STAT 4410/8416 Homework 5

Gerjol, Nicholas

Due on Nov 30, 2021

1. Working with Databases: Please follow the instruction below before answering the questions:

• Install the package sqldf using install.packages('sqldf') • Import the library using library('sqldf') • Read the file https://raw.githubusercontent.com/dsindy/kaggle-titanic/master/data/train.csv and store it in an object called titanic

We can now start writing SQL Script using SQLDF library right inside R. See example below:

```
library(sqldf)
library(data.table)
titanic <- fread("https://raw.githubusercontent.com/dsindy/kaggle-titanic/master/data/train.csv")
sqldf("SELECT passengerid, name, sex
FROM titanic
limit 5", drv="SQLite")
```

```
##   PassengerId                                                Name    Sex
## 1           1                             Braund, Mr. Owen Harris   male
## 2           2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female
## 3           3                              Heikkinen, Miss. Laina female
## 4           4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female
## 5           5                            Allen, Mr. William Henry   male
```

Answer the following questions. Write SQL Script where applicable. a) What does the following command do in MySQL? i) show databases;

```
#Show data bases lists all the databases accessible by MySQL
```

```
    ii) show tables;
```

```
#Shows the data tables inside the databases accessible by MySQL
```

Write SQL script to answer the following questions based on titanic data. Display the results of your script. i. What is the average age of passengers who survived? Group the data by Sex. Display only the column Sex, AverageAge

```
sqldf("select sex, avg(age) as AverageAge from titanic where Survived = 1 group by sex
", drv="SQLite")
```

```
##      Sex AverageAge
## 1 female   28.84772
## 2   male   27.27602
```

ii. What is the percentage of passengers who survived in each Passenger Class or Pclass? Group the data by Sex. Display Pclass, Sex, percentage value.

```
#sqldf("select sex, Pclass, sum(Survived) where Survived = 1 / sum(survived) as percentagevalue from ti
```

iii. What is the average age of all the passenger (survived and not survived)? Group the data by Pclass, Sex, Survived. After that use ggplot to generate a line plot to show average age vs pclass, facet by sex and color it by survived.

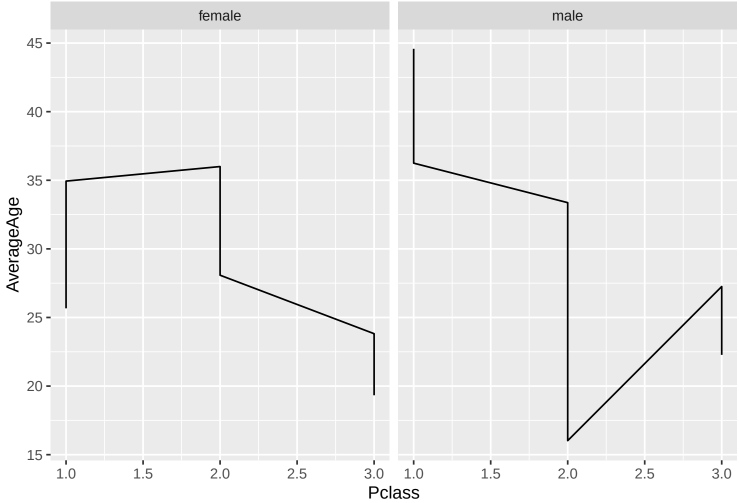
```
library(ggplot2)
sqldf("select avg(age) as AverageAge from titanic
", drv="SQLite")
```

```
##   AverageAge
## 1   29.69912
```

```
plotdat <- sqldf("select Pclass, sex, Survived, avg(age) as AverageAge from titanic group by Pclass, Se
", drv="SQLite")
plotdat
```

```
##    Pclass    Sex Survived AverageAge
## 1       1 female        0   25.66667
## 2       1 female        1   34.93902
## 3       1   male        0   44.58197
## 4       1   male        1   36.24800
## 5       2 female        0   36.00000
## 6       2 female        1   28.08088
## 7       2   male        0   33.36905
## 8       2   male        1   16.02200
## 9       3 female        0   23.81818
## 10      3 female        1   19.32979
## 11      3   male        0   27.25581
## 12      3   male        1   22.27421
```

```
g <- ggplot(plotdat, aes(x=Pclass, y = AverageAge), color = Survived) + geom_line() + facet_wrap(~Sex)
g
```



