An Approach to Predictive Policing with Dallas City by Team RAAS

0. Load Required packages

```
library(tibble)
library(dplyr)
library(magrittr)
library(readr)
library(zipcode)
library(lubridate)
library(stringr)
library(ggplot2)
library(forcats)
library(plotly)
library(scales)
library(MESS)
library(caret)
data("zipcode")
```

1. Data Reading

Objective : Read CSV from data source - Following chunk reads the source csv and loads only required attributes into a dataframe object for the use *Following chunk performs Data Reading*

```
#Uncomment below import line only when running the script for first time, to avoid multip
le time download of 400+Mb sized dataset
dallas<-read_csv('Police_Incidents.csv')
dallas%<>%select(`Service Number ID`,`Type  Location`,`Division`,`Sector`,`Council Distri
ct`,`Call Received Date Time`,`Victim Gender`,`Victim Age at Offense`,`Offense Status`,`N
IBRS Crime Category`,`Zip Code`)
#as_tibble(dallas)
#summary(dallas)
```

2. Data Pre-processing

Objective : Generate dataframes dallas_incidents,dallas_crime_type and dallas_crime_rate. Dataframe dallas_indcidents must be suitable for Exploratory data analysis. - Filter the required attributes and ignore the non-NA values - Transforms the attribute ‘call received date time’ string to R datetime object and sort them in ascending order
- Compute a new attribute ‘week of the day’(name of the weekday, incident occured viz Mon,Tue and so on), ‘rounded time’(Hours being rounded off to closest value and only hour value is extracted from the rounded date) and week number from ‘call recieved date time’ - Transform values of ‘rounded time’ to four ordinal values and compute ‘time slot of occurence’ attribute. - Clean the dataset to include data only from the city “Dallas” - To remove short head and to keep dataset symmetric, filter rows corresponding to value

from “12/31/2016 23:59:59” to “01/06/2019 00:00:00” (This date range consists of equal number of Mon,Tues,Wed etc of 105 counts) - Unselect the attributes that are not further required. *Following chunk performs Transformation/Cleaning*

```
time_slot_vec=seq(0,24,6)
labels_vec=c("0-6","7-12","13-18","19-23")

dallas$`Zip Code`=clean.zipcodes(dallas$`Zip Code`)

#Dataframe dallas_incidents suitable for visualizing data
dallas_incidents<-dallas%>%
   filter(!is.na(`Offense Status`) & !is.na(`Division`) & !is.na(`Call Received Date Time
`) & !is.na(`Zip Code`)  & !is.na(`Victim Age at Offense`) & !is.na(`NIBRS Crime Category
`) & !is.na(`Type  Location`) & !is.na(Sector) & !is.na(`Council District`))%>%
  inner_join(zipcode, by = c("Zip Code" = "zip"))%>%
  filter(`city`=="Dallas")%>%
  mutate(`Division`=str_to_upper(str_replace(`Division`," ","")))%>%
  mutate(`Call Received Date Time`= as_datetime(mdy_hms(`Call Received Date Time`)))%>%
  filter(`Call Received Date Time`>as_datetime(mdy_hms("12/31/2016 23:59:59")) &  `Call R
eceived Date Time`<as_datetime(mdy_hms("01/06/2019 00:00:00")))%>%
  mutate(`week of the day`=lubridate::wday(`Call Received Date Time`,label = TRUE, abbr
= FALSE),`week number of the day`=lubridate::wday(`Call Received Date Time`),`rounded tim
e`=hour(round_date(`Call Received Date Time`,"hour")))%>%
  mutate(`time slot of occurence`=cut(`rounded time`,breaks = time_slot_vec,labels = labe
ls_vec,include.lowest = TRUE))%>%
  arrange(`Call Received Date Time`)%>%
  select(-`city`,-`state`,-`latitude`,-`longitude`)
```

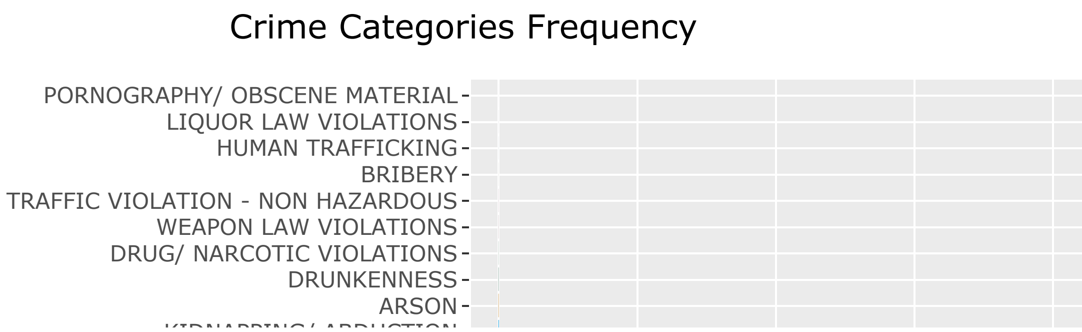
3. Exploratory Data Analysis

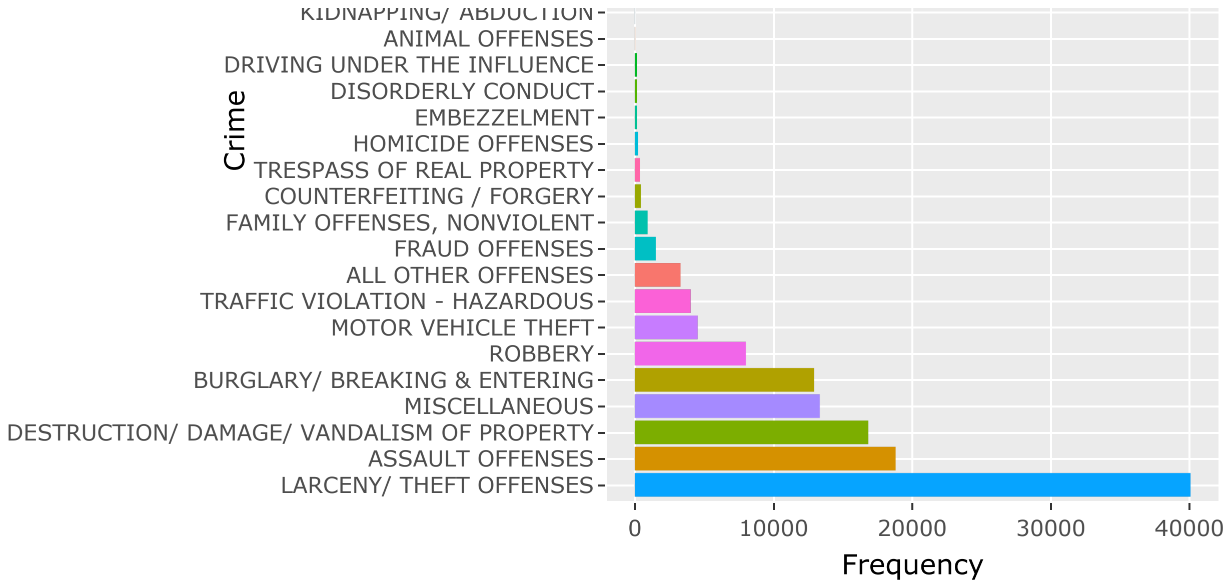
Objective : To evaluate the pattern/trend in the dataset that could 1. Answer some basic questions 2. Help in selecting attributes for predictive analysis

3.1 What crimes are frequent?

Following can be interpreted from the graph

- Larcency/Theft offences are the highest frequent in occurence, followed by Assault offences and Property related crimes.
- Numerous crimes are very negligible in occurence. Crimes such as Human trafficking, bribery, drunkeness among some others are very rare in appearance.
- Some categories such as ‘Miscellenaous’ and ‘Other crimes’ are ambigous. However, they have a decent frequency of occurence.



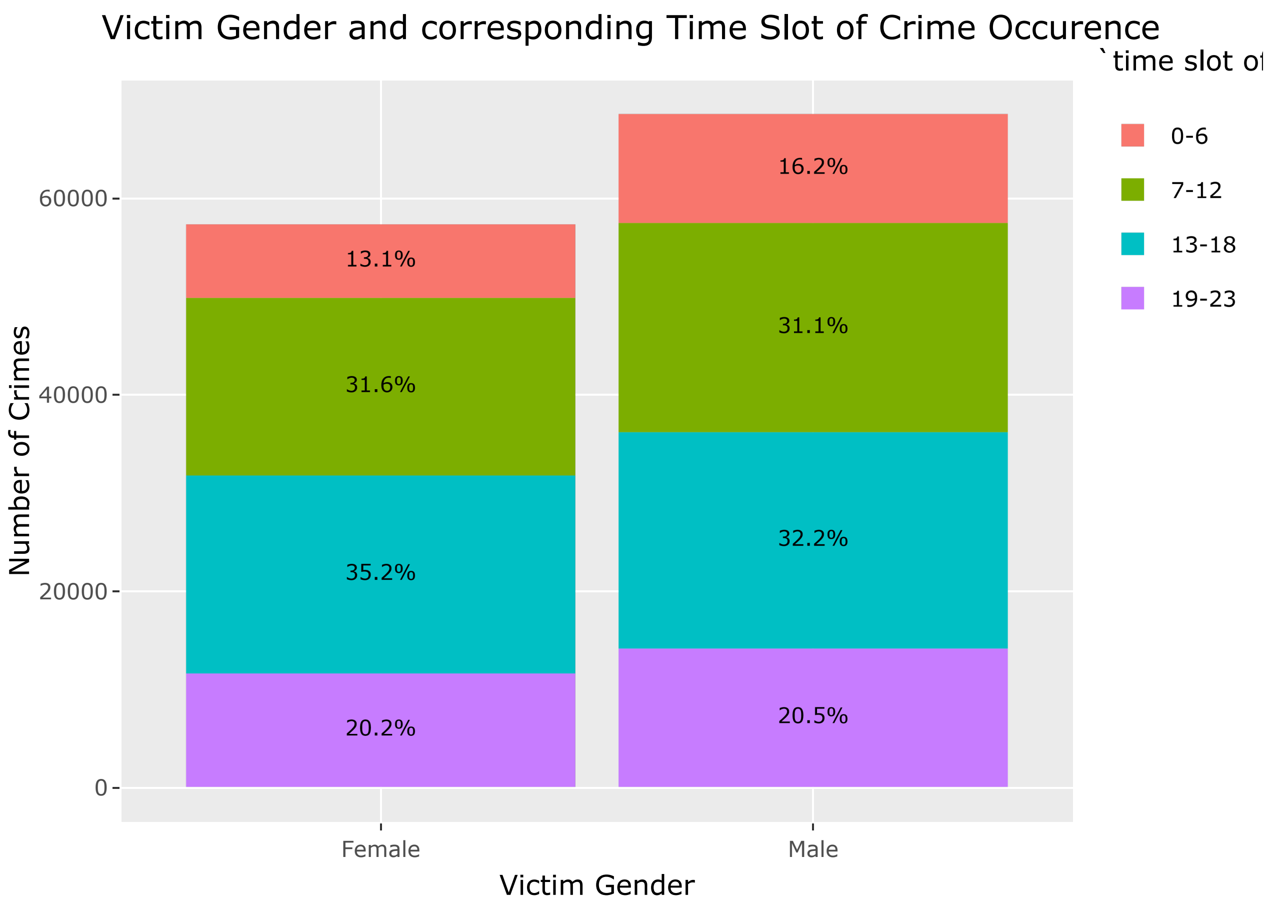


3.2 How victim gender and the time of crime are related and volume of crime for each time slot.

Following can be interpreted from the graph

- Interesting observation is Male are more prone to be victim during late night (00:00 to 06:00).
- Females victims are higher during the afternoon slot (13:00 to 18:00)

