BUSINESS ANALYTICS

GROUP 7

LAIBA SAQIB - 23110354
HADIYA AZHAR - 23110125
MUHAMMAD SAAD - 23110064
ABDUL HANNAN ANJUM - 23110058
MUHAMMAD JUNAID RIAZ - 23110164

INTRODUCTION

BACKGROUND

Along with the positive impact globalization has across the globe, the world has also seen crimes emerging through new means creating disruptions in the welfare and well-being of society. USA is no exception in this regard. With advancements in technologies, it is now easy to record a variety of information about the crime's subject, site, and kind in real time. It is also possible to obtain data that can be utilized to understand the attributes or for predictive purposes by evaluating these acquired data using different data mining techniques. Hence, in our project, we focus our analysis on crime data of Maryland, USA. This is because Maryland's economy is still performing better than the nation as a whole which is why this state holds immense importance for the USA. Our aim is to predict the characteristics of crimes that have a high probability of occurring in various counties of Maryland. For this purpose, we will first analyze one of the counties of Maryland, Montgomery County, in detail because it is not only the most affluent county in Maryland with the largest population but has the seventh largest sheriff office which provides us critical information on crimes that may not be available for other counties (MCSO, n.d.). This data is then used to make predictions about other counties of Maryland.

LITERATURE REVIEW

The best and most beneficial thing that law enforcement authorities can do from the standpoint of society as a whole is to prevent crime. Crime-related societal expenses and suffering can be entirely avoided if crimes are successfully (and justly) averted before they happen. According to Graham Ferrell, crime prevention is a method of shielding an event that has not yet occurred or has chances of occurring in future. He claims that the prevention of crime depends on two bases: a reliable prediction of victimization and a cost-effective method of prevention that can be practically implemented (Farrell).Therefore, this paper evaluates the former and tries to develop reliable predictions of criminal behavior. In addition to this, research and analysis on crime statistics can not only identify trends that help with operational response to crime and crime prevention but is also a means of accountability to the public and police leadership (Schwabe). Since we will be focusing on the reliable prediction of crimes or the type of crimes, authors tend to believe that socioeconomic factors such as poverty, employment, education and community safety can become important indicators of future crime. Generally speaking, the literature places high emphasis on poverty as a predictor of crime. Effective strength and administrative capabilities of law enforcement agencies also is seen as important. The authors note that both socioeconomic factors and police accountability are equally important in order to prevent crime.

PROBLEM STATEMENT

Crime analysis is a concept that entails preventing future crimes, recognizing current crimes and crime trends, and implementing the required countermeasures. In our project, we are going to build on the relationship between a crime and the criminal, the number of criminals' active regions, and the ability to classify the following type of crimes before they happen. Our analysis is going to be beneficial for Maryland’s police and officials in addressing security challenges in the state.

DATA SELECTION

For the given problem, we have chosen 3 datasets from http://www.data.gov that were in interest to the topic we have selected.
The first data named “Crime” contains 3,16,904 observations of each criminal incident that has been reported majorly in Montgomery County of Maryland. The data has 30 variables namely:
Incident ID, Offence code, CR Number, Dispatch Date / Time, NIBRS Code, Victims, Crime Name1, Crime Name2, Crime Name3, PRA, Address Number, Street Prefix, Street Name , Street Suffix, Street Type, Start_Date_Time, End_Date_Time, Latitude, Longitude, Police District Number and Location.

Useful information that we can extract from this data is the number of victims in each incident, what type of crime is occurring in terms of crime against person, property or society which is further categorized into robbery, rape, shoplifting, driving under influence etc., places where these crimes are taking place, time of crime and the exact locations of crime sites.

The second dataset named “Violent_Crime___Property_Crime_by_County__1975_to_Present” contains 1104 observations. It gives us data from 1975 to 2016 on total crimes categorized into 7 categories: Murder, Rape, Robbery, Aggravated Assault, Breaking & Entering, Larceny Theft and Motor Vehicle Theft. The data also merges data into violent and property crime while calculating percent changes in these types of crimes.
The third dataset named “Maryland_Counties_Socioeconomic_Characteristics” gives data on 41 socioeconomic characteristics of 24 counties in Maryland. These socioeconomic characteristics include education levels, employment status, proportion of different types of races and income levels of households etc.

DATA PROCESSING

DATASET 1

1) Identifying and removing garbage data

From crime data, we are removing 19 of the 30 variables that are not useful for analysis by using NULL function. These variables are for example, Incident ID and which offense code has been applied on that specific incident. While the type of streets in which crimes are occurring are important, the name of the streets do not add any value to our analysis. Moreover, the date and time at which a specific crime started are important, but the end time or dispatch time of report are not useful in regard to our analysis.

2) Treating null entries of different variables

The data initially does not contain NA values, rather it has empty cells. So, first of all, we use the “mutate” function to write NA in all the empty cells. Applying, “col sums” tells us that we have NA values in Crime Name 1, Crime Name 2, City and Street Type columns (286, 286, 1275 and 342 observations in each column respectively). We then used the “na.omit” function to omit all observations containing NA values. The reason behind using this technique was we were unable to impute data by taking mean or median considering the nature of variables; they are string data rather than numeric. Secondly, since we have over 3,00,000 observations in this dataset, removing these observations would not have a substantial impact on our analysis.

3) Removing observations for states other than Maryland

By using “table” function for state column, we observe the number of observations we have for each state. The output shows that out of total observations, 99.3% of observations are of Maryland state while the remaining 0.7% are of other states. Since the scope of our analysis is limited to Maryland state only, we remove the observations that are of other states.

4) Separating Place variable to extract useful information from each variable

The place variable contained hyphenated words where the first word gave overall classification of where the crime incident took place whereas the second word delves into more classification of place mentioned in the first word. We believe that only the first word of the hyphenated word is useful for our analysis because the second part contains too much classification. Having such deep classifications would make the data difficult to handle and hence, it would not be possible to perform analysis on all of them. For this purpose, we use “sapply” function in order to reduce data and provide the first word for each observation in another column.

5) Separating date and time from the original Start_Date_Time column

In Start_Date_Time column, we have the date on which the crime occurred as well as the time at which it occurred. Since each observation has 2 pieces of unique information, we need to extract useful information. In order to implement this, we use the “sapply” function to create separate columns of Date and Time.

6) Treating date column and creating day, month, and year variables

First of all, date type is changed from character type to date according to this format: "%m/%d/%Y" in order of month, day and year. Then we use “stfrtime” function to convert date and time objects to their string representation. We then create columns to have day of week, month and year in separate columns. The original Start_Date_Time column is removed because useful information in extracted from each observation and hence, presented in separate columns.

Note: The data cleaning of this dataset was efficient because we lost only 0.67% of data after the whole process which will have a negligible effect on the analysis.

DATASET 2

1) Removing junk data that is not necessary for analysis

In this dataset, only total number of population, murder, rape, robbery, aggravated assault, breaking and entering, larceny theft and motor vehicle theft were important with regards to our analysis, so we removed all other variables that calculated percentage changes in crimes. Moreover, there were no NA values seen in the data, so their treatment was not required.

2) Filtering and arranging data

We plan to merge this dataset with the 3rd dataset. Since the 3rd dataset is based on the year 2020, it only made sense that the observations of 2020 are filtered from this dataset as well. Next, we also arranged the dataset in alphabetical order of name of Jurisdictions.

DATASET 3

1) Identifying and removing garbage data

In this dataset, we figured out the most important socioeconomic factors of people living in each county that will affect crime rate in US such as education levels, employment status, racial identification, and gender etc. The rest of the variables are removed that might not have significant impact on occurrence of crime.

2) Merging Dataset 2 and 3

After arranging dataset 3 in alphabetical order of counties, it is merged with dataset 2 using "cbind" function.

3) Changing desired variables to numeric

Except for year and jurisdiction, all other variables have data in the form of numbers, so these variables are changed from character data type to numeric data type using "mutate" function. Moreover, one variable "Families in Poverty" is added to the dataset by computing it from "Percent Families in Poverty".

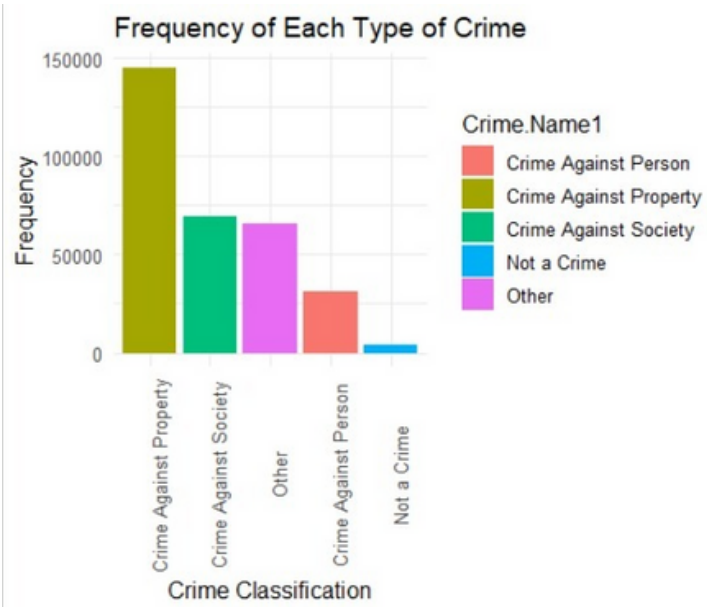
DATA EXPLORATION

For initial data exploratory analysis, we created different frequency charts using ggplot2 to get a gist of data. We observed what kind of crimes are most prevalent, which places are more vulnerable to different violent crimes or if these crimes are happening during the day or night. Furthermore what's the distribution of crimes across different hours in a day.

