CAPSTONE PROJECT

SUDIKSHA ASLESHA

OVERVIEW

PROJECT OBJECTIVE

WEBSCRAPING & DATA PREPROCESSING

EXPLORATORY DATA ANALYSIS

MODEL BUILDING

BUSINESS INSIGHTS

RECOMMENDATIONS

PROJECT OBJECTIVE

- **Web-Scraping:** Scrape product data such as **price, MRP, discounts, ratings, and reviews** from Snapdeal
- **Data Cleaning & Preprocessing:** Perform **data cleaning and preprocessing** to handle missing values, outliers, and inconsistent formats
- **Exploratory Data Analysis:** Conduct **exploratory data analysis (EDA)** to identify key trends, patterns, and relationships
- Apply **feature engineering** to create meaningful variables that improve model performance
- **Machine Learning Model:** To build and evaluate machine learning models to predict product prices or performance using features like **ratings, reviews, discounts, and MRP,** and identify **key factors influencing pricing** on Snapdeal.
- **Business Objective:** To convert analytical and model insights into actionable business recommendations for optimizing pricing, discounts, and product positioning.

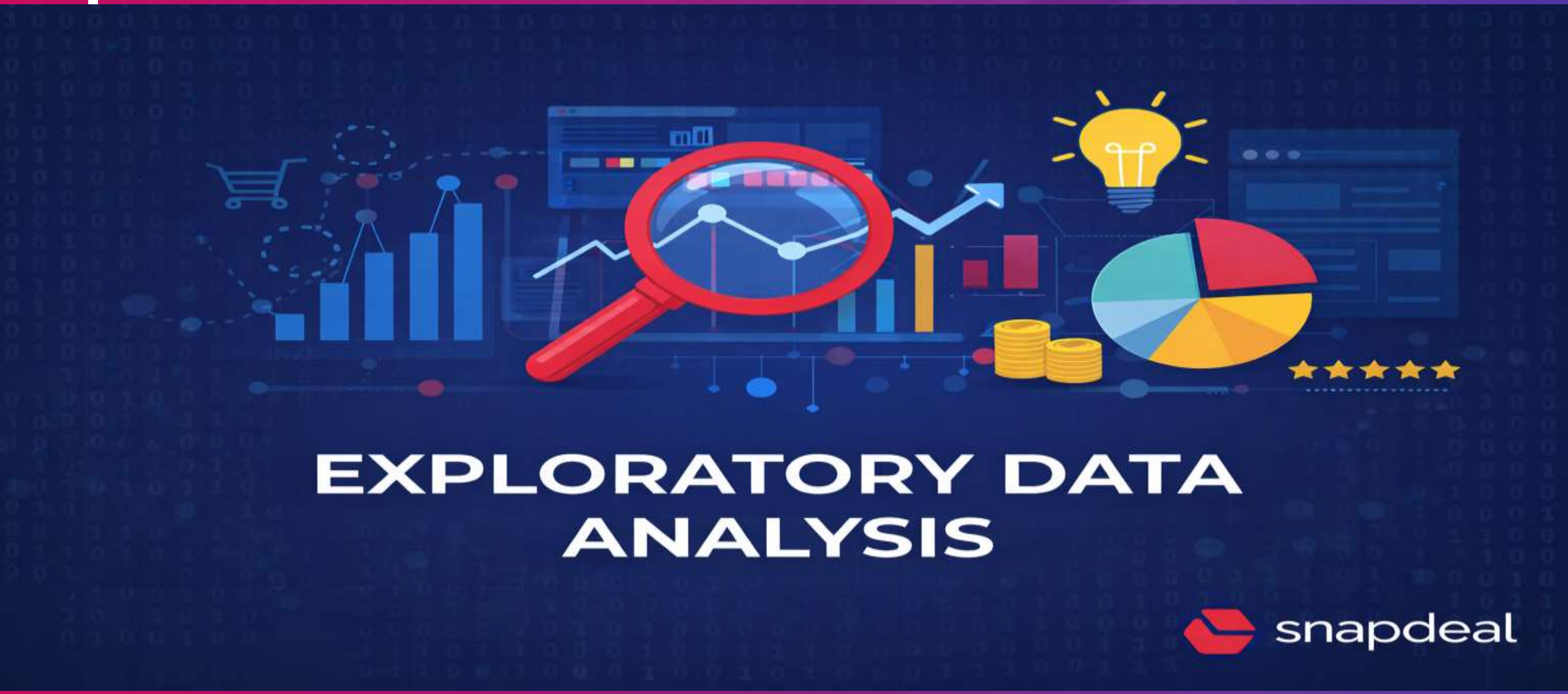
DATA COLLECTION & PREPROCESSING

Web Scraping Method

- Identified relevant Snapdeal product listing and detail pages
- Used Python libraries (Requests, BeautifulSoup) to fetch and parse HTML content
- Extracted key product details such as price, MRP, discount, ratings, and reviews
- Handled pagination to scrape multiple product pages
- Cleaned and stored scraped data in structured format (CSV/DataFrame)
- Ensured data consistency and avoided duplicate entries

Data Cleaning Method

- After collecting raw product data from Snapdeal, systematic data cleaning and preprocessing steps were performed to ensure data accuracy, consistency.
- Missing values in key fields, including ratings, reviews were addressed using appropriate techniques such as removal or statistical imputation.
- Converted data types (price, MRP, discount) into numerical format
- Removed special characters (₹, %, commas) for accurate analysis
- Discount values were standardized into percentage format, and rating values were validated to ensure they fall within acceptable limits.
- Prepared a clean and reliable dataset for EDA and model building