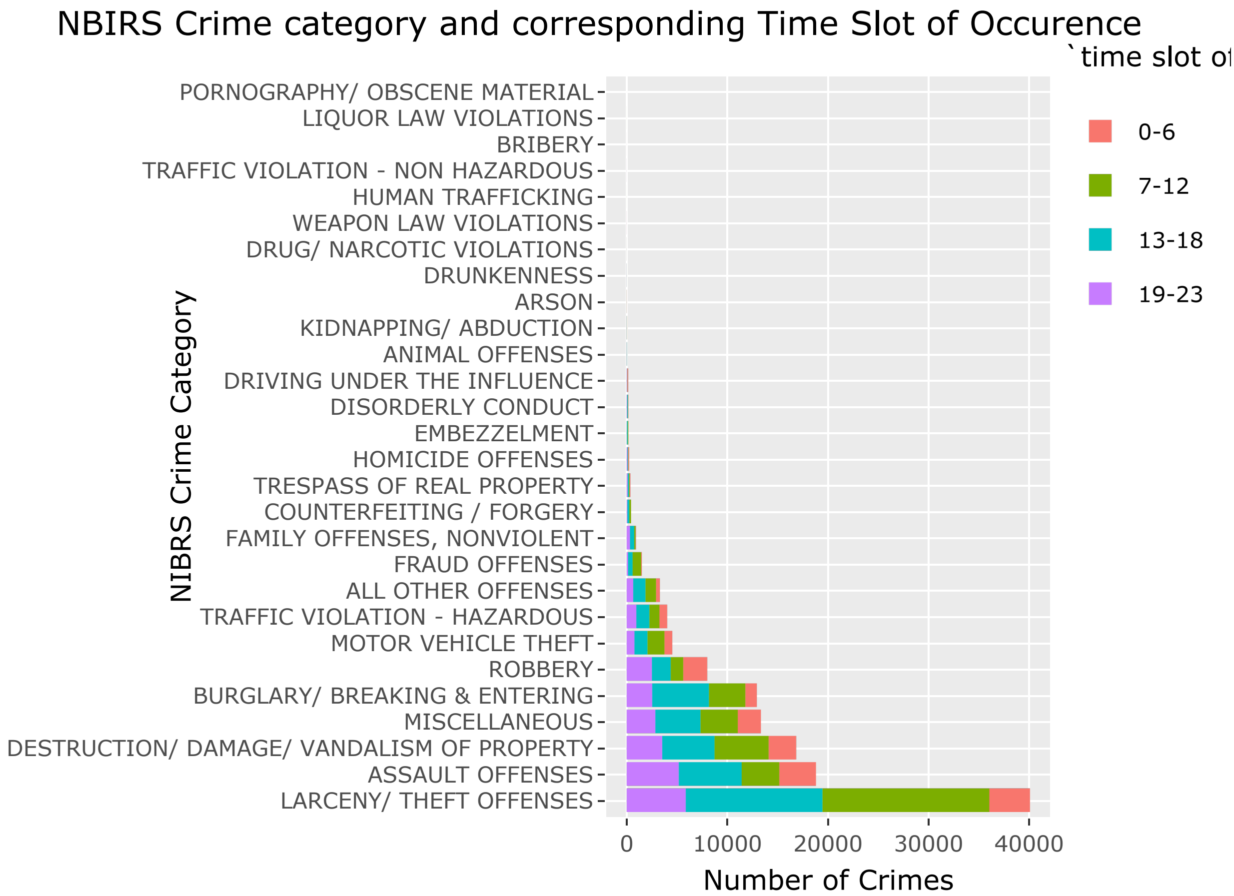
Please note : above observations could also be misleading as the relative population size of the city is not being considered.



3.3 How crimes relate to the time of day

Following can be interpreted from the graph

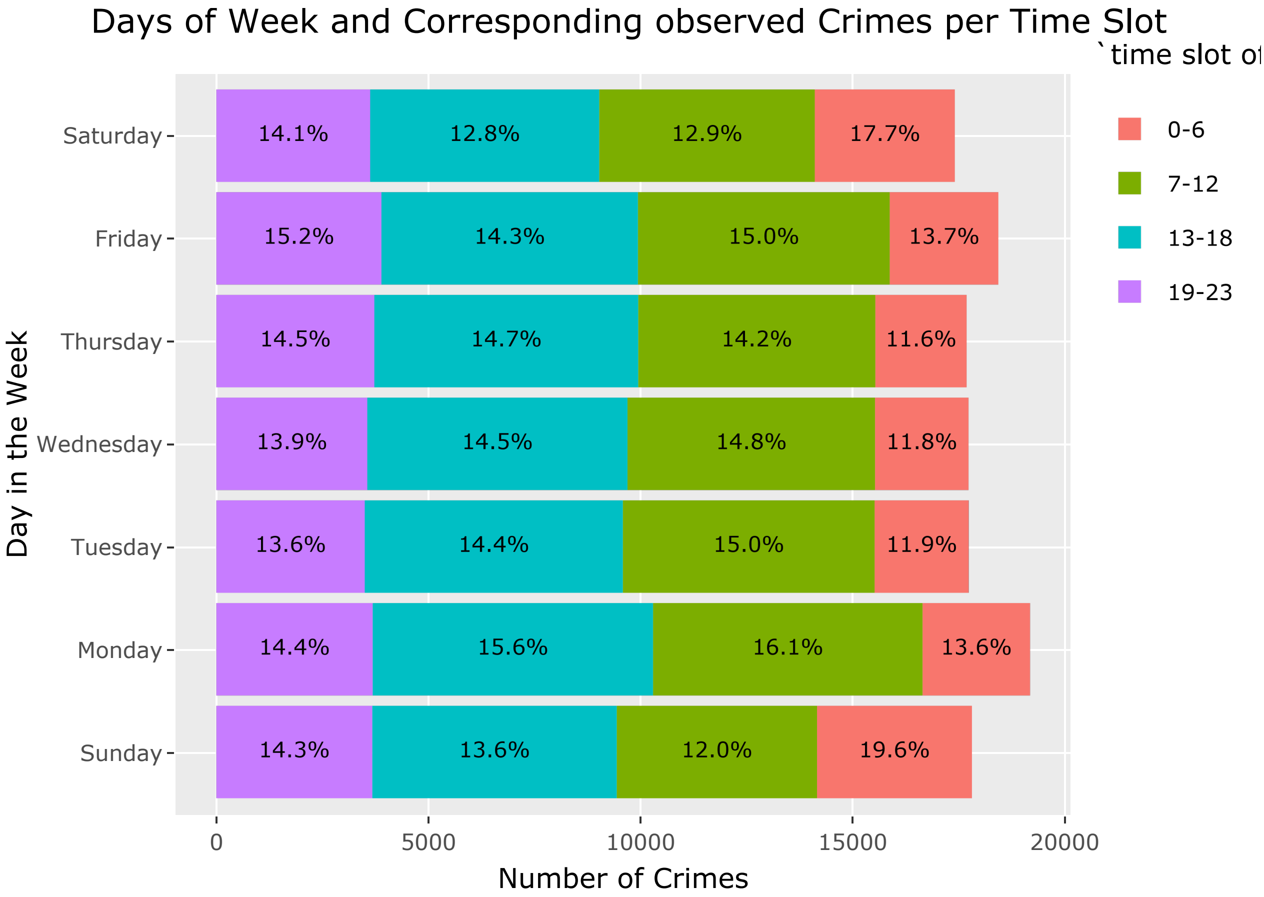
- The top 3 crimes mentioned earlier i.e, Larcency/Theft offences , Assault offences and Property related crimes seems to have equal number of occurences during late night slight (00:00 to 6:00). This means lesser possibilities of occurence of Larcency/Theft offences during this time slot in comparison to Assault offences and Property related crimes.
- Another interesting observation is - Burglary /Breaking and entering is relatively very low during evening times (19:00 to 6:00) in comparison to day time (7:00 to 18:00) - This observation also related directly to the above observation



3.4 How crimes are related to days of the week and time slot of occurence

Following can be interpreted from the graph

- Late night (00:00 to 6:00) on weekends has higher occurence of crime in comparison to that of weekdays - possible reason could be higher number of people staying outdoors on weekends leading to higher chances of crimes.
- Monday and Friday seems to have highest number of crimes in the week.
- Day time (7:00 to 18:00) on weekends seems to have lesser frequency of crimes in comparison to that of weekdays - possible reason could be fewer number of people being outdoors leading to lesser chances of crimes.



4. Feature Selection for Predictive Analysis

- Two types of problems were designed for the given dataset : 1. Regression, 2. Classification
- Objectives :
-
    1. Evaluate attributes and their correlation
-
    2. Generate datasets dallas_crime_rate(regression) and dallas_crime_type(classification) suitable to solve the designed problems

4.1 Feature Selection for Regression problem

4.1.1 Following chunk performs the generation of dataframe dallas_crime_rate

4.1.1.1 Following steps taken to generate dallas_crime_rate dataframe

- All steps to generate dallas_incidents dataframe
- group by 'Division', 'rounded time' and 'week number of the day' and summarize the frequency of records to new attribute 'freq'

