Project Objective

The objective of this project is to analyze an e-commerce product dataset to:

- Understand **product popularity**, **pricing**, and **value-for-money patterns**
- Identify **top-performing brands, categories, and product series**
- Build a **machine learning model** to predict whether a product is *popular* based on its attributes
- Present insights through an **interactive Power BI dashboard** for business decision-making

---

Data Collection Method

- Data of category laptops was scraped from an **open and scraping-friendly e-commerce test website** (webscraper.io)
- Python libraries used:
    - requests and BeautifulSoup for data extraction
    - pandas for data storage and manipulation
- Extracted fields include:
    - Product name
    - Price
    - Rating
    - Number of reviews

The scraped data was stored in a pandas DataFrame and later exported as CSV for dashboarding after performing the Exploratory Data Analysis in python.

Data Cleaning & Feature Engineering

Several cleaning and preprocessing steps were performed to make the data analysis-ready:

Cleaning Steps

- Removed duplicate rows, special characters and inconsistent formatting from product names
- Parsed product names to extract:
    - **Brand**
    - **Series**
- Handled missing and inconsistent category labels
- Standardized category names (Consumer, Business, Gaming, Premium, Convertible)

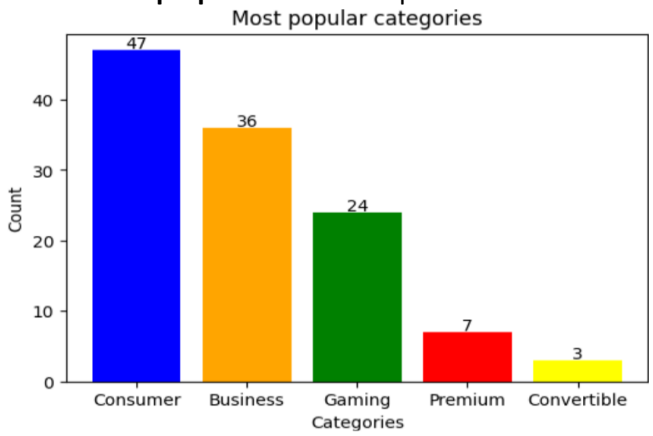
Feature Engineering

- **Popularity (Target Variable):**
    - Defined using review count (products with higher reviews classified as popular)
- **Value Score:**
- value_score = rating / price
- **Scaled Value Score:**
    - Applied MinMaxScaler for better interpretability and modeling
- Created derived fields such as:
    - Product category
    - Brand-series grouping

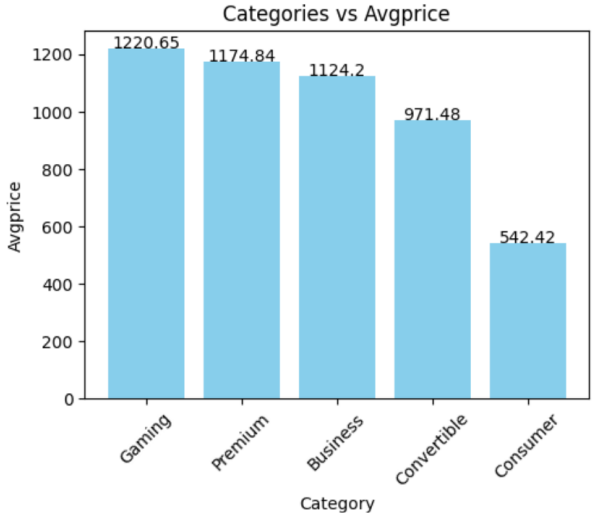
Exploratory Data Analysis (EDA) – Key Findings

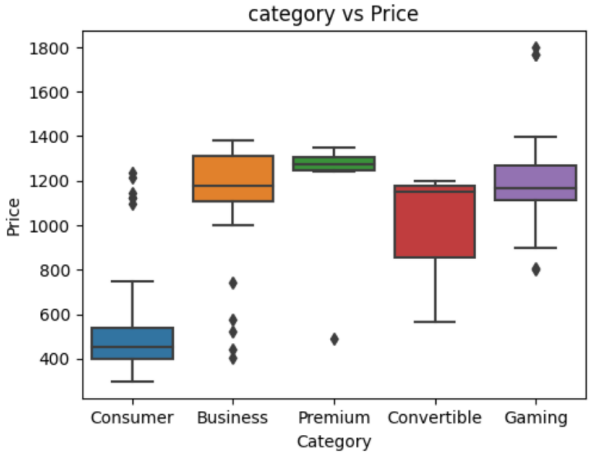
➢ **Category Insights**

- **Consumer laptops** dominate the platform

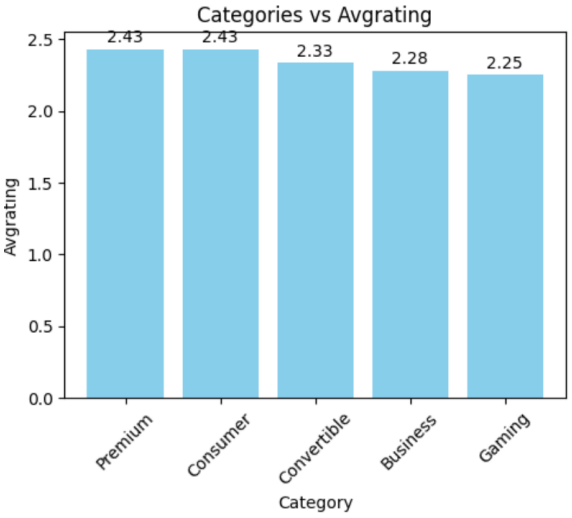


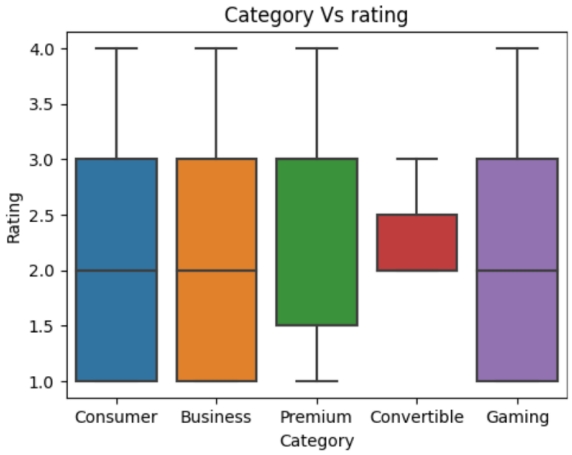
- Gaming and Premium categories have fewer listings but higher prices



