Analysing Online Shopper Behaviour: A Hybrid Clustering Approach

Mehmet Furkan Akpinar

July, 2025

1. Introduction

This project aims to create segments of e-commerce customers based on their purchasing, viewing, or cart behaviour and examine their purchasing habits.

The dataset used in this project is “E-commerce Behaviour Data from Multi Category Store”, which is available on Kaggle and contains 2019 October – 2020 April online customer behaviour data. However, only the 2019 October and 2019 November datasets have been used in this project due to hardware limitations.

The main focus of this project is to segment customers based on their online behaviour—such as viewing products, adding items to cart, or making purchases—to identify distinct user profiles. This segmentation allows for a deeper understanding of customer types and supports more targeted and timely marketing strategies. In addition to behaviour-based segmentation, the project also explores when customers are more likely to make purchases, particularly across different times of a day.

Unfortunately, the dataset features are limited due to the absence of demographic information. Research shows that customers who are working from home tend to make online purchases more often, and it has been revealed that older generations may show distinct time preferences centred on convenience and urgency. Therefore, any demographic information could leverage the analysis results and needs to be addressed for better evaluation in future research.

2. Objective and Technical Approach

The primary objective of this project was to segment e-commerce customers based on their behavioural patterns to uncover actionable insights that can inform marketing strategies and improve user engagement. Both unsupervised learning techniques and manual segmentation methods were utilised to identify meaningful customer groups. The overall approach consisted of the following key steps:

- **Data Cleaning and Preparation:**

  The dataset was filtered to include only relevant columns (e.g., user_id, event_type, event_time) and cleaned to remove null values and duplicates. Event_time fields were converted into usable formats, and new features were derived to represent behavioural patterns (e.g., session duration, purchase hour).

- **Feature Engineering:**

  Custom features such as average cart count, average purchase count, total amount spent, session duration, conversion rates, and purchase hour distribution were calculated to summarise customer behaviour at the user level.

3. Exploratory Data Analysis (EDA)

As part of the data exploration process, we started by combining two datasets from October and November 2019. After merging, the dataset contained over 4 million event records. Some columns, like category_code and brand, had too many missing values, so we decided to leave them out of the analysis. Similarly, any rows missing user_session information were removed to make the dataset cleaner and more reliable for analysis.

We converted the event_time column into a datetime format and sorted the records by time. This was a key step for doing time-based analysis and understanding how user behaviour changes over time.

Then, we focused on four main types of user actions: view (browsing a product), cart (adding to cart), remove_from_cart, and purchase. Looking at the distribution, viewing products was by far the most common action, while purchases were much rarer.

We also looked at the price distribution and noticed some outliers, which we kept in mind during visualisations and further analysis.

This initial exploration gave us a solid foundation for understanding how users interact with the platform and helped us move forward with deeper insights.

4. Segmentation & Insights

To uncover useful insights and help shape marketing strategies, we aimed to group users into meaningful segments. We used both machine learning and rule-based (manual) approaches to do this.

4.1.K-Means Clustering

We initially applied K-Means clustering using features such as purchase count, cart count, and total spending. While the clusters appeared visually balanced and lacked strong behavioural contrast at first glance, the silhouette score of 0.57 suggests a moderate level of separation between segments.

This indicates that, despite the perceived similarity between groups, there may be underlying patterns differentiating user behaviours that are not immediately visible without further profiling.

clusters	user_id	total_spent	cart_count	purchase_count	avg_purchase_price	conversion_rate
	None Count 0	None Count 0	None Count 0	None Count 0	None Count 0	None Count 0
	535807136.0 – 541937408.0	91.6 – 11606.0	3.1 – 14.5	0.6 – 17.9	59.5 – 920.6	0.1 – 3.3
0	541,059,985.87	91.60	3.11	0.63	59.51	0.08
1	539,859,664.30	2,176.21	4.54	2.90	920.55	0.72
2	541,937,397.28	587.72	3.35	2.78	218.14	1.03
3	535,807,124.37	11,606.02	14.54	17.91	811.65	3.33

4.2.Rule-Based Segmentation

To create more meaningful groups, we switched to a manual segmentation based on user behaviour. Users were grouped into five categories based on their purchase and cart activity:

1. **Browsers Only**

   Users with purchase_count == 0 and cart_count == 0. These users just viewed products but didn't interact further.