4.1.1.2 Following steps were completed done during preliminary features selection phase :

- Reduction of Location Type attribute to 4 categorical values from 73

```
type_location_bins<-tribble(
  ~Sub,~LocationType,~LocNum,
  "Highway, Street, Alley ETC","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",
1,
  "Airport - Love Field","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Medical Facility","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Financial Institution","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Bank/Savings And Loan","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Construction Site","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Religious Institution","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Government Facility","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Church/Synagogue/Temple/Mosque","Public Locations (Hospitals/Parks/ATMs/Streets/School
s)",1,
  "Shopping Mall","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Parking Lot (Park)","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Airport - All Others","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Jail/Prison/Penitentiary/Corrections Fac","Public Locations (Hospitals/Parks/ATMs/Stre
ets/Schools)",1,
  "School - Elementary/Secondary","Public Locations (Hospitals/Parks/ATMs/Streets/School
s)",1,
  "ATM Separate from Bank","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Daycare Facility","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Military Installation","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Outdoor Area Public/Private","Public Locations (Hospitals/Parks/ATMs/Streets/School
s)",1,
  "Amusement Park","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "PHARM","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Park","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Arena/Stadium/Fairgrounds/Coliseum","Public Locations (Hospitals/Parks/ATMs/Streets/Sc
hools)",1,
  "School/Daycare","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "City Park/Rec/Tennis/Golf/Trail","Public Locations (Hospitals/Parks/ATMs/Streets/Schoo
ls)",1,
  "School - College/University","Public Locations (Hospitals/Parks/ATMs/Streets/School
s)",1,
  "Government/Public Building","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",
1,
  "Community/ Recreation Center","Public Locations (Hospitals/Parks/ATMs/Streets/School
s)",1,
  "Dock/Wharf/Freight/Modal Terminal","Public Locations (Hospitals/Parks/ATMs/Streets/Sch
ools)",1,
  "Shelter - Mission/Homeless","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",
1,
  "School/College","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Camp/Campground","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Lake/Waterway/Beach","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",1,
  "Apartment Complex/Building","Private/Individual Locations (Residences and others)",4,
  "Convenience Store","Commercial Establishments (Restaurants/Stores)",2,
  "Gas or Service Station","Commercial Establishments (Restaurants/Stores)",2,
  "Bar/NightClub/DanceHall ETC.","Commercial Establishments (Restaurants/Stores)",2,
  "Parking Lot (Apartment)","Private/Individual Locations (Residences and others)",4,
  "Entertainment/Sports Venue","Commercial Establishments (Restaurants/Stores)",2,
  "Parking (Business)","Commercial Establishments (Restaurants/Stores)",2,
  "Storage Facility","Commercial Establishments (Restaurants/Stores)",2,
  "Single Family Residence - Vacant","Private/Individual Locations (Residences and other
```

```
s)",4,
  "Department/Discount Store","Public Locations (Hospitals/Parks/ATMs/Streets/Schools)",
1,
  "Condominium/Townhome Residence","Private/Individual Locations (Residences and other
s)",4,
  "Shopping Mall","Commercial Establishments (Restaurants/Stores)",2,
  "Grocery/Supermarket","Commercial Establishments (Restaurants/Stores)",2,
  "Specialty Store (In a Specific Item)","Commercial Establishments (Restaurants/Store
s)",2,
  "Personal Services","Private/Individual Locations (Residences and others)",4,
  "Tribal Lands","Private/Individual Locations (Residences and others)",4,
  "Restaurant/Food Service/TABC Location","Commercial Establishments (Restaurants/Store
s)",2,
  "Apartment Residence","Private/Individual Locations (Residences and others)",4,
  "Single Family Residence - Occupied","Private/Individual Locations (Residences and othe
rs)",4,
  "Retail Store","Commercial Establishments (Restaurants/Stores)",2,
  "Business Office","Commercial Establishments (Restaurants/Stores)",2,
  "Motor Vehicle","Private/Individual Locations (Residences and others)",4,
  "Commercial Property Occupied/Vacant","Commercial Establishments (Restaurants/Stores)",
2,
  "Industrial/Manufacturing","Commercial Establishments (Restaurants/Stores)",2,
  "Hotel/Motel/ETC","Commercial Establishments (Restaurants/Stores)",2,
  "Auto Dealership New/Used","Commercial Establishments (Restaurants/Stores)",2,
  "Liquor Store","Commercial Establishments (Restaurants/Stores)",2,
  "Rental Storage Facility","Commercial Establishments (Restaurants/Stores)",2,
  "Other","Others",3,
  "Cyberspace","Others",3
)

dallas_crime_rate<-dallas_incidents%>%
  select(`rounded time`,`week number of the day`,`Division`,`Type  Location`)%>%
  inner_join(type_location_bins, by = c("Type  Location" = "Sub"))%>%
  mutate(`rounded time`=factor(`rounded time`),`week number of the day`=factor(`week numb
er of the day`),`Division`=factor(`Division`),LocationType = factor(LocationType))%>%
  group_by(`Division`,`rounded time`,`week number of the day`)%>%
  summarise(freq=n())
```

4.1.2 Following chunk performs evaluation of variable importance using boxplot visualizations and anova method

4.1.2.1 Following attributes were considered(in various combinations) for the evaluation

- frequency ~ (ZipCode, LocationType, rounded time, NBIRS Category, time slot of occurence, week number of day and Division)

4.1.2.2 Following interpretations can be drawn from the tests

- There were too many outliers for ZipCode and NBIRS Category in boxplot - hence can be rejected
- Another reason for ignoring NBIRS category completely is - it is not a meaningful variable in describing the response variable
- There were not too many differences in mean level of ‘time slot of occurence’ in comparison to that of ‘rounded time’ from boxplot - hence time slot of occurence can be rejected

- Anova method shows combinations of ‘Division’, ‘rounded time’ and ‘week number of the day’ gave much favourable p-value (less than 0.05) than that of ‘LocationType’, ‘rounded time’ and ‘week number of day’. Thus we select the combination that has relatively lower p-value.

Please note : following chunk consists only those variables that were finally selected.

```
d_mod=lm(dallas_crime_rate$`freq` ~ dallas_crime_rate$`Division`+dallas_crime_rate$`week
number of the day`+dallas_crime_rate$`rounded time`, data = dallas_crime_rate)
summary(d_mod)
```

```
##
## Call:
## lm(formula = dallas_crime_rate$freq ~ dallas_crime_rate$Division +
##     dallas_crime_rate$`week number of the day` + dallas_crime_rate$`rounded time`,
##     data = dallas_crime_rate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.832 -10.635  -1.707   9.205  89.583
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                  56.45748    2.83540  19.912
## dallas_crime_rate$DivisionNORTHCENTRAL      -23.97619    1.76818 -13.560
## dallas_crime_rate$DivisionNORTHEAST          20.47619    1.76818  11.580
## dallas_crime_rate$DivisionNORTHWEST          -3.98810    1.76818  -2.255
## dallas_crime_rate$DivisionSOUTHCENTRAL        9.10714    1.76818   5.151
## dallas_crime_rate$DivisionSOUTHEAST          23.02381    1.76818  13.021
## dallas_crime_rate$DivisionSOUTHWEST          11.17857    1.76818   6.322
## dallas_crime_rate$`week number of the day`2   7.04762    1.76818   3.986
## dallas_crime_rate$`week number of the day`3  -0.19048    1.76818  -0.108
## dallas_crime_rate$`week number of the day`4   0.35714    1.76818   0.202
## dallas_crime_rate$`week number of the day`5  -0.08333    1.76818  -0.047
## dallas_crime_rate$`week number of the day`6   4.01786    1.76818   2.272
## dallas_crime_rate$`week number of the day`7  -1.74405    1.76818  -0.986
## dallas_crime_rate$`rounded time`1            -8.53061    3.27403  -2.606
## dallas_crime_rate$`rounded time`2           -13.04082    3.27403  -3.983
## dallas_crime_rate$`rounded time`3           -22.10204    3.27403  -6.751
## dallas_crime_rate$`rounded time`4           -30.83673    3.27403  -9.419
## dallas_crime_rate$`rounded time`5           -31.67347    3.27403  -9.674
## dallas_crime_rate$`rounded time`6           -20.18367    3.27403  -6.165
## dallas_crime_rate$`rounded time`7             1.06122    3.27403   0.324
## dallas_crime_rate$`rounded time`8            33.89796    3.27403  10.354
## dallas_crime_rate$`rounded time`9            49.32653    3.27403  15.066
## dallas_crime_rate$`rounded time`10           52.53061    3.27403  16.045
## dallas_crime_rate$`rounded time`11           50.38776    3.27403  15.390
## dallas_crime_rate$`rounded time`12           48.75510    3.27403  14.891
## dallas_crime_rate$`rounded time`13           51.06122    3.27403  15.596
## dallas_crime_rate$`rounded time`14           49.42857    3.27403  15.097
## dallas_crime_rate$`rounded time`15           51.42857    3.27403  15.708
## dallas_crime_rate$`rounded time`16           60.67347    3.27403  18.532
## dallas_crime_rate$`rounded time`17           60.34694    3.27403  18.432
## dallas_crime_rate$`rounded time`18           55.04082    3.27403  16.811
## dallas_crime_rate$`rounded time`19           43.75510    3.27403  13.364
## dallas_crime_rate$`rounded time`20           32.51020    3.27403   9.930
## dallas_crime_rate$`rounded time`21           24.65306    3.27403   7.530
## dallas_crime_rate$`rounded time`22           12.10204    3.27403   3.696
## dallas_crime_rate$`rounded time`23            4.89796    3.27403   1.496
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## dallas_crime_rate$DivisionNORTHCENTRAL       < 2e-16 ***
## dallas_crime_rate$DivisionNORTHEAST          < 2e-16 ***
## dallas_crime_rate$DivisionNORTHWEST         0.024292 *
## dallas_crime_rate$DivisionSOUTHCENTRAL      3.06e-07 ***
## dallas_crime_rate$DivisionSOUTHEAST          < 2e-16 ***
```

```
## dallas_crime_rate$DivisionSOUTHWEST         3.70e-10 ***
## dallas_crime_rate$`week number of the day`2 7.15e-05 ***
## dallas_crime_rate$`week number of the day`3 0.914233
## dallas_crime_rate$`week number of the day`4 0.839966
## dallas_crime_rate$`week number of the day`5 0.962418
## dallas_crime_rate$`week number of the day`6 0.023253 *
## dallas_crime_rate$`week number of the day`7 0.324170
## dallas_crime_rate$`rounded time`1           0.009293 **
## dallas_crime_rate$`rounded time`2           7.23e-05 ***
## dallas_crime_rate$`rounded time`3           2.34e-11 ***
## dallas_crime_rate$`rounded time`4            < 2e-16 ***
## dallas_crime_rate$`rounded time`5            < 2e-16 ***
## dallas_crime_rate$`rounded time`6           9.78e-10 ***
## dallas_crime_rate$`rounded time`7           0.745896
## dallas_crime_rate$`rounded time`8            < 2e-16 ***
## dallas_crime_rate$`rounded time`9            < 2e-16 ***
## dallas_crime_rate$`rounded time`10           < 2e-16 ***
## dallas_crime_rate$`rounded time`11           < 2e-16 ***
## dallas_crime_rate$`rounded time`12           < 2e-16 ***
## dallas_crime_rate$`rounded time`13           < 2e-16 ***
## dallas_crime_rate$`rounded time`14           < 2e-16 ***
## dallas_crime_rate$`rounded time`15           < 2e-16 ***
## dallas_crime_rate$`rounded time`16           < 2e-16 ***
## dallas_crime_rate$`rounded time`17           < 2e-16 ***
## dallas_crime_rate$`rounded time`18           < 2e-16 ***
## dallas_crime_rate$`rounded time`19           < 2e-16 ***
## dallas_crime_rate$`rounded time`20           < 2e-16 ***
## dallas_crime_rate$`rounded time`21          1.03e-13 ***
## dallas_crime_rate$`rounded time`22          0.000229 ***
## dallas_crime_rate$`rounded time`23          0.134930
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.21 on 1140 degrees of freedom
## Multiple R-squared:  0.8252, Adjusted R-squared:  0.8199
## F-statistic: 153.8 on 35 and 1140 DF,  p-value: < 2.2e-16
```

