

Analyzing Taxi Data

Dataset: 2015 July – Dec Taxi Data and Weather Data

Total: 69006887 data points

Tool:

- **Python:** I choose Python because it is convenient and I free at ease using it.
- **Pandas:** Provide high-performance, easy-to-use data structures and data analysis tools for Python.
- **Google Map Api:** Because I want to draw the data point on map.
- **Scikit Learn:** Simple and efficient tools for data mining and analysis in Python.
 1. Kmeans
 2. MiniBatchKmeans
 3. DBSCAN

Get Rid of Noise Data:

1. Location

Some longitude and latitude data is missing or locate at somewhere car can't drive to.

➔ 69006887 – 1083663(Location Noise) = 67923224 data points

2. Time

A few of data records have strange picktime and droptime gap. For example,

Picktime: 2015-07-02 14:04:06 Droptime: 2015-12-09 07:33:08

➔ 67923224 – 435(Time Noise) = 67922789 data points

3. Tip

Few data have negative tip, but I am not sure it is noisy or not. So I didn't get rid of it at first.

4. Passenger

Few data have zero passengers.

➔ 67923224 – 2842(Passenger Noise) = 67919947 data points

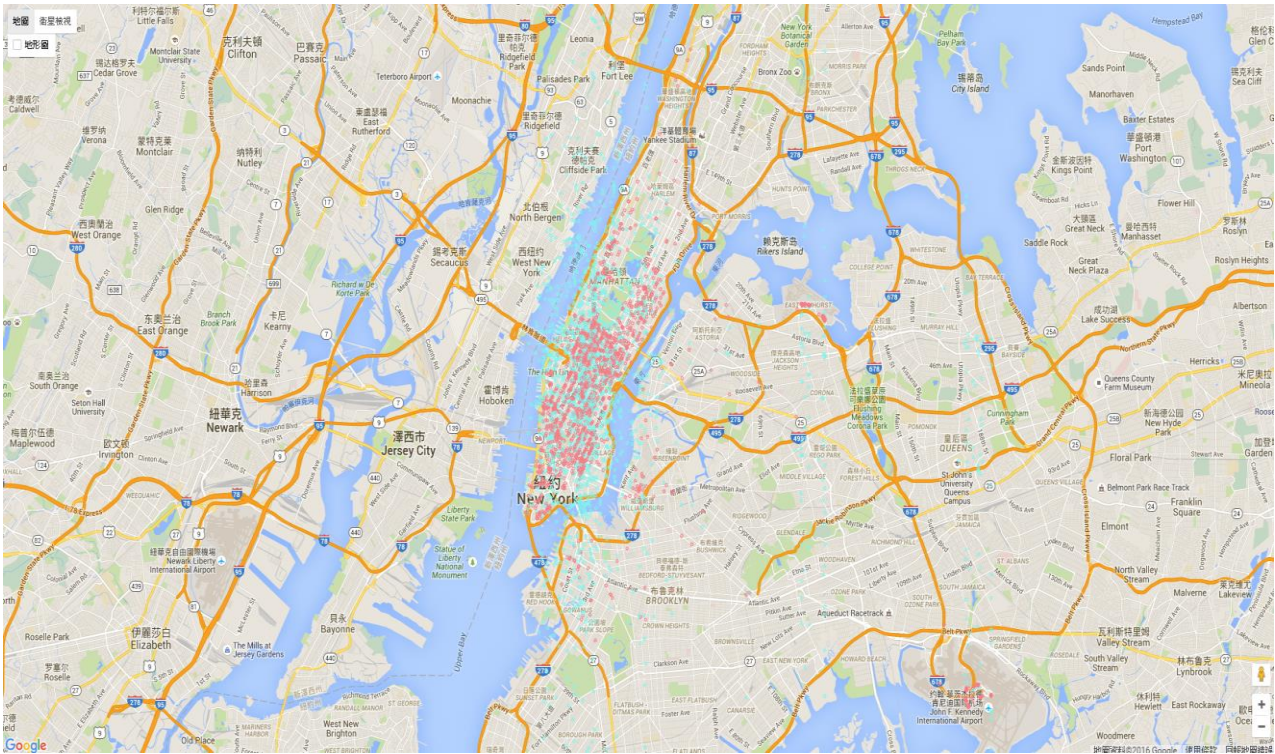
5. Speed

Few data have disproportionate trip distance and time. Like drive more than 1000 miles in 5 minutes.

➔ 67919947 – 93(Speed Noise) = 67919854 data points

Visualization:

Figure 1. Plot the location of drop and pick on Google Map



*Pink is pick and blue is drop

Figure 2. Plot the path on Google Map

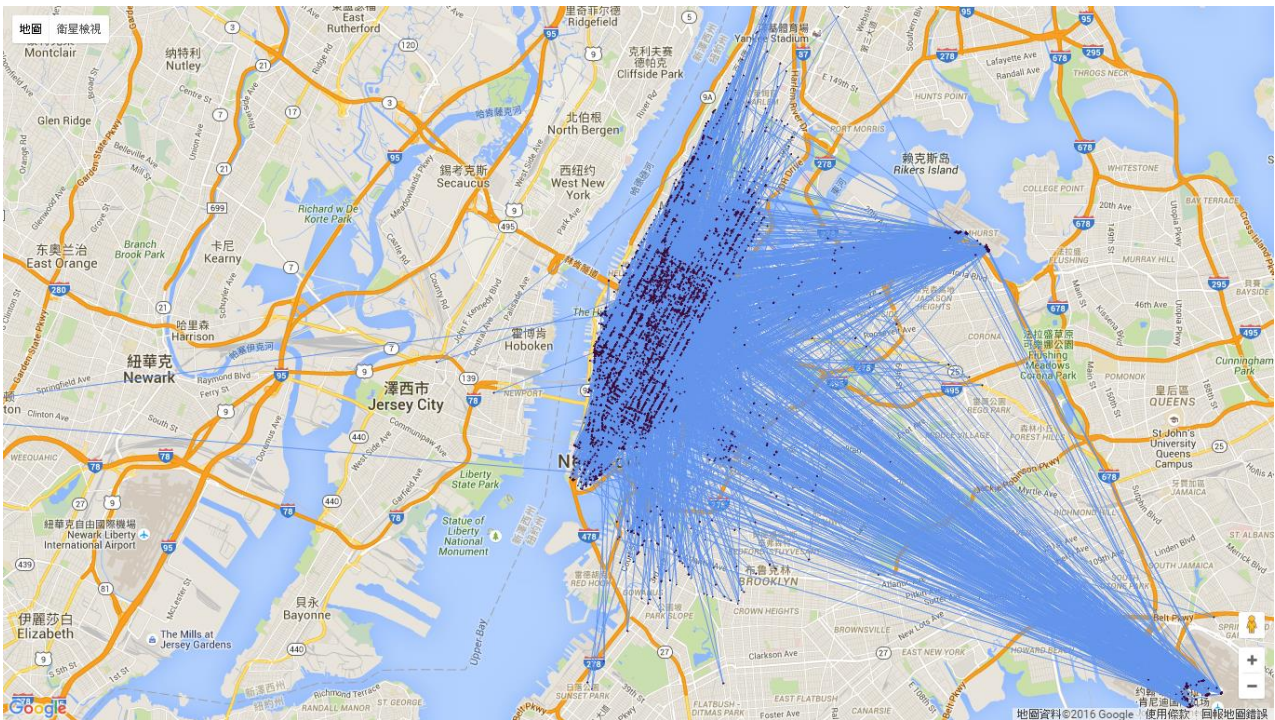
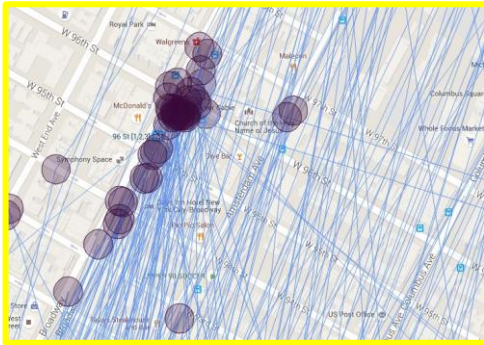
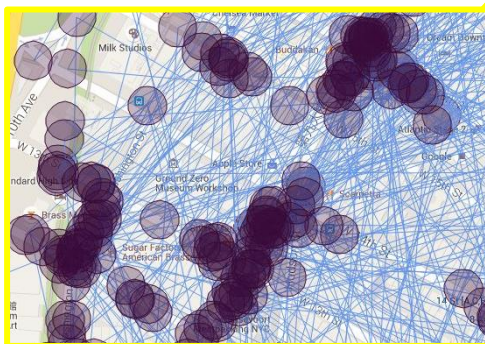
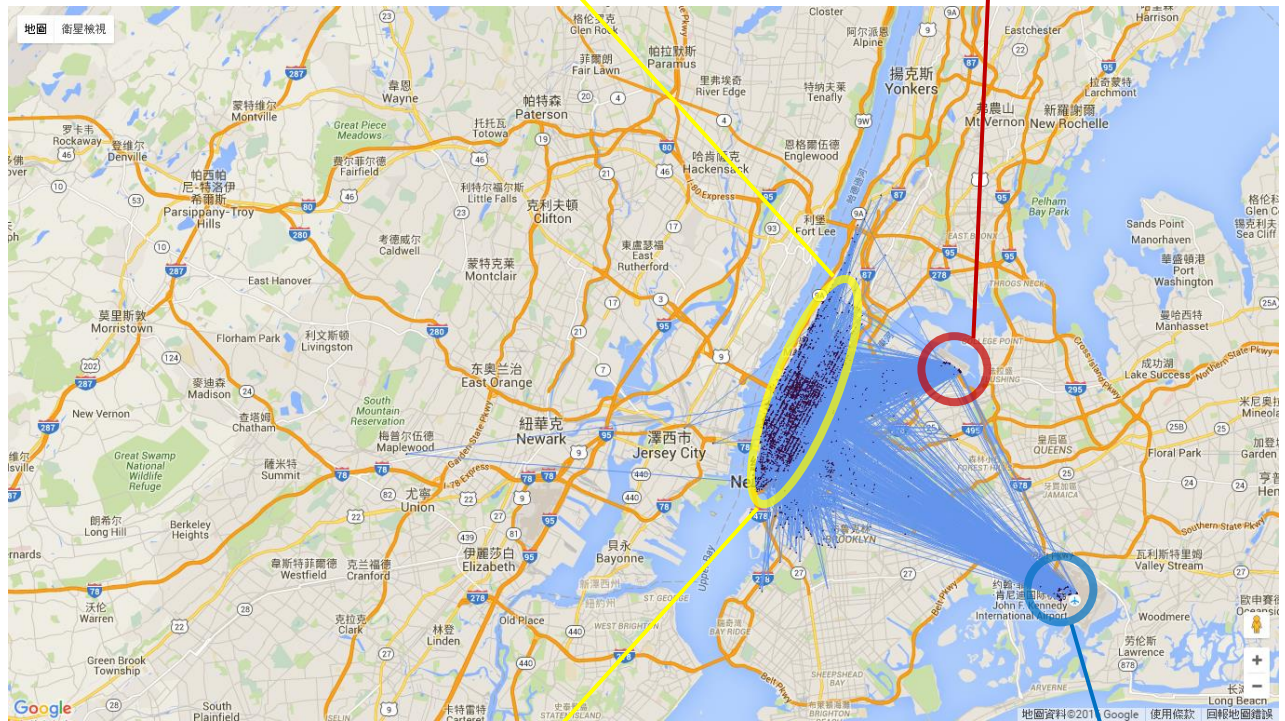
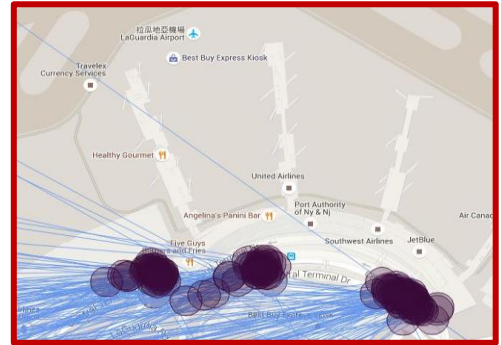


