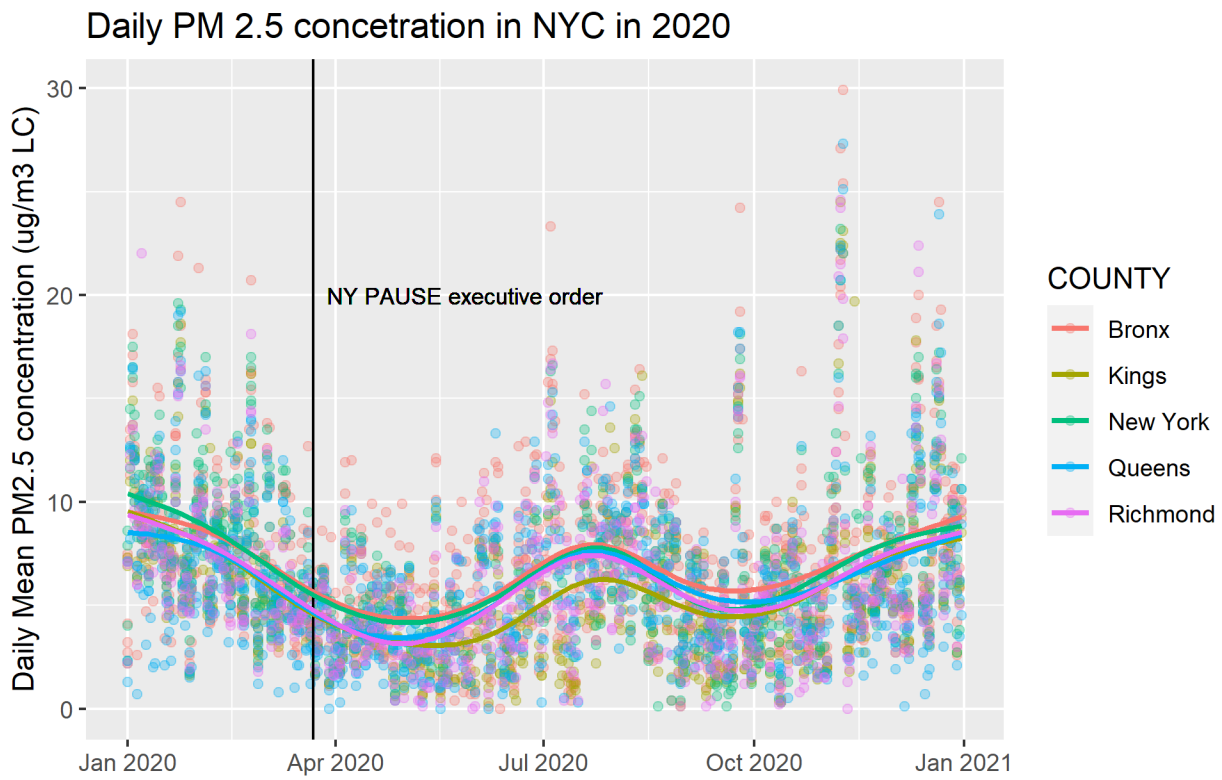
Somehow there are daily mean pm2.5 readings which are negative, these reading probably caused by faulty sensors. we will exclude these negative values. The percentage of data left is 99.717%, which is reasonable. Also, this dataset contains all data for New York state. Since we are focusing our study in NYC, we will select the data only in NYC.

```
## Rows: 193,975
## Columns: 20
## $ Date                              <chr> "01/01/2004", "01/04/2004", "01/07/20~
## $ Source                            <chr> "AQS", "AQS", "AQS", "AQS", "AQS", "A~
## $ `Site ID`                         <dbl> 360010005, 360010005, 360010005, 3600~
## $ POC                               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ `Daily Mean PM2.5 Concentration`  <dbl> 5.8, 10.9, 5.0, 10.3, 12.5, 2.5, 4.6,~
## $ UNITS                             <chr> "ug/m3 LC", "ug/m3 LC", "ug/m3 LC", "~
## $ DAILY_AQI_VALUE                   <dbl> 24, 45, 21, 43, 52, 10, 19, 52, 19, 5~
## $ `Site Name`                       <chr> "ALBANY COUNTY HEALTH DEPT", "ALBANY ~
## $ DAILY_OBS_COUNT                   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ PERCENT_COMPLETE                  <dbl> 100, 100, 100, 100, 100, 100, 100, 10~
## $ AQS_PARAMETER_CODE                <dbl> 88101, 88101, 88101, 88101, 88101, 88~
## $ AQS_PARAMETER_DESC                <chr> "PM2.5 - Local Conditions", "PM2.5 - ~
## $ CBSA_CODE                         <dbl> 10580, 10580, 10580, 10580, 10580, 10~
## $ CBSA_NAME                         <chr> "Albany-Schenectady-Troy, NY", "Alban~
## $ STATE_CODE                        <dbl> 36, 36, 36, 36, 36, 36, 36, 36, 36, 3~
## $ STATE                             <chr> "New York", "New York", "New York", "~
## $ COUNTY_CODE                       <chr> "001", "001", "001", "001", "001", "0~
## $ COUNTY                            <chr> "Albany", "Albany", "Albany", "Albany~
## $ SITE_LATITUDE                     <dbl> 42.64225, 42.64225, 42.64225, 42.6422~
## $ SITE_LONGITUDE                    <dbl> -73.75464, -73.75464, -73.75464, -73.~
```

