INFS 5096 Customer Analytics in Large Organisations

Assignment 1- Supermarket Data Analysis

Q0 – BUSINESS CASE AND DATASET INTRODUCTION

Business Case Introduction

Supermarkets rely on understanding customer behaviour to maintain loyalty, predict demand, and decide where to invest in promotions. In today's competitive retail environment, the ability to turn sales data into insights is essential for effective decision-making. The supermarket dataset we analyse provides exactly this opportunity. By exploring the transactions of thousands of customers across multiple years, we can answer three important management questions:

- How do we categorise our regular customers into meaningful groups based on their value and behaviour?
- What can past sales patterns tell us about expected revenue in the early months of 2016?
- How much of a difference do promotions make to product sales, and are they worth the cost?

Addressing these questions helps managers allocate marketing resources more effectively, plan stock levels and staffing, and design promotions that genuinely increase revenue rather than simply shifting sales between time periods.

Dataset Introduction

The data comes directly from the supermarket's checkout system. Every time a customer purchased an item between 2013 and 2015, the system recorded it as a transaction line. This means that if one customer bought five items in a single visit, there would be five separate lines linked to the same receipt. Each line includes key details such as:

- Date and time of purchase – showing when the shopping trip occurred.
- Receipt number – a unique code that groups all items bought in one visit.
- Customer number – identifying each shopper through a loyalty card.
- Product details – such as department name, item value, and quantity.
- Promotion flag – showing whether the product was purchased under a special offer.

Because this is real-world supermarket data, it includes imperfections. For example, some loyalty cards were "generic" and used by many people, creating a few "super customers" with unrealistically high spending. Transactions also included very large negative amounts that represent supplier payments rather than customer purchases.

To make the dataset usable, we applied sensible cleaning steps: removing supplier payments, excluding the generic cards, and keeping only genuine purchases and returns. We ensured that key variables such as sales dates and transaction amounts were in the correct format, removed rows with missing customer IDs or missing dates, and eliminated duplicate records to avoid double counting. We also filtered out transactions with zero quantities, trimmed extreme positive outliers that did not reflect normal supermarket purchases, and standardised categories such as department names and promotional offers to ensure consistency. Together, these steps produced a reliable dataset that reflects real customer behaviour and provides a solid basis for segmentation and further analysis.

After this preparation, the dataset is ready for analysis and covers the following scope:

Year	Trading_Days	Customers	Trips	Line_Items	Total_Sales
2013	362	13513	448671	11304494	44546893
2014	362	13927	302990	11225349	45999603
2015	349	13702	711067	10631862	43686745

Table 1: Yearly Dataset Overview (2013–2015)

From 2013 to 2015, the supermarket traded almost every day, with slightly fewer days in 2015 with only public holidays missing from the records. The number of customers was steady at around 14,000 each year. Shopping trips varied: they dropped sharply in 2014 before more than doubling in 2015. Item volumes were stable in 2013 and 2014 but dipped in 2015, showing smaller baskets. Sales peaked in 2014 at about A$46m, compared with A$44.5m in 2013 and A$43.7m in 2015. Overall, customers were stable, but shopping frequency and spending patterns shifted year to year.

Metric	Value
Trading days (2013–2015)	1,073
Unique customers	16,493
Shopping trips (receipts)	1,364,145
Line-items (rows)	33,161,705
Total sales (sum of Item_Value)	$134,233,242

Table 2: Overall Dataset Summary (2013–2015)

The combined dataset spans 1,073 trading days across 2013–2015, reflecting nearly continuous supermarket operations aside from public holidays. Within this period, the supermarket engaged 16,493 distinct customers, highlighting a stable and sizeable loyalty base. These customers made about 1.36 million shopping trips, resulting in more than 33 million line-items recorded at checkout. In monetary terms, this translates to total sales of approximately A$134 million over three years. The scale of transactions and breadth of coverage make this dataset well-suited for examining customer behaviour, sales patterns, and the effects of marketing activities.

