COMP5070_SP2_2023
Statistical Programming for Data Science
Assignment 2

Enna H

Introduction

As we immerse ourselves in the digital age, online reviews are playing an increasingly pivotal role in shaping the trajectory of businesses. Consumer feedback, more than ever, is at the forefront of shaping the business landscape, and platforms like Yelp are instrumental in this process. It's against this backdrop that we delve into an analysis of the online sentiment encapsulated in Yelp's user reviews. The research will provide insights into user behaviours, revealing patterns and trends that are crucial for businesses and users alike to understand, manage and navigate this influential online space effectively. Through our data-driven approach, we aspire to add a quantitative lens to the qualitative nature of user sentiment.

The report will be structured in several key sections, each offering valuable insights into different facets of the dataset. Starting with a preliminary analysis, we will examine the overall sentiment of reviews, analysing the language and star ratings used, and their relation to the review length. Following this, we will dissect the sentiments expressed in the reviews, scrutinising the distribution and variation of sentiments, with a particular focus on outliers. Moving on, we will explore the relationship between star ratings and review length, as well as their influence on the perceived usefulness of a review. This is followed by a temporal analysis to track changes in review volume over time. Lastly, the report identifies the most impactful users and businesses, considering various criteria such as the number and type of votes received, and improvements in star ratings. The objective is to establish a robust analytical framework that businesses can use to maximise their online presence and understand user behaviours in a more meaningful way.

In-Depth Exploration of Selected Variables in Yelp Reviews Dataset

Having understood the broad overview of the Yelp reviews dataset, we now turn our attention to a more detailed examination. We will delve into the nuances of selected variables, including “Rating”, “Review_Length”, “Positive_Words”, “Negative_Words”, and “Net_Sentiment”. Our goal is to gain a deeper comprehension of the underlying trends and patterns shaping the dataset and, by extension, users’ behaviors and sentiments.

Firstly, we will conduct a summary statistic of the chosen variables and subsequently move onto the visualization of our findings. Let’s dive in.

Table 1: Summary Statistics for Selected Variables in Yelp Reviews Dataset

skim_type	skim_variable	n_missing	numeric.mean	numeric.sd	numeric.p0	numeric.p50	numeric.p100
numeric	Rating	0	3.74	1.31	1.00	4.00	5.00
numeric	Review_Length	0	125.60	115.50	0.00	92.00	1047.00
numeric	Positive_Words	0	7.07	5.93	0.00	6.00	94.00
numeric	Negative_Words	0	2.55	3.25	0.00	2.00	65.00
numeric	Net_Sentiment	0	4.52	5.24	-59.00	4.00	80.00

Table 1 provides a statistical summary of our chosen variables: “Rating”, “Review_Length”, “Positive_Words”, “Negative_Words”, and “Net_Sentiment”. The data shows an average rating of around 3.74 stars, suggesting overall favorable reviews by users. Interestingly, the average review length is about 125 characters, indicating a preference for succinct reviews.

Intriguingly, reviews typically contain roughly seven positive words compared to just two negative words on average, reaffirming the impression of generally positive reviews. The average net sentiment, determined as the difference between the count of positive and negative words, is approximately 4.52, denoting a mostly positive sentiment in the reviews. However, the presence of potential outliers in the data calls for a more thorough analysis.

Figure 1 illuminates the distribution of ratings and their correlation with review lengths. As observed, a significant number of reviews are in the 3 to 5-star range, with 5-star ratings being the most frequent. This further reemphasizes the overall positive sentiments expressed in the reviews.

The correlation between the ratings and review lengths is depicted through a violin plot. It showcases an intriguing pattern: the review length tends to decrease as the rating improves, indicating that users giving lower ratings often write more extensively to detail their experiences.

