MBA 720C – Introduction to Python for Business Analytics

Milestone 3 – Kung Fu Pandas

Kwabena Frimpong, Isabel Karst, Paul Nolan, William Flaugher, Hao Sun

Professor Sean Sanders

UNC KENAN-FLAGLER BUSINESS SCHOOL

Agenda

UNC KENAN-FLAGLER BUSINESS SCHOOL

- Topic Overview
- Project Goal
- Project Hypothesis
- Measurement of Success
- Data Cleaning and Exploration
- Plot and Analysis
- Findings and Takeaways

- **Amazon India | 2023**
  - **Product performance**
  - **Pricing strategy**
  - **Consumer behavior, demand, and quality perceptions.**
  - **Review and Rating Overview**

*Link to Dataset: https://www.kaggle.com/datasets/lokeshparab/amazon-products-dataset

Project Goals

UNC KENAN-FLAGLER BUSINESS SCHOOL

- **Our goal is to analyze Amazon India's product data to uncover meaningful patterns in pricing, discounts, customer ratings, and review patterns.**

- **This analysis will help identify which product categories are gaining traction and provide insight into Amazon's positioning in the competitive e-commerce space.**

- **Specifically, we aim to address three key topics:**

  1. **Impact of Discounts –** Analyze the relationship between discount price, discount percentages, and the number of reviews.
  2. **Impact of Prices –** Explore the relationship between purchase price and the quantity/value of reviews per product.
  3. **Impact of Category** – Explore the relationship between product categories/subcategories and the number of ratings per item to uncover patterns in customer interest.

Project Hypotheses

UNC KENAN-FLAGLER BUSINESS SCHOOL

- **Hypothesis #1: Larger discounts will lead to more product reviews, due to increased customer interest and reach.**
- **Hypothesis #2: Higher priced goods will receive better ratings due a higher perception of quality.**
- **Hypothesis #3: Categories with more personal or emotional significance (i.e., clothing, child goods) will have a larger average number of reviews per item.**

What Does Success Look Like, by Goal?

UNC KENAN-FLAGLER BUSINESS SCHOOL

Success:

1. Impact of Discounts:

- Success: Positive correlation between discount percentage and number of ratings.

2. Impact of Price:

- Success: Positive correlation between price and average rating.

3. Impact of Category:

- Success: Correlation between product type and number of ratings.

Methodology:

- **Bar Charts** or **Histograms** to visualize the distributions and compare trends.
- **Scatter Plots** to identify relationships or correlations between variables.
- **Descriptive Statistics** to understand the data, summarizing key metrics like mean, variance, and standard deviations.

Data Overview

UNC KENAN-FLAGLER BUSINESS SCHOOL

Data Details:

- **Number of Records: 334,964**
- **Price Range: ₹10 - ₹99,999 (Rupee)**
- **Key Stats:**
  - Category Count – 20
  - Subcategory Count – 112
  - Columns - 10

Data Cleaning and Exploration

UNC KENAN-FLAGLER BUSINESS SCHOOL

Cleaning:

- After importing the data into python, we identified and removed unnecessary columns with URL Links and Image links.
- We checked for any missing values that would need to be removed (dropna()), and found that the dataset was already complete.
- All of the columns had the “object” data type and many had special characters
  - We changed the “object” type to “float” with astype(float())
  - We removed the special characters with str.replace()

Exploration:

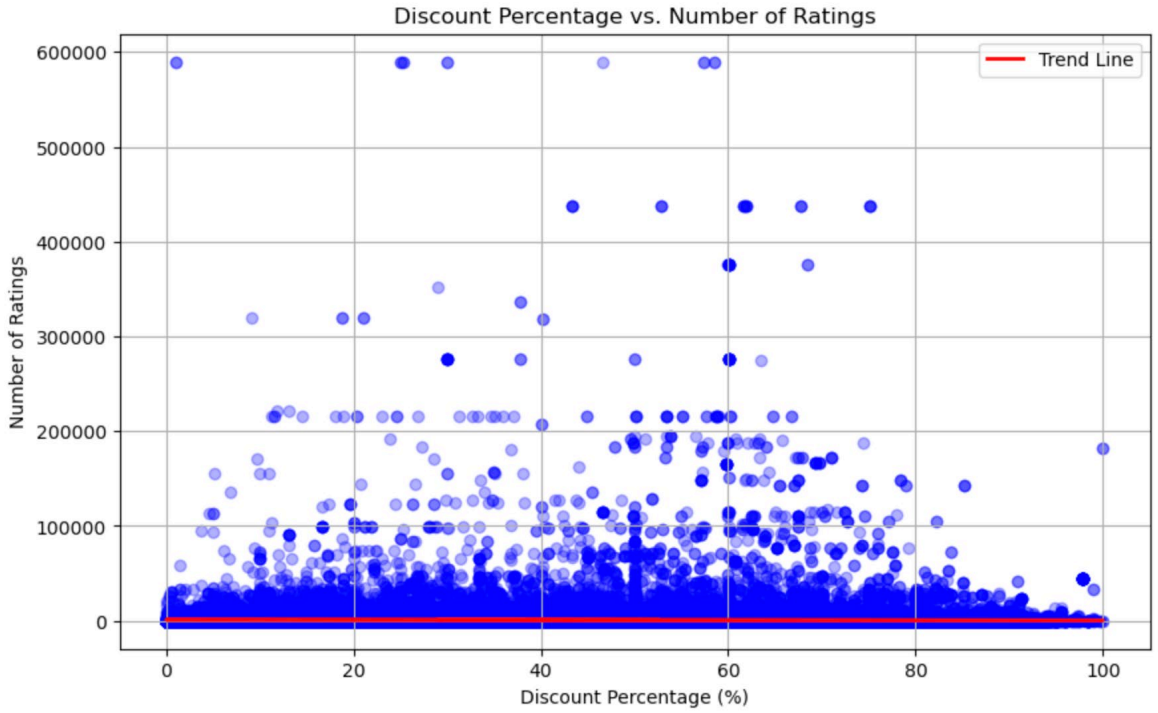
- We explored the dataset to understand the scope of our data with the following operations:
  - df.columns()
  - df.head()
  - df.count()
  - df.describe()
  - df.dtypes()

Hypothesis 1

Larger discounts will lead to more product reviews, due to increased customer interest and reach

UNC KENAN-FLAGLER BUSINESS SCHOOL

Methodology: A scatterplot was created to visualize the relationship between discount percentage and number of ratings.



Initial Observations:

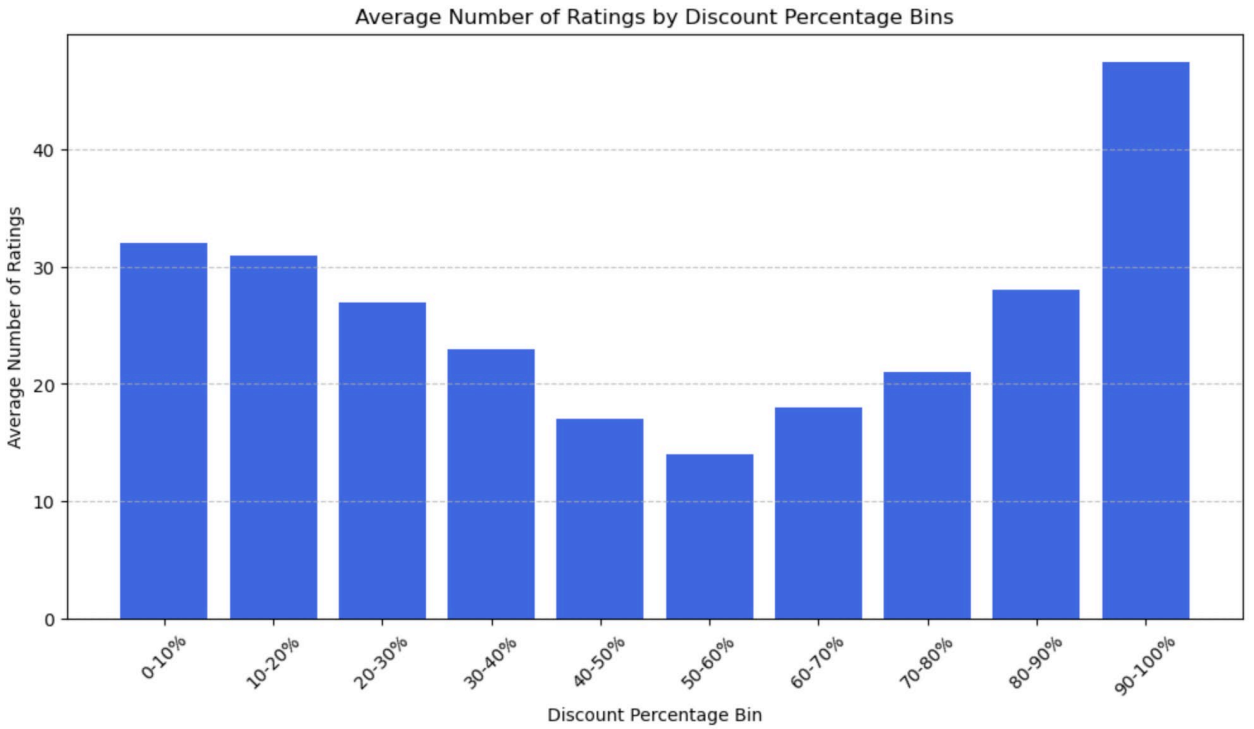
- **Discount Percentage vs. Ratings**
  - No clear indication to show relationship
  - Majority of products received <100K ratings
  - Few products with high ratings >300K ratings, which could point to other factors (i.e., brand, quality, marketing

Hypothesis 1

Larger discounts will lead to more product reviews, due to increased customer interest and reach

UNC KENAN-FLAGLER BUSINESS SCHOOL

Methodology: A bar graph showing discount percentage bins and average number of rating.



Initial Observations:

- **Discount Percentage vs. Ratings**
  - Lower end & Higher ends of range had a large average number of ratings
  - At 50-60% discount, the trend starts to increase
  - Extreme discounts tended to fare better

Hypothesis 1

Larger discounts will lead to more product reviews, due to increased customer interest and reach

UNC KENAN-FLAGLER BUSINESS SCHOOL

Conclusions:

Low R-score (0.0002) and Correlation (-0.01) show that there was no direct relationship between the level of discount and more product reviews.

Sellers should be tactical in aligning their pricing strategies. The data showed that slight or heavy discounts drove more engagement.

There were outliers which received many reviews (>300K). Discounting may be one of the factors, but there are likely a few other contributing factors.