Q1. MARKET SEGMENTATION ON 2014 REGULARS

Customer segmentation offers a practical solution by dividing shoppers into meaningful groups. This approach enables businesses to recognise distinct patterns of behaviour, better understand customer expectations, and tailor products or services to meet those needs more effectively (Das & Nayak, 2022).

For this analysis, transactional data from **2013–2015** is used. The year **2014** is the focus, but only customers who also purchased in **both 2013 and 2015** are retained. This step removes casual visitors and one-time shoppers, leaving a group of "regular" customers whose behaviour can be studied with greater confidence.

a) RFM Segmentation (Recency, Frequency, Monetary):

Here, each customer is evaluated on three dimensions: how recently they made a purchase, how frequently they visited, and how much money they spent. Scores are assigned to create categories such as **Champions, Loyal Customers, Potential Loyalists, and At Risk.** This method is straightforward, widely used in retail analytics, and provides direct insight into customer value and loyalty. A few examples are given in the table below.

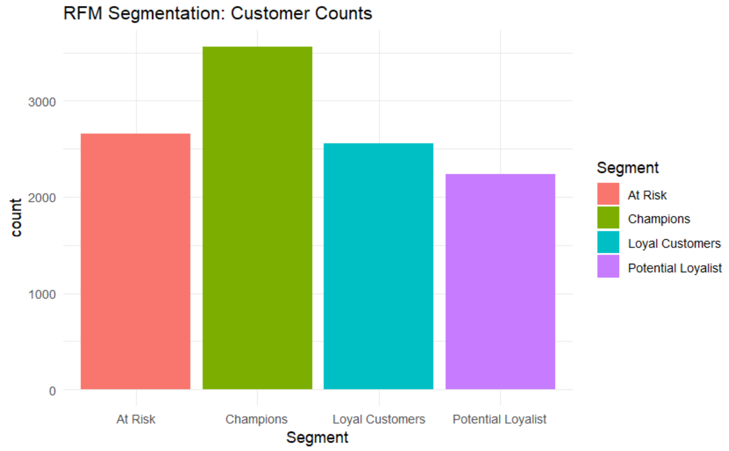
UniSA_Customer_No	Recency	Frequency	Monetary	R_score	F_score	M_score	RFM_Score	Segment
0000299827	0	133	6586.0800	4	5	4	13	Champions
0001925237	1	174	13612.4192	4	5	5	14	Champions
0002984536	8	6	742.8900	3	1	2	6	Potential Loyalist
0003004639	20	19	1824.3800	2	2	3	7	Potential Loyalist
0005225820	14	17	171.3100	2	2	1	5	At Risk

Table 3: RFM Segmentation of Regular Customers Based on Recency, Frequency, and Monetary Value

Interpretation of RFM Segmentation Results

Segment	n_customers	total_sales	avg_sales
At Risk	2658	1045929	393.5021
Champions	3558	29364583	8253.1150
Loyal Customers	2545	9817457	3857.5471
Potential Loyalist	2230	3531626	1583.6887

Table 4: Summary of RFM customer segments in 2014, showing number of customers, total revenue contribution, and average spend per customer.



The RFM analysis divides customers into four key groups.