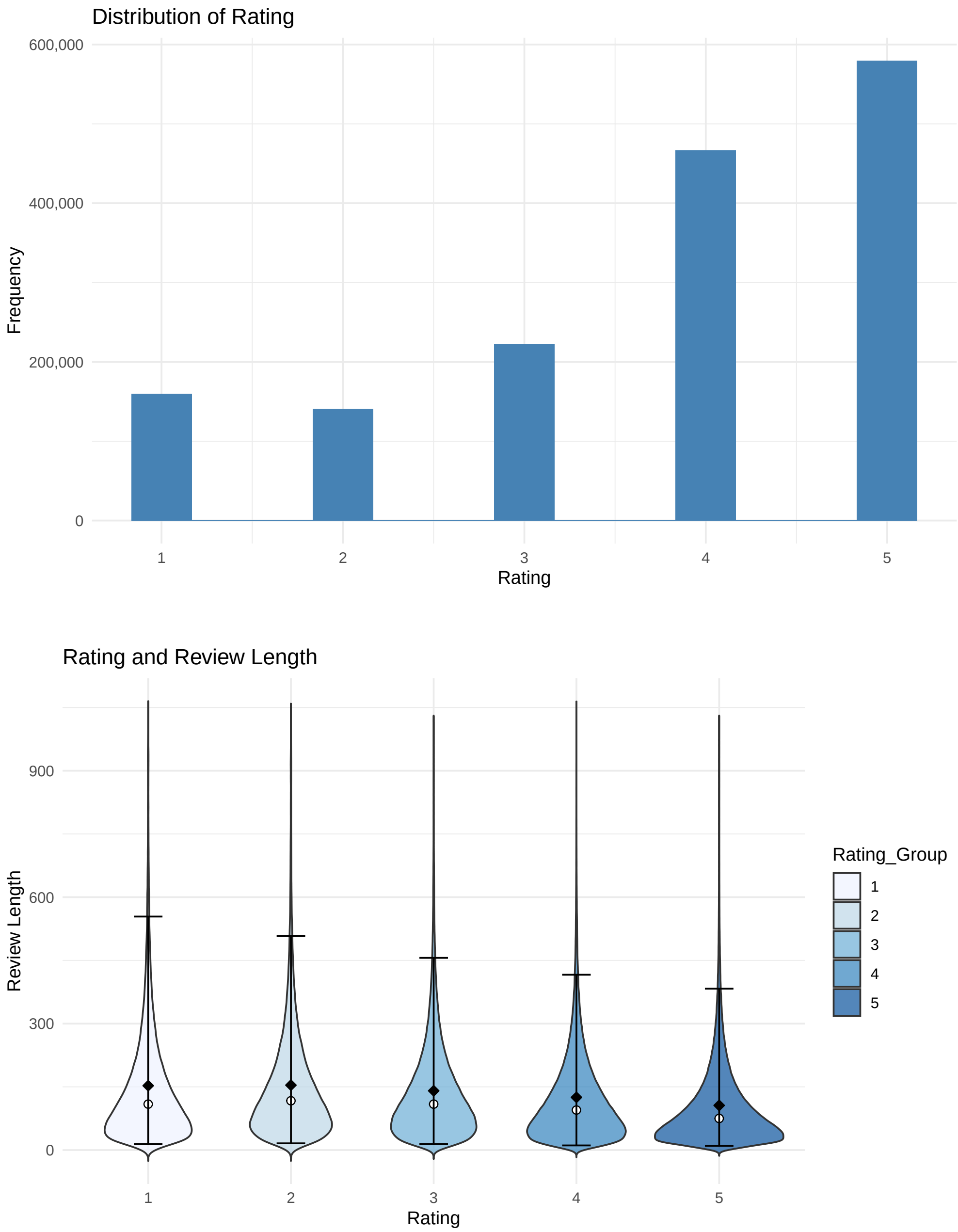


Figure 1. Visualizing the Distribution of Ratings and Correlation with Review Length.

Preliminary Impressions from Detailed Analysis

This in-depth exploration of our Yelp reviews dataset offers a comprehensive understanding of user sentiment. The predominance of positive reviews and greater use of positive words reaffirm the overall upbeat trend. While the dataset has some outliers, they do not significantly impact the largely positive sentiment. Interestingly, the inverse correlation between review length and rating provides a key insight into user behavior, suggesting a tendency among users to elaborate more on negative experiences. We anticipate that a more granular analysis of this dataset will reveal additional interesting trends and patterns, offering valuable insights for businesses on Yelp.

Understanding Review Sentiments Through Word Counts

To better grasp the overall sentiment of Yelp reviews, we're going to examine the number of positive and negative words used in each review. By doing so, we can understand how users generally feel about the businesses they review on Yelp.

Let's start by creating tables that count both positive and negative words in each review and visualize the first 20 entries to get a general idea of sentiment trends.

Table 2: Summary Statistics for Positive Words in the Yelp Reviews Dataset

column	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Positive_Words	85	42.19	25.02	42	42.00	21	0	94	94	0.05	1.88	2.71
n	85	18461.93	40673.70	362	7330.26	360	1	164596	164595	2.45	7.88	4411.68

Table 3: Summary Statistics for Negative Words in the Yelp Reviews Dataset

column	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Negative_Words	55	27.62	17.07	27	27.13	14	0	65	65	0.20	2.07	2.30
n	55	28532.07	82595.78	274	6274.04	272	1	436419	436418	3.63	15.98	11137.21

From these tables, we can immediately observe that reviewers tend to use more positive words compared to negative ones. This could suggest that businesses on Yelp are generally seen favorably by their reviewers.

Let's break this down further:

Positive Words: On average, a review contains around 42 positive words. Although there is some fluctuation (as seen from the standard deviation of 25), most reviews hover around this number. This suggests that users typically use a moderate number of positive words in their reviews.

Negative Words: Similarly, the average review has around 28 negative words. While there is also some variation here (with a standard deviation of 17), it's clear that reviews contain fewer negative words compared to positive ones.

Table 4. The counts of numbers of positive words per review and the counts of numbers of negative words per review.

Positive_Words	n	Negative_Words	n
0	50339	0	436419
1	103321	1	340959
2	140506	2	235540
3	162600	3	161944
4	164596	4	112062
5	151969	5	77610
6	132809	6	54395
7	113252	7	38704
8	94409	8	28010
9	78414	9	20340
10	64375	10	14880
11	52484	11	10980
12	43092	12	8302

Positive_Words	n	Negative_Words	n
13	35415	13	6341
14	29489	14	4749
15	24467	15	3744
16	20168	16	2938
17	16812	17	2164
18	13975	18	1810
19	11732	19	1459

One interesting thing to note is that both positive and negative words tend to follow a similar distribution. This suggests that while the total number of positive and negative words may differ, the way they are spread across reviews is quite similar.

