**MMAI 5040 – BUSINESS APPLICATIONS OF AI 1**

**(2021 Winter)**

**Impact of COVID-19 on New York City Bike Sharing Program**

Professor: Divinus Oppong-Tawiah



Submitted by:

Tian Wang (213990353)

Yijian Zhang (218337345)

Jordan Feng (218403204)

Dhrun Lauwers (218070342)

Joyce Otonglo (218420687)

Shaoshi Zhang (217875360)

Submitted on:

April 5th, 2021

**Table of Contents**

[**Executive Summary**](#_heading=h.v5h3lzx1w2nj)2

[**Problem Statement**](#_heading=h.ogcsqgsuxqsx)3

[**Methodology**](#_heading=h.7heffeukvwex)4

[Data source](#_heading=h.26rekebnk97k) 4

[Descriptive Analysis](#_heading=h.lt4opxzc6j9o) 6

[Supervised learning](#_heading=h.9qpk3ee7un89) 7

[Unsupervised learning](#_heading=h.qnstnt86qd4h) 8

[AI Implementation, Decision Rules and Cost Benefit](#_heading=h.qar6lct65dqe) 8

[**Findings & Interpretation**](#_heading=h.v4hb7cffheh9)10

[Descriptive Analysis](#_heading=h.u8whexwpxvpy) 14

[Supervised learning](#_heading=h.ej6wpc5ynvtq) 20

[Unsupervised Learning](#_heading=h.nxwqys1q1g3z) 27

[**Insights & recommendations**](#_heading=h.x1st2voua2l9)23

[**Conclusion**](#_heading=h.mlzbwx96uek9)25

# Executive Summary

As the COVID-19 virus becomes the most serious pandemic disease across the world, it is negatively impacting people’s daily life and changing behaviors. For this project, our group will focus on the bike-sharing program in the city of New York City, and use the methodologies we learned from this course to develop a series of findings, interpretations, insights, and recommendations. First, we will use two independent datasets: “Citi Bike Trip Histories” and “the Daily Cases, Hospitalizations and Deaths provided by the Government of NYC” to analyze the issue. Methodologies includes “Descriptive Analytics”, “Supervised Learning”, “Unsupervised Learning”, and “AI Implementation”. We tried to do the data cleaning, which will help us to analyze, identify, and correct dirty data in a data set. After dealing with the datasets, supervised learning should be used to develop prediction, RMS table, correlation matrix and feature importance. Those things will provide our group with full of insights whether the COVID-19 directly impact the number of bike rides and the average duration of trips, whether there is some positive relationships or negative relationships between each variables and which feature is the most important one. Unsupervised learning will provide insights with us that how the company should do station distribution and bike management. Lastly, AI implementation and cost benefit analysis will analyze whether our strategy is viable and how could we implement the solution. In fact, our group generated the full of insights and interpretations based on the analysis framework. Those findings give our group direction on how we should make recommendations to attract more customers for Citi Bike when the recovery begins. As the customers increase, we explore how should we manage bike availability properly. At the end our project, the conclusion summarizes all the issues we have already done again and explain to readers the importance of our project.

# Problem Statement

During the whole of year 2020, COVID-19 virus became the most serious pandemic all around the world. Most people who fall sick with this virus will experience moderate to serious symptoms, and an alarming number of deaths continue to occur. Because COVID-19 has a very powerful infectious feature, people need to be quarantined at home if necessary. People who go outside need to wear face masks. Under this situation, public transportation becomes very dangerous, as people could be easily infected by the COVID-19 virus, because public transportation is usually closed and crowded. If even one person carries the virus, all people are in danger. In this way, COVID triggered significant change in consumer behavior, as more and more people who want to go outside begin to choose bicycles as their transport means. Bicycles are relatively safer than mass transit facilities such as railway or bus. Our project will analyze the datasets in order to understand the extent of the impact of consumer behavior on the demand for biking-sharing programs. Thus, we could make the business plans for future expansion, suggestions and predictions. The change in consumer behavior has the potential to significantly disrupt the current and future demand of bike sharing programs. Plans to expand the program created before the start of the COVID pandemic may no longer reflect current demand. Our group tried to use an algorithm to determine how a company could improve and manage a bike-sharing program.

# Methodology

## Data source

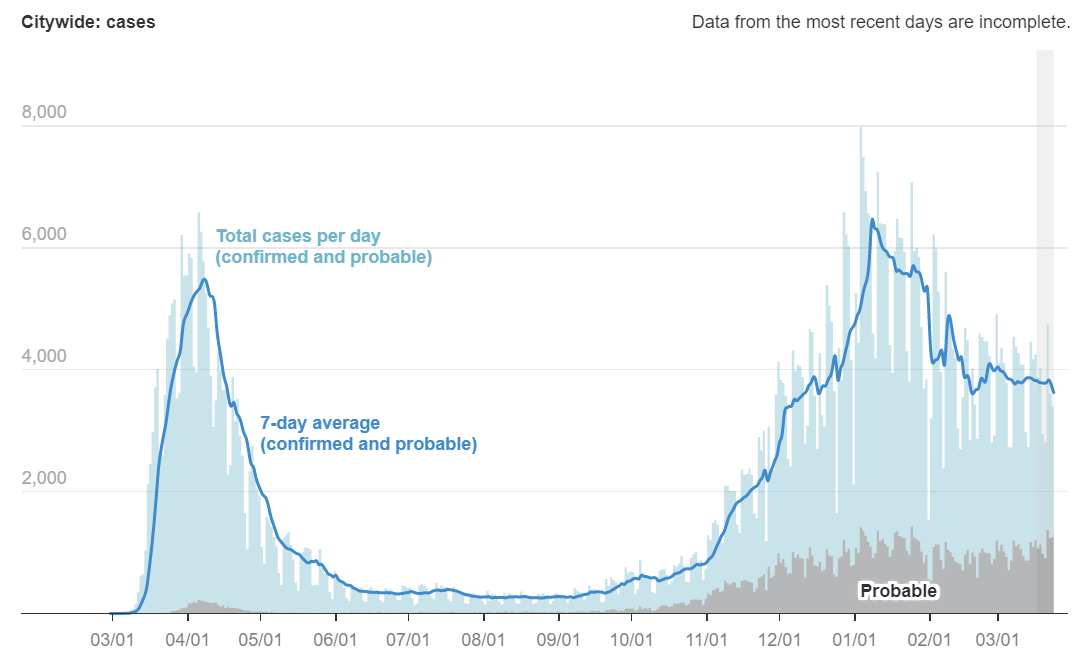
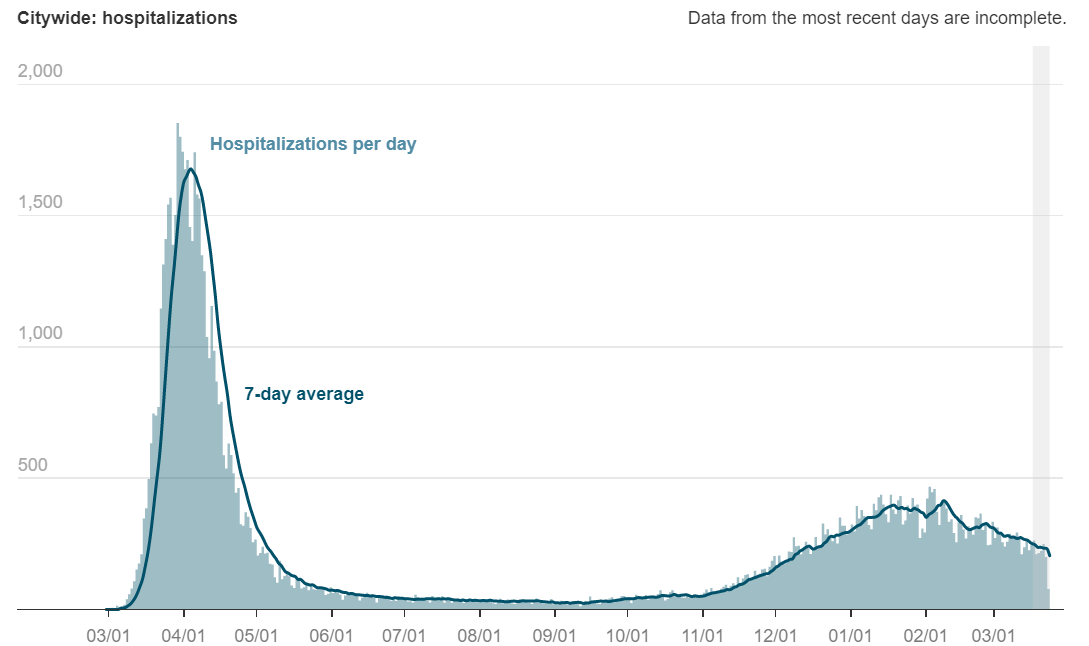
In this project, our group will use two independent datasets. The first one is Citi Bike Trip Histories which provides the details of each trip on a daily basis. The other dataset is the Daily Cases, Hospitalizations and Deaths provided by the Government of NYC to track the trend of COVID-19.

