Data Exploration Project
New York City Taxi Rides - 2013

Introduction

NYC is one of the busiest cities in the world and one of the densest in terms of population. Like any metropolitan city, NYC has a lot of taxis and the most common sight on the roads are the Yellow Taxis. A several interesting questions arise about the taxi rides. Let us focus on just two questions.

1. **How are the tips influenced by the time of day? Is there any strong correlation between tips and the number of passengers or the distance travelled etc?**

2. **Plot a map to identify the taxi hotspots around the city. Is it the same throughout the day? How about the weekends? How about holidays like 4th of July?**

Data Wrangling

Taxi Dataset

Figure 1: NYC Taxi & Limousine Commission

The data set is hosted by the **NYC Taxi and Limousine Commission (TLC)**. The original dataset for 2013 had more than **165 million** records which would take a huge amount of time to process. It would also require a lot more processing power than what can be offered by a personal computer.

A user on GitHub trimmed the data set by drawing one record for every 200 rows and uploaded a dataset with **846,000** records which is more manageable.

The data set consists of **2 CSV files**. **Trips and fares**. Together they contain all the taxi ride related details like the **driver's ID & medallion, pick up and drop off coordinates, total fare and tip amount, number of passengers, trip time and distance** etc [1]

Borough Boundaries

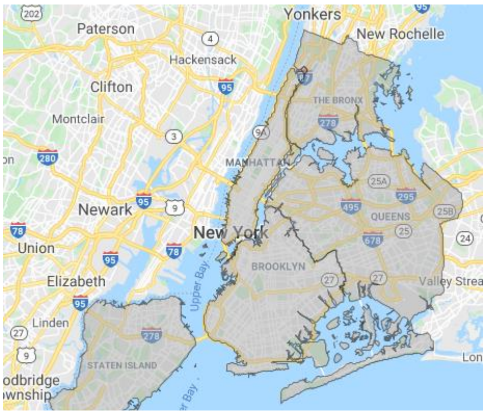
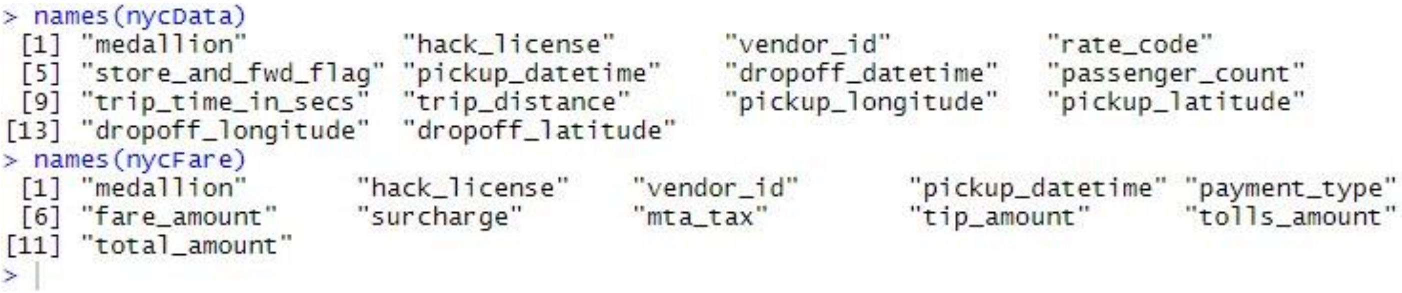


Figure 2: NYC Borough Boundaries plot

We have the **GPS coordinates** of the **pickup and drop off** locations. These coordinates can be classified into different boroughs of NYC if we have the shapefiles of the Borough Boundaries. We can get this data from the NYC Open Data site. The site also shows the boundaries plotted on a map. The plot on the site is shown in figure 2.

Both the data sources have been provided in the References [1& 2]

Transformations and Operations



```
> names(nycData)
 [1] "medallion"          "hack_license"       "vendor_id"          "rate_code"
 [5] "store_and_fwd_flag" "pickup_datetime"    "dropoff_datetime"   "passenger_count"
 [9] "trip_time_in_secs"  "trip_distance"      "pickup_longitude"   "pickup_latitude"
[13] "dropoff_longitude"  "dropoff_latitude"
> names(nycFare)
 [1] "medallion"       "hack_license"    "vendor_id"       "pickup_datetime" "payment_type"
 [6] "fare_amount"     "surcharge"       "mta_tax"         "tip_amount"      "tolls_amount"
[11] "total_amount"
> |
```

Figure 3: Names of all the columns in the two files

1. We start with exploring the datasets. The platform being used is R.

2. Figure 2 shows all the data we have in the file. There are a few columns that we won't be using today, like **vendor ID, rate code, store and forward flag etc**. So, the first step is to **remove them from the data frame.**

3. We then combine the two tables into a single dataframe. This can be done using a simple join. In R the function we use is **merge().** The join operation is performed on the **columns** mentioned below which are common to both the files.

   - Hack License
   - Medallion
   - Pick Up Date/Time

   These 3 columns can **uniquely identify** a single record because a driver (**hack license**) in a car (**medallion**) can only be starting one trip at a time (**pickup date/time**).

