University of Southern Denmark
Odense
Faculty of Business and Social Sciences
Hand-in date: 01-01-2026
Number of keystrokes (incl. spaces): 44,996

Data Science Exam Report

Customer Behavior, Churn Prediction, and Business Insights in an E-commerce Context

Authors:

Rasmus Larsen
22-06-1997,Data Driven Business Development, *[1st semester]* Eksamens nr: 4145631

Tancredi Calogero Piscitello
13-12-2001,Data Driven Business Development, *[1st semester]* Eksamens nr.:4155256

Supervisor:
Jonas Husum Dalstrup

Sworn statement:

The page following the title page must contain a sworn statement of the following wording: ” I hereby solemnly declare that I have personally and independently prepared this paper. All quotations in the text have been marked as such, and the paper or considerable parts of it have not previously been subject to any examination or assessment.”

1. Introduction and Business Context

1.1 Case background

The dataset analyzed in this project represents a mid-sized European e-commerce retailer operating across four geographical regions. The company sells consumer products within multiple categories and serves customers with diverse purchasing patterns, engagement levels, and loyalty behavior. As a digital retailer, data plays a central role in supporting operational planning, marketing initiatives, and strategic decision-making.

Customers interact with the retailer through several digital channels, including the website, mobile application, and email marketing. These interactions generate detailed data capturing both transactional behavior and customer engagement over time, reflecting different stages of the customer lifecycle.

At the customer level, the dataset includes information on regional attributes, loyalty indicators, and engagement metrics. Purchasing behavior is described using order-related variables such as purchase frequency and order value measures, while product-related variables enable analysis across categories and price segments. Operational performance is represented through delivery and return variables, offering insight into logistics reliability and return behavior.

These operational dimensions are particularly important in an e-commerce setting, as delayed deliveries and frequent returns may negatively influence customer satisfaction and increase the likelihood of customer churn. In addition, the dataset includes digital interaction variables such as email engagement, website activity, and customer support interactions, which together provide a broad view of customer behavior across channels.

Overall, the dataset is well suited for applying data science techniques including data preparation, exploratory analysis, dimensionality reduction, and predictive modeling, with the aim of supporting data-driven decisions related to customer retention, marketing effectiveness, and operational optimization.

1.2 Business Challenges

Operating in a highly competitive e-commerce market, the retailer faces several interconnected challenges related to customer retention, marketing efficiency, and operational performance. Customers can switch between online retailers with minimal effort, meaning that even relatively small declines in service quality or perceived value may lead to customer attrition. As a result, it is essential for management to understand which factors influence customer behavior and long-term loyalty.

Customer retention represents one of the most critical challenges. Acquiring new customers typically requires significant marketing investment, whereas retaining existing customers is generally more cost-effective. In an e-commerce context, customers may churn for several reasons, including

reduced purchasing frequency, declining engagement, delayed deliveries, or repeated product returns. Identifying customers at risk of churn before disengagement occurs is therefore an important business objective.

Another challenge concerns marketing effectiveness. The company invests in digital marketing activities such as email campaigns, website personalization, and promotional offers. Without a clear understanding of how different customer segments respond to these initiatives, marketing efforts may become inefficient or poorly targeted. Insight into variation in customer engagement and responsiveness is necessary to allocate marketing resources more effectively and improve return on investment.

Operational performance also plays a central role in the overall customer experience. Delivery reliability and return handling are key components of the e-commerce value proposition. High return rates, delayed deliveries, or inefficient logistics processes can increase operational costs and negatively affect customer satisfaction. These issues may also indirectly contribute to higher churn, emphasizing the importance of analyzing operational performance alongside customer behavior.

Finally, management faces the challenge of transforming large volumes of transactional, operational, and digital interaction data into actionable insights. While data is collected across multiple domains, creating business value requires systematic data preparation, appropriate analytical methods, and clear communication of results to non-technical stakeholders. These challenges motivate the application of data science techniques, including exploratory analysis, predictive modeling, and visualization, to support data-driven decision-making.

1.3 Project Objectives

The overall objective of this project is to apply the data science methods taught in the course to analyze customer behavior and support data-driven decision-making for a mid-sized e-commerce retailer.

The project begins with an analysis of customer behavior and business performance using structured SQL queries. This step focuses on identifying patterns related to purchasing activity, customer engagement, delivery performance, and return behavior in order to generate business-relevant insights from the underlying database.

Subsequently, the raw customer-level data is prepared and transformed into an analytical dataset suitable for statistical analysis and machine learning. This process includes data cleaning, handling of missing values, feature engineering, and the encoding of categorical variables.

Exploratory data analysis techniques are then applied to examine relationships between variables and assess the structure of the dataset. In particular, correlation analysis and Principal Component Analysis (PCA) are used to identify multicollinearity and uncover underlying patterns in customer behavior.

Finally, predictive models are developed to support business decision-making. Both regression and classification approaches are applied to predict future customer behavior, with a specific focus on

identifying customers at risk of churn. Model results are evaluated and communicated through visualizations and a Power BI dashboard to ensure interpretability for non-technical stakeholder

2. SQL Analysis

2.1 Database Implementation

Based on the specifications provided in the "exam_database_description.pdf", we created a relational database named retail_db. The schema consists of seven normalized tables designed to capture the entire customer lifecycle, from sign-up to order fulfillment and potential returns.

The SQL script begins by ensuring a clean environment using DROP TABLE IF EXISTS to prevent errors during iterative testing. The tables were created in a specific order (Customers/Products → Orders → Order Items/Returns) to respect foreign key constraints.

During the creation process, several technical decisions were made to ensure data integrity and query performance:

Data Types for Financials: We utilized DECIMAL(10, 2) for all monetary fields (price, total_amount, refund_amount) instead of FLOAT. This is crucial in a business context to avoid floating-point rounding errors when calculating financial KPIs.

Boolean Flags: Fields representing binary states, such as email_opt_in, complaint, and on_time, were defined as BIT. This optimizes storage and allows for easy filtering (e.g., WHERE email_opt_in = 1).

Primary and Foreign Keys: Every table has a PRIMARY KEY to ensure record uniqueness. FOREIGN KEYS were rigorously applied (e.g., linking orders to customers) to prevent "orphan records", ensuring we cannot have an order without a valid customer or a return without a valid order.

Date Handling: We used the DATE type for temporal data (order_date, signup_date). This allows for efficient cohort analysis and seasonality tracking (e.g., extracting Month or Year) without the overhead of timestamp data where it isn't needed.

2.2 Data loading and Analytical Queries

The dataset was loaded into the database using the *exam_data.sql* script. To enable a comprehensive analysis of the retailer's performance, a total of 16 analytical SQL queries were developed, each addressing a specific aspect of the business.

To ensure both coverage and depth, the analytical workload was divided evenly between the two authors, with each student responsible for eight queries within a defined business domain. This structure allowed for focused analysis while maintaining a coherent overall perspective on the dataset.

Tancredi concentrated on operations and product-related analysis, including the impact of discounts, return behavior, shipping performance, and customer payment method preferences. Rasmus focused on marketing and customer insights, examining seasonality patterns, engagement across digital channels, loyalty tier behavior, and demographic differences in spending.

Together, these queries provide a balanced view of customer behavior, operational efficiency, and commercial performance, forming the basis for the subsequent exploratory analysis and predictive modeling stages.

2.3 Operations & Product Analysis

Query 1

	discount_level	avg_quantity_per_item	total_lines_sold
1	High (>20%)	2.08	86
2	Medium (10-20%)	2.02	299
3	Low (<10%)	1.96	238
4	No Discount	1.94	108

The data shows that while offering a discount doubles the purchase volume compared to full price, deep discounts do not generate more sales than small ones. In fact, low discounts under 10% performed just as well as high discounts over 20%. This is vital for our bottom line because it proves we can stop sacrificing our profit margins with aggressive price cuts. Instead, the business should switch to a strategy of offering smaller, more frequent discounts, which will drive the same high sales volume while significantly increasing our overall profitability.

Query 2

	product_id	brand	category	return_count	total_refunded
1	71	Acme	Apparel	4	669.72
2	60	Globex	Home	3	951.16
3	57	Stark	Electronics	3	1138.48
4	21	Initech	Apparel	3	141.03
5	22	Soylent	Home	2	320.95
6	31	Soylent	Grocery	2	25.52
7	2	Soylent	Home	2	1065.59
8	13	Stark	Home	2	510.38
9	17	Umbrella	Home	2	428.75
10	20	Wonka	Apparel	2	378.23

This output shows a targeted list of high-risk inventory, distinguishing between products with frequent quality issues (like Acme Apparel) and those causing significant financial loss due to high refund values (like Stark Electronics). This insight is essential for the business to immediately address root causes, such as vendor defects or misleading descriptions, effectively stopping revenue leakage and preventing customer churn.

