The Effects of COVID-19 on NYC Transportation

Authors: Ignacio Calvera Maldonado, Jonathan Wiener, and Xuelei Guo

Github Repository: https://github.com/wienerjon/covid-big-data-project

Introduction

2020 has been marked by the emergence and soon after outbreak of a coronavirus called SARS-CoV-2, more commonly known as COVID-19, that spread across the world, became the first global pandemic in over a century, and ended with the lives of 1.57 million people and infecting a total of 69.1 million, as of the date of writing. It has been the cause of dislocations of magnitudes never seen before in many industries, from deeply straining and even overflowing the health services industries of many countries for an extended period of time, to bringing the global air travel industry to an almost-complete halt.

To help curb its spread, health guidelines and restrictions have been placed on a local, national, and international scale, usually involving some form of restriction on travel and mobility, use of personal protective equipment (PPE), and limits on congregation sizes. As a result, the transportation industry has been severely affected. In NYC, the MTA is suffering its worst financial crisis in history and the taxi workforce decreased up to 75% from pre-pandemic levels, amongst others.

In this paper, we will further discuss and analyze the effects of the COVID-19 pandemic on NYC's transportation industry through a wide range of city indicators, as well as the extent to which future case counts can be predicted through changes in transportation behaviours. The indicators we will be working with are CitiBike traffic and travels, for-hire taxis, MTA Subway turnstiles, pedestrian traffic across the Brooklyn Bridge, vehicle traffic across MTA bridges and tunnels, and national TSA air passenger travel data, as they cover the wide spectrum of transportation options in NYC and together can provide robust estimates of overall transportation volumes and distribution throughout NYC.

Problem formulation

Our first hypothesis is the following: It is expected to find that fewer people are traveling this past year than pre-COVID.

By observing the transportation data from the years 2019 to 2020, it is theorized that a drastic decrease in travel will take place in 2020 when compared to 2019.
This is what is being expected since during the COVID-19 pandemic, many of the residents in the greater New York City area were quarantined once coronavirus cases started to be reported.

Our second hypothesis is the following: It is expected to find a correlation between transportation and COVID-19 cases, when accounting for the necessary virus incubation periods specified by the CDC.

If it proves to be true, and predictions can be made with a high level of accuracy, these findings could be used to help health services employees better prepare for incoming surges in cases. Mortality rates were at their worst when hospitals and other health centers suddenly found themselves without enough capacity to handle the surges in infections. Therefore, a highly accurate predictor that could anticipate these surges some days in advance and provide an estimate of cases for said days could prove to be a great resource in making sure hospitals are aware of the amount of cases that will be coming in soon, and so anticipate and increase medical supplies if needed.

Related work

The topic chosen for this project was partly inspired by a New York Times article (2020) that explores how New York City has changed during the COVID-19 pandemic by observing the sound levels in different parts of the city. This article led us to explore the following articles which peak into how transportation in New York City has changed in the year 2020.

In the article, *How Coronavirus Has Changed New York City's Transit Landscape*, by the Wall Street Journal, the authors analyze how the COVID-19 pandemic affected the amount of transportation that takes place in New York City (Melgar, Rigdon, Rust, 2020). In the article, they explore three different forms of transportation metrics: the amount of traffic on roads, the amount of subway entries via Metrocard, and flight traffic.

By comparing the average speed of cars in different parts of Manhattan from before the pandemic and during, they observed that cars traveled faster once COVID-19 cases started to be recorded. They concluded from this that due to less congestion and vehicles on the road, the cars that were still traveling did not need to sit still in typical New York City traffic (Melgar et al., 2020).

They also observed that after schools and restaurants were closed down, there was a sharp decrease in ridership on the subways. While the typical amount of rides on a given day can vary between four to five million, the amount of riders recorded by the MTA on March 19th was 1.5 million (Melgar et al., 2020).

The last transportation metric observed was flight data. The Wall Street Journal reports that March 23, 2020 had about half the amount of flights fly into and out of the airports surrounding New York City compared to the previous year (Melgar et al., 2020).

Another work related to our topic was written by Caroline Spicack (2020) for Curbed New York. In addition to also exploring the differences in subway ridership, she also looks into buses and Citibikes. She reported that due to the pandemic, the amount of people riding buses is down 80 percent when compared to ridership in 2019.

Of the transit options provided in New York City, Citibikes seem to be the only one, besides walking, that is encouraged. Spicack writes that the company has provided “first responders, healthcare, and transit workers with a free month of Citi Bike” (2020). The bike sharing company has also installed new bike racks throughout the City to make biking available to more people.

Methods, architecture and design

A series of transformations were necessary to turn the large amounts of raw data into meaningful datasets from which we could perform trend analysis. While the specifics will be discussed below for each of the transportation indicators chosen, the general method was as follows: first, the dataset in question was downloaded and a preliminary cleaning effort was made. Then, it was uploaded to NYU’s High Performance Computing (hpc) clusters, where it was further cleaned if necessary, transformed and analysed through Spark processes, and then downloaded back to our local systems in a standard format. Once there, it was further analysed: first on its own on a Jupyter Notebook, and then grouped with the rest of datasets from which the final trend analysis was performed - all findings are detailed below in the Statistics and Results sections.

The Citibike trip data was downloaded from Citibike public data repository, and contained information about every single trip performed on a Citibike in NYC. We data was partitioned into monthly files, and all files pertaining to trips from 2019 and 2020 were downloaded for the purposes of this paper. At the time of download, there was only data for months up until and including November 2020.

The monthly files were uploaded to OpenRefine to perform preliminary distribution analysis and cleaning, but they were too big even on their own, so the original files were instead directly uploaded to the hpc. Once there, a python script using the pyspark module was run, which read said data files. It first joined all files together, and formatted the columns into a standard format agreed upon beforehand and shared between all the processed datasets. The Spark process then grouped the trips by day, and calculated both the total amount of trips per day and the total amount of hours traveled per day. This information was then written back to two .csv files (one for the amount of trips per day, and the other for the sum of hours traveled per day), which were downloaded from the hpc cluster back into our local system. This same python script then accessed and

processed the raw data from the MTA bridges and tunnels data, and the national TSA data, according to the design detailed below.

The MTA bridges and tunnels data was downloaded from Open Data NY. It shows the daily amount of traffic in each of the bridges and tunnels operated by the MTA. If a bridge or tunnel has bi-directional traffic, it is shown as two separate records on the data file. The last date for which there was data at the time of download was the 28th of November, 2020.

This file was accepted by OpenRefine's software, and so it was checked for duplicates, outliers, empty values and erroneous classification of attributes. A group of potential outliers was found during the week of the 13th to the 20th of September, 2020, but after query searches in the Internet it was found that it coincided with the beginning of the academic calendar for the NYC public school system. This event had been delayed for several weeks due to the pandemic, and the sudden announcement caused a lot of families to move back to the city from other parts of the country in a very short amount of time. The graph can be seen in the following section. Then, it was uploaded to the hpc, where the python script previously mentioned read it and filtered by date to records from 2019 and 2020. Then, the daily amount of traffic was computed for all bridges and tunnels (and in both directions when the case arose), and outputted to a file, which as mentioned above was downloaded back to the local system.

The national TSA daily passenger data was scraped from the official TSA website, which contains a daily-updated count of the amount of air passengers in all US airports since the beginning of march, 2020, and the same variable for the same weekday of 2019. At the time of download, the last date for which there was data was the 3rd of December, 2020. As such, data is available from the 1st of March, 2019 to the 3rd of December 2019, and the same range for 2020.

Since the scraping was done directly from the TSA webpage, we did not see the need to analyse it through OpenRefine, and instead just uploaded the original data into the hpc. There, it was reshaped and concatenated according to the date of the value using the pyspark.sql module in the python script mentioned above, and written to file. Since this data already provided a grouping by day, there was no need to perform grouping or arithmetic actions.

The above processed datasets were then plotted in a variety of ways, using a Jupyter Notebook available in the Github repository that accompanies this report. The results, as well as those for the datasets below, are shown in the next section.

The pedestrian count data was downloaded from Open Data NYC. It shows the hourly amount of pedestrians on the Brooklyn Bridge. It contains the amount of pedestrians

towards both Manhattan and Brooklyn, as well as a total amount. The data is from October 1st 2017 to June 30th 2020. However, records from August 2018 to December 2018 were missing for some unknown reasons.

This file was uploaded to OpenRefine for cleaning, but it was not accepted since it was too big. The original data was directly uploaded to the hpc instead. Before a Python script using the SparkSQL was run, the data was cleaned to eliminate duplicated and/or invalid rows by Microsoft Excel, and unrelated attributes such as temperature, weather, events, etc. were also removed from the data. The only columns kept were the time and the total count of pedestrians. The python script first partitioned the row of time in order to only take the date without the hour. Then it grouped up the data by day and sorted it from old to new. A script to group up the data by month was also run to output the count of pedestrians. The output file was then downloaded back to the local system and was analysed in a Jupyter Notebook.

The New York City For-Hire Taxi data was taken from the NYC.gov website. The data is categorized in clusters by month and the datasets show the datetime and location of each ride. Since the 2020 taxi dataset was only recorded until June 2020, the files for January to June were downloaded for the year 2019.

