COVID-19 Investigators

Big Data – Spring 2021

Avinash Authipudi(aa8382)
Zijie Dong(zd2036)
Zijian Gao(zg2125)

May 7th, 2021

Abstract

This report presents the impact and analysis of Covid-19 in New York City. It includes data cleaning, data quality issues detected, data analysis and visualization.

Introduction

This report is presented for the Big Data CS-GY-6513 as part of the Project Report. Covid-19 has led to great loss of human life worldwide and presents an unprecedented challenge to public health, economy and social disruption. Millions of people risked facing poverty. It also sparked fears of an impending economic crisis and recession. Travel restrictions, social distancing and self-isolation have taken a toll on the economy and well-being of the people. The need for manufactured products decreased and demand for food supply increased due to panic buying. In an attempt to understand this effect on the life of people, we try to analyze the movement of people before and during the pandemic. We focus on public transportation, crime rates, air quality and the economic impact on restaurants in New York City in particular.

Datasets

1. Dunkin-Donats.csv (coronavirus-data/DataCleanning/Cleaned at Zijian · z632101094/coronavirus-data · GitHub):
   This is the dataset I cleaned from part one of the project using covid restaurant.csv located in Zijian's branch DataCleanning/Source Folder. Dunkin_Donats.csv contains income, date etc for many Dunkin-Donuts stores located in New York City.

2. StarBucks Covid.csv(coronavirus-data/DataCleanning/Cleaned at Zijian · z632101094/coronavirus-data · GitHub):The dataset I cleaned from part one of the project using covid restaurant.csv located in Zijian's branch DataCleanning/Source Folder. It contains many information about Starbucks stores located in New York City including each store's income and level of income etc.

3. 2020_Green_Taxi_Trip_Data.csv(coronavirus-data/2020_Green_Taxi_Trip_Data.zip at Zijie · z632101094/coronavirus-data (github.com)): This dataset contains the data of NYC green taxi trips in 2020. It contains the number of trips, the number of passengers, amount of tips, and amount of prices.

4. NYPD.csv: This dataset includes all the crimes reported to NYPD. The dataset has been filtered to include crimes committed in the year 2020. It is located in the Data Cleaning folder on Github.

5. Covid_Analysis.csv: This dataset includes the number of Covid positive cases reported in New York City. It includes the number of cases reported along with the daily count of deaths reported.

6. AirQuality_2020Data.csv: This dataset contains the air quality levels reported in New York City which includes daily count of Particulate Matter, Nitrous Oxide, Carbon Monoxide, etc.

7. MTA_BUS_Dataset.csv: This dataset includes the daily count of subway, LIRR and Bus users of New York City. It also includes the approximate number of vehicles plying on the roads of the city.

Data Cleaning and Integration

(Zijian)

- use openrefine to check all the datasets column for incorrect value or null value. If a blank column is found, change that to N/A
- use openRefine to check all the clusters can be joined together.
- for the Covid Restaurant.csv, I use openRefine to first group the Restaurant by name. use sort by count to choose the first two records with most counts and then extract them out as two files so I can use those two datasets to compare the income before and after the COVID period for the same restaurant. Two files are Dunkin-Donats.csv and StarBucks Covid.csv Many data in the restaurants.csv only have one or two rows, so I skip those data because there is nothing you can do on analysis.

(Zijie)

- I removed all the data in 2019 since we don't need it.
- I only keep lpep_pickup_datetime, passenger_count, tip_amount, total_amount. I deleted all other columns.
- I did change on the lpep_pickup_datetime so it only show months.coronavirus-data/2020_Green_Taxi_Trip_Data_Cleaned.csv at Zijie · z632101094/coronavirus-data (github.com)

(Avinash)

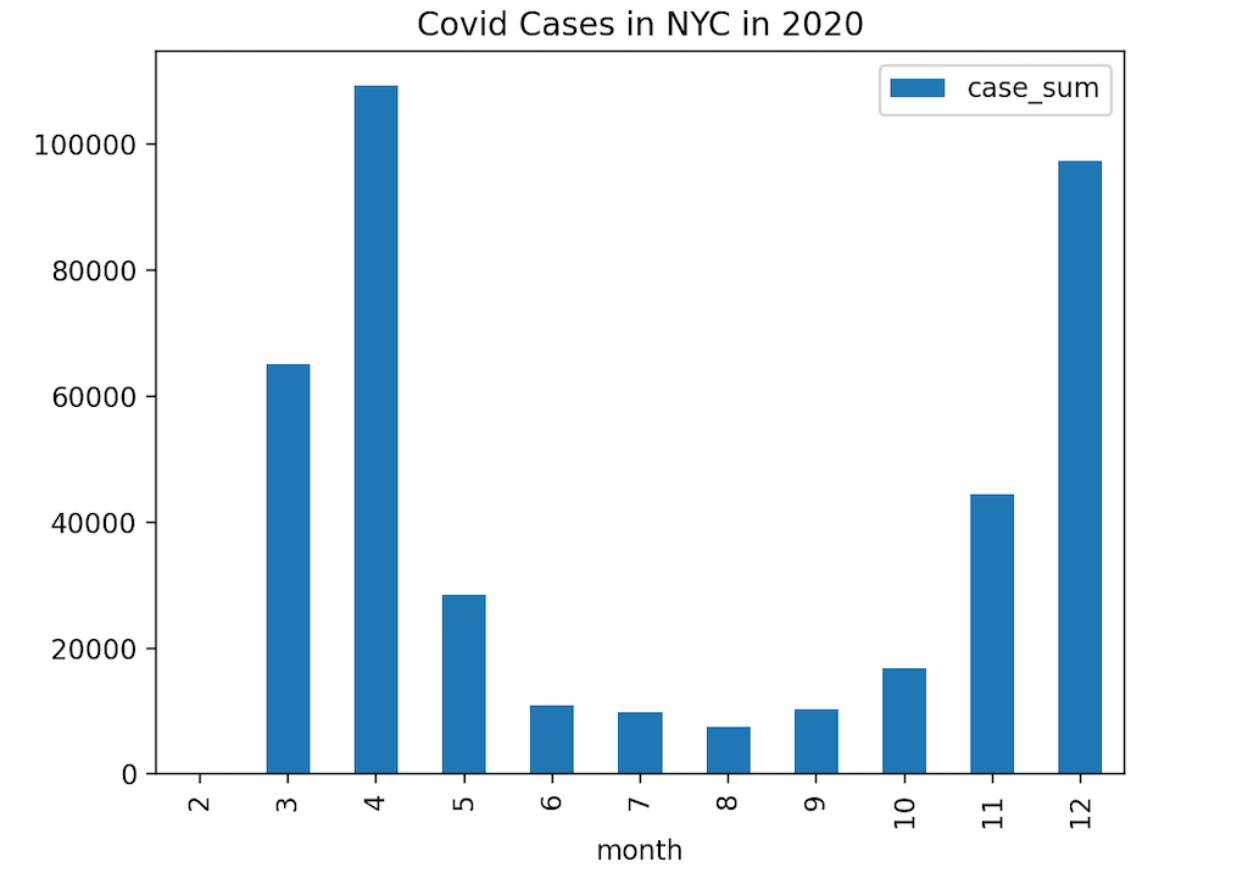
- The NYPD dataset included data from previous years which have been filtered using openrefine.
- The columns like crime report date, precinct number, etc have been retained and the unwanted columns have been removed.
- The Boroughs have been grouped together and the total crimes committed have been aggregated on a monthly basis.
- The records with null value fields have been removed from the datasets.
- The Covid_Analysis dataset included daily counts of covid cases and recoveries. The columns which reported the weekly averages of the cases and deaths have been removed using openrefine.
- A common change to the datasets include changing the date format in the dataset which is usually a string datatype into a timestamp datatype. This also has been performed using openrefine.

