NYC Public Transport Analysis: Studying the Relationship Between Citi Bikes and The Taxi and Subway Systems

Vaib Gadodia
Courant Institute of Mathematical Sciences
New York University
New York, New York
vaib@nyu.edu

Richa Pandey
Courant Institute of Mathematical Sciences
New York University
New York, New York
rp3261@nyu.edu

Kristin Liu
Courant Institute of Mathematical Sciences
New York University
New York, New York
jl11257@nyu.edu

***Abstract*—While there has been extensive work on the effect of bikeshare systems like New York City's Citi Bike System on areas like traffic, pollution, etc., only a little research is available that attempts to correlate and infer causal relationships between the Citi Bike system and the pre-existing public transportation options available in NYC. In this work, we aim to study the correlation and causal relationships between the Citi Bike System and Taxi and Subway usage in NYC.**

***Index Terms*—Analytic, Big Data, Bikeshare, Subway, Taxi, Transport**

I. Introduction

A. Overview

Transportation is a central aspect of human life; something that has immense effects on people's quality of life both in terms of direct effects like the ease of someone's daily commute to more indirect, longer horizon effects like climate change. Due to increasing urbanization and urban migration, governments have faced increased pressure to deploy more efficient and greener forms of public transport systems and bike share has been one of the most popular such systems.

In New York City, the popular Citi Bike System was introduced in June 2013 and since then has expanded to cover almost the entirety of Manhattan and the neighboring boroughs. At the time of writing, in December 2020, Citi Bike System operates over 15,000 bikes and over 1,000 stations. Citi Bike has been very popular. However, NYC also has its extensive subway system and the famous yellow taxis, which are an integral part of the city's mobility infrastructure.

An extensive body of research has shown the Citi Bike System has had influence on traffic, pollution, etc. and has seen accelerated adoption during the ongoing COVID-19 pandemic. However, not much work has focused on exploring and understanding the causal relationship, if any, between Citi Bikes and the subway and yellow cab system in Manhattan. With this work, we attempt to explore such a correlation and causal relationship.

B. Approach

We explore the bike-taxi and bike-subway correlation and causal effects using the count of trips taken in Manhattan by studying Citi Bike trip history, MTA Turnstile and TLC Trip Record data spanning a period of 8 years from 2012 to the end of 2020.

We preprocess the data to discard irrelevant feature columns and then we combine the three cleaned datasets on the basis of date, time and location. This ensures that for every locality in Manhattan, we have the related trip counts for all three modes of transport. We then use this combined data to create visualizations, perform correlation analysis using Pearson Correlation Coefficient and apply the Rubin Causal Model to extract any causal relationship that might exist between Citi Bikes and the Taxi system.

C. Paper Organization

This paper is organized as follows: in Section II, we describe the motivation behind this work. In Section III, we discuss related works. In Section IV, we briefly describe the data used here. In Section V, we describe our analytic, discussing preprocessing, combining data, visualizations and analysis. In Section VI, we provide details on the analysis, including correlation analysis and causal inference, as well as challenges and limitations of the analytic.

II. Motivation

Since its introduction in June 2013, Citi Bike's popularity has skyrocketed. We wanted to pursue this work to analyze the impact of Citi Bikes on the NYC mobility scene as we were interested in exploring the question of whether, due

to a superior user experience, Citi Bikes have established themselves as a true viable alternative to Taxis and Subways in a dense gridlocked urban center like NYC.

Through this work, we wanted to explore this question and we wanted to test our hypothesis that Citi Bikes have a negative impact on the usage of Taxis and Subways. We believe this work is important as the presence of this relationship between Citi Bikes and the other transportation systems has the potential of shaping the traffic future of NYC for the better.

III. Related Work

In a recent work, Teixeira et. al. [3] explored the subway operation and the bike share systems data during the ongoing COVID-19 pandemic in New York City and study the impact of bike share over urban transport systems. This study aimed to analyze the relationship between the MTA reducing ridership on major subway lines and the measures taken by the Citi Bike system to provide healthcare workers with a safe mode of transport.

They used the Citi Bike System data and the MTA Turnstile dataset and performed a time series analysis to examine the two system's daily ridership variation throughout the month of March 2020 along with the average daily trip duration in the case of Citi Bikes.

They found that in less than a month in March 2020, the subway ridership reduced by 90%, while the bike daily ridership started increasing on the day the subway ridership started to fall. However, with the declaration of a state of emergency in the city, the Citi Bike daily ridership also began to fall but not as drastically as the subway system. The study also found that a continued growth on the average trip duration of Citi Bike system occurred from an average of 13-minute daily average at the beginning of March to a 19-minute daily average by the end of the month.

They further concluded that the Citi Bike system proved to be more resilient to the COVID-19 pandemic than the subway system and compelling evidence was found of a transfer of ridership from subway to Citi Bike.

In a related work, Sobolevsky et. al. [2] discuss the changes in the landscape of urban mobility in New York City with the deployment of Citi Bikes. They quantitatively assess the impact of bike sharing on urban transportation, as well as associated economic, social and environmental implications. Their idea to perform this assessment is to largely benefit the Urban stakeholders who are considering a similar deployment. While the Citi Bike usage data is publicly available, they discuss that the main challenge of such an assessment is to provide an adequate baseline scenario of what would have happened in the city without the Citi Bike system.

