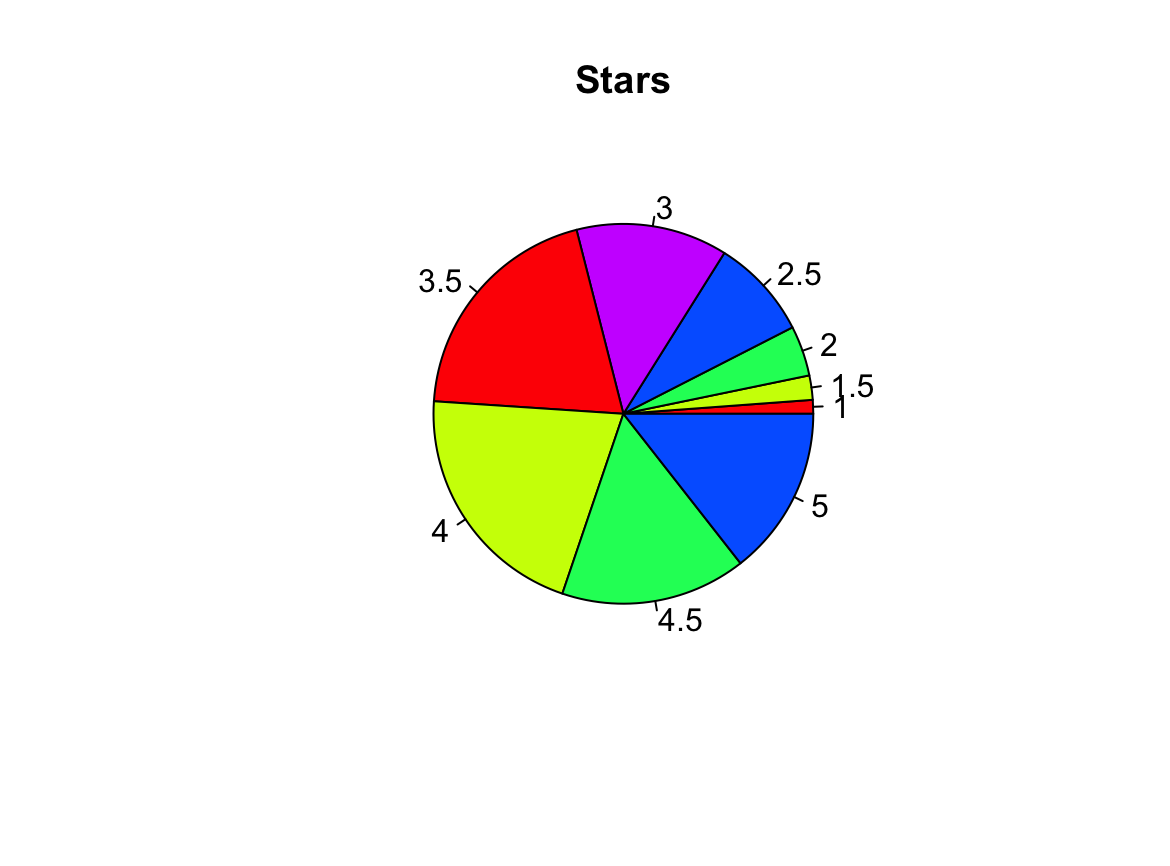
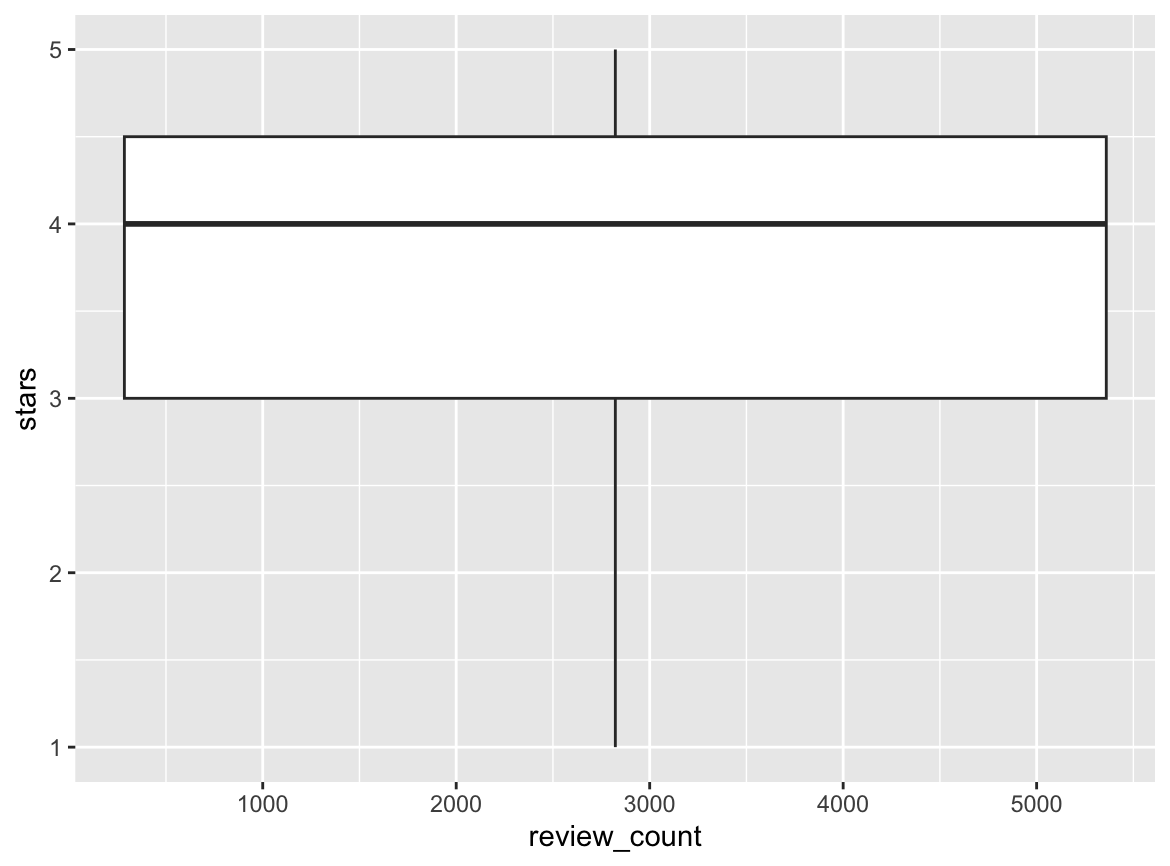
James Bean