2. **Added to Cart, No Purchase**

   Users with cart_count > 0 and purchase_count == 0. They seemed interested but didn't complete a purchase.

3. **One-Time Buyers**

   Users who made exactly one purchase. These are likely low-loyalty customers.

4. **Repeat Buyers**

   Users with more than one purchase. These are our loyal customers. We also calculated their average time between purchases.

5. **Non-Buyers (Combined Group 1 + 2)**

   All users who didn't make any purchases.

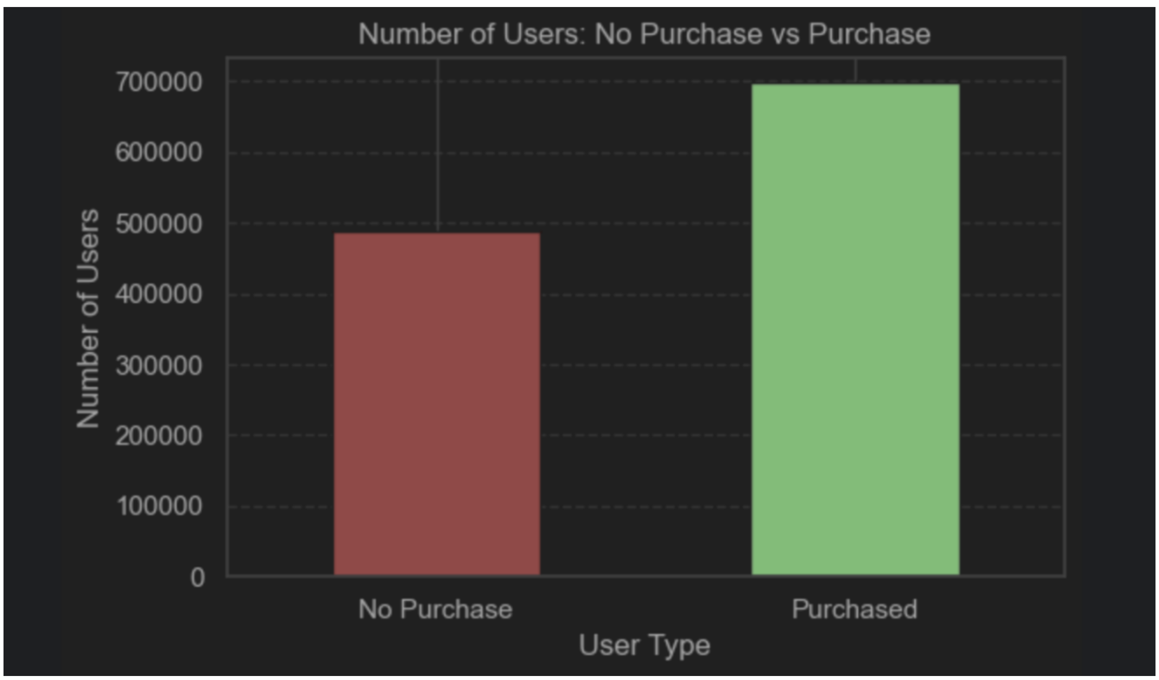
For each segment, we looked at key metrics like the number of users, average spend, and session count. This helped us generate actionable insights and spot behavioural differences between groups, like session counts.

5. Findings

Here are some of the key insights we found through our analysis:

