IST 707: Data Analytics

San Francisco Crime Analysis

Nikita Pralhad Dongare (SUID: 651838408)

Siddhant Prashant Bandiwadekar (SUID:364845292)

Problem:

The city of San Francisco is known for its technological advancements but at one point was infamous for having the most criminal activities taking place. By applying data mining techniques on the San Francisco crime data set which contains a record of the crimes across all of San Francisco's neighbourhoods to predict the type of crime that would take place in a particular area. This would help the San Francisco Police Department to curb the criminal incidents well in advance. The incidents that occurred in all the areas need to be analysed to discover connections between the different crimes in a specific area using classification techniques.

Dataset Information:

The dataset used is the San Francisco Crime dataset which contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2016 to 12/31/2016. The data has 150500 observations and 13 attributes. The dataset is obtained from the central clearinghouse for data published by the City and County of San Francisco (https://datasf.org/opendata/).

Dataset Link: https://www.kaggle.com/roshansharma/sanfranciso-crime-dataset

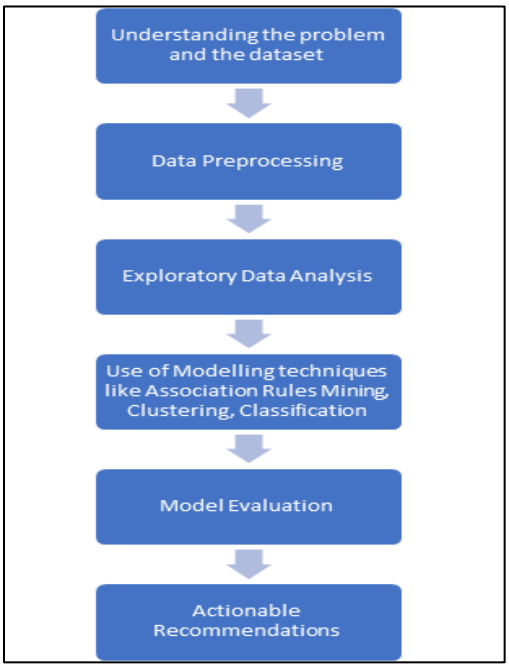
Attributes:

There are 13 attributes in the dataset which are as follows:

1. Incident number: ID of the incident
2. Dates: date of the crime incident
3. Time: time of the incident
4. Category: category of the crime incident like whether it is larceny, assault, etc. This is the target variable we are going to predict.
5. Descript: detailed description of the crime incident
6. DayOfWeek: the day of the week
7. PdDistrict: name of the Police Department District
8. Resolution: how the crime incident was resolved
9. Address: the approximate street address of the crime incident
10. X: Longitude
11. Y: Latitude
12. Location: location field in the form of a pair of coordinates
13. PdId: unique identifier for each complaint registered

Illustration of the tasks:

The tasks performed were performed in the following order:



Data Pre-processing:

The following steps were performed as a part of data preprocessing:

- One blank value was removed from the dataset.
- Columns like IncidentNum, Location, Time, Date , X , Y , Address and Resolution were removed since they were not considered significant attributes for prediction.
- There are 39 types of crime categories in the dataset. Several of these crime categories are similar to each other and hence were combined into 2 categories namely: Non Violent and Violent Crimes.

  Crimes which fall under the 2 categories are as follows:

  Violent Crimes: Assault, Burglary, Family Offenses, Kidnapping, Robbery, Sex Offenses (Forcible), Sex Offenses (Non Forcible), Suicide, Trea, Vandalism, Warrants, Weapon Laws.

  Non-Violent Crimes: Arson, Bad Checks, Bribery, Disorderly Conduct, Driving under the influence, Drug/ Narcotic, Drunkenness, Embezzlement, Extortion, Forgery, Fraud, Gambling, Larceny/ Theft, Liquor Laws, Missing Person, Loitering, Non-Criminal, Other Offenses, Pornography, Prostitution, Recovered Vehicle, Runaway, Secondary Codes, Stolen Property, Suspicious Occ, Trespass, Vehicle Theft.
- Crime Category, DayOfWeek, Weekend and PdDistrict were converted to numeric datatype.
- Month is extracted from the Date column and Hour is extracted from the Time column.
- Month and Hour variables were converted to factor using as.factor().
- The Target Variable (Crime Category) was converted to factor using as.factor().

