Team Report Submission: Phase 3

Team Name: Amazon

Team Members:

[Chidi Nna], [cnna1@unh.newhaven.edu]

[Venkata Naga Akhil Kuchimanchi], [vkuch4@unh.newhaven.edu]

Research Question

How do product attributes such as price, discount percentage, and review sentiment influence Amazon product ratings and review counts, and can these factors be used to build a predictive model for product popularity?

1. How to identify the importance of factors influencing Amazon product ratings and review counts?
2. How to build a predictive model for product popularity based on identified features?

Data Techniques We Used:

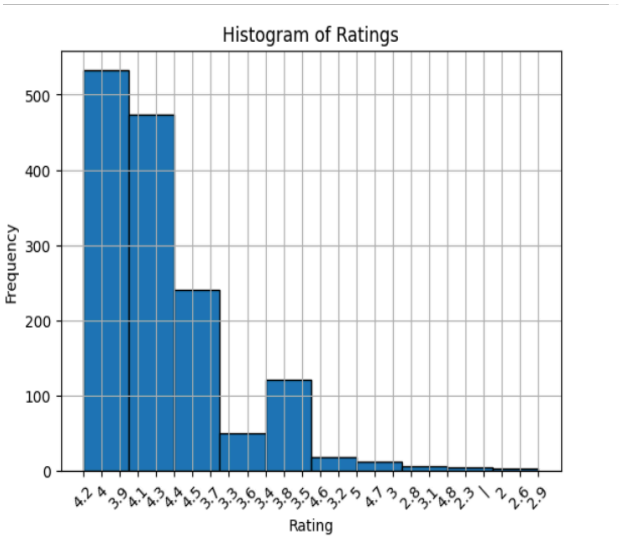
Univariate Analysis:

These techniques focus on a single variable to understand its distribution and summary statistics.

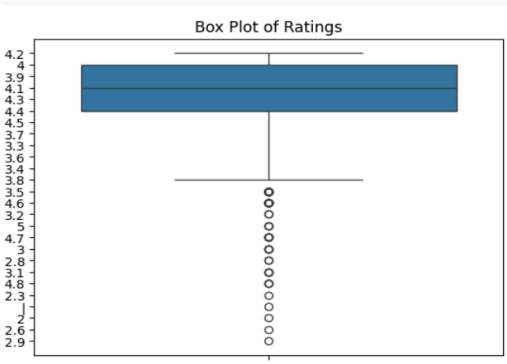
1. **Summary Statistics**:
    - **Mean**: Rating: `3.75358361774744, Rating Count: 77.48737201365188`
    - **Median**: The middle value when the data is sorted: Rating: `Median: 4.0, Rating Count 0`

- **Mode**: The most frequent value. Rating 4, Rating Count 0
- **Variance**: Measures the spread of the data from the mean. `rating: 0.2063139931740599, rating_count: 36942.00274343006`
- **Standard Deviation**: The square root of the variance, providing the spread of the data around the mean. `rating: 0.4542180018163744, rating_count: 192.2030248030193`
- **Range**: The difference between the maximum and minimum values. `rating: 5, rating_count: 992`
- **Quantiles**: Values that divide the data into equal portions (e.g., 25th, 50th, 75th percentiles). `Quantiles for rating:`
- `0.25    4.0`
- `0.50    4.0`
- `0.75    4.0`
- `Name: rating, dtype: float64`
- `Quantiles for rating_count:`
- `0.25    0.0`
- `0.50    0.0`
- `0.75    0.0`

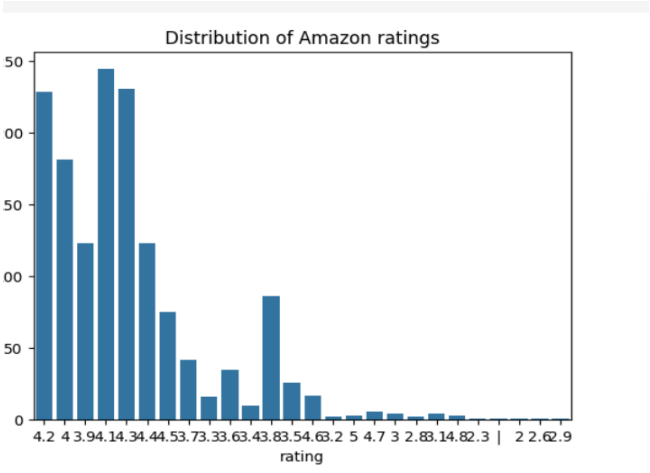
2. **Visualization Techniques**:
    - **Histogram**: Visualizes the distribution of a single variable.



- **Box Plot**: Shows the distribution of a variable, highlighting the median, quartiles, and outliers.



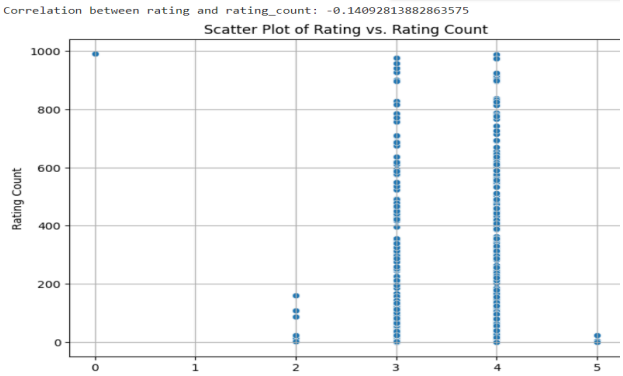
- **Count Plot**



Bivariate Analysis:

These techniques involve examining the relationship between two variables.

1. **Corr**elation:
    - Measures the strength and direction of the linear relationship between two variables. We calculated the Pearson correlation between rating and rating_count.
2. Scatter Plot:
    - A graphical representation showing the relationship between two variables (rating on the x-axis and rating_count on the y-axis)



Conclusion:

Bivariate Analysis: A correlation between rating and rating_count was carried out. In the end, the correlation was -0.14. This suggests that there is a very weak negative correlation between rating and rating_count, which means that the average rating tends to decline slightly as the number of ratings rises. *Univariate Analysis:* The box plot indicates that, with a few outliers on the lower end, the majority of ratings are concentrated toward the top end of the scale, or around 4.0. Both histograms demonstrate that the ratings are centered between 3.9 and 4.2, suggesting a skewed distribution in favor of higher ratings and a general decrease in low ratings. Customers may prefer to provide higher reviews, according to the

Amazon-specific histogram, but there is also a noteworthy cluster of average ratings that hover around 3.5.

GitHub: https://github.com/cnna4/Data-Mining