Affluence of taxis to Beijing's points of interest

77013 - Diego Hernandez
77076 - Rodrigo Pereira

Case Study

Is there any differences in affluences to certain points of interest along the day and/or week?

This is the research question we decided to study with the data provided in the class.

It is specially interesting because the time frame of the data provided coincides with the week of the chinese new year

Case Study

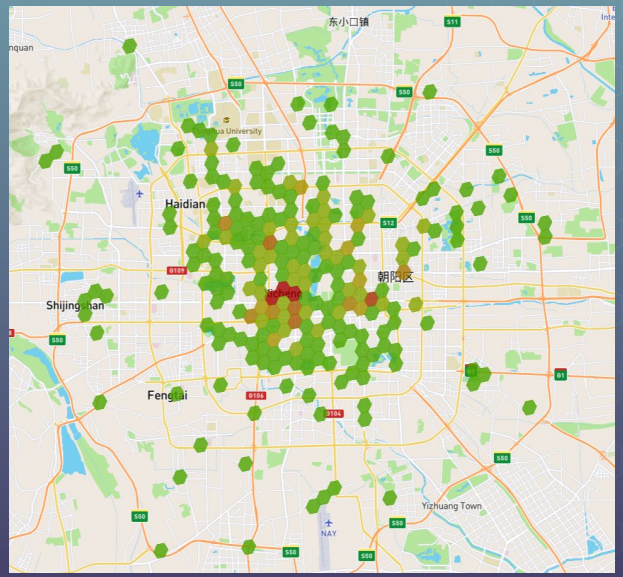
Religious places are popular at the time ?

Is there any differences in affluences to certain points of interest along the day and/or week?

Affluence to touristic places varies during the week

Methodology

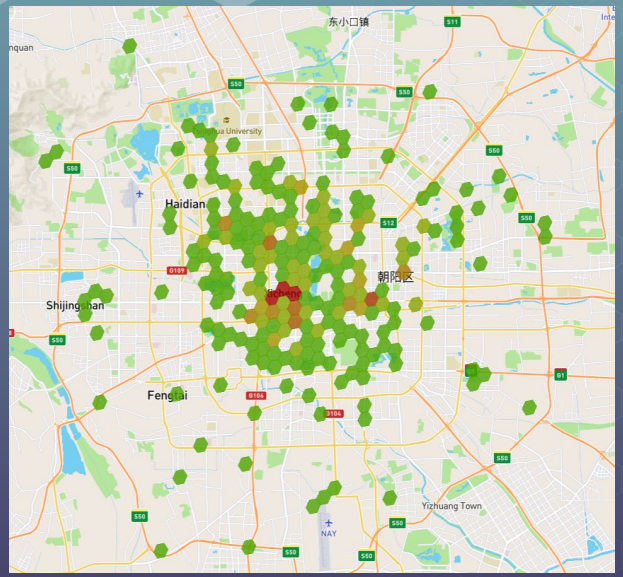
Approach to the problem



Since we wanted to work with counts of taxis near location, we decided that the best way to capture that concept was to divide the city of Beijing in small areas that would only contain a handful of locations and count the number of taxis at a certain point in time.

To tackle this problem we decided to use UBER's library "H3", that divides the globe into hexagons, with a parameterizable granularity (in our case we used hexagons that cover a city block).

Approach to the problem



- Divide the city of Beijing into small areas using Uber's H3 library
- Obtaining the number of taxis on the areas that contains the location we want to study
- Analyze results by types of point of interest

Pre processing

To achieve our goals we need to add additional information to the one provided.

For that purpose we mapped the coordinates of the given taxi to an H3 area, using the method **geoToH3(lat,long,zoom)**

We additionally restructured the given timestamps to include information about the day of the week.

```
tid,obsid,taxiid,segmentid,lon,lat,year,month,weekDayN,day,hour,minute,areaID
1,291748,1,8604,116.483,39.9195,2008,02,6,08,15,11,8831aa4667fffff
2,291749,1,699400,116.508,39.9313,2008,02,6,08,15,21,8831aa460bfffff
3,291750,1,10006,116.532,39.9154,2008,02,6,08,15,31,8831aa46e9fffff
4,291752,1,136649,116.547,39.9084,2008,02,6,08,15,51,8831aa468dfffff
5,291753,2,9222,116.364,39.8878,2008,02,0,02,13,33,8831aa4053fffff
6,291754,2,11356,116.375,39.8878,2008,02,0,02,13,37,8831aa42a5fffff
7,291755,2,11356,116.377,39.8879,2008,02,0,02,13,38,8831aa42a5fffff
8,291756,2,11357,116.38,39.888,2008,02,0,02,13,42,8831aa42a5fffff
9,291757,2,11352,116.394,39.8901,2008,02,0,02,13,43,8831aa42a3fffff
10,291758,2,179224,116.411,39.8915,2008,02,0,02,13,47,8831aa42b1fffff
11,291759,2,7706,116.421,39.8919,2008,02,0,02,13,48,8831aa476dfffff
12,291760,2,9120,116.422,39.8982,2008,02,0,02,13,52,8831aa4297fffff
13,291761,2,9120,116.422,39.899,2008,02,0,02,13,53,8831aa4297fffff
14,291762,2,11696,116.43,39.9073,2008,02,0,02,13,57,8831aa4293fffff
15,291763,2,1420,116.43,39.9118,2008,02,0,02,13,59,8831aa429bfffff
16,291764,2,1421,116.444,39.908,2008,02,0,02,14,02,8831aa4741fffff
17,291765,2,1421,116.444,39.9081,2008,02,0,02,14,04,8831aa4741fffff
18,291766,2,1421,116.444,39.9081,2008,02,0,02,14,04,8831aa4741fffff
19,291767,2,449850,116.455,39.9039,2008,02,0,02,14,09,8831aa4755fffff
20,291768,2,7378,116.455,39.8898,2008,02,0,02,14,12,8831aa4701fffff
21,291769,2,9301,116.455,39.8833,2008,02,0,02,14,14,8831aa4707fffff
22,291770,2,333972,116.46,39.8805,2008,02,0,02,14,17,8831aa4739fffff
23,291771,2,333972,116.458,39.8811,2008,02,0,02,14,19,8831aa4739fffff
24,291772,2,333972,116.458,39.8811,2008,02,0,02,14,19,8831aa4739fffff
```

