University of Johannesburg

School of Consumer Intelligence and Information Systems
Centre for Applied Data Science

MARKETING ANALYTICS

Assignment 2

Author: Penekitete Héritier Kaumbu
Student Number: 220080995
Supervisor: Dr Karel Nzita Mayemba

April 22, 2025

Contents

1 Exploratory Data Analysis **iv**
- 1.1 Dataset Overview . . . iv
- 1.2 Descriptive Statistics . . . iv
- 1.3 Data Quality & Missing Values . . . v
- 1.4 Visual Exploration . . . v

2 RFM Segmentation **x**
- 2.1 Data Preview & Preparation . . . x
- 2.2 Segment Definitions . . . x
- 2.3 Results & Distribution . . . xi
- 2.4 Deep Dive Visualizations . . . xii
- 2.5 Key Insights & Strategic Recommendations . . . xiii

3 Latent Class Analysis (LCA) **xiv**
- 3.1 Variable Selection & Rationale . . . xiv
- 3.2 Data Preparation & Encoding . . . xiv
- 3.3 Model Selection via BIC . . . xiv
- 3.4 Class Assignment & Size . . . xiv
- 3.5 Class Profiling . . . xv
- 3.6 Interpretation of Each Class . . . xvi
- 3.7 Marketing Actions by Class . . . xvi

4 Price Elasticity of Demand (PED) **xvii**
- 4.1 OLS Regression Specification . . . xvii
- 4.2 Elasticity at the Mean . . . xvii
- 4.3 Midpoint Method Calculation . . . xvii
- 4.4 Interpretation & Pricing Strategy . . . xvii

5 Multiple Regression for Units Demanded **xviii**
- 5.1 Model Specification & Variables . . . xviii
- 5.2 Estimation Results & Diagnostics . . . xviii
- 5.3 Extensions & Segment-Specific Models . . . xviii

6 Conclusions & Recommendations **xix**
- 6.1 Summary of Key Findings . . . xix
- 6.2 Strategic Recommendations . . . xix
- 6.3 Limitations & Future Work . . . xix

Appendix **xx**

Executive Summary

Introduction

A South African electronics retailer seeks to enhance marketing effectiveness and optimise product demand by gaining deeper insights into its customer base and price sensitivity. To support this goal, we analyse a dataset of 60 recent customer transactions and interactions, which includes metrics for recency, frequency, and monetary value alongside price (Rand), units demanded, advertising spend, home characteristics (bedrooms, proximity to Gautrain), and product/channel preferences. This report applies four core analytical techniques: Recency–Frequency–Monetary (RFM) segmentation to classify customers into behavioural groups; latent class analysis (LCA) to uncover hidden segments based on categorical attributes; price elasticity of demand estimation via ordinary least squares and the midpoint method to inform pricing strategy; and a multiple regression model predicting units demanded from price, ad spend, home size, and station proximity to identify key demand drivers. The insights produced will guide targeted marketing actions, pricing decisions, and operational enhancements for the retailer.

Chapter 1

Exploratory Data Analysis

1.1 Dataset Overview

The dataset consists of 60 observations and 12 variables. Table 1.1 summarises the field names and types:

Variable	Type
Customer_ID	Integer
Recency	Integer
Frequency	Integer
Monetary	Integer
Price_Rand	Integer
Units_Demanded	Integer
Ad_Spend	Integer
Bedrooms	Integer
Near_Gautrain	Integer (0/1)
Product_Preference	Categorical
Purchase_Channel	Categorical
Loyalty_Member	Categorical

Table 1.1: Dataset overview: variables and data types

1.2 Descriptive Statistics

Table 1.2 presents summary statistics for the seven numeric variables. Note the wide ranges in `Recency` (2–93 days), `Price_Rand` (1146–6855 ZAR), and `Units_Demanded` (1–99 units). The median values (25–75% quartiles) suggest moderate skew in spend and demand.