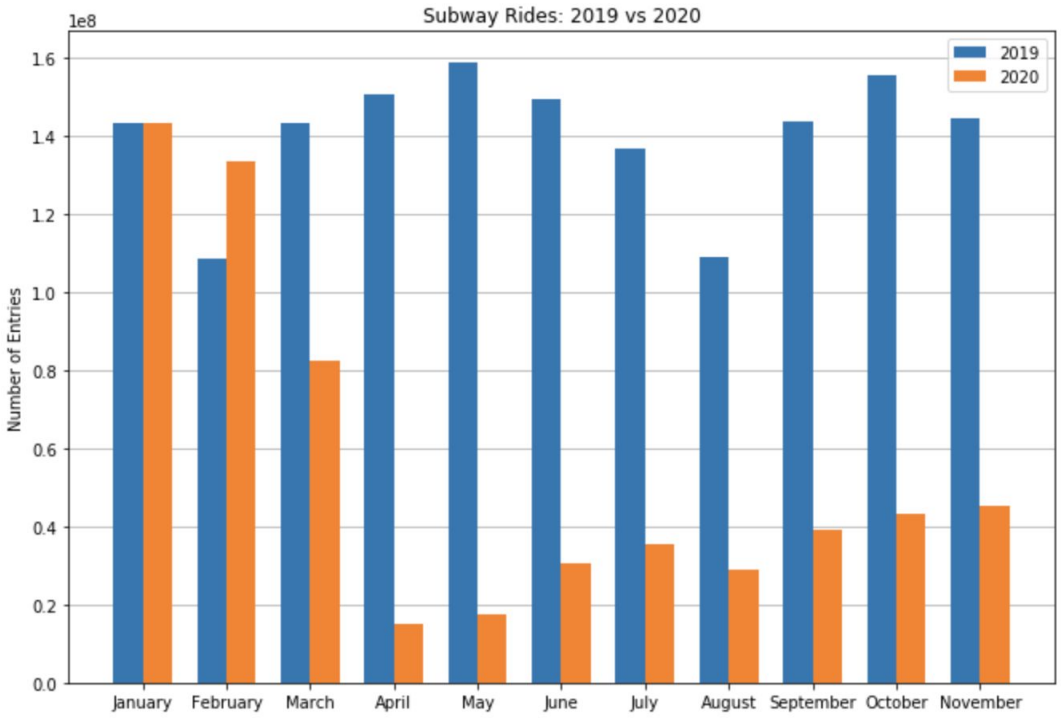
Since the datasets were too large to upload to OpenRefine, they were directly uploaded to the HPC. Once they were there, a SparkSQL script was executed in order to group together and count the number of taxi rides for each day. The output for each of the monthly rides were then merged and sorted by date. The output file was then downloaded back to the local machine where a Python script, running in Jupyter Notebook, accessed the file and performed analysis comparing the two years.

Datasets with information on the subway turnstile usage from 2019 until 2020 were taken from the Data.NY.gov website. The dataset held the information of how many entries and exits a specific subway turnstile had within a certain time frame.

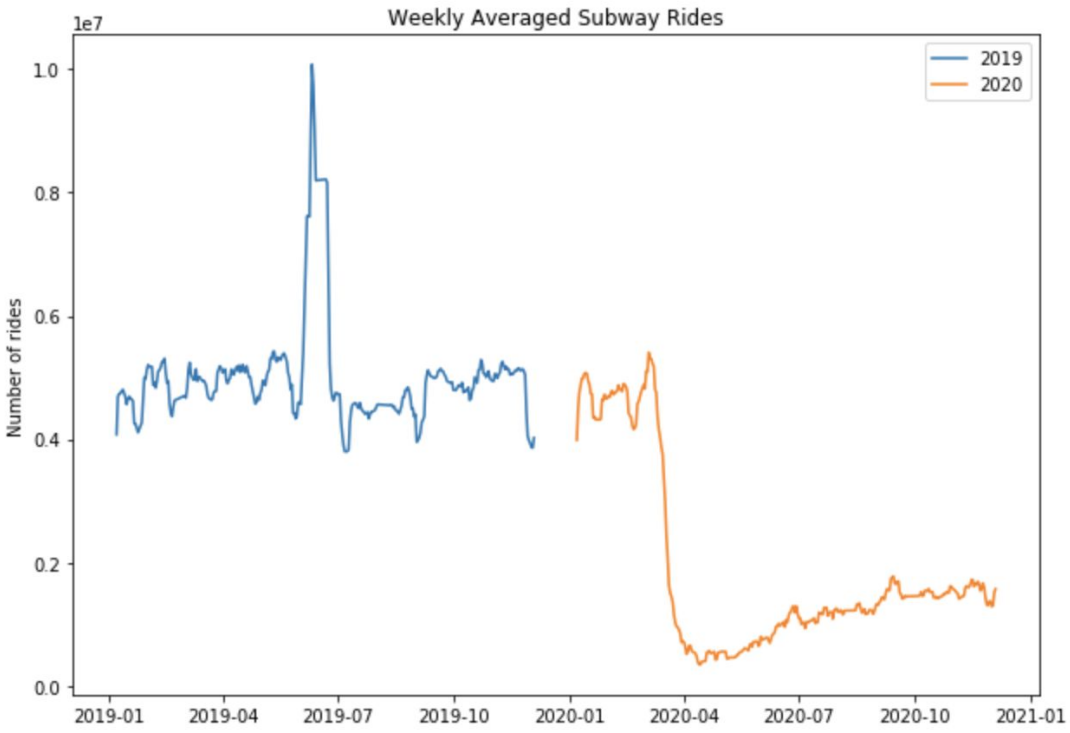
The subway turnstile dataset was too large to upload to OpenRefine so it was uploaded to the HPC. A few steps needed to be taken before the daily subway rides could be determined. The first query that needed to be performed was used to find the amount of entries a single turnstile had in a day. Since the entries for each day did not start at zero, the difference between the earliest collection of entries and the last collection in a day was found. Next, the dataset was grouped by day so the total number of entries for all turnstiles could be found. The output file was then downloaded back to the local machine where a Python script, running in Jupyter Notebook, accessed the file and performed analysis comparing the two years.

Statistical Analysis and Results for First Hypothesis

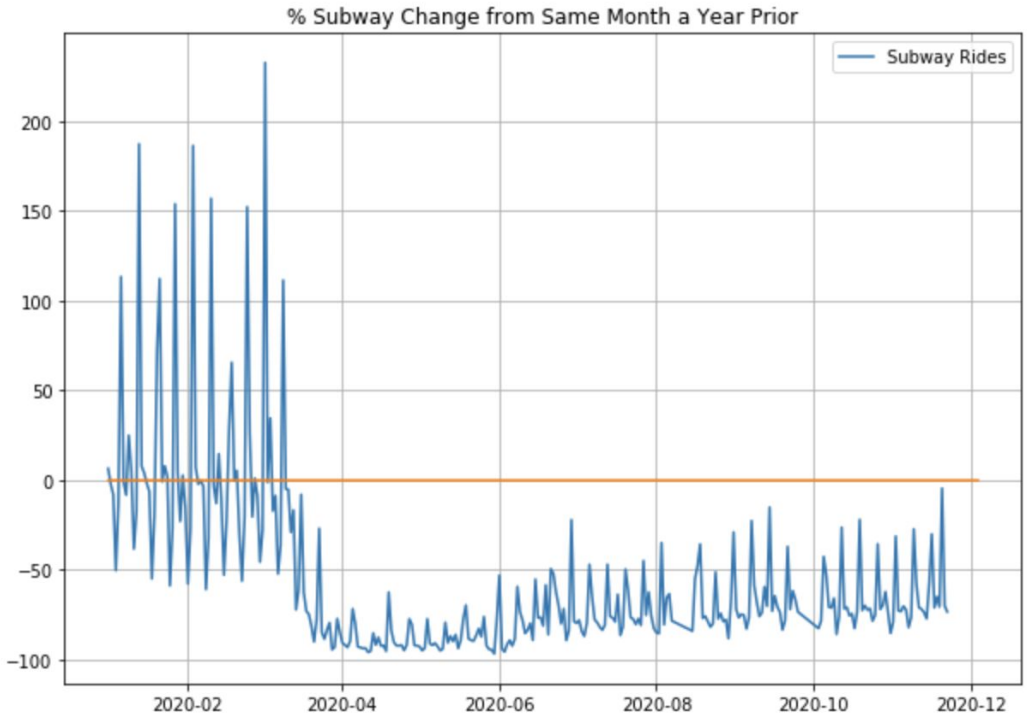
A series of graphs here show that fewer people are traveling compared to the pre-COVID time through subway rides, for-hire vehicle rides, Citibike trips, traffic volume through MTA bridges and tunnels, and TSA air passengers, thus overall confirming our first hypothesis.