4. We then check and remove any **duplicate rows** and check for **NULL values** in any of the columns

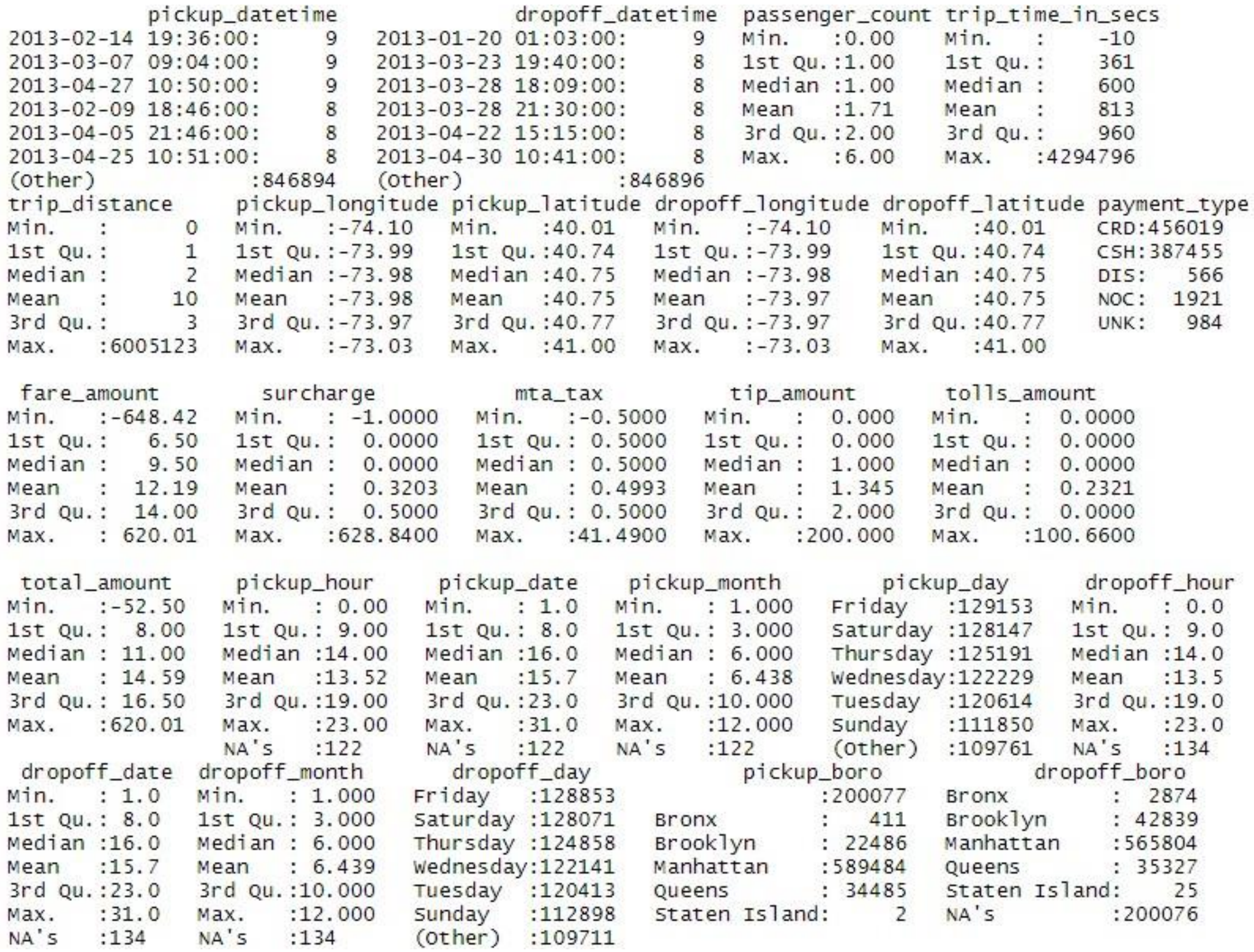
5. The Date/Time columns were read by **R** as **factors**. We need to convert them into valid date/time data type to be able to extract information out of them. This can be done using the **R function as.POSIXlt().** The output format needs to be specified like this **"%Y-%m-%d %H:%M:%S"**

   The columns were then used to extract details like the **day of week, month, hour etc.**

6. We even have the **latitude and longitudes** for each taxi ride. These co-ordinates can be converted into addresses or suburbs using the given shapefile data of NYC. RStudio couldn't handle the more than 850,000 records of data that needed to be converted. So, I used Python for this.

   New York City is divided into 5 boroughs. They are **Brooklyn, Bronx, Manhattan, Staten Island and Queens.** The NYC Shapefiles contain the borough boundaries in the form of polygons. The coordinates we have can be grouped into their boroughs using the **Shapely Package in Python**. (This took 5 hours!!)

