Alkım Karakurt

[E-posta adresi]

Abstract

This project aims to analyze Amazon cell phone reviews and items datasets which are taken from Kaggle. In this paper, I tried to use descriptive and sentiment analysis for dataset. It is aimed to answer different questions about datasets and mainly I tried to show that how sentiment of reviews effect brands’ ratings and success with using sentiment analysis tools.

TERM PROJECT

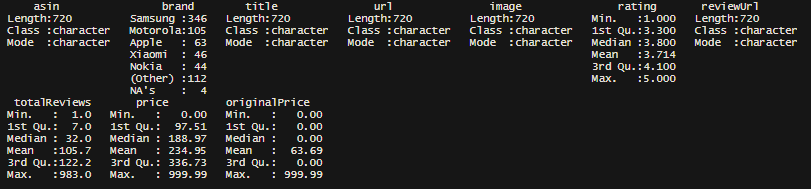
ADS 531



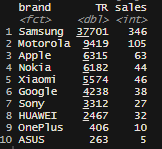
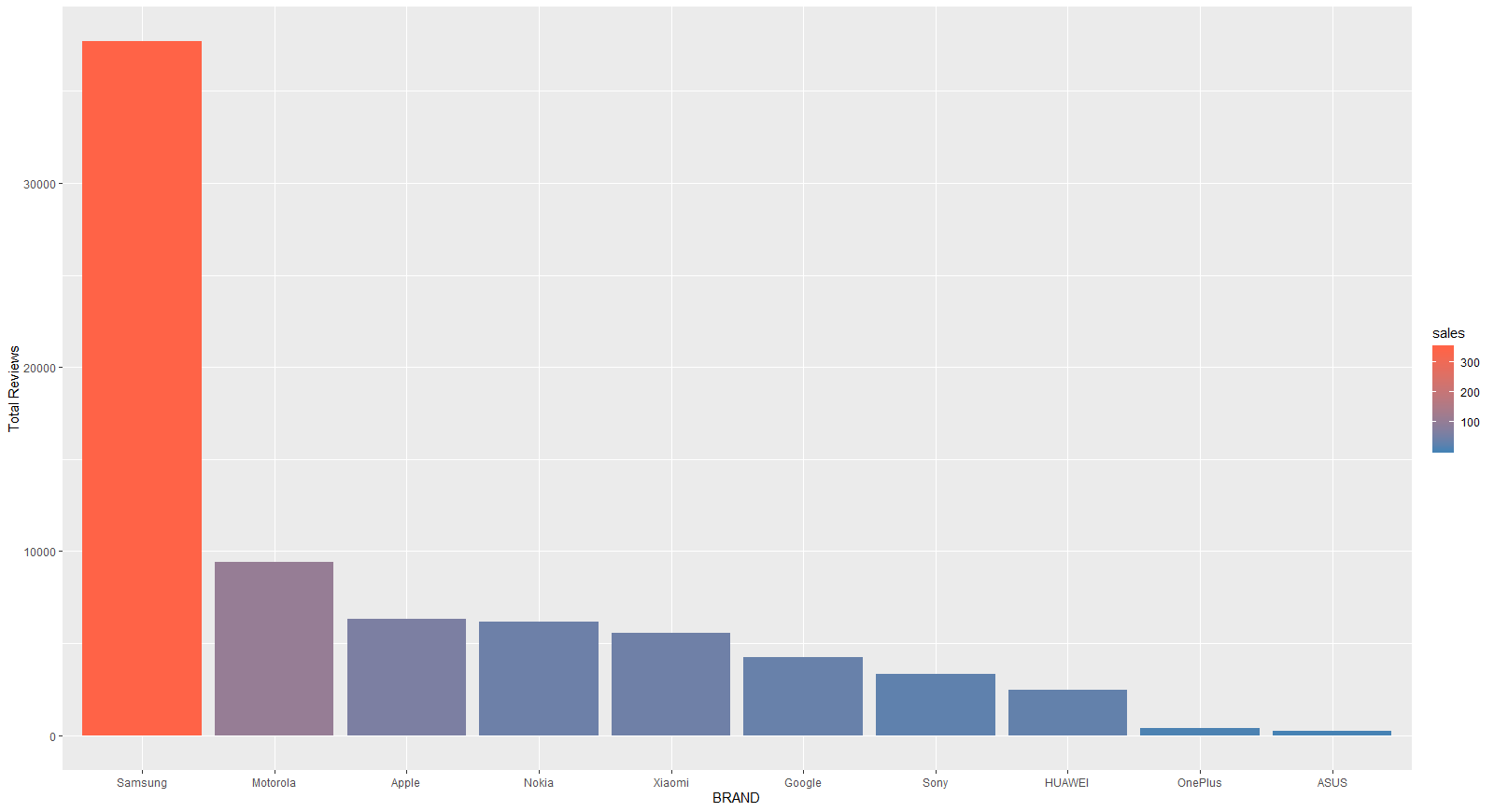
This project aims to analyze Amazon cell phone reviews and items datasets which are taken from Kaggle. In this paper, I tried to use descriptive and sentiment analysis for dataset. It is aimed to answer different questions about datasets and mainly I tried to show that how sentiment of reviews effect brands’ ratings and success with using sentiment analysis tools.

# Descriptive Analysis

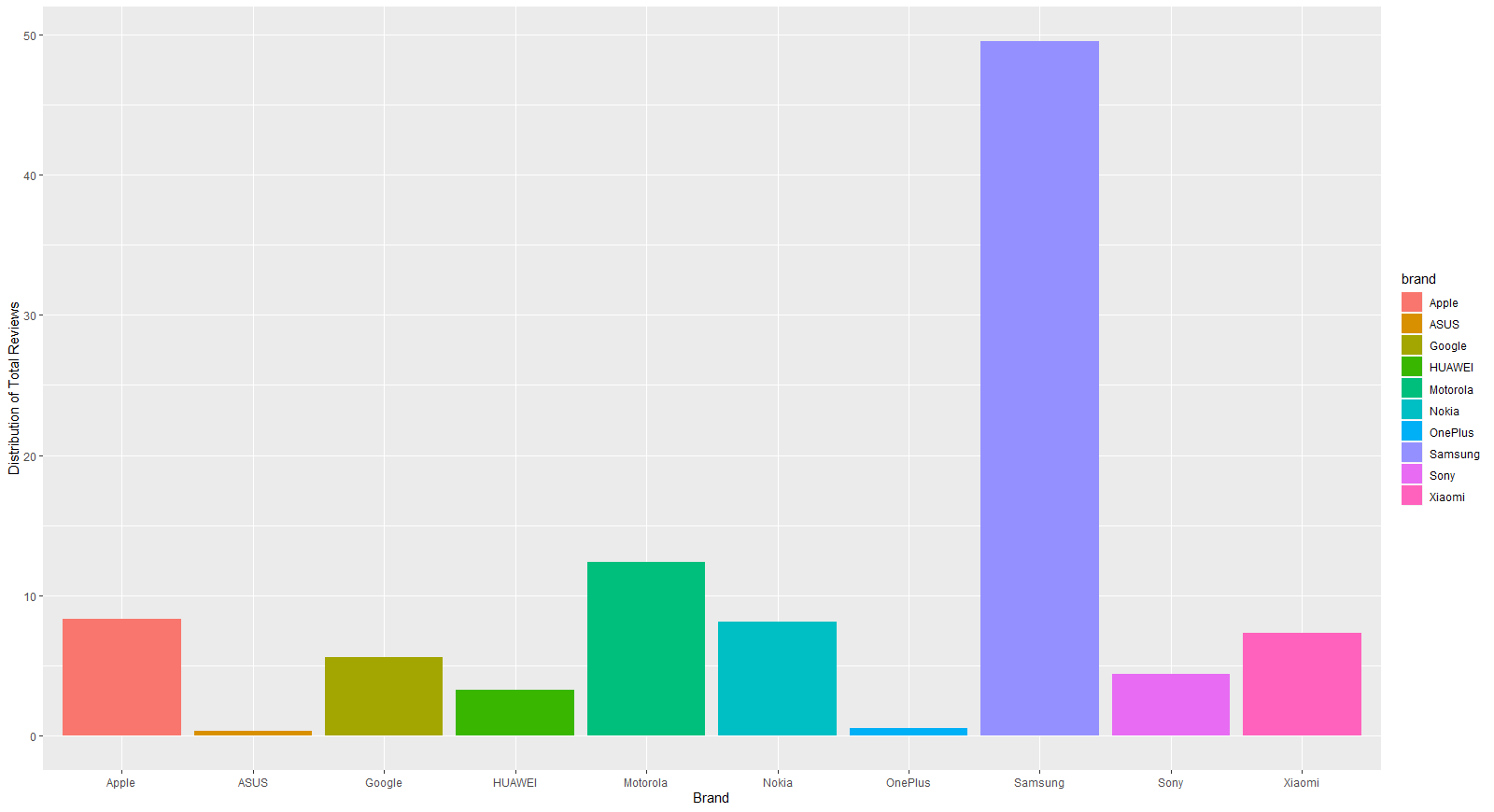
Before the start to descriptive analysis, we need to understand our datasets in general. For items datasets, there are 720 different notices in our datasets. Datasets includes 10 different cell phone brands. The variables like url, image, reviewUrl were not used for my project. These variables don’t have any importance for our project. Then I extract them my dataset.



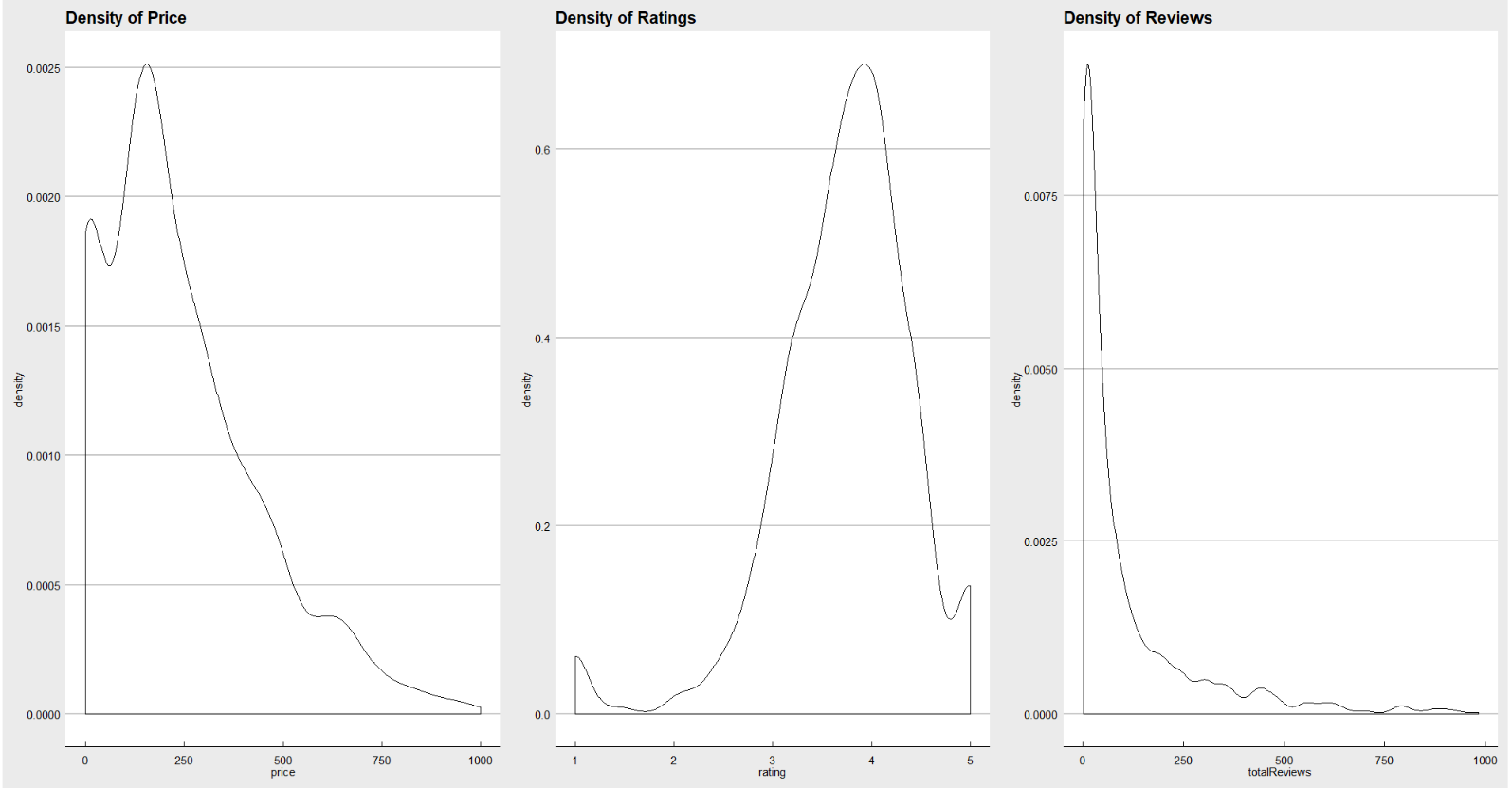
Then I check the how the total reviews and notices distribute between brands. I expect that, total reviews and notices(sales) should have positive correlation.



