Retailer Transaction

I. EDA

- Data cleaning
- Overview data
- Analysis:
  + Univariate analysis
  + Multivariate analysis
- RFM segments

II. Modeling

- Feature Engineering
- Setup & Training & Tuning
- Evaluation

III. Conclusion

- Insight analysis
- Evaluation of approaches to identify potential customers.

I. 1 - Data cleaning - Data overview

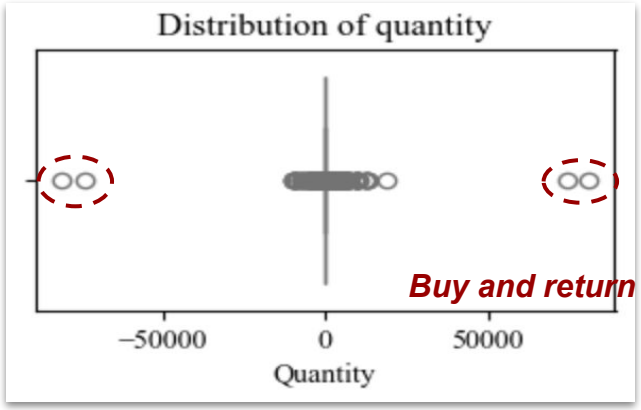
Dataset source: https://www.kaggle.com/datasets/cemalcici/online-retail-ii-uci-two-peroid

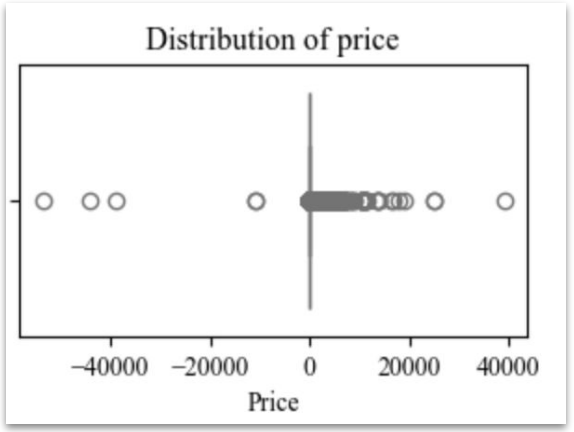
Basic information: covers transactions from a UK-based online retailer

- Total records: **1,067,371** total records, **34,335** were duplicates, leaving **1,033,036** unique entries.
- Time: 01/12/2009 - 09/12/2011
- Total columns: **8**

Column	Description
Invoice	Invoice number (C prefix = return order)
StockCode	Product code
Description	Product description
Quantity	Number of items
InvoiceDate	Date of transaction
Price	Price per item
CustomerID	Customer identifier
Country	Customer's country

I. 1 - Data cleaning - Outliers Detection





The chart indicates the presence of outliers in the data. However, our analysis reveals that only the following cases are considered true outliers:

- **Bad customers:** Customers with a purchase quantity of less than or equal to 0 (***99 customers***),
- **Returned products:** Items that were purchased and then returned, which hold no value in the transaction (*related* ***6,652 invoices***),
- **Bad debt:** Bad debt is a loan or outstanding balance that is considered uncollectible (*related* ***6 invoices***).

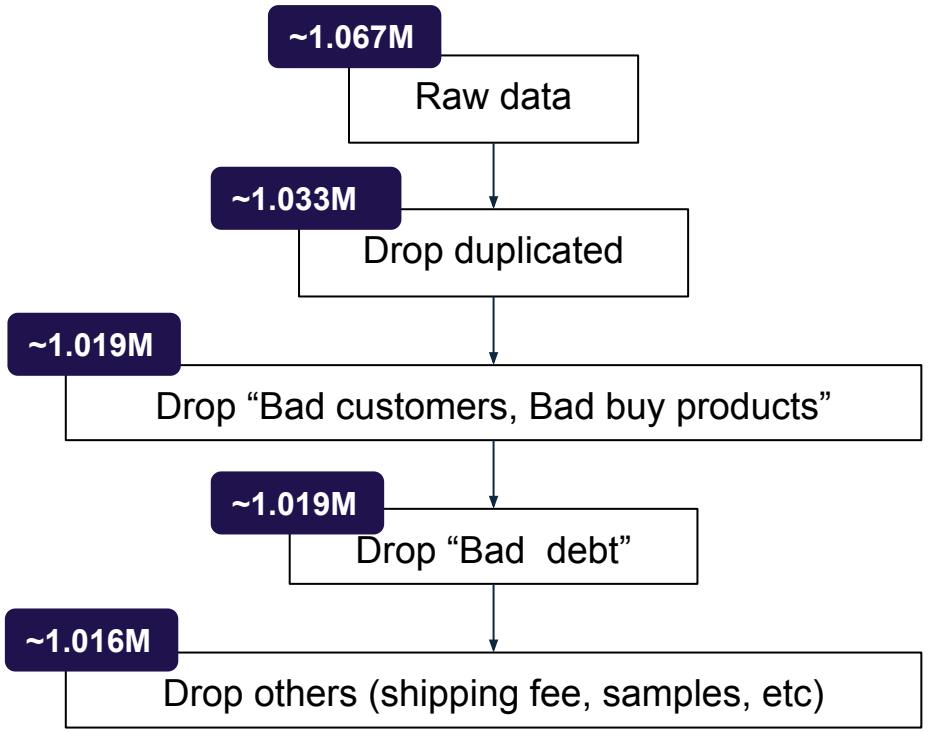
=> These outliers have been removed from the datas.

I. 1 - Data cleaning - Outliers Detection & Missing values

We removed all items that are not practically relevant, such as delivery charges and banking fees. Only information related to vouchers and gifts was retained to focus on evaluating the effectiveness of promotional programs.

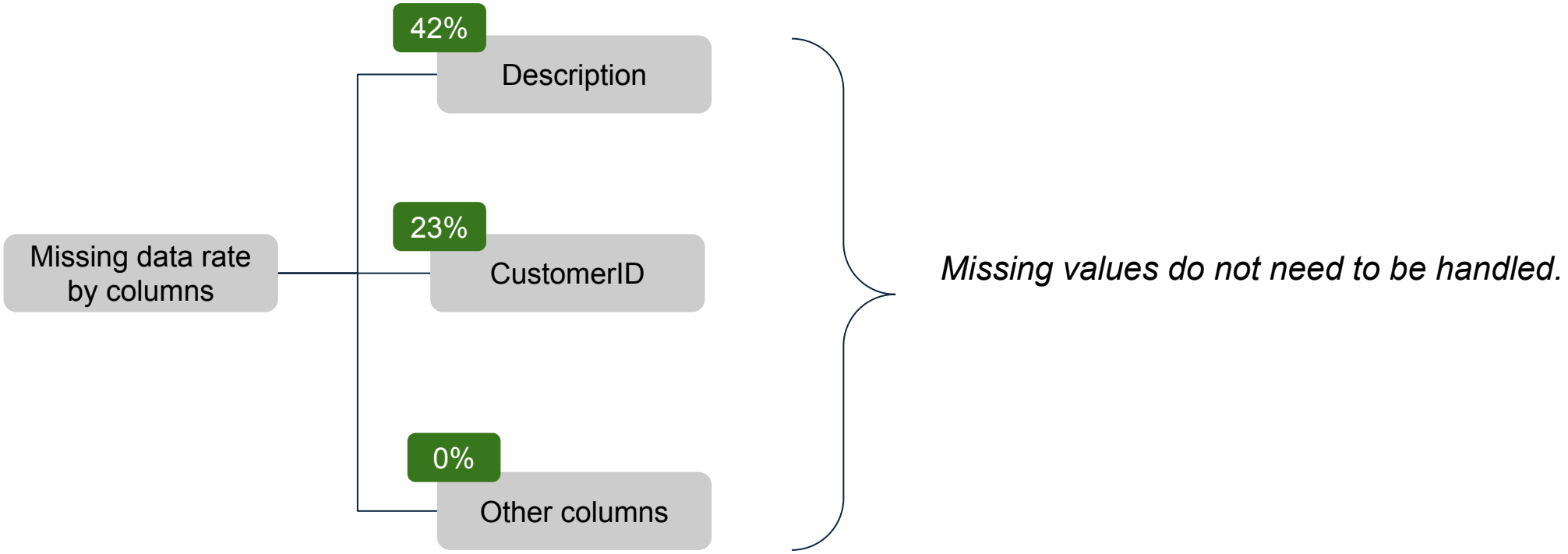
Stock code	Description
POST	POSTAGE (shipping fee)
D	Discount
DOT	Dot postage (shipping fee)
start: BANK CHARGES	bank charges
start: TEST	test products
start: GIFT, gift	Voucher
S	Samples
AMAZONEFREE	amazon free
CRUK	commission fee

Data Cleaning Flow Summary



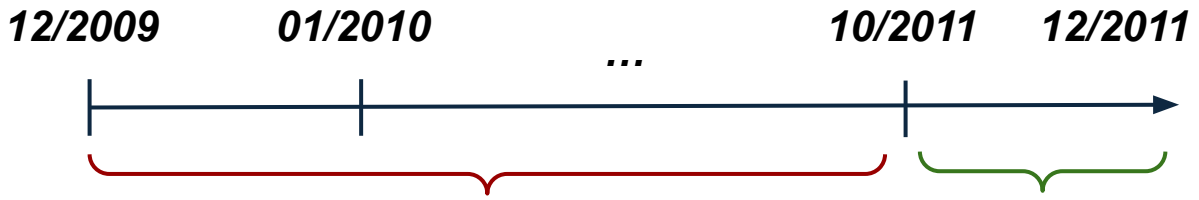
number of records remaining after each cleaning step

I. 1 - Data cleaning - Outliers Detection & Missing values



Missing value rate of each column.

I. 1 - Data cleaning - Conclusion



Training dataset

Testing dataset

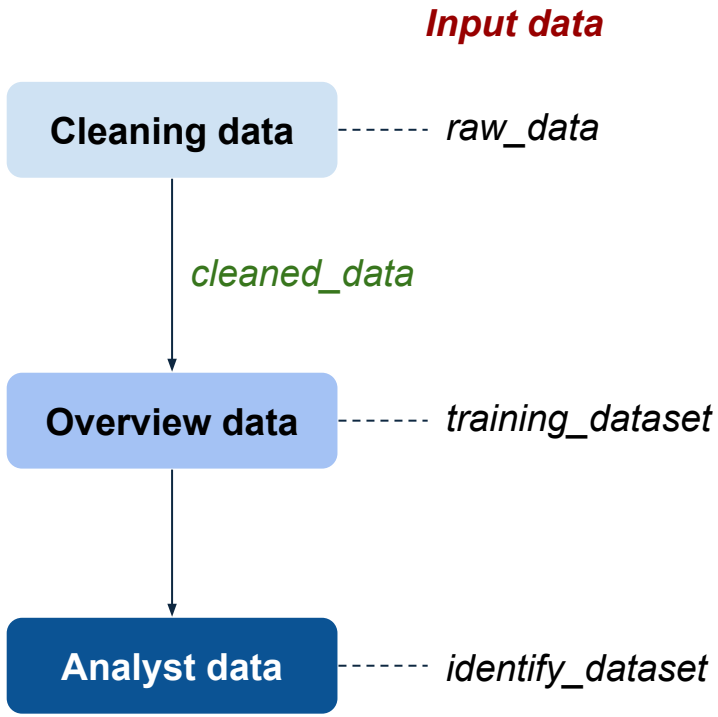
Training dataset: is utilized for analysis and extracting valuable insights.

Total rows: **849,542** rows.

Test dataset: is used to evaluate the model's performance and its ability to generalize to unseen data.

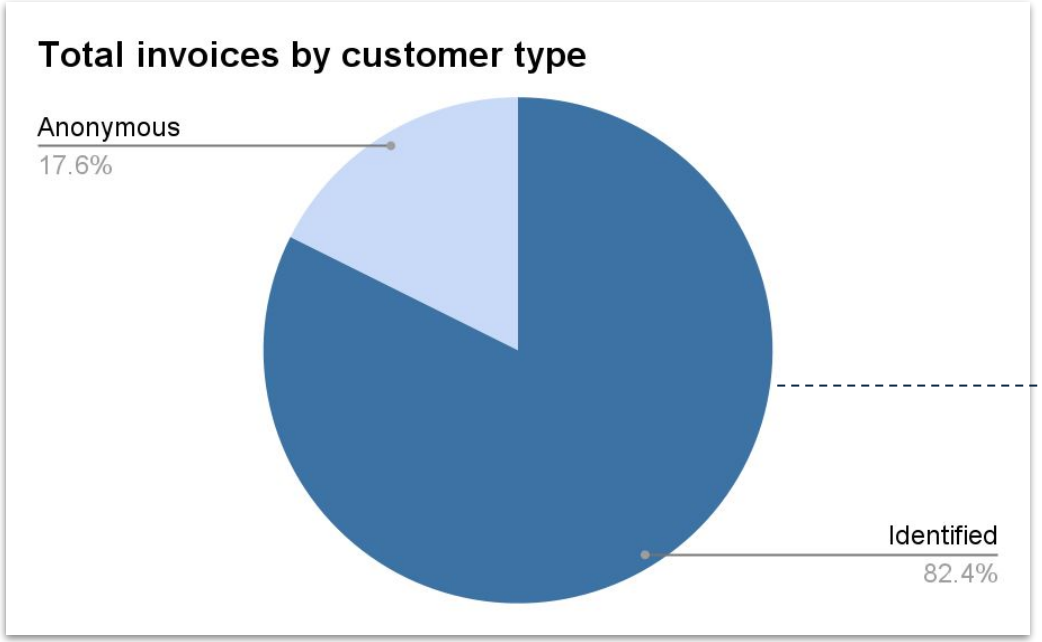
Total rows: **166,328** rows

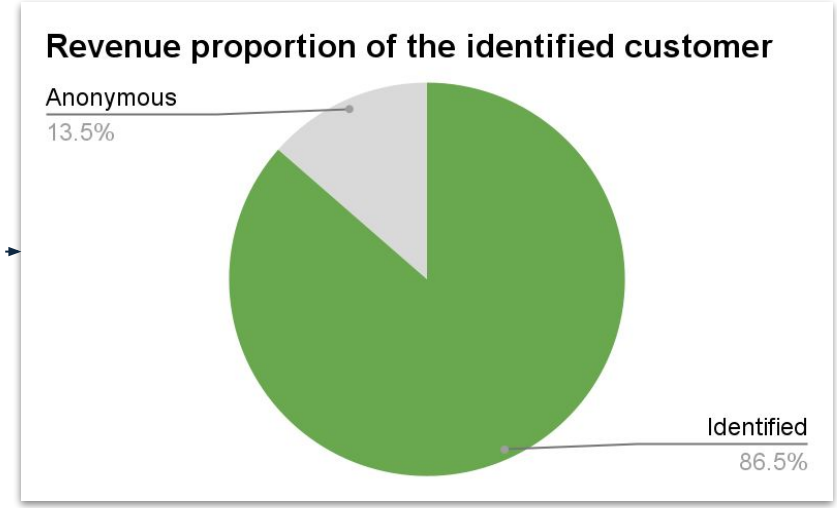
The following analyses will be conducted using only the **training dataset.**



