New York Taxi Data Exploration

Zi Wei

References for code are from: https://rpubs.com/shivanandiyer/BigRQuery

Information about Database

Taxi trips are recorded by the New York Taxi and Limousine Commision (TLC). There are 3 main categories for taxi trips: Yellow Medallion Taxicabs (yellow), For Hire Vehicles (fhv), and Street Hail Livery (green).

Yellow taxis provide transportation exclusively through street hails while fhv taxis provide transport exclusively through pre-arranged rides. Green taxis are a hybrid between yellow and fhv taxis in that they can accept hails in certain areas of New York City but can also handle pre arranged rides.

For this report I will be doing exploratory data analysis on the 2016 data for yellow taxis. The sql code to get the dataset is shown below:

```
"SELECT *
FROM [bigquery-public-data:new_york.tlc_yellow_trips_2016]
WHERE pickup_longitude is not NULL and pickup_latitude is not NULL
and dropoff_latitude  is not Null and dropoff_longitude is not NULL
and trip_distance >0"
```

```
## [1] "SELECT *  \nFROM [bigquery-public-data:new_york.tlc_yellow_trips_2016]\nWHERE pickup_longitude i
```

I used all rows (no limit) since I wanted to capture as much information as possible for the initial data exploration. I have also decided to take out Null values for pickup and dropoff related columns and any trips that are less than 0. This is done so that I can get as much complete information as possible.

Initial Data Exploration

Total Amount

```
## # Source:   lazy query [?? x 5]
## # Database: BigQueryConnection
##    Mean    Sd Minimum Maximum Number_of_Rows
##   <dbl> <dbl>   <dbl>   <dbl>          <int>
## 1  16.1  49.3   -958. 187442.       68982209
```

The minimum for Total fare is negative, this should be a recording error. Hence we will need to remove these values. At the same time, there are extremely large values in the dataset, with the maximum being more than 10000 times the mean. Such values could cause a shift in the standard deviation, leading to a very high standard deviation of 49, compared to the 16 of the mean.

Fare Amount

```
## # Source:   lazy query [?? x 5]
## # Database: BigQueryConnection
##    Mean    Sd Minimum Maximum Number_of_Rows
##   <dbl> <dbl>   <dbl>   <dbl>          <int>
## 1  12.8  48.7   -958. 187441.       68982209
```

In fare amount, we see that there is a negative minimum which does not make sense and should be an error in recording. We can also see that we have a very big maximum for fare amount, 187441. This means that either the distance travelled was extremely far or that there was an error in the meter.

Tolls Amount

```
## # Source:   lazy query [?? x 5]
## # Database: BigQueryConnection
##    Mean    Sd Minimum Maximum Number_of_Rows
##   <dbl> <dbl>   <dbl>   <dbl>          <int>
## 1 0.314  1.67  -100.0   1410.       68982209
```

From our data, there are negative values for the minimum toll amount paid which should be an error in recording.

Trip Distance

```
## # Source:   lazy query [?? x 5]
## # Database: BigQueryConnection
##    Mean    Sd Minimum   Maximum Number_of_Rows
##   <dbl> <dbl>   <dbl>     <dbl>          <int>
## 1  4.93 4036.    0.01 19072629.       68982209
```

The trip distance has a relatively low mean of 4.92 miles and a high standard deviation of 4036 miles. This is probably due to the presence of very large values for trip distance, as indicated by the maximum 19072629 miles which is more than 3 million times the mean. This high trip distance recorded may explain the high maxmum fare amount we observed earlier, which is determined by the trip distance.

