Data Wrangling and Visualisation

Group Assignment

Customer Segmentation and Heterogeneous Churn Dynamics in E-Commerce

Group A:
Khaled Albinali
Aayush Damani
Lena Sophie Stute
Giray Koca
Claire Niu

Submission Date: 23 February, 2026

Table of Contents

1. Introduction and Motivation

E-commerce firms operate in an environment characterised by low switching costs, high price transparency, and intense competitive pressure. Customers can compare alternatives instantly and migrate between platforms with minimal friction. At the same time, customer acquisition costs have risen steadily, driven by advertising saturation and increased digital competition. As a result, retaining existing customers has become strategically more important than replacing lost customers.

Customer churn therefore represents not only an operational metric, but a fundamental driver of long-term profitability and performance. Firms frequently rely on aggregate churn analyses to identify risk factors and design retention initiatives. However, effective churn management requires more than identifying which customers leave. It requires understanding how different types of customers disengage and whether the behavioural factors associated with churn operate uniformly across the customer base. When decisions are derived solely from overall relationships, meaningful structural differences across customer groups may be obscured, leading to retention strategies that are inefficient or misaligned with underlying behavioural dynamics.

Customers are not homogeneous. They differ substantially based on their tenure with the organisation, purchasing behaviour, service experience, and product preferences. These differences may generate distinct churn dynamics, meaning the same behavioural variable can have materially different implications across customer groups. A retention initiative that appears effective in aggregate may therefore fail to address, or even exacerbate, churn risk within specific segments. Recognising and modelling this heterogeneity is not only essential for developing targeted and economically efficient retention strategies, but also predicting churn. Against this backdrop, the study addresses the following research question: ***Does customer segmentation reveal meaningful heterogeneity in churn behaviour across lifecycle, service experience, purchasing behaviour, and transactional characteristics, and does this heterogeneity materially affect predictive modelling and strategic retention decisions?***

To address this question, the analysis is split into three stages. First, exploratory data analysis is conducted to identify key behavioural patterns associated with churn. Next, customers are segmented using clustering techniques based on lifecycle and purchasing characteristics. The resulting segments are profiled to examine differences in churn outcomes and structural attributes. Finally, logistic models are estimated both globally and within segments to evaluate whether churn drives vary across customer groups, and to address the predictive implications of incorporating segmentation. The contribution of this report lies in integrating segmentation with churn modelling to evaluate whether customer-wide relationships conceal meaningful behavioural heterogeneity. By combining structural profiling with segment-specific modelling, the analysis provides both interpretive insight and evidence on whether differentiated retention strategies are warranted.

2. Data Description and Preparation

The analysis is conducted using the E-commerce Customer Churn Analysis and Prediction dataset from Kaggle (Ankit Verma, 2021), comprising 5,329 customer-level observations. Each observation represents an individual customer. The data do not specify a fixed time horizon, and the variables capture recent account and behavioural characteristics.

The dataset contains 20 variables spanning lifecycle, behavioural, operational, and demographic dimensions. The binary target variable is Churn, defined as 1 if a customer has exited and 0 otherwise. For clarity, variables are grouped into four conceptual categories used throughout the analysis. These include **lifecycle variables** such as Tenure, **service experience** variables including Complain, WarehouseToHome, and SatisfactionScore, **purchasing behaviour** variables such as OrderCount, CouponUsed, CashbackAmount, DaySinceLastOrder, HourSpendOnApp, and OrderAmountHikeFromlastYear, and **customer and transactional** attributes including PreferredPaymentMode, PreferedOrderCat, MaritalStatus, CityTier, Gender, NumberOfAddress, and NumberOfDeviceRegistered. A full description of all variables is provided in ***Appendix E***.

CustomerID was removed prior to analysis as it carries no predictive information. Categorical variables were converted to factors, and numeric variables were retained in continuous form. For modelling purposes, numeric predictors were standardised to ensure comparability of coefficient magnitudes.

The dataset then underwent structured cleaning and preprocessing. Missing values were identified across several numeric variables, with missingness remaining below 6 percent for all affected features, as shown in ***Appendix F***. Numeric variables with missing observations were imputed using the median, while categorical variables were imputed using the mode. Inconsistent categorical labels were consolidated to ensure uniform category definitions across payment modes, login devices, and product categories. Duplicate customer records were checked and none were retained. Numeric variables were examined for extreme outliers using the three times interquartile range rule, and no anomalous observations required removal. Following these steps, the final dataset contained 5,329 complete customer observations and was used for all subsequent exploratory, clustering, and modelling analyses.

3. Exploratory Data Analysis

This section examines structural patterns in customer churn across customer tenure, purchasing behavior, engagement, and service experience. The objective is to identify which dimensions are strongly associated with retention risk, and whether meaningful heterogeneity exists within the customer base. These insights will directly inform the subsequent clustering design and modelling strategy. Overall, 16.8% of customers in the dataset are flagged as churned. This establishes the baseline retention benchmark against which subgroup differences are evaluated.

3.1 Customer Tenure and Churn Risk

Tenure reflects how long a customer has been active with the organisation and therefore offers a direct way to assess whether churn risk is concentrated during early onboarding or distributed evenly over time.

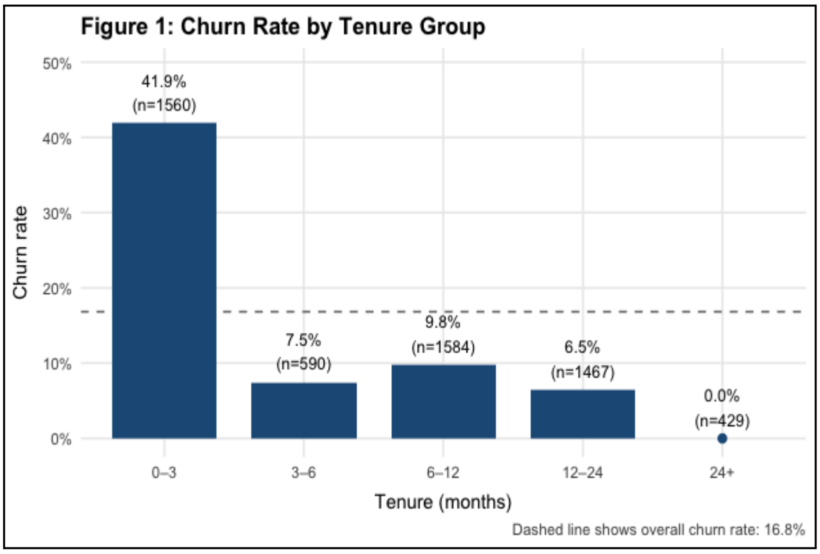


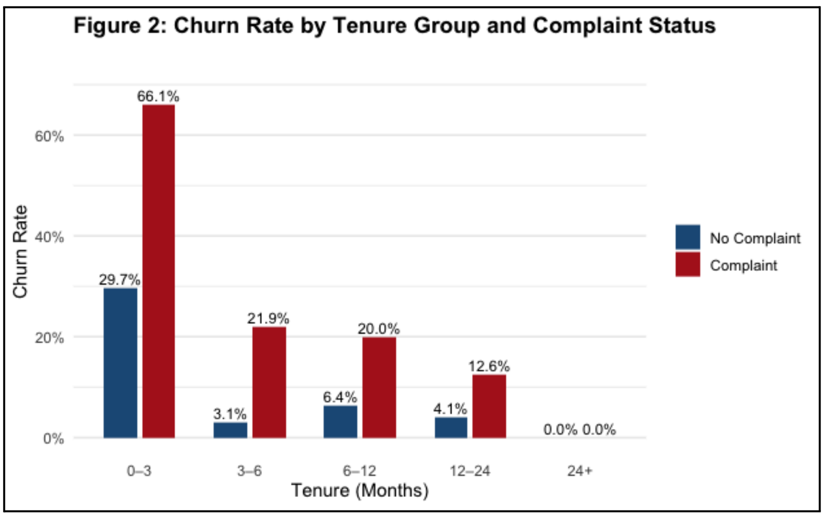
Figure 1 presents the churn rates by tenure groups. Customers were binned into five tenure categories (0-3, 3-6, 6-12, 12-24, and 24+ months) to assess how retention risks evolve over time. Based on the figure, churn is heavily concentrated within the earliest tenure stage. Customers who have been with the organisation for 0-3 months exhibit a churn rate of 41.9% (n = 1560). This is substantially above the baseline churn rate of 16.8% present among all customers in the dataset. Beyond three months, churn declines sharply and stabilises below 10% across subsequent tenure groups. Notably, customers with more than 24 months of tenure record a churn rate of 0.0% (n = 429).