Moreover, we also used other specific techniques such as association rules & exploratory charts in clustering analysis to analyze trends in the data set to support our hypothesis.

GRAPHS

FIGURE 1

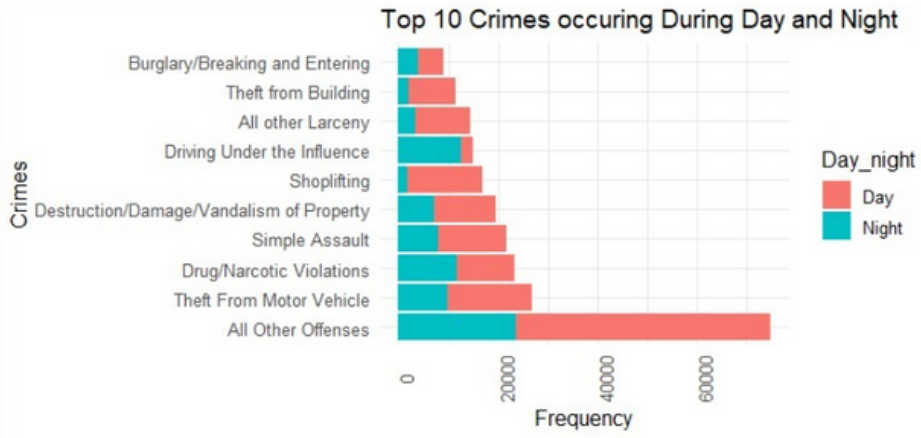


Q: Which type of crimes are the most prevalent in our data set?

Crimes against Property were one the most prevalent crimes in Maryland. These crimes against property include: burglary, larceny, theft, motor vehicle theft etc.

The Frequency of different types of crimes were analyzed with the help of these bar graphs. Moreover, these were further classified by the crime type. The most occurring crimes in terms of frequency were "all other offenses". Upon exploring, these offenses were mostly felonies labeled as "other" or "crimes against society" such as littering, police information and others which were not serious in nature. The most occurring crimes were those of theft from motor vehicles & Drug/Narcotic Violations. Simple Assaults (Crime against Person) were also prevalent. Overall, crime against property was most prevalent followed by Crime against Persons.

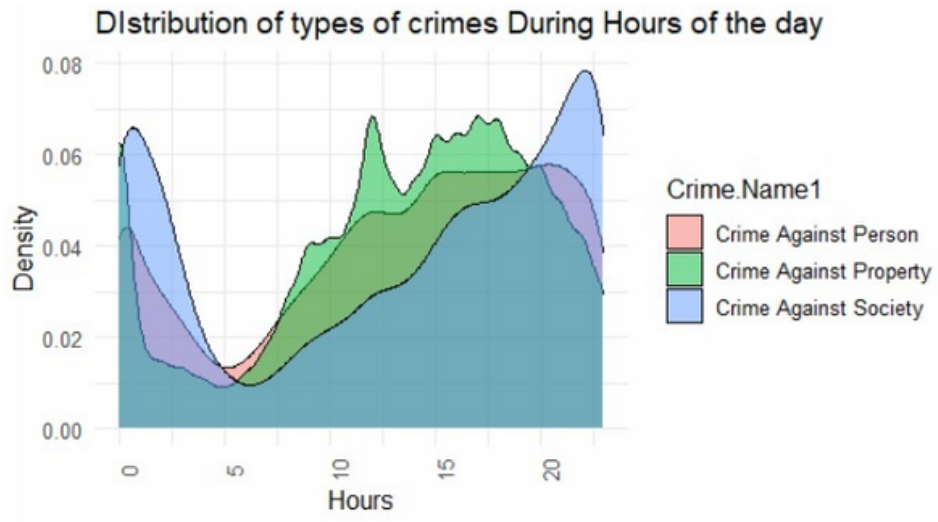
FIGURE 2



Q: Are violent crimes most prevalent during the day or night?

Crime against property & persons were frequent during the daytime rather than at night as shown in a graph in appendix (Exhibit 1-3).Whereas crimes against society were frequent during the night. We can classify these crimes further and can infer from Figure 2 that simple assaults were prevalent during daytime as a crime against a person. Moreover, Theft from Motor Vehicles is more prevalent during the day. It can be intuitively analyzed that the nature of these crimes are likely to happen during the day since there is a higher influx of traffic & people over the place of incident. Moreover, we can further analyze that driving under the influence is frequent during the night. Intuitively this makes sense as one cause of driving under influence is impaired driving or after drug influence. Considering that a driver's reaction depends on vision, and vision is limited at night, it is no surprise that the night driving accident rate is roughly three times that of daylight driving. Moreover, unnecessary use of drugs are prevalent during night so such crimes can occur.

FIGURE 3



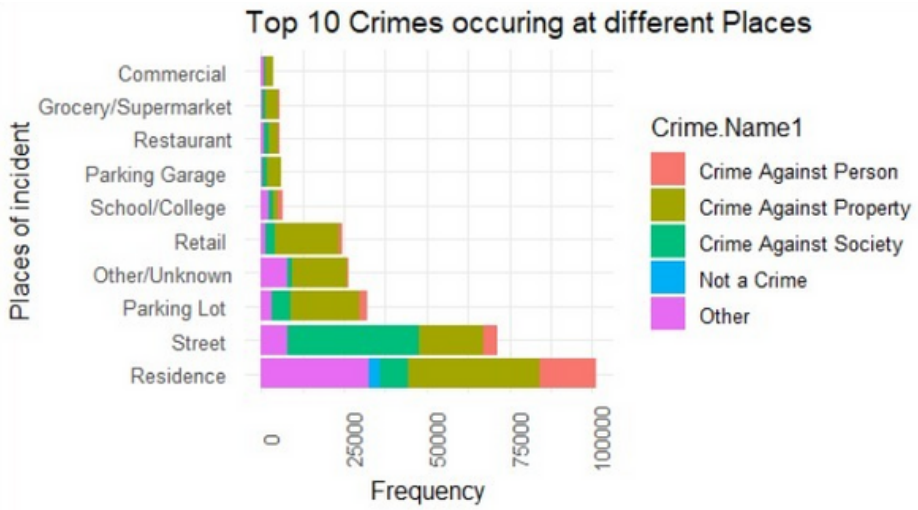
Q. Which is the distribution of different types of crimes that occur at different times of the day?

A density plot with hours on the x-axis, filled by crime type, helps us understand the frequency distribution of different types of crimes across 24 hours of the day.

Crimes against society usually occur in the late hours of the day in the range of 8:00 pm to 2:00 am peaking at 10:00 pm. That is because these crimes include drug/narcotic violations, driving under the influence, disorderly conduct, and trespass of real property. These crimes are associated with drinking and drug activities, usually done at night. The crimes against property occur primarily in the range of 11:00 am to 8:00 pm, peaking around noon. The crimes against property are motor vehicle thefts, vandalism of property, shoplifting, breaking and entering, etc.

Vehicles are parked outside during this time, houses are empty as people leave for their jobs, and shops are open, so these crimes occur in this time range. The crimes against person were most prevalent in the range of 12:00 pm to 10:00 pm, peaking at 9:00 pm. These crimes included simple and aggravated assault mainly.

FIGURE 4



Q. Which places are most vulnerable to crimes, at which time does the crime occur, and what is the type of those crimes?

From the figures (place-day_night, place-crimename1), we can identify places at which crimes occur, according to day/night, and classification of crimes as crimes against property, person and society. We can see that most crimes occur at Residences substantially greater than at any other place. The crimes that occurred during the daytime at residence are greater than those that occurred at Night. Also, most of the crimes at Residence are crimes against property owing to this type of crime. The second most occurred crime at Residence is a crime against a person.

We can also see that second to Residence, most crimes occurred on Streets. The proportion of crimes during the day and night for these crimes is similar. The types of crimes in the street are primarily crimes against society, followed by crimes against property. Parking lots experienced the third highest number of crimes. Daytime crimes in the Parking lot are more significant than crimes occurred at Night because most cars are parked during daytime and are vulnerable to theft. Crimes against property were most prevalent at parking lots which signifies the crimes against vehicles. We also checked concentration of different crimes in Montgomery. The maps are in appendix (Exhibit 11 and 12).

ASSOCIATION RULES (ARULES)

We wanted to see the general pattern of our dataset “crime clean”. We used variables such as “number of victims”, “Day or night time”, “Day of the week”, “Place”, “Street type” and “crime name 2”. Main reason for using these variables was to see the association of these variables. Primary reason for using these variables was that these are under our control to prevent the crimes.

We first created general rules without specifying any value of support and confidence to see the initial number of rules and whether they make any sense or not. Exhibit 4 shows the summary of this general rule. These rules did not give useful result because right hand side only contained variable of victim, and this only indicated that number of victims in each rule was 1 as shown in Figure 5. This meant that we cannot find any associations that can result in any place or time at which the crime will occur. We specifically wanted the time and place in Right Hand Side with high lift as well which suggests that the level of association between the ‘if part’ and ‘then part’ is higher than would be expected if these items were set independently. Although figure 6 shows that rules with high lift values had high confidence, number of rules made were very less (26), thus forcing us to try different values of confidence.