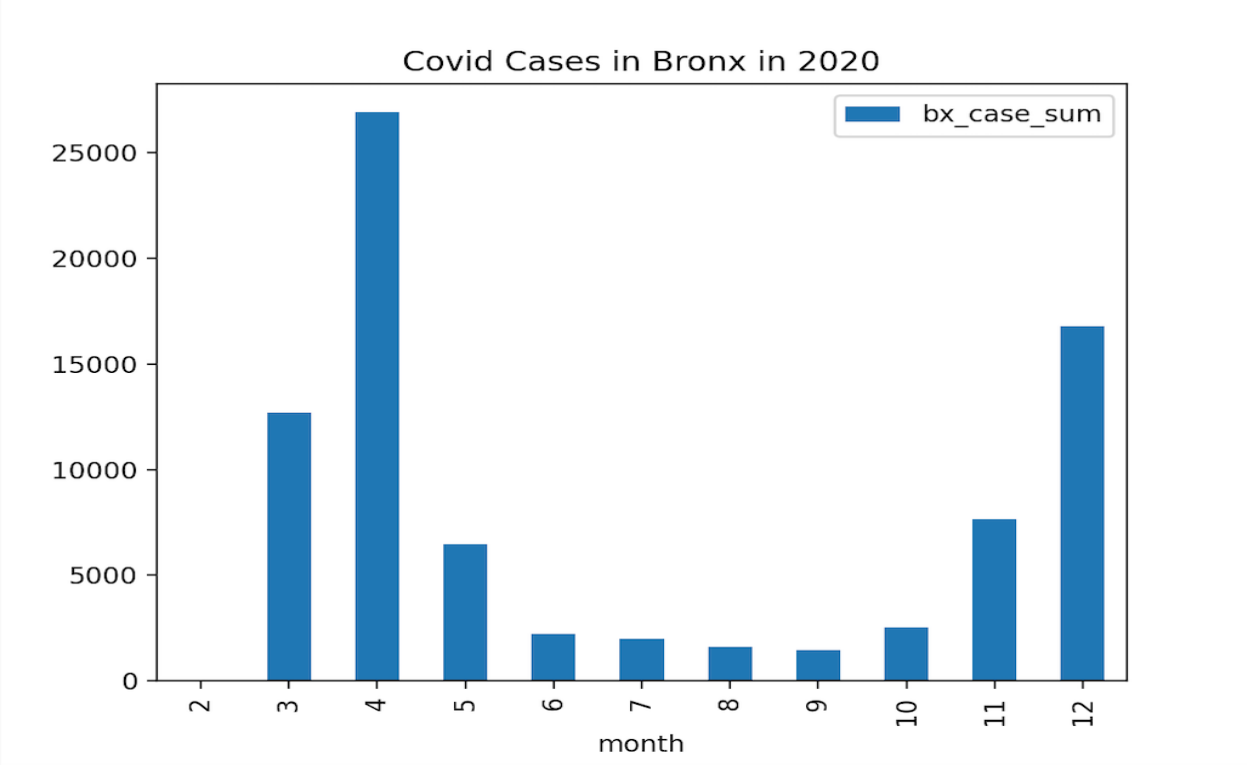
- The MTA_BUS Dataset also has been cleaned to exclude unwanted columns and only the data necessary for visualization has been retained. This is basically removal of average user data column and changing the date column into a timestamp datatype.
- The Air_Quality Dataset included air quality indexes from various cities around the world. I have filtered the dataset to include only the boroughs of New York City.

