STAT S – 670 Exploratory Data Analysis.

CHICAGO CRIMES DATA ANALYSIS

Team members: Ajinkya Pawale (ajpawale@iu.edu) , Anuhya Sankranti (ansankra@iu.edu) , Sharan Sumbad (ssumbad@iu.edu) , Shivani Vogiral (svogiral@iu.edu)

Introduction:

Almost every country is experiencing a tremendous rise in crimes. There is a pressing need to discover crime patterns and study various types of criminal activities. Security agencies across the globe are working hard to reduce these crimes; nevertheless, the volume of crime data is continually growing, making it challenging to manage such a large amount of data and keep track of crimes that occur across many geographies and over time periods. As a result, having a criminal information system that can process massive amounts of data in a short amount of time is critical. Through this project we worked on the Chicago crimes dataset to analyze different factors effecting the increase in crime rate in Chicago.

Statement of Goals:

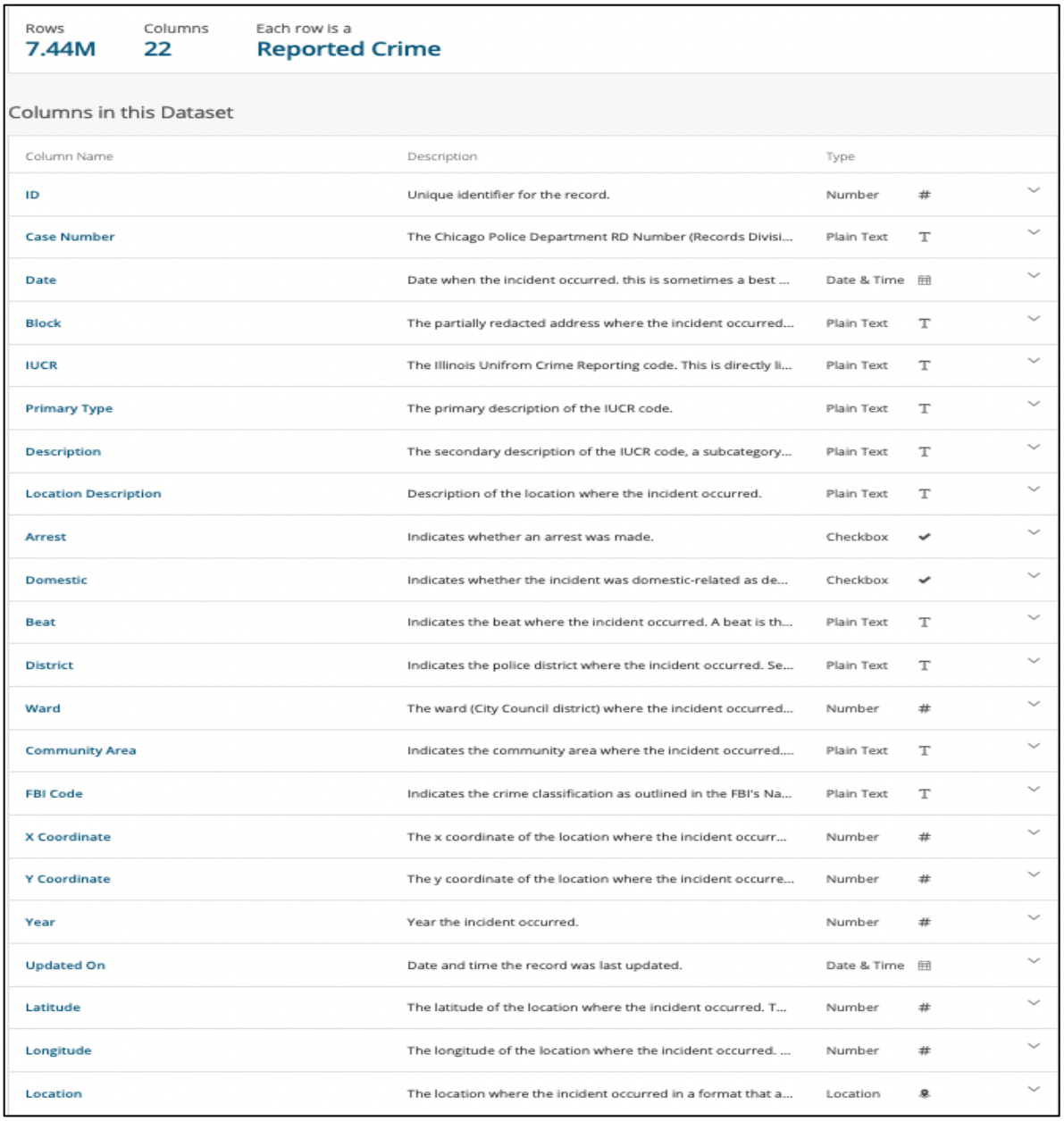
In this project we are trying to analyze the proportion of arrests made for a particular crime after it was reported to the Chicago Police Department. There are various variables in the dataset such as crime type, location, arrest made etc. Our goal is to predict whether the arrest was made or not for a particular crime depending on the other variables in the dataset like crime type, location, month, day, etc.

Theft is the most common property crime and assault is the most common violent crime in Chicago, according to reports most of these crimes go unsolved. This would make us think does the crime type and the location of the crime matter for a case to be marked solved in the city of Chicago? Perhaps this can be analyzed by understanding the interactions between location and crime type and other factors such as the month, weekday of the crime reported.

There are so many records and documentation in the police department that have been gathered during the years, which can be used as a valuable source of data for the data analytics tasks. Applying analytical task to these data bring us valuable information that can be used to increase the safety of our society and lower the crime rate.