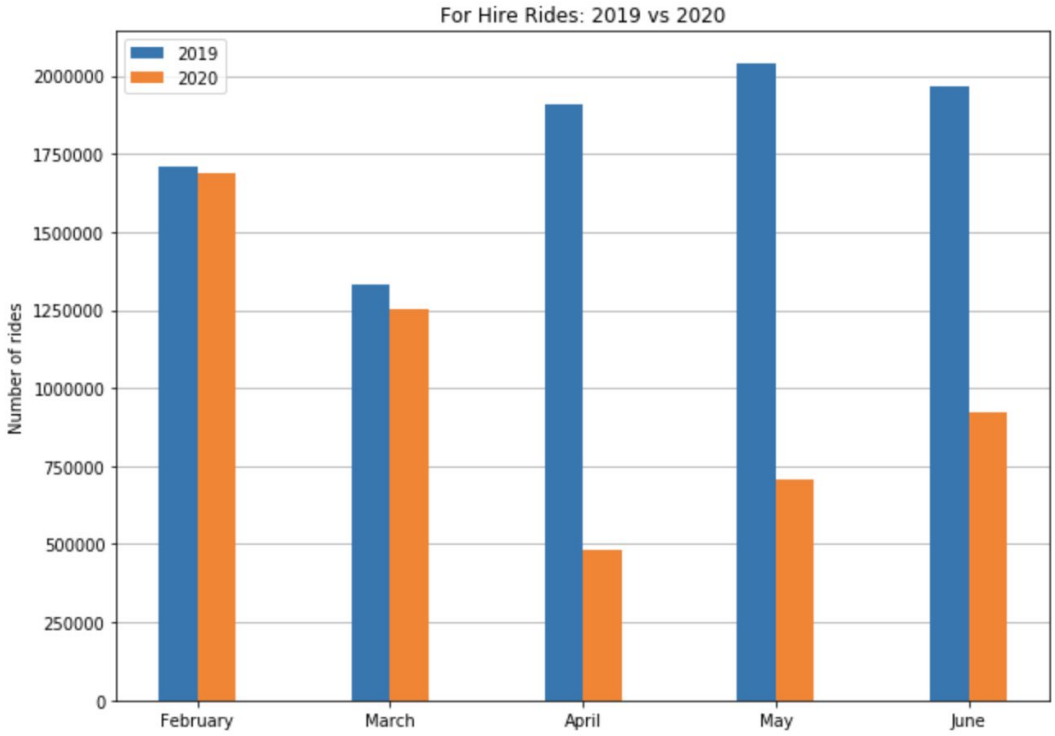
The amount of subway rides in 2019 vs 2020 compared by month: the x-axis is the months in a year, and the y-axis is the number of entries. In January and February, the number of entries of 2020 is about the same level as the one of 2019. However, starting from Marth, the number of entries has been slumping and reaches the lowest point in May. Though there was a small increase in the following months, the amount of 2020 is still much lower than the one in 2019, which means fewer people are traveling in 2020.



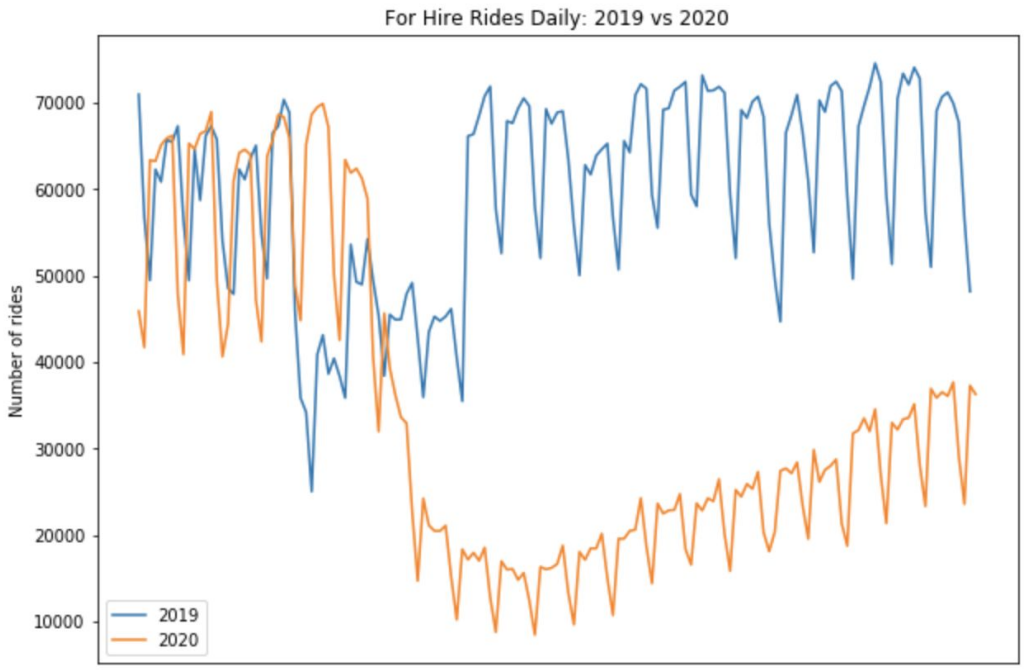
The weekly averaged subway rides in 2019 vs 2020 compared by month: the x-axis is the months in a year, and the y-axis is the number of entries. The overall trend of 2020 is going down (a slump in around March) and at a low level compared to the one of 2019, which means much fewer people are travelling in 2020.



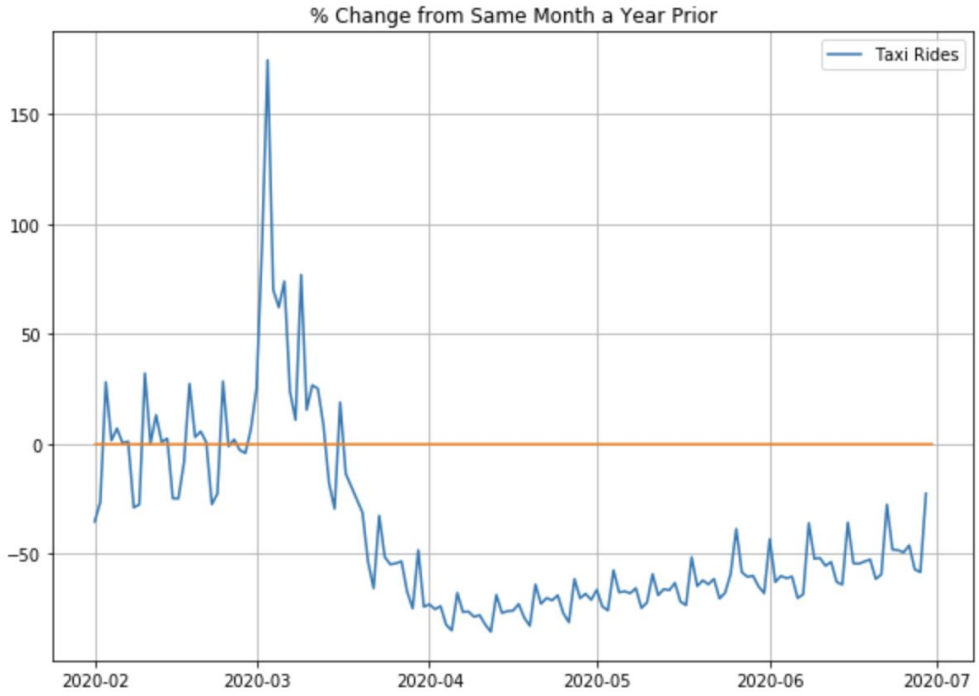
The percentage subway change from the same month a year prior: the x-axis is the months in a year, and the y-axis is the percentage change. Before around March, the percentage is fluctuating around 0-line. After that, the percentage change is always below the 0-line, which means in those months the subway rides of 2020 are fewer than the one of 2019.



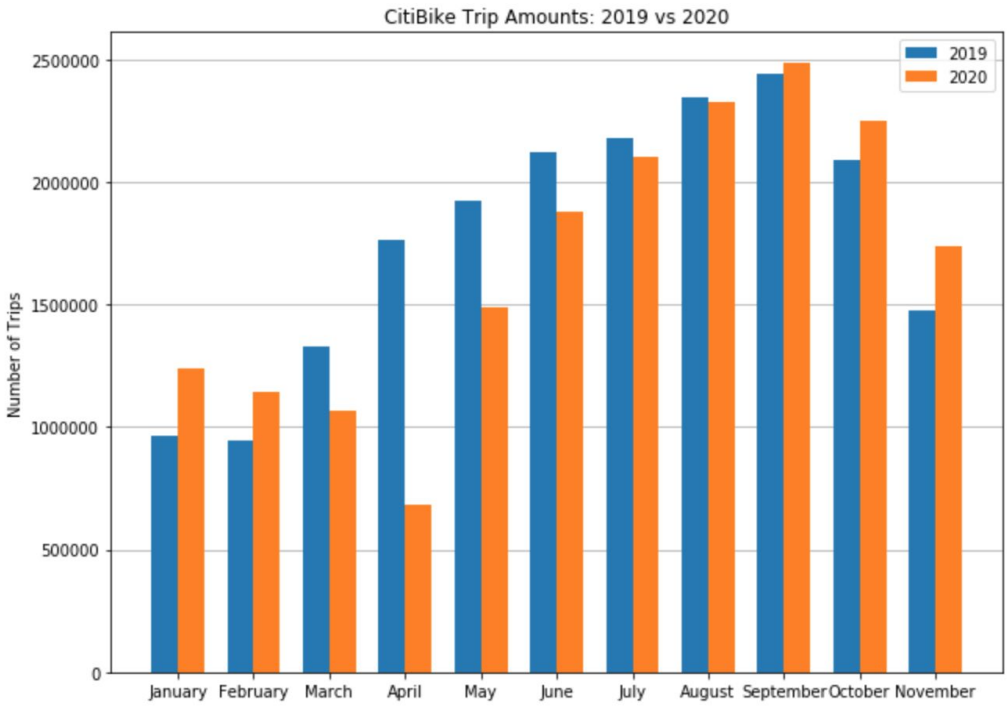
The amount of for hire rides in 2019 vs 2020 compared by month: the x-axis is the months from February to June, and the y-axis is the number of rides. In February and March, the number of rides of 2020 is about the same level as the one of 2019. However, starting from April, the number of rides has been slumping and reaches the lowest point. Though there was a small increase in the following months, the amount of 2020 is still much lower than the one of 2019, which means fewer people are traveling with for-hire rides in 2020.



