Figure 1: Logo

DESCRIPTIVE ANALYTICS PROJECT (CSD301)

Customer Review Analysis (Electronics)

By

Sr. No.	Registration No	Name of Students	Roll No
1	12303116	J V Purushotham	09

Submitted To

Mr. Madhav Dubey Sir, Assistant Professor
Lovely Professional University
Jalandhar, Punjab, India.

Delivered by:
Name of the student: J V Purushotham
Reg. No.: 12303116

Received by:
Name of the faculty: Mr. Madhav Dubey Sir

Acknowledgement

I would like to express my sincere gratitude to Mr. Madhav Dubey Sir for their invaluable guidance and support throughout this project. I also extend my appreciation to Flipkart for providing a platform that enabled the collection of electronic product reviews, which served as the foundation of this research. Finally, I thank my peers and family for their encouragement and constructive feedback during the course of this study.

Abstract

This research is intended to analyse Flipkart-scraped electronic product reviews in order to get insights into consumer preferences, sentiment, and product performance as a whole. The dataset contains prominent features like product URLs, titles, number of ratings and reviews, author names, ratings, review titles, and review text in detail. The aim of this analysis is to identify relevant patterns using descriptive statistics and regression analysis, which will enable business owners and consumers to make educated decisions.

With the use of statistical methods, we investigate central measures like mean, median, and mode along with regression analysis in order to develop relationships among variables like reviews and sentiments in the reviews. All these results deliver important information on customer behaviour as well as on market trends across the electronics sector. This research also involves numerous exploratory data analysis methods, visualization strategies, and predictive model techniques, alongside statistical and regression analysis, for deeper knowledge regarding customer reviews.

Through the use of machine learning models, we try to categorize customer sentiments, forecast ratings from review texts, and detect salient themes that arise from customer opinions. The study also looks into the effect of other variables like product type, price level, and reputation of the brand on customer satisfaction levels. In addition, this project analyses the real-world implications of the findings derived. Companies can apply these results to fine-tune marketing strategies, enhance product characteristics, and upgrade customer service.

Customers can gain from data-driven suggestions and make more educated buying choices. We also present the study's limitations, possible biases in customer reviews, and recommendations for future research in this area. Finally, this thorough analysis aims to close the gap between customer feedback and business strategy, creating a more data-driven e-commerce decision-making process. The findings to be produced out of this project can help drive improved product innovation, improved customer experiences, and improved business management in the world of electronics.

Contents

List of Figures

Acronyms and Abbreviations

- URL: Uniform Resource Locator
- ML: Machine Learning
- NLP: Natural Language Processing
- TF-IDF: Term Frequency-Inverse Document Frequency
- BoW: Bag of Words
- POS: Part of Speech
- LDA: Latent Dirichlet Allocation

Chapter 1

Introduction

Problem Statement

Online product reviews play a crucial role in shaping consumer decisions. The challenge lies in effectively analysing large volumes of customer feedback to identify meaningful patterns and trends. This project seeks to bridge this gap by conducting a statistical and predictive analysis of Flipkart electronics reviews, offering valuable insights into customer behaviour, product performance, and market trends.

Introduction about the Dataset

In today's digital era, online shopping platforms like Flipkart have become a primary destination for purchasing electronic products. With thousands of customer reviews available for each product, analysing these reviews provides valuable insights into customer satisfaction, product performance, and areas of improvement.

Customer Review Analysis involves extracting meaningful patterns and sentiments from user feedback to understand how consumers perceive electronic products. By leveraging techniques such as sentiment analysis, text mining, and machine learning, businesses can assess product quality, identify recurring issues, and enhance customer experience.

This analysis helps buyers make informed purchasing decisions while enabling brands to refine their products and services based on customer expectations. In this study, we will explore reviews of electronic products on Flipkart, categorize customer sentiments, and derive key insights to improve both customer satisfaction and business strategies.

Chapter 2

Extraction of Data Using Web Scraping Technique

2.1 Introduction

To perform exploratory data analysis (EDA) on real-world product reviews, raw data was extracted directly from an e-commerce platform. For this project, customer reviews were scraped from the Flipkart website, focusing on electronics such as mobiles, laptops, and tablets.

2.2 Web Scraping Tool Used

The data extraction was carried out using the Python library **BeautifulSoup**, which allows efficient parsing and extraction of information from HTML and XML files. The process also incorporated the **requests** module to fetch page content from Flipkart.

