Scientific Research Publishing

American Journal of Operations Research, 2019, 9, 192-199
http://www.scirp.org/journal/ajor
ISSN Online: 2160-8849
ISSN Print: 2160-8830

Big Data Analytics of Taxi Operations in New York City

Yuxin Tang

Shanghai World Foreign Language Middle School, Shanghai, China
Email: tommy19991230@outlook.com

How to cite this paper: Tang, Y.X. (2019) Big Data Analytics of Taxi Operations in New York City. *American Journal of Operations Research*, **9**, 192-199.
https://doi.org/10.4236/ajor.2019.94012

Received: April 18, 2019
Accepted: July 28, 2019
Published: July 31, 2019

Copyright © 2019 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

As a global financial center, the transportation system in New York City (NYC) has always been studied from various aspects. Since 2009, NYC Taxi and Limousine Commission have made public the information on NYC taxi operations, offering an opportunity for detailed analysis. Thus, this research project investigates taxi operations in New York City based on big data analysis. The correlation between taxi operations and different types of weather, including precipitation, snow depth, and snowfall is discussed in this paper. The research also evaluates taxi trip distribution in each NTA area using Geopandas, and presents its density on an NYC map.

Keywords

Taxi Operations, Big Data Visualization and Analysis, Linear Regression

1. Introduction

As a global financial center, New York City is frequently studied by researchers, and its transportation has become an increasingly important topic. A large amount of data related to transportation released by NYC Taxi and Limousine Commission makes more sophisticated analysis possible. Using big data analysis to study taxi operations in the city of New York, this research paper explores the statistics of taxi's payment type, daily and monthly trend of taxi operation, its long-term trend, and the impact of weather. To process the data, econometrics is used to find out comprehensive results.

Chris Whong and Todd W. Schneider have conducted similar research on taxi operations in the past. In 2013, Whong studied 170 million taxi trips in NYC and collected information of each trip's tip, total payment, number of passengers, trip start point, and trip end point [1]. He then visualized these data to

show how each component of taxi operation had changed over time. This visualization enabled audience to observe taxi's movement directly and clearly in NYC. In 2017, Schneider showed the trend of amount of yellow taxi, green taxi and Uber car between June 2014 and June 2015 by comparing taxi and Uber trips in NYC [2] [3]. With graphs and diagrams, Schneider concluded that people in Queens prefer to use Uber than those in Manhattan.

In this research paper, the impact of different types of weather, including precipitation, snow depth, and snowfall, will be evaluated to determine factors that affect taxi operations. In addition, distribution of taxi trips in each NTA area defined by Geopandas will be studied, and its density will be shown on a plotted NYC map.

2. Data and Methods

This research uses data of taxi operations between 2009 to 2015 from NYC Taxi and Limousine Commission, with a focus on the newest data from year 2015 [4]. Data of weather information is extracted from observation of central park in NYC. Geometrical information is obtained from map of NYC.

This research requires the use of Python for programming, and programming cells are run on Jupyter notebook. Numpy, Pandas, Geopandas, Matplotlib are applied to process data. Specifically, Numpy and Pandas are used to analyze array and data frame data, Geopandas is used for geometry data, and Matplotlib is used to plot graphs. Linear regression and linear algebra are later used to determine the functional relation of the selected data.

3. Data Analysis and Description

First, basic information of taxi operations in January 2015 is studied. Several columns of information related to the topic, such as pickup time, are selected, and the raw data is then read into Jupyter notebook using "read.csv". The data is grouped by day since daily trips are the main targets. The result, as plotted in **Figure 1**, shows total pickups corresponding to each day in January and signals that there are fluctuations of daily trips, especially on 27th January, when trip amount decreases sharply by around 150 thousand, and increases later to 500 thousand on 30th January.

Figure 1 illustrates amount of daily total trips on y-axis and days in January on x-axis.

Next, the average amount of hourly trip is learned: data of trip amount of January is grouped by 24 hours and then divided by 31, as there are 31 days in January. Consequently, the result of average trip amount of each hour in January is obtained. As shown in **Figure 2**, which graphs the relation between daily hour and number of trips, the amount of taxi trips is the lowest at 5 a.m. It then rises over the day, and decreases gradually again from 8 p.m. The amount of trips increases sharply from 6 a.m. to 10 a.m. because people begin to go out for work or need to move across the city. The amount of taxi trips arrives at its peak at around 5 p.m., when people leave their workplace and go back home.