Figure 3. Some observations on crowded spots

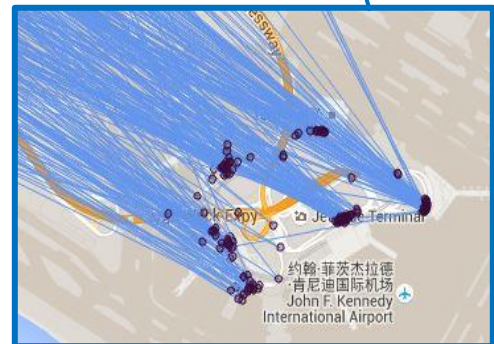
Car Rent Center



LaGuardia Airport



Apple Store, Google



John F. Kennedy International Airport

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

Thought: Group by the location of pickup and drop-off

Clustering Methods: Kmeans, MiniBatchKmeans, DBSCAN

Features: 1. pickup longitude and pickup latitude

2. drop-off longitude and drop-off latitude

Kmeans

Figure 4. Kmeans with 5 clusters (Pickup)

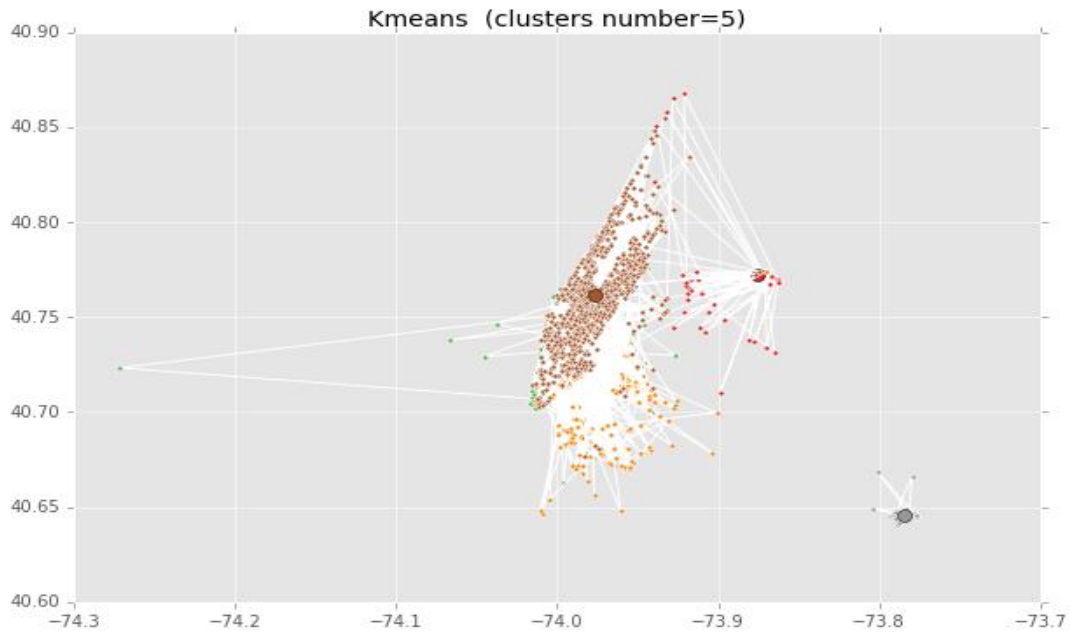


Figure 5. Plot the center above on map