The paper offers a balanced baseline scenario based on a transportation choice model to describe projected customer behavior in the absence of the Citi Bike system. The model also acknowledges the fact that Citi Bike might be used for recreational purposes and, therefore, not all the trips would have been actually performed, if Citi Bike would not be available. They have trained the model using open Citi Bike and other urban transportation data and it is applied to assess direct benefits of Citi Bike trips for the end users, as well as for urban stakeholders across different boroughs of New York City and the nearby Jersey City.

They have analysed the impact of Citi Bike deployment for the two key deployment phases: July 2013 in Manhattan and Brooklyn and September 2015 in Queens, Jersey City, and additional areas of Manhattan and Brooklyn. They constructed the baseline transportation model to describe transportation modes which would have been likely used to facilitate the given amount of Citi Bike ridership if Citi Bike was not available.

They conclude that the overall assessment of the Citi Bike as an urban innovation is positive. It turns out to be particularly beneficial for the end users, with an overall benefit-cost ratio from 3.33 to 8.27 depending on the subscription rate discount.

In a related work, Ting Ma et. al. [1] examined the impacts of the Capital Bikeshare (CaBi) program on Metrorail's ridership in Washington, D.C. as the case study of the question of how and to what extent bicycle sharing programs affect transit ridership.

Two sets of analysis are conducted: an Origin-Destination analysis to map quarterly CaBi trips; a regression analysis to estimate the effects of CaBi trips on transit ridership controlling for other variables.

When CaBi trips were mapped, it was observed that Metrorail stations had been important origins and destinations for CaBi trips. Six of seven CaBi stations producing more than 500 trips were located close to Metrorail stations.

This study conducted a regression analysis and found that public transit rider-ship was positively associated with CaBi ridership at the station level. A 10% increase in annual CaBi ridership contributed to a 2.8% increase in average daily Metrorail ridership.

They concluded that CaBi's impact on transit ridership demonstrates that bicycle sharing could be a complement to transit ridership, which could positively facilitate the rate of transit ridership. Therefore, they suggest that adding bicycle sharing program to transit station areas to provide convenience to transit riders so as to promote the usage of transit.

These works are very interesting and shed some light on the role that Citi Bikes could be playing in the NYC transportation landscape. However, our work attempts to understand the correlation and causal effects of Citi Bikes on Subways and Taxis.

IV. Datasets

The bike data we used was the trip history data obtained from the Citi Bike System at the following link: https://www.citibikenyc.com/system-data The dataset came in at about 56 GB and was collected once.

This dataset had records from the introduction of Citi bikes in New York City in June 2013 to September 2020. The data consisted of trip duration (seconds), start time and date, stop time and date, start station name, end station name, station ID, start/end station latitude and longitude, bike ID, user type

(customer = 24 hour pass or 3 day pass user; subscriber = Annual Member), gender (0 = unknown; 1 = male; 2 = female), rider's year of birth. After preprocessing, the data schema was the following:

- Year: Int - The year when the bike trip was initiated.
- Month: Int - The month when the bike trip was initiated.
- Day: Int - The day when the bike trip was initiated.
- Hour: Int - The hour when the bike trip was initiated.
- Station ID: Int - The ID of the start station.
- Start Latitude: Double - The latitude of the start station.
- Start Longitude: Double - The longitude of the start station.
- Starting Grid ID: String - The cell ID of the custom grid created over the Manhattan map where the starting bike station was located.
- Hour Bin: Int - The four hour period when the bike trip was initiated.
- Subway Zone: Int - The numerical code identifying which of the three pre-selected neighborhoods the trip was started in. This was used for subway analysis.

The second data that we used was the Yellow Taxi data made available by the NYC TLC. This data, which came in at about 180 GB was collected once from the following link: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

It had taxi trip records from 2012 to 2019. The data consisted of pickup datetime, dropoff datetime, trip distance, pickup longitude, pickup latitude, dropoff longitude, dropoff latitude, passenger count, taxi zone, fare amount etc. After cleaning, the data schema here was as follows:

- Year: Int - The year of the taxi trip (ranging from 2012 to 2019).
- Month: Int - The month of the taxi trip (ranging from 1 to 12).
- Day: Int - The day of the taxi trip (ranging from 1 to 31).
- Hour: Int - The pickup hour of the taxi trip (ranging from 0 to 23).
- Latitude: Double - The pickup latitude of the taxi trip (present for trips taken before July 2016).
- Longitude: Double - The pickup longitude of the taxi trip (present for trips taken before July 2016).
- Taxi zone: Int - The pickup zone (taxi zone) of the taxi trip (present for trips taken after July 2016).

The subway data that we used was the Subway turnstile data and Station Locations data acquired from New York MTA. These two datasets came in at about 12GB and were collected once from the following links: http://web.mta.info/developers/turnstile.html and http://web.mta.info/developers/developer-data-terms.html#data

The Subway turnstile dataset included the subway turnstile records from May 2010 to Nov 2020. It consisted of control area of a station, station unit, subunit channel position(presents turnstile number), station name, linename, division(represents the operating company of a station), date, time, description of a turnstile status (REGULAR, RECOVR AUD, OPEN), cumulative entry and exit register values for a turnstile in every 4 hours.