Dataset Description:

The data set we have worked with is taken from the data provided by Chicago police department. This dataset has the following columns:

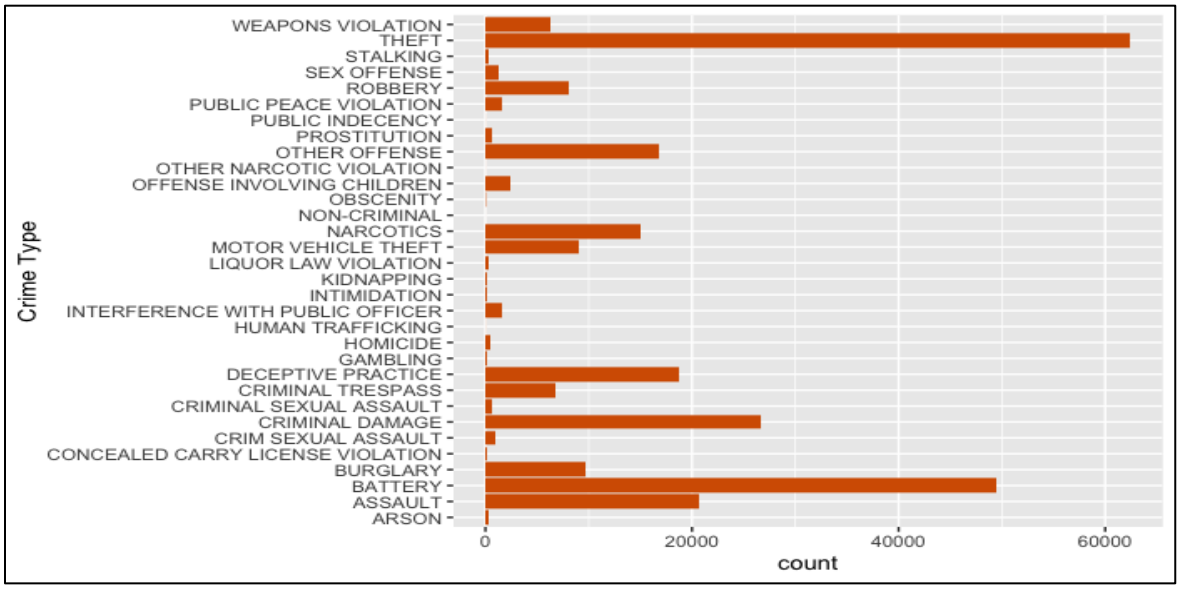


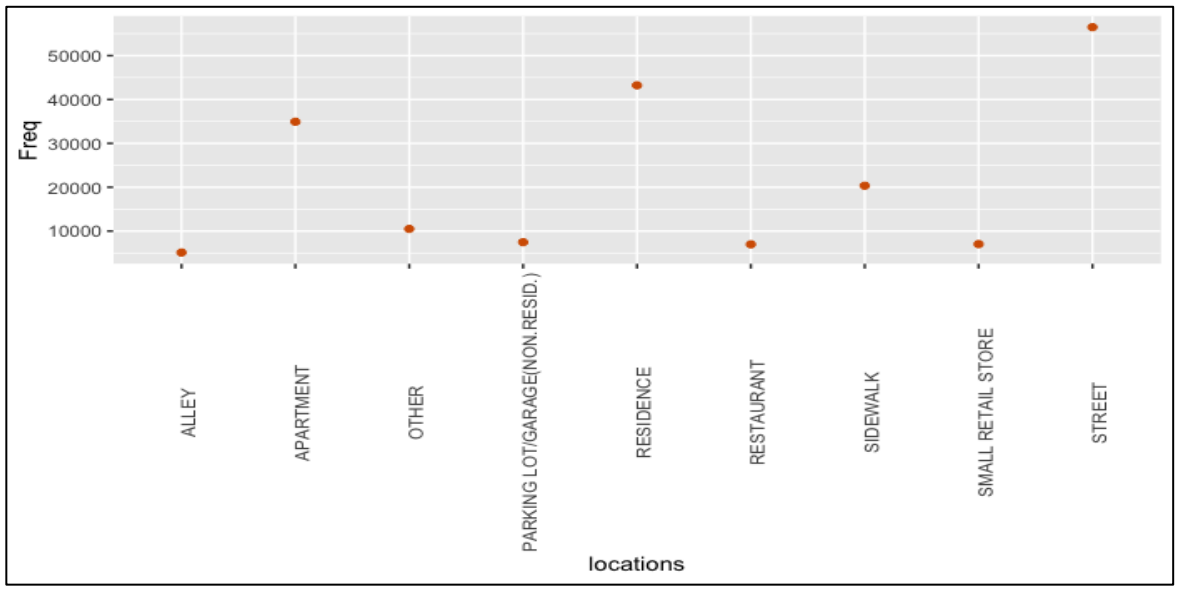
Rows 7.44M | Columns 22 | Each row is a Reported Crime