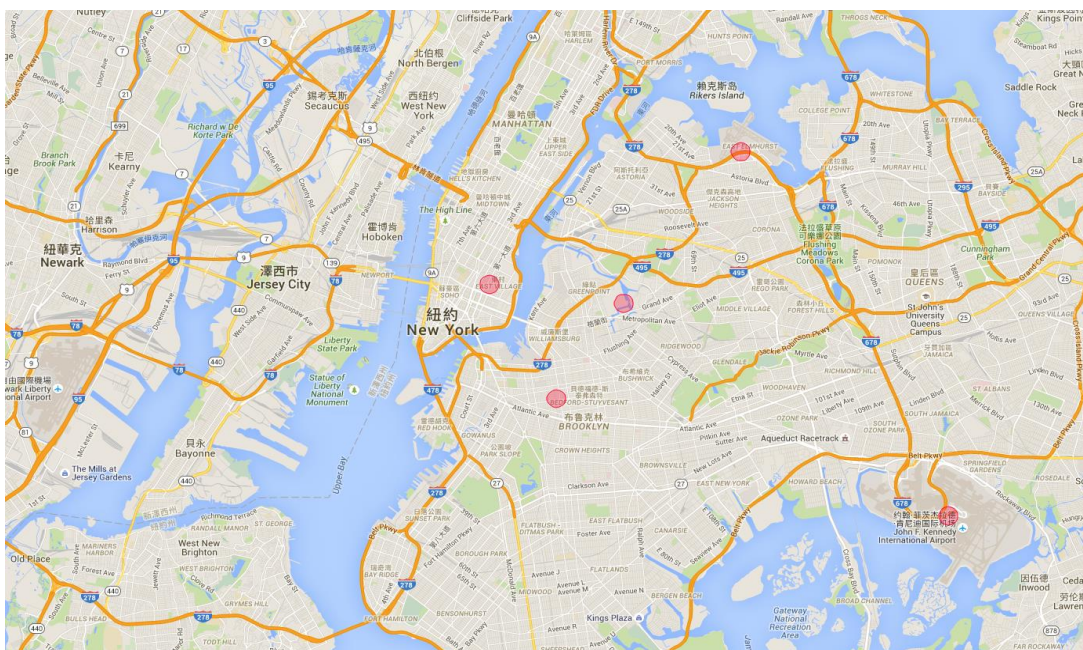
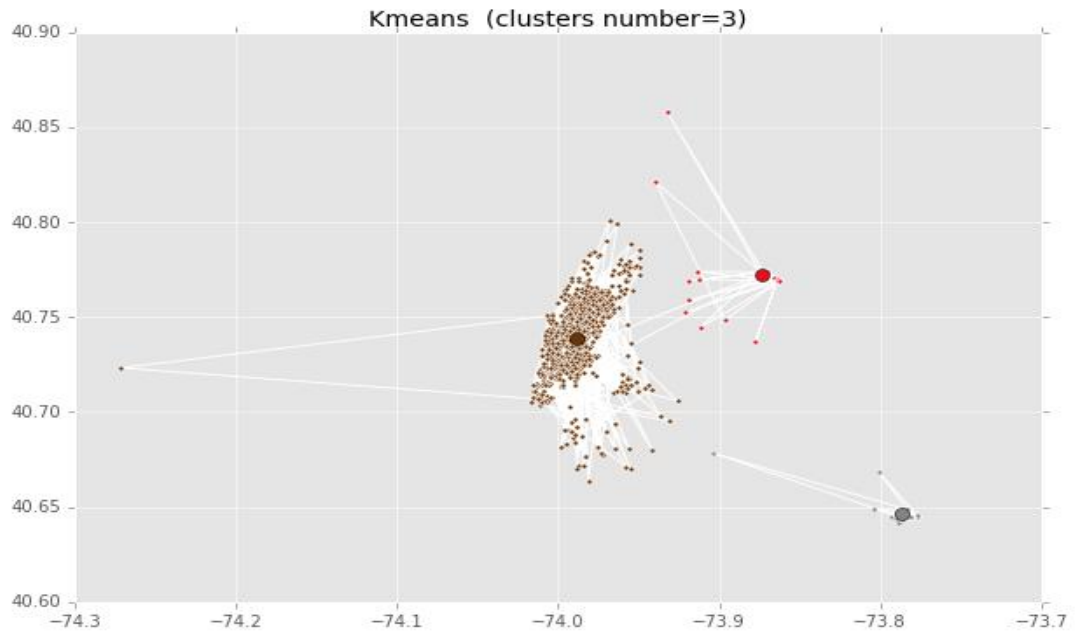
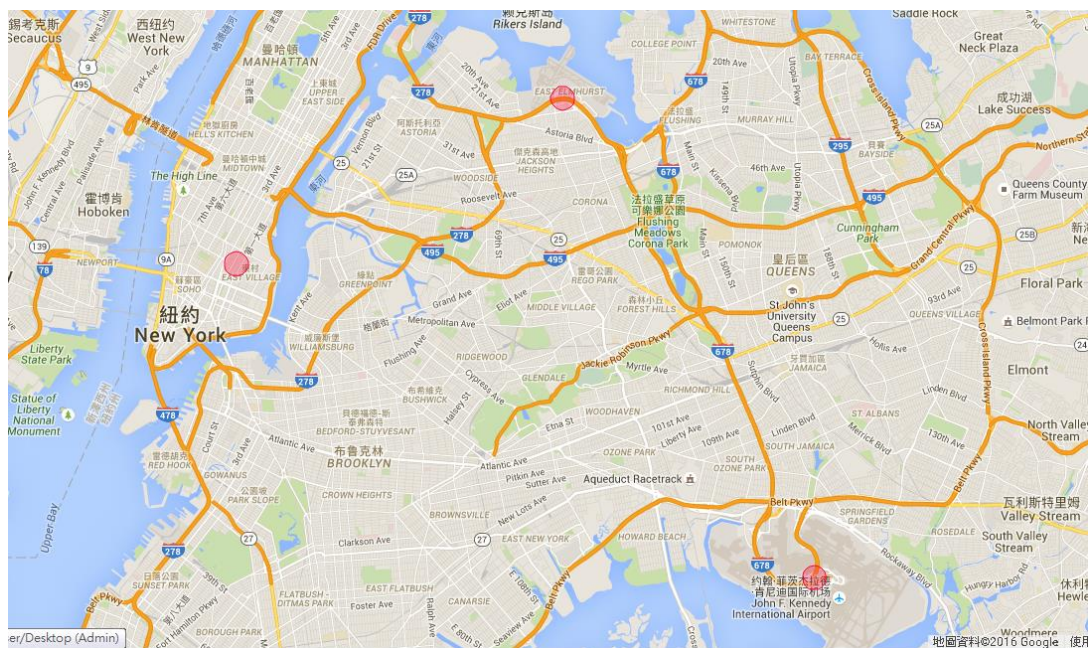


Figure 6. Kmeans with 3 clusters (Drop-off)



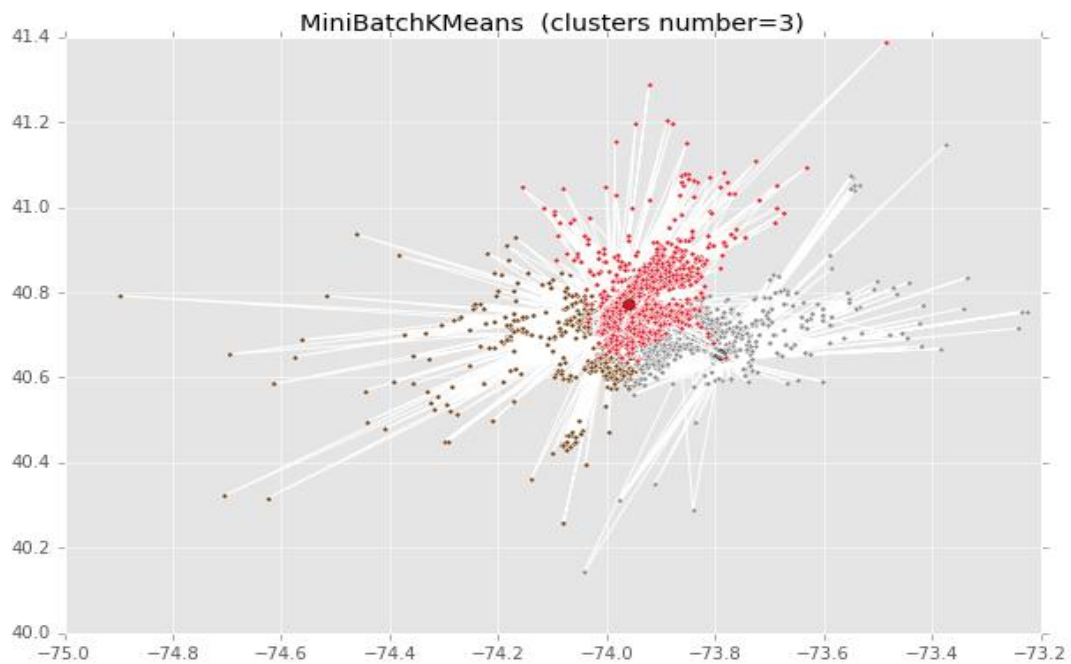
*Use 150000 data points because the exceeded cell block limit

Figure 7. Plot the center above on map



MiniBatchKmeans

Figure 8. MiniBatchKmeans with 3 clusters (Drop-off)



*Set batch size 45 and limit in 100 iterations

- ➔ Kmeans is easily be influenced by outlier, should be more carefully when data cleaning.
- ➔ Therefore, I try DBSCAN.

Figure 9. DBSCAN with eps=0.8 (Pickup)