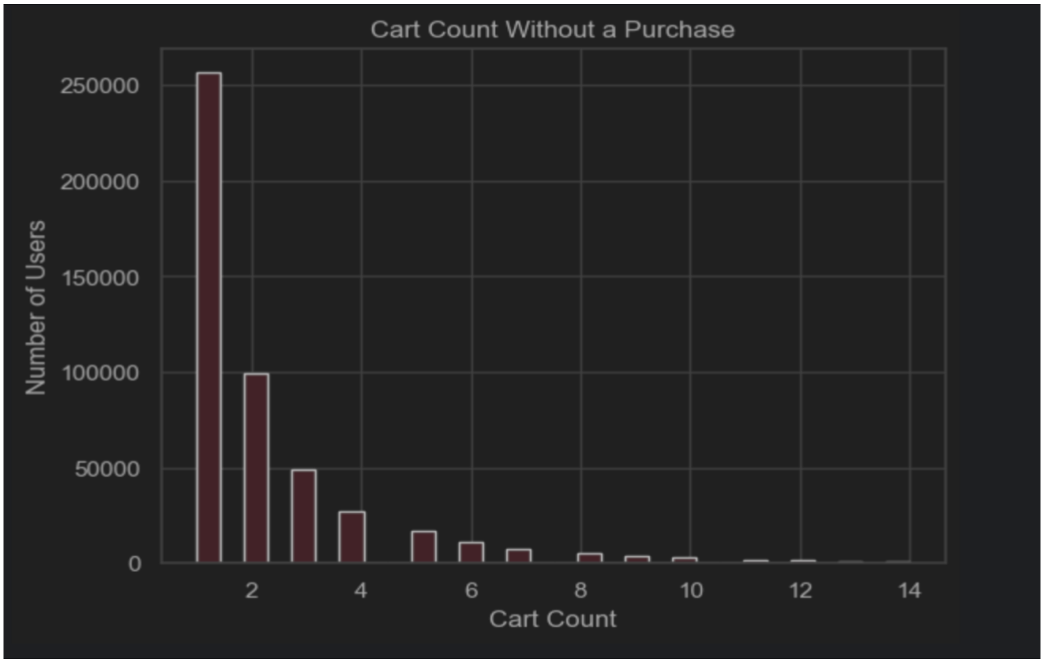
5.1 Purchase Distribution

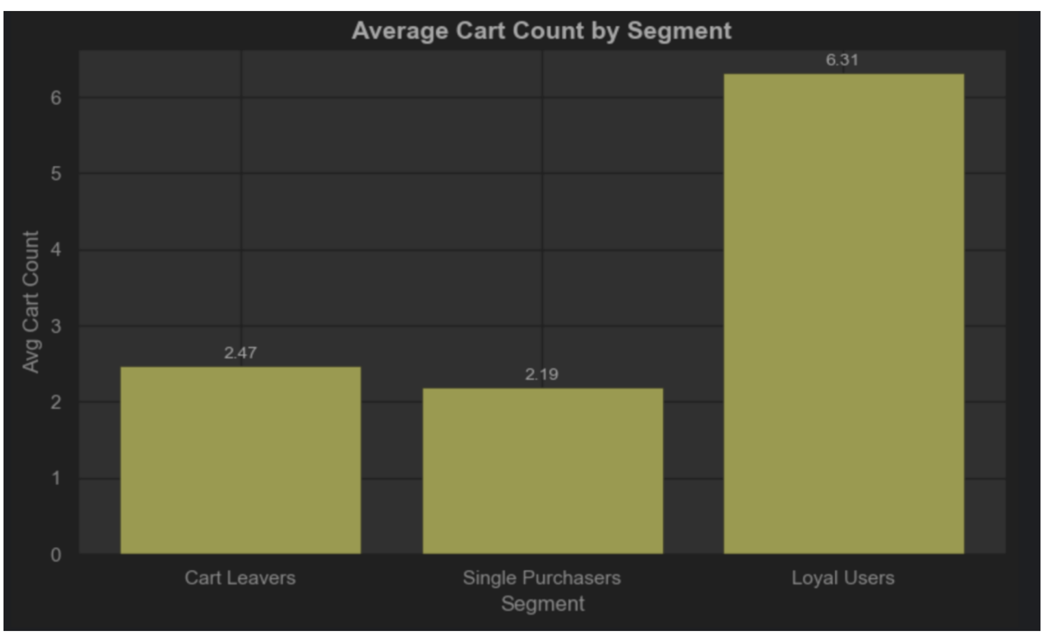
- Most users didn’t purchase at all.
- A large portion of those who did buy only purchased once.
- The repeat buyers, though fewer, contributed the highest total spend.



