IST 707 DATA ANALYTICS FINAL PROJECT

Boston Crime Data Analysis

Team Member:
Yucan Dai
Xuanran Ji

APRIL 30, 2020

SYRACUSE UNIVERSITY

Contents

Introduction

Since 1930, the FBI has been collecting data on the types, amounts, and impact of crime in the United States through the Uniform Crime Reporting Program. This program is the most reliable source of crime statistics for law enforcement administration, operation, and management. It is also used by researchers, politicians, and criminal justice professionals to gain a deeper understanding of crime and society.

Objective and Problems

In reality, due to limited financial and personnel resources, police officers cannot patrol in every community and street at any time. So, by analyzing the crime data, we want to help criminal justice professionals anticipate increased risk of crime and further focus on a specific area and allow police resources to be used more effectively. For example, helping determine which communities will receive criminal justice grants and where or when police officers will patrol based on areas or times that see higher crime.

Budgets are a crucial part of law enforcement at the local, state, and federal levels. Making sure enough money is allocated to the right locations, and programs can make a big difference in keeping communities safe. Researching crime data will help create more accurate budgets. Analyzing crime patterns to make prevention beforehand and better allocate police resources. In this project, we are aiming to answer the following questions.

1. Area and type of crimes that the Boston police department should focus on.
2. Factors contribute to the rate of crime.

Data Mining Task

In crime terminology, a cluster is a group of crimes in a geographical region or a hot spot of crime. In data mining terminology, a cluster is a group of similar data points which can be a possible crime pattern. Thus, appropriate clusters or a subset of the cluster will have a one-to-one correspondence to crime patterns. Thus, we would like to use clustering algorithms to identify groups of records that are similar between themselves but different from the rest of the data. Some of these clusters will be useful for identifying a crime committed by one of the same group of

suspects. These clusters will then be presented to the detectives to drill down using their domain expertise.

We will also use the association rules method to analyze crime patterns and to reduce the crimes as much as possible. We want to generate rules from the crime dataset based on the frequent occurrence of patterns to help the decision-makers of our security society to make a prevention action. Furthermore, since shooting is a severe crime, we'd like to predict as precisely as possible if there will be a shooting case in the future. We will create different prediction models to predict the chance of a shooting case, and compare their accuracy results, selecting the most accurate one.

Data Description

Link to the data set. (https://www.kaggle.com/AnalyzeBoston/crimes-in-boston)

In this analysis, we use crime.csv only. The data set is provided by the Boston Police Department to record the detail of police responded incidents. It contains 319073 observations and 17 variables that record incidents that happened in Boston from 2015 to 2018. The data set includes information such as the date of incidents, shooting cases or not, location of incidents, and so on.

