Basic Analytics 2

author: Dr. Marko Mitic

Problem Description: **AN ANALYTICAL DETECTIVE**

Crime is an international concern, but it is documented and handled in very different ways in different countries. In the United States, violent crimes and property crimes are recorded by the Federal Bureau of Investigation (FBI). Additionally, each city documents crime, and some cities release data regarding crime rates. The city of Chicago, Illinois releases crime data from 2001 onward here https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.

Chicago is the third most populous city in the United States, with a population of over 2.7 million people. There are two main types of crimes in the city: violent crimes, and property crimes. In this problem, we'll focus on one specific type of property crime, called "motor vehicle theft" (sometimes referred to as grand theft auto). This is the act of stealing, or attempting to steal, a car. In this problem, we'll use some basic data analysis in R to understand the motor vehicle thefts in Chicago.

Firstly, let explore the available dateset:

```
GTA=read.csv("mvtWeek1.csv")
str(GTA)
```

```
## 'data.frame':    191641 obs. of  11 variables:
##  $ ID                  : int  8951354 8951141 8952745 8952223 8951608 8950793 8950760 8951611 8951802
##  $ Date                : Factor w/ 131680 levels "1/1/01 0:01",..: 42824 42823 42823 42823 42822 4282
##  $ LocationDescription: Factor w/ 78 levels "ABANDONED BUILDING",..: 72 72 62 72 72 72 72 72 72 72 .
##  $ Arrest              : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ Domestic            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Beat                : int  623 1213 1622 724 211 2521 423 231 1021 1215 ...
##  $ District            : int  6 12 16 7 2 25 4 2 10 12 ...
##  $ CommunityArea       : int  69 24 11 67 35 19 48 40 29 24 ...
##  $ Year                : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
##  $ Latitude            : num  41.8 41.9 42 41.8 41.8 ...
##  $ Longitude           : num  -87.6 -87.7 -87.8 -87.7 -87.6 ...
```

One can observe that the dataset has 191641 observations. The description of the 11 variables are as follows:

- **ID**: a unique identifier for each observation
- **Date**: the date the crime occurred
- **LocationDescription**: the location where the crime occurred
- **Arrest**: whether or not an arrest was made for the crime (TRUE if an arrest was made, and FALSE if an arrest was not made)
- **Domestic**: whether or not the crime was a domestic crime, meaning that it was committed against a family member (TRUE if it was domestic, and FALSE if it was not domestic)
- **Beat**: the area, or "beat" in which the crime occurred. This is the smallest regional division defined by the Chicago police department.
- **District**: the police district in which the crime occured. Each district is composed of many beats, and are defined by the Chicago Police Department.
- **CommunityArea**: the community area in which the crime occurred. Since the 1920s, Chicago has been divided into what are called "community areas" of which there are now 77. The community areas were devised in an attempt to create socially homogeneous regions.
- **Year**: the year in which the crime occurred.

- **Latitude**: the latitude of the location at which the crime occurred.
- **Longitude**: the longitude of the location at which the crime occurred.

Summary statiscs of each variable is what we want to explore next:

```
summary(GTA)
```

```
##        ID                     Date
##  Min.   :1310022   5/16/08 0:00  :    11
##  1st Qu.:2832144   10/17/01 22:00:    10
##  Median :4762956   4/13/04 21:00 :    10
##  Mean   :4968629   9/17/05 22:00 :    10
##  3rd Qu.:7201878   10/12/01 22:00:     9
##  Max.   :9181151   10/13/01 22:00:     9
##                    (Other)       :191582
##                      LocationDescription   Arrest         Domestic
##  STREET                          :156564   Mode :logical   Mode :logical
##  PARKING LOT/GARAGE(NON.RESID.): 14852   FALSE:176105    FALSE:191226
##  OTHER                           :  4573   TRUE :15536     TRUE :415
##  ALLEY                           :  2308   NA's :0         NA's :0
##  GAS STATION                     :  2111
##  DRIVEWAY - RESIDENTIAL          :  1675
##  (Other)                         :  9558
##       Beat         District     CommunityArea        Year
##  Min.   : 111   Min.   : 1.00   Min.   : 0      Min.   :2001
##  1st Qu.: 722   1st Qu.: 6.00   1st Qu.:22      1st Qu.:2003
##  Median :1121   Median :10.00   Median :32      Median :2006
##  Mean   :1259   Mean   :11.82   Mean   :38      Mean   :2006
##  3rd Qu.:1733   3rd Qu.:17.00   3rd Qu.:60      3rd Qu.:2009
##  Max.   :2535   Max.   :31.00   Max.   :77      Max.   :2012
##                 NA's   :43056   NA's   :24616
##     Latitude       Longitude
##  Min.   :41.64   Min.   :-87.93
##  1st Qu.:41.77   1st Qu.:-87.72
##  Median :41.85   Median :-87.68
##  Mean   :41.84   Mean   :-87.68
##  3rd Qu.:41.92   3rd Qu.:-87.64
##  Max.   :42.02   Max.   :-87.52
##  NA's   :2276    NA's   :2276
```