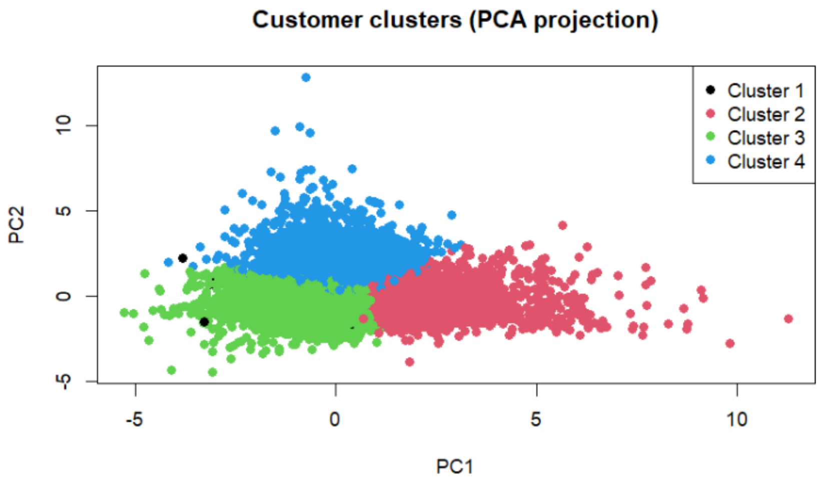
- **Champions**: 3,558 customers, contributing **A$29.4M** in sales, with an **average spend of ~A$8,250**. They are the most profitable group and need strong retention strategies
- **Loyal Customers**: 2,545 customers, generating **A$9.8M** in revenue, averaging **~A$3,860** each. They are steady contributors who can be further encouraged through rewards.
- **Potential Loyalists**: 2,230 customers, spending **A$3.5M** in total, averaging **~A$1,580** each. They show promise and can be developed into higher-value customers with targeted incentives.
- **At Risk**: 2,658 customers, contributing only **A$1.0M** overall, with a low **average spend of ~A$395**. They are disengaged and need reactivation campaigns.
- **Plot Insight**: The bar chart shows Champions as the largest and most valuable group, while At Risk customers are numerous but contribute little to overall sales.

b) Behaviour-based Clustering:

A broader feature set is developed to capture shopping habits, including average basket size, spend per trip, overall spending and visit frequency. These measures are standardised and grouped using k-means clustering to reveal distinct customer profiles or "personas" that describe shopping intensity and product preferences beyond simple RFM scores. To decide on the best number of clusters, here evaluated different values of **k** using the **Elbow** and **Silhouette** methods. Both methods indicated that **four clusters** strike a good balance between accuracy and interpretability. This choice also produced groups that are distinct and practical for business insights, making **k=4** the final selection.

cluster	**avg_basket_size**	**avg_spend_trip**	**total_spend**	**trips**
1	8.459161	40.83900	1127.329	48.12500
2	37.154557	156.67751	5139.281	34.89142
3	12.005327	49.71032	1707.831	36.62793
4	14.453117	60.35419	9124.931	170.88680

Table 5: Customer Clusters Based on Basket Size, Spending, and Trip Behaviour



- **Cluster 1 – Frequent Low Spenders**
  Customers in this group shop often but purchase only small baskets with low spend per trip. Their overall contribution to sales is limited, even though they visit regularly.
- **Cluster 2 – Bulk Buyers**
  These customers shop less frequently but make very large purchases when they do. Their high spend per trip makes them valuable, though their total trips remain few.
- **Cluster 3 – Mid-Level Shoppers**
  This group has moderate visit frequency and spends a reasonable amount per trip. They represent a stable but less engaged segment with growth potential.
- **Cluster 4 – Heavy Loyal Customers**
  The most important group, with very frequent visits and consistently high spending. They are the supermarket's main revenue drivers and should be prioritized for retention.

Comparison of RFM Segmentation vs. Clustering

The comparison between RFM and clustering shows that both methods complement each other. RFM highlights customer value and loyalty, while clustering uncovers shopping behaviours.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
At Risk	2	402	2,254	0
Champions	2	517	762	2,277
Loyal Customers	1	679	1,748	117
Potential Loyalist	3	511	1,716	0

Table 6: Cross-tabulation of RFM Segments and Clusters (Counts)

This table shows the number of customers in each RFM segment that fall into the four behavioral clusters. For instance, most At Risk customers belong to Cluster 3, while the majority of Champions are concentrated in Cluster 4. The raw counts highlight the scale of customer distribution across segments and clusters.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
At Risk	0.1%	15.1%	84.8%	0.0%
Champions	0.1%	14.5%	21.4%	64.0%
Loyal Customers	0.0%	26.7%	68.7%	4.6%
Potential Loyalist	0.1%	22.9%	77.0%	0.0%

Table 7: Percentage Distribution of RFM Segments Across Clusters

This table converts the counts into percentages within each RFM segment, making it easier to compare relative distributions. For example, nearly two-thirds of Champions belong to Cluster 4, while over three-quarters of Potential Loyalists are grouped in Cluster 3. These percentages highlight the dominant cluster patterns for each segment.

