E-Commerce Customer Behavior Analysis

Introduction

Study Facts :

- **Dataset** : Obtained from Kaggle (https://www.kaggle.com/ecommerce);
- **The goal and context of the dataset:**

  The goal of this dataset is to capture the nuances of user behavior across multiple categories.
- **9 Attributes**: event_time, event_type, product_id, category_id, category_code, brand , price, user_id, user_session
- **300K Records** in the Data Set.

Data description

Column Name	Data Purpose	Data Role	Data Type	Consolidation Type	Definition
event_time	Dimension	Input	DateTime	Time Difference	Time of user's event (UTC)
event_type	Dimension	Target	String	Count, Mode	One type of an action that a user does. [view, cart, remove_from_cart, purchase]
product_id	Dimension	ID	String	Count	ID of a product
category_id	Dimension	ID	String	Count	ID of product's category
category_code	Dimension	Input	String	Count, Mode	Product's category code name.
brand	Dimension	Input	String	Count, Mode	Brand name. Can be missed.
price	Measure	Input	Numeric	Count, Median, Mean, St.Dev., Min, Max	Float price of a product. Present.
user_id	Dimension	ID	String	Count	Permanent user ID.
user_session	Dimension	ID	String	Count	Temporary user's session ID. Same for each user's session. Is changed every time user come back to online store from a long pause.

Data preprocessing

Null Values Checking

1. Null values in two columns were identified:
   a. Category code
   b. Brand
2. They were replaced with:
   a. Category code: Nulls -> ‘undefined accessories’
   b. Brand: Nulls -> ‘undefined brand’

Duplicate Values Checking

1. Three duplicated values were detected.

2. The drop duplicates method was applied to cleanse and refine the data.

Data preprocessing

Outliers Checking

1. The distribution of the single numeric feature in the dataset, ‘price’, was examined.
2. Items with outlying high price were presented.
3. One decided not to drop them, as they don’t cause any abnormal tendencies in terms of customers behavior.

Inconsistent Values Checking

1. The lists out of all values from object data type features were created to make sure that all of them make sense;
2. It was found that items with zero price were presented;
   a. They were not available for purchase, so they were dropped;

Feature Engineering

1. New attributes were created by extracting date, day, month, year, weekday and time details from the “event time” attribute.

Research questions & EDA (visualizations, summarizations)

Finding the most popular products/product categories/brands in terms of the number of purchases

- Can we find the top-selling products?
- What are the product categories that generate the highest revenue?

Analyzing Sales differences over the time

- On what day of the month did most sales happen?
- Are there any specific times of the day or days of the week when purchases are more frequent?
- Is there any Week/Weekend Sales difference?

Research questions & EDA (visualizations, summarizations)

Pricing strategies analysis

- Do lower-priced products sell more frequently than higher-priced products or vice-versa?

Conversion analysis

- What are the most viewed products and do these views lead to satisfying purchase rates?

Research questions & Model Building

Segmentation of users based on behavior

- (K-means Clustering) Are there any segments of customers we can highlight (in terms of spending, conversion rate, number of purchases, number of views, lifetime value)?

Association rule analysis

- (Apriori & FP Growth) Are there any associations between the products that can be discovered using Market Basket Analysis?

Top 5 selling products, categories, brands

Products

category_code	Product _id	Purchase Rate
Electronics .smartphone	1004856	220
	1004767	162
	1004833	90
	1005115	83
	1004870	79

