RATING PREDICTIONS FROM REVIEWS GIVEN TO PRODUCTS IN ONLINE MARKETS

CPSC-6300 Applied Data Science

Spring 2023

Checkpoint 1

DMITRY LUKYANOV
SEJAL BANSAL
ADITHYA RAVI
SHAREEF SHAIK

The dataset peculiarities

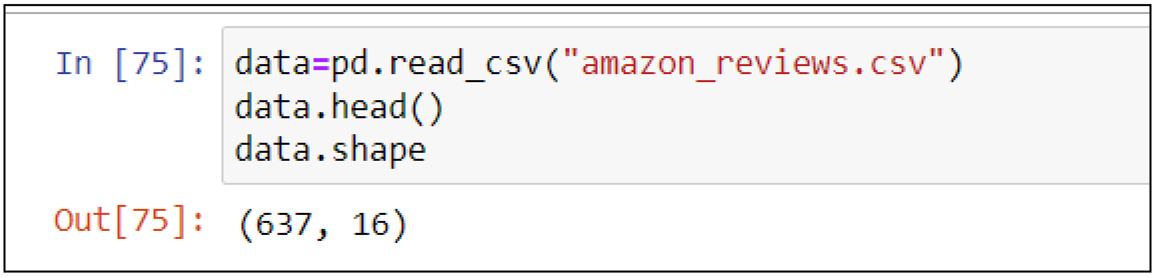
1. Almost all reviews contain only evaluation of products with average presence of non-product aspects in reviews approximately equal 1 to 100. It makes it virtually impossible to utilize the dataset for training a model that would be able to distinguish such facets. To be addressed during the next stages of the project.
2. The dataset is highly imbalanced in rating distribution. To be addressed during the next stages of the project.

EDA Unit

Our unit of analysis for the Amazon Review Dataset is at the review level. Each row in the dataset represents a single review, with features such as the product category, review text, rating and so on. Therefore, the analysis performed on this dataset focuses on the characteristics and distribution of individual reviews, rather than on the products or reviewers themselves.

Dataset size

The dataset contains a total of 637 reviews, with 637 reviews for each of the following categories: ‘wireless’, ‘video games’, ‘shoes’. Therefore, the total number of observations (i.e., the number of rows in the dataset) is 637.



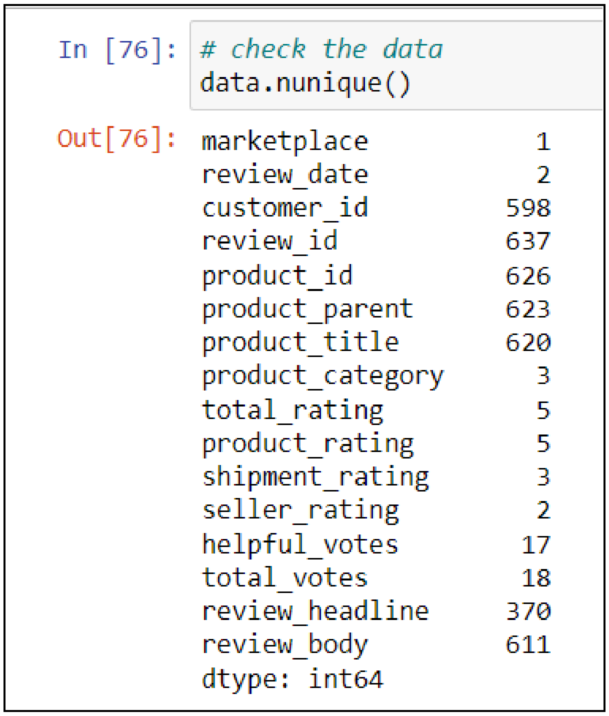
```
In [75]: data=pd.read_csv("amazon_reviews.csv")
         data.head()
         data.shape

Out[75]: (637, 16)
```

Fig 1: Dataset size

Unique observations

The below table specifies the number of unique observations:



```
In [76]: # check the data
         data.nunique()

Out[76]: marketplace           1
         review_date           2
         customer_id         598
         review_id           637
         product_id          626
         product_parent      623
         product_title       620
         product_category      3
         total_rating          5
         product_rating        5
         shipment_rating       3
         seller_rating         2
         helpful_votes        17
         total_votes          18
         review_headline     370
         review_body         611
         dtype: int64
```

Fig 2: No.of unique observations

Covered period of time

All reviews within the sample fall on 08/31/2015.

Data cleaning

Data cleaning is an important step in any data analysis project to ensure that the data is accurate, complete, and consistent. The following steps are included in the data cleaning procedure for our dataset:

1. Remove duplicates and null values: Deleted duplicate reviews to prevent skewing the analysis and replace null values with 0. We used drop duplicates() and replace() methods.
2. Getting rid of punctuation and special characters: Any punctuation marks, special characters, or non-alphanumeric characters were removed from the text.
3. Lowercase conversion: To guarantee data integrity, we converted all text to lowercase. This was accomplished by calling the lower() function on the 'review' column.
4. Stemming and Lemmatization: Used stemming and lemmatization to break words down into their root form.