FIGURE 5

```
> inspect(sort(Grules,by="lift")[1:10])
     lhs                                                  rhs                  support    confidence coverage  lift     count
[1]  {Crime.Name2=All Other Offenses}                  => {Victims=Single} 0.2361085 0.9933307  0.2376937 1.011957 74322
[2]  {Place=Residence , Crime.Name2=All Other Offenses} => {Victims=Single} 0.1139275 0.9931046  0.1147186 1.011727 35862
[3]  {Day_night=Day, Crime.Name2=All Other Offenses}    => {Victims=Single} 0.1619200 0.9929284  0.1630731 1.011548 50969
[4]  {Place=Street , Day_night=Night}                  => {Victims=Single} 0.1177239 0.9921022  0.1186610 1.010706 37057
[5]  {Place=Street }                                   => {Victims=Single} 0.2247069 0.9886781  0.2272801 1.007218 70733
[6]  {Street.Type=AVE, Day_night=Day}                  => {Victims=Single} 0.1318226 0.9861917  0.1336684 1.004684 41495
[7]  {Street.Type=AVE}                                 => {Victims=Single} 0.2101061 0.9849584  0.2133147 1.003428 66137
[8]  {Place=Street , Day_night=Day}                    => {Victims=Single} 0.1069830 0.9849376  0.1086191 1.003407 33676
[9]  {Day=Friday, Day_night=Day}                       => {Victims=Single} 0.1003466 0.9842333  0.1019541 1.002689 31587
[10] {Day=Thursday}                                    => {Victims=Single} 0.1451844 0.9838328  0.1475702 1.002281 45701
```

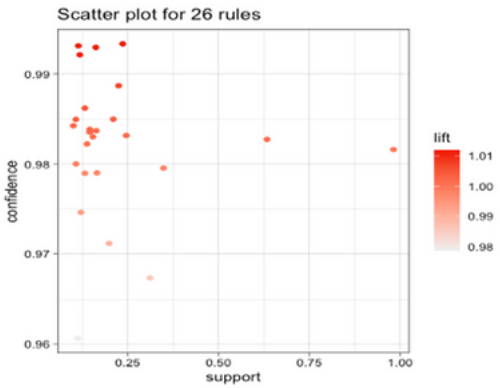


FIGURE 6

Then I set the conditions by setting support = 0.1 and confidence = 0.5 to see the new set of rules as shown in exhibit 5. We were keeping support = 0.1 because it helps us tell the minimum frequency for the occurrence of a particular association high. 42 rules were made with maximum confidence of 0.993, support ranging from 0.1 to 0.98 and lift ranging from 0.97 to 1.47 (exhibit 5). This didn't give satisfactory results because the plot had just two values with high lift, but its support and confidence were the lowest as shown in the plot of the particular rule as shown in exhibit 7. Moreover, rules with high lift had low confidence thus decreasing the chance of right hand side event to occur given the left hand side event is occurring. This showed that results were not much reliable as shown in exhibit 6.

After several hit and trials and keeping in view the range of support and confidence in first model, it was decided that support will be equal to 0.05 and confidence will be equal to 0.5. We found out first 15 association rules based on maximum lift. With these specifications, confidence for the occurrence of right-hand side event was high provided that left side events are occurring. Moreover, exhibit 9 showed that we got right hand sides to be either time or place, hence laying path for solving our upcoming hypothesis. Range of confidence turned out to be 0.52 to 1 and lift to be 0.809 to 3.0 (exhibit 8). Higher the lift, higher the chance for these events to occur together rather than separately.

In addition, rules made were also making a good sense, for example rue 2 shows that if crime is related to drugs and single victim is targeted, place of crime is more likely to be street. Similarly, rule 6 shows that if place of crime is retail and again single victim is targeted, time of crime is more likely to be day (exhibit 9a and 9b). Plot of general rules for these condition is shown in Figure 7. It showed that there were association rules with high lifts also having high confidence.

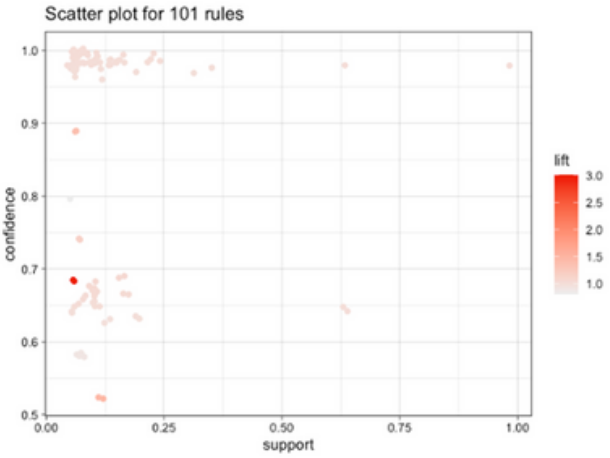


FIGURE 7

HYPOTHESIS

After exploring the data, we came up with 3 main hypotheses. Each hypothesis, however, may or may not lead to sub-hypothesis to strengthen our claim.

1) H-1: Features of crimes can be used to predict the type of crime in Montgomery County.

2) H-2: Other counties are associated with Montgomery County in terms of socioeconomic factors.

In order to test the above hypothesis, we have to test following sub-research questions as well:

- Which crime is the most prevalent across different counties?
- What is the trend of total crimes over the years? Can we identify top 5 counties with highest number of crimes?
- Can we make clusters of counties on the basis of socioeconomic conditions of each county?
- Within this cluster, which crimes are mostly correlated to each other?

3) H-3: Place and time of crime incidents can be used to check association of other characteristics of crime incidents for counties in the chosen cluster.

HYPOTHESIS 1: ANALYSIS

Features of crimes can be used to predict the type of crime in Montgomery County.

When we looked at our data initially, we planned on forming a machine learning algorithm that could be used to predict the nature of the crime (Whether it be against a person, society or property e.t.c). Given the nature of the problem, we knew we would be mostly dealing with discrete variables or variables that can be used as factors. We first employed a supervised classification method of Logistic Regression to model this problem.

Logistic Regression is a widely used machine learning classifier which is able to use both discrete and continuous variables. In layman's terms, it generally fits a model from the feature space provided by the data. It is a discriminative classifier as it learns which features can be used from the given data for classification. Since Logistic Regression also allows discrete features, we were able to use the place (Parking lot, building e.t.c), the street type (Avenue or Boulevard etc), Month, Year, Weekday, Hour, or a factor which told us whether it was currently day or night time.

We ran multi-class (or multinomial) logistic regression as we were predicting 5 different types of crime calls. After we ran the model, we found an accuracy of only 54.69% which is almost as good as flipping a coin.

FIGURE 8

Given the nature of the data, we want that the crime calls a police center receives to have a higher precision so even if the predicted value of the classifier is positive, we want the positivity to be as high as possible and we want the specificity to be as low as possible, meaning that when a specific crime does not take place we do not want a false alarm raised.

If we look at the detailed statistics in Exhibit 10, we see that only the crime against society class achieves a somewhat satisfactory result in our model, but generally none of these are satisfactory enough to be used. In essence, the model was as useful as flipping a coin, with discrepancies in each of the classes when looked at in detail.

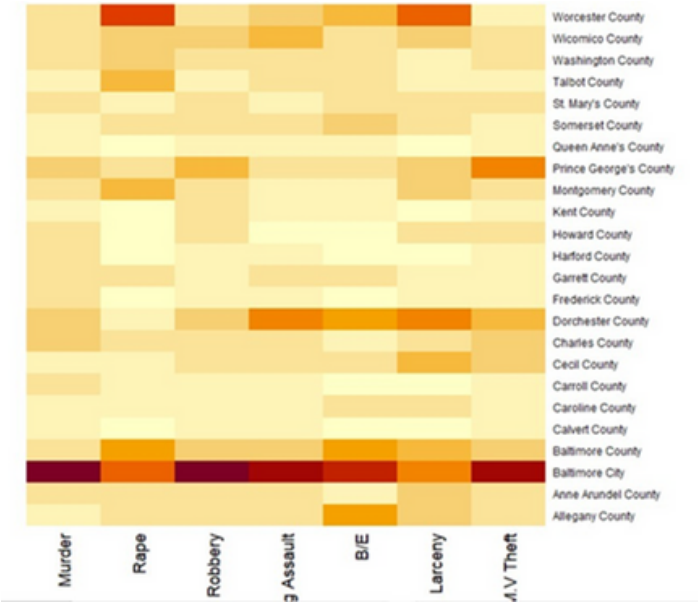
Logistic Regression is a sigmoid function based probability calculator so it may seem like a good machine learning model to base our prediction of our classes on. However, the nature of our problem is incredibly complex, and the variables we used due to the interest of dimensional complexity and unavailability of more useful variables we believe this was the best model we could produce. If we looked outside of course material and were not limited by our own personal machines, some algorithms like convoluted neural networks might have performed better.

HYPOTHESIS 2: ANALYSIS

Other counties are associated with Montgomery County in terms of socioeconomic factors.

Sub-hypothesis 1: Which crime is the most prevalent across different counties?

FIGURE 9



To check which crimes are most prevalent we constructed a heat map between frequent crimes such as theft, robbery and rape etc and different counties. The magnitude of occurrence of crimes is shown by the intensity of colors. A general observation could be drawn that Baltimore City is vulnerable to serious violent crimes such as murder & robbery are most frequent in, so their color is the most intense. However, other crimes such as Larceny & Theft are also occurring frequently. Moreover, we could analyze that Baltimore City, Baltimore County, Prince George and Dorchester County may be classified as Counties with the highest magnitude of crimes. If we draw a comparison of these with Montgomery County, we could further analyze that certain crimes such as rape & Larceny are most frequent than others. Such a trend could also be observed in other countries listed above.

Sub-Hypothesis 2: What is the trend of total crimes over the years? Can we identify the top 5 counties with the highest number of crimes?

We analyzed the time-series trend of the top 5 counties having greater numbers of crimes which are Baltimore City, Baltimore County, Prince George and Montgomery. Fluctuations in the trend have been observed over the years since 1970. This can be implied that several measures to control crime rates haven't been effective. That's why spikes after a couple of years are observed in the plot.

Another insight can be drawn that the trend line of these counties are similar in nature. For example, a sudden spike in the 1990s decade in Baltimore City is observed. Similarly, other counties also observed increased crime rates during the decade. It implies that these counties might have some common factors that causes disruptions in the crime rate trend.

