SUSTech Southern University of Science and Technology

Taxi data Analysis Project

Group member: 11811127周子越
11811105焦点
11811115孙翊航
11811509任博韬
11810935 卢斓

2021/05/28

SUSTech Southern University of Science and Technology

Table of Content:

SUSTech Southern University of Science and Technology

Data Understanding:

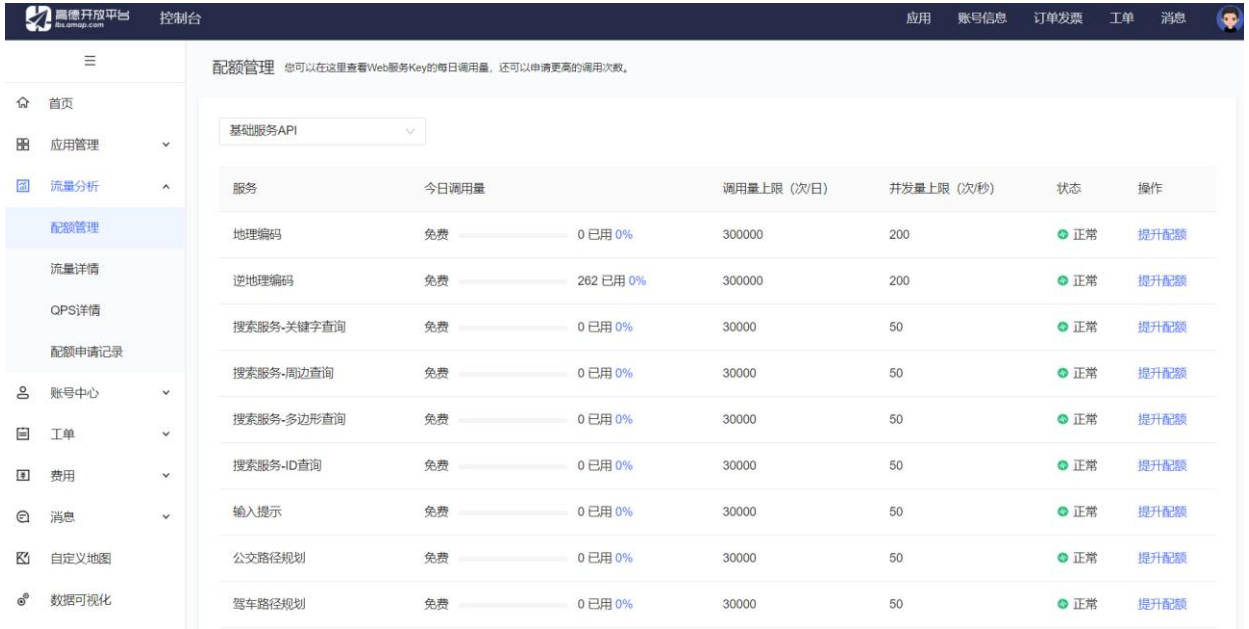
1. What is the Data we are using?
   1) The basic data provided by the teacher -- sample_taxi.csv

	A	B	C	D	E	F	G
1	taxi_id	time	lon	lat	is_passeng	speed	
2	22224	10:13:42	113.887	22.54752	0	0	
3	22224	19:31:36	113.9096	22.5822	1	71	
4	22224	17:06:32	113.9167	22.62355	1	100	
5	22224	15:07:19	113.9224	22.53538	1	21	
6	22224	14:55:18	113.9227	22.52003	1	10	
7	22224	14:55:03	113.9229	22.52053	1	18	
8	22224	10:25:57	113.9241	22.55533	1	60	
9	22224	14:52:18	113.9288	22.5248	1	26	
10	22224	6:52:26	113.929	22.55522	1	78	
11	22224	15:17:43	113.9379	22.54263	1	81	
12	22224	15:18:13	113.9424	22.54272	1	40	
13	22224	15:31:09	113.9454	22.53005	0	12	
14	22224	10:33:42	113.9607	22.56617	1	16	
15	22224	10:41:52	113.9613	22.56227	1	61	
16	22224	19:41:13	113.9619	22.55673	1	58	
17	22224	15:39:55	113.9666	22.57172	1	73	
18	22224	19:42:13	113.9667	22.55747	1	27	
19	22224	15:40:55	113.9673	22.58163	1	82	

SUSTech Southern University of Science and Technology

Data Understanding:

1. What is the Data we are using?
   2) Gaode API



- **返回结果参数说明**

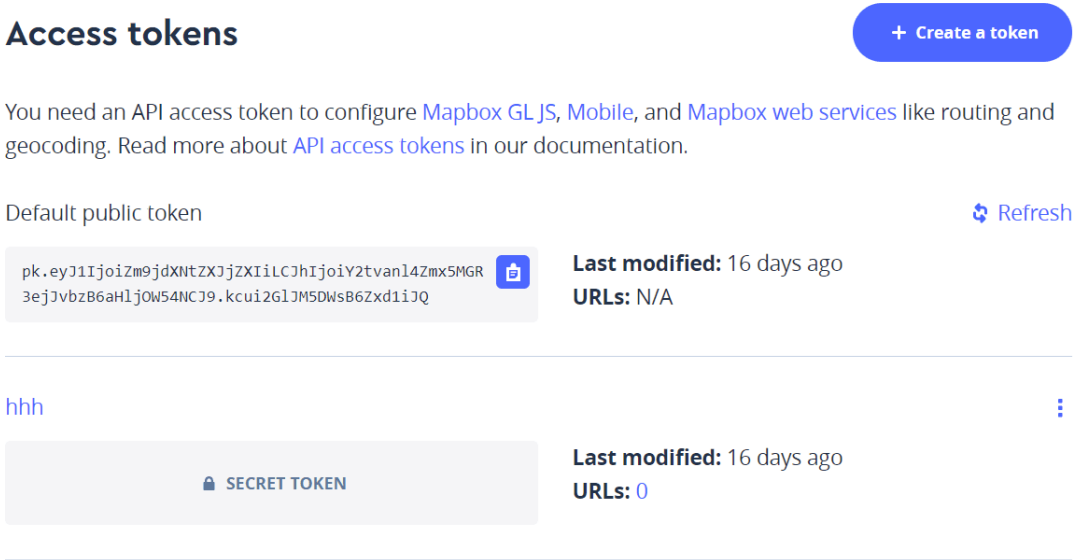
逆地理编码的响应结果的格式由请求参数output指定。

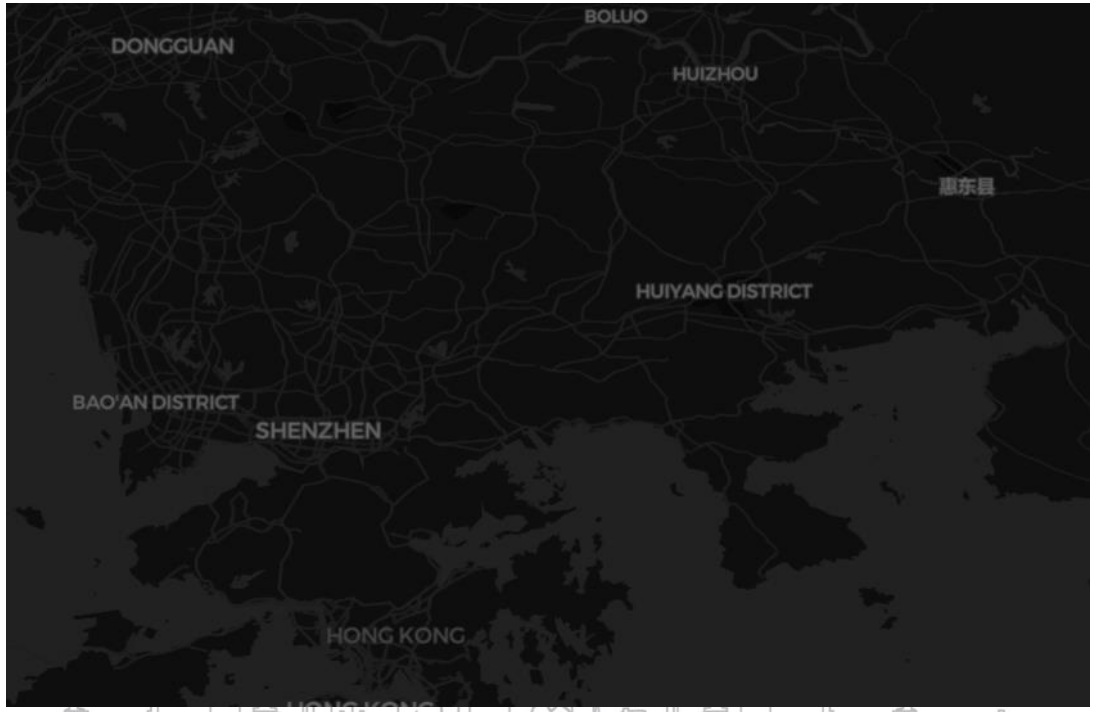
名称	含义	规则说明
status	返回结果状态值	返回值为 0 或 1，0 表示请求失败；1 表示请求成功。
info	返回状态说明	当 status 为 0 时，info 会返回具体错误原因，否则返回"OK"。详情可以参考 info状态表
regeocodes	逆地理编码列表	batch 字段设置为 true 时为批量请求，此时 regeocodes 标签返回，标签下为 regeocode 对象列表；batch 为false 时为单个请求，会返回 regeocode 对象；regeocode 对象包含的数据如下：
formatted_address	结构化地址信息	结构化地址信息包括：省份＋城市＋区县＋城镇＋乡村＋街道＋门牌号码 如果坐标点处于海域范围内，则结构化地址信息为：省份＋城市＋区县＋海域信息
addressComponent	地址元素列表	
province	坐标点所在省名称	例如：北京市
city	坐标点所在城市名称	请注意：当城市是省直辖县时返回为空，以及城市为北京、上海、天津、重庆四个直辖市时，该字段返回为空；省直辖县列表
citycode	城市编码	例如：010
district	坐标点所在区	例如：海淀区
adcode	行政区编码	例如：110108
township	坐标点所在乡镇/街道（此街道为社区街道，不是道路信息）	例如：燕园街道
towncode	乡镇街道编码	例如：110101001000

SUSTech Southern University of Science and Technology

Data Understanding:

1. What is the Data we are using?
   3) Mapbox map data





SUSTech Southern University of Science and Technology

Data Understanding:

2. What can we do with these data?

1) **sample_taxi.csv:**

The base of all analysis. After observing the data, we found that we can do the following things:

Path dynamic visualization
Popular area analysis with OD data
Seated rate over different time
Traffic trend direction analysis
Car speed analysis &Traffic jam analysis
Peak Observation and Prediction
Taxi utilization analysis
Standard for price

2) **Gaode api**

Can be used for reverse geocoding and be used popular area analysis

3) **Mapbox data**

Can be used as the base map layer of data visualization

Data Preprocessing:

1. Data Cleaning
   1) Drop data which missing from key attributes
   2) Fill the blank data
   3) Drop the data outside ShenZhen area

2. Data format conversion
   1) Combine “lat” and ”lon” as “coordinate”
   2) Convert “time” to “timestamp”
   3) Generate data in OD format
   4) Data format conversion to .json

SUSTech Southern University of Science and Technology

Data Preprocessing:

1	taxi_id	time	is_passeng	speed	coordinates	timestamp
2	22224	0:02:42	0	1	[[114.03508, 22.555201]]	[162]
3	22224	0:03:12	0	2	[[114.035164, 22.555]]	[192]
4	22224	0:06:04	0	3	[[114.031998, 22.54875]]	[364]
5	22224	0:07:30	0	9	[[114.032204, 22.547667]]	[450]
6	22224	0:07:32	0	9	[[114.032219, 22.547617]]	[452]
7	22224	0:07:33	0	6	[[114.032234, 22.547583]]	[453]
8	22224	0:09:03	0	1	[[114.034248, 22.547882]]	[543]
9	22224	0:12:40	0	4	[[114.042084, 22.553534]]	[760]
10	22224	0:14:26	1	1	[[114.039886, 22.553232]]	[866]
11	22224	0:16:30	1	1	[[114.033531, 22.553284]]	[990]
12	22224	0:16:49	1	15	[[114.033546, 22.553516]]	[1009]
13	22224	0:18:49	1	12	[[114.025284, 22.562782]]	[1129]
14	22224	0:20:19	1	13	[[114.020081, 22.560667]]	[1219]
15	22224	0:20:35	0	3	[[114.019218, 22.560617]]	[1235]
16	22224	0:21:20	0	13	[[114.019165, 22.560667]]	[1280]
17	22224	0:37:56	1	14	[[114.122452, 22.579384]]	[2276]

