Task 8: Analysis Report

This analysis was based on data that was collected from Amazon. The following product information was scraped from each product listing: the name, the category, the number of positive or negative ratings, price, the reviews of the product and a URL to the product's page.

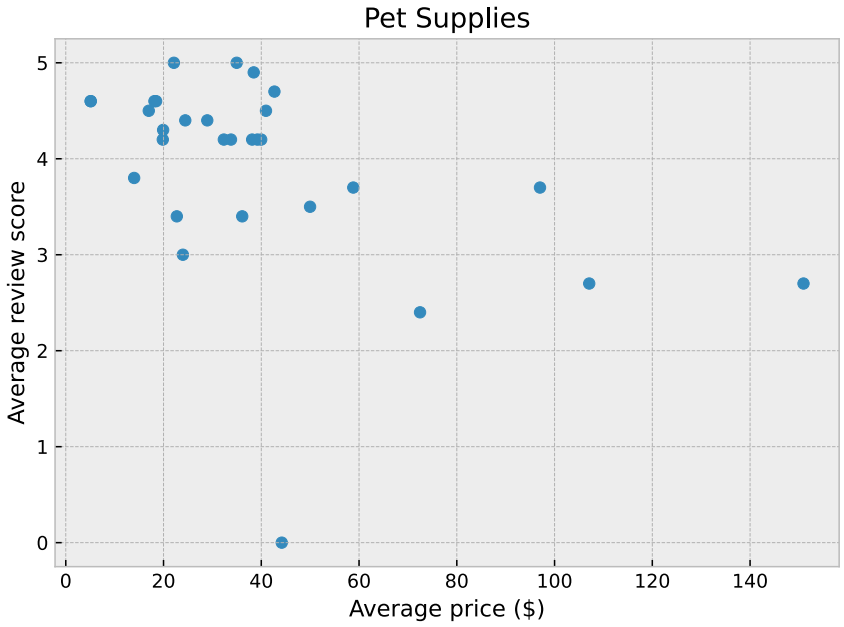

Figure 1: Task4

Figure 1 illustrates the relationship between the average review score and the average price for each product that is categorised as pet supplies. Based on the underlying data, I believed that a scatter plot was an effective way to visualise this relationship. Figure 1 highlights that more expensive listings tend to have a lower average review score, however more data is needed to reinforce this.

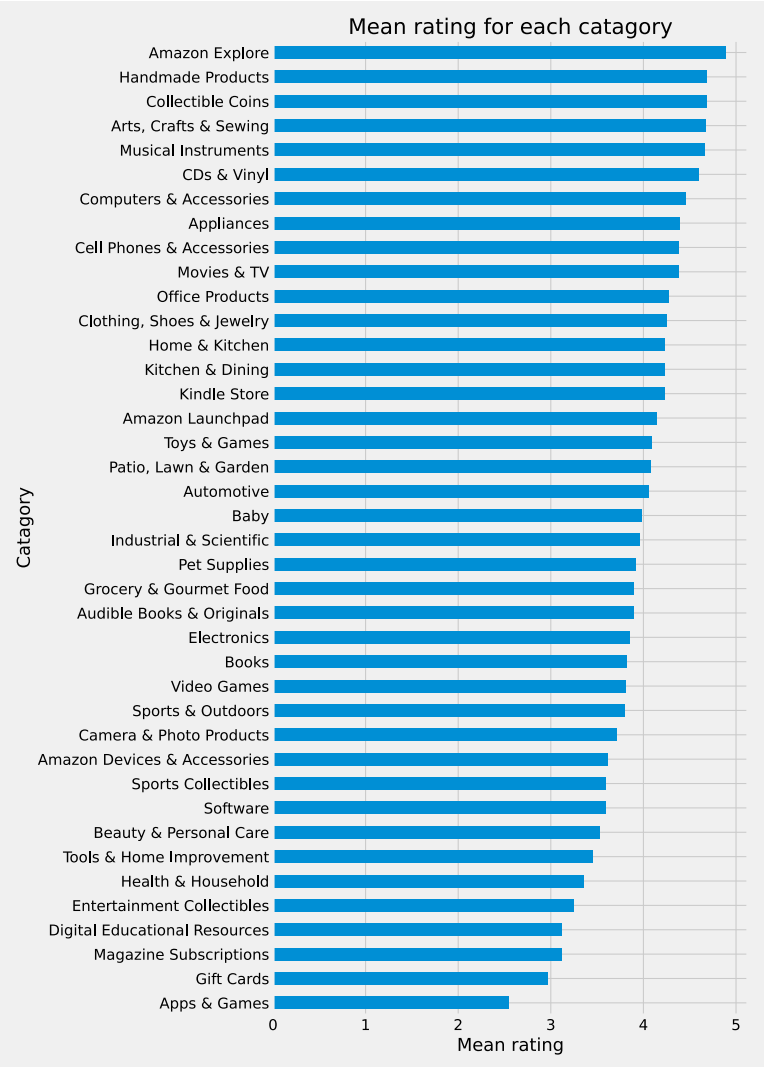

Figure 2: Task5

The category with the highest mean rating across its range was ‘Amazon Explore’ which had a mean rating of 4.88 and the category with the lowest mean rating 2.54 was ‘Apps and Games’. Interestingly, it appears that digital categories such as ‘Software’ and ‘Apps & Games’ tend to have mean rating that is below the average. Further analysis needs to be done to reinforce that statement.