- **At Risk customers** – The majority (about 85%) are grouped in Cluster 3, showing they mostly share a mainstream but low-value shopping pattern.
- **Champions** – Nearly two-thirds (around 64%) are in Cluster 4, which highlights their distinct high-spending behaviour compared to other groups.
- **Loyal Customers** – Most (about 69%) also sit in Cluster 3, with a smaller share in Cluster 4, reflecting steady shoppers who are valuable but not at the level of Champions.
- **Potential Loyalists** – Over three-quarters (roughly 77%) belong to Cluster 3, indicating they behave like average shoppers but have the potential to grow into higher-value segments.

Overall, Cluster 3 represents the common shopping style for most customer groups, while Cluster 4 is the home of high-value Champions. This shows RFM identifies who is valuable, and clustering explains how they shop.

Q2. MONTHLY SALES ANALYSIS AND FORECASTING

Accurate sales forecasting is essential for supermarkets to optimise inventory and plan promotional strategies. Monthly sales are shaped not only by customer demand but also by calendar factors such as trading days, weekends, and seasonal variations. In this study, three years of transaction data (2013–2015) were aggregated to the monthly level to capture these dynamics.

To model these patterns, a multiple linear regression approach was applied. The target variable was monthly sales in dollars, while explanatory variables included the calendar year, the specific month, the number of days in each month, and the number of weekend days. This framework allowed for both structural and seasonal influences to be incorporated into the analysis, leading to sales predictions for January, February, and March 2016, with corresponding error margins to quantify forecast uncertainty.

Regression Model Results

Predictor	Estimate	Std. Error	t-value	p-value	Significance
(Intercept)	74,355,811	131,932,362	0.564	0.577	Not significant
Year	-35,840	65,501	-0.547	0.588	Not significant
Month	46,670	16,074	2.903	0.0067	** Significant (1% level)
DaysInMonth	52,107	66,211	0.787	0.437	Not significant
WeekendDays	-38,611	70,618	-0.547	0.588	Not significant

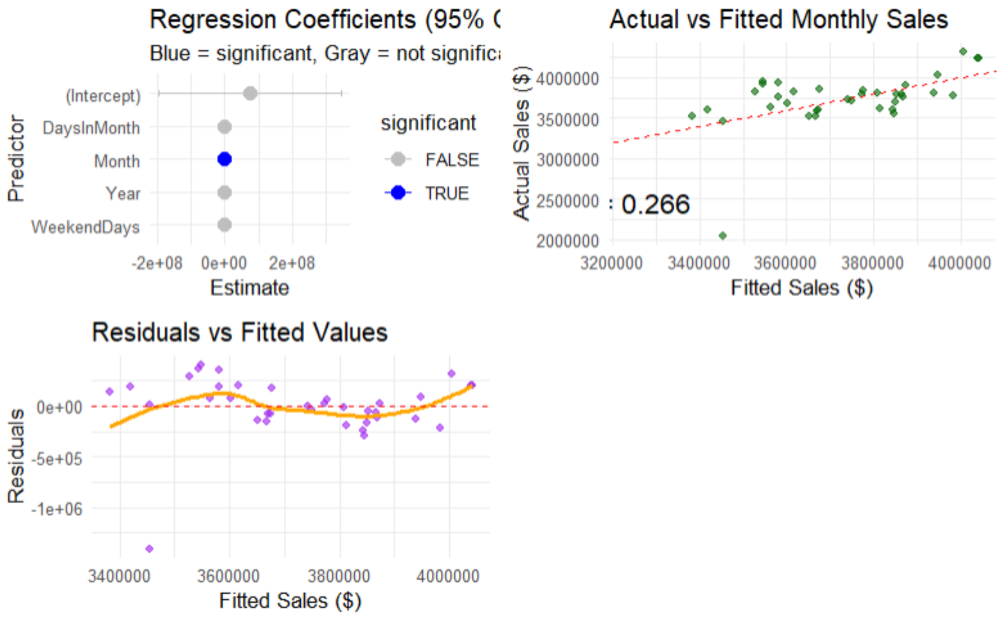
Table 8: Regression Coefficients for Monthly Sales Model

