Exploratory Data Analysis (EDA) Report

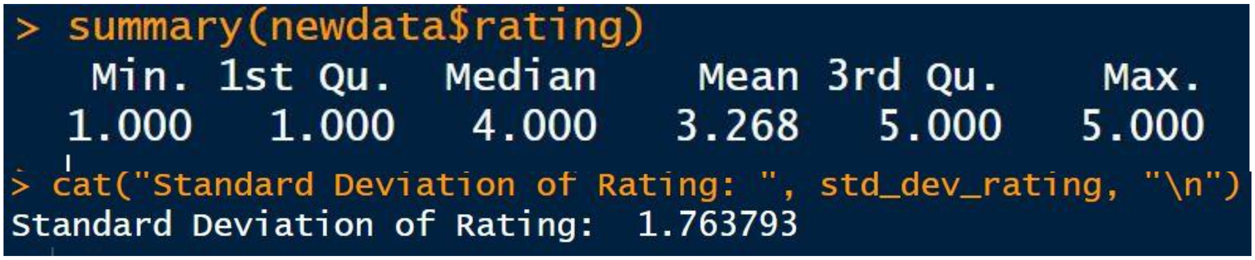
Dataset Overview

The dataset named "twitterdata" consists of 87,072 rows and 6 variables, namely date, username, review, rating, comments, and country. The variables are primarily of character type, except for the "rating" variable, which is an integer.

Descriptive Statistics

Numeric Variables

The "rating" variable, representing the ratings given, is of integer type. Here are some summary statistics for the ratings:



```
> summary(newdata$rating)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   4.000   3.268   5.000   5.000 
> cat("Standard Deviation of Rating: ", std_dev_rating, "\n")
Standard Deviation of Rating:  1.763793
```

