New York Yellow Taxi Analysis

Anjali Baldawa
925-577-5193
Department of Information System
California State University
Los Angeles
abaldaw@calstatela.edu

Dhruvi Patel
562-486-9904
Department of Information System
California State University
Los Angeles
dpatel86@calstatela.edu

Digesh Kansara
562-450-9556
Department of Information System
California State University
Los Angeles
dkansar@calstatela.edu

Abstract: As urban residents who are unable to drive themselves, their transportation within the city is mainly public transportation including buses, taxis, subways and railways which is very popular among people in recent years. The above three modes of transportation are important components of a city, and each has its own characteristics. New York City is a city with a complex transportation system and a large residential community. Its traffic problems are most obvious and prominent. Focusing on the analysis of the choice of transportation modes of residents in New York City and the analysis of traffic characteristics are very typical. This particular project highlights, the prevailing focus on the dataset of NYC taxi trips and fare.Apache Hive and Hadoop used for analysis of data set and tableau has been used for visualization. This data can be analysed for several purposes like avoiding traffic, lower rate where services are not functioning more frequency than a cab on crown location and many more. This information can be used by numerous authorities and industries for their own purpose. Government official can use this data to deliver supplementary public transport service. The company like Uber can use this data for their own taxi service[1].

1. Introduction

Taxis are a common public transportation in the city. Unlike public transportation such as buses and subways, taxi lines are not fixed and random. The density of taxis varies greatly from region to region. It is non-commuting and frequent. The main means of transportation, and its travel path, time, and get-off point information are closely related to human activities, which can better reflect the behavior patterns of urban residents [1].

There are many misperceptions in [2] TLC (Taxi and Limousine Commission) of the New York city, how the taxi services should be disseminated in the city that too based on certain assumptions, like most pickups, time, distance, airport timings. In order to provide very good taxi service and plan for effective integration in city transportation system, it's very important to analyse the demand for taxi transportation. The dataset provides the relating information such as where taxis are used, when taxis are used and factors which tend public to use taxis as divergent to other modes of transportation.

The core objective of this paper is to analyse the factors for demand for taxis, to find the most pickups, drop-offs of public based on their location, time of most traffic and how to overcome the needs of the public. Explicitly, the key contributions of this paper are as follows:

- Primitively to the best of our knowledge, we conduct the analysis that recommends the busiest day of the week based on trip count and total cab booked on the day of the week. Analysis was done with the help of Hive QL and Tableau.
- To Analysis whether long distance or short distance trips are preferred during which day of the week and which hour of the day, we used total trip count, distance and day of the week.
- To quantify the Total Trip count by Time of Day. we used Hive to analyse it. The third analysis was based on the total trip count for a day per hour. For executing such intricate query, we used Hive which is a part of Hadoop ecosystem. The final output consists of the total trip count for every hour of the day.
- Ultimately, we analysed the payment mode for fare and tip. This analysis consists of both payment type and tip amount to check which payment mode is widely used.
- Furthermore, to achieve our goal, we proposed our subsequent analysis i.e. Analysis on Region. Here we accumulate information which was allied with location like PickUp Latitude, PickUp Longitude and total trip count. Expending PickUp Latitude and PickUp Longitude we appraise the count of taxi booked from that particular location. This will help to provide more taxis on the most PickUp location and so on

2. Related Works

Identifying travel patterns from recorded taxi trips is important to understand human mobility and transportation planning. Existing approaches to trip purpose identification include traditional diary/phone based travel survey and more recently, GPS based travel survey as almost all taxicabs in cities of the developed countries have been equipped with GPS devices and different types of trip related information are recorded. Now we are going to fetch the section focusing Technologies which are used in

1

analysing huge dataset. The complete analysis is analyzed using BigData with Hadoop and visualization using tableau. Also, this section focus on earlier big data projects on NYC taxi dataset, namely to optimize taxi usage, and on big data infrastructures and applications for transport data events.

