DATA 607 Assignment #2

Anthony Conrardy

2024-02-04

Overview

This is Assignment #2 for DATA 607. For this particular assignment I chose to create some of the data that will be used in this project by randomly assigning values for the movie ratings, as well as missing data, for ten (10) participants. The data set is available on GitHub at:

https://github.com/Aconrard/DATA607/blob/main/Assignment2/movies_ratings.csv

The database for this assignment is located on Azure in an accessible database. The next code chunk with access the database from a server, and then we will access the required table from database for further analysis. This also requires the person using this Markdown to know the external database log on credentials. This will be provided in the comment sections of the submission on Blackboard.

```
username <- rstudioapi::askForPassword("Database username")
passworda <- rstudioapi::askForPassword("Database password")
```

```
mydb <- dbConnect(MySQL(), user = username, password = passworda, dbname = 'movie_ratings', host = "data
```

```
mrs <- dbReadTable(mydb, "movies_ratings")

mrs
```

```
##    person aquaman equalizer3 dune_part_2 wonka pawpatrol haunted_mansion tmnt
## 1       1       5          5           5     -         -               -    -
## 2       2       -          -           4     -         -               -    5
## 3       3       4          3           -     2         4               5    3
## 4       4       2          5           3     1         4               -    -
## 5       5       2          4           1     -         3               4    2
## 6       6       3          1           4     -         2               3    -
## 7       7       3          3           3     3         1               -    -
## 8       8       5          1           2     4         5               3    2
## 9       9       4          5           1     3         1               5    5
## 10     10       1          5           -     3         -               4    1
##    marvels
## 1        5
## 2        5
## 3        1
## 4        4
## 5        3
## 6        3
## 7        2
```

```
## 8        2
## 9        5
## 10       1
```

Missing Data

You will notice that some of the rows are missing data, which is logical since some people did not see all seven(7) movies. Using a form survey limits the values associated with a Likert Scale question, and those that did not get rated are simply given a default value. However, that default value may vary from platform to platform, so we must be know hoe the values are selected. In this case, the missing data is denoted by "-", but that is something that R does not recognize. Therefore, we must alter those entries into something recognizable as missing in R.

There may be other characters that are not recognizable, or that need to be changed. However, this example has only one and we are able to replace all values of "-" with NA with the following code chunk.

```
mrs[mrs == '-'] <- 'NA'
mrs
```

```
##    person aquaman equalizer3 dune_part_2 wonka pawpatrol haunted_mansion tmnt
## 1       1       5          5           5    NA        NA               NA   NA
## 2       2      NA         NA           4    NA        NA               NA    5
## 3       3       4          3          NA     2         4                5    3
## 4       4       2          5           3     1         4               NA   NA
## 5       5       2          4           1    NA         3                4    2
## 6       6       3          1           4    NA         2                3   NA
## 7       7       3          3           3     3         1               NA   NA
## 8       8       5          1           2     4         5                3    2
## 9       9       4          5           1     3         1                5    5
## 10     10       1          5          NA     3        NA                4    1
##    marvels
## 1        5
## 2        5
## 3        1
## 4        4
## 5        3
## 6        3
## 7        2
## 8        2
## 9        5
## 10       1
```

Reviewing Data

A look at the data after we handle the missing entries shows that while some of the movies have very few missing ratings, others have significantly more. So we should conduct a summary of the ratings for each movie and see what happens.

```
summary(mrs)
```

```
##      person        aquaman           equalizer3        dune_part_2
##  Min.   : 1.00   Length:10          Length:10          Length:10
```

```
##  1st Qu.: 3.25   Class :character   Class :character   Class :character
##  Median : 5.50   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 5.50
##  3rd Qu.: 7.75
##  Max.   :10.00
##     wonka            pawpatrol         haunted_mansion        tmnt
##  Length:10          Length:10          Length:10          Length:10
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     marvels
##  Min.   :1.00
##  1st Qu.:2.00
##  Median :3.00
##  Mean   :3.10
##  3rd Qu.:4.75
##  Max.   :5.00
```

We now note that the data type for the variables are listed as characters making an analysis of the ratings difficult for most of the movies. The Marvels has a proper summary because it did not have any missing values and was identified as numeric. We also note that the person variable was identified as a numeric type and was analyzed as such. However, the others need to be adjusted accordingly. We can do that with the following code chunk and then run the summary again.

```
mrs$person <- as.character(mrs$person)
mrs$aquaman <- as.numeric(mrs$aquaman)
mrs$equalizer3 <- as.numeric(mrs$equalizer3)
mrs$dune_part_2 <- as.numeric(mrs$dune_part_2)
mrs$wonka <- as.numeric(mrs$wonka)
mrs$pawpatrol <- as.numeric(mrs$pawpatrol)
mrs$haunted_mansion <- as.numeric(mrs$haunted_mansion)
mrs$tmnt <- as.numeric(mrs$tmnt)
mrs$marvels <- as.numeric(mrs$marvels)
summary(mrs, na.rm = TRUE)
```

```
##     person             aquaman        equalizer3     dune_part_2
##  Length:10          Min.   :1.000   Min.   :1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:1.750
##  Mode  :character   Median :3.000   Median :4.000   Median :3.000
##                     Mean   :3.222   Mean   :3.556   Mean   :2.875
##                     3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
##                     Max.   :5.000   Max.   :5.000   Max.   :5.000
##                     NA's   :1       NA's   :1       NA's   :2
##      wonka         pawpatrol     haunted_mansion      tmnt        marvels
##  Min.   :1.000   Min.   :1.000   Min.   :3.00    Min.   :1.0   Min.   :1.00
##  1st Qu.:2.250   1st Qu.:1.500   1st Qu.:3.25    1st Qu.:2.0   1st Qu.:2.00
##  Median :3.000   Median :3.000   Median :4.00    Median :2.5   Median :3.00
##  Mean   :2.667   Mean   :2.857   Mean   :4.00    Mean   :3.0   Mean   :3.10
##  3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.75    3rd Qu.:4.5   3rd Qu.:4.75
##  Max.   :4.000   Max.   :5.000   Max.   :5.00    Max.   :5.0   Max.   :5.00
##  NA's   :4       NA's   :3       NA's   :4       NA's   :4
```

Conclusion

While this was an exercise of some of the most basic functions, we were able to access a database from an external Azure connection through R, modify the data set to address missing elements, modify the data types to allow for a summary of the various ratings, and secured the password and log on credentials necessary to access the external database.