Layla Liu, Lejia Hu
Professor Suzanne McIntosh & Professor Ann Malavet
PBDBA-FA20

Crime Analysis in Boston and New York

Abstract

This paper aims to provide insights into police resource allocation in major cities in the United States, represented by New York City and Boston, to achieve maximum efficiency using big data analytics tools. We studied the top 20 most frequently committed crimes in both cities as well as the months and the exact times in a day the crimes were most likely to occur. Moreover, we investigated the month and the exact times in a day with the highest crime rate.

We used MapReduce to clean the data, dropping the unused columns and records t had missing values. We then applied Impala to extract the most common crimes, the months in which the crimes happened, and the times in a day when crimes happened. We further analyzed the month and time with the most number of committed crimes.

After the analytics, we found that the most common types of crime in New York were petit larceny and harassment with a percentage of 17% and 13% accordingly. The exact time it happened most was 3-8 pm, and crimes were more likely to happen in the second half of the year, especially in November and December, which are months at the end of the year. The most frequently committed crimes in Boston were human injury, investigating persons, and motor vehicle accidents(property damage). The month and the exact time with the highest crime rate were August and 5 PM.

From the research, we learned that New York City and Boston had different crime types, months when crimes happened, and times in a day when crimes happened. However, there are some similarities between the crimes in the two cities. Larceny, motor vehicle accidents, assault, etc. are all commonly committed crimes in both cities.

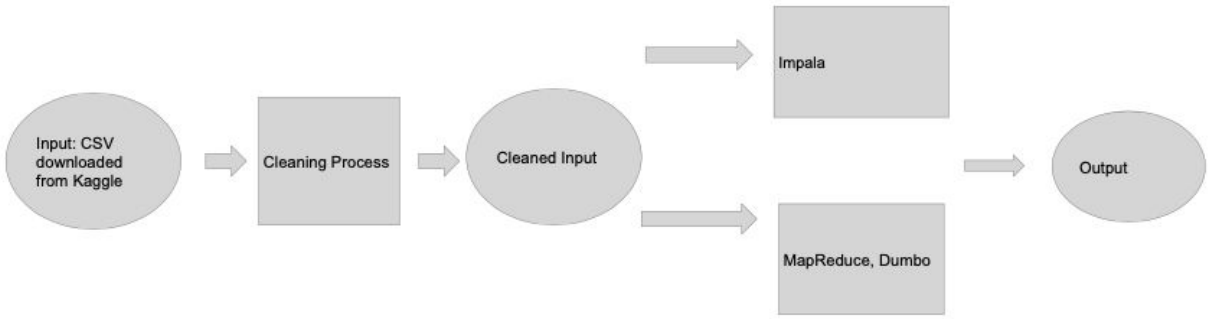
Introduction

By using big data analytic tools like MapReduce, this paper would study the crime rates in large cities.

As the crime rate in big cities kept staying at a higher rate than that of smaller cities and rural areas (Glaeser), it was important to find out an efficient way to deploy the limited human and capital resources to better prevent crimes from happening in big cities.

Although there are related studies about the crimes happening in other cities, which will be discussed in the following sections, we prefer to do research on metropolises, which are the cities we live in, since preventing crimes from happening and bringing peace to the place we live have always been an important topic in our daily lives.

After studying the crimes in NYC and Boston, we will obtain knowledge on when and where crimes are committed in two of the largest cities in the United States. We will analyze them individually and then make comparisons between them. Furthermore, when scholars research other major cities, they can take advantage of this research because NYC and Boston are good protocols when it comes to the analysis of metropolises.



Graph 1 : The analysis procedure

In our analysis, as Graph 1 shows, firstly we downloaded the datasets from Kaggle and uploaded them to HDFS. We cleaned the data with MapReduce and then exported the output files to create Impala tables. From there, we applied both Impala and MapReduce on the cleaned data to generate the ultimate output.

