

607_homework_2

Alec

9/4/2021

load RMySQL

RMySQL package allows for consumption of data from MySQL into R. For this assignment I experimented with MySQL, PostgreSQL, and SQLite.

```
library("RMySQL")
```

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

```
## Create connection 'mydb'
```

RMySQL requires a connection to be made to the SQL database. I am using my local connection, with username = "root" and no password setup. In future assignments I will look into building my databases in the cloud and implementing password authentication for tighter security.

```
mydb = dbConnect(MySQL(), user = 'root', dbname='data_607', host='localhost')
```

Select data from connection database

Using a simple select statement, we are able to load all of the movie preference data from MySQL into R.

```
data = dbGetQuery(mydb, "select * from data_607.movie_preferences")
```

Inspect Data

For this assignment I included 8 movies. Most of the films have respondent rankings for all respondents, but there are some exceptions. For instance, no one in my family watched Black Widow.

```
data
```

##	movie_name	respondent_1	respondent_2	respondent_3	respondent_4
## 1	1917	5	NA	5	NA
## 2	Black Widow	NA	NA	NA	NA
## 3	Catch Me If You Can	3	4	5	NA
## 4	Freaky Friday	3	1	2	5
## 5	Good Will Hunting	4	5	4	NA
## 6	Harry Potter 1	5	1	2	5
## 7	Honey I Shrunk The Kids	3	2	1	1

```
## 8      Lord Of The Rings 2      1      1      5      1
##  respondent_5
## 1          3
## 2          NA
## 3          NA
## 4          4
## 5          3
## 6          3
## 7          1
## 8          3
```

Dealing with Null Values

There are many methods to dealing with null values. These include marking the nulls with a new character or integer, imputing null values based on surrounding data, or removing columns / rows with excessive null values. 1. Because Black Widow has no rankings, I will choose to remove this row. 2. For all other NULL values, I will replace with '0', set to mark 'did not watch'

```
data <- data[-c(2), ]
```

```
data[is.na(data)] = 0
```

```
data
```

```
##      movie_name respondent_1 respondent_2 respondent_3 respondent_4
## 1      1917          5          0          5          0
## 3  Catch Me If You Can      3          4          5          0
## 4    Freaky Friday      3          1          2          5
## 5    Good Will Hunting      4          5          4          0
## 6    Harry Potter 1      5          1          2          5
## 7 Honey I Shrunk The Kids      3          2          1          1
## 8    Lord Of The Rings 2      1          1          5          1
##  respondent_5
## 1          3
## 3          0
## 4          4
## 5          3
## 6          3
## 7          1
## 8          3
```