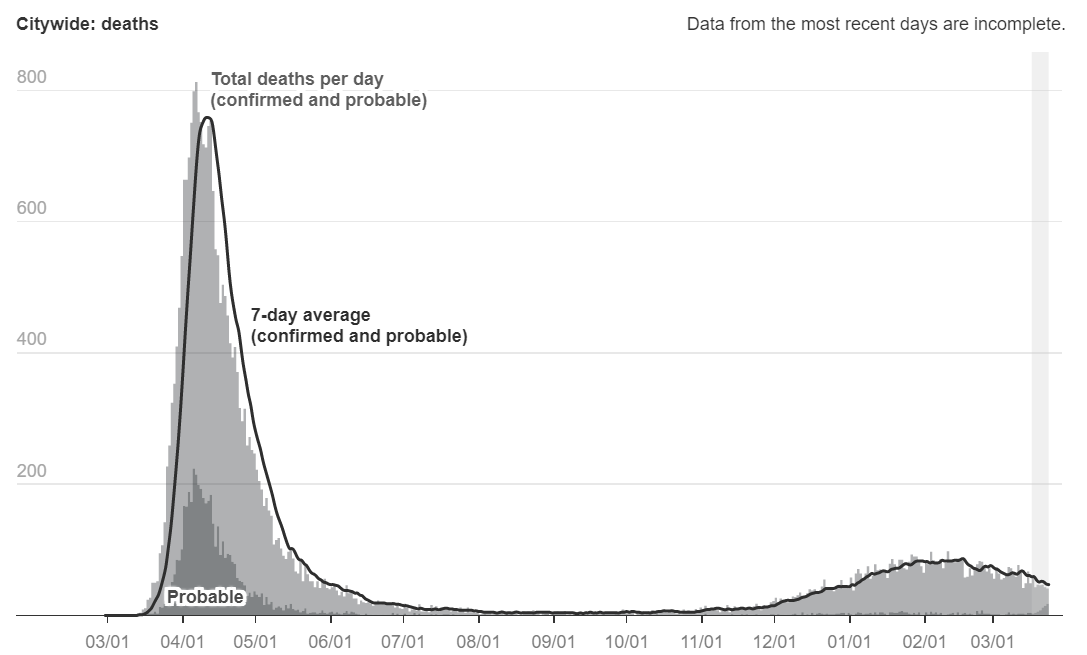
Where do Citi Bikers ride? When do they ride? How far do they go? Which stations are most popular? What days of week are most rides taken on? These kinds of questions are fundamental issues that our group as a data team need to know. Based on the database “Citibike data”, our group could discover the answers to these questions and maybe more interesting findings. Our group will use the data to do analysis, development and visualization.

Our group published the dataset from the “Citi Bike System Data” website, the data includes: Trip Duration(seconds), Start Time and Date, Stop Time and Date, Start Station Date, End Station Date, Station ID, Station Long, Bike ID, User Type(Customer = 24-hour pass or 3-day pass user; Subscriber = Annual member), Gender(Zero = unknown,; 1 = male; 2 = Female) and Year of Birth.

The second dataset shows the COVID-19 trends in New York City since the city’s first confirmed case was diagnosed on February 29, 2020. The website has already provided three charts with our group.

These charts show the daily number of confirmed and probable COVID-19 cases by diagnosis date, hospitalizations by admission date and deaths by date of death:





## Descriptive Analysis

After determining the data sources that need to be used, our group performed data cleaning and exploratory data analysis on these two data sets.

Data cleaning is the process of analyzing, identifying, and correcting dirty data in a data set. This is a necessary process for converting "wild" data into "manageable" data. Through removing duplicate data and structural errors, handling missing data, and verifying the accuracy of the data in the data set which can improve the data quality. So using correctly cleared data can more easily help projects generate valuable business insights and actions. At this stage, we first preview the data, and clean up the data based on the observation results through the following steps. First, correct outliers. Second, complete missing information. Third, create new analysis functions, And the fourth step is to convert the field into the correct format for calculation and presentation.

According to the 4C’s principle of data cleaning, we check the data to confirm whether there are any outliers, null values or missing data and then deal with them. For some potential outliers, we will wait until exploratory analysis is completed before determining whether to include or exclude from the data set. In addition, we will handle formatting. Our classification data is imported as objects, which makes mathematical calculations difficult. For this data set, we will convert the object data type to a categorical dummy variable. Besides, since the start time and stop time in the original data set both contain date and specific time, we convert them to date data type so that we can merge Citibike data with COVID data by using a left join on the matching date.

Finally, we summarize and organize the characteristics of the data set through descriptive statistical information, and conduct exploratory analysis. We will deploy descriptive and graphical statistics to find potential problems, patterns, classifications, correlations, and comparisons in the data set. Also, data classification, which includes qualitative or quantitative, is also important for understanding and choosing the correct hypothesis test or data model. Besides, we will also visually analyze these data to find trends in the data.

## Supervised learning

After dealing with the issue of datasets that need to be used, our group wants to use supervised learning- MLP first to find whether COVID cases directly impact the number of bike rides and the average duration of trips and then make the forecasting.

MLP is a neural network connecting multiple layers in a directed graph, that means the signal path will go through one way and each node has a nonlinear activation function. MLP uses back propagation as a supervised learning technique. Because there are multiple layers of neurons, which includes input layer hidden layers and output layer. There are three steps in MLP. It will compute the temporary outputs first and then compare outputs with desired targets, finally it will adjust the weights and repeat the process. For this algorithm, our group would like to change the number of hidden layers and nodes to see whether the RMS increases or decreases.

RMS serves two main purposes: one is to serve as a heuristic for training models and another one is to evaluate trained models for usefulness or accuracy. For our project, we want to find which condition has the smallest RMS error and based on that condition to choose the appropriate number of nodes and hidden layers.

Then our group also tried to develop the correlation matrix, which is a table showing correlation coefficients between variables that our group chose. Each cell in the table shows the correlation between two variables. Because our group tried to define whether there is a positive or negative relationship between each variable and to define feature importance next step.

Features are important to predictive models and influence models. The purpose of the feature would be much easier to understand in the context of a problem. Our group would like to use some machine learning model first to print the score and then compare them to check which model is the best one. The model includes ‘Decision Tree’, ‘Random Forest’, ‘Linear Regression’, ‘KNN’, ‘Stochastic Gradient Descent’, ‘MLPR’. Finally, our group wants to define the importance of each feature.

## Unsupervised learning

With respect to station-level detail forecasting, unsupervised learning is conducted through k-means clustering and hierarchy clustering.

The data is first grouped by starting station id, then by date, forming a multi-index dataframe. Then the mean and count are conducted to calculate the ride\_count and average trip duration for each station on each day. The data is then unstacked, forming a row for each station, with ride count and average trip duration for each day of the time period of interest (365 \*2) as columns. With little samples to cluster (52 rows with each station as a row), Elbow method is conducted to determine the number of clusters in the dataset. It is a heuristic used in determining the number of clusters in the dataset. This method could plot the variation as a function of the number of clusters, and then pick the elbow of the curve as the number of clusters to use.

Then K-means clustering is used as well as single and complete linkage. In simple speaking, K-means clustering is an algorithm that could group our objects based on features into K number of groups. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster. We also want to build another feature heatmap via this method and to look at the feature importance and interpret it, thus we could generate the recommendation.

## AI Implementation and Cost Benefit

The authorities of New York City have laid out an ambitious revamp and expansion programme for the city, and there are a number of important considerations that play into this decision and present a viable business case to justify these implementations from a commercial business perspective. These key factors include the location of bike stations in relation to their potential demand, provision of bike service across a diverse range of communities and demographics, realizing a balance between availability of bikes at the stations, achieving acceptable bike utilisation rates per station, and need for availability of slots for return of bikes. Besides these, the city could consider additional features to be made available to users, all the while ensuring that the benefits to be gained outweigh the cost of implementation and operation of the model.

