Bigdata HW01
Analyzing NYC Taxi Data

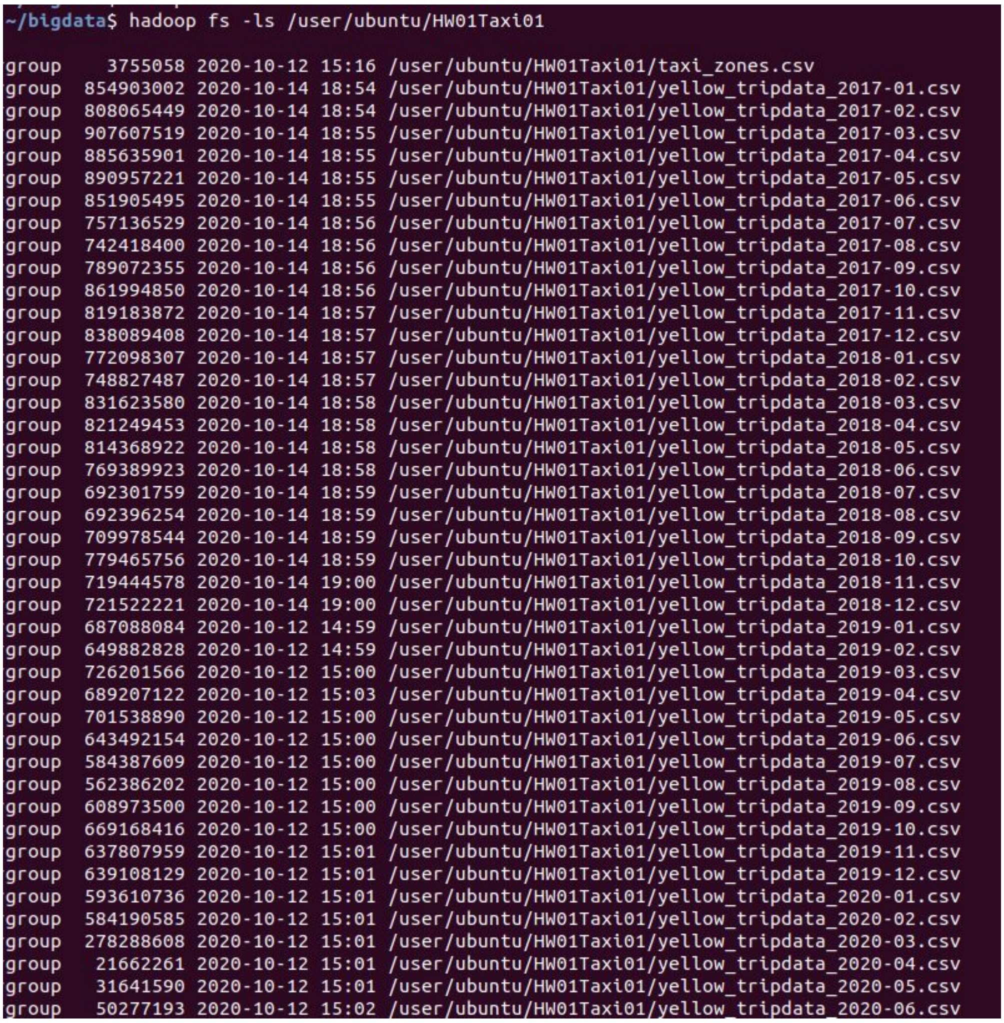
Student

資工碩一 0856169 林建興

Experiment

Scale of data :

