System analysis and Customer Segmentation of Citi-Bike in New York City

Vignesh Ramesh (vr839)
Dept. of Computer Science
New York University

Guruprasad Srinivasamurthy (gs2671)
Dept. of Computer Science
New York University

Jay M Patel (jmp840)
Dept. of Computer Science
New York University

Github Repo:
https://github.com/Vignesh6v/CitiBike_Data_Miner

Google Doc:
https://docs.google.com/document/d/1VD2mECZXfceuVZ6Mmi2i2W8eqU5oefoPNvCJ6lch5gM/edit?usp=sharing

System analysis and Customer Segmentation of Citi-Bike in New York City

Vignesh Ramesh (vr839)
Dept. of Computer Science
New York University

Guruprasad Srinivasamurthy (gs2671)
Dept. of Computer Science
New York University

Jay M Patel (jmp840)
Dept. of Computer Science
New York University

Abstract -

With the existing wealth of data pertaining to Citi Bike users in New York City, it is possible to identify, segment and categorize the customer domain based on several factors such as Age and Sex. We can analyse how different weather conditions and seasons affect the frequency and revenue. It is also possible to identify hot-spot locations and peak-demand hours, which can help the company to better manage the demand and supply. Identifying the hot-spots at peak hours would help the company to understand the market demand and thereby increase the bike availability at these hot-spots during peak hours will help in driving more revenue.

Keywords—CitiBikes, customer segmentation, marketing analytics, availability rebalancing

I. Introduction

Marketing analytics is a process to measure, manage and analyze marketing performance to optimize return on investment (ROI) and maximize its effectiveness. Beyond the obvious sales and lead generation applications, marketing analytics can offer profound insights into customer preferences and trends. Customer segmentation and categorization can significantly help any company to identify their prime and weak customer domains.

Citi Bike, a public bicycle sharing system that serves parts of New York City, is the largest bike-sharing program in the United States. With the existing wealth of data pertaining to Citi Bike users in New York City, it is possible to identify, segment and categorize the customer domain based on several factors such as Age and Sex, which will help the company to determine their potential customers that can be targeted and also the weak customer domain which needs to be improved. The prime objective of this project is to identify long-run customers and how the subway and yellow taxi takes the Citi bike customers. Also, potential audience that can be targeted to increase the company's customer base a and drive more revenue. Some of the specific targets of this project are to identify the percentage of men and women users in both the subscriber and casual usage category. The membership database can be coupled with the trip data to determine the mean and median age of Citi-bike users. This information will further help in customer segmentation and profiling to identify potential new customers who can be targeted. We are also interested in identifying the locations where usage of the Citi Bike is high and the popular routes in New York City. Other analytics objectives include the average trip duration and distance which will give useful insights in bike-availability rebalancing. Also, finding out how the customers are adapting to the concept of bike sharing.

Other deliverables of this project would be to identify hot-spot locations and peak-demand hours, which can be crucial information that can help the company to better, manage their business supply, which will also substantially help more customers. Identifying the hot-spots at peak hours would help the business to understand the market demand and thereby increase the bike availability at these hot-spots during peak hours will assist in driving more revenue. Our design would rely largely upon existing Citi Bike trip histories data and Citi Bike Daily Ridership and Membership Data. The impact of weather on the Citi Bike Ridership is also measured.

II. Motivation

In recent years, bike-sharing programs have surged in popularity, doubling worldwide since 2012. The NYC Citi Bike system with its 90,000 annual users consumes over 20-40,000 trips per day which result in load imbalance affecting the overall system health and thereby running the entire business with seamless customer service and bike availability. From the analytics, it is observed that different neighborhood of the city exhibits distinct behavior and at various times of the day. This depends on the several factors such as burrow nature which can be residential or business, subway availability, people's behavior and other factors. Our analytics on Bike availability and trip data will better help us understand these factors and thereby identify locations and time-slots which will require more supply due to peak demand and will also help us narrow down locations which have abundance availability throughout the day, thereby manage the demand-supply ratio for a seamless healthy system.