**AI Implementation**

The city plans to double the current service area by 35 square miles and triple the number of bikes to 40,000 over the next three years. There is a need to select the most suitable locations for the bike stations, as this will affect accessibility of the system, efficiency of bike use and ultimately profitability. Besides increasing capacity in the system, the expansion will foster inclusivity by bringing on board a diverse range of communities. For example, the South Bronx and northern Manhattan areas, known to be inhabited by low-income earners and minorities, will have stations in the next year. Rollout to similar areas is planned for the following years.

At the same time, the stations that are not viable towards providing support to the bike network will be closed and replaced. This way, the city is envisioning an equitable, accessible bikeshare system that works for its neighborhoods as well as the city’s overall transportation goals.

**Cost-Benefit Analysis**

The cost of doing this analysis is low for the city, and a lot of the data is available for free. That said, however, it’s hard to tell whether the findings from the models will work in the real world. Simulations and what-if analysis are still just estimates. We therefore recommend that the city starts the project with a pilot and in a phased approach. A small number of stations would be selected and their usage would be closely monitored to check such indicators as demand, new ridership patterns and feedback on user experience. Tests can be run on pricing structure while taking note of any limitations and barriers to scale, before bringing on board additional stations. This would allow proof of value before citywide rollout.

**Additional features**

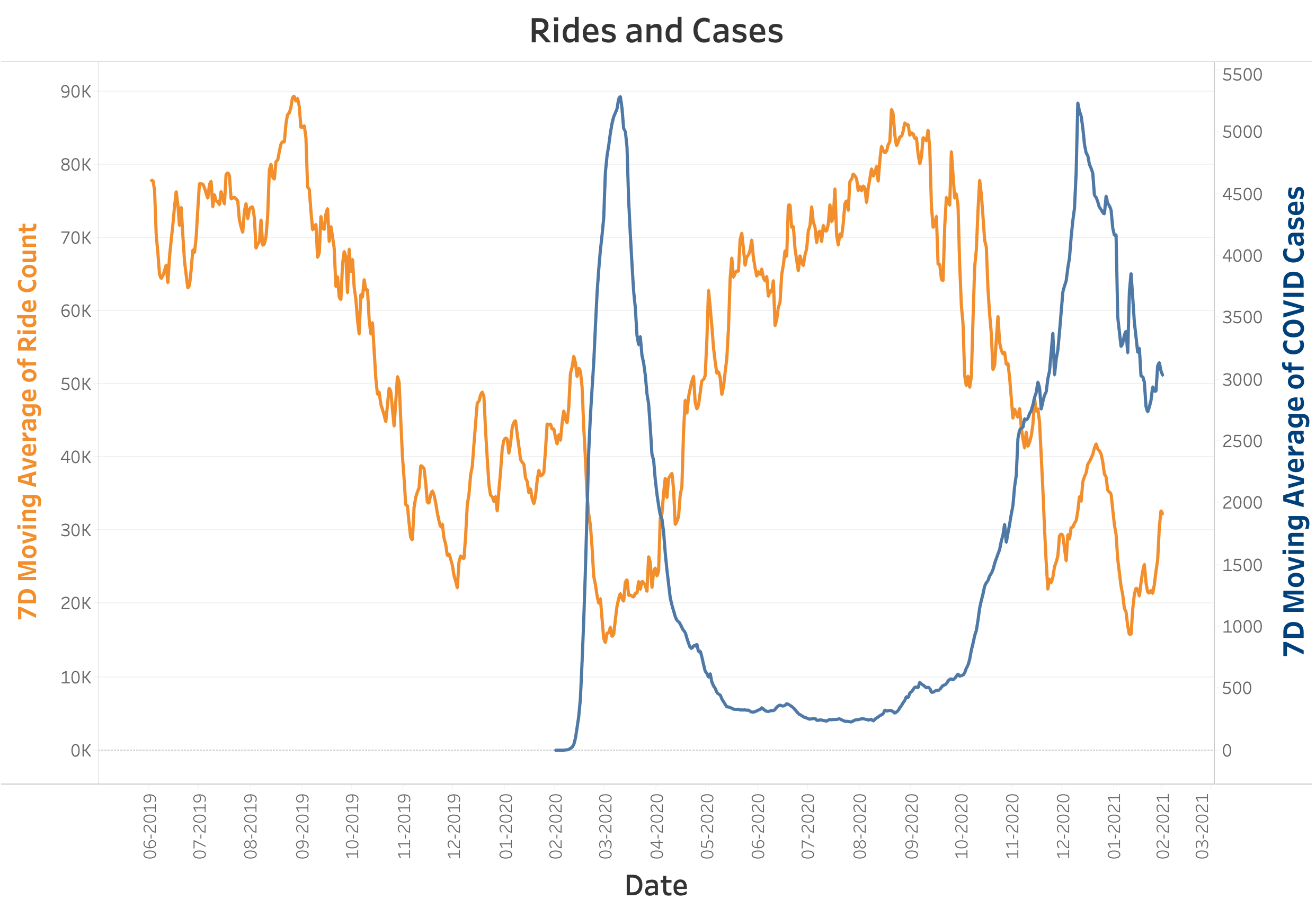
Some additional features could have been included that could make the model more robust. These include temperature checks in light of the covid pandemic, demographics e.g use by locals and tourists/ visitors. For example, the city is considering bike facilities for people with disabilities, as well as individuals who do not know how to ride traditional two-wheeled pedal bicycles.

