Crime analysis of New York City and Safe Borough Recommendation system

Jahnvi Arya
Computer Science
New York University
New York, USA
ja4158@nyu.edu

Rachit Jain
Computer Science
New York University
New York, USA
rj2219@nyu.edu

Shobhit Sinha
Computer Science
New York University
New York, USA
ss13881@nyu.edu

***Abstract*—Burglary has been replaced in NYC by robbery, which is followed by felonies. There are numerous criminal patterns that can be connected to the five boroughs of New York City. The authors use the complaints, arrests, and payroll dataset, which is available to a wide range of agents and stakeholders for usage in a variety of different applications. The authors process and analyze the data using Hive and the MapReduce architecture. The authors of this research want to examine the relationship between the quantity of complaints and the number of borough-level arrests. In order to understand the relationship between crime and the purchasing power of the neighborhood, the authors also intend to evaluate the pattern of city-wide payroll data at the borough level. The authors also recommend a safety index for a person to stay in based on various factors.**

***Index Terms*—NYC crime, complaints, arrests, payroll, fiscal, safety, recommendation, hive, mapreduce, hadoop, borough**

I. Introduction

New York City's crime index is 19, making it safer than 19% of American cities. The crime rate is 25.80 per 1000 people. About 5.80 violent crimes are committed per 1,000 people, which is 1.80 more than the national median. Burglary was the most common crime 20 years ago, according to statistics. It changed to robbery over the following ten years, and today it is felony assault. Due to changes in the city's population, judicial system, and demography, many components of these crime figures may have been updated.

NYC data for complaints, arrests, and payroll for a fiscal year is well documented and provides a lot of information. The features which are present in the data are offense description, victim age [1], perpetrator age, complaint date, base salary, etc. These features give the authors a rich representation of what kind of data is present and what kind of analysis can be computed on this.

The project consists of data from various NYC government sources and produces results based on complaints, arrests, and payroll information. There are a few major analysis that the authors depict in the project. The authors analyze types and severity of complaints in different boroughs in New York City [2]. The authors also analyze the arrests at a borough level and find correlation between the number of complaints to arrests made at a borough level. From the payroll data source, the authors analyze the city-wide payroll data at a borough level to understand the correlation between crime and average income of the boroughs. The authors then provide a recommendation engine based on an individual's age, working hours, and pay recommending the borough that can be chosen to live in.

The paper is divided into the following sections. Section II discusses the motivation behind this project. Section III then defines our proposed idea and Section IV discusses about the experimental setup. Section V delves deep into the methodology for the paper. Following this, Section VI talks about the results. Section VII discusses the coding challenged and Section VIII discusses the obstacles that we encountered in the project. Finally, in Section IX, we talk about the conclusions and Section X discusses the future scope of this work.

II. Motivation

The authors use the complaints, arrests, and payroll dataset, which is available to a wide range of agents and stakeholders for usage in a variety of different applications. Since the data is rich in features, there can be multiple queries that can be made (even from the raw data) to get various analytic points. In this section, the motivation of the project is discussed and what groups can be aided by this analysis.

As an individual, when one moves to a new city, they do not have much information regarding what kind of place they can live in. There are certain parameters that need to be considered before deciding on a place to live. One of those parameters is how safe a place is. The authors determine the most important individual level factors that need to be considered while moving and provide a recommendation based on that.

As a government official overseeing the functioning of the police [3], this analysis can be used to determine how effective the police force actually is in that area. Using the complaints and arrests data sources together, an analytic can be made based on how well the police forces are functioning in the boroughs. This can be used to improve the police forces. It can also be used to understand where the crime rate is more and accordingly deploy more forces at that time and place.

As a business, having the payroll [4] information can be utilized based on how the living situation is in these boroughs. The businesses can make decisions as to if it is required to adjust the pay based on the borough. Businesses can also use

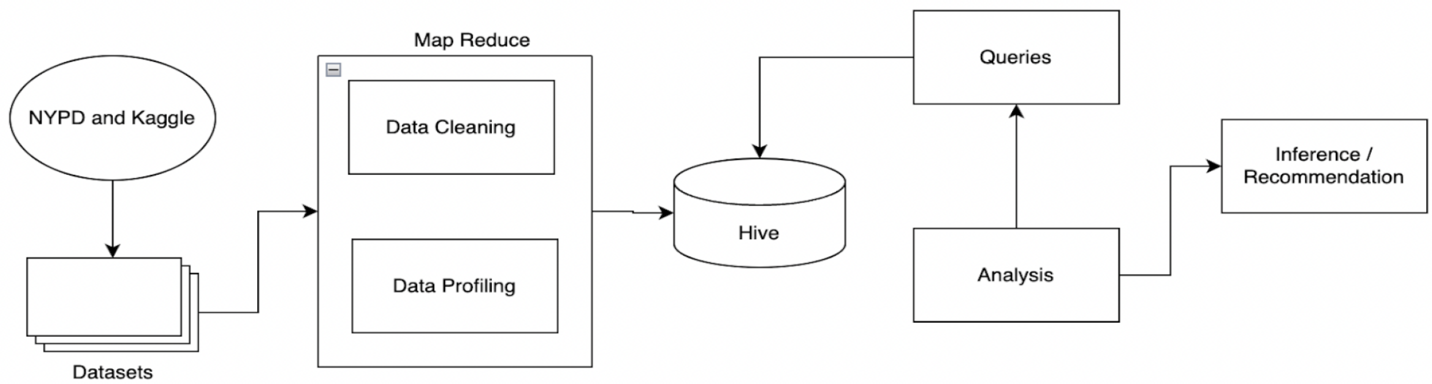

Fig. 1: Design Diagram

this analytic for placement of their offices. As a business, they would want to be situated in the safer parts of the neighbourhood. They can also decide their working hours to maximise safety of their employees.