Large Values: Fare Amount

In this section, we will try and understand the makeup of the large values we have observed for Fare amount.

```
## # Source:   lazy query [?? x 17]
## # Database: BigQueryConnection
##   Mean_fare Sd_fare Minimum_fare Maximum_fare Mean_distance Sd_distance
##       <dbl>   <dbl>        <dbl>        <dbl>         <dbl>       <dbl>
## 1   123415.  47464.       20044.      187441.       134468.     137817.
## # ... with 11 more variables: Minimum_distance <dbl>,
## #   Maximum_distance <dbl>, Mean_toll <dbl>, Sd_toll <dbl>,
## #   Minimum_toll <dbl>, Maximum_toll <dbl>, Mean_total <dbl>,
## #   Sd_total <dbl>, Minimum_total <dbl>, Maximum_total <dbl>,
## #   Number_of_Rows <int>
```

```
## # Source:   lazy query [?? x 17]
## # Database: BigQueryConnection
##   Mean_fare Sd_fare Minimum_fare Maximum_fare Mean_distance Sd_distance
##       <dbl>   <dbl>        <dbl>        <dbl>         <dbl>       <dbl>
## 1    24214.  50152.         1003      187441.        23782.      75610.
## # ... with 11 more variables: Minimum_distance <dbl>,
## #   Maximum_distance <dbl>, Mean_toll <dbl>, Sd_toll <dbl>,
## #   Minimum_toll <dbl>, Maximum_toll <dbl>, Mean_total <dbl>,
## #   Sd_total <dbl>, Minimum_total <dbl>, Maximum_total <dbl>,
## #   Number_of_Rows <int>
```

```
## # Source:   lazy query [?? x 17]
## # Database: BigQueryConnection
##   Mean_fare Sd_fare Minimum_fare Maximum_fare Mean_distance Sd_distance
##       <dbl>   <dbl>        <dbl>        <dbl>         <dbl>       <dbl>
## 1      184.   2233.         100.      187441.          68.0       3177.
## # ... with 11 more variables: Minimum_distance <dbl>,
## #   Maximum_distance <dbl>, Mean_toll <dbl>, Sd_toll <dbl>,
## #   Minimum_toll <dbl>, Maximum_toll <dbl>, Mean_total <dbl>,
## #   Sd_total <dbl>, Minimum_total <dbl>, Maximum_total <dbl>,
## #   Number_of_Rows <int>
```

Above, we show the summary statistics for fare amount for rides with a fare more than 10000 and 1000,100 respectively. The first thing that we notice is that the number of rows can be small, with 9 for fare more than 10000 and 51 for fare more than 1000. Their relatively small numbers seems to indicate that they are a special group of possible outliers in the data. For rides with a fare of more than 100, we have 31162 of such rides, however, this is less than 1% of the data in the database.

For this report, I will consider a fare greater than 100 as high as it is about 2 standard deviations away from the mean value of 12.84.

We observe that there is a minimum distance of 0.1 miles despite the high fare of at least 100.

Fare More than 1000, Short Distance

```
## # Source:   lazy query [?? x 4]
## # Database: BigQueryConnection
##   Mean_fare Mean_toll Mean_total Number_of_Rows
##       <dbl>     <dbl>      <dbl>          <int>
## 1     2517.         0      2518.              4
```

There were 4 rides where the fare amount was more than 1000 but the distance travelled was just 0.1 miles. This seems to be suspicious and warrants further investigation.

The following table will explore these 4 rides in detail.

```
## # A tibble: 4 x 10
##   fare_amount trip_distance total_amount tolls_amount pickup_datetime
##         <dbl>         <dbl>        <dbl>        <dbl> <dttm>
## 1       2020.           0.1        2021.            0 2016-02-12 15:49:02
## 2       2020.           0.1        2021.            0 2016-02-12 15:20:09
## 3       2020.           0.1        2021.            0 2016-02-12 15:28:38
## 4       4008.           0.1        4009.            0 2016-02-27 18:19:57
## # ... with 5 more variables: dropoff_datetime <dttm>,
```

```
## #   pickup_latitude <dbl>, pickup_longitude <dbl>, dropoff_latitude <dbl>,
## #   dropoff_longitude <dbl>
```

The pattern shared by these 4 rides is that they all occur in the month of February. More specifically, 3 of the 4 rides occur at about 1500 hours on the same day of 12 February and charge the same fare amount at 2020.37.

