

607HW2_Connin

Load Libraries

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr    0.3.4
## v tibble   3.0.6     v dplyr    1.0.3
## v tidyr    1.1.2     v stringr  1.4.0
## v readr    1.4.0     vforcats  0.5.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()

library(readr)
library(magrittr)

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##       set_names

## The following object is masked from 'package:tidyr':
##       extract

library(RMariaDB)
library(odbc) # --> interface btween db driver and r
library(DBI) # -- > standardizes func relates to db operations
```

```
import csv
```

```
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_character(),
##   'Top 10 Most Watched Netflix Shows in 2020 [Emily in Paris]' = col_character(),
##   '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_character(),
##   'Top 10 Most Watched Netflix Shows in 2020 [Dark Desire]' = col_character(),
##   '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_character(),
##   'Which TV and/or movie genres do you enjoy watching least?' = col_character(),
##   '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'
## )

m_survey%>%class()

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

m_survey%>%dim()

## [1] 12 15

(m_survey)

## # A tibble: 12 x 15
##       Id `Top 10 Most Wa~ `Top 10 Most Wa~ `Top 10 Most Wa~ `Top 10 Most Wa~
##     <dbl> <chr>          <chr>          <chr>          <chr>
## 1     1 Excellent      Poor          No opinion - I ~ No opinion - I ~
## 2     2 Excellent      Average       Good          No opinion - I ~
## 3     3 Excellent      No opinion - I ~ Good          Average
## 4     4 Good           No opinion - I ~ Excellent      No opinion - I ~
## 5     5 No opinion - I ~ No opinion - I ~ No opinion - I ~ No opinion - I ~
## 6     6 Poor           No opinion - I ~ No opinion - I ~ No opinion - I ~
## 7     7 No opinion - I ~ No opinion - I ~ No opinion - I ~ No opinion - I ~
## 8     8 No opinion - I ~ No opinion - I ~ No opinion - I ~ No opinion - I ~
## 9     9 No opinion - I ~ Average       Average        Good
## 10    10 No opinion - I ~ No opinion - I ~ No opinion - I ~ No opinion - I ~
## 11    11 Fair           Average       Good          No opinion - I ~
## 12    12 Excellent      No opinion - I ~ No opinion - I ~ No opinion - I ~
## # ... with 10 more variables: 'Top 10 Most Watched Netflix Shows in 2020 [Money
## #   Heist]' <chr>, 'Top 10 Most Watched Netflix Shows in 2020 [Dark
## #   Desire]' <chr>, 'Top 10 Most Watched Netflix Shows in 2020
## #   [Friends]' <chr>, 'Top 10 Most Watched Netflix Shows in 2020 [The
## #   Crown]' <chr>, 'Top 10 Most Watched Netflix Shows in 2020 [Ratched]' <chr>,
## #   'Top 10 Most Watched Netflix Shows in 2020 [Dark]' <chr>, 'Which TV and/or
## #   movie genres do you enjoy watching most?' <chr>, 'Which TV and/or movie

```

```

## #  genres do you enjoy watching least?' <chr>, 'On average, how many hours a
## #  week do you spend on Netflix each week?' <dbl>, 'What movie or TV show on
## #  Netflix or other streaming services would you highly recommend to adults
## #  that wasn't on this list?' <chr>

```

Tidy csv file

```

#rename columns

m_survey%>%dplyr::rename(Queens_Gambit="Top 10 Most Watched Netflix Shows in 2020 [The Queens Gambit]"
(m_survey)

## # A tibble: 12 x 15
##       Id Queens_Gambit Emily_in_Paris Lucifer The_Umbrella_Ac~ Money_Heist
##   <dbl> <chr>        <chr>      <chr>     <chr>      <chr>
## 1     1 Excellent    Poor       No opin~ No opinion - I ~ No opinion~
## 2     2 Excellent    Average    Good      No opinion - I ~ Good
## 3     3 Excellent    No opinion - ~ Good     Average      No opinion~
## 4     4 Good         No opinion - ~ Excell~ No opinion - I ~ No opinion~
## 5     5 No opinion  -- No opinion - ~ No opi~ No opinion - I ~ No opinion~
## 6     6 Poor         No opinion - ~ No opi~ No opinion - I ~ No opinion~
## 7     7 No opinion  -- No opinion - ~ No opi~ No opinion - I ~ No opinion~
## 8     8 No opinion  -- No opinion - ~ No opi~ No opinion - I ~ No opinion~
## 9     9 No opinion  -- Average    Average Good      No opinion~
## 10    10 No opinion -- No opinion - ~ No opi~ No opinion - I ~ No opinion~
## 11    11 Fair         Average    Good      No opinion - I ~ No opinion~
## 12    12 Excellent   No opinion - ~ No opi~ No opinion - I ~ No opinion~
## # ... with 9 more variables: Dark_Desire <chr>, Friends <chr>, The_Crown <chr>,
## #   Ratched <chr>, Dark <chr>, Genres_Liked <chr>, Genres_Disliked <chr>,
## #   Viewing_Hours <dbl>, Recommended <chr>

