Assignment 2

Abdelmalek Hajjam

9/7/2019

In this assignment, I'm using a local instance of MySQL. I created a schema (Database) called **movie-rating**. The accompanying file called **movie-rating.sql** will create the tables and populate them with data.

My tables are:

1. movies table
2. reviewers table
3. Surveys_movie_rating table

I'm normalizing my tables to make them very flexible, like so:

CREATE TABLE movies (movie_id int(6) NOT NULL AUTO_INCREMENT, movie_title varchar(50) NOT NULL, PRIMARY KEY (movie_id));

CREATE TABLE reviewers (reviewer_id int(6) NOT NULL AUTO_INCREMENT , reviewer_name char(50), PRIMARY KEY (reviewer_id)
);

CREATE TABLE Surveys_Movie_Rating (survey_id int(6) NOT NULL AUTO_INCREMENT, reviewer_id int(6), movie_id int(6), Rating int(2), PRIMARY KEY (survey_id));

In the Surveys_Movie_Rating table, the **Rating** field is a numeric field with values from 1 to 5. 1 being poor and 5 being excellent.

Let's load the RMySQL and DBI packages after installing them

```
#install.packages("RMySQL")
library(RMySQL)
```

```
## Loading required package: DBI
```

```
#let's also load DBI; normally the system will load it automatically for us. loading it manually after
library(DBI)
```

let's connect to the database using MySQL driver and return the tables in the database to make sure everything is solid

```
mydb <- dbConnect(MySQL(), user='root', password='theoracley', dbname='movie_rating', host='localhost')

#list of tables in the schema 'movie_rating'
tablesList <- dbListTables(mydb)
tablesList
```

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

Let's read each table data

```
#get data from the movies table
moviesData <- dbGetQuery(mydb, 'select * from movies;')
moviesData
```

```
##   movie_id      movie_title
## 1        1       Casablanca
## 2        2 King of New York
## 3        3       Spider-Man
## 4        4       The Matrix
## 5        5       Home Alone
## 6        6     wonder Woman
```

```
#get data from the reviewers table
reviewersData <- dbGetQuery(mydb, 'select * from reviewers;')
reviewersData
```

```
##   reviewer_id reviewer_name
## 1           1  Manal Hajjam
## 2           2   Hicham Nour
## 3           3   Soumia Hadi
## 4           4  Nancy Falour
## 5           5     Tony Atia
```

```
#get data from the surveys_movie_rating table
surveysData <- dbGetQuery(mydb, 'select * from surveys_movie_rating;')
surveysData
```

```
##    survey_id reviewer_id movie_id Rating
## 1          1           1        1      5
## 2          2           1        2      4
## 3          3           1        3      2
## 4          4           1        4      4
## 5          5           1        5      3
## 6          6           1        6      5
## 7          7           2        1      5
## 8          8           2        2      1
## 9          9           2        3      3
## 10        10           2        4      2
## 11        11           2        5      2
## 12        12           2        6      4
## 13        13           3        1      5
## 14        14           3        2      3
## 15        15           3        3      4
## 16        16           3        4      3
## 17        17           3        5      3
## 18        18           3        6      2
## 19        19           4        1      4
## 20        20           4        2      5
```

```
## 21         21           4         3      3
## 22         22           4         4      2
## 23         23           4         5      4
## 24         24           4         6      3
## 25         25           5         1      4
## 26         26           5         2      2
## 27         27           5         3      5
## 28         28           5         4      3
## 29         29           5         5      2
## 30         30           5         6      4
```

While the data returned from surveys is for storage, a user cannot really read it and make sense of it, we need to return data that is very readable for users. Let's do it by combining the other tables with the main table

```
#return surveys data in a format that can be easily readable
data <- dbGetQuery(mydb, 'SELECT s.survey_id, r.reviewer_name, m.movie_title, s.Rating FROM Surveys_Mov

data
```

```
##    survey_id reviewer_name      movie_title Rating
## 1          1  Manal Hajjam       Casablanca      5
## 2          2  Manal Hajjam King of New York      4
## 3          3  Manal Hajjam       Spider-Man      2
## 4          4  Manal Hajjam       The Matrix      4
## 5          5  Manal Hajjam       Home Alone      3
## 6          6  Manal Hajjam     wonder Woman      5
## 7          7   Hicham Nour       Casablanca      5
## 8          8   Hicham Nour King of New York      1
## 9          9   Hicham Nour       Spider-Man      3
## 10        10   Hicham Nour       The Matrix      2
## 11        11   Hicham Nour       Home Alone      2
## 12        12   Hicham Nour     wonder Woman      4
## 13        13   Soumia Hadi       Casablanca      5
## 14        14   Soumia Hadi King of New York      3
## 15        15   Soumia Hadi       Spider-Man      4
## 16        16   Soumia Hadi       The Matrix      3
## 17        17   Soumia Hadi       Home Alone      3
## 18        18   Soumia Hadi     wonder Woman      2
## 19        19  Nancy Falour       Casablanca      4
## 20        20  Nancy Falour King of New York      5
## 21        21  Nancy Falour       Spider-Man      3
## 22        22  Nancy Falour       The Matrix      2
## 23        23  Nancy Falour       Home Alone      4
## 24        24  Nancy Falour     wonder Woman      3
## 25        25     Tony Atia       Casablanca      4
## 26        26     Tony Atia King of New York      2
## 27        27     Tony Atia       Spider-Man      5
## 28        28     Tony Atia       The Matrix      3
## 29        29     Tony Atia       Home Alone      2
## 30        30     Tony Atia     wonder Woman      4
```