Variable	Mean	Std. Dev.	Min	25%	50%	75%	Max
Recency	50.12	29.41	2	21.75	54.00	75.00	93.00
Frequency	9.52	5.24	1	5.00	9.00	14.00	19.00
Monetary (ZAR)	2739.3	1275.2	406	1720.5	2882.5	3792.5	4979.0
Price_Rand (ZAR)	4086.7	1708.4	1146	2800.0	3987.5	5415.8	6855.0
Units_Dem.	45.22	28.73	1	20.75	38.00	68.25	99.00
Ad_Spend (ZAR)	1474.1	818.2	116	749.75	1394.5	2167.3	2906.0
Bedrooms	2.07	0.78	1	1.00	2.00	3.00	3.00

Table 1.2: Descriptive statistics for numeric variables

The three categorical fields break down as follows:

- **Product_Preference:** TV (20), Mobile (15), Tablet (13), Laptop (12).
- **Purchase_Channel:** In-store (33), Online (27).
- **Loyalty_Member:** No (34), Yes (26).
- **Near_Gautrain:** Yes (32), No (28).

1.3 Data Quality & Missing Values

A completeness check revealed *zero* missing values across all 12 fields. Data types align with expectations (e.g. binary and categorical flags stored as integers or strings), so no imputation or type-conversion steps were required.

1.4 Visual Exploration

Figure 1.1 shows histograms with kernel density estimates for each numeric variable. Most distributions exhibit moderate skew:

- `Monetary` and `Price_Rand` are right-skewed, indicating a cluster of lower-spend customers and a tail of high-value outliers.
- `Units_Demanded` likewise has a long upper tail (max = 99).
- `Frequency` and `Recency` appear approximately symmetric around their medians.

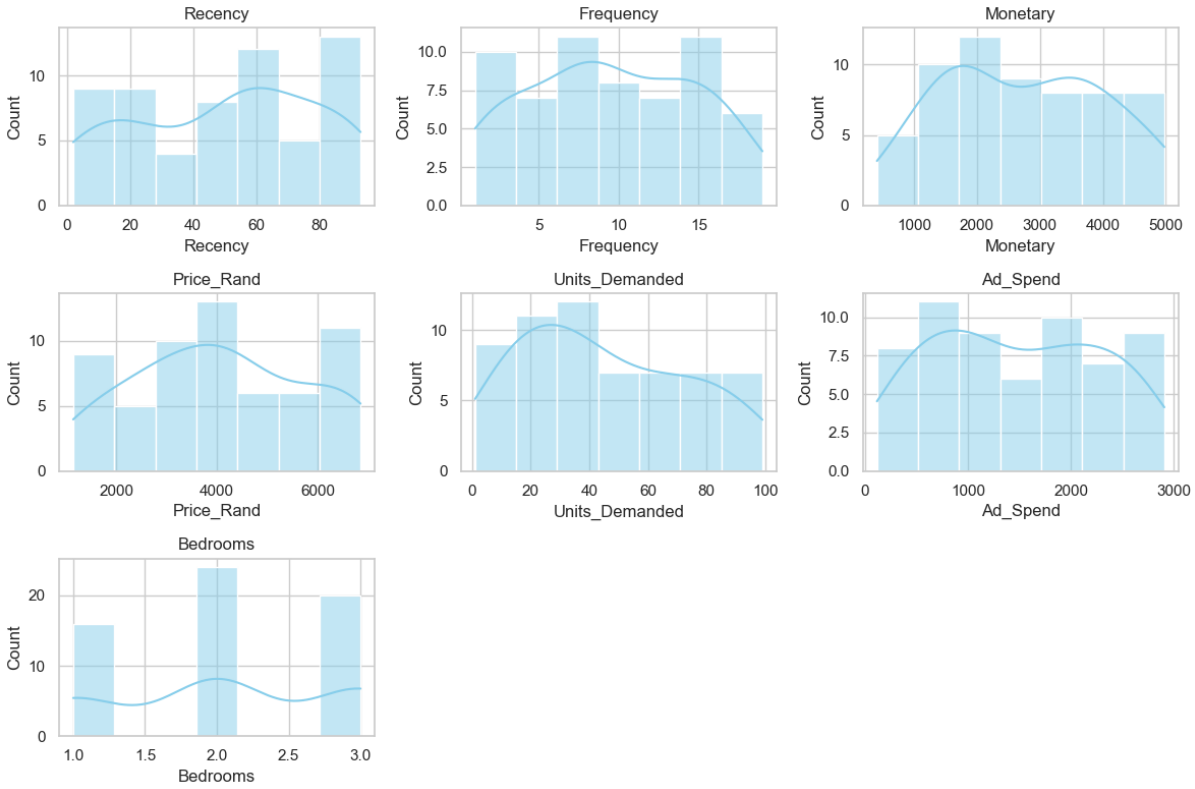

Figure 1.1: Histograms with KDE for numeric variables

Boxplots in Figure 1.2 confirm the presence of a few outliers and particularly in spend and demand but no critical data-entry errors.

