Ethical Data Science Analysis of Amazon E-Commerce Data

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

Exploratory Data Analysis (EDA) and Initial Insights

Exploratory Data Analysis (EDA) is the initial step where it is possible to identify meaningful patterns in a set of data and interpret them. In the Amazon product dataset, the EDA is performed by creating descriptive statistics, visualizations, and identification of trends, and it describes the distribution of pricing, the number of reviews, discounts, and brand representation. Although the practical goal of EDA is to get a better feel of the dataset's characteristics and find ways to prepare it for other analyses, it also becomes a decision point involving some ethical, privacy, and social justice considerations in data science.

Technically, EDA will enable us to examine some important data set features. Another example is the mobile products available in thousands of items, which are well-known brands, including Motorola, Samsung, Apple, and Xiaomi. Frequency counts cannot intensify the observation that some brands populate the data, and others have a negligible number of entries. Price distributions are expected to show that most products occupy a median range price band where some high prices models such as Apple iPhones distort the average price. There is also an extreme gap in the number of reviews: some products have tens of thousands while others have less than 100. Such a distribution has been referred to as the long tail distribution commonly seen in e-commerce where a few products take up most of consumer interest.

Nonetheless, such patterns could not be interpreted outside ethical considerations. As an example, it is not merely a technical result to be able to specify the power of brands in the database; it is also the way the large corporations have structured consumer decisions. In comparison, smaller brands with limited reviews stand a chance of being ditched in Amazon's

algorithms and researchers' studies. Without Mindfulness, Mindfulness to fairness, the insights gained during the process of EDA may, therefore, contribute to the prevalence of these disparities by favoring best-selling products and disregarding the minority. An ethically responsible attitude to EDA would thus encompass the following question: which voices are louder and quieter in the data?

Ethical implications can also be included in visualizations performed during EDA. Even a rudimentary comparison of the average ratings of various brands, e.g., using a bar chart, can potentially lead to the conclusion that certain companies are objectively better at producing products. Yet, such an interpretation still has no social and economic contexts, e.g., marketing budgets, brand loyalty, and cultural attitudes towards quality. Ethical EDA raises the need among researchers to challenge the measurement that the data truly represents. Ratings cannot be objective in quality determination; they are social influences of consumers' perceptions. Likewise, the number of reviews is not always a clean indicator of popularity because the platform algorithm dictates how particular products will be prioritized.

In the context of non-disclosure, EDA must also evaluate the risks of re-identifying or inadvertent disclosure of sensitive information. Although individual reviewers are not disclosed, regional patterns in demand may occasionally be indirectly revealed by high-level distributions of reviews. If these insights are incorporated with other data sets, they may be potentially attributed to certain communities or demographics. For example, disproportional review of budget smartphones in a particular area may be used to estimate a population's income rates or buying power. This might appear innocuous regarding market research, but worries arise when corporations or governments exploit this information to be exploitative. It is the responsibility of the EDA to impose restraint when it comes to interpretation because of these risks.

Such reflections are further enriched by a social justice approach that brings into the foreground the role that EDA can play in unmasking structural inequalities. For example, at an initial analysis, the cheaper models rank lower regarding ratings than the premium models. Superficially, this may be an indication that cheap devices are inferior. Nevertheless, a socially reasoned interpretation now examines the rating structures' bias. There is an expectation that more upper-end consumers buying the more expensive devices would have greater expectations, which translates into satisfaction and access to better after-sales services. In the meantime, low-income consumers can have no support, increase the number of defects, or live less long with their products, which impacts the review. Therefore, EDA is more than a technical process; it functions as a tool to see disparities in consumer experience.

The other important aspect of EDA is the relationship between the percentage discount and price. Pretend credits are often projected onto the original prices of many products and big discounts. Although doing this may give the appearance of savings, such practices have created ethical issues with respect to the art of deception with regard to product pricing. Detection of such trends in EDA directly relates to consumer protection and fairness. A responsible socially aware data scientist would not only report these outcomes but also interpret how these pricing strategies manipulate the behavior of the consumers and particularly disproportionately impact vulnerable populations, who are more price-sensitive.

Lastly, EDA transparency is an ethical concern in and of itself. In many cases, analysts do a lot of tests and visualizations and mention only the most interesting or desirable analysis. Such a case is called cherry-picking, which is deceptive to the audience and reduces trust in data science. Responsible EDA demands that one records every action, makes important and trivial descriptive conclusions, and takes precise differences between correlation and causality.

To sum up, EDA on Amazon's data does not simply describe the distribution and trends. It can be an ethical gateway where the researcher must consider the extent to which their content and conclusions are fair, their privacy implications, and their social justice. This combination of proportional rigor and principled consideration suggests that EDA is an instrument to comprehend data and a vehicle to increase ethical and fair data science operations.