Exploratory Data Analysis:

The insights generated by our analysis will be used to answer the following business questions:

1. **Which category of crime has the highest number of crimes?**

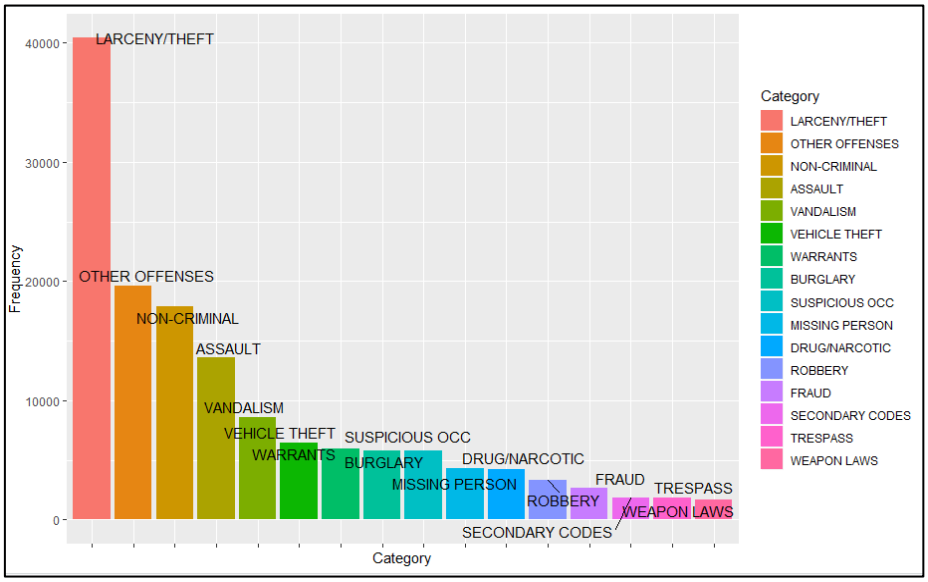


Figure 1: Distribution of Crimes by categories

From the above bar chart, we can see the distribution of crimes across various categories. Larceny/ theft has the highest number of crimes followed by other offenses, non-criminal, assault, vandalism and so on. Larceny/theft and vehicular theft accounted for almost 27% of the San Francisco city's different crimes.

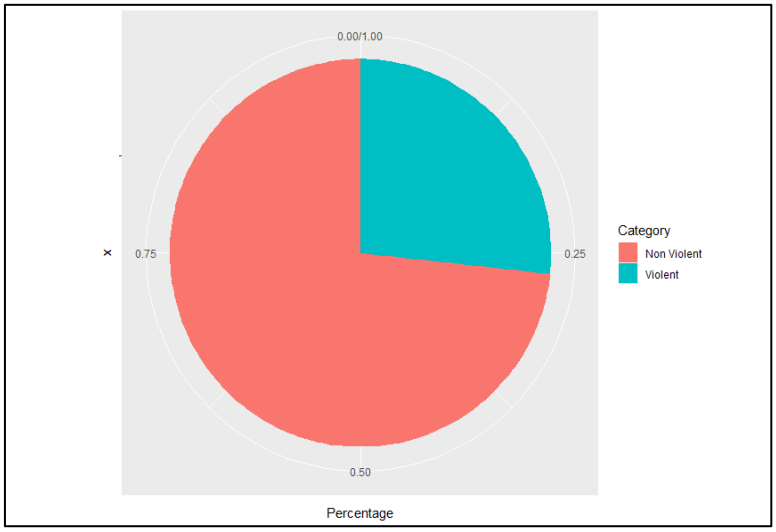


Figure 2: Distribution of crimes into violent and nonviolent crime categories

Since the 37 crime categories were classified into 2 crime categories namely violent and non-violent for better accuracy and prediction, the above pie chart shows the percentage of crimes in each category. We can see that non-violent crimes account for the 70 % of the crimes whereas violent crimes account for around 30% of the crimes.

