E-COMMERCE ANALYTICS

Yazdan Riazi
Student id : 16270796
Email : myrk5f@umsystem.edu
Professor : Syed Jawad Hussain Shah

PROJECT OVERVIEW / PROBLEM STATEMENT

Goal: Analyze customer behavior and build recommendation + segmentation models for an e-commerce platform, using 67M+ historical events.

Questions we aim to answer:

When do users shop most? (hour & weekday trends)

Which products and categories are most popular?

How do customers behave? (views, carts, purchases, sessions)

Can we segment users into meaningful clusters?

Can we build recommendation systems? (CF + time-aware)

How efficient is the purchase funnel?

DATA SAMPLING & PREPROCESSING

The original dataset contains **over 67 million rows**, which is too large to load into memory directly.
To solve this, I implemented **chunk-based random sampling**, ensuring that the sample is representative across all dates and hours.

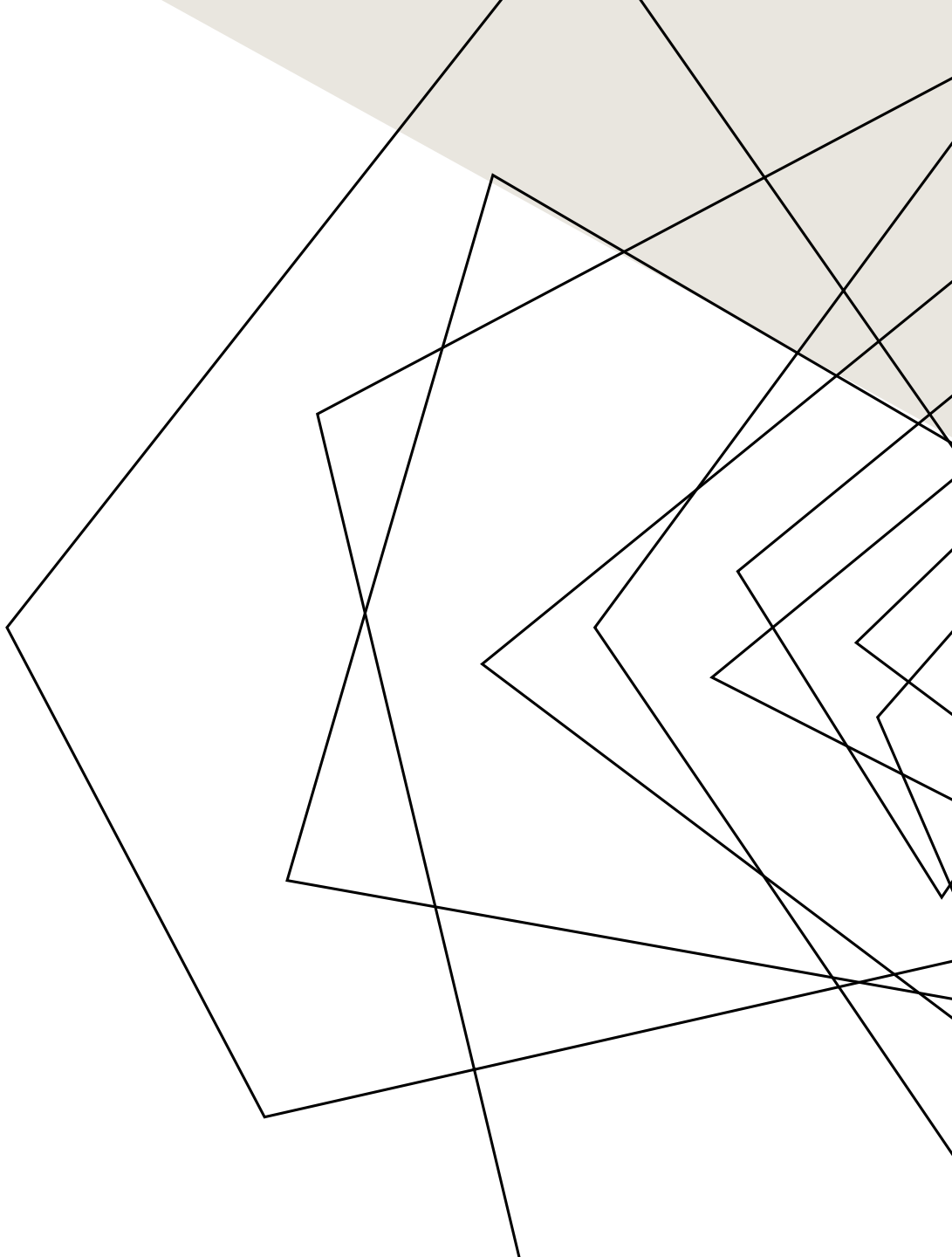
Sampling strategy

Read file in **2 million, row chunks**

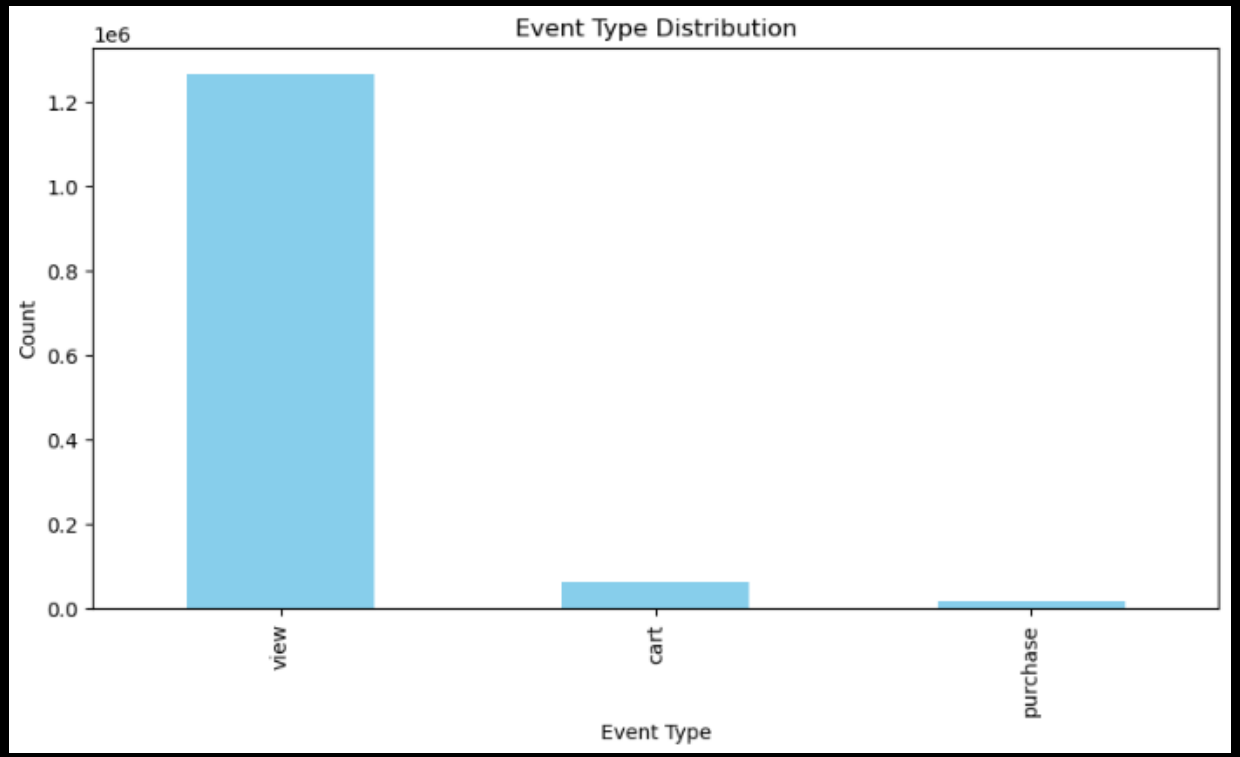
Randomly sample **2%** of each chunk (frac = 0.02)

Combine into one dataframe

Final sample size: **≈ 1.3 million records**



EXPLORATORY DATA ANALYSIS (EDA)