Query 3

	carrier	avg_shipping_days	total_shipments	late_percentage
1	UPS	4.03	64	60.94
2	Local	4.22	23	65.22
3	DHL	4.23	79	62.03
4	FedEx	4.39	74	67.57

This output shows a systemic failure in logistics performance, where every listed carrier exceeds a 60% late delivery rate, peaking with FedEx at 67.57%. This pervasive unreliability presents a severe risk to customer retention and operational efficiency, signaling an urgent need for the business to renegotiate Service Level Agreements (SLAs) or diversify shipping partners to ensure delivery standards meet consumer expectations.

Query 4

	order_id	distinct_categories_in_cart	cart_value
1	32	2	22659.35
2	37	2	19542.55
3	79	2	19474.55
4	217	2	18886.04
5	66	3	18279.50
6	200	2	17783.30
7	169	2	15983.04
8	209	3	15760.45
9	152	3	15041.85
10	132	3	14771.60

This output reveals that the highest-revenue orders are driven by multi-category shopping, as every top cart (valued $14k-$22k) contains items from at least 2 or 3 distinct categories. This validates cross-selling as the primary driver for high Average Order Value (AOV), indicating the business should prioritize cross-category bundling strategies to replicate this behavior and maximize revenue per transaction.

Query 5

	payment_method	avg_transaction_value	max_transaction_value
1	PayPal	1559.311136	4721.51
2	Transfer	1278.478000	3008.37
3	Card	1119.311761	4531.87
4	Other	923.250000	3908.51

This output shows that PayPal drives the highest revenue per transaction, with an average value of ~$1,559 and a maximum of ~$4,721, significantly outperforming cards and bank transfers. This insight is essential for checkout optimization, indicating that prioritizing PayPal visibility is key to converting high-ticket shoppers, while also necessitating a review of payment processing fees to ensure margins on these premium orders remain protected.

Query 6

	region	category	units_sold
1	East	Electronics	231
2	East	Grocery	126
3	East	Home	108
4	East	Apparel	89
5	North	Electronics	106
6	North	Grocery	70
7	North	Home	62
8	North	Apparel	43
9	South	Electronics	126
10	South	Grocery	76
11	South	Apparel	64
12	South	Home	59
13	West	Electronics	111
14	West	Grocery	78
15	West	Apparel	63
16	West	Home	48

This output shows a distinct regional hierarchy where the "East" dominates sales volume across every single category, led by Electronics (231 units), which nearly doubles the volume of the next-highest region. This insight is critical for inventory distribution, dictating that the bulk of stock, particularly for the universal "anchor" category of Electronics, must be allocated to Eastern warehouses to prevent stockouts, while signaling that underperforming regions like the North (lowest volume) require targeted marketing or promotional pricing to capture untapped market share.

Query 7

	region	inactive_customers
1	East	3
2	North	8
3	South	6
4	West	5

This output outlines the distribution of inactive customers across regions, identifying the North as having the highest number of dormant users (8), followed by the South (6). Monitoring these figures is important for the business because it highlights exactly where retention rates are softening, allowing the marketing team to focus their resources on specific areas with targeted re-engagement campaigns to recover valuable relationships before they are lost entirely.

Query 8

	gross_sales	total_refunds	net_revenue
1	289621.21	20832.71	268788.50

This output presents a high-level summary of financial performance, revealing that while gross sales reached roughly $289,621, refunds accounted for $20,832, resulting in a net revenue of $268,788. Monitoring these aggregate figures is essential for the business because it quantifies the exact

impact of returns on the bottom line (approximately a 7.2% revenue leakage), serving as the ultimate "source of truth" metric for accurate financial forecasting and cash flow management.

2.4 Marketing & Customer Insights

Query 1

	sales_year	sales_month	monthly_revenue
1	2024	1	9165.44
2	2024	2	14253.81
3	2024	3	15813.31
4	2024	4	25368.70
5	2024	5	20923.28
6	2024	6	27804.39
7	2024	7	19602.45
8	2024	8	13675.84
9	2024	9	9928.34
10	2024	10	20284.81
11	2024	11	10049.48
12	2024	12	13396.97
13	2025	1	12140.60
14	2025	2	12293.38
15	2025	3	18999.84
16	2025	4	14305.47
17	2025	5	10045.93
18	2025	6	21569.17

This output tracks monthly revenue performance over an 18-month period, revealing significant year-over-year volatility that demands immediate strategic review. While 2025 began with promising growth, January revenue increased from $9,165 to $12,140, the second quarter shows a concerning downward trend, with May revenue effectively halving from $20,923 in 2024 to $10,045 in 2025. Monitoring these variances is critical for the business because it isolates specific underperforming periods (like Q2 2025), suggesting the leadership to investigate whether the drop was caused by reduced marketing spend, stockouts, or competitive pressure, and to adjust forecasts accordingly.

Query 2

	email_opt_in	customer_count	avg_lifetime_spend
1	0	18	1343.467291
2	1	80	1172.576979

This output reveals a counter-intuitive trend where customers who opted *out* of emails actually have a higher average lifetime spend ($1,343) compared to subscribers ($1,173). While the opted-in group is significantly larger (80 vs. 18), this value discrepancy is critical for segmentation strategy as it suggests that the current email marketing content may be attracting lower-spending, discount-driven shoppers, while a cluster of high-value "VIP" customers prefers a low-touch relationship, necessitating a different, non-intrusive retention approach for this profitable segment.

Query 3

	customer_id	first_name	last_name	number_of_orders	total_spent
1	47	Name47	Surname47	5	10067.81
2	80	Name80	Surname80	2	8630.02
3	26	Name26	Surname26	4	8301.99
4	16	Name16	Surname16	4	7183.85
5	69	Name69	Surname69	5	6767.46
6	32	Name32	Surname32	4	6764.48
7	78	Name78	Surname78	3	6298.35
8	4	Name4	Surname4	4	6102.14
9	51	Name51	Surname51	3	5851.62
10	25	Name25	Surname25	3	5677.43

This output identifies the top-tier revenue drivers, highlighting that high spending does not always correlate with high frequency for instance, Customer 80 spent over 8,630 across just two transactions, nearly matching Customer 47 who required five orders to reach $10,067. Recognizing these "whales" is critical for Key Account Management strategies, as it allows the business to deploy white-glove service and exclusive loyalty perks to lock in this elite segment, whose retention is far more cost-effective and impactful on the bottom line than acquiring new customers.

Query 4

	channel	avg_session_time	longest_session
1	App	6.116421	29.45
2	Web	7.660333	27.36

This output reveals a distinct engagement gap where Web users spend more time on average (7.66 minutes) compared to App users (6.12 minutes), despite the App recording the single longest session. This distinction is vital for platform strategy because it suggests that customers use the Web interface for deep research and browsing, while the App is likely treated as a transactional tool for speed; consequently, the business should optimize the Web experience for content-rich discovery (e.g., detailed product guides) while streamlining the App specifically for friction-free, rapid checkout.

Query 5

	birth_year	total_spent_by_cohort
1	1964	23724.30
2	1971	21172.40
3	1989	16085.55
4	1968	15605.29
5	1949	13335.67
6	1974	12285.63
7	1986	11450.99
8	2000	9040.69
9	1975	8627.49
10	1982	8457.28

This output highlights a strong revenue concentration among older generations, with the 1964 and 1971 age groups leading total spend at approximately $23,724 and $21,172 respectively. Understanding this demographic split is vital for marketing alignment, as it indicates that while younger shoppers (e.g., born in 2000) are active, the highest disposable income, and thus the immediate revenue opportunity, lies with Gen X and Boomers. Consequently, the business should ensure its user interface and product recommendations are optimized for these mature, high-value customers rather than focusing exclusively on youth-oriented trends.

Query 6

	total_interactions	total_orders	conversion_ratio_proxy
1	416	240	0.5769230769230

This output establishes a critical baseline for funnel efficiency, revealing a conversion proxy of approximately 57.7% derived from 240 orders against 416 interactions. Monitoring this ratio is indispensable for the business because it acts as the primary check on user experience, maintaining such a high conversion rate is key to profitability, and any sudden fluctuation would serve as an immediate alert to investigate technical friction (like payment failures) or shifts in traffic quality that could threaten revenue goals.

