Bing-Je_Wu_HW5

Bing-Je Wu

5/6/2019

Step 1: Load the data

```
library(curl)
library(RSocrata)
Temp_JSON <- "https://opendata.maryland.gov/resource/pdvh-tf2u.json"
MyJSON_Full <- read.socrata(Temp_JSON)
```

Step 2: Clean the data

```
MyJSON_org <- data.frame(MyJSON_Full$case_number,MyJSON_Full$barrack,
                         MyJSON_Full$acc_date,MyJSON_Full$acc_time,
                         MyJSON_Full$acc_time_code,MyJSON_Full$day_of_week,
                         MyJSON_Full$road,MyJSON_Full$intersect_road,
                         MyJSON_Full$dist_from_intersect,MyJSON_Full$dist_direction,
                         MyJSON_Full$city_name,MyJSON_Full$county_code,
                         MyJSON_Full$county_name,MyJSON_Full$vehicle_count,
                         MyJSON_Full$prop_dest,MyJSON_Full$injury,
                         MyJSON_Full$collision_with_1,MyJSON_Full$collision_with_2
                         )
```

```
colnames(MyJSON_org)<-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")
```

#analyze the dataset

```
str(MyJSON_org)
```

```
## '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           : POSIXct, format: "2012-01-01" "2012-01-01" ...
##  $ 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 ...
```

#normalize the columns

```
MyJSON_org$DAY_OF_WEEK <- gsub('\\ ','',MyJSON_org$DAY_OF_WEEK)
```

#remove NAs

```
colSums(is.na(MyJSON_org))
```

```
##         CASE_NUMBER             BARRACK            ACC_DATE
##                   0                 730                   0
##            ACC_TIME       ACC_TIME_CODE         DAY_OF_WEEK
##                   0                   0                   0
##                ROAD      INTERSECT_ROAD DIST_FROM_INTERSECT
##                   0                   1                  13
##      DIST_DIRECTION           CITY_NAME         COUNTY_CODE
##                 381                 196                  34
##         COUNTY_NAME       VEHICLE_COUNT           PROP_DEST
##                  34                1251                   1
##              INJURY    COLLISION_WITH_1    COLLISION_WITH_2
##                   1                   1                   1
```

#create a new dataframe without NAs

```
MyJSON_clean <- na.omit(MyJSON_org)
colSums(is.na(MyJSON_clean))
```

```
##         CASE_NUMBER             BARRACK            ACC_DATE
##                   0                   0                   0
##            ACC_TIME       ACC_TIME_CODE         DAY_OF_WEEK
##                   0                   0                   0
##                ROAD      INTERSECT_ROAD DIST_FROM_INTERSECT
##                   0                   0                   0
##      DIST_DIRECTION           CITY_NAME         COUNTY_CODE
##                   0                   0                   0
##         COUNTY_NAME       VEHICLE_COUNT           PROP_DEST
##                   0                   0                   0
##              INJURY    COLLISION_WITH_1    COLLISION_WITH_2
##                   0                   0                   0
```

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

How many accidents happen on SUNDAY

```
library(sqldf)
library(gsubfn)
library(proto)
sqldf("select DAY_OF_WEEK, count(*) as Accidents from MyJSON_org
      where DAY_OF_WEEK='SUNDAY' ")
```

```
##   DAY_OF_WEEK Accidents
## 1      SUNDAY      2373
```

How many accidents had injuries (might need to remove NAs from the data)

```
sqldf("select injury, count(*) as Accidents from MyJSON_clean
      where INJURY = 'YES'")
```

```
##   INJURY Accidents
## 1    YES      5639
```

List the injuries by day

```
sqldf("select DAY_OF_WEEK,count(injury) as Injuries from MyJSON_org
      group by DAY_OF_WEEK")
```

```
##   DAY_OF_WEEK Injuries
## 1      FRIDAY     3014
## 2      MONDAY     2554
## 3    SATURDAY     2731
## 4      SUNDAY     2373
## 5    THURSDAY     2671
## 6     TUESDAY     2676
## 7   WEDNESDAY     2618
```

Step 4: Understand the data using tapply

How many accidents happen on SUNDAY

```
tapply(MyJSON_org$INJURY,MyJSON_org$DAY_OF_WEEK,length)[4]
```

```
## SUNDAY
##   2373
```

How many accidents had injuries (might need to remove NAs from the data)

```
tapply(MyJSON_clean$INJURY,MyJSON_clean$INJURY,length)[2]
```

```
##  YES
## 5639
```

List the injuries by day

```
tapply(MyJSON_org$INJURY,MyJSON_org$DAY_OF_WEEK, length)
```

```
##    FRIDAY    MONDAY  SATURDAY    SUNDAY  THURSDAY   TUESDAY WEDNESDAY
##      3014      2554      2732      2373      2671      2676      2618
```