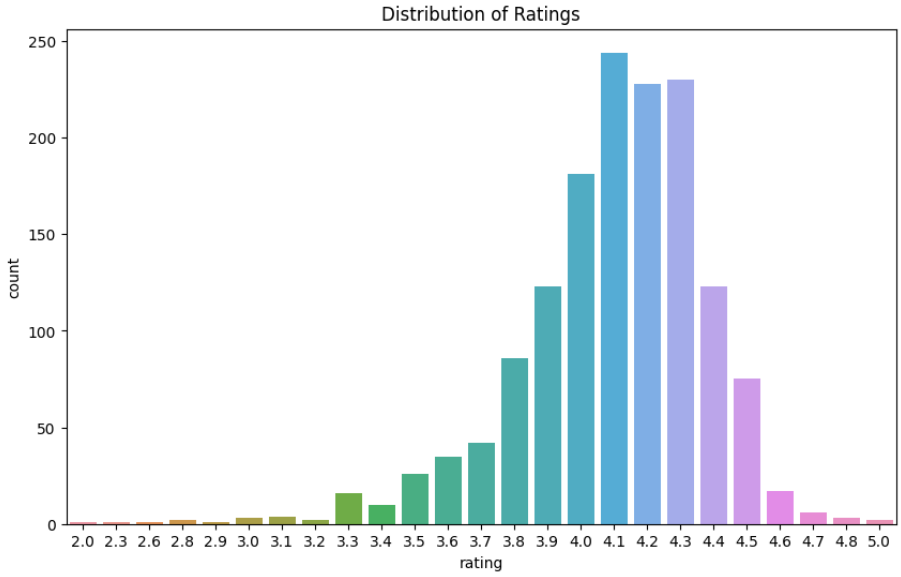
Amazon Product Reviews Analysis

In this project, we conducted an exploratory data analysis and predictive modeling on a dataset of Amazon product reviews.

1. Data Cleaning and Exploration

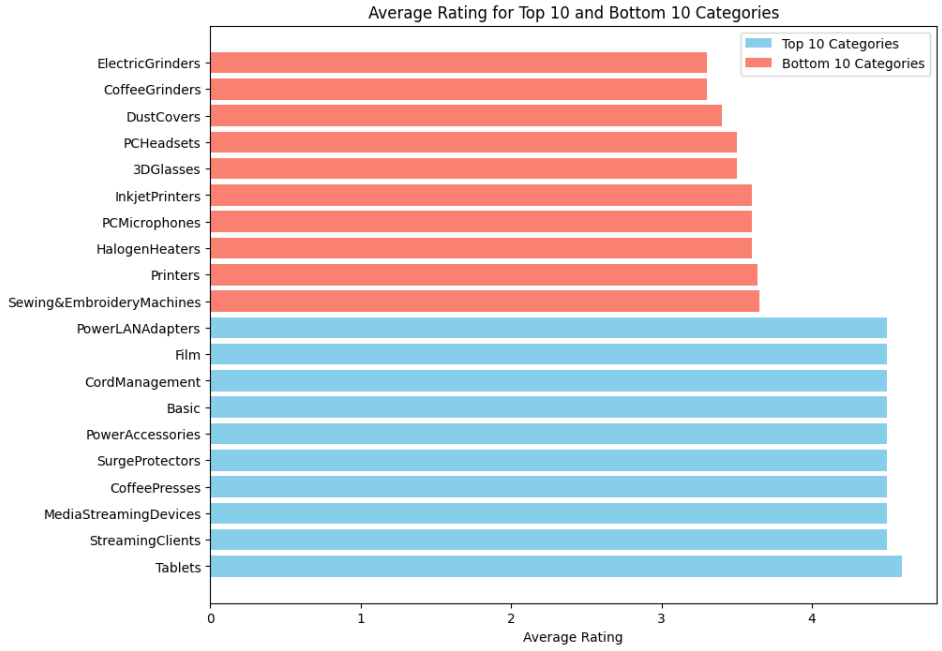
The dataset consisted of several columns such as product_id, product_name, category, discounted_price, actual_price, discount_percentage, rating, rating_count, about_product, user_id, user_name, review_id, review_title, review_content, img_link, and product_link. We started by cleaning the data, particularly the actual_price, discounted_price, discount_percentage, and rating columns, which involved removing special characters and converting the data types.