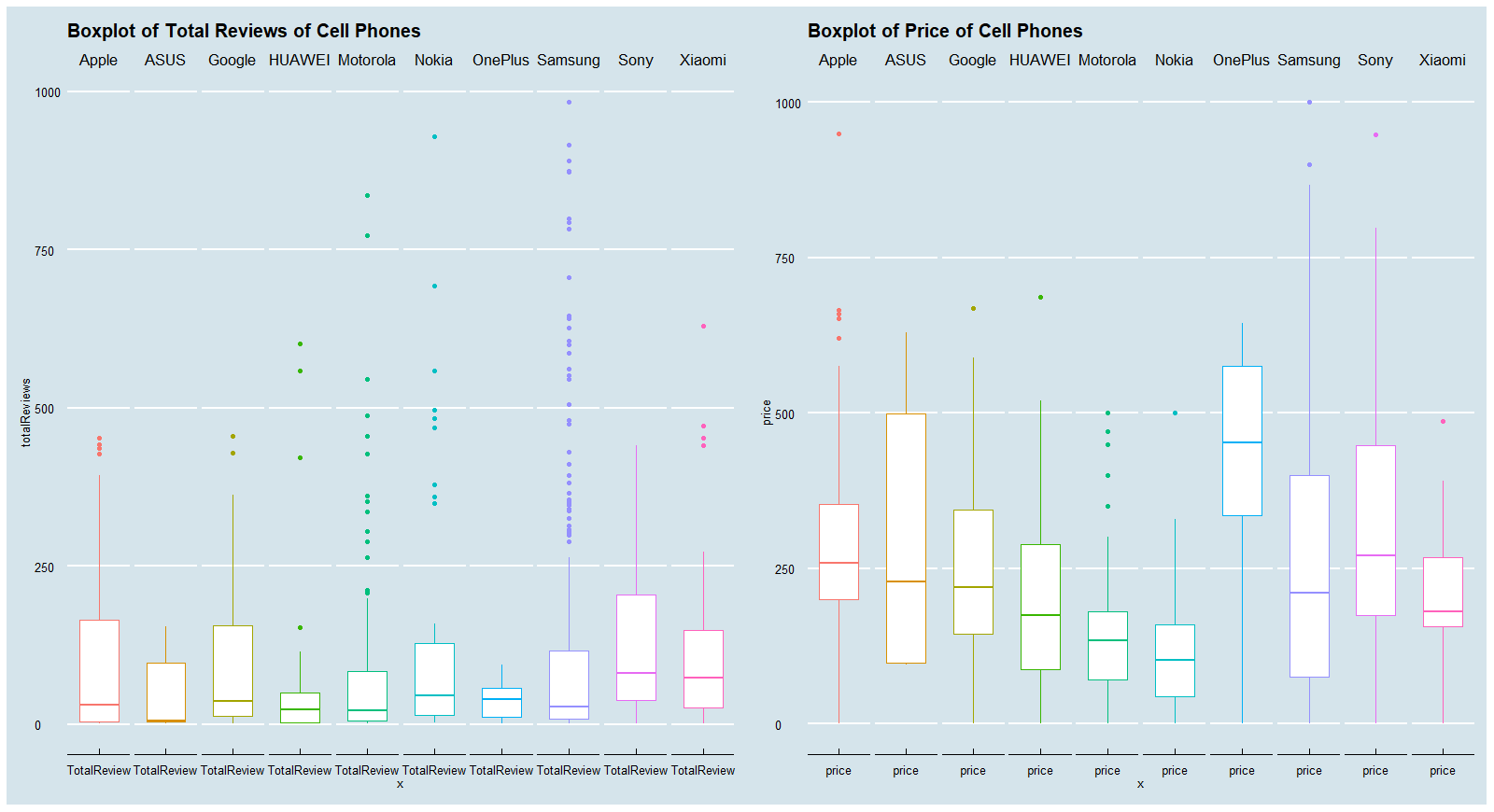
These table shows that, Samsung has higher reviews and sales. Thus, when it is examined generally, it can be said that, when total reviews increase, notices increase. After that, it can be said that, total reviews and sales are positively correlated with each other. The graph shows it with different way. The colors represent the notices of brands and bars shows the number of total reviews. Below graph also the distribution of total reviews in percentages. Samsung has the half of the total reviews. Besides, the percentage of ASUS and One Plus is very low.



After that, I want to understand the distribution of variables. To analyze that I used density graphs for each variable. After density graphs, price is nearly left-skewed, and prices are mostly between 180-350$. The mean and median of prices are respectively 234.95 and 188.97. It is also described this issue in different ways. In addition, total reviews are left-skewed. Even, mean and median of total reviews 105.7 and 122.2, the maximum number of total reviews for one product is 983. There is outlier issue in total reviews. It also means that, some of the product are higher demand than others. In addition, mean and median of ratings are 3.8, 3.714. With density graph of rating, we can say that most of the customer satisfied the phones in Amazon.

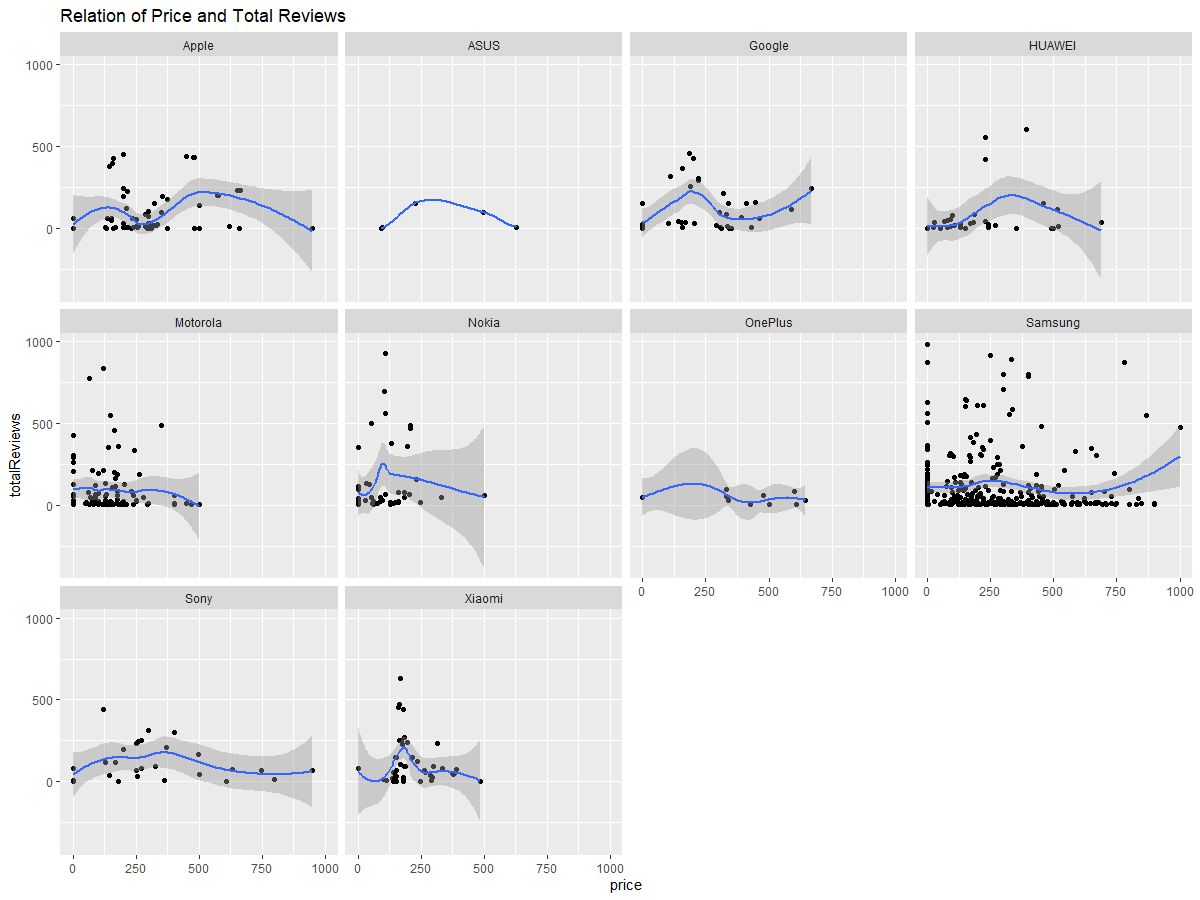


Because of the distribution of variables, I want to understand the outliers. To do this, I plot some boxplot graphs.

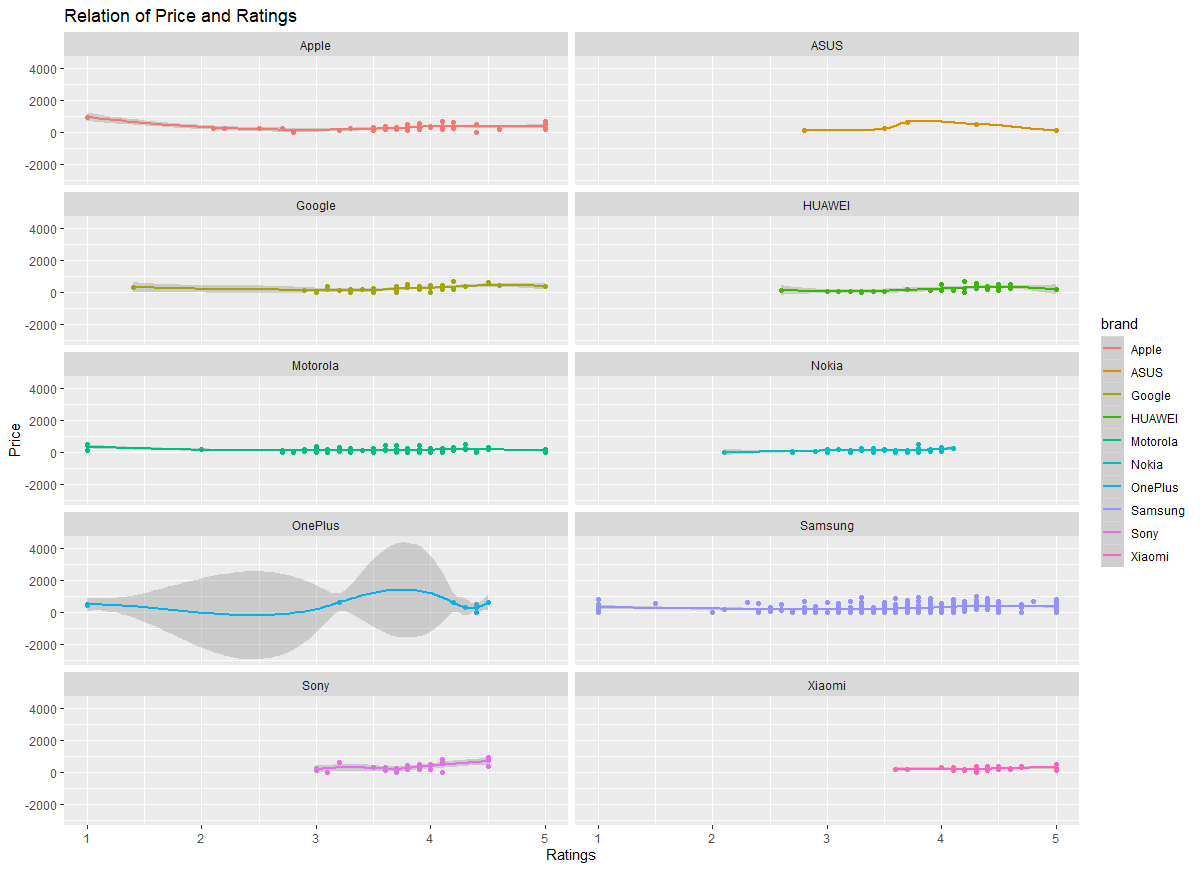


In this graph, we can see that there are significant outliers in Total Reviews, especially in Samsung. Also, generally, lots of firms have similar total reviews. However, because of some product, there is an important difference between brands. Even the prices include some outliers, it is not high as much as total reviews. The price level of firms differs from each other. For example, OnePlus have higher prices than others. Also, there are 13 products which have higher prices than 750$. 10 products of them is related with Samsung, 2 is from Sony and there is a phone of Apple.

