

**CUSTOMER BEHAVIOR ANALYSIS AND SALES
TREND FORECASTING
FOR SME POINT-OF-SALE SYSTEMS**

Nawagamuwage Kasun Malshan Perera

(IT22151810)

A thesis submitted in partial fulfilment of the requirements for the
B.Sc. (Hons) Degree in Information Technology

Department of Information Technology
Faculty of Computing
Sri Lanka Institute of Information Technology
Sri Lanka

Individual Component: Customer Behavior Analysis and Sales Trend Analysis

Group ID: R26 IT-026

September 2026

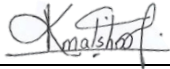
DECLARATION AND SUPERVISOR CERTIFICATION

I declare that this thesis is my own work and does not incorporate, without acknowledgement, material previously submitted for a degree or diploma at any university or institute of higher learning. To the best of my knowledge, it contains no material previously published or written by another person except where acknowledgement is made in the text. I grant the Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute this thesis, in whole or in part, in print, electronic, or other media, while retaining the right to use the content in future work.

I have disclosed the use of generative artificial intelligence and other automated tools in accordance with the applicable module policy. I remain responsible for the accuracy, originality, citations, code, data, analyses, interpretation, and conclusions presented in this thesis.

Student name: Nawagamuwage Kasun Malshan Perera

Student ID: IT22151810

Signature: 

Date: 07/10/2026

Supervisor certification. The above candidate carried out this research for the bachelor's degree thesis under my supervision.

Role and name	Signature	Date
Supervisor: Mr. Deemantha Siriwardhana		08/10/2026
Co-supervisor: Mrs. Ayesha Wijesooriya		08/10/26

ABSTRACT

Small and medium-sized enterprises frequently use point-of-sale systems that capture substantial transactional information but provide limited support for converting that information into customer-level and forward-looking business intelligence. This individual research investigates a customer behavior and sales trend analytics component within the SmartCost SME POS platform. The work combines Recency, Frequency and Monetary (RFM) analysis with K-Means clustering to segment identified customers and combines temporal sales analysis with regression-based trend modelling and short-term forecasting evaluation. The study uses the project transaction history containing 19,253 order records and 47,930 order-item records covering March 2024 to August 2026. Customer-linked transactions are used to derive behavioral features, while order-item data supports product and temporal analysis. The research uses chronological hold-out validation to evaluate forecasting methods rather than treating historical trend visualization as evidence of predictive accuracy. The implemented analytical service is exposed through Python and FastAPI and integrated with the wider SmartCost application [18]. Independent reproduction of the RFM and K-Means procedure produced four customer groups containing 16, 54, 97 and 14 customers. The application presents interpretable VIP/Loyal, Regular, Occasional and At Risk/Lapsed segments together with peak-hour, weekend/weekday, product and revenue-trend views. A 52-response user evaluation showed 84.6% positive responses to the statement that customer and sales insights help users understand sales patterns. In the historical hold-out experiment, simple exponential smoothing achieved lower MAE, RMSE and MAPE than the other tested lightweight baselines. The findings support the use of integrated customer analytics and carefully validated short-term forecasting as complementary capabilities within an SME-oriented POS environment.

Keywords: Customer Behavior Analysis, RFM, K-Means Clustering, Sales Trend Analysis, Short-Term Forecasting

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Mr. Deemantha Siriwardhana, and my co-supervisor, Mrs. Ayesha Wijesooriya, for their guidance, feedback and academic support throughout the research and development of this individual component.

I also acknowledge the support of the Faculty of Computing and the Department of Information Technology of the Sri Lanka Institute of Information Technology. The advice received during the research and development process helped me improve the analytical focus of the work and distinguish the research contribution from the implementation of the software product.

I am grateful to the members of the SmartCost project team for their cooperation in developing the wider POS platform and for making the shared transactional data, system interfaces and project artefacts available for individual research. The quality assurance and project-management records were also useful for understanding the system context and documenting the development process.

Finally, I thank my family, friends and everyone who supported me during the completion of this research and thesis

TABLE OF CONTENTS

Declaration and Supervisor Certification	i
Abstract	ii
Acknowledgements	iii
List of Figures	ix
List of Tables	x
List of Abbreviations	xi
1 INTRODUCTION	1
1.1 Overall Project Context	1
1.2 Individual Component Context	1
1.3 Problem Statement	2
1.4 Research Gap	3
1.5 Aim and Objectives	3
1.6 Research Questions or Hypotheses	4
1.7 Scope, Limitations and Assumptions	5
1.8 Individual Contribution	5
1.9 Thesis Structure	6
2 LITERATURE REVIEW AND RELATED WORK	7
2.1 Review Method and Scope	7
2.2 Domain and Theoretical Foundations	7
2.2.1 Point-of-sale analytics	7
2.2.2 Customer behavior analysis	7
2.2.3 RFM analysis	8
2.2.4 K-Means clustering	8
2.2.5 Temporal sales analysis	9

2.2.6	Sales trend analysis	9
2.3	Related Research	9
2.4	Existing Systems, Products and Techniques	10
2.5	Gap Synthesis and Conceptual Framework	11
3	RESEARCH METHODOLOGY	13
3.1	Research Design	13
3.2	Data, Participants and Study Materials	13
3.3	Data Sources and Validation.....	14
3.3.1	Source-level validation.....	14
3.3.2	Record-level validation.....	14
3.3.3	Analytical reproduction validation	16
3.4	Experimental Procedure.....	16
3.5	Forecasting Design	17
3.6	Forecasting Approaches.....	18
3.6.1	Linear regression	18
3.6.2	Naive baseline	18
3.6.3	Three-period moving average.....	18
3.6.4	Simple exponential smoothing	18
3.7	Measures and Evaluation Metrics	18
3.8	Data Analysis.....	19
3.9	Ethics, Privacy, Security and Responsible AI	19
3.10	Reproducibility and Research Artefacts.....	19
4	PRODUCT AND SOLUTION DESIGN	21
4.1	Requirements	21
4.1.1	Non-functional requirements.....	21
4.2	Component Architecture Within the Complete Solution	22
4.3	Design Alternatives and Decisions	22
4.3.1	Customer segmentation alternatives	22
4.3.2	Sales forecasting alternatives	22
4.3.3	Database and service architecture.....	23
4.4	Detailed Design	23
4.4.1	RFM feature pipeline.....	23
4.4.2	Cluster interpretation.....	23
4.4.3	Sales trend pipeline	23
4.4.4	Insight generation.....	24

4.5	Standards and Professional Practice	24
5	IMPLEMENTATION, INTEGRATION AND TESTING	25
5.1	Development Environment and Technology Stack	25
5.2	Component Implementation.....	25
5.2.1	FastAPI segmentation endpoint.....	25
5.2.2	RFM feature construction.....	25
5.2.3	K-Means configuration.....	26
5.2.4	Sales trend implementation	26
5.3	Integration With the Other Components	27
5.4	Verification and Software Testing.....	27
5.4.1	Analytical verification test cases	28
5.5	Deployment and Operation	29
6	RESULTS AND VALIDATION	30
6.1	Research Results	30
6.1.1	RFM and K-Means results	30
6.1.2	At-risk customer result	31
6.1.3	Peak-hour and temporal results	31
6.1.4	Weekend versus weekday results.....	32
6.1.5	Product results	33
6.1.6	Sales trend results.....	35
6.2	Product and Technical Results	38
6.3	Forecasting Validation Results.....	40
6.3.1	Forward projection after the most recent complete month	42
6.4	User or Stakeholder Validation.....	43
6.5	Integration Results	44
6.6	Summary of Findings.....	45
7	DISCUSSION	46
7.1	Interpretation of Findings	46
7.2	Comparison With Literature Baselines and Existing Solutions	46
7.3	Research Contribution	46
7.4	Product Contribution.....	47
7.5	Limitations and Threats to Validity	47
7.6	Implications	48
8	COMMERCIALISATION AND SUSTAINABILITY	49

8.1	Value Proposition and Target Users	49
8.2	Market and Competitive Position.....	49
8.3	Business or Sustainability Model.....	49
8.4	Scalability, Intellectual Property and Compliance	49
8.5	Contribution to the Sustainable Development Goals.....	50
9	CONCLUSION AND FUTURE DIRECTIONS	51
9.1	Conclusion	51
9.2	Achievement of Objectives	51
9.3	Individual Research and Product Contributions.....	52
9.4	Future Directions	52
	References	53
	Glossary	54
A	RESEARCH INSTRUMENTS AND EXTENDED RESULTS	57
A.1	Dataset Data Dictionary	57
A.2	RFM Calculation Example	57
A.3	Extended RFM Results	58
A.4	Forecasting Experiment Detail.....	58
A.5	User Evaluation Instrument	58
A.6	User Evaluation Extended Results	59
A.7	QA Evidence.....	59
B	ACADEMIC INTEGRITY AND AI USE DISCLOSURE	60
C	EXTENDED ANALYTICAL EVIDENCE AND VALIDATION	61
C.1	Purpose of Extended Validation	61
C.2	Data Quality Dimensions	61
C.2.1	Completeness	61
C.2.2	Consistency	61
C.2.3	Validity.....	61
C.2.4	Uniqueness	61
C.2.5	Timeliness	62
C.3	Customer Segmentation Sensitivity	62
C.4	Revenue Concentration	62
C.5	Temporal Analysis Considerations.....	63
C.6	Forecasting Period Definition	63

C.7	Hold-Out Design.....	63
C.8	Forecasting Error Interpretation.....	64
C.9	Implications for Model Selection.....	64
C.10	Research Traceability Matrix.....	64
C.11	Research Contribution Versus Product Function.....	65
C.12	Threats to Internal Validity.....	65
C.13	Threats to External Validity.....	65
C.14	Threats to Construct Validity.....	65
D	ADDITIONAL RESULTS TABLES AND ANALYTICAL INTERPRE-	
	TATION	67
D.1	Full Monthly Revenue Table	67
D.2	Forecast Prediction Table	67
D.3	Forecasting Discussion by Method	68
D.3.1	Linear regression	68
D.3.2	Naive forecast.....	68
D.3.3	Moving average.....	68
D.3.4	Simple exponential smoothing	68
D.4	Customer Segment Revenue Distribution	68
D.5	User Evaluation Detailed Interpretation.....	69
D.6	User Evaluation Limitations	70
E	SYSTEM SCREENSHOT EVIDENCE	71
E.1	Customer Behavior Main Screen	71
E.2	Busiest Hours Analysis.....	72
E.3	Weekend and Weekday Analysis	73
E.4	Top-Selling Products.....	74
E.5	Slow-Moving Products	75
E.6	Customer Segmentation.....	76
E.7	Sales Trend Trajectory	77
E.8	Custom Report Configuration.....	78
F	QUALITY ASSURANCE AND PROJECT EVIDENCE	79
F.1	Relevant Quality Assurance Finding.....	79
F.2	Interpretation of QA Evidence.....	79
F.3	Project Management Evidence	79

LIST OF FIGURES

2.1	Conceptual framework for the individual research	12
4.1	Architecture of the individual analytics component within SmartCost .	22
5.1	Implemented sales trend and revenue trajectory interface	27
5.2	Implemented customizable customer behavior report interface	29
6.1	Implemented RFM and K-Means customer segmentation interface . .	31
6.2	Implemented busiest-hours analysis	32
6.3	Implemented weekend-versus-weekday comparison	33
6.4	Implemented top-selling product analysis	34
6.5	Implemented slow-moving product analysis	35
6.6	Implemented sales trend and revenue trajectory	36
6.7	Monthly revenue observations and fitted linear trend	37
6.8	Implemented customer behavior and sales insights interface	39
6.9	Implemented report feature selection	40
6.10	Historical hold-out forecast comparison	41
6.11	Forward revenue projection after July 2026	43
D.1	Customer segment revenue share from raw-data reproduction	69
E.1	Implemented Customer Behavior and Sales Insights overview interface	71
E.2	Implemented busiest-hours analysis interface	72
E.3	Implemented weekend-versus-weekday sales comparison	73
E.4	Implemented top-selling product analysis	74
E.5	Implemented slow-moving product analysis	75
E.6	Implemented RFM and K-Means customer segmentation interface . .	76
E.7	Implemented sales trend and revenue trajectory interface	77
E.8	Implemented customer behavior report customization interface	78

LIST OF TABLES

2.1	Analytical approach comparison used in the research	11
3.1	Primary dataset characteristics	14
3.2	Data-quality validation summary for the supplied project dataset	15
3.3	Reproduced RFM/K-Means cluster profiles	16
4.1	Functional requirements for the individual component	21
5.1	Technology stack used by the individual component	25
5.2	Relevant QA evidence for the individual analytics component	28
5.3	Analytical verification test cases	28
6.1	Customer segment results from raw-data reproduction	30
6.2	Monthly revenue used in the 11-period trend analysis	38
6.3	Historical hold-out forecasting comparison	40
6.4	Forward revenue projection from the latest complete month	42
6.5	User evaluation results for analytics-related questions	44
A.1	Orders sheet analytical data dictionary	57
A.2	Order Items sheet analytical data dictionary	57
A.3	Extended RFM cluster profile	58
A.4	Selected user evaluation distributions	59
C.1	Interpretation of RFM cluster characteristics	62
C.2	Historical forecasting split	63
C.3	Traceability from research question to evidence	64
C.4	Distinction between product functionality and research contribution . .	65
D.1	Extended monthly revenue observations	67
D.2	Forecast values for the historical hold-out experiment	67
D.3	Customer-level revenue concentration	68

LIST OF ABBREVIATIONS

Abbreviation	Meaning
AI	Artificial Intelligence
API	Application Programming Interface
BI	Business Intelligence
CSV	Comma-Separated Values
ETL	Extract, Transform and Load
FastAPI	Python framework for building REST APIs
K-Means	K-Means clustering algorithm
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
POS	Point of Sale
RFM	Recency, Frequency and Monetary
RMSE	Root Mean Squared Error
SME	Small and Medium-sized Enterprise
UI	User Interface
UX	User Experience

1 INTRODUCTION

1.1 Overall Project Context

Point-of-sale systems are increasingly used by small and medium-sized enterprises to record transactions, manage products and support daily business operations. A modern POS system can capture product identifiers, quantities, prices, timestamps, customer identifiers and payment information for every transaction. However, the presence of data does not automatically provide useful knowledge. A business owner still needs analytical mechanisms that can transform raw records into interpretable information about customer behavior, product demand and sales performance.

The SmartCost project was developed as an integrated SME-oriented POS and business intelligence platform. The overall research proposes lightweight architecture that connects operational POS functions with analytical services. The research paper describes a workflow involving data collection, preprocessing, aggregation, pattern analysis, visualization and evaluation. The wider system uses MongoDB, Node.js, FastAPI, Python analytical libraries and a React Native interface to support this workflow [17][20].

The individual component covered by this thesis is concerned specifically with customer behavior analysis and sales trend analysis. The component is intended to answer practical questions such as: Which customers contribute most to revenue? Which customers have become inactive? When are sales strongest? Which products are selling frequently or slowly? Does the historical revenue series exhibit a stable or changing trajectory? Can historical information be used to construct and validate a short-term forecasting process?

These questions are important because a POS application that only records transactions remain largely descriptive. An analytical POS component can support decision-making by organizing historical data into customer segments, temporal patterns and forward-looking estimates. The distinction between descriptive analysis and predictive analysis is particularly important because a research-grade analytical component should evaluate whether historical information can support future-oriented estimation.

1.2 Individual Component Context

The individual component is implemented as part of the SmartCost customer behavior and sales insights interface. The implemented interface provides a time filter, key

insights, busiest-hour analysis, weekday-versus-weekend comparison, top-selling and slow-moving products, customer segmentation, sales trend visualization and report export functionality.

