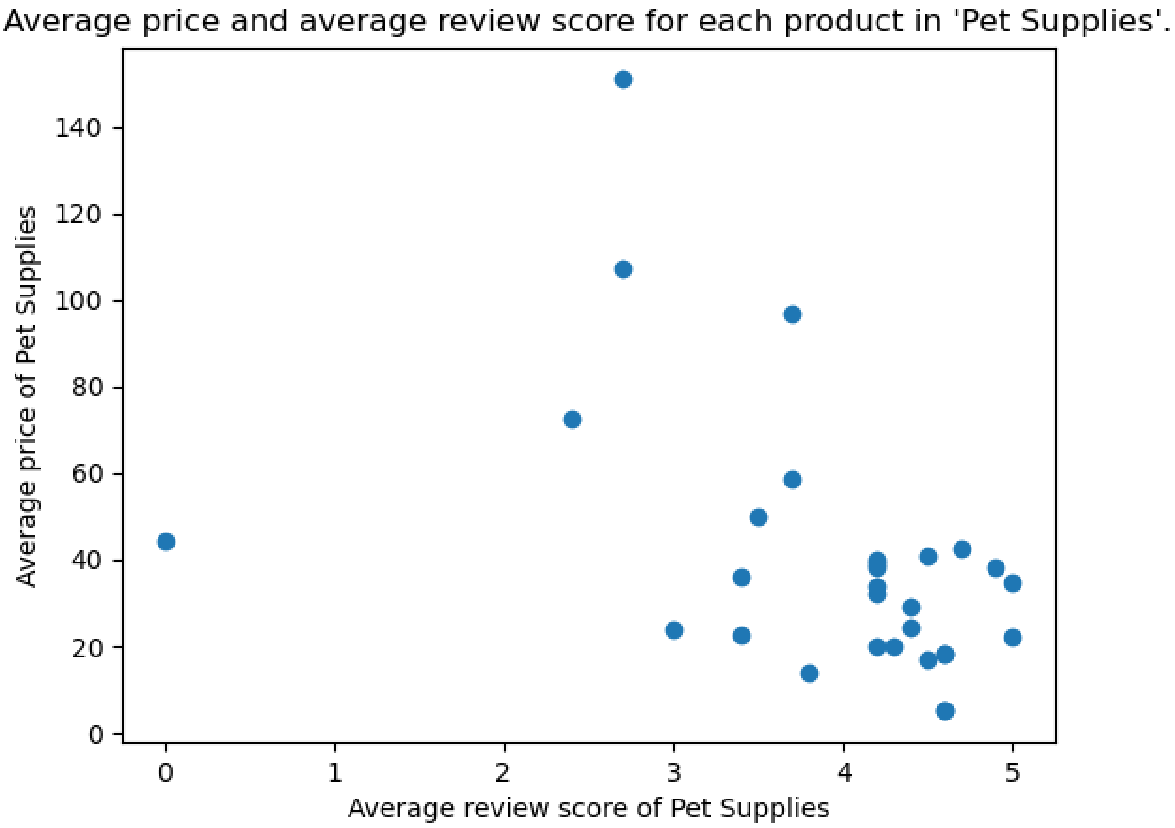
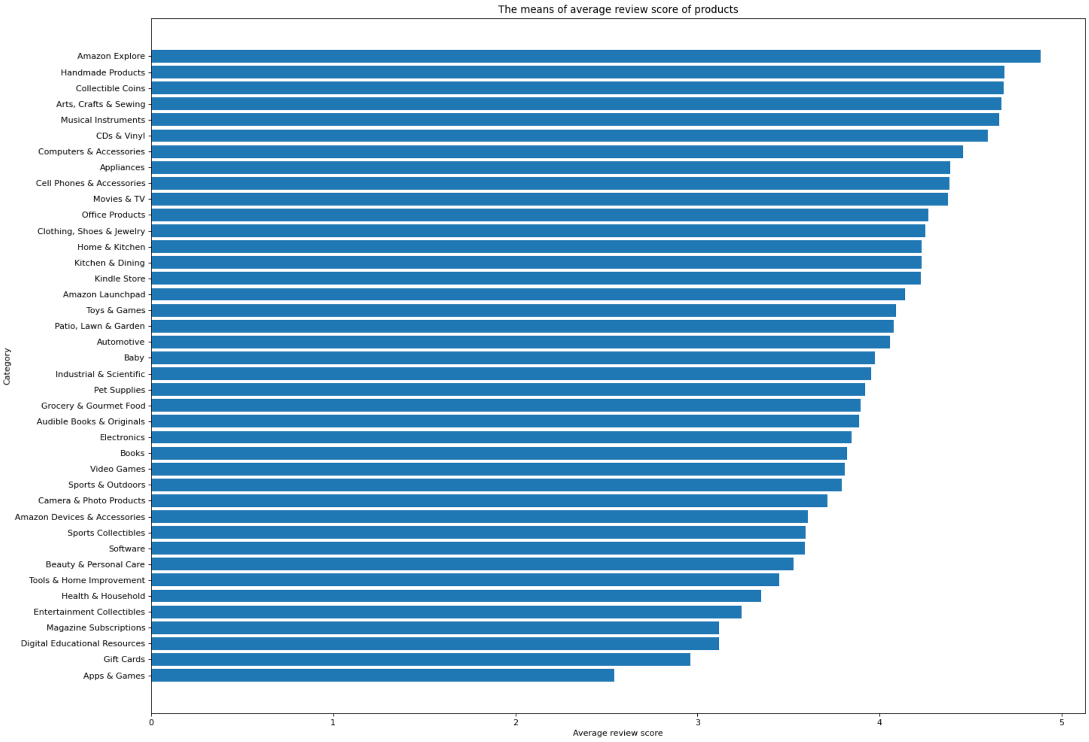
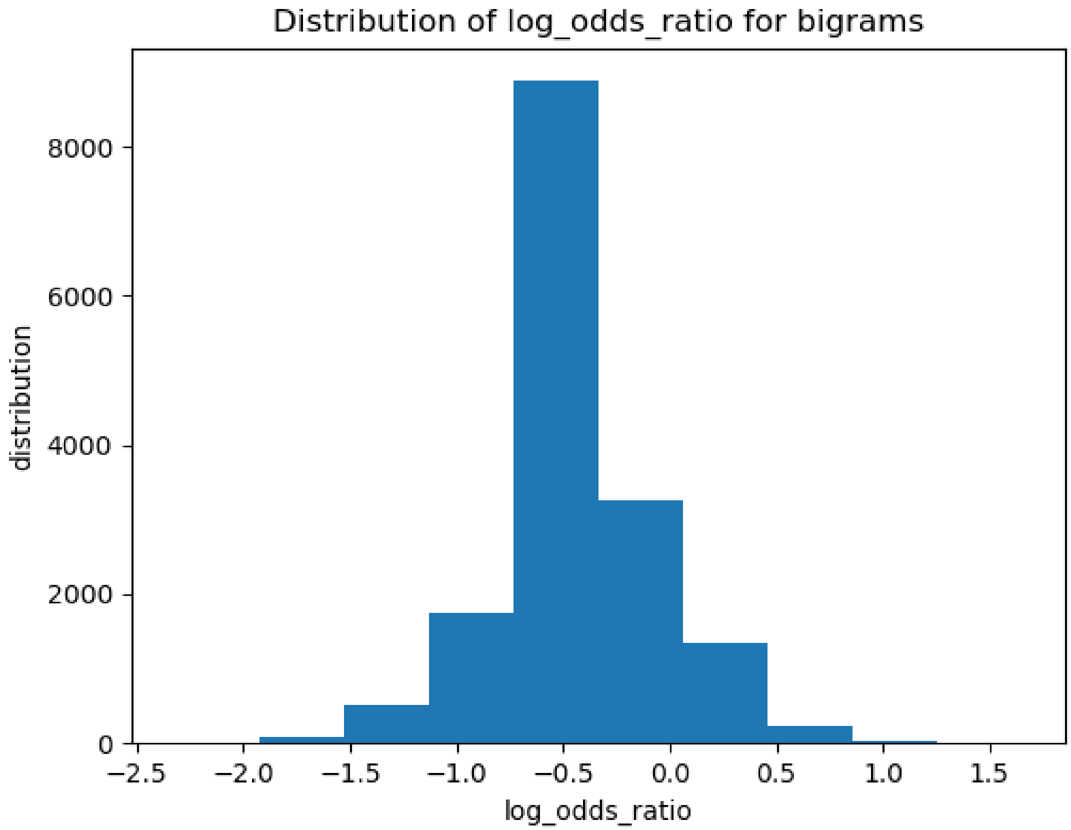