Data Checking



```
            pickup_datetime               dropoff_datetime   passenger_count trip_time_in_secs
 2013-02-14 19:36:00:      9   2013-01-20 01:03:00:      9   Min.   :0.00    Min.   :     -10
 2013-03-07 09:04:00:      9   2013-03-23 19:40:00:      8   1st Qu.:1.00    1st Qu.:     361
 2013-04-27 10:50:00:      9   2013-03-28 18:09:00:      8   Median :1.00    Median :     600
 2013-02-09 18:46:00:      8   2013-03-28 21:30:00:      8   Mean   :1.71    Mean   :     813
 2013-04-05 21:46:00:      8   2013-04-22 15:15:00:      8   3rd Qu.:2.00    3rd Qu.:     960
 2013-04-25 10:51:00:      8   2013-04-30 10:41:00:      8   Max.   :6.00    Max.   :4294796
 (Other)             :846894   (Other)             :846896
 trip_distance      pickup_longitude pickup_latitude dropoff_longitude dropoff_latitude payment_type
 Min.   :      0   Min.   :-74.10   Min.   :40.01   Min.   :-74.10    Min.   :40.01    CRD:456019
 1st Qu.:      1   1st Qu.:-73.99   1st Qu.:40.74   1st Qu.:-73.99    1st Qu.:40.74    CSH:387455
 Median :      2   Median :-73.98   Median :40.75   Median :-73.98    Median :40.75    DIS:   566
 Mean   :     10   Mean   :-73.98   Mean   :40.75   Mean   :-73.97    Mean   :40.75    NOC:  1921
 3rd Qu.:      3   3rd Qu.:-73.97   3rd Qu.:40.77   3rd Qu.:-73.97    3rd Qu.:40.77    UNK:   984
 Max.   :6005123   Max.   :-73.03   Max.   :41.00   Max.   :-73.03    Max.   :41.00

  fare_amount        surcharge           mta_tax           tip_amount        tolls_amount
 Min.   :-648.42   Min.   : -1.0000   Min.   :-0.5000   Min.   :  0.000   Min.   :  0.0000
 1st Qu.:   6.50   1st Qu.:  0.0000   1st Qu.: 0.5000   1st Qu.:  0.000   1st Qu.:  0.0000
 Median :   9.50   Median :  0.0000   Median : 0.5000   Median :  1.000   Median :  0.0000
 Mean   :  12.19   Mean   :  0.3203   Mean   : 0.4993   Mean   :  1.345   Mean   :  0.2321
 3rd Qu.:  14.00   3rd Qu.:  0.5000   3rd Qu.: 0.5000   3rd Qu.:  2.000   3rd Qu.:  0.0000
 Max.   : 620.01   Max.   :628.8400   Max.   :41.4900   Max.   :200.000   Max.   :100.6600

  total_amount      pickup_hour      pickup_date    pickup_month          pickup_day     dropoff_hour
 Min.   :-52.50   Min.   : 0.00   Min.   : 1.0   Min.   : 1.000   Friday   :129153   Min.   : 0.0
 1st Qu.:  8.00   1st Qu.: 9.00   1st Qu.: 8.0   1st Qu.: 3.000   Saturday :128147   1st Qu.: 9.0
 Median : 11.00   Median :14.00   Median :16.0   Median : 6.000   Thursday :125191   Median :14.0
 Mean   : 14.59   Mean   :13.52   Mean   :15.7   Mean   : 6.438   Wednesday:122229   Mean   :13.5
 3rd Qu.: 16.50   3rd Qu.:19.00   3rd Qu.:23.0   3rd Qu.:10.000   Tuesday  :120614   3rd Qu.:19.0
 Max.   :620.01   Max.   :23.00   Max.   :31.0   Max.   :12.000   Sunday   :111850   Max.   :23.0
                  NA's   :122     NA's   :122    NA's   :122      (Other)  :109761   NA's   :134
  dropoff_date   dropoff_month        dropoff_day              pickup_boro           dropoff_boro
 Min.   : 1.0   Min.   : 1.000   Friday   :128853                   :200077   Bronx        :  2874
 1st Qu.: 8.0   1st Qu.: 3.000   Saturday :128071   Bronx          :   411   Brooklyn     : 42839
 Median :16.0   Median : 6.000   Thursday :124858   Brooklyn       : 22486   Manhattan    :565804
 Mean   :15.7   Mean   : 6.439   Wednesday:122141   Manhattan      :589484   Queens       : 35327
 3rd Qu.:23.0   3rd Qu.:10.000   Tuesday  :120413   Queens         : 34485   Staten Island:     25
 Max.   :31.0   Max.   :12.000   Sunday   :112898   Staten Island:      2   NA's         :200076
 NA's   :134    NA's   :134      (Other)  :109711
```

*Figure 4: Summary statistics of the dataset generated by the **summary()** function in **R***

We now have a dataframe of data that can be used to generate any kind of visualizations. But the data is not perfect yet. The figure 4 shows the summary statistics of the data. This can be generated using the R function **summary().**

We can notice a few abnormalities in the data. Some are detailed below –

1. The **maximum trip distance** is **6005123 miles**. That is 15.6 times the distance from the Earth to the Moon. That cant be right.

2. The **fare amount** and the **total amount** columns both have their **minimum values in negative.**

3. Some of the **Date/Time** entries were **erroneous** and that caused a few **NULLS** in the extracted fields like the **month, day, hour etc**

4. **Box plots** were generated for the **tip, trip distance and trip time** columns to check for and **remove outliers**. The upper hinge in a box plot is the value outside which the data points are considered outliers. The upper hinges for **tips, trip distances and trip times** are **$5, 5.75 miles and 1560 seconds respectively.**

NYC is a huge city covering a wide area. But the 5 boroughs are the most densely populated. Therefore, most of the cab rides originate and end here. But we still have a lot of data that is spread over larger and isolated areas. So, in order to gain meaningful visualizations from the graphs we filter out any points outside the 5 boroughs. This leaves us with a little more than **450,000** records. This data is saved to a file which will be our source.