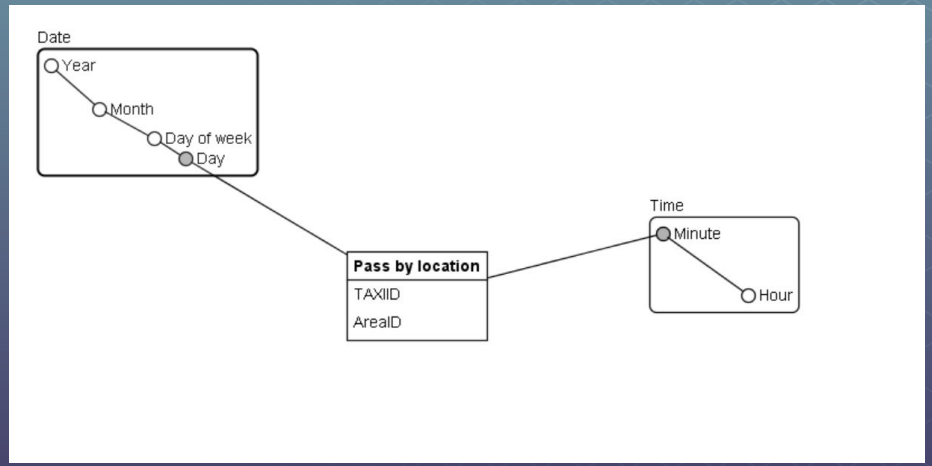
Pre processing

- Add additional H3's AreaID to given data;
- Add information about day of the week to given data;

```
gid,obsid,taxiid,segmentid,lon,lat,year,month,weekDayN,day,hour,minute,areaID
1,291748,1,8604,116.483,39.9195,2008,02,6,08,15,11,8831aa4667fffff
2,291749,1,699400,116.508,39.9313,2008,02,6,08,15,21,8831aa460bfffff
3,291750,1,10006,116.532,39.9154,2008,02,6,08,15,31,8831aa46e9fffff
4,291752,1,136649,116.547,39.9084,2008,02,6,08,15,51,8831aa468dfffff
5,291753,2,9222,116.364,39.8878,2008,02,0,02,13,33,8831aa4053fffff
6,291754,2,11356,116.375,39.8878,2008,02,0,02,13,37,8831aa42a5fffff
7,291755,2,11356,116.377,39.8879,2008,02,0,02,13,38,8831aa42a5fffff
8,291756,2,11357,116.38,39.888,2008,02,0,02,13,42,8831aa42a5fffff
9,291757,2,11352,116.394,39.8901,2008,02,0,02,13,43,8831aa42a3fffff
10,291758,2,179224,116.411,39.8915,2008,02,0,02,13,47,8831aa42b1fffff
11,291759,2,7706,116.421,39.8919,2008,02,0,02,13,48,8831aa476dfffff
12,291760,2,9120,116.422,39.8982,2008,02,0,02,13,52,8831aa4297fffff
13,291761,2,9120,116.422,39.899,2008,02,0,02,13,53,8831aa4297fffff
14,291762,2,11696,116.43,39.9073,2008,02,0,02,13,57,8831aa4293fffff
15,291763,2,1420,116.43,39.9118,2008,02,0,02,13,59,8831aa429bfffff
16,291764,2,1421,116.444,39.908,2008,02,0,02,14,02,8831aa4741fffff
17,291765,2,1421,116.444,39.9081,2008,02,0,02,14,04,8831aa4741fffff
18,291766,2,1421,116.444,39.9081,2008,02,0,02,14,04,8831aa4741fffff
19,291767,2,449850,116.455,39.9039,2008,02,0,02,14,09,8831aa4755fffff
20,291768,2,7378,116.455,39.8898,2008,02,0,02,14,12,8831aa4701fffff
21,291769,2,9301,116.455,39.8833,2008,02,0,02,14,14,8831aa4707fffff
22,291770,2,333972,116.46,39.8805,2008,02,0,02,14,17,8831aa4739fffff
23,291771,2,333972,116.458,39.8811,2008,02,0,02,14,19,8831aa4739fffff
24,291772,2,333972,116.458,39.8811,2008,02,0,02,14,19,8831aa4739fffff
```

Temporary dimensional model

- Necessity for temporary dimensional model to hold our csv data, that reflected the event of passing by an area.
- Use of Microsoft SSIS tools to do it

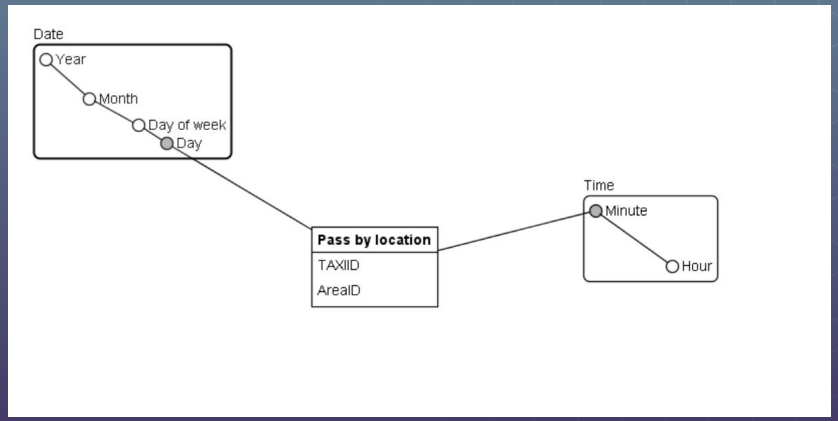


Temporary dimensional model

After the pre processing, we had a csv, so we figured that we should translate that information to a dimensional model.

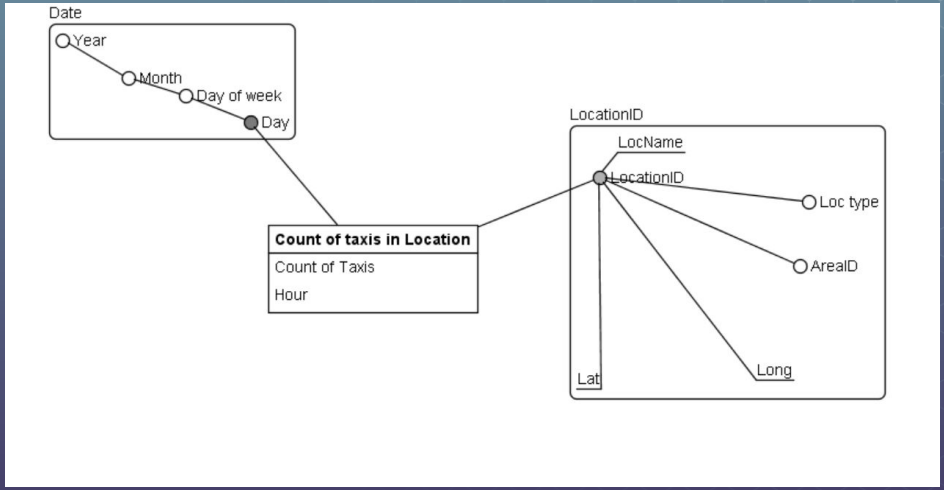
For that we used Microsoft's SISS tool.

