University at Buffalo
The State University of New York

Department of Industrial and Systems Engineering
School of Engineering and Applied Sciences

Exploratory Data Analysis of NYC Taxi Trip Data

Programming and Database Fundamentals for Data Scientists

Submitted by

GROUP 5

Sai Kiran Putta

Likhith Kumar Miryala

HariKrishna Rangineeni

Project Guide

Dr. Varun Chandola

Date

12/15/2017

ACKNOWLEDGEMENT

We would like to thank to the course instructor Dr. Varun Chandola for letting us undertake the Project, reviewing our work throughout the process of conducting the project and for providing knowledge and material necessary for the project.

We would also like to thank SreeLekha Guggilam, the teaching assistant for taking the time to meet us whenever required and helping in resolving the queries we had throughout the course.

Sincerely,

Sai Kiran Putta
Likhith Kumar M
Harikrishna Rangineeni

Abstract

Dataset is New York City Taxi Data. New York city taxi data set is a breakdown of every taxi trip in NYC by location and trip duration. Each record represents a trip duration in NYC by pickup-datetime, dropoff-datetime, pickup-latitude, drop off latitude, pickup longitude and drop off longitude. The database includes the date and time, location (latitudes, longitudes), passenger count and different vendors and related factors for all 1458644 taxi trips in New York city during January to June 2016.

The objective of the project is to identify the estimated trip duration to travel from one place to another place in New York and plotting the various contributing factors versus trip duration. Based on individual trip attributes, we group the individual points into high number of clusters. Using this we calculate the average time taken to travel from one cluster to another and interpret the obtained results.

Matplotlib is used for plotting, Pandas, SQL are used for data querying and numpy is used for analysis. These reports/visualizations can be used by the public to see how long does it take to travel from one cluster to another cluster. These reports can be used by the cab companies to improve their estimated destination time.

Data Set Details:

The data set has 14 million rows which represent 14 million taxi trips. There are 13 columns which includes the date and time (pickup date time, dropoff date time), location (pickup and dropoff latitudes and longitude coordinates), vendor-id, passenger count for all the taxi trips. The data is collected from 2016 january to june.

Data fields:

id - a unique identifier for each trip

vendor_id - a code indicating the provider associated with the trip record

pickup_datetime - date and time when the meter was engaged

dropoff_datetime - date and time when the meter was disengaged

passenger_count - the number of passengers in the vehicle (driver entered value)

pickup_longitude - the longitude where the meter was engaged

pickup_latitude - the latitude where the meter was engaged

dropoff_longitude - the longitude where the meter was disengaged

dropoff_latitude - the latitude where the meter was disengaged

store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memorybefore sending to the vendor because the vehicle did not have a connection to the server - Y=storeand forward; N=not a store and forward trip

trip_duration - duration of the trip in seconds

#We shall be working on all the data fields except store_and_fwd_flag. We feel this feature is redundant.

Introduction:

Cities like New York are flooded with thousands of vehicles on streets every day. The major proportion of these vehicles are taxis which are the widely recognized icons of the city.

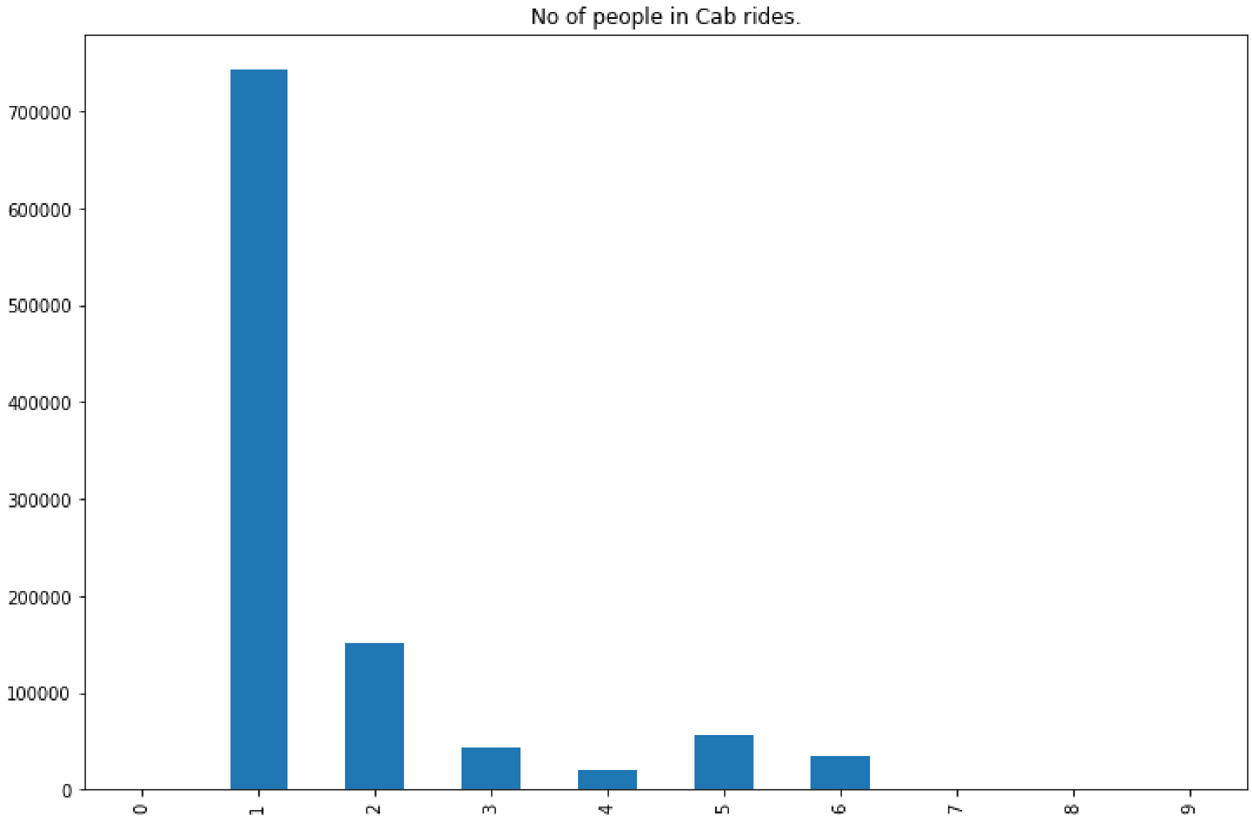
Economic and social benefits of a region are guaranteed by efficient, safe and well-maintained transportation system. In the last few years, there is an enormous change in the transportation system of NYC due to increased congestion. Eventually there are significant delays in NYC, particularly in the most happening places like Manhattan. Due to this loss of time, there are often imposed costs on shippers, manufacturers and businesses, which in turn are imposed on consumer.

With an intention to give an estimated trip duration from one location to another in a taxi, the project proposes a method to show when a person can start to reach the destination at correct time. The dataset for the study contains trip duration and other related factors. In addition, they encode information about movement: a trip is associated with pickup and drop-off locations and times.

Exploratory data analysis

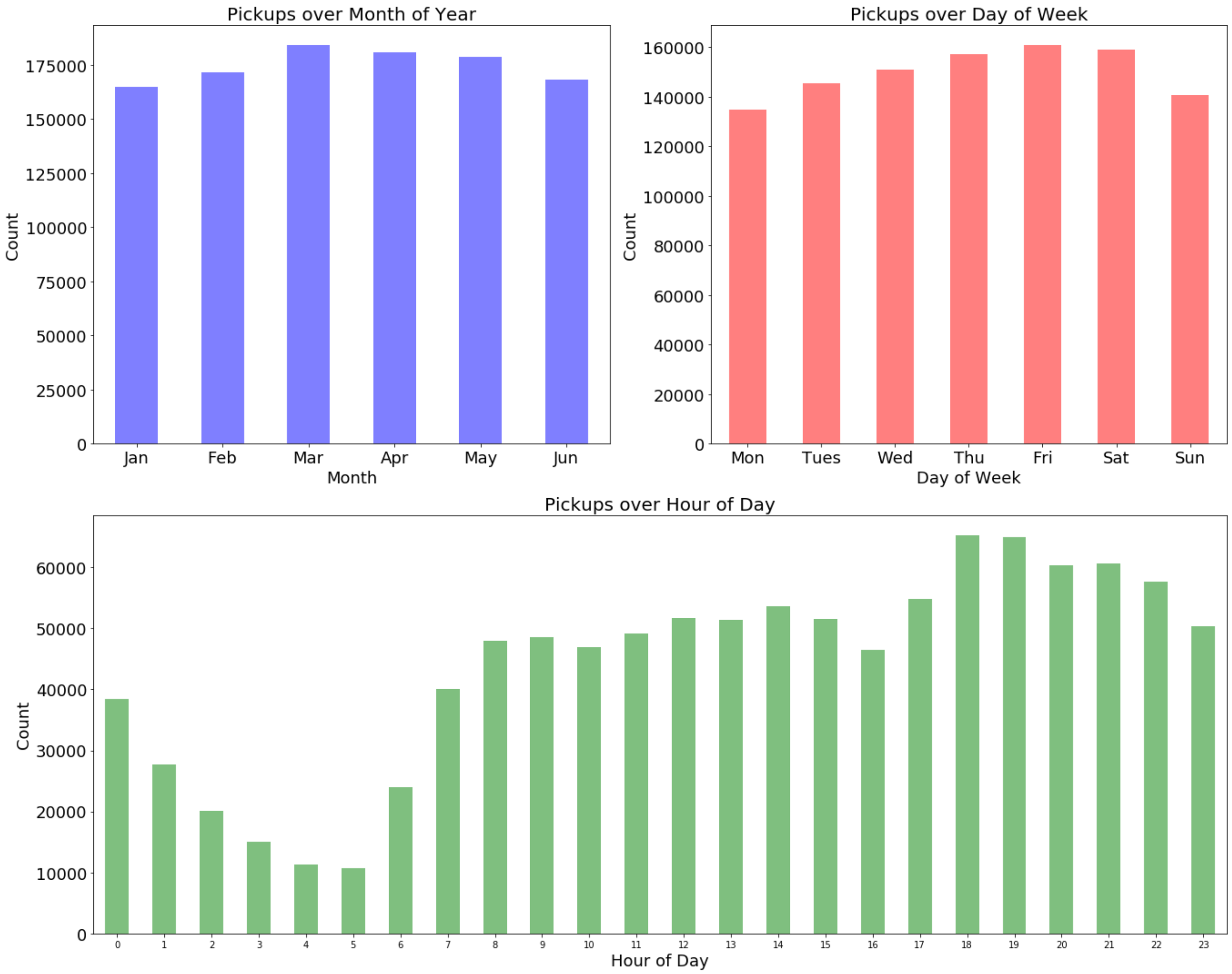
This is an Exploratory Data Analysis for the NYC taxi trip duration dataset. The data comes in the shape of 10 million training observations, where each row contains one taxi trip.