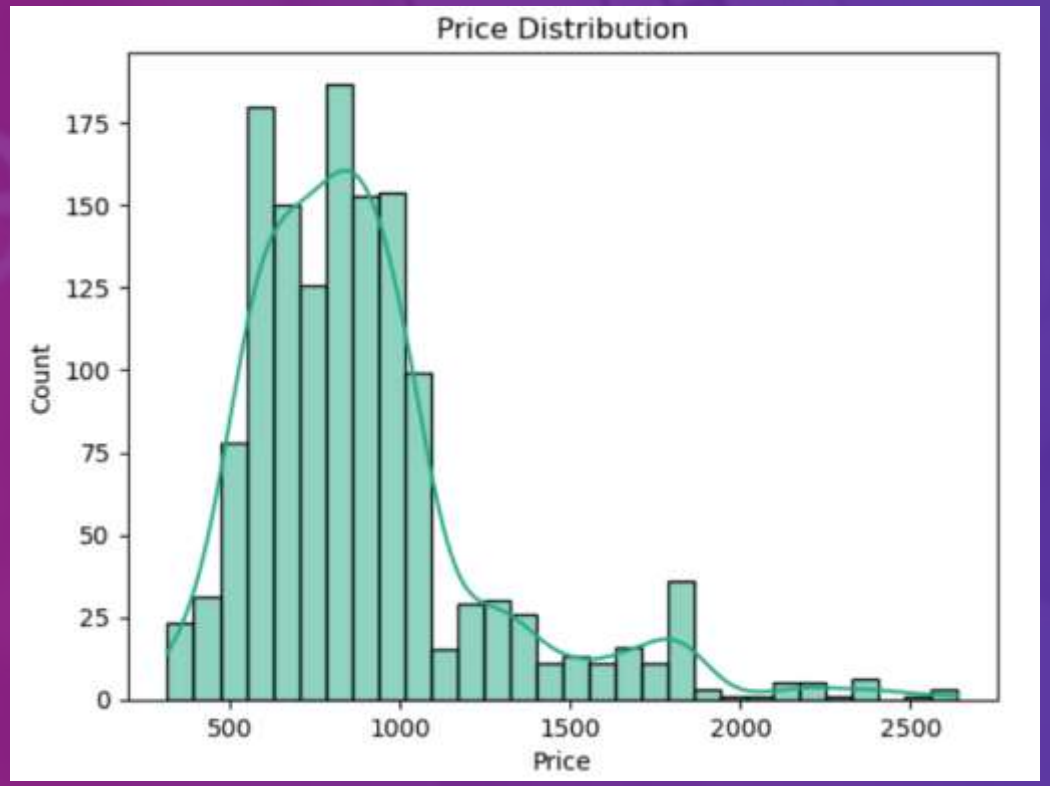

EXPLORATORY DATA
ANALYSIS
snapdeal

PRICING ANALYSIS

INSIGHTS:

- Majority of products are priced in the **mid-range segment.**
- The distribution is **right-skewed**, indicating fewer high-priced products
- Most products fall within a **common price band**, showing pricing consistency
- A small number of **premium-priced products** exist at the higher end
- Extremely high prices act as **outliers** and may influence model performance

Price Distribution

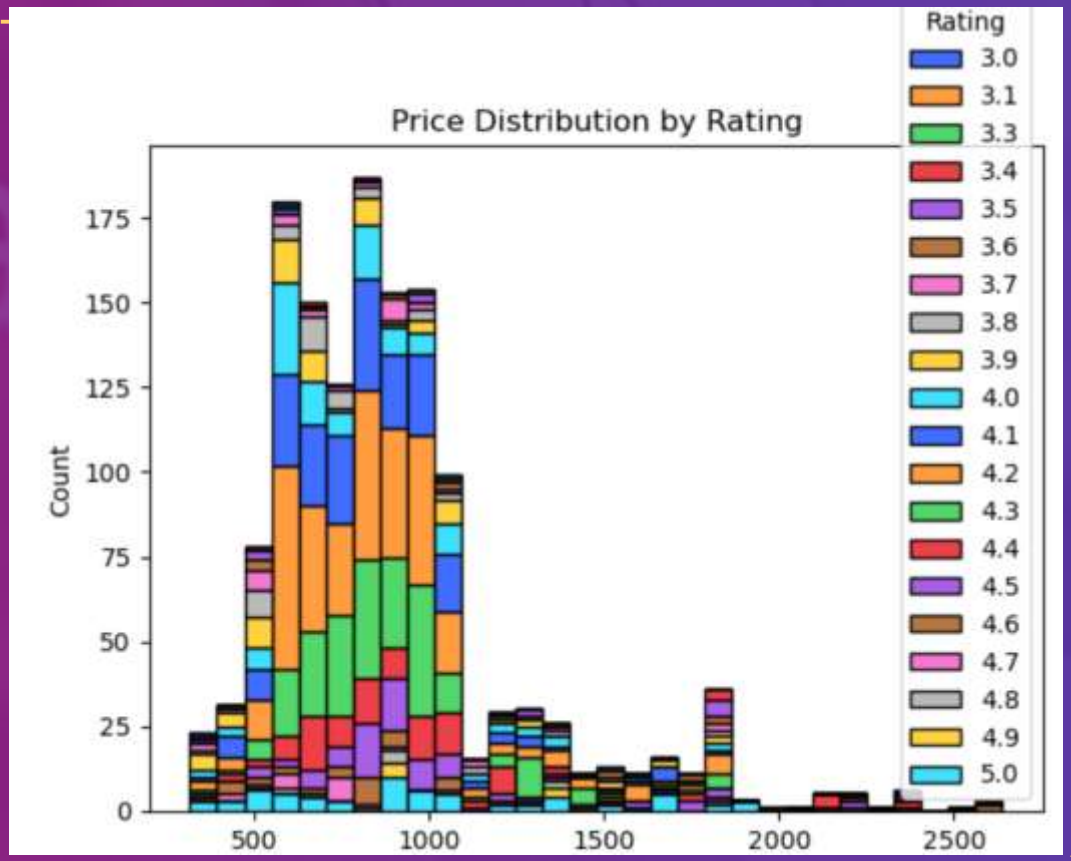


PRICING ANALYSIS

INSIGHTS:

- Most products across all ratings are concentrated in the mid-price range
- Higher-rated products (4.0 and above) are more frequent in the mid to higher price bands
- Lower-rated products are mostly concentrated in the lower price segments
- Premium-priced products tend to have moderate to high ratings, indicating perceived quality

