Visualization Report

Tip Amount In New York City Taxi Services

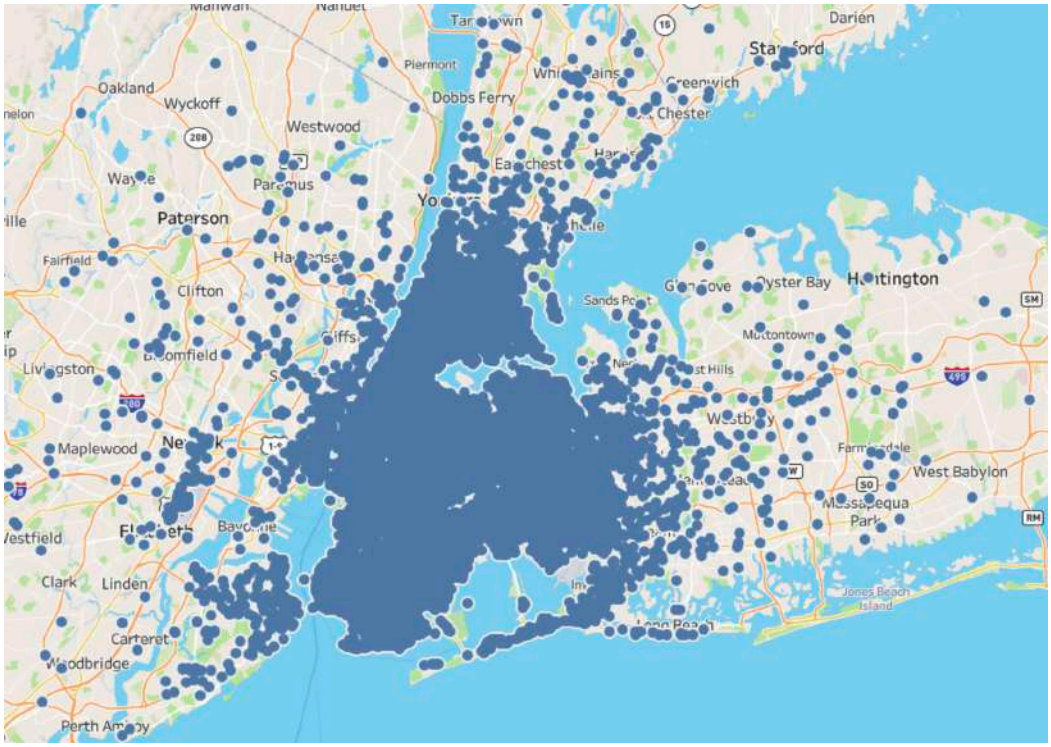

Table of Contents

Introduction

Gratuity culture in the United States has covered a range of areas, even includes taxis. This report aims to discover the factors which are likely to correlate with the tip amount so that we can find out the circumstances which result in a high tip amount.

To achieve this goal, the New York Taxi and Limousine Trip Record Dataset (also known as TLC) are investigated thoroughly. We preprocess the dataset and attempt a set of various geographical visualization methods to demonstrate our findings.

In the end, our investigation suggests several circumstances which are likely to achieve more tips. Without a further understanding of the local county and culture, however, it is hard to make any assertive conclusions regards this issue.

Methodology

Data Period and Attribute Selection

To start with, understandings towards the dataset itself is urgent. TLC is raised by an agency of the New York City government who regulates the vehicle hiring industries.

TLC experiences several modifications in which the recorded data are directly affected. The latest change in 2016 is the location information due to privacy concerns. The dataset is considered to be less geographically precise after 2016. Therefore, the data released before 2016 is considered. Therefore, TLC of 2015 August is determined in this project.

In TLC, there are three main genres of the vehicles. For-hire vehicles such as Uber, green taxi which is only allowed to pick up in boroughs and yellow taxi which has no pick up restrictions. According to Todd's research report (2018), over 80% of the market share belongs to the yellow and green taxi in 2015. Hence, the main analysis object in this investigation are the yellow and green taxis.

At this stage, two datasets are chosen to be the analysis object, which is "Yellow Taxi August 2015" and "Green Taxi August 2015". Throughout the attributes in the dataset, most of the attributes are essential for describing a taxi ride. The attributes that can be used for analysis are kept, whereas attribute "F22", "PickupCell" and "DropoffCell" are dropped.

Data Cleaning and Preprocessing

A reasonable but precise data aggregation is difficult to achieve when given complicate points with only longitude and latitude. The huge amount of data records usually results in a messy plot. To produce insightful geographical visualizations, the New York Taxi Zone Shapefile provided from the New York City Taxi & Limousine Commission is adopted. This investigation aims to aggregate points into the official taxi zones. It is believed that this method would provide clearer demonstration regards the location than square or hex binning.

Various data cleaning steps are taken before joining the shapefile and the dataset. Firstly, the records with a tip amount below 0 are eliminated. Secondly, only the records without a payment type of cash are considered since cash tips are not recorded. Thirdly, rows with a null value in "tip amount" column, " latitude" or " longitude" are dropped. Furthermore, it is found that a certain number of trips are recorded outside of New York City. To save computational power, the records with a pickup or dropoff location outside New York City are excluded.

Nevertheless, the lack of computational power has been a severe issue when visualizing dataset with more than millions of the records. In this case, the green taxi dataset holds more than 1.5 million of the records and the yellow taxi dataset holds more than 12 million of the records. Due to the huge data volume, the visualization process becomes incredibly slow and inefficient.

In response to this difficulty, we have to adopt Python for random sampling. The data size has been reduced to 800,000 records. The data integrity has been damaged for both datasets. For the green taxi dataset, the sample size is 56.7% of the original size. For the yellow taxi, the sample size is less than 10% of the original size in which it may result in a bias visualization. It becomes difficult to assert any findings from the reduced yellow taxi dataset. With this in mind, we visualize the original data at some stages for comparisons before we conclude findings.

The spatial joining for the datasets is performed in Tableau using the built-in function *MAKEPOINT*. We inner join the spatial shapefile to the dataset by detecting intersections. For each dataset, there are two options for joining in which we can choose to use whether the dropoff location or pickup location. Both joining options are performed for different purposes of the analysis.

Green Taxi Analysis

Green taxi is a taxi genre that only allowed to pick up customers in boroughs. The pickup points of the green taxi are evenly distributed over the five boroughs. However, the distribution of the dropoff points is skewed. There are few records recorded in Staten Island. Therefore, the analysis is based on the pickup location.