Mean Rating: 3.27

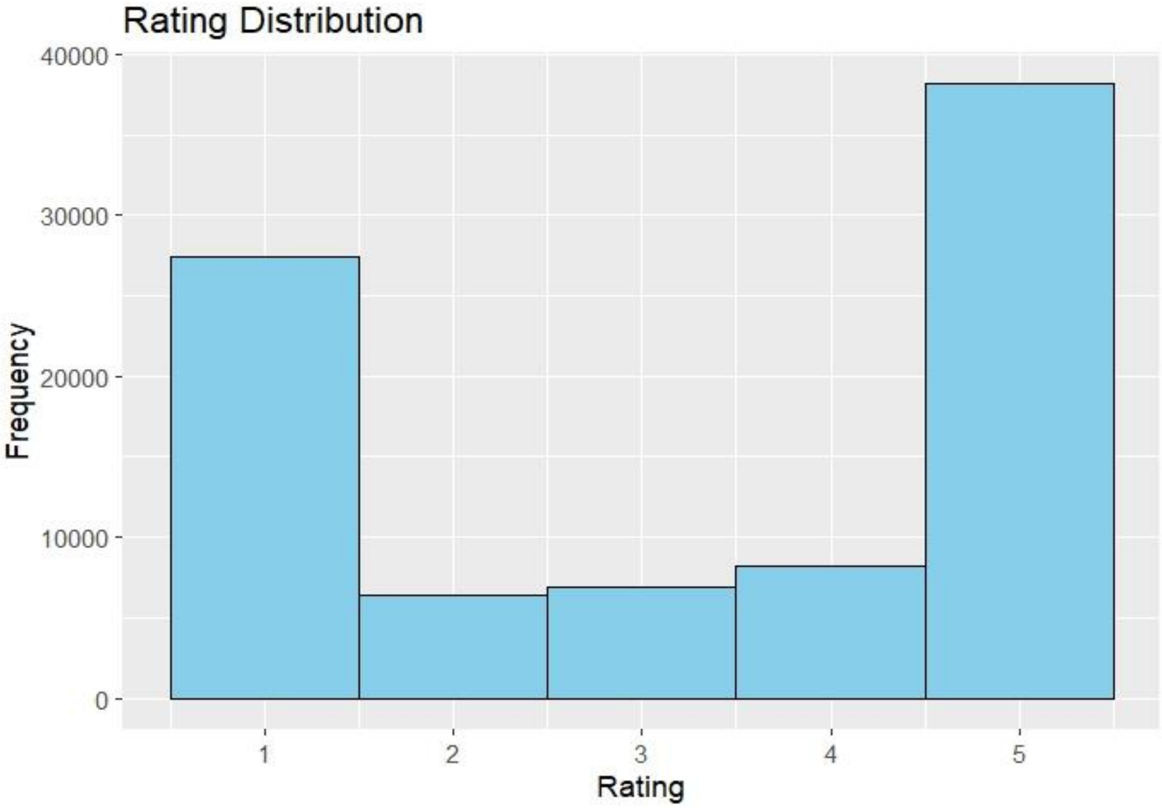
Median Rating: 4.00

Standard Deviation: 1.76

These statistics provide an overview of the central tendency and variability in the distribution of ratings. The mean rating indicates the average score given by users, while the median rating represents the middle value, showing that at least 50% of the ratings are 4.00 or below. The standard deviation of 1.76 signifies the degree of deviation or spread of ratings around the mean.

Rating Distribution

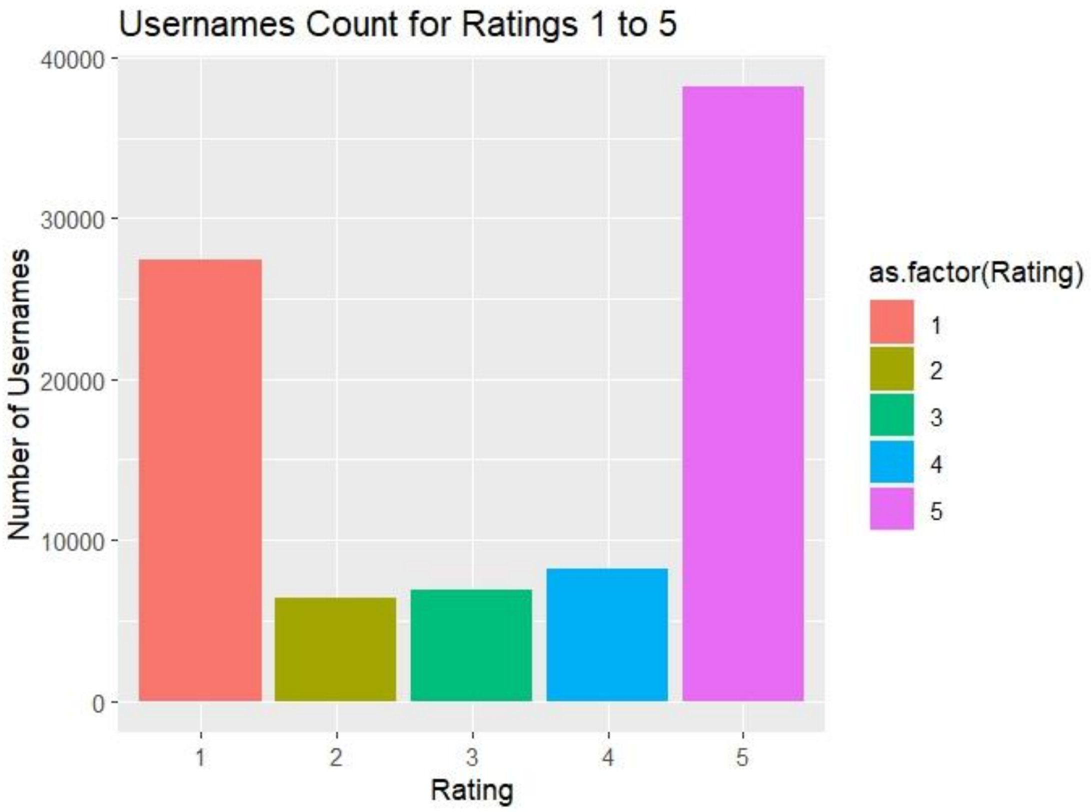
A histogram was created to visualize the distribution of ratings. The majority of ratings fall within the range of minimum 8000 to maximum 39000 users.



From the above histogram we can see the number to ratings are distributed in a range of 1 -5. minimum rating ranges 8000 and the maximum rating ranges 39000

Usernames Count for Ratings 1 to 5

The count of unique usernames giving ratings in the range of 1 to 5 was examined. The distribution is as follows:



A number of 39000 user gives the twitter app 5-star rating.

A number of 28000 user gives the twitter app 1-star rating.