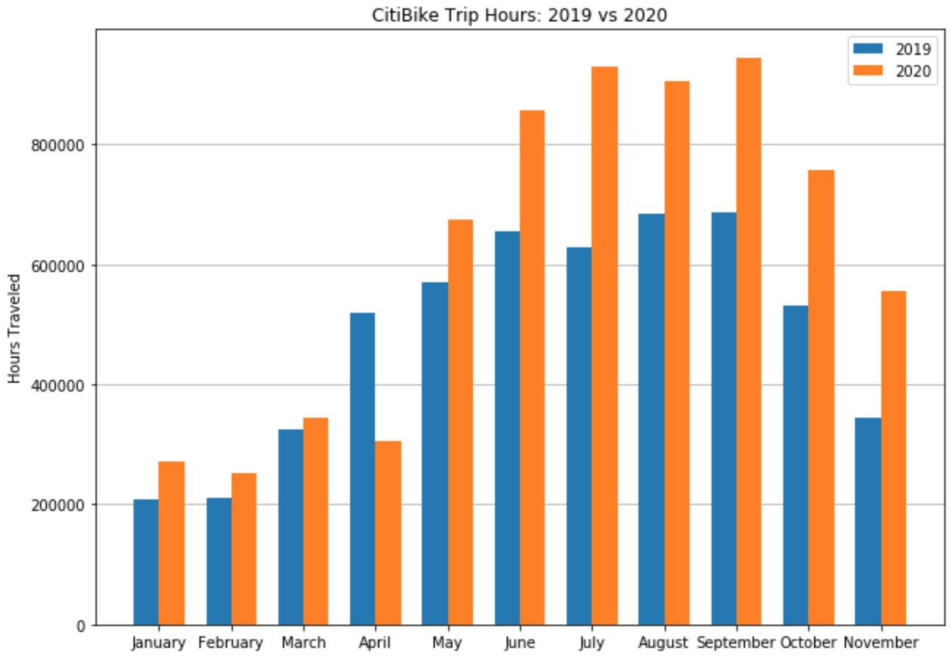
The for hire rides usage in 2019 vs 2020 compared by day: the x-axis is the timeline in a year, and the y-axis is the number of rides. The overall trend of 2020 is going down and a slump happened at around ⅓ of the year. Though the pattern and horizontal level of the 2020 line is similar to the 2019 line in the first third or a year, it is at a much lower level in the following months, which means much fewer people are travelling with for hire ride in 2020.



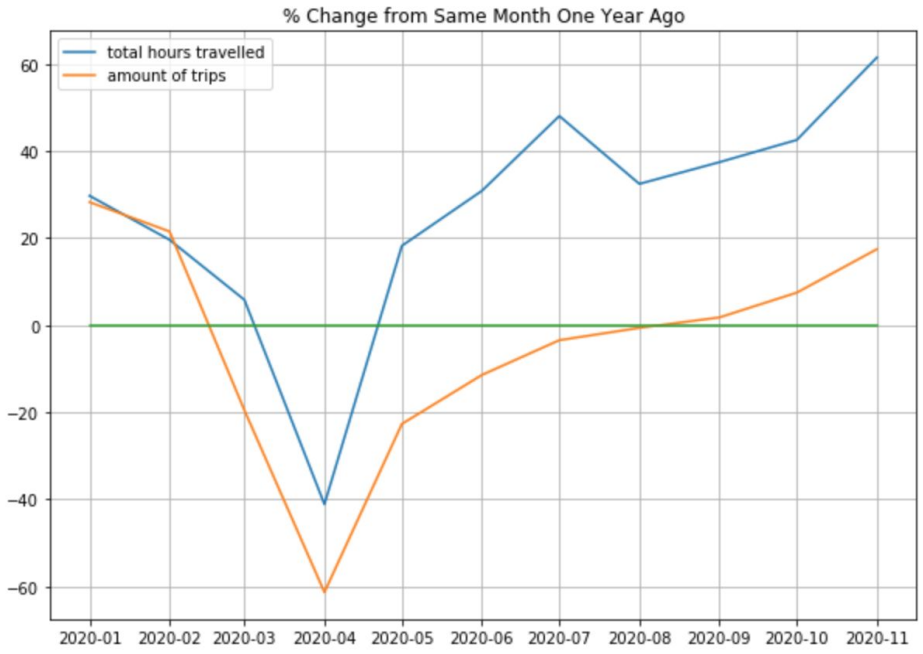
The percentage change of taxi rides from the same month a year prior: the x-axis is the months in a year, and the y-axis is the percentage change. Before early March, the percentage is fluctuating around 0-line and reaches the highest point. After that, the percentage change is always below the 0-line, which means in those months the taxi rides of 2020 are fewer than the one of 2019.



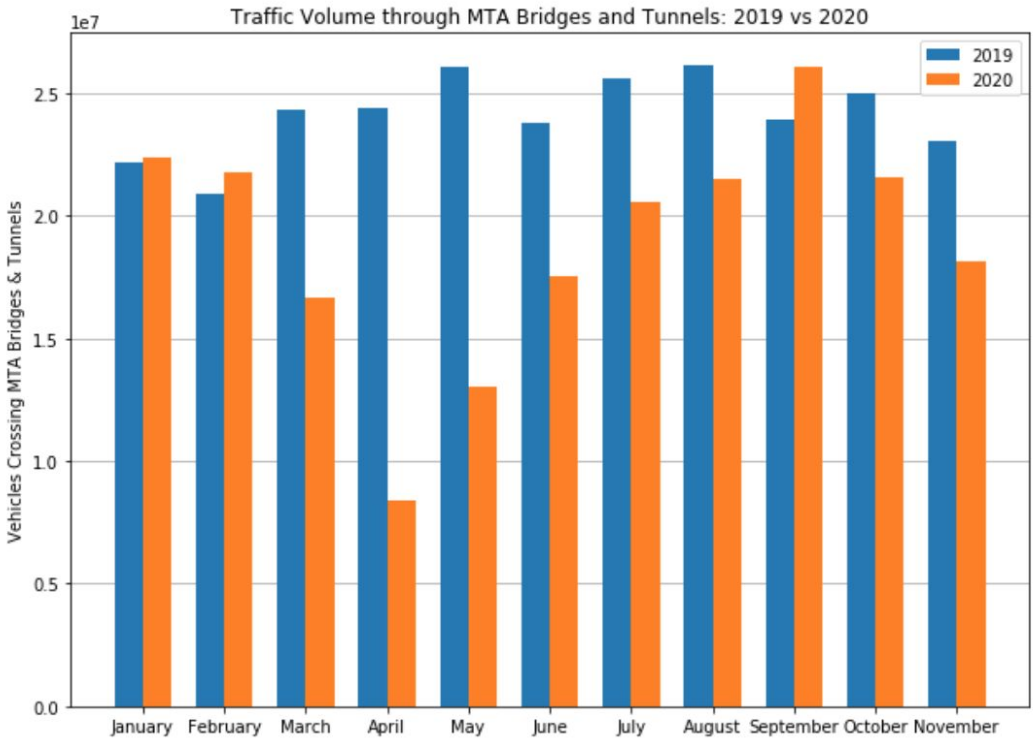
The amount of CitiBike trips in 2019 vs 2020 compared by month: the x-axis is the months in a year, and the y-axis is the number of trips. The number of trips of 2020 is around the same level as the one of 2019. Though it is against our hypothesis (fewer people traveling in 2020), it supports previous findings from related works that walking and biking were encouraged.



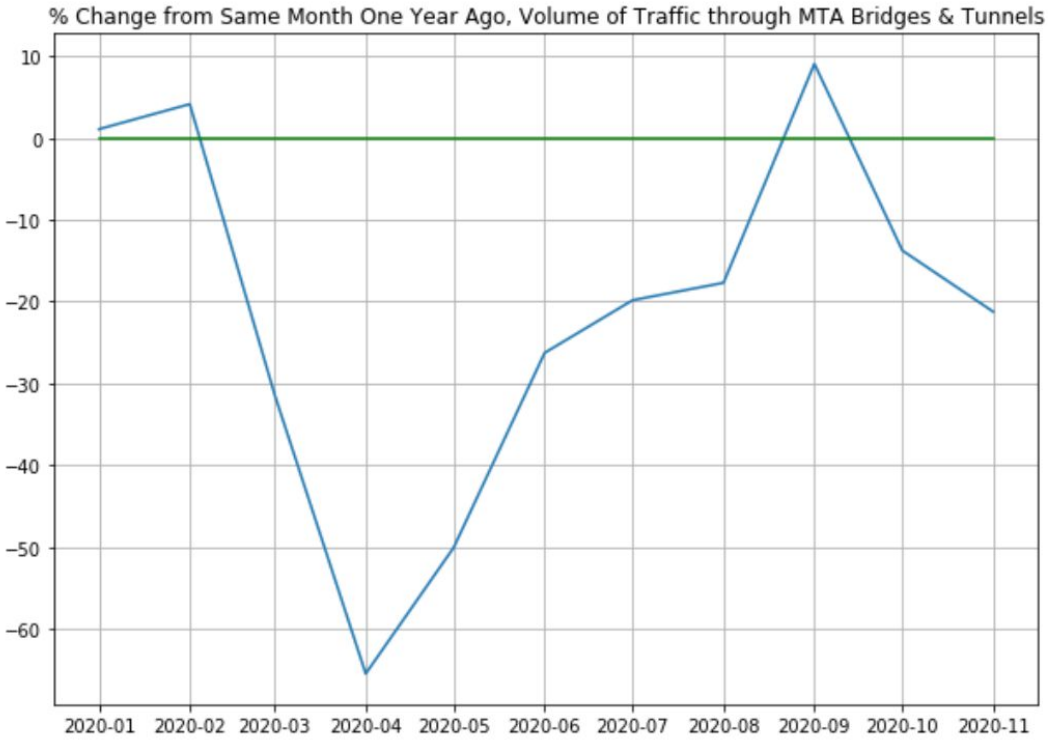
The using hours of CitiBike trips in 2019 vs 2020 compared by month: the x-axis is the months in a year, and the y-axis is the hours travelled. Similar to the number of trips of 2020, the hours traveled is around the same level as the one of 2019 and even higher in several months. Though it is against our hypothesis (fewer people traveling in 2020), it supports previous findings from related works that walking and biking were encouraged.



