MAST30034. Applied Data Science
Y2019S1 Assignment 1

Yin Zhou Zheng (911261)

INTRODUCTION

The aim of this project is to gain an initial insight into the New York City Taxi & Limousine Service Trip Record datasets featured in this subject. The datasets contain trip details from green & yellow taxis and for-hire-vehicles (FHV) operating in the New York City area.

The ultimate purpose of analysing these datasets is to understand "what constitutes a profitable taxi driver in New York". For our initial data inspection, we chose to look at the relationship between sporting events and trip frequencies; as well as the relationship with trip tip rates in USD per mile.

For our sporting event, we focused on the 2015 US Tennis Open. We observed the geographic distribution of trip frequencies and average tip rates over several days. We compared trips on normal weekdays and weekends against the final two days of the Tennis Open featuring the Men's and Women's Singles Finals.

We believed it was reasonable to observe an increase in the number of taxi trips made during the event compared to typical weekdays/weekends; particularly in the number of drop-offs at the event venue – Billie Jean King National Tennis Centre (located in New York City). Additionally, we thought it would also be possible that passengers excited by ongoing major sporting events would be more willing to pay higher tips.

Relating back to the ultimate purpose of this data analysis: if we observe the above hypothesised relationships in our visualisations, then this supports the suggestion that "a profitable taxi driver" will work during major sporting events such as the US Open.

DATA PERIOD SELECTION

As previously mentioned, we chose to focus on the effect of the 2015 US Tennis Open on trip frequencies and tip rates. We chose to look at the 2015 data since it was the last year of the Tennis Open when yellow taxi drop-off longitude and latitude values were still provided. We specifically chose to look at the last two days of the event: the 12th and 13th of September, the days of the Women's and Men's Singles Finals respectively.

Additionally, we further inspected taxi trip activity for two business days (weekdays) and two non-business days (weekends) to compare with the last two days of the event. Due to the arrangement of the datasets, it was more convenient to choose dates from the same dataset, and therefore the same month. Hence, we chose to inspect Monday the 21st, Tuesday the 22nd, Saturday the 26th and Sunday the 27th.

To summarise, we carried out our analysis on six days in September 2015, the 12th, 13th, 21st, 22nd, 26th and 27th.

DATASET & ATTRIBUTE SELECTION

For our analysis, we chose to only inspect the datasets concerning the yellow taxis. Unlike green taxis, yellow taxis were permitted to respond to street hails anywhere within New York City [1]. Without these restrictions, more data could be collected, and hopefully, more 'potential' relationships could be observed. Additionally, the 'yellow taxi' data was preferred over the 'for-hire-vehicle' data due to the former's greater abundance of available data.

Since the datasets were arranged by months, we decided to look at the September 2015 yellow taxi trip data [2] since it contains all our dates of interest. The original dataset itself has a size of 1.8GB, containing more than 11 million trip records.

To test our hypotheses, we only needed a fraction of the available attributes contained in the original dataset. These included:

- 'tpep_dropoff_datetime': The date and time of a trip's drop-off.
  This was needed to determine which dates each trip was made.
- 'passenger_count': The number of passengers within the taxi during the trip.
  Since we only wished to observe the movement of passengers, we ignored any item deliveries ('passenger_count' = 0) [3].
- 'trip_distance': A trip's distance in miles.
  This was required to calculate the tip rate in USD per mile.
- 'dropoff_longitude' & 'dropoff_latitude': The longitude and latitude values of the drop-off location respectively.
  These were required to produce our geographic distribution heatmaps of trip frequencies and average tip rates.

- ‘payment_type’: The method used to pay the taxi fare and tip amount.
  Tips were only recorded for payments made by credit card; otherwise, they would default to zero [3]. This attribute allowed us to determine which fares were paid by credit card.
- ‘tip_amount’: The tip amount in USD.
  In conjunction with ‘trip_distance’, this allowed us to calculate the tip rate in USD per mile.

It should be noted that we chose to focus on the ‘drop-off’ details rather than the ‘pick-up’ details. We believed the former to be more suitable in answering our hypotheses. For instance, if we were to observe a relatively higher frequency and average tip rate for drop-offs at the event venue during the US Open compared to typical weekdays and weekends, then this would support our hypotheses.

DATA PRE-PROCESSING & CLEANSING

Addressing Errors & Filtering Relevant Data

We began our pre-processing by loading all the data into a Python Pandas data-frame. From there, we could easily minimise our dataset by selecting only the relevant attributes – which were listed above. Just this step alone significantly reduced the size of our dataset. Following up with a quick perusal of the dataset allowed us to identify several possible errors; these included:

- ‘dropoff_longitude’ & ‘dropoff_latitude’: We observed longitude & latitude values of zero which would indicate drop-off locations quite distant from New York City. We assumed that these were GPS tracking errors.
- ‘trip_distance’: There were several records with negative or zero trip distances. Negative distances would have been unreasonable, and zero distances would suggest that the taxi never moved. Therefore, we concluded that these must be input errors as well.
- ‘tip_amount’: There were also several records with negative tip amounts which, again, would not make any sense.