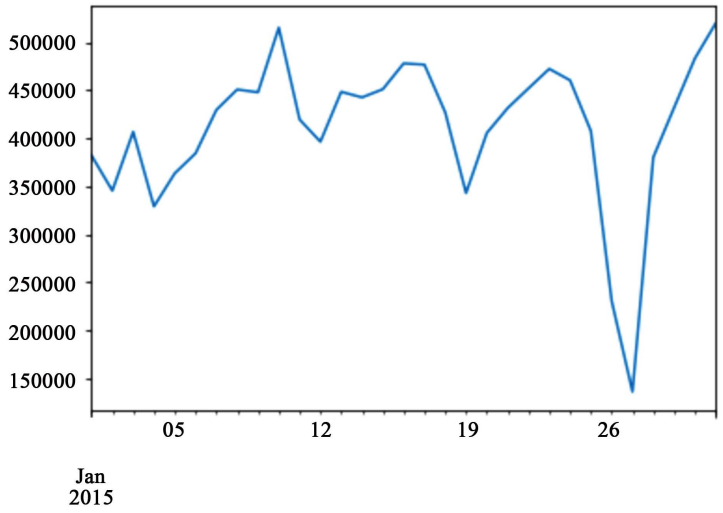

Figure 1. Daily pickup amount in January, 2015 in NYC.

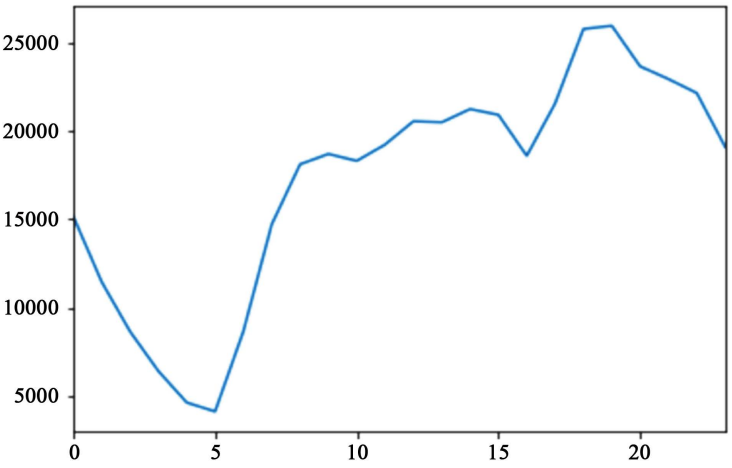

Figure 2. Average number of trips for each hour in January, 2015 in NYC.

Figure 2 shows the amount of average hourly trip on the y-axis and daily hour on the x-axis. At around 5 a.m., trip amount is at its lowest level during the day, and it then increases at around 7:30 a.m., when the rush hour begins in morning. The second increase in the trip amount during the day is another start of the rush hour, at around 5 p.m.

The impact of weather is then considered. Data of average snow depth in NYC, January 2015, is read into Jupyter notebook and is arranged by time, which corresponds to each day of January 2015. Linear regression is applied to build a functional relationship between snow depth and daily trips in January. Figure 3 is a graph showing that daily trip amount decreases gently with an increase in the snow depth in January 2015. The reason is probably that people prefer to stay at home instead of going out when there is snow on the ground.

In Figure 3, x-axis displays snow depth and y-axis displays the total amount of daily trips in January, 2015. The plots and lines show results of linear regression of snow depth against daily trips in that month.

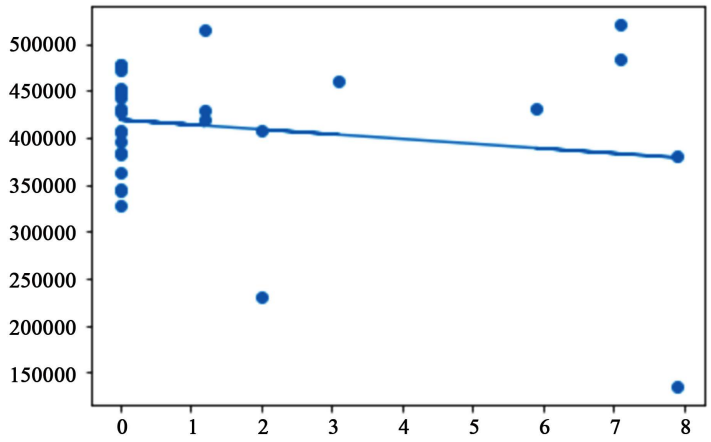

Figure 3. Relationship between snow depth and daily trip amount in January, 2015 in NYC.

