CPE 351: Big Data Experience

New York Trip Duration

Prepared by

Thanchanok	Eiamsakulchai	57070503416
Wasunan	Rojkanok	57070503432
Pacharapol	Boonrut	57070503455
Pichaya	Piyawanmongkon	57070503456
Tantikorn	Phuprasurt	57070503468

Submitted to

Assoc. Prof. Dr.Tiranee Achalakul
Asst Prof Dr.Santitham Prom-on

B. Eng. Computer Engineering
Academic Year 2017
King Mongkut's University of Technology Thonburi

Introduction

New York City Taxi Trip Duration

The competition dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this playground competition. Based on individual trip attributes, participants should predict the duration of each trip in the test set. **What we will predict from this dataset is trip duration of New York city taxi based on given features.**

Given Data fields

There are 2 datatsers which are

1. **train.csv** : the training set contains 1458644 trip records and 11 attributes

Attributes	Type	Decription
id	Character	a unique identifier for each trip
vendor_id	Integer	a code indicating the provider associated with the trip record
pickup_datetime	Character	date and time when the meter was engaged
dropoff_datetime	Character	date and time when the meter was disengaged
passenger_count	Integer	the number of passengers in the vehicle (driver entered value)
pickup_longtitude	Double	the longitude where the meter was engaged
pickup_latitude	Double	the latitude where the meter was engaged
dropoff_longitude	Double	the longitude where the meter was disengaged
dropoff_latitude	Double	the latitude where the meter was disengaged
store_and_fwd_flag	Character	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
trip_duration	Integer	duration of the trip in seconds

2. test.csv : the testing set contains 625134 trip records and 9 attributes

Attributes	Type	Decription
id	Character	a unique identifier for each trip
vendor_id	Integer	a code indicating the provider associated with the trip record
pickup_datetime	Character	date and time when the meter was engaged
passenger_count	Integer	the number of passengers in the vehicle (driver entered value)
pickup_longtitude	Double	the longitude where the meter was engaged
pickup_latitude	Double	the latitude where the meter was engaged
dropoff_longitude	Double	the longitude where the meter was disengaged
dropoff_latitude	Double	the latitude where the meter was disengaged
store_and_fwd_flag	Character	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip

Overview of the data sets

- **Train data**

```
> summary(train)
      id              vendor_id     pickup_datetime    dropoff_datetime   passenger_count pickup_longitude
 Length:1458644     Min.   :1.000   Length:1458644     Length:1458644     Min.   :0.000   Min.   :-121.93
 Class :character   1st Qu.:1.000   Class :character   Class :character   1st Qu.:1.000   1st Qu.: -73.99
 Mode  :character   Median :2.000   Mode  :character   Mode  :character   Median :1.000   Median : -73.98
                    Mean   :1.535                                         Mean   :1.665   Mean   : -73.97
                    3rd Qu.:2.000                                         3rd Qu.:2.000   3rd Qu.: -73.97
                    Max.   :2.000                                         Max.   :9.000   Max.   : -61.34
 pickup_latitude dropoff_longitude dropoff_latitude store_and_fwd_flag trip_duration
 Min.   :34.36   Min.   :-121.93   Min.   :32.18    Length:1458644     Min.   :      1
 1st Qu.:40.74   1st Qu.: -73.99   1st Qu.:40.74    Class :character   1st Qu.:    397
 Median :40.75   Median : -73.98   Median :40.75    Mode  :character   Median :    662
 Mean   :40.75   Mean   : -73.97   Mean   :40.75                       Mean   :    959
 3rd Qu.:40.77   3rd Qu.: -73.96   3rd Qu.:40.77                       3rd Qu.:   1075
 Max.   :51.88   Max.   : -61.34   Max.   :43.92                       Max.   :3526282
```

```
> glimpse(train)
Observations: 1,458,644
Variables: 11
$ id                <chr> "id2875421", "id2377394", "id3858529", "id3504673", "id2181028", "id0801584"...
$ vendor_id         <int> 2, 1, 2, 2, 2, 2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 1, 2, 1, 2, 1, 1, 2, 2...
$ pickup_datetime   <chr> "2016-03-14 17:24:55", "2016-06-12 00:43:35", "2016-01-19 11:35:24", "2016-0...
$ dropoff_datetime  <chr> "2016-03-14 17:32:30", "2016-06-12 00:54:38", "2016-01-19 12:10:48", "2016-0...
$ passenger_count   <int> 1, 1, 1, 1, 1, 6, 4, 1, 1, 1, 1, 4, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ pickup_longitude  <dbl> -73.98215, -73.98042, -73.97903, -74.01004, -73.97305, -73.98286, -73.96902,...
$ pickup_latitude   <dbl> 40.76794, 40.73856, 40.76394, 40.71997, 40.79321, 40.74220, 40.75784, 40.797...
$ dropoff_longitude <dbl> -73.96463, -73.99948, -74.00533, -74.01227, -73.97292, -73.99208, -73.95741,...
$ dropoff_latitude  <dbl> 40.76560, 40.73115, 40.71009, 40.70672, 40.78252, 40.74918, 40.76590, 40.760...
$ store_and_fwd_flag <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "...
$ trip_duration     <int> 455, 663, 2124, 429, 435, 443, 341, 1551, 255, 1225, 1274, 1128, 1114, 260, ...
```

- **Test data**

```
> summary(test)
      id              vendor_id     pickup_datetime    passenger_count pickup_longitude  pickup_latitude
 Length:625134      Min.   :1.000   Length:625134      Min.   :0.000   Min.   :-121.93   Min.   :37.39
 Class :character   1st Qu.:1.000   Class :character   1st Qu.:1.000   1st Qu.: -73.99   1st Qu.:40.74
 Mode  :character   Median :2.000   Mode  :character   Median :1.000   Median : -73.98   Median :40.75
                    Mean   :1.535                      Mean   :1.662   Mean   : -73.97   Mean   :40.75
                    3rd Qu.:2.000                      3rd Qu.:2.000   3rd Qu.: -73.97   3rd Qu.:40.77
                    Max.   :2.000                      Max.   :9.000   Max.   : -69.25   Max.   :42.81
 dropoff_longitude dropoff_latitude store_and_fwd_flag
 Min.   :-121.93   Min.   :36.60    Length:625134
 1st Qu.: -73.99   1st Qu.:40.74    Class :character
 Median : -73.98   Median :40.75    Mode  :character
 Mean   : -73.97   Mean   :40.75
 3rd Qu.: -73.96   3rd Qu.:40.77
 Max.   : -67.50   Max.   :48.86
```

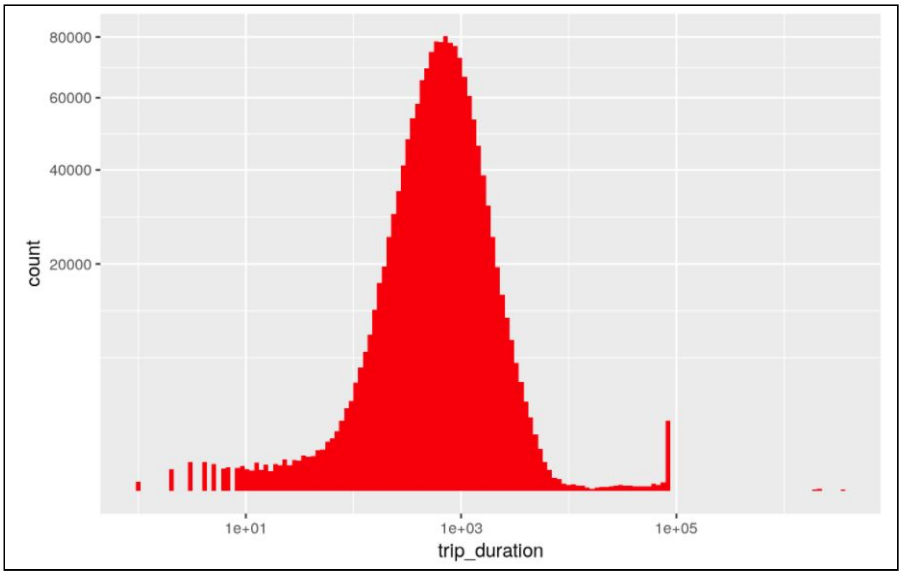
```
> glimpse(test)
Observations: 625,134
Variables: 9
$ id                <chr> "id3004672", "id3505355", "id1217141", "id2150126", "id1598245", "id0668992"...
$ vendor_id         <int> 1, 1, 1, 2, 1, 1, 1, 1, 2, 2, 1, 2, 1, 2, 1, 1, 2, 2, 2, 2, 1, 2, 2, 1, 2...
$ pickup_datetime   <chr> "2016-06-30 23:59:58", "2016-06-30 23:59:53", "2016-06-30 23:59:47", "2016-0...
$ passenger_count   <int> 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 4, 1, 1, 1, 1, 1, 1, 5, 1, 1, 1, 1, 5, 1, 1, 3...
$ pickup_longitude  <dbl> -73.98813, -73.96420, -73.99744, -73.95607, -73.97021, -73.99130, -73.97831,...
$ pickup_latitude   <dbl> 40.73203, 40.67999, 40.73758, 40.77190, 40.76147, 40.74980, 40.74155, 40.701...
$ dropoff_longitude <dbl> -73.99017, -73.95981, -73.98616, -73.98643, -73.96151, -73.98051, -73.95207,...
$ dropoff_latitude  <dbl> 40.75668, 40.65540, 40.72952, 40.73047, 40.75589, 40.78655, 40.71700, 40.719...
$ store_and_fwd_flag <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "...
```