# Findings & Interpretation

## Descriptive Analysis

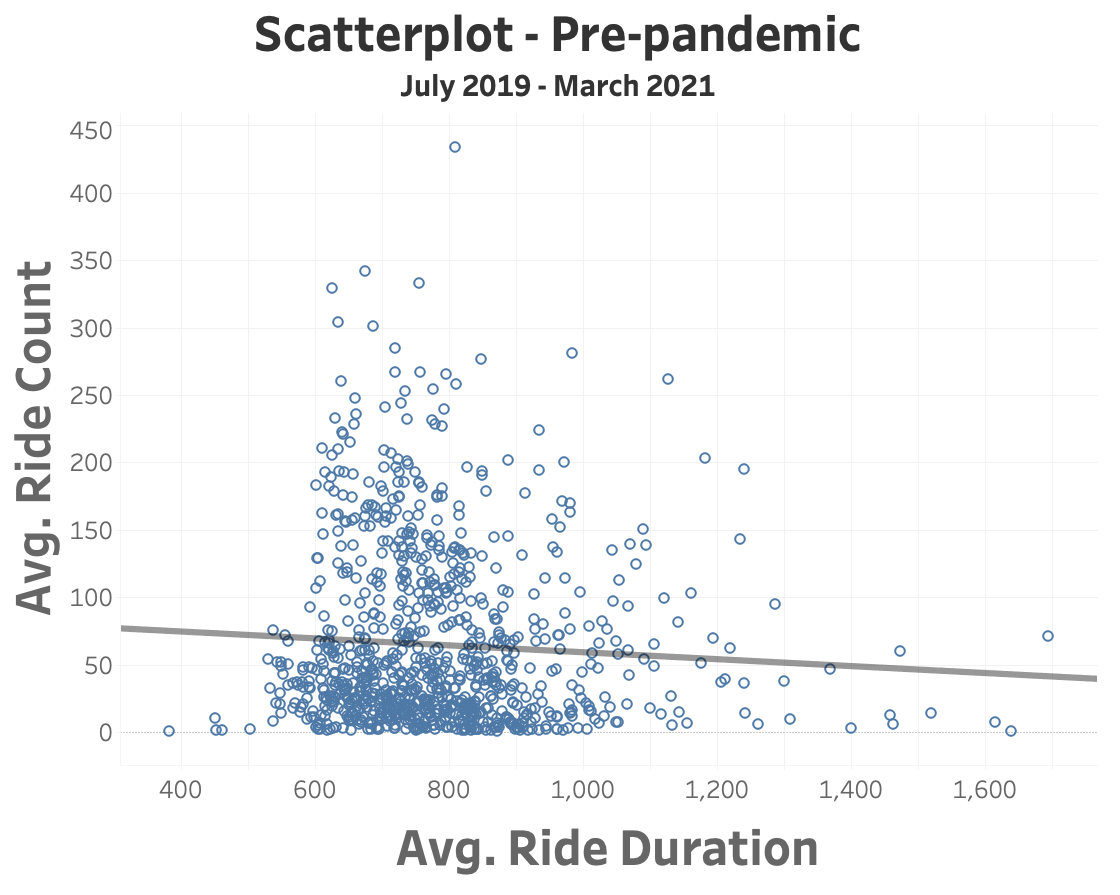
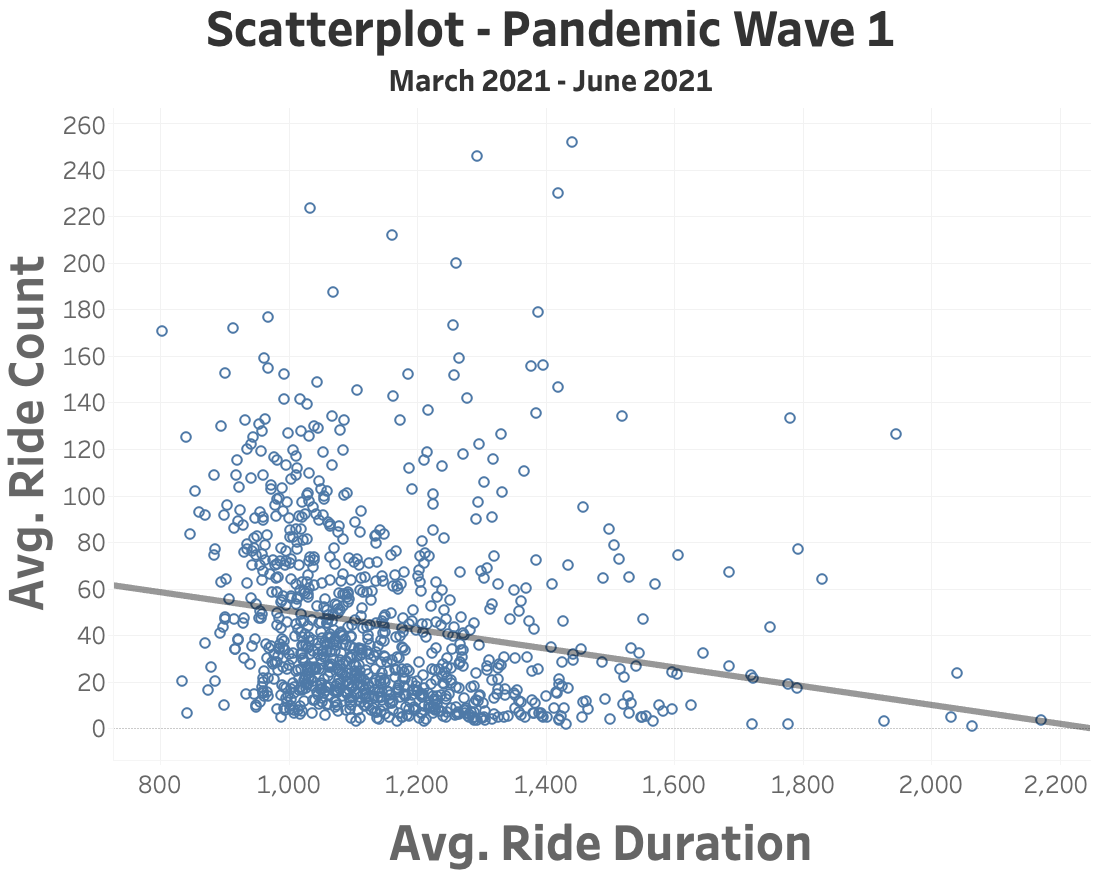
The Citibike network currently contains approximately 1,300 stations, stretching across NYC’s iconic neighbourhoods including Manhattan, The Bronx, Queens and Brooklyn. The network is continuously expanding, with more than 100 stations added per year as part of a multi-year expansion plan that started in 2019.[[1]](#footnote-1) The goal of the plan is the double the area serviced and triple the number of bikes in circulation by 2023.

Before the pandemic, the most popular stations were seeing an average of more than 300 rides per day, while less popular stations were seeing less than a dozen rides per day. The average riders per day at each station was around 80.

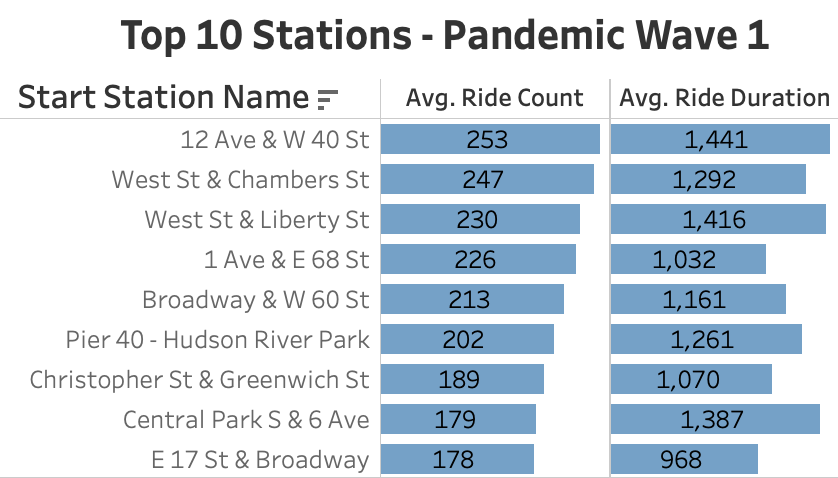
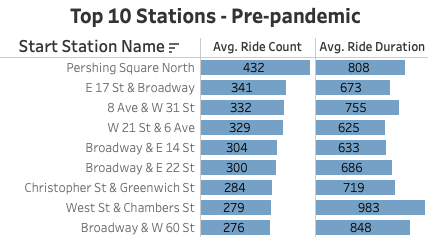
The average duration of a ride was a bit longer than 13 minutes, with many outliers that were likely due to customers who fail to return their bike. 

Initial visual inspection of the 7 day moving averages shows an inverse relationship between the number of overall rides taken on the Citibike network and the number of COVID-19 cases. Specifically, there is a noticeable drop in the number of rides taken that coincides with the COVID-19 outbreak in early 2020. As the number of cases declines throughout the summer of 2020, the ridership begins to recover to levels greater than the average of the previous summer, before the beginning of the pandemic. During the winter of 2020 and early 2021, the number of cases again rose to a peak with the ridership declining over the same time period. This relationship suggests that the on-going pandemic is of importance to the future of the Citibike network, and that a review of future plans is warranted.

**Ride Count and Duration**

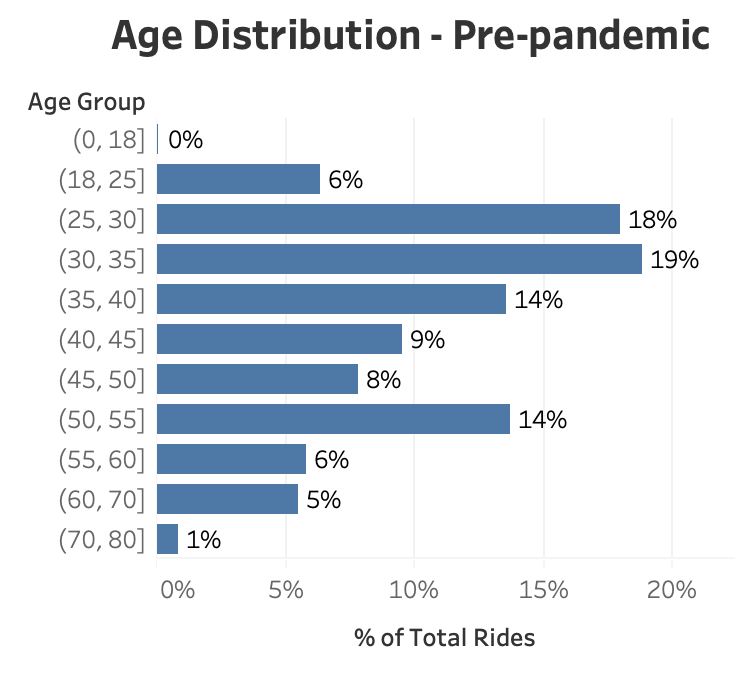
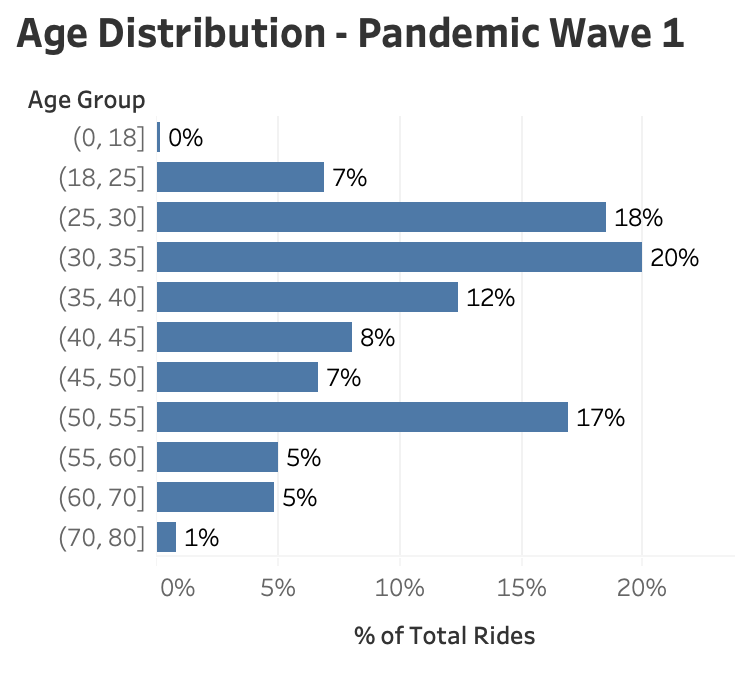