The Station Locations dataset consisted of divison, station name, borough, linename, station latitude, station longitude etc. After profiling and cleaning, the data schema was shown as below:

- Control Area: String - The control area number.
- Remote Unit: String - The unit number for a station.
- Subunit Channel Position: String - The turnstile device number.
- Station: String - Station name.
- Year: Int - The year of cumulative subway entry/exit registered.
- Month: Int - The month of cumulative subway entry/exit registered.
- Day: Int - The day of cumulative subway entry/exit registered.
- Hour: Int - The hour of cumulative subway entry/exit registered.
- Entries: Int - cumulative subway entry register value for a turnstile in every 4 hours.
- Exits: Int - cumulative subway exit register value for a turnstile in every 4 hours.
- Latitude: Double - The latitude of the subway station.
- Longitude: Double - The longitude of the subway station.
- Grid ID: String - The cell ID of the custom grid created over the Manhattan map where the subway station was located.
- Subway Zone: Int - The numerical code identifying which of the three pre-selected neighborhoods the subway located at.

V. Description of Analytic

A. Preprocessing

We preprocessed the bike data using a Hadoop MapReduce pipeline. We filtered out feature columns that were deemed to be non-essential based on domain knowledge. After that the remaining columns were further processed and that processing is described below.

- starttime: The values here indicated the date and time when the bike trip was initiated. The values included the date - in YYYY-MM-DD format - concatenated with time - in HH:MM:SS format - with a space. We further processed this column and split it into five different features indicating year, month, day, hour (based on a 24-hout clock starting at 0 upto 23) and hour bin (based on a 24-hour clock starting at 0 at 00:00 hours and ending at 20 hours as the measurements were classified in 4 hour bins in order to be able to work with the MTA data). Any rows with illegal date/time values were discarded.
- start station latitude: This column had the latitude of the starting bike station. Any rows with illegal values were discarded and a bounding box was created around Manhattan to get the min/max latitudes to generate a criterion for filtering out illegal latitudes. Any latitudes less than

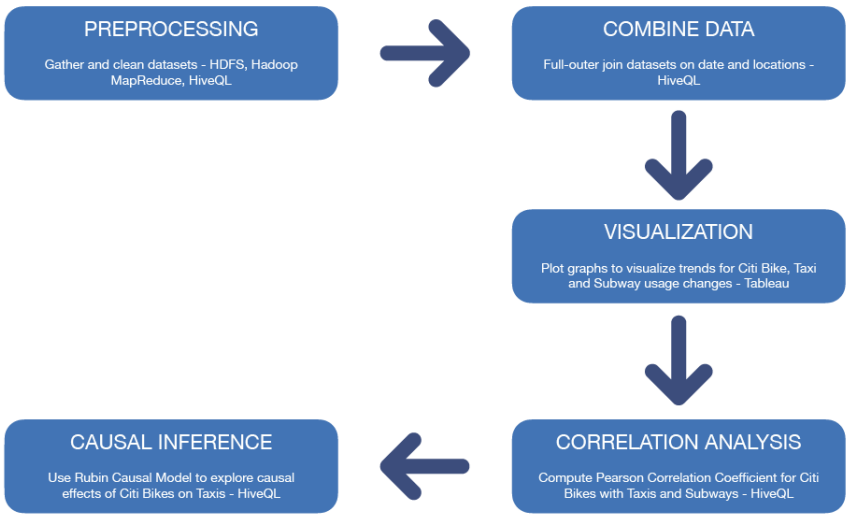

Fig. 1: Analytic Design

40.69715 or greater than 40.861752 were discarded. This column was used to generate the row for the grid ID.

- start station longitude: This feature column had the longitude of the starting bike station. Any rows with illegal values were discarded and a filtering criterion similar to the one for latitudes was used to filter out illegal entries. Any entries less than -74.022208 or greater than -73.924361 were discarded. This column was used to generate the column for the grid ID.
- birth year: This was the rider's birth year. We filtered this column out to remove any illegal value and to limit the data where the riders were born between 1943 and 1998. This was done to avoid getting bad data with very old or very young riders.
- gender: This column indicated the rider's gender and no filter was applied here, except one to discard any illegal values, i.e. not 0, 1, or 2.

We preprocessed the taxi data using Hive. We filtered out all feature columns that were non-essential. We also filtered out taxi trip records which had trip distance higher than 5 miles. We further filtered out trips taken outside of Manhattan by filtering on the pickup latitude and longitude. As a consequence of this filtering, our data was left with four feature columns that were further processed and are described below. All the trips taken before July 2016 have pickup latitudes and pickup longitudes, while the trips started after July 2016 don't have latitudes and longitudes. They instead have pickup taxi zones.

- PICKUP_DATETIME: The values here indicated the date and time when the taxi trip was started. The values included the data - in YYYY-MM-DD format - concatenated with time - in HH:MM:SS format. A small set of records had different formats which were cleaned accordingly. We further processed this column and split it into four different features indicating year, month, day and hour. We further grouped the hours into groups of 6 hours.
- PICKUP_LATITUDE: This feature column had pickup latitude of taxi trips. Any rows with illegal values were discarded and a bounding box was created around Manhattan to get the min/max latitudes to generate a criterion for filtering out illegal latitudes. Any latitudes less than 40.69715 or greater than 40.861752 were discarded.
- PICKUP_LONGITUDE: This feature column had pickup longitude of taxi trips. Any rows with illegal values were discarded and a filtering criterion similar to the one for latitudes was used to filter out illegal entries Any longitudes less than -74.022208 or greater than -73.924361 were discarded
- PICKUP_TAXI_ZONE: This feature column had NYC taxi zones where the taxi trip started. Any taxi zones that lied outside Manhattan were discarded.

