ACADGILD

SESSION 12: Generalized Linear Models

Assignment 1

Submitted by: Munmun Ghosal
Login Id: munmun55@gmail.com
(M):+91-8007178659

Data Analytics

Table of Contents

1. Problem Statement

1. Use the given link below:
https://archive.ics.uci.edu/ml/machine-learning-databases/communities/

Perform the below operations:

a) Find out top 5 attributes having highest correlation (select only Numeric features).
b) Find out top 3 reasons for having more crime in a city.
c) Which all attributes have high correlation with crime rate?

2. Solution

a) Find out top 5 attributes having highest correlation (select only Numeric features).

The R-script for the given problem is as follows:

```
library(readr)
Crimes <- read_csv("E:/munmun_acadgild/acadgild data analytics/supporting
files/communities.csv ")
View(Crimes)

names(Crimes) <- c("Case", "Number", "Date", "Block", "IUCR", "Primary Type",
"Description",
                   "Location Desc", "Arrest", "Domestic", "Beat", "District", "Ward",
"Community Area",
                   "FBI Code", "X Coordinate", "Y Coordinate", "Year", "Updated On",
                   "Latitude", "Longitude", "Location")
head(Crimes)
str(Crimes)

#a. Find out top 5 attributes having highest correlation (select only Numeric features).

Crimes <- na.omit(Crimes)
names(Crimes)
c <- cor(Crimes[c(11,12,13,14,18,20,21)])
c
library(reshape2)
m <- melt(c)
```

```
library(dplyr)
m
top <- m%>%select(Var1, Var2, value)%>%filter(value != 1)
top[order(top$value, decreasing = T)[1:10],]
```

