MichaelY-DATA607-Week02-Movies

Michael Y.

September 8, 2019

Contents

Assignment 2: Movies Database

Already done (using MySQL Workbench):

Choose six recent popular movies.

Ask at least five people that you know (friends, family, classmates, imaginary friends) to rate each of these movies that they have seen on a scale of 1 to 5.

Take the results (observations) and store them in a SQL database. (See attached SQL script for database creation and loading.)

To be done here:

Load the information into an R dataframe, and examine it.

Load up some libraries

```
library(RMySQL)
```

```
## Loading required package: DBI
```

```
library(dplyr)
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

Connect to the MySql database and retrieve the data:

```
# I created "stduser" as a read-only account in my database which only has "select" privilege
connstd <- dbConnect(MySQL(), user="stduser", password="password",
                     dbname="Week2_Movies", host="localhost")

# Create a query which joins the 3 database tables,
# replacing the auto-generated ID codes with the movie names and the reviewers' names

query <- 'Select M.Movie_title, F.Friend_name, R.Rating
          From Movies as M, Friends as F, Ratings as R
          Where (M.Movie_id = R.Movie_ID AND F.Friend_id = R.Friend_ID);'

# Execute the query
result <- dbGetQuery(connstd, query)

# Close the database connection
discard <- dbDisconnect(connstd) # this function returns "TRUE", so assignment suppresses printing
```

The dimensions of the results dataframe are 30, 3 .

```
# structure of the results dataframe
str(result)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ Movie_title: chr  "Aladdin (2019)" "Aladdin (2019)" "Aladdin (2019)" "Aladdin (2019)" ...
##  $ Friend_name: chr  "Andrew" "Bernard" "Charlie" "Dilbert" ...
##  $ Rating     : int  1 5 1 2 4 4 1 3 2 1 ...
```

```
# summary of the results dataframe
summary(result)
```

```
##  Movie_title        Friend_name           Rating
##  Length:30          Length:30          Min.   :1.00
##  Class :character   Class :character   1st Qu.:1.00
##  Mode  :character   Mode  :character   Median :2.00
##                                        Mean   :2.52
##                                        3rd Qu.:4.00
##                                        Max.   :5.00
##                                        NA's   :1
```

List the results (there are only 30 rows):

```
result
```

```
##                  Movie_title Friend_name Rating
## 1             Aladdin (2019)      Andrew      1
## 2             Aladdin (2019)     Bernard      5
## 3             Aladdin (2019)     Charlie      1
## 4             Aladdin (2019)     Dilbert      2
## 5             Aladdin (2019)     Ernesto      4
## 6          Avengers: Endgame      Andrew      4
## 7          Avengers: Endgame     Bernard      1
## 8          Avengers: Endgame     Charlie      3
## 9          Avengers: Endgame     Dilbert      2
```

```
## 10         Avengers: Endgame     Ernesto      1
## 11            Captain Marvel      Andrew      4
## 12            Captain Marvel     Bernard      5
## 13            Captain Marvel     Charlie      5
## 14            Captain Marvel     Dilbert      2
## 15            Captain Marvel     Ernesto      4
## 16 Spider-Man: Far from Home      Andrew      2
## 17 Spider-Man: Far from Home     Bernard      1
## 18 Spider-Man: Far from Home     Charlie      1
## 19 Spider-Man: Far from Home     Dilbert      2
## 20 Spider-Man: Far from Home     Ernesto      2
## 21      The Lion King (2019)      Andrew      2
## 22      The Lion King (2019)     Bernard      1
## 23      The Lion King (2019)     Charlie      2
## 24      The Lion King (2019)     Dilbert      5
## 25      The Lion King (2019)     Ernesto      3
## 26               Toy Story 4      Andrew      3
## 27               Toy Story 4     Bernard      2
## 28               Toy Story 4     Charlie      2
## 29               Toy Story 4     Dilbert      1
## 30               Toy Story 4     Ernesto     NA
```

Describe the results:

```
describe(result$Rating)  %>% kable() %>%  kable_styling(c("striped", "bordered"))
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	29	2.51724	1.4046	2	2.44	1.4826	1	5	4	0.563286	-1.09891	0.260828

(Note that there is one "NA" value, which we will have to exclude later.)