Query 7

	brand	category	items_sold
1	Globex	Electronics	92
2	Stark	Electronics	92
3	Wayne	Electronics	92
4	Wonka	Apparel	81
5	Soylent	Home	72
6	Acme	Electronics	69
7	Soylent	Grocery	67
8	Aperture	Electronics	54
9	Acme	Apparel	52
10	Acme	Grocery	46

This output highlights a highly competitive "power trio" within the Electronics sector, where Globex, Stark, and Wayne are tied for market dominance with exactly 92 units sold each. Identifying these volume leaders is critical for supply chain resilience, as it confirms that Electronics is the primary anchor category driving throughput; consequently, the business must prioritize stock availability for these specific vendors to prevent costly stockouts, while simultaneously leveraging the cross-category popularity of brands like "Acme" and "Soylent" to negotiate better trade terms or volume discounts.

Query 8

	loyalty_tier	support_requests
1	Plus	8
2	Basic	6
3	Premium	3

This output reveals a discrepancy in support volume, with the "Plus" loyalty tier generating the highest number of inquiries (8), nearly triple the volume of the top-tier "Premium" segment (3). Monitoring these metrics is vital for operational efficiency, as it suggests that the mid-tier "Plus" experience may be causing specific friction or confusion, necessitating a targeted review of onboarding materials or benefits clarity for this group to reduce service costs and improve satisfaction.

3. Data Preparation and Feature Engineering

3.1 Dataset overview

The dataset used for data preparation and feature engineering is provided in the file *exam.csv*. It represents a customer-level dataset from a mid-sized European e-commerce retailer and combines information related to customer demographics, transactional behavior, digital engagement, and operational performance.

In its original form, the dataset consists of 6,000 observations and 37 variables, with each observation representing an individual customer. After data preparation and the encoding of categorical variables, the final analytical dataset contains 6,000 observations and 39 variables.

A large proportion of the dataset consists of numerical variables describing key dimensions of customer behavior and business performance. These include demographic and lifecycle indicators such as *age* and *tenure_months*, transactional measures such as *total_orders*, *total_spent*, *avg_order_value*, *median_order_value*, and *max_order_value*, as well as recency and activity metrics including *recency_days*, *days_since_first_order*, *orders_last_3m*, and *orders_last_1m*.

Product-related behavior is captured through variables such as *distinct_products*, *category_diversity*, and category share measures, while operational performance is reflected in variables including *avg_delivery_days* and *on_time_rate*. Customer engagement is measured using digital interaction variables such as *email_open_rate*, *email_click_rate*, *website_sessions_3m*, and *app_sessions_3m*. In addition, the dataset includes a forward-looking variable, *future_3m_spend*.

Alongside numerical features, the dataset contains several categorical variables, most notably customer region and loyalty tier. These variables provide contextual information related to customer segmentation and geographic distribution but require transformation before they can be incorporated into statistical or machine learning models.

The dataset also includes identifier and contact-related variables, such as *customer_id*, *first_name*, *last_name*, *email*, and *phone*. While relevant for operational and customer relationship management purposes, these variables do not contribute analytical value for modeling customer behavior or predicting future outcomes. They were therefore removed during the data preparation process.

An initial inspection of the dataset revealed missing values across multiple variables, as well as substantial variability and extreme values in several numerical features. Outlier detection was performed using the interquartile range (IQR) method, which identified a considerable number of extreme observations across multiple variables.

For example, high variability was observed in engagement-related variables such as *email_open_rate* (485 potential outliers) and *email_click_rate* (171 potential outliers), as well as in transactional and behavioral variables including *total_spent* (417 outliers), *returns_count* (382 outliers), *orders_last_1m* (307 outliers), and *future_3m_spend* (353 outliers). These patterns reflect the inherent heterogeneity of e-commerce customer behavior, where a relatively small subset of customers may be highly active or high-spending, while others exhibit limited engagement. For this reason, extreme values were retained instead of being removed.

Following data preparation steps including the removal of irrelevant identifiers, imputation of missing values, inspection of outliers, and encoding of categorical variables, the final prepared dataset was saved as *exam_prep.csv*. This dataset serves as the standardized input for subsequent correlation analysis, principal component analysis, and predictive modeling.

The prepared dataset provides a solid foundation for analyzing customer behavior, understanding variation in spending and engagement patterns, and developing predictive models related to future spending and customer churn.

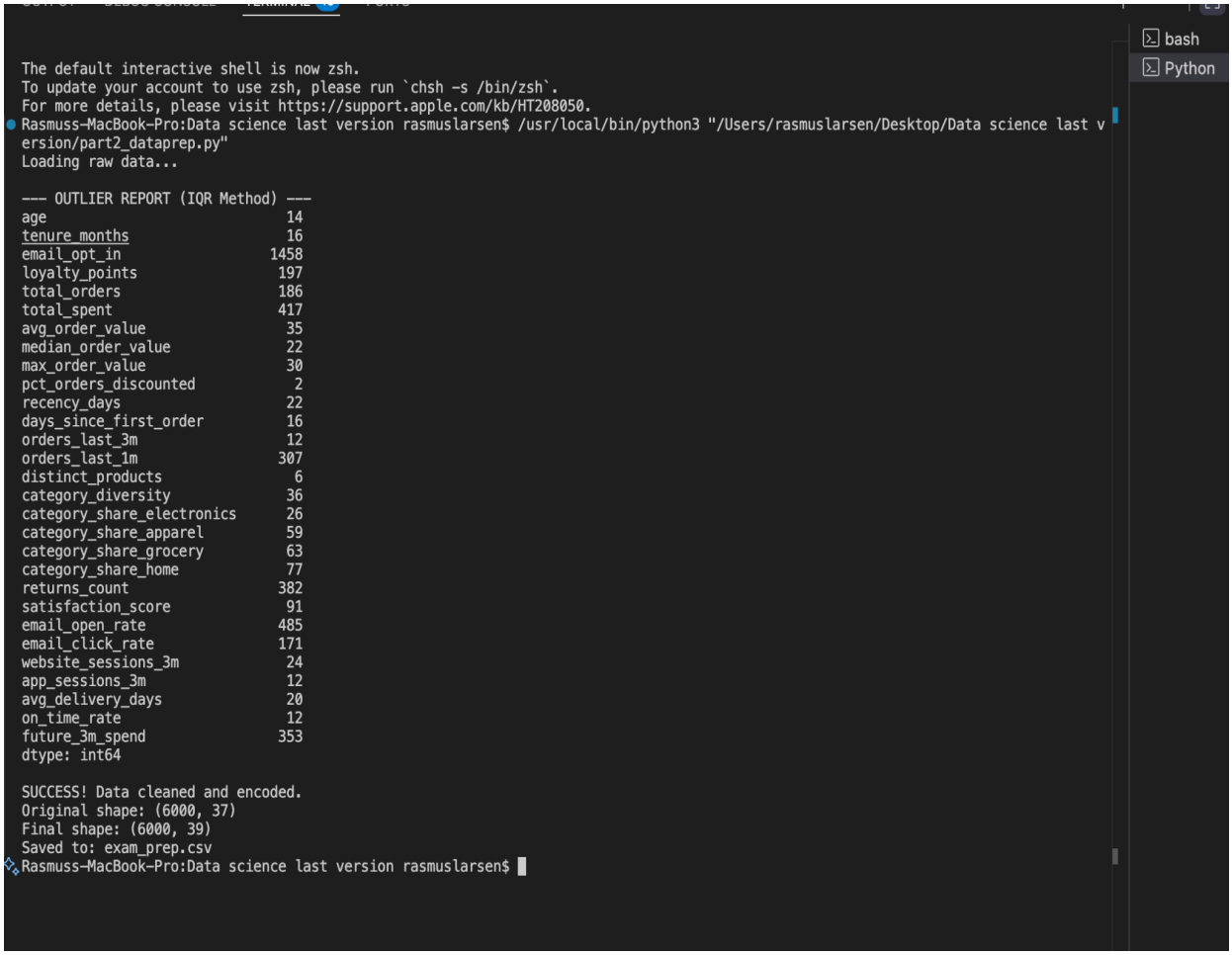


Fig. 17(output from code)

4. Exploratory Analysis, Correlation, and PCA

This section presents the exploratory data analysis conducted on the prepared dataset, with a particular focus on correlation analysis and dimensionality reduction using Principal Component Analysis (PCA). The purpose of this analysis is to identify underlying relationships between variables, assess multicollinearity, and evaluate whether a reduced set of components can capture the

majority of the variance in the data.

4.1 Principal Component Analysis (PCA)

To address multicollinearity and gain a structured understanding of the high-dimensional feature space, Principal Component Analysis (PCA) was applied to the fully prepared dataset (*exam_prep.csv*). Prior to PCA, all variables were standardized to ensure that features measured on different scales contributed equally to the analysis.