Data Exploration

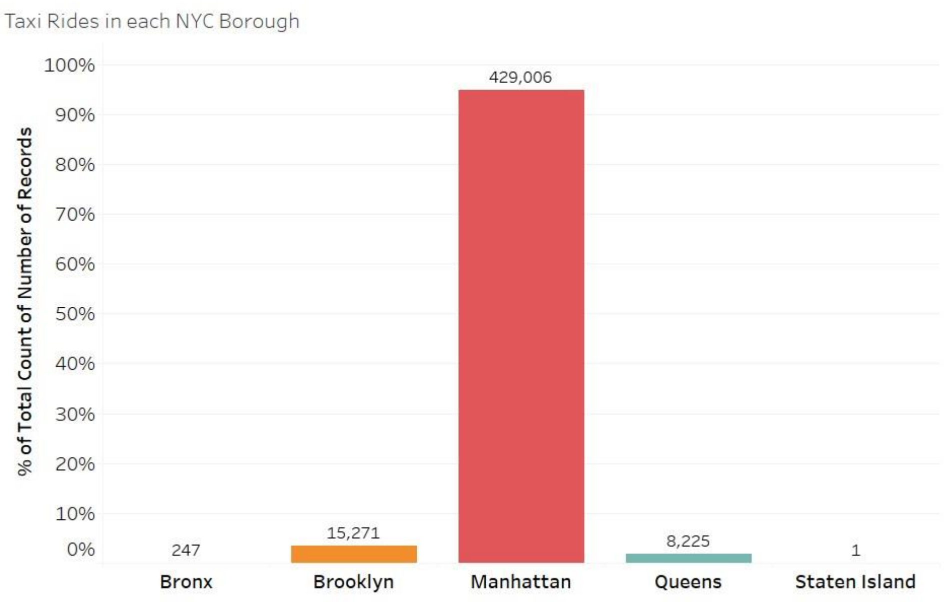


Figure 5: Simple Bar Plot showing the number of taxi rides in each borough

The graph in figure is a **Simple Bar Plot** generated using **Tableau**. It shows the number of taxi rides in each NYC borough. **Manhattan** has the highest number of taxi rides of the 5 boroughs. Manhattan is the most **densely populated** borough of the 5 at over **70,000 people per square mile [3]**. Brooklyn takes the 2nd spot at a far lower 37,000 people per sq. mile [4]. That could explain the far higher frequency of taxi rides.

We should note that the original dataset has 165 million records. This is only a sample of 450,000 records. But we take this as a representation of the city with reasonable confidence because the records were taken as a random sample.

The table also contains the **tip amounts** of all the cab rides. We can explore any possible correlations between the tips and other attributes. Using **Tableau,** we can generate a few basic statistics to see what the tip amounts are like. **'No tip'** is the most common tip, but at 220,000 it's only half of the overall rides.

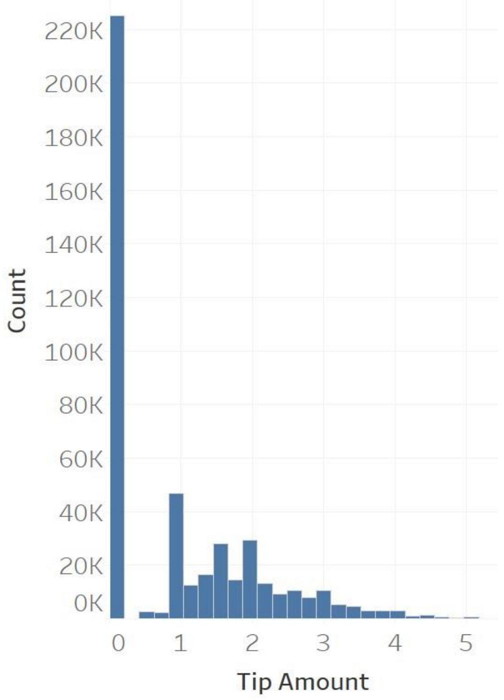


Figure 6: Histogram showing the number of rides for various amounts of tip given

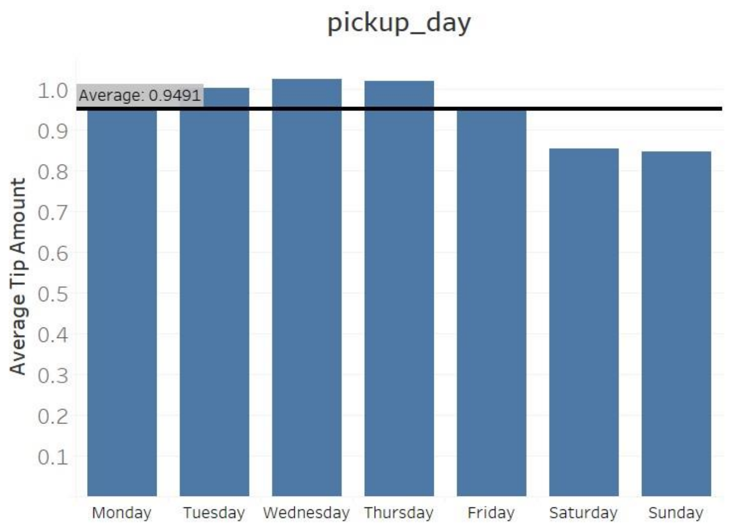


Figure 7: Simple Bar Plot showing the distribution of taxi rides through the week

It is nice to see that about half the population offers a tip. But they aren't as generous during the **weekends** as seen in figure 7.