Coordinate and timestamp

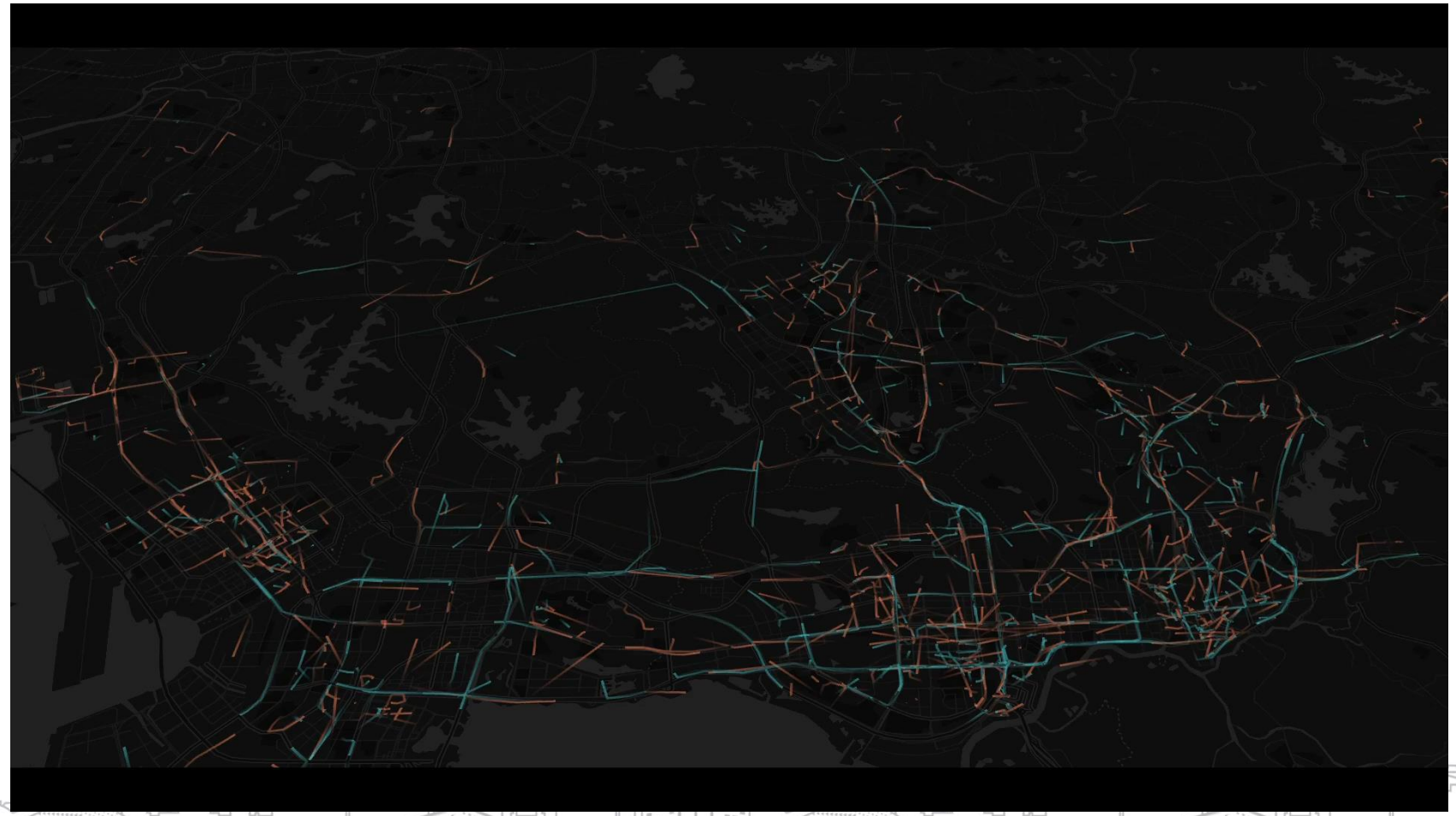
1	taxi_id	start_tim	start_lor	start_lat	end_time	end_lon	end_lat	distance
2	22224	208	114.0343	22.55412	364	114.032	22.54875	0.007103
3	22224	836	114.0404	22.55143	1235	114.0192	22.56062	0.030357
4	22224	1506	114.0254	22.55575	2492	114.121	22.58148	0.124452
5	22224	3791	114.1032	22.54925	4098	114.108	22.56612	0.021702
6	22224	4684	114.1053	22.55185	5711	114.0286	22.55982	0.12011
7	22224	23249	114.1006	22.61278	25212	113.887	22.54742	0.278347
8	22224	37346	113.9051	22.55498	38188	113.9664	22.5663	0.082478
9	22224	38203	113.9662	22.56642	39407	114.0451	22.5257	0.126611
10	22224	39468	114.0464	22.52577	40250	114.0419	22.51297	0.029636
11	22224	40284	114.0414	22.51323	40581	114.0487	22.5235	0.024266
12	22224	40660	114.0498	22.5258	41144	114.0615	22.5498	0.035765
13	22224	41453	114.0718	22.54145	41852	114.0646	22.53017	0.024919
14	22224	42288	114.0749	22.5346	43008	114.084	22.54478	0.028103
15	22224	43053	114.0836	22.54518	44683	114.0547	22.62522	0.132515
16	22224	47541	114.1023	22.55585	47942	114.1195	22.57382	0.029877
17	22224	48048	114.1203	22.57377	48677	114.1131	22.54982	0.034534

OD format data

SUSTech Southern University of Science and Technology

Data Analysis and Visualization

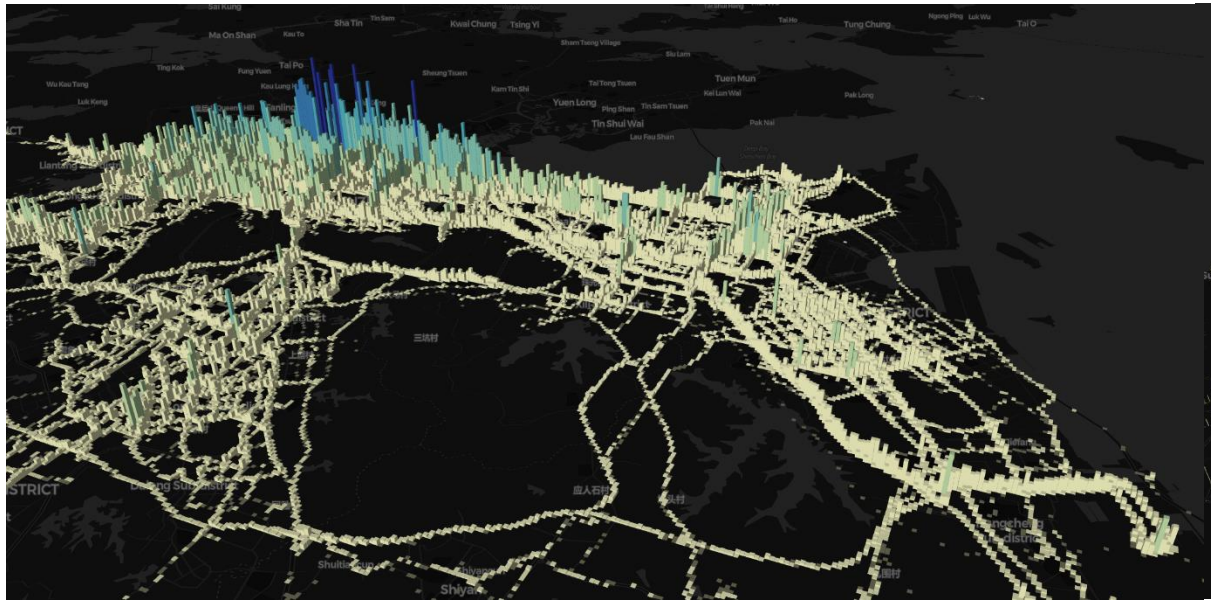
Path Dynamic Visualization

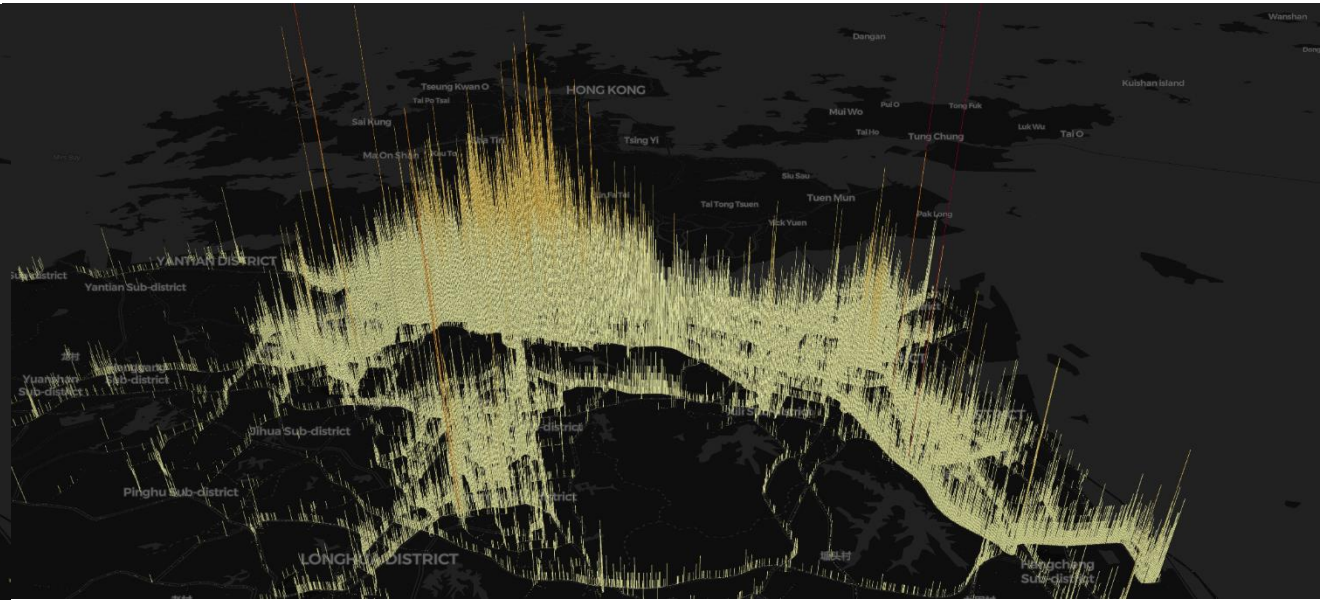


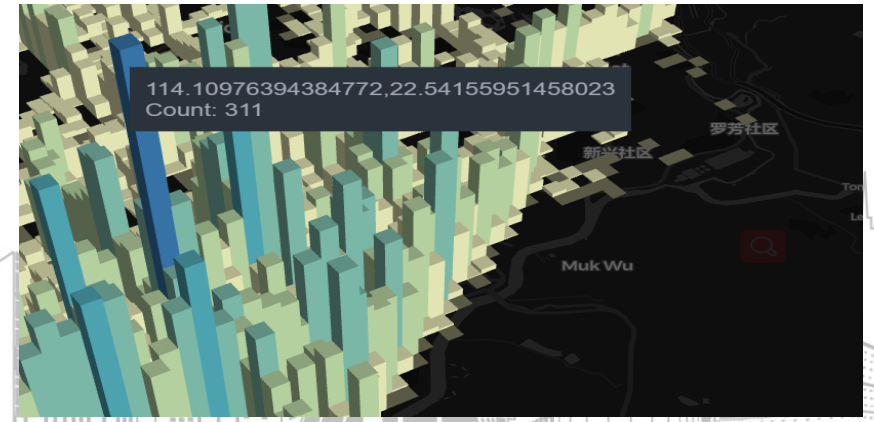
SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Popular area analysis







Weight diagram of morning peak traffic flow

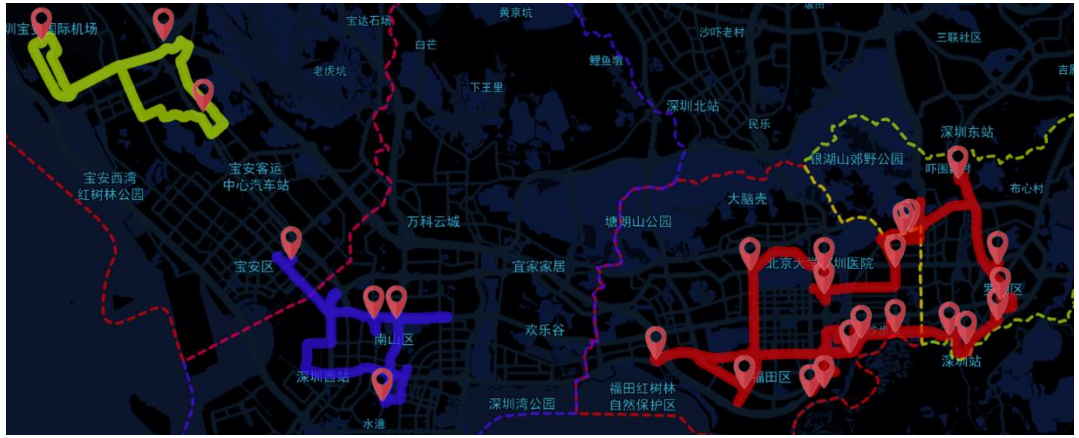
&

Density diagram of morning peak traffic flow

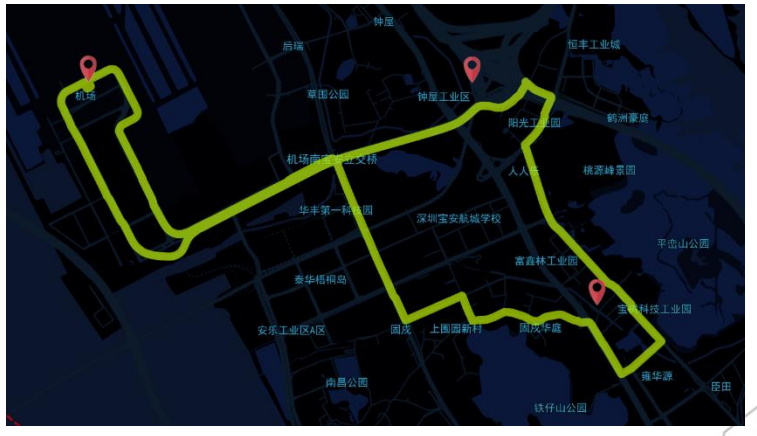
SUSTech Southern University of Science and Technology

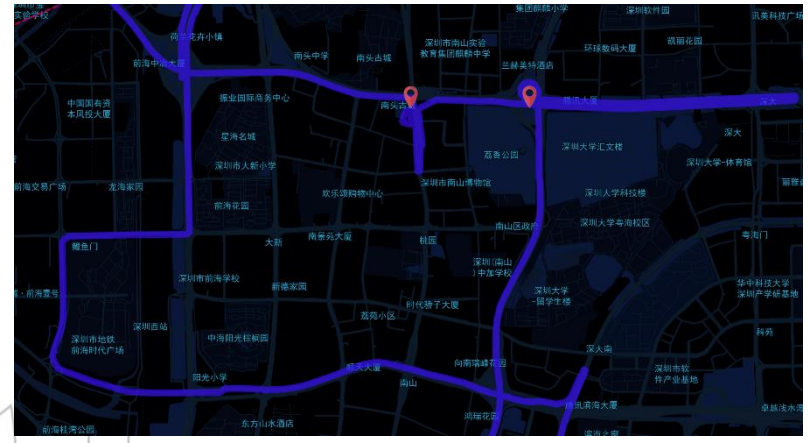
Data Analysis and Visualization

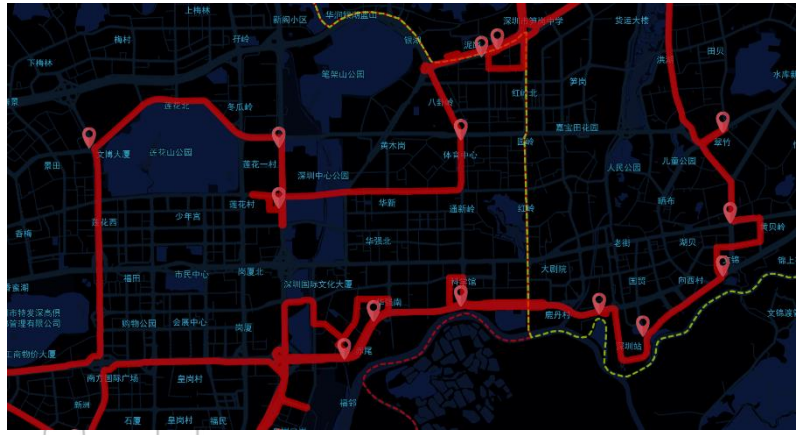
Popular area analysis



Red: Avg Count >= 100
(Huaqiang North Commercial Area)
Blue: Avg Count >= 80
(Nanshan District Museum & Shenzhen West Railway Station)
Green: Avg Count >=50
(Bao 'an International Airport)