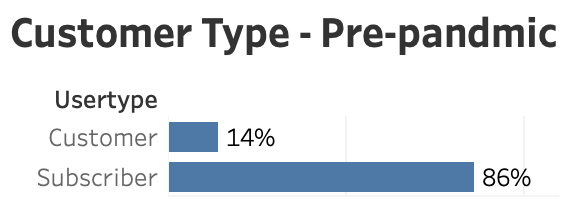
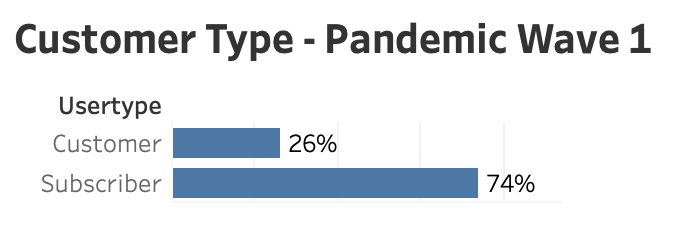
During the first wave of the pandemic, the average rides per day across all stations declined ~25% to around 60 rides per day while the average duration increased ~25% from 13 minutes to more than 15 minutes. One could argue that longer rides may suggest that customers are using the bikes as a means of exercise rather than a means of getting from one place to another.

**Changes to the most popular stations**

Perhaps more interesting is what happened to the list of the 10 most popular stations before and after the first wave of the pandemic. The majority of the top 10 stations has changed, likely due to a change in commuting patterns during this time. For example, Pershing Square North was by far the most popular station before the beginning of the pandemic, with more than 400 riders departing from this station each day. After the pandemic, it is no longer part of the top 10 stations as daily ridership fell to approximately 130 riders per day. Looking at things from a different perspective, the most popular station during the first wave of the pandemic saw little change in ridership compared to pre-pandemic levels. Overall, these findings suggest that bike traffic in certain neighbourhoods has been more affected by COVID than others, and that having the appropriate number of bikes at each station is critical.

**Changes in rider demographics**

** **

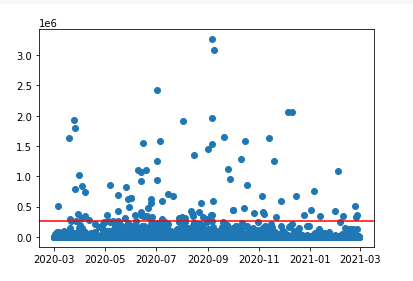
** **

When it comes to the age demographic of customers, there do not appear to be any major changes in the distribution across the age groups selected. However, when it comes to the number of customers who are subscribers versus one-time customers, there does appear to be a significant shift with the overall number of subscribers declining from 86% to 74%. This is likely due to the fact that commuters are more likely to be subscribers, and that commuter traffic declined as employees work from home. Also, assuming that subscribers have more regular travel patterns than one-time customers, it may become more challenging to have the right number of bikes at each station.

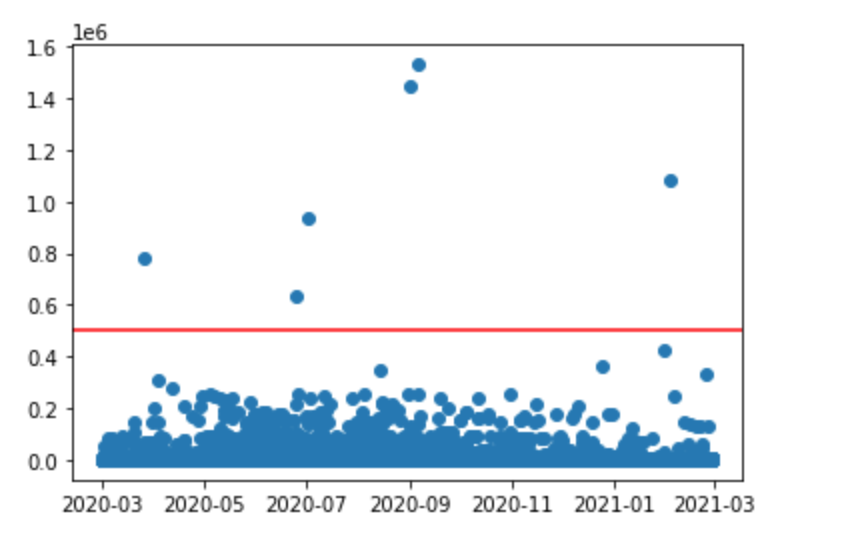
**Data cleaning**

After obtaining an appropriate data source, we first carried out data cleaning. The first step is to read the data and keep the required columns. In this step, we first read the data to get the daily number of confirmed and COVID-19 cases in New York City from March 2020 to the end of February 2021. At the same time, we read the trip records of Citi Bike during this period. In the first data set, we only kept the date, the number of confirmed cases in the city and the average number of confirmed cases within seven days. And we deleted hospitalization dates, hospitalized cases, death cases and other duplicated data to ensure simple and clear predictor variables. For the second data set, we only keep valid data such as the date and the duration of each trip, and other data related to customer levels are not used. By removing unwanted observations from the data set, including duplicate or irrelevant observations, can help us create a more manageable and more efficient data set.

Secondly, we try to correct the outliers. Since the COVID data comes from government websites, we believe that all the data are true and valid. Therefore, through plot the histogram, we have not found any outliers in these confirmed cases. However, for Citi bike data, the more than 300,000 trip records are too concentrated to track patterns through histograms. As a result, we use the describe() function and the scatter plot for further inspection. Through the scatter plot, we can see the trend more clearly, and indeed found some outliers in the Tripduration column. From the scatter plot of trip duration, we found there are 3 extremely obvious outliers that each trip is significantly longer than 2300000 seconds which is about 26.7 days. Also there are dozens of trips that are longer than 259200 seconds which is 3 days, the maximum length to hold the bike for customers.

