**Unexecuted Code**

#Homework 5 - 2 Aug 2018

#Step 1: Load the data

#-------------------------------------------------------------------------

#Read in the following JSON dataset

#Library for http get requests

library(RCurl)

#Library for json parsing

library(jsonlite)

#HTTP get request

urlObj <- getURL("http://data.maryland.gov/api/views/pdvh-tf2u/rows.json?accessType=DOWNLOAD")

#Converting the url string to json

jsonObj <- fromJSON(urlObj)

#Getting the data from the JSON list

data <- jsonObj$data

#Converting the list into the data frame

df <- data.frame(data, stringsAsFactors = F)

#Viewing the structure of the data frame, making sure all looks good

str(df)

#Step 2: Clean the data

#-------------------------------------------------------------------------

#deleting the first 8 columns

df <- df[,-1:-8]

#names of the columns

namesOfColumns <- c("CASE\_NUMBER","BARRACK","ACC\_DATE","ACC\_TIME","ACC\_TIME\_CODE",

"DAY\_OF\_WEEK","ROAD","INTERSECT\_ROAD","DIST\_FROM\_INTERSECT","DIST\_DIRECTION",

"CITY\_NA ME","COUNTY\_CODE","COUNTY\_NAME","VEHICLE\_COUNT","PROP\_DEST","INJURY",

"COLLISION\_WITH\_1","COLLISION\_WITH\_2")

#Assigning the names of teh coulmns to the data frame

colnames(df) <- namesOfColumns

#making sure the column headers make sense

head(df)

#Deleting the space at the end of the day of the week

df$DAY\_OF\_WEEK <- gsub(" ", "", df$DAY\_OF\_WEEK)

#Converting the distance from intersection into numeric format

df$DIST\_FROM\_INTERSECT <- as.numeric(df$DIST\_FROM\_INTERSECT)

#Converting vehicle count into numeric

df$VEHICLE\_COUNT <- as.numeric(df$VEHICLE\_COUNT)

#Making day of week a factor

df$DAY\_OF\_WEEK <- as.factor(df$DAY\_OF\_WEEK)

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

#-------------------------------------------------------------------------

#Loading library fir sql commands

library(sqldf)

#How many accidents happen on SUNDAY

sqldf("SELECT count(CASE\_NUMBER) AS 'Sunday Crashes' FROM df WHERE df.DAY\_OF\_WEEK = 'SUNDAY'")

#How many accidents had injuries

sqldf("SELECT count(CASE\_NUMBER) AS 'Accidents with Injuries' FROM df WHERE INJURY = 'YES'")

# INJURIES by day

#Creating the select statement with paste

selectStr = paste("SELECT DAY\_OF\_WEEK, count(CASE\_NUMBER) AS Injuries FROM df WHERE INJURY = 'YES'",

"GROUP BY DAY\_OF\_WEEK ORDER BY Injuries DESC")

#Selecting the number of accidents with injuries by day

sqldf(selectStr)

#Step 4: Understand the data using tapply

#-------------------------------------------------------------------------

#How many accidents happen on SUNDAY

#True are accidents on Sunday

tapply(df$CASE\_NUMBER, df$DAY\_OF\_WEEK == "SUNDAY", length)

#How many accidents had injuries

#True had accidents

tapply(df$CASE\_NUMBER, df$INJURY == "YES", length)

#Injuries by day

tapply(df[df$INJURY == "YES",]$INJURY, df[df$INJURY =="YES",]$DAY\_OF\_WEEK, length)

**Console log w/plot**

**Executed code**

> #Homework 5 - 2 Aug 2018

>

> #Step 1: Load the data

> #-------------------------------------------------------------------------

> #Read in the following JSON dataset

> #Library for http get requests

> library(RCurl)

> #Library for json parsing

> library(jsonlite)

> #HTTP get request

> urlObj <- getURL("http://data.maryland.gov/api/views/pdvh-tf2u/rows.json?accessType=DOWNLOAD")

> #Converting the url string to json

> jsonObj <- fromJSON(urlObj)

> #Getting the data from the JSON list

> data <- jsonObj$data

> #Converting the list into the data frame

> df <- data.frame(data, stringsAsFactors = F)

> #Viewing the structure of the data frame, making sure all looks good

> str(df)

'data.frame': 18638 obs. of 26 variables:

$ X1 : chr "1" "2" "3" "4" ...