Motivation

Our motivation came from articles in the US news. “Oakland spent 41 percent of the city's general fund on policing in Fiscal Year 2017. Chicago spent nearly 39 percent, Minneapolis almost 36 percent, Houston 35 percent” (Neuhauser). “With these much money spent, the crime rates in big cities were still much higher than small cities or suburb areas” (Glaeser). Therefore, we would like

to examine how to better distribute police resources to achieve the most efficient use. The team agreed on this topic when browsing through Kaggle, which has abundant datasets to choose from to support this research. Using our knowledge as well as the big data tools, we wished to provide guidance to the police department and hopefully lower crime rates in big cities.

Related Work

The previous related works and papers had done research on various determinants like geographical and economical reasons for the crime that happened in Arizona in the 1990s (Cahill), but this research not only looked into the reasons of crimes happening but also dedicated to finding the solutions. Currently, there is no other similar research on how often crimes are happening in major cities, and how can we resolve this problem, but we found out that NYPD indeed had also made great efforts to develop a better system, by “(1) innovative policing, (2) public-private collaboration, (3) a vibrant infrastructure of alternative-to-incarceration programs, and (4) a major philosophical shift in the judicial role and mindset” (Lippman). Our big data analytics would add more innovative and insightful ideas to the police systems by giving the police department a rank of crime types in the cities, the month of when a crime happened, and the time in a day of when a crime happened so that it would guide the police to pay special attention to a specific time or specific types of crimes.

Description of Datasets

The data of NYC comes from Kaggle and it has a size of 253.42 MB. It is indeed a large file and needs to be cleaned. This Kaggle dataset comes from NYC Open Data, and it was uploaded to Kaggle. It records crimes in NYC between 2014 and 2015. The data set includes a variety of parameters that are related to crimes that are reported to the police department, including the time, location, crime type, etc.

Link:
https://www.kaggle.com/ADAMSCHROEDER/CRIMES-NEW-YORK-CITY?SELECT=NYPD_COMPLAINT_DATA_HISTORIC.CSV

Crimes in Boston’s original data was provided by the Boston Police Department(BPD) to document the initial details surrounding an incident to which BPD officers respond. It was then uploaded to Kaggle by Analyze Boston. The dataset contains records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. Records begin from June 14, 2015, and continue to September 3, 2018.

Link: https://www.kaggle.com/AnalyzeBoston/crimes-in-boston

Analytics Stages

Our analytic stages are divided into three parts: the ingestion stage, cleaning and profiling stage, and the analytic stage. In the following sections, we will give a detailed description of what we did in each stage.

Ingestion stage

Our data came from Kaggle, a professional website for scholars to share information and datasets, so we did not struggle much with collecting data from the public. However, since there are numerous datasets on the same topic on Kaggle, we need to do research to decide which datasets to use. After reviewing multiple datasets, we decided that we would study the datasets on crimes in NYC and Boston, which have larger sizes. It contains more information on the topic we'll be studying, which makes the results we derived more trustworthy. We ingested the data by downloading the datasets from Kaggle and then uploading them to HDFS.

Cleaning and profiling stage

Since the size of the datasets selected was large, we gave up a lot of unrelated information. Through comparing the datasets we had we found that the overlapping parts in these two datasets are months of when the crime happened, times in a day it happened and the crime types, so we dropped unnecessary columns and records that contained no useful information(missing or wrong values). Kaggle's data was already in a standard format, so the profiling shows that the number of data after cleaning was very close to the number before cleaning.