III. Related Work

Analyzing the data available from Citi-Bike users, it is seen that in the U.S. there 90,000 annual users with an average of 20-40,000 trips per day. About 5.5 million trips were observed from July 1, 2013, to February 28, 2014. It was noted that the mean distance traveled by the users using Citi-Bike was 1.81 km whereas the median distance traveled was 1.42 km resulting in the observation that most trips are less than 1 mile [4]. With the data on the availability of the number of bikes at any station at any time, it was found that most stations were 57% full at most times. To analyze the current status of the system, the stations which the ratio of the number of available bikes and the total capacity of the station around 80% were termed as "starved" and the ones with 20% were termed "congested". This helps in identifying the stations which have most demands at a particular time interval and helps in balancing the bike demand-supply throughout the day.

It was also observed without vehicle transportation to move the bikes to keep the demand-supply ratio in control has proven to be very effective. Some of the significant findings from this research are that local re-routing can be a very effective method to increase bike availability and that offering incentives to users using these re-routes can improve the overall system health. Another observation is that the New York City's bike sharing program suffers from global and local imbalance [4]. Also, the other main cause for the imbalance is due to people's usage of Citi bikes to travel from home to their workplace and hence most the bikes are accumulated near the work zone of the city especially Midtown and a very few are available near the residential areas.

To solve the imbalance [5], the initial step is analyzing the given data and determine the patterns of the bike usage and the how long the bike is idle at a particular station, how much bikes are collected during the rush hours at the hot spots locations. Then a concept of integer programming is used to remove the imbalance, where all the routes are represented as a point in the space and then repeated calculations are performed to eliminate the unnecessary points. Sometimes, the optimal solution requires a truck full of bikes to be transferred.

A major issue with Bike Sharing Systems (BSS) is the unbalanced distribution of space, time and number of the bikes among the stations. Bike Sharing systems are mostly used for medium-short distances and one-way trips, and this causes an unbalanced distribution of the bikes over a time and consequently to the increase of the probability to find a full or empty station. The BSS reallocation can be done either during the night when the bikes demand is less or during the day when the bikes distribution among the stations changes frequently due to the higher demand level. The method for the purpose mentioned above is Fuzzy Inference System [1] which minimizes the vehicles repositioning costs by keeping a high-level user's satisfaction, assuming that it increases the probability to find an available bike or a free docking point in any station at any time.

The problem is to relocate the bicycles from overcrowded station to the stations with a shortage of bikes and to predict the future demand; a Bayesian network can be used. The relocation paths are computed by solving a Traveling Salesman Problem as an optimization problem and lost users costs are also calculated. The solution algorithm is based on a Branch and Bound procedure where the outputs are the optimal relocation matrix and the optimal relocation path and bike distribution among the stations. This method is modular [2] so that it can be extended to wider and real sized actual systems and can also be used for real-time management and the strategic design that is to determine the optimal layout of the BSS.

The most important factor for a success of the bike sharing program is the location of the bike stations and their relation to trip demand [2]. For the approval of a user, the distance between station and origin and destination of user bike trips should be small, and further the distance between the stations should be suited with the best interest of bike transportation. Juan Carlos proposes a GIS (Geographic Information System) based method to find the ideal locations for stations and scrutinize the accessibility of each station [2].

The method mentioned by Juan Carlos [2] has four stages for finding the optimal station location in a bike-sharing program. First, it is important to learn about the distribution of the future user demand. The spatial distribution of demand is the most important part in optimal location models and produced by creating a layer of points which contains population and employment associated with each building number and building a layer of polygons which includes the number of trips generated. The number of inhabitants in each building multiplied with the ratio of trips per inhabitant in the building's transport zone gives the number of trips on building basis. This could help in inferring about zones that attract a high number of trips per used for medium-short distances and one-way trips, and this job (for instance, banking). This data is used in building kernel density maps and show the spatial distribution of bike station demand. Secondly, the location location-allocation models are calculated using candidate locations for bike-stations and at the demand point data. The locations that are considered have minimal impedance and maximum coverage. Hence from this stage, candidate locations, the number of stations to be located and the solution answers are chosen. Once bike-station location has been derived, the characteristics like station capacity, distribution of bicycles at each station about travel demand, accessibility from bike stations are described. The final stage involves the analysis of station use with regards to accessibility to potential destinations. Optimal methods for locating new bike stations have not been developed to date including GIS. Yet, GIS is still a useful tool for developing methods for bike station location. The limitations of this model include consideration of only workdays, the presence of localities with less population and scare jobs but may attract a considerable number of trips (e.g., large parks) and the presence of isolated stations due to maximizing coverage approach.

