Katherine HW 5

Katherine Penney

8/13/2021

```
library(jsonlite)
library(sqldf)
```

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

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

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

Step 1: Load the data

```
theURL <- "http://opendata.maryland.gov/api/views/pdvh-tf2u/rows.json?accessType=DOWNLOAD"
thedata <- theURL
json_data <- fromJSON(thedata)
```

Step 2: Clean the data

Use 2nd dataset

```
thedataset <- json_data[[2]]
```

Get rid of 1st 8 column

```
cleandata <- thedataset[,-1:-8]
cleanerdata <- data.frame(cleandata)
```

names columns

```
namesOfColumns <-c("CASE_NUMBER","BARRACK","ACC_DATE","ACC_TIME","ACC_TIME_CODE","DAY_OF_WEEK","ROAD","
colnames(cleanerdata) <- namesOfColumns
colnames(cleanerdata)
```

```
##  [1] "CASE_NUMBER"         "BARRACK"             "ACC_DATE"
##  [4] "ACC_TIME"            "ACC_TIME_CODE"       "DAY_OF_WEEK"
##  [7] "ROAD"                "INTERSECT_ROAD"      "DIST_FROM_INTERSECT"
## [10] "DIST_DIRECTION"      "CITY_NAME"           "COUNTY_CODE"
## [13] "COUNTY_NAME"         "VEHICLE_COUNT"       "PROP_DEST"
## [16] "INJURY"              "COLLISION_WITH_1"    "COLLISION_WITH_2"
```

Remove NAs

```
removena <- na.omit(cleanerdata)
cleanerdata <- removena
```

Clearing spaces

```
cleanerdata$DAY_OF_WEEK <- gsub(" ", "", cleanerdata$DAY_OF_WEEK)
```

Step 3: Understandinf the data using SQL

How many accidents happened on SUNDAY

```
accidents <- sqldf("SELECT COUNT(cleanerdata.DAY_OF_WEEK) FROM cleanerdata WHERE DAY_OF_WEEK = 'SUNDAY'
accidents
```

```
##   COUNT(cleanerdata.DAY_OF_WEEK)
## 1                           2061
```

How many accdents had injuries

```
injuries <- sqldf("SELECT COUNT(cleanerdata.INJURY) FROM cleanerdata WHERE INJURY = 'YES'")
injuries
```

```
##   COUNT(cleanerdata.INJURY)
## 1                      5639
```

The list of injuries by day

```
inj_day <- sqldf("SELECT DAY_OF_WEEK, COUNT(cleanerdata.INJURY) FROM cleanerdata WHERE INJURY = 'YES' G
inj_day
```

```
##   DAY_OF_WEEK COUNT(cleanerdata.INJURY)
## 1      FRIDAY                       915
## 2      MONDAY                       795
## 3    SATURDAY                       827
## 4      SUNDAY                       705
## 5    THURSDAY                       864
## 6     TUESDAY                       748
## 7   WEDNESDAY                       785
```

Step 4: Understand data using tapply

Accidents on Sunday

```
tapacc <- tapply(cleanerdata$CASE_NUMBER, cleanerdata$DAY_OF_WEEK == 'SUNDAY', length)
tapacc
```

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

Accidents with injury

```
tapinj <- tapply(cleanerdata$CASE_NUMBER, cleanerdata$INJURY == 'YES', length)
tapinj
```

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

Injuries by day of week

```
tapinjday <- tapply(cleanerdata$DAY_OF_WEEK, cleanerdata$DAY_OF_WEEK, length)
tapinjday
```

```
##    FRIDAY    MONDAY  SATURDAY    SUNDAY  THURSDAY   TUESDAY WEDNESDAY
##      2630      2207      2377      2061      2356      2331      2301
```