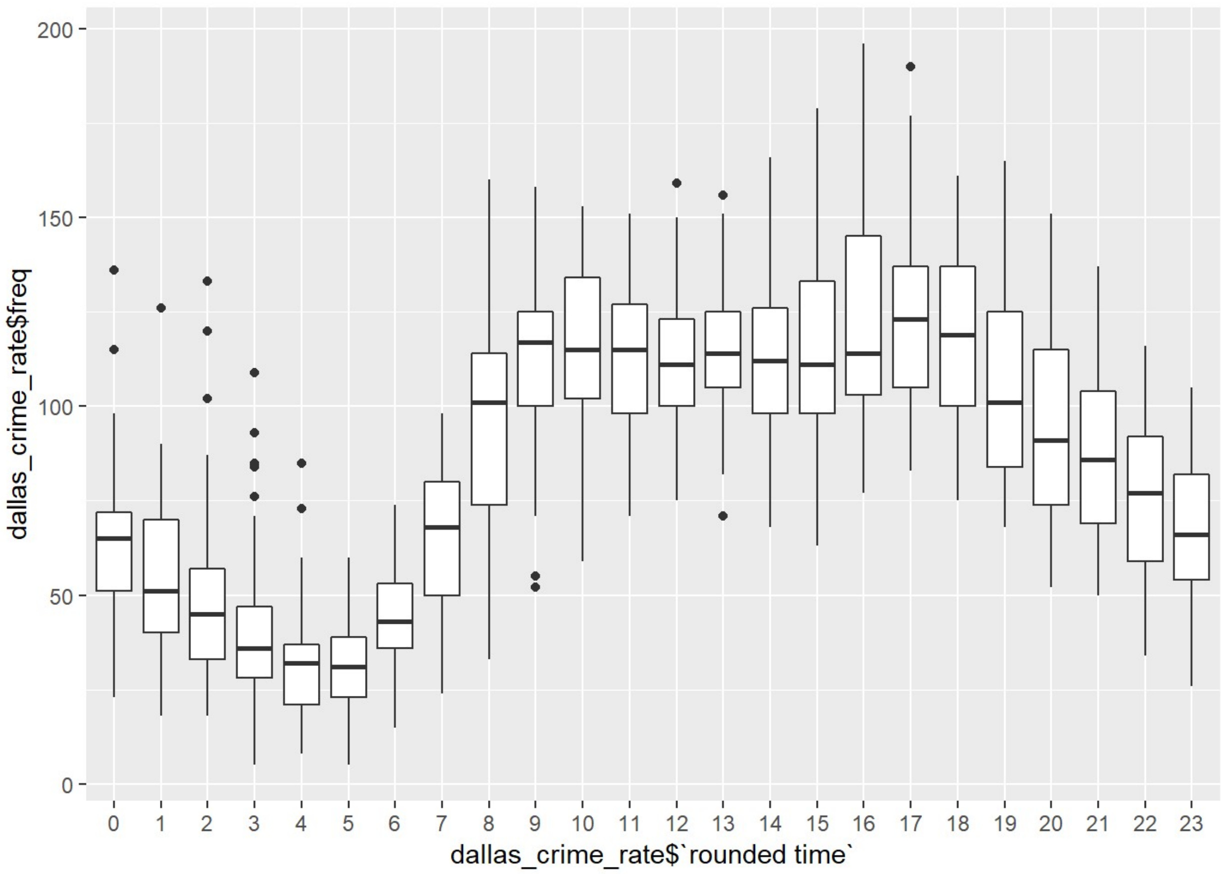
```
anova(d_mod)
```

```
## Analysis of Variance Table
##
## Response: dallas_crime_rate$freq
##                                               Df  Sum Sq Mean Sq F value
## dallas_crime_rate$Division                     6  262873   43812 166.826
## dallas_crime_rate$`week number of the day`     6    9473    1579   6.012
## dallas_crime_rate$`rounded time`              23 1141310   49622 188.949
## Residuals                                   1140  299390     263
##                                               Pr(>F)
## dallas_crime_rate$Division                  < 2.2e-16 ***
## dallas_crime_rate$`week number of the day` 3.333e-06 ***
## dallas_crime_rate$`rounded time`            < 2.2e-16 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

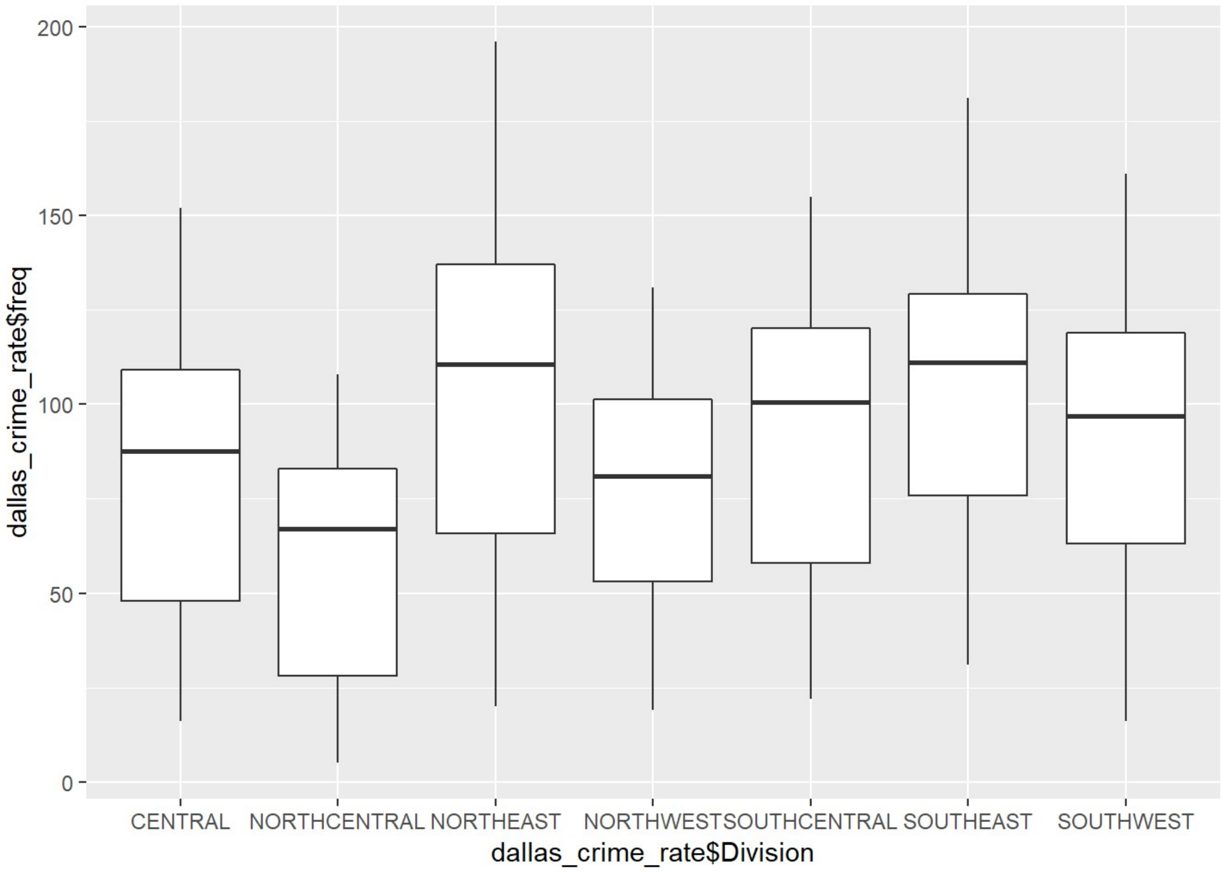
```
#confint(d_mod)
print(as_tibble(dallas_crime_rate))
```

```
## # A tibble: 1,176 x 4
##    Division `rounded time` `week number of the day`  freq
##    <fct>    <fct>          <fct>                    <int>
##  1 CENTRAL  0              1                           68
##  2 CENTRAL  0              2                           51
##  3 CENTRAL  0              3                           30
##  4 CENTRAL  0              4                           44
##  5 CENTRAL  0              5                           42
##  6 CENTRAL  0              6                           66
##  7 CENTRAL  0              7                           76
##  8 CENTRAL  1              1                           88
##  9 CENTRAL  1              2                           42
## 10 CENTRAL  1              3                           36
## # ... with 1,166 more rows
```

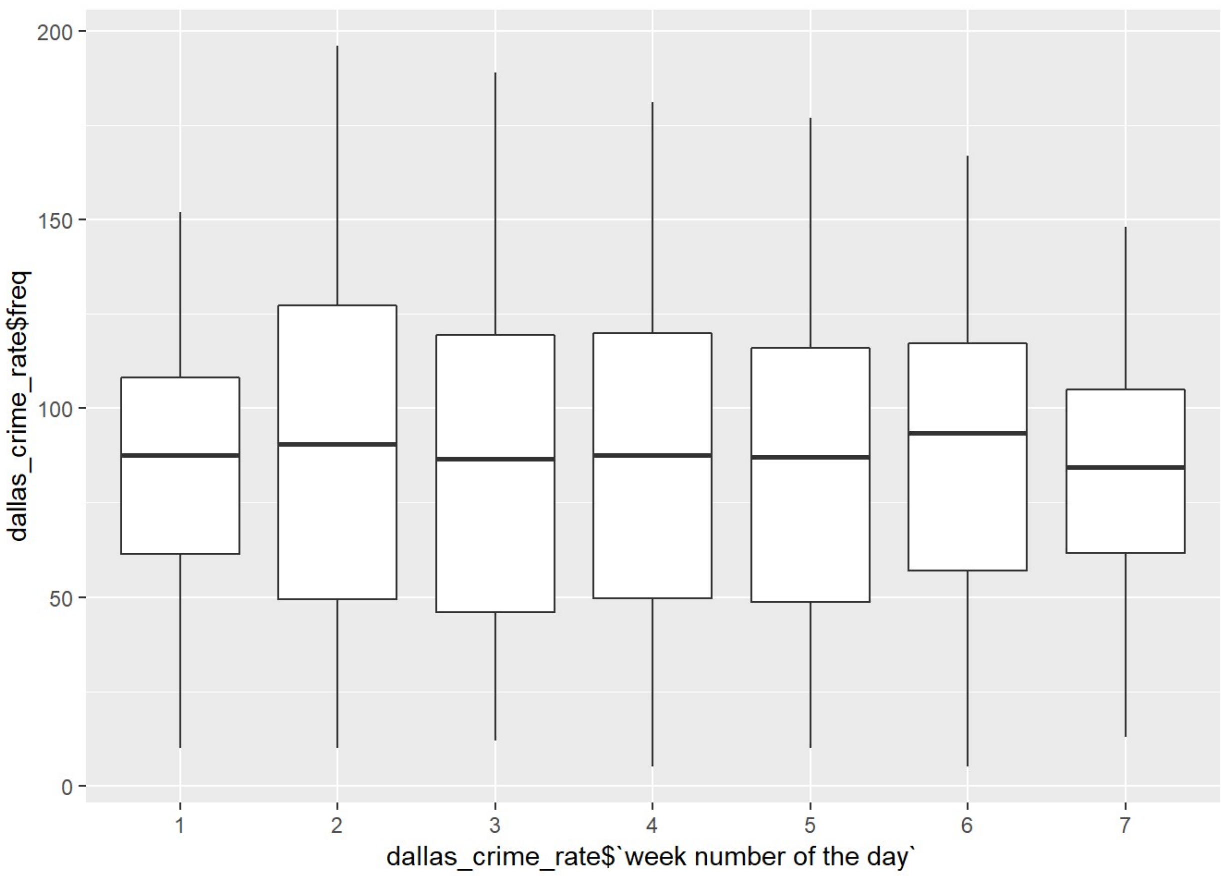
```
ggplot(dallas_crime_rate, aes(x=dallas_crime_rate$`rounded time`, y=dallas_crime_rate$`fr
eq`)) + geom_boxplot()
```




```
ggplot(dallas_crime_rate, aes(x=dallas_crime_rate$`Division`, y=dallas_crime_rate$`freq
`)) + geom_boxplot()
```




