R Notebook

Title: "IST687 – JSON & tapply Homework: Accident Analysis"
Name: Sathish Kumar Rajendiran
Week: 5
Date: 05/06/2020

Exercise: Accident Analysis

Step 1: Load the data

```
# Intall Packages

# install.packages("RSQLite")
# install.packages("jsonlite")
# install.packages("RCurl")
# install.packages("RJSONIO")
```

#Step 1: Load the data

```
  <!-- Read in the following JSON dataset -->
  <!-- http://data.maryland.gov/api/views/pdvh-tf2u/rows.json?accessType=DOWNLOAD -->
```

#Step 2: Clean the data

```
library(RJSONIO)
library(RCurl)
library(jsonlite)
```

```
##
## Attaching package: 'jsonlite'

## The following objects are masked from 'package:RJSONIO':
##
##     fromJSON, toJSON
```

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load shared object '/Library
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Library not loaded: /opt/X1
##   Referenced from: /Library/Frameworks/R.framework/Resources/modules//R_X11.so
##   Reason: image not found
```

```
## Could not load tcltk.  Will use slower R code instead.
```

```
## Loading required package: RSQLite
```

```
    marylandData <- "http://opendata.maryland.gov/api/views/pdvh-tf2u/rows.json?accessType=DOWNLOAD"
    jsonResult <- fromJSON(marylandData)
    # summary(jsonResult)

    jsonData <- jsonResult$data
    # summary(jsonData)

# Convert to Dataframe
    dfTemp <- data.frame(jsonData,stringsAsFactors = FALSE)

    # head(dfTemp)

# delete columns 1 through 8

    marylandDF <- dfTemp[, -c(1:8)]
    head(marylandDF)
```

```
##           X9              X10                  X11   X12 X13       X14
## 1 1363000002        Rockville 2012-01-01T00:00:00  2:01   1 SUNDAY
## 2 1296000023           Berlin 2012-01-01T00:00:00 18:01   5 SUNDAY
## 3 1283000016 Prince Frederick 2012-01-01T00:00:00  7:01   2 SUNDAY
## 4 1282000006      Leonardtown 2012-01-01T00:00:00  0:01   1 SUNDAY
## 5 1267000007            Essex 2012-01-01T00:00:00  1:01   1 SUNDAY
## 6 1267000006            Essex 2012-01-01T00:00:00  1:01   1 SUNDAY
##                         X15                           X16  X17 X18
## 1  IS 00495 CAPITAL BELTWAY IS 00270 EISENHOWER MEMORIAL    0   U
## 2 MD 00090 OCEAN CITY EXPWY  CO 00220 ST MARTINS NECK RD 0.25   W
## 3          MD 00765 MAIN ST             CO 00208 DUKE ST  100   S
## 4  MD 00944 MERVELL DEAN RD      MD 00235 THREE NOTCH RD   10   E
## 5    IS 00695 BALTO BELTWAY    IS 00083 HARRISBURG EXPWY  100   S
## 6 IS 00083 HARRISBURG EXPWY        MD 00137 MT CARMEL RD 0.25   S
##              X19 X20        X21  X22 X23 X24       X25             X26
## 1 Not Applicable  15 Montgomery    2 YES  NO       VEH OTHER-COLLISION
## 2 Not Applicable  23  Worcester    1 YES  NO FIXED OBJ OTHER-COLLISION
## 3 Not Applicable   4    Calvert    1 YES  NO FIXED OBJ       FIXED OBJ
## 4 Not Applicable  18  St. Marys    1 YES  NO FIXED OBJ OTHER-COLLISION
## 5 Not Applicable   3  Baltimore    2 YES  NO       VEH OTHER-COLLISION
## 6 Not Applicable   3  Baltimore <NA>  NO YES FIXED OBJ OTHER-COLLISION
```

```
# Assign Column Names
```

```r
colnames(marylandDF) <- c("CASE_NUMBER","BARRACK","ACC_DATE","ACC_TIME","ACC_TIME_CODE","DAY_OF_WEE

