Sharat_Sripada_HW5.R

ssharat

2020-02-16

```
#
#      Course: IST-687
#      Name: Sharat Sripada
#      Homework #4
#      Due Date: 2/9/2020
#      Date Submitted: 2/9/2020
#      Topic: JSON & tapply Homework: Accident Analysis


# install.packages("RCurl")
# install.packages("curl")
# install.packages("stringr")
library("RCurl")
library("sqldf")

## Loading required package: gsubfn

## Loading required package: proto

## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to
load shared object
'/Library/Frameworks/R.framework/Resources/modules//R_X11.so':
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6):
Library not loaded: /opt/X11/lib/libSM.6.dylib
##   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

library("jsonlite")
library("stringr")

# Load the data
url <- "https://opendata.maryland.gov/resource/pdvh-tf2u.json"
document<-fromJSON(txt=url)
str(document)

## 'data.frame':    1000 obs. of  18 variables:
##  $ case_number        : chr  "1363000002" "1296000023" "1283000016"
"1282000006" ...
```

```
##  $ barrack            : chr  "Rockville" "Berlin" "Prince Frederick"
"Leonardtown" ...
##  $ acc_date           : chr  "2012-01-01T00:00:00.000" "2012-01-
01T00:00:00.000" "2012-01-01T00:00:00.000" "2012-01-01T00:00:00.000" ...
##  $ acc_time           : chr  "2:01" "18:01" "7:01" "0:01" ...
##  $ acc_time_code      : chr  "1" "5" "2" "1" ...
##  $ day_of_week        : chr  "SUNDAY   " "SUNDAY   " "SUNDAY   " "SUNDAY
" ...
##  $ road               : chr  "IS 00495 CAPITAL BELTWAY" "MD 00090 OCEAN
CITY EXPWY" "MD 00765 MAIN ST" "MD 00944 MERVELL DEAN RD" ...
##  $ intersect_road     : chr  "IS 00270 EISENHOWER MEMORIAL" "CO 00220 ST
MARTINS NECK RD" "CO 00208 DUKE ST" "MD 00235 THREE NOTCH RD" ...
##  $ dist_from_intersect: chr  "0" "0.25" "100" "10" ...
##  $ dist_direction     : chr  "U" "W" "S" "E" ...
##  $ city_name          : chr  "Not Applicable" "Not Applicable" "Not
Applicable" "Not Applicable" ...
##  $ county_code        : chr  "15" "23" "4" "18" ...
##  $ county_name        : chr  "Montgomery" "Worcester" "Calvert" "St.
Marys" ...
##  $ vehicle_count      : chr  "2" "1" "1" "1" ...
##  $ prop_dest          : chr  "YES" "YES" "YES" "YES" ...
##  $ injury             : chr  "NO" "NO" "NO" "NO" ...
##  $ collision_with_1   : chr  "VEH" "FIXED OBJ" "FIXED OBJ" "FIXED OBJ" ...
##  $ collision_with_2   : chr  "OTHER-COLLISION" "OTHER-COLLISION" "FIXED
OBJ" "OTHER-COLLISION" ...
```

```
# > str(document)
# 'data.frame': 1000 obs. of  18 variables:
# .
# .

# Cleansing the data (2x Steps as below)
document_cleanse <- document

# Step-1: Omit all NAs
document_cleanse_omit_nas <- na.omit(document)
str(document_cleanse)
```

```
## 'data.frame':    1000 obs. of  18 variables:
##  $ case_number        : chr  "1363000002" "1296000023" "1283000016"
"1282000006" ...
##  $ barrack            : chr  "Rockville" "Berlin" "Prince Frederick"
"Leonardtown" ...
##  $ acc_date           : chr  "2012-01-01T00:00:00.000" "2012-01-
01T00:00:00.000" "2012-01-01T00:00:00.000" "2012-01-01T00:00:00.000" ...
##  $ acc_time           : chr  "2:01" "18:01" "7:01" "0:01" ...
##  $ acc_time_code      : chr  "1" "5" "2" "1" ...
##  $ day_of_week        : chr  "SUNDAY   " "SUNDAY   " "SUNDAY   " "SUNDAY
" ...
##  $ road               : chr  "IS 00495 CAPITAL BELTWAY" "MD 00090 OCEAN
```

```
CITY EXPWY" "MD 00765 MAIN ST" "MD 00944 MERVELL DEAN RD" ...
##  $ intersect_road     : chr  "IS 00270 EISENHOWER MEMORIAL" "CO 00220 ST
MARTINS NECK RD" "CO 00208 DUKE ST" "MD 00235 THREE NOTCH RD" ...
##  $ dist_from_intersect: chr  "0" "0.25" "100" "10" ...
##  $ dist_direction     : chr  "U" "W" "S" "E" ...
##  $ city_name          : chr  "Not Applicable" "Not Applicable" "Not
Applicable" "Not Applicable" ...
##  $ county_code        : chr  "15" "23" "4" "18" ...
##  $ county_name        : chr  "Montgomery" "Worcester" "Calvert" "St.
Marys" ...
##  $ vehicle_count      : chr  "2" "1" "1" "1" ...
##  $ prop_dest          : chr  "YES" "YES" "YES" "YES" ...
##  $ injury             : chr  "NO" "NO" "NO" "NO" ...
##  $ collision_with_1   : chr  "VEH" "FIXED OBJ" "FIXED OBJ" "FIXED OBJ" ...
##  $ collision_with_2   : chr  "OTHER-COLLISION" "OTHER-COLLISION" "FIXED
OBJ" "OTHER-COLLISION" ...

# > str(document_cleanse)
# 'data.frame': 876 obs. of  18 variables:
# .
# .

# Step-2: Remove spaces from a few columns like day_of_week
document_cleanse$day_of_week <- str_replace(document_cleanse$day_of_week, "\
.*","")
document_cleanse_omit_nas$day_of_week <-
str_replace(document_cleanse_omit_nas$day_of_week, "\ .*","")

# Use the sqldf function of R to interpret the data-frame
# using SQL commands
# How many accidents happen on SUNDAY
sqldf("select count(day_of_week) from document_cleanse where
day_of_week=='SUNDAY'")

##   count(day_of_week)
## 1                 95

# How many accidents had injuries
sqldf("select count(injury) from document_cleanse where injury=='YES'")

##   count(injury)
## 1           301

# Remove NAs from the data & get the counts again
sqldf("select count(day_of_week) from document_cleanse_omit_nas where
day_of_week=='SUNDAY'")

##   count(day_of_week)
## 1                 86
```

```
sqldf("select count(injury) from document_cleanse_omit_nas where
injury=='YES'")

##   count(injury)
## 1           272

# Using tapply to achieve the same tasks
tapply(document_cleanse$day_of_week, document_cleanse$day_of_week=='SUNDAY',
length)

## FALSE  TRUE
##   905    95

tapply(document_cleanse$injury, document_cleanse$injury=='YES', length)

## FALSE  TRUE
##   699   301
```