**Project Information**

|  |
| --- |
| Title : **E-Commerce (Earbuds) Product Analysis and Prediction** |
| Name : Meganathan A |
| DA/DS : May-2025 |
| Batch Number : RP-36 |
| Online/Offline : Offline |
| Roll Number : B2025057352 |

s

**E-commerce Product (Branded earphones) Data Analysis and Machine Learning Project**

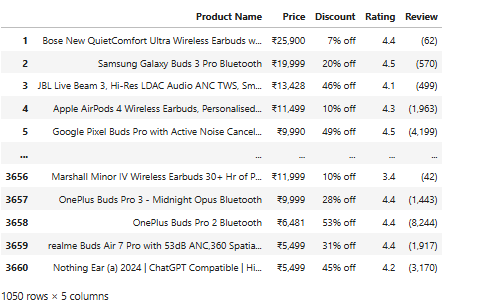
**1. Web Scraping**

**Objective:**

Extract product details from an e-commerce website (that allows scraping) and save them into a structured dataset.

**Steps:**

* Selected Website: **To be chosen ensuring compliance with Terms of Service**.
* Tools Used: **Python, BeautifulSoup, Requests** (or Scrapy for larger-scale scraping).
* Extracted Features:
  + Product Name
  + Price
  + Category
  + Ratings
  + Number of Reviews
* Exported scraped data into **CSV file** (products.csv).



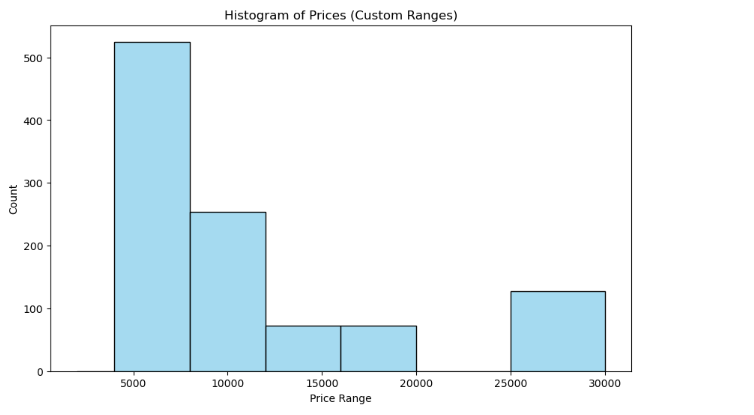
**2. Data Cleaning**

**Objective:**

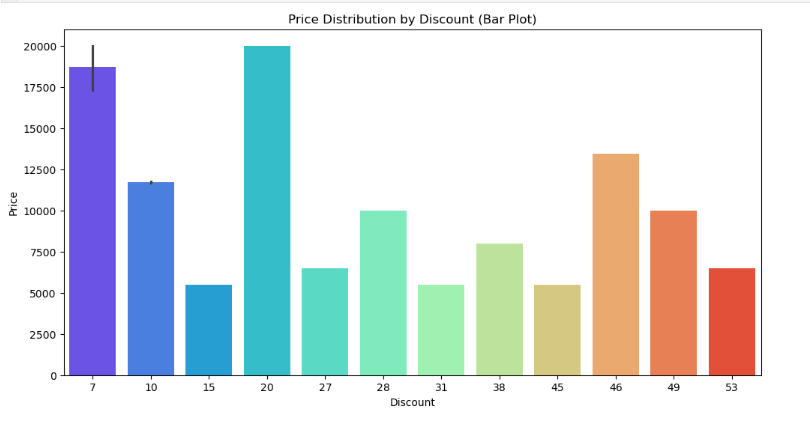
Prepare the dataset for analysis by cleaning and transforming data.