Similarly to previous problem (Basic_Analytics1) the `Date` is a factor variable, so we want to transform it into a more convenient format:

```
DateConvert = as.Date(strptime(GTA$Date, "%m/%d/%y %H:%M"))
```

Next, we transform the `Date` variable in the original dataframe. Months and days can be extracted from the variable `DateConvert` as follows:

```
GTA$Date = DateConvert
GTA$Month = months(DateConvert)
GTA$Weekday = weekdays(DateConvert)
```

We can find the month in which most of the thefts has occured with (October)

```r
table(GTA$Month)
```

```
##
##     April    August  December  February   January      July      June
##     15280     16572     16426     13511     16047     16801     16002
##     March       May  November   October September
##     15758     16035     16063     17086     16060
```

Similarly, Friday is the month in which most of the motor vehicle thefts occured:

```r
table(GTA$Weekday)
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##     29284     27397     27118     26316     27319     26791     27416
```

Next, we can explore the number of arrest per month and per day:

```r
table(GTA$Month, GTA$Arrest)
```

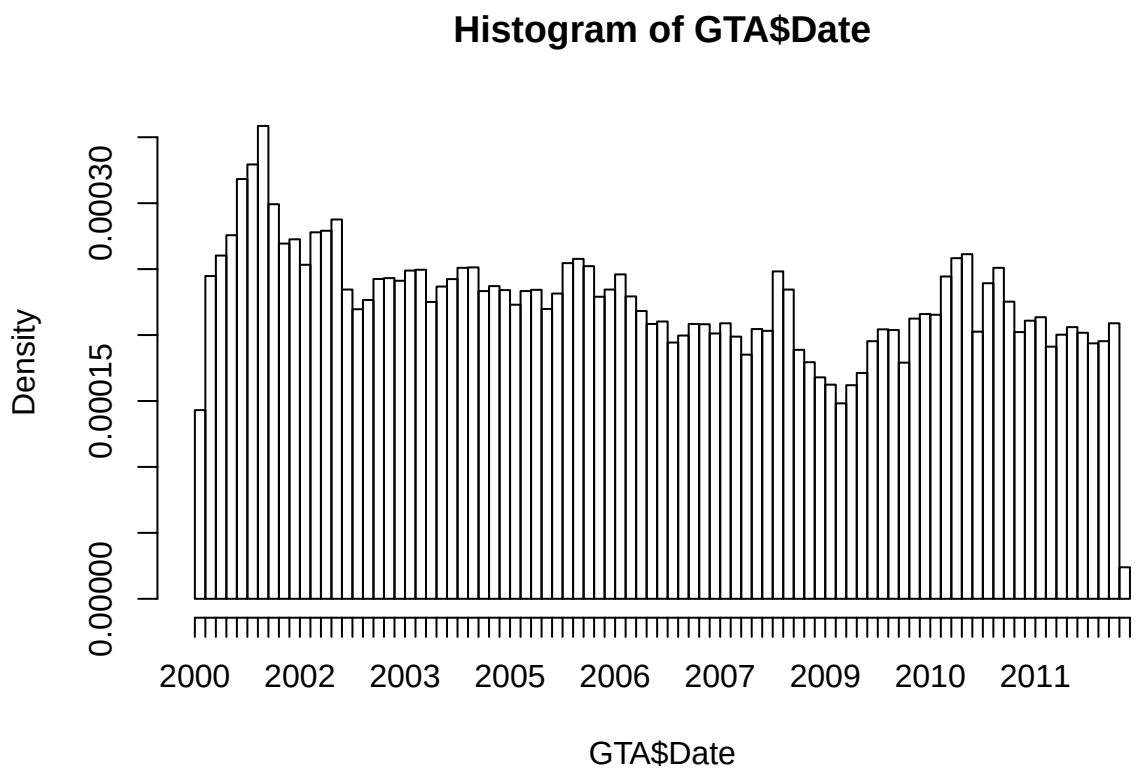
```
##
##             FALSE  TRUE
##   April     14028  1252
##   August    15243  1329
##   December  15029  1397
##   February  12273  1238
##   January   14612  1435
##   July      15477  1324
##   June      14772  1230
##   March     14460  1298
##   May       14848  1187
##   November  14807  1256
##   October   15744  1342
##   September 14812  1248
```

```r
table(GTA$Weekday, GTA$Arrest)
```

```
##
##             FALSE  TRUE
##   Friday    26914  2370
##   Monday    25221  2176
##   Saturday  24863  2255
##   Sunday    23986  2330
##   Thursday  25232  2087
##   Tuesday   24683  2108
##   Wednesday 25206  2210
```