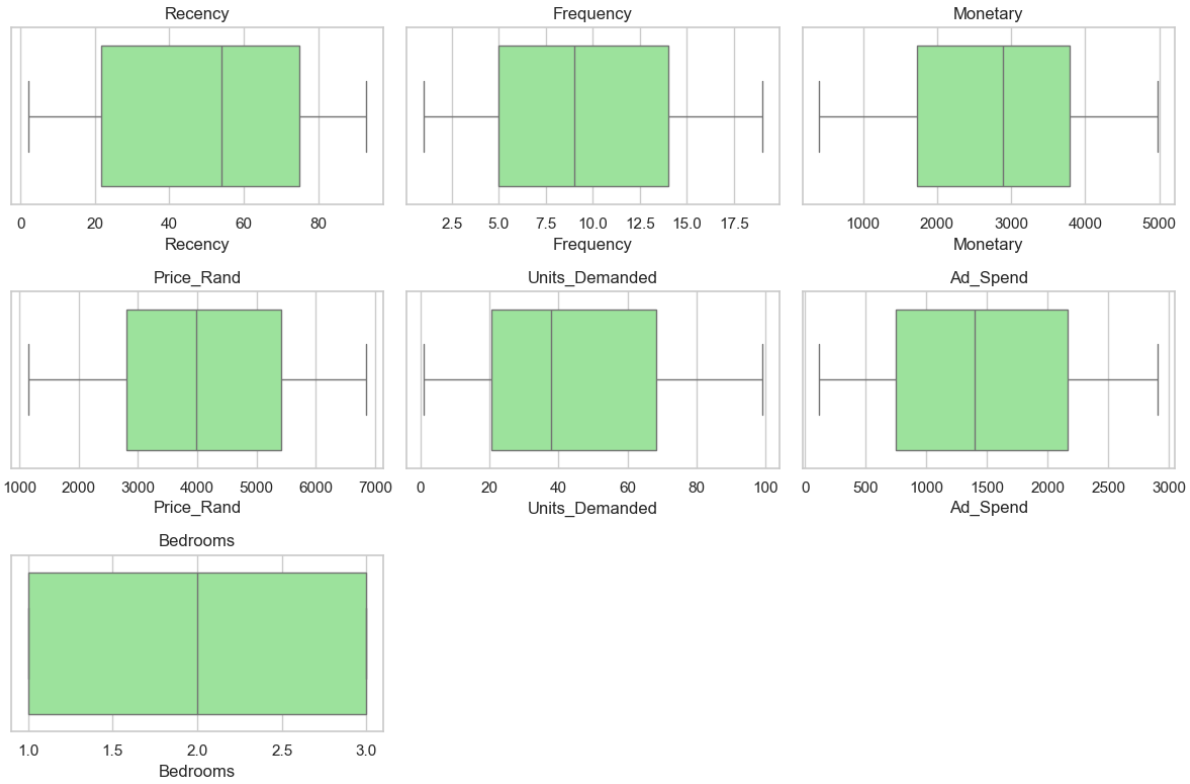

Figure 1.2: Boxplots for numeric variables

Categorical counts (Figure 1.3) show a slight tilt toward TV buyers and in-store purchases, with roughly equal loyalty membership and Gautrain proximity.

