# **Supermarket Sales Analysis**

**Introduction**

This report presents an analysis of transaction data from a supermarket chain collected over three consecutive years: 2013, 2014, and 2015. The dataset encompasses daily transaction records from each trading day, excluding public holidays when the supermarket was closed. Each entry in the dataset corresponds to an individual sale transaction and includes detailed information of that transaction.

The datasets for each of the three years (2013, 2014, 2015) contain similar variables, structured around supermarket sales transactions. Here's an overview based on the rows from each dataset and their structure.

### **Variables Across Datasets**

Sale\_Date and Sale\_Time: The date and time of each transaction.

UniSA\_Receipt\_No1: A unique identifier for the receipt.

Receipt\_Line\_No: The line number on the receipt.

UniSA\_Customer\_No: A unique identifier for the customer (has some missing values).

Total\_Sale\_Amount\_InclusiveGST: The total sale amount for the transaction including GST.

Total\_Receipt\_Lines: The total number of items in the transaction.

Barcode\_Item: The barcode of the item sold (has some missing values).

Item\_Description: Description of the item.

Quantity\_Sold: Quantity of the item sold.

Item\_Value: Value of the item.

Department\_Name: The department of the item (has some missing values).

Commodity\_Name: The commodity or type of the item (has some missing values).

Offer: Information on whether the sale was part of a special offer.

Each year's dataset follows a similar format with 14 columns.

There are some missing values in UniSA\_Customer\_No, Barcode\_Item, Department\_Name, and Commodity\_Name, which will need to be considered during analysis.

### **Research Objective:**

The primary goal of this research is to analyse transactional data from a supermarket chain between 2013 and 2015 to gain insights into customer purchasing patterns and sales trends. This study focuses on identifying regular customers by aggregating 2014 data based on continued patronage across the three years, analysing metrics such as visit frequency, purchase volume, and total spending, and employing cluster analysis to potentially segment the market. Additionally, we will analyse monthly sales data from all three years to predict first quarter sales for 2016 using various forecasting techniques including time series analysis and regression models. Lastly, by examining the 2013 data, we aim to assess the impact of sales promotions on the volume of purchases, providing insights into how price changes influence consumer behaviour. This comprehensive analysis will help optimize marketing strategies and enhance business performance.

# **Data Preprocessing**

Prior to conducting our comprehensive analysis, the dataset underwent a crucial phase of data preprocessing to ensure accuracy and reliability. We encountered two significant issues that required rectification. Firstly, the presence of 'super customers' in the transactional data was identified. These outliers were initially perceived as customers making excessively large purchases. Upon further investigation, it was revealed that these were not actual customers, but transactions linked to generic loyalty cards used by staff when shoppers did not have or did not present a loyalty card. To address this anomaly, we disassociated these transactions from the analysis to prevent skewing our insights on customer behaviour and spending patterns.

Secondly, our data integrity checks revealed instances of negative values in the 'Quantity\_Sold' and 'Item\_Value' columns. Such entries are inconsistent with the nature of transactional sales data, as they may represent returns, refunds, or data entry errors. To maintain the integrity of our analysis, which focuses on sales rather than returns, we removed these negative entries from the dataset. This step was necessary to preserve the validity of our analysis, focusing on actual sales activity and customer spending patterns.

# **Question 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index** | **UniSA\_Customer\_No** | **Number\_of\_Trips** | **Total\_Purchases** | **Total\_Spent** |
| 1 | 299827 | 133 | 1735.0154 | 6599.91 |
| 2 | 1925237 | 174 | 2840.4109 | 13625.1692 |
| 3 | 2984536 | 6 | 166.9898 | 742.89 |
| 4 | 3004639 | 19 | 461.5841 | 1824.38 |
| 5 | 5225820 | 17 | 37.4722 | 171.31 |
| 6 | 5316222 | 1 | 4.61 | 32.69 |
| 7 | 5714956 | 54 | 299.7333 | 1427.04 |
| 8 | 6312990 | 41 | 925.373 | 3638.6 |
| 9 | 6652575 | 1 | 26.16 | 82 |
| 10 | 7180839 | 32 | 247.8683 | 1211.96 |
| 10997 | 9990742549 | 44 | 416.6245 | 1768.5706 |
| 10998 | 9991166278 | 99 | 1235.4566 | 4563.34 |
| 10999 | 9993661500 | 32 | 1049.2876 | 3754.16 |
| 11000 | 9993879211 | 6 | 18.693 | 80.94 |
| 11001 | 9993952646 | 29 | 565.6243 | 2100.64 |
| 11002 | 9994391298 | 9 | 284.0341 | 1449.58 |
| 11003 | 9994971948 | 61 | 822.0876 | 2673.2594 |
| 11004 | 9996877212 | 5 | 90.4837 | 540.29 |
| 11005 | 9997118084 | 85 | 1166.3321 | 4281.4 |
|  | **Total** | **814260** | **11517428.25** | **47633829.11** |

