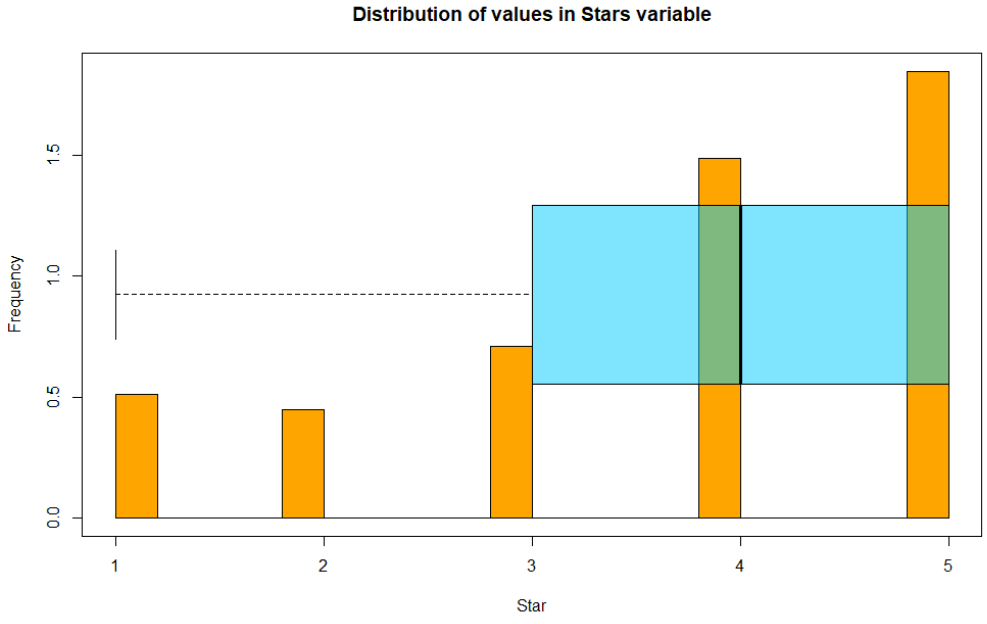
ASSIGNMENT 2

Yelp is an online review website where users can post reviews of any kind of businesses like restaurants, barbers, hotels, doctors, retailers etc. Reviews on this website has many different variables like the business id, user id, review's length, star rating, number of positive and negative words used, net sentiment, different kind of votes on the review like useful or cool or funny, and the date review was posted on. We will perform the different analysis on the Yelp reviews dataset to find:

1. Statistical summary of five variables like stars, review length, positive and negative words used and overall sentiment of the review.
2. Counts of positive and negative words in reviews
3. The overall sentiment of the reviews
4. Review's average length and which kind of average is better suitable here
5. Finding relations between useful votes and review length or number of stars
6. Evolution of number of daily reviews
7. Finding best business and user.

Statistical Summary

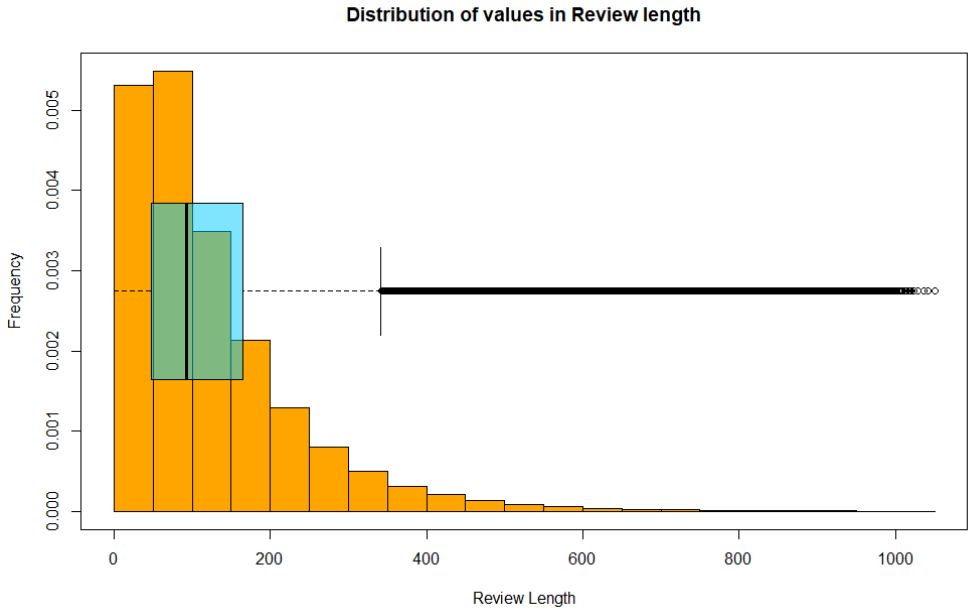
Stars




```
     stars
Min.   :1.000
1st Qu.:3.000
Median :4.000
Mean   :3.743
3rd Qu.:5.000
Max.   :5.000
```