Distribution of the number of passengers across the variables in the dataset.



Inference: From the above plot, we can observe that passengers who are travelling alone are travelling in cabs more than going in groups. Nearly 70,000 cabs have been booked by passengers who are travelling alone. Also we can observe 5 passengers as a group are travelling more compared to 3 and 4 passengers.

Let’s take a close look at the pickups. The below plots show the distribution of taxi demand over pickup time in months, days of week, hours of the day.

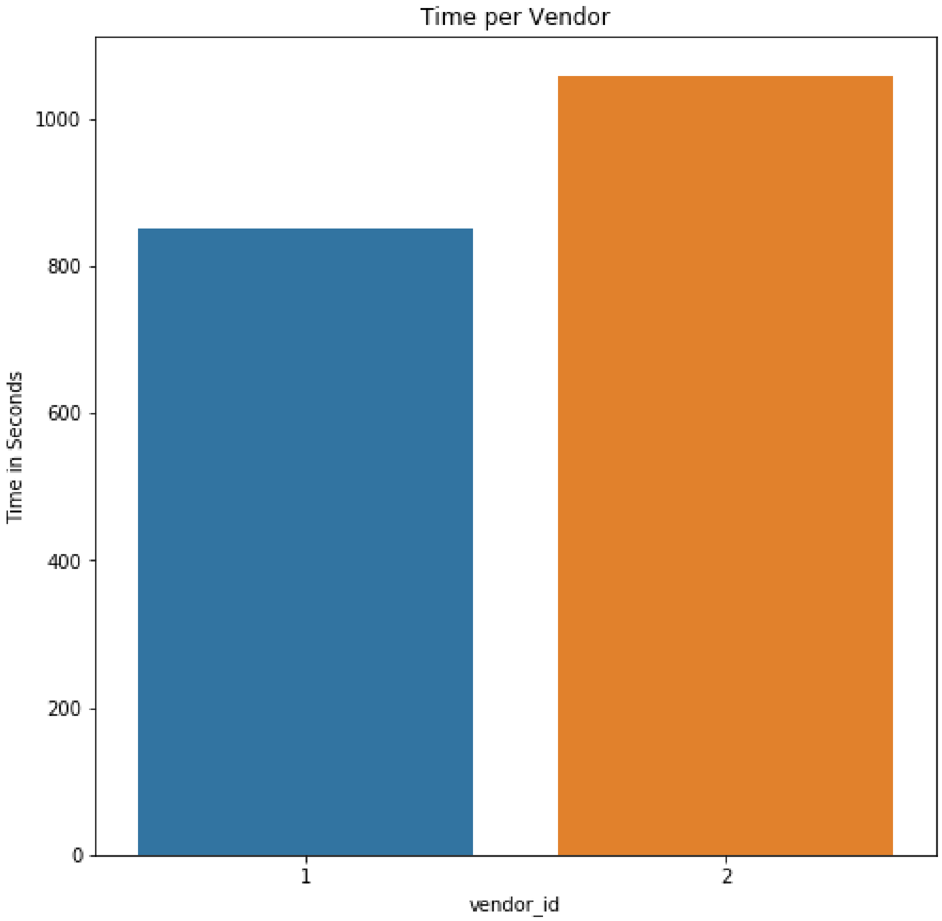


Inference: For pickups over month of year, the number of pickups for every month is almost the same. Expected because there is no major festive season around these months.

Let's move on to the number of pickups on different days of the week. Everything looks good here, except that number of pickups in weekdays is same as in the weekends. Also it can be observed that pickups are slightly higher on Fridays compared to other days of the week.

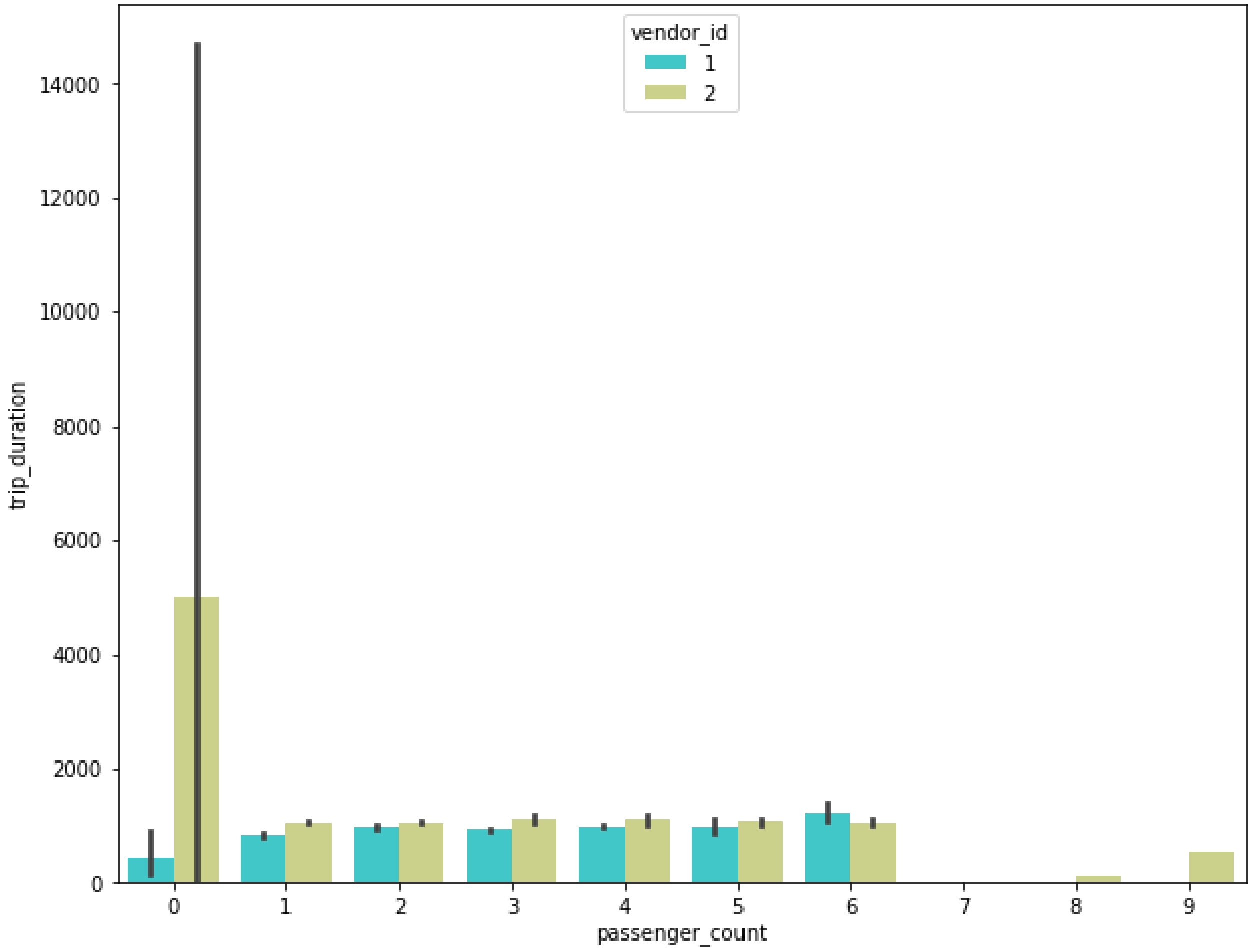
For hour of the day, everything looks good here too. As expected, the number of pickups are less during the morning hours. The highest number of pickups are around 6pm and 7pm in the evening which makes sense as many people are on their way to home from office. After mid night less number of trips are taken.

Vendor id takes on only two values i.e. 1 and 2 (This represents data from two different taxi companies). The number of observations in the dataset from each of the two companies i.e. 1 and 2, seems to be comparable in terms of trip duration.



