

# eCommerce Customer Segmentation

---

Dcard Data Analyst Intern Take Home Quiz - Chu, Po Yen



# Content

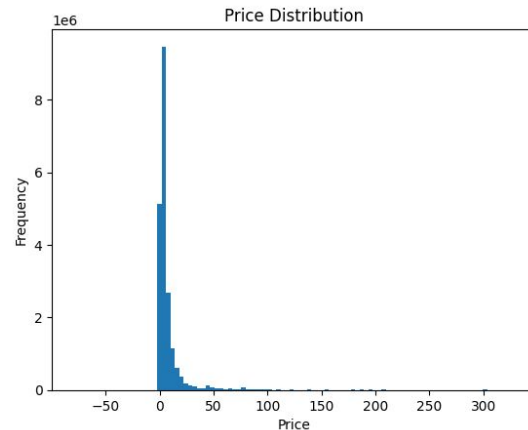
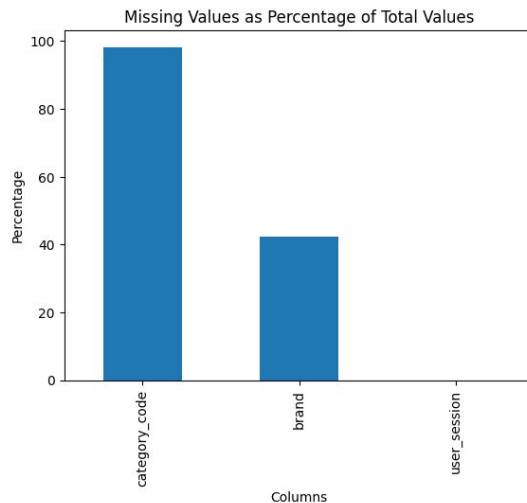
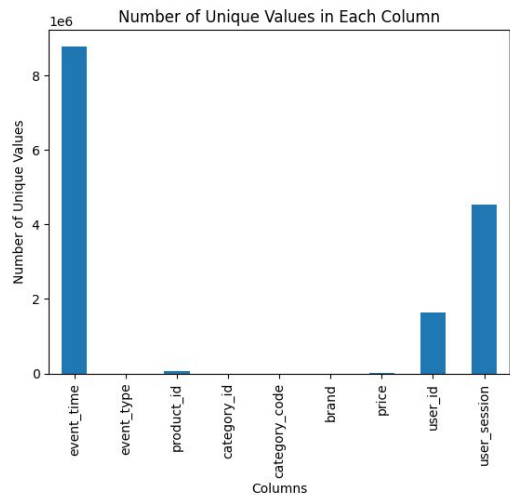
1. Target
2. EDA
3. RFM Analysis
4. K-means Clustering
5. Market Basket Analysis
6. Client Persona
7. Conclusion
8. Limitations



# The overall data quality is good, no complex preprocessing needed



## Key Takeaways



Only four columns contain missing values

*category\_code* provides description of *category\_id*, however the missing values make this column useless

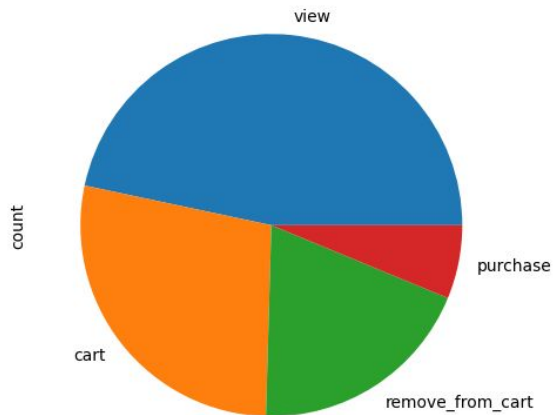
Negative price should be eliminated

The pricing structure tends to follow a log distribution, aligning well with the dynamics of eCommerce