A map view of these 4 points is shown below

The numbers indicate the 4 different rides, the pickup point is in red while the dropoff point is in blue. It is clear that there seems to be no observable pattern between the 4 rides as their pickup points and destination points are spaced far apart with no common destination.

We can also see that the third ride, which occurs at 1800 on 27 February travels very far compared to the others. While its distance is recorded at 0.1 miles (160.1 meters), the distance travelled could not have been that short. This record is most likely a mistake, and the distance travelled was definitely further than 0.1 miles.

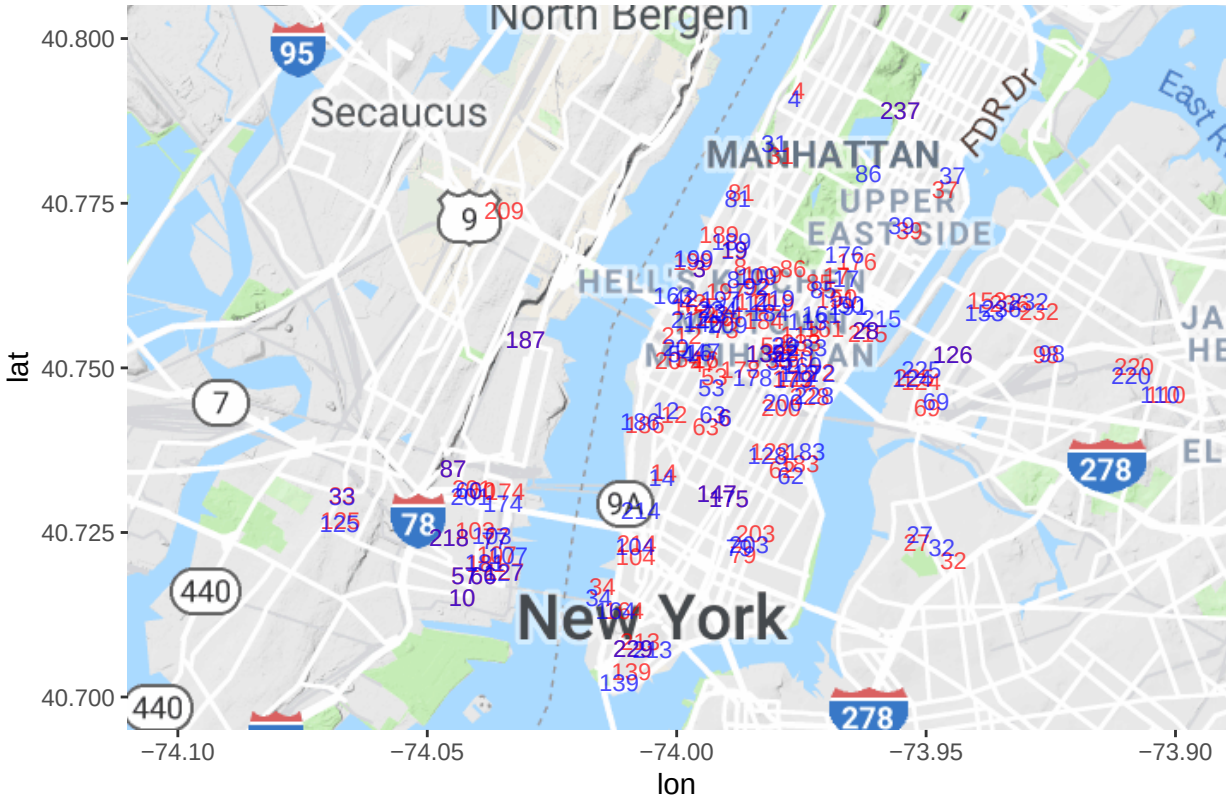
1500 hours on 12 February 2016 was a special time as we had 3 cases of rides going for just 0.1 miles but charged 2020.37 which is more than 100 times the mean fare amount at 12.84.