R-Score	Correlation
0.002	-0.01

Hypothesis 2

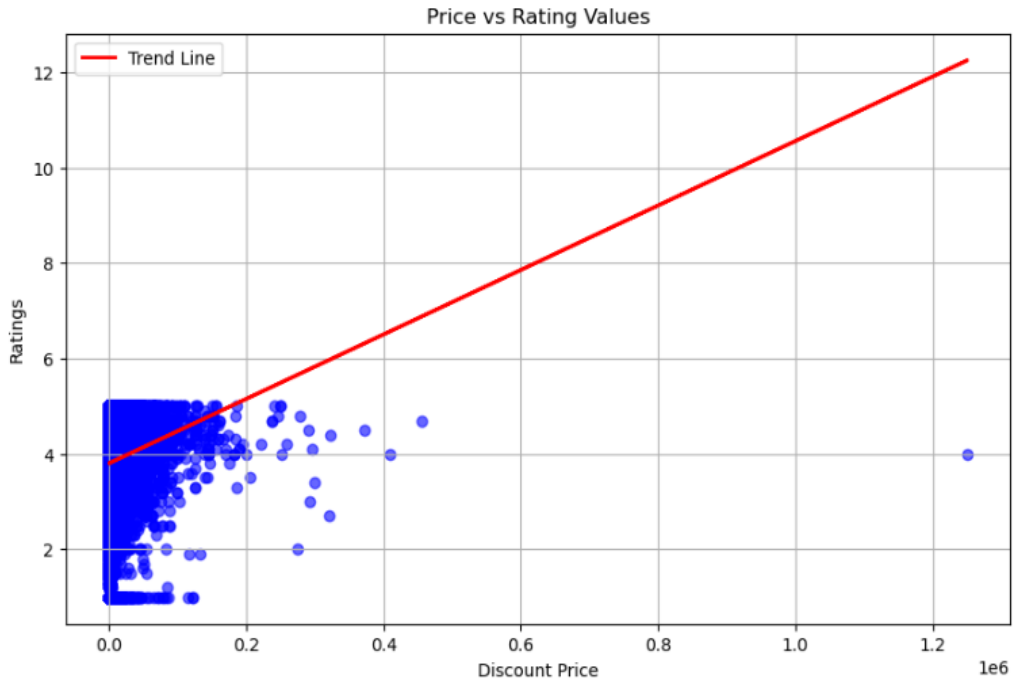
Higher priced goods will receive better ratings due a higher perception of quality

UNC KENAN-FLAGLER BUSINESS SCHOOL

Methodology: Scatterplot created to visualize the relationship between discount price and rating levels.

Initial Observations:

- **Ratings Distribution by Price:**
  - Lower-priced products show a wide range of ratings.
  - Higher-priced products are more concentrated around 4-5 stars, suggesting a more positive perception.



Hypothesis 2 – Methodology

Higher priced goods will receive better ratings due a higher perception of quality

UNC KENAN-FLAGLER BUSINESS SCHOOL

Next Steps:

- **Removal of extreme outliers**
- Focus on the **typical range** of prices and ratings
- Avoid distortion

How we identified and removed outliers:

- Lower Bound: set at 5th percentile of discount column
- Upper Bound: set at 95th percentile of the discount column

Hypothesis 2

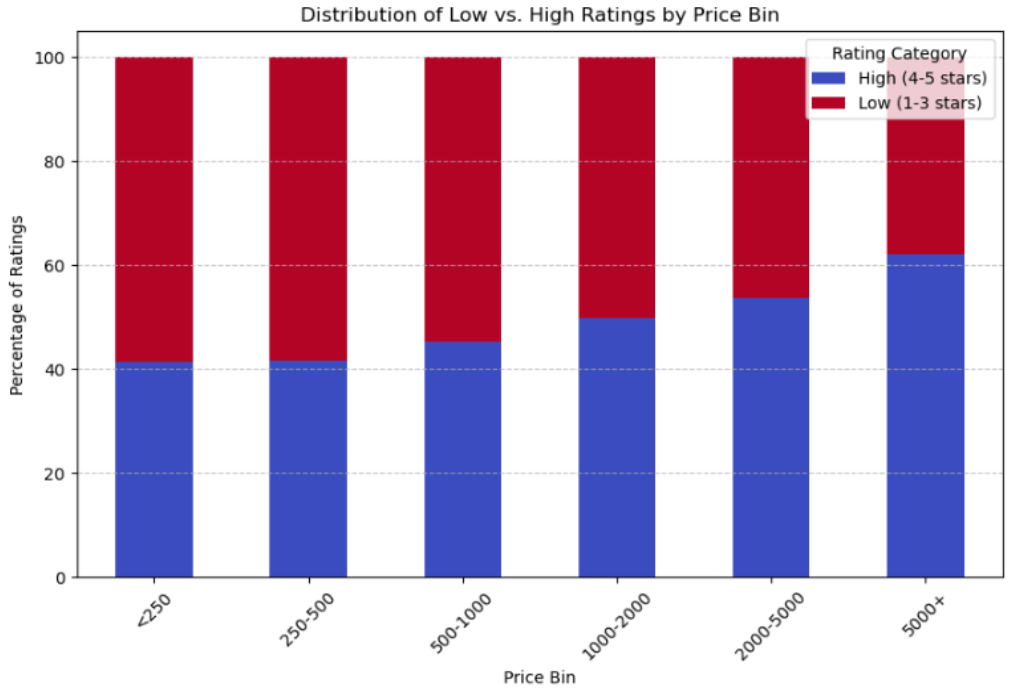
Higher priced goods will receive better ratings due a higher perception of quality

