How Safe is Bike Sharing: A Big Data Analysis?

Mohit Nihalani
New York University
New York City, USA
mn2643@nyu.edu

Hardik Bharath Rokad
New York University
New York City, USA
hbr244@nyu.edu

Rohit Saraf
New York University
New York City, USA
rs6785@nyu.edu

***Abstract*—**

With the increasing popularity of the bike sharing programs in the NYC, there has been increasing concerns regarding the safety of riders, mainly because the NYC has high traffic and also it is not mandatory for bike sharing riders to wear helmets. This paper performs the big data analysis on NYC Vision Zero, Citi Bike, and Weather dataset to analyze the pattern between the accidents and city bikes. We analyzed which areas are more prone to accidents and which are not and what is the probability of the Citi bike riders to meet an accident and is there any correlation between growth of city bike popularity and increasing bike accidents in NYC. This research can be further used to identify the reasons of high accidents in the given area compared to other and resolve safety issues, by efficient planning of bicycle lanes, regulating traffic or even posting traffic officers in the areas of high bicycle accidents.

Keywords—analytics, bike sharing, accidents, safety

I. INTRODUCTION

Bike sharing is one of the fastest growing alternative mode of transport. As of 2016 there are over 1000 cities with the bike share program with over 5 million global fleet size of bikes and over 200,000 alone in USA as of 2017. As of 2018 there are more than 48 million reported bike share trips across USA, among them largest was reported by Citi Bike NYC. The project will analyze the growth of Citi Bikes since 2015 will also investigate the travel patterns of the users, that take advantage of these shared bicycle in New York city. We will be analyzing various travel behaviors, like peak timings of travel, months, average duration of travel, frequent origin/destination and by which age group this system is dominated. Our analytic will also analyze, how the growth has been for this industry and what future trends can we expect. There has been a lot of concerns regarding the safety of these bike share programs, especially when NYC is not known for its bike friendliness and bike share operators don't provide helmets to the riders, and most of these riders are occasional users, many of them don't carry helmets with them, which can increase the risk of fatal injuries in case of accident One of the bicycle researchers in New York Times also predicted that because of increase bike share, number of injuries among cyclist and fatalities will rise [12]. However, there have been lot of research on the technological advancements in bike share systems such as rebalancing, bike share facilitators, impact quantification but the research on bike safety is scarce and research on crash risk is also rare. Even there is no proper methodology to collect bike accidents reports, which makes it difficult to distinguish accidents involving bike share riders and personal users. We don't even know out of all reported accidents, how much constitutes Citi bikes, which is important for local authority and operators to expand bike sharing and at the same keeping it safe for its users, and increase in fatalities involving bikes is demanding for improvement. To bridge the gap between safety and bike sharing programs, we analyzed the NYC Vision dataset of accidents, to understand the patterns of accidents, where bikes were involved and also map these patterns with the weather dataset. We analyzed in which areas most bicycles accidents occur and which one are the safest ones. We also analyzed if the increasing accidents are directly related to increase in bike sharing riders. To summarize study on bicycle safety is scarce and according to [14], chances of injury for bike share riders is much less compared for cycling in general, this study sets out to examine the patterns of Citi bike growth and the impact of this growth on cyclist crash risk. Based on [14] we hypothesize that increasing growth of bike share programs is not associated with higher fatalities of bikers, there are many other factors like lack of infrastructure and regulations and on contrary Citi bike are one of the safest means of transportation.

II. MOTIVATION

With the increasing concerns of environmental and health issues, bike sharing has become one of the most used alternatives for transportation. Over the past many years there is a significant growth in bike sharing programs around the world, and lot of campaigns are being run to encourage bike sharing. There has been lot of research to tackle bike sharing issues, bicycle rebalancing, impacts of bike share but the research on crash risk of bike share users is scarce. NYC Citi bike sharing program has seen nearly 60 million trips, but the surprising part is cycling has suddenly become less dangerous, and only one reported deadly accident involving Citi bike, while the data has around 60 million trips. The main aim of this paper is to analyze the Citi bike, NYC Vision and NYC weather data to find the correlation between the accidents involving bikes. We would be analyzing in which areas the most accidents occur, what are total number of trips started in these areas? How the trips are affected by weather? Which are the safest areas for renting bike? There are many unsolved questions, which we will try to answer through our analysis. This research can be further used to identify the reasons of high accidents in the given area compared to other and resolve safety issues, by efficient

planning of bicycle lanes, regulating traffic or even posting traffic officers in the areas of high bicycle accidents.