The component also contains a Python analytical service. The source code supplied for the individual research contains Fast API routing and an RFM/K-Means segmentation implementation. The implementation constructs Recency, Frequency and Monetary features, standardizes those features using StandardScaler, and applies K-Means with four clusters, ten initializations and a fixed random state of 42. The resulting clusters are mapped into interpretable customer labels used by the application. The sales trend interface uses a linear regression trajectory over period-level revenue. The screenshot evidence shows eleven periods, labelled from October to August, with a period revenue bar, rolling average and target baseline. The interface therefore provides a visual historical trajectory. The research contribution in this thesis extends the analytical evaluation beyond the interface by defining a reproducible forecasting experiment and a hold-out evaluation procedure.

1.3 Problem Statement

Many SME POS systems capture detailed transactional records but primarily use those records for billing, stock operations and basic reporting. The research paper identifies a gap between transaction processing and integrated behavioral intelligence. Business owners may therefore have access to large quantities of historical information without having a systematic mechanism for identifying customer segments, temporal demand patterns or forward-looking revenue behavior.

The problem addressed by this individual research can be stated as follows:

How can transactional data from an SME POS environment be transformed into validated customer behavior segments and sales trend information, and how can historical sales data be evaluated for short-term future trend prediction in a lightweight analytical architecture?

The problem has four related dimensions. First, customer transactions need to be transformed into meaningful behavioral features. Second, those features need to be segmented using a reproducible clustering procedure. Third, historical sales records need to be aggregated and analyzed across time. Fourth, trend analysis needs to be evaluated as a predictive process rather than presented only as a retrospective chart.

1.4 Research Gap

The research paper identifies a gap in SME-focused analytical systems because many traditional POS platforms concentrate on transaction processing and basic reporting. It also identifies a fragmentation problem in which reporting, visualization and analytical functions may exist separately rather than as an integrated decision-support workflow.

Two methodological gaps are addressed in this individual study. First, the analytical dataset and processing pipeline need explicit validation so that customer and sales results can be reproduced from source records. Second, analytical approaches need to be compared experimentally so that the research can distinguish descriptive visualization from predictive performance.

Consequently, the research gap addressed by this thesis is not simply the absence of a dashboard. The deeper gap concerns the validation of an integrated analytical workflow that combines customer segmentation with sales trend analysis and evaluates whether historical information can support short-term prediction under the constraints of an SME POS environment.

1.5 Aim and Objectives

The aim of this individual research is:

To design, implement and experimentally evaluate a customer behavior and sales trend analytics component that transforms SME POS transaction data into validated customer segments, temporal sales insights and short-term sales trend predictions.

The objectives are:

- O1.** To prepare and validate a multi-period transactional dataset for customer behavior and sales trend analysis.
- O2.** To derive Recency, Frequency and Monetary features from identified customer transactions.
- O3.** To apply and evaluate K-Means clustering for customer segmentation using standardized RFM features.
- O4.** To analyse sales behavior across temporal periods and product-level dimensions.
- O5.** To investigate alternative trend and forecasting approaches and define a reproducible short-term forecasting horizon.

- O6.** To evaluate forecasting performance using held-out historical periods and error-based metrics.
- O7.** To integrate analytical outputs into an SME-oriented POS interface that presents interpretable business insights.
- O8.** To validate the usefulness of the analytical outputs through user feedback and comparison with conventional POS functionality.

1.6 Research Questions or Hypotheses

The research is guided by the following questions:

- RQ1:** How can RFM features derived from actual SME POS transactions be used to identify meaningful customer behavior segments?
- RQ2:** How reproducible are the resulting customer segments when the clustering procedure is applied to the validated dataset?
- RQ3:** What temporal and product-level patterns can be identified from the transaction history?
- RQ4:** Which forecasting approach provides the lowest error under the defined short-term hold-out evaluation procedure?
- RQ5:** To what extent do the analytical outputs provide information that is not normally available through transaction-focused POS reporting?
- RQ6:** How do users perceive the usefulness of customer and sales analytics for understanding business performance and sales patterns?

The forecasting experiment uses a short-term horizon defined in this thesis as the next three monthly periods after the most recent complete historical period. The use of a three-month horizon is intended to match operational planning decisions such as stock procurement, staffing and promotional planning. Because the supplied dataset ends during August 2026, forecasts beyond the available observation period are treated as model outputs rather than verified future observations. Historical hold-out evaluation is therefore used to estimate predictive performance.

1.7 Scope, Limitations and Assumptions

The scope includes customer-level segmentation, temporal sales analysis, product performance analysis and short-term revenue forecasting within the SmartCost POS environment. The main customer segmentation variables are Recency, Frequency and Monetary value. The primary clustering algorithm implemented by the component is K-Means.

The scope does not include medical, financial-credit or other high-risk customer profiling. Customer identifiers are used only for grouping transactional records in the analytical dataset. The thesis does not attempt to infer personal characteristics that are not present in the POS data.

The main dataset contains 19,253 order records and 47,930 order-item records covering March 2024 to August 2026. The data contains 181 unique customer-mobile identifiers and 20 unique products. Customer-linked orders account for 12,410 records. These values describe the supplied project dataset and are not claimed to represent the wider SME population.

A further limitation concerns the difference between raw data precomputation and the application screenshot. Re-running the supplied RFM/K-Means process over the complete dataset produces 97 customers in the lower-spending residual segment, while the application screenshot displays 98 customers in the Occasional segment. This one-customer discrepancy is treated as a reproducibility finding and is discussed rather than hidden.

The sales trend model is limited by the length, completeness and business-specific nature of the dataset. August 2026 is incomplete in the raw dataset, so it should not be treated as a fully observed month when estimating future performance. Forecasting results are therefore evaluated using historical hold-out periods and are interpreted as short-term analytical estimates rather than guarantees.

1.8 Individual Contribution

The individual contribution includes the analytical implementation and evaluation of customer segmentation and sales trend functionality within the SmartCost project. The contribution includes the Python/FastAPI analytical service, RFM feature construction, K-Means clustering configuration, customer segment interpretation, sales trend visualization support, analytical insight presentation and research-oriented evaluation. The individual contribution is distinguished from other project modules. The group project also contains POS operations, table management, inventory, order management, product recommendations, kitchen display and other functions. These are treated as system context rather than as the core research contribution of this thesis.

The MS Planner evidence records several completed tasks assigned to the author, including table and floor-plan management, order-hub work, product search, offline cart synchronization, kitchen display work, receipt printing and customer credit functionality. Those records support the wider development history but are not used as evidence that the author implemented unrelated analytics tasks. The analytics contribution in this thesis is supported primarily by the source code, dataset, system screenshots and analytical results.

1.9 Thesis Structure

Chapter 1 introduces the problem, gap, objectives, questions, scope and individual contribution. Chapter 2 reviews the theoretical foundations and related work. Chapter 3 describes the research design, data, preprocessing, analytical methods, evaluation metrics, ethics and reproducibility. Chapter 4 presents the product and solution design. Chapter 5 explains implementation, integration and testing. Chapter 6 presents research results, technical results, user validation and integration evidence. Chapter 7 discusses the findings, comparisons, contributions and limitations. Chapter 8 addresses commercialization and sustainability. Chapter 9 concludes the research and identifies future directions.

2 LITERATURE REVIEW AND RELATED WORK

2.1 Review Method and Scope

The literature review is focused on the analytical concepts required by the individual component: POS analytics, customer behavior analysis, RFM segmentation, clustering, temporal sales analysis, regression and forecasting, business intelligence and decision support. The original research paper supplied for this project provides a foundation of twenty references covering data mining, business intelligence, association rule mining, ETL, machine learning, forecasting and the technologies used in SmartCost.

The review is organized around the research questions rather than around individual software technologies. This approach is important because the research contribution is methodological and analytical. The objective is to understand which techniques can transform transactional records into customer and sales intelligence, how those techniques can be evaluated, and what limitations remain when they are applied to SME environments.

2.2 Domain and Theoretical Foundations

2.2.1 Point-of-sale analytics

A POS system records operational events such as sales, products, quantities, prices, timestamps and payment details. Traditional POS implementations prioritize transaction completion and operational accuracy. However, the same records can support analytical activities when they are organized into consistent datasets.

Business intelligence systems extend this concept by integrating data, applying analytical procedures and presenting outputs through visual interfaces [9]. Turban, Sharda and Delen describe business intelligence and decision-support systems as mechanisms for transforming organizational data into information that can support managerial decisions [9]. In the SME context, the challenge is to provide useful analytical functionality without imposing the cost and complexity of enterprise platforms. structure in transactional datasets [1], [2], [12].

For this thesis, customer behavior is operationalized through RFM variables. This provides an interpretable representation because each feature has a direct business meaning. Recency reflects inactivity, Frequency reflects engagement and Monetary reflects revenue contribution.

2.2.2 Customer behavior analysis

Customer behavior analysis aims to identify recurring patterns in customer purchasing activities. Transactional attributes such as purchase frequency, monetary contribution and time since the most recent purchase provide useful behavioral signals. Data mining literature identifies clustering and aggregation as suitable approaches for discovering structure in transactional datasets [1], [2], [12]. For this thesis, customer behavior is operationalized through RFM variables. This provides an interpretable representation because each feature has a direct business meaning. Recency reflects inactivity, Frequency reflects engagement and Monetary reflects revenue contribution.

2.2.3 RFM analysis

RFM analysis represents each customer using three measures. Recency is the elapsed time since the customer's most recent purchase. Frequency is the number of transactions during the selected analysis period. Monetary is the total monetary value associated with the customer's transactions.

The basic definitions are:

$$R_i = D_{\text{ref}} - D_{\text{last},i} \quad (2.1)$$

$$F_i = |\{o : o \text{ belongs to customer } i\}| \quad (2.2)$$

$$M_i = \sum_{o \in O_i} V_o \quad (2.3)$$

where D_{ref} is the reference date, $D_{\text{last},i}$ is the last transaction date for customer i , O_i is the set of transactions belonging to customer i , and V_o is the transaction value.

RFM is attractive in SME systems because the features can be computed directly from transaction records and explained to business users without requiring complex model interpretation.

2.2.4 K-Means clustering

K-Means partitions observations into k clusters by minimizing the within-cluster sum of squared distances:

$$J = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (2.4)$$

where C_j is cluster j and μ_j is its centroid.

In the SmartCost implementation, RFM variables are standardized before clustering. Standardization is important because Monetary values can have a much larger numerical scale than Recency or Frequency, which would otherwise distort distance-based

clustering [3], [16]. Without scaling, the clustering procedure could be dominated by the monetary dimension.

The implementation uses $k = 4$, $n_init=10$ and $random_state=42$. The fixed random state improves reproducibility. The resulting clusters are interpreted using their RFM profiles rather than assuming that a numerical cluster label has an intrinsic meaning.

2.2.5 Temporal sales analysis

Temporal analysis aggregates transaction records across hours, days, weeks or months. It can identify demand concentration, changes in sales volume and differences between weekdays and weekends. Such analysis can support operational planning, staffing and inventory decisions.

The individual system includes a busiest-hours visualization and a weekend-versus-weekday comparison. The screenshots demonstrate that these outputs are intended to be directly interpretable by an SME user.

2.2.6 Sales trend analysis

Trend analysis attempts to identify systematic movement in a time series. Linear regression provides a simple baseline in which period index t is used to explain revenue y_t :

$$\hat{y}_t = \beta_0 + \beta_1 t \quad (2.5)$$

The slope β_1 indicates the direction and magnitude of the fitted linear trajectory. However, a fitted line is not automatically a reliable forecast. Forecast quality must be tested against observations that were not used during model fitting.

This distinction is central to the revised research methodology. The application visualization is treated as a descriptive and diagnostic component, while the thesis introduces a hold-out forecasting experiment to evaluate predictive performance.

2.3 Related Research

Data mining research provides a foundation for extracting patterns from large transactional datasets. Han, Kamber and Pei describe data mining as a process for discovering useful patterns from data [1]. Witten, Frank, Hall and Pal discuss practical machine learning workflows involving preprocessing, model construction and evaluation [2]. Mitchell provides theoretical foundations for machine learning approaches [3].

Customer analytics research highlights clustering and segmentation as ways of identifying groups with similar behavior [8], [12]. RFM analysis is widely used because it converts transactional history into interpretable customer-level variables [8]. The SmartCost research paper adopts RFM and K-Means as core behavioral analytics techniques.

Association rule mining is another established approach for transactional analysis. Agrawal and Srikant introduced efficient algorithms for mining association rules [7], and later recommender-system research has used relationships between products to support cross-selling. Although association analysis exists in the wider SmartCost research, it is not the primary individual contribution of this thesis and is therefore treated as system context [6], [7], [14].

Forecasting research demonstrates that model choice should be linked to the structure and purpose of the time series [15]. Taylor and Letham discuss scalable forecasting and the importance of practical forecasting workflows. The individual research therefore treats forecasting as an experimentally evaluated process rather than as a visual extension of historical reporting.

2.4 Existing Systems, Products and Techniques

The original SmartCost research compares traditional POS systems with business intelligence platforms and the proposed framework. Traditional POS systems are strong in transaction processing but commonly provide limited behavioral intelligence. Enterprise BI platforms provide rich analytical functionality but can require greater infrastructure, configuration and technical expertise.

The individual research adds a methodological comparison. Four forecasting baselines are considered for the hold-out experiment: linear regression, a naive last-value baseline, a three-period moving-average baseline and simple exponential smoothing. These methods provide different assumptions about the time series. Linear regression emphasizes long-term direction, the naive baseline assumes persistence, moving average smooth short-term fluctuations, and exponential smoothing places greater weight on recent observations.

Table 2.1: Analytical approach comparison used in the research

Approach	Main assumption	Research role
Linear regression	Revenue follows an approximately linear trajectory over the selected periods	Production visualization baseline and forecasting baseline
Naive baseline	The next value is similar to the most recent observed value	Minimal benchmark
Three-period moving average	Recent values provide a smoothed estimate of the next period	Simple smoothing benchmark
Simple exponential smoothing	Recent observations receive greater effective weight	Alternative short-term forecasting approach

2.5 Gap Synthesis and Conceptual Framework

The literature and project evidence indicate that the analytical problem has three connected layers. The first layer is behavioral representation, where raw transactions are converted into RFM features. The second layer is pattern discovery, where customers and sales periods are grouped or aggregated. The third layer is predictive validation, where historical data is separated into training and evaluation periods.

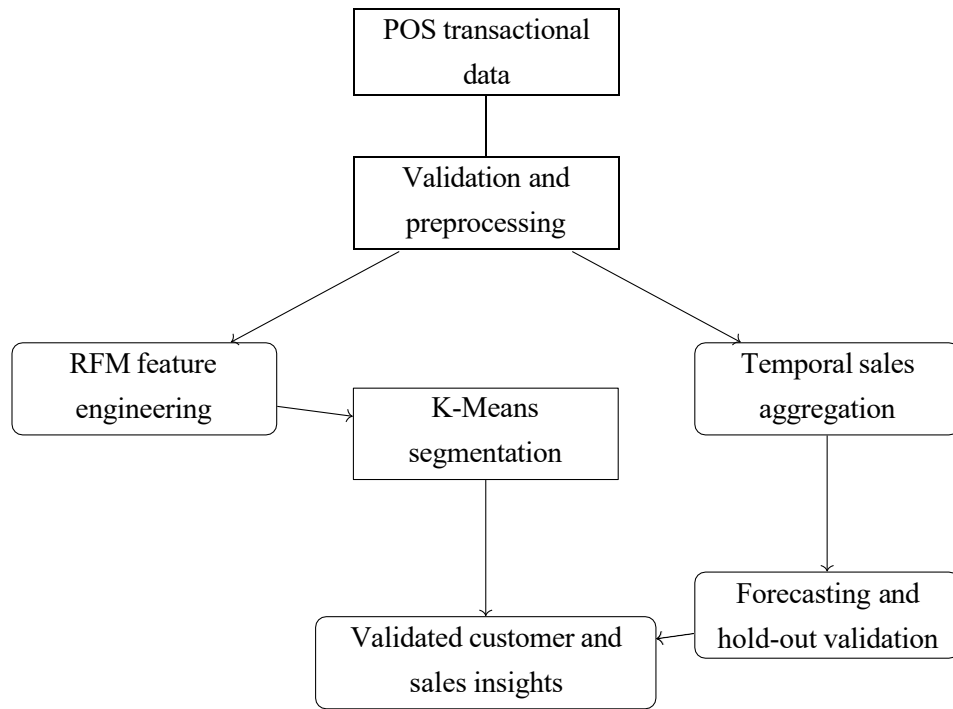


Figure 2.1: Conceptual framework for individual research

The conceptual framework in Figure 2.1 connects data validation, analytical modelling, predictive evaluation and decision-support outputs so that the research contribution can be traced from source records to evidence.

