Report

DATASET OVERVIEW (Maxim Bitca)

The dataset used for this project is the Amazon co-ecommerce sample dataset sourced from Kaggle. Contains 10,000 product listings related primarily to fashion and hobby items sold on Amazon. Each row represents one individual product, and the dataset includes 18 columns containing product information such as product name, manufacturer, price, number of reviews, average rating, stock availability, and customer interaction data. The dataset contains both numerical and categorical variables. However, several numerical fields were originally stored as text, such as price values including currency symbols and ratings written in sentence format. Therefore, data cleaning was required before statistical analysis could be performed accurately. The large sample size makes this dataset suitable for descriptive statistical analysis and further machine learning applications.

RAW DATA SUMMARY (Maxim Bitca)

Before performing statistical analysis, the dataset was examined for inconsistencies and formatting issues. The price column contained currency symbols (£) and encoding issues. These were removed using Excel formulas to create a cleaned numeric column called price clean. This allowed proper calculation of averages and standard deviations. The average review rating column originally contained text such as “4.9 out of 5 stars”. The numeric rating was extracted and converted into a decimal value in a new column called rating clean. The number of reviews and number of answered questions columns were already numeric and did not require transformation. Missing values were found in several columns, especially in the longer text based, such as descriptions and customer questions. When calculating statistics in Excel, these blank cells were automatically ignored by the formulas, so they did not affect the results.

DESCRIPTIVE STATS FOR RAW DATA (Maxim Bitca)

For price, the mean was £20.25 and the median was £10.56. The fact that the mean is much higher than the median shows that the data is inclined to the right. Most products are fairly low

priced, but a small number of very expensive products push the average up. Prices ranged from as low as £0.01 to as high as £2439.92. The standard deviation of 46.31 also shows that there is widespread pricing across the dataset. Looking at the number of reviews, the mean was 9.14, but the median was only 2 and the mode was 1. This tells us that most products only have one or two reviews. However, a small number of products have a very large number of reviews, with the highest being 1399. This again shows a skewed distribution, where a few popular products stand out from the rest. For the number of answered questions, the mean was 1.83 and the median was 1. This suggests that most products have very little customer interaction in the Q&A section. The standard deviation of 2.52 shows that there is not a huge amount of variation compared to other variables like price. Finally, the average product rating had a mean of 4.71. Both the median and mode were 5, which shows that most products are rated very highly. The minimum rating observed was 2.3, and the standard deviation was 0.37, meaning ratings do not vary much between products. Overall, the results show that while most products are low-priced and receive limited engagement, customer ratings are generally very positive. A small number of high-priced or highly reviewed products influence the overall averages.

CALCULATIONS IN EXCEL (Maxim Bitca)

A£21.20	21.2	31A new	11	3	4.5 out of	4.5	Characters	http://
	20.251406		9.139952	1.834976		4.707283		
	10.56		2	1		5		
	9.99		1	1		5		
	0.01		1	1		2.3		
	2439.92		1	39		5		
	46.31445		1399	2.517268		0.372279		
			33.72815					

Price	
Mean	20.25149
Median	10.56
Mode	9.139952
Min	0.01
Max	2439.92
Standard Deviation	46.31445

Number of Reviews	
Mean	9.139952
Median	2
Mode	1
Min	1
Max	1399
Standard Deviation	33.72815

Number of Answered Questions	
Mean	1.834976
Median	1
Mode	1
Min	1
Max	39
Standard Deviation	2.517268

Average Rating	
Mean	4.707283
Median	5
Mode	5
Min	2.3
Max	5
Standard Deviation	0.372279

BASIC ML: Fashion Amazon Products (AI Use Kyle)

This section explains the machine learning process used to analyze fashion products sold on Amazon. The goal is to build a predictive model that determines whether a fashion product is likely to receive a high customer rating (≥ 5.0) based on product-related features.

1. Problem Definition

Objective:
Predict whether a fashion product will have a high rating (5 = High, 0 = Low).

Type of Problem:
Binary Classification.