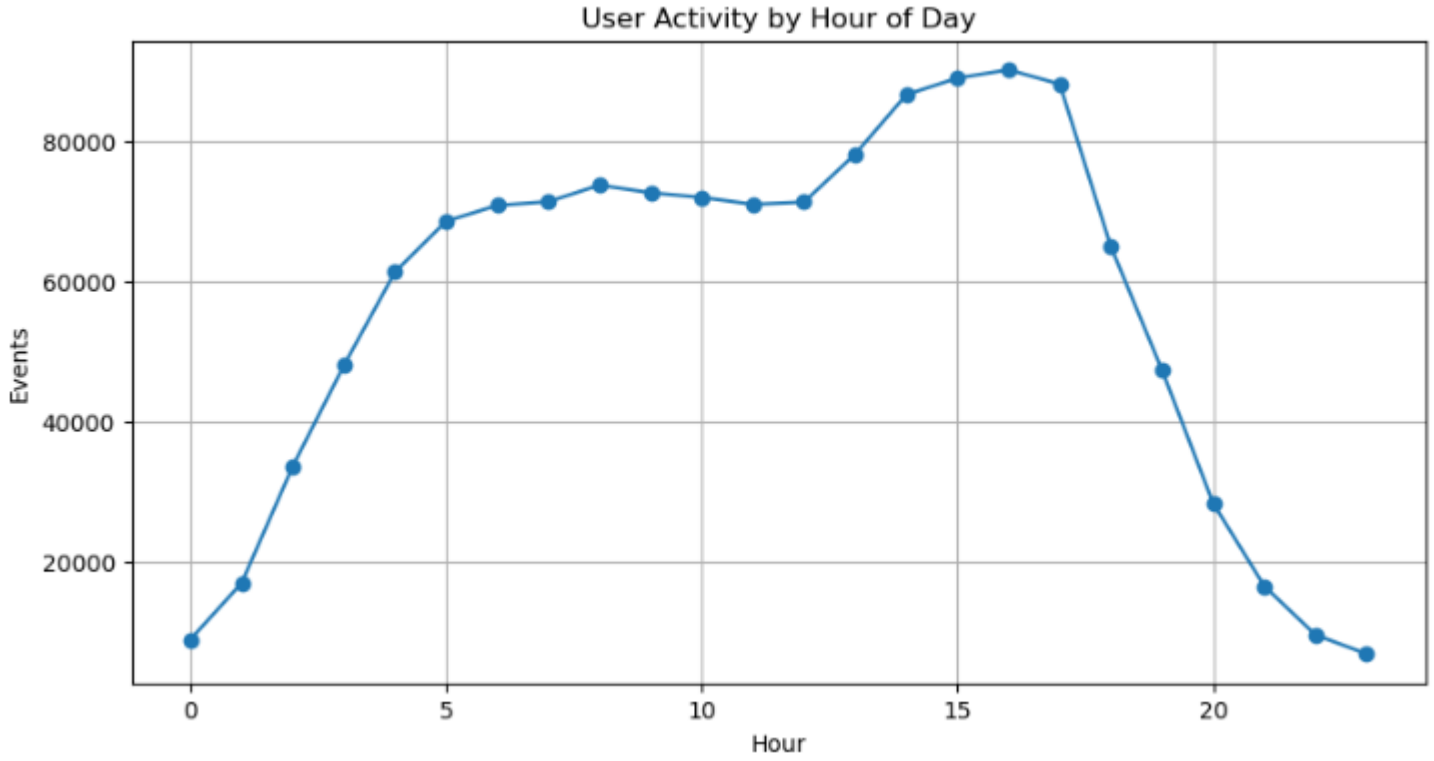


Event Type Distribution

Views dominate the dataset → ~93% of all events
Add-to-cart events are ~4–5%
Purchases are ~1–2%
This is expected for large e-commerce platforms because browsing is far more common than purchasing. It confirms typical “wide funnel” user behavior.