Let's look at the results, grouped by Movie :

```
describeBy(result$Rating,group = result$Movie_title )
```

```
##
##  Descriptive statistics by group
## group: Aladdin (2019)
##    vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 5  2.6 1.82      2     2.6 1.48   1   5     4 0.27    -2.08 0.81
## ------------------------------------------------------------------------------------------
## group: Avengers: Endgame
##    vars n mean  sd median trimmed  mad min max range skew kurtosis   se
## X1    1 5  2.2 1.3      2     2.2 1.48   1   4     3 0.26    -1.96 0.58
## ------------------------------------------------------------------------------------------
## group: Captain Marvel
##    vars n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 5    4 1.22      4       4 1.48   2   5     3 -0.65     -1.4 0.55
## ------------------------------------------------------------------------------------------
## group: Spider-Man: Far from Home
##    vars n mean   sd median trimmed mad min max range  skew kurtosis   se
## X1    1 5  1.6 0.55      2     1.6   0   1   2     1 -0.29    -2.25 0.24
## ------------------------------------------------------------------------------------------
## group: The Lion King (2019)
##    vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 5  2.6 1.52      2     2.6 1.48   1   5     4 0.54    -1.49 0.68
## ------------------------------------------------------------------------------------------
## group: Toy Story 4
##    vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 4    2 0.82      2       2 0.74   1   3     2    0    -1.88 0.41
```

We need to drop the item with the NA rating in order to obtain non-NA summary results.

Subsetting using !is.na(result$Rating) :

```
result[!is.na(result$Rating),] %>%
  group_by(Movie_title) %>%
  summarize(count=n(),
            min=min(Rating),
            mean=mean(Rating),
            median=median(Rating),
            max=max(Rating),
            sd=sd(Rating),
```

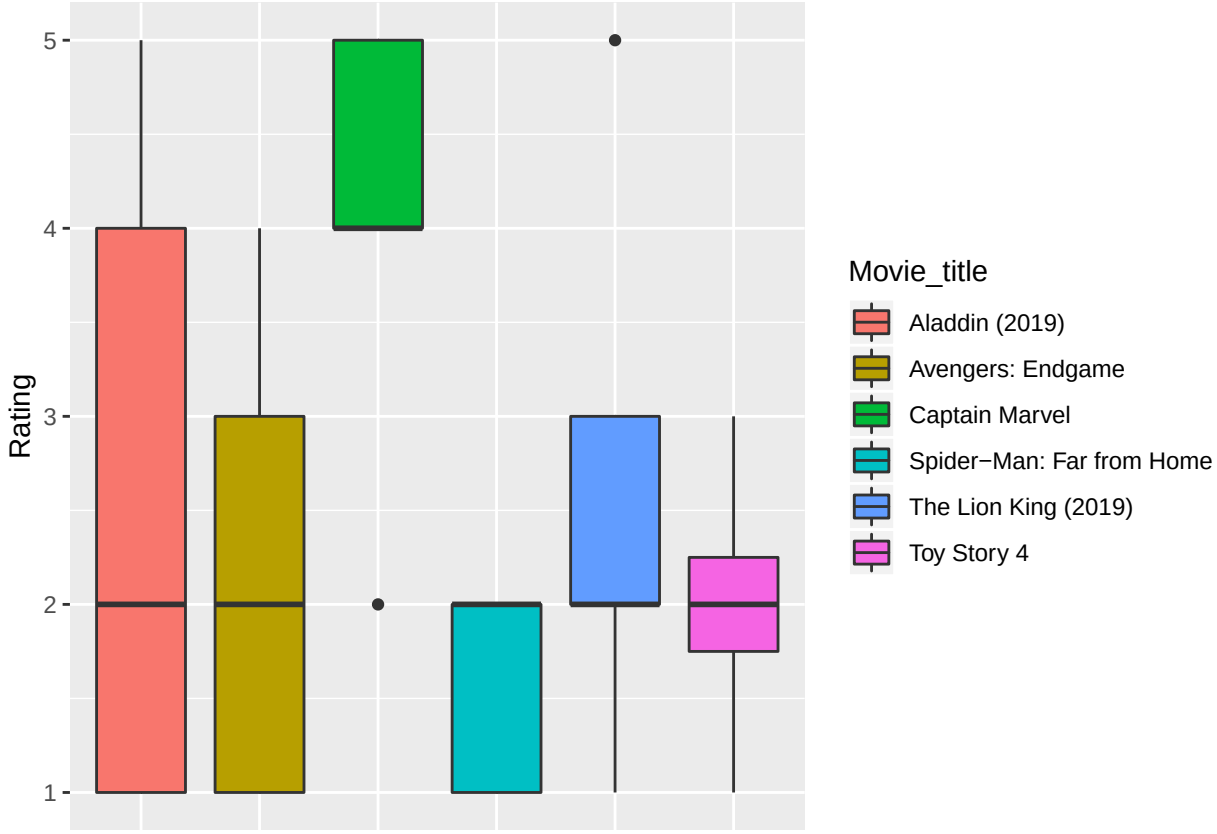
```
            IQR=IQR(Rating)
            )
```

```
## # A tibble: 6 x 8
##   Movie_title               count   min  mean median   max    sd   IQR
##   <chr>                     <int> <int> <dbl>  <dbl> <int> <dbl> <dbl>
## 1 Aladdin (2019)                5     1   2.6      2     5 1.82    3
## 2 Avengers: Endgame             5     1   2.2      2     4 1.30    2
## 3 Captain Marvel                5     2   4        4     5 1.22    1
## 4 Spider-Man: Far from Home     5     1   1.6      2     2 0.548   1
## 5 The Lion King (2019)          5     1   2.6      2     5 1.52    1
## 6 Toy Story 4                   4     1   2        2     3 0.816   0.5
```

