Undergraduate Group 4 Report

Justin Neal
Computer Science
Georgia State University
Atlanta, United States
jneal27@student.gsu.edu

Zachary Benator
Computer Science
Georgia State University
Atlanta, United States
zbenator1@student.gsu.edu

Brian Cabigon
Computer Science
Georgia State University
Atlanta, United States
briancabigon@gmail.com

Lee Klarich
Computer Science
Georgia State University
Atlanta, United States
lklarich1@student.gsu.edu

Timothy Lewis
Computer Science
Georgia State University
Atlanta, United States
tlewis58@student.gsu.edu

***Abstract*—We used Python with supported libraries (Matplotlib, NumPy, Pandas) and R to answer various meaningful questions about NYPD complaint data.**

I. Introduction

With this project we aimed to provide insight about crimes committed in New York City by answering meaningful questions using Python and R. Answers to these questions will be supported by visualizations to provide ease of understanding about the dataset.

II. The Dataset

The dataset [1] that we used contains reported crimes in New York between the years of 2006 and 2017. It consists of a single table of 6.5 million complaints. For each complaint the table includes data such as date, time, description, location, whether or not the crime was completed or interrupted, level of offense, and more.

III. Preprocessing

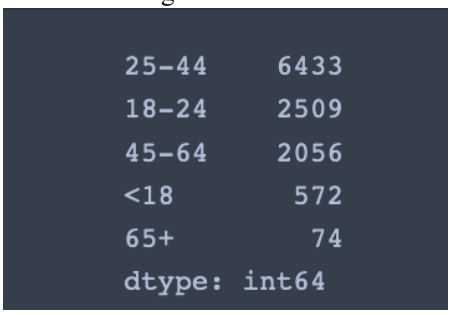
A. Preprocessing was accomplished entirely using Python with pandas.
   1) There were some mistyped dates that were either before 2006 or after 2017. To fix these, we replaced the first digit of the year with a two for all values, then we replaced the second digit with a zero for all values. Finally, we replaced the third digit with a zero if it was not already the number one.
   2) "From" dates and "from" times were combined into a single column in datetime format.
   3) "To" dates and "to" times were combined into a single column in datetime format.
   4) All columns that we decided wouldn't be necessary for our project were removed from the data set.
   5) The size of the original dataset was 2.06 GB uncompressed. The dataset was ultimately reduced to 904MB uncompressed.

B. Further preprocessing was performed on a per-question basis.

IV. Questions We Attempted to Answer

A. What is the average age of suspects that committed burglary?

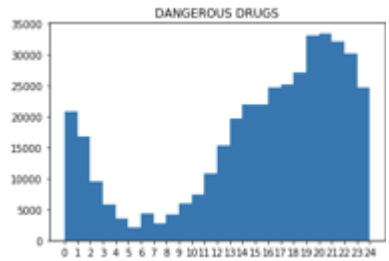
1) Method: Question A can be answered easily with an average reduction over every entry with a known perpetrator age in the dataset with pandas.
2) Background information regarding the data: After analysis of the data, it was concluded that this question could not be properly answered because the suspect ages were in a range, and this range was a string format as well. The best suited question for our data is: *What age group commits the most burglaries?*
3) Preprocessing
   a) Unique values of the suspect age group were displayed so the bad values could be filtered out.
   b) Bad age group values were converted to null and dropped.
   c) All columns were dropped excluding suspect age group and offense description.
   d) This data was then filtered to only contain burglaries.
4) Solution
   a) Value counts were performed on the suspect age group column and sorted in descending order.
   b) It was concluded that the age group that committed the most burglaries is 25-44 as shown in the figure below.



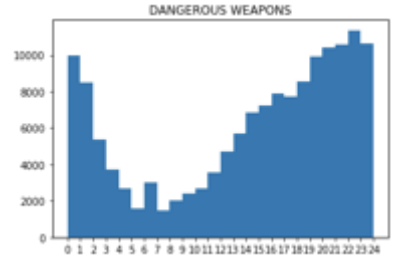
B. *Is there any correlation between the time of the offense and the type of offense that was committed?*

