Nigel Loh Yesheng Student Number: 1300656

Assignment 1 Analysis Report

Input Dataset/Parameters:

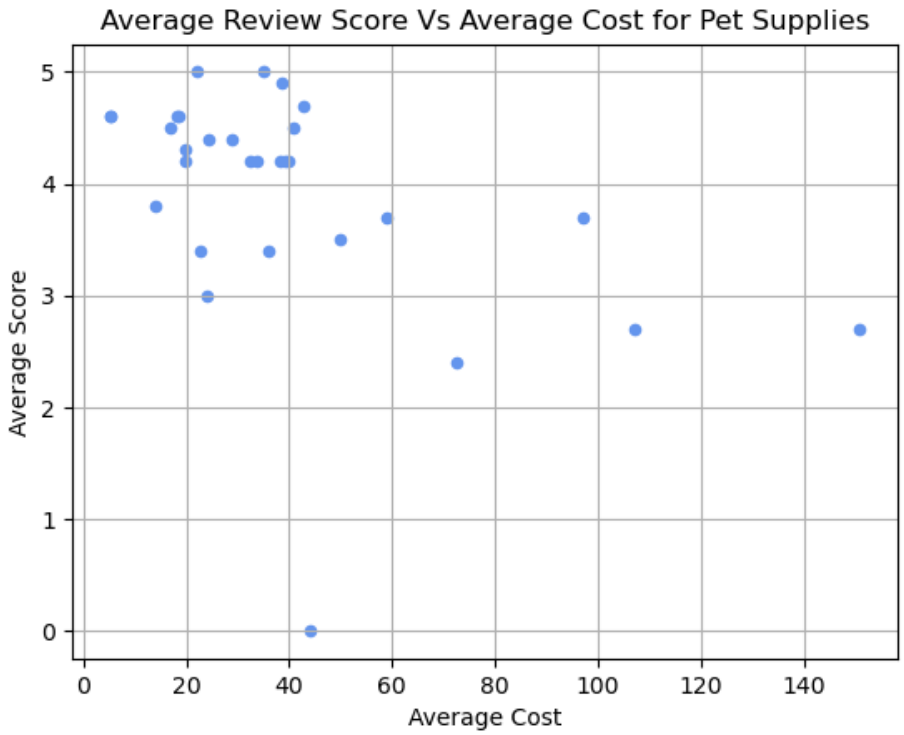
A dataset comprising reviews of more than 1000 products:

- `/course/data/dataset.csv`

The dataset is a CSV file containing several fields:

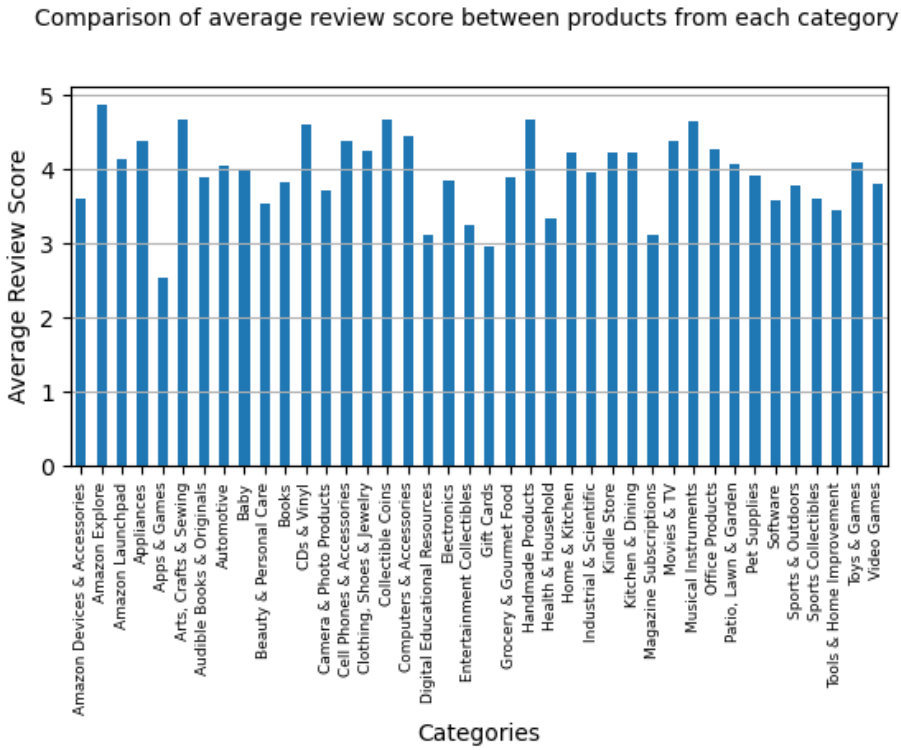
- ID: Indicates a unique ID number for each product in the dataset
- product_name: The name of the product, as displayed on the Amazon website
- category: Each product is assigned to a single category indicating the type of product
- noRatings: This represents the number of positive or negative ratings (not reviews) of the product
- cost: How much the product sells for. Note that many products have a price range rather than a single price, typically meaning they can be customized when purchased.
- REVIEWLIST: A list of reviews for the product, expressed as a JSON string
- product_url: A link to the product's page on Amazon

