NYC Transportation Method Selection In Rush Hours

Jiaqi Liu
Courant Institute of Mathematical Science, New York University
New York, New York
jiaqi.liu@nyu.edu

Xuebo Lai
Courant Institute of Mathematical Science, New York University
New York, New York
xl1638@nyu.edu

Yunhan Yang
Courant Institute of Mathematical Science, New York University
New York, New York
yy1316@nyu.edu

Abstract

This is an interesting analytics project of travel efficiencies during rush hours on two of the most commonly known transportation methods, Citibike and For-Hire Vehicles (FHV), in New York City. Our primary objective is to develop a comprehensive understanding of both Citibike and FHV traffic patterns and to be able to recommend traffic options with more efficient travel time during rush hours for people. In this project, we approach the project through partitioning the Manhattan Island into 48 independent and non-overlapping zones and then analyze travel efficiency for in-zone and cross-zone traffic activities. Analyze approaches include basic statistic modeling, mean trip duration comparison, correlations with different weather conditions, and traffic flow predictions.

Keywords— Big data analytics; Travel time prediction; Traffic analytics;

I. INTRODUCTION

New York City is the third most congested city in the world in terms of traffic and the second worst in the United States, according to the INRIX Global Traffic Scorecard. For people living in the city, it can be quite challenging to select an efficient transportation method during rush hours. As stated in the New York Times, the rush hours in New York City has two major periods, from 7 to 9 A.M. and from 4 to 6 P.M. Aiming at maximizing the efficiency on commuting within the New York City, we have conducted a big data analytics project that focuses on analyzing the traffic activities within Manhattan Island for traffic options of For-Hire Vehicles (including yellow taxis) and Citibike. The objective of this project is to develop a comprehensive understanding of both Citibike and FHV traffic patterns in NYC and to be able to recommend traffic options with more shorter travel time during rush hours for people. In order to achieve the analytics, the project has partitioned the Manhattan Island into 48 independent and non-overlapping zones by adopting taxi zones compartmentalized by the New York City's Taxi and Limousine Commission. In this project, we have extracted, cleaned, and analyzed batches of raw data regarding traffic activities in recent half years with the two potential traffic options stated above. Our primary analytic strategies are as following: (1) Calculate the mean trip duration for all Citibike and FHV with respect to the same available pickup and drop-off locations. (2) Evaluate correlations between weather and traffic activities. In particular, we would like to see how the weather impacts on the number of users for Citibike and FHV, as well as how the traffic speed changes. (3) Establish a traffic flow pattern of the pedestrian movement based on the most frequent pickup and drop-off locations for both transportation means in morning rush hours and evening rush hours. Each traffic activity dataset contains roughly 5 Gigabytes of data. Various data filtering, statistical modeling, and analyzing algorithms were implemented to improve the accuracy of data collection and ETL. The technologies used for analytics are MapReduce Programming Paradigm, Apache Impala SQL query engine, and Hadoop Distributed File System on Dumbo machine from NYU High Performance Computing. Our desired typical user for the project application is the residents commuting within the Manhattan Island on a daily basis. We believe there are several actionable business insights from the analytics and other parties can benefit from this analytic result. This project analytics can serve as a good reference for NYC's famous yellow taxis and FHV (For-Hire Vehicles businesses owners) organizations such as Uber and Lyft to increase their businesses in less congested zones or zones with relatively high commute efficiency. For Citibike, as well as its other competitors, they can relocate and deploy a reasonable number of bikes at zones with high demand and much more congested areas so the efficiency of riding a bike is better than driving and taking a train. The result in the analytics can also be used by the NYC government to produce better public transportation headway control scheduling or to create special public bus/MTA routes during rush hours that passes hotspot regions to enhance service and space utilizations of public transportation.

We organize our paper as follows: The next section presents the motivations of performing this analytics; Section III presents the related work covering aspects from data filtering algorithms to dynamic travel time prediction model, and path prediction model; It is followed by Section IV which includes the detailed design and implementation of our analytic strategy models; In Section V, we present our results acquired from the models that we implemented and constructed; All future works that can expand the project are in Section VI. Finally, we conclude this analytics project in Section VII.

II. MOTIVATION

We select this topic to conduct a traffic condition study because we live in the New York City on a regular basis, and we all believe the level of traffic congestion is extremely

inconvenient for people commuting in the city, especially during rush hours. This phenomenon got us thinking: what are the potential solutions to travel efficiently transport from one point to another within New York City. Perhaps there are other public transportation methods that are more efficient than driving/taking a cab or taking a crowded and sometimes unreliable subway train.

