NYC Crime Analysis

IBM Data Science Capstone Project

Tengyu Zhang

2020-12-24

Introduction Problem

It is important to analize the crime data for the police system as well as the public so that we can know the trend and locations of the crimes in order to get preparied for them, or, more importantly, find ways to decrease the amount.

This project is trying to answer the following questions by analyzing the crime data of NYC:

1. What are the most common crime types? Which borough had the most crime number?
2. Did some of the efforts of the police system and the social education system work? What is the trend of crimes declined?
3. What is the crime trend during a year? which month is the peak of crime?
4. Which area has the most concentrated crime that needs special attention? Is this area related to it's economic prosperity?
5. Can the crime trend of 2020 be predicted?

About the DATA

The data was collected and will be used as follows:

Data name	Collected from	Used for
NYPD_Complaint_Data_Historic.csv	NY Open Data	Analize the crime features and trends
NYC venues information	Foursquare	Analize the relationship between crime and economic prosperity

There are a lot of information in the Complaint Data, here we mainly focus on some of them to answer the questions, ie. date, description, borough

name and location(latitude and longitute). The "date" is used for analize the number and trend. The crime type can be known from "description". From "borough name" we can get the idea of which aera is more dangerous to draw the public's attention. The"location" info can be used to get the visualization on maps. In order to simply evaluate and visualize the economic prosperity, the Foursquare API is used to acquire venue information. The economic prosperity is evaluated based on the desicity of venues on the map when serching the whole city.

Answer Questions

(Methodology, Results and Discussion)

Due to the large size of the data, only data after year 2012 was used. Let's try to analyze the data to answer the following questions.

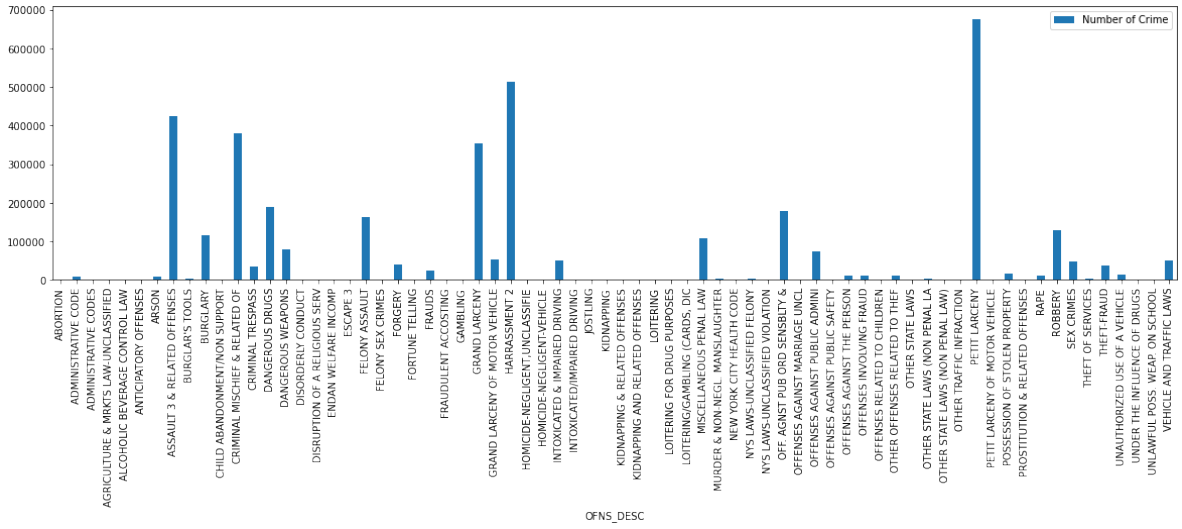
1. **What are the most common crime types? Which borough had the most crime number?**