Columns in this Dataset

Column Name	Description	Type
ID	Unique identifier for the record.	Number
Case Number	The Chicago Police Department RD Number (Records Divisi...	Plain Text
Date	Date when the incident occurred. this is sometimes a best ...	Date & Time
Block	The partially redacted address where the incident occurred...	Plain Text
IUCR	The Illinois Unifrom Crime Reporting code. This is directly li...	Plain Text
Primary Type	The primary description of the IUCR code.	Plain Text
Description	The secondary description of the IUCR code, a subcategory...	Plain Text
Location Description	Description of the location where the incident occurred.	Plain Text
Arrest	Indicates whether an arrest was made.	Checkbox
Domestic	Indicates whether the incident was domestic-related as de...	Checkbox
Beat	Indicates the beat where the incident occurred. A beat is th...	Plain Text
District	Indicates the police district where the incident occurred. Se...	Plain Text
Ward	The ward (City Council district) where the incident occurred...	Number
Community Area	Indicates the community area where the incident occurred....	Plain Text
FBI Code	Indicates the crime classification as outlined in the FBI's Na...	Plain Text
X Coordinate	The x coordinate of the location where the incident occurr...	Number
Y Coordinate	The y coordinate of the location where the incident occurre...	Number
Year	Year the incident occurred.	Number
Updated On	Date and time the record was last updated.	Date & Time
Latitude	The latitude of the location where the incident occurred. T...	Number
Longitude	The longitude of the location where the incident occurred. ...	Number
Location	The location where the incident occurred in a format that a...	Location

This dataset includes recorded crime episodes in the City of Chicago from 2001 to the present (with the exception of murders where data exists for each victim). As this is a huge data to work with, we limited the year to 2019 and worked with the data containing different crimes of 2019. We also limited the columns in the dataset by selecting the columns that were appropriate for prediction and analysis.

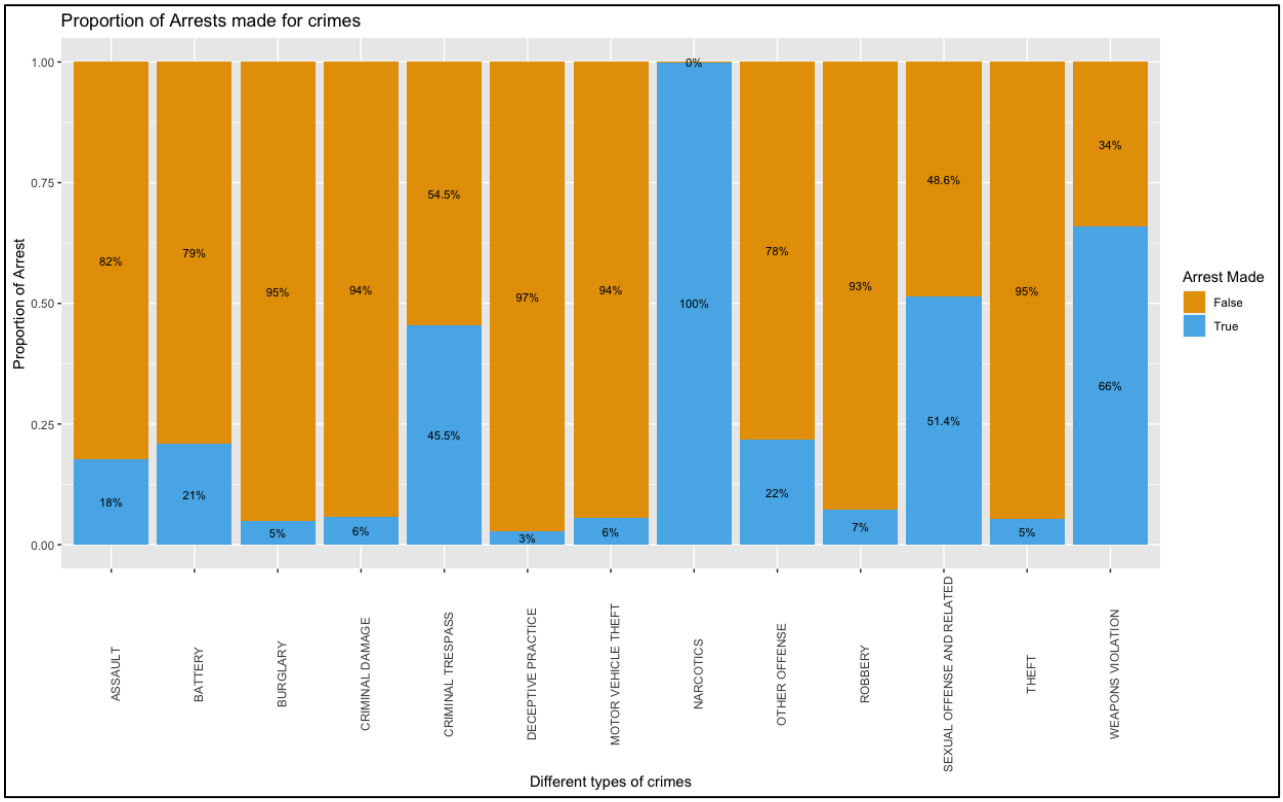




Out of all the different types of crimes and different locations of crimes taking place, we chose the types of crime and location that has frequency greater than 5000. There were few crime types such as ‘Obscenity’, ‘Criminal Sexual Assault’, ‘Sex Offense’, ‘Prostitution’ and ‘Crim Sexual Assault’ that did not have much frequency in the dataset individually but when grouped together they had a good amount of proportion in the dataset. So, we grouped all these crime types to “Sexual Offense and Related”. We also converted a few categorical columns to their respective numerical encodings to make use of them while modelling if required. After all this preprocessing we were finally left with 184330 rows and 19 columns.

