Lab 2 Seung Min Song

2022-09-11

Contents


```
install.packages('dplyr')
library(tidyverse)
library(RMySQL)
library(DBI)
library(dplyr)
```

Survey software

I used Google Forms to gather the movie reviewers' survey response.

CSV

Please visit https://github.com/seung-m1nsong/607 homework2 to view my mysql query

View GitHub CSV file to RStudio

```
movie <- data.table::fread( "https://raw.githubusercontent.com/seung-m1nsong/607/main/homework2/Movie_Ra
              select=c("firstname","lastname","age", "sex", "location","SpiderMan","Thor","Top_Gun","Mo

movie
```

```
##    firstname lastname age    sex       location SpiderMan Thor Top_Gun Morbius
## 1:       Ben    Brand  32   Male  North America         5    3       5       2
## 2:    Samuel      Kim  22   Male  North America         4    3       3       1
## 3:       Ray    Watts  55   Male  North America         4    2       5       2
## 4:     Kelly   Kovack  37 Female  North America         4    3       4       0
## 5:      Reno  Jimenez  40   Male  North America         4    1       5       2
## 6:     Maria     Riva  27 Female Eastern Europe         5    4       4       0
## 7:    Daniel    Hanry  38   Male  North America         4    3       5       1
```

```
## 8:   Raymund      Suh  40   Male           Asia         5    3       5       0
##    Toronto Doctor_Strange
## 1:       1              1
## 2:       1              1
## 3:       0              1
## 4:       3              0
## 5:       1              1
## 6:       2              2
## 7:       0              2
## 8:       0              1
```

mysql

Please visit https://github.com/seung-m1nsong/607 homework2 to view my mysql query

Connect RStudio with my local mysql

```
con <- DBI::dbConnect(
  RMySQL::MySQL(),
  user="root",
  password="",
  host="localhost",
  dbname="moviedatabase"
)
```

View movie list

```
 dfMovielist <- DBI::dbGetQuery(con,
"select * from movie")
dfMovielist
```

```
##   movieid                                  movietitle releaseyear  genre
## 1       1                   Spider-Man: No Way Home        2021 Action
## 2       2                    Thor: Love and Thunder        2022 Action
## 3       3                         Top Gun: Maverick        2022 Action
## 4       4                                   Morbius        2022 Action
## 5       5                                   Toronto        2022  Comic
## 6       6 Doctor Strange in the Multiverse of Madness        2022 Action
```

View all movie reviewer

```
 dfMovieReviewer <- DBI::dbGetQuery(con,
"select * from reviewer")
dfMovieReviewer
```

```
##   reviewerid firstname lastname age    sex       location
## 1          1       Ben    Brand  32   Male  North America
## 2          2    Samuel      Kim  22   Male  North America
## 3          3       Ray    Watts  55   Male  North America
## 4          4     Kelly   Kovack  37 Female  North America
## 5          5      Reno  Jimenez  40   Male  North America
```

```
## 6          6     Maria     Riva  27 Female Eastern Europe
## 7          7    Daniel    Hanrt  38   Male  North America
## 8          8   Raymund      Suh  40   Male           Asia
```

Show movie title with rating from reviewers.
I joined two table: Table called 'move' and 'movierating'

```
 dfMovieReviewAverage <- DBI::dbGetQuery(con,
"select movie.movietitle, avg(movierating.rating) as avg_rating, min(movierating.rating) as min_rating,
from  movie
inner join movierating
ON movierating.movie = movie.movietitle
group by movie.movietitle;")

dfMovieReviewAverage
```

```
##                                   movietitle avg_rating min_rating max_rating
## 1 Doctor Strange in the Multiverse of Madness      1.125          0          2
## 2                                     Morbius      1.000          0          2
## 3                       Spider-Man: No Way Home      4.375          4          5
## 4                        Thor: Love and Thunder      2.750          1          4
## 5                             Top Gun: Maverick      4.250          3          5
## 6                                       Toronto      1.000          0          3
```

About missing data

I gathered my data from Google Forms. I made the entire field a required field to avoid any missing data. If it is really important data, it is important not to omit it in the first place. The best way to avoid any missing important data is to ensure reviewer to not to missing data insertion. I gave reviewers a clear instruction which movie to rate. This way I can avoid multiple selections.

Please visit the link below to check the Goole Forms:
https://forms.gle/eRduhTBvFjz1khL26

Standarizing data

Data standardization is an important due to the fact that it provides a structure for creating and maintaining data quality

For example, the analysis of vulnerable areas for Covid-19 is a standard analysis model developed by New York City. However, if New Jersey data is inputed according to the data format, it can be used in New Jersey. This model can be easily used to analyze areas where there is shortage of COVID-19 screening clinics.