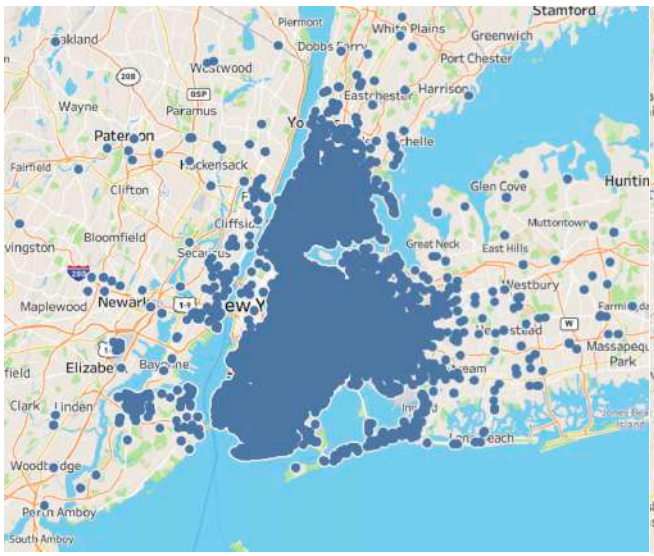

Figure 1 pick up distribution

Figure 2 dropoff distribution

First of all, the data records are aggregated by the boroughs. There are six distinct boroughs in NYC, which are Manhattan, Brooklyn, Queens, Bronx, Staten Island and the EWR airport. Figure 3 consists of two parts, the upper map

indicates the number of taxi records, whereas the lower map demonstrates the average tips per trip. As figure 3 indicates, there are few records recorded in Staten Island and EWR. Nevertheless, EWR and Staten Island show a dramatically high average tip per trips. There are only 24 green taxi records in EWR, however, it has the highest average tip amount of 9.153$. The lack of population size indicates weak statistical power and biased results. Thus, exclude the data from these two boroughs and produce another view of the data.

Figure 3 Trip amount and average tips

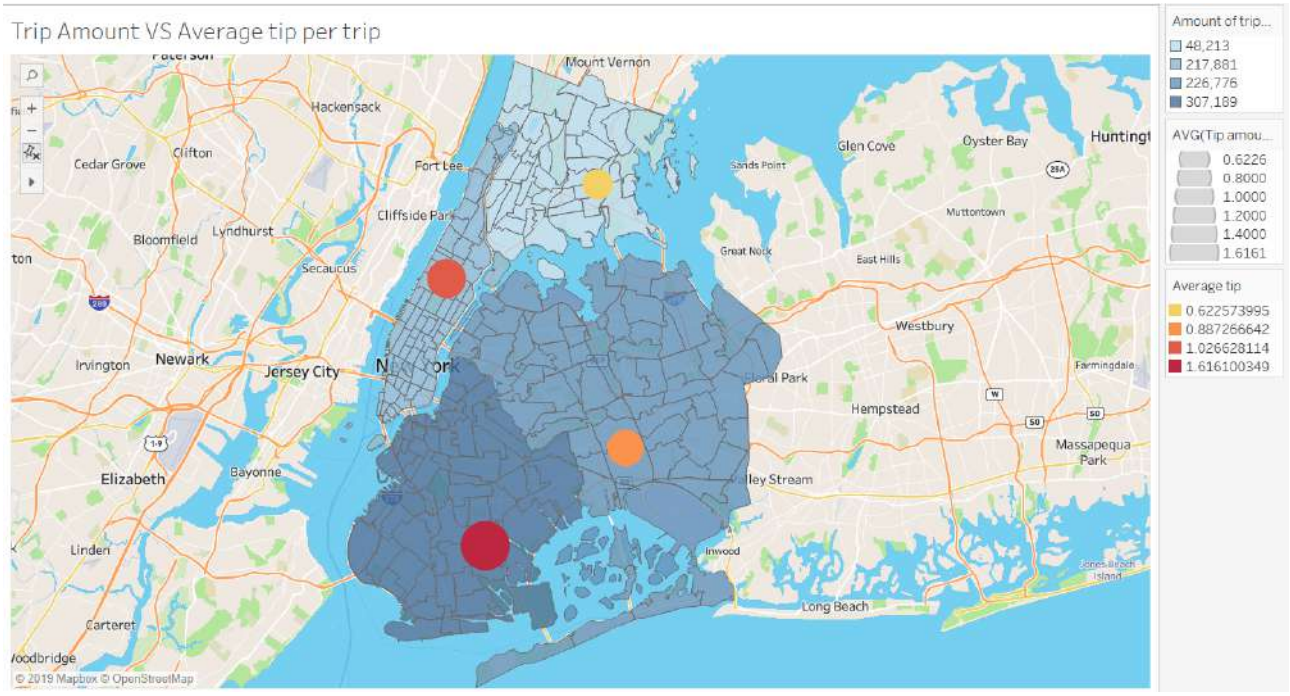

Figure 4 Trip amount and average tip without outlier

As figure 4 indicates, it appears that the average tip amount is positively correlated with the number of trips in

each borough. In other words, large trip amounts usually result in a large average tip. Brooklyn holds the largest amount of green taxi records, it also has the largest average tips of 1.616$. The borough which holds the least number of records also has the lowest average tip.

In figure 5, the plot is remade by adopting taxi zones instead of the borough. Taxi zones divide the NYC into over a hundred blocks. They are developed based on the NYC Department of City Planning' s Neighborhood Tabulation Areas. The purpose of this map is to gain a clear geographical indication of the customer' s start point and destination. Since the area has been divided into small blocks, the difficulty of conducting a bivariate visualization is rapidly increased.

Figure 5 Number of trips and Average tip

In figure 5, the upper-level map is the number of trips in each zone and the lower level on is the average tip amount in each zone. As figure 5 indicates, the average tip amount in each zone is close to each other. Concerning the number of trips, there are several highlighted zones indicating a large number of trips. For example, at the edge between the Bronx and Manhattan and the edge between Queens and Brooklyn. However, the high number of trips in the upper level does not reflect into the average tip amount at the lower level.

To validate whether the number of trips and the average tip amount is positively correlated, the original data is determined. Figure 6 aggregates the raw data records by the taxi zone, and plot the points with the corresponding value. The data does not support a positive correlation at the taxi zone level. If we determine data at boroughs level like figure 4, a not strong positive correlation seems to exist. There are several concerns at the boroughs level such as the unequal sample sizes and culture differences between counties. At this stage, we are not confident enough to conclude a positive correlation

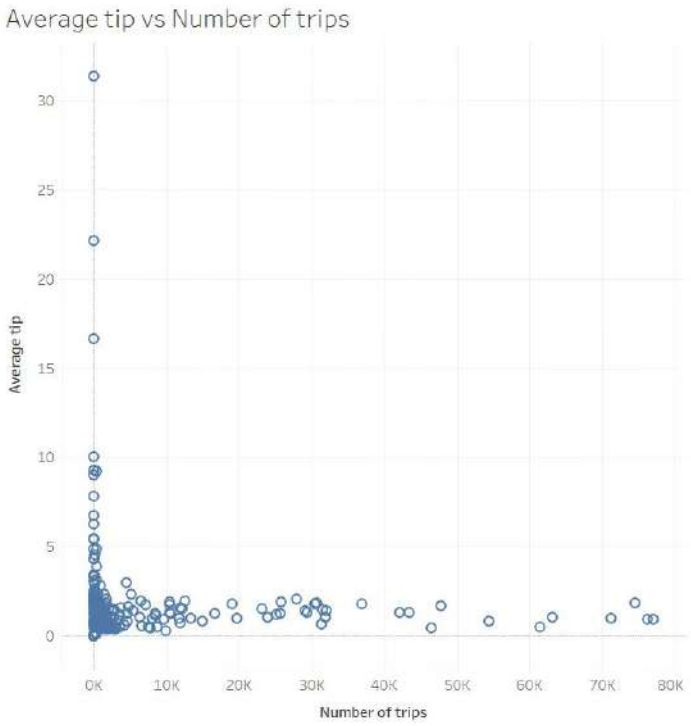

Figure 6 Average tip Vs Number of trips

The next analysis object is the trip distance. Figure 7 is plotted at the taxi zone level. The upper map indicates the average tips paid for each mile. The lower map indicates the average travel distance. Both maps do not demonstrate a dramatic difference in each taxi zone.