DSC-140

Technical Report 2

The analysis was based on 2 data sets, the Business dataset and the Users dataset. The business dataset includes information about some businesses listed on Yelp. These include things like business names, location and star ratings etc. The users data set contains information about the activities and engagement of users, such as votes received, number of users and reviews written. The columns in the user data are as follows, “user\_id” to identify each user, “name”, “review count” written by the user, “yelping\_since” when the user created their account, “useful\_votes”, “funny\_votes”, “cool\_votes”, “fans”, “average\_stars”, and “elite” when the user reaches elite status on yelp. The user data set has 11 columns and 552339 rows. The categorical data columns are user\_id, name, and elite. The numerical data is review\_count, useful\_votes, funny\_votes, cool\_votes, fans\_ and average\_stars. There is also date data for \_yelping\_since. The user data set provides a lot of information about popularity metrics, user behavior and business reviews. This analysis focuses on user popularity and identifying insights companies can use to enhance the user experience.

Many different methods of analysis were used in this report. The distribution of star rating was demonstrated by using a pie chart. Most of the business had ratings in the range of 4 to 5 stars.

A box plot was used to analyze how review counts vary across star ratings. Both businesses with high and low ratings had many reviews, this implies that review count is not a good indication of the quality of business.

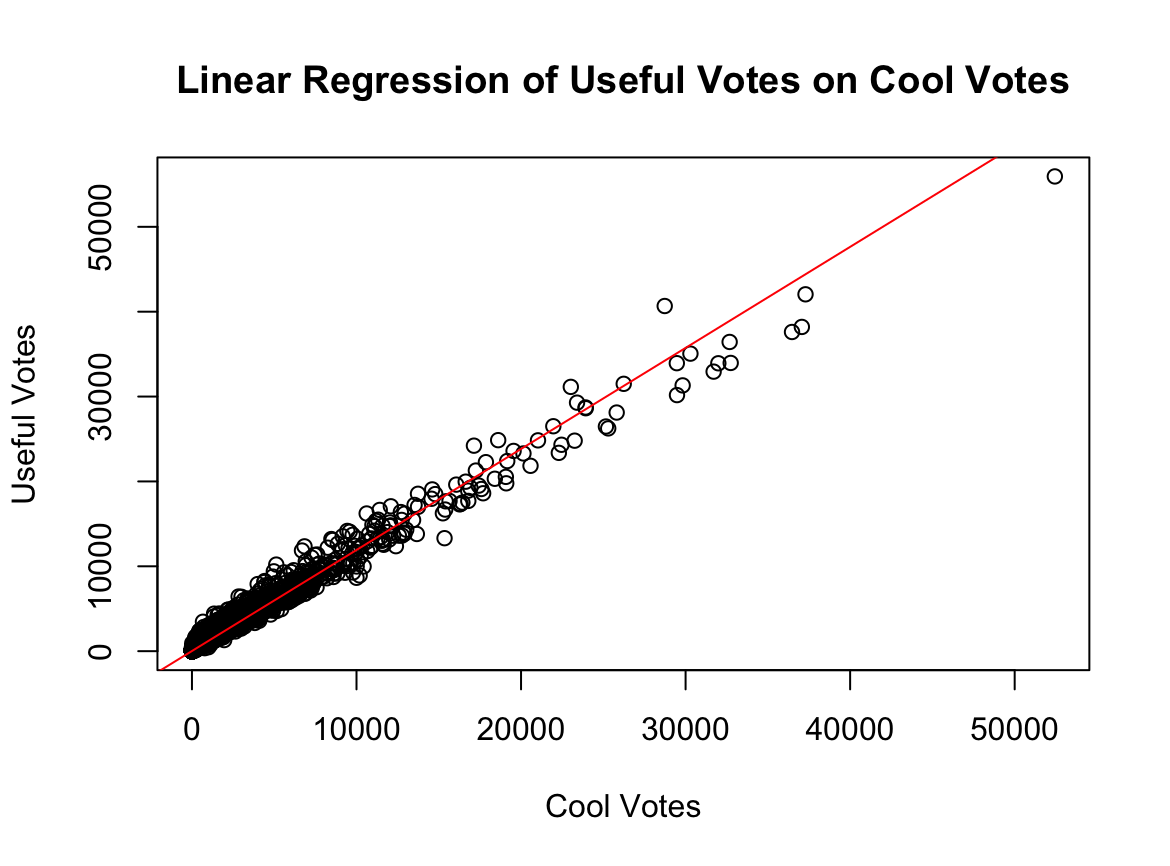
Here is the correlation matrix for cool\_votes, funny\_votes, and useful\_votes. Strong positive correlations, mostly between cool\_votes and useful\_votes.

