Project Report

SHUYI ZHANG

1. PREPATATION

The first step is to upload the Police_Incidents_cleaned.tsv file into VirtualBox using command:
scp -P 2222 /User/zoe/Desktop/Police_Incidents_cleaned.tsv root@127.0.0.1:~/

In order to manage data directly and clearly, I created a file named “Police_project” to write and save all sequent files. Command: mkdir Police_project

After changing the Police_Incidents_cleaned.tsv into Police_project, I created another createT.hive file to name all variables of Police_Incidents_cleaned.tsv
Command1: mv Police_Incidents_cleaned.tsv ./Police_project/
Command2: vi createT.hive

2. PROBLEM SOLUTION

Q1) As the first question I was curious is whether the type of location is relevant to the number of crimes, especially in the weekdays and weekends.

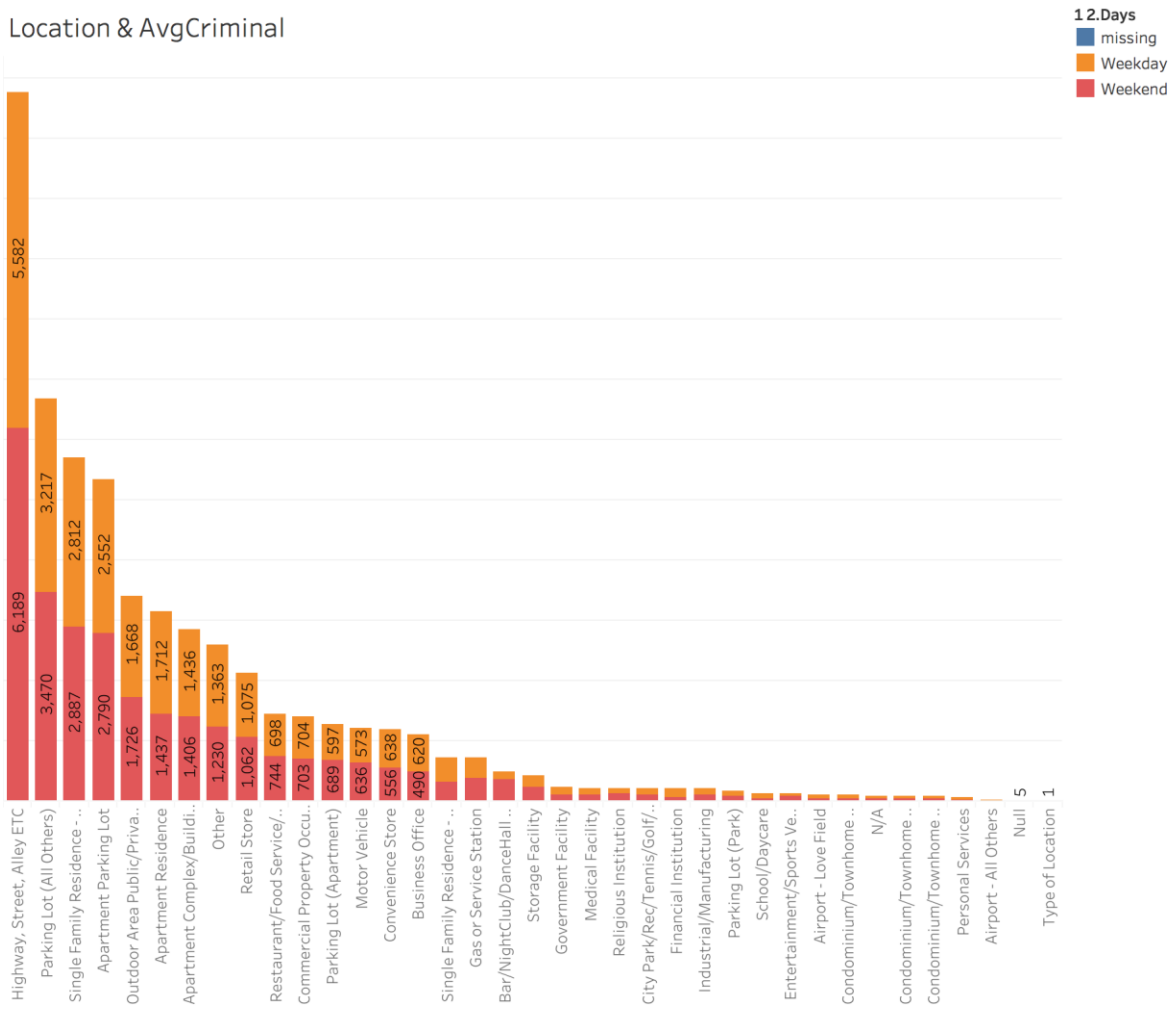
Process:

Create a “LocaVSCrime.hive” file: Group Type Of Location , and Day1Week variables and using “case” “when” queries to find out the average criminals from different type of locations in weekends and weekdays.

