Amazon Ratings Analysis

Data

The data used contains Amazon Ratings over three categories (Amazon Prime Video; Clothing, Shoes and Jewellery; Electronics) over the period May 1996 to October 2018 and containing 32M, 8.8M, 21M ratings respectively. Credit goes to Jianmo Ni, Jiacheng Li and Julian McAuley for compiling and publishing the data here.

Each data point contains the User ID, Product ID, Rating Given, Unix Timestamp. For the purpose of analysis, the data is processed into three main forms (functions for the same are defined in data_processing.py):

1. Each rating is amended to include the following information:

   i. Total Ratings Received by that product

   ii. Average Rating Achieved by that product

   iii. The rating number, i.e. chronologically, what was the rating number (e.g. 3rd Rating)

Optionally, the following information may also be amended:

   iv. Previous and Subsequent Rating Received by the product

   v. Time elapsed between previous and next ratings

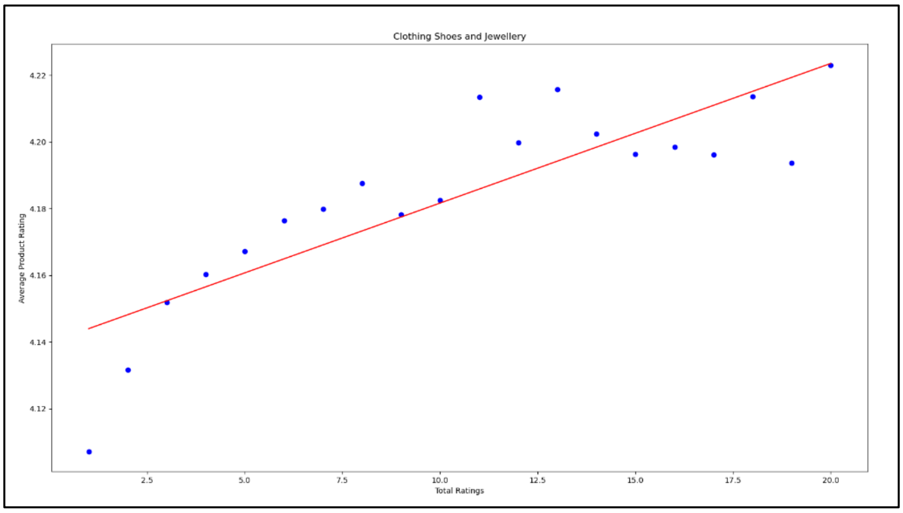
   v. Average product rating thus far

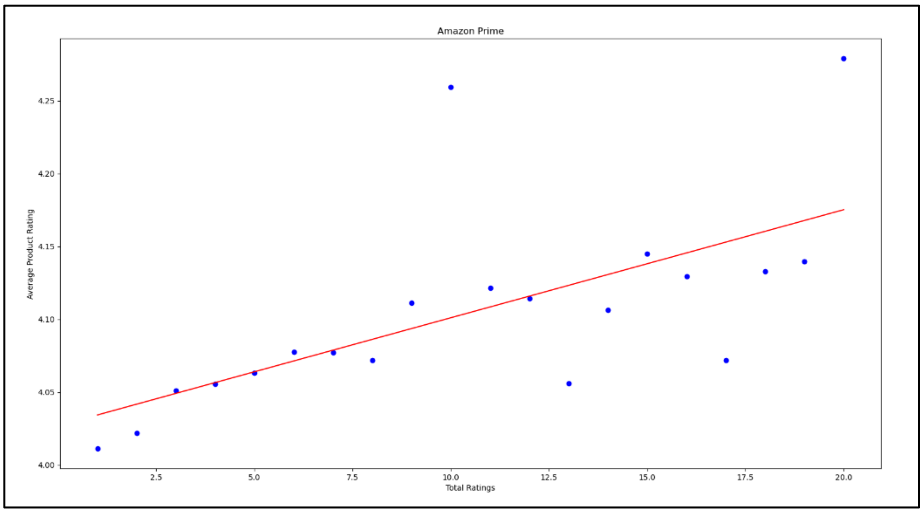
2. All products including their average rating and total ratings achieved