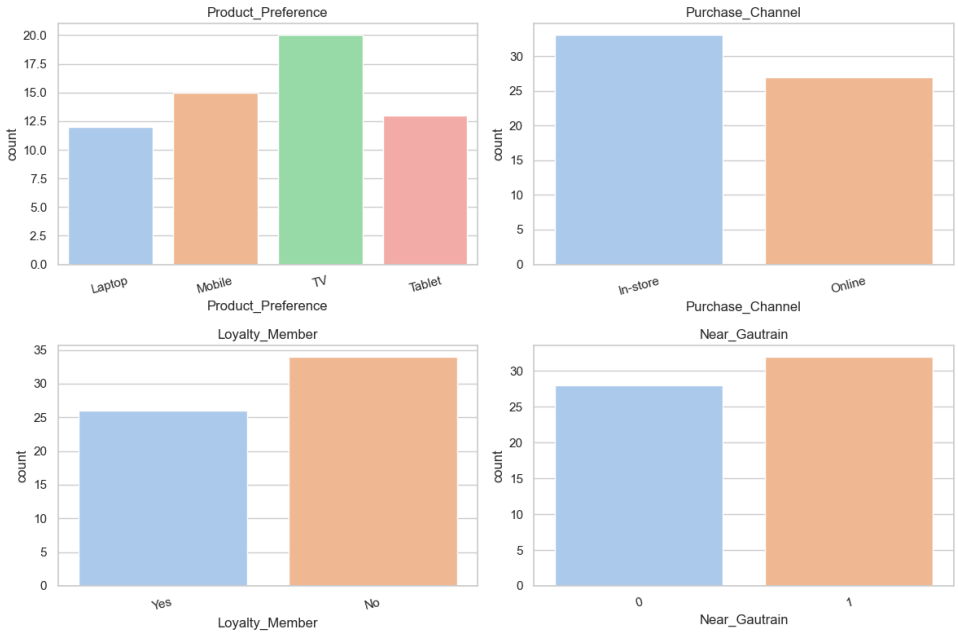

Figure 1.3: Bar charts for categorical variables

The pairplot in Figure 1.4 illustrates weak correlations among R, F, and M, and the correlation matrix in Figure 1.5 quantifies these relationships. Notably, Recency and

Frequency show a modest positive correlation (r=0.36), while most other feature pairs are near zero.