In general, our analytical findings would aid social, financial, and government analysts in their research and monitoring of company trends, arrest trends, etc. whether they be general or specific to a certain industry or type of organization. Our findings also indirectly assist communities and scholars.

III. Proposed Idea

This study's objective is to analyze how crime varies among New York City's boroughs. Data collection is the first step, and MapReduce [5] is used to process and clean the data. In order to help the writers learn about the data, the MapReduce framework will eliminate pointless rows, drop unimportant features, generate aggregated rows, etc. To finally develop our recommendation system, we worked with three important data sets to produce analytical findings that build upon one another. Our data has been cleansed and enhanced, and it has been imported into Hive, where the authors have written queries to generate their findings. The data are analyzed and conclusions are drawn using various queries. Fig 1 shows a pictorial representation of the proposed method.

IV. Experimental Setup

All the experiments have been carried out on NYU Dataproc. The authors have used google gcloud SDK to login via shell. The data is preprocessed using Map Reduce. Owing to the size of the dataset the authors used Hive for analysis and building the recommendation system. Following are the datasets that were used:

A. NYPD Arrests Data

The authors have used the Arrest Data of NYC [6] to analyze the pattern of arrests. This data contains records representing an arrest effected in NYC by the NYPD and includes information about the type of crime, Perpetrator's age, and borough along with other details. The data ranges from 2006 to the last financial year and is updated quarterly. The dataset has about 5.3M rows and 19 columns. Each row depicts an arrest done by NYPD.

B. NYPD Complaint Data

The NYPD Complaint Data Historic dataset [7] includes all valid felony, misdemeanor and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of the year 2021. This data was collected to study the relationship between the number of complaints in different boroughs of New York City. This dataset has 35 columns and about 7.8M rows.

C. NY Citywide Payroll Data

The public's curiosity about how the City spends its money on salaries and overtime pay for all municipal employees has led to the collection this data. The appropriate user Agencies input data into the City's Personnel Management System ("PMS") [8]. It has data from 2014 to 2020. There are 17 columns and close to 4M rows. The authors are using the data to understand the payroll structure and overtime hours of the employees of the New York City.

V. Methodology

A. Data Understanding

The data understanding phase involved finding the correct data and understanding the useful columns. The complaints data of NYPD contains all the complaints registered with NYPD which serves as an indication of the number of crimes happening at the borough level. Using "VIC AGE GROUP," conclusions could be derived about the age group that society considers to be the most "targeted." The complaint time may provide information about the most dangerous time of day.
The arrests data is a breakdown of each arrest made in NYC by the NYPD from 2006 through the end of the previous calendar year (2021). Before being released on the NYPD website, this information is personally gathered each quarter and examined by the Office of Management Analysis and Planning. A borough's safety may be inferred from the arrest statistics in conjunction with the complaint data. How effective a police department is in a certain borough is determined

by how many complaints result in arrests. The nature of the crime also provides interesting information about the types of offenses that are more prevalent in the borough.
The NY payroll data offers details on the salaries of New York City's municipal employees. It acts as a clue to comprehend the quality of life in a neighborhood. It also sheds insight on the number of overtime hours that residents of each borough worked. The difference in pay between total compensation and overtime compensation is another insight provided by this dataset.

B. Data Preparation

All the datasets were preprocessed using Map Reduce.

1) NYPD Complaint Data:

- Preprocessing started with converting all the ',' with ';' to have the ability to split on ','.
- Many columns have been dropped in the final output since they are not required for our analysis. Relevant columns that helped in our study of Number of Complaints in each year partitioned at borough level are kept. Columns kept for final analysis are - Complaint_Year, Complaint_date, OFNC_DESC, LAW_CAT_CD, SUSP_AGE_GROUP, VIC_AGE_GROUP and all other columns are dropped. The BORO_NM is used as the key for partitioning the data at the borough level.
- To get the Complaint_Time, two columns namely CMPLNT_FR_TM and CMPLNT_TO_TM were analysed. If CMPLNT_FR_TM was present, it was taken as the Complaint_Time otherwise CMPLNT_TO_TM was taken as the Complaint_Time.
- If OFNC_DESC, LAW_CAT_CD, SUSP_AGE_GROUP, VIC_AGE_GROUP are each present in the dataset, they were taken as it as. In case any field was not presnet, the respective field was marked as UNKNOWN. These columns give important information for relevant analysis but are not the basis of our study. Dropping rows for missing data in these columns may reduce the size of dataset and may result in loss of important information for our study.