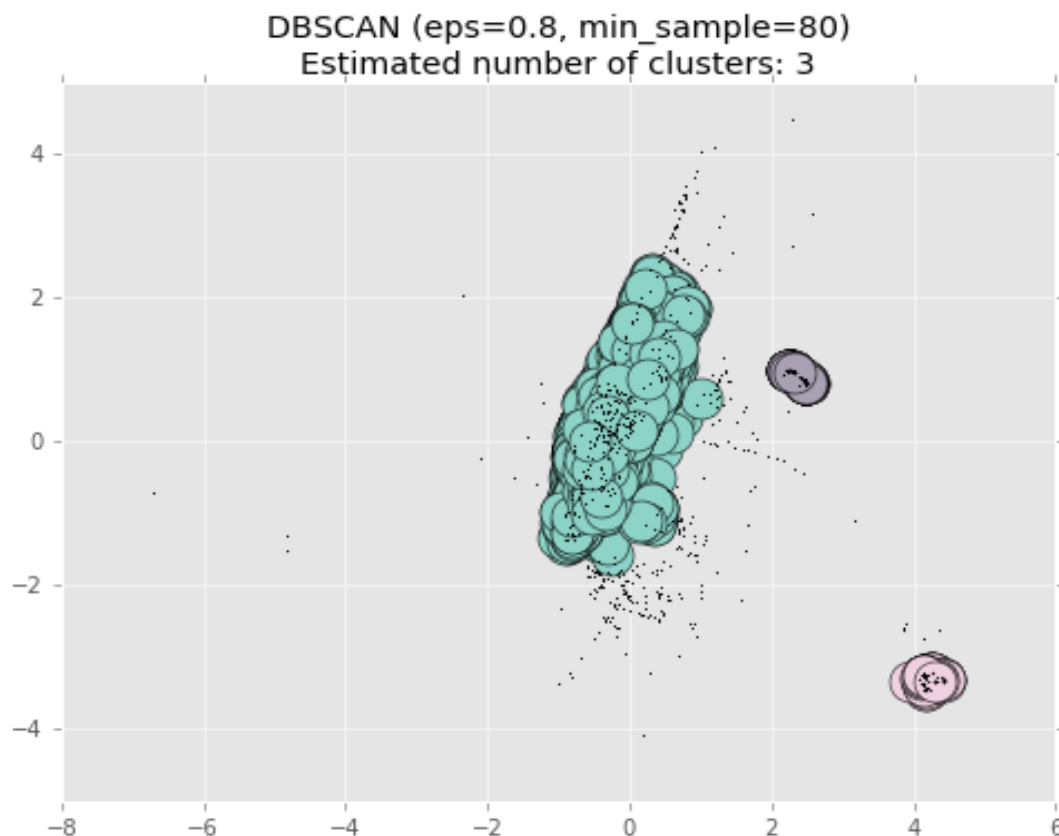
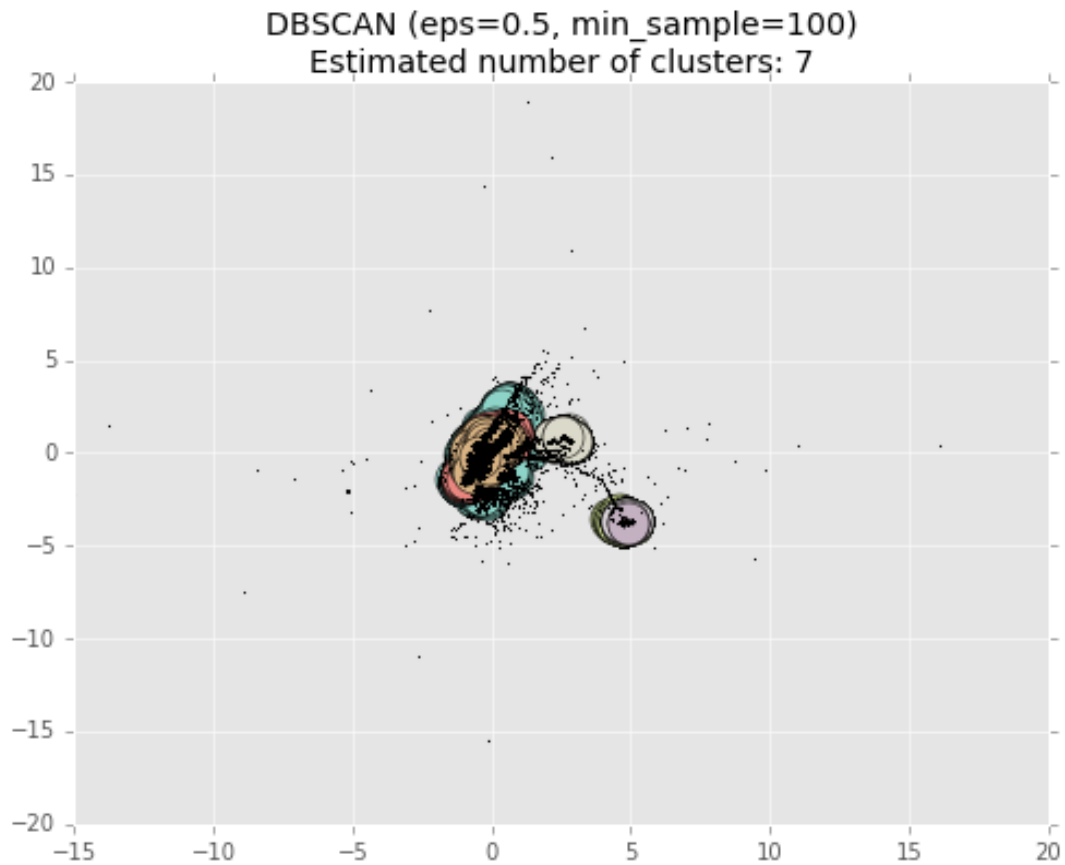
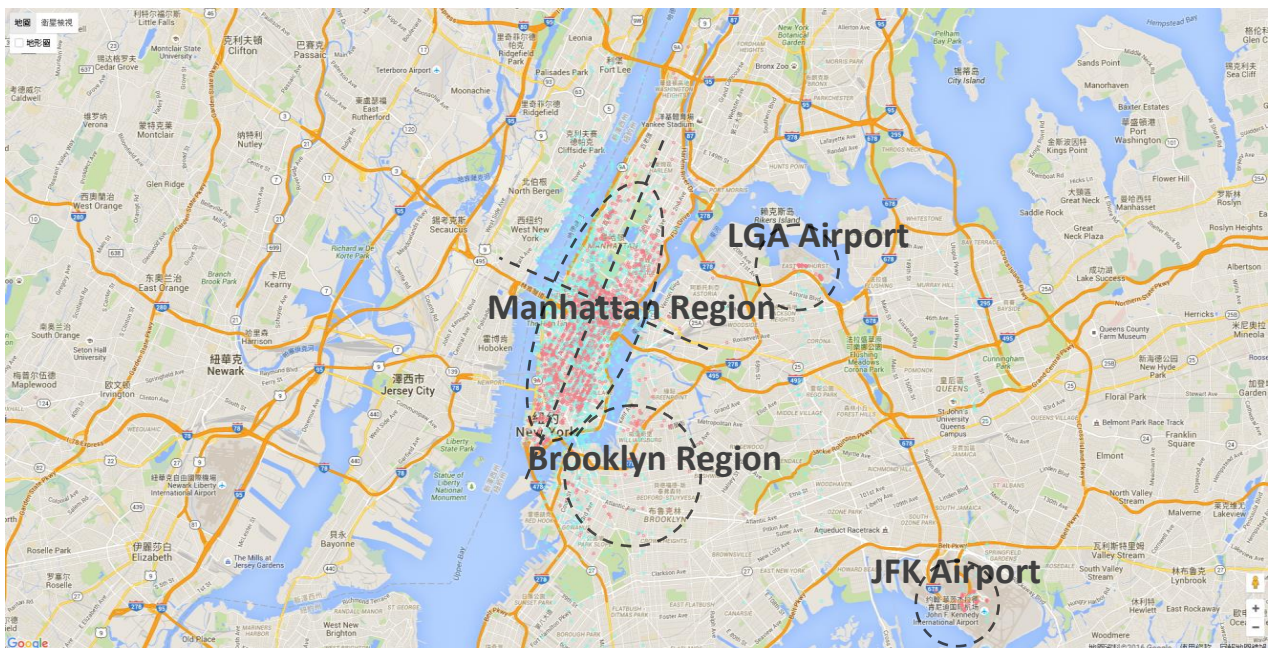


Figure 10. DBSCAN with eps=0.5 (Drop-off)



*Use 1500000 data points because the exceeded cell block limit

Figure 11. Clustering Summary Map



I define the region into four main parts: John F. Kennedy International Airport (LKF), LaGuardia Airport (LGA), Manhattan Region, and Brooklyn Region. The Manhattan Region is divided in to four parts: right up (RU.), left up (LP.), right down (RD.), left down (LD.).

Pickup Location Clustering


Three Centers	JFK	Manhattan 	LGA
Kmeans	27952297	38286723	1684204
MiniBatchKmeans	28537996	37700532	1684696

Table 12. Number of data points in three pickup location cluster


Five Centers	JFK	LD.	RD. 	Brooklyn	LGA
Kmeans	19238840	2121201	29487420	15427177	1648586
MiniBatchKmeans	20804979	2067721	28035428	15365061	1650032

Table 13. Number of data points in five pickup location cluster

When clustering into three cluster (Figure 4 and Figure 5), the most pickup location region is locate at **the Manhattan region**; when clustering into five cluster, the most pickup location region is locate at **the right down area of Manhattan region**. Some nodes belong to Brooklyn may be classify into Manhattan region when clustering into three center. Manhattan region has more frequent records and is strong enough to gather to be a cluster center.

Drop-off Location Clustering


Three Centers	JFK	Manhattan 	LGA
Kmeans	32472187	32409941	2741096
MiniBatchKmeans	25767428	41043162	1142634

Table 14. Number of data points in three drop-off location cluster


Five Centers	JFK	LU. 	LD.	RD.	LGA
Kmeans	16378785	33578110	14530379	974142	2461798
MiniBatchKmeans	1340871	29818073	19983930	1228308	3484202

Table 15. Number of data points in five drop-off location cluster

When clustering into three cluster (Figure 6 and Figure 7), the most drop-off location region is locate at **the Manhattan region**; when clustering into five cluster, the most drop-off location region is locate at **the left up area of Manhattan region**.

In Conclusion, the most pickups and drop-offs region is **Manhattan Region**. But the region cover over large area and sometimes the Manhattan region can separate into more clusters. Concerned about the density, the most pickups and drop-offs point is **John F. Kennedy International Airport**. It can always gather to a cluster center under different clustering conditions.

Q2: What is the best time to take taxi?

Thought: How to define Best Time?