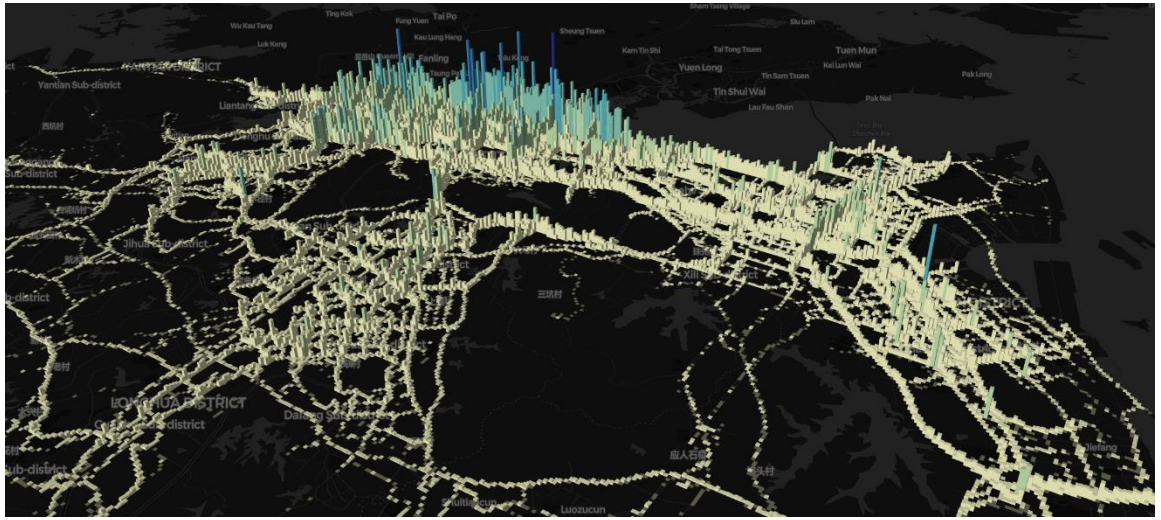




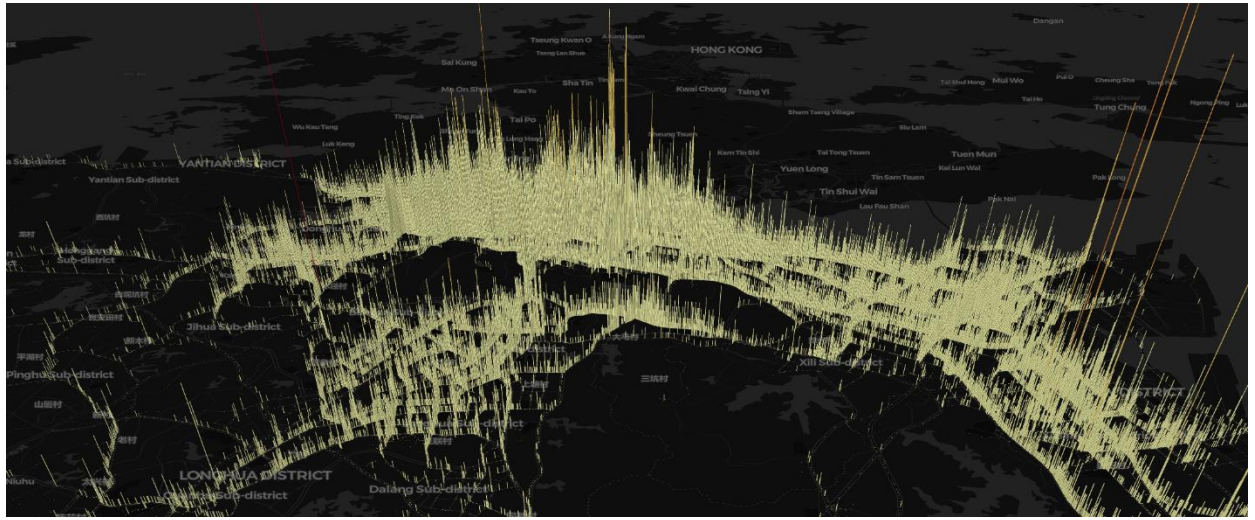
SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Popular area analysis



Weight diagram of night peak traffic flow

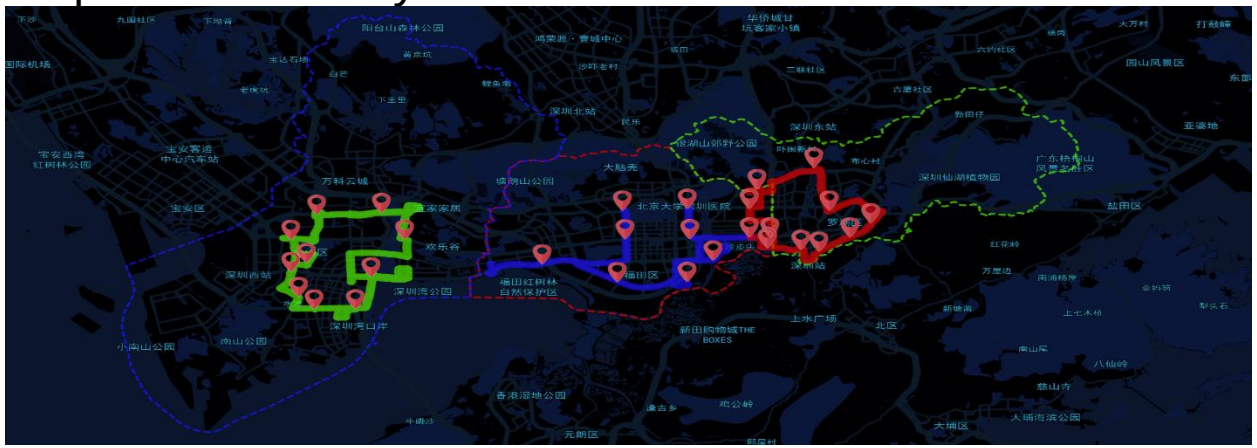


Density diagram of night peak traffic flow

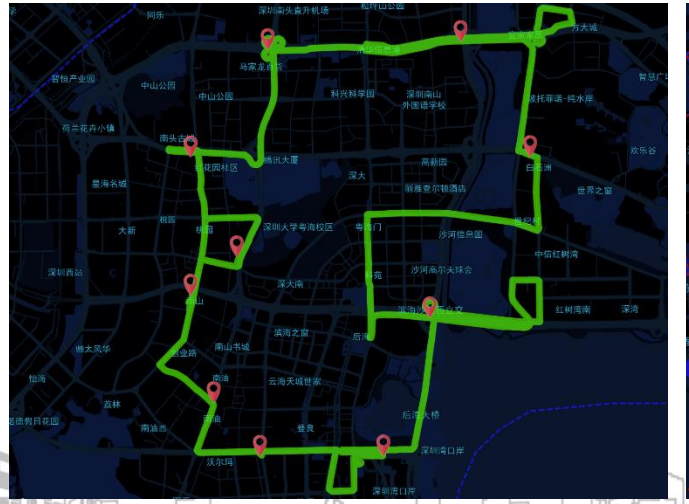
SUSTech Southern University of Science and Technology

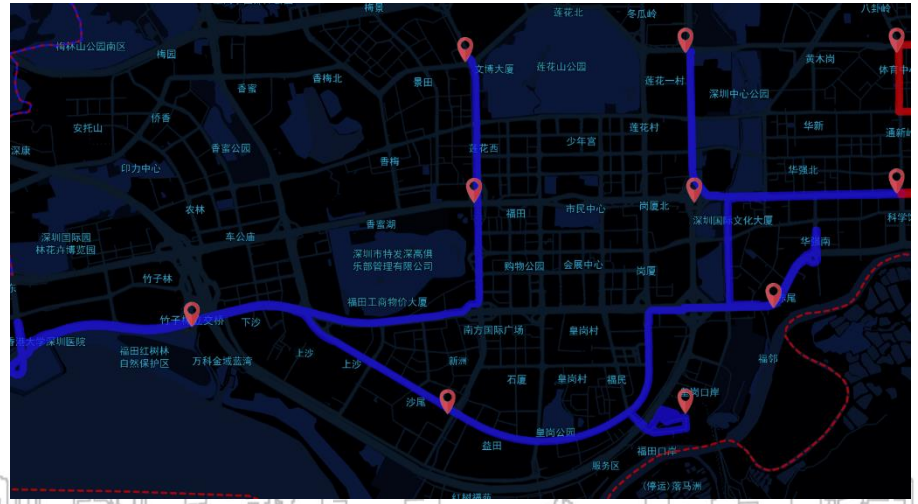
Data Analysis and Visualization

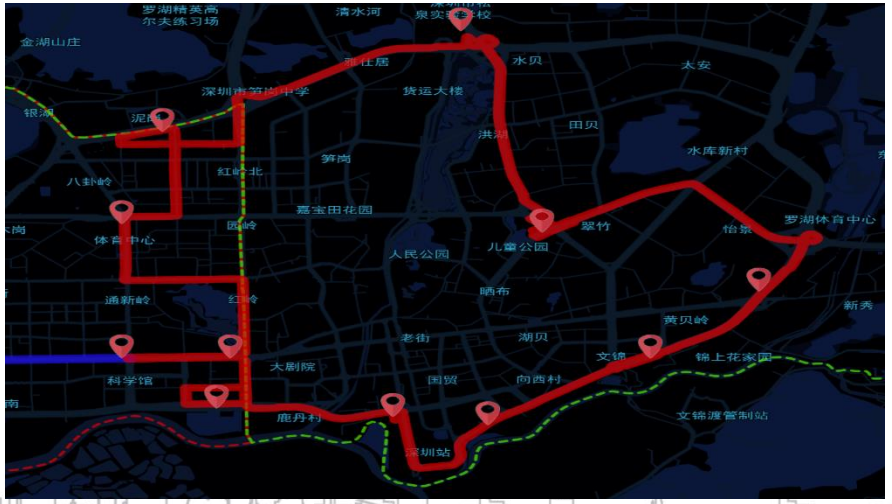
Popular area analysis



Red: Avg Count >= 120
(Shenzhen Station (Luohu Port))
Blue: Avg Count >= 100
(Civic Center & Convention Center)
Green: Avg Count >=80
(Shenzhen University & Yuehai Building)