2) NYPD Arrests Data:

- The initial step was to convert ',' to ';' in descriptions to be able to split on ','.
- There are many columns that we do not need for our analysis and hence have been dropped. The columns that are kept are - ARREST_DATE, OFNS_DESC, LAW_CAT_CD, and AGE_GROUP. The ARREST_BORO is also saved and used as a key in mapper and reducer.
- Since we are doing the evaluation on per year basis we have extracted the year from ARREST_DATE. If the ARREST_DATE is missing then the row is dropped as it is of no use to our analysis.
- The AGE_GROUP, OFNS_DESC, and LAW_CAT_CD have a fallback to "UNKNOWN" in case of missing values. These columns are good to have but are not essential as we don't plan to make them the base of our analysis. Dropping rows based on missing data in these columns will lead to a reduction in data size and the loss of a good chunk of data.
- The ARREST_BORO is converted from initials to the complete value to make it in line with our other data source. B(Bronx), S(Staten Island), K(Brooklyn), M(Manhattan), Q(Queens).
- The LAW_CAT_CD is converted from initials to the complete value to make it in line with our other data source. F(Felony), M(Misdemeanor), V(Violation).

3) NY Citywide Payroll Data:

- The first step of cleaning the data was reading the input file which had, in the descriptions. This was solved by checking if the text was in a string format indicating one column in the data and the comma was converted to a semi-colon.
- If the borough was not an NYC borough, it was not included in the final HDFS cluster and the row was dropped.
- Based on the pay type, following mathematical calculations were done:
  - per annum or prorated annual = base salary + total OT paid + other allowances
  - per day = (base salary * number of days * working weeks) + total OT paid + other allowances
  - per hour = (base salary * number of hours * number of days * working weeks) + total OT paid + other allowances
- The average working hours are 7 and the average number of working weeks are 49.
- Identifying and removing outliers.

C. Analysis

Using map reduce job, data cleaning and data profiling was performed. The profiled data files were then stored in HDFS. In the second phase, Hive was to create tables using the HDFS files, and hive queries were performed for data analysis. Top K queries were performed to get the most prevalent crimes in each borough. A comparative analysis was then performed between overtime hourly pay and standard hourly pay. To study the relationship between crimes in the day vs crimes in the night, total number of crimes in day and night were calculated for each borough.

D. Recommendation System

Through this paper, the authors propose a novel recommendation system which takes income, age, and working hours as input and provides a safety index for each borough. The algorithm works on 4 aspects:

- Safety Ratio (w_1): The safety ratio is determined by taking the ratio of the number of arrests to the number of complaints. Since it's a good measure, the higher ratio is better. It is a measure of the effectiveness of the Police Force in that borough. The ratio indicates how many complaints resulted in arrests.

Feature	Value
Age	18-24
Income	$70000
Working Hours	8 am - 6 pm

TABLE I: Sample Input for Recommendation System

Borough	Safety Index
Brooklyn	0.624
Bronx	0.644
Manhattan	0.773
Staten Island	0.718
Queens	0.628

TABLE II: Safety Index by Recommendation System

- Age (w_2): This is a categorical feature. The age filter is divided into 5 categories:
  - <18
  - 18-24
  - 25-44
  - 45-64
  - 65+

  The authors take the ratio of crimes targeted to the given age group with the total number of complaints in a particular borough. Since this is a bad measure, the lesser the better. This ratio is a measurement of what percentage of crimes happening are against people of the same age group. It gives insights on how safe is the borough for a given age group.
- Income (w_3): Average income is a good measure of the standard of living of a borough. Many studies have found relatively richer neighborhoods to be safer in terms of dangerous crimes involving arms and drugs. The algorithm is designed in a way to give weighted scores depending on the income and average income of the borough.
  - If the given income is more than the average income of the borough then 1 is added to the final score.
  - If the given income is less than the average income of the borough by not more than 10000 then 0.75 is added to the final score.
  - In other cases we add 0.5 to the score.
- Working hours (w_4): The working hours of an individual are a very important contributing factor in deciding the borough to live in. It gives insights on the duration of the day the person is most exposed to crimes. The ratio of complaints during the given working hours to the total number of complaints talks about the probability of a crime happening during the given hours. We take the negation of the ratio.