According to New York City's official economic development corporation, NYCEDC, there are 55 percent of commuters using the public transportation system, compared to 4.9 percent nationally. 86 percent of NYC commuters drive to work, which is roughly 28 percent of New Yorkers. The in-depth analytics of two of the popular transportation methods in NYC project makes it possible for potentially all residents in the city to get to destination efficiently instead of wasting it on a traffic heavy road or standing with a crowd on subway. Beyond this, it is also possible for us to expand the current work to other cities with similar traffic conditions and provide proper suggestions.

The richness of data available also motivates us to apply numerous analytical tools and techniques on the data to provide better suggestions and generalize our code as much as possible so it is adaptable for larger volume of data in future. We hope that our code can help the city government to make smarter decisions of road construction planning and potentially design more alternative routes and services for New Yorkers.

III. Related Work

There are numerous excellent related researches that provide important advises for this analytic project.

We found a paper named *Estimating Pedestrian Densities, Wait Times, and Flows with Wi-Fi and bluetooth Sensors* [5] to be quite critical and helpful for us to develop the filtering algorithms used for this analytics. The paper uses Wi-Fi connections to determine pedestrian flows, and introduces two major data filtering algorithms that we found quite useful.

Moving blocks algorithm was used to counting people in the crowds. It filters out people that move in circles or going back and forth to avoid double counting. For each connected device, it detects whether a device that shows up multiple times in every 5 minutes and only treats it as one count and if the same devices show up after the 5 minutes cool-down period, it may get one more count. Based on their idea, we have implemented a similar algorithm to filter Citibike data with same starting bike station id and ending bike station.

The estimation of pedestrian flows algorithm focuses on finding overlapping devices from recorded from multiple sensors. The researchers assume the sensor devices have a known distance, and so can base on the connecting time data to draw the flow path and corresponding moving speeds. If an individual walk significantly slower than a standard walking speed (normal distribution of 1.2m/s), then the data should be filtered out. However, in our case, there may be some drawbacks of implementing this strategy since in our dataset, zone numbers are the only information we have related to the user's geographic location, and so we can only estimate a rough distance for cross-zone traffic activities and calculate a relatively less accurate speed based the information. Since there is no standard traffic speed, and this speed highly depends on the region, this approach was not used in the analytics project.

Ideas from a famous paper, *Dynamic Travel Time Prediction with Real-Time and Historic Data* [6], written way back in 2003 helped us to resolve this filtering issue. This paper focused on the modeling a way to perform dynamic travel time predictions. It discusses numerous models that have been implemented and tested on the traffic data from New York State Thruway.

The way that they perform extreme data handling has been integrated into out analytics project. Since there are numerous traffic conditions on the road, there could have vehicles with abnormal characteristics. The researchers came up with a basic statistical model to filter out data based on the performance of all vehicles. They have calculated an average time of all vehicles passing through the same region and remove all data that have been highly deviated (50% higher or lesser) from the mean travel time is considered as biased and thus removed.

The prediction accuracy verification is also quite interesting and important. This paper introduces the user of two commonly practiced approach to examine the goodness of travel time-related data: linked-based and path-based. The path-based prediction approach focuses on calculating the travel time when a vehicle passes a particular path, and the link-based is the total sum of travel times of vehicles in the consecutive individual links that constitute the whole path. According to two of their error prediction models, mean absolute relative model and root relative square model, they find the path-based prediction model to be less accurate to their hypothesis since it is very sensitive to road conditions. However, from the standing point of our analytics project, we prefer exactly the model that reflects the congested road conditions during rush hours in New York City. The path-based model is used in our accuracy examination to check for the goodness of cross zone traffic activities.

Different from the mature approaches adopted by the paper above, the research paper Travel-Time prediction with Support Vector Regression [7] by Chun-Hsin Wu ; Jan-Ming Ho ; D.T. Lee, explores a new way, Support Vector Regression(SVR), to predict travel-time. This research believes that SVR has been underrated in predicting travel-time throughout history, especially given the success of SVR in multiple fields including Financial Market, predictive maintenance and electricity price forecasts. Also, the past application example has indicated that SVR approach works particularly good for time series predictions, which further implies its potential in travel time prediction. Moreover, the paper manifests a significant improvement in accuracy of travel time prediction by using SVR, compared with other traditional modeling methods, when applying both SVR and traditional modeling ways on Sun Yet-Sen Highway traffic data from Intelligent Transportation Web Services Project.

The research paper also gave a detailed description in the math and program implementation on SVR which offers us the opportunity to reproduce their algorithm. For this paper, however, we would only adopt and test the traditional methods for travel time predictions due to limited time and technology restriction. We expect to apply SVR algorithm to predict both

Citibike and Yellow Taxi's travel-time and compare their accuracies with those obtained by the traditional ways in the future.

