PROJECT REPORT

End-to-End E-Commerce Product Engagement Analysis

Submitted By: Irfan Khan J **Date:** January 31, 2026 **Tech Stack:** Python, PostgreSQL, Power BI

---

1. Executive Summary

The objective of this project was to analyze a raw dataset of Amazon product listings to identify key drivers of customer engagement and pricing effectiveness. By leveraging **Python** for data cleaning, **PostgreSQL** for backend analytics, and **Power BI** for visualization, we transformed raw data into actionable business intelligence.

The analysis revealed significant insights:

1. **Niche Dominance:** While "Electronics" holds the most inventory, **"Musical Instruments"** generates the highest per-product engagement.
2. **Discount Strategy:** A **"Medium" discount** strategy yields higher engagement than aggressive "Very High" discounts.
3. **Brand Power:** Amazon's private labels (Amazon Basics) dominate the top engagement slots, proving brand trust drives volume.

---

2. Introduction

E-commerce platforms generate massive amounts of data, but raw data is rarely analysis-ready. This project demonstrates the end-to-end Data Science pipeline: cleaning messy data, engineering meaningful features, storing it in a relational database, and presenting it via an interactive dashboard.

We aimed to answer three core business questions:

- Which product categories generate the most customer interaction?
- Does a higher discount percentage correlate with higher engagement?
- Which specific products are the "Star Performers"?

---

3. Phase 1: Data Cleaning & Preprocessing (Python)

The initial dataset contained formatting issues common in web-scraped data, such as currency symbols (₹) and text-based numeric columns. This phase involved standardizing the data to ensure mathematical accuracy.

Data Cleaning Methodology:

- **Type Conversion:** We removed currency symbols (₹) and commas from price columns using .str.replace() and converted strings to float64.
- **Handling Nulls:** Missing values in rating_count were filled with 0 to prevent calculation errors.

	product_id	product_name	category	discounted_price	actual_price	discount_percentage	rating	rating_count
0	B07JW9H4J1	Wayona Nylon Braided USB to Lightning Fast Cha...	Computers&Accessories|Accessories&Peripherals|...	₹399	₹1,099	64%	4.2	24,269
1	B098NS6PVG	Ambrane Unbreakable 60W / 3A Fast Charging 1.5...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹349	43%	4.0	43,994
2	B096MSW6CT	Sounce Fast Phone Charging Cable & Data Sync U...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹1,899	90%	3.9	7,928
3	B08HDJ86NZ	boAt Deuce USB 300 2 in 1 Type-C & Micro USB S...	Computers&Accessories|Accessories&Peripherals|...	₹329	₹699	53%	4.2	94,363
4	B08CF3B7N1	Portronics Konnect L 1.2M Fast Charging 3A 8 P...	Computers&Accessories|Accessories&Peripherals|...	₹154	₹399	61%	4.2	16,905

about_product	user_id	user_name	review_id
High Compatibility : Compatible With iPhone 12...	AG3D6O4STAQKAY2UVGEUV46KN35Q,AHMY5CWJMMK5BJRBB...	Manav,Adarsh gupta,Sundeep,S.Sayeed Ahmed,jasp...	R3HXWT0LRP0NMF,R2AJM3LFTLZHFO,R6AQJGUP6P86,R1K...
Compatible with all Type C enabled devices, be...	AECPFYFQVRUWC3KGNLJIOREFP5LQ,AGYYVPDD7YG7FYNBX...	ArdKn,Nirbhay kumar,Sagar Viswanathan,Asp,Plac...	RGIQEG07R9HS2,R1SMWZQ86XIN8U,R2J3Y1WL29GWDE,RY...
【 Fast Charger& Data Sync】-With built-in safet...	AGU3BBQ2V2DDAMOAKGFAWDDQ6QHA,AESFLDV2PT363T2AQ...	Kunal,Himanshu,viswanath,sai niharka,saqib mal...	R3J3EQQ9TZI5ZJ,R3E7WBGK7ID0KV,RWU79XKQ6I1QF,R2...
The boAt Deuce USB 300 2 in 1 cable is compati...	AEWAZDZZJLQUYVOVGBEUKSLXHQ5A,AG5HTSFRRE6NL3M5S...	Omkar dhale,JD,HEMALATHA,Ajwadha,amar singh ...	R3EEUZKKK9J36I,R3HJVYCLYOY554,REDECAZ7AMPQC,R1...
[CHARGE & SYNC FUNCTION]- This cable	AE3Q6KSUK5P75D5HFYHCRAOLODSA,AFUGIFH5ZAFXRDSZH...	rahuls6099,Swasat Borah,Ajay Wadke,Pranali,RVK...	R1BP4L2HH9TFUP,R16PVJEXKV6QZS,R2UPDB81N66T4P,R...