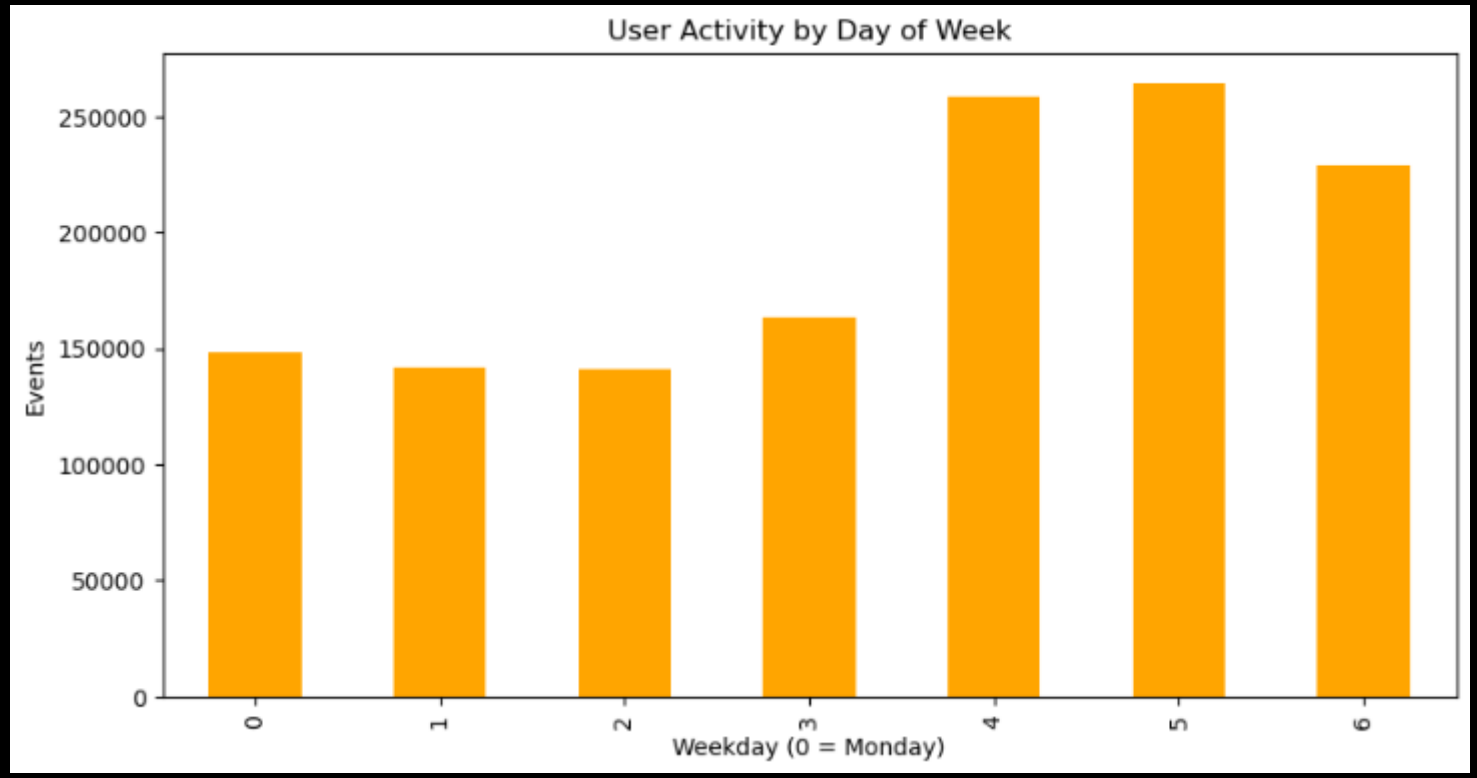
USER ACTIVITY BY HOUR

- **Activity rises sharply after 12 PM**
- **Peaks between 15:00–17:00**
- **Declines after 17:00**



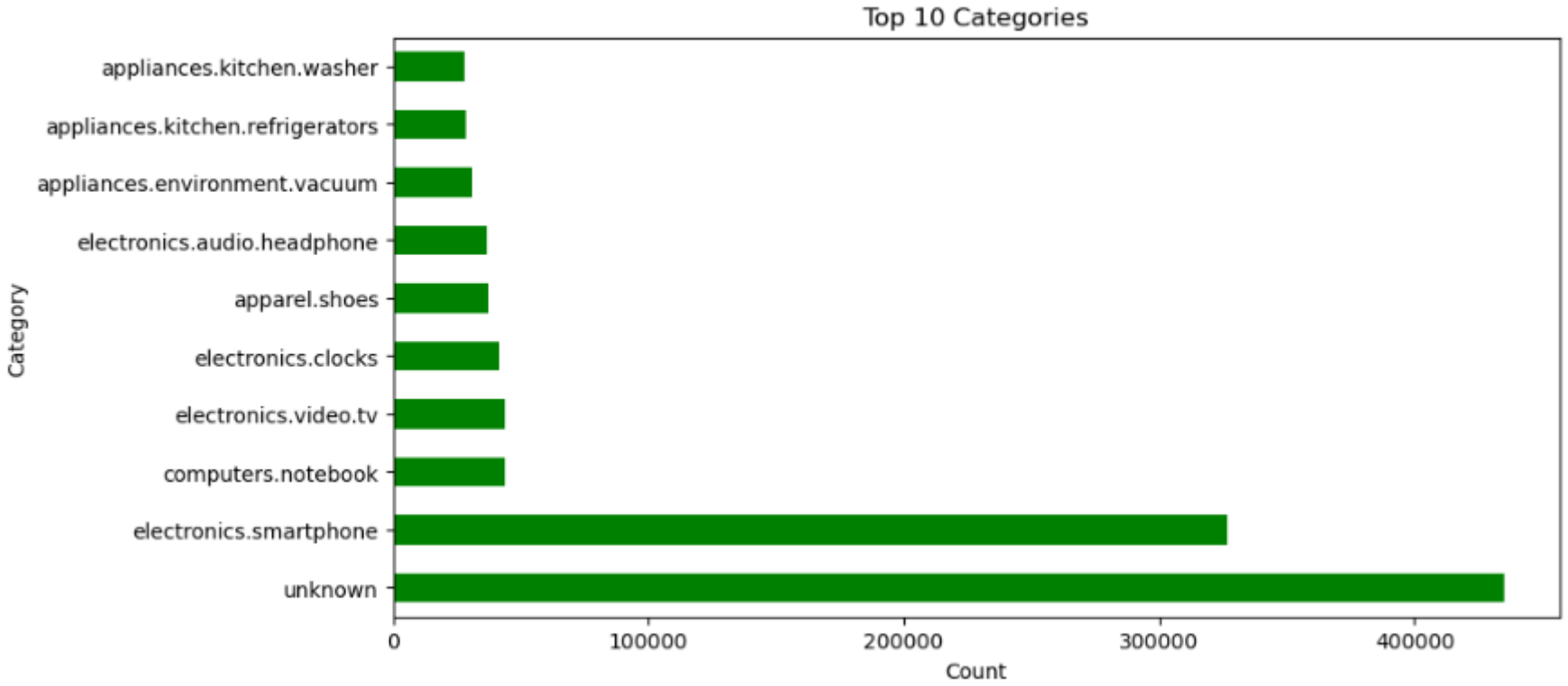
The afternoon peak suggests users shop after work/school hours.
Marketing campaigns should target 3–5 PM for maximum engagement.