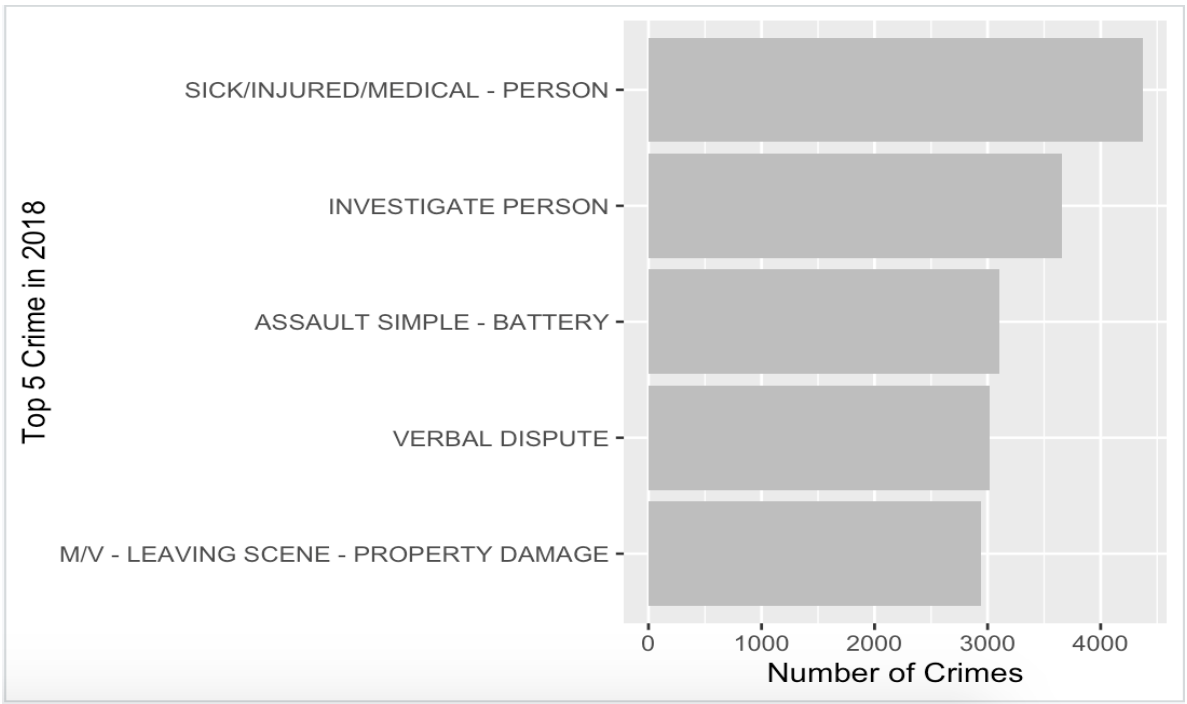
```
> str(crime_raw)
'data.frame':   319073 obs. of  17 variables:
 $ INCIDENT_NUMBER    : Factor w/ 282517 levels "142052550","I010370
82515 282514 282513 282512 282511 282510 282509 282508 ...
 $ OFFENSE_CODE       : int  619 1402 3410 3114 3114 3820 724 3301 3
 $ OFFENSE_CODE_GROUP : Factor w/ 67 levels "Aggravated Assault",..:
 65 ...
 $ OFFENSE_DESCRIPTION: Factor w/ 244 levels "A&B HANDS, FEET, ETC.
 130 231 223 124 124 165 22 232 207 232 ...
 $ DISTRICT           : Factor w/ 13 levels "","A1","A15",..: 9 7 10
 $ REPORTING_AREA     : int  808 347 151 272 421 398 330 584 177 364
 $ SHOOTING           : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1 1 1
 $ OCCURRED_ON_DATE   : Factor w/ 233229 levels "2015-06-15 00:00:00
 233228 233226 233227 233229 233224 233225 233222 ...
 $ YEAR               : int  2018 2018 2018 2018 2018 2018 2018 2018
 $ MONTH              : int  9 8 9 9 9 9 9 9 9 9 ...
 $ DAY_OF_WEEK        : Factor w/ 7 levels "Friday","Monday",..: 4 6
 $ HOUR               : int  13 0 19 21 21 21 21 20 20 20 ...
 $ UCR_PART           : Factor w/ 5 levels "","Other","Part One",..:
 $ STREET             : Factor w/ 4658 levels "","  ALBANY ST ",..: 2
 3100 2461 2742 2505 ...
 $ Lat                : num  42.4 42.3 42.3 42.3 42.3 ...
 $ Long               : num  -71.1 -71.1 -71.1 -71.1 -71.1 ...
 $ Location           : Factor w/ 18194 levels "(-1.00000000, -1.000
2 11292 2034 4673 6752 10012 10748 5513 ...
```

