**CAPSTONE PROJECT: END-TO-END DATA SCIENCE PROJECT**

**E-COMMERCE PRODUCT DATA ANALYSIS**

**NANTHINI H**

**DA&DS - 2025**

1. **Introduction**

This project focuses on analyzing watch products available on Flipkart by building an end-to-end data pipeline.

The dataset was collected through web scraping and includes key attributes such as product name, price, ratings, and customer reviews.

The primary objective of this project is to gain meaningful insights into product trends, pricing patterns, and customer preferences, while also applying clustering and classification techniques to enhance understanding of the data.

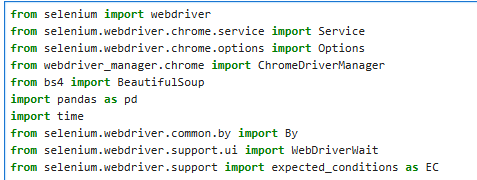
The project covers the entire workflow—data acquisition, cleaning, exploratory analysis, unsupervised clustering, and supervised modeling—providing both analytical insights and predictive solutions for real-world e-commerce scenarios.

**2. Project Workflow**

* **Data Collection** – Scraped watch product data (price, rating, reviews, name) using Python libraries like Requests and BeautifulSoup.
* **Data Cleaning** – Handled missing values, removed duplicates, standardized formats, and engineered new features (e.g., log-price).
* **Exploratory Data Analysis** – Used visualization tools to analyze distributions, correlations, and category-based patterns.
* **Unsupervised Learning** – Applied KMeans clustering, validated results with Elbow Method and Silhouette Score.
* **Supervised Learning** – Implemented models (Logistic Regression, SVM, KNN, Random Forest) and compared performance metrics.
* **Evaluation** – Measured accuracy, precision, recall, F1-score, and ROC-AUC; identified the most effective model.

1. **Data Understanding**

**Source**: Flipkart e-commerce website, scraped using BeautifulSoup.



**Key Features:**

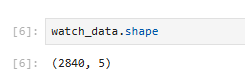
Product Name

Price

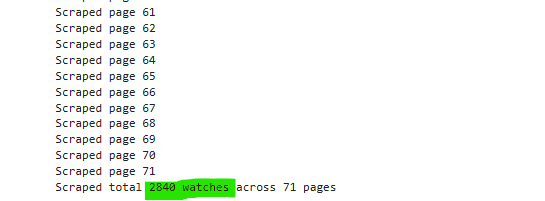
Ratings

Number of Reviews

**Dataset Size:** 2804 Rolws and 5 Columns.



**Challenges:** Presence of missing values, inconsistent formats in price.





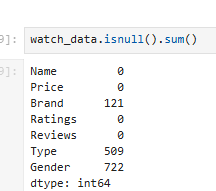
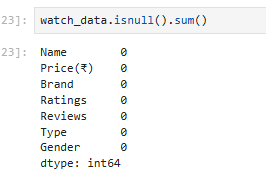
1. **Data Cleaning**

After collected the data, Gender and Type columns were extracted from Name column.



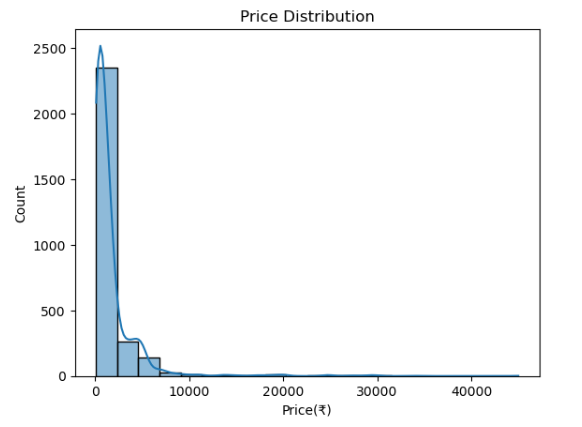
Before building models, it was necessary to clean and prepare the dataset. The first step was to identify missing values across all columns.

Columns such as Brand, Type, Gender had a few missing entries, which were handled using appropriate techniques such as imputation or removal of incomplete rows.