The magnitude of the differences in churn rates across tenure groups suggests that churn is structurally front-loaded, with early or newer customers representing higher churn risk relative to more established customers. This has important implications for interpretation. For instance, while the overall churn rate is 16.8%, this aggregate figure masks substantial temporal heterogeneity. Once disaggregated by tenure groups, it becomes clear that churn risk is not evenly distributed across the customer base. As a result, all subsequent analyses should be interpreted with this structural imbalance in mind.

3.2 Service Experience and Operational Factors

Service quality is a central determinant of customer retention. Customers who encounter delivery issues or show product dissatisfaction are more likely to disengage from the organisation. In this dataset, complaint behavior provides a direct observable indicator of service dissatisfaction. However, as established in ***Section 3.1***, churn risk is heavily concentrated among early tenure customers. Therefore, examining complaints without accounting for tenure may mask important lifecycle differences. For this reason, complaint effects are analysed within tenure groups, as opposed to the full customer base.

Figure 2 presents churn rate by tenure groups, split by whether a customer raised a complaint in the past month. Across all tenure bands, complaints exhibit materially higher churn than those who did not. The effect is most pronounced in the earliest tenure group. Among customers in their first three months, churn increases from 29.7% among non-complainers to 66.1% for those who complained. This indicates that negative early service experiences are strongly associated with immediate disengagement from the organisation.



Although overall churn declines sharply among subsequent tenure groups, the effect of complaints on churn rate persists in the 3-6, 6-12, and 12-24 month groups. As established earlier, while churn risk is structurally higher among early tenure customers, complaints further amplify this risk across all tenure groups, with the strongest effect observed among customers in their first three months.

Warehouse-to-home distance serves as a proxy for delivery logistics and potential service friction. Greater distance may reflect longer delivery times and increased chances of delay. This can negatively affect customer experience, and hence, retention. As with complaints, however, the impact of delivery distance should be assessed within tenure groups.

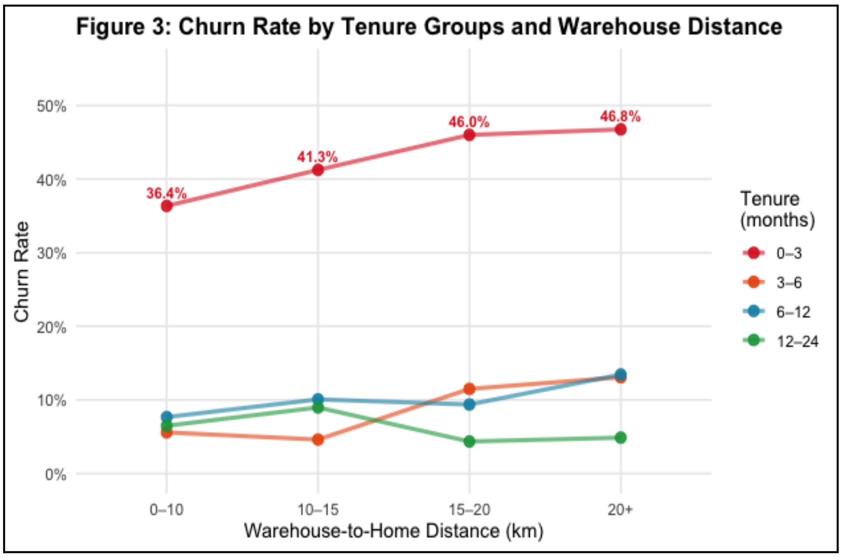


Figure 3: Churn Rate by Tenure Groups and Warehouse Distance

Figure 3 presents churn rates by warehouse-to-home distance across tenure groups. For customers in the earliest tenure band (0-3 months), churn increases consistently as delivery distance increases. Churn increases from 36.4% in the 0-10km band to 46.8% for customers located more than 20 km away from the warehouse.

A similar upward pattern is visible for customers with 3-6 months of tenure, where churn increases as warehouse distance increases. However, the magnitude of increase is smaller compared to the earliest tenure band. The 6-12 month group also shows a positive relationship between distance and churn, however the slope flattens even further. In contrast, for customers with 12-24 months of tenure, churn remains low and relatively stable across all warehouse distance bands, suggesting that warehouse distance has little additional impact on churn risk at this stage of the customer relationship. Overall, longer warehouse distances are associated with higher churn rates. However, as customers become more established, they appear to become less sensitive to delivery-related factors, as the effect decreases and eventually diminishes over time.

3.3 Purchasing Behaviour and Promotional Intensity

Customer purchasing behavior may reflect underlying engagement and commitment to the organisation. In this dataset, purchasing behavior is captured through three measures: total order count, coupon usage, and average cashback (all in the past month). Unlike tenure, which reflects the length of a customer's relationship with the organisation, these variables capture recent transactional activity and promotional exposure. The key question is whether higher short-term activity corresponds to lower churn risk, or whether it reflects temporary, promotion-driven engagement.

Figure 4 presents customer tenure across average cashback levels in the past month, with mean churn rates shown above each bin. The plot shows a clear upward shift in median tenure as average cashback increases. Customers in the highest cashback bands are predominantly more established, while lower cashback bins contain a larger concentration of customers with early tenures. Mean churn rates are lowest in the highest cashback tiers (4.3% and 6.4%). However, because cashback is strongly associated with tenure, differences in churn across bins largely reflect lifecycle composition as opposed to cashback itself. Higher cashback groups churn less, however they are also composed of longer-tenure customers, who are structurally less likely to churn.

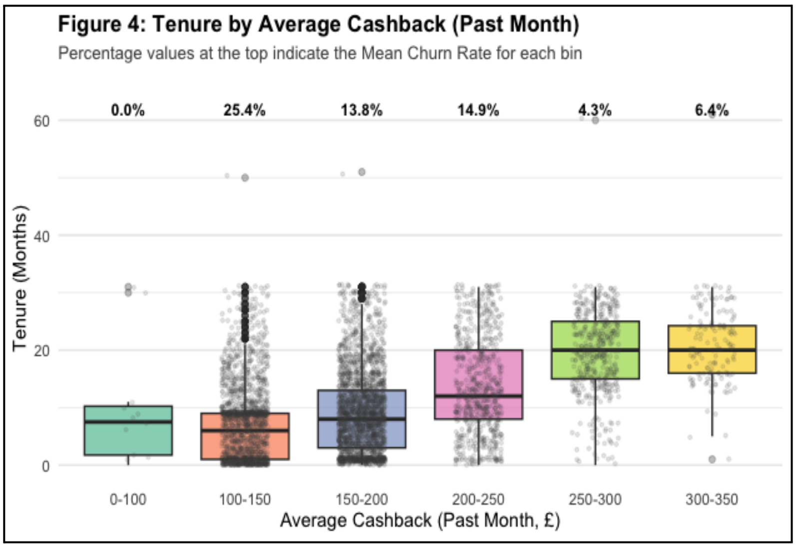


Figure 4: Tenure by Average Cashback (Past Month)
Percentage values at the top indicate the Mean Churn Rate for each bin