Tips vs Time of day

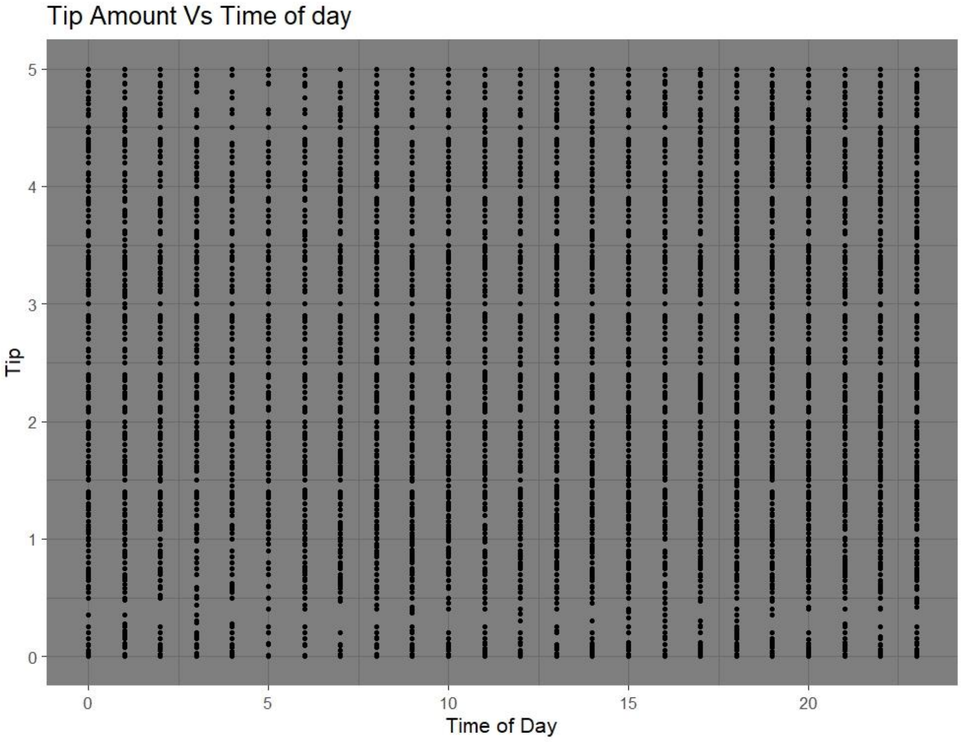


Figure 8: Correlation plot between Tips and Time of day

This is a **scatter plot** or **correlation plot** between two variables generated using **ggplot2**, **R**'s graphics package.

It shows all the recorded **tip amounts** for different **times of day by the hour (0 to 23)**. Its clear that there is **absolutely no relation** between the tips and the time of day.

Although there is a fascinating pattern in the above data. The lines are all broken near the **integers and the ½ dollar marks.** That is, nobody offers tips like 3.1 dollars or 2.9 or 1.4 dollars. But all other decimals are common.