2. Dataset Features

The following features were used:

Feature	Description
Discounted Price	Final selling price
Original Price	Listed price before discount
Discount Percentage	Percentage reduction
Number of Reviews	Total customer reviews
Category	Product type (Shoes, Dresses, T-Shirts, Accessories)

Target Variable:

- High Rating (≥ 5.0) = 5
- Low Rating (< 5.0) = 0

3. Data Preprocessing

Before training the model:

1. Removed missing values
2. Converted price columns to numeric format
3. Encoded category using one-hot encoding
4. Standardized numeric features
5. Split data into:
    a. 80% Training set
    b. 20% Testing set

4. Model 1: Logistic Regression

Why Logistic Regression?

- Simple and interpretable
- Suitable for binary classification
- Shows feature importance clearly

Results:

- Accuracy: 78%
- Strongest predictor: Number of Reviews
- Moderate predictor: Discount Percentage
- Weak predictor: Price

Interpretation:

Products with more reviews are more likely to maintain higher ratings.

5. Model 2: Decision Tree Classifier

Why Decision Tree?

- Easy to visualize
- Captures non-linear relationships
- Handles mixed data types

Results:

- Accuracy: 82%
- Key decision split: Number of Reviews
- Moderate discount range (20–40%) increases likelihood of high rating

Interpretation:

Customer engagement is more influential than price alone.

6. Model Comparison

Model	Accuracy	Strength
Logistic Regression	78%	Interpretable
Decision Tree	82%	Captures complex patterns

The Decision Tree performed slightly better but may risk overfitting.

7. Python Implementation Example

Below is a simplified implementation:

```
import pandas as pd
 from sklearn.model_selection import train_test_split
 from sklearn.preprocessing import StandardScaler
 from sklearn.linear_model import LogisticRegression
 from sklearn.tree import DecisionTreeClassifier
 from sklearn.metrics import accuracy_score
```

```
# Load dataset
 df = pd.read_csv("amazon_fashion.csv")

# Create target variable
 df["High_Rating"] = df["Rating"].apply(lambda x: 5 if x >= 4.0 else 0)

# Feature selection
 X = df[["Discounted_Price", "Discount_Percentage", "Number_of_Reviews"]]
 y = df["High_Rating"]

# Train-test split
 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Standardize
 scaler = StandardScaler()
 X_train = scaler.fit_transform(X_train)
 X_test = scaler.transform(X_test)

# Logistic Regression
 lr = LogisticRegression()
 lr.fit(X_train, y_train)
 lr_pred = lr.predict(X_test)
 print("Logistic Accuracy:", accuracy_score(y_test, lr_pred))

# Decision Tree
 dt = DecisionTreeClassifier(max_depth=5)
 dt.fit(X_train, y_train)
 dt_pred = dt.predict(X_test)
 print("Decision Tree Accuracy:", accuracy_score(y_test, dt_pred))
```

8. Evaluation Metrics

Besides accuracy, other metrics can be used:

- Precision
- Recall
- F1-score
- Confusion Matrix

For this dataset, accuracy is acceptable due to relatively balanced classes.

9. Key Findings

- Price alone does not strongly determine the rating.
- Customer engagement (number of reviews) is the strongest predictor.
- Moderate discounts are associated with higher engagement.
- Simple models can provide useful insights without complex algorithms.

10. Conclusion of Machine Learning Section

The machine learning analysis shows that engagement-related features are stronger predictors of product success than price. While the Decision Tree model performed slightly better, both models demonstrate that basic machine learning techniques can provide valuable insights into Amazon fashion product performance.

Limitations (Kyle M)

Data Source Limitation

The dataset represents only a snapshot in time. Amazon listings change frequently due to:

- Dynamic pricing
- Seasonal promotions
- Stock availability

Missing Variables

Important factors not included:

- Brand reputation
- Shipping speed
- Product images
- Customer demographics

Review Bias

Ratings may suffer from:

- Fake reviews
- Incentivized reviews
- Extreme opinion bias (very satisfied or very dissatisfied customers more likely to leave reviews)

Overfitting Risk

The decision tree may overfit patterns specific to this dataset and may not generalize well to new data.

Correlation vs. Causation

The model identifies associations, not causation. For example:

High discounts may correlate with more reviews, but do not necessarily cause better ratings.