Categories

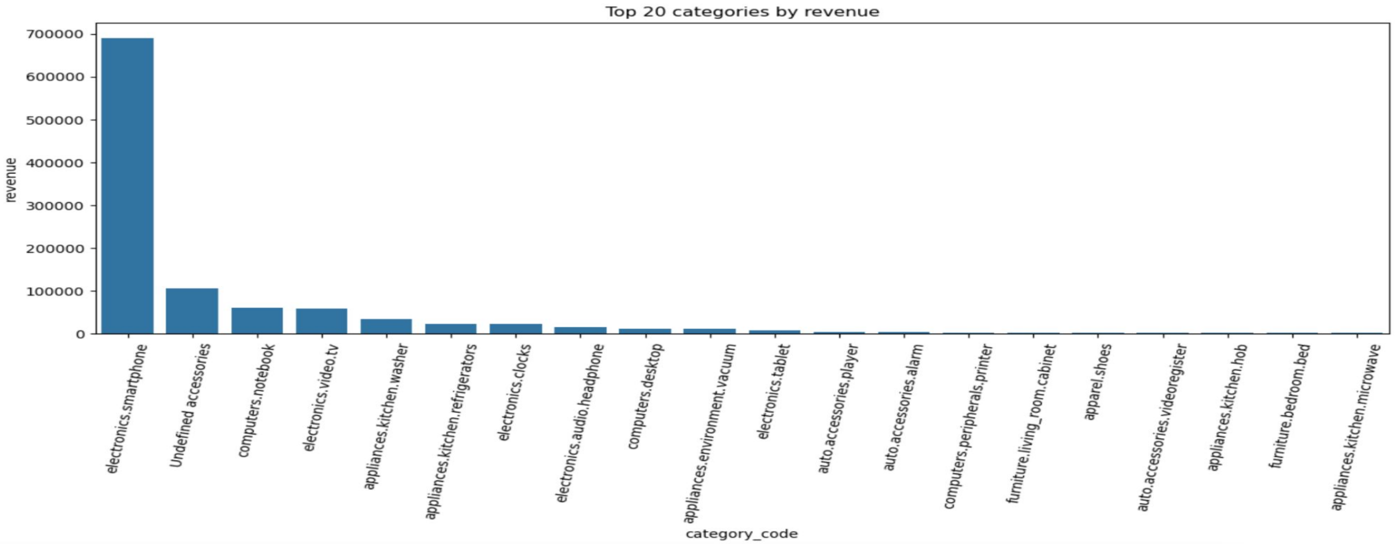
category_code	Purchase Rate
electronics.smartphone	2445
Undefined accessories	1271
Electronics.audio. headphone	204
electronics.video.tv	157
electronics.clocks	137

Brands

brand	Purchase Rate
samsung	1240
apple	1008
Undefined brand	445
xiaomi	415
huawei	173

- We had to use a list and the sort method in order to display the top selling products, categories and brand.
- In conclusion, we found that Electronic Smartphone with the brand Samsung is the top most selling category and brand. Which can be used to find out which product in the market is the most opted or sought after.