When doing the research on compare speed Citibike and FHV in a city scale, we discovered a paper called "An optimization approach for the placement of bicycle-sharing stations to reduce short car trips: An application to the city of Seoul" [8], this paper took the approach which extracted taxi trajectories similar to bicycle riding patterns as an applicable dataset. It helps us part of the data clean and profiling algorithm design. In order to find similar trajectory of taxi and bike, we use location mapping on bike pickup site and destination to match taxi's pickup and drop off zones. Besides, in the future, for our algorithm design, we found that the approach that the paper took in calculating the distance of each travel by using these pick-up and drop-off sites and then extracted trips of a distance no greater than three miles and illustrates the traffic flow in the New York City during rush hours with illustrations are useful.

When we trying to explore the data analytics on possible parameters, we found a paper called "Is There a Limit to Adoption of Dynamic Ridesharing Systems? Evidence from Analysis of Uber Demand Data from New York City" [9], The paper presents multiple possibilities for us to explore on parameters, such as Time Variables: summer, winter, the difference in weeks from the initial week. Environmental Variables: Weekly Precipitation (in.), Weekly Average Max Temperature (o F), Built Environment Variables, Demographic Variables, and Interaction Variables. When we analyze our data, we feel those variables are useful for us to select some of them to find the pattern in the big data, such as the environment data, time variables. Though the panel model that the paper used for Uber weekly pickup demand with week index is not applicable in our case, our datasets focus on weekdays with weather factor. For the part that the random effects structure was employed that allows consistent estimation of coefficients associated with time-invariant variables, it relates to the statistical model we built for eliminate outliners when calculating mean trip duration comparison on certain dates of the trip with the same pickup and drop off locations. From the influence of environmental variable explanation in the paper, it exposes that these riders may potentially opt for an Uber trip if it means avoiding exposure to the elements, which we also going to explore the potential association commuters' selection of transportation when meeting different weather conditions.

IV. Design and Implementation

A. Description of Datasets

In this research, we mainly made use of three different datasets, TLC Trip Record Data, Citi Bike Data and NOAA Weather Report Data to find the correlations between weather and different traffic ways. Each of the dataset and the specific approaches that we adopted have been detailed below.

The first traffic way that we research is for For-Hire Vehicles ("FHV") service. In order to better understand this traffic way, we analyze New York City's FHV trip data collected from yellow and green taxi trip records. We pulled obtained the data from New York City Taxi & Limousine Commission (NYCTLC) Official Website. The FHV data has been collected since 2009 until now. Sized 1 Gigabyte, each raw FHV data file contains roughly a million records of FHV trip data in a month. Because our research goal is to analyze and find the most efficient travel ways under certain weather conditions for people currently living in New York, we decided to make use of the latest data ranged from January 1st 2018 to June 30th. Before we perform data processing, we wrote a few python scripts to combine all the FHV data files from first half of 2018 and obtained a merged file of roughly six million records in size six Gigabytes.

In the raw FHV data file, each trip record from NYCTLC is defined by 17 fields, including Vender ID, Trip Pick-up and Drop-off time, and Passenger Count. We can mainly divide the fields into three categories, trip travel information (pick-up and drop-off time/zone), payment information (payment type, fare amount, tax, tolls amount), and passenger information (passenger count). Because our research is only interested in analyzing traffic information, we would be mainly using the trip travel information relevant fields.

The second traffic way that we dived into is biking. In our case, we decided to make use of shared Citibike data for the analysis. This is because, shared Citibike is a strong alternative for subway, on-demand mobility services and FHV services. According to our analysis of Citibike data published, there are 418 Citibike stations in Manhattan. What's more, according to Citibike official site, there are currently 143,000 registered bike members and 50 million trips taken place for the past five years. [10] Abundant bike stations and substantial users indicate that Citibike can be a potential stable alternatives solution to New York City crowded subway and FHV services, particularly in short/medium range distance travel.

We obtained Citibike data from Citibike official site. The data is available from June 2013 until now. Because Citibike clients increased exponentially every year and are highly subjective to weather conditions, the data available varies greatly in size ranging from few hundred Kilobytes to few hundred Megabytes. Because of the research goal for this project, we made use of the latest data which is the Citibike data from January 2018 to the end of June 2018. Each raw data file contains roughly 30 thousand entries of data. Similar to the approach that we adopted for FHV data, we used python scripts to combine all the available data in the beginning. The merged Citibike data file has roughly 180 thousand entries of data.

In the raw Citibike data, there are 15 fields to define each of the riding record, Trip Duration, Start Time, Stop Time, Start Station ID, Start Station Name, Start Station Coordinates, End Station ID, End Station Coordinates, Bike ID, User Type, User Birth Year, User Gender. It is very interesting to note that the riding data from Citibike contains some basic information about users. This project would not dig into users' data since it is irrelevant to our current

research goal, but we expect to learn more about users' behaviors in the future work.

In order to understand how the weather conditions would affect public choices among different traffic ways, we obtained weather data for New York Central Park from National Centers for Environmental Information (NOAA).