Data Preparation

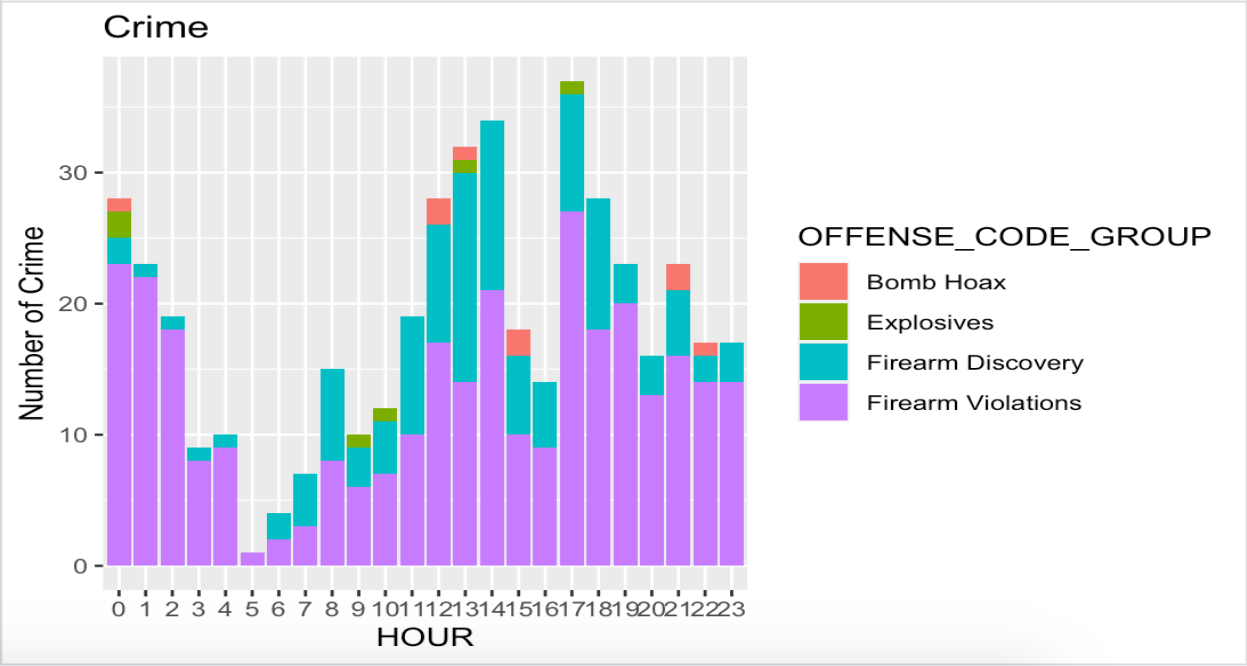
- In the data set, columns REPORTING_AREA has 20250 NA, The Lat column has 19999 NA, the Long column has 19999 NA. We drop all rows containing NA.
- Encoding categorical data to make our dataset easier to be understood and analyzed. We assigned labels, “No” and “Yes” to the SHOOTING variable.
- Convert OFFENSE_CODE, REPORTING_AREA, YEAR, MONTH, HOUR from numeric to factor variables.
- Remove 22 duplicate records in the data set.
- Drop variables INCIDENT_NUMBER, which has no predictive value, and LOCATION, which is redundant. However, for different analysis, we select different variables which will be explained in each analysis.

Exploratory Analysis

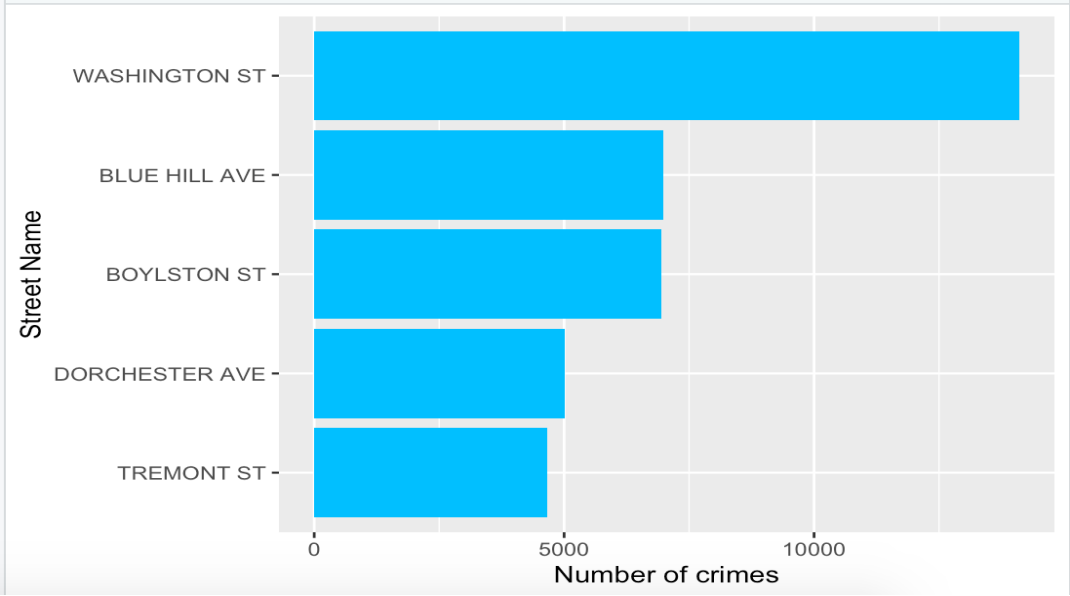
Top 5 Crimes in the year of 2018 at Boston.