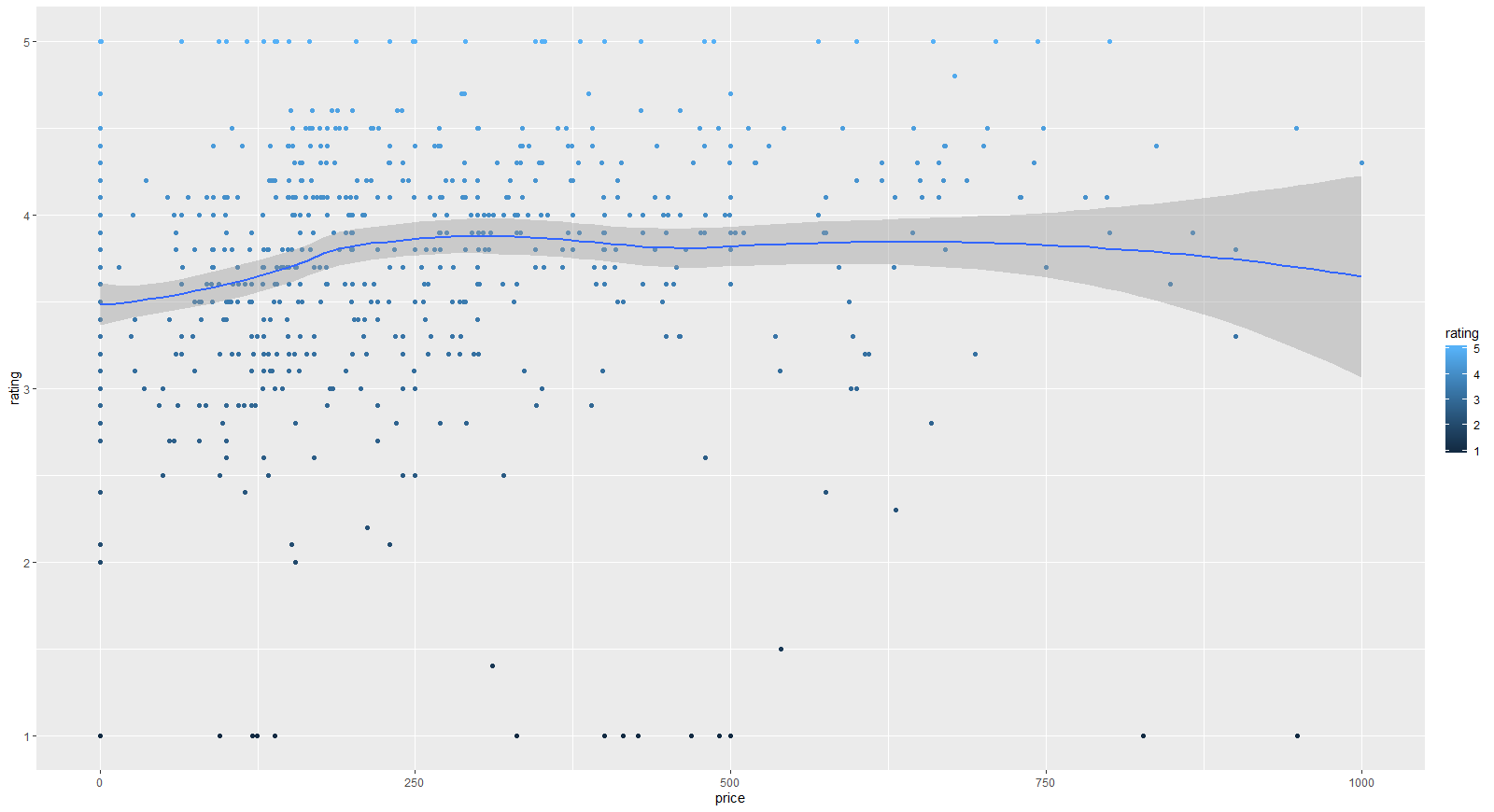
After these, I tried to check the relation between price and total reviews in brand level. Graphics shows that there is no relation between price and total reviews. However, in Motorola and Xiaomi after a point there is decrease in total reviews when price increases. It can be related with the income level of people. Because of the higher prices, customer cannot afford these items. However, for others brands the situation is always changing. Thus, it cannot be said that there is a relation between price and total reviews.

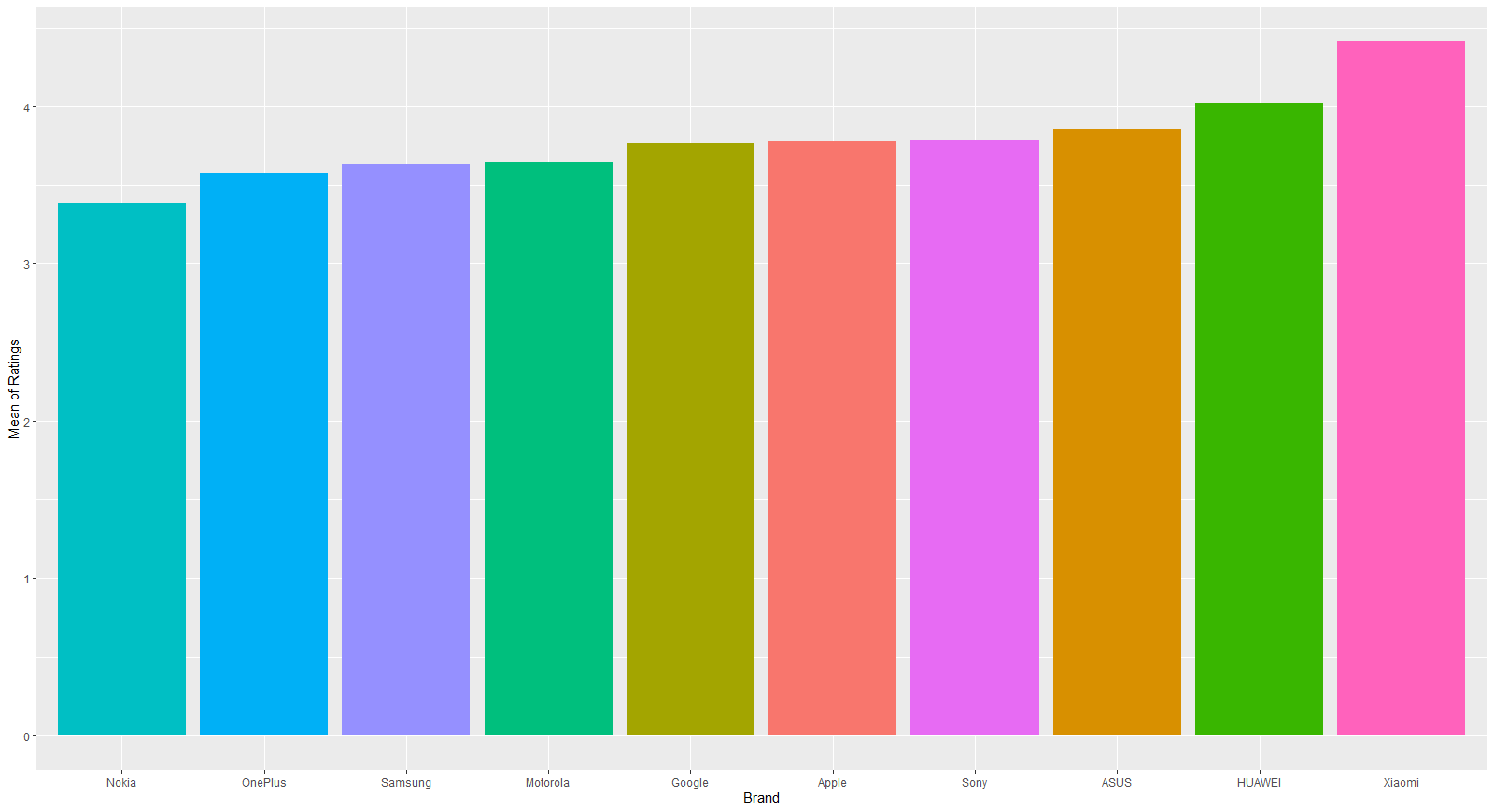


In addition, I expect that ratings and prices can have relationship in many ways. Thus, it is controlled the relation of price and ratings by brand level. However, it is seen that there is no relation between both variable. In addition, graph also shows the distribution of the ratings for each brand. Xiamoi and Sony have higher ratings than others. There are no ratings which is lower than 3 for these brands.



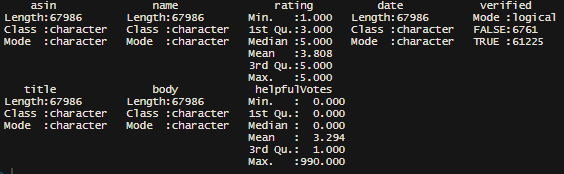
If we check the relation in all brands, again we cannot see any significant relation between price and ratings after the mean. However, until the mean there is slightly increase in ratings. Below graphics shows the ratings and prices for all brands. It can be seen ratings are distributed between 3 and 4. In addition, most of the products which prices are higher than 500, have higher ratings value.





Then, it is examined the average ratings of brands for each notice. Xiamoi is the first place for average ratings and Nokia has a lowest value. For Samsung, even brand have highest total reviews numbers, average rating value is lower.

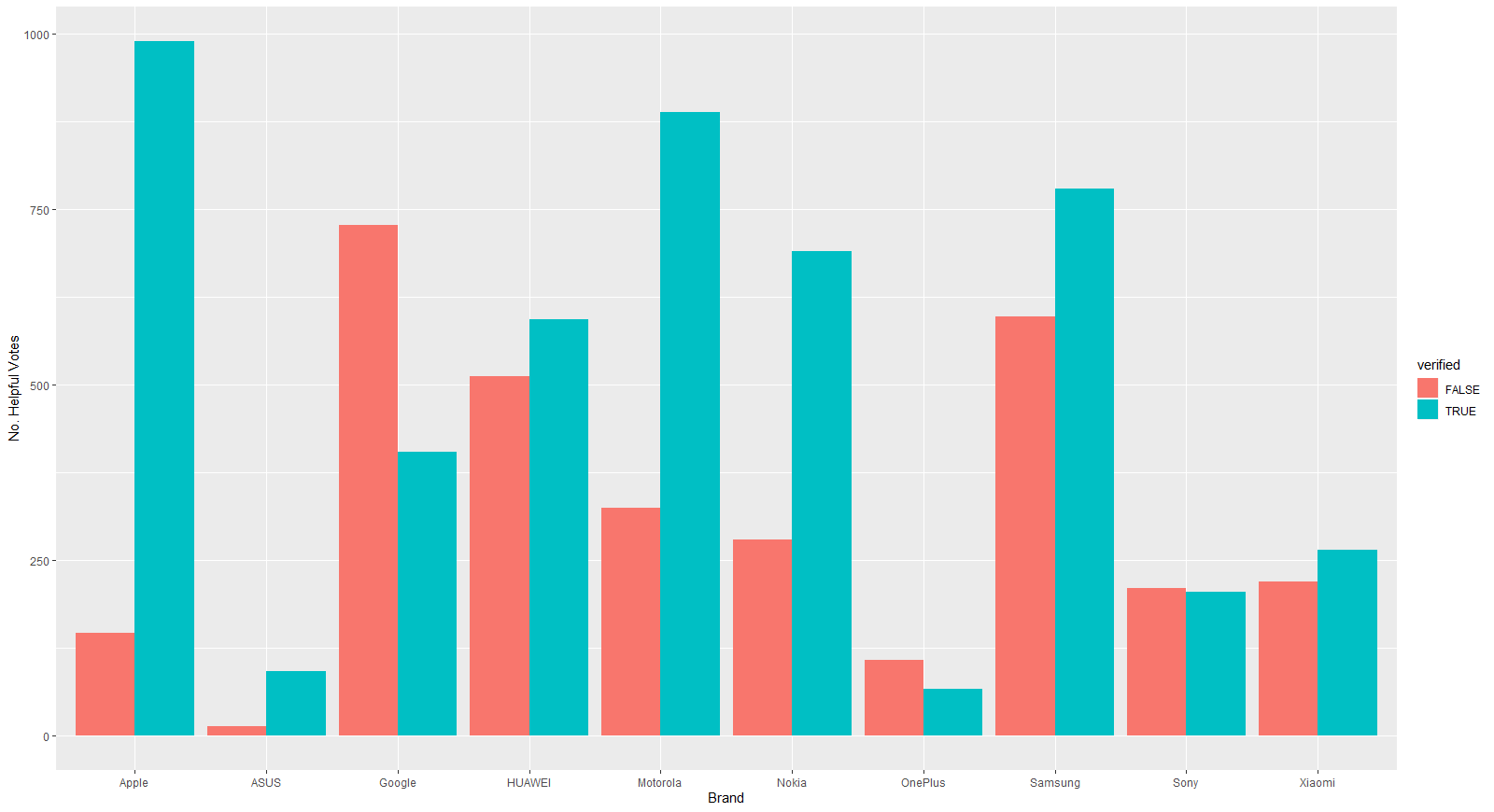
After that, I start to check reviews datasets which includes individual reviews from customers.