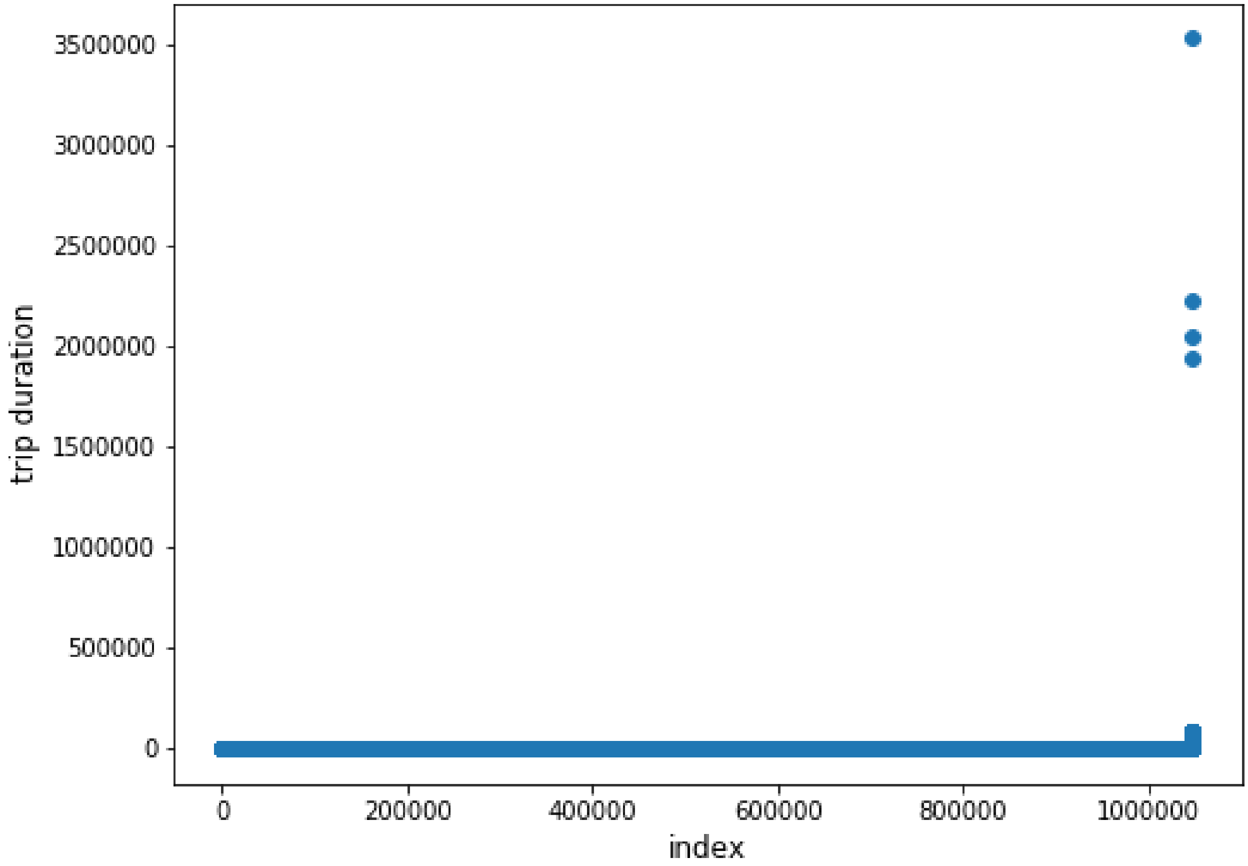
Inference: In the above graph, vendor-id was plotted against mean trip duration (in seconds) for each company or vendor-id. We can see that vendor 2 takes more time compared to vendor 1.

Let’s take a look at the average trip duration per passenger count according to vendor-id



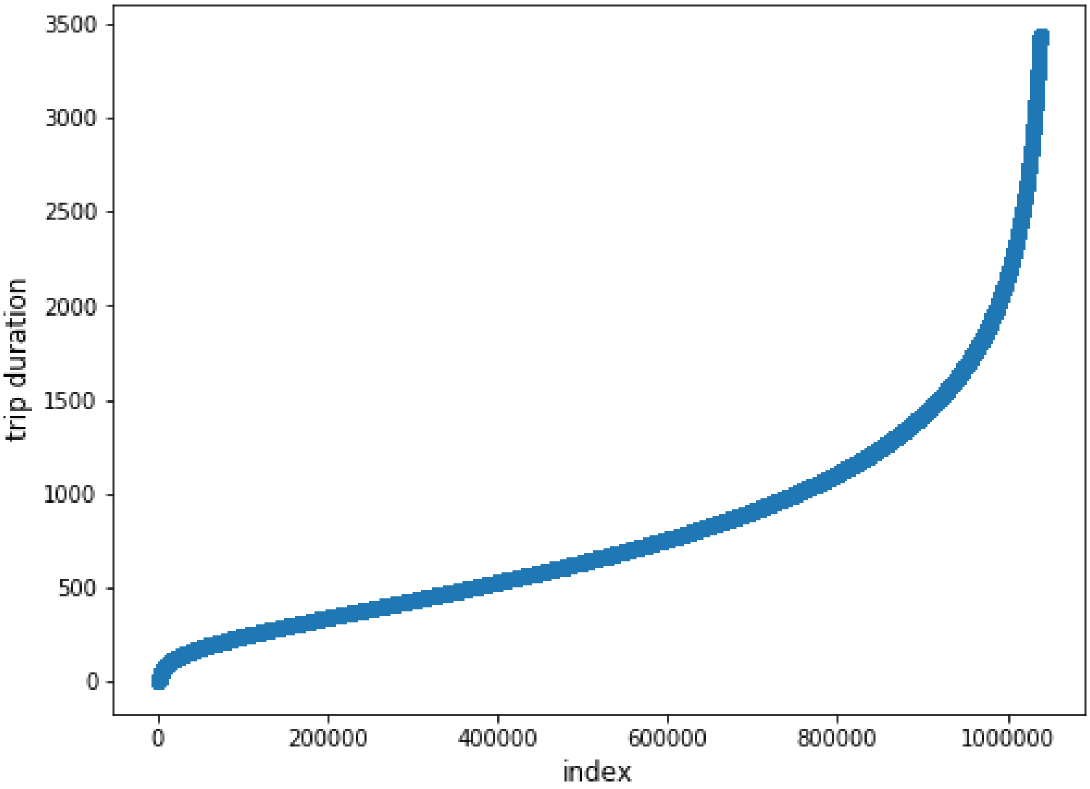
Inference: From the above plot, we can observe that more passengers are travelling in vendor-id 2 except for passenger group of six people. And passenger groups who have more than six people are less and no one is preferring vendor-id 1. And vendor-id 0 means can driver is travelling alone and it infers that most of the times cab driver is travelling alone without a passenger.

As our main goal is to find out the trip duration, let’s take a look at the trip duration column.



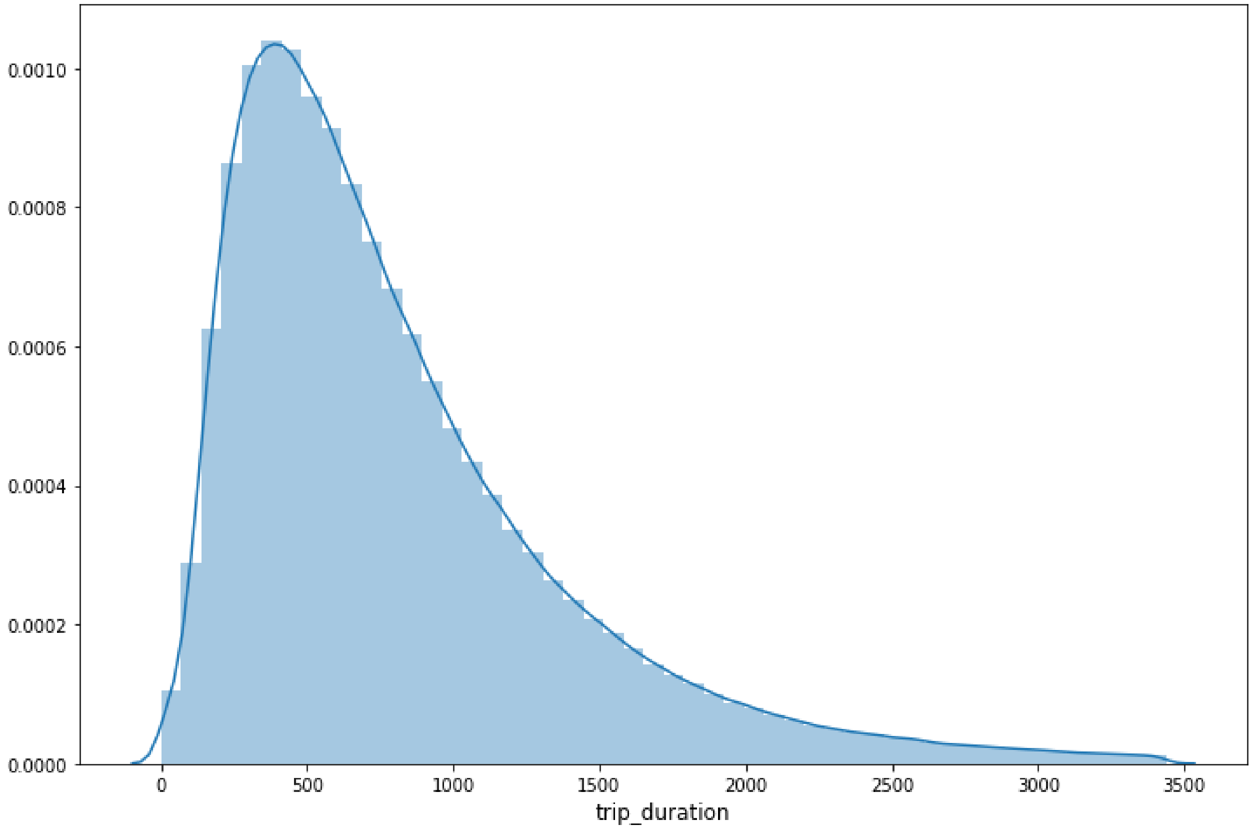
Inference: It can be observed that few trip durations in dataset are very high (20 days or more), we considered them as outliers and removed them before plotting.

Plotting the trip duration column after removing outliers:



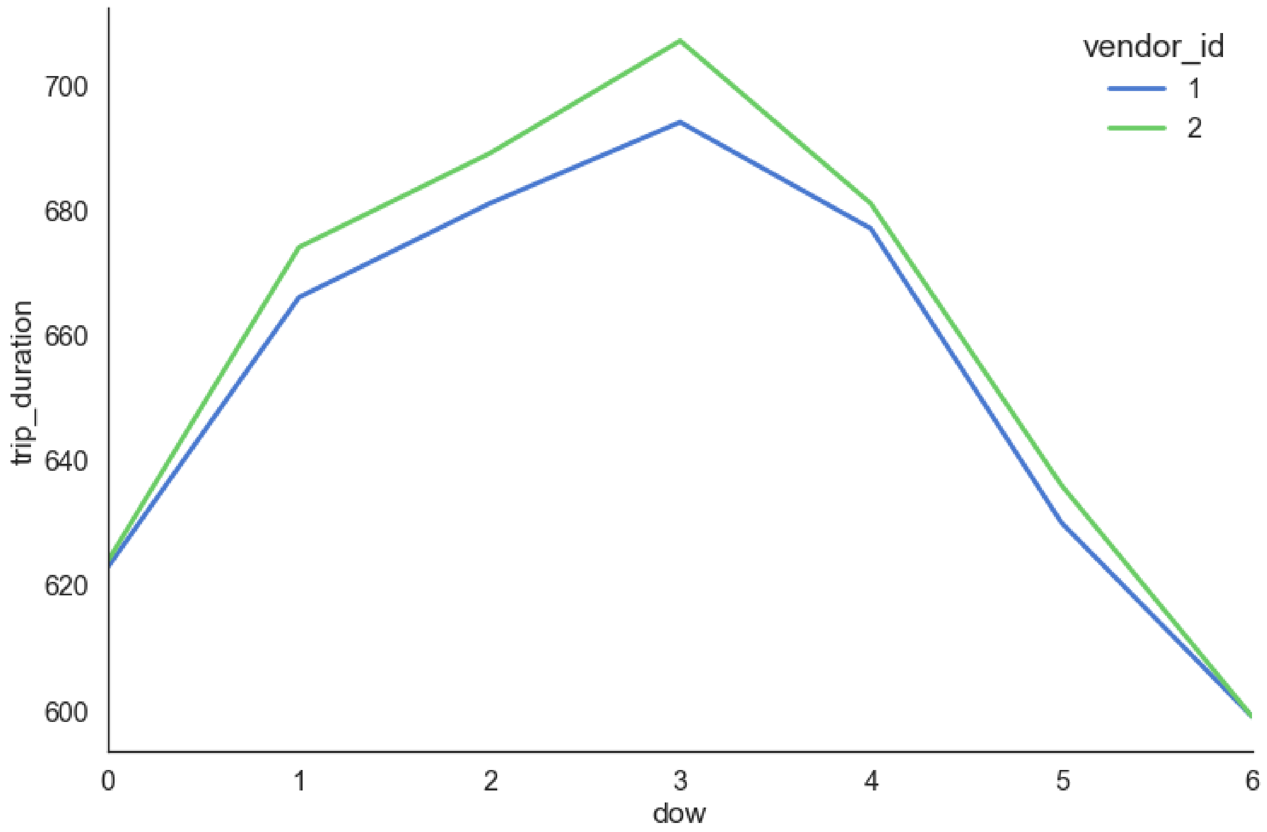
Inference: It looks better after removing outliers. It is interesting to see that a lot of trips, have a trip duration nearing 23 hours.

Trip duration density plot:



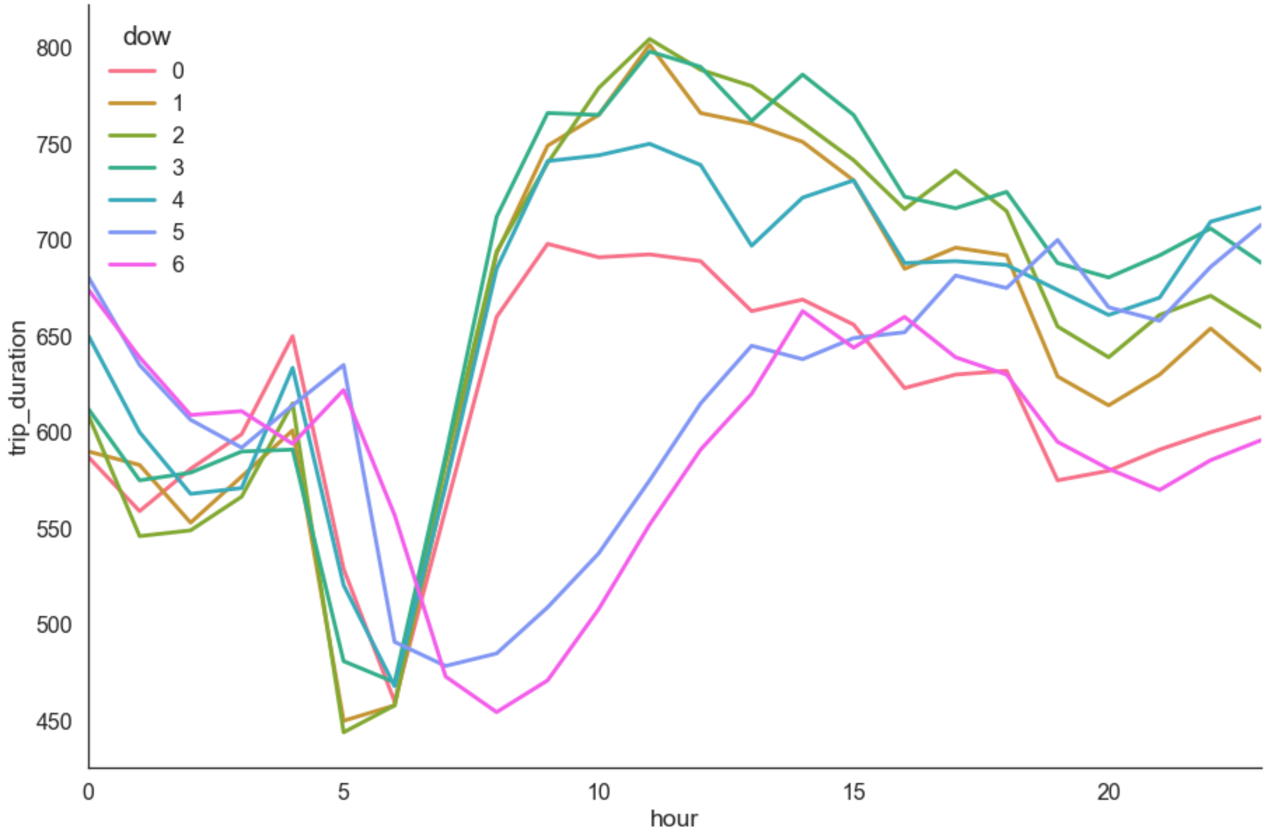
Inference: We plotted a density curve for the trip duration. It has Gaussian Distribution skewed to the right with mean 807.46 and standard deviation 575.42. The distribution is a right tailed. So, using mean would be inappropriate. Hence using median for the analysis.

Trip duration over days of week:



Inference: It's clear that the vendor 1 is taking more time than vendor 2 on all the days of the week, we can also subset data frame based on the month and that will also give us the same results.

Trip duration over days of week for every particular hour:



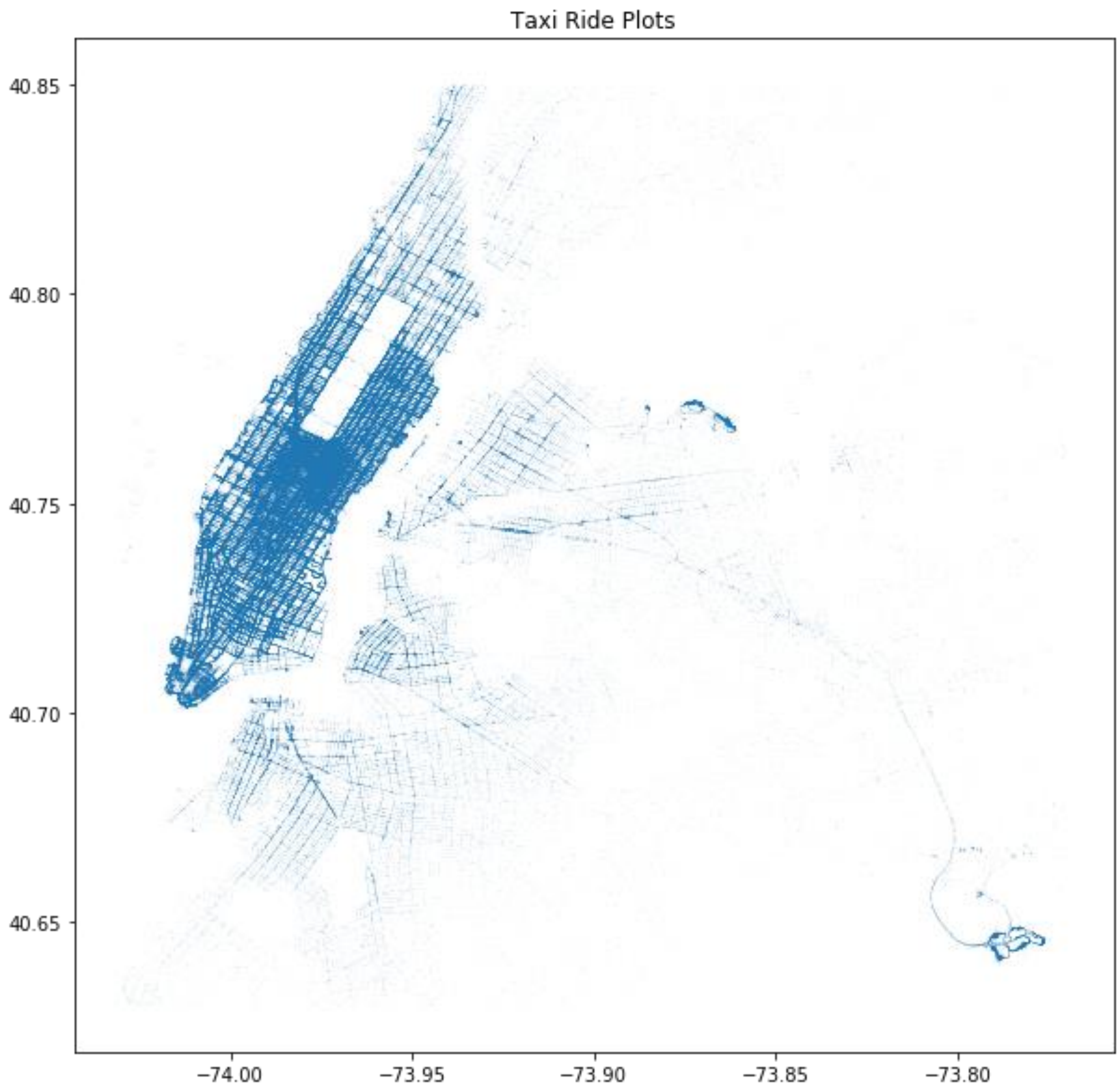
Findings:

- It’s clear from the above plot that on day 0, that is Sunday and day 6 that is Saturday, the trip duration is very less that all the weekdays at 5 AM to 15 AM time.
- See this, on Saturday around midnight, the rides are taking far more than usual time, this is obvious through now verified using given data.