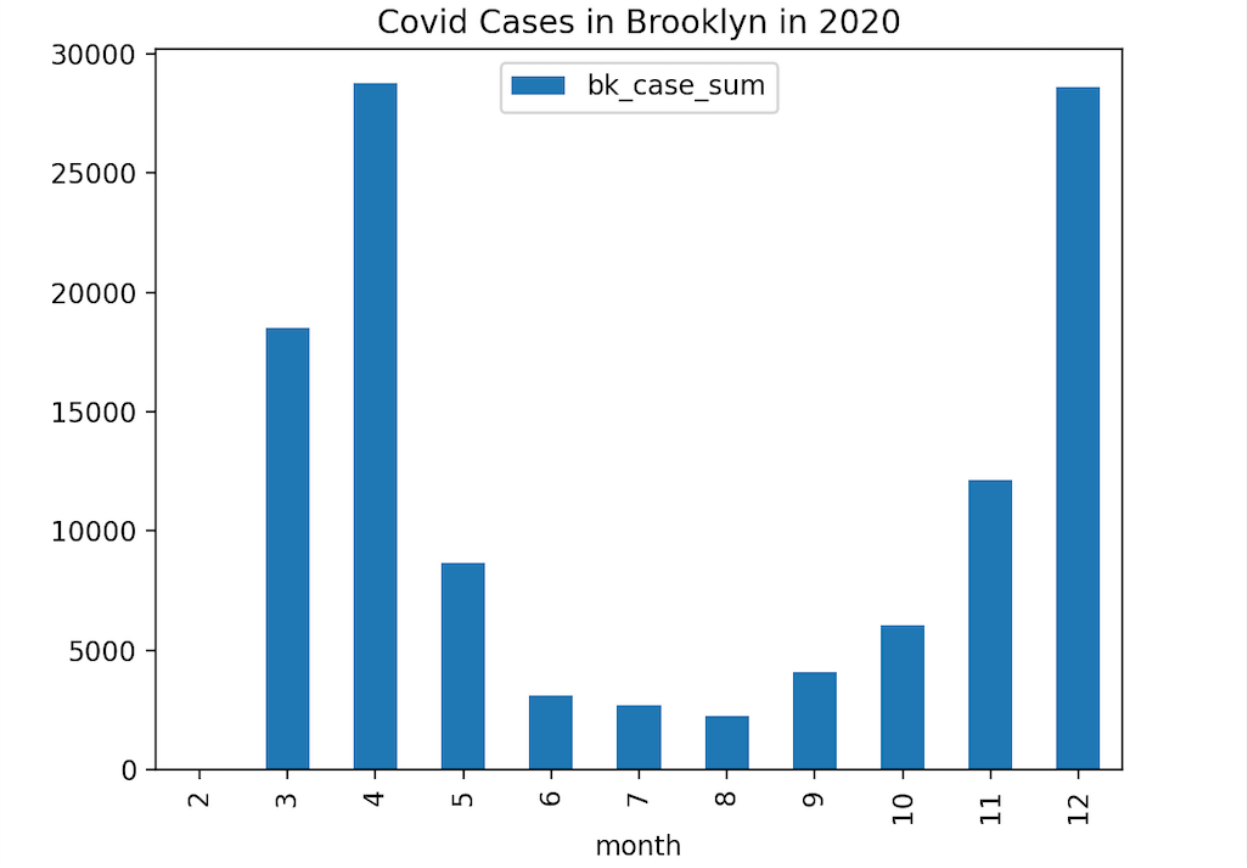
Data Analysis and Findings

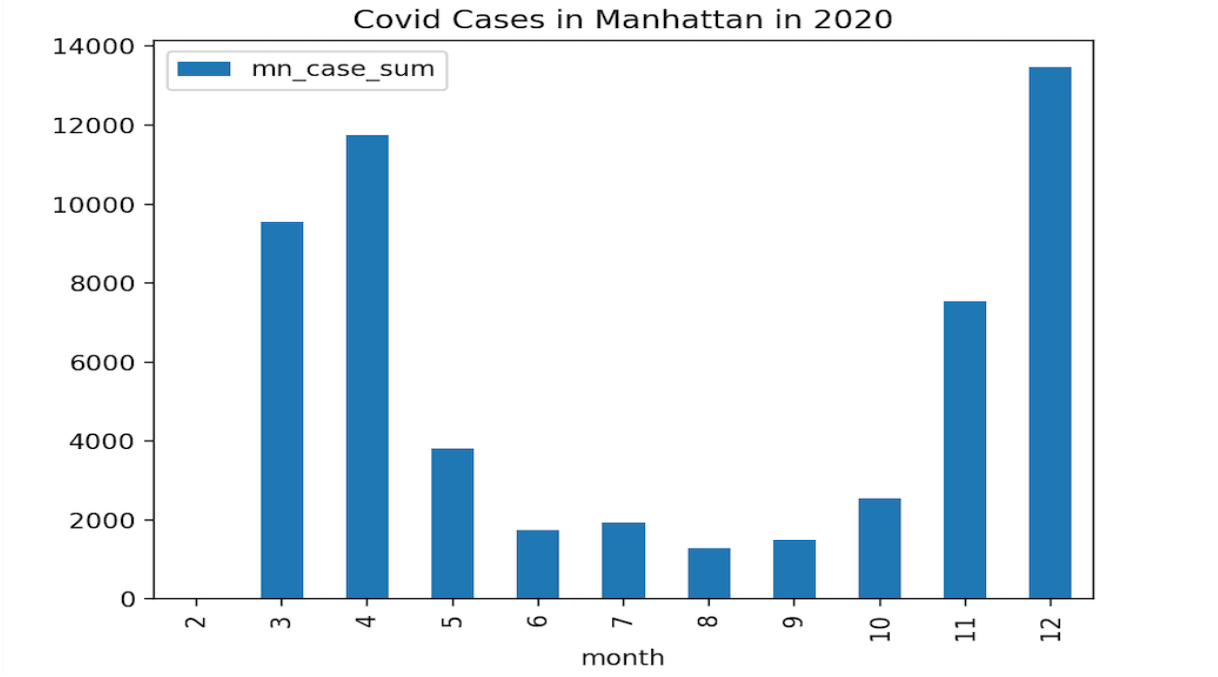
Covid_Analysis(Avinash):

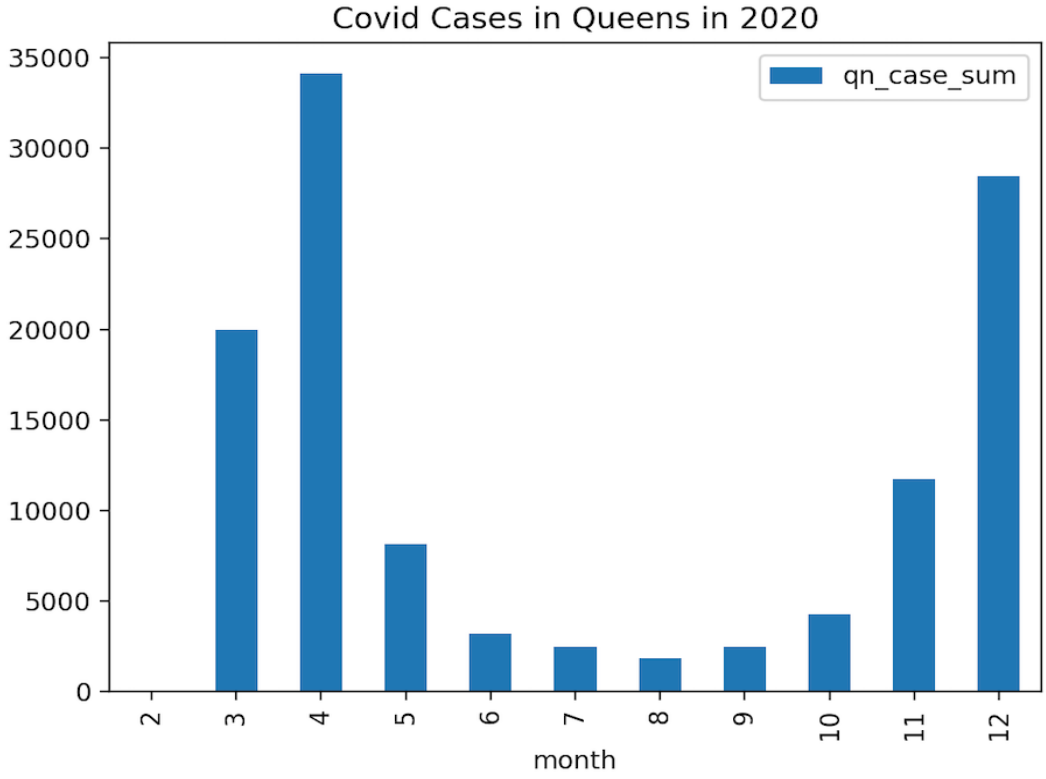
Firstly, we analyze the number of Covid-19 Cases reported in the city of New York. NYC saw its peak Covid time during the month of April. These are visualized in the below graphs. The graphs also show borough wise Covid Cases reported.

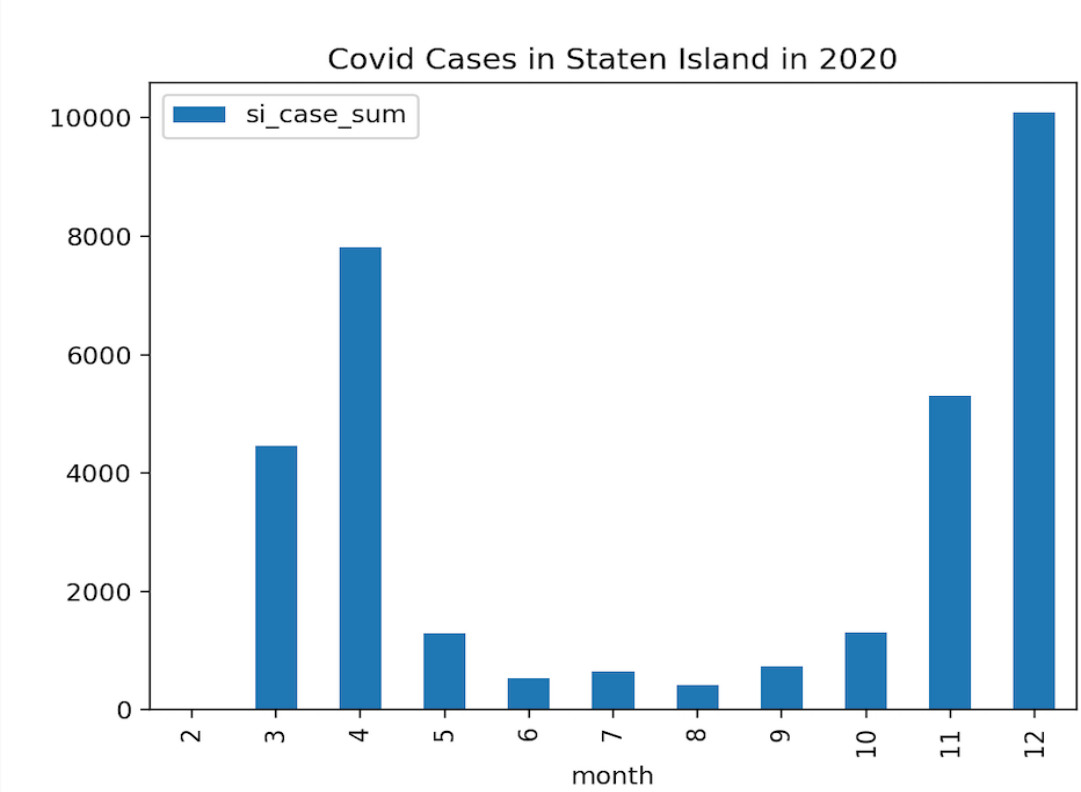



Covid Cases in Bronx in 2020
bx_case_sum
25000
20000
15000
10000
5000
0
2
3
4
5
6
7
8
9
10
11
12
month


Covid Cases in Brooklyn in 2020
bk_case_sum
30000
25000
20000
15000
10000
5000
0
2
3
4
5
6
7
8
9
10
11
12
month


Covid Cases in Manhattan in 2020
mn_case_sum
14000
12000
10000
8000
6000
4000
2000
0
2
3
4
5
6
7
8
9
10
11
12
month


Covid Cases in Queens in 2020
qn_case_sum
35000
30000
25000
20000
15000
10000
5000
0
2
3
4
5
6
7
8
9
10
11
12
month



As indicated by the above graphs, April was the peak Covid month where NYC registered a record number of cases. As a result of that, the Governor announced lockdown restrictions in NYC. Our report summarizes and visualizes the effect of the lockdown and impact of Covid-19 in NYC.