3 RESEARCH METHODOLOGY

3.1 Research Design

The research uses a quantitative, experimental and design-oriented approach. Quantitative analysis is used because the primary evidence consists of transactional records, customer-level aggregates, clustering results, time-series values and user survey responses. Experimental analysis is used because alternative forecasting approaches are evaluated on held-out historical observations. Design-oriented development is used because the research output is implemented as a functional analytical component within a working POS system.

The methodology is organized into the following stages:

1. Define the analytical problem and research questions.
2. Extract and validate the actual group transaction dataset.
3. Preprocess order and order-item records.
4. Construct customer-level RFM features.
5. Standardize RFM features and apply K-Means.
6. Interpret and validate customer clusters.
7. Aggregate sales by temporal period and product.
8. Define a short-term forecasting horizon.
9. Compare forecasting baselines using a historical hold-out period.
10. Integrate validated analytical outputs into the POS interface.
11. Evaluate technical behavior, usefulness and limitations.

3.2 Data, Participants and Study Materials

The primary study material is the actual SmartCost transaction history supplied for the group project. The workbook contains two relevant sheets: *Orders* and *Order Items*. The *Orders* sheet contains 19,253 records and 23 fields. The *Order Items* sheet contains 47,930 records and 9 fields.

The data spans from 1 March 2024 to 27 August 2026. There are 181 unique customer-mobile identifiers in the complete dataset and 12,410 orders associated with an identified customer. The item data contains 20 unique products.

Table 3.1: Primary dataset characteristics

Property	Value
Orders	19,253
Order-item records	47,930
Unique identified customers	181
Customer-linked orders	12,410
Unique products	20
Start date	2024-03-01
End date	2026-08-27
Primary customer identifier	customerMobile
Primary transaction value	totalAmount

A separate user feedback survey contains 52 responses. The survey is used only for user/stakeholder validation of the analytical and decision-support functionality. The survey is not treated as evidence of model accuracy.

3.3 Data Sources and Validation

To strengthen research traceability, the study distinguishes three levels of evidence.

3.3.1 Source-level validation

The raw dataset is checked for the existence of the expected Orders and Order Items sheets and for the expected analytical fields. Date fields are converted to a consistent datetime representation. Customer identifiers are normalized for grouping.

3.3.2 Record-level validation

The research checks for missing identifiers, invalid timestamps, duplicate order identifiers, negative or zero quantities where inappropriate, and numerical consistency between item-level and order-level values. Records that cannot support the intended analysis are excluded from the corresponding analytical calculation rather than silently treated as valid.

Table 3.2: Data-quality validation summary for the supplied project dataset

Validation item	Result	Interpretation
Orders	19,253	Complete order-level population in the supplied workbook
Order-item records	47,930	Item-level observations available for product analysis
Missing order identifiers	0	No missing order IDs detected
Duplicate order identifiers	0	Order-level uniqueness check passed
Missing order timestamps	0	All order timestamps were parseable
Missing total transaction values	0	Revenue field available for all orders
Orders linked to a customer identifier	12,410	64.46% of all orders support customer-level grouping
Unique customer identifiers	181	Customer population available for RFM analysis
Unique products	20	Product dimension available for sales-volume analysis
Non-positive quantities	0	No invalid quantity values detected
Non-positive item totals	0	No invalid line-total values detected
Repeated exact item-row combinations	958	Treated as a data-quality warning; not used to invalidate order-level RFM calculations

The validation results show that the fields required for order-level revenue and customer-level RFM calculations are complete at the order level. Customer identifiers are not available for every order, so customer-level analysis is explicitly restricted to the 12,410 customer-linked orders. Repeated exact item-row combinations are retained as a data-quality warning because an identical item row can arise from source-system duplication; they are not treated as proof of an invalid transaction without additional domain evidence.

3.3.3 Analytical reproduction validation

The strongest validation for the RFM/K-Means component is reproduction. The raw dataset is independently aggregated by customers and transformed into the same three RFM variables used by the Python implementation. The same standardization and K-Means configuration are then applied. Agreement between the independent calculation and the application output is treated as evidence that the implementation is operating on the intended data structure.

The reproduction produced four clusters with sizes of 14, 54, 16 and 97 customers. Their mean monetary values were approximately Rs.55,321, Rs.361,872, Rs.1,010,126 and Rs.111,456, respectively. Using a 30 September 2026 reference time, the reproduced mean Recency values were approximately 167, 52, 47 and 66 days for the same groups. These values correspond closely to the application segments shown in the supplied screenshots.

Table 3.3: Reproduced RFM/K-Means cluster profiles

Cluster interpretation	Customers	Mean Recency	Mean Frequency	Mean Monetary
At Risk/Lapsed	14	167.36	15.64	55,321
Regular	54	52.24	93.70	361,872
VIP/Loyal	16	46.75	267.56	1,010,126
Occasional	97	66.46	29.38	111,456

The application screenshot displays 98 customers in the Occasional group. The raw-data reproduction produced 97. This discrepancy is not treated as an error to be hidden; it is documented as a reproducibility difference that may arise from application-side filtering or a different customer-source snapshot.

3.4 Experimental Procedure

The customer segmentation experiment uses the following procedure.

Algorithm 1 RFM and K-Means customer segmentation

- 1: Load customer-linked orders.
- 2: Parse transaction timestamps.
- 3: Define the latest transaction timestamp as the analysis reference date.
- 4: Group transactions by customer identifier.
- 5: Calculate Recency, Frequency and Monetary values.
- 6: Standardize the three features using StandardScaler.
- 7: Fit K-Means with $k = 4$, $n_init=10$, and $random_state=42$.
- 8: Calculate cluster profiles.
- 9: Rank and interpret clusters using their RFM characteristics.
- 10: Return customer segment labels and RFM profiles.

For sales analysis, transaction values are aggregated by time period. Hourly analysis uses the hour extracted from the transaction timestamp. Daily analysis uses calendar date. Monthly analysis is used for the forecasting experiment because the research requirement requested an explicit forecasting period and because monthly aggregation reduces high-frequency noise for short-term business planning.

3.5 Forecasting Design

The forecasting experiment defines short-term forecasting as a three-month horizon. The purpose is not to claim long-term business forecasting accuracy, but to test whether a small set of transparent forecasting baselines can provide useful near-term estimates from historical POS revenue.

The latest observed period in the dataset is August 2026, but that month is incomplete because the data ends on 27 August. Therefore, the experiment avoids treating August as a fully observed training month when defining completed-period performance.

For historical validation, the monthly sequence from October 2025 through August 2026 is divided into a training segment and a three-month test segment. The first eight periods are used for model fitting, and the final three observed periods are used as a pseudo-future test set. This procedure reproduces the situation of forecasting unseen future months while preserving actual values for evaluation.

3.6 Forecasting Approaches

3.6.1 Linear regression

Linear regression models revenue as a function of period index. It is transparent and computationally lightweight, making it suitable as a baseline in an SME analytics environment.

3.6.2 Naive baseline

The naive method predicts that the next period will have the same revenue as the most recently observed training period. This provides a minimum benchmark against which more complex methods can be evaluated.

3.6.3 Three-period moving average

The moving-average method predicts the next value from the mean of the previous three values. The procedure is repeated recursively for the multi-step forecast.

3.6.4 Simple exponential smoothing

Simple exponential smoothing updates the estimated level using recent observations. It is useful when recent history should have greater influence than older observations while no explicit seasonal pattern is modelled.

3.7 Measures and Evaluation Metrics

For customer segmentation, evaluation focuses on reproducibility, cluster size, cluster profile separation and business interpretability. The research does not treat a cluster number as meaningful by itself; the segment interpretation is based on the RFM profile.

For forecasting, the primary metrics are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (3.1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (3.2)$$

$$MAPE = \frac{100}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \quad (3.3)$$

Lower values indicate smaller prediction errors. Because the number of monthly observations is limited, the forecasting experiment is treated as a validation study rather than a claim of universal model superiority.

3.8 Data Analysis

The research analysis consists of four connected analytical views:

1. **Customer-level analysis:** RFM values and K-Means clusters.
2. **Temporal analysis:** hourly, weekday/weekend and monthly revenue behavior.
3. **Product analysis:** product quantities and revenue contributions.
4. **Predictive analysis:** regression and alternative forecasting methods on monthly revenue.

The interface combines these outputs into a customer behavior and sales insights screen. This allows research to connect quantitative analysis with decision-support presentation.

3.9 Ethics, Privacy, Security and Responsible AI

The dataset contains customer-linked identifiers. Customer identifiers are used for analytical grouping but should not be exposed in the thesis in identifiable form. The thesis therefore reports aggregate customer segment results rather than publishing individual customer names or mobile numbers.

The analytical models are used for business decision support, not for automated high-stakes decisions about individuals. Segment labels such as VIP/Loyal or At Risk/Lapsed are analytical categories derived from transaction behavior and should not be interpreted as personal judgments about customers.

The forecasting results are presented with uncertainty and validation limitations. A model output is not treated as a guaranteed future result. Users remain responsible for

business decisions made using analytical information.

3.10 Reproducibility and Research Artefacts

The research artefacts include the SmartCost transaction workbook, the Python analytical source code, the application screenshots, the QA report, the user feedback survey and the project-management evidence. The Python implementation fixes the K-Means random state at 42 and uses a fixed number of initializations, which supports repeatable clustering.

4 PRODUCT AND SOLUTION DESIGN

4.1 Requirements

The functional requirements of the individual component are derived from the research problem and the actual application interface.

Table 4.1: Functional requirements for the individual component

ID	Requirement	Description
FR01	Customer identification	Identify transactions that can be associated with a customer identifier.
FR02	RFM calculation	Calculate Recency, Frequency and Monetary features.
FR03	Customer segmentation	Apply K-Means clustering to standardized RFM features.
FR04	Segment presentation	Present interpretable segment labels and customer counts.
FR05	Temporal analysis	Aggregate sales by hour, day and period.
FR06	Product analysis	Identify high- and low-volume products.
FR07	Trend analysis	Display revenue trajectories across periods.
FR08	Forecast evaluation	Compare forecasting approaches using a historical hold-out.
FR09	Insight generation	Convert analytical results into concise business statements.
FR10	Report export	Allow selected analytical sections to be included in an A4 report.

4.1.1 Non-functional requirements

The component should be reproducible, understandable, responsive enough for normal SME use and capable of operating with a relatively lightweight analytical stack. The output should be understandable to non-technical business users. The analytical service should expose stable API behavior and separate analytical processing from the mobile user interface.

4.2 Component Architecture Within the Complete Solution

The overall architecture described in the research paper follows a React Native to Node.js to FastAPI to MongoDB communication flow [17][20]. The mobile interface requests analytics from the operational backend, which communicates with the Python analytical service. MongoDB provides transactional data sources.

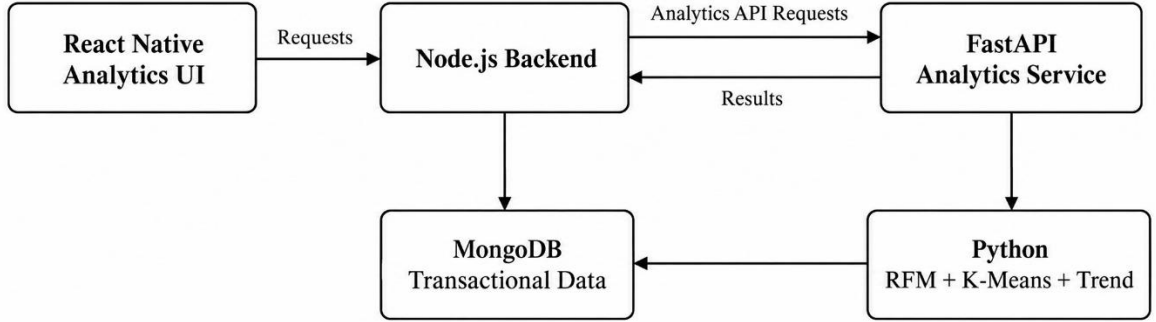


Figure 4.1: Architecture of the individual analytics component within SmartCost

4.3 Design Alternatives and Decisions

4.3.1 Customer segmentation alternatives

Possible customer segmentation approaches include rule-based segmentation, RFM scoring with fixed thresholds, hierarchical clustering and K-Means clustering. A rule based system is easy to explain but can be sensitive to manually selected thresholds. Hierarchical clustering can reveal nested structure but can become more computationally expensive and less convenient for repeated operational segmentation. K-Means was selected for the implemented component because it is computationally lightweight, widely studied, and straightforward to integrate with standardized numerical RFM features.

The research nevertheless evaluates the segmentation output rather than assuming that K-Means is universally optimal. Cluster profiles are inspected and the reproducibility procedure is documented.

4.3.2 Sales forecasting alternatives

The implemented interface uses a linear regression trajectory. The research evaluates this approach against naive, moving-average and simple exponential-smoothing baselines. The comparison is deliberately transparent so that the thesis can explain why a particular method is suitable for the available data.

4.3.3 Database and service architecture

MongoDB is used by the wider system because the POS environment produces transactional and semi-structured records. Node.js handles operational API routing while FastAPI provides an analytical service boundary. Python provides access to data-processing and machine-learning libraries. This separation keeps the analytical code modular and allows the analytical service to be evaluated independently.

4.4 Detailed Design

4.4.1 RFM feature pipeline

The feature pipeline first selects customer-linked transactions. Timestamps are parsed and ordered. For each customer, the most recent transaction determines Recency, the number of distinct orders determines Frequency, and the sum of transaction value determines Monetary.

The resulting feature matrix is standardized before K-Means. Standardization transforms each feature approximately as:

$$z = \frac{x - \mu}{\sigma} \quad (4.1)$$

where μ and σ are the feature mean and standard deviation.

4.4.2 Cluster interpretation

Cluster interpretation is performed after K-Means rather than before it. This avoids assigning a business meaning to arbitrary numerical labels. The cluster profiles are compared by Recency, Frequency and Monetary values. High frequency and high monetary contribution combined with low recency are interpreted as high-value active behavior, while high recency and low monetary contribution indicate an at-risk or lapsed profile.

4.4.3 Sales trend pipeline

The sales trend pipeline extracts transaction periods, aggregates revenue and constructs a chronological series. The interface displays period revenue, a rolling average and a target baseline. The research layer additionally creates a train/test split for forecasting evaluation.

4.4.4 Insight generation

The user interface converts numerical outputs into statements such as peak-period observations, product performance statements and customer segment summaries. Insight generation is intended to reduce the interpretation burden on SME users. However, the thesis treats generated statements as derived from analytical measurements rather than as independent facts.

4.5 Standards and Professional Practice

The implementation follows modular API design, input validation, separation of analytical logic from presentation, fixed model configuration for reproducibility and privacy-conscious reporting. The thesis also follows the SLIIT documentation requirements for headings, margins, pagination, figures, tables and references.

5 IMPLEMENTATION, INTEGRATION AND TESTING

5.1 Development Environment and Technology Stack

The analytical component uses Python and FastAPI. Scikit-learn provides K-Means and StandardScaler. The wider application uses Node.js, MongoDB and React Native. The research paper identifies pandas and NumPy for analytical processing.

