607 Extra Credit2 (Movie Rating) Seung Min Song

2022-09-20

Contents

Team member: Ted Kim
GitHub: https://github.com/seung-m1nsong/607

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

```
##    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

View movie list

```
##   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
```

```
##        reviewer                                       movie rating
## 1     Ben Brand                     Spider-Man: No Way Home      5
## 2     Ben Brand                      Thor: Love and Thunder      3
## 3     Ben Brand                           Top Gun: Maverick      3
## 4     Ben Brand                                     Morbius      2
## 5     Ben Brand                                     Toronto      1
## 6     Ben Brand Doctor Strange in the Multiverse of Madness      1
## 7    Samuel Kim                     Spider-Man: No Way Home      4
## 8    Samuel Kim                      Thor: Love and Thunder      3
## 9    Samuel Kim                           Top Gun: Maverick      3
## 10   Samuel Kim                                     Morbius      1
## 11   Samuel Kim                                     Toronto      1
## 12   Samuel Kim Doctor Strange in the Multiverse of Madness      1
## 13    Ray Watts                     Spider-Man: No Way Home      4
## 14    Ray Watts                      Thor: Love and Thunder      2
## 15    Ray Watts                           Top Gun: Maverick      5
## 16    Ray Watts                                     Morbius      2
## 17    Ray Watts Doctor Strange in the Multiverse of Madness      1
## 18 Kelly Kovack                     Spider-Man: No Way Home      4
## 19 Kelly Kovack                      Thor: Love and Thunder      3
## 20 Kelly Kovack                           Top Gun: Maverick      4
## 21 Kelly Kovack                                     Toronto      3
## 22 Reno Jimenez                     Spider-Man: No Way Home      4
## 23 Reno Jimenez                      Thor: Love and Thunder      1
## 24 Reno Jimenez                           Top Gun: Maverick      5
## 25 Reno Jimenez                                     Morbius      2
## 26 Reno Jimenez                                     Toronto      1
## 27 Reno Jimenez Doctor Strange in the Multiverse of Madness      1
## 28   Maria Riva                     Spider-Man: No Way Home      5
## 29   Maria Riva                      Thor: Love and Thunder      4
## 30   Maria Riva                           Top Gun: Maverick      4
## 31   Maria Riva                                     Toronto      2
## 32   Maria Riva Doctor Strange in the Multiverse of Madness      2
## 33 Daniel Hanry                     Spider-Man: No Way Home      4
## 34 Daniel Hanry                      Thor: Love and Thunder      3
## 35 Daniel Hanry                           Top Gun: Maverick      5
## 36 Daniel Hanry                                     Morbius      1
## 37 Daniel Hanry Doctor Strange in the Multiverse of Madness      2
## 38  Raymund Suh                     Spider-Man: No Way Home      5
## 39  Raymund Suh                      Thor: Love and Thunder      3
## 40  Raymund Suh                           Top Gun: Maverick      5
## 41  Raymund Suh Doctor Strange in the Multiverse of Madness      1
```

View all movie reviewer

```
##   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'

```
##                                    movietitle avg_rating min_rating max_rating
## 1 Doctor Strange in the Multiverse of Madness     1.2857          1          2
## 2                                     Morbius     1.6000          1          2
## 3                     Spider-Man: No Way Home     4.3750          4          5
## 4                      Thor: Love and Thunder     2.7500          1          4
## 5                           Top Gun: Maverick     4.2500          3          5
## 6                                     Toronto     1.6000          1          3
```

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.