USER ACTIVITY BY WEEKDAY



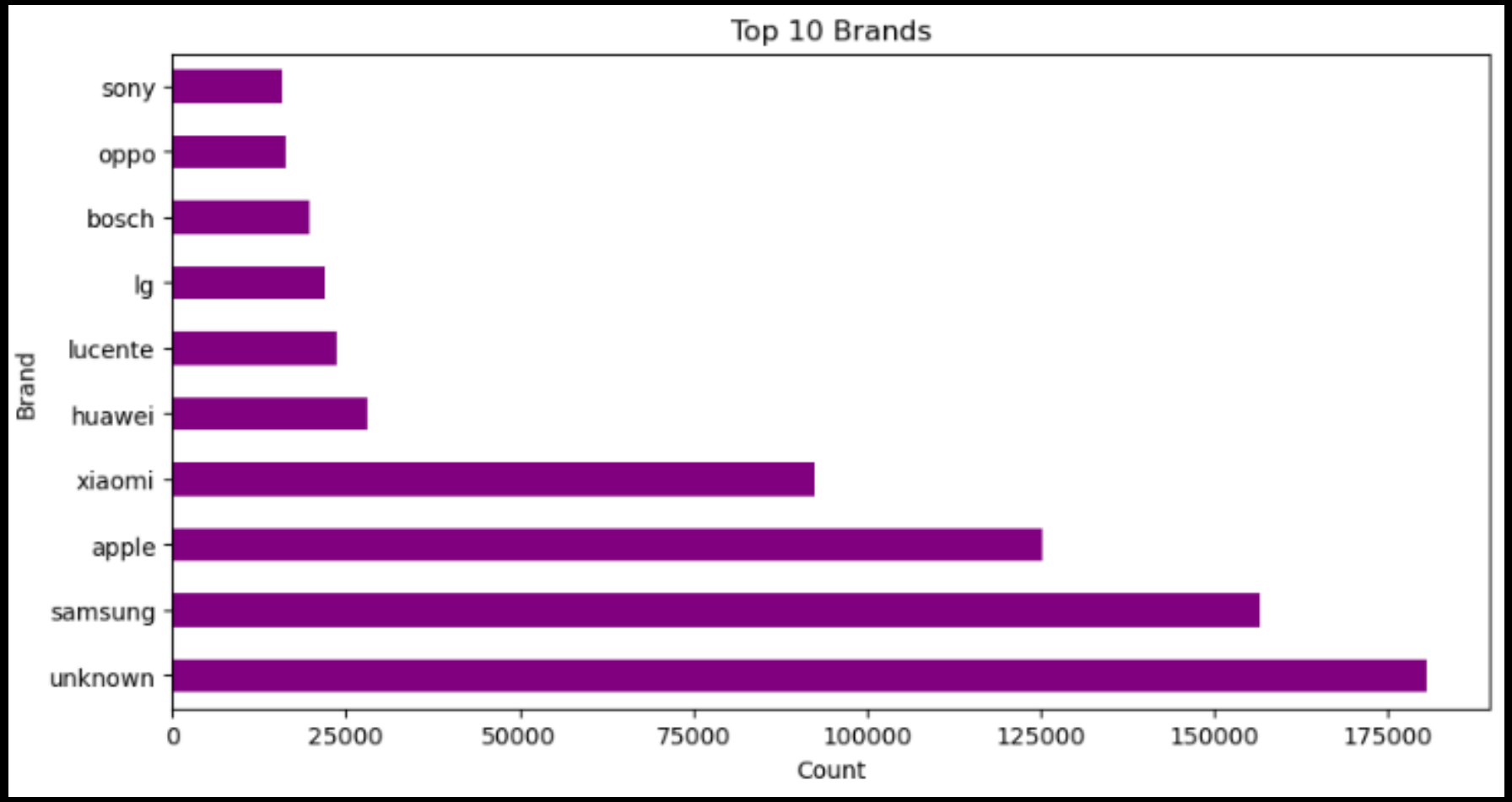
- Highest traffic on Saturday and Friday
- Lowest on early weekdays
- Traffic is highest on weekends, indicating leisure-driven shopping behavior.