Now let's visualize these word counts to better understand these trends:

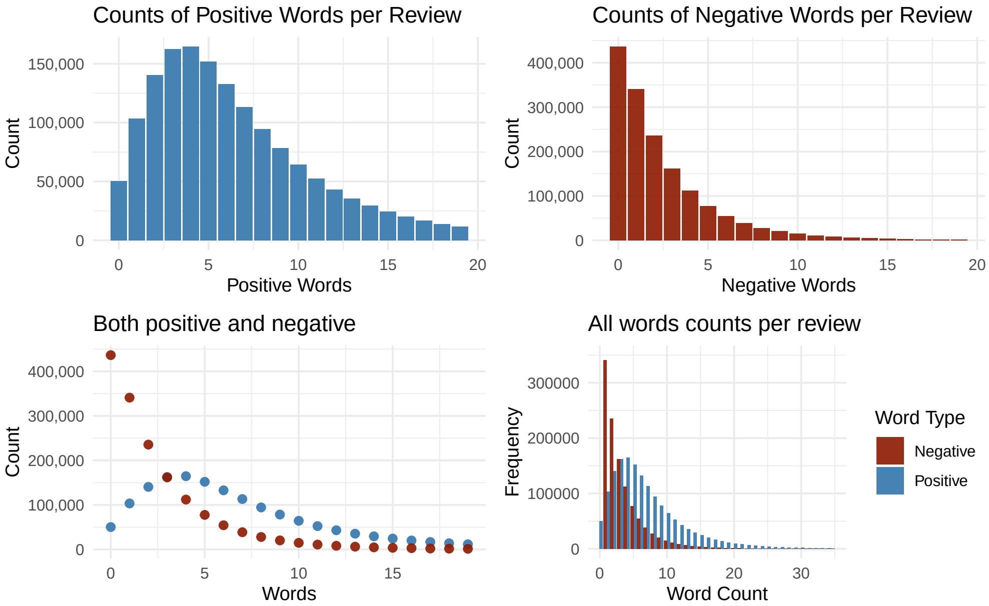


Figure 2. The counts of positive words per review and the counts of negative words per review.

From these visualizations, we can draw a few conclusions:

1. Reviews typically contain a small to moderate number of positive words. The number of reviews tends to decrease as the number of positive words increases.
2. A similar trend can be seen with negative words, with most reviews containing a small number of such words.
3. Comparing positive and negative words directly, we can see that reviews generally contain more positive words.
4. When looking at the total word count across all reviews, it's clear that positive words are used more frequently than negative ones.

In conclusion, both the data and visualizations suggest that reviewers on Yelp tend to express more positive sentiments than negative ones. This insight can be incredibly valuable for businesses looking to understand their customer feedback on Yelp better. It's also a useful starting point for more in-depth analyses, like examining the specific positive and negative words that are used most often.

An In-Depth Look at Reviews Sentiment

To understand how customers feel about businesses on Yelp, we analyzed the sentiment expressed in the reviews. For context, we calculated a “net sentiment score” for each review, which reflects the balance between positive and negative words used.

Table 5: Summary Statistics for Net Sentiment in the the Yelp reviews dataset

column	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Net_Sentiment	117	17.85	34.54	18	18.00	29	-59	80	139	-0.04	1.92	3.19
n	117	13412.51	35380.86	97	3354.86	96	1	167212	167211	3.10	11.83	3270.96

Let’s take a look at the big picture first. On average, reviews on Yelp are mostly positive, with an average net sentiment score of about 18 (on a scale that goes from -59, extremely negative, to 80, extremely positive). It’s important to note, however, that the sentiment scores vary quite a bit from review to review.

To give an idea of this variation, imagine a group of people asked to guess the number of candies in a jar. If everyone’s guesses were close to each other, we would say there’s low variability. But if some guessed very low and others very high, that’s high variability. In the case of our Yelp reviews, there’s a moderate level of variability. This means that while the average sentiment leans positive, there are still plenty of reviews that are neutral or negative.

To better understand this, we grouped the reviews into categories based on their sentiment scores, similar to putting the candy jar guesses into buckets. Most reviews had a relatively neutral sentiment, falling within the “-10 to 10” category, meaning they had a similar number of positive and negative words. But a significant portion of reviews fell in the “10 to 20” and “20 to 40” categories, indicating a generally positive sentiment.

Table 6. The counts of net sentiment by bin in the the Yelp reviews dataset.

Sentiment_Bin	n	Sentiment_Bin	n
< -20	206	< -40	3
-20 to -10	3895	-40 to -30	16
-10 to 10	1399464	-30 to -20	187
10 to 20	145722	-20 to -10	3895
20 to 40	19302	-10 to 0	271006
> 40	675	0 to 10	1128458
		10 to 20	145722
		20 to 30	16562
		30 to 40	2740

However, it’s worth noting that there are some outliers - a few reviews that are either extremely positive or negative. These are like those far-off guesses for the candy jar, few but significant. Even though they’re rare, these reviews can strongly affect a business’s overall sentiment perception and should be taken into account.

In the chart below, you can see the distribution of these sentiment scores, which illustrates this mix of neutral, positive, and a few highly negative or positive reviews.