We then explored the dataset and found that the majority of the reviews were positive, with most ratings being around 4. We also found that the length of the review content varied significantly, with some reviews being much longer than others.



2. Category vs. Rating

We further analyzed the relationship between product categories and ratings. We found that certain categories tended to have higher average ratings than others. A bar graph was used to visualize this, showing the average rating for each category.



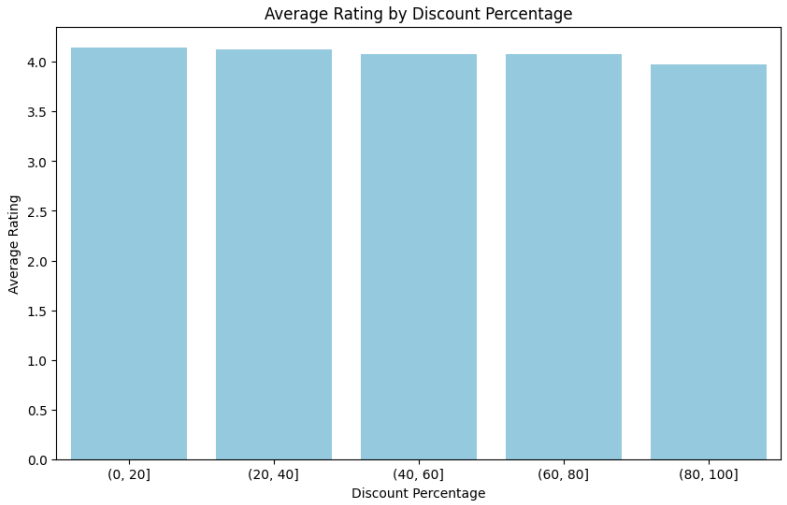
3. Topic Modelling

We performed topic modelling on the review content to identify common themes in the reviews. We found that topics related to "sound", "battery", "cable", and "charging" were frequently discussed.

```
Topic: 0
Words: 0.015*"remot" + 0.013*"easi" + 0.009*"sound" + 0.009*"amazon" + 0.009*"batteri" + 0.009*"need" + 0.008*"time" + 0.008*"look" + 0.008*"great" + 0.007*"pictur"
Topic: 1
Words: 0.025*"nice" + 0.022*"water" + 0.011*"heat" + 0.011*"easi" + 0.011*"sound" + 0.011*"amazon" + 0.010*"time" + 0.010*"imag" + 0.010*"look" + 0.008*"instal"
Topic: 2
Words: 0.041*"cabl" + 0.023*"charg" + 0.012*"sound" + 0.011*"nice" + 0.009*"month" + 0.009*"origin" + 0.009*"better" + 0.009*"look" + 0.008*"time" + 0.008*"wire"
Topic: 3
Words: 0.023*"phone" + 0.018*"batteri" + 0.013*"camera" + 0.009*"time" + 0.008*"nice" + 0.008*"featur" + 0.008*"charg" + 0.008*"screen" + 0.007*"heat" + 0.007*"look"
Topic: 4
Words: 0.025*"watch" + 0.014*"easi" + 0.012*"time" + 0.010*"clean" + 0.010*"featur" + 0.009*"look" + 0.008*"batteri" + 0.007*"great" + 0.007*"power" + 0.006*"come"
Topic: 5
Words: 0.013*"mous" + 0.013*"sound" + 0.012*"connect" + 0.009*"watch" + 0.008*"feel" + 0.008*"time" + 0.008*"look" + 0.008*"batteri" + 0.008*"earphon" + 0.007*"build"
Topic: 6
Words: 0.038*"charg" + 0.027*"cabl" + 0.017*"phone" + 0.016*"fast" + 0.012*"devic" + 0.012*"connect" + 0.011*"issu" + 0.011*"power" + 0.009*"speed" + 0.009*"time"
Topic: 7
Words: 0.027*"cabl" + 0.017*"phone" + 0.015*"charg" + 0.013*"camera" + 0.012*"look" + 0.009*"easi" + 0.009*"laptop" + 0.008*"instal" + 0.008*"nice" + 0.008*"mobil"
Topic: 8
Words: 0.024*"charg" + 0.017*"cabl" + 0.015*"sound" + 0.015*"money" + 0.013*"valu" + 0.013*"phone" + 0.010*"best" + 0.010*"amazon" + 0.009*"fast" + 0.008*"nice"
Topic: 9
Words: 0.015*"coffe" + 0.013*"easi" + 0.008*"nois" + 0.008*"speed" + 0.008*"time" + 0.007*"issu" + 0.007*"clean" + 0.007*"overal" + 0.007*"come" + 0.007*"month"
```