Price Distribution by Rating

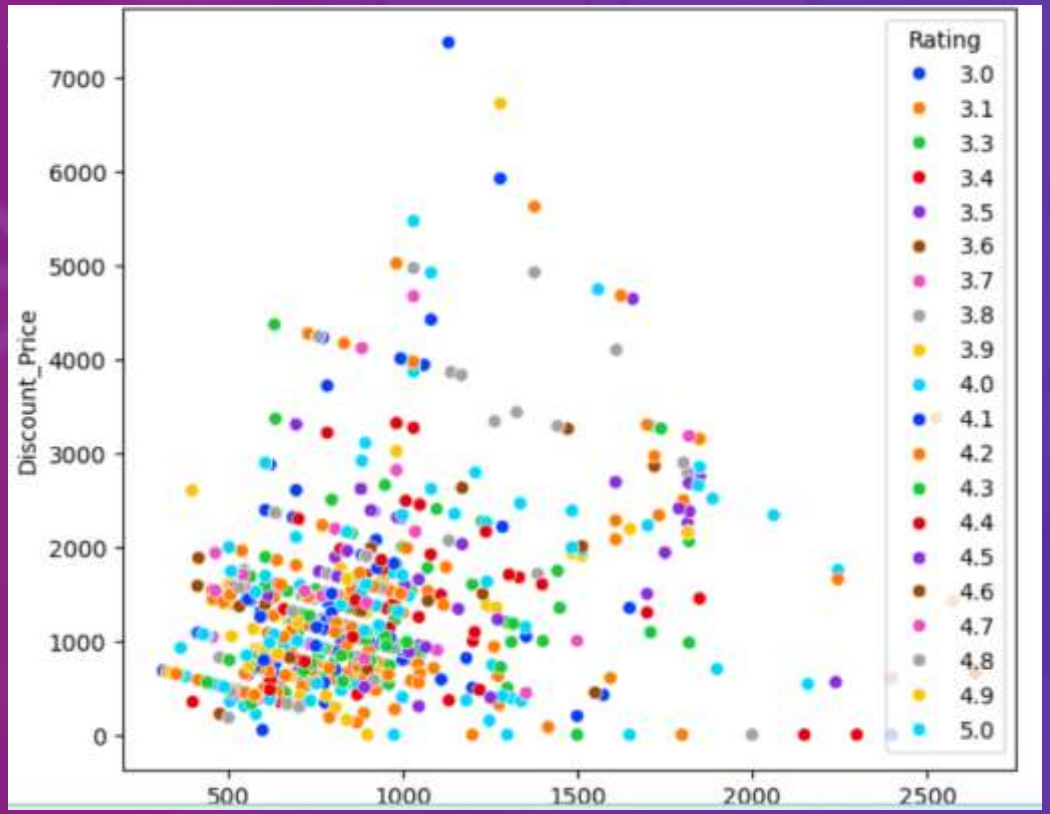


DISCOUNT ANALYSIS

INSIGHTS:

- Most products cluster in the **lower to mid-price range**, indicating high product availability in this segment
- **Higher discounts** are more common for **mid-priced products** rather than very low or very high prices
- Products with **higher ratings (4.0+)** are spread across price ranges, suggesting rating is not solely price-driven
- Premium-priced products tend to have **moderate discounts**, likely to maintain brand value
- The scattered pattern indicates **no strong linear relationship** between price and discounted price, highlighting the role of other factors

Price Vs Discount_Price

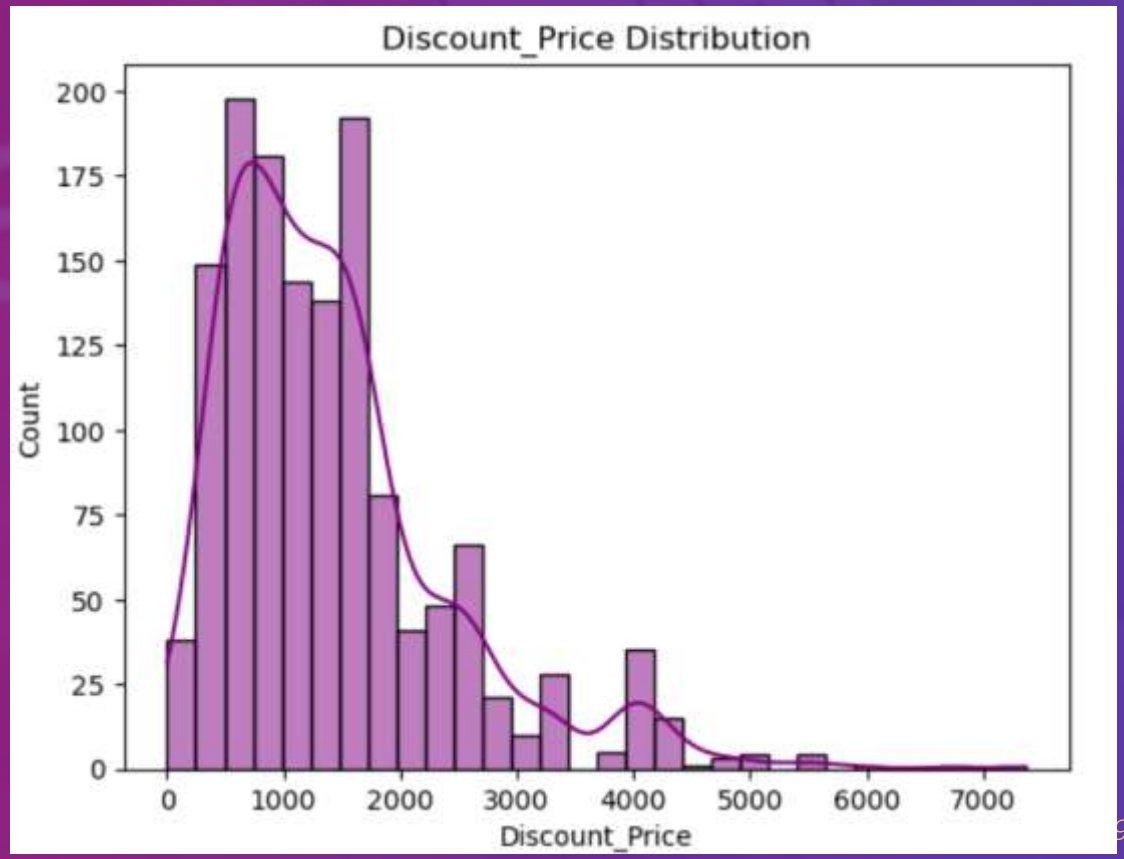


DISCOUNT ANALYSIS

INSIGHTS:

- Most products are concentrated in the low to mid discounted price range
- The distribution is right-skewed, indicating fewer high-priced discounted products
- A small number of high-value outliers exist at the upper price end
- Mid-range pricing dominates customer offerings on Snapdeal