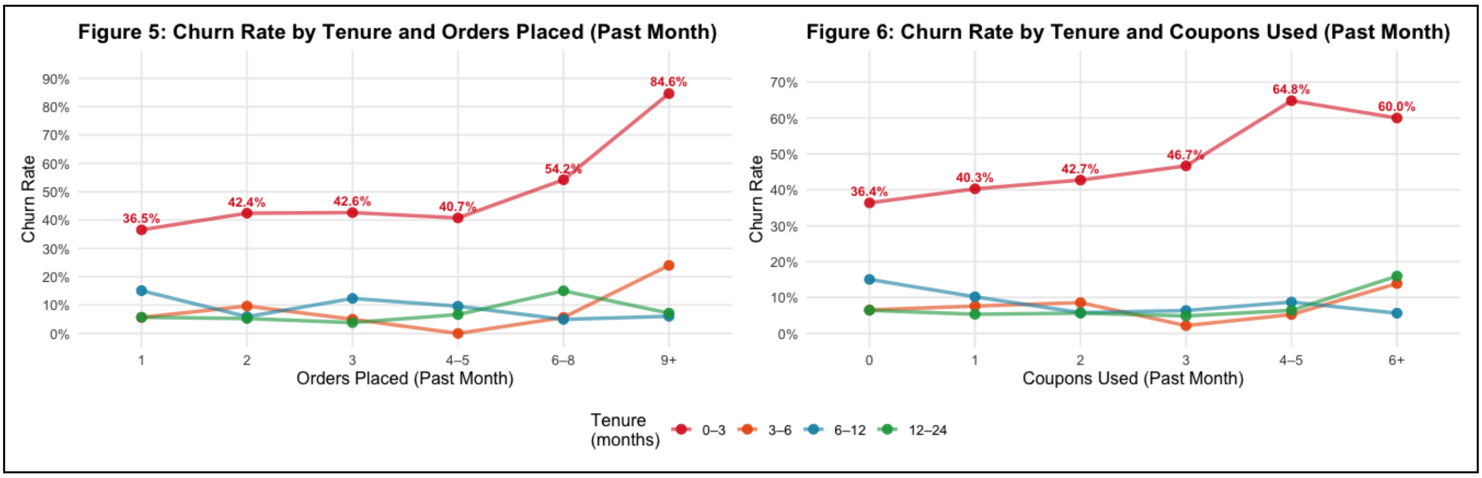


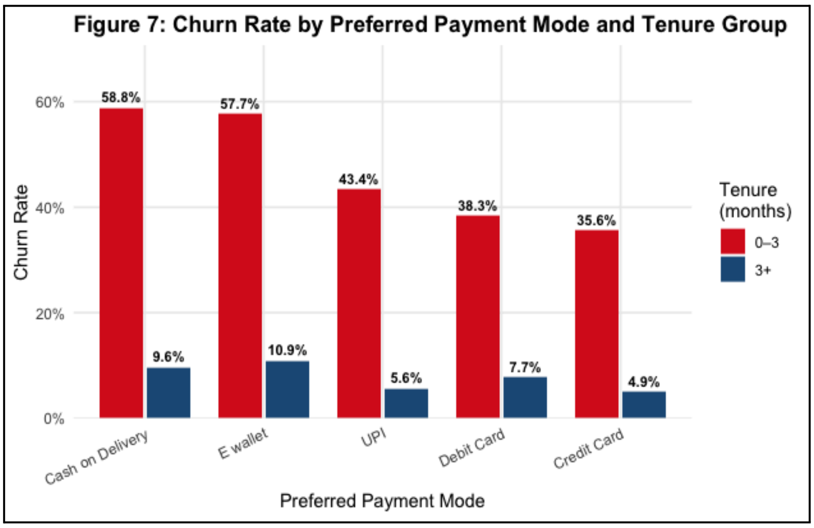
Figure 5: Churn Rate by Tenure and Orders Placed (Past Month)
Figure 6: Churn Rate by Tenure and Coupons Used (Past Month)

While cashback reflects promotional rewards received, order count directly capture recent purchase frequency. Churn rates are therefore evaluated across orders placed in the past month (within tenure groups) to assess whether higher transactional intensity corresponds to increased retention risk. ***Figure 5*** (above) presents churn rates by order count in the past month, across tenure groups. A clear pattern emerges among customers in their first three months. For this group, churn remains relatively consistent across one to five orders, however sharply increases beyond five orders. Customers placing six to eight orders exhibit a churn rate of 54.2% (n = 107), while those placing nine or more orders reach 84.6% (n = 26). Although the number of observations in the highest-order bin is smaller, the sample size is not negligible, and the magnitude of the increase is substantial. A similar, though less extreme pattern is observed among customers with three to six months of tenure, where churn increases at higher order levels. In contrast, for customers with six to twelve and twelve to twenty-four months of tenure, churn remains relatively low and stable across order counts. These results indicate that elevated purchase frequency is associated with higher churn risk primarily during the early stages of the customer relationship. However, much like the other interactions explored, the effect diminishes as tenure increases. The pattern observed in ***Figure 5*** is not immediately intuitive. Higher purchase frequency would typically be interpreted as a signal of stronger customer commitment and therefore lower churn risk. However, churn increases as order count rises, particularly among early-tenure customers.

A possible explanation to the above is that elevated early order volumes are driven by promotional intensity rather than sustained commitment. ***Figure 6*** (above) therefore analyses churn rates across coupon usage within tenure groups. The plot shows a clear positive relationship between coupon usage and churn among early-tenure customers. For those in their first three months, churn rises steadily with coupon intensity, and increases sharply beyond four coupons. This mirrors the findings from ***Figure 5***, where higher order counts were also associated with elevated churn among early customers. A plausible explanation is that aggressive onboarding promotions drive short-term purchasing spikes rather than sustained commitment. Customers may use multiple coupons to place several early orders, but once incentives diminish, they disengage. The relationship holds for customers with three to six months of tenure, although the magnitude is weaker. Beyond this stage, the pattern becomes less consistent. This suggests that promotional intensity is most strongly linked to churn risk for customers who are relatively new to the organisation, compared to those with more established tenure.

3.4 Customer and Transactional Attributes

Beyond purchasing behaviour, customer and transactional characteristics may also capture differences in customer commitment and purchasing intent. Preferred payment methods, in particular, may reflect different levels of upfront commitment, trust, and friction in the transaction process. ***Figure 7*** therefore presents churn rates by preferred payment method, split between early (0-3 months) and established (3+ months) customers.



Debit Card and Credit Card payments account for the majority of the dataset (41.1% and 31.5% respectively), while Cash on Delivery and E-wallet represent smaller shares (9.1% and 10.9%; see ***Appendix A.1***). Despite their relatively limited presence, both Cash on Delivery and E-wallet users exhibit materially higher churn rates. Tenure remains the dominant driver of churn. However, clear differences arise across payment methods within early-tenure customers. Among those in their first three months, churn is the highest for those who prefer Cash on Delivery (58.8%) and E-wallet (57.7%), compared to other payment methods. Interestingly, this ordering persists among established customers, where Cash on Delivery and E-wallet users continue to exhibit higher churn rates than other payment methods. As the proportion of early-tenure customers is broadly similar across payment modes (approximately 25–31%; see ***Appendix A.2***), these differences cannot be explained by lifecycle composition alone. Payment method therefore appears to capture behavioural differences beyond tenure. These results suggest that customers using card-based payments may be more committed at the point of purchase, whereas Cash on Delivery and E-wallet users may exhibit lower transactional commitment, which aligns with their higher observed churn.

Following preferred payment mode, the next question is whether the type of product purchased, another transactional attribute, influences churn risk. ***Figure 8*** (below) shows substantial variation in churn rate across product categories. Customers whose preferred category is Mobile Phone exhibit the highest churn rate (27.4%), which is far higher than other categories. In contrast, Grocery customers show the lowest churn rate (4.9%). A likely explanation is that mobile phone purchases are typically infrequent, high-value, and one-off transactions. In contrast, grocery purchases are typically repeated. This structural difference in purchase frequency appears closely linked to retention outcomes.