The output of the R-Script (from Console window) is given as follows:

```
> library(readr)
> Crimes <- read_csv("E:/munmun_acadgild/acadgild data
analytics/supporting files/communities.csv")
Parsed with column specification:
cols(
  .default = col_character(),
  ID = col_double(),
  Arrest = col_logical(),
  Domestic = col_logical(),
  Beat = col_double(),
  District = col_double(),
  Ward = col_double(),
  `Community Area` = col_double(),
  `X Coordinate` = col_double(),
  `Y Coordinate` = col_double(),
  Year = col_double(),
  Latitude = col_double(),
  Longitude = col_double()
)
See spec(...) for full column specifications.
|=================================================================
==================================| 100%  216 MB
> View(Crimes)
> names(Crimes) <- c("Case", "Number", "Date", "Block", "IUCR",
"Primary Type", "Description",
+                    "Location Desc", "Arrest", "Domestic", "Beat",
"District", "Ward", "Community Area",
+                    "FBI Code", "X Coordinate", "Y Coordinate",
"Year", "Updated On",
+                    "Latitude", "Longitude", "Location")
> head(Crimes)
# A tibble: 6 x 22
    Case Number Date  Block IUCR  `Primary Type` Description `Location
Desc` Arrest Domestic  Beat District  Ward
   <dbl> <chr>  <chr> <chr> <chr> <chr>          <chr>       <chr>
<lgl>  <lgl>    <dbl>    <dbl> <dbl>
1 1.05e7 HZ250~ 5/3/~ 013X~ 486   BATTERY        DOMESTIC B~ APARTMENT
TRUE   TRUE      1022       10    24
2 1.05e7 HZ250~ 5/3/~ 061X~ 486   BATTERY        DOMESTIC B~ RESIDENCE
FALSE  TRUE       313        3    20
3 1.05e7 HZ250~ 5/3/~ 053X~ 470   PUBLIC PEACE ~ RECKLESS C~ STREET
FALSE  FALSE     1524       15    37
4 1.05e7 HZ250~ 5/3/~ 049X~ 460   BATTERY        SIMPLE      SIDEWALK
FALSE  FALSE     1532       15    28
5 1.05e7 HZ250~ 5/3/~ 003X~ 820   THEFT          $500 AND U~ RESIDENCE
FALSE  TRUE      1523       15    28
6 1.05e7 HZ250~ 5/3/~ 082X~ 041A  BATTERY        AGGRAVATED~ STREET
FALSE  FALSE      631        6     8
# ... with 9 more variables: `Community Area` <dbl>, `FBI Code` <chr>,
`X Coordinate` <dbl>, `Y Coordinate` <dbl>,
```

```
#   Year <dbl>, `Updated On` <chr>, Latitude <dbl>, Longitude <dbl>,
Location <chr>
> str(Crimes)
Classes ‘spec_tbl_df’, ‘tbl_df’, ‘tbl’ and 'data.frame':      1048575
obs. of  22 variables:
 $ Case           : num  10508693 10508695 10508697 10508698 10508699
...
 $ Number         : chr  "HZ250496" "HZ250409" "HZ250503" "HZ250424" ...
 $ Date           : chr  "5/3/2016 23:40" "5/3/2016 21:40" "5/3/2016
23:31" "5/3/2016 22:10" ...
 $ Block          : chr  "013XX S SAWYER AVE" "061XX S DREXEL AVE"
"053XX W CHICAGO AVE" "049XX W FULTON ST" ...
 $ IUCR           : chr  "486" "486" "470" "460" ...
 $ Primary Type   : chr  "BATTERY" "BATTERY" "PUBLIC PEACE VIOLATION"
"BATTERY" ...
 $ Description    : chr  "DOMESTIC BATTERY SIMPLE" "DOMESTIC BATTERY
SIMPLE" "RECKLESS CONDUCT" "SIMPLE" ...
 $ Location Desc : chr  "APARTMENT" "RESIDENCE" "STREET" "SIDEWALK" ...
 $ Arrest         : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
 $ Domestic       : logi  TRUE TRUE FALSE FALSE TRUE FALSE ...
 $ Beat           : num  1022 313 1524 1532 1523 ...
 $ District       : num  10 3 15 15 15 6 1 2 24 7 ...
 $ Ward           : num  24 20 37 28 28 8 3 3 40 17 ...
 $ Community Area: num  29 42 25 25 25 44 35 38 1 67 ...
 $ FBI Code       : chr  "08B" "08B" "24" "08B" ...
 $ X Coordinate   : num  1154907 1183066 1140789 1143223 1139890 ...
 $ Y Coordinate   : num  1893681 1864330 1904819 1901475 1901675 ...
 $ Year           : num  2016 2016 2016 2016 2016 ...
 $ Updated On     : chr  "5/10/2016 15:56" "5/10/2016 15:56" "5/10/2016
15:56" "5/10/2016 15:56" ...
 $ Latitude       : num  41.9 41.8 41.9 41.9 41.9 ...
 $ Longitude      : num  -87.7 -87.6 -87.8 -87.7 -87.8 ...
 $ Location       : chr  "(41.864073157, -87.706818608)" "(41.782921527,
-87.60436317)" "(41.894908283, -87.758371958)" "(41.885686845, -
87.749515983)" ...
 - attr(*, "spec")=
  .. cols(
  ..   ID = col_double(),
  ..   `Case Number` = col_character(),
  ..   Date = col_character(),
  ..   Block = col_character(),
  ..   IUCR = col_character(),
  ..   `Primary Type` = col_character(),
  ..   Description = col_character(),
  ..   `Location Description` = col_character(),
  ..   Arrest = col_logical(),
  ..   Domestic = col_logical(),
  ..   Beat = col_double(),
  ..   District = col_double(),
  ..   Ward = col_double(),
  ..   `Community Area` = col_double(),