Comparison between average price with average review score for each product in 'Pet Supplies' (task4.png):



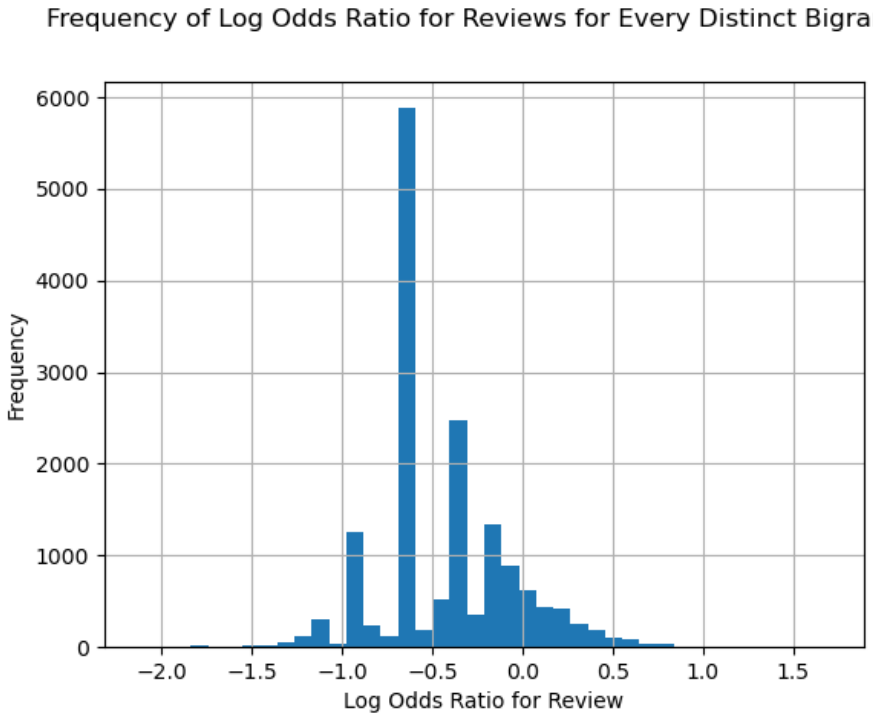
According to the observation in 'task4.png' above, it seems that the products in ('Pet Supplies') tend to have a higher average review score with a price range of $20.00 to $40.00. Therefore, we can observe that a lower average price range tends to give a higher review score.

Comparison of average review scores between products from each category (task5.png)