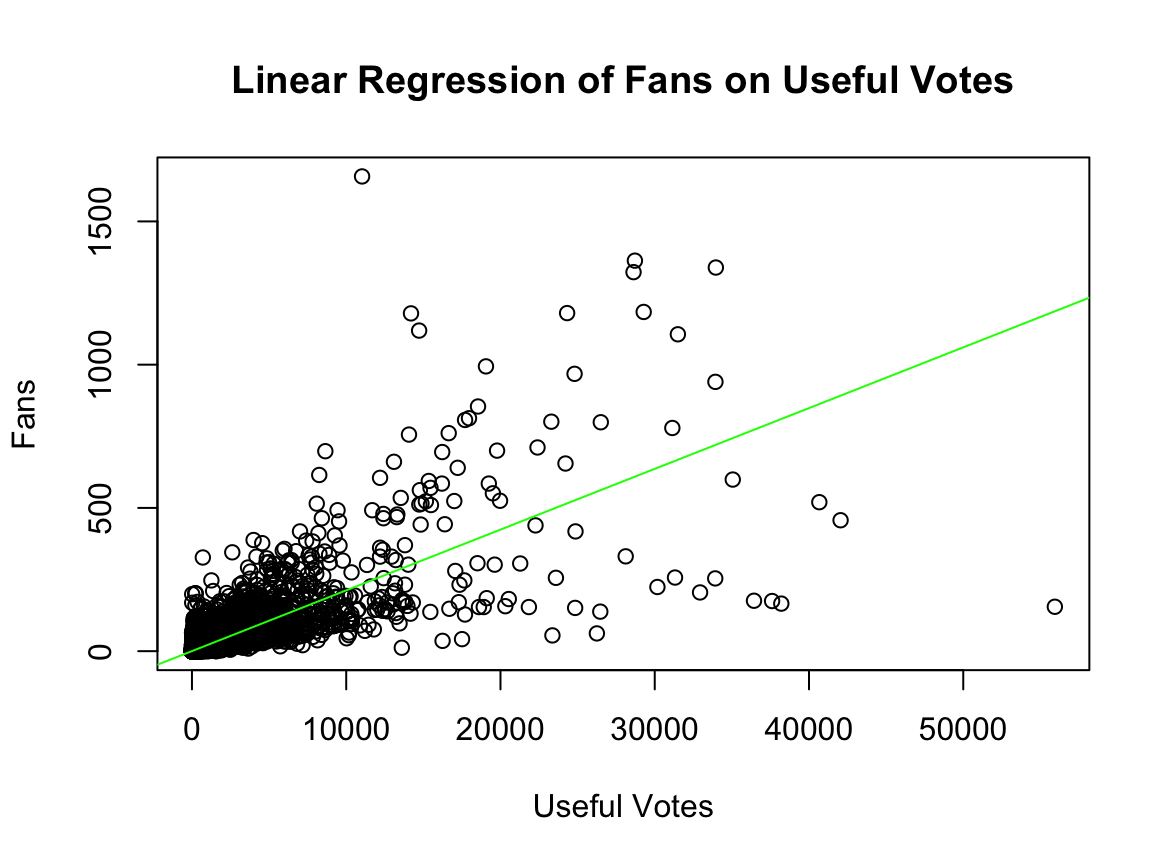