  ..   `FBI Code` = col_character(),
  ..   `X Coordinate` = col_double(),
  ..   `Y Coordinate` = col_double(),
  ..   Year = col_double(),
  ..   `Updated On` = col_character(),
  ..   Latitude = col_double(),
  ..   Longitude = col_double(),
  ..   Location = col_character()
  .. )
```

```
> Crimes <- na.omit(Crimes)
> names(Crimes)
 [1] "Case"           "Number"         "Date"           "Block"
"IUCR"           "Primary Type"
 [7] "Description"    "Location Desc"  "Arrest"         "Domestic"
"Beat"           "District"
[13] "Ward"           "Community Area" "FBI Code"       "X Coordinate"
"Y Coordinate"   "Year"
[19] "Updated On"     "Latitude"       "Longitude"      "Location"
> c <- cor(Crimes[c(11,12,13,14,18,20,21)])
> c
                      Beat     District         Ward Community Area
Year     Latitude    Longitude
Beat            1.00000000  0.996402087  0.687144016    -0.49621344 -
0.012652765  0.575284245 -0.479976546
District        0.99640209  1.000000000  0.691655842    -0.49621461 -
0.008529942  0.576344843 -0.483244475
Ward            0.68714402  0.691655842  1.000000000    -0.54302431 -
0.004215319  0.592008238 -0.397964013
Community Area -0.49621344 -0.496214608 -0.543024307     1.00000000
0.001632430 -0.691892413  0.221028077
Year           -0.01265277 -0.008529942 -0.004215319     0.00163243
1.000000000 -0.002721412 -0.004346718
Latitude        0.57528424  0.576344843  0.592008238    -0.69189241 -
0.002721412  1.000000000 -0.209999084
Longitude      -0.47997655 -0.483244475 -0.397964013     0.22102808 -
0.004346718 -0.209999084  1.000000000
> library(reshape2)
> m <- melt(c)
> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> m
             Var1           Var2        value
1            Beat           Beat  1.000000000
2        District           Beat  0.996402087
3            Ward           Beat  0.687144016
4  Community Area           Beat -0.496213439
5            Year           Beat -0.012652765
6        Latitude           Beat  0.575284245
7       Longitude           Beat -0.479976546
8            Beat       District  0.996402087
9        District       District  1.000000000
10           Ward       District  0.691655842
11 Community Area       District -0.496214608
12           Year       District -0.008529942
13       Latitude       District  0.576344843
14      Longitude       District -0.483244475
15           Beat           Ward  0.687144016
16       District           Ward  0.691655842
17           Ward           Ward  1.000000000
```

```
18 Community Area           Ward -0.543024307
19           Year           Ward -0.004215319
20       Latitude           Ward  0.592008238
21      Longitude           Ward -0.397964013
22           Beat Community Area -0.496213439
23       District Community Area -0.496214608
24           Ward Community Area -0.543024307
25 Community Area Community Area  1.000000000
26           Year Community Area  0.001632430
27       Latitude Community Area -0.691892413
28      Longitude Community Area  0.221028077
29           Beat           Year -0.012652765
30       District           Year -0.008529942
31           Ward           Year -0.004215319
32 Community Area           Year  0.001632430
33           Year           Year  1.000000000
34       Latitude           Year -0.002721412
35      Longitude           Year -0.004346718
36           Beat       Latitude  0.575284245
37       District       Latitude  0.576344843
38           Ward       Latitude  0.592008238
39 Community Area       Latitude -0.691892413
40           Year       Latitude -0.002721412
41       Latitude       Latitude  1.000000000
42      Longitude       Latitude -0.209999084
43           Beat      Longitude -0.479976546
44       District      Longitude -0.483244475
45           Ward      Longitude -0.397964013
46 Community Area      Longitude  0.221028077
47           Year      Longitude -0.004346718
48       Latitude      Longitude -0.209999084
49      Longitude      Longitude  1.000000000
> top <- m%>%select(Var1, Var2, value)%>%filter(value != 1)
> top[order(top$value, decreasing = T)[1:10],]
       Var1     Var2     value
1  District     Beat 0.9964021
7      Beat District 0.9964021
8      Ward District 0.6916558
14 District     Ward 0.6916558
2      Ward     Beat 0.6871440
13     Beat     Ward 0.6871440
17 Latitude     Ward 0.5920082
33     Ward Latitude 0.5920082
11 Latitude District 0.5763448
32 District Latitude 0.5763448
```