We removed the above errors while also removing irrelevant records that were unlikely to contribute to answering our hypotheses. This involved:

- Removing records with *zero* values for drop-off longitude or latitude.
- Removing records with *negative or zero* trip distances.
- Removing records with *negative* tip amounts.
- Removing records with passenger counts of *zero*.
- Removing records that did not involve a *credit card* payment.

After this initial pre-processing and cleansing, we managed to reduce the original dataset of more than 11 million records to seven million. More significantly, this reduced the original file size of 1.8GB to a mere 0.5GB; largely a consequence of the attribute selection, ignoring a large proportion of the original attribute set. This significant reduction in size translated to faster processing speeds.

Tip Rate Attribute

As part of our analysis, we wished to inspect the effect of a major sporting event on the tip rate in USD per mile. The original dataset did not provide such an attribute; however, the attribute was easily derived by combining two existing attributes: 'tip_amount' and 'trip_distance'. By dividing the former by the latter, we could determine each trip's tip rate measured in USD per mile.

Ultimately, we wished to create a heatmap overlay over New York City to determine the distribution of tip rates across different drop-off areas. We employed square binning to help construct our heatmap. Consequently, drop-off coordinates had to be assigned to their respective bins; therefore, each coordinate's associated tip rate had to somehow contribute to the tip rate representing their respective bin.

We chose to use the *mean average* to summarise the coordinates' tip rates in each square bin. However, a major weakness of the 'mean' is its sensitivity to outliers. For instance, a very high tip rate would drastically increase its respective bin's average tip rate. This would be misleading since the resultant bin's average tip rate would be *more* representative of the outlier than the more frequent and 'normal' tip rates. Therefore, we chose to remove such outliers to avoid misleading information in our heatmaps.

We initially looked at a boxplot (see *Figure 1*) representing the distribution of tip rates for trips made in September 2015. It is evident that we have several outliers, with the largest outlier at a tip rate of around 10,000 USD per mile. Despite this record possibly being an error, it is still entirely possible that a generous individual may have paid a 100 USD tip for a 0.01 mile trip. Nevertheless, such occurrences tend to be rare and do not reflect the general trend of tip rates observed in such areas; which we will require in order to fairly compare our event days against non-event days.

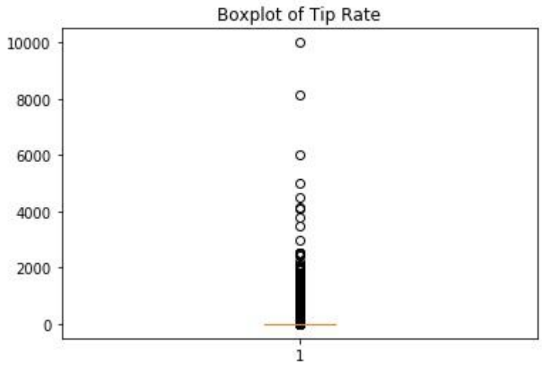


Figure 1: Boxplot of tip rates distribution.

A closer look at F*igure 1* also suggests that most of the outliers occur beyond 2500 USD per mile. Therefore, we produce an additional boxplot that looks at the distribution of tip rates below 2500 USD per mile (in *Figure 2*).

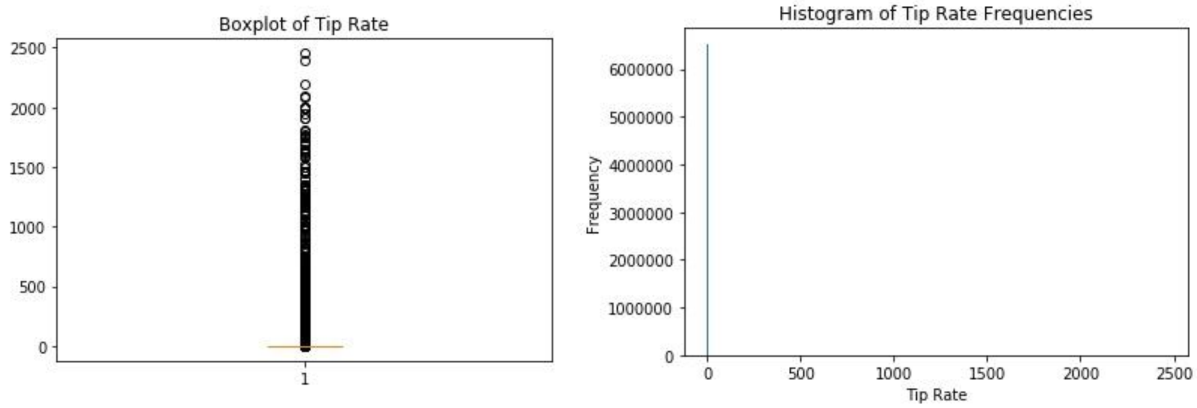


Figure 2: Boxplot & Histogram of Records with Tip Rates < 2500.

At this point, it becomes more apparent that our boxplot lacks the ability to convey the density of observations at given tip rates when there are a lot of records. Therefore, we produce an equivalent histogram which better represents the distribution of these records. It is evident in *Figure 2* that most of the records have tip rates well below 100 USD per mile.