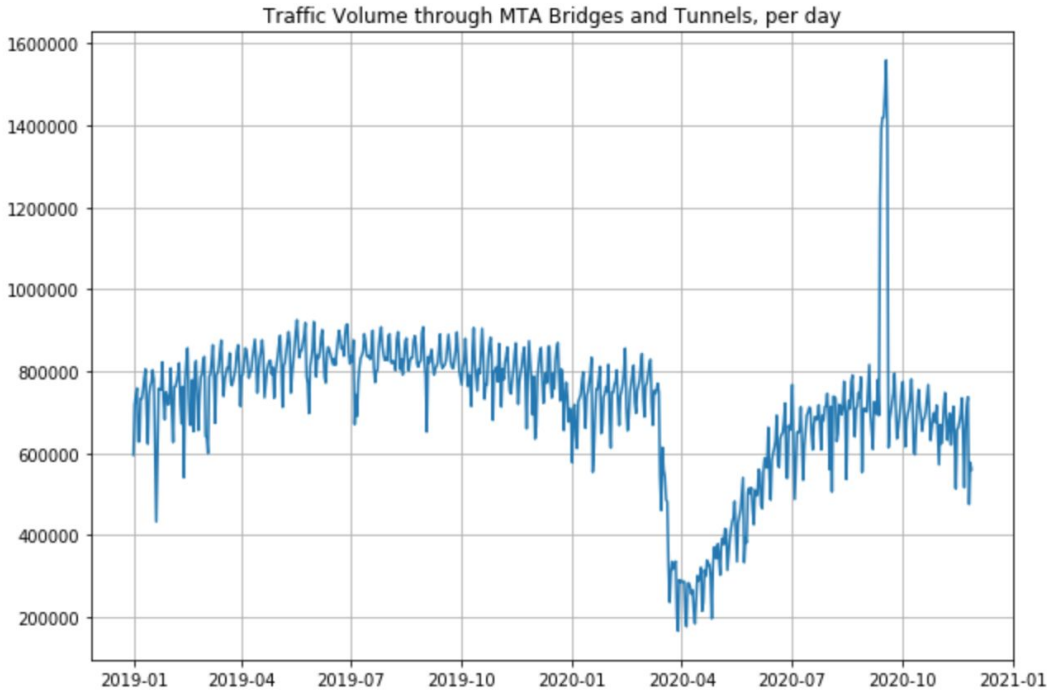
The percentage change of CitiBike total hours traveled and amount of trips from the same month a year prior: the x-axis is the months in a year, and the y-axis is the percentage change. Only in some months, March, April, and May, the percentage change is below 0-line. The percentage change in other months is always above the 0-line, which means the Citibike is being used more.



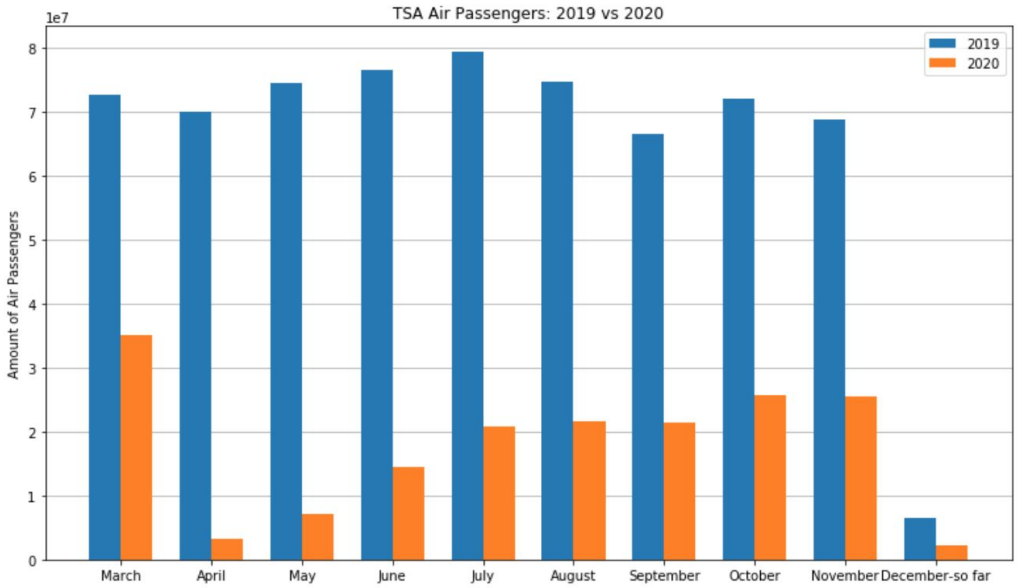
The traffic volume through MTA bridges and tunnels in 2019 vs 2020 compared by month: the x-axis is the months in a year, and the y-axis is the number of vehicles crossing MTA bridges and tunnels. In most months, the number of vehicles of 2020 is about the same level as the one of 2019. However, in March, April, and May, the number of 2020 is significantly lower than the one of 2019, which means fewer people are traveling in 2020 to some extent.



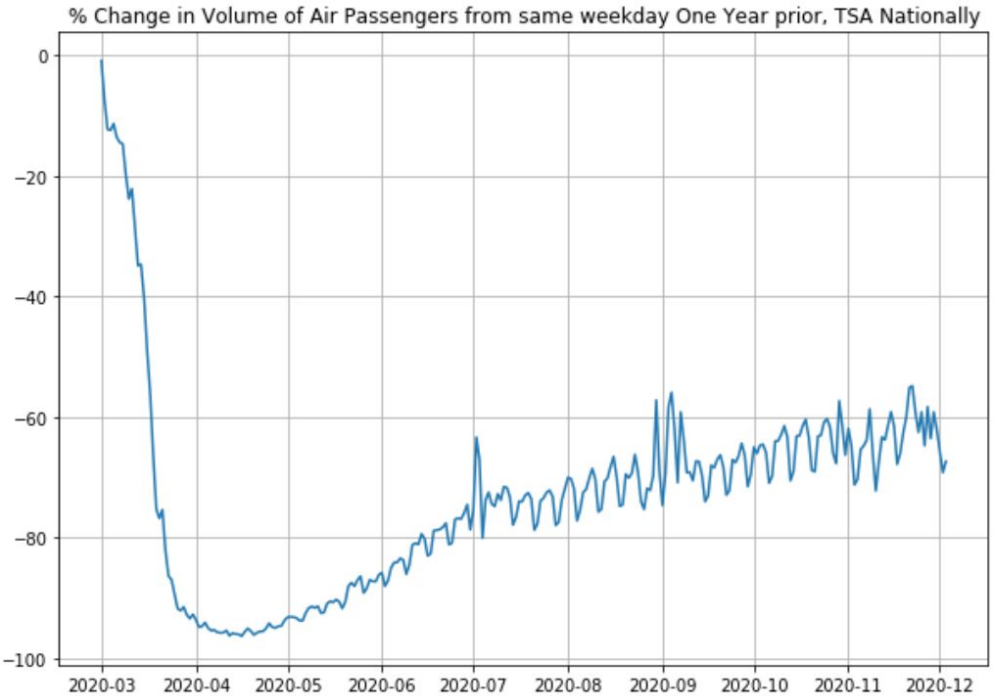
The percentage change of the volume of traffic through MTA bridges and tunnels from the same month a year prior: the x-axis is the months in a year, and the y-axis is the percentage change. In most months the percentage change is below or significantly below the 0-line, which means the volume of 2020 is much lower than the one of 2019.



The traffic volume through MTA bridges and tunnels by day from January 2019 to December. There is a significant slump at the time around April 2020, which causes the amount of traffic being of 2020 being fewer than the one of 2019. Here, we can notice the surge in traffic during the week of the 13th to the 20th of September, which was confirmed to be caused by the reopening of the NYC Public School system on the 21st of the same month.



The amount of TSA air passengers in 2019 vs 2020 compared by month: the x-axis is the months in a year, and the y-axis is the number of passengers. In each month of a year, the number of passengers is much smaller than the one of 2019, which means much fewer people are traveling by air in 2020.



The percentage change of the volume of air passengers from the same weekday a year prior: the x-axis is the timeline in a year, and the y-axis is the percentage change. The percentage change has always been below the 0-line, which means the volume of 2020 has always been lower than the one of 2019. It further demonstrates just how few people have travelled by air in 2020, almost bringing the airline industry to a complete halt during the second half of March and the month of April.

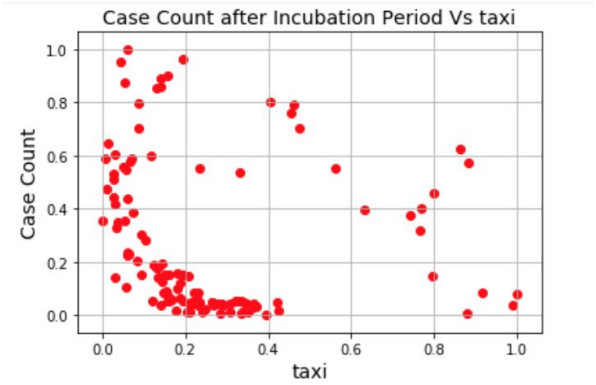
Statistical Analysis and Results for Second Hypothesis

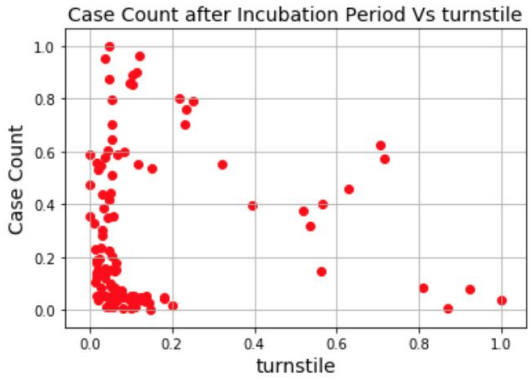
After successfully proving our initial hypothesis, we then tackled the second and more complex hypothesis: analysing the correlation between our indicators and COVID-19 new cases in NYC, with the goal of developing a machine learned model that can accurately predict future case counts given the current values for the transportation indicators and virus incubation averages. In order to do so, we outer joined all the processed datasets mentioned above, and filtered by the dates where we had data for all indicators. Unfortunately, this heavily reduced out available data points for training and testing for the period between March 1st and June 30th 2020, but since our datasets had daily frequency it still provided us with 119 records to split between training and testing. The lack of more recent data for some of our datasets proved to be a real challenge during this analysis, as the more complex the models the more data points are usually required to effectively train and test the necessary parameters. We were thus limited on how complex we could allow our models to become.