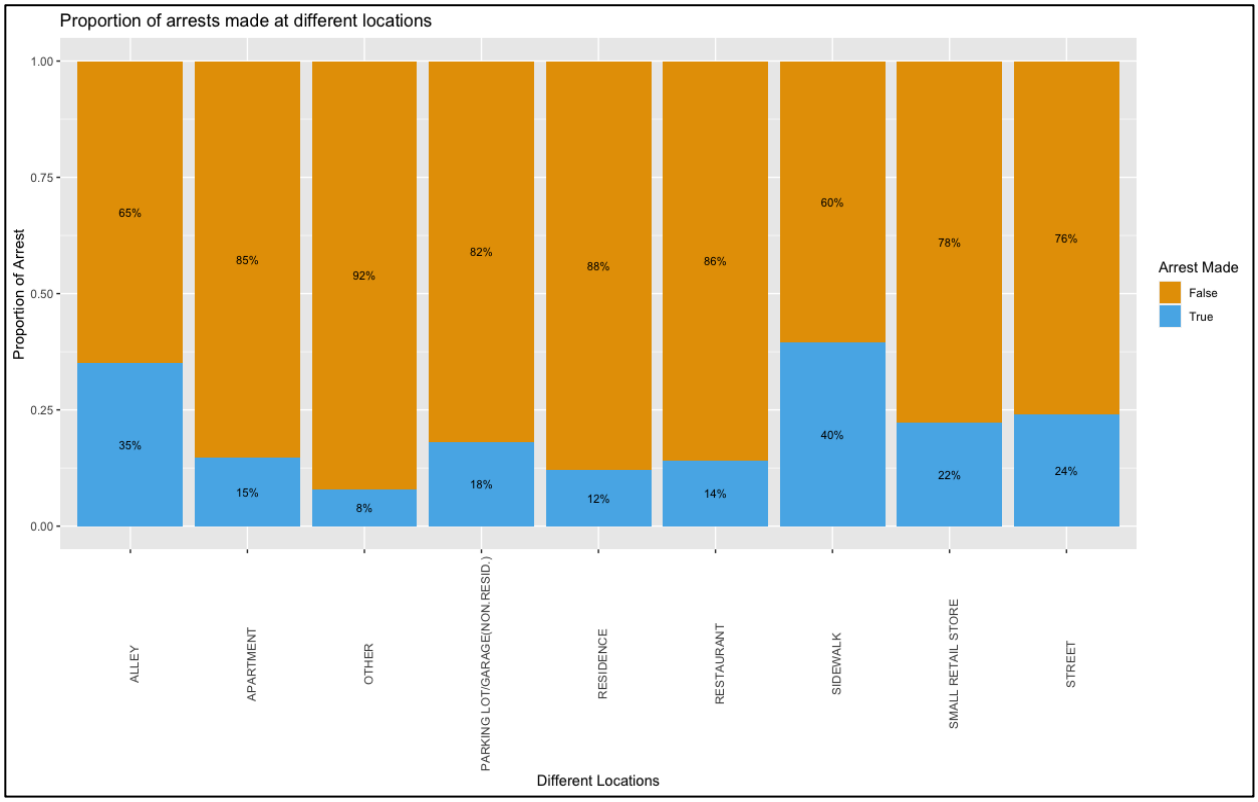
To further analyze the data, we have plotted proportion of arrests with respect to different columns of the dataset. Initially, we made a plot showing the proportion of arrest, i.e., if the particular crime led to arrest or not, with respect to different types of crimes we have in our dataset.

The plot below shows the proportion of arrest for different crime types:



From the above plot we can see the following observations: All the crimes in Narcotics lead to arrests. This is not because of its low proportion because it is well above 5000 which we set as the threshold while reducing the dataset. One thing to note here is that the data is only for 2019 so if we compare the data from other years, we can draw better conclusions. Other crimes like Weapons Violation, Sexual Offense and Related, Criminal Trespass show a significant number of crimes leading to arrests as compared to others. Theft is the most common property crime in Chicago but the percentage of arrests for it is only 5%. Also, Assault is the most violent crime but the percentage of arrests for it is only 18, which should be a major concern for the Police Department.

The plot below shows the proportion of arrest for crimes at different locations:



From the above plot we can see the following observations: The frequency of crimes at Street is the highest in the dataset, but the number of arrests made are only 24% from the above plot. Similarly, the next highest frequencies of crimes are at Residence and Apartment, but the number of arrests made for them are only 12% and 15% respectively. So, this is also a major concern that the Police Department should consider.