$ X2 : chr "FC3C1F5D-7CB9-4B08-907C-64857BEDDD56" "D16F0B23-8EBD-4A2A-8333-B090A1F4073E" "6A8ED8B0-D91F-44C9-ACE1-612A59A0B99B" "55B44317-F0C6-4923-AEC1-4DD6B6CB340B" ...

$ X3 : chr "1" "2" "3" "4" ...

$ X4 : chr "1367579703" "1367579703" "1367579703" "1367579703" ...

$ X5 : chr "697617" "697617" "697617" "697617" ...

$ X6 : chr "1367579703" "1367579703" "1367579703" "1367579703" ...

$ X7 : chr "697617" "697617" "697617" "697617" ...

$ X8 : chr "{\n}" "{\n}" "{\n}" "{\n}" ...

$ X9 : chr "1363000002" "1296000023" "1283000016" "1282000006" ...

$ X10: chr "Rockville" "Berlin" "Prince Frederick" "Leonardtown" ...

$ X11: chr "2012-01-01T00:00:00" "2012-01-01T00:00:00" "2012-01-01T00:00:00" "2012-01-01T00:00:00" ...

$ X12: chr "2:01" "18:01" "7:01" "0:01" ...

$ X13: chr "1" "5" "2" "1" ...

$ X14: chr "SUNDAY " "SUNDAY " "SUNDAY " "SUNDAY " ...

$ X15: chr "IS 00495 CAPITAL BELTWAY" "MD 00090 OCEAN CITY EXPWY" "MD 00765 MAIN ST" "MD 00944 MERVELL DEAN RD" ...

$ X16: chr "IS 00270 EISENHOWER MEMORIAL" "CO 00220 ST MARTINS NECK RD" "CO 00208 DUKE ST" "MD 00235 THREE NOTCH RD" ...

$ X17: chr "0" "0.25" "100" "10" ...

$ X18: chr "U" "W" "S" "E" ...

$ X19: chr "Not Applicable" "Not Applicable" "Not Applicable" "Not Applicable" ...

$ X20: chr "15" "23" "4" "18" ...

$ X21: chr "Montgomery" "Worcester" "Calvert" "St. Marys" ...

$ X22: chr "2" "1" "1" "1" ...

$ X23: chr "YES" "YES" "YES" "YES" ...

$ X24: chr "NO" "NO" "NO" "NO" ...

$ X25: chr "VEH" "FIXED OBJ" "FIXED OBJ" "FIXED OBJ" ...

$ X26: chr "OTHER-COLLISION" "OTHER-COLLISION" "FIXED OBJ" "OTHER-COLLISION" ...

>

> #Step 2: Clean the data

> #-------------------------------------------------------------------------

> #deleting the first 8 columns

> df <- df[,-1:-8]

> #names of the columns

> namesOfColumns <- c("CASE\_NUMBER","BARRACK","ACC\_DATE","ACC\_TIME","ACC\_TIME\_CODE",

+ "DAY\_OF\_WEEK","ROAD","INTERSECT\_ROAD","DIST\_FROM\_INTERSECT","DIST\_DIRECTION",

+ "CITY\_NA ME","COUNTY\_CODE","COUNTY\_NAME","VEHICLE\_COUNT","PROP\_DEST","INJURY",

+ "COLLISION\_WITH\_1","COLLISION\_WITH\_2")

> #Assigning the names of teh coulmns to the data frame

> colnames(df) <- namesOfColumns

> #making sure the column headers make sense

> head(df)

CASE\_NUMBER BARRACK ACC\_DATE ACC\_TIME ACC\_TIME\_CODE DAY\_OF\_WEEK ROAD

1 1363000002 Rockville 2012-01-01T00:00:00 2:01 1 SUNDAY IS 00495 CAPITAL BELTWAY