# No dominant brand and category exceeded 15%, while increasing conversion rate is priority



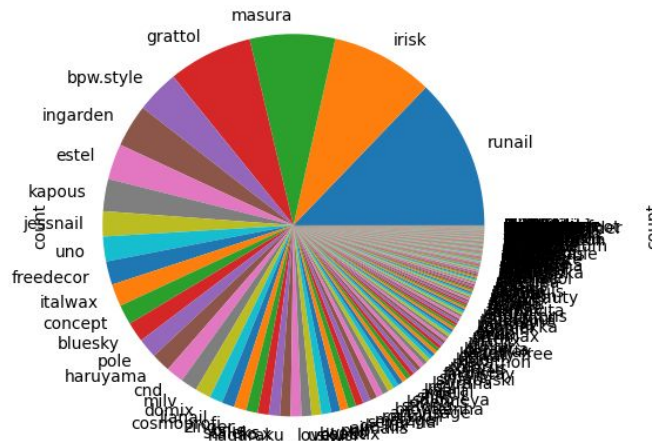
Event Type Distribution



Event type == 'purchase' demonstrates as a significantly smaller portion

Increase conversion rate from view-to-cart can also be an issue

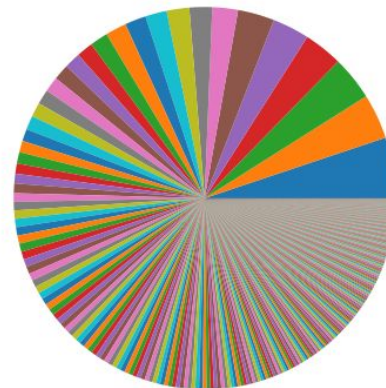
Brand Distribution



No brand and category exceeded 15%, though this platform is specific for cosmetics. Customers' disperse preference may result in difficulties on promotions or brand cooperation

It appears to be a Russian platform featuring numerous local brands after conducting research

Category ID Distribution



# Analyzing temporal shifts in online behavior



## Key Takeaways

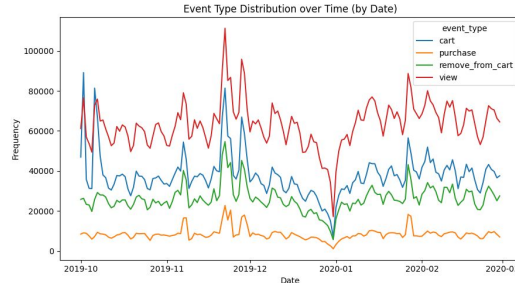
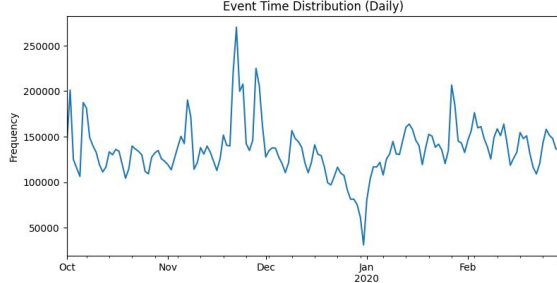
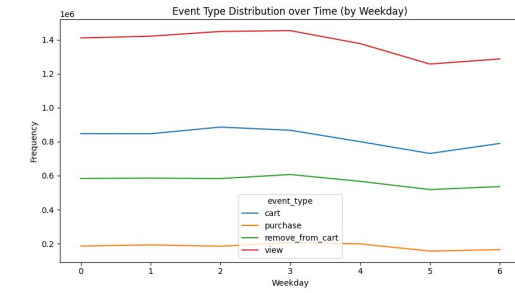
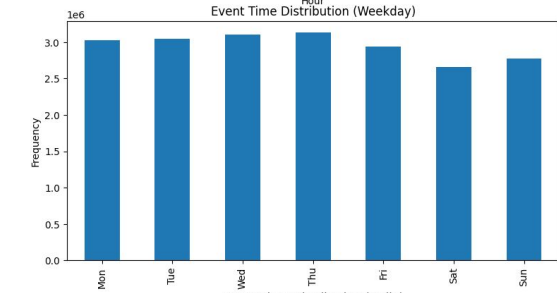
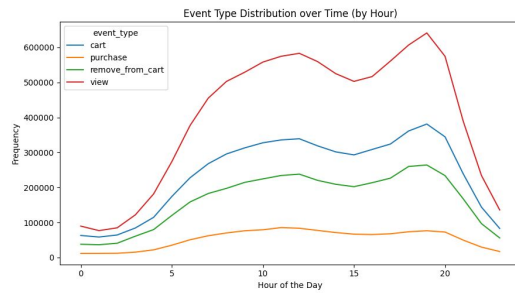
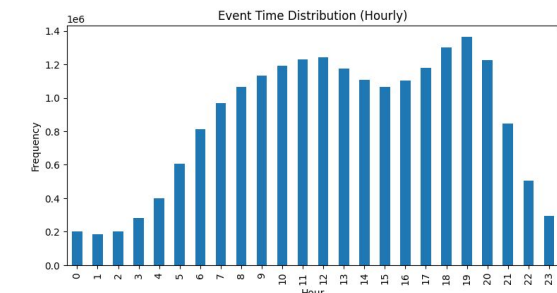
All event types exhibit similar fluctuations without any noticeable abnormal growth or decline