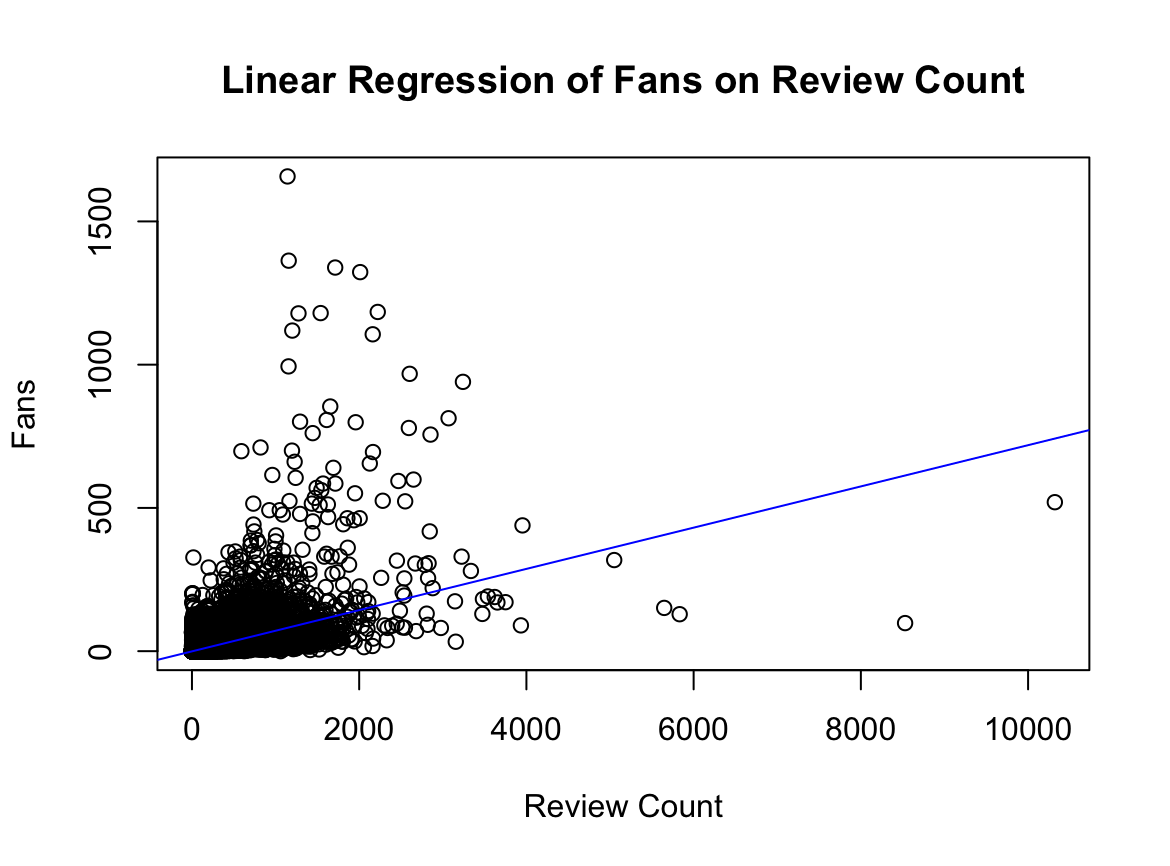
cool\_votes funny\_votes useful\_votes

