Statistical report

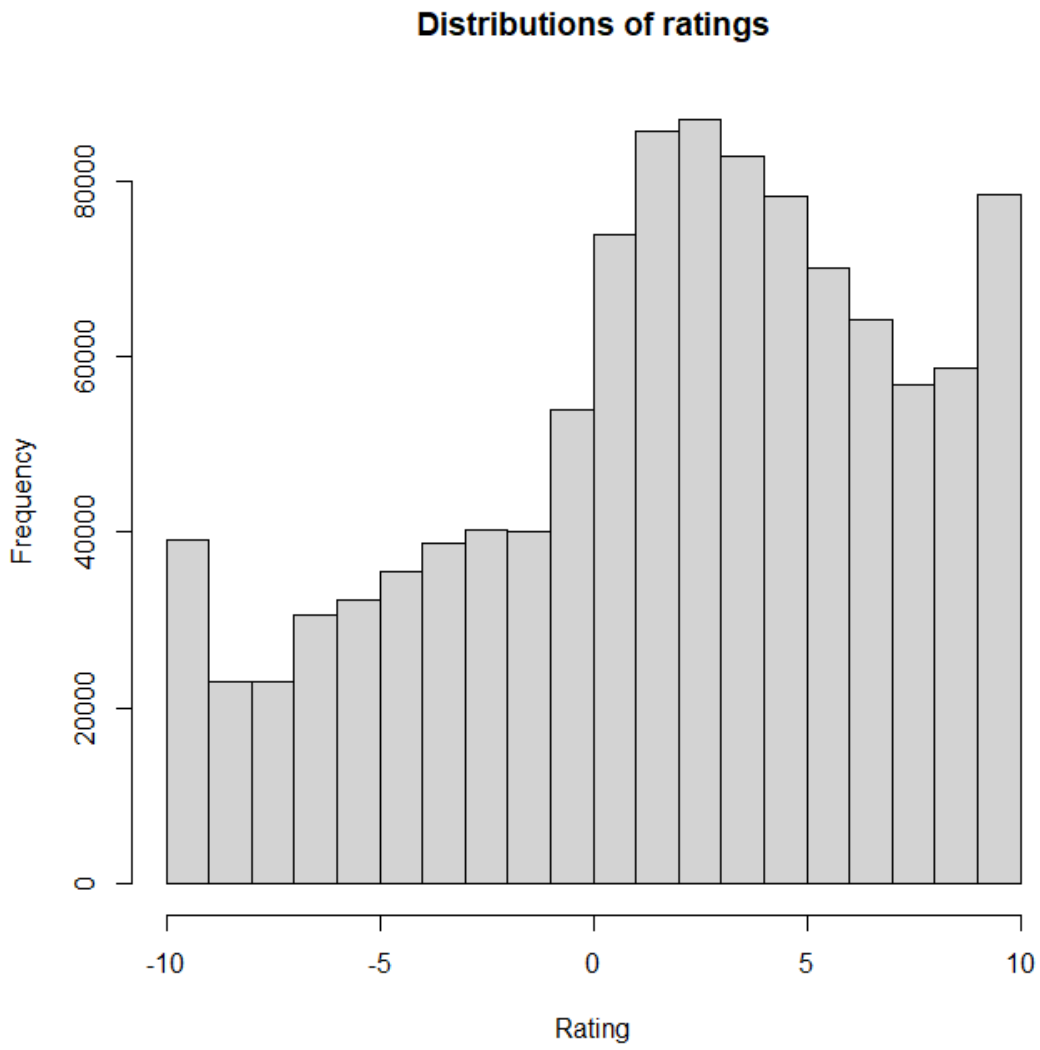
This document is a statistical report on the dataset used in our project.
The dataset contains anonymous ratings(-10 to 10) provided by a total of 41,000 users. Train file contains 1.1 million ratings for 139 jokes.