iv. What is the name, age, sex and pclass of the 5 oldest and 5 youngest persons who died?

```
info <- sqldf("select name, age, sex, Pclass from titanic where survived = 0  order by age desc limit 5
", drv="SQLite")
info
```

```
##                          Name  Age  Sex Pclass
## 1         Svensson, Mr. Johan 74.0 male      3
## 2   Goldschmidt, Mr. George B 71.0 male      1
## 3     Artagaveytia, Mr. Ramon 71.0 male      1
## 4        Connors, Mr. Patrick 70.5 male      3
## 5 Mitchell, Mr. Henry Michael 70.0 male      2
```

```
info <- sqldf("select name, age, sex, Pclass from titanic where survived = 0  order by age asc limit 5
", drv="SQLite")
info
```

```
##                          Name Age  Sex Pclass
## 1            Moran, Mr. James  NA male      3
## 2     Emir, Mr. Farred Chehab  NA male      3
## 3         Todoroff, Mr. Lalio  NA male      3
## 4         Kraeff, Mr. Theodor  NA male      3
## 5 Rogers, Mr. William John  NA male      3
```

v. On average which Passenger Class is more expensive?

```
info <- sqldf("select Pclass, avg(fare) as avgfare from titanic group by Pclass order by avgfare desc li
", drv="SQLite")
info
```

```
##   Pclass  avgfare
## 1      1 84.15469
```

c. Notice the following R codes and explain what it is doing.

```
library(RSQLite)
conn <- dbConnect(RSQLite::SQLite(), "titanicDB")
dbWriteTable(conn, name = "titanic", value = titanic, overwrite=TRUE)
dbListTables(conn)
```

```
## [1] "titanic"
```

```
#The code is connecting to the titanic data base then renaming the titanicdb as "titanic"
```

d. Use package dplyr to obtain the same result as you did in 1b(iii) above. For this use the connection string conn and the function tbl(). Store the result in an object called meanAge.

```
library(dplyr)
#conDplyr = src_mysql(dbname = "trainingDB", user = "training", password = "training123", host = "local
# meanAge <- conDplyr %>%
#   tbl("titanic") %>%
#   select(avg(age)) %>%
#   collect()
#I commented this because I'm still unable to get mysql on my machine so localhost connects to nothing
```

e. Show the SQL query to create meanAge in 1(d) using the fiunction show_query()

2. Extracting twitter data: In this problem we would like to extract data from twitter. For this refer to the documentation in the e following link. https://github.com/geoffjentry/twitteR/

a. Twitter API: Set up twitter API using any of the following methods. Make sure you installed all the packages as mentioned in the class. Method 2: If you don't like creating an account with twitter and going through all the trouble, you can use my keys (ssh, don't tell anyone). For this download the hw5-twitter-auth file from blackboard and load it as follows.

```
load("hw5-twitter-auth")
library(twitteR)
setup_twitter_oauth(api_key,api_secret,access_token, access_token_secret)
```

```
## [1] "Using direct authentication"
```