The raw data file that we obtained from NOAA contains nine fields, which detail all the parameters of current weather and information of observation stations such as temperature range, precipitations and observation station ID. The parameter fields that we are interested in the most are date, precipitation in millimeters, and snowfall depth in millimeters, because the three major weathers that greatly influence people's choices of transportation ways are sunny, snowing and raining.

B. Design Details

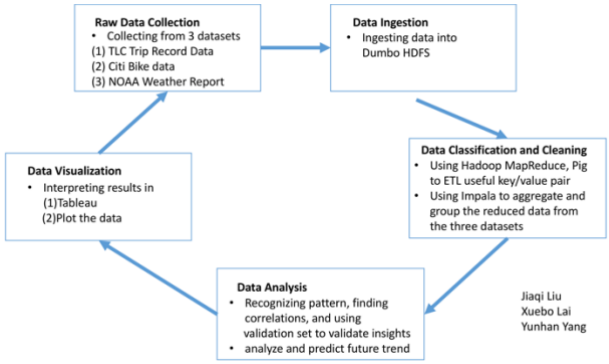

Figure 1. Design Diagram for NYC Transportation Method Selection In Rush Hours

Given the size and number of records, we decided to make use of MapReduce Programming Paradigm that is available in Courant Dumbo Hadoop Distributed File System (HDFS) to clean, process and analyze our data.

B.1 Data Ingestion

We first used command "scp" to upload all the data to our servers. From our servers, we used Dumbo to inject the data into Hadoop Distributed File System (HDFS).

B.2 Data Cleaning, Profiling and Aggregation

In this part, we mainly perform three operations on all the data, data cleaning, profiling and aggregation. These three MapReduce operations were customized and designed to cater each set of data. The data cleaning operations to retain only useful fields and erase invalid records are written in PIG, a high-level built-in tool in HDFS. The data profiling and aggregations are written in Java.

B.2.1 FHV Data

For FHV Data, we used Pig Program to retain only, Vender ID, Pick-up Time, Drop-off Time, Trip Distance, Pick-up Location ID and Drop-off location ID since we are only interested in the correlation between travel time using FHV in any two given zones and weather conditions. Fields such as trip payment type, passenger count, fare amount, which would not create any value for our research, have all been dropped from the data file. Also, after we random sample 1% of all the data, we realized there are a huge amount of records that are missing value in one or multiple fields. We, therefore, wrote codes in Pig to eliminate all the records with one or more missing values.

In the data profiling and aggregation step, we wrote MapReduce function to further process the data. Because we are only interested in the weekday rush-hour transportation in our research, we first used mapping function to erase all the data that do not fall into the desired time range (Desired data ranges are from either 7am-9am or 5pm-7pm during weekday).Then, we format the key as the combination of trip date, pick-up location id, and drop-off location id; we format the value as combination of vender id, trip distance, and trip duration. In the reduce phrase, we aggregate the data by the trips that share the same date, pick-up and drop-off location. Then, we calculate the average trip duration as the value of the reduce function. At last, we output the key as the input key from mapping function and value as specified above.

B.2.2 Citibike

Similar to what we did for FHV Data, we used PIG script for Citibike data to erase all the data entries that have one or multiple missing values. Then, we retain only the fields that later contribute to our analysis, which are trip duration, start and end time, start and end location and bike id.

In the MapReduce Programs, we perform data profiling and aggregation. For the Mapping phase, we filtered out all the data that are not within rush hour during weekdays. For data that are in the morning rush hour, we label it as 0; we label the data in the evening rush hour as 1. Furthermore, in order to keep the consistency of the data, we translate (map) all the data from bike zone to taxi zone. We output the key of the mapping function as a combination of data's data, rushing hour labels, start and end taxi zone; we output the travel duration as value output. In the reduce phase, we further erase all the data with travel duration that is out of 2 standard deviation since we consider that as outliers. We outputted the key as it is from the mapping phase, and the value for the reduce phase as the average without outliers of all the travel duration data that share the same date, rush hour, start zone and end zone.

B.2.3 Weather

In the PIG file, we iterated through all rolls of weather records and retained only two fields, which are the date, precipitations and snowfall depth for each data entries since we found out the most relevant weather patterns that would affect people's choices of transportations are snows, raining and sunny. In the MapReduce function, we translate the precipitation and snowfall depth into the dummy variables that can represent weather conditions. In this case, when

there is no precipitation and snowfall, we label the data entry as 0. If there is only precipitation, we label the data entry as 1. If there are both precipitation and snowfall, we label the data as 2. In this way, we make the value discrete which would help us later when we do the analysis on those data.

V. Results

A. Taxi and CitiBike Analytics

A.1 Data fields Summary:

The major data fields that we used in the analytics include: pickup location, drop-off location, rush hour(discrete values), all taxi average trip duration, all bike average trip duration, speed comparation, speed difference, number of taxi, number of bikes, count comparatino and count difference.