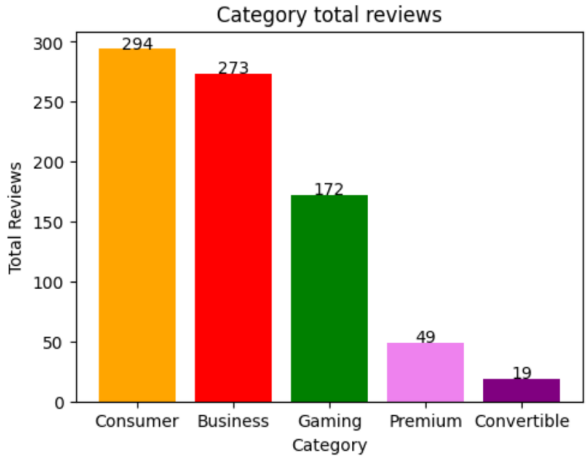


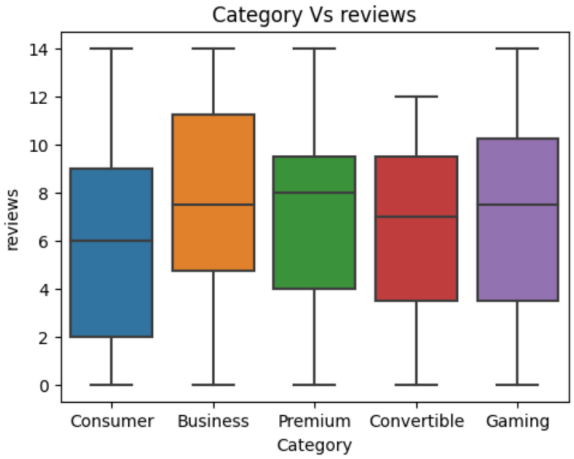
- Despite similar average ratings across categories, the box plot reveals that Convertible products have the tightest and most consistent customer satisfaction range.



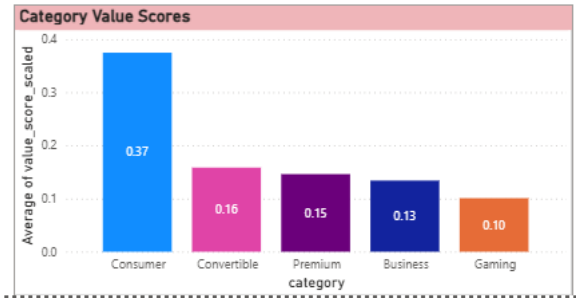


- Consumer segment has got highest engagement. Box plot suggests that Business laptops demonstrate high and consistent customer engagement, rivaling Gaming in median reviews but outperforming it in reliability and spread.



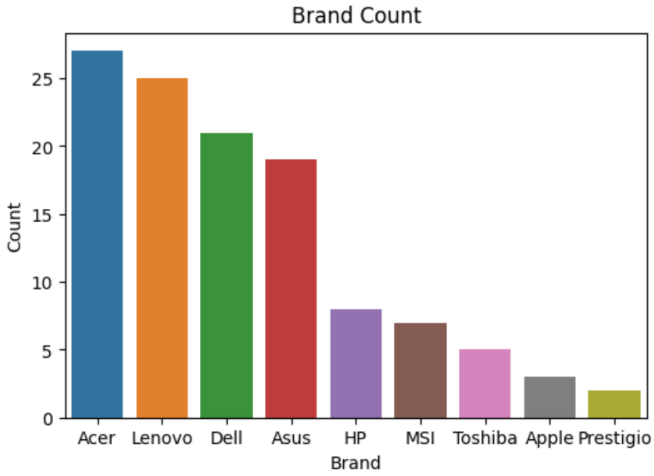


- Consumer category shows strong value-for-money positioning.

