Assignment week2

waheeb Algabri

You might ask yourself.Am I able to use a password without having to share the password with people who are viewing my code? Yes it is possible to use a password without having to share it with others who are viewing the code. by using this code in R by using the Sys.getenv function:

```
library(knitr)
opts_chunk$set(fig.width=7, fig.height=5)
```

In case you need to create a normalized set of tables that corresponds to the relationship between your movie viewing friends and the movies being rated.

```
{r, language='sql'}

CREATE TABLE Friend (
  FriendID INT PRIMARY KEY,
  Name VARCHAR(50) NOT NULL
);

CREATE TABLE Movie (
  MovieID INT PRIMARY KEY,
  Name VARCHAR(50) NOT NULL
);

CREATE TABLE Rating (
  FriendID INT NOT NULL,
  MovieID INT NOT NULL,
  Rating INT NOT NULL,
  PRIMARY KEY (FriendID, MovieID),
  FOREIGN KEY (FriendID) REFERENCES Friend (FriendID),
  FOREIGN KEY (MovieID) REFERENCES Movie (MovieID)
);
```

My goal in this assignment is to determine the overall popularity of the six movies among the five people who participated in the survey. To achieve this, I will calculate the average rating for each movie and compare the results to determine which movie was rated the highest and which was rated the lowest. This will give me an idea of which movies are most popular among the people in my survey, and whether there are any significant differences in the ratings for each movie.

Fist of all i,m goiing to connect R to MYSQL to get my data of my survey, so i will use a packet called RMYSQL

```
library(RMySQL)

# Connect to the database using the environment variables
```

```
con <- dbConnect(MySQL(),
                 host = "localhost",
                 username = "root",
                 password = "Alex9297248844",
                 dbname = "Movie_Ratings")
```

Load data from the database into an R dataframe

```
df <- dbGetQuery(con, "SELECT * FROM Ratings")
```

```
str(df)
```

```
## 'data.frame':    27 obs. of  4 variables:
##  $ id    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Movie : chr  "M3GAN" "M3GAN" "M3GAN" "Vengeance" ...
##  $ Rating: int  4 5 3 3 4 5 5 4 3 5 ...
##  $ Person: chr  "Shoshana" "Sammy" "Jason" "Shoshana" ...
```

Analizing

First of all I have to clean and handling missing data , So I have to make sure the data is clean and consistent.

```
dplyr::filter(df)
library(dplyr)
# Check for missing values
sum(is.na(df))

# Impute missing values with the mean
df <- df %>% mutate_if(is.numeric, funs(ifelse(is.na(.), mean(., na.rm = TRUE), .)))

# check for duplicates
sum(duplicated(df))

# Remove duplicates
df <- unique(df)

# Remove irrelevant columns
df <- select(df, Movie, Rating)

# Convert data type
df$Rating <- as.numeric(df$Rating)
```

Second, I will calculate the average rating for each movie and compare the results to determine which movie was rated the highest and which was rated the lowest.

```
# Load required libraries
library(dplyr)

# Group by movie name and calculate the average rating for each movie
df_avg <- df %>%
  group_by(Movie) %>%
  summarise(average_Rating = mean(Rating))

# Sort the data by average rating in descending order
df_avg <- arrange(df_avg, desc(average_Rating))

# Print the top movie
df_avg[1, ]
```

```
## # A tibble: 1 x 2
##   Movie     average_Rating
##   <chr>              <dbl>
## 1 Nomadland           4.25
```

```
# Print the bottom movie
df_avg[nrow(df_avg), ]
```