An event model, that translates the passage of a taxi in a certain area.



Definitive Dimensional Model

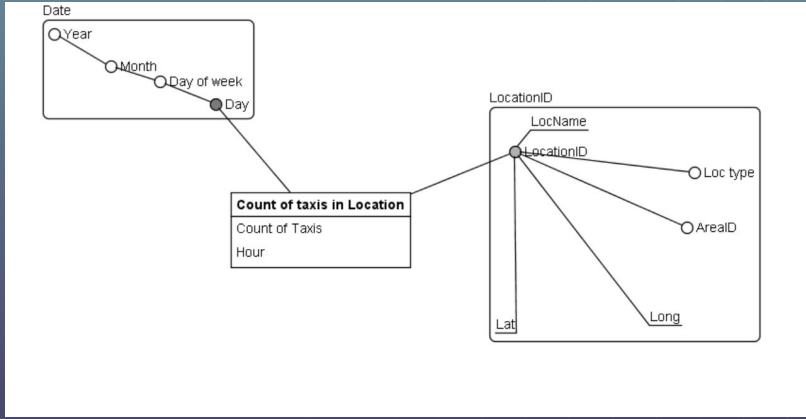
- Use google's API to gather Points of interests and their types
- Join point of interest data with the the one stored in ou temporary data model to get data mart that could give us counts of taxis by hour on different locations