Table 5.1: Technology stack used by the individual component

Technology	Role
Python	Analytical processing and machine learning
FastAPI	REST API layer for analytical functionality
scikit-learn	StandardScaler and K-Means clustering
pandas	Tabular data transformation and aggregation
NumPy	Numerical computation
MongoDB	Transactional data storage in the wider system
Node.js	Operational backend and API routing
React Native	Mobile analytics interface

5.2 Component Implementation

5.2.1 FastAPI segmentation endpoint

The Python implementation exposes a segmentation endpoint through FastAPI. The router accepts the required request structure and returns customer segment information. The segmentation service requires a sufficient number of customers with orders before clustering is attempted. The supplied implementation requires at least 12 customers with orders and uses four clusters by default.

5.2.2 RFM feature construction

The implementation derives three numerical features for each customer:

- Recency: days since the customer's last update/transaction.
- Frequency: total number of orders.
- Monetary: total sales attributed to the customer.

The three features are standardized using StandardScaler. This is important because the monetary feature is measured in currency and may be much larger numerically than the other variables.

5.2.3 K-Means configuration

The implementation uses:

- Number of clusters: $k = 4$
- Number of initializations: 10
- Random state: 42
- Standardized RFM features

The cluster labels are subsequently mapped to business-oriented labels. The mapping is based on the ranked cluster profiles rather than the raw numerical label produced by K-Means.

5.2.4 Sales trend implementation

The application presents a sales trend and revenue trajectory component based on linear regression across periods. The supplied screenshot displays 11 periods from October through August. It also displays a rolling average and target baseline. Visualization allows users to tap a period and inspect detailed sales and order information.



Figure 5.1: Implemented sales trend and revenue trajectory interface

5.3 Integration With Other Components

The analytical component is integrated into the SmartCost mobile application through the wider Node.js and FastAPI architecture. The mobile interface provides the user facing controls, while the backend and analytical service provide data and analytical calculations.

The component is not an isolated research prototype. It is connected to operational transaction data and presents its results alongside other POS capabilities. This integration is important for commercialization because an SME user can access analytics within the same application used for daily operations.

5.4 Verification and Software Testing

The group QA report contains 42 recorded issues across the wider SmartCost application. The issue directly relevant to the analytics interface is BUG-031, which concerns

cropping of the lower section of the Analytics screen on both Android and iOS. It was classified as a medium-severity P2 UI/UX issue and was recorded as open.

Table 5.2: Relevant QA evidence for the individual analytics component

Bug ID	Area	Issue	Status
BUG-031	Analytics screen	Lower section of analytics content is cropped, limiting visibility of complete content.	Open

This issue is reported honestly because it is relevant to the implemented analytics screen. The QA report does not provide independent algorithm-level test cases for RFM or K-Means, so this thesis does not claim that the QA team independently validated the clustering algorithm. Algorithm-level validation is instead addressed through the raw-data reproduction experiment described in Chapter 3 and Chapter 6.

5.4.1 Analytical verification test cases

Table 5.3: Analytical verification test cases

ID	Test	Expected validation	Evidence
T01	Load customer-linked transactions	Customer records can be grouped by identifier	Dataset reproduction
T02	Calculate RFM	Every eligible customer receives R, F and M values	RFM table
T03	Standardize RFM	Features are transformed before K-Means	Source code
T04	K-Means execution	Four clusters are produced with fixed configuration	Reproduction
T05	Cluster profiling	Cluster statistics are interpretable	Cluster table
T06	Monthly aggregation	Chronological revenue series is produced	Sales data
T07	Forecast hold-out	Unseen historical periods can be predicted	Forecast metrics
T08	UI analytics rendering	Analytical outputs are visible in the interface	Screenshots

5.5 Deployment and Operation

The analytics service is intended to operate as a modular service accessed by the wider Smart-Cost application. The separation of analytical logic from the mobile interface makes it possible to modify analytical procedures without redesigning the entire POS interface.

Operationally, the system should validate input, handle missing customer identifiers appropriately and avoid exposing customer personal information in analytical reports. The analytical report feature allows users to select a time frame and select which analytical sections should be included.

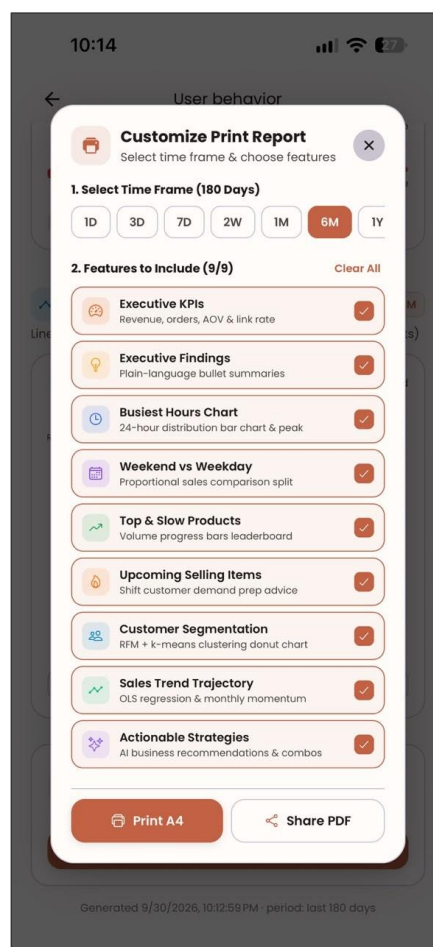


Figure 5.2: Implemented customizable customer behavior report interface

6 RESULTS AND VALIDATION

6.1 Research Results

The full project dataset contains 19,253 orders and 47,930 order-item records. The dataset spans March 2024 to August 2026. There are 181 unique customer-mobile identifiers, of which 12,410 order records are linked to an identified customer. This corresponds to 64.46% of the supplied order records.

These characteristics provide substantially more temporal coverage than the 12-month simulated dataset described in the original research paper. The extended dataset therefore provides a stronger empirical basis for validating the individual component.

6.1.1 RFM and K-Means results

The independent raw-data reproduction produced the cluster profiles shown in Table 3.3. The 16-customer high-value cluster has a mean monetary contribution of approximately Rs.1.01 million per customer and a mean frequency of approximately 268 orders. The 14-customer At Risk/Lapsed cluster has the highest mean recency, approximately 167 days, and the lowest mean monetary value.

Table 6.1: Customer segment results from raw-data reproduction

Segment	Customer Count	Mean Recency	Mean Frequency	Mean Monetary
At Risk/Lapsed	14	167.36	15.64	55,321
Regular	54	52.24	93.70	361,872
VIP/Loyal	16	46.75	267.56	1,010,126
Occasional	97	66.46	29.38	111,456

The results indicate that customer count alone does not represent revenue contribution. The VIP/Loyal group contains fewer customers than the Regular and Occasional groups but contributes approximately one-third of the revenue represented in the customer-level RFM calculation. The Regular group contributes the largest revenue share in the raw-data reproduction.



Figure 6.1: Implemented RFM and K-Means customer segmentation interface

6.1.2 At-risk customer result

The application interface reports 14 customers as At Risk/Lapsed and indicates an average of 167 days since their last order. This result is consistent with the raw-data observation that the corresponding cluster has substantially higher recency than the other groups. The exact application display and raw-data reference date are treated separately because application-side filters can differ from a full-dataset reproduction.

6.1.3 Peak-hour and temporal results

The supplied application screenshot for the six-month interface reports a peak period of 11 AM–12 PM, with 234 orders and Rs.933,360 in sales. The interface reports that this period is 10.34% higher than other hours in the selected comparison.

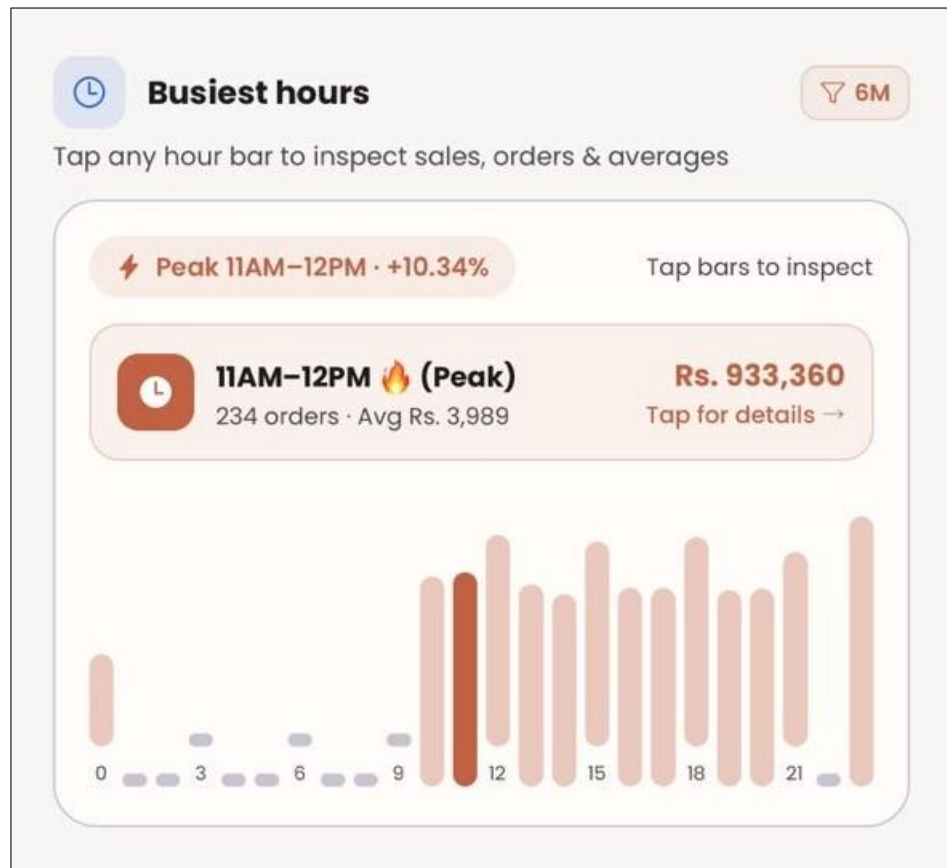


Figure 6.2: Implemented busiest-hours analysis

The raw dataset contains timestamps that support independent temporal analysis. Differences between raw-data precomputation and displayed interface values can arise when the application applies a rolling 180-day filter, uses a particular transaction-status rule or uses a different reference timestamp. The thesis therefore distinguishes between the displayed application result and the raw-data analytical reproduction rather than treating them as identical without verification.

6.1.4 Weekend versus weekday results

The application reports average sales of Rs.3,753 for weekdays and Rs.3,813 for weekends, describing weekend average sales as 1.6% higher. This illustrates how temporal analytics can be converted into an interpretable business statement.

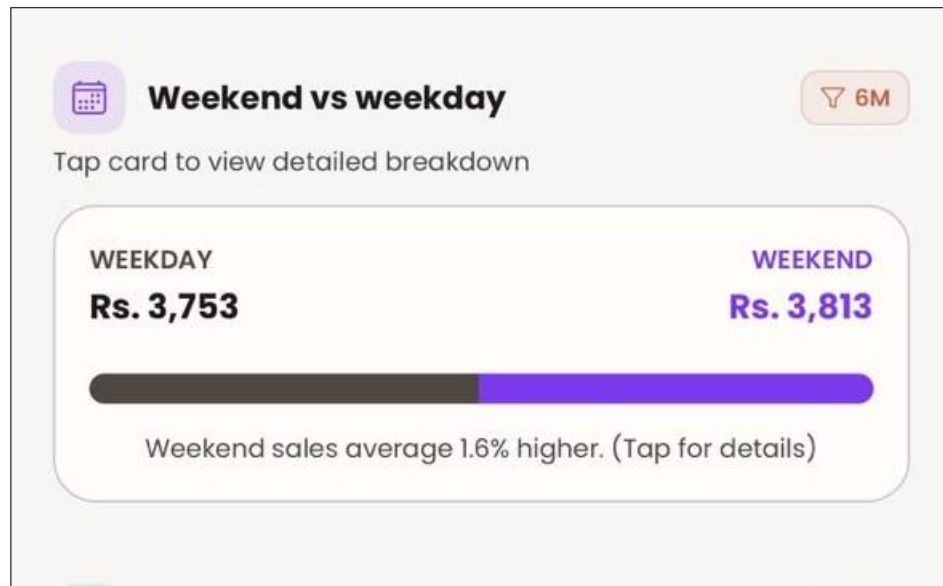


Figure 6.3: Implemented weekend-versus-weekday comparison

6.1.5 Product results

The application screenshot identifies Chicken Kottu as the highest-selling product in the selected period with 1,461 units. The other displayed top products are Fresh Lime Juice, Chicken Rice & Curry, Chicken Fried Rice and Soft Drink.

The slow-moving product view displays Crab Curry with 256 units, Cuttlefish Curry with 290, Grilled Seer Fish with 416, Chocolate Biscuit Pudding with 522 and Egg Rice & Curry with 583 units.

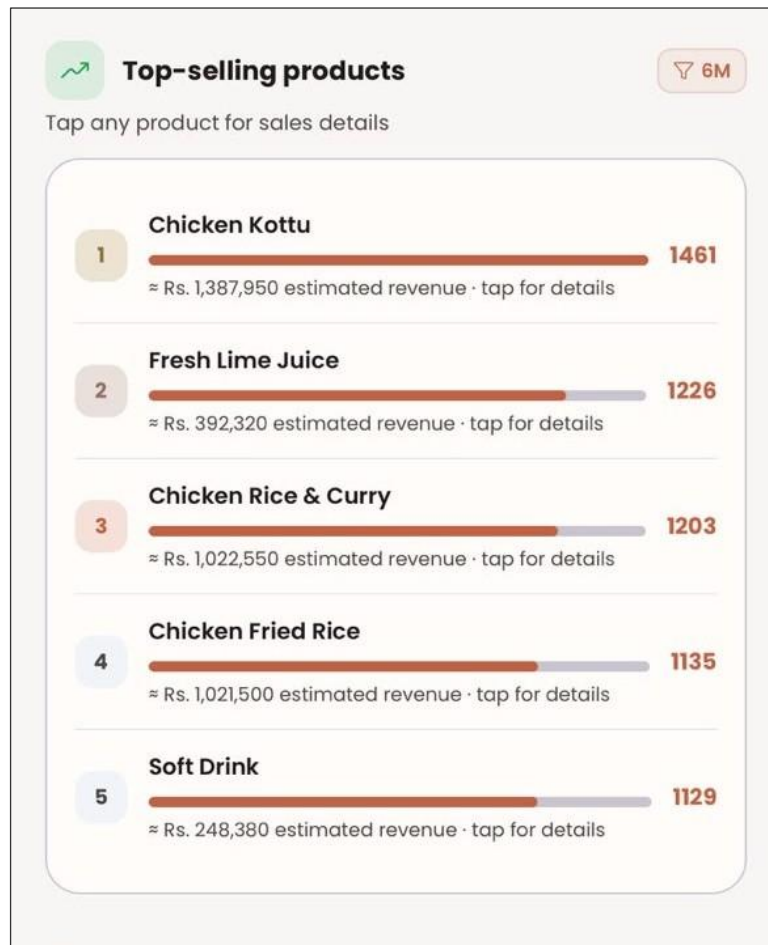


Figure 6.4: Implemented top-selling product analysis

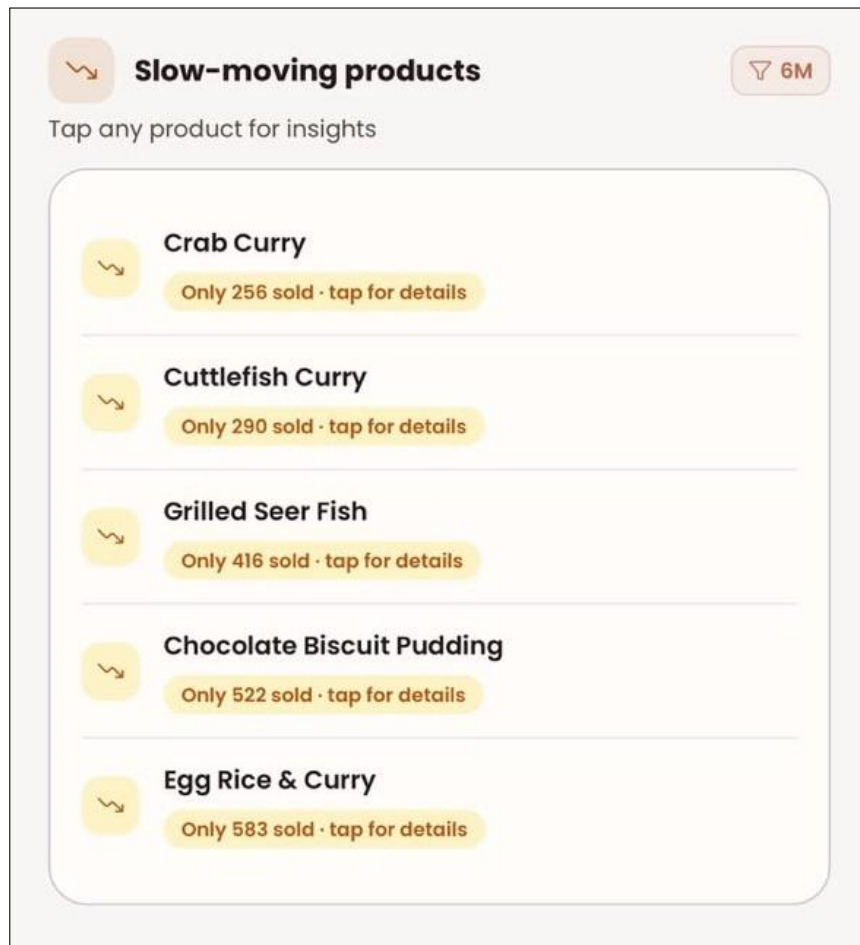


Figure 6.5: Implemented slow-moving product analysis

These results demonstrate the practical connection between customer behavior analytics and inventory planning. A slow-moving product may require different stock or promotional decisions from a high-volume product. However, quantity sold alone should not be treated as a complete profitability measure because margin, spoilage, cost and price are not represented by volume alone.

