Superstore Analytics Dashboard: Business Intelligence, Forecasting, and Customer Segmentation

Seminar Paper

"Business Intelligence Dashboard implementation"
"*In the context of Advanced Analytics*"
Winter semester 26/27

Supervisor: Prof. Dr. Christian Leubner

Course: Informatics Business (M.Sc.)

Submitted By: Md Abdullah Al Noman Mridha
Matriculation Number: 30494592
Haydnstraße, 63
44145, Dortmund
Nomanmridha.mdabdullahal@fh-swf.de
https://github.com/nomanmridha/Superstore-data-Analysis

Deadline: 2027-01-31

Table of Contents

List of Abbreviations

Abbreviation	Full Form	Definition
BI	Business Intelligence	Technologies and processes used to collect, analyse and visualize business data for decision-making.
EDA	Exploratory Data Analysis	Initial data investigation process used to summarize patterns, detect anomalies and validate assumption.
KPI	Key Performance Indicator	A measurable metric used to evaluate business performance (e.g.. Sales, Profit, margin)
RFM	Recency, Frequency, Monetary	A customer segmentation framework evaluating purchase recency, purchase frequency, and total spending.
MAE	Mean Absolute Error	Average of absolute differences between actual and predicted values.
RMSE	Root Mean Squared Error	Square root of average squared forecast errors; penalizes large deviations.
MAPE	Mean Absolute Percentage Error	Percentage-based forecasting accuracy metric.
SARIMA	Seasonal Autoregressive Integrated Moving average	Time series model capturing trend, seasonality and autoregressive behaviour.
ARIMA	Autoregressive integrated moving average	Statistical forecasting model combining AR and MA components with differencing.
K-means	K-Means Clustering	Unsupervised learning algorithm used to group customers based on similarity.
Dash	Plotly Dash Framework	Python-based web framework for building analytical dashboards.
Prophet	Facebook Prophet	Additive time-series forecasting model handling trend and seasonality.

List of Symbols

Symbol	Name	Definition / Use in Project
μ	Mean (Average)	Statistical average value of a dataset.
σ	Standard Deviation	Measure of variability in a dataset.
Σ	Summation	Represents total aggregation (e.g. Total Sales, total monetary value).
Δt	Time Difference	Difference in time used to calculate recency in RFM analysis.
$\geq$	Greater Than or Equal	Used in filtering conditions (e.g. Sales thresholds)
$\leq$	Less than or Equal to	Used in conditional data filtering.
%	Percentage	Used in Profit margin and MAPE calculations.
K	Number of Clusters	Represents number of clusters in K-means algorithm.
y_t	Actual Value at Time t	Observed sales value in time-series forecasting.
ŷ_t	Predicted Value at Time t	Forecasted sales value in time-series value.
e_t	Forecast Error	Difference between actual and predicted values.
R	Recency	Time since last customer purchase
F	Frequency	Number of customer purchase.
M	Monetary	Total revenue generated by a customer.

List of Figures

Introduction

The rapid digitalization of retail operations has transformed how organizations capture and utilize transactional data. Modern retail enterprises no longer rely solely on descriptive summaries of past sales; instead, they increasingly integrate predictive modelling and customer intelligence to guide strategic decisions. Business Intelligence (BI) systems serve as the backbone of this transformation, enabling interactive visualization, forecasting, and segmentation in a unified analytical environment. (Sharda, 2018)

The Superstore dataset, obtained from Kaggle, provides a structured representation of retail transactions across multiple regions, categories, and customer segments. While often used for introductory analytics exercises, this project approaches the dataset from a decision-support perspective. Rather than performing isolated statistical analyses, the objective is to design and implement a fully functional analytics dashboard that integrates descriptive analytics, forecasting, and RFM-based customer segmentation.

The research questions guiding this project are:

1. How can transactional retail data be transformed into an interactive Business Intelligence dashboard?
2. To what extent can future sales be forecasted reliably using time-series method?
3. How effectively can RFM segmentation reveal actionable customer groups?
4. How can these analytical outputs inform managerial decision-making?

The report proceeds by presenting theoretical foundations, methodology, empirical findings and business implications.

Theoretical Background

Business Intelligence in Retail

Business Intelligence encompasses tools and processes used to transform raw data into meaningful information for decision-making (Sharda, 2018). In retail environments, BI dashboards typically present KPIs such as total sales, profit margins, regional performance, and customer behaviour metrics. Effective BI systems are interactive, allowing users to filter across time periods, product categories, and geographic regions.

Time Series Forecasting

Time-series forecasting models aim to capture temporal patterns such as trend and seasonality. The ARIMA framework developed by (Box, 1976) remains foundational in modelling linear dependencies in time-series data. Seasonal (SARIMA) extends this approach by explicitly incorporating periodic patterns (Hyndman, 2018).

Forecast accuracy is evaluated using metrics such as MAE, RMSE, and MAPE. While MAPE provides interpretability in percentage terms, it may exaggerate errors when actual values are small (Hyndman, 2018).

RFM Customer Segmentation

RFM analysis evaluates customers based on recency, frequency, and monetary value (Fader, 2005). This behavioural segmentation approach is widely used in retail marketing to identify high-value customers and those at risk of churn. When combined with clustering algorithms such as K-means (MacQueen, 1967), RFM enables automated grouping of customers into actionable segments.

Dataset Preparation and Exploratory Analysis

Raw Dataset Overview

Figure 01 presents a preview of the raw Superstore dataset. It contains order-level transactional data, including order date, ship mode, region, category, sales, profit, and discounts.

FIGURE 1: RAW DATASET OVERVIEW

The dataset spans 2014-2017 and includes over 9,000 transactions.

Data Cleaning

Figure 2 shows the cleaned dataset structure after pre-processing.

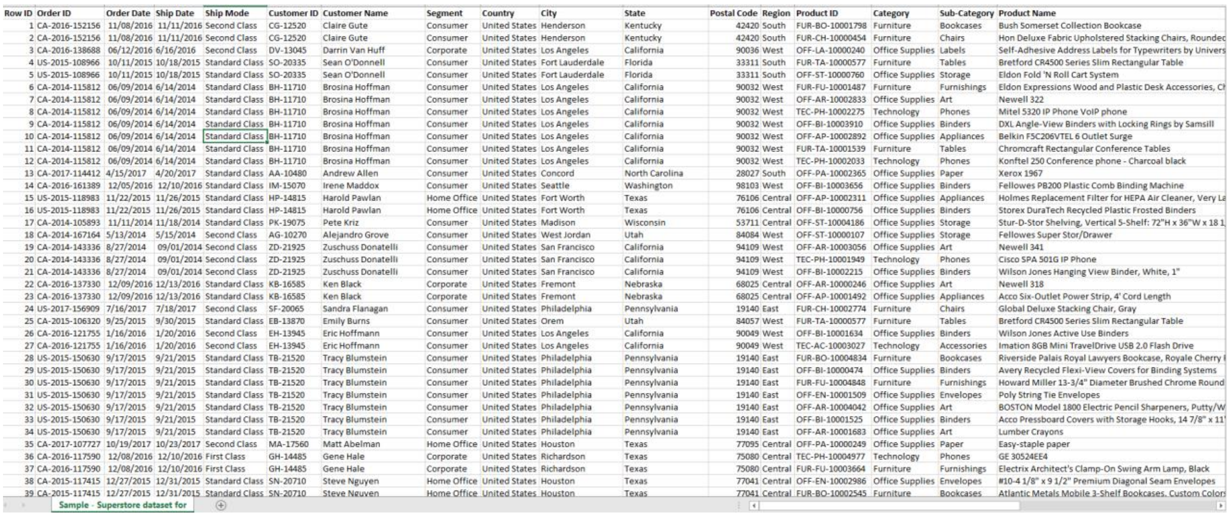

FIGURE 2: CLEANED DATASET

Cleaning Involved:

- Converting date columns into datetime format.
- Removing duplicate records.
- Correcting profit scaling anomalies

- Verifying discount consistency.
- Aggregating monthly sales.

This step ensured that forecasting and segmentation would rely on consistent, reliable data.

Exploratory Analysis

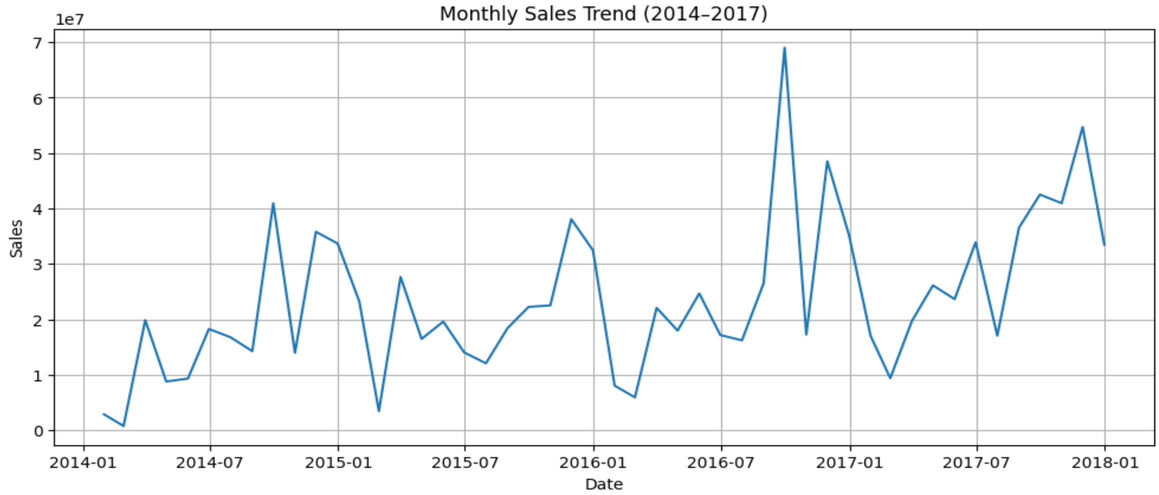

FIGURE 3: MONTHLY SALES TREND (2014-2017)

Sales show clear seasonality, with peaks near year-end periods. This pattern suggests inventory planning must account for seasonal demand spikes.

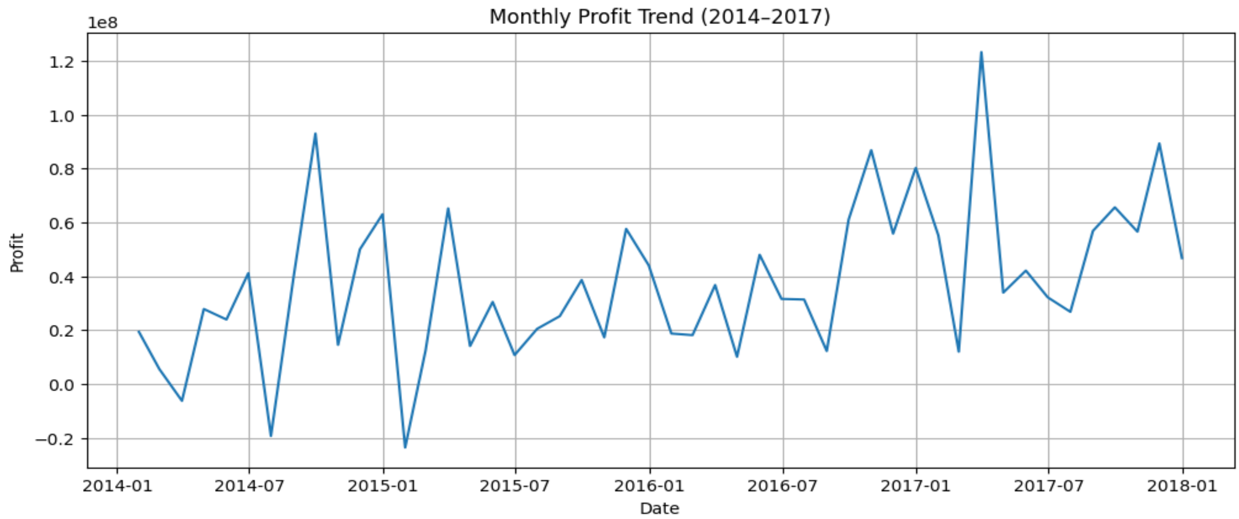

FIGURE 4: MONTHLY PROFIT TREND (2014-2017)

Profit does not increase proportionally with sales, indicating that discounting and cost structures influence margins.

Descriptive Dashboard Analytics

Category-Level Performance

Office Supplies generate the highest profit despite moderate sales volume. Furniture exhibits higher revenue but weaker profitability.