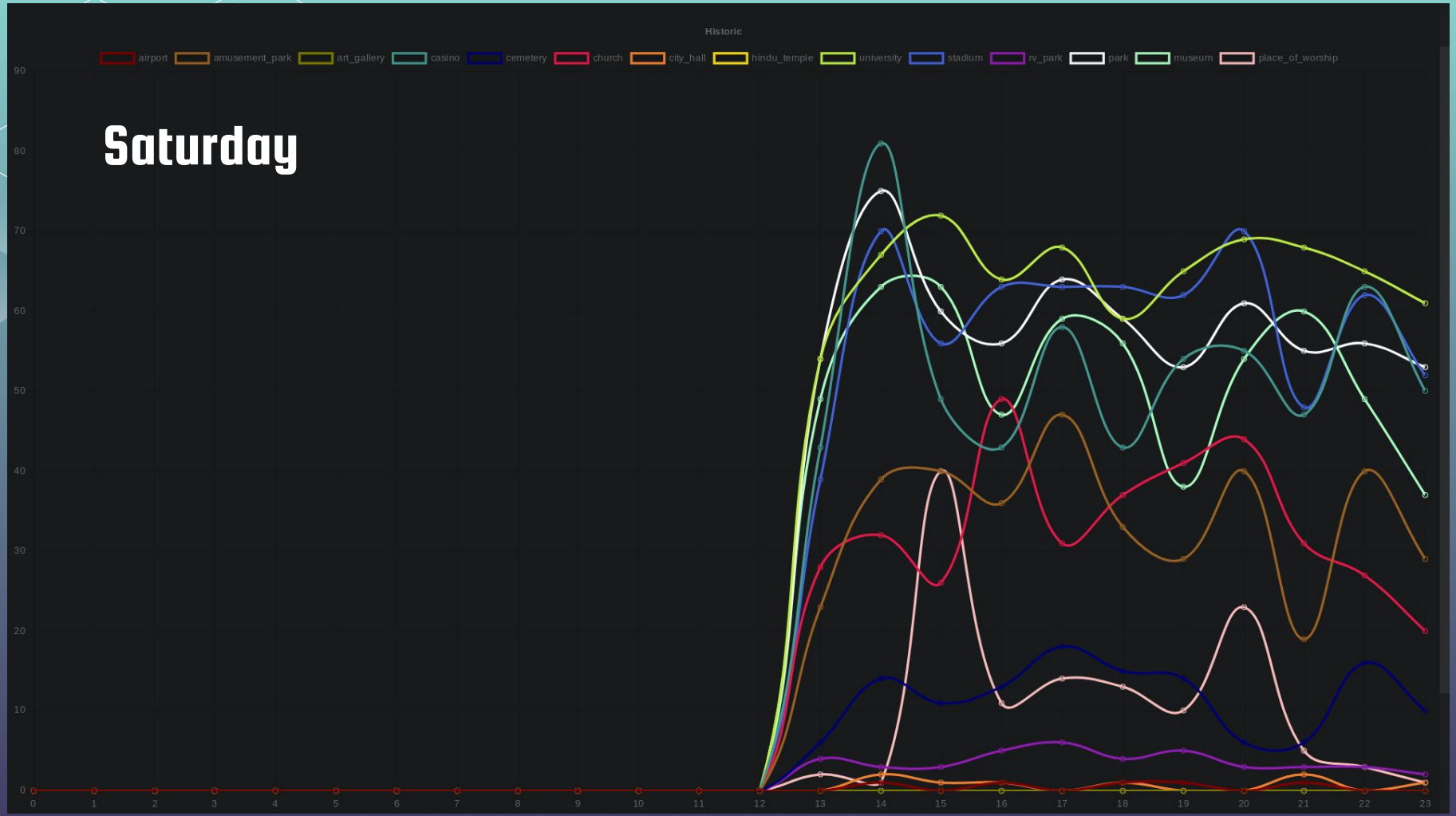
Definitive Dimensional Model

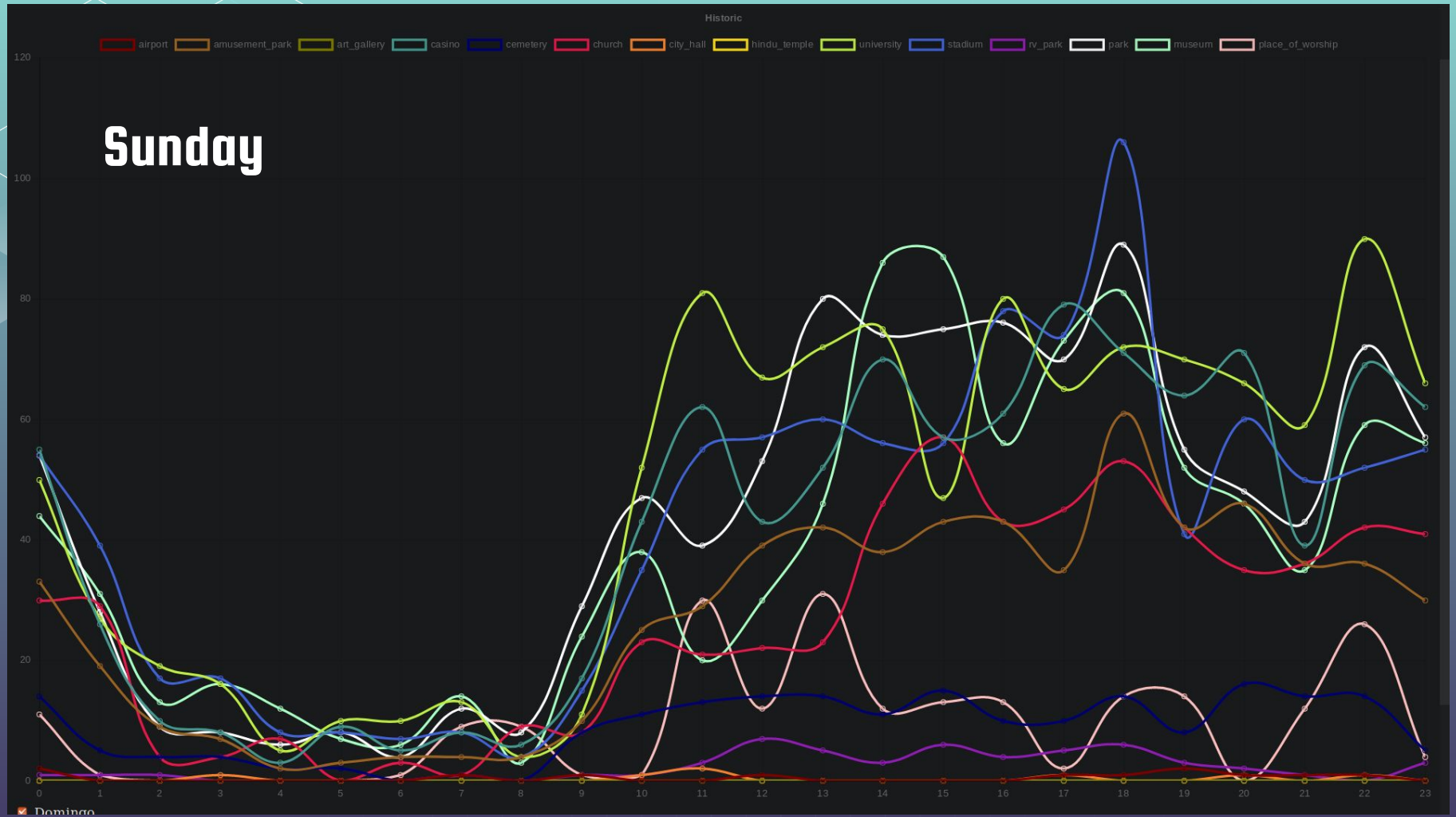
To gather our points of interest we made use of the google's API, to get a list of these points, with their coordinates, their types and their names. The