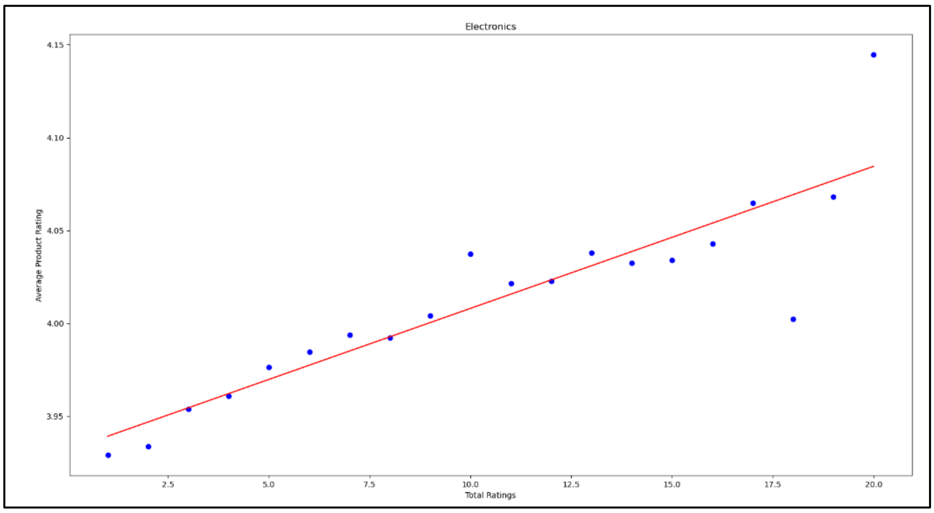
Average Rating vs. Total Ratings Analysis

The aim is to establish the impact of average product rating on the total ratings achieved by a product. Simply put, “Do people tend to buy higher rated products more?”

The following graphs plot the average product rating vs the total ratings achieved by a product for products with up to 20 total ratings across the three categories. (Code for the same is under avg_rating_total_ratings_analysis.py)







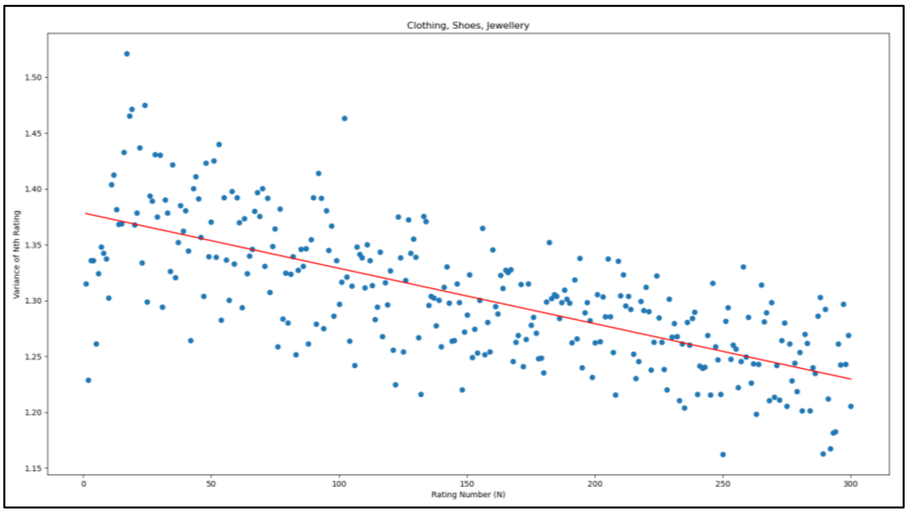
Across all three categories, we see that there is a positive correlation between the average product rating and total reviews. The above graphs have been plotted by looking at all products with N total reviews and then calculating the average of their product ratings for N between 1 and 20. The red line shows the line of best fit plotted through Ordinary Least Squares Linear Regression.