Method

1. Exploration data analysis (EDA)

1.1. Individual visualization (relationship between the number of trips and each attribute)

a. Trip Duration



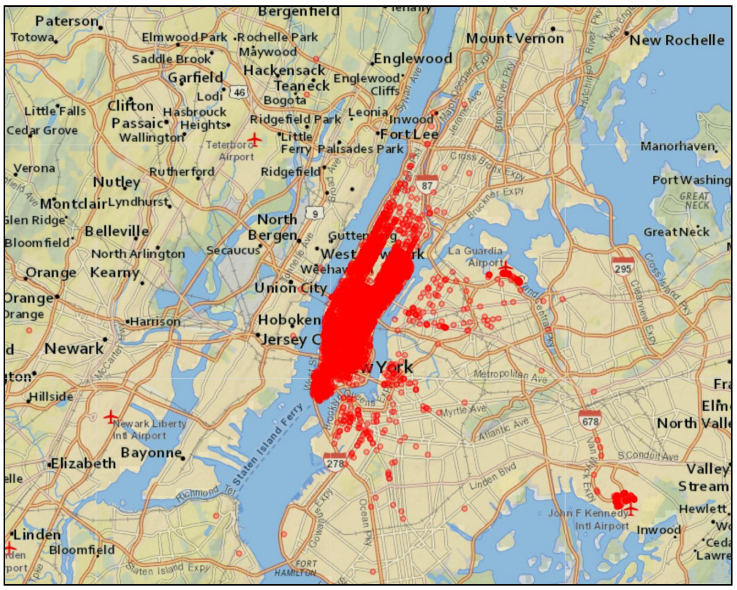
trip duration histogram

```
# A tibble: 10 x 11
   trip_duration     pickup_datetime    dropoff_datetime        id vendor_id passenger_count pickup_longitude
           <int>              <dttm>              <dttm>     <chr>     <fctr>          <fctr>            <dbl>
 1       3526282 2016-02-13 22:46:52 2016-03-25 18:18:14 id0053347         1               1        -73.78391
 2       2227612 2016-01-05 06:14:15 2016-01-31 01:01:07 id1325766         1               1        -73.98379
 3       2049578 2016-02-13 22:38:00 2016-03-08 15:57:38 id0369307         1               2        -73.92168
 4       1939736 2016-01-05 00:19:42 2016-01-27 11:08:38 id1864733         1               1        -73.78965
 5         86392 2016-02-15 23:18:06 2016-02-16 23:17:58 id1942836         2               2        -73.79453
 6         86391 2016-05-31 13:00:39 2016-06-01 13:00:30 id0593332         2               1        -73.78195
 7         86390 2016-05-06 00:00:10 2016-05-07 00:00:00 id0953667         2               1        -73.99601
 8         86387 2016-06-30 16:37:52 2016-07-01 16:37:39 id2837671         2               1        -73.99228
 9         86385 2016-06-23 16:01:45 2016-06-24 16:01:30 id1358458         2               1        -73.78209
10         86379 2016-05-17 22:22:56 2016-05-18 22:22:35 id2589925         2               4        -74.00611
# ... with 4 more variables: pickup_latitude <dbl>, dropoff_longitude <dbl>, dropoff_latitude <dbl>,
#   store_and_fwd_flag <chr>
```

Top 10 Maximum trip duration records

We use logarithmic for x-axis and square-root for y-axis. As you can see from the trip duration histogram and top 10 maximum trip duration records, there are some trips which is strange because it is less than 100 seconds and longer than one day which is highly unlikely .

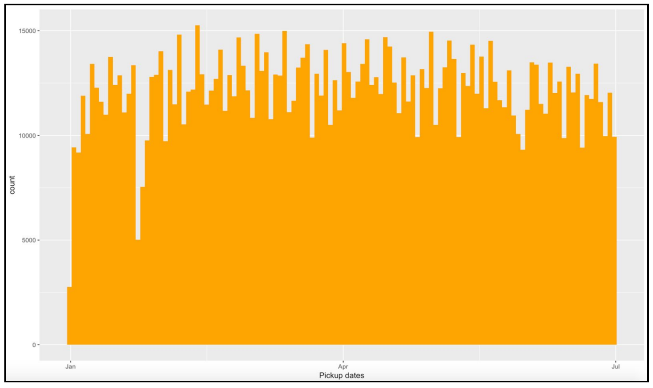
b. pickup_longtitude, pickup_latitude, dropoff_longtitude, dropoff_latitude



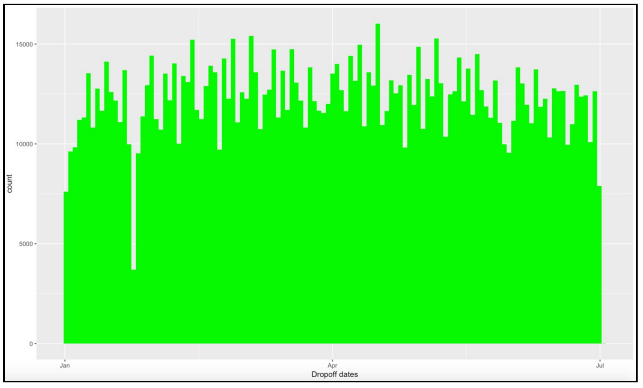
pickup and dropoff area

This graph is shown the pickup and dropoff area. Most of the trip is located in Manhattan and another popular places is JFK airport and LGA airport.

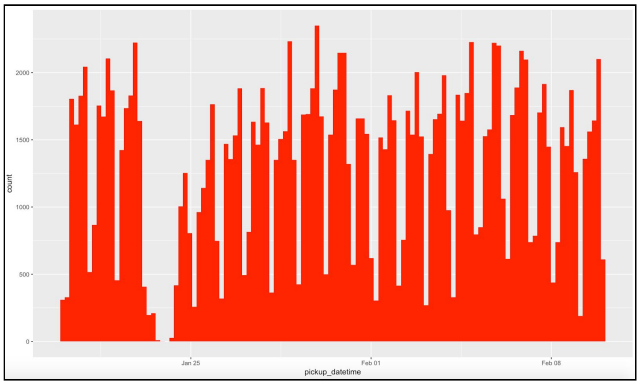
c. pickup and dropoff datetime



graph between the number of trips & pickup date from January to July 2016



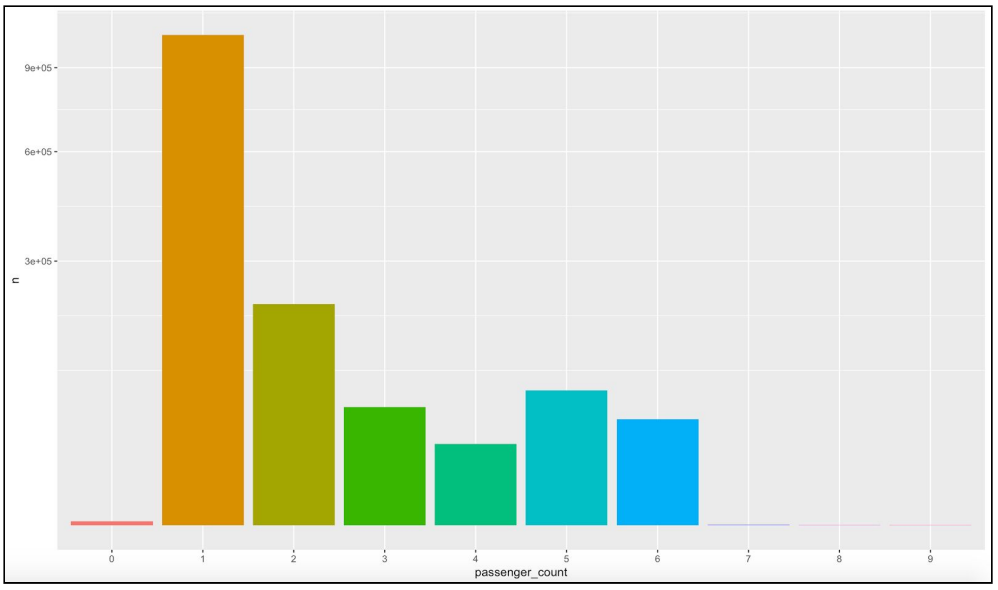
graph between the number of trips & dropoff date from January to July 2016



graph between the number of trips & pickup date between January and February 2016

Both graphs show the number of the trips during January 2016 to July 2016 for pickup and dropoff. Normally the number of trips is steady but it was drop rapidly around late January to early February from the zoom in graph from January and February 2016.