The star category ranges from 1 to 5 in value with an average of 3.7 stars given to businesses. From the graph it can be observed that 75% of all reviews gave 3 or higher stars (1st quantile range is at 3). Generally, number of reviews are increasing with increase in star value except for 1-star reviews which can be due to angry users giving the lowest possible rating to businesses.

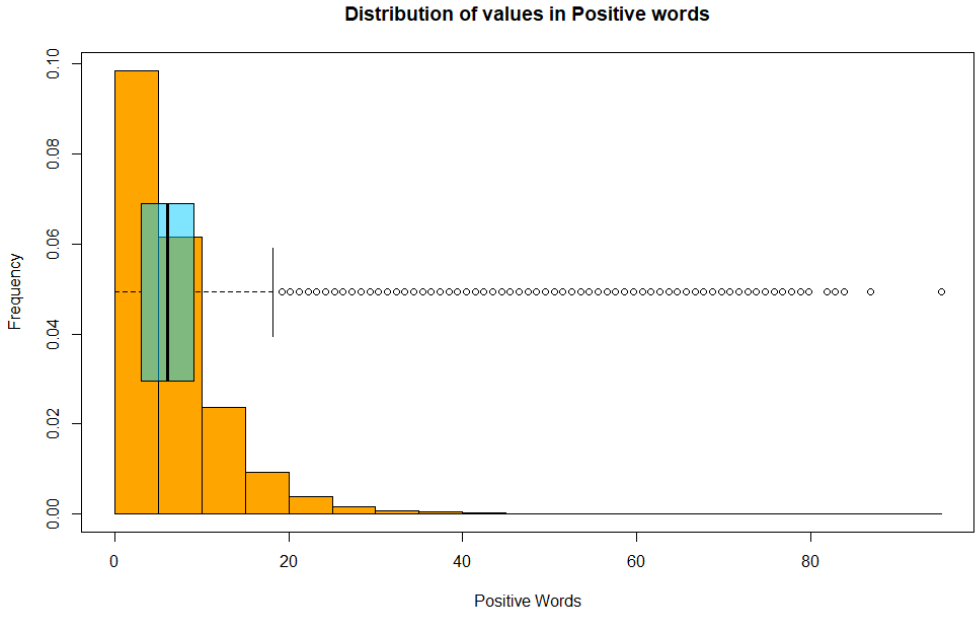
Review Length




```
review_length
Min.    :   0.0
1st Qu.:  48.0
Median :  92.0
Mean   : 125.6
3rd Qu.: 165.0
Max.   :1047.0
```

The length of reviews can vary vastly from 0 words to 1047 words. However at an average the reviews had 125.6 words and half of all the reviews contained 92 or less number of words. The third quantile range is at 165 words indicating only a quarter of all reviews were of more than 165 words. There are a lot of outliers in the distribution which can be seen in the boxplot.