The other important goal for a bike sharing provider is to ensure high bike availability, is to satisfy customers but tends to be difficult because customer trips are highly dynamic, and redistributing bikes is costly. Recent studies show that rides have similarity in regards with spatiotemporal dependencies in bike usage and one-way use, and short trip causes imbalances in the spatial distribution of bikes. With the help of data mining, ride data can be analyzed and processed to support station location decisions [3]. First, preprocessing is done by gathering ride and customer data as well as location factors and is properly cleaned and selected to be suitable for mining. Secondly, data mining is done with cluster analysis on station activity, ride patterns and customer base segmentation. Then post processing is performed, where the solutions of cluster analysis are validated by visualization with a geographical information system for the cluster interpretation.

IV. Design

The Citi Bike membership dataset has basic customer information such as age, date of birth and gender. By parsing these values, we were able to extract information such as the number of subscribers who take the annual plan and the number of customers who take the 24-hour or 7-day passes. In the reducer phase, we sorted out this information by age thereby generating further analysis on age-gender wise customer segmentation. By doing so, we were able to identify the most popular and least popular customer segments of Citi Bike users in New York City.

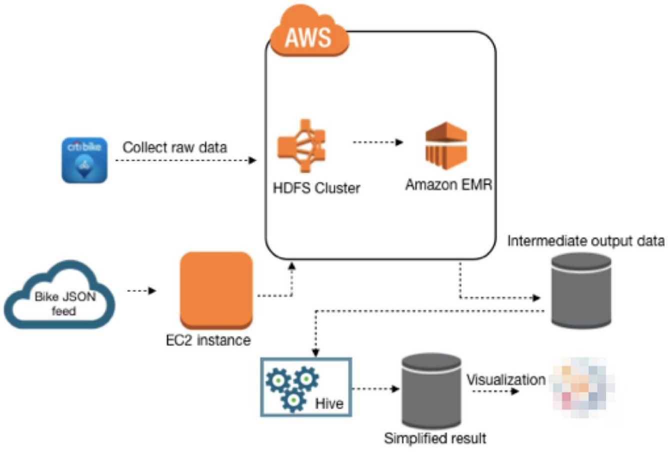


Fig 1. Design Architecture Diagram

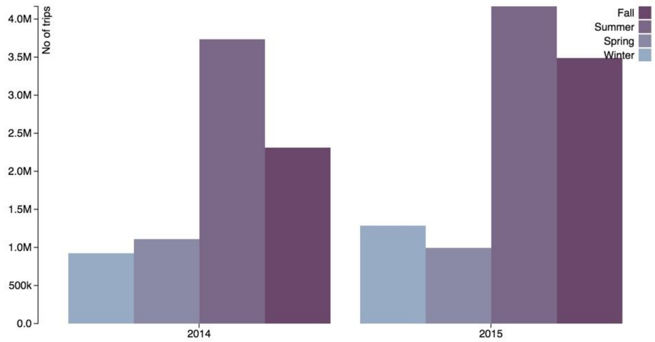
On considering the scalability and fault tolerance, we have designed the cluster with three larger mapper and two reducers. Also, by cleaning the entire dataset, we reduce the probability of failure.

To assist in our processing of data we have used third party modules like matplotlib, rtree, numpy, geopy and shapefile in our mapper and reducer programs. For faster processing, we stored intermediate results in S3 and used these as input in other mapreduce programs. To portray the generated results in a concise and precise manner, we have used D3 in generating charts that use the mapreduce results. We also used Google Maps to plot latitude and longitude.

V. Results

1. *Season-Wise Trip Segmentation*

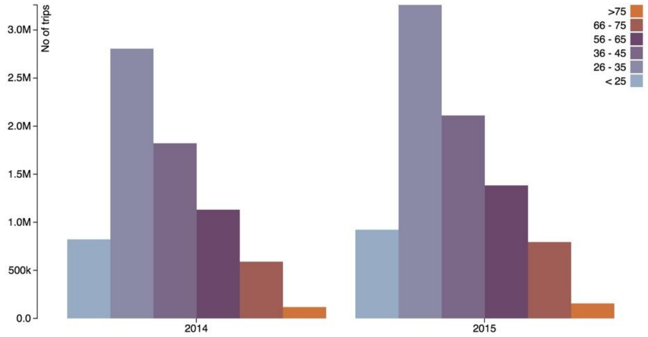
Data: We have analysed trips across all stations for the entire year of 2014 and 2015.



