Chunjie_Nan_Assignment2

Chunjie Nan

9/5/2021

I have asked five friends and get a simple survey of rating recent movies they have seen. Six movies were picked, and each friend of mine have rated according to their satisfaction of movies. The table was created with MySQl, and exported to Github as a CSV file.SQL code on git hub is here[linked phrase]https://raw.githubusercontent.com/nancunjie4560/Data607/master/Movies.sql

```
library(kableExtra)
library(RMySQL)
```

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

```
library(RCurl)
library(DBI)
library(ggplot2)
```

Importing data from Github

```
movie_data<- read.csv(url('https://raw.githubusercontent.com/nancunjie4560/Data607/master/Movies.csv'))
str(movie_data)
```

```
## 'data.frame':    5 obs. of  7 variables:
##  $ respondent     : chr  "Misun" "John" "Juyong" "Kevin" ...
##  $ F_F_Nine       : int  5 4 4 4 5
##  $ still_water    : chr  "4" "NULL" "NULL" "3" ...
##  $ shang_chi      : chr  "3" "2" "1" "NULL" ...
##  $ dont_breath_Two: chr  "NULL" "5" "NULL" "5" ...
##  $ black_widow    : int  2 3 3 4 4
##  $ rogue_hostage  : chr  "NULL" "5" "4" "4" ...
```

There were many NULL in our dataset, it because my friends have not seen the new movies yet. So, I have replace the NULL value to the median value of 3 in rating.

Replace NULL to "3"

```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:RCurl':
##
##     complete
```

```
data<-gather(movie_data,movies,rating,-respondent)
data$rating<- ifelse(data$rating=='NULL', '3',data$rating)
data$rating
```

```
##  [1] "5" "4" "4" "4" "5" "4" "3" "3" "3" "4" "3" "2" "1" "3" "3" "3" "5" "3" "5"
## [20] "4" "2" "3" "3" "4" "4" "3" "5" "4" "4" "3"
```

```
data
```

```
##    respondent          movies rating
## 1       Misun        F_F_Nine      5
## 2        John        F_F_Nine      4
## 3      Juyong        F_F_Nine      4
## 4       Kevin        F_F_Nine      4
## 5        Juju        F_F_Nine      5
## 6       Misun     still_water      4
## 7        John     still_water      3
## 8      Juyong     still_water      3
## 9       Kevin     still_water      3
## 10       Juju     still_water      4
## 11      Misun       shang_chi      3
## 12       John       shang_chi      2
## 13     Juyong       shang_chi      1
## 14      Kevin       shang_chi      3
## 15       Juju       shang_chi      3
## 16      Misun dont_breath_Two      3
## 17       John dont_breath_Two      5
## 18     Juyong dont_breath_Two      3
## 19      Kevin dont_breath_Two      5
## 20       Juju dont_breath_Two      4
## 21      Misun     black_widow      2
## 22       John     black_widow      3
## 23     Juyong     black_widow      3
## 24      Kevin     black_widow      4
## 25       Juju     black_widow      4
## 26      Misun   rogue_hostage      3
## 27       John   rogue_hostage      5
## 28     Juyong   rogue_hostage      4
## 29      Kevin   rogue_hostage      4
## 30       Juju   rogue_hostage      3
```

Sort the rate from High to Low.

```
library(dplyr)
```

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

```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

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

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