1. Easy to Take Taxi (Few people take)
2. Cheaper
3. Faster

Easy to Take Taxi (Few people take)

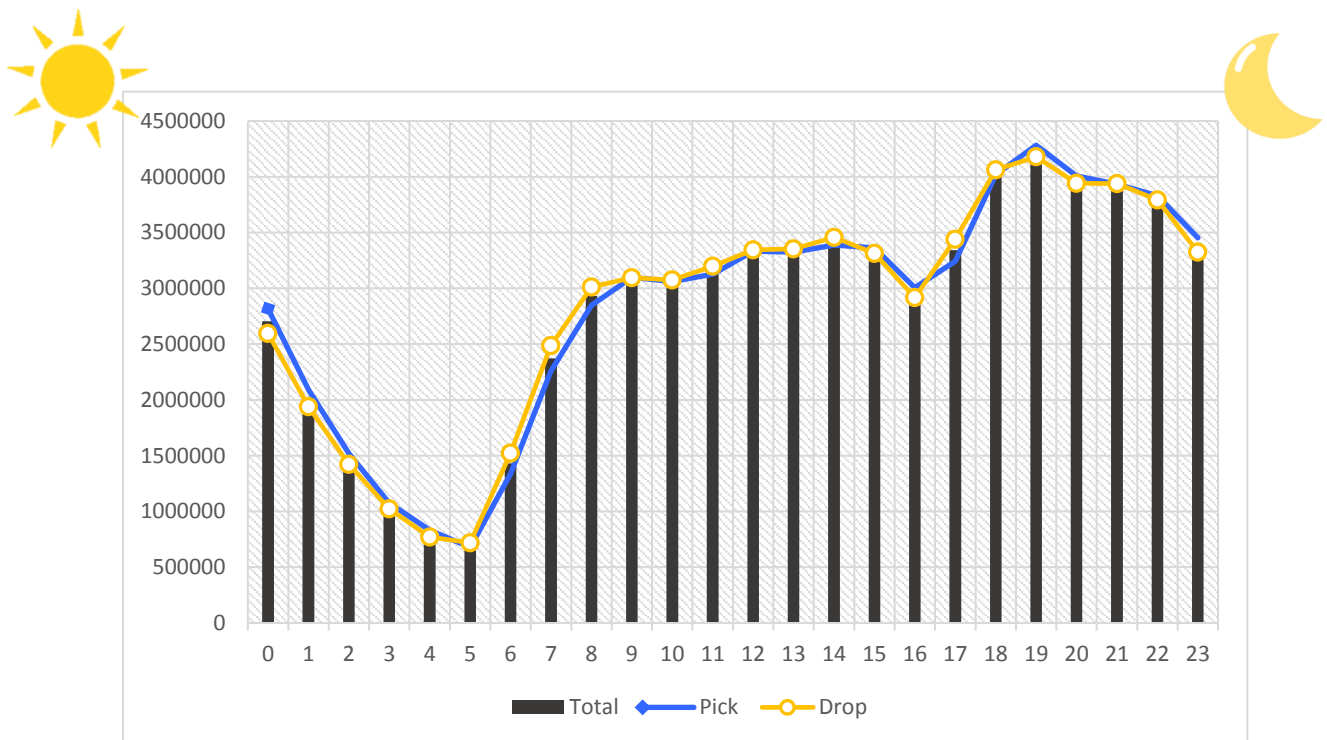


Chart 16. Count number of record by hour

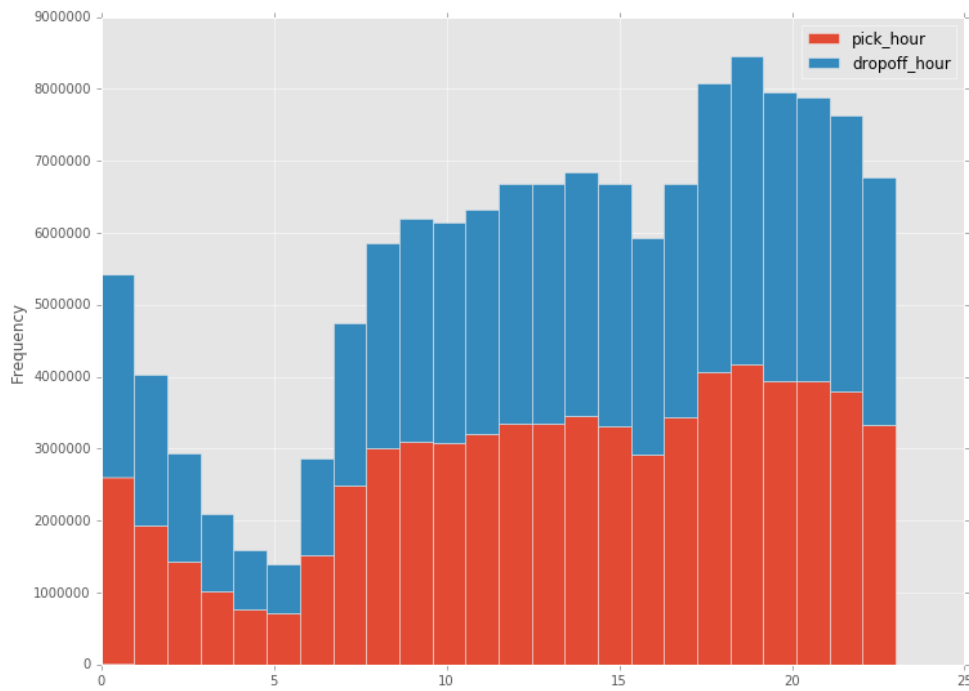


Chart 17. Stacked count number of record by hour

The trend of taxi taking frequency of pickup and drop-off hour looks similar. Both low at 4-5 o'clock and high at 18-20 o'clock.

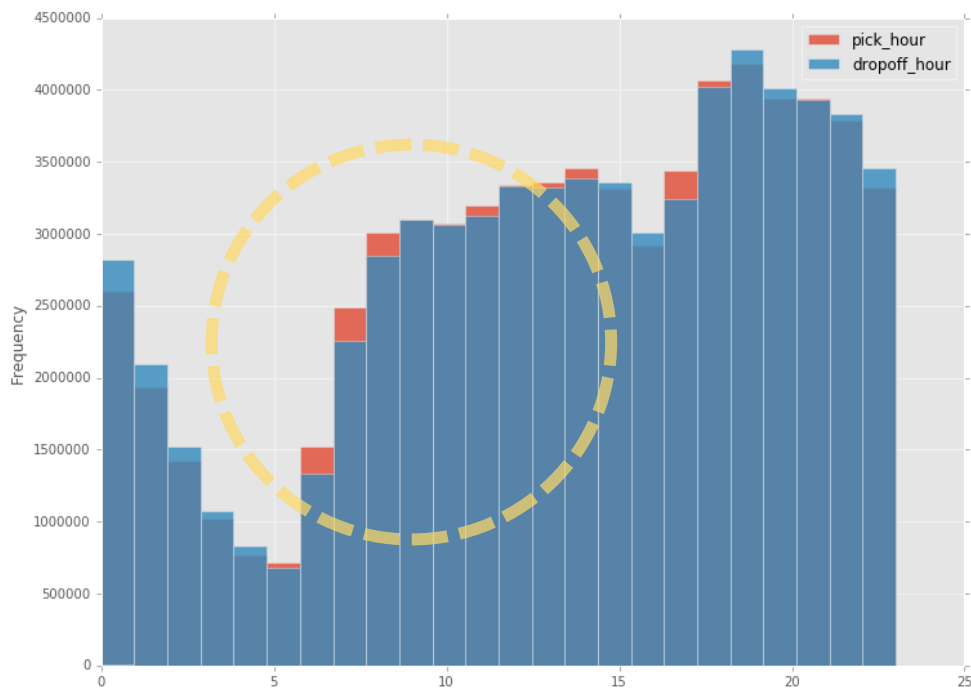


Chart 18. Overlapped count number of record by hour

The pickup mostly happen on the morning and drop-off mostly occur in evening when compared with each other.

Because the table is too big, I left only the **top 2** and **last 2**.

Hour	4	5	18	19
Record Number	770009	719121	4062481	4179552
Record Number	829894	678033	4018056	4282472

Table 19. Count number of record by hour

There are around 4200000 records at 19 o'clock and there are only 70000 records in 5 o'clock. It is about 6 time decrease! Therefore, if we want to call a taxi easily, we can choose **5 o'clock**.

