STAT- S 670 EXPLORATORY DATA ANALYSIS- FINAL PROJECT

ANALYSIS OF CRIMES IN CHICAGO

By Viswa Suhaas Penugonda (vpenugon@iu.edu), Harsha Sai Gade (harsgade@iu.edu) & Manasa Gudise (mgudise@iu.edu)

STATEMENT OF GOALS

Abstract:

In this project, we aim to perform Exploratory Data Analysis to the crimes data of Chicago to gain a better understanding of the patterns and trends of criminal activity in the city.

There are several reasons why we should care about the crimes taking place in Chicago. First, crime has a significant impact on the safety and well-being of residents, particularly those who live in high-crime neighborhoods. Second, crime can have a negative impact on the city's economy, as businesses may be hesitant to invest in areas with high crime rates. Finally, understanding crime patterns can help law enforcement agencies develop more effective strategies for preventing and reducing crime.

By exploring previous crime data, we can gain insights into the types of crimes that are most prevalent in certain areas, the times, and days when crimes are most likely to occur, and the demographics of both victims and perpetrators. This information can be used to develop predictive models that can help law enforcement agencies allocate resources more effectively and prevent future crimes.

Research Questions:

1. How accurately can we predict the likelihood of an arrest being made for a particular crime in Chicago based on various variables, such as the type of crime, location, month, and day?
2. Is there a significant interaction between the type of crime and its location in determining the likelihood of a case being solved in Chicago?
3. How does the overall trend of crime in Chicago vary across different months in the year 2022? Are there any seasonal patterns in the data?
4. Are there any specific types of crime that are more likely to be solved in Chicago?
5. How does the location of a crime impact the likelihood of an arrest being made in Chicago?

DATA DESCRIPTION

The project aims to analyze the crimes data of Chicago obtained from the CLEAR system of the Chicago Police Department. The dataset contains 7,786,062 records of reported crimes and has 22 columns.

To make the analysis more manageable, the study limited the analysis to the year 2022, and the data was filtered accordingly. The resulting dataset was then used to identify the crime trend in each month of the year, with a focus on understanding the variation in criminal activity over time.

The columns of interest in the dataset include date, primary type, description, location description, and arrest. The ‘date’ column is a time-series variable that records the date when the incident occurred. The ‘primary type’ column is a categorical variable that indicates the type of crime committed, while the ‘description’ column provides a more specific description of the crime. The ‘location description’ column is a categorical variable that records the location where the crime took place. The ‘arrest’ column is a binary variable that indicates whether an arrest was made or not.

Other columns such as ID, case number, block, IUCR, beat, district, ward, community area, FBI code, and coordinates were discarded as they do not provide any meaningful insights for the research questions of the project.

By focusing on the year 2022 and specific columns of interest, the study aims to provide valuable insights into the crime trend in the city and identify factors that may be influencing criminal activity in the area.

“UNDERSTANDING THE DATA: EXPLORING THE CHICAGO CRIMES DATASET”

To better understand the data and the situation of crime in Chicago, we need to understand distributions and characteristics of various classes and the individual class variables. We tried focusing on the crimes that are more prominent or the crimes whose distribution seemed abnormally different from other crimes.

We first tried to understand how the crime is distributed among various types of crimes. The types of crimes that are included in this dataset are, Weapons violation, theft, stalking, sex offense, robbery, public peace violation, public indecency, prostitution, narcotic violation, offense involving children, obscenity, Non-criminal, motor vehicle theft, liquor law violation, kidnapping, intimidation, interference with public officer, human trafficking, homicide, gambling, deceptive practice, criminal trespassing, criminal sexual assault, criminal damage, concealed carry license violation, burglary, battery, assault and arson.

We plotted count of each type of crime to understand the frequency of each type of crime and compare it to the other type of crimes.

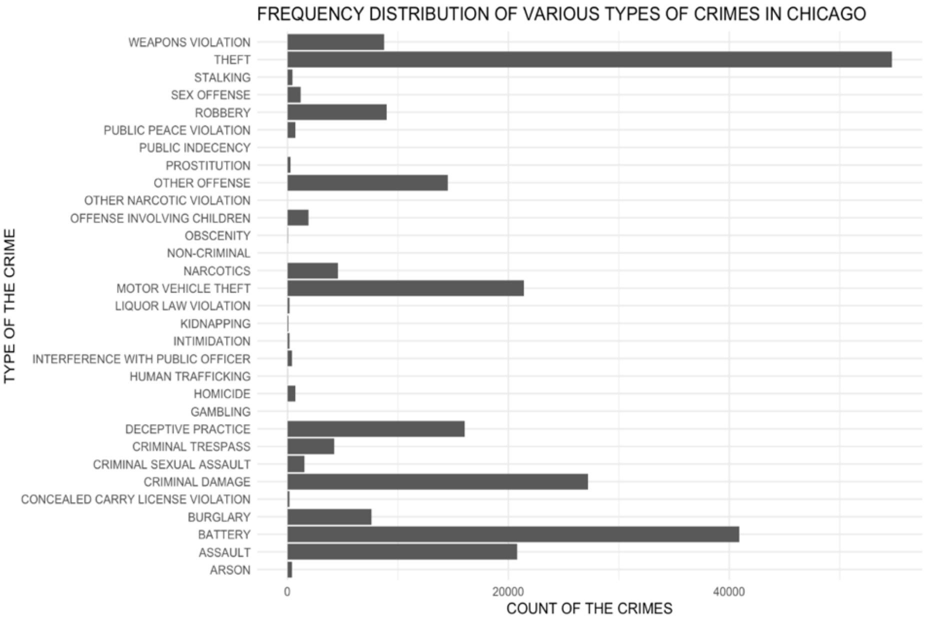


Figure.1

From this graph, we can understand that, in the year 2022, theft is the most frequent occurring crime in Chicago, followed by Battery, Criminal damage, Assault, Motor vehicle theft and deceptive practice. The Battery, Criminal damage and assault are like each other in terms of type of the crimes. Also, theft and motor vehicle are under the same category too.

We then tried understanding the crime based on the location of occurrence of crime.

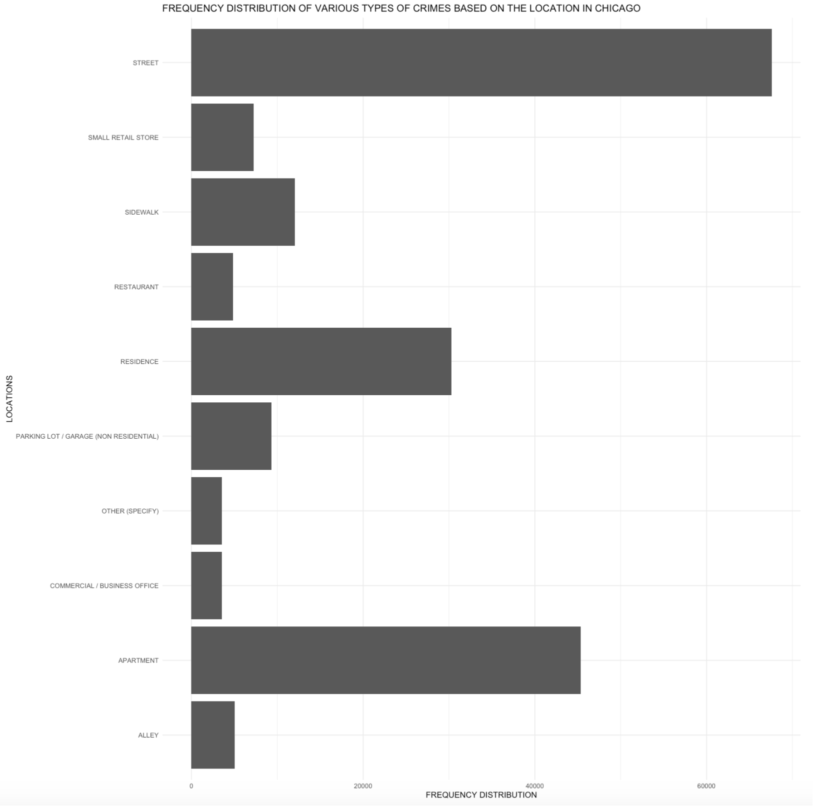


Figure. 2

From this graph, it is understood that most of the crime happens on the street, followed by apartments, residence, sidewalk and then alley.

