Assignment 2: Sql and R

Carol Campbell

2023-09-17

I chose six movies now playing, and asked ten "friends" to rate each of the movies they had seen on a scale of 1 to 5. Not everyone saw every movie.

Using MySQL workbench, I created the "movies" database which has three individual tables to store data:

Table 1: movie Table 2: reviewer Table 3: rating

Install needed packages and/or libraries.

I created a new MySQL database connection, "Data607", which I will be using throughout this course. The password is "Fall2023!" (Reference - Youtube video, "MySQL Workbench Add User and Connect to Database", https://www.youtube.com/watch?v=P7whjxMqYU4)

```
# Connect MySQL to R to upload my "movies" database. Use "Fall2023!" for password.

mydb = dbConnect(RMySQL::MySQL(),
       dbname='movies',
       host='127.0.0.1',
       port=3306,
       user='Data607',
       password=rstudioapi::askForPassword("Enter password"))
```

What tables are in the "movies" database?

```
# See the database tables.

dbListTables(mydb)
```

```
## [1] "movie"    "reviewer" "score"
```

Show each table:

```
#movie table
movie_tbl <- dbSendQuery(mydb, "SELECT * FROM movie")
dbFetch(movie_tbl)
```

```
##   movie_id               title            genre
## 1        1     The Equalizer 3 Action/Adventure
## 2        2       Grand Turismo Action/Adventure
## 3        3 The Retirement Plan           Comedy
## 4        4          The Nun II           Horror
## 5        5          Dumb Money     Comedy/Drama
## 6        6         Expend4bles  Action/Thriller
```

```r
#reviewer table
reviewer_tbl <- dbSendQuery(mydb, "SELECT * FROM reviewer")
dbFetch(reviewer_tbl)
```

```
##    reviewer_id first_name movie_id
## 1          201      Craig        1
## 2          201      Craig        2
## 3          201      Craig        3
## 4          201      Craig        4
## 5          201      Craig        5
## 6          201      Craig        6
## 7          202     Andrea        1
## 8          202     Andrea        2
## 9          202     Andrea        3
## 10         202     Andrea        4
## 11         202     Andrea        5
## 12         202     Andrea        6
## 13         203     Darryl        1
## 14         203     Darryl        2
## 15         203     Darryl        3
## 16         203     Darryl        4
## 17         203     Darryl        5
## 18         203     Darryl        6
## 19         204    Beverly        1
## 20         204    Beverly        2
## 21         204    Beverly        3
## 22         204    Beverly        4
## 23         204    Beverly        5
## 24         204    Beverly        6
## 25         205     Maysie        1
## 26         205     Maysie        2
## 27         205     Maysie        3
## 28         205     Maysie        4
## 29         205     Maysie        5
## 30         205     Maysie        6
## 31         206      Karen        1
## 32         206      Karen        2
## 33         206      Karen        3
## 34         206      Karen        4
## 35         206      Karen        5
## 36         206      Karen        6
## 37         207    Jassiem        1
## 38         207    Jassiem        2
## 39         207    Jassiem        3
## 40         207    Jassiem        4
## 41         207    Jassiem        5
## 42         207    Jassiem        6
## 43         208       Marc        1
## 44         208       Marc        2
## 45         208       Marc        3
## 46         208       Marc        4
## 47         208       Marc        5
## 48         208       Marc        6
```