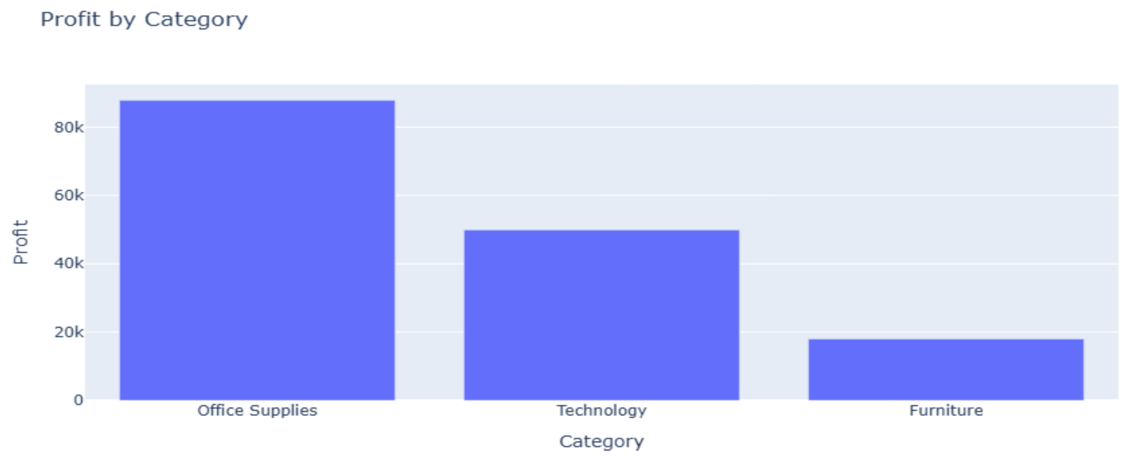

FIGURE 5: PROFIT BY CATEGORY

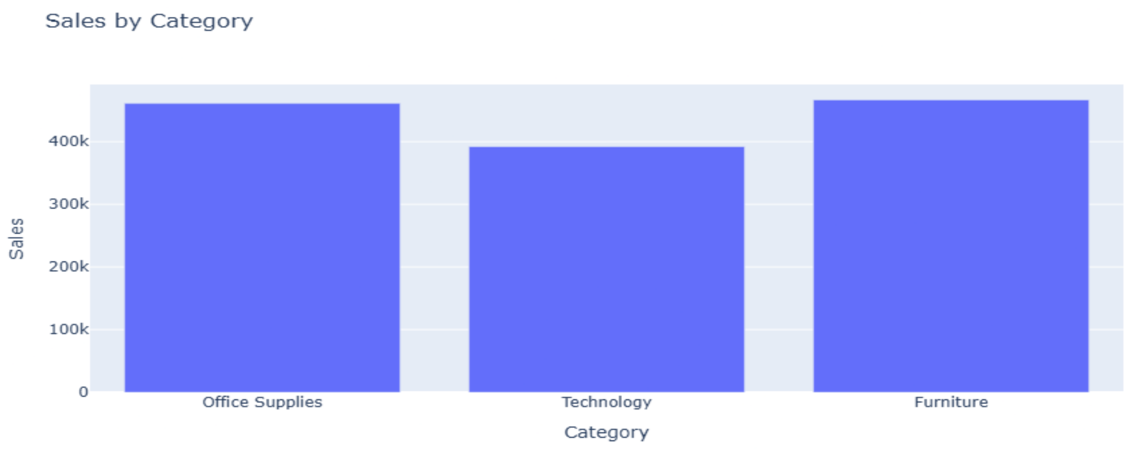

FIGURE 6: SALES BY CATEGORY

This divergence between revenue and profitability highlights the importance of margin-focused strategy.

Discount Impact

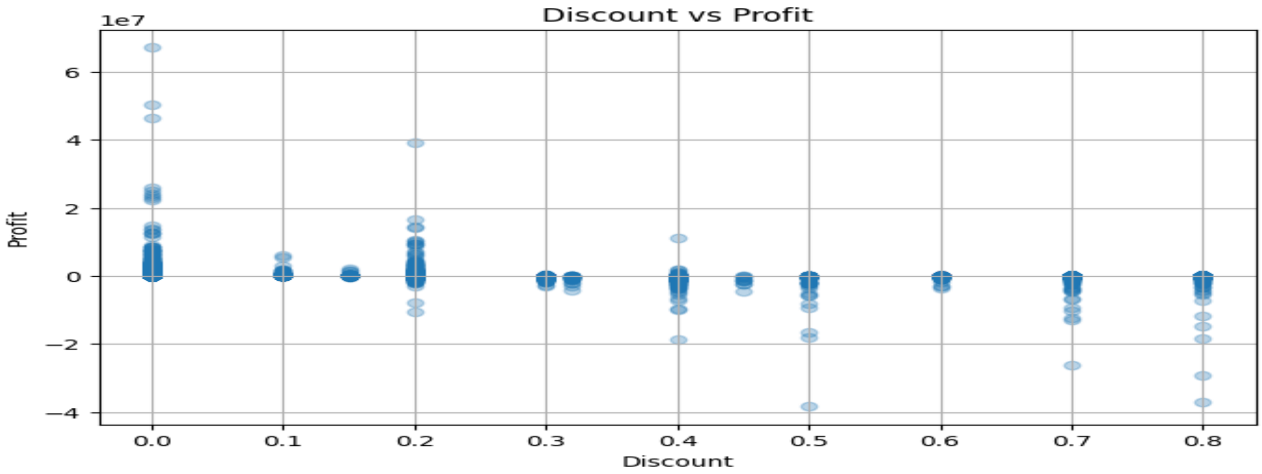

FIGURE 7: DISCOUNT-PROFIT RELATIONSHIP

A negative correlation (~ -0.38) indicates that heavy discounting reduces profitability.

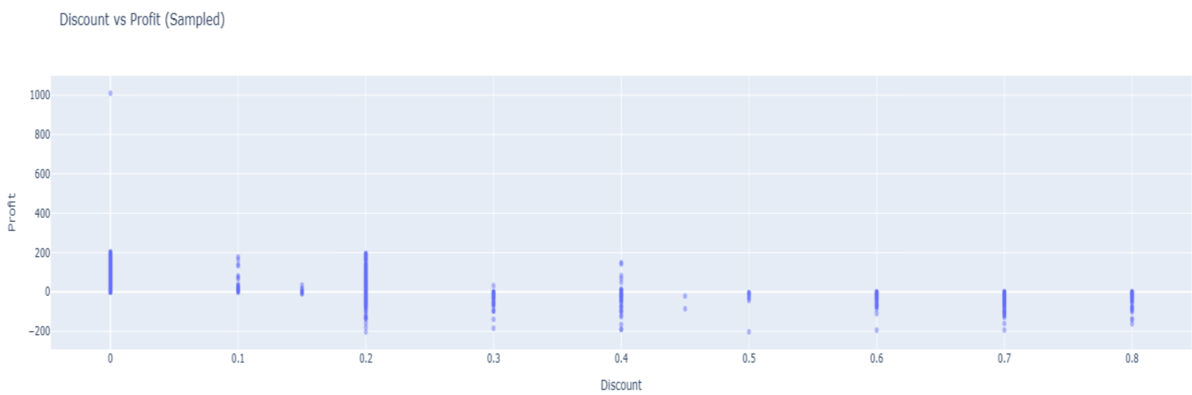

FIGURE 8: OVERALL DISCOUNT-PROFIT RELATIONSHIP

The scatterplot reinforces the nonlinear relationship between discount levels and profit outcomes.