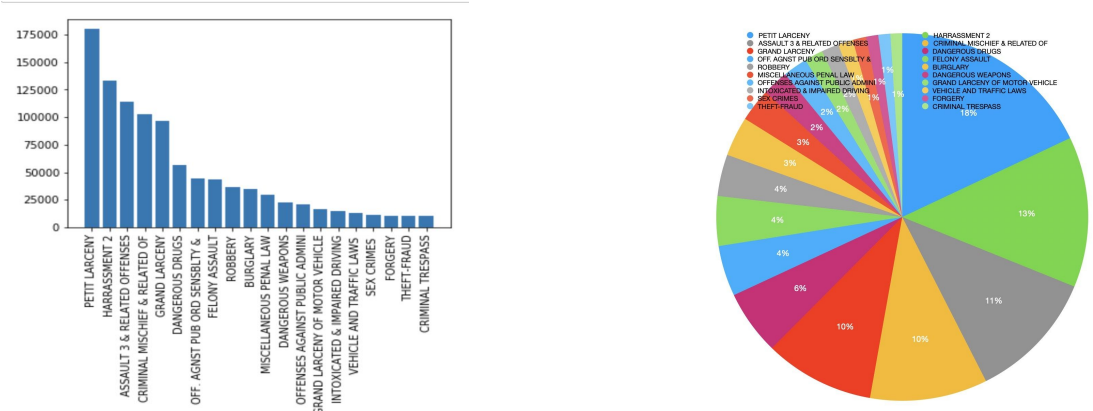
Analytics stage

By performing analysis using these two tools, we easily sorted out the rank for the top 20 types of crimes happening in both Boston and NYC and the rank for the months and times in a day that crimes were more likely to happen. Moreover, we also found the month and time with the most number of crimes happening for each of the top 20 crimes.

For NYC, findings were that the most common crimes were larceny and harassment from the bar chart. As we can see in Graph 4, petit larceny contributed 18% of the sum of the top 20, and harassment contributed 13%, assault, criminal mischief and grand larceny are 11%, 10% 10% accordingly.

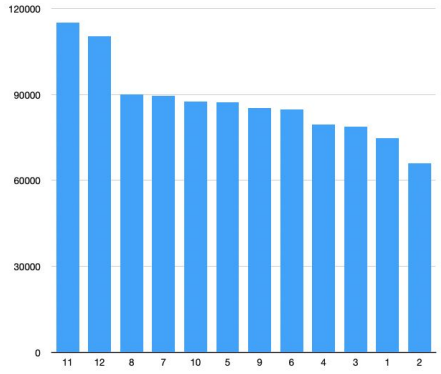
```java
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;
import java.util.Arrays;
import java.util.List;

public class CrimeCountMapper extends
        Mapper<Object, Text, Text, IntWritable> {

    private final static IntWritable one = new IntWritable(1);

    private Text word = new Text();

    public void map(Object key, Text value, Context context)
            throws IOException, InterruptedException {

        String line = value.toString();

        String[] fields = line.split(",");

        if (fields.length == 4) {
            String crime = fields[2].trim();
            word.set(crime);
            context.write(word, one);
        }
    }
}
```

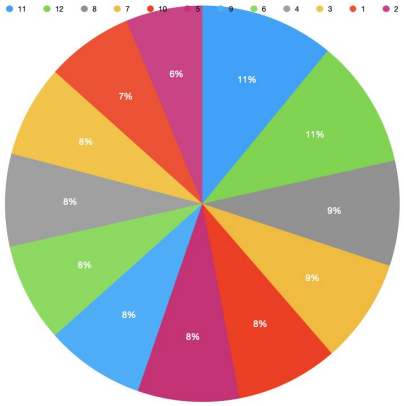
Graph 2: Mapper for the top 20 crime count



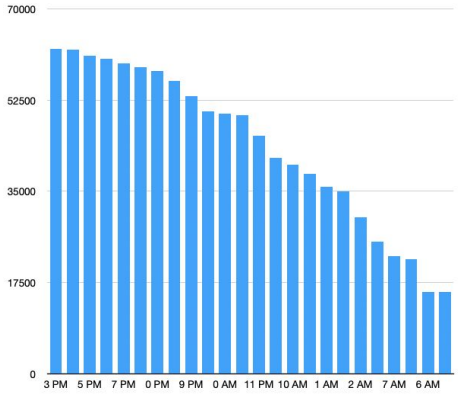
Graph 3 and Graph 4: Bar and Pie Chart for NYC Crime Types Ranking