Cheaper

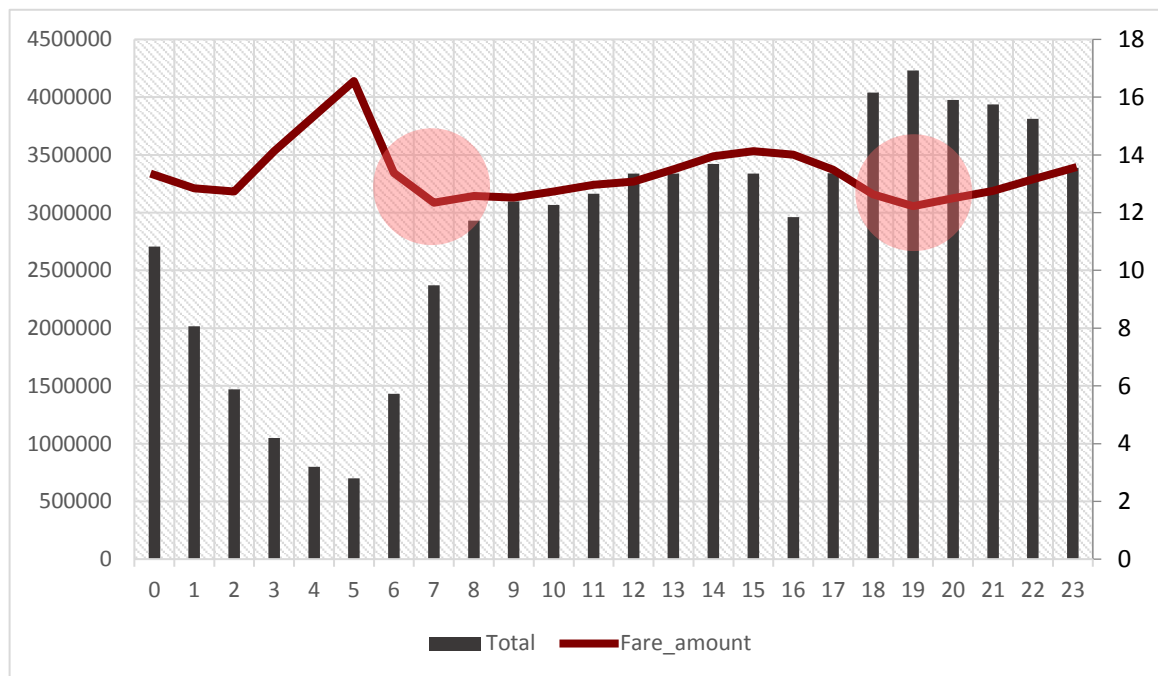


Chart 20. Fare amount by hour

If budget is our concerned, we can choose the lowest time-and-distance fare time interval. It is around **7 o'clock and 19 o'clock**.

Faster

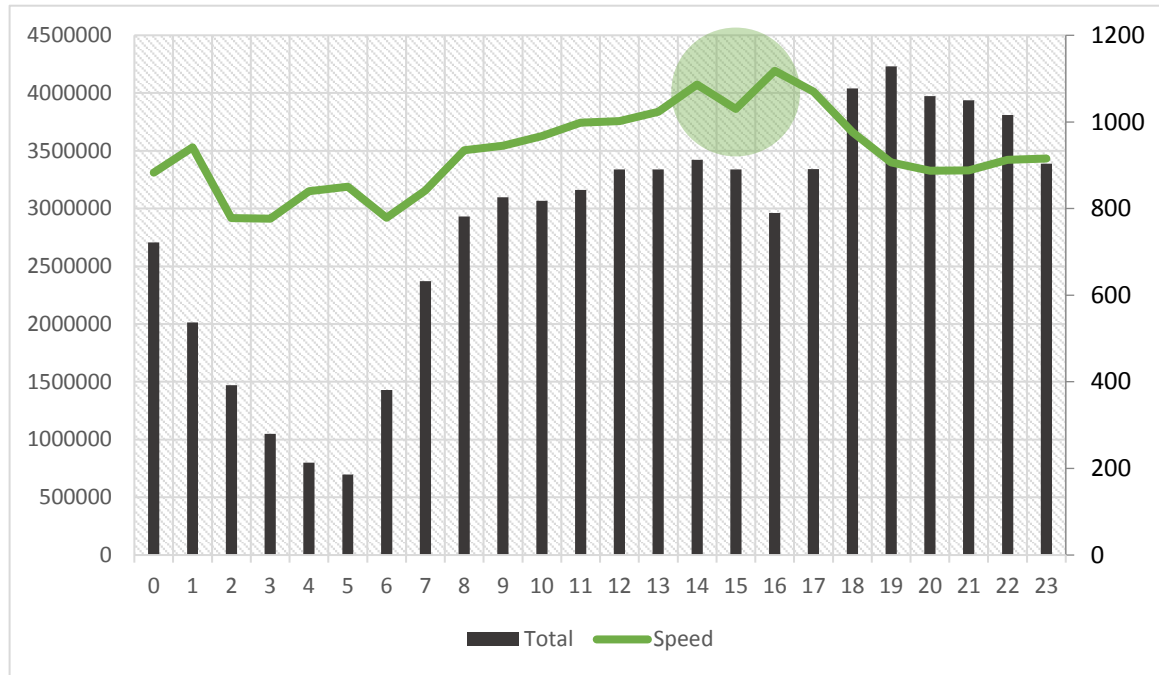


Chart 21. Speed by hour

If we are running out of time, we can choose the faster one. It is around **17 o'clock**.

For me, **I would choose take taxi at 16 o'clock as best time. Because it performs well on these three aspects.**

Q3: Whether weather affects customers to take taxi or not?

Thought: Count the record number and passenger by days and map to the weather

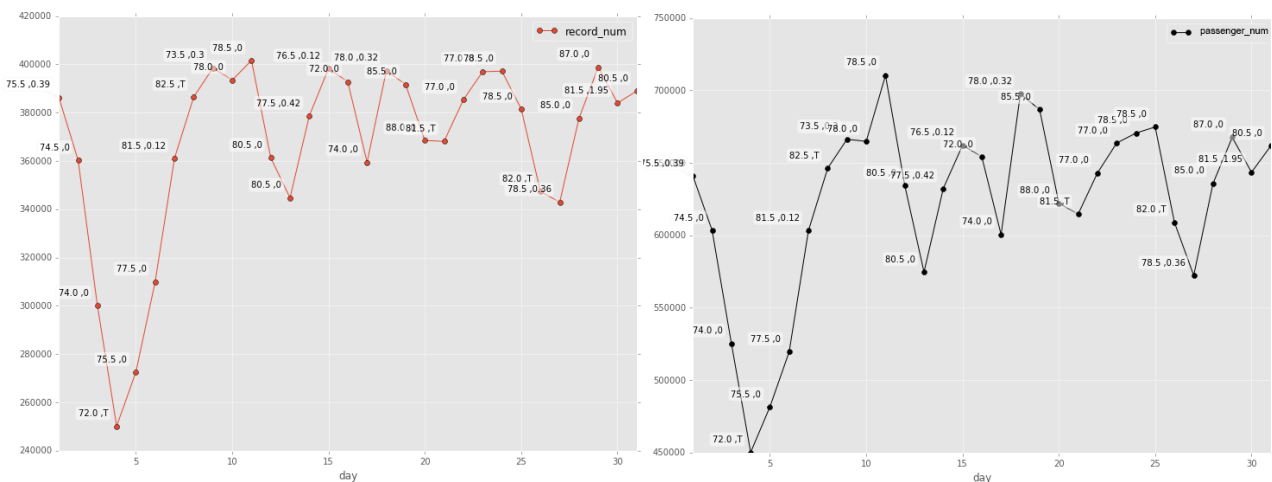


Chart 22. Count the record number and passenger by days

The relation seems not obvious. I find I need average temperature data which not provided.

Therefore, I go find data on [Weather Underground](http://www.wunderground.com/history/airport/KNYC/2015/7/1/CustomHistory.html?dayend=31&monthend=12&yearend=2015&req_city=&req_state=&req_statename=&reqdb.zip=&reqdb.magic=&reqdb.wmo=)

(http://www.wunderground.com/history/airport/KNYC/2015/7/1/CustomHistory.html?dayend=31&monthend=12&yearend=2015&req_city=&req_state=&req_statename=&reqdb.zip=&reqdb.magic=&reqdb.wmo=).

The Weather Underground provide high, low, and average of temperature, dew point, humidity, press, visibility, wind, precipitation, and summary.