Model summary

- Residual standard error: **320,900 (df = 31)**
- Multiple R-squared: **0.266**
- Adjusted R-squared: **0.171**
- F-statistic: **2.807 (p = 0.0425)**

Interpretation

- **Month** is the only statistically significant predictor of sales (**p = 0.0067**). This indicates a strong seasonal pattern, with certain months consistently driving higher or lower revenue, likely due to consumer shopping cycles and seasonal promotions.
- **Year** is not significant (**p = 0.588**), suggesting that across 2013–2015 there was no clear upward or downward trend in monthly sales once seasonal variation is taken into account.
- **Days in the Month** is not significant (**p = 0.437**), showing that having more calendar days in a month does not automatically lead to higher sales volumes.
- **Weekend Days** is also not significant (**p = 0.588**), indicating that the number of weekends in a month does not systematically influence overall sales when compared to weekdays.
- The model explains about **26.6% of the variation in monthly sales (R^2 = 0.266)**, meaning that calendar effects alone account for only part of sales variation, while other factors such as promotions, holidays, or economic conditions likely play a larger role.
- Despite its modest explanatory power, the overall regression is **statistically significant (F = 2.807, p = 0.0425)**, confirming that the predictors collectively contribute to explaining sales variation.



Regression Coefficients Plot

- The plot highlights Month as the only variable with a coefficient that is clearly different from zero, confirming its role as a significant driver of sales.

- Year, Days in Month, and Weekend Days remain close to zero with wide confidence ranges, consistent with their non-significance in the regression table.

Actual vs Fitted Sales Plot

- The scatter shows an overall upward trend, but with visible spread around the fitted line.

- This matches the modest explanatory power of the model ($R^2 \approx 0.27$), meaning seasonality is captured, but other influences are missing.

- The closeness of many points to the line indicates the model reasonably reflects monthly patterns, though external factors such as promotions or holidays are not fully explained.

Residuals vs Fitted Plot

- Residuals are mostly centered around zero, though the smoother suggests some minor non-linear patterns not captured by the model. This reinforces the modest R² value.