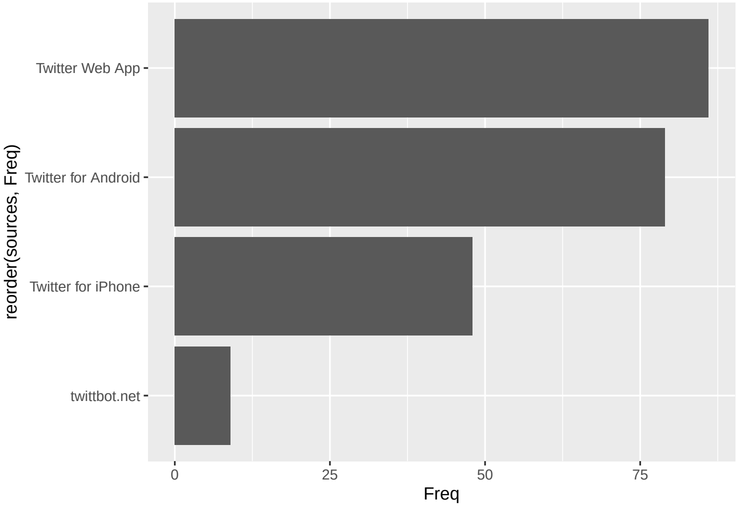
b. Now search twitter messages for “data science job”. Display few job informations.

```
dsTweet <- searchTwitter("data science job")
dfTweet <- twListToDF(dsTweet)
tweetText <- dfTweet$text
head(tweetText)
```

```
## [1] "RT @emilieschario: @tayloramurphy @seanjtaylor time is a flat circle. Come to my talk on why "da
## [2] "@tayloramurphy @seanjtaylor time is a flat circle. Come to my talk on why "data scientist" is a
## [3] "Analyst II, Data Science - Somerville, Massachusetts - Liberty Mutual Insurance https://t.co/hM
## [4] "Many years ago I applied for a data engineering role and accidentally got routed to a data scien
## [5] "RT @BDataScientist: Our AI for Good Research Lab is searching for an experienced Data and Applic
## [6] "RT @BDataScientist: Our AI for Good Research Lab is searching for an experienced Data and Applic
```

c. Search 300 tweets using the hash tag #chess and save them in an object called rTweets. Show the top 7 sources of tweets (such as android or iphone) in a ordered bar plot.

```
library(ggplot2)
rTweets <- searchTwitter("#chess", n=300)
sources <- sapply(rTweets, function(x) x$getStatusSource())
sources <- gsub("</a>", "", sources)
sources <- strsplit(sources, ">")
sources <- sapply(sources, function(x) ifelse(length(x) > 1, x[2], x[1]))
source_table = table(sources)
df <- data.frame(names(source_table),source_table)
ggplot(df[df$Freq>8,], aes(reorder(sources,Freq),Freq)) +
  geom_bar(stat="identity") + coord_flip()
```



d. Notice that the object rTweets is a list. Convert it into a data frame using function twListToDF and store it in an object called dTweets. Display some data from dTweets.