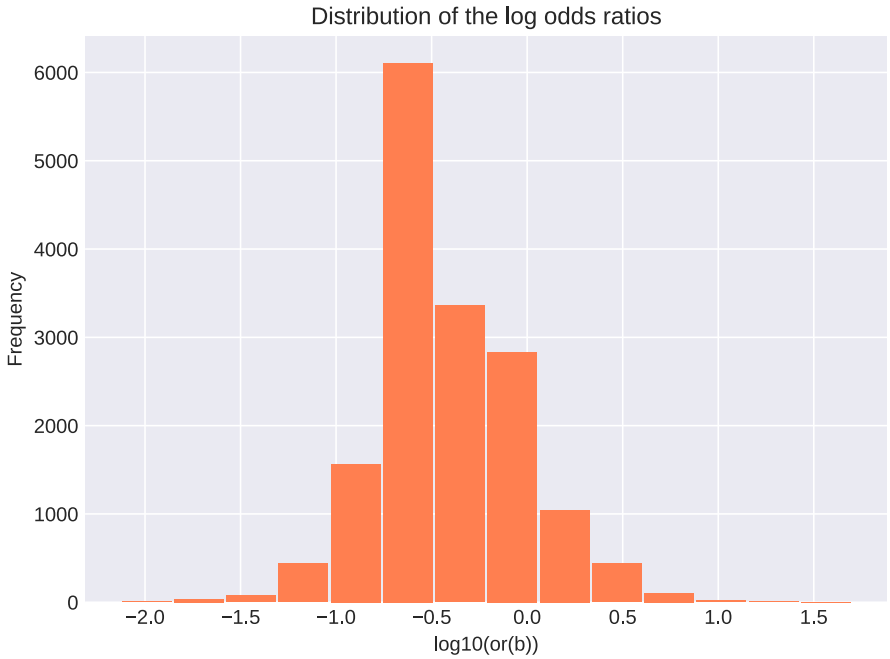

Figure 3: Task7b

Figure 3 depicts the distribution of the log odds ratio for bigrams in the Amazon reviews vocabulary. The distribution of the log odds ratios for all bigrams is approximately symmetrical. Closely looking at the histogram, it appears that most of the bigrams in the vocabulary are frequently used in negative reviews, since more than half of the bigrams have a negative log odds ratio. Furthermore, a significant number of bigrams have a lor(b) in the -1 to -0.5 range.

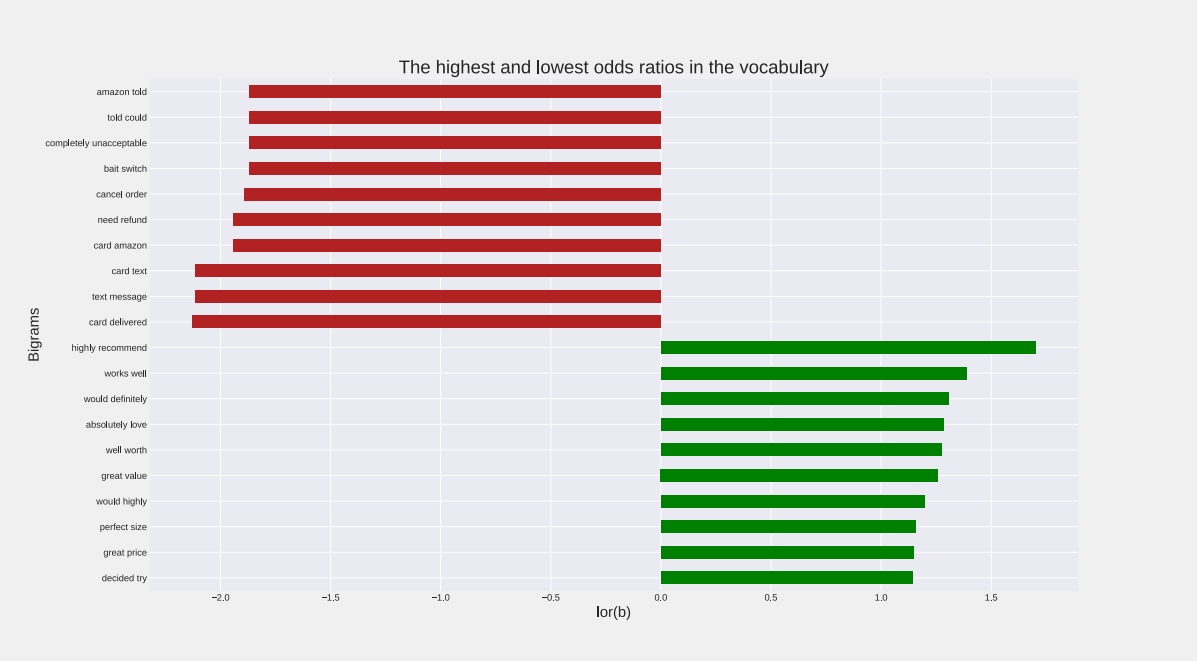

Figure 4:Task 7c

The bigrams that my program determined are the most indicative of a positive review seems correct. 'Highly recommend' and 'well worth' both have a positive meaning. The order of the positive bigrams was not surprising.

However, the output for the ten most negative bigrams was unexpected. Six out of the ten bigrams were negative, whilst the other four were neutral. Interestingly, three out of the four most negative bigrams include the word 'card'. A potential reason for this might be the reviews from the 'Gift Card' category. Figure 2 highlights that the 'Gift Cards' category has the second lowest mean rating. My speculation is that most occurrences of the bigrams containing the word 'card' came from reviews that were about gift cards, which were mostly negative.

Limitations

With current processing techniques, it's hard to determine which bigrams are indicative of a positive review. From the analysis above, it is evident that further text processing is needed. One modification to the text processing step that will increase the accuracy of the sentimental analysis, is to add the names of products and categories into the list of stop-words. Additionally, the accuracy of this report was also limited by the dataset. Products with missing attributes were discarded. This could be an issue if those values were not missing completely at random.