

Assignment Report: Data Analysis on Flipkart Products Dataset

1. Introduction

The objective of this assignment was to perform data analysis on a real-world dataset from Kaggle. The dataset used was Flipkart Product Listing, containing product information such as product names, prices, ratings, and specifications. The primary tools used were NumPy and Pandas in Python.

2. Dataset Overview

- Source: Kaggle (Flipkart Product Listings)
- File: flipkart_products_20250405.csv
- Key Columns:
 - product_name
 - brand
 - retail_price
 - discounted_price
 - product_rating
 - product_category_tree

3. Tools & Libraries Used

- NumPy: For numerical computations and array manipulations.
- Pandas: For data loading, cleaning, transformation, and analysis.
- Google Colab: Used for coding and generating this analysis.

4. NumPy Operations

- Created numerical arrays from product price columns.
- Performed basic statistics like mean, median, and standard deviation on product prices.

- Applied vectorized operations to compare retail and discounted prices.
- Filtered arrays using boolean indexing to find products with significant discounts.

5. Pandas Operations

- Loaded the dataset and displayed the first few entries using `.head()`.
- Handled missing values and removed irrelevant columns.
- Used `.groupby()` to analyze average product ratings by brand.
- Sorted data based on product price and ratings.
- Used `.apply()` and `.map()` to clean string columns.
- Filtered top 10 most expensive and top-rated products.

6. Combination Dataset Operations

- Merged data from multiple columns to create new insights (e.g., price difference).
- Created new columns to show discount percentages.
- Combined NumPy array operations with Pandas DataFrame to enhance analysis.
- Exported cleaned and processed datasets to new CSV files.

7. Key Insights

- Many products had large price discrepancies between retail and discounted prices.
- Certain brands consistently had higher ratings.
- A significant number of products had missing or zero ratings.
- Discounts were found to vary significantly across product categories.

8. Conclusion

This assignment successfully demonstrated how to use NumPy and Pandas for practical data analysis. By combining their capabilities, it was possible to extract valuable insights from a raw dataset and prepare it for further use or visualization.