First look of the NY PAUSE executive order impact

First, we want to look at how the NY pause executive order affects the PM2.5 level in 2020 in NYC

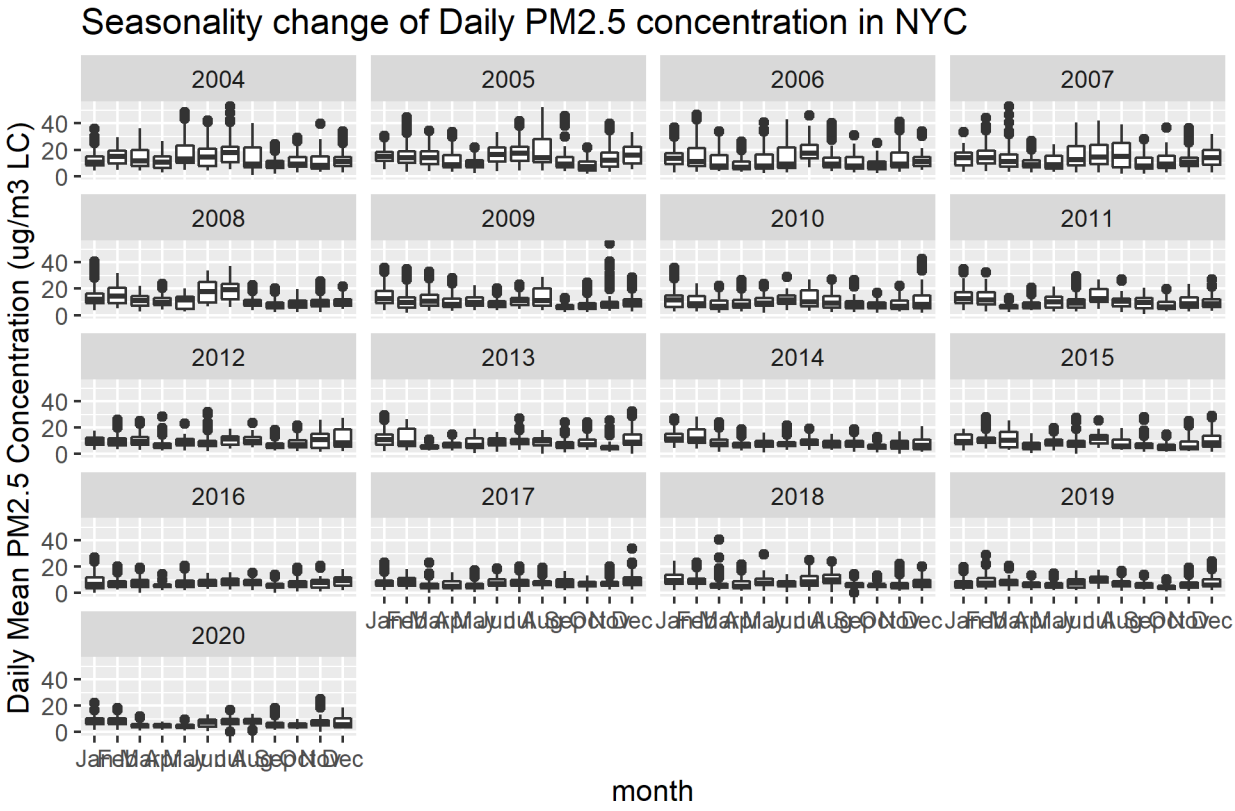
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



graph 1.1

From this graph we can see that the PM2.5 level went down after the NY PAUSE executive order. Later during the summer, the PM2.5 went up agian and drops back after around September. However, the PM2.5 level already starts to went down before the executive order. Could this phenomenon caused by the economy already slowed before the executive order or it is just seasonal change?