```
ggplot(dallas_crime_rate, aes(x=dallas_crime_rate$`week number of the day`, y=dallas_crim
e_rate$`freq`)) + geom_boxplot()
```



4.2 Feature Selection for Classification problem

4.2.1 Following chunk performs the generation of dataframe dallas_crime_type

4.2.1.1 Following steps taken to generate dallas_crime_type dataframe

- All steps to generate dallas_incidents dataframe
- group by ‘Division’, ‘rounded time’ and ‘week number of the day’ and summarize the frequency of records to new attribute ‘freq’

4.2.1.2 Following steps were completed done during preliminary features selection phase :

- Reduction of Location Type attribute to 4 categorical values from 73
- Usage of ‘NIBRS Crime Category’(28 categorical values) instead of ‘Category Type’(903 categorical values) attribute
- Further reduction of ‘NIBRS Crime category’ to consist 8 categorical values in new attribute ‘Category’
- Usage of ‘Division’(13 categorical values) instead of ‘Zip Code’ (122 categorical values)
- Cleaning the ‘Division’ attribute - bringing values to consistent format, thus reducing to 8 categorical values

```
category_bins = tribble(
  ~Sub,~Category,~CatNum,
  "BRIBERY","ALL OTHER OFFENSES",1,
  "HUMAN TRAFFICKING","ALL OTHER OFFENSES",1,
  "PORNOGRAPHY/ OBSCENE MATERIAL","ALL OTHER OFFENSES",1,
  "FAMILY OFFENSES, NONVIOLENT","ALL OTHER OFFENSES",1,
  "DRUG/ NARCOTIC VIOLATIONS","ALL OTHER OFFENSES",1,
  "ARSON","DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY",2,
  "DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY","DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERT
Y",2,
  "TRAFFIC VIOLATION - NON HAZARDOUS","TRAFFIC VIOLATION",3,
  "DRIVING UNDER THE INFLUENCE","TRAFFIC VIOLATION",3,
  "TRAFFIC VIOLATION - HAZARDOUS","TRAFFIC VIOLATION",3,
  "ROBBERY","BURGLARY/ BREAKING & ENTERING",4,
  "MOTOR VEHICLE THEFT","LARCENY/ THEFT OFFENSES",5,
  "KIDNAPPING/ ABDUCTION","ASSAULT OFFENSES",6,
  "ANIMAL OFFENSES","ASSAULT OFFENSES",6,
  "HOMICIDE OFFENSES","ASSAULT OFFENSES",6,
  "WEAPON LAW VIOLATIONS","ASSAULT OFFENSES",6,
  "KIDNAPPING/ ABDUCTION","ASSAULT OFFENSES",6,
  "HOMICIDE OFFENSES","ASSAULT OFFENSES",6,
  "EMBEZZELMENT","FRAUD OFFENSES",7,
  "COUNTERFEITING / FORGERY","FRAUD OFFENSES",7,
  "DRUNKENNESS","DRUNKENNESS/TRESPASSING/NUISANCE",8,
  "DISORDERLY CONDUCT","DRUNKENNESS/TRESPASSING/NUISANCE",8,
  "LIQUOR LAW VIOLATIONS","DRUNKENNESS/TRESPASSING/NUISANCE",8,
  "TRESPASS OF REAL PROPERTY","DRUNKENNESS/TRESPASSING/NUISANCE",8
  )

dallas_crime_type<-dallas_incidents%>%
  inner_join(category_bins, by = c("NIBRS Crime Category" = "Sub"))%>%
  select(`Division`,`week of the day`,`time slot of occurence`,`Category`,`NIBRS Crime Ca
tegory`,`rounded time`)%>%
  filter( !is.na(`Division`) & !is.na(`week of the day`) & !is.na(`time slot of occurence
`) & !is.na(`Category`) & !is.na(`NIBRS Crime Category`) & !is.na(`rounded time`))%>%
  mutate(`Category`=factor(Category),`Division`=factor(`Division`),`week of the day`=fact
or(`week of the day`),`time slot of occurence`=factor(`time slot of occurence`),`NIBRS Cr
ime Category`=factor(`NIBRS Crime Category`),`rounded time`=factor(`rounded time`))
```

4.2.2 Following chunk performs evaluation of variable importance using chi-square test

- Null hypothesis : There is no association between 2 variables

4.2.2.1 Following attributes were considered(in various combinations) for the evaluation

- Crime Category ~ (rounded time, time slot of occurence, week of the day and Division) #### 4.2.2.2 Following interpretations can be drawn from the tests
- both combinations of (rounded time, week of the day and Division) and (time slot of occurence, week of the day and Division) passes chi-square test. However, we pick the combination that contains time slot of occurence as it has only factor levels in comparison to that of 24 in rounded time for building model with better accuracy.
- Another reason to reject rounded time is the duration of traning the model is higher.

```
#Following code does not include `NIBRS Crime Category` as it had lower significance than `Category`
tbl_dallas_zrwt<-dallas_crime_type%>%
  categorize(`Category`,`Division`,`week of the day`,`time slot of occurence`)

if(chisq.test(table(tbl_dallas_zrwt),simulate.p.value = TRUE)[[3]]<0.05){
    print("p-value is significant  - Null Hypothesis rejected")
  }else{
    print("Null hypothesis sustained - no significant association observed")
  }
```

```
## [1] "p-value is significant  - Null Hypothesis rejected"
```

5. Model Training

- Both classification and regression models are trained using the explanatory variables that had highest importance(from our inferential statistics tools) in predicting the future outcome
- Divide the data into training(75%) and testing(25%)
- Parameters auto-tune length set to 10
- Resampling method chosen : Cross validation of chunk size 10

5.1 Regression Model

- Using the explanatory variables `week of the day` , `rounded time` and `division` for predicting the `crime frequency`
- generating dummy variables of the dataframe
- Using the log transformation for the response variable `crime frequency` for normalization
- Using linear regression model and gradient boosting algorithms.
- preprocess the target attribute to scale and center
- Using RMSE metrics for cross validation evaluation

```
dmy <- dummyVars(freq  ~., data = dallas_crime_rate,fullRank = T)
reg_train_transformed <- data.frame(predict(dmy, newdata = dallas_crime_rate))
reg_train_transformed$`freq`<-(dallas_crime_rate$freq)
reg_intrain <- createDataPartition(y = (dallas_crime_rate$freq), p= 0.75, list = FALSE)
reg_training <- reg_train_transformed[reg_intrain,]
reg_testing <- reg_train_transformed[-reg_intrain,]
set.seed(100)

trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3,verboseIter = FALS
E)

  lm_reg_model <- train(`freq` ~., data = reg_training, method="lm", metric="RMSE",
                        preProcess=c("BoxCox"), tuneLength = 10,
                      trControl=trctrl)

  gbm_reg_model <- train(`freq` ~., data = reg_training, method="gbm", metric="RMSE",
                        preProcess=c("BoxCox"), tuneLength = 10,
                      trControl=trctrl)


  save(lm_reg_model, file = "lm_regression.rda")
  save(gbm_reg_model, file = "gbm_regression.rda")
```

5.2 Classification Model

- Using the explanatory varibales `week of the day` , `division` and `time slot of occurence` for predicting `Category` (crime type)
- generating dummy variables of the dataframe
- Using the algorithms SVM,Random Forest and Naive Bayes.
- preprocess the target attribute to scale and center

```
dmy <- dummyVars(Category  ~., data = dallas_crime_type,fullRank = T)
cls_train_transformed <- data.frame(predict(dmy, newdata = dallas_crime_type))

cls_train_transformed$`Category`<-dallas_crime_type$Category



cls_intrain <- createDataPartition(y = dallas_crime_type$`Category`, p= 0.75, list = FALS
E)
cls_training <- cls_train_transformed[cls_intrain,]
cls_testing <- cls_train_transformed[-cls_intrain,]

trctrl <- trainControl(method = "repeatedcv",number = 10, repeats = 3,verboseIter = FALS
E)
 set.seed(111)
   svm_Linear_cls_model <- train(`Category` ~., data = cls_training, method = "svmLinea
r",
                    trControl=trctrl,
                    preProcess = c("center", "scale"),
                    tuneLength = 10)
   random_Forest_cls_model<-train(`Category` ~., data = cls_training, method = "rf",
                    trControl=trctrl,
                    preProcess = c("center", "scale"),
                    tuneLength = 10)
   naive_bayes_cls_model<-train(`Category` ~., data = cls_training, method = "nb",
                    trControl=trctrl,
                    preProcess = c("center", "scale"),
                    tuneLength = 10)

 save(svm_Linear_cls_model, file = "svm_classification.rda")
 save(random_Forest_cls_model, file = "rf_classification.rda")
 save(naive_bayes_cls_model, file = "nb_classification.rda")
```

6 Model Prediction and Evaluation

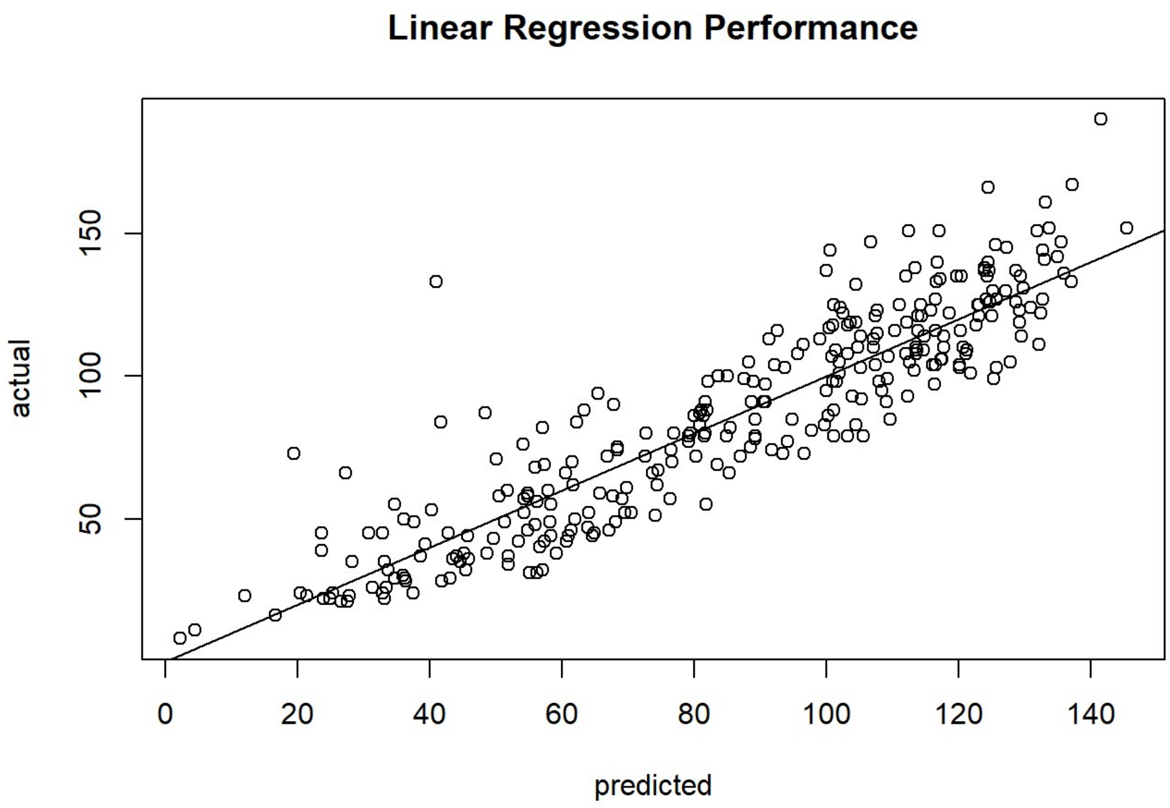
6.1 Regression Model

- Evaluation using `Predicted v/s Actual Dataset Plot` and `RMSE Mean`

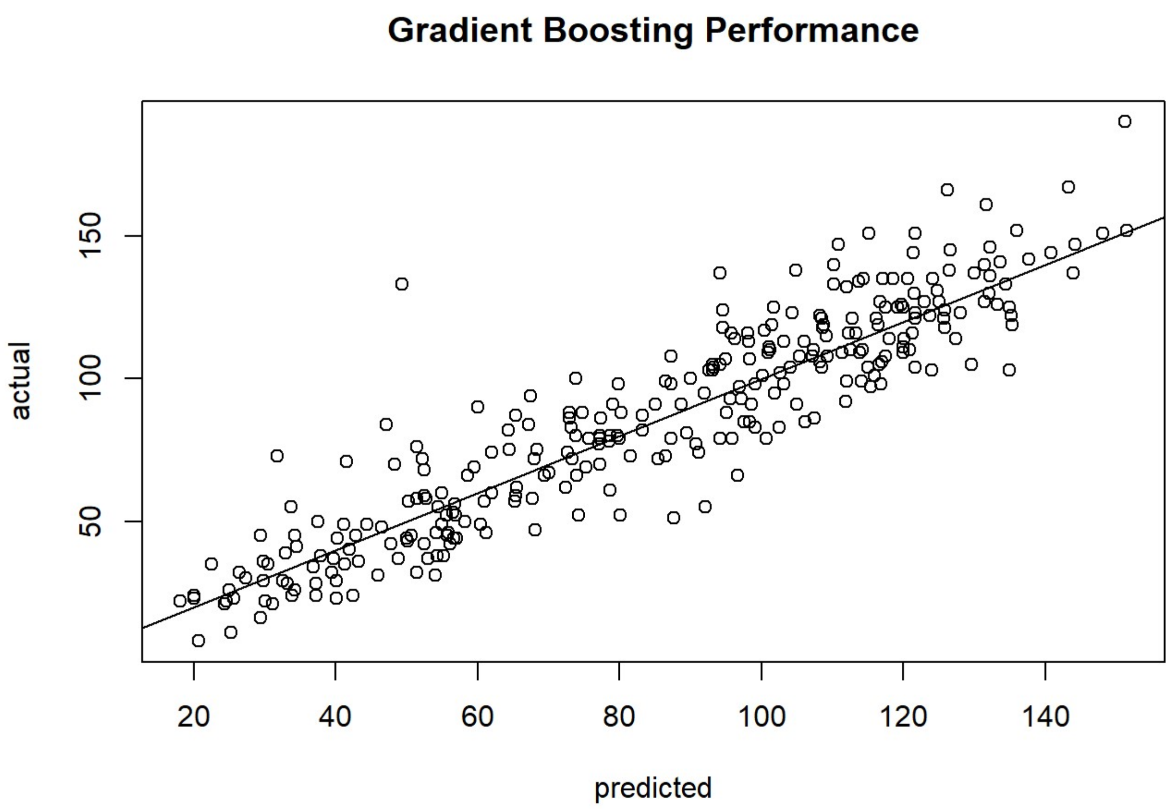
```
load("lm_regression.rda")
load("gbm_regression.rda")
lm_test_pred <- predict(lm_reg_model, newdata = reg_testing)
gbm_test_pred <- predict(gbm_reg_model, newdata = reg_testing)
print("Linear Regression Model Performance :")
```