Fare More than 100, Short Distance

We will now widen the scope of visualization and look at rides where the fare was more than 100 but the distance travelled was just 0.1 miles to see if we can identify a pattern in the data.

The red dots in the figure above represent pickup points while the blue dots represent dropoff points. The pickup and dropoff points seem to cluster near 3 points: Midtown Manhattan, south of state route 9A, and state route 78.

In the figure below, we will take a closer look at the pickup and dropoff points. Now, pickup points and dropoff points are numbered. A blue colored number represents a dropoff point and a red colored number represents a pickup point.



From the figure, we see that in most cases, the distance between pickup and dropff are close due to the low distance tavelled of 0.1 miles. However, there is a clear anomalous result, 208, whose pickup and dropoff points seem to suggest a distance greater than 0.1 miles as we can see from the figure below.

Outside of this one anomalous result, there really does not seem to be a clear explanation as to why the fares are so high for a short distance trip based on the location of pickup and dropoffs.

Another possible reason for the high fare despite travelling only a short distance would be the time of the ride. In the section "Fare More than 1000, Short Distance", we saw that 3 of rides of the 4 rides we identified travelled at about 1500 hours on the same day, 12 February 2016. The fare could be inflated greatly due to slow traffic.

```
## # A tibble: 10 x 11
##    fare_amount trip_distance total_amount tolls_amount pickup_longitude
##          <dbl>         <dbl>        <dbl>        <dbl>            <dbl>
##  1         126           0.1         152.            0            -73.8
##  2         140           0.1         140.            0            -73.9
##  3         350           0.1         420.            0            -74.0
##  4        2020.          0.1        2021.            0            -74.0
##  5         180.          0.1         200.            0            -74.4
##  6         180           0.1         200.            0            -74.0
##  7         111           0.1         134.            0            -74.3
##  8         180           0.1         180.            0            -74.0
##  9         120           0.1         125.            0            -73.7
## 10         120           0.1         145.           15            -74.0
## # ... with 6 more variables: pickup_latitude <dbl>,
## #   pickup_datetime <dttm>, dropoff_longitude <dbl>,
## #   dropoff_latitude <dbl>, dropoff_datetime <dttm>, number <int>
```

```
## Table of Months where Fare>100, Trip distance=0.1
```

```
##
##  1  2  3  4  5  6
## 49 29 40 32 37 51
```

We can see that the most of these rides occur in the first half of the year. However, if we check in the database, the data is only recorded till June, so we should expect data up till June. The sql query used to extract the finding is shown below:

```
"SELECT unique(MONTH(pickup_datetime)) as mon
FROM [local-concord-245816:project1.yellow_taxi_2016b]"
```

```
## [1] "SELECT unique(MONTH(pickup_datetime)) as mon \nFROM [local-concord-245816:project1.yellow_taxi_
```

We take a look at date month combos to see if we can discover anything special.

```
## Table of Dates where Fare>100, Trip distance=0.1
```