Seasonality changes of of PM2.5 concentration in NYC

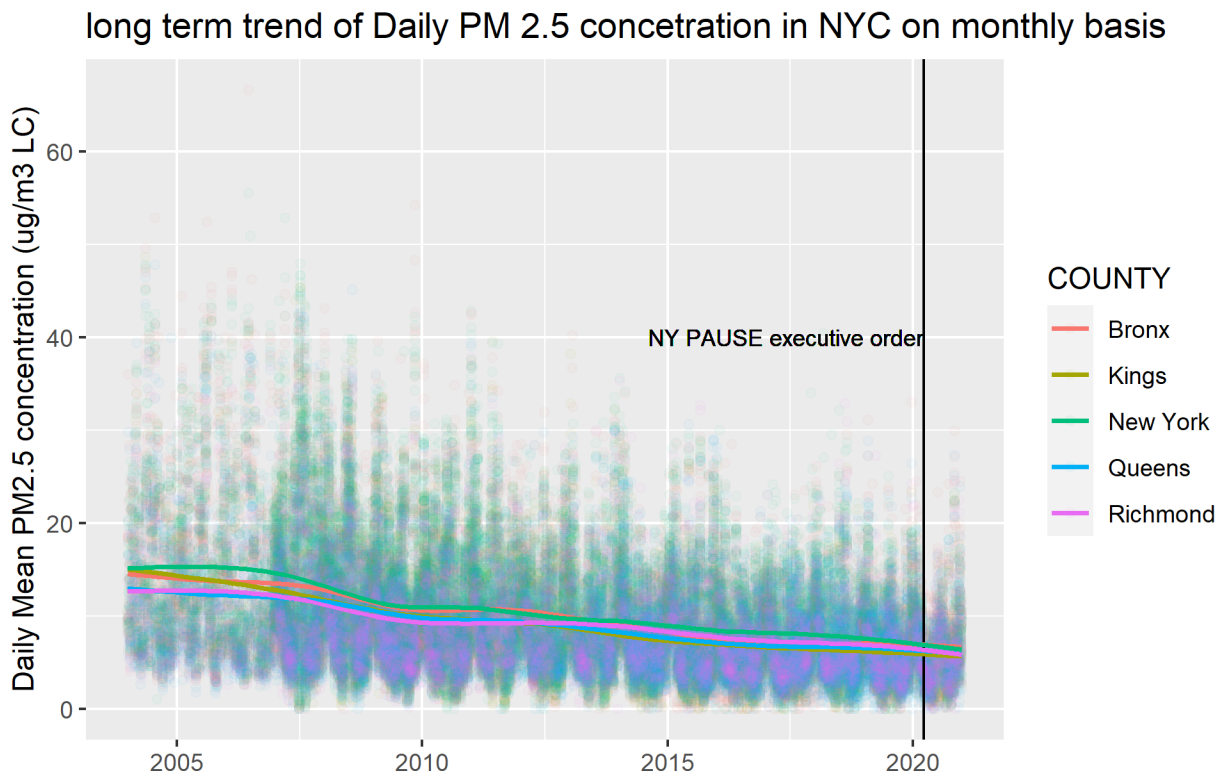


graph 1.2

From this graph, we will exam if there is any seasonality change for the PM2.5 level in New York City. We looked back the data from 2004 to 2020 on an annual basis. The level of PM2.5 tends to rise in the winter and summer months and fall in the spring and fall months. It confirms there is a seasonality change of PM2.5 levels in New York City. Furthermore, we can see the PM2.5 levels getting lower and lower over time. Next, we will take a deeper look at the long term trend for the PM2.5 levels in New York City. This result is also inline with other researchers' conclusion that PM2.5 has a strong seasonal pattern in the United States.[3]

Long term trends of PM2.5 concentration in NYC

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



graph 1.3

This scatter plot shows us the long term trend of PM2.5 levels in New York City has a clear downward trend. Also, the days of extremely high level of PM2.5 is also significantly less in recent years. That is good news for New York City. However, we can not see a significant impact of the NY pause executive order on the level of PM2.5. We will further analysis these hypothesis through modeling.