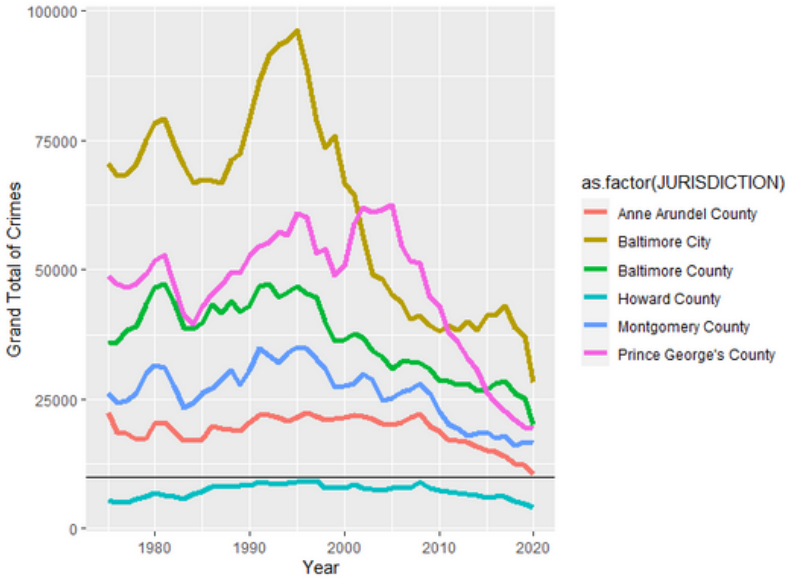


FIGURE 10

Sub-Hypothesis 3: Can we make clusters of counties on the basis of socioeconomic conditions of each county?

Multilinear Regression

Mathematically investigating further, we performed multilinear regression using the grand total of crimes as the predicted variable and various socioeconomic factors as predictors. We found that High School Students with no Diploma, people with an associates degree, people who attended some college but had no degree and employed males were significant impactors of the grand total of crimes variable. (Exhibit 13)

Recursive Feature Selection

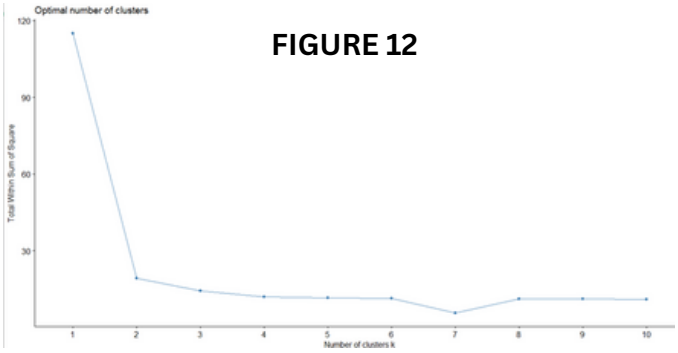
The next method we employed was of recursive feature selection in R, using a 10 fold cross validation technique (cross validation uses different portions of a dataset to train and test accordingly). In essence, it is a backward search feature selection method, where it starts with a subset of the entire predictors, keeps removing certain predictors and checks to see which predictors are the one which highly impact our predicted variable. According to this, the most impactful socioeconomic factors were Black people who live alone, families in poverty, people who have studied less than 9th grades and females. We felt that this method brought us predictors that make sense with our predicted variable. People, especially racially profiled black people, families in poverty, people with poor education are all more enticed or socioeconomically more likely to commit crime.

FIGURE 11

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables RMSE Rsquared  MAE RMSESD RsquaredSD MAESD Selected
         4 1526   0.9879 1200   2666    0.01768  1952        *
         8 1997   0.9178 1547   2956    0.22471  2147
        16 1963   0.9144 1538   3351    0.23393  2415
        20 1962   0.9082 1547   3269    0.25175  2361

The top 4 variables (out of 4):
   Black.Alone, Families.in.Poverty, Less.than.9th.Grade, Female
```

On the basis of these four variables, we decided to divide the counties into clusters to perform inter and intra cluster analysis. From the K-means clustering using total sum square method, we found that the best number of clusters is 2, and similarly using the elbow method we found 2 to be the optimal number of clusters. (Refer to Figure 12)

FIGURE 12



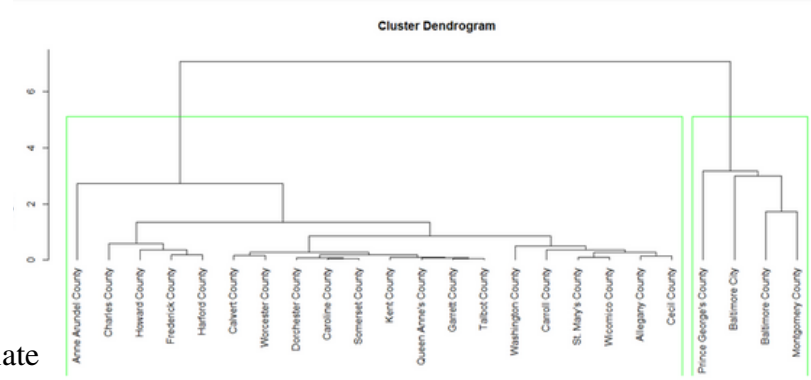


FIGURE 13 a

The given dendrogram in Figure 13 a gives an appropriate division of the given clusters, so we expect that any findings we have with respect to Montgomery County will be similar to Baltimore and Prince George's Colony. Clusters were tried using complete, average and single methods and every method recommended 2 clusters, Complete method was chosen because clusters made more sense depending on population and location This division in hindsight also makes sense because these are the largest counties in the state (in terms of Population).

By subtracting the maximum cluster mean from the minimum value of the cluster mean for each element, we can get an idea of the variables which contributed most to the cluster formation which in our case are variables Families.in.Poverty, and Less.than.9th.Grade. Using these 2 variables, we can plot the clusters. The plot is given in figure 13 b.

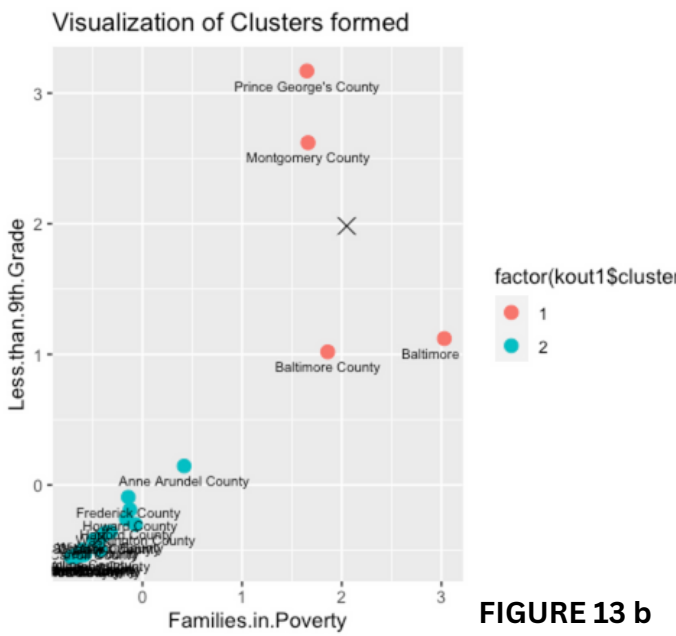


FIGURE 13 b

Note: *Our further analysis would be limited to the counties in the cluster of Montgomery i.e. Baltimore City, Baltimore County and Price George's County.*

Sub-Hypothesis 4: Within this cluster, which crimes are mostly correlated to each other?

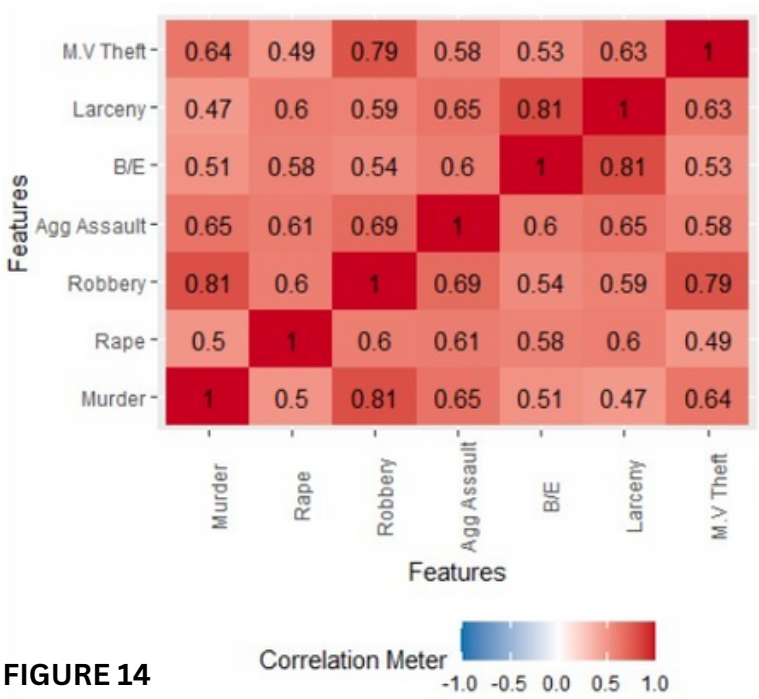


FIGURE 14

Correlation Plot highlights the most correlated crimes according to their value. We could infer that all these crimes are positively correlated to each other as all of these factors have a value greater than 0. Murder & Robbery are highly correlated to each other. It can be implied that Robbery and Murder are likely to occur under similar conditions. One can also infer that Robbery might lead to Murder in some circumstances. Rape is, however, positively correlated but value of 0.5 implies that rape and murder are not likely to happen together 50% of the times. Larceny (a felony) is least likely to result in murder as compared to other crimes since the correlation is less relatively. Another insight can be that violent crimes have high correlation with each other while felonies or non violent crimes are correlated with each other so we can classify these two crimes into two types.

HYPOTHESIS 3: ANALYSIS

Place and time of crime incidents can be used to check association of other characteristics of crime incidents for counties in the chosen cluster.

Place and time of the crime are most important aspects of a crime to determine the action plan for preventing any future crime. We wanted to show different factors which results in a crime at any particular place or time. We took into account the item frequency plot of top 10 occurring items and general rules made in data exploration to form subrules. From general rules we decided to keep time of the day, i.e day or night, and place, i.e street, on right hand side. Looking at the item frequency plot, we made another subrule by keeping single victim on right hand side.

Data Preparation:

A-rules only work on categorical variables. Hence our all variables were converted into factor variables. We used variables such as “number of victims”, “Day or night time”, “Day of the week”, “Place”, “Street type” and “crime name 2”. Number of victims were made into three levels namely single, pair and group. Day or night time was made by using time of the day and classifying different ranges into the two levels. The association rules were made by keeping support and confidence constant that was 0.05 and 0.5 respectively. Consistency was maintained to ensure that there is a high accuracy of analysis given to the security teams for prevention of crimes.

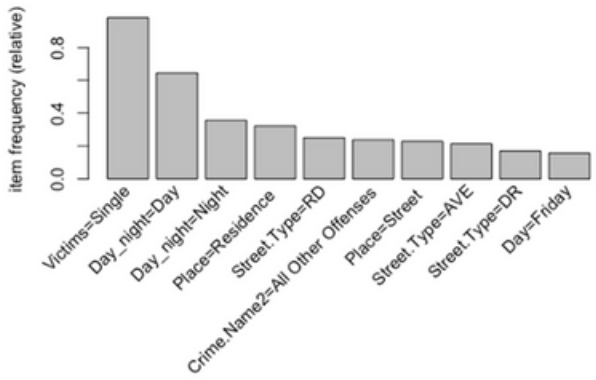


FIGURE 15

We kept one of the places constant in right hand side and it was street. This was the only place considered because rules with street had highest lift, thus showing that crime occurring on streets are most likely to be predicted if other events are occurring.

Street:

```
> inspect(sort(Street_rule,by="lift"))
    lhs                                                rhs             support   confidence coverage   lift     count
[1] {Crime.Name2=Drug/Narcotic Violations}              => {Place=Street } 0.0505434 0.6820715  0.07410278 3.001017 15910
[2] {Crime.Name2=Drug/Narcotic Violations, Victims=Single} => {Place=Street } 0.0505434 0.6820715  0.07410278 3.001017 15910
```

If we see towards the prediction of crime occurring on street, then knowing that crime occurring is related to narcotics/drugs will increase the chances that security forces catch the criminal on a street by 3 folds [lift = 3.00]. If we see towards the confidence, then crime of narcotics is most predictive of crime being occurring on a street [68.2%].
If we add another aspect on left hand side of the preceding association rule related to number of victims being single, chances that security forces catch the criminal on a street by same 3 folds [lift = 3.00]. Similarly, Confidence of predicting crime being occurring on a street remains the same as well [68.2%].
Since the number of crimes was introduced on the left hand side, We decided to see some rules in which the crime number was seen on right hand side.

```
> inspect(sort(victim_single,by="lift")[1:10])
     lhs                                                  rhs                 support confidence   coverage     lift count
[1]  {Crime.Name2=Shoplifting}                         => {Victims=Single} 0.05323100  1.0000000 0.05323100 1.018752 16756
[2]  {Crime.Name2=Destruction/Damage/Vandalism of Property} => {Victims=Single} 0.06205306  1.0000000 0.06205306 1.018752 19533
[3]  {Crime.Name2=Drug/Narcotic Violations}             => {Victims=Single} 0.07410278  1.0000000 0.07410278 1.018752 23326
[4]  {Crime.Name2=Theft From Motor Vehicle}             => {Victims=Single} 0.08553938  1.0000000 0.08553938 1.018752 26926
[5]  {Place=Street ,
      Crime.Name2=Drug/Narcotic Violations}             => {Victims=Single} 0.05054340  1.0000000 0.05054340 1.018752 15910
[6]  {Day_night=Day,
      Crime.Name2=Theft From Motor Vehicle}            => {Victims=Single} 0.05459703  1.0000000 0.05459703 1.018752 17186
[7]  {Place=Other/Unknown,
      Day_night=Day}                                   => {Victims=Single} 0.06188469  0.9977464 0.06202447 1.016456 19480
[8]  {Place=Other/Unknown}                             => {Victims=Single} 0.08349032  0.9975707 0.08369364 1.016277 26281
[9]  {Day_night=Night,
      Crime.Name2=All Other Offenses}                  => {Victims=Single} 0.07418856  0.9942101 0.07462061 1.012853 23353
```

In these associations, we looked at rule number 5 in which left hand side had street place. If we see towards the prediction of crimes resulting in one victim, then knowing that crime of Drugs/Narcotics occurring on street will increase the chances that security forces catch the criminal by 1 folds [1.02]. If we see towards the confidence then crime of drugs occurring on street is most predictive of single victim being affected [100%].

Day:

For this, we kept the time of day constant in the right hand side. These both times had high lift ratios in our general rules, thus showing that there are high chances of events on left hand side to occur with either time of the day. These both events are also present in 2nd and third position of our itemfrequency plot.

```
> inspect(sort(Day_rule,by="lift"))
     lhs                                rhs               support confidence   coverage     lift  count
[1]  {Place=Retail }                 => {Day_night=Day} 0.06927718  0.8846294 0.07831209 1.372515  21807
[2]  {Place=Retail ,
      Victims=Single}                => {Day_night=Day} 0.06874664  0.8845289 0.07772119 1.372359  21640
[3]  {Place=Other/Unknown,
      Victims=Single}                => {Day_night=Day} 0.06188469  0.7412199 0.08349032 1.150013  19480
[4]  {Place=Other/Unknown}           => {Day_night=Day} 0.06202447  0.7410894 0.08369364 1.149811  19524
[5]  {Crime.Name2=All Other Offenses} => {Day_night=Day} 0.16307314  0.6860641 0.23769375 1.064438  51332
[6]  {Crime.Name2=All Other Offenses,
      Victims=Single}                => {Day_night=Day} 0.16191995  0.6857862 0.23610851 1.064007  50969
[7]  {Day=Tuesday,
      Victims=Single}                => {Day_night=Day} 0.09836425  0.6797884 0.14469834 1.054701  30963
[8]  {Day=Tuesday}                   => {Day_night=Day} 0.09993360  0.6792701 0.14711909 1.053897  31457
[9]  {Day=Monday,
      Victims=Single}                => {Day_night=Day} 0.09246487  0.6712174 0.13775697 1.041403  29106
[10] {Day=Monday}                    => {Day_night=Day} 0.09403423  0.6704720 0.14025078 1.040247  29600
```

If we see towards the prediction of crime occurring at day time, then knowing that crime occurring in place of retail will increase the chances that security forces catch the criminal at day time by 1.4 folds [lift = 1.37]. If we see confidence, then crime occuring in retail place is most predictive of crime occurring at day time [88.5%]. Lift and confidence remains approximately same if on left hand side we add another event that victim of any crime will be 1.
We had an additional feature of day in rule 7. We see that if crime occurs on Friday and single victim is targeted, chances for security forces to catch criminals at day time will increase by 1.1 folds [lift = 1.05]. This means that more security can be deployed on friday in Montgomery at day time to catch the criminal. If we see at the confidence, then crime occurring on friday is most predictive of crime occuring at day time [68%].

Night:

```
> inspect(sort(Night_rule,by="lift"))
    lhs                                 rhs                 support   confidence coverage  lift     count
[1] {Place=Street , Victims=Single} => {Day_night=Night} 0.1177239 0.5238997  0.2247069 1.473829 37057
[2] {Place=Street }                 => {Day_night=Night} 0.1186610 0.5220916  0.2272801 1.468743 37352
```

If we see towards the prediction of crime occurring at night time, then knowing that crime occurring in place of Street and single victim being targetted will increase the chances that security forces catch the criminal at night time by 1.4 folds [lift = 1.47]. If we see confidence, then crime occuring in street place and single victime is most predictive of crime occurring at night time [52.4%].
We can conclude from this sub-hypothesis that if we want to catch criminals, then we will have to go to different places for different times of the day. For day time, we will have to deploy security forces in retail places while at night time security forces will be deployed on streets. There are different types of street in our dataset. Item frequency plot showed that RD (road) type of street is most occurring event. Hence, most focus must be on this street at night.

ASSUMPTIONS & LIMITATIONS

ASSUMPTIONS

The data we used was for latest 2020, so we are assuming that our analysis would also extend to current timeline. Our dataset is only for the crimes which are reported to relevant authorities, however, there might also be unreported crimes, including them give us slightly different results. We removed some variables in cleaning which seemed irrelevant to us assuming that these would not effect our analysis such as Police District Name, and address.

LIMITATIONS

The dataset used by us did not include observations from most recent years (2021-22). Also, we were limited only to R modules we were taught in class. We were limited in regards to data mining and machine learning techniques within the constraints of the course content. Since we were on our personal computer, due to the large size of some datasets we faced problems in regards to processing time. Moreover, the nature of our observations, which was mostly categorical limited the extent of analysis. There were no appropriate datasets related to US criminal agencies in order to analyze their administration and enforceability abilities to prevent crime which would have helped us to better analyse the questions at hand. In case of a major crime (like from a mass shooting in a specific county), the results might get skewed, since there were not enough representations of these crimes in our data set.

CONCLUSION & RECOMMENDATION

After cleaning of data, we made visualizations of data to get a proper understanding of data. That included different distributions from which we can get an idea about frequency of different types of crimes, along with the time of these occurences. When we answered our **Hypothesis 1** of whether certain variables could be used to classify the crime that occured, we found that our logistic regression model turned out to be pretty weak in terms of its accuracy (only 54.69%). For further research, we recommend that if we analyse each class of crime we might find that certain indicators are better for said specific classes. Another possible addition is to use some sort of other data available to the police which can be more helpful for prediction of the type of cime. Other than this, due to limitations of high dimensionality, we had to specificially pick some indicators for our prediction model which may have not been highly correlated with the type of crime that occured. Some sort of correlation analysis might help in regards feature selection for this.

Next, through our **hypothesis 2**, based on the results of logistic regression, we have established comparison of Montgomery County to other counties. For this, after running a regression model, we did not believe that the results were as accurate as they should be so we approached a random forest feature selection, given our dataset we decided to use this method (even though other uni and bivariate correlation methods were available). On the basis of this, we found four variables, Black.Alone, Families in Poverty, Less than 9th Grade education and Female, and based on these socioeconomic indicators we decided to proceed with clustering to perform inter and intra cluster analysis. This cluster analysis established that Montgomery County is similar to 3 other counties in terms of socioeconomic conditions i.e. Baltimore City, Baltimore County and Prince George's County.

In our hypothesis 3, we focus on this cluster specifically and generate A-rules to check associations of certain characteristics of crime incidents so that we can recommend the time and place to security forces to deploy more security at that time. We were able to identify the factors which lead to crime occurrence in street and retail places during day or night time. We used these events specifically because of their occurrence in top 10 item frequency plot. Rules having these events had high lift and confidence as well. Hence, we recommend that security forces will need to deploy security forces in retail places at day time and on streets at night time to increase the chance of catching the criminals. We also recommend the security forces to pay a careful heed to street types road and avenue at night time when crime occurring on streets has high lift and confidence. Moreover, it was seen that if crime occurs on Friday and single victim is targeted, chances for security forces to catch criminals at day time will increase.

APPENDIX A

EXHIBIT 1

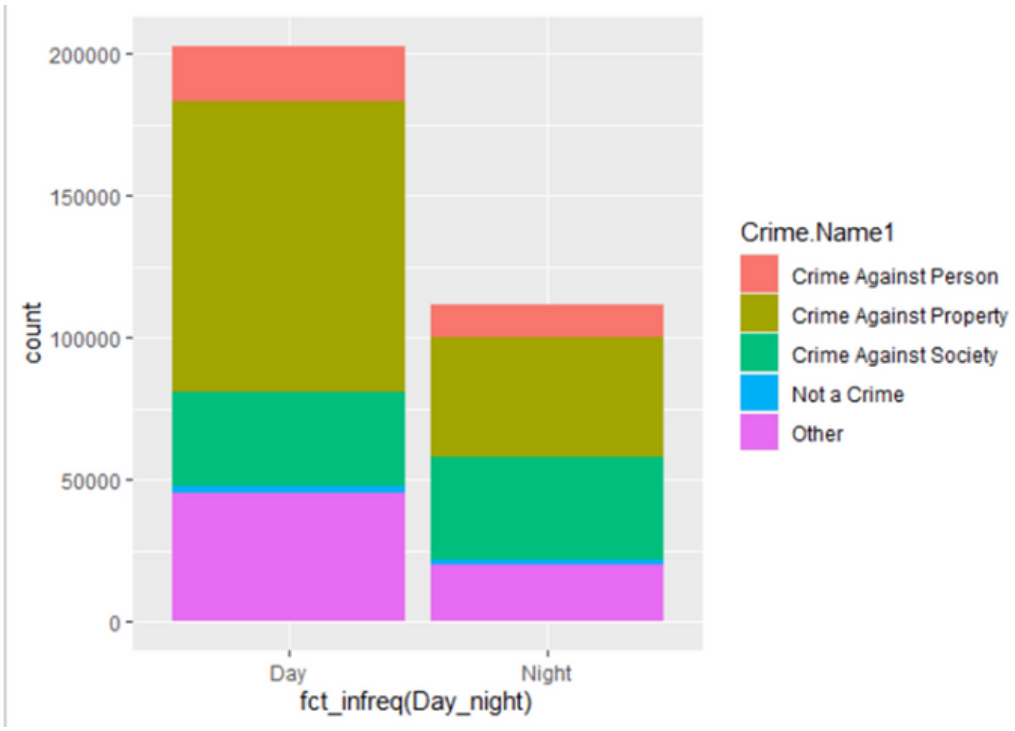


EXHIBIT 2

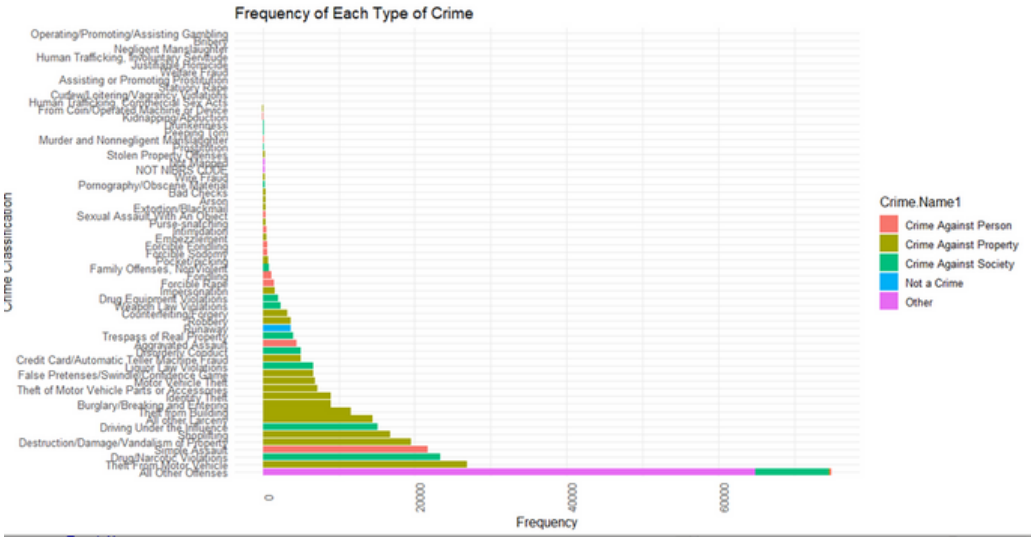


EXHIBIT 3

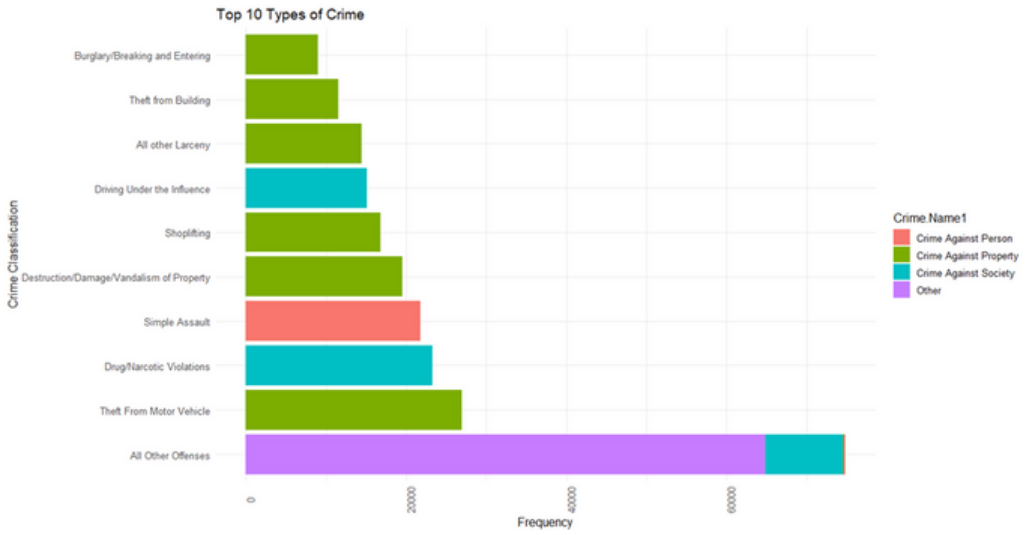


EXHIBIT 4

```
> summary(Grules)
set of 26 rules

rule length distribution (lhs + rhs):sizes
 1  2  3
 1 15 10

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   2.000   2.346   3.000   3.000

summary of quality measures:
    support         confidence        coverage            lift           count
 Min.   :0.1003   Min.   :0.9606   Min.   :0.1020   Min.   :0.9786   Min.   : 31587
 1st Qu.:0.1238   1st Qu.:0.9797   1st Qu.:0.1266   1st Qu.:0.9980   1st Qu.: 38974
 Median :0.1505   Median :0.9833   Median :0.1530   Median :1.0018   Median : 47366
 Mean   :0.2174   Mean   :0.9823   Mean   :0.2214   Mean   :1.0007   Mean   : 68431
 3rd Qu.:0.2211   3rd Qu.:0.9850   3rd Qu.:0.2238   3rd Qu.:1.0034   3rd Qu.: 69584
 Max.   :0.9816   Max.   :0.9933   Max.   :1.0000   Max.   :1.0120   Max.   :308985

mining info:
        data ntransactions support confidence                          call
 crime_arules        314779     0.1        0.8 apriori(data = crime_arules)
```

EXHIBIT 5