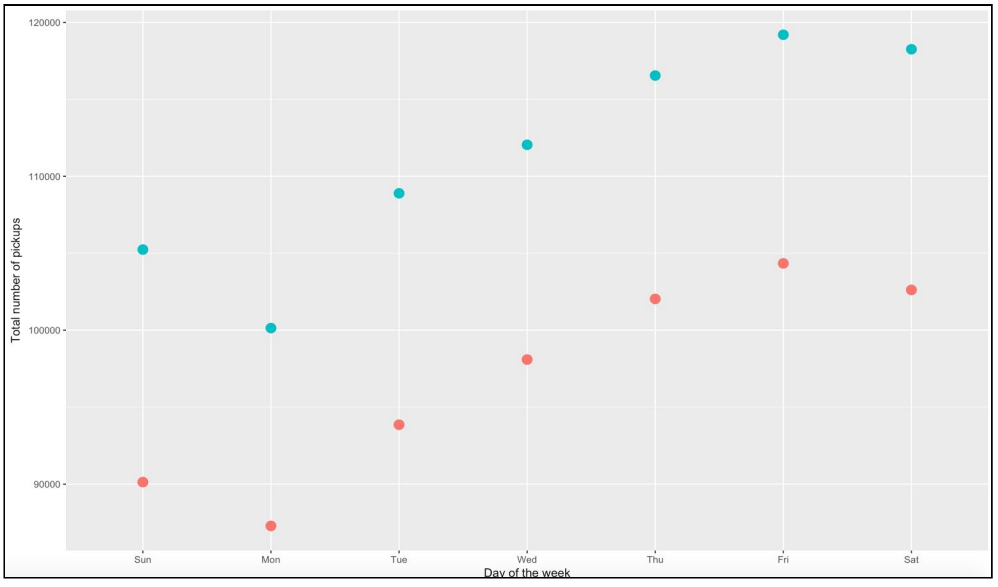
d. passenger counts



graph between the number of trips and passenger counts

From the graph, we noticed that there are some trips which has no passenger and also the vast majority of trips has only one passenger.

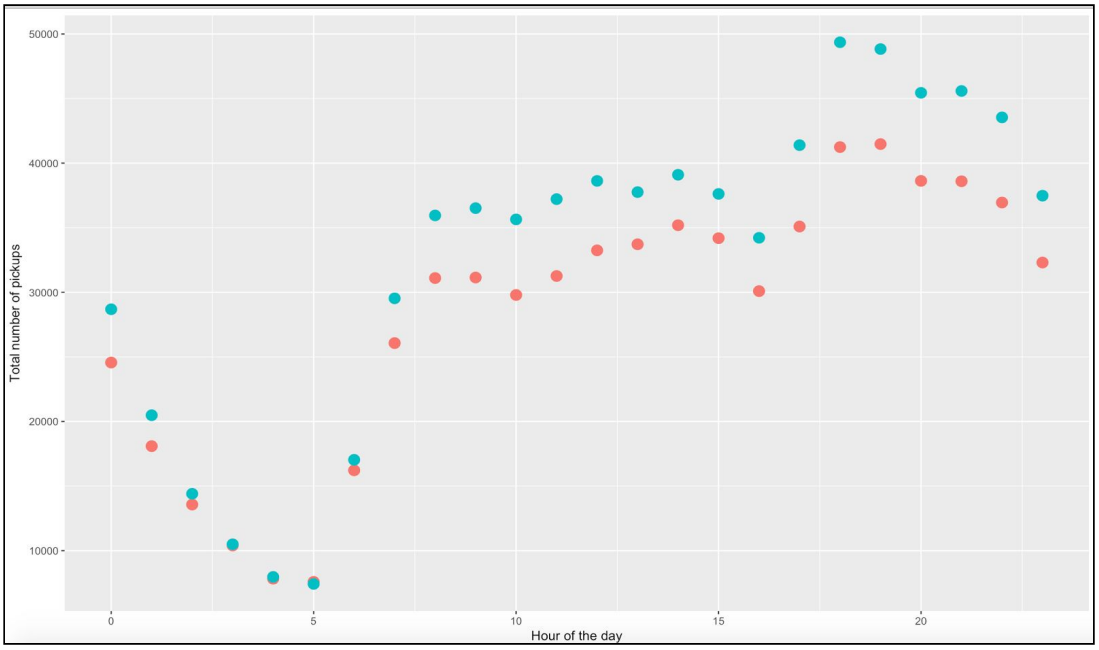
e. Weekday between two vendors



graph between the number of trips during day in a week by two vendors

From this graph, we can tell that the pattern of how frequent do people use taxi is not different although vendor two has more number of trips. People use the taxi the most on Friday and use taxi the least on Monday.

f. Hour of the day between two vendors

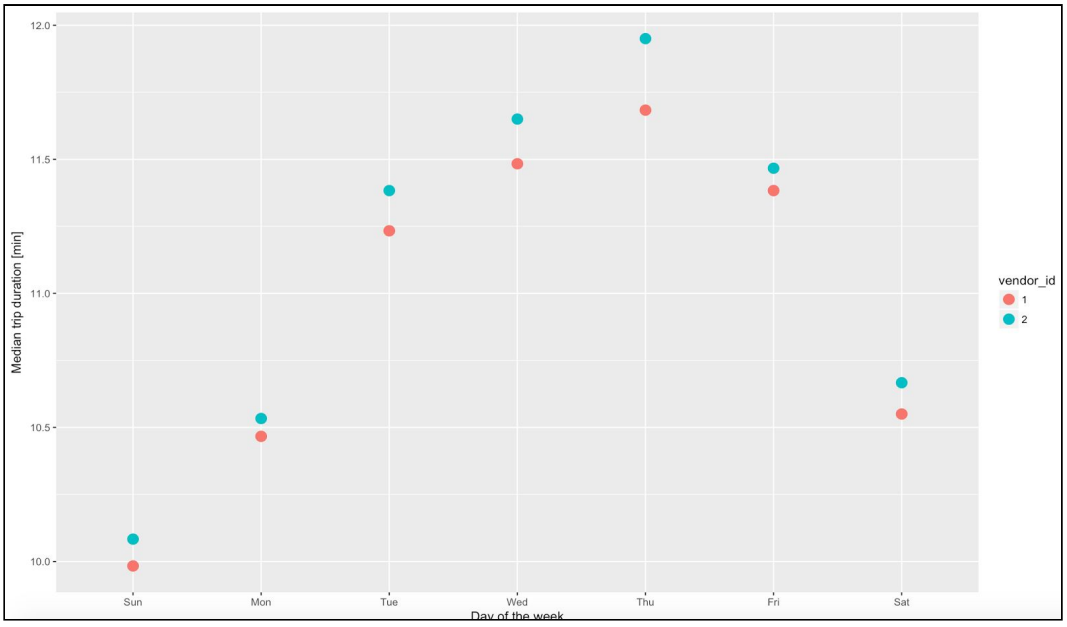


graph between the number of trips during hour of the day by two vendors

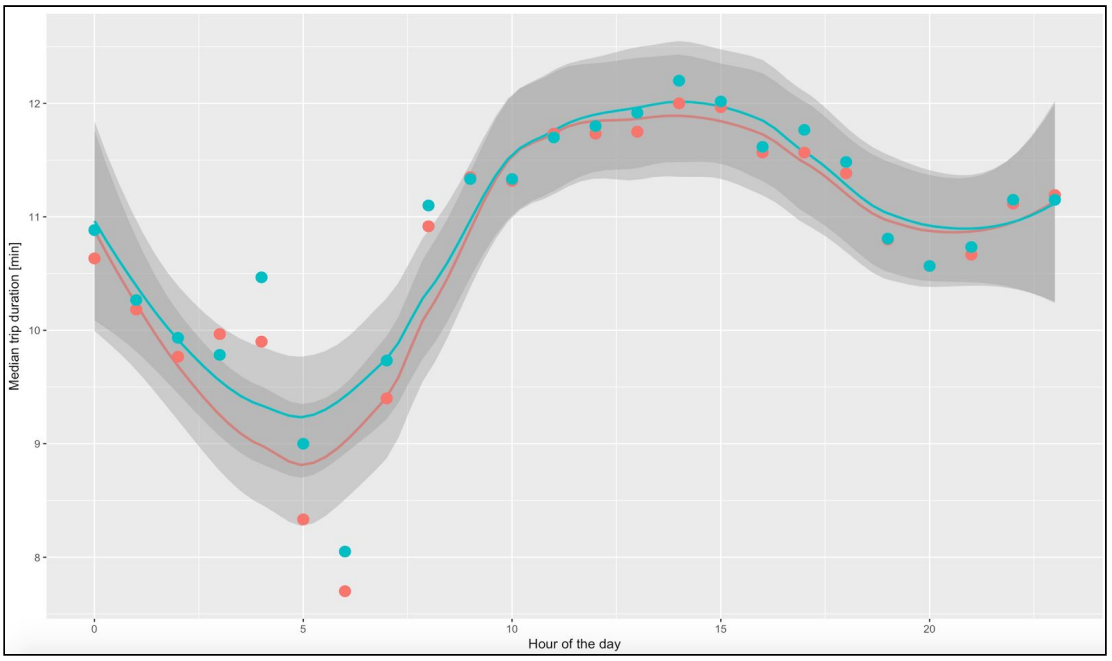
Same as the previous graph, we can tell that the pattern of how frequent do people use taxi is not different much. The graph will increase rapidly during 5am to 7.30am which is the rush hour and rise again around 4pm and stabilize until before midnight and decrease steadily until 5am.