2. On which day of the week, do most crimes take place?

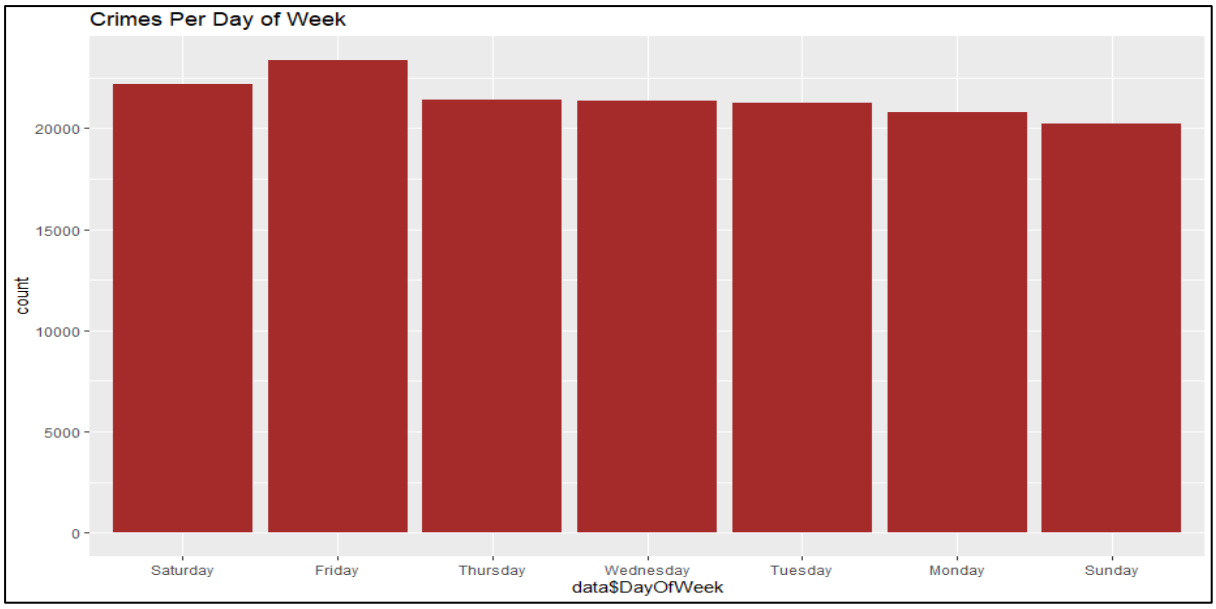


Figure 3: Distribution of crimes by day of the week

From the bar graph above, we can see that the maximum number of crimes take place on Friday followed by Saturday. This implies that as the weekend approaches, the number of crimes also increase.