The PCA produced 39 principal components, corresponding to the number of input features. The explained variance ratios show that the first principal component (PC1) accounts for 19.36% of the total variance, while the first two components combined (PC1 + PC2) explain 28.93% of the variance. These results indicate that no single component dominates the variance structure; however, the leading components still capture a meaningful share of the information contained in the dataset.

This variance distribution is typical for behavioral and transactional datasets, where customer characteristics are influenced by multiple interacting factors rather than a single underlying dimension

4.2 Component Loadings and Interpretation

An examination of the component loadings for the first three principal components reveals distinct underlying patterns in the data.

The first principal component (PC1) is primarily driven by customer value and engagement-related variables, most notably *loyalty_points*, which exhibits a substantially higher loading than the remaining features. This suggests that PC1 can be interpreted as a customer value or loyalty dimension, capturing variation related to accumulated rewards and long-term engagement behavior.

The second principal component (PC2) shows relatively higher loadings for *tenure_months*, while loyalty-related variables contribute negatively. This pattern indicates that PC2 reflects a customer lifecycle or tenure dimension, separating newer customers from more established ones.

The third principal component (PC3) displays more moderate and distributed loadings across several variables, suggesting a more nuanced component that captures secondary behavioral patterns rather than being dominated by a single feature.

Categorical variables introduced through one-hot encoding, such as region and gender indicators, exhibit relatively small loadings across all principal components. This indicates that numerical behavioral variables contribute more strongly to the overall variance structure than demographic attributes in this dataset.

Overall, The PCA results show that variance is distributed across multiple latent dimensions, with customer value, engagement, and tenure emerging as the most influential structures. Although PCA is not used directly for dimensionality reduction in subsequent modeling steps, it provides valuable insight into the underlying structure of the data and confirms the presence of correlated behavioral features.

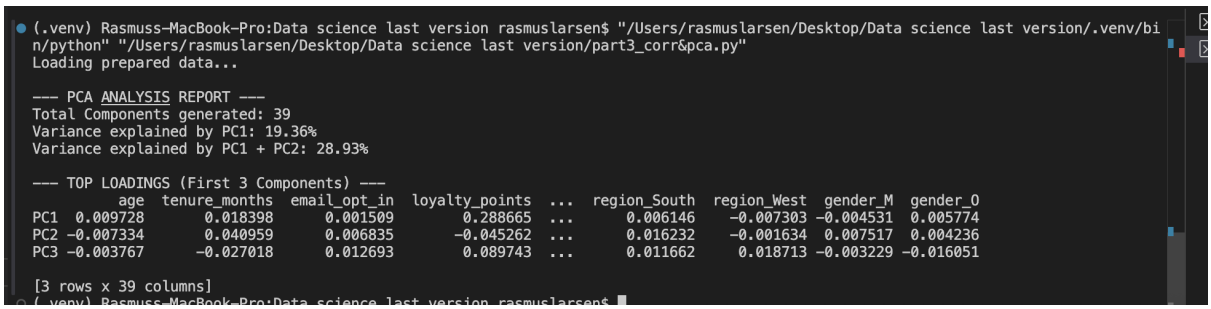


Fig. 18 (PCA analysis)

4.3 Screeplot

The figure under here presents the PCA scree plot with cumulative explained variance. The first principal component (PC1) explains approximately 19.4% of the total variance, while the first two components combined explain 28.9%. The cumulative variance increases gradually across components, reflecting a dataset where information is distributed across multiple correlated behavioral features.

The scree plot shows that approximately 21 principal components are required to reach 90% cumulative explained variance, indicating that no sharp “elbow” is present. This suggests that customer behavior in the dataset is multi-dimensional and cannot be effectively summarized by only a small number of components.

Given this structure, PCA is not used as a strict dimensionality reduction technique for modeling, but rather as an exploratory tool to understand variance structure and confirm the presence of multicollinearity among behavioral variables.

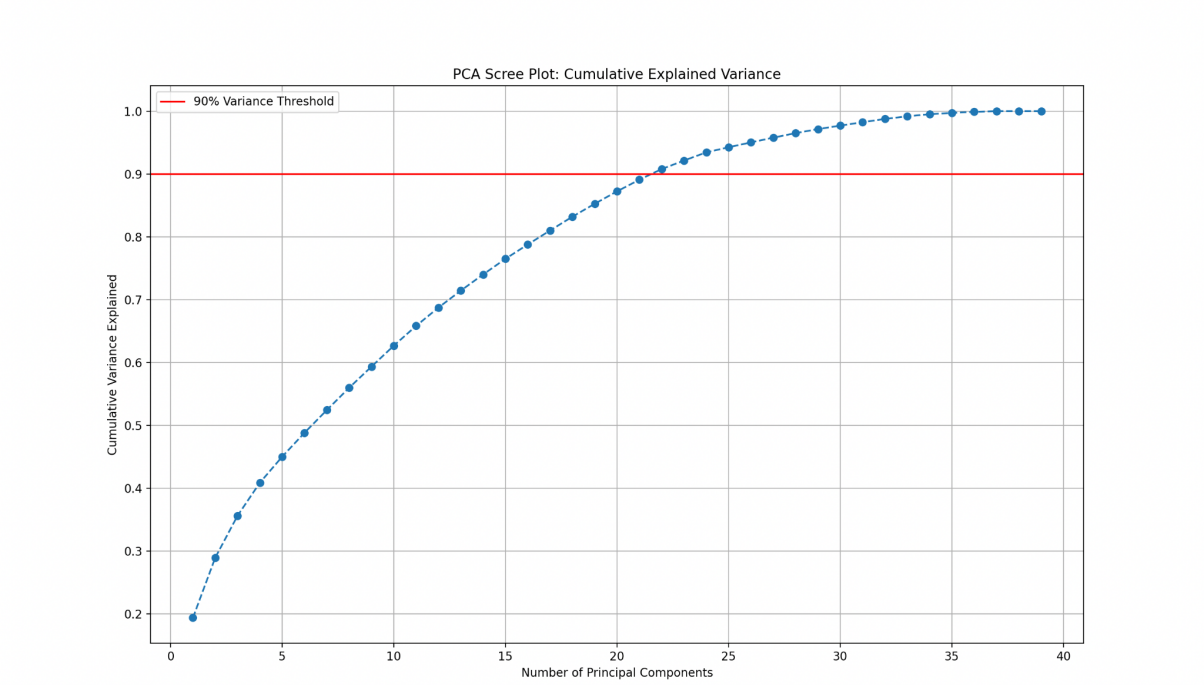


Fig. 19 (Scree plot)

4.4 Correlation matrix

Figure 20 shows the correlation matrix including both numerical and one-hot encoded variables. Several clear correlation structures are evident, confirming the presence of multicollinearity in the dataset. Strong positive correlations are observed among customer value and purchasing behavior variables, including *total_orders*, *total_spent*, *average_order_value*, *median_order_value*, and *max_order_value*. These relationships indicate overlapping information content related to purchasing intensity and customer value.

Engagement-related variables such as *email_open_rate*, *email_click_rate*, *website_sessions_3m*, and *app_sessions_3m* also display moderate positive correlations, reflecting consistent behavioral engagement patterns across channels.

Conversely, recency-based variables (e.g., *recency_days* and *days_since_first_order*) show negative correlations with spending and engagement metrics, which aligns with expected customer behavior: more recent activity is associated with higher engagement and future spending.

The correlation matrix further reveals meaningful relationships between returns-related variables (*returns_count*, *returns_rate*), *complaint_count*, and *satisfaction_score*. Higher return and complaint rates are negatively correlated with satisfaction and positively associated with churn-related indicators.

Notably, *churn_90d* shows negative correlations with spending, engagement, and satisfaction metrics, while being positively associated with recency and lower activity levels. This supports the behavioral interpretation of churn risk in the dataset.