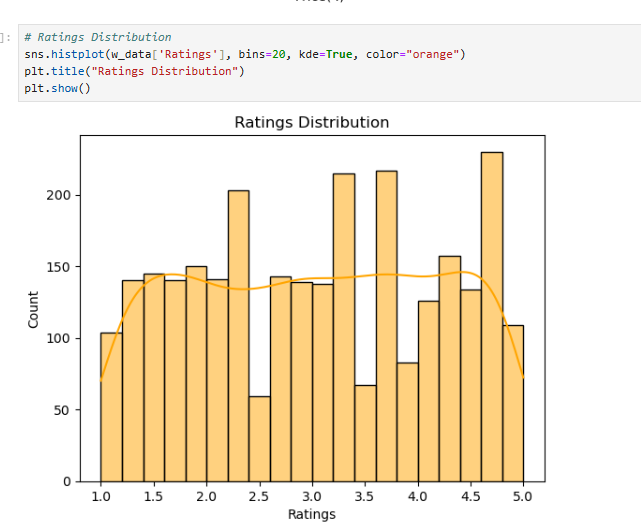
****

1. **Exploratory Data Analysis (EDA)**

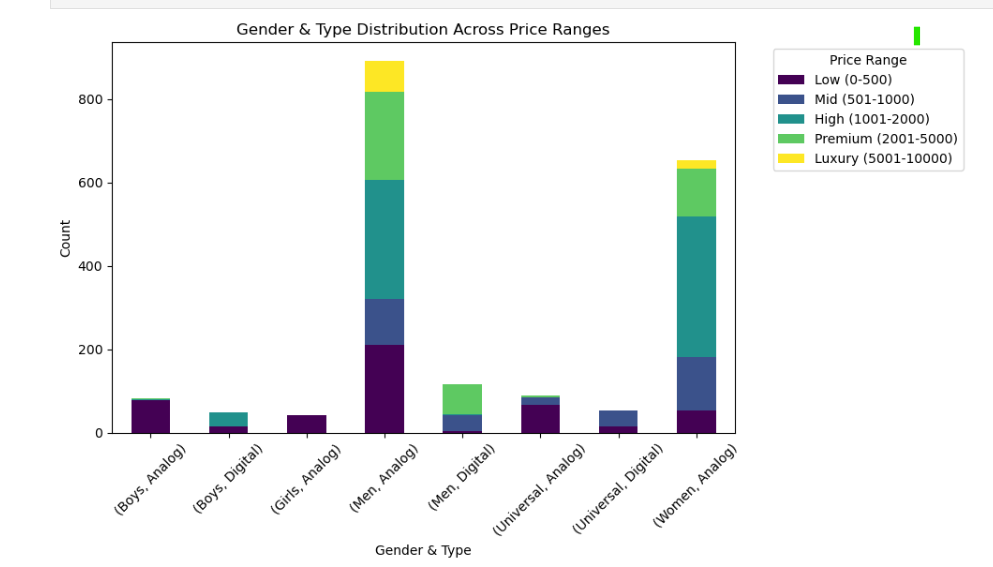
* **Price Distribution**: Most watches fell in the affordable to mid-range price categories, with a small number of premium watches. The distribution was right-skewed, justifying the need for log transformation.



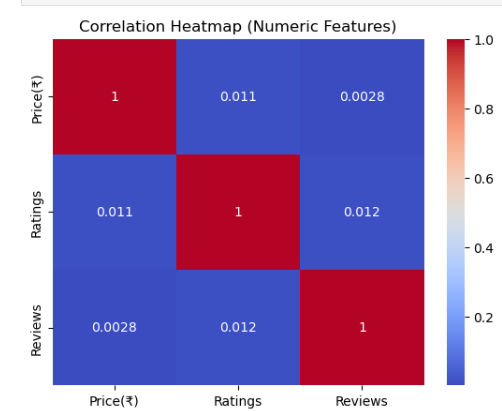
* **Ratings Distribution**



* The **analysis of gender and type distribution** across different price ranges highlights a clear dominance of **Men’s Analog watches**, which span all price categories from low to luxury, indicating wide consumer demand and availability.
* **Women’s Analog watches** form the next significant group, primarily concentrated in the mid to high ranges with some premium and luxury representation.
* In contrast, **Digital watches** (across Men, Boys, and Universal categories) show limited presence, suggesting lower popularity compared to analog models.
* Watches designed for **Boys and Girls** remain a niche, mostly confined to affordable ranges. Overall, the trend reveals that analog watches, especially for men, drive the majority of the market, while other segments remain comparatively small.