Positive Words

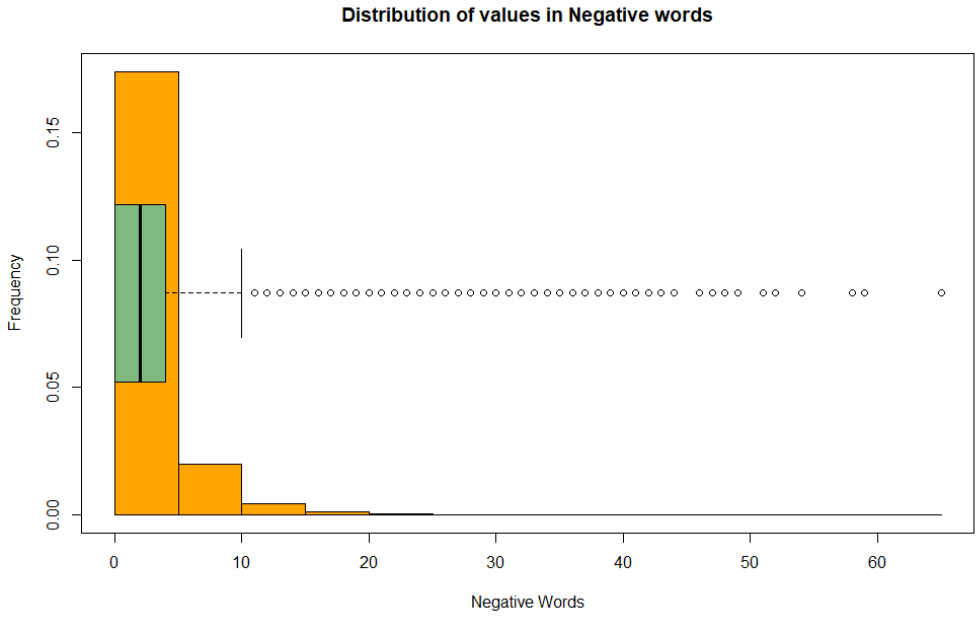



```
  pos_words
Min.    : 0.000
1st Qu.: 3.000
Median : 6.000
Mean   : 7.072
3rd Qu.: 9.000
Max.   :94.000
```

Number of positive words used in reviews ranges from 0 to 94 words with average of 7 positive words used in reviews posted. The third quantile range is at 9 showing only a quarter of all reviews have more than 9 positive words int them. The histogram and boxxplot shows the same thing where both of them

are rightly skewed indicating that majority of all reviews were at the lower end of the positive word count spectrum.

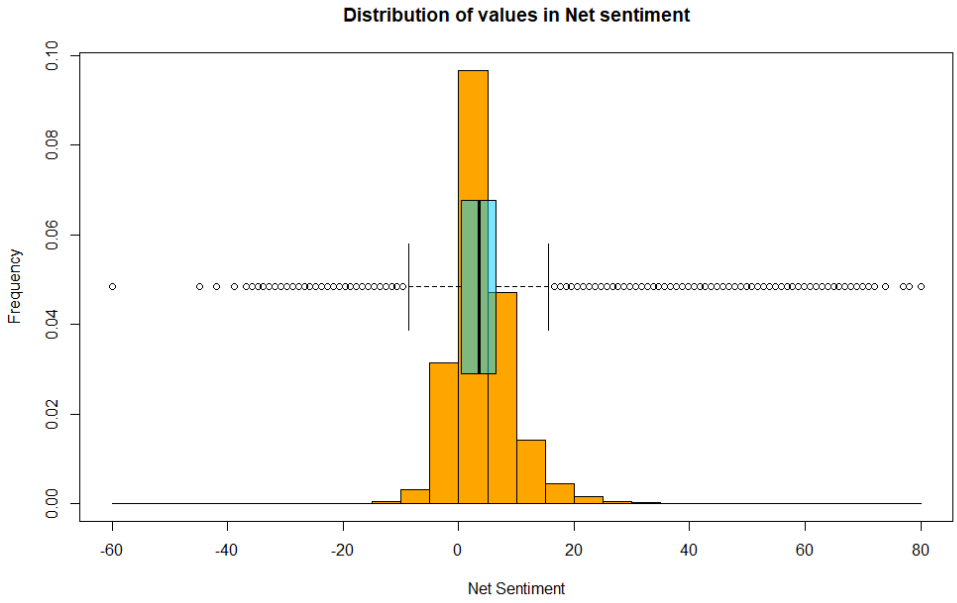
Negative words




```
  neg_words
Min.   : 0.00
1st Qu.: 0.00
Median : 2.00
Mean   : 2.55
3rd Qu.: 4.00
Max.   :65.00
```

Number of negative words used in reviews ranges from 0 to 65. However the average number of negative words is just 2 and the 3rd quantile is at 4 meaning only a quarter of all reviews contained more than 4 negative words. Infact, around 25% of all reviews didn't contain even a single negative word in them. Overall, the graph is higly rightly skewed with majority of all values accumulated towards lower end of spectrum with some outliers present.