review_title	review_content	img_link	product_link
Satisfied,Charging is really fast,Value for mo...	Looks durable Charging is fine tooNo complains...	https://m.media-amazon.com/images/W/WEBP_40237...	https://www.amazon.in/Wayona-Braided-WN3LG1-Sy...
A Good Braided Cable for Your Type C Device,Go...	I ordered this cable to connect my phone to An...	https://m.media-amazon.com/images/W/WEBP_40237...	https://www.amazon.in/Ambrane-Unbreakable-Char...
Good speed for earlier versions,Good Product,W...	Not quite durable and sturdy,https://m.media-a...	https://m.media-amazon.com/images/W/WEBP_40237...	https://www.amazon.in/Sounce-iPhone-Charging-C...
Good product,Good one,Nice,Really nice product...	Good product,long wire,Charges good,Nice,I bou...	https://m.media-amazon.com/images/I/41V5FtEWPk...	https://www.amazon.in/Deuce-300-Resistant-Tang...
As good as original,Decent,Good one for second...	Bought this instead of original apple, does th...	https://m.media-amazon.com/images/W/WEBP_40237...	https://www.amazon.in/Portronics-Konnect-POR-1...

Figure 1: Initial dataset inspection revealing string formatting issues in price and discount columns.

```
df.info()
```

```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   product_id           1465 non-null   object
 1   product_name         1465 non-null   object
 2   category             1465 non-null   object
 3   discounted_price     1465 non-null   object
 4   actual_price         1465 non-null   object
 5   discount_percentage  1465 non-null   object
 6   rating               1465 non-null   object
 7   rating_count         1463 non-null   object
 8   about_product        1465 non-null   object
 9   user_id              1465 non-null   object
 10  user_name            1465 non-null   object
 11  review_id            1465 non-null   object
 12  review_title         1465 non-null   object
 13  review_content       1465 non-null   object
 14  img_link             1465 non-null   object
 15  product_link         1465 non-null   object
dtypes: object(16)
```

Figure 2: Data type analysis showing all columns were initially loaded as 'object', necessitating type conversion.

Text Data Validation: We also validated the text-heavy columns (Reviews and Descriptions) to ensure they were correctly parsed and not truncated.

---

4. Phase 2: Feature Engineering (Python)

To analyze engagement effectively, we engineered new metrics that combine volume and quality.

- **Engagement Score:** We calculated Rating × Rating Count. This metric captures both the quality of a product and its popularity.
- **Price Binning:** Used pd.qcut(q=5) to categorize actual_price into quantiles.
- **Discount Categorization:** Used pd.cut() to bin discount_percentage into four logical levels (Low, Medium, High, Very High).

	rating	rating_count	engagement_score
0	4.2	24269	101929.8
1	4.0	43994	175976.0
2	3.9	7928	30919.2
3	4.2	94363	396324.6
4	4.2	16905	71001.0

Figure 5: The 'engagement_score' column was created to quantify product popularity.

Initial Analysis (Python): Before moving to the database, we ran a quick visualization in Python to test our hypothesis about discounts.

```
df.groupby('discount_level', observed=True)['engagement_score'] \
  .mean() \
  .plot(kind='bar', title='Engagement by Discount Level')
```

```
<Axes: title={'center': 'Engagement by Discount Level'}, xlabel='discount_level'>
```