I. 2 - Overview data - Customers

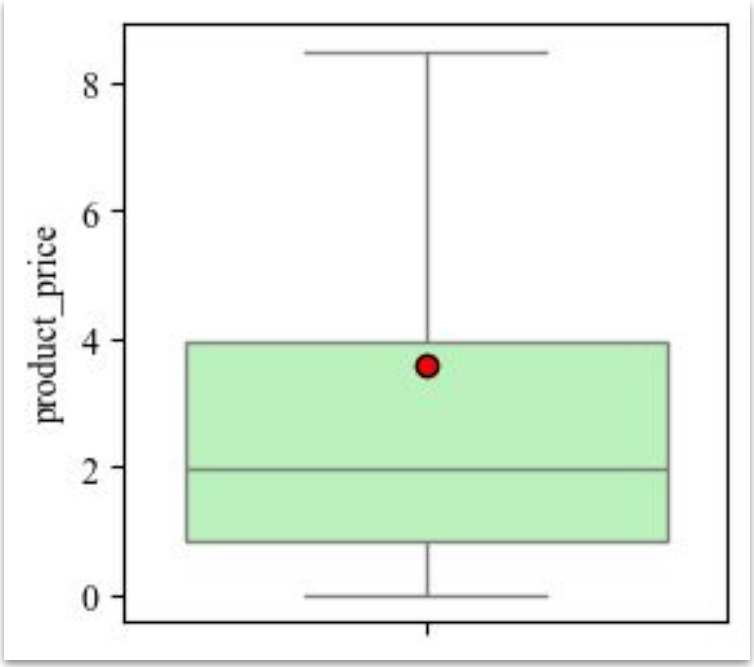
👥 Total customers: **5,402** 📦 Total products: **5,133** 🌍 Total country: **43** 🧾 Total invoices: **43,969**
(including 12% return invoices)





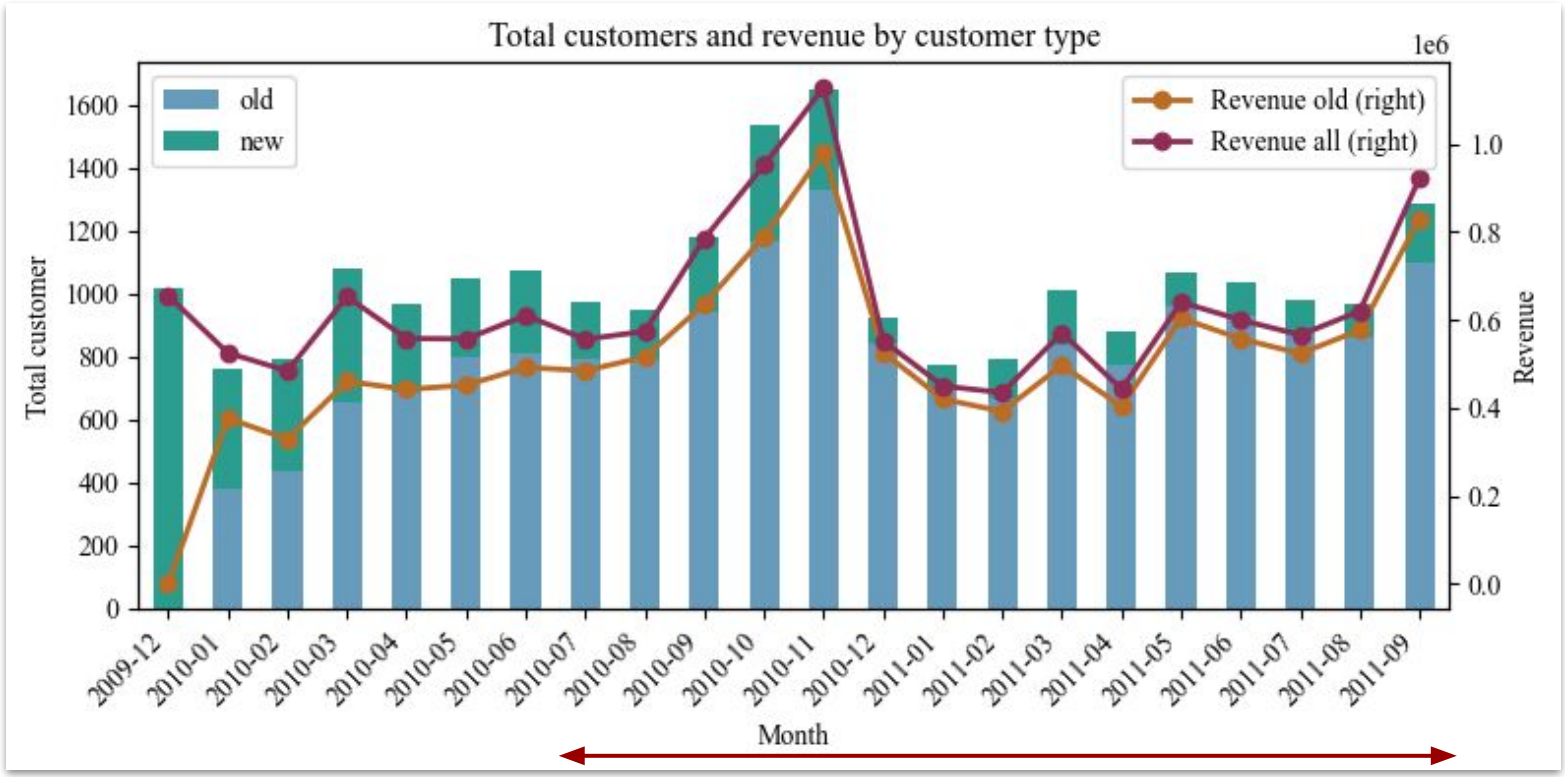
- **Identified customers account for 82.4% of the data and contribute 86.5% of the total revenue.** Therefore, the following analysis will focus exclusively on this group.
- **From section 3**, anonymous customers will not be included in the analysis to avoid potential information noise**.**

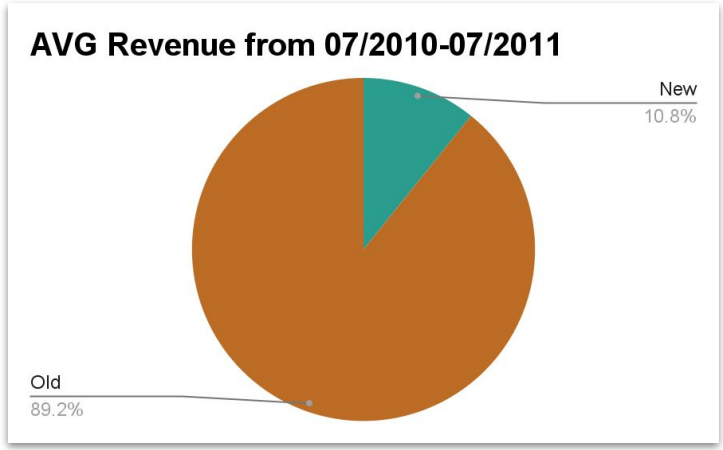
I. 2 - Overview data - Price per product (USD)



Based on the image, we observe that **the mean is higher than the median**, which typically indicates the presence of a few products with very high prices, pulling the mean price above the more common price level of the majority of products in the dataset.

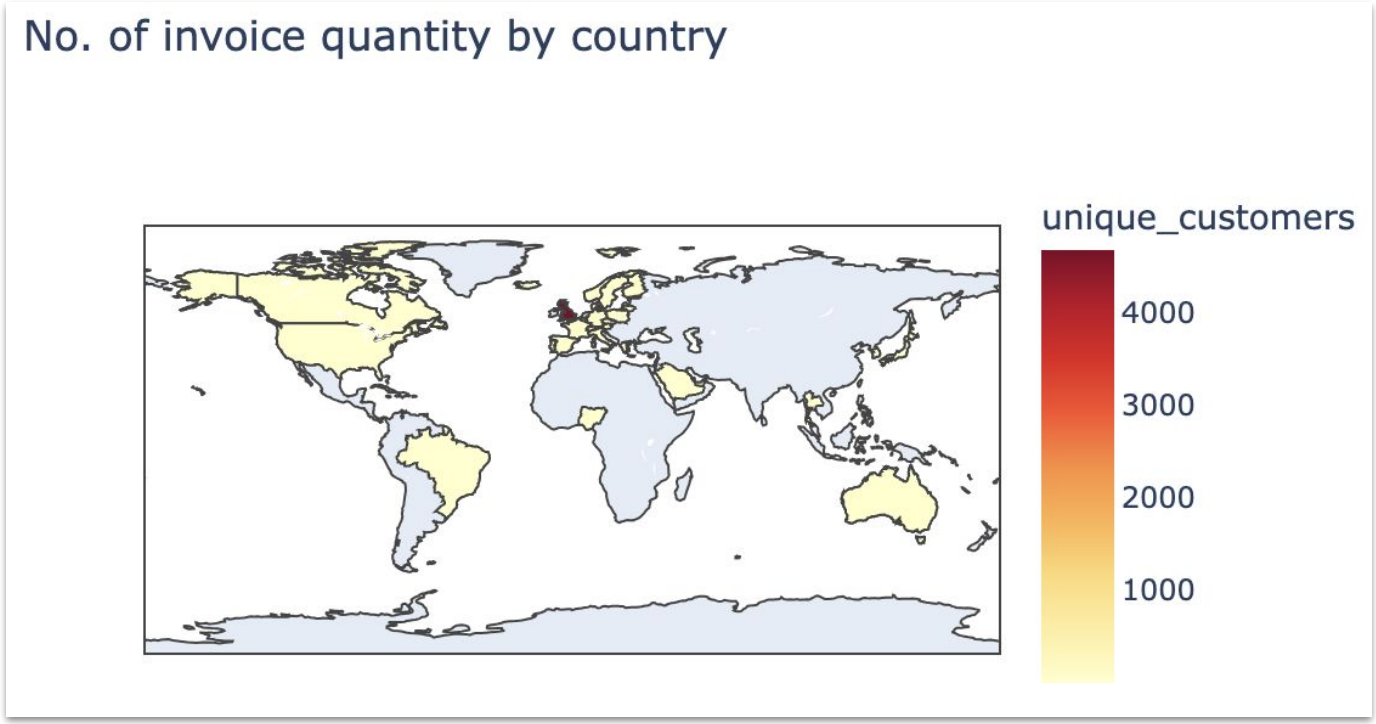
I. 2 - Overview data - Customers

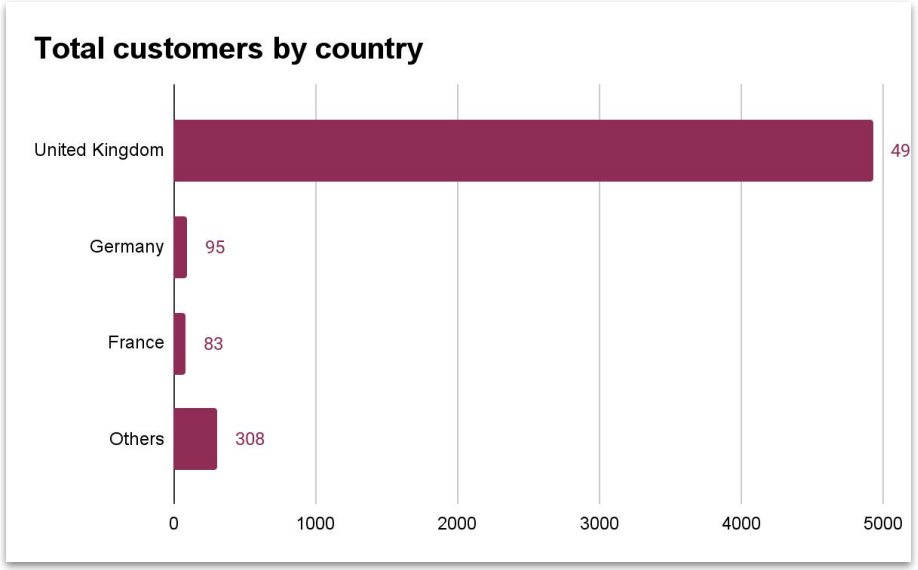




- 89% of revenue comes from existing customers; therefore, it's essential to maintain strong relationships and increase customer lifetime value.
- New customers contribute only 11% of the revenue, indicating the need to strengthen marketing efforts and expand the potential customer base to ensure sustainable growth.