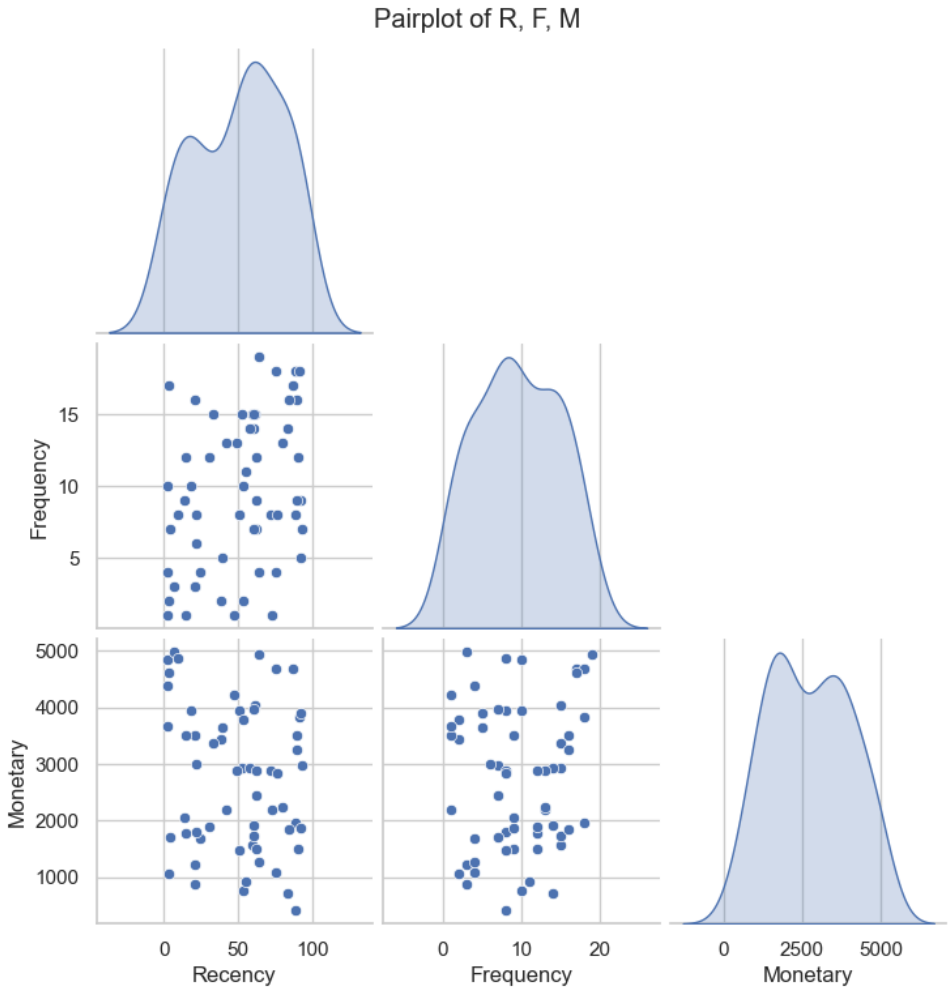

Figure 1.4: Pairwise scatter plots and KDE diagonals for Recency, Frequency, Monetary

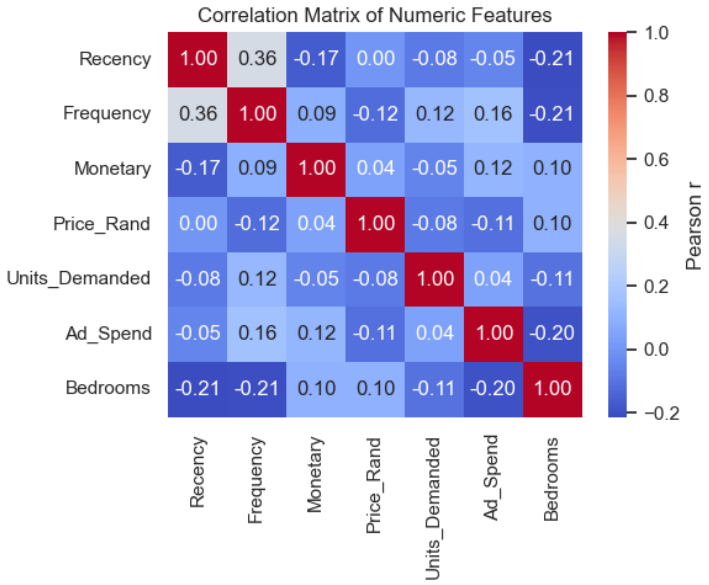

Figure 1.5: Correlation heat map of numeric features

Key takeaways:

- No missing data – ready for segmentation and modelling.
- Spend and demand are right-skewed; consider log-transforms in later regression if needed.
- Weak linear correlations suggest multivariate techniques (RFM, LCA, regression) will yield distinct insights.

Chapter 2

RFM Segmentation

In this section we compute Recency–Frequency–Monetary (RFM) scores for each customer, assign them to behavioural segments, and examine the distribution of those segments.

2.1 Data Preview & Preparation

We begin with a quick preview of the raw data. Table 2.1 shows the first five rows of the dataset, illustrating the key fields used in the RFM computation.

Customer_ID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score	Segment
1	52	15	2931	3	4	3	10	New Customers
2	93	7	2973	1	2	3	6	At Risk
3	15	12	1770	5	4	2	11	New Customers
4	72	8	2890	2	2	3	7	Need Attention
5	61	15	4040	2	4	5	11	New Customers

Table 2.1: Sample of computed RFM scores and assigned segments

The three raw metrics are:

- **Recency:** days since last purchase (lower is better)
- **Frequency:** total number of purchases (higher is better)
- **Monetary:** total spend (higher is better)

Each metric is binned into five quintiles, with Recency inverted (5 = most recent) and Frequency/Monetary labeled naturally (5 = highest). The sum of the three scores yields an overall RFM_Score in the range 3–15.

2.2 Segment Definitions

Customers are assigned to one of five segments based on their RFM_Score:

Champions $13 \leq \text{RFM\_Score} \leq 15$

New Customers $10 \leq \text{RFM\_Score} \leq 12$

Need Attention $7 \leq \text{RFM\_Score} \leq 9$

At Risk $5 \leq \text{RFM\_Score} \leq 6$

Lost $\text{RFM\_Score} < 5$

2.3 Results & Distribution

Table 2.2 reports the final counts in each segment:

Segment	Count
New Customers	24
Need Attention	21
At Risk	9
Lost	3
Champions	3

Table 2.2: Distribution of customers across RFM segments

Figure 2.1 shows the same information graphically, and Figure 2.2 maps each customer's Recency vs. Frequency score coloured by segment.