1) Method: For question B we used matplotlib to create a series of histograms to compare the time and type of offense.
2) Background information regarding the data: Date/time of offense is logged in a FROM and TO column. If the time of offense is precisely known, it is logged in the FROM column, and the TO column is empty. Otherwise, the event happened at some point between the time in FROM and the time in TO.
3) Preprocessing
    a) Rows were dropped if FROM was null.
    b) Rows were dropped if the time between FROM and TO was greater than three hours (using functions from Python's datetime and timedelta libraries).
    c) Similar Offense Descriptions were combined (i.e. ADMINISTRATIVE CODE and ADMINISTRATIVE CODES, INTOXICATED/IMPAIRED DRIVING and INTOXICATED & IMPAIRED DRIVING).
4) Solution: We plotted a separate histogram for each offense description with time of day on the x-axis and the number of recorded offenses on the y-axis.
5) Results
    a) All 60 histograms can be seen in the notebook. This report contains only 6.
    b) Most offenses seemed to follow a vaguely sinusoidal sort of shape, with peaks and troughs at different times of day for different offenses.
        i) Peak around 6pm, minimum around 6am

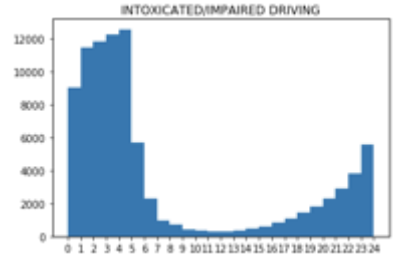
        ii) Peak around 8pm, minimum around 5am



        iii) Peak around 10pm, minimum around 7am



    c) Some crimes did not follow the common sinusoidal pattern.
        i) Very sharp decline from 5am to 7am following a fairly steady increase (bars close at 4am in New York)



        ii) Very irregular with a huge spike at noon

        iii) irregular shape

C. *Is there any noted correlation to the type of offense and the location of the offense?*

i.e. Are assaults more prevalent at home or in a specific borough?

1) Method: For question C we will first group the data by the type of offense and then we will use ggplot2 to plot the data on a map using the given coordinates.
2) Background information regarding the data:
A shapefile [2] was used to plot the boroughs.
3) Preprocessing
    a) Null values for offense description, latitude, and longitude were dropped.
    b) Shapefile was converted into a dataframe and exported as a csv for further processing with python.

c) Coordinates in the shapefile were using spatial reference EPSG: 2263 to plot the boroughs, so the dataset's coordinates were also converted to spatial reference coordinates using the pyproj library.

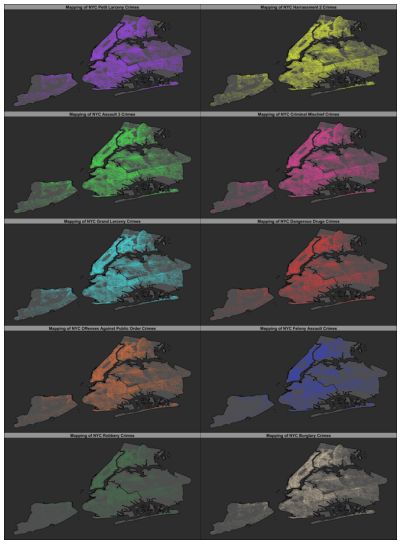
d) To mitigate outliers, all coordinates greater than or less than the max and min coordinates of the shapefile were dropped.

4) Solution

a) The coordinates of the top 10 offenses were plotted using ggplot2 with transparency to show the concentration of crimes.

5) Results

i)



D. Were there any rises of specific crimes on certain holidays?

i.e. Were robberies more common on Christmas compared to other crimes?

1) Method: For question D we will create a line graph using matplotlib and compare the relative occurrence rates of each crime on each holiday.
2) Background information regarding the data: The data was across the span of the years 2006 to 2017 and the question called for an aggregation of all of the crimes on given days across all the years.
3) Preprocessing

e) The data was normalized to a single year in order to facilitate aggregation

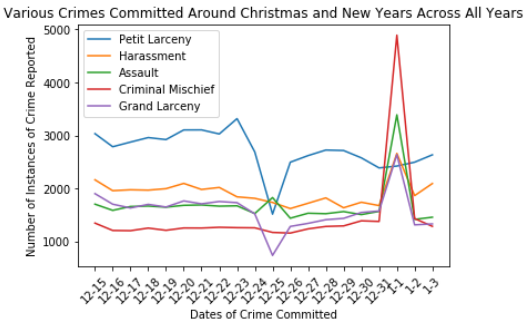
f) After aggregation, the dates were filtered by the top 5 most reported crimes and then separated.