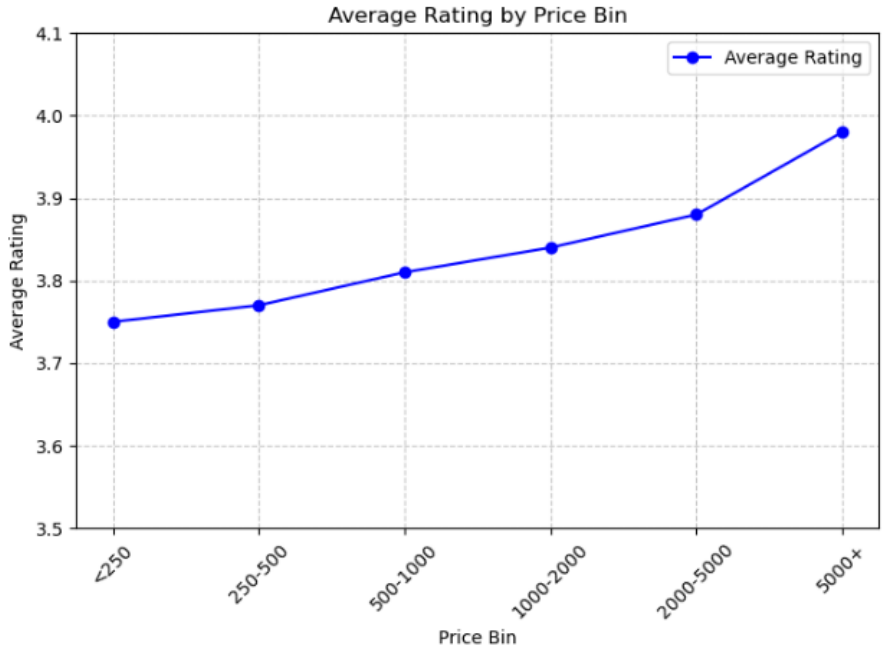
UNC KENAN-FLAGLER BUSINESS SCHOOL

Further analyzing by price bins:

Binning isolates patterns within different price ranges & simplifies data. Grouping by price makes it easier to see patterns

We divided the products into price ranges as follows:





As price increases, the portion of **high ratings (4-5 stars)** also increases.

Hypothesis 2

Higher priced goods will receive better ratings due a higher perception of quality

UNC KENAN-FLAGLER BUSINESS SCHOOL

Results - Price bins show a clear trend:

- **Lower-priced products**: More variation in ratings, with a mix of low and high ratings.
- **Higher-priced products**: Consistently higher ratings, with fewer low ratings.

Conclusion:

- Higher-priced goods tend to receive higher ratings.

amazon

Hypothesis 3

More emotionally-significant categories will get more reviews.

UNC KENAN-FLAGLER BUSINESS SCHOOL

- **Hypothesis:**
  - Categories with more personal or emotional significance will have more reviews per item.
- **Methodology v1:**
  - New column with binary labeling of significance for each product category. (e.g., Appliances = “no”)
  - Calculate the avg. number of reviews per label.
- **Findings v1:**
  - Insignificant – 1258.4
  - Significant – 358.4

Hypothesis 3

More emotionally-significant categories will get more reviews.