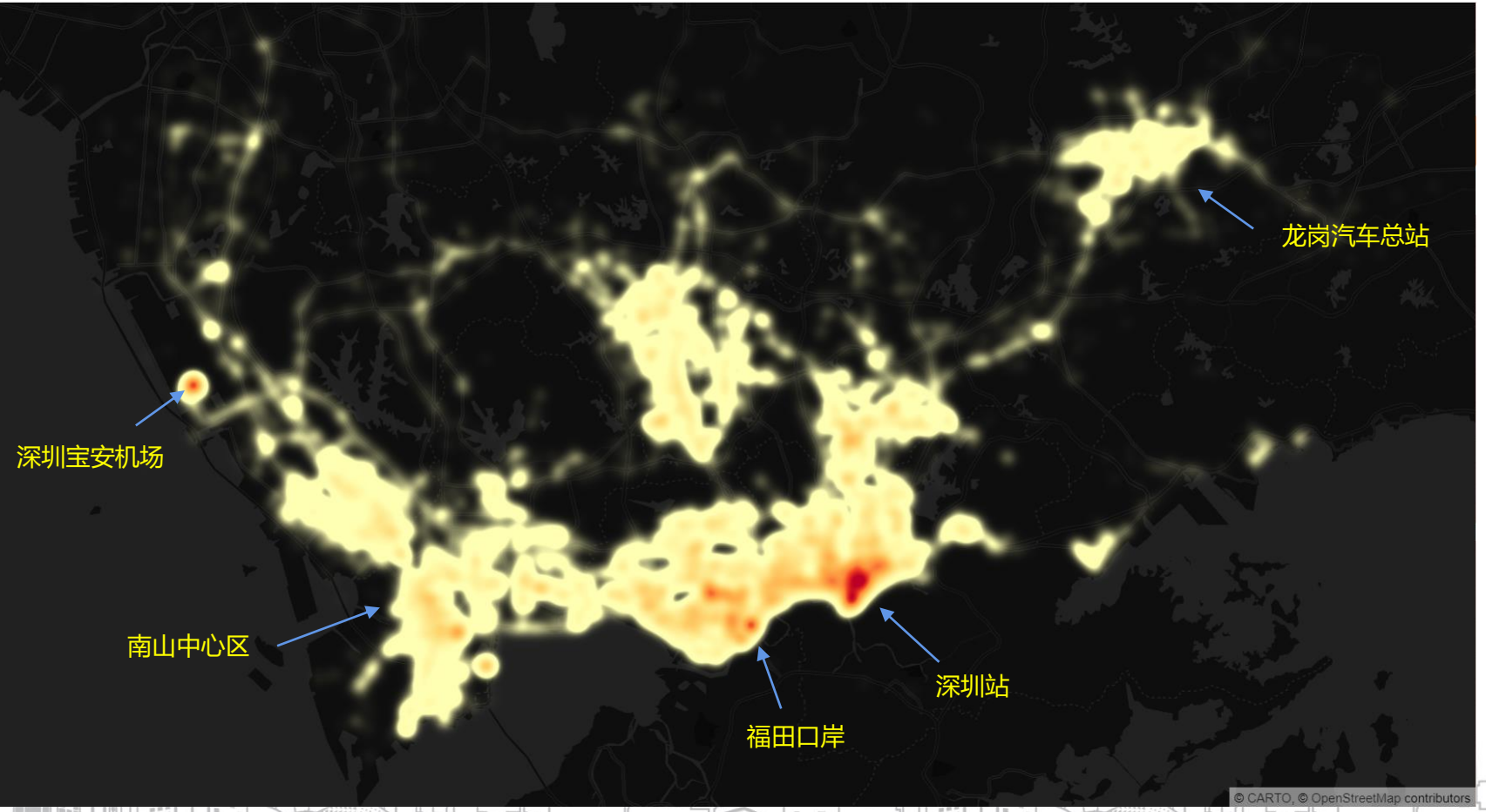




SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Popular area analysis



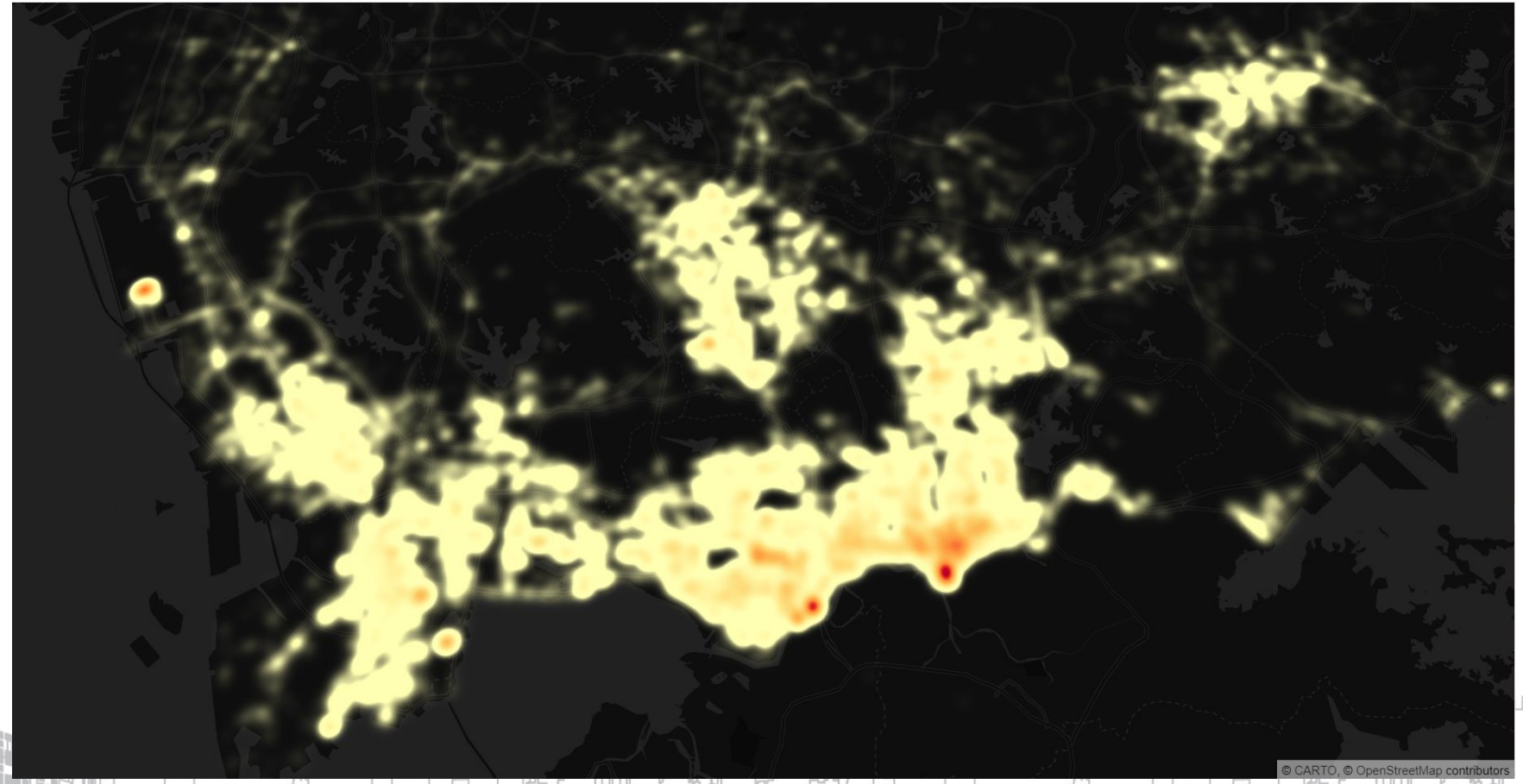
Heat map analysis
(Pickup place)

Combine most heated places' Longitude and latitude data with Gaode API and get to know what Place it is.

SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Popular area analysis



Heat map analysis
(Left place)

SUSTech Southern University of Science and Technology

Data Analysis and Visualization

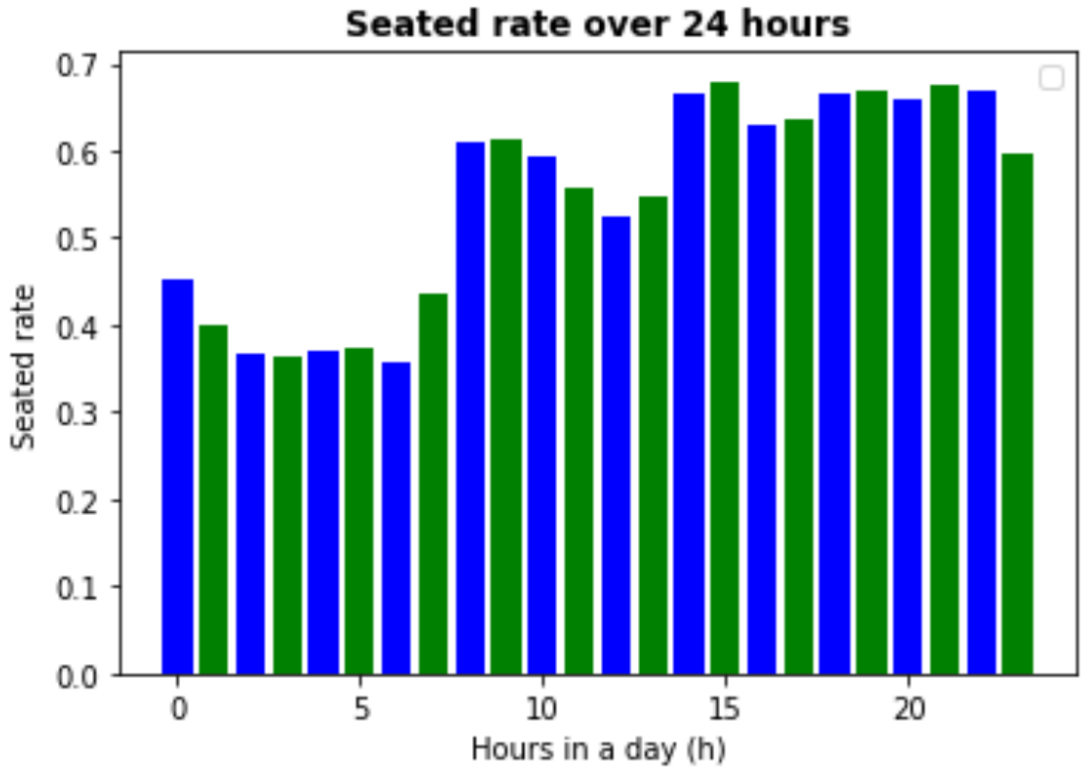
Popular area analysis

compare with heat map

SUSTech Southern University of Science and Technology

Data Analysis and Visualization

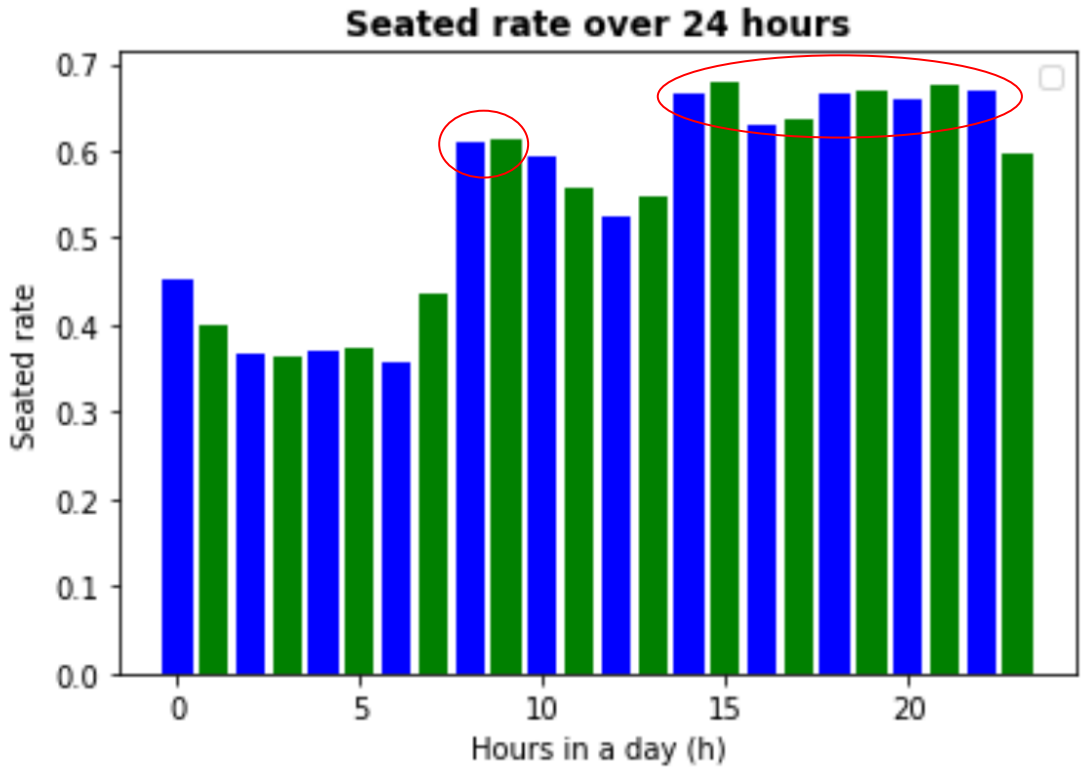
Seated rate over different time



Seated rate calculation method
For every hour we inspect on, for each car , we sum up the time it consumes during the time period when it is seated and divided by the whole time period. Then calculate the avg of all car to represent the seated rate during this time period.

Data Analysis and Visualization

Seated rate over different time



We can conclude from the graph:

1) During 0:00~7:00, the seated rate is pretty low
2) It can be infer that there is a morning peak during 7:00~10:00 since there is a obvious increase of seated rate
3) In the afternoon till night, the seated rate is pretty high

0:00~7:00 can be scheduled less taxis since the seated rate is pretty low.
Taxi driver can drive out during hours that have higher seated rate

Data Analysis and Visualization

Traffic trend direction analysis

0:00 —— 1:00
Car Num: 2814
(5.49, 12.37)

1:00 —— 2:00
Car Num: 2050
(9.53, 7.17)

2:00 —— 3:00
Car Num: 1519
(5.96, 2.26)

3:00 —— 4:00
Car Num: 1108
(4.82, 2.47)

4:00 —— 5:00
Car Num: 785
(-0.29, 0.94)

5:00 —— 6:00
Car Num: 732
(-8.17, 2.67)

Data Analysis and Visualization

Traffic trend direction analysis

6:00 —— 7:00
Car Num: 1271
(-10.23, 0.18)

7:00 —— 8:00
Car Num: 2290
(-10.62, 0.72)

8:00 —— 9:00
Car Num: 3220
(-7.42, 3.53)

9:00 —— 10:00
Car Num: 3053
(-3.90, 5.58)

10:00 —— 11:00
Car Num: 3344
(-5.00, 1.51)

11:00 —— 12:00
Car Num: 3015
(-4.88, 6.67)

SUSTech Southern University of Science and Technology

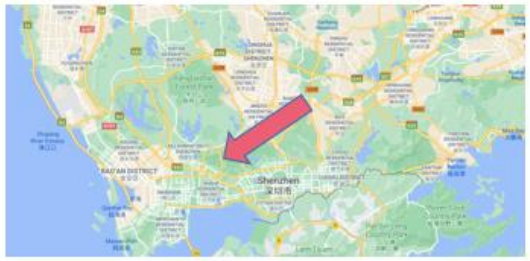
Data Analysis and Visualization

Traffic trend direction analysis

12:00 —— 13:00
Car Num: 2967
(-2.02, 2.40)

13:00 —— 14:00
Car Num: 3297
(0.27, 6.59)

14:00 —— 15:00
Car Num: 3559
(4.17, 0.39)



15:00 —— 16:00
Car Num: 3419
(-2.61, -1.02)

16:00 —— 17:00
Car Num: 3039
(2.40, 0.52)

17:00 —— 18:00
Car Num: 2995
(-1.02, 1.11)

Data Analysis and Visualization

Traffic trend direction analysis

18:00 —— 19:00
Car Num: 2674
(2.37, -3.36)

19:00 —— 20:00
Car Num: 3316
(-3.69, 6.36)

20:00 —— 21:00
Car Num: 3765
(2.35, 5.52)

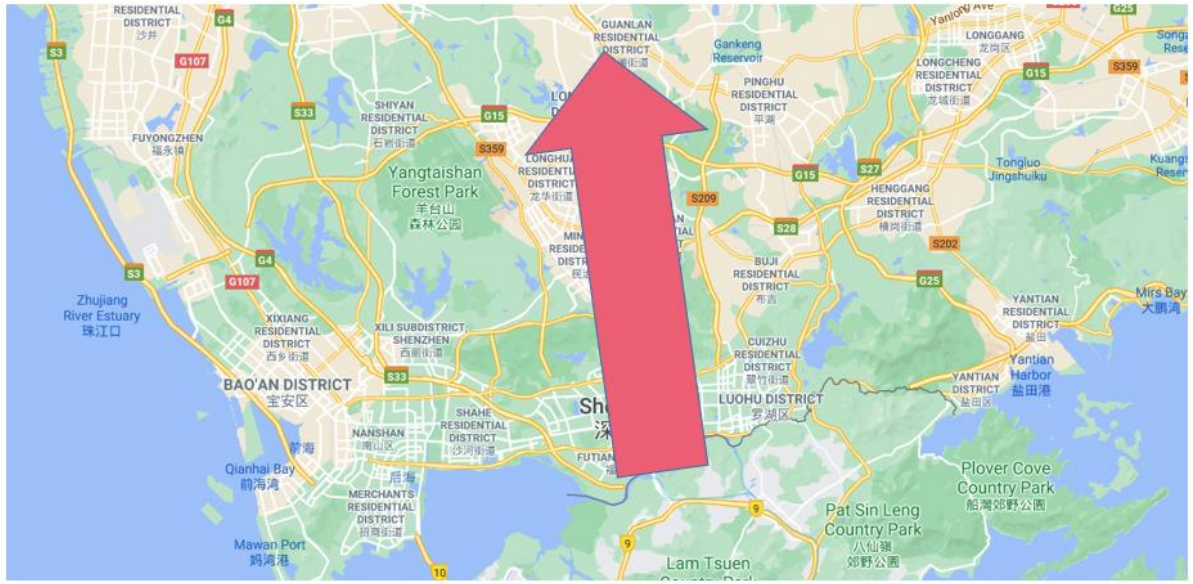
21:00 —— 22:00
Car Num: 3877
(10.66, 8.21)

22:00 —— 23:00
Car Num: 4025
(8.76, 13.05)

23:00 —— 24:00
Car Num: 3020
(0.93, 6.29)

Data Analysis and Visualization

Traffic trend direction analysis



24小时内 总人流量：(-2.12, 92.13)

We can conclude that, from 20:00 to the f ollowing 7:00,there is a tendency that peo ple move to northeast, while in other time, there is a tendency than people move to northwest. Moreover, the sum of the vecto r shows that, there is a tendency that peo ple want to move to north, which means t here are more people take a taxi in south and get off in north.