Figure 8: Churn Rate by Product Category

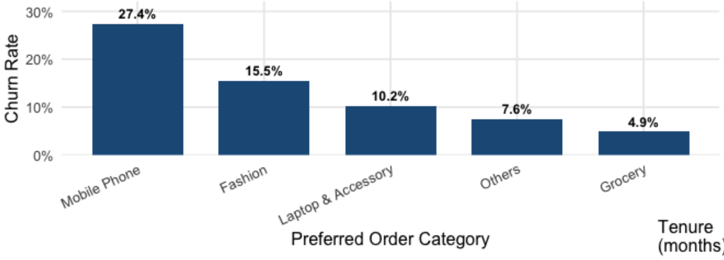
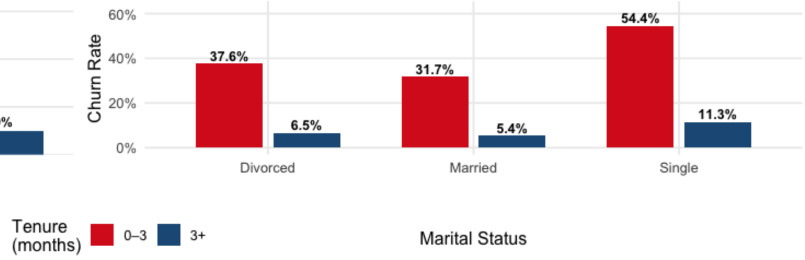


Figure 9: Churn Rate by Marital Status and Tenure Group



Turning to demographic characteristics, marital status exhibits a similarly consistent pattern. Figure 9 shows that single customers experience the highest churn across both tenure groups. For early-tenure customers, the churn rate is 54.4% for singles, compared to 37.6% for divorced and 31.7% for married customers. Interestingly, this pattern persists even among established customers (3+ months), where singles continue to display the highest churn rate (11.3% versus 6.5% and 5.4% respectively). This pattern is also reflected in overall churn rates (see ***Appendix B.1***). Importantly, singles account for approximately 31.9% of the dataset (see ***Appendix B.2***), indicating that the elevated churn is not simply a composition artefact but reflects a structurally higher attrition risk within this group.

4. Customer Segmentation (K-Prototypes Clustering)

The exploratory analysis in ***Section 3*** demonstrated that churn risk is not uniformly distributed across the customer base. Instead, it varies systematically across tenure, service experience, purchasing behavior, and transactional attributes. While variable-level relationships provide insight into individual drivers, they do not capture how these characteristics combine with distinct customer types. To further explore this, a segmentation approach was implemented with an intention to identify homogenous groups of customers with shared behavioral and transactional profiles.

4.1 Methodology

To identify distinct customer sections, a clustering approach was applied using the K-prototypes algorithm. This method was selected as the dataset contains both numerical and categorical variables. K-prototypes extend the traditional K-means framework by combining the euclidean distance for numerical features with a dissimilarity measure for categorical variables. This makes it more appropriate for data with mixed-type variables.

All behavioural, transactional, operational, and demographic variables were included in the clustering procedure. `CustomerID` and `Churn` were excluded to prevent identifier bias and leakage of outcomes. Categorical variables, such as `PreferredLoginDevice`, `PreferredPaymentMode`, `Gender`, `PreferedOrderCat`, and `MaritalStatus` were converted to factors. Additionally, numerical variables were standardised using z-score normalisation to ensure comparability across different scales and to prevent high-variance variables from influencing distance calculations.

The number of clusters was determined using the elbow method based on the total within-cluster sum of squares for k-values ranging from 2 to 10 (see ***Appendix C***). The reduction in within-cluster variance diminishes beyond $k = 4$. Therefore, a four-cluster solution was selected for interpretability. The final K-prototypes model assigns each customer to one of four segments. These segments are profiled and interpreted in the following sections.

4.2 Cluster Summary Statistics

Table 1 reports the key statistics that most clearly differentiate the identified segments, focusing on churn rate, average tenure, order count, coupon usage, cashback amount, days since last order (DSLO), and the dominant product category and its share. These variables were selected based on their observed variation across clusters and their relevance to customer retention dynamics.

Table 1: Summary Statistics of Clusters

Cluster	n	Share	Churn Rate	Tenure	Order Count	Coupons Used	Cashback	DSLO	Top Category	Top Share %
1	1466	26.04%	11.46%	9.09	1.83	0.86	164.41	4.26	Laptop & Acc.	65.28%
2	769	13.66%	14.69%	12.43	**8.76**	**4.94**	205.21	**8.15**	Laptop & Acc.	37.32%
3	**1953**	**34.69%**	**27.85%**	5.25	1.90	1.20	145.31	2.70	**Mobile Phone**	**86.12%**
4	1442	25.61%	8.53%	**16.59**	2.46	1.57	**218.54**	5.07	Laptop & Acc.	41.05%

The four clusters differ meaningfully in size and behavioural characteristics. Cluster 3 represents the largest segment (34.69%), exhibits the highest churn rate (27.85%), and shortest average tenure (5.25 months). In contrast, Cluster 4 displays the longest tenure (16.59 months), and the lowest churn rate (8.53%).

Engagement intensity also varies across segments. Cluster 2 records the highest average orders in the past month (8.76), coupon usage (4.94), and elevated cashback levels. Cluster 1 shows comparatively lower purchase intensity. Differences are also observed in dominant product categories across all clusters. This indicates further structural segmentation within the customer base. Overall, the summary statistics show clear quantitative separation across the four segments when it comes to tenure, purchasing behavior, and retention. These differences are examined further in the subsequent section through visual comparison and segment-level interpretation. In ***Section 4.4***, each segment will be characterised.

4.3 Comparative Analysis of Cluster Outcomes

4.3.1 Churn Rate Across Clusters

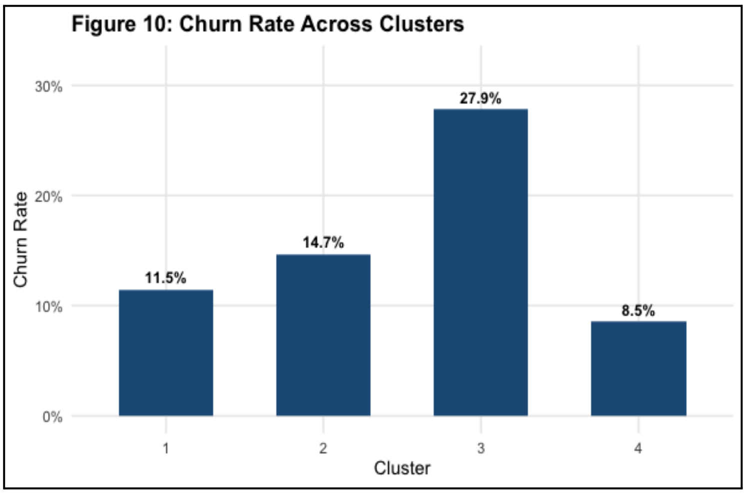


Figure 10: Churn Rate Across Clusters

Figure 10 illustrates variation in churn outcomes across the four identified clusters. Cluster 3 exhibits the highest churn rate at 27.9%, more than three times that of Cluster 4 (8.5%). Clusters 1 and 2 display intermediate churn levels of 11.5% and 14.7% respectively. The magnitude of this spread indicates that churn risk is not evenly distributed across the customer base. Particularly, the elevated churn rate observed in Cluster 3 suggests the presence of a structurally higher-risk group. In contrast, Cluster 4 represents a lower-risk and stable segment. The clear separation in churn rate across segments reinforces the structural differences in Table 1. Additionally, it confirms that the clustering procedure has meaningfully segmented customers according to their risk profiles.

4.3.2 Tenure Across Clusters

Based on the analysis in ***Section 3.1*** and subsequent sections, tenure emerged as a central determinant of churn risk. Given its structural importance in explaining retention outcomes, ***Figure 11*** examines the distribution of the average tenure across the four identified clusters.