I. 2 - Overview data - Country

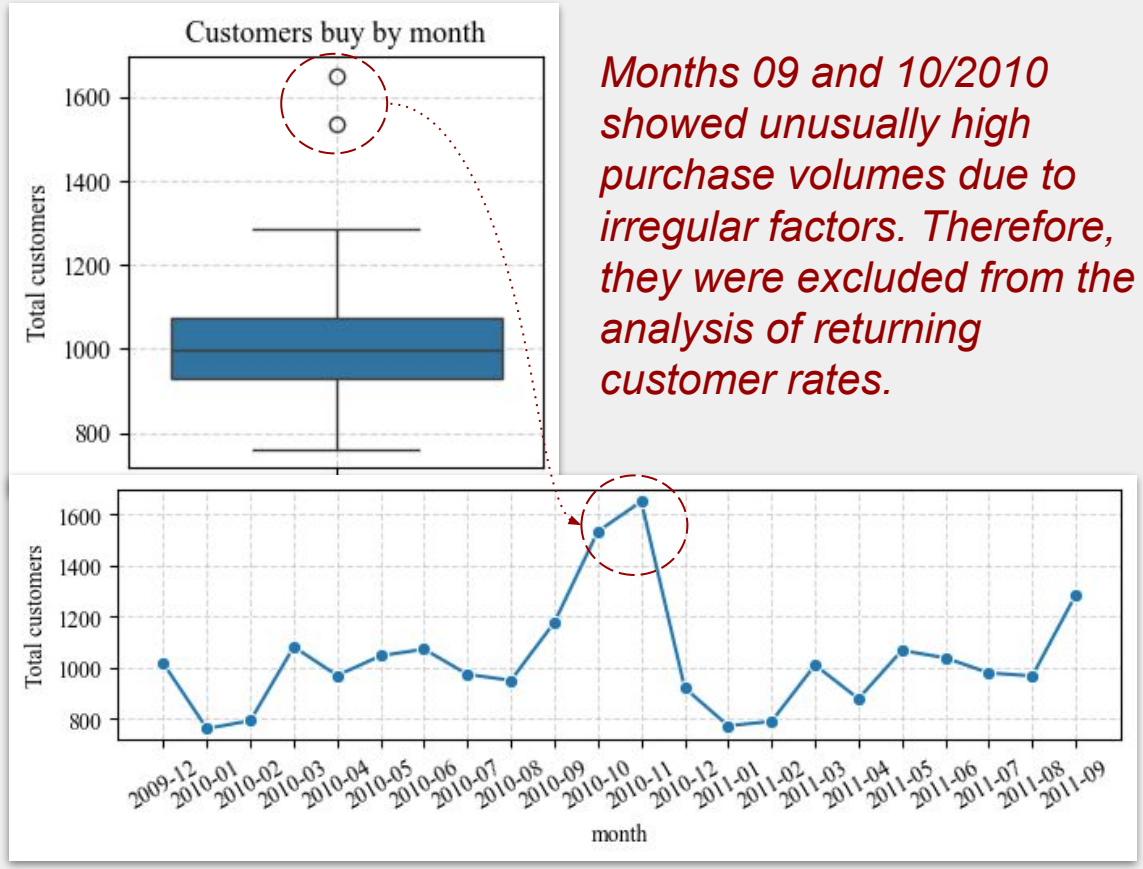




The majority of orders come from the United Kingdom, while other countries contribute only a small proportion.

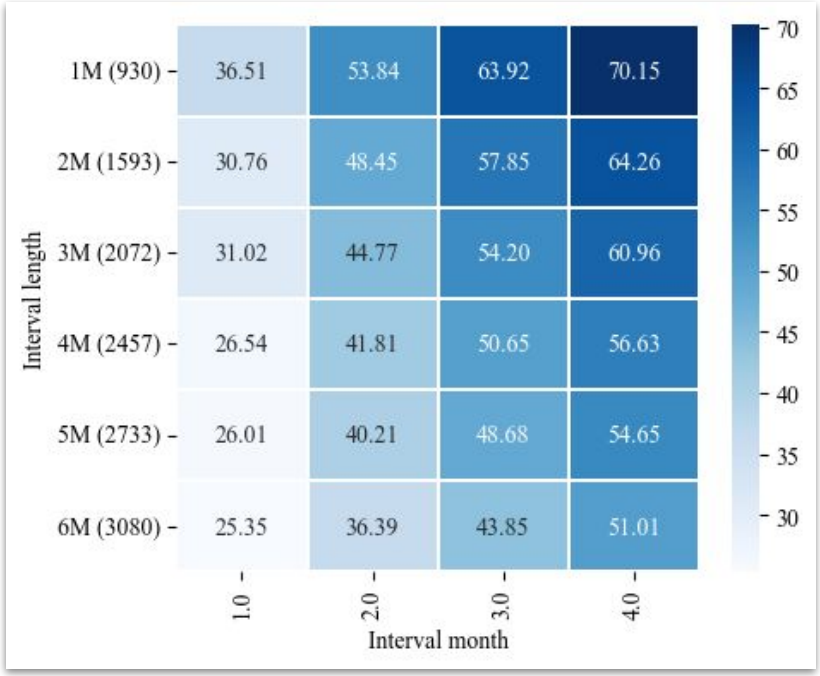
I. 2 - Overview data - Repurchase

Only consider purchase orders, excluding return orders.



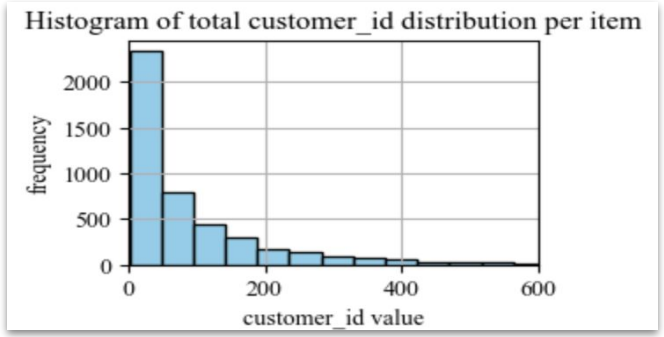
Months 09 and 10/2010 showed unusually high purchase volumes due to irregular factors. Therefore, they were excluded from the analysis of returning customer rates.

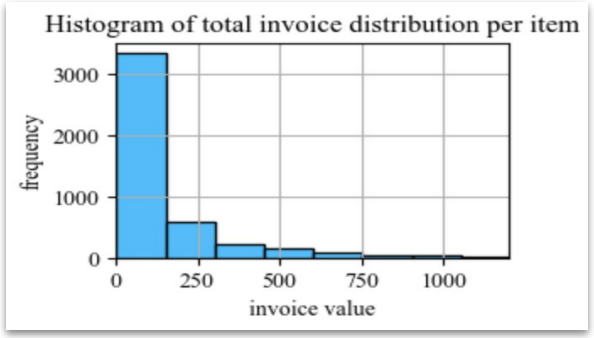
Monthly Cumulative Returning Customer Percent

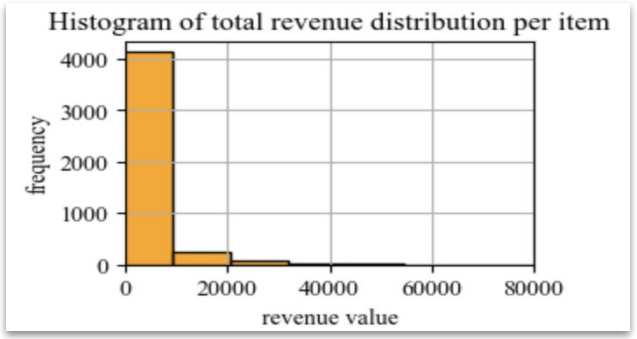


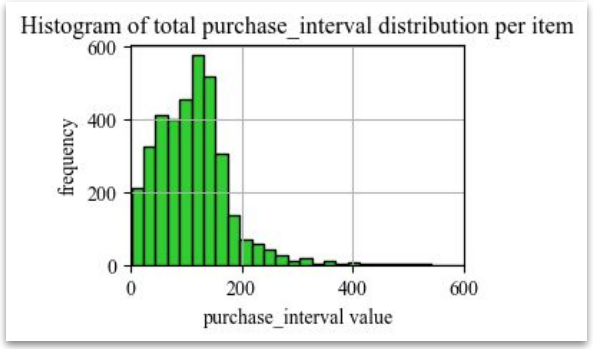
Most customers tend to return within the first month. As the observation period extends, the sample size increases, causing the return rate to decrease. **Therefore, no significant difference is observed when considering longer periods.**

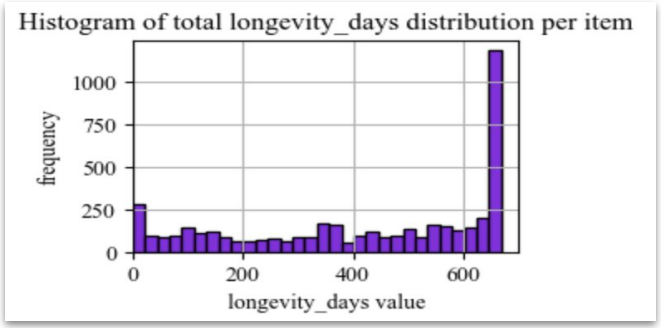
I. 2 - Overview data - Products

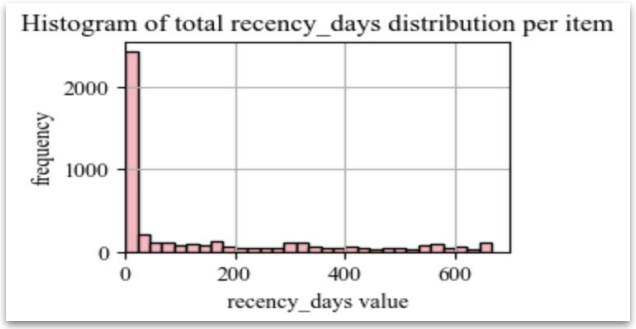












Each product attracts an average of approximately **90 customers**, generating a total of **143 orders**, indicating a certain level of **customer repeat purchases** (as the number of orders exceeds the number of customers). The total revenue reaches **3,068 USD**, corresponding to an **average spending of about 21.5 USA per order** (3,068 / 143 ≈ 21.5).

The **average return time for a customer to repurchase the product is 111 days**, which is significantly shorter than the **average product lifespan of 420 days**. This suggests there is potential to further **monetize the existing customer base throughout the product's lifecycle**. However, the **average recency of purchases across all products is 140 days**, which is higher than the 111-day return time. This may indicate a **recent decline in purchase frequency** or that **the product's sales cycle is slowing down compared to expectations**.

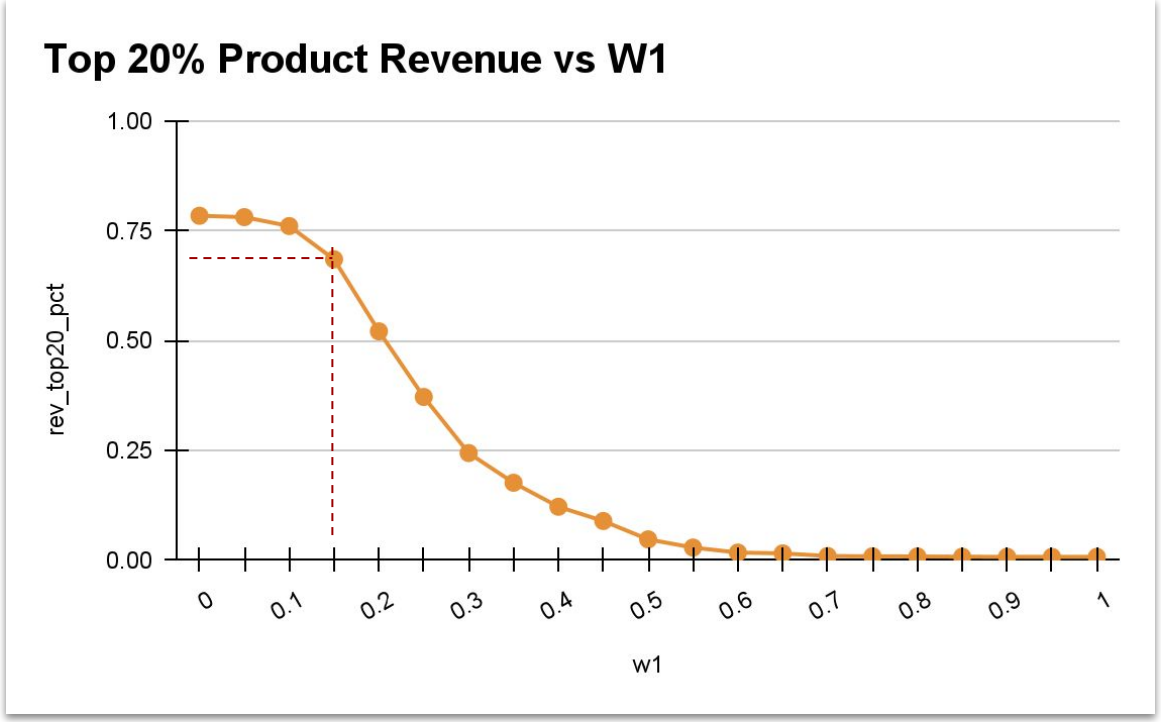
I. 2 - Overview data - Products

We aim to build a scoring system for products based on two key criteria:

1. **Low number of customers** (to prioritize ease of implementation)
2. **High revenue** (to maximize business impact)

product_score = w_1 *(1- **total_customers_norm**) + w_2 * **total_revenue_norm**

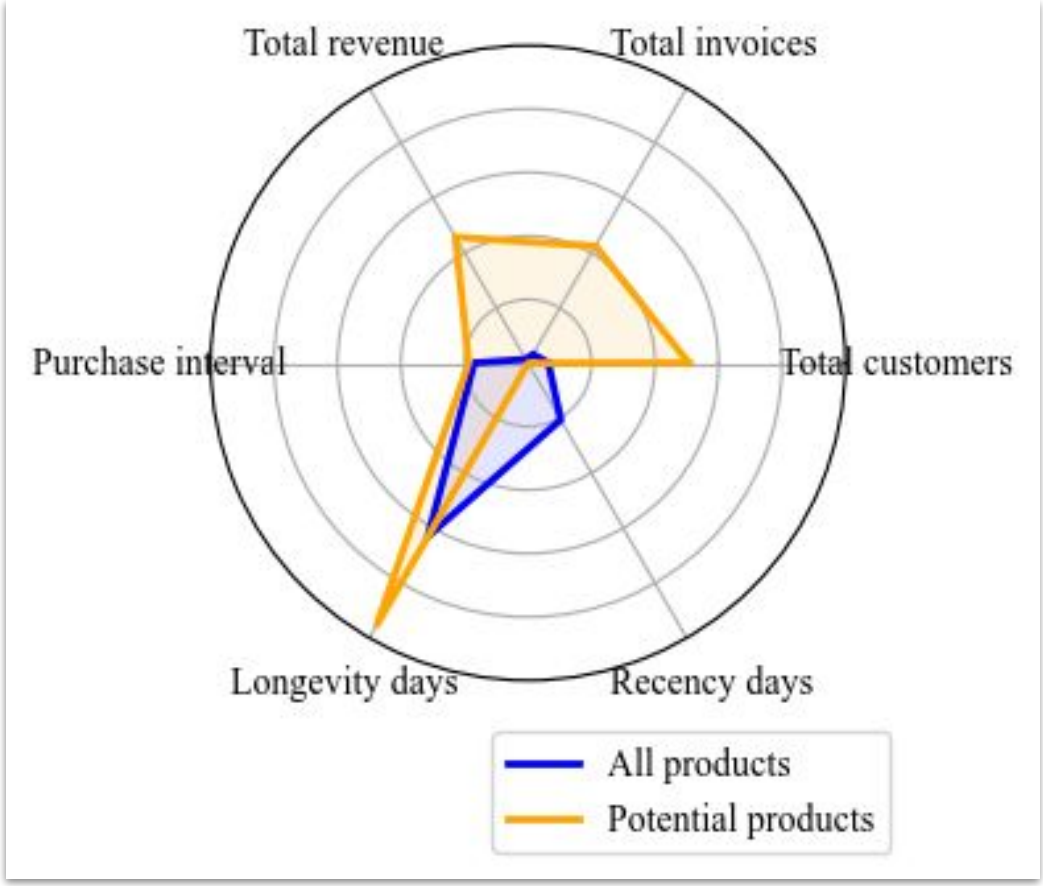
To determine appropriate weights, we apply an **grid search** over combinations of weights (w_1 , w_2).