Results: We see a general trend where there are maximum trips happening in the summer than compared to other seasons. However, we also see a significant increase in the number of trips in the fall of 2015 when compared with the same in 2014

2. *Age-wise Customer Segmentation*

Data: We have analysed trips across all stations for the entire year of 2014 and 2015.



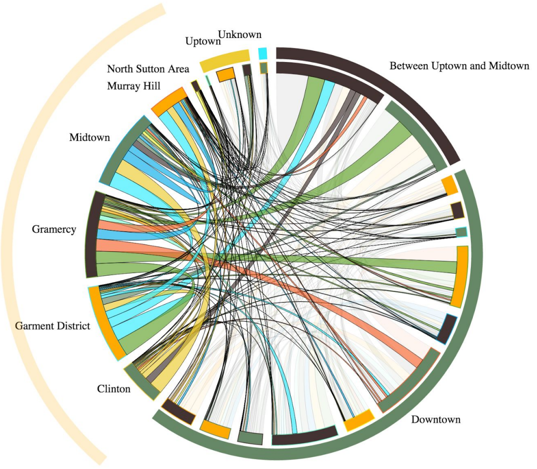
Results: We see that people in the age group of 26-35 are the most common users. We also see an increase in the number of customers in this age group in 2015 when compared to 2014. With regards to other age-groups, we see a stepwise decrease in the number of customers when compared with increasing age groups.

3. *Trip Distribution across different neighbourhood*

Data: We have analysed trips that have happened in the month of January 2015. We have

- Consolidated trips across all stations within a neighbourhood.
- Considered only those that start in a neighbourhood and end in another one.
- Considered trips only within Manhattan.

Results:

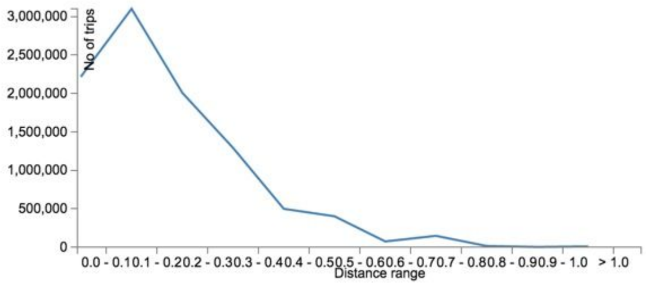


We can identify the popular neighbourhood destinations from a particular neighbourhood.

4. *Trip distribution over proximity to Subway stations*

Data: We have analysed trips that have happened in the month of January 2015.

Results:



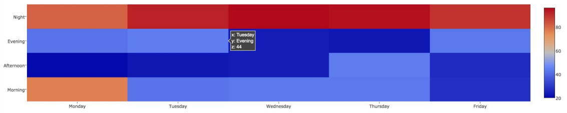
The x-axis indicates the distance between the citibike station and its closest subway station. The number of trips are consolidated across all the stations that have subway stations at the same distance. We see that lesser the distance between subways and bike stations, the more the number of trips.

5. *Daily Trip Distribution over the time of the day*

Data: We have analysed trips for the year 2014 and 2015.

- Only considered trips on weekdays.
- Morning:4:00AM-11:00AM, Afternoon:11:00AM -3:00PM, Evening:3:00PM- 7:00 PM, Night: 7:00 PM - 4:00 AM

Results:

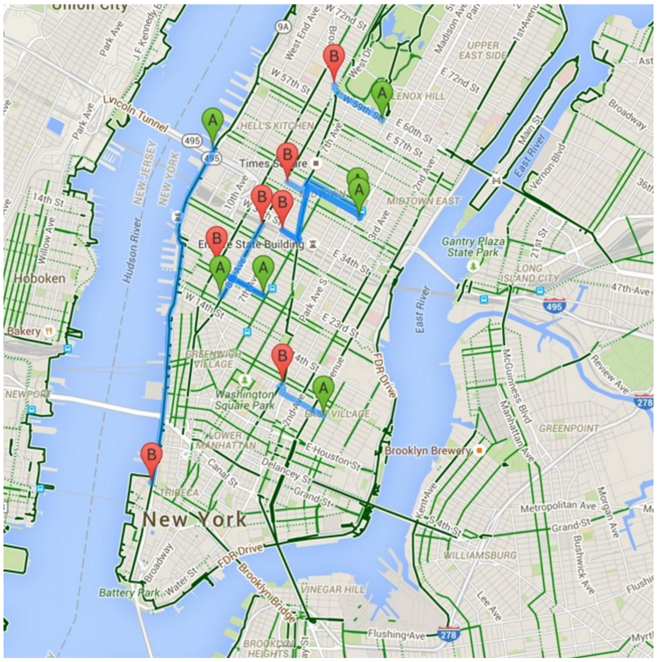


We see that more trips happen in the later part of the day than in the morning. We can also see significant increase in the number of trips on Monday mornings when compared to mornings of other days.

6. Hot Routes

Data: We have analysed data for the entire year of 2014 and 2015 and chosen the top 5 routes of each year.

Results:



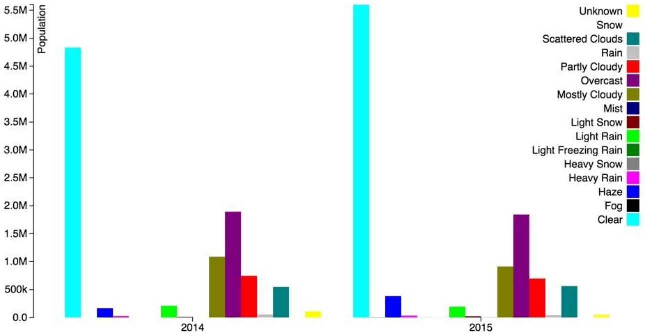
The thicker blue line indicates that this particular route was in the top 5 routes both i the year 2014 and 2015. We see that majority of the routes are in manhattan, particularly midtown.

7. Trip distribution across different weather conditions

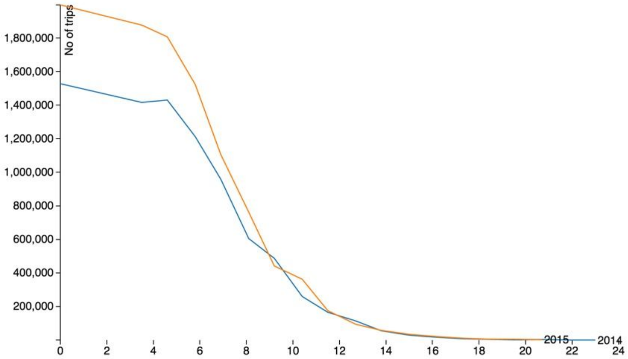
Data:We have analysed trips for the year 2014 and 2015. We have also used weather data for the entire year of 2014 and 2015.

Results:

a. Weather Conditions

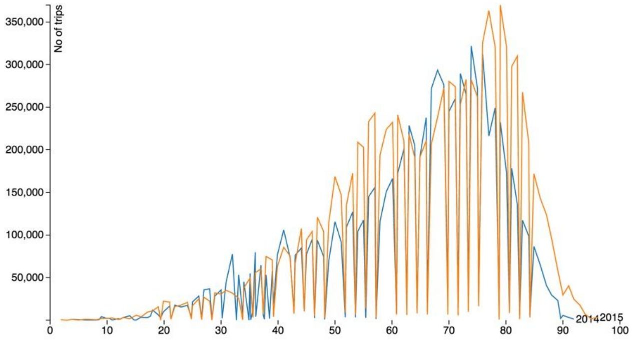


b. Wind



The wind speed in in miles per hour.