1. **Restaurant Dataset(Zijian):**
   After checking the counts for each store in datasets I choose the two biggest results to be mine analysis target. Here is the count check results for starbucks and dunkin donuts.

Counts:

Dunkin Donuts Starbucks

```
: data1.groupBy("STREET").count().show(100)
```

```
+--------------------+-----+
|              STREET|count|
+--------------------+-----+
|    EAST  170 STREET|    2|
|  COLLEGE POINT BLVD|    1|
|           48 STREET|    2|
|          150 STREET|    1|
|    EAST  149 STREET|    1|
|    ROSSVILLE AVENUE|    2|
|    EAST  138 STREET|    2|
|  QUEENS PLAZA NORTH|    1|
|    EAST   46 STREET|    1|
|     GARRISON AVENUE|    1|
|           BROADWAY |    2|
|           45 AVENUE|    1|
|         QUEENS BLVD|    1|
|        CANAL STREET|    6|
|   BEACH  129 STREET|    2|
| EAST TREMONT AVENUE|    2|
|          PARK PLACE|    1|
|   VANDERBILT AVENUE|    1|
|  NORTHERN BOULEVARD|    1|
|           26 AVENUE|    2|
|         JACKSON AVE|    1|
|      WEBSTER AVENUE|    1|
|    QUEENS BOULEVARD|    4|
|    EAST   86 STREET|    1|
|  BRUCKNER BOULEVARD|    2|
|            2 AVENUE|    3|
|        QUEENS PLZ S|    1|
|    WEST   72 STREET|    1|
|YELLOWSTONE BOULE...|    1|
|            5 AVENUE|    1|
|     CHAMBERS STREET|    1|
| CLINTONVILLE STREET|    1|
|      GREENPOINT AVE|    2|
|     FLATBUSH AVENUE|    1|
```

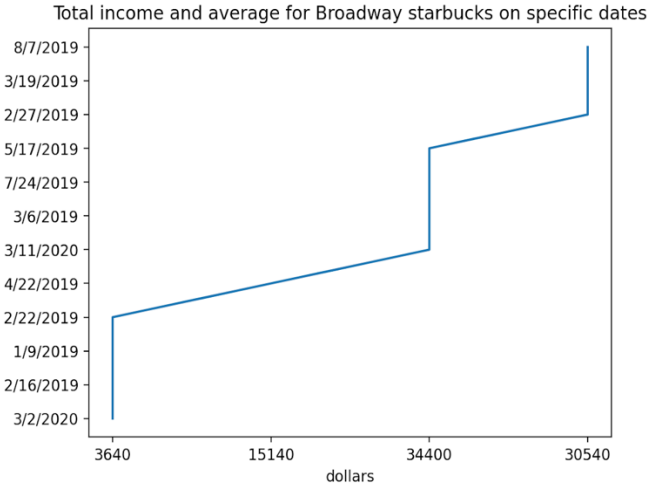
```
11]: data.groupBy("STREET").count().show(100)
```

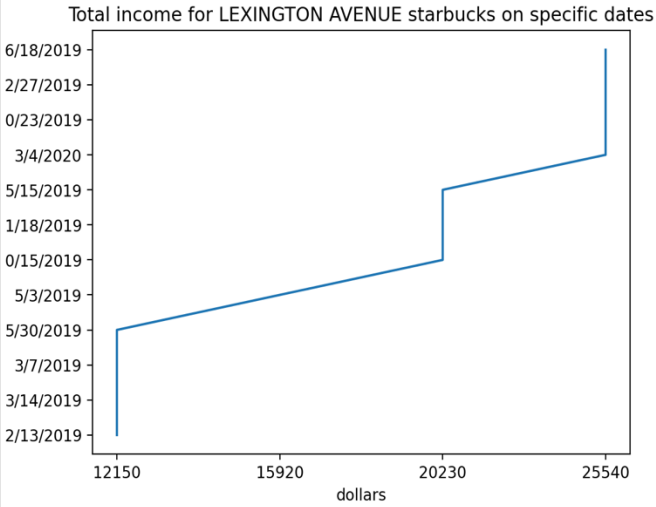
```
+--------------------+-----+
|              STREET|count|
+--------------------+-----+
|    GREENWICH STREET|    1|
|    EAST  149 STREET|    1|
|         PAGE AVENUE|    1|
|            BROADWAY|   12|
|    EAST   53 STREET|    1|
|         YORK AVENUE|    1|
|    EAST   42 STREET|    3|
|        CANAL STREET|    1|
|    EAST   80 STREET|    1|
|    QUEENS BOULEVARD|    3|
|    WEST   73 STREET|    1|
|           35 AVENUE|    1|
|            2 AVENUE|    3|
|            5 AVENUE|    1|
|         PARK AVENUE|    1|
|     COLUMBUS AVENUE|    3|
|     FLATBUSH AVENUE|    2|
|    EAST   51 STREET|    1|
|  CONTINENTAL AVENUE|    2|
|      MADISON AVENUE|    6|
|       AUSTIN STREET|    2|
|            3 AVENUE|    6|
|    LAFAYETTE STREET|    1|
|     PARK AVE FRNT 1|    2|
|AVENUE OF THE AME...|    1|
|            1 AVENUE|    5|
|       CHURCH STREET|    1|
|         MAIN STREET|    3|
|    WEST   66 STREET|    1|
|    LEXINGTON AVENUE|   12|
|    EAST   69 STREET|    1|
|        WORTH STREET|    2|
```

Graphs:

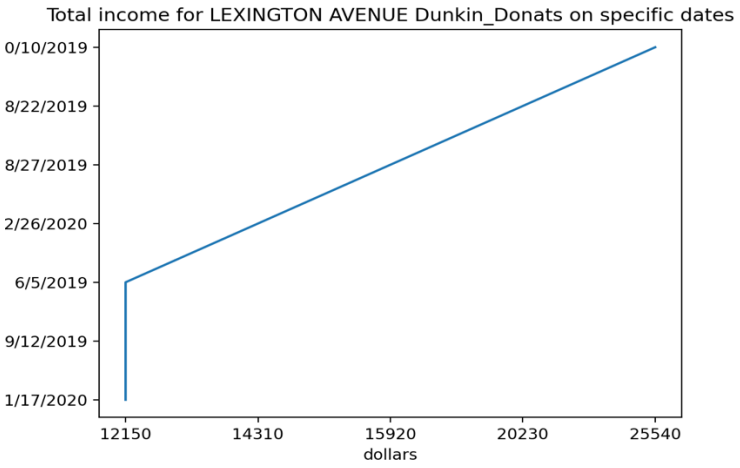
Total Income of Starbucks on Broadway and Lexington Avenue

12

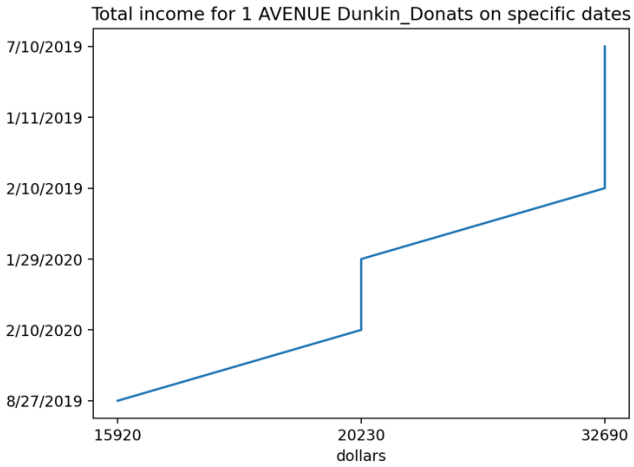




Total Income of Dunkin Donats on specified dates



6



Analysis

For Starbucks, we can easily find that the store in Broadway has lower income throughout the pandemic period than normal. Broadway Starbucks experienced a huge decrease in income due to the Covid impact. Moving on to the Starbucks at Lexington ave. Although during the covid period its income has decreased a little bit compared to its maximum point, Lexington Ave Starbucks are still higher than Average income level. I guess the reason is that uptown doesn't have that many workers than the midtown, same with the visitors during the pandemic.