2.2.1 Libraries Used

- **BeautifulSoup (bs4)** – For parsing the HTML structure of web pages and locating desired tags.
- **requests** – For sending HTTP requests to the Flipkart server and receiving page content.
- **pandas** – To structure the extracted data into DataFrames.

2.3 Source Website

The reviews were extracted from the product review section of the Flipkart website (`https://www.flipkart.com`). Flipkart is a popular Indian e-commerce platform where users post genuine reviews and ratings for electronics and other products.

2.4 Data Collection Strategy

To ensure fair representation and avoid bias:

- **Two consecutive review pages** were scraped for each product category (mobiles, laptops, and tablets).
- The approach avoided overrepresentation of highly positive or negative feedback that typically appears on the first page.
- Each page contained around 10 reviews, resulting in a sample of approximately 20 reviews per category.

2.5 Fields Extracted

The following review fields were extracted:

- **Product URL**
- **Title**
- **Number of Ratings Reviews**
- **Authors**
- **Review Titles**
- **Review Texts**

2.6 Challenges Faced

- Flipkart employs dynamic content loading via JavaScript, which occasionally required handling asynchronously loaded elements.
- Inconsistent HTML tag structure across different product categories was handled by designing robust scraping logic.
- Frequent access requests could trigger CAPTCHA, requiring time intervals between requests.

2.7 Conclusion for webscrapping the data

Using BeautifulSoup and related libraries, a clean and relevant dataset of customer reviews was successfully extracted from Flipkart. The data collected serves as the foundation for further preprocessing, sentiment analysis, and dashboard visualization in subsequent chapters.

Chapter 3

Data Understanding

The dataset used in this project was collected from product review pages on Flipkart. It consists of user reviews, ratings, and other relevant details about mobile products listed on the e-commerce platform. The data was extracted in a structured tabular format with the following attributes:

Column Name	**Description**
Product URL	The web address (link) of the product's review page on Flipkart. Each entry points to a specific product. Example: `https://www.flipkart.com/motorola-g05-forest-g...`
Title	The full title of the product, including brand name, model, variant, and storage capacity.
Number of Ratings & Reviews	The total number of ratings and reviews received for the product, possibly including symbols (e.g.,). Example: 4.3 5,105 Ratings & 413 Reviews
Authors	Names and information about users who submitted the reviews, sometimes including status (e.g., Certified Buyer).
Ratings	Comma-separated list of numerical ratings given by users (scale 1-5). Example: 5, 5, 5, 5, 5, 4, 5, 4, 3, 3
Review Titles	Short summary titles given by users about their experience. Example: Great product, Terrific, Terrific purchase, Just wow!
Review Texts	Full detailed review text written by users, containing feedback and observations.

Chapter 4

Preprocessing and Cleaning of Data

The dataset was collected from Flipkart using the BeautifulSoup library in Python. Reviews were scraped and separated into three categories: **mobiles**, **laptops**, and **tablets**, and processed using different Python files.

4.1 Initial Preprocessing Steps

- Assigning unique Product IDs for each item.
- Total number of products considered:
  - Mobiles: 61
  - Laptops: 74
  - Tablets: 53
- Changing datatype of the `Number of Ratings & Reviews` column to extract:
  - `Overall Rating`
  - `Number of Ratings`
  - `Number of Reviews`
- Extracting product-specific information from the `Title` column:
  - For Mobiles: Model, Colour, Storage
  - For Laptops/Tablets: Brand, Model, Specifications (and Colour/WiFi Type for tablets)

4.2 Text Processing Steps

- Cleaning the `Review Title` and `Review Text` columns:
  - Removing punctuation
  - Converting text to lowercase
  - Removing common phrases like "READ MORE"
  - Removing emojis using regex

  - Removing stop words
  - Removing numbers
  - Stripping extra spaces

4.3 Handling Missing Values

- Checked for null values in key columns such as `No_of_Ratings`.
- Filled missing values using:
  - Mean values for ratings
  - Mapping based on `Product ID` for filling `Model` and `Overall Rating`
- Dropped irrelevant or redundant columns: `Number of Ratings & Reviews`, `Authors`, `Ratings`
- Reordered and renamed columns for better clarity
- Replaced remaining null values with the term `Unknown`

4.4 Export for Further Analysis

- Exported cleaned data to CSV format for each category (mobiles, laptops, tablets)
- Cleaned datasets were then used for Exploratory Data Analysis (EDA)

Chapter 5

Exploratory Data Analysis

To enhance the analytical capabilities of our datasets, a new column `Price` was added to each of the datasets (Mobiles, Laptops, and Tablets). This will significantly contribute to more meaningful insights, especially in price-based segmentation and comparisons. Following this, the order of the columns was adjusted for logical readability, and the datatypes of each column were changed appropriately to reflect their correct data types, such as converting numerical strings to integers/floats, and categorical data to category type.