Tips vs Passengers

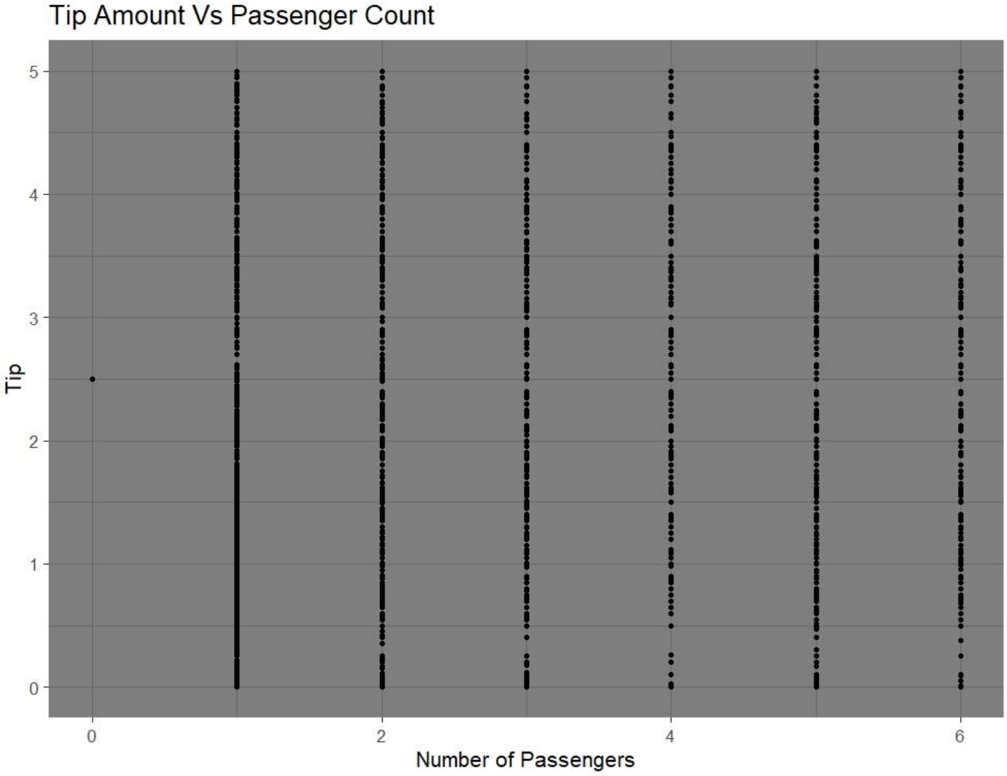


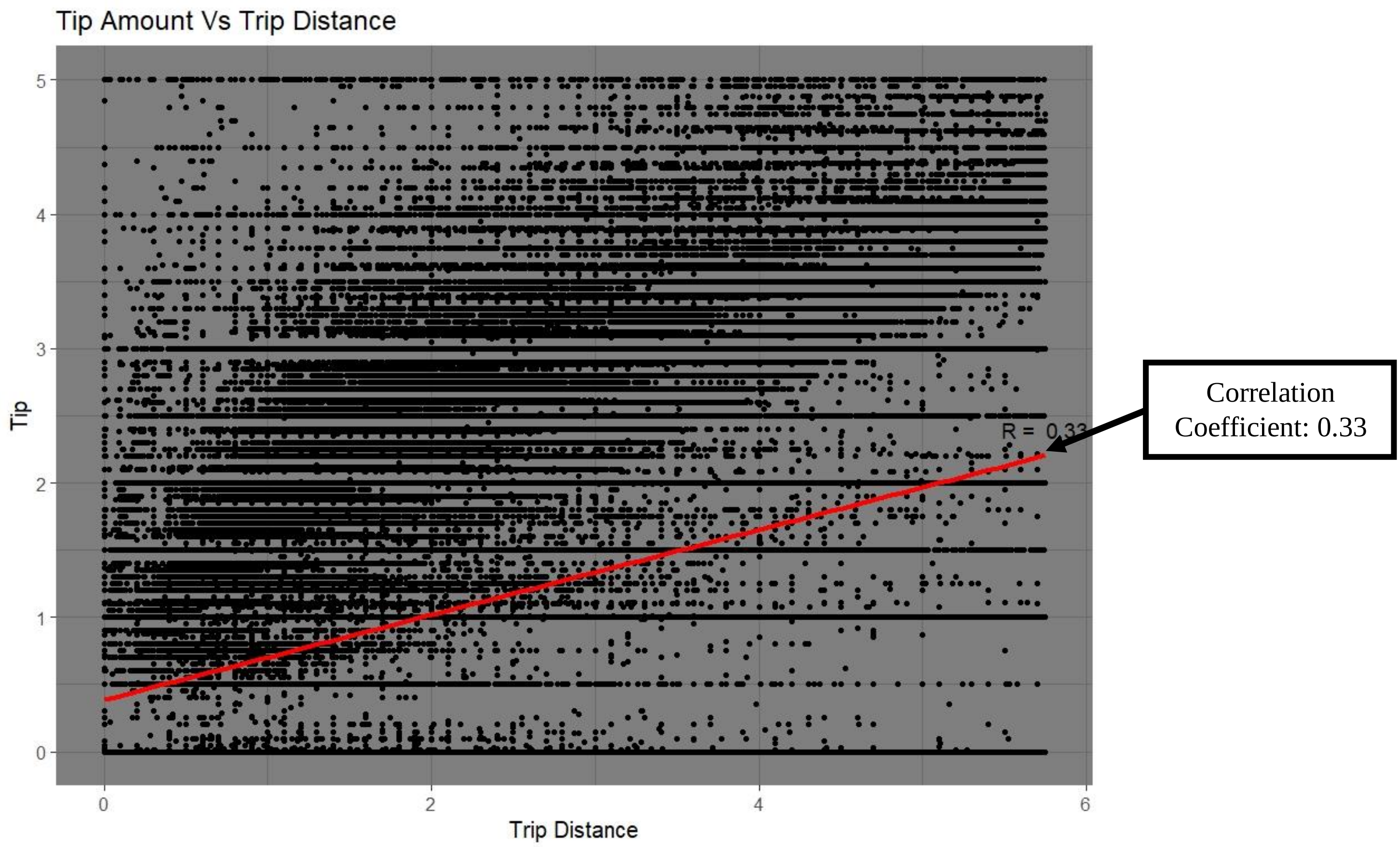
Figure 9: Correlation plot between Tips and the number of passengers

This is also a **scatter plot** made using ggplot2 on R. This graph shows the **tips** given for rides against the **passenger counts.**

Just like time of day there is **no correlation** between the number of passengers and the tips given. We can observe the gaps around integers and ½ dollar marks in this graph as well.

Also, there seems to be an error in the data. There is a taxi ride recorded with **no passengers** but a tip of **2.5 dollars**.

Tips vs Distance



Correlation Coefficient: 0.33

Figure 10: Correlation plot between Tips and trip distance

The above graph shows the **correlation** between **tips and the trip distance.** The graph is quite unusual to look at. Like the previous two graphs, the tips are all over the place. High tips for short trips, low tips for long trips and the other way around. But we can still see a weak pattern. Plotting a **Linear Regression** line, we see that the **correlation coefficient is 0.33**. This signifies a **very weak correlation.**

Tips vs Total charge

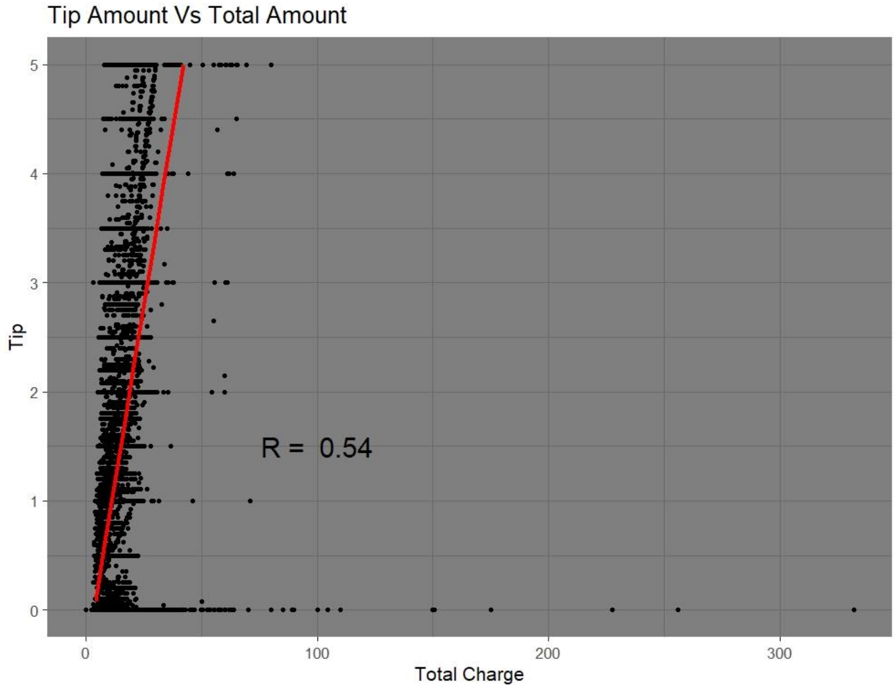


Figure 11: Correlation plot between tips and total charge

The **tip amounts** showed the most correlation with the total **taxi charge**. The **correlation coefficient is 0.54.** So, we can say that there is a moderate correlation between the two variables.