1.2. Feature relations (relationship between the target attributes;trip_duration and each attribute)

a. pickup weekday and hour of the day



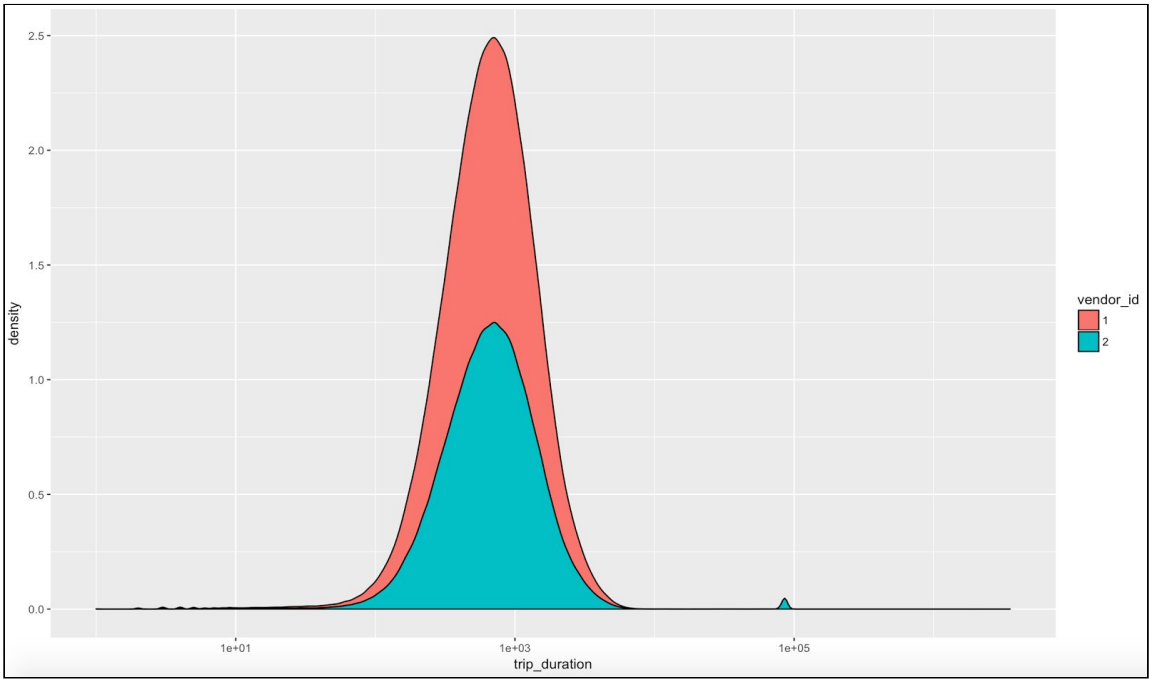
graph between the average trip duration and day of the week by two vendors



graph between the average trip duration and hour of the day by two vendors

From both graphs, there is a similar pattern for the average trip duration between both day of the week and hour of the day and also both vendor_id is also has the same pattern which mean vendor_id ,weekday, and hour should be important features to predict the trip duration and vendor_id.

b. trip_duration between two vendors id



graph show trip duration histogram between two vendors

```
> train %>%
+     group_by(vendor_id) %>%
+     summarise(mean_duration = mean(trip_duration),
+               median_duration = median(trip_duration))
# A tibble: 2 x 3
  vendor_id mean_duration median_duration
     <fctr>         <dbl>           <dbl>
1         1      845.4382             658
2         2     1058.6432             666
```

mean trip duration value and median trip duration value

As you can see from above figure, median value of trip duration is 658 meanwhile mean value of trip duration is about 854 which mean that our dataset contains many outliers so we have to clean the data first.

3. Feature engineering

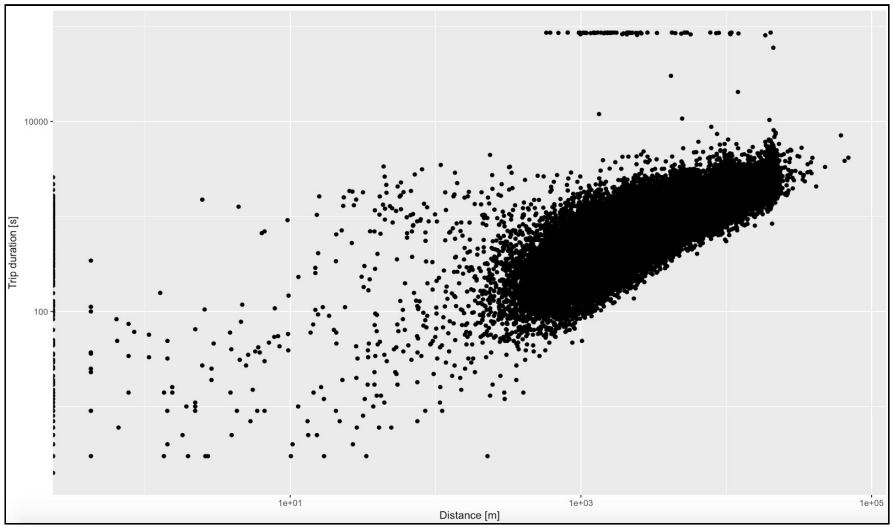
We have create new features to make our model work better Finally, we came up with the train set with 27 attributes which represents in the below table

No.	Attributes	Decription
1.	id<chr>	a unique identifier for each trip
2.	vendor_id<fctr>	a code indicating the provider associated with the trip record
3.	pickup_datetime<dttm>	date and time when the meter was engaged
4.	dropoff_datetime<dttm>	date and time when the meter was disengaged
5.	passenger_count<fctr>	the number of passengers in the vehicle (driver entered value)
6.	pickup_longtitude<dbl>	the longitude where the meter was engaged
7.	pickup_latitude<dbl>	the latitude where the meter was engaged
8.	dropoff_longitude<dbl>	the longitude where the meter was disengaged
9.	dropoff_latitude<dbl>	the latitude where the meter was disengaged
10.	store_and_fwd_flag<chr>	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
11.	trip_duration<int>	duration of the trip in seconds
12.	distance<dbl>	The shortest distance computed from the direct distance between the two points, and compare with our trip_durations.
13.	bearing<dbl>	It is the direction that tells about tstart out for instance in the direction of North-West or South-East.