- Transdec (Demiryurek et al. 2010) is a project of the University of California to create a big data infrastructure adapted to transport. It's built on three tiers comparable to the MVC (Model, View,Controller) model for transport data.

- (Yuan et[2] al. 2013), (Ge et al. 2010), (Lee et al. 2004)worked a transport project to help taxi companies optimize their taxi usage. They work on optimising the odds of a client needing a taxi to meet an empty taxi, optimizing travel time from taxi to clients, based on historical data collected from running taxis.
- Apart from this there were certain analytical problem and visualization that we did in tableau , checking on the previous work none of the visualization work was done in tableau , it was either done using python or powerBI, thus we tried to visualize those analytical problems using tableau.

3. Analytic Problems

The statistics and analysis of the data set will be divided into two parts, the first is the data set from Trip Data and other data set is Fare Data. The Trip dataset contains data for the entire year from January 1, 2013 to December 31, 2013. Analyze the travel characteristics of the user. The second part is analysis of payment method, Tip amount, pickup date and many more.

3.1 Analysis for particular day of a week

This analysis will be on total trip count on each day of the week.The column of the graph represents a week and row represents trip count. From the Figure1, we can conclude that Monday is having the lowest number of booked trips, while Friday and Saturday have the highest number of trips in a week.

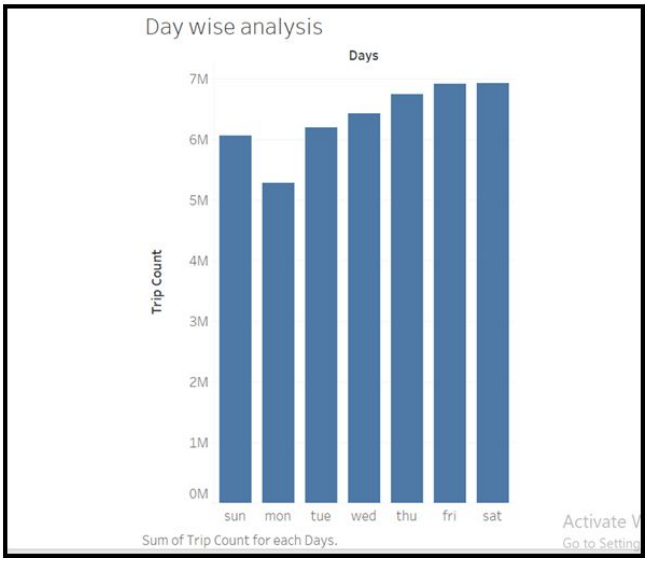

Figure 1. Day Wise Analysis

3.2 Analysis of trip count during three months on a particular day

The below graph shows the analysis of trip count during the three months on a particular day. The row represents the trip count and the columns represents dates of three months.The orange line in figure 2 represents the trip count on a particular day for the three months.It can be depicted that trip count for Monday is low and for Saturday it's the highest. Also from the graph we can conclude that on 21 January on the occasion of Martin Luther King day the trip count was low and thus assuming that trips are low on public holidays.

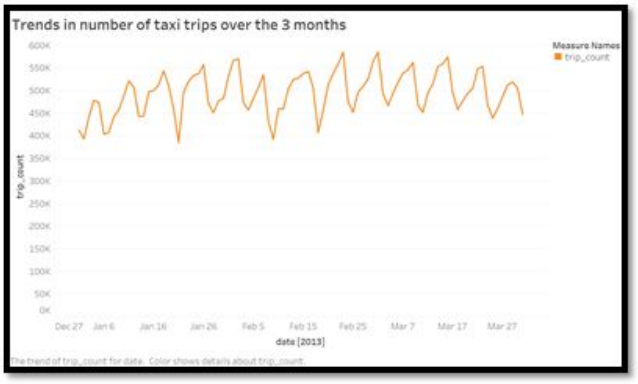

Figure 2. Analysis during three months

3.3 Analysis of trip count during each day for different time slots

In Figure 3 the column represents time (24-hour format) and row represents trip count. The figure depicts the trip count during each day of the week for different time slot. The graph is representing that the peak hours are higher during the night for all the days whereas in the morning the trip count is not that high as compared to the night .During