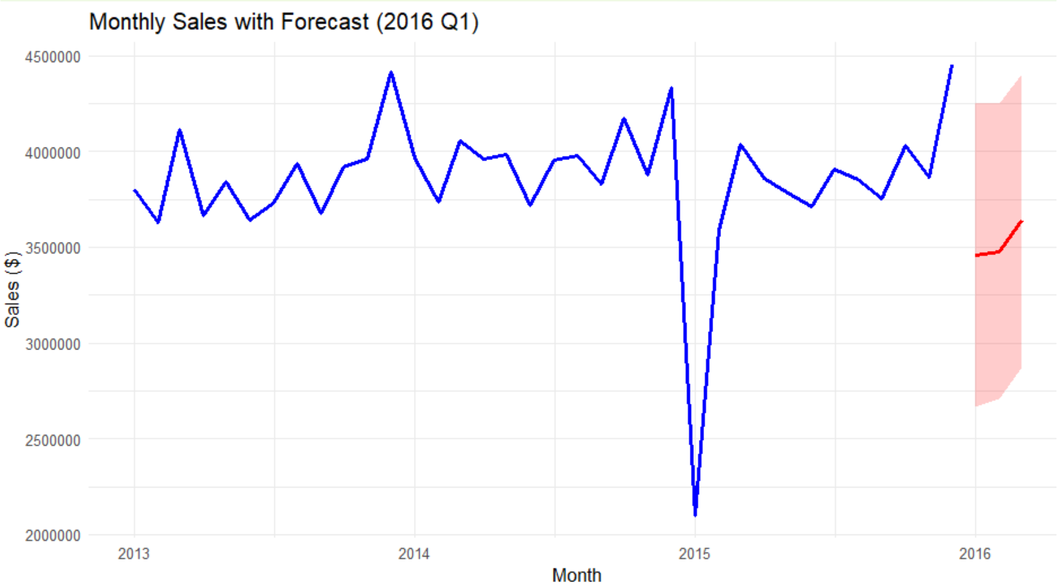
Predicted Sales for Jan–Mar 2016

Month (2016)	Predicted Sales ($)	95% Lower Bound	95% Upper Bound
January	3,379,251	2,612,089	4,146,413
February	3,398,929	2,653,485	4,144,372
March	3,549,813	2,809,586	4,290,041

Table 9: Forecasted sales values with lower and upper prediction limits.

- **January 2016**: Sales are forecast at about $3.38 million, with possible values ranging between $2.61 million and $4.15 million. This gives an **error margin of ±$0.77 million**.
- **February 2016**: Expected sales are close to $3.40 million, similar to January, with an interval of $2.65 million to $4.14 million. The **error margin here is ±$0.75 million**.
- **March 2016**: Predicted sales rise to around $3.55 million, with a range between $2.81 million and $4.29 million. The **error margin is ±$0.74 million**.

Overall, the first quarter of 2016 is projected to show stable sales, with March standing out as the strongest month.



- **Blue line:** Shows past monthly sales from 2013–2015, mostly between $3.5M and $4.0M.
- **Red line:** Forecast for Jan–Mar 2016 indicates steady sales, starting near $3.45M and rising to around $3.6M.

- **Shaded band:** Represents prediction intervals, ranging roughly from $2.7M to $4.4M, capturing forecast uncertainty.

  The plot shows steady sales predictions for early 2016, with March the strongest month, while the shaded range illustrates expected variability around the forecasts.

In conclusion, the regression analysis shows that supermarket sales are mainly shaped by seasonality. The month variable was the only significant predictor, highlighting strong recurring patterns tied to festive periods and consumer habits. Other factors such as year, number of days in the month, and weekends were not significant, suggesting they do not systematically influence sales.

Although the model explains about 27% of the variation in monthly sales, it remains statistically valid and confirms that seasonal cycles are an important driver. Forecasts for January–March 2016 predict steady sales of around $3.38M–$3.55M, with March projected as the strongest month. Error margins of about ±$0.74M–$0.77M provide a realistic range of expected values.

Overall, sales appear stable with clear seasonal effects, and management should account for these patterns while also considering promotions, holidays, and other external influences to improve planning and forecasting.

Q3. IMPACT OF PROMOTIONS ON ITEM SALES

Promotions are an essential element of the marketing mix, often used to stimulate demand and influence customer decisions. In retail settings, these may involve direct price reductions, product bundling, or enhanced product visibility through catalogues and in-store placement. While traditional marketing literature highlights price-based promotions as a way to temporarily reduce selling price and encourage purchases, supermarket promotions can extend beyond discounts(Niu et al., 2024).

In this analysis, all types of promotions recorded in the dataset were assessed for their influence on the number of items sold. The approach involved aggregating daily item-level transactions and applying a quasi-Poisson regression model, which is appropriate for modelling sales counts. The model included not only the promotion variable but also price, department, day of the week, and month, allowing the effect of promotions to be examined while controlling for pricing differences and seasonal shopping behaviours.