For the Dunkin_Donats, both lexington ave and 1 ave stores were facing a huge decrease in their income. That is pretty usual to me since during the pandemic,

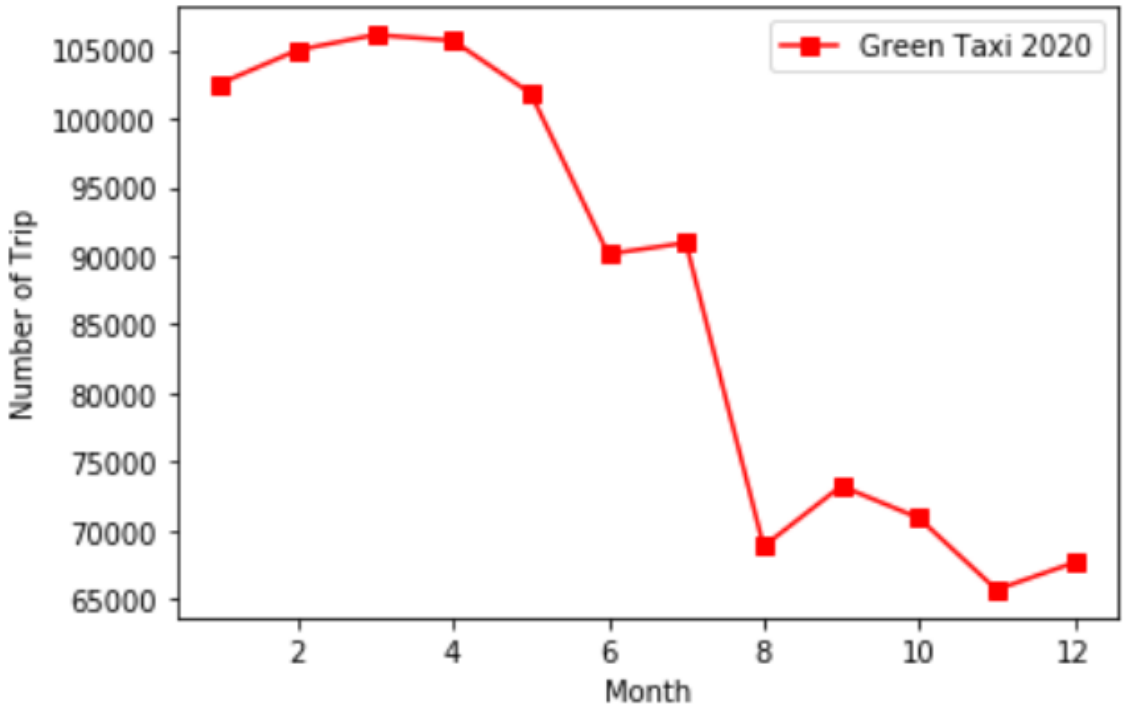
everyone is worried about the food they eat every day. So they probably buy food from outside restaurants as least as possible. In this way, they can reduce their chance to get infected by the covid.

Both StarBucks and Dunkin Donuts are facing a hard time during the pandemic, the income for each store has dropped significantly.

Challenge

When I tried to sort the data based on the dates using regular expressions, it didn't go the way I expected. I guess the date format in the dataset is so disordered so it makes sorting dates even harder for me. In the end, I just sort the value and draw all the data on a graph.

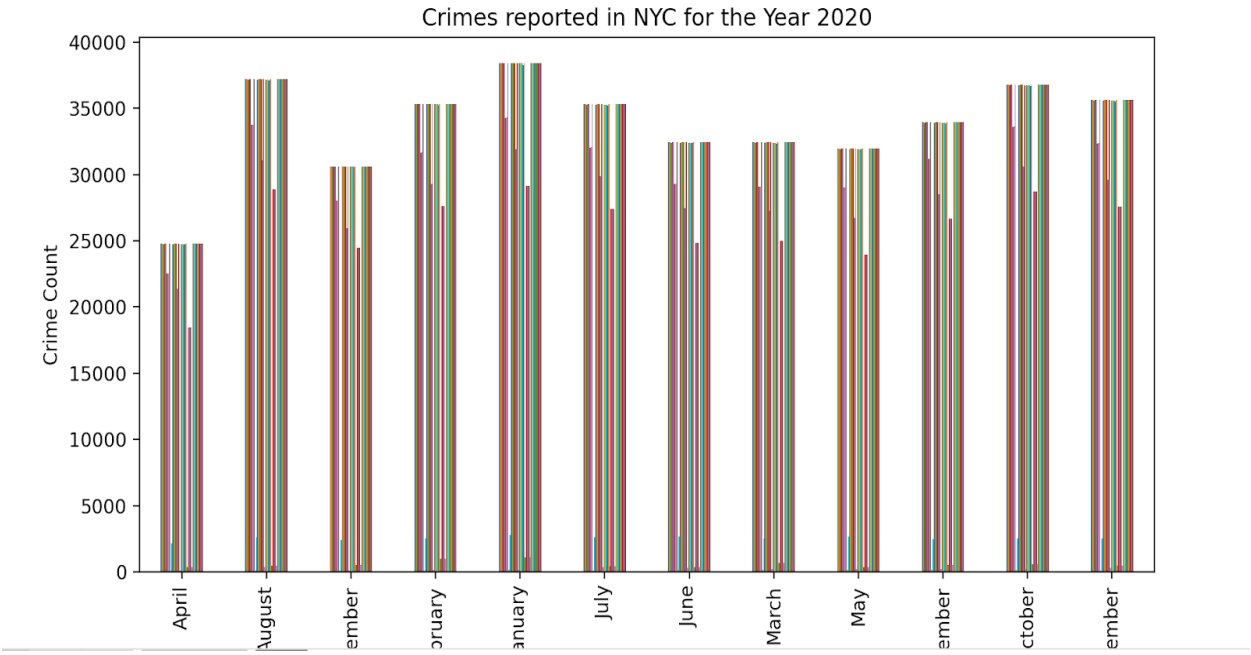
2. **Taxi Dataset(Zijie):**



-From this graph, we can see that before the outbreak. There are more than 100000 trips per month. The trip decreases a lot from May to June. It increases a little bit in July but then decreases a lot. At the lowest point, there are only half trips compared with the amount before the outbreak.