Sales vs Profit Overtime

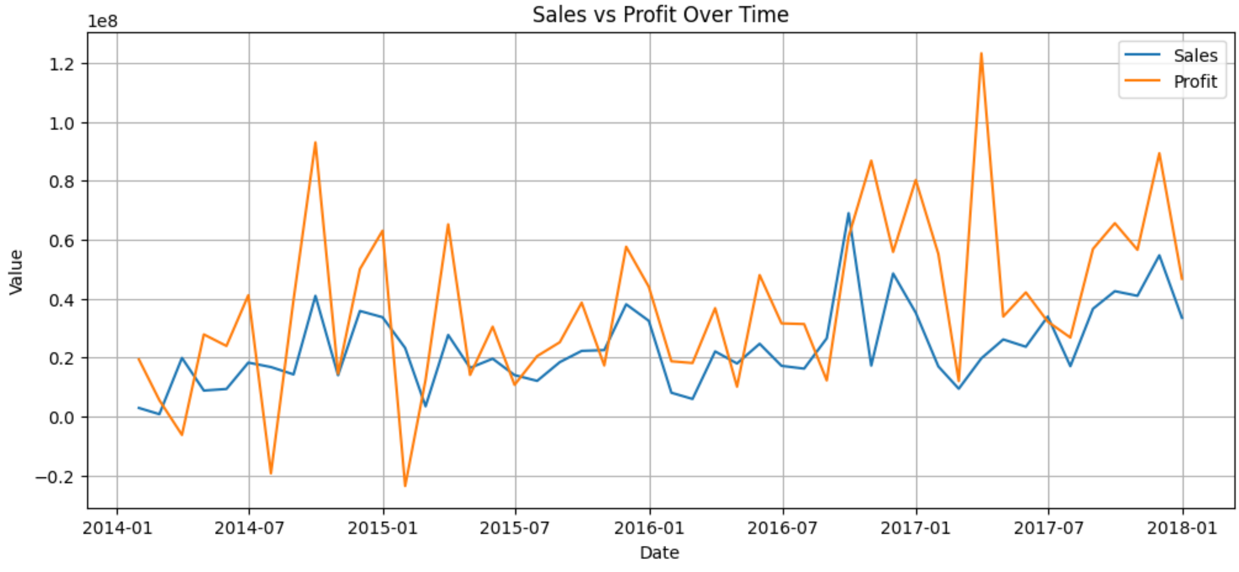

FIGURE 9: SAMPLED DASHBOARD VIEW

Periods of high sales do not always align with high profitability, suggesting promotional strategies require optimization.

Forecasting Results

Six-Month Sales Forecast

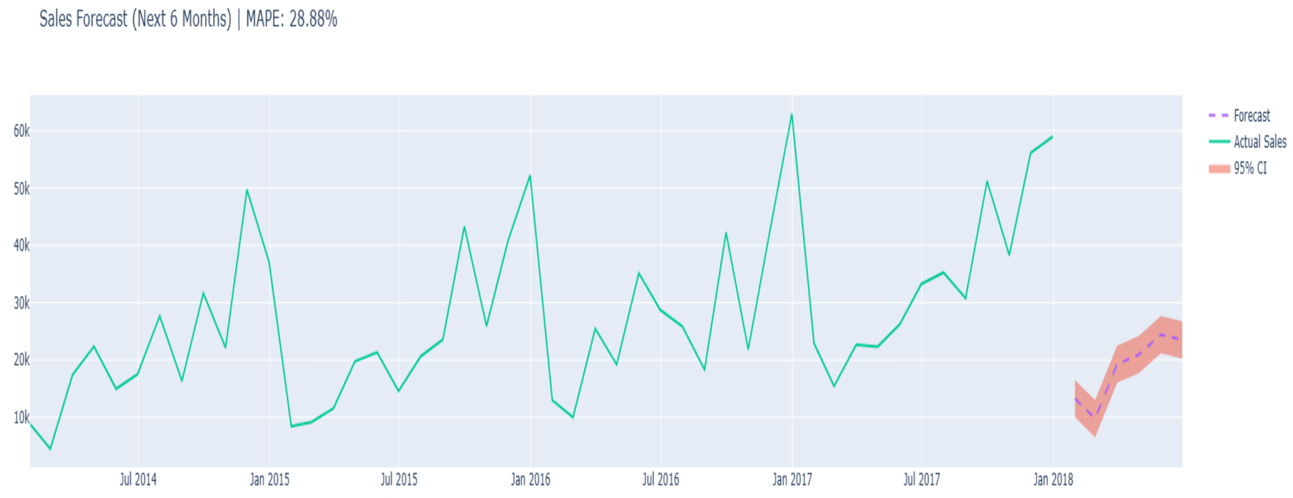

FIGURE 10: SIX MONTH SALES FORECAST

The model achieved a MAPE of 28.88%. While moderate, this accuracy level is acceptable for retail forecasting (Makridakis, 2020). Forecasts indicate short-term stabilization rather than aggressive growth.

Segment-Specific Forecasts

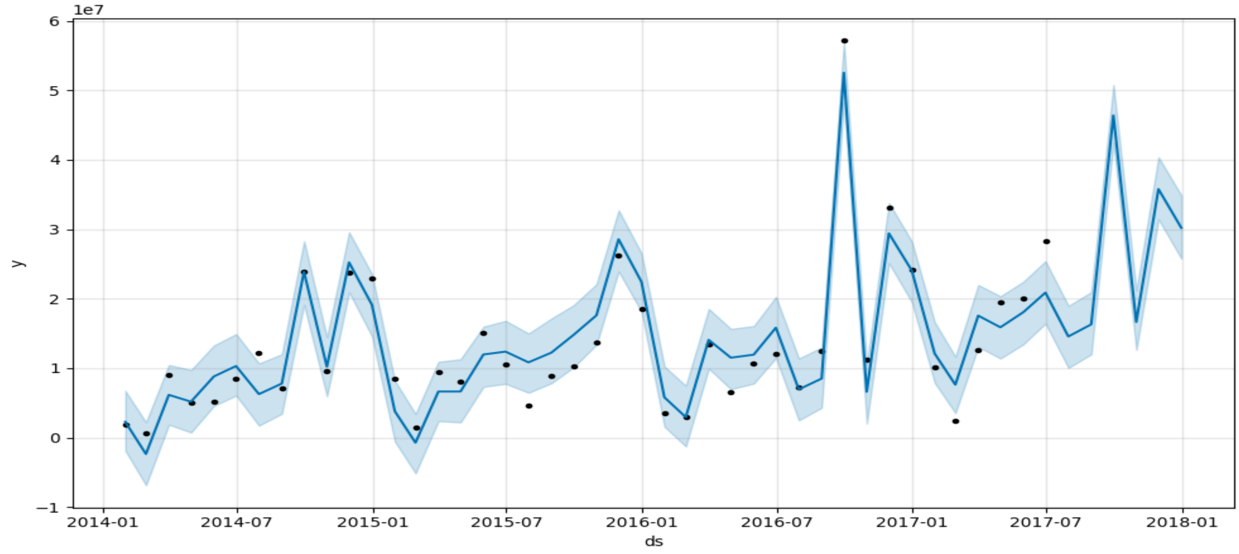

FIGURE 11: CHAMPION SEGMENT FORECAST

Champions show stable purchasing behaviour with consistent projected sales.