Key Regression Results on Units Sold

Variable	Estimate	Std. Error	t-value	p-value	Key Insight
Promotion (Offer)	-0.0441	0.00237	-18.58	<0.001	Items under promotion were sold ~4.3% fewer units compared to non-promoted items.
Average Price (log)	-0.7626	0.00106	-721.65	<0.001	Higher prices sharply reduce unit sales, showing strong price sensitivity.
Fruit & Vegetables Dept	0.7910	0.00320	247.30	<0.001	This department consistently sells more items than the baseline category.

Fresh Meat Dept	0.7306	0.00438	166.68	<0.001	Strong positive demand: items in this category sell at much higher volumes.
Frozen Foods Dept	-0.4319	0.00458	-94.26	<0.001	Fewer units sold relative to baseline, suggesting lower customer preference or storage issues.
Grocery Dept	-0.8301	0.00287	-289.60	<0.001	Grocery products sell significantly fewer units compared to the reference group.
Day of Week (trend)	0.2547	0.00204	124.90	<0.001	Weekly shopping patterns are evident, with sales peaking closer to weekends.
Month (trend)	0.0588	0.00252	23.38	<0.001	Seasonal variations across months strongly influence sales volumes.

Table 10: Regression Results: Effect of Promotions and Other Factors on Unit Sales

Condition	**Expected Impact on Units Sold**
No promotion (baseline)	100% (reference level)
With promotion	~95.7% of baseline (**-4.3%**)

Table 11: Promotion Effect (Interpretation in % terms)

Effect of Promotions

- Promotions reduced sales volumes by approximately **4.3%** compared to non-promoted items.
- This suggests that not all promotional activities work as intended. Some may cause product substitution, encourage stockpiling earlier, or make customers perceive less value in the offer.

Price Sensitivity

- The strong negative coefficient for **log(Average Price)** confirms that sales volumes fall sharply when prices increase.
- Customers display high price elasticity, meaning their purchasing behavior is very sensitive to price changes.

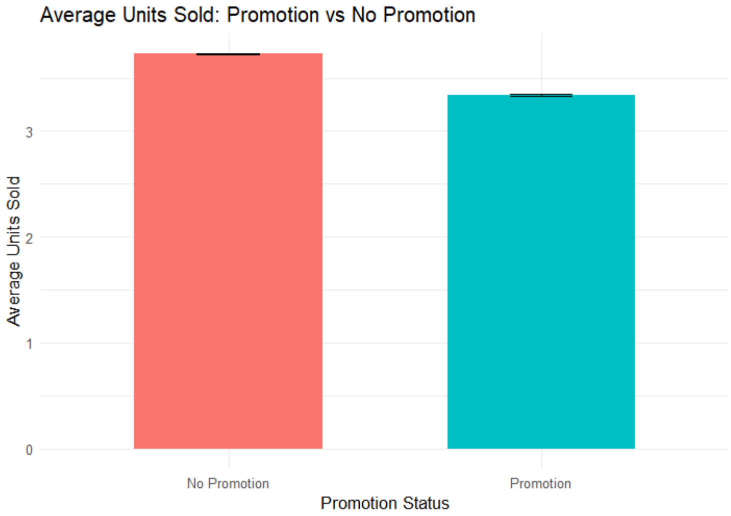
Category-Level Insights

- **Fruit & Vegetables (+79%)** and **Fresh Meat (+73%)** record substantially higher sales than the baseline, reflecting their role as essential household products.
- **Frozen Foods (-43%)** and **Grocery (-83%)** have much lower sales volumes, suggesting these categories may have different purchase cycles, storage limitations, or consumer habits.

Shopping Patterns

- **Day of the Week effect:** Sales increase closer to weekends, reflecting typical grocery shopping routines.

- **Month effect:** Seasonal cycles are evident, with higher demand during particular times such as holidays or festive periods.



- The chart shows that products on promotion were sold in slightly smaller quantities compared to those without promotion.
- On average, non-promoted items achieved higher unit sales.
- The confidence intervals are narrow, confirming that the difference, though modest, is meaningful.

