Data Science Final Research Project: Amazon Dataset

Does a higher discount percentage relate to a lower rating per product category? Consumer perception of pricing and its impact on purchasing decisions play a pivotal role in today's competitive marketplace. As businesses strive to attract customers and maximize sales, understanding the relationship between discount percentages and ratings per product category becomes crucial. In this study, the objective is to explore the relationship between mean discount percentages and mean ratings per product category. By examining the relationship between these two variables, I aim to gain insights into consumer behavior. To find the answers to my research, I downloaded an Amazon sales Dataset from Kaggle.com. It contains 1465 rows of different products, and 16 columns of different information ranging from product ID to discounted pricing. When dealing with a dataset it is important to clean and process the data before analysis. When doing so I found two unknown numbers in the rating\_count. And after finding the corresponding rows, I decided to remove them completely from the dataset. Removing rows with missing values helps maintain the overall quality and accuracy of the dataset. The next step in the research is to process the data and clean it to fit the needs when graphing and grouping the data. To do so, I split the category column into three separate columns and labeled them category\_1 category\_2 and category\_3. Making it easier to see overall trends based on categories rather than individual products. I decided to use caterory\_2 when analyzing patterns because there is a wider variety of categories, and it is more specific than category\_1. Then I also transformed the discount parentages into decimals to be easily manipulated when needed. As well as doing some smaller transformations through functions like groupby and sort\_values to prepare for graphing. One key component to data science is the visualization of the research. Putting together all the data and plotting graphs make it easier to see trends and patterns instead of comparing numbers in a table. To get a good idea of the size of the categories, I plotted the number of products in each category. The highest categories are electronics, computers & accessories, and home & kitchen for the main product category column. For category two the highest are accessories & peripherals, kitchen & home appliances, and home theater, TV & video. After interpreting the two bar plots, I then wanted to see the distribution of the ratings of the products. I created a histogram of a distribution and I found that the distribution is slightly skewed to the left, indicating that a larger proportion of products have higher ratings compared to those with lower ratings. This skewness suggests that most products in the dataset tend to receive positive ratings from users, although there are still variations in the ratings provided. The next step of my research was to apply these findings to the mean discount percentage. I created two plots to compare, the mean discount percentage per category and the mean rating per category. These two graphs give insight to my hypothesis. In the graph of mean discount percentages per Category, overall, it is clear to see that there is a wide variety of discount percentages for the different categories of products. The lowest mean discount of zero percent is for the arts and crafts category, and the highest mean discount, which is seventy percent, is in the wearable technology. There is no clear pattern or recognizable reason for the varying percentages. But when analyzing the mean ratings per category graph, it is easy to see that all the categories are around the mean rating of 4.0. The highest rated category of products is tablets and the lowest is the car accessories category. When comparing these two graphs to see if a higher discount percentage relates to a lower rating per product category, the graphs appear to be slightly related or dependent on each other, as there is a small observable link between them. For example, the wearable technology category has the highest discount percentage over all the categories, and it has one of the lowest ratings. Then for the arts and crafts category with the lowest discount percentage, it also has one of the highest ratings. These observations prove that there is a slightly visible relationship between the discount percentage of the product category and the mean ratings. If the discount percentage is higher than it can be expected that the rating will be lower. Then also, if the discount percentage is lower, we can predict that the rating will be higher. We can make some assumptions about consumers and business tactics as to why these patterns are visible but, further research and analysis is recommended to explore additional variables and potential factors that might contribute to the observed relationship.