TOP CATEGORIES

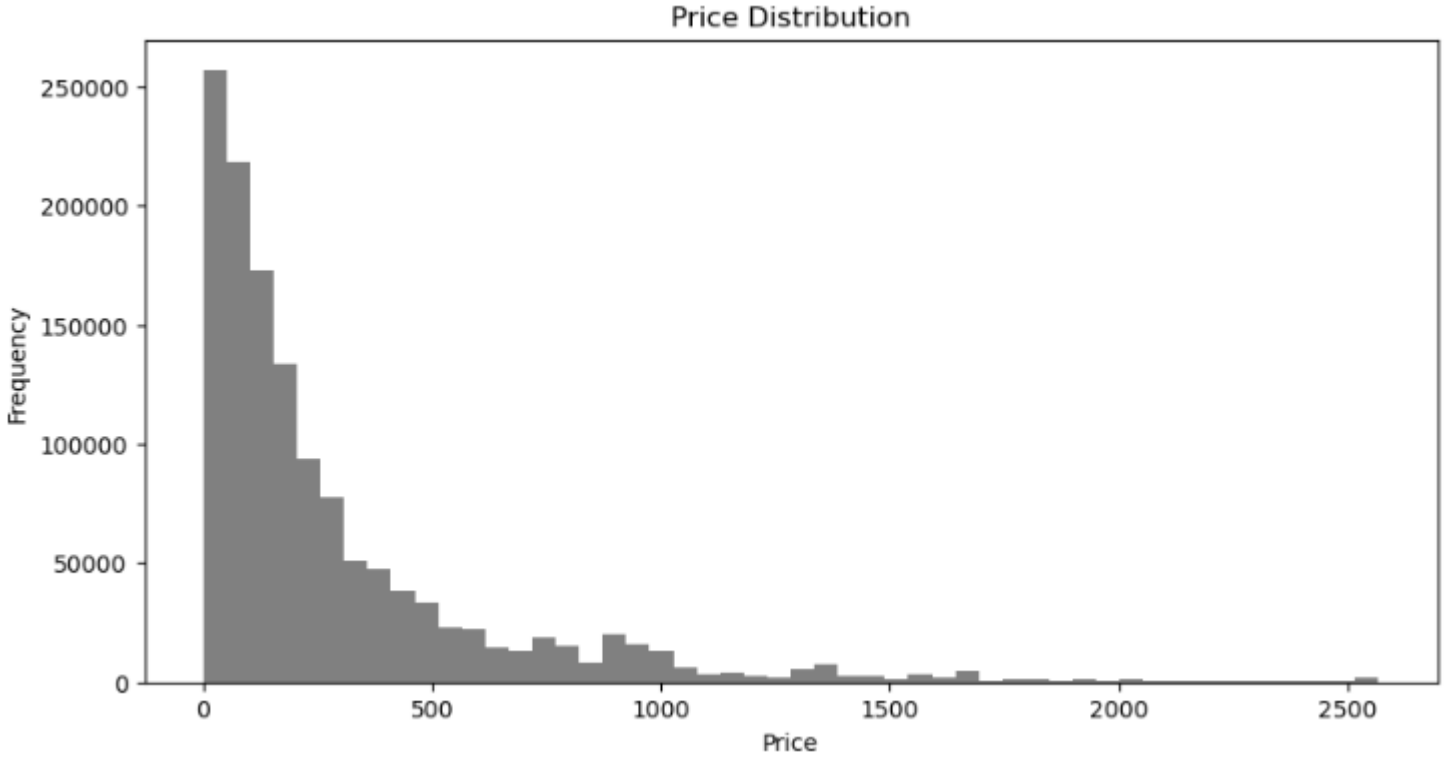


- **“Electronics.smartphone” is the clear leader**
- **Followed by notebooks, TVs, clocks**
- **“Unknown” appears frequently → category metadata missing**

TOP BRANDS


Top 10 Brands
Brand
sony
oppo
bosch
lg
lucente
huawei
xiaomi
apple
samsung
unknown
0
25000
50000
75000
100000
125000
150000
175000
Count

PRICE DISTRIBUTION



- Highly right-skewed
- Most products priced under $300
- Long tail up to $2500+

CUSTOMER BEHAVIORAL FEATURES

user_id	num_cart	num_purchases	num_views	total_spent	num_sessions	conversion_rate
31198833	0	0	1	0.0	1	0.0
49484535	0	0	1	0.0	1	0.0
82678143	0	0	1	0.0	1	0.0
94566147	0	0	1	0.0	1	0.0
107837897	1	0	0	0.0	1	0.0

Feature	**Meaning**
num_views	Total views
num_cart	Total add-to-carts
num_purchases	Total purchases
total_spent	Sum of all purchased product prices
num_sessions	Unique sessions
conversion_rate	purchases / sessions

This produced **769,000+** unique user profiles.