III. Related Work

Bike-sharing systems have been growing rapidly for the last 10 years in urban environments. It is one of the smart-transportation methods that provide a low-cost, environmentally friendly transportation alternative for cities. The increasing population has produced several mobility problems due to the increase in congestion and CO2 emissions and bike-sharing is one of the solutions to reduce the impact of these on society. Impacts of bike sharing on environment has been discussed in the paper [5], which have analyzed energy savings and CO2 and NOx emission in Shanghai, China in 2016. The methodology and experiments discussed in this paper, were used to transform the Citi bike data. The problems faced by them in analyzing the data were similar to us, as the Citi bike data has lot of bias, which restricts us to provide accurate results, we choose to analyze and deal with data similarly as discussed in this paper, like which fields to remove, how to deal with missing data. They calculated the distance between the stations and found the average speed, which is not provided by the dataset, we use the similar technique, but instead of Euclidean distance, we choose haversine distance to calculate distance between each station and speed of the riders.

Analysis regarding the safety of bikes were presented was performed by Elliot Fishman and Paul Schepers for the International transport forum [9]. This paper goes into depth of researching on crash risk of bike share users. Data was not only gathered from police statistics but also hospital data was considered for victims treated at emergency departments or admitted to hospital. They performed 2 studies, Study 1: is a secondary analysis of longitudinal hospital injury data study from 10 North American cities, divided into two categories; 5 cities with bike share programs and 5 cities without and study 2 examines injury risk for bike share programs based on data provided by bike share operators who were contacted for this study in two major cities Paris and London. The research paper made an hypothesis that introduction to bike share programs are associated with lower injury risk and both the studies supported that, as study 1 indicated that the introduction of bike share programs are associated with lower injury risk, while Study 2 found that bike share users are less likely than other cyclists to sustain fatal or severe injuries. The results of the paper makes sense because due to higher bicycles on road there has been noted increased driver awareness cautiousness towards cyclists, the speed of bikes on bike share programs has been capped on lower speeds, which reduces the chance of any crash, while the upright position of bike share bikes may increase the visual profile of the rider in traffic and improve their field of vision. Finally, one can also analyze the bike share riders generally rides in nearby city centers, where motor vehicles speed is lower which reduces the chances of crash, and bike share infrastructure are mostly on inner city, where there are more dedicated cycle paths.

Predicting Bike Usage for New York City's Bike Sharing System The above paper is on the same area as the project we choose, it predicts the bike demand and tells what factors influence this. They take weather, taxi usage and spatial variables as covariates to predict bike demand. As per this paper no study has used other factors to decide the bike demand. They processed raw data to get the number of bike trips between each stations during the morning rush hours they also simultaneously obtained taxi data for the same duration and to every taxi trip they assign pickup and drop-off bike station id using the taxis trip pickup and drop-off location and they only include trips where pickup and drop-off location are within quarter mile of a bike station, this is now grouped by pickup and drop-off station id to get the count of taxi trips for each station pair. they also gathered NYC precipitation data and realized that the bike trips on the weekdays is 26% more with rain less than 1mm than the on a weekday with rain greater than 1mm. They have used regression analysis to predict bike trips during morning rush hours from station to neighborhood pair. With this they were able to predict the bike usage pattern of New York City during the morning rush hour. The project we chose uses similar concept and analyze travel behavior from the data we will find out the peak timings of the bookings, and what is the duration of the average booking. What are peak days, are they on working days or on the weekends? In what areas this system is most popular, and which are the most frequent origin/destination stops? and Why are these so popular? does the number of bookings in these stops change on day to day basis which will help us to analyze if there is some rebalancing problem? What is the user frequency, means the user booked ones, rebook another time and what is the turnover rate? what age group and which gender is this system dominated? Is there any increase in membership on year to year basis? Which city has seen the huge growth in recent years and what could be the growth in future?

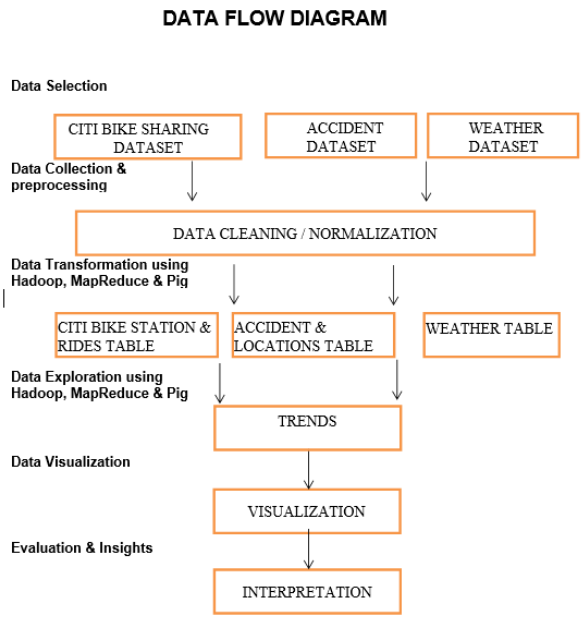
Pedestrian deaths from 1998 to 2002 described in the paper [11] were extracted from the Fatality Analysis Reporting System database of the National Highway Traffic Safety Administration. Fatal pedestrian motor vehicle crashes in New York City and the US every year, there are about 5000 pedestrian deaths in the US and about 200 in New York City due to motor vehicle crashes. People aged >65 years constituted 12% of the total population and 22% of the pedestrian deaths. In New York City, 38% of pedestrian deaths were among people aged >65 years, this group comprised the same 12% of the population. Age-specific death rates varied most between the youngest and oldest age groups. In New York City, the death rate for the oldest age group (>85 years) was 11 times higher than for the youngest age group, birth to 14 years. People aged >65 years had higher age specific death rates than those aged (65 years, nationally (3.1/100 000 v 1.5/100 000) and in New York City (7.2/100 000 v 1.5/ 100 000), older pedestrian death rates in New York City were 4.8 times higher than the younger age group. In contrast with national characteristics of motor vehicle crashes, almost half of all motor vehicle crash deaths in New York City were among pedestrians. This may be an important consideration especially in neighborhoods with large populations of elderly people. Enforcement of traffic laws, those controlling vehicle speed, can reduce pedestrian