Unnecessary columns were dropped

1. id: has no value to the project
2. marketplace: all values are the same

Columns’ format has been unified

1. review_date: had two different formatting

Types were adjusted

Column	**Original**	**Target**
review_date	object	datetime

customer_id	int	string
review_id	object	string
product_id	object	string
product_parent	int	string
product_title	object	string
product_category	object	string
total_rating	int	string
product_rating	int	string
shipment_rating	float	string
seller_rating	float	string
review_headline	object	string
review_body	object	string

Missed values

There were no missing values except missing labels shipment_rating and seller_rating. Corresponding units were labeled manually with marking lack of relevant aspects in reviews as “NA”.

Reviews cleaning

Texts of reviews were cleaned from html-tags.

Reviews normalizing

- Text was lowercase.
- Punctuation marks were removed.
- Contradictions in text were expanded.

Visualizations and Analysis

Words

i) In total, the dataset contains 22496 words with the distribution on the unique and non-unique words. The following visualizations are on the unique and non unique words in which, the pie chart depicts the percentages of both respectively, and further on the bar graph plots frequently used words vs. the number of times they have been used.

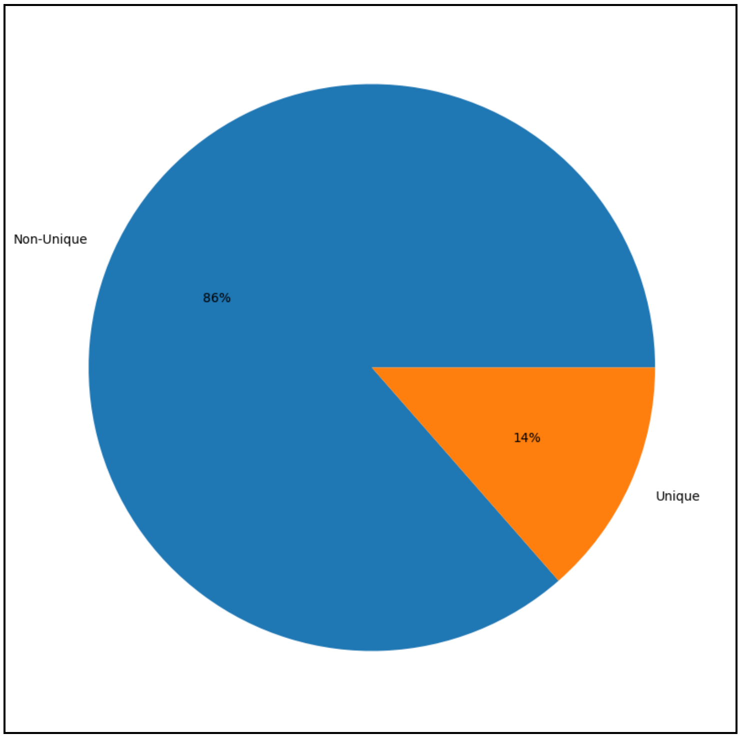


Fig 3: Proportion of unique words in the dataset

The blue slice represents the number of words that appear more than once in the dataset, or non-unique words. This includes common words like "the", "and", "a", etc., as well as frequently occurring words like "good", "product", "quality", etc.The orange slice represents the number of words that appear only once in the dataset, or unique words. These are less common words that are used only in a single review or a few reviews.

The percentages on the chart indicate the proportion of each slice relative to the total number of words in the dataset. For example, if the blue slice is 80% and the orange slice is 20%, it means that 80% of the words in the dataset are non-unique (appear more than once) and 20% are unique (appear only once).

ii) The next words are the most frequent in the reviews:

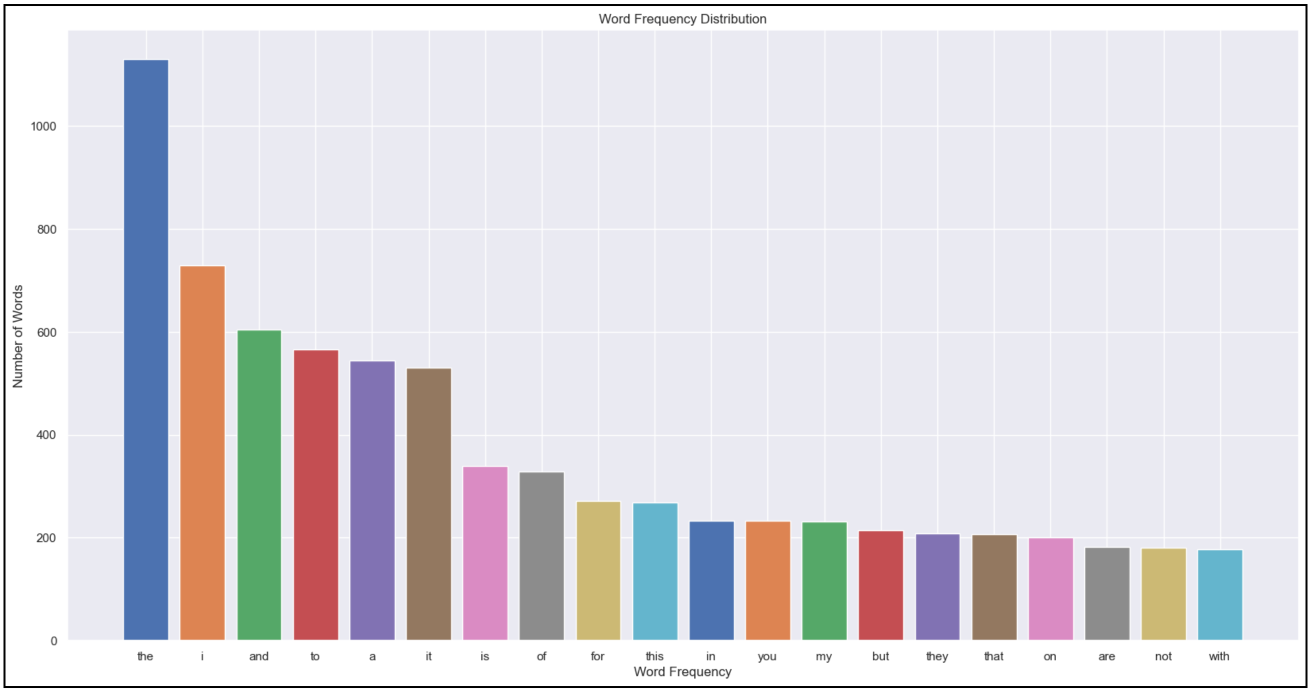


Fig 4: Frequency Distribution of Top Words

The chart displays the frequency of the top words in the dataset on the y-axis and the actual words on the x-axis. Each bar represents a single word and its frequency in the dataset. The height of the bar represents the frequency of that word, while the x-axis label displays the word itself. Overall, this chart provides a visual representation of the most common words in the Amazon review dataset and how often they appear in the reviews.

iii) Reviews' lengths distribution is the next, the chart displays the number of reviews on the y-axis and the length of the reviews (in terms of the number of words) on the x-axis. The histogram is divided into bins, where each bin represents a range of review lengths. The height of each bar represents the number of reviews that fall within that particular range of review lengths.

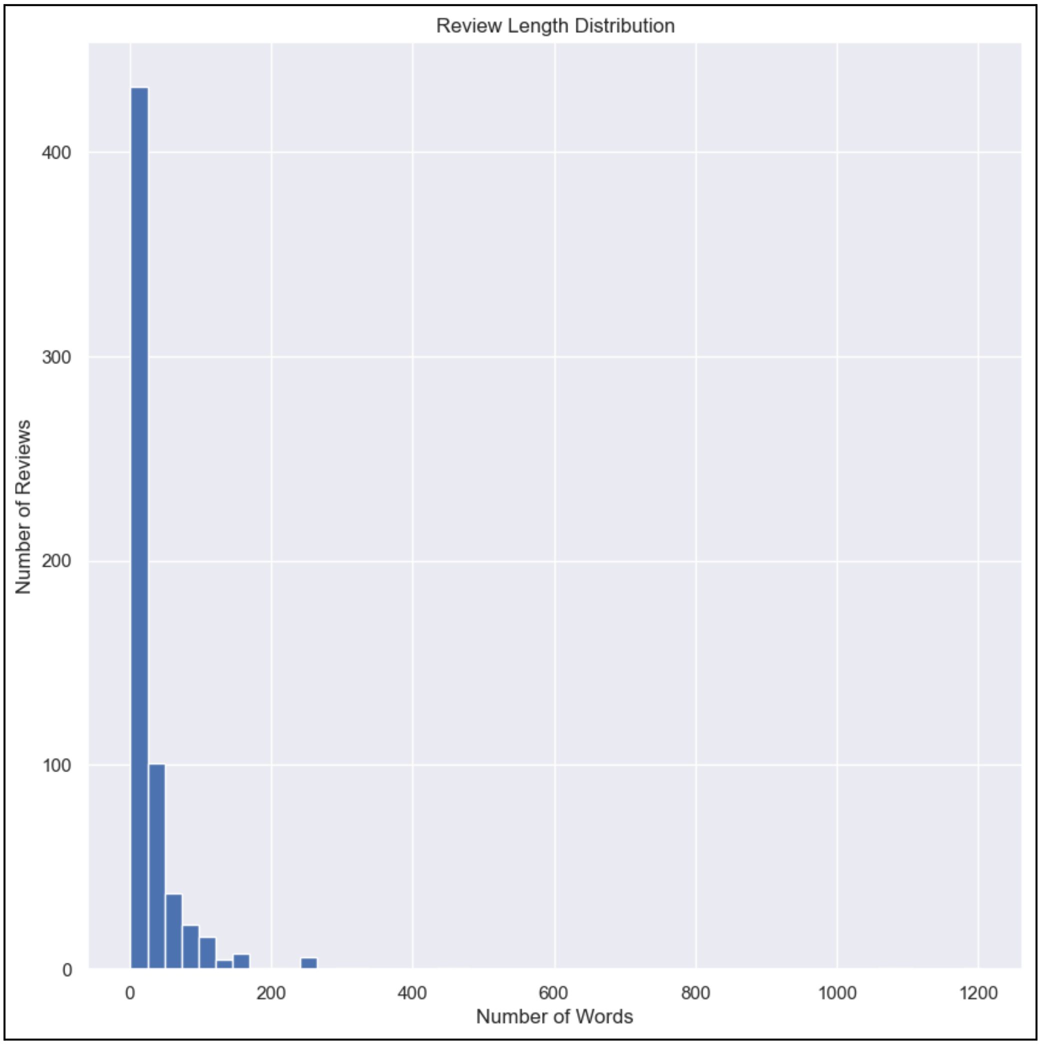


Fig 5:: Distribution of review lengths in the dataset

Categories

iv)Categories of products in the dataset sample are distributed in the next way.