Now, let's make a boxplot by Movie :

```
ggplot(result, aes(x=Movie_title, y=Rating, fill=Movie_title)) +
  geom_boxplot() +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



We can see that Captain Marvel was quite popular, with mean and median ratings of 4:

```
result[result$Movie_title=="Captain Marvel",]
```

```
##       Movie_title Friend_name Rating
## 11 Captain Marvel      Andrew      4
## 12 Captain Marvel     Bernard      5
## 13 Captain Marvel     Charlie      5
## 14 Captain Marvel     Dilbert      2
## 15 Captain Marvel     Ernesto      4
```

while Spider-Man: Far from Home was at the opposite end of the spectrum, receiving the lowest ratings:

```
result[result$Movie_title=="Spider-Man: Far from Home",]
```

```
##                  Movie_title Friend_name Rating
## 16 Spider-Man: Far from Home      Andrew      2
## 17 Spider-Man: Far from Home     Bernard      1
## 18 Spider-Man: Far from Home     Charlie      1
## 19 Spider-Man: Far from Home     Dilbert      2
## 20 Spider-Man: Far from Home     Ernesto      2
```

Now, Let's look at how each friend tended to rate the films:

```
describeBy(result$Rating,group = result$Friend_name )
```

```
##
##  Descriptive statistics by group
## group: Andrew
##    vars n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 6 2.67 1.21    2.5    2.67 1.48   1   4     3 -0.04    -1.88 0.49
## ------------------------------------------------------------------------------------------
## group: Bernard
##    vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 6  2.5 1.97    1.5     2.5 0.74   1   5     4 0.45    -1.98 0.81
## ------------------------------------------------------------------------------------------
## group: Charlie
##    vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 6 2.33 1.51      2    2.33 1.48   1   5     4 0.71    -1.15 0.61
## ------------------------------------------------------------------------------------------
## group: Dilbert
##    vars n mean   sd median trimmed mad min max range skew kurtosis   se
## X1    1 6 2.33 1.37      2    2.33   0   1   5     4 1.07    -0.43 0.56
## ------------------------------------------------------------------------------------------
## group: Ernesto
##    vars n mean  sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 5  2.8 1.3      3     2.8 1.48   1   4     3 -0.26    -1.96 0.58
```

Again, we have to exclude the item with the NA:

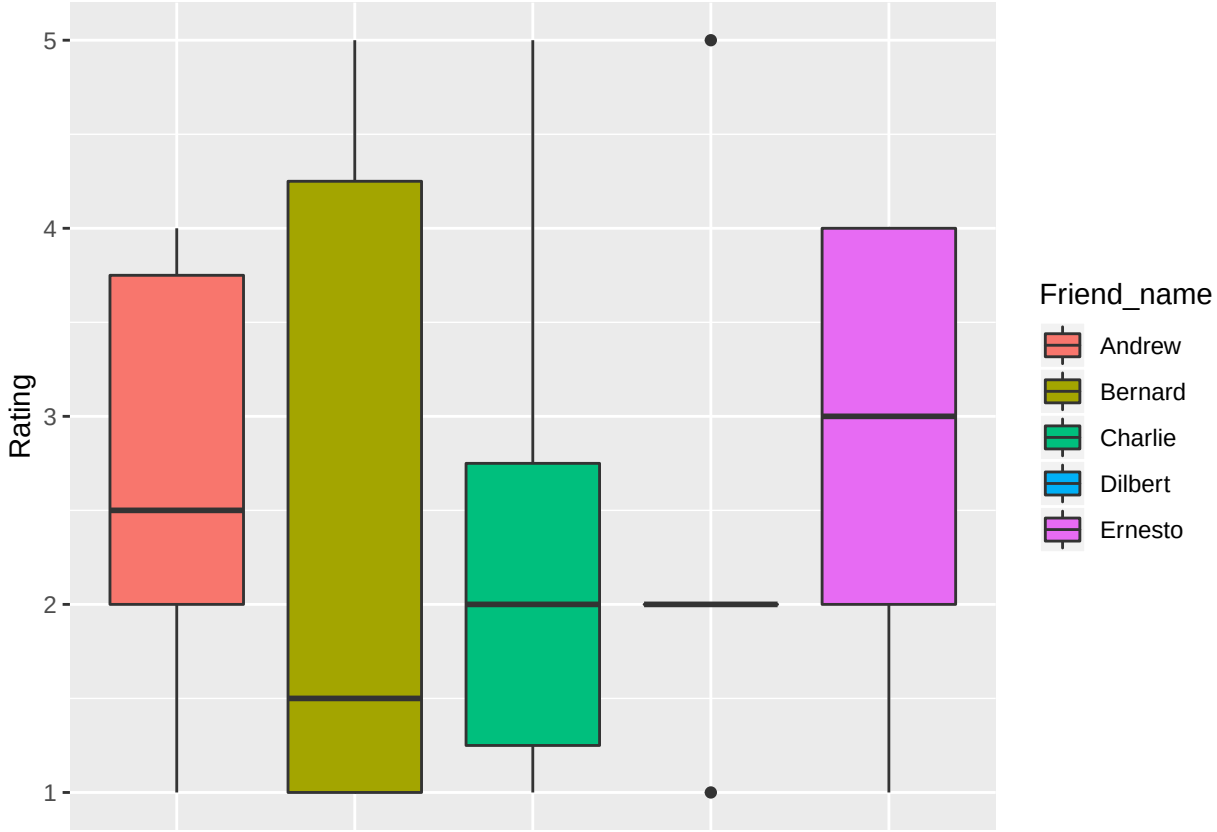
```
result[!is.na(result$Rating),] %>%
  group_by(Friend_name) %>%
  summarize(count=n(),
            min=min(Rating),
            mean=mean(Rating),
            median=median(Rating),
            max=max(Rating),
            sd=sd(Rating),
            IQR=IQR(Rating)
            )
```