Table 1: PM2.5 statistical trend in NYC

Year	PM2.5_mean_SD	IQR(min-max)
2004	13.74±8.05	10.9(1-52.8)
2005	14.67±8.49	10.925(1.6-52.4)
2006	12.79±8.34	10(2.4-66.6)
2007	12.94±7.98	10.2(0-52.8)
2008	11.86±6.55	7.6(0-45.1)
2009	10.35±5.95	6.725(0.1-54.2)
2010	10.23±5.98	7.7(0-42.6)
2011	10.74±5.78	7.3(0-42.8)
2012	9.42±4.95	6.8(0-31.9)
2013	8.92±5.22	6.3(0-40.2)
2014	8.30±4.53	5.4(0-35.6)
2015	8.50±5.13	6.9(0-32.7)
2016	7.18±3.93	5.1(0-30)
2017	7.09±3.79	5(0-34)
2018	7.33±4.33	5.3(0-40.4)
2019	6.75±3.80	4.7(0-33)
2020	6.41±3.80	4.675(0-29.9)

From this summary table we can conclude that the PM2.5 concentration in NYC generally have a downward trend from 2004 to 2020. Not only the average level of PM2.5 went down, but also the extreme high level of PM2.5 went down in recent years. However the PM2.5 data is quite noisy as the standard deviation is large.