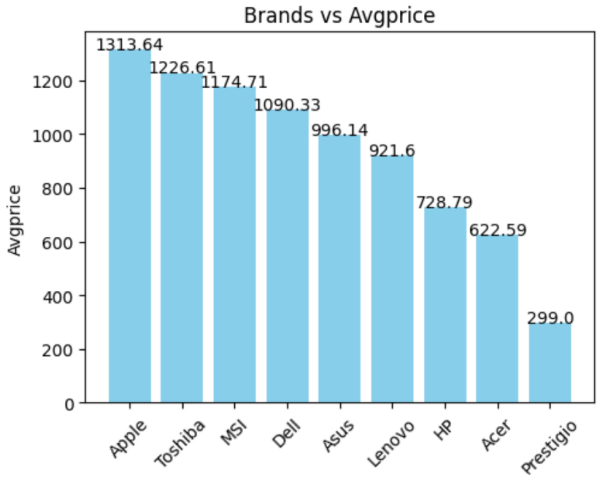


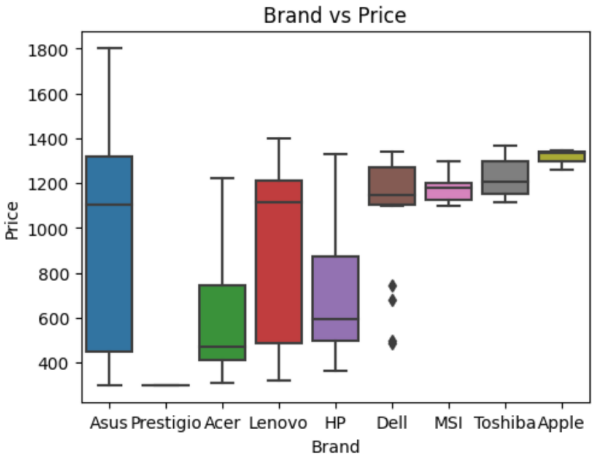
➢ **Brand Insights**

- Acer, Lenovo, and Dell have the **highest product presence**

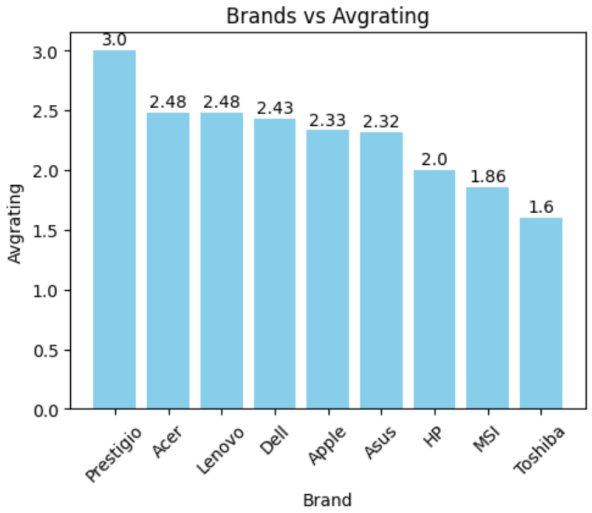


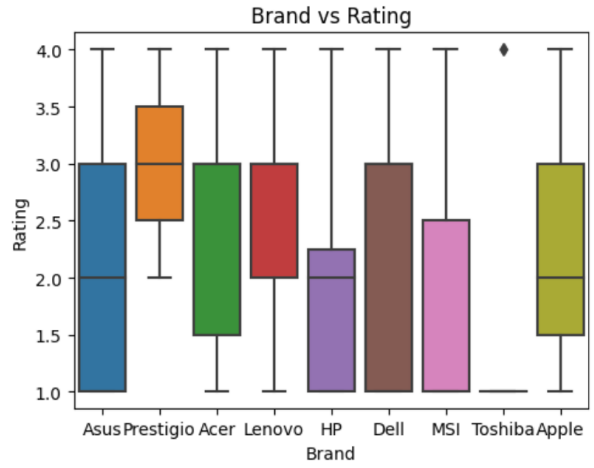
- Apple stands out with the highest median price around 1300 indicating a premium pricing strategy and consistent high-end positioning. Prestigio has the lowest price range centered around 400 making it the most affordable brand in the comparison. Asus spans from 400 to $1800 suggesting a diverse product lineup—from entry-level to high-performance models.



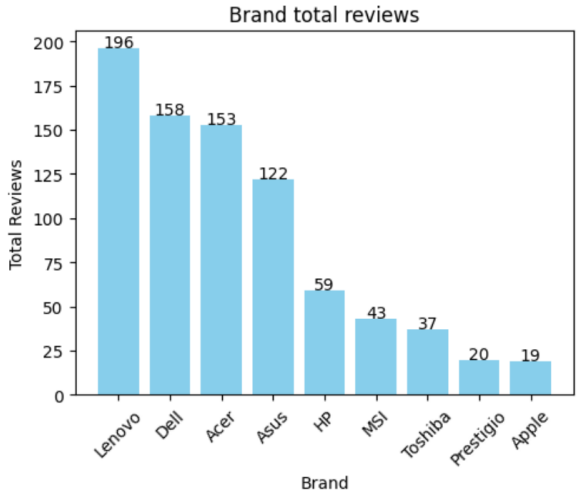


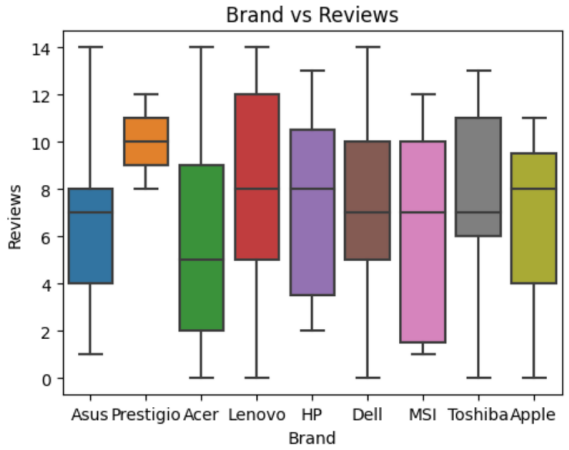
- Prestigio Leads in Customer Satisfaction Despite Low Pricing. It shows the highest median rating among all brands, even though it typically operates in the lowest price segment. This suggests that affordability doesn't compromise perceived quality — users are highly satisfied with Prestigio products relative to their cost.



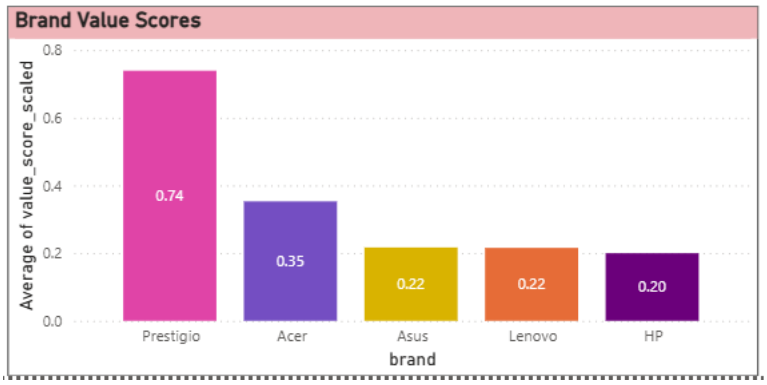


- Lenovo leads in **total reviews**, indicating strong engagement. Boxplot suggests Prestigio has a high median review count with a tight interquartile range, meaning most of its products consistently receive more reviews than other brands.





- Prestigio shows strong **value-for-money positioning**



- **Price, Rating & Reviews**
  - Higher-priced products do **not always** receive higher ratings
  - Popularity is more strongly linked to **review volume** than rating alone
  - Value_score varies significantly across brands and categories
- **Correlation Analysis**
  - Weak correlation between price and rating
  - Weak correlation between price and reviews
  - Moderate relationship between reviews and popularity

Model Performance

- **Problem Type**
  - **Binary Classification**: Predict whether a product is *popular* or *not popular*
- **Models Evaluated**
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Gradient Boosting
- **Evaluation Strategy**
  - Initial Train–Test split
  - Improved using **Stratified K-Fold Cross-Validation (5 folds)**
  - Metric used: **F1 Score** (balanced precision and recall)
- **Optimization Process**
  - GridSearchCV with 5-fold stratified cross-validation was applied to the Random Forest classifier to explore optimal hyperparameter combinations.
  - The tuned model achieved comparable F1 performance to the baseline model but did not result in a significant improvement, indicating that the original configuration was already well-generalized.

- This outcome highlights the stability of the baseline model and confirms that additional model complexity did not yield meaningful performance gains for the given dataset and target formulation.

➢ **Final Cross-Validated Results**

Model	Mean F1 Score
Decision Tree	**0.65**
Random Forest	0.60
Gradient Boosting	0.59
Logistic Regression	0.57

➢ **Interpretation**

- Moderate performance is expected due to:
  - Small dataset size
  - Popularity being derived from reviews (limited signal)
- Cross-validation provided a **more reliable estimate** than a single train-test split

---

Key Business Insights

- High product count does **not guarantee popularity**
- Brands with competitive pricing often achieve:
  - Higher value_score
  - Better popularity
- Popularity is influenced more by:
  - Product visibility and engagement
  - Price accessibility
- Premium-priced products require strong branding or differentiation to succeed

---

Recommendations

➢ **For Business Teams**

- Focus on **value-for-money positioning** rather than price alone
- Increase visibility for mid-priced, high-rating products
- Optimize pricing strategies by category

➢ **For Data & Product Teams**

- Collect additional features such as:
  - Specifications
  - Discounts
  - Seller ratings
- Use larger datasets for improved model performance
- Consider regression models for price optimization in future work

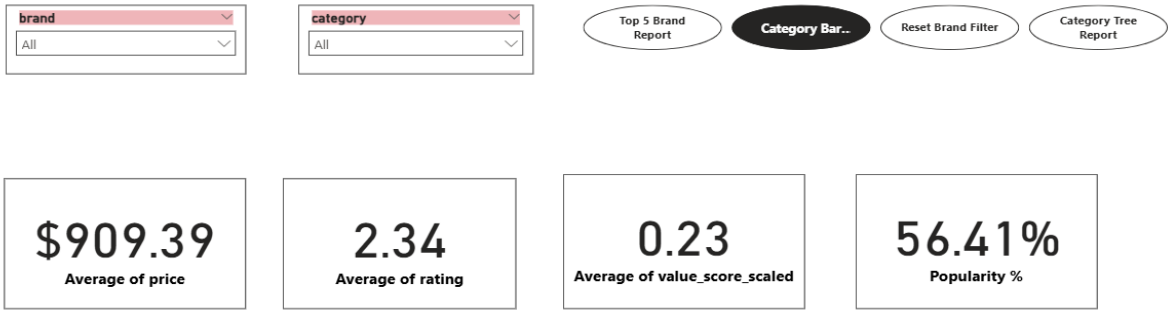
---

WebScraping Code

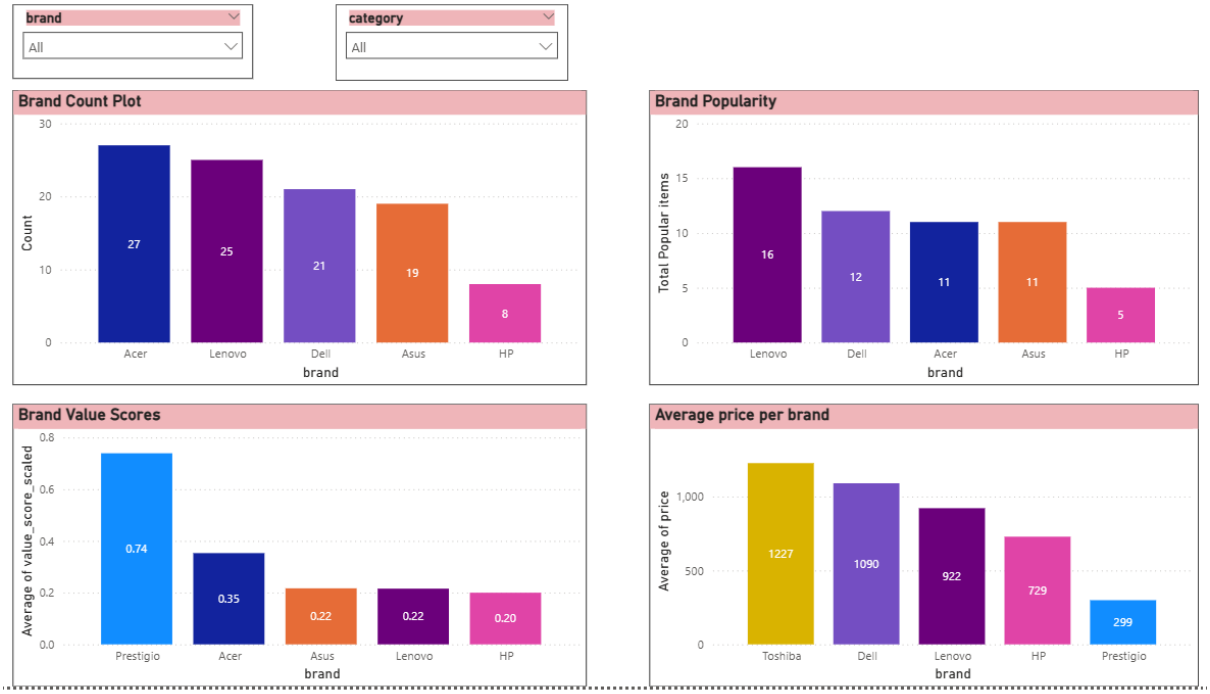
```
url = "https://webscraper.io/test-sites/e-commerce/static/computers/laptops"
response = requests.get(url)
soup = BeautifulSoup(response.text, "html.parser")  # converting raw html data into structured format
products = soup.find_all("div", class_="thumbnail")
base_url = "https://webscraper.io/test-sites/e-commerce/static/computers/laptops?page="

data = []

for page in range(1,21):
    url = base_url + str(page)
    response = requests.get(url)
    soup = BeautifulSoup(response.text, "html.parser")  # converting raw html data into structured format
    products = soup.find_all("div", class_="thumbnail")

    if not products:
        break

    for product in products:
        # extracting product name
        #name = product.find("a", class_="title").text.strip()
        product_tag = product.find("a", class_="title")
        name = product_tag["title"].strip()

        #extracting product price
        price = product.find("span", itemprop="price").text.strip().replace("$", "")

        #extracting rating
        rating_tag = product.find("p", attrs={"data-rating": True})
        rating = int(rating_tag["data-rating"]) if rating_tag else 0

        #extracting product reviews
        reviews = product.find("span", itemprop="reviewCount").text.strip()

        data.append([name, float(price), rating, int(reviews), "Laptop"])

df = pd.DataFrame(data, columns=["product_name", "price", "rating", "reviews", "category"])
```

Dashboard Screenshots

- **KPIS & Bookmark Links**



➢ Brand Report



➢ Brand Takeaways

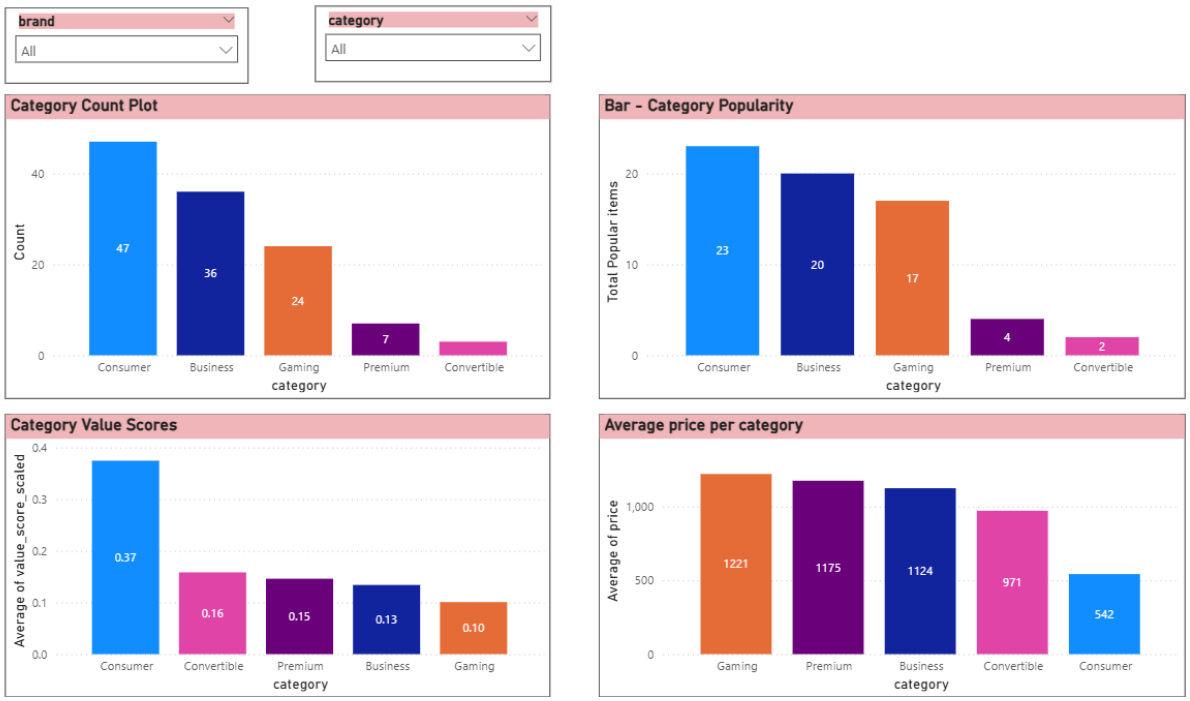
Observations:

1. **Market Presence:** Acer Leads the Pack. This suggests Acer is aggressively targeting multiple segments, possibly with a wide range of models and price points.
2. **Popularity:** Lenovo Captures Consumer Attention. Despite Acer's volume, **Lenovo** tops the popularity chart with 16 highly reviewed products.
3. **Value Perception:** Prestigio Surprises, though absent in the count/popularity charts, scores highest in value perception (0.74).
4. **Pricing Strategy:** Toshiba and Dell Aim High. These brands may be positioning themselves as premium or performance-oriented.

Recommendations:

- **Lenovo** balances volume and popularity well, making it a consumer favorite.
- **Acer** is a volume player with decent value perception, ideal for mass-market penetration.
- **Prestigio** is a sleeper hit—low price, high value, but needs visibility.
- **Dell** and **Toshiba** seem to target premium buyers, but Dell also maintains popularity.
- **HP** appears underrepresented across all metrics, suggesting a need to rethink its market approach.

➢ Category Report:



➢ Category Takeaway:

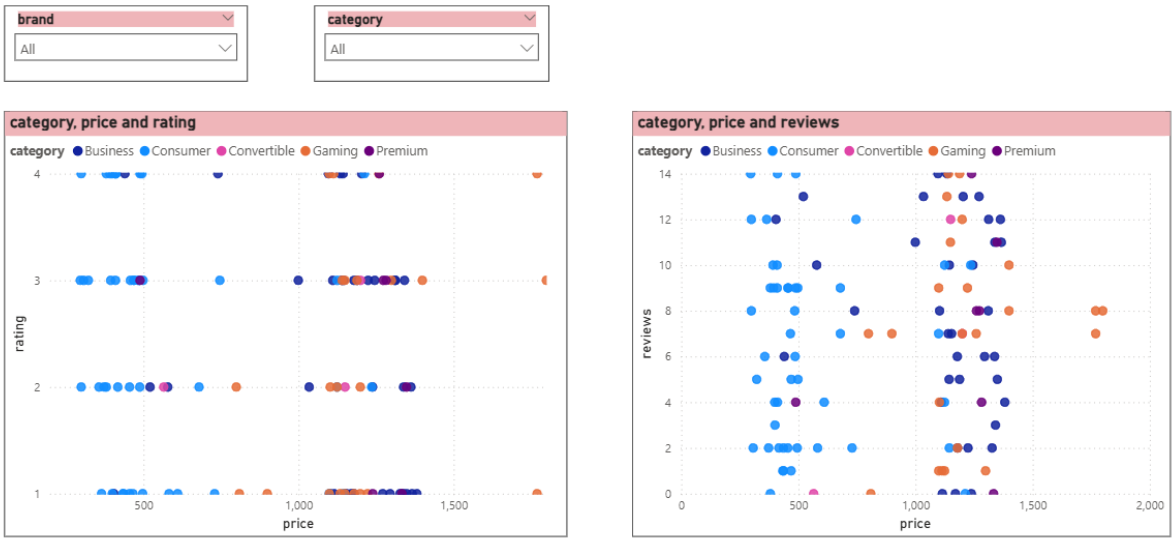
Observations:

1. **Consumer Category: The Mass-Market Leader**
   - With **47 products**, the Consumer category dominates in volume.
   - It also leads in **popularity** (23 popular items) and **value perception** (average score of 0.37).
   - Despite its strength, it has the **lowest average price** (₹542), suggesting high accessibility and strong value-for-money appeal.
2. **Business Category: High Stakes, Moderate Value**
   - The second-largest in volume (**36 products**) and popularity (**20 popular items**).
   - Although Business laptops have a relatively high average price (₹1124), their value score remains low (0.13) because the ratings do not scale proportionately with the cost—indicating that consumers perceive them as offering less value for money compared to other categories.
3. **Gaming Category: Premium Power, Niche Appeal**
   - With **24 products**, Gaming is a mid-sized category.
   - It scores well in popularity (**17 items**) but has the **lowest value score** (0.10).
   - It commands the **highest average price** (₹1221), reflecting its performance-driven nature.
4. **Premium & Convertible: Niche but Noteworthy**
   - **Premium** (7 products) and **Convertible** (3 products) are the smallest categories.
   - Both have modest popularity (4 and 2 items respectively) and low value scores (0.15 and 0.16).
   - Yet, they're priced high—₹1175 for Premium and ₹971 for Convertible.

Recommendations:

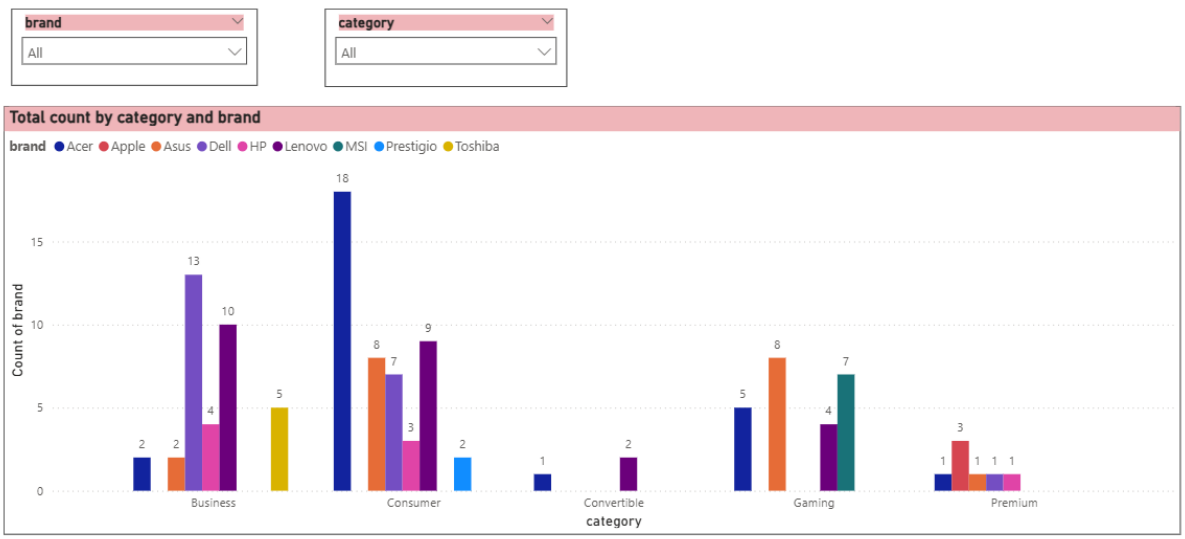
- **Consumer** is the sweet spot: high volume, high popularity, high value, and low price.
- **Business and Gaming** categories need to justify their pricing with better value perception.
- **Premium and Convertible** products should focus on clearer differentiation and consumer education to boost perceived value.
- Consider targeted marketing and feature enhancements to elevate underperforming categories.

➢ Correlation Plots:



The above scatter plots shows there is no correlation between price and rating, high price items got both high ratings as well as low ratings and viceversa
And also no correlation between price and reviews high price products got both high and low total reviews and the same way low price customers also got both low and high total reviews

➢ Clustered Column Chart:

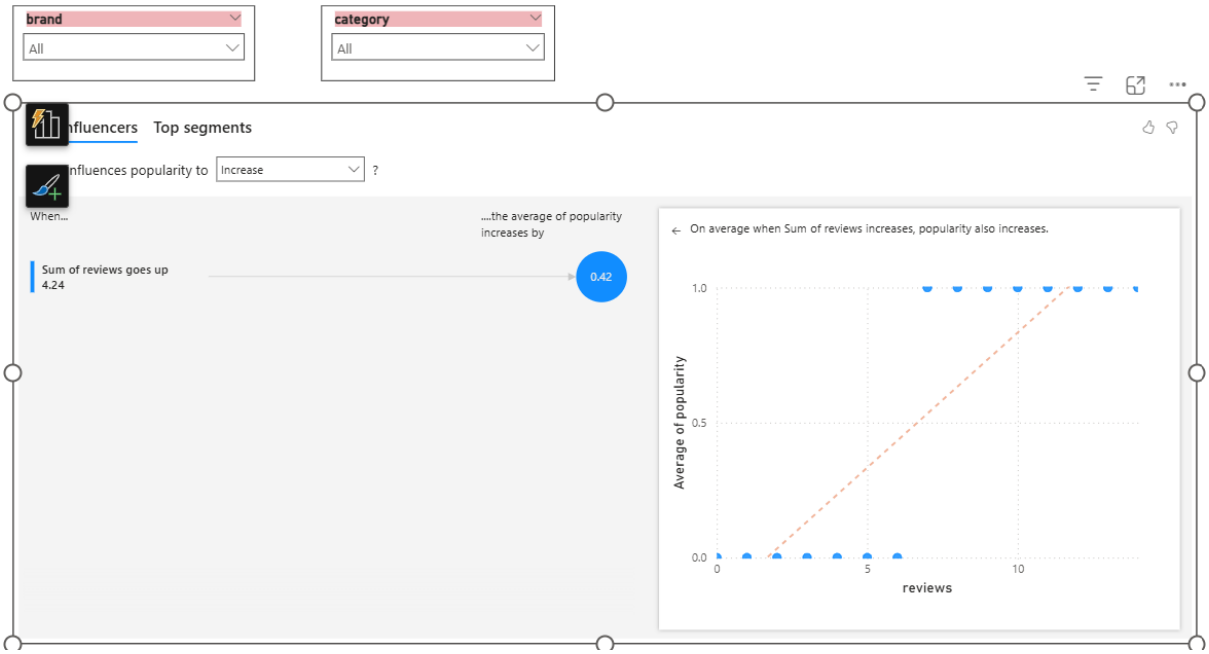


- **Dell emerges as the leader in the Business category, suggesting a strong foothold in enterprise solutions.**
- **Acer dominates the Consumer segment, indicating mass-market appeal and affordability.**
- **Asus and MSI shine in Gaming, pointing to performance-driven innovation.**

➢ KeyInfluencer AI Visual:

In this chart I have tried to analyze popularity based on the following fields

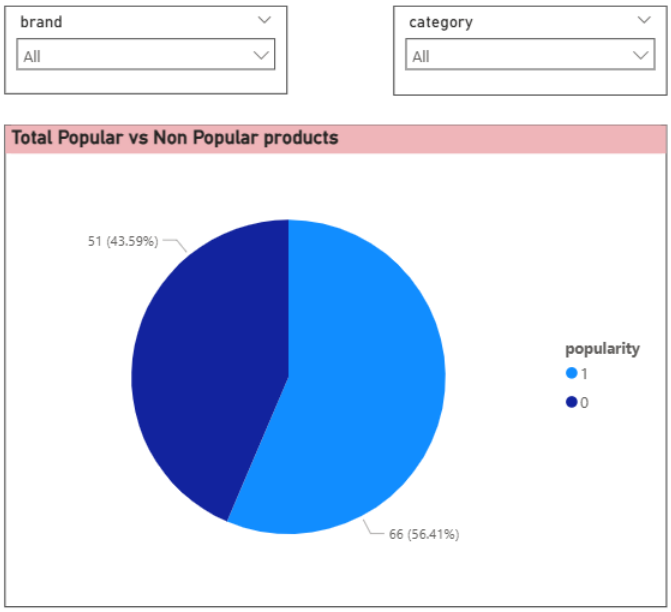
- Price
- Rating
- Value_Score
- Reviews



Among all the analyzed fields — Price, Rating, Value_Score, and Reviews — only Reviews show a meaningful relationship with Popularity. Specifically, when the sum of reviews increases by 4.24, the average popularity rises by 0.42. This suggests that higher customer engagement through reviews tends to boost perceived popularity.