Again, we restrict our dataset to records with tip rates below 100 USD per mile. *Figure 3* below shows us that a tip rate cap of 100 USD per mile is still too high to clearly represent the distribution of our records; hence, we further restrict our dataset to records with tip rates below 10 USD per mile.

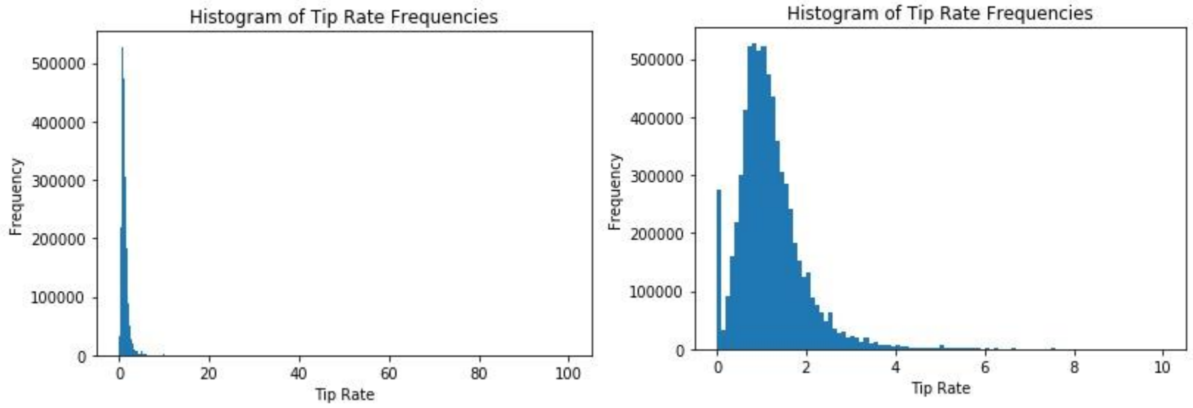


Figure 3: Histograms of Records with Tip Rates < 100 (Left) & Tip Rates < 10 (Right).

However, the right tail is still a bit empty; we can still obtain a better representation of the distribution of typical tip rates for trips made in September 2015. To finalise our histogram, we chose to restrict our dataset to records with a *tip rate* of *at most 7 USD per mile* (see *Figure 4* below).

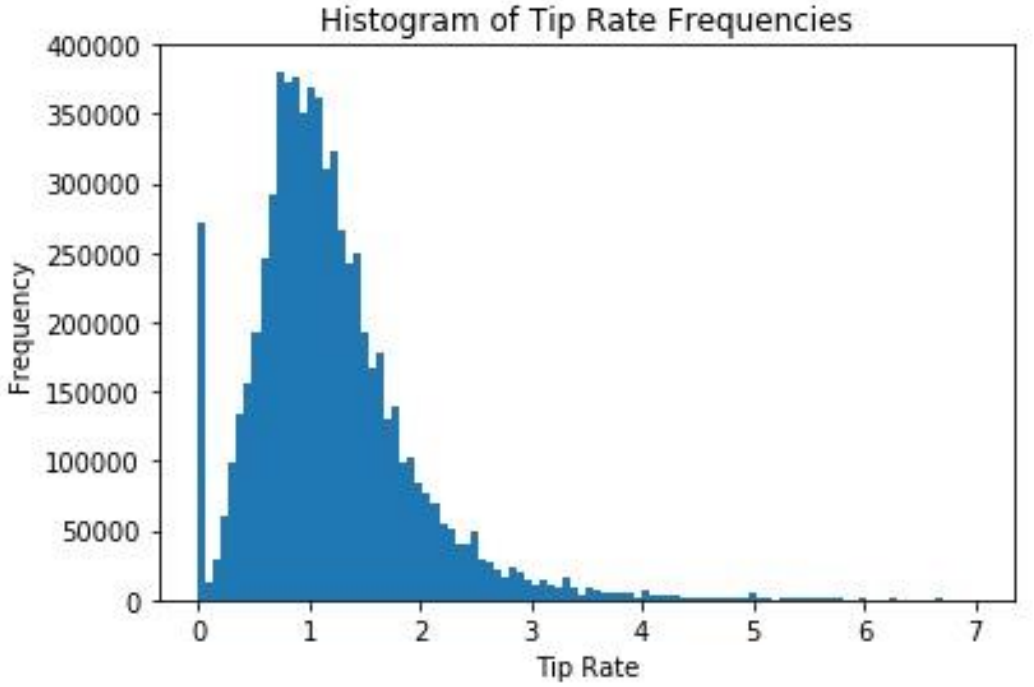


Figure 4: Histograms of Records with Tip Rates < 7.

Date Specific Datasets

Since we wish to compare the trip frequency and tip rates between dates of interest, we chose to assign our records to smaller datasets representing each of our dates of interest. Therefore, we created six separate data-frames containing the records of trips made on the 12th, 13th, 21st, 22nd, 26th and 27th of September 2015.