Here we load the data and create the dataframe.

```
library("ggplot2")
trainSet <- read.csv("train.csv")
df <- data.frame(trainSet)
```

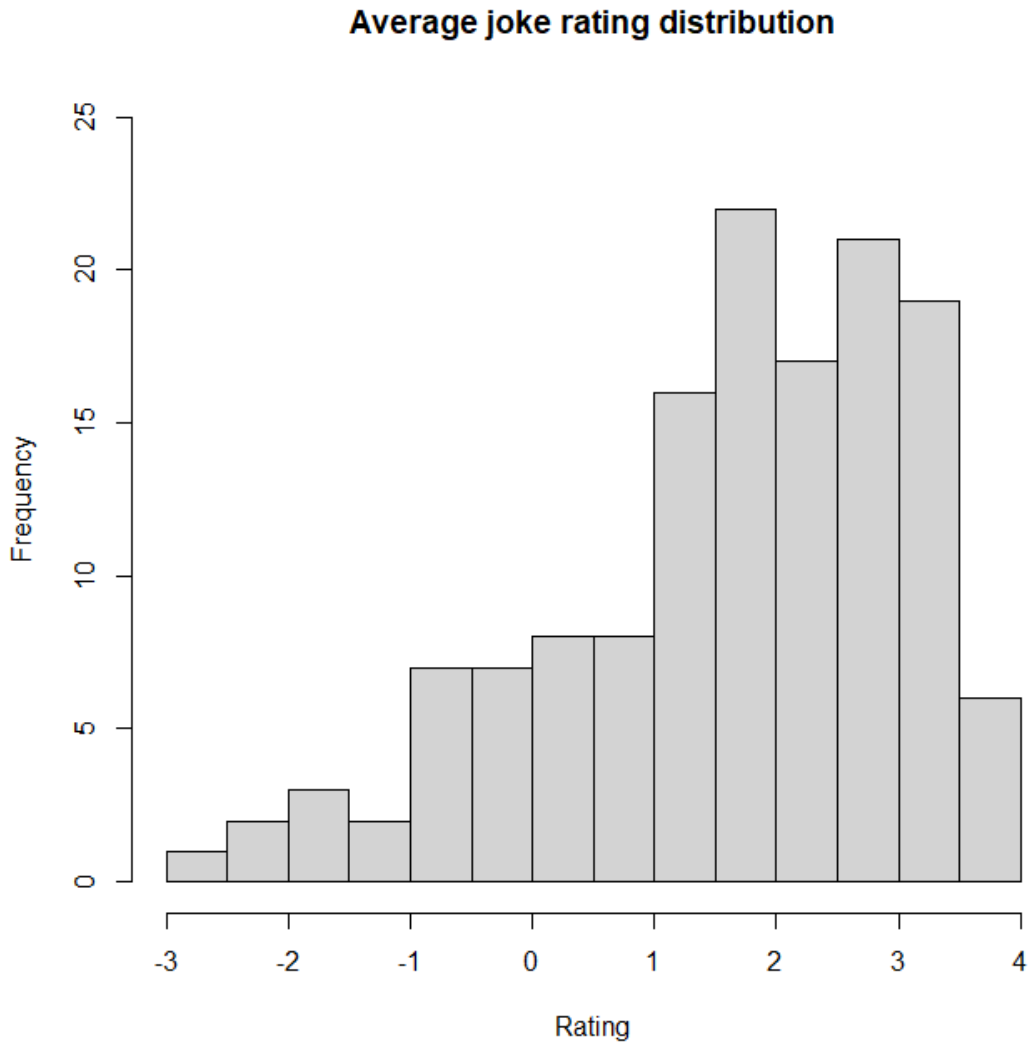
Here is a histogram representing distribution of the grade. The distribution is little skewed to the right.

```
hist(df$Rating,main = "Distributions of ratings", xlab = "Rating")
```