About missing record

```
#step 0: insert the old survey data into a new data frame for easier viewing.

df <- dfMovierating

# There is no record for the review with no rating in the table because Null values are not stored in t

#step 1. Use the pivot_wider() function to create a data frame identical to Excel format. missing recor

pv_wider <- df %>%
  pivot_wider(
    names_from = movie,
    values_from = rating
  )
print(pv_wider)
```

```
## # A tibble: 8 x 7
##   reviewer     ‘Spider-Man: No Way Home‘ Thor:~1 Top G~2 Morbius Toronto Docto~3
##   <chr>                            <int>   <int>   <int>   <int>   <int>   <int>
## 1 Ben Brand                            5       3       3       2       1       1
## 2 Samuel Kim                           4       3       3       1       1       1
## 3 Ray Watts                            4       2       5       2      NA       1
## 4 Kelly Kovack                         4       3       4      NA       3      NA
## 5 Reno Jimenez                         4       1       5       2       1       1
## 6 Maria Riva                           5       4       4      NA       2       2
## 7 Daniel Hanry                         4       3       5       1      NA       2
## 8 Raymund Suh                          5       3       5      NA      NA       1
## # ... with abbreviated variable names 1: ‘Thor: Love and Thunder‘,
## #   2: ‘Top Gun: Maverick‘, 3: ‘Doctor Strange in the Multiverse of Madness‘
```

```
#step 2: if change the data frame created in step 1 to its original form using the pivot_longer() funct

pv_longer <- pv_wider %>%
              pivot_longer(
                cols = colnames(pv_wider)[2:7],
                names_to = 'movie',
                values_to = 'rating'
              )

print(pv_longer)
```

```
## # A tibble: 48 x 3
##    reviewer   movie                                       rating
##    <chr>      <chr>                                        <int>
##  1 Ben Brand  Spider-Man: No Way Home                          5
##  2 Ben Brand  Thor: Love and Thunder                           3
##  3 Ben Brand  Top Gun: Maverick                                3
##  4 Ben Brand  Morbius                                          2
##  5 Ben Brand  Toronto                                          1
##  6 Ben Brand  Doctor Strange in the Multiverse of Madness      1
##  7 Samuel Kim Spider-Man: No Way Home                          4
##  8 Samuel Kim Thor: Love and Thunder                           3
##  9 Samuel Kim Top Gun: Maverick                                3
## 10 Samuel Kim Morbius                                          1
## # ... with 38 more rows
```

```
#step 3: use is.na()function to exclude NA value. After that calculate movie_avg for each movie. (ident

movie_avg <- pv_longer %>%
                filter(!is.na(rating)) %>%
                group_by(movie) %>%
                summarise(movie_avg = mean(rating))
print(movie_avg)
```

```
## # A tibble: 6 x 2
##   movie                                      movie_avg
##   <chr>                                          <dbl>
## 1 Doctor Strange in the Multiverse of Madness      1.29
## 2 Morbius                                          1.6
## 3 Spider-Man: No Way Home                          4.38
## 4 Thor: Love and Thunder                           2.75
## 5 Top Gun: Maverick                                4.25
## 6 Toronto                                          1.6
```

```
#step 4: use is.na() function to exclude NA value. After that calculate movie_mean.

movie_mean <- mean((pv_longer %>%
                      filter(!is.na(rating)))$rating)

print(movie_mean)
```

```
## [1] 2.829268
```

```
#step 5: Using mutate, add column named sub_avg_mean in the movie_avg created in step 3 and insert movi
movie_compute <- movie_avg %>%
                              mutate(subs_avg_mean = movie_avg - movie_mean)

print(movie_compute)
```

```