2017/01 ~2020/06 Yellow Taxi (28.4GB)
data rows : 317,547,921 rows



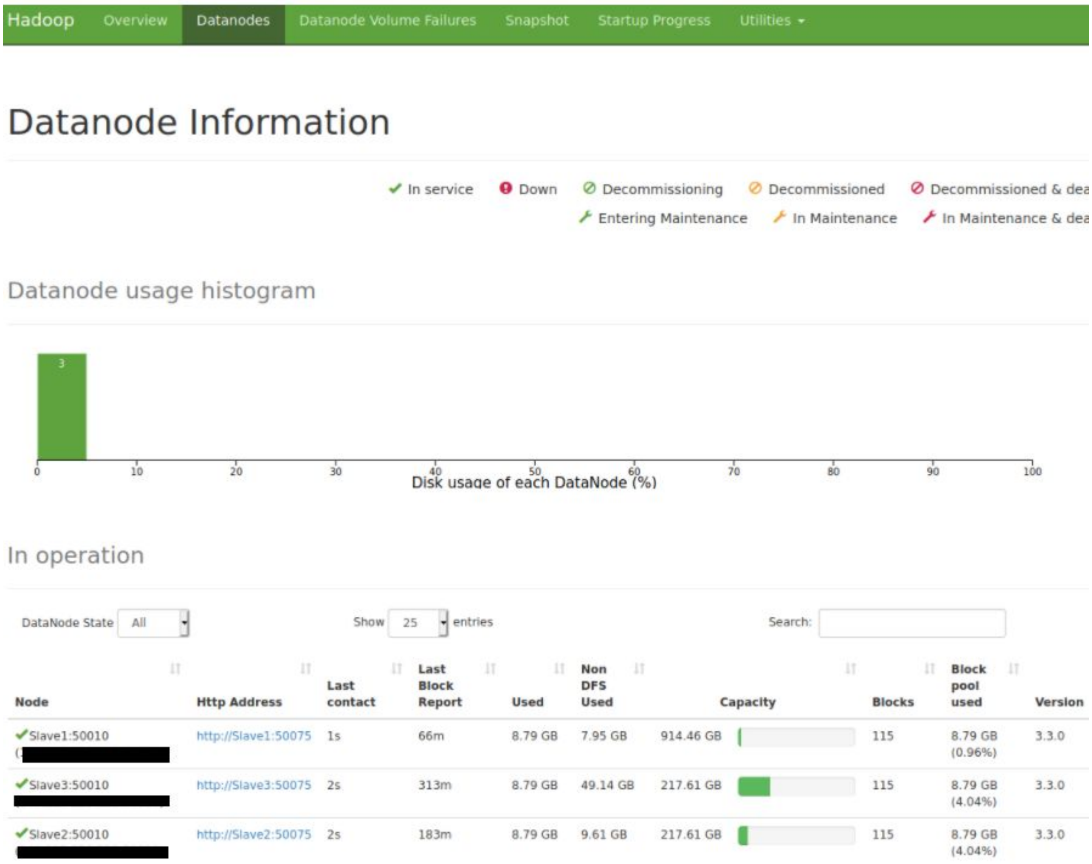
```
~/bigdata$ hadoop fs -ls /user/ubuntu/HW01Taxi01

group    3755058 2020-10-12 15:16 /user/ubuntu/HW01Taxi01/taxi_zones.csv
group  854903002 2020-10-14 18:54 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-01.csv
group  808065449 2020-10-14 18:54 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-02.csv
group  907607519 2020-10-14 18:55 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-03.csv
group  885635901 2020-10-14 18:55 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-04.csv
group  890957221 2020-10-14 18:55 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-05.csv
group  851905495 2020-10-14 18:55 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-06.csv
group  757136529 2020-10-14 18:56 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-07.csv
group  742418400 2020-10-14 18:56 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-08.csv
group  789072355 2020-10-14 18:56 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-09.csv
group  861994850 2020-10-14 18:56 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-10.csv
group  819183872 2020-10-14 18:57 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-11.csv
group  838089408 2020-10-14 18:57 /user/ubuntu/HW01Taxi01/yellow_tripdata_2017-12.csv
group  772098307 2020-10-14 18:57 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-01.csv
group  748827487 2020-10-14 18:57 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-02.csv
group  831623580 2020-10-14 18:58 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-03.csv
group  821249453 2020-10-14 18:58 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-04.csv
group  814368922 2020-10-14 18:58 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-05.csv
group  769389923 2020-10-14 18:58 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-06.csv
group  692301759 2020-10-14 18:59 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-07.csv
group  692396254 2020-10-14 18:59 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-08.csv
group  709978544 2020-10-14 18:59 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-09.csv
group  779465756 2020-10-14 18:59 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-10.csv
group  719444578 2020-10-14 19:00 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-11.csv
group  721522221 2020-10-14 19:00 /user/ubuntu/HW01Taxi01/yellow_tripdata_2018-12.csv
group  687088084 2020-10-12 14:59 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-01.csv
group  649882828 2020-10-12 14:59 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-02.csv
group  726201566 2020-10-12 15:00 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-03.csv
group  689207122 2020-10-12 15:03 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-04.csv
group  701538890 2020-10-12 15:00 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-05.csv
group  643492154 2020-10-12 15:00 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-06.csv
group  584387609 2020-10-12 15:00 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-07.csv
group  562386202 2020-10-12 15:00 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-08.csv
group  608973500 2020-10-12 15:00 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-09.csv
group  669168416 2020-10-12 15:00 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-10.csv
group  637807959 2020-10-12 15:01 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-11.csv
group  639108129 2020-10-12 15:01 /user/ubuntu/HW01Taxi01/yellow_tripdata_2019-12.csv
group  593610736 2020-10-12 15:01 /user/ubuntu/HW01Taxi01/yellow_tripdata_2020-01.csv
group  584190585 2020-10-12 15:01 /user/ubuntu/HW01Taxi01/yellow_tripdata_2020-02.csv
group  278288608 2020-10-12 15:01 /user/ubuntu/HW01Taxi01/yellow_tripdata_2020-03.csv
group   21662261 2020-10-12 15:01 /user/ubuntu/HW01Taxi01/yellow_tripdata_2020-04.csv
group   31641590 2020-10-12 15:01 /user/ubuntu/HW01Taxi01/yellow_tripdata_2020-05.csv
group   50277193 2020-10-12 15:02 /user/ubuntu/HW01Taxi01/yellow_tripdata_2020-06.csv
```

