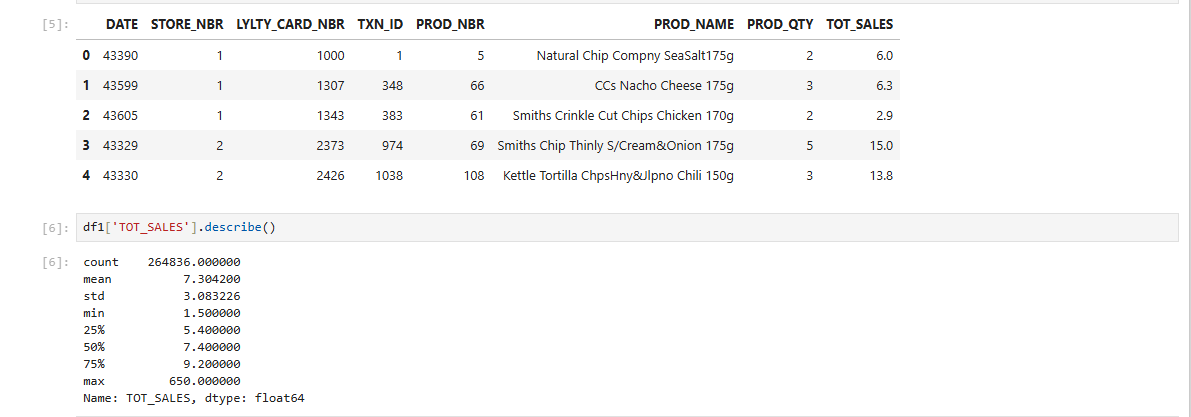
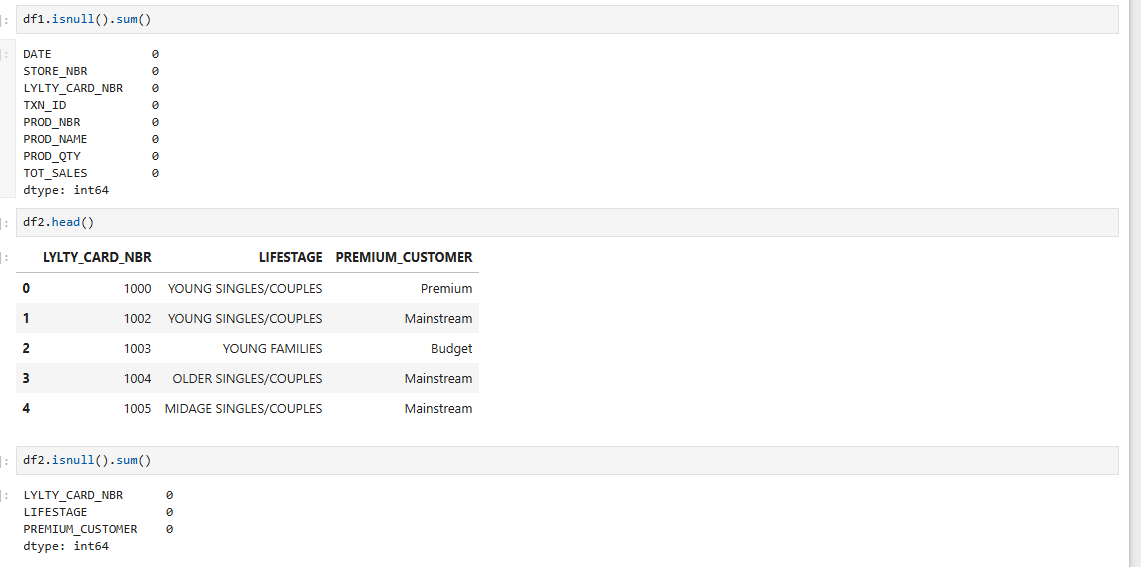
Problem 2

**1. Importing Data and Summary**

Start by loading the transaction and customer data from the provided CSV files. We want to:

1. Generate summaries of the data to get an understanding of its shape, size, and general structure.
2. Use describe() for numeric columns and value\_counts() for categorical columns.
3. Check for missing values with isnull().sum().