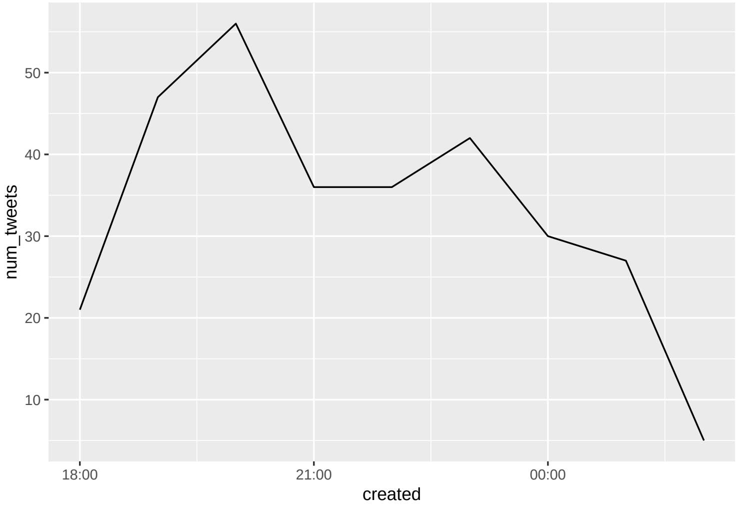
```
dTweets <- twListToDF(rTweets)
head(dTweets)
```

```
##
## 1                                       Other than experience, is there a way to stop falling for thi
## 2 RT @echaguen: España octava potencia mundial de #ajedrez según el ranking de la  Federación Interna
## 3                A great insight into an underlining and probably oldest gaming cultures. \n#wargame
## 4                                 White to play and win: Gutkovich, Polina vs Johnson, Frank B. Titl
## 5                   top down #cryptocurrency in last 15 minute :\n\n#ERN  13.235 | 12.949 USDT (-2.1
## 6                                            @bigdybbukenergy @manunderbridge Game is garbage try 9x
##   favorited favoriteCount       replyToSN             created truncated
## 1     FALSE             1            <NA> 2021-12-01 01:08:58     FALSE
## 2     FALSE             0            <NA> 2021-12-01 01:06:30     FALSE
## 3     FALSE             1            <NA> 2021-12-01 01:03:19      TRUE
## 4     FALSE             0            <NA> 2021-12-01 01:00:47      TRUE
## 5     FALSE             0            <NA> 2021-12-01 01:00:02      TRUE
## 6     FALSE             0 bigdybbukenergy 2021-12-01 00:57:15     FALSE
##            replyToSID                  id replyToUID
## 1                <NA> 1465850375068536841       <NA>
## 2                <NA> 1465849754584244227       <NA>
## 3                <NA> 1465848953434349571       <NA>
## 4                <NA> 1465848316328042498       <NA>
## 5                <NA> 1465848125889953797       <NA>
## 6 1464689070534234119 1465847426879737866 2183239617
##
## 1                            <a href="http://twitter.com/download/android" rel="nofollow">Twitter fo
## 2                                         <a href="https://mobile.twitter.com" rel="nofollow">Twitte
## 3                                                  <a href="http://www.linkedin.com/" rel="nofollow"
## 4                                            <a href="https://12chess.com" rel="nofollow">1 2 Chess!
## 5 <a href="https://help.twitter.com/en/using-twitter/how-to-tweet#source-labels" rel="nofollow">moon
## 6                                         <a href="https://mobile.twitter.com" rel="nofollow">Twitte
##        screenName retweetCount isRetweet retweeted longitude latitude
## 1      OttoSilver            0     FALSE     FALSE      <NA>     <NA>
## 2      ugogarciap           27      TRUE     FALSE      <NA>     <NA>
## 3 JeffersonSolves            0     FALSE     FALSE      <NA>     <NA>
## 4     onetwochess            0     FALSE     FALSE      <NA>     <NA>
## 5   moon_or_earth            0     FALSE     FALSE      <NA>     <NA>
## 6    ISolvedChess            0     FALSE     FALSE      <NA>     <NA>
```

e. dTweets has a column showing the time the tweet was created. Generate a plot showing number of tweets on each of the hours. Add a smooth line overlaid on your plot.

```
library(lubridate)
library(data.table)
dTweets$created <- ceiling_date(dTweets$created, unit="hour")
dTweets$num_tweets <- 1
df = subset(dTweets, select = c(created, num_tweets))
test <- aggregate(. ~ created, df, sum)

g <- ggplot(test, aes(x=created, y=num_tweets)) + geom_line()
g
```