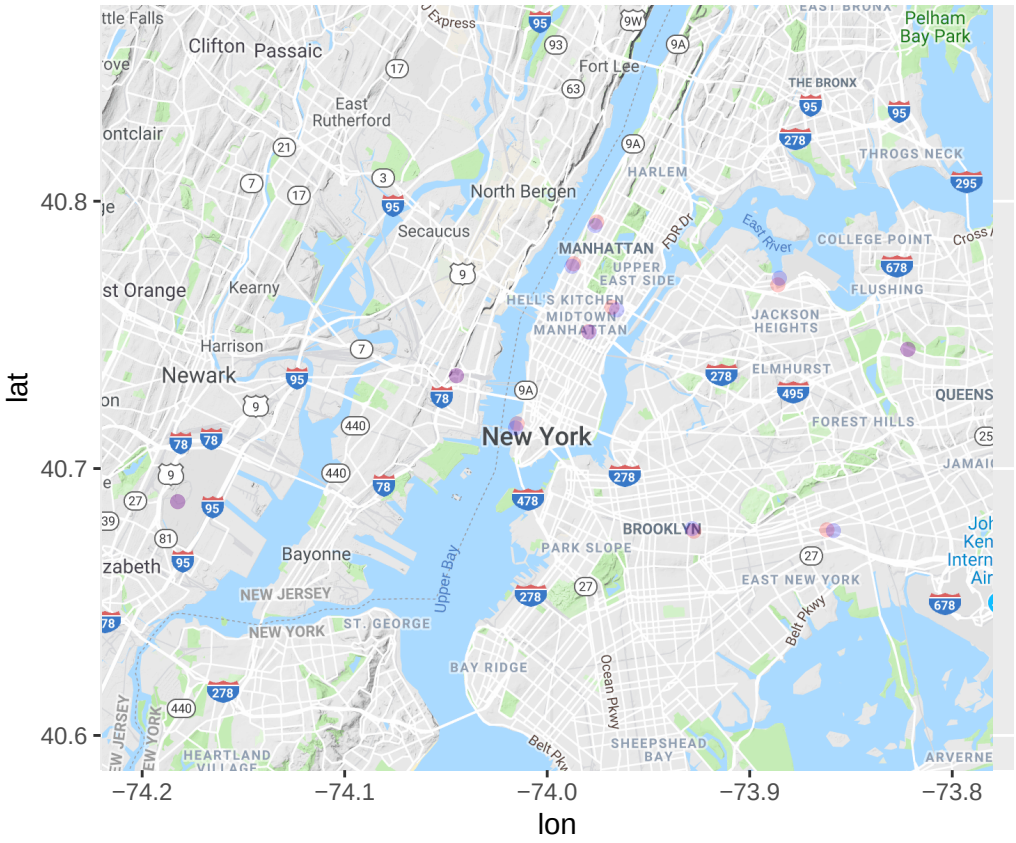
```
##
## 01/01 01/02 01/03 01/04 01/05 01/06 01/07 01/09 01/12 01/13 01/14 01/15
##     6     2     1     2     2     4     1     1     2     1     3     1
## 01/16 01/19 01/20 01/21 01/22 01/23 01/24 01/25 01/26 01/27 01/28 01/29
##     1     3     1     2     1     1     2     4     3     1     2     1
## 01/30 02/02 02/04 02/06 02/08 02/09 02/12 02/14 02/15 02/17 02/18 02/19
##     1     1     1     2     5     2     3     1     1     1     1     2
## 02/21 02/23 02/25 02/26 02/27 02/28 03/01 03/02 03/03 03/06 03/09 03/11
##     2     1     2     2     1     1     3     1     2     3     2     2
## 03/12 03/14 03/15 03/17 03/19 03/20 03/21 03/22 03/23 03/24 03/25 03/26
##     2     1     1     1     2     1     3     1     3     1     2     2
## 03/27 03/29 03/30 03/31 04/01 04/02 04/03 04/04 04/05 04/08 04/09 04/10
##     2     3     1     1     1     2     1     1     1     3     1     1
## 04/12 04/14 04/15 04/16 04/17 04/19 04/20 04/22 04/23 04/25 04/26 04/27
##     2     2     1     1     1     2     1     4     1     1     1     1
## 04/28 04/29 05/01 05/03 05/04 05/07 05/08 05/10 05/11 05/12 05/13 05/14
##     1     2     2     1     1     2     1     1     1     2     2     4
## 05/15 05/16 05/17 05/18 05/19 05/20 05/21 05/22 05/25 05/26 05/27 05/28
##     1     1     3     1     3     1     2     1     1     2     1     1
## 05/29 05/31 06/01 06/02 06/03 06/05 06/06 06/07 06/08 06/09 06/10 06/11
##     1     1     1     2     2     2     3     1     3     2     3     1
## 06/13 06/14 06/15 06/17 06/18 06/20 06/21 06/22 06/23 06/24 06/25 06/27
##     4     2     2     2     4     1     2     1     3     1     1     2
## 06/28 06/29 06/30
##     2     1     3
```

```
## Table of Hours where Fare>100, Trip distance=0.1
```

```
##
## 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
## 13 19 13 13 13  4  2  5  3  6  8 13  6  5  6 15  6 13  8  5  7 13 21 21
```