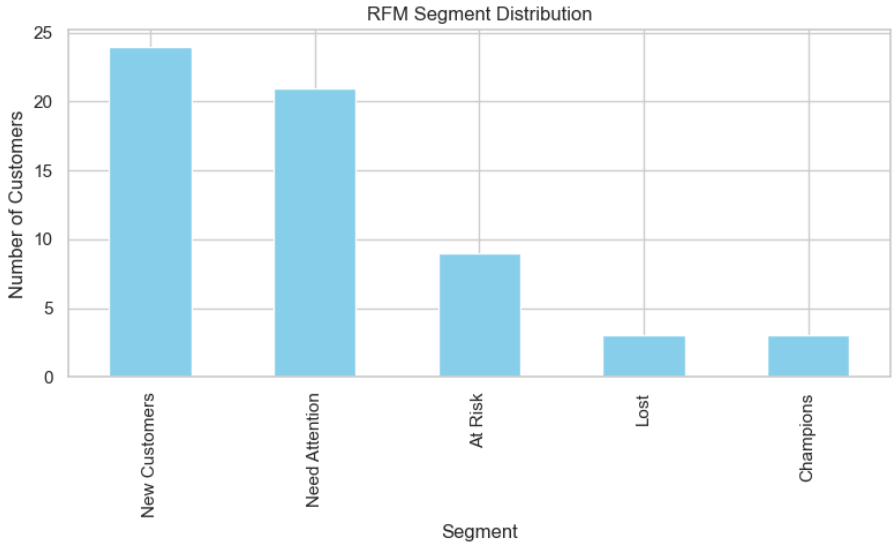

Figure 2.1: Bar chart of RFM segment counts

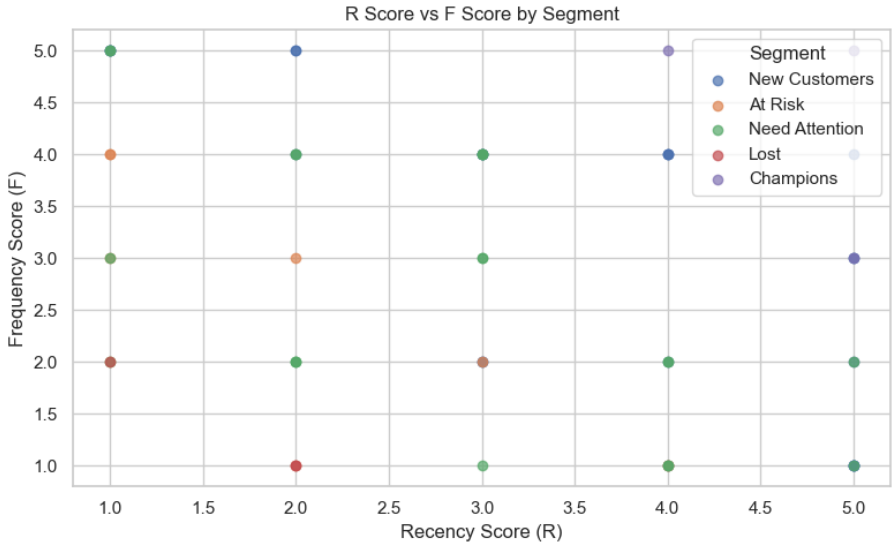

Figure 2.2: Scatter plot of R_Score vs F_Score by segment

2.4 Deep Dive Visualizations

To further explore how spend and scores vary by segment, we produced:

- A boxplot of Monetary spend by segment (Figure 2.3).
- A heatmap of the average R, F, M scores per segment (Figure 2.4).