Figure 7 Tip per mile and travel distance

As usual, we exclude EWR and Staten Island, then produce other plots to determine whether the tip amount is

correlated with the travel distances. Is there a possibility that people who travelled a long distance will pay more tips? According to figure 8, there is no evidence of extreme contrast in both maps. For example, there is no block which has a deep green colour in the upper map and a light red colour in the lower map. The colour of the underneath map seems lighter than the upper map. It means that high trip distance intends to result in high tip amount.

Figure 8 Tips and Travel distance

Figure 9 Fare without tips and tips amount

Figure 9 is then produced to determine whether people paid a high amount of basic fare is willing to pay fewer tips. The plot shows a surprising result. Similar to figure 8, there is no evidence of any apparent contrast in the map. However, it is found that the overall colour in the upper map is visibly lighter than the lower map. In other words,

people are intended to pay more tips when given a high enough basic fare. To determine whether the finding from figure 9 and figure 10 is reasonable, the original data is accessed.

Both figure 10 and figure 11 are aggregated at the taxi zone level. Figure 10 shows the relationship between trip distance and the tip amount. Figure 11 shows the relationship between the basic fare amount and the tip amount. In common sense, high trip distance means high fares. According to the two plots, there seems to exist in positive relationships. Overall, it seems that a reasonably long trip can sometimes result in a higher trip than the short trip.

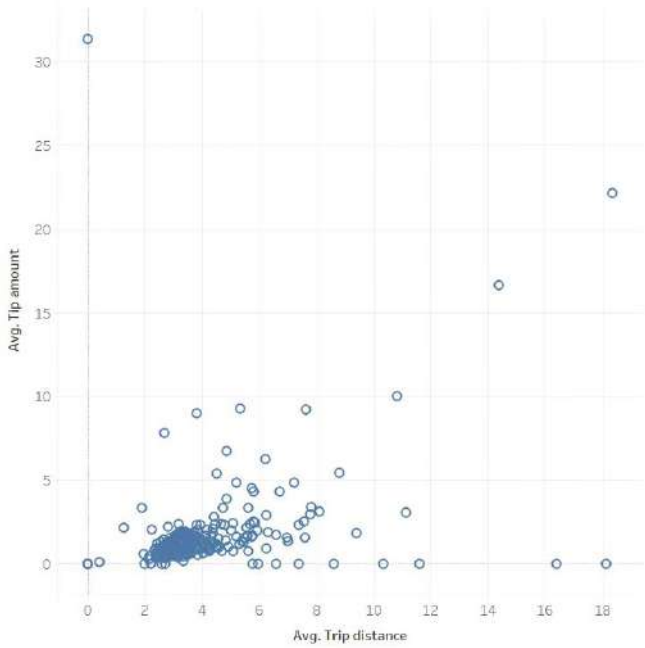

Figure 10 Tip VS Distance

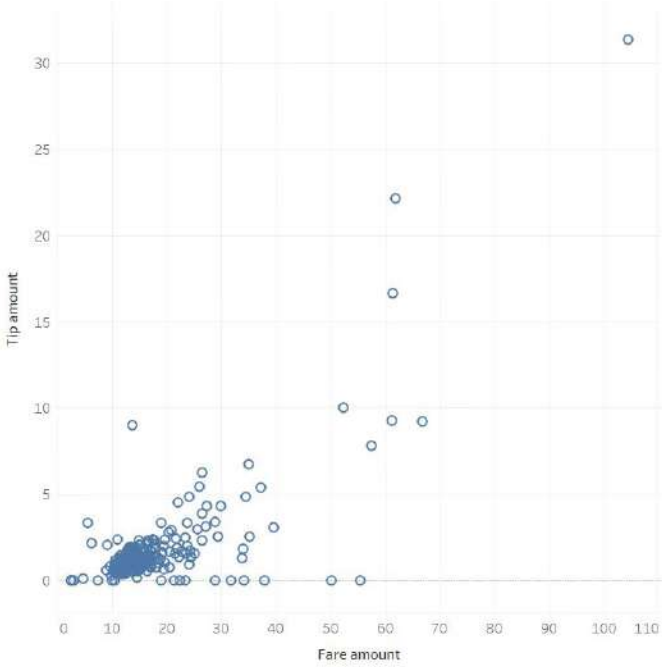

Figure 11 Tip VS Fare Amount

At the current stage, there are several facts can be concluded. In NYC, the green taxis are usually called outside Staten Island and EWR. The trip duration seems to have a positive relationship with the tip amount. Besides, the other attributes such as the trip amount are evenly distributed among the taxi zones. The original size of the green taxi data is about 10% of the yellow taxi data size. Todd (2018) indicates that the market share of the yellow taxi is more than four times the green taxi. In the real-life, there are also some restrictions regarding the green taxi. It is considered that local citizens may use green taxis for particular purposes. In sum, without a deeper understanding of the nature and characteristic of the green taxi, it is hard for us to make further conclusions.

Yellow Taxi Analysis

Before analyzing the actual dataset, two plots are produced to determine the point distribution. As Figure 12 demonstrates, the pickup locations are aggregated at the areas close to Manhattan. As figure 13 shows, the dropoff location is evenly distributed among the boroughs. Therefore, the analysis is operated on the dropoff location.

Figure 14 consists of two components. The upper map indicates the number of trips recorded in each borough. The lower map represents the average amount per trip. The record distribution of the yellow taxi is remarkably different from the green taxi. The number of green taxis is evenly distributed among the four main boroughs. The yellow taxi records in Manhattan, however, occupied around 88.6% of the whole dataset. The huge difference in the data size makes it extremely difficult to interpret the NYC yellow taxi data as a whole.