Figure 23. Pearson correlation coefficient for record count and weather features

	Count	Month	Day	OempH	OemA	OemL	LuH	LuA	LuL	MoH	MoA	MoL	PrH	PrA	PrL	SeeH	SeeA	SeeL	WindH	WindA	WindL	PreSum	Summ
Count	1	0.067177	-0.14659	-0.13057	-0.1247	-0.11235	-0.14778	-0.17723	-0.16076	-0.15165	-0.13355	-0.09567	-0.17196	-0.19732	-0.21256	-0.07264	-0.00814	0.028175	0.247268	0.203196	0.252645	0.064092	-0.03778
Month	0.067177	1	-0.00543	-0.84157	-0.8468	-0.83956	-0.67081	-0.70013	-0.7128	0.121888	0.200348	0.239306	0.500743	0.365976	0.198778	-0.32689	-0.19192	-0.08038	0.244563	0.320933	0.190769	0.017663	0.037867
Day	-0.14659	-0.00543	1	-0.07847	-0.10677	-0.13107	-0.04984	-0.10164	-0.16788	-0.02003	-0.0343	-0.03707	0.212909	0.184378	0.142389	0.210487	0.070874	-0.04391	-0.00897	0.024687	0.014407	0.065283	0.030742
OempH	-0.13057	-0.84157	-0.07847	1	0.989894	0.959019	0.845612	0.867493	0.864797	0.007947	-0.10815	-0.19794	-0.57533	-0.45145	-0.29008	0.201057	0.110878	0.00895	-0.32163	-0.42709	-0.24981	-0.01103	-0.04489
OemA	-0.1247	-0.8468	-0.10677	0.989894	1	0.986941	0.866783	0.897819	0.901086	0.032064	-0.0588	-0.1347	-0.60408	-0.48213	-0.32383	0.198859	0.072191	-0.02954	-0.29388	-0.39858	-0.22347	0.013473	-0.01143
OemL	-0.11235	-0.83956	-0.13107	0.959019	0.986941	1	0.873515	0.913314	0.920589	0.070494	0.001037	-0.06636	-0.63537	-0.52038	-0.3691	0.183438	0.022462	-0.08234	-0.24907	-0.35595	-0.18116	0.04429	0.034715
LuH	-0.14778	-0.67081	-0.04984	0.845612	0.866783	0.873515	1	0.964093	0.885465	0.465846	0.371786	0.219943	-0.59298	-0.54713	-0.45653	0.144407	-0.21894	-0.37669	-0.19286	-0.25768	-0.09238	0.246414	0.304373
LuA	-0.17723	-0.70013	-0.10164	0.867493	0.897819	0.913314	0.964093	1	0.959893	0.406335	0.368177	0.266078	-0.57789	-0.51173	-0.4047	0.121943	-0.24051	-0.35654	-0.23688	-0.31872	-0.14602	0.248024	0.27856
LuL	-0.16076	-0.7128	-0.16788	0.864797	0.901086	0.920589	0.885465	0.959893	1	0.279874	0.294759	0.25372	-0.5445	-0.44888	-0.3289	0.090702	-0.19684	-0.26434	-0.28909	-0.36769	-0.21571	0.208282	0.171314
MoH	-0.15165	0.121888	-0.02003	0.007947	0.032064	0.070494	0.465846	0.406335	0.279874	1	0.896988	0.651598	-0.11234	-0.2205	-0.29605	-0.08194	-0.58897	-0.74856	0.097052	0.096052	0.154314	0.455318	0.64683
MoA	-0.13355	0.200348	-0.0343	-0.10815	-0.0588	0.001037	0.371786	0.368177	0.294759	0.896988	1	0.91919	-0.04793	-0.16326	-0.26077	-0.15158	-0.70682	-0.75299	0.093509	0.140123	0.145232	0.561132	0.663099
MoL	-0.09567	0.239306	-0.03707	-0.19794	-0.1347	-0.06636	0.219943	0.266078	0.25372	0.651598	0.91919	1	0.021207	-0.07989	-0.18019	-0.19004	-0.6921	-0.62733	0.072775	0.156691	0.110351	0.557438	0.560468
PrH	-0.17196	0.500743	0.212909	-0.57533	-0.60408	-0.63537	-0.59298	-0.57789	-0.5445	-0.11234	-0.04793	0.021207	1	0.952451	0.835659	-0.08081	0.040825	0.162491	0.011129	0.103035	-0.06913	-0.08335	-0.17731
PrA	-0.19732	0.365976	0.184378	-0.45145	-0.48213	-0.52038	-0.54713	-0.51173	-0.44888	-0.2205	-0.16326	-0.07989	0.952451	1	0.954357	-0.06166	0.164029	0.271732	-0.09831	-0.02328	-0.21125	-0.18619	-0.29495
PrL	-0.21256	0.198778	0.142389	-0.29008	-0.32383	-0.3691	-0.45653	-0.4047	-0.3289	-0.29605	-0.26077	-0.18019	0.835659	0.954357	1	-0.02728	0.271733	0.354706	-0.18994	-0.13289	-0.31756	-0.26079	-0.37173
SeeH	-0.07264	-0.32689	0.210487	0.201057	0.198859	0.183438	0.144407	0.121943	0.090702	-0.08194	-0.15158	-0.19004	-0.08081	-0.06166	-0.02728	1	0.370716	0.120699	0.124369	0.078327	0.148214	-0.14073	0.032864
SeeA	-0.00814	-0.19192	0.070874	0.110878	0.072191	0.022462	-0.21894	-0.24051	-0.19684	-0.58897	-0.70682	-0.6921	0.040825	0.164029	0.271733	0.370716	1	0.792899	-0.1229	-0.11675	-0.1579	-0.69288	-0.64366
SeeL	0.028175	-0.08038	-0.04391	0.00895	-0.02954	-0.08234	-0.37669	-0.35654	-0.26434	-0.74856	-0.75299	-0.62733	0.162491	0.271732	0.354706	0.120699	0.792899	1	-0.19562	-0.14449	-0.24983	-0.59474	-0.80416
WindH	0.247268	0.244563	-0.00897	-0.32163	-0.29388	-0.24907	-0.19286	-0.23688	-0.28909	0.097052	0.093509	0.072775	0.011129	-0.09831	-0.18994	0.124369	-0.1229	-0.19562	1	0.826565	0.895566	0.179569	0.260748
WindA	0.203196	0.320933	0.024687	-0.42709	-0.39858	-0.35595	-0.25768	-0.31872	-0.36769	0.096052	0.140123	0.156691	0.103035	-0.02328	-0.13289	0.078327	-0.11675	-0.14449	0.826565	1	0.780109	0.214264	0.247414
WindL	0.252645	0.190769	0.014407	-0.24981	-0.22347	-0.18116	-0.09238	-0.14602	-0.21571	0.154314	0.145232	0.110351	-0.06913	-0.21125	-0.31756	0.148214	-0.1579	-0.24983	0.895566	0.780109	1	0.224623	0.343334
PreSum	0.064092	0.017663	0.065283	-0.01103	0.013473	0.04429	0.246414	0.248024	0.208282	0.455318	0.561132	0.557438	-0.08335	-0.18619	-0.26079	-0.14073	-0.69288	-0.59474	0.179569	0.214264	0.224623	1	0.640592
Summ	-0.03778	0.037867	0.030742	-0.04489	-0.01143	0.034715	0.304373	0.27856	0.171314	0.64683	0.663099	0.560468	-0.17731	-0.29495	-0.37173	0.032864	-0.64366	-0.80416	0.260748	0.247414	0.343334	0.640592	1

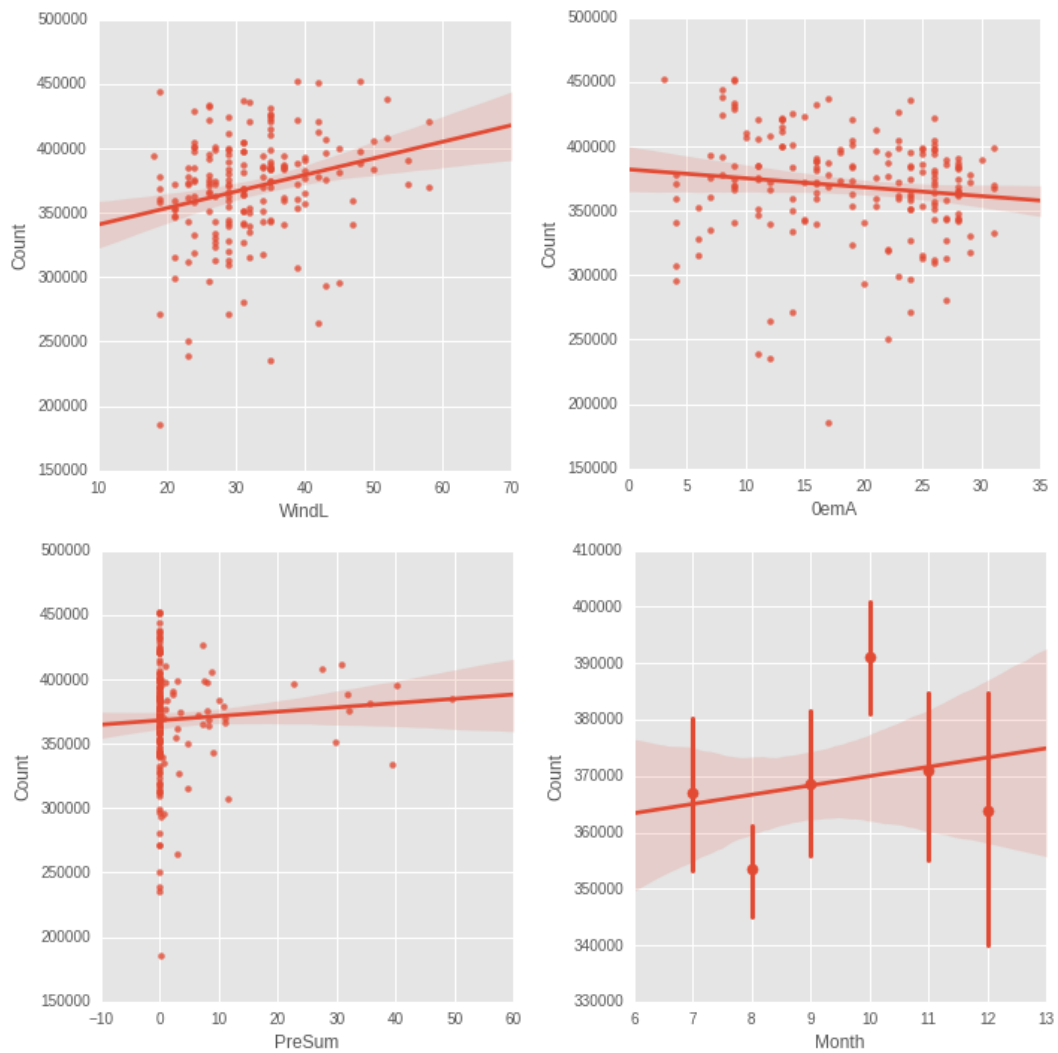
<http://pandas.pydata.org/pandas-docs/version/0.9.0/computation.html>

Month	Day	TempH	TemA	TemL	LuH	LuA	LuL	MoH	MoA	MoL
0.067177	-0.14659	-0.13057	-0.1247	-0.11235	-0.14778	-0.17723	-0.16076	-0.15165	-0.13355	-0.09567
PrH	PrA	PrL	SeeH	SeeA	SeeL	WindH	WindA	WindL	PreSum	Summ
-0.17196	-0.19732	-0.21256	-0.07264	-0.00814	0.028175	0.247268	0.203196	0.252645	0.064092	-0.03778

Table 24. Relation of record count and weather features

I think the weather seems has just little relation with customer take taxi or not. The most relate one is the wind (Charts on next page). The **bigger wind** cause the more taxi records; and also, when winter coming people have more chance to take a taxi. Maybe because the weather is cold, people want to take a taxi.