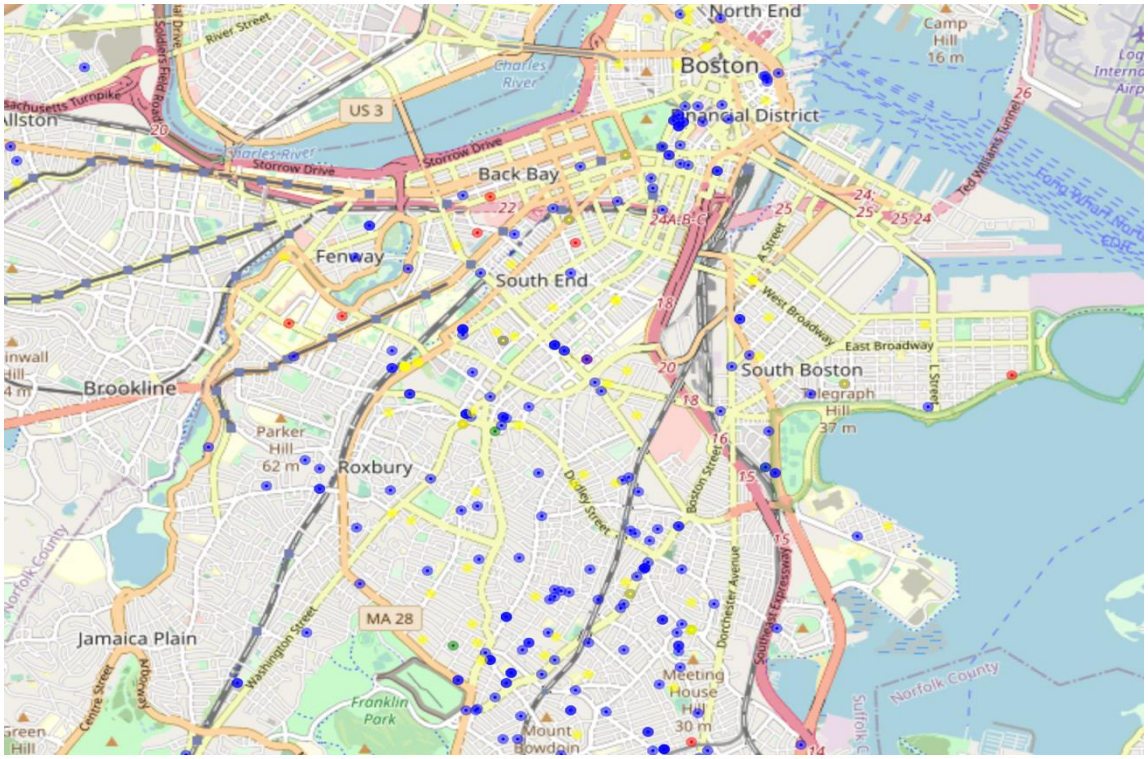
We consider bomb hoax, explosives, firearm discovery, firearm violations, human trafficking as serious crimes. The following plot is for the year 2018 that Serious crime happened at different hours. We observe that Firearm violations are more often than other serious crimes. Boston had the highest number of serious crimes at around 5 PM.



The below plot shows the street name in Boston that has the highest number of crimes. Washington St has the highest number of crimes among other streets. The police department should focus more on patrolling the area.