Although this leads to an interesting question. Since tips are part of the total charge how does this influence the bias of the graph. Since the tips are only a tiny portion of most of the cab rides, we can disregard this.

Taxi Hot Spots for Different Days

The following plots show **the pickup coordinates** of the taxi rides in the form of a **density map**. These graphs have been generated using **Tableau.** Since more than **90 percent** of the data is from **Manhattan**, we would get much more meaning full graphs by focusing only on that area.

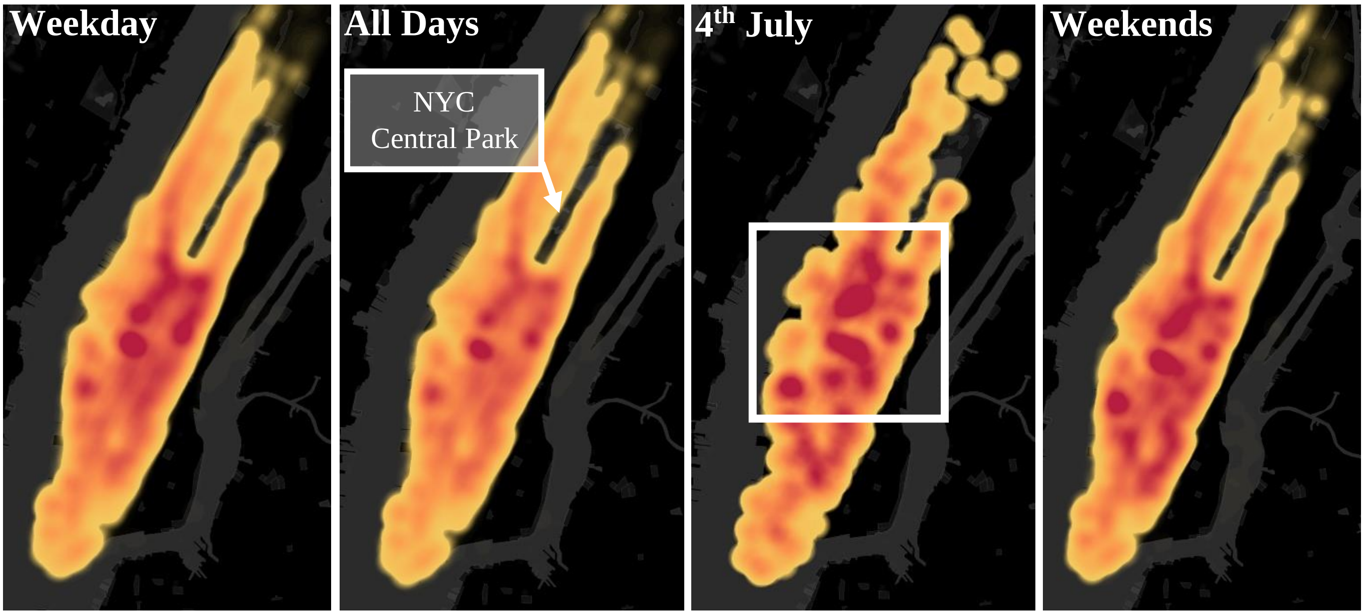


Figure 12: Tableau Density plots showing the taxi hotspots in Manhattan. Left to Right the maps shown are for Weekdays, All days, 4th July and Weekends

These graphs show a density map of taxi pick ups in the **Manhattan Borough.** From left to right the four graphs above show taxi densities for **Weekdays, All Days, 4th July and Weekends** respectively.

Observations

Weekdays and **All Days** naturally look similar since they are made of roughly the same data plus or minus 104 days for the weekends. The red regions are areas where the taxi traffic is highest, and the yellow regions are the other end of the spectrum. The long empty section in the middle is the **New York City Central Park**.

Most of the **high taxi traffic** areas lie to the **south of the park**. The area highlighted above in the 3rd map has been expanded in the figure below. These popular areas are even more crowded during the **holidays**. The **3rd and 4th maps** show the **4th of July** and the **weekends** respectively. A quick search online shows that there are a lot of popular tourist attractions in those areas and hence the high taxi volume. The 3 biggest spots clearly visible in the figure below are centred around **Times Square** [5], **Chrysler Building** [6] and the **Empire State Building** [7] (going clockwise from the top).