The chart displays the number of products on the y-axis and the product categories on the x-axis. Each bar represents a product category, and the height of the bar represents the number of products in that category. The product categories are ordered by frequency, with the most common category at the top and the least common category at the bottom.

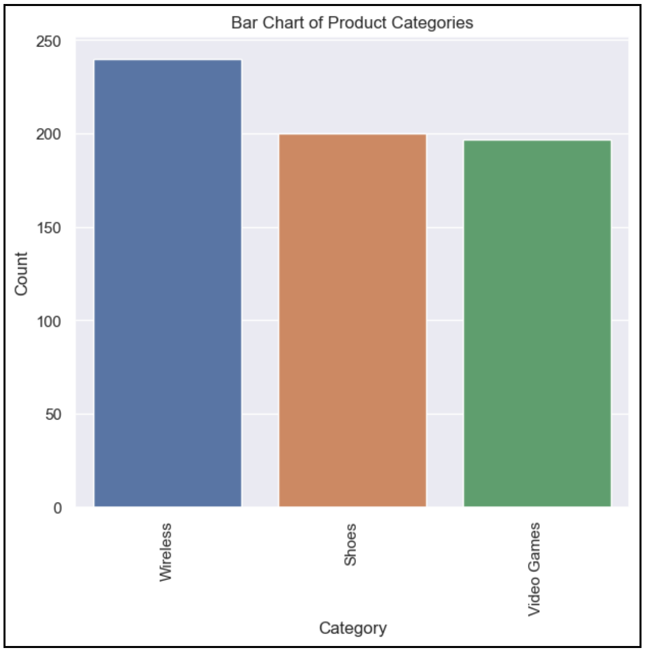


Fig 6: Bar chart of most common product categories

v) Correlation

The correlation among features if we consider rating-related ones as quantitative is the next. The chart displays a color-coded matrix that shows the pairwise correlation coefficients between all features in the dataset, except for the 'id', 'customer_id', and 'product_parent' columns. The correlation coefficient measures the strength of the linear relationship between two variables, ranging from -1 (strong negative correlation) to 1 (strong positive correlation).

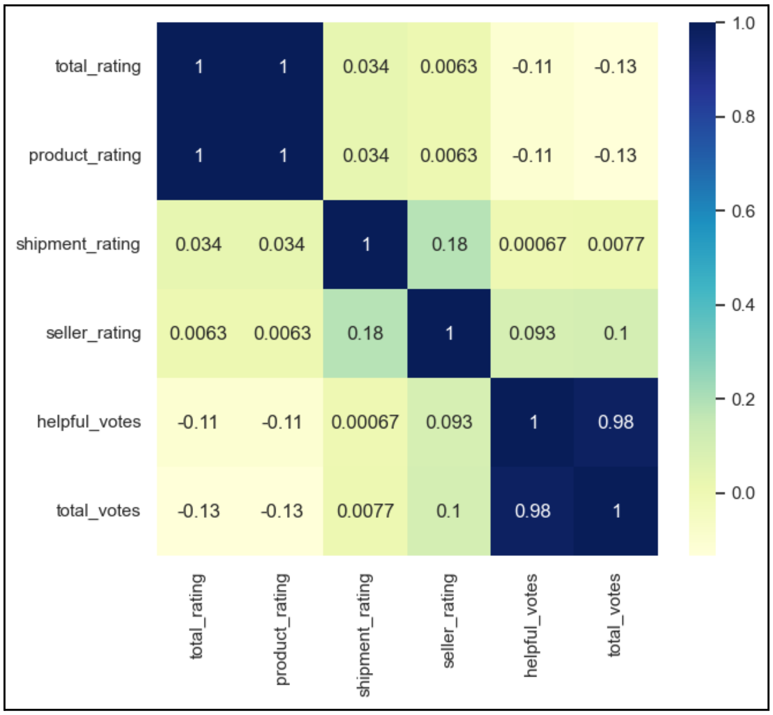


Fig 7: Correlation heatmap of Dataset features

The numbers inside each cell of the heatmap represent the correlation coefficient between the corresponding pair of features. The chart provides a visual summary of how different features in the Amazon review dataset are related to each other, and are helpful in identifying patterns or relationships in the data.

vi) Rating

The rating distribution among reviews is the next, the chart displays the number of reviews on the y-axis and the total rating (ranging from 1 to 5) on the x-axis. Each bar represents a total rating, and the height of the bar represents the number of reviews with that particular rating. The chart provides a useful summary of the distribution of total ratings in the Amazon review dataset and can be helpful in understanding how customers rate the products they have purchased.

