Assignment 1 IS 607

Talha Muhammad

September 9, 2016

Step 1. Initial Setup

Load the different packages:

```
library(bitops)
library(stringr)
library(XML)
library(RCurl)
library(RMySQL)
library(reshape2)
library(ggplot2)
```

```
inputdir="C:/Users/talha/Documents/Training/CUNY Classes/IS607/Week2"
outputdir="C:/Users/talha/Documents/Training/CUNY Classes/IS607/Week2"
```

Step 2. Load URLs and Scrapped Data from the Web

we load the data on movies from (the web) using BoxOffice Mojo

```
#load URLs
box<-getURL("http://www.boxofficemojo.com/quarterly/?view=releasedate&yr=2016&quarter=Q3")
# Parse the URLs
box_parsed<-htmlParse(box,encoding="UTF-8")
tables<-readHTMLTable(box_parsed,stringsAsFactors=FALSE)
```

Step 3. Clean the data

The scrapped data require some cleaning and reformainting

```
#Select and clean the different tables
movie_table<-tables[[4]]
str(movie_table)
```

```
## 'data.frame':    105 obs. of  9 variables:
##  $ V1: chr  "Filter26 Aries5&2A24Abr.AEFAlcArgo.BGBSMBSTBVCineGalaxyCJCLCohenCol.Collective ECZDistri
##  $ V2: chr  NA "The Secret Life of Pets" "Suicide Squad" "Jason Bourne" ...
##  $ V3: chr  NA "Uni." "WB" "Uni." ...
##  $ V4: chr  NA "$361,837,775" "$307,407,853" "$158,771,290" ...
##  $ V5: chr  NA "4,381" "4,255" "4,039" ...
##  $ V6: chr  NA "$104,352,905" "$133,682,248" "$59,215,365" ...
##  $ V7: chr  NA "4,370" "4,255" "4,026" ...
##  $ V8: chr  NA "7/8" "8/5" "7/29" ...
##  $ V9: chr  NA "-" "-" "-" ...
```

```
# clean and name the different columns
movie_table<-movie_table[2:101,]
colnames(movie_table)<-c("Rank","Title","Studio","Gross_Q3","Total_Theaters","Opening","Open_theatre","

# all the data are strings and need to be converted to numeric
movie_table$Gross_Q3<-str_replace_all(movie_table$Gross_Q3,"[$,]","")
movie_table$Opening<-str_replace_all(movie_table$Opening,"[$,]","")
movie_table$Total_Theaters<-str_replace_all(movie_table$Total_Theaters,"[,]","")
movie_table$Open_theatre<-str_replace_all(movie_table$Open_theatre,"[,]","")
# convert to numeric
movie_table$Gross_Q3<-as.numeric(movie_table$Gross_Q3)
movie_table$Total_Theaters<-as.numeric(movie_table$Total_Theaters)
movie_table$Opening<-as.numeric(movie_table$Opening)
```

```
## Warning: NAs introduced by coercion
```

```
movie_table$Rank<-as.numeric(movie_table$Rank)
movie_table$Gross_Q3<-movie_table$Gross_Q3/1000000
movie_table$Opening<-movie_table$Opening/1000000
movie_table[1:10,]
```

```
##    Rank                    Title Studio  Gross_Q3 Total_Theaters   Opening
## 2     1  The Secret Life of Pets   Uni. 361.83778           4381 104.35291
## 3     2             Suicide Squad     WB 307.40785           4255 133.68225
## 4     3              Jason Bourne   Uni. 158.77129           4039  59.21536
## 5     4         Star Trek Beyond   Par. 156.58063           3928  59.25321
## 6     5      Ghostbusters (2016)   Sony 126.70289           3963  46.01875
## 7     6     The Legend of Tarzan     WB 126.40863           3591  38.52786
## 8     7                  Bad Moms    STX 107.52627           3215  23.81734
## 9     8             Sausage Party   Sony  93.18258           3135  34.26353
## 10    9 The Purge: Election Year   Uni.  79.04244           2821  31.51511
## 11   10     Pete's Dragon (2016)     BV  70.01665           3702  21.51410
##    Open_theatre Open_date close_date
## 2          4370       7/8          -
## 3          4255       8/5          -
## 4          4026      7/29          -
## 5          3928      7/22          -
## 6          3963      7/15          -
## 7          3561       7/1          -
## 8          3215      7/29          -
## 9          3103      8/12          -
## 10         2796       7/1       8/18
## 11         3702      8/12          -
```

Step 4. Export the top ranked movies

The top ranked movies are then exported as a table to be used to develop the survey and input into the database

```
movie_names<-subset(movie_table,Rank<=10,Rank:Title)
write.table(movie_names,file.path(outputdir,"movie_names.csv"),row.names=FALSE,col.names=FALSE,quote=TRU
```

Step 5. Develop a Survey Instrument

Survey can be accessed at the link below https://www.surveymonkey.com/r/WJQ6PBG

The survey is conducted for six respondents and export the data. The data is exported and a SQL database is created. Please see MySQL code.

Step 6. Run SQL Query in R and Import the data

Using the SQL database we run the query in R.

```
rmysql.settingsfile<-"C:/Program Files/MySQL/MySQL Server 5.0/my.ini"
con <- dbConnect(RMySQL::MySQL(), dbname = "moviesurvey", username="root", password="password")
surveydata<-dbGetQuery(con,"SELECT * from survey")
dbDisconnect(con)
```

```
## [1] TRUE
```

Step 7. Merge and Analyze the data

Develop some initial exploratory plots

```
survey_reshape<-dcast(surveydata[,2:4],movie_id~survey_id,value.var="score")
survey_reshape$avgrank<-apply(survey_reshape[,2:7],1,mean)
survey_combined<-merge(movie_table,survey_reshape,by.x="Rank",by.y="movie_id")
# Develop some plots
ggplot(survey_combined,aes(avgrank,Gross_Q3,color=Studio, label=Title))+geom_point(size=3)+xlab("Averag
```