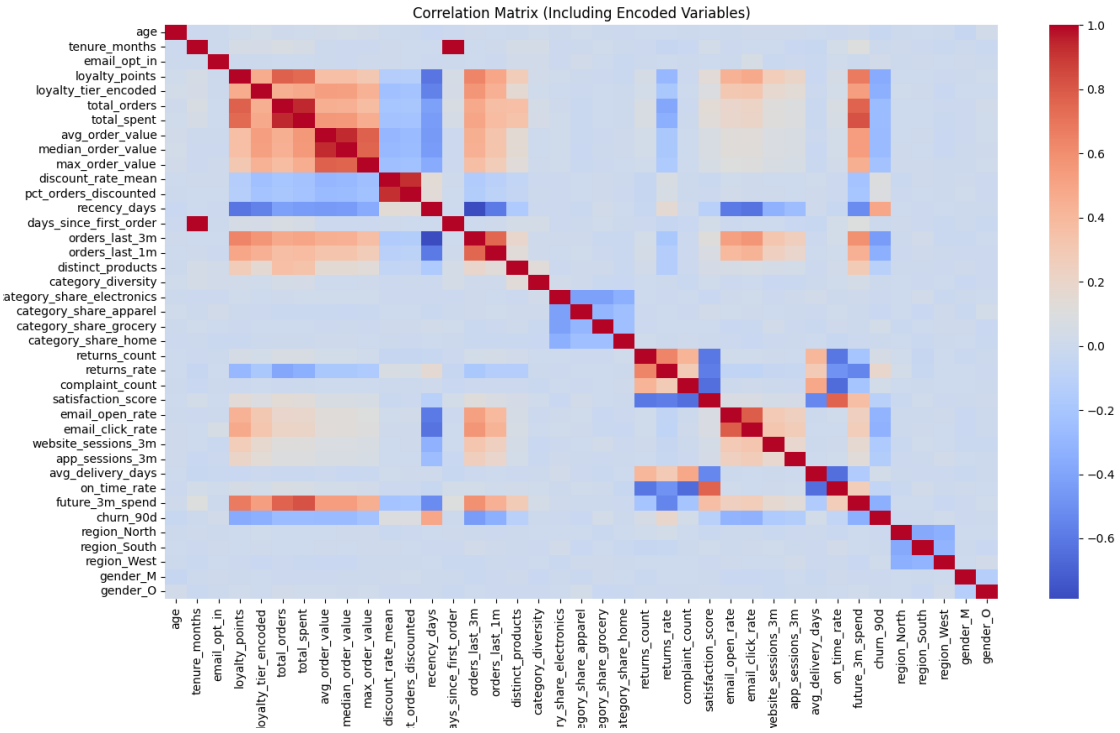


Fig. 20(correlations matrix)

4.5 Summary of Exploratory Findings

Together, the PCA and correlation analysis confirm that the dataset contains substantial multicollinearity, particularly among value, engagement, and purchasing-related variables. PCA demonstrates that variance is distributed across multiple latent dimensions rather than dominated by a single factor, while the correlation matrix provides interpretable, domain-relevant relationships between key features.

These findings justify the use of regularized regression models and multivariate techniques in later modeling stages, where correlated predictors must be handled carefully to ensure stable and interpretable results.

5. Prediction Goals

The purpose of this part is to define and justify the prediction targets used in the predictive modeling part. The selected targets are chosen based on the relevance for the business and the availability in the dataset, and alignment with common decision-making challenges in an e-commerce store.

For this analysis, churn_90d is selected as the primary target variable for the classification task. This variable defines a binary classification problem, where the objective is to predict whether a customer will stop purchasing within the next 90 days (1) or remain active (0).

5.1 Business Justification

Customer churn prediction is an important component of data-driven decision-making in retail, as customer retention is a key driver of long-term profitability. It is well documented in marketing and customer analytics literature that acquiring new customers is significantly more expensive than retaining existing ones. Estimates from *Harvard Business Review* suggest that customer acquisition costs can be between five and twenty-five times higher than retention costs (Reichheld & Sasser, 2014). As a result, even small improvements in retention rates can have a substantial impact on business performance.

By predicting customer churn before it occurs, companies can shift from a reactive to a proactive retention strategy. Instead of responding after customers have already disengaged, marketing teams can intervene earlier through targeted actions such as personalized offers or re-engagement campaign.

6. Predictive Modelling

As we chose churn_90d, which is a binary variable, we performed classification modelling to proactively identify customers at risk of leaving. We implemented both Logistic Regression and Decision Tree algorithms to train models capable of differentiating between retained and churned users based on historical data. To further enhance our strategy, we utilized unsupervised K-Means clustering to statistically divide the customer base into distinct segments. This combination allows us to predict exactly which customers might leave and group them by behavior, making our marketing campaigns much more effective and personalized.

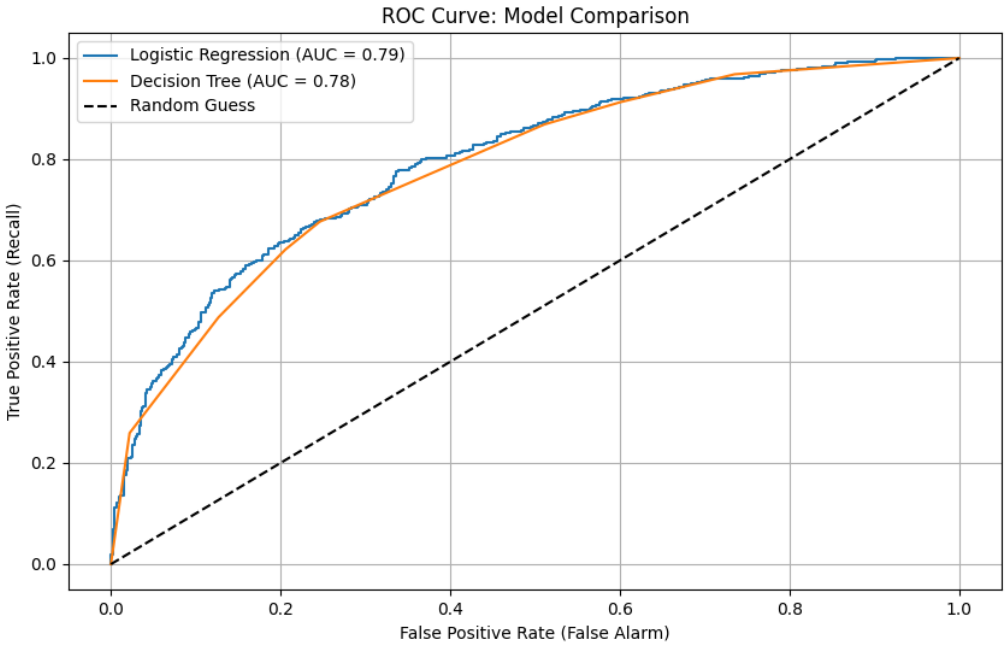


Fig. 21 (ROC Curve)

The first thing we did was conduct an ROC curve analysis to evaluate and compare the performance of two classification models: Logistic Regression and a Decision Tree. The results indicate that both models perform well, significantly outperforming the random guess baseline. The Logistic Regression model achieved a slightly higher Area Under the Curve (AUC) of 0.79, compared to 0.78 for the Decision Tree, suggesting a marginal advantage in predictive capability. Visually, the Logistic Regression curve appears smoother, reflecting its continuous probability output, whereas the Decision Tree curve exhibits a stepped pattern typical of discrete leaf-node predictions. To complement the ROC analysis, we calculated the overall accuracy scores, which yielded consistent results: the Logistic Regression model attained an accuracy of 71.33%, edging out the Decision Tree's 71.17%. This convergence of metrics confirms that, while performance is comparable, the Logistic Regression model offers a slightly more robust and stable solution for this specific dataset.

