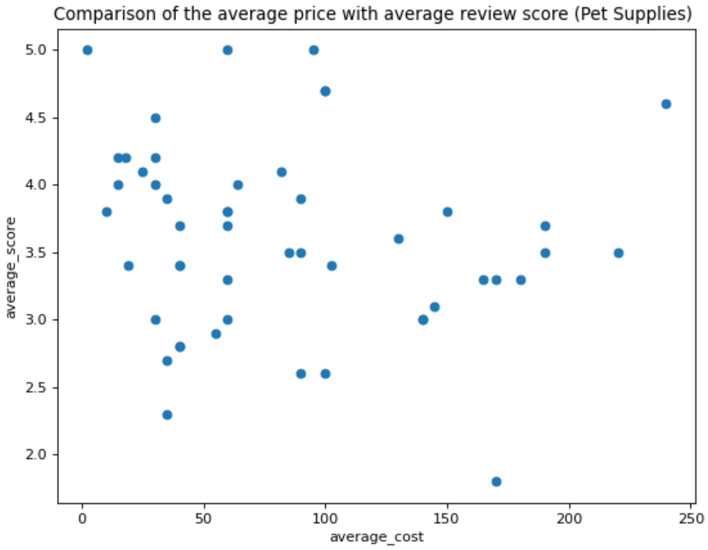
Task 8: Analysis Report

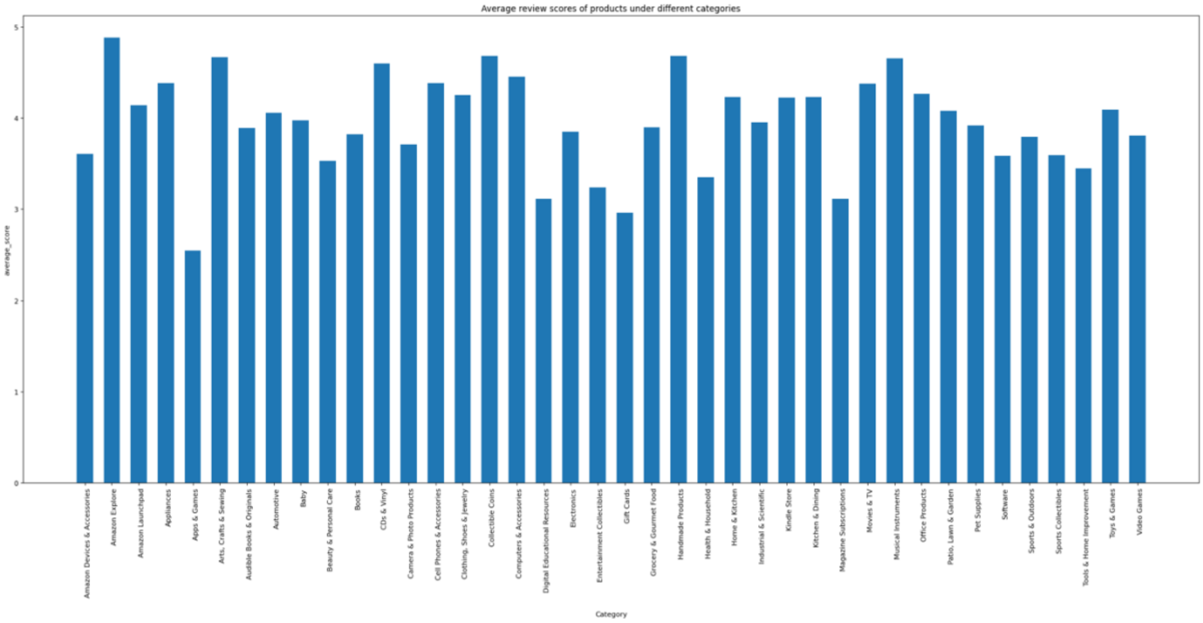
The data source for this Amazon product review analysis was retrieved from the Amazon online store.

The plot generated for task 4 is a scatterplot representing the average price comparison with the average review score on the Pet Supplies categories.



According to the scatterplot generated for task 4, there appears to be no relationship between the review scores and the average price. The data points are more clustered as a group of data points.

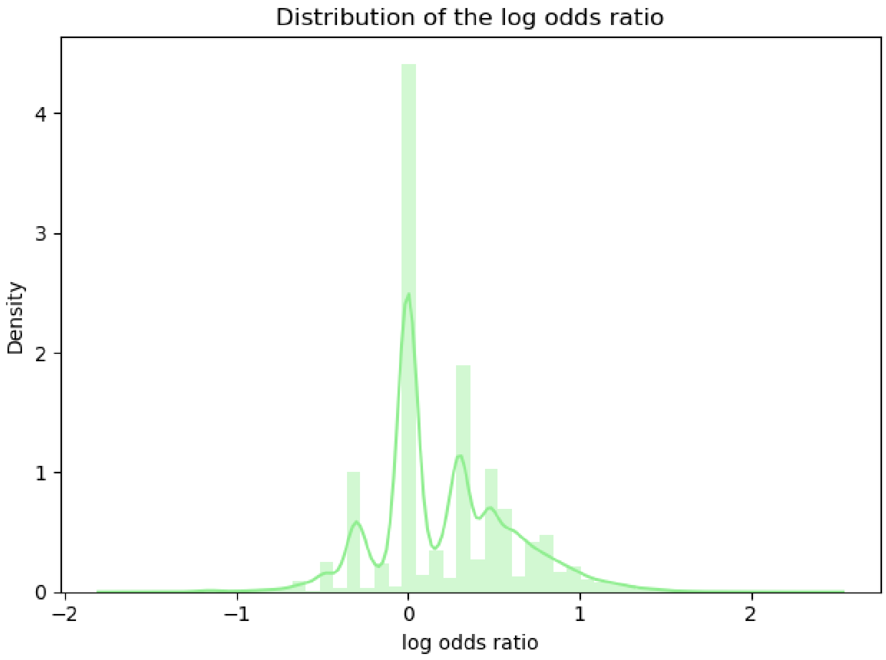
The plot for task 5 is a bar graph representing the average review scores of products under different categories.