## # A tibble: 6 x 3
##   movie                                      movie_avg subs_avg_mean
##   <chr>                                          <dbl>         <dbl>
## 1 Doctor Strange in the Multiverse of Madness      1.29       -1.54
## 2 Morbius                                          1.6        -1.23
## 3 Spider-Man: No Way Home                          4.38        1.55
## 4 Thor: Love and Thunder                           2.75       -0.0793
## 5 Top Gun: Maverick                                4.25        1.42
## 6 Toronto                                          1.6        -1.23
```

```
#step 6: use is.na()function to exclude NA value. After that calculate each person's user_avg and user_

user_compute <- pv_longer %>%
                  filter(!is.na(rating)) %>%
                  group_by(reviewer) %>%
                  mutate(
                    user_avg = mean(rating),
                    sub_user_avg_mean_movie = mean(rating) - movie_mean
                  )

print(user_compute)
```

```
## # A tibble: 41 x 5
## # Groups:   reviewer [8]
##    reviewer   movie                                        rating user_~1 sub_u~2
##    <chr>      <chr>                                         <int>   <dbl>   <dbl>
##  1 Ben Brand  Spider-Man: No Way Home                           5     2.5  -0.329
##  2 Ben Brand  Thor: Love and Thunder                           3    2.5   -0.329
##  3 Ben Brand  Top Gun: Maverick                                3    2.5   -0.329
##  4 Ben Brand  Morbius                                          2    2.5   -0.329
##  5 Ben Brand  Toronto                                          1    2.5   -0.329
##  6 Ben Brand  Doctor Strange in the Multiverse of Madness      1    2.5   -0.329
##  7 Samuel Kim Spider-Man: No Way Home                          4    2.17  -0.663
##  8 Samuel Kim Thor: Love and Thunder                           3    2.17  -0.663
##  9 Samuel Kim Top Gun: Maverick                                3    2.17  -0.663
## 10 Samuel Kim Morbius                                          1    2.17  -0.663
## # ... with 31 more rows, and abbreviated variable names 1: user_avg,
## #   2: sub_user_avg_mean_movie
```

```
#step 7: use is.na()function to bring record with NA value and merge(m1) with the movie_compute crated


m1 <- merge(pv_longer[is.na(pv_longer$rating),], movie_compute)
print(m1)
```

```
##                                          movie     reviewer rating movie_avg
## 1 Doctor Strange in the Multiverse of Madness Kelly Kovack     NA  1.285714
## 2                                     Morbius Kelly Kovack     NA  1.600000
## 3                                     Morbius   Maria Riva     NA  1.600000
## 4                                     Morbius  Raymund Suh     NA  1.600000
## 5                                     Toronto    Ray Watts     NA  1.600000
## 6                                     Toronto Daniel Hanry     NA  1.600000
## 7                                     Toronto  Raymund Suh     NA  1.600000
##   subs_avg_mean
## 1     -1.543554
## 2     -1.229268
## 3     -1.229268
## 4     -1.229268
## 5     -1.229268
## 6     -1.229268
## 7     -1.229268
```

```
m2 <- merge(m1,
            unique(user_compute[,c('reviewer', 'user_avg', 'sub_user_avg_mean_movie')]),
            by.x=c('reviewer'),
            by.y=c('reviewer')) %>%
      mutate(rating = round(movie_mean + subs_avg_mean + sub_user_avg_mean_movie, 0)) %>%
      select('reviewer', 'movie', 'rating', 'user_avg', 'sub_user_avg_mean_movie')
print(m2)
```

```
##       reviewer                                       movie rating user_avg
## 1 Daniel Hanry                                     Toronto      2      3.0
## 2 Kelly Kovack Doctor Strange in the Multiverse of Madness      2      3.5
## 3 Kelly Kovack                                     Morbius      2      3.5
## 4   Maria Riva                                     Morbius      2      3.4
## 5    Ray Watts                                     Toronto      2      2.8
## 6  Raymund Suh                                     Morbius      2      3.5
## 7  Raymund Suh                                     Toronto      2      3.5
##   sub_user_avg_mean_movie
## 1              0.17073171
## 2              0.67073171
## 3              0.67073171
## 4              0.57073171
## 5             -0.02926829
## 6              0.67073171
## 7              0.67073171
```