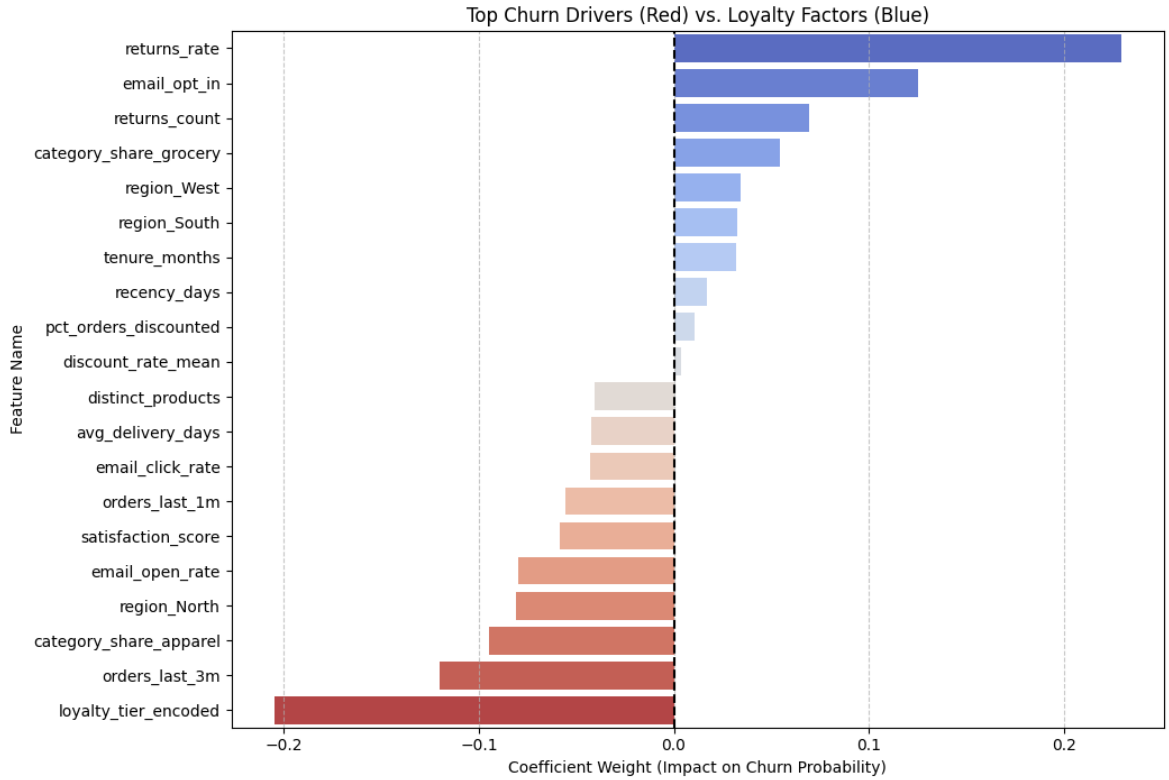


Fig. 22 (Top churn driver)

To understand the 'why' behind the predictions, we analyzed the Logistic Regression coefficients to identify the specific features that most strongly influence churn probability. The analysis reveals a clear distinction between risk factors and retention drivers. On the positive side of the axis (represented by the **Blue bars**), we observe the factors that increase the likelihood of churn; notably, **return behavior** (returns_rate and returns_count) emerges as the strongest predictor of customer attrition, indicating that customers who frequently return items are at the highest risk of leaving. Interestingly, email_opt_in also appears as a strong churn driver, whereas email_open_rate (found on the opposite side) is a retention factor; this suggests that while simply being on the mailing list correlates with higher churn (perhaps due to "inbox fatigue"), actual *engagement* with the content is protective. Conversely, the negative coefficients (represented by the **Red bars**) highlight the key drivers of **loyalty**. The variable loyalty_tier_encoded shows the most significant protective effect, followed by orders_last_3m (recent order volume) and category_share_apparel. This indicates that customers with a higher loyalty status and those shopping for apparel are significantly more likely to be retained compared to those shopping for groceries (category_share_grocery), which appears as a risk factor.

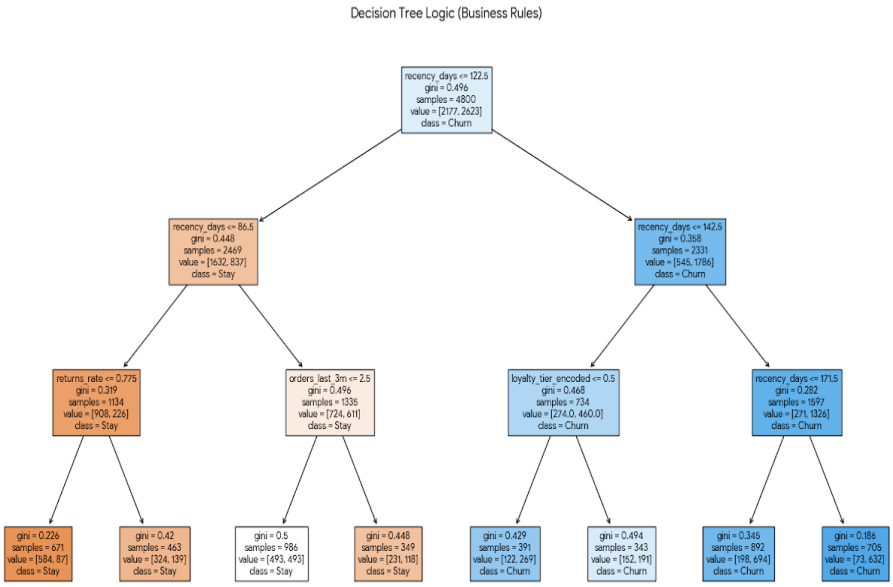


Fig. 23(Decision tree logic)

To complement the statistical insights from the Logistic Regression, we visualized the Decision Tree to uncover the specific logic and "business rules" the model uses to classify customers. The tree structure immediately identifies Recency(recency_days) as the dominant splitting criterion at the root node. The primary rule is simple yet powerful: customers who have not made a purchase in over 122.5 days (the right branch, predominantly **Blue**) are immediately pushed toward high-risk churn categories. Conversely, customers with recent activity (left branch, **Orange**) are generally categorized as "Stay," but with caveats. For these active users, the model subsequently evaluates Returns (returns_rate) and Purchase Frequency (orders_last_3m) to refine its prediction. This visualization clarifies the hierarchy of risk: while "High Returns" is a significant behavioral driver, **Inactivity** is the foundational signal. Essentially, if a customer goes silent for four months, their specific return habits matter less, they are already in the danger zone.

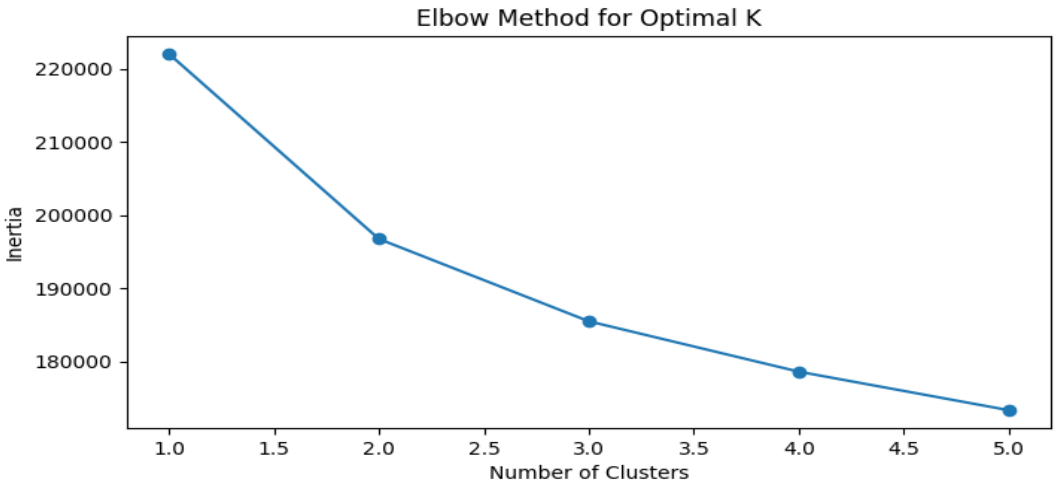


Fig. 24 (Elbow method for optimal K)

Moving beyond the binary prediction of "Churn" vs. "Stay," we employed Unsupervised Learning (Cluster Analysis) to uncover natural distinct groupings within the customer base. To determine the optimal number of segments, we utilized the Elbow Method, which plots the "Inertia" (a measure of how internally coherent the clusters are) against the number of clusters. The chart reveals a steep decline in inertia as we move from 1 to 2 clusters, confirming that significant differences exist among customers. The critical "elbow" point, where the curve begins to flatten and the marginal gain of adding another cluster diminishes, appears most distinct at K=3. While the curve continues to smooth out slightly towards 4, selecting **3 clusters** offers the ideal balance between statistical precision and operational practicality. This suggests the customer base naturally organizes into three primary archetypes, allowing the marketing team to design three targeted strategies rather than diluting efforts across too many micro-segments.