From the task5.png plot, the Amazon Explore category appears to have the highest review score, whereas Apps & Games appears to have the lowest review score. Most of the categories exceed an average review score of 3. Only Apps & Games, and Gift Cards have average scores below 3. These categories are digital goods, which the digital goods may not be delivered to the expected quality level.

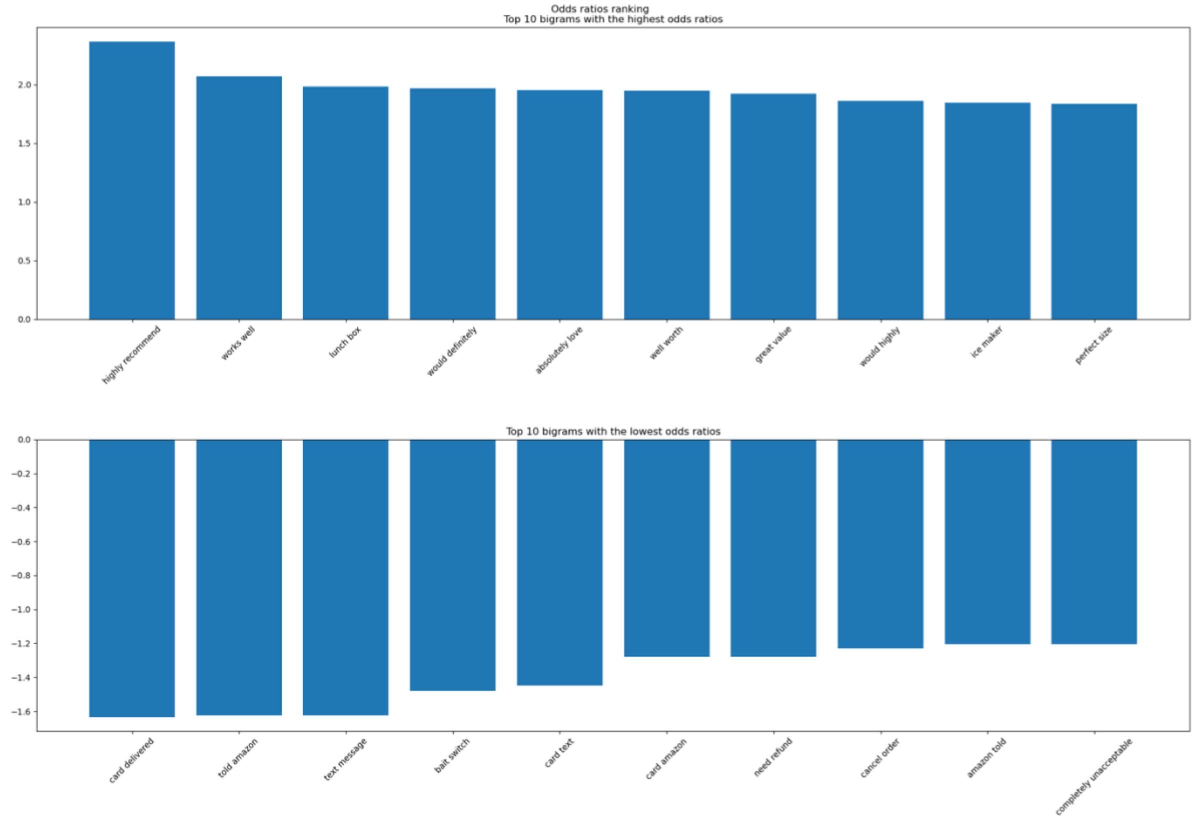
On the other hand, the categories with higher review scores are Arts, Craft & Sewing, Collectible Coins, CDs & Vinyls, and Handmade products. Most of these categories are artistic. The subjective interpretation of the values of art products and collectibles may have impacted the review score.

The plot generated for task 7b is a density graph representing the distribution of the log odds ratio.



The distribution of the log odds ratios resembles a normal distribution. The reason behind the zero log odds ratio having the highest density may originate from the neutral and descriptive words of the products in a review. The bigrams with very high or very low odds ratios are not common as they primarily represent customers' strong sentiment toward the products.

The plot generated for task 7c is a bar graph representing the top 10 bigrams with the highest and lowest odds ratios.



I agree with most of the indicative bigrams for positive and negative reviews. Bigrams such as "highly recommend", "words well", "would definitely", "absolutely love", "well worth", "great value", "would highly", "perfect size" all have positive connotations attached to them. Although bigrams such as "lunch box" and "ice maker" may be the product names that do not indicate any positive meaning.

With the negative reviews, bigrams such as "bait switch", "need refund", "cancel order", and "completely unacceptable" appear to have negative connotations attached. However, with the negative reviews, the bigrams with the lowest odds ratios appear more random and not indicative.
The limitation of the dataset is that it may have an imbalanced weight of different categories. Some bigrams from task7c.png, like "lunch box", and "ice maker", may have come from the kitchen appliances categories. This limitation may be fixed with a larger and more balanced dataset.

With the pre-processing steps, unigram may be more helpful and accurate than bigram in capturing the exact words that appear the most in positive or negative reviews.

In addition, the number of positive and negative reviews is also an essential factor to consider. If there are only a few positive reviews but hundreds of negative reviews, then the analysis of n-grams of positive reviews will not be accurate. There may be bias in the result.