$$w_{score} = (w_1 + (1 - w_2) + w_3 + (1 - w_4))/4$$

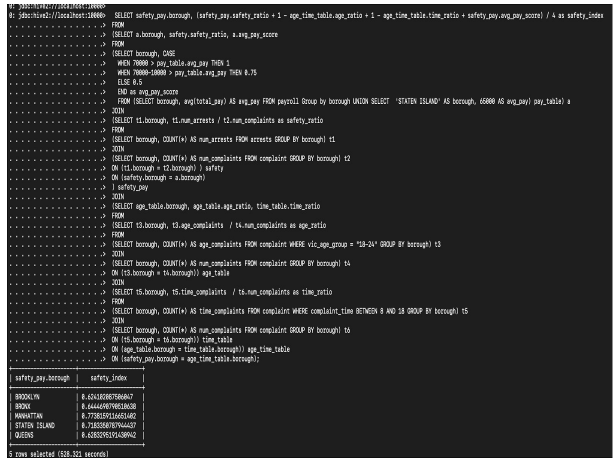

Fig. 2: Recommendation System

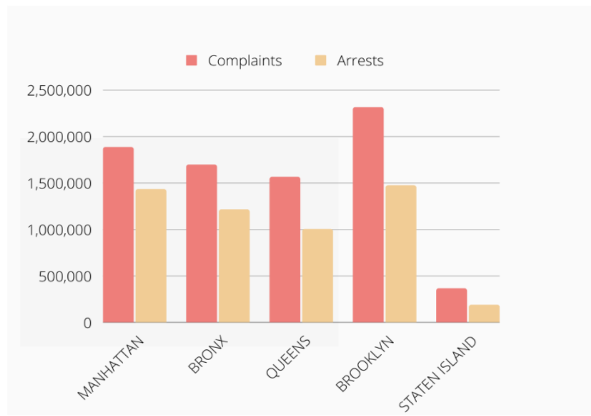

Fig. 3: Arrests and Complaints

VI. Result

Fig. 3 shows the number of complaints and number of arrests in all five boroughs across New york city. It is observed that the number of complaints is more than the number of arrests and this trend is the same across all five boroughs with the number of complaints in Brooklyn being the maximum and the number of complaints in Staten island being the minimum. An important observation is that the number of complaints and arrests is significantly lower in the case of Staten island borough. The ratio of the number of arrests to the number of complaints is maximum in the case of manhattan.

Fig. 4 shows the number of misdemeanor, felony and violation committed across all five boroughs for the complaint dataset. It is observed that the number of complaints for misdemeaner is maximum and the trend is similar for all 5 boroughs. In contrast, the number of complaints is minimum for petty violation across all boroughs. This result looks promising since people don't generally get arrested for petty violations. The trend of misdemeanor to felony to violation is similar for all boroughs.