14.	jfk_dist_pick<dbl>	A number of trips began or ended at either of the two NYC airports. The pick up point is JFK airport.
15.	jfk_dist_drop<dbl>	A number of trips began or ended at either of the two NYC airports. The destination drop off point is JFK airport.
16.	lg_dist_pick<dbl>	A number of trips began or ended at either of the two NYC airports. The pick up point is La Guardia airport
17.	lg_dist_drop<dbl>	A number of trips began or ended at either of the two NYC airports. The destination drop off point is La Guardia airport.
18.	date<date>	The date time data that will display on Month/Day/Year
19.	work<lgl>	Delays rate during work hours, its result will display on True or False.
20.	jfk_trip<lgl>	It indicating journeys to JFK airport or not created by define a pickup or dropoff distance of less than 2 km from the JFK airport
21.	lg_trip<lgl>	It indicating journeys to La Guardia airport or not created by define a pickup or dropoff distance of less than 2 km from the LGA airport
22.	pickup_weekday<int>	The number of pickup day in one week which covers a peak in raw trip_duration distribution.
23.	pickup_month<int>	The number of pickup month on the highest pickup day which covers with pickup_wday (which day in the week).
24.	pickup_hour<int>	The number of pickup hour on the highest pickup day which covers with pickup_wday (which day in the week).
25.	night_trip<lgl>	The feature checks which trip is on night trip or not. (Result will display on True or False)
26.	rush_hour<lgl>	The feature checks which trip is on rush hour or not. (Result will display on True or False)
27.	weekday<lgl>	The feature checks which trip is on weekend day or not. (Result will display on True or False)

```
> summary(train)
      id            vendor_id  pickup_datetime              dropoff_datetime             passenger_count
 Length:1458644     1:678342   Min.   :2016-01-01 00:00:17   Min.   :2016-01-01 00:03:31   1      :1033540
 Class :character   2:780302   1st Qu.:2016-02-17 16:46:04   1st Qu.:2016-02-17 17:05:32   2      : 210318
 Mode  :character              Median :2016-04-01 17:19:40   Median :2016-04-01 17:35:12   5      :  78088
                               Mean   :2016-04-01 10:10:24   Mean   :2016-04-01 10:26:24   3      :  59896
                               3rd Qu.:2016-05-15 03:56:08   3rd Qu.:2016-05-15 04:10:51   6      :  48333
                               Max.   :2016-06-30 23:59:39   Max.   :2016-07-01 23:02:03   4      :  28404
                                                                                           (Other):     65
 pickup_longitude  pickup_latitude dropoff_longitude dropoff_latitude store_and_fwd_flag trip_duration
 Min.   :-121.93   Min.   :34.36   Min.   :-121.93   Min.   :32.18    Length:1458644     Min.   :      1
 1st Qu.: -73.99   1st Qu.:40.74   1st Qu.: -73.99   1st Qu.:40.74    Class :character   1st Qu.:    397
 Median : -73.98   Median :40.75   Median : -73.98   Median :40.75    Mode  :character   Median :    662
 Mean   : -73.97   Mean   :40.75   Mean   : -73.97   Mean   :40.75                       Mean   :    959
 3rd Qu.: -73.97   3rd Qu.:40.77   3rd Qu.: -73.96   3rd Qu.:40.77                       3rd Qu.:   1075
 Max.   : -61.34   Max.   :51.88   Max.   : -61.34   Max.   :43.92                       Max.   :3526282

    distance          bearing         jfk_dist_pick     jfk_dist_drop      lg_dist_pick      lg_dist_drop
 Min.   :      0   Min.   :-180.000   Min.   :    310   Min.   :    432   Min.   :    197   Min.   :    147
 1st Qu.:   1233   1st Qu.:-126.873   1st Qu.:  20646   1st Qu.:  20587   1st Qu.:   8464   1st Qu.:   8347
 Median :   2096   Median :   8.167   Median :  21283   Median :  21272   Median :   9716   Median :   9603
 Mean   :   3445   Mean   : -16.454   Mean   :  20801   Mean   :  20967   Mean   :   9734   Mean   :   9770
 3rd Qu.:   3880   3rd Qu.:  53.420   3rd Qu.:  21962   3rd Qu.:  22014   3rd Qu.:  11116   3rd Qu.:  11093
 Max.   :1242299   Max.   : 179.996   Max.   :4128727   Max.   :4128726   Max.   :4118057   Max.   :4118057

     speed               date                month          wday           hour          work
 Min.   :   0.000   Min.   :2016-01-01   Mar    :256189   Mon:187418   Min.   : 0.00   Mode :logical
 1st Qu.:   9.131   1st Qu.:2016-02-17   Apr    :251645   Wed:210136   1st Qu.: 9.00   FALSE:1117165
 Median :  12.806   Median :2016-04-01   May    :248487   Fri:223533   Median :14.00   TRUE :341479
 Mean   :  14.439   Mean   :2016-03-31   Feb    :238300   Sat:220868   Mean   :13.61
 3rd Qu.:  17.865   3rd Qu.:2016-05-15   Jun    :234316   Sun:195366   3rd Qu.:19.00
 Max.   :9285.227   Max.   :2016-06-30   Jan    :229707   Tue:202749   Max.   :23.00
                                         (Other):     0   Thu:218574
  jfk_trip         lg_trip             dist          blizzard
 Mode :logical   Mode :logical   Min.   :      0   Mode :logical
 FALSE:1416450   FALSE:1403827   1st Qu.:   1233   FALSE:1407411
 TRUE :42194     TRUE :54817     Median :   2096   TRUE :51233
                                 Mean   :   3445
                                 3rd Qu.:   3880
                                 Max.   :1242299
```

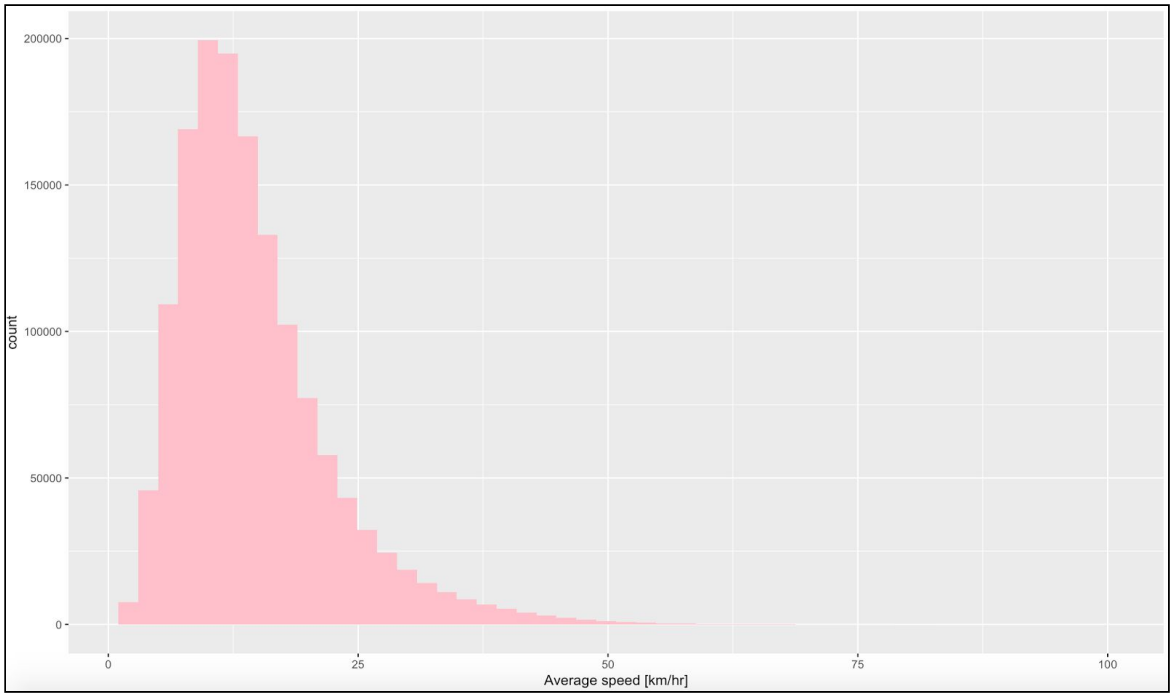
a. Direct distance of the trip



Relationship between distance and trip duration graph

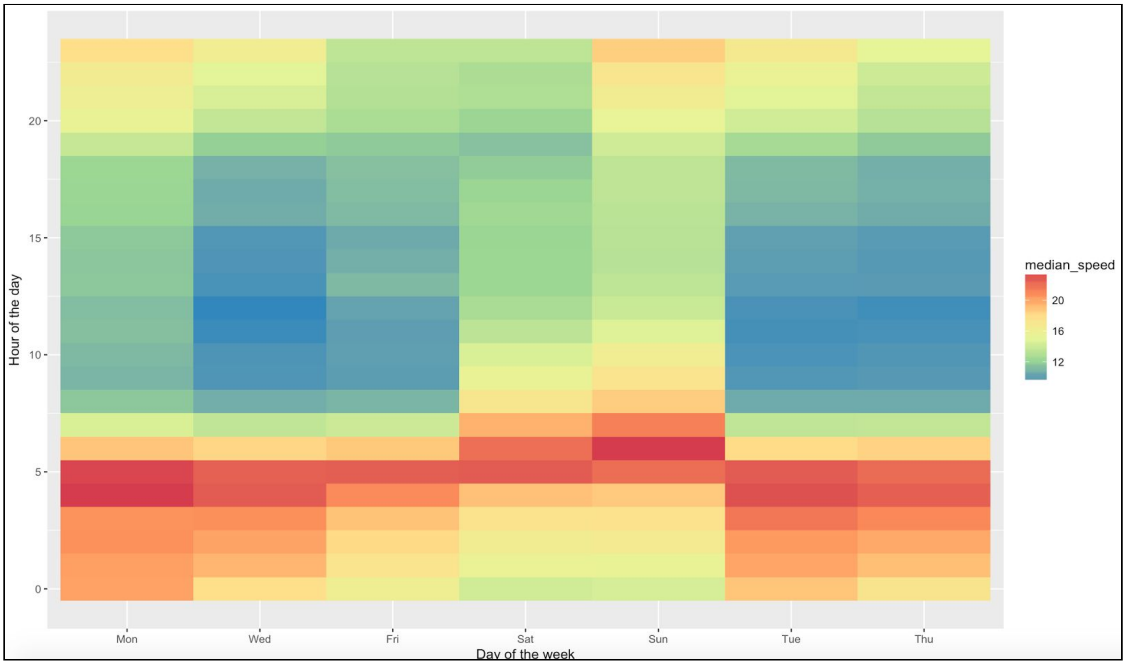
From this graph, we noticed that the more distance increases, the more trip duration increases. Value between trip duration and distance increase as a exponential graph. Most of the trips have the trip duration between 100 seconds to 1000 seconds.

