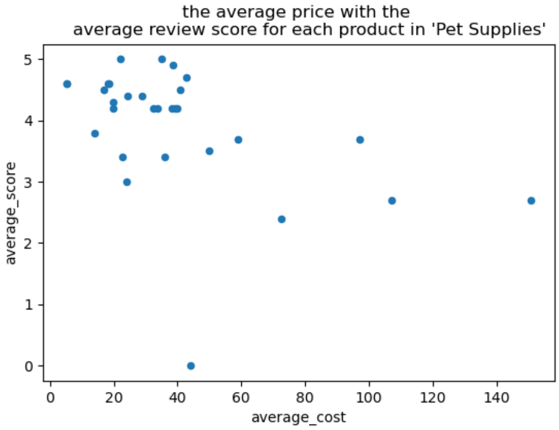
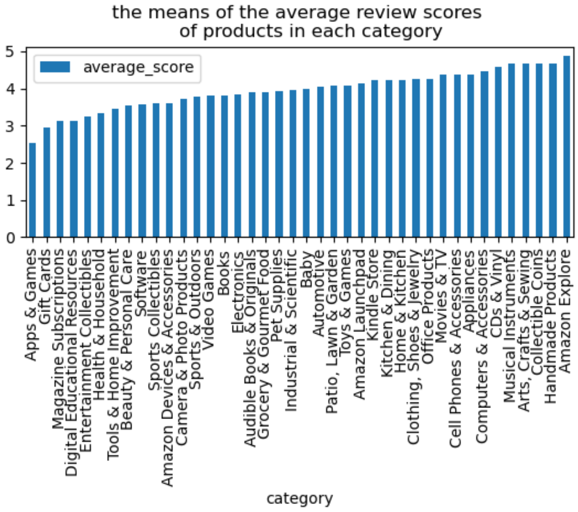
The raw data adopted was the reviews from Amazon products. This dataset was a CSV file containing the information of category, rating, cost, along many other product details. For Task4, a Pandas Dataframe was produced using the average price and average review score for each product in ‘Pet Supplies’. A scatter plot was introduced to compare numerical variables. For Task5, a dictionary with category name as key and mean average review score as the value was employed to generate Pandas series. A bar chart was plotted to highlight the numerical observations. Logistic regression and binning preprocessing strategy were used in Task7, three bar charts determined the distribution and reliability of bigram for reviews respectively.



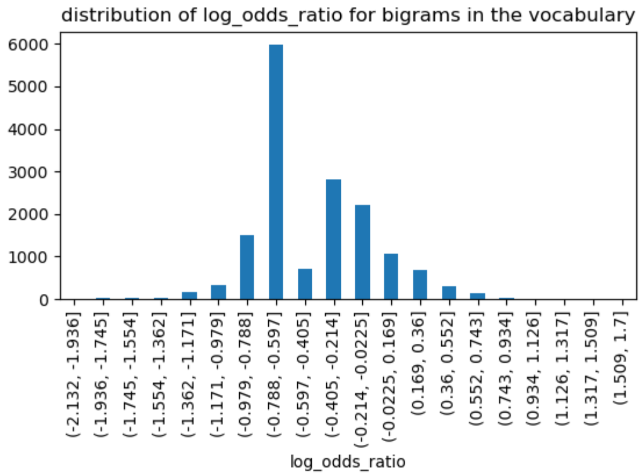
TASK4

The scatter plot in Task4 demonstrated a moderate linear negative correlation between ‘average_cost’ and ‘average_score’. There was an outlier on the bottom left. We concluded from the obvious trend –The relationship between the two variables was negative, as the average price increased, the average score of the product review decreased.



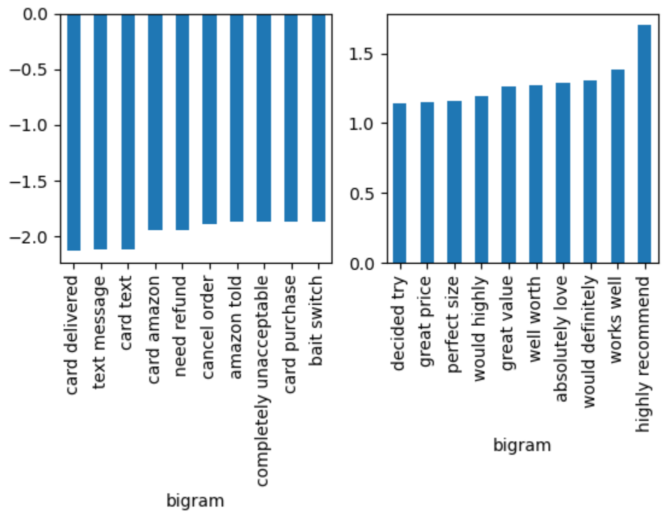
TASK5

The bar chart outlined the differences between the mean average score received by each category. We noticed that ‘Apple & Games’ had the worst mean review score of 2.6, and ‘Amazon Explore’ received the highest mean review score of 4.9. Generally, all categories from the examined Amazon products achieved satisfying review scores, such as none of them were scored below the median of 2.5.



TASK7B

A preprocessing method of binning was employed to enhance visuality. After dividing the range into 20 equidistant intervals, the visualization of the bar chart illustrated a normal distribution of the log odds ratios over all bigrams.



TASK7C

Both bar charts gave us a general idea regarding the two review types. However, the graphs presented a less indicative result, since the small value of odd ratios severely affected reliability. In the Top 10 lowest odd ratios graph, bigrams such as ‘amazon told’, ‘card purchase’ and ‘bait switch’ shared the same ratio of ‘-1.8665’. As the difference between bigrams might be minimized with rounding, the graph would miss other meaningful bigrams. Additionally, the highly correlated bigrams also decrease readability and precision, thus they were less indicative.

The main limitation of the dataset was integrity issues, such as products missing reviews and others missing prices. To solve this problem, an accurate data cleaning would be required to remove any missing or corrupted data from preprocessing step. Another limitation was the unreliability of social network data resources because we could not guarantee all product reviews were real and precise.

The processing techniques had limitations against outliers. In Task 5, the mean of average review scores was susceptible to the influence of outliers. We should take deviant data out from consideration. Moreover, the log odds ratio method from Task7 showed less accuracy. When building a binary logistic regression model to indicate positive and negative review, there might not always be linearity hence hard to fit the dataset into the correct distribution. The tree-based method would be a wise alternative to solve such problems.

(Words:499)