Purchases didn't decrease as significantly during troughs, while they increased substantially during peaks. This pattern suggests that there may be steady demand or essential purchases driving the observed fluctuations

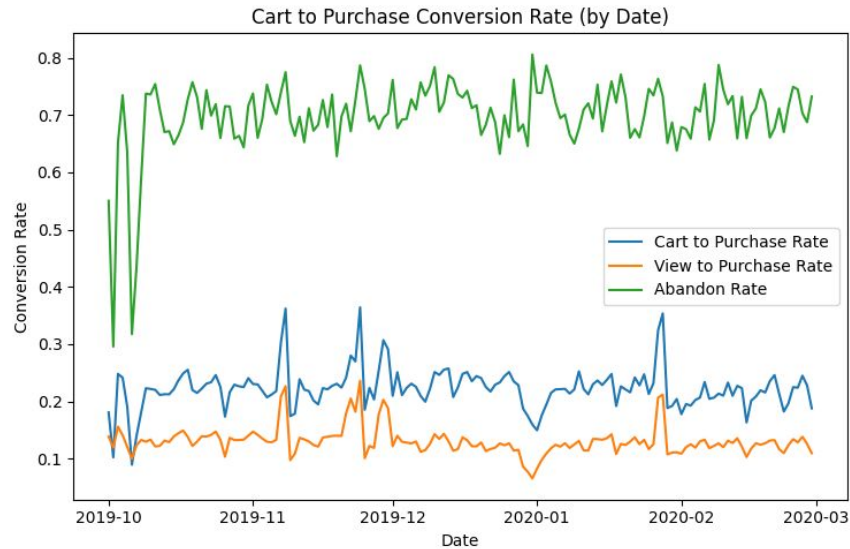
The anomaly of more carts than views in October seems improbable, but since the cause cannot be identified, I have chosen not to address it

1111 & Thanksgiving Day likely featured a promotion, as evidenced by the notable increase in purchases during that period

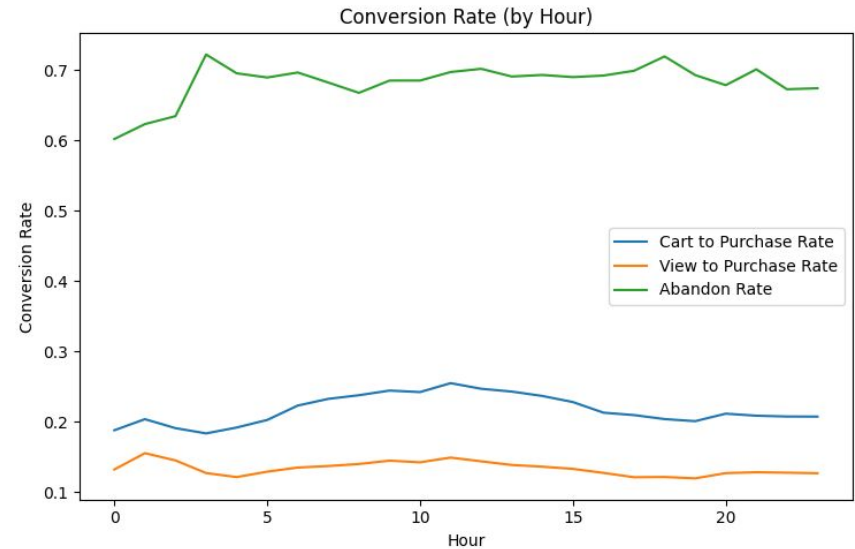
Long holidays (Christmas - New Year) is observed a reduction in online purchasing, possibly due to online stores or logistics systems being closed for the holiday period



## Promotions did increase conversion rate, and impulse purchasing mostly happens in midnights (1-2 a.m.)



Given the assumption of promotions occurring on 1111 and Thanksgiving, there was indeed an increase in conversion rates during these promotional periods