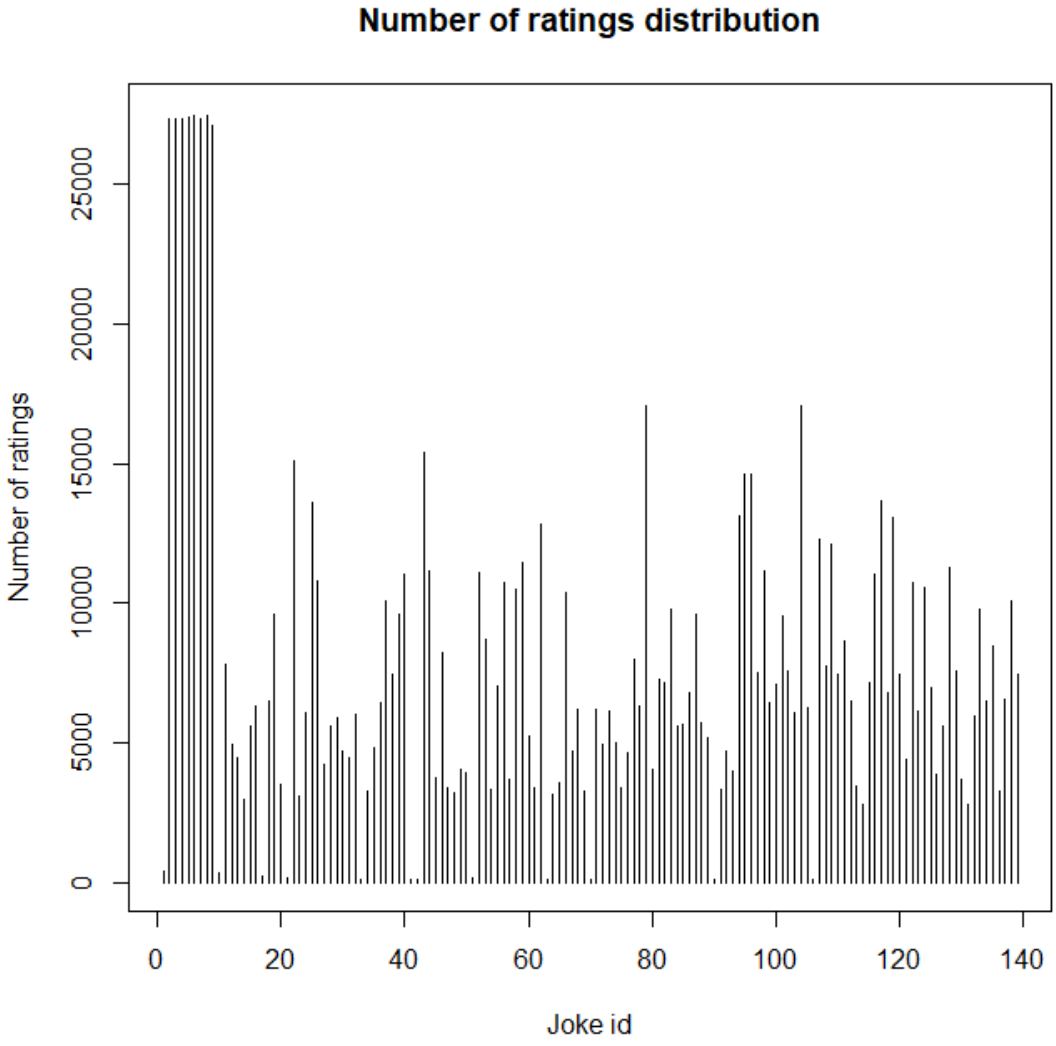
Here is a histogram representing average joke rating distribution. Average joke rating ranges from -3 to 4, the most popular are within [1.5, 2]. The distribution of average ratings is a little skewed to the right.

```
average_ratings <- aggregate(df$Rating ~ df$joke_id, df, mean)
hist(average_ratings$`df$Rating`,ylim=c(0,25), breaks = 20, main = "Average joke
rating distribution", xlab = "Rating")
```



Here is a histogram representing amount of ratings distribution.

```
ratings_numbers <- aggregate(df$Rating ~ df$joke_id, df, length)
plot(ratings_numbers, type = "h", breals = 139, main = "Number of ratings
distribution", xlab = "Joke id", ylab = "Number of ratings")
```


Number of ratings distribution
Number of ratings
0
5000
10000
15000
20000
25000
0
20
40
60
80
100
120
140
Joke id