Fig. 5 shows the number of arrests in case of misdemeanor,

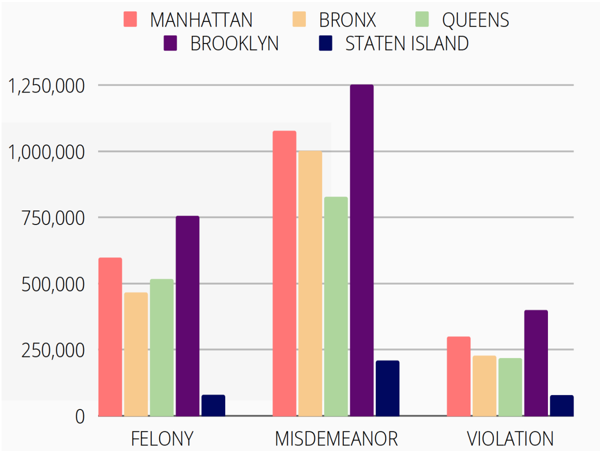

Fig. 4: Complaints

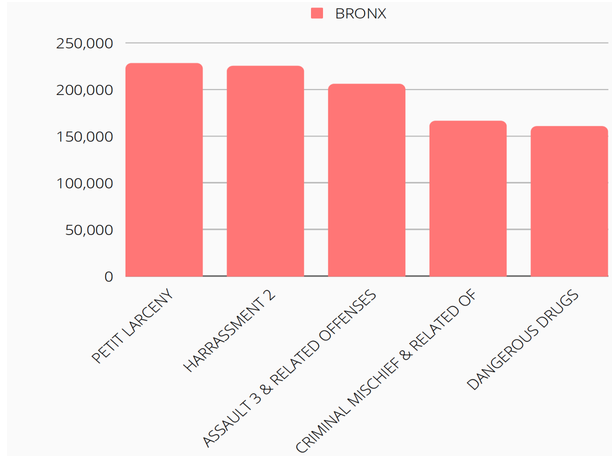

Fig. 7: Top 5 Complaints

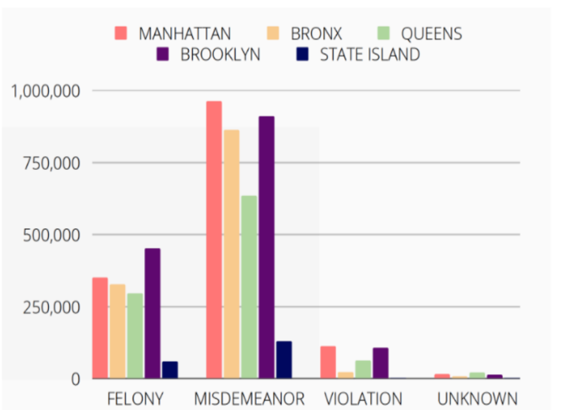

Fig. 5: Arrests

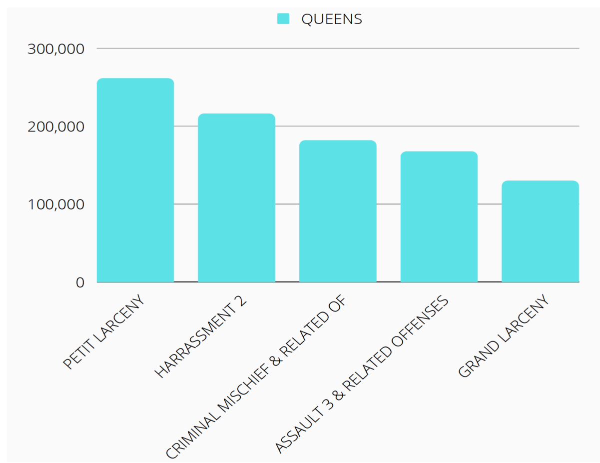

Fig. 8: Top 5 Complaints

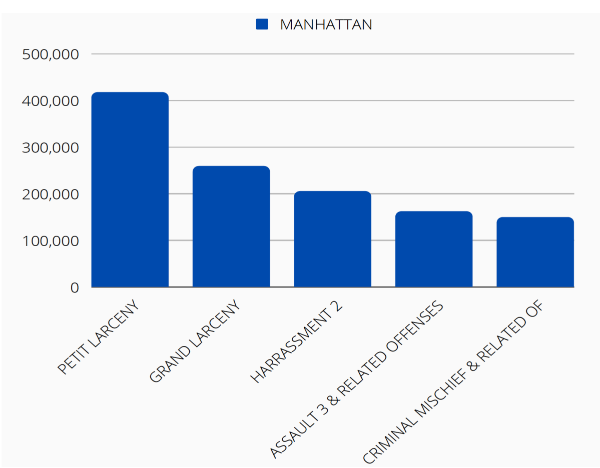

Fig. 6: Top 5 Complaints

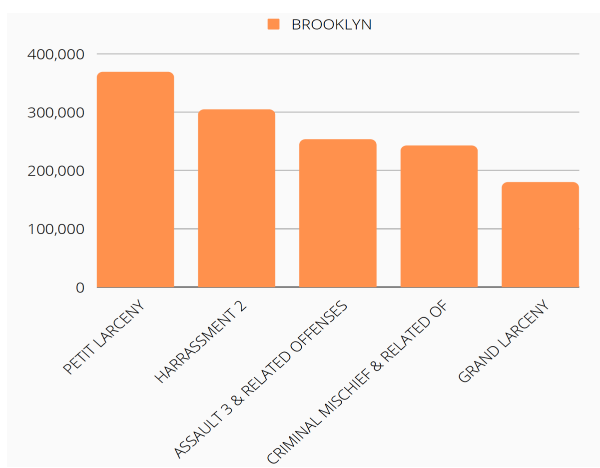

Fig. 9: Top 5 Complaints

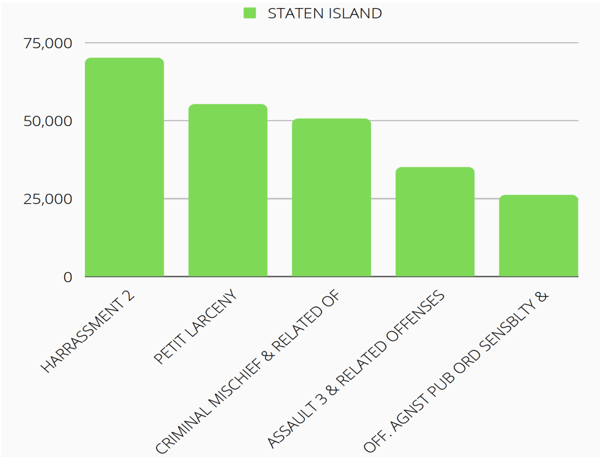

Fig. 10: Top 5 Complaints

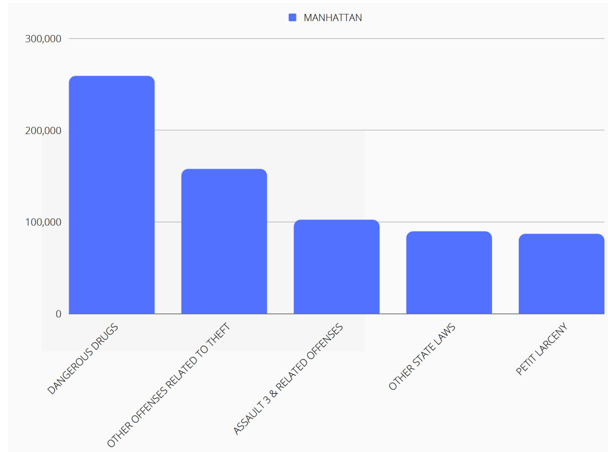

Fig. 11: Top 5 Arrests

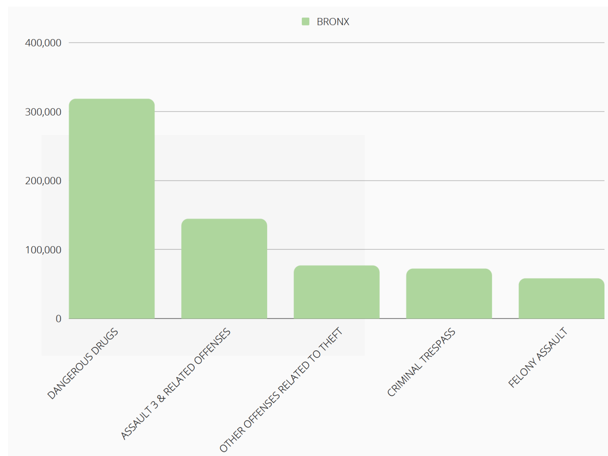

Fig. 12: Top 5 Arrests

felony, and violation, and a trend similar to complaints was recorded. After comparing the complaints and arrests in general categories, analysis was performed at the borough to find the top five crimes in each borough. Fig. 6 shows number of complaints in top five crimes in manhattan. Fig. 7 shows number of complaints in top five crimes in bronx. Fig. 8 shows the number of complaints in top five crimes in queens. Fig. 9 shows number of complaints in the top five crimes in Brooklyn. Fig. 10 shows the number of complaints in the top five crimes in Staten island. In each graph, the x-axis shows the type of crime and the y-axis shows the absolute number of complaints. After careful analysis, it can be observed that petit larceny, harassment, assault & related offenses, and criminal mischief & related of have the maximum number of complaints registered and are part of the top five of all the five boroughs. A similar analysis was performed for number of arrests as well where the y-axis showed the total number of arrests and the x-axis showed the type of crime but no concrete results were found. Fig. 11 shows the arrests in manhattan. Fig. 12 shows the arrests in the Bronx. Fig. 13 shows the arrests in queens. Fig. 14 shows the arrests in Brooklyn. Fig. 15 shows the arrests in Staten island.