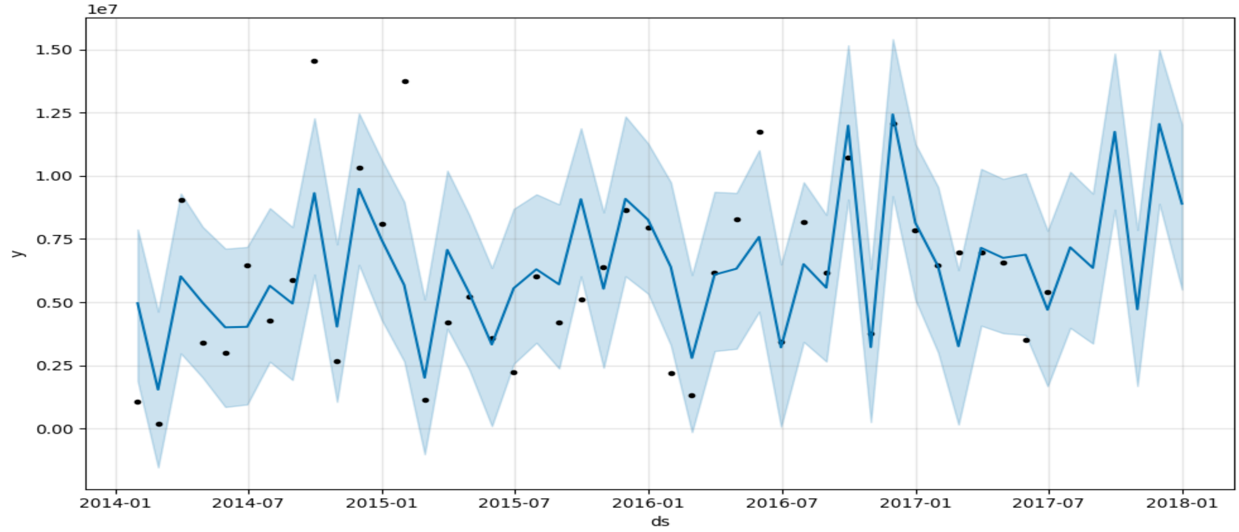

FIGURE 12: LOYAL SEGMENT FORECASTS

Loyal customers demonstrate moderate but predictable revenue patterns.

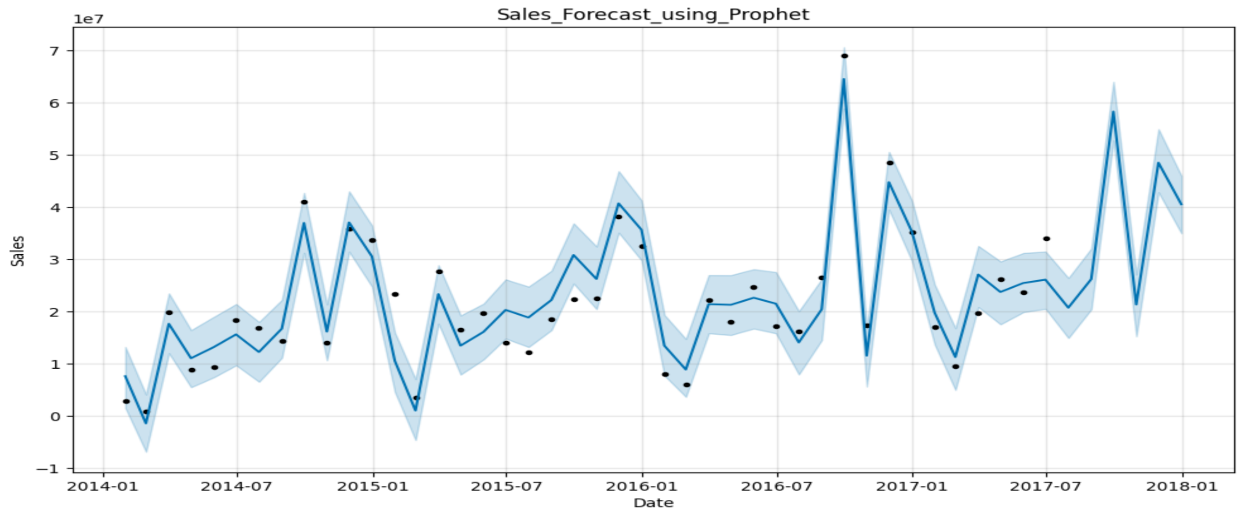

FIGURE 13: PROPHET-BASED FORECASTING

The Prophet model captures trend components similarly but introduces smoother seasonal adjustments.

Segment-level Forecasting Evaluation

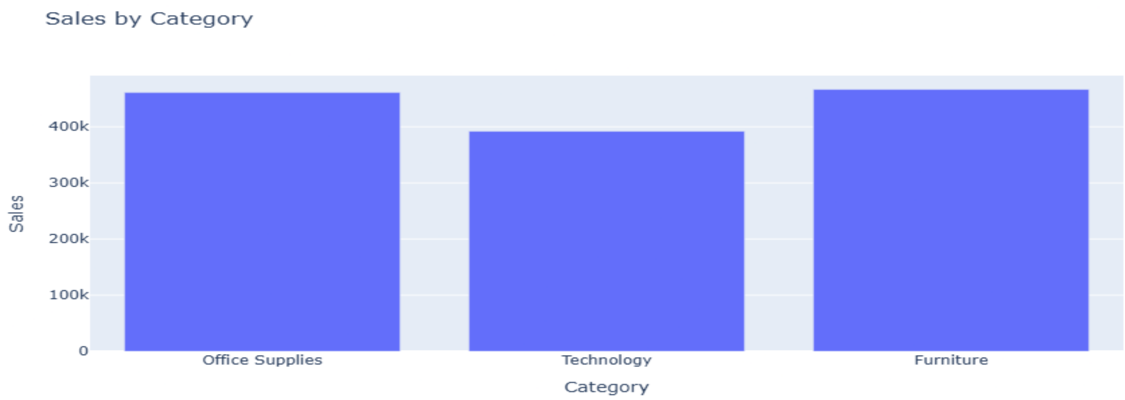

FIGURE 14: CATEGORY-LEVEL FORECAST EVALUATION

Technology category demonstrates higher volatility.

Segment evaluation (forecast KPIs)

Region	months	forecast_sum	last6_actual_sum	growth_pct	mape	rmse	status
West	48	36625.92	96452.87	-62.03	20.19	4147.01	ok
East	48	26224.59	82208.23	-68.1	24.64	3980.97	ok
Central	48	25137.72	53664.69	-53.16	14.91	1328.14	ok
South	48	23183.34	38309.75	-39.48	18.42	2258.75	ok

FIGURE 15: REGION-LEVEL EVALUATION

Central region shows lowest MAPE (~14.09%), indicating greater predictability compared to East (~24.6%).

Segment evaluation (forecast KPIs)

Segment	months	forecast_sum	last6_actual_sum	growth_pct	mape	rmse	status
Consumer	48	55939.77	132316.69	-57.72	13.21	3994.39	ok
Corporate	48	33972.1	80832.86	-57.97	19.94	3219.61	ok
Home Office	48	21259.69	57486	-63.02	36.21	3945.71	ok

FIGURE 16: SEGMENT-LEVEL EVALUATION

Consumer segment exhibits lower forecast error than home office.

