Data Storytelling: Sentiment Analysis of Amazon Product Reviews

Springboard Capstone Project 1

Exploratory Data Analysis

The objective of my Capstone project 1 (sentiment Analysis of Amazon product reviews) is to determine the polarity of a review text i.e. whether a given review is positive (rating: 4 or 5) or negative(rating 1 or 2) based on ratings and reviews, which can help businesses make a decision based on customer reviews. In this exploration of the Amazon product reviews dataset, I will analyze the features of review text with respect to ratings. For this project, I will use three product categories: baby products, grocery and gourmet foods, and cellphones dataset.

These are the questions I will look in the data:

- What does the distribution of data based on review categories (or star ratings) and distribution of review length look like?
- What does the distribution of the number of ratings in different product categories look like?
- How do the reviews change across high and low star ratings based on review length and percentage of uppercase words per review for different product categories?
- What are the most frequently used words for high and low ratings?

Data Source: http://jmcauley.ucsd.edu/data/amazon/

Distribution Plots

1. Rating Distribution

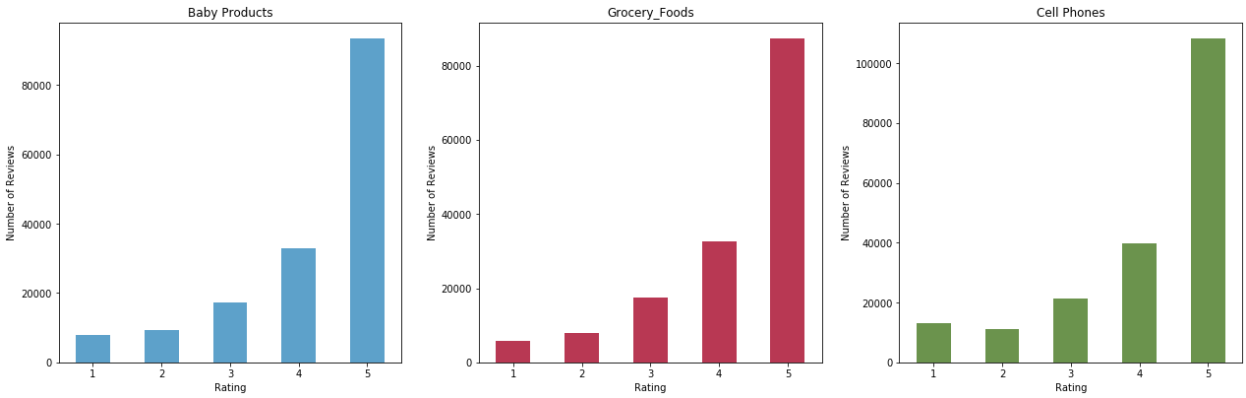


Fig 1. Distribution of rating by product categories

I have plotted the distribution of ratings for the three different product categories and noticed that most of the reviews are five stars in all three categories. I also see that cell phones tend to have more ratings of 1 than the other categories.

2. Review Length Distribution

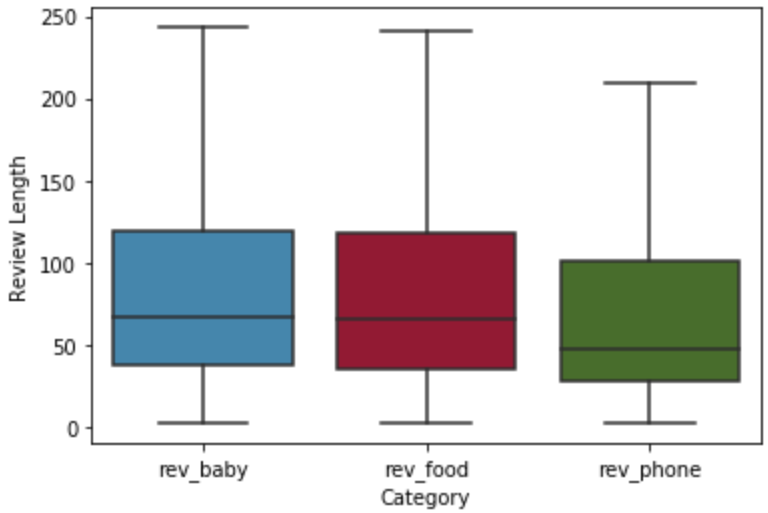


Fig 2. Distribution of review length by product categories

From the review length distribution, we can see the cell phone reviews are shorter in length compared to the other two product categories - 74.8% of cell phone reviews are of length less than equal to 100, whereas baby products and grocery food categories have about 56% and 53% of reviews of length less than or equal to 100 respectively.