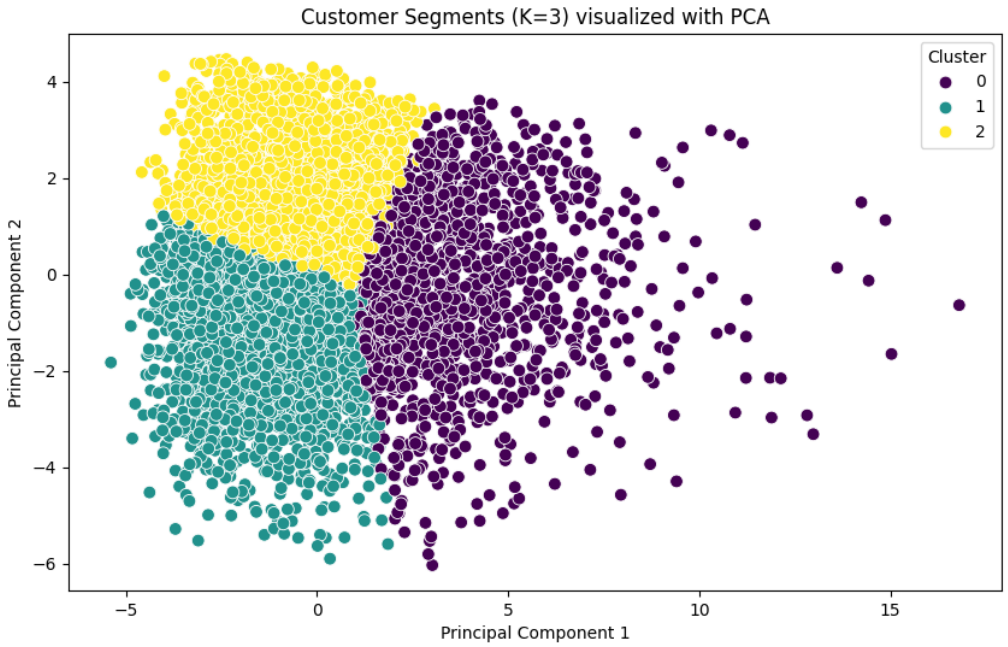
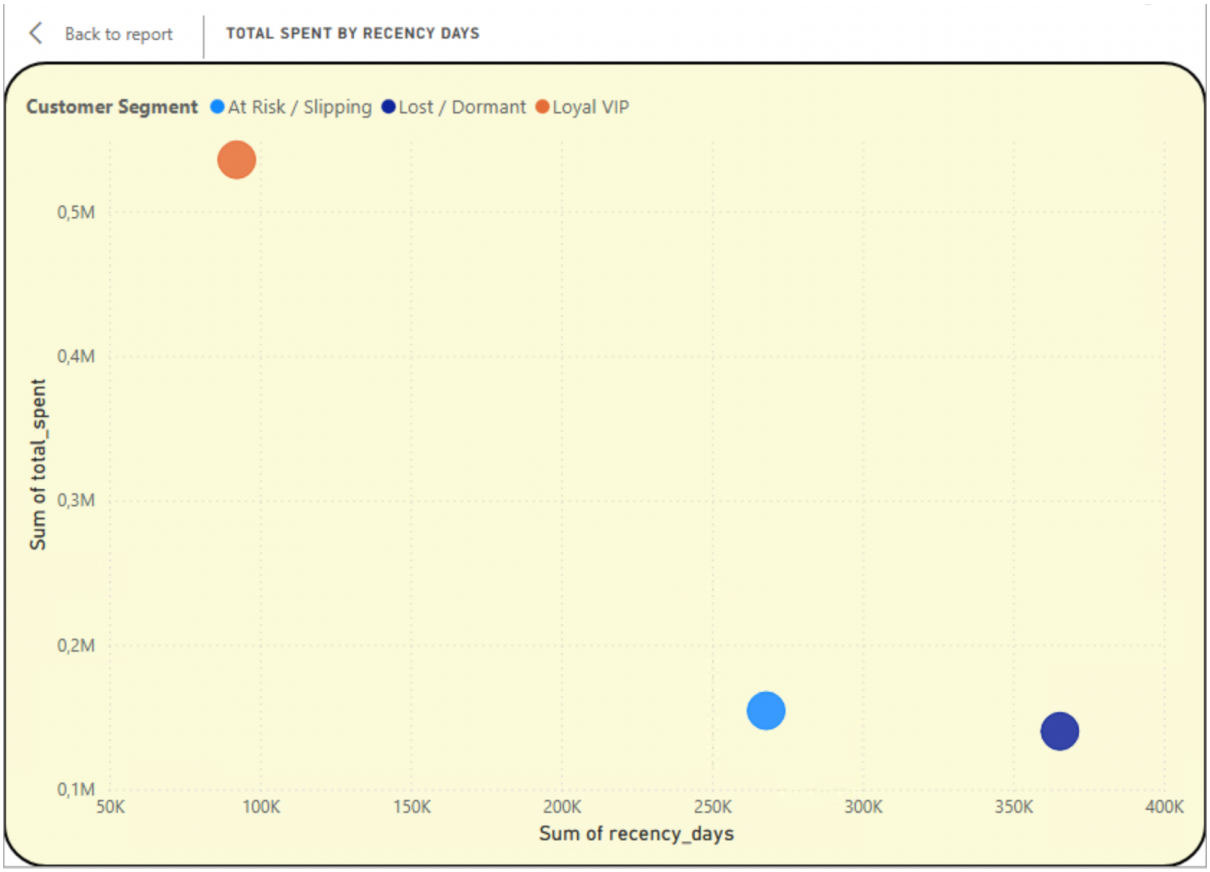


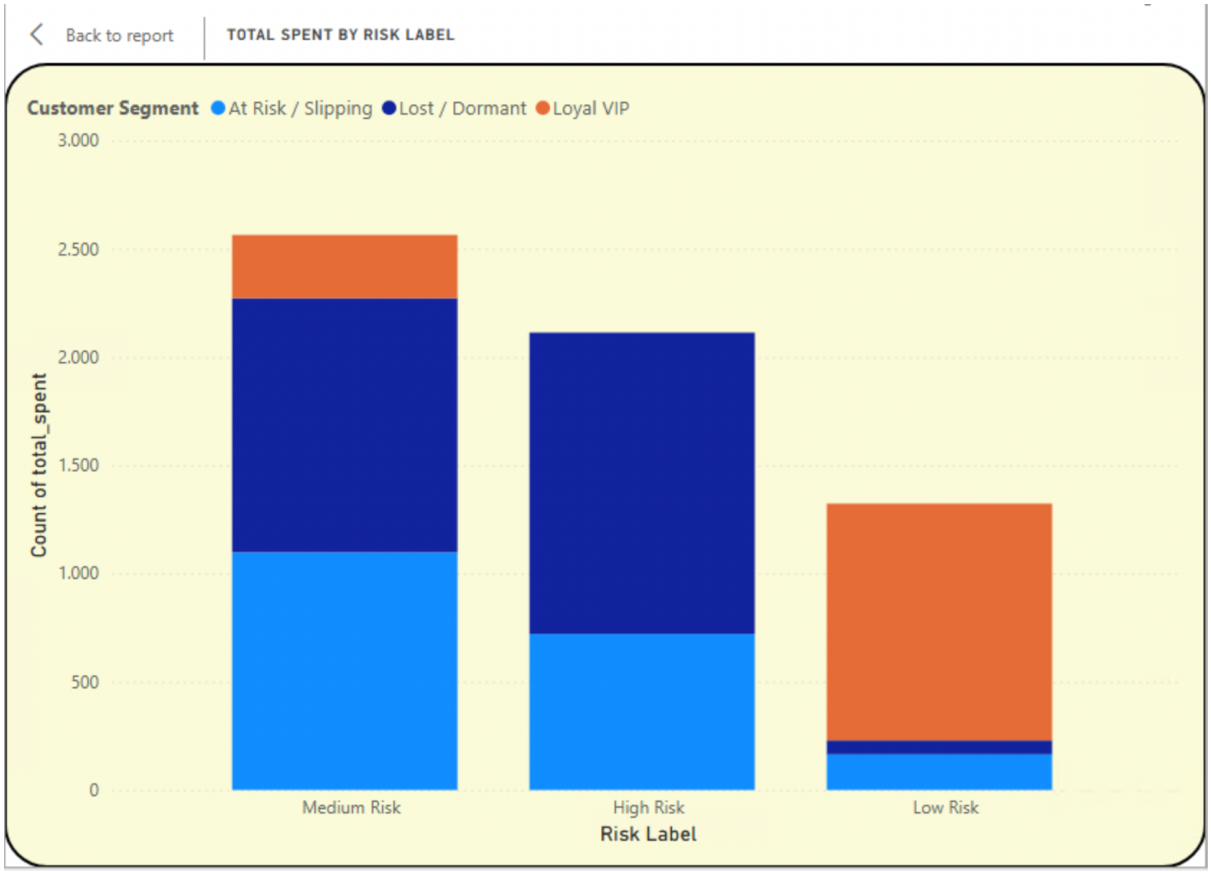
Fig. 25 (Customer segment visualized with PCA)

To validate the mathematical findings of the Elbow Method, we visualized the resulting segments using **Principal Component Analysis (PCA)** to project the multi-dimensional customer data into a 2D space. The resulting scatter plot confirms that the chosen **K=3** configuration yields three distinct and well-separated customer groups. The visual distinction is sharp: the **Yellow (Cluster 2)** and **Teal (Cluster 1)** groups appear densely clustered in the upper-left and lower-left quadrants respectively, indicating high internal similarity, these customers likely share very consistent behaviors regarding spending and frequency. In contrast, the **Purple (Cluster 0)** group spreads broadly across the right side of the chart, suggesting a more heterogeneous group with higher variance in their behavioral patterns. This clear spatial separation validates that the segmentation is not just a statistical artifact, but reflects real, underlying differences in customer behavior, providing a solid foundation for targeted marketing strategies.