2

Saturday's and Sunday's the trip count is the highest during 12:00AM – 2.00 AM compared to other days in the week. It can be concluded that on weekdays, the trip count gradually start increasing from 5 AM and reaches peak at 9AM and there is a slight drop from 10 AM to 5PM(17:00) and again it reaches peak point at 6PM (18:00) to 8 PM (20:00).

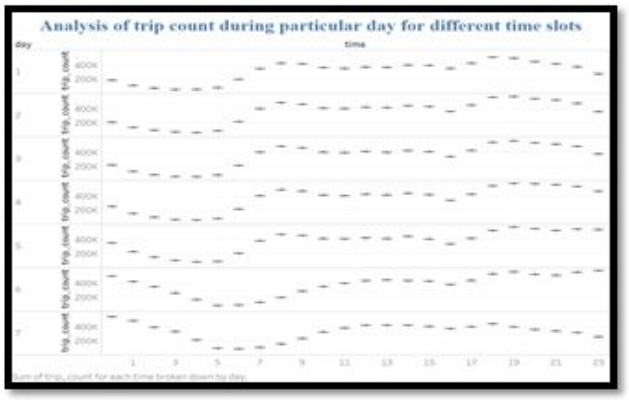

Figure 3. Analysis during each day for different time slots.

3.4 Analysis based on distance

This analysis will focus on the distance of a trip and total count of taxi for every hour of the day .The row represents the trip count and column represents the hours (24 Hour Format). We have divided distance based on miles, the value of miles is less than 40 is considered as short distance trip and the range between 40-150 is considered as long distance trip. From the figure, we can conclude that people prefer taking short distance trip early morning compared to long distance trip for weekdays.However, the trips for long distance and short distance for early mornings and late nights on weekends are usually high.[3] Fig 7 gives the comparison chart between long distance and short distance trip per hour for each day of the week.