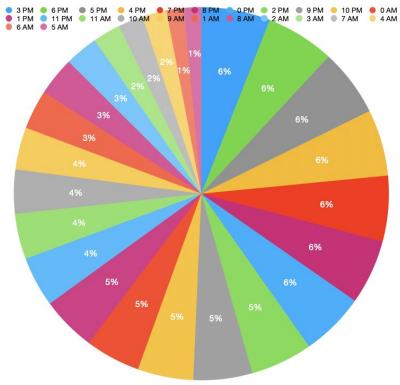
The statistics for all crimes is that they are more likely to happen in the afternoon, especially 3-8 pm as we can see in Graph 7; and a significant increase in the number of crimes happen toward the end of the year especially in November and December. One reason is that New York, unlike other major cities, still has a sufficient number of people during holiday seasons due to the work types of local citizens as well as tourists, so that the crime rate would increase.





Graph 5 and Graph 6: Bar and Pie Chart for NYC Month Ranking





Graph 7 and Graph 8: Bar and Pie Chart NYC for Time in a day Ranking

After listing out the time and month that a specific crime is most likely to happen as it was shown in Graph 9 with a mapper code and reducer code in Graph 10 and 11 respectively, we can see the months are still pretty much all Nov and Dec. One highlight from the month analysis is that we found that theft-fraud in March surprisingly had the highest rate in March. However, through this analysis, the time when a crime happens has a major change. The appearance of 12 AM and 10 PM starts to increase. Other than petit larceny and harassment that would most likely happen in the afternoon, other crimes were more likely to happen at night. And The previous result “highest crime rate at 3 pm” was probably due to the large base number of petit larceny and harassment.

Crime	More likely to happen in	More likely to happen i
PETIT LARCENY 180246	NOV	4PM
HARRASSMENT 2 133179	NOV	3PM
ASSAULT 3 & RELATED OFFENSES 114430	NOV	9PM
CRIMINAL MISCHIEF & RELATED OF 102771	NOV	12AM
GRAND LARCENY 96232	DEC	12PM
DANGEROUS DRUGS 56868	NOV	8PM
OFF. AGNST PUB ORD SENSBLTY & 44772	NOV	12PM
FELONY ASSAULT 43921	NOV	10PM
ROBBERY 36801	DEC	3PM
BURGLARY 34994	DEC	8AM
MISCELLANEOUS PENAL LAW 29221	NOV	6PM
DANGEROUS WEAPONS 22953	NOV	10PM
OFFENSES AGAINST PUBLIC ADMINI 21353	NOV	8PM
GRAND LARCENY OF MOTOR VEHICLE 16223	NOV	10PM
INTOXICATED & IMPAIRED DRIVING 15169	NOV	3AM
VEHICLE AND TRAFFIC LAWS 13050	NOV	6PM
SEX CRIMES 11780	NOV	12AM
FORGERY 10591	NOV	4PM
THEFT-FRAUD 10472	MAR	12PM
CRIMINAL TRESPASS 10292	NOV	8PM

Graph 9: The Time and Month that a Specific Crime is Most Likely to Happen

Graph 10: Mapper for month count and time count based on the top 20 crimes

```java
import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class DualCrime {
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "word count");
        job.setNumReduceTasks(1);
        job.setJarByClass(DualCrime.class);
        job.setMapperClass(DualCrimeMapper.class);
        job.setCombinerClass(IntSumReducer.class);
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
     }
    }
```

```java
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class IntSumReducer extends
        Reducer<Text, IntWritable, Text, IntWritable> {
    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values,
                       Context context) throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        result.set(sum);
        context.write(key, result);
    }
}
```