5.2 Cart Behaviour

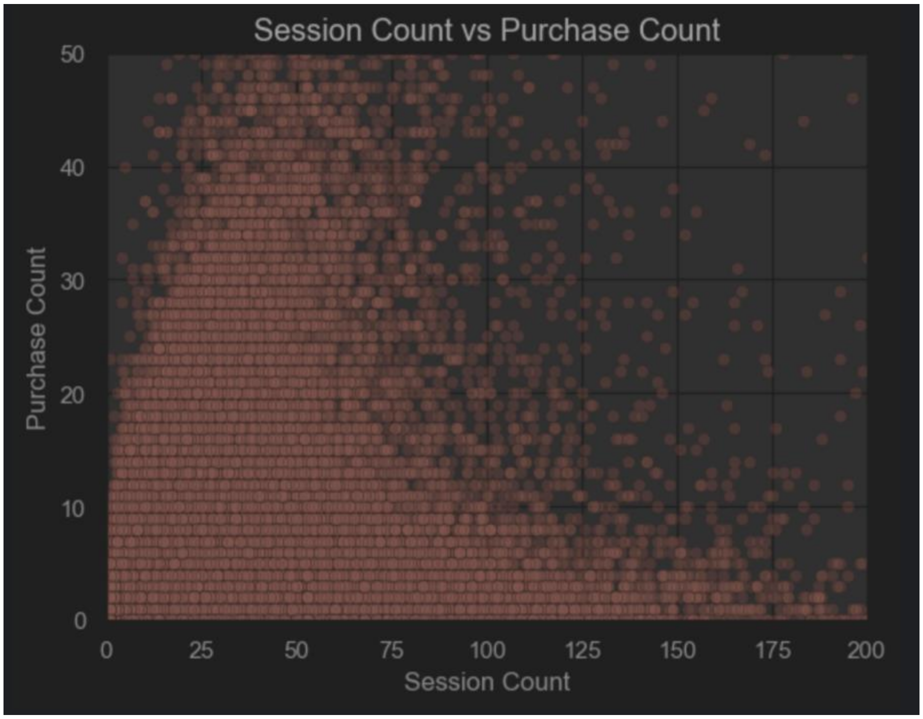
- Many users added items to their cart but never bought them.
- These users clearly showed intent, but something stopped them from completing the purchase.
- This group could be targeted with campaigns to convert them into customers.