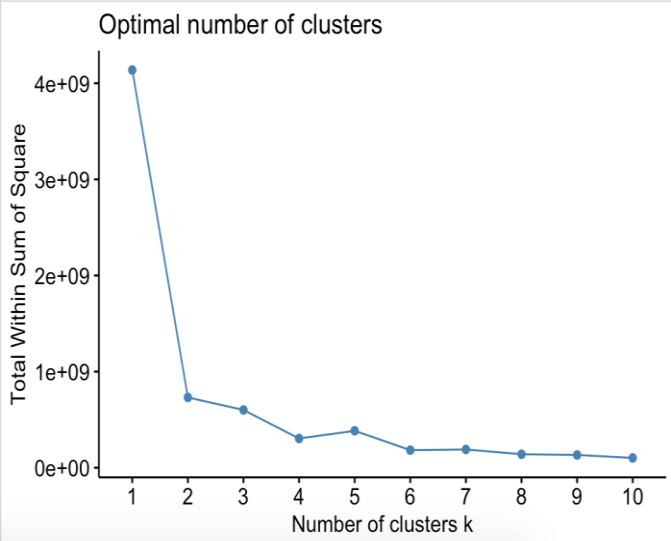
The following plot is the serious crime map in Boston in the year 2018. Each color circle represents a serious type of crime. We observe that most of the crime has happened in downtown Boston.



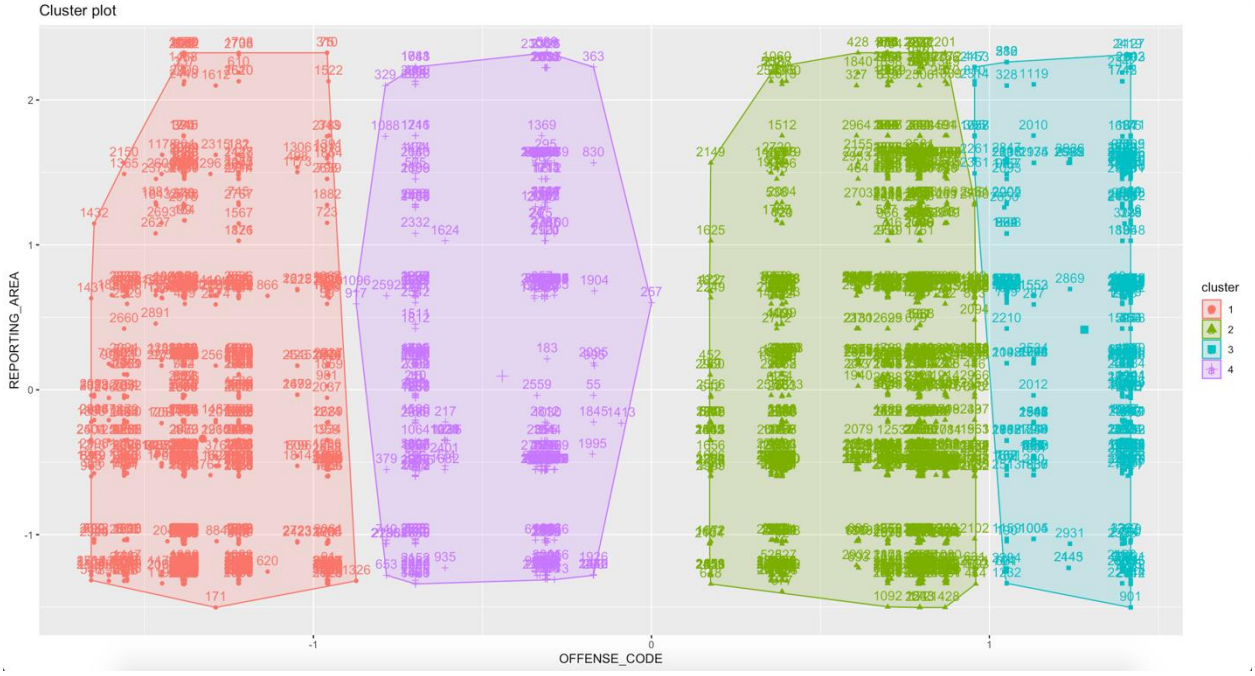
Data Analysis

1.K-means clustering analysis

We are performing K means clustering analysis based on a partial dataset because the dataset is too large to cluster. We choose the year 2018 and WASHINGTON ST as the condition to filter the dataset.