But the relation of taking taxi and temperature seems not obvious as I thought before.



**Chart 25. Relations between taking a taxi or not (a) Wind
(b) Temperature (c) Precipitation (d) Summary)**

Q4: Does long distance trip imply more tip?

At first the correlation between distance and tip I count is 0.000003. Then, I remove the noise records where tips is negative and disproportionate trip distance and time. I get **0.730463**. The tip seems have strong relation with trip distance.

Additional:

Night additional taxi fare

Not only distance, we may want to know the relation between tip amount and the pickup/drop-off time and the duration of taking taxi.

	trip_distance	tip_amount	total_amount	fare_amount	pick_hour	dropoff_hour	duration
trip_distance	1.000	0.730	0.919	0.921	-0.016	-0.022	0.019
tip_amount	0.730	1.000	0.852	0.775	0.007	0.006	0.015
total_amount	0.919	0.852	1.000	0.987	0.008	0.006	0.020
fare_amount	0.921	0.775	0.987	1.000	-0.002	-0.004	0.020
pick_hour	-0.016	0.007	0.008	-0.002	1.000	0.929	0.001
dropoff_hour	-0.022	0.006	0.006	-0.004	0.929	1.000	0.001
duration	0.019	0.015	0.020	0.020	0.001	0.001	1.000

Table 26. The relation of money and time features

I especially curious about whether there have the night additional taxi fare in New York or not. Therefore, count the mean fare amount by hours. It meets the lowest fare at 19'o clock and **highest fare at 5'o clock**. As mentioned before, although 5 o'clock has least passenger. The fare is also the highest.

I think there has night additional taxi fare as shown by chart 24.

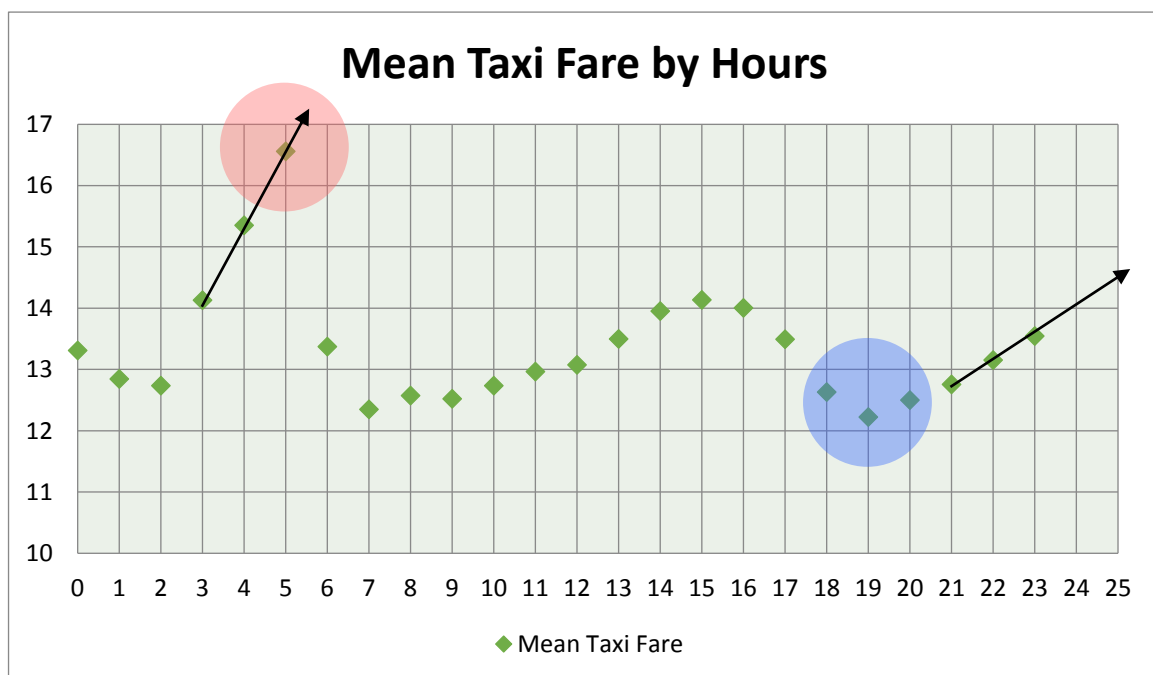
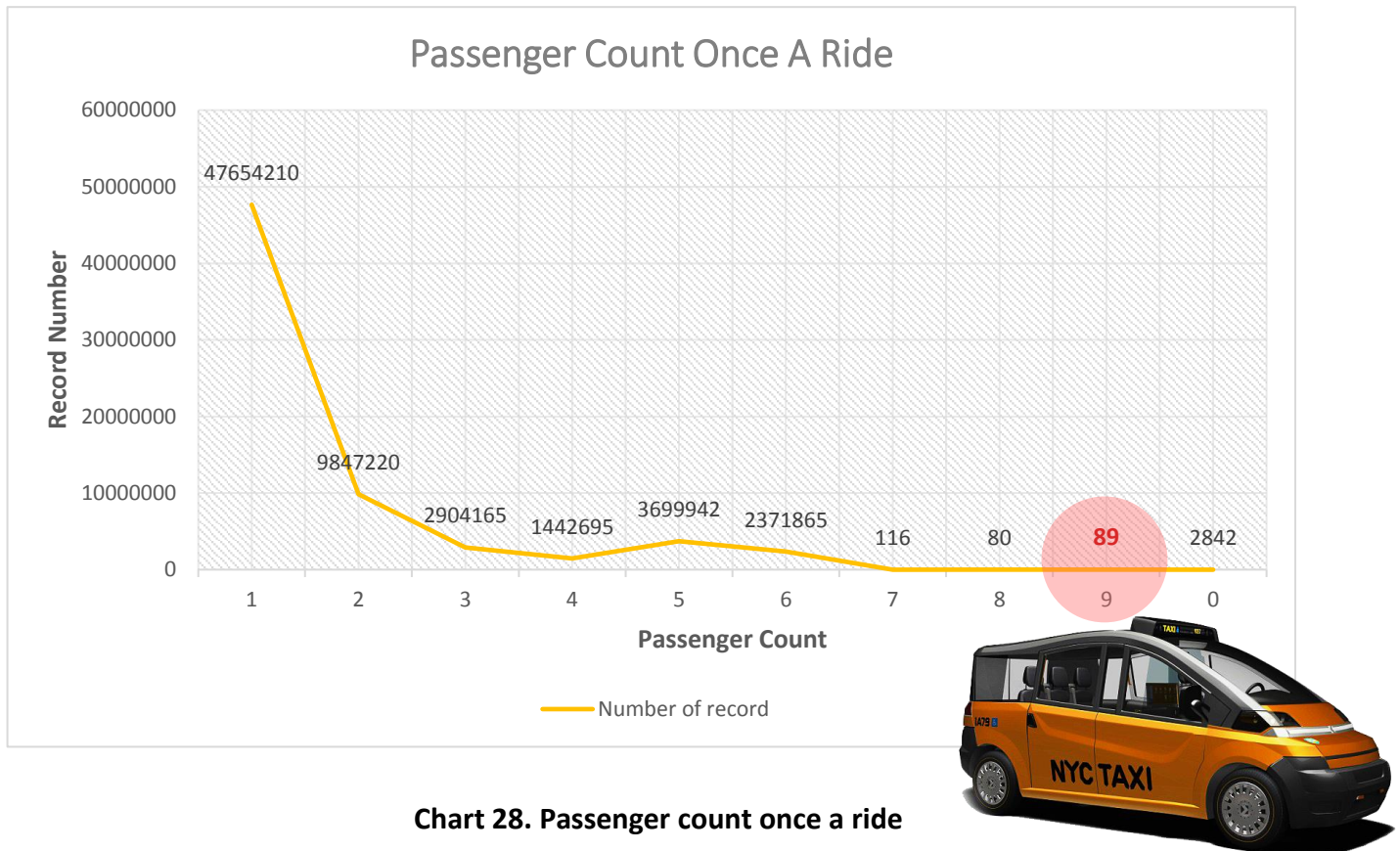


Chart 27. Relations between taxi fare and time

Passenger Count

When counting the passenger number (chart 25), there are 89 records take 9 passengers at once. At first I thought those records may be input error.



And subsequently learned that there are nine seats taxi in USA, looks like the picture above :)