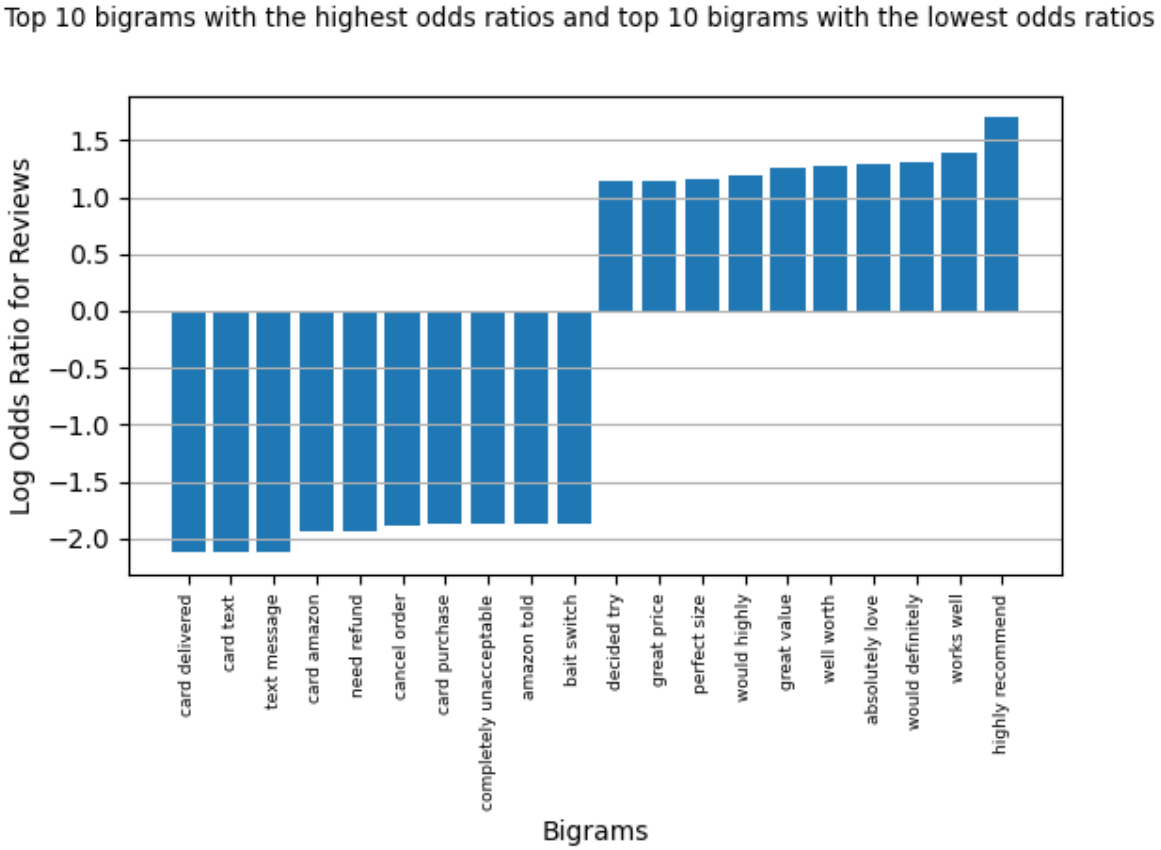
From the information above (task5.png), we can observe that majority of the products have an average review score of 3 and above. There were only two categories, 'Apps and Games' and 'Gift Cards' which have an average review score below 3.

Frequency of 'Log Odds Ratio' for Reviews for every distinct bigram (task7b.png)



The histogram above (task7b.png) is skewed to the left, which means that more distinct bigrams appear in negative reviews more frequently. This could also mean the positive reviews uses a more diversified and wider range of vocabulary.

The Top 10 bigrams with the highest odds ratios, and the top 10 bigrams with the lowest odds ratios (task7c.png)



I agree that the bigram ‘highly recommended’ is the most indicative of positive reviews and the bigram ‘card delivered’ is the most indicative of negative reviews. For ‘card delivered’, this seems to be true as shown in task5.png. The category ‘Gift Cards’ is the second lowest for total average review score which is below 3. ‘The bigram ‘card text’ is also appearing as well for negative reviews. It seems that the word ‘Card’ happens to be common for the negative reviews, hence odd ratios look out for that word and associate it with the negative reviews.

There are some limitations of the dataset as although we look at the frequency of the bigrams’ appearances, the word in the bigram could possibly denote either positivity or negativity reviews. For example, the bigrams such as ‘card purchase’ and ‘amazon told’ is under negative reviews, however, definitions do not necessarily mean it is negative feedback. Basically, it is important that we go beyond the frequency, and account for the meaning of the word. Lemmatization or stemming could be used to group appropriate words together to improve analysis. Since we only consider the frequency of the bigrams and not their meaning, the processing methods employed are rather constrained. Analyzing the meaning of the bigrams from the reviews would be much more accurate and produce better results.