Once we finished separating our dates of interest, we saved each dataset as a CSV file. From the initial pre-processing and cleaning, we managed to reduce the dataset to roughly seven million records in 0.5GB. With this additional pre-processing, we managed to obtain six datasets consisting of only ~250,000 records each, with file sizes of roughly 20MB; yet another significant reduction in size.

Square Binning

Before proceeding to produce our square bins, we further reduced the size of each date-specific dataset by filtering out unnecessary attributes. Many of the initially selected attributes were used to help clean out errors, filter relevant records, or derive a new attribute. After completing these pre-processing and cleansing tasks, many of these attributes were no longer necessary. We only required the drop-off coordinates – longitude and latitude values – and the derived tip rate; all other attributes were removed.

To produce our square bins, we merely reduced the precision of our longitude and latitude values. We initially rounded each value to three decimal places; however, the resultant heatmap's square bins were far too small when overlayed on a map of New York City. Therefore, we chose to round all our values to *two decimal places*, consequently producing decently sized square bins for our heatmap.

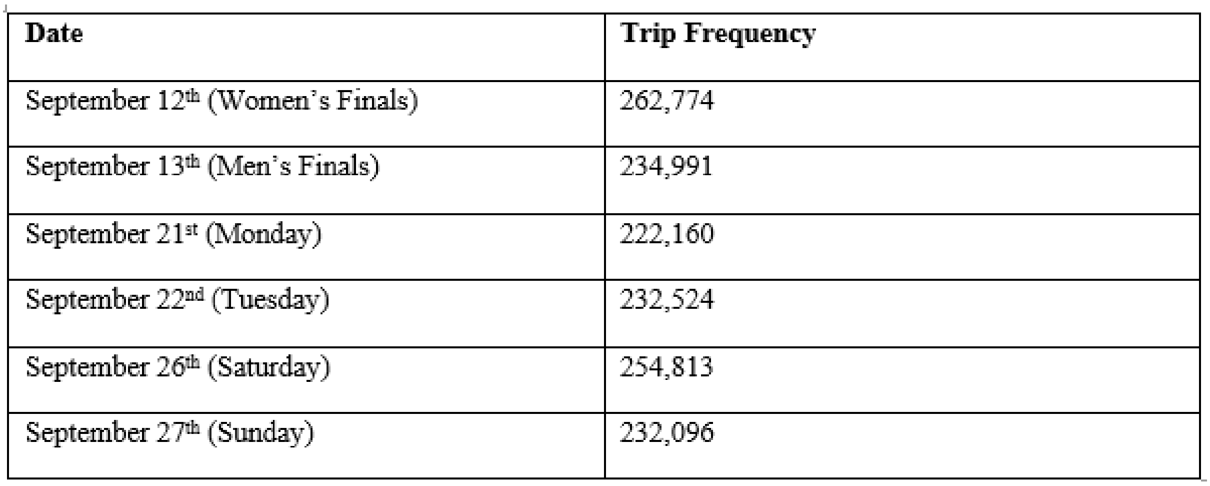
After square binning, we used Python's 'groupby' function to assign each record to their respective square bin. Additionally, we calculated each bin's associated average tip rate and drop-off frequency. This newly assembled data was used to plot our heatmap as an overlay on top of a New York City base-map using Matplotlib's 'Basemap' toolkit [4]; with the heatmap's colour bar indicating the value of a square bin's average tip rate or drop-off frequency.

This finalised our pre-processing stage; allowing us to proceed towards creating visualisations which would help answer our hypotheses.

VISUALISATIONS & ANALYSIS

Date-Specific Trip Frequencies

We initially observed the trip frequencies for each date (see *Table 1* below). We believed that we would observe a higher trip frequency for the dates of the US Open Finals compared to typical weekdays and weekends. While this was evident on the 12th of September for the Women's Finals, the same was not observed for the Men's Finals on the 13th.



Date	Trip Frequency
September 12th (Women's Finals)	262,774
September 13th (Men's Finals)	234,991
September 21st (Monday)	222,160
September 22nd (Tuesday)	232,524
September 26th (Saturday)	254,813
September 27th (Sunday)	232,096

Table 1: Frequency of Trips for Each Date of Interest.

However, it is worth noting that September 12th is also a Saturday; and Saturday the 26th is the date with the second highest trip frequency among our dates of interest. This may suggest that more taxi trips are made on Saturday's compared to other days; however, further analysis should be made before reaching such conclusions.

Heatmaps

We initially inspect *Figure 5*, our heatmap for the drop-off frequencies on the 12th of September – the day of the Women's finals. Contrary to our hypothesis claim, we do not observe a larger drop-off frequency at the event venue. Instead, most of the drop-off frequencies in the Brooklyn and Queens regions are between 2000 to 5000. Majority of the drop-offs occur within Midtown Manhattan, where the centre square bin peaks at over 12,000 drop-offs.