Net Sentiment

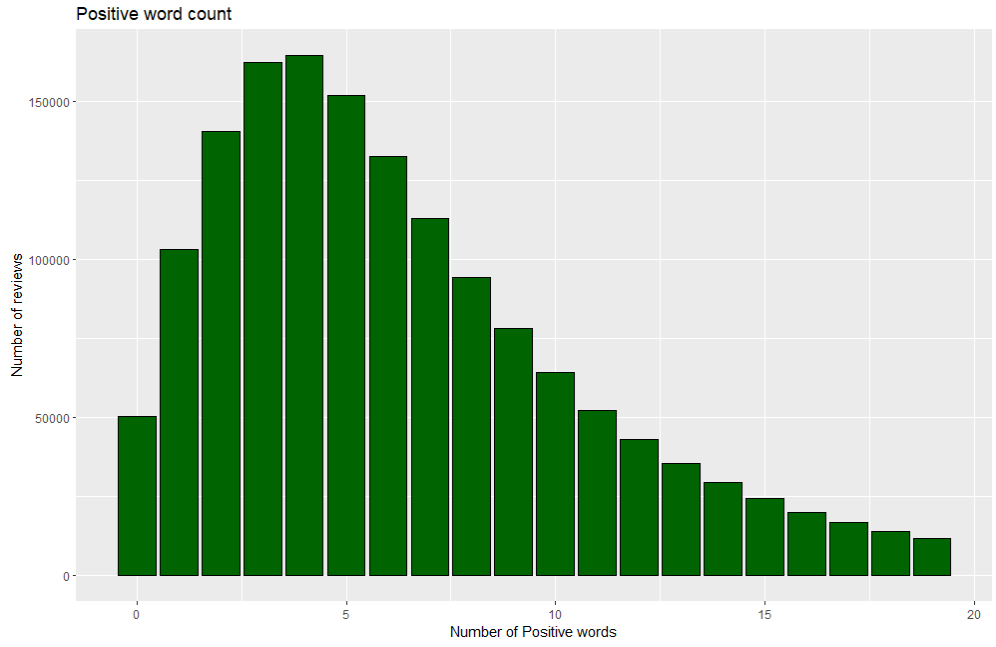



```
net_sentiment
Min.   :-59.000
1st Qu.:  1.000
Median :  4.000
Mean   :  4.522
3rd Qu.:  7.000
Max.   : 80.000
```

Net sentiment shows the overall mood of the user posting review. It is derived by subtracting number of negative words from positive words in a review. It ranges from -89 to 80 with average at 4.5 meaning at an average the sentiment of reviews are positive. The 1st quantile is at 1 meaning only a quarter of all reviews were of negative sentiment. The boxplot shows that eventhough mainly all values lies between -10 to 20, there are a lot of outliers present.

Count of Positive and Negative words

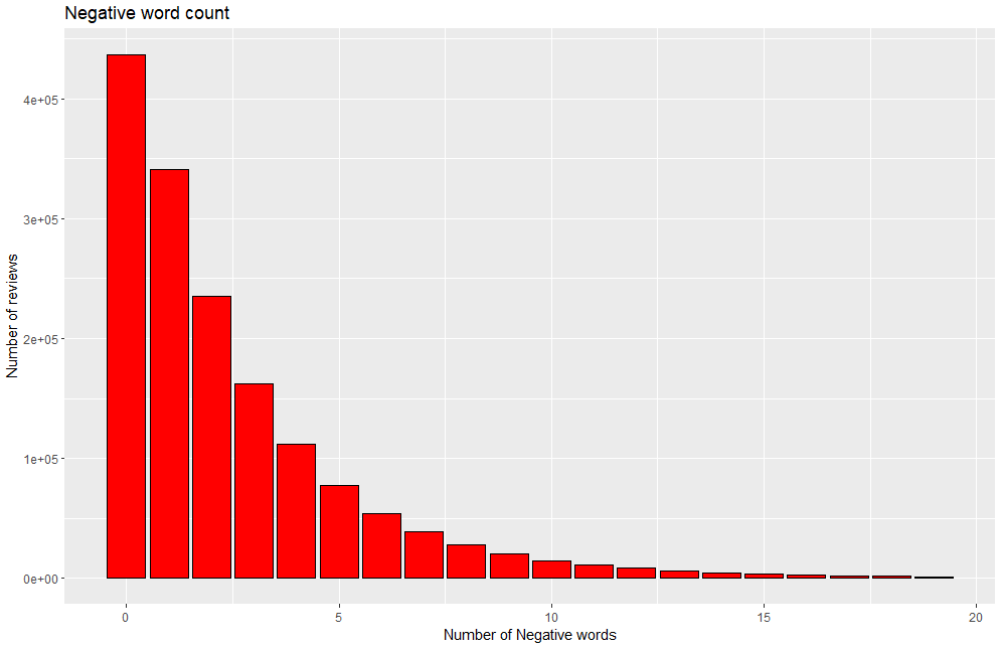
Positive words




```
     Count
Min.   :     1
1st Qu.:    24
Median :   362
Mean   : 18462
3rd Qu.:  8164
Max.   :164596
```