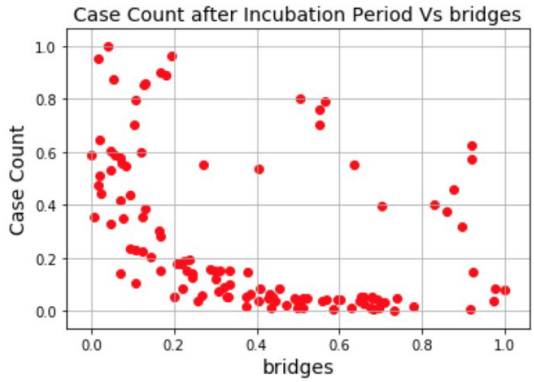
First, we modified the indicator data (the x-values) into a four day rolling average, and then shifted the COVID-19 case count (the y-value) forward 7 days. This was done to match the transportation indicators on a given day with the possible COVID-19 case count appearing 7-10 days later. This timeframe was confirmed by the CDC to be the updated virus incubation period.

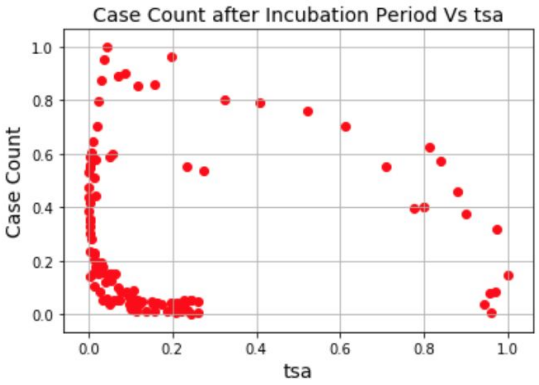
Then, we performed a MinMax scaling on each attribute, to scale the ranges of values to the range [0, 1], but while maintaining their original distribution. On contrast, performing a Standard Scaler on the data would change the distribution of the values to that with a mean of 0 and variance of 1, which would distort the relationships found and plotted in our previous section. Alternatively, keeping the attributes in their absolute scale would negatively affect the performance of some predictive models, as they would unnaturally bias the weightings towards the largest-valued attributes. Once the preprocessing was finished, we split the x and y attributes into training and testing samples, to prevent overfitting of the algorithms, and started with the actual regressions from the sklearn python module.

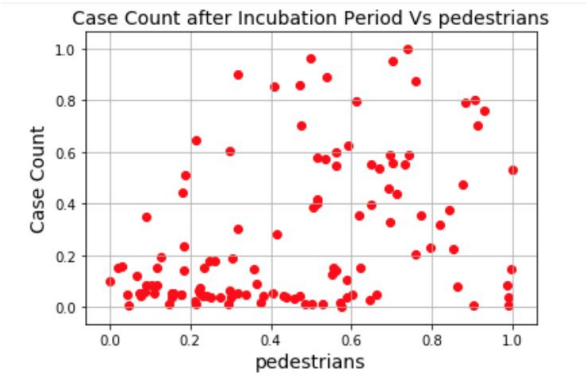
Before doing so, we plotted the relationship between each x-attribute and y-values, shown below. Since no single variable provided visual proof of a strong correlation with the y-values, we decided to use all of the indicators together to predict the case count.

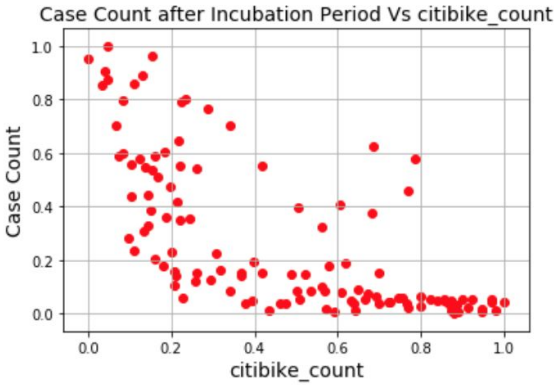


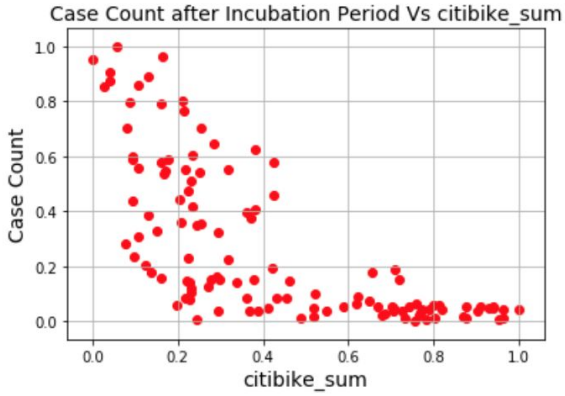




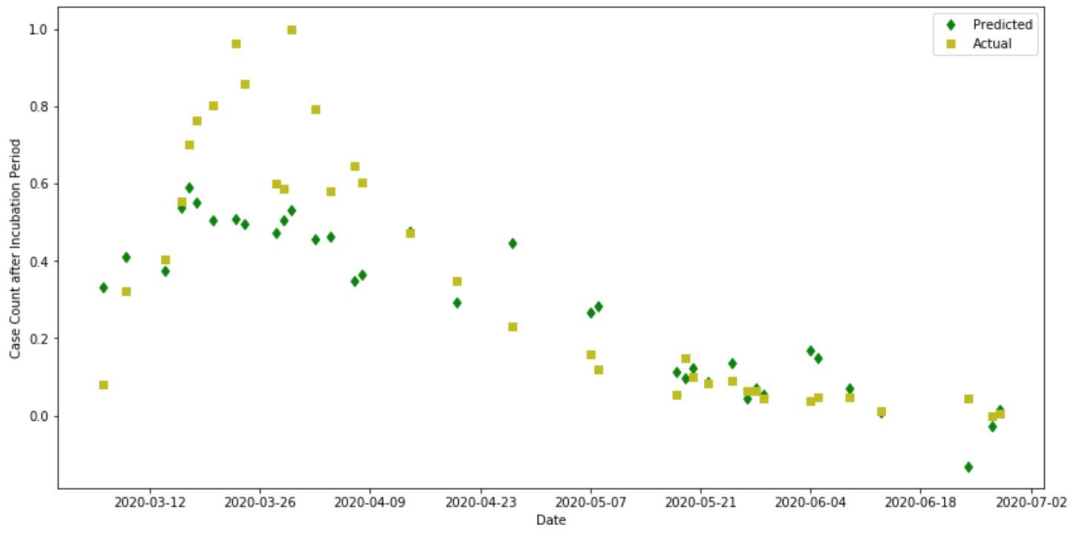






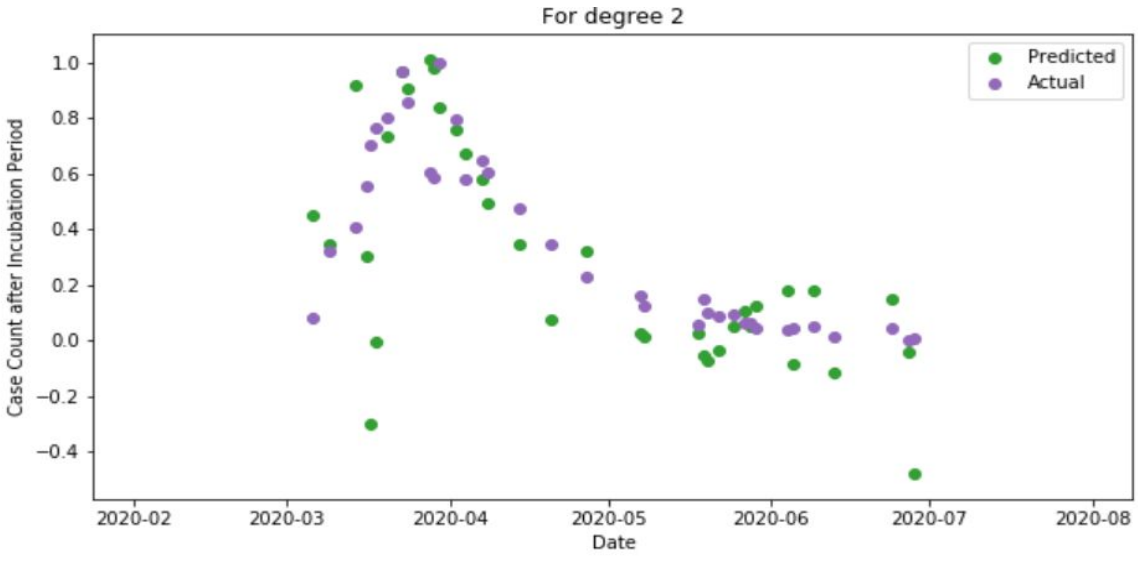


We began with the Multi-Linear Regression model, which estimates the weights for the x-attributes as they form a linear combination meant to predict the y-value. The weights that minimize the mean squared error, or MSE, will be chosen to fit the line. The MSE is the mean of the squared differences between the y-value predicted from a set of x-values, and the actual y-value associated with them. We used this score on the testing data to assess the predictive value of the different regressions. After fitting with the training data, the plot between the predicted y-values and the true testing y-values and the associated MSE score is shown below.