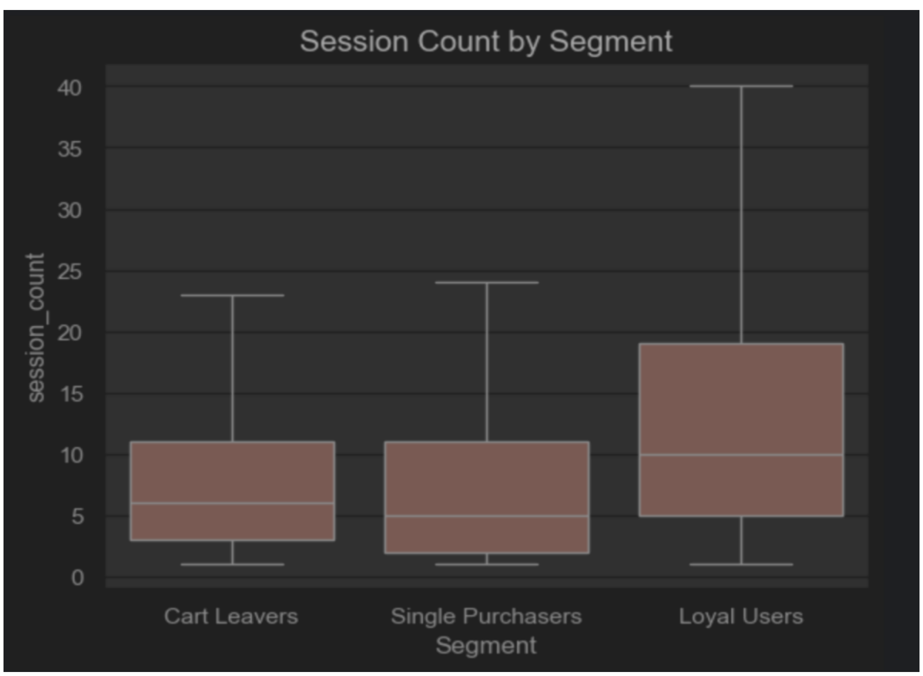
- 

- 

5.3 Session Count

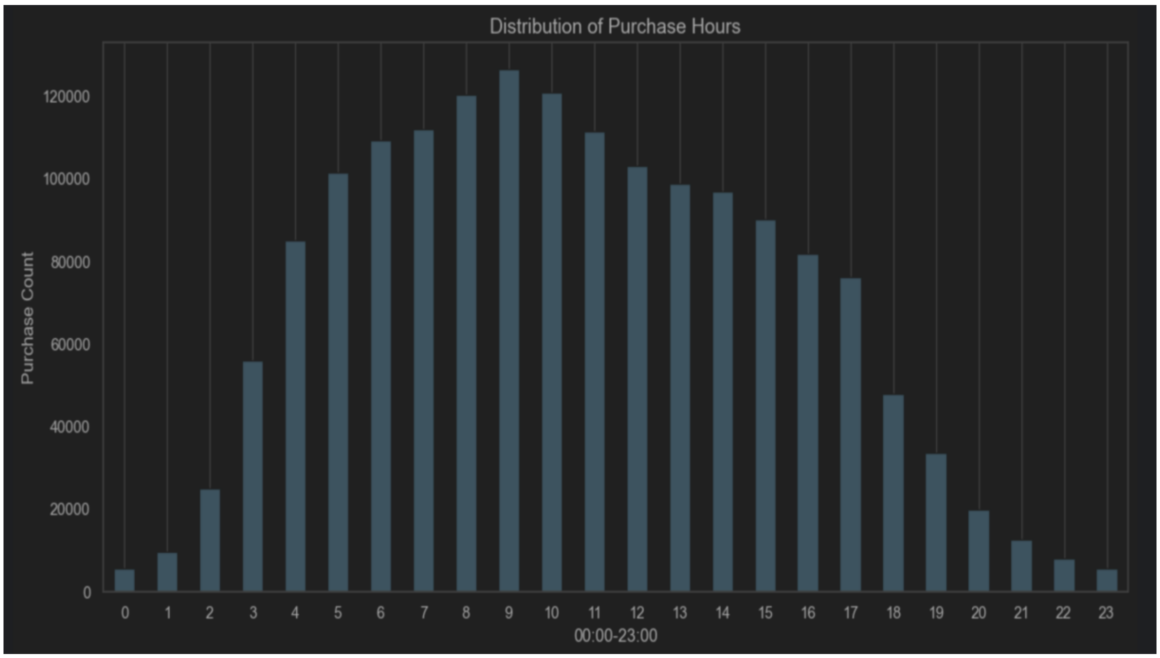
- There was a weak but positive correlation between session count and purchases.
- A scatter plot showed that some users had lots of sessions but still didn't buy anything.
- This might point to issues in the user journey or friction in the buying process.