Tool :

OS: Linux, IDE : spyder, Lib : pyspark

Spec of platform :

hadoop HDFS



Datanode Information

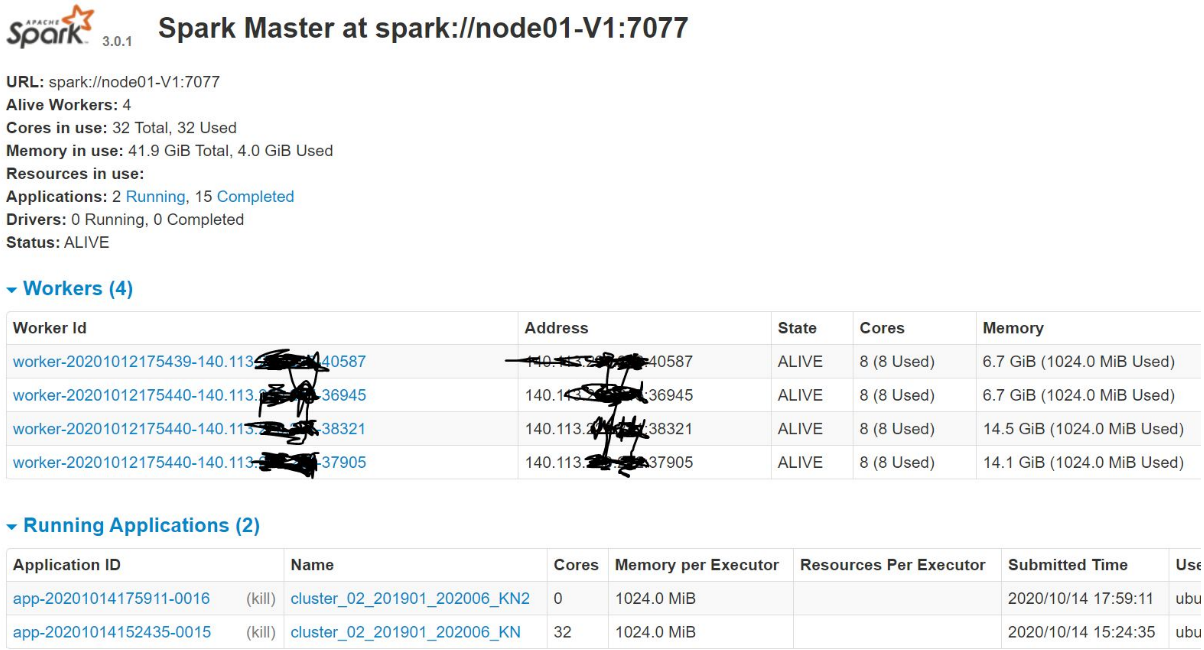
✓ In service ● Down ⊘ Decommissioning ⊘ Decommissioned ⊘ Decommissioned & dea
Entering Maintenance In Maintenance In Maintenance & dea