Optimal Clusters = 4



In this cluster analysis, we are using 4 clusters to group OFFENSE_CODE and REPORTING AREA. We can see all OFFENSE_CODE has been grouped into 4 different reporting areas. This cluster plot is based on the incidents that happened in 2018 at WASHINGTON ST.

2.Association Rules Analysis

For this analysis, we only keep OFFENSE_CODE_GROUP, SHOOTING, YEAR, MONTH, DAY_OF_WEEK, HOUR, and STREET, because we want to know when and where is each crime type more likely to happen. The reason we keep OFFENSE_CODE_GROUP instead of OFFENSE DESCRIPTION is that OFFENSE_CODE_GROUP has much less distinct values which are easier to generate effective rules.

We replace "Monday" to "Sunday" with numbers 1 to 7 in DAY_OF_WEEK to make the results clearer. And we group some variables to reduce distinct values based on the distinct levels of each variable, by doing which, each category will have a similar number of records. For example, we divide 12 months into 4 quarters, and we categorize different hours into four groups, morning, afternoon, evening, and night. We also divide 7 different days per week into two groups, Monday to Thursday and Friday to Sunday.

After completing these preparations, we apply the apriori algorithm to generate crime rules and get the following interesting findings.

Rule-1:

```
    lhs                                          rhs                        support    confidence lift      count
[1] {OFFENSE_CODE_GROUP=Fraud}                => {DAY_OF_WEEK=Mon to Thu} 0.01221664 0.6687253  1.1505832 3898
[2] {OFFENSE_CODE_GROUP=Residential Burglary} => {DAY_OF_WEEK=Mon to Thu} 0.01051797 0.5986443  1.0300045 3356
[3] {OFFENSE_CODE_GROUP=Violations}           => {DAY_OF_WEEK=Mon to Thu} 0.01140491 0.5970468  1.0272558 3639
```

```
    lhs                                      rhs                        support    confidence lift      count
[1] {OFFENSE_CODE_GROUP=Aggravated Assault} => {DAY_OF_WEEK=Fri to Sun} 0.01150520 0.4702190  1.1227920 3671
[2] {OFFENSE_CODE_GROUP=Vandalism}          => {DAY_OF_WEEK=Fri to Sun} 0.02220495 0.4596173  1.0974770 7085
[3] {OFFENSE_CODE_GROUP=Property Lost}      => {DAY_OF_WEEK=Fri to Sun} 0.01389024 0.4545175  1.0852997 4432
```

We find that crimes like fraud, residential burglary, and violations are more likely to happen from Monday to Thursday. However, crimes like aggravated assault, vandalism and property loss happen more frequently on Friday and weekends.

Rule-2:

```
    lhs                                                rhs          support    confidence lift     count
[1] {OFFENSE_CODE_GROUP=Aggravated Assault}         => {HOUR=night} 0.01221664 0.4992955  1.333454 3898
[2] {OFFENSE_CODE_GROUP=Larceny From Motor Vehicle} => {HOUR=night} 0.01527550 0.4493408  1.200041 4874
[3] {OFFENSE_CODE_GROUP=Verbal Disputes}            => {HOUR=night} 0.01803349 0.4392702  1.173146 5754
[1] {OFFENSE_CODE_GROUP=Property Lost}             => {HOUR=afternoon} 0.01164937 0.3811917  1.0909023 3717
```

Property Lost happens more in the afternoon while aggravated assault, larceny-theft from Motor Vehicles and Verbal Disputes take place more at night.