Clear differences in customer maturity are observed across segments. Cluster 4 exhibits the longest average tenure (16.69 months), followed by Cluster 2 (12.43 months) and Cluster 1 (9.09 months). Cluster 3 records the shortest average tenure at 5.25 months. This ordering closely mirrors the churn pattern observed in ***Figure 10***, where clusters with shorter tenures display higher churn rates. The alignment between tenure and churn across segments reinforces the structural role of customer duration in an organisation, particularly when it comes to shaping retention outcomes. Additionally, it confirms that the clustering procedure captures meaningful differences in customer lifecycle positioning.

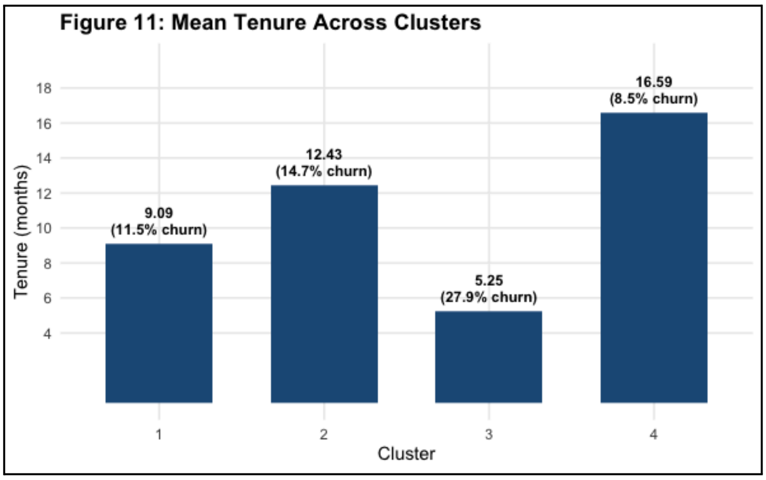


Figure 11: Mean Tenure Across Clusters

4.3.3 Purchasing Behaviour Across Clusters

Throughout ***Section 3.3***, purchasing behaviour was shown to vary meaningfully across customers and churn rate. Thus, differences in order frequency and coupon usage across the identified clusters are examined.

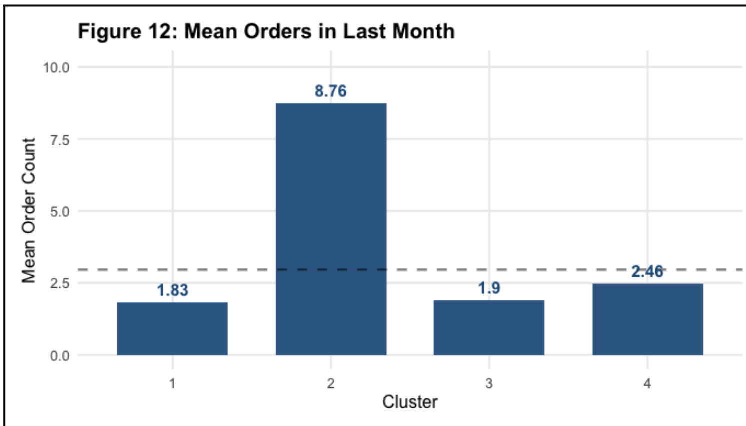


Figure 12: Mean Orders in Last Month

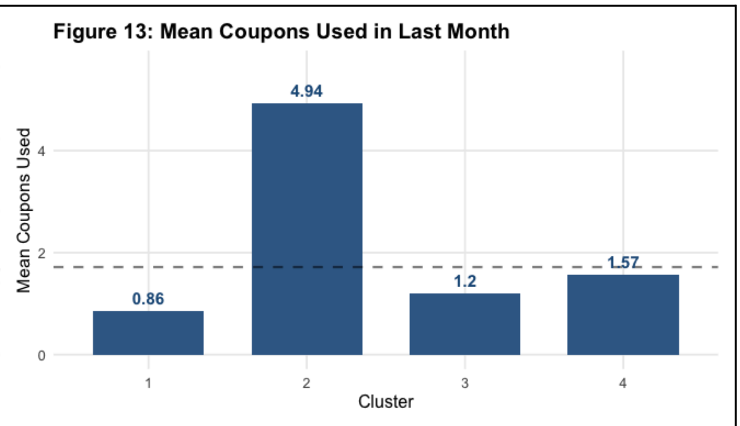


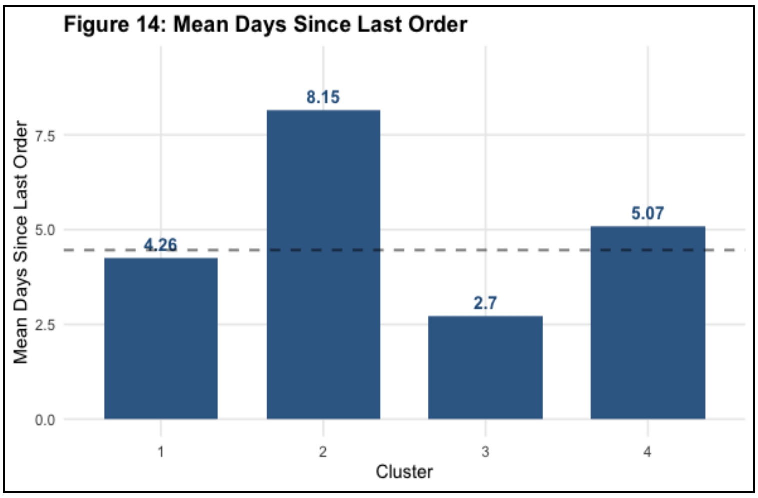
Figure 13: Mean Coupons Used in Last Month

Figure 12 presents the mean number of orders placed in the last month across the four identified clusters. The dashed horizontal line represents the overall dataset average. This provides a benchmark against which cluster-level purchasing frequency can be compared to. Cluster 2 stands out clearly compared to the order clusters, with a mean of 8.76 orders placed in the past month. This is markedly higher than all other segments. In contrast, Clusters 1 and 3 exhibit relatively low order frequencies which are both below the dataset mean. Cluster 4 shows slightly higher order frequency compared to Clusters 1 and 3, although also slightly below the overall benchmark. These differences reinforce the behavioural heterogeneity identified in ***Table 1***, highlighting Cluster 2 as a distinctly high-activity segment, while Clusters 1 and 3 appear comparatively low-frequency purchasers.

Figure 13 displays the mean number of coupons used in the last month across the four clusters. Similar to the previous figure, the dash horizontal line represents the overall dataset average, which will also serve as a reference point. A pattern similar to that observed in ***Figure 12*** is identified. Cluster 2 exhibits substantially higher coupon usage (4.94) compared to other segments, and well above the dataset average. In contrast, Clusters 1 and 3 show comparatively low coupon usage, both below the overall benchmark. Cluster 4 remains closer to the average, but still below. When considered alongside ***Figure 12***, the two figures reveal a consistent pattern. Cluster 2 represents a high-volume and promotion-intensive segment. Their elevated order frequency is accompanied by substantially high coupon usage. This indicates that purchasing activity in this segment is likely closely tied to promotional incentives. This contrasts with Clusters 1 and 3, who both have lower purchasing intensity and lower coupon usage. This points to structurally different behavioural profiles across segments.

Interestingly, despite exhibiting the highest order frequency and coupon usage, Cluster 2 also records the second-highest churn rate of 14.7%, as shown earlier. In ***Section 3.3***, it was found that higher purchasing frequency combined with higher coupon usage was associated with greater churn rates. The behavioral profile of Cluster 2 appears to be more consistent with this pattern. Their heavy engagement is likely driven by incentives, which may not necessarily translate into long-term retention. This may indicate that Cluster 2 is a more promotion-sensitive segment whose activity is conditional on promotions as opposed to underlying loyalty to the organisation.