Following the analysis of taxi operations in January 2015, data of the whole year is studied. Data of taxi trip operations from February to December of 2015 and data of weather are read into Jupyter notebook respectively. After daily trip amount is selected and combined to a data frame, linear regression is applied to show the relationship between snow depth and the amount of daily trips in the year of 2015. According to Figure 4, daily trips slightly increase with the rise in snow depth during the year. This result, which is opposite to that of the analysis in January 2015, signals that the previous statistics do not represent the situation of the whole year. The reason could be that January is the coldest month in winter, greatly different from other seasons. Based on the overall trend of 2015, it is found that more people need taxi rides when snow depth goes up.

Figure 4 illustrates daily trips (y-axis) against snow depth (x-axis) in 2015. According to the graph, the amount of daily trip rises gradually as snow depth is larger in 2015.

Using the same method, linear regression is applied to test the relationship between snowfall and the amount of daily trips in 2015. Figure 5 shows that trip amount slightly falls as snowfall goes up, which is opposite to the result gained from linear regression between snow depth and daily trips. The reason for such difference is probably that snow stays for a long time throughout the year in the city and therefore has less impact on people's lives than snowfall.

Figure 5 shows the relationship between daily trips (y-axis) and snowfall (x-axis) in year 2015. It can be observed from the graph that daily trip amount decreases when snowfall increases. The possible reason is that people prefer to stay at home instead of going out on snowy days.

Another liner regression is done to test the relationship between daily trips and precipitation in 2015. In Figure 6, x-axis represents precipitation, while y-axis represents the amount of daily trips in 2015. As illustrated in the graph, the amount of daily trips decreases gently as precipitation increases. The reason is probably that people chose to stay at home instead of going out on rainy days if not necessary.

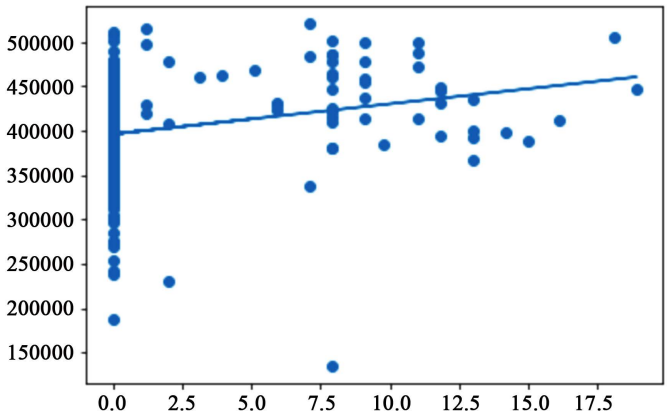

Figure 4. Relationship between snow depth and daily trips in NYC in 2015.

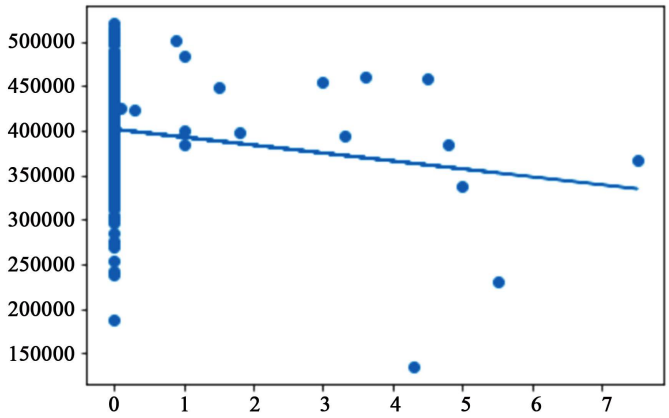

Figure 5. Relationship between snowfall and daily trips in NYC in 2015.