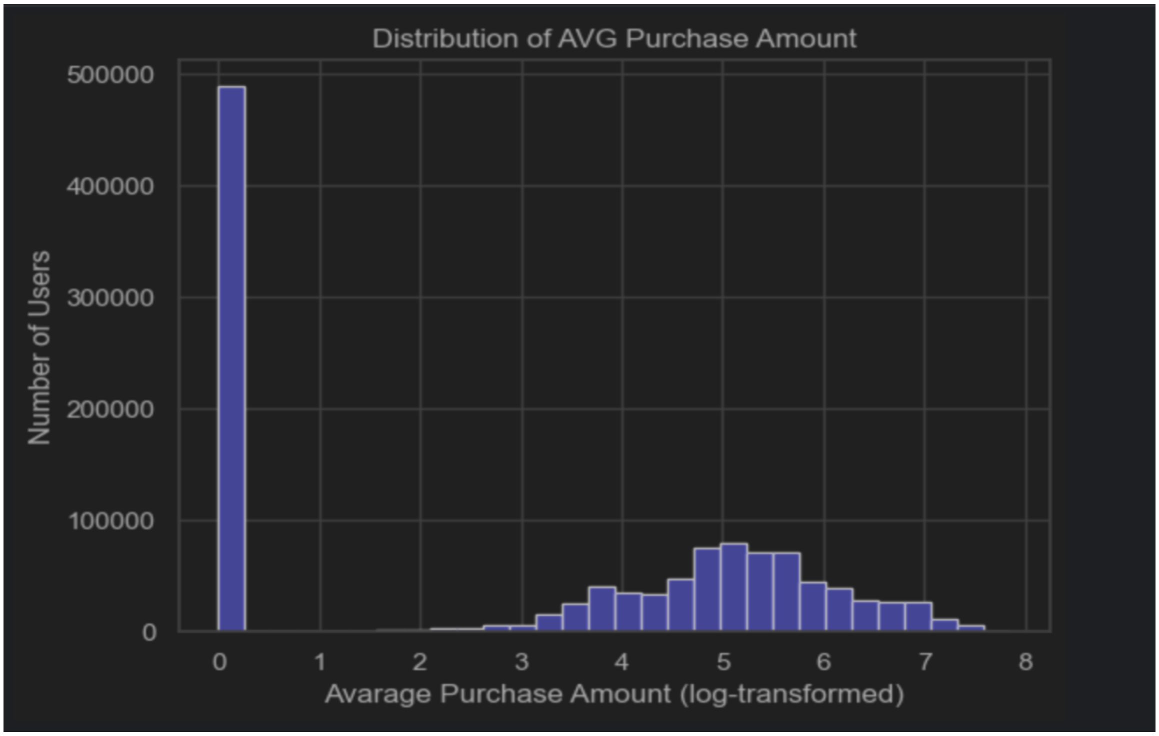
5.4 Purchase Timing

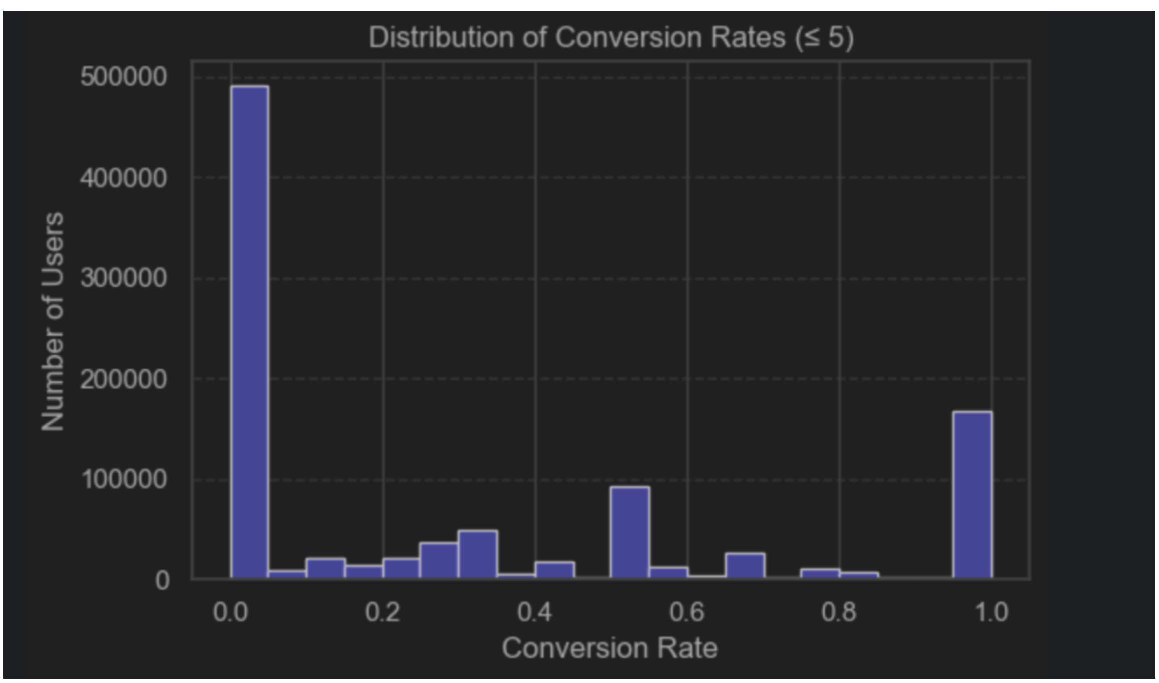
- Purchases tended to occur more frequently during morning hours, peaking between 8 AM and 11 AM.
- Activity gradually declined throughout the afternoon and evening.
- These insights can inform the optimal timing of promotional campaigns and customer engagement strategies.