Discount_Price Distribution

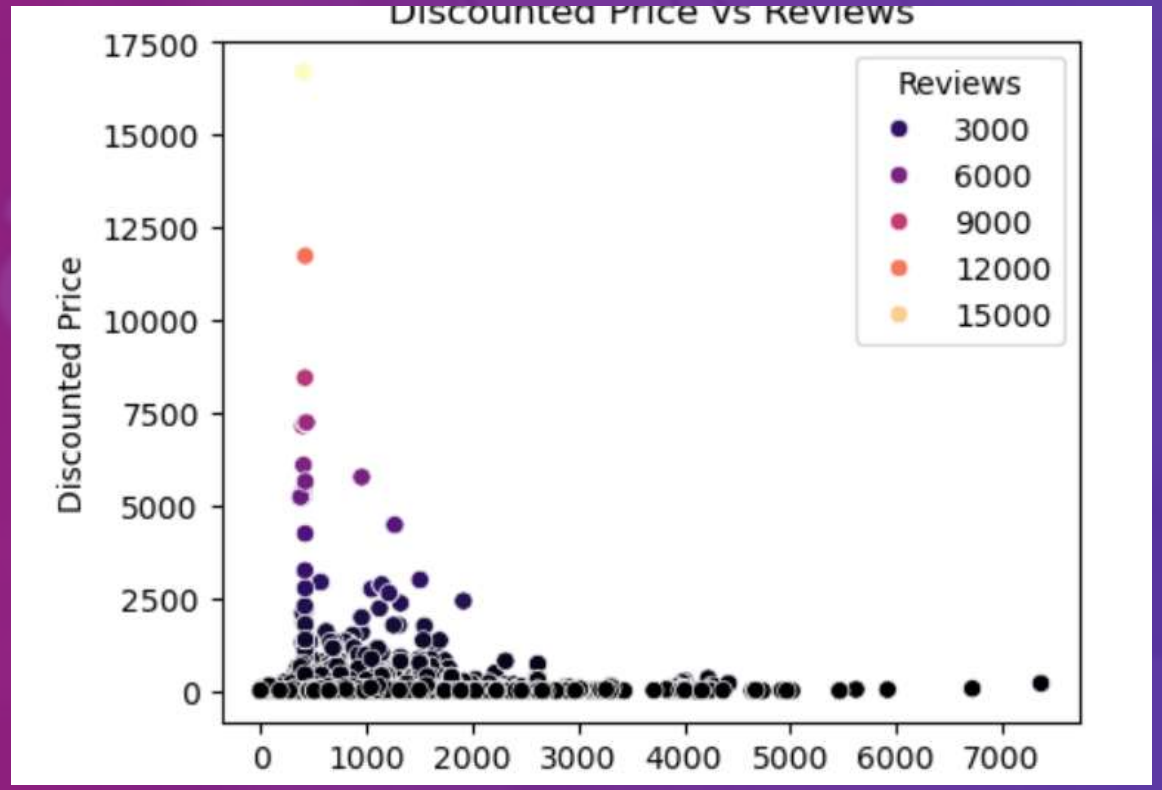


DISCOUNT ANALYSIS

INSIGHTS:

- **Lower-priced discounted products receive significantly more reviews**, indicating higher customer engagement in budget segments.
- **High discounted prices rarely achieve high review volumes**, suggesting limited demand for expensive items even after discounts.
- Products with **moderate prices and strong discounts** perform best in terms of visibility and interaction.
- A few **high-price outliers** exist but fail to scale in customer engagement, highlighting price sensitivity on Snapdeal.

Discount_Price Vs Reviews

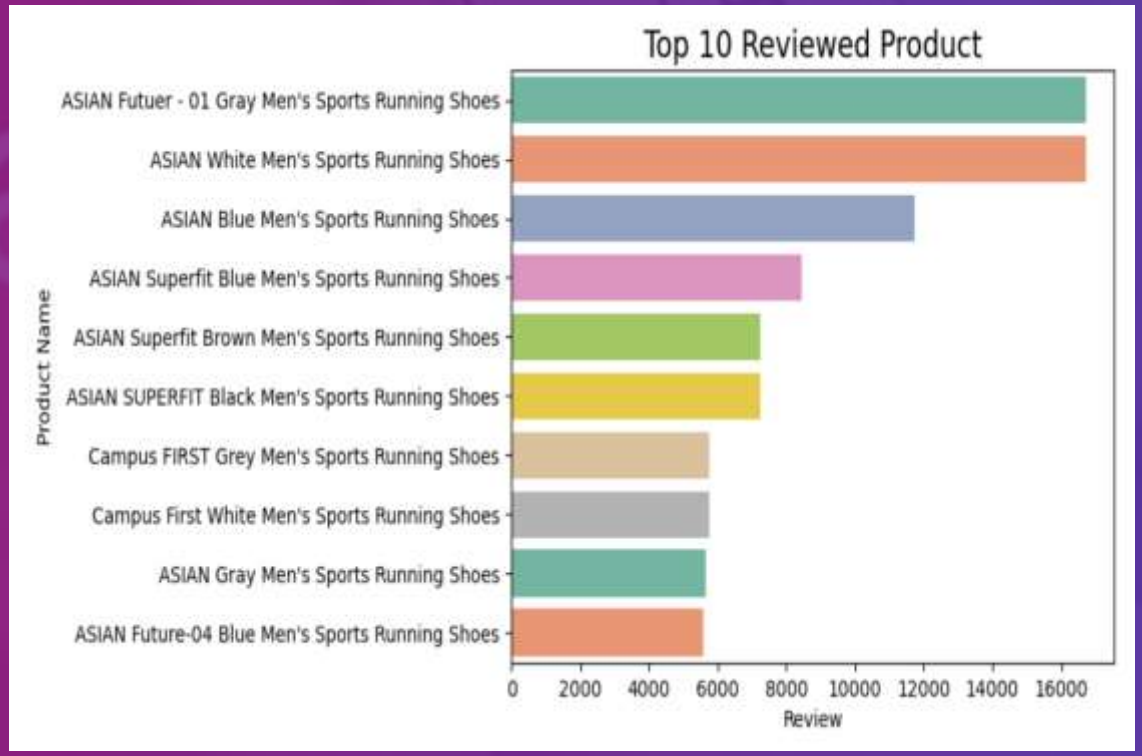


REVIEW & RATING ANALYSIS

INSIGHTS:

- Products with a large volume of reviews demonstrate **high visibility and sustained customer engagement**, indicating consistent demand.
- These items are commonly found in **well-established and highly competitive product categories**, where customer interaction is frequent.
- A higher number of reviews suggests that these products are **regularly purchased and actively shared by customers through feedback**.
- Most of the highly reviewed products also show **favorable ratings**, implying that popularity is generally backed by positive customer experience.