From the table of dates, we can see that there seems to be no special date where the fares were just higher than usual. However, when we look at the hours where these occur, we can see that many of them occur

from 0000 to 0400, 1100, 1500, 1700 and 2100 to 2300. The only timing that seems special is 1500, which we showed earlier contain 3 of the highest fares for a distance travelled of 0.1 miles.



A quick look reveals no pattern for the rides that occured at 1500 hours. So my conclusion is that 1500 is probably a special hour in New York, where it is similar to a rush hour and as a result many taxis charged high fares.

Length of a Drive

We saw in the section Trip Distance that the maximum distance travelled by a single ride is 19072629 miles.

```
## # A tibble: 1 x 8
##   fare_amount trip_distance pickup_datetime     pickup_longitude
##         <dbl>         <dbl> <dttm>                        <dbl>
## 1         2.5     19072629. 2016-03-07 19:57:50           -74.0
## # ... with 4 more variables: pickup_latitude <dbl>,
## #   dropoff_datetime <dttm>, dropoff_longitude <dbl>,
## #   dropoff_latitude <dbl>
```

We can see that there is acutally an error in the record for this trip distance. The ride lasted about 18 minutes from pickup to dropoff and it is not possible to have travelled 19072629 miles in that time period.

With the possiblity of such errors being present, it might be better to estimate the length of a ride using the duration of the ride instead. The sql code used to generate the table being queried is as follows:

```
"SELECT fare_amount,trip_distance,dropoff_longitude,dropoff_latitude,pickup_longitude,pickup_latitude,T
FROM `project1.yellow_taxi_2016b`"
```

```
## [1] "SELECT fare_amount,trip_distance,dropoff_longitude,dropoff_latitude,pickup_longitude,pickup_lati
```