Let's see which movie got the highest Rating

```
#return the best movie on the top of the list with it's average
theBest <- dbGetQuery(mydb, 'select s.movie_id, m.movie_title, avg(Rating) as AVG_Rating
from Surveys_Movie_Rating s, movies m where s.movie_id = m.movie_id
group by movie_id
order by avg(Rating) desc;')
```

```
## Warning in .local(conn, statement, ...): Decimal MySQL column 2 imported as
## numeric
```

```
theBest
```

```
##   movie_id      movie_title AVG_Rating
## 1        1       Casablanca        4.6
## 2        6     wonder Woman        3.6
## 3        3       Spider-Man        3.4
## 4        2 King of New York        3.0
## 5        4       The Matrix        2.8
## 6        5       Home Alone        2.8
```

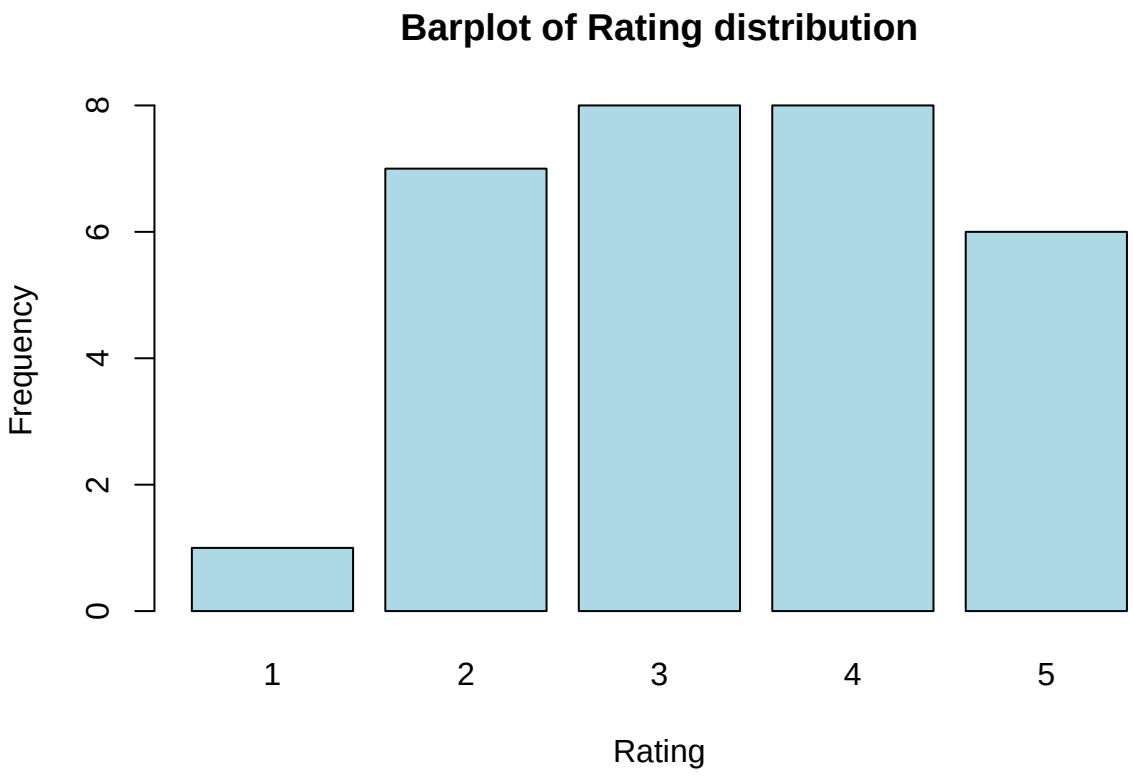
It looks like the movie 'Casablanca' is the winner with an average rating of 4.6.

Let's do some ditribution ploting

```
#let's see the ditribution for Rating
table(data$Rating)
```

```
##
## 1 2 3 4 5
## 1 7 8 8 6
```

```
#Let's plot the Rating ditribution
barplot(table(data$Rating),xlab='Rating', ylab='Frequency', main='Barplot of Rating distribution', col=
```



We see that Rating 1 was given only 1 time, Rating 2 was given 7 times, Rating 3 was given 8 times, Rating 4 was given 8 times too, and finally Rating 5 was given 6 times.

Finally, let's close the connection and release resources

```
dbDisconnect(mydb)
```

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