Run hive -f LocaVSCrime.hive , then using “vi view1_2.hive” command to create a “1_2.hive” file defining a table with variable name. Then creating .tsv format of the table. Command:
hive -f view1_2.hive > /root/Police_project/1_2.tsv

Note: all documents with name “number_number.hive” are to creating a table with variables’ name.
All documents with name “number_number.tsv” are the result table of each question.

Results:



As we can see above, Highway a street in Alley ETC is the highest area of incidents, which happened 11,771 times in total. But the frequent incident ratio of weekends and weekdays in each area is rough half to half. Therefore, whether weekends or weekdays basically has nothing to do with the number of incidents. But places like Highway. Street in Alley ETC, Parking lot(excluding Apt Parking lot), and Single family residence are target happening incidents.

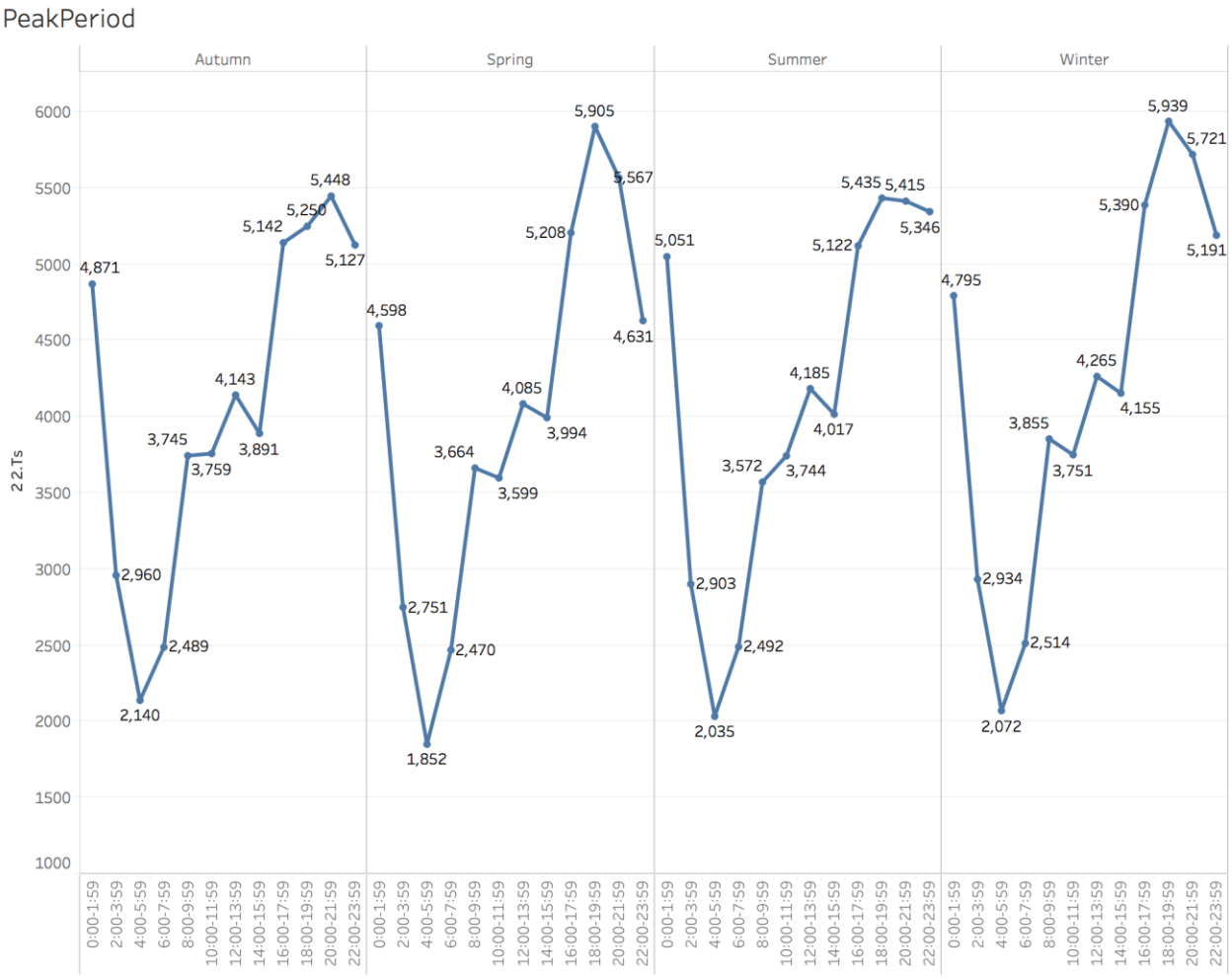
Q2) To see if peak incident period is the same or roughly around specific time period of a day in each season.