f. Arrange the dataframe dTweets based on the retweetCount. While doing this select only columns text, screenName, retweetCount. Store the data in a object called mostTweets. Display five texts that are most retweeted.

```
head(dTweets)
```

```
##
## 1                                       Other than experience, is there a way to stop falling for thi
## 2 RT @echaguen: España octava potencia mundial de #ajedrez según el ranking de la  Federación Interna
## 3                A great insight into an underlining and probably oldest gaming cultures. \n#wargames
## 4                                 White to play and win: Gutkovich, Polina vs Johnson, Frank B. Titl
## 5                   top down #cryptocurrency in last 15 minute :\n\n#ERN  13.235 | 12.949 USDT (-2.16
## 6                                            @bigdybbukenergy @manunderbridge Game is garbage try 9x9
##   favorited favoriteCount       replyToSN             created truncated
## 1     FALSE             1            <NA> 2021-12-01 02:00:00     FALSE
## 2     FALSE             0            <NA> 2021-12-01 02:00:00     FALSE
## 3     FALSE             1            <NA> 2021-12-01 02:00:00      TRUE
## 4     FALSE             0            <NA> 2021-12-01 02:00:00      TRUE
## 5     FALSE             0            <NA> 2021-12-01 02:00:00      TRUE
## 6     FALSE             0 bigdybbukenergy 2021-12-01 01:00:00     FALSE
##            replyToSID                  id replyToUID
## 1                <NA> 1465850375068536841       <NA>
## 2                <NA> 1465849754584244227       <NA>
## 3                <NA> 1465848953434349571       <NA>
## 4                <NA> 1465848316328042498       <NA>
## 5                <NA> 1465848125889953797       <NA>
## 6 1464689070534234119 1465847426879737866 2183239617
##
## 1                            <a href="http://twitter.com/download/android" rel="nofollow">Twitter fo
## 2                                         <a href="https://mobile.twitter.com" rel="nofollow">Twitte
## 3                                                  <a href="http://www.linkedin.com/" rel="nofollow">
## 4                                            <a href="https://12chess.com" rel="nofollow">1 2 Chess!
## 5 <a href="https://help.twitter.com/en/using-twitter/how-to-tweet#source-labels" rel="nofollow">moon
## 6                                         <a href="https://mobile.twitter.com" rel="nofollow">Twitte
##        screenName retweetCount isRetweet retweeted longitude latitude
## 1      OttoSilver            0     FALSE     FALSE      <NA>     <NA>
## 2      ugogarciap           27      TRUE     FALSE      <NA>     <NA>
## 3 JeffersonSolves            0     FALSE     FALSE      <NA>     <NA>
## 4     onetwochess            0     FALSE     FALSE      <NA>     <NA>
## 5   moon_or_earth            0     FALSE     FALSE      <NA>     <NA>
## 6    ISolvedChess            0     FALSE     FALSE      <NA>     <NA>
##   num_tweets
## 1          1
## 2          1
## 3          1
## 4          1
## 5          1
## 6          1
```

```
dTweets <- as.data.table(dTweets)
testsort <- dTweets[order(-retweetCount),]
head(testsort)
```