3. Crime Dataset(Avinash):

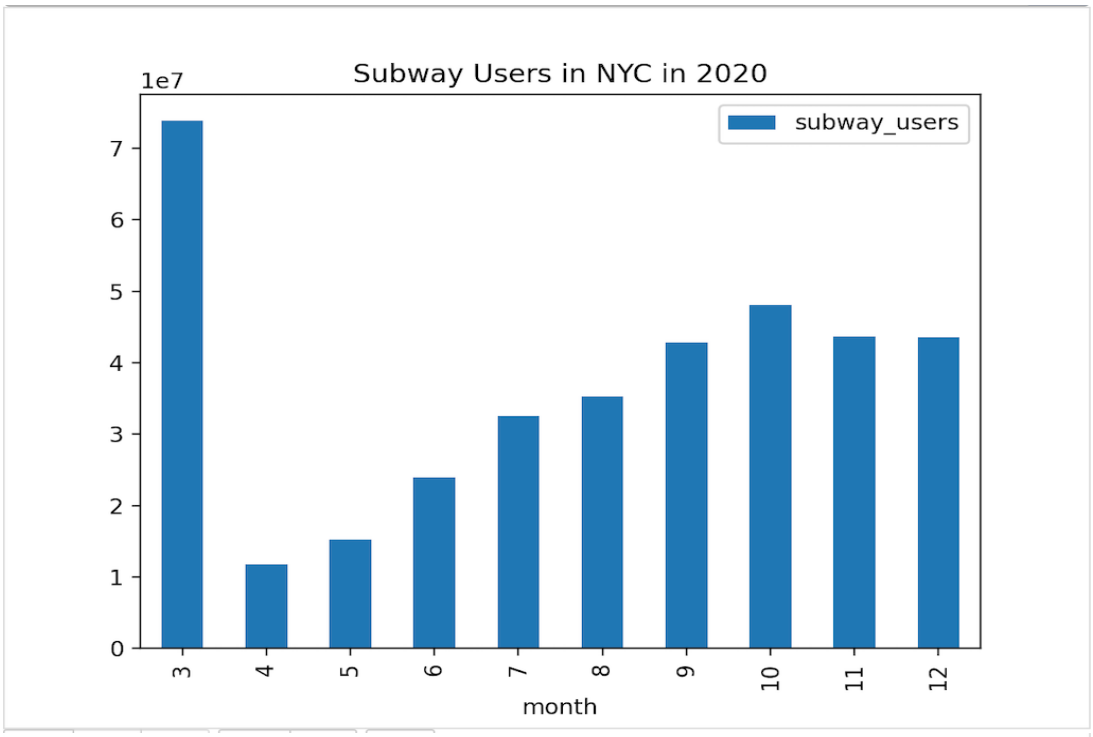
The crime dataset was sorted per the date and grouped to show the total count of the crimes reported per month. This was done using Spark coupled with Jupyter. The dataset was queried to output the count of the crimes. The Ipython file has been uploaded to Github.
A bar graph was plotted reflecting the monthly crime reports. The graph is as shown below:

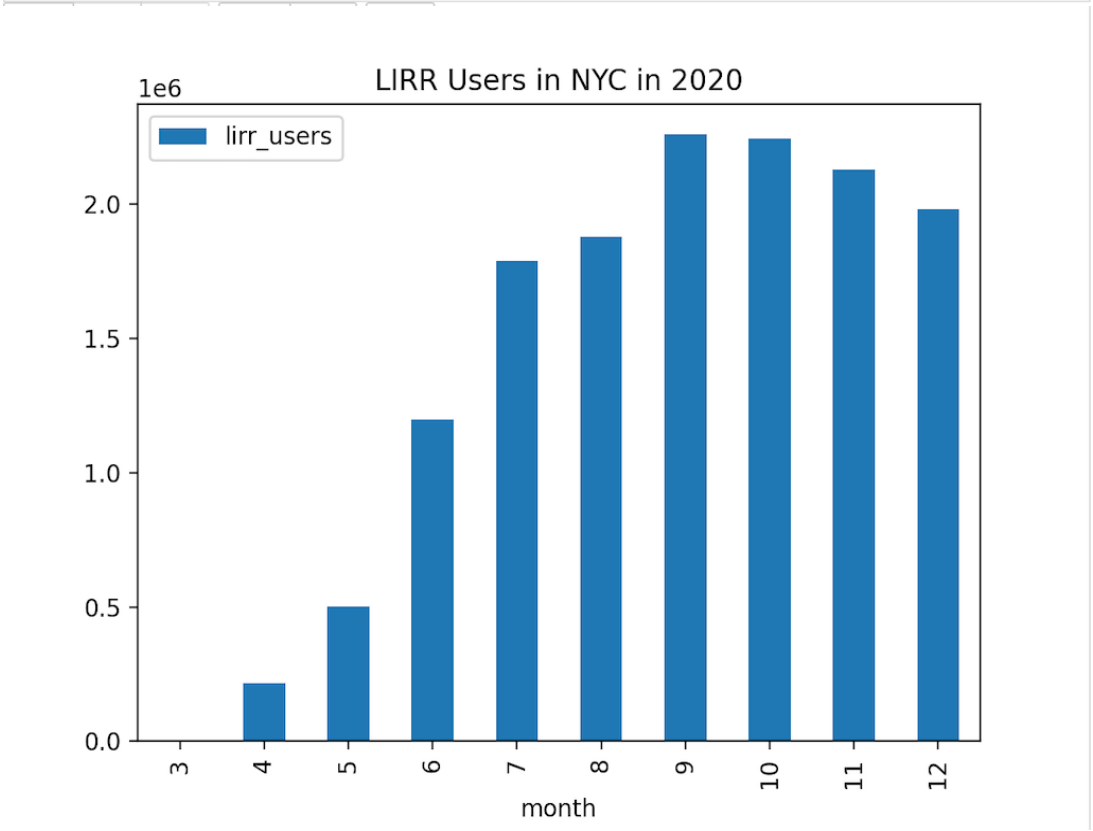


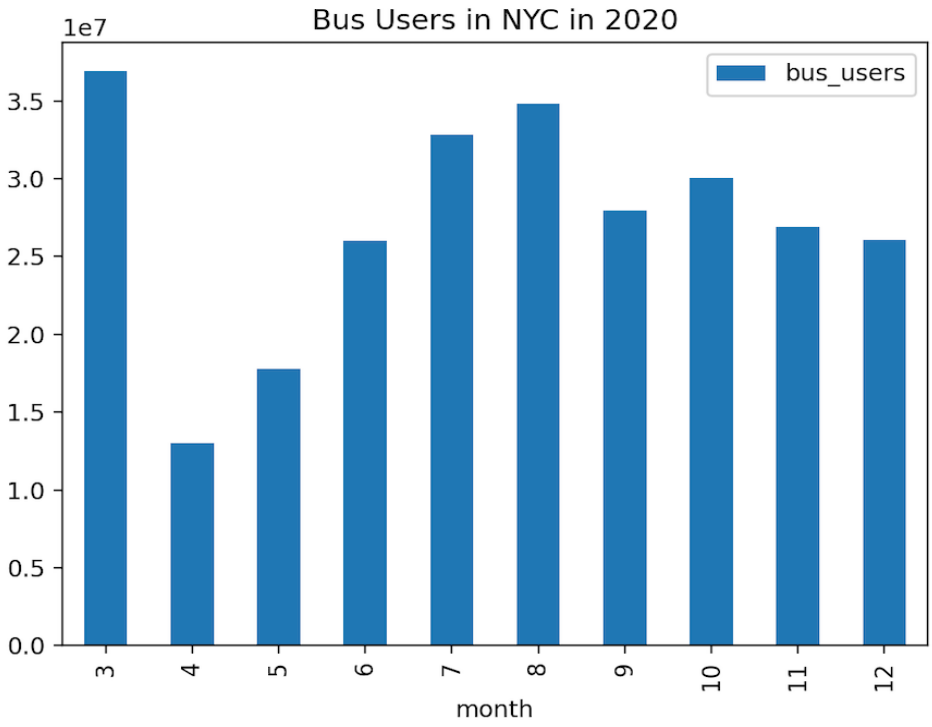
Crime reports significantly decreased when lockdown was announced in April in NYC. Once the restrictions were eased, the graph shows it returning to the normal average.