Based on the chart, we select the weight combinations that yield a revenue share **above 60%** to ensure sufficient revenue contribution. Among these, we choose the combination with the highest value of `w1`, as it helps minimize the number of customers required for deployment while still maintaining strong revenue performance.

I. 2 - Overview data - Potential products

Select the top 10 products with the highest scores, ensuring that their most recent purchase occurred within the last 5 days, as potential products.



Analysis shows that most potential products share three key characteristics: a large customer and invoice base, a long lifespan, and recent purchase activity.

I. 2 - Overview data - Potential products

The average time for customers to repurchase a potential product is 60 days. Therefore, we analyze how many months of customer history should be considered.

Looking at Figure (b), we observe that extending the observation period does not significantly impact the number of long-term customers. As the observation window increases, the sample size grows, leading to a natural decline in the return rate. Hence, the return behavior becomes less distinguishable over longer durations.

Figure (a) Purchase interval of customers to repurchase a potential product

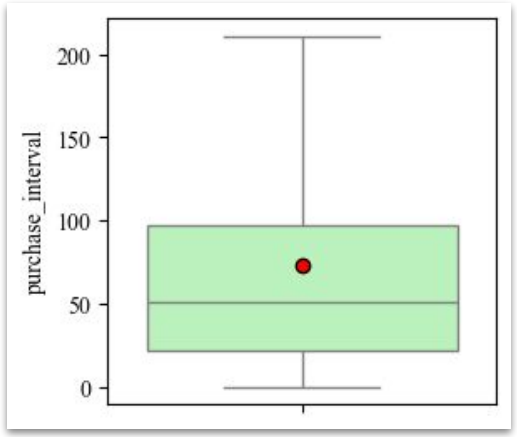
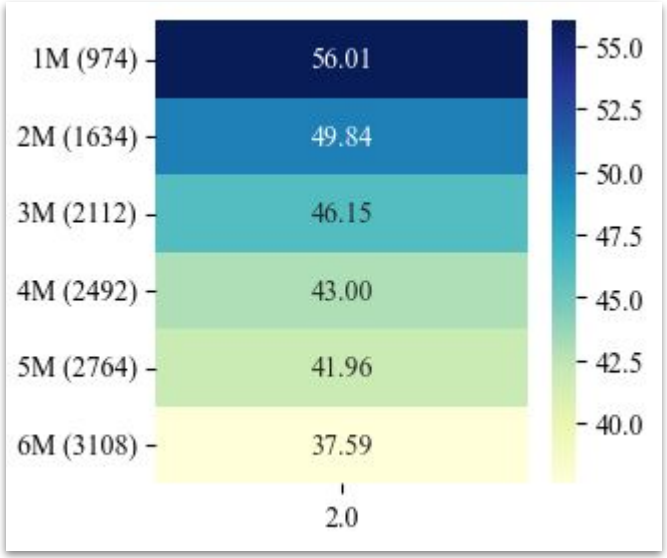


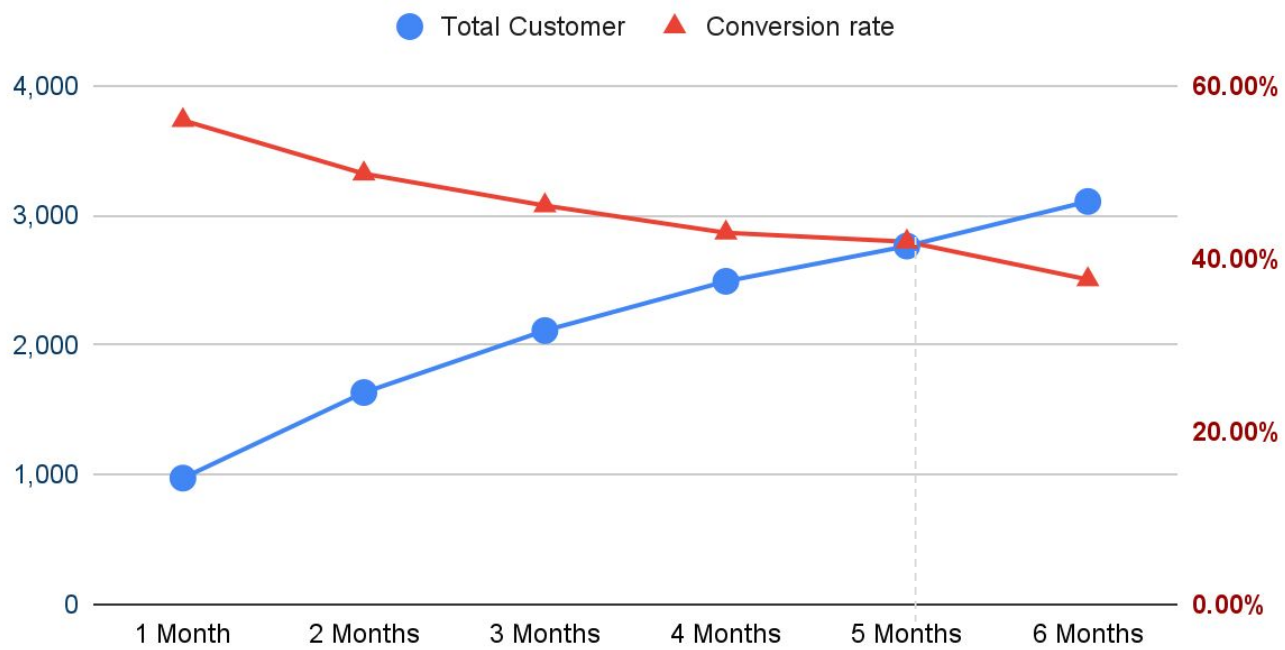
Figure (b) Returning Customer Percent in 2 Months



I. 2 - Overview data - Potential products

To address this, we analyze both **Gross Customer Count** and **Gross Customer Conversion Rate (CRGrossCustomer)**. As shown in Figure (c), we identify the **intersection point** between these two metrics and select it as the optimal cutoff for defining the **historical data** (input) and the **label data** (output) used in our modeling.

Figure (c). Total Customer per interval months and Conversion rate in 2 next months



We selected the **historical data** as the group of customers who purchased potential products over a period of **5 months,** and the **labeled data** corresponds to the following **2 months**.

I. 2 - Overview data - Data modeling

Only **24.2%** of customers returned buy top 10 potential products **(label 1)** in the following two months.

I. 3 - Analysis

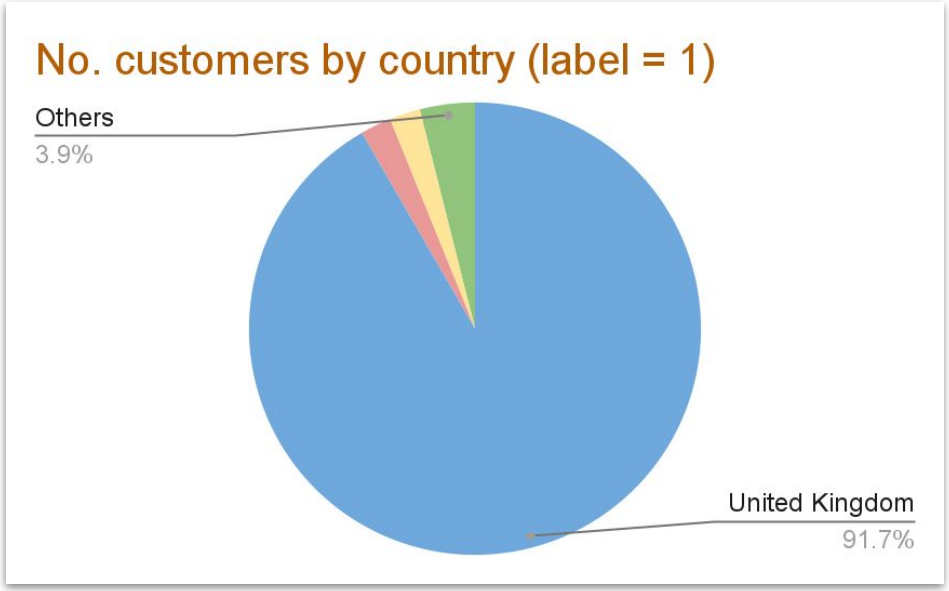
Category	Variable	Sign
I. Country	1.Customer Country	customer_country
II. Invoice	2. Number of Invoices per Customer	num_invoices
	3. Average Monthly Invoices per Customer	avg_monthly_invoices
	4. Average Quantity per invoice per customer	avg_invoice_quantity
	5. Average Product Types per Invoice per Customer	avg_invoice_product_types
	6. Average Revenue per Invoice per Customer	avg_invoice_revenue
III. Purchase Interval	7. Average Purchase Interval per Customer	avg_purchase_interval
IV. Product	8. Average Product Price per Customer	avg_price
	9. Number of Product Types per Customer	num_product_types
V. Revenue	10. Revenue per Customer	revenue

KH đã từng mua bao nhiêu sản phẩm ở top 10?
Mua top 10 cách đây bao lâu?

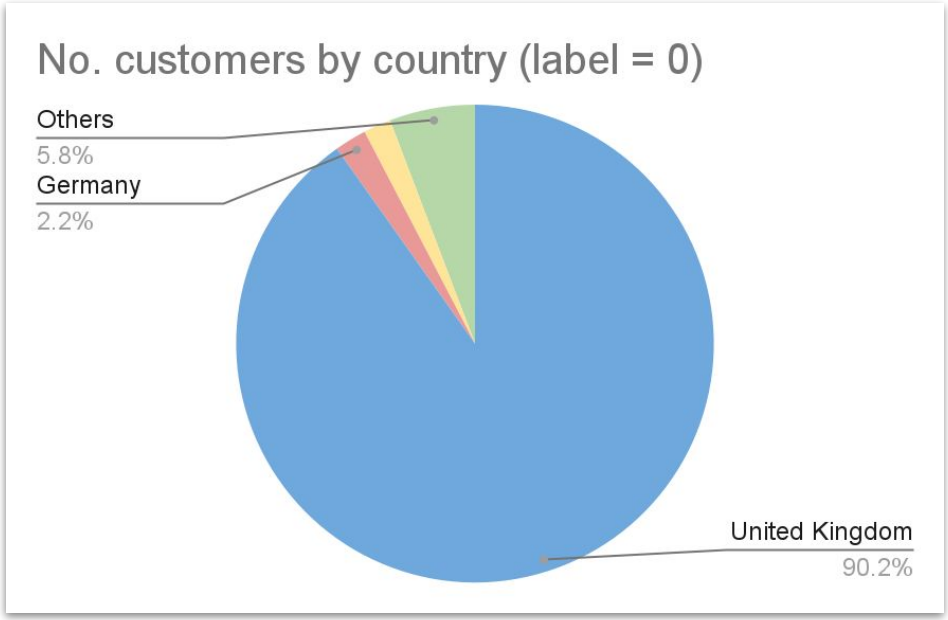
I. 3 - Analysis - Univariate Analysis

1. Customer Country

Purchase



No Purchase



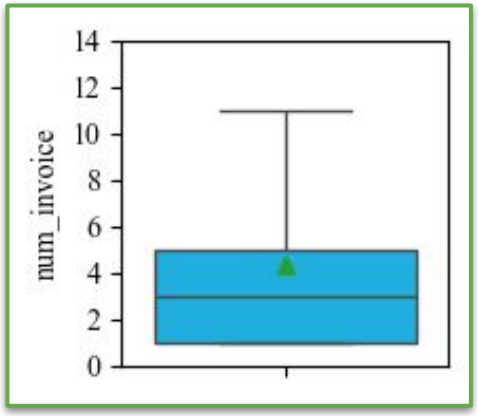
There is **no difference** in country between the two groups: **returning and non-returning customers.**

I. 3 - Analysis - Univariate Analysis

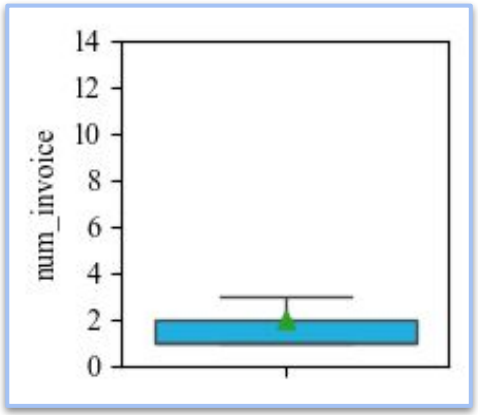
2. Number of Invoices per Customer

Non-returning customers typically placed around **2 orders over the 5-month period**, whereas **returning customers had an average of 5 orders** during the same timeframe.