Conclusion/Interpretation:

District~Beat, Ward~District, Ward~Beat, Latitude ~Ward, Latitude~District are top5 attributes with highest correlations

b) Find out top 3 reasons for having more crime in a city.

The R-script for the given problem is as follows:

```
x <- as.data.frame(table(Crimes$Description))
x[order(x$Freq, decreasing = T)[1:3],]
```

The output of the R-Script (from Console window) is given as follows:

```
> x <- as.data.frame(table(Crimes$Description))
> x[order(x$Freq, decreasing = T)[1:3],]
                        Var1   Freq
279                   SIMPLE 107887
1             $500 AND UNDER  97476
118 DOMESTIC BATTERY SIMPLE   93001
```

Conclusion/Interpretation:

Simple, **$500 and under** and Domestic **Battery Simple** are the top 3 reasons for having more crime

c) Which all attributes have high correlation with crime rate?

The R-script for the given problem is as follows:

```
crime <- Crimes
head(crime)
table(is.na(crime))

crime$Date <- as.POSIXlt(crime$Date, format= "%m/%d/%Y %H:%M:%S")
crime$`Updated On` <- as.POSIXlt(crime$`Updated On`, format= "%m/%d/%Y
%H:%M:%S")

install.packages("chron")
library(chron)

crime$Time <- time(format(crime$Date,"%H:%M:%S"))
crime$Date <- as.POSIXct(crime$Date)
crime$`Updated On` <- as.POSIXct(crime$`Updated On`)

# There could be certain time intervals of the day where criminal activity is more
prevalent

time.tag <-  chron::chron(time=c("00:00:00", "06:00:00", "12:00:00",
"18:00:00","23:59:00"))
time.tag
crime$time.tag <- cut(crime$Time, breaks= time.tag,
```

```
                labels= c("00-06","06-12", "12-18", "18-00"), include.lowest =TRUE)
table(crime$time.tag)

# date variable to contain just the date part
crime$date <- as.POSIXlt(strptime(crime$Date, format = "%Y-%m-%d"))
crime$date <- as.POSIXct(crime$date)

# days and months could be predicatble variable
crime$day <- as.factor(weekdays(crime$Date, abbreviate = TRUE))
crime$month <- as.factor(months(crime$Date, abbreviate = TRUE))
str(crime$day)
str(crime$month)

# converting Arrest yes / no to binary varibale
crime$Arrest <- ifelse(as.character(crime$Arrest) == "true",1,0)

# The data contain about 31 crime types, not all of which are mutually exclusive. We can
combine
# two or more similar categories into one to reduce this number and make the analysis a
bit easier.7
crime$crime <- as.character(crime$`Primary Type`)
crime$crime <- ifelse(crime$crime %in% c("CRIM SEXUAL
ASSAULT","PROSTITUTION", "SEX OFFENSE","HUMAN TRAFFICKING"), 'SEX',
crime$crime)
crime$crime <- ifelse(crime$crime %in% c("MOTOR VEHICLE THEFT"), "MVT",
crime$crime)
crime$crime <- ifelse(crime$crime %in% c("GAMBLING", "INTERFEREWITH
PUBLIC OFFICER", "INTERFERENCE WITH PUBLIC OFFICER",
"INTIMIDATION",
                                "LIQUOR LAW VIOLATION", "OBSCENITY", "NON-
CRIMINAL", "PUBLIC PEACE VIOLATION",
                                "PUBLIC INDECENCY", "STALKING", "NON-CRIMINAL
(SUBJECT SPECIFIED)","NON - CRIMINAL"),
                  "NONVIO", crime$crime)
crime$crime <- ifelse(crime$crime == "CRIMINAL DAMAGE",
"DAMAGE",crime$crime)
crime$crime <- ifelse(crime$crime == "CRIMINAL TRESPASS","TRESPASS",
crime$crime)
crime$crime <- ifelse(crime$crime %in% c("NARCOTICS", "OTHER NARCOTIC
VIOLATION", "OTHER NARCOTIC VIOLATION"), "DRUG", crime$crime)
crime$crime <- ifelse(crime$crime == "DECEPTIVE PRACTICE","FRAUD",
crime$crime)
crime$crime <- ifelse(crime$crime %in% c("OTHER OFFENSE",
"OTHEROFFENSE"), "OTHER", crime$crime)
```

```
crime$crime <- ifelse(crime$crime %in% c("KIDNAPPING", "WEAPONS