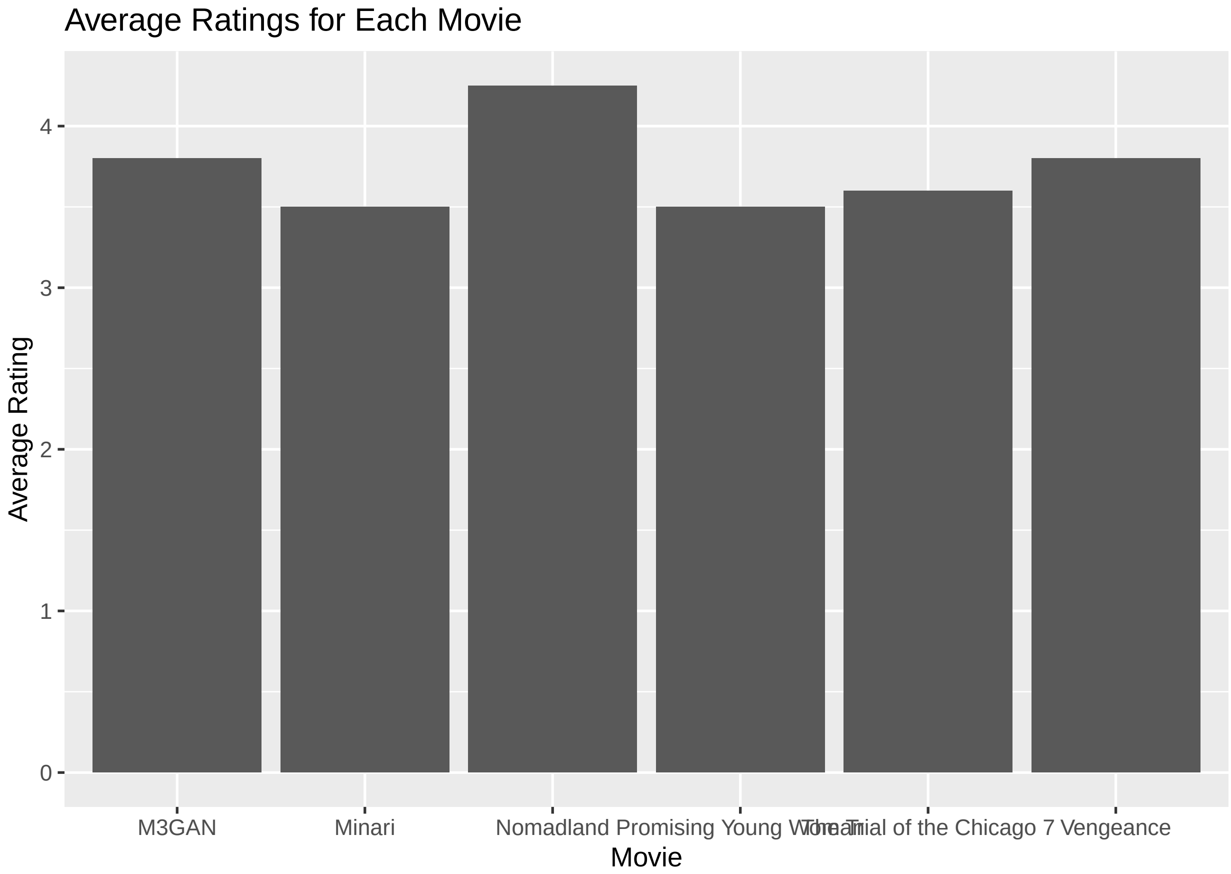
```
## # A tibble: 1 x 2
##   Movie                 average_Rating
##   <chr>                          <dbl>
## 1 Promising Young Woman            3.5
```

In the previous code, I grouped the data by movie name and calculate the average rating for each movie using the group_by and summarise functions from the dplyr library. Then, i star sorting the data by average rating in descending order using the arrange function. Finally, i print the top and bottom movies by accessing the first and last rows of the df_avg data frame, respectively.

Third, i will visualizing the results: I,m going to create charts or plots to help visualize the results and make the insights easier to understand. For example, i created a bar chart that shows the average rating for each movie.

```
# Load required libraries
library(ggplot2)

# Create the bar chart
ggplot(df_avg, aes(x = Movie, y = average_Rating)) +
  geom_col() +
  ggtitle("Average Ratings for Each Movie") +
  xlab("Movie") +
  ylab("Average Rating")
```