Top 10 Reviewed Product

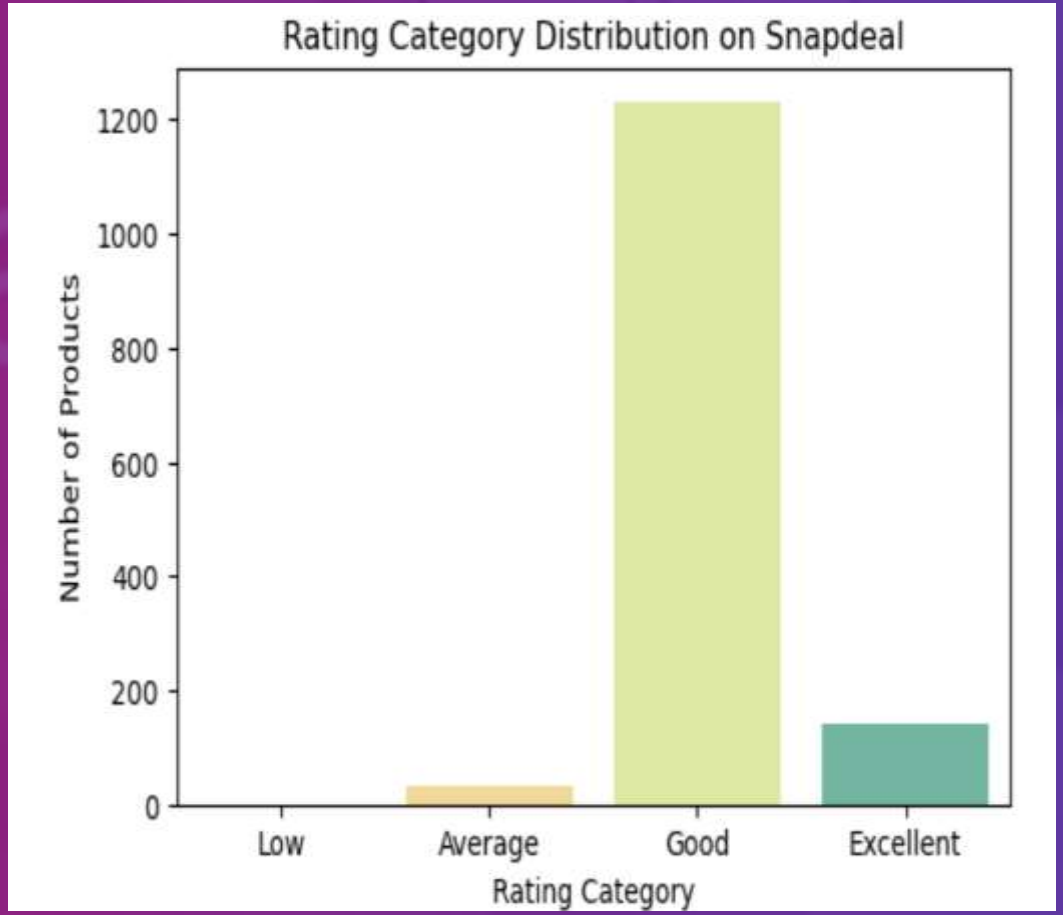


REVIEW & RATING ANALYSIS

INSIGHTS:

- The majority of products fall under the **“Good” rating category**, reflecting acceptable but improvable quality standards.
- **“Excellent” rated products are relatively fewer**, presenting an opportunity to promote and expand high-quality offerings.
- Very few products are rated **Low or Average**, indicating effective filtering or customer preference toward better-rated products.
- Improving products from **Good → Excellent** can significantly boost trust and conversions.