6.1.6 Sales trend results

The application displays an 11-period sales trend from October through August. The screenshot labels the visualization as a linear regression trajectory and presents period revenue, a rolling average and a target baseline.



Figure 6.6: Implemented sales trend and revenue trajectory

An independent regression calculation on the October 2025–August 2026 monthly revenue sequence gives a slope of approximately -Rs.47,750 per month and an in-sample R^2 of approximately 0.119. The low R^2 indicates that a simple straight-line relationship explains only a small portion of the month-to-month variation in this particular 11-period series. This is an important research finding because a visually smooth trend line should not automatically be interpreted as a strong predictive model.

The monthly values used for this calculation are shown in Table 6.2.

Figure 6.7 provides an independently generated visualization of the same eleven-period sequence and its fitted linear trend. It is used as a research figure rather than a replacement for the application screenshot.

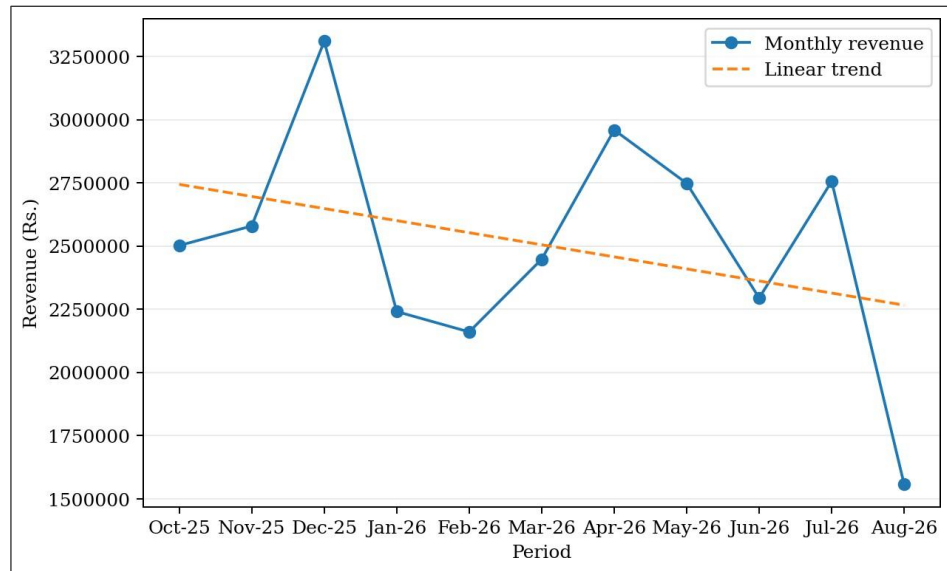


Figure 6.7: Monthly revenue observations and fitted linear trend

Table 6.2: Monthly revenue used in the 11-period trend analysis

Period	Revenue (Rs.)
October 2025	2,502,050
November 2025	2,579,170
December 2025	3,309,990
January 2026	2,240,830
February 2026	2,160,400
March 2026	2,447,880
April 2026	2,959,040
May 2026	2,747,900
June 2026	2,294,680
July 2026	2,755,690
August 2026	1,556,960

August is incomplete in the raw dataset and is therefore a limitation for direct monthly comparison. The research avoids treating the August value as a completed-month observation.

6.2 Product and Technical Results

The implemented component successfully combines customer segmentation, temporal analysis, product ranking, slow-moving product identification, trend visualization and report generation within the SmartCost interface.

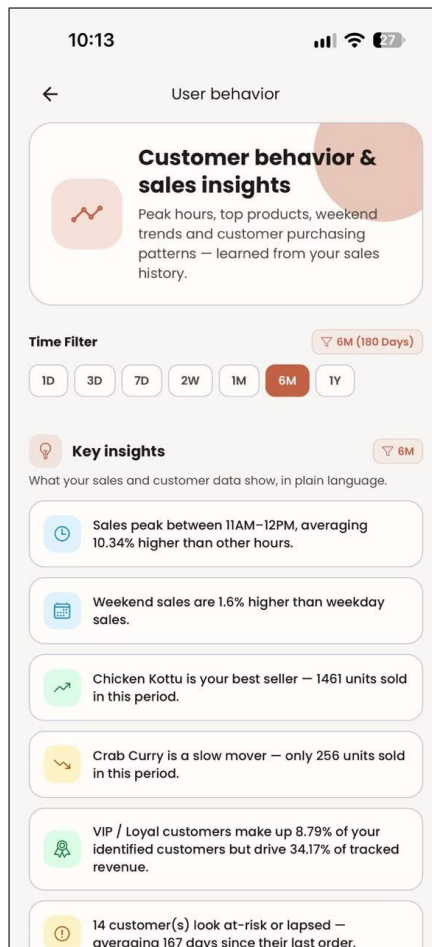


Figure 6.8: Implemented customer behavior and sales insights interface

The customer behavior screen provides a six-month filter, key insights and multiple analytical cards. This design presents analytical outputs in business-oriented terms so that users do not need statistical training to interpret the main findings.

The customer segmentation screen uses a chart and segment cards. The application labels the method as “RFM + K-means segmentation” and identifies four customer groups. Independent raw-data reproduction yielded 16, 54, 97 and 14 customers, while the supplied interface snapshot displays 16, 54, 98 and 14. The small difference is retained as a reproducibility observation because the application view may use a different filter or reference timestamp.

The report customization interface allows users to include nine analytical sections: Executive KPIs, Executive Findings, Busiest Hours Chart, Weekend vs Weekday, Top & Slow Products, Upcoming Selling Items, Customer Segmentation, Sales Trend Trajectory and Actionable Strategies.

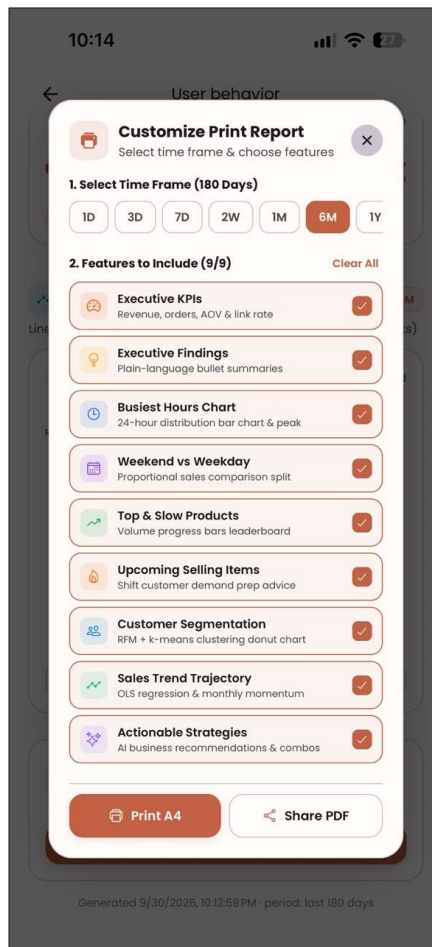


Figure 6.9: Implemented report feature selection

6.3 Forecasting Validation Results

The forecasting experiment uses October 2025 through May 2026 as the training period and June through August 2026 as the historical hold-out period. The three test months therefore act as a pseudo-future period.

Table 6.3: Historical hold-out forecasting comparison

Method	MAE (Rs.)	RMSE (Rs.)	MAPE
Linear Regression	528,297	691,578	30.65%
Naive baseline	550,650	735,710	32.17%
Three-period moving average	559,180	736,019	32.51%
Simple exponential smoothing	507,393	645,400	29.08%

Figure 6.10 visualizes the historical hold-out forecasts against the observed June–August 2026 values.

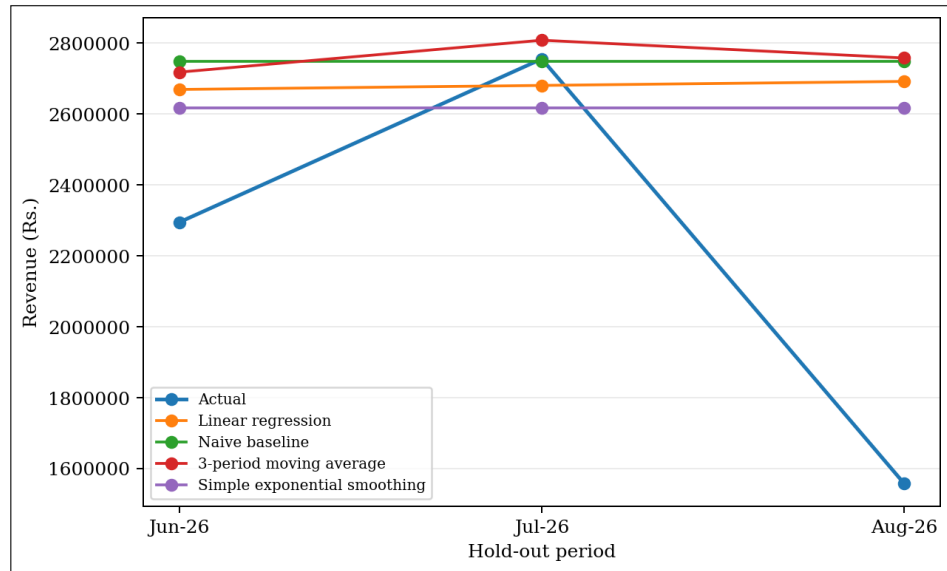


Figure 6.10: Historical hold-out forecast comparison

In this particular historical hold-out experiment, simple exponential smoothing produced the lowest MAE, RMSE and MAPE among the four tested approaches. The result should not be generalized beyond this dataset because only three monthly periods were used for the hold-out evaluation and the underlying series contains substantial variation.

The important research contribution is therefore not that one forecasting method is universally superior. Rather, the experiment demonstrates that forecasting quality can differ from visual trend quality and that model evaluation should be performed against unseen historical observations.

The production interface currently displays a linear regression trajectory. The experimental comparison indicates that the analytical layer could be extended to expose an evaluated short-term forecasting model while retaining linear regression as a transparent trend visualization.

6.3.1 Forward projection after the most recent complete month

Because August 2026 is an incomplete month in the supplied data, July 2026 is treated as the latest complete monthly observation for a forward-looking illustration. Linear regression and simple exponential smoothing were fitted using October 2025 through July 2026 and projected three months ahead. These estimates for August, September and October 2026 are model outputs and are not presented as validated observations.

Table 6.4: Forward revenue projection from the latest complete month

Method	Aug 2026	Sep 2026	Oct 2026
Linear regression	2,597,196	2,596,729	2,596,263
Simple exponential smoothing	2,599,721	2,599,721	2,599,721

The two projections are close to one another because the complete-month training series from October 2025 to July 2026 has a relatively flat fitted trajectory. This exercise demonstrates the distinction between future prediction and retrospective trend visualization. It also shows why future estimates should be accompanied by a clearly stated cutoff date and an explicit indication that ground truth is not yet available.

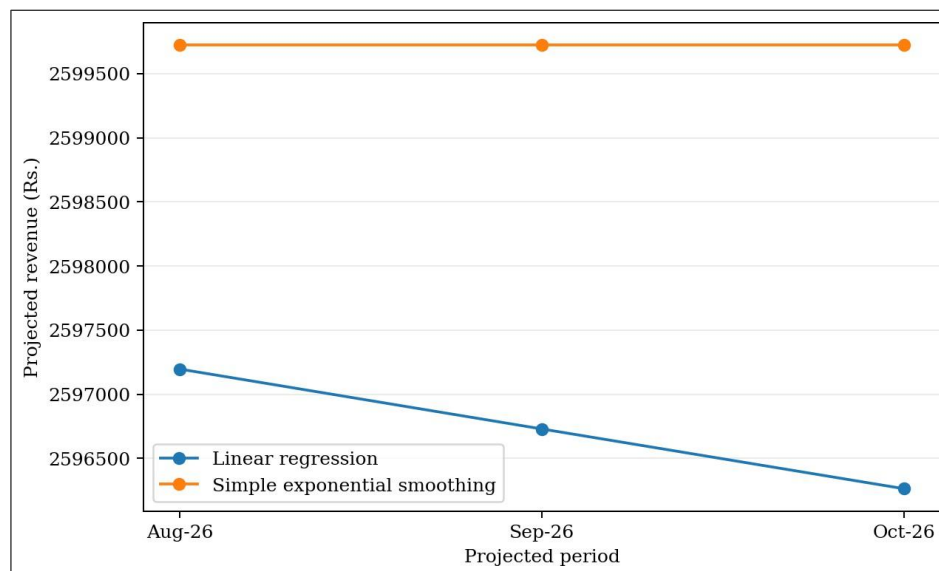


Figure 6.11: Forward revenue projection after July 2026

6.4 User or Stakeholder Validation

The user feedback survey contains 52 responses. The analysis below focuses only on questions relevant to the individual component.

Table 6.5: User evaluation results for analytics-related questions

Statement	Positive	Positive Rate
Sales reports and analytics help me understand business performance.	41/52	78.8%
Product demand predictions are useful for planning future stock or production.	45/52	86.5%
Business performance/profit predictions provide useful decision information.	40/52	76.9%
Customer and sales insights help me understand sales patterns.	44/52	84.6%
Alerts and predictive insights can help identify potential problems earlier.	37/52	71.2%
AI-based features help me make better business decisions.	44/52	84.6%

The results indicate generally positive user perceptions of analytics-related functionality. In particular, 84.6% of respondents agreed or strongly agreed that customer and sales insights help them understand sales patterns. The same positive-response rate was recorded for the statement that AI-based features help users make better business decisions.

For overall satisfaction, 45 of the 52 respondents selected a rating of 4 or 5, corresponding to 86.5%. The mean satisfaction score was 4.27 on the five-point scale.

Regarding comparison with previous/manual methods, 44 respondents (84.6%) reported that SmartCost made business management either slightly or significantly easier. For recommendation intention, 46 respondents (88.5%) selected either “Definitely” or “Probably”.

These results support usability and perceived usefulness but do not establish causal business impact. The survey is a perception-based evaluation and should therefore be interpreted alongside the analytical and technical validation.

6.5 Integration Results

The analytical outputs are integrated into the customer behavior screen and report generation workflow. The screenshots show that the system can combine customer segmentation, temporal analytics and product performance into a single report-selection interface.

