Ethical Data Science Analysis of Amazon E-Commerce Data

Data collected from https://www.kaggle.com/datasets/grikomsn/amazon-cell-phones-reviews

Statistical Analysis and Insights

After the preparation for the exploratory phase, statistical analysis becomes the next step of any organized data scientist's work. Step beyond the descriptive summaries with statistical techniques and produce deeper information about associations among variables. The statistical analysis has the opportunity to answer questions, in case of the Amazon product dataset, like: Are more expensive products rated better? Is there a correlation between the number of reviews and improvement in ratings? Do discounts correlate with better consumer engagement? These questions look technical, but the answers produce deep ethical, privacy, and social justice ramifications to data-driven decision-making.

Key Statistical Analysis

1. Correlation Analysis

   A correlation matrix may indicate the intensity of associations within the prices, discount percentage, rating, and review of the numbers. For example, the number of reviews might moderately correlate with ratings (popular products might have slightly higher ratings). The lagged price can be weakly correlated with ratings, but much more with the discounts offered, indicating that companies often change prices to influence consumer opinion. Ethically, such knowledge brings up questions of algorithmic amplification. Amazon algorithms may increase the visibility of products with large rating and review counts, and the latter are likely to be mechanically associated with price and brand dominance; hence, inferior or cheaper brands may become an afterthought in the market.

In this manner, a neutral statistical connection may be coding larger economic inequalities.

2. Regression Analysis

   Regression models can predict product ratings based on correlating factors like price, discount and number of reviews. A linear regression analysis would indicate that the counts of reviews have been the most influential predictor of ratings. This shows a cycle of support: products with greater reviews will continue to get more reviews and have an even higher average rating, whereas the lesser-known products remain under the radar. The ethical implication is obvious: rating products can easily be perceived by the consumer as a quality measure, although it is partially a legacy of a statistical positive-feedback loop. This serves to erode equity and observance in the digital marketplaces.

3. Distribution Analysis

   The distribution of prices reveals that only several products cost more than 1000 dollars whereas most are between 100 and 400 dollars. Ratings are also skewed towards the positive and numerous products would be 3.5 to 4.5 stars. This creates suspicion about false ratings and the possibility of companies promoting favourable ratings and making them visible. Regarding social justice, the inflated rating is detrimental to consumers who are less digitally literate and will not have access to the tools of criticism to recognize a manipulative rating pattern.

4. Hypothesis Testing

   Hypothesis testing can compare the average ratings of high-quality and low-quality brands. Let us take the scenario where we are interested in testing the null hypothesis which conveys that we have no difference in ratings between the high-end brands (Apple,

Samsung) and the pocket friendly brands (Xiaomi, Motorola). When the null hypothesis is rejected through statistical analysis, it indicates some systematic difference in consumers' perceptions of the quality of the brands. However, rejecting the null hypothesis does not prove that premium brands are superior, which could be because of cultural inclinations, marketing, or socio-economic factors. As ethical scientists, data scientists are obliged not to exaggerate causal interpretations.

Analysis Statistical Ethics

Statistical techniques are not neutral in regard to values. Interpretation of correlations or model construction may tend to entrench existing inequalities. For instance:

- Manipulation of Consumer: Showing that historical positive reviews count when the discount is applied may bias the companies to inflate the original prices and display phony discounts. Consumers feel highly exploited because such practices take advantage of the psychological aspect of consumers. At the same time, low-income groups are disproportionately impacted, which raises the issue of fairness.
- Bias Amplification: Statistical models will support already market-dominant products and brands. This only fuels inequality over and above giving consumers varied and balanced options.
- Transparency and Accountability: Companies do not explain how their regression models operate even as they continue to recommend their products, thus leaving consumers in the dark. Such opacity is against ethical data science.

Privacy Considerations

Although this data does not contain personal identifiers per se, statistical interpretations are occasionally able to provide sensitive data about consumer cohorts. For example, it could

happen that, after analysis, it appears that, in the case of budget phones, higher numbers of reviews were uploaded; this could induce the idea that poorer consumers are more active in online reviewing. It is not harmful in itself, but in combination with other datasets, it might be used to profile the vulnerable classes. In an ethical statistical setup, inference should be made cautiously because it may be wrong and cause acknowledgment of the suffering of a certain population.

Aspects of Social Justice

From a social justice point of view, the statistical analysis indicates inequality of access, quality, and representation:

- Access Disparity: Premium products can have more presence, which might mean cheaper options are inaccessible.
- Quality Perception Bias: Research findings on statistical differences in ratings cause the establishment of cultural perceptions between the notions of good and bad brands, which may stigmatize low price-focused alternatives.
- Platform Power: Statistical wisdom- Amazon as a platform has tremendous power to influence consumer behaviour in its algorithmic choices which are often non-democratic, and nearly transparent.

Conclusion

Looking at the Amazon dataset through statistical analysis, the numbers hide further truths of a systemically manipulative process of consumers, brand hegemony, and rating inflation. Interpretation of these findings will be refracted through ethical, privacy, and social justice perspectives, preventing data scientists from presenting misleading ideas on data issues and representing the encouragement of greater transparency and equity of actions in digital

marketplaces. Rather the responsible statistical analysis needs to inquire who gains access to these insights and who is excluded because of these predictions?