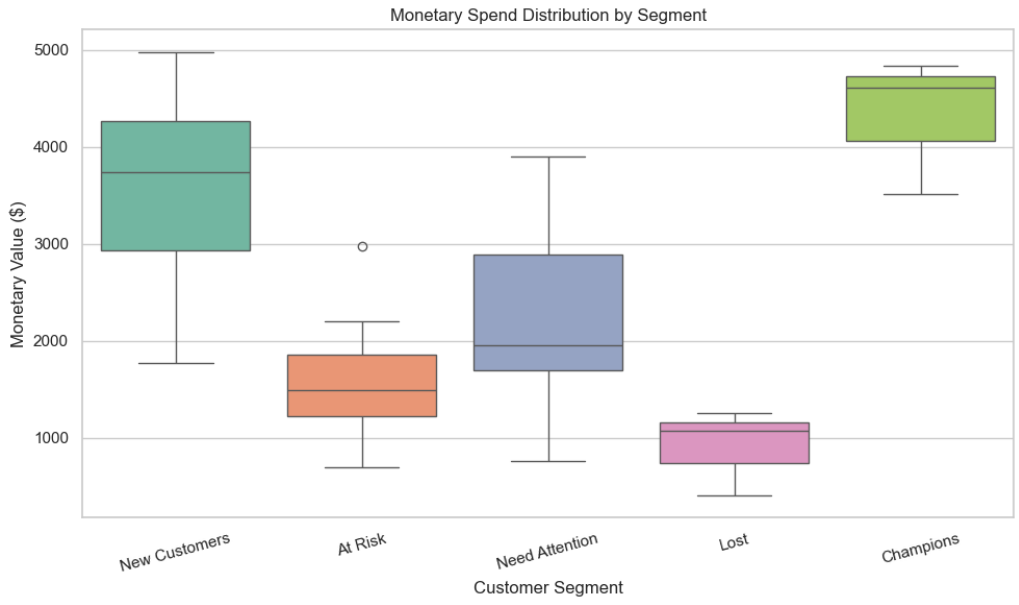

Figure 2.3: Monetary spend distribution by RFM segment

Figure 2.4: Average R, F, M scores by segment

2.5 Key Insights & Strategic Recommendations

- **New Customers (24):** Send a short welcome series (email/SMS) highlighting popular products and a one-time 10% cross-sell coupon to boost second purchases.
- **Need Attention (21):** Issue a limited-time 15% re-engagement offer via each customer's preferred channel, paired with a one-question survey to capture why they lapsed.
- **Champions (3):** Reward top buyers with VIP perks (e.g. early access to new launches) and a simple referral incentive (e.g. "Give R100, Get R100").
- **At Risk (9):** Trigger a "We miss you" campaign featuring a targeted discount bundle and, if possible, a brief personal outreach (call or in-store invite).
- **Lost (3):** Deploy a final steep-discount email; if there's no response, suppress these contacts to avoid wasted marketing spend.

Cross-Segment Tip: Tailor messaging by each group's dominant channel (in-store vs. online) and product preference, and recompute RFM quarterly to capture movement between segments.

Chapter 3

Latent Class Analysis (LCA)

We apply a model-based clustering (Gaussian Mixture) on three key categorical variables to uncover hidden customer segments. Below we describe the inputs, model selection, class assignment, profiling, and recommended marketing actions.

3.1 Variable Selection & Rationale

We select the following categorical attributes for LCA:

- **Product_Preference:** captures the main product category each customer buys (*Laptop*, *Mobile*, *TV*, *Tablet*).
- **Purchase_Channel:** indicates whether the transaction was in–store or online.
- **Loyalty_Member:** flags whether the customer belongs to the retailer's loyalty program.

These dimensions reflect product interest, channel behaviour, and program engagement which all critical for tailored marketing.

3.2 Data Preparation & Encoding

No missing values were found in the three selected fields, so we proceed with `OneHotEncoder(sparse_out` to convert each category into dummy variables. The resulting feature matrix has 60 rows and 10 columns (4 + 2 + 2 + intercept dropped by GMM).

3.3 Model Selection via BIC

We fit Gaussian Mixture Models for $k = 2$ through 6 classes and compute the Bayesian Information Criterion:

The minimum BIC at $k = 3$ indicates that a three-class solution best balances fit and parsimony.

3.4 Class Assignment & Size

Fitting the GMM with $k = 3$ yields the following class membership:

# Classes k	BIC
2	–2772.71
3	**–2845.41**
4	–2766.33
5	–2806.12
6	–2761.97

Table 3.1: BIC for different numbers of latent classes

Class Label	# Customers
0	10
1	25
2	25

Table 3.2: Distribution of customers across the three latent classes

Figure 3.1 visualizes these counts.