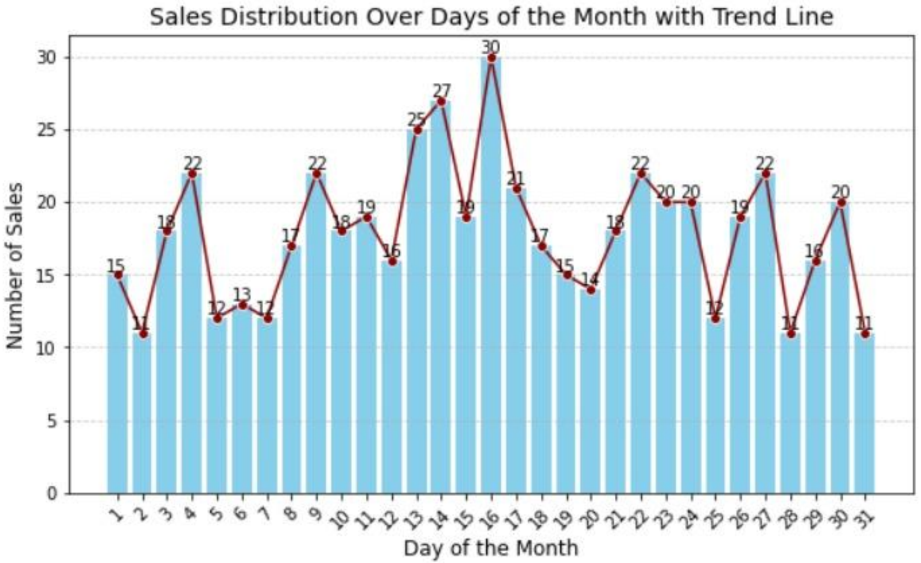
Top product categories by revenue*

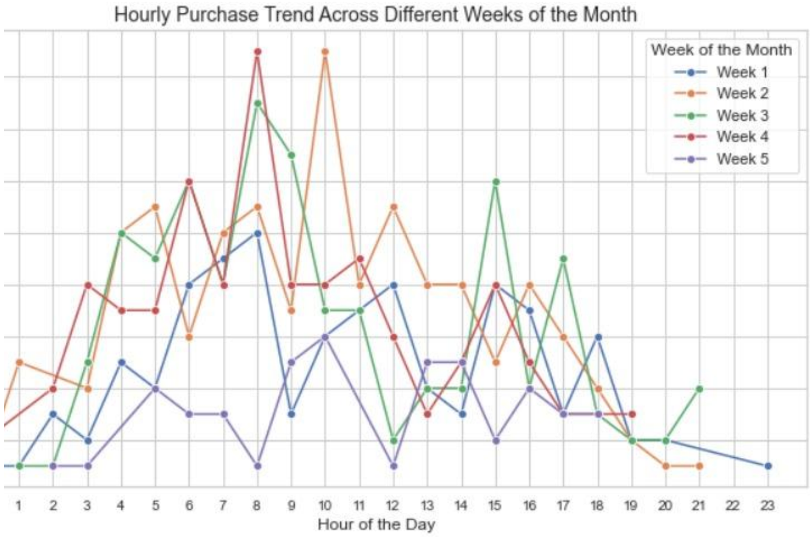


Revenue = Number of Transactions * Median Price of Transactions

- We found out the top product categories by revenue in order to see which product category is dominating the E-Commerce website and how it might affect the revenue generated.
- The E-commerce platform can use this product to boost sales by focusing on these top product categories.

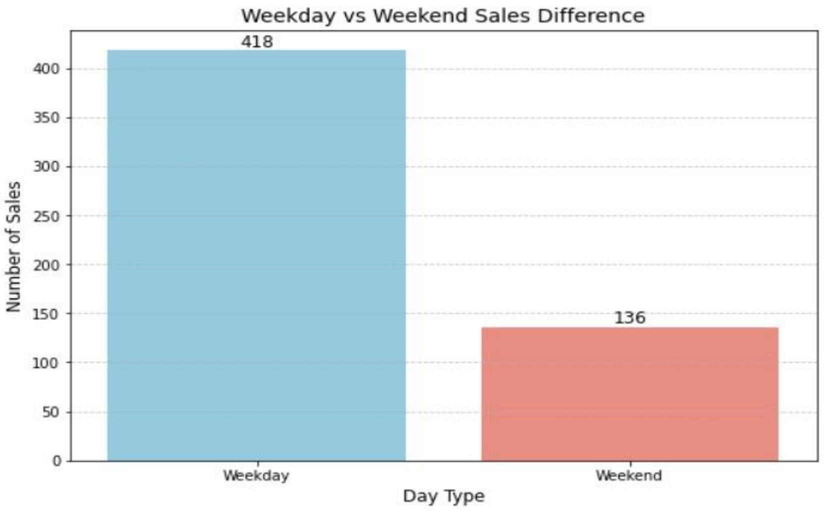
Analysing Sales Differences Over the Time





- Align inventory levels with anticipated demand on specific days, reducing the risk of stockouts and ensuring a seamless shopping experience for customers.
- Tailor promotions and discounts to coincide with peak hours, encouraging increased sales and customer participation during these periods

Analysing Sales (Weekend vs Weekdays)



- Implement weekend-only deals to stimulate purchasing.
- Evaluate weekend pricing or discounts to boost sales.

Conversion Analysis

18 items with top views took place in top purchases

Top 20 Items by Views

category_code	product_id	Views	Purchases	Convers	Convers_rank	Place
Electronics. smartphone	1004856	2938	220	0.075	12	1
	1004767	2807	162	0.058	14	2
	1005115	2328	83	0.036	18	3
	1004249	1398	73	0.052	15	4
	1005105	1361	44	0.032	18	5
	…	…	…	…	…	…

Top 20 Items by Purchases

category_code	product_id	Views	Purchases	Convers	Convers_rank	Place
Electronics. smartphone	1004856	2938	220	0.075	12	1
	1004767	2807	162	0.058	14	2
	1004833	1345	90	0.067	12	3
	1005115	2328	83	0.036	18	4
	1004870	1343	79	0.059	13	5
	…	…	…	….	…..	….