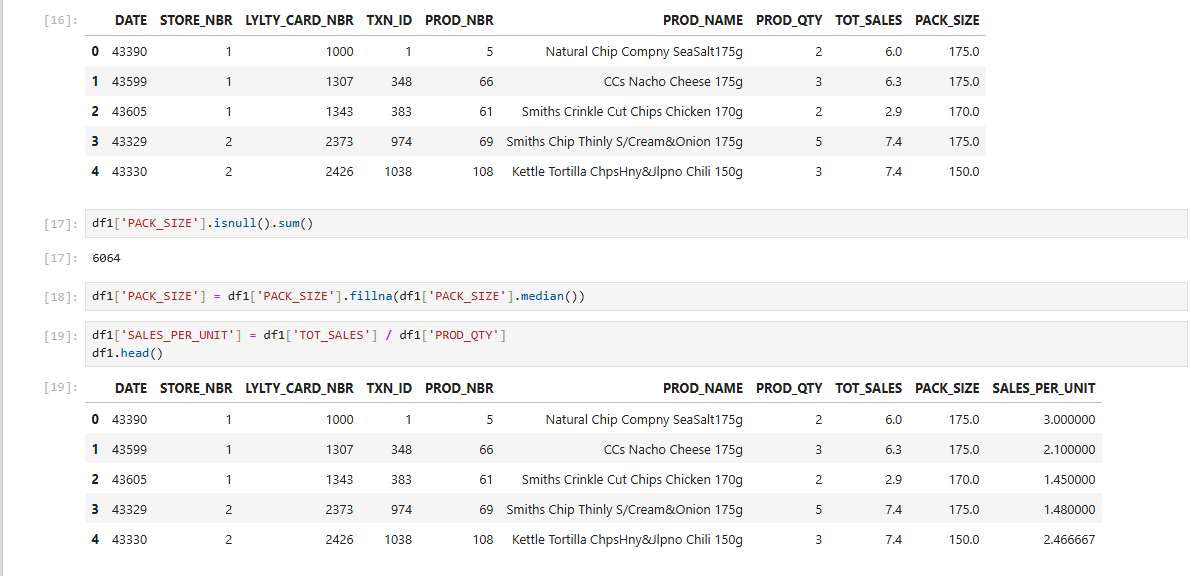
**2. Data Cleaning**

1. Missing data: If there are any null values in critical columns(TOTAL\_SALES), decide whether to replace missing values or drop the rows, depends on the data distribution.
2. Outliers: Use boxplots/standard deviation methods to detect any extreme values in TOTAL\_SALES.