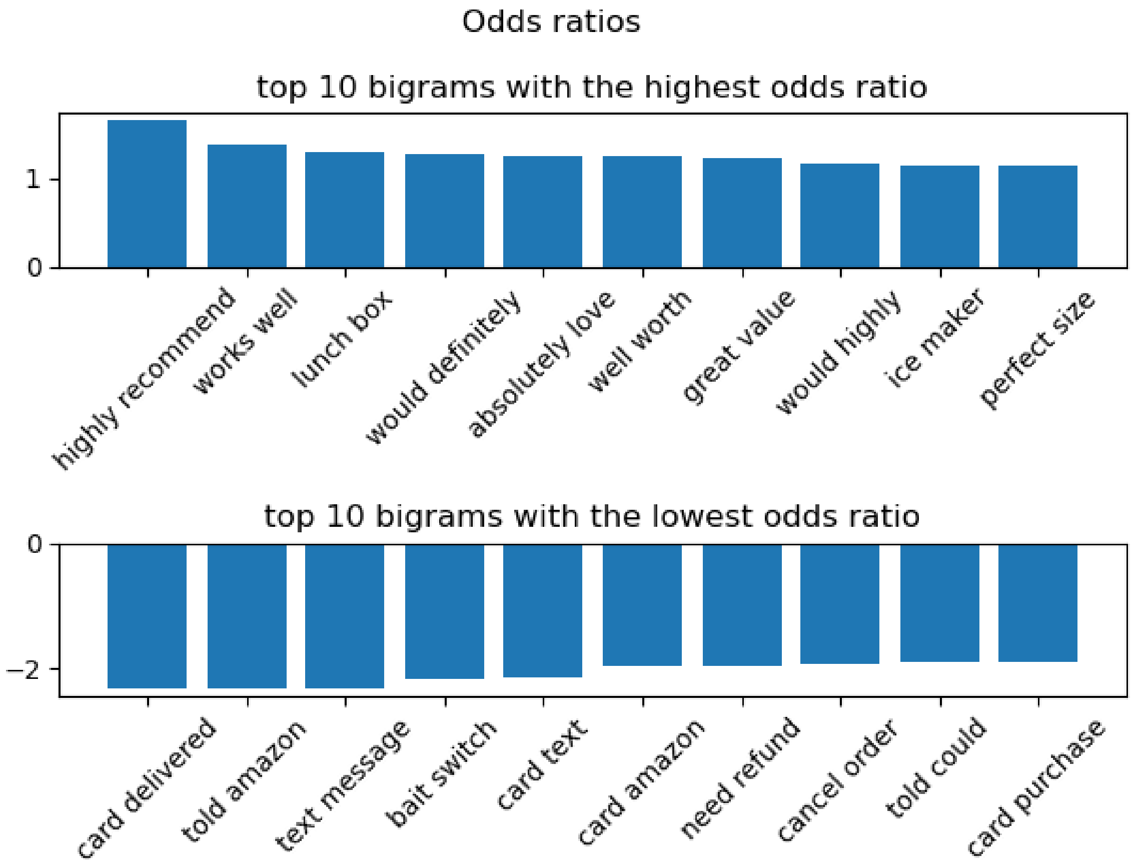
- Above is task4.png, a scatter plot which shows the relationship between the average price and average review score for each product in ‘Pet Supplies’. The data source of this graph comes from the calculated average score and average price in task2 and task3 separately.
  By using the scatter plot, it could be observed that most of the dots are clustered in the bottom right corner. Therefore, the relationship between average price and average review score in ‘Pet Supplies’ should be a negative correlation.