```
> summary(Grules)
set of 42 rules

rule length distribution (lhs + rhs):sizes
 1  2  3
 2 23 17

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   2.000   2.357   3.000   3.000

summary of quality measures:
    support         confidence        coverage            lift           count
 Min.   :0.1003   Min.   :0.5221   Min.   :0.1020   Min.   :0.9722   Min.   : 31587
 1st Qu.:0.1180   1st Qu.:0.6509   1st Qu.:0.1472   1st Qu.:0.9980   1st Qu.: 37131
 Median :0.1505   Median :0.9790   Median :0.1866   Median :1.0021   Median : 47366
 Mean   :0.2121   Mean   :0.8501   Mean   :0.2586   Mean   :1.0264   Mean   : 66766
 3rd Qu.:0.2031   3rd Qu.:0.9837   3rd Qu.:0.2438   3rd Qu.:1.0104   3rd Qu.: 63922
 Max.   :0.9816   Max.   :0.9933   Max.   :1.0000   Max.   :1.4738   Max.   :308985

mining info:
        data ntransactions support confidence
 crime_arules        314779     0.1        0.5
                                                                                call
 apriori(data = crime_arules, parameter = list(support = 0.1, conf = 0.5))
```

EXHIBIT 6

```
> inspect(sort(Grules,by="lift")[1:10])
     lhs                                                rhs                 support    confidence coverage  lift     count
[1]  {Place=Street , Victims=Single}                 => {Day_night=Night} 0.1177239 0.5238997  0.2247069 1.473829 37057
[2]  {Place=Street }                                 => {Day_night=Night} 0.1186610 0.5220916  0.2272801 1.468743 37352
[3]  {Crime.Name2=All Other Offenses}                => {Day_night=Day}   0.1630731 0.6860641  0.2376937 1.064438 51332
[4]  {Crime.Name2=All Other Offenses, Victims=Single} => {Day_night=Day}   0.1619200 0.6857862  0.2361085 1.064007 50969
[5]  {Street.Type=RD, Victims=Single}                => {Day_night=Day}   0.1631875 0.6638322  0.2458264 1.029945 51368
[6]  {Street.Type=RD}                                => {Day_night=Day}   0.1658942 0.6634734  0.2500389 1.029389 52220
[7]  {Crime.Name2=All Other Offenses}                => {Victims=Single}  0.2361085 0.9933307  0.2376937 1.011957 74322
[8]  {Place=Residence , Crime.Name2=All Other Offenses} => {Victims=Single} 0.1139275 0.9931046 0.1147186 1.011727 35862
[9]  {Day_night=Day, Crime.Name2=All Other Offenses} => {Victims=Single}  0.1619200 0.9929284  0.1630731 1.011548 50969
[10] {Day=Friday, Victims=Single}                    => {Day_night=Day}   0.1003466 0.6514798  0.1540287 1.010780 31587
```

EXHIBIT 7

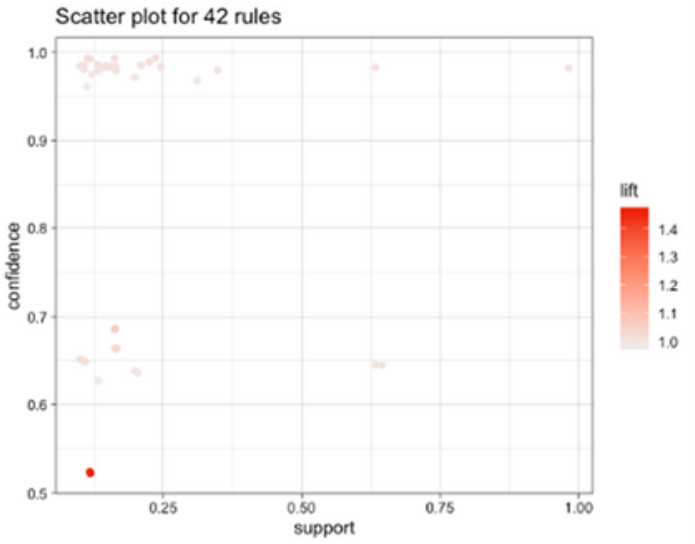


EXHIBIT 8

```
> summary(Grules)
set of 101 rules

rule length distribution (lhs + rhs):sizes
 1  2  3  4
 2 45 52  2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   2.535   3.000   4.000

summary of quality measures:
    support         confidence        coverage            lift            count
 Min.   :0.05054   Min.   :0.5221   Min.   :0.05054   Min.   :0.8087   Min.   : 15910
 1st Qu.:0.06513   1st Qu.:0.6667   1st Qu.:0.07588   1st Qu.:0.9990   1st Qu.: 20501
 Median :0.09680   Median :0.9772   Median :0.11472   Median :1.0041   Median : 30471
 Mean   :0.13030   Mean   :0.8541   Mean   :0.15912   Mean   :1.0608   Mean   : 41014
 3rd Qu.:0.13367   3rd Qu.:0.9843   3rd Qu.:0.16307   3rd Qu.:1.0188   3rd Qu.: 42076
 Max.   :0.98159   Max.   :1.0000   Max.   :1.00000   Max.   :3.0010   Max.   :308985

mining info:
         data ntransactions support confidence
 crime_arules        314779    0.05        0.5
                                                                                call
 apriori(data = crime_arules, parameter = list(support = 0.05, conf = 0.5))
```

EXHIBIT 9a

