HW5

Diego Valdes

February 11, 2019

```
rm(list=ls()) # clear work space
#dev.off(dev.list()["RStudioGD"]) # clear plots

# import necessary libraries
suppressWarnings(library("jsonlite"))
suppressWarnings(library("RCurl"))
```

```
## Loading required package: bitops
```

```
#library(data.table)
suppressWarnings(library(plyr))
suppressWarnings(library('sqldf'))
```

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

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

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

```
suppressWarnings(library(stringi))


# Load the data
# get json file
jsonUrl ="http://data.maryland.gov/api/views/pdvh-tf2u/rows.json"
jsonObject = getURL(jsonUrl)
json_file = fromJSON(jsonUrl)

# extract each element from the list
length(json_file)
```

```
## [1] 2
```

```
meta = json_file[[1]] # first element of the list
data = json_file[[2]] # second element of the list

# Clean the data
#remove last 8 cols
data = data[, c(-1:-8)]

# names for cols
namesOfColumns = c("CASE_NUMBER","BARRACK","ACC_DATE","ACC_TIME","ACC_TIME_CODE",
                   "DAY_OF_WEEK", "ROAD","INTERSECT_ROAD","DIST_FROM_INTERSECT",
  "DIST_DIRECTION","CITY_NAME","COUNTY_CODE","COUNTY_NAME","VEHICLE_COUNT",
  "PROP_DEST","INJURY","COLLISION_WITH_1","COLLISION_WITH_2")

# assign col names
colnames(data) = namesOfColumns
```

```r
# What are we dealing with?
summary(data)
```

```
##      CASE_NUMBER            BARRACK                     ACC_DATE
##  1257000644:    3   Forestville : 1911   2012-07-21T00:00:00:  113
##  1262006287:    3   College Park: 1536   2012-01-21T00:00:00:  111
##  1266001445:    3   Frederick   : 1501   2012-12-09T00:00:00:   98
##  1250003311:    2   Bel Air     : 1385   2012-06-12T00:00:00:   95
##  1250005131:    2   Rockville   : 1381   2012-08-26T00:00:00:   92
##  1251005139:    2   (Other)     :10194   2012-02-29T00:00:00:   86
##  (Other)   :18623   NA's        :  730   (Other)            :18043
##     ACC_TIME     ACC_TIME_CODE    DAY_OF_WEEK
##  17:11  :  160   1:1665        FRIDAY   :3014
##  17:12  :  149   2:2645        MONDAY   :2554
##  17:08  :  136   3:3330        SATURDAY :2732
##  16:05  :  133   4:4109        SUNDAY   :2373
##  16:08  :  130   5:4540        THURSDAY :2671
##  14:06  :  129   6:2349        TUESDAY  :2676
##  (Other):17801                 WEDNESDAY:2618
##                           ROAD                         INTERSECT_ROAD
##  IS 00095 CAPITAL BELTWAY   : 1163   IS 00695 BALTO BELTWAY  :  173
##  IS 00495 CAPITAL BELTWAY   :  874   MD 00185 CONNECTICUT AVE:  153
##  IS 00695 BALTO BELTWAY     :  840   MD 00100 NO NAME        :  133
##  US 00301 CRAIN HWY         :  672   IS 00095 CAPITAL BELTWAY:  129
##  IS 00095 NO NAME           :  618   MD 00201 KENILWORTH AVE :  112
##  IS 00095 J F K MEMORIAL HWY:  568   (Other)                 :17937
##  (Other)                    :13903   NA's                    :    1
##  DIST_FROM_INTERSECT DIST_DIRECTION          CITY_NAME      COUNTY_CODE
##  0      :4941        E   :2656      Not Applicable:18170   16     :3453
##  100    :1399        N   :4493      Mount Airy    :   59   12     :1659
##  0.25   :1381        S   :4586      Westminster   :   20   3      :1597
##  0.5    :1226        U   :3974      Berlin        :   19   10     :1502
##  500    :1224        W   :2548      Leonardtown   :   10   15     :1387
##  (Other):8454        NA's: 381      (Other)       :  164   (Other):9006
##  NA's   :  13                       NA's          :  196   NA's   :  34
##           COUNTY_NAME   VEHICLE_COUNT  PROP_DEST     INJURY
##  Prince Georges:3453   2      :7816   NO  : 7071   NO  :12204
##  Harford       :1659   1      :6864   YES :11566   YES : 6433
##  Baltimore     :1597   3      :2083   NA's:    1   NA's:    1
##  Frederick     :1502   4      : 480
##  Montgomery    :1387   5      : 105
##  (Other)       :9006   (Other):  39
##  NA's          :  34   NA's   :1251
##         COLLISION_WITH_1        COLLISION_WITH_2
##  VEH             :10675   OTHER-COLLISION:13644
##  FIXED OBJ       : 4299   FIXED OBJ      : 2644
##  OTHER-COLLISION: 2205    VEH            : 1759
##  ANIMAL          :  846   NON-COLLISION  :  559
##  NON-COLLISION   :  465   PED            :   17
##  (Other)         :  147   (Other)        :   14
##  NA's            :    1   NA's           :    1
```

```r
str(data)
```

```
##  chr [1:18638, 1:18] "1363000002" "1296000023" "1283000016" ...
```

```
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:18] "CASE_NUMBER" "BARRACK" "ACC_DATE" "ACC_TIME" ...
```

```
# Lets turn it into a dataframe
data = as.data.frame(data)
#data = data.table(data)
summary(data)
```

```
##      CASE_NUMBER            BARRACK                     ACC_DATE
##  1257000644:    3   Forestville : 1911   2012-07-21T00:00:00:  113
##  1262006287:    3   College Park: 1536   2012-01-21T00:00:00:  111
##  1266001445:    3   Frederick   : 1501   2012-12-09T00:00:00:   98
##  1250003311:    2   Bel Air     : 1385   2012-06-12T00:00:00:   95
##  1250005131:    2   Rockville   : 1381   2012-08-26T00:00:00:   92
##  1251005139:    2   (Other)     :10194   2012-02-29T00:00:00:   86
##  (Other)   :18623   NA's        :  730   (Other)            :18043
##     ACC_TIME     ACC_TIME_CODE    DAY_OF_WEEK
##  17:11  :  160   1:1665        FRIDAY   :3014
##  17:12  :  149   2:2645        MONDAY   :2554
##  17:08  :  136   3:3330        SATURDAY :2732
##  16:05  :  133   4:4109        SUNDAY   :2373
##  16:08  :  130   5:4540        THURSDAY :2671
##  14:06  :  129   6:2349        TUESDAY  :2676
##  (Other):17801                 WEDNESDAY:2618
##                          ROAD                         INTERSECT_ROAD
##  IS 00095 CAPITAL BELTWAY   : 1163   IS 00695 BALTO BELTWAY  :  173
##  IS 00495 CAPITAL BELTWAY   :  874   MD 00185 CONNECTICUT AVE:  153
##  IS 00695 BALTO BELTWAY     :  840   MD 00100 NO NAME        :  133
##  US 00301 CRAIN HWY         :  672   IS 00095 CAPITAL BELTWAY:  129
##  IS 00095 NO NAME           :  618   MD 00201 KENILWORTH AVE :  112
##  IS 00095 J F K MEMORIAL HWY:  568   (Other)                 :17937
##  (Other)                    :13903   NA's                    :    1
##  DIST_FROM_INTERSECT DIST_DIRECTION          CITY_NAME      COUNTY_CODE
##  0      :4941        E   :2656      Not Applicable:18170   16     :3453
##  100    :1399        N   :4493      Mount Airy    :   59   12     :1659
##  0.25   :1381        S   :4586      Westminster   :   20   3      :1597
##  0.5    :1226        U   :3974      Berlin        :   19   10     :1502
##  500    :1224        W   :2548      Leonardtown   :   10   15     :1387
##  (Other):8454        NA's: 381      (Other)       :  164   (Other):9006
##  NA's   :  13                       NA's          :  196   NA's   :  34
##          COUNTY_NAME   VEHICLE_COUNT  PROP_DEST     INJURY
##  Prince Georges:3453   2      :7816   NO  : 7071   NO  :12204
##  Harford       :1659   1      :6864   YES :11566   YES : 6433
##  Baltimore     :1597   3      :2083   NA's:    1   NA's:    1
##  Frederick     :1502   4      : 480
##  Montgomery    :1387   5      : 105
##  (Other)       :9006   (Other):  39
##  NA's          :  34   NA's   :1251
##         COLLISION_WITH_1        COLLISION_WITH_2
##  VEH             :10675   OTHER-COLLISION:13644
##  FIXED OBJ       : 4299   FIXED OBJ      : 2644
##  OTHER-COLLISION: 2205    VEH            : 1759
##  ANIMAL          :  846   NON-COLLISION  :  559
##  NON-COLLISION   :  465   PED            :   17
```

