DATA607 WEEK TWO ASSIGNMENT

John K. Hancock

September 7, 2018

Week Two Assignment DATA607 Importing Databases into R Fall 2018

John K. Hancock

This project entails connecting R to an instance of MYSQL, displaying tables/fields, creating dataframes, using a SQL query to create a combination of dataframes, and plotting insight gained from the data.

Import the necessary Libraries

```
options(warn=-1)
library(DBI)
library(RMySQL)
library(stringr)
library(ggplot2)
```

Establish a Connection to the MySql Instance through the user, Student01

```
conn<- dbConnect(RMySQL::MySQL(),
                 dbname="movie_reviews",
                 host="DESKTOP-FGDMVA4",
                 user="Student01",
                 password="CUNY_DATA607"
                 )
```

Printout list of tables in the connection

```
dbListTables(conn)
```

```
## [1] "movies"    "reviewers" "reviews"
```

Read in one of the tables into a dataframe

```
Reviewers <- dbReadTable(conn, 'Reviewers')
```

Note: The import appended a “” after each entry in the Gender column. So, the column needs a bit of cleaning. After research, this seems to be a bug in the code. R interprets “\n” as “\r”.

```
Reviewers
```

```
##   ReviewerID         USERNAME       City State   Zip        Occupation
## 1          1         Joe Chan   Brooklyn    NY 11217   Project Manager
## 2          2         Joe Liao     Newark    NJ 17101      Data Analyst
## 3          3      Mario Nikac     Camden    NJ 17175          Salesman
## 4          4    Gloria Stivic    Queens     NY 11368        Home Maker
## 5          5     Carrie Coons Haverstraw    NY 10927           Actress
## 6          6 Ismael Rodriguez   New York    NY 10014 Assistant Manager
## 7          7      Zazie Beets   New York    NY 10027             Model
## 8          8    Cynthia Nixon   New York    NY 10014         Candidate
##    Income   Gender
## 1   75000   Male\r
## 2  125000   Male\r
## 3  125000   Male\r
## 4   50000 Female\r
## 5   75000 Female\r
## 6   75000   Male\r
## 7  500000 Female\r
## 8 1200000 Female\r
```

```
Reviewers$Gender<- str_replace_all(Reviewers$Gender, "\r", "")
```

```
Reviewers
```

```
##   ReviewerID         USERNAME       City State   Zip        Occupation
## 1          1         Joe Chan   Brooklyn    NY 11217   Project Manager
## 2          2         Joe Liao     Newark    NJ 17101      Data Analyst
## 3          3      Mario Nikac     Camden    NJ 17175          Salesman
## 4          4    Gloria Stivic    Queens     NY 11368        Home Maker
## 5          5     Carrie Coons Haverstraw    NY 10927           Actress
## 6          6 Ismael Rodriguez   New York    NY 10014 Assistant Manager
## 7          7      Zazie Beets   New York    NY 10027             Model
## 8          8    Cynthia Nixon   New York    NY 10014         Candidate
##    Income Gender
## 1   75000   Male
## 2  125000   Male
## 3  125000   Male
## 4   50000 Female
## 5   75000 Female
## 6   75000   Male
## 7  500000 Female
## 8 1200000 Female
```

```
Reviews <- dbReadTable(conn, 'Reviews')
Movies <- dbReadTable(conn, 'movies')

Movies$Trailer <- str_replace_all(Movies$Trailer, "\r", "")
Reviews$Comments <- str_replace_all(Reviews$Comments, "\r", "")
```

```
Reviews
```

```
##    ReviewID MovieID Rating                                  Comments
## 1         1       1      4                 I Loved It. Best Movie Ever.
## 2         2       1      5               Great movie. Will See it Again.
## 3         3       1      5                       Soo many superheroes!!!
## 4         4       1      5                                It was great.
## 5         5       1      4                                      Amazing
## 6         6       1      2                                   Hated it
## 7         7       1      1                             Worst Movie Ever
## 8         8       1      3                                 It was okay.
## 9         1       2      2                                       So so.
## 10        2       2      3                                   Allright
## 11        3       2      2                               Disappointing
## 12        4       2      4                                    Loved it
## 13        5       2      4                                  Funny movie
## 14        6       2      4                               Really liked it
## 15        7       2      5                                     Grea film
## 16        8       2      3                                Average movie
## 17        1       3      4                      Fun times at the movie
## 18        2       3      5                      Me and my kids loved it
## 19        3       3      3                                 It was okay.
## 20        4       3      4                                  A good time
## 21        5       3      5                     Best movie of the summer
## 22        6       3      5                                    Loved it
## 23        7       3      4                           My kids loved it.
## 24        8       3      5                 Whole family had a great time
## 25        1       4      1                                  Fell asleep
## 26        2       4      1                                       Boring
## 27        3       4      2                                         Dull
## 28        4       4      1                                         Awful
## 29        5       4      5                                    Inspiring
## 30        6       4      5                                I admire her
## 31        7       4      5                           A remarkable woman
## 32        8       4      5                                      My hero
## 33        1       5      1                             Hated this movie
## 34        2       5      4                             Tom Cruise rocks
## 35        3       5      5                           Stunts were awesome
## 36        4       5      5                                    Fantastic
## 37        5       5      1                                Terrible movie
## 38        6       5      1                                       Horrible
## 39        7       5      1                           Want my money back
## 40        8       5      1                         Do not waste your time
## 41        1       6      1                                           Weid
## 42        2       6      1                          Did not understand it
## 43        3       6      1                                   Not my taste
## 44        4       6      1                                   Over my head
## 45        5       6      1          I am not the audience for this film
## 46        6       6      2                      Wait until its on Netflix
## 47        1       7      1                             Dumbest movie ever
## 48        2       7      1 Can't belive I wasted my money on that film.
## 49        3       7      1                    Worst Star Wars movie ever
## 50        4       7      1                                       Horrible
## 51        5       7      1                                         Boring
## 52        6       7      2                                  Dont see it
```

```
## 53        7       7      1                            Totally forgettable
## 54        8       7      1                                    Truly Awful
```

```
Movies
```

```
##   ID                       Title             Genre Box_Office
## 1  1     Avengers Infinity Wars Super Hero Action  2.046e+09
## 2  2          Crazy Rich Asians   Romantic Comedy  3.000e+07
## 3  3              Incredibles 2  Animation Family  1.167e+09
## 4  4                        RBG       Documentary  1.390e+07
## 5  5 Mission Impossible Fallout            Action  1.780e+08
## 6  6        Sorry to Bother You            Comedy  1.660e+07
## 7  7     Solo A Star Wars Story      SciFi Action  2.000e+08
##           Male_Lead         Female_Lead   Length
## 1  Robert Downey Jr  Scarlett johansson 02:29:00
## 2     Henry Golding        Constance Wu 02:10:00
## 3   Craig T. Nelson        Holly Hunter 02:05:00
## 4                NA Ruth Bader Ginsburg 01:37:00
## 5        Tom Cruise   Rebecca Ferguson  02:28:00
## 6 Lakeith Stanfield      Tessa Thompson 01:51:00
## 7  Alden Ehrenreich        Emila Clarke 02:15:00
##                                    Trailer
## 1 https://www.youtube.com/watch?v=Xe5MeKNFjGQ
## 2 https://www.youtube.com/watch?v=ZQ-YX-5bAs0
## 3 https://www.youtube.com/watch?v=i5qOzqD9Rms
## 4 https://www.youtube.com/watch?v=biIRlcQqmOc
## 5 https://www.youtube.com/watch?v=wb49-oV0F78
## 6 https://www.youtube.com/watch?v=enH3xA4mYcY
## 7 https://www.youtube.com/watch?v=jPEYpryMp2s
```

```
dbListFields(conn, 'movies')
```

```
## [1] "ID"          "Title"       "Genre"       "Box_Office"  "Male_Lead"
## [6] "Female_Lead" "Length"      "Trailer"
```

```
dbListFields(conn, 'reviews')
```

```
## [1] "ReviewID" "MovieID"  "Rating"   "Comments"
```

```
dbListFields(conn, 'reviewers')
```

```
## [1] "ReviewerID" "USERNAME"   "City"       "State"      "Zip"
## [6] "Occupation" "Income"     "Gender"
```