The analysis shows that promotions did not deliver the expected lift in sales. Instead, promoted items sold around 4% fewer units than those without offers. Sales were more strongly influenced by price levels, essential categories like fresh produce and meat, and natural shopping cycles across weeks and months. The bar chart visually reinforces this finding, with non-promoted products achieving higher average sales than those under promotion. This means that promotions on their own are not a reliable way to increase volumes. A more effective approach would be to align pricing with customer sensitivity, focus on high-demand categories, and account for timing patterns such as weekends and holiday periods when planning campaigns.

Q4. CONCLUSION

The analysis of supermarket data from 2013 to 2015 offered valuable insights into customer behaviour, promotional impacts, and sales forecasts. Several themes emerged that help explain how customers interact with the store and what factors influence sales outcomes.

To begin with, customer segmentation revealed meaningful differences in shopping behaviour. Through **RFM analysis**, customers were grouped based on their recency, frequency, and monetary value, producing categories such as *Champions*, *Loyal Customers*, *Potential Loyalists*, and *At Risk*. This showed that a relatively small group of customers accounted for a large share of sales, while others displayed signs of declining engagement. In addition, **clustering with behavioural measures** like average basket size, spend per trip, and departmental preferences highlighted diverse shopping styles. For example, some clusters represented small but frequent buyers, while others showed large and varied purchasing patterns. Together, these approaches provided a clearer picture of customer diversity, offering practical opportunities for targeted loyalty campaigns, retention programs, and resource prioritisation.

The **promotion analysis** offered a surprising finding. Instead of increasing demand, promotions were linked to a reduction of about **4% in units sold**. This points to possible substitution effects, where customers shift purchases across time or categories, or to perceptions that discounts reduce the product's value. At the same time, the results confirmed the importance of **price sensitivity**, as higher unit prices were strongly associated with lower sales volumes. Differences across categories were also significant: products in *Fruit & Vegetables* and *Fresh Meat* showed consistently higher unit sales, reflecting everyday household needs, while categories such as *Frozen Foods* and *Grocery* reported lower sales relative to the baseline. In addition, **seasonal cycles** and **weekly patterns** strongly influenced demand, showing that natural shopping habits and calendar effects play a larger role than promotions alone.

The **sales forecast model** extended the analysis into 2016 by predicting monthly sales using regression with calendar-based predictors (year, month, days per month, and weekend counts). The model suggested relatively stable sales in the first quarter: approximately **$3.38 million in January**, **$3.40 million in February**, and **$3.55 million in March**. Each estimate carried an error margin of about ±$0.75 million, underscoring the inherent uncertainty in forecasting. These results identified March as the strongest month, while January and February were projected to remain steady. Including error margins provided a realistic view of possible fluctuations, supporting more informed planning and expectation management.

Overall, this study shows that supermarket performance is shaped by a mix of customer differences, pricing strategies, product category dynamics, and seasonal behaviour. Promotions, although widely used, do not necessarily guarantee higher sales and should be designed with greater care. Instead, management can achieve stronger results by focusing on **customer retention**, aligning prices with demand sensitivity, and considering seasonal and weekly shopping trends when making business decisions. By combining segmentation, promotional evaluation, and forecasting, supermarkets can adopt a more data-driven approach that balances short-term tactics with long-term strategy, ultimately supporting sustainable growth and improved performance.

REFERENCES

1. Niu, J., Jin, S., Chen, G., & Geng, X. (2024). How Can Price Promotions Make Consumers More Interested? An Empirical Study from a Chinese Supermarket. *Sustainability*, *16*(6), 2512–2512. https://doi.org/10.3390/su16062512
2. Das, S., & Nayak, J. (2022). Customer Segmentation via Data Mining Techniques: State-of-the-Art Review. *Computational Intelligence in Data Mining*, 489–507. https://doi.org/10.1007/978-981-16-9447-9_38