Process:

Create a “peakPeriod.hive” file to classify the incidents numbers’ trend in each season by dividing a day into 12 periods. And group total incidents in each time period based on each season.

Run hive -f peak_period.hive then using “vi 2_2.hive” command to create a “2_2.hive” file defining a table with variable name. Then creating .tsv format of the table. Command:
hive -f 2_2.hive > /root/Police_project/2_2.tsv

Results:



According to the above map, incidents’ numbers in 4:00-5:59am period are in off-peak in each season. And then the incidents fluctuate with an upward trend until getting peak in 20:00-21:59 period in Autumn, while in 18:00-19:59 period in other seasons. But incidents in 18:00-19:59 period gets 2nd high in Autumn. Consequently, the peak period is almost the around the same period.

Q3) To find out if the type of incidents is relevant to other factors, like the local fertility in each area (based on zip code).

Process:

I found a resource about the fertility in Dallas area using fertility related zipcode. from the website https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

Then making the data cleaning by only saved the data of total women, women had a birth in the past 12 months (including unmarried teenagers) in each area, and zip code.
And join 2 tables using zip code.

Result:

Fertility

3 2.Zipcode		
76018	3 2.ZipcodeCounts	1
	3 2.Hadbirth	354
	3 2.Hadbirunmateen	10
	3 2.Hadbirunmarried	129
	3 2.Nothadbir	7,480
	3 2.Total	7,834

Unfortunately, the result turned out with only 1 zip code, at first, I thought there was a issue in the process of my commands, but after comparing 2 files' zip code, there is only one same zip code. And I searched all other resources in different years from the U.S. census website, the zip code resources are all the same.

With only 1 result, it is hard to tell anything.

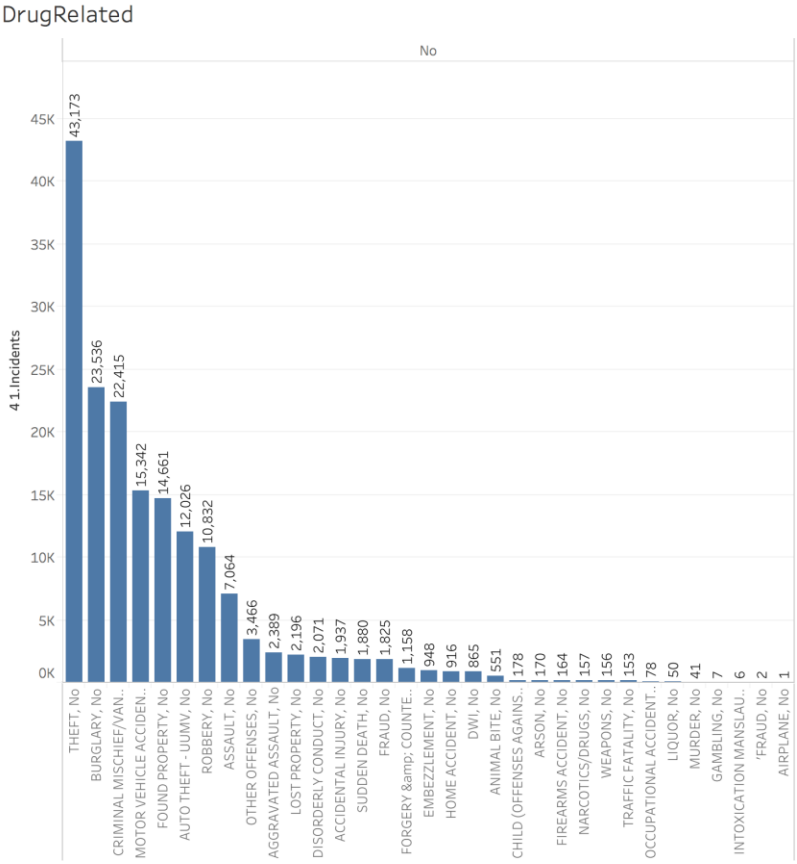
Q4) Assumed that the drug is related to the type of incidents. Compare to the type of no-drug-related incidents.

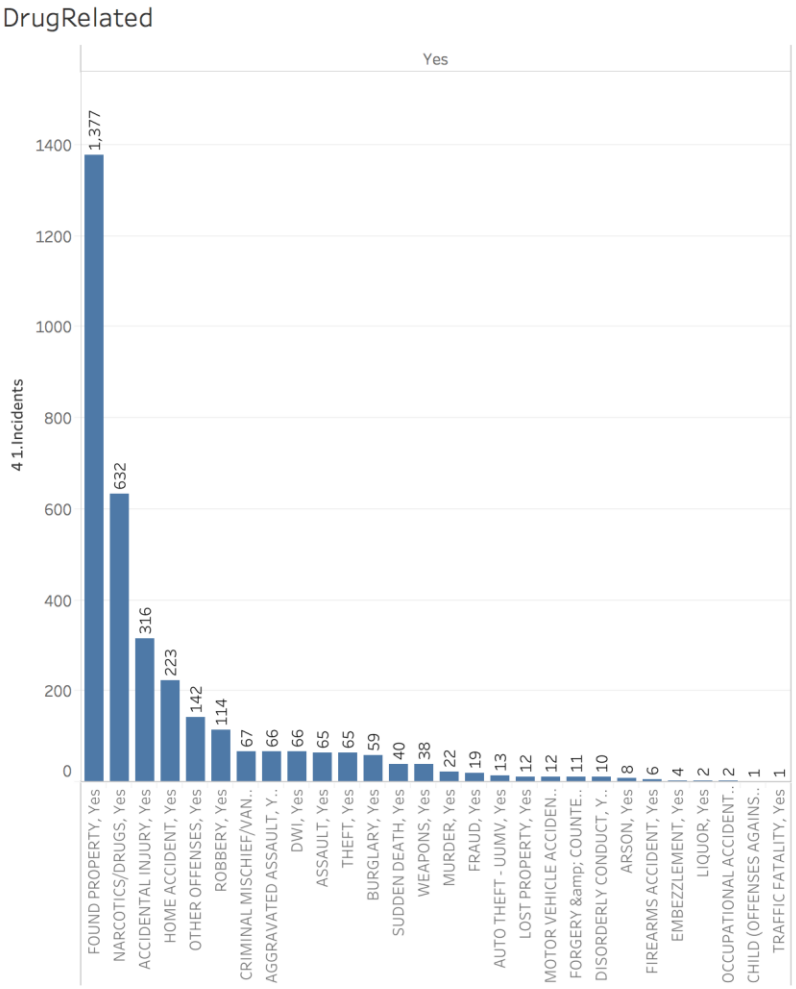
Process:

Create a file named "drugRelate.hive". Select 2 variables about the type of incidents and drug related incidents, and compare what types of incidents relating to the drugs with non-drug related incidents.

Run hive -f drugRelate.hive then using “vi 4_1.hive” command to create a “4_1.hive” file defining a table with variable name. Then creating .tsv format of the table. Command: hive -f 4_1.hive > /root/Police_project/4_1.tsv

Results:





From the table, the criminal type, found property, relates to the drug, and the criminal numbers are almost the total amount of other drug-related criminals. And theft happened with largest times in the whole criminal numbers, and it is no-drug related criminals.

Q5) To figure out whether the type of incident is involved with a particular time period.

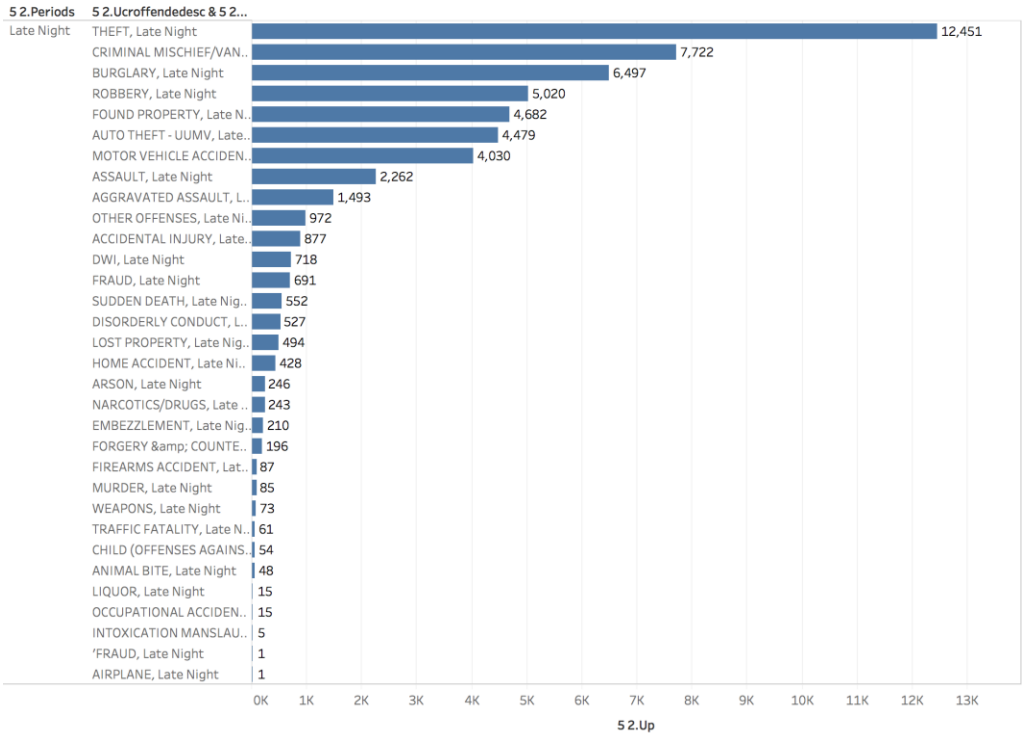
Process:

Create a file named “timeRelaType.hive”. Select 2 variables about the type of incidents and time period regarding as “late night”, “day”, “evening” by using variable “Day1Occur”.

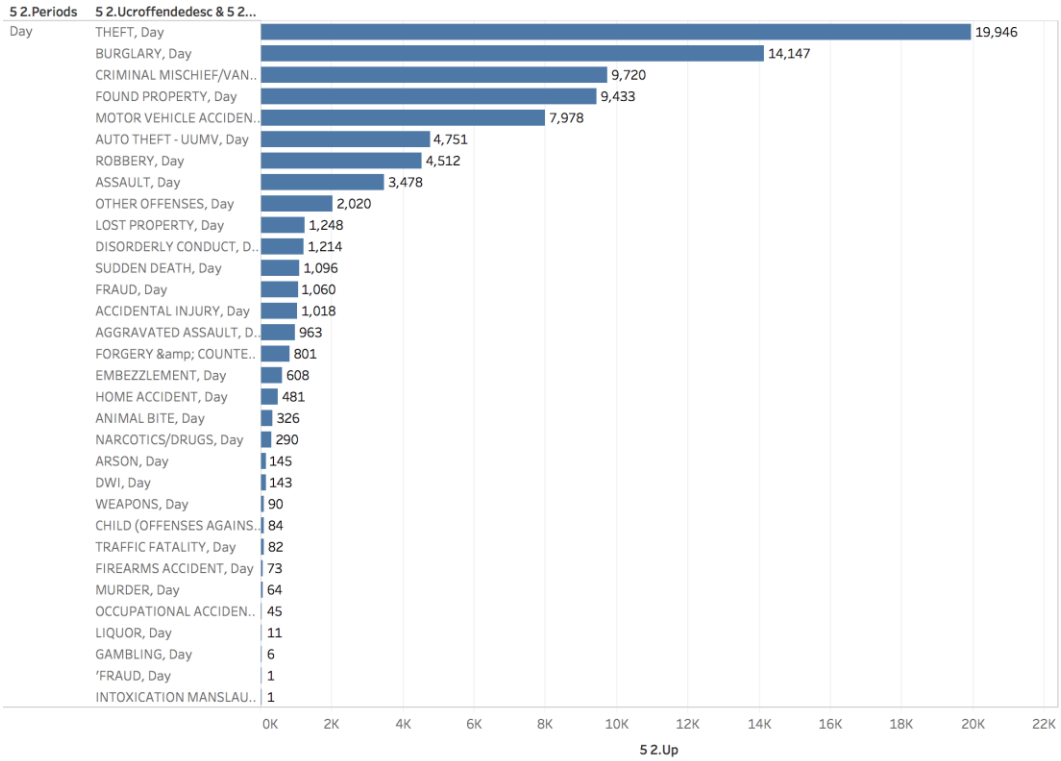
Run hive -f timeRelaType.hive then using “vi 5_2.hive” command to create a “5_2.hive” file defining a table with variable name. Then creating .tsv format of the table. Command:
hive -f 5_2.hive > /root/Police_project/5_2.tsv

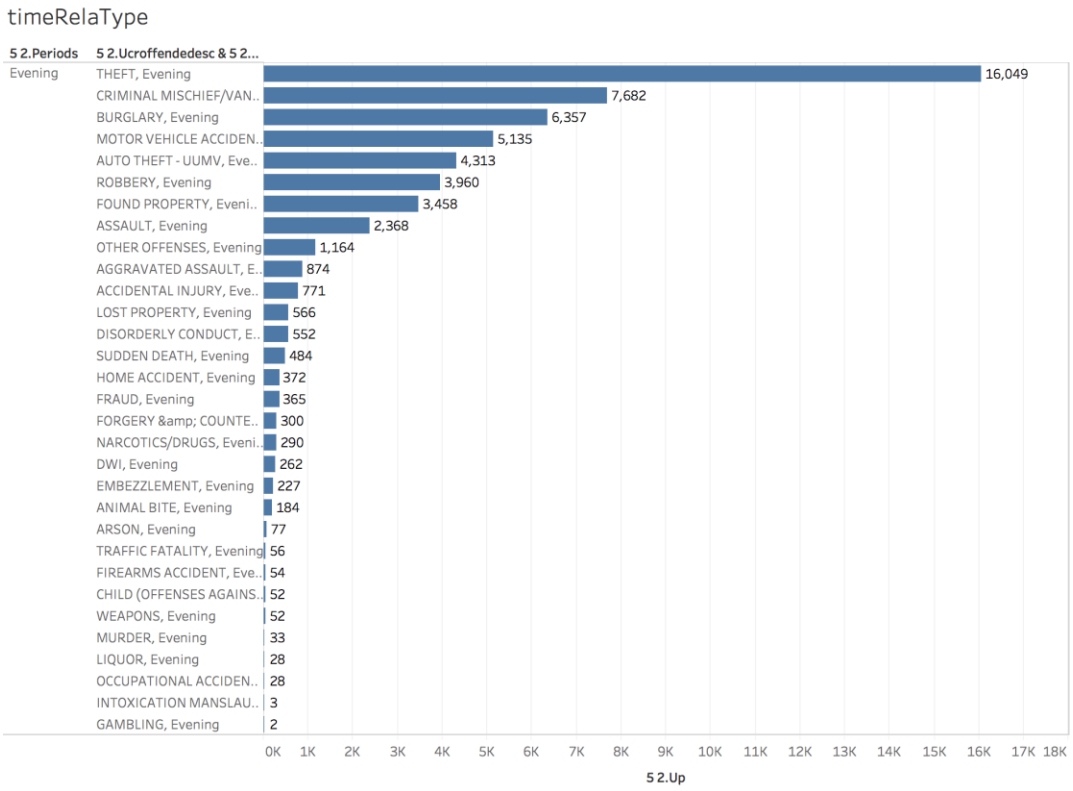
Results:

timeRelaType



timeRelaType





Interestingly, whenever the time, theft is the most happened crimes.