3. Review length distribution by rating

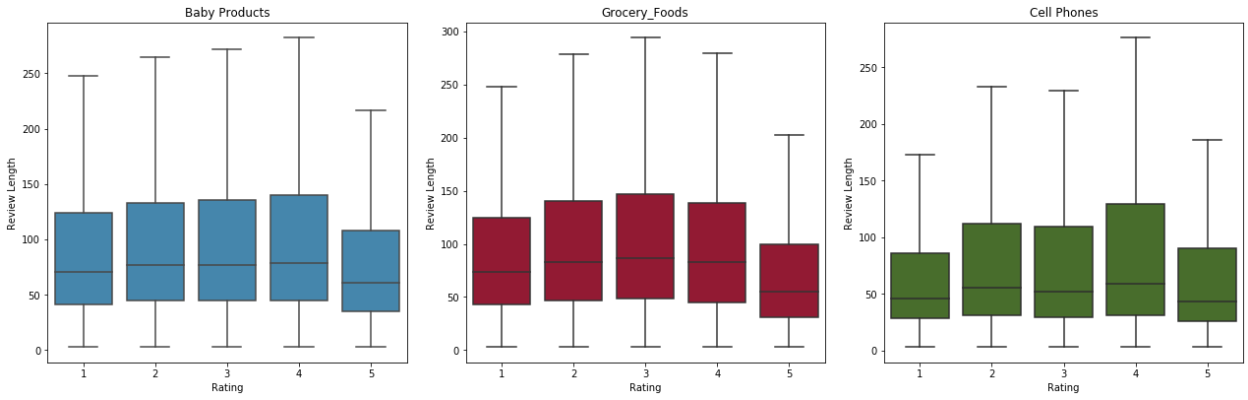


Fig 3. Distribution of Review Length by Rating

In all of the three product categories, five star rated reviews are found to be the shortest length followed by one star rated reviews. Two, three and four star rated reviews are longer and show different trends in different products categories. In the case of baby products and cell phones categories four star rated reviews are longer and in grocery_foods three star rated reviews are longer. From this exploration we can see that, when people are highly satisfied or dissatisfied with the product, they write very short reviews and, in other cases, they have to explain where and which part of the product they are unsatisfied with hence the reviews are longer. As we can see an example of a short 5 star review below:

Example of 5 star review with less than 5 words:

'Love this product'

'Great little headset.'

'Nice Case. Great Price.'

4. Distribution of % of capitals by rating

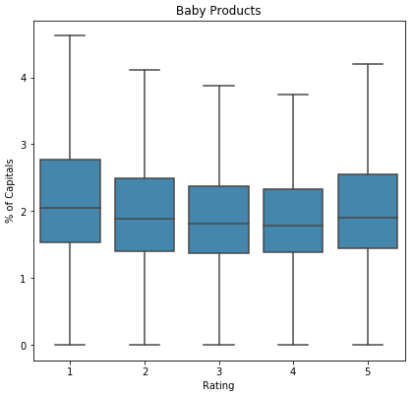


Fig 4. Distribution of % of Capitals by Rating

In all the three product categories, reviews corresponding to 1-star and 5-star ratings have more uppercase letters and three star rated reviews have least upper case letters. It might be due to highly satisfied or dissatisfied customers writing reviews of short sentences which contain more uppercase letters.

5. Distribution of % of Capitals

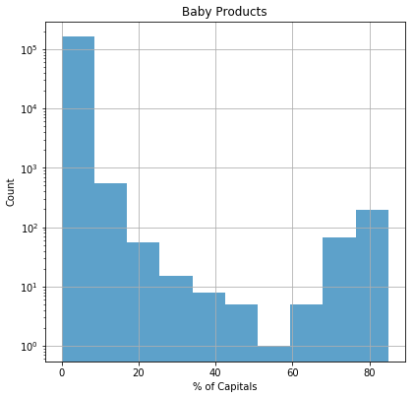


Fig 5. Distribution of % of Capitals

For all the three product categories the % of capitals show bimodal distribution one peak centered at around 20% and the other peak centered at around 70% of capitals.

6. Distribution of % caps (> 0.6) vs ratings

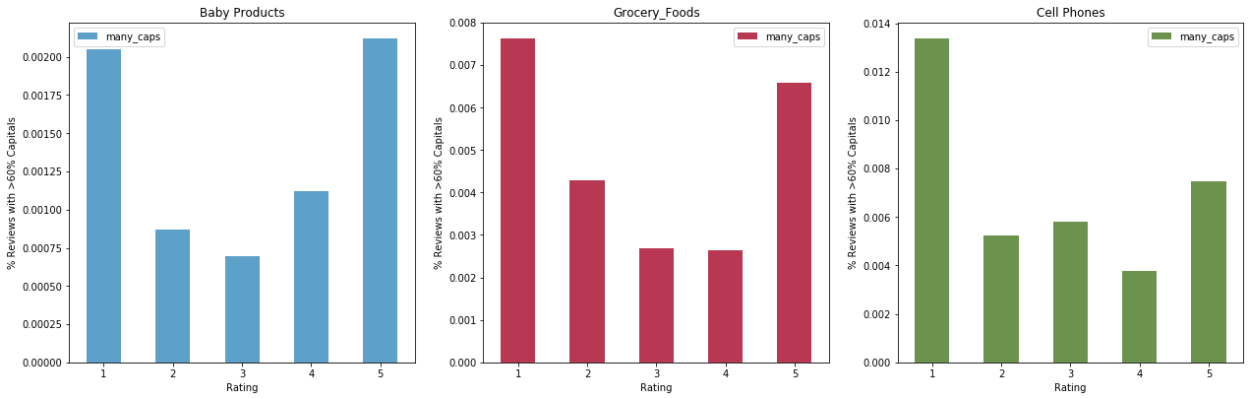


Fig.6 Distribution of % of capitals (> 60%) by rating

In fig.6 we have plotted the average number of reviews greater than 60% capital letters grouped by rating. 1 and 5 star reviews are more likely to contain reviews with >60% capital letters across product categories then other star ratings.