Data Analysis and Visualization

Car speed analysis

When the speed is slow, it can reflect the congestion situation of the road and get the area with high traffic jam incidence. When the speed is fast, it can be assumed that it is on the highway, so as to get the distribution of the high-speed.

Set 0 < speed <= 15 is low speed, speed >= 80 is high speed.

SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Car speed analysis

Divide the data into two csv file according to the taxi speed.

Visualize the data by pydeck.

```
low = data[(data['speed'] <= 15) & (data['speed'] > 0)]
high = data[data['speed'] >= 80]

low.to_csv('low.csv', index=False)
high.to_csv('high.csv', index=False)
```

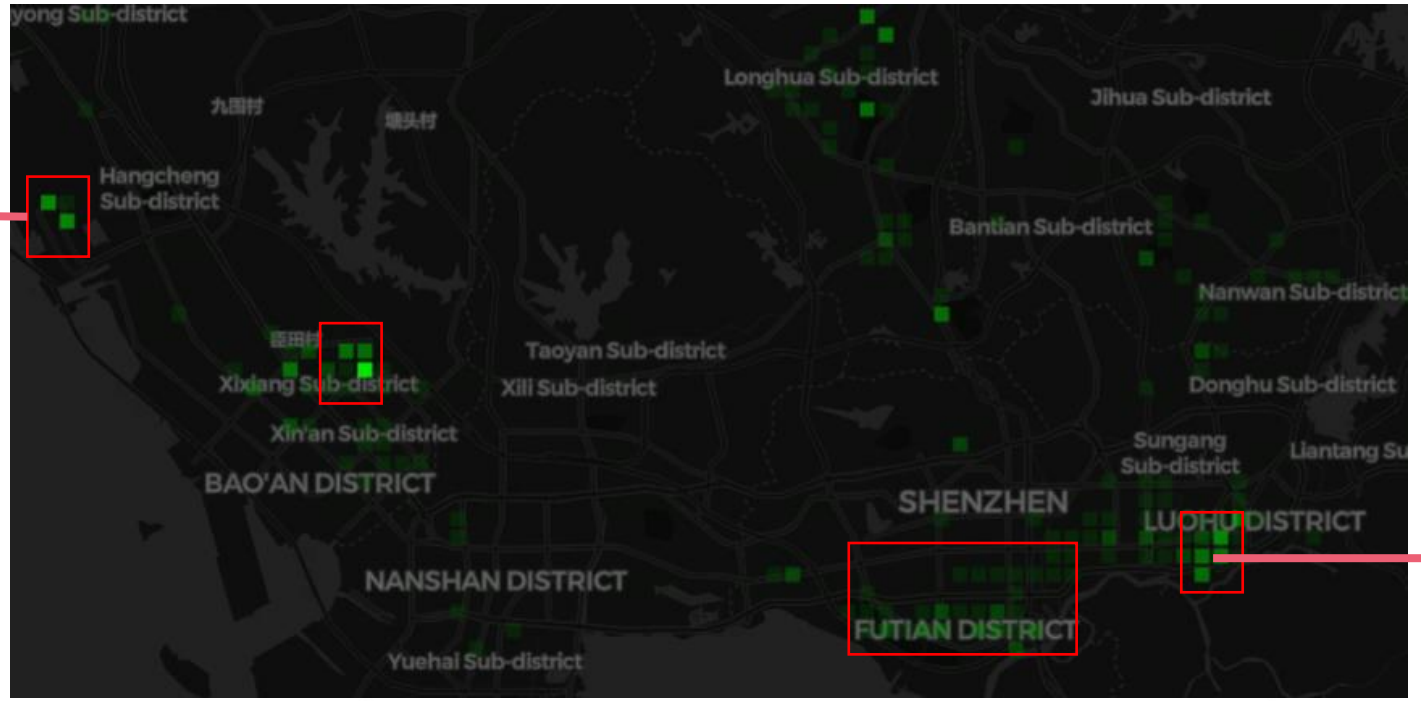
```
view_state = pdk.ViewState(
    longitude=114.023465,
    latitude=22.632468,
    zoom=11,
    pitch=0,
    bearing=0
)

r = pdk.Deck(layers=[layer], initial_view_state=view_state)
r.to_html('low.html')
```

SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Car speed analysis and traffic jam



SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Car speed analysis and traffic jam

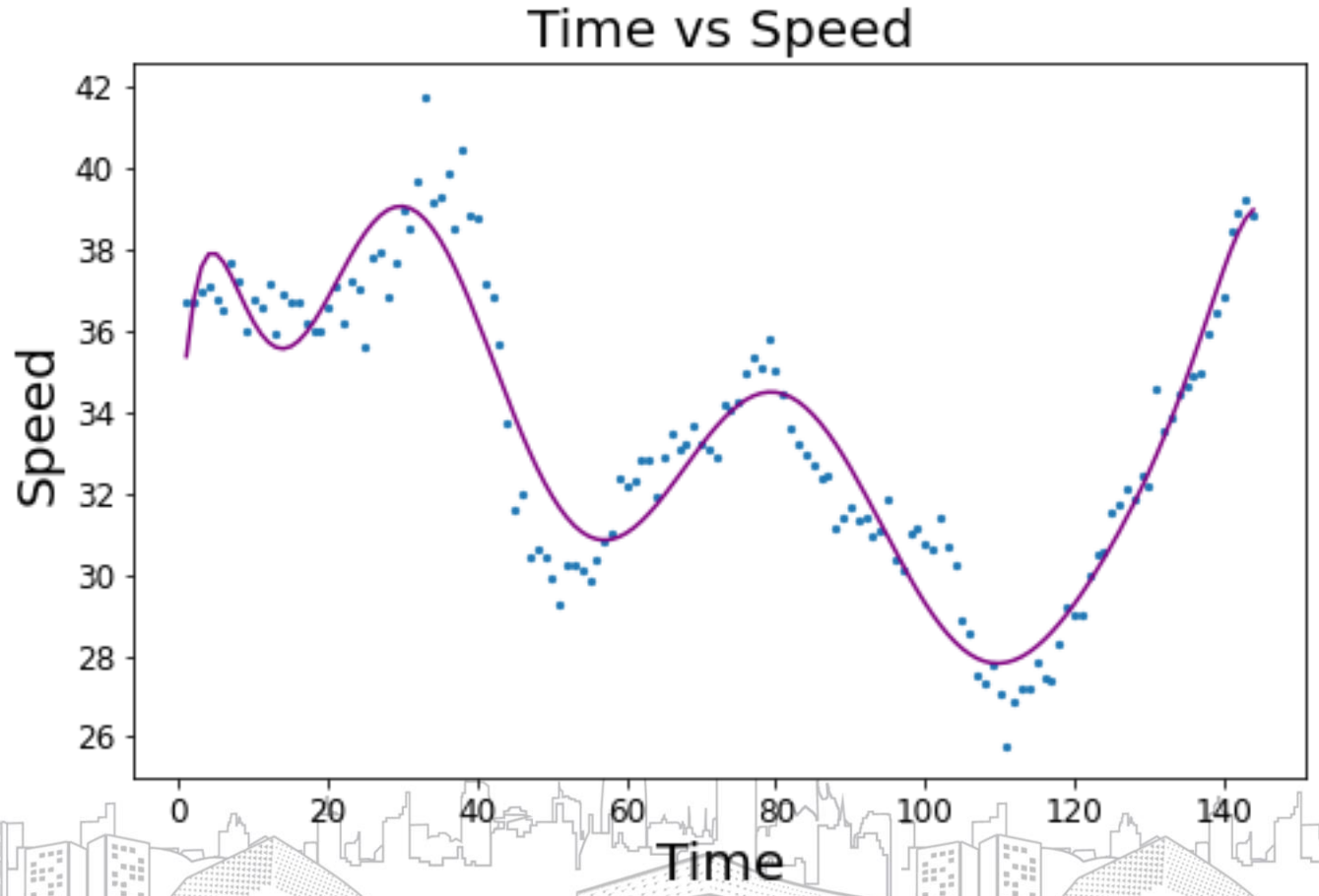
ShenHai expressway

Jinggang'ao expressway

Riverside avenue

Data Analysis and Visualization

Peak Observation and Prediction



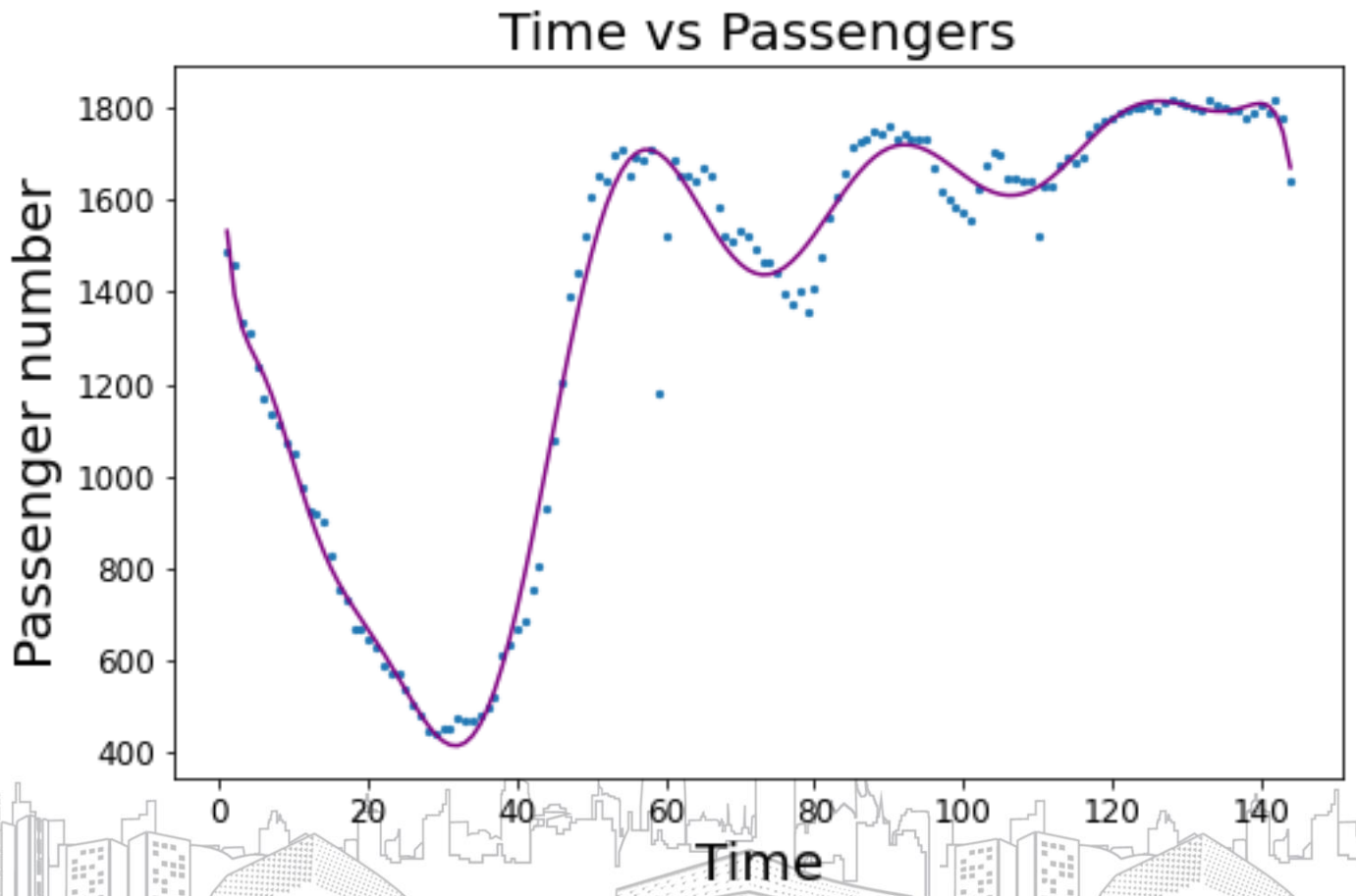
Time: every 10 minutes

Degree: 10 (Adjust to avoid bad fit and overfit)

Observe speed peak and valley
Predict speed situation in other days for this city according to the timeline