b. Travel speed



relationship between average travel speed and the number of trips graph

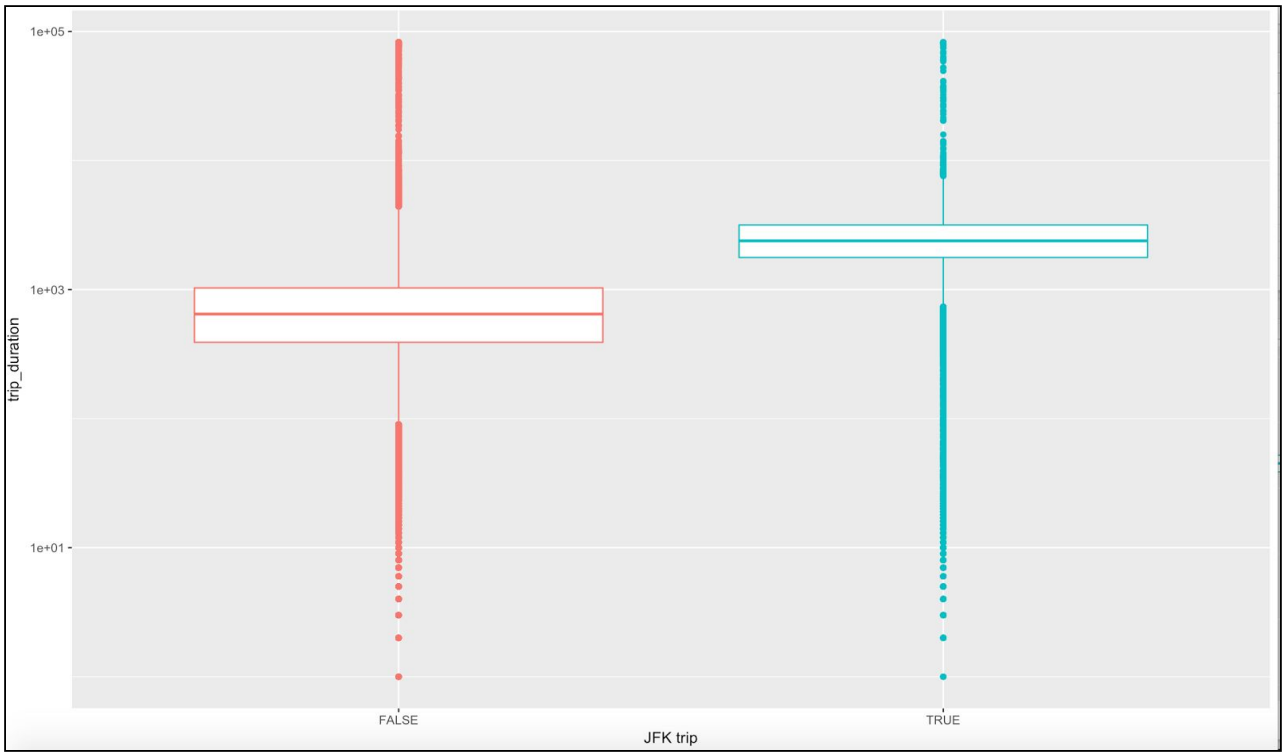
From this graph we filtered the speed between 3km/hr and 100km/hr and found that the speed is right-skewness. The average speed which has the highest number of trips is about 13 km/hr.



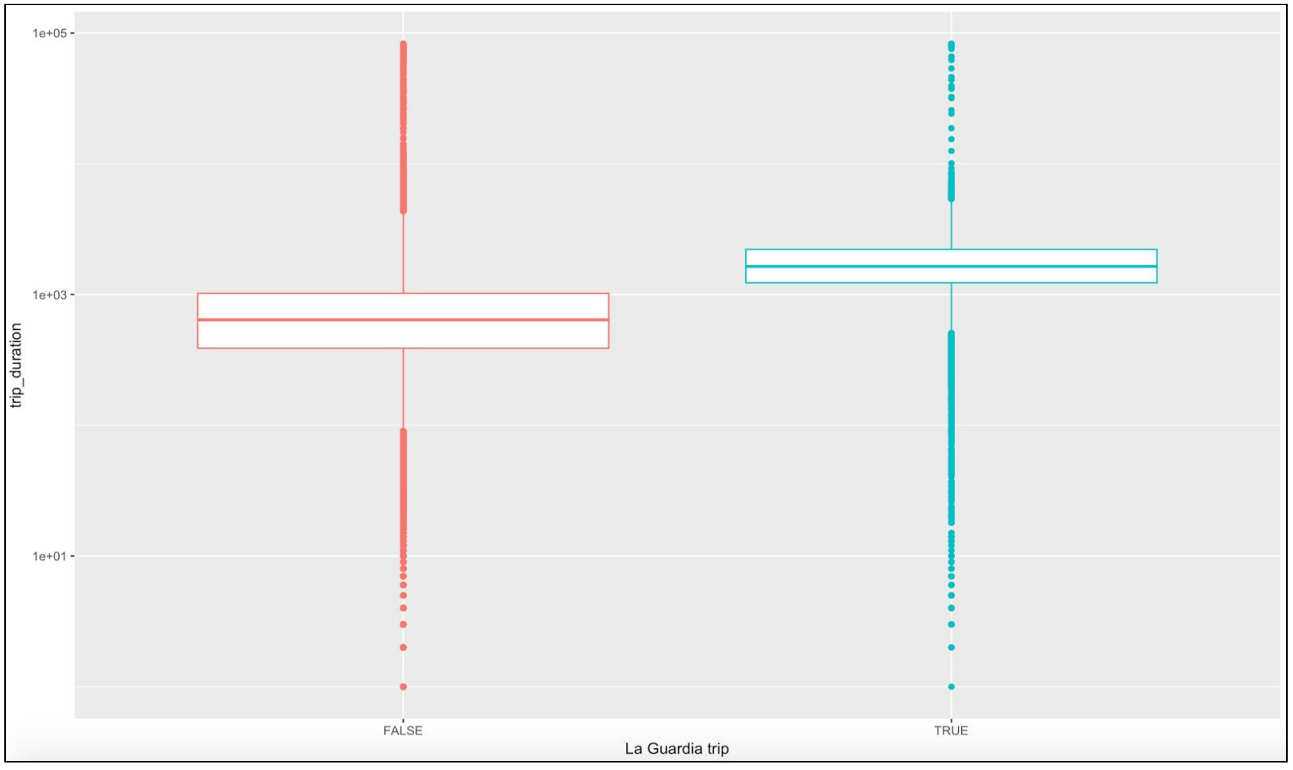
relationship between average travel speed during day of the week and hour of the day

From this graph, we found that the average travel speed is tend to be faster during the night time until the early morning time (0am-6am) while on the weekend the overall average speed is slower than the weekday.

c. Airport distance



trip duration boxplot between JFK trip and not JFK trip



trip duration boxplot between LGA trip and not LGA trip

From the first map which displayed the pickup and dropoff, we noticed that beside the Manhattan area, there are some trips that pickup or dropoff at JFK airport and LGA airport. Our hypothesis is the longer trip duration is the trips to the airports and from this graph we can see that our hypothesis is true since you can see that Q1 of both airport trips is higher than the others so both jfk_trip and lg_trip should be one of our feature for prediction model.