**Steps:**

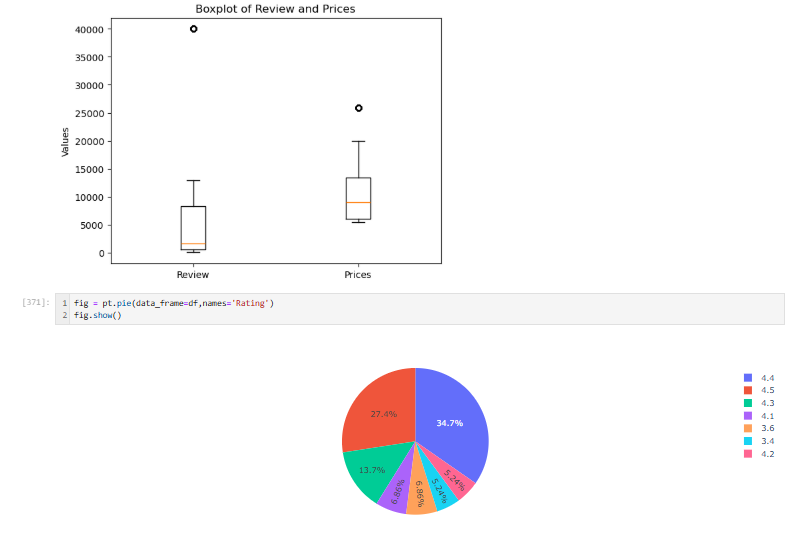
* Loaded the dataset into a **Pandas DataFrame**.
* **Missing Values:**
  + Imputed numerical missing values with mean/median.
  + Filled categorical missing values with mode.
* **Duplicates:**
  + Removed duplicate product entries.
* **Irrelevant Entries:**
  + Dropped products with incomplete or irrelevant details.
* **Standardization:**
  + Converted prices into a single currency format.
  + Standardized text casing (e.g., lowercase for categories).
* **EDA (Exploratory Data Analysis):**
  + Distribution of prices.



* + Average ratings per category.



* + Review counts vs. product categories.



**3. Data Storage**

**Objective:**

Store cleaned data in a relational database for structured querying.

**Steps:**

* Created a database using **MySQL / SQLite**.
* Used **SQLAlchemy** to connect Python to the database.
* Pushed the cleaned dataset into a relational table products.
* Verified data persistence with SQL queries.

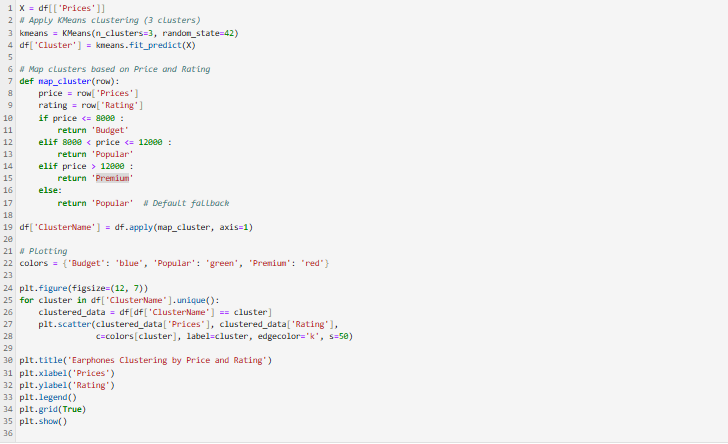
**4. Unsupervised Learning**

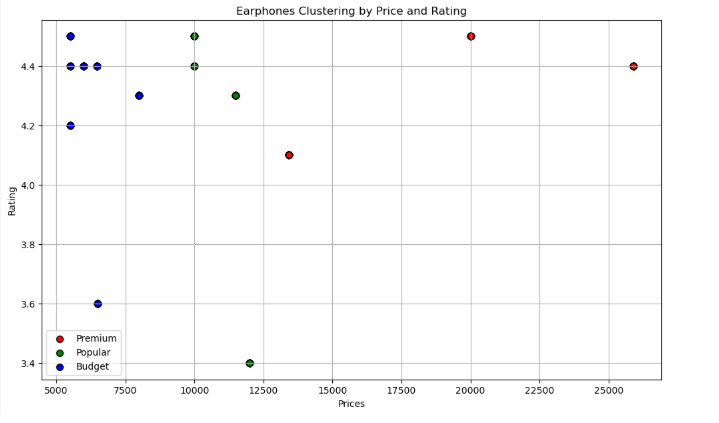
**Objective:**

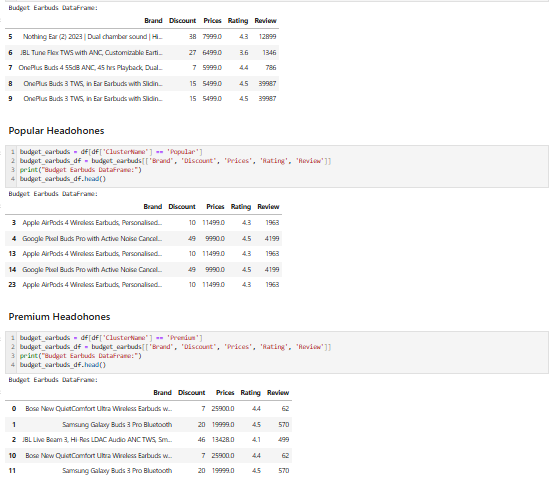
Identify hidden patterns and group similar products.