So, we can understand that in private places, apartments and residence places are more prone to crime and in public places, places like street, alley and sidewalk are more exposed to crime. So, in the above-mentioned places, people in general need to be more careful than they usually are.

To better understand the scenario, we plotted frequency of crime against type of crime faceted by the location type of the crime.

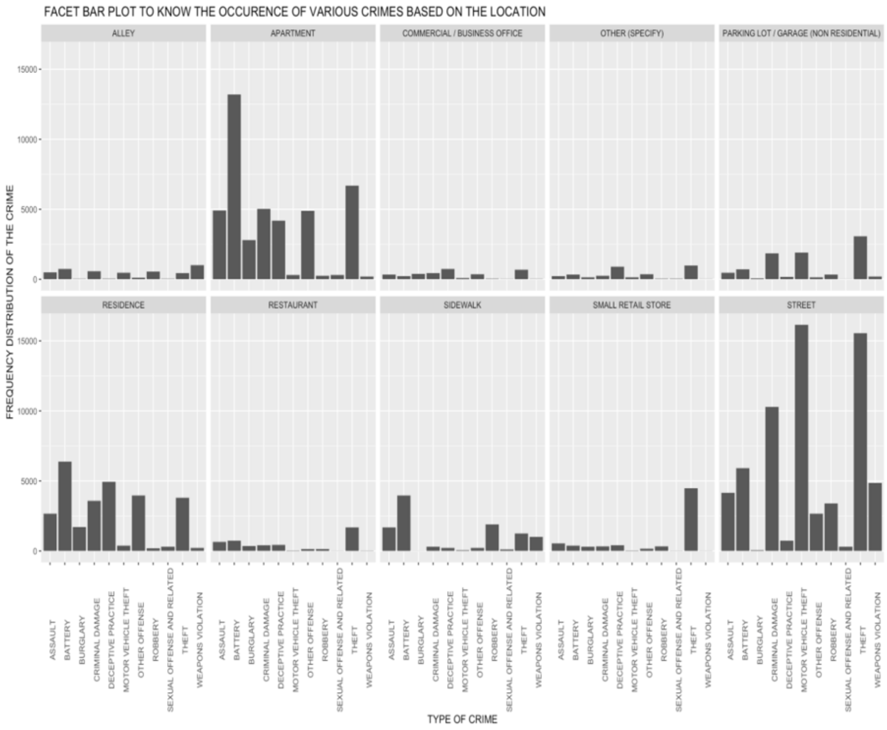


Figure. 3

This graph has shown few interesting distributions and relations between the variables.

We understood that the distribution of the apartments and residence are like each other, in both scenarios, battery is the highest occurring crime followed by theft and criminal damage. But even when comparing the individual type of crime in these areas against the crime in the street, it is understood that the crime is more frequent in the streets despite the type of crime. This could be due to the security in apartments being more secure but when the apartments or residence are themselves considered as entities, there could be crime happening due to fights breaking out between 2 neighbors or the heists which are planned and executed.

We then tried to understand how crime has varied over the year. So, we plotted count of crime against the month.

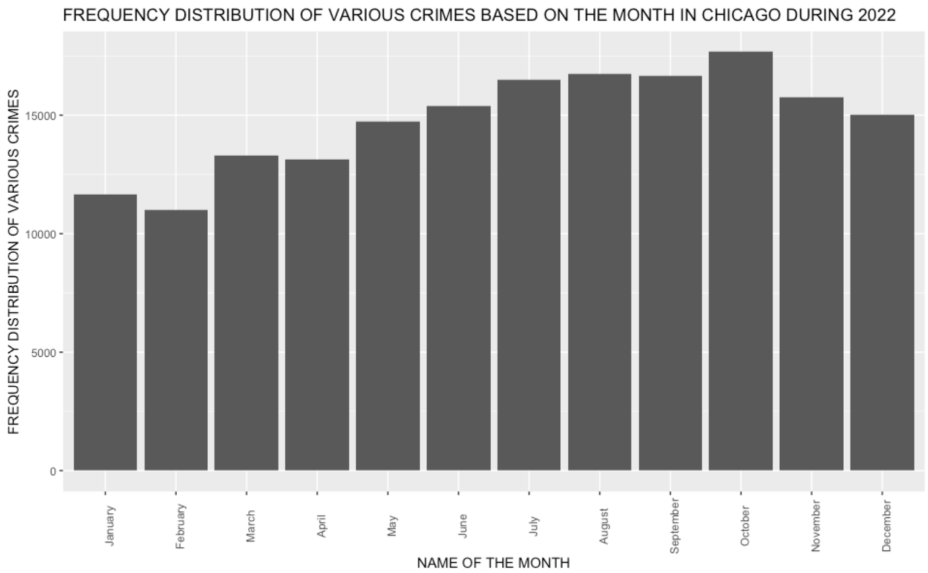


Figure. 4

From this graph, we could infer that the crime overall increased over the year 2022 from January till October and then taking a small dip towards the end of the year. We tried understanding this situation better by faceting the above plot by the type of location of the crime.