Create a dataframe that shows average rating per movie genre by gender

```
genre_ratings_by_gender<- dbGetQuery(conn, "SELECT movies.Genre, reviewers.Gender, AVG(reviews.R
ating) AS 'Average_Rating'
                                            FROM movies
                                            INNER JOIN reviews ON reviews.MovieID = movies.ID
                                            INNER JOIN reviewers ON reviewers.ReviewerID = revie
ws.ReviewID
                                            GROUP BY movies.genre, reviewers.gender")
```

Clean up the “”

```
genre_ratings_by_gender$Gender <- str_replace_all(genre_ratings_by_gender$Gender, "\r", "")
```

```
genre_ratings_by_gender
```

```
##                Genre Gender Average_Rating
## 1  Super Hero Action   Male           4.00
## 2  Super Hero Action Female           3.25
## 3    Romantic Comedy   Male           2.75
## 4    Romantic Comedy Female           4.00
## 5   Animation Family   Male           4.25
## 6   Animation Family Female           4.50
## 7         Documentary   Male           2.25
## 8         Documentary Female           4.00
## 9              Action   Male           2.75
## 10             Action Female           2.00
## 11             Comedy   Male           1.25
## 12             Comedy Female           1.00
## 13      SciFi Action   Male           1.25
## 14      SciFi Action Female           1.00
```

