SI618 Project Report
Lechen Zhang
leczhang@umich.edu
October 27, 2022

The Criminal Incidents Data Analysis of New York

I. Motivation

Public safety is one of the most important concerns of governments, especially in modern cities with high population density. Though many cities have invested a lot in police deployment, it can't cover the whole city all the time. However, if we are able to analyze the factors and patterns in crime data (like crime time, place, type, criminal and victim characteristics), then we can predict the communities of high risks, and distribute reasonable police strength in advance. Therefore, my research goal is to find out the influential factors and patterns of crimes.

I will mainly focus my analysis on the following 3 questions:

1. Is the crime data dependent on time? More specifically, are they influenced by weekdays, months, or real-time events like COVID pandemic?
2. Is the crime data dependent on geographic factors? More specifically, are they influenced by geographic proximity or geographic properties like house price?
3. Are the fields in crime data correlated? More specifically, is the gender of suspect related to the felony crime?

II. Data Source

1. NYPD Complaint Data Historic

My primary dataset is NYPD Complaint Data Current from New York Police Department (NYPD). It provides detailed crime statistics of all valid felony, misdemeanor, and violation crimes from 2006 to 2021. It has 35 columns in total, covering data like precinct, occurrence time, location, level of offense, offense type, suspect & victim characteristics, and etc. These data are helpful and sufficient for us to analyze the crimes in New York. More information on the dataset can be found at: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i. However, for the simplicity and readability of research, I only downloaded the data from 2019 to 2021. The data size is 429MB, saved at
./dataset/ NYPD_Complaint_Data_Historic.csv

The important variable is:

ADDR_PCT_CD: The precinct in which the incident occurred
    Number (nullable = true)
CMPLNT_FR_TM: Exact time of occurrence for the reported event
    string (nullable = true)
LAW_CAT_CD: Level of offense: felony, misdemeanor, violation
    string (nullable = true)
SUSP_SEX: Suspect's Sex Description
    string (nullable = true)
SUSP_RACE: Suspect's Race Description
    string (nullable = true)
VIC_SEX: Victim's Sex Description
    string (nullable = true)
VIC_RACE: Victim's Race Description
    string (nullable = true)

Record Num: 1310287 rows

Time Period: Jan. 1 2019 – Dec. 31 2021

2. NYC House Price

This dataset directly came from the codes of my Project 1, but the time range is modified to make my analysis more reasonable. The dataset is NYC House Price from the Department of Finance of New York city government. The Department of Finance collects data and values properties every year. Since it's the output of one part of my Project 1, it only contains 5 useful columns, and I'll only use 3 columns for this project. It was saved at
./dataset/house_price.csv.

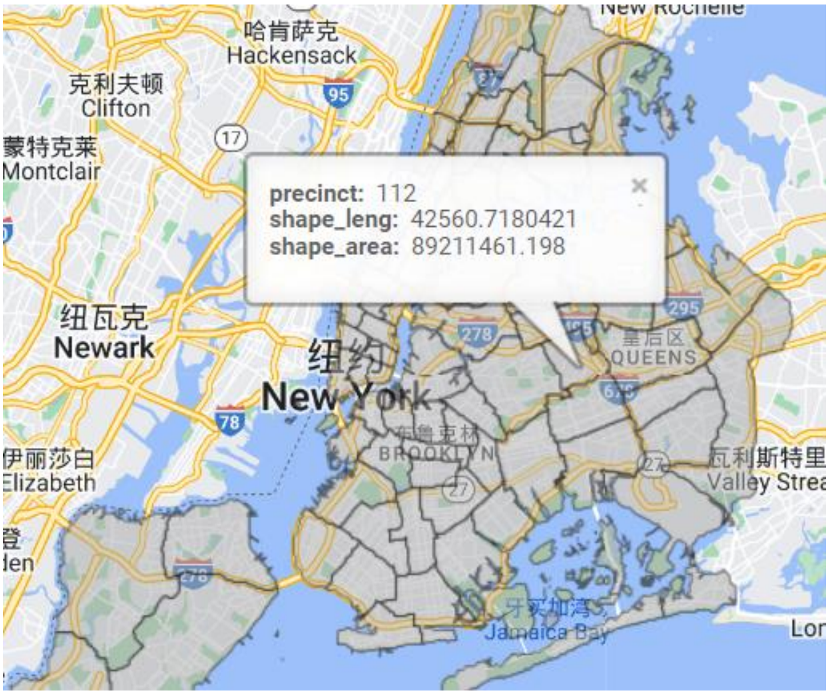
The important variable is:
neighborhood_of_crime: The name of neighborhood
    string (nullable = true)
precinct: The id of NYPD precinct
    number (nullable = true)
median_price: The median sale price of each precinct
    number (nullable = true)

Record Num: 237 rows

Time Period: Jan-Dec 2019

3. Police Precincts

This dataset is used for map demonstration of this project. Police Precincts is the GIS data of precincts in New York City from Department of City Planning (DCP). It contains the zone of each precinct and its number. More information on the dataset can be found at: https://data.cityofnewyork.us/Public-Safety/Police-Precincts/78dh-3ptz. The data was saved at **./dataset/Police Precincts.kml and ./ dataset/nypp.csv**



precinct: 112
shape_leng: 42560.7180421
shape_area: 89211461.198

Fig. 2 Police Precincts GIS

The important variable is:

the_geom: The zone data of precincts

MULTIPOLYGON (nullable = false)

precinct: The number of precincts

Number (nullable = false)

Record Num: 77 rows

Time Period: Created on 2013

III. Methods

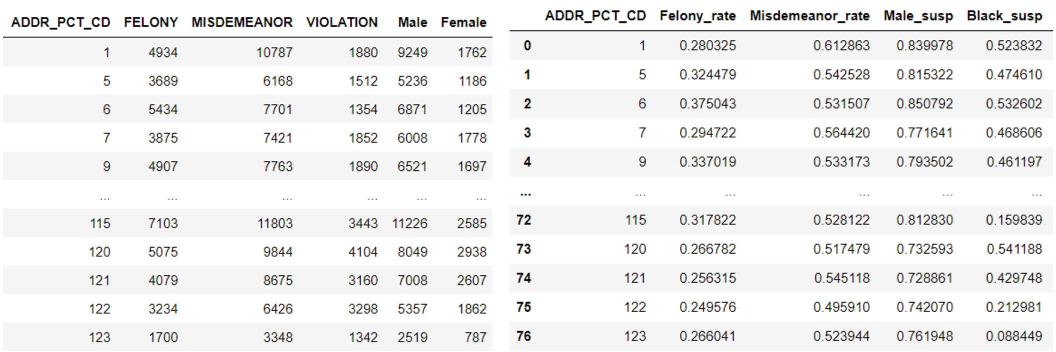
Question 1

I picked up useful columns like dates and crime type, and group them by dates and crime type to get the counts of each crime type in every day. Use pd.to_datetime to convert the date data into Python Datetime, and use “weekday()” function to classify them into different weekdays.

I dropped all rows that has missing values. The challenge I met is that for the date Jan 1st, the crime rates are usually extremely high. I guess the reason is that NYPD will attribute all crimes whose occurrence date is unknown to Jan 1st of the year for simplicity. Therefore, I also dropped all Jan 1st data.

Question 2

I picked up all useful properties of each crime incident for this question, including the characteristics of suspects and victims. I read the GIS data first by geopandas package for plotting the map. Then I used pd.crosstab to turn every categorical data into numerical data of each precinct. In this case, the index should be the precinct number, and the column names should be the number of specific crimes (e.g., crimes whose suspect is Asian). To prevent the population difference among precincts, I turned these data into proportions (e.g., what percent of crimes are caused by Asian) to prevent bias (sample data shown below).



ADDR_PCT_CD	FELONY	MISDEMEANOR	VIOLATION	Male	Female
1	4934	10787	1880	9249	1762
5	3689	6168	1512	5236	1186
6	5434	7701	1354	6871	1205
7	3875	7421	1852	6008	1778
9	4907	7763	1890	6521	1697
...	...	...	...	...	...
115	7103	11803	3443	11226	2585
120	5075	9844	4104	8049	2938
121	4079	8675	3160	7008	2607
122	3234	6426	3298	5357	1862
123	1700	3348	1342	2519	787

	ADDR_PCT_CD	Felony_rate	Misdemeanor_rate	Male_susp	Black_susp
0	1	0.280325	0.612863	0.839978	0.523832
1	5	0.324479	0.542528	0.815322	0.474610
2	6	0.375043	0.531507	0.850792	0.532602
3	7	0.294722	0.564420	0.771641	0.468606
4	9	0.337019	0.533173	0.793502	0.461197
...	...	...	...	...	...
72	115	0.317822	0.528122	0.812830	0.159839
73	120	0.266782	0.517479	0.732593	0.541188
74	121	0.256315	0.545118	0.728861	0.429748
75	122	0.249576	0.495910	0.742070	0.212981
76	123	0.266041	0.523944	0.761948	0.088449

Fig. 1. Comparison of conversion from crime counts
to crime proportion in each precinct

I dropped all rows of crime data that has missing values. But for house price data, some precincts don't have house price data. If I simply drop them, the plot will have many holes. Therefore, I filled these precincts by the house price of its surrounding precincts. The problem I met is that the house price is skewed, which will make the plot more light-colored. The solution is to put log scale on house price and normalize them into 0-1. This makes the plots look better.

Question 3

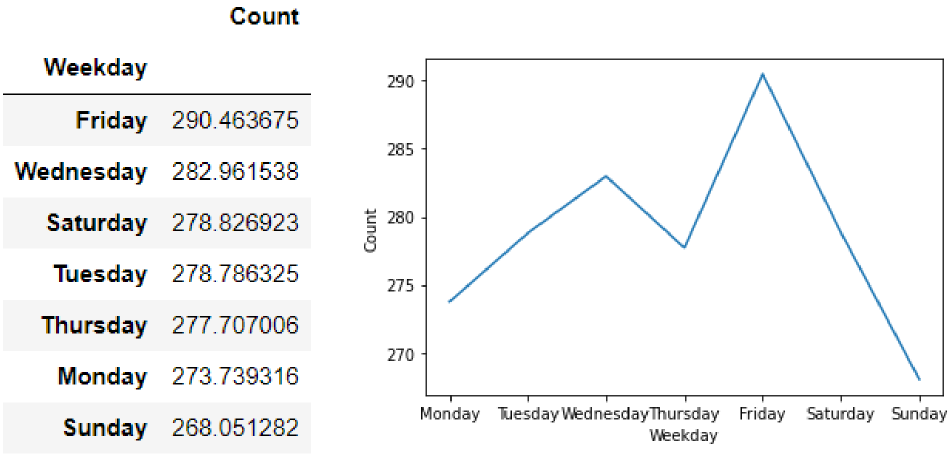
In this part I primarily use crosstabs to turn useful categorical data into numerical data, and do analysis based on the relationship inside these numerical data. Also, I used the proportional values I've created in Q2 for analysis. I didn't do much data processing here since the data in Q1 and Q2 are sufficient for my analysis.

I dropped all rows of crime data that has missing values. I didn't meet much problem in data processing here. Probably the only challenge here is how to analyze a dataset that almost all columns are categorical, and my answer is use crosstabs to turn them into numerical. It really works well in my dataset.

IV. Analysis and Results

Question 1: Is the crime data dependent on time? More specifically, are they influenced by weekdays, months, or real-time events like COVID pandemic?

Firstly, I'm going to find the crime rate pattern in different weekday. Since we added a new column of "Weekday" in dataset, we can just group the data by weekday, and find the average crime count in each weekday. The results are shown below.



	Count
Weekday	
Friday	290.463675
Wednesday	282.961538
Saturday	278.826923
Tuesday	278.786325
Thursday	277.707006
Monday	273.739316
Sunday	268.051282

Fig. 2. The Week-of-Day effect (Relationship between Weekday and Crime Rate)

Finding 1.1: Based on the results above, we can see that there's a clear Day-of-Week effect in crime data. The crime rate will reach a peak at Friday, and rapidly decrease on weekends. This reminded us to put more emphasis on the social security on Friday.

Also, I can use ggplot to plot the crimes rates over time. To see the impact of Covid-19 pandemic, I chose the time range from June 2019 to Dec 2020 for plotting. I used 3 different colors for different crime types (Felony, Misdemeanor and violation).

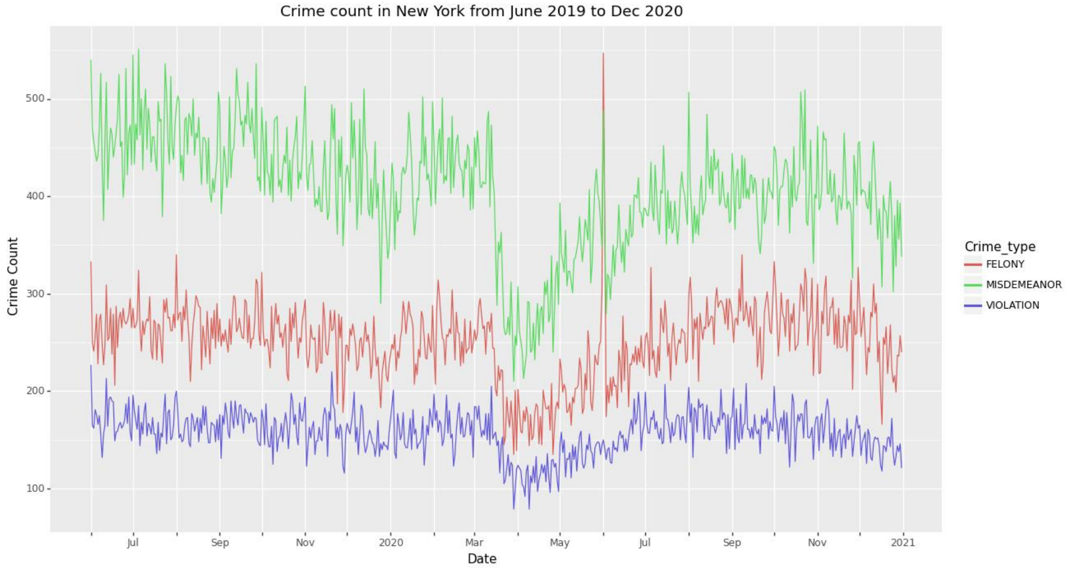


Fig. 3. Crime counts in New York from June 2019 to Dec 2020

To remove the Day-of-Week effect, I can use a regression model to fit the weekday values, and plot a new Time v.s. Crime Rate plot for comparison.

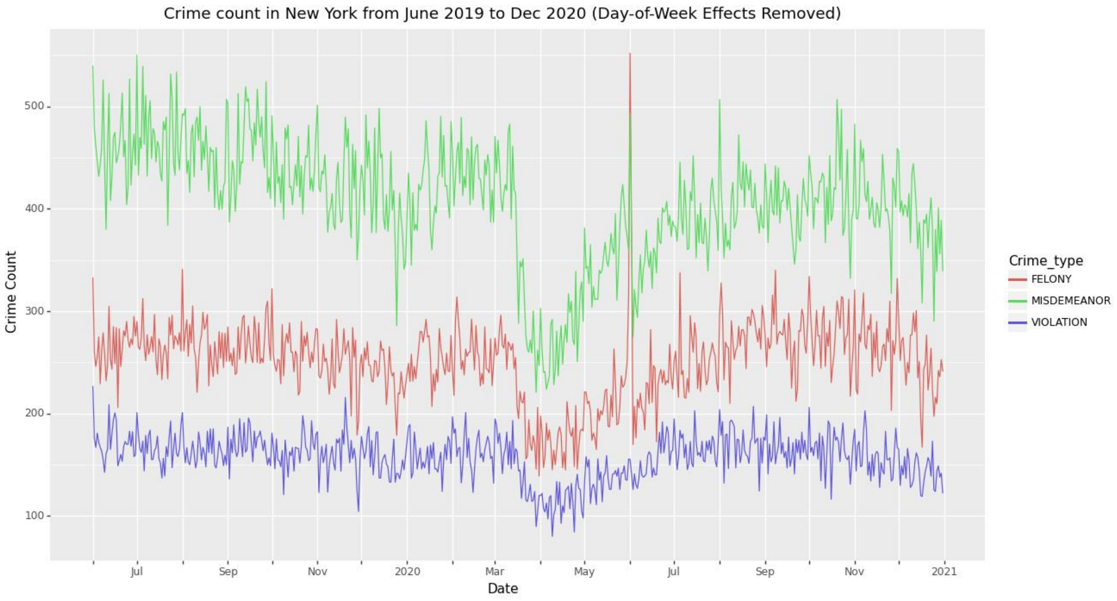


Fig. 4. Crime counts in New York from June 2019 to Dec 2020
(Day-of-Week Effect Removed)

However, the removal of Day-of-Week Effect is not very effective. The probable reason is that the crime rate data has many noises that can't be easily erased by removing the Day-of-Week Effect. To see a much clearer pattern, I used the smoothing function "geom_smooth" in ggplot as follows:

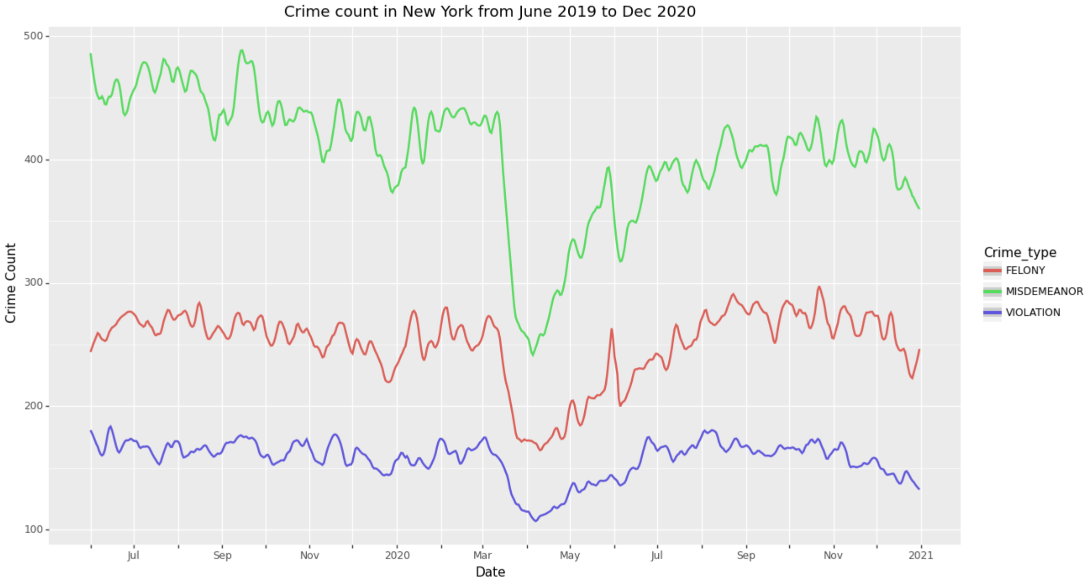


Fig. 5. Crime counts in New York from June 2019 to Dec 2020
(Smoothed by geom_smooth(span = 0.2))

Finding 1.2: Based on the plots above, we noticed that the Covid-19 pandemic significantly influenced the crime rate of each type, especially the misdemeanor crimes. The crime rate dramatically decreased from Mar-Apr 2020, and soon increased to the common level in 2 months. It indicates that the crime rate can be influenced by real-time events. Also, we can see some monthly pattern here. The crime rate will slightly increase in the first half of the year, and will slightly decrease in the second half of the year. However, more data are needed to support this hypothesis, and I won't do it here.

Conclusion: Based on the analysis above, we get enough evidence that the crime data is related to the time factors.

Question 2: Is the crime data dependent on geographic factors? More specifically, are they influenced by geographic proximity or geographic properties like house price?

Firstly, I tried to plot the felony rate of different precincts on the map. The plot can be done by joining the DataFrame of GIS map and the precinct proportion table (shown before in Fig. 1) based on the precinct number. Then, I will use plt.fill() to fill in each precinct with the color. The deeper the color, the higher felony rate it is.

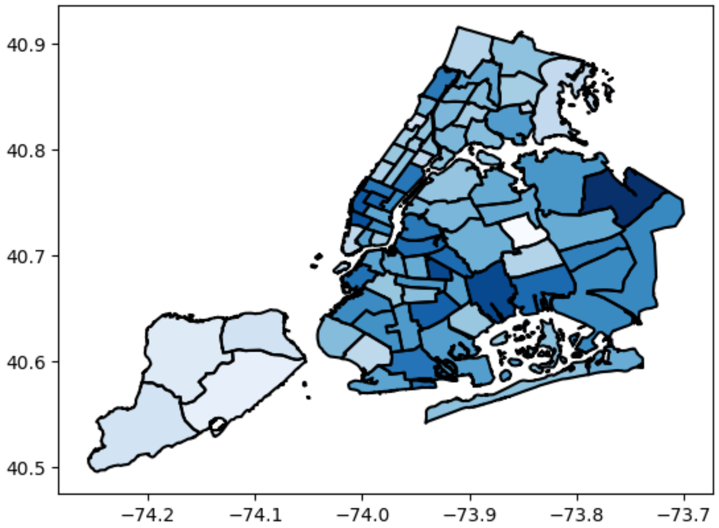
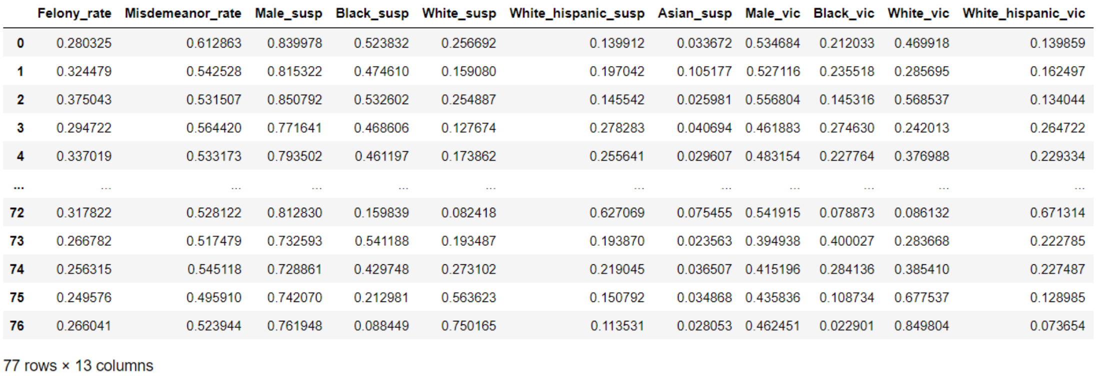


Fig. 6. Felony rate plot in New York Precinct map
(Darker = Higher)

Finding 2.1: We can see that the felony rate has shown an interesting pattern here. There're many deep color areas in the central urban area. Generally speaking, the central areas have darker color than surrounding, but it's also worth noticing that the right-up corner also has a dark area. This reminded me to try adding more features and do a clustering on these precincts.

To do so, I will use more columns for our cluster. Other than the felony rate, I added columns like "male suspect rate", "black suspect rate", "male victim rate", "Asian victim rate", and etc. There're 12 columns and 77 rows in total for clustering.



	Felony_rate	Misdemeanor_rate	Male_susp	Black_susp	White_susp	White_hispanic_susp	Asian_susp	Male_vic	Black_vic	White_vic	White_hispanic_vic
0	0.280325	0.612863	0.839978	0.523832	0.256692	0.139912	0.033672	0.534684	0.212033	0.469918	0.139859
1	0.324479	0.542528	0.815322	0.474610	0.159080	0.197042	0.105177	0.527116	0.235518	0.285695	0.162497
2	0.375043	0.531507	0.850792	0.532602	0.254887	0.145542	0.025981	0.556804	0.145316	0.568537	0.134044
3	0.294722	0.564420	0.771641	0.468606	0.127674	0.278283	0.040694	0.461883	0.274630	0.242013	0.264722
4	0.337019	0.533173	0.793502	0.461197	0.173862	0.255641	0.029607	0.483154	0.227764	0.376988	0.229334
...	...	...	...	...	...	...	...	...	...	...	...
72	0.317822	0.528122	0.812830	0.159839	0.082418	0.627069	0.075455	0.541915	0.078873	0.086132	0.671314
73	0.266782	0.517479	0.732593	0.541188	0.193487	0.193870	0.023563	0.394938	0.400027	0.283668	0.222785
74	0.256315	0.545118	0.728861	0.429748	0.273102	0.219045	0.036507	0.415196	0.284136	0.385410	0.227487
75	0.249576	0.495910	0.742070	0.212981	0.563623	0.150792	0.034868	0.435836	0.108734	0.677537	0.128985
76	0.266041	0.523944	0.761948	0.088449	0.750165	0.113531	0.028053	0.462451	0.022901	0.849804	0.073654

77 rows × 13 columns

Fig 7.1. The new data structure for clustering

Using Rule of Thumb gives a suggested cluster number of 6 as follows:

$$k \approx \sqrt{\frac{n}{2}} = \sqrt{77/2} \approx 6$$

Also, we can perform Silhouette Method to test the optimized cluster number. The plot is shown below.

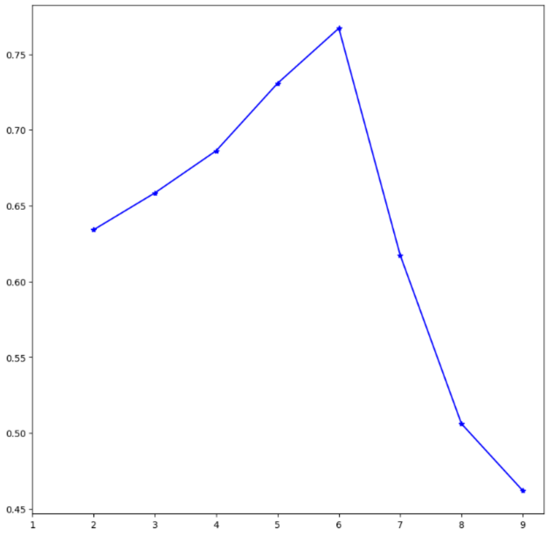


Fig. 7.2. Silhouette Method of clustering

It can be seen from the plot that the best cluster number is still 6.

Then, I used K-Means classifier to classify these precincts. I appended the classification data to the map data, and used different colors to show the cluster results as follows.

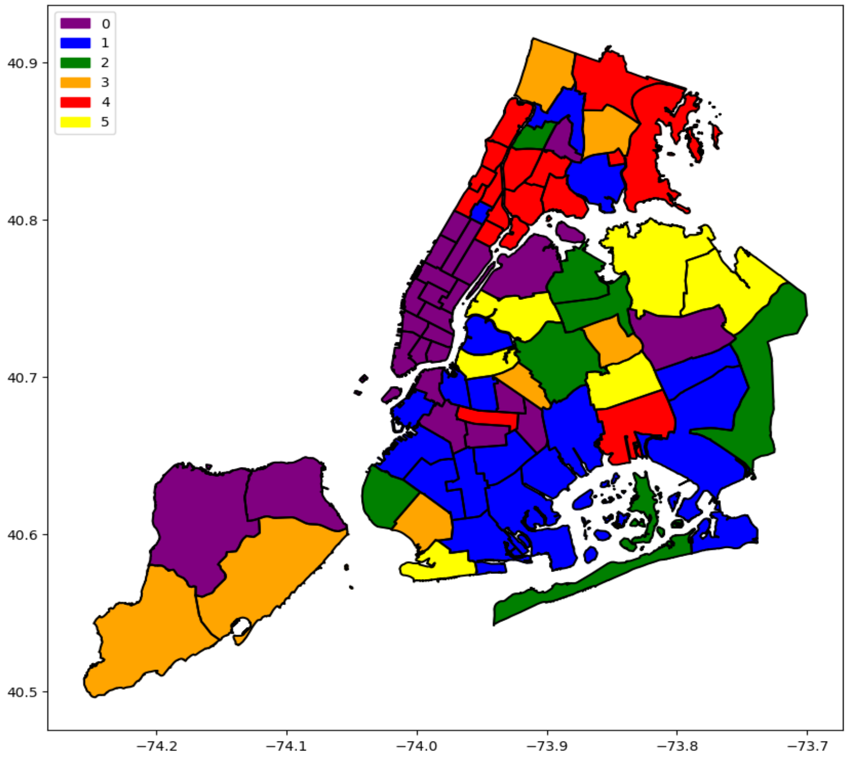
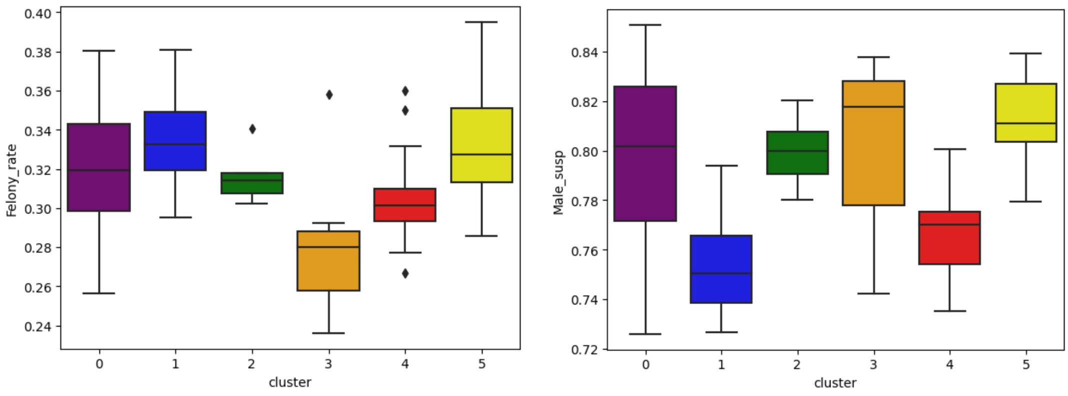


Fig. 8. Cluster result of precincts on map

Finding 2.2: Based on the plot above, we can see that the areas of each cluster are geographically close/adjacent to each other. The central urbans are clustered together, and the north and south area are separated by different clusters. Therefore, we get the evidence that the crime data is dependent on geographical factors.

For each of these clusters, they have the following properties for each influencing variable.



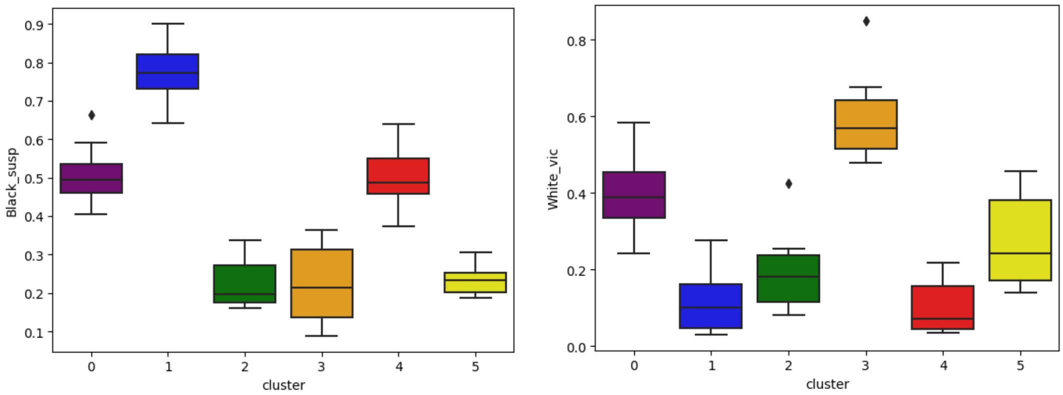


Fig. 9. Proportion of properties in each cluster

Finding 2.3:

Based on the plot above, we may notice that the blue areas have the highest black suspect rate, highest felony crime rate, and lowest white victim rate. This may indicate that the area has a larger proportion of black people.

However, the black people are not always related to high felony crime rate. In yellow areas (cluster 5), it has the lowest black crime rate, but also has the highest felony crime rate. After checking the map, we noticed that the yellow areas are mostly in Queens, which is the rich area of New York. Therefore, I will suggest the police in Queens be more careful about the felony crimes.

Also, many people may have the bias that areas with high black crime rate will also have high male crime rate, but the blue areas (cluster 1) showed the lowest male crime rate and highest black crime rate. It can also be a worth noticing problem for social security.

Generally speaking, it seems like racial data may a larger influence on the clusters since the distribution of racial data has a relatively small variance. Census data of race composition can be a good reference to confirm this hypothesis, but I can't include it here due to the page limit.

Finally, I joined the house price data from my Project 1 and tried to plot it on the map. Noticed that I applied a log scale to the median house price to prevent the negative plotting effect of skewed distribution.

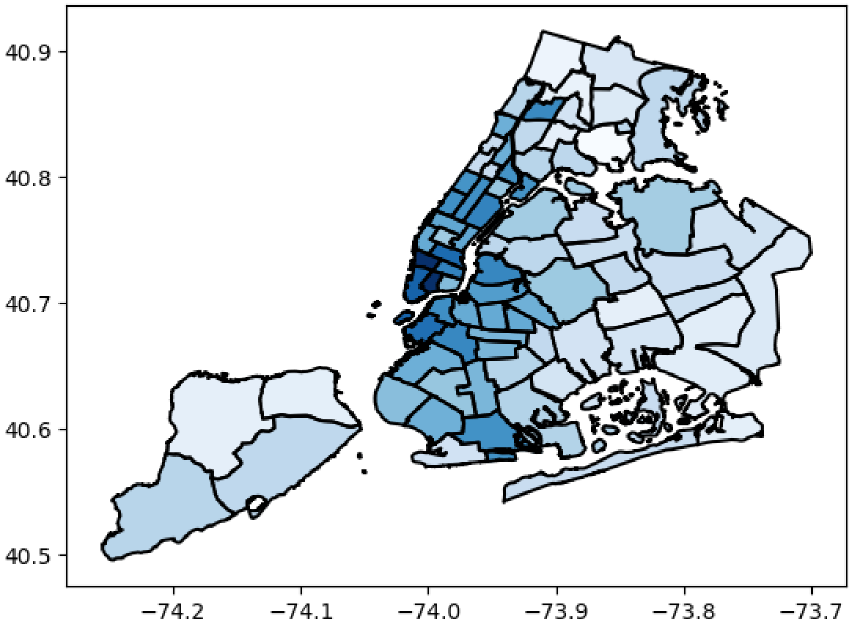
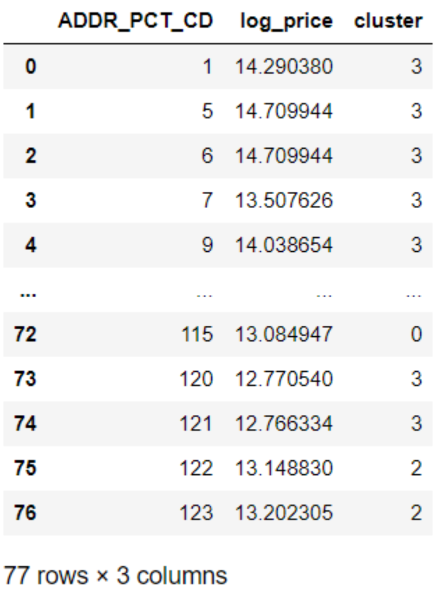


Fig. 10. Median house price plot of precincts (From my Project 1)
(Darker = Higher)

Finding 2.4: I noticed that some areas in Fig. 10 may be correlated to the cluster plot in Fig. 8. Each cluster might be related to a house price level.

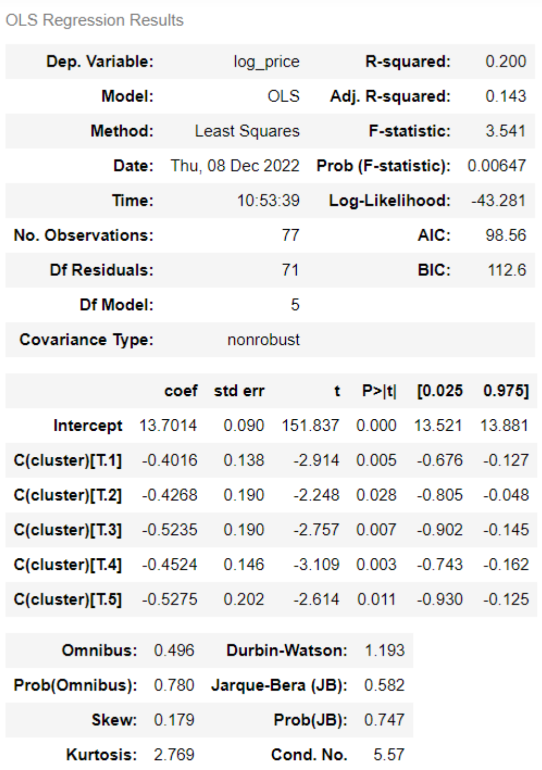
To confirm my Finding 2.3, I performed an OLS regression between clusters and median house price. The OLS regression codes and results are shown below:

```
model1 = smf.ols('log_price ~ C(cluster)', data=price_vs_cluster).fit()
```



	ADDR_PCT_CD	log_price	cluster
0	1	14.290380	3
1	5	14.709944	3
2	6	14.709944	3
3	7	13.507626	3
4	9	14.038654	3
...	...	...	...
72	115	13.084947	0
73	120	12.770540	3
74	121	12.766334	3
75	122	13.148830	2
76	123	13.202305	2

77 rows × 3 columns



OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.200
Model:	OLS	Adj. R-squared:	0.143
Method:	Least Squares	F-statistic:	3.541
Date:	Thu, 08 Dec 2022	Prob (F-statistic):	0.00647
Time:	10:53:39	Log-Likelihood:	-43.281
No. Observations:	77	AIC:	98.56
Df Residuals:	71	BIC:	112.6
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	13.7014	0.090	151.837	0.000	13.521	13.881
C(cluster)[T.1]	-0.4016	0.138	-2.914	0.005	-0.676	-0.127
C(cluster)[T.2]	-0.4268	0.190	-2.248	0.028	-0.805	-0.048
C(cluster)[T.3]	-0.5235	0.190	-2.757	0.007	-0.902	-0.145
C(cluster)[T.4]	-0.4524	0.146	-3.109	0.003	-0.743	-0.162
C(cluster)[T.5]	-0.5275	0.202	-2.614	0.011	-0.930	-0.125

Omnibus:	0.496	Durbin-Watson:	1.193
Prob(Omnibus):	0.780	Jarque-Bera (JB):	0.582
Skew:	0.179	Prob(JB):	0.747
Kurtosis:	2.769	Cond. No.	5.57

Fig. 11 LEFT. The price-cluster dataframe to fit
Fig. 11 RIGHT. OLS Regression Result between cluster and house price

Finding 2.5: I noticed that the p value = 0.006, which means that the relationship between clusters and house prices is significant. We can also find the house price properties from the fitted coefficients. For example, cluster 0 (purple), which is the central urban area, has the largest positive influence to the house price. For cluster 3 (orange), which is the places far from the city center, has the lowest house price, even if the felony crime rate is the also lowest there.

Conclusion: We can now conclude that the cluster is quite success, and it proves the strong relationship between crime data and geographical data like adjacency or house price. More research can be done on its relationship with Census of each area, but I won't cover them in this report due to the page limit.

Question 3: Are the fields in crime data correlated? More specifically, is the gender of suspect related to the felony crime?

For this question, I will reuse the crime count table created in Q1 and the crime rate table created in Q2 to do investigation (shown in Fig. 1).

Firstly, I will plot a mosaic plot to show the relationship between crime type and suspect gender in crime count. Also, I will do a chi-square analysis on this data to support my arguments.

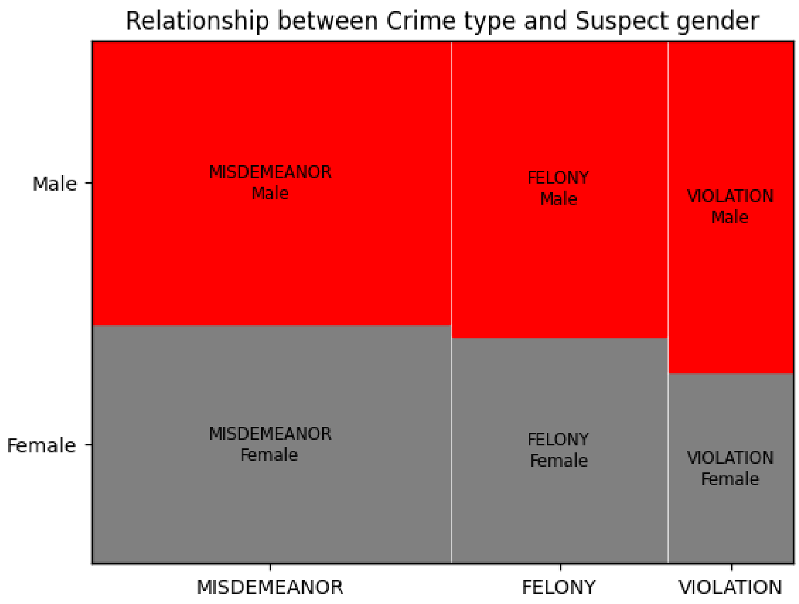
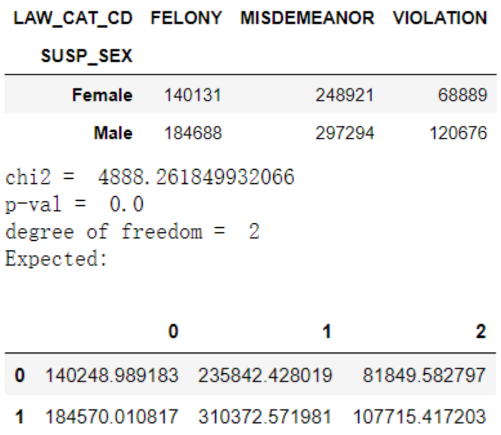


Fig. 12. Mosaic plot of crime type v.s. suspect gender



LAW_CAT_CD	FELONY	MISDEMEANOR	VIOLATION
SUSP_SEX			
Female	140131	248921	68889
Male	184688	297294	120676

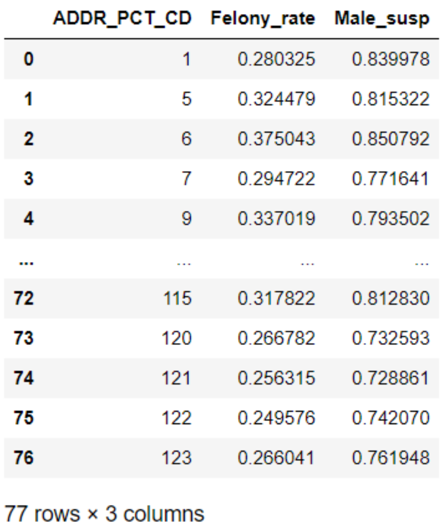
```
chi2 =  4888.261849932066
p-val =  0.0
degree of freedom =  2
Expected:
```

	0	1	2
0	140248.989183	235842.428019	81849.582797
1	184570.010817	310372.571981	107715.417203

Fig. 13. Chi-square result of crime type v.s. suspect gender

Finding 3.1: Based on the results above, I noticed that the crime type is related to the gender of suspects. Male and Female suspects share similar number of misdemeanor crimes, but male suspects have more felony crimes and far more violation crimes than female suspects. The Chi-square result also supported this argument. P-val = 0.0 means that the relationship is significant. Therefore, we can conclude that the crime count of felony crime is related to suspect's gender.

By reusing the table in Question 2 (shown in Fig. 1), we can also find the relationship between crime rate of felony crime and suspect's gender in different precincts. These data can be plot into regression plot as follows:



	ADDR_PCT_CD	Felony_rate	Male_susp
0	1	0.280325	0.839978
1	5	0.324479	0.815322
2	6	0.375043	0.850792
3	7	0.294722	0.771641
4	9	0.337019	0.793502
...	...	...	...
72	115	0.317822	0.812830
73	120	0.266782	0.732593
74	121	0.256315	0.728861
75	122	0.249576	0.742070
76	123	0.266041	0.761948

77 rows × 3 columns

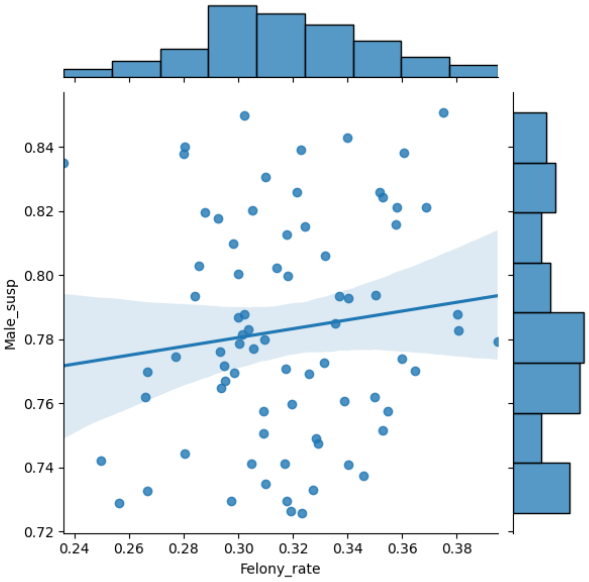


Fig. 14. Regression plot between Felony crime rate and Male suspect crime rate

Finding 3.2: Based on the regression plot above, we can conclude that area which has higher Felony rate usually also has a higher Male suspect crime rate.

Conclusion: Based on the findings above, we may conclude that there exists a relationship between the gender of suspect and the felony crime. Male criminals are more likely to commit a felony crime than female.

V. Additional notes and Self-reflection

Dear GSI and Professor, Thanks for your patience to read my verbose report. Actually, it has an even longer version before because I also made an analysis of the relationship between the race suspect and victims, making my Question 3 as verbose as Question 2. I deleted these parts because the page is limited, and I think Question 2 is a more fun challenge to me.

At first, I have no idea about how to process this dataset, because almost all data are categorical. However, by turning these categorical data into counts and proportional data, I suddenly found out that I could apply any analysis tool on it. Thanks to the numerical data, I can do a clustering in Question 2, and I learnt a lot plotting tools to plot them as a map. I'm very satisfied with the map plot I made, though the analysis might still be not sufficient (I really want to add an in-depth analysis of Census dataset to it but it will take too many pages). Also, it's very joyful that I can reuse my house price codes/data in Project 1 to plot a house price map for comparison.

Finally, I should say this course provided me a lot food for thought this semester. I really enjoy the feeling of learning things on practice, and I feel like I'm getting more experienced on data manipulation. Thanks for the patient and responsible GSIs. cya!