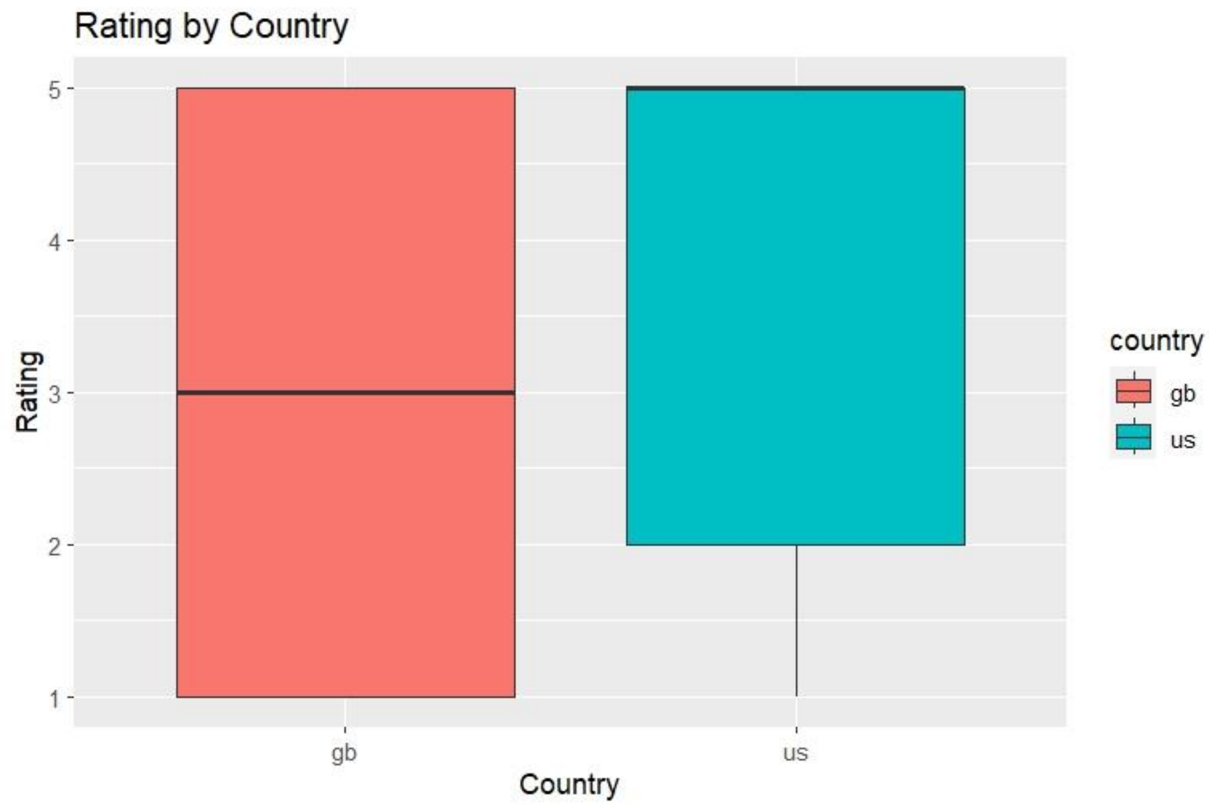
A number of 9800 user gives the twitter app 4-star rating.

A number of 9500 user gives the twitter app 3-star rating.

A number of 8000 user gives the twitter app 2-star rating.

Country Rating Distribution

The distribution of ratings across different countries is as follows:



From the above boxplot we can see that number to ratings are distributed in a range of 1 -5. between Germany and USA