The integration also demonstrates the separation between computation and presentation. The Python service performs clustering, while the mobile interface displays the results in business-oriented language. This supports maintainability and allows the analytical procedure to be evaluated independently.

6.6 Summary of Findings

The principal findings are:

1. The actual group dataset provides a substantially larger and longer historical basis than the simulated dataset described in the original paper.
2. RFM and K-Means reproduce the principal customer segmentation structure shown by the application, with a small one-customer discrepancy that is documented as a reproducibility issue.
3. Customer value is unevenly distributed; the VIP/Loyal group represents 16 customers but approximately 34.18% of the reproduced customer-level revenue.
4. The application identifies clear peak-hour, weekend/weekday and product-level patterns.
5. The 11-period linear regression trajectory has a low in-sample R^2 , demonstrating why visual trend lines should not be treated as sufficient evidence of predictive accuracy.
6. Historical hold-out forecasting shows that simple exponential smoothing produced lower error than the other tested baselines for the selected test period.
7. User feedback provides positive evidence for the usefulness of customer and sales insights, with 84.6% positive responses to the customer-and-sales-pattern statement.

7 DISCUSSION

7.1 Interpretation of Findings

The results demonstrate that customer behavior analytics can provide a more structured view of POS activity than transaction totals alone. RFM features transform a sequence of individual purchases into a customer-level behavioral representation. K-Means then identifies groups with different purchasing intensity and monetary contribution.

The reproduced cluster profiles show a clear distinction between active high-value customers and customers with longer periods since their last transaction. The At Risk/Lapsed group has a substantially higher mean Recency than the other groups. This makes the segment operationally interpretable because a business can investigate retention or reactivation strategies without requiring the user to understand the mathematics of clustering.

The results also demonstrate why customer count should not be used as the only measure of customer value. The VIP/Loyal group contains only 16 customers but contributes approximately 34.18% of the customer-level revenue in the reproduced dataset. The Regular segment contains more customers and contributes a larger share overall. This indicates that segment-specific strategies may be more informative than a single customer-average metric.

7.2 Comparison With Literature Baselines and Existing Solutions

The research paper identifies traditional POS systems as primarily transaction-oriented and describes a gap in integrated behavioral analytics. The implemented SmartCost component addresses this gap by bringing customer segmentation, temporal sales analysis and trend visualization into the same interface.

The comparison should not be interpreted as a claim that all commercial POS systems lack analytics. Modern commercial systems vary substantially in their capabilities. The relevant distinction is that the research investigates an integrated analytical workflow specifically designed for an SME-oriented POS architecture.

The forecasting experiment also strengthens the comparison because it moves beyond descriptive charts. A traditional reporting screen may display monthly sales history, while the research procedure explicitly separates training and evaluation periods and calculates forecast errors.

7.3 Research Contribution

The individual research contribution has four parts.

First, it provides a reproducible customer segmentation pipeline based on RFM and K-Means. The use of a real multi-period transaction dataset allows the method to be evaluated

against operational data rather than only a small synthetic demonstration.

Second, it provides a validated sales trend analysis workflow. The research distinguishes a descriptive regression trajectory from predictive forecasting and demonstrates that the two should not be treated as equivalent.

Third, it introduces an explicit historical hold-out evaluation and compares several lightweight forecasting baselines. This provides the required methodological comparison and experimental validation.

Fourth, it connects the analytical results to an integrated SME POS interface. The research therefore links methodological analysis with practical decision support.

7.4 Product Contribution

The product contribution is the implementation of analytics functionality that can be consumed by an SME user without direct interaction with the underlying statistical algorithms. The interface provides segment cards, peak-hour information, weekend/weekday comparisons, product rankings, slow-moving product information and sales trend visualization.

The report-generation feature further allows users to select which analytical sections to include in an A4 report. This supports operational communication because analytical outputs can be shared or printed rather than remaining only inside the application.

7.5 Limitations and Threats to Validity

Several limitations affect the interpretation of the results.

The first is dataset representativeness. The transaction history belongs to the SmartCost project environment and cannot be assumed to represent all SMEs, restaurants or retail businesses.

The second is customer identification. Only 181 unique customer-mobile identifiers are available in the full dataset. Orders without an identified customer cannot contribute to customer-level RFM segmentation.

The third is temporal completeness. The dataset ends on 27 August 2026, making August an incomplete monthly period. This can distort monthly comparisons and trend estimation.

The fourth is the small forecasting evaluation set. Only three monthly periods are used for the historical hold-out. The resulting error comparison is useful as an experiment but insufficient to establish general forecasting superiority.

The fifth is cluster sensitivity. K-Means requires a chosen number of clusters and can be affected by feature scaling, initialization and the underlying distribution of the data. The implementation fixes $k = 4$ and the random state, but future work should evaluate alternative cluster counts and stability measures.

The sixth is the application-versus-raw-data discrepancy. The application screenshot reports 98 Occasional customers while the complete raw-data reproduction produces 97. This suggests

that the application may apply a different filter or data snapshot. The discrepancy is small but demonstrates the importance of versioned analytical data and reproducibility.

7.6 Implications

For SME decision support, the findings imply that customer segmentation should be combined with sales analysis rather than treated as an isolated machine-learning feature. A segment is more useful when it can be related to time periods, products and revenue.

For system design, the findings imply that transparency is important. Business users can interpret labels such as VIP/Loyal or At Risk/Lapsed more easily than raw cluster numbers. Similarly, forecasting output should include the historical period, model name and error information where appropriate.

For research methodology, the findings demonstrate that implementation screenshots alone are insufficient evidence. Reproducible data transformations, held-out evaluation and comparison with alternative approaches provide stronger evidence of research contribution.

8 COMMERCIALISATION AND SUSTAINABILITY

8.1 Value Proposition and Target Users

The target users are small and medium-sized businesses that already use or intend to use a POS system but require more accessible analytics. The value proposition is the integration of transaction processing and business analytics in a single lightweight platform.

For business owners, the customer segmentation component can support retention-oriented planning. For operations managers, peak-hour and product analysis can support staffing and inventory decisions. For management users, the sales trend and forecasting views can support short-term planning.

8.2 Market and Competitive Position

The competitive context contains three broad categories: traditional POS systems, enterprise business intelligence tools and specialized analytics products.

Traditional POS products typically provide strong transaction processing but vary in the depth of customer and sales analytics. Enterprise BI tools provide extensive analytical functionality but may require additional configuration, data integration and technical expertise. SmartCost is positioned as an integrated SME-oriented system that combines operational POS functionality with embedded analytics.

The research does not claim that SmartCost is universally superior to commercial alternatives. Its proposed position is instead based on integration, accessibility, lightweight deployment and domain-specific analytical workflows.

8.3 Business or Sustainability Model

A potential sustainability model is a software-as-a-service or subscription model in which SMEs pay for access to the POS and analytical functions. Alternative models include onetime deployment with support fees, branch-based subscriptions or feature-based tiers.

The analytics component can be offered as part of the core POS subscription because its value depends on the availability of transaction history. A premium analytics tier could provide extended forecasting, historical comparisons and automated reports.

8.4 Scalability, Intellectual Property and Compliance

The modular architecture supports scalability by separating the mobile interface, operational backend, analytical service and database. Additional analytical models can be introduced without replacing the POS interface.

Customer data should be handled according to applicable privacy and data protection requirements. Customer identifiers should be minimized in reports and protected in storage and transmission. Analytical results should be aggregated where individual-level information is not necessary.

The source code, UI design and analytical workflow may be protected as project intellectual property according to the ownership rules applicable to the university project. Third-party libraries should remain subject to their respective licenses.

8.5 Contribution to the Sustainable Development Goals

The research can contribute indirectly to Sustainable Development Goal 8, Decent Work and Economic Growth, by supporting productivity and data-driven management in small businesses. It can also support responsible digitalization by enabling businesses to make more informed use of operational data.

The contribution should be interpreted as an alignment rather than as a direct measurement of SDG impact. A future longitudinal study would be required to quantify sustained business outcomes.

9 CONCLUSION AND FUTURE DIRECTIONS

9.1 Conclusion

This thesis investigated customer behavior analysis and sales trend forecasting within the SmartCost SME POS environment. The research addressed a gap between transaction recording and validated analytical decision support.

The study used the actual SmartCost transaction history containing 19,253 orders and 47,930 order-item records across March 2024 to August 2026. Customer-linked transactions were transformed into RFM features and segmented using the implemented K-Means configuration. The raw-data reproduction demonstrated that the application segmentation is substantially reproducible and identified four distinct behavioral profiles.

The research also analyzed temporal sales behavior, product performance and revenue trajectories. The implemented interface presents these results through customer segments, peak-hour analysis, weekend/weekday comparison, top and slow-moving products and sales trend visualizations.

A key improvement over the original research framing was the explicit forecasting validation. The research defined a short-term three-month horizon and evaluated several lightweight approaches on a historical hold-out. Simple exponential smoothing produced the lowest error in the selected experiment, although the limited number of test periods prevents broad generalization.

The user evaluation of 52 responses provided additional evidence that users perceive sales analytics and customer/sales insights as useful. However, user perception does not replace objective model validation.

Overall, the research demonstrates that an SME POS platform can combine interpretable customer segmentation, temporal sales analysis and forward-looking forecasting evaluation in a lightweight analytical architecture.

9.2 Achievement of Objectives

Objective O1 was addressed by validating the actual group dataset, documenting its structure and identifying customer and product coverage.

Objective O2 was addressed by deriving Recency, Frequency and Monetary values from customer-linked transactions.

Objective O3 was addressed through standardized RFM features and K-Means clustering with a fixed configuration.

Objective O4 was addressed through hourly, weekday/weekend, monthly and product-level analysis.

Objective O5 was addressed by defining a three-month short-term forecasting horizon and comparing linear regression, naive, moving-average and exponential-smoothing approaches.

Objective O6 was addressed through MAE, RMSE and MAPE evaluation on a historical hold-out period.

Objective O7 was addressed through integration into the SmartCost customer behavior interface and report workflow.

Objective O8 was addressed through user feedback, QA evidence, raw-data reproduction and comparison with transaction-focused POS functionality.

9.3 Individual Research and Product Contributions

The individual research contribution is the validated analytical workflow connecting customer behavior segmentation with sales trend analysis and short-term forecasting evaluation.

The individual product contribution is the implemented customer behavior interface and Python analytical service that present RFM/K-Means results, temporal sales information, product performance and trend information within the SmartCost platform.

The research contribution is therefore distinct from the general development of a POS application. It lies in the transformation and validation of transactional data, the analytical comparison of methods and the experimental evaluation of predictive performance.

9.4 Future Directions

Future work should extend the forecasting evaluation using a larger number of completed monthly observations and rolling-origin validation. Additional models such as seasonal methods and more advanced time-series approaches can be investigated when sufficient historical data is available.

Customer segmentation can also be improved by testing alternative cluster counts and stability measures. Silhouette analysis, repeated clustering and alternative clustering algorithms could provide stronger evidence about the structure of the customer population.

The data pipeline can be extended to include profit margin, product cost, promotions, holidays and external demand variables. These features could improve the distinction between high sales volume and high business value.

The application can also expose forecast intervals rather than point estimates only. This would help users understand uncertainty when planning stock and staffing.

Finally, a longitudinal deployment study across multiple SMEs could test whether the insights generated by the system lead to measurable improvements in inventory efficiency, customer retention or sales performance.

REFERENCES

- [1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Amsterdam, Netherlands: Elsevier, 2012.
- [2] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.
- [3] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.
- [4] M. Chen, S. Mao, and Y. Liu, 'Big data: A survey,' Mobile Networks and Applications, vol. 19, no. 2, pp. 171-209, 2014.
- [5] A. McAfee and E. Brynjolfsson, 'Big data: The management revolution,' Harvard Business Review, vol. 90, no. 10, pp. 60-68, 2012.
- [6] G. Adomavicius and A. Tuzhilin, 'Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,' IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, 2005.
- [7] R. Agrawal and R. Srikant, 'Fast algorithms for mining association rules,' in Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487-499.
- [8] P. Kotler and K. L. Keller, Marketing Management, 15th ed. Upper Saddle River, NJ, USA: Pearson, 2016.
- [9] E. Turban, R. Sharda, and D. Delen, Decision Support and Business Intelligence Systems, 10th ed. Upper Saddle River, NJ, USA: Pearson, 2015.
- [10] P. Vassiliadis, 'A survey of extract-transform-load technology,' Journal of Data Warehousing and Mining, vol. 5, no. 3, pp. 1-27, 2009.
- [11] D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining. Cambridge, MA, USA: MIT Press, 2001.
- [12] C. C. Aggarwal, Data Mining: The Textbook. Cham, Switzerland: Springer, 2015.
- [13] D. Olson and D. Delen, Advanced Data Mining Techniques. Berlin, Germany: Springer, 2008.
- [14] F. Ricci, L. Rokach, and B. Shapira, Recommender Systems Handbook, 3rd ed. New York, NY, USA: Springer, 2022.

- [15] S. J. Taylor and B. Letham, “Forecasting at scale,” *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.
- [16] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA, USA: O’Reilly Media, 2022.
- [17] MongoDB Inc., “MongoDB Documentation,” 2025.
- [18] FastAPI, “FastAPI Framework Documentation,” 2025.
- [19] OpenJS Foundation, “Node.js Documentation,” 2025.
- [20] React Native, “React Native Documentation,” 2025.
- [21] J. A. McCarty and M. Hastak, “Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression,” *Journal of Business Research*, vol. 60, no. 6, pp. 656–662, 2007, doi: 10.1016/j.jbusres.2006.06.015. ScienceDirect
- [22] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7. DOI
- [23] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021. Monash University
- [24] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and Machine Learning forecasting methods: Concerns and ways forward,” *PLoS ONE*, vol. 13, no. 3, p. e0194889, 2018, doi: 10.1371/journal.pone.0194889. DOI
- [25] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose, “A state space framework for automatic forecasting using exponential smoothing methods,” *International Journal of Forecasting*, vol. 18, no. 3, pp. 439–454, 2002, doi: 10.1016/S0169-2070(01)00110-8.

GLOSSARY

Term	Definition
Customer behavior analysis	Analysis of purchasing activities and patterns associated with identified customers.
Forecast horizon	The future period for which a forecasting model generates estimates.
K-Means	An unsupervised clustering algorithm that partitions observations around centroids.
Monetary	Total transaction value associated with a customer in RFM analysis.
Recency	Time elapsed since a customer's latest transaction.
Frequency	Number of transactions associated with a customer during the analysis period.
Sales trend	Directional or systematic movement in sales values over time.
Short-term forecasting	Forecasting over a relatively near operational horizon; in this thesis, three monthly periods.
SME	Small and medium-sized enterprise.

A RESEARCH INSTRUMENTS AND EXTENDED RESULTS

A.1 Dataset Data Dictionary

Table A.1: Orders sheet analytical data dictionary

Field	Type	Use
orderid	Identifier	Transaction identity and order counting
cartNumber	Identifier	Cart/order context
checkoutTime	Datetime	Temporal analysis
status	Categorical	Transaction validity/filtering
paymentOption	Categorical	Payment analysis
orderType	Categorical	Operational segmentation
customerName	Text	Customer context; not reported in thesis
customerMobile	Identifier	Customer grouping
amount	Numeric	Transaction value
discountedAmount	Numeric	Discount-related analysis
taxAmount	Numeric	Transaction accounting context
serviceChargeAmount	Numeric	Transaction accounting context
totalAmount	Numeric	Primary revenue measure
itemCount	Numeric	Basket-size context
createdAt	Datetime	Record creation time
updatedAt	Datetime	Record update time

Table A.2: Order Items sheet analytical data dictionary

Field	Type	Use
orderId	Identifier	Link item to order
cartNumber	Identifier	Cart context
checkoutTime	Datetime	Temporal analysis
productName	Text	Product analysis
qty	Numeric	Sales volume
unitCost	Numeric	Unit price/cost field
lineTotal	Numeric	Item revenue
warrantyMonths	Numeric	Product metadata where applicable
warrantyExpiresAt	Datetime	Product metadata where applicable

A.2 RFM Calculation Example

For each identified customer, the system groups eligible transactions. Suppose customer i has a latest purchase at date $D_{last,i}$, n_i orders and transaction values $v_{i1}, \dots, v_{in_i}$. Then:

$$R_i = D_{ref} - D_{last,i}$$

$$F_i = n_i$$

$$M_i = \sum_{j=1}^{n_i} v_{ij}$$

These values are then standardized before clustering.