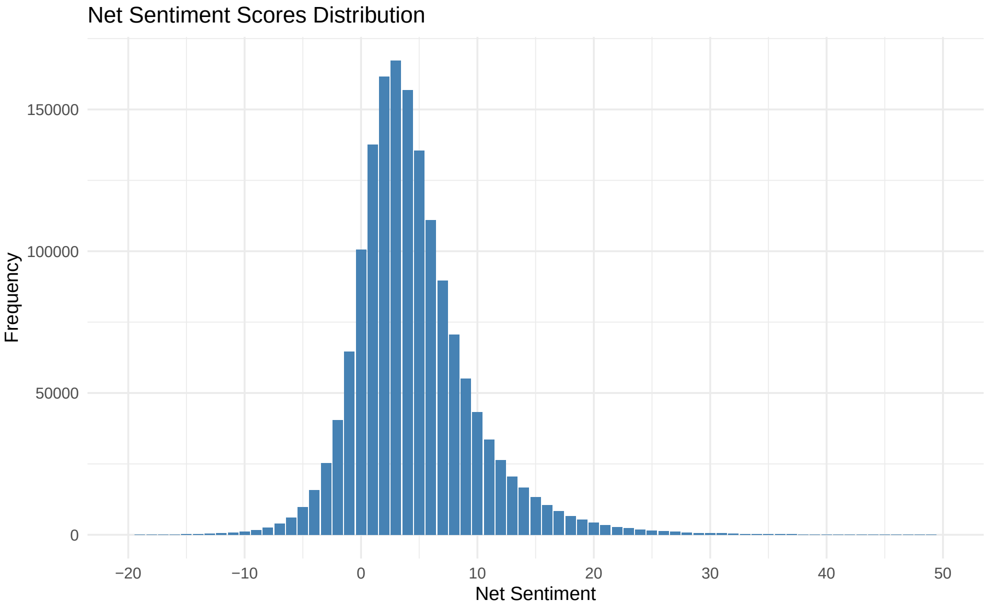


Figure 3. The distribution of net sentiment in the Yelp reviews dataset.

This understanding of review sentiment can be valuable for businesses. It can help them gauge their overall performance, identify any potential issues, and strategize on how to better satisfy their customers.

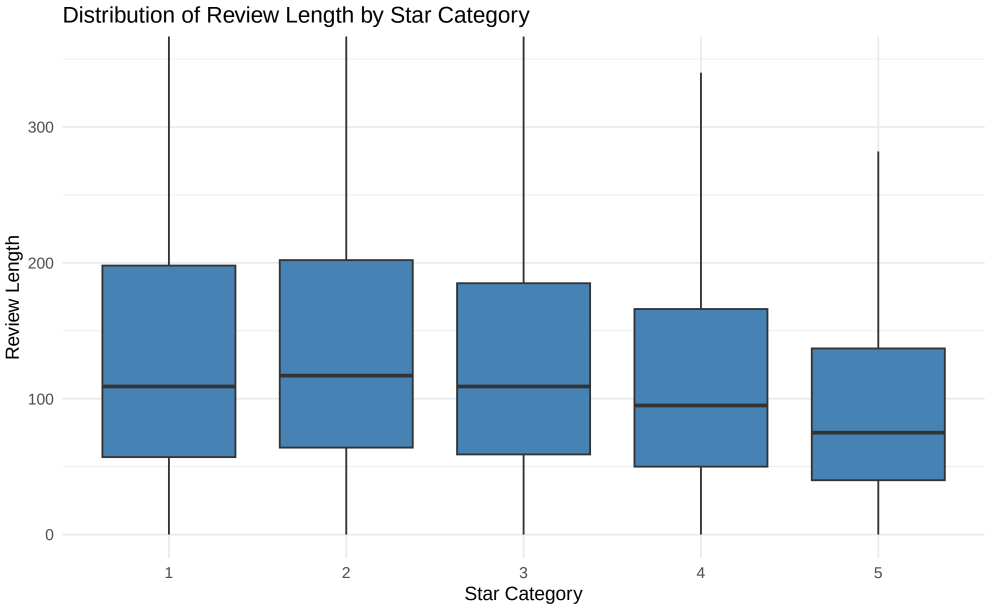
Insights on Review Length and Star Rating

We wanted to see if there was any correlation between the length of a review and the star rating given by a customer. We broke down the reviews into their star rating categories and looked at the average length of a review in each category.

Data Representation

To represent this data, two types of visualisation techniques were employed - a box plot and a bar plot. The box plot provides a detailed view of the distribution of review lengths per star category, showcasing the median, quartiles, and potential outliers in the data. The bar plot, on the other hand, provides a more concise view of the average review length per star category, highlighting the general trend in the data.

We displayed our findings in two ways: a box plot and a bar plot. The box plot gives us an in-depth view into the distribution of review lengths for each star category. It shows us the most common length of reviews (median), how spread out the lengths are (quartiles), and any outliers. The bar plot is a simple representation of the average review length in each star category, which helps us understand the general pattern in the data.