- Above is task5.png, a bar chart which shows the means of average review score of each product.
  The data source of this graph comes from the calculated average review score of task2.
  By using the bar chart, it shows that: The overall average review score is between 2.5 and 5. Besides, “Amazon Explore” has the highest mean of average review score which almost reach 5 while “Apps & Games” has the lowest mean of average review score which is about 2.5.



- Above is task7b.png, a histogram which shows the distribution of log odds ratio for bigrams in the amazon dataset.
  The data source of this graph is from task6's data of score and bigrams.
  This histogram tends to be a shape of symmetric unimodal. Because the log odds ratio appears at regular frequencies and the most data converge of distribution is more than 8000 bigrams.



- Above is task7c.png. It is a bar chart which shows the top 10 bigrams with the highest odds ratios, and the top 10 bigrams with the lowest odds ratios.
  The data source of this graph is from task6's data of score and bigrams. All of the 5-star reviews are considered as positive review while all of 1-star reviews are considered as negative review.
  For positive review in the top 10 bigrams, it could be the "most indicative" bigrams since almost all of the reviews could reflect proper positive reviews except for "lunch box" and "ice maker".
  However for negative review in the top 10 bigrams, only "cancel order" and "need refund" have the sense of negative review. For other bigrams, it is not clear what they are intended to convey. So it could hardly be the "most indicative" bigrams.
  In conclusion, I agree with the "most indicative" bigrams for positive reviews but disagree for negative reviews.

Limitations:

- The first limitation could be the format of the dataset which is the format of csv. Since when reviewing the dataset, it must perform an additional format conversion to complete the following series of analysis. But if the dataset is a json format, from the point of view of code implementation, it could be easier and more efficient as it can be manipulated directly.

- Secondly, there is too much content that is not relevant to the information needed to be filtered out. Through the top 10 negative review of task7c.png, most of the bigrams have no connection with negative reviews which could be confusing. Therefore, in the future, to better define the dividing line between positive and negative reviews. The dataset could make the products be representative first by using statistical model, and delete the irrelevant words.