Exploration on New York Crime Open Data Based on PolyGamy Thoughts

BINQIAN ZENG (BZ866), LEI XU (LX557), AND YANG SUN (YS2603)

Center of Data Science, New York University

Compiled May 10, 2017

New York City (NYC), the most populous city in the United States, comprises 5 boroughs: Bronx, Brooklyn, Manhattan, Queens and Staten Island. The city has been considered as the world's center of finance, technology and culture. More opportunities come with more risks. In this project, we will follow the process of polygamy[1] taught in lecture and analyze data of NYPD crime: investigate spatial and temporal patterns of various categories of reported crimes, bring up potential causes and verify hypotheses with other datasets. To be more specific, we will dig into the relationship between the number of recorded crimes and yearly economic trend, weather conditions and many other geographical information. Through our exploration, we found that there exist temporal patterns for crime occurrence and interesting spatial distribution. We will first discuss the data preparation part in which we checked data validity, then we cleaned the data, find interesting features, look into correlations, conduct statistical test and show the temporal and spatial characteristics.

1. INTRODUCTION

The primary dataset we used is NYPD Complaint Data Historic (Public Safety) which has more than 5.1 million rows and 20 columns. The big dataset contains large amount of information which might be useful for the future New York City planning Among all patterns we discovered,we first discuss a possible pattern between the crimes and unemployment rate. Then we look at the correlation between average monthly temperature and different categories of crimes by running statistical test. Finally, we discuss a surprising relationship between the duration gap to complete one case and the number of recorded crimes in each borough; we also investigate impact of the geographical information on crime rates.

2. PART I DATA UNDERSTANDING

A. Dataset Description

The main dataset for this project is crime dataset from NYC Open Data. The dataset includes records of felony, misdemeanor, and violation crimes reported from 2006 to 2015. For the second part, we will also use other public data files to explore interesting feature. The sizes of datasets are shown in the above table.

For details of data quality issue and a brief data summary, refer to Appendix A.

Names	Rows	Columns
Crime	5,101,231	24
Temperature	13	5
Unemployment Statistics	15	13
GDP Growth	260	70
NY Population and Num of Crimes	65	12

B. Data Preparation

Based on the conclusion from our data quality issue exploration, we first clean data in preparation for further analysis. We reconstruct our dataset as follows.

1. **Handling Invalid Data** During the first part of this project, we checked the validity of the data. Using conclusion, we first removed invalid data. For example, after aggregating the occurrences of crime related to the crime date, we filtered out invalid reporting date.

2. **Handling Multiple Spaces in Text** In part 1, we found that some columns may have more than one spaces which give us an incorrect unique value. For example, in column LOC-OF-OCC-DES, the correct unique value should be 5, but the

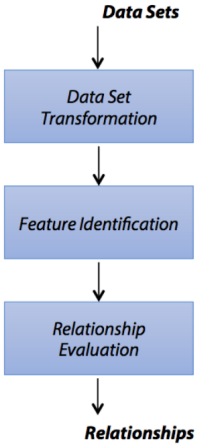
Spark.SQL returns 6 because multiple spaces are considered as one. We wrote scripts and combined the multiple spaces before further exploration work.

3. **Handling Mis-Mapping** In part 1, we also detected that some attributes correspond to multiple classification codes, such as 'OFNS_DESC', which may be caused by typo. In the second part of the project, we fixed some problem, such as combining 'OFNS_DESC' ends with '..law' and '..la'.

3. PART II DATA EXPLORATION

A. Overview

Following the process in polygamy algorithm[1], our work flow is basically as below:



A.1. Dataset Transformation

To explore the temporal and spatial pattern, we first aggregate available data attributes related to time and location. We use PySpark to finish this part of work.

A.2. Feature Identification

In this part, we brainstorm the attributes that we can explore, analyze potential causes, and come up with the features to be tested.

A.3. Relationship Evaluation

To test relationship between features, we draw plots and detect patterns. For those which indicates no obvious pattern during plot, we will not move on to the next stage. For patterns clearly shown in the plot, we run statistical analysis and test if the Pearson Correlation or ANOVA-test value is significantly different from zero.

With the idea above, we explored the dataset from temporal and spatial perspectives with weather, economy dataset. We have the result as below.

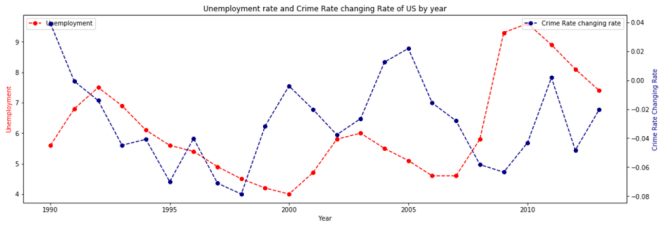
B. Experimental Techniques and Methods

Due to the massive scale of the dataset, big data techniques become essential, and we used Pyspark (including PysparkSql) in Dumbo on NYU HPC to carry out data analysis in our project.

C. Temporal Exploration Work

C.1. Unemployment Rate vs Crime Rate

1. **Pattern Plot**



It seems that there are correlation between two variables, especially around 2007, when unemployment rate rises, the crime rate drops. We will implement statistical test to testify our hypothesis.

2. **Hypothesis**
There is a correlation between crime and unemployment rate.

3. **Correlation**
Using Pearson correlation, we have results as follows As

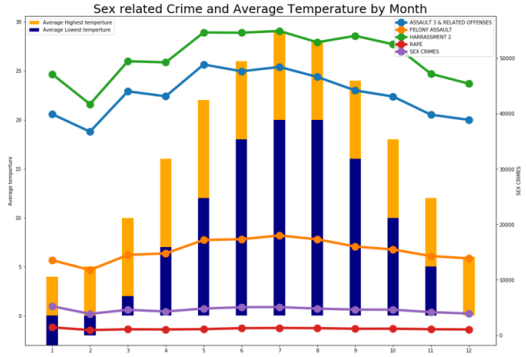
Pearson correlation	p-value
−0.4213	0.1969

shown in the above table, even though the Pearson Correlation indeed negative as claimed in the hypothesis, but the p-value is high. We cannot conclude that there is a significant relationship between crime rates and unemployment rate.

C.2. Temperature vs Crime Rate

1. Sex-related Crime versus Average Temperature by Month

(a) **Pattern Plot**



It is quite obvious that there are some correlation between two variables. We will implement statistical test to testify our hypothesis.

(b) **Hypothesis**
There is a positive correlation between sex-related crime and temperature.

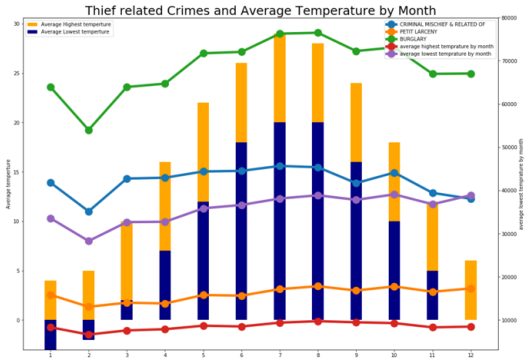
(c) **Correlation**
Using Pearson correlation, we have results in the following table. During data exploration, we found that crimes take place in January is abnormally high. The reason is that if one does not remember the exact data of crime, the record would be filled as in January. Therefore, when we calculate the correlation, we remove the data of January.

Pearson correlation	p-value
0.9009	0.0002

As shown in the table, the p-value is 0.00015, which is much smaller than 0.01, so we can conclude that the sex-related crime is positively correlated with temperature with 99% confidence level.

2. Thief-related Crime versus Average Temperature by Month

(a) **Pattern Plot**



It is quite obvious that there are correlation between two variables. We will implement statistical test to verify our hypothesis.

(b) **Hypothesis**
There is a positive correlation between number of larceny-related crimes and temperature.

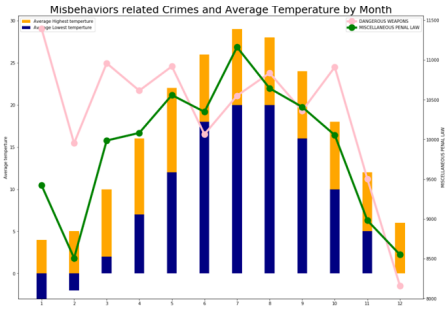
(c) **Correlation**
Using Pearson correlation, we have results as follows. We adopted the same approach as above.

Pearson correlation	p-value
0.8276	0.0017

As shown in the table, the p-value is 0.00166, which is much smaller than 0.01, so we can conclude that the thief-related crime is positively correlated with temperature with 99% confidence level.

3. Misbehavior Crime versus Average Temperature by Month

(a) **Pattern Plot**



It is quite obvious that there are correlation between two variables. We will implement statistical test to testify our hypothesis.

(b) **Hypothesis**
There is a positive correlation between crimes related to misbehaviors and temperature.

(c) **Correlation**
Using Pearson correlation, we have results as follows. Similarly, We removed the data for January.

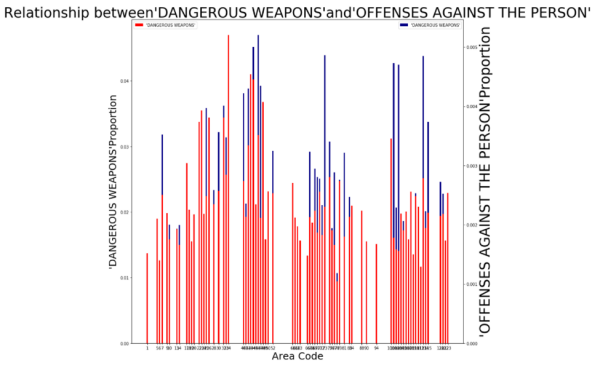
Pearson correlation	p-value
0.7212	0.0122

As shown in the table, the p-value is 0.0122, which is smaller than 0.05, so we can conclude that the misbehavior-related crime is positively correlated with temperature with 95% confidence level.

D. Spatial Exploration Work

D.1. Crimes of Dangerous Weapons and Offense Against the Person at two regions

1. **Pattern Plot**



We drew numbers of kinds of crimes over 74 precincts. From all graphs, it is quite obvious that Dangerous Weapons and Offense Against the Person have similar distribution. We will implement statistical test to testify our hypothesis.

2. **Hypothesis**
Two categories of crime has similar distribution among different areas.

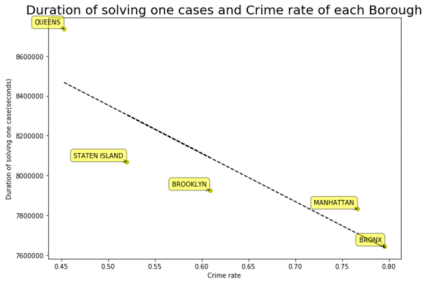
3. **Correlation**
Using Pearson correlation, we have results as follows As

Pearson correlation	p-value
0.4910	5.7896×10^{-6}

shown in the table, the p-value is 5.7896×10^{-6}, which is very close to 0, so we can conclude that our claimed hypothesis is true with 95% confidence level.

D.2. Efficiency of Police vs Crime Rate

1. **Pattern Plot**



As shown in the above plot, the relationship between crime rates and average time period to complete a case is almost linear over five boroughs . We will implement statistical test to testify our hypothesis.

2. **Hypothesis**
   There is a negative correlation between crime rates and average time period to complete a case.

3. **Correlation**
   Since we are comparing the difference between 5 boroughs, we are using ANOVA test. The results are as follows,

ANOVA	p-value
1852	9.4×10^{-11}

From the above table, the p-value is much smaller than 0.01, so we can conclude that there is a significant difference in terms of case handling efficiency between five boroughs.

D.3. Crime Distribution and US citizen Distribution

1. **Pattern Plot**

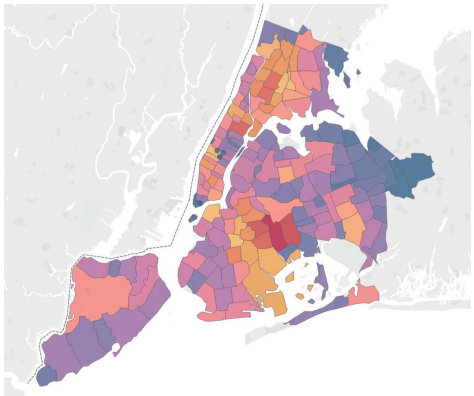


Figure C3.1 Crime Distribution over different areas by zip-codes

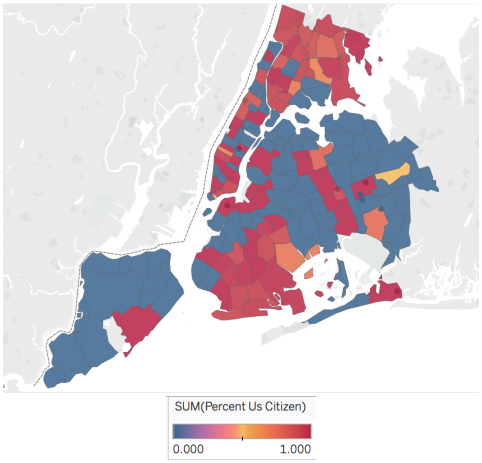


Figure C3.2 US Citizen Percentage over corresponding areas

We plot each variable in different districts over different zip-codes. We can see in certain areas where the crime frequency is extremely high, usually the US citizen percentage is low, which reflects the potential influence of floating population. We will implement statistical test to testify our hypothesis.

2. **Hypothesis**
   Floating population has a positive effect on recorded number of crimes, the more the proportion of residence of citizen is, the less likely the crime is going to occur.

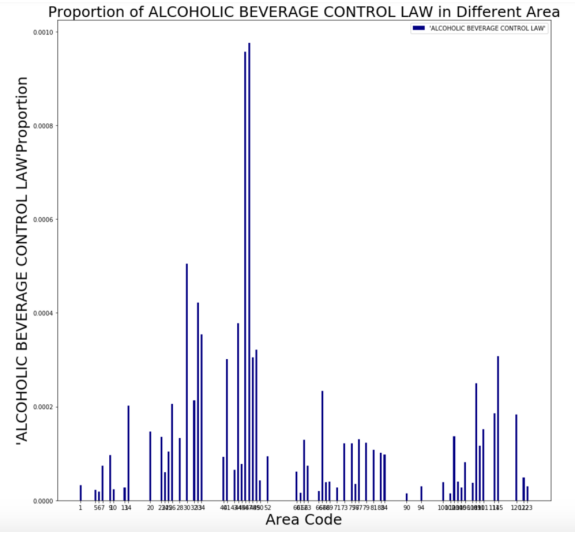
3. **Correlation**
   Since the pattern over all areas will be influenced by some areas where no US citizen gathering, we filtered those areas with population percentage lower than 0.03. The results are as follows,

Pearson correlation	p-value
0.5812	0.0483

From the above table, the p-value is slightly smaller than 0.05, so we can conclude that there is a significant influence of floating population on number of crimes at 95% confidence level.

D.4. Crimes of Dangerous Weapons and Offense Against the Person at two regions

1. **Pattern Plot**



From the bar chart we found that two precincts, 46 and 75, have an extremely high proportion of alcoholic beverage control law crime.

2. **Potential Cause**
Precinct 46 includes an area called University Heights, where many universities and colleges are located. Students in universities may drink and party more often than average.

Precinct 76 is Red Hook in Brooklyn. From the Brooklyn Safety Map provided by Google[9], we can see that the most crime-dense area is Red Hook Houses. A federally funded project overseen by the Roosevelt Administration, the rent was less than 6 dollars month when they opened[10].

There are more work we have done and please see more exploration work and script in the Appendix B.

E. Challenges/Issues

1. Large data amount:
Millions of data restrict the flexibility of applying data wrangling techniques. For example, when we wanted to retrieve the zip-code from latitude and longitude, we tried different python packages such as sklearn.KDTree and geopy. However, due to the lack of packages on Dumbo, we had to switch other ideas of exploring distribution. Also, because of the large amount, we had to debug with sample data first, which might still throw errors when running jobs on the cluster.

2. Data quality issue:
Problems such as mishandling and typos increased the difficulty of getting information from the data, both technically and understanding. For example, when we doing zip-code mapping, it went smoothly for the sample set, but when we run the code on the whole test (which we gave up later since too time-consuming), there is one entry with encode error and several hour's work lost.

4. INDIVIDUAL CONTRIBUTION

Everyone contributed to all tasks, including data cleaning, pattern exploration, statistical tests, result analysis and report writeup.

5. CONCLUSION

1. In this analysis, the most important data quality issue is missing data and inconsistency. In part1, we use some filtering and clustering method to solve these issues. In this stage, we use bigdata tool PySpark and PySparkSql to clean original dataset efficiently.

2. In part 2, we use PySpark to aggregate the information we need and visualize the pattern and potential relation by Matplotlib and Geoplot. Furthermore, we use coorelation analysis to test if our hypothesis is right.

3. In the end, we found that crime activities are associated with factors like labor participation rate, temperature, characteristics of different areas. Labor participation rate are negatively associated with crime activities over time. Temperature are positively associated with crime activities over time. Some kinds of crime are coorelated with characteristics of areas which make some areas have higher proportion of certain kinds of crimes than other areas. In addition, we found an interesting pattern that there is a negative correlation between crime rates and average time to complete a case.

REFERENCES

1. Fernando Chirigati, Harish Doraiswamy, Theodoros Damoulas, Juliana Freire, Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets
2. NYC Open Dataset, NYPD_Complaint_Data_Historic *https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i* (2005-2015).
3. NYC Open Dataset, The population of the boroughs of New York City *https://www.citypopulation.de/php/usa-newyorkcity.php*
4. NYPD Crime *http://www.nyc.gov/html/nypd/html/home/poa_crime.shtml*
5. New York Population and Number of Crimes *http://www.disastercenter.com/crime/nycrime.htm*
6. World Bank National accounts data *http://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?start=2006*
7. Local Area Unemployment Statistics *https://www.bls.gov/lau/data.htm*
8. New York City Temperatures: Averages by Month *https://www.currentresults.com/Weather/New-York/Places/new-york-city-temperatures-by-month-average.php* 2006
9. Brooklyn Safety Map *(https://www.google.com/maps/d/viewer?mid=1TOOhlgGoaANKGleiKq1Fw &hl=en&ll=40.674595900531656%2C-74.00590075973508&z=14)*
10. Red Hook Houses Infomation *(http://www.atlasobscura.com/places/red-hook-houses)* 2006

A. Acknowledgement

We would like to thank Prof. Juliana Freire, who prepares this comprehensive course. Also, we would like to thank Dr. Nicholas Knight and Dr. Erin Carson, who provide us with excellent lab guidance, and to our classmates who give us many helpful suggestions.

Appendix A

	Column Name In csv	Base Type	Semantic Data Type	# of Unique Value	# of Valid Value	# of Invalid Value	# of Null Value
Col_1	CMPLNT_NUM	Integer	ID	5,101,231	5,101,231	0	0
Col_2	CMPLNT_FR_DT	Date	Occurring date	6,317	5,100,576	0	655
Col_3	CMPLNT_FR_TM	Time	Occurring Time	1,442	5,100,280	903	48
Col_4	CMPLNT_TO_DT	Date	Ending date	4,827	5,101,053	31	147
Col_5	CMPLNT_TO_TM	Time	Ending time	1,441	4,885,891	215,328	12
Col_6	RPT_DT	Date	Report date	3,652	5,101,231	0	0
Col_7	KY_CD	Integer	Offense classification code	74	5,101,231	0	0
Col_8	OFNS_DESC	String	Offense code description	70	5,082,391	0	18,840
Col_9	PD_CD	Integer	Internal classification code	415	5,096,657	0	4,574
Col_10	PD_DESC	String	Internal classification code description	403	5,096,657	0	4,574
Col_11	CRM_ATPT_ CPTD_CD	String	Crime status	2	5,101,224	0	7
Col_12	LAW_CAT_CD	String	Level of offense	3	5,101,231	0	0
Col_13	JURIS_DESC	String	Responsible Jurisdiction	25	5,101,231	0	0
Col_14	BORO_NM	String	Borough	5	5,100,768	0	463
Col_15	ADDR_PCT_CD	Integer	Precinct code	77	4,286,476	814,365	390
Col_16	LOC_OF_ OCCUR_DESC	String	Location of event	6	3,973,890	213	1,127,128
Col_17	PREM_TYP_DESC	String	Premise Type Description	70	5,067,952	0	33,279
Col_18	PARKS_NM	String	Park name	863	7,599	0	5,093,632
Col_19	HADEVELOPT	String	Housing Development	278	253,205	0	4,848,026

Col_20	X_COORD_CD	Integer	NY X-coordinate	69,532	4,913,085	0	188,146
Col_21	Y_COORD_CD	Integer	NY Y-coordinate	72,316	4,913,085	0	188,146
Col_22	Latitude	Decimal	Latitude	112,803	4,913,085	0	188,146
Col_23	Longitude	Decimal	Longitude	112,807	4,913,085	0	188,146
Col_24	Latitude, Longitude	Decimal	Latitude, Longitude	112,826	4,913,085	0	188,146

Appendix B

https://github.com/xlxulei1005/BigData2017Project