3. At what hour of the day, do maximum crimes take place?

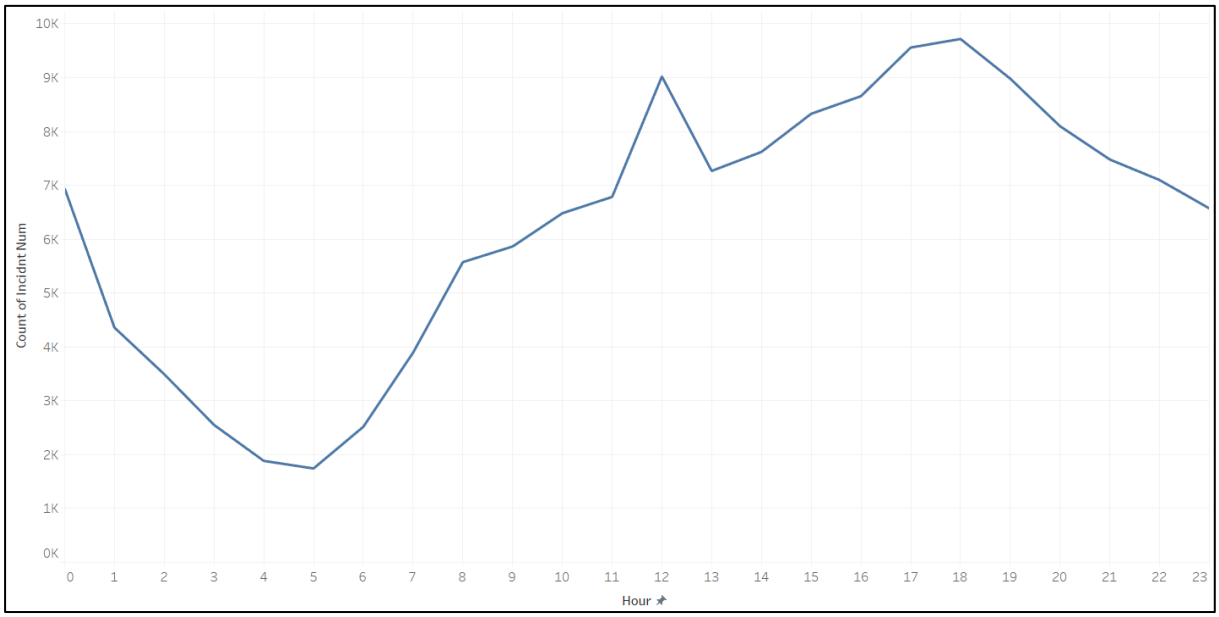


Figure 4: Distribution of crimes by hour

The above line graph shows the distribution of crimes at each hour of the day. It can be observed that 5 am is the safest part of the day whereas 6 pm is the most dangerous hour with the highest reported incidents. In addition to this, 4 pm to 8 pm are the peak hours that are prone to crimes. Moreover, 12 pm is the second dangerous hour during the day.

4. In which month, do maximum crimes take place?

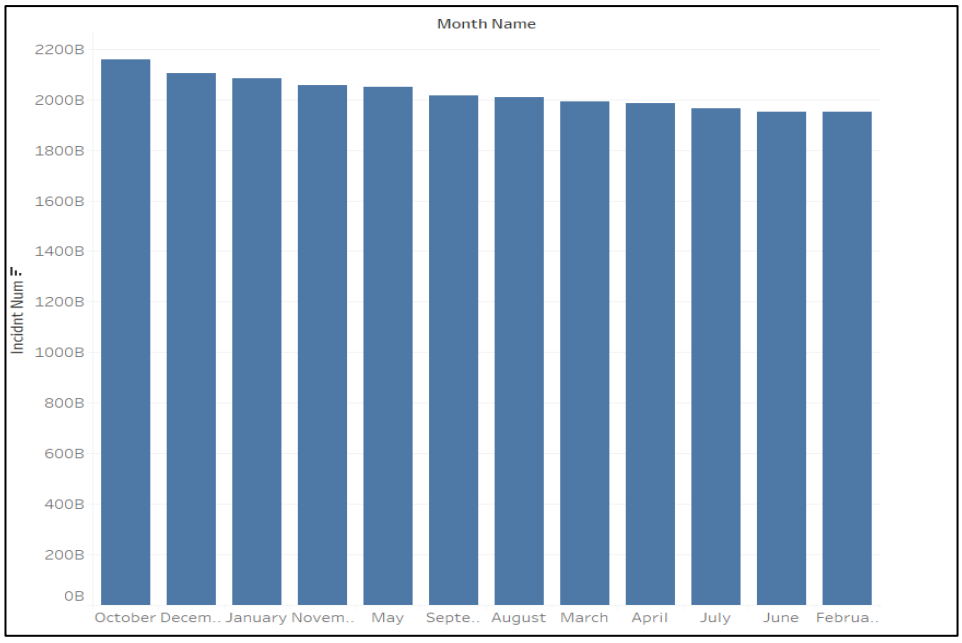


Figure 5: Distribution of crimes over months

From the above graph, we can see that October has the maximum number of crimes whereas February is the safest month.