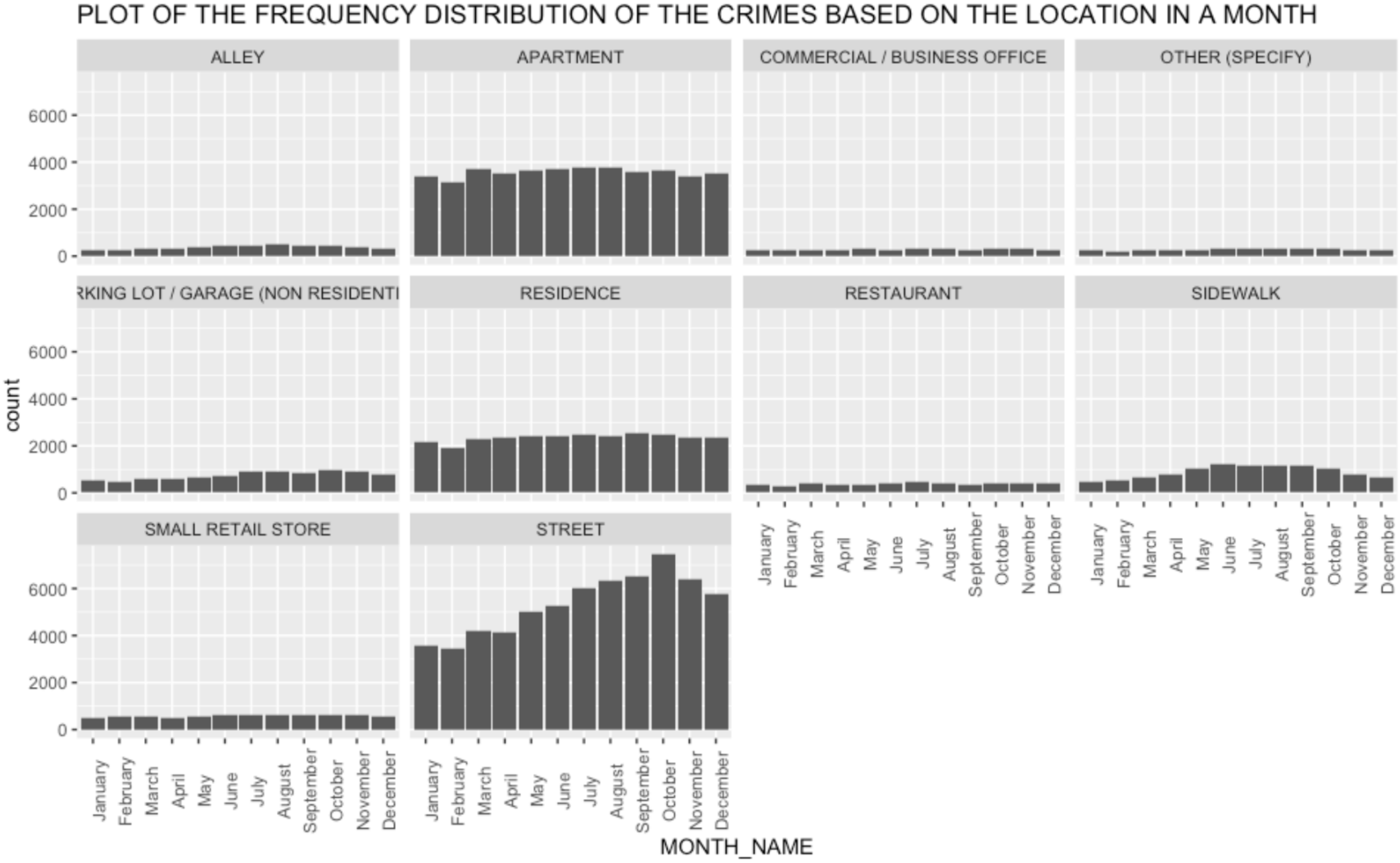


Figure. 5

Here we have observed that the crime in private areas like apartment, residence and restaurants has remained constant over the year with minor dips and increases over the year. But the crime in public areas like street, alley, sidewalk, and parking spaces has shown an increase in frequency from January to October and then a minor dip till December.

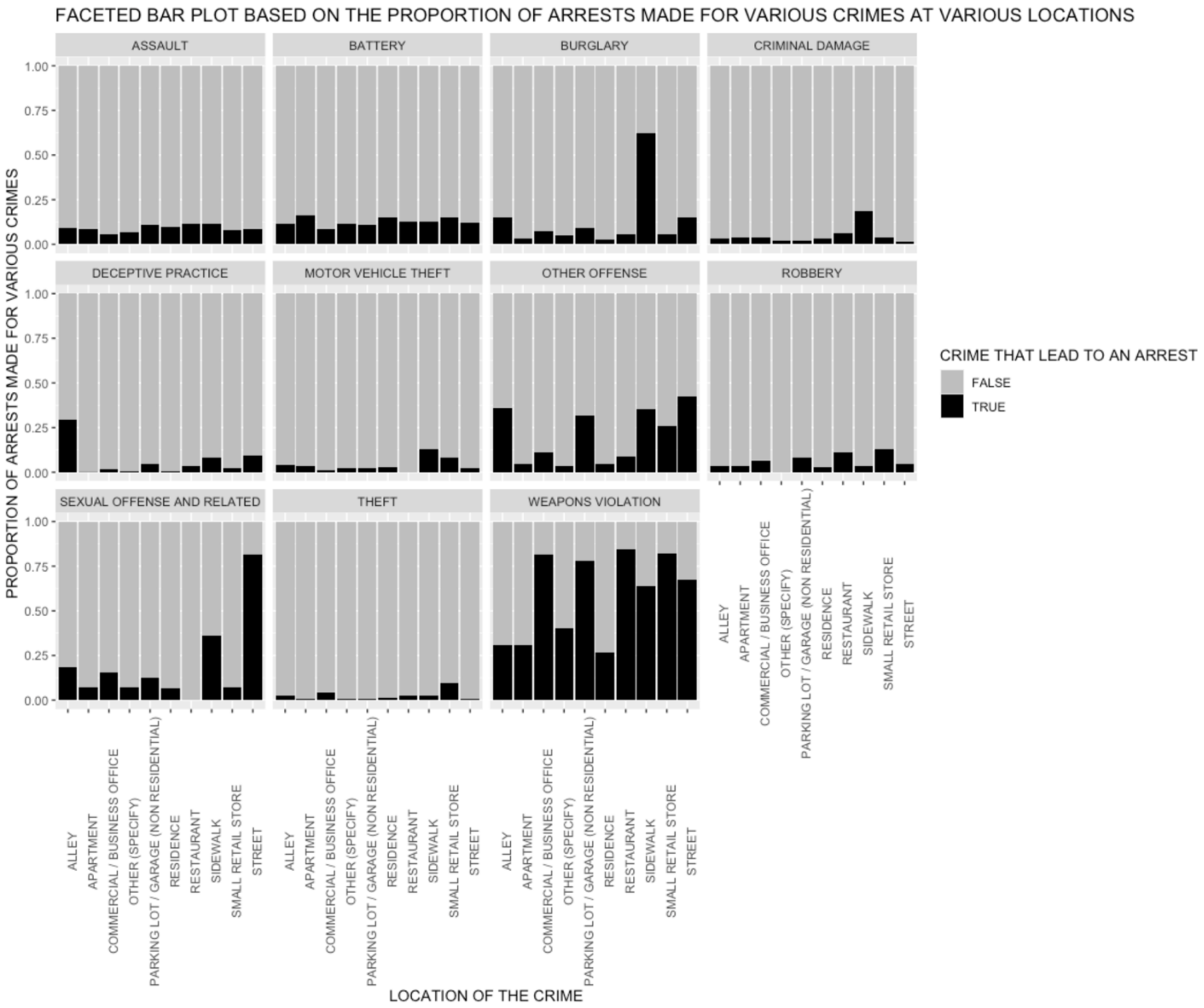


Figure. 6

After we understood how various crimes have varied based on the location type over the year, we then started understanding how the arrests have varied depending on how the type of crime and the location type of the crime. This can be best understood using a graph, where we plot, proportion of arrests in each type of location faceted by the type of crime.

From this graph, we have observed that the weapon violation is the type of crime which resulted in the highest chance of getting arrested despite any location. This could be due to seriousness of the issue and the extent of harm this crime can cause. We can also observe that the crimes committed in public areas like street, alley and sidewalk usually result in higher arrest proportions than the crimes done in the public area. Out of various arrests that are made, sexual offense on the streets usually results in an arrest about 80% of the time.

Since weapon violation resulted in the most proportion of arrests being made, we tried understanding how the weapon related arrests has varied over the year 2022.