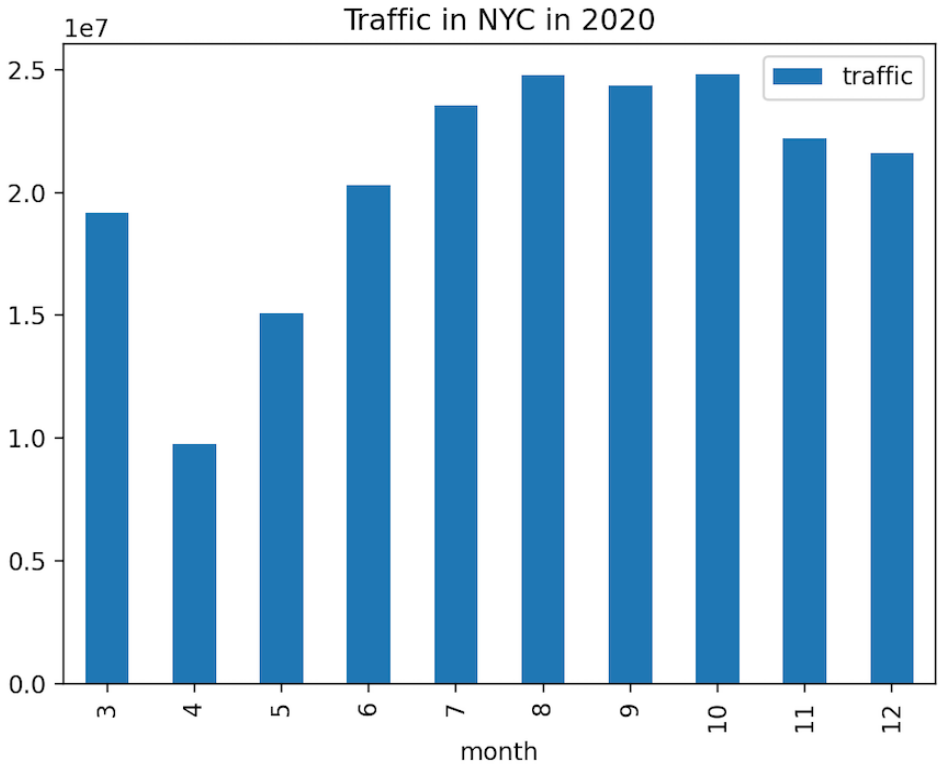
4. MTA_BUS Dataset(Avinash):

This dataset was properly cleaned to include only the necessary columns required for visualization. The below graph shows the number of subway, LIRR, Bus users and traffic of NYC.


Subway Users in NYC in 2020
1e7
subway_users
0
1
2
3
4
5
6
7
3
4
5
6
7
8
9
10
11
12
month


LIRR Users in NYC in 2020
1e6
lirr_users
0.0
0.5
1.0
1.5
2.0
3
4
5
6
7
8
9
10
11
12
month

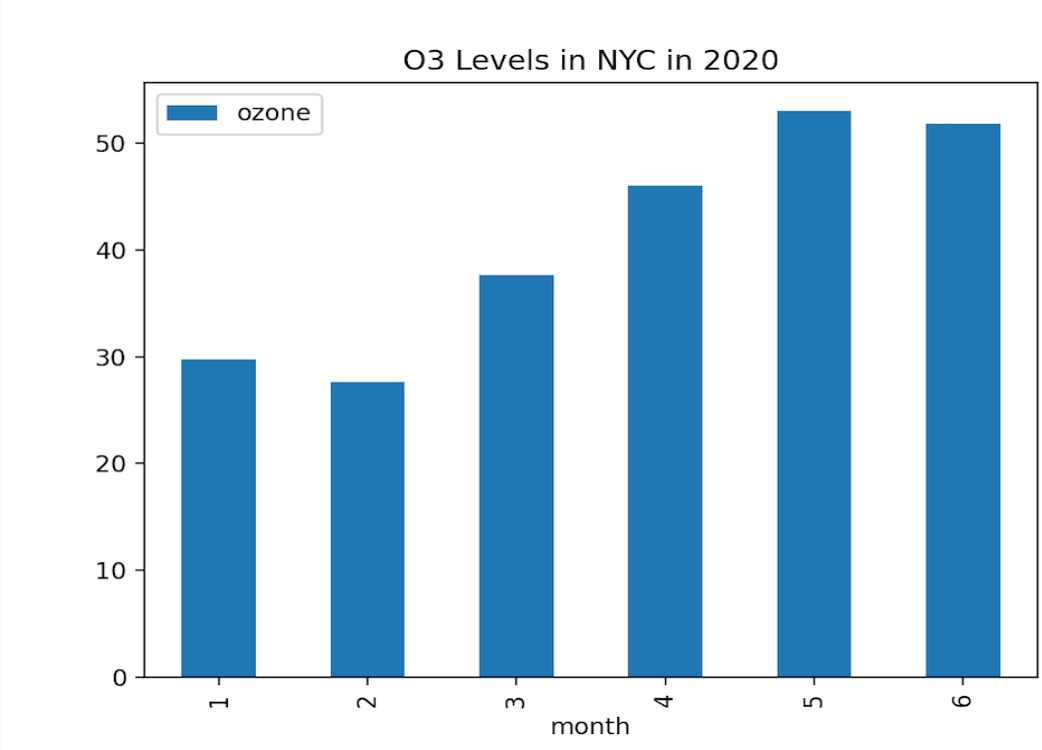

Bus Users in NYC in 2020
1e7
bus_users
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3
4
5
6
7
8
9
10
11
12
month

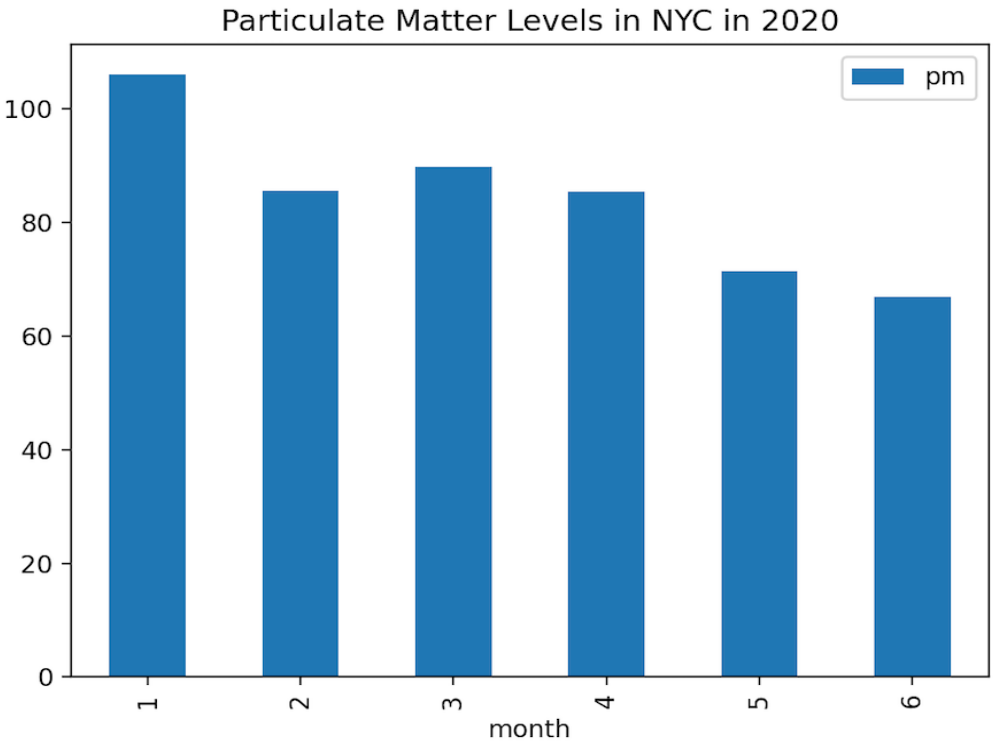

Traffic in NYC in 2020
1e7
traffic
2.5
2.0
1.5
1.0
0.5
0.0
3
4
5
6
7
8
9
10
11
12
month