For Germany, the user of Twitter is given the

minimum rating of **1-star**

maximum rating of **5-star**

For the USA, the user of Twitter is given the

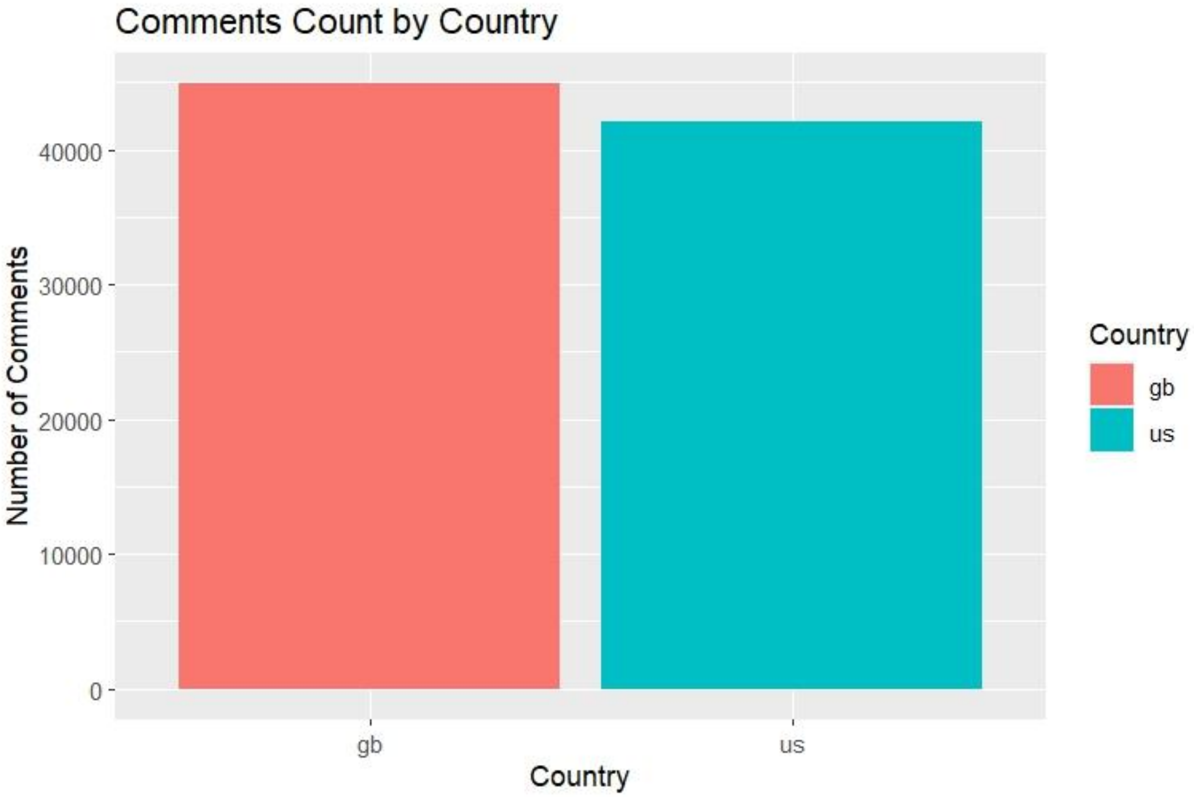
minimum rating of **2-star**

maximum rating of **5-star**

Comments Analysis

Comments Count by Country

The number of comments was analyzed across different countries. The distribution is as follows:



From the above analysis we can see that, the maximum number of users from Germany have made their comments on the Twitter app store. And the US is just below the number.

For Germany, above 50000 users comments on the Twitter app .

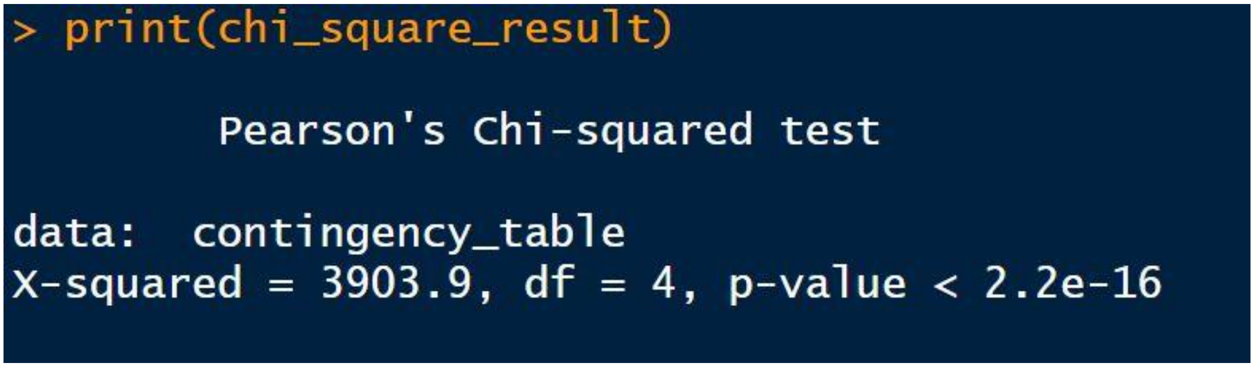
For USA, about 39000 of users comments on the Twitter app.