As you can see, date is a character format like October 11, 2005. I convert them into year and month format to analyze easily. Comments were started from 2003 and it finished at 2009. The mean of individual ratings is obviously very similar the general mean of products. Helpful Votes represent the how the comments are beneficial for other customers. Also, the mean of helpful votes 3.294 and the max helpfulvotes is 990. It is obvious that there is an outlier in helpfulVotes. When we filter the dataset, it is commented for Apple Iphone SE in 2017 and customers gave 1 point for rating. This dataset mostly used for frequency and sentiments analysis for this project because of it includes all reviews. For this section, I merge datasets to analyze them together for using helpfulVotes and dates. I expect that over years, there is a should increase in total reviews. In addition, verified votes should be higher than non-verified votes.

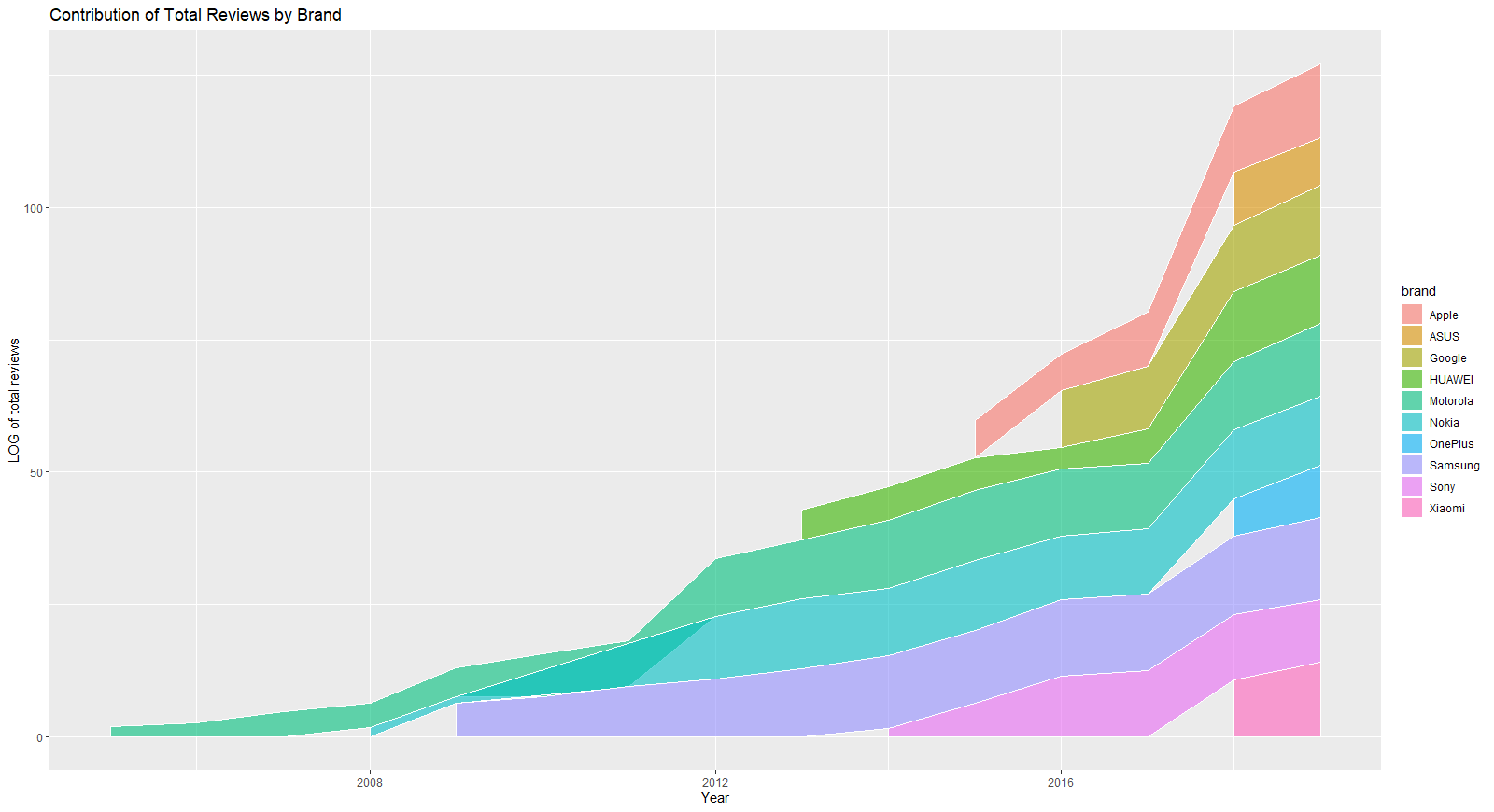
Then I want to check the verified votes and their helpful votes points. It is expected that verified votes have higher helpful votes rates than not verified votes. For this case, there are NA values in dataset for helpful votes and I replace them 0.

This table explain the frequency of helpful votes for verification by customer. People care more verified votes than others. Also, verified customers comment better than not verified customers. They show more importance than other people when they comment about product. It is also very important, because the right criticism is very important for both customer and producer.

Then I want to check this relation for every brand. For Apple, number of verified helpful votes are so much higher. It is very good for the customer who want to buy Apple. They can find effective votes for Apple products. However, only Google have higher number of non-verified helpful votes. Thus, customer should be more careful when they buy Google products by reading reviews.



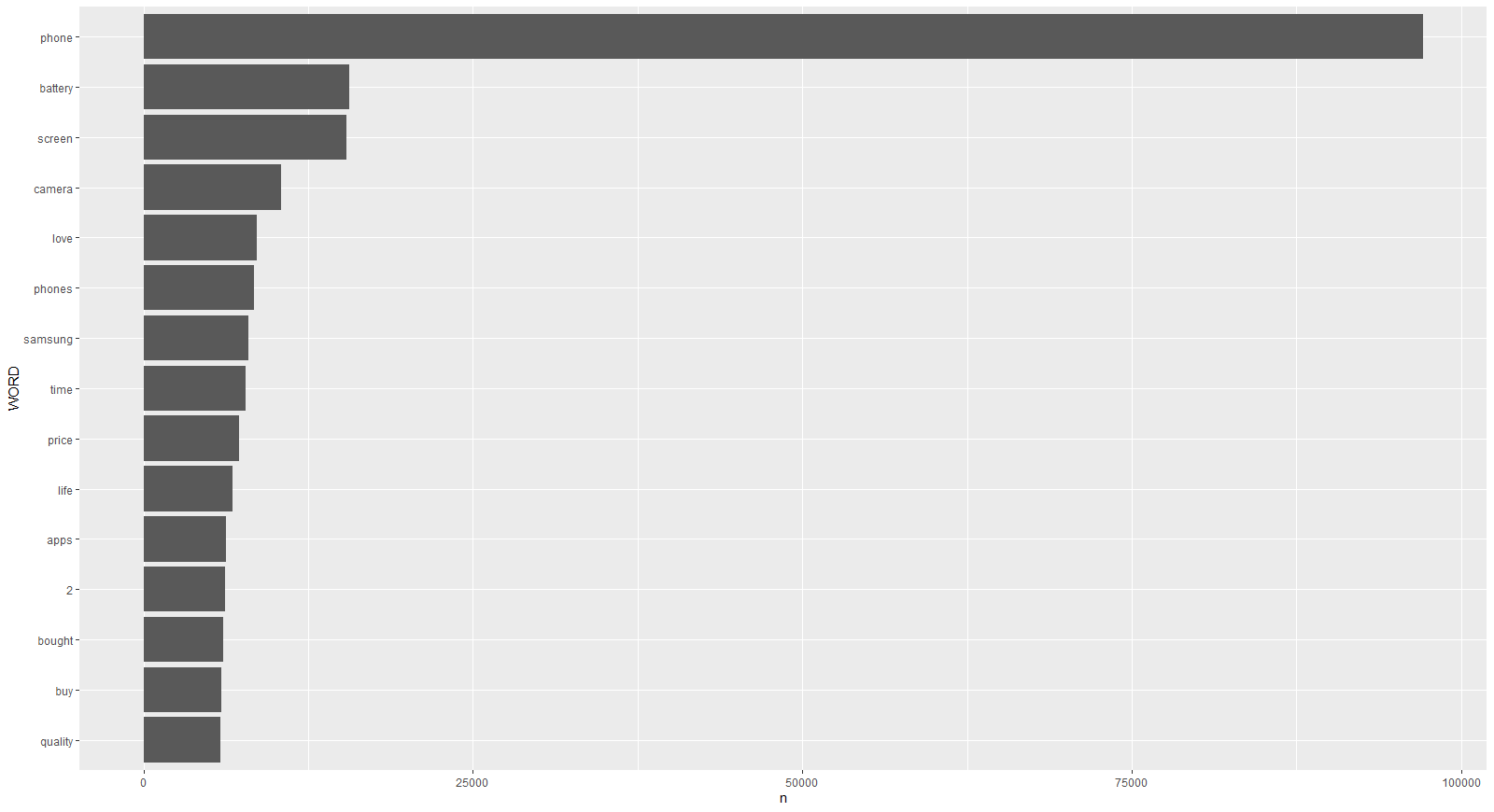
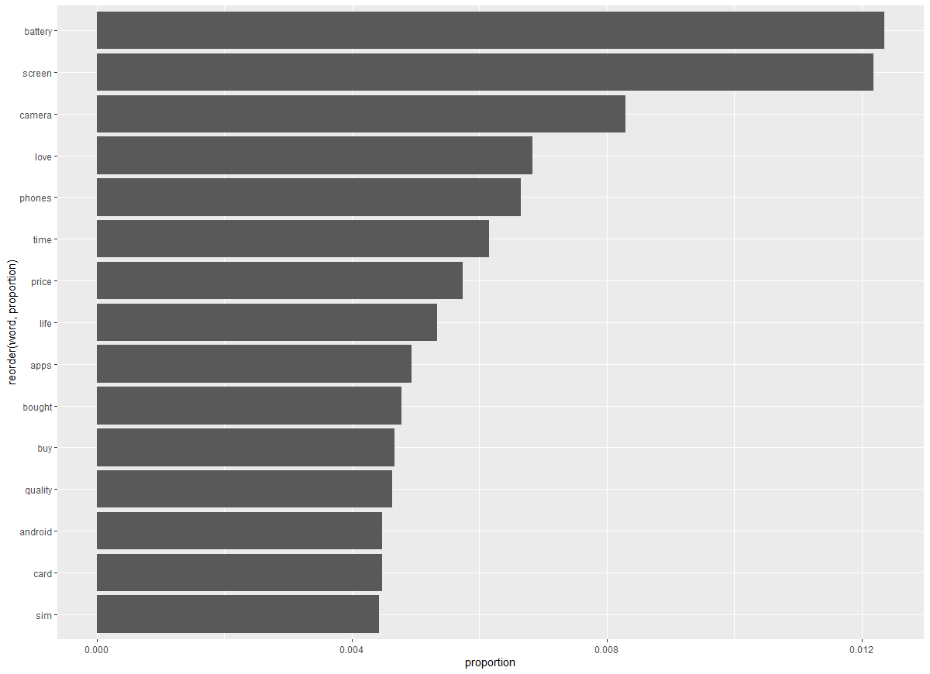
Then I check the how number of reviews changes in years with comparing brands.



This graph shows the change of total reviews between years. Obviously, year and total reviews are positively correlated. After the increase in technological development and usage of online shopping, number of total reviews are increases. It is also showed the when the brands started to online shopping. For example, reviews of Samsung started with 2009. However, reviews of Xiaomi start at the 2018. As I explained before, Xiaomi have higher average ratings than others, it can be related with them. Because of Xiaomi starts late, the technology level of products is higher, but others firm also have some products which are used old technology. However, people can comment for these products now and because of the new technologies, they cannot be satisfied from these products. It can cause to decrease in brands average ratings.