A.2 Data Analytics

A.2.1 Define questions and hypothesis

We define the same trip as the ones that share the same pickup location, drop off location, rush hour, and traffic way. Within the framework with respect to trip, we have raised several questions: Which transportation method is more optimal in time and in what percentage? What is the traffic trend during rush hours? How do the commuters choose between For-Hire Vehicles (FHV) and Citibikes?

In the experiment, we set up the purpose of the experiments based on the fundamental analytic questions above. For the first query, we would like to know whether Citibike or FHVs is faster and in what percentage, with respect to the same trip as defined above?

A.2.2 Query Data

The first hypothesis that we raised is that, people should make decision when they travel.

We ran the Impala to query both bike data and taxi data with respect to the same trip as defined above. The query returns 2706 different comparation with respect to the same trip. For all the comparations, there are 1478 cases where For-Hire Vehicles (FHV) is faster than the Citibike. On the other hand, there are 1228 comparations where Citibike is faster than FHV. In total, the FHV is faster in 20.35% cases.

To further confirm our result, we also wrote query to check how many FHV trip is faster than Citibike trip. We compared the data from FHV and Citibike data that have the same start location, same end location and same rush hour on the same day. In total, the query returns 2,356,382 total trip data for FHV, with 1,156,102 entries of FHV data faster in the comparison to bike data. On the other hand, the query returns 606,427 entries of bike data, with 311,930 entries of data faster than FHV data. From above data, we can easily calculate that the probability that publics successfully choose FHV as a faster transportation tool is fast is 49.06% and the probability of that for the Citibike is 51.44%.

The ratio on the type of trips that taxi is faster than Citibike is 120%, however, the number of trips that people make the choice is in low success rate. Since it only 49.06% success rate when people choosing the taxi, 51.44% success rate when people choosing the bike. From that, we got an insight that Some commuters should change their options on transportations during rush hours. It does prove our hypothesis. It's necessary using NYC Citibike and FHV from the pick-up, drop-off location, and rush hour to predict the average trip duration times and give suggestion to residents commuting within the Manhattan Island on a daily basis which public transportations options are more optimal.

For the second question, we make the hypothesis that the both the bikes and FHV pick-up locations around train stations are more popular than any other areas. we ran the Impala query on both taxi and bike table. We grouped the taxi's and bike's trips data by the same pick up location, drop-off location, same rush hour at the same date. The result that our queries output contains fields including pick-up (pulocationid), drop-off location(dolocationid), rush_hour (0 denoting 7am-9am; 1 denoting 5pm-7pm), number of taxis, number of bikes, average speed comparison.

The top five trips in Manhattan during rush hours are shown below,

TABLE I.
TOP 5 MANHATTAN TAXI TRIPS IN RUSH HOUR

pulocationid	dolocationid	rush_hour	all_taxi_commuter_count	all_bike_commuter_count	speed_compare
186	234	0(7-9am)	18589	760	1
107	170	0	12685	684	0.85
107	162	0	11238	914	0.84
140	162	0	11222	530	1.39
229	162	0	11159	646	1.7

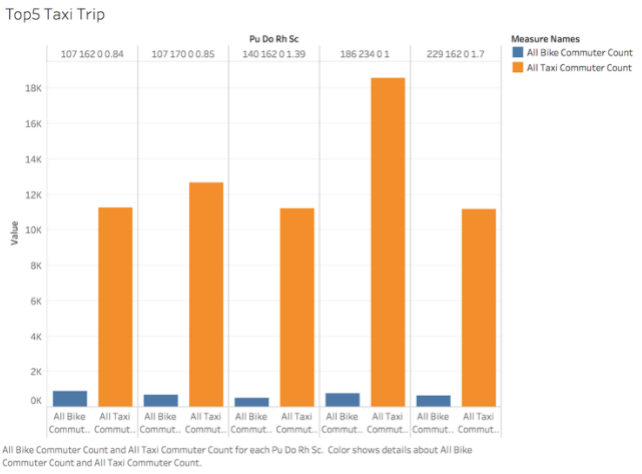

Figure 2. Top 5 taxi pickup diagram with information of pickup drop off location id rush hour and speed compare value

Here are the top 5 Citibike trips in Manhattan during rush hour:

TABLE II.
TOP 5 MANHATTAN CITIBIKE TRIPS IN RUSH HOUR

pulocationid	dolocationid	rush_hour	all_taxi_commuter_count	all_bike_commuter_count	speed_compare
43	43	0(7-9am)	4893	21289	0.74
43	43	1(4-6pm)	6248	19879	0.72
43	164	0	1140	6007	1.35
68	68	1	7516	3841	0.88
164	43	1	1006	3511	1.25