In this paper, we also aim to compare the average overtime hours in all five boroughs and analyze if there is any relationship between average overtime hours and the crime rate in New York City. Fig. 16 shows the trend of average overtime hours over the period of the year 2015 to 2020. It is clearly observed that the number of overtime hours significantly increase in the year 2020. This can be attributed to the onset of covid 19 pandemic which required people to stay at home. People were relatively free during these years and could afford more overtime hours. Money could be another motivating factor to work overtime. Fig 14 also shows that the number of average overtime hours is significantly less for manhattan borough across all five years. Fig. 17 shows the average total pay for all the boroughs from the year 2015 to the year 2020. It is observed that the average borough has been increasing in every borough. The rate of increment is maximum for the manhattan borough while in the case of queens the total pay is almost comparable except for the year 2020. A key takeaway is that even though the cost of living is increasing, no significant increase in total pay is observed. Another important observation is that the total pay is slightly higher in 2020 which could be due to more overtime hours shown in the previous graph. Fig. 18 here very similar to the previous figure, as average total pay for all the years combined. An insight that the authors draw from here is that average total pay in Queens is higher compared to the other boroughs which is mainly due to the fact that average total pay has been higher for all the years otherwise as well. Fig. 19 shows the average overtime hours across all the boroughs for all the years.

VII. Coding Challenges

The authors faced a few coding challenges while implementation. A few of these were related to the data sources that were utilized and a few were due to the complex queries and the recommendation engine that was built.