```
## 49         209      Trudy        1
## 50         209      Trudy        2
## 51         209      Trudy        3
## 52         209      Trudy        4
## 53         209      Trudy        5
## 54         209      Trudy        6
## 55         210        Kim        1
## 56         210        Kim        2
## 57         210        Kim        3
## 58         210        Kim        4
## 59         210        Kim        5
## 60         210        Kim        6
## 61         201      Craig        6
## 62         202     Andrea        6
## 63         203     Darryl        6
## 64         204    Beverly        6
## 65         205     Maysie        6
## 66         206      Karen        6
## 67         207    Jassiem        6
## 68         208       Marc        6
## 69         209      Trudy        6
## 70         210        Kim        6
```

```
#score table
score_tbl <- dbSendQuery(mydb, "SELECT * FROM score")
dbFetch(score_tbl)
```

```
##    movie_id reviewer rating
## 1         1    Craig      5
## 2         2    Craig      4
## 3         3    Craig      4
## 4         4    Craig     NA
## 5         5    Craig     NA
## 6         6    Craig      5
## 7         1   Andrea     NA
## 8         2   Andrea      4
## 9         3   Andrea      3
## 10        4   Andrea     NA
## 11        5   Andrea      5
## 12        6   Andrea     NA
## 13        1   Darryl      4
## 14        2   Darryl     NA
## 15        3   Darryl     NA
## 16        4   Darryl      4
## 17        5   Darryl      3
## 18        6   Darryl      5
## 19        1  Beverly      2
## 20        2  Beverly      3
## 21        3  Beverly     NA
## 22        4  Beverly      1
## 23        5  Beverly     NA
## 24        6  Beverly      3
## 25        1   Maysie     NA
## 26        2   Maysie     NA
```

```
## 27        3   Maysie      5
## 28        4   Maysie     NA
## 29        5   Maysie      4
## 30        6   Maysie     NA
## 31        1    Karen      5
## 32        2    Karen      2
## 33        3    Karen      4
## 34        4    Karen     NA
## 35        5    Karen     NA
## 36        6    Karen     NA
## 37        1  Jassiem     NA
## 38        2  Jassiem      4
## 39        3  Jassiem      2
## 40        4  Jassiem      3
## 41        5  Jassiem     NA
## 42        6  Jassiem     NA
## 43        1     Marc     NA
## 44        2     Marc      3
## 45        3     Marc     NA
## 46        4     Marc     NA
## 47        5     Marc      5
## 48        6     Marc     NA
## 49        1    Trudy      2
## 50        2    Trudy     NA
## 51        3    Trudy      4
## 52        4    Trudy     NA
## 53        6    Trudy     NA
## 54        1      Kim     NA
## 55        2      Kim     NA
## 56        3      Kim      2
## 57        4      Kim      3
## 58        5      Kim     NA
## 59        6      Kim      4
## 60        5    Trudy     NA
```