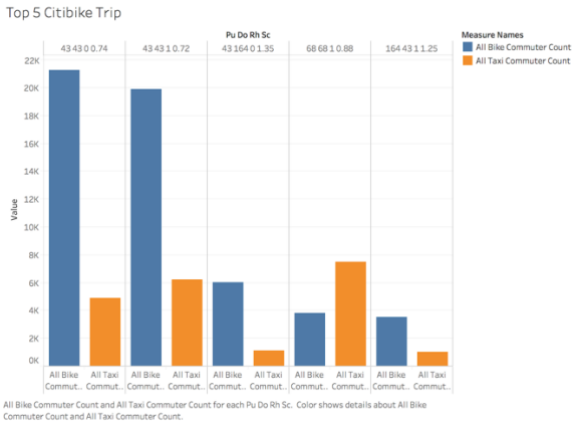

Figure 3. Top 5 CitiBike pickup diagram with information of pickup drop off location id rush hour and speed compare value

From the observation above, during rush hour morning 7-9 am evening 4-6pm, we found that the most frequent taxi pickup location is 186: Manhattan Penn Station/Madison Sq West, drop off location is 234. The locations like Manhattan Gramercy(107), Manhattan Murray Hill(170), Manhattan Midtown East(162) are also among the top pick-up and drop-off locations.

What out of our expectation is that NYC FHV are busier in morning rush hours (7-9 am) than evening rush hours (4-6pm). The top five FHV trips all happened in the morning rush hours (7-9 am). The pickup location at Manhattan Penn Station/Madison Sq West(186) meets our hypothesis. This analytics result provides helpful tips for FHV drivers to find zones with higher demand.

For the Citibike data, it's really unexpected that Manhattan Central Park (43) has the most pick-up and drop-off locations, which was different from our hypothesis of the train station. The rest of the top pick-up and drop-off locations for bike are East Chelsea(68), and Midtown South(164). It's interesting to find midtown area and central park are the most popular Citbike pick-up zones for NYC residents. The actionable insight here is that Citbike can set up more bike stations at the top pick-up zones specified above. During rush hours, FHV drivers need to be careful of bikers around their area. And the bike riders commuting in those areas need to be careful of other bikers and pedestrians.

B. Taxi Citibike Weather Analytics

B.1 Data Fields Summary:

The data fields for this analytics are rush hour, taxi average trip duration, bike average trip duration, speed comparison, speed different, number of taxi commuter, number of bike commuter, the count number comparison, count number's difference and weather conditions.

B.2 Data Analysis

B.2.1 Hypothesis and Questions

We have raised several questions: Would weather be a factor influencing people's transportation options? (The weather patterns that we analyzed here include snowday, sunny day and rainy day). Also, what is the impact of weather's influence on the speed of FHVs and Citibike.

B.2.1 Query Data

We designed the experiment to compare commuter number of sunny day, rainy day, and snowy day . Our Hypothesis is that the weather would not have a large influence on people usage of transportation during rush hour since we are doing the research on the weekdays.

We ran the Impala query on table bike, FHV and weather tables. We group the data by same pick up location, drop off location, rush hour, weather. The output fields that we have include pick-up location(pulocationid), drop-off location(dolocationid), rush_hour, average FHV(Taxi) speed, average bike speed, speed_comparison, speed difference, number of taxi trip, number of bike trip, count number comparison, count number difference, weather.

The result obtained from the query is shown below:

TABLE III.
TAXI CITIBIKE WEATHER IMPACT SUMMARY

	Taxi Daily Commuters number	Taxi Faster Rate	CitiBike Daily Commuters number	CitiBike Faster Rate
Sunny	22532 per/day	48.9%	5842 per/day	49%
Rainy	22061 per/day	48.15%	5803 per/day	53%
Snowy	10794 per/day	53.7%	1715 per/day	48%

The analytics result have disproved our hypothesis that the weather does not have a large influence on people usage of transportation during rush hours. From the chart, we

can see for both taxi and Citibike, snow has the largest impact on commuters' choice on public transportation. Also, with the weather as a categorize, the success rate of people chose the correct transportation options during rush hour is lower what we expected. It's, therefore, necessary to use NYC Citibike and FHV data from the past to predict the average trip duration and provide suggestions to residents commuting within the Manhattan Island on a daily basis.

C. Special Events analysis

C.1 Hypothesis and Questions

The experiment in this part is designed to discover special events happened on the first half of 2018.

Using the bike table, taxi table and weather table, we queried to find the days that have maximum and minimum number of taxi trips. Similarly, we query to find the days that have maximum and minimum number of bike trips. From the impala script, we obtained the following result.