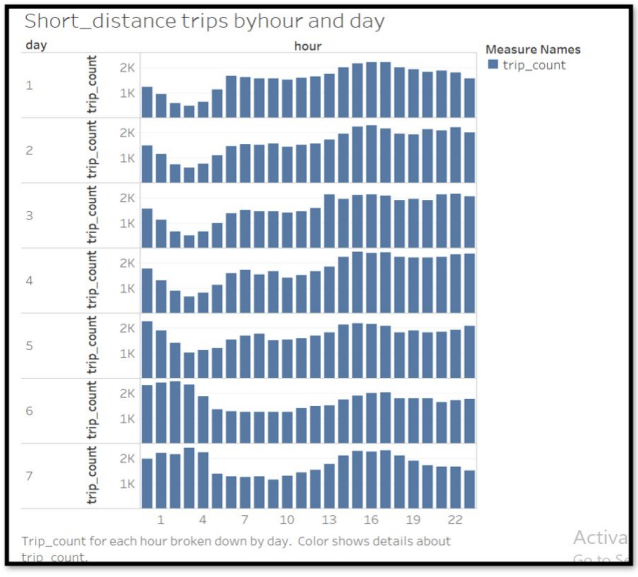

Figure 5. Analysis of Short Distance Trip

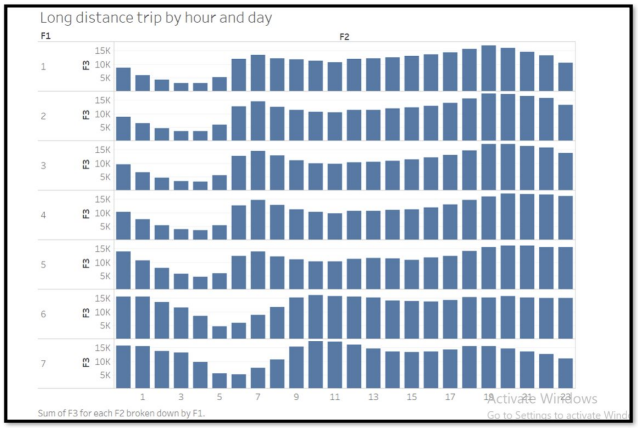

Figure 6. Analysis of Long Distance Trip

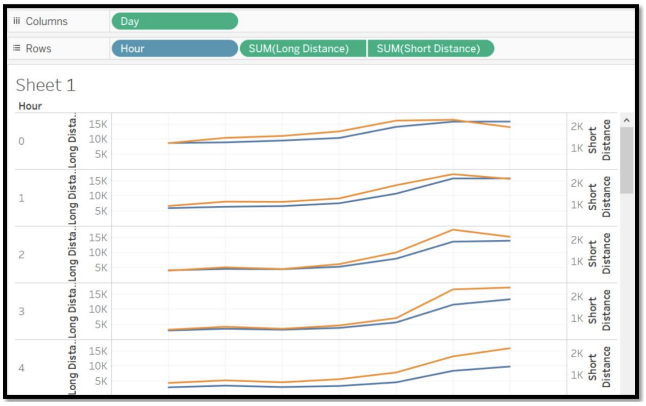

Figure 7. Comparison of Long Distance and Short Distance Trip

3.5 Different Type of Payment Analysis

The bar graph in the figure7 represents the payment method that is used by people while paying their fare amount. Through the graph the most convenient way of paying the fares is either through cash or through card. The blue color symbolizes the payment done through card and the orange color symbolizes the payment done through cash. By analyzing the graph it can be concluded that there is very little difference between the fares paid by cash and card whereas people prefer paying tip through card rather than cash.

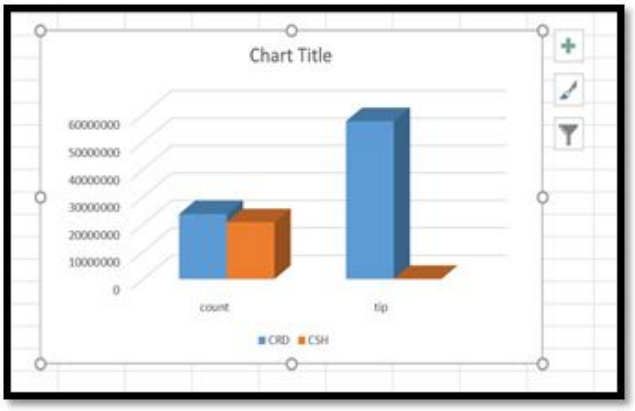

Figure 8. Analysis of Payment

[3]

3.6 Analysis based on time and location

This section will through some light on how we determine some complex analysis by using Hive. The hive is a part of Hadoop ecosystem which is a software that facilitates querying and managing large dataset using simple SQL like commands. It is built on top of the Hadoop. In this analysis we will determine average of total pickup by months for each day based on location. Fig 9 shows the Total Drop-offs by each Day of four months based on location.