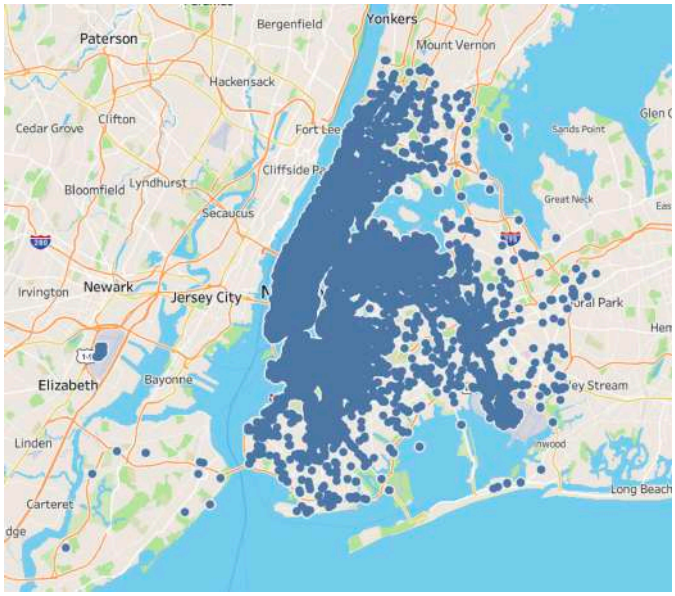

Figure 12 Pick up Locations

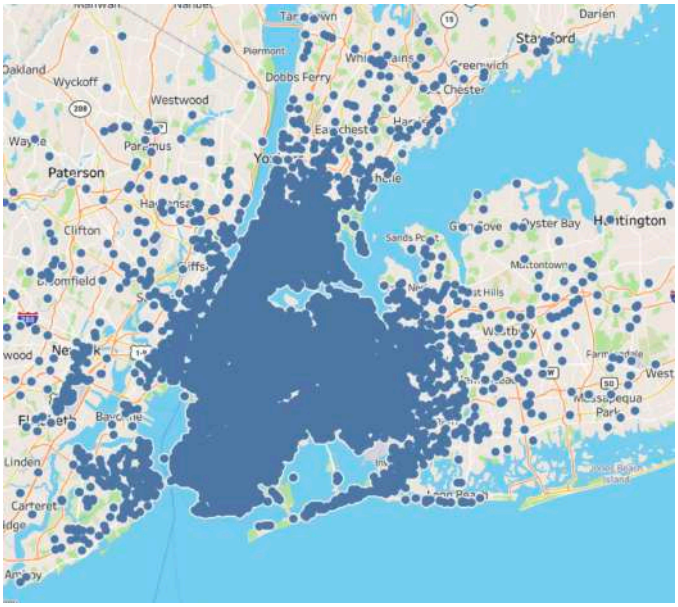

Figure 13 Dropoff Location

Figure 14 Trip Amount and Average Tip per trip

In figure 14, EWR and Staten Island hold the highest average tips per trip. Nevertheless, the small sample has a huge negative impact on its statistical power. The sample size of Brooklyn and Queens are less than 5% of the sample size of Manhattan. This makes the result from Brooklyn and Queens less convincible. Hence, a decision is made to overcome this issue, the analysis regards the yellow taxi focuses on the records in Manhattan. All the taxi zones and boroughs outside Manhattan are filtered to efficiently analyze the dataset.

In figure 15, the map is replotted. The map on the left is the number of trips and the map on the right shows the average tips peri trip. This figure demonstrates a dramatic contrast in colour. In other words, the taxi with a huge amount of records offers a tiny average tip per trip. The next analysis purpose would be the factors that are likely to cause a high tip amount.

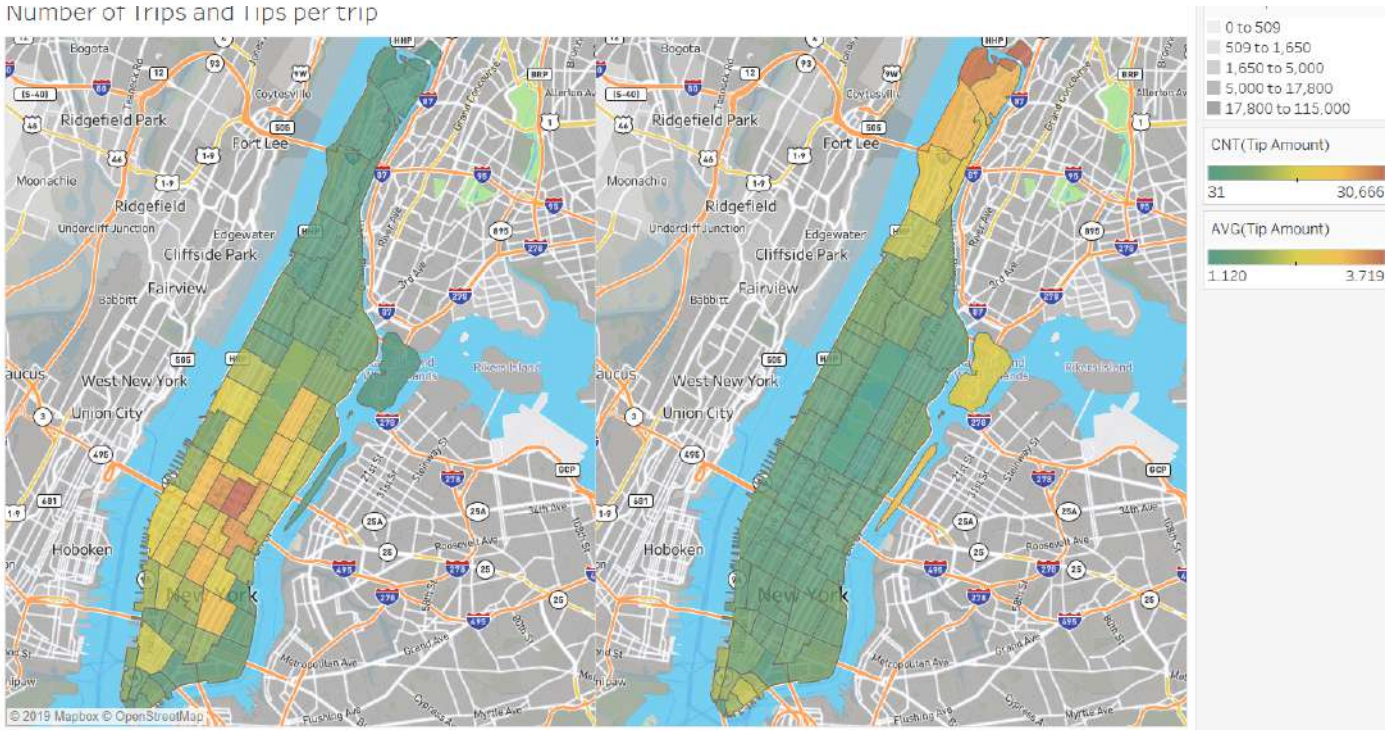

Figure 15 Number of trips and Average tips

In Tableau, the Datetime time difference can be calculated through built-in function *Dateiff*. The time duration for each trip in minute is calculated for analysis. There are two components in figure 16, the map on the left indicates the average tips per trip. The map on the right demonstrates the average tips earned per minute in each area.

From figure 16, we can conclude several interesting facts. The color distribution of the tips per minute seems to contradict to the tips per trip on the left. The taxi zone with high tips per minute does not necessarily result in a high tips rate. There must be hidden factors affecting the tips gain. To understand the logic behind this fact, we still need other complementary attributes to help us comprehend the story.