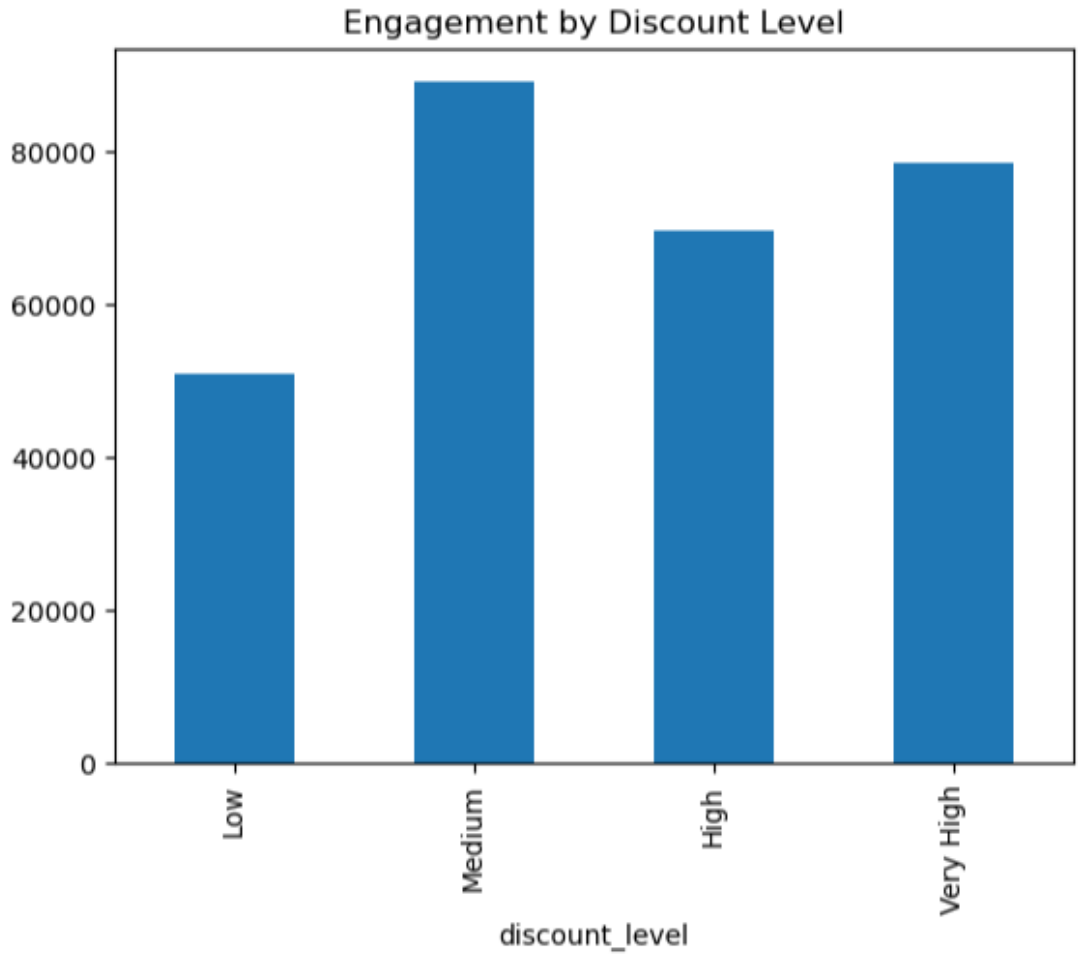

Figure 6: Python bar chart indicating that 'Medium' discount levels drive higher engagement than 'Very High' discounts.

---

5. Phase 3: Database Integration (PostgreSQL)

The cleaned DataFrame was migrated to **PostgreSQL** to enable complex querying and persistent storage. We utilized the sqlalchemy library for the migration.

- **Challenges & Solutions:** We resolved dependency errors by clearing existing views before data migration to ensure a clean load.

```
table_name = "customer"
df.to_sql(table_name, engine, if_exists="replace", index=False)
print(f"✅ Data successfully loaded into table '{table_name}'.")
```

```
✅ Data successfully loaded into table 'customer'.
```

Figure 7: Confirmation of successful data ingestion into the customer table.

---

6. Phase 4: SQL Analytics & Insights

With data in the database, we executed analytical queries to extract business intelligence.

Category Analysis: Using SPLIT_PART to extract the main category, we aggregated the average engagement score.

	main_category text	avg_engagement_score double precision
1	musicalinstruments	175742.2
2	electronics	124480.4441064639
3	computers&accessori...	72103.53885209712
4	toys&games	68228.09999999999
5	home&kitchen	27374.24362416109
6	officeproducts	20892.50967741936
7	homeimprovement	18628.25
8	health&personalcare	14652
9	car&motorbike	4248.4

Figure 8: Category analysis revealing "Musical Instruments" as the highest engagement niche (Avg Score: 175k) despite lower product volume.