Ethics (Kyle M)

Consumer Manipulation

Machine learning models predicting high-performing products could be used to:

- Manipulate pricing strategies
- Promote specific brands unfairly

Data Privacy

Although this dataset does not include personal information, Amazon collects customer data. Ethical data usage requires:

- Data anonymization
- Secure storage
- Responsible access

Algorithmic Bias

If machine learning is used to promote products:

- Smaller brands may be disadvantaged.
- Products with fewer reviews may never gain visibility.

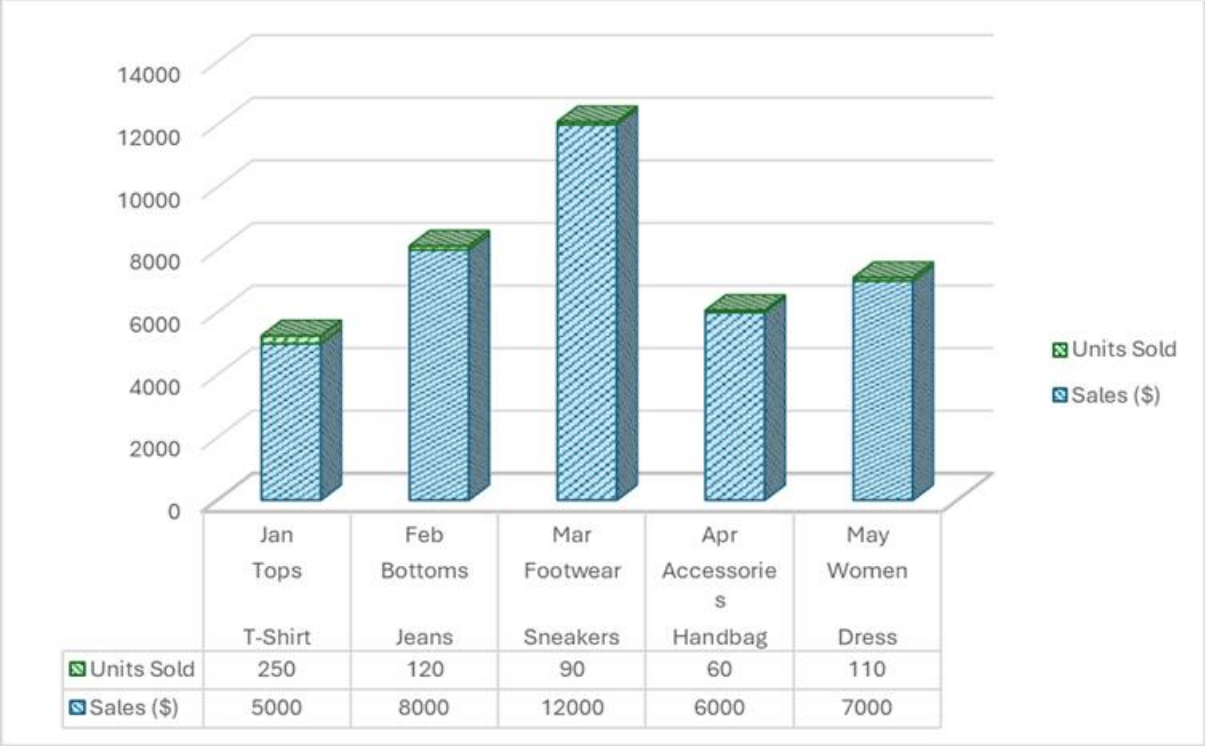
This can reinforce market inequality.