2 1296000023 Berlin 2012-01-01T00:00:00 18:01 5 SUNDAY MD 00090 OCEAN CITY EXPWY

3 1283000016 Prince Frederick 2012-01-01T00:00:00 7:01 2 SUNDAY MD 00765 MAIN ST

4 1282000006 Leonardtown 2012-01-01T00:00:00 0:01 1 SUNDAY MD 00944 MERVELL DEAN RD

5 1267000007 Essex 2012-01-01T00:00:00 1:01 1 SUNDAY IS 00695 BALTO BELTWAY

6 1267000006 Essex 2012-01-01T00:00:00 1:01 1 SUNDAY IS 00083 HARRISBURG EXPWY

INTERSECT\_ROAD DIST\_FROM\_INTERSECT DIST\_DIRECTION CITY\_NA ME COUNTY\_CODE COUNTY\_NAME

1 IS 00270 EISENHOWER MEMORIAL 0 U Not Applicable 15 Montgomery

2 CO 00220 ST MARTINS NECK RD 0.25 W Not Applicable 23 Worcester

3 CO 00208 DUKE ST 100 S Not Applicable 4 Calvert

4 MD 00235 THREE NOTCH RD 10 E Not Applicable 18 St. Marys

5 IS 00083 HARRISBURG EXPWY 100 S Not Applicable 3 Baltimore

6 MD 00137 MT CARMEL RD 0.25 S Not Applicable 3 Baltimore

VEHICLE\_COUNT PROP\_DEST INJURY COLLISION\_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

6 <NA> NO YES FIXED OBJ OTHER-COLLISION

> #Deleting the space at the end of the day of the week

> df$DAY\_OF\_WEEK <- gsub(" ", "", df$DAY\_OF\_WEEK)

> #Converting the distance from intersection into numeric format

> df$DIST\_FROM\_INTERSECT <- as.numeric(df$DIST\_FROM\_INTERSECT)

> #Converting vehicle count into numeric

> df$VEHICLE\_COUNT <- as.numeric(df$VEHICLE\_COUNT)

> #Making day of week a factor

> df$DAY\_OF\_WEEK <- as.factor(df$DAY\_OF\_WEEK)

>

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

> #-------------------------------------------------------------------------

> #Loading library fir sql commands

> library(sqldf)

> #How many accidents happen on SUNDAY

> sqldf("SELECT count(CASE\_NUMBER) AS 'Sunday Crashes' FROM df WHERE df.DAY\_OF\_WEEK = 'SUNDAY'")

Sunday Crashes

1 2373

> #How many accidents had injuries

> sqldf("SELECT count(CASE\_NUMBER) AS 'Accidents with Injuries' FROM df WHERE INJURY = 'YES'")

Accidents with Injuries

1 6433

> # INJURIES by day

> #Creating the select statement with paste

> selectStr = paste("SELECT DAY\_OF\_WEEK, count(CASE\_NUMBER) AS Injuries FROM df WHERE INJURY = 'YES'",

+ "GROUP BY DAY\_OF\_WEEK ORDER BY Injuries DESC")

> #Selecting the number of accidents with injuries by day

> sqldf(selectStr)

DAY\_OF\_WEEK Injuries

1 FRIDAY 1043

2 THURSDAY 968

3 SATURDAY 950

4 MONDAY 915

5 WEDNESDAY 896

6 TUESDAY 843

7 SUNDAY 818

>

> #Step 4: Understand the data using tapply

> #-------------------------------------------------------------------------

> #How many accidents happen on SUNDAY

> #True are accidents on Sunday

> tapply(df$CASE\_NUMBER, df$DAY\_OF\_WEEK == "SUNDAY", length)

FALSE TRUE

16265 2373

> #How many accidents had injuries

> #True had accidents

> tapply(df$CASE\_NUMBER, df$INJURY == "YES", length)

FALSE TRUE

12204 6433

> #Injuries by day

> tapply(df[df$INJURY == "YES",]$INJURY, df[df$INJURY =="YES",]$DAY\_OF\_WEEK, length)

FRIDAY MONDAY SATURDAY SUNDAY THURSDAY TUESDAY WEDNESDAY

1043 915 950 818 968 843 896