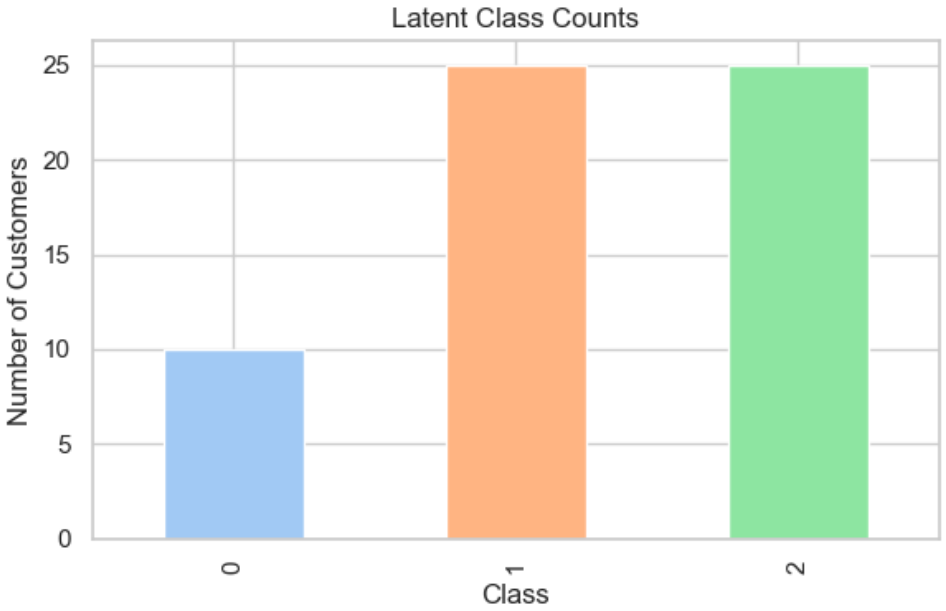

Figure 3.1: Bar chart of latent class sizes

3.5 Class Profiling

Table 3.3 cross-tabulates each class by product preference, purchase channel, and loyalty membership.

	Class 0	Class 1	Class 2
Product Preference			
Laptop	0	0	12
Mobile	0	15	0
TV	10	10	0
Tablet	0	0	13
Purchase Channel			
In-store	3	13	16
Online	7	12	9
Loyalty Member			
Yes	3	11	13
No	7	14	12

Table 3.3: Class profiles by product, channel and loyalty

3.6 Interpretation of Each Class

Class 0 (n=10) *TV-only Shoppers:* All members purchased TVs, with 70% online transactions, and low loyalty membership (30%).

Class 1 (n=25) *Mobile / TV Mixers:* Split between mobile (60%) and TV (40%) preferences, fairly balanced in-store vs. online (52% vs. 48%), moderate loyalty (44%).

Class 2 (n=25) *Tech Bundlers:* Primarily laptop (48%) and tablet (52%) buyers, mostly in-store (64%), with high loyalty membership (52%).

3.7 Marketing Actions by Class

- **Class 0 (TV-only Shoppers)** Deploy targeted online video ads showcasing TV bundles and accessory cross-sells, plus email promotions timed to peak viewing hours.
- **Class 1 (Mobile / TV Mixers)** Send mobile push notifications with flash deals on mobile accessories; run "bundle and save" campaigns combining mobiles and entry-level TVs.
- **Class 2 (Tech Bundlers)** Invite to in-store "bundle days" at Gautrain-adjacent locations, offering exclusive laptop + tablet package discounts and loyalty point bonuses.

Chapter 4

Price Elasticity of Demand (PED)

4.1 OLS Regression Specification

4.2 Elasticity at the Mean

4.3 Midpoint Method Calculation

4.4 Interpretation & Pricing Strategy

Chapter 5

Multiple Regression for Units Demanded

5.1 Model Specification & Variables

5.2 Estimation Results & Diagnostics

5.3 Extensions & Segment-Specific Models

Chapter 6

Conclusions & Recommendations

6.1 Summary of Key Findings

6.2 Strategic Recommendations

6.3 Limitations & Future Work

Appendix

Code Listings

Full Regression Output