```
## # A tibble: 5 x 8
##   Friend_name count   min  mean median   max    sd   IQR
##   <chr>       <int> <int> <dbl>  <dbl> <int> <dbl> <dbl>
## 1 Andrew          6     1  2.67    2.5     4  1.21  1.75
## 2 Bernard         6     1  2.5     1.5     5  1.97  3.25
## 3 Charlie         6     1  2.33    2       5  1.51  1.5
## 4 Dilbert         6     1  2.33    2       5  1.37  0
## 5 Ernesto         5     1  2.8     3       4  1.30  2
```

Boxplot by friend:

```
ggplot(result, aes(x=Friend_name, y=Rating, fill=Friend_name)) +
  geom_boxplot() +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



We observe that Bernard either likes a film or hates it – with Bernard, there is no middle ground.

Bernard gave the widest disperion among his ratings, using mostly "1"s and "5"s, which explains his large IQR and standard deviation.

His Median is the lowest, as half his ratings were "1"s:

```
result[result$Friend_name=="Bernard",]
```

```
##                 Movie_title Friend_name Rating
## 2            Aladdin (2019)     Bernard      5
## 7         Avengers: Endgame     Bernard      1
## 12           Captain Marvel     Bernard      5
```

```
## 17 Spider-Man: Far from Home     Bernard      1
## 22      The Lion King (2019)     Bernard      1
## 27                Toy Story 4     Bernard      2
```

Because Dilbert gave so many ratings of "2", his IQR = 0, thus his box is flat, with outliers at "1" and "5" :

```
result[result$Friend_name=="Dilbert",]
```

```
##                  Movie_title Friend_name Rating
## 4              Aladdin (2019)     Dilbert      2
## 9           Avengers: Endgame     Dilbert      2
## 14             Captain Marvel     Dilbert      2
## 19 Spider-Man: Far from Home     Dilbert      2
## 24      The Lion King (2019)     Dilbert      5
## 29                Toy Story 4     Dilbert      1
```

Conclusion:

With a small data set (6 movies and 5 reviewers) the aggregated figures display interesting results across both movie and reviewer.

It would be interesting to see the results across a larger sample, for example using the data assembled by "Rotton Tomatoes" which tabulates published movie reviews and scores films on a scale of 0%-100% based upon the percentage of reviews which are favorable vs. unfavorable.

Furthermore, it would be interesting to compare/contrast such "professional" assessments with opinions from individuals, such as those assembled by firms like Amazon.