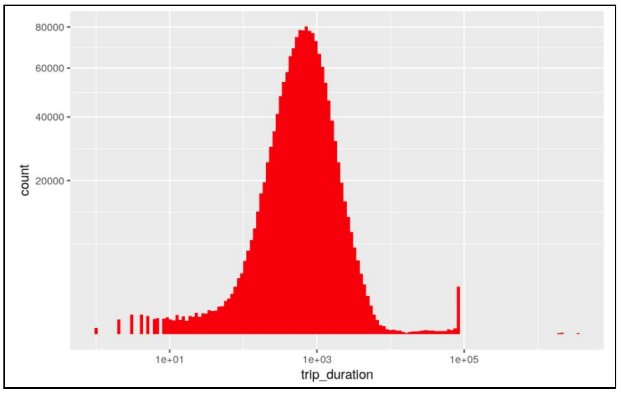
4. Data Cleaning

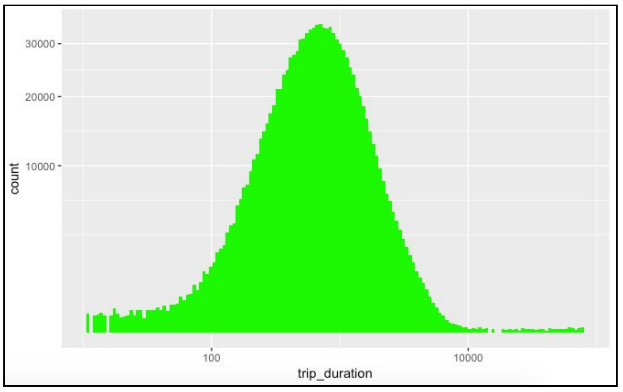
```
train <- train %>%
  filter(trip_duration < 22*3600,
         dist > 0 | (near(dist, 0) & trip_duration < 60),
         jfk_dist_pick < 3e5 & jfk_dist_drop < 3e5,
         trip_duration > 10,
         speed < 100)
```

final code that we used to clean data

- Formatting Criteria
  - trip duration longer than 22 hours and less than 10 seconds
  - trip distance equal zero which trip duration is more than a minutes
  - long distance location which pickup or dropoff locations more than 300 km away from NYC (JFK airport)
  - travel speed more than 100 km/hr

After cleaning the data we have 1048575 columns from 1458644 columns



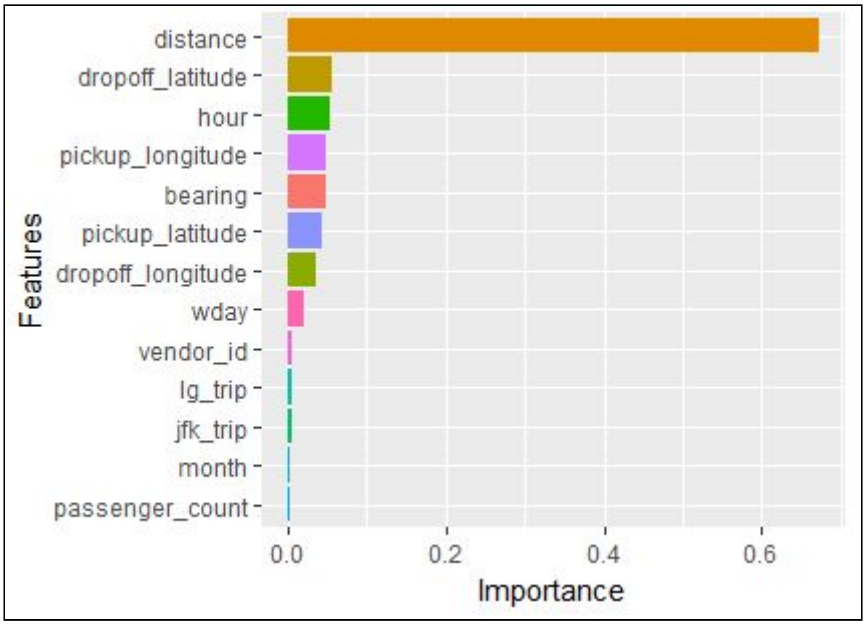


before data cleaning trip duration histogram *after data cleaning duration histogram*

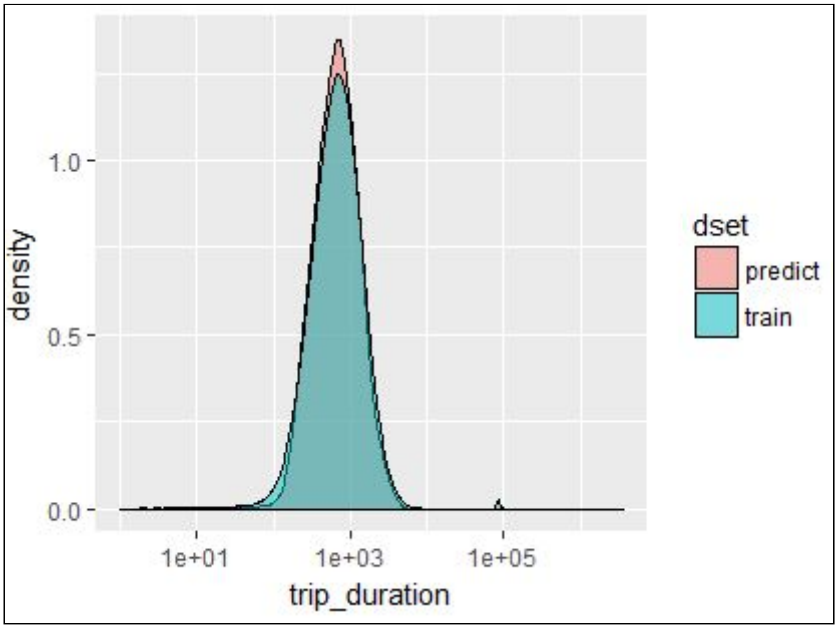
From the graph, after data cleaning duration histogram, we have clean the suspicious extreme trips and also the strange delta-shaped peak of trip duration before 100000 seconds is deleted.

5. Training model

- **Model 1: XGBoost parameters and fitting result**



Feature importance ranking from xgboost



relation between predict and train value graph

```
[251]   train-rmse:0.386086     valid-rmse:0.385573
[256]   train-rmse:0.385781     valid-rmse:0.385295
[261]   train-rmse:0.385537     valid-rmse:0.385029
[266]   train-rmse:0.385273     valid-rmse:0.384692
[271]   train-rmse:0.384980     valid-rmse:0.384401
[276]   train-rmse:0.384662     valid-rmse:0.384020
[281]   train-rmse:0.384428     valid-rmse:0.383766
[286]   train-rmse:0.384052     valid-rmse:0.383398
[291]   train-rmse:0.383666     valid-rmse:0.383092
[296]   train-rmse:0.383397     valid-rmse:0.382835
[300]   train-rmse:0.383108     valid-rmse:0.382591
```

rmlse value from XGBoost model (train set)

```
[990]   train-rmse:0.375571+0.000594    test-rmse:0.397627+0.002093
[991]   train-rmse:0.375555+0.000596    test-rmse:0.397627+0.002091
[992]   train-rmse:0.375535+0.000601    test-rmse:0.397634+0.002086
[993]   train-rmse:0.375519+0.000594    test-rmse:0.397629+0.002086
[994]   train-rmse:0.375502+0.000591    test-rmse:0.397638+0.002082
[995]   train-rmse:0.375483+0.000589    test-rmse:0.397636+0.002082
[996]   train-rmse:0.375462+0.000594    test-rmse:0.397623+0.002070
[997]   train-rmse:0.375445+0.000595    test-rmse:0.397617+0.002072
[998]   train-rmse:0.375422+0.000592    test-rmse:0.397604+0.002074
[999]   train-rmse:0.375401+0.000594    test-rmse:0.397602+0.002071
[1000]  train-rmse:0.375382+0.000591    test-rmse:0.397606+0.002079
```

rmlse value from XGBoost model (test set)

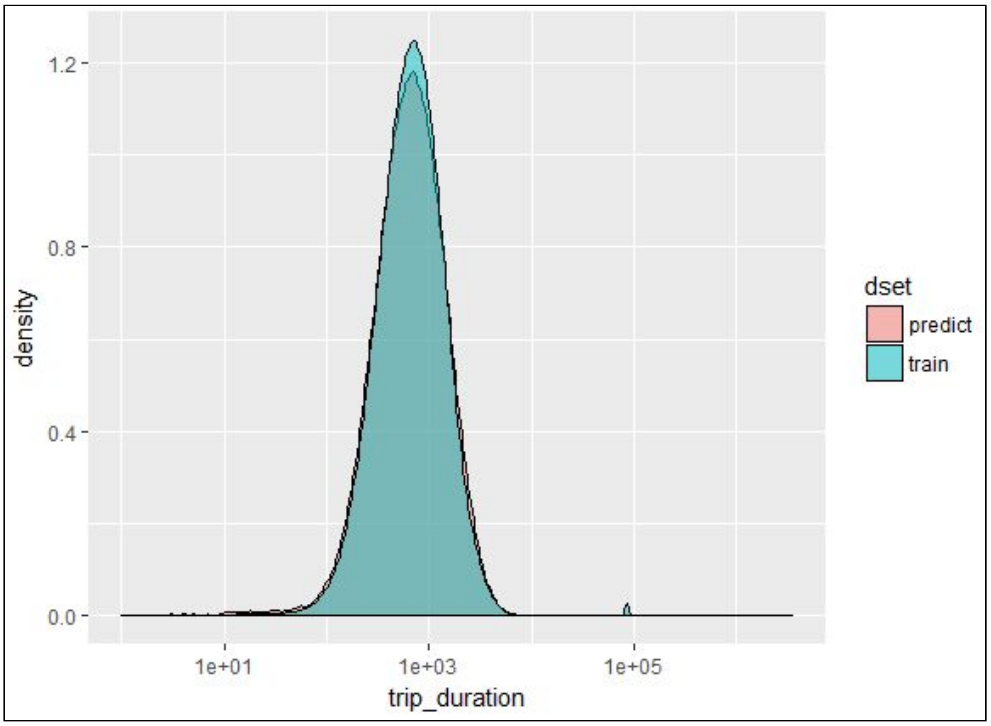
Submission and Description	Private Score	Public Score
submit2.csv a few seconds ago by May Junejuly xgboost with eta 0.1	0.39993	0.40219

kaggle score - rmsle (test set)