Rating Category Distribution

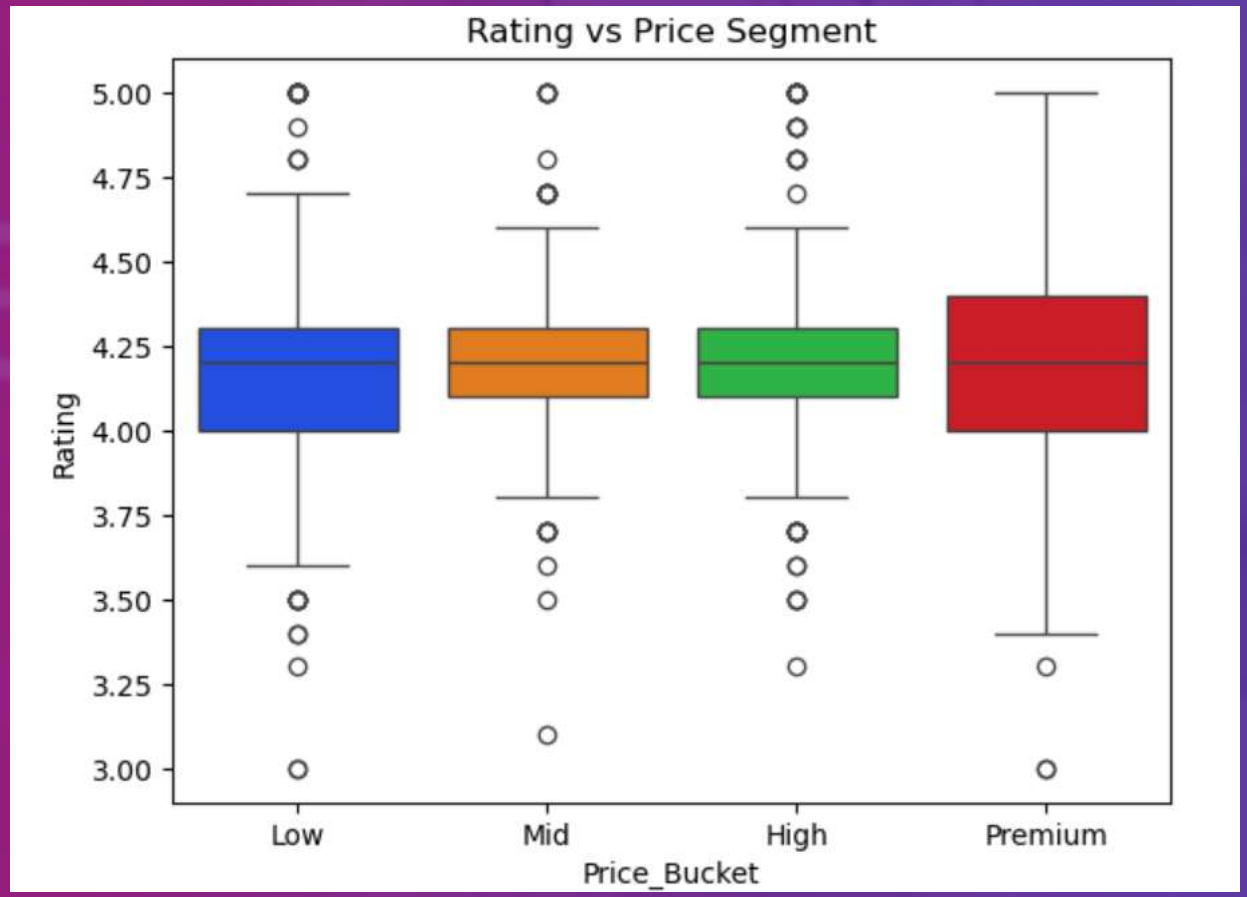


REVIEW & RATING ANALYSIS

INSIGHTS:

- **Low and mid-priced products achieve ratings comparable to premium products, highlighting strong value-for-money offerings.**
- **The premium price segment shows greater variability in ratings, reflecting mixed customer experiences at higher prices.**
- **Several outliers are observed across all segments:**
  - **Some low-priced products receive exceptionally high ratings, suggesting strong quality at affordable prices.**
  - **A few high-priced products have lower ratings, indicating that higher price does not always guarantee better customer satisfaction.**

Rating Vs Price Segment

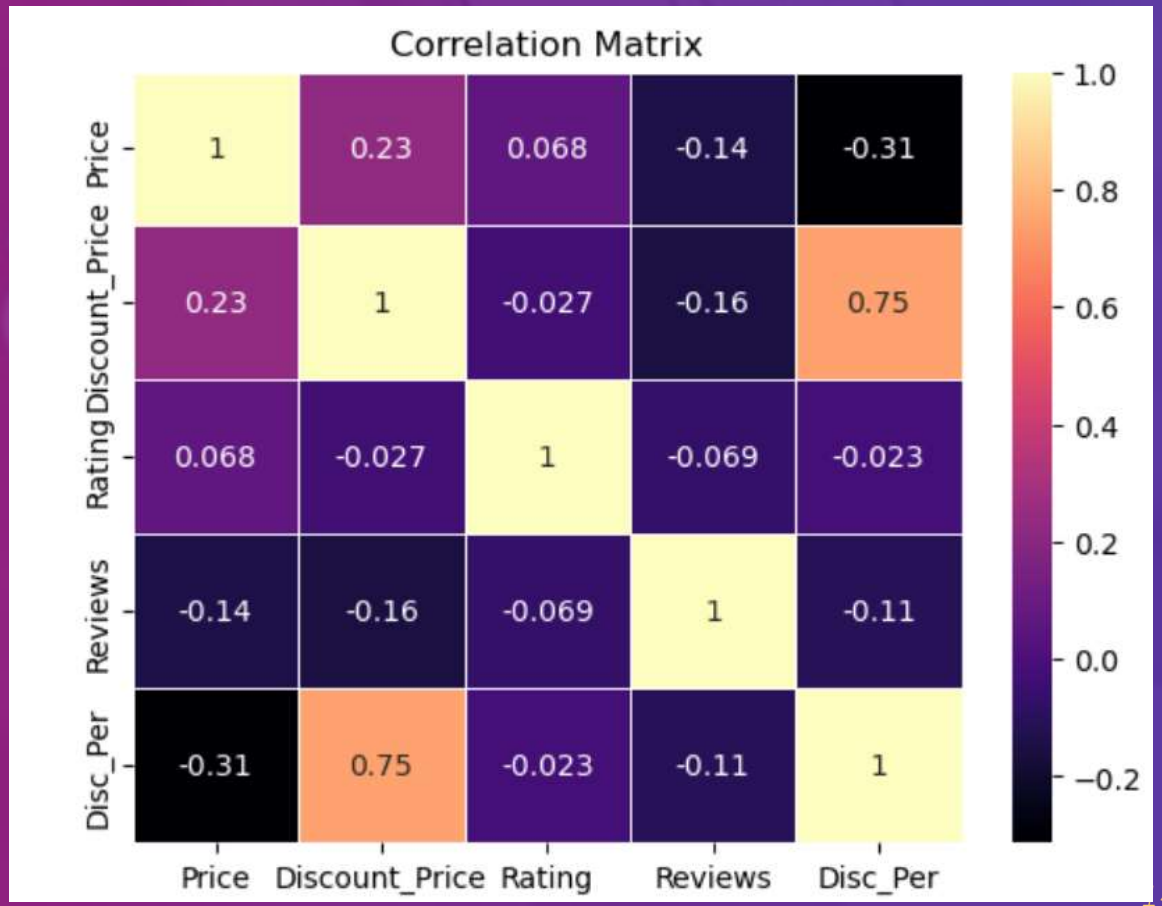


CORRELATION ANALYSIS

INSIGHTS:

- Discount Price and Discount Percentage show a strong positive correlation, as expected
- Product price has a moderate positive correlation with discounted price
- Ratings and reviews show weak correlation with price, suggesting customer satisfaction is not price-dependent
- Reviews have a slight negative correlation with price, indicating higher engagement for affordable products

Rating Vs Price Segment



MODEL BUILDING

RANDOM FOREST REGRESSOR- MODEL

- ✓ ***$R^2 \approx 0.97$ indicates an excellent fit, explaining almost all variability in the target.***
- ✓ ***Low MAE (~18) shows predictions are very close to actual values on average.***
- ✓ ***RMSE > MAE suggests a few larger errors, but overall prediction accuracy is very strong.***
- ✓ ***The model captures complex non-linear patterns effectively.***

- **Random Forest Regressor** was selected due to the **non-linear relationships** observed between price, ratings, reviews, and discounts during EDA.
- Price data showed **skewness and outliers**, which Random Forest handles better than linear models due to its tree-based structure.
- Features such as **discount percentage and MRP** had varying influence across price ranges, making ensemble models more suitable.
- Random Forest effectively captured **complex interactions** between features like ratings, review count, and discounts without requiring strict assumptions.
- The model demonstrated **strong predictive performance** with improved R^2 and low error.
- **Feature importance analysis** revealed that MRP and discount-related variables were the most influential predictors of price.