# head(marylandDF)

marylandDF <- na.omit(marylandDF)  # Remove NAs from the data

# summary(marylandDF)
# length(marylandDF)
# colnames(marylandDF)
# nrow(marylandDF)

head(marylandDF)
```

```
##   CASE_NUMBER          BARRACK            ACC_DATE ACC_TIME ACC_TIME_CODE
## 1  1363000002        Rockville 2012-01-01T00:00:00     2:01             1
## 2  1296000023           Berlin 2012-01-01T00:00:00    18:01             5
## 3  1283000016 Prince Frederick 2012-01-01T00:00:00     7:01             2
## 4  1282000006      Leonardtown 2012-01-01T00:00:00     0:01             1
## 5  1267000007            Essex 2012-01-01T00:00:00     1:01             1
## 7  1267000005            Essex 2012-01-01T00:00:00     1:01             1
##   DAY_OF_WEEK                      ROAD               INTERSECT_ROAD
## 1   SUNDAY     IS 00495 CAPITAL BELTWAY IS 00270 EISENHOWER MEMORIAL
## 2   SUNDAY    MD 00090 OCEAN CITY EXPWY  CO 00220 ST MARTINS NECK RD
## 3   SUNDAY             MD 00765 MAIN ST             CO 00208 DUKE ST
## 4   SUNDAY     MD 00944 MERVELL DEAN RD      MD 00235 THREE NOTCH RD
## 5   SUNDAY       IS 00695 BALTO BELTWAY    IS 00083 HARRISBURG EXPWY
## 7   SUNDAY             IS 00070 NO NAME       IS 00695 BALTO BELTWAY
##   DIST_FROM_INTERSECT DIST_DIRECTION      CITY_NAME COUNTY_CODE COUNTY_NAME
## 1                   0              U Not Applicable          15  Montgomery
## 2                0.25              W Not Applicable          23   Worcester
## 3                 100              S Not Applicable           4     Calvert
## 4                  10              E Not Applicable          18   St. Marys
## 5                 100              S Not Applicable           3   Baltimore
## 7                 1.5              S Not Applicable           3   Baltimore
##   VEHICLE_COUNT PROP_DEST INJURY COLLI SION_WITH_1 COLLISION_WITH_2
## 1             2       YES     NO               VEH  OTHER-COLLISION
## 2             1       YES     NO         FIXED OBJ  OTHER-COLLISION
## 3             1       YES     NO         FIXED OBJ        FIXED OBJ
## 4             1       YES     NO         FIXED OBJ  OTHER-COLLISION
## 5             2       YES     NO               VEH  OTHER-COLLISION
## 7             1       YES     NO         FIXED OBJ  OTHER-COLLISION
```

Step 3: Understand the data using SQL (via SQLDF)

```r
# Answer the following questions:
# • How many accidents happen on SUNDAY
# • How many accidents had injuries (might need to remove NAs from the data)
# • List the injuries by day

    marylandDF <- na.omit(marylandDF)  # Remove NAs from the data

    sqlResults <- sqldf("select  * from marylandDF limit 5")
    # sqlResults
```

```
      nbrofsudayAccidentsSQL <- sqldf("select count(*)  from marylandDF where lower(trim(Day_of_Week)) =
      cat( "\n Number of accidents happenened on Sunday are: ",unlist(nbrofsudayAccidentsSQL))
```

```
##
##  Number of accidents happenened on Sunday are:  2061
```

```
      nbrofInjuriesSQL <- sqldf("select count(*)  from marylandDF where lower(trim(INJURY)) = 'yes'")
      cat( "\n Number of accidents with injuries are: ",unlist(nbrofInjuriesSQL))
```

```
##
##  Number of accidents with injuries are:  5639
```

```
      injuriesBydaySQL <- sqldf("select trim(DAY_OF_WEEK) as Day, count(*) as Accidents  from marylandDF

      injuriesBydaySQL
```

```
##         Day Accidents
## 1    FRIDAY       915
## 2    MONDAY       795
## 3  SATURDAY       827
## 4    SUNDAY       705
## 5  THURSDAY       864
## 6   TUESDAY       748
## 7 WEDNESDAY       785
```

```
      injuriesByday <- injuriesBydaySQL$Accidents
```

Step 4: Understand the data using tapply