The figure above shows number of reviews for number of positive words in the reviews. There are around 50000 reviews with no positive words in them, the number of reviews then increses with increse in number of positive words. Reviews containing 4 positive words are highest in number and after that the trend reverses and the number of reviews decreses with increase in positive words. The graph indicates that most reviews contained 3 to 7 positive words in them which is inline with the previous finding in statistical summary.

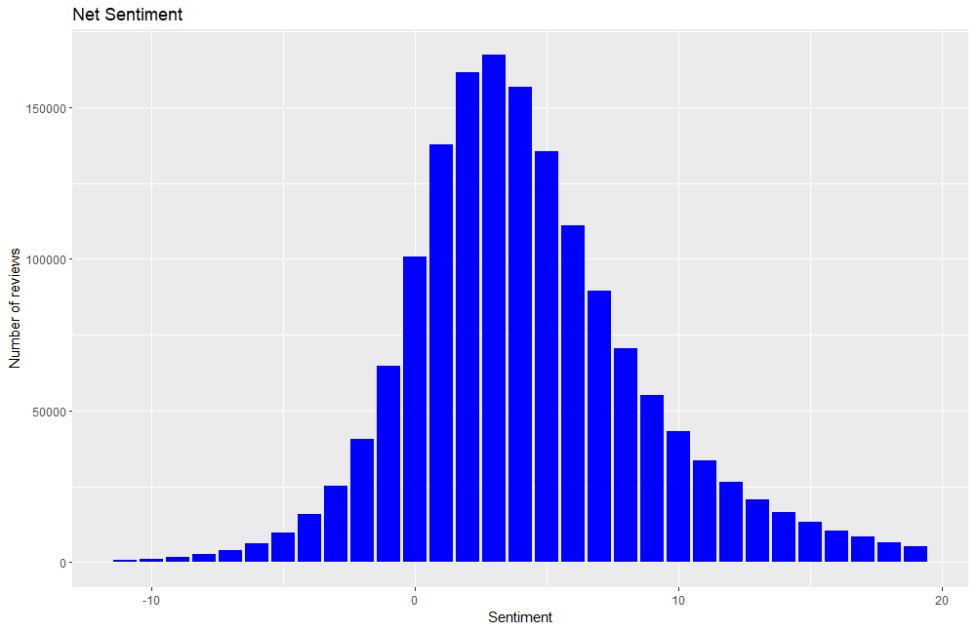
Negative word



	Count
Min.	1
1st Qu.	15
Median	274
Mean	28532
3rd Qu.	5545
Max.	436419

The number of reviews by negative word content is shown above. Most number of reviews had no negative words and as number of negative words increases, the number of reviews decreases exponentially. Surprisingly, nearly half of all reviews had no negative word in them.