After filtering out irrelevant feature columns, the taxi trip data was further processed and the observations with irregular data or missing entries were discarded.

For the subway data, we used Hadoop MapReduce and Hive to preprocess. We filtered out all feature columns that were non-essential. We formatted the turnstile data prior to Oct 2014 because the number of sets (including date, time, description of turnstile status, cumulative entries, cumulative exits) varies in each line (up to 5 sets per line), and combined the data prior to Oct 2014 and the data after Oct 2014 into one data. And we mapped the two datasets by using linename, division and station name fields to obtain subway station coordinates. The remaining columns were profiled and cleaned and are described as follows.

- date: The values here indicated the date when cumulative subway entries/exits were registered. The value was in MM-DD-YY format for the data prior to Oct 2014, and in MM/DD/YYYY format for the data after Oct 2014. We processed this column and split it into three different

features indicating year, month and day, and formatted year feature of the data prior to Oct 2014 from YY to YYYY. Any rows with illegal date value were discarded.

- time: The values here indicated the time when cumulative subway entries/exits were registered. The values included the time in HH:MM:SS format. We further processed this column and split it into one feature indicating hour. Any rows with illegal time values were discarded.
- cumulative entries: The values here indicated cumulative subway entry register values (we call it cumulative subway entries after that) for a turnstile in every 4 hours. As we needed net subway entry register values (we call it subway entries after that) to do our further analysis, we added one column using LAG() function in Hive and ordered by control area, remote unit, subunit channel position, year, month, day, hour so as to get the previous 4 hours period cumulative subway entries in each line, and then used the cumulative subway entries minus the previous one to get net subway entries in every 4 hours. Any rows with illegal subway entries were discarded and maximum subway entries value was set to 14400 (which represented that 3,600 persons were passing through a turnstile per hour or 60 persons per minute) to exclude the errant calculations of the turnstile devices. Any subway entries larger than 14400 were discarded.
- cumulative exits: The values here indicated cumulative subway exit register values (we call it cumulative subway exits after that) for a turnstile in every 4 hours. We used a same way as we did for subway entries to obtain net subway exit register values (we call it subway exits after that) to help our further analysis. Any rows with illegal subway exits were discarded and a filtering criterion similar to the one for subway entries was used to filter out illegal subway exits. Any subway exits larger than 14400 were discarded.
- latitude: This column had the latitude of the subway station. Any rows with illegal values were discarded and a bounding box was created around Manhattan to get the min/max latitudes to generate a criterion for filtering out illegal latitudes. Any latitudes less than 40.69715 or greater than 40.861752 were discarded. This column was used to generate the row for the grid ID and subway zone.
- longitude: This feature column had the longitude of the subway station. Any rows with illegal values were discarded and a filtering criterion similar to the one for latitudes was used to filter out illegal longitude. Any longitude less than -74.022208 or greater than -73.924361 were discarded. This column was used to generate the column for the grid ID and subway zone.

B. Combining Data

In order to compare the bike, taxi and subway counts, we joined the individual Hive tables of each dataset into different purpose driven tables using HiveQL. We joined the bike and taxi tables on year, month, and gridID. This generated a combined dataset for our analysis. We also formed a dataset of bike and subway by joining a table on year, month, subway zone of bike and subway data for analyzing the relationship between these two.

C. Visualizations

We build some visualizations regarding the usage levels of the three modes of transportation. We used Tableau for this.

All visualizations can be found in the Appendix.

In Fig 2, we can see the change in the Citi bike usage since its introduction in June 2013. We can see the monthly and hourly trends for the period of 7 years. We see the usage has increased significantly over these years while the pattern has remained the same. We can also observe that bike usage is more in warmer months as compared to cooler months, which makes sense as bikes are preferable in a warm weather. The hourly trend shows that there has been more usage during the day than the night time. In Fig 3, we have similar visualisations for Taxis. We see the usage has decreased significantly over the period of 8 years. Similar to bikes, we can see that the pattern has remained the same over the years. We can also observe the pattern and notice that the usage has been more in cooler months which is opposite of what we observed with bikes. In Fig 4, we have similar visualisations for the subways. We can see that the subways have mostly remained stable over the 8 years. The hourly trend shows us that the subway usage has been very consistent over the years. These are some interesting insights which helped us in our further analysis.

As we can see in Fig 4, the subway usage of whole Manhattan area seems constant during these years. We would like to know whether the trend would remain the same or show different in some particular areas — with low density subway stations, because in our common sense, the transportation of these regions are inconvenient and people may tend to use citi bike as an alternative to go to other places or take citi bike as a complement to help them get to other places from subway stations. We pick the east and west side of Manhattan because these regions have low subway availability. We divide them into three neighborhoods, middle to lower east Manhattan, lower to middle west Manhattan and middle to upper west Manhattan (since subway stations did not open till Jan 2017 in the upper east Manhattan (2nd AVE and east of that area), we did not pick that area; and since two important subway stations located in the middle east Manhattan region shutdown/partially shutdown in the past 3-4 years, we also did not choose that zone) and visualize yearly subway entries/exits and bike trips to study both substitute and complementary relationship between bikes and subway.