Insight: This suggests a "Niche Gem" strategy. While we sell fewer instruments, the customers are highly passionate.

Top Product Identification: We identified the top 5 products by engagement_score.

	product_name text	main_category text	discount_level text	engagement_score double precision
1	AmazonBasics Flexible Premium HDMI Cable (Black, 4K@60Hz, 18Gbps), 3-Foot	electronics	Very High	1878681.2000000002
2	Amazon Basics High-Speed HDMI Cable, 6 Feet (2-Pack),Black	electronics	Very High	1878681.2000000002
3	Amazon Basics High-Speed HDMI Cable, 6 Feet - Supports Ethernet, 3D, 4K video,B...	electronics	High	1878681.2000000002
4	AmazonBasics Flexible Premium HDMI Cable (Black, 4K@60Hz, 18Gbps), 3-Foot	electronics	Very High	1878676.8
5	boAt Bassheads 100 in Ear Wired Earphones with Mic(Furious Red)	electronics	Very High	1491223.2999999998

Figure 9: Top 5 product identification, highlighting Amazon Basics HDMI Cables as market leaders.

Optimization (Views): To prepare for Power BI, we created a permanent SQL View (product_dashboard) to pre-calculate metrics and improve dashboard load times.

```sql
CREATE OR REPLACE VIEW product_dashboard AS
SELECT
    product_id,
    product_name,
    SPLIT_PART(category, '|', 1) AS main_category,
    discounted_price,
    discount_level,
    price_range,
    rating::NUMERIC AS rating,
    rating_count::NUMERIC AS rating_count,
    engagement_score
FROM customer
WHERE rating IS NOT NULL
  AND engagement_score IS NOT NULL;
```

Data Output | Messages | Notifications

```
CREATE VIEW

Query returned successfully in 83 msec.
```

Figure 10: Creating the product_dashboard view to optimize Power BI data retrieval.

---

7. Phase 5: Data Visualization (Power BI)

The final phase involved building a "One-Pager" executive dashboard in Power BI.

Dashboard Features:

- **KPI Cards:** Displaying Total Products (1.46k), Average Rating (4.10), and Avg Price.
- **Dual-Axis Analysis:** The center visualizations contrast Volume (Bar Chart) vs. Engagement (Line Chart).
- **Interactive Slicers:** A category menu on the left allows for instant filtering.

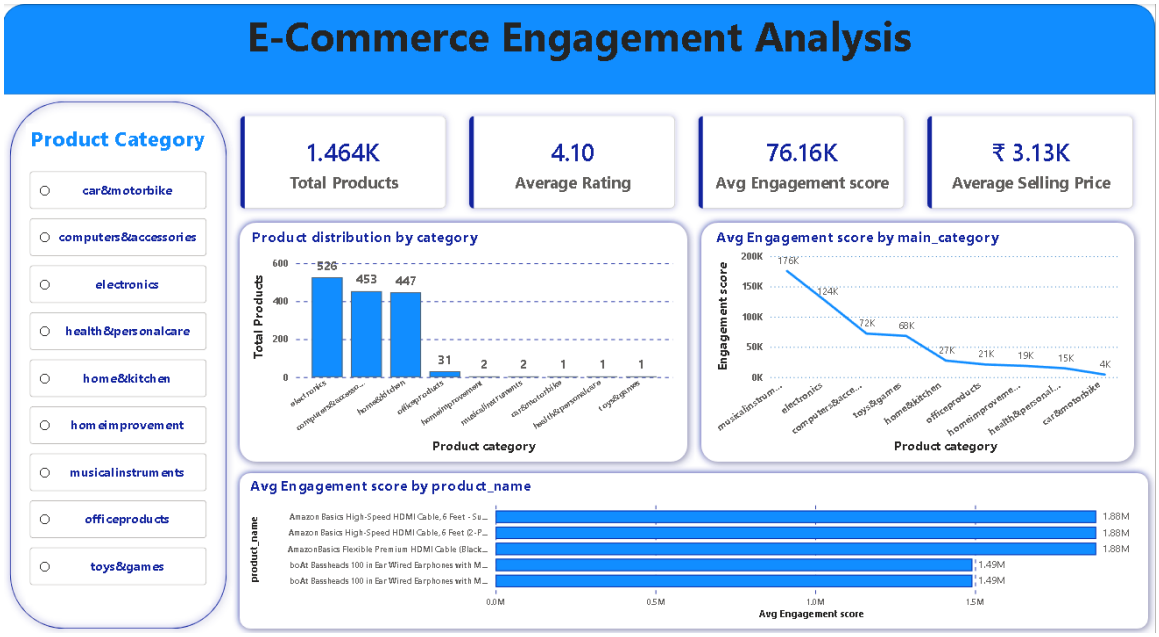

Figure 11: Full Dashboard Overview showing the dominance of Electronics in volume, but the high engagement trend line for Musical Instruments.