following were mapped to the respective H3 areas.

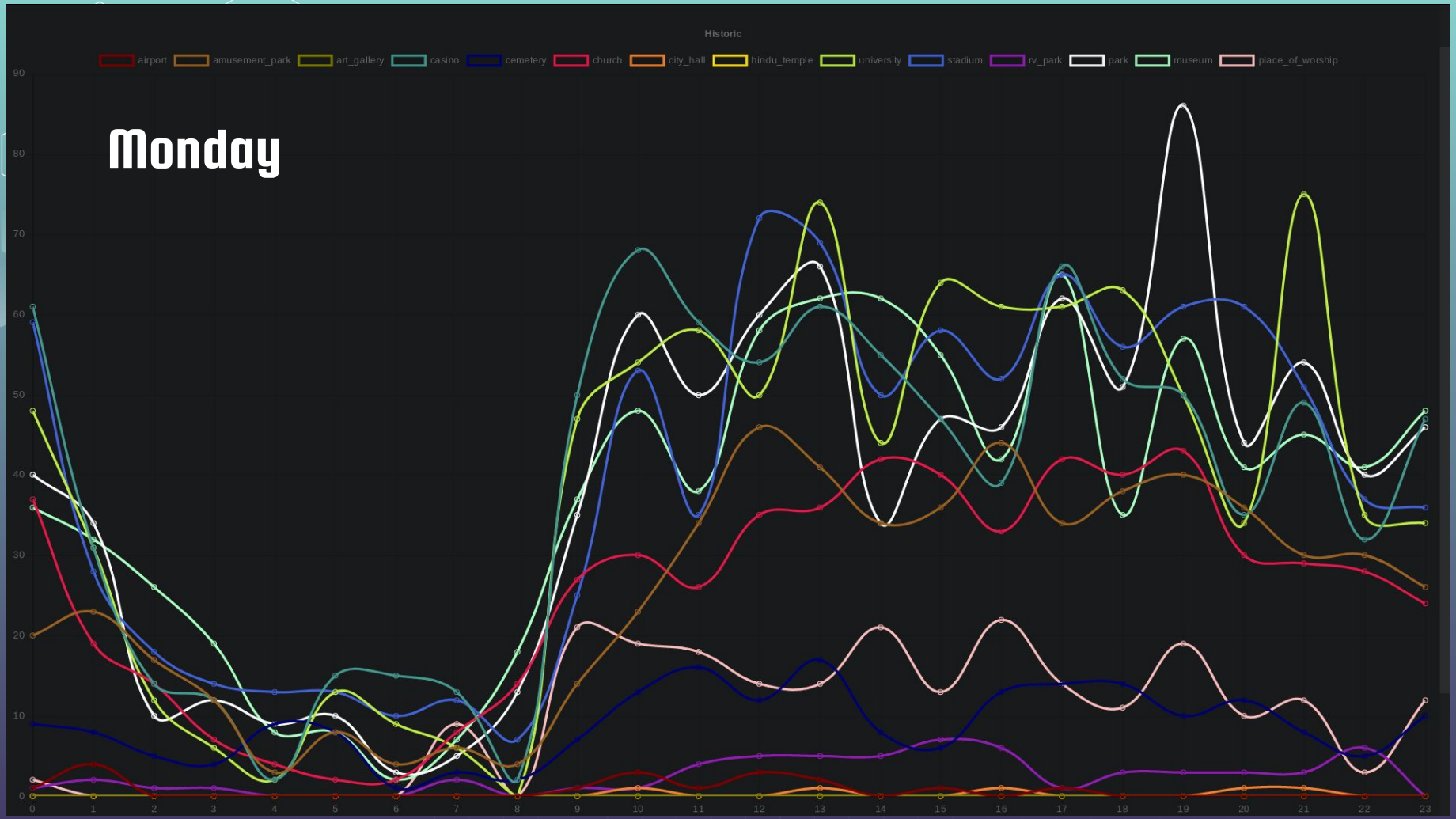
We then joined this data to the data present on our "temporary" dimensional model. To generate the final dimensional model that was going to be able to answer our questions