```
  marylandDF$DAY_OF_WEEK <- trimws(marylandDF$DAY_OF_WEEK)

  unique(marylandDF$DAY_OF_WEEK)
```

```
## [1] "SUNDAY"    "MONDAY"    "TUESDAY"   "WEDNESDAY" "THURSDAY"  "FRIDAY"
## [7] "SATURDAY"
```

```
# Subset Method
  # sundayAccidents <- subset(marylandDF,marylandDF$DAY_OF_WEEK=="SUNDAY")
  # NbrOfsundayAccidents <- nrow(sundayAccidents)
  # cat( "\n Number of accidents happen on Sunday are ",NbrOfsundayAccidents)
  #
  # unique(marylandDF$INJURY)
  # injuryAccidents <- subset(marylandDF,marylandDF$INJURY=="YES")
  # NbrOfInjuryAccidents <- nrow(injuryAccidents)
  # cat( "\n Number of accidents happen on Sunday are: ",NbrOfInjuryAccidents)

# tapply Method
  sundayAccidents <- tapply(marylandDF$DAY_OF_WEEK,marylandDF$DAY_OF_WEEK=="SUNDAY",length)
  cat( "\n Number of accidents happen on Sunday are: ",sundayAccidents[[2]])
```

```
##
##  Number of accidents happen on Sunday are:  2061
```

```
  sundayAccidents
```

```
## FALSE  TRUE
## 14202  2061
```

```
  injuryAccidents <- tapply(marylandDF$INJURY,marylandDF$INJURY=="YES",length)
  cat( "\n Number of accidents had injuries are: ",injuryAccidents[[2]])
```

```
##
##  Number of accidents had injuries are:  5639
```

```
  injuryAccidents
```

```
## FALSE  TRUE
## 10624  5639
```

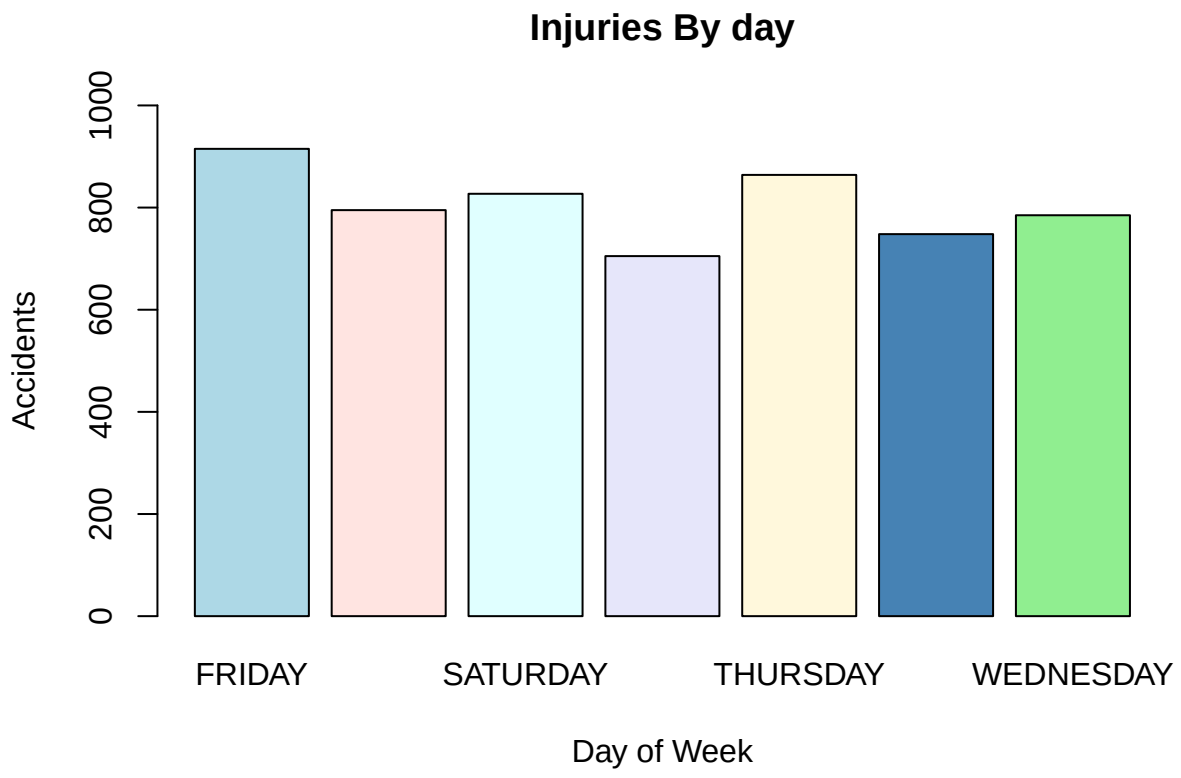
```
  injuriesByday <- tapply(marylandDF$INJURY=="YES",marylandDF$DAY_OF_WEEK,sum)
  cat("\n Injuries by day:",sort(injuriesByday,decreasing = TRUE))
```