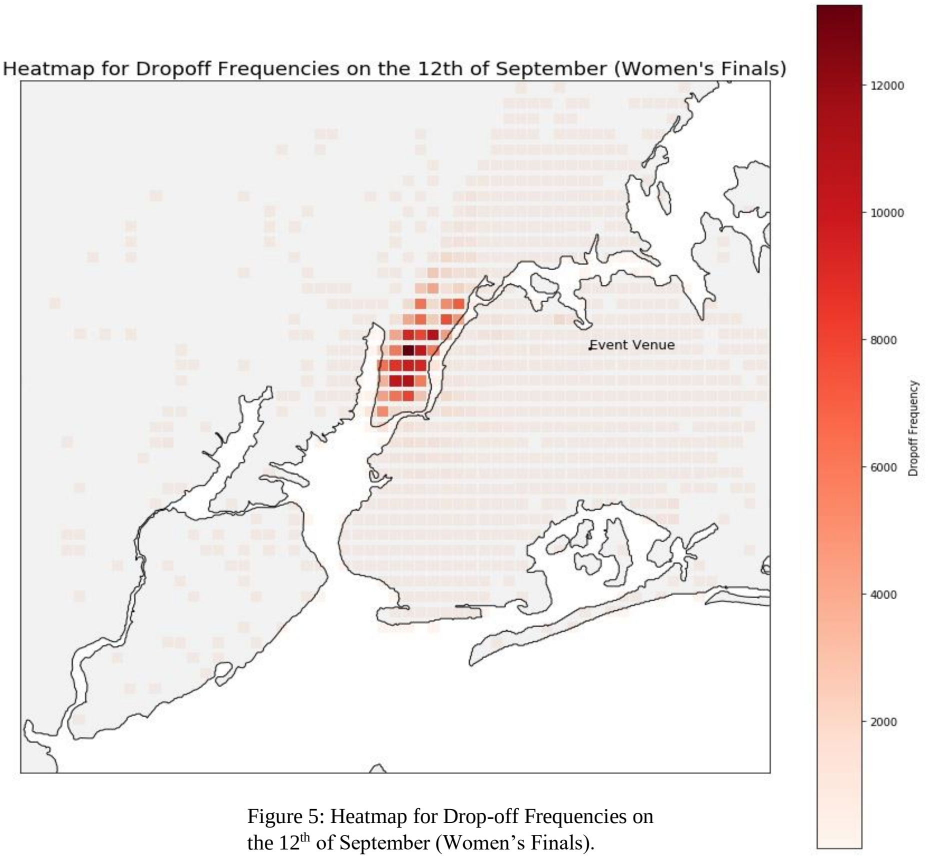

Heatmap for Dropoff Frequencies on the 12th of September (Women's Finals)
Event Venue
Dropoff Frequency
12000
10000
8000
6000
4000
2000

Figure 5: Heatmap for Drop-off Frequencies on the 12th of September (Women's Finals).

Similarly, for *Figure 6* below, we do not observe what we initially hypothesised. The average tip rate in the square bin containing the event venue is barely any different from its surrounding areas; with the majority of Brooklyn and Queens consisting of average tip rates of one to four USD per mile. However, we observe Manhattan with slightly higher average tip rates; suggesting that a taxi driver will be more profitable serving customers in Manhattan.

Additionally, higher tip rates are observed beyond New York City; with the highest observed in the west, near Summit, with an average tip rate of more than 6 USD per mile. This is likely an anomaly, since it does not re-occur in the other dates of interest.

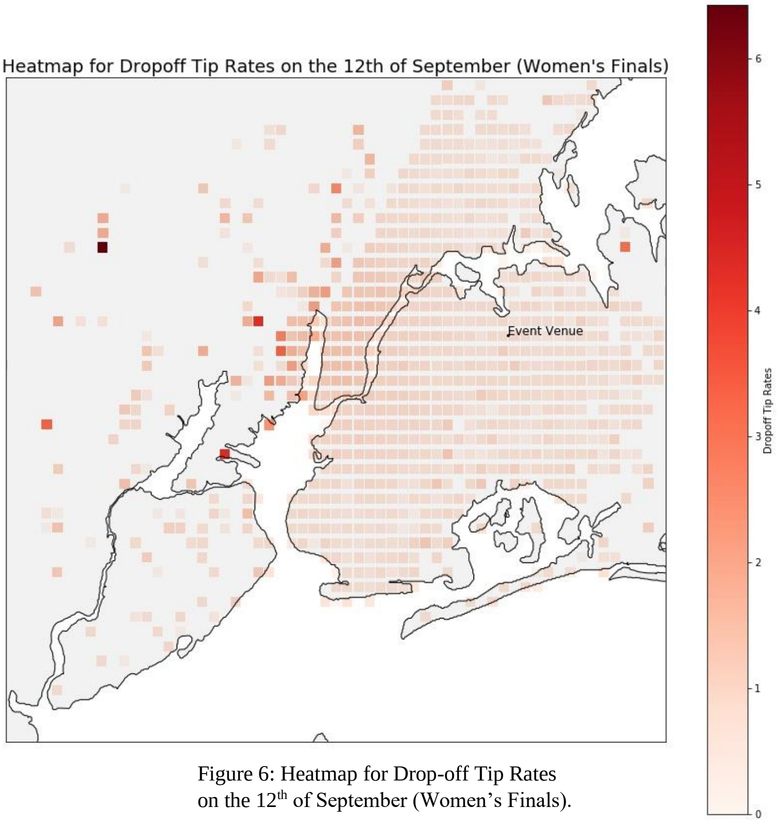

Heatmap for Dropoff Tip Rates on the 12th of September (Women's Finals)
Event Venue
Dropoff Tip Rates
6
5
4
3
2
1
0

Figure 6: Heatmap for Drop-off Tip Rates
on the 12th of September (Women's Finals).