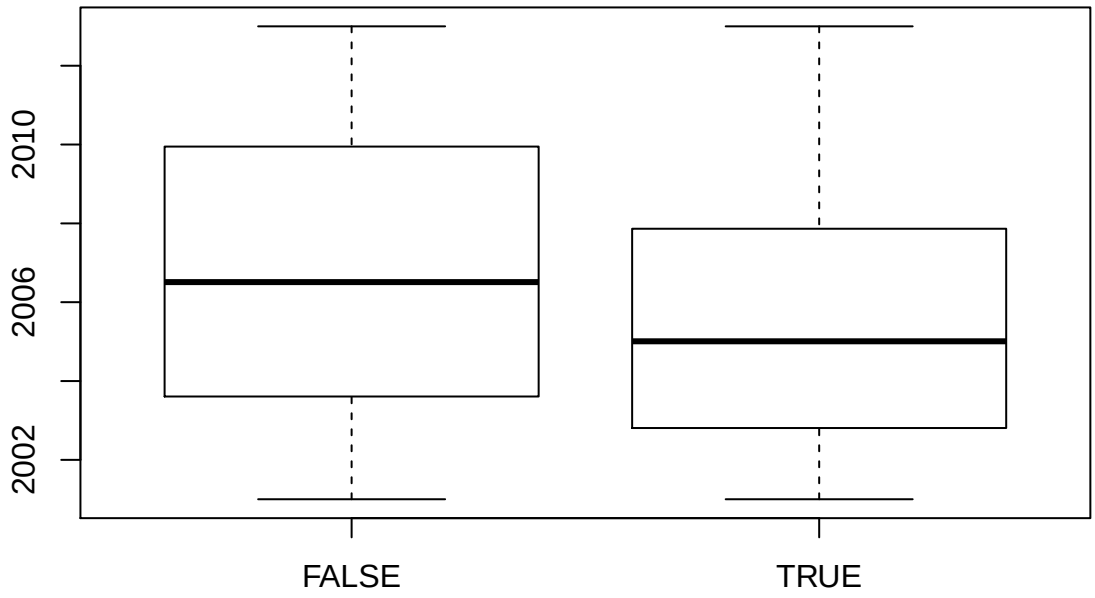
As it known, visualisations help understanding patterns in data better. We can plot histogram of the crimes to see general trend over time:

```
hist(GTA$Date, breaks = 100)
```



We can observe that crime decreses during 2005-2008 period, but increses during 2009-2011. Let's now plot boxplot to see other patterns:

```
boxplot(GTA$Date ~ GTA$Arrest)
```



The `arrest boxplot` is located towards the bottom of the graph, which indicate that there were more crimes for which arrests were made in the first half of the time period. To further confirm this, we can determine proportion of the arrests in selected years:

```
proportion = table(GTA$Arrest,GTA$Year)
proportion2001 = proportion[2,1]/sum(proportion[,1])
proportion2001
```

```
## [1] 0.1041173
```

```
proportion2007 = proportion[2,7]/sum(proportion[,7])
proportion2007
```

```
## [1] 0.08487395
```

```
proportion2012 = proportion[2,12]/sum(proportion[,12])
proportion2012
```

```
## [1] 0.03902924
```

One can easily confirm that the arrests are conducted more frequently in the earlier years of the time period. Finally, we can analyze the locations in which the crimes happened (results suppressed due to reasons of clariness):

```
sort(table(GTA$LocationDescription))
```