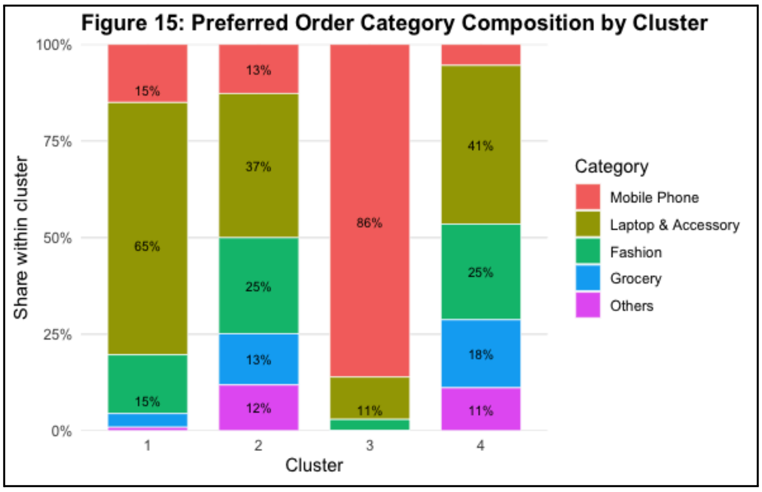
Building on the analysis of order frequency and coupon usage, to further explore the purchasing behavior across clusters, ***Figure 14*** examines recency of activity through the mean days since last order (DSLO) across clusters. Once again, the dashed horizontal line indicates the overall dataset average.



Cluster 2 records the highest DSLO (8.15 days), despite having the highest order frequency and coupon usage. This suggests that customers in this segment tend to place multiple orders within short windows, followed by longer inactive gaps. One plausible explanation is that activity is closely tied to promotional availability, with customers increasing purchases when incentives are present and reducing engagement once those incentives expire. In contrast, Cluster 3 exhibits the lowest DSLO (2.7 days), while also having the highest churn rate. This indicates that churn in this segment is unlikely to be driven by prolonged activity but rather early onboarding instability, which is explained by its shorter average tenure (***Section 4.3.2***).

4.3.4 Customer Transactional Attributes Across Clusters

Figure 15 illustrates the distribution of preferred product categories within each cluster, and clear compositional differences emerge. Cluster 3 is overwhelmingly concentrated in the Mobile Phone category (86%). This is notable given that ***Section 3.4*** identified Mobile Phone purchases as having the highest churn rate (27.4%). This alignment is consistent with Cluster 3 having the highest churn rate and relatively low order frequency. As discussed earlier, mobile phone purchases tend to be rare and high-value, typically one-off as opposed to recurring. This may naturally limit retention.



In contrast, Cluster 4 exhibits the lowest concentration of customers whose preferred category is Mobile Phone and highest share preferring Grocery. In ***Section 3.4***, it was identified that Grocery is the category with the lowest churn rate (4.9%). The concentration of preferences in the Mobile Phone and Grocery categories therefore aligns with Cluster 4 having the lowest churn rate and longest tenure. This segment appears to prefer categories which are likely to have more repeated purchases, contributing to a structurally more stable customer profile.

Cluster 2, the high-frequency segment, displays a more balanced distribution of preferred categories. Rather than being dominated by a single product type, customer preferences in this segment are more diversified. This heterogeneity may partially support its elevated order volume.

4.4 Segment Characterisation

The clustering procedure identifies four structurally distinct customer segments. These segments differ not only in churn outcomes, but also in tenure, purchasing behavior, and transactional attributes. The following characterisation of segments is based on the patterns observed and explored in ***Section 4.3***.

Cluster 1: Core Customers. This cluster represents the balanced centre of the customer base. It exhibits moderate tenure, moderate churn, and steady but not intensive purchasing behavior. Order frequency and coupon usage are below those observed in the high-activity segment (Cluster 2), while recently remains close to the dataset average. Overall, this segment appears stable but not highly engaged. It does not display severe churn risk, nor does it exhibit the structural stability of the lowest-risk segment (Cluster 4). As a result, Cluster 1 can be interpreted as the core and mainstream customer group.

Cluster 2: Deal Seekers. This cluster is defined by the highest order frequency and coupon usage across all segments. However, it also records the highest days since last orders. As mentioned, this could suggest that purchasing activity may occur in concentrated bursts as opposed to sustained engagement. This pattern is consistent with behaviour that is highly responsive to promotional incentives. Despite strong purchasing behavior, this segment also records the second-highest churn rate. As a result, these characteristics suggest a high-volume group, but one whose volume is driven by promotions. Their engagement may fluctuate depending on incentive availability.

Cluster 3: High-Risk New Customers. This cluster exhibits the highest churn rate and shortest average tenure across all segments. Additionally, the preferred category of customers in this segment is heavily concentrated in the Mobile Phone category, which was identified as having the highest churn rate. Compared to other segments, order frequency is relatively low, yet recency of activity is high. This indicates that it is unlikely that churn is driven by prolonged inactivity. Instead, this segment appears to be structurally unstable, potentially reflecting newer customers whose purchasing activity is concentrated in infrequent and high-value product categories (mobile phones). The behavioural and compositional alignment reinforces its classification as the highest-risk segment within the customer base.

Cluster 4: Loyal Customers. This cluster demonstrates the lowest churn rate and longest average tenure. Additionally, this cluster exhibits moderate purchase frequency and relatively low purchase recency. Product preferences are primarily concentrated in the Grocery, the category which was associated with the lowest churn rate. The concentration of preferences in categories that are likely to involve more repeated purchases, combined with the low churn and longer tenure, positions this segment as the most stable group within the dataset.

5. Logistic Regression Model

To complement the exploratory analysis in ***Section 3*** and the clustering analysis in ***Section 4***, we estimate logistic regression models to examine the determinants of churn more formally. There are two objectives to this. First, to validate the key drivers of churn identified earlier. Secondly, to assess whether the segment profiles derived from the clustering analysis exhibit distinct behavioural sensitivities. This modelling framework therefore serves to both reinforce the exploratory findings and test the robustness of the proposed customer segmentation structure. A global logistic regression model is first estimated using all customers, followed by cluster-specific models for each of the four segments identified from ***Section 4***. This approach therefore extends the earlier analysis by formally modelling churn risk at both the global and segment levels. The global logistic regression is estimated on all numeric predictors, with churn specified as the binary outcome.

5.1 Global Logistic Regression Model

The results of the global model reinforce earlier exploratory findings. Tenure emerges as the dominant negative predictor of churn, indicating that longer-tenure customers are substantially less likely to leave the organisation (as discussed in ***Section 3.1***). Similar to the findings in ***Section 3.2***, Complaint exhibits the strongest positive associated with churn, highlighting dissatisfaction as a critical determinant of customer retention. Warehouse-to-home distance also remains positively associated with churn, although its magnitude is considerably smaller in comparison. Other behavioural variables display more moderate effects. A visualisation of the standardised coefficients from the global model is provided in ***Appendix D***.

5.2 Segment-Specific Logistic Regression Models

Separate logistic regression models are next estimated within each identified customer segment, again using standardised numeric features. This allows for direct comparison of coefficient magnitudes across segments. The heatmap below contrasts the segment-specific coefficients with those from the global model, highlighting differences in behavioural sensitivities across customer groups. ***Figure 16*** presents the standardised logistic regression coefficients from both the global model and the segment-specific models.

Figure 16: Segment-Specific Churn Drivers

Standardised logistic regression coefficients comparing segment-level and global effects..

The segment specific models reveal meaningful heterogeneity in churn drivers across customer groups. Complaint remains a consistently strong positive predictor in all segments, with the largest magnitude observed in Cluster 3. This aligns with the earlier finding in ***Section 3.2*** that complaints substantially intensify churn risk. Given that Cluster 3 (the *"High-Risk New Customers"* segment) is characterised by shorter tenure and higher overall churn, dissatisfaction within this group appears particularly destabilising.

Coupon usage displays a more segment-dependent effect. It is positive only in Cluster 2 (the *"Deal Seekers"* segment), while remaining negligible or negative elsewhere. This pattern is consistent with the findings from ***Section 4*** and the segment characterisation of Cluster 2, where engagement may be driven by incentives rather than intrinsic loyalty. In such cases, churn risk may increase once promotional intensity declines.