deaths. This may be of importance for older pedestrians, who are even more susceptible to death from vehicles at greater speeds than younger pedestrians. Pedestrian safety can serve as a part of neighborhood revitalization efforts, which benefit all ages. The paper was used to understand who the major portion were involved in the motor vehicle crash. The paper suggests that the pedestrian fatalities due to crash is increasing but the data used in our analytics showed that the fatalities are increasing but the number of accidents are decreasing. These means there are many motor vehicles crashes that goes unreported.

The research paper [8], examined two bike-sharing systems of Chicago and Budapest. Analysis of approximately two million transaction data records associated with bike trips made over a three-month period was investigated. It aimed at demonstrating how big-data collected by smart-device transactions can be used to profile the usage of such systems. The usage of relevant open-source data for research purposes has been demonstrated to be a feasible strategy to gain a deeper understanding of the functioning of those Bike-Sharing System that has proved to be successful. We can use the parameters indicated by them in the research paper for our analytics. Also, the model followed by them can be furthered improved and used, which can give better results in our project.

IV. Design and Implementation

A. Design Details



We begin by collecting the different datasets needed for our analytics. We now have to ingest this data either using the curl command or WinSCP. We then clean the data and normalize it by dropping unnecessary columns and formatting data if required, profiling code. In our case we had latitude and longitude, from these we figured out what were the zip codes and used these for our analytics. We achieve this by using Hadoop MapReduce. We now ingest the cleaned data into Hive and create a table out of it. We do the same for the other datasets. We will now have tables from Citi bike dataset, Accident dataset and Weather dataset. Now we can perform our analytics and figure out the trends in the data for each of these tables, look for patterns and correlations between these table's and the tables from the other datasets. We can do that by using join operations which will join data from multiple tables, and we will have combined analytics. In our case we combined the accident dataset and Citi bike dataset using zip code to get total bike accidents near Citi-Bike station. We can find out which areas have the highest number of accidents and if they are because of a bike sharing station being present or not. We then use weather data to see if on a rainy day the number of total rides get impacted (it should, ideally), We perform many such analytics and try getting as much insight out of it as possible. We go into details of the analytics in the results section. We achieve the above by using hive queries and join operations. Once we have the trends, we can make graphs out of it as shown in the results section and visualize our trends. Finally, we now are in a position to make meaningful insights out of our analytics with the trends and visualization and draw conclusions out, make decisions based on the that.

V. Datasets

A. Citi Bike Dataset [4]

The data is made public by Washington D.C. bike sharing company Citi Bike Dataset have the historical trip data, where each trip is anonymized and includes, client id, trip start day and time, end day and time, rider type (member, single, pass), gender, and year of birth, bike number. Data has been processed to remove any trips lasting less than 60 seconds. Data used is from 2015-2019.

Schema:

- Duration
- Start/End Time
- Start/End Station
- Start/End Station Id
- Start Latitude
- End Latitude
- Bike Id
- User Type
- Gender
- Age
- Start Longitude
- End Longitude

B. Weather Dataset

This data is available at https://www.ncdc.noaa.gov/ . The Weather Dataset contains data about the daily weather conditions like date(mm/dd/yyyy), Average temperature, Dew, visibility, windspeed, max windspeed, gust speed, max temperature, min temperature, frshtt (describes the weather that day ex-fog/snow/tornado). The period that this data captures is between 01/2016 -11/2019.

Schema:

- Year (int)
- Month (int)
- Average temperature(double)
- Weather_State (String)
- Dew(double)
- Visibility (double)
- Windspeed (double).
- Max temperature (double)

C. Accident Dataset [7]

The motor vehicle collision data is provided by NYC OpenData. Each row in the data set contains a crash event. The crash events contain information like accident date, time, latitude, longitude, type of vehicle involved in the accident, and borough. Data was processed to bring date in certain format, remove irrelevant information and empty values.