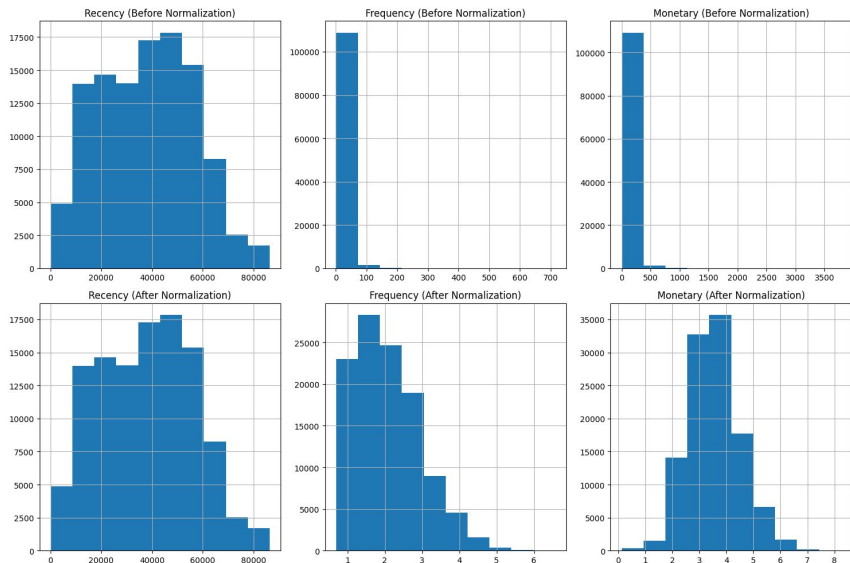
Conversion rates surge in the midnight hours, particularly between 1-2 a.m. My speculation is that customers tend to engage in impulse purchasing before going to sleep

# RFM Analysis: Filter out purchase records & cluster by K-means



## Preprocessing

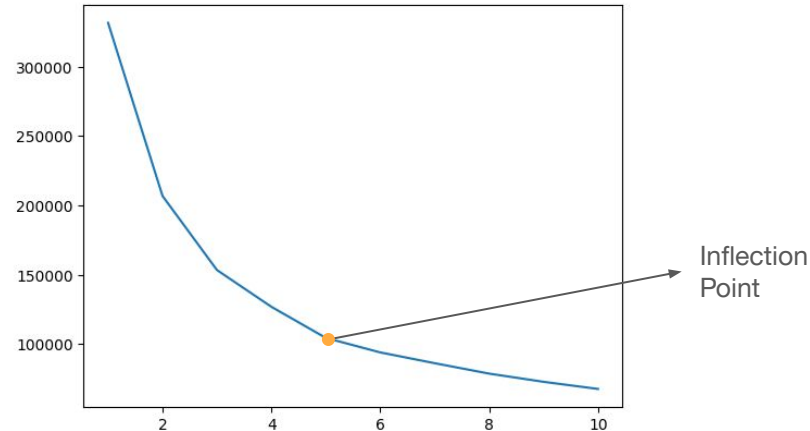
1. Filtered out 'purchase' records
2. Discovered that frequency and monetary data is severely skewed
3. Normalize the data is by log



## Find Optimal N\_Clusters (Elbow Point)

Since the data is not normally distributed, it is hard to do customer segmentation just by RFM index, I chose to conduct K-means clustering instead.