VIOLATION", "CONCEALED CARRY LICENSE VIOLATION","OFFENSE
INVOLVING CHILDREN"), "VIO", crime$crime)
table(crime$crime)

# A potential important indicator of criminal activity in a particular area could be the
history of criminal activities in the past.

temp <- aggregate(crime$crime, by=list(crime$crime, crime$time.tag), FUN=length)
names(temp) <- c("crime", "time.tag", "count")
library(dplyr)
temp <- ddply(crime, .(crime, day), summarise, count = length(date))

#install.packages("doBy")
library(doBy)
crime.agg <- ddply(crime, .(crime, Arrest, Beat, date, `X Coordinate`, `Y Coordinate`,
time.tag, day, month),
                summarise, count=length(date), .progress='text')

beats <- sort(unique(crime.agg$Beat))
dates <- sort(as.character(unique(crime.agg$date)))
temp <- expand.grid(beats, dates)
names(temp) <- c("Beat", "date")

model.data <- aggregate(crime.agg[, c('count', 'Arrest')], by=
                    list(crime.agg$Beat, as.character(crime.agg$date)), FUN=sum)
names(model.data) <- c("Beat", "date", "count", "Arrest")
model.data <- merge(temp, model.data, by= c('Beat', 'date'), all.x= TRUE)
#View(model.data)
model.data$count[is.na(model.data$count)] <- 0
model.data$Arrest[is.na(model.data$Arrest)] <- 0
model.data$day <- weekdays(as.Date(model.data$date), abbreviate= TRUE)
model.data$month <- months(as.Date(model.data$date), abbreviate= TRUE)
pastDays <- function(x) {c(0, rep(1, x))}
model.data$past.crime.1 <- ave(model.data$count, model.data$Beat,
                        FUN=function(x) filter(x, pastDays(1), sides= 1))
model.data$past.crime.7 <- ave(model.data$count, model.data$Beat,
                        FUN=function(x) filter(x, pastDays(7), sides= 1))
model.data$past.crime.30 <- ave(model.data$count, model.data$Beat,
                         FUN=function(x) filter(x, pastDays(30), sides= 1))

meanNA <- function(x){mean(x, na.rm= TRUE)}
model.data$past.crime.1 <- ifelse(is.na(model.data$past.crime.1),
                           meanNA(model.data$past.crime.1), model.data$past.crime.1)
model.data$past.crime.7 <- ifelse(is.na(model.data$past.crime.7),
```

```
                          meanNA(model.data$past.crime.7), model.data$past.crime.7)
model.data$past.crime.30 <- ifelse(is.na(model.data$past.crime.30),
                           meanNA(model.data$past.crime.30), model.data$past.crime.30)
# past variables for arrests
model.data$past.arrest.30 <- ave(model.data$Arrest, model.data$Beat,
                         FUN= function(x) filter(x, pastDays(30), sides= 1))
model.data$past.arrest.30 <- ifelse(is.na(model.data$past.arrest.30),
                            meanNA(model.data$past.arrest.30), model.data$past.arrest.30)
# arrests per crime
model.data$policing <- ifelse(model.data$past.crime.30 == 0, 0,
                      model.data$past.arrest.30/model.data$past.crime.30)

# trend
model.data$crime.trend <- ifelse(model.data$past.crime.30 == 0, 0,
                         model.data$past.crime.7/model.data$past.crime.30)

# season could be another reason
model.data$season <- as.factor(ifelse(model.data$month %in% c("Mar", "Apr", "May"),
"spring",
                               ifelse(model.data$month %in% c("Jun", "Jul", "Aug"),
"summer",
                                      ifelse(model.data$month %in% c("Sep", "Oct","Nov"),
"fall", "winter"))))

model.cor <- cor(model.data[, c("count", "past.crime.1", "past.crime.7",
                          "past.crime.30","policing", "crime.trend")])
model.cor
library(psych)
psych::cor.plot(model.cor)
```