Net Sentiment



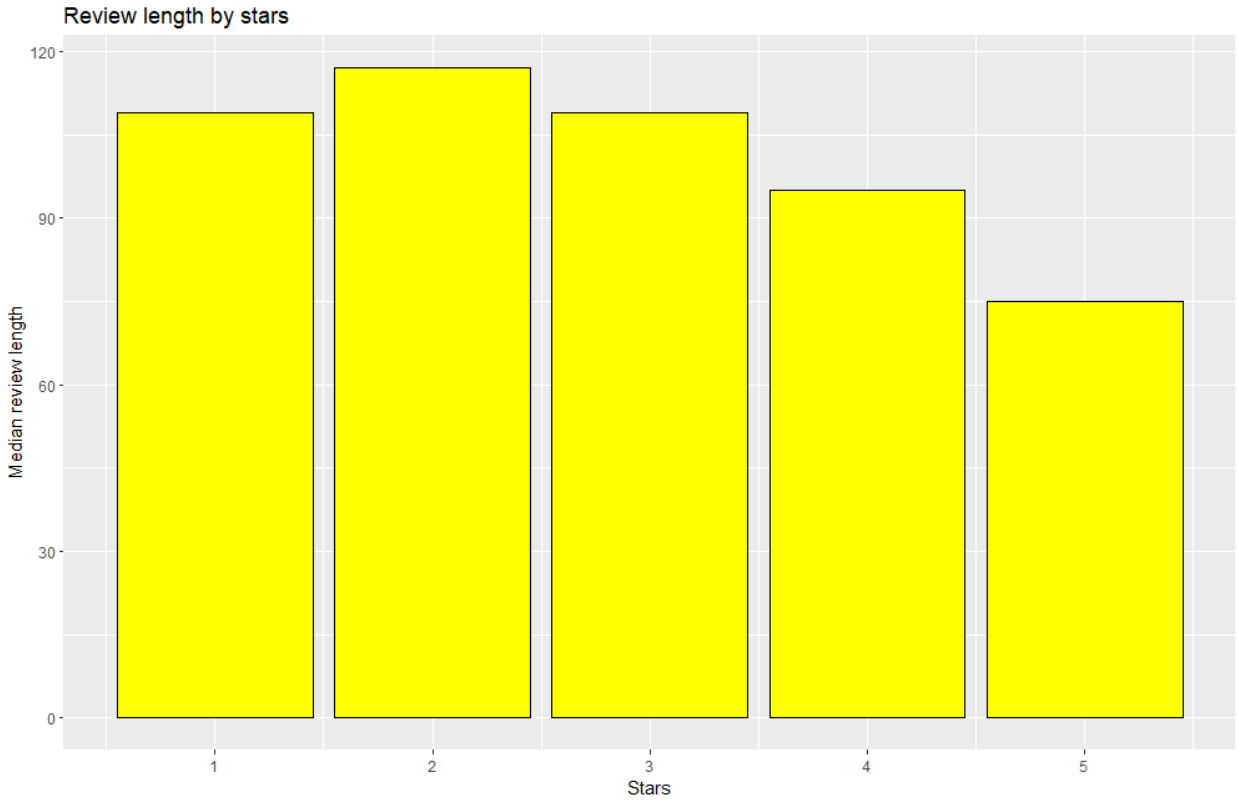
	Count
Min.	1
1st Qu.	8
Median	97
Mean	13413
3rd Qu.	2810
Max.	167212

The overall sentiment of reviews lies somewhat on the positive side with very few reviews having overall negative sentiment. The net sentiment graph peaks at around 4 showing mainly all values had somewhat positive sentiment.

Average Review Length

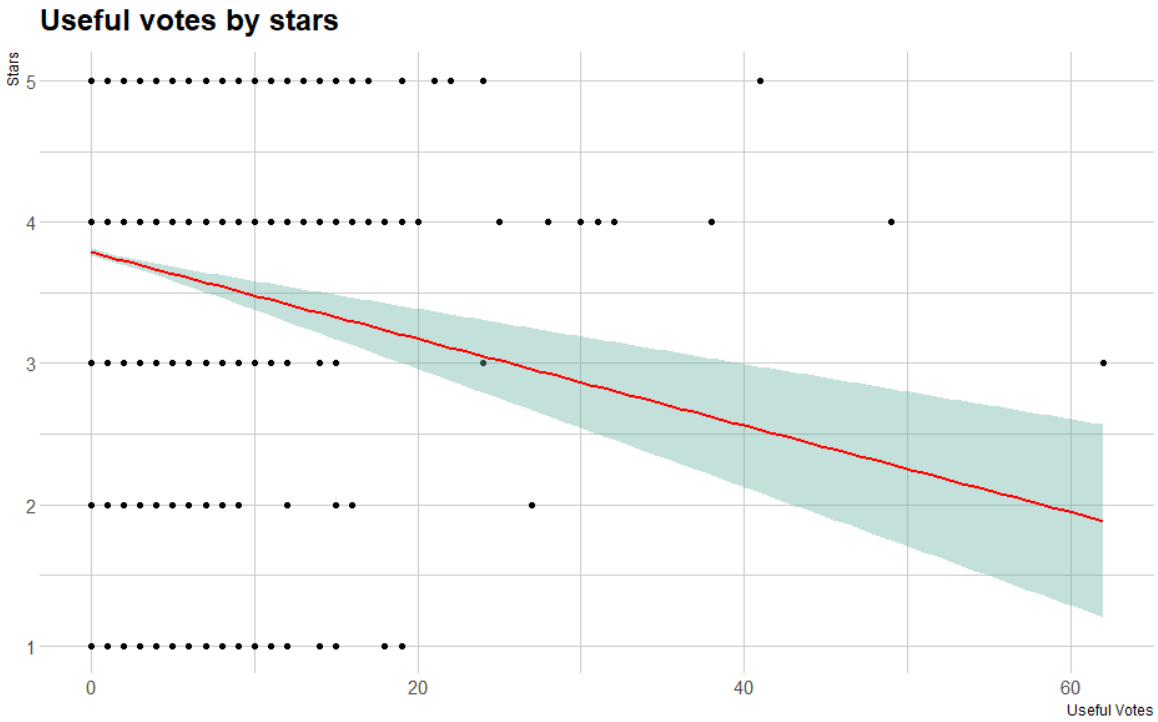
	stars	mean_length	median_length
	<int>	<dbl>	<dbl>
1	1	153.	109
2	2	154.	117
3	3	141.	109
4	4	125.	95
5	5	106.	75