7. PowerBI Dashboard



Now we moved on to Power BI for additional visualization to translate our statistical clusters into actionable business segments. By plotting Total Spent against Recency Days, the chart reveals a distinct and highly informative contrast between the three identified groups. The **Loyal VIP** segment (Orange bubble) occupies the ideal 'North-West' position, characterized by high cumulative spending (approx. 0.55M) and low recency scores, confirming these are our most active and valuable customers. Conversely, the **Lost / Dormant** segment (Dark Blue bubble) is isolated in the bottom-right, representing customers with the highest inactivity (over 350K aggregate recency days) and low monetary value. Crucially, the **At Risk / Slipping** segment (Light Blue bubble) sits in the middle ground; these customers show low spend and dangerously increasing inactivity, marking them as the priority target for immediate re-engagement campaigns before they drift permanently into the 'Dormant' category.



To bridge the gap between our machine learning output and business strategy, we generated a **'Risk Label'** variable based directly on the churn_prob (churn probability) column calculated by our Logistic Regression model. We then cross-referenced these risk tiers with our customer segments to validate the targeting. The resulting stacked bar chart reveals a crucial insight regarding the **Medium Risk** category, which stands out as the highest column by volume. While the **Low Risk** column is correctly dominated by **Loyal VIPs** (Orange) and **High Risk** by **Lost/Dormant** users (Dark Blue), the **Medium Risk** group represents a complex 'Uncertainty Zone'. Here, we see a significant presence of **Lost/Dormant** customers (Dark Blue) alongside the expected **At Risk** users (Light Blue). This anomaly suggests that many 'Lost' customers historically spent large amounts; this high monetary value likely acted as a protective factor in the model, suppressing their calculated churn probability just enough to keep them out of the 'High Risk' bucket and landing them in the Medium tier instead. Consequently, this Middle column represents the strategic 'battleground': it contains high-value customers who have drifted away but, according to the model's probability scores, are not yet fully 'unrecoverable.'

Finally, we used the Decomposition Tree in Power BI to understand exactly *why* customers are at risk.

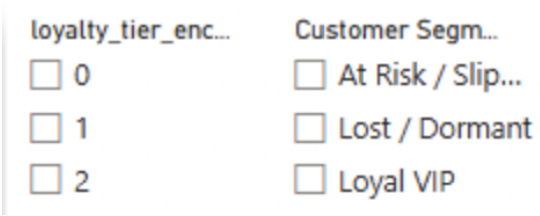
To clarify the starting point: the "Count of Risk Label" simply represents the number of customers (6,000) in our analysis. From here, the tool breaks down this total to find the root causes of churn.

The chart clearly shows that the main driver is Returns. The largest branch reveals that 3,404 customers have a returns_rate of 1.0, meaning they return 100% of what they buy.

Drilling down further, we see a surprising pattern, these customers typically have a low Complaint Count (often just 1) and a relatively decent Satisfaction Score of 6.7.

In conclusion, this is a critical insight. The score of 6.7 proves that these customers are not leaving because of bad service or anger. Instead, the issue is purely behavioral, because they return every single item they buy, they never complete a successful purchase cycle. This habit prevents them from being "active" customers, leading the model to eventually classify them as "Lost / Dormant" despite their positive satisfaction scores.

Customer Segment	Risk Label	%GT Sum of churn_prob	total_spent	email_opt_in
Lost / Dormant	High Risk	15,80%	0,00	1
At Risk / Slipping	High Risk	7,23%	0,00	1
Lost / Dormant	Medium Risk	6,65%	0,00	1
At Risk / Slipping	Medium Risk	5,12%	0,00	1
Lost / Dormant	High Risk	4,71%	0,00	0
Lost / Dormant	Medium Risk	2,45%	0,00	0
At Risk / Slipping	Medium Risk	2,08%	0,00	0
At Risk / Slipping	High Risk	1,61%	0,00	0
Loyal VIP	Low Risk	0,39%	0,00	1
At Risk / Slipping	Low Risk	0,32%	0,00	1
Loyal VIP	Medium Risk	0,18%	0,00	1
At Risk / Slipping	Low Risk	0,14%	0,00	0
Lost / Dormant	Low Risk	0,11%	0,00	1
Lost / Dormant	Low Risk	0,10%	0,00	0
Loyal VIP	Low Risk	0,08%	0,00	0
Lost / Dormant	High Risk	0,05%	51,30	1
Lost / Dormant	High Risk	0,05%	50,46	1
At Risk / Slipping	High Risk	0,05%	12,12	1
Lost / Dormant	High Risk	0,05%	42,38	1
Lost / Dormant	High Risk	0,05%	89,74	1
Lost / Dormant	High Risk	0,05%	107,94	1
Lost / Dormant	High Risk	0,05%	15,96	1
At Risk / Slipping	Medium Risk	0,04%	150,47	1
At Risk / Slipping	Medium Risk	0,03%	338,15	0
At Risk / Slipping	Medium Risk	0,03%	57,73	1



Finally, to turn these insights into a practical tool for the marketing team, we created a Detailed Customer List with dynamic filters.

We added Slicers for Customer Segment and Loyalty Tier. This allows the team to instantly filter the data, for example, selecting only "At Risk" customers to generate an immediate contact list.

The table itself is sorted by the Contribution to Churn Probability (%GT Sum of churn_prob) to show the biggest risks first. The results confirm our previous findings: the top rows are filled with "Lost" and "At Risk" customers who currently have a total_spent of 0.00. Crucially, notice that the email_opt_in column is frequently 1.

This table proves that our highest-risk users are technically "subscribers" (they receive emails) but they spend nothing. This list provides the specific targets for a "Win-Back" campaign to reactivate them, or alternatively, to clean them from the database to improve email engagement metrics.

8. Conclusion and Recommendations

This project applied a full data science workflow to analyze customer behavior, identify churn drivers, and support data-driven decision-making for a mid-sized European e-commerce retailer. By combining SQL-based business analysis, exploratory data analysis, PCA, predictive modeling, clustering, and visualization, the project demonstrates how analytical techniques can be translated into actionable business insights.

The SQL analysis revealed several critical operational and commercial findings. Deep discounts were shown to provide limited additional sales volume compared to small discounts, indicating that aggressive price reductions unnecessarily decrease profit margins. Logistics performance emerged as a major risk factor, with consistently high late delivery rates across carriers, highlighting the need for improved service-level agreements or alternative shipping strategies. In addition, high-value orders were strongly associated with cross-category purchases, confirming cross-selling as a key driver of revenue growth.

Exploratory analysis confirmed substantial multicollinearity among spending, engagement, and activity-related variables. PCA showed that customer behavior is multi-dimensional, with variance distributed across several latent factors related to customer value, engagement, and lifecycle stage. These findings justified the use of regularized and multivariate modeling approaches in subsequent analyses.

For predictive modeling, churn within 90 days was selected as the primary classification target due to its direct business relevance. Both Logistic Regression and Decision Tree models achieved strong performance (AUC ≈ 0.79 and 0.78), demonstrating that churn risk can be predicted reliably using historical customer data. Model interpretation revealed that return behavior and customer inactivity are the strongest churn drivers, while loyalty status, recent purchases, and engagement act as key retention factors.

Customer segmentation using K-means clustering identified three distinct and interpretable customer groups: Loyal VIPs, At Risk / Slipping customers, and Lost / Dormant customers. Combining these segments with churn probabilities in Power BI revealed a strategically important “Medium Risk” group, consisting largely of previously high-value customers who have become inactive. This group represents the highest potential return on targeted retention efforts.

Based on these findings, the key recommendations are to shift toward proactive churn prevention, reduce excessive return behavior, optimize discount strategies, improve logistics reliability, and focus marketing efforts on high-risk but high-value customer segments.

List of References

Reichheld, F. F., & Sasser, W. E. (2014). *The value of keeping the right customers*. Harvard Business Review.
https://hbr.org/2014/10/the-value-of-keeping-the-right-customers