Data Analysis and Visualization

Peak Observation and Prediction



Time: every 10 minutes

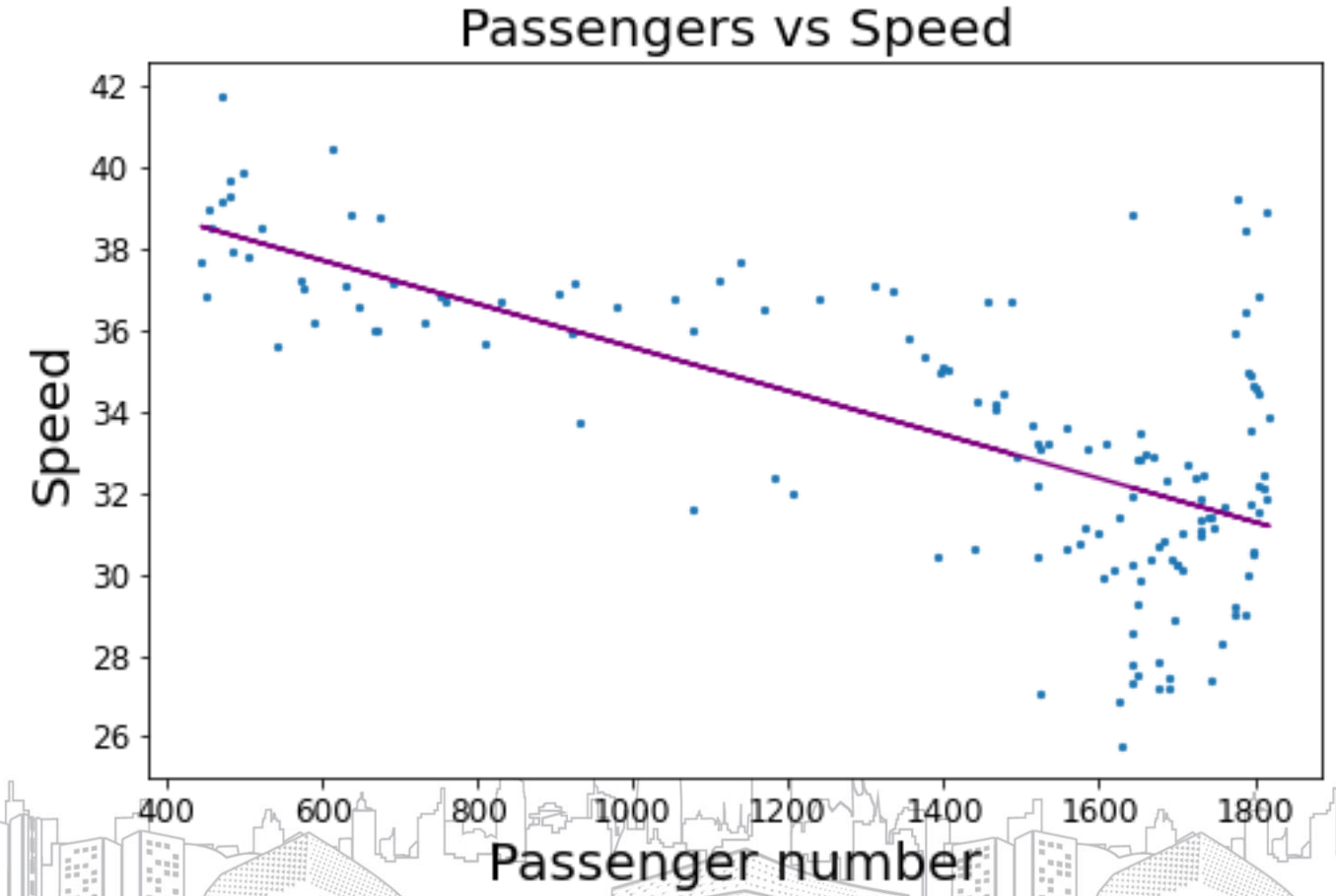
Degree: 15 (Adjust to avoid bad fit and overfit)

Observe traffic peak and valley
Predict traffic situation in other days for this city according to the timeline

Question: Does traffic has a relation with the speed?

Data Analysis and Visualization

Peak Observation and Prediction



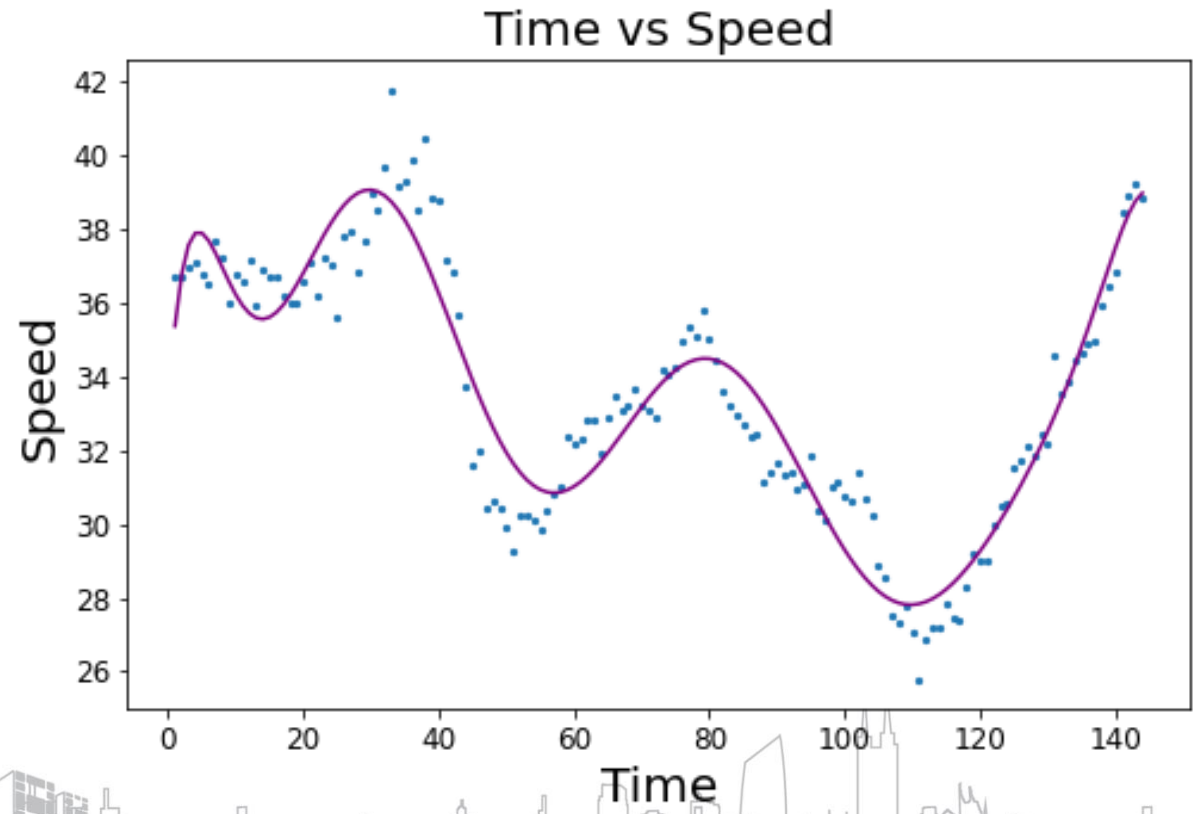
Question: Does traffic has a relation with the speed?

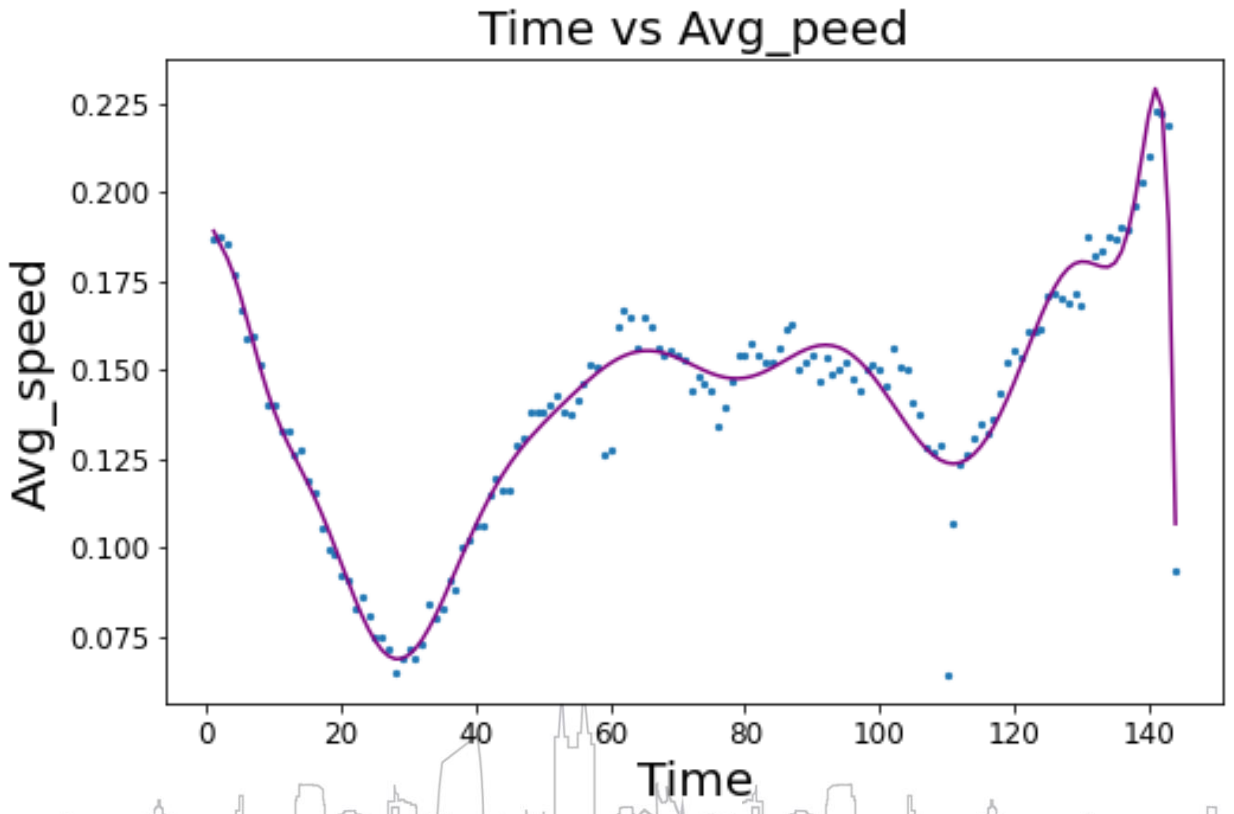
Approximately, the slower the speed, the more congested the traffic

Use speed data observed to predict traffic situation like traffic accidents; use traffic situation to predict the driving speed and make plan ahead of time.