- **Model 2: Decision tree**

```
Feature ranking Decision tree:
1. feature distance (0.679895)
2. feature bearing (0.063712)
3. feature dropoff_latitude (0.051363)
4. feature pickup_hour (0.051241)
5. feature dropoff_longitude (0.034659)
6. feature pickup_longitude (0.033078)
7. feature pickup_latitude (0.030436)
8. feature pickup_weekday (0.016866)
9. feature pickup_month (0.013026)
10. feature weekday (0.008122)
11. feature passenger_count (0.006291)
12. feature rush_hour (0.004465)
13. feature vendor_id (0.003336)
14. feature work (0.001758)
15. feature night_trip (0.001347)
16. feature lg_trip (0.000317)
17. feature jfk_trip (0.000088)
```

Feature importance from decision tree



relation between predict and train value graph

```
With model: DecisionTreeRegressor(criterion='mse', max_depth=None, max_features=None,
           max_leaf_nodes=None, min_impurity_decrease=0.0,
           min_impurity_split=None, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.0,
           presort=False, random_state=None, splitter='best')
Train RMSLE: 0.000276448745885
Val. RMSLE: 0.482543207943
```

rmlse value from decision tree model - rmsle (train set)

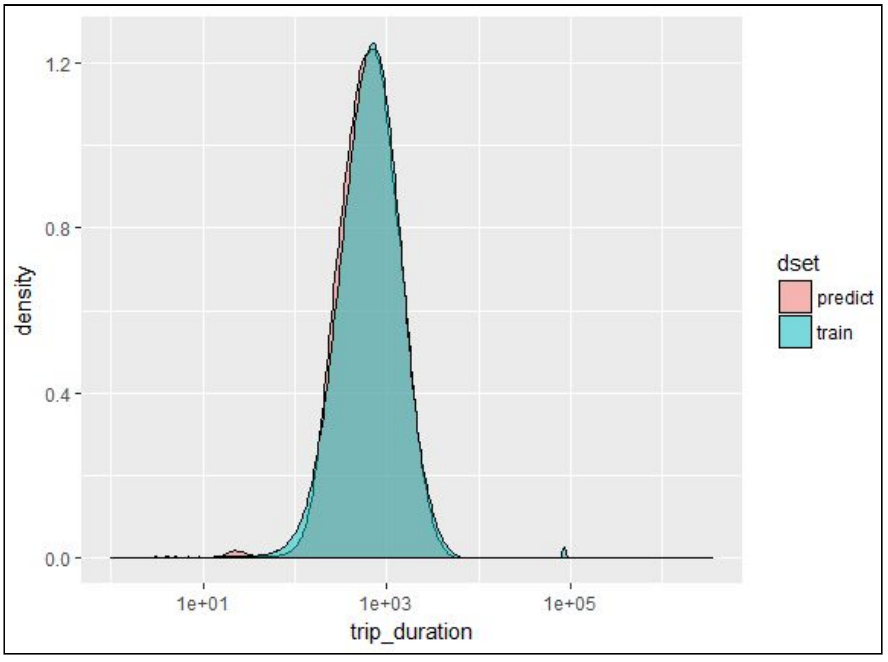
submission-dt.csv a day ago by May Junejuly using decision tree	0.63833	0.63711

kaggle score from decision tree model - rmsle (test set)

- **Model 3: Random forrest**

```
Feature ranking Random Forest:
1. feature distance (0.680009)
2. feature bearing (0.064019)
3. feature dropoff_latitude (0.051785)
4. feature pickup_hour (0.049699)
5. feature dropoff_longitude (0.034452)
6. feature pickup_longitude (0.033023)
7. feature pickup_latitude (0.030092)
8. feature pickup_weekday (0.017583)
9. feature pickup_month (0.013074)
10. feature weekday (0.007332)
11. feature passenger_count (0.006287)
12. feature rush_hour (0.005450)
13. feature vendor_id (0.003300)
14. feature work (0.001923)
15. feature night_trip (0.001576)
16. feature lg_trip (0.000282)
17. feature jfk_trip (0.000114)
```

Feature importance from decision tree



relation between predict and train value graph

```
With model: RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_decrease=0.0, min_impurity_split=None,
           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, n_estimators=50, n_jobs=1,
           oob_score=False, random_state=None, verbose=0, warm_start=False)
Train RMSLE: 0.128683034987
Val. RMSLE: 0.340619719201
```

rmlse value from random forrest model - rmsle (train set)

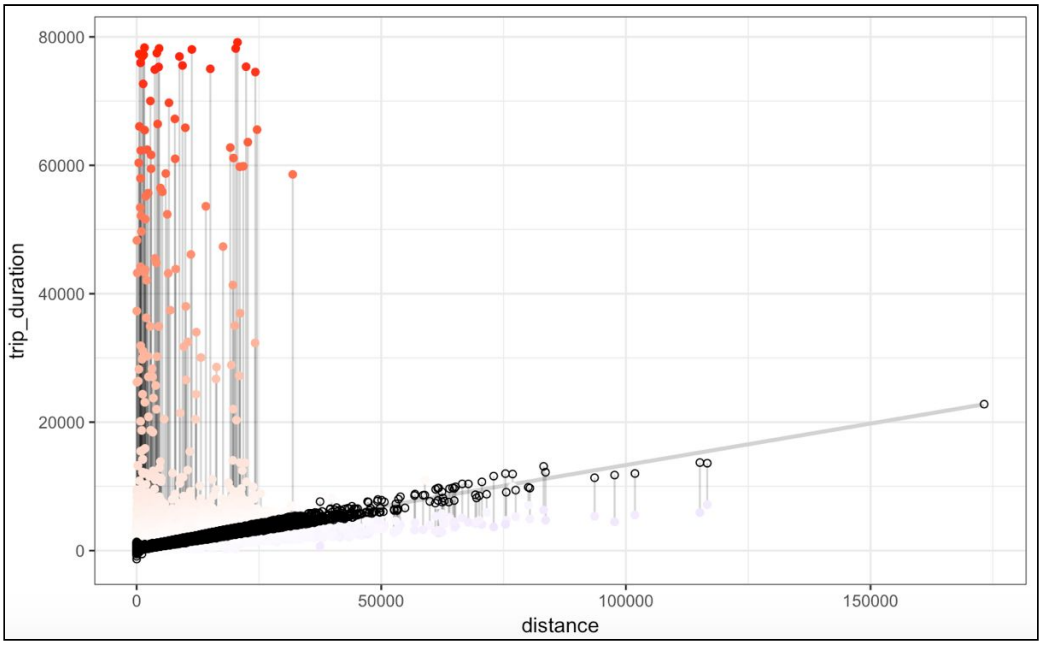
kaggle score from random forrest model - rmsle (test set)

- **Model 4: Linear regression**

```
> fit <- lm(trip_duration ~ distance+pickup_latitude+dropoff_latitude+pickup_hour, data = train)
> summary(fit)

Call:
lm(formula = trip_duration ~ distance + pickup_latitude + dropoff_latitude +
    pickup_hour, data = train)

Residuals:
   Min     1Q Median     3Q    Max
 -8356   -242    -88    151  77754

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.338e+04  1.090e+03   30.62   <2e-16 ***
distance        1.291e-01  1.719e-04  750.93   <2e-16 ***
pickup_latitude  7.700e+02  2.646e+01   29.11   <2e-16 ***
dropoff_latitude -1.581e+03  2.202e+01  -71.77   <2e-16 ***
pickup_hour      4.053e+00  1.008e-01   40.20   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 660.7 on 1048570 degrees of freedom
Multiple R-squared:  0.3772,    Adjusted R-squared:  0.3772
F-statistic: 1.588e+05 on 4 and 1048570 DF,  p-value: < 2.2e-16
```

So our final linear equation for this model is

Trip_duration = 3.338e+04+(1.291e-01*distance)+(7.700e+02*pickup_latitude)
-(1.581e+03*dropoff_latitude)+(4.053+00*pickup_hour)



graph displayed relation between predicted value and actual value

kaggle score from linear regression - rmsle (test set)

Conclusion

	XGBoost	Decision tree	Random Forest	Linear Regression
Kaggle Score (RMLSE)	0.39993	0.63833	0.53662	0.59504

From the result, we concluded that XGBoost has the best performance to predict the trip duration.