Similar trends are observed in both the heatmaps generated for the Men's Finals on the 13th (see *Figure 7*). Again, most of the drop-offs occur in Midtown Manhattan. And the tip rates are also slightly higher in Manhattan compared to Queens and Brooklyn. As we mentioned before, the highest average tip rate occurs in a different square bin for the 13th; again, this is likely another outlier, perhaps a single point that entirely defines its respective square bin.

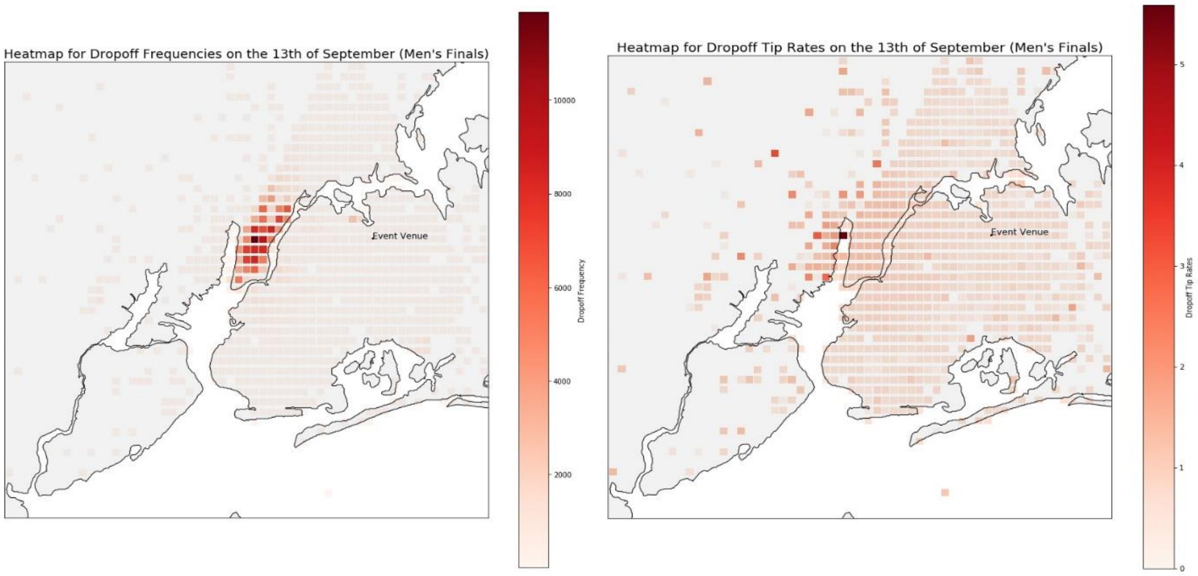

Heatmap for Dropoff Frequencies on the 13th of September (Men's Finals)
Event Venue
Dropoff Frequency
10000
8000
6000
4000
2000
Heatmap for Dropoff Tip Rates on the 13th of September (Men's Finals)
Event Venue
Dropoff Tip Rates
5
4
3
2
1
0

Figure 7: Heatmap for Drop-off Frequency & Average
Tip Rates on the 13th of September (Men's Finals).

We now produce the frequency heatmaps for all our dates of interest in *Figure 8* to allow ease of comparison. Evidently, there is very little difference in the drop-off frequencies between our event days, weekdays and weekends; suggesting that a major sporting event such as the US Open has very little influence on the overall taxi drop-off frequency in New York City.