Join the three tables movie, reviewer, rating to make one table called "movie_ratings", which I loaded into a data frame.

```
movie_ratings <- dbSendQuery(mydb, "SELECT
    m.title As 'Title',
    m.genre As 'Genre',
    r.first_name AS 'Reviewer',
    s.rating As 'Rating'
FROM movie m
JOIN reviewer r
ON m.movie_id = r.movie_id
JOIN score s
ON r.first_name = s.reviewer
AND r.movie_id = s.movie_id");

#dbFetch(movie_ratings) and create data frame
movie_ratings<-fetch(movie_ratings)
print(movie_ratings)
```

```
##                  Title            Genre Reviewer Rating
## 1      The Equalizer 3 Action/Adventure    Craig      5
## 2        Grand Turismo Action/Adventure    Craig      4
## 3  The Retirement Plan           Comedy    Craig      4
## 4           The Nun II           Horror    Craig     NA
## 5           Dumb Money     Comedy/Drama    Craig     NA
## 6          Expend4bles  Action/Thriller    Craig      5
## 7      The Equalizer 3 Action/Adventure   Andrea     NA
## 8        Grand Turismo Action/Adventure   Andrea      4
## 9  The Retirement Plan           Comedy   Andrea      3
## 10          The Nun II           Horror   Andrea     NA
## 11          Dumb Money     Comedy/Drama   Andrea      5
## 12         Expend4bles  Action/Thriller   Andrea     NA
## 13     The Equalizer 3 Action/Adventure   Darryl      4
## 14       Grand Turismo Action/Adventure   Darryl     NA
## 15 The Retirement Plan           Comedy   Darryl     NA
## 16          The Nun II           Horror   Darryl      4
## 17          Dumb Money     Comedy/Drama   Darryl      3
## 18         Expend4bles  Action/Thriller   Darryl      5
## 19     The Equalizer 3 Action/Adventure  Beverly      2
## 20       Grand Turismo Action/Adventure  Beverly      3
## 21 The Retirement Plan           Comedy  Beverly     NA
## 22          The Nun II           Horror  Beverly      1
## 23          Dumb Money     Comedy/Drama  Beverly     NA
## 24         Expend4bles  Action/Thriller  Beverly      3
## 25     The Equalizer 3 Action/Adventure   Maysie     NA
## 26       Grand Turismo Action/Adventure   Maysie     NA
## 27 The Retirement Plan           Comedy   Maysie      5
## 28          The Nun II           Horror   Maysie     NA
## 29          Dumb Money     Comedy/Drama   Maysie      4
## 30         Expend4bles  Action/Thriller   Maysie     NA
## 31     The Equalizer 3 Action/Adventure    Karen      5
## 32       Grand Turismo Action/Adventure    Karen      2
## 33 The Retirement Plan           Comedy    Karen      4
## 34          The Nun II           Horror    Karen     NA
## 35          Dumb Money     Comedy/Drama    Karen     NA
## 36         Expend4bles  Action/Thriller    Karen     NA
## 37     The Equalizer 3 Action/Adventure  Jassiem     NA
## 38       Grand Turismo Action/Adventure  Jassiem      4
## 39 The Retirement Plan           Comedy  Jassiem      2
## 40          The Nun II           Horror  Jassiem      3
## 41          Dumb Money     Comedy/Drama  Jassiem     NA
## 42         Expend4bles  Action/Thriller  Jassiem     NA
## 43     The Equalizer 3 Action/Adventure     Marc     NA
## 44       Grand Turismo Action/Adventure     Marc      3
## 45 The Retirement Plan           Comedy     Marc     NA
## 46          The Nun II           Horror     Marc     NA
## 47          Dumb Money     Comedy/Drama     Marc      5
## 48         Expend4bles  Action/Thriller     Marc     NA
## 49     The Equalizer 3 Action/Adventure    Trudy      2
## 50       Grand Turismo Action/Adventure    Trudy     NA
## 51 The Retirement Plan           Comedy    Trudy      4
## 52          The Nun II           Horror    Trudy     NA
## 53          Dumb Money     Comedy/Drama    Trudy     NA
```

```
## 54         Expend4bles  Action/Thriller    Trudy     NA
## 55     The Equalizer 3 Action/Adventure      Kim     NA
## 56       Grand Turismo Action/Adventure      Kim     NA
## 57 The Retirement Plan           Comedy      Kim      2
## 58          The Nun II           Horror      Kim      3
## 59          Dumb Money     Comedy/Drama      Kim     NA
## 60         Expend4bles  Action/Thriller      Kim      4
## 61         Expend4bles  Action/Thriller    Craig      5
## 62         Expend4bles  Action/Thriller   Andrea     NA
## 63         Expend4bles  Action/Thriller   Darryl      5
## 64         Expend4bles  Action/Thriller  Beverly      3
## 65         Expend4bles  Action/Thriller   Maysie     NA
## 66         Expend4bles  Action/Thriller    Karen     NA
## 67         Expend4bles  Action/Thriller  Jassiem     NA
## 68         Expend4bles  Action/Thriller     Marc     NA
## 69         Expend4bles  Action/Thriller    Trudy     NA
## 70         Expend4bles  Action/Thriller      Kim      4
```

```r
# Checked the structure of the data. 50 rows. 4columns.

str(movie_ratings)
```

```
## 'data.frame':    70 obs. of  4 variables:
##  $ Title   : chr  "The Equalizer 3" "Grand Turismo" "The Retirement Plan" "The Nun II" ...
##  $ Genre   : chr  "Action/Adventure" "Action/Adventure" "Comedy" "Horror" ...
##  $ Reviewer: chr  "Craig" "Craig" "Craig" "Craig" ...
##  $ Rating  : int  5 4 4 NA NA 5 NA 4 3 NA ...
```