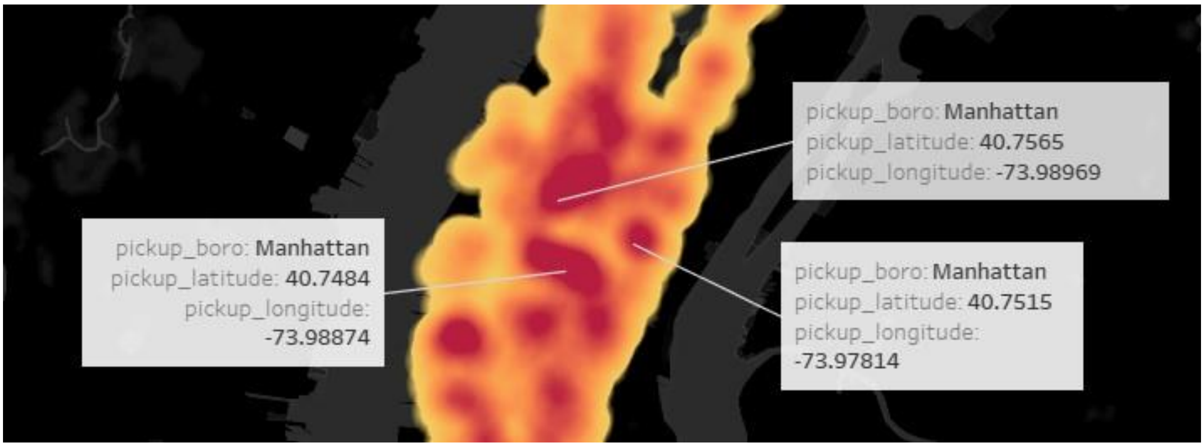
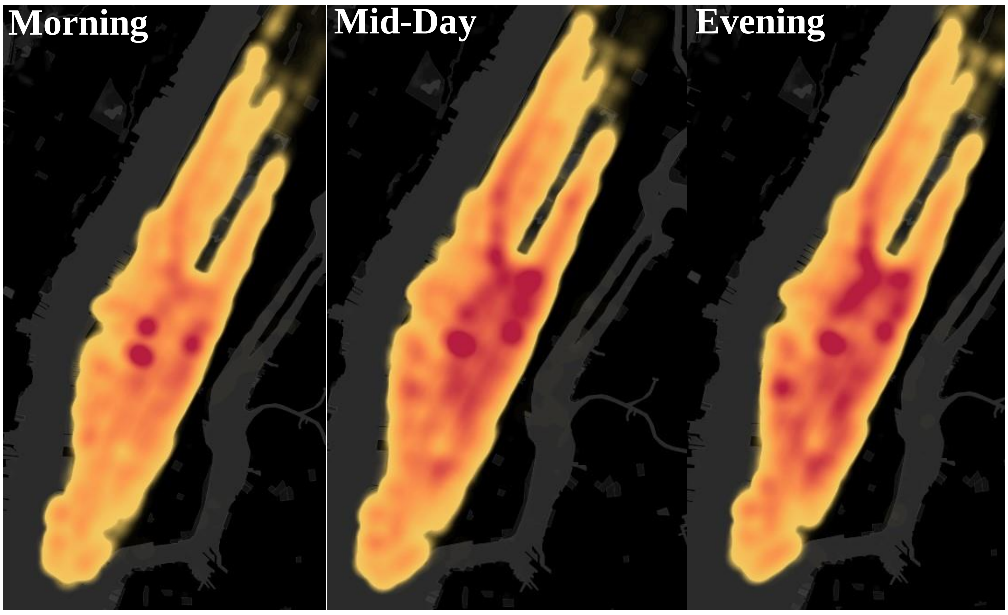


Figure 13: Magnification of the 3rd map in figure 12. Shows the locations of 3 major tourist attractions in NYC

Taxi Hot Spots during different times of the day

The graphs below show the same area throughout the year but at different times of the day. From **left to right** the graphs show the **taxi traffic densities** in the **Morning (6am to 10am), Mid-Day (11am to 4pm) and Evening (5pm to 10pm)**



Morning times show the least taxi traffic. Which is no surprise. Afternoon and evening are much denser with evening being slightly more spread out. We can see from the evening map that the taxi rides shift southwards. These are residential areas and have a few parks and other places that might be visited by people after work or kids after school that leads to the high taxi density.

Conclusion

We explored two aspects of the data set and answered two questions. The first one concerned the tips. We saw that the **amount of tip** given had **no correlation** with the **number of passengers or the time of day**. On the other hand, tips were **very weakly positively correlated** with the **trip distance** and **moderately correlated** to the **total charge.**

The other aspect we explored is the **GPS coordinate data**. We plotted a **density map** on **Tableau** to visualize how the taxi hotspots around Manhattan change with different dates. Manhattan has a lot of historical landmarks and we were able to identify a few such landmarks based on the coordinates of the taxi hot spots.

Reflection

This project has improved my proficiency with Tableau and R graphics. It would be interesting to use the full dataset with all the 165 million records. Online services from Google or Amazon might help deal with the large size of the data.

I wanted to use Reverse Geocoding to convert the GPS coordinates into addresses. This would give us the local county names and that could be used to generate more aggregated visualizations. But the process needed an online API from Google or any other service provider and the daily transactions were limited to 2,500. Unlimited services were also available but the conversion speed was too low, going at the rate of 2 records every second. It would take too long to convert all the values we have.

References

Data Sources

[1] ipython-books/minibook-2nd-data. (2020). Retrieved 28 April 2020, from https://github.com/ipython-books/minibook-2nd-data/blob/master/nyc_taxi.zip

[2] Borough Boundaries. (2020). Retrieved 28 April 2020, from https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm

Other Information

[3] Manhattan Population 2020. (2020). Retrieved 28 April 2020, from https://worldpopulationreview.com/boroughs/manhattan-population/

[4] Brooklyn Population 2020. (2020). Retrieved 28 April 2020, from https://worldpopulationreview.com/boroughs/brooklyn-population/

[5] Empire State Building, New York, USA. (2020). Retrieved 28 April 2020, from https://www.latlong.net/place/empire-state-building-new-york-usa-5312.html

[6] Chrysler Building, New York, NY, USA. (2020). Retrieved 28 April 2020, from https://www.latlong.net/place/chrysler-building-new-york-ny-usa-14548.html

[7] Times Square, Manhattan, NYC, USA. (2020). Retrieved 28 April 2020, from https://www.latlong.net/place/times-square-manhattan-nyc-usa-7560.html