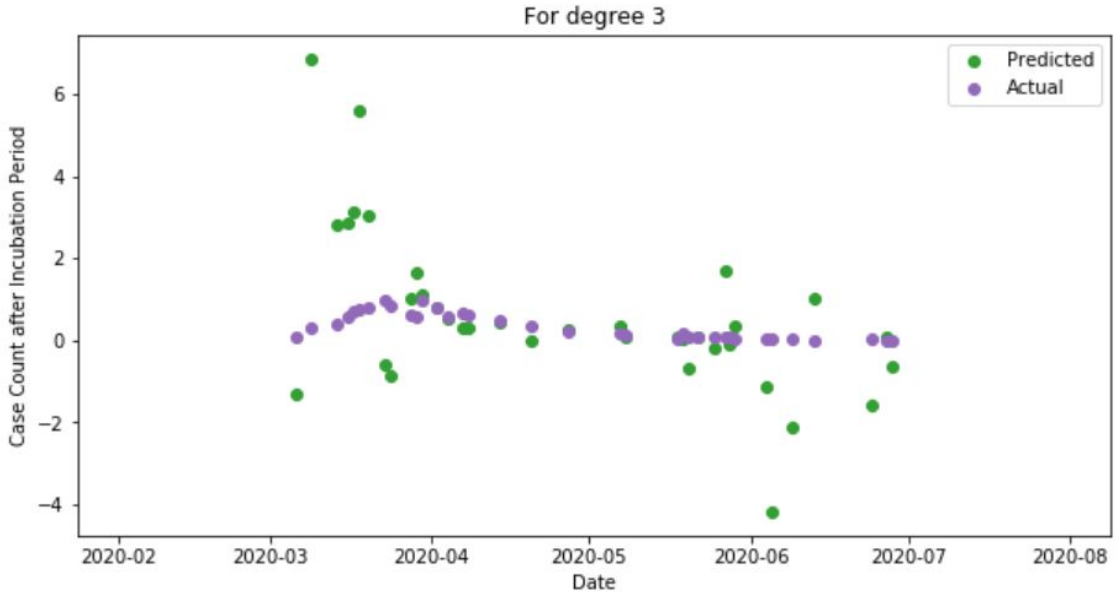



```
The mean squared error is 0.03419377349282487
```

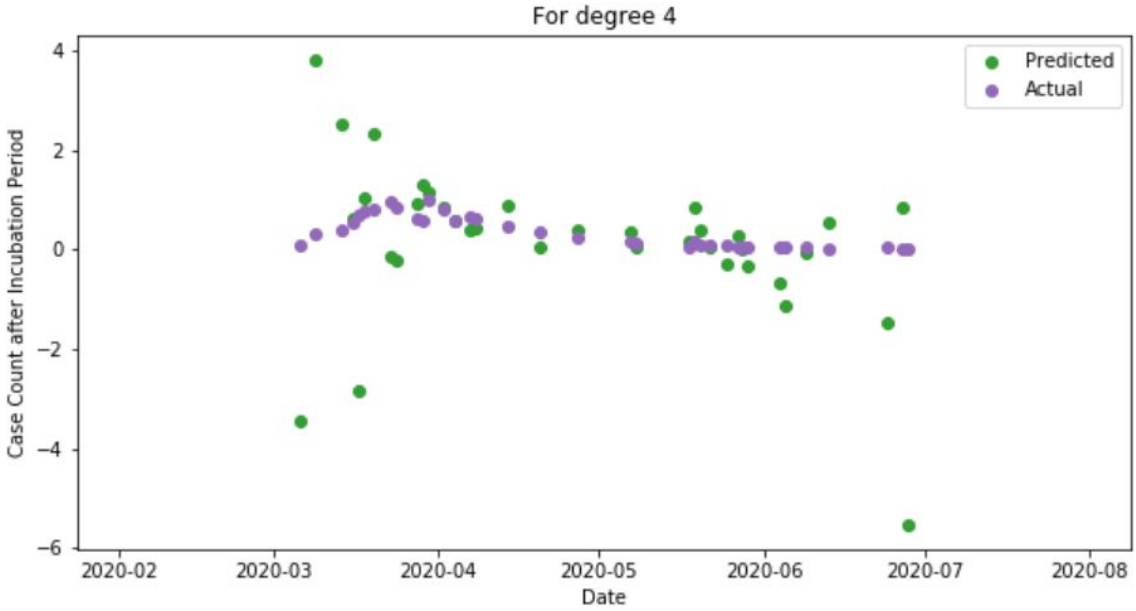
We established this as our benchmark. Then, we tried the Polynomial Regression. Given that we have 7 indicators, the possible maximum degree of the polynomial would be 7, but since Polynomial models are much more susceptible to overfitting than Linear ones, and the amount of testing data to correct such overfitting was a scarce resource for us, we decided to limit the maximum degree to be on the range [2, 5], and performed a regression with each possible max degree. The results are shown below.



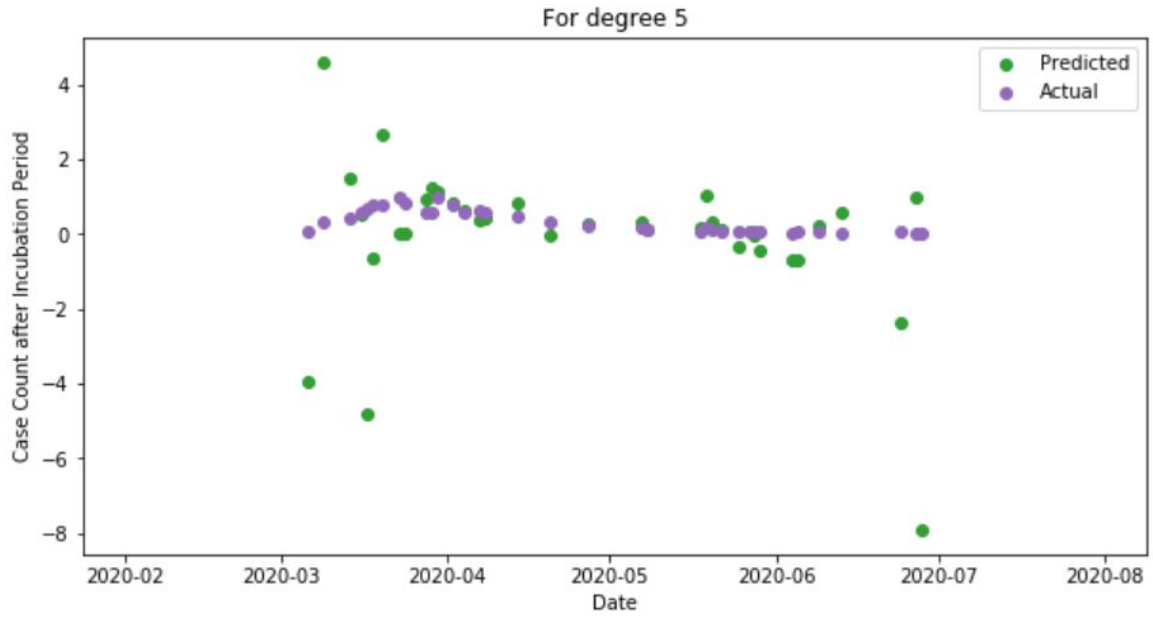

```
The mean squared error is 0.08329375640783764
```




```
The mean squared error is 3.594024006944973
```



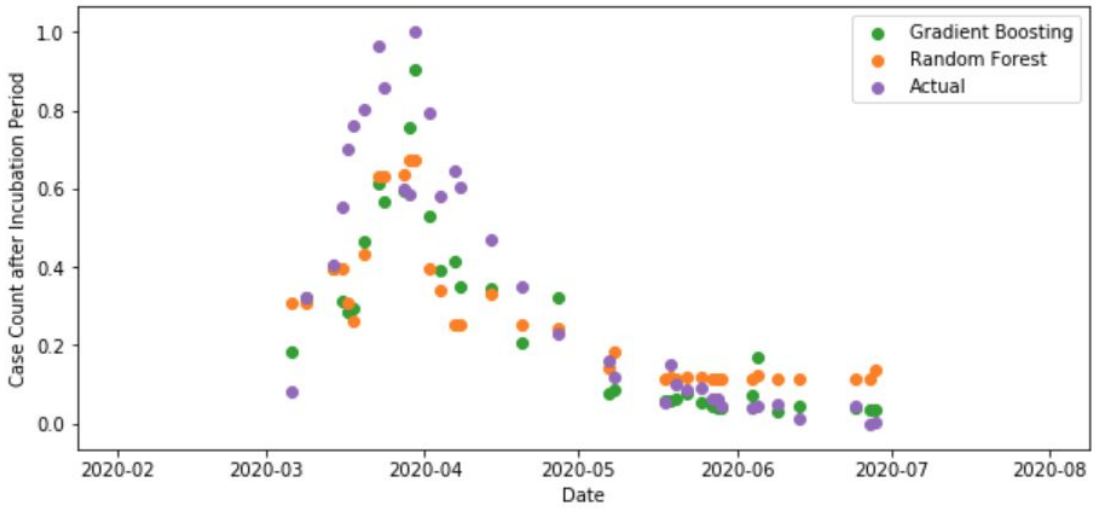

```
The mean squared error is 2.3347481302309676
```




```
The mean squared error is 4.070542382125195
```