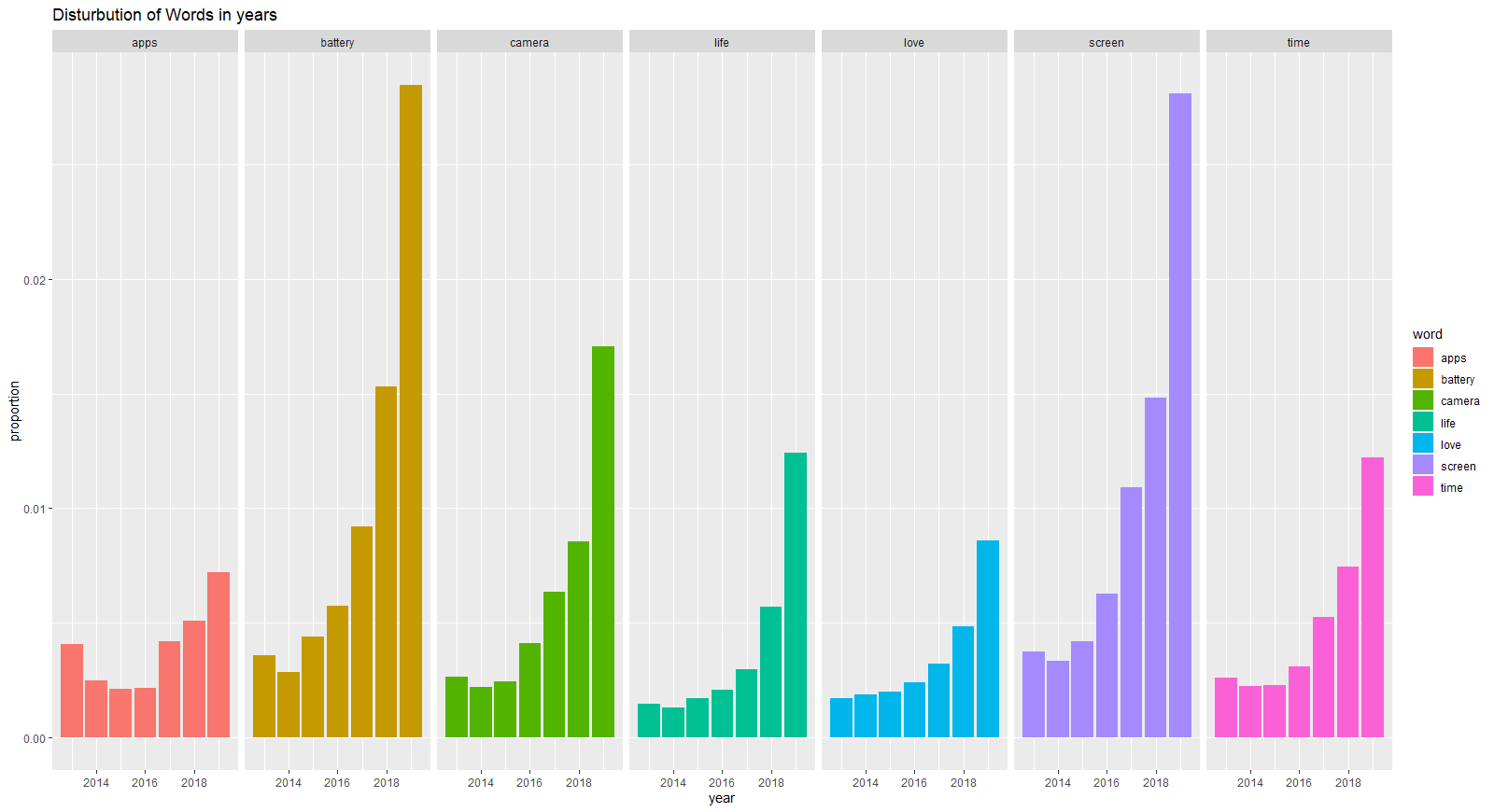
## Sentiment Analysis

To started sentiment analysis, I need to check the frequency of words to understand which words are important for me. Otherwise, there can be other stop words for this project. This graph shows the most used words in reviews. Obviously, phone is the most used word and I need to extract it from dataset because it has no meaning for us. After that there is words like battery, screen and camera which are futures of cell phones. After that, in the future steps, I want to check that how it is affected total reviews of brands.



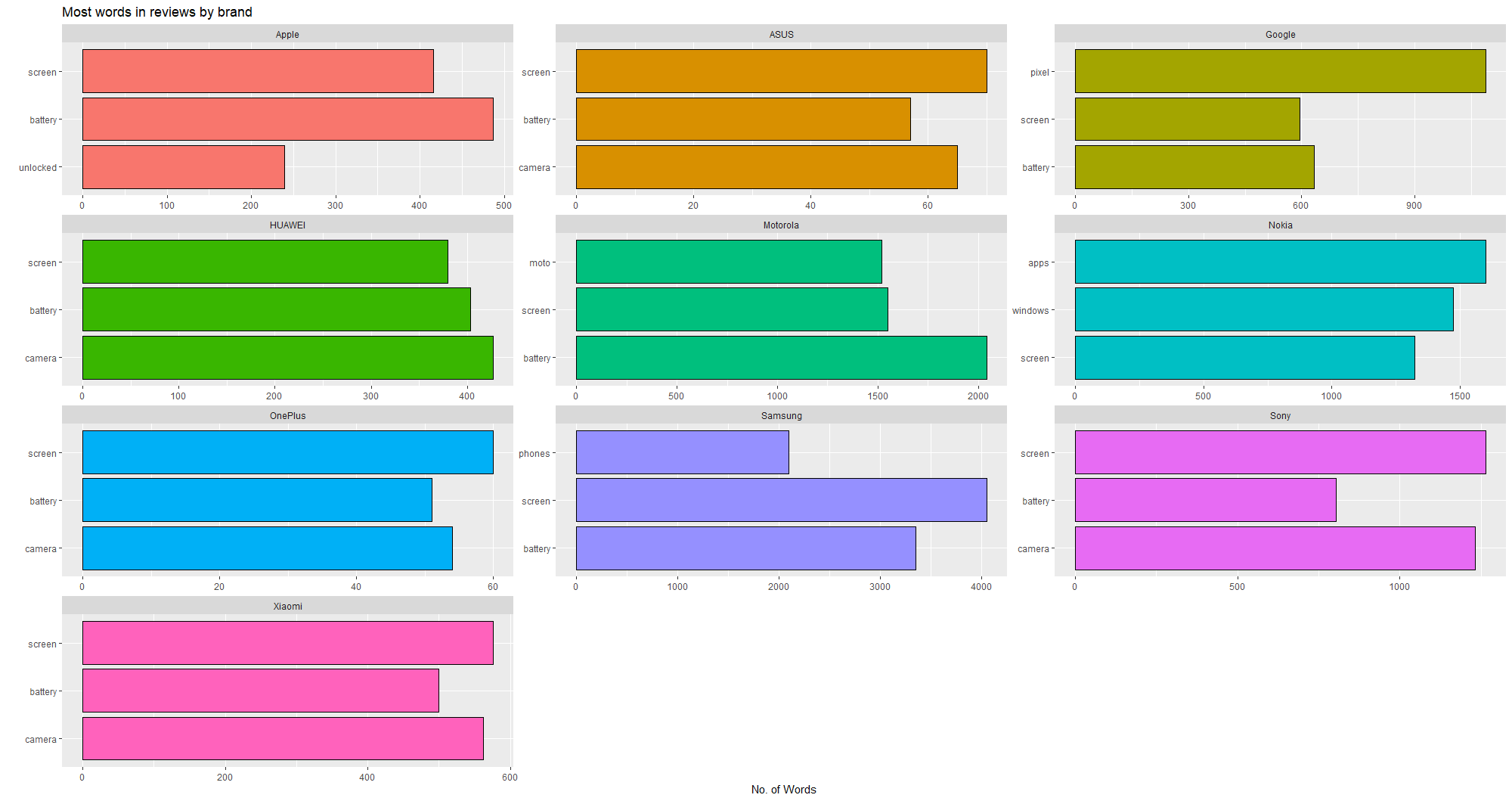
After I extract the stop words with using customization for stop words data in R packages. I again check the distribution of words.

Then, futures like battery screen and camera is very used in reviews. It can be said that these futures are very important for producer for their sales. If the brands improve their futures, it can be affected their sales positively. Then, I looked the how the usage of these words changes in years. I expect that battery should be increased after years because of new phones have lots of battery problem if we compare the older phones. Also, camera and screen should be used more year by year because in older phones there is not different screens and camera futures like today.



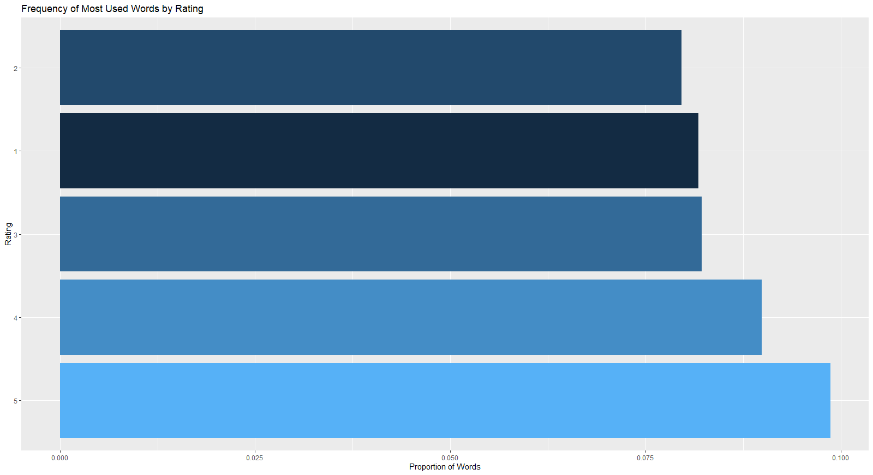
This graphs how these futures increase their importance in years. In years, camera, battery and screen increase their frequency too much, especially in last years. It can be related with the technological improvements because in the past, we cannot have these futures in our phones in different specific factors. However, in today, firms racing each other to develop these futures for increase their market share and attract the customers. This word analysis also shows the why the brands focused on these futures in mainly because most of the customers to show attention of these futures.

After that, it is checked that which words are mostly used for brands. It is important because we can understand that which words are most important for brands.



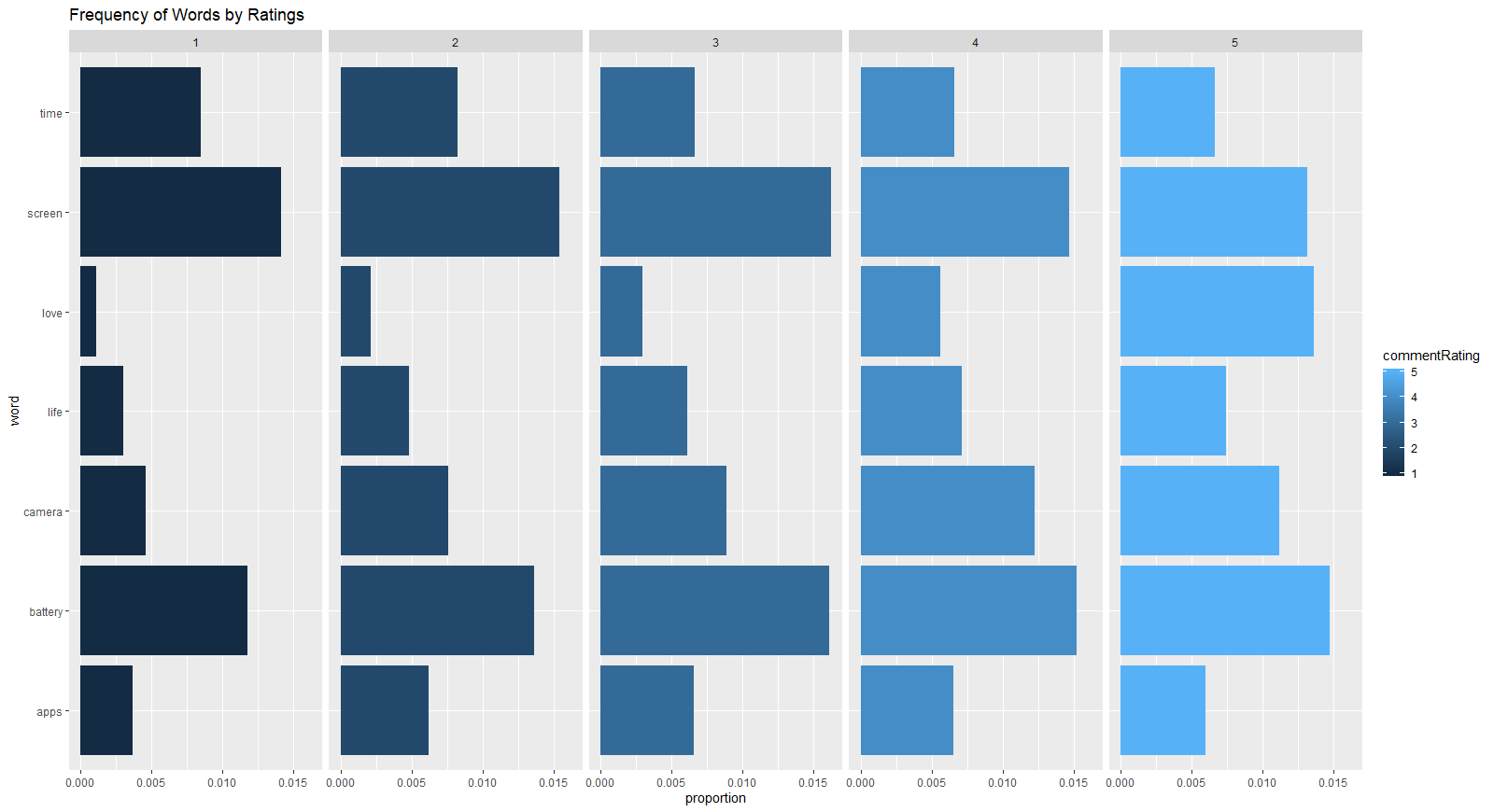
Generally, people used same words for each brand. Because of that we worked in same type products, this is very possible. Besides, there is some specific issue for Apple which is “unblocked” word. Some Iphones can blocked the before and customers need to unblock phone before using. Because of this issue, there can be problem for buying Iphones and people mentioned this issue in their reviews.