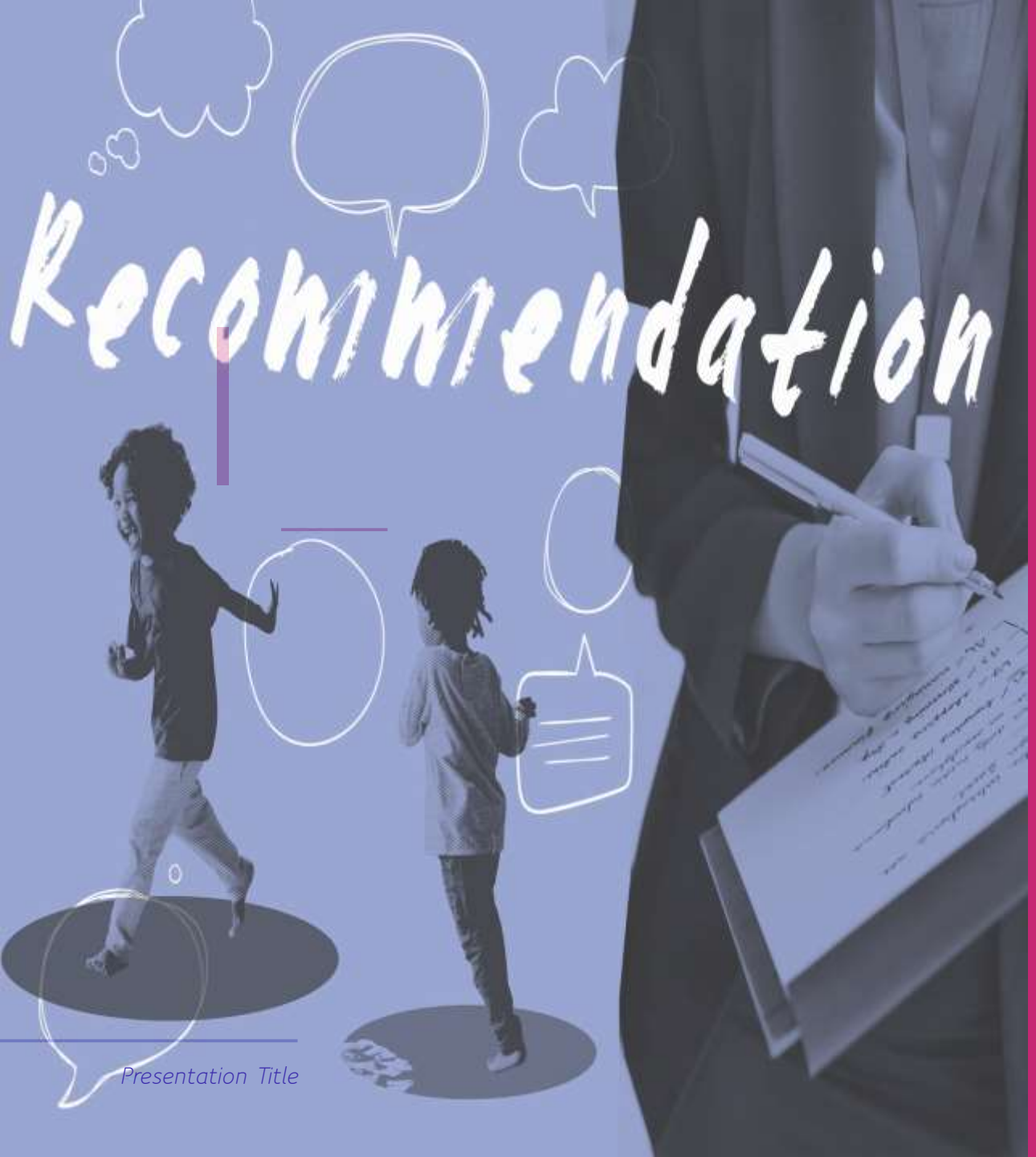
K-MEANS CLUSTERING

- K-Means clustering was applied to segment Snapdeal products based on **price, discount percentage, ratings, and review volume**, resulting in three well-defined product segments.
- **Cluster 1 – Budget High-Volume Segment** consists of **low-priced products (≈ ₹628)** with **very high review counts** and higher discounts, representing **price-sensitive customers** and mass-market demand.
- **Cluster 2 – Value-for-Money Segment** includes **mid-range products (≈ ₹778)** with **moderate reviews and consistently high ratings**, indicating balanced demand driven by quality and affordability.
- **Cluster 0 – High-Price, Low-Engagement Products** comprises **high-priced products (≈ ₹1,609)** with **lower review volumes but strong ratings**, reflecting **niche or brand-driven demand** rather than volume-based sales.

Overall, clusters with **higher ratings and review volumes** represent **high-performing products**, while premium clusters highlight quality-focused products with limited reach, enabling targeted **pricing and promotional strategies**.

BUSINESS INSIGHTS

- ***Engagement depends more on customer feedback than pricing***
  Products that accumulate strong ratings and a large number of reviews tend to perform better on Snapdeal, even when they are not the lowest priced. This highlights the importance of **trust, credibility, and perceived quality**.
- ***Discounts improve visibility but do not guarantee satisfaction***
  Higher discounts increase product exposure and interaction; however, products with poor ratings do not benefit equally, showing that **discounts work best when combined with good quality**.
- ***Final selling price influences customers more than MRP***
  User behavior aligns closely with the discounted price rather than the original listed price, showing that **effective price presentation and realistic discounts** are key drivers.
- ***Product satisfaction is consistent across price levels***
  Highly rated products appear in both budget and premium segments, reinforcing that **customer satisfaction is driven by product value rather than price alone**.
- ***Customer interaction decreases for higher-priced products***
  As product prices increase, review counts generally decline, suggesting that **lower-priced products encourage higher participation and feedback**.
- ***Random Forest effectively models pricing behavior***
  The Random Forest model successfully captured complex, non-linear relationships between pricing, discounts, ratings, and reviews, validating its suitability for this dataset.
- ***Moderately priced products often outperform premium items***
  Several mid-priced products achieve better engagement and visibility than higher-priced alternatives due to **strong feedback and competitive positioning**.



- **Promote high-rated and well-reviewed products** as they drive higher customer trust and better conversions than low-priced alternatives.
- **Apply discounts strategically**; moderate discounts increase visibility without harming perceived quality.
- **Focus on mid-price segments**, which show stronger ratings and consistent customer engagement.
- **Improve quality for low-rated but high-traffic products** through seller checks and better product information.
- **Use Random Forest–based price prediction** to support data-driven pricing and competitive positioning.
- **Encourage verified customer reviews** to improve product credibility and model accuracy.