The mean and median length of reviews in different star categories is shown above. Here, it can be clearly observed that the mean values are for greater than the median values. This is due to mean values being heavily influenced by the outliers present in the distributions. So for better representation of the review length we shall take median as the true average here.



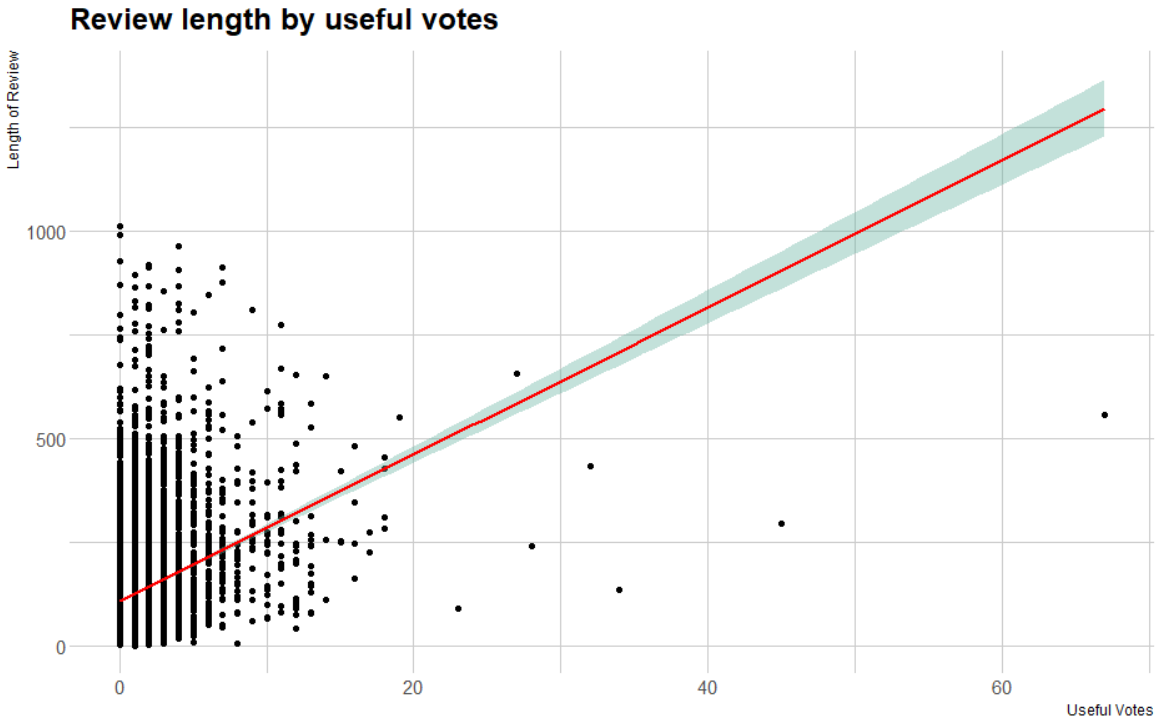
It can be observed from the above graph that average length of review is longest in 2 star category. The length of review decreases as the number of stars increases. This behaviour can be attributed to users giving lower star rating justifying their ratings and users giving higher stars having no need to justify themselves.

Relationship between Useful votes and star or review length.