C. Temperature

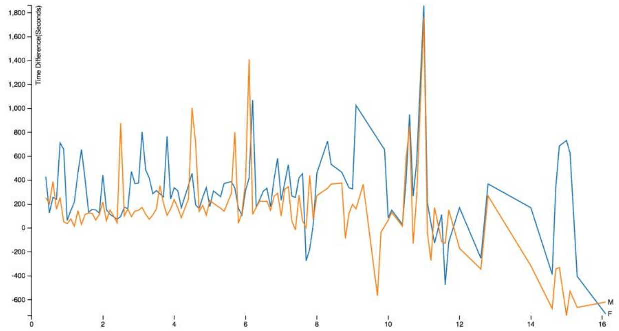


The temperature is in fahrenheit.

8. Extra time taken by customers to reach a destination

Data: We have analysed data for the entire year of 2014 and 2015

Results:

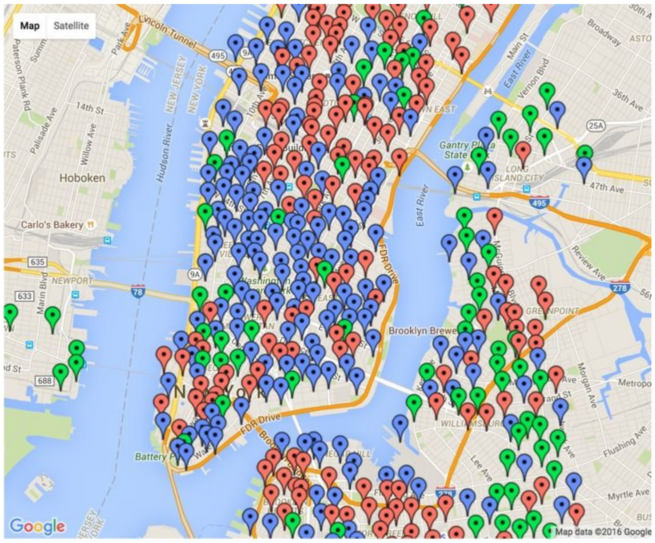


We see that on an average, customers take an extra 800 seconds more than what google estimates is the time needed to reach a destination from a particular source. Here, the blue line indicates female customers and the red one indicates male customers. Zero time difference indicates that the trip duration was exactly the time estimated by Google. The X-axis indicates the distance of the trips.

9. Bike Rebalancing

Data:We have analysed trips for the entire year of 2014 and 2015. We have considered bike remaining at the end of the day for our calculations.

Results:



Red stations indicates that there is a high probability of bike deficiency. Blue indicates high probability of Bike availability. Green station are those that have sufficient number of bikes.

10. Other results

a. We found the list of stations that are the most active.
b. We also found the amount of calories being burnt per month by New Yorkers by using Citi Bike.
c. We also found list of most used bikes. These help in identifying bikes that might need servicing.
d. We also found out the most likely destination of a trip which starts at a particular station.

VI. Future Work

With the membership data analysis, we were able to categorize and segment the users based on several factors such as Age, Sex. With this wealth of information, Citi-Bike could target their customers to drive more revenue and also convert their 'Customer' type users to 'Subscriber' type users since they are the only source with the member details such as Member ID, email and phone number which is essential for targeted advertising.

VII. Conclusion

The analytic insights generated from this project can help better make business decisions at Citi-Bike which will improve the overall quality of the service and drive more revenue through efficient rebalancing mechanisms and targeted advertising. With deep insights on the 'likely users' and their profiling, Citi-Bike can redesign their marketing strategies so to increase their customer base and drive more revenue.
The observations from this project prove the fact that understanding your current system behavior and user activities from the existing data will prove to be an effective mechanism for business scaling and expansion.

Contributions

As a group of three authors, each person has contributed to the distinct areas of the project. The author, Vignesh Ramesh has done the data clean part of the Citi Bike and Weather Dataset and also, did various basic stats from the data like age-wise segmentation, Trips based on weather, gender-wise segmentation, rebalancing, analyzed how efficient the bike sharing idea works by comparing the estimated time with the google map time calculations. The author, Guruprasad has contributed by gathering the neighbourhood and subway data set , cleaning them to remove outliers and processing it along with the results gathered previously to analyze how these factors are affecting the CitiBike business. The author, Jay Patel analyzed trips on the timely basis, bikes that are likely to attention for services, found the list of stable Citi Bike stations, geocoding for the Citi Bikes stations, data smoothing. Also, provided a new way to look at the analyzed data by visualization using D3.