Graph 11 and 12: Reducer, Runner for month count and time count based on the top 20 crimes

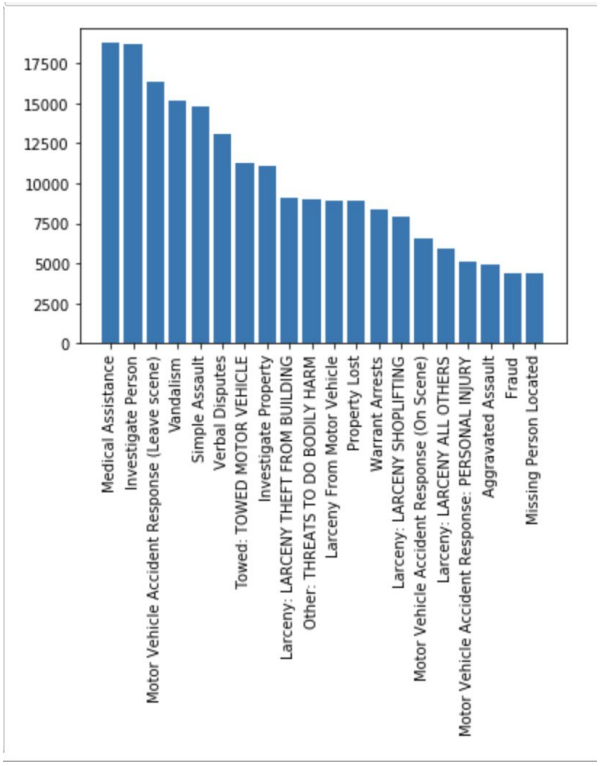
As for Boston, graph 13 illustrates that these are the top 20 most frequently committed crimes in Boston: medical assistance (human injury), investigating person, motor vehicle accidents, vandalism, assault and larceny, etc. Besides, the period with the highest crime rate is 4-6 pm. The month with the highest crime rate is June-August (summer).

```
# Create a table
create external table w1 (OFNS string, TIME int, YEAR int, MONTH int)
row format delimited fields terminated by ','
location'/user/sl6166/impalaInput/';

# Create a view with top 20 most frequent crime types
CREATE VIEW IF NOT EXISTS w2 AS select OFNS, count(OFNS) AS OFNSnum from w1 group by OFNS order by OFNSnum desc limit 20;
select * from w2;

# Display the number of times each crime occur in different months in descending order. Repeat 20 times with different crime types.
select month, count(month) AS MONTHnum from w1 where OFNS = 'Medical Assistance: SICK/INJURED/MEDICAL - PERSON' group by month order by MONTHnum desc limit 12;
select month, count(month) AS MONTHnum from w1 where OFNS = 'Investigate Person: INVESTIGATE PERSON' group by month order by MONTHnum desc limit 12;

# Display the number of times each crime occur at different times in descending order. Repeat 20 times with different crime types.
select time, count(time) AS TIMEnum from w1 where OFNS = 'Medical Assistance: SICK/INJURED/MEDICAL - PERSON' group by time order by TIMEnum desc limit 25;

# Create a view of months in which crimes happen most frequently.
CREATE VIEW IF NOT EXISTS w3 AS select month, count(month) AS MONTHnum from w1 group by month order by MONTHnum desc limit 12;
select * from w3;

# Create a view of times at which crimes happen most frequently.
CREATE VIEW IF NOT EXISTS w4 AS select time, count(time) AS TIMEnum from w1 group by time order by TIMEnum desc limit 25;
select * from w4;
```