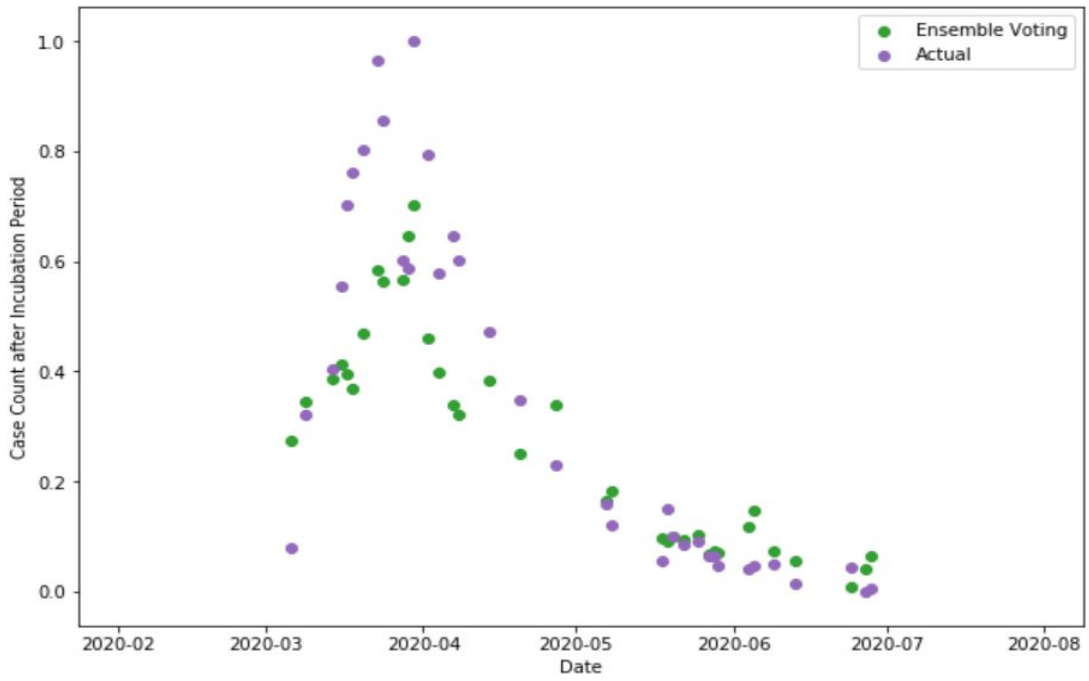
As seen, the Polynomial Regression with max-degree of 2 returned the smallest MSE between the testing data and predicted values, and even such value was much higher than our benchmark. This was due to the overfitting of these regressions on the training data, which then did not yield accurate results on the testing data. Due to this, we shifted towards decision tree-based regression models, which are much less prone to overfitting while still being capable of inferring highly accurate predictions on complex data relationships.

We fitted both Gradient Boosting and Random Forest models, as they both use a large amount of decision trees to assess the relative importance of each feature, and then decide the final prediction from a combination of the results from these trees. This is called ensemble regression, and the two models vary in the way the combination is performed. The results for the predicted and actual testing data are shown below on the same graph, for comparison purposes.



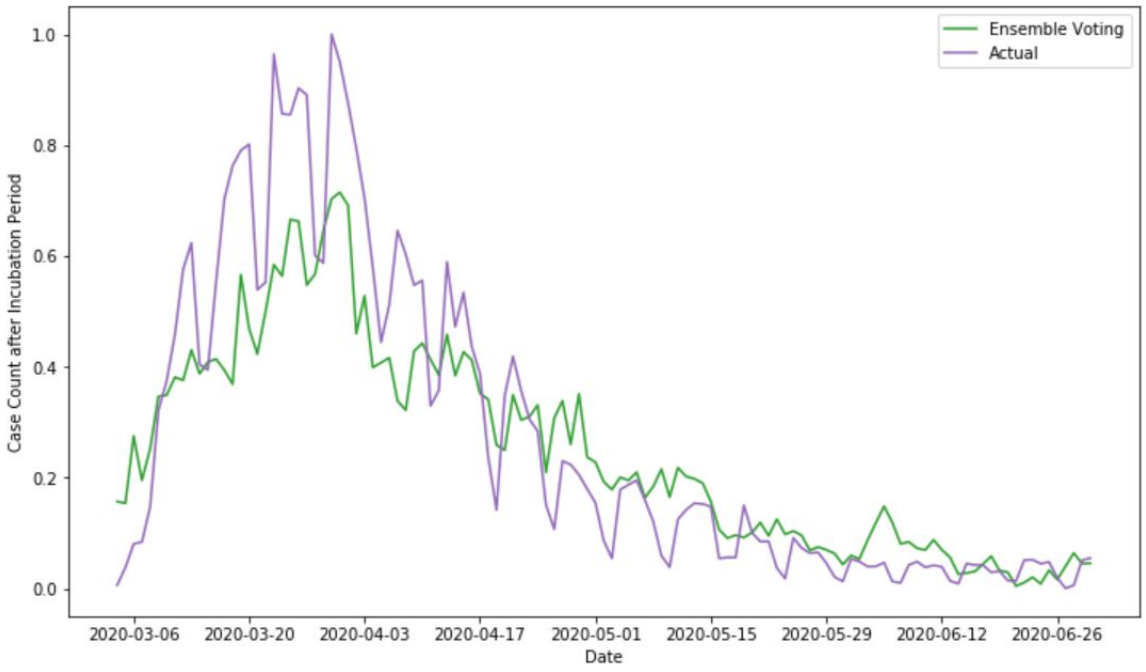

```
The mean squared error is 0.03135172442572189 for Random Forest
The mean squared error is 0.04184058392382553 for Gradient Boosting
```

The Random Forest regressor performed better than the benchmark, while the Gradient Boosting regressor performed just slightly worse. However, we still thought there was room for improvement and so took the concept of ensembles one final step further. This was done through sklearn's Voting Regressor, part of the sklearn.ensemble submodule. The predictors introduced into the Voting Regressor were the top 3 predictors so far: the Linear, Random Forest, and Gradient Boosting regressors. The results for the ensemble of these regressors according to the Voting regressor are shown below.




```
The mean squared error is 0.031262097877437584 for the Enseble Voting
```

As seen, the Voting Regressor performed better than any single predictor on the basis of MSE scores, and brought said value down to 0.03127. However, this score had been calculated so far on the scaled y-values, so we now transformed this value back to the absolute scale of the y-values to understand its actual accuracy in terms of case counts (and so its practical significance). Scaled back to the original range, the MSE score became 314.86. This means that, on average, the differences between the predicted case count after incubation period and the actual case count for that day were 314 cases. The plot for the whole data set is shown below to show its accuracy.



As such, we managed to successfully prove hypothesis two by showing a statistically significant correlation between the transportation indicators on a given day and the amount of new COVID-19 cases after the incubation period of virus infections for that day. As of the date of writing, there are over 10,000 case counts per day in NYC, so a mean squared error of 314 would be considered as highly accurate. It thus has practical significance and could potentially be of help to health services employees and other frontline workers that need to make sure they have enough medical supplies to best battle the virus before spikes in infections occur.

Conclusion and Further Improvements

Throughout this paper, we were able to successfully provide significant insights into the effects of COVID-19 on transportation in NYC. By partitioning the task into two different hypotheses, one to compare transportation indicators during 2020 to those in 2019, and the other to find and capitalize from a relationship between said indicators and new infection cases in NYC, we were able to not just look at the effects with a reactive lens, but instead turn said insights into proactive measures and predictors. These could in turn help frontline workers and health service employees better prepare themselves for future case count surges and against the subsequent shortages in medical supplies that have unfortunately happened on repeated occasions on a local, national and even global scale.

References

Articles

Barry, D., & Heisler, P. T. (2020, August 21). The Sounds of New York City Have Changed. Listen for Yourself. Retrieved from https://www.nytimes.com/interactive/2020/08/20/nyregion/nyc-sights-sounds-coronavirus.html

Melgar, L., Rigdon, R., & Rust, M. (2020, March 28). How Coronavirus Has Changed New York City's Transit Landscape. Retrieved from https://www.wsj.com/articles/how-coronavirus-has-changed-new-york-citys-transit-landscape-11585404000

Spivack, C. (2020, May 21). Here's how the coronavirus pandemic is affecting public transit. Retrieved from https://ny.curbed.com/2020/3/24/21192454/coronavirus-nyc-transportation-subway-citi-bike-covid-19

Datasets

Citibike (2020). System Data. Retrieved from https://www.citibikenyc.com/system-data

Transportation Security Administration (TSA) (2020). TSA checkpoint travel numbers for 2020 and 2019. Retrieved from http://www.tsa.gov/coronavirus/passenger-throughput

Taxi and Limousine Commission (TLC), (2020). TLC Trip Record Data. Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Metropolitan Transportation Authority (MTA) (2019). Turnstile Usage Data: 2019. Retrieved from https://data.ny.gov/Transportation/Turnstile-Usage-Data-2019/xfn5-qji9

Metropolitan Transportation Authority (MTA) (2020). Turnstile Usage Data: 2020. Retrieved from https://data.ny.gov/Transportation/Turnstile-Usage-Data-2020/py8k-a8wg

Department of Transportation (DOT), NYC OpenData (2020). Brooklyn Bridge Automated Pedestrian Counts Demonstration Project. Retrieved from

https://data.cityofnewyork.us/Transportation/Brooklyn-Bridge-Automated-Pedestrian-Counts-Demons/6fi9-q3ta

Police Department (NYPD) (2020). Crash-Pedestrians-2020. Retrieved from https://data.cityofnewyork.us/Public-Safety/Crash-Pedestrians-2020/anau-hnpq#__sid=js0