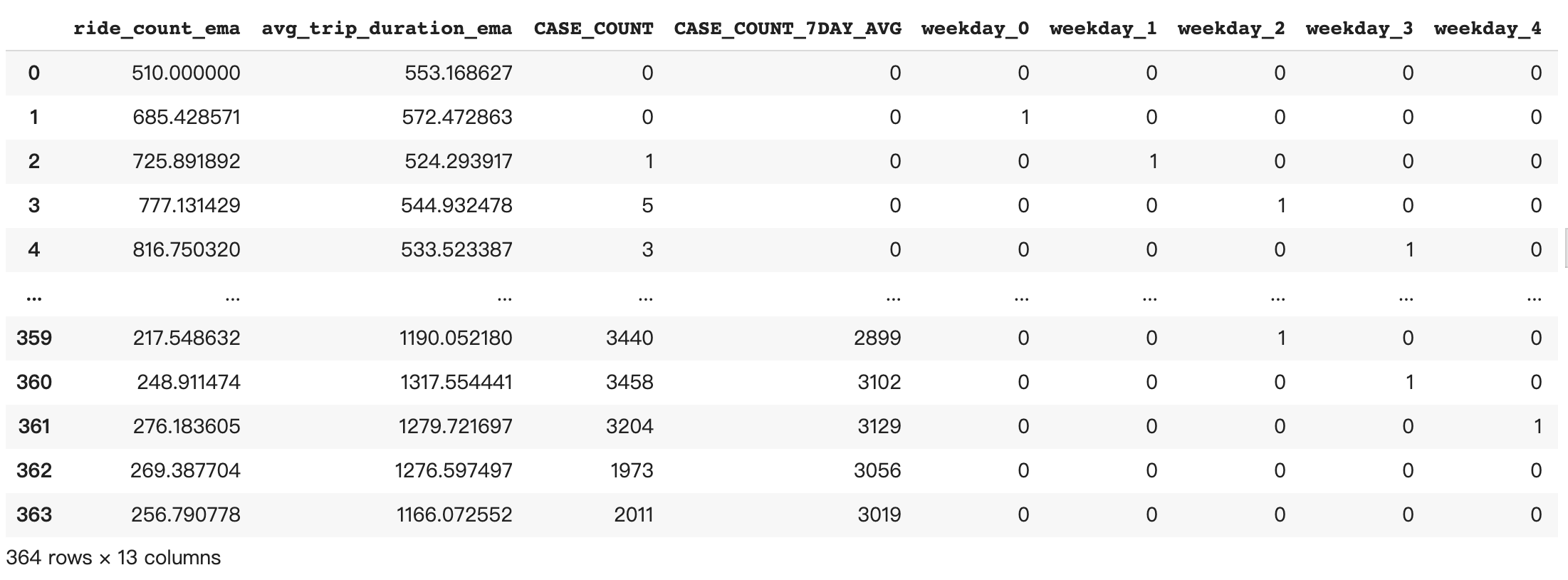


Based on the user type content, customers are either using a 3-day pass or 24-hour pass, they are not allowed to hold the bike more than 3 days. And most of the trips in the plot seem to exceed 500,000 Seconds (5.8 days), we speculate that it is because the user forgot to return to the bicycle. So we added an auxiliary line to the scatter chart according to the maximum duration of three days. The abnormal trips displayed on the horizontal line are regarded as outliers, and these outliers are replaced with the median of all trip durations. Finally, we filtered these outliers, as shown in the figure below, but there are still several trip records that exceed three days. These are the trip duration from subscribers, so we assume that these subscribers are allowed to return bicycles for more than three days.



In the third step, we use the isnull() function to find missing values and fill in the missing values with appropriate guesses. However, neither of these two data sets has any missing value.

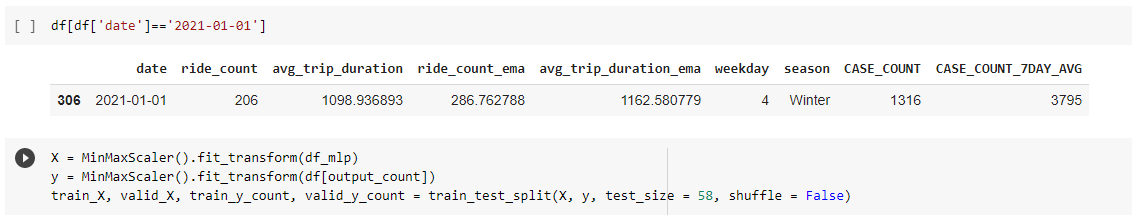
In the fourth step, we try to create new columns to get more data information so that it improves data quality. By adding a few columns to the bicycle data frame, we hope to see the number of trips viewed per day and the average trip duration. So we group the bike data frame by date and create two columns: ride\_count and ave\_trip\_duration. Besides, we would like to see the moving average of these two new columns to track with the trends. Therefore, we calculate the exponential moving average of these two columns, and then create 2 new columns: ride\_count\_ema and avg\_trip\_duration\_ema. At the same time, we are also interested in the influence of weekdays and seasons on cycling data, so we created two more columns called weekday and season. The last step is converting and merging the dataset. Since both the start time and the stop time in the original data set contain date and specific time, we convert them to date data type so that we can merge citibike data with covid data by using a left join on the matching date. The final complete data result as shown in the figure below, these data processing work helped us to lay a good foundation for our MLP prediction.



## Supervised learning



At first, our group chose the predictors and made sure the output.

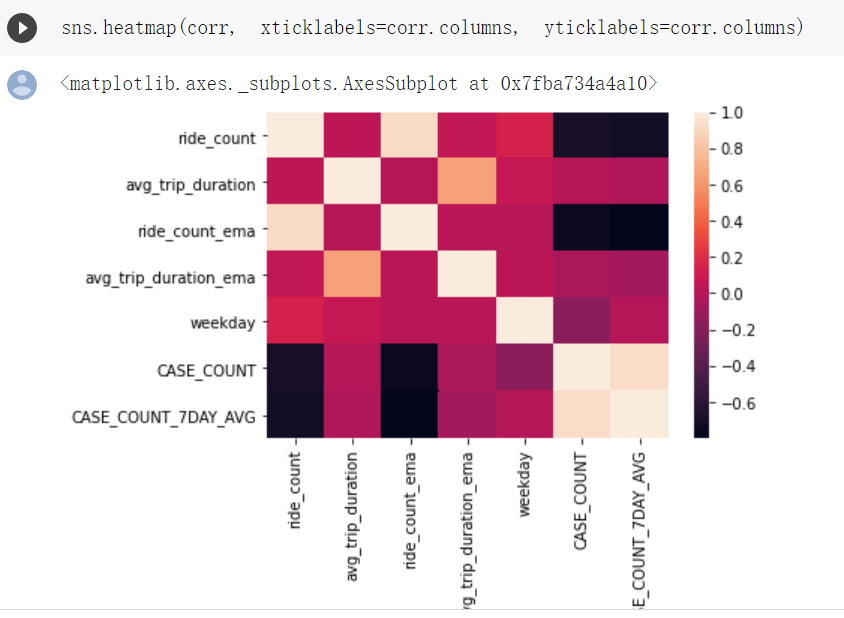
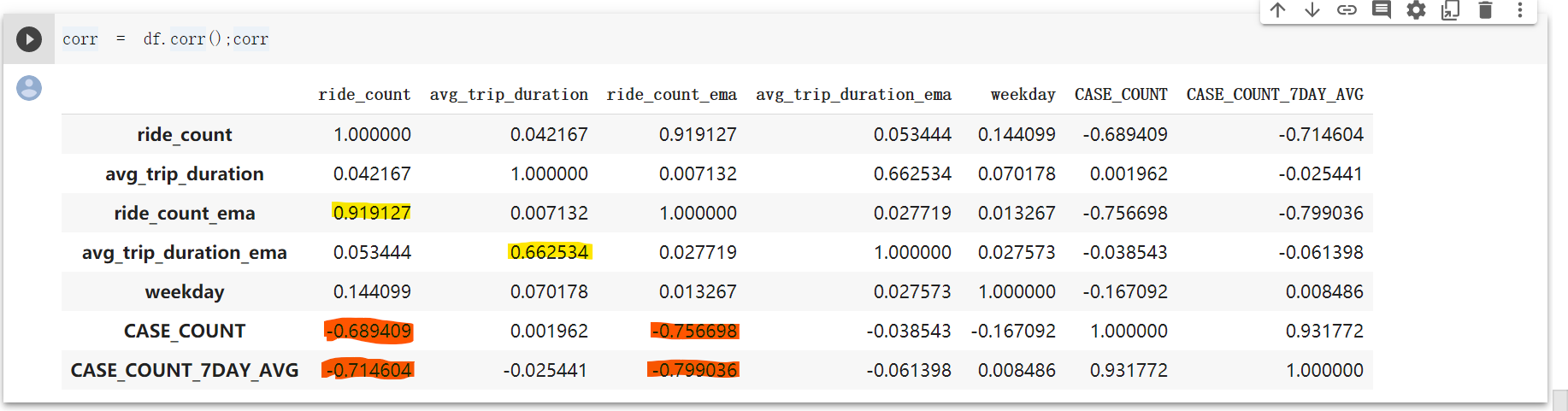


Then our group split the dataset to train data and valid data based on the date that data in 2020 are used for training and the data in 2021 are used for validation. And we will generate the table to calculate the RMS Error.