UNC KENAN-FLAGLER BUSINESS SCHOOL

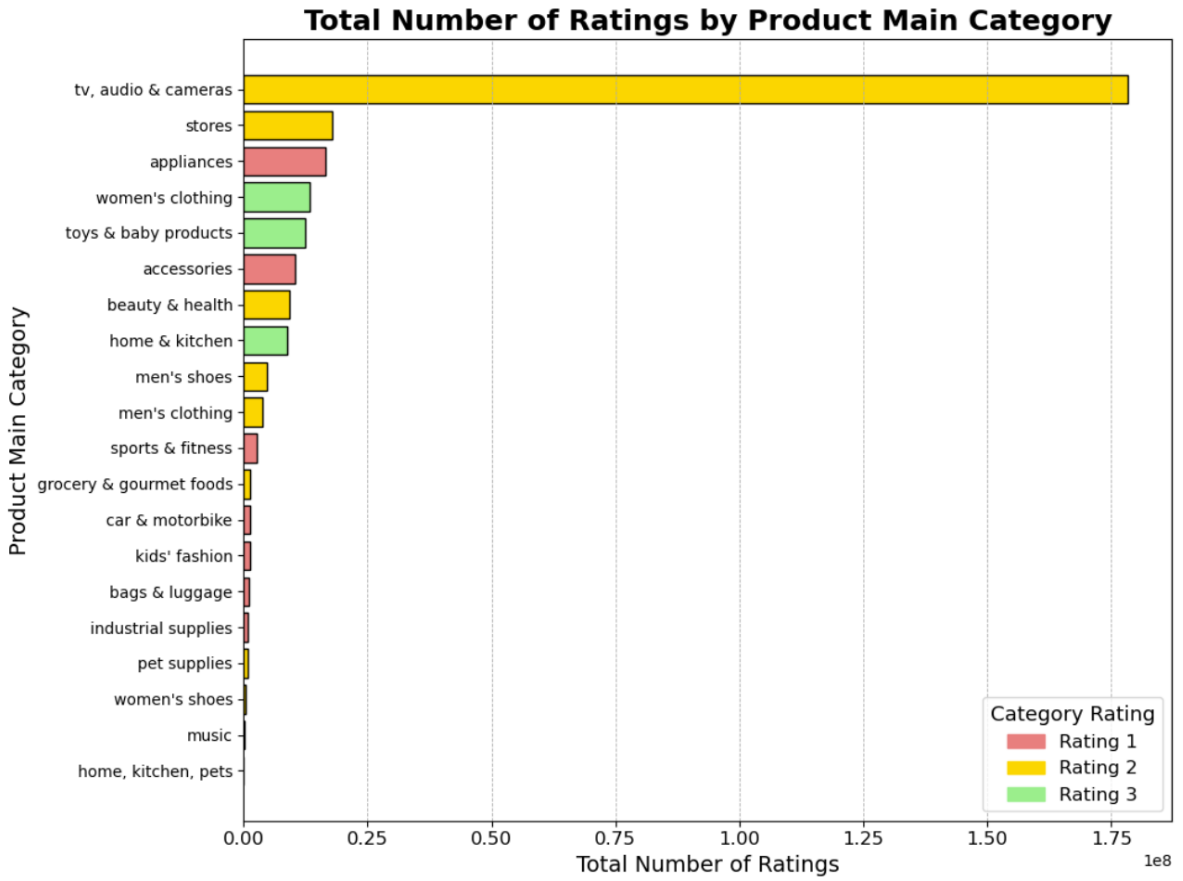
- **Evaluating Results:**
  - To ensure validity, we pulled the number of reviews per category.
  - “tv, audio & cameras” had 1,250% more reviews than the average category.
- **Methodology v2:**
  - We removed this outlier category and repeated the previous exercise.
- **Findings v2:**
  - Insignificant – 390.3
  - Significant – 358.4

Hypothesis 3

More emotionally-significant categories will get more reviews.