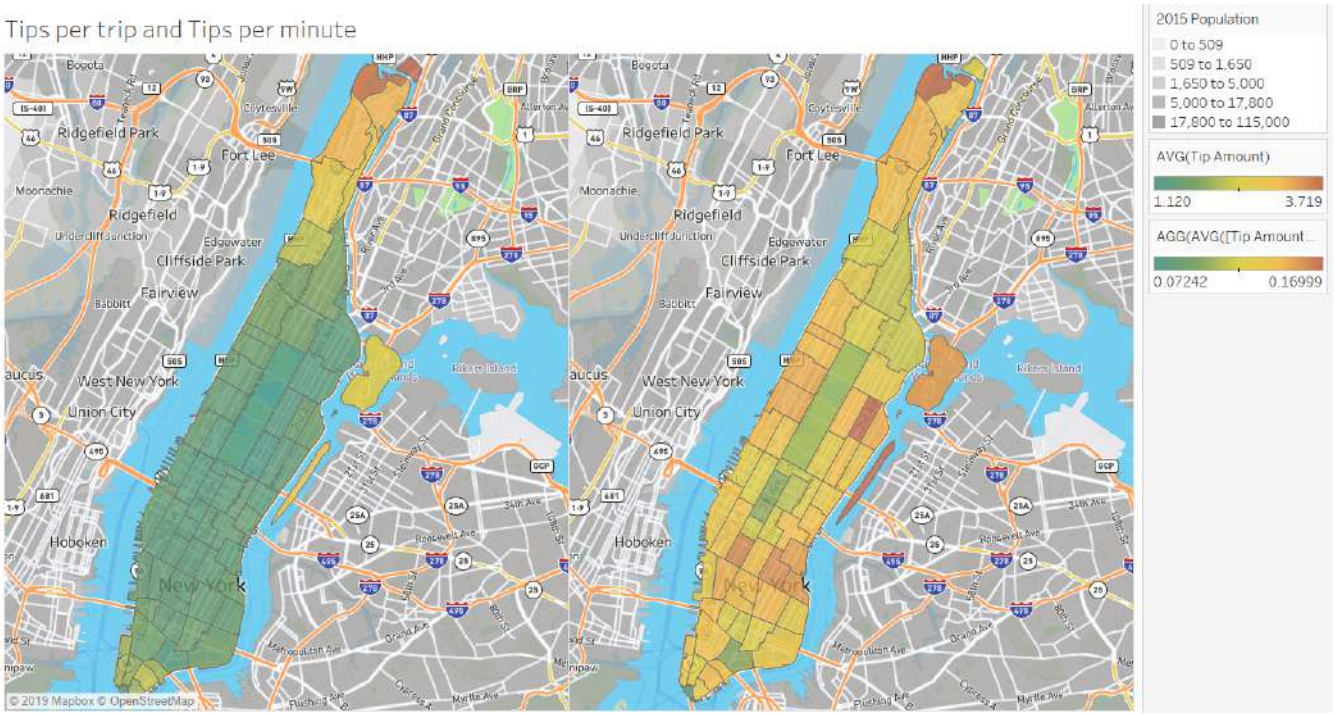

Figure 16 Tips per trip and Tips per Minute

To help understand the map, several taxi zones are excluded to make a clear instruction. In figure 17, there are three main components. From left to right, they are the number of trips, tips per minute and average trip duration. The combined three maps are difficult to interpret because some of the taxi zone have three different colors. Some of the zones has a low tips per min minute, a short trip duration and an extremely high number of trips. It is not easy for the readers to comprehend the logics at once. Based on figure 17, three representative taxi zone are picked up to explain the different scenarios.