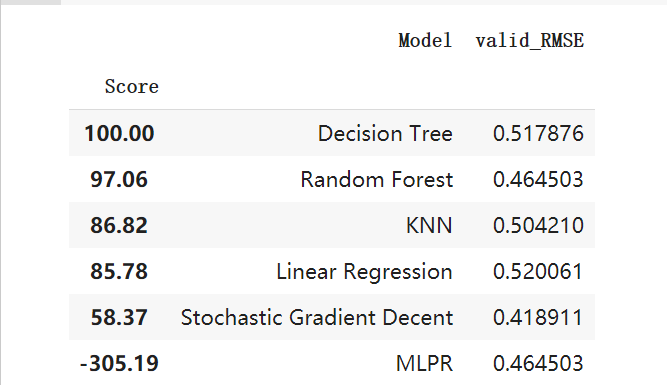
|  |  |  |
| --- | --- | --- |
| **COVID vs Daily Trip count** | | |
| **Model** | **Training Set Score (RMS)** | **Validation Set Score (RMS)** |
| 1 hidden layer 2 nodes | 0.06979 | 0.05732 |
| 1 hidden layer 5 nodes | 0.07118 | 0.06687 |
| 5 hidden layer 2 nodes | 0.06880 | 0.05070 |
| **5 hidden layer 5 nodes** | **0.06866** | **0.04348** |
| 100 hidden layer 2 nodes | 0.07676 | 0.04966 |

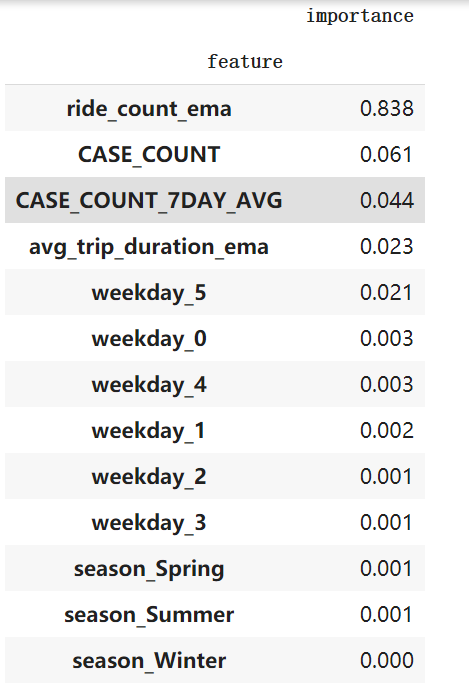
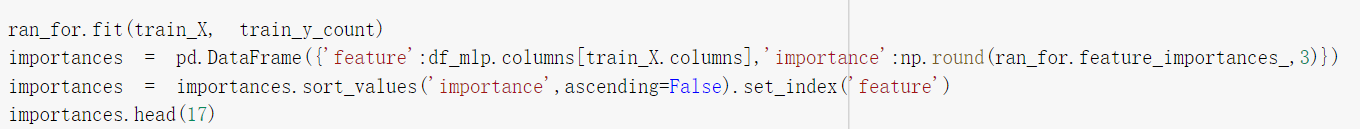
|  |  |  |
| --- | --- | --- |
| **COVID vs AVG duration** | | |
| **Model** | **Training Set Score (RMS)** | **Validation Set Score (RMS)** |
| 1 hidden layer 2 nodes | 0.01634 | 0.12772 |
| 1 hidden layer 5 nodes | 0.01703 | 0.12867 |
| 5 hidden layer 2 nodes | 0.01990 | 0.12772 |
| **5 hidden layer 5 nodes** | **0.01905** | **0.12766** |
| 100 hidden layer 2 nodes | 0.02044 | 0.12933 |

As we mentioned before, our group wants to find which condition has the smallest RMSE in validation datasets, because the role for training datasets is only for training the model. Based on the data above, **0.04348** is the smallest RMSE in “COVID vs Trip count”, and **0.12766** is the smallest RMSE in “COVID vs AVG trip duration”. So our group believes that we need to choose the MLP structure that contains 5 hidden layers with 5 nodes for both research.



The charts above are correlation table and heatmap. Our group tries to find the relationship between each variable. The two main variables that our group tried to look at are “ride\_count” and “avg\_trip\_duration”. “ride\_count” has a very strong positive relationship with its seven day exponential average. So if we want to forecast the demand, it should be a good feature to pay more attention to. This factor is important because it is easy to interpret, and it is just like people use technical analysis to predict stock price by EMA. “ ride\_count” has a negative relationship with “CASE\_COUNT” and “CASE\_COUNT\_7DAY\_AVG”. That means people prefer to stay at home than go outside during the pandemic. In this way, as the recovery begins, more and more people will go outside and “ride\_count” would increase.

The disappointing issue is that there is no obvious relationship between trip duration and other factors except “avg\_trip\_duration\_ema”. But of course they should have positive correlation. In this way, other variables do not affect average trip duration a lot, our group will not pay more attention to them in this dimension. 

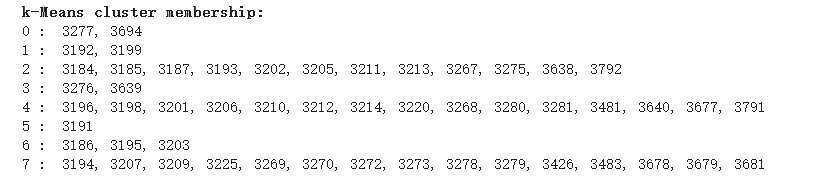
Based on the result above, our group wants to do the model selection. We decided to choose “Random forest” because it has a higher score and relatively lower RMSE. In this way, “Random Forest” is appropriate for us to calculate the feature importance.

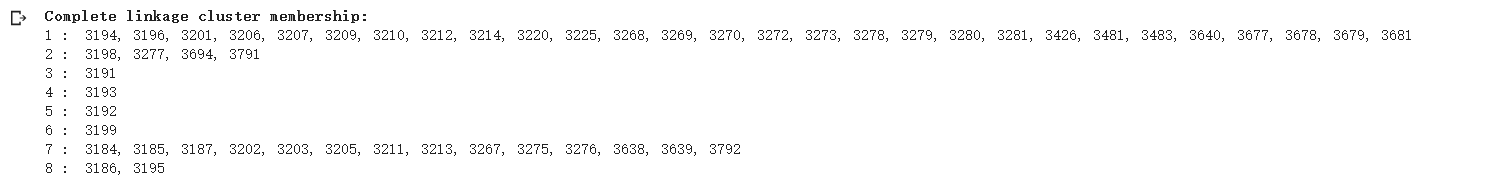
The chart above shows the score of feature importance. Just like the result from the correlation table. “ride\_count\_ema” is the most important feature. “CASE\_COUNT” is relatively important. Other features are not that important. So those two features will help us to forecast demand on bike-sharing programs. One interesting feature is season, almost all features about season are 0, which means season does not influence the bike-sharing program a lot based on the existing dataset.

## Unsupervised Learning

For unsupervised learning, the data is structured by start station as rows, and ride counts as well as average trip duration for each day for the interest (365\*2 columns).

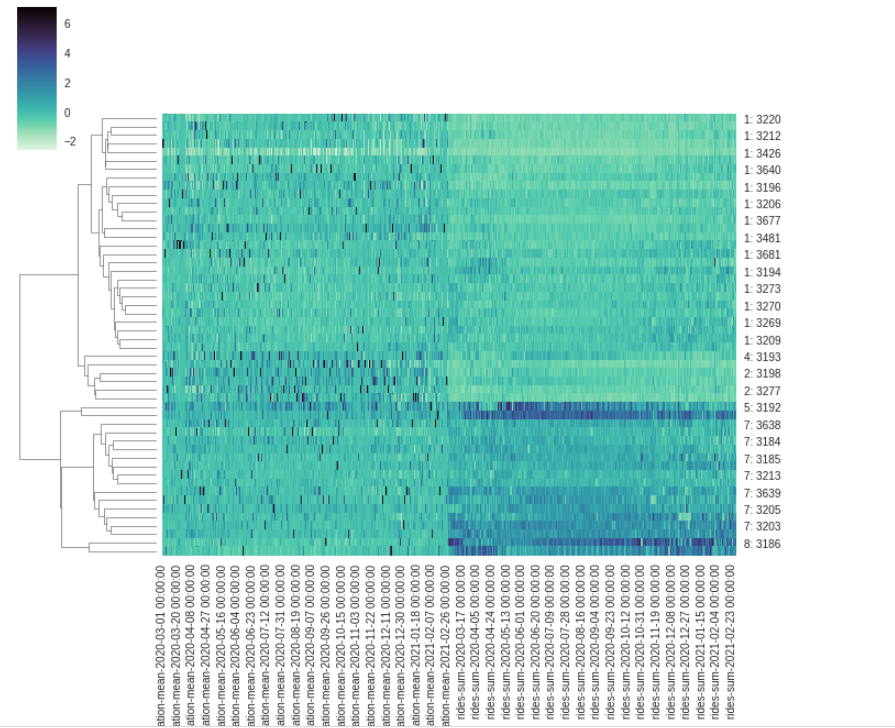
Elbow method suggested that the optimal number of clusters is 8, therefore 8 clusters are initialized for K-means, and the results are as follows (each of the 4-digit number represents the id for a station:



Then hierarchy clustering is used, with complete linkage. For a matrix of this size, fastcluster is used to represent better results. 