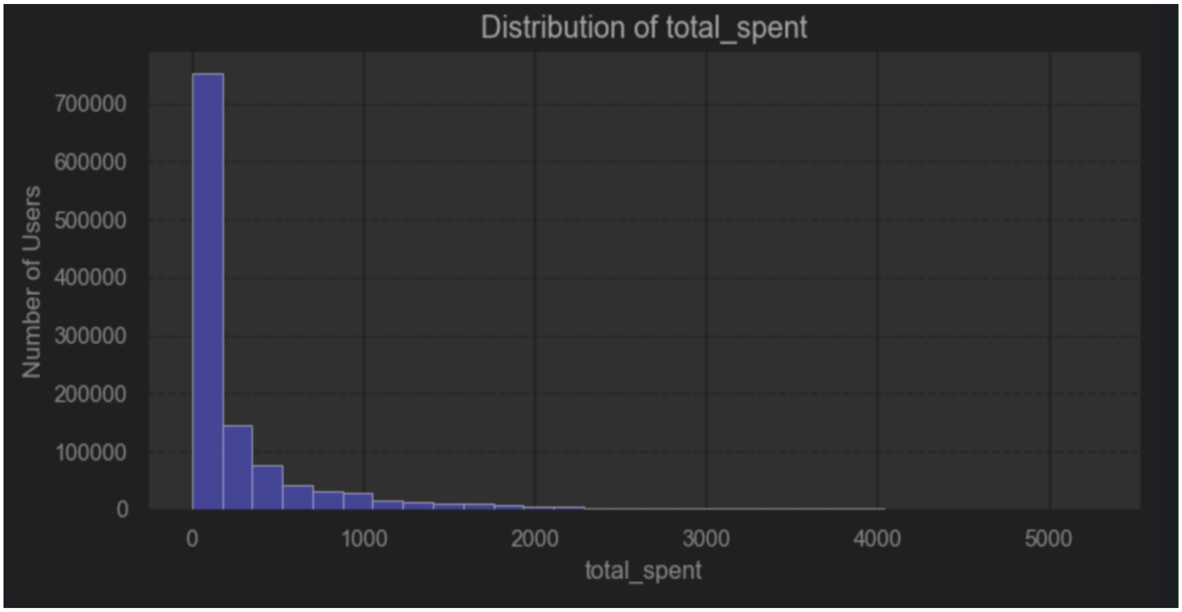


5.5 Repeat Buyer Behaviour

- For users who bought more than once, we looked at the average time between purchases.
- Some shopped frequently, others less so — showing a variety of patterns even among loyal users.