Datanode usage histogram

In operation

Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Version
✓Slave1:50010 ([redacted]	http://Slave1:50075	1s	66m	8.79 GB	7.95 GB	914.46 GB	115	8.79 GB (0.96%)	3.3.0
✓Slave3:50010 [redacted]	http://Slave3:50075	2s	313m	8.79 GB	49.14 GB	217.61 GB	115	8.79 GB (4.04%)	3.3.0
✓Slave2:50010 [redacted]	http://Slave2:50075	2s	183m	8.79 GB	9.61 GB	217.61 GB	115	8.79 GB (4.04%)	3.3.0

spark standalone



Spark Master at spark://node01-V1:7077

URL: spark://node01-V1:7077
Alive Workers: 4
Cores in use: 32 Total, 32 Used
Memory in use: 41.9 GiB Total, 4.0 GiB Used
Resources in use:
Applications: 2 Running, 15 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

▾ Workers (4)

Worker Id	Address	State	Cores	Memory
worker-20201012175439-140.113[redacted]40587	[redacted]40587	ALIVE	8 (8 Used)	6.7 GiB (1024.0 MiB Used)
worker-20201012175440-140.113.[redacted]36945	140.1[redacted]36945	ALIVE	8 (8 Used)	6.7 GiB (1024.0 MiB Used)
worker-20201012175440-140.113[redacted]38321	140.113.[redacted]38321	ALIVE	8 (8 Used)	14.5 GiB (1024.0 MiB Used)
worker-20201012175440-140.113[redacted]37905	140.113.[redacted]37905	ALIVE	8 (8 Used)	14.1 GiB (1024.0 MiB Used)

▾ Running Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	Use
app-20201014175911-0016 (kill)	cluster_02_201901_202006_KN2	0	1024.0 MiB		2020/10/14 17:59:11	ubu
app-20201014152435-0015 (kill)	cluster_02_201901_202006_KN	32	1024.0 MiB		2020/10/14 15:24:35	ubu

Data preparation

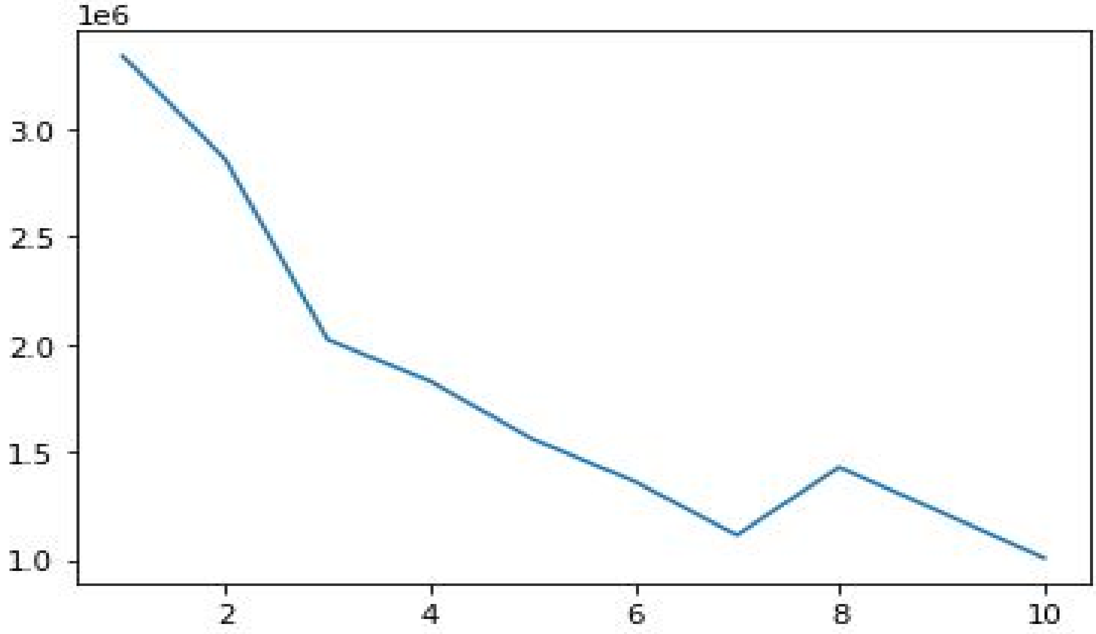
I combine data between 201701 to 202006.
I also merge longitude and latitude by locationID.
Limit trips distance in 0 to 200 miles.