A.3 Extended RFM Results

Table A.3: Extended RFM cluster profile

Cluster	N	Mean R	Median R	Mean F	Mean M	Revenue Share
At Risk/Lapsed	14	167.36	167	15.64	55,321	1.64%
Regular	54	52.24	52	93.70	361,872	41.32%
VIP/Loyal	16	46.75	47	267.56	1,010,126	34.18%
Occasional	97	66.46	66	29.38	111,456	22.86%

A.4 Forecasting Experiment Detail

The monthly revenue sequence used for the experiment is:

Oct 2025	2,502,050
Nov 2025	2,579,170
Dec 2025	3,309,990
Jan 2026	2,240,830
Feb 2026	2,160,400
Mar 2026	2,447,880
Apr 2026	2,959,040
May 2026	2,747,900
Jun 2026	2,294,680
Jul 2026	2,755,690
Aug 2026	1,556,960

The first eight periods are used for historical model fitting and the final three periods are treated as a pseudo-future validation set.

A.5 User Evaluation Instrument

The user feedback survey contained 52 responses. The following analytics-related statements were analysed:

1. Sales reports and analytics help me understand business performance.
2. Product demand predictions are useful for planning future stock or production.
3. Business performance/profit predictions provide useful information for decision-making.
4. Customer and sales insights help me understand sales patterns.
5. SmartCost's alerts and predictive insights can help me identify potential business problems earlier.
6. Overall, the AI-based features help me make better business decisions.
7. How satisfied are you with SmartCost overall?
8. Compared with your previous/manual method, has SmartCost made managing your business easier?
9. Would you recommend SmartCost to another small business owner?

A.6 User Evaluation Extended Results

Table A.4: Selected user evaluation distributions

Question	SA	A	N	D	SD
Sales reports and analytics	18	23	4	4	3
Product demand prediction	16	29	6	0	1
Business performance prediction	12	28	5	4	3
Customer and sales insights	16	28	6	1	1
Alerts and predictive insights	17	20	11	3	1
AI-based decision support	17	27	6	1	1

A.7 QA Evidence

The group QA report contains 42 issues. The issue most directly related to the analytics interface is BUG-031, concerning cropping of the lower section of the Analytics screen on Android and iOS. It is classified as Medium severity and P2 priority and is recorded as Open.

B ACADEMIC INTEGRITY AND AI USE DISCLOSURE

This thesis was prepared with the use of generative AI tools as an assisting technology for drafting, language refinement, structural organization and LaTeX formatting. The analytical results were based on the supplied project dataset, source code, survey evidence, QA evidence and system screenshots. Generated text was reviewed and adapted by the student.

The student remains responsible for verifying the numerical results, source references, claims, figures, tables, code descriptions and conclusions presented in this thesis.

C EXTENDED ANALYTICAL EVIDENCE AND VALIDATION

C.1 Purpose of Extended Validation

This section documents additional validation considerations that connect the implemented system to the research questions. The purpose is to make the analytical evidence traceable from raw records to model output and finally to the user-facing insight.

The validation chain is:

1. raw transaction records;
2. data-quality checks;
3. customer and time feature construction;
4. analytical model execution;
5. independent reproduction;
6. predictive hold-out evaluation;
7. user-facing presentation;
8. stakeholder feedback.

This chain reduces the risk that a visually attractive dashboard is interpreted as evidence of model correctness without examining the underlying data and evaluation procedure.

C.2 Data Quality Dimensions

Five data-quality dimensions are particularly relevant to the component.

C.2.1 Completeness

Completeness concerns whether the fields required for a particular analysis are available. Customer segmentation requires a customer identifier and transaction timestamp. Sales trend analysis requires a valid timestamp and transaction value. Product analysis requires product name and quantity.

A missing customer identifier does not necessarily invalidate an order for revenue analysis, but it makes the order unsuitable for customer-level RFM analysis. The research therefore uses analysis-specific inclusion criteria.

C.2.2 Consistency

Consistency concerns whether equivalent concepts are represented using compatible values. Timestamp parsing, customer identifier normalization and numerical conversion are examples of consistency operations.

C.2.3 Validity

Validity concerns whether values fall within expected structural ranges. Dates should be parse able, quantities should be numerical and transaction values should be compatible with the transaction status.

C.2.4 Uniqueness

Duplicate transaction identifiers can inflate Frequency and Monetary values. Therefore, order identifiers are treated as important uniqueness controls during customer aggregation.

C.2.5 Timeliness

Timeliness is particularly important for Recency and forecasting. A customer segment can change as new transactions arrive. Similarly, a monthly sales series is affected by whether the current month is complete. The incomplete August 2026 period is therefore treated cautiously.

C.3 Customer Segmentation Sensitivity

The K-Means configuration uses four clusters. This choice is part of the implemented component and is not presented as an objectively optimal universal value. A stronger future experiment could calculate silhouette scores for several values of k and examine cluster stability.

The current research instead focuses on reproducibility and interpretability. The four clusters have clearly different mean Recency, Frequency and Monetary values. The separation is particularly visible between the At Risk/Lapsed cluster and the VIP/Loyal cluster. The At Risk/Lapsed group has the longest mean time since purchase, while the VIP/Loyal group has the largest mean frequency and monetary contribution.

Table C.1: Interpretation of RFM cluster characteristics

Segment	Observed profile	Potential business interpretation
VIP/Loyal	Low recency, high frequency and high monetary contribution	Maintain relationship, prioritize retention and recognize high-value behavior.
Regular	Low recency and substantial frequency and monetary contribution	Encourage continued repeat purchasing and monitor movement toward higher value.
Occasional	Moderate recency, lower frequency and moderate monetary contribution	Encourage repeat purchases and monitor whether engagement increases or de-clines.
At Risk/Lapsed	High recency, low frequency and low monetary contribution	Consider reactivation or retention-oriented communication.

These interpretations are analytical recommendations, not automatic decisions. The system does not independently send customer communications or make irreversible decisions based on the segment.

C.4 Revenue Concentration

The raw data reproduction shows that the customer segments have different contributions to total customer-linked revenue. The 16-customer VIP/Loyal segment contributes approximately 34.18% of the reproduced customer-level revenue, while the 54-customer Regular segment contributes approximately 41.32%.

This illustrates why customer count and revenue contribution should be considered together. A segment with many customers may not necessarily contribute the largest revenue share, and a small high-value segment may require disproportionate retention attention.

C.5 Temporal Analysis Considerations

Hourly sales analysis is sensitive to time zone and transaction timestamp interpretation. The application may also apply to a rolling period filter. Therefore, a reproducible analysis should be documented:

- the source timestamp field;
- time zone assumptions;
- the start and end dates of the selected window;
- transaction-status filtering;
- whether refunded or reversed transactions are included;
- whether the analysis uses order-level or item-level aggregation.

The supplied application screenshot reports a peak of 11 AM–12 PM in the selected six-month view. This is retained as an application result. Independent precomputation should use the same filter and status rules before claiming exact numerical agreement.

C.6 Forecasting Period Definition

The production-oriented research horizon is defined as short-term forecasting over three monthly periods.

The rationale is operational. An SME may need to estimate demand for the next few months to plan procurement, staffing and promotional activity. A three-month horizon is long enough to support operational planning but short enough that recent transaction behavior remains relevant.

Long-term forecasting over years is beyond the evidence supported by the current dataset. Such a forecast would require substantially more historical observations, stable business conditions and additional explanatory variables.

C.7 Hold-Out Design

The hold-out experiment follows a chronological split. Random train/test splitting is avoided because it could allow information from later periods to influence the model used to predict earlier periods.

The training sequence contains October 2025 through May 2026. June, July and August 2026 form the historical evaluation set. This design approximates the real decision problem in which a model is trained using information available before the forecast period.

Table C.2: Historical forecasting split

Dataset portion	Periods	Purpose
Training	Oct 2025–May 2026	Model fitting
Validation/Test	Jun 2026–Aug 2026	Historical pseudo-future evaluation
Future	After Aug 2026	Forecast only; no ground truth in supplied dataset

C.8 Forecasting Error Interpretation

The forecasting experiment produced MAPE values between approximately 29.08% and 32.51%. These errors are relatively large for a precise forecasting application and demonstrate that the current dataset contains substantial month-to-month variation.

This is an important research finding. The presence of a trend line does not guarantee accurate forecasting. The system can therefore use the regression trajectory for trend visualization while a future production version should expose a separately validated forecasting model and communicate uncertainty.

C.9 Implications for Model Selection

The experiment suggests that simple models can be useful benchmarks in SME environments. A lightweight method with lower error may be more practical than a complex model that requires more infrastructure, particularly when the available time series is short.

However, model selection should not be based on a single three-period hold-out. A future study should use rolling-origin evaluation over several historical windows. This would provide a distribution of errors rather than one estimate.

C.10 Research Traceability Matrix

Table C.3: Traceability from research question to evidence

Question	Method	Evidence	Chapter
RQ1	RFM feature construction and K-Means	Cluster profiles and reproducibility	3, 6
RQ2	Fixed configuration and raw-data reproduction	Cluster counts and RFM means	3, 6
RQ3	Temporal and product aggregation	Peak hours, weekend comparison and product results	6
RQ4	Chronological hold-out forecasting	MAE, RMSE and MAPE	3, 6
RQ5	Baseline comparison	Traditional POS versus integrated analytics discussion	2, 7
RQ6	User survey	52-response analytics-focused evaluation	6

C.11 Research Contribution Versus Product Function

Table C.4: Distinction between product functionality and research contribution

Area	Product functionality	Research contribution
Customer segmentation	Display customer groups	Reproduce and evaluate RFM/K-Means from raw data
Sales trends	Display revenue trajectory	Evaluate whether historical trends provide predictive information
Forecasting	Trend visualization	Compare forecasting baselines on held-out observations
Data processing	Prepare dashboard data	Define validation and reproducibility procedures
User interface	Present insights	Evaluate perceived usefulness through structured feedback

C.12 Threats to Internal Validity

Internal validity can be affected by data filtering, application-side transformations, incomplete periods and differences between raw-data calculations and the live application snapshot. The documented 97-versus-98 Occasional-customer discrepancy is one example.

The research reduces this threat by retaining the exact raw-data reproduction results, documenting application screenshot results separately and avoiding unsupported reconciliation.

C.13 Threats to External Validity

The transaction history represents one project environment. Customer behavior in another SME may differ because of business type, product mix, customer demographics, seasonality and pricing. The results therefore demonstrate feasibility and analytical behavior within the study dataset rather than population-level generalization.

C.14 Threats to Construct Validity

RFM is a useful but simplified representation of customer behavior. It does not include satisfaction, product preference depth, marketing response, profitability or demographic characteristics. Similarly, revenue trend is not identical to profit trend.

The research therefore uses careful terminology. It reports “customer purchasing behavior” based on transaction data rather than claiming to measure complete human behavior.

D ADDITIONAL RESULTS TABLES AND ANALYTICAL INTERPRETATION

D.1 Full Monthly Revenue Table

Table D.1: Extended monthly revenue observations

Period	Revenue (Rs.)	Interpretation context
Jun 2025	2,132,300	Historical context
Jul 2025	2,582,590	Historical context
Aug 2025	2,818,180	Historical context
Sep 2025	2,317,560	Historical context
Oct 2025	2,502,050	Forecasting training period
Nov 2025	2,579,170	Forecasting training period
Dec 2025	3,309,990	Forecasting training period
Jan 2026	2,240,830	Forecasting training period
Feb 2026	2,160,400	Forecasting training period
Mar 2026	2,447,880	Forecasting training period
Apr 2026	2,959,040	Forecasting training period
May 2026	2,747,900	Forecasting training period
Jun 2026	2,294,680	Hold-out period
Jul 2026	2,755,690	Hold-out period
Aug 2026	1,556,960	Hold-out period; incomplete month

D.2 Forecast Prediction Table

Table D.2: Forecast values for the historical hold-out experiment

Method	Jun 2026	Jul 2026	Aug 2026
Actual	2,294,680	2,755,690	1,556,960
Linear regression	2,669,490	2,680,842	2,692,193
Naive	2,747,900	2,747,900	2,747,900
3-period moving average	2,718,273	2,808,404	2,758,193
Simple exponential smoothing	2,618,130	2,618,130	2,618,130

The prediction table shows why the final August observation is important. The actual value is substantially lower than the model estimates, but August is incomplete. This illustrates the danger of interpreting an incomplete operational period as a complete demand observation.

D.3 Forecasting Discussion by Method

D.3.1 Linear regression

Linear regression is highly interpretable. Its main advantage is that the user can understand the concept of an increasing or decreasing trajectory. Its weakness in this dataset is that the month-to-month values do not closely follow a single straight line.

D.3.2 Naive forecast

The naive forecast provides a strong conceptual baseline because it asks whether a complex model performs better than simply repeating the last known value. Its performance is useful for judging the value added by analytical modelling.

D.3.3 Moving average

The moving average reduces noise but can lag behind sudden changes. It is appropriate when short-term fluctuations are considered noise, but it may respond slowly to structural changes.

D.3.4 Simple exponential smoothing

Simple exponential smoothing gives more importance to recent information and therefore can adapt more quickly than a long historical average. In the selected experiment it produced the lowest error metrics among the tested methods.

D.4 Customer Segment Revenue Distribution

The raw data reproduction gives the following revenue shares:

Table D.3: Customer-level revenue concentration

Segment	Revenue (Rs.)	Share
VIP/Loyal	16,162,020	34.18%
Regular	19,541,080	41.32%
Occasional	10,811,200	22.86%
At Risk/Lapsed	774,500	1.64%
Total	47,288,800	100.00%

The distribution demonstrates the value of combining segmentation with revenue analysis. The at-risk group contributes only a small proportion of customer-linked revenue, while the Regular and VIP/Loyal groups represent most of the monetary contribution.

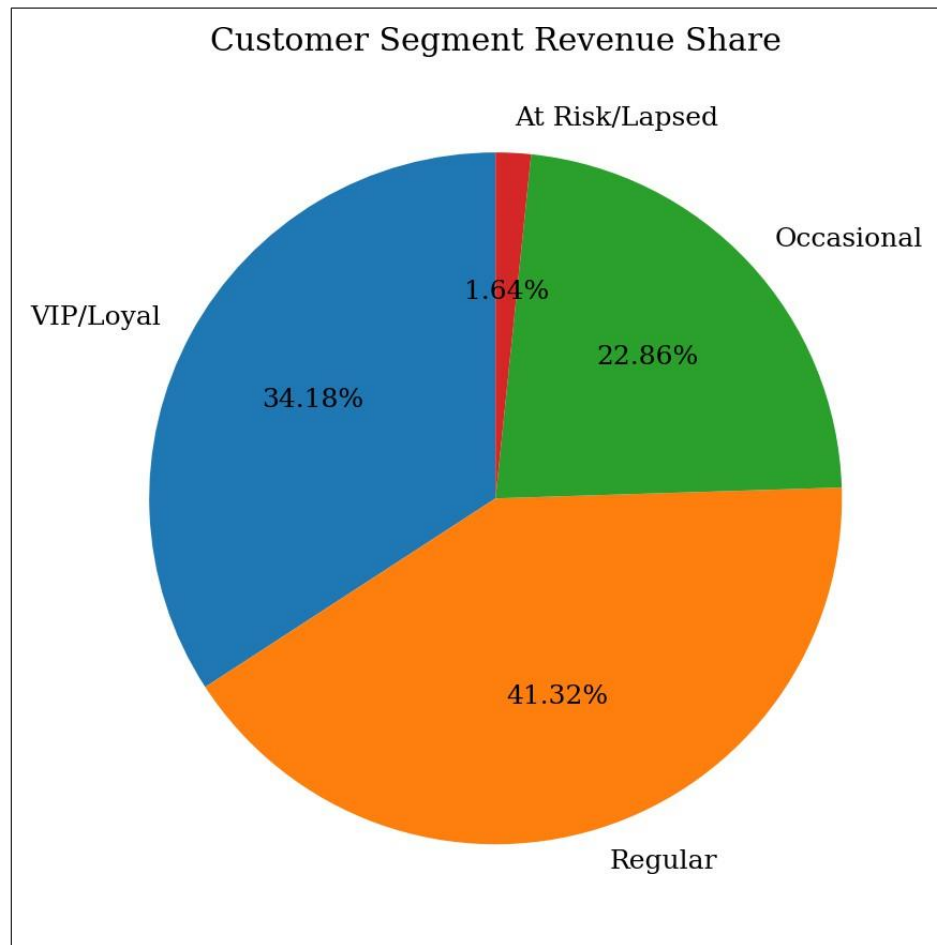


Figure D.1: Customer segment revenue share from raw-data reproduction

D.5 User Evaluation Detailed Interpretation

The survey results provide evidence from 52 respondents. The strongest analytics-related positive rate among the selected questions was for product demand predictions, with 45 respondents agreeing or strongly agreeing. Customer and sales insights received 44 positive responses, while AI-based decision support also received 44 positive responses.

