607HW2 SQL and R

Sean Connin

Assignment Overview

Develop a database and relational tables for viewer movie ratings (Likert scale 1-5) that can be queried using SQL and R. The latter should be able to generate the SQL tables and data from provided code.

Approach Taken:

1. Collected viewer data using Google Forms survey instrument
2. Downloaded ratings data into csv format
3. Subset, cleaned, and created csv subsets in Rstudio
4. Built database and tables in MySQL using MYSQL workbench
5. Connected Rstudio and MYSQL database to demonstrate data queries
6. Uploaded scripts and data files to Github:

Github repository: https://github.com/sconnin/607HW2

1.Data import

1. import as as csv, remove any spaces in column headers
2. review table attributes

```
#Step 1

m_survey <- read_csv('hw2_DB.csv',  na = c(" ", "", "NA"))

##
## -- Column specification --------------------------------------------------
------
## cols(
##   Id = col_double(),
##   `Top 10 Most Watched Netflix Shows in 2020 [The Queens Gambit]` = col_ch
aracter(),
##   `Top 10 Most Watched Netflix Shows in 2020 [Emily in Paris]` = col_chara
cter(),
##   `Top 10 Most Watched Netflix Shows in 2020 [Lucifer]` = col_character(),
##   `Top 10 Most Watched Netflix Shows in 2020 [The Umbrella Academy]` = col
_character(),
##   `Top 10 Most Watched Netflix Shows in 2020 [Money Heist]` = col_characte
r(),
##   `Top 10 Most Watched Netflix Shows in 2020 [Dark Desire]` = col_characte
r(),
##   `Top 10 Most Watched Netflix Shows in 2020 [Friends]` = col_character(),
```

```
##   `Top 10 Most Watched Netflix Shows in 2020 [The Crown]` = col_character(
),
##   `Top 10 Most Watched Netflix Shows in 2020 [Ratched]` = col_character(),
##   `Top 10 Most Watched Netflix Shows in 2020 [Dark]` = col_character(),
##   `Which TV and/or movie genres do you enjoy watching most?` = col_charact
er(),
##   `Which TV and/or movie genres do you enjoy watching least?` = col_charac
ter(),
##   `On average, how many hours a week do you spend on Netflix each week?` =
col_double(),
##   `What movie or TV show on Netflix or other streaming services would you
highly recommend to adults that wasn't on this list?` = col_character()
## )
```

```
#Step 2

m_survey%>%class()

## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"

m_survey%>%dim()

## [1] 12 15
```

2.Clean and subset data for viewer ratings csv

Steps:

1. Rename columns
2. Subset columns
3. Pivot longer and recode_factor categorical variables
4. Repivot to tidy format
5. Write to csv file

```
#Step 1

m_survey%<>%dplyr::rename(Queens_Gambit="Top 10 Most Watched Netflix Shows in
2020 [The Queens Gambit]", Emily_in_Paris="Top 10 Most Watched Netflix Shows
in 2020 [Emily in Paris]", Lucifer = "Top 10 Most Watched Netflix Shows in 20
20 [Lucifer]", The_Umbrella_Academy = "Top 10 Most Watched Netflix Shows in 2
020 [The Umbrella Academy]", Money_Heist= "Top 10 Most Watched Netflix Shows
in 2020 [Money Heist]", Dark_Desire="Top 10 Most Watched Netflix Shows in 202
0 [Dark Desire]",Friends="Top 10 Most Watched Netflix Shows in 2020 [Friends]
",The_Crown="Top 10 Most Watched Netflix Shows in 2020 [The Crown]", Ratched=
"Top 10 Most Watched Netflix Shows in 2020 [Ratched]", Dark="Top 10 Most Watc
hed Netflix Shows in 2020 [Dark]", Genres_Liked="Which TV and/or movie genres
do you enjoy watching most?", Genres_Disliked="Which TV and/or movie genres d
o you enjoy watching least?", Viewing_Hours = "On average, how many hours a w
eek do you spend on Netflix each week?", Recommended="What movie or TV show o
n Netflix or other streaming services would you highly recommend to adults th
at wasn't on this list?")
```

```
#Step 2

m_rating <- m_survey%>%select(-c(Genres_Liked,Genres_Disliked, Recommended, V
iewing_Hours))

#Step 3

m_rating<-m_rating%>%pivot_longer(cols=2:11, names_to = 'Movies', values_to =
'Rating')

m_rating<-m_rating%>%mutate(Rating=recode_factor(Rating, "No opinion - I have
n't seen it" = '0', "Poor"='1', "Fair"='2', "Average"='3', "Good"='4', "Excel
lent"='5', .ordered=TRUE ))

#Step 4

m_rating%<>%pivot_wider(names_from = Movies, values_from = Rating)%>% rename_
all(make.names)

#Step 5

write_csv(m_rating, path="m_rating.csv")

## Warning: The `path` argument of `write_csv()` is deprecated as of readr 1.
4.0.
## Please use the `file` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

3. Clean and subset data for viewing hours csv

1. Subset columns
2. Add column for foreign key in db table
3. Write to csv