➢ Decomposition Tree AI Visual:

➢ Bar Chart (ML Problem Chart)

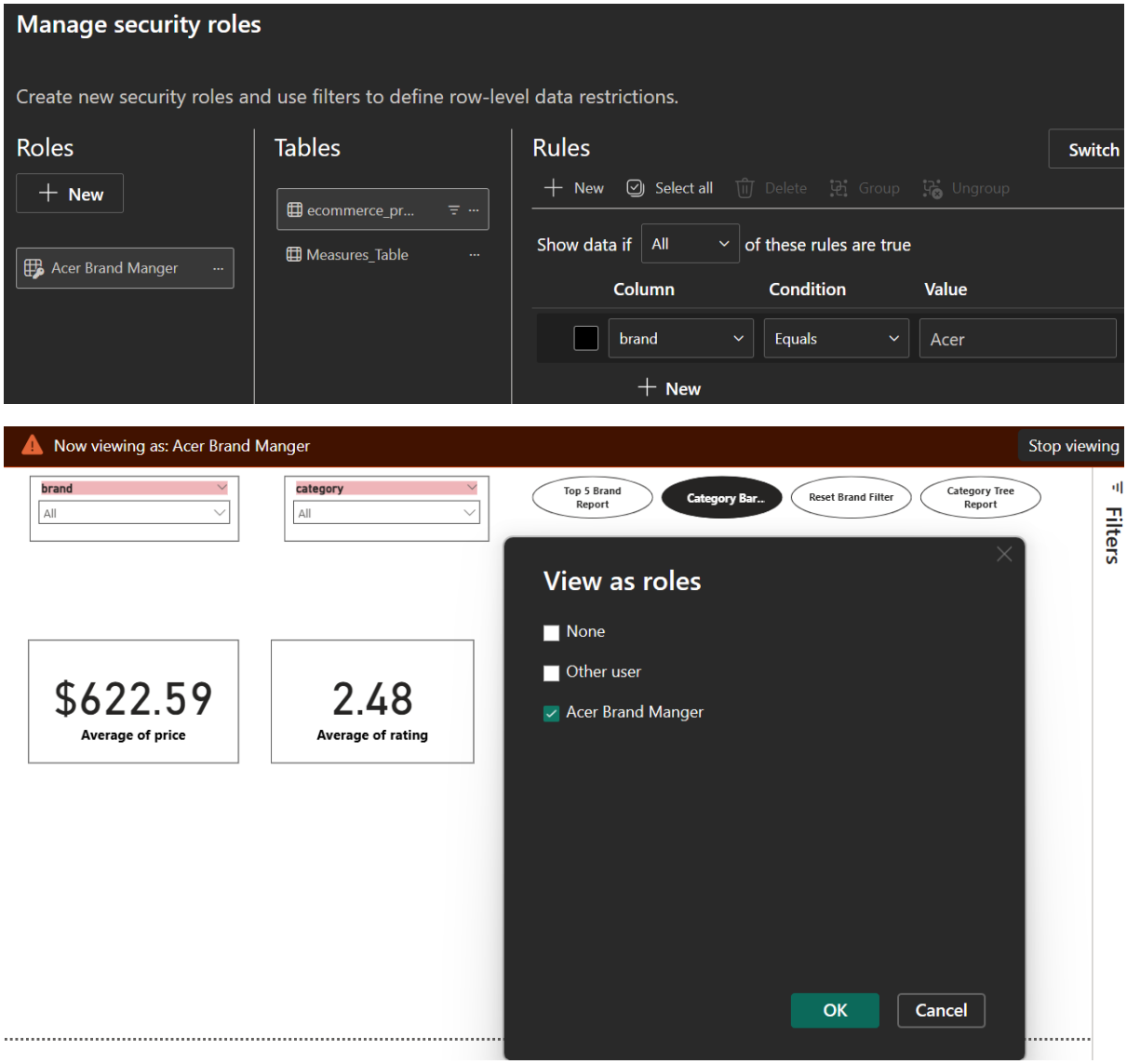


- Over half the products (56.41%) are considered popular, meaning they've attracted significant consumer attention.
- This suggests strong visibility, marketing, or brand loyalty driving review activity.
- The remaining 43.59% are less reviewed, possibly due to niche targeting, limited availability, or lack of awareness.

➢ Drill Through Page:

product_name	Average of price	Average of rating	Average of reviews	Average of value_score_scaled
Acer Aspire 3 A315-21	393.88	3.00	9.00	0.55
Acer Aspire 3 A315-31 Black	391.05	3.33	7.33	0.62
Acer Aspire 3 A315-51	436.29	1.00	1.00	0.13
Acer Aspire 3 A315-51 Black	494.71	4.00	2.00	0.59
Acer Aspire 7 A715-71G	1,123.87	2.00	4.00	0.09
Acer Aspire A315-31-C33J	379.94	2.00	0.00	0.37
Acer Aspire A315-51-33TG	457.38	3.00	9.00	0.47
Acer Aspire A515-51-5654	679.00	2.00	9.00	0.19
Acer Aspire ES1-572 Black	445.17	2.40	6.20	0.39
Acer Aspire ES1-732 Black	410.46	4.00	14.00	0.72
Acer Extensa 15 (2540) Black	439.73	4.00	6.00	0.67
Acer Nitro 5 AN515-51	974.81	2.00	7.00	0.11
Acer Predator Helios 300 (PH317-51)	1,177.81	2.00	8.00	0.09
Acer Spin 5	564.98	2.00	0.00	0.23
Acer Swift 1 SF113-31 Silver	488.64	3.00	4.00	0.44
Acer TravelMate P645-S-511A Black	1,170.10	1.00	0.00	0.02
Aspire E1-510	306.99	3.00	2.00	0.72
Aspire E1-572G	581.99	1.00	2.00	0.09
Total	**622.59**	**2.48**	**5.67**	**0.35**

➢ Role Manager:



Project Video Link

- **Webscraping and EDA (Video Part 1):**

https://www.loom.com/share/702ebfaa0b1348afb1467b819bcbe6e2

- **Modeling, Interactive Dashboard and Project report (Video Part 2):**

https://www.loom.com/share/a7025b95781b489d86d123e09cfd1c5b