Analysis of trip duration:

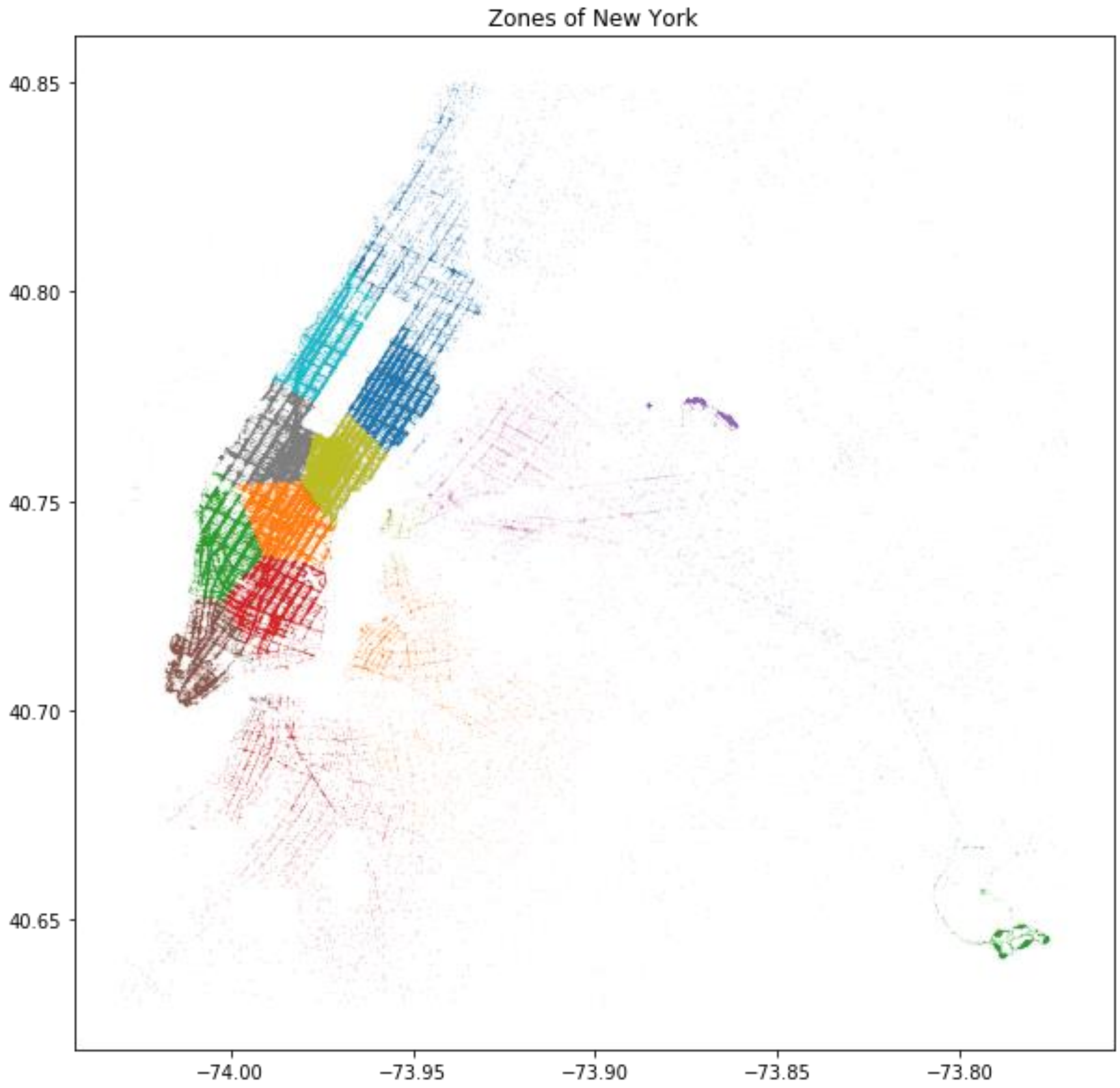
Removed rides to and from far away areas and plotted the rest taxi rides.

Taxi Rides Plot:



Now we used K-means clustering to cluster New York into different groups based on location, and analyze the traffic into and out of every cluster based on the trip duration. One can expect that residential areas would have more incoming traffic in the evening, whereas commercial areas would mostly attract people during the day, and areas with rich nightlife would show more traffic in the night.

We divided the New York city into a total of 15 clusters based on the pick-up and drop-off points of each taxi ride. The plot looks as follows:

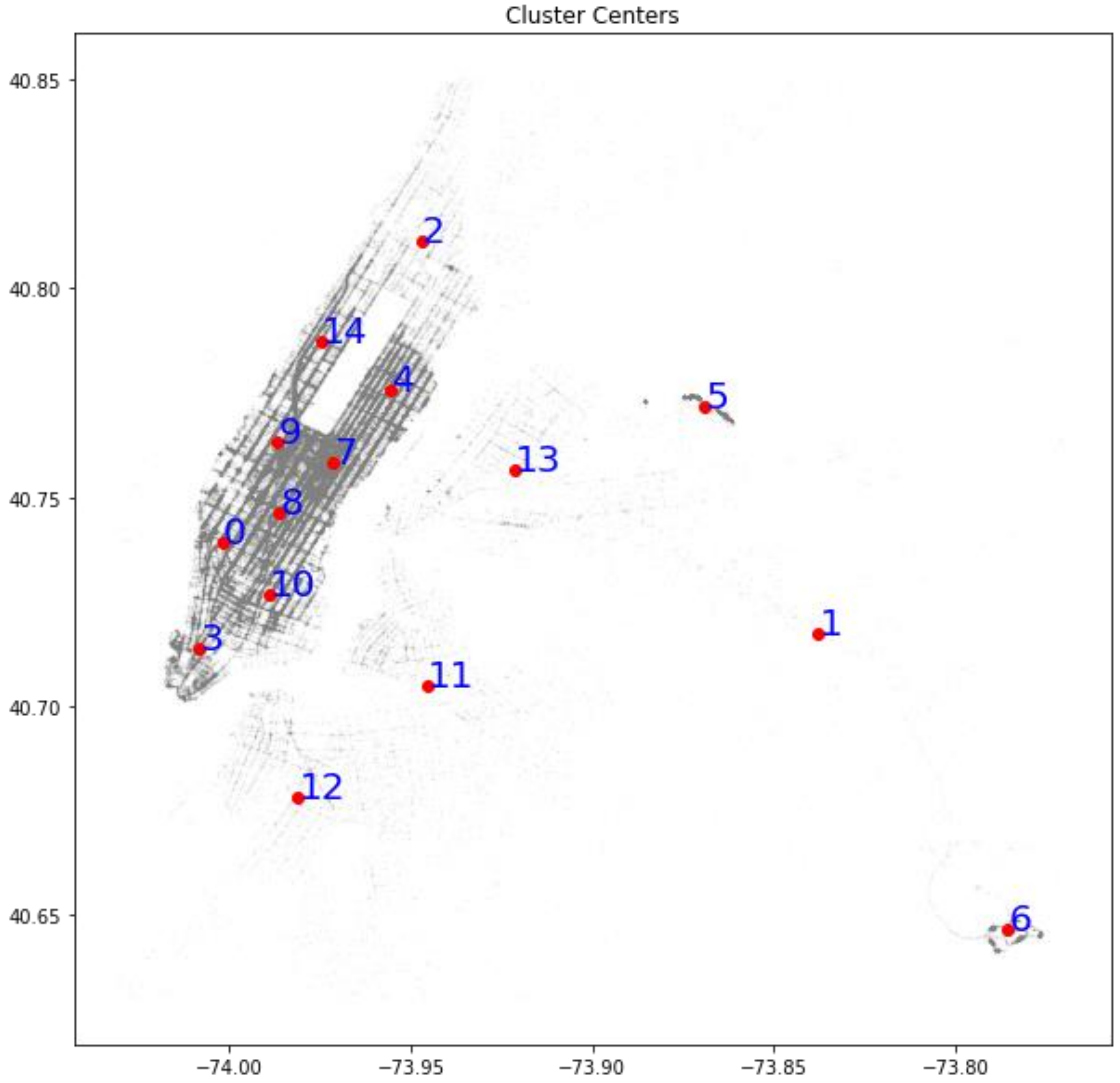


As we can see, the clustering results in a partition which is somewhat similar to the way NY is divided into different neighborhoods. We can see Upper East and West side of Central park in dark blue and sky blue respectively. West midtown in gray, Chelsea and West Village in brown, downtown area in red.

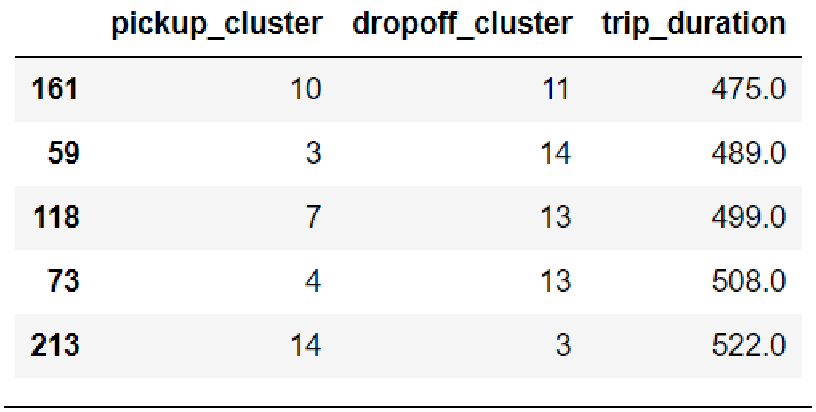
The airports JFK and La LaGuardia have their own cluster. Brooklyn is divided into 2 clusters, and the Bronx has too few rides to be separated from Harlem.

As we have to consider a particular point in each cluster to represent them. We are considering the center or the mid-point of each cluster that we have divided earlier.

