Background

As less than 10% of China’ s 1.4 billion citizens own automobiles, the frequency at which Chinese citizens commute on taxis, buses, trains, and planes is the highest in the world. Didi Chuxing, the dominant ride-hailing company in China, processes over 11 million trips, plans over 9 billion routes and collects over 50TB of data per day. To meet needs of riders, Didi must continually innovate to improve cloud computing and big data technologies and algorithms in order to process this massive amount of data and uphold service reliability.

Supply-demand forecasting is critical to enabling Didi to maximise utilisation of drivers and ensure that riders can always get a car whenever and wherever they may need a ride. Supply-demand forecasting helps to predict the volume of drivers and riders at a certain time period in a specific geographic area. For instance, demand tends to surge in residential areas in the mornings and in business districts in the evenings. Supply-demand forecasting allows Didi to predict demand surges and guide drivers to those areas. The end result is higher earnings for drivers and no surge pricing for riders!

File name (after registration can be downloaded)	File format	Download
citydata.tar.gz	.tar.gz(600MB)	Download

Definition and Evaluation Criteria

1. Definition

A passenger calls a ride(request)by entering the place of origin and destination and clicking “Request Pickup” on the Didi app. A driver answers the request (answer) by taking the order.

Didi divides a city into n non-overlapping square districts $D=\{d_1,d_2,\cdots,d_n\}$ and divides one day uniformly into 144 time slots $t_1,t_2,\cdots,t_{144}$, each 10 minutes long.

In district d_i, and time slot t_j , the number of passengers' requests is denoted as a_{ij}, and drivers' answers as a_{ij} .

In district d_i , and time slot t_j , the demand is denoted as $demand_{ij} = r_{ij}$ and the supply as $supply_{ij} = a_{ij}$, and the demand supply gap is: gap_{ij} : $gap_{ij} = r_{ij} - a_{ij}$

Given the data of every district and time slot t_j , t_{j-1} , you need to predict $gap_{i,j+1}$, $\forall d_i \in D$.

2. Evaluation Metrics

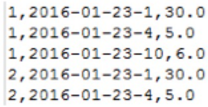
Consider d_i districts and t_j time slots, and the supply-demand gap gap_{ij} , and your prediction is s_{ij} , we use as the evaluation metrics:

$$MAPE = \frac{1}{n}\sum_{d_i}\left(\frac{1}{q}\sum_{t_j}\left|\frac{gap_{ij} - s_{ij}}{gap_{ij}}\right|\right), \quad \forall gap_{ij} > 0$$

The lowest MAPE will be the best.

3. Format of your results

The result you submitted needs to be in the following format: district ID, time slot, prediction value. For example:



```
1,2016-01-23-1,30.0
1,2016-01-23-4,5.0
1,2016-01-23-10,6.0
2,2016-01-23-1,30.0
2,2016-01-23-4,5.0
```

The detailed description of each field is as follows:

Data name	Data type	Example
District ID	string	1,2,3,4 (the same as district mapping ID)
Time slot	string	2016-01-23-1 (The first time slot on Jan. 23rd, 2016; one day is uniformly divided into 144 ten minute time slots)
Prediction value	double	6.0

Data Format

The training set contains three consecutive weeks of data for City M in 2016, and you need to forecast the supply-demand gap for a certain period in the fourth and fifth weeks of City M. The test set contains the data of half an hour before the predicted time slot. The specific time slots where you need to predict the supply-demand gap are shown in the explanation document in the test set.

The Order Info Table, Weather Info Table and POI Info Table are available in the database, while the District Definition Table and Traffic Jam Info Table are derived from other tables in the database. All sensitive data has been anonymised.

Order Info Table

Field	Type	Meaning	Example
order_id	string	order ID	70fc7c2bd2caf386bb50f8fd5dfef0cf
driver_id	string	driver ID	56018323b921dd2c5444f98fb45509de
passenger_id	string	user ID	238de35f44bbe8a67bdea86a5b0f4719
start_district_hash	string	departure	d4ec2125aff74eded207d2d915ef682f
dest_district_hash	string	destination	929ec6c160e6f52c20a4217c7978f681
Price	double	Price	37.5
Time	string	Timestamp of the order	2016-01-15 00:35:11

The Order Info Table shows the basic information of an order, including the passenger and the driver (if driver_id =NULL, it means the order was not answered by any driver), place of origin, destination, price and time. The fields order_id, driver_id, passenger_id, start_hash, and dest_hash are made not sensitive.

District Info Table

Field	Type	Meaning	Example
district_hash	string	District hash	90c5a34f06ac86aee0fd70e2adce7d8a

district_id	string	District ID	1

The District Info Table shows the information about the districts to be evaluated in the contest. You need to do the prediction given the districts from the District Definition Table. In the submission of the results, you need to map the district hash value to district mapped ID.

POI Information Table

Field	Type	Meaning	Example
district_hash	string	District hash	74c1c25f4b283fa74a5514307b0d0278
poi_class	string	POI class and its number	1#1:41 2#1:22 2#2:32

The POI Info Table shows the attributes of a district, such as the number of different facilities. For example, 2#1:22 means in this district, there are 22 facilities of the facility class 2#1. 2#1 means the first level class is 2 and the second level is 1, such as entertainment#theater, shopping#home appliance, sports#others. Each class and its number is separated by \t.

Traffic Jam Info Table

Field	Type	Meaning	Example
district_hash	string	Hash value of the district	1ecbb52d73c522f184a6fc53128b1ea1
tj_level	string	Number of road sections at different congestion levels	1:231 2:33 3:13 4:10
tj_time	string	Timestamp	2016-01-15 00:35:11

The Traffic Jam Info Table shows the overall traffic status on the road in a district, including the number of roads at different traffic jam levels in different time periods and different districts. Higher values mean heavier traffic.

Weather Info Table

Field	Type	Meaning	Example
Time	string	Timestamp	2016-01-15 00:35:11
Weather	int	Weather	7
temperature	double	Temperature	-9
PM2.5	double	pm25	66

The Weather Info Table shows the weather info every 10 minutes each city. The weather field gives the weather conditions such as sunny, rainy, and snowy etc; all sensitive information has been removed. The unit of temperature is Celsius degree, and PM2.5 is the level of air pollutions.