4) Solution

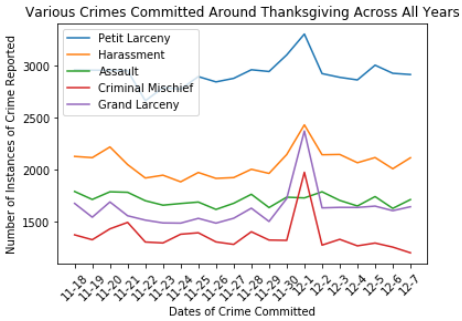
b) We created line graphs for 4 different holidays and graphed the crime report rates for 5 different crimes from 10 days before to 10 days after the holiday to see if there is any significant change in crime reports on the holidays' dates.

5) Results

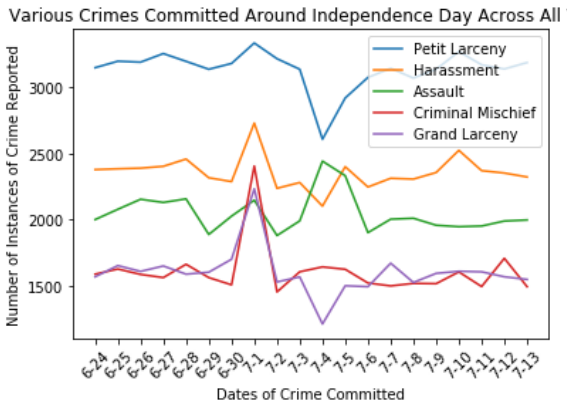
a) On Christmas, reports of both types of larceny decreased while assault increased and harassment and criminal mischief remained relatively unphased. On New Years, with New York being an extremely populated on New Years night, all of the top 5 crimes except petit larceny spiked on New Years day.



b) On Thanksgiving, reports of all types of crimes slightly increased on Thanksgiving. On December 1st, all crimes except assault drastically increased in number of reports. .



c) On Independence Day, petit larceny, harassment, and grand larceny all decreased, assault increased, and criminal mischief remained unchanged. The first of the month also has a spike in the number of reports on all crimes.



ii

E. Are there any specific date ranges or time ranges where crime is more prevalent in one borough than the other?

i.e. Does Manhattan see more crimes than the other 4 boroughs?

1) Method: For question E we will create a histogram using matplotlib to group dates and times by borough.

F. What crimes are easiest to get away with?

1) Method: For question F we will group the crimes by type and determine which type of crime tends to have the least suspect information. We will also consider which crimes were most commonly reported to have been completed.
2) Background information regarding the data: The Suspect Age Group column ('SUSP_AGE_GROUP') contains all of the age group ranges for the crimes. All of the crimes except the ones involving traffic violations contain ranges <18, 18-24, 25-44, 45-64, and 65+. The traffic violations column contains a wide range of entries in this column that are not necessarily even range or valid ages, ranging from -980 to 2016.

3) Preprocessing: We selected the column 'SUSP_AGE_GROUP' and removed rows that didn't have a valid age ranged as mentioned prior. Then the rows that were not of a crime that had more than 10,000 reports were removed.
For determining the likelihood for a crime to be completed, the column 'CRM_ATPT_CPTD_CD' was used and the crimes that did not have more than 10k reports were filtered out.
Crimes that have >10k reports:

OFNS_DESC	
PETIT LARCENY	80667
HARRASSMENT 2	69554
GRAND LARCENY	49242
CRIMINAL MISCHIEF & RELATED OF	42718
ASSAULT 3 & RELATED OFFENSES	38709
ROBBERY	21152
OFF. AGNST PUB ORD SENSBLTY &	19925
FELONY ASSAULT	16528
BURGLARY	15525

4) Solution: The rows were separated by those that did not suspect information at all, meaning the 'SUSP_AGE_GROUP' field was non-null, and those that did. Then those rows were grouped by the crime committed. Then the crimes were ranked by the percentage of reports that did not have suspect information descending and it was found that OFF. AGAINST PUB ORD SENSIBILITY was found to have the highest percentage of reports without suspect information.

OFNS_DESC	
OFF. AGNST PUB ORD SENSBLTY &	87.808087
ASSAULT 3 & RELATED OFFENSES	86.712778
HARRASSMENT 2	84.534586
FELONY ASSAULT	83.019120
ROBBERY	65.151490
CRIMINAL MISCHIEF & RELATED OF	63.790020
PETIT LARCENY	59.313955
GRAND LARCENY	44.108600
BURGLARY	42.857669

Evaluating how easy a crime was to get away with was also analyzed through the percentage of crimes that were deemed completed and those that were attempted, meaning that they were interrupted and thus not allowed to complete.
For this analysis, the crimes were separated by the crimes that had been completed versus attempted and then each grouped by the type of crime committed. Then the percentage of the crime reports that were actually completed were calculated and then sorted ascending to show the crimes that were most likely to be left incomplete.