Modeling

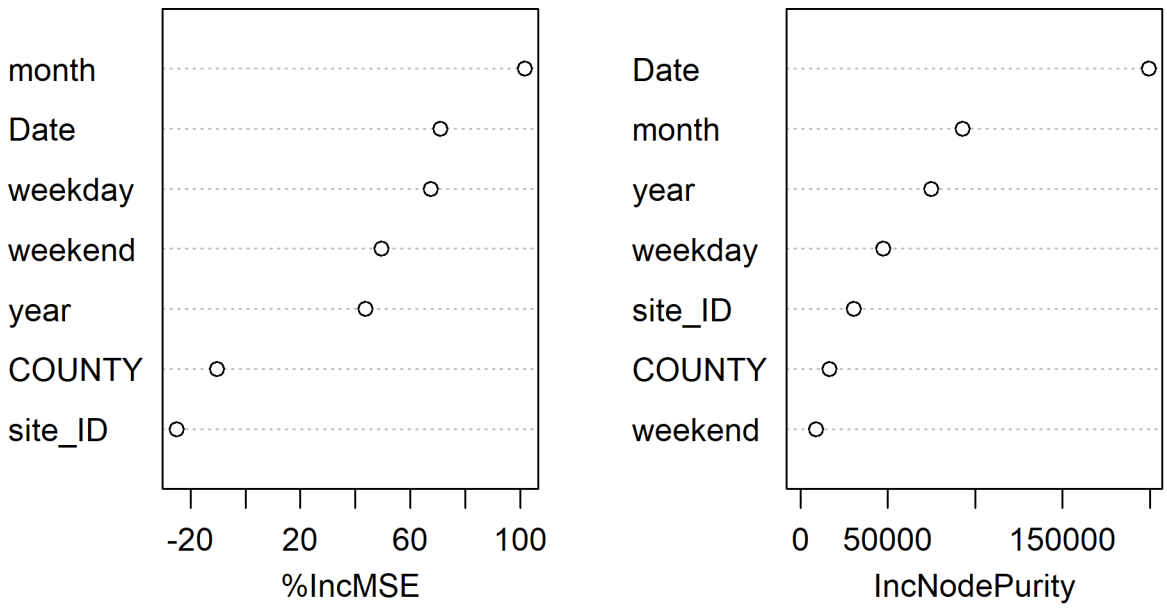
Random Forest Model

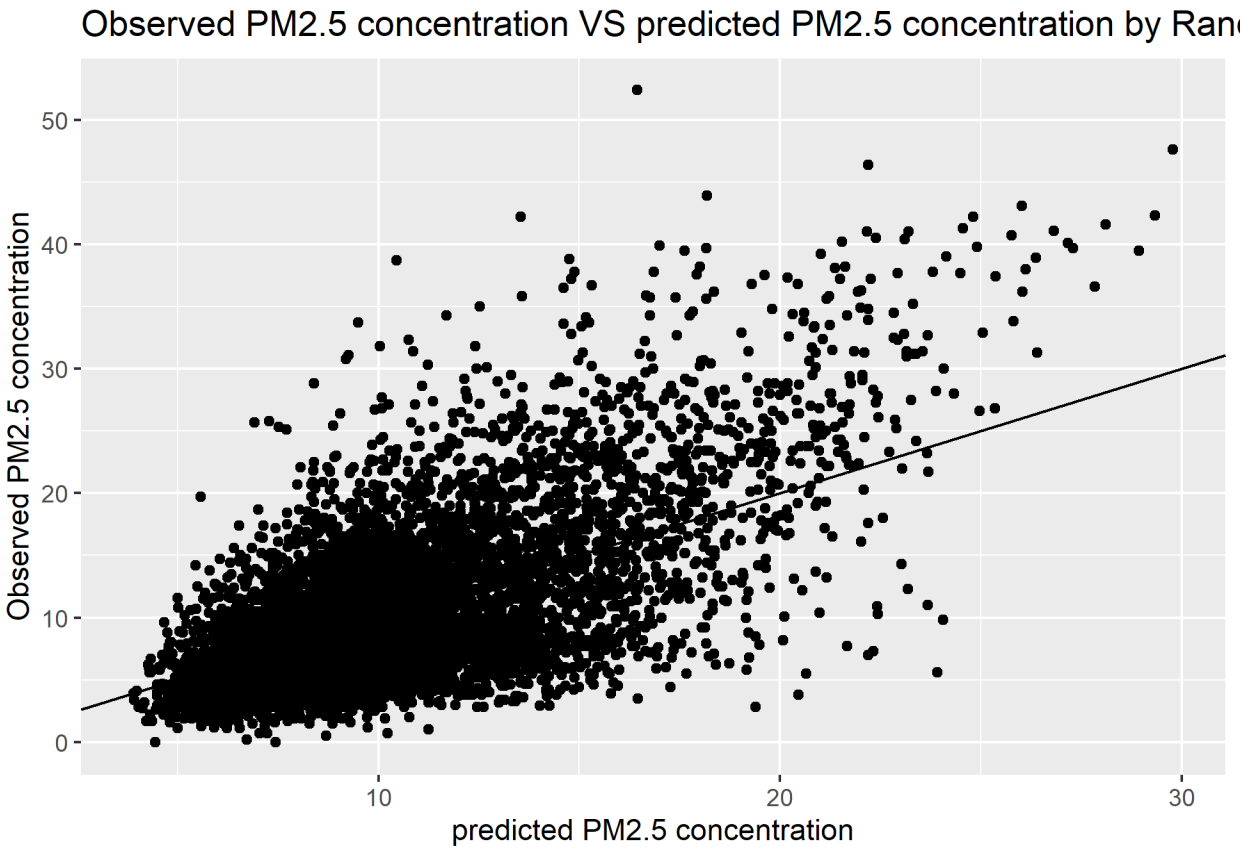
```
##
## Call:
##  randomForest(formula = PM2.5_concentration ~ Date + COUNTY +      site_ID + year + month + weekday +
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 24.74909
##                     % Var explained: 42.81
```

First, we set seed for reproducible result. The data was randomly split into 70% training set and 30% testing set. The training set was used to fit the random forest model. The testing set was used to validate the model The random forest model is used in this part of analysis. The explanatory variables are date, the county, collecting site, year, month, weekday and weekend. The dependent variable is the level of PM2.5. The RMSE of the training set is 4.103 The RMSE of the testing set is 4.958.

Table 2: importance of variables in Random Forest Model

	%IncMSE	IncNodePurity
Date	70.78798	199212.191
COUNTY	-10.36690	16738.709
site_ID	-25.18209	30684.024
year	43.75198	74692.088
month	101.58905	92813.081
weekday	67.36367	47506.297
weekend	49.42857	8870.814


PM2.5_rf_mdl2
month
Date
weekday
weekend
year
COUNTY
site_ID
-20
20
60
100
%IncMSE
Date
month
year
weekday
site_ID
COUNTY
weekend
0
50000
150000
IncNodePurity



From the importance plot we can see that the variables associate with time overshadow the variables associate with geographical location. I think it is reasonable as our study was focused in New York city. The monitoring station are relatively close. The date when the data was collected is the most important predictor for the PM 2.5 concentration according to our random forest model.