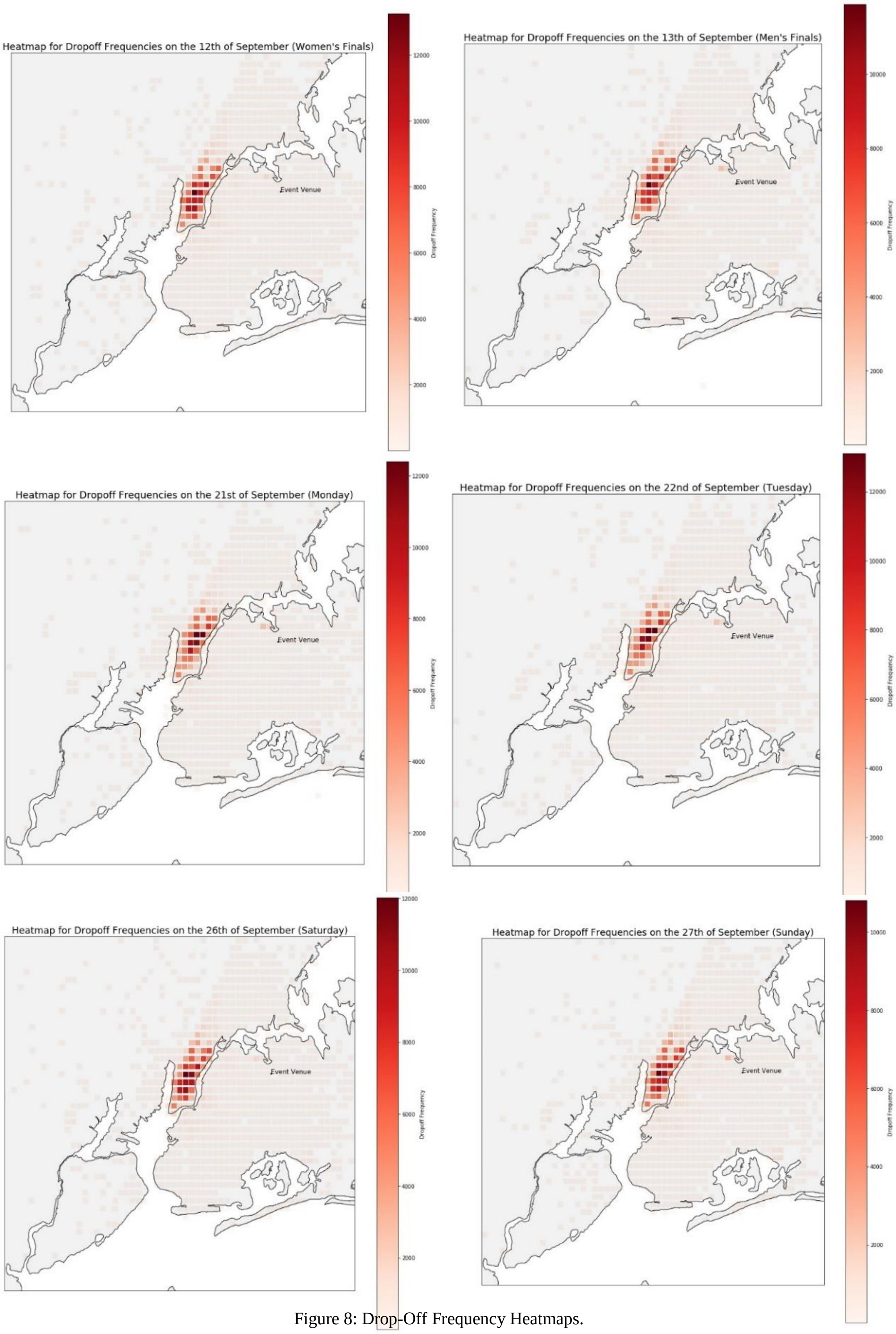

Heatmap for Dropoff Frequencies on the 12th of September (Women's Finals)
Event Venue
Dropoff Frequency
12000
10000
8000
6000
4000
2000
Heatmap for Dropoff Frequencies on the 13th of September (Men's Finals)
Event Venue
Dropoff Frequency
10000
8000
6000
4000
2000
Heatmap for Dropoff Frequencies on the 21st of September (Monday)
Event Venue
Dropoff Frequency
12000
10000
8000
6000
4000
2000
Heatmap for Dropoff Frequencies on the 22nd of September (Tuesday)
Event Venue
Dropoff Frequency
12000
10000
8000
6000
4000
2000
Heatmap for Dropoff Frequencies on the 26th of September (Saturday)
Event Venue
Dropoff Frequency
12000
10000
8000
6000
4000
2000
Heatmap for Dropoff Frequencies on the 27th of September (Sunday)
Event Venue
Dropoff Frequency
10000
8000
6000
4000
2000

Figure 8: Drop-Off Frequency Heatmaps.

Now we do the same for our average tip rates in *Figure 9* below. Again, we observe the general trend with slightly higher tip rates in Manhattan compared to Queens and Brooklyn. However,