```
#step 8: combine m2 created in step 7 and user_compute with no NA record using the union() function.
final <- union(user_compute, m2)

print(final)
```

```
## # A tibble: 48 x 5
## # Groups:   reviewer [8]
##    reviewer   movie                                       rating user_~1 sub_u~2
##    <chr>      <chr>                                        <dbl>   <dbl>   <dbl>
##  1 Ben Brand  Spider-Man: No Way Home                          5    2.5   -0.329
##  2 Ben Brand  Thor: Love and Thunder                           3    2.5   -0.329
##  3 Ben Brand  Top Gun: Maverick                                3    2.5   -0.329
##  4 Ben Brand  Morbius                                          2    2.5   -0.329
##  5 Ben Brand  Toronto                                          1    2.5   -0.329
##  6 Ben Brand  Doctor Strange in the Multiverse of Madness      1    2.5   -0.329
##  7 Samuel Kim Spider-Man: No Way Home                          4    2.17  -0.663
##  8 Samuel Kim Thor: Love and Thunder                           3    2.17  -0.663
##  9 Samuel Kim Top Gun: Maverick                                3    2.17  -0.663
## 10 Samuel Kim Morbius                                          1    2.17  -0.663
## # ... with 38 more rows, and abbreviated variable names 1: user_avg,
## #   2: sub_user_avg_mean_movie
```

```
#step 9: use pivot_wider() function to make the data frame created in step 8 identical to Excel format

final %>%
  pivot_wider(
    names_from = movie,
    values_from = rating
  ) %>%
  select(1, 4:9, 2, 3)
```

```
## # A tibble: 8 x 9
## # Groups:   reviewer [8]
##   reviewer    Spider--~1 Thor:~2 Top G~3 Morbius Toronto Docto~4 user_~5 sub_u~6
##   <chr>            <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Ben Brand            5       3       3       2       1       1    2.5  -0.329
## 2 Samuel Kim           4       3       3       1       1       1    2.17 -0.663
## 3 Ray Watts            4       2       5       2       2       1    2.8  -0.0293
## 4 Kelly Kovack         4       3       4       2       3       2    3.5   0.671
## 5 Reno Jimenez         4       1       5       2       1       1    2.33 -0.496
## 6 Maria Riva           5       4       4       2       2       2    3.4   0.571
## 7 Daniel Hanry         4       3       5       1       2       2    3     0.171
## 8 Raymund Suh          5       3       5       2       2       1    3.5   0.671
## # ... with abbreviated variable names 1: ‘Spider-Man: No Way Home‘,
## #   2: ‘Thor: Love and Thunder‘, 3: ‘Top Gun: Maverick‘,
## #   4: ‘Doctor Strange in the Multiverse of Madness‘, 5: user_avg,
## #   6: sub_user_avg_mean_movie
```

```
# mean_movie
movie_mean
```

```
## [1] 2.829268
```

```
# movie_avg
merge(movie_compute, dfMovierating) %>%
  select('movie', 'movie_avg') %>%
  pivot_wider(
    names_from = movie,
    values_from = movie_avg
  )
```

```
## # A tibble: 1 x 6
##   Doctor Strange in the Multiverse of ~1 Morbius Spide~2 Thor:~3 Top G~4 Toronto
##   <list>                                 <list>  <list>  <list>  <list>  <list>
## 1 <dbl [7]>                              <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## # ... with abbreviated variable names
## #   1: ‘Doctor Strange in the Multiverse of Madness‘,
## #   2: ‘Spider-Man: No Way Home‘, 3: ‘Thor: Love and Thunder‘,
## #   4: ‘Top Gun: Maverick‘
```

```
# avg-mean
merge(movie_compute, dfMovierating) %>%
  select('movie', 'subs_avg_mean') %>%
  pivot_wider(
    names_from = movie,
    values_from = subs_avg_mean
  )
```

```
## # A tibble: 1 x 6
##   Doctor Strange in the Multiverse of ~1 Morbius Spide~2 Thor:~3 Top G~4 Toronto
##   <list>                                 <list>  <list>  <list>  <list>  <list>
## 1 <dbl [7]>                              <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## # ... with abbreviated variable names
## #   1: ‘Doctor Strange in the Multiverse of Madness‘,
## #   2: ‘Spider-Man: No Way Home‘, 3: ‘Thor: Love and Thunder‘,
## #   4: ‘Top Gun: Maverick‘
```