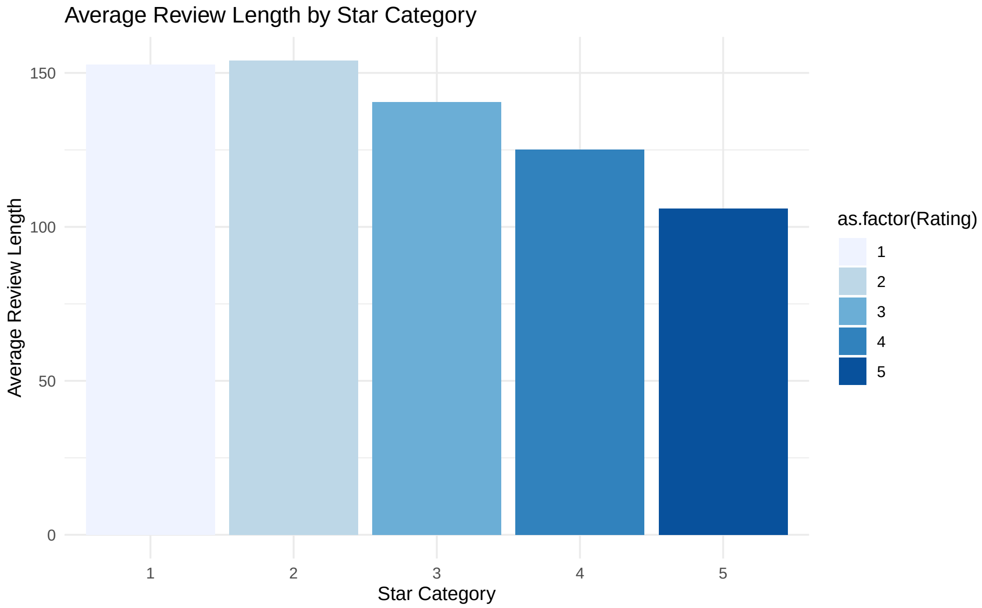


Figure 4. The distribution of review length by star category.

The type of average used in this analysis is the arithmetic mean, a measure of central tendency that calculates the sum of all review lengths and divides it by the total number of reviews in each star category. This form of average was chosen due to its ability to provide a fair representation of the typical review length within each category.

Table 7. Average Review Length by Star Category.

Star Category	Average Review Length
1	153
2	154
3	141
4	125
5	105

We found an interesting trend: as the star rating goes up, the length of the reviews goes down. This indicates that customers tend to leave shorter reviews when they had a positive experience and rated a restaurant higher. This trend is especially clear in the 5-star category, which has the shortest average review length.

This finding suggests that highly satisfied customers may not provide as much detailed feedback. Such insight could be valuable to businesses wanting to encourage customers to leave more detailed, positive reviews. It also helps potential customers and the broader community interpret Yelp reviews more effectively.

Insights on Review Usefulness, Star Rating, and Review Length

Next, we wanted to understand whether the perceived usefulness of a review is linked to the star rating and the length of the review.

Star Rating and Review Usefulness

We first looked at whether there's a connection between the star rating of a review and how useful other users found it. Here's what we found:

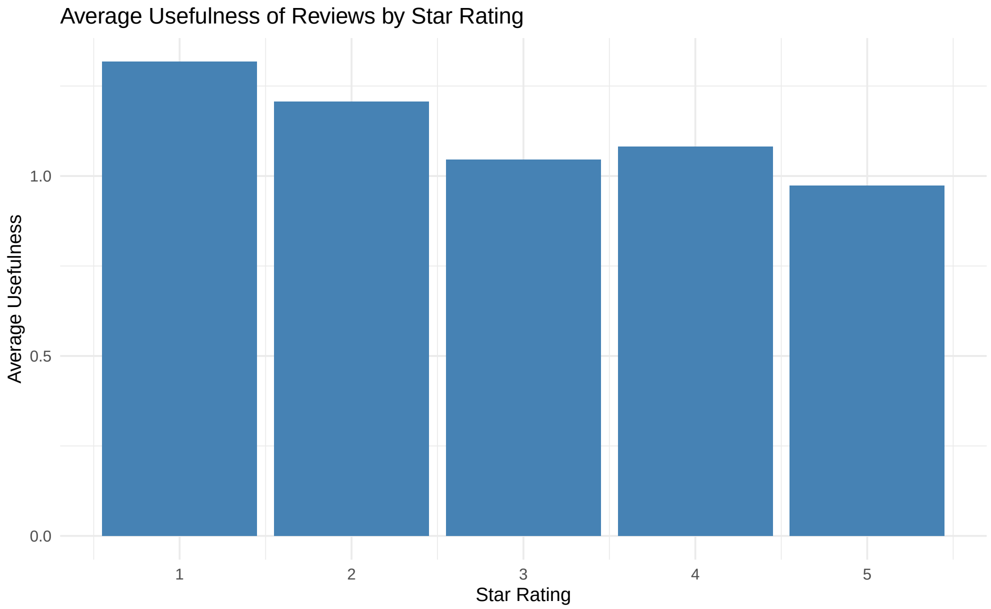


Figure 5: Average Usefulness of Reviews by Star Rating.

It seems that higher-rated reviews are slightly less useful to others, although this trend isn't very strong. We did, however, find that the star rating does influence how useful a review is to others, albeit not by a lot.

Review Length and Review Usefulness

Then, we looked at whether the length of a review impacts its usefulness.