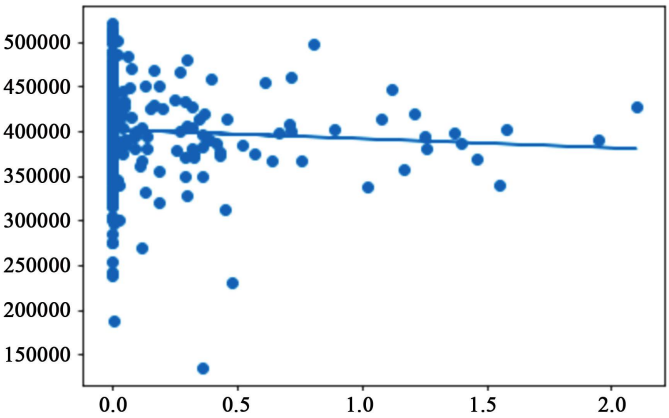

Figure 6. Relationship between precipitation and amount of daily trips.

In addition to the study of daily trips, the trend of monthly average trips in 2015 is examined and graphed. The number of days per month is standardized to 30 days. As shown in **Figure 7**, the first half of the year generally has greater amount of taxi trips than the second-half of the year. This trend is probably caused by the cold weather of the first 6 months of year, when people prefer to take taxi than using other means of transportation.

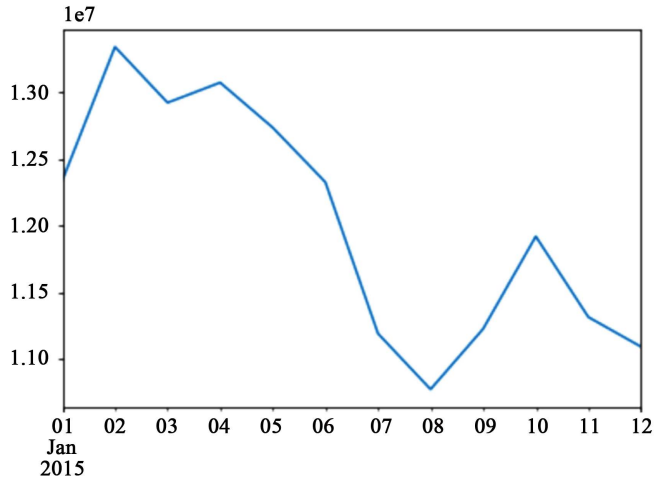

Figure 7. Monthly taxi trips in NYC in 2015.

In Figure 7, x-axis represents 12 months in a year and y-axis represents monthly trip amount. The graph shows the general trend of monthly trip amount during 2015.

In a similar manner, it is found that the trend of average hourly trips throughout the year is almost the same with that in January. Figure 8 depicts the relationship between daily trips and 24 hours in a day.

In Figure 8, y-axis represents amount of daily trips in 2015, while x-axis represents 24 hours per day. Hourly average trips are shown as a result.

Data of payment to taxis in year 2015 is also decomposed to analyze variation of payment in weekday. Figure 9 shows the trend of daily trips on weekday. According to the graph, the amount of trips is at its peak on Friday, while at its lowest level on Sunday, when people enjoy their weekends at home. Graphically, the trend of weekday payment is similar to the trend of trip amount in 2015, however, the highest taxi payment appears on Thursdays, while the largest amount of trips appears on Fridays.

In Figure 9, x-axis represents 7 days respectively during a week, and y-axis represents amount of daily trips.

The location of taxi trips is then carefully studied. A map of NYC is used and read into the Jupyter notebook. It is converted to Geopandas format in order for python to analyze. To find out the trip amount in each NTA area, a function is set up to select which NTA area each trip belongs to. The result is shown in Figure 10, a graph that illustrates the amount of taxi trips in 195 NTA areas. The density of each area is gained by dividing trip amount in an area by the corresponding area. The graph is drawn based on the NYC map, and the differentiation in the darkness of the color distinguishes the density of taxi trips.

Figure 10 shows the trip density and distribution from 8 a.m. to 10 a.m. in January 2015, NYC. According to the graph, it can be also observed that Manhattan and the airport are the busiest areas during rush hours. It implies that the taxi company may distribute more taxis to these areas to meet the need of passengers.

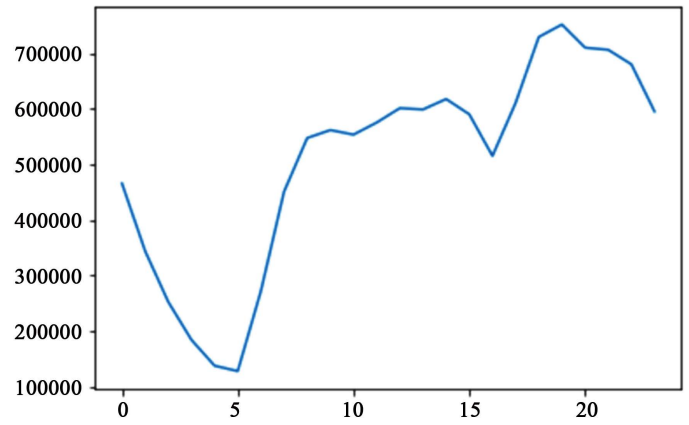

Figure 8. Average amount of trips in 24 hours in NYC in 2015.