TABLE IV.
MAX TAXI BIKE PICK UP TABLE

Max(Taxi)

date	taxi_commuter_count	bike_commuter_count	weather
2018-05-02	29987	9286	0(Sunny)

Max(CitiBike)

date	taxi_commuter_count	bike_commuter_count	weather
2018-05-10	29307	12863	1(Rainy)

TABLE V.
MIN TAXI BIKE PICK UP TABLE

Min(Taxi)

date	taxi_commuter_count	bike_commuter_count	weather
2018-01-04	282	63	2(Snowy)

Min(CitiBike)

date	taxi_commuter_count	bike_commuter_count	weather
2018-01-04	282	63	2(Snowy)

It is unexpected that the minimum taxi and Citibike commute date is the same day. We did research on that day and discovered that 'Bomb cyclone' winter storm brought snow, fierce winds to New York State that day. The weather that day was extremely bad. Also, we found that the peak of the number of both taxi and Citibike trips happened in May. From the result above, we can provide actionable insights that the FHV and Citibike managements need to warm publics about the extreme weather in January, and increase the number of Traffic polic in May.

VI. Future Work

Given time, we would like to expand our analytics to more transportation methods, we wanted to include datasets such as Buses and NYC Subway Train. With the data for other transportation methods, this analytic can provide a much more detailed and thorough comparison that truly convers every aspects of transportation methods in the city.

Also, we would also like to expand our analytics into a different city that has a similar traffic condition and study the similarities in traffic between New York and other cities to find out the common facts that is shared between multiple cities for rush hour transportations.

What's more, the travel time prediction and analytics that we have done in this paper are mostly based on the past average travel time between two zones. We would like to implement the a more accurate method for predicting the travel time. As discussed in Related Work Section, there is an underestimated highly accurate travel time prediction model called Support Vector Regression(SVR), discussed in the research paper *Travel-Time prediction with Support Vector Regression* [7]. In the future, we will program this algorithm to try to increase our prediction accuracy as well as test the arguments about SVR made by that paper.

VII. Conclusion

In the experiments conducted in the Result section, we have applied big data tools to fulfill the analytics inquires and to explore optimal solutions for our desired users. NYC has high demands for public transportation system compared to other regions in the nation. In order to discover the traffic patterns during rush hour, we have analyzed all traffic activities within the Manhattan Island for Taxi and Citibike as well as their correlations to weather. Our approach can be partitioned into three parts: 1. Transportation efficiencies, 2. Traffic trends within Manhattan Island, 3. Weather impactions.

From the Traffic trend analytics for the Rush hour Taxi top 5 trips in Manhattan and Rush hour CitBike top 5 trips in Manhattan, we discovered during rush hours the most traffic trends happened on 186: Manhattan Penn Station/Madison Sq West, drop off location is 234. The taxi top pickup locations are like 107 Manhattan Gramercy,170 Manhattan Murray Hill, 162 Manhattan Midtown East. It helps the taxi commuters in Midtown East and NYPD's Traffic Enforcement Agents to commuted and managed in a more efficient way. The bike top pickup locations are 43 Manhattan Central Park, which not similarly as our hypothesis for the train station. The rest are 68 East Chelsea, and 164 Midtown South. It helps the Citibike commuters near uptown east, central park area and Citibike management staffs to commuted and managed the Citbike stations in a more efficient way.

In the Traffic Weather Analytics, we found that During rush hours, rain doesn't have too much impact on number of traffic trips per day. The number of Taxi on sunny day is 22,532 trips per day, on rainy date is 22,061 per day. The number of Citibike on sunny day is 5842 per day, on rainy day is 5803 per day. However, snow has a huge impact on number of traffic trips. The number of Taxi on snowy day is 10794 per day, the number of Citibike on snowy day is 1715 per day. The success rate for choosing the right transportation tool varies greatly due to different weather patterns. Therefore, we need to combine the weather condition to predict the average trip duration for publics.

From the special day analysis, we discovered that the peak of taxi and Citibike happened on May 2018. The lowest for taxi and Citibike happened on the same day Jan-04 2018. We researched and found that 'Bomb cyclone' winter storm brought snow, fierce winds to New York State. This discovery provides helpful suggestions to NYPD's Traffic Enforcement Agent, TLC, and Citibike management about the special weather/events which would cause huge influence on the traffic.

To check the accuracy of this analytics, we must argue the correctness of our filtered data, analytic code, and then based on the healthiness and correctness of the data and program, we can eventually argue for accuracies of the analytics result. One verification model has been applied for each aspect.

Although we have implemented data filtering algorithms targeting extreme data and potentially invalid data as mentioned in the previous sections, we would still have to verify the remaining data after filtering has a logical distribution and to avoid potential over filtering. Therefore, we have calculated the standard deviation: $\sigma_{Filtered\_Data} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ for data prepared for our analytic. The result shows data points different within 3 minutes from average value lies within ± 1.7 standard divisions. This represents the data is healthy concentrating within an acceptable range.