5. Which district has the maximum number of crimes?

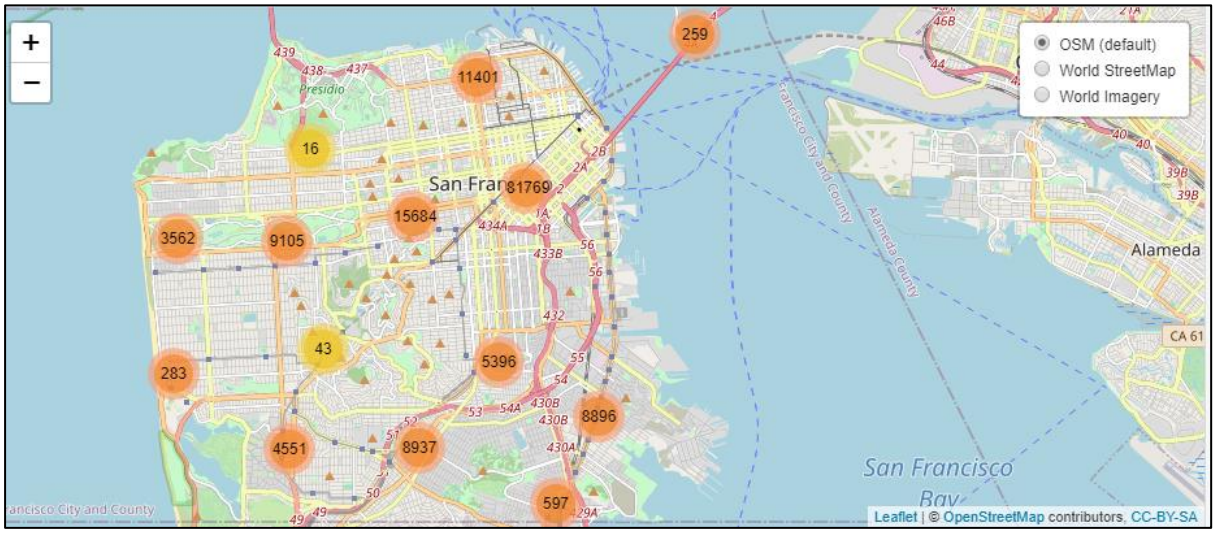


Figure 6: Distribution of crimes across various districts in San Francisco

As we can see from the above heat map, Southern district of San Francisco has the highest number of crimes which is around 81769 followed by Tenderloin district which has around 15684 crimes. These 2 districts are a part of Southern San Francisco where most of the most famous neighbourhoods exist. Thus, more security forces could be deployed in those areas as well as increased investments in technology surveillance tools to track down crimes could prove helpful.

6. What are the top streets where the highest number of crimes take place?

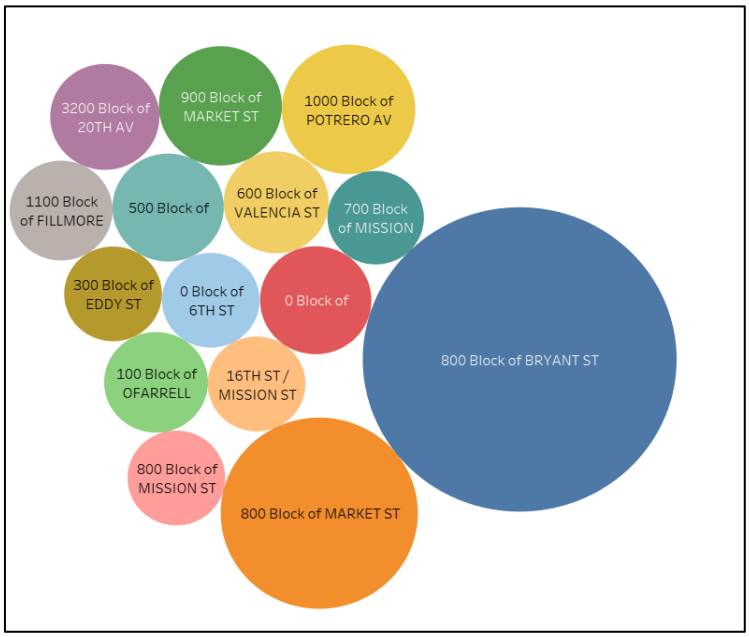


Figure 7: Distribution of crimes across streets in San Francisco

From the above graph, we can conclude that Bryant st and market street have the highest number of crime incidents. Both of those streets are in famous neighbourhoods. Thus, additional security could be deployed during peak hours on those streets.