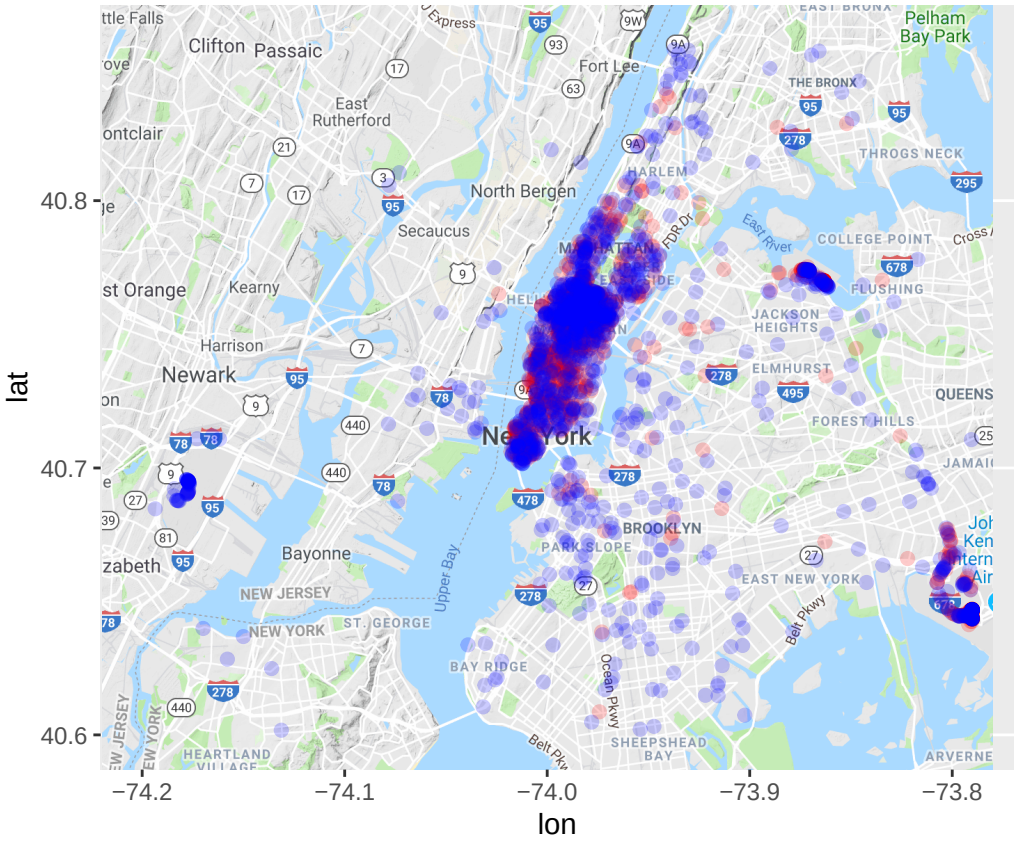
Time taken for a Ride

```
## # Source:   lazy query [?? x 5]
## # Database: BigQueryConnection
##   Mean_time Sd_time Min_time Max_time Number_of_Rows
##       <dbl>   <dbl>    <int>    <int>          <int>
## 1      13.5    10.8        1       96       68621042
```

The summary statistics for the time needed for a ride to be completed (difference between dropoff and pickup time) are shown above. For the data, I have filtered out rides that go for more than 8 hours on the assumption that the taxi driver works about 8 hours a day. This filtering is done in an effort to further clean the dataset as we have noticed unrealistically long amounts of time being taken taken for a ride.

Under these conditions, the mean duration of 1 ride is about 13.5 minutes, the shortest ride is 1 minute while the longest ride is 96 minutes which is about 8 hours. The standard deviation of the time taken for each ride is about 11 minutes.

The graph below shows the pickup points (red) and dropoff points (blue) for rides that took 96 minutes.



From the figure above, we can see that most pickup and dropoff points occur in the city. There is a small cluster of points near the airport at (-73.8,40.66). However, most of these points seem to indicate a distance that requires less than 8 hours.