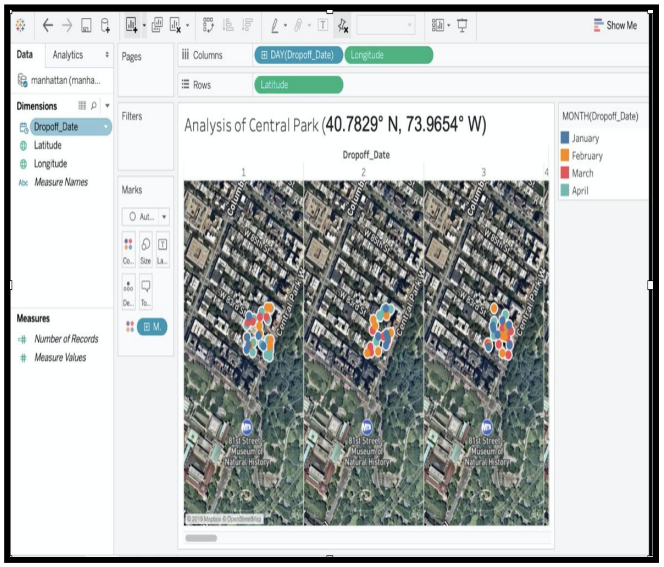

Figure 9. Total Drop-offs of each day for four months based on location.[4]

We have added another spatial analysis using latitude and longitude for a short distance trip (miles <40). In figure 10 the blue dots indicate the drop_off taxi for that particular location.

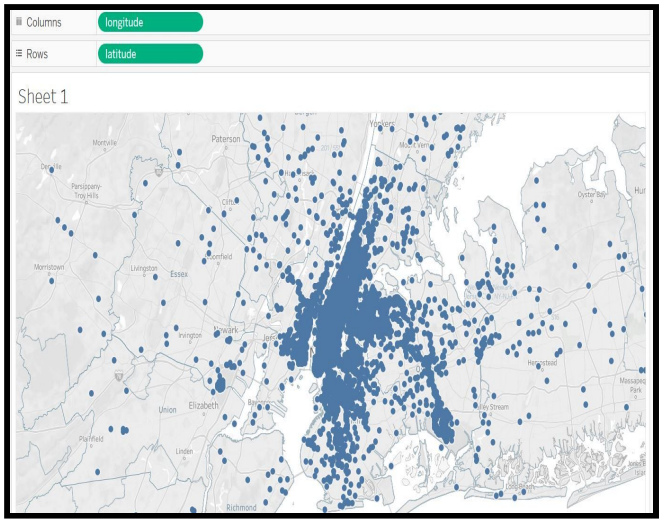

Figure 10. Total Drop-offs taxi for short distance trip (miles < 40).

4. Conclusion

In this assignment, we generated new system that ropes in the visual exploration of big origin-destination and spatiotemporal data. The most vital section of this project is a visual query model that sanctions the users to quickly select data slices and use them.According to this project we demonstrate how taxi trip count increases and decreases in a week and varies at a particular time of a day. This analysis can be used for making business decisions which can be further used as a key element for marketing. We had downloaded the dataset and uploaded to the HDFS , later the data was manipulated and analyzed in HDFS using Hadoop and visualization of the results was done in excel and tableau maps.Not only the analysis can be used in making the business decisions but it can also be used in industries for the near future to identify how many taxi are approximately used.

References

[1]CHENG Jing, LIU Jiajun, GAO Yong. Analyzing the Spatio-Temporal Characteristics of Beijing′s OD Trip Volume Based on Time Series Clustering Method[J]. Journal of Geoinformation Science , 2016, 18 (9): 1227-1239.

[2]NYC Taxi & Limousine Commission. http://www.nyc.gov/html/tlc/html/about/about.shtml.

K. Hwang, *Computer Arithmetic*, John Wiley, 1997.

[3]https://flrec.ifas.ufl.edu/geomatics/hochmair/pubs/NYC_Taxi_Hochmair_2016_TRR_DraftLayout.pdf

[4]https://www.google.com/search?q=nyc+taxi+analysis+ieee+research+paprs&oq=nyc+taxi+analysis+ieee+research+paprs&aqs=chrome..69i57j33l2.35123j1j7&sourceid=chrome&ie=UTF-8

[5] Demiryurek, U., Banaei-Kashani, F. & Shahabi, C., 2010. TransDec:A Spatiotemporal Query Processing Framework for Transportation Systems. IEEE, pp.1197–1200.

[6] Ge, Y. et al., 2010. An energy-efficient mobile recommender system. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’10. N