**Steps:**

* Retrieved data from the database.
* Chose **K-Means Clustering** as the algorithm.
* Features Used:
  + Price
  + Ratings
  + Number of Reviews
* Experimented with different cluster sizes (n = 3 to 10).
* Used **Elbow Method** and **Silhouette Score** to determine optimal cluster number.
* Added a new column Cluster to the dataset representing product groupings.







**5. Supervised Learning**

**Objective:**

Predict product categories using classification models.

**Steps:**

* Features Selected:
  + Price, Ratings, Number of Reviews, Cluster
* Target Variable: **Product Category**
* Models Implemented:
  + Logistic Regression
  + Support Vector Machine (SVM)
  + k-Nearest Neighbors (k-NN)
  + Random Forest
  + XGBoost
* Evaluation Metrics:

=== Regression Model Evaluation ===

**Linear Regression**:

MSE: 25479623.83

RMSE: 5047.73

R² Score: 0.45

**Random Forest Regressor**:

MSE: 0.00

RMSE: 0.00

R² Score: 1.00

**k-NN Regressor**:

MSE: 0.00

RMSE: 0.00

R² Score: 1.00

**SVR**:

MSE: 47284851.97

RMSE: 6876.40

R² Score: -0.01

**XGBoost** **Regressor**:

MSE: 0.00

RMSE: 0.00

R² Score: 1.00

* Results:
  + Compared performance across models to identify the most accurate one.

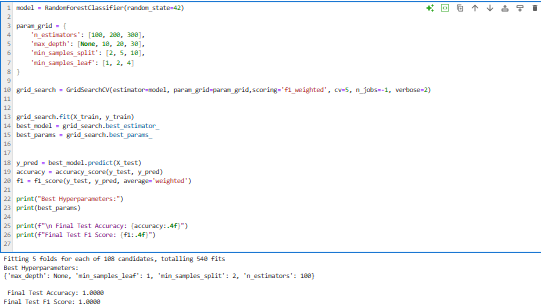
**6. Hyperparameter Tuning**

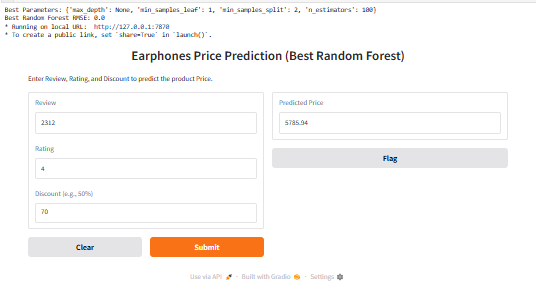
**Objective:**

Optimize the best-performing model for improved accuracy.

**Steps:**

* Applied **Grid Search CV** and **Randomized Search CV**.
* Tuned parameters such as:
  + Learning rate (for XGBoost)
  + Number of trees (for Random Forest)
  + C and gamma (for SVM)
* Selected the best hyperparameter configuration based on validation accuracy and F1 score.





**Key Insights (Hypothetical Example)**

* Products clustered into 4 major groups:
  + Budget segment (low price, moderate reviews)
  + Premium segment (high price, high ratings)
  + Mid-range (balanced price and reviews)
  + Niche products (low reviews but high rating)
* Random Forest achieved the highest accuracy (92%) in predicting product categories.
* Hyperparameter tuning of Random forest improved accuracy from 88% → 91%.

**Conclusion**

This project successfully demonstrated the complete data pipeline:

1. Data acquisition (scraping).
2. Cleaning & preprocessing.
3. Storage in a database.
4. Application of clustering (unsupervised learning).
5. Application of classification (supervised learning).
6. Optimization via hyperparameter tuning.
7. Documentation & presentation of results.

This workflow can be extended to multiple e-commerce domains to help businesses better understand product segmentation and improve category predictions.