Order count exhibits further heterogeneity. While it is positively associated with churn in most segments, it turns negative in Cluster 3. This aligns with the segment profile identified earlier, where Cluster 3 is characterised by relatively infrequent, high-value purchases, with a strong concentration in the mobile phone category, and elevated overall churn risk. Within this group, additional purchasing activity may therefore signal stronger engagement and reduce the likelihood of exit. Interestingly, Cashback amount is negatively associated with churn in this cluster, suggesting that incentives may partially offset exit risk within this otherwise high-churn segment.

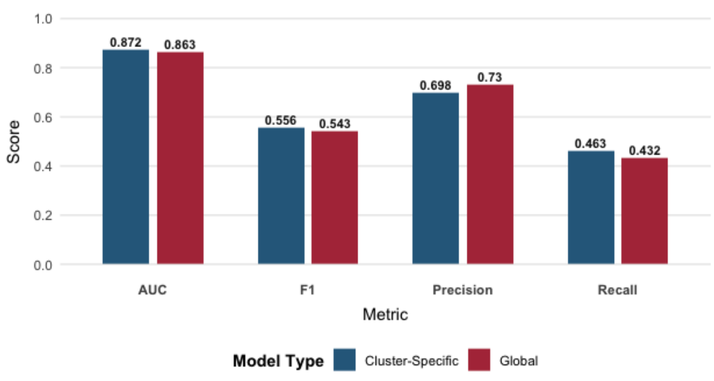
Days since the last order (DSLO) also varies in importance across segments. It is close to zero in Cluster 2, the high-frequency and promotion-sensitive segment. This indicates that recency does not meaningfully differentiate churn risk within this segment. This is consistent with the episodic purchasing pattern previously identified, where customers in this segment likely concentrate their activity within short promotional windows and reduce engagement once incentives diminish. However, it is negative in other clusters, suggesting that more recent activity is associated with lower churn probability in segments where engagement is more stable and less promotion-driven.

5.3 Predictive Performance Comparison: Global vs Segment-Specific Models

To quantify the benefit of segment-specific modelling, the global logistic regression is compared with an ensemble of segment-level models using repeated train–test splits. In each iteration, the global model is trained on the full customer base, while separate logistic regressions are estimated within each segment and combined for prediction. Average performance across 20 splits (80/20) is reported to assess differences in predictive accuracy and classification balance.

Figure 17: Global vs Segment-Specific Model Performance

Average performance over 20 train/test splits



The segment-specific logistic regression models outperform the global model across most evaluation metrics. The AUC increases from 0.863 for the global model to 0.872 for the segment-specific ensemble, indicating improved overall discrimination between churners and non-churners. The improvement is more pronounced for the F1 score (0.543 to 0.556) and Recall (0.432 to 0.463), suggesting that the segment-level models identify a greater proportion of true churners. Precision decreases moderately from 0.730 to 0.698, implying a small trade-off in false positive rate. Overall, segmentation delivers a modest but consistent improvement in churn detection performance.

6. Recommendations and Implications

6.1 Segment-Specific Strategic Recommendations

The segmentation analysis from ***Section 4*** reveals that churn risk is not uniformly distributed across the customer base, and the logistic regression from ***Section 5*** reveals that behavioural drivers that drive churn are not consistent across segments. As a result, retention strategy should therefore be differentiated rather than uniformly applied.

Cluster 1: Core Customers. This segment represents a relatively balanced profile with moderate churn risk and average purchasing behavior. This segment may serve as the baseline audience for standard retention initiatives. However, interventions should remain calibrated to avoid inadvertently shifting behaviour toward the more promotion-sensitive dynamics observed in Cluster 2.

Cluster 2: Deal Seekers. This segment displays high order frequency and elevated coupon usage, alongside above-average churn risk. Despite having the highest average number of orders, this segment also exhibits the longest average time since last order, suggesting that purchasing activity may be concentrated in short, intensive bursts followed by periods of inactivity. Within this group, coupon usage is positively associated with churn, indicating that engagement may be closely tied to promotional incentives rather than sustained loyalty. Behaviour therefore appears episodic and incentive-driven. Broad discounting strategies may therefore be inefficient for this segment. Instead, targeted loyalty mechanisms or non-price incentives may better encourage sustained engagement and reduce reliance on promotions.

Cluster 3: High-Risk New Customers. This segment exhibits the highest churn rate, lowest tenure, and is heavily concentrated in the Mobile Phone category (where purchases are relatively infrequent). Segment-specific modelling indicates that complaints exert a particularly strong positive effect on churn within this group, relative to other segments. Retention efforts for this segment should prioritise service quality and post-purchase engagement rather than broad promotional incentives. This suggests that churn risk may stem from narrow, one-off purchases rather than sustained engagement. Retention strategy for this segment should therefore focus on increasing the depth of engagement and encourage cross-category purchasing. The aim should be to lengthen tenure rather than relying on isolated high-value transactions.

Cluster 4: Loyal Customers. This segment demonstrates the lowest churn rate, longer tenure, and a higher concentration of preference in repeat-consumption categories such as Grocery. The behavioural profile suggests structurally stable demand. Retention strategy for this group should focus on maintaining service consistency and reinforcing loyalty rather than aggressive promotional activity. Protecting this segment from service failures is likely to yield greater marginal benefit than offering incremental discounts.

6.2 Implications for Churn Modelling and Retention Design

The analysis demonstrates that aggregate churn relationships may conceal materially different behavioural dynamics across customer segments. Variables such as coupon usage and order frequency do not exert uniform effects. In some cases, their association with churn varies in magnitude or direction across segments. This highlights the risk of designing retention policies based solely on customer-wide estimates. Incorporating segmentation into churn modelling provides two benefits. First, it improves interpretability by revealing structurally distinct behavioural sensitivities. Second, it enhances predictive performance by accounting for heterogeneity that a single global model cannot fully capture. While the improvement in accuracy is incremental, the strategic value lies in preventing misaligned interventions and enabling differentiated retention design. Overall, segmentation should be viewed not as a replacement for predictive modelling, but as a structural framework that strengthens both decision-making and model specification.

7. Conclusion

This study examined whether customer segmentation reveals meaningful heterogeneity in churn behaviour and whether accounting for this heterogeneity influences predictive modelling and retention strategy. The findings demonstrate that churn dynamics are not uniform across the customer base. Distinct behavioural segments exhibit materially different purchasing patterns, service sensitivities, and risk profiles. In particular, the analysis shows that promotional intensity, purchasing frequency, and product concentration do not exert consistent effects across segments, reinforcing the importance of modelling behavioural structure rather than relying solely on aggregate relationships. Segment-specific modelling further confirms that certain drivers of churn vary in magnitude and, in some cases, direction across customer groups. While the improvement in predictive performance from incorporating segmentation is incremental, the primary contribution lies in the enhanced interpretability and strategic clarity it provides. By identifying structurally distinct behavioural profiles, the analysis supports differentiated retention design and reduces the risk of misaligned interventions derived from aggregate estimates. Several limitations should be acknowledged. The dataset does not include a defined time horizon, limiting the ability to examine dynamic churn behaviour over time. In addition, the analysis relies on observational data, meaning that identified relationships should not be interpreted as strictly causal. Future research could extend the framework using longitudinal data, alternative clustering techniques, or more advanced modelling approaches to further evaluate the stability and generalisability of segment-level churn dynamics. Overall, the findings suggest that customer segmentation provides meaningful structural insight into churn behaviour and serves as a valuable complement to predictive modelling in e-commerce retention strategy.

8. References

Verma, A. (2021) 'Ecommerce Customer Churn Analysis and Prediction'. Kaggle.