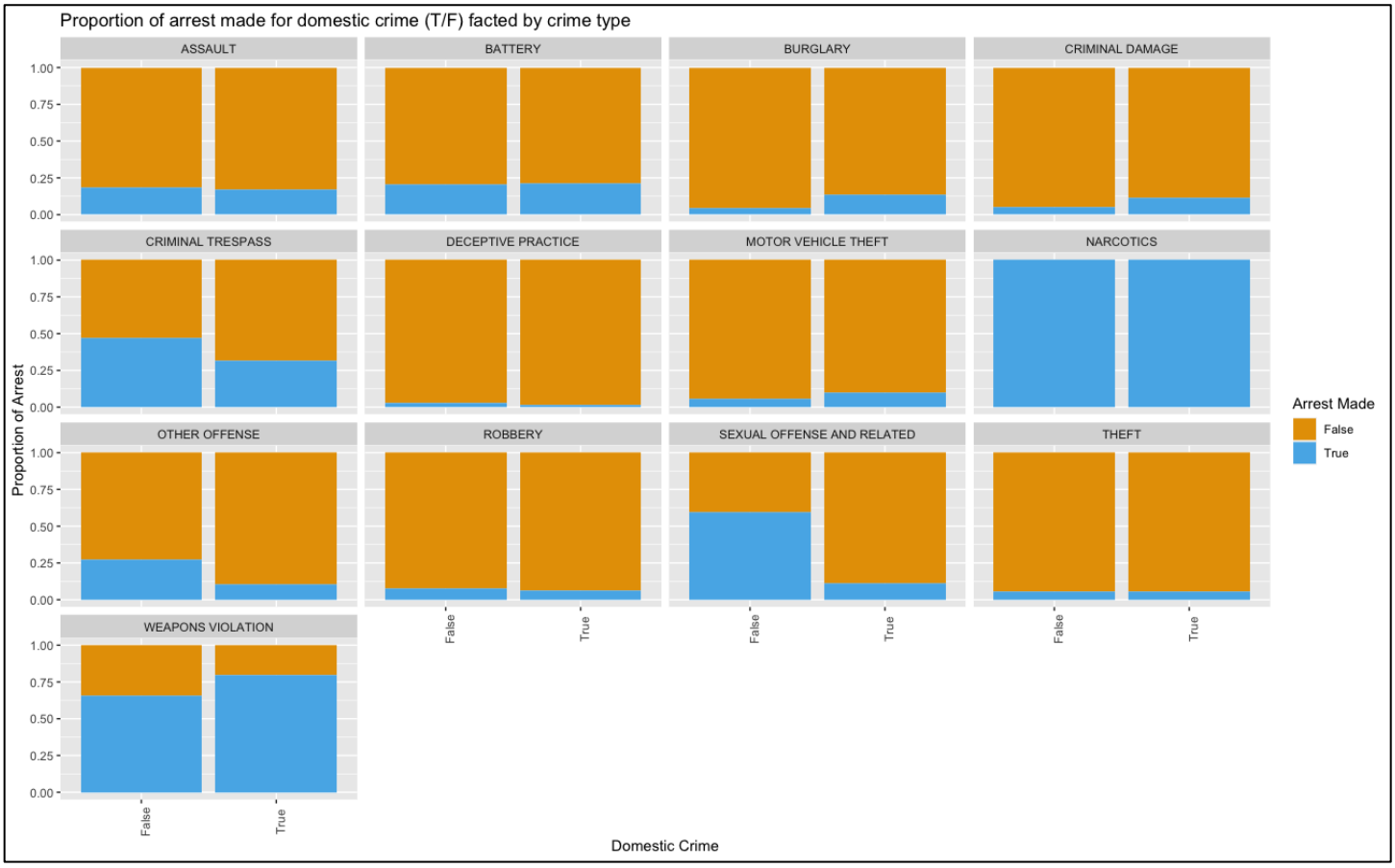
Initially, we tried to plot domestic crime, month and week variables against the arrest made, but we did not get much variation with any of these variables. So, we tried to combine domestic crime variable with the crime type and location variable. It made more sense to plot the interaction between domestic crime and crime type variable against the arrest made variable.

Similarly, we plotted month and week variables against the target, arrest made but there was not much interaction between them and didn't make much difference even when combined with some other variables in dataset. So, we didn't consider these attributes for further analysis.

This plot shows the proportion of arrest at different locations for different types of crimes:

From the plot we can observe that there is good interaction between different locations and crime type with respect to the proportion of arrest made. There are few interpretations that can be made from this plot like, Burglary at Alley has good number of arrests made while Burglary at an Apartment has low number of arrests made. In the same way if we look at sexual offense and related, the highest proportion of arrests made can be observed at street and least at small retail store. Similar interpretations can be made from this plot for different types of crimes at different locations.

This plot shows the proportion of arrest for domestic and non-domestic types of crimes for crime types:



From the above plot we can see the following observations: There exists interaction between crime type and domestic crime with respect to proportion of arrest made. We can see that for Sexual Offense and Related, the proportion of arrest made for Domestic Crime = True is very less, which is a concern and needs to be addressed.

Results and insights: (Modeling and answering our questions)

We considered an additive model with the explanatory variables: crime type, location, and domestic crime. The target variable was as mentioned before, arrest made. The reason why we chose an additive model is because it is simple to interpret as compared to interactive/multiplicative models. Here, in our case there are a lot of categorical values inside each categorical variable, so applying a multiplicative model becomes difficult and cannot be understood properly while visualizing the results or comparing them with the ground truth.

This is our model below:

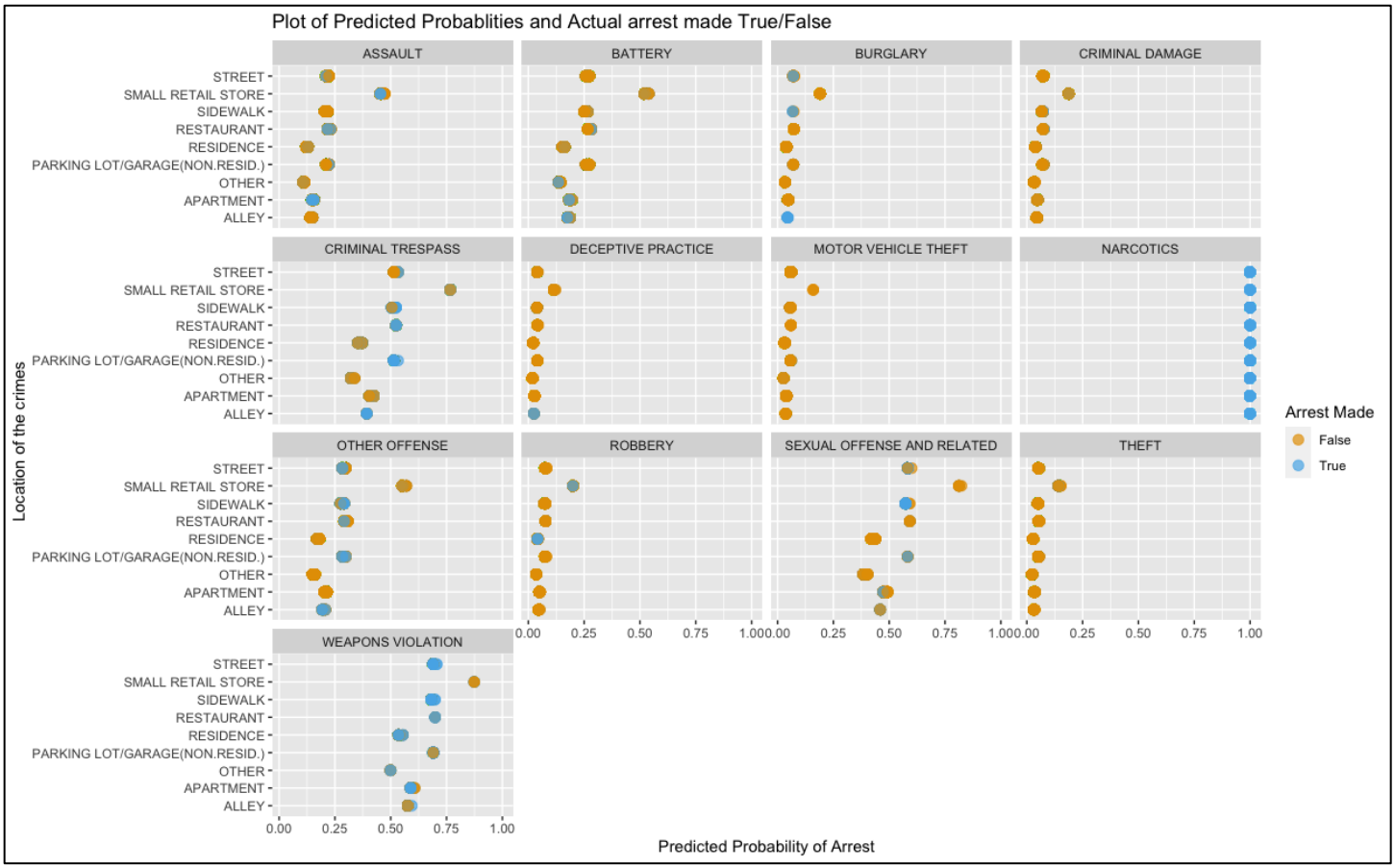
```
Call:
glm(formula = arrest_num ~ primary_type_x + location_description_x +
    domestic_x, family = "binomial", data = crimes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0369  -0.5954  -0.3220  -0.2169   2.8291
```

Where, primary_type_x is the crime type, location_description_x is the location of the crime and domestic_x is the domestic crime (T/F). Target variable is arrest_num which is arrest made or not.

```
Coefficients:
                                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                                 -1.81446    0.05233 -34.675  < 2e-16 ***
primary_type_xBATTERY                        0.26420    0.02540  10.401  < 2e-16 ***
primary_type_xBURGLARY                      -1.26132    0.06049 -20.850  < 2e-16 ***
primary_type_xCRIMINAL DAMAGE               -1.27695    0.03714 -34.380  < 2e-16 ***
primary_type_xCRIMINAL TRESPASS              1.37762    0.04142  33.260  < 2e-16 ***
primary_type_xDECEPTIVE PRACTICE            -1.85822    0.05876 -31.622  < 2e-16 ***
primary_type_xMOTOR VEHICLE THEFT           -1.47441    0.05508 -26.771  < 2e-16 ***
primary_type_xNARCOTICS                      9.46876    0.50053  18.917  < 2e-16 ***
primary_type_xOTHER OFFENSE                  0.39049    0.02997  13.030  < 2e-16 ***
primary_type_xROBBERY                       -1.20337    0.05434 -22.145  < 2e-16 ***
primary_type_xSEXUAL OFFENSE AND RELATED     1.65361    0.05534  29.883  < 2e-16 ***
primary_type_xTHEFT                         -1.61269    0.03296 -48.932  < 2e-16 ***
primary_type_xWEAPONS VIOLATION              2.12208    0.03743  56.690  < 2e-16 ***
location_description_xAPARTMENT              0.05069    0.05183   0.978 0.328088
location_description_xOTHER                 -0.31067    0.06375  -4.873 1.10e-06 ***
location_description_xPARKING LOT/GARAGE(NON.RESID.)  0.49325    0.06212   7.940 2.03e-15 ***
location_description_xRESIDENCE             -0.16609    0.05177  -3.208 0.001337 **
location_description_xRESTAURANT             0.53385    0.06260   8.528  < 2e-16 ***
location_description_xSIDEWALK               0.45639    0.05265   8.668  < 2e-16 ***
location_description_xSMALL RETAIL STORE     1.62778    0.05938  27.415  < 2e-16 ***
location_description_xSTREET                 0.49366    0.04990   9.892  < 2e-16 ***
domestic_xTrue                               0.07625    0.02013   3.787 0.000152 ***
```