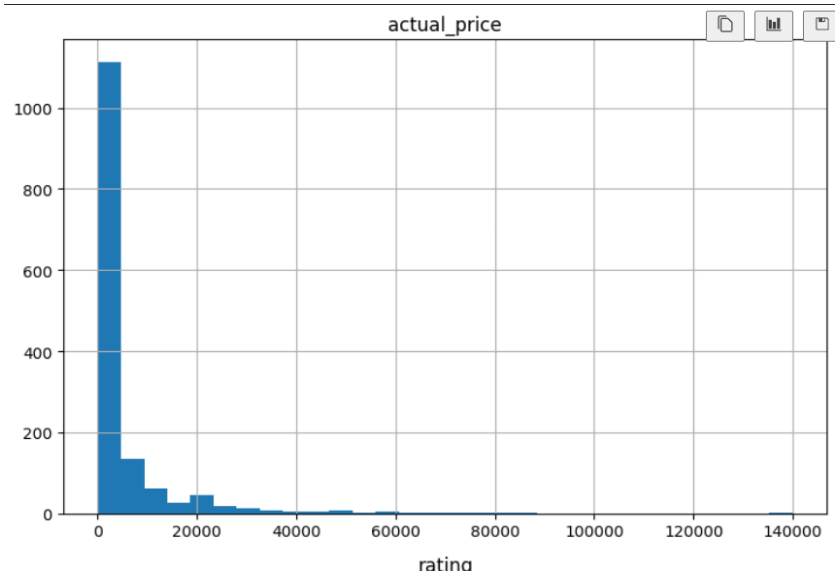
4. Discount vs. Rating

We explored the relationship between the discount percentage and the product rating. We found that there was no clear correlation between these two variables.



5. Price Distribution

We looked at the distribution of the actual prices of the products. We found that most products were priced below 10,000, with a small number of products having a much higher price.



6. Predictive Modelling

We built a predictive model to estimate the product rating based on other features in the dataset. We used a Random Forest Regressor and found that the model achieved a Mean Squared Error (MSE) of 0.069 and an R-squared (R^2) value of 0.157 on the test set.

We then attempted to improve the model by adding a new feature (discounted ratio) and using a Gradient Boosting Regressor, but found that the model's performance slightly worsened.

We also built a model to predict the actual price of the products. The model achieved a MSE of approximately 48,336,192.16 and an R^2 value of approximately 0.344.

Lastly, we built a Random Forest Classifier to predict whether a product has a high rating. The model achieved an accuracy of 0.730, a precision of 0.727, a recall of 0.620, and an F1 score of 0.669.

This project provided valuable insights into the dataset and demonstrated the application of various data analysis and machine learning techniques. However, the performance of the predictive models could be improved with additional relevant features and more sophisticated feature engineering and model tuning strategies.