Example of a reviews with >= 60% capitals:

'AMAZING SOUND, MOVIE THEATER QUALITY' (5-star)

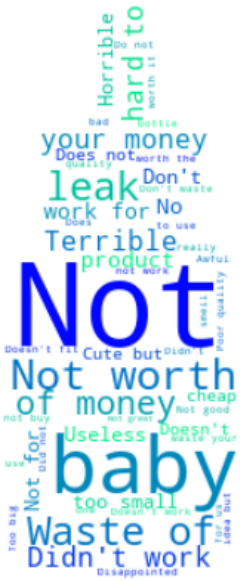
"COULDN'T DOWNLOAD ANYTHING" (1-star)

7. Word clouds from review summaries for the high and low rating

The word clouds from review summaries are plotted by high and low ratings. In this case I have taken 5 stars ratings as high and 1- 2 stars as low ratings. In the word clouds the font size of a word indicates its frequency and importance in the review. The words used more often in the reviews are greater in font size and darker in color. The size of the word decreases as the frequency of the word in the reviews decreases. As we can see from the word cloud of high rated grocery_foods summary, the words like ‘Delicious’, ‘Yummy’, ‘Testy’, ‘Excellent’, ‘the best’, ‘Love this’ looks bigger and darker as they were used more frequently used in the reviews. On the other hand in the word cloud of low rated reviews the words like ‘Not good’, ‘Terrible taste’, ‘Disappointing’, ‘Yuck’ are the most used words.

High Rating (Baby Products)

Low Rating (Baby Products)



High Rating (Grocery_food)

Low Rating (Grocery_food)

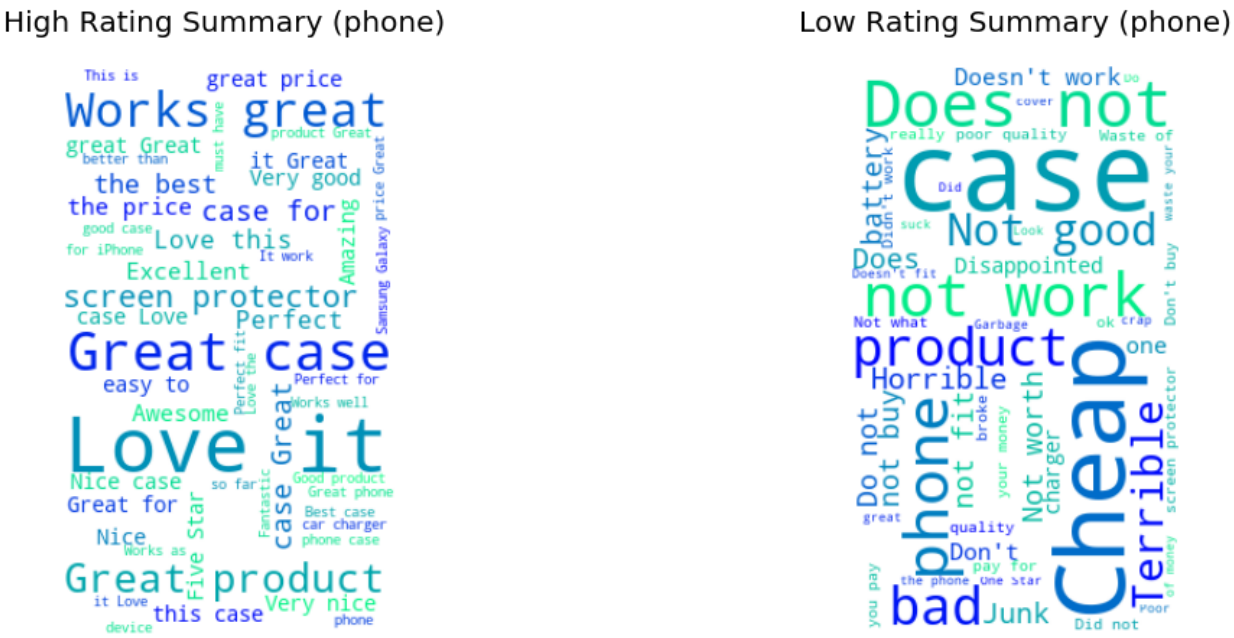


Fig 7. Word clouds from review summaries for high vs low rating