

DATA607_Assignment SQL and R

Gabriel Santos

2022-09-10

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.9
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Calling the necessary packages

```
library(DBI)
library(RMySQL)
library(RODBC)
```

Establish MySQL connection

```
mydb = dbConnect(MySQL(), user='root', password='Sysadm123', dbname='movie_data', host='localhost')
```

Table list

```
dbListTables(mydb)

## [1] "movies"                "people"
## [3] "peoplemovieratingsresults" "peoplemoviesratings"
```

Table fields

```
dbListFields(mydb, 'movie_data.people')

## [1] "NameID"      "FirstName" "LastName"
```

```
dbListFields(mydb, 'movie_data.movies')
```

```
## [1] "MovieID" "MovieTitle"
```

```
dbListFields(mydb, 'movie_data.PeopleMoviesRatings')
```

```
## [1] "RatingID" "NameID" "MovieID" "Rating"
```

```
dbListFields(mydb, 'movie_data.PeopleMovieRatingsResults')
```

```
## [1] "FirstName" "LastName" "MovieTitle" "Rating"
```

```
PeopleMoviesRatings = dbSendQuery(mydb, "select * from movie_data.PeopleMoviesRatings")
```

Fetch table people from database

```
PeopleMoviesRatings = fetch(PeopleMoviesRatings, n = 5)  
print(PeopleMoviesRatings)
```

```
## RatingID NameID MovieID Rating  
## 1 1 1 1 4  
## 2 2 1 2 4  
## 3 3 1 3 2  
## 4 4 1 4 5  
## 5 5 1 5 3
```

movies table fields

Reading a csv file to R

```
mysql <- read.csv("https://raw.githubusercontent.com/GabrielSantos33/DATA607-Assignment-2/main/movie_ratings.csv")  
head(mysql)
```

I exported my sq data into a csv file and I placed the csv file on Github. I then made R read the csv file on Github.

```
## movie Fri_name Stars  
## 1 Eternals Ahmed 4  
## 2 Shang-Chi Ahmed 4  
## 3 Spider-Man No Way Home Ahmed 5  
## 4 Dune Ahmed 3  
## 5 Venom Ahmed 1  
## 6 No Time to Die Ahmed NULL
```

Average Rating of Movies My Friends Watched

```

avg_mov <- mysql %>%
  group_by(movie) %>%
  filter(Stars != "NULL") %>%
  summarise(avg= mean(as.integer(Stars)))
avg_mov

```

I first filtered out the null values and then I had to make the chr values under Stars into integer values to calculate the average. To know the average rating among the movies my friends had seen.

```

## # A tibble: 6 x 2
##   movie          avg
##   <chr>        <dbl>
## 1 "Dune"         3
## 2 "Eternals"    2.8
## 3 "No Time to Die" 3
## 4 "Shang-Chi"   3.6
## 5 "Spider-Man No Way Home" 4.8
## 6 "Venom "     2.2

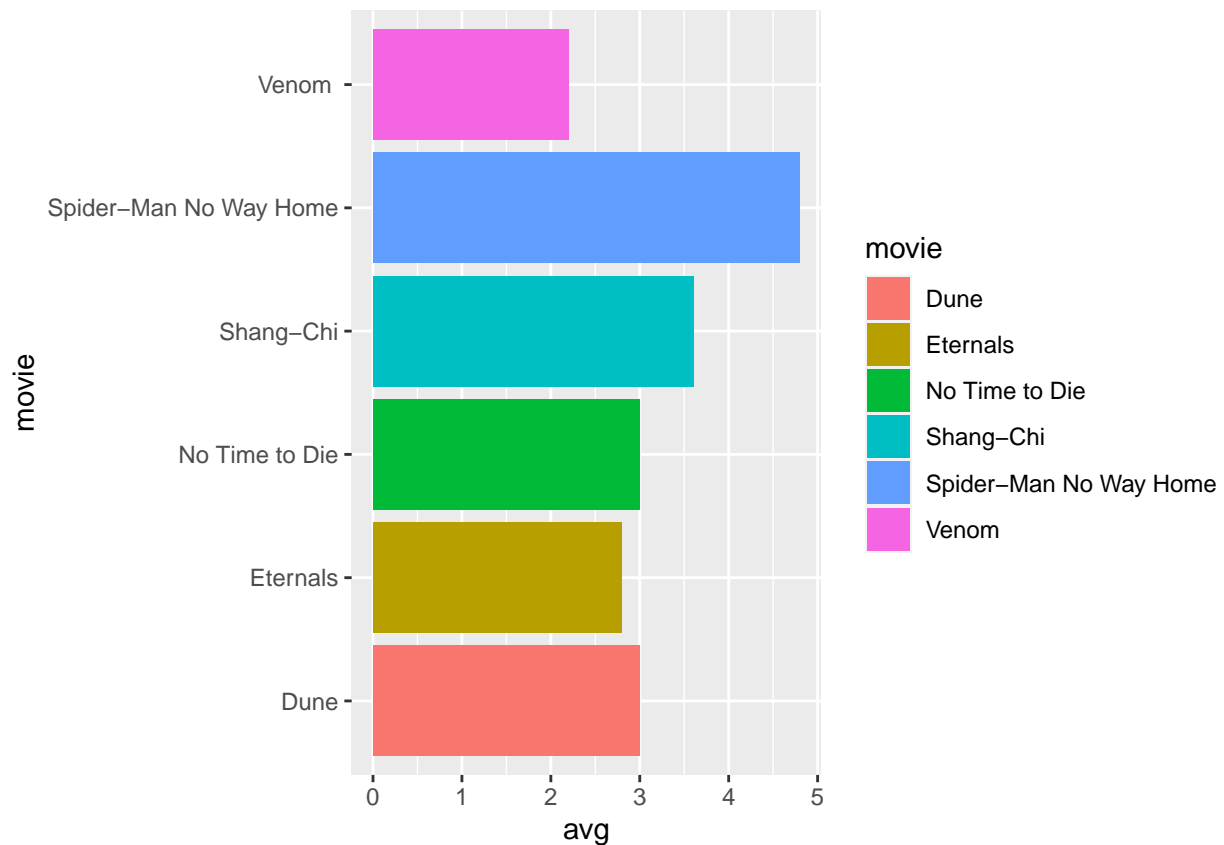
```

Spider-Man was watched by all my friends and they also rated it very highly with a rating of 4.8

```

library(ggplot2)
ggplot(data=avg_mov, aes(x=movie,y=avg , fill=movie)) +
  coord_flip() +
  geom_bar(stat="identity")

```



Movies my Friends did not watched and how many didn't watch.

Movies my friends did not watch so I aggregated the data by movie, I filtered the condition where the stars were null and I counted how many people did not watch the movies.

4 of my friends did not watch No Time to Die and one of my friend did not watch dune.

```
nul <- mysql %>%
  group_by(movie) %>%
  filter(Stars=="NULL") %>%
  count(Stars,sort=TRUE)
nul
```

```
## # A tibble: 2 x 3
## # Groups:   movie [2]
##   movie      Stars     n
##   <chr>      <chr> <int>
## 1 No Time to Die NULL     4
## 2 Dune        NULL     1
```

Results

```
library(ggplot2)
ggplot(data=nul,aes(x=movie,y=n,fill=movie)) +
  geom_bar(stat="Identity")
```

