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

The relationship between House Price and Criminal Incidents in Neighborhoods of New York

I. Motivation

Public safety is one of the most important concerns of governments, especially in modern cities with high population density. The spread of guns may cause mass casualties as long as the vicious incident happens. If the police can reach the crime scene much faster, then we can save a lot of innocent lives. But the fact is, 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 factor of every neighborhood (like housing price) and the patterns in crime data (like crime time and criminal features), then we can predict the communities of high risks, and distribute reasonable police strength in advance. This can not only improve the social security level, but also cut down the expenses of police deployment.

Since New York is a typical modern city with high population density and high crime rate, I will use its data for the following research questions.

Research Questions:

1. Is house price related to felony crime rate&frequency of neighborhood?
2. Does house price of neighborhood affect felony crime rate of each race?
3. Does felony crime rate of different time in a day differ in different neighborhood class?

II. Data Source

I met with some trouble when I was using my original dataset in project proposal. Therefore, I did some minor modification, and added 2 additional data sources to support my project.

1. NYC Calendar Sales

My first dataset is NYC Calendar Sales from the Department of Finance of New York city government. The Department of Finance collects data and values properties every year. It has 19 columns in total, covering data like neighborhood, sale price, land area

and etc. More information on the dataset can be found at: https://data.cityofnewyork.us/Housing-Development/NYC-Calendar-Sales-Archive-/uzf5-f8n2. For simplicity, I only used the data of whole New York in 2015, and saved it at **./raw_dataset/2015_newyork.csv**.

The important variable is:
NEIGHBORHOOD: The name of neighborhood
    string (nullable = true)
SALE PRICE: The sale price
    Number (nullable = true)

Record Num: 108202 rows

Time Period: Jan-Dec 2015

2. NYPD Complaint Data Current

My second 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 every quarter. It has 36 columns in total, covering data like precinct, occurrence time, location, level of offense, offense type, response time, 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-Current-Year-To-Date-/5uac-w243. The data size is 139MB, saved at
./raw_dataset/ NYPD_Complaint_Data_Current__Year_To_Date_.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_RACE: Suspect's Race Description
    string (nullable = true)

Record Num: 396978 rows

Time Period: Jan. 1 - Oct. 19 in 2022

3. Neighborhood Names GIS

This dataset is additional to the project, and the reason will be explained further in data manipulation chapter.

Neighborhood Names GIS is the GIS data of neighborhoods in New York City from Department of City Planning (DCP). It contains the coordinates of each neighborhood center, the neighborhood name, and etc. More information on the dataset can be found at: https://data.cityofnewyork.us/City-Government/Neighborhood-Names-GIS/99bc-9p23. The data was saved at **./raw_dataset /Neighborhood Names GIS.kml**.

Fig. 1 Neighborhood Names GIS

The important variable is:

the_geom: The coordinate of neighborhood center

POINT (nullable = false)

Name: The name of neighborhood

string (nullable = false)

Record Num: 299 rows

Time Period: Created on Feb 26, 2020

4. Police Precincts

This dataset is also additional to the project, and the reason will be explained further in data manipulation chapter.

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 **./raw_dataset/Police Precincts.kml and ./raw_dataset/nypp.csv**

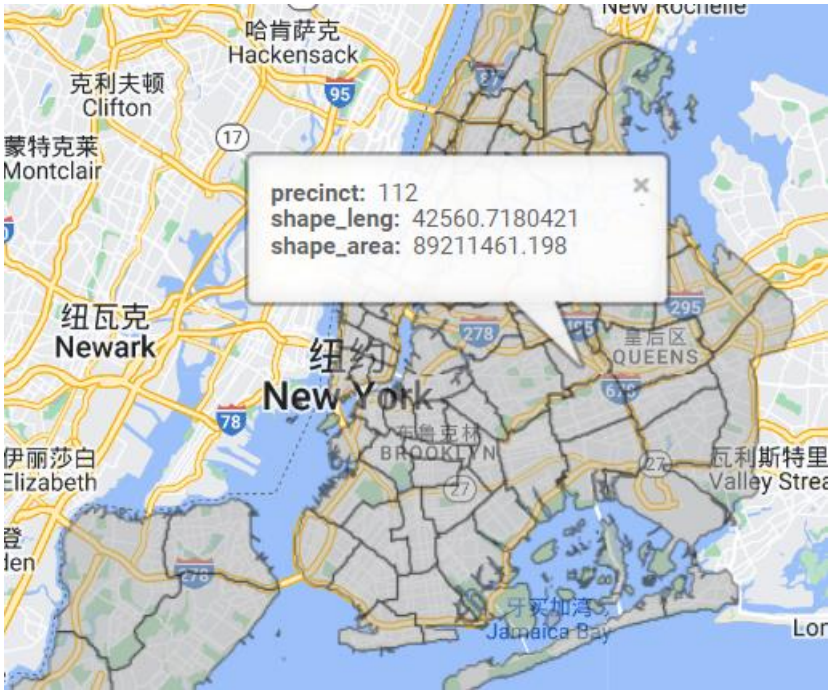


Fig. 2 Police Precincts GIS

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. Data Manipulation

Data manipulation is very important in my project. The reason is that I want to join these 2 data set by neighborhood, but the crime dataset only has precinct to indicate the location of crime. Therefore, I have to build the relationship between precinct and neighborhood, which brings 2 extra datasets into this project. The following part of this chapter will be mostly about the joining process.

However, the overall 4 dataset can't match perfectly since they are from different source, which can also lead to multiple unmatched data. For simplicity, I will drop most of unreasonable and missing data during SparkSQL process.

Since the data manipulation is quite complicated, I used the following diagram to show the manipulation process. Then I will explain the steps in detail.

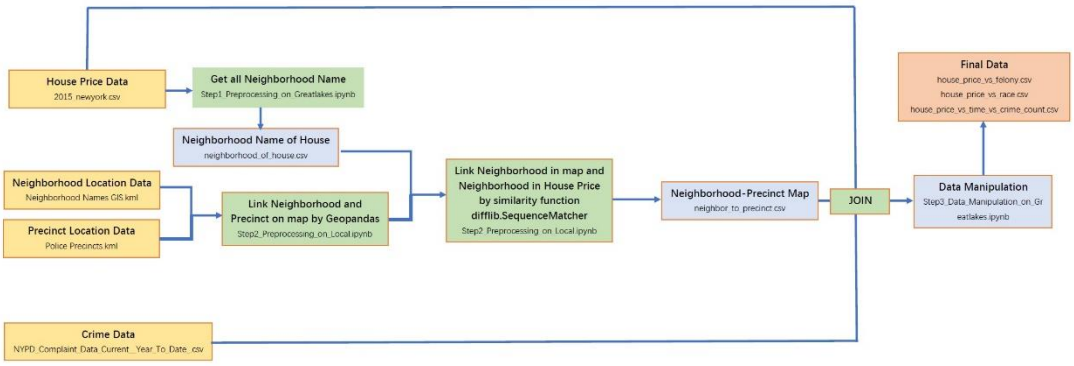


Fig. 3 Workflow of data manipulation

Ultimate Goal: Join House Price dataset and Crime dataset by a map from neighborhood name in House Price dataset to precinct number in Crime dataset (neighborhood_to_precinct.csv)

Step 1. Get all neighborhood names of House Price dataset.

The code of this part is shown in Step1_Preprocessing_on_Greatlakes.ipynb.

To build up the ultimate map, we have to start with what we have in House Price dataset. Hence, I used SparkSQL here to find the unique neighborhood names in House Price dataset.

```
q1 = sqlContext.sql('''
    SELECT NEIGHBORHOOD as neighborhood_of_house
    FROM
        houses
    WHERE NEIGHBORHOOD is not NULL
    GROUP BY neighborhood
    ORDER BY neighborhood ASC
    ''')
```

The output is a list of neighborhood name in neighborhood_of_house.csv.

Step 2_1. Match neighborhood in GIS dataset with precinct in GIS dataset

The code of this part is shown in Step2_Preprocessing_on_Local.ipynb.

	geometry	Precinct
0	MULTIPOLYGON (((-74.04388 40.69019, -74.04351 ...	1
1	MULTIPOLYGON (((-73.98864 40.72293, -73.98869 ...	5
2	MULTIPOLYGON (((-73.99968 40.73855, -73.99684 ...	6
3	MULTIPOLYGON (((-73.97346 40.71896, -73.97357 ...	7
4	MULTIPOLYGON (((-73.97161 40.72672, -73.97163 ...	9
...	...	...
72	MULTIPOLYGON (((-73.85892 40.76241, -73.85931 ...	115
73	MULTIPOLYGON (((-74.05357 40.60370, -74.05407 ...	120
74	MULTIPOLYGON (((-74.15946 40.64145, -74.15975 ...	121
75	MULTIPOLYGON (((-74.05051 40.56642, -74.05047 ...	122
76	MULTIPOLYGON (((-74.16983 40.56109, -74.16983 ...	123

	Name	Description	geometry
0	Wakefield		POINT (-73.84720 40.89471)
1	Co-op City		POINT (-73.82994 40.87429)
2	Eastchester		POINT (-73.82781 40.88756)
3	Fieldston		POINT (-73.90564 40.89544)
4	Riverdale		POINT (-73.91259 40.89083)
...	...	...	...
294	Lighthouse Hill		POINT (-74.13793 40.57651)
295	Richmond Valley		POINT (-74.22957 40.51954)
296	Malba		POINT (-73.82668 40.79060)
297	Highland Park		POINT (-73.89028 40.68249)
298	Madison		POINT (-73.94842 40.60938)

Fig. 4 Dataset view of neighborhood and precinct in GIS dataset

Based on the above 2 dataset, I noticed that I can join them by determining whether the POINT is inside the MULTIPOLYGON. Here I used the package of geopandas, and the MULTIPOLYGON has an attribute contains, and MULTIPOLYGON.contains(POINT) will return True/False. By using a "for loop", I mapped all POINTs to a MULTIPOLYGON, and builded up a mapping between neighborhood and precinct.

Step 2_2. Match neighborhood in GIS dataset with neighborhood in House Price dataset

neighborhood_of_house	neighborhood_of_GIS
ANNADALE	Annadale
ARDEN HEIGHTS	Arden Heights
ARROCHAR	Arrochar
ARROCHAR-SHORE ACRES	Shore Acres
ARVERNE	Arverne
ASTORIA	Astoria
BATH BEACH	Bath Beach
BATHGATE	East Flatbush
BAY RIDGE	Bay Ridge
BAYCHESTER	Baychester
BAYSIDE	Bayside
BEDFORD PARK/NORWOOD	Bedford Park

Fig. 5 Dataset comparison between neighborhood in GIS dataset and in House dataset

The code of this part is shown in Step2_Preprocessing_on_Local.ipynb.

If we carefully watch the difference between GIS dataset and House Price dataset, we can observe that there exists not only capitalization problems in between, but also some other difference like "ARROCHAR-SHORE ACRES" and "Shore Acres". This can't be easily solved by join function. Therefore, I used a similarity function from difflib package for joining these 2 datasets. The function is difflib.SequenceMatcher(None, a, b) .quick_ratio(), which returns the similarity (0-1) between a and b. For example, "ARROCHAR-SHORE ACRES" and "Shore Acres" has the similarity of 0.71.

Here I designed a “for loop” to find a best match “neighborhood of GIS” for each “neighborhood of house”. The detailed code is shown in notebook.

Step 2_3. Join two table of Step 2_1 and Step 2_2 to neighbor_to_precinct.csv

Now we almost completed our ultimate goal, and neighbor_to_precinct.csv is the map from neighborhood name in House Price dataset to precinct number in Crime dataset.

Step 3. Join House Price dataset and Crime dataset

The code of this part is shown in Step3_Data_Manipulation_on_Greatlakes.ipynb.

The key idea is just

“SELECT * FROM
crime
INNER JOIN neighbor_to_precinct
ON crime.precinct = neighbor_to_precinct.precinct
INNER JOIN house
ON neighbor_to_precinct.neighborhood = house.neighborhood”

But I will not join them together at once in practice, since it will create to an unnecessary large dataset. Instead, I will join house and “neighbor_to_precinct” first to get temp table “price_with_precinct”, and then join the combined data with crime dataset.

IV. Analysis and Visualization

Q0. Data preparation of median house price in each neighborhood

Since our research object is house price of each neighborhood, and house price is a highly-skewed data, I will use the median price of each neighborhood as indicator (The average price is used as a reference). The calculation is done by SparkSQL. Before calculation, I filtered out all invalid house price data (<$1000 or NULL) and neighborhoods with <10 house price data. Then I grouped the data by “neighborhood”, and calculated the median house price of each neighborhood by percentile_approx(`SALE PRICE`, 0.5). Finally, I joined the table with “neighbor_to_precinct” to get temp table “price_with_precinct”. The resulting table is shown below.

neighborhood_of_house	neighborhood_of_crime	precinct	neighborhood	avg_price	median_price
MIDTOWN CBD	Midtown	14	MIDTOWN CBD	24375971.18	995000.0
FASHION	Madison	61	FASHION	18634471.52	3265000.0
FLATIRON	Flatiron	13	FLATIRON	12763218.49	2240150.0
JAVITS CENTER	Jamaica Center	103	JAVITS CENTER	10684839.18	610000.0
SOHO	Soho	1	SOHO	8620889.23	2952925.0
LITTLE ITALY	Little Italy	5	LITTLE ITALY	8528344.77	3115000.0
KIPS BAY	Prince's Bay	123	KIPS BAY	8208160.88	850000.0
FINANCIAL	Financial District	1	FINANCIAL	5398535.87	1095000.0
BUSH TERMINAL	Butler Manor	123	BUSH TERMINAL	5226644.37	770000.0
CLINTON	Clinton	10	CLINTON	4552447.31	1053888.0
CHELSEA	Chelsea	10	CHELSEA	4467666.58	1236000.0
NAVY YARD	Navy Yard	88	NAVY YARD	4375354.00	970000.0
GREENWICH VILLAGE...	Greenwich Village	1	GREENWICH VILLAGE...	4293513.60	1395000.0
CHINATOWN	Chinatown	5	CHINATOWN	4063978.05	1697852.0
HARLEM-WEST	East Harlem	23	HARLEM-WEST	4018326.92	575000.0
CIVIC CENTER	Civic Center	5	CIVIC CENTER	3940556.27	2525000.0
LONG ISLAND CITY	Long Island City	108	LONG ISLAND CITY	3904138.18	930000.0
MOUNT HOPE/MOUNT ...	Mount Eden	44	MOUNT HOPE/MOUNT ...	3663337.71	510000.0
DOWNTOWN-FULTON MALL	Downtown Flushing	109	DOWNTOWN-FULTON MALL	3621201.24	1058980.0
PELHAM PARKWAY SOUTH	Pelham Parkway	49	PELHAM PARKWAY SOUTH	3601999.54	350000.0

only showing top 20 rows

Fig. 6 Combined data of neighborhood, precinct and median house price (price_with_precinct)

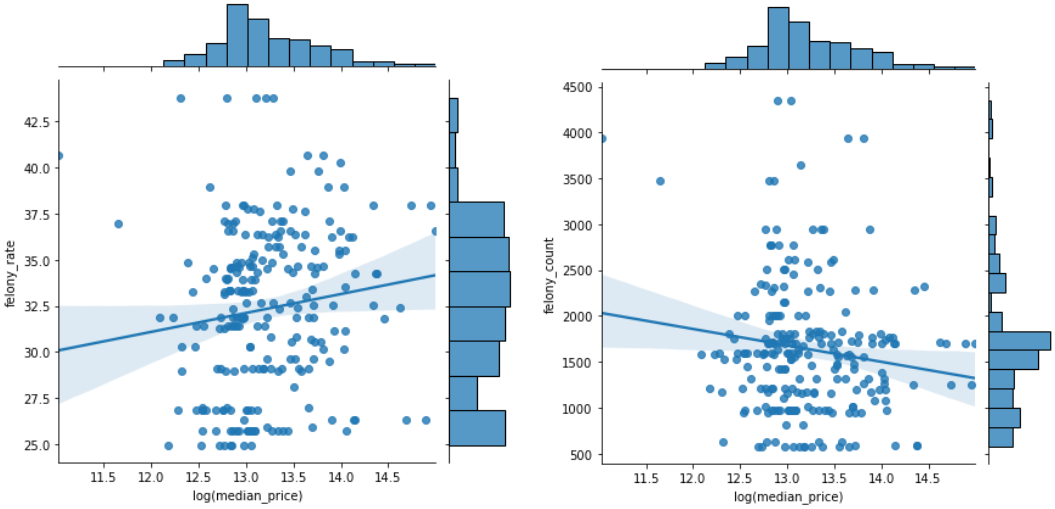
Q1. Is house price related to felony crime rate&frequency of neighborhood?

Now we already have the house price data in "price_with_precinct", we should only manipulate the crime data and join them together. The crime data is joined by precinct, and the felony rate&frequency is calculated in each precinct as follows:

felony rate: SUM(cast((LAW_CAT_CD == 'FELONY') as int))/COUNT(*)*100
felony frequency: SUM(cast((LAW_CAT_CD == 'FELONY') as int))

Then, the crime data is joined with "price_with_precinct" through "precinct". The resulting data and its visualization are shown below:

neighborhood	avg_price	median_price	felony_rate	felony_count
MIDTOWN CBD	24375971.18	995000.0	40.64	3936
FASHION	18634471.52	3265000.0	36.55	1703
FLATIRON	12763218.49	2240150.0	32.39	1714
JAVITS CENTER	10684839.18	610000.0	37.99	2507
SOHO	8620889.23	2952925.0	26.31	1702
LITTLE ITALY	8528344.77	3115000.0	37.96	1246
KIPS BAY	8208160.88	850000.0	29.11	581
FINANCIAL	5398535.87	1095000.0	26.31	1702
BUSH TERMINAL	5226644.37	770000.0	29.11	581
CLINTON	4552447.31	1053888.0	38.98	1321
CHELSEA	4467666.58	1236000.0	38.98	1321
NAVY YARD	4375354.00	970000.0	30.56	951
GREENWICH VILLAGE...	4293513.60	1395000.0	26.31	1702
CHINATOWN	4063978.05	1697852.0	37.96	1246
HARLEM-WEST	4018326.92	575000.0	30.82	1469
CIVIC CENTER	3940556.27	2525000.0	37.96	1246
LONG ISLAND CITY	3904138.18	930000.0	34.58	1710
MOUNT HOPE/MOUNT ...	3663337.71	510000.0	37.64	3643
DOWNTOWN-FULTON MALL	3621201.24	1058980.0	37.09	2941
PELHAM PARKWAY SOUTH	3601999.54	350000.0	33.29	1759

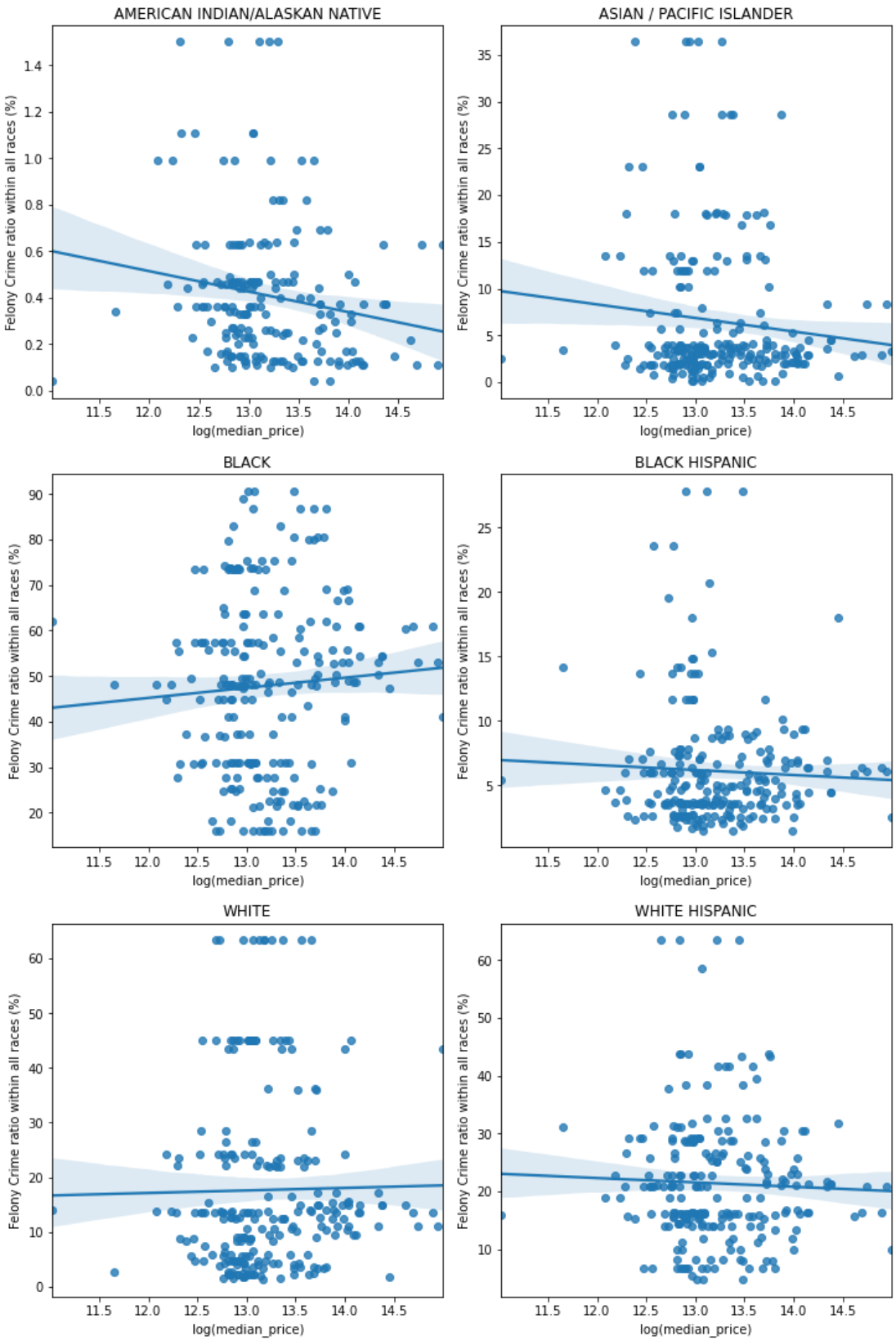


Based on the visualization above, we can observe that the log(price) distribution is less skewed and normally distributed, which helps our analysis. It can be observed that there exist many outliers for both house price and felony data, but the whole dataset follows a clear pattern. The felony crime frequency is lower in high house price neighborhoods, and higher in low house price neighborhoods. However, the felony crime rate in all crime type is higher in high house price neighborhoods, and lower in low house price neighborhoods. This suggested that high house price neighborhoods usually perform better in public security, but they should put more emphasis on felony crimes to reduce the felony rate. For middle house price neighborhoods, their felony crime data has a large variance. Therefore, those outliers with high felony frequency should invest more police strength to improve the social security.

Q2. Does house price of neighborhood affect felony crime rate of each race?

I used SparkSQL here. The crime data in Q2 is GROUP BY (precinct and SUSP_RACE) ON precinct, so that I can get the crime data of each race in each neighborhood. To better compare the crime rate of each race, I used race_count/SUM(race_count) OVER(PARTITION BY neighborhood) to calculate the relative felony crime rate. The data and its visualization are shown below:

neighborhood	avg_price	median_price	race	race_ratio
LITTLE ITALY	8528344.77	3115000.0	AMERICAN INDIAN/A...	0.63
SOHO	8620889.23	2952925.0	AMERICAN INDIAN/A...	0.11
CIVIC CENTER	3940556.27	2525000.0	AMERICAN INDIAN/A...	0.63
TRIBECA	3253793.64	2400000.0	AMERICAN INDIAN/A...	0.11
FLATIRON	12763218.49	2240150.0	AMERICAN INDIAN/A...	0.22
FORDHAM	3198945.36	1900000.0	AMERICAN INDIAN/A...	0.15
CARROLL GARDENS	2188451.93	1750000.0	AMERICAN INDIAN/A...	0.37
COBBLE HILL	2851194.50	1745000.0	AMERICAN INDIAN/A...	0.37
UPPER EAST SIDE (...	2274575.03	1700000.0	AMERICAN INDIAN/A...	0.37
CHINATOWN	4063978.05	1697852.0	AMERICAN INDIAN/A...	0.63
RED HOOK	1382044.15	1400000.0	AMERICAN INDIAN/A...	0.37
GREENWICH VILLAGE...	4293513.60	1395000.0	AMERICAN INDIAN/A...	0.11
GREENWICH VILLAGE...	3175467.13	1375000.0	AMERICAN INDIAN/A...	0.11
WILLIAMSBURG-NORTH	3538125.96	1350000.0	AMERICAN INDIAN/A...	0.13
SOUTHBRIDGE	2672503.30	1305000.0	AMERICAN INDIAN/A...	0.13
TODT HILL	1266955.33	1275000.0	AMERICAN INDIAN/A...	0.47
UPPER WEST SIDE (...	3179013.17	1250000.0	AMERICAN INDIAN/A...	0.17
EAST VILLAGE	3339575.35	1250000.0	AMERICAN INDIAN/A...	0.12
BOERUM HILL	1780772.93	1240000.0	AMERICAN INDIAN/A...	0.33
CHELSEA	4467666.58	1236000.0	AMERICAN INDIAN/A...	0.30



It can be seen that the felony crime rate of American Indian and Asian decreases when the house price of neighborhood increases. This reminded our police to pay more attention to American Indian and Asian in poor zone. However, Black people has higher

felony crime rate in richer zone. This differs from our intuition, because according to the census data, there're fewer black people in rich zone. Therefore, our police should pay more attention to the felony crime of black people in rich zone.

For the remaining data, there's no obvious relationship. The crime rate of white people slightly increases when the house price increases, and the crime rate of Black/White Hispanic slightly decreases when the house price increases.

Q3. Does felony crime rate of different time in a day differ in different neighborhood class?

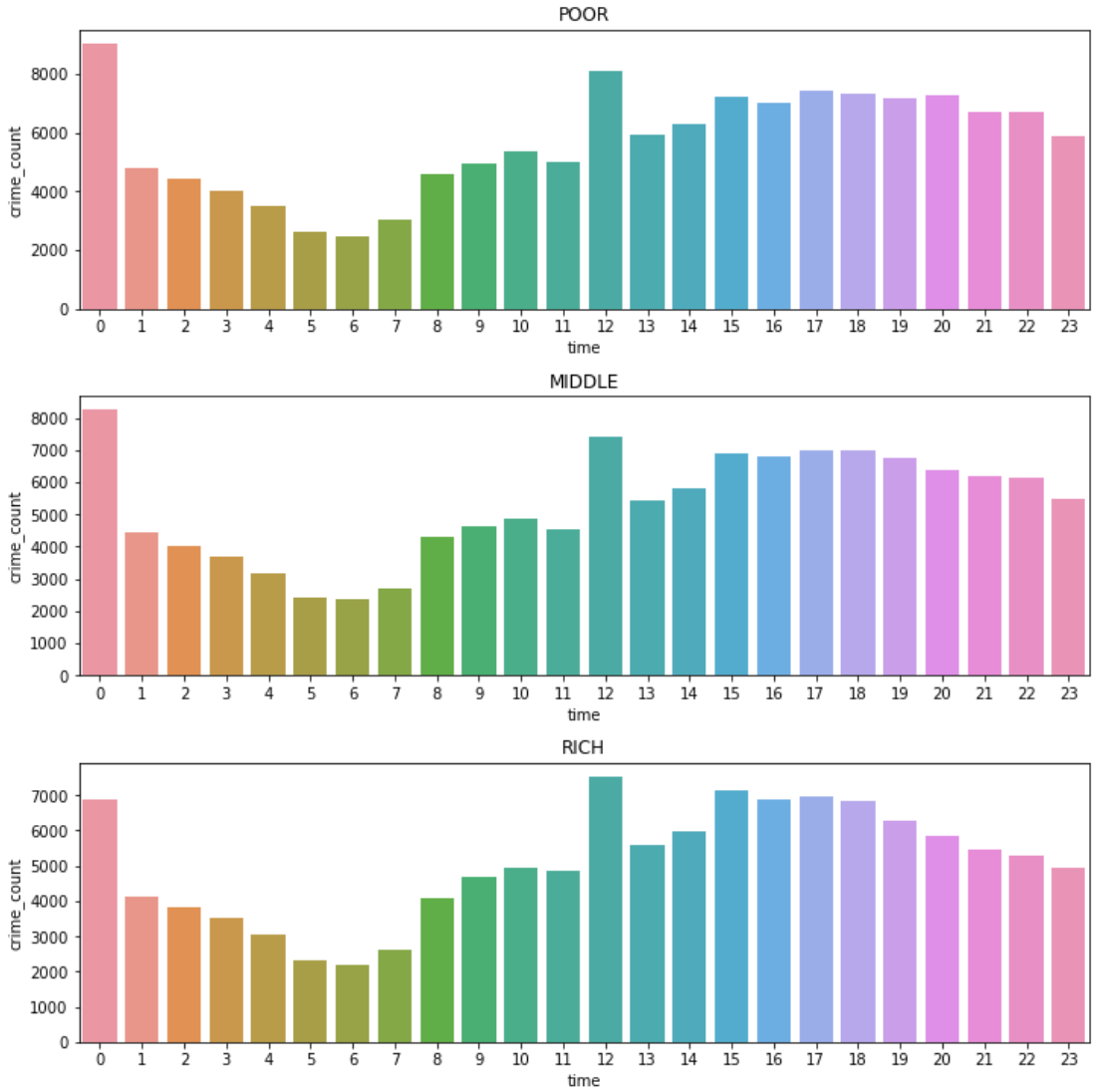
To investigate this problem, we should first classify our house prices into 3 classes (Poor, Median, Rich). To do so, I used a percentile function in SparkSQL to find the 33.3%, 66.6% value of house price, which is as follows:

percentile_approx(median_price, (1 / 3), 10000)	percentile_approx(median_price, (2 / 3), 10000)
418700.0	640479.0

Also, since our original time data format is like "12-00-00", I used cast(LEFT(CMPLNT_FR_TM,2) as int) to get the first 2 digit (12) as the time stamp.

Then, I used a (CASE…WHEN…END) structure to do the classification in SparkSQL. After that, I used GROUP BY class, time, and used COUNT(*) to count the frequency of crime. The result and its visualization are shown below:

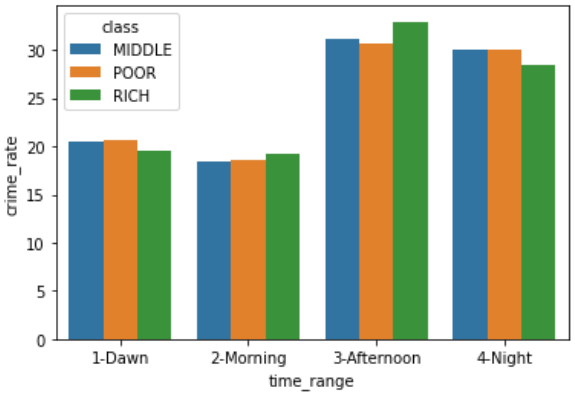
class	time	crime_count
MIDDLE	0	8263
POOR	0	9059
RICH	0	6876
MIDDLE	1	4444
POOR	1	4767
RICH	1	4130
MIDDLE	2	4035
POOR	2	4429
RICH	2	3836



As we can see, there's no clear difference among different classes, but we can still observe some small difference. For example, for rich neighborhoods, the felony crime rate will decrease a lot between 18-24. To make this relationship clearer, I further classified the time data as follows:

```
(SELECT cast(ADDR_PCT_CD as int) as precinct, CASE
WHEN cast(LEFT(CMPLNT_FR_TM,2) as int)<6 THEN '1-Dawn'
WHEN cast(LEFT(CMPLNT_FR_TM,2) as int)>=6 AND cast(LEFT(CMPLNT_FR_TM,2) as int)<12 THEN '2-Morning'
WHEN cast(LEFT(CMPLNT_FR_TM,2) as int)>=12 AND cast(LEFT(CMPLNT_FR_TM,2) as int)<18 THEN '3-Afternoon'
WHEN cast(LEFT(CMPLNT_FR_TM,2) as int)>=18 AND cast(LEFT(CMPLNT_FR_TM,2) as int)<24 THEN '4-Night'
```

class	time_range	crime_rate
MIDDLE	1-Dawn	20.56
MIDDLE	2-Morning	18.44
MIDDLE	3-Afternoon	31.06
MIDDLE	4-Night	29.95
POOR	1-Dawn	20.72
POOR	2-Morning	18.58
POOR	3-Afternoon	30.70
POOR	4-Night	30.00
RICH	1-Dawn	19.48
RICH	2-Morning	19.16
RICH	3-Afternoon	32.91
RICH	4-Night	28.45



It can be observed that Rich class has lowest crime rate at night, but it has highest crime rate on afternoon. Therefore, our police should pay more attention to the social security of afternoon in rich zone. Also, it can be seen that most criminal incidents happen on afternoon and at night. Therefore, our government should invest more police strength on afternoon and night.

V. Challenges and Conclusion

The biggest challenges I met are in the data manipulation process. The main reason is that my datasets are from different sources without cleaning, so there exists many mismatching problems when joining two tables. That's also the reason why I have such a long project report. I spent most of my time on complicated data manipulations.

The first challenge I met is joining these 2 datasets that were geographically related to each other (through neighborhood and precinct). It was very hard to build a map in between. My solution is to find extra map dataset of neighborhood and precinct, and use python packages to find the intersection of these 2 areas.

The second challenge is to join two neighborhood name datasets from different source. Since they are from different source, there exists some minor difference in between, and I can't easily JOIN the table by the same key. For example, the key is "ARROCHAR-SHORE ACRES" on left side and "Shore Acres" on right side. I will lose a lot of data by joining them without preprocessing. Therefore, I used a similarity function here to match the "most similar neighborhood name", and created an intermediate table which serves as a bridge between two datasets.

In conclusion, criminal incidents, especially felony crimes, are related to house price of neighborhoods. Therefore, it's reasonable for our government to distribute different police strength to each neighborhood based on the house price, time, and race of criminals.