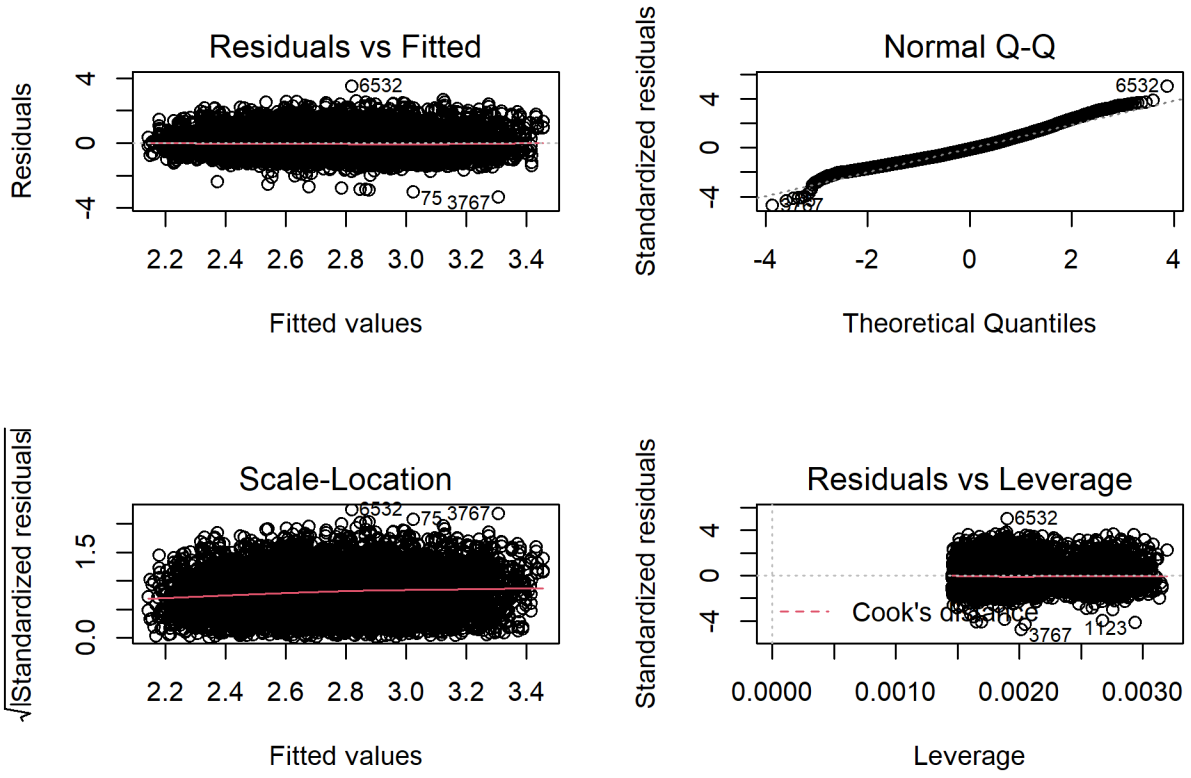
Linear Regression Model

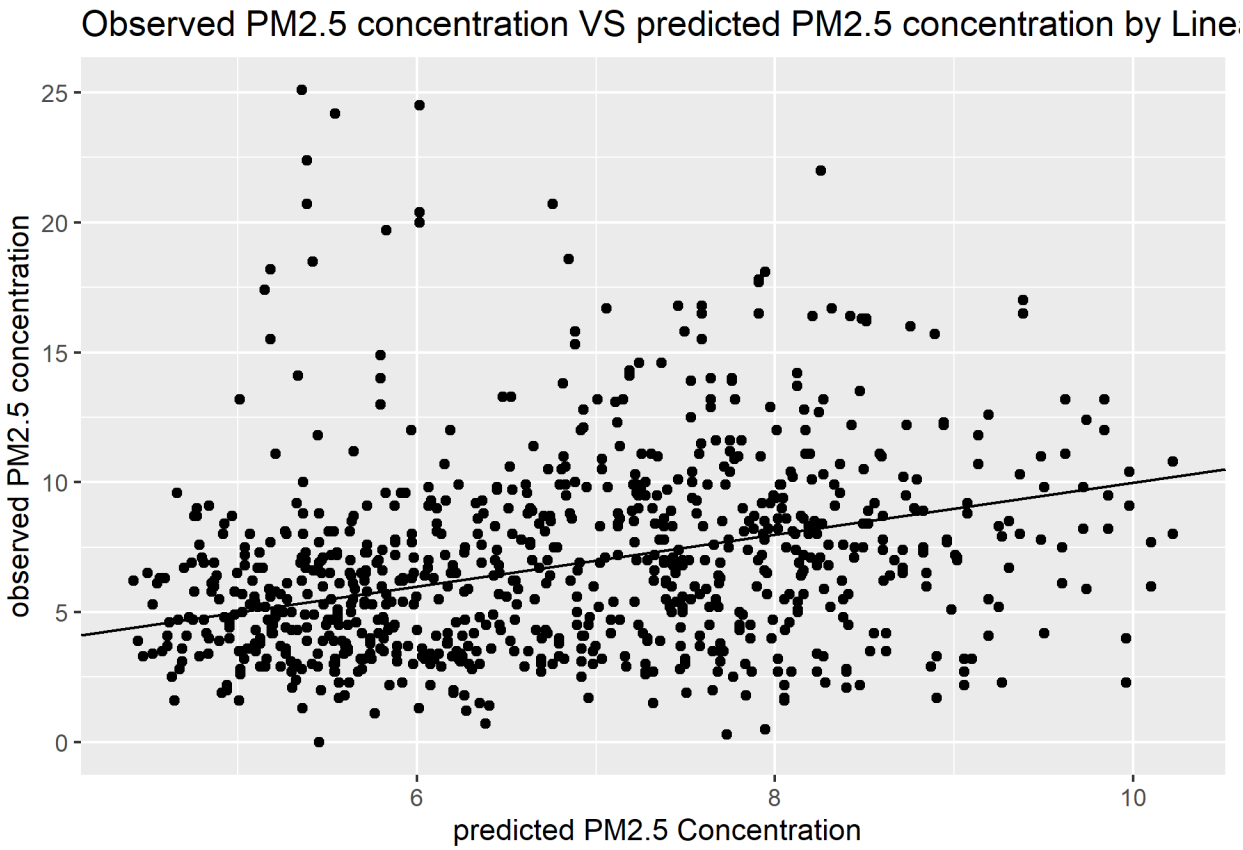
Table 3: The linear regression model

term	estimate	std.error	statistic	p.value
(Intercept)	-4504.2294582	604.9726310	-7.4453442	0.0000000
Date	-0.0063697	0.0008408	-7.5760042	0.0000000
I(month)2	0.2515545	0.0451385	5.5729539	0.0000000
I(month)3	-0.0017222	0.0617840	-0.0278742	0.9777631
I(month)4	0.0230716	0.0835871	0.2760183	0.7825404
I(month)5	0.3960000	0.1072179	3.6934136	0.0002226
I(month)6	0.6935993	0.1317204	5.2656929	0.0000001
I(month)7	1.1573416	0.1564554	7.3972602	0.0000000
I(month)8	1.1690963	0.1819479	6.4254462	0.0000000
I(month)9	1.0991743	0.2071241	5.3068402	0.0000001
I(month)10	1.1600937	0.2324710	4.9902717	0.0000006
I(month)11	1.4307290	0.2578831	5.5479749	0.0000000
I(month)12	2.0005497	0.2832501	7.0628392	0.0000000
year	2.2888924	0.3070984	7.4532857	0.0000000
I(COUNTY)Kings	0.0000560	0.0298761	0.0018750	0.9985040

term	estimate	std.error	statistic	p.value
I(COUNTY)New York	0.1466526	0.0223225	6.5697083	0.0000000
I(COUNTY)Queens	-0.1317111	0.0210000	-6.2719666	0.0000000
I(COUNTY)Richmond	-0.0980923	0.0304997	-3.2161683	0.0013037

As we observed from graph 1.3, the PM2.5 concentration and time seems to have a linear association from 2013-2020. Thus we tried to fit a linear regression model using the square root of PM2.5 concentration as dependent variable. We tried to using the date, month, year and the county where the data was taken.