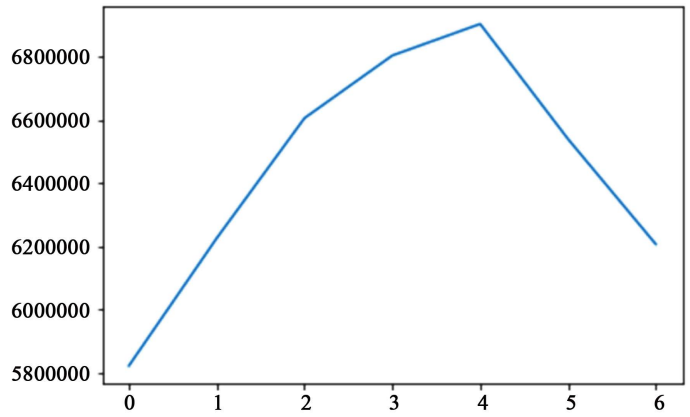

Figure 9. Average amount of daily trips in a week in NYC in 2015.

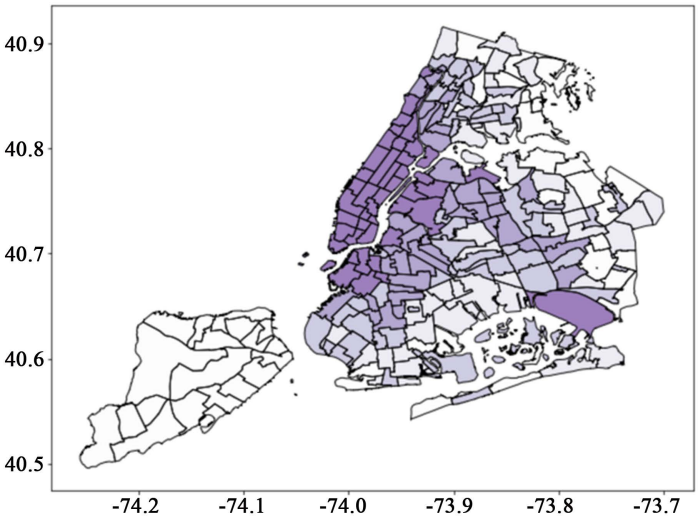

Figure 10. Density of taxi trips in 195 NTA areas in NYC in 2015.

4. Conclusions

This research paper mainly analyzed basic information of taxi trips in 2015 in the city of New York. The trend of the amount of daily trip and average hourly

trip in January is first studied. The weather's impact on the trip amount is then discussed using linear regression to find out the relationship between snow depth, snowfall, precipitation and trip amount. The difference between the impact caused by snowfall and snow depth is later compared. Furthermore, monthly trend, weekday trip amount, hourly average trip, and weekday payment are examined based on the data of 2015. In addition, the distribution of taxi trips in each NTA area is discussed and their density is shown on an NYC map. Finally, by looking at the trips from 8 a.m. to 10 a.m. in **Figure 10**, it's found that a certain area has a greater need for taxis. It is hence meaningful for taxi companies to study the trend and redistribute taxis in the city to satisfy the need of passengers.

The limitation of this research is that it takes a long time to run such a great number of data (2 GB for one month). It takes an afternoon to run only one cell of a year's data, which greatly restricts the amount of data used in the research process. Another limitation is the lack of visualization. The results are mostly shown through graphs, but not by animations which would allow readers to understand more comprehensively and directly.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Whong, C. (2013) NYC Taxis: A Day in the Life.
http://chriswhong.github.io/nyctaxi/#

[2] Schneider, T. (2016) Taxi, Uber, and Lyft Usage in New York City.
http://toddwschneider.com/posts/taxi-uber-lyft-usage-new-york-city/

[3] Schneider, T. (2015) Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance.
http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/

[4] TLC Trip Record Data (2017) NYC Taxi & Limousine Commission.
http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml