Chicago crime rate vs. New York crime rate

Hai Anh Le
Neha Shashidhara
Jeevan Sai
MSIS graduate students
California State University, Los Angeles

Abstract: This project will illustrate the usage of Hadoop, MapReduce, and Hive on big data. We will utilize the knowledge learned in class, extensive researches and development of HiveQL in order to generate data and visualize it on Tableau. Specifically, we are using Chicago Crime Rate Data Portal and NYC Open Data as the foundation to compare and generate results. Elements of this project include a report paper, a tutorial on the queries, and one group presentation.
URL: https://data.cityofnewyork.us/api/views/qgea-i56i/rows.csv?accessType=DOWNLOAD&bom=true&format=true
https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD&bom=true&format=true
Dataset size: 2.98GB
Cluster version: IOP4.2
No of nodes: 2 (management and data nodes)
Memory size: RAM 72GB, data disk 1TB SATA
CPU Speed: 2.10GHz

1. Introduction

Based on the generous list of data provided by our instructor, we have done some researches and exclusively decided which data we are using for this project. We are going to manipulate and filter through the two datasets below following with steps:

- Chicago Crime Rate; data size is 1.52GB.
- New York Crime Rate; data size is 1.46GB.
- Narrowing down the information to have a detailed comparison in safety between the two cities.
- From each dataset, sort out the type crimes, to see which locations those crimes happen often, whether it is inside a residence or outside
- Conclude which city is safer in terms of the highest crime rates and the lowest ones, and yearly crime rate in both cities.
- The tools we are using is HiveQL, Putty, IBM Bluemix BigInSights, and Tableau.

2. Manipulating datasets

2.1 Tools and data processing

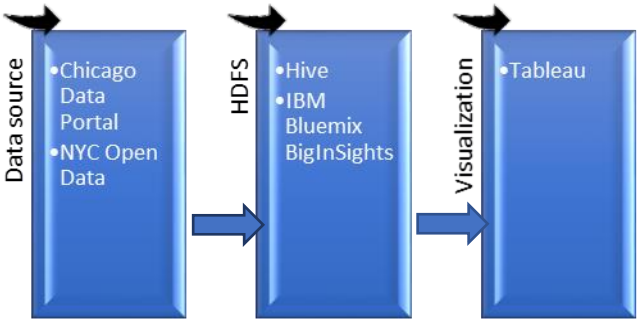

Figure 1. Data processing.

- In order to unify the time of the crime rate data in Chicago and New York, we have chosen to analyze from 2006 until 2017 which is the present.
- Basic commands to connect to the data source such as wget, mkdir, hive were also used to start up the project. We sucessfully downloaded and uploaded data from the Chicago Data Portal and NYC Open Data to HDFS file system. From then, we connected to Hive and started writing more complex codes designed for each dataset.
- For both of the large datasets, we picked out the top 5 streets for the highest crime rates, the top three locations for the crimes, and the crime rate yearly.

2.2 Chicago Crime Rate

In the Chicago dataset, after creating tables and running queries, we received very inquisitive and interesting results.

Table 1. Chicago filtered crime rate from 2006 till 2017.

Top three locations		Top five streets		Top three types of crimes	
Type of locations	Number of crimes	Name of the streets	Number of crimes	Type of crimes in the top 5 cities	Number of crimes
Street	294157	W Ohare St	2828	Felony	795
Residence	188366	N State St	2268	Misdemeanor	1667
Apartment	90788	S Cicero Av	1275	Violation	366
		S Federal St	1122		
		N State St	1005		

- The top 3 types of locations: street, residence, apartment.
- Afterwards, we also narrowed the data down to retrieve the top streets that these types of locations reside in:

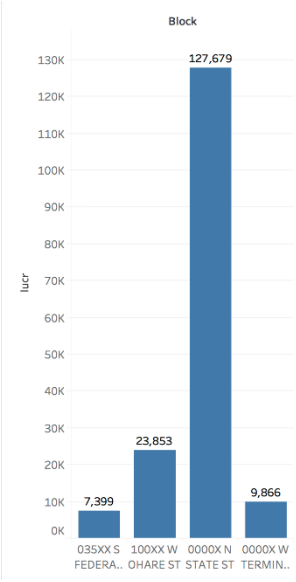

Figure 2 Chicago unsafe areas.

To put the whole result in a deeper perspective, we wrote queries to acquire that Ohare St. is the most dangerous street in these 3 crime categories:

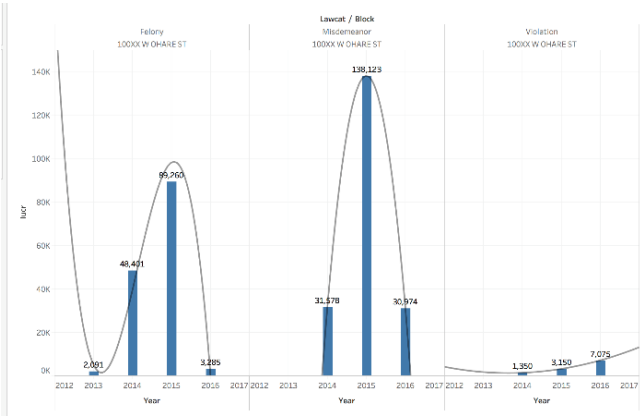

Figure 3. Ohare st. crimes in Chicago.

We also analyzed the crime rate yearly to discover the trend of the crimes in Chicago. 2015 is the year with the highest crime in both categories of felony and misdemeanor. Then in 2016, there is a significant decline in the crime rate.

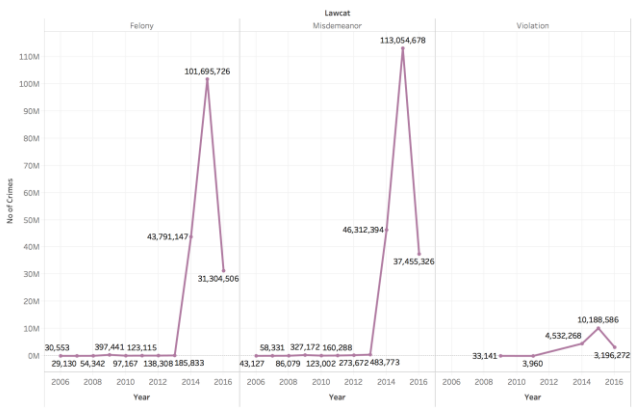

Figure 4. Yearly crime rate in Chicago.

Felony is the combination of sexual assaults, assaults, sex offense, criminal damage, homicide, etc. Misdemeanor includes ritualism, robery, theft, weapon violations, etc. Lastly, violation is a compilation of arson, carry license violation, gambling, public indecency, etc.

2.3 New York Crime Rate

Table 2. NYC filtered crime data from 2006 till 2017.

Top three locations		Top five streets		Top three types of crimes	
Type of locations	Number of crimes	Name of the streets	Number of crimes	Type of crimes in the top 5 cities	Number of crimes
Street	320256	Brooklyn	315635	Felony	76455
Residence	229485	Manhattan	244739	Misdemeanor	141374
Apartment	97445	Bronx	227473	Violation	26910
		Queens	211958		
		Staten Island	48743		

The top cities that the 3 crimes happened the most in New York city are below:

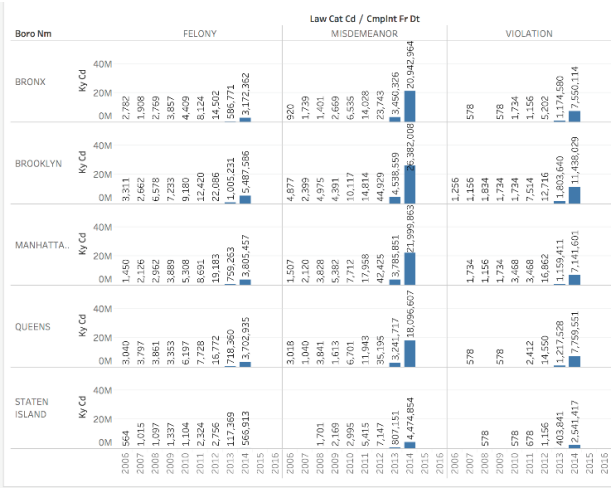

Figure 5. The most dangerous areas in NYC based on the 3 crimes.

Brooklyn is the most dangerous city in these 3 crime categories in New York City:

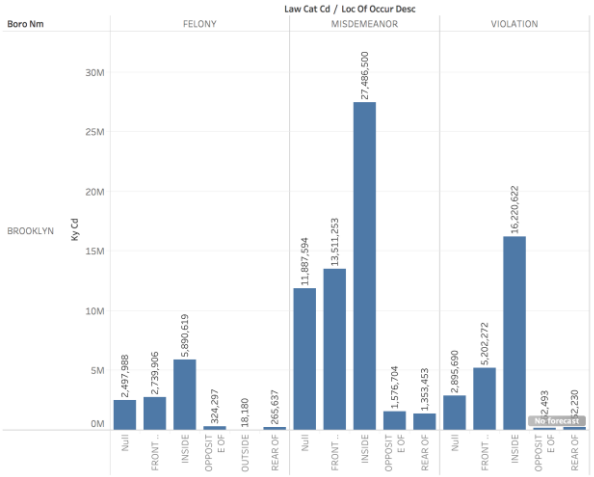

Figure 6. Brooklyn has the highest crime rate.

Until this point of the data analyzing process, we realized that in each individual area listed for Chicago and New York city, New York city areas seem to have the highest crime rates:

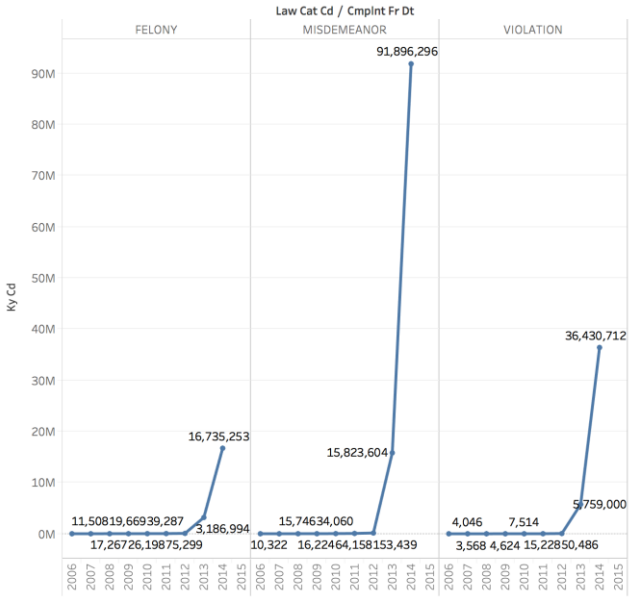

Figure 7. Top 3 crimes in NYC.

2.4 Comparing and forecasting

a. Comparing

As we have seen, Ohare street and Brooklyn are the representatives for Chicago and New York city in the 3 crime categories of felony, misdemeanor, and violation. Brooklyn has a significant higher number of crimes compared to Ohare. With misdemeanor, Ohare street has 138,123 number of crimes and Brooklyn has 27,486,500 number of crimes. We could easily assume that with all the misdemeanor crimes, New York city is more dangerous.

However, after analyzing total crimes yearly in both cities (Figure 8 and 9), we have come to a counter conclusion.

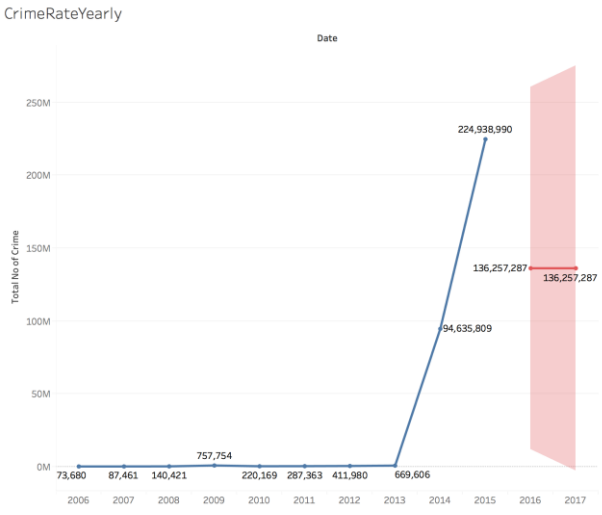

Figure 8. Chicago crime rate yearly.

From 2006 to 2017, in Chicago, the total amount of crimes in the 3 categories is 224,938,990. (Figure 8)

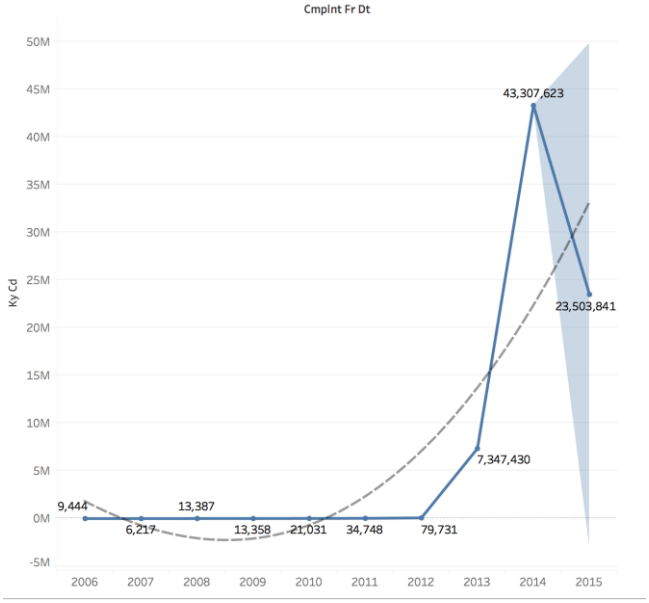

Figure 9. NYC crime rate yearly.

On the contrary with Chicago, from 2006 to 2017, New York has the total amount of crime rate in the 3 crime categories is only 43,307,623. It is only 19% compared to Chicago. Hence, Chicago has a way higher crime rate that New York city does in term of felony, misdemeanor, and violation from 2006 to 2017. This is a surprising result after assuming that since Brooklyn is more dangerous than Ohare, then New York city has higher crime rate.

b. Forecasting

In the recent years, the 3 crime categories in Chicago and New York city have tremendously decreased and seemed to continue the trend. At the same time, Chicago remains having a higher total crime amount over New York city. For the future usage and research, our queries could automatically update data from Chicago data portal and NYC open data as their data updates rather frequently. This data analysis could be utilized by the police department in both cities to determine which area needs security tightening. It could also provide insights for people who study criminal justice or have related researches to this matter.

3. Summary

- We successfully used many tools learned in class such as HiveQL, IBM Bluemix, and Tableau to use and manipulate data.
- Individually, New York city has Brooklyn as the region with the higher crime rate than Ohare street of Chicago.
- In total, Chicago turned out to have a higher yearly crime than New York City.
- Crime rates are declining, and Chicago crime rate remains higher than New York city.

4. Github URL

5. References

http://www.calstatela.edu/centers/hipic/related-site

https://data.cityofnewyork.us/api/views/qgea-i56i/rows.csv?accessType=DOWNLOAD&bom=true&format=true

https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD&bom=true&format=true

https://console.bluemix.net/data/bic/