OFNS_DESC	
ROBBERY	88.885926
BURGLARY	92.763371
FELONY ASSAULT	95.289164
GRAND LARCENY	98.148565
PETIT LARCENY	98.760325
ASSAULT 3 & RELATED OFFENSES	99.362424
CRIMINAL MISCHIEF & RELATED OF	99.625515
OFF. AGNST PUB ORD SENSBLTY &	99.693594
HARRASSMENT 2	99.730861

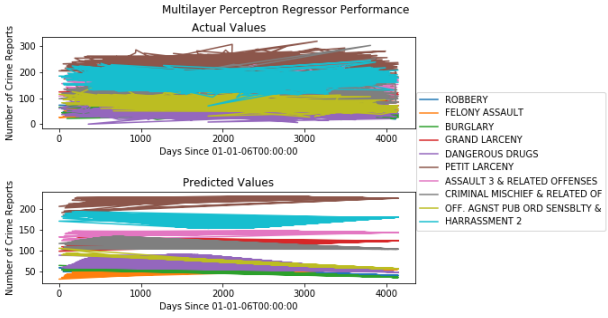
It can be seen that robberies followed by burglaries are the most likely crimes to be interrupted or left incomplete, which makes sense because the amount of time it takes to commit the act of robbery/burglary is significantly longer than it could be for assault or larceny.

G. What crime will be most likely to be committed on a given day in the future?

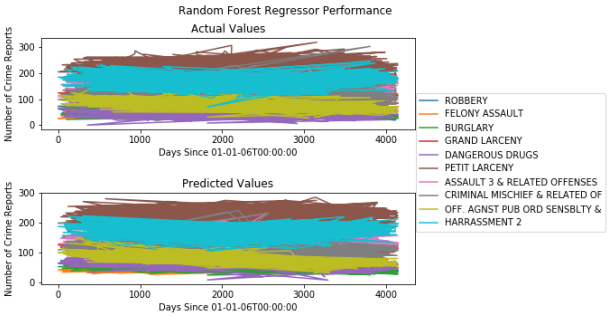
1) Method: For question G we will compare models using: deep neural network regression, random forest regression, and support vector regression. These regressors will be trained on the association of date and time of day with the type of crime from a randomly sampled training subset from the dataset, and the data entries that were not sampled for the training subset will be used as a validation set.
2) Background information regarding the data: The dataset contains columns for the date and time the

complaint was filed as well as the date and time the crime from the complaint was completed. Also, the crimes in the complaints are already organized into specific codes for each different type of crime. The time span for this ranges from 2006 to 2017.

3) Preprocessing
   a) The start and end dates and times for the complaints were combined into a new column with the combined date and time for the actual occurrence of the crime.
   b) All of the columns in the dataset excluding the complaint datetime and the crime code were dropped to shrink the size of the dataset used for this portion.
   c) Every row was one-hot encoded for the top 10 most prevalent crimes.
   d) The data was grouped by the day of occurrence, and the value counts for each of the top 10 most prevalent crimes is aggregated for each group.
   e) After the aggregation, there were only rows with the count of each crime occurrence and the day on which it occurred.
4) Solution
   a) A multilayer perceptron regressor and a random forest regressor were used to try to forecast the number of occurrences of each of the top 10 most prevalent crimes for a given day in the future.
   b) Each model was trained with the complaint day as the predictor and the counts of the crimes as the targets. This means these models are one-to-many regressors.
   c) The multilayer perceptron model was trained on a variety of different layer counts and sizes as well as different batch sizes. For the activation function, ReLU was used across the board because of its speed of execution as well as its ability to resist gradient vanishing.
   d) The random forest regressor model was trained on the same data in the same way as the previous model.
   e) The best models of each type were used to create the graphs.
5) Results
   a) Multilayer Perceptron Regressor



   b) Random Forest Regressor



V. REFERENCES

[1] A. Sellanes, *NYD Complaint Data Historic.* May 19, 2019., City of New York. Version 1. from url: https://www.kaggle.com/agustinsellanes/nypd-complaint-data-historic

[2] https://www1.nyc.gov/assets/planning/download/zip/data-maps/open-data/nybb_19d.zip