```
> inspect(sort(Grules,by="lift")[1:15])
     lhs                                                    rhs                    support    confidence coverage   lift     count
[1]  {Crime.Name2=Drug/Narcotic Violations}                  => {Place=Street }        0.05054340 0.6820715  0.07410278 3.001017 15910
[2]  {Crime.Name2=Drug/Narcotic Violations, Victims=Single} => {Place=Street }        0.05054340 0.6820715  0.07410278 3.001017 15910
[3]  {Place=Street , Victims=Single}                        => {Day_night=Night}      0.11772386 0.5238997  0.22470686 1.473829 37057
[4]  {Place=Street }                                        => {Day_night=Night}      0.11866103 0.5220916  0.22728009 1.468743 37352
[5]  {Place=Retail }                                        => {Day_night=Day}        0.06927718 0.8846294  0.07831209 1.372515 21807
[6]  {Place=Retail , Victims=Single}                        => {Day_night=Day}        0.06874664 0.8845289  0.07772119 1.372359 21640
[7]  {Place=Other/Unknown, Victims=Single}                  => {Day_night=Day}        0.06188469 0.7412199  0.08349032 1.150013 19480
[8]  {Place=Other/Unknown}                                  => {Day_night=Day}        0.06202447 0.7410894  0.08369364 1.149811 19524
[9]  {Crime.Name2=All Other Offenses}                       => {Day_night=Day}        0.16307314 0.6860641  0.23769375 1.064438 51332
[10] {Crime.Name2=All Other Offenses, Victims=Single}       => {Day_night=Day}        0.16191995 0.6857862  0.23610851 1.064007 50969
[11] {Day=Tuesday, Victims=Single}                          => {Day_night=Day}        0.09836425 0.6797884  0.14469834 1.054701 30963
[12] {Day=Tuesday}                                          => {Day_night=Day}        0.09993360 0.6792701  0.14711909 1.053897 31457
[13] {Day=Monday, Victims=Single}                           => {Day_night=Day}        0.09246487 0.6712174  0.13775697 1.041403 29106
[14] {Day=Monday}                                           => {Day_night=Day}        0.09403423 0.6704720  0.14025078 1.040247 29600
[15] {Day=Wednesday, Victims=Single}                        => {Day_night=Day}        0.09829118 0.6690019  0.14692213 1.037966 30940
```

EXHIBIT 9b

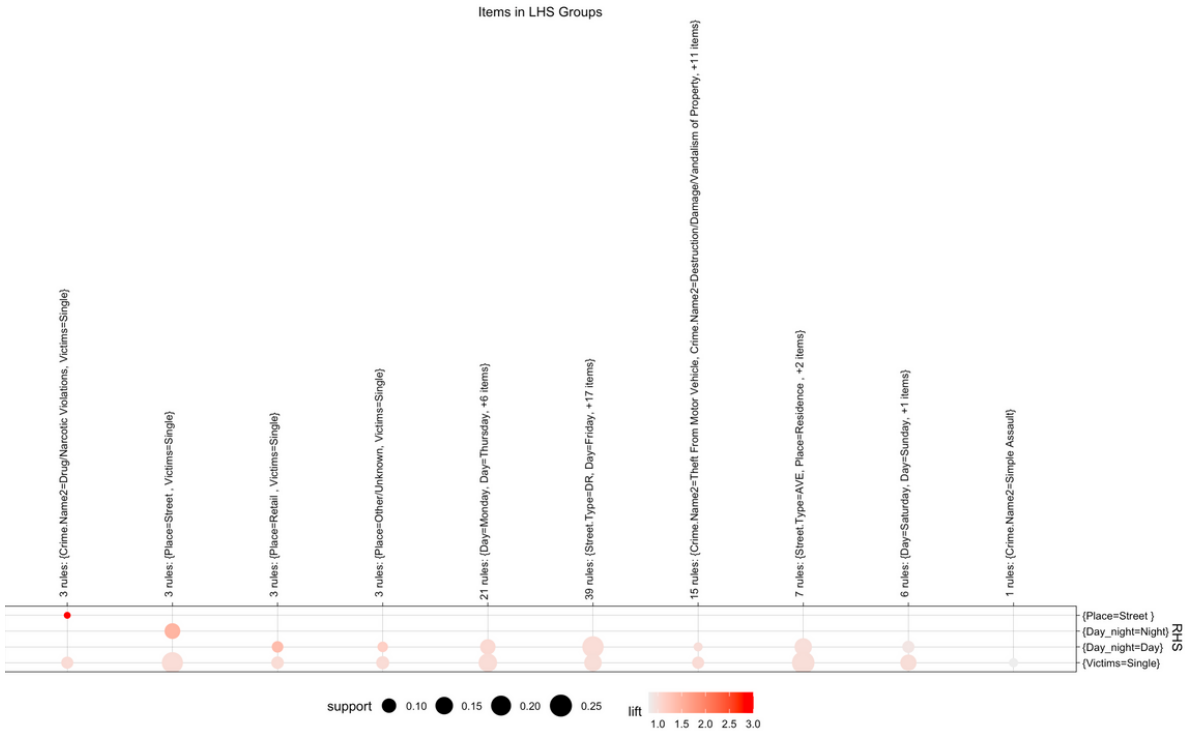


EXHIBIT 10

	Class: Crime Against Person	Class: Crime Against Property
Sensitivity	0.4035088	0.5546
Specificity	0.9012572	0.7515
Pos Pred Value	0.0073588	0.8331
Neg Pred Value	0.9988008	0.4300
Prevalence	0.0018108	0.6910
Detection Rate	0.0007307	0.3832
Detection Prevalence	0.0992947	0.4600
Balanced Accuracy	0.6523830	0.6530
	Class: Crime Against Society	Class: Not a Crime Class: Other
Sensitivity	0.5880	NA 0.39853
Specificity	0.8787	0.98831 0.81179
Pos Pred Value	0.5689	NA 0.17864
Neg Pred Value	0.8868	NA 0.92928
Prevalence	0.2140	0.00000 0.09315
Detection Rate	0.1258	0.00000 0.03712
Detection Prevalence	0.2212	0.01169 0.20780
Balanced Accuracy	0.7334	NA 0.60516

EXHIBIT 11

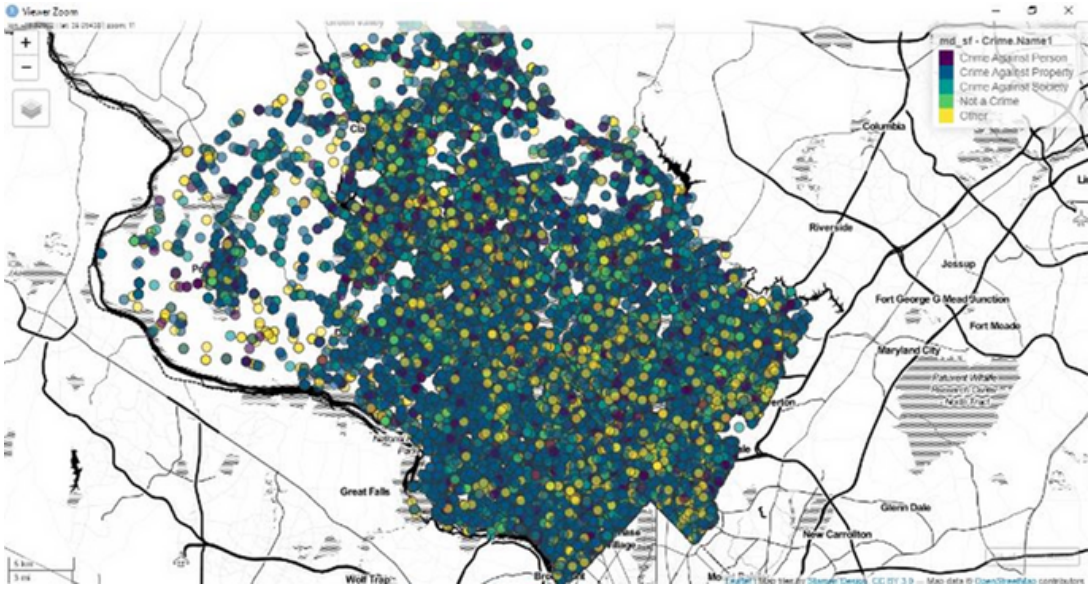


EXHIBIT 12

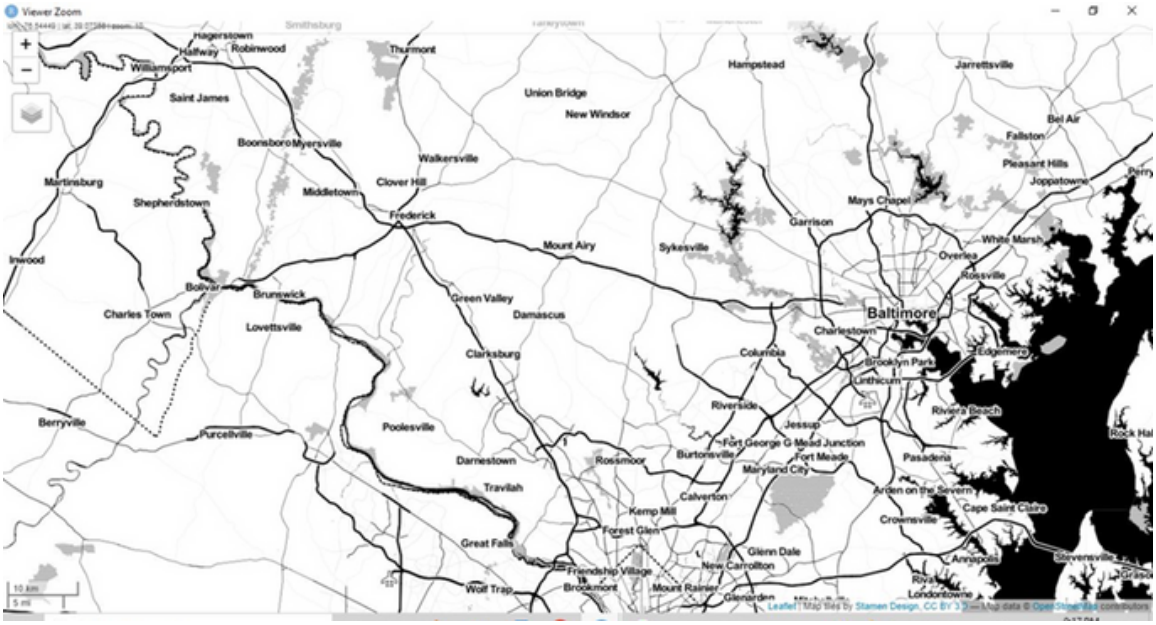


EXHIBIT 13

```
                          Estimate   Std. Error    t value   Pr(>|t|)
High.School.no.Diploma   -1.23506129  0.29783237 -4.1468337 0.01429697
Associates.degree        -1.39349965  0.34994000 -3.9821102 0.01637227
Some.College.no.degree    0.32654623  0.09289949  3.5150487 0.02455960
Employed                 -0.39194292  0.11405393 -3.4364702 0.02637883
Male                     -0.22393565  0.06609077 -3.3883046 0.02757311
Bachelor.s.degree         0.26817283  0.08009603  3.3481412 0.02861824
Families.in.Poverty       2.09861956  0.65311451  3.2132490 0.03248834
High.School.Diploma       0.25813355  0.09065384  2.8474642 0.04651492
White.Alone               0.47510988  0.18169756  2.6148391 0.05911873
Black.Alone               0.46402908  0.18449525  2.5151275 0.06569707
Median.Household.Income.... -0.02147742  0.01072159 -2.0031928 0.11569406
Hispanic.or.Latino..of.any.race.  0.37403706  0.18724955  1.9975325 0.11644418
Voting.Age.Population    -0.42687011  0.21484024 -1.9869188 0.11786512
Graduate.or.Professional  0.31943701  0.17313256  1.8450430 0.13878454
(Intercept)           1253.98842981 743.65490991  1.6862505 0.16702572
POPULATION                0.14851885  0.09710853  1.5294110 0.20089823
Less.than.9th.Grade      -0.66075050  0.46734232 -1.4138469 0.23029945
Unemployed               -0.19144033  0.24643502 -0.7768390 0.48063272
Asian.Alone              -0.06062372  0.13284045 -0.4563649 0.67180833
Families                 -0.07233050  0.16710846 -0.4328357 0.68745029
```

APPENDIX B

CODEBOOK

Incident ID	Specific identification number of the crime that occurRed
Offense Code	A code which identifies what type of crime occurRed
CR Number	Crimes associated with the criminal division's unique ID
Dispatch Date/Time	Time of police dispatch
NIBRS Code	Code according to the US National Incident Based Reporting System
Victims	Number of Victims
Crime Name 1	Divisions of Crime based on Property, society, person, Other,
Crime Name 2	Further Division of the crime based on crime 1, such as a crime against property may be considered as Robbery
Crime Name 3	Further division from Crime Name 2, such as Robbery involving guns
Police District Name	Police District from which stations are assigned to specific crimes
Block Address	Address of the crime
City	City the crime took place in
State	State the crime took place in
ZIP Code	Geographical location identification number
Agency	Police agency responsible for the crime given in its jurisdiction
Place	Place of the crime
Sector	Sector the Crime took place
PRA	Unique code assigned to the paperwork reduction act so police have to deal with an appropriate amount of specific paperwork
Address Number, Street Prefix, Street Name, Street Suffix, Street Type	Details about the address
Start Date Time, End Date Time	Time it took

Latitude	Latitude of the location
Longitude	Longitude of the location
Police District Number	The police District under which the Crime Occurred
Location	Location Based on Latitude and Longitude

REFERENCES

Farrell, G. (1995). Preventing Repeat Victimization. Crime and Justice, 19, pp.469–534. doi:10.1086/449236.

MCSO (n.d.). Montgomery County Sheriff's Office. [online] www.mctxsheriff.org. Available at: https://www.mctxsheriff.org/about_the_department/index.php [Accessed 9 Dec. 2022].

Schwabe, W., Davis, L.M., Jackson, B.A., Science And Technology Policy Institute (Rand Corporation, Rand Corporation and United States. Office Of Science And Technology Policy (2001). Challenges and choices for crime-fighting technology : federal support and local law enforcement. Santa Monica, Ca: Rand.