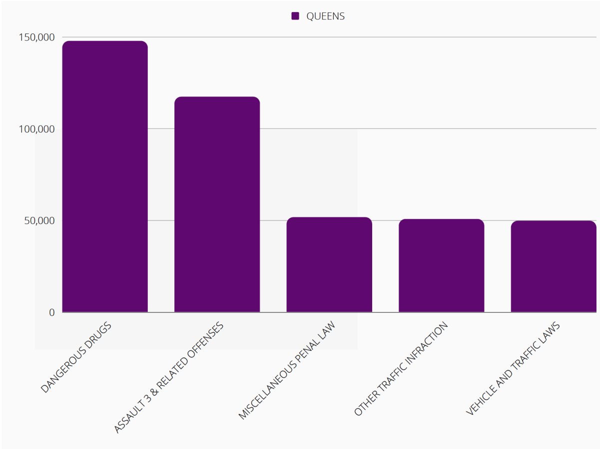

Fig. 13: Top 5 Arrests

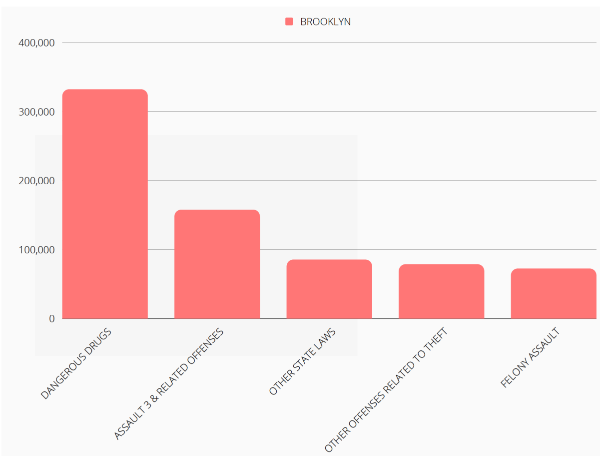

Fig. 14: Top 5 Arrests

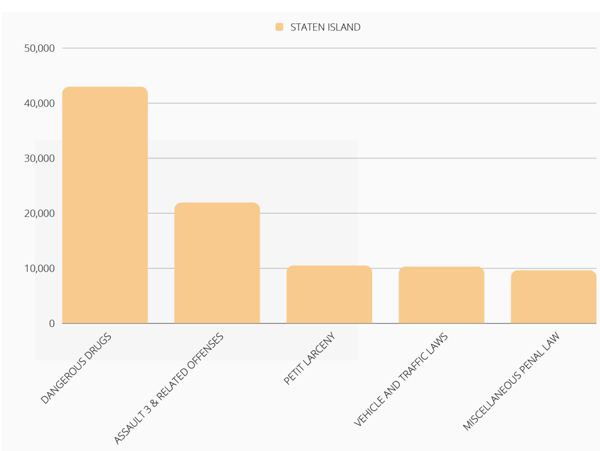

Fig. 15: Top 5 Arrests

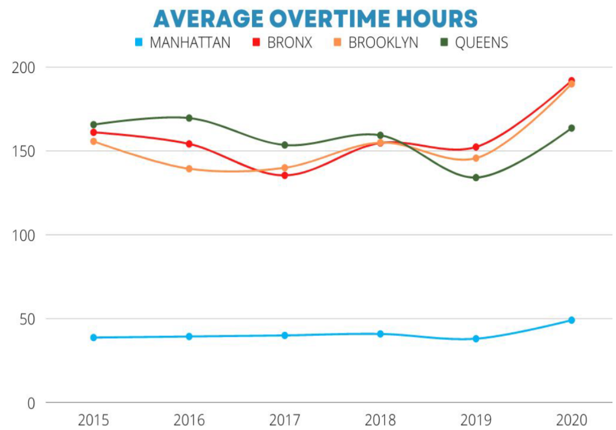

Fig. 16: Average Overtime Hours across all years

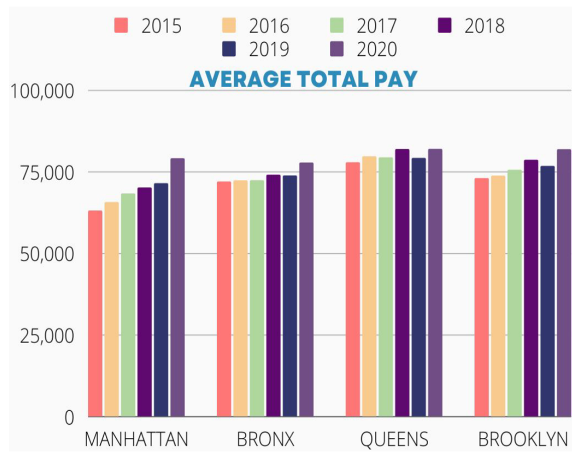

Fig. 17: Average Total Pay across all years

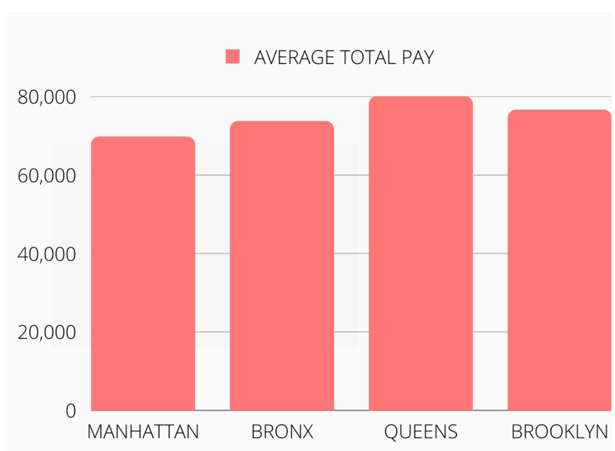

Fig. 18: Average Total Pay

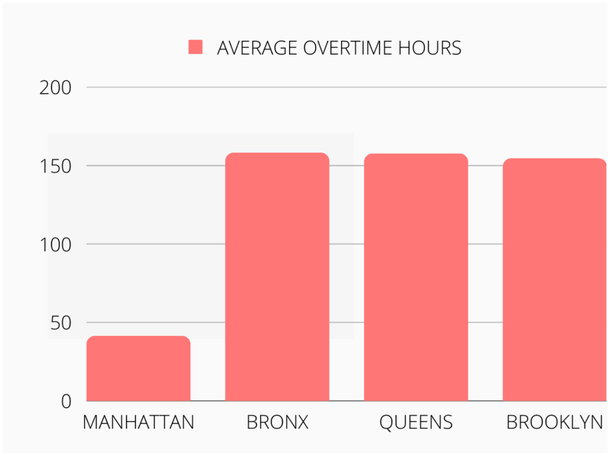

Fig. 19: Average Overtime Hours

A. Recommendation Algorithm

The recommendation algorithm defined in the study is a novel algorithm which takes into account the age, salary, and age and returns a safety index for each of the borough. The algorithm was a very complex query which had multiple case, joins, and unions statements. This made the query quite complex and required a lot of nesting of these queries. While developing this algorithm, as mentioned before, the data for Staten Island was missing from the payroll dataset. Due to this, the query also specifically needed to handle the missing data while building the safety index.