Data Analysis and Visualization

Taxi utilization analysis

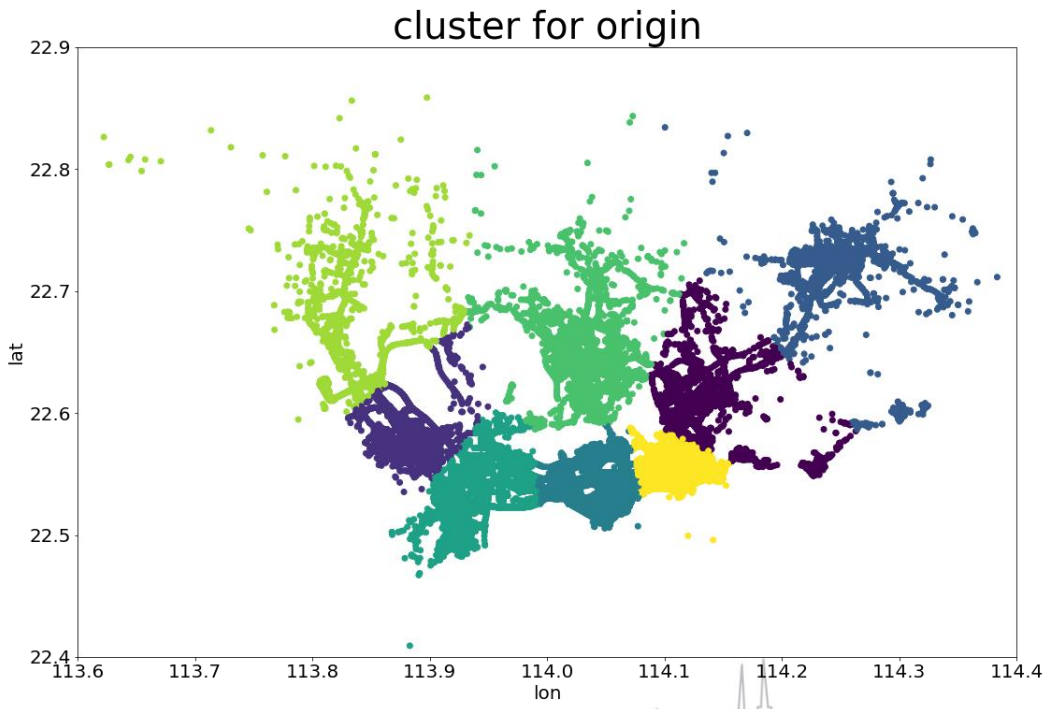


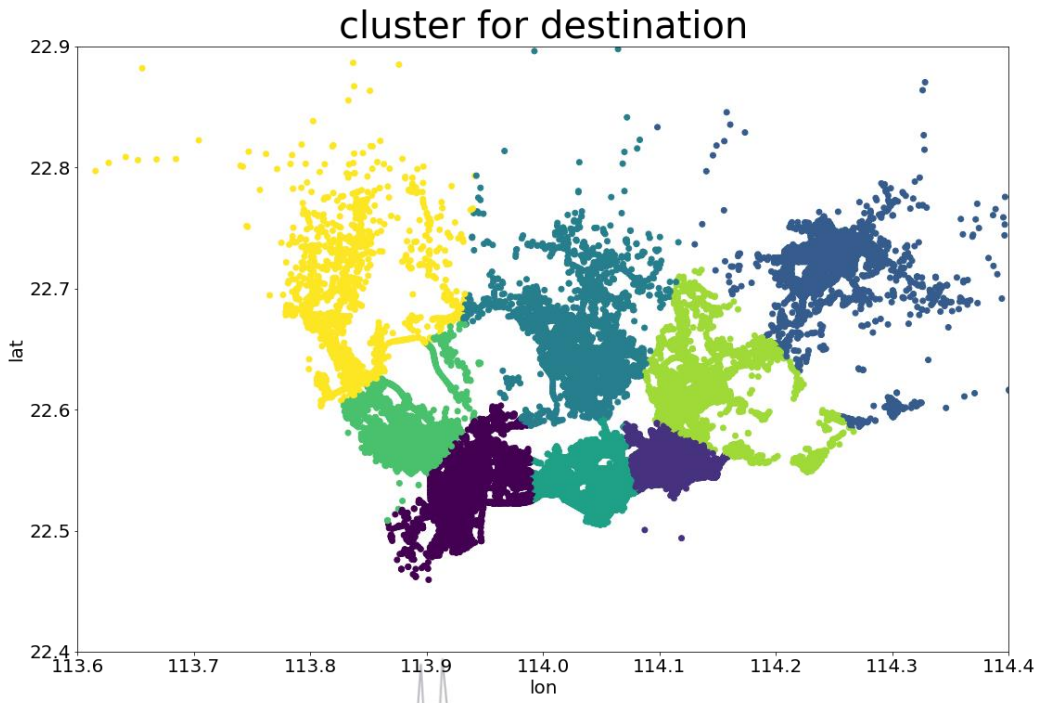


SUSTech Southern University of Science and Technology

Data Analysis and Visualization

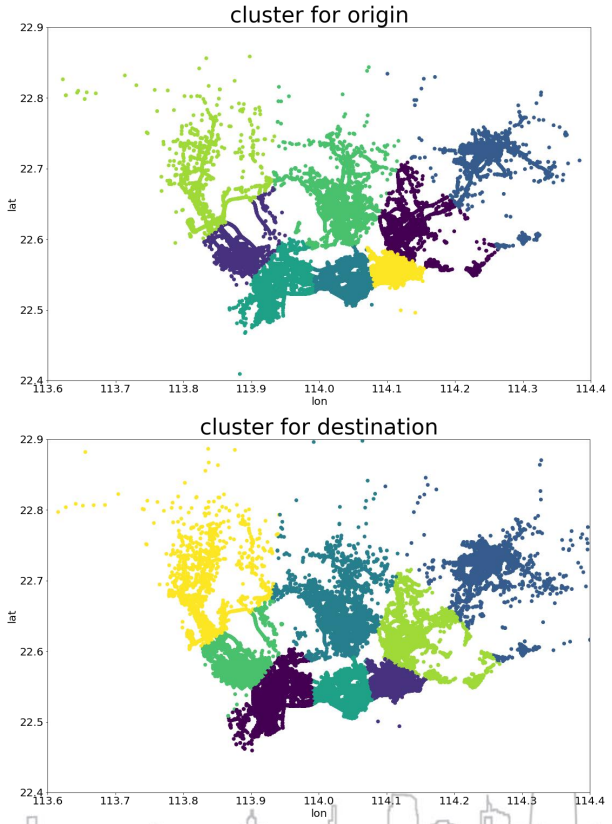
Standard for price -- Region





Data Analysis and Visualization

Standard for price -- Region



1. 8 clusters(decided by k-means training)

    Clusters are similar for origins and destinations

    Rationable to divide Shenzhen into 8 small regions(in reality: 6 Administration regions and 4 Functional regions)

2. Set different price standard for different regions

    As shown before, a taxi whose destination locates in popular areas has higher possibility to get a next order

    Backup level(Ability to get a next order quickly):

$$\sum_{i \in R} \frac{possibility(i)}{distanceTo(i)}$$

R: regions

possibility(i): the possibility to get a next order in region i

distanceTo(i): the distance from the current position(destionation of the previous order) to region i

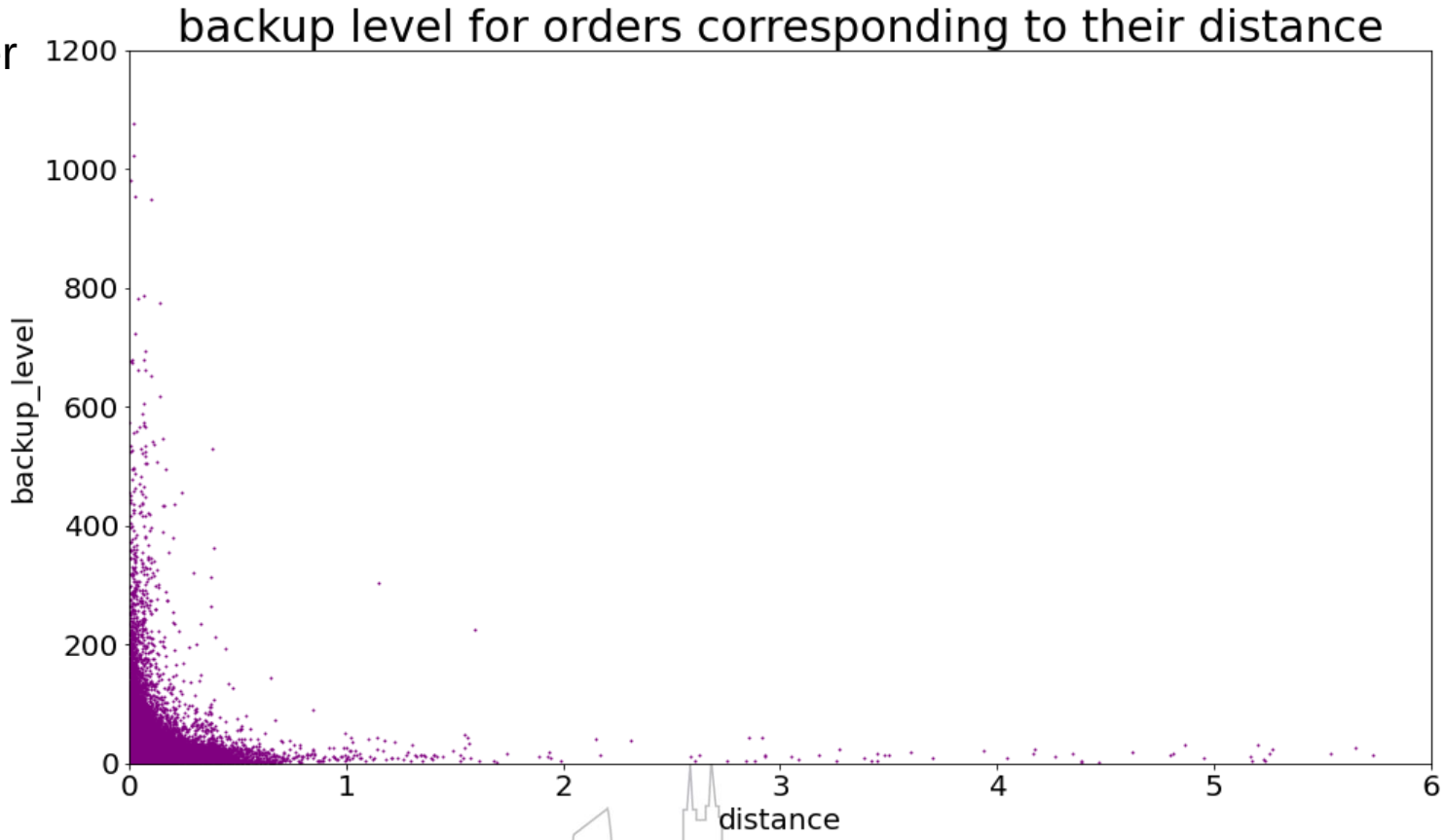
Data Analysis and Visualization

Standard for price -- Distance per Order

Distance: Travel distance for every order
Backup level: Ability to get a next order quickly. The higher the level, the stronger the ability.

$$\sum_{i \in R} \frac{possibility(i)}{distanceTo(i)}$$

R: regions
possibility(i): the possibility to get a next order in region i
distanceTo(i): the distance from the current position(destionation of the previous order) to region i



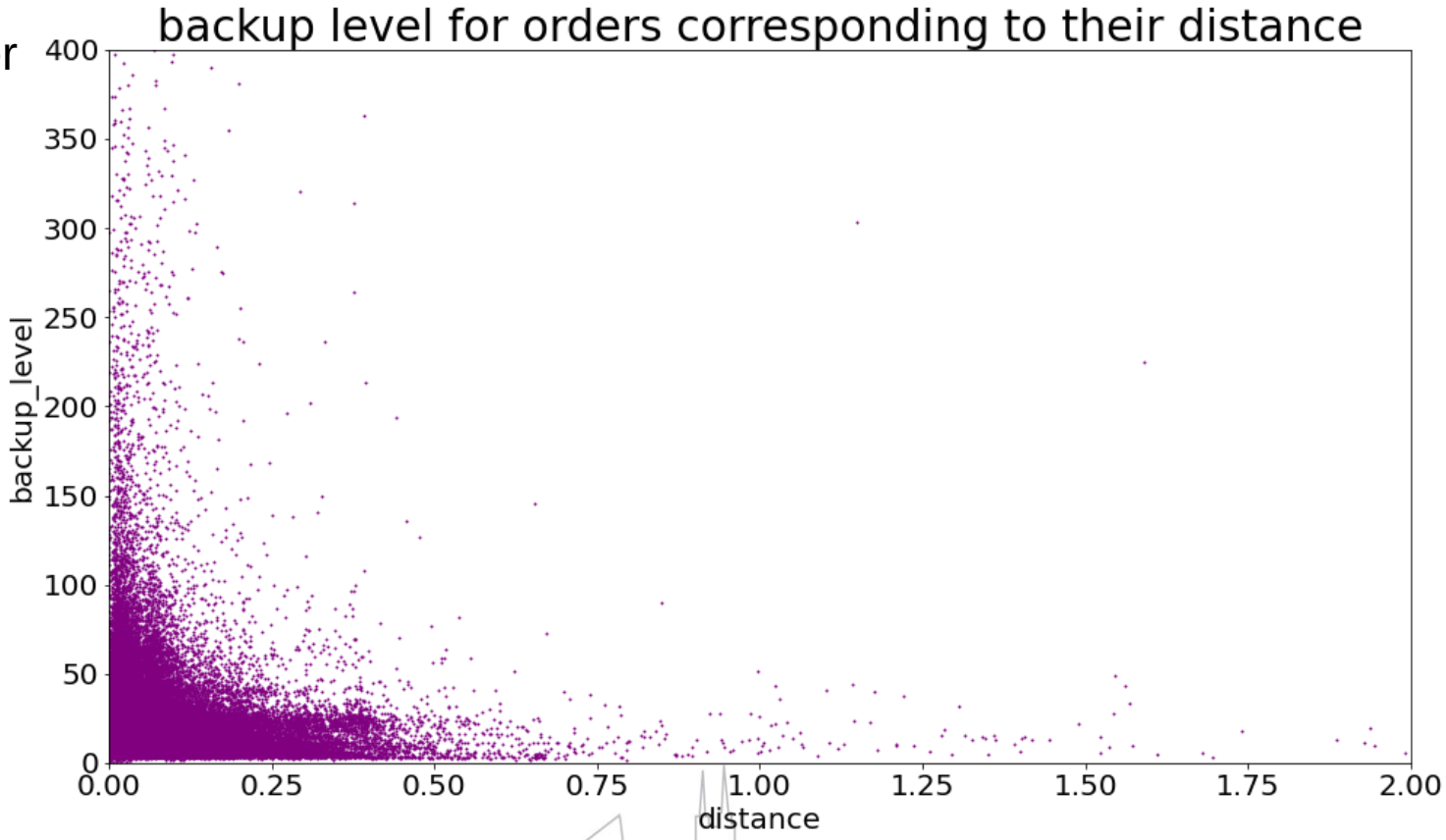
Data Analysis and Visualization

Standard for price -- Distance per Order

Distance: Travel distance for every order
Backup level: Ability to get a next order quickly. The higher the level, the stronger the ability.

$$\sum_{i \in R} \frac{possibility(i)}{distanceTo(i)}$$

R: regions
possibility(i): the possibility to get a next order in region i
distanceTo(i): the distance from the current position(destionation of the previous order) to region i



SUSTech Southern University of Science and Technology

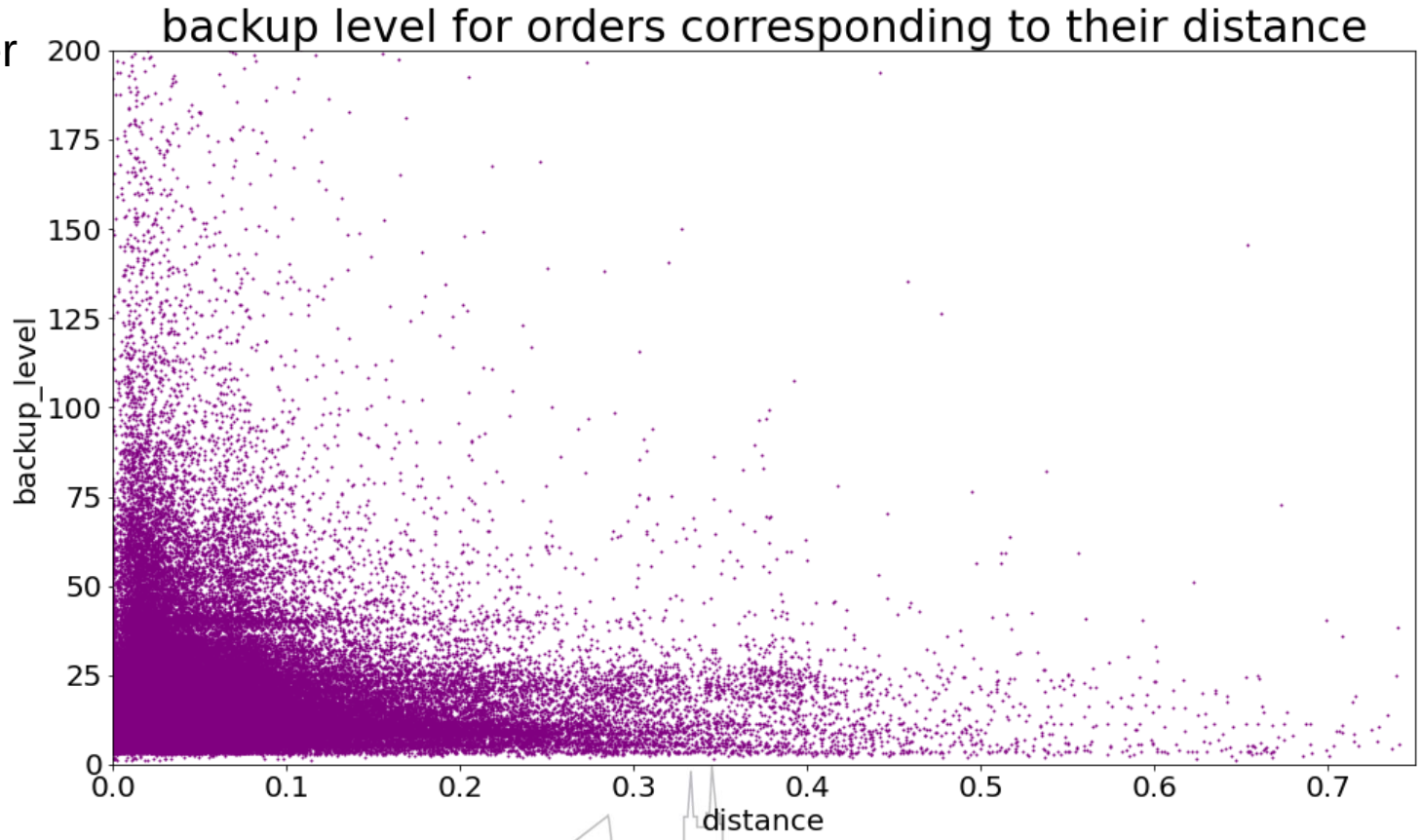
Data Analysis and Visualization

Standard for price -- Distance per Order

Distance: Travel distance for every order
Backup level: Ability to get a next order quickly. The higher the level, the stronger the ability.