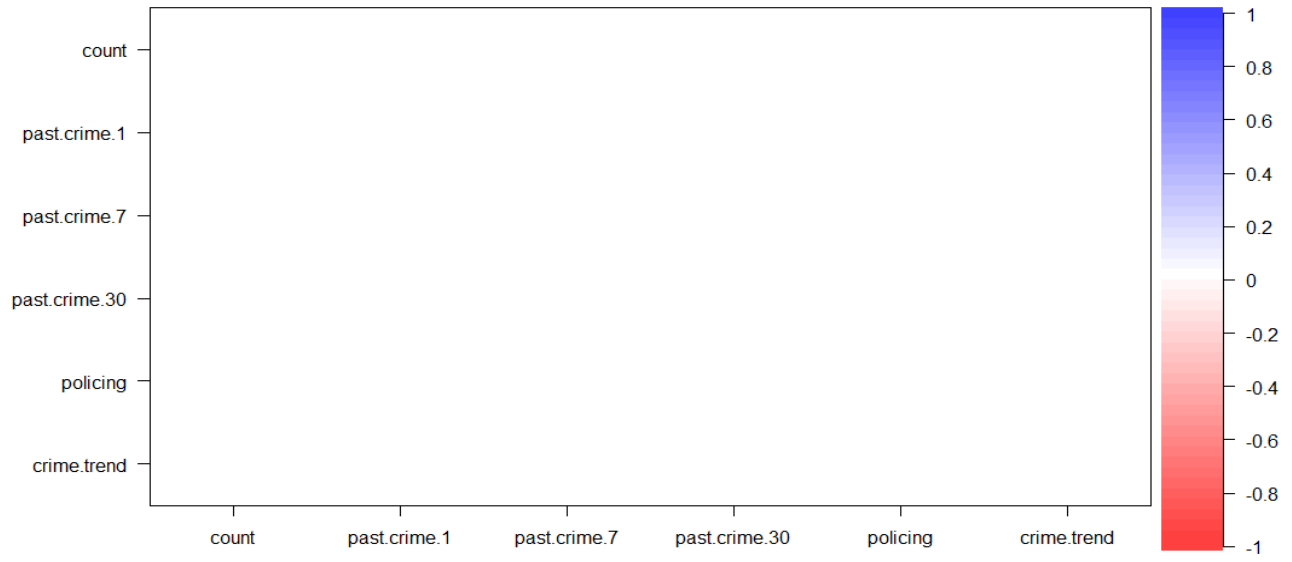
The output of the R-Script (from Console window) is given as follows:

```
> crime <- Crimes
> head(crime)
# A tibble: 6 x 22
    Case Number Date  Block IUCR  `Primary Type` Description `Location
Desc` Arrest Domestic  Beat District  Ward
   <dbl> <chr>  <chr> <chr> <chr> <chr>          <chr>       <chr>
<lgl>  <lgl>    <dbl>    <dbl> <dbl>
1 1.05e7 HZ250~ 5/3/~ 013X~ 486   BATTERY        DOMESTIC B~ APARTMENT
TRUE   TRUE      1022       10    24
2 1.05e7 HZ250~ 5/3/~ 061X~ 486   BATTERY        DOMESTIC B~ RESIDENCE
FALSE  TRUE       313        3    20
3 1.05e7 HZ250~ 5/3/~ 053X~ 470   PUBLIC PEACE ~ RECKLESS C~ STREET
FALSE  FALSE     1524       15    37
4 1.05e7 HZ250~ 5/3/~ 049X~ 460   BATTERY        SIMPLE      SIDEWALK
FALSE  FALSE     1532       15    28
```

```
5 1.05e7 HZ250~ 5/3/~ 003X~ 820   THEFT          $500 AND U~ RESIDENCE
FALSE  TRUE      1523       15    28
6 1.05e7 HZ250~ 5/3/~ 082X~ 041A  BATTERY        AGGRAVATED~ STREET
FALSE  FALSE      631        6     8
# ... with 9 more variables: `Community Area` <dbl>, `FBI Code` <chr>,
`X Coordinate` <dbl>, `Y Coordinate` <dbl>,
#   Year <dbl>, `Updated On` <chr>, Latitude <dbl>, Longitude <dbl>,
Location <chr>
> table(is.na(crime))

   FALSE
22863082

> crime$Date <- as.POSIXlt(crime$Date, format= "%m/%d/%Y %H:%M:%S")
> crime$`Updated On` <- as.POSIXlt(crime$`Updated On`, format=
"%m/%d/%Y %H:%M:%S")
> library(chron)
> crime$Time <- time(format(crime$Date,"%H:%M:%S"))
> crime$Date <- as.POSIXct(crime$Date)
> crime$`Updated On` <- as.POSIXct(crime$`Updated On`)
> # There could be certain time intervals of the day where criminal
activity is more prevalent
>
> time.tag <-  chron::chron(time=c("00:00:00", "06:00:00", "12:00:00",
"18:00:00","23:59:00"))
> time.tag
[1] 00:00:00 06:00:00 12:00:00 18:00:00 23:59:00
> crime$time.tag <- cut(crime$Time, breaks= time.tag,
+                       labels= c("00-06","06-12", "12-18", "18-00"),
include.lowest =TRUE)
> table(crime$time.tag)

00-06 06-12 12-18 18-00
    0     0     0     0

> # date variable to contain just the date part
> crime$date <- as.POSIXlt(strptime(crime$Date, format = "%Y-%m-%d"))
> crime$date <- as.POSIXct(crime$date)
> # days and months could be predicatble variable
> crime$day <- as.factor(weekdays(crime$Date, abbreviate = TRUE))
> crime$month <- as.factor(months(crime$Date, abbreviate = TRUE))
> str(crime$day)
 Factor w/ 0 levels: NA NA NA NA NA NA NA NA NA NA ...
> str(crime$month)
 Factor w/ 0 levels: NA NA NA NA NA NA NA NA NA NA ...
> # converting Arrest yes / no to binary varibale
> crime$Arrest <- ifelse(as.character(crime$Arrest) == "true",1,0)
> # The data contain about 31 crime types, not all of which are mutually
