**Yelp**

**Introduction**

Yelp is a website that allows users to review local business founded in 2014. Yelp as of 2021 exist in 32 countries and over 219+ cities. As of 31 December 2020, Yelp cumulated of 224 million reviews (YELP, 2020). Business types reviewed includes restaurants, home services, auto services and more. The dataset which we will be exploring was provided by UniSA which contains 1,569,264 rows of reviews. Dataset can also be obtained from yelp website directly <https://data.world/brianray/yelp-reviews>.

The focus on this data exploratory analysis is sentiment analysis. “Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material” (towardsdatascience, 2018). We will be analysing the positive and negatives words in reviews, lengths of reviews and the overall feel of the review based on the net sentiment. Through graphs and tables, we will show the relationship between star ratings and review lengths, and how useful people found the star ratings.

**General Summary**

**Star Ratings**

**A graph of a bar chart

Description automatically generated with medium confidence**

Based of the data, people generally provide positive reviews with majority of reviews between 4 to 5 stars with makes up 67% of all reviews. The least common rating was 2 stars making up 9% of the total reviews. The mean rating was 3.7 stars with the median sitting at 4 stars.

**Review Length**

A graph of a number of points

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Q1 | Median | Mean | Quartile 3 | Standard Deviation |
| 48.0 | 92.0 | 125.6 | 165.0 | 115.5 |

The review lengths histogram shows there is right skewed. With majority of the lengths of between 48 to 165 characters. The most frequent amount of character for a review is 42 characters.

**Positive and negative word summary**

**A graph with a line

Description automatically generatedA graph with a line

Description automatically generated**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | | Median | Mean | 3rd Quartile | Max |  | Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| 0.0 | 3.0 | 6.0 | | 7.1 | 9 | 94 |  | 0.0 | 3.0 | 5.0 | 6.0 | 8.0 | 17.0 |

**A graph with a line

Description automatically generated A graph with a line and a rectangle

Description automatically generated with medium confidence**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | | Median | Mean | 3rd Quartile | Max |  | Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| 0.0 | 0.0 | 2.0 | | 2.6 | 4.0 | 65 |  | 0.0 | 0.00 | 1.0 | 2.4 | 3.0 | 14.0 |

To visualise the review length two sets of boxplots was used was generated to show the statistical summary. With and without outliers, due to the large variances from the Median value and the Max value outliers was removed. The IQR 1.5 rule was used in the above right graphs to reflect the numbers without outliers. There are most positive words in reviews then there are in negatives words. The average review has 6 positives words compare 2 words.

A graph with lines and numbers

Description automatically generatedA grid with a rectangle

Description automatically generated

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Min | 1st Quartile | | Median | Mean | 3rd Quartile | Max |  | Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
| -59.0 | 1.0 | 4.0 | | 4.5 | 7.0 | 80 |  | 2.0 | 3.0 | 4.0 | 3.8 | 5.0 | 6.0 |

Net sentiment is calculated by subtracting Negative words from Positive words. As per the above graph this shows that reviews tend to be favourable to positive words. With the mean sitting at 3.8 net sentiment. Further to that, when comparing total sum of words, there was 2.75x more positive words to Negative word as per below graph. However, this reflects the problem with Net sentiment, as this only captures the total positive and negative words and not the outcome of the sentiment of each unique review.

A graph with a red and blue rectangle

Description automatically generated

**Positive vs Negative words**

A graph of a number of positive words

Description automatically generatedA graph of a number of negative words

Description automatically generated

In continuing from the above analysis, the two histograms used to represent the frequency of words. Positive words are had the highest concentration at 164,596 at 4 positive words compare to 340,959 reviews with 1 negative words, which implies more positive words are found in reviews than Negative. Positive words tendency be more normal distributed whereas Negative words are right skewed.

**Net Sentiment**

To visualise the net sentiments, a frequency distribution table was generated to show the number of reviews per net sentiment. The frequency table was then visualised using a bar chart. Based on the bar chart, it is clear that the data was skewed to the left. This implies that there is a tendency for reviews to be geared towards an overall positive sentiment.

A graph with purple bars

Description automatically generated

**Review lengths and stars**

We summarised the average review length per star category using a mean, standard deviation and the five number summary. The results are as shown below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Stars | Mean | SD | Median | IQR | Q1 | Q3 |
| 1 | 153 | 143 | 109 | 141 | 57 | 198 |
| 2 | 154 | 131 | 117 | 138 | 64 | 202 |
| 3 | 141 | 118 | 109 | 126 | 59 | 185 |
| 4 | 125 | 109 | 95 | 116 | 50 | 166 |
| 5 | 106 | 102 | 75 | 97 | 40 | 137 |

The summary tables shows that the review length highly deviated from the mean scores in each star category. Large deviations reflects that the data deviates from the normal distribution. Therefore, median was the appropriate measure to use to visualise data. Once again when graphing grouped box plot, it reflects a high number of outliers in the average review length per star category as per the below boxplots.

A graph with numbers and lines

Description automatically generated with medium confidence

Therefore, to visualise the summarised median scores, a dot plot was used. The dot plot showed that negative reviews are lengthier on average than positive reviews.

**A graph with black dots

Description automatically generated**

**Are longer review more useful to people?**

To analyse the relationship between vote useful, review length and star rating provided, a grouped scatter plot and a correlation plot was used. The scatter showed that the review length in characters were more in all the star rating provided than the number of votes received indicating the review was useful. This is clearly evident from the many scatter points, which are densely scattered at the bottom of the scatter plot. The graph also provided some insights into the relationship between votes useful and length of review per star rating. In particular, the scatter plot shows weak linear relationship between the variables.

A chart with many dots

Description automatically generated

To gain more insights into the relationship between votes useful and review length per each star rating, we measured each stars rating group individually. Based on the scatter plot, similar findings are identified in that there was a weak linear relationship between vote useful and review length per each star rating category. In particular, the scatter plot shows that the length of reviews in characters are more in higher number of votes received indicating the review was useful. However, this pattern was more favouring in 4- and 5-star ratings, which implies that positive star ratings were associated with more characters than negative star rating.

A graph of a number of dots

Description automatically generated with medium confidence

As the scatterplot can be difficult to interpret, a correlation plot was used to reinforce the analysis the relationship between votes useful and review length as per below. The correlation between variables voted as useful and star rating provided was -0.0490 which shows a fairly negative correlation, while variables voted as useful and length of the review was 0.33 which was a weak positive correlation. The correlation between star rating provided and length of the review was -0.15 which signifies a weak negative correlation. Further to that the correlation plot colour gradient shows the strength of the relationship. Darker the colour the stronger, the fader the colour the weaker the relationship. The colours are overall fade showing an overall weak relationship between the variables.

A screenshot of a graph

Description automatically generated

**Conclusion**

In conclusion reviewers are more positive! With this analysis we confirmed that there are more positive words identified in reviews with 2.75x more positive words than negative words. We identified with the net sentiment that there was a positive trend. This was further reinforced by the average review reflected more positive words used than negative words. The length of majority reviews lies between 48 to 165 characters and that longer reviews tend to have a higher star rating. Finally, there was a weak positive relationship between length of a review and people finding it useful.