SQL and R

Zachary Palmore

9/5/2020

Assignment 2

For Assignment 2 of DATA607.

Background

An survey was given to six close friends and family. Each of them were asked to rank the movies selected on a scale of 1 - 5 based on how much they enjoyed it. A score of 1 indicates a bad movie, something they would not watch again and thought was a complete waste of time. A score of 5 indicated that they would definitely watch it again and loved the movie.

Method

The data were collected with Google forms then written into the database in SQL. This can be found and reproduced by extracting the information needed from the script located at this repository:

[GitHub Repository]

For movies that had not been seen by those same friends and family they were asked to skip it, leaving the score blank.

Connecting to a Database

```
library(RMariaDB)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(RSQLite)
```

For purposes of this assignment, I will connect to the database via the RMariaDB package written by MySQL's creators. This local password will also be changed before posting.

```
localuserpassword <- "Eight80"
movies_db <- dbConnect(RMariaDB::MariaDB(), user='root', password=localuserpassword, dbname='movies', h
# Lists the tables in the database
dbListTables(movies_db)
```

```
## [1] "details" "ratings"
```

Loading to a Dataframe

To create a data frame from the database with only the information needed we write in SQL.

```
ratings <- dbGetQuery(movies_db, "SELECT * FROM ratings")
glimpse(ratings)
```

```
## Rows: 30
## Columns: 4
## $ ID        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17...
## $ Score     <int> 4, 4, 2, 3, 4, 4, NA, 3, NA, NA, 5, 5, NA, NA, NA, NA, NA...
## $ Critic    <chr> "Melanie", "Melanie", "Melanie", "Melanie", "Melanie", "M...
## $ MovieName <chr> "Spiderman Far From Home", "Dolittle", "Irresistible", "T...
```

There are 30 rows for each of 4 columns. The variables include an *ID*, *Score*, the *Critic*'s name, and the name of the movie as *MovieName*.

However, there are a few missing data points that need to be dealt with. For example, the only person who has seen all movies is Melanie. All other critics have not seen at least one of the movies.

Handling missing data

There are 12 missing values in this data set.

```
sum(is.na(ratings))
```

```
## [1] 12
```

Where did they come from?

Those who have not seen all the movies and were therefore unable to rank their experience will have their case removed from the data set. This was complete with the *na.omit* function and the result was reassigned to the same data frame.

```
ratings <- na.omit(ratings)
glimpse(ratings)
```

```
## Rows: 18
## Columns: 4
## $ ID        <int> 1, 2, 3, 4, 5, 6, 8, 11, 12, 18, 20, 21, 23, 25, 26, 27, ...
## $ Score     <int> 4, 4, 2, 3, 4, 4, 3, 5, 5, 4, 4, 5, 3, 5, 1, 3, 5, 5
## $ Critic    <chr> "Melanie", "Melanie", "Melanie", "Melanie", "Melanie", "M...
## $ MovieName <chr> "Spiderman Far From Home", "Dolittle", "Irresistible", "T...
```

This leaves us with 18 rows of data for each of the same 4 column variables, *ID*, *Score*, *Critic*, and *MovieName*. With incomplete cases removed, the data is now clean and ready for analysis.

Discussion

Is there any benefit in standardizing ratings? How might you approach this?

There is a great benefit to standardizing ratings in survey research like this. The standard allows a researcher to reproduce the work of another and thereby validate or disprove existing knowledge. Standardization when accompanied by repeat iterations of the experiment, produces confidence in the collected results of all researchers performing the experiment.

To standardize one must generate and administer the survey questions in a way that does not unintentionally influence the results of the survey. For example, the researcher should ask all questions in the same manner without variations. It is also best to stick to one method of recording information and to avoid using language that may be perceived as unclear.
