Group members:

Jingyuan Liu 69763183

Jinyi Wu 28812972

Cici Liu 77155471

Zefan Liu 17237355

Online shopping has been a prevalent behavior worldwide, and how prices and sales volume of online goods are associated with consumer behaviors and profit optimization has also become a popular topic in both economics and machine learning. In particular, with traditional economic models, we may ignore some important lurking information behind the data, such as what dominates the variances between sale information of goods, how these goods are naturally clustered and how people’s rating on a good affects the discounted price. Insights into these topics may help us reveal the deeper economic pattern of selling and purchasing in online shopping.

Amazon, as one of the most world-renowned e-commerce companies, can potentially be an ideal resource of data for our analysis. We found a dataset called “Amazon Sales Dataset”, which is retrieved from Kaggle (<https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset>). It contains over 1K+ ratings and reviews for Amazon products listed on the official website of Amazon India. This real dataset is frequently updated by contributors and provides credible and valid data.

Although the price or reviews of a product may fluctuate from time to time, the data provides an insight for the market of online goods and consumers’ preferences. Moreover, it does not have any missing values. It has one categorical variable (i.e., category) and several useful numerical variables such as discounted\_price, actual\_price, discount\_percentage, rating, and rating\_count. Since our dataset is rather clean and straightforward, we plan to spend more time coding and trying to present a more comprehensive, detailed, and visible project by answering multiple subquestions that explain our main topic.

There are there

1. Use multivariable linear regression and/or other regression models to fit the discounted price based on its rank, number of reviews, original price and other features.
2. Conducting PCA to determine the structure of variance behind goods. (mainly shown by visualization)
3. Exploring the natural clusters of goods by both Kmeans and Hierarchy clustering algorithms (mainly shown by visualization)
4. Does the count of rating and the rating score of a product positively correlate to the discount rate of a product? (mainly shown by visualization)
5. We can also filter some most common categories among those products and reanalyze our previous questions from a classification perspective, i.e., what category of product can be predicted by our regression model most accurately? Does the average actual price of each category affect this relationship? (mainly analyzed by classification)
6. More to come if we have time…

Visualization:

There will be multiple parts in our project that will include visualization, such as the scree-plot for PCA, the intuitive checking of clustering results, and also the results of the regression model/classification model.

Implementation:

We aim to use our analysis to determine customer preferences in Amazon India and use it as an effective tool for price setting. We can also help customers to determine whether this is the best price to buy this product and if it will be discounted in the future, etc.

Conclusion:

The Amazon Sales Dataset provides a wealth of data that can be used to predict product prices, customer preferences, and ranks and sales. Our conclusion will be based on what dominates the variances between sale information of goods, how these goods are naturally clustered and how people’s rating on a good affects the discounted price.

The distribution of work of each group member:

Each group member will be mainly in charge of 1-2 sections of the analysis:

| Introduction | Jinyi Wu |
| --- | --- |
| Data Description/EDA? | Cici Liu |
| Section 1 | Zefan Liu |
| Section 2 | Zefan Liu |
| Section 3 | Jingyuan Liu |
| Section 4 | Jingyuan Liu Jinyi Wu |
| Section 5 | Cici Liu |
| Discussion | Jinyi Wu |
| Combine section;  Ensure the flow is smooth | Zefan Liu, Jinyi Wu, Jingyuan Liu, Cici Liu |

We will reach out to other teammates if anyone encounters issues with their sections. Everyone will edit and proofread.