Tidy the data Since every reviewer did not see every movie, we have to exclude the blank fields from any calculations. We do this my filtering them out by using "!is.na" function.

```r
# Group by title to see the average score for each movie rated; use is.na to exclude blank fields.

new_ratings <- movie_ratings %>%
  filter(!is.na(Rating)) %>%
   group_by(Title) %>%
    summarise(Avg_Score = mean(as.numeric(Rating))) %>%
      arrange(desc(Avg_Score)
      )

new_ratings
```

```
## # A tibble: 6 x 2
##   Title               Avg_Score
##   <chr>                   <dbl>
## 1 Dumb Money               4.25
## 2 Expend4bles              4.25
## 3 The Equalizer 3          3.6
## 4 The Retirement Plan      3.43
## 5 Grand Turismo            3.33
## 6 The Nun II               2.75
```

#See average rating per reviewer

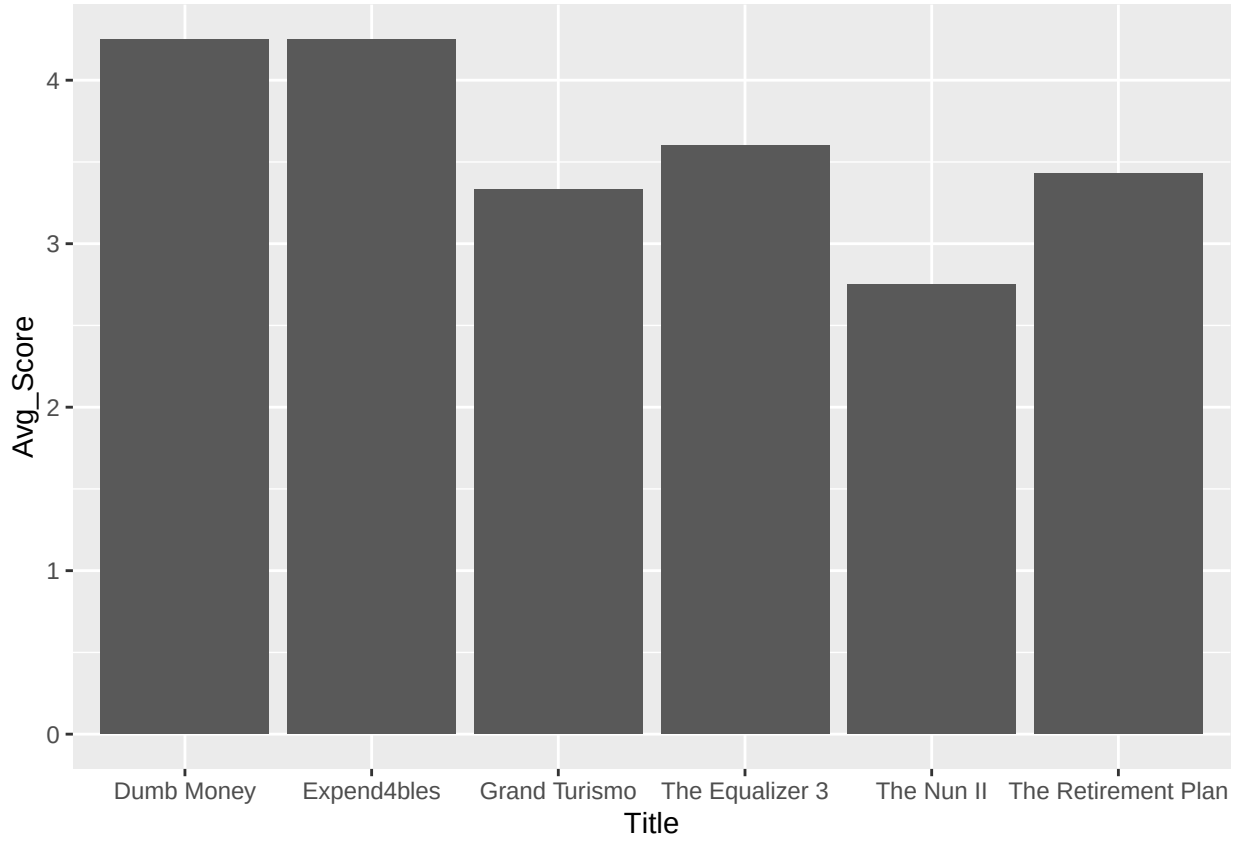
```
reviewer_ratings <- movie_ratings %>%
  group_by(Reviewer)%>%
    summarise (Avg_rating = mean(Rating, na.rm=TRUE)
  )

reviewer_ratings
```

```
## # A tibble: 10 x 2
##    Reviewer Avg_rating
##    <chr>         <dbl>
##  1 Andrea         4
##  2 Beverly        2.4
##  3 Craig          4.6
##  4 Darryl         4.2
##  5 Jassiem        3
##  6 Karen          3.67
##  7 Kim            3.25
##  8 Marc           4
##  9 Maysie         4.5
## 10 Trudy          3
```