```
## [1] "Linear Regression Model Performance :"
```

```
plot(as_data_frame(lm_test_pred)$value,as_data_frame(reg_testing$freq)$value,
      xlab="predicted",ylab="actual",main="Linear Regression Performance")
 abline(a=0,b=1)
```




```
plot(as_data_frame(gbm_test_pred)$value,as_data_frame(reg_testing$freq)$value,
     xlab="predicted",ylab="actual",main="Gradient Boosting Performance")
abline(a=0,b=1)
```



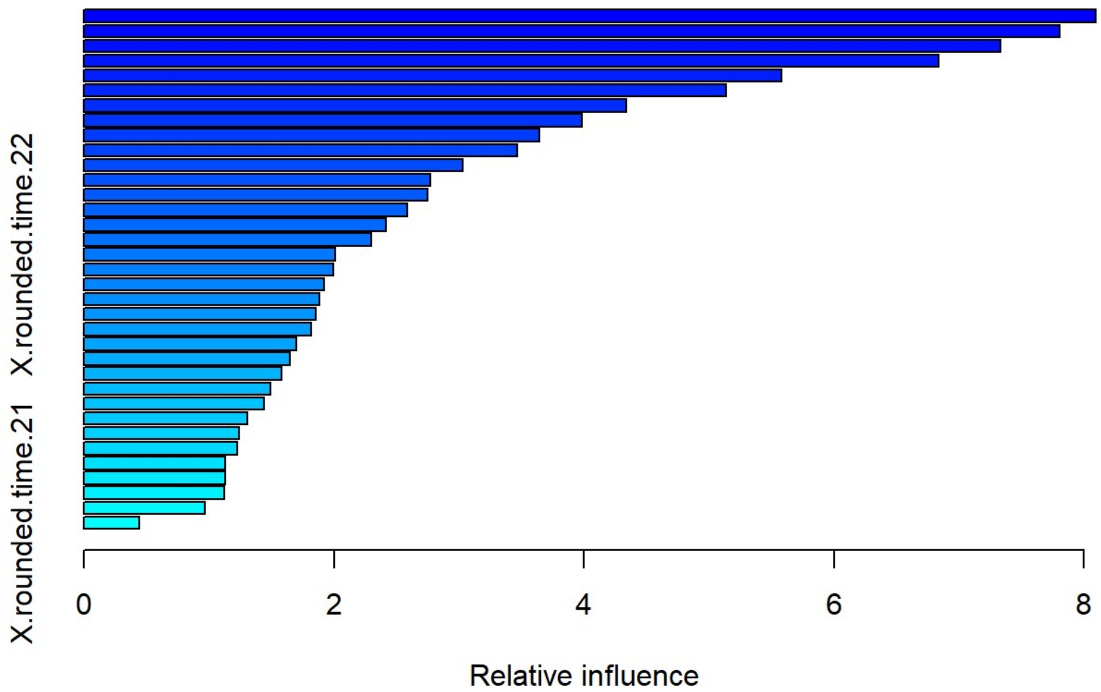
```
summary(lm_reg_model)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.248 -10.601  -1.928   9.279  79.316
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  56.00186    3.49422  16.027  < 2e-16 ***
## Division.NORTHCENTRAL       -23.04897    2.05862 -11.196  < 2e-16 ***
## Division.NORTHEAST           20.46433    2.04564  10.004  < 2e-16 ***
## Division.NORTHWEST           -3.48907    2.07106  -1.685 0.092419 .
## Division.SOUTHCENTRAL         9.45627    2.08776   4.529 6.76e-06 ***
## Division.SOUTHEAST           24.40195    2.04291  11.945  < 2e-16 ***
## Division.SOUTHWEST           12.06414    2.12115   5.688 1.77e-08 ***
## X.rounded.time.1             -8.68765    4.04146  -2.150 0.031867 *
## X.rounded.time.2            -14.94847    3.93178  -3.802 0.000154 ***
## X.rounded.time.3            -26.31831    3.93757  -6.684 4.21e-11 ***
## X.rounded.time.4            -33.07186    3.88676  -8.509  < 2e-16 ***
## X.rounded.time.5            -33.41327    4.03760  -8.276 4.94e-16 ***
## X.rounded.time.6            -20.88839    3.93587  -5.307 1.42e-07 ***
## X.rounded.time.7             -0.07135    3.90904  -0.018 0.985442
## X.rounded.time.8             30.78169    3.98421   7.726 3.13e-14 ***
## X.rounded.time.9             46.04906    3.98081  11.568  < 2e-16 ***
## X.rounded.time.10            49.96062    3.86344  12.932  < 2e-16 ***
## X.rounded.time.11            47.53415    3.86531  12.298  < 2e-16 ***
## X.rounded.time.12            47.26274    3.88568  12.163  < 2e-16 ***
## X.rounded.time.13            47.06667    4.03846  11.655  < 2e-16 ***
## X.rounded.time.14            46.58521    4.06859  11.450  < 2e-16 ***
## X.rounded.time.15            50.88558    4.04421  12.582  < 2e-16 ***
## X.rounded.time.16            59.39989    3.93298  15.103  < 2e-16 ***
## X.rounded.time.17            56.13190    4.00788  14.005  < 2e-16 ***
## X.rounded.time.18            54.50645    3.86740  14.094  < 2e-16 ***
## X.rounded.time.19            44.18055    4.06659  10.864  < 2e-16 ***
## X.rounded.time.20            31.99302    3.82760   8.359 2.59e-16 ***
## X.rounded.time.21            22.09977    4.25213   5.197 2.53e-07 ***
## X.rounded.time.22             9.76386    3.88584   2.513 0.012166 *
## X.rounded.time.23             3.28520    3.93066   0.836 0.403510
## X.week.number.of.the.day.2    8.89571    2.02620   4.390 1.27e-05 ***
## X.week.number.of.the.day.3    1.72069    2.02403   0.850 0.395495
## X.week.number.of.the.day.4    2.36545    2.02747   1.167 0.243658
## X.week.number.of.the.day.5    1.36593    2.05726   0.664 0.506898
## X.week.number.of.the.day.6    4.59968    2.01465   2.283 0.022670 *
## X.week.number.of.the.day.7    1.12378    2.04223   0.550 0.582277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.36 on 848 degrees of freedom
## Multiple R-squared:  0.8257, Adjusted R-squared:  0.8185
## F-statistic: 114.8 on 35 and 848 DF,  p-value: < 2.2e-16
```

```
print("Gradient Boosting Model Performance :")
```

```
## [1] "Gradient Boosting Model Performance :"
```

```
summary(gbm_reg_model)
```



```
##                                                       var   rel.inf
## X.rounded.time.3                         X.rounded.time.3 8.0920359
## X.rounded.time.4                         X.rounded.time.4 7.8039801
## Division.NORTHCENTRAL               Division.NORTHCENTRAL 7.3338245
## X.rounded.time.5                         X.rounded.time.5 6.8396764
## X.rounded.time.2                         X.rounded.time.2 5.5774770
## X.rounded.time.6                         X.rounded.time.6 5.1342691
## X.rounded.time.1                         X.rounded.time.1 4.3370913
## Division.SOUTHEAST                     Division.SOUTHEAST 3.9861202
## X.rounded.time.16                       X.rounded.time.16 3.6474665
## Division.NORTHEAST                     Division.NORTHEAST 3.4700320
## X.rounded.time.7                         X.rounded.time.7 3.0306502
## X.rounded.time.8                         X.rounded.time.8 2.7717700
## X.rounded.time.18                       X.rounded.time.18 2.7477211
## X.rounded.time.17                       X.rounded.time.17 2.5871394
## X.rounded.time.10                       X.rounded.time.10 2.4198187
## X.rounded.time.23                       X.rounded.time.23 2.2963296
## X.rounded.time.22                       X.rounded.time.22 2.0079331
## X.rounded.time.15                       X.rounded.time.15 1.9982201
## X.rounded.time.9                         X.rounded.time.9 1.9194459
## Division.NORTHWEST                     Division.NORTHWEST 1.8838930
## X.week.number.of.the.day.2 X.week.number.of.the.day.2 1.8569564
## X.rounded.time.19                       X.rounded.time.19 1.8200742
## X.rounded.time.11                       X.rounded.time.11 1.7004011
## X.rounded.time.12                       X.rounded.time.12 1.6458098
## X.rounded.time.14                       X.rounded.time.14 1.5822248
## Division.SOUTHWEST                     Division.SOUTHWEST 1.4965695
## X.rounded.time.13                       X.rounded.time.13 1.4409334
## X.rounded.time.20                       X.rounded.time.20 1.3084281
## X.week.number.of.the.day.6 X.week.number.of.the.day.6 1.2400862
## X.week.number.of.the.day.7 X.week.number.of.the.day.7 1.2298774
## X.week.number.of.the.day.3 X.week.number.of.the.day.3 1.1326219
## X.week.number.of.the.day.4 X.week.number.of.the.day.4 1.1295638
## Division.SOUTHCENTRAL               Division.SOUTHCENTRAL 1.1220124
## X.week.number.of.the.day.5 X.week.number.of.the.day.5 0.9680776
## X.rounded.time.21                       X.rounded.time.21 0.4414690
```