- 90% of top 20 items by views are included into top 20 by purchases;
- Sight-attractive items yield high purchases;

Conversion Analysis

*Conversion = # of Purchases / # of Views

But only 1 item with top views took place in top conversions

Top 20 Items by Views

category_code	Product_id	Views	Purchases	Conversion	Convers_rank	Place
electronics.smartphone	1004856	2938	220	0.075	12	1
	1004767	2807	162	0.058	14	2
	1005115	2328	83	0.036	18	3
	1004249	1398	73	0.052	15	4
	1005105	1361	44	0.032	18	5
	…	…	…	…	…	…

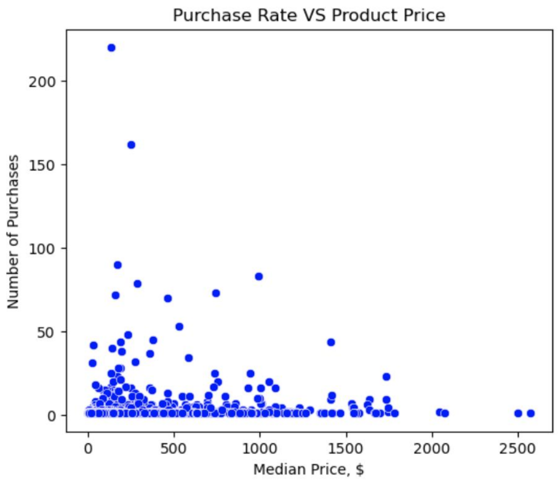
Top 20 Items by Conversion* (>= 98th views percentile)

category_code	Product_id	Views	Purchases	Convers	Convers_rank	Place
	…	…	…	…	…	…
electronics.smartphone	1004856	2938	220	0.075	12	16
	1005143	53	4	0.075	12	17
undefined accessories	12701602	43	3	0.07	12	18
	26400194	57	4	0.07	12	19
electronics.smartphone	1005113	72	5	0.069	12	20

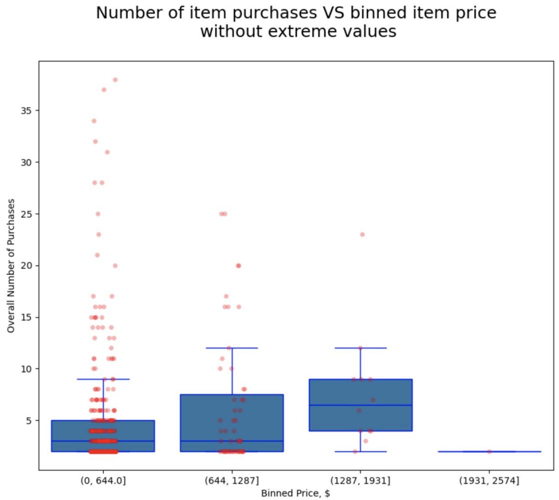
- Only 5% of top 20 items by views are included into top 20 by conversion;
- Sight-attractive items don't yield high conversion;

Pricing Strategy Analysis

The purchase rate against price per item



- Prices binning;
- Filtering out zero-purchases;
- Filtering out extreme values (1th, 99th percentile).



- No interesting tendency except the 3rd bin (sharp incline of the purchase rate mean).
- Explanation: Dealers of high-price products can't stay with bad-sellers on market unlike cheap-goods dealers.
- Reason: Markup = Gross Profit/Sales Price.
- Example: AVG Markup of HDMI Cables ~ 1000%, AVG Markup of Smartphones ~ 79% [source].