We can next subset a dataframe with 5 locations with most crimes:

```
Top5=subset(GTA, GTA$LocationDescription=="STREET" | GTA$LocationDescription=="PARKING LOT/GARAGE(NON.R
            | GTA$LocationDescription=="ALLEY" | GTA$LocationDescription=="GAS STATION"
            | GTA$LocationDescription=="DRIVEWAY - RESIDENTIAL")
str(Top5)
```

```
## 'data.frame':    177510 obs. of  13 variables:
##  $ ID                  : int  8951354 8951141 8952223 8951608 8950793 8950760 8951611 8951802 8950706
##  $ Date                : Date, format: "2012-12-31" "2012-12-31" ...
##  $ LocationDescription: Factor w/ 78 levels "ABANDONED BUILDING",..: 72 72 72 72 72 72 72 72 72 72 .
##  $ Arrest              : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Domestic            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Beat                : int  623 1213 724 211 2521 423 231 1021 1215 1011 ...
##  $ District            : int  6 12 7 2 25 4 2 10 12 10 ...
##  $ CommunityArea       : int  69 24 67 35 19 48 40 29 24 29 ...
##  $ Year                : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
##  $ Latitude            : num  41.8 41.9 41.8 41.8 41.9 ...
##  $ Longitude           : num  -87.6 -87.7 -87.7 -87.6 -87.8 ...
##  $ Month               : chr  "December" "December" "December" "December" ...
##  $ Weekday             : chr  "Monday" "Monday" "Monday" "Monday" ...
```

To make our tables a bit nicer to read, we can refresh this factor variable with:

```
Top5$LocationDescription = factor(Top5$LocationDescription)
```

Finally, using `table` function on Top 5 we can observe that the most arrest happend on `Gas Station` location.

```
table(Top5$Arrest, Top5$LocationDescription)
```

```
##
##          ALLEY DRIVEWAY - RESIDENTIAL GAS STATION
##   FALSE   2059                   1543        1672
##   TRUE     249                    132         439
##
##         PARKING LOT/GARAGE(NON.RESID.) STREET
##   FALSE                          13249 144969
##   TRUE                            1603  11595
```

```
str(Top5)
```

```
## 'data.frame':    177510 obs. of  13 variables:
##  $ ID                  : int  8951354 8951141 8952223 8951608 8950793 8950760 8951611 8951802 8950706
##  $ Date                : Date, format: "2012-12-31" "2012-12-31" ...
##  $ LocationDescription: Factor w/ 5 levels "ALLEY","DRIVEWAY - RESIDENTIAL",..: 5 5 5 5 5 5 5 5 5 5
##  $ Arrest              : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Domestic            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Beat                : int  623 1213 724 211 2521 423 231 1021 1215 1011 ...
```

```
##  $ District           : int  6 12 7 2 25 4 2 10 12 10 ...
##  $ CommunityArea      : int  69 24 67 35 19 48 40 29 24 29 ...
##  $ Year               : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
##  $ Latitude           : num  41.8 41.9 41.8 41.8 41.9 ...
##  $ Longitude          : num  -87.6 -87.7 -87.7 -87.6 -87.8 ...
##  $ Month              : chr  "December" "December" "December" "December" ...
##  $ Weekday            : chr  "Monday" "Monday" "Monday" "Monday" ...
```

Using following code we can calculate proportion of the arrest conducted in Top5 locations:

```
ans=table(Top5$Arrest, Top5$LocationDescription)

rate=ans[2,]/colSums (ans, na.rm = FALSE, dims = 1)
rate
```

```
##                          ALLEY         DRIVEWAY - RESIDENTIAL
##                     0.10788562                     0.07880597
##                    GAS STATION PARKING LOT/GARAGE(NON.RESID.)
##                     0.20795831                     0.10793159
##                         STREET
##                     0.07405917
```

Saturday is marked at the day with highest motor vehicle thefts at `Gas Station`. Interestingly, Saturday is the day in which fewest motor vehicle thefts in `residential driveways` location happened.

```
table(Top5$Weekday, Top5$LocationDescription)
```

```
##
##             ALLEY DRIVEWAY - RESIDENTIAL GAS STATION
##   Friday      385                    257         332
##   Monday      320                    255         280
##   Saturday    341                    202         338
##   Sunday      307                    221         336
##   Thursday    315                    263         282
##   Tuesday     323                    243         270
##   Wednesday   317                    234         273
##
##             PARKING LOT/GARAGE(NON.RESID.) STREET
##   Friday                             2331  23773
##   Monday                             2128  22305
##   Saturday                           2199  22175
##   Sunday                             1936  21756
##   Thursday                           2082  22296
##   Tuesday                            2073  21888
##   Wednesday                          2103  22371
```