After that, I want to analyze that, how frequency of most used words effects the rating of phones. Because of that the most used words include futures of phones, there can be difference between ratings.



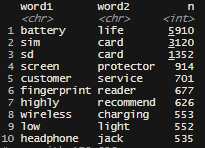
The graph shows that how the most used words distribute between ratings and we saw that they mostly used in higher ratings and lower ratings the usage of most used words is lower. Thus, people more care about these features in high rate phones. It can be related with the expectation from phones or maybe the high ratings phones were more successful about in these topics, which is more sensible.

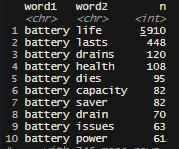
Then, I want to expect that if frequency of positive and words which are related with futures of phones increases, there should be increased in ratings. I mean that, ratings and these words are positively correlated. I checked the distribution of these words for every rating.

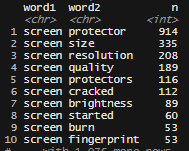


This graph proved my expectation. We can directly see that frequency of positive words increases by ratings. For example, if we look the love, we can see the direct increase between ratings. Also, there are same situations for battery and camera. Thus, firms should be careful about these futures when they produce the product. However, screen have smaller decreases with increase in the ratings. To understand these words effects on ratings we should analyze deeply.

The analyze effects of these word deeper way, I used n-grams tokenization tools. I used 2 words for sentiment analysis to see that, how the most common words effect the brands in general or specific ways. For example, I try to examine the words which are using with battery, screen and camera.

These two words are most used words in our reviews. We can see that battery life and screen protector is very important which we try to understand. Also, there is a customer service in most used words. It also tells us that buyers show importance on the customer services of brands. Then I check the futures of phones differently.

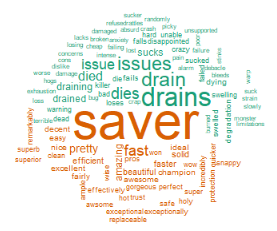
This table shows that usage of battery words with others. We can see that both positive and negative words. The life of battery and the usage time is very important consequences for batteries. Before I mentioned that, because of new technological improvements, battery times are getting shorter. Thus, people started to show more importance batteries of phones.



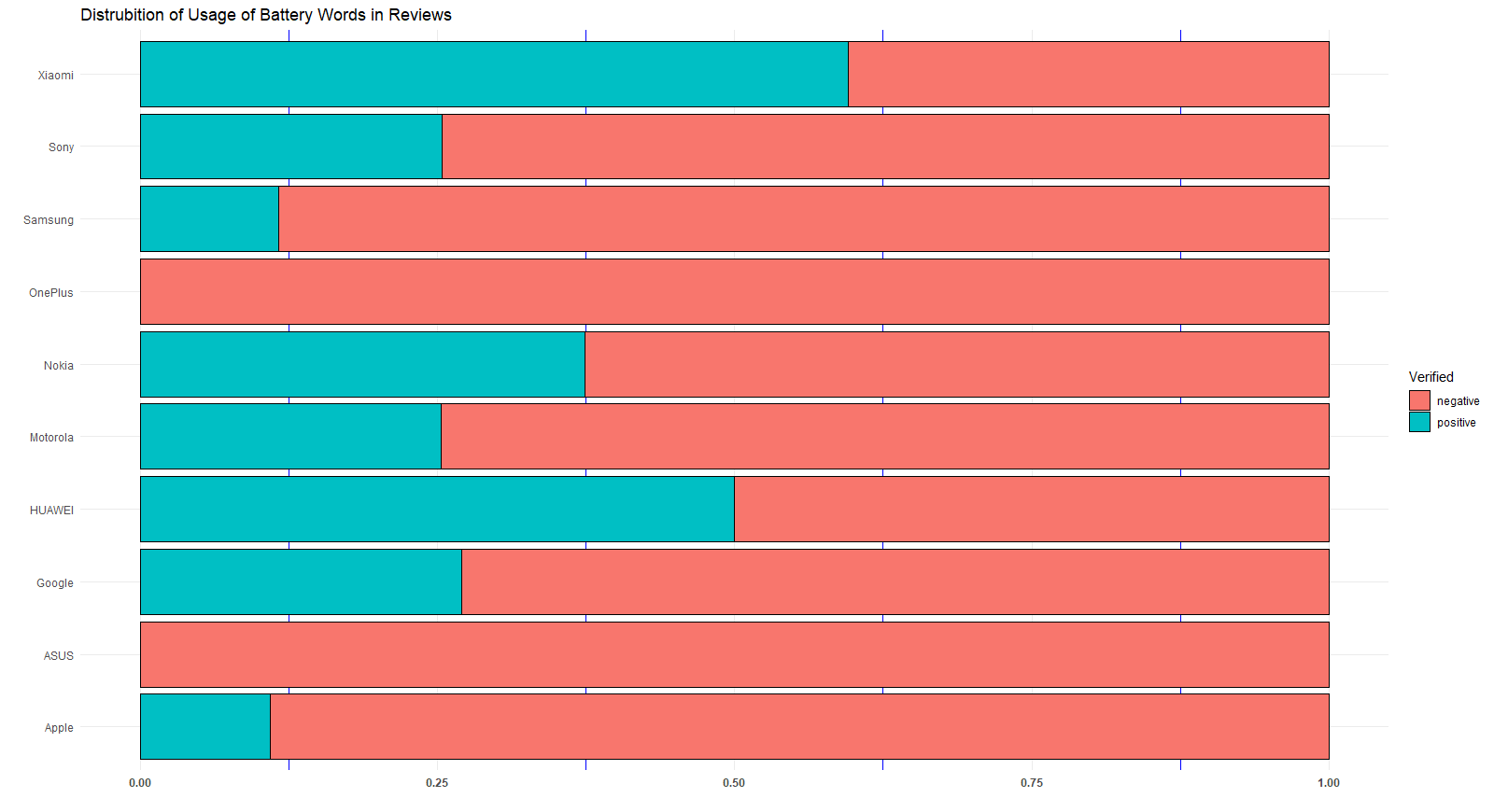
This table shows that usage of screen words with others. It is seen that the protection of screen is very important because the buying new screen is extra cost of customers and lots of people drop their phones regularly. Screen quality and resolution is the other important futures for screens. Because of there are lots of different screen type of new phones, people have lots of opportunity about this. Thus, if some brand or product have better quality about screen it can get an advantage against its competitors because of the frequency of screen words which also represent the importance level of customers.

After that I want to check that distribution of positive and negative words which are using with these words. I check the comparison the negative and positive sentiments of words which are used after battery. It is very important to understand that are people satisfied from batteries of phones.

When we check the wordcloud graph of battery word, we see that saver is most used word with battery. Most of the words are related with drain word and the also number of the negative words higher than positive words.

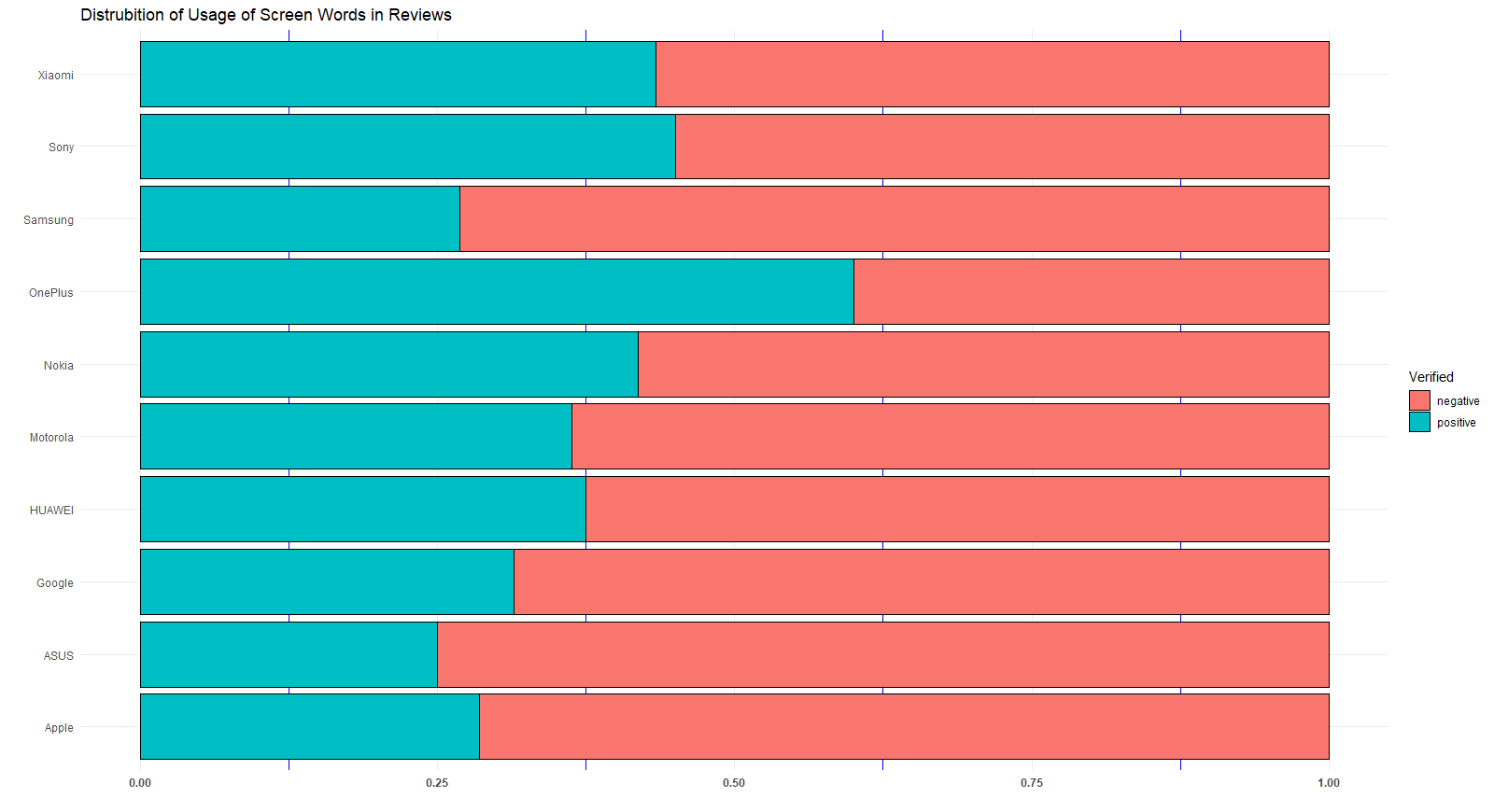


Below graph shows that OnePlus and ASUS don’ t have any positive sentiment about battery. Besides, Xiamoi has highest positive sentiments and before I explained, brand have higher average rating. However, we can say that, mostly people don’t satisfy from the battery of phones. It shows that, if some brands improvement their batteries of products, they can increase their market share and get important benefit from this investment.