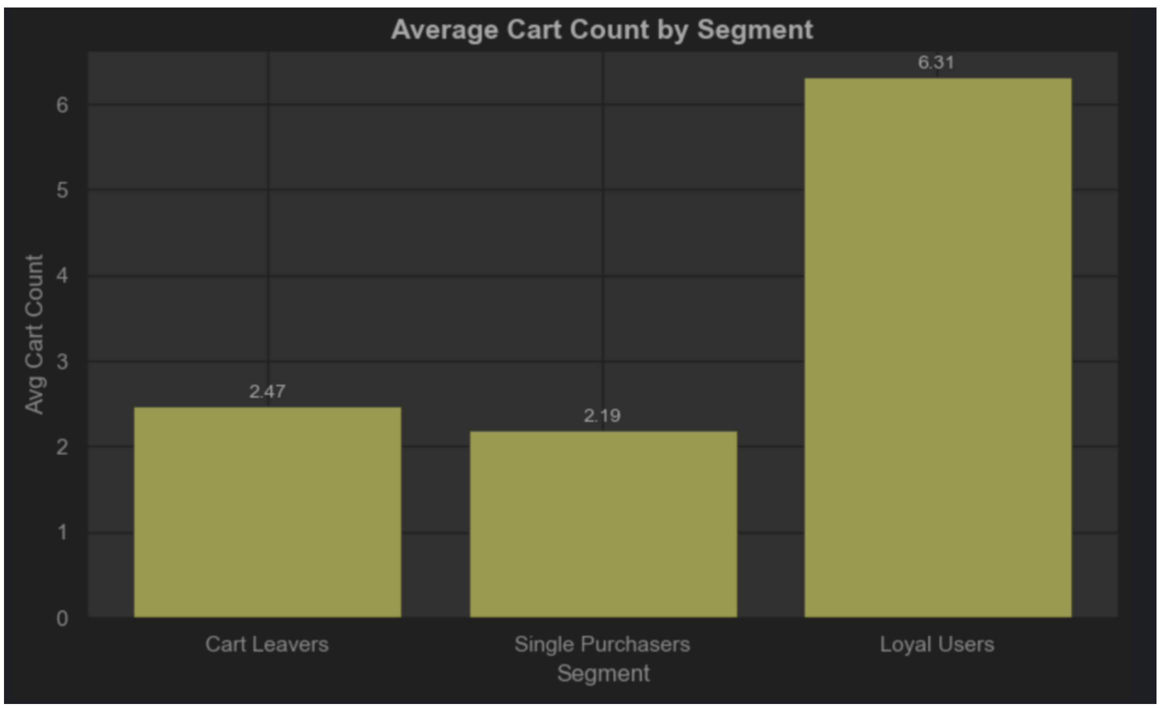


5.6 Segment Differences

- Repeat buyers had the highest average cart count and total spend.
- One-time buyers usually had fewer sessions.
- Browsers could have many sessions but zero interaction.
- Cart abandoners had higher session counts, suggesting they spent more time researching before giving up.

	Segment	User Count	Avg Cart Count	Avg Purchase Count	Avg Total Spent	Avg Session Count
0	Cart Leavers	417294	2.47	0.00	0.00	9.14
1	Single Purchasers	362329	2.19	1.00	265.67	8.42
2	Loyal Users	274504	6.31	4.31	1,368.91	15.05


Customer Segments Evaluation
User Count by Segments
Number of Users
400000
350000
300000
250000
200000
150000
100000
50000
0
Cart Leavers
Single Purchasers
Loyal Users
Average Total Spent by Segments
Avg Total Spent
1400
1200
1000
800
600
400
200
0
Cart Leavers
Single Purchasers
Loyal Users


Average Cart Count by Segment
Avg Cart Count
6
5
4
3
2
1
0
2.47
2.19
6.31
Cart Leavers
Single Purchasers
Loyal Users
Segment

6. Conclusion

This analysis provided valuable insights into how users interact with the e-commerce platform. A significant portion of users did not complete any purchases, highlighting opportunities to optimise the customer journey and improve conversion rates. Notably, users who added items to their carts but didn't finalise the purchase—accounting for a cart abandonment rate of 41.18%—represent a high-potential group for targeted re-engagement strategies such as reminders or personalised offers.

Interestingly, even greater potential may lie with users who made only one purchase. Since acquiring new customers is often more costly and time-consuming than retaining existing ones, focusing on strategies to turn first-time buyers into repeat customers could significantly boost long-term revenue. This is also supported by our findings: only 24.91% of users made more than one purchase, but these repeat buyers showed notably higher overall spending.

The segment-based approach allowed for a more nuanced understanding of customer behaviours, opening the door to personalised marketing efforts. For instance, while the correlation between session count and purchase count was relatively weak (0.14), it still suggests that users with frequent visits may need just a small nudge—such as relevant offers or a better user experience—to convert. Additionally, users showed a clear tendency to purchase in the morning hours, especially between 8 AM and 11 AM. This insight can inform the timing of marketing campaigns, such as sending push notifications or emails when users are most active.

Lastly, we examined the average time between purchases among repeat buyers. On average, these users made another purchase every 150.77 hours, which dropped to 98.09 hours when excluding outliers. This behavioural pattern can guide the scheduling of follow-up promotions or retention efforts.

Overall, this project lays a strong foundation for data-driven decision-making. Incorporating additional data sources in future—such as product-level information or user demographics—could further enhance the depth and value of these insights.

Dataset:

https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

Code:

https://github.com/mehmetfurkanakpinar/ecommerce-customer-segmentation.git

Portfolio:

mehmetfurkanakpinar.com