"The correlation coefficient between Votes useful and Stars is: -0.049"

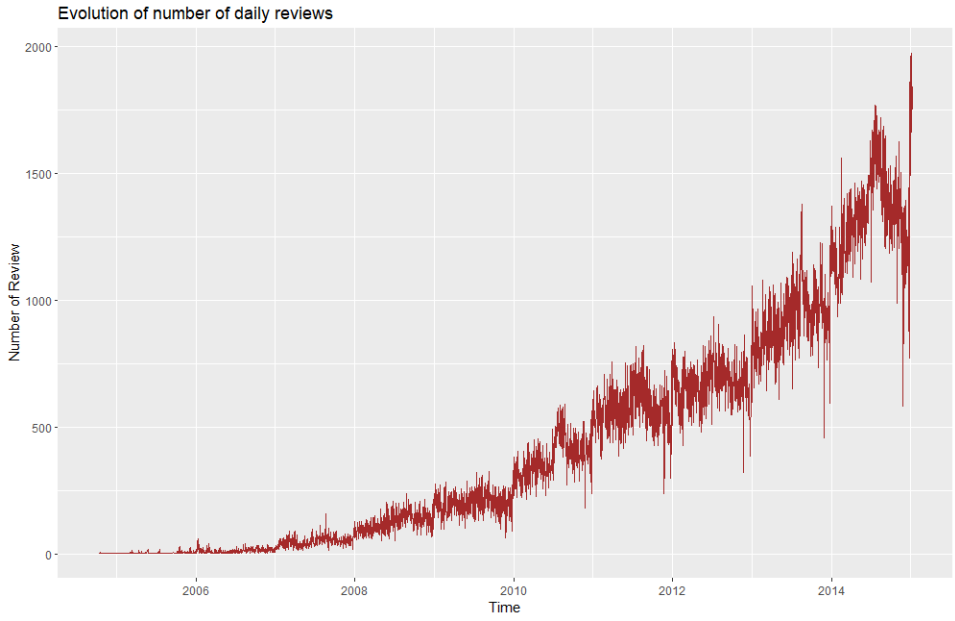
As it can be seen above that as number of stars incrases the number of useful votes decreases. At an average, people do not find reviews with higher star category as useful. This is further confirmed by the pearson correlation coefficient of -0.049 showing a weak negative linear relationship.



"The correlation coefficient between Votes useful and Review length is: 0.326"

In the above given plot it can be seen that as number of useful votes increases the length of the reviews also increases. This can be due to people being able to find their relevant information in lengthy reviews. This is further confirmed by the pearson correlation coefficient of 0.326 which shows a positive linear relationship.

Evolution of daily reviews



	total
Min.	1.0
1st Qu.	71.0
Median	331.0
Mean	455.5
3rd Qu.	713.0
Max.	1972.0

In the figure above it can be seen that number of daily reviews starts from near zero in 2005 and goes to around 2000 reviews per day 2015. The daily reviews has increased gradually over the decade but it has experienced a lot of small spikes and dips along the way. The average number of reviews on yelp was around 455/day during the given decade.

Best Business/User

Any business is a good business if it is able to satisfy it's customers needs and provide good service along the way. A satisfied customer will always write positive things about the business. Net sentiment which shows overall sentiment of the user writing the review can be a good indicator of which business is the best. Here we will find the business with most average net sentiment value and to eradicate businesses which have just one or two highly positive reviews we filter businesses by the average number of reviews.

By the above criteria the best business is "egaieBcnSZLYGl1N-_CtvQ" with star rating of 4.52, average net sentiment of 17.8 and a total of 48 reviews.

A users job is to provide review that can help other people to make decisions. If any review is helping people than they will vote it as useful or funny or cool so we can take total votes as the criteria for finding best user. However to avoid oneoff case where a user has posted only one review and gets 100s of votes we will filter out the users with less than average number of reviews.

By the above criteria the best user is "WJSNywtir04BgDDpZVZMpg" with average of 138 votes per review and total of 15 reviews.

Conclusion

To conclude, reviews depend heavily on the businesses but overall, more than half (around 75%) of the reviews had 3 or higher star rating, was of length less than 165 words, had less than 7 positive and 4 negative words in them. Mainly all reviews are of positive sentiments. However, exceptionally large or small values, also known as outliers are present in every variable. Number of reviews decreases as the number of positive or negative words increases showing that majority of reviews are of balanced tone. Average length of reviews also decreases with increase in star rating. There is some positive rlation between votes useful and review length indicating that people find lengthy reviews more useful. The relation between star rating and votes useful is a negative one but not strong enough to conclusively say anything.

As the popularity of the website has increased, so has the number of reviews being posted on the website on a daily basis.