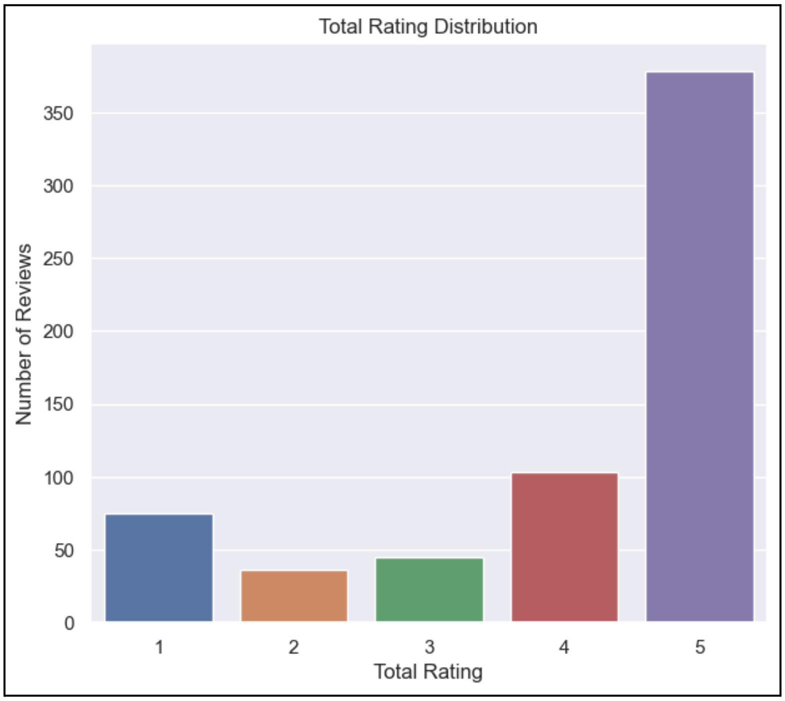


Fig 8:: Distribution of total ratings in the dataset

The chart displays the frequency of the ratings of the product in the dataset on the y-axis and the ratings on the x-axis. Each bar represents a rating and its frequency in the dataset. The height of the bar represents the frequency of that rating, while the x-axis label displays the rating itself. Overall, this chart provides a visual representation of the most ratings in the Amazon review dataset and how often they appear in the reviews.

vii) The current shipment rating distribution that to be address on the next stages of the project is the next,

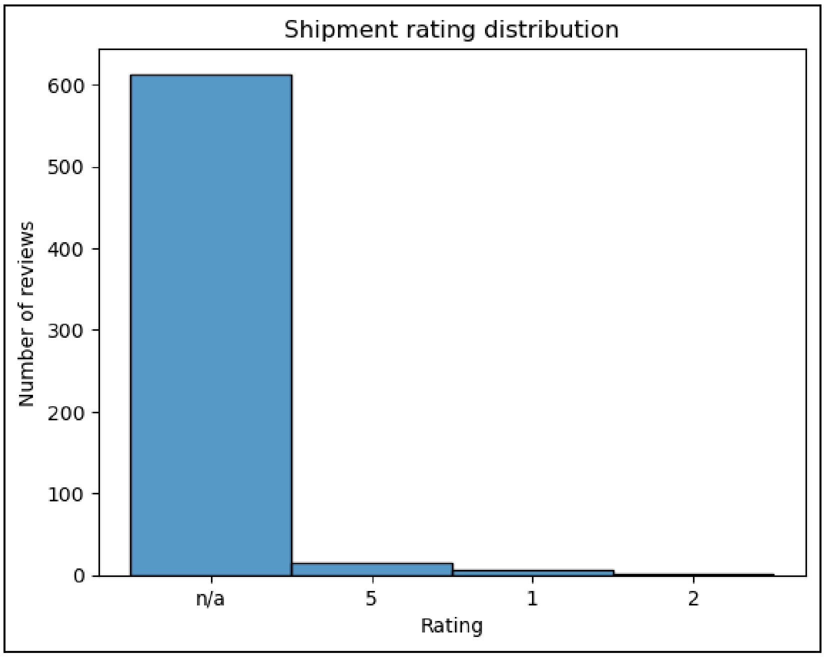


Fig 9: Shipment rating histogram

The chart displays the frequency of the **shipment ratings** of the product in the dataset on the y-axis and the ratings on the x-axis. Each bar represents a rating and its frequency in the dataset. The height of the bar represents the frequency of that rating, while the x-axis label displays the rating itself.

viii) The current seller rating distribution that to be address on the next stages of the project is the next,