Purchase

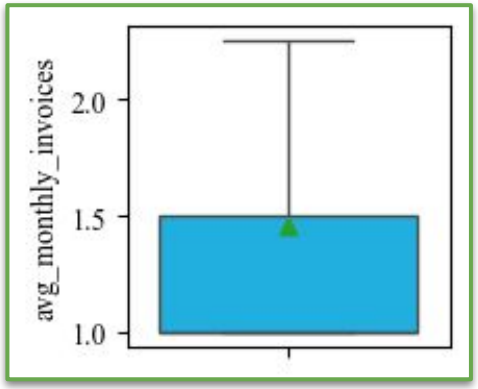


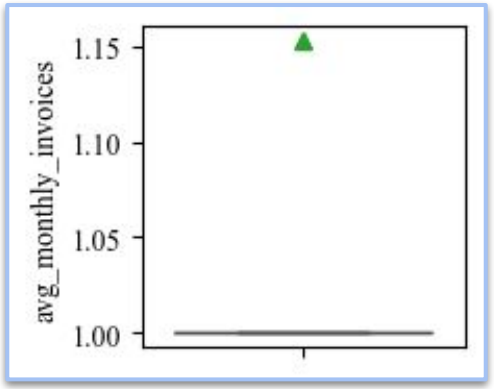
No Purchase



3. Average Monthly Invoices per Customer

Non-returning customers typically placed around **1 orders over per month**, whereas **returning customers had an average of 2 orders** during the same timeframe.





I. 3 - Analysis - Univariate Analysis

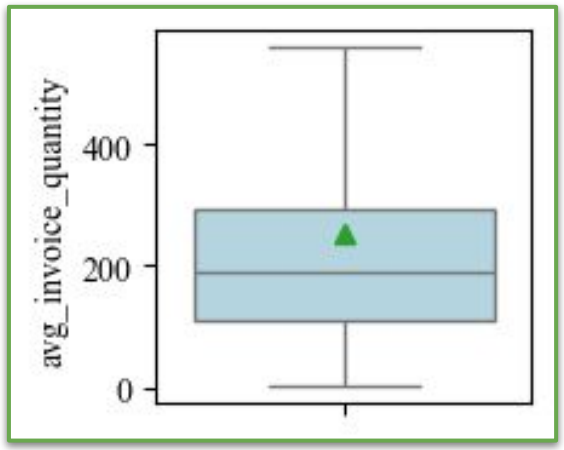
4. Average Quantity per invoice per customer

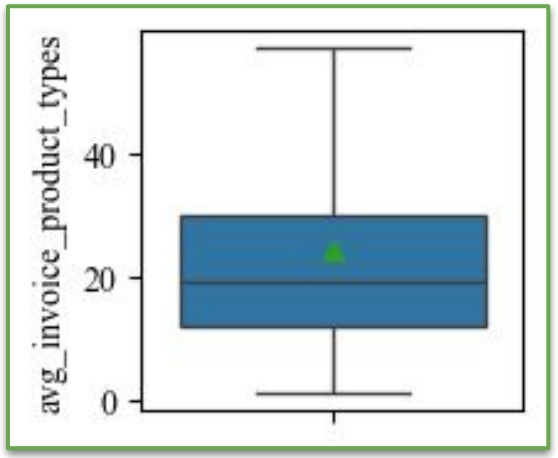
Over a 5-month period, **the total quantity of products purchased** by customers who did **not return** was typically around **200**, while **returning customers** tended to purchase around **250**.

5. Average Product Types per Invoice per Customer

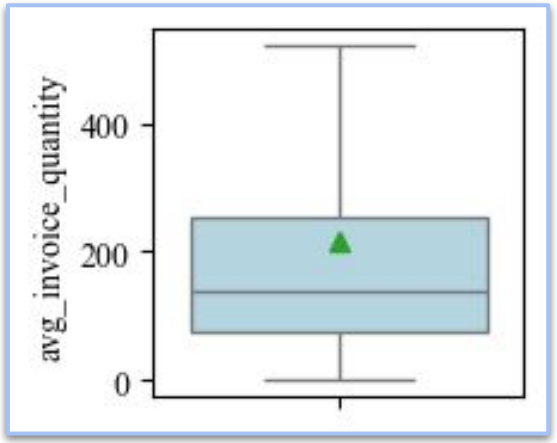
There is **no significant** difference in the average number of product types per invoice per customer between the two groups.

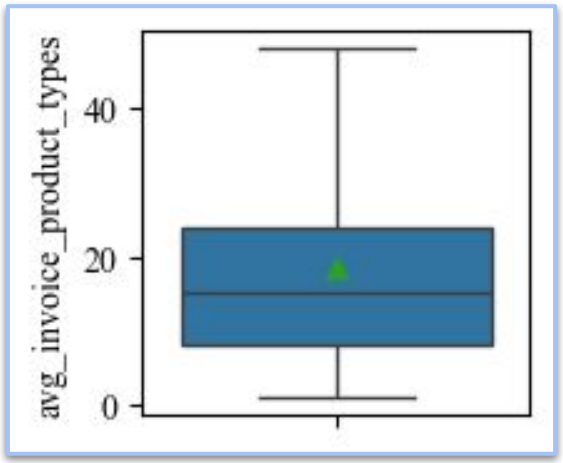
Purchase





No Purchase

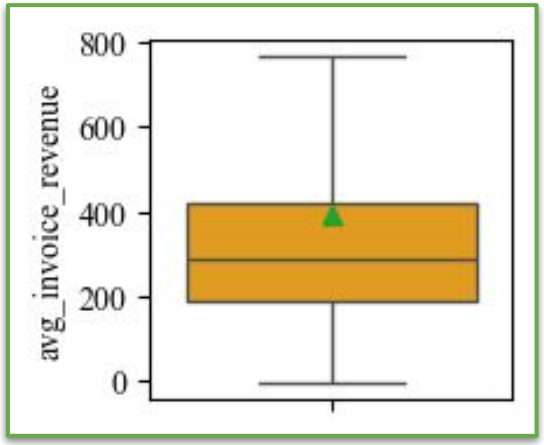




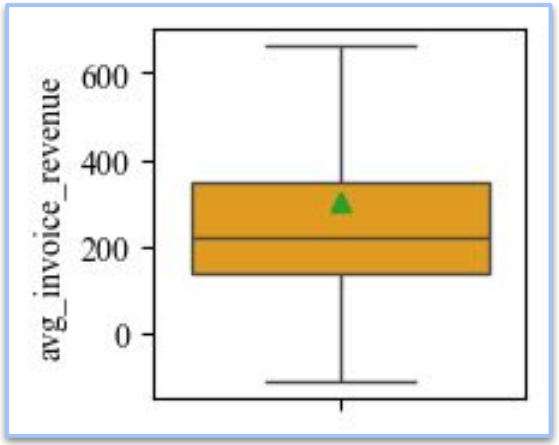
I. 3 - Analysis - Univariate Analysis

6. Average Revenue per Invoice per Customer

Purchase



No Purchase

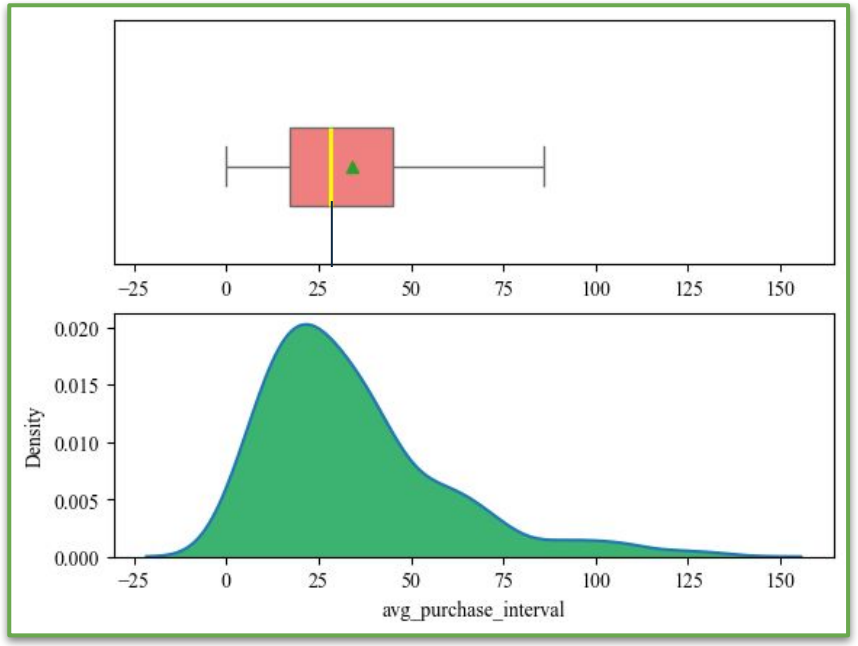


Non-returning customers brought in an average revenue of **344 (USD)** over per invoice, compared to **410 (USD)** from **returning customers.**

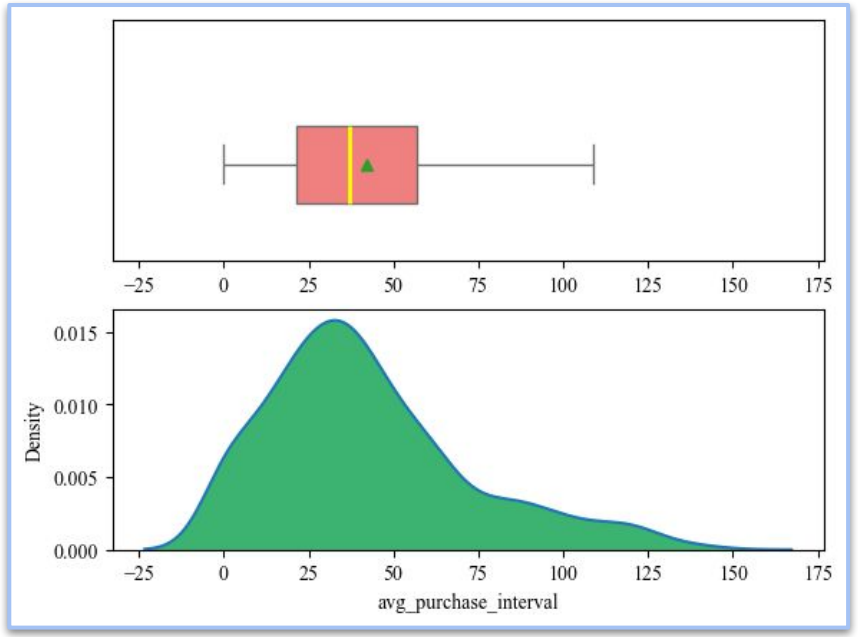
I. 3 - Analysis - Univariate Analysis

7. Average Purchase Interval per Customer

Purchase



No Purchase



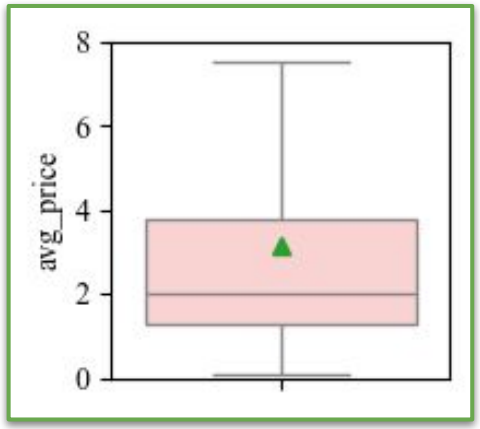
Non-returning customers had an average repurchase interval of **42 days**, whereas **returning customers** repurchased after an average of **33 days**.

I. 3 - Analysis - Univariate Analysis

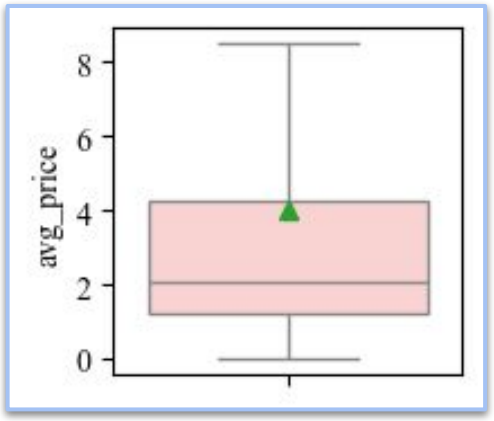
8. Average Product Price per Customer

A portion of customers **who do not return** tend to make **high-value purchases**. As a result, there is a significant gap between the mean and median of this customer group.

Purchase

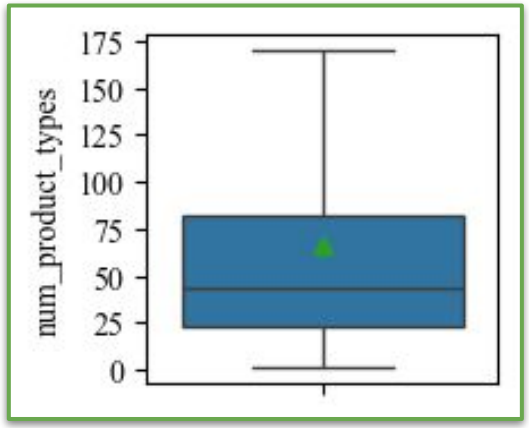


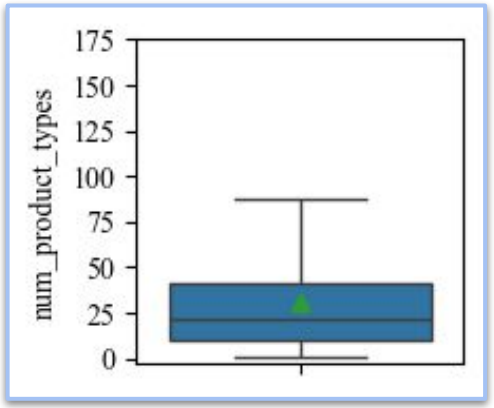
No Purchase



9. Number of Product Types per Customer

Customers who do not return tend to purchase **a less diverse range of products** compared to returning customers.