These variations support differentiated regional inventory strategies.

Customer Segmentation (RFM)

Optimal Clustering

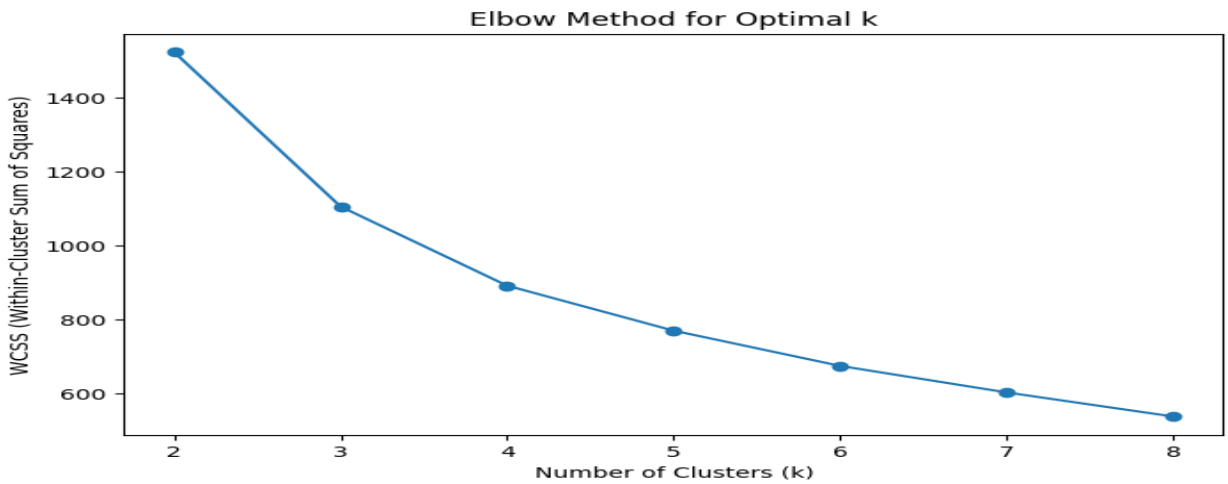

FIGURE 17: ELBOW METHOD CLUSTERING

The inflection point suggests k = 4 clusters.

Cluster Distribution

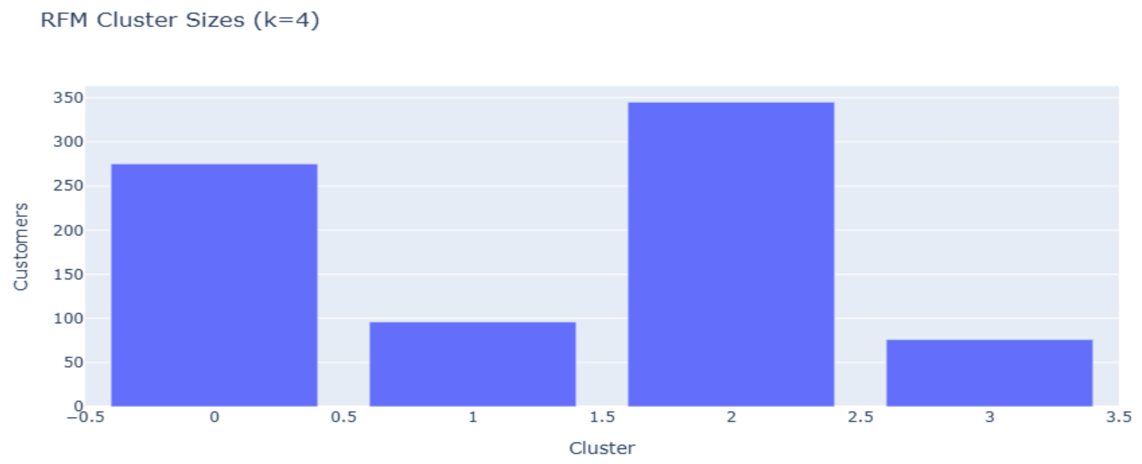

FIGURE 18: CLUSTER SIZE DISTRIBUTION

Cluster sizes reveal a concentration of revenue in high-value segments.

Behavioral Visualization

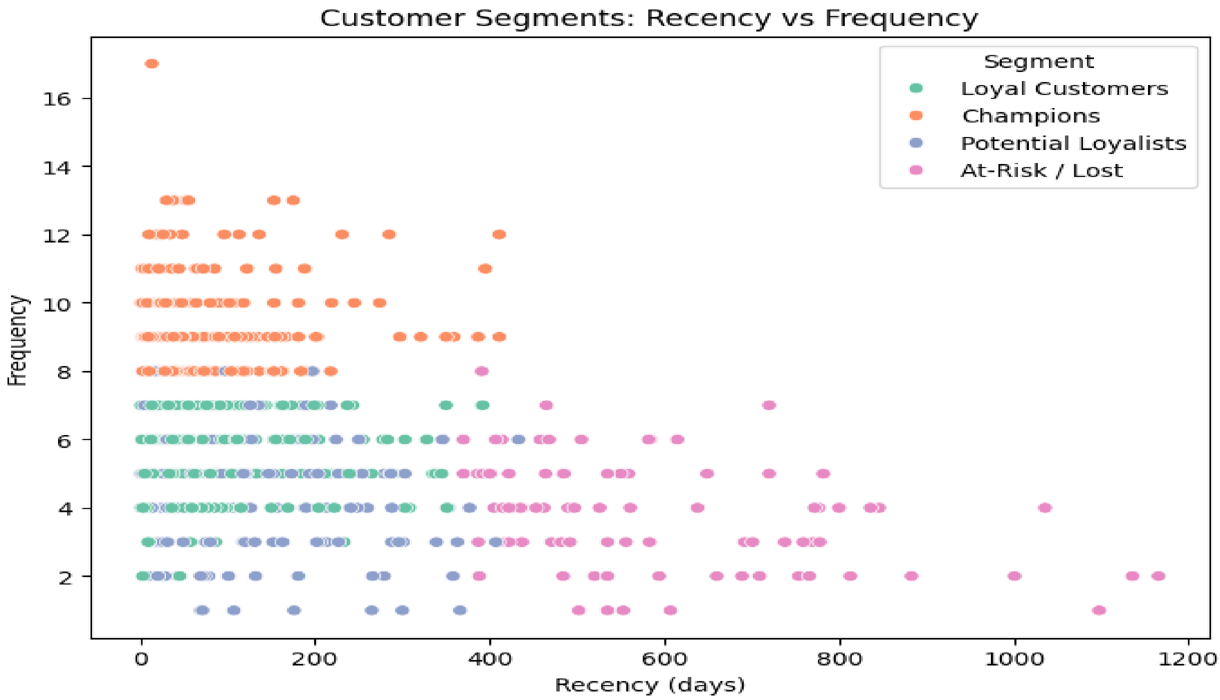

FIGURE 19: RECENCY VS FREQUENCY

Customers with low recency and high frequency represent Champions.