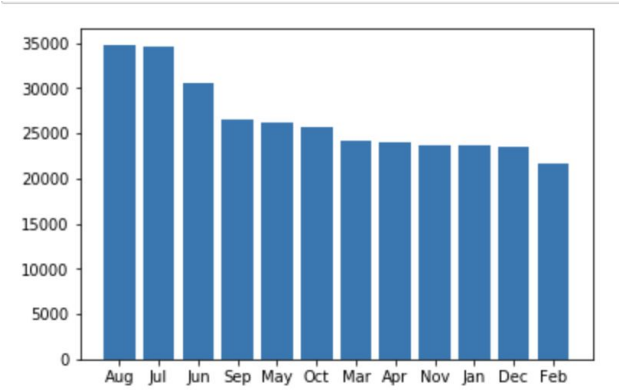
Graph 13: Impala Commands to Generate Output

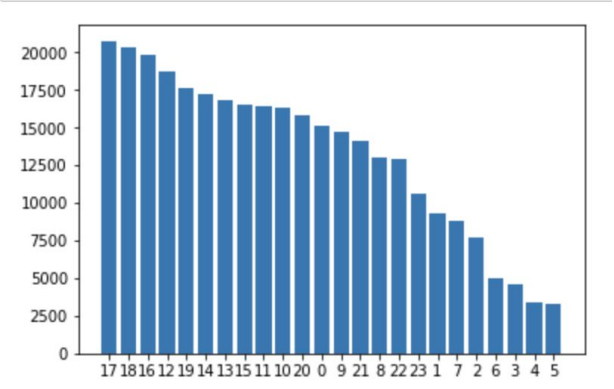
Common crimes in Boston, such as human injury, investigating person, motor vehicle accidents, vandalism, assault and larceny, happen most frequently during Summer. Our interpretation was that the activity overall is lower during other seasons in Boston because of the climate and most people spend more time indoors when it's cold. Therefore, crimes happened most frequently during the summer when the weather was suitable for human activity. Crimes were likely to occur

at dusk in Boston because of the crime types. Crimes such as human injury, motor vehicle accidents and so on happen frequently during peak hours.



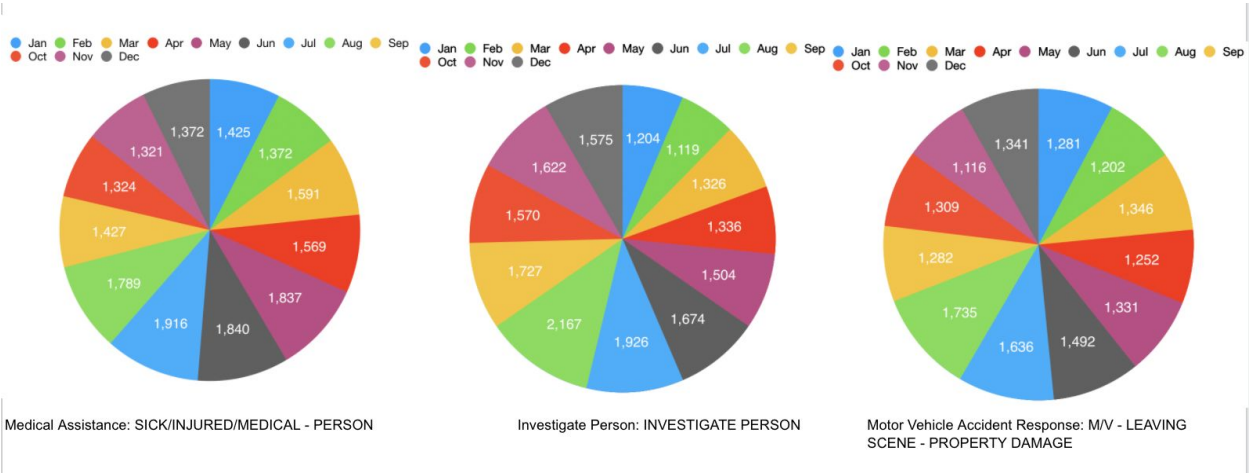
Graph 14: Top 20 Most Frequently Committed Crimes in Boston





Graph 15 and 16: Bar Charts for Boston for Month Ranking and Time in a Day Ranking

Graph 17 is the pie charts for the number of cases reported in each month. Only the top 3 crimes are shown here, which is medical assistance, I think it just means human injury. Investigating person and motor vehicle accident response. You can see that still most of them happen during the summer and early fall.