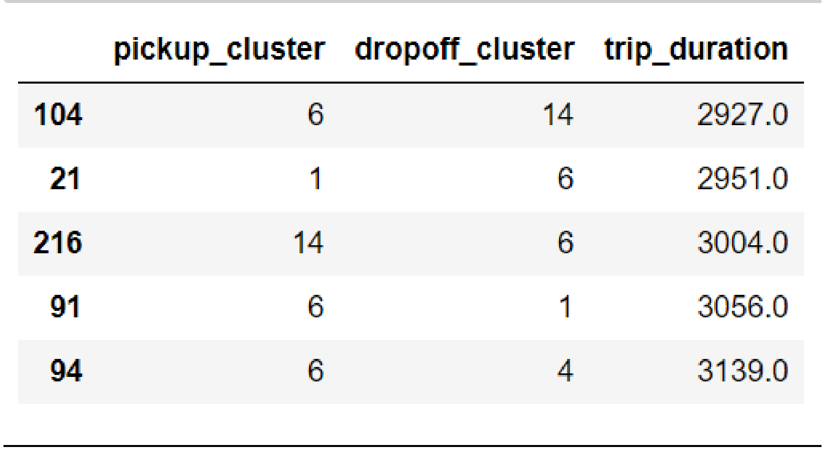
Plot of the cluster centers:



Here are the results that we obtained. The trip duration here is the average trip duration.



	pickup_cluster	dropoff_cluster	trip_duration
161	10	11	475.0
59	3	14	489.0
118	7	13	499.0
73	4	13	508.0
213	14	3	522.0

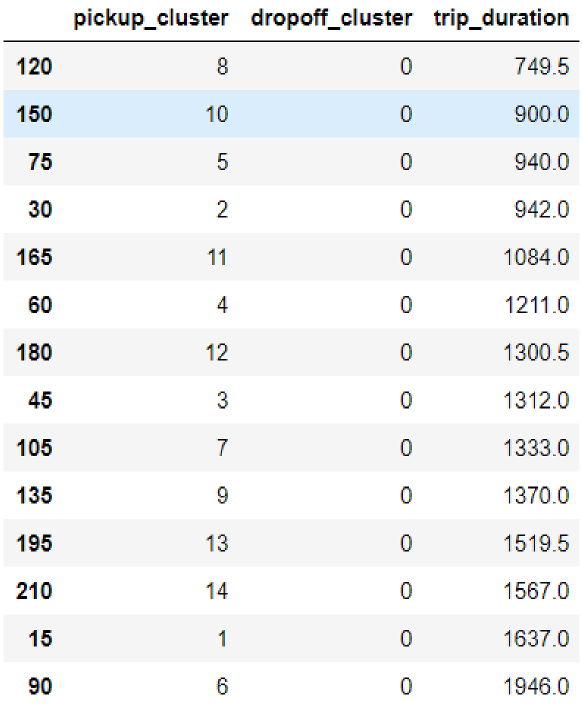


	pickup_cluster	dropoff_cluster	trip_duration
104	6	14	2927.0
21	1	6	2951.0
216	14	6	3004.0
91	6	1	3056.0
94	6	4	3139.0

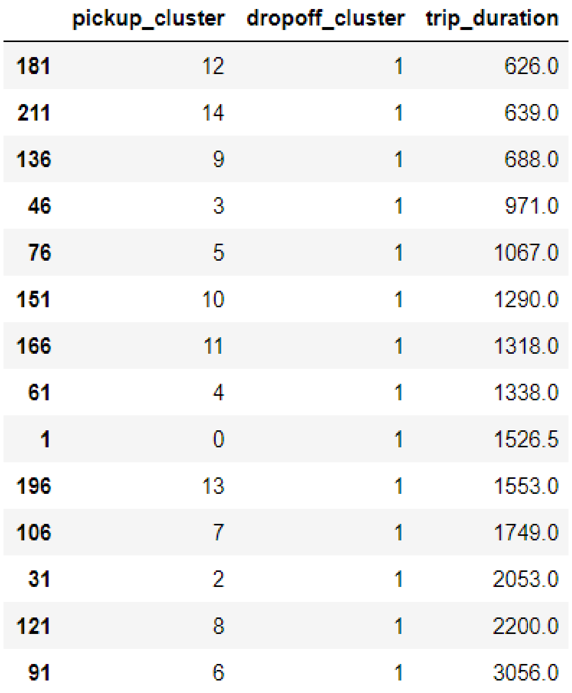
So from above we can observe that the average time taken to travel from cluster 10 to 11 is 475 seconds. While to travel from cluster 6 to 4 it took 3139 seconds which is approximately 52 minutes. Cluster 6 corresponds to airport and cluster 4 is the central park area.

The complete set of results can be observed below. This is the analysis for the average time taken to travel from one cluster to another cluster.

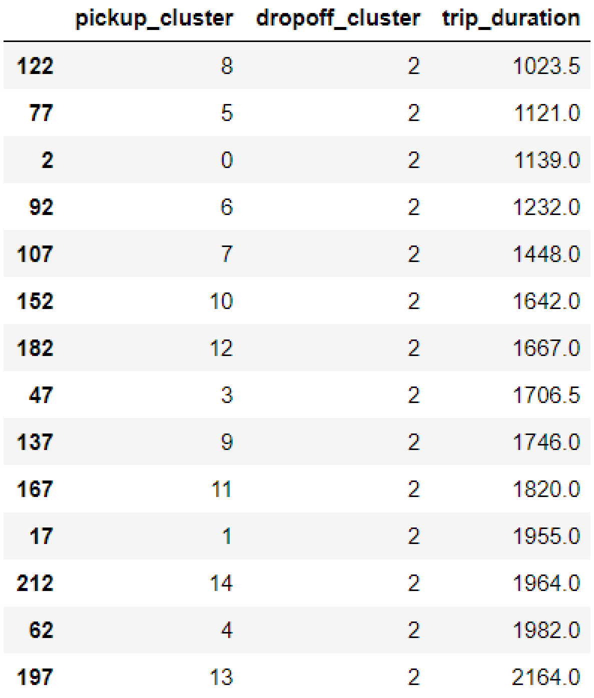
Trip duration (in sec) from different clusters to a particular cluster:



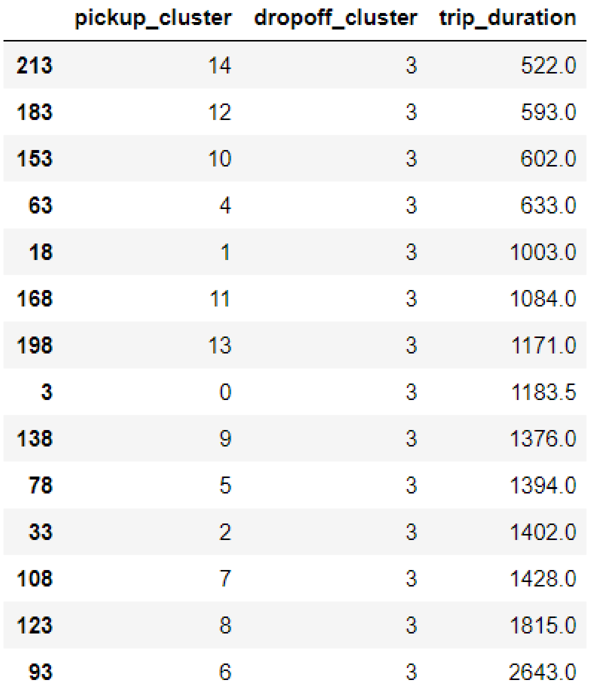
	pickup_cluster	dropoff_cluster	trip_duration
120	8	0	749.5
150	10	0	900.0
75	5	0	940.0
30	2	0	942.0
165	11	0	1084.0
60	4	0	1211.0
180	12	0	1300.5
45	3	0	1312.0
105	7	0	1333.0
135	9	0	1370.0
195	13	0	1519.5
210	14	0	1567.0
15	1	0	1637.0
90	6	0	1946.0



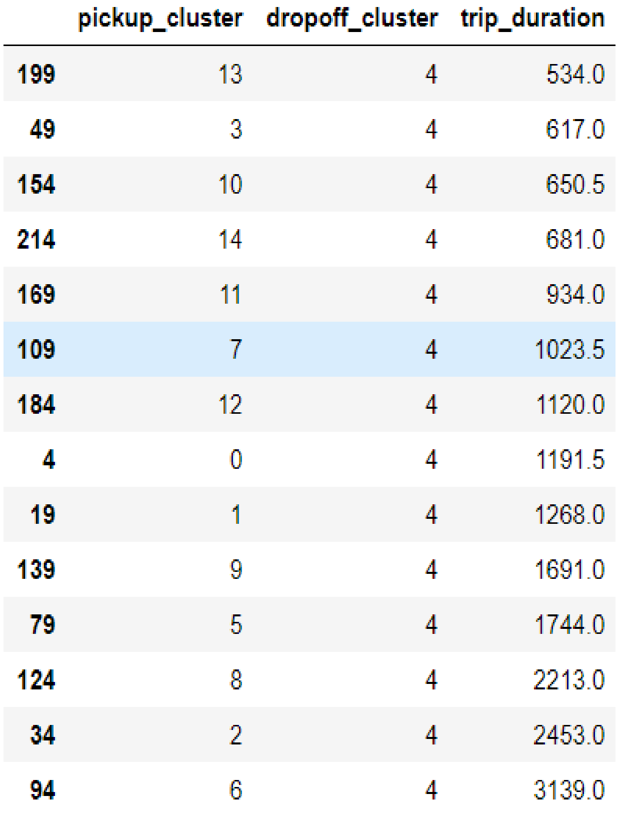
	pickup_cluster	dropoff_cluster	trip_duration
181	12	1	626.0
211	14	1	639.0
136	9	1	688.0
46	3	1	971.0
76	5	1	1067.0
151	10	1	1290.0
166	11	1	1318.0
61	4	1	1338.0
1	0	1	1526.5
196	13	1	1553.0
106	7	1	1749.0
31	2	1	2053.0
121	8	1	2200.0
91	6	1	3056.0