Schema:

- Date
- Time
- Latitude
- Number of cyclists injured
- Type of Vehicle
- Borough
- Longitude
- Number of cyclists killed

VI. RESULTS

Bike share systems are one of the fast-growing alternative modes of transport, and one of the earlier bike share system is Citi Bike in New York City. Though it started in 2013, but the sharpest increase can be seen from 2016, which is about 15 million rides and since then it has been continuously increasing reaching 18 million till the end of October 2019. This sharp increase is due to various reasons: 1. People are become more aware of the green transportation. 2. They are choosing bikes, as they can keep them healthy and 3. Increase in infrastructure in inner city, as according to [10], 1,240 miles of dedicated bike lanes has been installed 1,240, which definitely is a biggest reason, as NYC was never exactly laid out as a bike friendly city.

Year	Total Rides (In Millions)
2015	9.9
2016	15.8
2017	16.3
2018	17.5
2019 (till October 2019)	12.5+

While analyzing the data, we found out that there was lot of anomalies , as there were some rides with duration of 0 seconds and while there were some with 1900 hours, this could be when the bikes are not docked properly so to remove any bias we dropped all the rides data which had a duration of more than 2 hour. After removing these trips, we found out that mean duration of the rides was 20 mins, which clearly shows that bikes are often used for last mile transportation.

One of the biggest challenges was to calculate average distance travelled by users, as there is no reliable measure to calculate bike routes since we can't predict what route was taken by which rider. We could have used google maps API, using start and end latitude and longitude, but this would have required lot more API calls then the daily limit, so we decided to calculate distance using Haversine Formula.

After calculating distances, we found an anomaly as distance travelled, ranges from 0 to 8675 miles, which is clearly not possible, so we decided it is safe to remove these trips also and keep the trips which are in between 0.2 miles to 20 miles. After removing these trips, average distance travelled was 2 miles, which is close to what we predicted, as it clearly shows that people use bicycles for very short distances especially for last mile commute.

Based on the distance travelled average speed is close to 11.2 mph, which matches with the analysis [6].

Citi Bike riders birth year data is not accurate, as they don't have to provide any sort of identification for riding Citi bikes, because of that most of them provide wrong birth year, which can be clearly seen after analyzing the data as age ranges from -25475 to 162, which is clearly not possible and makes impossible for us to analyze impact of Citi bikes based on age groups. So, for these analytics we are not considering age groups. But after removing these ages and keeping the age between 5 to 80, average age of Citi bike users was around 40. There is about 10% of data which doesn't lie in these ranges, so we decided that it is safe to replace all the ages outside these range with 40.

To get much better results with the accidents database, we have to map, all the stations with their zip code. Which is one of the biggest hurdles, because if we map it into the Hadoop, we will be making unnecessary API calls, as there are about 25,822,188 trips, and only 846 unique stations, so if we make google API calls to find Zip code, we will be making many redundant calls, so we decided to map out all the unique stations ID's and their latitude and longitude and reverse geocode using Geopy python API, to reverse geocode locally each station and map each station Zip code with the station ID. After that we created the station id and zip code table in Hive, to further carry our research.

Other reason for increase in bike share riders is also due to increase in bike stations. Since inception of Citi bike share program, total number of stations has almost tripled, from 332 in 2015 to 846 stations in 2019, and largest increase is from 2015-2016 almost doubling the number of stations. This increase of stations from 2015-2016 also correlates with the total rides in those years. This shows that if there is a bike station nearby, then people are more inclined to use bikes. So local authority can make plans to increase the bike stations in

more populated area, which will help in increasing more riders using bikes as their last mile transport.

Year	Total Bike Stations
2015	332
2016	612
2017	799
2018	810
2019	846

Year	Jan-April	May-Aug	Sept-Dec
2016	3003380	5609920	5213042
2017	3544807	6815423	5995658
2018	3844377	7666956	6849653
2019	5124644	7278648	N. A

To get more insight on the factors that impact the number of rides we analyzed a weather dataset to see the correlations.
One of the correlations that we see between the number of rides from the above table and the temperature and dew table is that the number rides is directly proportional to the temperature and dew. If the temperature is less and the dew is less (which means higher probability of rain) then the number of rides is less. If the temperature is more and the dew is more (which means sunny weather) then the number of rides is more.
For instance, in 2018- Quarter 1 the no if rides =1053216 with temperature =35.1 and dew = 24.5 vs 2018-Quarter 2 – no of rides =2625746, temperature =58.5, dew=47.4. we clearly see an increase in the number of rides based on the above example.