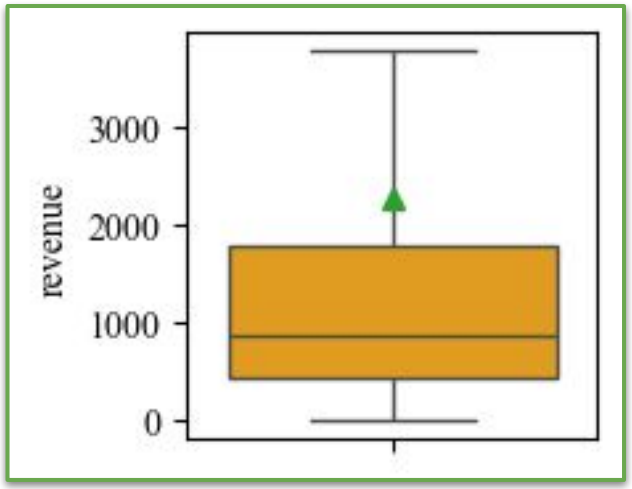




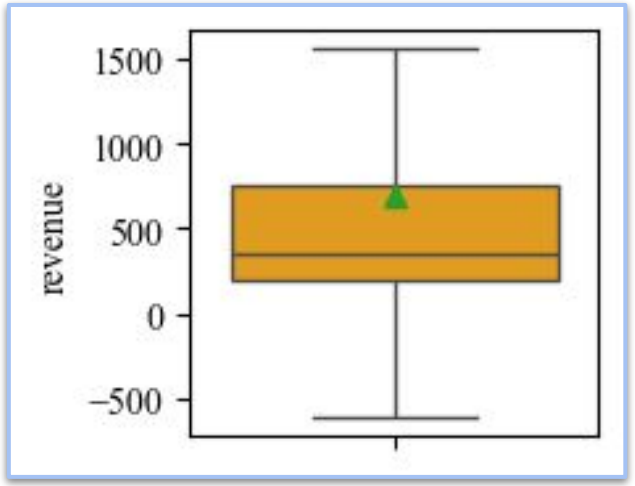
I. 3 - Analysis - Univariate Analysis

10. Revenue per Customer

Purchase



No Purchase



Non-returning customers brought in an average revenue of **699 (USD)** over 5 months, compared to **2251 (USD)** from **returning customers.**

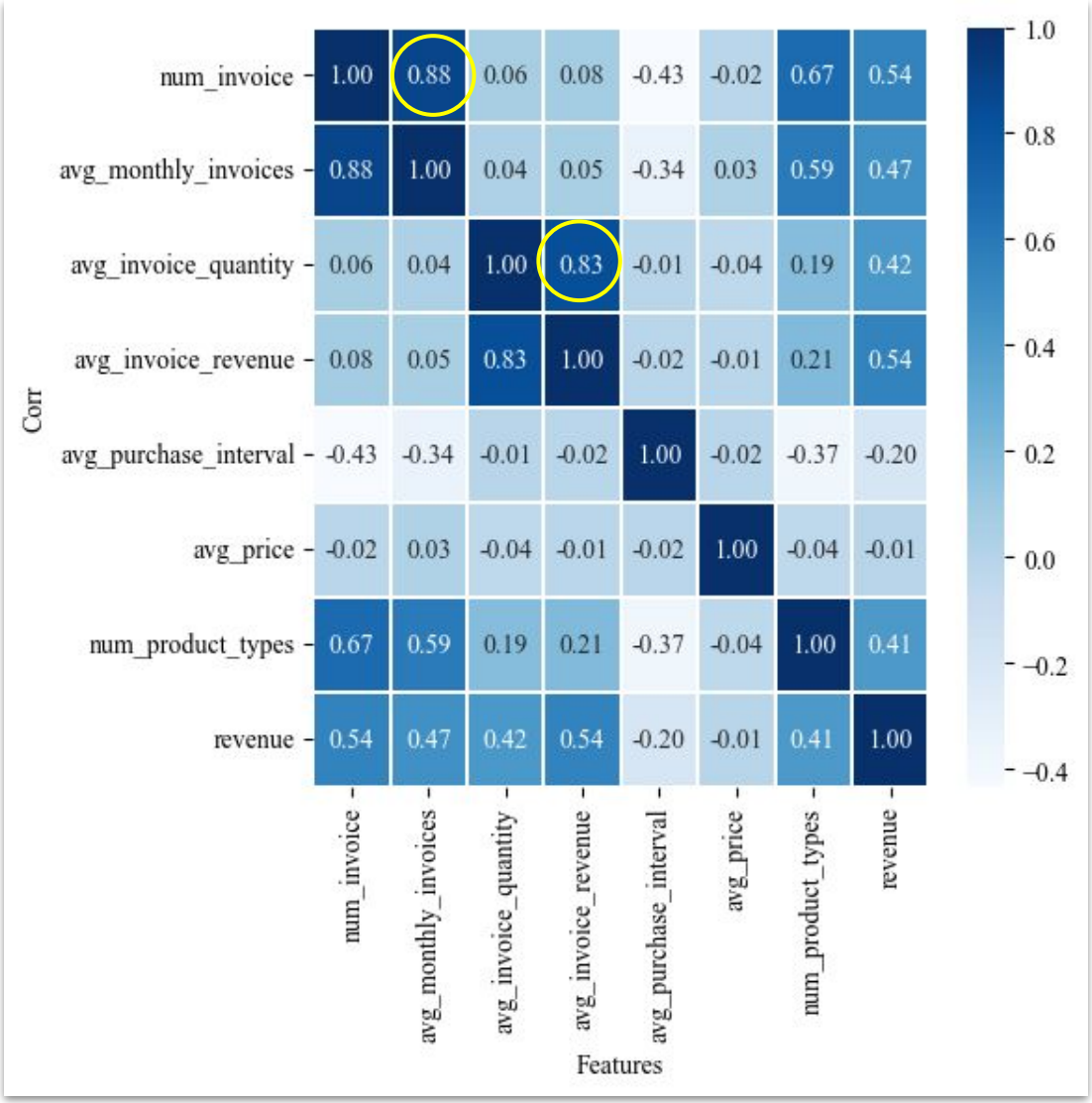
I. 3 - Analysis - Univariate Analysis - Conclusion

	Non-return	Return
1. customer_country	no difference	
2. num_invoices	Low	High
3. avg_monthly_invoices	Low	High
4. avg_invoice_quantity	Low	High
5. avg_invoice_product_types	no difference	
6. avg_invoice_revenue	Low	High
7. avg_purchase_interval	Long	Short
8. avg_price	High (1 small group)	Low
9. num_product_types	Low	High
10. revenue	Low	High

To encourage non-returning customers to engage more with the brand, strategies can be implemented to promote a wider range of products.

*Features that show **no difference** between the two groups will be excluded and not used in the multivariate analysis.*

I. 4 - Analysis - Multivariate Analysis



Variables with **corr > 0.8** are:

- avg_monthly_invoices (remove) and num_invoices,
- avg_invoice_quantity (remove) and avg_invoice_revenue.

I. 4 - Analysis - Multivariate Analysis

Hypothesis Testing (LLR p-value**):**

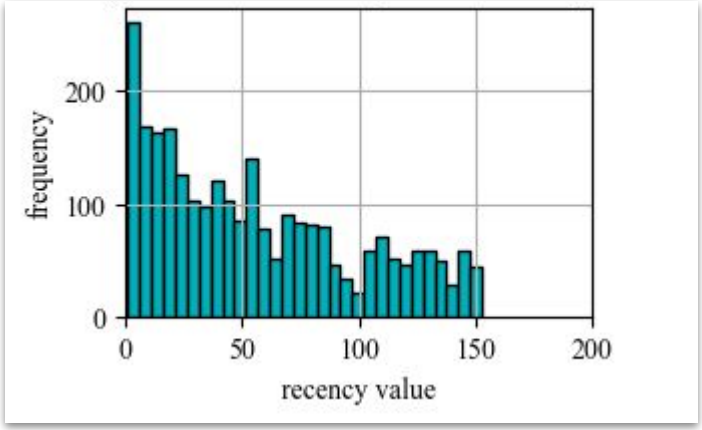
- **H_0 (null hypothesis):** The model is not better than the null model (which only includes the intercept).
- **H_1 (alternative hypothesis):** The model has at least one independent variable that helps explain the dependent variable better.

num_invoices	1
avg_invoice_revenue	2
avg_purchase_interval	3
avg_price	4
num_product_types	5
revenue	6

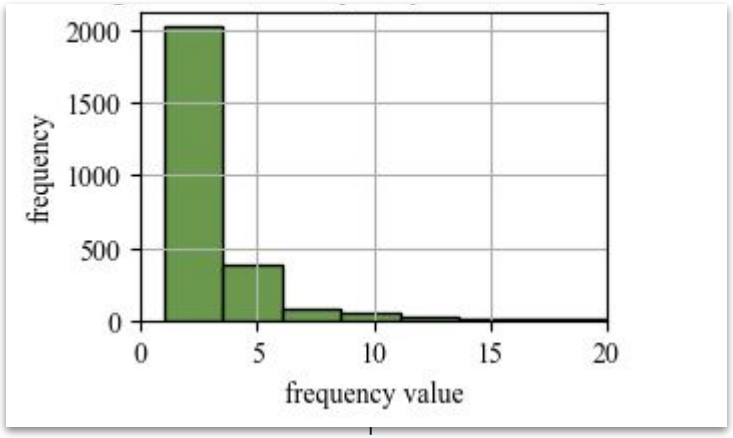
p-values	**1**	**2**	**3**	**4**	**5**	**6**
1						
2	0.00					
3	0.00	0.00				
4	0.00	0.00	0.00			
5	0.00	0.00	0.00	0.00		
6	0.00	0.00	0.00	0.00	0.00	

=> The combination of variables is statistically significant.

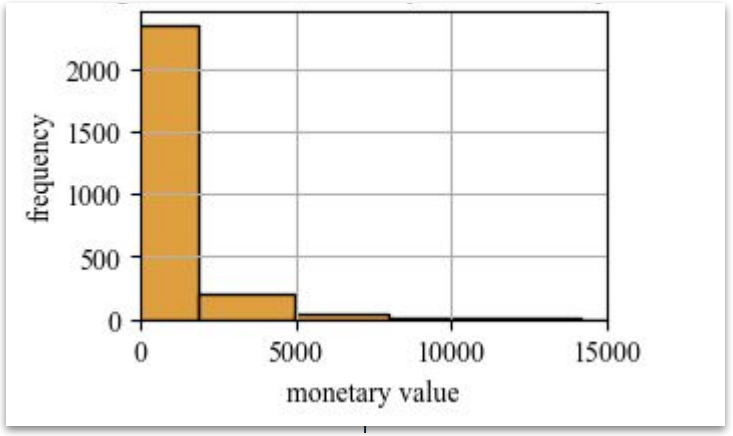
I. 5 - RFM Segments



- (0.999, 14.0]: **5**
- (14.0, 34.0] : **4**
- (34.0, 59.0] : **3**
- (59.0, 102.0] : **2**
- (102.0, 153.0]: **1**

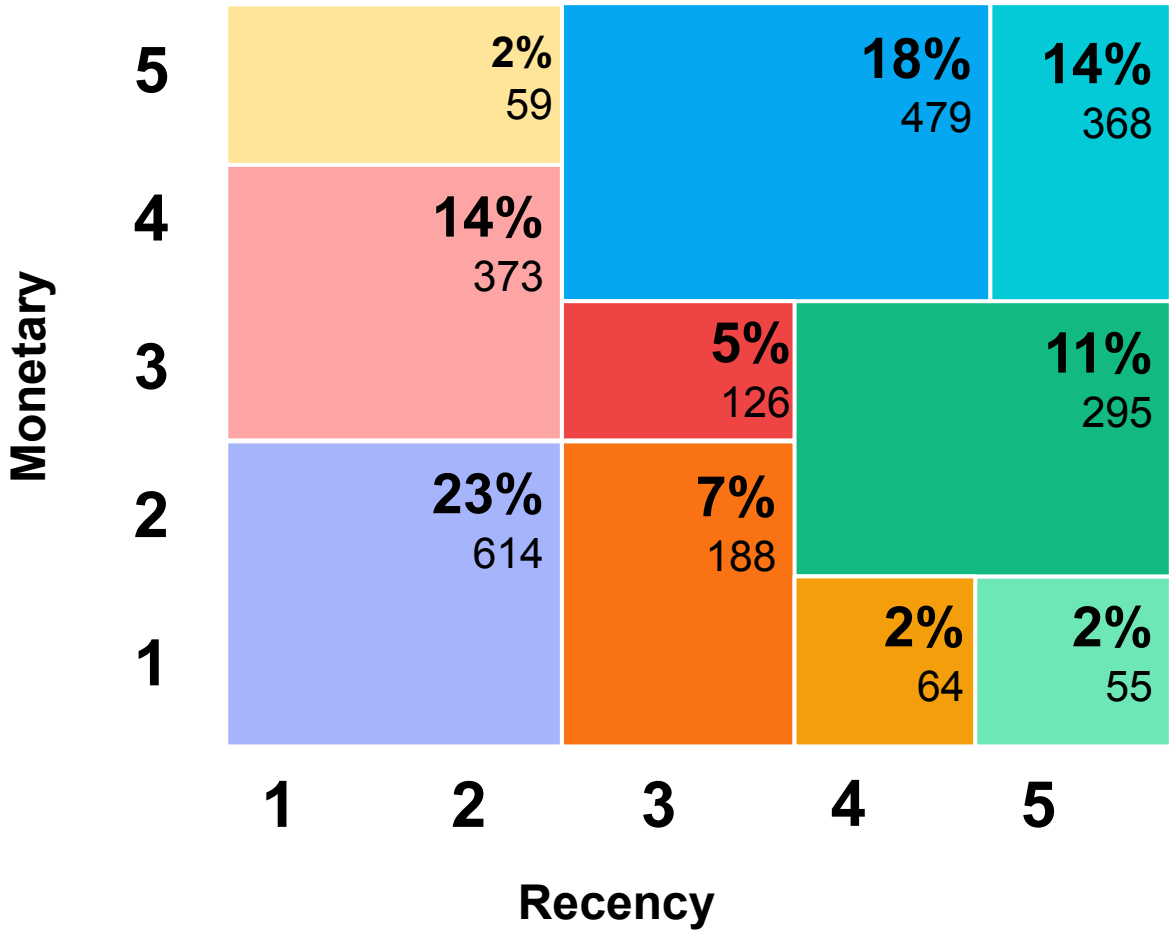


It can only be divided into two groups because the data is quite similar, with n**o significant differences.** Therefore, this variable will not be used in the RFM scoring.