7. In which districts, is the number of Larceny/ Theft crimes highest?

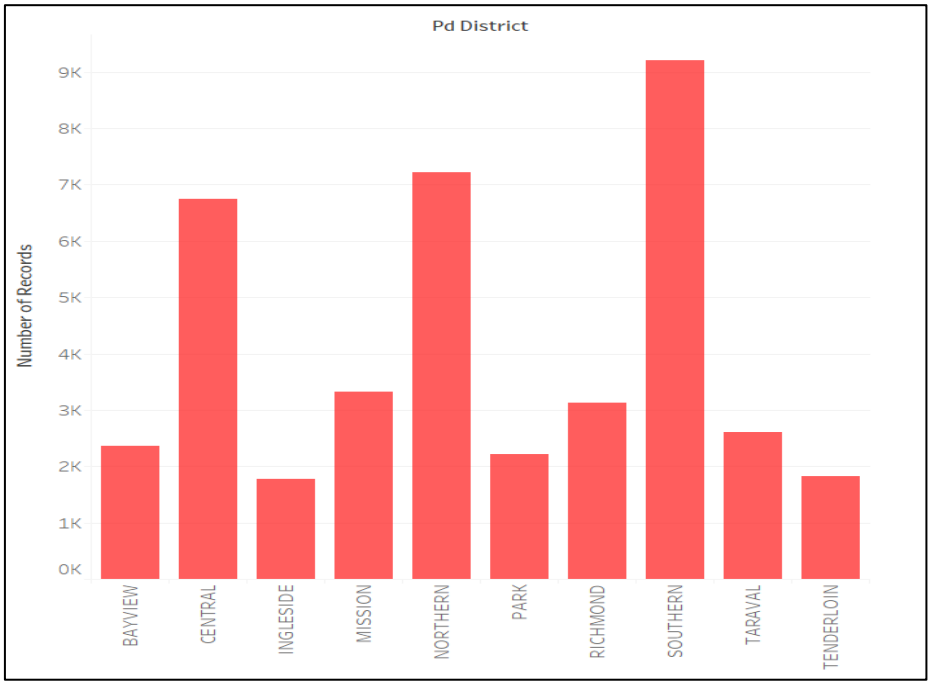


Figure 8: Distribution of Larceny/ Theft related crimes across districts in San Francisco

Larceny/ Theft accounts for almost 27% of San Francisco's crimes. From the above graph, we can see that Larceny/ Theft takes place the most in the Southern, Northern and Central districts of San Francisco. Thus, increasing security forces in these districts could help in curbing the crimes.

Modelling Techniques:

Techniques such as classification and association rule mining were used to create a prediction model in R which helped us anticipate the location of incidents occurring and detect the links between the various categories of crime that have been taking place in specific locations around the city. We used the following three techniques:

1) Association Rule Mining
2) Naive Bayes
3) CART

1) Association Rule Mining:

Association rule mining among large sets of data items finds interesting associations and relationships. This rule demonstrates how often a transaction happens with an item set. We may find rules that predict the occurrence of an item based on the occurrence of other items in the transaction, given a collection of transactions.

For RHS set to Crime.Category = Violent, and support and confidence set to 0.001 and 0.95 respectively, we obtained the following 6 interesting rules:

```
> inspect(Rules)
    lhs                    rhs                          support confidence     coverage     lift count
[1] {DayOfWeek=Thursday,
     Month=1,
     Hour=5,
     PdDistrict=BAYVIEW} => {Crime.Category=Violent} 5.980106e-05          1 5.980106e-05 3.743663     9
[2] {DayOfWeek=Sunday,
     Month=9,
     Hour=3,
     PdDistrict=CENTRAL} => {Crime.Category=Violent} 4.651194e-05          1 4.651194e-05 3.743663     7
[3] {DayOfWeek=Tuesday,
     Month=2,
     Hour=0,
     PdDistrict=TARAVAL} => {Crime.Category=Violent} 4.651194e-05          1 4.651194e-05 3.743663     7
[4] {DayOfWeek=Thursday,
     Month=1,
     Weekend=No,
     Hour=5,
     PdDistrict=BAYVIEW} => {Crime.Category=Violent} 5.980106e-05          1 5.980106e-05 3.743663     9
[5] {DayOfWeek=Sunday,
     Month=9,
     Weekend=Yes,
     Hour=3,
     PdDistrict=CENTRAL} => {Crime.Category=Violent} 4.651194e-05          1 4.651194e-05 3.743663     7
[6] {DayOfWeek=Tuesday,
     Month=2,
     Weekend=No,
     Hour=0,
     PdDistrict=TARAVAL} => {Crime.Category=Violent} 4.651194e-05          1 4.651194e-05 3.743663     7
```

Figure 9: Association Rules for Crime Category= Violent, support= 0.001, confidence= 0.95

From the above analysis, we learnt that maximum number of violent crimes could occur in the Taraval District of San Francisco between 00:00 to 01:00 am during the month of February. From rules 1 and 4, we can also imply that Bayview district is prone to violent crimes on Thursdays whereas Central district is more likely to violent crimes on Sundays.