Questions

Q1: What are the most pickups and drop offs regions?

1.1 How many clusters should it choose?

7 is the best by elbow method.

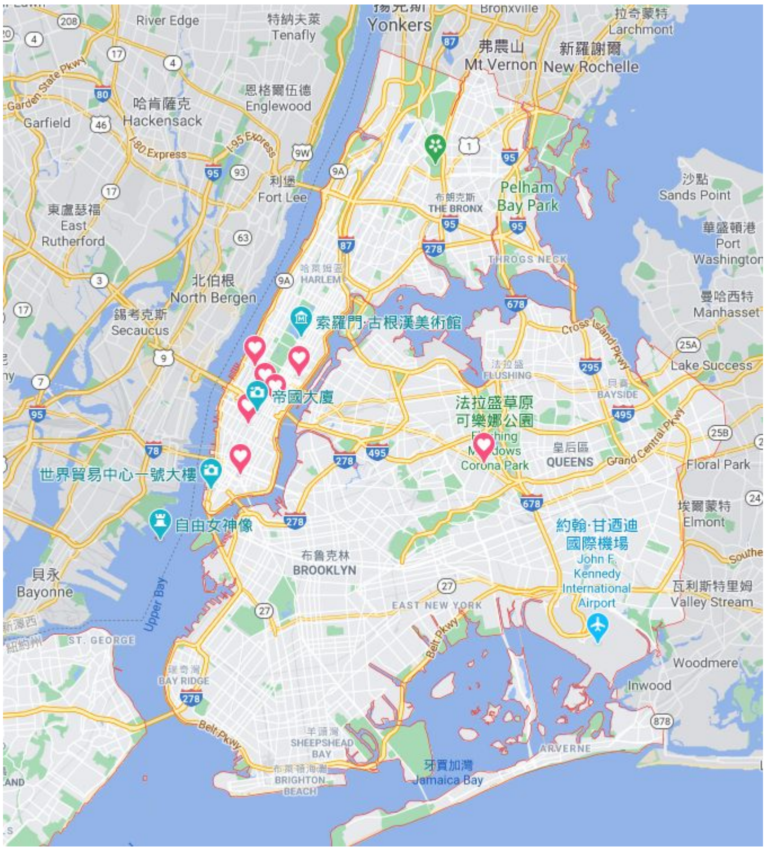


So I put 3 years data to KDD version K-mean by Spark MLlib.