Categorical Variables

Correlation measures the linear relationship between numeric variables.

For categorical variables, we have to measures the association, such as the chi-squared test This test is suitable for analyzing the association between two categorical variables.

For the association between two categorical variables, a chi-squared test was performed.



```
> print(chi_square_result)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 3903.9, df = 4, p-value < 2.2e-16
```

The chi-squared test was conducted to examine the association between the "country" and “rating” variables.

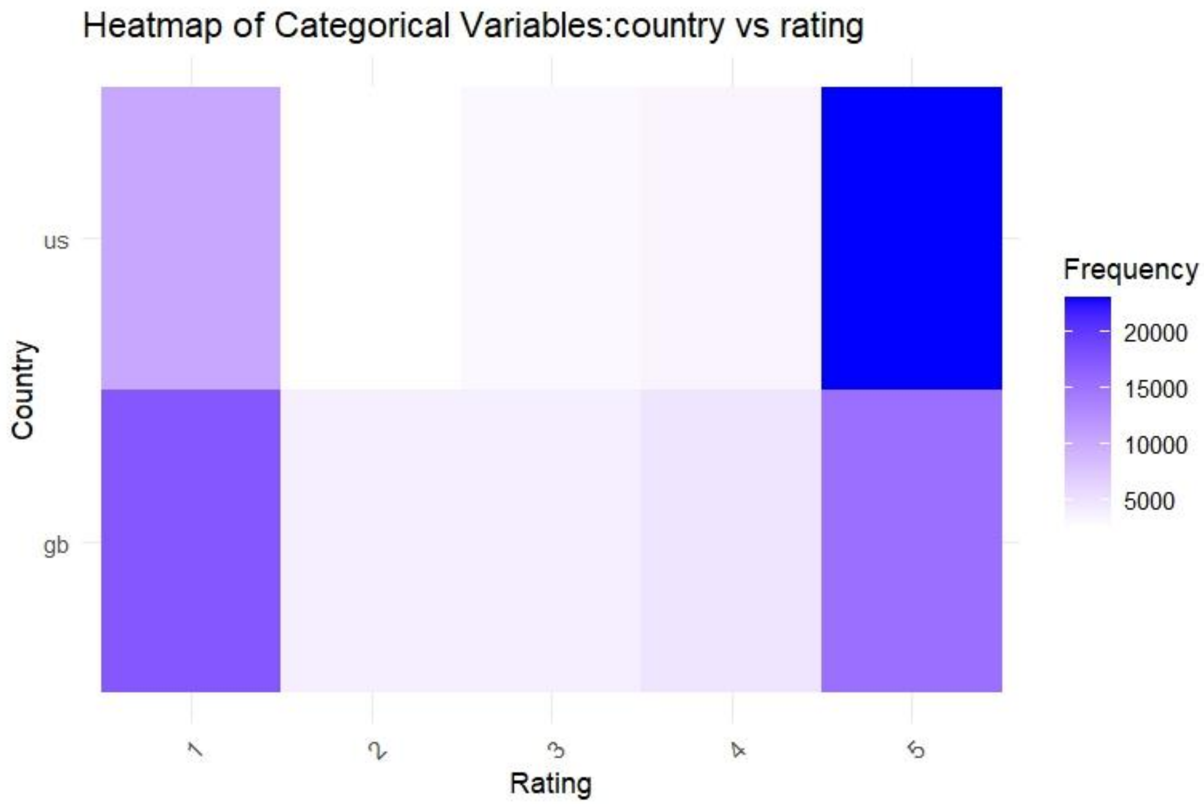
The chi-squared test yielded a test statistic of : 3903.9

The degrees of freedom for the test were : 4

The p-value associated with the chi-squared statistic was : 2.2e-16 which is <0.05

The low p-value (2.2e-16 < 0.05[p-value]) suggests that there is a statistically significant association between the "country" and "rating" variables.

The heatmap below visualizes the observed frequencies of categories across countries.



The heat map successfully describes the categorical variables.