Pricing Strategy Analysis

Reinforcement of the explanation

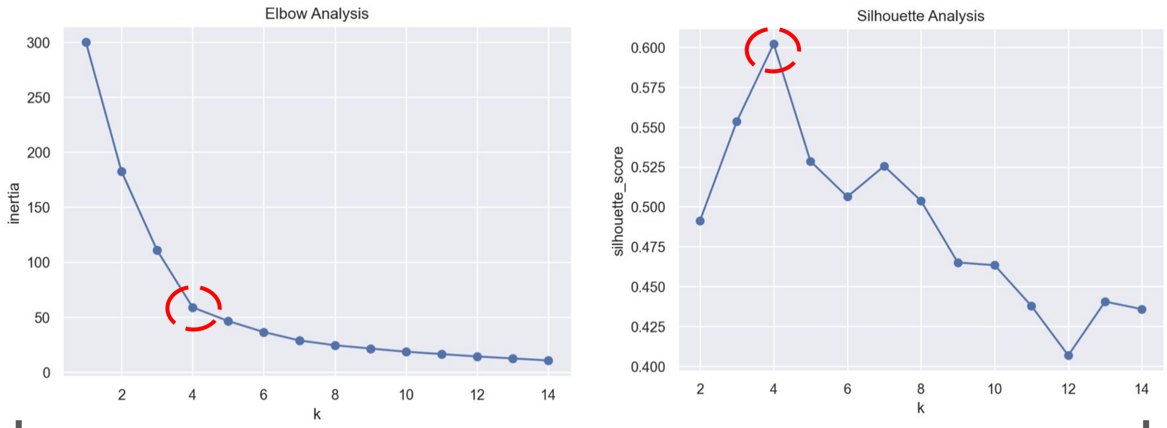
category_code	product_id	event_type_purchase	price	price_binned
Electronics.smartphone	1005106	7	1516.214313	(1287, 1931]
	1005124	9	1641.785616	
	1005129	9	1407.409	
	1005130	6	1611.706667	
	1005132	12	1415.847844	
	1005133	3	1621.781512	
	1005135	23	1731.114	
	1005138	2	1744.514634	
	1005140	4	1521.914118	
	1005141	4	1725.206354	
	1005143	4	1546.449831	
	1005144	9	1723.266185	

- All items in the 3rd price bin are smartphones.

Customer Segmentation

The approach description

Regular Customers		
Total Spend.	Convers. Rate	CV*
8454.46	0.18	66.05
2200.56	1.00	241.04
…	…	…

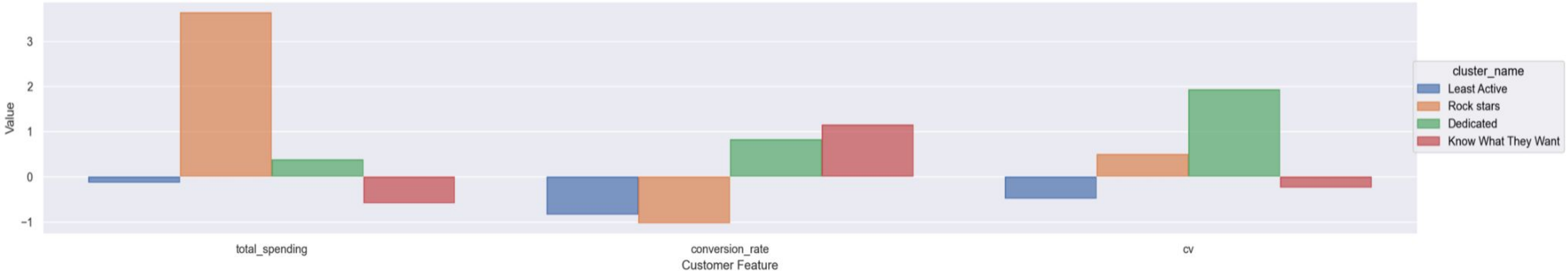


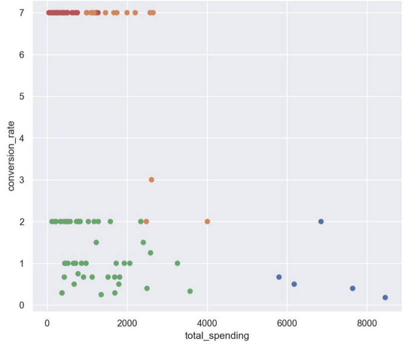
Number of Clusters = 4

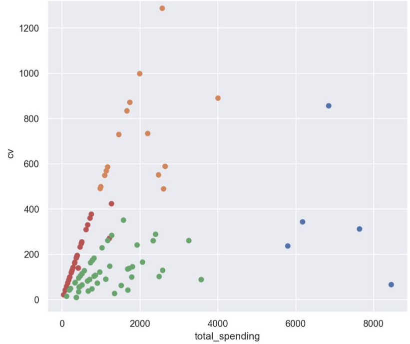
*Customer Value = Avg Transaction Price * Avg Number of Transactions