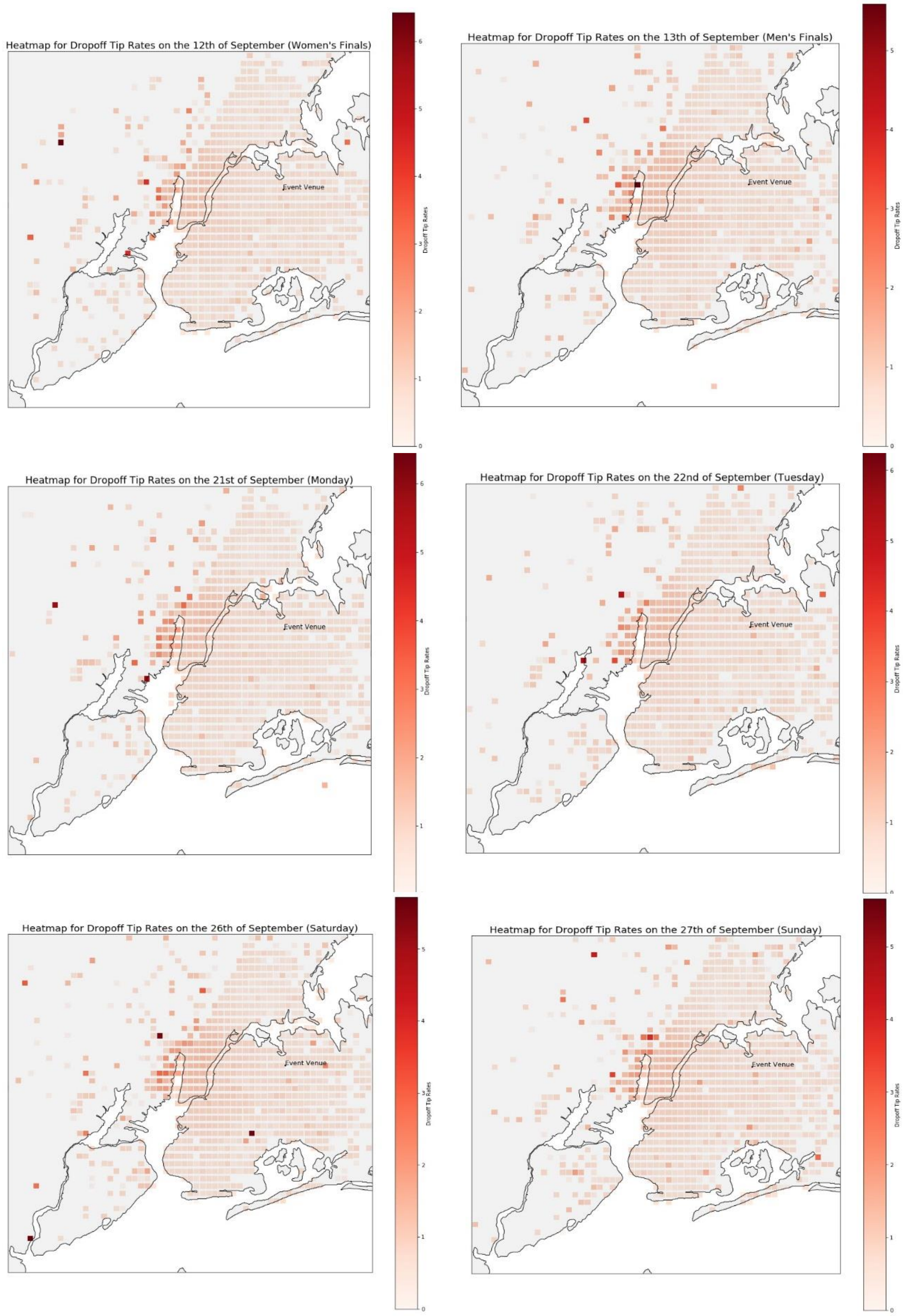

Heatmap for Dropoff Tip Rates on the 12th of September (Women's Finals)
Heatmap for Dropoff Tip Rates on the 13th of September (Men's Finals)
Heatmap for Dropoff Tip Rates on the 21st of September (Monday)
Heatmap for Dropoff Tip Rates on the 22nd of September (Tuesday)
Heatmap for Dropoff Tip Rates on the 26th of September (Saturday)
Heatmap for Dropoff Tip Rates on the 27th of September (Sunday)
Event Venue
Dropoff Tip Rates
0
1
2
3
4
5
6

Figure 8: Average Tip Rate Heatmaps.

we also note that a lot of the higher tip rates are found at drop-offs slightly west of Manhattan in all our dates of interest. This suggests that passengers dropped off in these areas provide higher average tip rates.

CONCLUSION & REFLECTION

In conclusion, we lacked the evidence to support our hypotheses. There was no significant evidence to support our belief that a major sporting event such as the US Tennis Open would increase drop-off frequencies and tip rates at the event venue.

However, this initial exploration of the datasets was useful in gaining some insight into answering the ultimate question of "what constitutes a profitable taxi driver in New York City". Particularly in our tip rate heatmaps, the drop-offs in the region slightly west of Manhattan held the highest average tip rates around the New York City area. This could be worth further exploration.

In retrospect, many things could have been improved upon in this analysis.

The 'Basemap' toolkit was not very suitable in producing city-scaled maps due to its resolution limitations. Importing a map of New York City would have been a better option; providing a more precise map. A New York City map would also feature reference points such as city and borough names which would be helpful in interpreting visualisations and observing trends.

Additionally, the removal of outliers for excessive tip rates which would otherwise heavily influence the mean averages may have inadvertently affected our analysis. The restriction to records with tip rates below seven USD per mile only removed ~16,000 records – a small proportion compared to the existing seven million. However, perhaps many of these records were drop-offs at the event venue. A separate set of heatmaps should have been produced using only the removed records to determine if such a situation existed.

Despite a lot of room for improvement, a lot of insight was gained from this project with respect to the data analysis process, particularly when working with larger datasets.

RESOURCES

[1] New York City Taxi and Limousine Commission. (2019). *TLC Trip Records User Guide*. Retrieved from https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf

[2] New York City Taxi and Limousine Commission. (2019). *2015 September Yellow Taxi Trip Records [CSV]*. Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[3] New York City Taxi and Limousine Commission. (2018). *Data Dictionary – Yellow Taxi Trip Records*. Retrieved from https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

[4] VanderPlas, J. (2016). *Python Data Science Handbook; Essential Tools for Working with Data*. California, United States: O’Reilly Media.