1.2 Pickups regions cluster by longitude and latitude.

Within Set Sum of Squared Error = 4344119.726058249

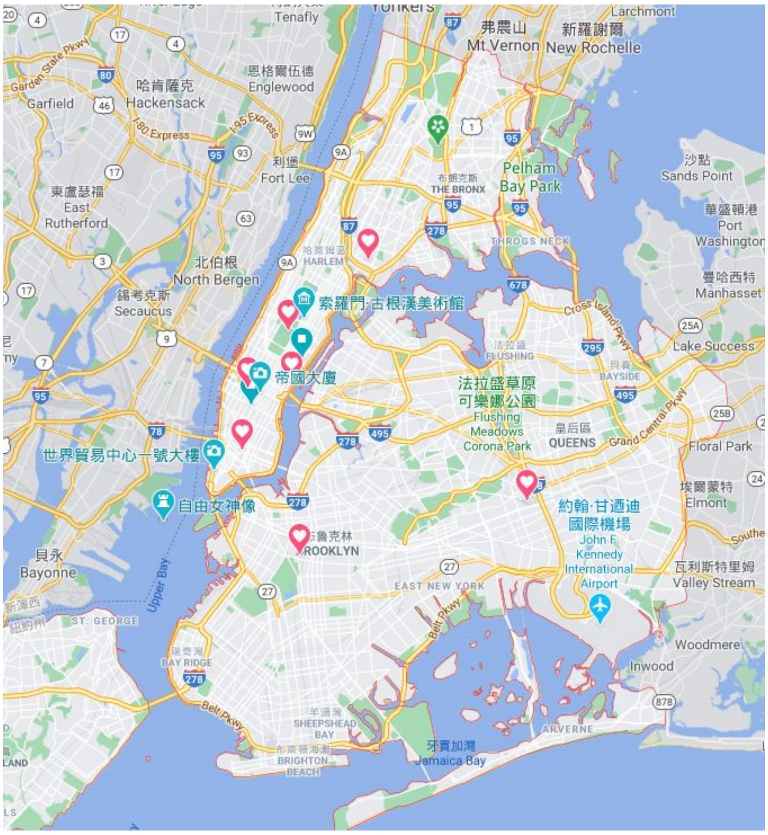
Class	Count	Cluster Center Zone	longitude	latitude
3	**69,038,847**	**Midtown/ Manhattan**	**-73.99124083**	**40.74283413**
6	56,707,777	Manhattan/ Upper East Side	-73.95980444	40.76428019
0	55,999,661	Manhattan/ Little Italy/ NoLiTa	-73.99579419	40.7192809
2	53,692,060	Manhattan/ Midtown Center	-73.97434821	40.75162814
5	**30,336,400**	**Manhattan/ Central Park**	**-73.95886571**	**40.7948278**
1	29,087,603	Manhattan/ Lincoln Square East	-73.98670798	40.76906572
4	18,314,090	Queens/ Forest Hills	-73.84749976	40.72468127



Heart point is the Pickups cluster center

1.3 Drop off regions cluster by longitude and latitude.

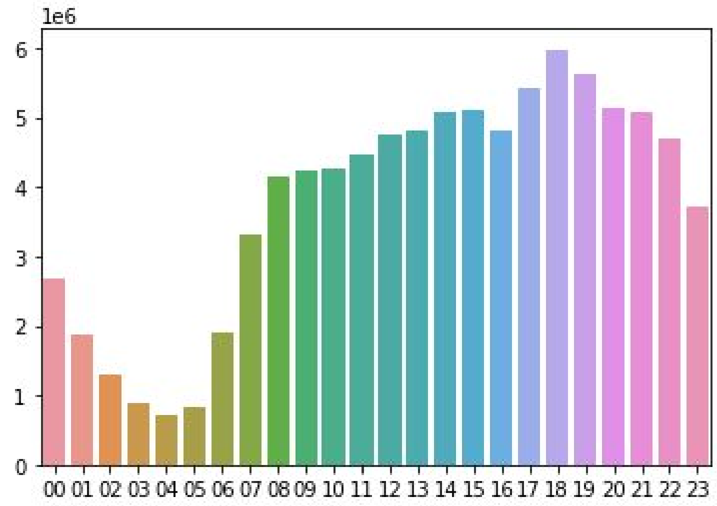
Class	Count	Cluster Center Zone	longitude	latitude
3	**54,313,979**	**Manhattan/ Sutton Place/ Turtle Bay North**	**-73.96595111**	**40.75292648**
1	52,323,941	Manhattan/ Little Italy / NoLiTa	-73.99636818	40.72195639
4	**46,106,717**	**Manhattan/ Central Park**	**-73.96824718**	**40.77683572**
0	44,946,260	Manhattan / Penn Station/ Madison Sq West	-73.99324856	40.75063729
5	21,834,979	Bronx / Mott Haven/Port Morris	-73.91941332	40.80967138
6	8,025,546	Brooklyn / Prospect Heights	-73.9616439	40.67340502
2	5,929,182	Queens/ Richmond Hill	-73.82269243	40.69872953



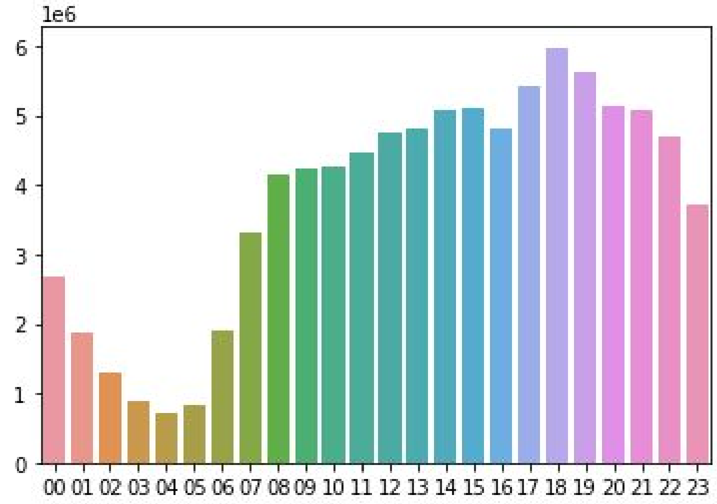
Heart point is the Drop off cluster center.