exclusive. We can combine
> # two or more similar categories into one to reduce this number and make
the analysis a bit easier.7
> crime$crime <- as.character(crime$`Primary Type`)
> crime$crime <- ifelse(crime$crime %in% c("CRIM SEXUAL
ASSAULT","PROSTITUTION", "SEX OFFENSE","HUMAN TRAFFICKING"), 'SEX',
crime$crime)
> crime$crime <- ifelse(crime$crime %in% c("MOTOR VEHICLE THEFT"), "MVT",
crime$crime)
> crime$crime <- ifelse(crime$crime %in% c("GAMBLING", "INTERFEREWITH PUBLIC
OFFICER", "INTERFERENCE WITH PUBLIC OFFICER", "INTIMIDATION",
```

```
+                                          "LIQUOR LAW VIOLATION",
"OBSCENITY", "NON-CRIMINAL", "PUBLIC PEACE VIOLATION",
+                                          "PUBLIC INDECENCY", "STALKING",
"NON-CRIMINAL (SUBJECT SPECIFIED)","NON - CRIMINAL"),
+                        "NONVIO", crime$crime)
> crime$crime <- ifelse(crime$crime == "CRIMINAL DAMAGE",
"DAMAGE",crime$crime)
> crime$crime <- ifelse(crime$crime == "CRIMINAL TRESPASS","TRESPASS",
crime$crime)
> crime$crime <- ifelse(crime$crime %in% c("NARCOTICS", "OTHER NARCOTIC
VIOLATION", "OTHER NARCOTIC VIOLATION"), "DRUG", crime$crime)
> crime$crime <- ifelse(crime$crime == "DECEPTIVE PRACTICE","FRAUD",
crime$crime)
> crime$crime <- ifelse(crime$crime %in% c("OTHER OFFENSE", "OTHEROFFENSE"),
"OTHER", crime$crime)
> crime$crime <- ifelse(crime$crime %in% c("KIDNAPPING", "WEAPONS VIOLATION",
"CONCEALED CARRY LICENSE VIOLATION","OFFENSE INVOLVING CHILDREN"), "VIO",
crime$crime)
> table(crime$crime)

   ARSON  ASSAULT  BATTERY BURGLARY   DAMAGE      DRUG    FRAUD HOMICIDE
MVT   NONVIO    OTHER  ROBBERY      SEX
    1448    63675   187643    61045   108508   109738    46558       76
43785    19536    61262    39491    13796
   THEFT TRESPASS      VIO
  234716    27458    20496


> temp <- aggregate(crime$crime, by=list(crime$crime, crime$time.tag), FUN=length)
> names(temp) <- c("crime", "time.tag", "count")
> library(dplyr)
> temp <- ddply(crime, .(crime, day), summarise, count = length(date))
> install.packages("doBy")
Error in install.packages : Updating loaded packages
> library(doBy)
> length(Case ~ crime + month)
[1] 3
> length(crime)
[1] 28
> install.packages("doBy")
Installing package into 'C:/Users/Munmun/Documents/R/win-library/3.5'
(as 'lib' is unspecified)

> temp <- aggregate(crime$crime, by=list(crime$crime, crime$time.tag), FUN=length)
> names(temp) <- c("crime", "time.tag", "count")
> library(dplyr)
> temp <- ddply(crime, .(crime, day), summarise, count = length(date))
> library(doBy)
> # temp <- summaryBy(Case ~ crime + month, data = crime, FUN= length)
> # names(temp)[3] <- "count"
>
> crime.agg <- ddply(crime, .(crime, Arrest, Beat, date, `X Coordinate`, `Y
Coordinate`, time.tag, day, month),