#


###Create table for viewer ratings

# remove select columns

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

# combine movies into single col and create col for review values

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

# replace category value in Rating col

m_rating<-m_rating%>%mutate(Rating=recode_factor(Rating, "No opinion - I haven't seen it" = '0', "Poor"=1, "Good"=2, "Excellent"=3))

# pivot back to tidy

m_rating%>%pivot_wider(names_from = Movies, values_from = Rating)%>% rename_all(make.names)

```

```

#write to new csv file

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.

### Create csv for viewer reviews

# remove select columns

m_hrs <- m_survey%>%select(c(Id, Viewing_Hours))
m_hrs%>% mutate(Id2 = Id)

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

```

Create csv for genres

```

#Note - genres not included in db, including code for later use

library(stringr)

split2<-str_split_fixed(m_survey$Genres_Liked, ", ", 4)%>%data.frame()
split2%>%dplyr::rename(First_Choice='X1', Second_Choice='X2', Third_Choice='X3', Fourth_Choice='X4')

# create an id column and relocate id column to front of table

split2%>% mutate(Id = row_number())%>%relocate(Id)

split2%>%pivot_longer(c(First_Choice, Second_Choice, Third_Choice, Fourth_Choice), values_to = 'Favorite_Genres')
  mutate(Favorite_Genres = na_if(Favorite_Genres, ""))
  
split2

## # A tibble: 48 x 2
##       Id Favorite_Genres
##   <int> <chr>
## 1     1 "Comedy"
## 2     1 "Drama"
## 3     1 "Action and Adventure"
## 4     1 <NA>
## 5     2 "Drama"
## 6     2 <NA>
## 7     2 <NA>
## 8     2 <NA>
## 9     3 "Comedy"
## 10    3 "Action and Adventure"
## # ... with 38 more rows

```

```

split3<-str_split_fixed(m_survey$Genres_Disliked, "", 4)%>%data.frame()

split3%>%dplyr::rename(First_Choice='X1', Second_Choice='X2', Third_Choice='X3', Fourth_Choice='X4')

split3%>% mutate(Id = row_number())%>%relocate(Id)

split3%>%pivot_longer(c(First_Choice, Second_Choice,Third_Choice,Fourth_Choice), values_to = 'Disliked')
  mutate(Disliked_Genres = na_if(Disliked_Genres, ""))
  
sp <- inner_join(split2, split3, by = 'Id')

```

Query SQL database

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

```

#Open connection to mysql

con <- dbConnect(RMariaDB::MariaDB(), user='root', password='Lupine20$', dbname='607hw2', host='localhost')

# List tables

dbListTables(con)

## [1] "viewer_ratings" "viewing_hours"

# query a table join

sql <- 'SELECT *
FROM viewer_ratings vr
LEFT JOIN viewing_hours vh
ON vr.Id = vh.Id2'
com_table <- dbGetQuery(con,sql)

com_table%>%dplyr::select(-c(Id..12,Id2))
com_table

```

	Id	Queens_Gambit	Emily_in_Paris	Lucifer	The_Umbrella_Academy	Money_Heist
## 1	1	5	1	0	0	0
## 2	2	5	3	4	0	4
## 3	3	5	0	4	3	0
## 4	4	4	0	5	0	0
## 5	5	0	0	0	0	0
## 6	6	1	0	0	0	0
## 7	7	0	0	0	0	0
## 8	8	0	0	0	0	0
## 9	9	0	3	3	4	0
## 10	10	0	0	0	0	0
## 11	11	2	3	4	0	0
## 12	12	5	0	0	0	0
##		Dark_Desire	Friends	The_Crown	Ratched	Dark
						Viewing_Hours

## 1	0	2	0	0	0	20
## 2	3	3	4	0	0	2
## 3	0	4	0	0	0	1
## 4	0	5	0	0	0	8
## 5	0	0	0	0	0	5
## 6	0	1	0	0	0	0
## 7	0	4	4	0	0	1
## 8	0	0	0	0	5	5
## 9	0	2	0	4	0	12
## 10	0	2	5	0	0	4
## 11	0	0	0	0	0	6
## 12	0	5	0	0	4	5