Q2: When are the peak hours and off-peak hours for taking taxi?

pick up_hour: I count the number of pickups in different hours of day.



drop off_hour: I count the number of dropoffs in different hours of day.

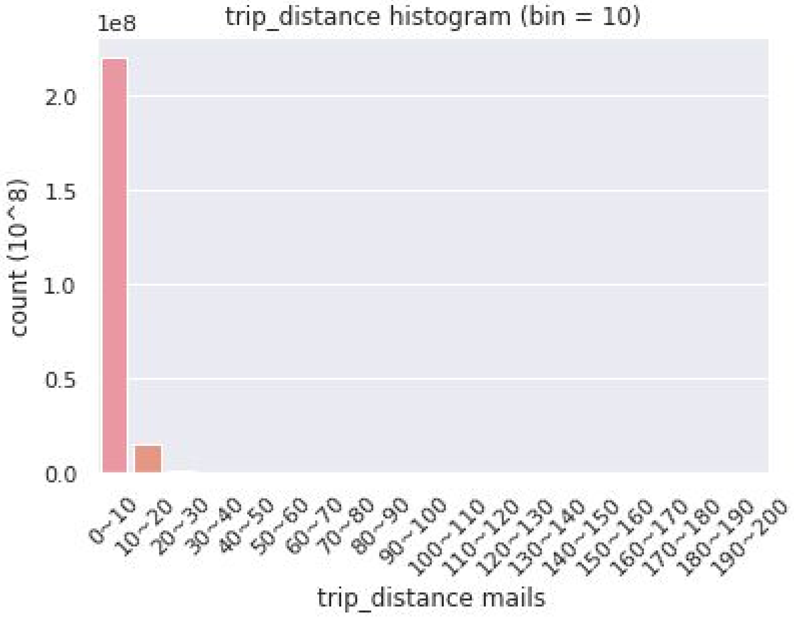


As you could see, The peak hour both **in the 18:00**.

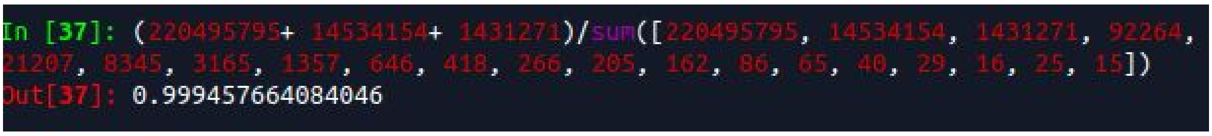
Q3: What are the differences between short and long distance trips of taking taxi?

3.1 defintion of short distance

I define short distance as **trips distance trips lower than 30 miles**. According to the result by the histogram of distance trips column. There are 99.99% datas lower than 30 miles.



([1.0, 10., 20., 30., 40., 50., 60., 70., 79., 89., 99., 109., 119.40399780273437, 129., 139., 149., 158., 168., 178., 188., 198.],
[**220495795**, **14534154**, **1431271**, 92264, 21207, 8345, 3165, 1357, 646, 418, 266, 205, 162, 86, 65, 40, 29, 16, 25, 15])



```
In [37]: (220495795+ 14534154+ 1431271)/sum([220495795, 14534154, 1431271, 92264,
21207, 8345, 3165, 1357, 646, 418, 266, 205, 162, 86, 65, 40, 29, 16, 25, 15])
Out[37]: 0.999457664084046
```