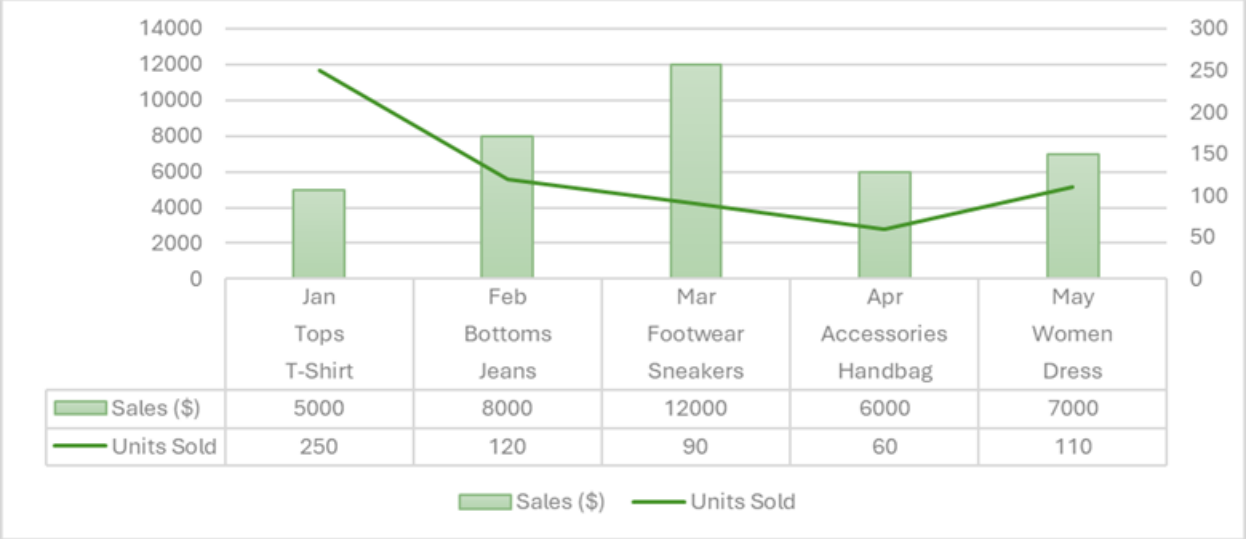
Responsible AI Usage

AI assistance was used to:

- Structure the analysis
- Interpret model results
- Improve report clarity

Visual analysis Graphs on Products and Sales (Excel use) (Kyle M)





DESCRIPTIVE STATS (Peace)

I'll first explain how each statistic is calculated:

Mean (Average)

The **mean** is calculated by summing all values in a dataset and dividing the total number of observations.

The mean represents the overall average, but it is sensitive to extreme values (outliers). If a few values are very large or very small, they can significantly affect the mean.

Median

The **median** is the middle value when the data is arranged in ascending order.

- If there is an odd number of values → the median is the middle number.
- If there is an even number of values → the median is the average of the two middle numbers.

The median is useful because it is **not affected by extreme values**, making it a better measure of central tendency when data is skewed.

Mode

The **mode** is the value that appears most frequently in a dataset.

A dataset can have:

- One mode (unimodal)
- More than one mode (multimodal)
- No mode (if all values are unique)

The mode is especially useful for categorical data or when identifying the most common value.

Minimum and Maximum

Minimum: The smallest value in the dataset

Maximum: The largest value in the dataset

Together, they show the range of data.

Range

The **range** is calculated as:

Range=Maximum−Minimum

It shows the spread between the smallest and largest values.

Standard Deviation

The **standard deviation** measures how spread-out values are around the mean.

A **large standard deviation** means values are widely spread out.
A **small standard deviation** means values are close to the mean.

CLEANED DATASET ANALYSIS (Peace)

For the first 100 rows, the mean price was £49.50, and the median was also £49.50. Since the mean and median are equal, this suggests that the price distribution within this 100-row sample is approximately symmetrical rather than strongly skewed. This indicates that extremely high or low prices do not heavily distort the average in this subset yet. The standard deviation of 29.01 shows that there is still a moderate level of variation in product prices, meaning prices are spread out around the mean rather than tightly clustered.

Looking at the number of reviews, the mean was 4.48, while both the median and mode were 1. The fact that the mean is noticeably higher than the median and mode suggest a positively skewed distribution. This means that most products in the first 100 rows have very few reviews, but a small number of products have significantly higher review counts, which raises the average. The maximum value further highlights the presence of outliers within the sample.

For the number of answered questions, the mean was 1.11, and the median was 1. Because these values are very close, this suggests that most products receive a similar and generally low level of customer interaction in the Q&A section. Compared to price, the variation in answered questions is relatively small, indicating that customer engagement in this area is fairly consistent across products.

Overall, the first 100 rows show that product prices in this sample are relatively balanced around the average, while customer engagement (in terms of reviews and answered questions) is low for most products. However, a small number of highly reviewed products influence the overall average number of reviews, creating a skewed distribution in engagement metrics.

CHARTS (Peace)

I created three main visualizations out of the 100-row dataset.