```
##  (Other)        :  147   (Other)        :   14
##  NA's           :    1   NA's           :    1
```

```
str(data)
```

```
## 'data.frame':    18638 obs. of  18 variables:
##  $ CASE_NUMBER        : Factor w/ 18571 levels "1056008704","1057002761",..: 18555 18168 17488 17116
##  $ BARRACK            : Factor w/ 22 levels "Bel Air","Berlin",..: 19 2 17 14 7 7 7 4 4 4 ...
##  $ ACC_DATE           : Factor w/ 366 levels "2012-01-01T00:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ ACC_TIME           : Factor w/ 288 levels "0:01","0:02",..: 145 121 253 1 13 13 13 241 73 241 ...
##  $ ACC_TIME_CODE      : Factor w/ 6 levels "1","2","3","4",..: 1 5 2 1 1 1 1 2 4 2 ...
##  $ DAY_OF_WEEK        : Factor w/ 7 levels "FRIDAY   ","MONDAY   ",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ ROAD               : Factor w/ 2460 levels "  "," ECI Annex Parkin Lot",..: 900 1200 1770 1810 9
##  $ INTERSECT_ROAD     : Factor w/ 6750 levels "  "," Columbia Park Rd",..: 2922 1098 1035 3489 2862
##  $ DIST_FROM_INTERSECT: Factor w/ 137 levels "0","0.0050000000000000001",..: 1 22 47 46 47 22 44 22 1
##  $ DIST_DIRECTION     : Factor w/ 5 levels "E","N","S","U",..: 4 5 3 1 3 3 3 3 1 NA ...
##  $ CITY_NAME          : Factor w/ 71 levels "Aberdeen","Accident",..: 47 47 47 47 47 47 47 47 47 47
##  $ COUNTY_CODE        : Factor w/ 26 levels "0","1","10","11",..: 8 17 20 11 19 19 19 9 9 9 ...
##  $ COUNTY_NAME        : Factor w/ 26 levels "Allegany","Anne Arundel",..: 16 26 5 21 3 3 3 18 18 18
##  $ VEHICLE_COUNT      : Factor w/ 10 levels "1","10","2","3",..: 3 1 1 1 3 NA 1 3 1 1 ...
##  $ PROP_DEST          : Factor w/ 2 levels "NO","YES": 2 2 2 2 2 1 2 2 2 2 ...
##  $ INJURY             : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 2 1 1 2 1 ...
##  $ COLLISION_WITH_1   : Factor w/ 7 levels "ANIMAL","BICYCLE",..: 7 3 3 3 7 3 3 7 3 3 ...
##  $ COLLISION_WITH_2   : Factor w/ 7 levels "ANIMAL","BICYCLE",..: 5 5 3 5 5 5 5 5 5 5 ...
```

```
# sqldf
#number of sundays
data$DAY_OF_WEEK = trimws(data$DAY_OF_WEEK, which = c("both", "left", "right") ) # get rid of white spa
result = sqldf(stri_paste("select count(*) from data where data.DAY_OF_WEEK = 'SUNDAY'"))
result
```

```
##   count(*)
## 1     2373
```

```
# number of injuries
result = sqldf(stri_paste("select count(*) from data where data.INJURY = 'YES'"))
result
```

```
##   count(*)
## 1     6433
```

```
# injures by day
data$INJURY0 = ifelse(data$INJURY == 'YES', 1, 0)
result = sqldf(stri_paste("select data.DAY_OF_WEEK, sum(data.INJURY0) from data group by data.DAY_OF_WE
result
```

```
##   DAY_OF_WEEK sum(data.INJURY0)
## 1      FRIDAY              1043
## 2      MONDAY               915
## 3    SATURDAY               950
## 4      SUNDAY               818
## 5    THURSDAY               968
## 6     TUESDAY               843
## 7   WEDNESDAY               896
```

```
#tapply

# injuries on sunday
```

```r
data$Sunday = ifelse(data$DAY_OF_WEEK == 'SUNDAY', 1, 0)
result = tapply(data$Sunday , data$DAY_OF_WEEK == 'SUNDAY', sum) # how many happened on sunday
result
```

```
## FALSE  TRUE
##     0  2373
```

```r
# accidents with injuries
result = sum(sapply(data$INJURY0, sum, na.rm = TRUE))
result
```

```
## [1] 6433
```

```r
# injuries by day
tapply(data$INJURY0, data$DAY_OF_WEEK, sum, na.rm = TRUE) # accidents by day
```

```
##    FRIDAY    MONDAY  SATURDAY    SUNDAY  THURSDAY   TUESDAY WEDNESDAY
##      1043       915       950       818       968       843       896
```