Deep Dive: Niche Analysis: By filtering for "Musical Instruments", we can see the specific performance of this high-value category.

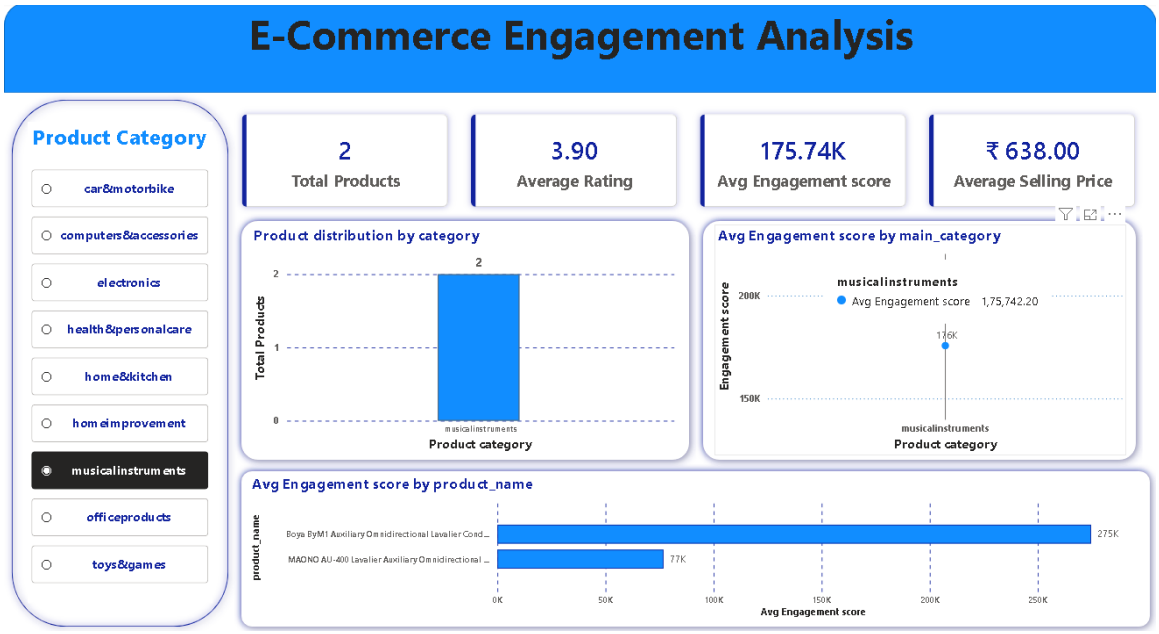

Figure 12: Dashboard filtered for 'Musical Instruments'. Although there are only 2 products, they possess a massive Average Engagement Score of 175.74k.

8. Business Recommendations

Based on the analysis, I recommend the following strategies:

1. **Target the "Niche" Market:**
    - **Action:** Launch a specific marketing campaign for **Musical Instruments**. Since the engagement rate is disproportionately high, this audience is loyal and likely to recommend products.
2. **Optimize Discounting Strategy:**
    - **Action:** Reduce the frequency of "Very High" discounts (>60%). Move pricing towards the "Medium" range. The data (Figure 6) proves that massive discounts do not guarantee higher engagement.
3. **Leverage Brand Trust:**
    - **Action:** Replicate the success of the "Amazon Basics" line (Figure 9) by creating "Bundle" offers for trusted categories (Cables, Chargers).

---

9. Conclusion

This project successfully demonstrated the complete data analytics lifecycle. We moved from raw, unstructured text data in a CSV file to a structured database and finally to an interactive, executive-grade dashboard. The insights derived from engagement_score and discount_level provide a clear roadmap for pricing and inventory strategies.