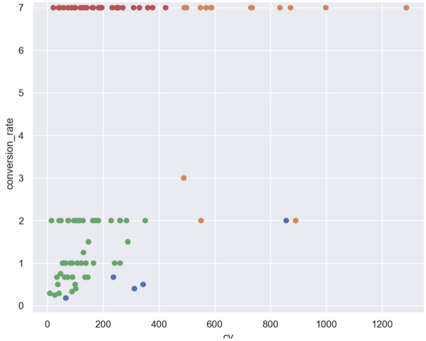
Customer Segmentation

The visual description of the customer k-means clustering





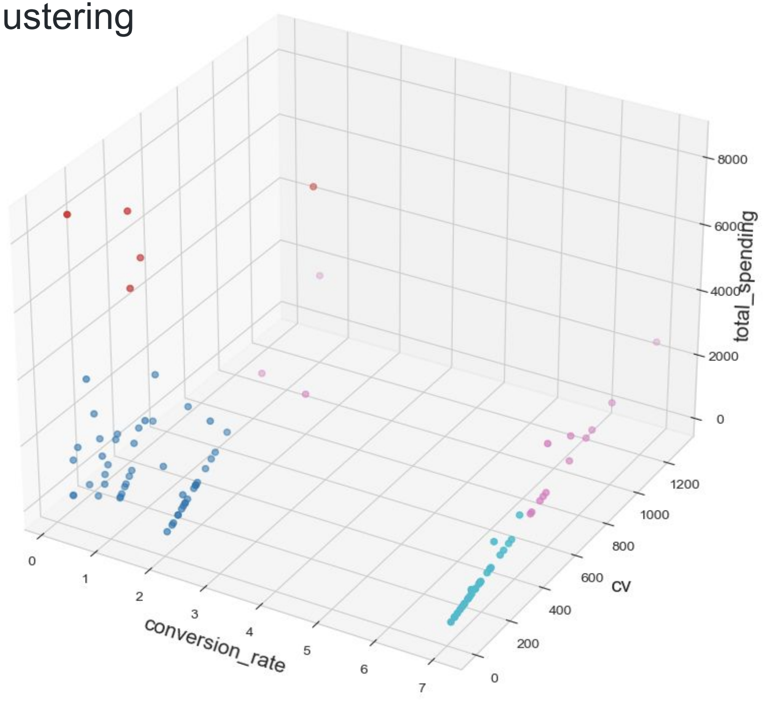




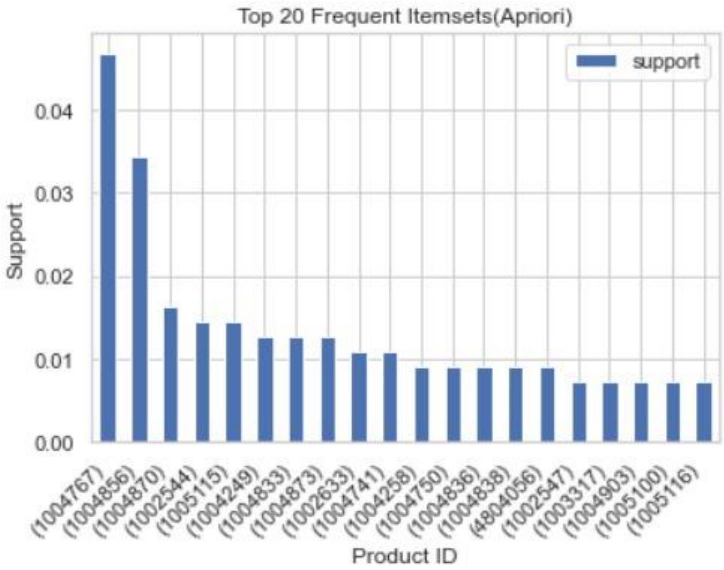
Customer Segmentation

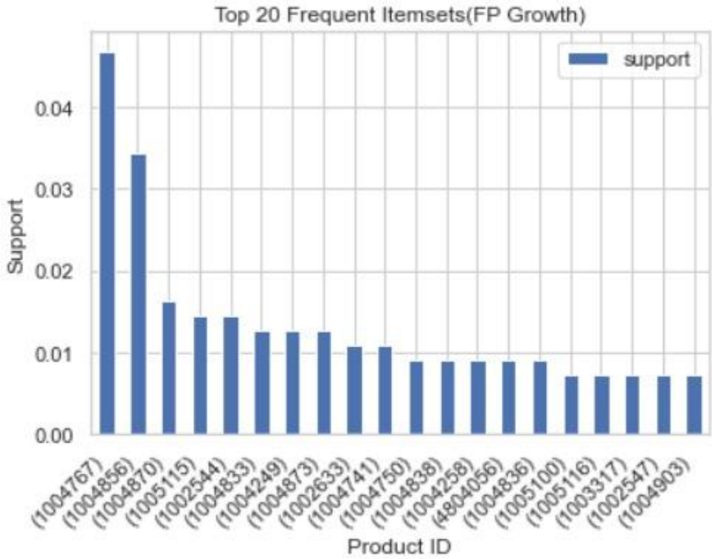
The visual description of the customer k-means clustering

- 4 customer clusters associated with distinctly different behavior patterns were formulated.
- The e-commerce platform can treat them differently in terms of user experience design and marketing campaigns.



Market Basket Analysis





- Identified the products frequently bought together to optimize cross-selling strategies, suggesting complementary items to customers and potentially increasing average order value

Market Basket Analysis

Association Rules:

	antecedents	consequents	antecedent support	consequent support	support
0	(1004659)	(5701166)	0.003155	0.001113	0.000186
1	(5701166)	(1004659)	0.001113	0.003155	0.000186
2	(1004873)	(1005129)	0.008350	0.001670	0.000186
3	(1005129)	(1004873)	0.001670	0.008350	0.000186

	confidence	lift	leverage	conviction	zhangs_metric
0	0.058824	52.833333	0.000182	1.061317	0.984177
1	0.166667	52.833333	0.000182	1.196215	0.982166
2	0.022222	13.306173	0.000172	1.021019	0.932635
3	0.111111	13.306173	0.000172	1.115606	0.926394

- We can maximize revenue potential by capitalizing on product associations, driving sales through strategic product placements and targeted marketing efforts

Discussions - Conversion Analysis

Limitations and concerns:

- We **subsetted 300K rows** out of the original file and analyzed the **single month** only;

Usage:

- That's a **hint** both to dealers and the e-store that