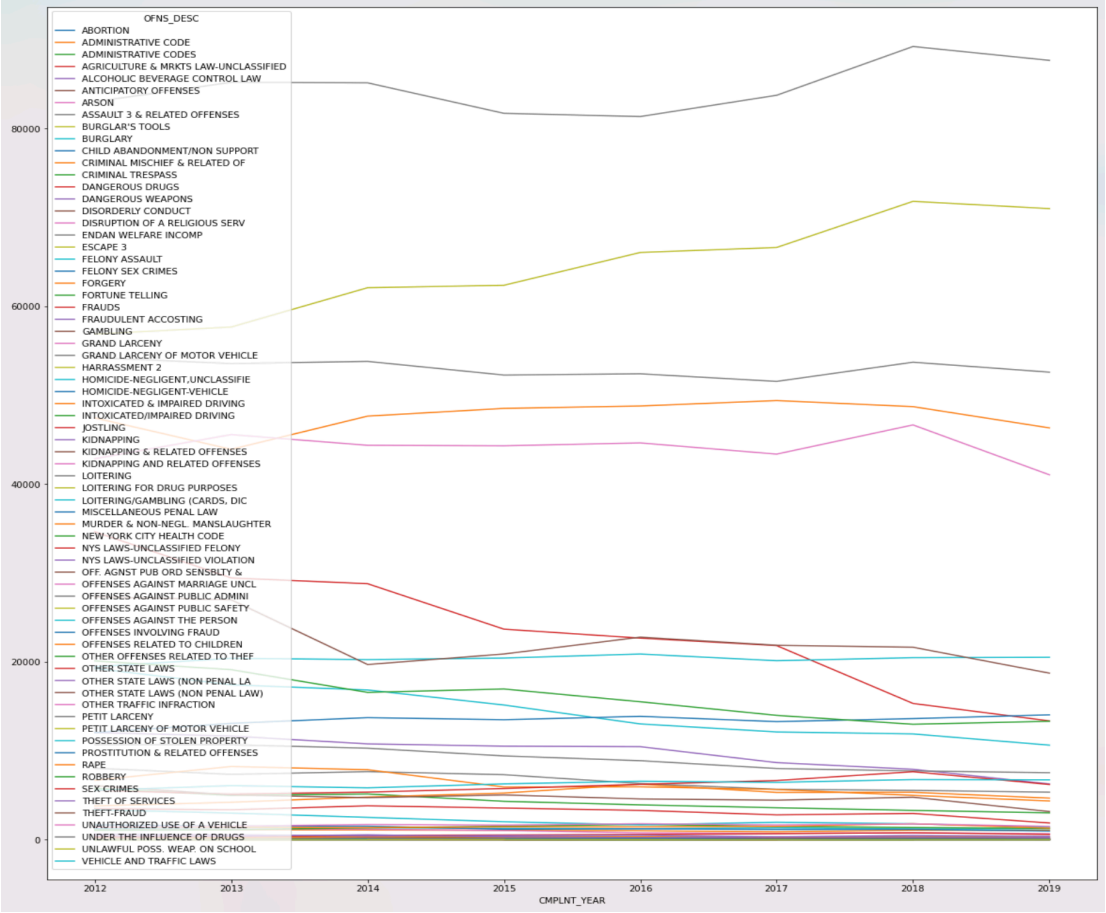
For the first question, we can simply group the data by crime type and borough and then summarize them.

The result of crime types is as follow:

```
PETIT LARCENY                       676860
HARRASSMENT 2                       514354
ASSAULT 3 & RELATED OFFENSES        424031
CRIMINAL MISCHIEF & RELATED OF      380670
GRAND LARCENY                       352753
                                     ...
FORTUNE TELLING                          4
KIDNAPPING AND RELATED OFFENSES          3
OFFENSES AGAINST MARRIAGE UNCL           2
UNDER THE INFLUENCE OF DRUGS             2
LOITERING FOR DRUG PURPOSES              1
Name: OFNS_DESC, Length: 70, dtype: int64
```

The result shows that “Petit larceny” is the most common crime type. And the top 5 types are petit larceny, harassment 2, assault 3 & related offences, criminal mischief & related of and grand larceny. Let’s visualize it:





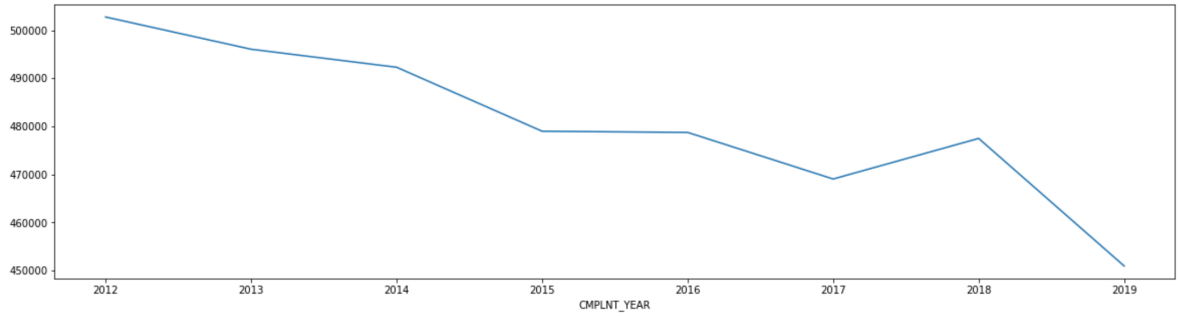
The result of borough that has the most crime is as follow:

	BORO_NM	count
1	BROOKLYN	1144279
2	MANHATTAN	926791
0	BRONX	833774
3	QUEENS	764509
4	STATEN ISLAND	174446

From the result we know that Brooklyn is the most dangerous borough.

2. Did some of the efforts of the police system and the social education system work? What is the trend of crimes declined?

For this question, we can count the number by year and draw a figure to visualize it. The result is as follows:

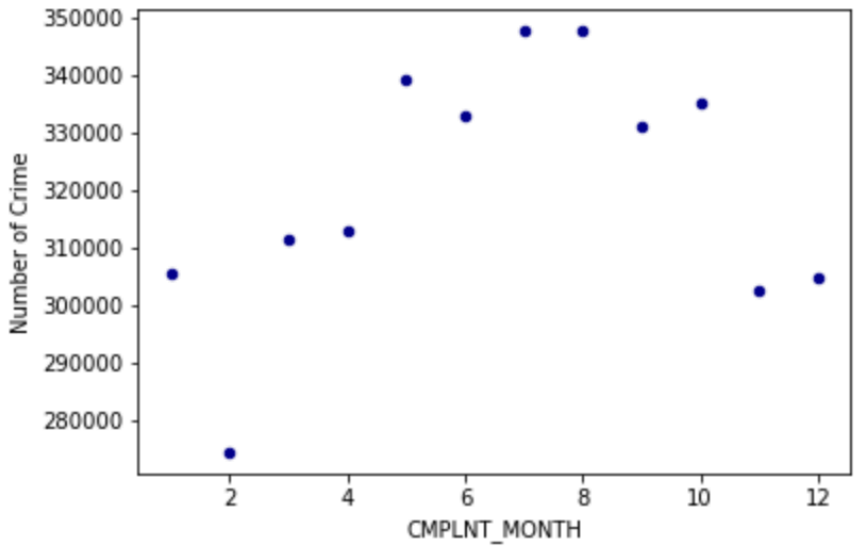


The figure shows that there is an obvious trend that the crime number is descending year by year, which tells that the efforts of the police system and the social education system do work!

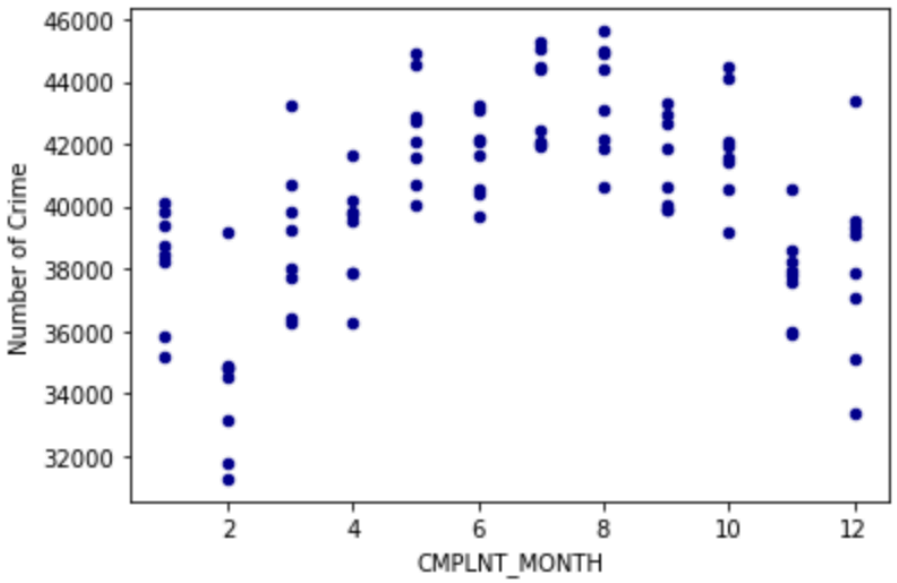
3. What is the crime trend during a year? which month is the peak of crime?

First, let's count the total number of each month since 2012 and draw a scatter plot.

	CMPLNT_MONTH	Number of Crime
0	8	347684
1	7	347652
2	5	339394
3	10	335357
4	6	332877
5	9	331210
6	4	312878
7	3	311535
8	1	305770
9	12	304898
10	11	302668
11	2	274565



Let’s do more work, count the number of each month year by year, and draw a scatter plot.



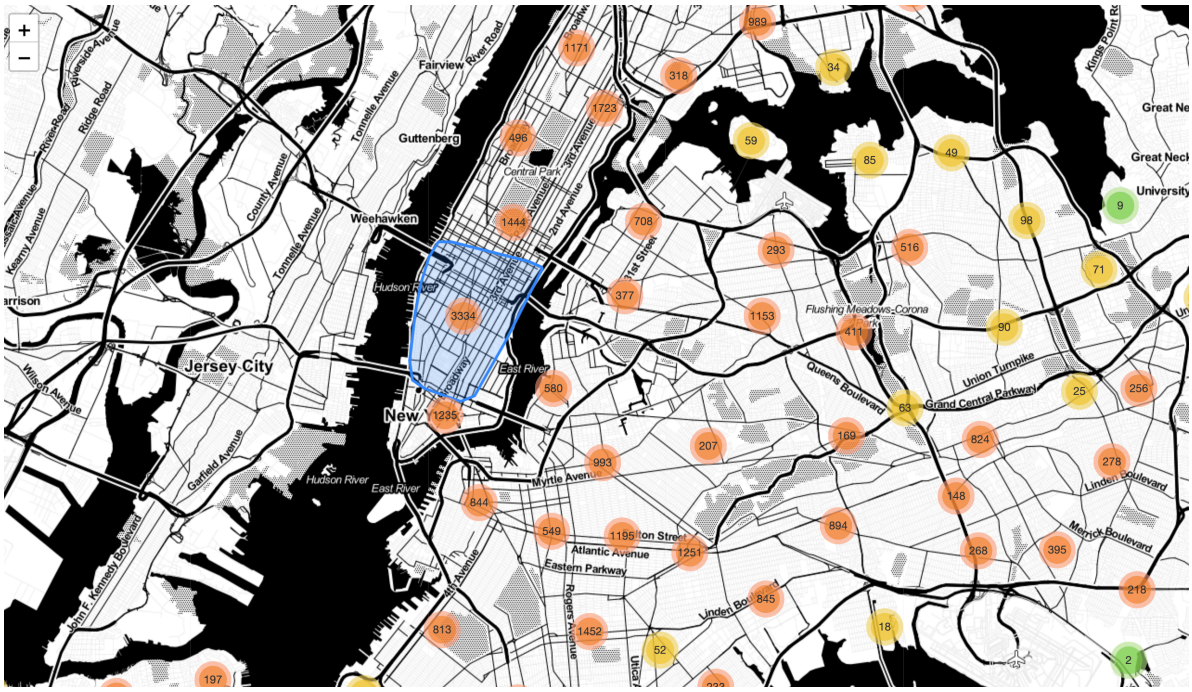
From the above chart and figures we know that there is a curve trend through out a year. And the peak of a year is July and August, which is the busiest time of the police system.

4. Which area has the most concentrated crime that needs special attention? Is this area related to it's economic prosperity?

For this question, we need to draw a map.

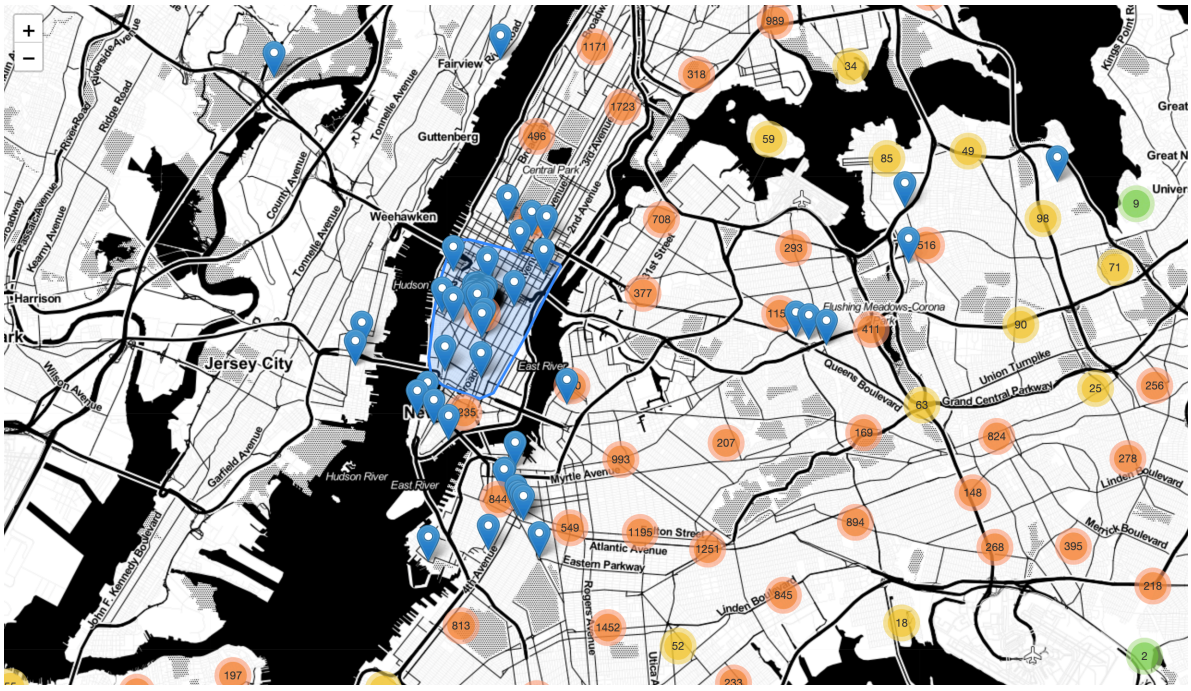
Geopy library is used to get the latitude and longitude values of New York City. And then import folium library to visualize the data on the map.

Due to the large amount of data, let's focus on Aug of 2019 to get the visualization. And we cannot visualize all the cases due to the large amount, so we can cluster them for better visualization.



Next, foursquare API is used to search the whole city, get the top 100 venues. Here we simply evaluate the economic prosperity by the density of the returned venues.

The venues and clustered crimes are showed on the map.



From the above map we can conclude that the area which has the most top venues has the most concentrated crime. Therefore, crime has a direct relationship with economic prosperity.

5. Can the crime trend of 2020 be predicted?

Two ways are used here to predict the number of 2020. The first is using the historic data grouped by month, then draw a curve to fit the data to figure out the trend during a year. This curve is the average prediction of 2020. The second way is using the most recent 2 years data.

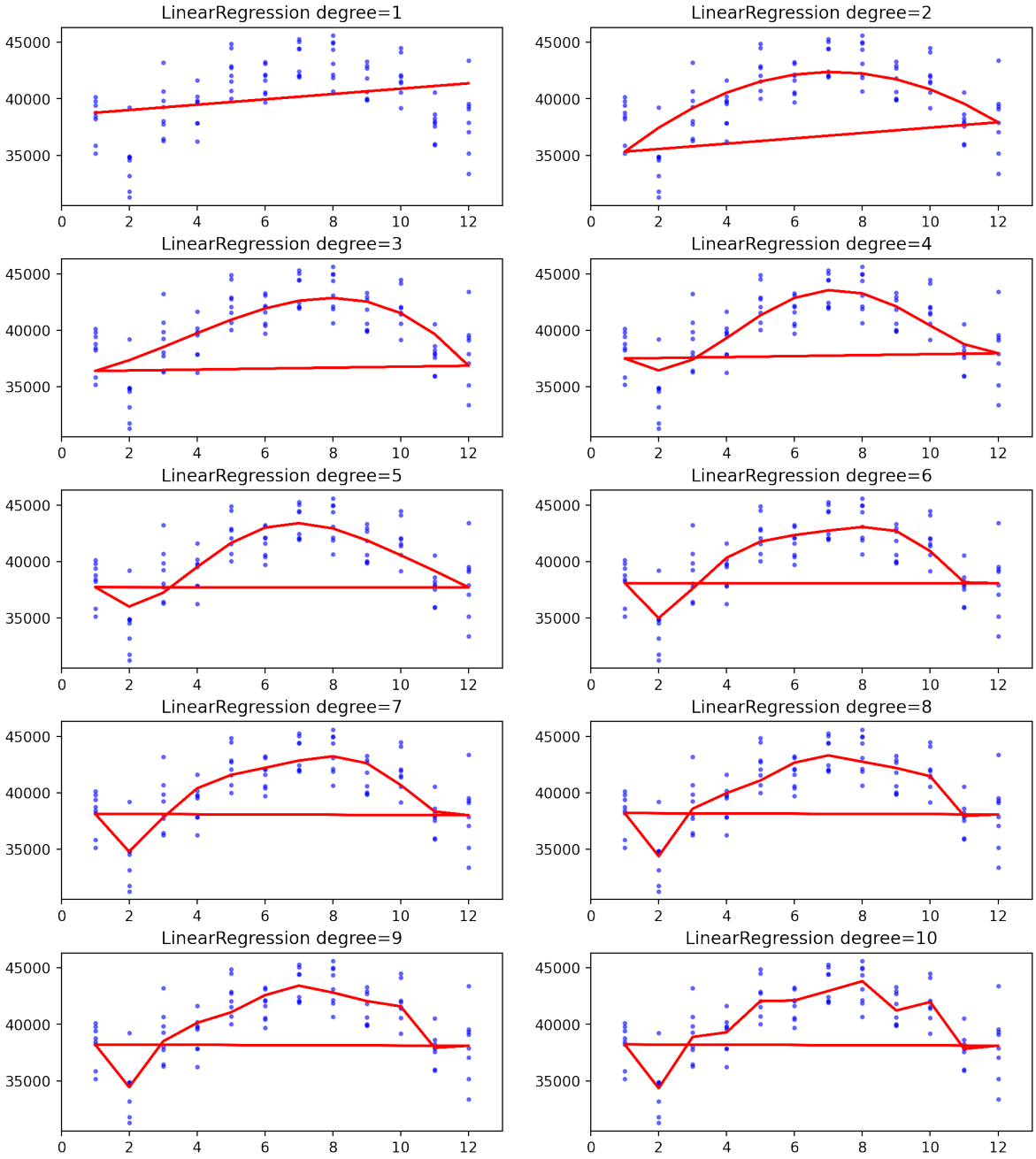
From the trend figure we know that the trend is not linear throughout the year. So polynomial regression methodology is used.

For the first way of using historic data, let's try the model with degree from 1 to 10 to find the most fit.

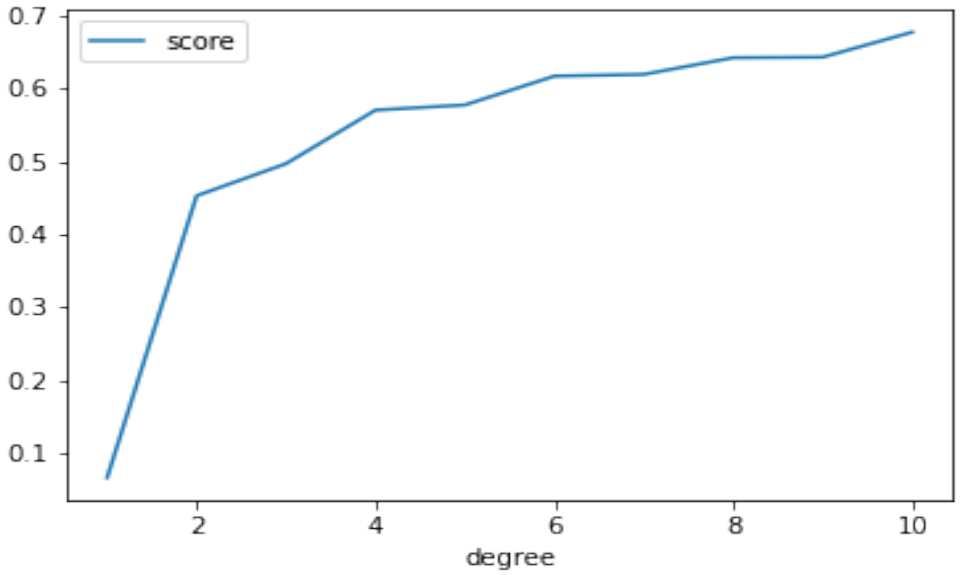
First, get the score and MSE for each degree.

```
degree: 1; train score: 0.06558032532203195; mean squared error: 9506880.509946218
degree: 2; train score: 0.45283034984523385; mean squared error: 5566959.497597453
degree: 3; train score: 0.4969764217050640 6; mean squared error: 5117812.886574367
degree: 4; train score: 0.5705710496119784; mean squared error: 4369053.680571869
degree: 5; train score: 0.5776630468574382; mean squared error: 4296898.981546839
degree: 6; train score: 0.6172927633491955; mean squared error: 3893702.227946779
degree: 7; train score: 0.6196882748307984; mean squared error: 3869330.051254703
degree: 8; train score: 0.642561285357248; mean squared error: 3636617.7230894826
degree: 9; train score: 0.6433352693265352; mean squared error: 3628743.13171258
degree: 10; train score: 0.6777380689903367; mean squared error: 3278725.55987143
```

Then, draw the fitted curve with different degrees.

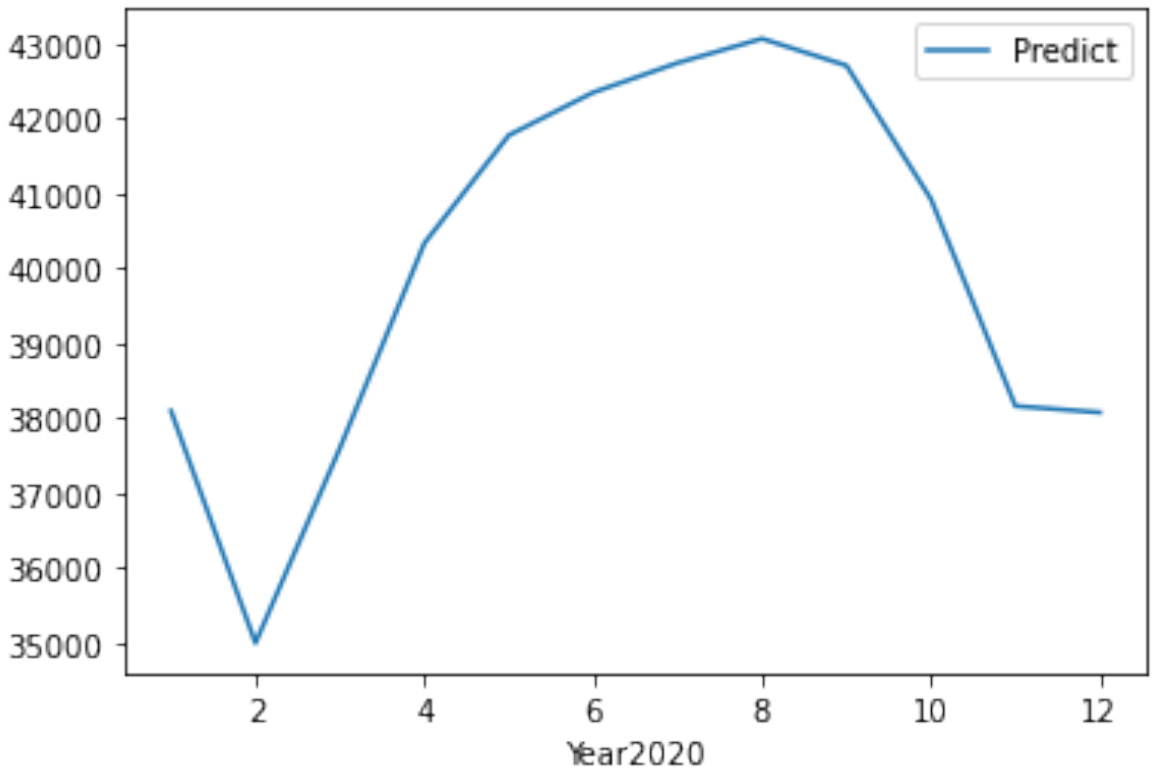


Next, draw a figure to visualize the scores with different degrees.



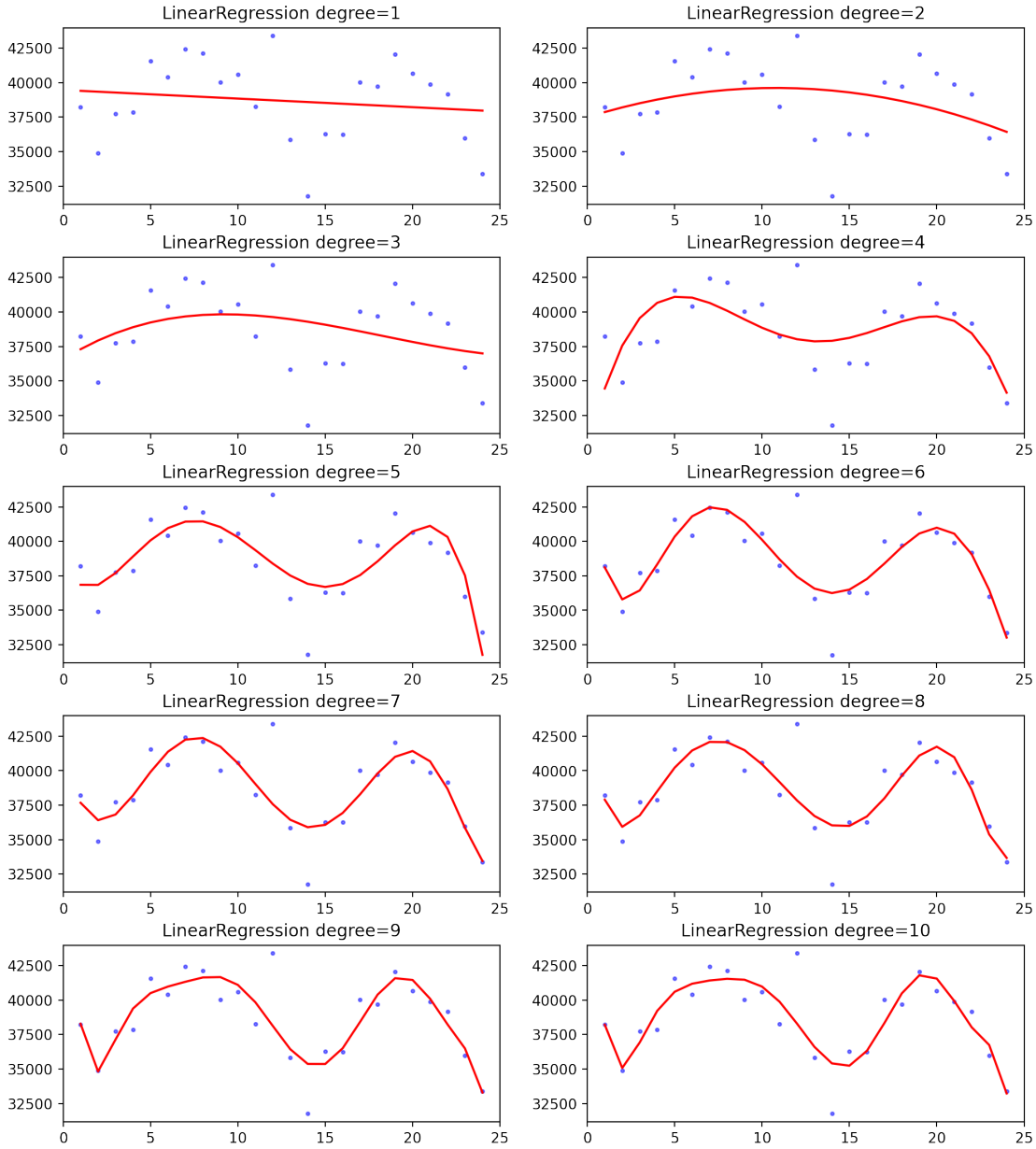
Since that a high degree may cause overfitting, we select degree 6. Let's get the prediction of 2020 with degree 6.

	Year2020	Predict
0	1	38093.359328
1	2	34994.544745
2	3	37598.049784
3	4	40332.958507
4	5	41775.122438
5	6	42347.341843
6	7	42741.893211
7	8	43065.402929
8	9	42706.067168
9	10	40923.217964
10	11	38159.235498
11	12	38073.806585

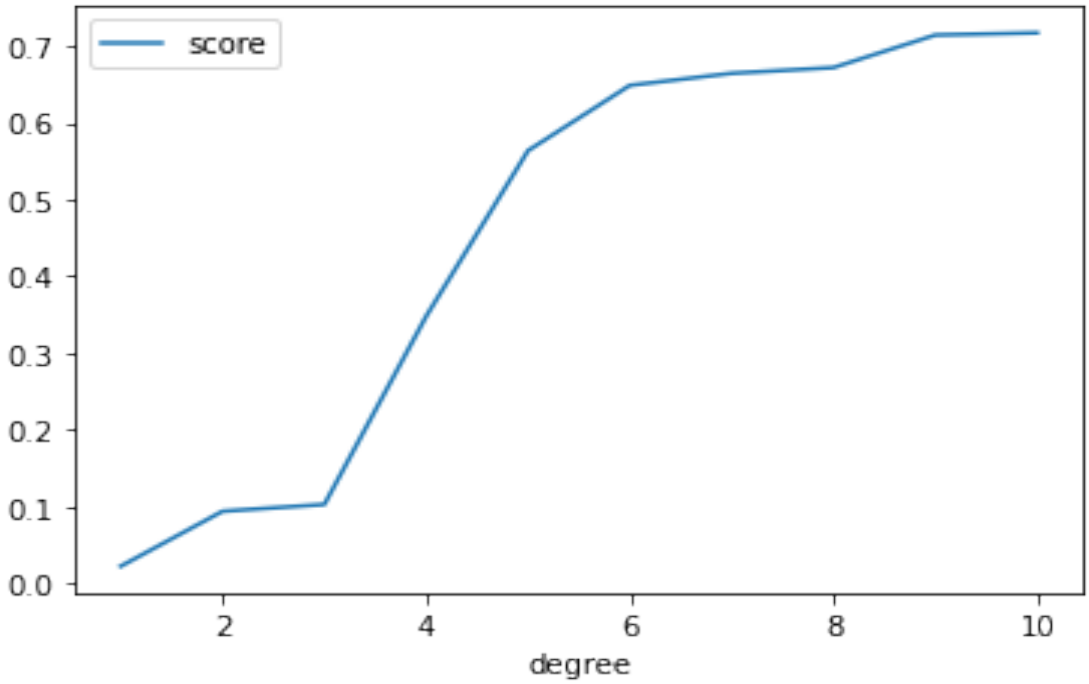


For the second way, we use only the recent 2 years data to build the model.

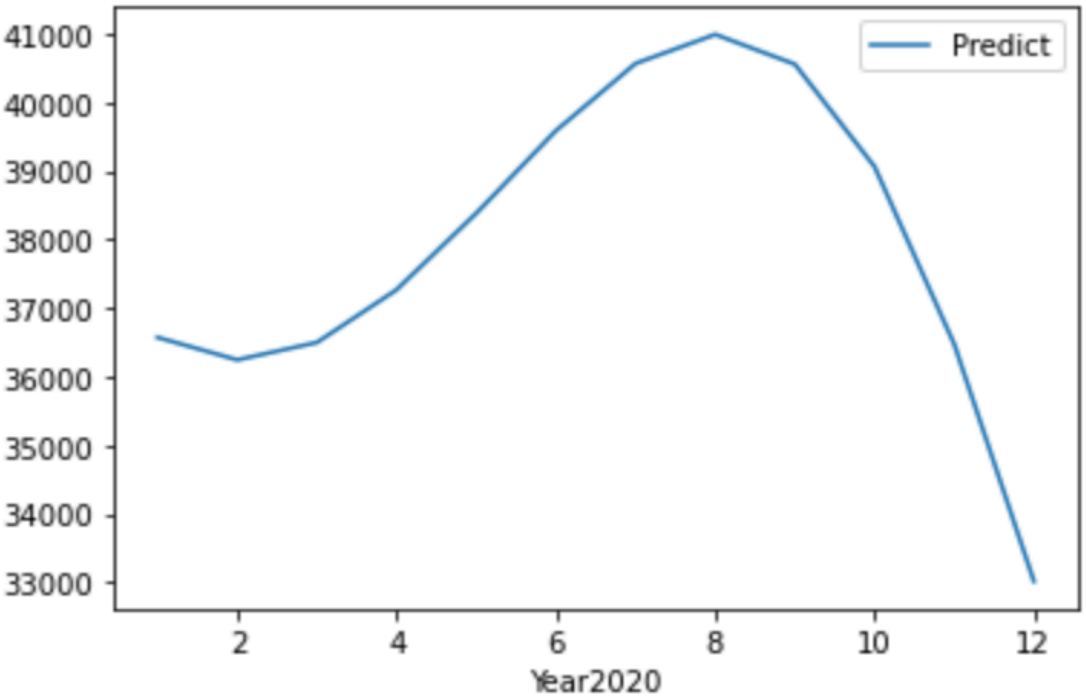
Also, try different degrees and draw the fitted curve.



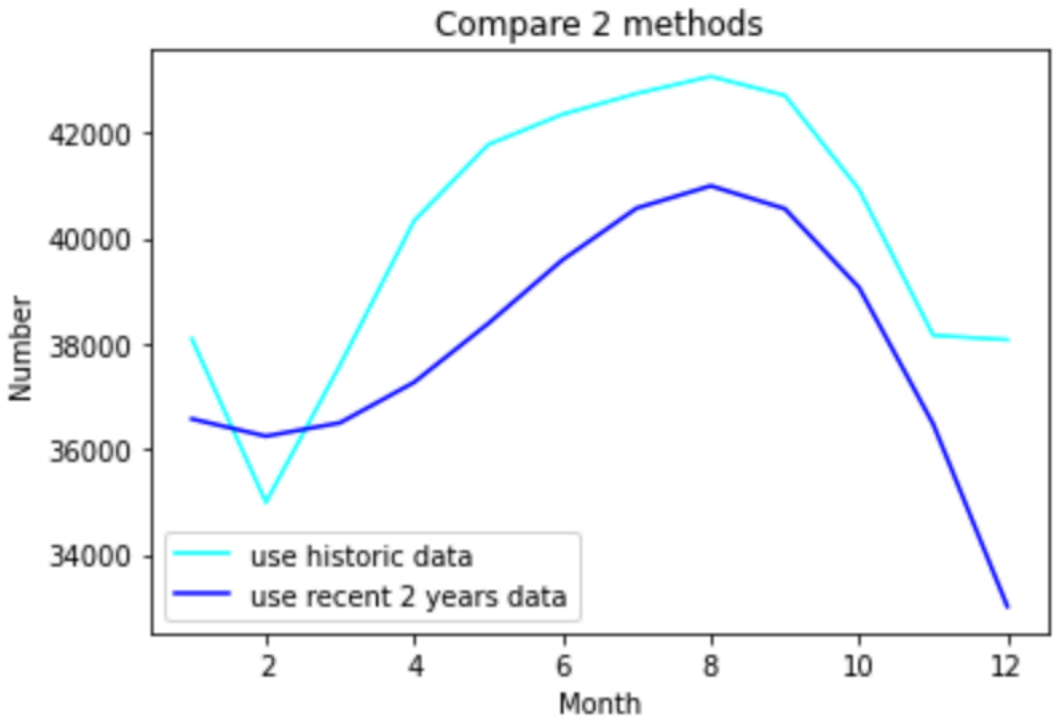
Draw a figure to visualize the scores with different degrees.



It is more obvious from this figure that degree 6 is appropriate. So let's get the prediction with degree 6.



Finally, we can compare the 2 ways of prediction.



Apparently, the second prediction is lower than the first one. This is due to the crime number decreasing year by year. Thus, it is believed that the second prediction is more precise.

Conclusion

From the above analysis, we answered the five questions we mentioned at the beginning. We got the most common type of crime, the area which has the most number of crimes, the trend of the crime cases, the relationship between crime and economic, and the prediction of 2020.

For the prediction, one thing I need to mention is that there is no actual data to verify the accuracy of the prediction. Therefore, this prediction only gives us a general idea.