```
##
##  Injuries by day: 915 864 827 795 785 748 705
```

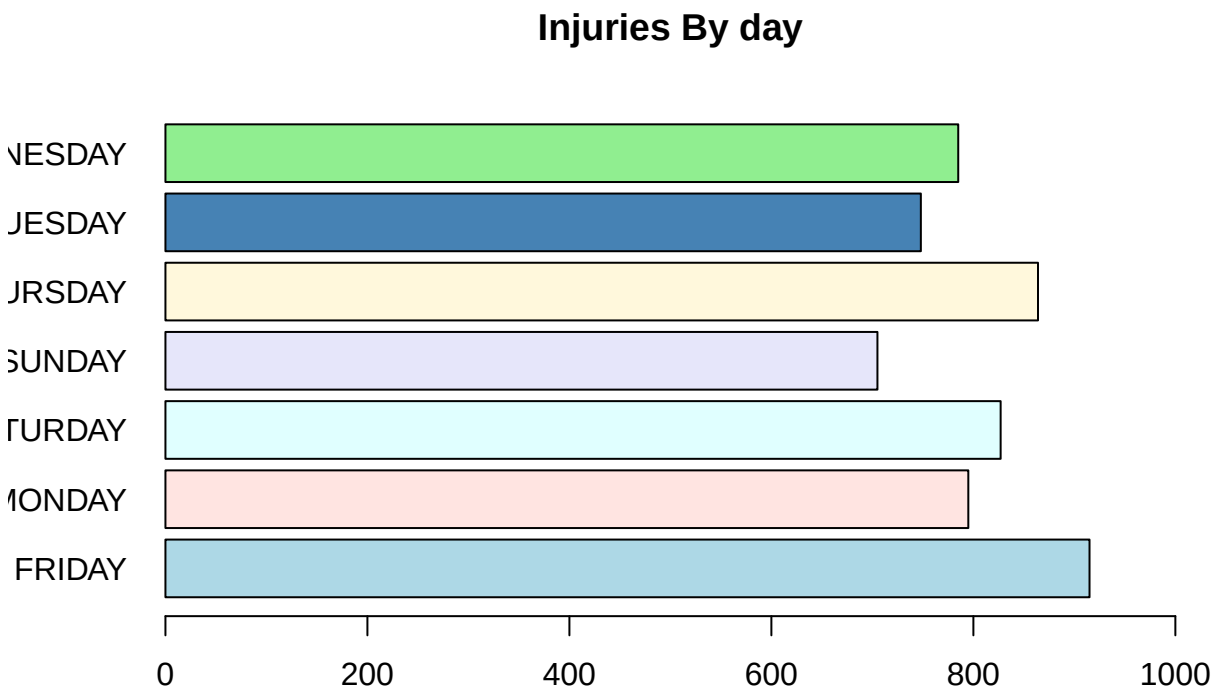
```
  sort(injuriesByday,decreasing = TRUE)
```

```
##    FRIDAY  THURSDAY  SATURDAY    MONDAY WEDNESDAY   TUESDAY    SUNDAY
##       915       864       827       795       785       748       705
```

```
#Graphs

      days <- c(injuriesBydaySQL$Day)
      accidents <- injuriesBydaySQL$Accidents

      colors <- c("lightblue","mistyrose","lightcyan","lavender","cornsilk","steelblue","lightgreen")

# Barplot
      barplot(accidents, main = "Injuries By day"
              , names.arg = days, xlab = "Day of Week", ylab = "Accidents"
              ,ylim = c(0,1000), beside=TRUE, col = colors,border = "black")
```




```
# Horizontal barplot
      barplot(main = "Injuries By day"
              , height=accidents, names=days,xlim = c(0,1000)
              , col = colors,border = "black",horiz=T, las=1)
```




```
# boxplot(split(injuriesBydaySQL$Accidents,injuriesBydaySQL$Day),main='Injuries By day', border =
```