Graph 17: Pie Charts for Boston Month Distribution of Top 3 Crimes

time	freq
12	1087
16	1078
13	1068
14	1057
17	1052
18	1036
19	1030
10	1012
11	998
15	991
21	985
20	981
22	869
9	795
0	772
8	672
23	636
1	635
2	495
7	397
3	313
4	278
6	274
5	272

Medical Assistance: SICK/INJURED/MEDICAL - PERSON

time	timenum
16	1257
18	1185
17	1178
12	1170
19	1145
14	1119
13	1100
15	1086
11	1022
20	1020
10	954
21	894
9	818
0	766
22	763
23	597
8	592
1	447
7	402
2	391
3	262
6	213
4	200
5	169

Investigate Person: INVESTIGATE PERSON

time	timenum
17	1180
18	1082
16	1080
19	959
20	882
15	830
21	827
14	801
12	796
13	773
22	771
10	749
9	726
11	713
8	700
23	666
0	552
7	530
2	415
1	392
6	323
3	255
5	179
4	142

Motor Vehicle Accident Response: M/V - LEAVING SCENE - PROPERTY DAMAGE

Graph 18: Boston Times in a Day Distribution of Top 3 Crimes

Graph 18 contains the impala results of the number of cases reported in each time period. Again, Only the top 3 crimes are shown here. We can see that still most of them happen during the afternoon. One thing we found interesting was that we thought incidents like larceny or vandalism would most likely to happen at night. But from the results. They still happen most frequently in the afternoons. This might be due to the fact that the Boston dataset records on these types of crimes are somewhat biased.

Conclusion

The results of our research were that: New York City and Boston had different results in terms of crime types, months of when crimes happen, and time in a day of when crimes happen. These

differences proved that the police force should be arranged differently due to various characteristics of the cities.

In New York, crimes happened the most in November and December, whereas in Boston they were July and August. In New York, the most common crimes were petit larceny and harassment, and in Boston, they were medical assistance and investigating people. In New York, 3 pm was the time with the highest crime rate, and the corresponding time in Boston is 5 pm.

Hence, we can draw the conclusion that this project had been designed to provide insight on how to better allocate police resources for the most efficient use. A significant amount of the city's general fund had been spent on policing. However, the crime rates in big cities were still considerably higher than those of smaller cities. As contributors, it is our ardent hope that the conclusions from this project might act as a guide on police resources allocation optimization.

For further research, we would like to collect data sets that include the zip code, so that we can give a more detailed suggestion to the police department on where they should deploy the resources.

Works Cited

Glaeser, Edward L. “Why Is There More Crime in Cities?” Journal of Political Economy, vol. 107, no. S6, 1 Dec. 1999. JSTOR, www.jstor.org/stable/10.1086/250109?refreqid=search-gateway:8329ef4d6d0b113e08301832ba92ba75.

Lippman, Hon. Jonathan. “HOW ONE STATE REDUCED BOTH CRIME AND INCARCERATION.” Hofstra.edu, law.hofstra.edu/pdf/academics/journals/lawreview/lrv_issues_v38n04_bb1_lippman_final.pdf.

Meagan E. Cahill & Gordon F. Mulligan (2003) The Determinants of Crime in Tucson, Arizona1, Urban Geography,24:7, 582-610, DOI: 10.2747/0272-3638.24.7.582

Neuhauser, Alan. “Cities Spend More and More on Police. Is It Working?” U.S. News & World Report, U.S. News & World Report, 2017, www.usnews.com/news/national-news/articles/2017-07-07/cities-spend-more-and-more-on-police-is-it-working.