ggplot of the Average Score for each movie

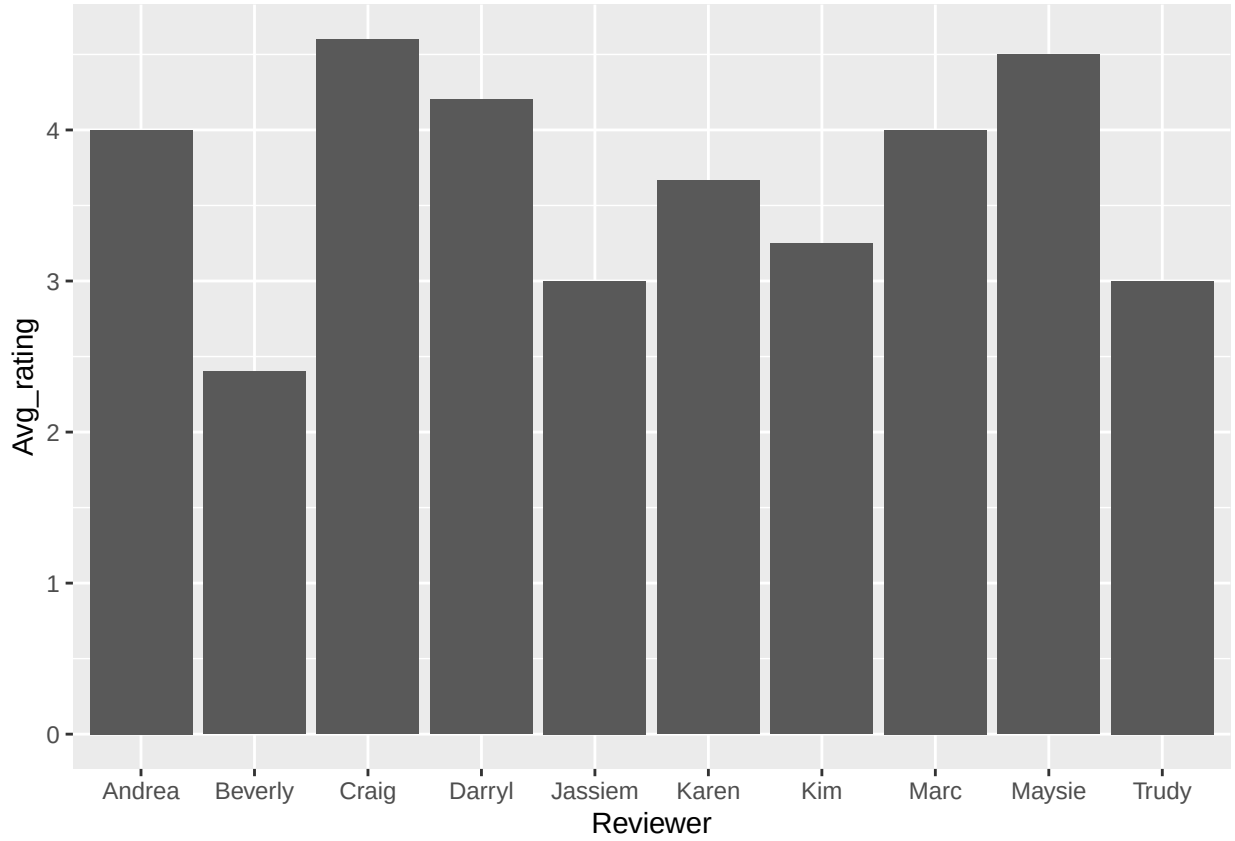
```
new_ratings %>%
ggplot +
  coord_flip() +
  geom_col(aes(Avg_Score, Title)
   )
```



Graphically, "The Nun II" has the lowest rating; "Dumb Money and Expend4bles" are equally tied at a rating of 4 each.

#Lets see how the reviewer ratings for each movie

```
reviewer_ratings %>%
ggplot +
  geom_col(aes(Reviewer, Avg_rating)
   )
```



At first glance, we can easily conclude that Beverly "grades" movies more harshly than others with an on-average rating of approximately 2.4, which would be true if everyone saw/reviewed every movie, but such is not the case. Some reviewed two movies while others reviewed as many as four, in essence skewing the results. More data exploration is necessary, but I am unsure how to illustrate this graphically.

```
# Lets look at the count for each rating per movie.
options(dplyr.summarise.inform = FALSE)  #silences warning message

count_reviewed <- movie_ratings
count_reviewed %>% group_by(Title, Rating) %>% summarise(count = n())%>%
arrange(desc(Title))
```

```
## # A tibble: 25 x 3
## # Groups:   Title [6]
##    Title               Rating count
##    <chr>                <int> <int>
##  1 The Retirement Plan      2     2
##  2 The Retirement Plan      3     1
##  3 The Retirement Plan      4     3
##  4 The Retirement Plan      5     1
##  5 The Retirement Plan     NA     3
##  6 The Nun II               1     1
##  7 The Nun II               3     2
##  8 The Nun II               4     1
##  9 The Nun II              NA     6
## 10 The Equalizer 3          2     2
```

```
## # i 15 more rows
```