WEBSCRAPING CODE

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import time
```

PHASE 1: WEB SCRAPING

```
base_url = "https://www.snapdeal.com/search?keyword=SHOES&santizedKeyword=&catId=&categoryId=0&suggested=false&vertical=p&noOfRe
```

```
headers = {
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) "
                  "AppleWebKit/537.36 (KHTML, like Gecko) "
                  "Chrome/115.0.0.0 Safari/537.36"}
```

```
# Scrap First Page

def scrape_snapdeal_first_page(url):
    products_data = []

    response = requests.get(url, headers=headers)
    soup = BeautifulSoup(response.text, "html.parser")
```

```
    for item in items:
        # Name
        title = item.find("p", class_="product-title")
        name = title.text.strip() if title else None

        # Price
        price_tag = item.find("span", class_="product-price")
        price = (
            price_tag.text.replace("₹", "").replace(",", "").strip()
            if price_tag else None
        )

        # MRP
        mrp_tag = item.find("span", class_="lfloat product-desc-price strike")
        mrp = (
            mrp_tag.text.replace("₹", "").replace(",", "").strip()
            if mrp_tag else price
        )

        # Discount
        discount_tag = item.find("div", class_="product-discount")
        discount = discount_tag.text.strip() if discount_tag else "0%"

        # Rating
        rating = None
        rating_tag = item.find("div", class_="filled-stars")
```

RAW DATA

	Name	Price	MRP	Discount	Rating	Reviews
0	Campus SNIPER LIGHT GREY Men's Sports Running ...	Rs. 807	Rs. 1949	59% Off	4.3	163
1	ASIAN TITAAN-06 Off White Men's Sports Running...	Rs. 717	Rs. 1999	64% Off	4.3	712
2	hotstyle Gray Men's Sports Running Shoes	Rs. 512	Rs. 2249	77% Off	4.0	826
3	PENNEN Black Men's Sports Running Shoes	Rs. 354	Rs. 999	65% Off	3.7	48
4	Bersache N-SPO-S-9197 White Men's Sports Runni...	Rs. 1061	Rs. 4999	79% Off	4.1	121
5	Clymb Gray Men's Sports Running Shoes	Rs. 632	Rs. 2100	70% Off	4.1	538
6	Campus ZURIK PRO Blue Men's Sports Running Shoes	Rs. 827	Rs. 1999	59% Off	4.3	653
7	HotStyle (tm) Gray Men's Sports Running Shoes	Rs. 512	Rs. 2249	77% Off	4.4	19
8	ASIAN DOMINATOR-03 Navy Men's Sports Running S...	Rs. 839	Rs. 2499	66% Off	4.2	1266
9	Campus WELLS Black Men's Sports Running Shoes	Rs. 602	Rs. 899	33% Off	4.1	337
10	ASIAN NAVIGATOR-02 White Men's Sports Running ...	Rs. 841	Rs. 2999	72% Off	4.3	217
11	Clymb Sports Olive Men's Sports Running Shoes	Rs. 503	Rs. 999	50% Off	4.2	365
12	Campus MIKE (N) Navy Blue Men's Sports Running...	Rs. 950	Rs. 1999	52% Off	4.3	950
13	Clymb Sports Blue Men's Sports Running Shoes	Rs. 503	Rs. 999	50% Off	4.2	107

DASHBOARD

snapdeal

Average Price	Average Rating	Total Reviews	Total Discount
₹ 1.3M	4.2 ★	488.5K	₹ 2.0M

Price Range

₹ 315 — ₹ 2,639

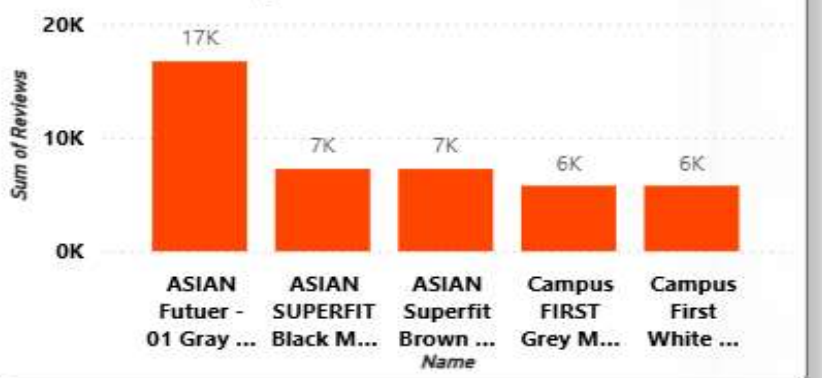
Average Discount %

57.12%

Top Products by Discount %

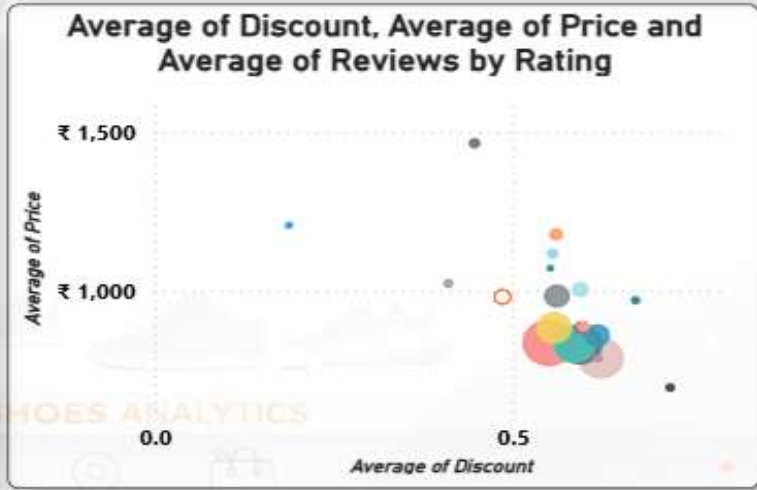
Average of Price and Average of MRP by Rating

Top 5 Reviewed Product



Count of Reviews by Rating

Average of Discount, Average of Price and Average of Reviews by Rating



Average of Price by Rating

VIDEO LINK

Part 1- https://www.loom.com/share/85a508322f644af6a375feee1df2adf3

Part 2- https://www.loom.com/share/6912271fa5ae483ebafca00041427366