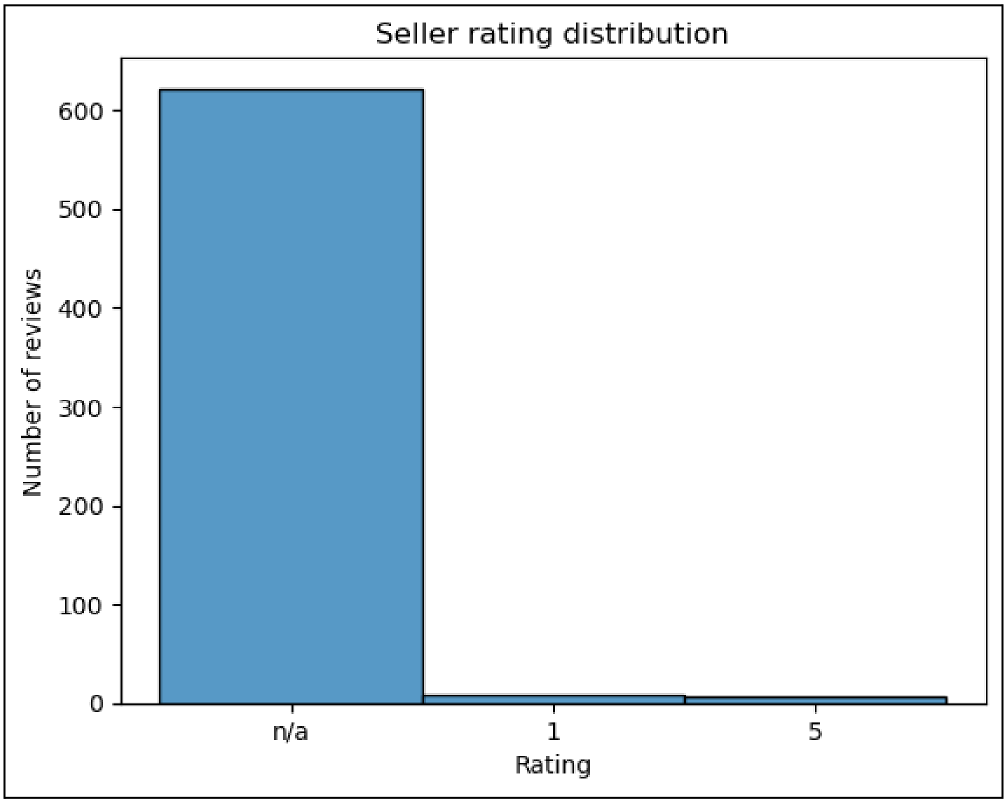


Fig 10: Seller rating histogram

The chart displays the frequency of the seller ratings of the product in the dataset on the y-axis and the ratings on the x-axis. Each bar represents a rating and its frequency in the dataset. The height of the bar represents the frequency of that rating, while the x-axis label displays the rating itself.

ix) There is the next distribution of ratings across categories of the products,
The chart displays the total rating on the y-axis and the product categories on the x-axis. Each box represents a product category, and the horizontal line inside the box represents the median rating for that category. The top and bottom of the box represent the upper and lower quartiles, respectively, and the whiskers extend to the most extreme data points that are not considered outliers. Outliers are displayed as individual points beyond the whiskers.

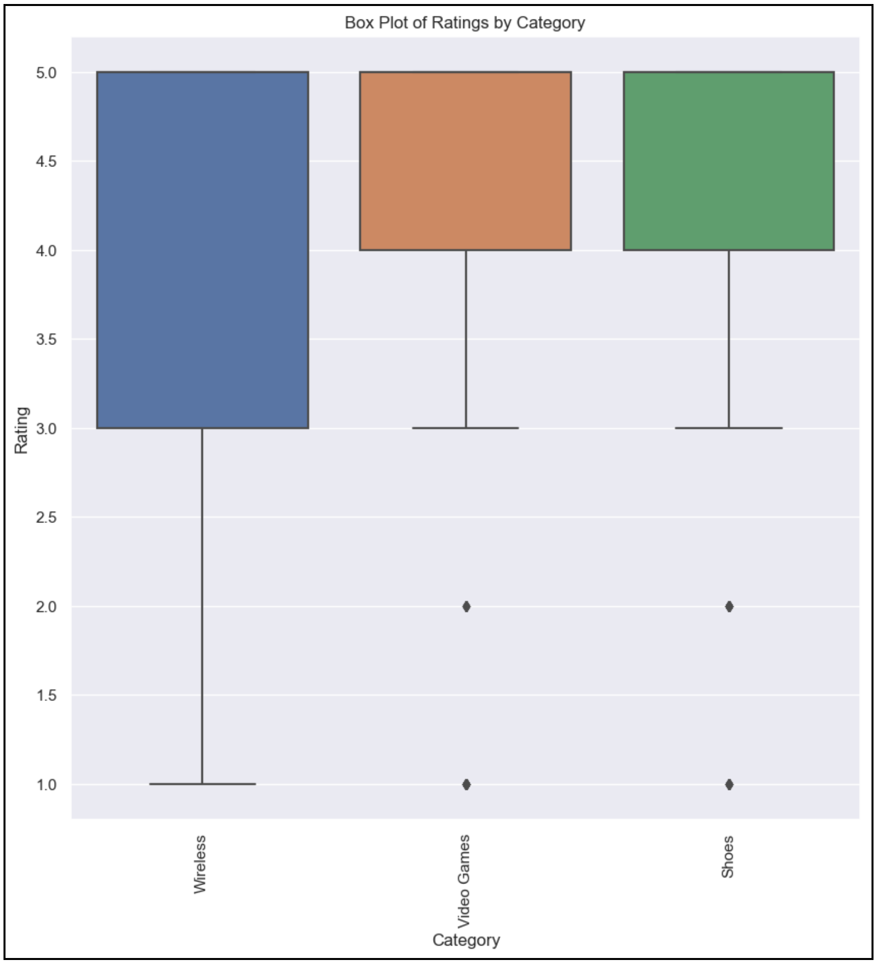


Fig 11: Box Plot of Ratings by Product Category

Overall, this chart provides a summary of how ratings are distributed across different product categories in the Amazon review dataset and can be helpful in identifying any differences in ratings across categories.

x) Sentiment

The sentiment distribution across all reviews with categorization is the next

The Sentiment Label Distribution Pie Chart shows the distribution of sentiment labels in the Amazon review dataset. The chart is divided into three sections, each representing a sentiment label: Positive, Negative, and Neutral. The size of each section is proportional to the number of reviews with that sentiment label. The chart is also color-coded to differentiate between the sentiment labels, with blue representing Positive, orange representing Negative, and green representing Neutral. The chart shows that the majority of reviews in the dataset are Positive, followed by Negative and then Neutral.