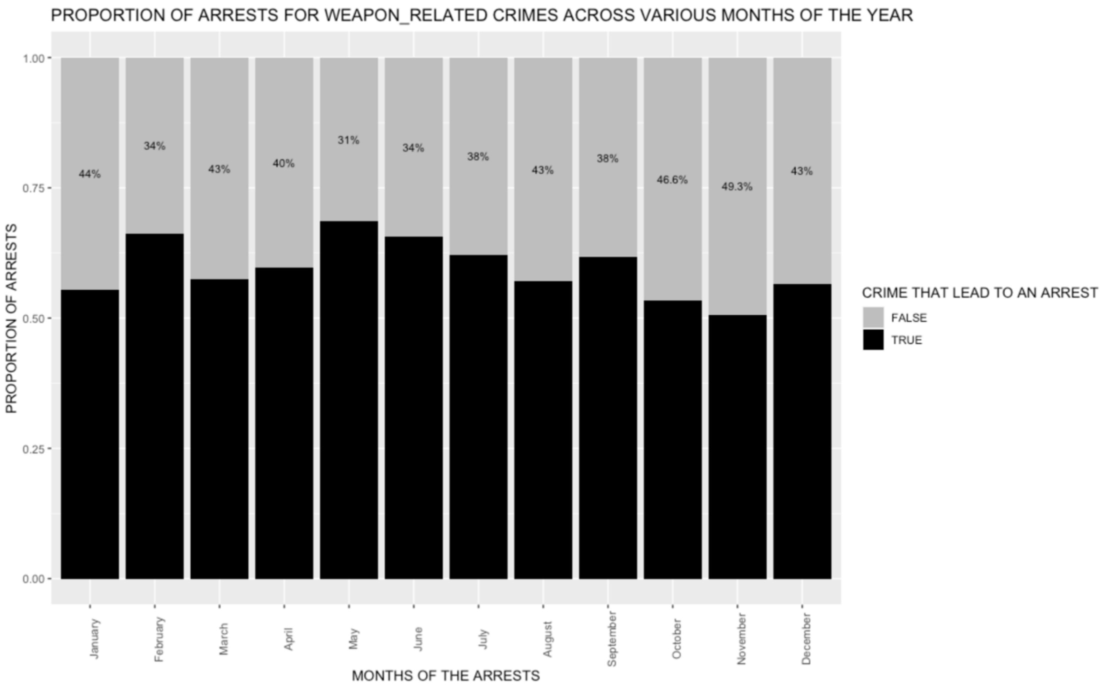


Figure. 7

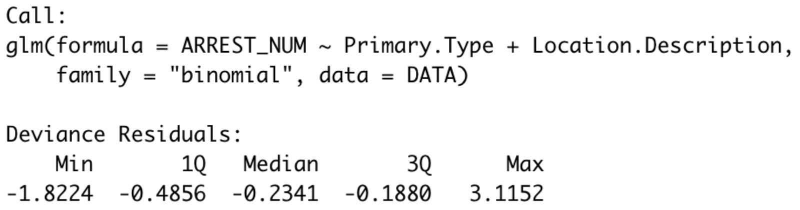
From this graph we could understand that the weapon violation arrests had been consistent over the year with only minor variations between the arrest percentages between various months over the year 2022.

FITTING A MODEL TO THE DATA

The logistic regression model used in this project aimed to investigate whether there was a significant relationship between the likelihood of an arrest being made, crime type, and location.

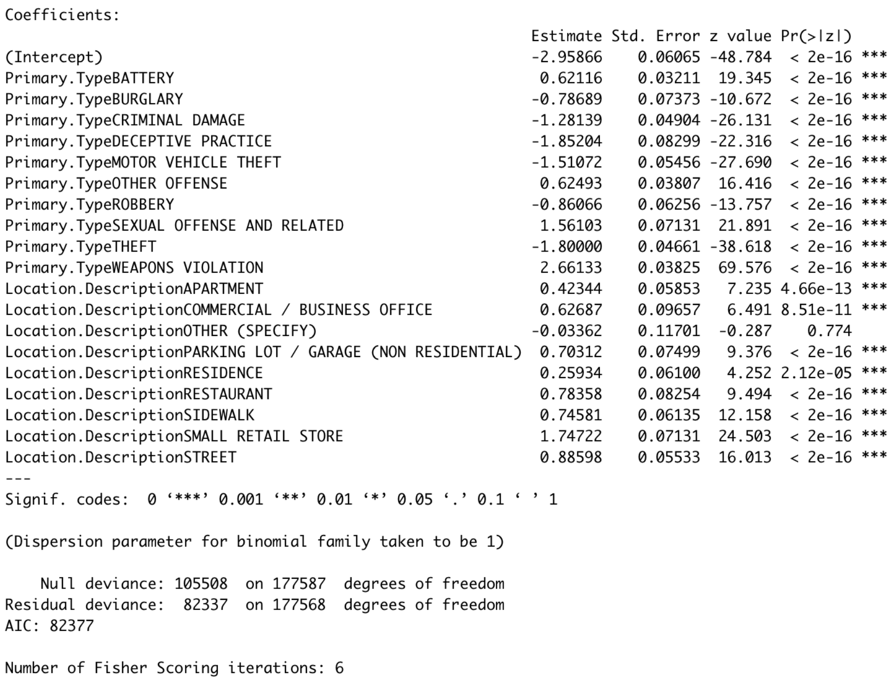
The model used binary response variables, making it a suitable choice for analyzing the dataset. The model's binomial family is ideal for binary responses, and its additive form was used to simplify the interpretation of the results.

The predictor variables used in the model were the crime type and the location description, which are both categorical variables. The target variable, which was whether an arrest was made or not, was represented by the "ARREST_NUM" variable. By using these variables, the model can predict the probability of an arrest being made given the crime type and location description in the city of Chicago.



```
Call:
glm(formula = ARREST_NUM ~ Primary.Type + Location.Description,
    family = "binomial", data = DATA)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8224  -0.4856  -0.2341  -0.1880   3.1152
```



```
Coefficients:
                                                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                               -2.95866    0.06065 -48.784  < 2e-16 ***
Primary.TypeBATTERY                                        0.62116    0.03211  19.345  < 2e-16 ***
Primary.TypeBURGLARY                                      -0.78689    0.07373 -10.672  < 2e-16 ***
Primary.TypeCRIMINAL DAMAGE                               -1.28139    0.04904 -26.131  < 2e-16 ***
Primary.TypeDECEPTIVE PRACTICE                            -1.85204    0.08299 -22.316  < 2e-16 ***
Primary.TypeMOTOR VEHICLE THEFT                           -1.51072    0.05456 -27.690  < 2e-16 ***
Primary.TypeOTHER OFFENSE                                  0.62493    0.03807  16.416  < 2e-16 ***
Primary.TypeROBBERY                                       -0.86066    0.06256 -13.757  < 2e-16 ***
Primary.TypeSEXUAL OFFENSE AND RELATED                     1.56103    0.07131  21.891  < 2e-16 ***
Primary.TypeTHEFT                                         -1.80000    0.04661 -38.618  < 2e-16 ***
Primary.TypeWEAPONS VIOLATION                              2.66133    0.03825  69.576  < 2e-16 ***
Location.DescriptionAPARTMENT                              0.42344    0.05853   7.235 4.66e-13 ***
Location.DescriptionCOMMERCIAL / BUSINESS OFFICE           0.62687    0.09657   6.491 8.51e-11 ***
Location.DescriptionOTHER (SPECIFY)                       -0.03362    0.11701  -0.287    0.774    
Location.DescriptionPARKING LOT / GARAGE (NON RESIDENTIAL)  0.70312    0.07499   9.376  < 2e-16 ***
Location.DescriptionRESIDENCE                              0.25934    0.06100   4.252 2.12e-05 ***
Location.DescriptionRESTAURANT                             0.78358    0.08254   9.494  < 2e-16 ***
Location.DescriptionSIDEWALK                               0.74581    0.06135  12.158  < 2e-16 ***
Location.DescriptionSMALL RETAIL STORE                     1.74722    0.07131  24.503  < 2e-16 ***
Location.DescriptionSTREET                                 0.88598    0.05533  16.013  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 105508  on 177587  degrees of freedom
Residual deviance:  82337  on 177568  degrees of freedom
AIC: 82377

Number of Fisher Scoring iterations: 6
```

Figure. 8

The coefficients of the model indicate the effect of each predictor variable on the probability of an arrest being made. The intercept coefficient is negative, indicating that the probability of an arrest being made is low in general. The model shows that crime types such as ‘weapons violations’, ‘sexual offenses and related’ have a high positive effect on the probability of an arrest being made, while crime types such as ‘Deceptive Practice’, ‘Criminal Damage’ have a negative effect. The location descriptions such as small retail stores, streets, and parking lots have a positive effect, while locations such as ‘other’ have a negative effect.