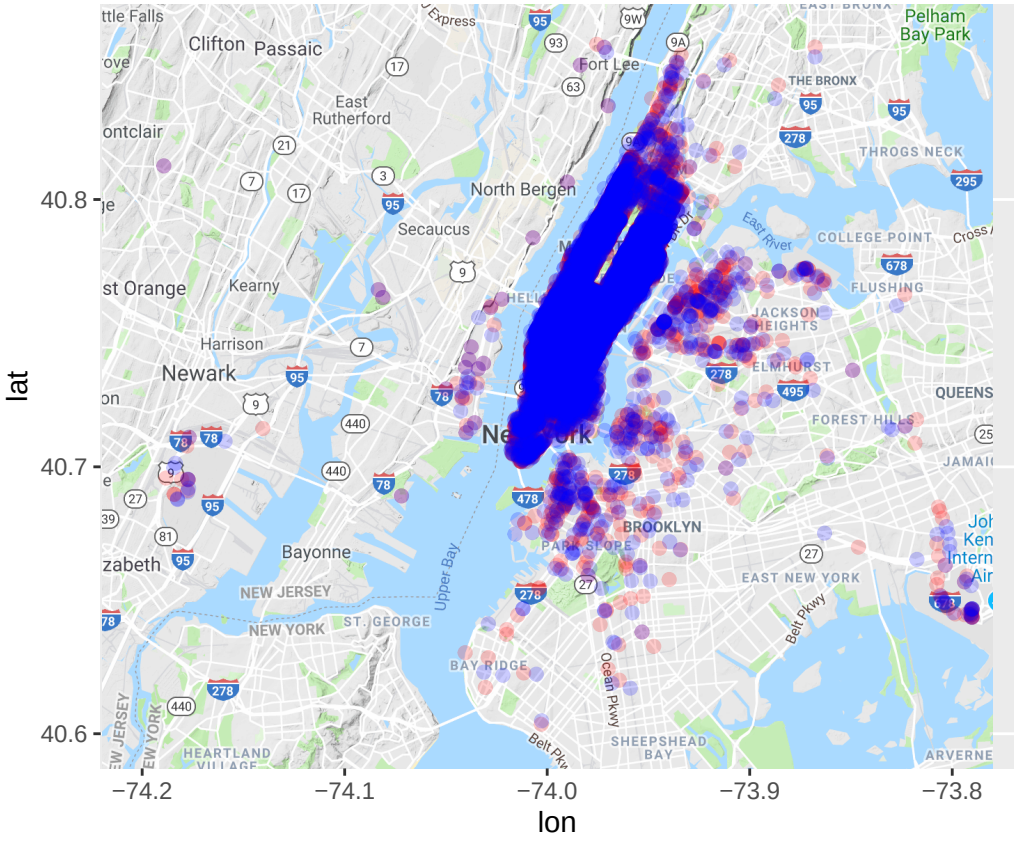
Below, we will show the summary statistics for fare amount for trips. The mean of $52 for a fare seems to suggest that most of these trips go only as far as the airport from New York city (vice versa).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   52.00   52.00   59.55   57.50  450.00
```

We can see that there is a maximum fare of 450 and upon checking, it is revealed that this driver went all the way from New York city to Harriman state park. However, he seems to be the exception to our observation that the the length of a ride does not seem to add up with the fare amount generated.

A possible explanation for this unexpected observation is that the drivers were waiting or parked at a location for a very long time before picking up a passenger. It is possible that they had the timer on from waiting or parking all the way until a passenger is picked up, from which the ride went as long as the fare indicates.

The next figure explores the short rides that occured (time taken=1).



Based on a 10000 row sample, we can see that most of these rides occured in the City, which is expected. The summary statistics of the fares for short rides is shown below.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.010    3.500    3.500    4.036    3.500 2550.200
```

Even with a short ride 1 minute, a driver can make 4 dollars on average. We also notice a very large fare amount of over 2000. We shall take a look at this ride in particular below:

It is clear that fare amount has nothing to do with the distance travelled or time taken for the ride in this case.

Based on the cases we have observed so far for short rides (by time or distance) with high fares, a speculation I have for this occurence is that the taxi drivers only engaged the meter for 1 ride but reported their total earnings for a large number of rides. However, if this does not turn out to be true, and that the driver did indeed make a large amount of money off of 1 quick drive, I would still not bank on it as such occurences are few and far between: Only 5 out of the 872941 rides which took a minute got a fare amount of over 1000.

Conclusion

This report highlights several interesting cases for rides that occured in New York City. In particular, short rides that made more than 100 in fares. A 1 minute ride hardly seems silly as it indicates a place within walking distance. The fact that there are rides that costed 100 dollars just to ferry a passenger a short distance away seems otherwordly.

We also discovered a special hour in New York city, 1500 where there seems to be a higher occurence of rides which costed over 100 dollars but went for just 0.1 miles.

We also found out about a driver who went on an 8 hour journey to ferry a passenger from New York city to Harriman state park. The fact that this was a ride hail and not a pre-arranged drive is surprising. The passenger must have just decided to go on a journey, and the taxi driver must be really dedicated to his work since an 8 hour drive to the park meant and 8 hour drive back to the city.