a) Substitute Case: We visualize the yearly subway entries and bike trips to see whether they have a substitute connection. In Fig 5 (1), we could see that subway usage in these three selected areas has been constant while bike usage has grown year by year. In Fig 5 (2), Fig 6 (1) and Fig 6 (2), we could find out that subway usage in each area has been almost stable while at the same time bike usage has increased consistently. Therefore, as we could discover in these four graphs, bike does not seem to act as a substitute to subway.

b) Complement Case: We visualize the yearly subway exits and bike trips to find out whether they have a complementary relation. First, we see all three regions as a whole. From Fig 7 (1), we could see that subway usage pattern has remained the same while bike usage has gone up year wise. Then we obverse these selected areas separately. As we can see from Fig 7 (2), subway usage has also displayed a similar pattern of no significant growth while bike usage has risen yearly. In Fig 8 (1), we could discover that the usage of subway has grown at a minimal rate while, in the meantime, the usage of bike has increased per year. From Fig 8 (2), we could notice that subway usage has dropped minimally but the trend remained almost stable, meanwhile, bike usage has kept increasing each year. So, as all these four graphs show, subway usage seems to have not been influenced much by bike usage as a complement.

D. Pearson Correlation Coefficient

Pearson Correlation Coefficient, also known as Pearson's r is a statistic that measures the linear correlation between two variables and we use it to quantify the linear relationship between the Citi Bike and Taxi, and Citi Bike and Subway datasets.

Pearson's r can be easily calculated using the following formula

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

Here, x and y are the two variables of interest. In our case, we took the total bike trips in a year as x and the total taxi trips or the total subway entries/exits in a year as y. $\bar{x}$ and $\bar{y}$ were the averages of the two measures.

This coefficient varies in value from -1 to $+1$, with -1 being an indicator of perfect negative correlation and $+1$ indicating a perfect positive correlation. A coefficient value of 0 shows the absence of any correlation.

We used Apache Hive for this computation and since, in 2013 there were no bikes in the first 5 months of the year, we decided not to include that year in our calculations.

E. Rubin Causal Model

In order to quantify the causal relationship between Citi Bike and Taxi data, we used the Rubin Causal Model to perform causal inference.

The Rubin Model is an approach to the statistical analysis of cause and effect based on the framework of potential outcomes. It is a widely used method for causal inference where the central idea is that since, we can never directly observe causal effects at a unit level as any one unit can never both be exposed to a treatment and not, we use randomized experiments to observe population level causality effects, based on the assumption that the population is more or less homogeneous. Here, once a random assignment of treatment and control units, i. e., units with and without bikes respectively are obtained, a difference between the means of those two cohorts is used as a measure of the causal effects.

In order to apply the Rubin Model to our datasets, we started by dividing the Manhattan data into 30 randomly selected zones of around 5X5 blocks where half of them had Citi Bikes and the other half did not. This gave us a treatment and a control set to compute causal effects. Once we had these sets, we computed the average changes in Taxi usage over a two year period from 2013 to 2015 for each of them. Finally, we computed the difference between these averages to get the causal effect of interest.

We used Apache Hive for this analysis.

VI. Analysis

A. Bike and Taxi

a) Pearson Correlation Coefficient: Given the combined Bike and Taxi datasets, we computed the Pearson Correlation Coefficient for total yearly trips in both types of transport solutions for the years 2014 to 2019.

We obtained a Pearson's r value of $\mathbf{-0.975}$, which is a very strong indicator of high negative correlation between Citi Bikes and Taxis. This implies that with the increase in Citi Bike usage, Taxi usage tends to decrease. This result fits in well with our observations in the visualization stage of the analysis and with our initial hypothesis that Bike might be acting as an alternative to Taxis.

b) Rubin Causal Model: With a highly negative Pearson's r, we observed that the Citi Bike and Taxi datasets had very high negative correlation. However, correlation does not automatically imply causation. Therefore, we applied the Rubin Causal Model to extract the causal effects of Citi Bikes on Taxi usage.

For 30 randomly selected 5X5 block zones in Manhattan, where half of them had Citi Bike stations and the other half did not, we created a treatment and a control group and performed causal inference for the two separate cases of relatively warmer and cooler months.

In the analysis for warmer months, we included data from April to November, keeping the other months in the cooler months cohort.

The results of the analysis can be seen in **TABLE 1**. We observe that the causal inference results in negative values, which suggests the presence of causal effects of Bikes on Taxi usage. This is so because these values are the difference in the percentage change in taxi usage in areas with Citi Bikes and areas without. Since, this is a negative value that means that the areas with Citi Bikes have seen a larger taxi usage decline than the other areas when everything else except the presence of Citi Bikes was the same. This is a strong indication of causal effects.

Furthermore, we also observe that the causal effect for warmer months is large (more negative) than for cooler months. This also falls in line with our initial hypothesis as this suggests that Citi Bike usage in warmer weather is responsible for larger decline in taxi usage than in the cooler months as people are more willing to use bikes as a taxi alternative in those months.

TABLE I: Rubin Causal Model - Citi Bikes and Taxis

	Avg. % change in taxis from 2013 to 2015
Warmer Months	-13.12%
Cooler Months	-12.53%

B. Bike and Subway

a) Pearson Correlation Coefficient for substitute case: Given the joined Bike and Subway dataset, we computed the Pearson Correlation Coefficient for total yearly bike trips and total yearly subway entries from 2014 to 2019.