We take the squared root of PM2.5 concentration because the data is skewed, the linear model fit better when take the square root of the PM 2.5 concentration to mitigate the impact of extreme values. We confirm these by looking at the Q-Q plot as the residues are almost normally distributed. Next we used the linear regression model build upon data from 2013 to 2019 to predict the PM2.5 concentration in 2020. Then we compare the predicted value to the observed value to calculate RMSE, which is 3.702. From our model, the NY pause executive order does not have a significant impact on the level of PM2.5 concentration.

```
##
## Call:
## lm(formula = sqrt_PM2.5_concentration ~ Date + I(month) + year +
##     I(COUNTY), data = PM2.5_NYC_lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3066 -0.5001 -0.0747  0.4150  3.5374
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4.504e+03  6.050e+02  -7.445 1.06e-13 ***
## Date              -6.370e-03  8.408e-04  -7.576 3.92e-14 ***
## I(month)2          2.516e-01  4.514e-02   5.573 2.58e-08 ***
## I(month)3         -1.722e-03  6.178e-02  -0.028 0.977763
## I(month)4          2.307e-02  8.359e-02   0.276 0.782540
## I(month)5          3.960e-01  1.072e-01   3.693 0.000223 ***
## I(month)6          6.936e-01  1.317e-01   5.266 1.43e-07 ***
## I(month)7          1.157e+00  1.565e-01   7.397 1.51e-13 ***
## I(month)8          1.169e+00  1.819e-01   6.425 1.38e-10 ***
```

```
## I(month)9          1.099e+00  2.071e-01   5.307 1.14e-07 ***
## I(month)10         1.160e+00  2.325e-01   4.990 6.14e-07 ***
## I(month)11         1.431e+00  2.579e-01   5.548 2.97e-08 ***
## I(month)12         2.001e+00  2.833e-01   7.063 1.75e-12 ***
## year               2.289e+00  3.071e-01   7.453 9.94e-14 ***
## I(COUNTY)Kings     5.602e-05  2.988e-02   0.002 0.998504
## I(COUNTY)New York  1.467e-01  2.232e-02   6.570 5.32e-11 ***
## I(COUNTY)Queens   -1.317e-01  2.100e-02  -6.272 3.73e-10 ***
## I(COUNTY)Richmond -9.809e-02  3.050e-02  -3.216 0.001304 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7007 on 9015 degrees of freedom
## Multiple R-squared:  0.1248, Adjusted R-squared:  0.1231
## F-statistic: 75.61 on 17 and 9015 DF,  p-value: < 2.2e-16
```

This model does not show us a strong correlation as we can see from linear regression summary as the adjusted R is only 0.1231. As some other researchers suggests, weather and other meteorological variations contribute to the variability of the PM2.5 concentration too.For example, the northeast wind in the winter are strongly associated with low PM2.5 level[4] Furthermore, PM 2.5 has multiple sources and they tends to vary depend on the season. [5]

Conclusion and further analysis

The outbreak of the COVID-19 pandemic has adversely affected all aspects of life and poses a severe threat to human health and economic development. New York City administration enacted a strict isolation decision at the end of March 2020 to tackle the COVID-19, creating a unique opportunity to assess air quality. Therefore, we investigated the impact of the lockdown on air quality in New York City. We evaluated the air pollutants concentration-PM2.5, during the lockdown and compared them with pre-COVID-19. According to our analysis, the COVID-19 related lockdown does not significantly impact the PM2.5 level in New York city. However, we discover the PM2.5 level in New York city went down gradually over the past 17 years. For future investigation, it may be a good idea to include meteorological data and digs further on the source of PM2.5 in order to achieve a better model.

Reference

[1]Daniella Rodríguez-Urrego, Leonardo Rodríguez-Urrego, Air quality during the COVID-19: PM2.5 analysis in the 50 most polluted capital cities in the world, Environmental Pollution, Volume 266, Part 1, 2020, 115042, ISSN 0269-7491

2]Shaolong Feng, Dan Gao, Fen Liao, Furong Zhou, Xinming Wang, The health effects of ambient PM2.5 and potential mechanisms, Ecotoxicology and Environmental Safety, Volume 128, 2016, Pages 67-74, ISSN 0147-6513, https://doi.org/10.1016/j.ecoenv.2016.01.030.

[3]Michelle L. Bell, Francesca Dominici, Keita Ebisu, Scott L. Zeger, and Jonathan M. Samet, Spatial and Temporal Variation in PM2.5 Chemical Composition in the United States for Health Effects Studies, Environmental Health Perspectives Vol. 115, No.7 2007

[4]Arthur T. DeGaetano, Owen M. Doherty, Temporal, spatial and meteorological variations in hourly PM2.5 concentration extremes in New York City, Atmospheric Environment, Volume 38, Issue 11, 2004, Pages 1547-1558, ISSN 1352-2310

[5]Kazuhiko Ito, Nan Xue, George Thurston, Spatial variation of PM2.5 chemical species and source-apportioned mass concentrations in New York City, Atmospheric Environment, Volume 38, Issue 31, 2004, Pages 5269-5282, ISSN 1352-2310