Secondly, I used same process for screen word.

In the word cloud graph, we see that, protection, clarity, fast are most used words. In general, cracked is most used negative word.

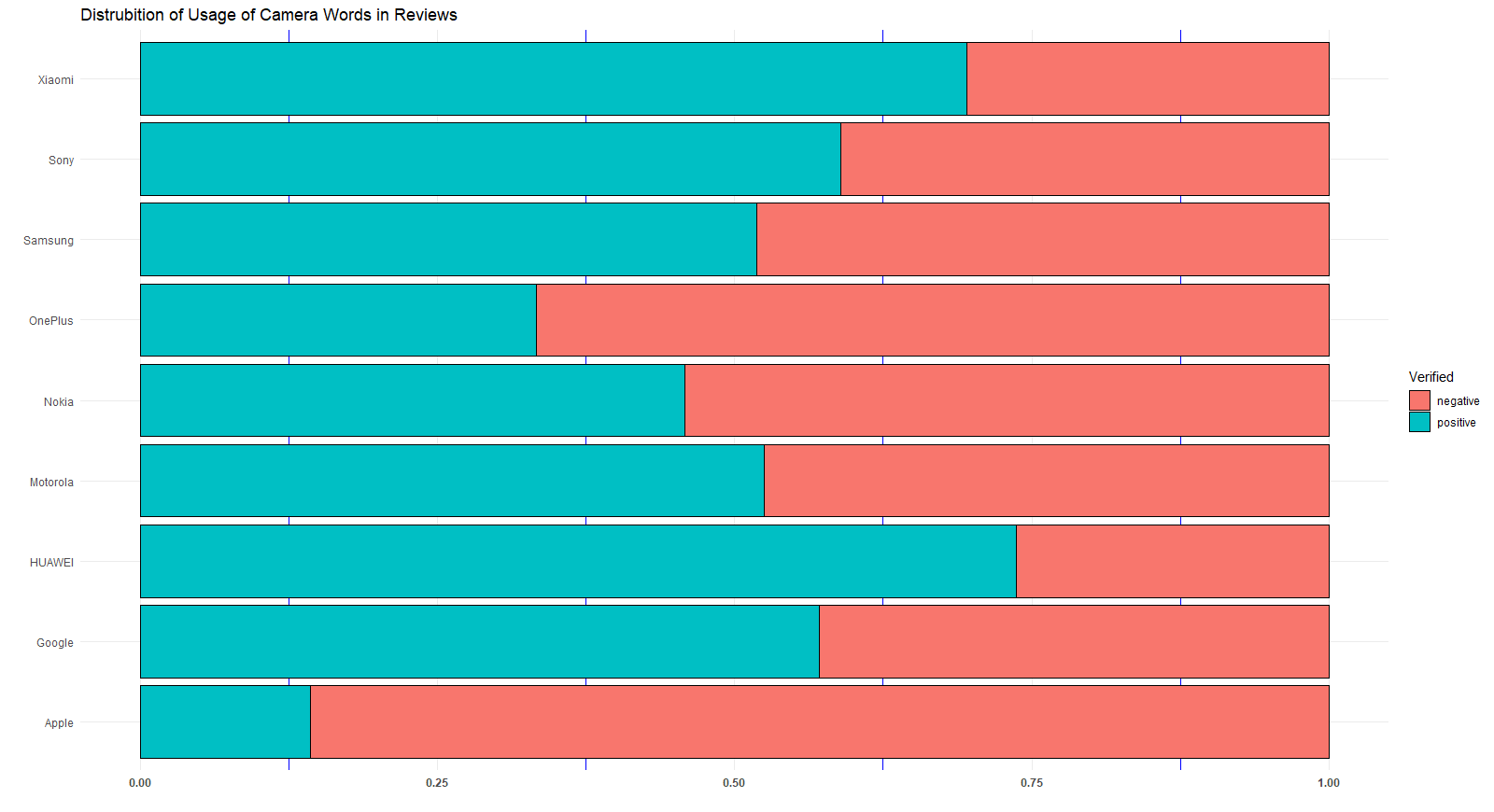


Even OnePlus don’t have any positive comment about battery, it has highest positive sentiment about screens. For screens we can said that in general, positive and negative sentiments are close to each other. The difference between them not high as much as battery. Xiamoi and Sony is very close to each other. Even Xiaomi is under the 0.5 positive sentiment about screen, in general rating is higher.

Thirdly, same process is used for camera.

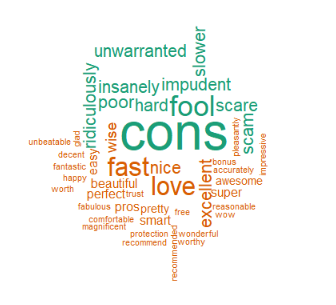


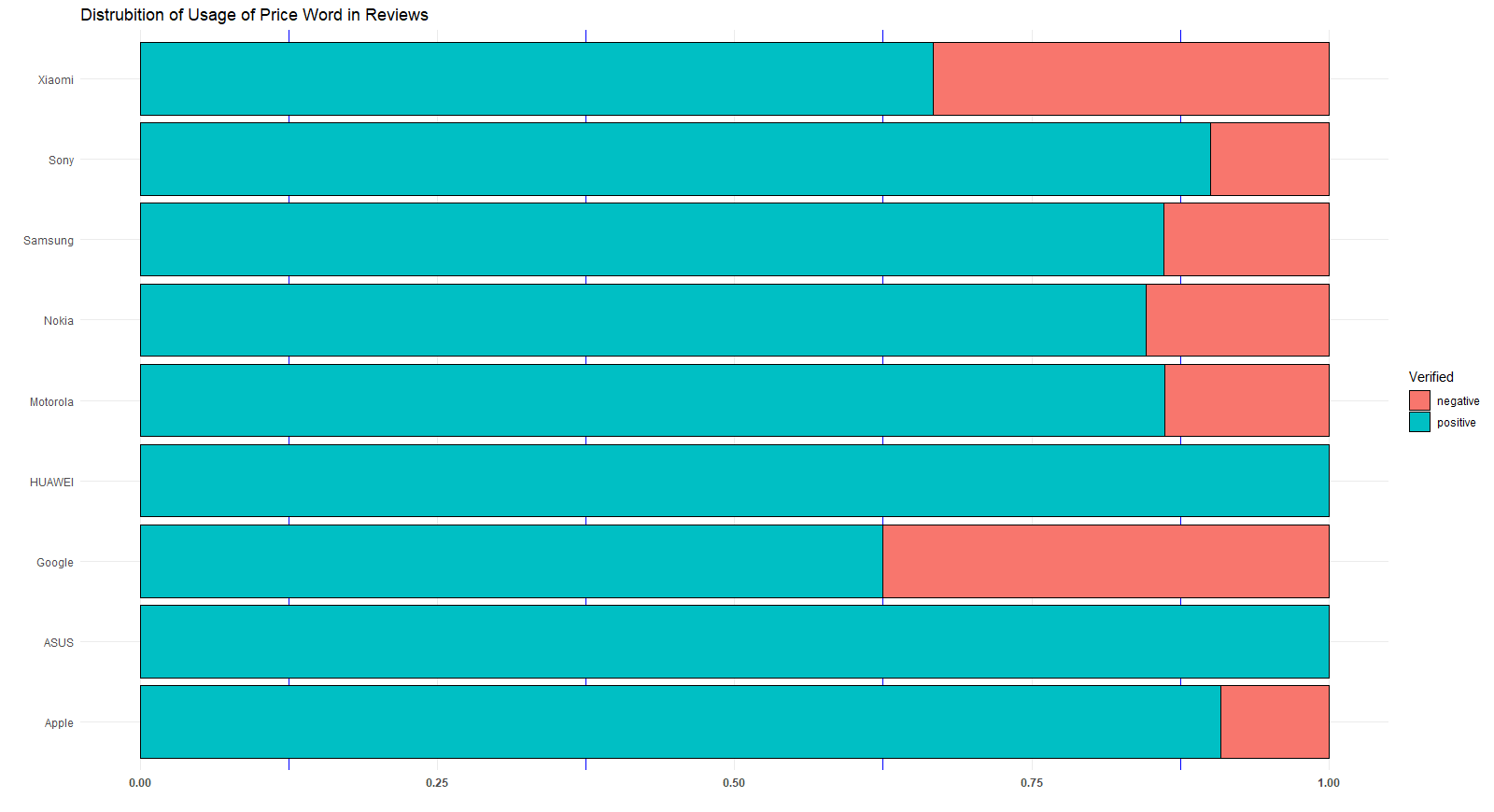
This word cloud graph shows the most used words with camera. The distribution of positive and negative words is not different than each other. Fast is mostly used positive words for camera. Bump and failed are the negative words for camera.



Most of the people are not satisfied from apple phones cameras. The most successful brands are Xiaomi and Huawei. In addition, people are usually satisfied from cameras of phones. Because of that Apple should show more importance about their camera technologies.

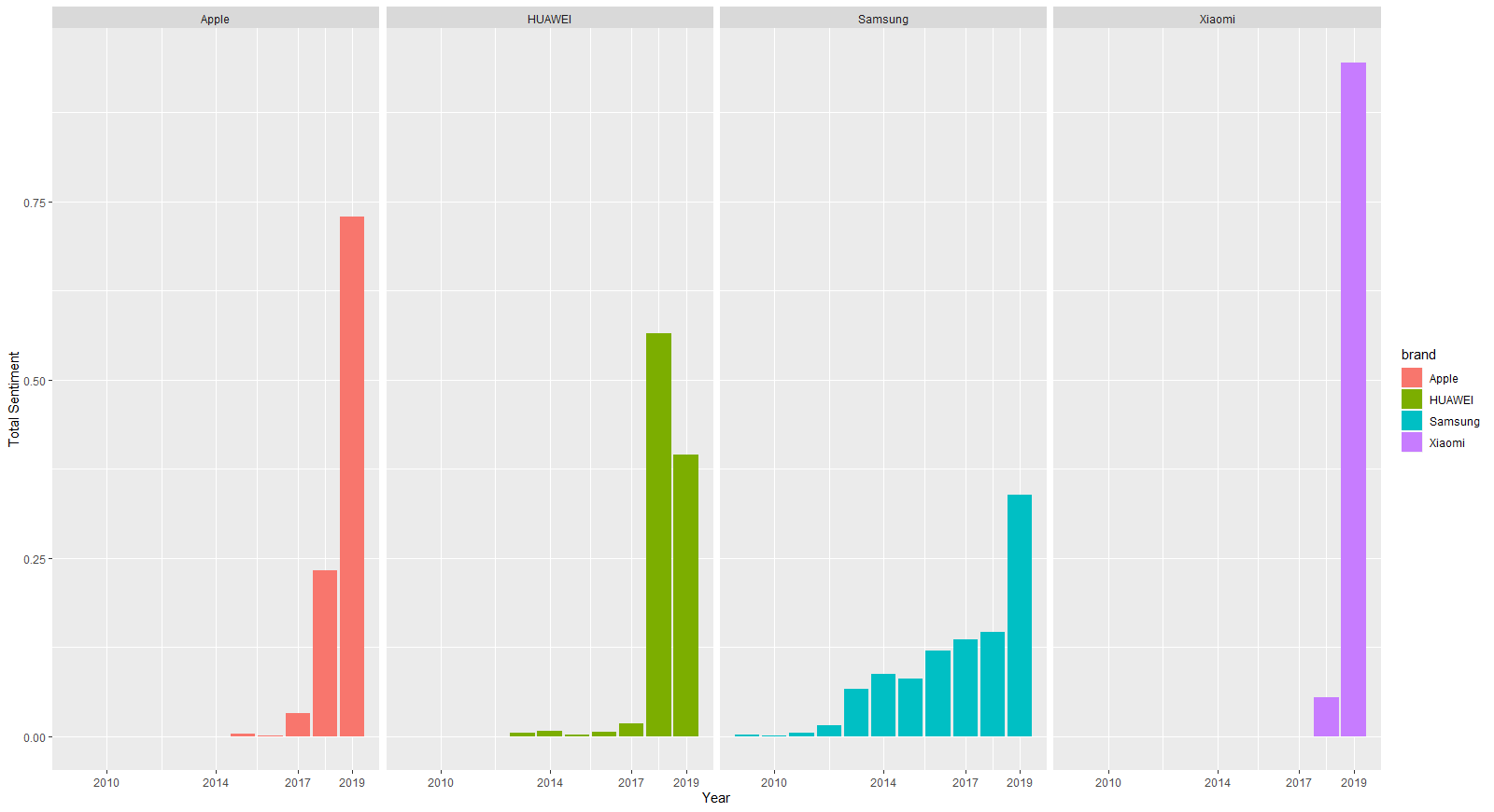
Lastly, same process is used for price. Because of the prices of phones are very close to each other and distribution of prices are not high even there are some outliers, I expect that most of the people should have positive sentiment about the prices of products.