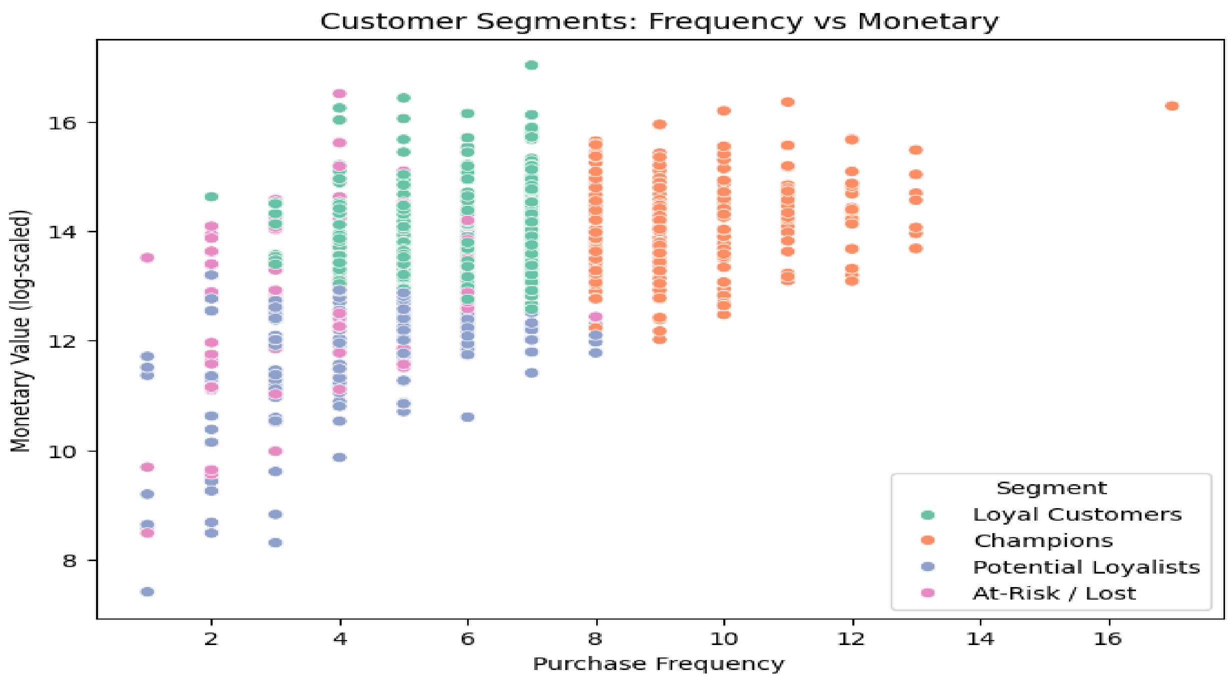

FIGURE 20: FREQUENCY VS MONETARY

High-monetary clusters represent big spenders.

Final Dashboard

The dashboard integrates:

- KPIs
- Forecast visualization
- RFM Clustering
- Segment evaluation
- Interactive filters

It functions as a unified retail decision-support tool.

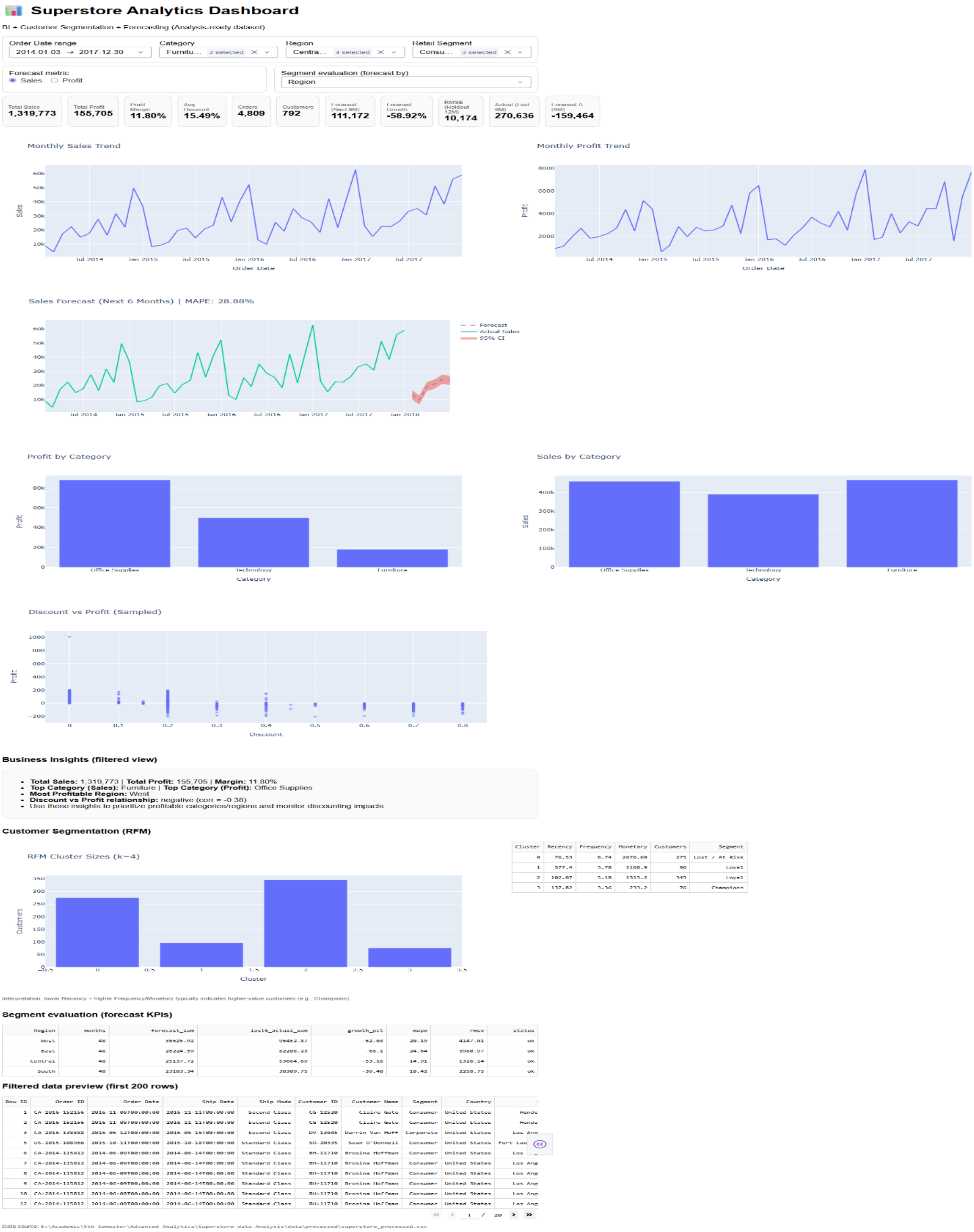

FIGURE 21: FINAL DASHBOARD PREVIEW

Discussion

The results of this project demonstrate how combining descriptive analytics, forecasting, and customer segmentation in a single dashboard creates a more complete decision-support system than using each method independently. From a retail management perspective, the dashboard does not only describe "what happened", but also supports answering "what is likely to happen next" and "which customers contribute most to value and risk".