The output above, reveals the first and last ten entries of subset of **11,005 regular customers** who shopped at the supermarket consistently across the years 2013, 2014, and 2015. The data reveals significant variations in customer engagement levels. For instance, Customer ID 299827 visited the supermarket 133 times in 2014, purchasing items worth a total of $6,599.91. This indicates a high level of engagement and loyalty. On the other end of the spectrum, Customer ID 5316222 made just one trip, spending a mere $32.69.

Across all regular customers in 2014, there were **814,260 trips** to the supermarket, resulting in **purchases of approximately 11,517,428** items, and total **spending was around $47,633,829.** This underscores the substantial impact regular customers have on the supermarket's overall performance.

A pie chart with numbers and text

Description automatically generated

The donut chart above gives the insights of proportion percentages of different departments.

* In an analysis of departmental sales within the supermarket, it is apparent that the Grocery department is preeminent, commanding 47.55% of total purchases, thereby affirming its critical role in driving the supermarket's revenue and serving as a primary shopping destination. Close behind is the Fruit & Veg department, which garners 16.03% of purchases, reflecting a consumer preference for fresh produce and highlighting its significance in attracting consistent shopper traffic. The Dairy department also shows strong performance with 12.29% of sales, indicating sustained consumer demand for dairy products.
* The Bakery department, with a 7.38% share of sales, suggests a reliable demand for its offerings. The Fresh Meat department, holding a smaller yet substantial 4.40% of the sales, signifies its lesser prominence relative to other food categories. Departments such as Variety and Frozen, with respective shares of 4.28% and 3.76%, cater to niche markets but are notable contributors to the overall sales mix.
* The Deli, with a 2.35% sales proportion, points to a specialized demand among the supermarket's clientele. Meanwhile, Seafood and Poultry, at 1.29%, indicates a more selective customer base. The Tobacco department, accounting for the smallest slice at 0.66%, is suggestive of the least frequent purchases, which may be influenced by factors such as public health awareness and pricing strategies.

**Cluster Analysis**

**A pie chart with numbers and a number of data points with Crust in the background

Description automatically generated**

The provided pie chart illustrates the distribution of customer segments based on a cluster analysis, which grouped customers according to their number of trips to the supermarket and money spent.

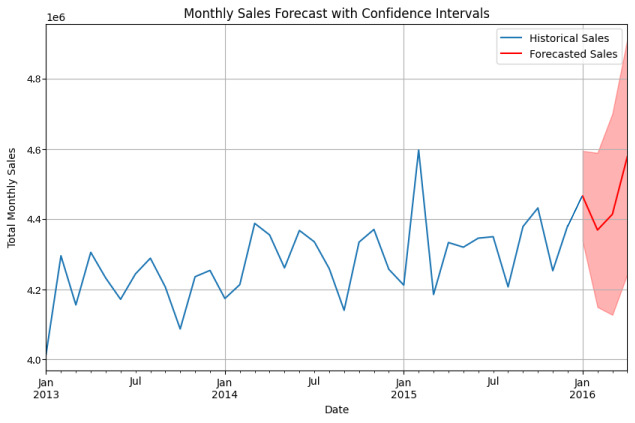
Cluster 0 which is represented by colour green comprising 10% of the dataset, may represent customers with a high number of trips and high expenditure, which could be indicative of highly engaged and possibly more affluent customers.

Cluster 2, in orange, accounts for a significant 37.7% of the dataset and represents what might be considered the core customer base those who shop with moderate frequency and have average-sized transactions.

The largest segment, Cluster 3, is indicated in blue and makes up 44% of the data, likely composed of customers who visit the supermarket regularly but spend less per visit, which could be reflective of budget-conscious individuals or smaller households. Lastly, Cluster 4, marked in red and making up 8.2% of the dataset, could indicate a less common shopping behaviour, such as customers who make seasonal visits or prefer to purchase in bulk.