The lowest positive rate among the selected analytics/prediction questions was the statement concerning alerts and predictive insights helping users identify business problems earlier, with 37 positive responses. This suggests that predictive alerts may require clearer explanation, stronger evidence or better integration into operational workflows.

Overall satisfaction was high in the survey, with 45 respondents selecting a score of four or five. However, satisfaction should not be interpreted as a direct measure of analytical accuracy.

D.6 User Evaluation Limitations

The survey was conducted in the context of the SmartCost project and is not a randomized sample of the SME population. Responses are self-reported and may be affected by familiarity with the project. The questionnaire also evaluates the complete SmartCost experience, so only the questions directly relevant to the individual component are used in this thesis.

E SYSTEM SCREENSHOT EVIDENCE

E.1 Customer Behavior Main Screen

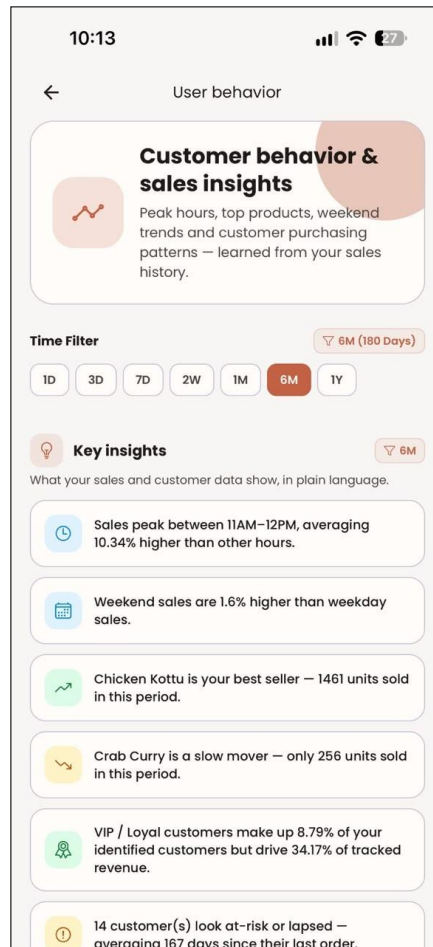


Figure E.1: Implemented Customer Behavior and Sales Insights overview interface

E.2 Busiest Hours Analysis

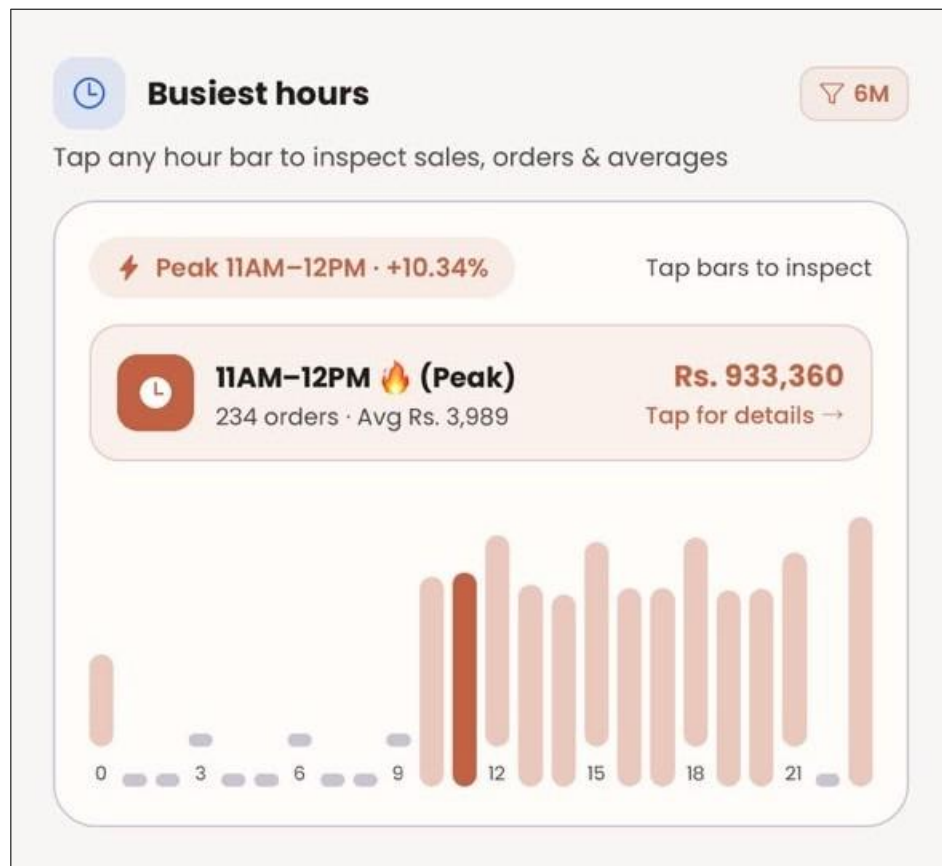


Figure E.2: Implemented busiest-hours analysis interface

E.3 Weekend and Weekday Analysis

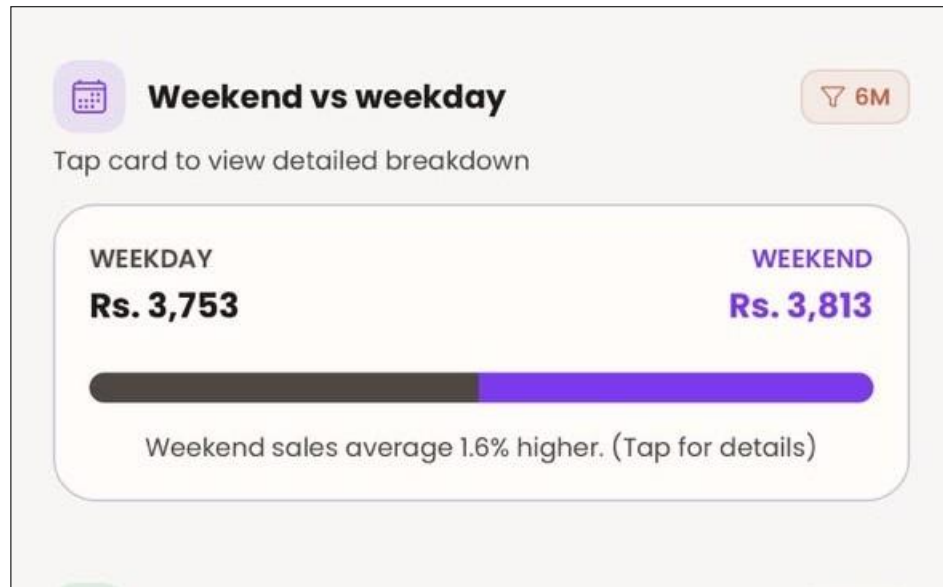


Figure E.3: Implemented weekend-versus-weekday sales comparison

E.4 Top-Selling Products

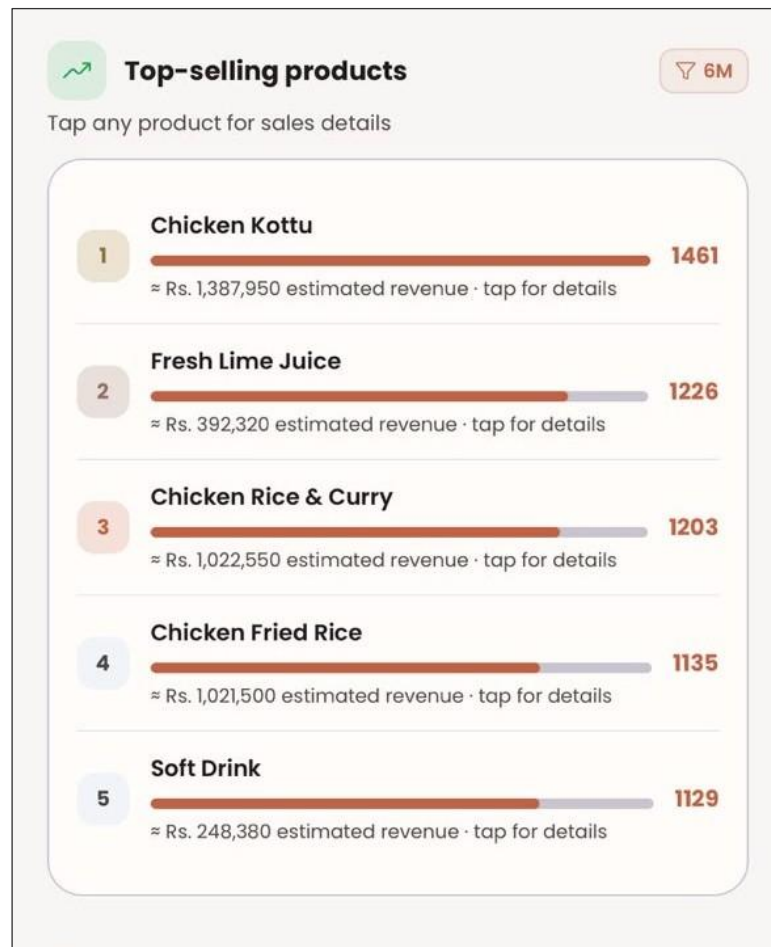


Figure E.4: Implemented top-selling product analysis

E.5 Slow-Moving Products

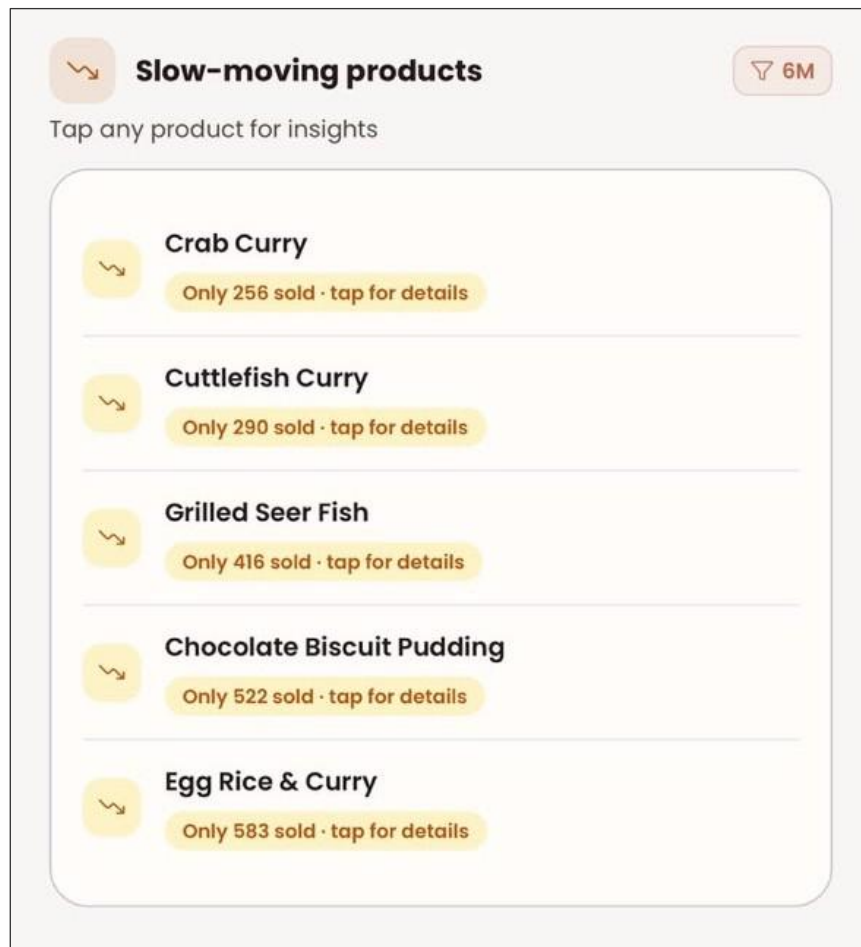


Figure E.5: Implemented slow-moving product analysis

E.6 Customer Segmentation



Figure E.6: Implemented RFM and K-Means customer segmentation interface

E.7 Sales Trend Trajectory



Figure E.7: Implemented sales trend and revenue trajectory interface

E.8 Custom Report Configuration

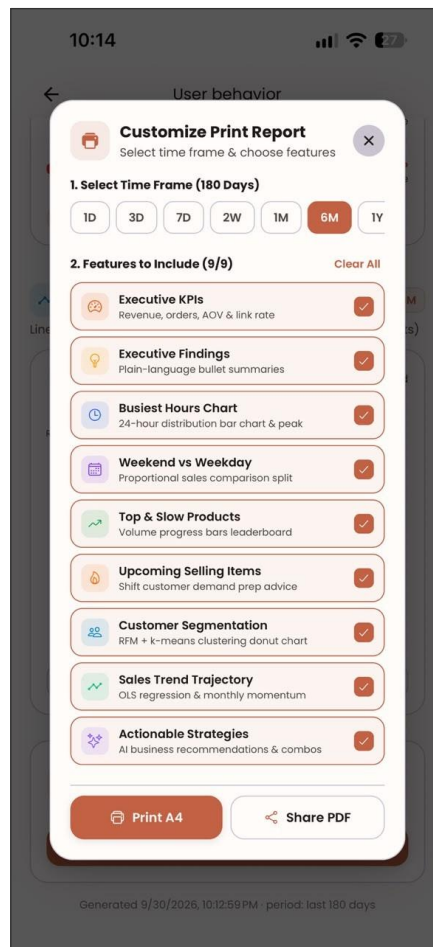


Figure E.8: Implemented customer behavior report customization interface

F QUALITY ASSURANCE AND PROJECT EVIDENCE

F.1 Relevant Quality Assurance Finding

The group QA evidence contains BUG-031, which concerns the Analytics screen being cropped at the lower section on both Android and iOS. It is recorded as a medium severity, P2 issue and was open in the supplied report.

The issue demonstrates that the analytics interface was included in system-level quality assurance. It also illustrates the distinction between analytical correctness and user-interface correctness.

F.2 Interpretation of QA Evidence

The QA report should not be interpreted as an algorithm-validation report. Its purpose is software quality assurance across the group application. Algorithm-level validation is provided separately through data reproduction and forecasting experiments.

This distinction strengthens the research argument because each evidence source is used for the purpose it actually supports.

F.3 Project Management Evidence

The MS Planner export records completed tasks assigned to IT22151810 within the wider SmartCost project. These records demonstrate participation in broader product development and integration activities. They are not used as proof of ownership of analytics algorithms that were assigned elsewhere in the group.

The principal evidence for the individual analytical contribution remains the Python source code, dataset analysis, analytical screenshots, forecasting experiment and user evaluation.