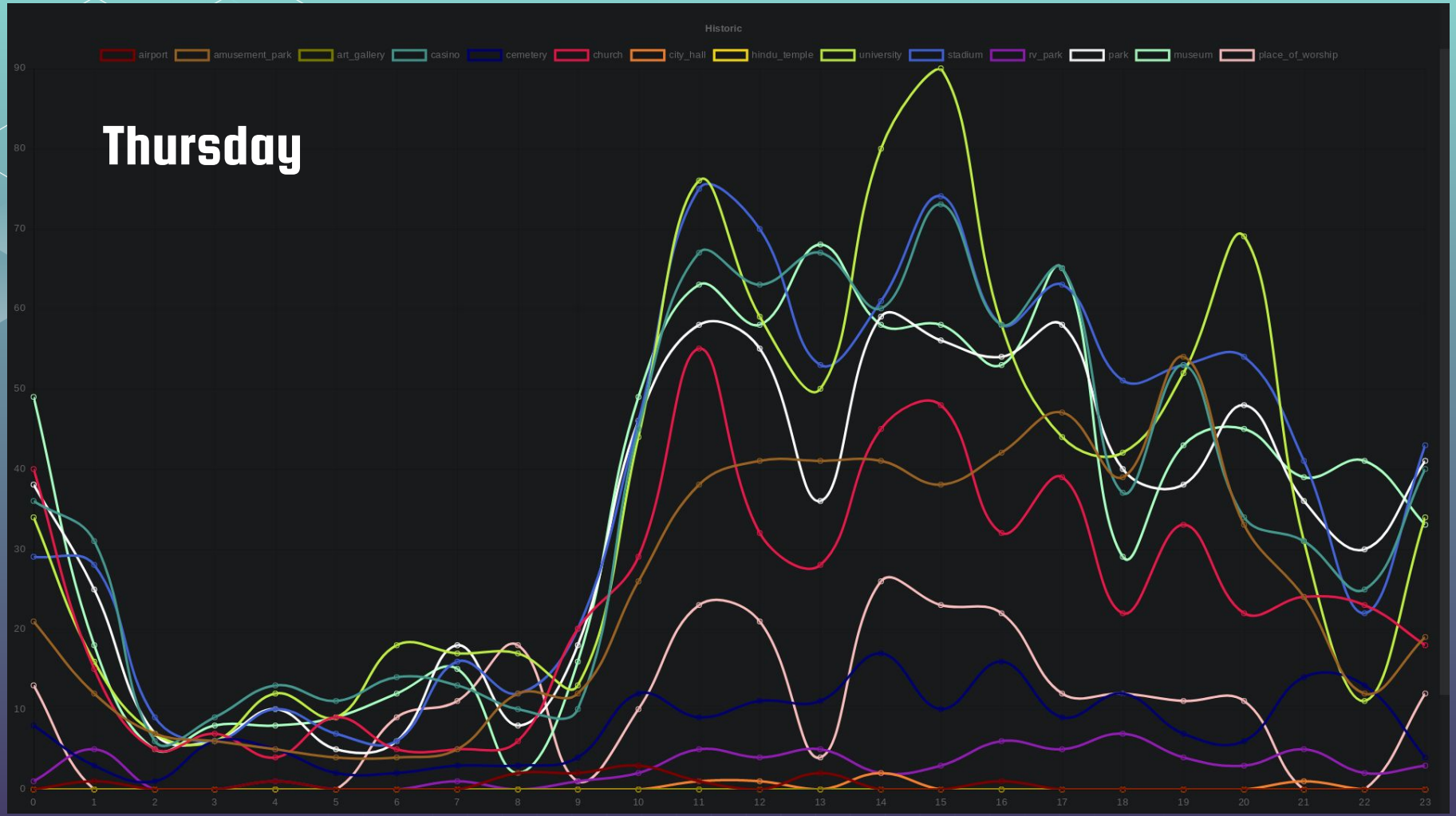


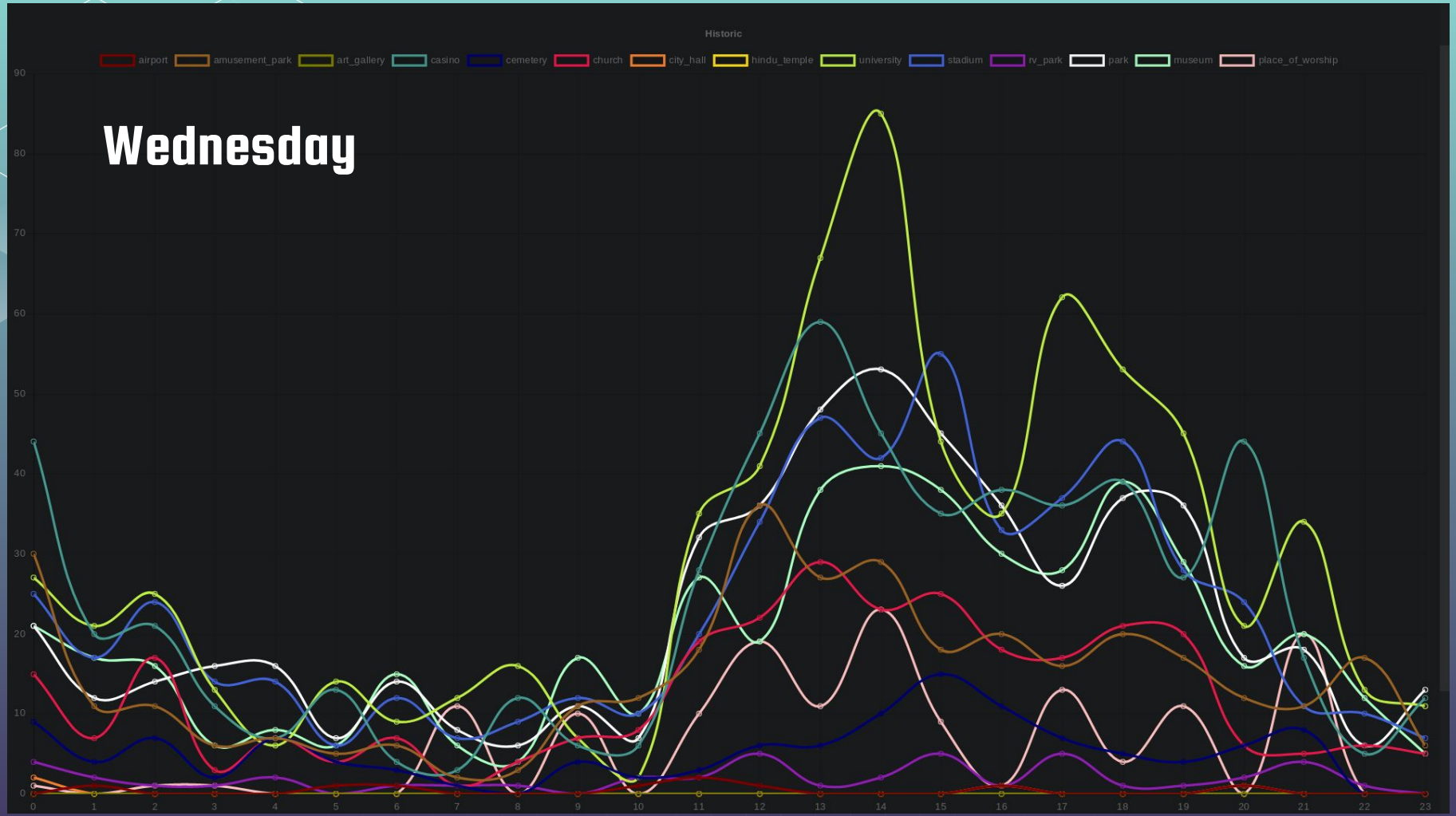
Analysis

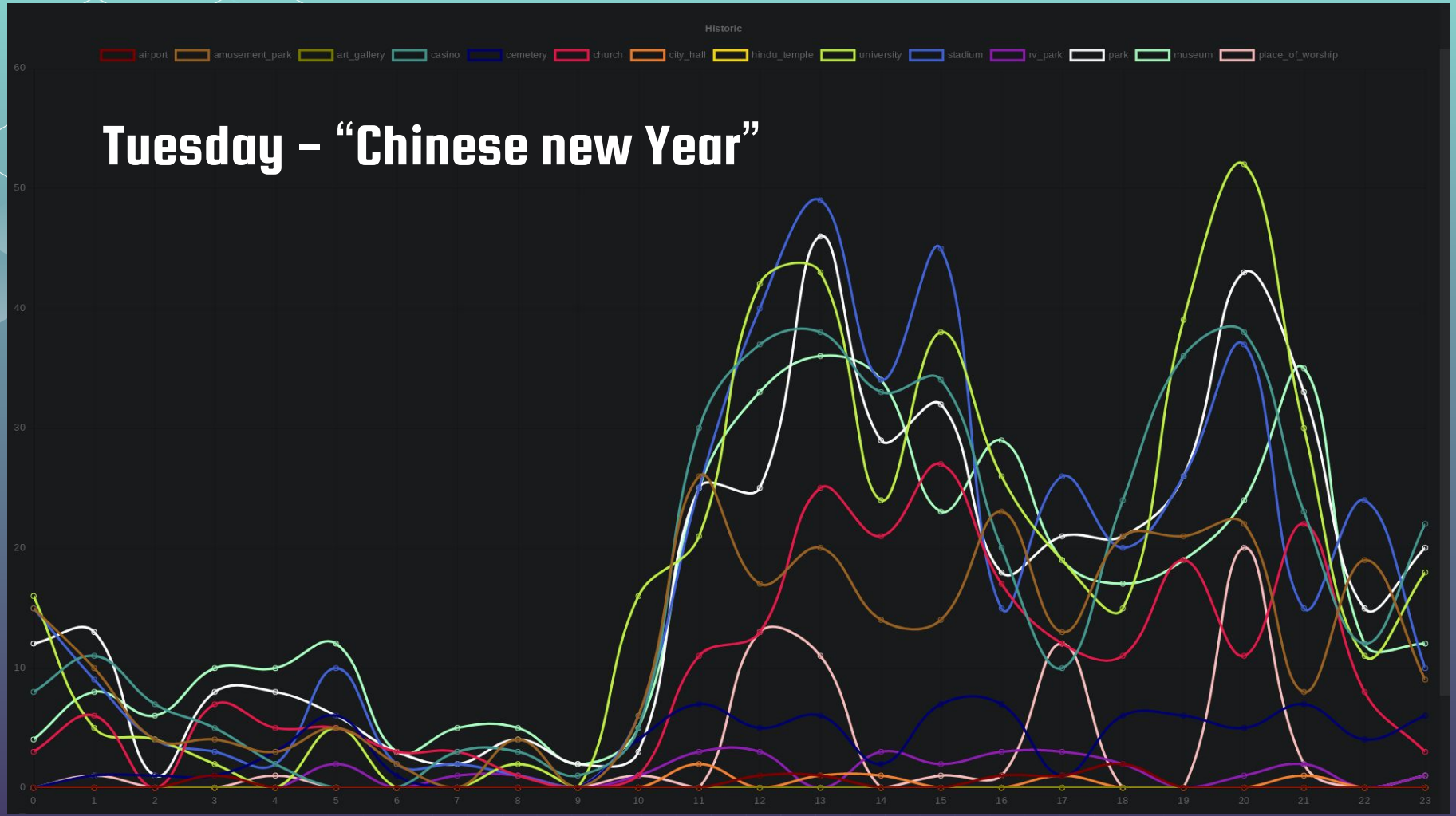

Historic
airport
amusement_park
art_gallery
casino
cemetery
church
city_hall
hindu_temple
university
stadium
rv_park
park
museum
place_of_worship
Saturday
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23