B. Payroll dataset

Staten island data was missing which caused some calculations to be changed accordingly. Another issue that the authors faced was that the data was recorded incorrectly at certain points and in certain agencies. Due to this, there were rows which had to be dropped. Moreover, the pay rate that was defined was not the same for every entry. There were some payroll records with *per Day* pay basis while some records had pay basis of *per Hour*. This resulted in all the salaries not being in the same unit. The authors wrote an algorithm based on the average number of working hours in a week and working weeks in a year to process all the data records in the same unit. This helped in getting a total pay for each record thereby helping the authors analyze the results.

C. Hive Queries

Certain Hive queries were complex and needed further analysis of the data before processing. Queries such as getting the day time and the night time [9] and then using it to calculate the number of complaints. The authors also realized the importance of overtime pay and included those queries as well in their final calculation while comparing overtime and standard hourly rates.

VIII. Obstacles

Following are a few of the obstacles that the authors faced:

1) Data was read in the CSV format due to which there were certain strings present in the data. The strings were under double quotes. This resulted in the MapReduce not being able to distinguish between a comma present in the separator or comma present in a string. Processing this required the authors to define a method to differentiate between the strings and separator.
2) Data collection posed a challenge while collecting the crimes dataset. This was because all the data had to be collected so that the years matched in the other datasets.
3) Lots of missing values in the dataset were present due to which certain records or columns had to be dropped.
4) The payroll dataset only had data present from the years 2015-2020.
5) The pay rates were not standardized across all departments.
6) Selecting features for developing the recommendation algorithm was a challenging task. The authors analyzed the graphs and trends and finally decided on the input parameters to the recommendation algorithm.

IX. Conclusions

In this paper, the authors provide a strong analysis of how the complaints and crime are related. Using an additional dataset of payroll, the authors strengthen their argument by analyzing the purchasing power for defining a safety index. Based on the analysis, Manhattan is deemed as the safest borough in NYC. The intermediate results show various improvements to the analysis and provides various stakeholders a way of improving their services. Analyzing the number of complaints during the day time and the night time helped the authors analyze and recommend how the police force can also be varied.

X. Future Work

Several questions can be answered based on the experiments and results provided by our study. In future work, the authors would like to include more features in the dataset. The current study supports analysis and recommendation using Hive. Using a Machine Learning algorithm could improve the accuracies of this recommendation even more. The authors saw some missing data in the columns due to which certain records had to be dropped. The authors would like to work on this by imputing the data as it could provide more insights.

Acknowledgments

The authors thank Professor Yang Tang for his efforts and for his rapid responses to any and all inquiries about this study. His knowledge was crucial in the development of the authors' approach to criminal analysis and the creation of a suggestion engine. Finally, we express our gratitude to our course classmates for the discussions we had with them.

REFERENCES

[1] Institute, U. (2001, March 8). Risk of violence highest for 18- to 24-year-olds, says Urban Institute study of D.C. crime. urban institute nonprofit social and economic policy research. Retrieved December 19, 2022, from https://webarchive.urban.org/publications/900403.html

[2] Extra Space Storage. (2022, October 31). What's the best borough in NYC for your lifestyle? Extra Space Storage. Retrieved December 19, 2022, from https://www.extraspace.com/blog/moving/city-guides/best-boroughs-nyc/

[3] Greene, J. A. (1999). Zero Tolerance: A Case Study of Police Policies and Practices in New York City. Crime & Delinquency, 45(2), 171–187. https://doi.org/10.1177/0011128799045002001

[4] Freeman, R. B. (1999). Chapter 52 The Economics of Crime. Handbook of Labor Economics, 3529–3571. https://doi.org/10.1016/s1573-4463(99)30043-2

[5] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107–113. https://doi.org/10.1145/1327452.1327492

[6] Publisher data.cityofnewyork.us. (2022, December 16). NYPD arrests data (historic). Catalog. Retrieved December 19, 2022, from https://catalog.data.gov/dataset/nypd-arrests-data-historic

[7] Publisher data.cityofnewyork.us. (2022, December 16). NYPD Complaint Data historic. Catalog. Retrieved December 19, 2022, from https://catalog.data.gov/dataset/nypd-complaint-data-historic

[8] York, C. of N. (2021, January 1). NY citywide payroll data (fiscal year). Kaggle. Retrieved December 19, 2022, from https://www.kaggle.com/datasets/new-york-city/ny-citywide-payroll-data-fiscal-year

[9] Study finds crime in big cities is more likely during the day. (n.d.). Retrieved December 20, 2022, from https://www.usnews.com/news/cities/articles/2019-06-12/study-finds-crime-in-big-cities-is-more-likely-during-the-day