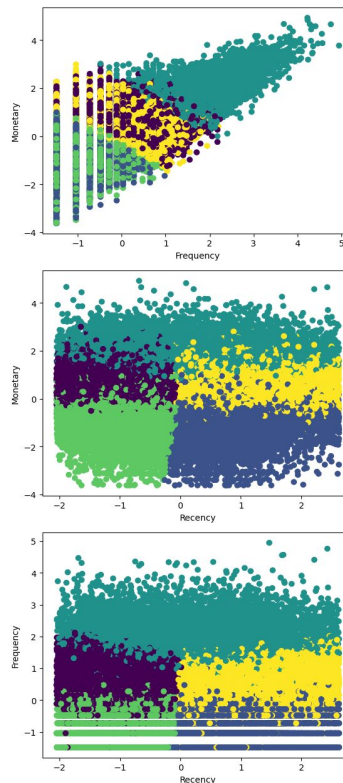
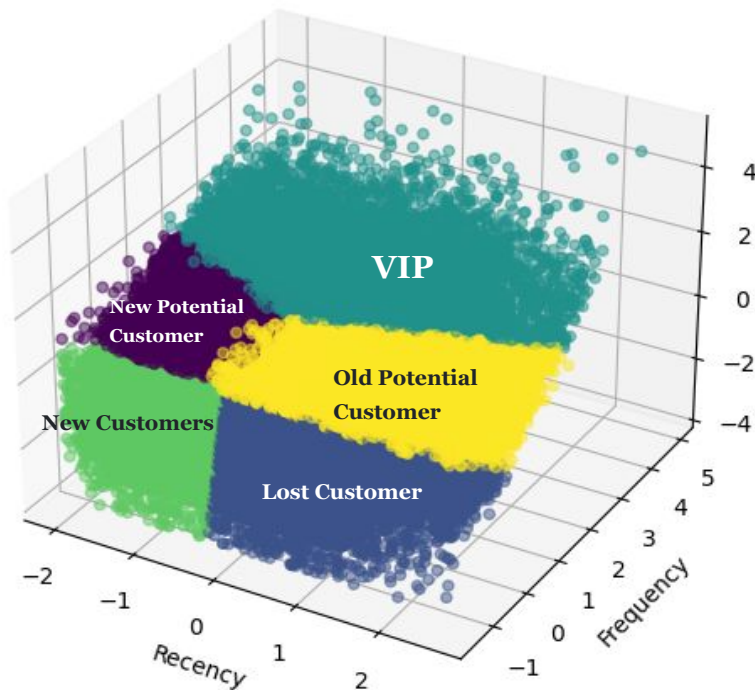
After standard-scaling, conduct K-means in different n\_clusters, discovering that **5** is the optimal number (the elbow method: find the inflection point where the slope start to become mild)



# K-means Clustering into 5 customer types



Define each cluster:



## Key Takeaways

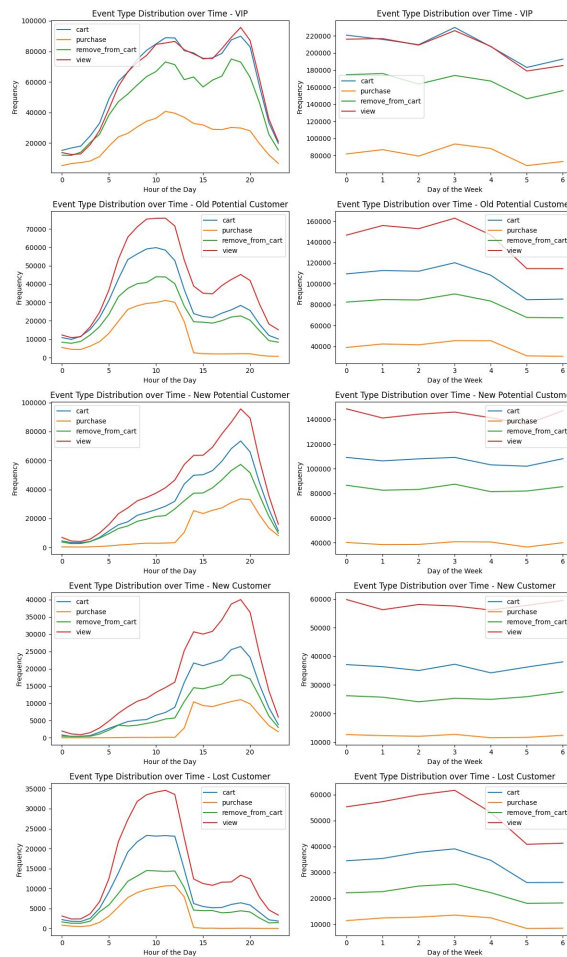
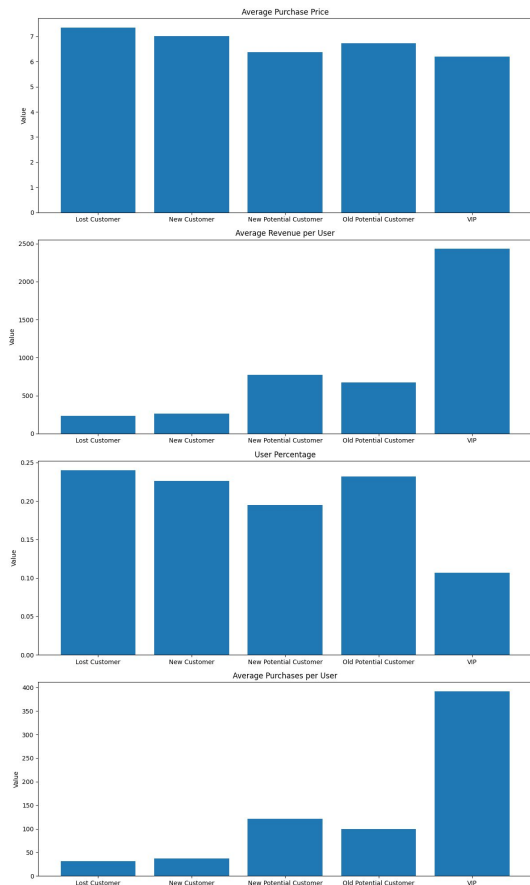
No significantly low purchase price users according to the Frequency-Monetary scatter plot, while high-end consumer exists