Temperature				
Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2016	37.3	58.4	73.3	51.0
2017	36.6	59.6	71.0	51.8
2018	35.1	58.5	73.3	51.3
2019	33.9	59.0	71.9	57.0

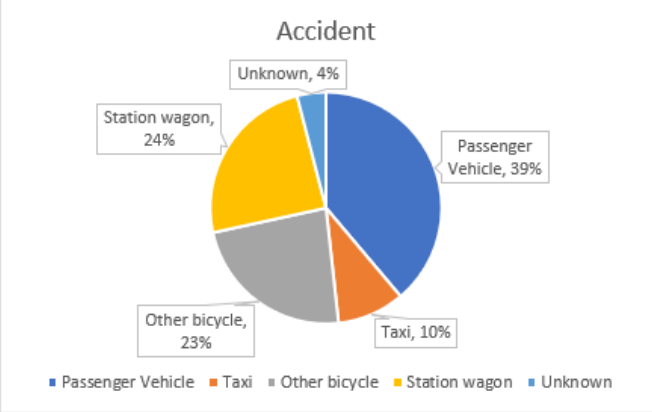
Dew				
Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
2016	24.7	45.3	62.8	35.7
2017	24.8	48.3	61.8	53.5
2018	24.5	47.4	65.2	42.8
2019	21.6	48.4	61.9	48.0

Most popular stations for the rides are Pershing square north in Manhattan, it should be one of the busiest due to its proximity to grand central. Which shows that people who are travelling from nearby areas for work in NYC, tends to choose bikes for their last mile transportations as it is one of cheapest and fastest, as it helps to avoid, terrible NYC traffic.

Station Id	Station Name	Total Rides Out Flow
519	"Pershing Square North"	206221
514	"12 Ave & W 40 St"	179318
426	West St & Chambers St	174393
3255	8 Ave & W 31 St	132251
459	W 20 St & 11 Ave	129567
281	Grand Army Central	128543
2006	Central Park S & 6 Ave	126339
3002	South End Ave & Liberty St	122476

Over the last few years there have been increasing concerns regarding the bike safety especially when Citi bike riders is growing rapidly, but on the contrary, there has been only 1 reported fatality as of 2019 since it began in 2013. So, we decided to analyze the NYC Accidents data and find correlation between Citi bike.

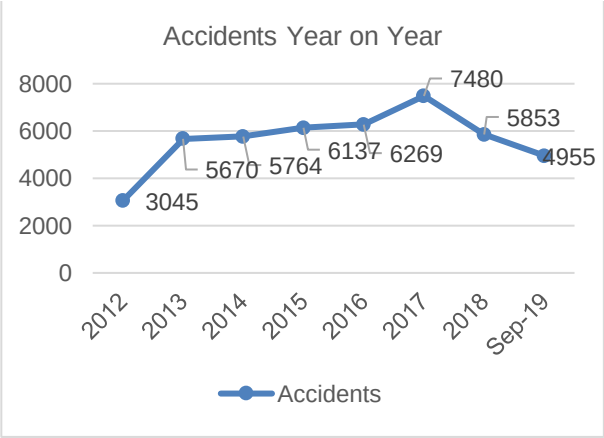
Accident Data:



The above graph shows the distribution of vehicle type that was involved in an accident. The passenger vehicles were the one most involved in an accident with bikes. This analysis raises a question: Is it because of the increase in the usage of passenger vehicles that has led to increase in the number of accidents?

Firstly, we found out the accident trend with year. The below chart shows the accidents on year to year basis. These are only the accidents where bicycle is involved, either as a primary vehicle or as a secondary vehicle. We can analyze that there has been increase in reported incidents from 2012 to 2013, this could be because of Citi bike started in 2013 and the city was not bike friendly and motorist were not ready for such an increase in bicycle riders and this growing trend of increase in incidents can be seen till 2017, with the total of 7480 highest recorded incidents. The major reason is because of the sudden increase in popularity of Citi bike riders as we can see from the data above that in 2017, we saw the highest percentage increase

in total riders and due to this sudden increase, the higher incidents is justified.



But this trend didn't continue as total incidents decreased from 7480 to 5853 in between 2017 – 2018, but total number of riders didn't decrease, this drop is striking because as we expect that increase in bike share riders should increase the number of injuries among cyclist, but the data is not supporting that.

One of the reasons could be measures taken by local authority to improve bicycle rider's safety, by monitoring traffics around higher traffic zone, and increase in bike dedicated lanes, as around 66.1 new lanes have been installed and 20.04 miles of protected bike lanes have been installed in 2018, which directly shows that the measures taken by the local authority has been effective in reducing bike related incidents, and this dropping trend is looking to continue in 2019, as on quarterly basis total number of incidents reported in 2019 has been less than in 2018.