We obtained a Pearson's r value of $\mathbf{0.09}$, which indicates minimal to no correlation between Citi Bike usage and subway entries in the selected neighborhoods. This implies that Citi Bike seems not to have an impact on Taxi usage in substitute manner, which matches with our observations in the visualization stage of the analysis that bike does not seem to act as a substitute to subway.

b) Pearson Correlation Coefficient for complement case: Given the combined Bike and Subway dataset, we computed the Pearson Correlation Coefficient for total yearly bike trips and total yearly subway exits from 2014 to 2019.

We obtained a Pearson's r value of $\mathbf{0.26}$, which shows a very weak positive correlation between Citi Bike usage and subway exits in the selected areas. This could be an indication that the Citi Bikes act as a complement to the Subway system in a limited capacity. This result fits in with our observations in the visualization stage of the analysis that subway usage has not been influenced much by bike usage as a complementary mode of transport.

C. Challenges and Limitations

The datasets used for this work were very rich in the information they offered. However, as a direct consequence of that richness, they were also very large. The TLC Trip Records Data (Taxi Data) came in at about 180 GB. It took significant time for us to download this data and then upload it to HDFS. Data cleaning also took considerable amount of time. Besides the challenge posed by the size of this data, there was another challenge related to the schema. There was a change in the data fields in 2016. To accommodate for this, we had to split the data and process them in separate tables.

Furthermore, our subway data came without any information about the station location in terms of its coordinates. As a result, in order to join the three datasets and use the Subway data for analysis, we had to map each subway station to the correct location by cross-referencing the stations to an external table that mapped station names, linename, division to coordinates. This was particularly challenging because not only did station names change over the years, new stations built and old ones shut down, but also new subway lines were added in during the years, which made our task more difficult.

Finally, while our analysis provides a reliable framework for understanding the effects and the relationship between Citi Bikes and Taxi and Subway systems in NYC, due to the nature of correlations, there were challenges in interpreting the results of Pearson'r r in the case of Bike and Subway data.

VII. Conclusion

Citi Bikes, since their introduction in NYC in June 2013, have seen a tremendous rise in their popularity. At the same time, Taxi usage has declined by a large margin and Subway usage has mostly remained stable.

Our analysis shows that while there is minimal correlation between Citi Bikes and the Subway system in both substitute and complementary terms, Citi Bikes tend to exert a strong negative correlation with Taxi usage. Furthermore, through causal inference we observe that the relationship between Citi Bikes and Taxis is a causal one and the presence of Citi Bike stations in a neighborhood leads to almost 13 percentage point higher drop in taxi use in the area.

VIII. Future Work

1) More complex modelling techniques: Given the structure of our data and the trends seen through visualization, a correlation study and using the Rubin Causal Model for causal inference was the best way of modelling a relationship between Citi Bikes and the Taxi and Subway systems. However, it would be interesting to further explore this relationship between the three datasets by using more advanced modern techniques like deep learning, etc.

2) Exploring the relationship between Taxi and Subway usage: While our work focused on the effect of Citi Bikes on the other modes of public transport in NYC, it would be interesting to explore the relationship between Taxi and Subway usage to further understand the reasons behind why Taxi usage has declined so greatly.

References

[1] Ting Ma, Chao Liu, Sevgi Erdoğan, "Bicycle Sharing and Transit: Does Capital Bikeshare Affect Metrorail Ridership in Washington, D.C." Published 2015.

[2] Stanislav Sobolevsky, Ekaterina Levitskaya, Henry Chan, Marc Postle and Constantine Kontokosta, "Impact Of Bike Sharing In New York City" Published 2018.

[3] Joao Filipe Teixeira, Miguel Lopes, "The link between bike sharing and subway use during the COVID-19 pandemic: The case-study of New York's Citi Bike" Transportation Research Interdisciplinary Perspectives, Volume 6, 2020.

APPENDIX A
TRENDS

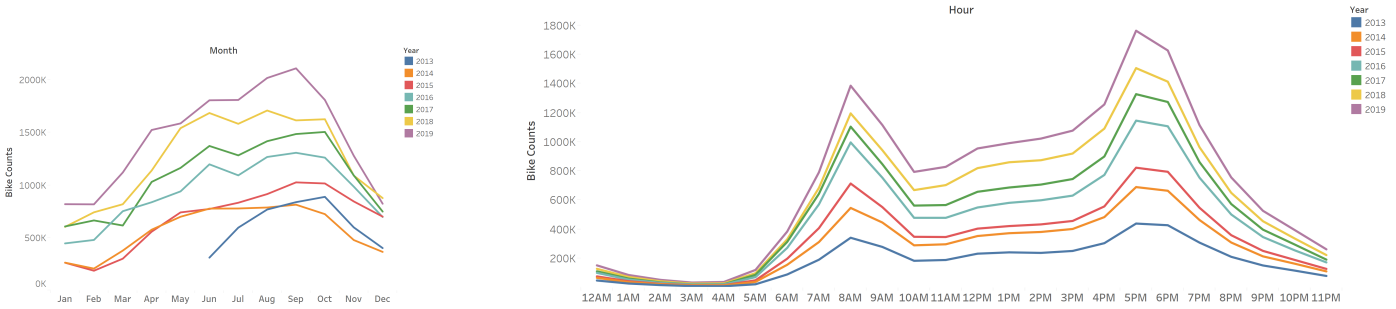

Fig. 2: Bike Trends - Monthly and Hourly over 7 years.