9. Appendices

Appendix A: Composition of Payment Method in Dataset

PreferredPaymentMode	n	share	share_pct
Debit Card	2314	0.41101243	41.1%
Credit Card	1774	0.31509769	31.5%
E wallet	614	0.10905861	10.9%
Cash on Delivery	514	0.09129663	9.1%
UPI	414	0.07353464	7.4%

Appendix A.1: Composition of Preferred Payment Method in Dataset

PreferredPaymentMode	Tenure_Group	n	share_within_mode	share_pct
Cash on Delivery	0–3	160	0.3112840	31.1%
Cash on Delivery	3+	354	0.6887160	68.9%
Credit Card	0–3	536	0.3021421	30.2%
Credit Card	3+	1238	0.6978579	69.8%
Debit Card	0–3	579	0.2502161	25.0%
Debit Card	3+	1735	0.7497839	75.0%
E wallet	0–3	156	0.2540717	25.4%
E wallet	3+	458	0.7459283	74.6%
UPI	0–3	129	0.3115942	31.2%
UPI	3+	285	0.6884058	68.8%

Appendix A.2: Composition of Tenure Groups Within Each Preferred Payment Method

Appendix B: Churn Rate by Marital Status (Overall)

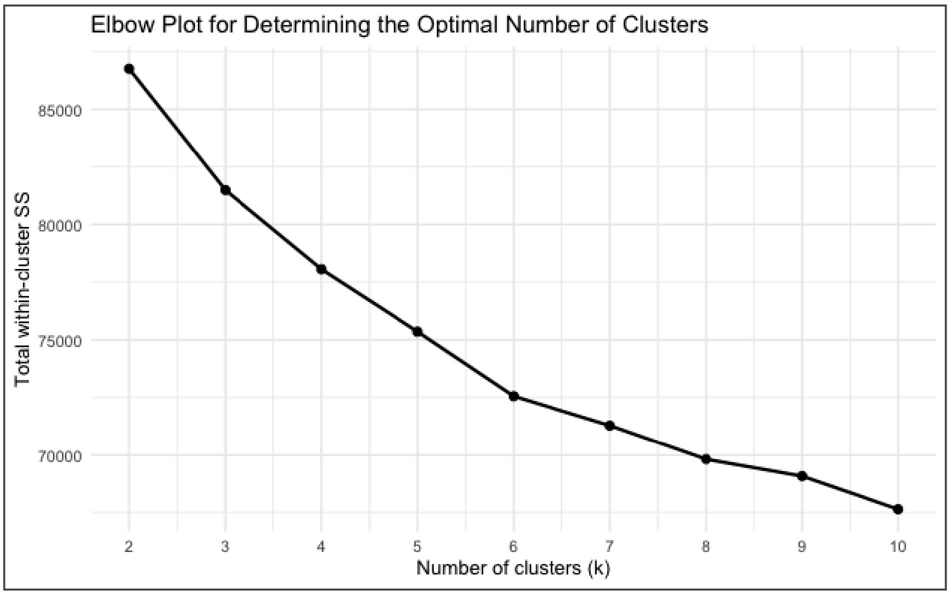
MaritalStatus	n	churn_rate	churn_rate_pct
Single	1796	0.2672606	26.7%
Divorced	848	0.1462264	14.6%
Married	2986	0.1152043	11.5%

Appendix B.1: Overall Churn Rate by Marital Status

MaritalStatus	n	share	share_pct
Married	2986	0.5303730	53.0%
Single	1796	0.3190053	31.9%
Divorced	848	0.1506217	15.1%

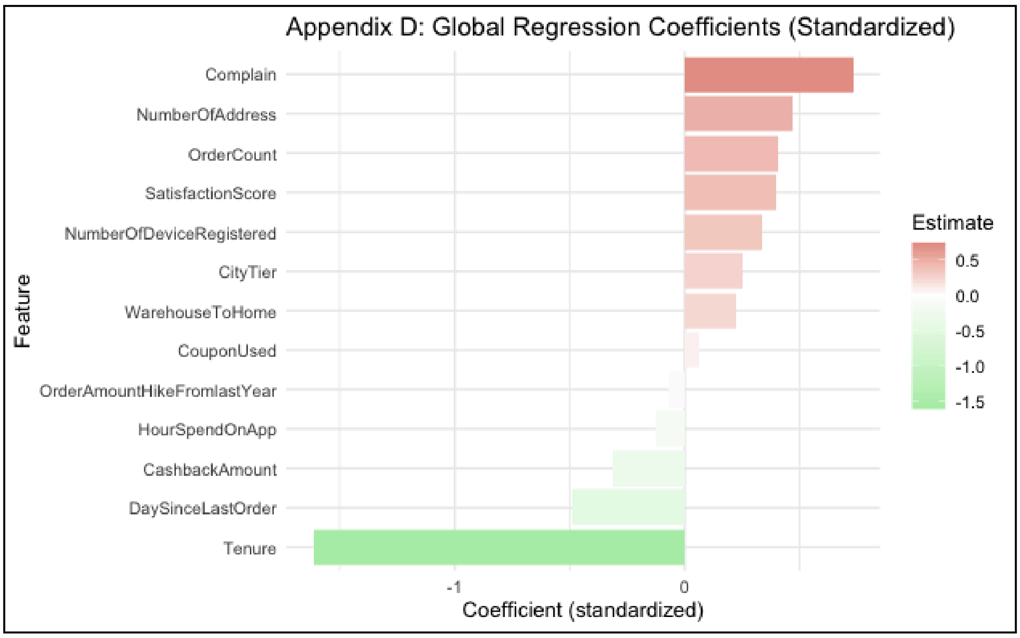
Appendix B.2: Composition of Marital Status in Dataset

Appendix C: Elbow Plot for Determining the Optimal Number of Clusters (K-Prototypes)



Appendix C: Elbow Plot for Determining the Optimal Number of Clusters (K-Prototypes)

Appendix D: Standardized Coefficients from the Global Logistic Regression Model



Appendix D: Standardized Coefficients from the Global Logistic Regression Model

Appendix E: Full Description of Dataset Variables

Variable Name	Data Type	Definition
CustomerID	Integer (ID)	Unique identification number for each customer account.
Churn	Binary (0/1)	Target variable: 1 if customer churned, 0 if active.
Tenure	Numeric	Total duration of the customer relationship in months.
PreferredLoginDevice	Categorical	Primary device (Mobile/Computer) used to access the platform.
CityTier	Ordinal (1–3)	Categorization of the customer's city by development level.
WarehouseToHome	Numeric	Distance between the fulfillment warehouse and customer home.
PreferredPaymentMode	Categorical	Most frequently used transaction method (e.g., UPI, Credit Card).
Gender	Categorical	Self-identified gender of the account holder.
HourSpendOnApp	Numeric	Average hours spent by the customer on the mobile application.
NumberOfDeviceRegistered	Numeric	Total number of unique devices linked to the customer's account.
PreferedOrderCat	Categorical	Product category with the highest purchase frequency.
SatisfactionScore	Ordinal (1–5)	Customer-reported satisfaction rating on a 1 to 5 scale.
MaritalStatus	Categorical	Current marital status (Single, Married, or Divorced).
NumberOfAddress	Numeric	Total number of shipping addresses stored in the profile.
Complain	Binary (0/1)	Indicates if a formal complaint was lodged in the last month.
OrderAmountHike	Percentage	Percentage increase in order value compared to the previous year.
CouponUsed	Numeric	Total promotional coupons redeemed in the last 30 days.
OrderCount	Numeric	Total number of orders placed within the last 30 days.
DaySinceLastOrder	Numeric	Recency metric: days elapsed since the customer's last order.
CashbackAmount	Numeric	Average monthly cashback rewards earned by the customer.
Variable Name	Data Type	Behavioral segment ID (1–4) generated via K-Means clustering.

Appendix F: Missing Values in Dataset

Variable	Missing_Count	Missing Percentage
DaySinceLastOrder	307	5.452931
OrderAmountHikeFromlastYear	265	4.706927
Tenure	264	4.689165
OrderCount	258	4.582593
CouponUsed	256	4.547069
HourSpendOnApp	255	4.529307
WarehouseToHome	251	4.458259