A key descriptive insight is the clear separation between sales volume and profitability. The category analysis (Figure 5 and 6) suggests that higher sales do not automatically imply higher profit. In practice, this is important because managers can over-prioritize revenue while overlooking margin efficiency. In this dataset, Office supplies appears to generate strong profit relative to its sales, whereas Furniture shows comparatively weaker profitability. This pattern is consistent with real retail environments, where product mix, discount sensitivity, and fulfilment costs can produce different margin structures even when sales are high. Therefore, category-level decisions (e.g. Promotion intensity, shelf focus, bundling) should incorporate profit contribution rather than sales alone.

Discounting emerges as another major lever. The negative discount–profit relationship (Figure 7 and Figure 8) indicates that discounting tends to erode profitability in this dataset. While discounts can drive short-term volume, the plots suggest diminishing or negative returns at higher discount levels. From a managerial standpoint, this implies the need for a more disciplined discount policy: discounts should be targeted to specific customer segments, time windows, or categories rather than applied broadly. The sampled dashboard view (Figure 9) also reinforces that even when sales are high, profitability may not follow—suggesting that promotional campaigns can inflate revenue at the expense of margin. A practical recommendation is to set guardrails such as maximum discount thresholds by category, and to monitor profitability impact immediately during campaign periods.

The forecasting component adds forward-looking capability. The six-month forecast (Figure 10) provides a short-term projection that can support inventory planning, staffing, and budgeting. The reported forecasting accuracy (MAPE ≈ 28.88%) indicates moderate performance. In retail forecasting, this level can be acceptable for high-level planning, especially when the underlying demand is volatile or affected by promotions. However, it also signals that forecast should be used as directional guidance rather than precise targets. This is exactly why the project includes holdout validation metrics and segment evaluation: the value is not only producing a forecast line, but helping decision-makers understand where the forecast is more reliable and where it is riskier.

The segment-level forecasting evaluation (Figures 14–16) is one of the most practically valuable parts of the dashboard because it reveals that forecast reliability is not uniform across business dimensions. For example, the regional evaluation shows Central as more predictable (lower MAPE) compared to East (higher MAPE). Operationally, this supports differentiated planning: regions with stable demand can operate with leaner safety stock and tighter replenishment cycles, while volatile regions require buffer inventory and more frequent forecast updates. Similarly, category-level differences imply that certain categories may require enhanced modelling or additional explanatory variables (e.g., promotional calendars), while others can be managed effectively with simpler models.

The segmentation component provides customer-level insight that complements the aggregate time-series view. The RFM analysis and clustering (Figures 17–20) identify meaningful behavioural groupings: customers who purchase recently and frequently (Champions) differ fundamentally from customers with long inactivity (At Risk/Lost). This segmentation is directly actionable. Champions represent the customers most likely to generate consistent future revenue, so retention strategies (loyalty offers, priority service, exclusive promotions) should focus there. At Risk customers should receive reactivation efforts (targeted discounting, win-back campaigns) but those efforts should be measured carefully given the earlier discount–profit findings. The Frequency vs Monetary visualization (Figure 20) also helps interpret "Big Spenders" patterns—customers who contribute high monetary value but may not purchase frequently. This can guide relationship strategies such as personalized outreach or premium bundles.

An important contribution of this project is the integration of these outputs into a unified interface (Figure 21). Instead of requiring separate notebooks for EDA, forecasting, and segmentation, the dashboard enables interactive exploration and decision support. Filters (date range, category, region, segment) allow users to test business questions dynamically, such as whether discount–profit behaviour changes by segment, or whether forecasting reliability shifts when focusing on a subset of the business.

At the same time, several methodological limitations should be considered when interpreting results. First, the forecasting approach is intentionally lightweight and emphasizes interpretability rather than maximum predictive accuracy. It does not include external regressors such as promotions, holidays, macroeconomic indicators, or competitor dynamics, which are known drivers of retail demand. Second, the dataset spans 2014–2017, which limits generalization beyond this period. Third, K-Means clustering assumes spherical clusters and relies on the fixed choice of k=4 for interpretability; alternative clustering methods (e.g., hierarchical or Gaussian mixture models) could produce different segmentation boundaries.

These limitations do not invalidate the results, but they define how the dashboard should be used: as a strong decision-support tool for exploration and short-term planning, not as a fully optimized enterprise forecasting system.

Overall, the project achieves its objective: it transforms transactional retail data into a structured, interactive analytics system that supports descriptive monitoring, predictive planning, and customer intelligence. The strongest practical contribution is not any single chart, but the combined workflow that enables managers to understand performance drivers (category/discount), anticipate near-term demand (forecasting), and prioritize customer strategy (RFM clusters) within one coherent framework.

Limitations

- Forecasting does not include external variables
- Dataset limited to 2014-2017
- Static clustering (k = 4)
- No demographic segmentation

Future research could integrate macroeconomic indicators and dynamic clustering technique.

Conclusion

The project successfully demonstrates the transformation of transactional retail data into a predictive segmentation-enabled Business Intelligence dashboard. The integration of forecasting and customer analytics elevates the system beyond descriptive reporting toward actionable strategic insight.

References

Box, G. E. (1976). *Time series analysis: Forecasting and control.* San Francisco: Holden-Day.

Chen, H. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188. doi:https://doi.org/10.2307/41703503

Fader, P. S. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4), 415–430. doi:https://doi.org/10.1509/jmkr.2005.42.4.415

Hyndman, R. J. (2018). *Forecasting: Principles and practice (2nd ed.).* OTexts. doi:https://doi.org/10.5072/zenodo.593231

James, G. W. (2013). *An introduction to statistical learning: with applications in R.* New York: Springer.

Kuhn, M. &. (2019). *Feature engineering and selection: A practical approach for predictive models.* Boca Raton, FL: CRC Press / Taylor & Francis.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, (pp. 281–297). Berkeley, CA.

Makridakis, S. S. (2020). The M4 competition: Results, findings, and conclusions. *International Journal of Forecasting*, 36(1), 54–74. doi:https://doi.org/10.1016/j.ijforecast.2019.04.014

Montgomery, D. C. (2015). *Introduction to time series analysis and forecasting (2nd ed.).* Hoboken, NJ: Wiley. doi:https://doi.org/10.1002/9781118745113

Moubray, J. (1997). *Reliability-centred maintenance.* Oxford: Butterworth-Heinemann.

Sharda, R. D. (2018). *Business intelligence, analytics, and data science: A managerial perspective (4th ed.).* New York, NY: Pearson.

Wedel, M. &. (2000). *Market segmentation: Conceptual and methodological foundations (2nd ed.).* New York, NY: Springer. doi:https://doi.org/10.1007/978-1-4615-4651-1