$$\sum_{i \in R} \frac{possibility(i)}{distanceTo(i)}$$

R: regions
possibility(i): the possibility to get a next order in region i
distanceTo(i): the distance from the current position(destionation of the previous order) to region i

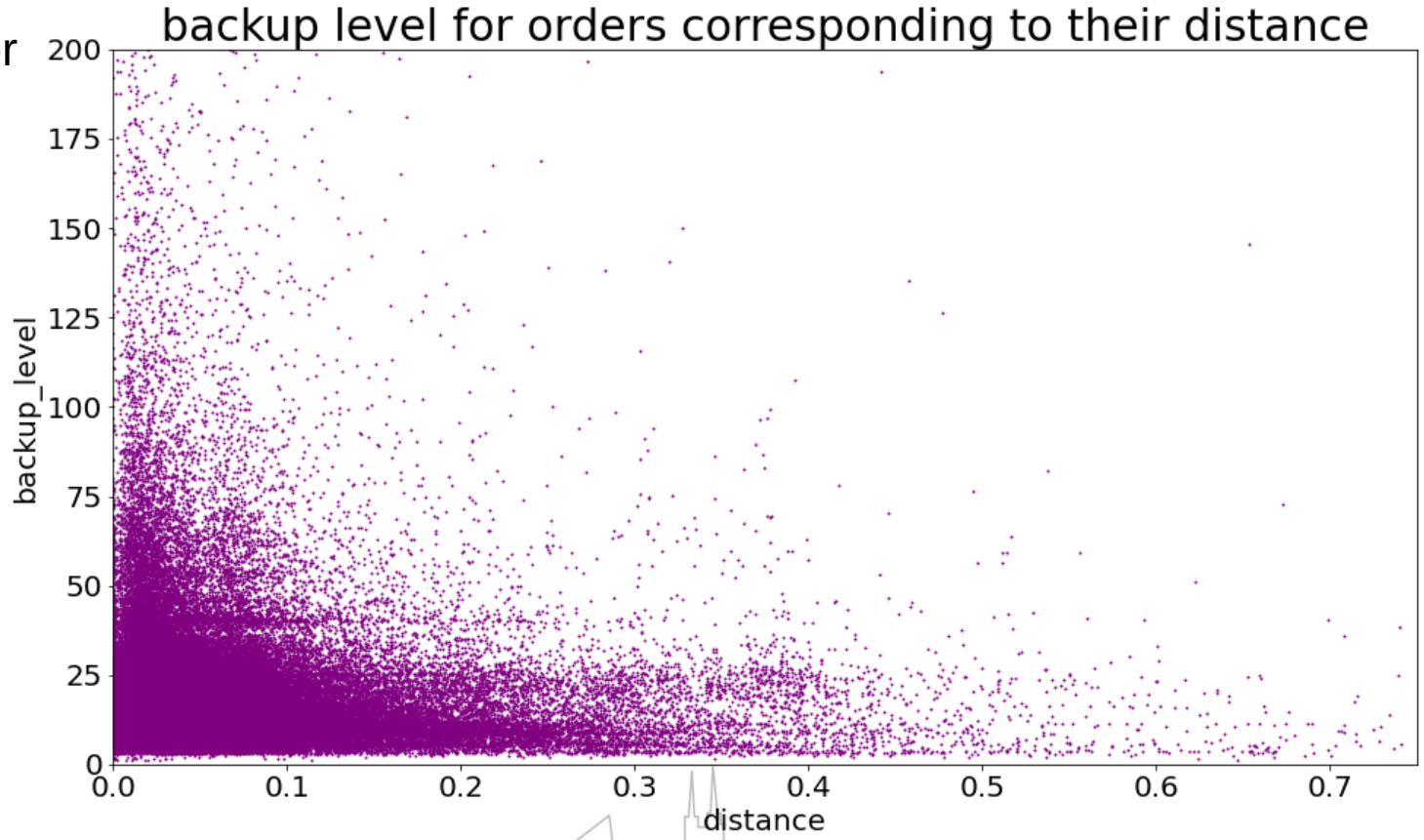


Data Analysis and Visualization

Standard for price -- Distance per Order

Conclusion:
The longer the distance, the harder a taxi can quickly get another order after the previous one.

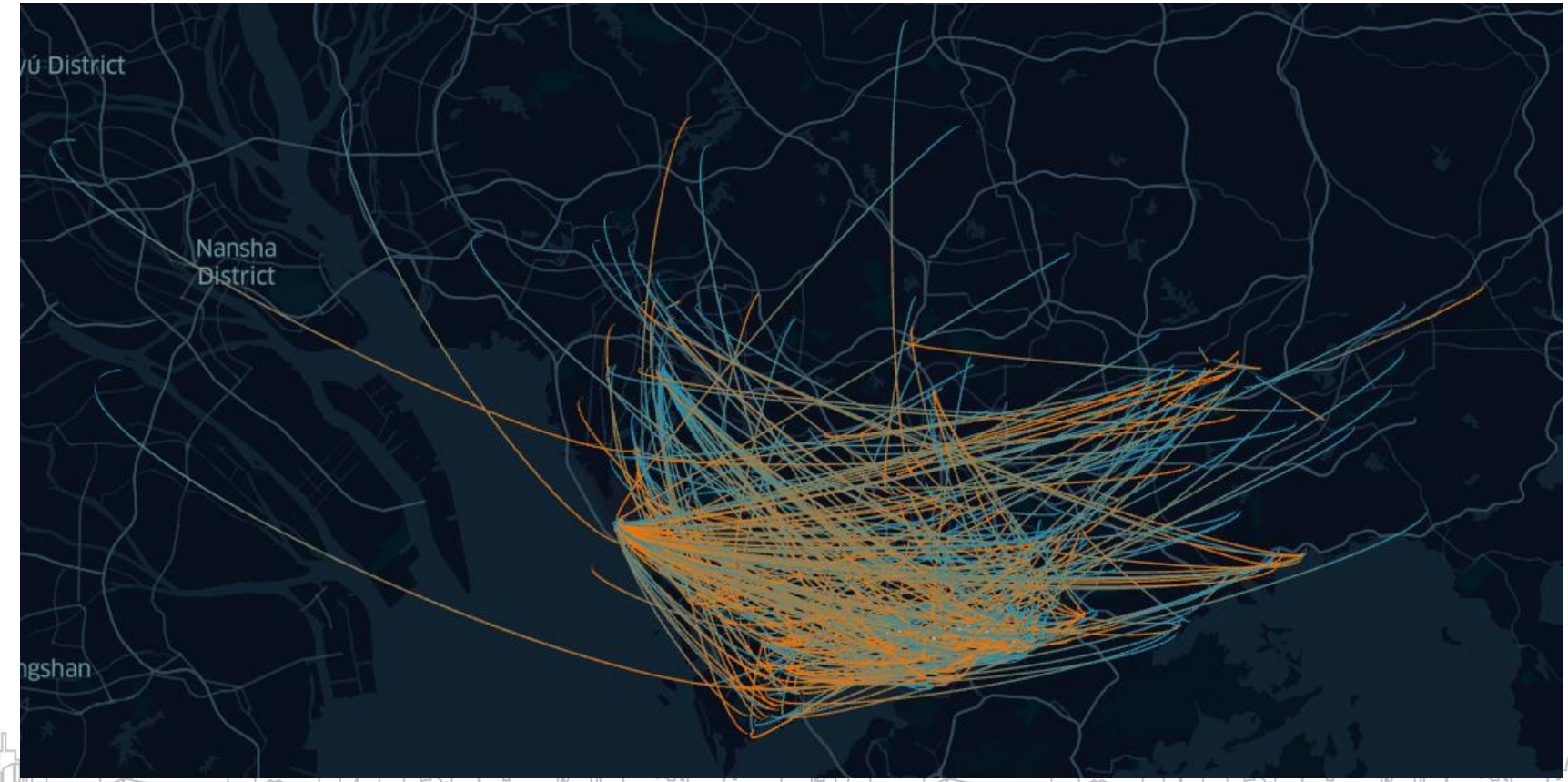
Why?



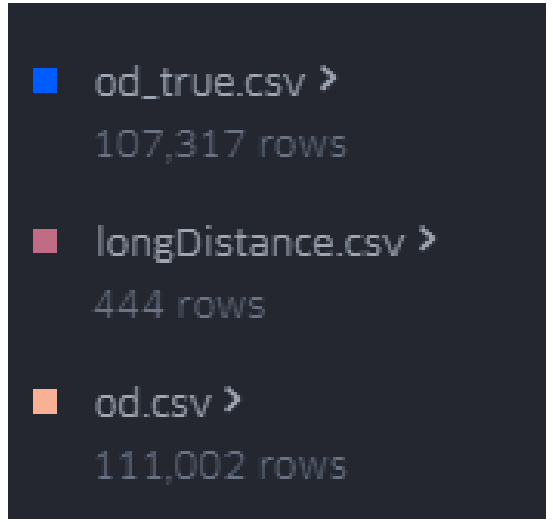
SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Standard for price -- Distance per Order



Tool: kelper.gl

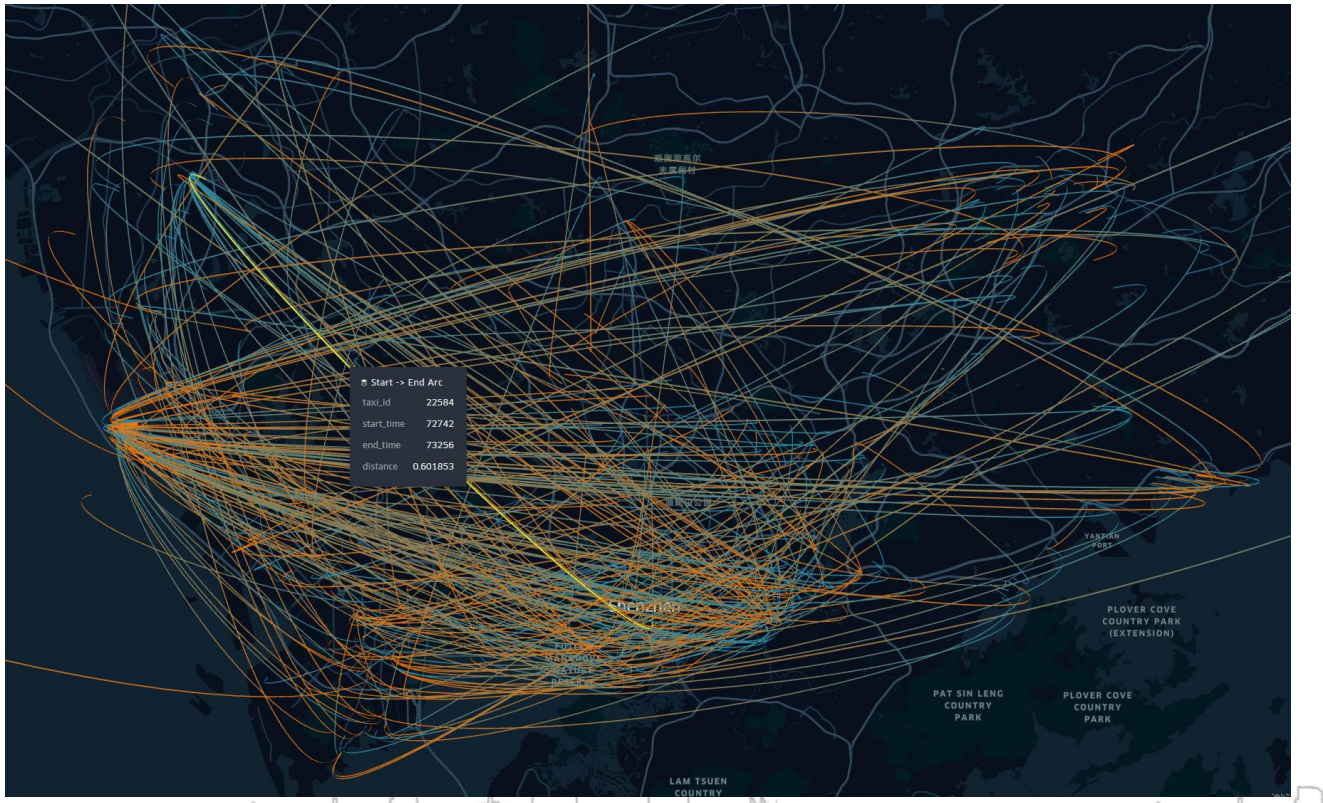


Validation result

SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Standard for price -- Distance per Order



For an order travellling long distance, usually either its origin or destination is far away from the popular areas while one point can locate in popular areas.

Origin not in popular areas:
costly to get this order

Destination not in popular areas:
costly to get next order

SUSTech Southern University of Science and Technology

Data Analysis and Visualization

Standard for price, taking into account Region and Distance per Order

Conclusion:
1. An order ends in region without many popular areas should pay an extra cost
2. An order travels a long distance should pay an extra cost

SUSTech Southern University of Science and Technology

Conclusion

Difficulties we've met and how we conquer it

1) The data is large, it is time-consuming to perform several operation over it.
   Regarding time-consuming operation, we sample it over certain time period.
2) The conversion of data format is complex, hard to fit the data format into the model.
   we wrote python script to fix this.
1) When trying to convoke Gaode API, we face quota limitation problem and efficiency problem.
   We decided not to generate query for all data, but for important data like the data in popular area.

What we learnt

We are more familiar with the whole set of process of data analysis.
While brainstorming ideas, we manage to enhance our creativity mind and ability to extract insight from large set of data.
We've learnt to use tools like deck.gl and many python libraries(Pandas, matplotlib ….). Also we learnt to find data online and convoking commercial api for our use.

Thank you!