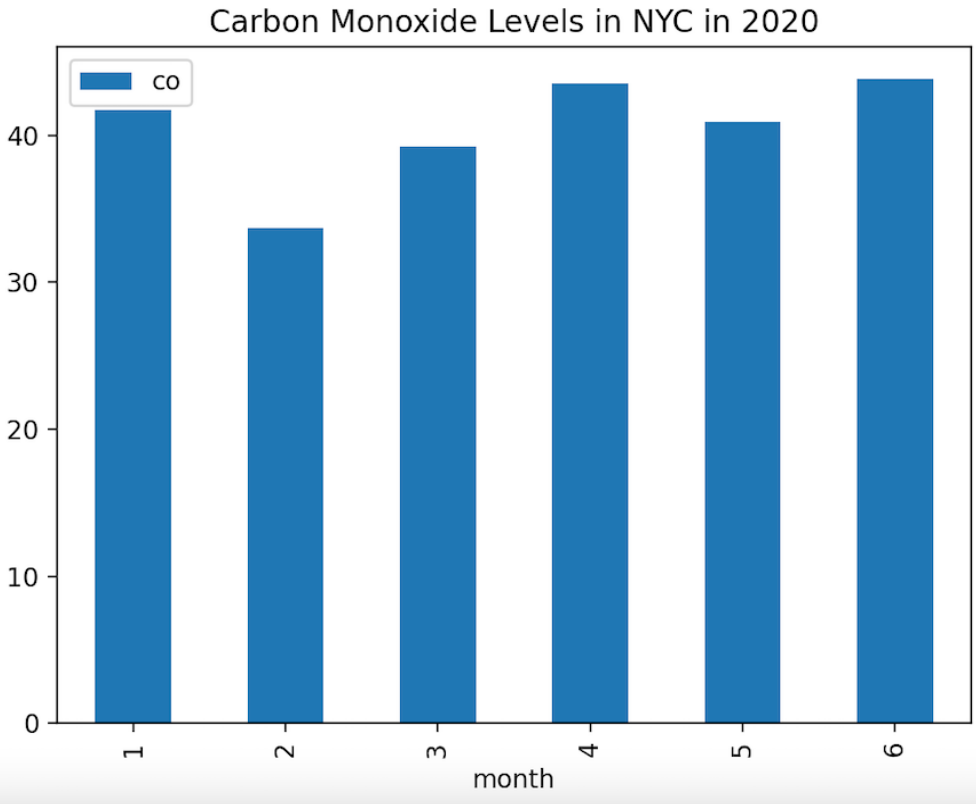
As indicated by the above graphs, there was significant decrease in the number of Subway, LIRR and Bus users in NYC during the month of April when lockdown was announced. Traffic too reduced. Slowly, after the peak Covid time was over, users started picking up.

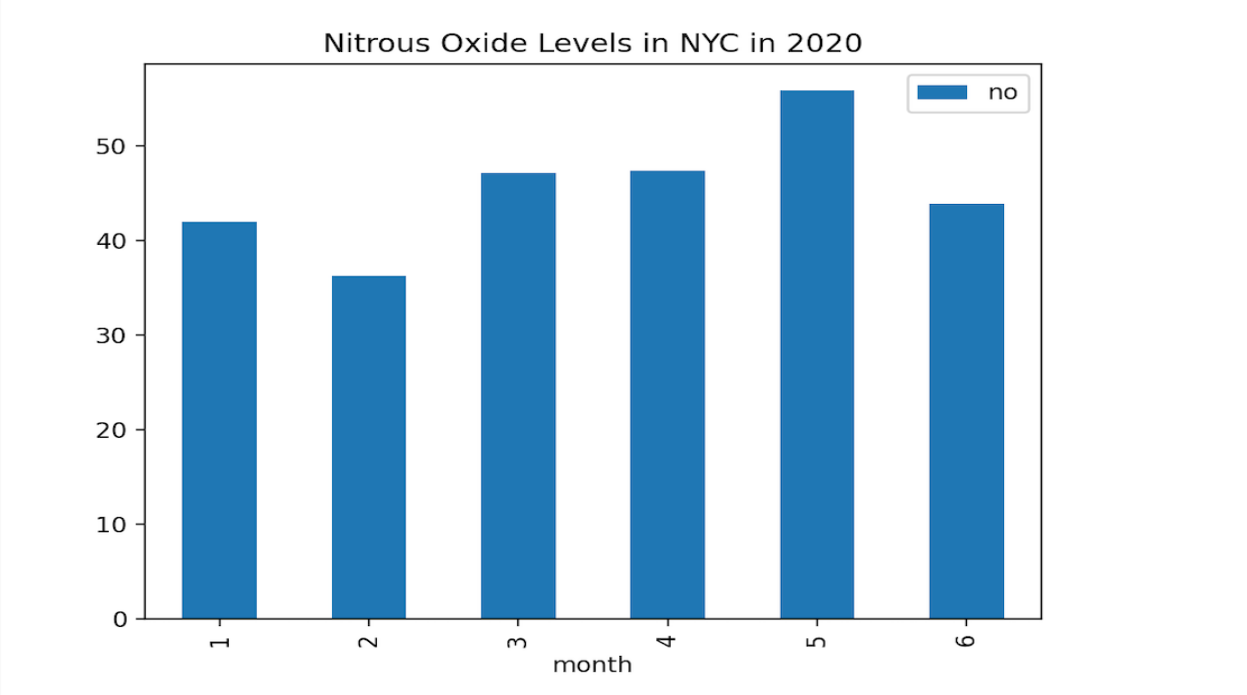
5. **Air_Quality Dataset(Avinash):**

Air Quality of NYC improved during the lockdown phase. This dataset included the Air Indexes reported in NYC on a daily basis. These are visualized by the below graphs.




Particulate Matter Levels in NYC in 2020
pm
100
80
60
40
20
0
1
2
3
4
5
6
month


Carbon Monoxide Levels in NYC in 2020
co
40
30
20
10
0
1
2
3
4
5
6
month



Nitrous Oxide and Carbon Monoxide levels were almost the same whereas the ozone and particulate matter levels decreased noticeably.

Challenges

I referred to various data sets to make sure that the ones I picked for the project were accurate. Filtering out rows with null values were a challenge.
Clustering and aggregating the dataset based on the boroughs was a challenge.
OpenRefine was a new tool and difficult to use in particular.

Conclusion

Based on the information above, the COVID did impact the economy for the restaurants in New York. Pandemic destories many restaurants' income compared to the normal period. But it is lucky to see that nowadays more and more restaurants are surviving.
Also, the amount of taxi trips decreases a lot due to COVID. At the lowest point, there are only half trips compared with the amount before the outbreak.

There was a significant decrease in the crimes reported during the month of April when NYC faced extreme difficulties with Covid-19. The crime rates returned to the normal average once the restrictions were eased in the city of New York.
There was a significant decrease in the number of users of public transportation and traffic during the month of April when NYC peaked with Covid-19. These rates are returning to the normal average after the restrictions were eased in the city of New York.
The air quality of NYC showed improvements in the levels of Ozone and Particulate matter. Nitrous Oxide and Carbon Monoxide levels were somewhat the same but improved slightly.

- Github link: GitHub - z632101094/coronavirus-data
- Github link 2: https://github.com/aa8382/coronavirus-data