```
print("Comparison on performance : ")
```

```
## [1] "Comparison on performance : "
```

```
res <- resamples(list(lm = lm_reg_model, gbm = gbm_reg_model))
summary(res)
```

```
##
## Call:
## summary.resamples(object = res)
##
## Models: lm, gbm
## Number of resamples: 30
##
## MAE
##          Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lm  10.824641 11.97518 13.17711 12.85080 13.54861 14.50971    0
## gbm  8.884341 10.90838 11.28763 11.44677 11.95069 14.55138    0
##
## RMSE
##         Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lm  13.55192 15.22779 16.94521 16.68425 17.87928 19.07530    0
## gbm 11.21397 13.74198 14.80004 14.93577 15.61135 19.66246    0
##
## Rsquared
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm  0.7372684 0.7891697 0.8106980 0.8135152 0.8475670 0.8696961    0
## gbm 0.7378512 0.8299602 0.8584688 0.8490330 0.8742155 0.9136934    0
```

6.2 Classification Model

- Evaluation using `Confusion Matrix`

```
load("svm_classification.rda")
load("rf_classification.rda")
load("nb_classification.rda")
test_pred_svm <- predict(svm_Linear_cls_model, newdata = cls_testing)
test_pred_rf <- predict(random_Forest_cls_model, newdata = cls_testing)
test_pred_nb <- predict(naive_bayes_cls_model, newdata = cls_testing)
print("SVM Model Performance :")
```

```
## [1] "SVM Model Performance :"
```

```
confusionMatrix(test_pred_svm, factor(cls_testing$`Category`))
```

```
## Confusion Matrix and Statistics
##
##                                               Reference
## Prediction                                     ALL OTHER OFFENSES
##   ALL OTHER OFFENSES                                          234
##   ASSAULT OFFENSES                                              0
##   BURGLARY/ BREAKING & ENTERING                                 0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                    0
##   DRUNKENNESS/TRESPASSING/NUISANCE                              0
##   FRAUD OFFENSES                                                0
##   LARCENY/ THEFT OFFENSES                                       0
##   TRAFFIC VIOLATION                                             0
##                                               Reference
## Prediction                                     ASSAULT OFFENSES
##   ALL OTHER OFFENSES                                        0
##   ASSAULT OFFENSES                                        150
##   BURGLARY/ BREAKING & ENTERING                             0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                0
##   DRUNKENNESS/TRESPASSING/NUISANCE                          0
##   FRAUD OFFENSES                                            0
##   LARCENY/ THEFT OFFENSES                                   0
##   TRAFFIC VIOLATION                                         0
##                                               Reference
## Prediction                                     BURGLARY/ BREAKING & ENTERING
##   ALL OTHER OFFENSES                                                       0
##   ASSAULT OFFENSES                                                         0
##   BURGLARY/ BREAKING & ENTERING                                         2001
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                               0
##   DRUNKENNESS/TRESPASSING/NUISANCE                                         0
##   FRAUD OFFENSES                                                           0
##   LARCENY/ THEFT OFFENSES                                                  0
##   TRAFFIC VIOLATION                                                        0
##                                               Reference
## Prediction                                     DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERT
Y
##   ALL OTHER OFFENSES
0
##   ASSAULT OFFENSES
0
##   BURGLARY/ BREAKING & ENTERING
0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                                          421
5
##   DRUNKENNESS/TRESPASSING/NUISANCE
0
##   FRAUD OFFENSES
0
##   LARCENY/ THEFT OFFENSES
0
##   TRAFFIC VIOLATION
0
##                                               Reference
## Prediction                                     DRUNKENNESS/TRESPASSING/NUISANCE
##   ALL OTHER OFFENSES                                                          0
##   ASSAULT OFFENSES                                                            2
```

```
##   BURGLARY/ BREAKING & ENTERING                                            0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                               0
##   DRUNKENNESS/TRESPASSING/NUISANCE                                       135
##   FRAUD OFFENSES                                                           0
##   LARCENY/ THEFT OFFENSES                                                  0
##   TRAFFIC VIOLATION                                                        0
##                                               Reference
## Prediction                                     FRAUD OFFENSES
##   ALL OTHER OFFENSES                                        0
##   ASSAULT OFFENSES                                          0
##   BURGLARY/ BREAKING & ENTERING                             0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                0
##   DRUNKENNESS/TRESPASSING/NUISANCE                          0
##   FRAUD OFFENSES                                          153
##   LARCENY/ THEFT OFFENSES                                   0
##   TRAFFIC VIOLATION                                         0
##                                               Reference
## Prediction                                     LARCENY/ THEFT OFFENSES
##   ALL OTHER OFFENSES                                                 0
##   ASSAULT OFFENSES                                                   0
##   BURGLARY/ BREAKING & ENTERING                                      0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                         0
##   DRUNKENNESS/TRESPASSING/NUISANCE                                   0
##   FRAUD OFFENSES                                                     0
##   LARCENY/ THEFT OFFENSES                                         1133
##   TRAFFIC VIOLATION                                                  0
##                                               Reference
## Prediction                                     TRAFFIC VIOLATION
##   ALL OTHER OFFENSES                                           0
##   ASSAULT OFFENSES                                             0
##   BURGLARY/ BREAKING & ENTERING                                0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                   0
##   DRUNKENNESS/TRESPASSING/NUISANCE                             0
##   FRAUD OFFENSES                                               0
##   LARCENY/ THEFT OFFENSES                                      0
##   TRAFFIC VIOLATION                                         1046
##
## Overall Statistics
##
##                Accuracy : 0.9998
##                  95% CI : (0.9992, 1)
##     No Information Rate : 0.4648
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9997
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: ALL OTHER OFFENSES Class: ASSAULT OFFENSES
## Sensitivity                             1.0000                 1.00000
## Specificity                             1.0000                 0.99978
## Pos Pred Value                          1.0000                 0.98684
## Neg Pred Value                          1.0000                 1.00000
## Prevalence                              0.0258                 0.01654
## Detection Rate                          0.0258                 0.01654
```

```
## Detection Prevalence                    0.0258              0.01676
## Balanced Accuracy                       1.0000              0.99989
##                        Class: BURGLARY/ BREAKING & ENTERING
## Sensitivity                                        1.0000
## Specificity                                        1.0000
## Pos Pred Value                                     1.0000
## Neg Pred Value                                     1.0000
## Prevalence                                         0.2206
## Detection Rate                                     0.2206
## Detection Prevalence                               0.2206
## Balanced Accuracy                                  1.0000
##                        Class: DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY
## Sensitivity                                                     1.0000
## Specificity                                                     1.0000
## Pos Pred Value                                                  1.0000
## Neg Pred Value                                                  1.0000
## Prevalence                                                      0.4648
## Detection Rate                                                  0.4648
## Detection Prevalence                                            0.4648
## Balanced Accuracy                                               1.0000
##                        Class: DRUNKENNESS/TRESPASSING/NUISANCE
## Sensitivity                                          0.98540
## Specificity                                          1.00000
## Pos Pred Value                                       1.00000
## Neg Pred Value                                       0.99978
## Prevalence                                           0.01511
## Detection Rate                                       0.01489
## Detection Prevalence                                 0.01489
## Balanced Accuracy                                    0.99270
##                        Class: FRAUD OFFENSES Class: LARCENY/ THEFT OFFENSES
## Sensitivity                          1.00000                         1.0000
## Specificity                          1.00000                         1.0000
## Pos Pred Value                       1.00000                         1.0000
## Neg Pred Value                       1.00000                         1.0000
## Prevalence                           0.01687                         0.1249
## Detection Rate                       0.01687                         0.1249
## Detection Prevalence                 0.01687                         0.1249
## Balanced Accuracy                    1.00000                         1.0000
##                        Class: TRAFFIC VIOLATION
## Sensitivity                               1.0000
## Specificity                               1.0000
## Pos Pred Value                            1.0000
## Neg Pred Value                            1.0000
## Prevalence                                0.1153
## Detection Rate                            0.1153
## Detection Prevalence                      0.1153
## Balanced Accuracy                         1.0000
```

```
print("Random Forest Model Performance :")
```

```
## [1] "Random Forest Model Performance :"
```

```
confusionMatrix(test_pred_rf, factor(cls_testing$`Category`))
```

```
## Confusion Matrix and Statistics
##
##                                               Reference
## Prediction                                     ALL OTHER OFFENSES
##   ALL OTHER OFFENSES                                          234
##   ASSAULT OFFENSES                                              0
##   BURGLARY/ BREAKING & ENTERING                                 0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                    0
##   DRUNKENNESS/TRESPASSING/NUISANCE                              0
##   FRAUD OFFENSES                                                0
##   LARCENY/ THEFT OFFENSES                                       0
##   TRAFFIC VIOLATION                                             0
##                                               Reference
## Prediction                                     ASSAULT OFFENSES
##   ALL OTHER OFFENSES                                          0
##   ASSAULT OFFENSES                                          150
##   BURGLARY/ BREAKING & ENTERING                               0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                  0
##   DRUNKENNESS/TRESPASSING/NUISANCE                            0
##   FRAUD OFFENSES                                              0
##   LARCENY/ THEFT OFFENSES                                     0
##   TRAFFIC VIOLATION                                           0
##                                               Reference
## Prediction                                     BURGLARY/ BREAKING & ENTERING
##   ALL OTHER OFFENSES                                                       0
##   ASSAULT OFFENSES                                                         0
##   BURGLARY/ BREAKING & ENTERING                                         2001
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                               0
##   DRUNKENNESS/TRESPASSING/NUISANCE                                         0
##   FRAUD OFFENSES                                                           0
##   LARCENY/ THEFT OFFENSES                                                  0
##   TRAFFIC VIOLATION                                                        0
##                                               Reference
## Prediction                                     DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERT
Y
##   ALL OTHER OFFENSES
0
##   ASSAULT OFFENSES
0
##   BURGLARY/ BREAKING & ENTERING
0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                                            421
5
##   DRUNKENNESS/TRESPASSING/NUISANCE
0
##   FRAUD OFFENSES
0
##   LARCENY/ THEFT OFFENSES
0
##   TRAFFIC VIOLATION
0
##                                               Reference
## Prediction                                     DRUNKENNESS/TRESPASSING/NUISANCE
##   ALL OTHER OFFENSES                                                          0
##   ASSAULT OFFENSES                                                            2
```