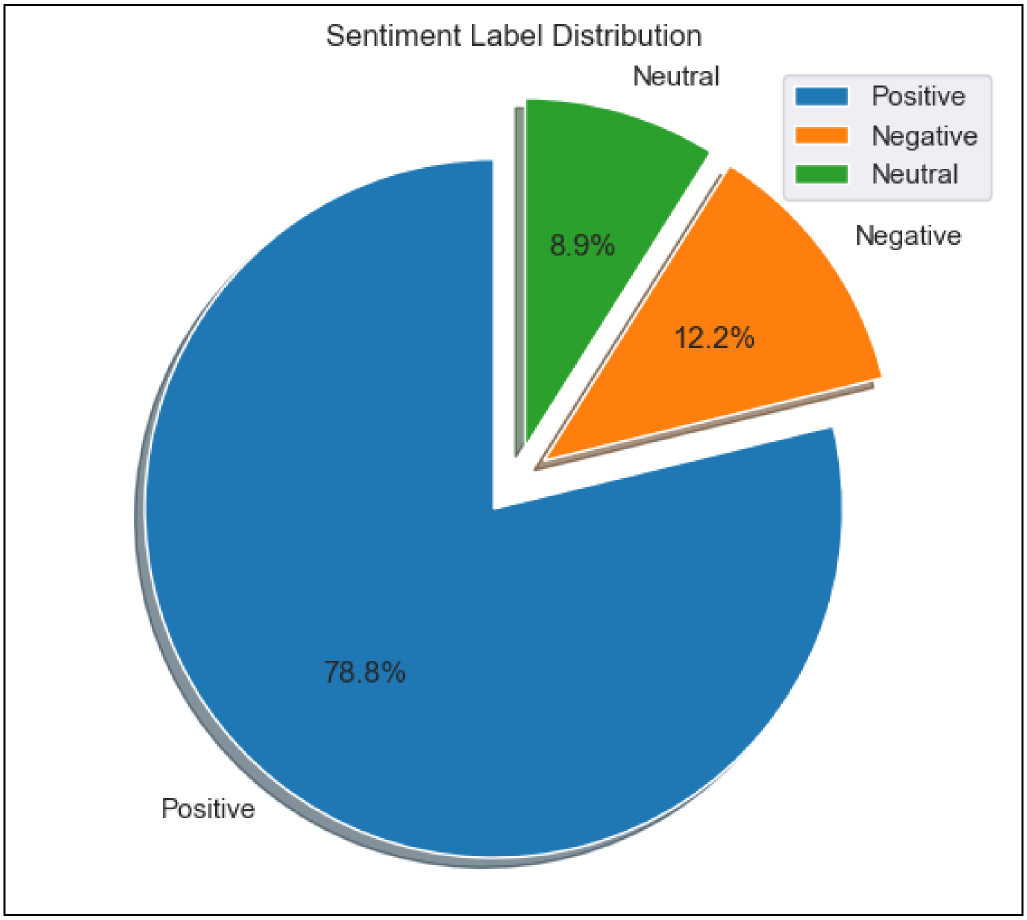
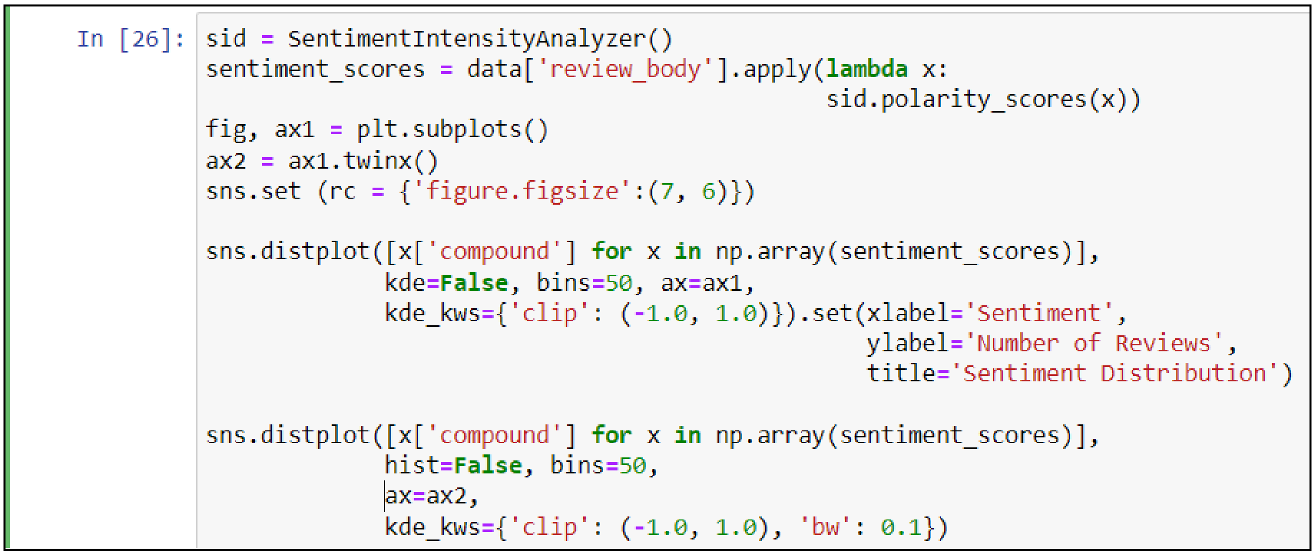