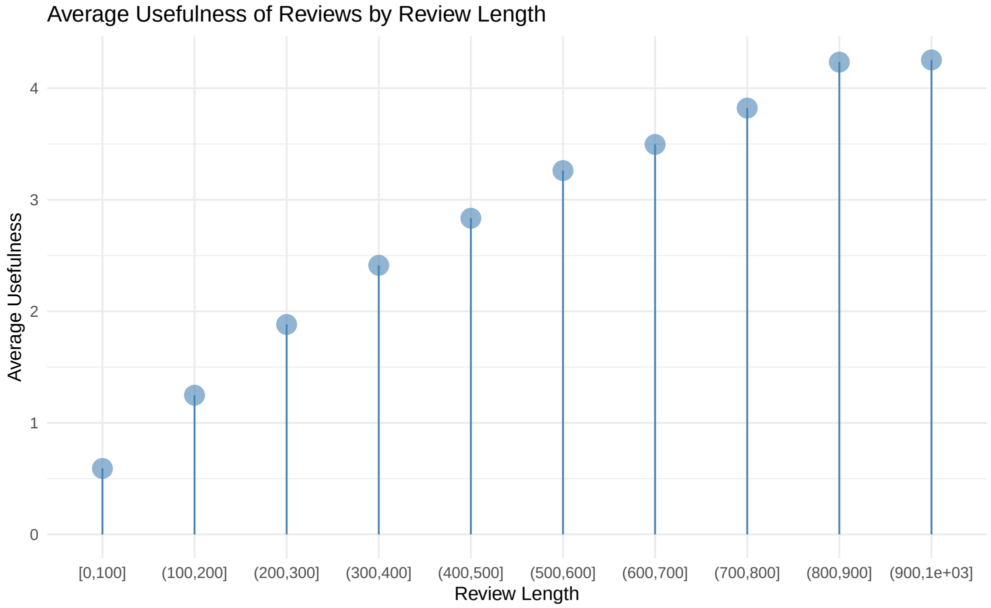


Figure 6: Average Usefulness of Reviews by Review Length.

Longer reviews are generally perceived as more useful by other users. This relationship was statistically significant, meaning the length of a review does impact how useful it is perceived by other Yelp users.

In summary, both the star rating and the length of a review seem to influence how useful other users find a review. However, the relationships are not very strong: higher-rated reviews are only slightly less useful, and longer reviews are only moderately more useful.

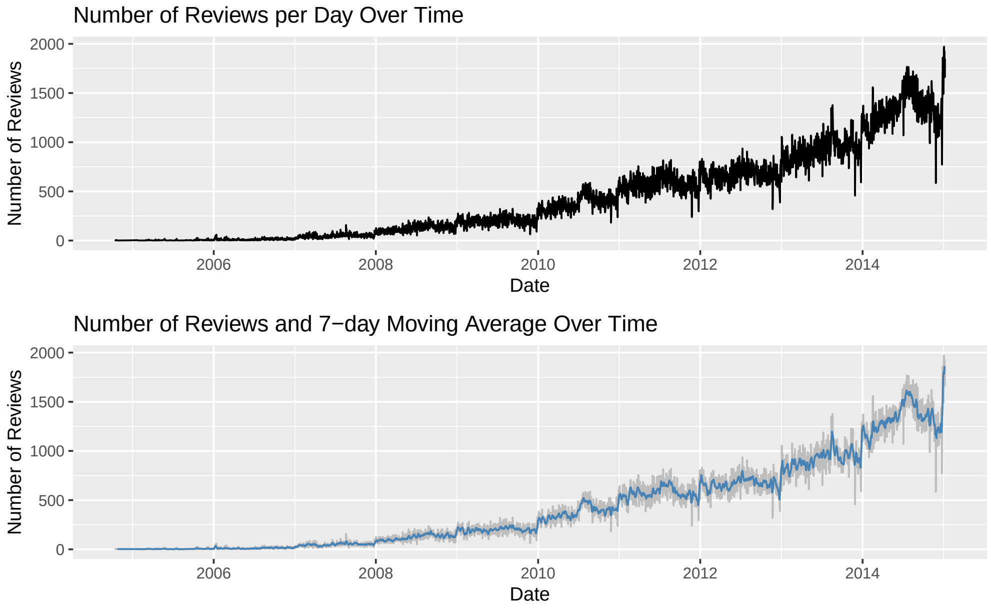
These findings can help guide Yelp users who want their reviews to be useful to others: they might consider writing more detailed reviews. Businesses looking for useful feedback might also encourage their customers to provide more detailed reviews.

Monitoring Daily Review Trends

Now let's examined how the number of daily reviews on Yelp changed over time. We will counted the reviews for each day. This allowed us to track the changing volume of reviews generated on Yelp each day and how this has changed over time.

Our findings were visualized in two ways:

1. A timeseries analysis graph showing the number of reviews submitted each day over the years.
2. A similar timeseries analysis graph, but with a 7-day moving average added in blue, giving a smoother overall view of the trends.



Number of Reviews per Day Over Time

Number of Reviews and 7−day Moving Average Over Time

Figure 7: Number of Reviews per Day Over Time. The grey line in both plots represents the daily count of reviews, whereas the blue line in the second plot depicts the 7-day moving average of reviews.

Insights from the timeseries analysis reveal a general increase in the number of reviews per day over the years. The trend does indicate a few periods of decline, contributing to some fluctuation in review volume. It is noteworthy to mention that by the end of 2015, the platform was seeing close to 2000 reviews a day. The second graph, with the 7-day moving average, shows that despite some short-term fluctuations, the general trend has been a steady growth in daily reviews.