Historic
airport
amusement_park
art_gallery
casino
cemetery
church
city_hall
hindu_temple
university
stadium
rv_park
park
museum
place_of_worship
Sunday
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Domingo

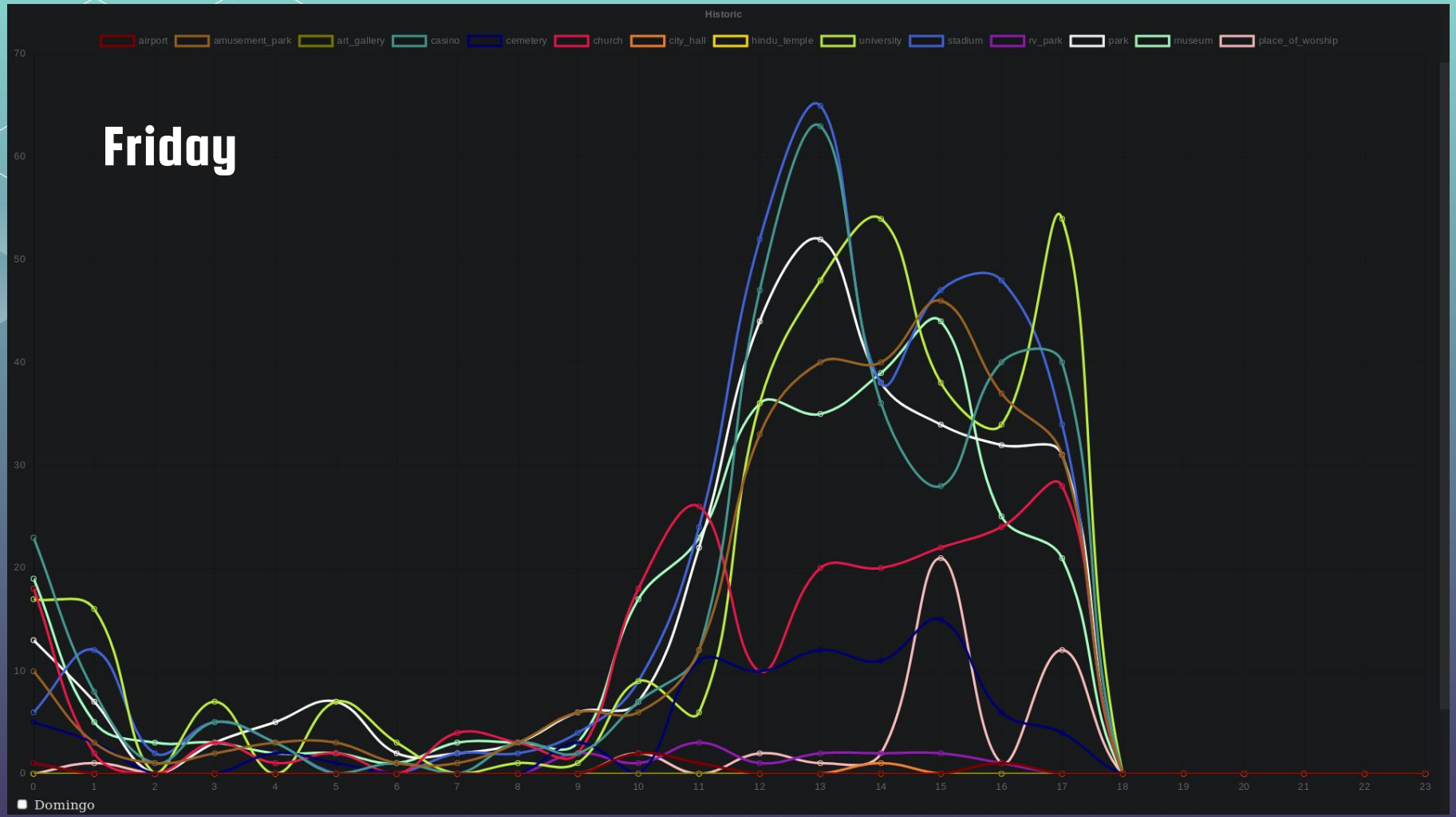

Historic
airport
amusement_park
art_gallery
casino
cemetery
church
city_hall
hindu_temple
university
stadium
rv_park
park
museum
place_of_worship
Monday
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