- (-4287.631, 189.47]: **1**
- (189.47, 335.01]: **2**
- (335.01, 597.07]: **3**
- (597.07, 1187.16]: **4**
- (1187.16, 87962.41]: **5**

I. 5 - RM Segments



2,621
customers

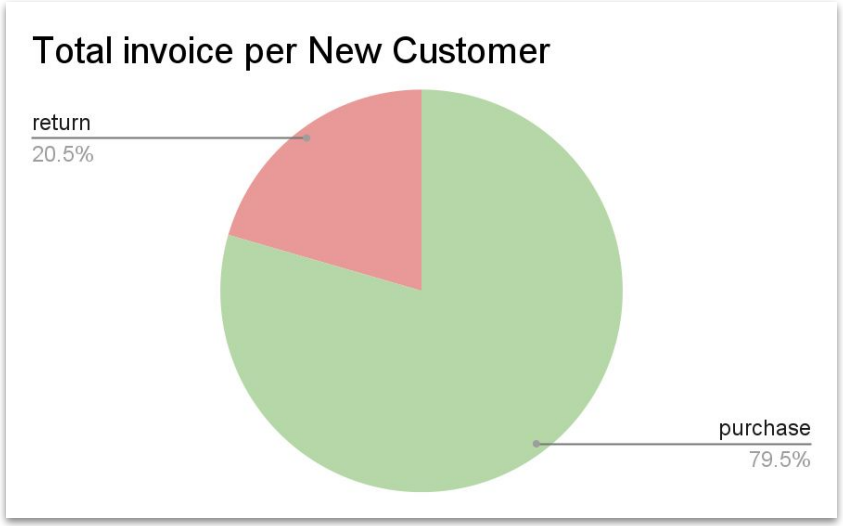
I. 5 - RM Segments

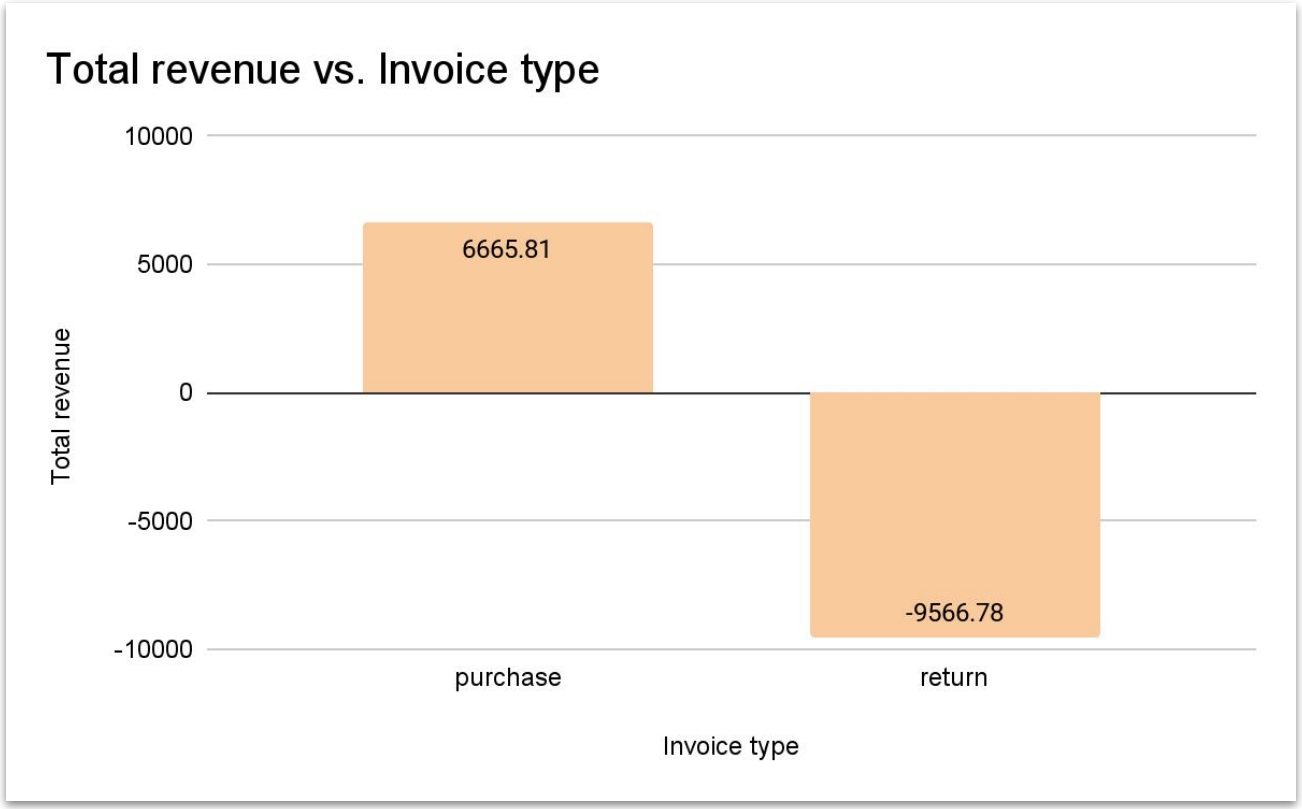
Historical data (5 Months) → *Labeled data (2 Months)*

Customer Segment	Total customers	AVG Revenue per customer (USD)	Percent customer	The non-purchase rate in the next 2 months	The non-purchase rate of the top 10 in the next 2 months	
Champions	368	43.01	14.04%	72.28%	46.20%	Take care to boost revenue (upsell).
Loyal Customer	479	34.08	18.28%	67.22%	39.04%	
Potential Loyalist	295	21.36	11.26%	41.69%	17.29%	
New Customers	55	-123.16	2.10%	32.73%	3.64%	Check behavior of customers.
Promising	64	15.35	2.44%	26.56%	6.25%	
Needs Attention	188	19.31	7.18%	28.19%	10.64%	
About to Sleep	126	27.22	4.8%	53.17%	26.98%	
Can't Lose Them	59	168.69	2.25%	47.46%	27.12%	Take care to boost return (voucher, promotion).
At Risk	373	36.90	14.23%	41.55%	21.98%	
Hibernating	614	21.63	23.42%	26.38%	11.24%	

I. 5 - RM Segments

New Customers Revenue < 0

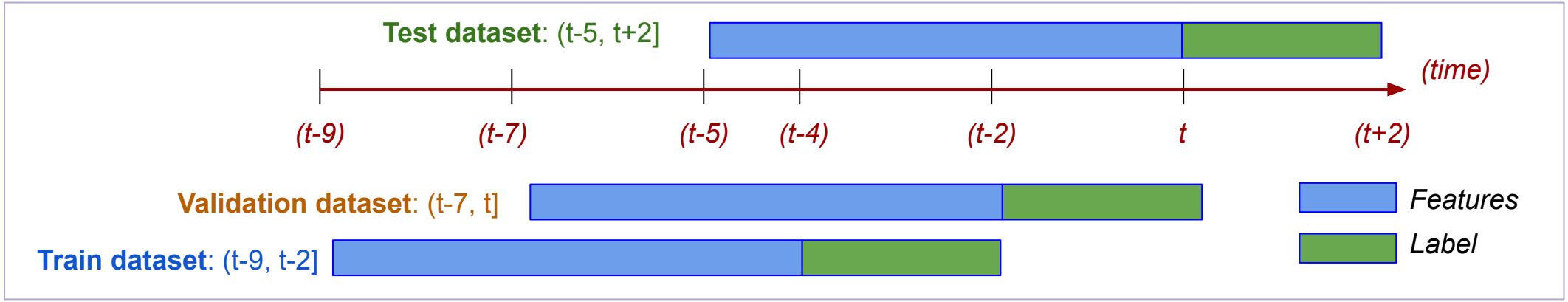




Due to the high value of returned orders, the total revenue became negative.

II. 1 - Feature Engineering

1.1. Data Splitting:



1.2. Feature engineering: Some features extracted from EDA were used for feature engineering in the model, such as:

Index	Feature name	Type
1	num_invoices	Analysis
2	avg_invoice_revenue	Analysis
3	avg_purchase_interval	Analysis
4	avg_price	Analysis
5	num_product_types	Analysis
6	revenue	Analysis

II. 2 - Setup & Training & Tuning

2.1. Set up hyperparameters for XGBoost:

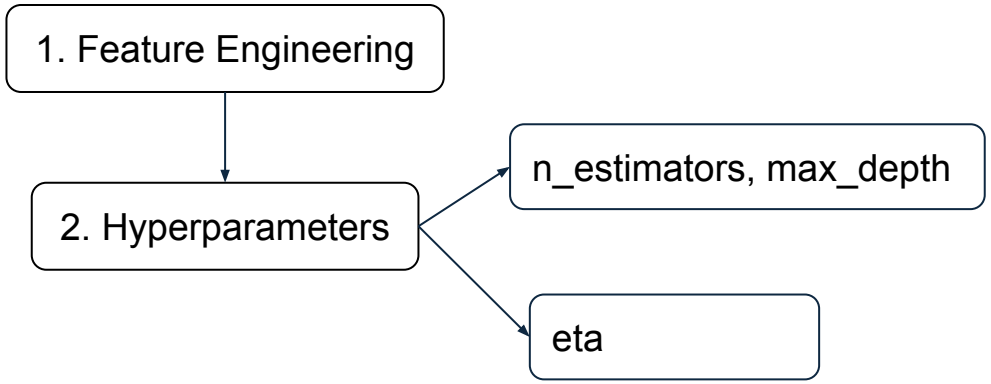
Hyperparameters	Description	Values
eta	learning rate	0.1
max_depth	max depth of a tree	6
subsample	sample ratio of training data	0.8
colsample_bytree	sample ratio of features	0.8
alpha	L1 regularization term	default
n_estimators	number of estimators for boosting	40
scale_pos_weight	weight value of labels	default

These parameters will be optimized after the feature engineering process has been finalized.

II. 2 - Setup & Training & Tuning

2.2. Model training and tuning workflow:

First, we trained the model and performed feature engineering to evaluate the model's precision. We chose **precision** as the evaluation metric. Next, after fixing the feature engineering process, we optimized the hyperparameters, specifically: n_estimators, max_depth, and eta.



II. 2 - Setup & Training & Tuning

2.2. Model training and tuning workflow:

Feature Engineering

We will incorporate the feature engineering components shown in the table below:

Index	Feature name	Add Type
7	top10_products_purchase_count	Label features
8	recency_top10_purchase_days	Label features
9	recency_days	RFM features
10	purchase_frequency	RFM features
11	revenue	RFM features
12	total_invoices_autumn_past	Past features
13	total_revenue_past	Past features
14	revenue_trend	Trend features

We will use precision to evaluate whether these additional features are beneficial.

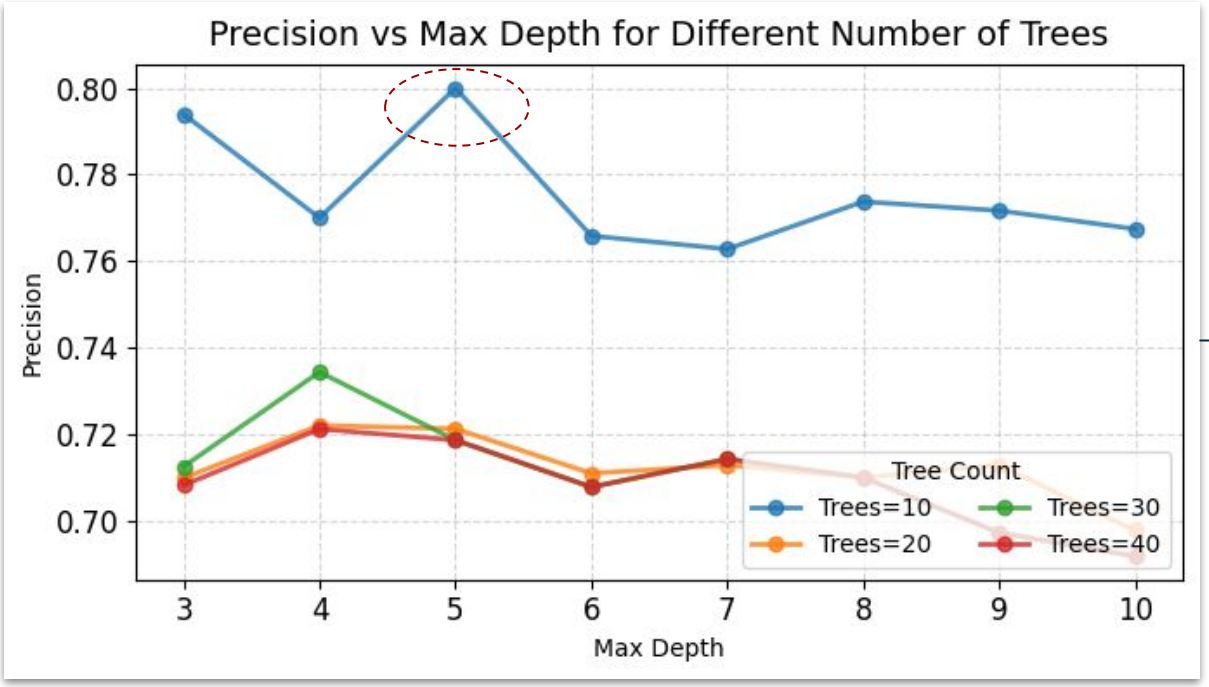
Feature type	Precision (validation dataset)	Note
Analysis features	64%	
+ Label features	69%	Added feature engineering
+ RFM features	69%	Not added feature engineering
+ Past features	71%	Added feature engineering
+ Trend features	69%	Not added feature engineering