Rule-3:

```
    lhs                                          rhs           support    confidence lift      count
[1] {OFFENSE_CODE_GROUP=Missing Person Reported} => {SHOOTING=No} 0.01190010 1.0000000  1.0032039 3797
[2] {OFFENSE_CODE_GROUP=Harassment}              => {SHOOTING=No} 0.01255825 1.0000000  1.0032039 4007
[3] {STREET=COMMONWEALTH AVE}                    => {SHOOTING=No} 0.01295628 1.0000000  1.0032039 4134
[4] {STREET=HYDE PARK AVE}                       => {SHOOTING=No} 0.01086899 0.9994236  1.0026256 3468
```

Missing person and harassment are almost unlikely to cause shooting. And shooting has a very small chance to take place at COMMONWEALTH AVE and HYDE PARK AVE.

Rule-4:

```
    lhs                                          rhs          support    confidence lift      count
[1] {OFFENSE_CODE_GROUP=Medical Assistance} => {YEAR=2018} 0.01756965 0.2381478  1.1568325 5606
[2] {OFFENSE_CODE_GROUP=Simple Assault}     => {YEAR=2018} 0.01121373 0.2260837  1.0982293 3578
```

```
    lhs                                          rhs                support    confidence      lift count
[1] {OFFENSE_CODE_GROUP=Larceny From Motor Vehicle} => {MONTH=quarter_3} 0.01087839  0.3199963 1.0644293  3471
[2] {OFFENSE_CODE_GROUP=Verbal Disputes}            => {MONTH=quarter_3} 0.01229813  0.2995649 0.9964665  3924
[3] {OFFENSE_CODE_GROUP=Towed}                      => {MONTH=quarter_3} 0.01056498  0.2986622 0.9934638  3371
```

Simple Assault is more likely to happen in 2018 and Larceny From Motor Vehicle is more likely to happen in the third quarter.

Rule-5:

```
    lhs                                rhs                 support    confidence lift      count
[1] {OFFENSE_CODE_GROUP=Larceny} => {HOUR=morning}    0.02317025 0.2850588  1.0323194 7393
[2] {OFFENSE_CODE_GROUP=Larceny} => {MONTH=quarter_2} 0.01988573 0.2446501  0.9654711 6345
```

Larceny happens more in the morning and in the second quarter.

Prediction Models

1.Predicting shooting cases

Since shooting is a severe crime, we'd like to predict as precisely as possible if there will be a shooting case in the future. So we apply three different models, SVM, Naïve Bayes, and decision trees to make a prediction.

In this part, we drop INCIDENT_NUMBER, OCCURRED_ON_DATE and Location which is either not predictive or redundant.

SVM:

We firstly normalize the data except prediction variable SHOOTING and then combine SHOOTING with standardized data. Then splitting 80% of the dataset into the training set and the rest 20% into the test set. After scaling the features, proceed to fitting the SVM classifier data to the training set. The prediction is defined in the variable crime_pred. From this the data of the test set results are predicted. In order to find how accurate the predictions are, we use the confusion matrix and create an accuracy function defined the accuracy level as sum(diag(x) / (sum(rowSums(x))) *100, which is shown below.

```
> print(conf_matrix_svm)
                  crime_pred
crime_test_labels Pred:No Pred:Yes
       Actual:No    59062        0
       Actual:Yes     195        0

> accuracy(conf_matrix_svm)
[1] 99.67092
```

Decision Tree:

We use Boosting C5.0 Decision Tree Accuracy to train our model and use test set to make prediction. Next, we adjust the complexity parameter, the Minimum Split Size, the Minimum Bucket Size, and the Maximum depth, comparing their accuracy level to find the most accurate one. Finally, pruning the tree using the best cp and got the following output.

```
> print(conf_matrix_decision_tree)
                  crime_pred
crime_test_labels    No   Yes
              No  59055     7
              Yes   169    26

> accuracy(conf_matrix_decision_tree)
[1] 99.70299
```

Naïve Bayes Model:

We build a Naive Bayes model and apply the model to predict test data, then adjust the model with Laplace=1, and got the below output.

```
> print(conf_matrix_Bayes)
                  crime_test_pred_1
crime_test_labels Pred:No Pred:Yes
       Actual:No    58495      567
       Actual:Yes       1      194

> accuracy(conf_matrix_Bayes)
[1] 99.04146
```