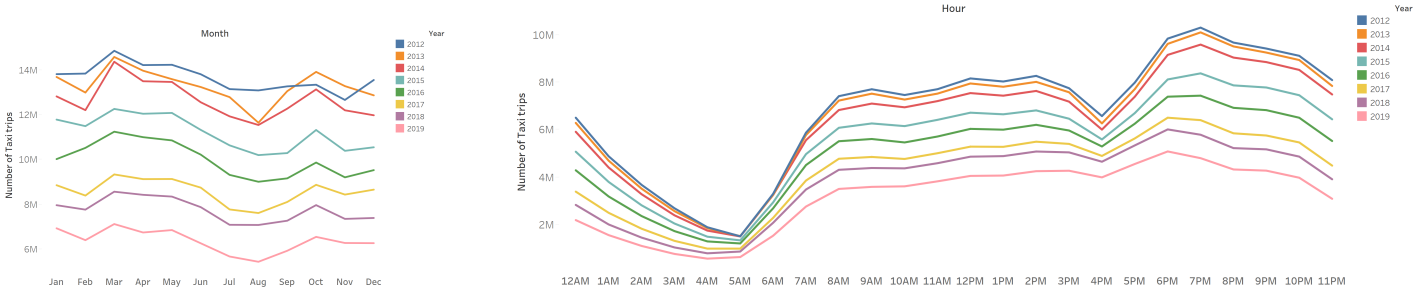

Fig. 3: Taxi Trends - Monthly and Hourly over 8 years.

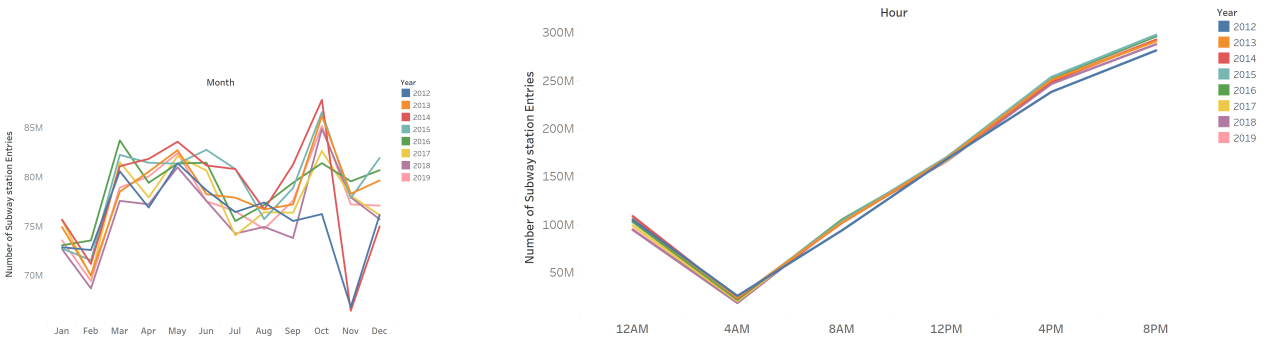

Fig. 4: Subway Trends - Monthly and Hourly over 8 years.

APPENDIX B
SUBWAYS ANALYTIC GRAPHS

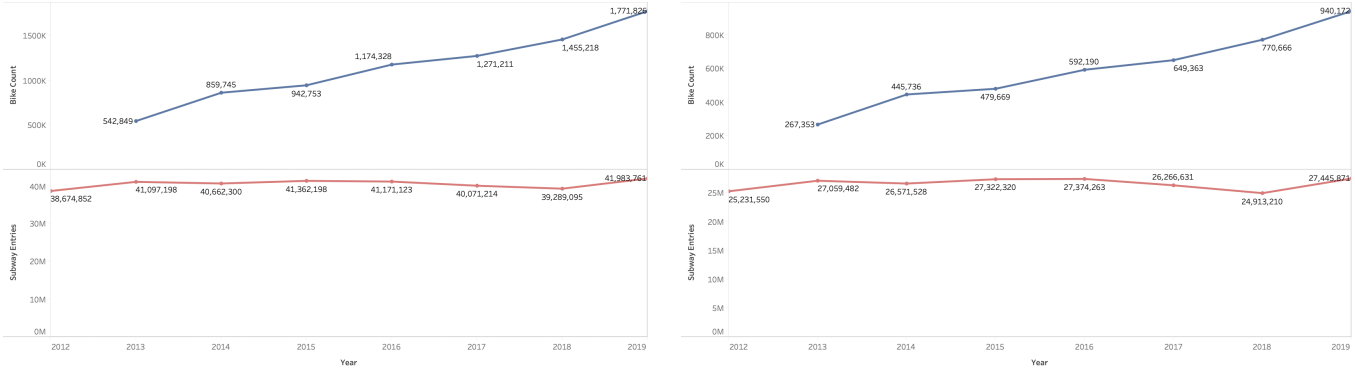

Fig. 5: (1) Substitute Case of Bike and Subway in three selected areas (left) and (2) Middle to Lower East Manhattan (right)