Acknowledgment

This material is based on the data and the APIs provided by Citi-Bike without whose efforts this analytic project would not be possible. We would also like to acknowledge Amazon AWS for providing us with resources to run our analytics on their cloud instances. Also, our sincere gratitude to Professor. Juliana Freire for her continued support and guidance throughout our work.

References

[1] Caggiani, Leonardo and Ottomanelli, Michele. A Modular Soft Computing based Method for Vehicles Repositioning in Bike-sharing Systems

[2] Juan Carlos García-Palomares, Javier Gutiérrez, Marta Latorre Optimizing the location of stations in bike-sharing programs: A GIS approachx

[3] Patrick Vogela, Torsten Greisera, Dirk Christian Mattfelda. Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns

[4] Jahaziel Guzman, Donald Hanson II, Franky Rodriguez. Self-Balancing Bikes: Locally Re-routing Users to Improve the Flow of Bike Share Programs

[5] Eoin O'Mahony , David B. Shmoys. Data Analysis and Optimization for (Citi)Bike Sharing

Log Reports

Started to analyze the business of Citi Bike by using Citi Bike ridership data. Initially, started with automating the progress of getting the Citi Bike membership dataset and the weather data for two years 2014 and 2015. Due to the high volume of data, data cleaning process becomes tedious. First, started cleaning by each and every attribute of the dataset and then with list of the available Citi Bike station JSON file, we were able to validate the data. Likewise, after cleaning the weather data by checking it the condition of weather on the hourly basis and validating it with timing condition and with few samples of standard weather data from known source.

Thus, after collecting the cleaned data. We began to do our analysis, which has been outlined in the following table.

Date	Progress
April 4	Started with the Idea to choose Weather and Citi Bike data Set
April 11	Produced the automation of fetching both the data set
April 15	Began the process of cleaning the data set
April 18	Understood the dataset based on various attributes
April 22	Started with the basic stats from the cleaned data set
April 24	Faces problems in merging two different type of data set
April 27	After merging, produce the results based on various scheme.
May 1	Produced the status on what we have worked with the dataset
May 4	Spatial analysis were done with neighborhood shape files
May 7	Analysis of how subway affecting the business and list of popular stations for rebalancing
May 10	Used D3 to project the analyzed data in various forms.
May 13	A Final report was made with all the outcomes of the ridership data set analysis.

Status Report For Citi Bike

➢ Done:

1. Basic Citi-bike report, which includes the month-wise rides, Number of active stations, Total number of bikes and customers.
2. As a part of data clean, the trips that have the trip duration less than 60 seconds and more than 8 hours has been removed, which are possible errors in tracking the trips.
3. The Membership dataset has basic customer information such as age, date of birth and gender. By parsing these values from the dataset, we were able to extract information such age, gender and age-gender wise customer segmentation.
4. We extracted the information such as the number customers who have taken 3 different services that Citi-Bike offer viz Annual Memberships, 7-day passes, and 24-hour passes.
5. Identified the peak demand hour for each station which will considerably help in load rebalancing of Citi-bikes.
6. How the weather has affected the business of the Citi-Bikes.
7. Citi-bike availability distribution of a particular station.
8. Citi-bike availability distribution of a particular station.
9. List of bikes that needs maintenance check based on the usages.
10. '*Season of Citi-bike*', where you can target the potential customer.

➢ To Do:

1. Need to find how the yellow taxi will be affecting the Citi Bike business.
2. Is it good to take yellow taxi or Citi Bike in specific routes.

➢ Issues/Comments:

1. Understanding how to merge the results of membership data with the NYC weather data.
2. Found hard to find the unique number of users (Customers and Subscribers) of the citi bike.
3. Even after cleaning the trip data set we found few false positives, which were showing some erratic values.
4. Unable to figure, is the same person is using the pass to ride. Because, we found we riders have an age of more than 150, and also in some trips, few rider ride more than 8 hours.
5. As we were unable to find the slopes in New York City, we faced problem in finding the probability of people take more bikes for downslope than the uphill.
6. Unable to calculate the revenue generated by the Citi Bike, as there was no pricing details.
7. Unable to categorize the customer based on their passes they have purchased.