CUSTOMER SEGMENTATION (K-MEANS)

I tested K = 3 to K = 6 and found that K = 4 provided the most interpretable clusters with clear behavioral differences.

Segment 0 – Low-intent Browsers

- Low views
- Almost no cart/purchase activity
- Very low engagement

→ *Represents majority of users.*

Segment 1 – High-Value Buyers

- Higher cart + purchase activity
- Very high `total_spent`
- Highest conversion rate (0.88!)

→ *Primary revenue drivers.*

Segment 2 – Medium Browsers

Higher activity than Segment 0
Very little purchasing
→ *Potential to convert with better recommendations.*

Segment 3 – Power Viewers

Hundreds of views
No purchases
→ *Comparison shoppers / researching users*

This segmentation is extremely useful for personalized marketing.

Segment	User per segment
Segment 0	696690
Segment 2	56300
Segment 1	16668
Segment 3	2

COLLABORATIVE FILTERING RECOMMENDATION SYSTEM

Built a **user–item matrix** using only **purchase events**.

Collaborative Filtering Pipeline

1. Build user–item matrix (using purchases only)
2. Compute item-item cosine similarity
3. Recommend similar products based on user history

Filtering logic:

- Keep users with ≥2 purchases
- Keep products bought by ≥5 users
- Results: **1078 users × 13 products**

```
Number of purchase rows: 18356
Unique users with purchases: 17471
Unique products purchased: 5990
Filtered purchases for CF: (1078, 13)
```

All recommendations are logically similar to the user's purchase history.

TIME-AWARE RECOMMENDATIONS

Hour popularity (conditional probability)

Given an hour, what categories does the platform sell most?

Weekday popularity

Given a weekday, what categories sell most?

Final_score ranks products by combining global popularity (base_score), hour-specific demand patterns (hour_pop), and weekday patterns (weekday_pop).

Higher scores indicate products more likely to be purchased *at this specific time*.

Final_score = base_score + α * hour_pop + β * weekday_pop

product_id	category_code	brand	price	score
5100610	electronics.clocks	apple	334.34	0.433013
5701086	auto.accessories.player	pioneer	127.42	0.102062
1005009	electronics.smartphone	xiaomi	85.97	0.102062
4804718	electronics.audio.headphone	apple	360.09	0.072169
1005105	electronics.smartphone	apple	1348.28	0.039284
4804055	electronics.audio.headphone	apple	188.94	0.036084
1005238	electronics.smartphone	oppo	282.89	0.034021
1003317	electronics.smartphone	apple	928.18	0.034021
1004249	electronics.smartphone	apple	722.40	0.020833
1005100	electronics.smartphone	samsung	139.68	0.018042

SESSION-PRODUCT FUNNEL ANALYSIS

Constructed a funnel per (user, session, product):

- Did the user **view** the product?
- Did the user **add to cart**?
- Did the user **purchase**?

	user_id	user_session	product_id	first_time	last_time	viewed	added_to_cart	purchased
0	31198833	dbd84cb9-75db-42f6-88ec-dad77d71ccfb	1004767	2019-11-13 02:11:07+00:00	2019-11-13 02:11:07+00:00	True	False	False
1	49484535	e225514f-8ade-4f6c-9605-9481f4e608ea	12702774	2019-11-19 16:19:23+00:00	2019-11-19 16:19:23+00:00	True	False	False
2	82678143	98aa2dc0-46aa-465d-8694-e7ab9275fc8a	26011858	2019-11-21 03:04:24+00:00	2019-11-21 03:04:24+00:00	True	False	False
3	94566147	3c4da186-6885-4170-9346-a2c835c44694	1005007	2019-11-12 08:03:18+00:00	2019-11-12 08:03:18+00:00	True	False	False
4	107837897	8c62cf51-971c-46fc-8d06-9ec85e3a0740	4700557	2019-11-29 05:00:04+00:00	2019-11-29 05:00:04+00:00	False	True	False

Funnel totals (product-level):

- Views: **1,252,647**
- Add-to-cart: **59,504**
- Purchases: **18,280**

Product-level cart abandonment rate: **99.24%**

Product-level abandonment is always extremely high because users often:

- Add product A to cart
- But buy product B
- Or buy without adding to cart

FINAL BUSINESS INSIGHT

1. Smartphones dominate the entire marketplace

Demand, views, and purchases are heavily skewed toward electronics.