Although both methods are initialized with 8 clusters, the results show a difference between the two. This could be a result of noises on each column, since the ride count and average trip duration is on daily level. Most of the stations, however, are clustered the same way.

A cluster map is then plotted, with columns on the X-label, and stations grouped with complete linkage mentioned above.



Since the cluster map is grouped by clusters, it is clear to see the intuition of complete linkage clusters. Group 5 and 8 are heavily affected by the ride counts across the entire time, while group 7 is also affected, but not as significantly. Group 2 is affected by the average trip duration across the entire time period of interest. (more bikes on more trip duration, more bike moving frequency on more ride count.

# 

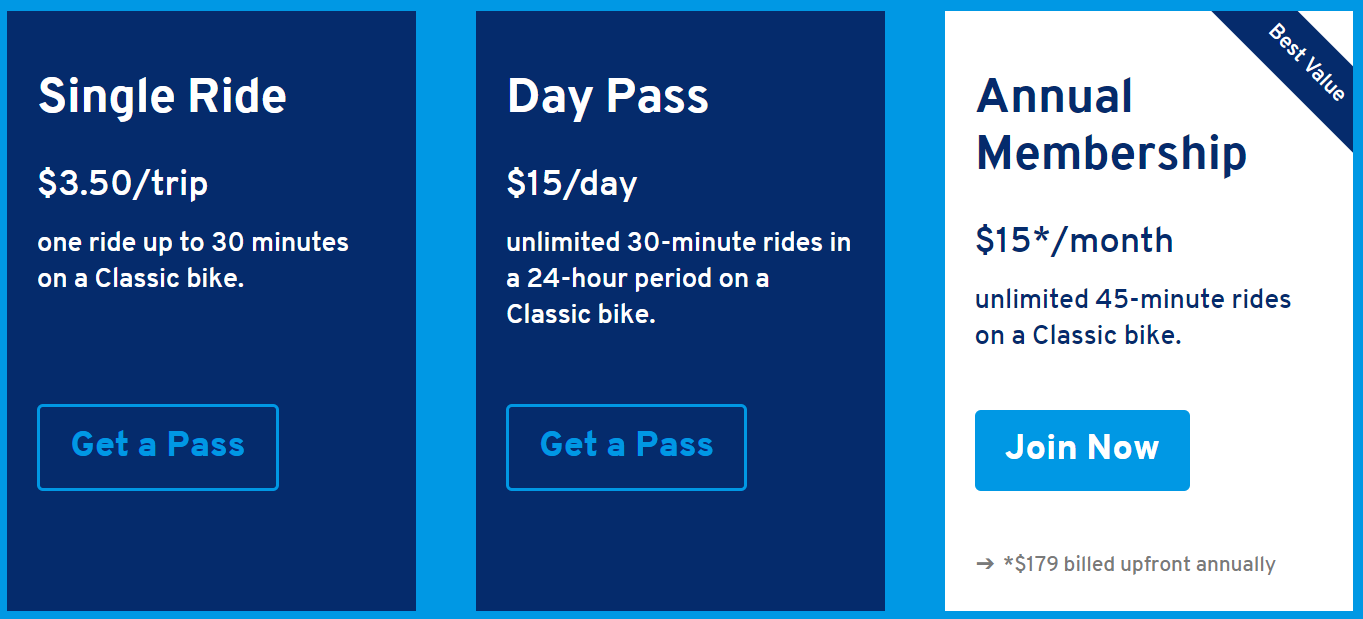
# Insights & recommendations

## Bike Availability

The management team of the Citibike network can use the forecasting model we prototyped to forecast the demand for bikes at each station the next day. Based on demand at each station, the available bikes can be distributed such that rider demand can be not only met, but exceeded, to allow for some margin of safety. In addition, based on the over and under-supply of bikes at each station at the end of the day, a route-planning exercise can be done to make sure bikes can be re-distributed as efficiently as possible.

The redistribution of bikes and quantity tuning of the bikes specific to each station could be derived from the findings from the unsupervised learning model. Based on the clustermap of plotting complete linkage clusters with respect to the ride counts and average trip duration, the Group 5 ,7 and 8 are more likely to be affected by the number of ride counts than other groups. These groups could be identified as stations at busy locations, where people constantly take bikes away. In order to achieve bike quantity equilibrium in these stations, more monitoring is needed to ensure that there are no empty or full racks at all times. Furthermore, Group 2 and 4 have significantly more trip durations than other clusters, implying that people who start from this station take longer trips than other stations. This could be a result from a station beside a park or monument, where the purpose of these bikes are not solely for commuting. These stations could often face a bike shortage as a result of many bikes being taken for long trips and result in empty racks. Adding the capacity for bike racks, as well as adding more bikes could be a solution to ensure that everyone can take the bike for a ride without having to wait for current bike users to return.

## Promotion



Here is the existing plan of Citi Bike. We could see the single ride is 3.5 dollars and has a 30 minutes limit. Day Pass is 15 per day and the limit is also 30 minutes. The best value for the existing plan is annual membership, 179 per year and 45 minutes limit. But the membership is not enough. Based on our analysis, as the recovery begins, more and more people will go outside. Ride count will increase, we need to think about how to promote the subscription.

1. Bundling subscription: If two or more people want to subscribe together, both of them will get the promotion code, which can reduce the price of the membership fee. For example, two people get 15% off and three people get 25% off.
2. New fitness membership: In fact, not all people take bike-sharing programs as transportation. Some people believe that riding is a good exercise to train themselves. In this way, Citi Bike could provide fitness membership. If the fitness subscriber achieved their mission, e.g. 200 minutes of riding per week, they will get a cash bonus back. In this way, win-win strategies materialize: the company gets more subscribers and subscribers get more benefits. This membership also encourages customers to do exercise everyday.

# 

# Conclusion

The COVID-19 pandemic has had a huge impact on people's lifestyle and transportation. Compared with the general decline in transportation, some studies and reports indicate that cycling has increased during this challenging period. This is because it does not involve proximity to others and ensures a safe distance between people. In this report, we focused on the relationship between consumer behavior and the demand for bike sharing programs to verify whether the coronavirus is an important variable among them. Based on the changes in demand and review of future expansion plans, suggestions and forecasts are made to adjust the plan accordingly. Through data processing and merging two data sets of Citi Bike Trip Histories and the Covid-19 Cases provided by the New York City Government, we use a supervised learning model to compare a large amount of training data, adjust weights, calculate RMS errors and improve accuracy. At the same time, we performed unsupervised learning on the detailed predictions at the station level. Finally, according to the results, "ride\_count" has a strong positive correlation with the value of its 7-day cycling exponential average index, and the most important features are "CASE\_COUNT" and "CASE\_COUNT\_7DAY\_AVG". So we believe that as the epidemic improves, the number of cyclists will increase. What’s more, through the MLP prediction of case count, RMSE is also within an acceptable range, so we believe that COVID has indeed affected people’s demand for bike sharing systems. In addition, for the bike sharing system planned by New York City in the future, we suggest that association rules be used to display popular connections established by users between stations to effectively support the bike network. Also, the cost-effectiveness is controlled by piloting, testing and implementing the project in stages. The management team of the Citibike network could use our prototype model to predict the bike demand at each station so that they can effectively redistribute the number of bikes at each station. Lastly, for the promotion plan of Citi Bike, we suggest promoting the subscription of shared bikes through the two aspects of bundling subscription and new fitness membership, thereby promoting the widespread use of this ideal travel mode in cities and encouraging bike sharing to become a more effective and sustainable travel behavior.

1. <https://www.citibikenyc.com/blog/major-citi-bike-expansion-map-revealed> [↑](#footnote-ref-1)