Looking at the coefficients of the model fitted, the crime with the highest coefficient value (2.66) was ‘Weapons Violation’ which corresponds to the proportion of arrests made for committing the crime (at around 65%) with ‘Deceptive Practice’ having the lowest coefficient value (-1.85) also corresponds to the proportion of arrests made for committing the crime (at around 5%).

In this project, the predict function was used to calculate the probability of an arrest being made. Since there was no separate testing data, the training data was used to evaluate the performance of the model.

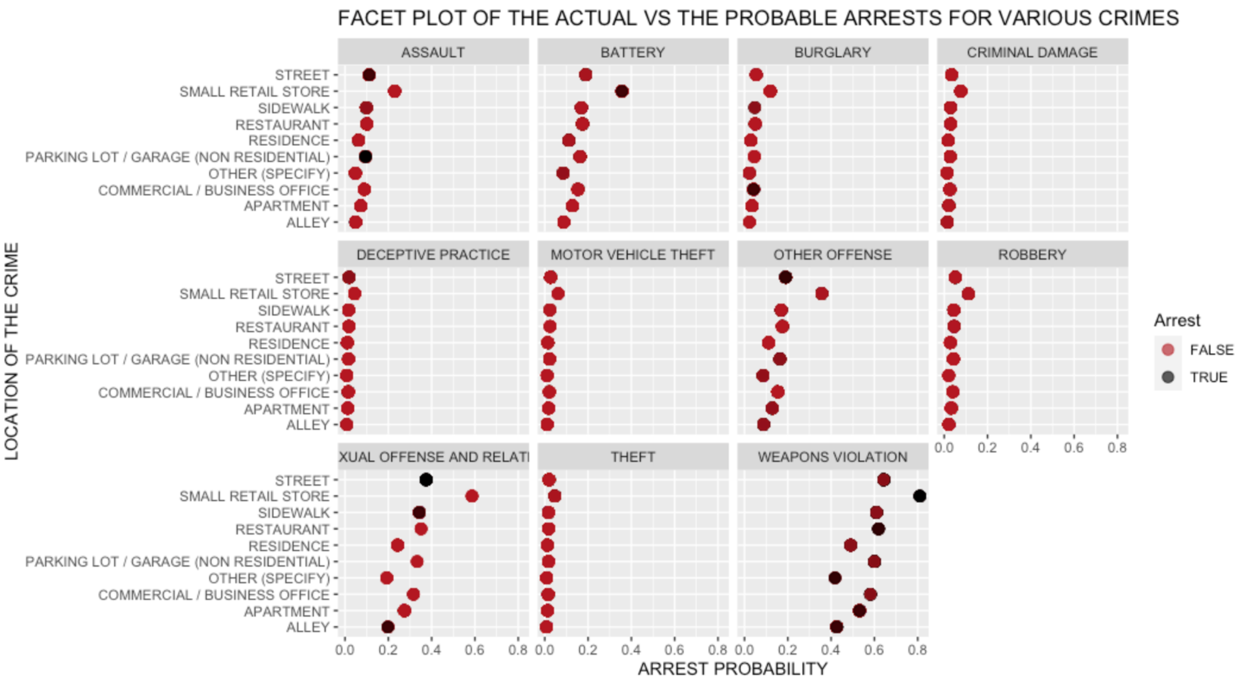


Figure. 9

On the X-axis are the projected probability of the model for making an arrest, and on the Y-axis are the actual values for making an arrest (True or False). The predictions of the model should ideally include points which are close to 1 for "arrest made = True" and points which are close to 0 for "arrest made = False."

For all areas, the predicted probability of an arrest for the crime type of ‘Weapons Violation’ is greater than 0.4, which is consistent with our earlier analysis in which we noted a high number of arrests for ‘Weapons Violation’ at over 40% (Figure. 7). On the other hand, crimes such as ‘Deceptive Practice’, ‘Motor Vehicle Theft’, and ‘Theft’ have a very low probability of being arrested (likely in between 0 and 0.1), which is also in parallel with the very low proportion of arrests, which is evident with our previous analysis (Figure. 6). For the crime ‘Sexual Offense and Related’, we can see that most of the predicted probabilities of the crimes lie in between 0.4 and 0.6, which is consistent with the proportion of arrests made for same at around 45%.

The results of the model showed a significant relationship between the predictor variables and the target variable. The coefficients for the predictor variables were all statistically significant, indicating that they were useful in predicting the probability of an arrest being made. This analysis provides insights into the factors that influence the likelihood of an arrest being made for different crime types and locations in Chicago. The model's simplicity and interpretability make it a useful tool for law enforcement agencies in Chicago to optimize their resource allocation and improve the arrest rate for different types of crimes in various locations.

CONCLUSION

Our analysis of the crimes data of Chicago for the year 2022 revealed several interesting patterns and trends in criminal activity in the city. The highest proportion of arrests was made for 'Weapons Violation' crimes, indicating that this crime type is being prioritized by law enforcement in the city. We also found that there was a significant relationship between arrests and the crime type and location of the crime, suggesting that the likelihood of an arrest being made for a particular crime type depends on the location where the crime was committed.

Our logistic model was able to predict the probability of arrest in different locations based on the type of crime committed, and the model was found to be a good fit for the data as evidenced by the low residual deviance and AIC values. The predicted probabilities of arrest across different locations faceted by crime type were consistent with the proportions of arrests made for each crime type and location. These findings provide valuable insights for law enforcement agencies and policymakers in the city who can use the information to better allocate resources and prioritize efforts to combat crime.

Overall, our project demonstrates the importance of exploratory data analysis in understanding patterns and trends in criminal activity in a city. By examining the crimes data of Chicago, we were able to identify factors that influence the likelihood of an arrest being made for different crime types and locations and provide valuable insights for stakeholders in the city.

LIMITATIONS

It should be noted that the conclusions and insights gained from this project only pertain to the year 2022 and cannot be generalized to other years. Analyzing data from multiple years may reveal new crime categories and locations that were not previously considered, leading to a more comprehensive understanding of crime trends in Chicago.

FUTURE WORK

In the future, we plan to expand our analysis by incorporating data from multiple years to gain a better understanding of how crime trends have changed over time in Chicago. This will enable us to identify patterns and trends that may not be apparent in a single year of data. Additionally, we can fit the data with various other models and compare the prediction accuracy of the future crimes, selecting the best model to base our resources on to prevent crimes at places.