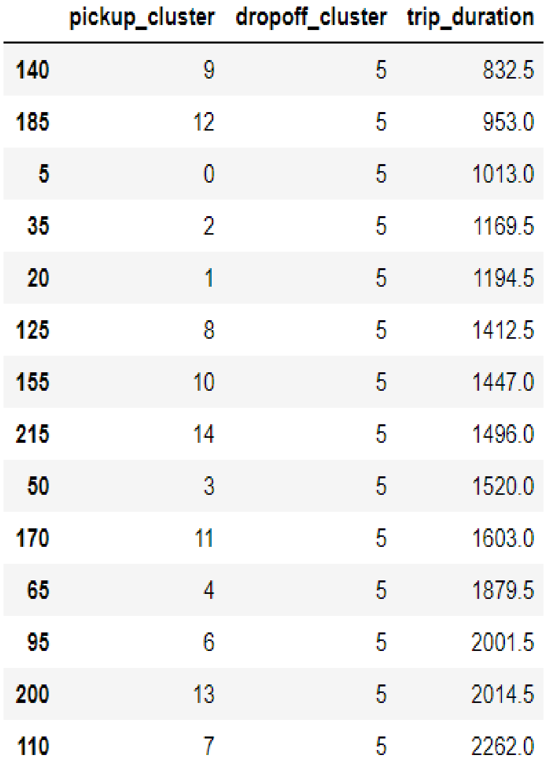
	pickup_cluster	dropoff_cluster	trip_duration
122	8	2	1023.5
77	5	2	1121.0
2	0	2	1139.0
92	6	2	1232.0
107	7	2	1448.0
152	10	2	1642.0
182	12	2	1667.0
47	3	2	1706.5
137	9	2	1746.0
167	11	2	1820.0
17	1	2	1955.0
212	14	2	1964.0
62	4	2	1982.0
197	13	2	2164.0



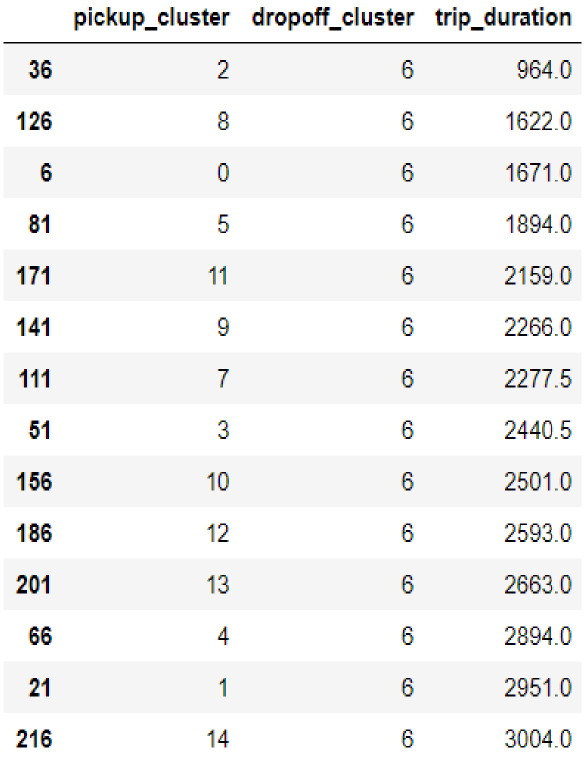
	pickup_cluster	dropoff_cluster	trip_duration
213	14	3	522.0
183	12	3	593.0
153	10	3	602.0
63	4	3	633.0
18	1	3	1003.0
168	11	3	1084.0
198	13	3	1171.0
3	0	3	1183.5
138	9	3	1376.0
78	5	3	1394.0
33	2	3	1402.0
108	7	3	1428.0
123	8	3	1815.0
93	6	3	2643.0



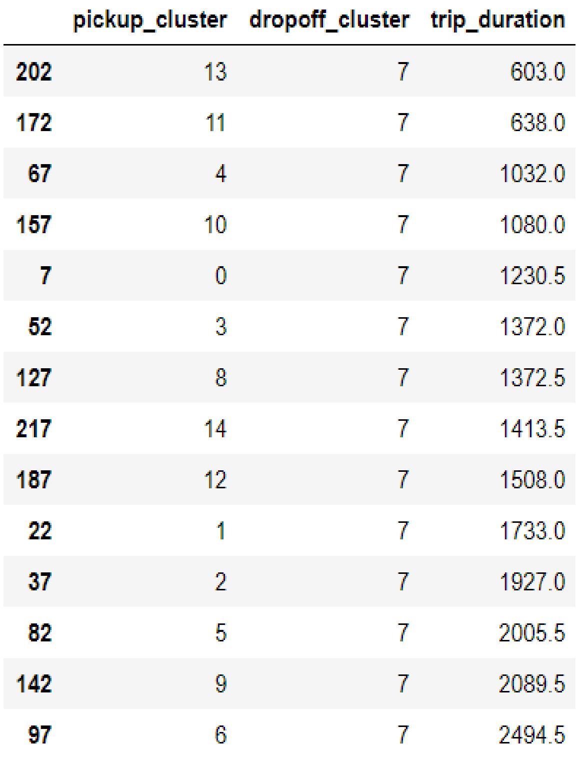
	pickup_cluster	dropoff_cluster	trip_duration
199	13	4	534.0
49	3	4	617.0
154	10	4	650.5
214	14	4	681.0
169	11	4	934.0
109	7	4	1023.5
184	12	4	1120.0
4	0	4	1191.5
19	1	4	1268.0
139	9	4	1691.0
79	5	4	1744.0
124	8	4	2213.0
34	2	4	2453.0
94	6	4	3139.0



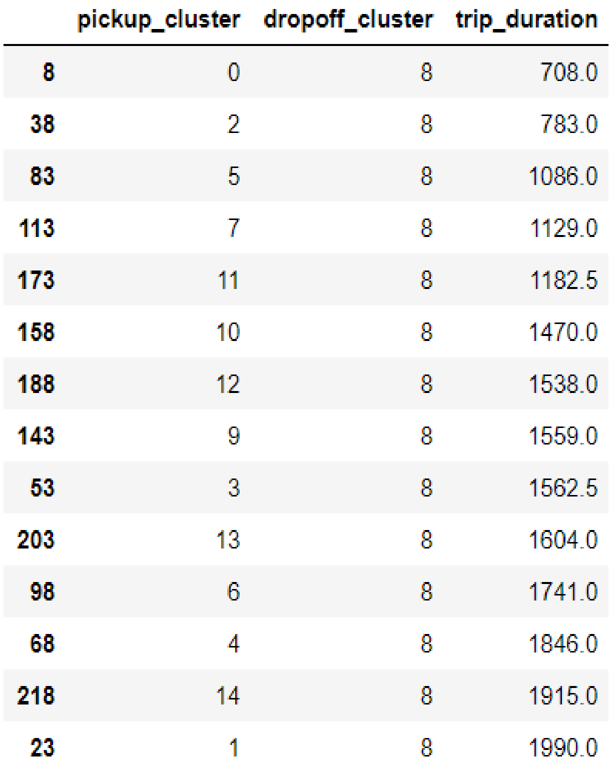
	pickup_cluster	dropoff_cluster	trip_duration
140	9	5	832.5
185	12	5	953.0
5	0	5	1013.0
35	2	5	1169.5
20	1	5	1194.5
125	8	5	1412.5
155	10	5	1447.0
215	14	5	1496.0
50	3	5	1520.0
170	11	5	1603.0
65	4	5	1879.5
95	6	5	2001.5
200	13	5	2014.5
110	7	5	2262.0



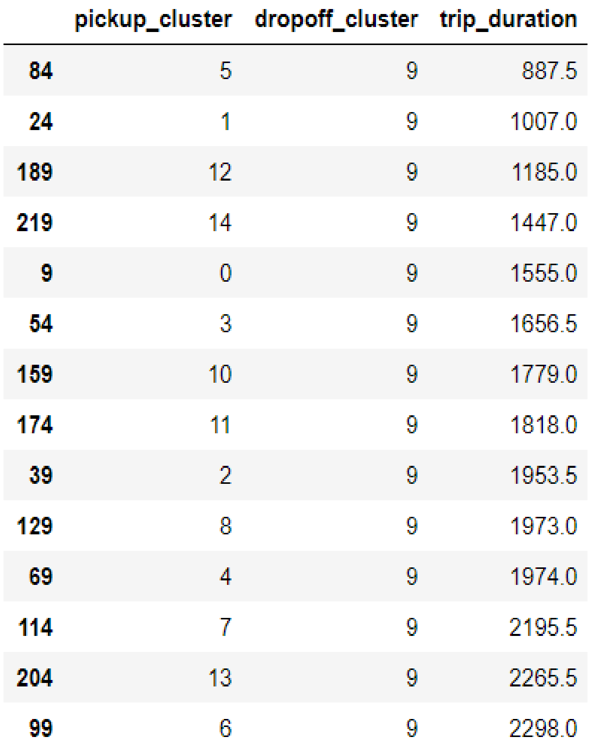
	pickup_cluster	dropoff_cluster	trip_duration
36	2	6	964.0
126	8	6	1622.0
6	0	6	1671.0
81	5	6	1894.0
171	11	6	2159.0
141	9	6	2266.0
111	7	6	2277.5
51	3	6	2440.5
156	10	6	2501.0
186	12	6	2593.0
201	13	6	2663.0
66	4	6	2894.0
21	1	6	2951.0
216	14	6	3004.0



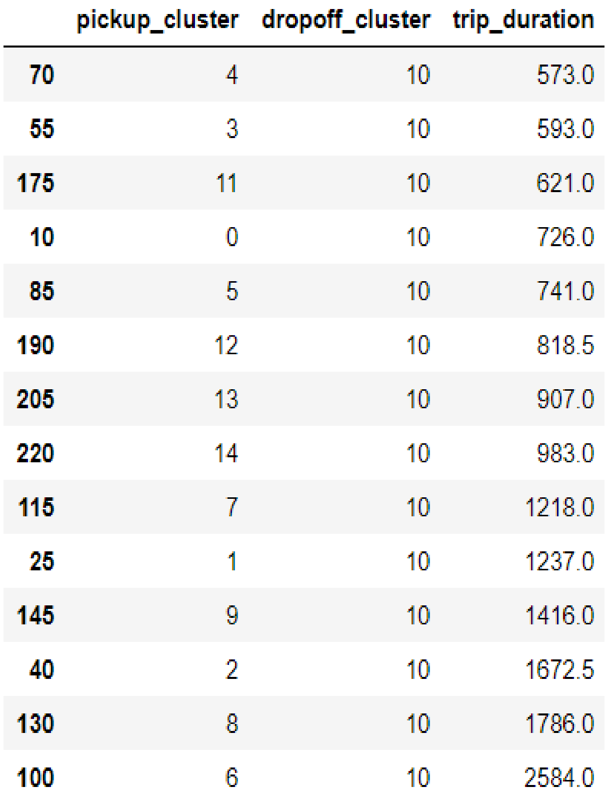
	pickup_cluster	dropoff_cluster	trip_duration
202	13	7	603.0
172	11	7	638.0
67	4	7	1032.0
157	10	7	1080.0
7	0	7	1230.5
52	3	7	1372.0
127	8	7	1372.5
217	14	7	1413.5
187	12	7	1508.0
22	1	7	1733.0
37	2	7	1927.0
82	5	7	2005.5
142	9	7	2089.5
97	6	7	2494.5



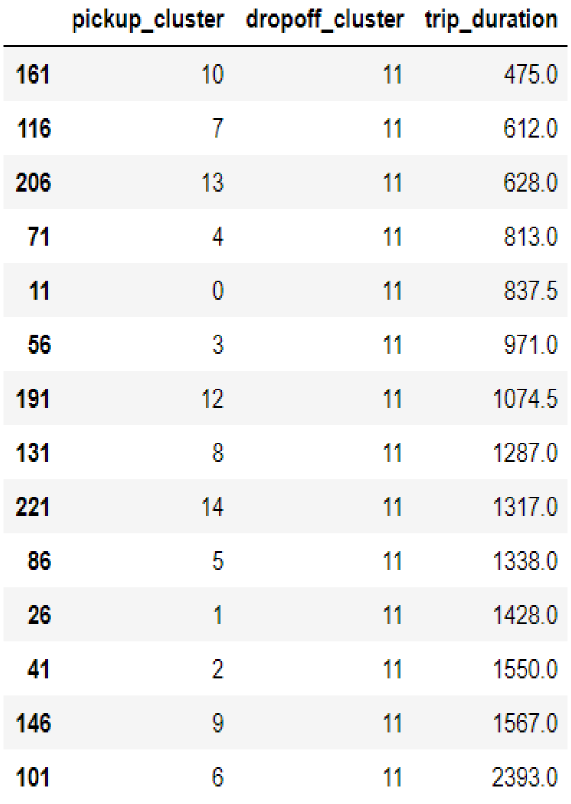
	pickup_cluster	dropoff_cluster	trip_duration
8	0	8	708.0
38	2	8	783.0
83	5	8	1086.0
113	7	8	1129.0
173	11	8	1182.5
158	10	8	1470.0
188	12	8	1538.0
143	9	8	1559.0
53	3	8	1562.5
203	13	8	1604.0
98	6	8	1741.0
68	4	8	1846.0
218	14	8	1915.0
23	1	8	1990.0



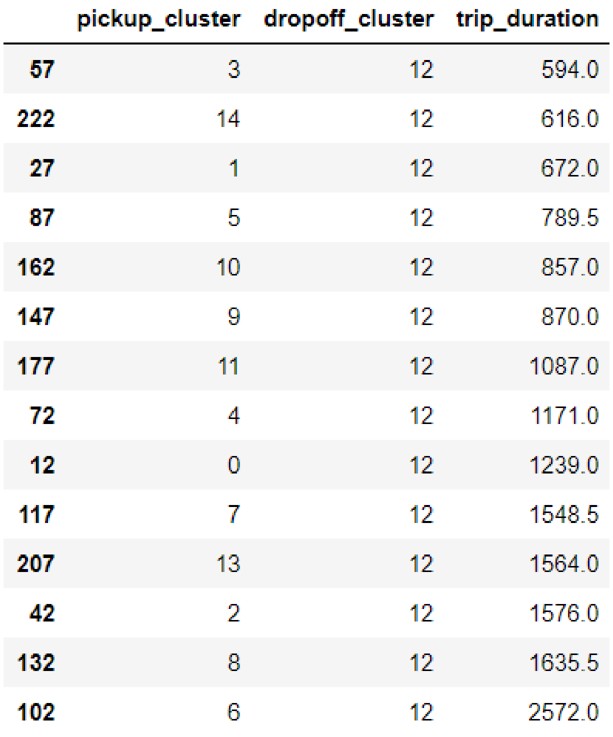
	pickup_cluster	dropoff_cluster	trip_duration
84	5	9	887.5
24	1	9	1007.0
189	12	9	1185.0
219	14	9	1447.0
9	0	9	1555.0
54	3	9	1656.5
159	10	9	1779.0
174	11	9	1818.0
39	2	9	1953.5
129	8	9	1973.0
69	4	9	1974.0
114	7	9	2195.5
204	13	9	2265.5
99	6	9	2298.0



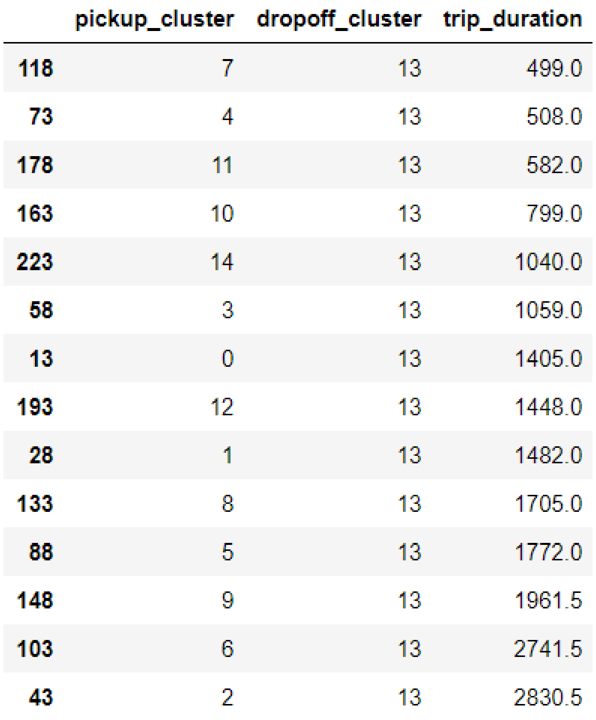
	pickup_cluster	dropoff_cluster	trip_duration
70	4	10	573.0
55	3	10	593.0
175	11	10	621.0
10	0	10	726.0
85	5	10	741.0
190	12	10	818.5
205	13	10	907.0
220	14	10	983.0
115	7	10	1218.0
25	1	10	1237.0
145	9	10	1416.0
40	2	10	1672.5
130	8	10	1786.0
100	6	10	2584.0



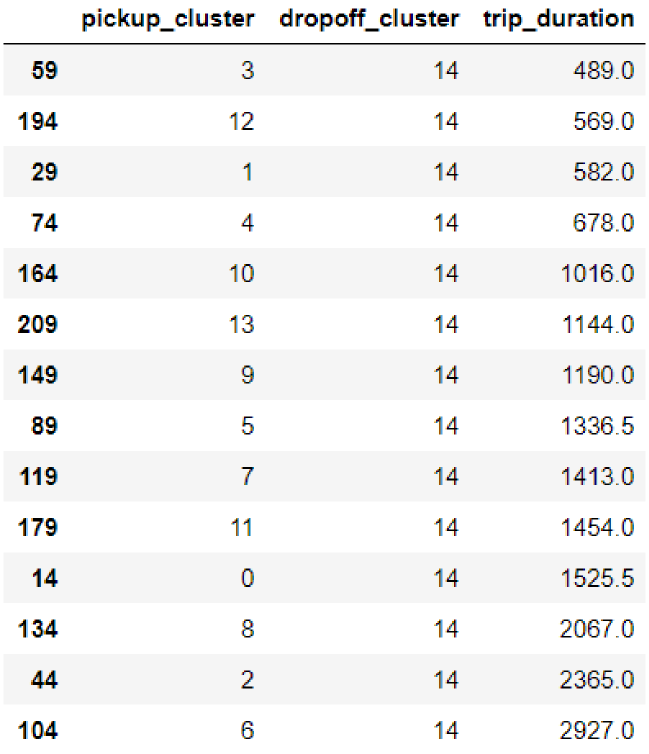
	pickup_cluster	dropoff_cluster	trip_duration
161	10	11	475.0
116	7	11	612.0
206	13	11	628.0
71	4	11	813.0
11	0	11	837.5
56	3	11	971.0
191	12	11	1074.5
131	8	11	1287.0
221	14	11	1317.0
86	5	11	1338.0
26	1	11	1428.0
41	2	11	1550.0
146	9	11	1567.0
101	6	11	2393.0



	pickup_cluster	dropoff_cluster	trip_duration
57	3	12	594.0
222	14	12	616.0
27	1	12	672.0
87	5	12	789.5
162	10	12	857.0
147	9	12	870.0
177	11	12	1087.0
72	4	12	1171.0
12	0	12	1239.0
117	7	12	1548.5
207	13	12	1564.0
42	2	12	1576.0
132	8	12	1635.5
102	6	12	2572.0



	pickup_cluster	dropoff_cluster	trip_duration
118	7	13	499.0
73	4	13	508.0
178	11	13	582.0
163	10	13	799.0
223	14	13	1040.0
58	3	13	1059.0
13	0	13	1405.0
193	12	13	1448.0
28	1	13	1482.0
133	8	13	1705.0
88	5	13	1772.0
148	9	13	1961.5
103	6	13	2741.5
43	2	13	2830.5



	pickup_cluster	dropoff_cluster	trip_duration
59	3	14	489.0
194	12	14	569.0
29	1	14	582.0
74	4	14	678.0
164	10	14	1016.0
209	13	14	1144.0
149	9	14	1190.0
89	5	14	1336.5
119	7	14	1413.0
179	11	14	1454.0
14	0	14	1525.5
134	8	14	2067.0
44	2	14	2365.0
104	6	14	2927.0

CONCLUSION:

All the estimated times are calculated to travel from one zone of New York city to another zone.

It can be observed that the travel time to airport (cluster 6) from all other zones are high. So those who are travelling to airport should start off early than usual expected time.

FUTURE RESEARCH DIRECTIONS:

- Calculate the actual distance with precision between two data points using Google API.
- Analyze the Average time taken, Average speed between two different zones in New York city.
- Over speeding by taxis.