"Acquiring a new customer can be five times more expensive than retaining an existing one." Thus, this concept should be kept in mind when we do further analysis

Our later analysis assume that groups classified as old customers (or lost customers) will no longer use the platform



# Targeting recent customers and VIPs with night and weekend engagement



## Key Takeaways

If increasing profit is the top goal, combined with the concept mentioned in the last slide, we should gave up old potential customers and lost customers (assuming they no longer use the platform) as they also did not perform five-time greater revenue

Revenue variation mainly depends on frequency instead of purchase price

VIP group should be tried to retain regardless of their recency, their average value per user is so high that it worth the cost)

Recent customers tends to be active at night and weekends. My guess is that old customers came for the previous promotion, might averaged an advanced age

# K-means Clustering: Investigate User Behavior



## Feature Selection

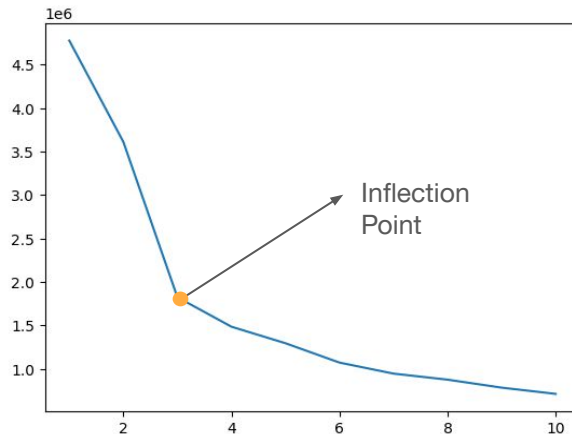
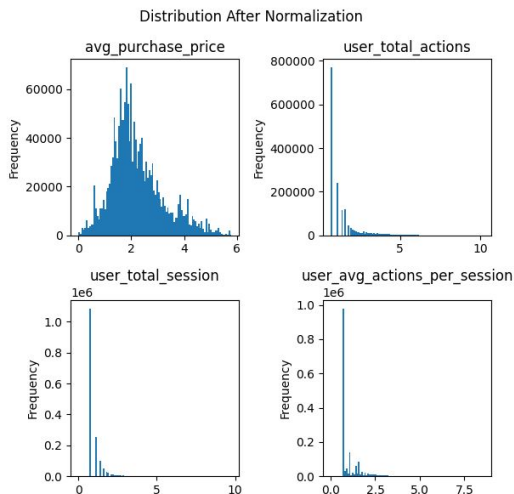
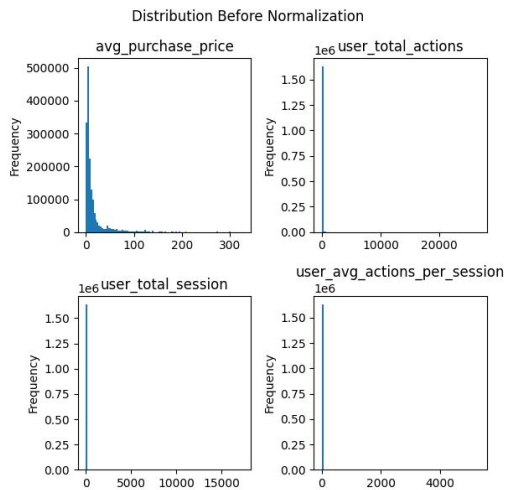
Choose 4 numerical columns that reflects the 'surfing behavior' of users, mostly focus on frequency and avg\_purchase\_price

## Preprocessing

1. Consider all kinds of event type
2. All data is severely skewed, normalize the data is by log
3. use standard-scaler to ensure normality

## Find Optimal N\_Clusters (Elbow Point)

Conduct K-means in different n\_clusters, discovering that **3** is the optimal number (the elbow method: find the inflection point where the slope start to become mild)