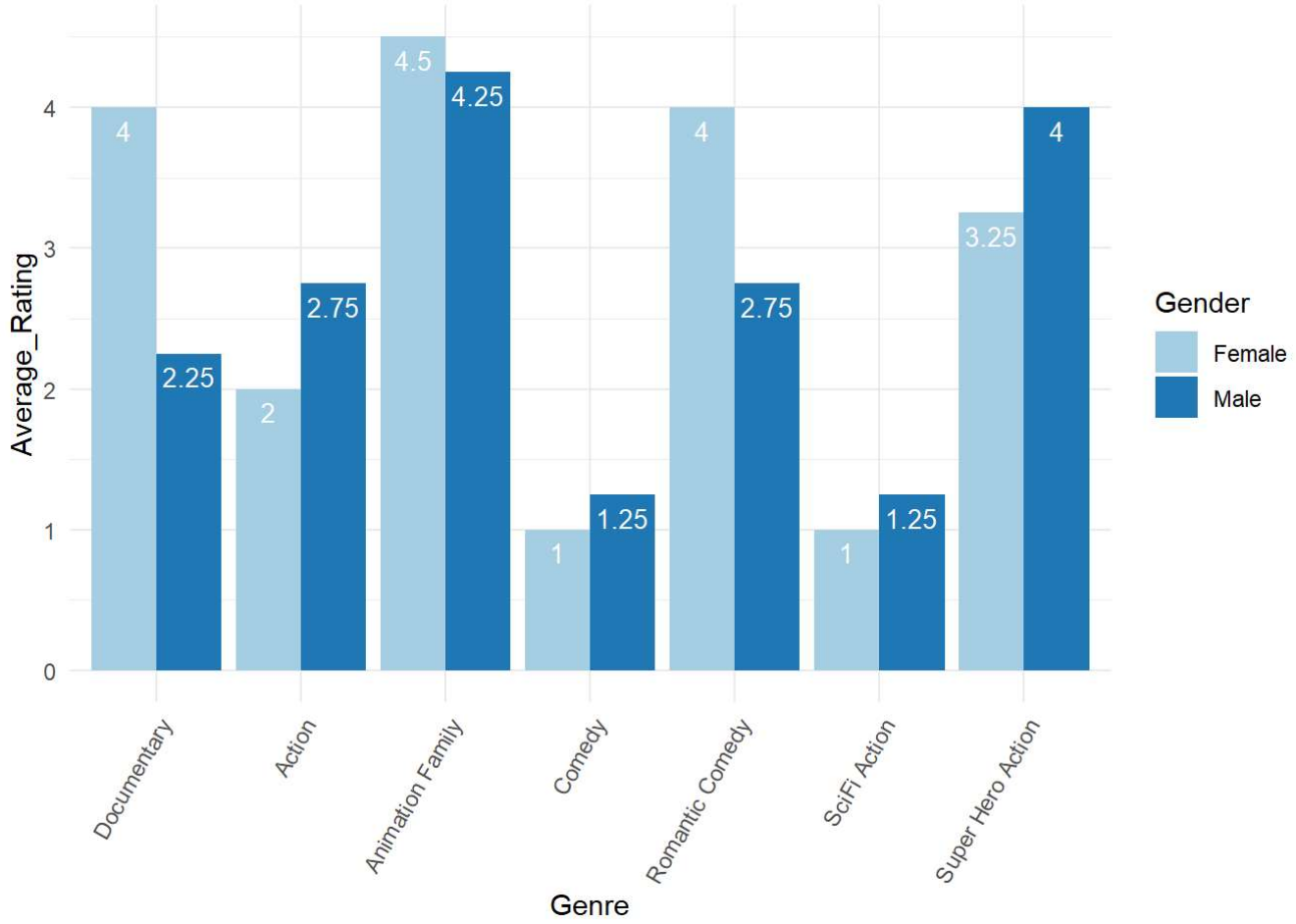
Change the Gender variable to a factor

```
genre_ratings_by_gender$Gender <- as.factor(genre_ratings_by_gender$Gender)
genre_ratings_by_gender
```

```
##                Genre Gender Average_Rating
## 1  Super Hero Action   Male           4.00
## 2  Super Hero Action Female           3.25
## 3    Romantic Comedy   Male           2.75
## 4    Romantic Comedy Female           4.00
## 5   Animation Family   Male           4.25
## 6   Animation Family Female           4.50
## 7         Documentary   Male           2.25
## 8         Documentary Female           4.00
## 9              Action   Male           2.75
## 10             Action Female           2.00
## 11             Comedy   Male           1.25
## 12             Comedy Female           1.00
## 13      SciFi Action   Male           1.25
## 14      SciFi Action Female           1.00
```

Below, we can gain insight from the data by plotting the average rating by gender for each current genre of movies. Females in this sample prefer Animation Family, Documentaries and Romantic Comedies whereas Males also prefer Animation Family and Super Hero Action movies.

```
ggplot(data=genre_ratings_by_gender, aes(x=Genre, y=Average_Rating, fill=Gender)) +
  geom_bar(stat="identity", position=position_dodge())+
  geom_text(aes(label=Average_Rating), vjust=1.6, color="white",
            position = position_dodge(0.9), size=3.5)+
  scale_fill_brewer(palette="Paired")+
  theme_minimal() + theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



Finally, the connection is closed.

```
dbDisconnect(conn)
```

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