Model Evaluation

	Accuracy	Precision	Recall	F1 Score
SVM	99.67%	0	0	0

	Accuracy	Precision	Recall	F1 Score
Decision Tree	99.70%	78.79%	13.33%	22.81%

	Accuracy	Precision	Recall	F1 Score
Naïve Bayes	99.04%	25.49%	99.49%	40.59%

Since police department wants to know more about the probability of correctly predicting a shooting case among actual cases, recall is the best measure to evaluate the model accuracy, and Naïve Bayes model has the highest recall, 99.49%, showing it's a good model because it almost won't miss a shooting crime. Therefore, although the overall accuracy for Bayes model is a little lower than the other two models, we still select Naïve Bayes model in this case. However, the precision value for Bayes model is pretty low, which indicates there may waste a lot of police resources to investigate a nonexistent shooting crime.

2.Predicting UCR_PART code for incidents

Naïve Bayes model

Naive Bayes' model is based on all incidents that happened in 2018 in Boston. We conduct the Bayes model by splitting the data set (60712 records) into the training(45534 records) and testing the data set(15177 records). Our Naive Bayes classifier is based on the training data set and labels. Prediction is based on the classifier and test data set. In our classifier we are setting UCR_PART as our dependent variable and all other variables as independent variables. UCR_PART means Uniform Crime Reporting which compiles official data on crime, UCR is developed by FBI. For example: Part I offenses included murder, non-negligent homicide, rape, robbery, aggravated assault and so on.

Below is our confusion matrix:

NB_Predictions	Other	Part One	Part Three	Part Two
4	0	0	1	0
Other	0	68	7	88	74
Part One	0	0	2689	6	5
Part Three	0	0	1	7618	12
Part Two	0	0	0	66	4539

	Accuracy	Precision	Recall	F1 Score
Naive Bayes	98.29%	100%	80%	88.89%

Tuning by using laplace = 1 but it decreased our model accuracy. Laplace Smoothing is a technique to smooth categorical data. Laplace helps us avoid unseen value or feature that may result the probability to 0.

NB_Predictions	Other	Part One	Part Three	Part Two
4	0	23	155	157
Other	0	68	6	133	103
Part One	0	0	2668	0	0
Part Three	0	0	0	7491	0
Part Two	0	0	0	0	4370

	Accuracy	Precision	Recall	F1 Score
Naive Bayes (Laplace)	96.20%	100%	1.18%	2.33%

KNN Model

The KNN model is based on limited data. We are focusing on incidents that happened in January 2018 at WASHINGTON ST. Our dependent variable is still different UCR_PART crimes. Independent variables are OFFENSE_CODE, REPORTING_AREA,HOUR. Start with a KNN model by splitting the data by a proportion of 70%. Confusion matrix is shown below.

	test_pred			
crimes_raw7_test_labels	Part One	Part Three	Part Two	Row Total
Part One	15	0	7	22
Part Three	0	46	0	46
Part Two	2	11	23	36
Column Total	17	57	30	104

	Accuracy	**Precision**	**Recall**	**F1 Score**
KNN	80.77%	80.90%	80.77%	80.83%

After comparing these two models, Naive Bayes model has higher accuracy and precision, so we decide to select Naïve Bayes model in this case. But Naive Bayes model also has some limitations. For example, it is robust to irrelevant attributes and noise points.

Conclusions

1.Time and location have a big impact on the number of crimes.
2.Among serious crimes, firearm violation crimes happen more frequently.
3.Crimes tend to happen more at night than in the morning.
4.The highest number of crimes in Boston is injury related to person and investigate person.
5.Most of the serious crimes appear in the downtown of Boston.
6.COMMONWEALTH AVE almost won't have shooting crime.

Recommendations

This analysis indicates there are some patterns in crime. Certain areas and certain time have higher risks of some crimes happening. Therefore, the police department can assign more police officers to patrol at night and in downtown area, and focus more on preventing firearm violation crimes whereas COMMONWEALTH AVE can be allocated less manpower.

Since shooting case may require more human and financial resources to handle, the accuracy level of prediction is very important. Although we don't want to overlook a shooting case, we also want to allocate the precious resources efficiently. Therefore, in order to increase the precision value, the police department need to consider other factors that may be relevant to shooting problem when they collect the data.