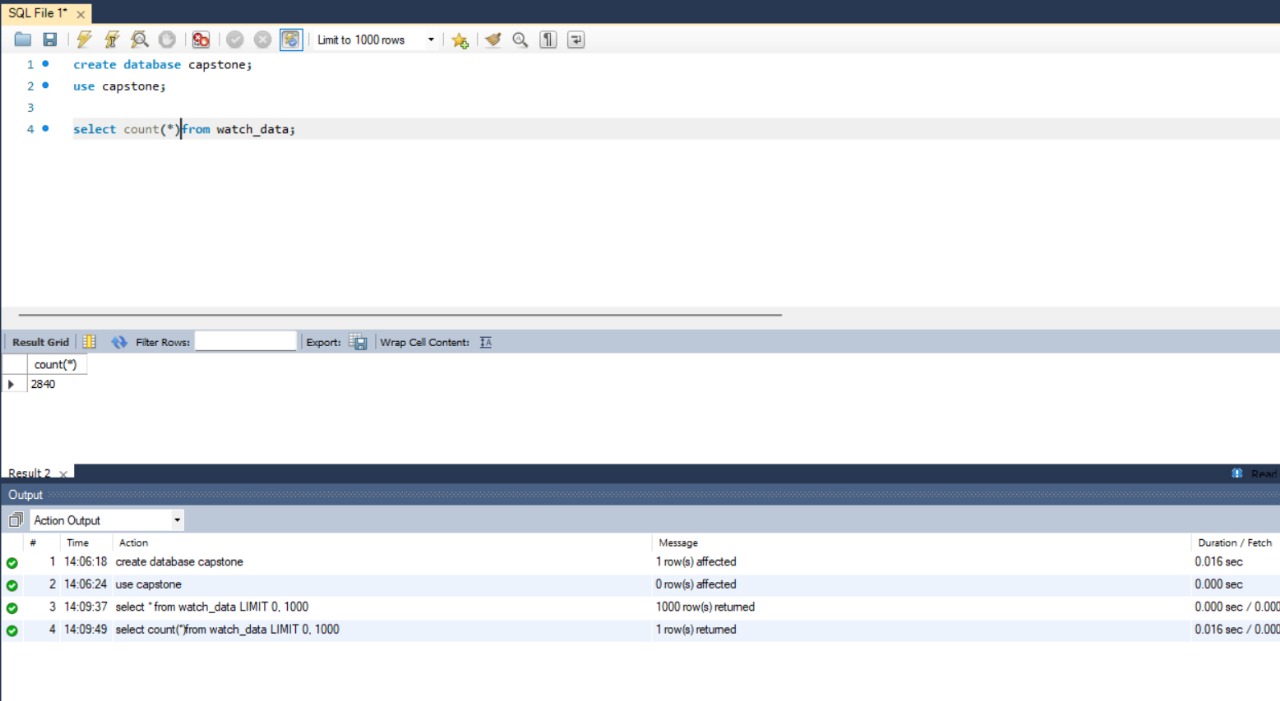


* **Correlation Analysis**: Heatmaps revealed relationships between price, reviews, and ratings. Although price was not strongly correlated with rating, review counts provided valuable insights into product visibility and demand.



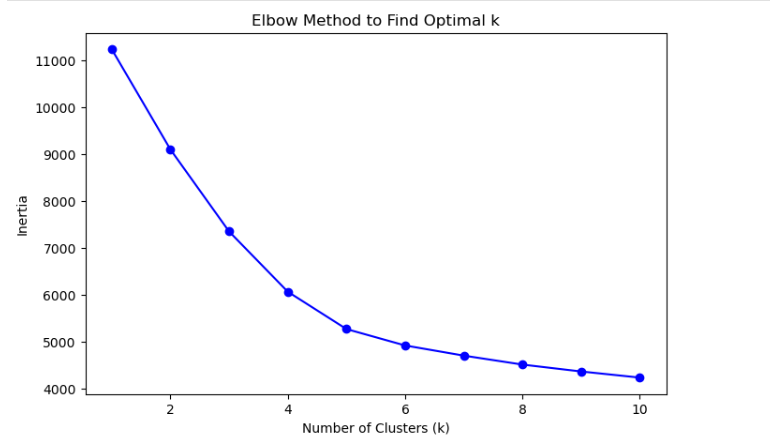
1. **Data Storage**

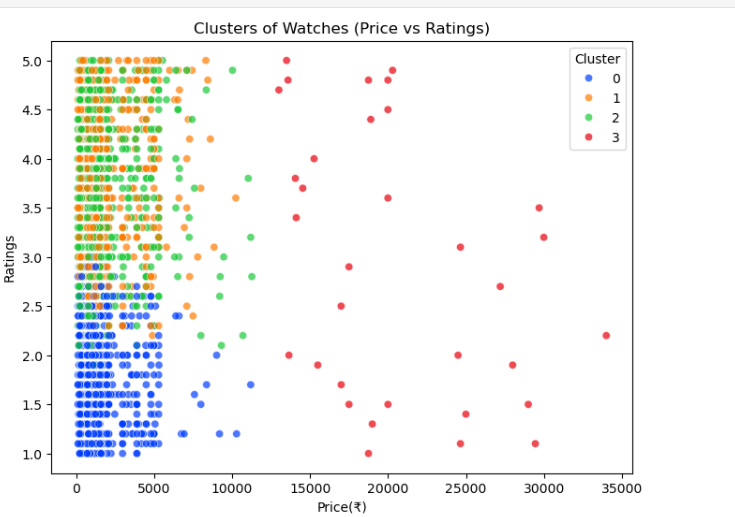
* The cleaned dataset is stored in a relational database for structured querying and efficient access.
* Tool **SQLAlchemy** is used to manage the data.
* This step ensures that the dataset is well-organized, secure, and ready for further analysis.