    - Hight views-rate yields high purchases but doesn't define high conversion;
    - Dealers of eye-catching items should ponder on how to encourage viewers to buy.

Discussions - Pricing Strategy Analysis

Limitations and concerns:

- As we **subsetted 300K rows** out of the original file and analyzed the **single month** only;
- The explanation through the Markup is just an **assumption**.

Usage:

- That's a **hint** both to dealers and the e-store that
  - The price doesn't define the purchase rate of an item;

Discussions - Customer Segmentation

Limitations and concerns:

- As we picked 300K random rows randomly, the number of regular customers is relatively small compared to original dataset.
- Customers with **infinity conversion** were found; this value was replaced with high integer number;

Usage:

- Customer segmentation model can be implemented into **recommendation system** on the e-commerce platform or might be of use in **advertising and UX**.

Conclusions

What we have accomplished:

- Conducted **Conversion Analysis** and formulated a **weak point** dealers of sight-attractive items should focus on;
- Analysed **Pricing Strategies** and explained why there is **no explicit correlation** between price and purchases;
- Conducted **Customer Segmentation** and formulated **recommendations** to the e-store on how to use this model.
- In-depth **Time Series Analysis** was conducted to unreveal sales patterns to show insights of customer behavior over specific time periods.
- Applied **Market Basket Analysis** to uncover support(frequency) of top products and tried to find out the association rules.
- Conducted all the above mentioned analysis for effective market strategies to enhance business outcomes.