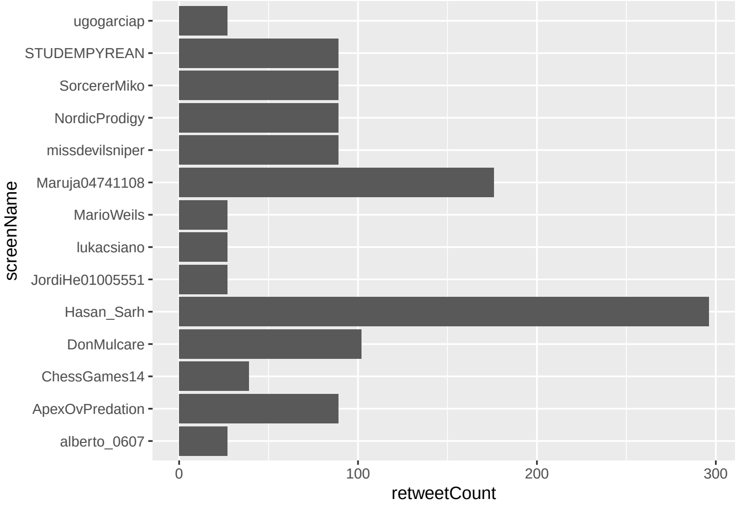
```
##
## 1:                    RT @borsaressami: #chess geçen gün bir hayli x yaptirmisti bize ama bugün girme
## 2:         RT @neo4j: Miss the 3rd episode of Discover #Neo4j Aura Free /w @alexandererdl &amp; @ElLa
## 3:                   RT @StreetArtDream: ... together can be better than alone. Be the first change.
## 4: RT @MedicalLynx: <U+275D>Mon3tr and I will protect everyone.<U+275E>\n\nArknights/Multiverse. \n\n
## 5: RT @MedicalLynx: <U+275D>Mon3tr and I will protect everyone.<U+275E>\n\nArknights/Multiverse. \n\n
## 6: RT @MedicalLynx: <U+275D>Mon3tr and I will protect everyone.<U+275E>\n\nArknights/Multiverse. \n\n
##    favorited favoriteCount replyToSN             created truncated replyToSID
## 1:     FALSE             0      <NA> 2021-11-30 19:00:00     FALSE       <NA>
## 2:     FALSE             0      <NA> 2021-12-01 00:00:00     FALSE       <NA>
## 3:     FALSE             0      <NA> 2021-11-30 19:00:00     FALSE       <NA>
## 4:     FALSE             0      <NA> 2021-12-01 01:00:00     FALSE       <NA>
## 5:     FALSE             0      <NA> 2021-12-01 00:00:00     FALSE       <NA>
## 6:     FALSE             0      <NA> 2021-11-30 23:00:00     FALSE       <NA>
##                     id replyToUID
## 1: 1465750985234685960       <NA>
## 2: 1465818470222807042       <NA>
## 3: 1465750154003234818       <NA>
## 4: 1465838912236965888       <NA>
## 5: 1465823630307250176       <NA>
## 6: 1465816805130129411       <NA>
##                                                                             statusSource
## 1:   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
## 2:              <a href="https://mobile.twitter.com" rel="nofollow">Twitter Web App</a>
## 3: <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 4: <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 5: <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 6: <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
##         screenName retweetCount isRetweet retweeted longitude latitude
## 1:      Hasan_Sarh          209      TRUE     FALSE      <NA>     <NA>
## 2:  Maruja04741108          176      TRUE     FALSE      <NA>     <NA>
## 3:      DonMulcare          102      TRUE     FALSE      <NA>     <NA>
## 4: missdevilsniper           89      TRUE     FALSE      <NA>     <NA>
## 5:    STUDEMPYREAN           89      TRUE     FALSE      <NA>     <NA>
## 6:    SorcererMiko           89      TRUE     FALSE      <NA>     <NA>
##    num_tweets
## 1:          1
## 2:          1
## 3:          1
## 4:          1
## 5:          1
## 6:          1
```

```
mostTweets <- subset(testsort, select = c(text, screenName, retweetCount))
head(mostTweets, 5)
```

```
##
## 1:                    RT @borsaressami: #chess geçen gün bir hayli x yaptirmisti bize ama bugün girm
## 2:         RT @neo4j: Miss the 3rd episode of Discover #Neo4j Aura Free /w @alexandererdl &amp; @ElLa
## 3:                   RT @StreetArtDream: ... together can be better than alone. Be the first change.
## 4: RT @MedicalLynx: <U+275D>Mon3tr and I will protect everyone.<U+275E>\n\nArknights/Multiverse. \n\
## 5: RT @MedicalLynx: <U+275D>Mon3tr and I will protect everyone.<U+275E>\n\nArknights/Multiverse. \n\
##          screenName retweetCount
## 1:      Hasan_Sarh          209
## 2:  Maruja04741108          176
## 3:      DonMulcare          102
## 4: missdevilsniper           89
## 5:    STUDEMPYREAN           89
```

g. Generate a bar chart showing top 15 screen names and count of retweets from mostTweets. Order the bars based on the retweet counts.

```
topTweets <- head(mostTweets, 15)
bartweets <- subset(topTweets, select = c( screenName, retweetCount))
b <- ggplot(bartweets, aes(x=screenName, y=retweetCount)) + geom_bar(stat = "identity") + coord_flip()
b
```


ugogarciap
STUDEMPYREAN
SorcererMiko
NordicProdigy
missdevilsniper
Maruja04741108
MarioWeils
lukacsiano
JordiHe01005551
Hasan_Sarh
DonMulcare
ChessGames14
ApexOvPredation
alberto_0607
screenName
retweetCount