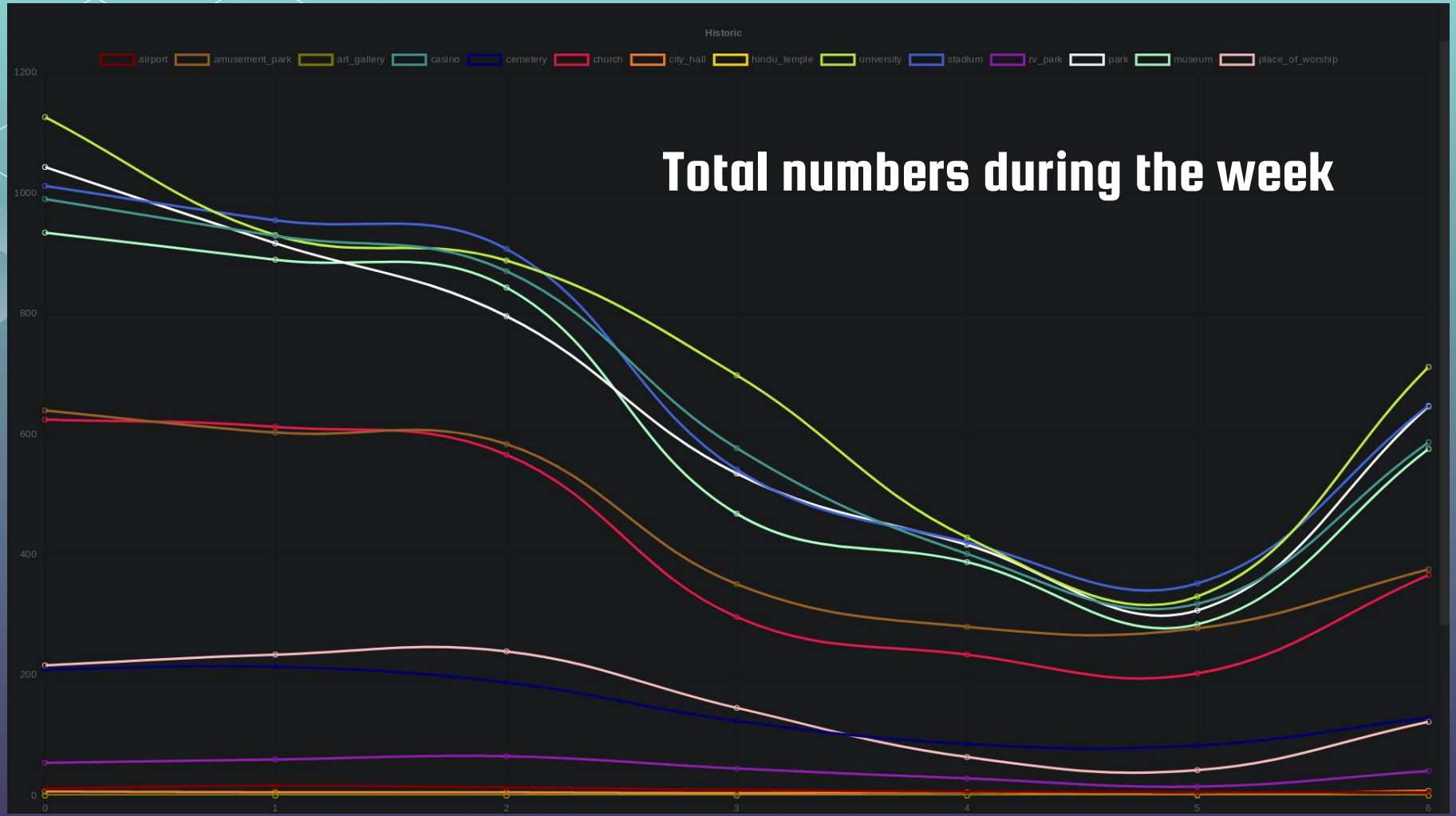

Historic
airport
amusement_park
art_gallery
casino
cemetery
church
city_hall
hindu_temple
university
stadium
rv_park
park
museum
place_of_worship
Thursday
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23


Historic
airport
amusement_park
art_gallery
casino
cemetery
church
city_hall
hindu_temple
university
stadium
rv_park
park
museum
place_of_worship
Wednesday
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23



Tuesday – “Chinese new Year”


Historic
airport
amusement_park
art_gallery
casino
cemetery
church
city_hall
hindu_temple
university
stadium
rv_park
park
museum
place_of_worship
Friday
70
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Domingo

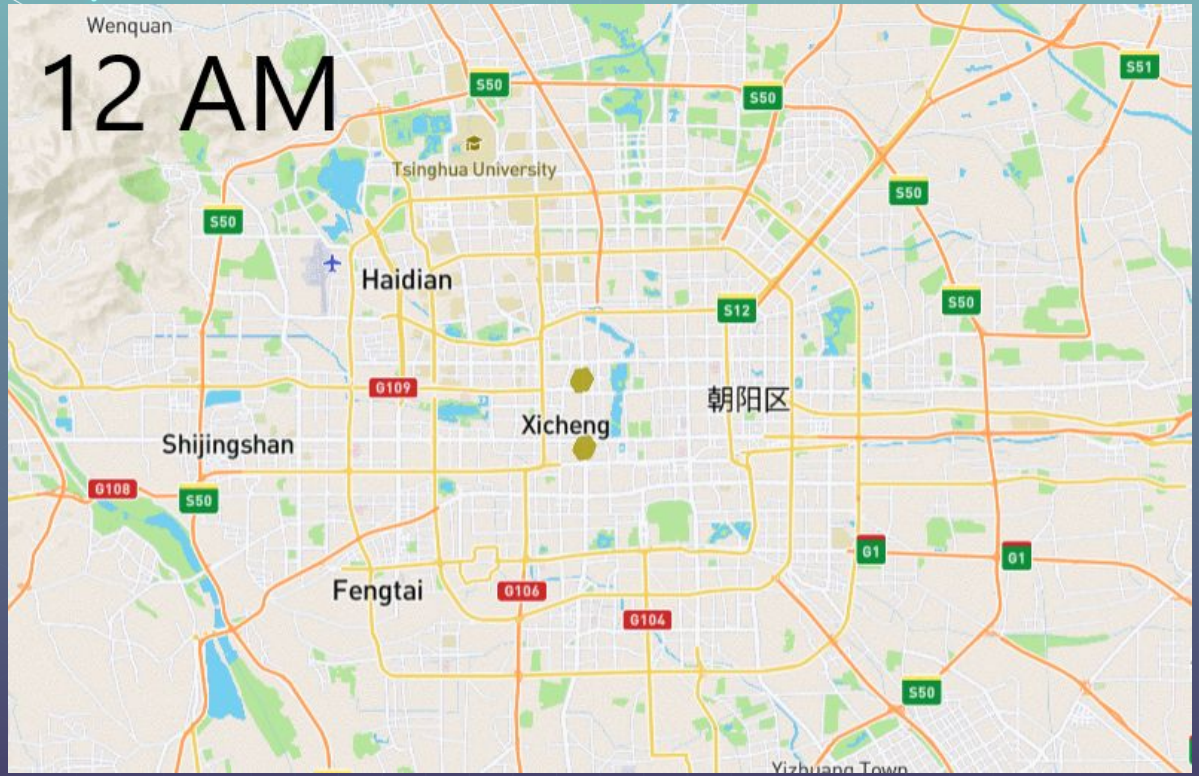


Total numbers during the week

Affluence to different types of POIs

- One aspect that is obvious looking at the data is that it starts being read at 12 PM of 2 February (Saturday) and stops being read adt 6 PM of 8 February (Friday);
- The late night hours are the hours with less taxis circulating overall (despite existing some with still some affluence like casinos)
- The type of locations with more taxi affluence near them are overall universities (probably due to an elevate concentration of them, 65 in total, 34 unique areas with at least one)
- The data shows a relevant dip in the total number of taxis near points of interest around the time of the start of the chinese new year (tuesday), probably since its a day where most people stay at home with family
- Despite having small variances on the affluences of the different types, it does seem to follow a pattern along the week

Affluence to Universities during the day



Conclusions

Our hypothesis was partially supported by the data obtained and we can still infer some more interesting facts about it, some of those can be interesting especially to the taxi industry. The most obvious one is the points of interest with most affluence, what can help taxists better decide on places to search for passengers.