For RHS set to Crime.Category = Nonviolent and support and confidence set to 0.001 and 0.85 respectively, we obtained the following 12 interesting rules:

```
> inspect(Rules)
     lhs                                rhs                                support confidence    coverage     lift count
[1]  {weekend=Yes,
      Hour=9,
      PdDistrict=CENTRAL}  => {Crime.Category=Non Violent} 0.001056485  0.8641304 0.001222599 1.179085   159
[2]  {weekend=No,
      Hour=11,
      PdDistrict=RICHMOND} => {Crime.Category=Non Violent} 0.001993369  0.9063444 0.002199350 1.236685   300
[3]  {weekend=No,
      Hour=13,
      PdDistrict=PARK}     => {Crime.Category=Non Violent} 0.001654496  0.8556701 0.001933568 1.167542   249
[4]  {weekend=No,
      Hour=13,
      PdDistrict=RICHMOND} => {Crime.Category=Non Violent} 0.001966790  0.8654971 0.002272440 1.180950   296
[5]  {weekend=No,
      Hour=13,
      PdDistrict=CENTRAL}  => {Crime.Category=Non Violent} 0.003681088  0.8536210 0.004312321 1.164746   554
[6]  {weekend=No,
      Hour=14,
      PdDistrict=RICHMOND} => {Crime.Category=Non Violent} 0.002093037  0.8630137 0.002425265 1.177562   315
[7]  {weekend=No,
      Hour=12,
      PdDistrict=PARK}     => {Crime.Category=Non Violent} 0.002511645  0.8630137 0.002910318 1.177562   378
[8]  {DayOfWeek=Thursday,
      Hour=19,
      PdDistrict=NORTHERN} => {Crime.Category=Non Violent} 0.001176088  0.8676471 0.001355491 1.183884   177
[9]  {weekend=No,
      Hour=12,
      PdDistrict=CENTRAL}  => {Crime.Category=Non Violent} 0.004199363  0.8552097 0.004910332 1.166913   632
[10] {DayOfWeek=Friday,
      Month=2,
      PdDistrict=CENTRAL}  => {Crime.Category=Non Violent} 0.001136220  0.8636364 0.001315623 1.178411   171
[11] {DayOfWeek=Thursday,
      weekend=No,
      Hour=19,
      PdDistrict=NORTHERN} => {Crime.Category=Non Violent} 0.001176088  0.8676471 0.001355491 1.183884   177
[12] {DayOfWeek=Friday,
      Month=2,
      weekend=No,
      PdDistrict=CENTRAL}  => {Crime.Category=Non Violent} 0.001136220  0.8636364 0.001315623 1.178411   171
```

Figure 10: Rules for Crime Category= Non-Violent, support= 0.001, confidence= 0.85

From the above analysis, we learnt that maximum crimes could occur in the Central district of San Francisco on Thursdays and Fridays i.e. as the weekend approaches, the number of crimes also increase. Moreover, non-violent crimes can take place in Richmond district on weekdays during the hours 11 am to 1 pm. These rules tell us where and when additional security could be required to help the SF Police Department in alleviating crime rate.

2) Naive Bayes:

Naive Bayes classifiers are based on techniques of Bayesian classification. These depend on the Bayes theorem which is an equation that explains the relationship of statistical quantities with conditional probabilities. It is a family of algorithms where a common concept is shared by all of them, i.e. each pair of features being classified is independent of each other.

Splitting into Train and Test data:

90% of the data was used for training and 10% was used for the test set. The “Churn” attribute from the dataset was used to predict the customer churn from the dataset. The value in the set.seed() was changed to a random number.