These trends suggest increased user engagement and trust in Yelp as a reliable source for business reviews. For businesses, it's a signal of the growing importance of maintaining a positive online presence, as customers increasingly turn to platforms like Yelp to inform their choices.

Identifying the Best Users and Businesses on Yelp

Our next step was to identify the best users and businesses on Yelp. For users, looked at the number of ‘useful’, ‘funny’, and ‘cool’ votes they received. User ‘kGgAARL2UmvCcTRfiscjug’ stood out, receiving the most ‘useful’ votes. Even when considering a weighted score of all three types of votes, this user remained in the top spot, indicating that their reviews are not only useful, but also entertaining and interesting.

For businesses, focused on average star rating and the total number of reviews. Here, business ’tAdd__IgXQEknDDicEbRgQ’ excelled, achieving an impressive 5-star average rating from a substantial 69 reviews. This suggests consistent high-quality service from this business.

Finally, by examining how businesses had improved over time by looking at changes in their average star rating, business ‘eGevCRobYnA_HSj60s2EWvQ’ stood out, showing a significant increase in its average star rating.

Table 8: Best User by Useful Votes

user_id	total_votes_useful
kGgAARL2UmvCcTRfiscjug	8785
fczQCSmaWF78toLEmb0Zsw	7878
1BW2HC851fJKPfJeQxjkTA	6665
4ozupHULqGyO42s3zNUzOQ	5994
C8ZTiwa7qWoPSMIivTeSfw	5806
0bNXP9quoJEgyVZu9ipGgQ	5741

Table 9: Best User by Weighted Score

user_id	total_votes_useful	total_votes_funny	total_votes_cool	value_score
kGgAARL2UmvCcTRfiscjug	8785	4026	6554	6911.1
fczQCSmaWF78toLEmb0Zsw	7878	3630	6507	6329.4
C8ZTiwa7qWoPSMIivTeSfw	5806	7321	5806	6260.5
1BW2HC851fJKPfJeQxjkTA	6665	4479	5799	5836.0
4ozupHULqGyO42s3zNUzOQ	5994	2692	4582	4721.0
0bNXP9quoJEgyVZu9ipGgQ	5741	2816	4517	4618.7

Table 10: Best Business by Average Star Rating and Total Reviews

business_id	avg_star_rating	total_reviews
tAdd__IgXQEknDDicEbRgQ	5.00	69
nGF9IIg9AMqZNNedyfkHKQ	5.00	52
7EY1sCIfoYvfU6Hs6LvWHw	5.00	51
mKuujaUzhahl3mIZPdt74Q	4.98	62
Z0DD_sieK8ijAETu7OwGXw	4.98	59
-bVdJy8LzTNRYcSxJt2XNw	4.97	96

Table 11: Best Business by Improvement Over Time

business_id	avg_star_rating	total_reviews	total_improvement
eGevCRobYnA_HSj60sEWvQ	4.06	318	4
pCSbslNkq0fL5ZZ3RhUZMw	3.36	320	4
BjWJRBhTC0sev3qEI3m2Hg	3.34	172	4
zgab_7ppM5aqe65dTYXT5g	3.19	72	4
0NlTmTzKYXbE6Hx2iGEb5A	3.16	82	4
1o2Lx_YMC2xpuBG2q00LAQ	2.75	68	4

The 'best_user' table shows the top users sorted by the sum of 'useful' votes they've received. User 'kGgAARL2UmvCcTRfiscjug' has received the most 'useful' votes with a total of 8785.

The 'best_user_advanced' table, however, considers not only 'useful' votes but also 'funny' and 'cool' votes, giving different weights to each type of vote. User 'kGgAARL2UmvCcTRfiscjug' remains the top user when considering these weights, with a weighted 'value score' of 6911.1.

The 'best_business' table lists the top businesses based on their average star rating and the total number of reviews, but only for businesses with more than 50 reviews. Business 'tAdd___IgXQEknDDicEbRgQ' has an average star rating of 5.00, with a total of 69 reviews, making it the top business by this measure.

Finally, the 'best_business_advanced' table shows businesses that have demonstrated improvement over time. Business 'eGevCRobYnA_HSj60s2EWvQ' stands out with a total improvement of 4.00 in its average star rating, even though its overall average star rating is 4.06.