5.1 Mobile Data Analysis

Univariate Analysis

- Histograms were plotted for `Overall Rating`, `No of Reviews`, `No of Ratings`, `Price`, and `Average Rating` to understand the distribution of these numerical variables.
- Histplots were used to examine how `Overall Rating` and `Average Rating` vary with `Storage_GB`.
- An ECDF (Empirical Cumulative Distribution Function) plot was created for `Price` to better visualize cumulative probability distribution.
- Stripplot and violinplot were used to study the relationship between `No of Ratings` and `Average Rating`, and between `No of Ratings` and `Overall Rating`.
- Price categorization was performed: Budget (¡10000), Mid-Range (10000-20000), and Premium (¿20000). A count plot was then used to visualize product distribution across these categories.
- Count plots were also used for most popular smartphone colors and common storage options.
- A bar plot visualized the top 10 most reviewed smartphone models.
- Histogram of review length (in words) to see the distribution.

Bivariate Analysis

- Pairplot among `Price`, `Overall Rating`, `Average Rating`, `No of Ratings`, and `No of Reviews`.
- Line plots: `Price` vs `Storage` and `Color` vs `Price`.
- Box plots for price across different storage and review categories.
- Regplot: `Price` vs `No of Reviews`.
- Parallel coordinates plot for `Price`, `No of Ratings`, `Overall Rating`, and `Model`.
- Bubble plots and scatter plots for multiple combinations, with size or color representing key metrics.
- Count and bar plots based on grouped price and rating.
- Sentiment vs `Overall Rating` scatter plot and heatmap.

Multivariate Analysis

- 3D scatter plots for `Price`, `Model`, `Overall Rating` vs `Review Length`.
- Sentiment analysis using `SentimentIntensityAnalyzer` to classify reviews.
- Encoding of sentiment and model, followed by 3D scatter visualizations.
- Bar plots for keyword frequency, both positive and negative, from extracted keywords.
- Bar plots for n-gram analysis (bigrams and trigrams).
- Topic modeling using LDA (Latent Dirichlet Allocation) to extract hidden topics in text data. Each topic is visualized using bar plots showing top terms.
- Correlation matrix for understanding relationships among numerical columns.

5.2 Laptop Data Analysis

Univariate Analysis

- Histplots for `Overall Rating`, `No of Reviews`, `No of Ratings`, `Price`, and `Average Rating`.
- Box plots and KDE plots combined for `Price`, `Overall Rating`, and `Average Rating`.
- ECDF plot for `Price`.
- Boxen plot for `No of Ratings`—useful for long-tailed data.
- Brand distribution visualized with pie charts, bar plots, and stacked barplots (brand vs number of models).

Bivariate Analysis

- Pairplot and heatmap of correlation matrix among main numeric variables.
- Bubble plot and hexbin plot for `Price` and `Overall Rating`.
- Violin, regression, joint, and error-bar bar plots for multiple bivariate relationships.

Multivariate Analysis

- 3D scatter plots and FacetGrid visualizations for multiple features segmented by brand.
- Hierarchical clustering heatmap to identify similar product groupings.
- Radar chart to show multiple metrics for each brand.
- KDE-based contour plots for pairwise density.
- Price buckets created and analyzed using cross-tab heatmaps and box plots.
- Extracted `RAM`, `Storage`, `OS Version` from specification column.
- New metrics such as `Price per GB RAM/Storage`, `Review Engagement Score`, and `Rating per 1000 Price`.
- Sentiment analysis with histograms and label distributions.
- Word frequency plots, keyword extraction, n-gram bar plots, treemaps, and average review length.
- Topic modeling (LDA) with 5 topics visualized via bar plots.
- Clustering with KMeans (k=3) to group laptops.

5.3 Tablet Data Analysis

Univariate Analysis

- Histplots for `Overall Rating`, `No of Reviews`, `No of Ratings`, `Price`, `Average Rating`.
- ECDF for `Price`.
- Word and character count analysis for review titles and texts.
- New metric: `Rating Density = No of Ratings / No of Reviews`.
- Sentiment analysis using `SentimentIntensityAnalyzer` and plotting results.
- Keyword extraction and visualization.
- Distribution of features like `battery, display, performance, camera` in reviews.
- Extracting `RAM_GB`, `Storage_GB`, `Screen_Size_Inches` and analyzing their distribution.
- Calculating `Popularity Score = No of Ratings  Average Rating`.