# K-means Clustering Results

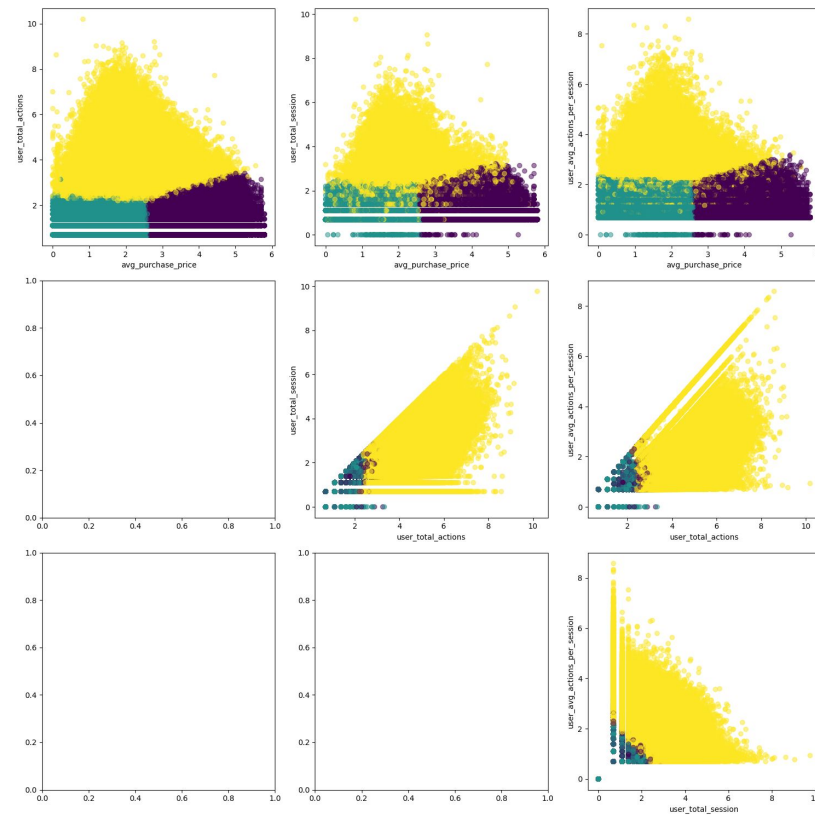
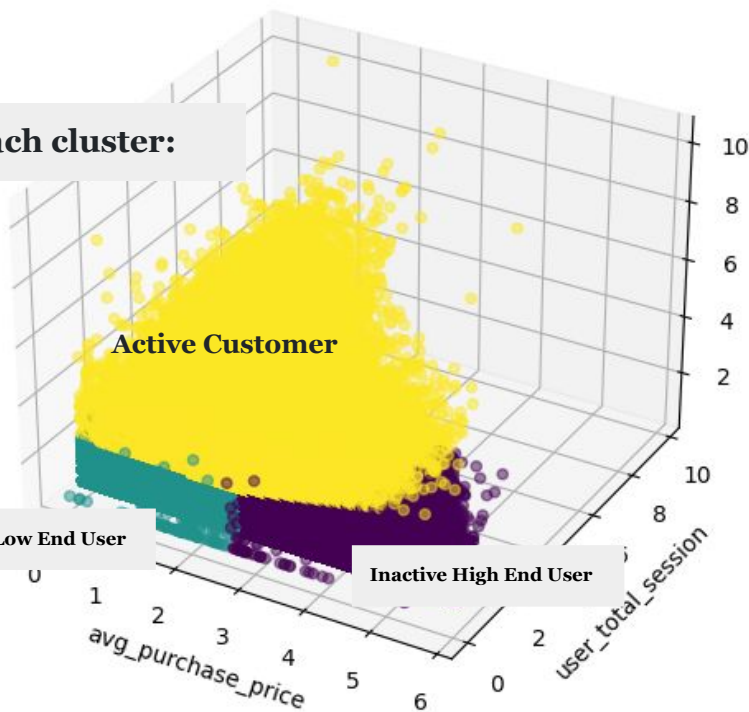


Define each cluster:

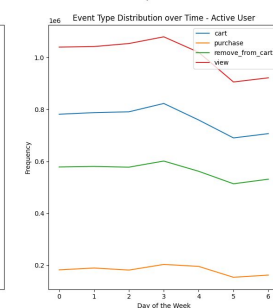
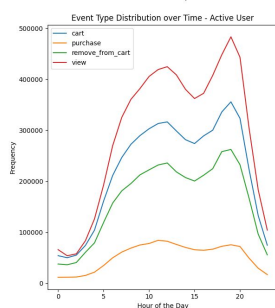
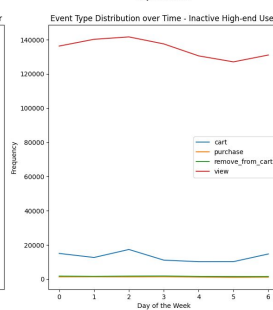
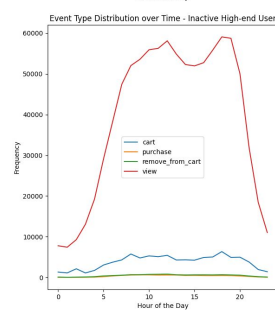
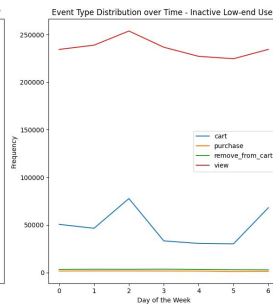
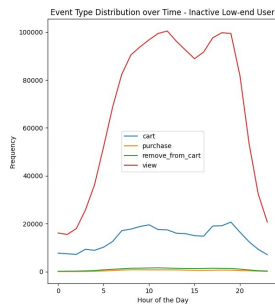
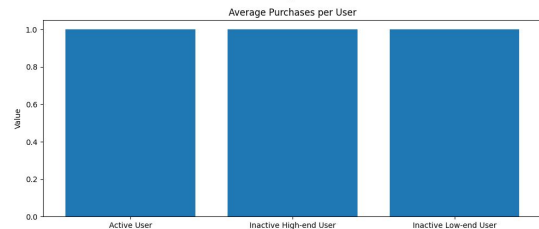
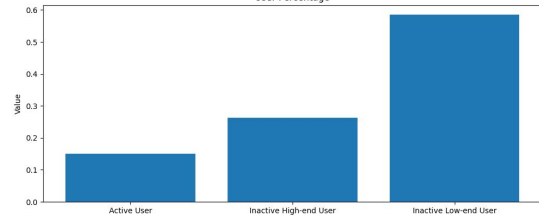
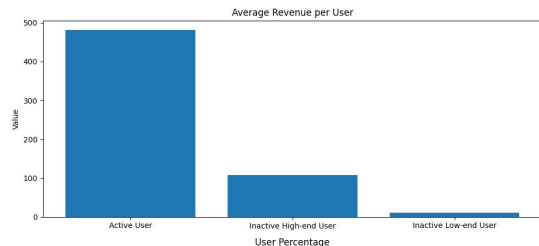
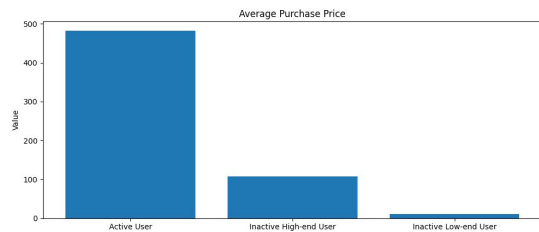
**Active Customer**

**Inactive Low End User**

**Inactive High End User**



# Engaging in interactive online activities could further boost profits



## Key Takeaways

Active users contribute the highest average revenue and price, as this encompasses all browsing activities

Inactive users yield a low conversion rate, yet clustering results may introduce significant bias

Low-end users are more likely to put items in cart, but not purchasing them

Low-end users also have more actions comparing to high-end ones

**Engaging in interactive online activities could further boost profits**



# Conclusion, Suggestions & Limitation Summary



Conclusion	Suggestions	Limitations
The activity level of users strongly correlates with profits (not only purchases). Also, the significance of VIP and recent customers should override old & lost customers as retaining them creates more profits.	1. Encourage engagement in interactive online activities 2. Given that recent customers are more active during nights and weekends, scheduling promotions and campaigns during these times should be considered	Considering the skewed distributions and presence of outliers, K-medoids may offer a more robust approach, as for datasets with categorical variables where K-modes could be considered. However, constraints on time and memory complexity prevent their immediate application in the current scenario
Association rules mainly compose of several dominant brands	Promotions of bundling dominant brands could increase niche profits	Market basket analysis may provide limited insights due to the absence of descriptive data  Sequential analysis can be conducted in future research for understanding and predicting customer behavior, such as forecasting demand surges or analyzing the number of items in a cart before purchase using techniques like LSTM