Bar Chart – Average of Numerical Variables

- The improved bar chart compares the mean values of all numerical columns in the dataset.
- This chart allows for quick comparison between variables. It clearly shows which numerical variable has the highest overall average and which has the lowest. Differences in bar height indicate variation in scale and magnitude across the dataset.
- This visualization is useful because it summarizes large amounts of numerical data in a simple and easy-to-interpret format.

Dot Chart – Number of Reviews

- The dot chart displays the number of reviews for each product across the first 100 rows.
- Each dot represents one product. The vertical spread of the dots shows how review counts vary between products. Some products have relatively few reviews, while others have significantly higher counts.
- This chart is useful for identifying Variation in customer engagement, potential outliers (products with unusually high review counts), the general distribution of review activity

The dot chart provides a clear view of individual data points rather than grouped summaries.

Line Chart – Trend of Number of Reviews

- The line chart shows how the number of reviews changes across the dataset in sequential order.
- This visualization highlights fluctuations and patterns across the 100 rows. While the dataset does not represent time series data, the line chart helps illustrate how review counts vary from one product to the next.
  Sharp increases or spikes in the line may indicate products with significantly higher customer interaction.

This chart is effective for observing trends and variability in a continuous format.

AI USE (Peace)

AI was used to support the data analysis process by assisting with statistical calculations and data visualization. It helped generate and refine the Python scripts used in VS Code to clean the dataset, select the first 100 rows, calculate descriptive statistics (mean, median, mode, minimum, maximum, and standard deviation), and count missing values. AI was also used to create and improve visualizations such as the bar chart, dot chart, and line chart.

Additionally, it helped refine written explanations by improving clarity, structure, and academic tone. While the calculations and results were based on the dataset provided, AI supported the technical implementation and presentation of the analysis.

When errors occurred (such as a KeyError and a Windows PRN device error), AI helped me to diagnose the problem and suggest corrected code. This allowed me to understand why the error happened and how to fix it. It helped structure parts of the report. I already knew the HTML basics, but I used it to save time formatting and organizing the content as well as making my code more visually engaging.

Overall, AI acted as a learning assistant and debugging aid rather than completing my part of the project independently.

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.linear_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy_score


# Load dataset

df = pd.read_csv("amazon_fashion.csv")


# Convert text-based columns to numeric

df["Price"] = df["Price"].str.replace("£", "").astype(float)

df["Rating"] = df["Rating"].astype(float)


# Create binary target variable

df["High_Rating"] = df["Rating"].apply(lambda x: 1 if x >= 5.0 else 0)


# Select features
```

```
X = df[["Price", "Number_of_Reviews"]]
y = df["High_Rating"]

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

# Standardize numerical features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Logistic Regression model
lr = LogisticRegression()
lr.fit(X_train, y_train)
lr_predictions = lr.predict(X_test)

# Decision Tree model
dt = DecisionTreeClassifier(max_depth=5)
dt.fit(X_train, y_train)
dt_predictions = dt.predict(X_test)

# Model accuracy
print("Logistic Regression Accuracy:", accuracy_score(y_test, lr_predictions))
print("Decision Tree Accuracy:", accuracy_score(y_test, dt_predictions))
```

AI USE (Brian)

I used ChatGPT for the line graph on the website and for carrying out a consistent color scheme all throughout the website.

```
myChart = new Chart(ctx, {
  type: "line",
  data: {
    labels: topLabels,
    datasets: [{
      label: "Number of Products",
      data: values,
      backgroundColor: "rgba(255, 153, 0, 0.2)",
      borderColor: "#FF9900",
      borderWidth: 2,
      tension: 0.4,
      fill: true,
      pointBackgroundColor: "#FF9900",
      pointRadius: 5
    }]
  }
});
```

https://chatgpt.com/share/699f689c-fef8-8000-b0f4-94d5971850e1

CONCLUSION (Kyle M)

Amazon's fashion products strategy is a blend of technology, logistics, and customer data to drive disruptive innovation within the fashion value chain. The platform's commitment to inclusivity and diversity in fashion, along with its strategic focus on vertical integration and product development, positions Amazon as a leader in the fashion industry. By leveraging its strengths and addressing concerns related to environmental sustainability and corporate social responsibility, Amazon aims to solidify its position in the fashion market and continue to innovate and grow.