We can also see the cluster 1 which is contributing 0.0 % that means there is an outlier in the data which can be ignored in this analysis.

# **Question 2**

****

|  |  |
| --- | --- |
| **Month** | **Sale** |
| 2016-01 | 4.46E+06 |
| 2016-02 | 4.38E+06 |
| 2016-03 | 4.58E+06 |

The graph and table provided illustrates the historical monthly sales from January 2013 to December 2015, along with forecasted sales for January, February, and March of 2016. The predictions are represented by the red line, and the shaded area around this line indicates the error margins for the expected values, providing a range where the actual sales figures are expected to fall. ARIMA model was used for this forecasting, and it was set to (7, 2, 4).

* **January 2016:** Estimated sales of approximately $4.46 million.
* **February 2016:** Estimated sales of approximately $4.38 million.
* **March 2016:** Estimated sales of approximately $4.58 million.

The prediction for January 2016 shows a stable start to the year when compared to the historical data, which may reflect the typical post-holiday sales period. The slight decrease in February's sales forecast might reflect a seasonal normalization following January's shopping activities. The slight decrease in February's sales forecast might reflect a seasonal normalization following January's shopping activities. Historically, February often shows a dip after the surge in January. The increase for March is notable and could be attributed to various factors such as the onset of spring.

# **Question 3**

Our analysis explores the relationship between sales promotions and purchasing patterns, focusing on mangoes, bananas, and lettuce in 2013. During a year of stable prices, any notable shifts likely reflect promotional activities. By examining these price movements, we aim to understand how discounts and pricing strategies impact sales volume for these key products.

A graph showing a line going up

Description automatically generated

Fig 1

In the line graph in Fig 1above, the price of mangoes appears relatively stable until around May when a significant increase is observed. This peak could suggest a sales promotion that temporarily increased the price, potentially as part of a strategy to make a subsequent discount seem more appealing. Following this peak, the price drops below the previous level, which may indicate the promotion period where the price is reduced to stimulate sales.

Another notable peak occurs in September, with the price returning to its pre-promotion level shortly thereafter. The pattern of peaks and troughs in the graph could represent repeated promotional cycles.

A graph showing a line of bananas

Description automatically generated

Fig 2

The price trend for bananas is characterized by high variability with several peaks and drops throughout the year as shown in the line graph above in Fig 2. Such fluctuations may be related to promotions or could be attributed to supply issues or changes in demand.

The dotted line, potentially representing the average price over the year, crosses through many of these fluctuations, which may indicate that the promotions did not significantly deviate from the average price level.

A graph showing a line of blue lines

Description automatically generated with medium confidence

Fig 3

By examining the line graph in Fig 3 above, lettuce prices show a similar pattern of variability as seen with bananas, with frequent changes in pricing. These could correlate with sales promotions aimed at increasing the volume of sales.

The frequent ups and downs in the price trend may also suggest that lettuce is a product with more elastic demand, where price changes are used to manage sales volumes.

# **Conclusion**

The in-depth analysis of the supermarket's transaction data from 2013 to 2015 has culminated in substantive insights. Notably, the behavior of 11,005 regular customers indicates a diverse range of engagement, with an astonishing 814,260 total trips made and a significant $47,633,829 spent in 2014 alone. Such figures underscore the value of customer loyalty and the potential returns on investment in targeted retention strategies.

Departmental sales analysis revealed that the Grocery, Fruit & Veg, and Dairy departments are central to revenue, with the Grocery department alone accounting for 47.55% of total purchases. This quantitative insight into departmental dominance could direct marketing resources and inventory optimization efforts.

Through cluster analysis, customers were segmented into distinctive groups, where Cluster 3 (44%) represented frequent but smaller-spending shoppers, and Cluster 0 (10%) likely reflected a segment with high spending and visit frequency. Such differentiation offers a blueprint for tailored marketing initiatives.

Our predictive models projected a positive sales trajectory for early 2016, with January expected to bring in around $4.46 million, February slightly less at approximately $4.38 million, and March anticipating a rise to about $4.58 million. These forecasts are pivotal for planning and demonstrate the value of predictive analytics in retail.

The price trend analysis for select products during the promotions confirmed the impact of pricing strategies on sales volumes. For instance, mangoes experienced a significant price increase in May, followed by a decrease, suggestive of promotional activity. Meanwhile, bananas showed a high price variability throughout the year, and lettuce prices depicted frequent changes, possibly in response to elastic demand.