```
##   BURGLARY/ BREAKING & ENTERING                                        0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                           0
##   DRUNKENNESS/TRESPASSING/NUISANCE                                   135
##   FRAUD OFFENSES                                                       0
##   LARCENY/ THEFT OFFENSES                                              0
##   TRAFFIC VIOLATION                                                    0
##                                             Reference
## Prediction                                   FRAUD OFFENSES
##   ALL OTHER OFFENSES                                      0
##   ASSAULT OFFENSES                                        0
##   BURGLARY/ BREAKING & ENTERING                           0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY              0
##   DRUNKENNESS/TRESPASSING/NUISANCE                        0
##   FRAUD OFFENSES                                        153
##   LARCENY/ THEFT OFFENSES                                 0
##   TRAFFIC VIOLATION                                       0
##                                             Reference
## Prediction                                   LARCENY/ THEFT OFFENSES
##   ALL OTHER OFFENSES                                               0
##   ASSAULT OFFENSES                                                 0
##   BURGLARY/ BREAKING & ENTERING                                    0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                       0
##   DRUNKENNESS/TRESPASSING/NUISANCE                                 0
##   FRAUD OFFENSES                                                   0
##   LARCENY/ THEFT OFFENSES                                       1133
##   TRAFFIC VIOLATION                                                0
##                                             Reference
## Prediction                                   TRAFFIC VIOLATION
##   ALL OTHER OFFENSES                                         0
##   ASSAULT OFFENSES                                           0
##   BURGLARY/ BREAKING & ENTERING                              0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                 0
##   DRUNKENNESS/TRESPASSING/NUISANCE                           0
##   FRAUD OFFENSES                                             0
##   LARCENY/ THEFT OFFENSES                                    0
##   TRAFFIC VIOLATION                                       1046
##
## Overall Statistics
##
##                Accuracy : 0.9998
##                  95% CI : (0.9992, 1)
##     No Information Rate : 0.4648
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9997
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: ALL OTHER OFFENSES Class: ASSAULT OFFENSES
## Sensitivity                             1.0000                 1.00000
## Specificity                             1.0000                 0.99978
## Pos Pred Value                          1.0000                 0.98684
## Neg Pred Value                          1.0000                 1.00000
## Prevalence                              0.0258                 0.01654
## Detection Rate                          0.0258                 0.01654
```

```
## Detection Prevalence                    0.0258              0.01676
## Balanced Accuracy                       1.0000              0.99989
##                        Class: BURGLARY/ BREAKING & ENTERING
## Sensitivity                                        1.0000
## Specificity                                        1.0000
## Pos Pred Value                                     1.0000
## Neg Pred Value                                     1.0000
## Prevalence                                         0.2206
## Detection Rate                                     0.2206
## Detection Prevalence                               0.2206
## Balanced Accuracy                                  1.0000
##                        Class: DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY
## Sensitivity                                                     1.0000
## Specificity                                                     1.0000
## Pos Pred Value                                                  1.0000
## Neg Pred Value                                                  1.0000
## Prevalence                                                      0.4648
## Detection Rate                                                  0.4648
## Detection Prevalence                                            0.4648
## Balanced Accuracy                                               1.0000
##                        Class: DRUNKENNESS/TRESPASSING/NUISANCE
## Sensitivity                                          0.98540
## Specificity                                          1.00000
## Pos Pred Value                                       1.00000
## Neg Pred Value                                       0.99978
## Prevalence                                           0.01511
## Detection Rate                                       0.01489
## Detection Prevalence                                 0.01489
## Balanced Accuracy                                    0.99270
##                        Class: FRAUD OFFENSES Class: LARCENY/ THEFT OFFENSES
## Sensitivity                         1.00000                          1.0000
## Specificity                         1.00000                          1.0000
## Pos Pred Value                      1.00000                          1.0000
## Neg Pred Value                      1.00000                          1.0000
## Prevalence                          0.01687                          0.1249
## Detection Rate                      0.01687                          0.1249
## Detection Prevalence                0.01687                          0.1249
## Balanced Accuracy                   1.00000                          1.0000
##                        Class: TRAFFIC VIOLATION
## Sensitivity                               1.0000
## Specificity                               1.0000
## Pos Pred Value                            1.0000
## Neg Pred Value                            1.0000
## Prevalence                                0.1153
## Detection Rate                            0.1153
## Detection Prevalence                      0.1153
## Balanced Accuracy                         1.0000
```

```
print("Naive Bayes Performance :")
```

```
## [1] "Naive Bayes Performance :"
```

```
confusionMatrix(test_pred_nb, factor(cls_testing$`Category`))
```

```
## Confusion Matrix and Statistics
##
##                                             Reference
## Prediction                                   ALL OTHER OFFENSES
##   ALL OTHER OFFENSES                                          0
##   ASSAULT OFFENSES                                            0
##   BURGLARY/ BREAKING & ENTERING                               0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                234
##   DRUNKENNESS/TRESPASSING/NUISANCE                            0
##   FRAUD OFFENSES                                              0
##   LARCENY/ THEFT OFFENSES                                     0
##   TRAFFIC VIOLATION                                           0
##                                             Reference
## Prediction                                   ASSAULT OFFENSES
##   ALL OTHER OFFENSES                                        0
##   ASSAULT OFFENSES                                          0
##   BURGLARY/ BREAKING & ENTERING                             0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY              150
##   DRUNKENNESS/TRESPASSING/NUISANCE                          0
##   FRAUD OFFENSES                                            0
##   LARCENY/ THEFT OFFENSES                                   0
##   TRAFFIC VIOLATION                                         0
##                                             Reference
## Prediction                                   BURGLARY/ BREAKING & ENTERING
##   ALL OTHER OFFENSES                                                     0
##   ASSAULT OFFENSES                                                       0
##   BURGLARY/ BREAKING & ENTERING                                       2001
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                             0
##   DRUNKENNESS/TRESPASSING/NUISANCE                                       0
##   FRAUD OFFENSES                                                         0
##   LARCENY/ THEFT OFFENSES                                                0
##   TRAFFIC VIOLATION                                                      0
##                                             Reference
## Prediction                                   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERT
Y
##   ALL OTHER OFFENSES                                                                 
0
##   ASSAULT OFFENSES                                                                   
0
##   BURGLARY/ BREAKING & ENTERING                                                      
0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                                      421
5
##   DRUNKENNESS/TRESPASSING/NUISANCE                                                   
0
##   FRAUD OFFENSES                                                                     
0
##   LARCENY/ THEFT OFFENSES                                                            
0
##   TRAFFIC VIOLATION                                                                  
0
##                                             Reference
## Prediction                                   DRUNKENNESS/TRESPASSING/NUISANCE
##   ALL OTHER OFFENSES                                                        0
##   ASSAULT OFFENSES                                                          0
```

```
##   BURGLARY/ BREAKING & ENTERING                                               0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                                137
##   DRUNKENNESS/TRESPASSING/NUISANCE                                            0
##   FRAUD OFFENSES                                                              0
##   LARCENY/ THEFT OFFENSES                                                     0
##   TRAFFIC VIOLATION                                                           0
##                                           Reference
## Prediction                                 FRAUD OFFENSES
##   ALL OTHER OFFENSES                                    0
##   ASSAULT OFFENSES                                      0
##   BURGLARY/ BREAKING & ENTERING                         0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY          153
##   DRUNKENNESS/TRESPASSING/NUISANCE                      0
##   FRAUD OFFENSES                                        0
##   LARCENY/ THEFT OFFENSES                               0
##   TRAFFIC VIOLATION                                     0
##                                           Reference
## Prediction                                 LARCENY/ THEFT OFFENSES
##   ALL OTHER OFFENSES                                             0
##   ASSAULT OFFENSES                                               0
##   BURGLARY/ BREAKING & ENTERING                                  0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY                  1133
##   DRUNKENNESS/TRESPASSING/NUISANCE                               0
##   FRAUD OFFENSES                                                 0
##   LARCENY/ THEFT OFFENSES                                        0
##   TRAFFIC VIOLATION                                              0
##                                           Reference
## Prediction                                 TRAFFIC VIOLATION
##   ALL OTHER OFFENSES                                       0
##   ASSAULT OFFENSES                                         0
##   BURGLARY/ BREAKING & ENTERING                            0
##   DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY            1046
##   DRUNKENNESS/TRESPASSING/NUISANCE                         0
##   FRAUD OFFENSES                                           0
##   LARCENY/ THEFT OFFENSES                                  0
##   TRAFFIC VIOLATION                                        0
##
## Overall Statistics
##
##                Accuracy : 0.6854
##                  95% CI : (0.6757, 0.695)
##     No Information Rate : 0.4648
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.466
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: ALL OTHER OFFENSES Class: ASSAULT OFFENSES
## Sensitivity                             0.0000                 0.00000
## Specificity                             1.0000                 1.00000
## Pos Pred Value                             NaN                     NaN
## Neg Pred Value                          0.9742                 0.98346
## Prevalence                              0.0258                 0.01654
## Detection Rate                          0.0000                 0.00000
```

```
## Detection Prevalence                0.0000               0.00000
## Balanced Accuracy                   0.5000               0.50000
##                        Class: BURGLARY/ BREAKING & ENTERING
## Sensitivity                                    1.0000
## Specificity                                    1.0000
## Pos Pred Value                                 1.0000
## Neg Pred Value                                 1.0000
## Prevalence                                     0.2206
## Detection Rate                                 0.2206
## Detection Prevalence                           0.2206
## Balanced Accuracy                              1.0000
##                        Class: DESTRUCTION/ DAMAGE/ VANDALISM OF PROPERTY
## Sensitivity                                                 1.0000
## Specificity                                                 0.4122
## Pos Pred Value                                              0.5963
## Neg Pred Value                                              1.0000
## Prevalence                                                  0.4648
## Detection Rate                                              0.4648
## Detection Prevalence                                        0.7794
## Balanced Accuracy                                           0.7061
##                        Class: DRUNKENNESS/TRESPASSING/NUISANCE
## Sensitivity                                      0.00000
## Specificity                                      1.00000
## Pos Pred Value                                       NaN
## Neg Pred Value                                   0.98489
## Prevalence                                       0.01511
## Detection Rate                                   0.00000
## Detection Prevalence                             0.00000
## Balanced Accuracy                                0.50000
##                        Class: FRAUD OFFENSES Class: LARCENY/ THEFT OFFENSES
## Sensitivity                          0.00000                         0.0000
## Specificity                          1.00000                         1.0000
## Pos Pred Value                           NaN                            NaN
## Neg Pred Value                       0.98313                         0.8751
## Prevalence                           0.01687                         0.1249
## Detection Rate                       0.00000                         0.0000
## Detection Prevalence                 0.00000                         0.0000
## Balanced Accuracy                    0.50000                         0.5000
##                        Class: TRAFFIC VIOLATION
## Sensitivity                              0.0000
## Specificity                              1.0000
## Pos Pred Value                              NaN
## Neg Pred Value                           0.8847
## Prevalence                               0.1153
## Detection Rate                           0.0000
## Detection Prevalence                     0.0000
## Balanced Accuracy                        0.5000
```

7 Interpretations and Scope for Future Improvement

7.1 Interpretations

- Regression Model Evaluation – RMSE Mean : 14(gbm) > 17(lm) – Rsquared : 0.85(lm) > 0.82(gbm)

- Classification Model Evaluation – Accuracy : 100%(SVM) > 99.98% (RF) > 68%(Naive Bayes)

7.2 Scope for Future Improvement

- Predicted values in regression fit well with the actual values as per the plotted graphs of actual vs predicted.
- Accuracy is high for SVM,RF - High overfitting possibile (or over-simplified model), Accuracy is moderate for Naive Bayes method
- Better feature engineering and complex selection of explanatory attributes must be addressed