APPENDICES

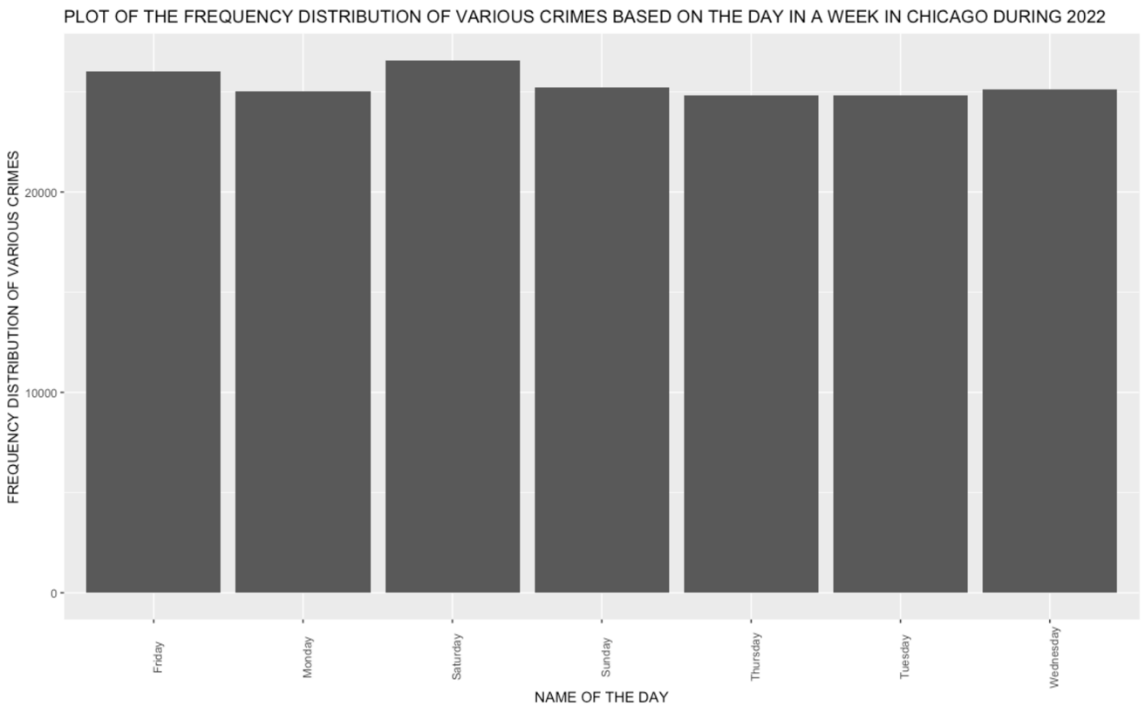


Figure I

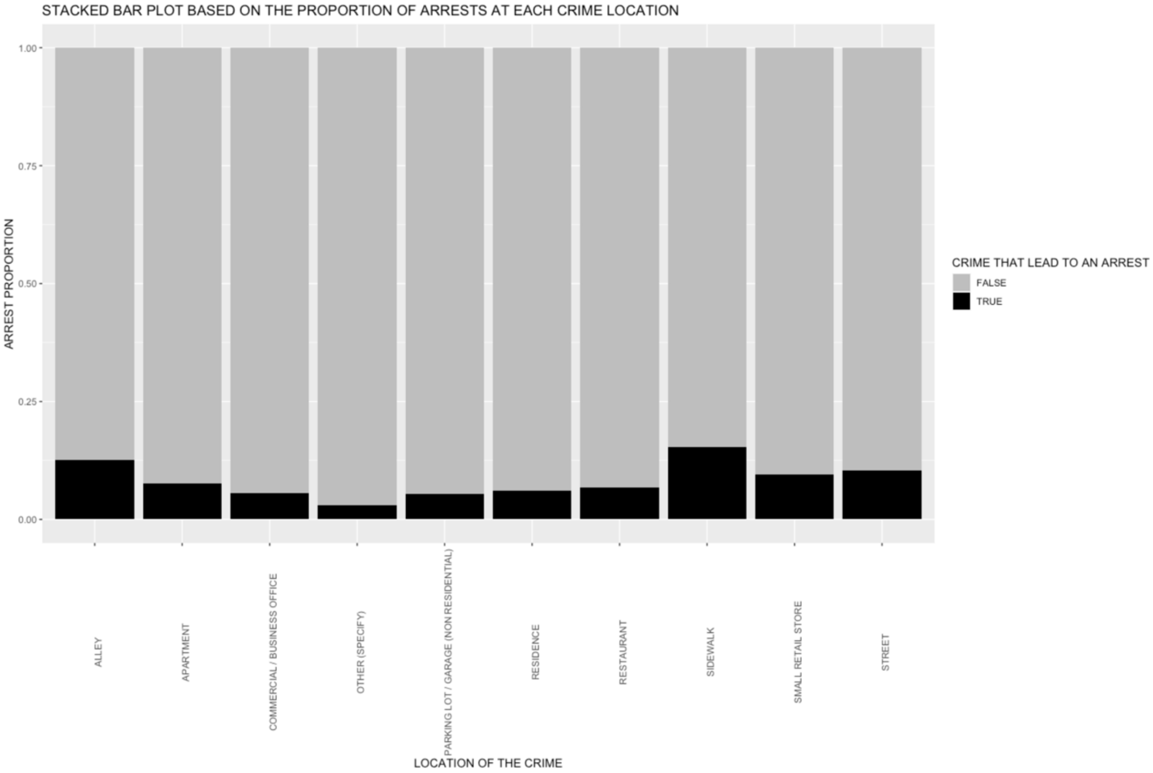


Figure II

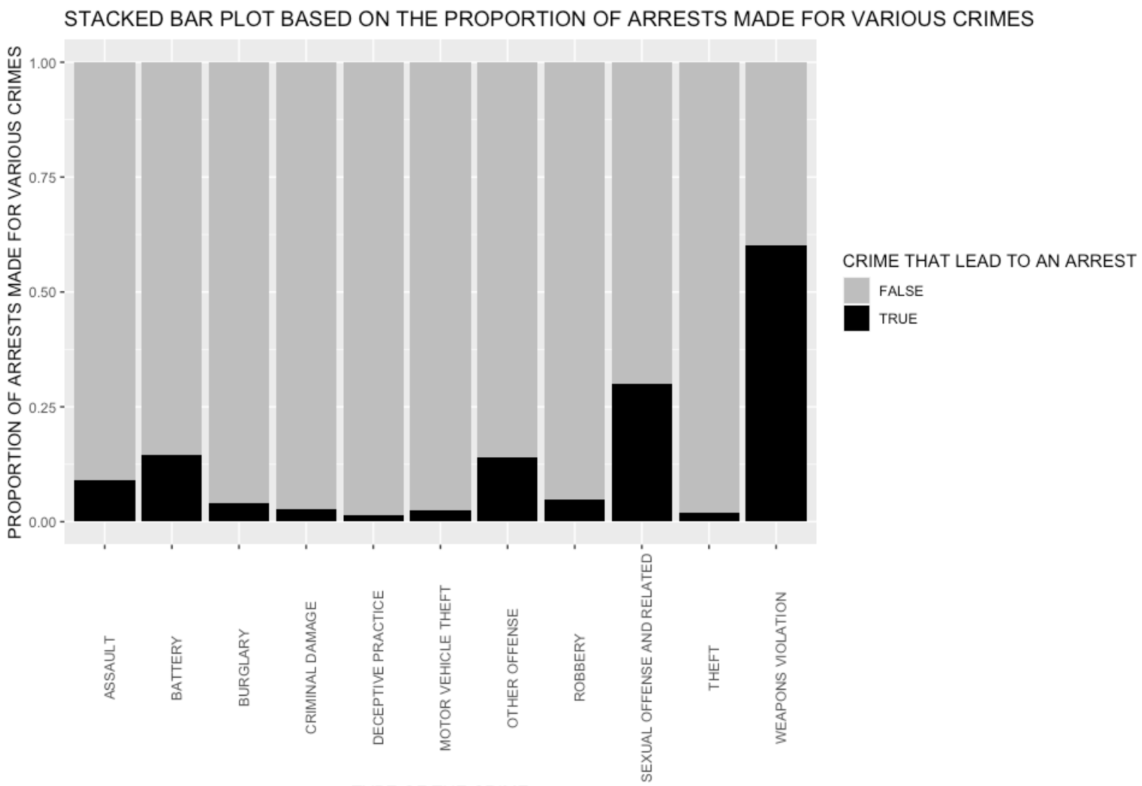


Figure. III

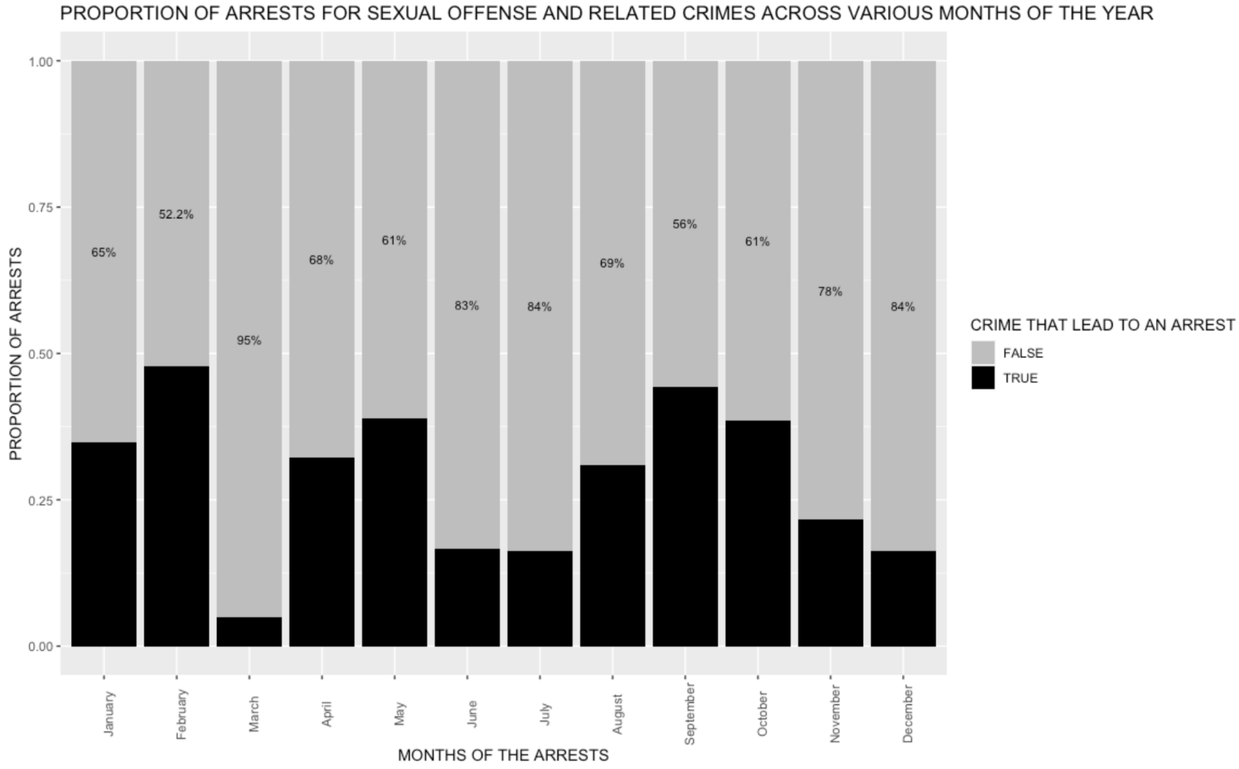


Figure. IV

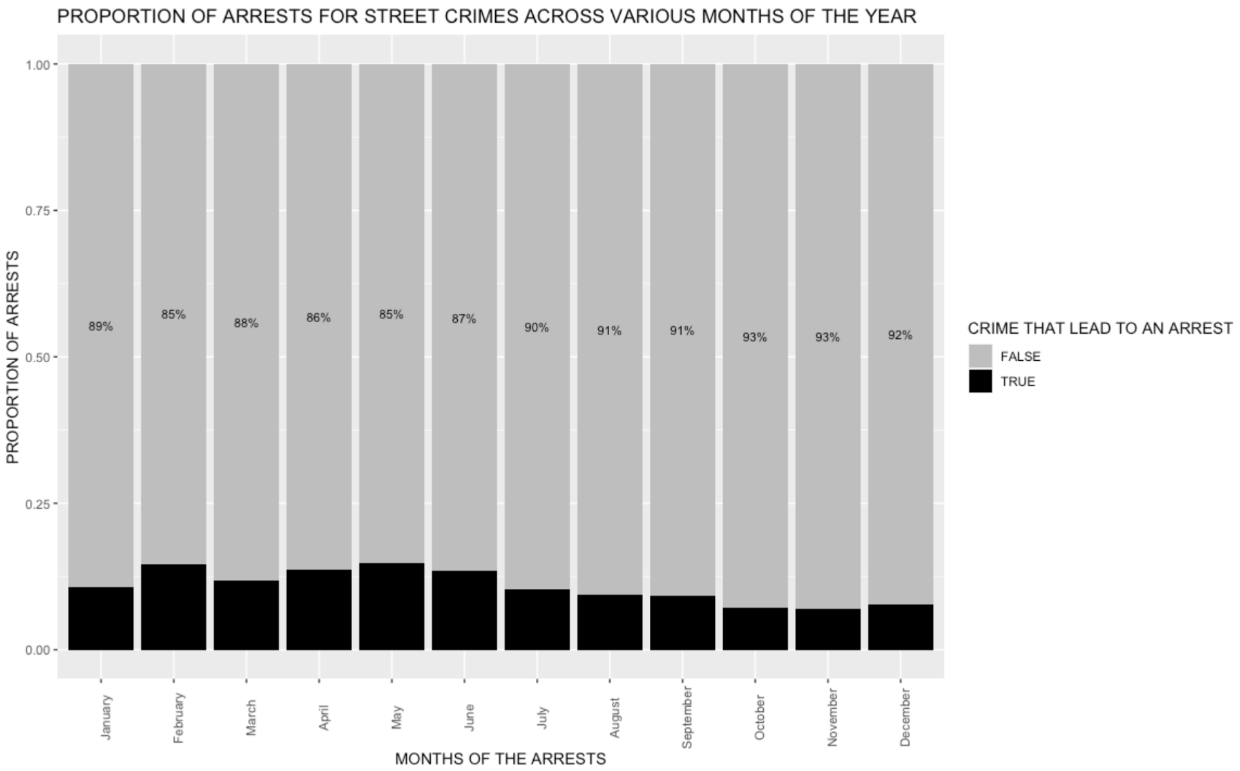