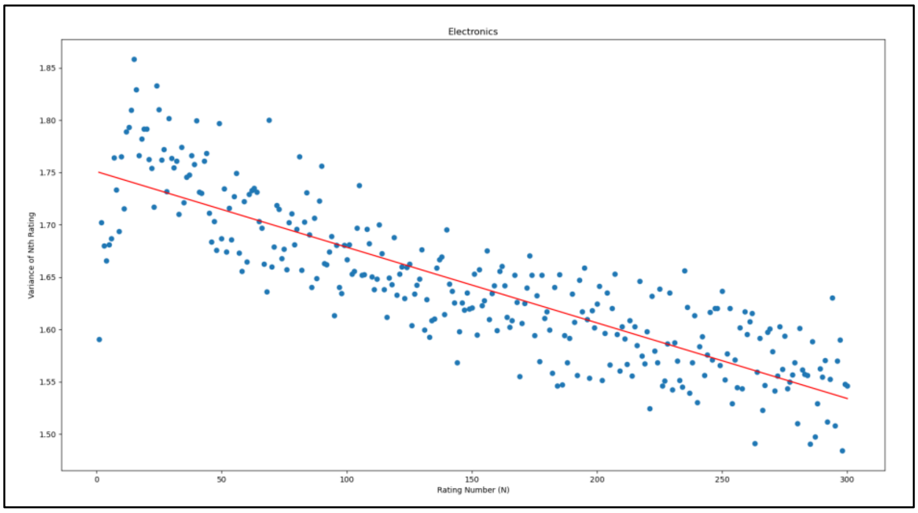
Variance Analysis across Increasing Rating Numbers

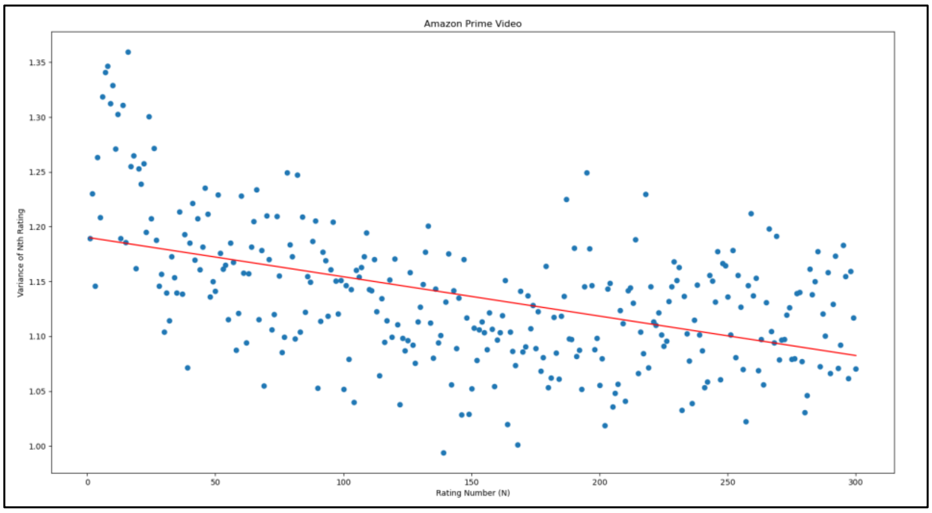
The aim of this was to answer the question “Are customers influenced by the existing product rating before giving their own?”

This analysis was carried out as follows:

All products with more than a certain number of total ratings (>300) were identified. Then the variance of the Nth rating received by the product was calculated by squaring the difference between the Nth product rating and the final average rating of the product, and then taking the mean of these across all identified products. The following graphs plot the variance of the Nth rating across all three categories.







All three graphs show reduction in variance as the rating number increases. This shows that as the average rating seen by a user becomes more “prominent” (i.e. is based on more ratings), the tendency to differ from it reduces. Thus, customers are influenced by existing ratings while giving their own. The effect of this is most pronounced in “Clothing, Shoes and Jewellery” followed by “Electronics” and lastly “Prime Video” (magnitudes of reduction in variances are 0.22, 0.20, 0.10 respectively)