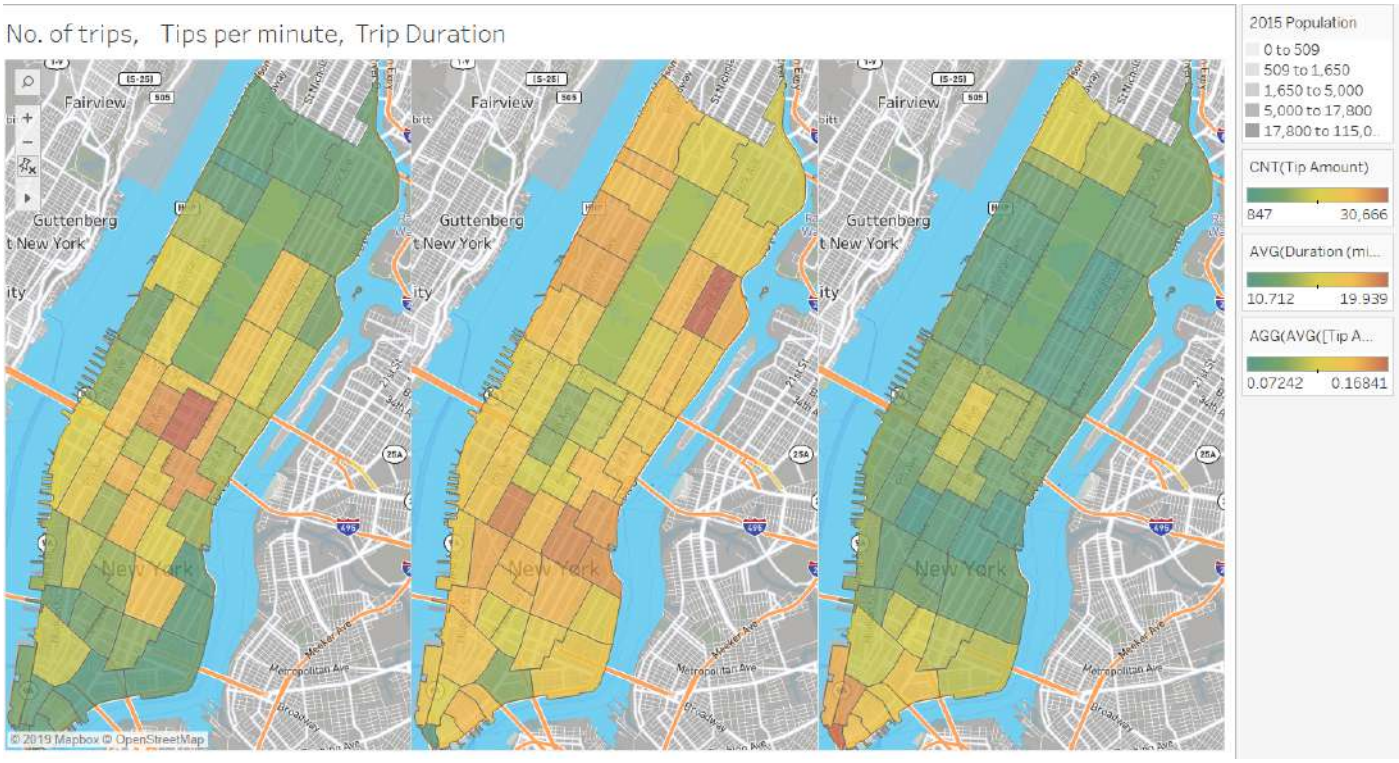

Figure 17 Number of Trips, Tips Per Min, Duration

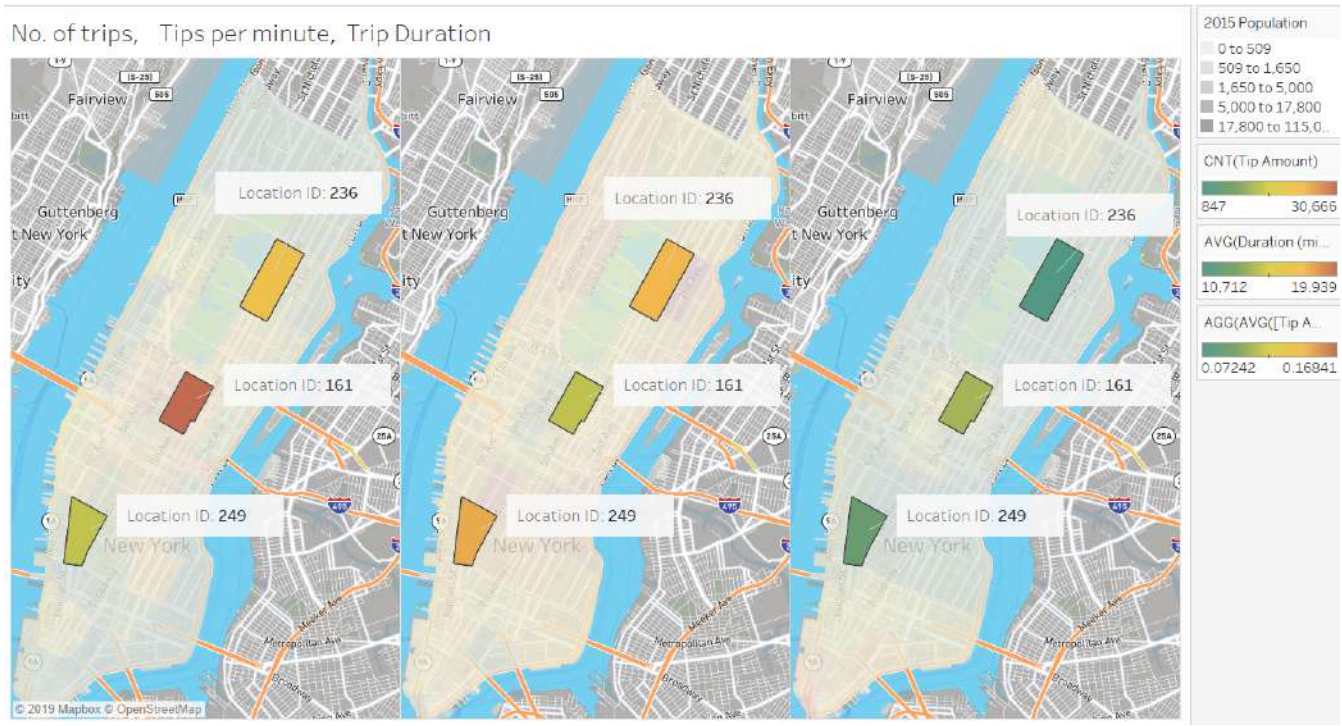

Figure 18 Three representative examples

Start with taxi zone 236 on the top-right corner. 236 has a middle-high level of the number of trips, a middle-high level of the tips per minute and a low level of trip duration. Due to its low average trip duration, its ability to make profits is not remarkable. The high quantity of trips also scales down the average tips it can gain. As a result, location 236 has low tips per trip.

Taxi zone 161 has an extremely high quantity of trips, middle level of the tips per minute and tips. The middle-level duration and tips per minute indicate profitability at the middle level. However, the overwhelming number of trips amount becomes a huge drawback in which the overall tips per trips is reduced.

Taxi Zone 249 theoretically has higher tips per trip than the above two taxi zones. Its middle-high tips per minute and middle-low trip duration indicate middle-level profitability. However, its relatively small number of trips has been a mean advantage for this area.

To conclude, it is impossible for us to assert any directly links between the tips rate and variables listed above. For example, an extremely large number of trips records seems to scale down the overall tips rate for that area. However, there exist certain circumstances that both the tips per unit of time and trip duration are high. In this case, the resulting tip rate for certain area is remarkable. Therefore, asserting a low tip rate base on an observation of a single related attribute is unsuggested.

The most interesting aspect to figure 17 its complexity. Each block in figure 17 has its unique colour set and characteristic to investigate. It is difficult to interpret the overall profitability by looking at one for two distinct variable or attribute. However, it is much efficient to analysis an object in a sequential process. Generally, tips per unit of time is an important index. A taxi zone with high tips per unit of time may eventually not result in high tips rate. However, taxi zones with high tips rate must have high tips per unit of time. For the yellow taxis, most taxis are called in Manhattan. Furthermore, a deep understanding regards the city planning of Manhattan is required for further analysis.

Taxi Tips On The Events Days

Data preprocessing

We are interested in the change in tips rate under event circumstances. Thus, one musical event happened in august 2015 are determined. The event is Billy Joel’ s concert on August 20th, MSG. MSG is in the heart of Manhattan. In the purpose of gaining as much sample size as possible, Billy Joel’ s concert joins the yellow taxi dataset.

To set up for analysis, several data cleaning steps are taken. First of all, the approximate time range is speculated. For the night concert, the event usually starts from 7:30 till mid-night. Assume people start to aggregate at least three hours before the events start. Furthermore, assume most people have been home after the event at three o’ clock in mid-night. In other words, we assume the taxi records drop customers in certain areas in between 4:30 p.m. to 9:00 a.m. are related to the event. Also, records pickup customers in certain areas in between 7:00 p.m. to 3:00 a.m. is related to the event. What’ s more, Spatial join is performed to allocate the related taxi records. It is found that the location of MSG is at the edge between three taxi zones, which are zone 100, zone 186 and zone 68. Within the selected time range, all records with a pickup or dropoff location in these areas remain. Due to the page limit, the following analysis only determines the after-event taxis.

Analysis

Figure 19 demonstrates the records that picks up the customers at certain locations within the given time range. Figure 20 demonstrate the corresponding dropoff locations.

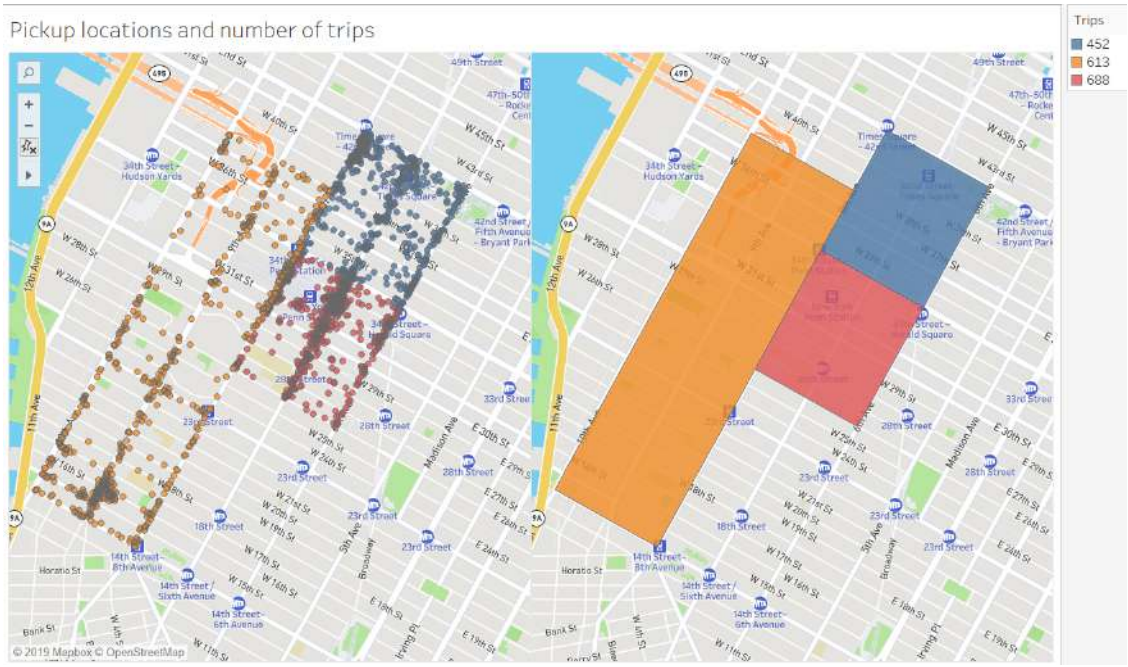

Figure 19 Pickup Locations

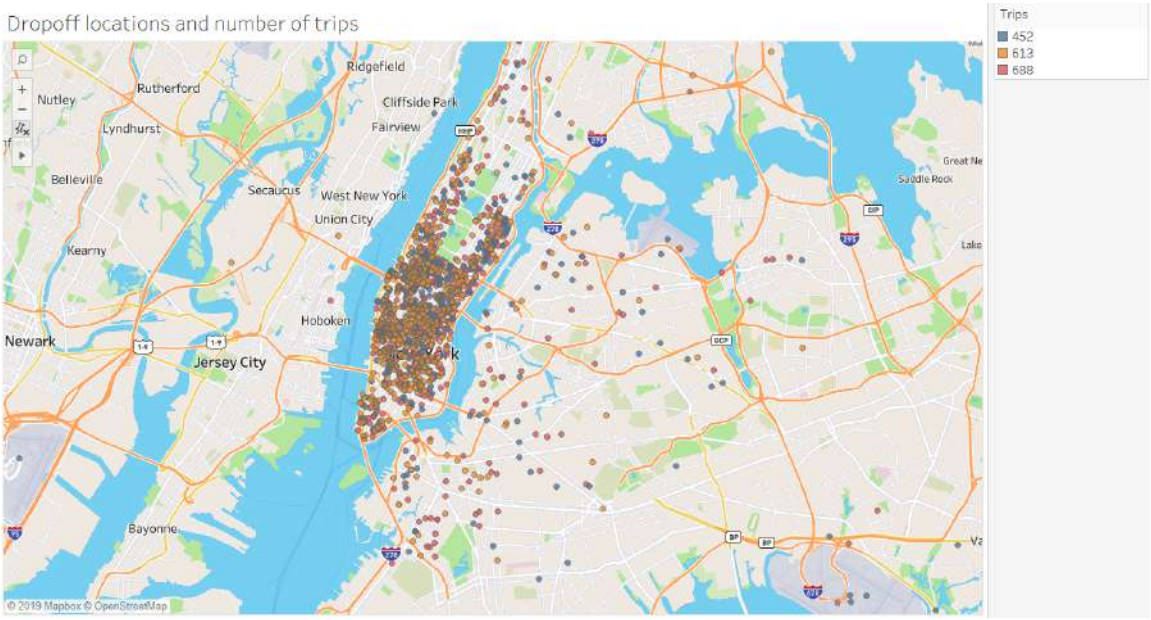

Figure 20 Drop off Locations

According to figure 19, the taxis pickup locations are aggregated into certain streets. MSG is in the centre area around the three taxi zones. The main concern is that we cannot suggest whether the taxi records are related to the event. As figure 19 shows, there are multiple aggregations exist on the map. There exists a possibility that multiple events happening at the same time and similar places. Concerning figure 20, the dropoff locations are evenly distributed among the map. Nevertheless, it is found that a certain amount of taxis drops the customers near the MSG. Since the investigation aims at people who participated in the event, the data purity is doubted.

With this in mind, figure 21 and figure 22 are produced. In figure 21, the tip rate and tips per minute for the customers who paid tips are determined. Figure 22 utilizes the same setup, however, using the original data.

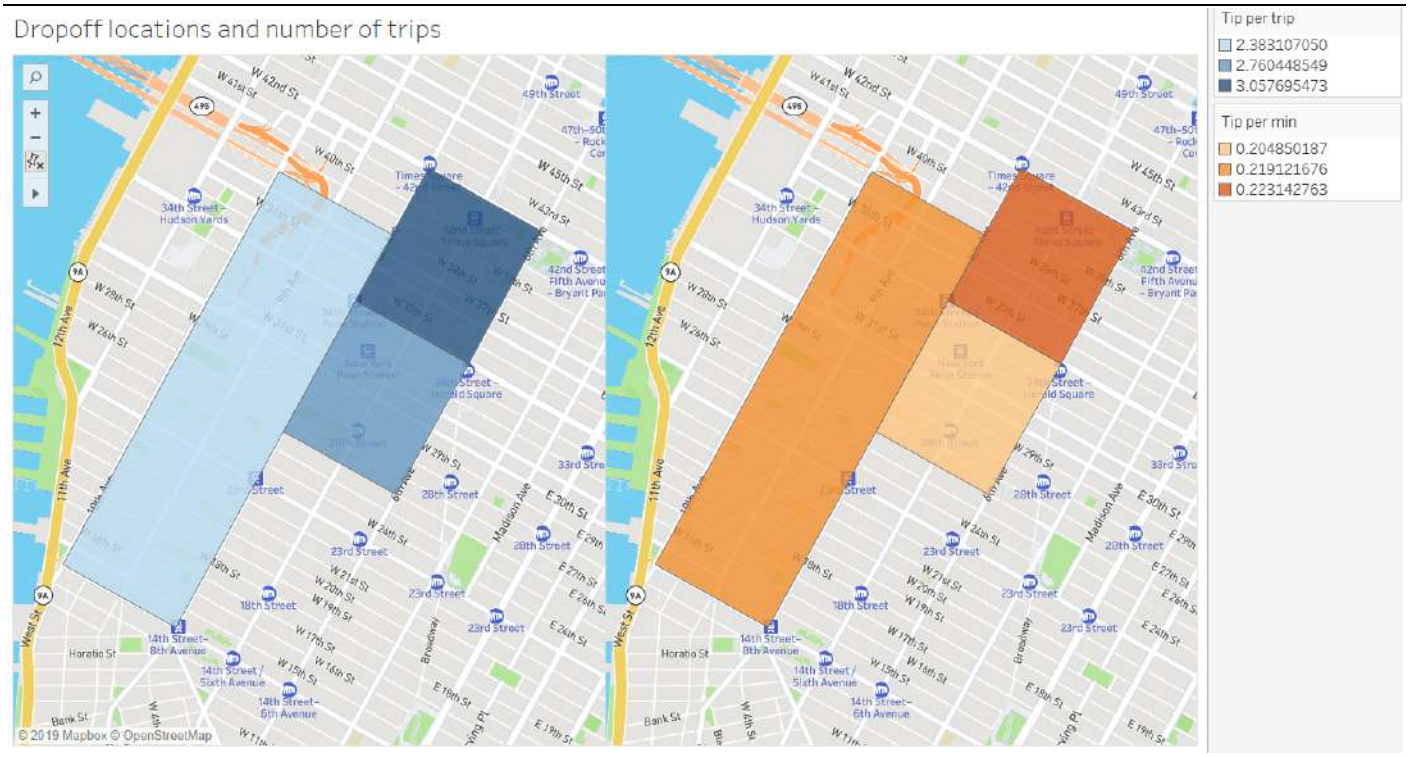

Figure 21 tip rate and tip per min (selected)