Accidents				
Year	**Quarter 1**	**Quarter 2**	**Quarter 3**	**Quarter 4**
2013	744	1643	2006	1277
2014	584	1665	2195	1320
2015	523	1735	2319	1560
2016	829	1364	2287	1789
2017	1085	2226	2543	1626
2018	1019	1723	1960	1151
2019	880	1698	1929	N.A.

To analyze what is the role of the increasing popularity of Citi Bike riders we combined Citi-Bike data with Accident data to get total bike accidents near Citi-Bike station. We combined these two datasets using zip code, like how many rides originate in and total accidents in the zip code. We can easily analyze that in the areas where the total accidents are highest, it is not necessary they are the most popular areas for Citi bike riders. For example, highest accidents are in Brooklyn (Zip code: 11211), but it is not even in the top 20 popular areas for Citi bike riders.

Total Trips	Total Accidents	Area Stations	Trips / Accidents	Area Zip code
386485	1404	30	275	11211
1046098	1095	24	959	10002
1158090	869	32	1332	10011
113445	954	22	118	11206
235992	868	23	271	11217

Contrary our analysis shows that the most popular areas for Citi bikes has significant lower reported incidents. Such as in 10019, which has the highest total rides but only 332 incidents since 2013. Which clearly shows that increase in the rides is not the main cause of these incidents. Most striking point is that highest incidents are in Brooklyn which has more than 300 miles of dedicated bike lanes highest compared to all NYC boroughs [14]. These analytics has confounded us because we were expecting it to be lower. But this also proves that only increasing dedicated and protected bike paths is not enough for rider safety, controlling the traffic and regulating the motorist is more important. Local authority have been implementing various regulations such as motorist speed, and making illegal to park in bike lanes and even open your cars doors in bike paths, but all of these rules are being violated and often go unnoticed, so it is much important for the local authority to not only spend on building bike lanes but also regulate traffic and motorist behavior.

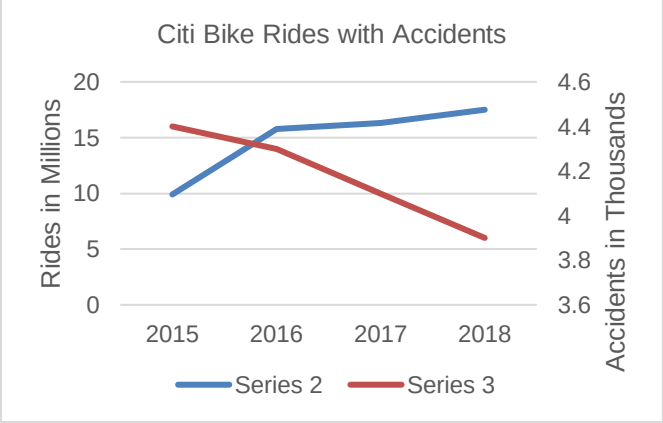
Year	Total Reported Incidents Involving Citi Bike	Trip/Accidents
2019	285	N/A
2018	239	75313
2017	227	71806
2016	215	69767

The above table shows total accidents involving Citi Bikes, and it can clearly be seen, that the numbers are growing each year, which should be the point of concern for the local authority regarding the safety of bikes. Though chances of accidents/trips are decreasing but that doesn't imply that Citi Bikes are getting safer, but it demands for protected bicycle infrastructure and other initiatives.

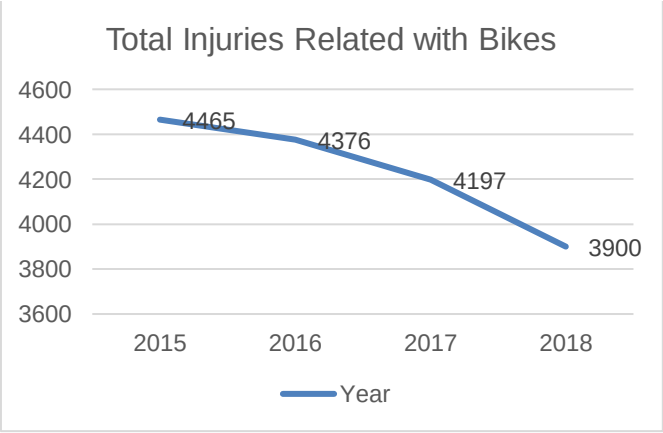
The analysis also takes a different curve here, what if most of the riders involved in these accidents are not Citi bike riders which is shown in [13], as only one fatal accident has been reported, so it's not only other motorist should be aware of cyclist, but cyclist should be responsible and control their speed limits and more regulations should be created such as mandatory helmets and safety lights which may increase awareness from other road users.

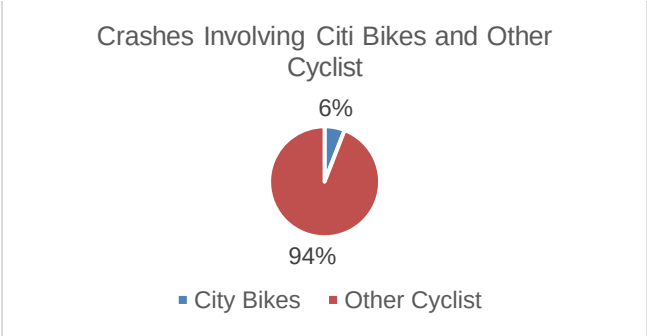
Busiest Zip codes for Citi Bike Riders

Zip Code	Rides	Accidents
10019	1461734	332
10009	1298258	669
10011	1158090	869
10002	1046098	954
10003	988688	742

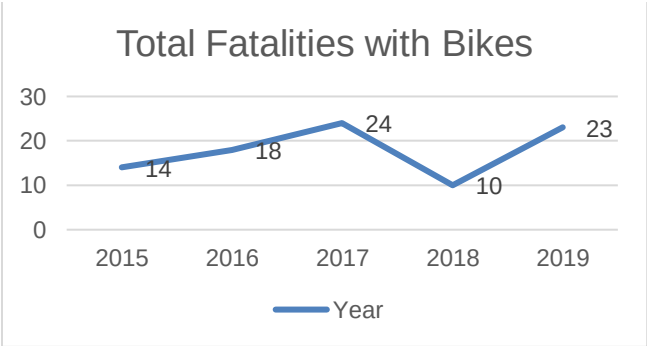


Above graph presents reported incidents from the year 2015-2018 and trips made by Citi bikes. It can clearly be seen that continuous increase in Citi bike share rides is not the cause of accidents. Since 2017 reported accidents are on decline while rides are on increase, and as out of the total incidents most of them are fatalities and since there is only one reported fatality of Citi bike riders, we can analyze that Citi Bike are pretty safe, and most of these incidents are mainly caused by the non-Citi bike riders or personal riders.





Most of the accidents that were in the dataset involved injuries. The above graph shows the decrease in injuries from year 2015-2018. One of the reasons might be increase in the bike lanes [12], that may have led to a better infrastructure. Our results also shows that 94% of the accidents involved other cyclist and only 6% were involving Citi Bikes, which points that Citi Bike riders are much safer than other cyclist.



The above graph shows the distribution of fatalities reported from year 2015-2019, which has increased abruptly in 2019. These brings out questions such as: Is the safety of bike riders improving? Are all the accidents reported? As no data can be found about whether a Citibike was involved in an accident reported by NYC motor vehicle collision, the analysis still remains just a hypothesis.

VII. Future Work

The scope for future work is massive given that we have three datasets involved in deciding the safety of bike sharing. A study to split the accidents between bike share users and general bike users since we at this point have no data to differentiate if a particular accident was caused by a bike share user or a general bike user. We can further our analytics by taking into account the accident data from the police department and the accident data from the bike sharing operators and see if there is any reporting difference between the two. We can also consider speed as a factor contributing to the accidents and be more specific about the trends in the accident data in terms of the average speeds of a bike share user vs a general bike user and see how this impacts the number of accidents curve. Based on the accident time and area we can get to know if there's a need for better bicycle infrastructure or the need of more streetlight (based on the accident time). Our analytics get much more insightful when we cover all the above scenarios and the insights would be more conclusive and solid.

VIII. Conclusion

Efficient traffic flow directly affects the urban economic development and well-being of residents. Bike share programs provide cheapest and sustainable means of transportation which solves problem of 'last-mile' and traffic congestion. There are more than 1000 cities which currently have bike share systems, and many are in process of expanding rapidly. One of the biggest bike share systems is Citi Bike NYC.

Our research shows there have been more than 17 million rides alone in 2018 and it will cross this number for the 2019. The average duration was 18 mins, while average distance was 2 miles and the average speed is 11.2 miles. These figures tentatively match the study [15]. Peak hours for the Citi Bike riders are between 8-9 AM and 5-7 PM, which clearly shows that Citi bike is mostly used for last-mile transportation, to and from the office.

The most popular stations for the Citi bike riders are near major transport stations such as Pershing Square, Grand Central Terminals, and near major tourist stops, such as Central Park and Union square. While the lease frequented destination stations are in Brooklyn.

One of the biggest points of concern is the rider's safety and increasing popularity of bikes has also increased the major crashes but according to New York Times report there has been only one major fatality involving Citi bike. Number of reported incidents involving crashed of Citi Bike has definitely increased from 215 in 2016 to 285 in 2019, but still it is far less when compared to crashes and fatalities involving non-bike share users. Growth in total rides is larger when compared to increase in crashes, which leads us to conclude that despite these increase in crashed Citi Bike users are much safer than non-Citi bike users. Increase in crashed hasn't affected in popularity of Citi Bikes in anyway.

While analyzing the accidents dataset, we found that most of the incidents are reported for injuries, but there can be a lot of underreporting, as minor crashes are probably not reported which can contribute to the low risk for bike share. Also, no proper connection between accident data reported by citibike and by the NYC motor vehicle collision can be found, which restricted us to just estimate the contribution of citibike to the accidents, using the citibike station and the accident location. These are some of the limitations for this research.

Other interesting point is that most of the accidents involving Cyclist didn't even involve Citi Bikes, almost 94% of total accidents involved personal cyclist, which tends to show Citi Bike riders are much safer or it could be the result of underreporting of accidents as the data for accidents were taken from two sources, Citi Bike crashes were based on Citi Bike monthly reports, while NYC crash data is used for all accidents involving cyclist and we were unable to distinguish between crashes involving Citi Bike and other cyclist, which lead a limitation in our results. But completely based on the data, we can say that Citi Bike riders are safer than personal bikers, this could be because they ride at lower speeds.

One of the biggest striking result was that the highest accidents were in Brooklyn [16], and Brooklyn is the least popular for Citi bike users and it has largest protected bike lane network in NYC, which is the biggest point of concern and this demands for new rules and regulations for protection of shared bike and non-shared bike users. This also shows that only building protected lanes will not solve this problem, there is a need for deeper solution. There has been another analysis that out of top 10 popular areas for Citi bikes, only 2 are in top 10 accident prone areas. Which shows that areas with high popularity of Citi bikes involve less fatalities. This could be because bike sharing operators often limit the speed of bicycles so bike share riders ride at lower speed when compared to other cyclist which helps them to avoid crashes. Bike sharing cycles are also equipped with safety lights which may increase awareness, and mostly bike share is popular in inner city and where there is better infrastructure. Finally, where there are high number of cyclist other motorist perceive with caution towards cyclist.

Our analysis shows there is a lot of scope of improving the safety of bike share riders and other cyclists. Despite increase in protected bike lanes, there is a need of improved infrastructure and other initiatives to improve a bicycle friendliness. Local authority and bike share operators should discuss and answer some important questions like why Brooklyn has maximum number of incidents, despite having largest protected bike lane network? What are the major causes of these accidents? Are rules and regulations being violated? Decision makers should reassess their plans and should come up with better set of rules and regulations such as motor vehicle speed limitation in areas with higher number of cyclists, mandatory helmets and safety lights. Answering these issues will create a safer environment for both bike share and private riders.

Acknowledgment

We would like to thank Prof. Suzanne McIntosh, Srishti Grover, Omkar Patinge for being available and helpful whenever needed. We would also thank NYU HPC team for being very responsive. We also thank Citi bike, National centers of Environment Information for keeping their data available for all. The IEEE, Springer and other journals which we could refer and gain more insights about our analytic.

References

1. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
2. J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implemenation, 2004.
3. S. Ghemawat, H. Gobioff, S. T. Leung. The Google File System. In Proceedings of the nineteenth ACM Symposium on Operating Systems Principles – SOSP '03, 2003.
4. https://www.citibikenyc.com/system-data
5. Y. Zhang and Z. Mi, "Environmental benefits of bike sharing: A big data-based analysis", *Applied Energy*, vol. 220, pp. 296-301, 2018. Available: 10.1016/j.apenergy.2018.03.101.
6. Predicting Bike Usage for New York City's Bike Sharing System
7. https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions-Crashes/h9gi-nx95
8. Soltani A., Mátrai T., Camporeale R., Allan A. (2019) Exploring Shared-Bike Travel Patterns Using Big Data: Evidence in Chicago and

Budapest. In: Geertman S., Zhan Q., Allan A., Pettit C. (eds) Computational Urban Planning and Management for Smart Cities. CUPUM 2019. Lecture Notes in Geoinformation and Cartography. Springer, Cham
9. Fishman, E. and P. Schepers (2018), “The Safety of Bike Share Systems”, ITF Discussion Papers, International Transport Forum, Paris.
10. https://www1.nyc.gov/html/dot/downloads/pdf/cycling-in-the-city.pdf
11. Nicaj L, Wilt S, Henning K. Motor vehicle crash pedestrian deaths in New York City: the plight of the older pedestrian. *Inj Prev*. 2006;12(6):414–416. doi:10.1136/ip.2005.010082
12. https://data.cityofnewyork.us/Transportation/Bicycle-Routes/7vsa-caz7
13. "No Riders Killed in First 5 Months of New York City Bike-Share Program", *Nytimes.com*, 2019.
14. Brooklyn Bike Paths, Bike Lanes & Greenways", *NYC Bike Maps*,
15. D. Singhvi, S. Singhvi and P. I. Frazier, "Predicting Bike Usage for New York City’s Bike Sharing System", 2015.
16. http://www.nycbikemaps.com/maps/brooklyn-bike-map/