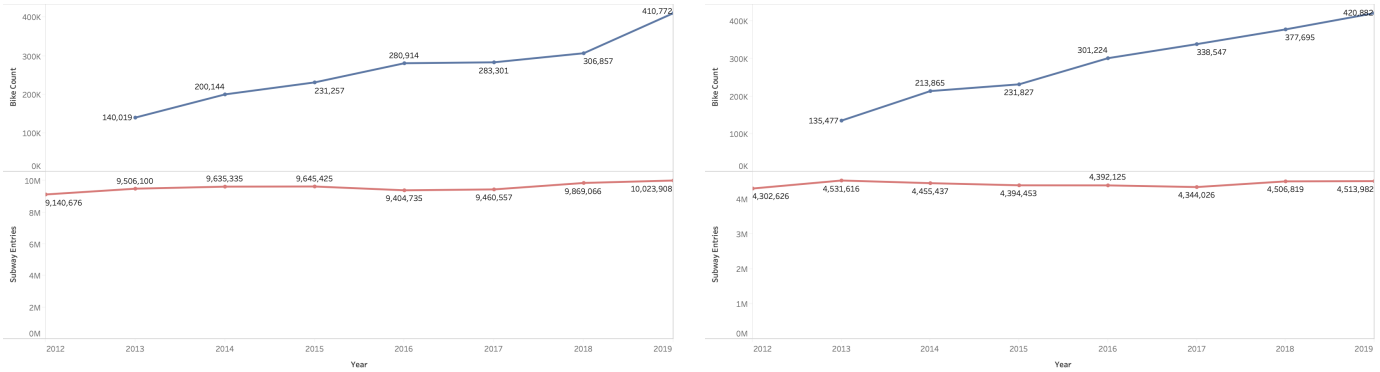

Fig. 6: (1) Substitute Case of Bike and Subway in Lower to Middle West Manhattan (left) and (2) Middle to Upper West Manhattan (right)

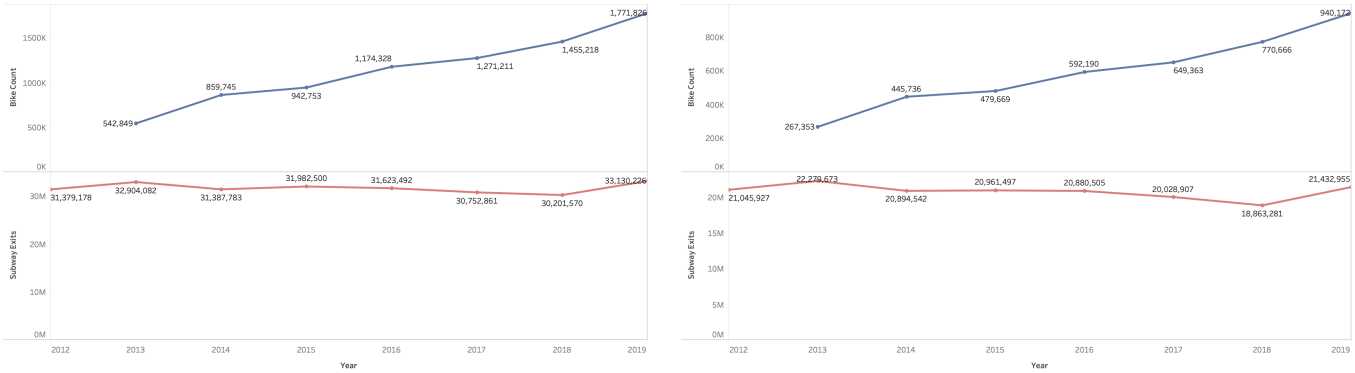

Fig. 7: (1) Complement Case of Bike and Subway in three selected areas (left) and (2) Middle to Lower East Manhattan (right)

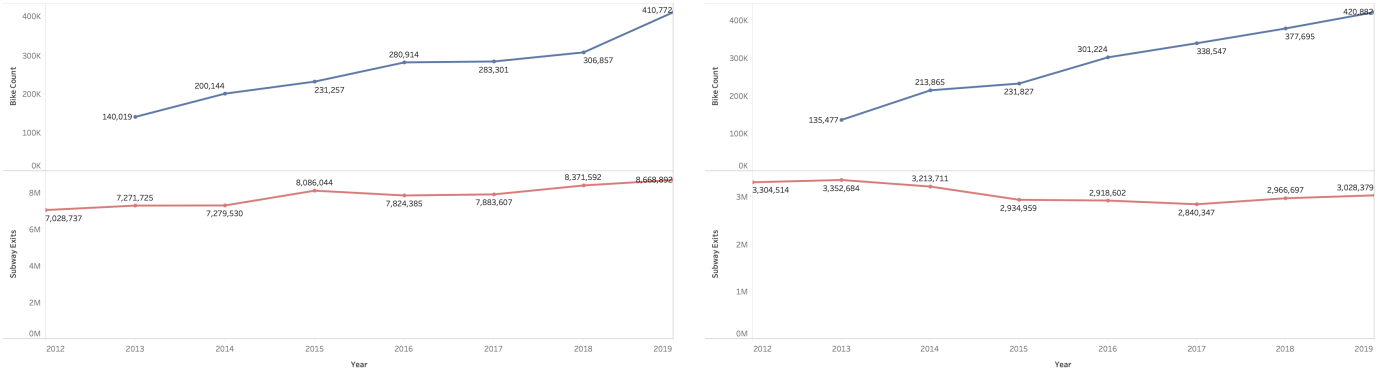

Fig. 8: (1)Complement Case of Bike and Subway in Lower to Middle West Manhattan (left) and (2) Middle to Upper West Manhattan (right)