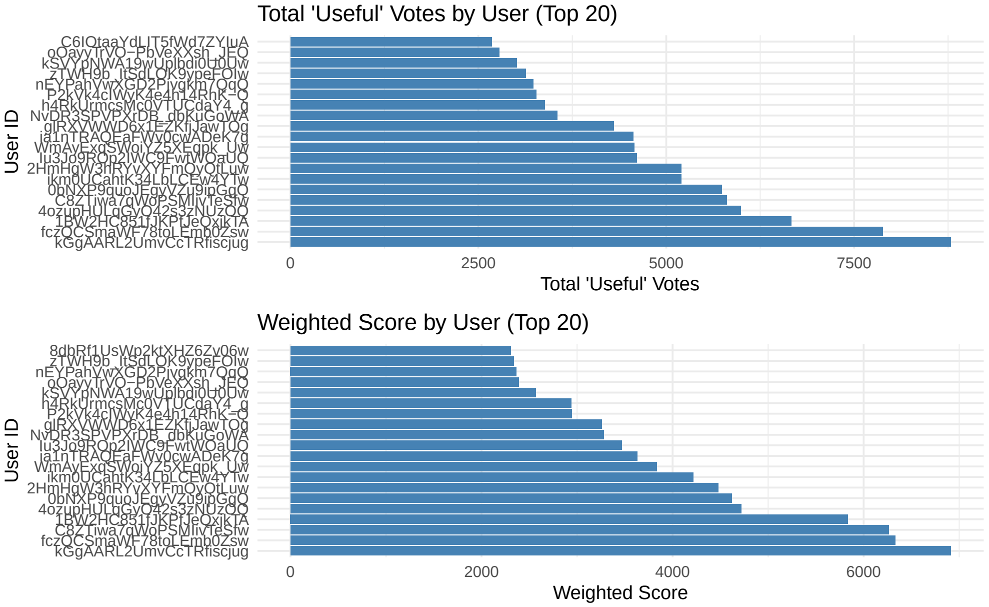


Figure 8. The top 20 users according to the different ranking metrics

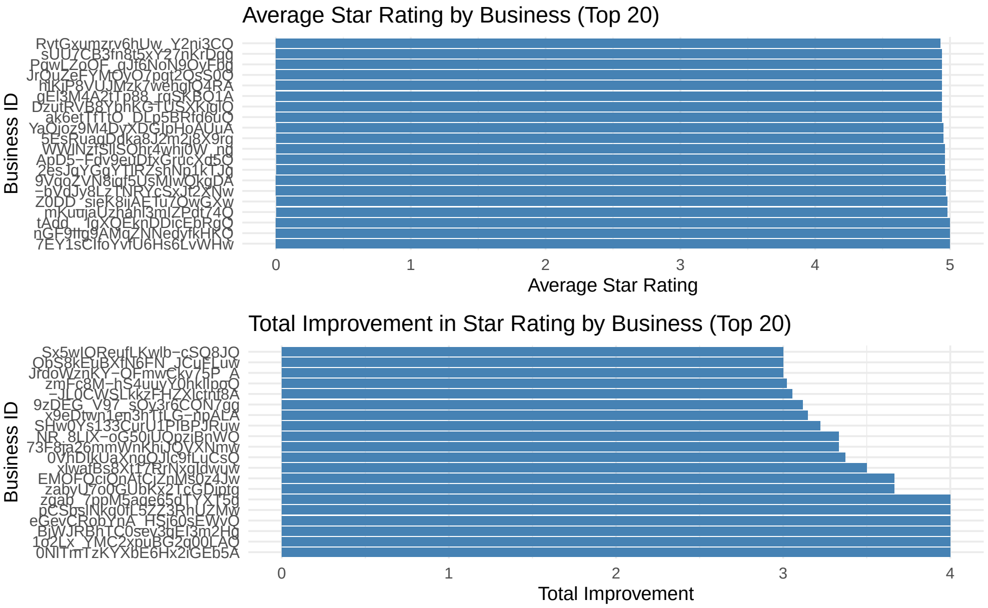


Figure 9. The top 20 businesses according to the different ranking metrics.

Taken together, these analyses highlight the multi-dimensional nature of success on Yelp. High-quality service and engaging, useful reviews are both vital. Moreover, businesses that strive to improve over time are rewarded with better ratings, suggesting that both businesses and users need to be proactive and dedicated to excel on this platform.

Conclusion

Through our comprehensive analysis of Yelp's user reviews, we have unearthed a myriad of insights that shed light on the complex dynamics of online sentiment. It's evident that Yelp reviewers exhibit a tendency to lean towards positivity, with the vast majority of reviews expressing neutral to positive sentiments. The significance of this finding cannot be overstated as it highlights the overarching tone of Yelp's online discourse, providing businesses with an invaluable reference point in understanding their online feedback.

Noteworthy is the inverse relationship between star ratings and review length. Customers seem to be more verbose when their experiences are unsatisfactory. Thus, a longer review could be indicative of a less positive dining experience. Businesses could use this insight to improve their services by giving more attention to the detailed feedback in longer reviews.

We also discovered that review usefulness is positively correlated with review length, but weakly negatively correlated with star ratings. In essence, detailed reviews tend to be deemed more useful by other users. Therefore, users aiming to offer useful feedback and businesses aiming for helpful feedback should focus on the provision and elicitation of detailed reviews.

The analysis of review volume over time unveiled a promising trend for Yelp and its registered businesses. There's a steady increase in user engagement over the years, reinforcing the growing relevance and trust in the platform as a credible source of business reviews. For businesses, this means that managing their online reputation on Yelp is becoming more critical than ever before.

Lastly, our identification of top users and businesses provided insights into how various factors such as the number and type of votes received, and improvements in star ratings contribute to the prominence of a user or business within the Yelp community. This understanding can guide businesses in strategising their interactions with the platform and its users.

In conclusion, this study affirms the crucial role of online reviews in today's digital age. The behavioural trends and patterns we uncovered in Yelp reviews offer a wealth of practical implications for both Yelp users and businesses. The ability to understand, interpret, and respond to these findings could profoundly influence business strategies, customer experiences, and ultimately, the trajectory of businesses in this online era.