These are the words which are used with prices.



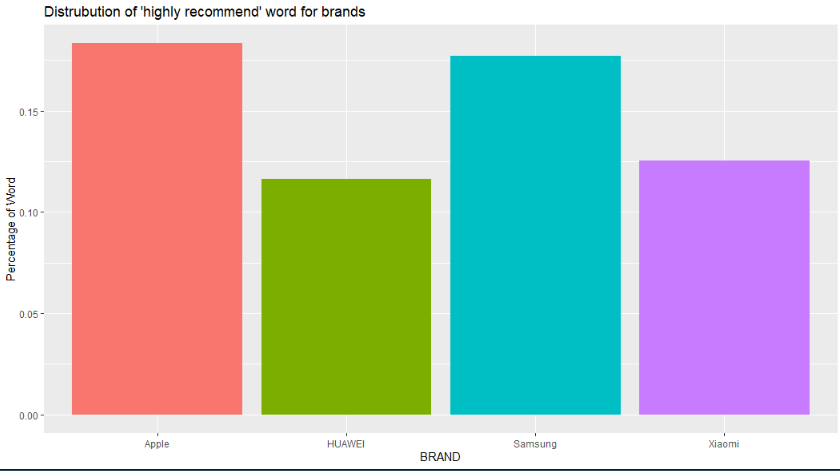
As expected, most of the people satisfied from prices. However, even Xiamoi has very successful about futures of phones, people are not happy about price of products. For Huawei phones, every customer has positive comment. Even Apple phones have lower sentiment about future, people have satisfied of prices.

After that, I want to analyze 4 different brands which are Apple, Samsung, Xiaomi and Huawei. I choose them because of some specific qualifications. For example, even Apple have some lowest sentiments values about futures, people are satisfied, and the total reviews number is high. Samsung has the highest total reviews and there are lots of different products in market. Xiaomi and Huawei are chosen because of the success in futures. To do this, I used afinn sentiments and I get the proportion of total sentiments values of brands for each other. I expect that with years, there should be increase in sentiments levels because of the development in futures.



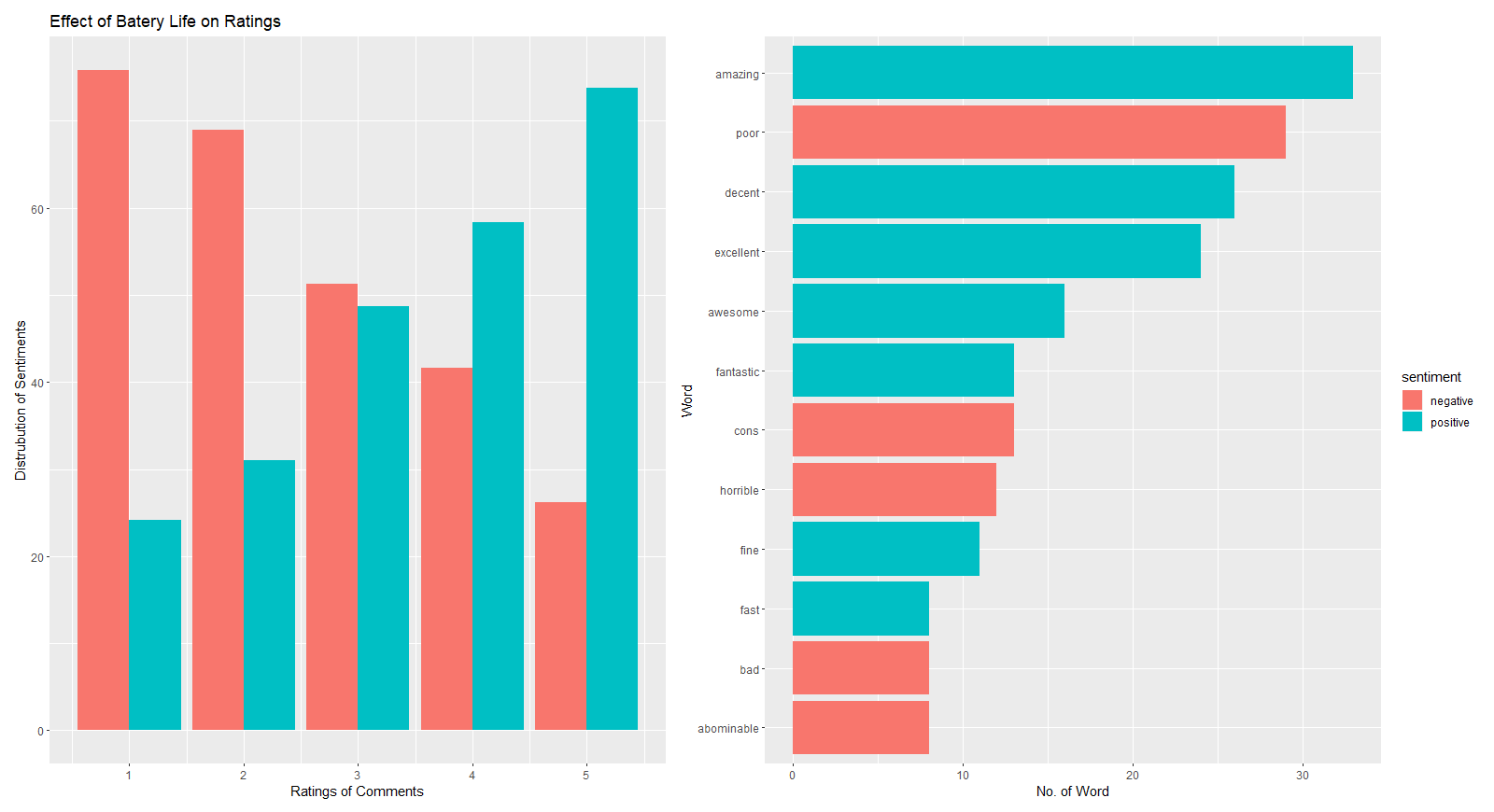
There is increase in total sentiments of all brands expect the 2019 in Sony. After the 2018, positive sentiments are decreased for Huawei. Unexpectedly, Apple always increased their sentiments values, even people are not satisfied from products features. It can be quiet success for Apple. With year 2019 Samsung increases positive sentiments very importantly. Furthermore, even Xiamoi is new in the market, after year, total sentiments go up with huge amount increases. It can be related with the there is no old for phone for Xiaomi and probably, all phones used new technological developments. However, for other brands, there are some older phones in the market and people can not be satisfied from them because of the increasing expectations.

After that, I want to check the “highly recommend” phrase for these brands because of also highly recommend one of the most used words in the beginning. It is also very important because people do not recommend things which they do not like. That is directly show that are people satisfied from phone or not? The percentage of highly recommend word for each brand shows at the below graph. I found the percentage of word with the count of highly recommend word divided by all two words using in brands reviews.



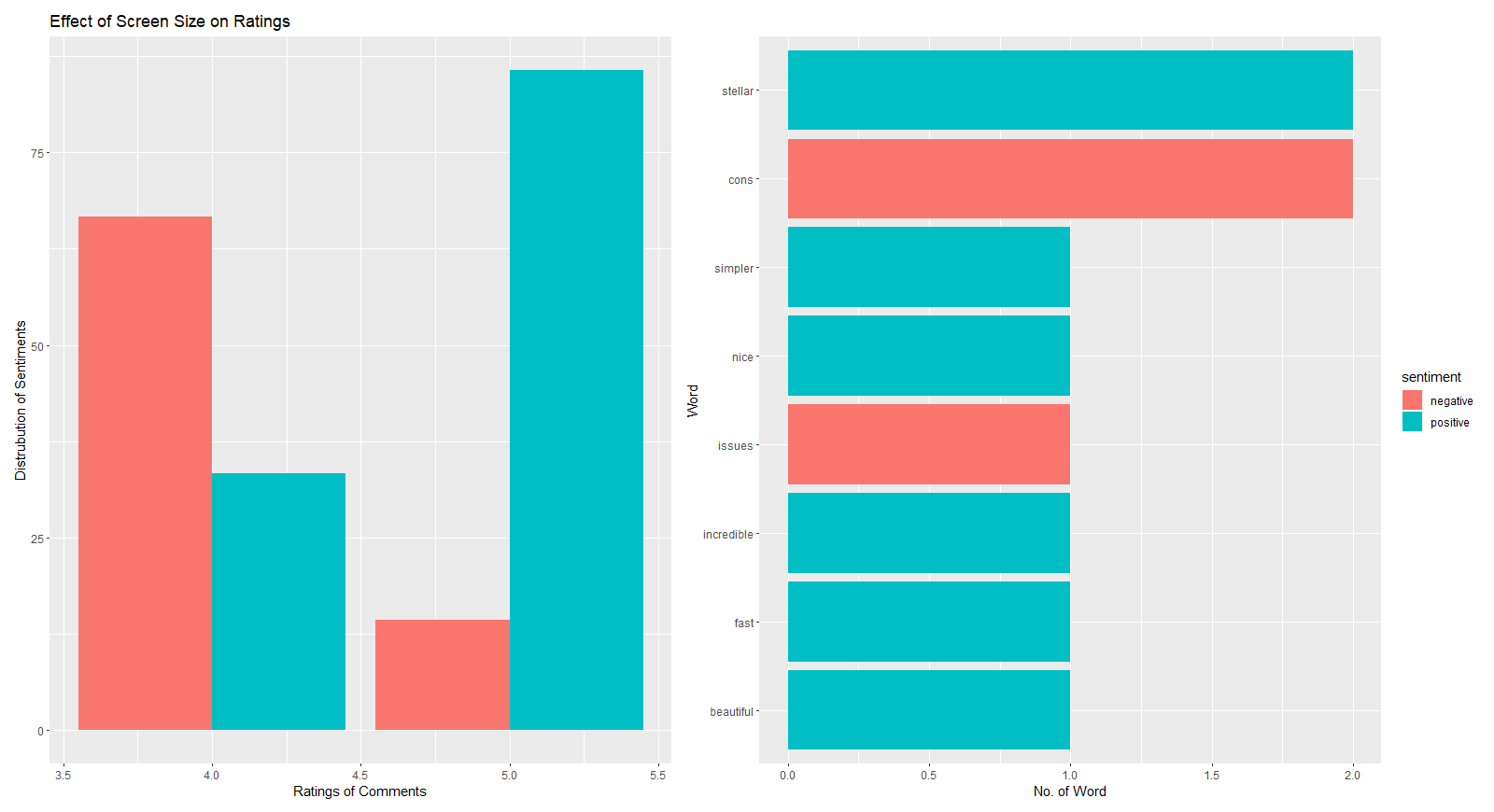
This graph tells us even people are not so much satisfied from Apple futures, they recommend phone and they like it in general. Also, Samsung have vey close to Apple and even they have very different products, in general people recommend its product.

Then I want to check that is there any relation with these futures and ratings. It is thought that positive sentiments can increase the ratings of brands. It is tried to examine in 2 words level. Firstly, I divided comments in 3 words with using n gram tokenization process. After that, I check the sentiments of battery life words which is one of the most used phrases in reviews. I analyze the words before the phrase to understand the positive and negative effect of the battery life.



These graphs show the which words are mostly used before the battery life and it is relation between ratings of comments. Most used words for battery life are amazing and poor. There are two different sentiments. It means that one of the phones have very good performance about battery life, others not. Ratings and sentiments of battery life is positively correlated. If people used negative words about battery life their ratings are lower otherwise higher. It can be seen from the graph the with distribution of sentiments, ratings are directly change. If we check the chi-square of between comment ratings and sentiments, we found that value of X-squared = 45.511, df = 4, p-value = 3.114e-09 which means the both variables are dependent to each other.

Secondly, same process used for screen resolution which is very important for the customer to choose phone as explained before and it is also one of the most used word which is used with screen.



In this graph we can see that, even there is no specific words with high usage about screen size. The effect of positive sentiments on ratings is obviously seen from the graph. In fact, even 3.5 is not have lower value for rating, the negative sentiments are higher than positive.

Finally, I tried to analyze that these 4 firms with using nrc sentiments because it has very different sentiments index for words like trust, joy and so on. It was expected that there should be significant differences between these firms, because of the average ratings and many other observations which were discussed before in the project. However, after the seeing plotting and distribution of words, these firms have only very smaller differences between each other in generally. Even they have so many different aspects, people feel same level similar sentiments.