UNC KENAN-FLAGLER BUSINESS SCHOOL

- **Methodology v3:**
  - We used a more nuanced approach, adding a “slightly significant” option.
  - Changed our ratings from T/F to a scale of 1-3.
  - “tv, audio & cameras” is clearly still skewing the results

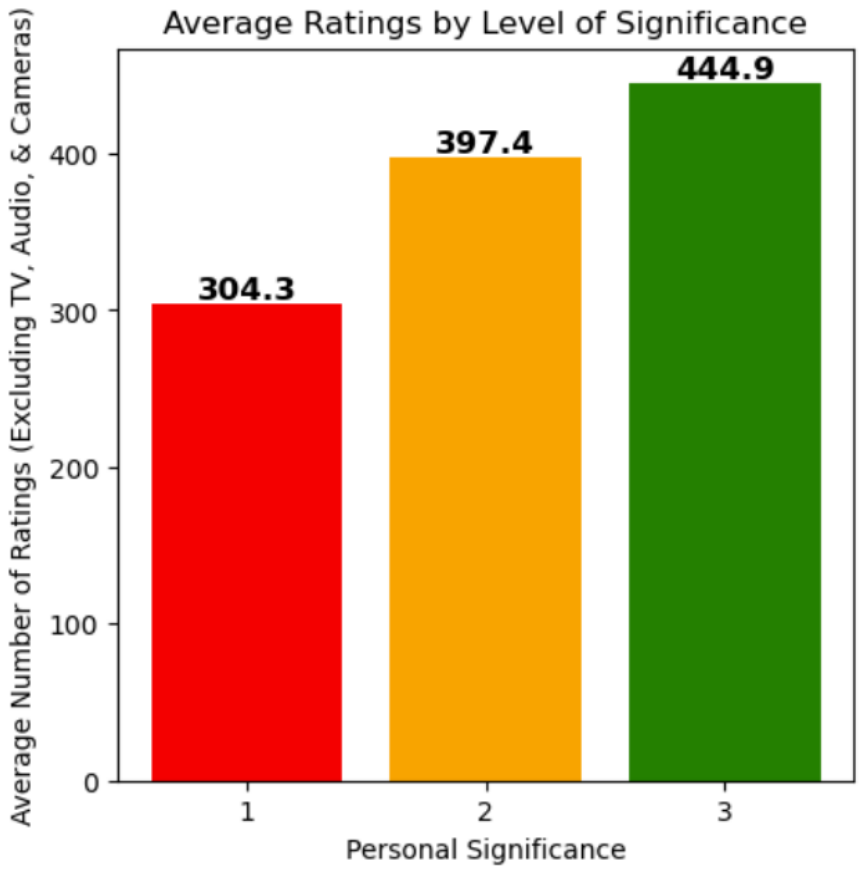


Hypothesis 3

More emotionally-significant categories will get more reviews.

UNC KENAN-FLAGLER BUSINESS SCHOOL

- **Methodology v4:**
  - We removed “tv, audio & cameras” from the dataset
- **Results v4:**
  - When adding a 3rd label option and removing the outlier, the number of ratings increase alongside personal significance.
- **Conclusion:**
  - These findings support hypothesis 3, that people are more likely to rate an item that’s more personally significant.
  - *Major caveat – our largest product category was removed.
  - *Could indicate that personal significance isn’t as straightforward as we initially believed.



Findings and Takeaways

UNC KENAN-FLAGLER BUSINESS SCHOOL

- **Hypothesis #1:**
  - No direct relationship between discount percentages and the number of reviews.
- **Hypothesis #2:**
  - Higher-priced goods tend to receive higher ratings.
- **Hypothesis #3:**
  - People are more likely to rate an item that's more personally significant.*

BUSINESS FOR LIFE.

UNC | KENAN-FLAGLER BUSINESS SCHOOL