Finally, the Feature Engineering data consists of three types: **analysis features, label features, and past features.**

II. 2 - Setup & Training & Tuning

2.2. Model training and tuning workflow:

2. Hyperparameters



- Trees = 10,
- Max depth = 5

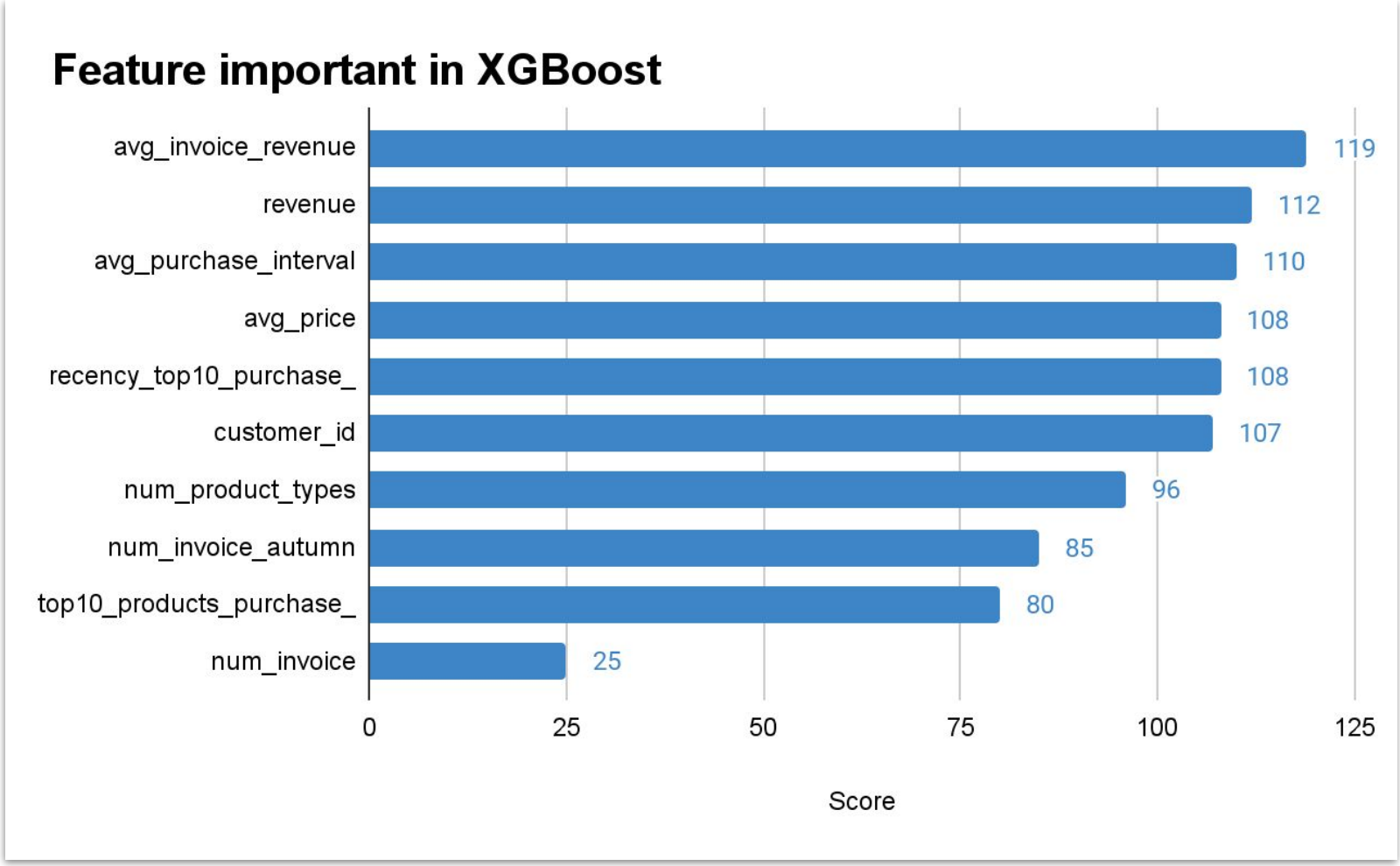
	lr	precision
0	0.1	0.800000
1	0.2	0.689024
2	0.3	0.666667
3	0.4	0.639896
4	0.5	0.628647
5	0.6	0.640097
6	0.7	0.611765
7	0.8	0.615566
8	0.9	0.563679
9	1.0	0.575949

The most optimal hyperparameters are:

- Number of trees = 10,
- Maximum depth = 5,
- Learning rate = 0.1.

II. 3 - Evaluation

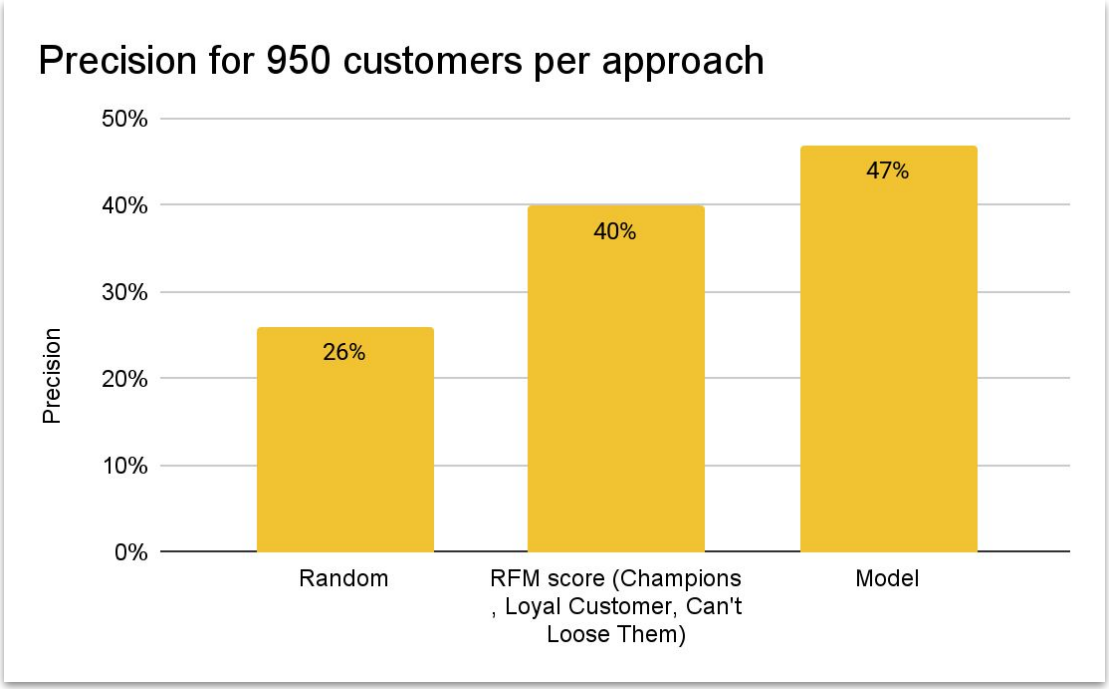
3.1. Feature important:



Among all features, **revenue** is the most important one, indicating that customer spending plays a key role in the prediction.

II. 3 - Evaluation

3.2. Compare with other approach:



The number of customers used for comparison was set to the total number of customers in the three key segments: **Champions, Loyal Customers, and 'Can't Lose Them'.**

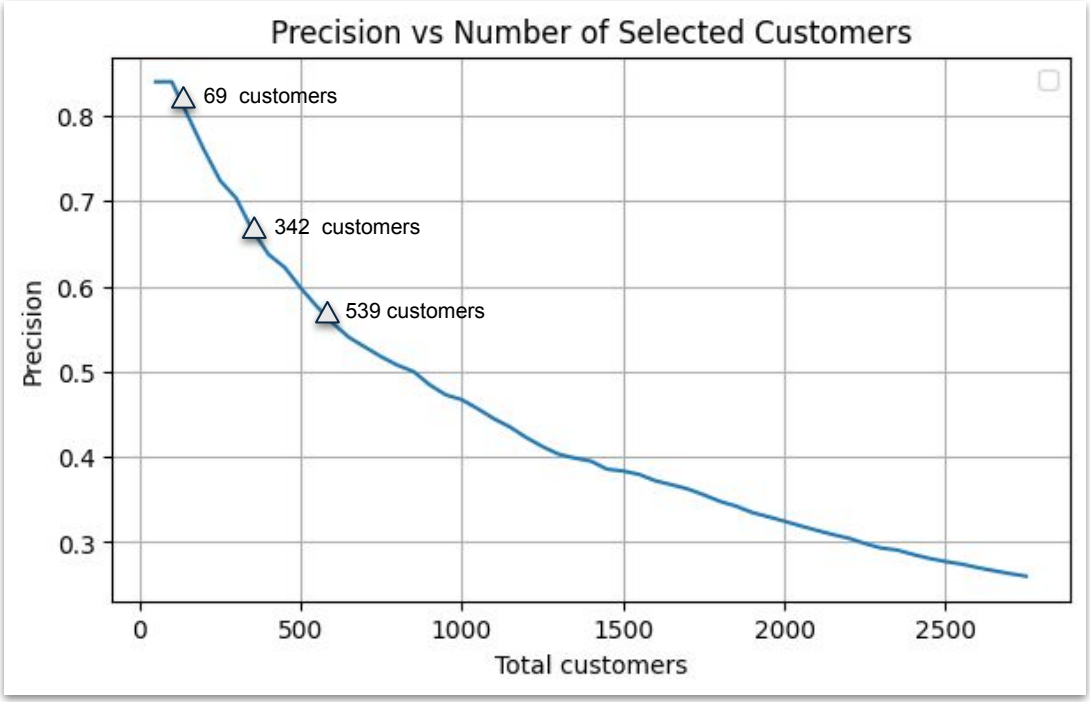
As shown in the figure:

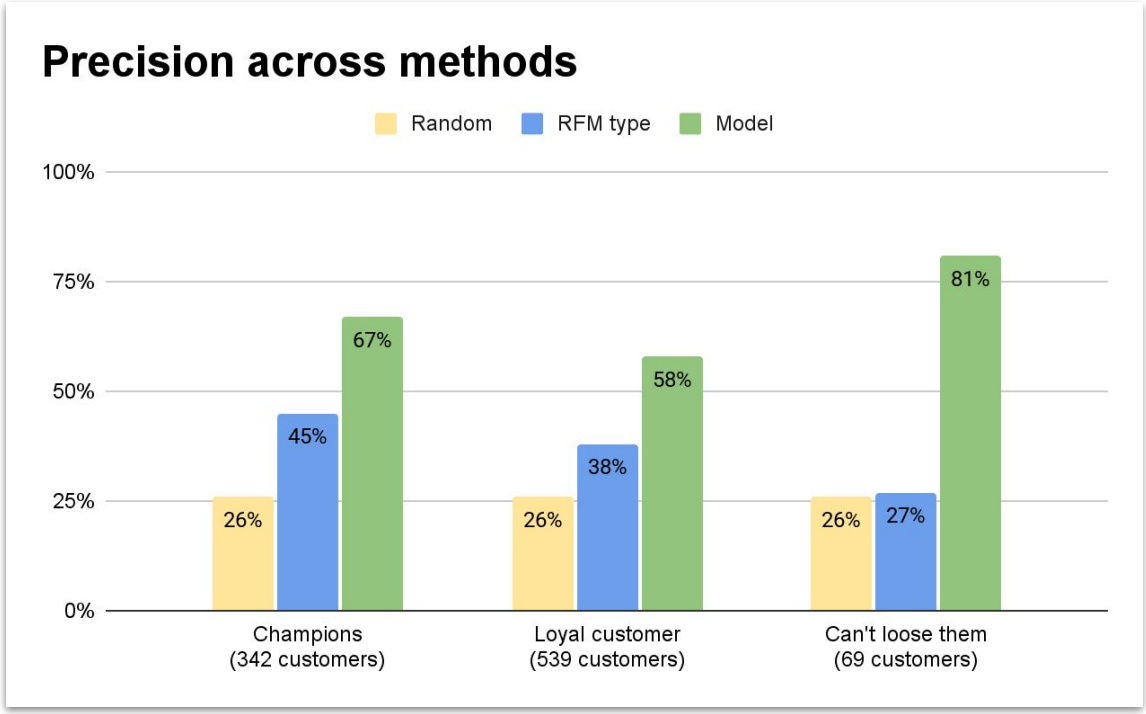
- Customers were randomly selected for comparison, the proportion of label-1 samples was the lowest, at only 26%.
- Next, when selecting only from the Champions, Loyal Customers, and 'Can't Lose Them' segments, the proportion increased to 40%.
- Finally, when the selection was based on the predicted probability of each customer being classified as label-1 by the model, the highest precision of 47% was achieved.
- Finally, by ranking customers based on the predicted probability by the model and selecting those with the highest probabilities, the highest precision of 47% was achieved.

=> These findings demonstrate the effectiveness of the proposed model.

II. 3 - Evaluation

3.3. Precision chart of modeling





From the chart, it is evident that customer groups selected by the model consistently achieve **higher precision** compared to traditional methods such as RFM segmentation or random selection. Specifically:

- The model performs **2 to 3 times better** than the RFM-based segmentation.
- Compared to random selection, the model shows an improvement of approximately **3 to nearly 4 times**.

This indicates that the model is effective at capturing key characteristics for identifying target customers, and it **can significantly outperform or complement traditional selection methods** in identifying high-value customer segments.

III. Conclusion

Analysis on retailer transaction data, including:

- General data overview
- Product scoring to identify potential products
- Customer segmentation using RFM
- Univariate and multivariate analysis comparing customers who did and did not purchase potential products

Built predictive model with:

- Feature engineering and hyperparameter tuning
- Model performance significantly better than baselines:
  - 3–4 times higher than random baseline
  - 2–3 times higher than RFM-based segmentation (Champions, Loyal Customers, Can’t Lose Them)

Results demonstrate the model’s effectiveness in identifying potential customers and supporting targeted strategies.