2. Most users are passive browsers (Segment 0)

→ Opportunity: retargeting, personalized discounts

3. Segment 1 contributes disproportionate revenue

→ Treat them as “VIP customers”

4. Product-level cart abandonment is extremely high

→ Evaluate checkout flow

→ Improve cart reminders

→ Offer real-time promotions

FINAL BUSINESS INSIGHT

5. Peak shopping hours are 3–5 PM

→ Schedule ads, push notifications during this window

6. Friday, Saturday and Sunday = highest traffic

→ Weekly campaigns should target these days

7. Time-aware recommendations outperform static CF

→ Higher responsiveness
→ Matches demand trends
→ Great fit for real production systems

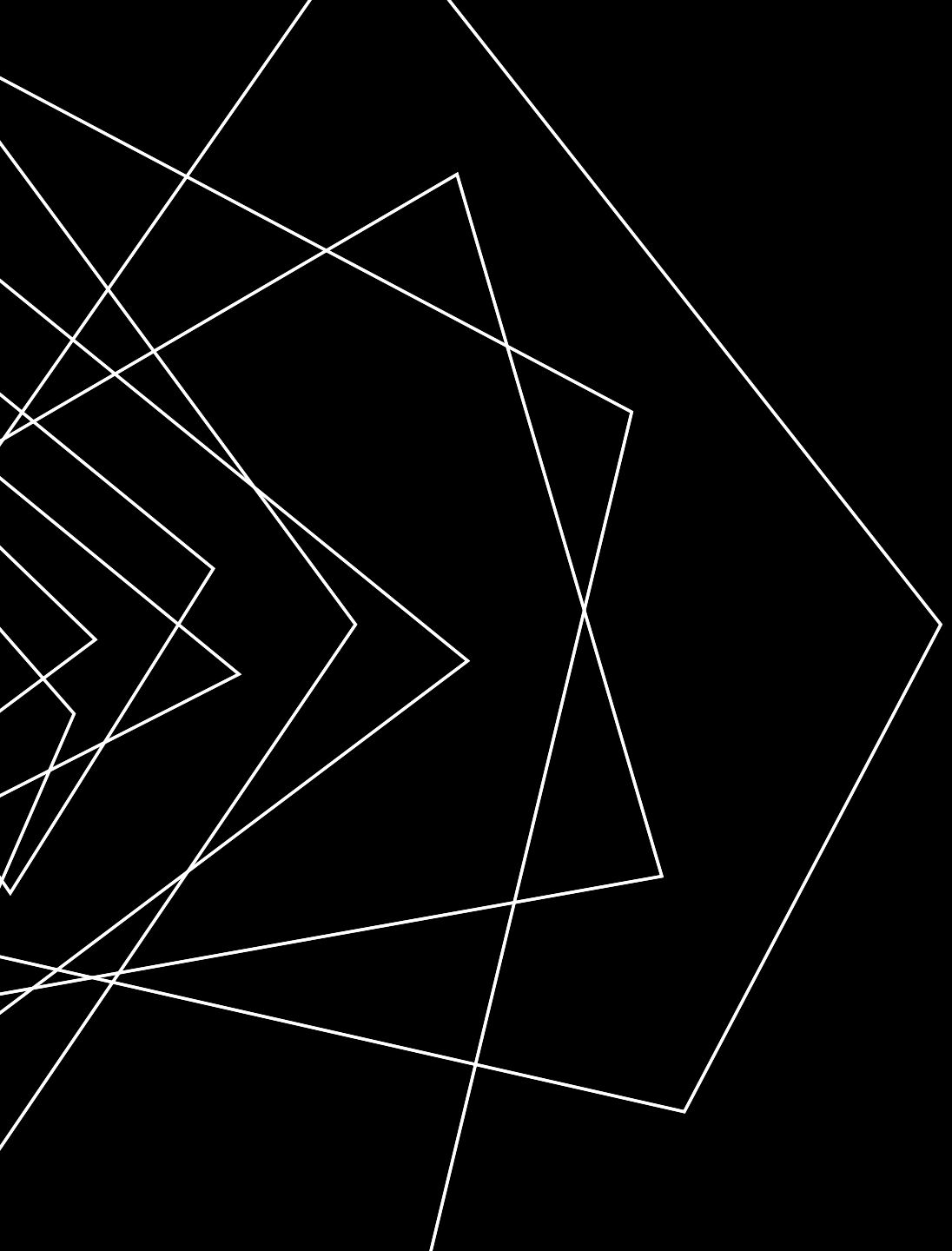
LIMITATIONS & FUTURE WORK

Limitations

- Sampled ~2% of data (memory constraints)
- Missing category metadata → many “Unknown” values
- Collaborative filtering limited due to sparse purchase matrix
- No demographic or textual data

Future Work

- Evaluate session-level funnels
- Add price elasticity / discount impact
- Use deep learning recommender models (LightFM / Neural CF)
- Test time-based cross-validation for stability



THANK YOU

Yazdan Riazi
Student id : 16270796
Email : Myrk5f@umsystem.edu