When impala performs the prediction time calculation, it iterates over more than ten thousand of data. It can be quite difficult to examine the correctness of our code from the execution result from impala due to the large volume of data. To show the program execution result is correct, we designed a primitive verification model named random sampling from uniform distribution. We decided to extract 30 data from Citibike data to be analyzed. For all data $n \in N$, each data point has the same probability to be selected as the verification sample. Then we compare the execution result from impala against the average calculation from a regular Python program that we implemented based on the selected data points. Our verification step proofs the SQL code that has been implemented for impala is fully function as expected.

The most critical part of the analytics is to verify the goodness of the analytic result. In order to do so, we have decided to use the path-based model as previously mentioned in the Related Work section to verify the consistency and accuracy of our data.

The path-based prediction approach focuses on calculating the travel time when a vehicle passes a particular path. To setup the model, we have randomly selected any pairs TLC standard zones that is not adjacent from each other, denoted as α and β. Treat each zone as an independent set. Then we hand pick a special middle zone γ based on the geography of New York City such that $\delta(\alpha, \beta) = \delta(\alpha, \gamma) + \delta(\gamma, \beta)$ to ensure all regular traffics from zone α to β passes zone γ. For instance, if one person departs from Greenwich Village North to Alphabet City, then a reasonable traffic route would definitely pass through East village, and so the data is correct. Hence, if the predicted time from Greenwich Village North to Alphabet City is consistent with the sum of prediction time from Greenwich Village North to East village and prediction time from East village to Alphabet City. The benefit of using path-based model is it is sensitive to the traffic conditions, so we can make zero assumptions of the road conditions. Continue from the examples above, if there were heavy congestions from East village to Alphabet City, then the additional time it takes for the trip would also show up on the trip from Greenwich Village North to Alphabet City during the same period of a day, since the former is a sub route of the later. Since the prediction time is extracted and generated independently for different keys (based on departure location and destinations), this technique allows us to cross reference the accuracy and consistency for one data based on the others. Each verification point associates with a deviation value from our predicted time to the sum of path-based prediction time. Based on the Central Limit Theorem, selecting 30 verification models each for FHV dataset guarantees we can have a Gaussian distributed data. The closer of the varication data mean to zero means the better consistency and accuracy of analyzed data.

Among the 30 models we have in the goodness checking step using the exact starting location and ending location in the above example, the overall mean deviation from path-based model is equal to 2.27 minutes from original prediction, which is acceptable considering it is roughly around 16 percent of prediction time difference based on the distance between the pickup and drop-off zone. This verification step reflects the analytic data is highly consistent. During the modeling of verification data points, we have noticed some outliers that is a little bit far from where most data points concentrated, and this case occurs on both directions. This may due to some extreme cases such as a person started from the furthest location from the destination within the departure zone and also stopped at the furthest location from the departure zone, so the trip has the maximum length possible between these two zones, and the vice versa cases with the minimum length possible between two zones. Based on the fact that our overall path-based model sum of verification data is close to the prediction data, the weight of outliers would gradually been reduced as the dataset gets larger, and so we believe the path-based model has demonstrated the correctness of this traffic analytic.

ACKNOWLEDGMENT

We would like to express our appreciation to Professor Suzanne McIntosh for providing us a wonderful series of lectures regarding the fundamental knowledge of big data analytics. In addition, we also want to thank all parties that makes data available for us to conduct this study. This study cannot be conducted without all of their efforts.

REFERENCES

1. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
2. A. Gates. Programming Pig. O'Reilly Media Inc.,Sebastopol, CA, October 2011.
3. J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implemenation, 2004.
4. 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.
5. A. Kurkcu and K. Ozbay, "Estimating pedestrian densities, wait times, and flows with Wi-Fi and Bluetooth sensors," Transp. Res. Rec., J. Transp. Res. Board, vol. 2644, pp. 72–82, Jul. 2017.
6. S.I.J. Chien, C.M. Kuchipudi, "Dynamic travel time prediction with real-time and historic data" Journal of Transportation Engineering, 129 (6) (2003), pp. 608-616
7. Chun-Hsin Wu ; Jan-Ming Ho ; D.T. Lee "Travel-time prediction with support vector regression" IEEE Transactions on Intelligent Transportation Systems, Volume 5, Issue 4, pp. 276-281, Dec.06 2004
8. ChungParkSo, YoungSohn, "An optimization approach for the placement of bicycle-sharing stations to reduce short car trips: An application to the city of Seoul" Transportation Research Part A: Policy and Practice, vol.105, pp. 154-166, Nov.2017.
9. Raymond Gerte, Karthik C. Konduri, Naveen Eluru "Is There a Limit to Adoption of Dynamic Ridesharing Systems? Evidence from Analysis of Uber Demand Data from New York City" ABJ70: Committee on Artificial Intelligence and Advanced Computing Applications, Aug. 2017.
10. Citibike Raw Data, https://s3.amazonaws.com/tripdata/index.html, Dec 2018