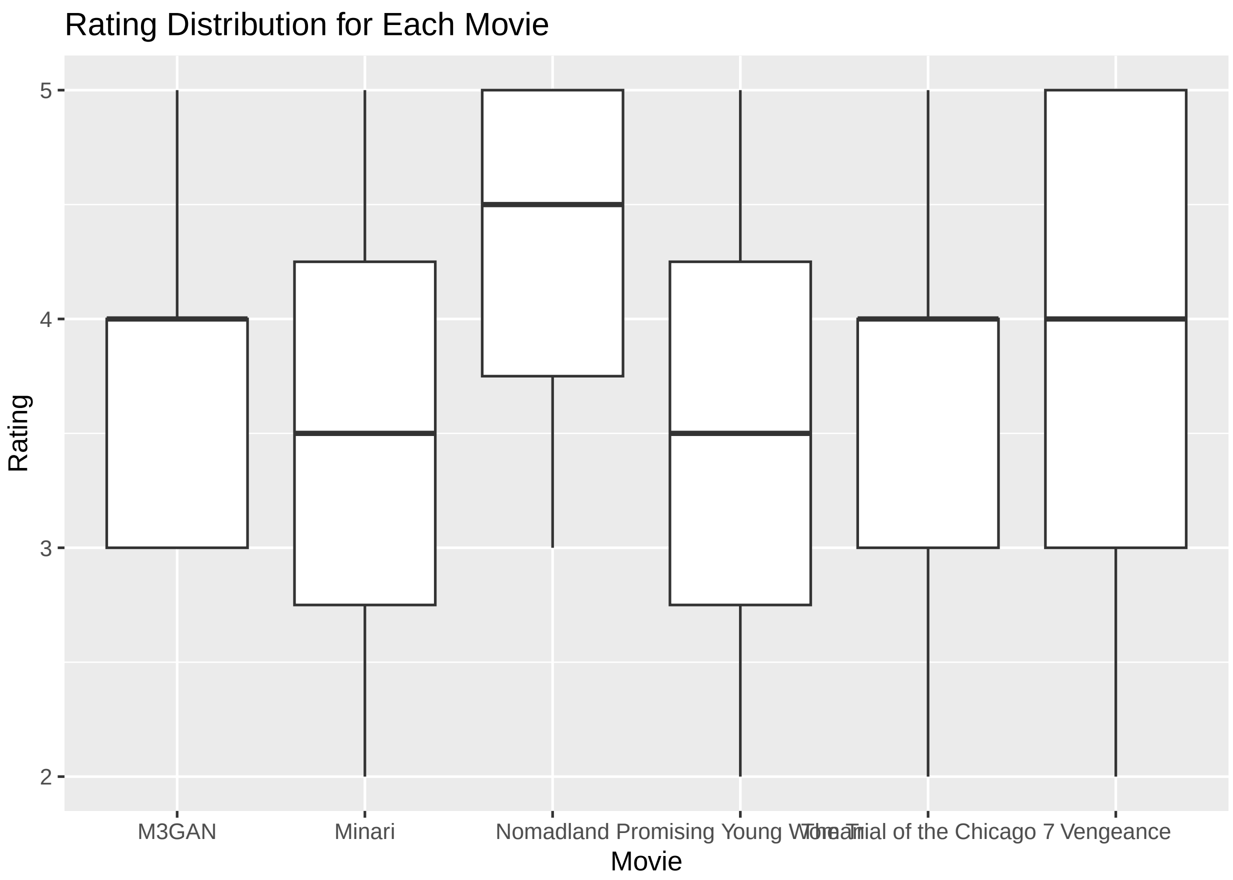


In the previous code,I used the ggplot2 library to create a bar chart that shows the average rating for each movie. The ggplot function creates the basic plot, and I specify the data to use (df_avg) and the variables to map to the x and y axes (movie and average_Rating, respectively). The geom_col function creates the columns for the bar chart, and the ggtitle, xlab, and ylab functions specify the title, x-axis label, and y-axis label, respectively.

finally , I will look for patterns: Examine the data to see if there are any patterns or trends in the ratings. For example, are the ratings for certain movies consistently high or low?

```
# Load required libraries
library(ggplot2)

# Create the box plot
ggplot(df, aes(x = Movie, y = Rating)) +
  geom_boxplot() +
  ggtitle("Rating Distribution for Each Movie") +
  xlab("Movie") +
  ylab("Rating")
```



In the code, I used the ggplot2 library to create a box plot that shows the distribution of ratings for each movie. The ggplot function creates the basic plot, and we specify the data to use (df) and the variables to map to the x and y axes (movie_name and rating, respectively). The geom_boxplot function creates the box plot, and the ggtitle, xlab, and ylab functions specify the title, x-axis label, and y-axis label, respectively. From the box plot, I can easily see the median, quartiles, and outliers for each movie, and make comparisons to see if there are any patterns or trends in the ratings.

In conclusion

After doing some work on my data set i found that the higher average rating and similar rating distribution for this movie indicates that the Movie which called (Nomadlan) is the most popular among the people I surveyed, but for further analysis to confirm this, I can do it by performing statistical tests, such as (t-test) to see if the difference in ratings between this movie and the others is statistically significant but no need for that since I have visualized the results when I created a bar chart which showed the average rating for each movie.