+                    summarise, count=length(date), .progress='text')

|=================================================================================
==================================| 100%
> beats <- sort(unique(crime.agg$Beat))
> dates <- sort(as.character(unique(crime.agg$date)))
> temp <- expand.grid(beats, dates)
> names(temp) <- c("Beat", "date")
> model.data <- aggregate(crime.agg[, c('count', 'Arrest')], by=
+                           list(crime.agg$Beat, as.character(crime.agg$date)),
FUN=sum)
> names(model.data) <- c("Beat", "date", "count", "Arrest")
> model.data <- merge(temp, model.data, by= c('Beat', 'date'), all.x= TRUE)
```

```
> View(model.data)
> model.data$count[is.na(model.data$count)] <- 0
> model.data$Arrest[is.na(model.data$Arrest)] <- 0
> model.data$day <- weekdays(as.Date(model.data$date), abbreviate= TRUE)
> model.data$month <- months(as.Date(model.data$date), abbreviate= TRUE)
> pastDays <- function(x) {c(0, rep(1, x))}
> model.data$past.crime.1 <- ave(model.data$count, model.data$Beat,
+                                FUN=function(x) filter(x, pastDays(1), sides= 1))
> model.data$past.crime.7 <- ave(model.data$count, model.data$Beat,
+                                FUN=function(x) filter(x, pastDays(7), sides= 1))
> model.data$past.crime.30 <- ave(model.data$count, model.data$Beat,
+                                 FUN=function(x) filter(x, pastDays(30), sides=
1))
> meanNA <- function(x){mean(x, na.rm= TRUE)}
> model.data$past.crime.1 <- ifelse(is.na(model.data$past.crime.1),
+                                   meanNA(model.data$past.crime.1),
model.data$past.crime.1)
> model.data$past.crime.7 <- ifelse(is.na(model.data$past.crime.7),
+                                   meanNA(model.data$past.crime.7),
model.data$past.crime.7)
> model.data$past.crime.30 <- ifelse(is.na(model.data$past.crime.30),
+                                    meanNA(model.data$past.crime.30),
model.data$past.crime.30)
> # past variables for arrests
> model.data$past.arrest.30 <- ave(model.data$Arrest, model.data$Beat,
+                                  FUN= function(x) filter(x, pastDays(30), sides=
1))
> model.data$past.arrest.30 <- ifelse(is.na(model.data$past.arrest.30),
+                                     meanNA(model.data$past.arrest.30),
model.data$past.arrest.30)
> # arrests per crime
> model.data$policing <- ifelse(model.data$past.crime.30 == 0, 0,
+                               model.data$past.arrest.30/model.data$past.crime.30)
> # trend
> model.data$crime.trend <- ifelse(model.data$past.crime.30 == 0, 0,
+
model.data$past.crime.7/model.data$past.crime.30)
```



Conclusion/Interpretation:

All the variables considered in the model have significant relation with the crime.