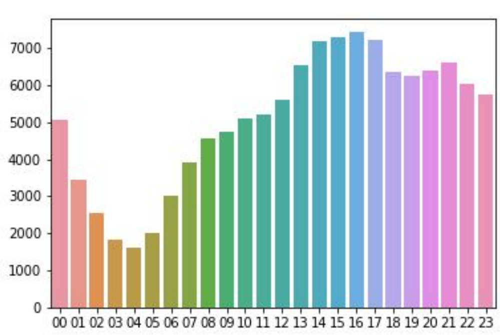
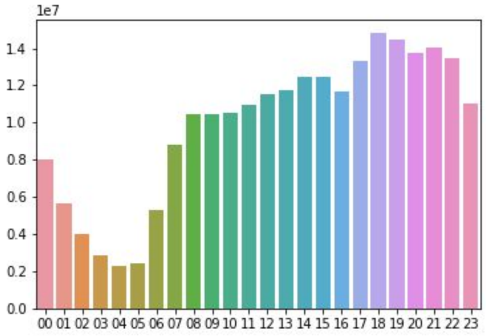
3.2 Top 5 most pickups and drop offs region between short and long distance trips.

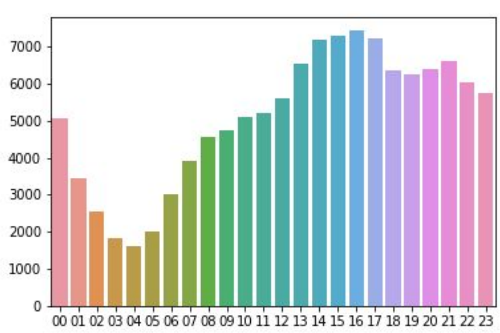
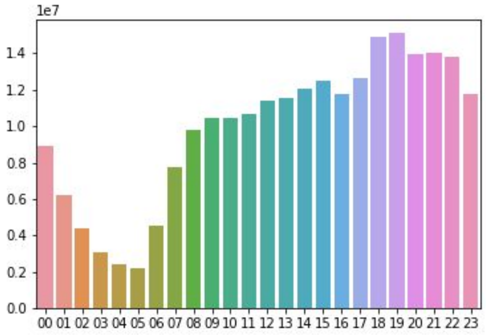
long_trip_PUL				short_trip_PUL			
ID	Borogh	Name	Count	ID	Borogh	Name	Count
132	Queens	JFK Airport	58788	161	Manhattan	Midtown Center	8856818
138	Queens	LaGuardia Airport	15176	186	Manhattan	Penn Station /Madison Sq West	8685190
230	Manhattan	Times Sq /Theatre District	1945	237	Manhattan	Upper East Side South	8335182
186	Manhattan	Penn Station /Madison Sq West	1941	138	Queens	LaGuardia Airport	8269230
48	Manhattan	Clinton East	1840	162	Manhattan	Midtown East	8216733

..

long_trip_DOL				short_trip_DOL			
ID	Borogh	Name	Count	ID	Borogh	Name	Count
1	EWR	Newark Airport	14705	236	Manhattan	Upper East Side North	8224325
132	Queens	JFK Airport	6821	161	Manhattan	Midtown Center	8137228
44	**Staten Island**	Charleston /Tottenville	1242	237	Manhattan	Upper East Side South	7223824
84	**Staten Island**	Eltingville/ Annadale/ Prince's Bay	800	162	Manhattan	Midtown East	6890128
23	**Staten Island**	Bloomfield/ Emerson Hill	687	170	Manhattan	Murray Hill	6770043

3.3 The peak hours and off-peak hours between short and long distance trips.

Long pick up hours	Short pick up hours
	

Long drop off hours	Short drop off hours
	

We can see the **peak** of pick up and drop off hours in **long distance** trips both are **16:00** .
The **off-peak** of pick up and drop off hours in **long distance** trips both are **04:00** .
The **peak** of **pick up** hours in **Short distance** trips both are **18:00** .
The **peak** of **drop off** hours in **Short distance** trips both are **17:00** .
The **off-peak** of **pick up** hours in **Short distance** trips, both are **04:00** .
The **off-peak** of **drop off** hours in **Short distance** trips, both are **05:00** .

Other Observations

1. In the “trip_distance” colums ,there are some error datas which are **negative number.** It is no reason. So I drop these negative numbers.
2. LocationID 264 is N/V. LocationID 265 is N/A.
3. K-mean seen like not a good cluster method.
4. 28.4GB data will overflow in single PC.