Bivariate Analysis

- Pairplot for primary metrics.
- Multiple scatterplots and boxplots for new derived metrics.
- Line and bar plots to compare features like `Storage`, `Screen Size`, `WiFi Type` vs other metrics.
- Correlation analysis and hexbin plots.

Multivariate Analysis

- 3D scatter and heatmaps combining multiple features.
- Bubble plots and violin plots combining sentiment, storage, and screen size.
- Parallel coordinates plots for multidimensional analysis.
- LDA Topic modeling with 5 topics visualized using bar plots.
- Product clustering using KMeans (k=3).

Chapter 6

Interactive Dashboard with Streamlit

To enhance user experience and enable dynamic data exploration, three separate interactive dashboards were developed using **Streamlit**—one each for **mobiles**, **laptops**, and **tablets**.

6.1 Dashboard Features

Each dashboard allows users to filter and visualize data based on various specifications, enabling focused analysis and deeper insights. Key features include:

- **Category-Based Dashboards**: Three dedicated dashboards were created, tailored to the unique attributes of mobiles, laptops, and tablets respectively.
- **Filter Options**:
  - **Mobiles**: Users can filter by storage capacity, color, price range, rating, and sentiment of reviews.
  - **Laptops**: Users can filter by brand, processor type, RAM size, price range, average rating, and number of reviews.
  - **Tablets**: Users can filter by screen size, storage, battery life, brand, and rating.
- **User Interaction**: Real-time data updates allow users to instantly observe how changes in filters affect the analysis, helping them draw comparisons and make informed decisions.
- **Responsive Design**: The dashboards are structured for intuitive navigation and responsive layout across devices.

6.2 Technology Used

- **Backend**: Python (pandas, numpy, seaborn, matplotlib, nltk, sklearn)
- **Frontend**: Streamlit for creating interactive user interfaces

6.3 Conclusion for streamlit dashboarding

These dashboards serve as practical tools for exploring the EDA results in an interactive environment, making the analysis accessible and actionable for a broader audience including product analysts, marketers, and consumers.

Learning Outcomes

Through the successful execution of this project, the following key learning outcomes were achieved:

1. Gained hands-on experience with **web scraping techniques** using Python libraries such as `BeautifulSoup` and `requests` to extract real-time data from e-commerce websites.
2. Understood the importance of **data preprocessing and cleaning** in preparing unstructured textual data for meaningful analysis.
3. Learned to perform **exploratory data analysis (EDA)** to derive insights from customer reviews and feedback.
4. Developed the ability to design and implement an **interactive dashboard** using `Streamlit` for visualizing and filtering large datasets based on user-defined specifications.
5. Improved skills in **data visualization and user interface design**, enhancing the interpretability of analytical results.
6. Understood the necessity of collecting **diverse and unbiased data** for more accurate and fair analysis by considering multiple review pages.
7. Strengthened the ability to work on an **end-to-end data science pipeline**, from data collection to user-centric visualization.
8. Improved project documentation and reporting skills in a structured LaTeX format.

Overall Conclusion

This project aimed to perform exploratory data analysis (EDA) on customer reviews for electronic products—specifically mobiles, laptops, and tablets—sourced directly from the Flipkart website using web scraping techniques.

The data extraction process utilized the **BeautifulSoup** library, combined with the **requests** module, to collect unbiased and relevant user feedback from two review pages for each product category. By carefully selecting multiple pages, the dataset captured a more balanced perspective of user sentiments, avoiding the skew commonly introduced by single-page reviews.

After successfully collecting the data, it was cleaned and preprocessed to ensure consistency, completeness, and suitability for analysis. This step involved handling missing values, standardizing formats, and organizing data for efficient filtering.

A user-friendly dashboard was then developed using **Streamlit**, providing interactive filters and category-wise exploration. Three separate dashboards were created to visualize reviews related to:

- Mobile Phones
- Laptops
- Tablets

These dashboards allow users to filter reviews based on various specifications and understand product feedback trends in an intuitive manner.

In conclusion, this project successfully demonstrated the end-to-end pipeline of:

1. Real-time data collection from a public platform,
2. Preprocessing of unstructured text data,
3. And effective visualization through dashboard interfaces.

The approach can be extended or modified for other product categories or platforms, showcasing its adaptability for broader applications in review analytics and customer sentiment research.