Figure. V

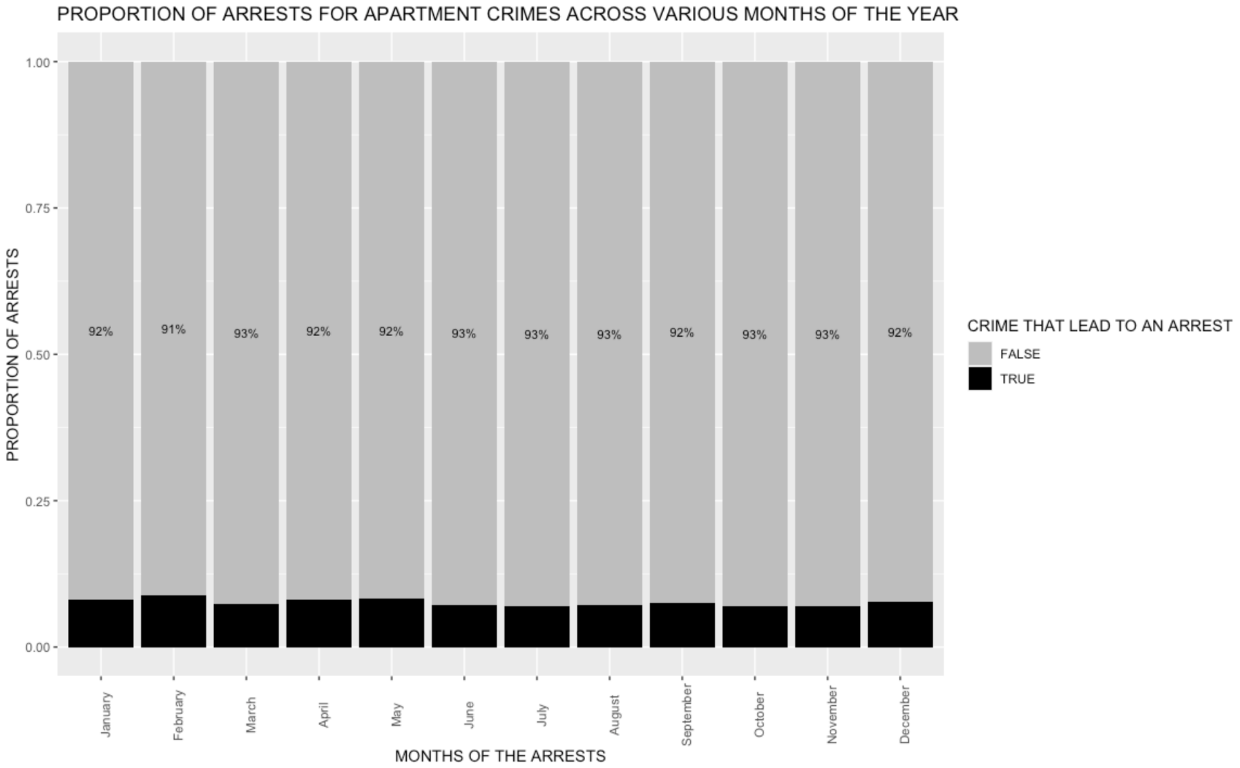


Figure. VI

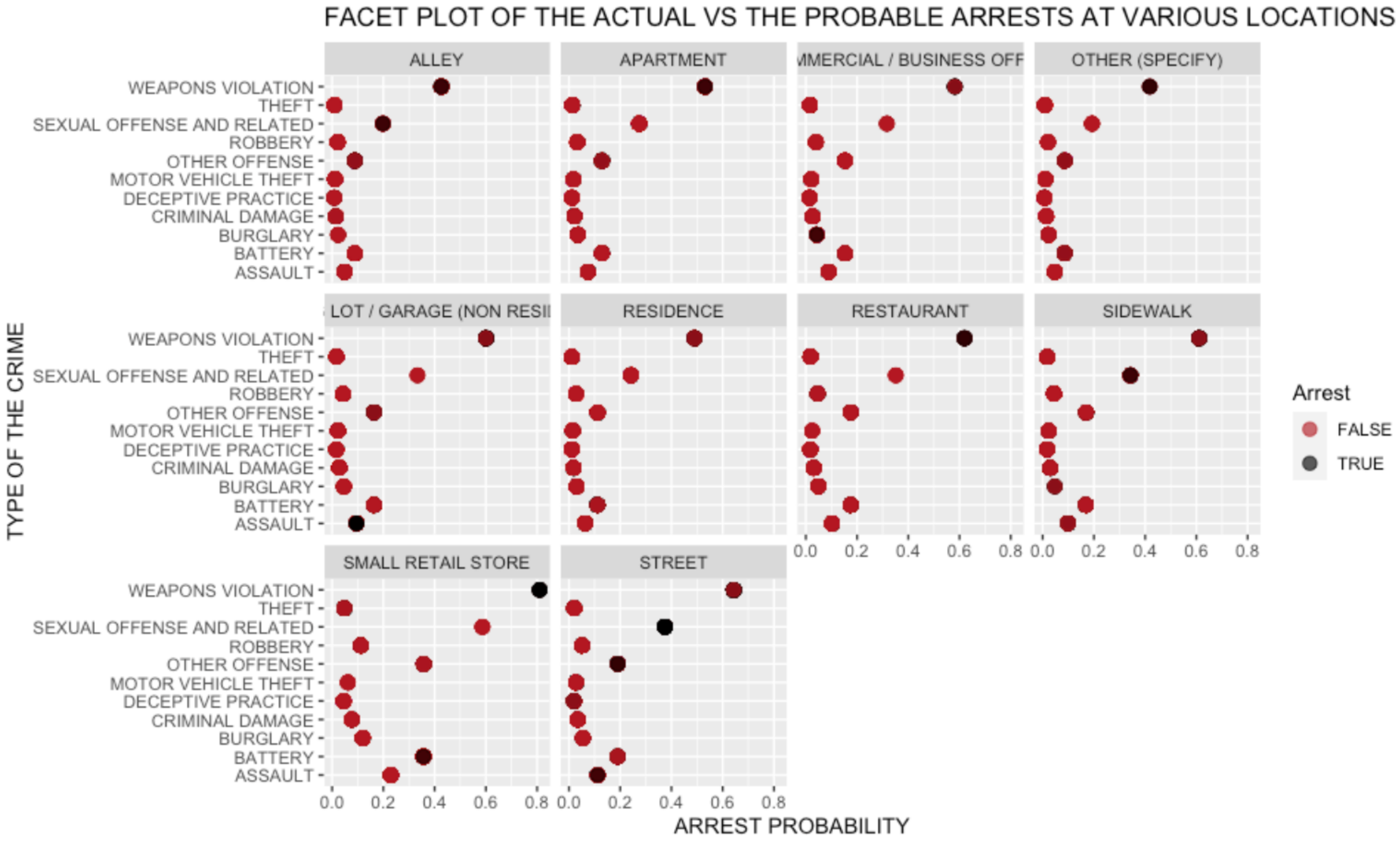

FACET PLOT OF THE ACTUAL VS THE PROBABLE ARRESTS AT VARIOUS LOCATIONS
ALLEY
APARTMENT
MMERCIAL / BUSINESS OFF
OTHER (SPECIFY)
LOT / GARAGE (NON RESII
RESIDENCE
RESTAURANT
SIDEWALK
SMALL RETAIL STORE
STREET
WEAPONS VIOLATION
THEFT
SEXUAL OFFENSE AND RELATED
ROBBERY
OTHER OFFENSE
MOTOR VEHICLE THEFT
DECEPTIVE PRACTICE
CRIMINAL DAMAGE
BURGLARY
BATTERY
ASSAULT
TYPE OF THE CRIME
0.0
0.2
0.4
0.6
0.8
ARREST PROBABILITY
Arrest
FALSE
TRUE

Figure. VII