Fig 12: Sentiment Label Distribution Pie Chart

xi) The sentiment distribution across all reviews after normalization is the next,
The sentiment score ranges from -1 to 1, where negative scores indicate negative sentiment, positive scores indicate positive sentiment, and neutral scores indicate neutral sentiment.



```
In [26]: sid = SentimentIntensityAnalyzer()
         sentiment_scores = data['review_body'].apply(lambda x:
                                                      sid.polarity_scores(x))
         fig, ax1 = plt.subplots()
         ax2 = ax1.twinx()
         sns.set (rc = {'figure.figsize':(7, 6)})

         sns.distplot([x['compound'] for x in np.array(sentiment_scores)],
                      kde=False, bins=50, ax=ax1,
                      kde_kws={'clip': (-1.0, 1.0)}).set(xlabel='Sentiment',
                                                         ylabel='Number of Reviews',
                                                         title='Sentiment Distribution')

         sns.distplot([x['compound'] for x in np.array(sentiment_scores)],
                      hist=False, bins=50,
                      ax=ax2,
                      kde_kws={'clip': (-1.0, 1.0), 'bw': 0.1})
```

Fig 13: Code snippet for Sentiment Distribution and Density Plot

The plot consists of two subplots: the first shows a histogram of sentiment scores, indicating the number of reviews that fall into different sentiment score ranges. The second subplot shows the density plot of the sentiment scores, which represents the probability distribution of sentiment scores in the dataset.
The x-axis of both subplots represents the sentiment score, and the y-axis of the first subplot represents the number of reviews with the corresponding sentiment score. The y-axis of the second subplot represents the probability density of the sentiment score.

The plot provides insight into the sentiment distribution of the Amazon product reviews dataset, showing that the majority of reviews have a positive sentiment score, with a spike of neutral and negative sentiment reviews. The density plot also indicates that the distribution is largely skewed towards positive sentiment, with a peak around a sentiment score of 0.8.

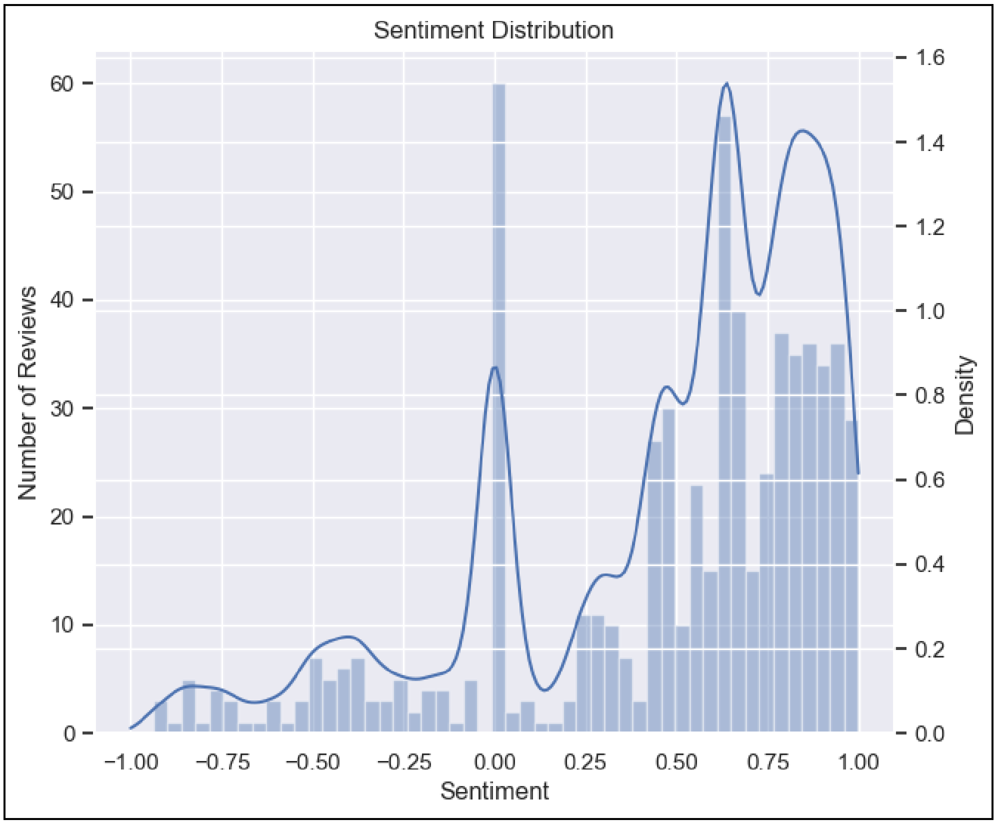


Fig 14: Sentiment Distribution and Density Plot

Key predictors

Considering the nature of the project and having features of the dataset, the most reasonable choice of predictors will include in order of importance

1. review_body - a text of a review
2. review_headline - a title of a review