1. **Unsupervised Learning – KMeans Clustering**

* Clustering was applied to segment watches into groups based on their features.
* Using the Elbow Method, the point where the reduction in inertia slowed down indicated the optimal number of clusters.
* The Silhouette Score further validated this choice by measuring the quality of cluster separation.
* The final clusters revealed distinct product groups such as budget watches, mid-range popular watches, and premium watches.
* These clusters not only enhanced the interpretability of the dataset but also provided a new feature for supervised classification.



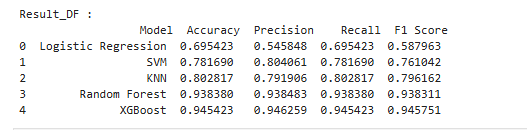


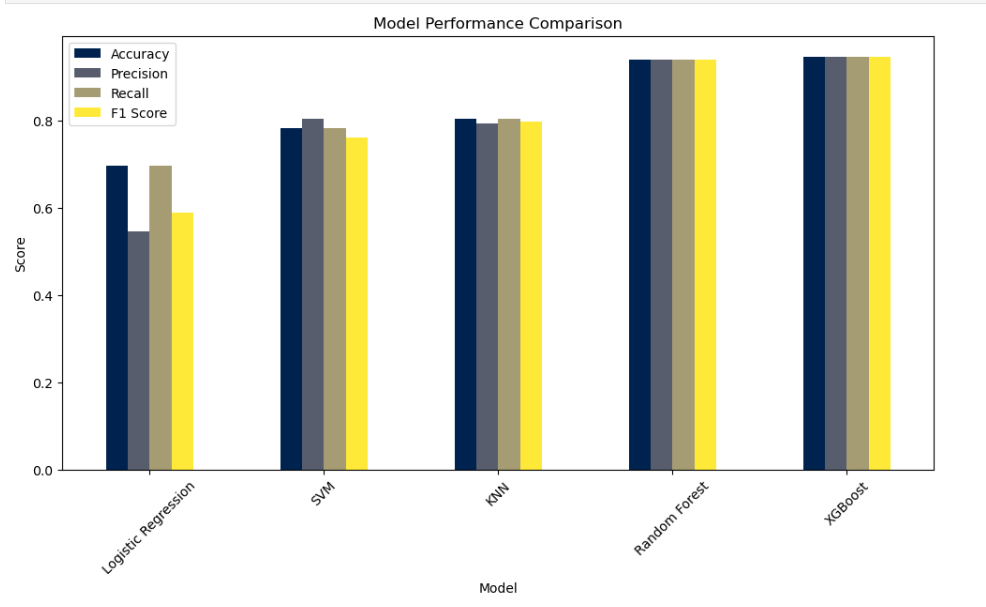
1. **Model Comparison**

For Supervised Learning, several machine learning models were trained and evaluated to predict the outcomes of IPL matches.

* **XGBoost** achieved the **highest performance** overall (Accuracy: 94.5%, F1-Score: 0.945), making it the best model for this dataset.
* **Random Forest** closely followed (Accuracy: 94.3%, F1-Score: 0.944), showing that ensemble techniques consistently outperform simpler models.
* **KNN and SVM** showed moderate results (Accuracy ~78–80%), indicating they can capture patterns but are less robust for this dataset.
* **Logistic Regression** had the lowest performance (Accuracy: 69.5%, F1-Score: 0.588), highlighting its limitations in modeling complex relationships.

**Overall Insight:** Ensemble models (**XGBoost, Random Forest**) are the most reliable and accurate, while Logistic Regression is the weakest performer for this classification task.





1. **Insights & Results**

* Watches could be effectively segmented into 3–4 groups using clustering, which reflected meaningful categories in terms of pricing and popularity.
* XGBoost achieved the highest predictive performance, outperforming Logistic Regression, SVM, and KNN.
* Review counts and ratings were key indicators of product perception, while price remained the primary driver of classification.
* EDA confirmed that customer trust (ratings + reviews) has a significant role in product differentiation on e-commerce platforms.

1. **Conclusion**

* This project successfully demonstrated the process of converting raw web-scraped data into meaningful insights using data science techniques.
* By following the full pipeline—data collection, cleaning, exploratory analysis, clustering, and supervised modeling—the project showcased both technical implementation and practical relevance.
* The combination of KMeans clustering for segmentation and Random Forest for classification provided a robust approach to analyzing the Flipkart watch dataset.
* Such a framework could be extended to other product categories, helping e-commerce businesses optimize pricing, improve recommendations, and better understand customer preferences.