cool\_votes 1.0000000 0.9764113 0.9832708

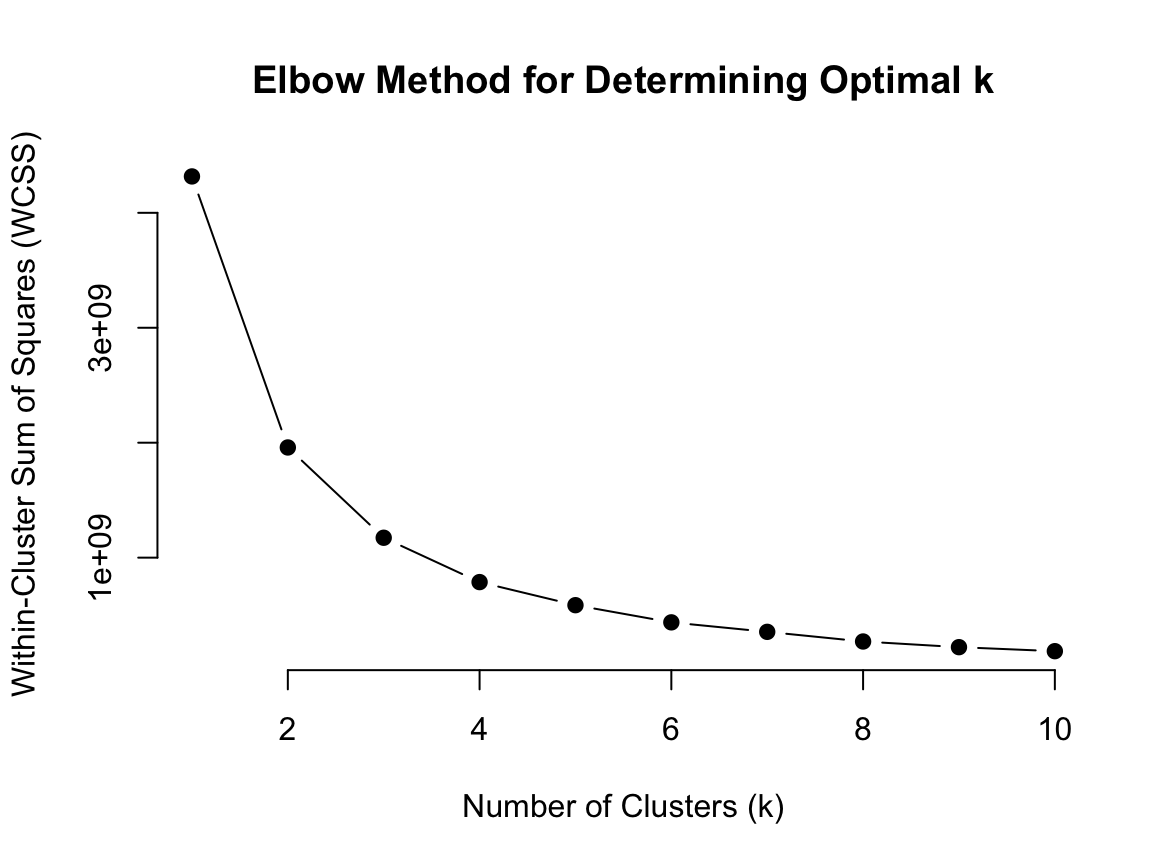
funny\_votes 0.9764113 1.0000000 0.9546541

useful\_votes 0.9832708 0.9546541 1.0000000

A linear regression model was used to predict useful\_votes based on cool\_votes. The model revealed a strong positive relationship.

A second linear regression model was used to predict a relationship between fans and review\_count. While there was a positive trend the relationship was very weak. This shows that writing more reviews does not mean more fans. However, there is a strong positive relationship between fans and useful\_votes. This suggest that reviews that are deemed useful attracts more fans.

The elbow method was used to determine the best number of clusters for the dataset then K-means clustering was applied. Users were put into different groups based on review\_count and fans.



From this analysis we can conclude that useful votes are the strongest driver of user popularity, while funny and cool votes contribute to the engagement. High review counts and fan counts showed a weak correlation, however useful votes and cool votes were important to attracting fans.