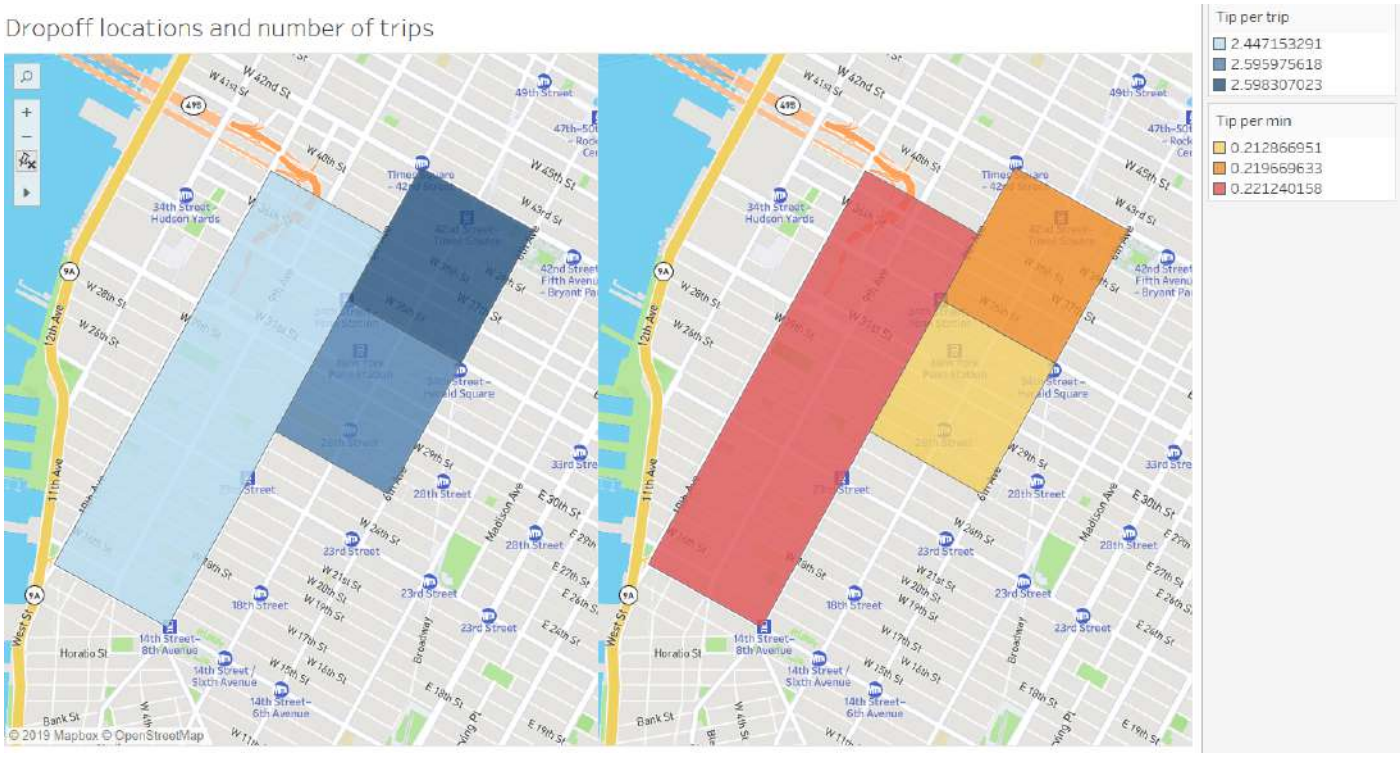

Figure 22 tip rate and tip per min (raw)

According to the above figures, there are no dramatic differences. The tip rate of the selected data is slightly higher than the raw data. However, the tips per minute of the raw data are higher than the selected data. We are confident enough to state that events have a positive correlation with the tip amount due to the low data purity. Neither we can conclude that there are no relationships at all. To improve the analysis regards the tips rate and events. The urgent change to make is an efficient method to filter the event-related records.

Conclusions

In summary, the following findings are analyzed during the investigation.

For the green taxi, the number of taxi records is evenly distributed among The Bronx, Manhattan, Brooklyn and Queens. If NYC is divided by the taxi zones, the tip rate and other indexes are seeming close to each other except for some outliers with small sample sizes. It seems that there exists a positive relationship between the tip rate and the trip distance. Furthermore, if the basic fare is large enough, the customers are likely to pay more tips than the average level.

For the yellow taxi, records aggregated at Manhattan. The biased distribution makes it difficult to interpret data in other boroughs. Once Manhattan is divided into taxi zones, the number of trips differs in each taxi zone. Furthermore, the tips per unit of time and average trip duration various in each taxi areas. The diversities may due to the city planning of Manhattan. For example, the parks, event grounds or housing estates. Further justifications require more background knowledge regards Manhattan City.

Regards the relationship between the events and taxi tips, there is no clear evidence indicating a strong correlation. Due to the difficulties of filtering noises and unrelated data, the statistical power of the results is doubted. More information is required to perform better data selections.

Related Works

Todd W.. 2019. Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. [ONLINE] Available at: https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/. [Accessed 6 August 2019].

nycinsiderguide. 2015. NYC events August 2015. [ONLINE] Available at: https://www.nycinsiderguide.com/new-york-city-events-august-2015. [Accessed 6 August 2019].

Stubbhub. 2015. MSG. [ONLINE] Available at: https://www.stubhub.com/madison-square-garden-tickets/venue/1282/?gcid=chAFF-_-geoUS-_-genAllTix-_-cmp142639-_-partAFFW&awc=7219_1565676139_0b067aeb31ee340cc2a3cafc13c440e3. [Accessed 6 August 2019].

Tableau. 2015. Create Dual-Axis (Layered) Maps in Tableau. [ONLINE] Available at: https://help.tableau.com/current/pro/desktop/en-us/maps_dualaxis.htm. [Accessed 6 August 2019].

NYC open data. 2015. NYC taxi zones. [ONLINE] Available at: https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc. [Accessed 6 August 2019].

NYC open data. 2015. NYC taxi zones. [ONLINE] Available at: https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc. [Accessed 6 August 2019].