```
group<-group_by(data, movies) %>%
    arrange(desc(rating))
group
```

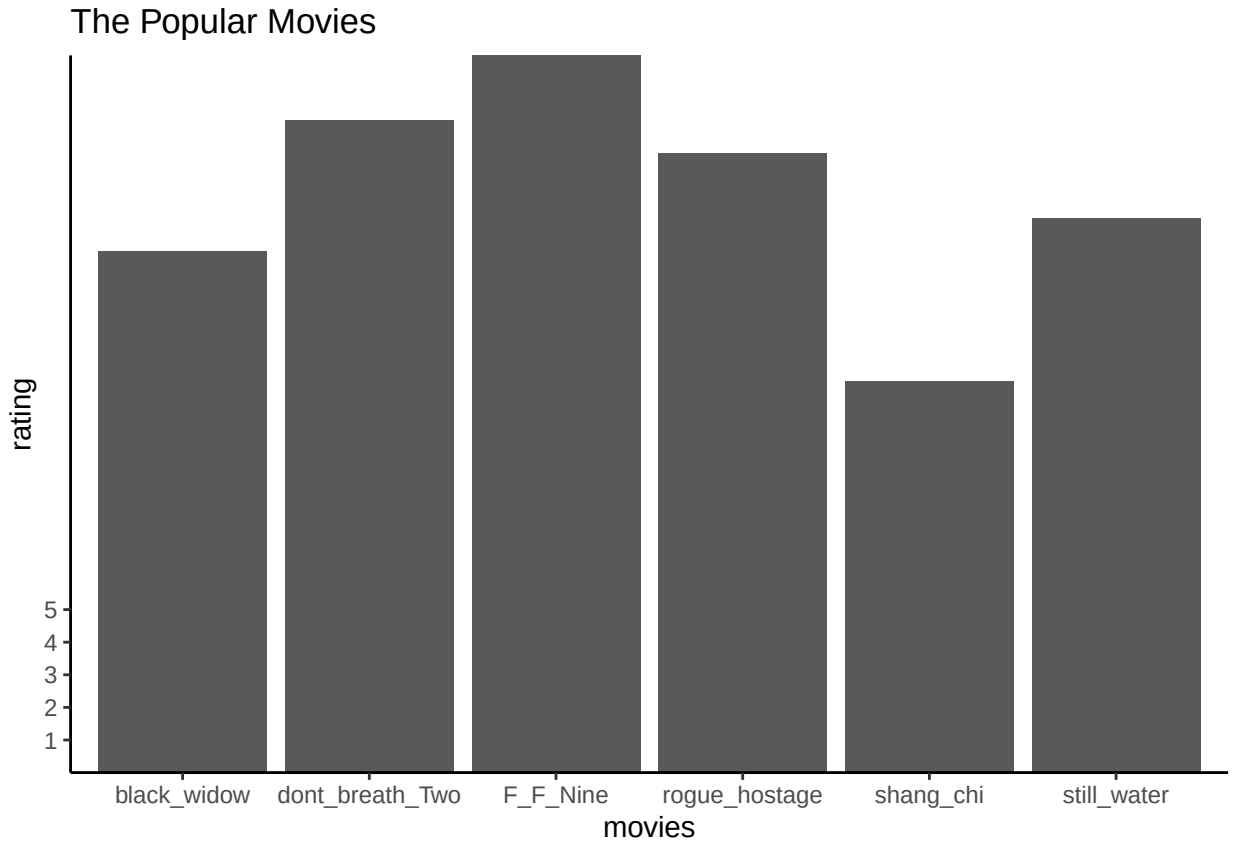
```
## # A tibble: 30 x 3
## # Groups:   movies [6]
##    respondent movies          rating
##    <chr>      <chr>           <chr>
##  1 Misun      F_F_Nine        5
##  2 Juju       F_F_Nine        5
##  3 John       dont_breath_Two 5
##  4 Kevin      dont_breath_Two 5
##  5 John       rogue_hostage   5
##  6 John       F_F_Nine        4
##  7 Juyong     F_F_Nine        4
##  8 Kevin      F_F_Nine        4
##  9 Misun      still_water     4
## 10 Juju       still_water     4
## # ... with 20 more rows
```

From the above table, I can see that F_F_Nine and dont_breath_two win the highest rating in this dataset. The respondent is removed from the group for plotting purpose.

```
g<-group[,-1]
g
```

```
## # A tibble: 30 x 2
## # Groups:   movies [6]
##    movies          rating
##    <chr>           <chr>
##  1 F_F_Nine        5
##  2 F_F_Nine        5
##  3 dont_breath_Two 5
##  4 dont_breath_Two 5
##  5 rogue_hostage   5
##  6 F_F_Nine        4
##  7 F_F_Nine        4
##  8 F_F_Nine        4
##  9 still_water     4
## 10 still_water     4
## # ... with 20 more rows
```

```
ggplot(g, aes(x = movies,rating)) +
  geom_col() +
  ggtitle('The Popular Movies')+
  theme_classic()
```



Each Block shows the rating, and the highest rating 5 is square, and as the rating gets lower, the height of cubic gets lower. According to the chart, we can see that F_F_Nine won among the movies with 5+5+4+4+4 = 22 ratings. Dont_Breath_two won the second position which is slightly higher than rougue_hostage. Shang_Chi gets the worst rate among the movies, and I'm definitely going to watch Shang_Chi and find out why is this given the worst rate by my friends.