Question 2

# Introduction

This report details the analysis of chip purchasing behavior using two datasets: transaction\_data and purchase\_behaviour. The primary aim is to uncover purchasing patterns, identify customer segments, and provide data-driven strategic recommendations for the retail store's chip category for the next six months. The approach involved cleaning and merging the datasets, performing exploratory data analysis (EDA), and finally offering actionable insights.

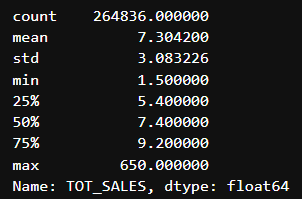
## Data Cleaning and Preparation:

The analysis began by examining both the transaction\_data and purchase\_behaviour datasets for inconsistencies, missing data, and outliers.

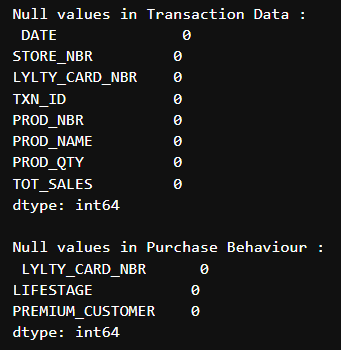
### Missing Data:

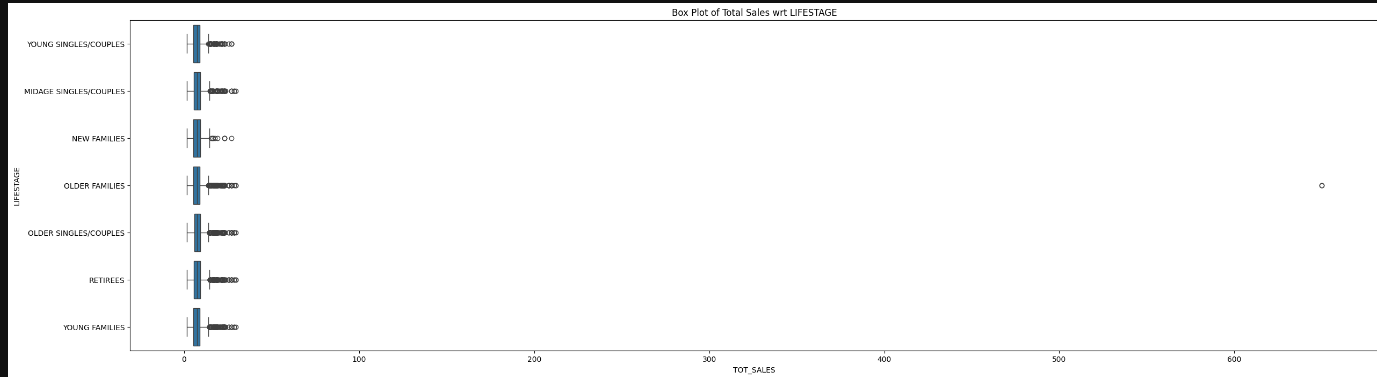
Both datasets were checked for null values. Any missing data in essential fields, needs to either be imputed with appropriate values or removed to maintain the integrity of the dataset.

Describing the dataset “transaction\_data.xlsx” to get more insight:

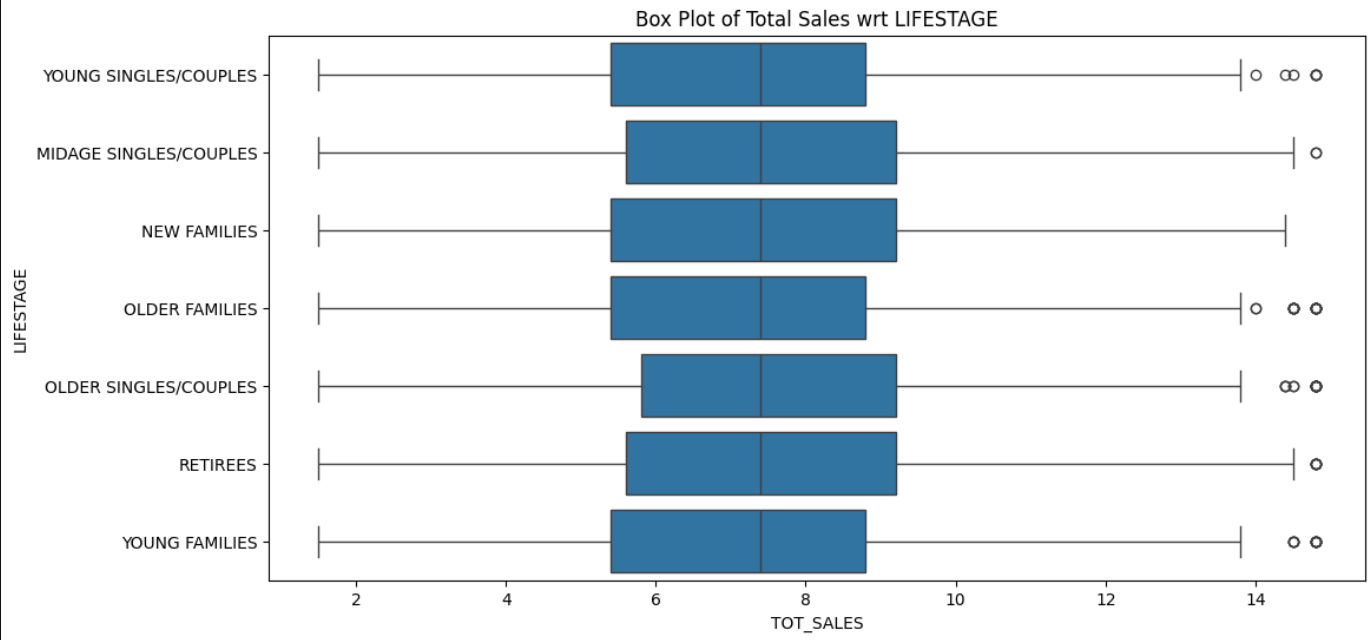


Checking for NULL values in both the datasets:

 As we can see there are no Missing values in the datasets, therefore there is no need to interpolate or remove any values from the datasets.

**Outliers**: Outliers were identified and removed to prevent them from skewing the analysis results. This was done using the Interquartile Range (IQR) method, focusing on TOT\_SALES to ensure reasonable values for all transactions.

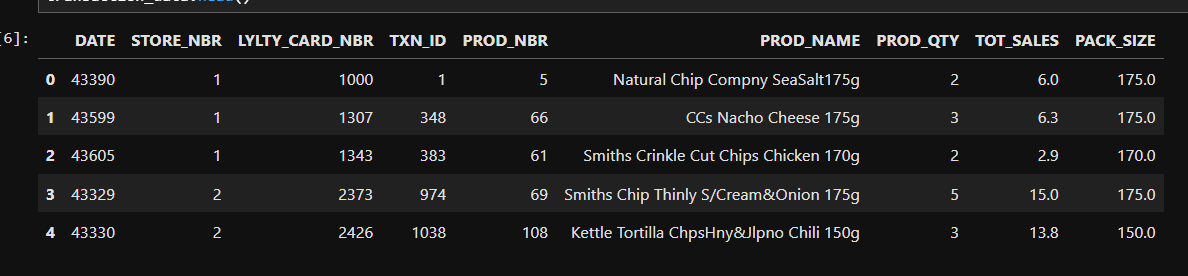
*Created a box plot for temperature across different years before removing outliers*



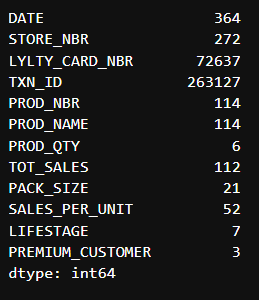
*Box plot for temperature across different years after removing outliers*

**Data Formats**: The data formats for numerical fields (e.g., sales values, pack sizes) were standardized, ensuring consistency across all records. Categorical variables, such as brand names and customer segments, were checked to ensure proper labeling.

### Feature Engineering

1. Extract PACK\_SIZE from the Product Name
2. Calculate various values such as 'Average prices' and 'Sales per product unit'

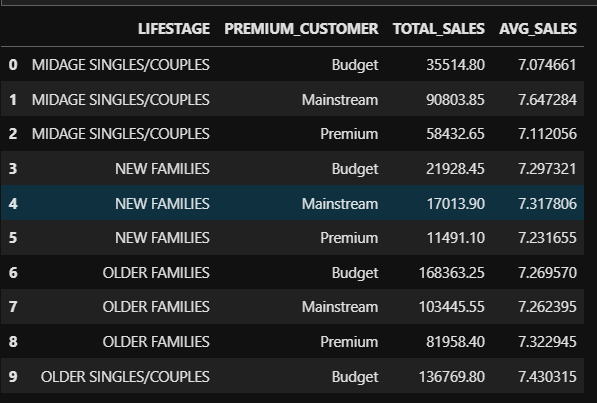
Analysing unique datapoints for merged dataset:



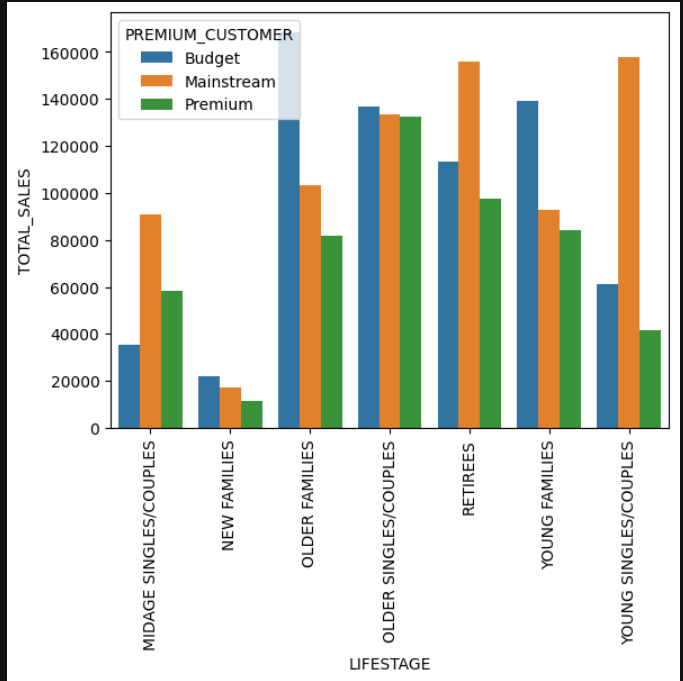
### Exploratory Data Analysis (EDA)

The next phase involved analyzing purchasing behavior by calculating key metrics such as total sales, sales drivers, and pack size preferences.

* **Total Sales and Sales Drivers** Total sales were calculated across different customer segments to identify which segments contributed most to chip sales. The segments were based on LIFESTAGE and PREMIUM\_CUSTOMER.



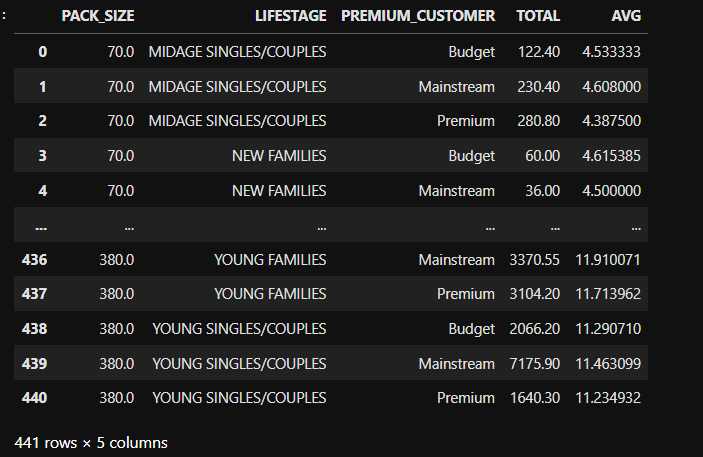
Plotting these over a boxplot for visualization:

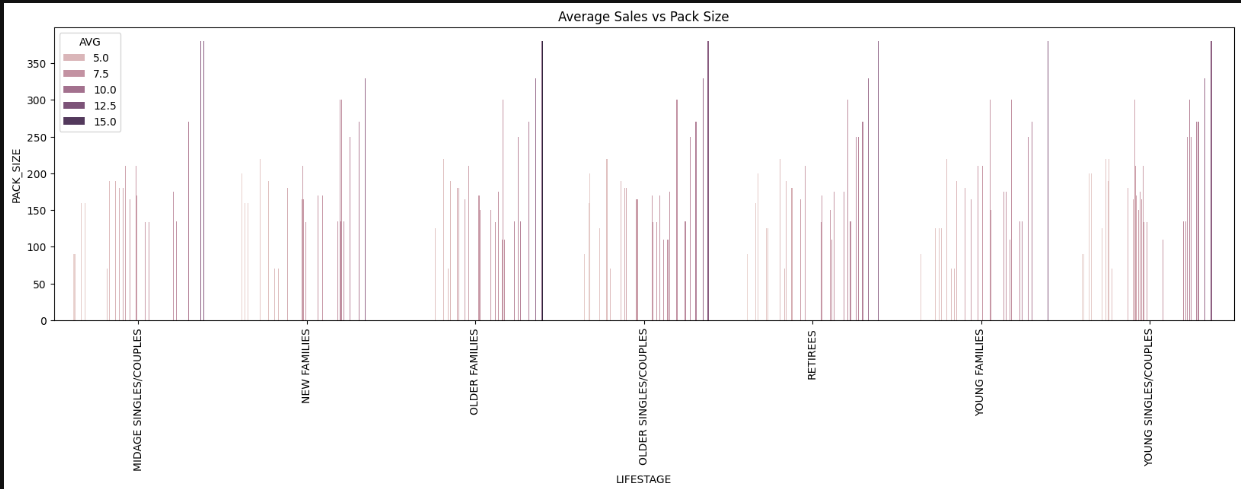


**Key Insight:**

1. Older Families with young children in the budget category were the highest spenders on chips.
2. Young singles and couples in the economy segment preferred smaller pack sizes and were more price-sensitive, favoring economy brands.

Analysing Correlation of PACK\_SIZE and plotting it using a barplot.





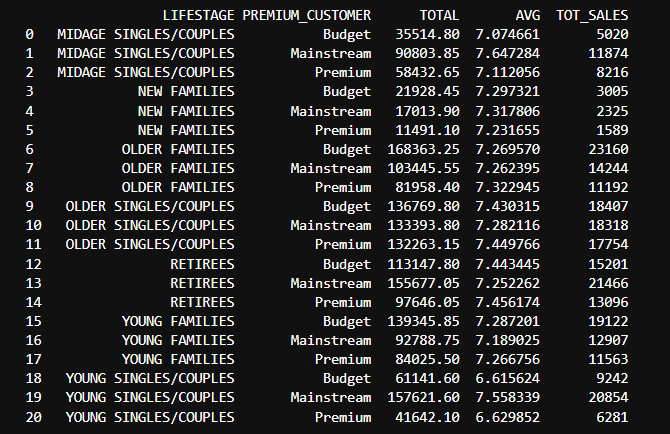
**Key Insight:**

1. Large pack sizes were preferred by family-oriented segments, especially those in the Older Families.

Customer Segmentation and Trends

Customer segmentation based on Lifestage and Premium Status revealed distinct purchasing patterns. By segmenting customers into categories such as young singles and couples, families with young children, and older families, we could identify variations in purchasing habits across these groups.

* **Average Sales per Customer Segment** The analysis showed that premium customers, particularly older families, had the highest average sales per transaction.



**Key Insight:** Premium older families and families with young children were the most valuable customer segments in terms of total revenue. Economy young singles and couples preferred smaller pack sizes and accounted for lower overall revenue, but had frequent, smaller transactions.

# **Summary**

This project involved the analysis of two primary datasets:

1. Transaction Data: This dataset includes detailed sales transaction records, encompassing product names, quantities sold, total sales amounts, and various other attributes.
2. Customer Data: This dataset provides information about customers, such as their life stage and premium customer status.

**Data Cleaning Process:**

To ensure the integrity and quality of the data, the following steps were undertaken:

1. Outlier Treatment: Outliers were identified and managed using the Interquartile Range (IQR) method, minimizing their potential impact on the analysis.
2. Missing Value Imputation: Missing values were addressed by imputing medians where applicable, ensuring the dataset remained robust for subsequent analyses.
3. Feature Extraction: Additional features were derived, such as the extraction of PACK\_SIZE from product names, and the calculation of SALES\_PER\_UNIT to provide deeper insights into product performance.

**Feature Engineering:**

Developed key metrics like SALES\_PER\_UNIT to facilitate a more nuanced understanding of product-level performance. Integrated the transaction data with customer information, enabling a comprehensive analysis that combines sales performance with customer demographics. Customer Segmentation Analysis:

The customer data was segmented based on LIFESTAGE and PREMIUM\_CUSTOMER status to analyze patterns in total and average sales. This segmentation allowed for the identification of specific customer groups that contribute significantly to overall sales.

**Correlation Analysis between Pack Size and Sales:**

An in-depth examination of the relationship between chip pack sizes and sales metrics (total and average) was conducted to identify any significant trends or patterns.

**Visualization Techniques:**

To effectively communicate the findings, a range of visualizations were employed:

1. Bar Plots: Illustrated total and average sales across different customer segments and premium status groups.
2. Box Plots: Showcased the distribution of prices within various customer segments, providing insights into spending behavior.

**Key Results:**

1. Older Families and Retirees in both mainstream and premium categories showed high total sales.
2. Young Singles/Couples exhibited lower average sales compared to other categories.
3. Sales were relatively affected by the size of chip packs.

#### Conclusion:

Targeting Older Families and Retirees in marketing campaigns may lead to higher returns due to their demonstrated purchasing power. Additional promotional efforts for Young Singles/Couples could help boost sales in that category.