Model:

The control parameter was set as follows: **Controls <- trainControl (method='repeatedcv', number=15)**

The following confusion matrix and statistics were obtained:

```
Confusion Matrix and Statistics

             Reference
Prediction    Non Violent Violent
  Non Violent       10994    3983
  Violent              35      37

               Accuracy : 0.733
                 95% CI : (0.7259, 0.7401)
    No Information Rate : 0.7329
    P-Value [Acc > NIR] : 0.4895

                  Kappa : 0.0088

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.996827
            Specificity : 0.009204
         Pos Pred Value : 0.734059
         Neg Pred Value : 0.513889
             Prevalence : 0.732873
         Detection Rate : 0.730547
   Detection Prevalence : 0.995216
      Balanced Accuracy : 0.503015

       'Positive' Class : Non Violent
```

As we can see from the above statistics obtained from the confusion matrix:

Test Set Accuracy = 73.3%

3) CART:

A CART tree is a binary decision tree that is built by repeatedly splitting a node into two child nodes, starting with the root node containing the entire sample of learning where y is the variable that is dependent, or the target variable.

Other methods like the Bagged Adaboost and Adaboost Classification tree took a lot of processing time on the system and were difficult to implement using the available hardware resources.

Splitting into Train and Test data:

80% of the data was used for training and 20% was used for the test set. The value in the set.seed() was changed to a random number.

The “Churn” attribute from the dataset was used to predict the customer churn from the dataset.

Model:

The control parameter was set as follows: **Controls <- trainControl (method='repeatedcv', number=10)**

The following confusion matrix and statistics were obtained from the CART Algorithm:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 19932  6854
         1  2127  1186

               Accuracy : 0.7016
                 95% CI : (0.6964, 0.7068)
    No Information Rate : 0.7329
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0628

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9036
            Specificity : 0.1475
         Pos Pred Value : 0.7441
         Neg Pred Value : 0.3580
             Prevalence : 0.7329
         Detection Rate : 0.6622
   Detection Prevalence : 0.8899
      Balanced Accuracy : 0.5255

       'Positive' Class : 0
```

As we can see from the above statistics obtained from the confusion matrix:

Test Set Accuracy = 70.16 %

Model Evaluation:

We have evaluated the models based on the performance metrics such as accuracy, precision, recall, f-measure which are obtained using the confusion matrix. Other aspects of evaluation that were considered were speed, robustness, scalability, and interpretability.

Model Evaluation Table:

Model Evaluation	**Naive Bayes**	**CART**
Accuracy	73.30%	70.16%
Precision	0.7340	0.7441
Recall	0.9968	0.9035
F-measure	0.8459	0.8161

These metrics helped in understanding how accurately the categories are being classified as well as evaluating the model will help the SFPD to significantly minimize the crime rate throughout the city which will be ultimately beneficial for the people living in the city.

From the model evaluation table, we can observe that using Naive Bayes we are getting the highest values of evaluation metrics when compared to the other classification algorithms. Since F-Measure is the highest (0.8454) and gives us the harmonic mean of precision and recall, it helps us in giving accurate results on the dataset as it depends on values of other important metrics (recall and precision). Additionally, aaccuracy of the model is the percentage of correctly classified data. This gives us an idea about how the model is performing. The accuracy of the Naïve Bayes model is higher compared to the CART model. Overall, by comparing the results of all the classification algorithms, we can conclude the Naive Bayes shows the best performance having significantly higher values of F-Score and accuracy.

Conclusion:

Based on our analysis, we conclude by making the following recommendations to the San Francisco security and police authorities:

- Southern and Tenderloin are the PD districts which require special attention.
- Famous neighborhoods like Bryant Street and Market Street have the highest crime rates. Thus, awareness campaigns could be implemented for the local people to be more cautious about their properties in the specific famous neighbourhoods.
- Violent crimes take place in the Tenderloin region during the evenings and thus, such places can be avoided by citizens and additional security can be provided by SFPD.
- Crimes usually increase as the weekend approaches and is the highest on Fridays. We recommend increasing security forces in the public areas during those days.
- October is the month with the highest number of crimes whereas February is the lowest.
- Larceny/theft and vehicular theft account for 27% of the City's reported crime incidents. So, additional care can be taken by the residents to avoid traveling with valuable things to the districts where maximum crimes occur.
- 5 am is the safest part of the day whereas 6 pm is the most dangerous hour with the highest reported incidents. Moreover, 12 pm is the second dangerous hour during the day and in fact is the hour where most of the incidents reported under some crime categories are maximum. Additional security could be made available during those hours in the hotspot regions.
- There are chances of violent crimes taking place in the Taraval district during midnight whereas non-violent crimes could be reported in the Central area of San Francisco on weekdays. So, additional security and public surveillance can be deployed in those areas.