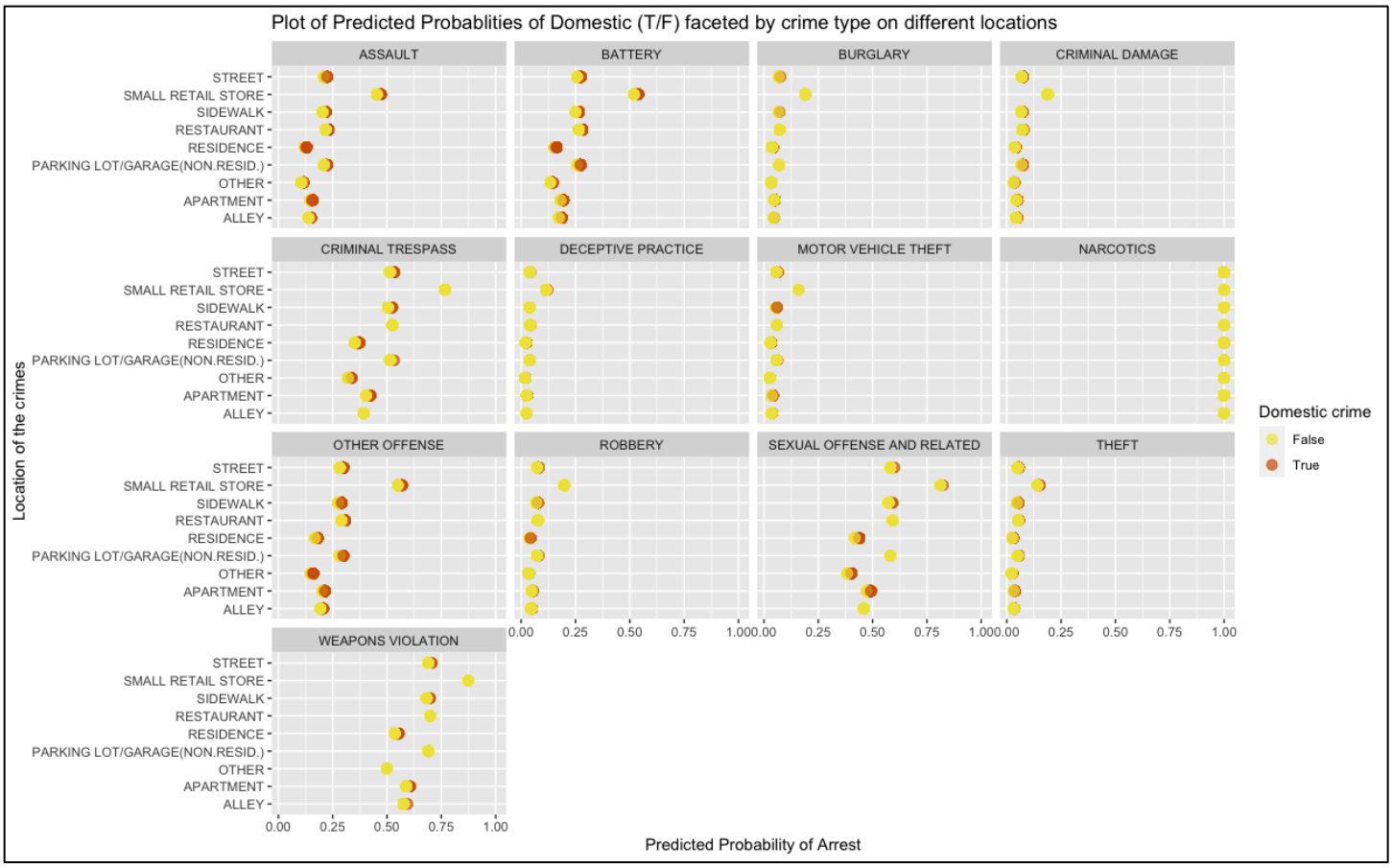
After modelling the data if we observe the coefficient values of different features in the model, Narcotics has the highest coefficient value of 9.46, this does make sense because in our data with 12160 entries of crime type narcotics, 12156 led to arrest, i.e., arrest_num is True. In the same way, crime type Deceptive Practice has the least coefficient value of -1.858. Deceptive Practice is the crime that has 12538 entries in total out of which 12176 are false for arrest made, which justifies the least coefficient value.

This is a plot of predicted arrest probabilities against the location of the crimes faceted by type of crime:



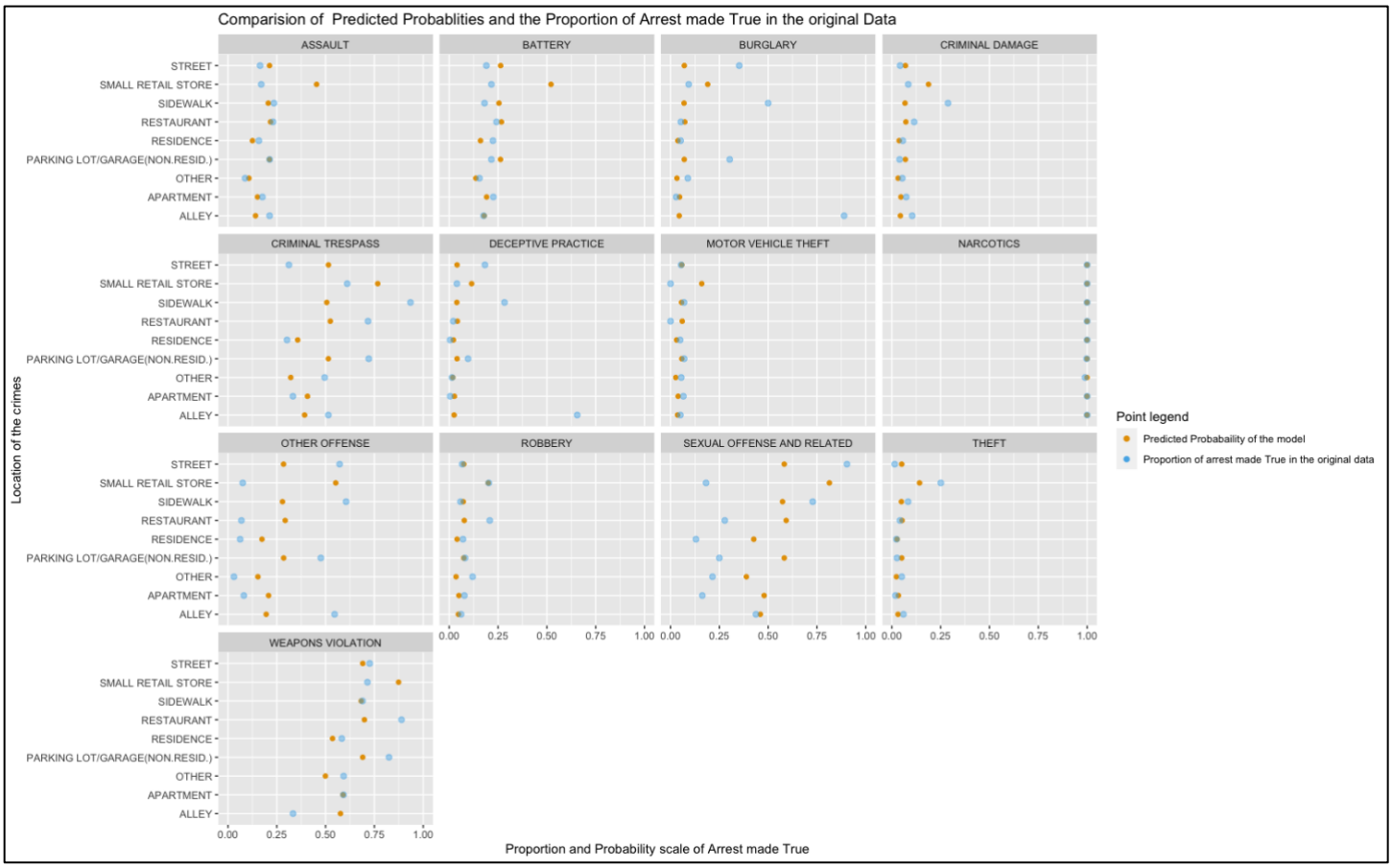
In this plot, x axis is the predicted probabilities of the model for arrest made and the label colors are the ground truth for arrest made (T/F). The points closer to 0 should be Arrest Made= False (light yellow) and points closer to 1 should be Arrest Made= True (light blue) for the model to be an ideal one. From the plot above showing proportion of arrests for different crimes we saw that narcotics has 100% of arrest made = True and here in this graph we can observe that the predicted probability of arrest for crime type narcotics is 1 for all the entries. Theft has 95% arrest made = False in the dataset and here, we can see that the predicted probability of arrest for theft is very less which does make sense. Also, if we look at Sex Offense and Related, it has 51.4% arrest made = True and 48.6% arrest made = False in the dataset hence, the predicted probabilities are distributed across the middle of the plot. Similarly, Criminal Trespass has 45.5% arrest made = True and 54.5% False in the dataset which resulted in predicted probabilities being distributed around 0.5. So, we can say that the prediction model is decent.

This is a plot of predicted arrest probabilities against the location of the crimes faceted by type of crime and domestic(T/F):



From this plot we can see that domestic type (T/F) is not contributing much as there is no much variation, the points are overlapping in many crime types and the plot as a whole doesn't describe much about the domestic crime type.

This is a plot of predicted arrest probabilities and proportion of arrests made in actual dataset against the location of the crimes faceted by crime type.



As the points in this plot are showing the predicted probabilities of arrest and proportion of arrest made from data set, the points must be overlapping which is the case for most of the crime types. But, if we look at crime types like sexual offence and related, Criminal Trespass and Other offense, the points are not overlapping, the reason for this could be the actual proportions of arrests made in the dataset. The proportions of arrest made for these crime types is not much different, the distribution is around 50% so, it does make sense.

Conclusion:

There is significant amount of interaction between the type of crime and location with respect to the target variable arrest made. Domestic crime variable doesn't contribute much individually towards the model. Although, it can be used as an interaction with the crime type variable, which we have shown above. The variation between type of crime and location of crime with factors like day of the week and month is almost constant hence, we have not included them during our model selection.

In Chicago, the kind of crime and the location of the crime are crucial variables in determining whether a case is considered solved. They can be further analyzed by looking at the interaction between location and crime type on a model and also the interaction of domestic crime variable with the crime type. As of now the results are in line with our primary research question.

Limitations:

In this project we have considered an additive model and not a multiplicative/interactive model because the interaction between the existing variables becomes complicated to interpret in terms of predicted probabilities and the target variable. Also, we have taken the data for 2019 in this case, so the interpretations are only limited to a particular year and general conclusions cannot be made from them. As we limited the year to 2019 there were few crimes that had frequency less than 5000 and were not considered because we chose only the ones above 5000, but on a broad spectrum if we consider the data from different years we may see some more different crime types with higher frequency of crimes apart from the once we have modeled.

Future work:

We would like to work on data from other years and see how the interpretations match to them as compared to our assumptions for 2019. Also, we can implement a similar model for major cities across U.S and see how the trend varies for each one of them.