DATA607 Assignment 2: SQL and R

James Williams

STEP 1: Load Movie_Rankings.sql into R dataframe

Data is stored here: https://github.com/jambawilliams/Movie_Rankings

There are some of the references I used: https://www.r-bloggers.com/accessing-mysql-through-r/ https://cran.r-project.org/web/packages/RMySQL/RMySQL.pdf http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization

Connect R to MySQL and pull in the Movie Ratings table.

```
library(RMySQL)
```

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

```
mydb <-dbConnect(MySQL(), user='root', password='Sourpie7!', dbname='Assignment2', host='localhost')
rs <- dbSendQuery(mydb, "select * from Movie_Rankings")
data <- fetch(rs, n=-1)
```

STEP 2: Explore Dataframe

Display the Movie Ratings table.

```
data
```

```
##    Friend Lion_King Little_Mermaid Snow_White Aladdin Pocahontas Jungle_Book
## 1     Sam         1              4          0       5          3           2
## 2   Eliza         2              4          3       5          5           1
## 3 Carrie         2              3          1       4          5           5
## 4     Joe         1              2          5       0          4           3
## 5 Gloria         4              5          1       5          2           3
## 6     Dan         3              1          2       4          5           5
```

STEP 3: Replace Null Values

Null values replaced with 0 since no ranking was orignally assigned. This will allow us to pull summary statistics later.

```
dbSendQuery(mydb, "UPDATE Movie_Rankings SET Snow_White = 0 where Snow_White is null")
```

```
## <MySQLResult:0,0,1>
```

```
dbSendQuery(mydb, "UPDATE Movie_Rankings SET Aladdin = 0 where Aladdin is null")
```

```
## <MySQLResult:0,0,2>
```

```
rs <- dbSendQuery(mydb, "select * from Movie_Rankings")
data_conditioned <- fetch(rs, n=-1)
data_conditioned
```

```
##   Friend Lion_King Little_Mermaid Snow_White Aladdin Pocahontas Jungle_Book
## 1    Sam         1              4          0       5          3           2
## 2  Eliza         2              4          3       5          5           1
## 3 Carrie         2              3          1       4          5           5
## 4    Joe         1              2          5       0          4           3
## 5 Gloria         4              5          1       5          2           3
## 6    Dan         3              1          2       4          5           5
```

STEP 4: Summary Statistics

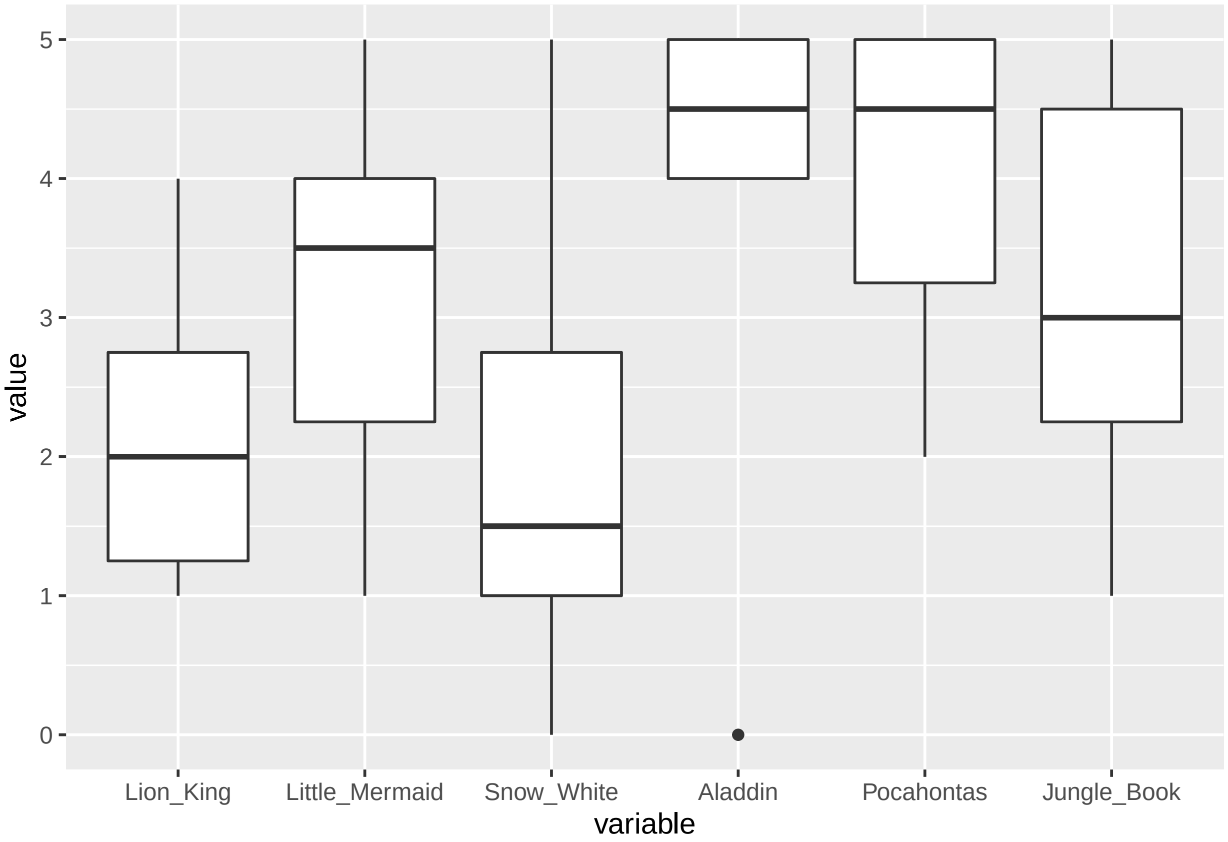
```
summary(data_conditioned)
```

```
##     Friend            Lion_King     Little_Mermaid    Snow_White
##  Length:6           Min.   :1.000   Min.   :1.000   Min.   :0.00
##  Class :character   1st Qu.:1.250   1st Qu.:2.250   1st Qu.:1.00
##  Mode  :character   Median :2.000   Median :3.500   Median :1.50
##                     Mean   :2.167   Mean   :3.167   Mean   :2.00
##                     3rd Qu.:2.750   3rd Qu.:4.000   3rd Qu.:2.75
##                     Max.   :4.000   Max.   :5.000   Max.   :5.00
##     Aladdin        Pocahontas    Jungle_Book
##  Min.   :0.000   Min.   :2.00   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:3.25   1st Qu.:2.250
##  Median :4.500   Median :4.50   Median :3.000
##  Mean   :3.833   Mean   :4.00   Mean   :3.167
##  3rd Qu.:5.000   3rd Qu.:5.00   3rd Qu.:4.500
##  Max.   :5.000   Max.   :5.00   Max.   :5.000
```

STEP 5: Graph Data Distribution

```
library(reshape2)
library(ggplot2)
ggplot(melt(data_conditioned), aes(variable, value)) + geom_boxplot()
```

```
## Using Friend as id variables
```


5
4
3
2
1
0
value
Lion_King
Little_Mermaid
Snow_White
Aladdin
Pocahontas
Jungle_Book
variable