Introduction, Problem Statement and Objectives

In the B2C e-commerce sector, businesses face the challenge of optimizing their operations and enhancing customer understanding to drive revenue growth.

Our dataset from Kaggle, *transactions.csv*, contains a one-year record of e-commerce sales transactions comprising 500,000 rows and 8 columns. The descriptions of the dataset are explained below (Ramos, 2021).

Column Name	**Description**
CustomerNo	An identification number for each unique customer
TransactionNo	An identification number for each unique transaction
Date	The date on which the transaction was made.
ProductNo	An (alpha)numeric code for each unique product
ProductName	Name of Product
Price	Unit Price of the specific product
Quantity	Quantity purchased for a single product within the transaction
Country	Country where the customer is based in

Our first objective is to understand the contributing factors to customer loyalty to gain actionable insights for nurturing loyal customer relationships for sustained revenue growth. Our second objective is to understand customer behaviour through effective segmentation to recommend tailored customer targeting strategies.

Exploratory Data Analysis (EDA)

In our analysis of customer distribution across countries, we observed a dominant presence in the United Kingdom (Figure 1), leading us to focus our efforts on understanding the behaviour of UK-based customers. We also found that there was statistically a significant difference in transaction volumes across the different months of the year (Figure 2). Additionally, our examination of the average amount spent per transaction highlighted notable variance, prompting further investigation (Figure 3).

Delving into customer preferences, we found a great variance in the number of unique products purchased by each customer, limiting to 500 products, suggesting characterisation by purchase variety (Figure 4). Lastly, analyzing the average basket size, within 5000 items, allowed us to understand patterns in bulk purchases (Figure 5). This multifaceted exploration of our dataset provides us insights to incorporate both temporal and product-related factors for customer analysis.

Feature Engineering

We feature-engineered three variables, Recency, Frequency, and Monetary (RFM), as they are benchmark metrics for customer profiling (Murphy & Kvilhaug, 2022). In addition to RFM, we decided to feature engineer other variables based on our EDA findings to gather more granular insights into our model, shown in the table below.

Variable	**Description**
Recency	The number of days since the customer's last order
Frequency	The number of transactions made by each customer
Monetary	Cumulative revenue generated from each customer
duration	The number of days between the first and last purchase
loyal_customer	A binary variable with value 1 if that customer's Duration was above median duration and 0 if otherwise.
unique_products	The number of unique products each customer bought
average_basket_size	Customer's average number of items per transaction
avg_spend_per_trxcn	Customer's average transaction value
festive_spender	The categorical variable if the customer's highest spending month was within October to December period

Logistic Regression to Predict Customer Loyalty

Prior to running a regression, for all numeric variables except loyal_customer, we plotted a correlation matrix (Figure 6) and found a strong correlation value of

0.99025119 between average_basket_size and avg_spend_per_trxcn. We decided to remove the latter as we believe that transaction value can be indirectly inferred from Monetary and Frequency and the former gives its own unique insight. For our logistic regression model, we chose loyal_customer as a proxy for Customer loyalty which is our target variable. We first fit a base model, **Model_1**, with the standardised RFM variables to test their sufficiency in explaining customer loyalty. We then fit a second model, **Model_2**, with the standardised RFM variables along with standardised values for unique_products, average_basket_size and festive_spender to test whether these additional variables improve the explanatory power of the model. These models were fit on a training dataset. We then used the testing set to evaluate the models based on their Area under Receiver Operating Characteristic (AROC) values. We picked Model 2 as it had a higher AROC of 0.938. We also plotted the confidence intervals of **Model_2**'s predictors, showcasing their statistical significance (Figure 7). We recreated **Model_2** in the form of an ML pipeline consisting of six stages. In the first two stages, we indexed and encoded the festive_spender variable. In the next two stages, we collected and standardised all numeric variables of **Model_2**. For the last two stages, we assembled all variables into a single vector and ran the logistic regression. We put **Model_2** through a cross-validation pipeline to find the optimise hyper-parameters. Based on the AROC values, we chose our alpha and lambda values as 0 (Figure 8).

K-means Clustering Model

For our second model, we used K-means clustering for customer profiling. To find the optimal number of clusters, we obtained the silhouette scores for the respective number of clusters ranging from 2 to 6. We proceeded with 4 clusters as it had the highest silhouette score (Figure 9). For this model, we decided only to use the RFM

variables as adding further predictors would make generating insights complicated. We then built a three-stage ML pipeline. In the first stage, we would obtain the RFM variables. In the second stage, we would standardise the RFM variables. In the last stage, we would conduct K-means clustering for 4 clusters. We subsequently fitted our pipeline into a model. We then generated predictions of the clusters that our customers would fall into based on their average RFM values (Figure 10)

Communication of Results

From our logistic regression model, we found that the optimal way to increase customer loyalty would be through a rewards programme for making frequent purchases as customer Frequency had the highest positive coefficient and a gift prize for customers who spend a certain significant amount as customer Monetary value had the second highest positive coefficient as shown below.

```
Coefficients:
                    (Intercept)                   M_standardized                   F_standardized
                      1.1479583                        2.7102174                        5.1857637
                R_standardized average_basket_size_standardized     unique_products_standardized
                     -0.8392270                       -1.0126157                        0.2810471
            festive_spender_0
                      0.6504016
```

From our predicted K-means clustering model, Figure 10, each cluster had its own characteristics and advised the relevant marketing strategies as below:

Cluster	Characteristics	Recommended Actions
0	High frequency & monetary, most recent: **most loyal and valuable customers**	Attractive discounts and premium loyalty programs
1	Relatively recent, but lower Monetary and Frequency than cluster 2: **Potential to become part of most valuable**.	Inform them about tier-based loyalty programmes
2	Not very recent, with frequency near 1, low monetary value: **lost customers**	Send them a conversion incentive such as an irresistible discount
3	Relatively recent, but frequency and monetary low: **could be newer customers**	Welcome offers, email to join loyalty programs