```
#Step 1

m_hrs <- m_survey%>%select(c(Id, Viewing_Hours))

#Step 2

m_hrs%<>% mutate(Id2 = Id)

#Step 3

write_csv(m_hrs, path="m_hrs.csv")
```

4. Create df for viewer names and save to csv

1. Create name list

2. Convert list to df
3. Add Id col
4. Write to csv

```
#Step 1

names<- list(Name=c('Sam', "Sarah", "Jennifer", "Euardo","Laura", "Mary", "Ta
ylor", "Kiesha","Rob", "Andrew", "Beth","Maria"))

#Step 2

viewer<-as.data.frame(names)

#Step 3

viewer%<>%mutate(Id=row_number())%>%relocate(Id, .before=Name)

#Step 4

write_csv(viewer, path="names.csv")
```

5. Query SQL database

1. Open mysql connection and list tables
2. Left join viewer name, rating and viewing hrs tables
3. Drop duplicate Id columns
4. Identify three highest ranked movies
5. Disconnect mysql

*viewer ratings key: +No opinion - I haven't seen it = 0 +Poor = 1 +Fair = 2 +Average = 3 +Good = 4 +Excellent = 5

```
#Step 1

con <- dbConnect(RMariaDB::MariaDB(),user='root', password='XXXXX', dbname='6
07hw2',host='localhost')

dbListTables(con)

## [1] "viewer"        "viewer_rating" "viewing_hrs"

# Step 2

sql <- 'SELECT *
FROM viewer v
LEFT JOIN viewer_rating vr
ON v.Id = vr.Id2
Left JOIN viewing_hrs vh
ON v.Id = vh.Id2'
```

```
join_table <- dbGetQuery(con,sql)

#Step 3

join_table%<>%dplyr::select(-c(Id..3,Id2,Id..15,Id2..17))

join_table
```

```
##    Id     Name Queens_Gambit Emily_in_Paris Lucifer The_Umbrella_Academy
## 1   1      Sam             5              1       0                     0
## 2   2    Sarah             5              3       4                     0
## 3   3 Jennifer             5              0       4                     3
## 4   4   Euardo             4              0       5                     0
## 5   5    Laura             0              0       0                     0
## 6   6     Mary             1              0       0                     0
## 7   7   Taylor             0              0       0                     0
## 8   8   Kiesha             0              0       0                     0
## 9   9      Rob             0              3       3                     4
## 10 10   Andrew             0              0       0                     0
## 11 11     Beth             2              3       4                     0
## 12 12    Maria             5              0       0                     0
##    Money_Heist Dark_Desire Friends The_Crown Ratched Dark Viewing_Hours
## 1            0           0       2         0       0    0            20
## 2            4           3       3         4       0    0             2
## 3            0           0       4         0       0    0             1
## 4            0           0       5         0       0    0             8
## 5            0           0       0         0       0    0             5
## 6            0           0       1         0       0    0             0
## 7            0           0       4         4       0    0             1
## 8            0           0       0         0       0    5             5
## 9            0           0       2         0       4    0            12
## 10           0           0       2         5       0    0             4
## 11           0           0       0         0       0    0             6
## 12           0           0       5         0       0    4             5
```

```
#Step 4

rating_sum<-join_table%>%select(c(2:12))%>%summarize_if(is.numeric, sum, na.rm=TRUE)
rating_sum
```

```
##   Queens_Gambit Emily_in_Paris Lucifer The_Umbrella_Academy Money_Heist
## 1            27             10      20                     7           4
##   Dark_Desire Friends The_Crown Ratched Dark
## 1           3      28        13       4    9
```

```
rating_sum%<>%pivot_longer(cols=1:10, names_to = 'Movies', values_to = 'Rating_Sum')%>%arrange(desc(Rating_Sum))%>%slice_max(Rating_Sum, n = 3)

rating_sum
```

```
## # A tibble: 3 x 2
##   Movies        Rating_Sum
##   <chr>              <int>
## 1 Friends               28
## 2 Queens_Gambit         27
## 3 Lucifer               20

#Step 5

dbDisconnect(con)
```