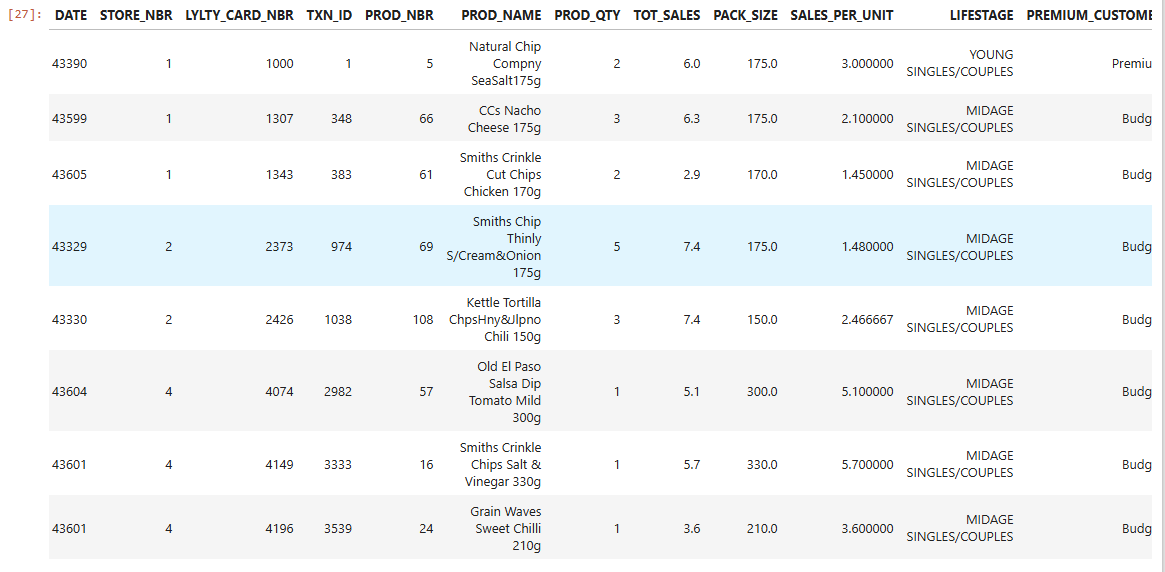
**3. Feature Engineering**

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

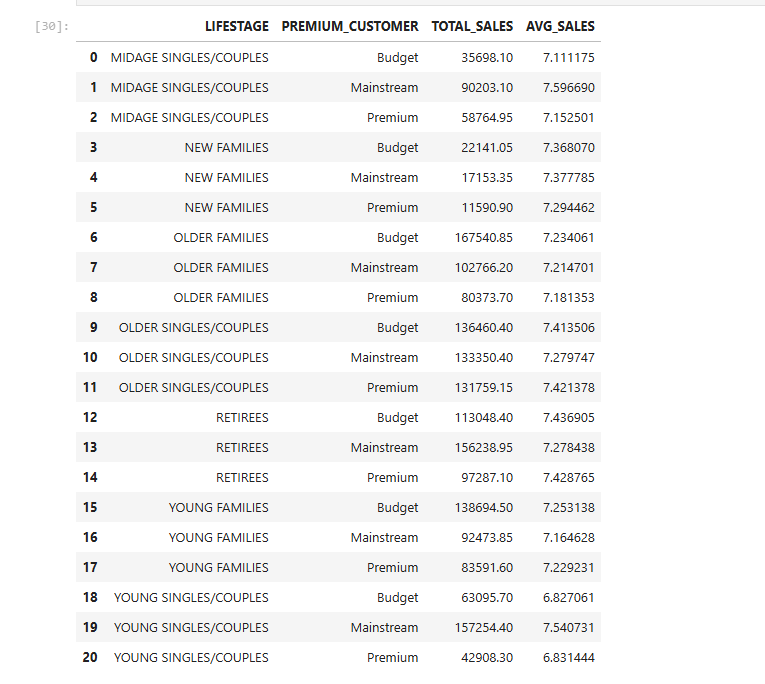


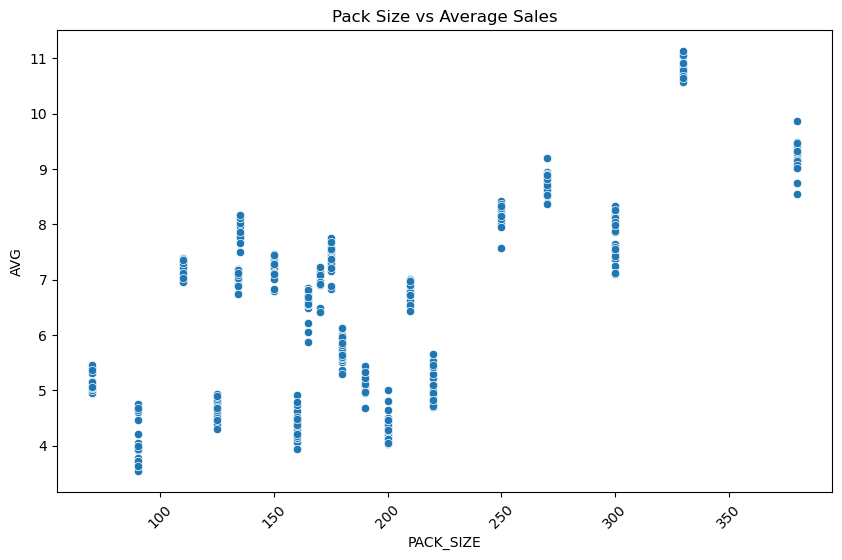
**4. Merging Datasets**

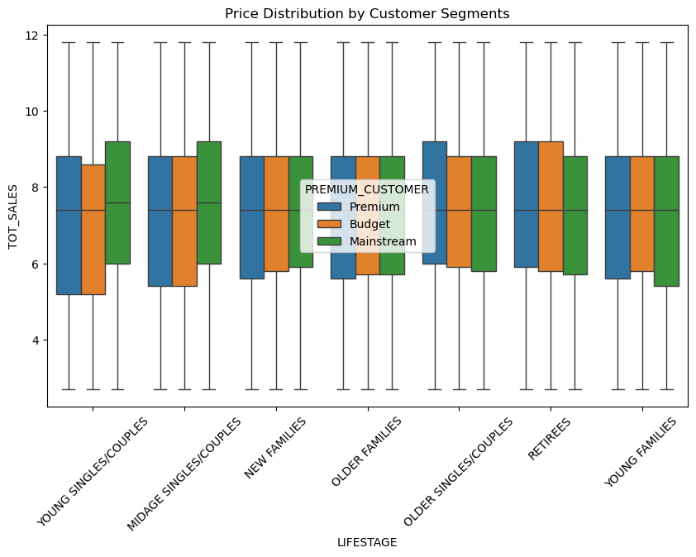
Now, once the data cleaning and processing is done on both the datasets, we'll merge the datasets on a shared key(LYLTY\_CARD\_NBR)

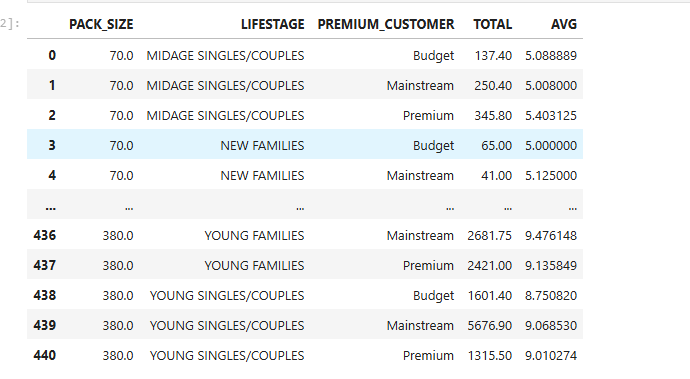


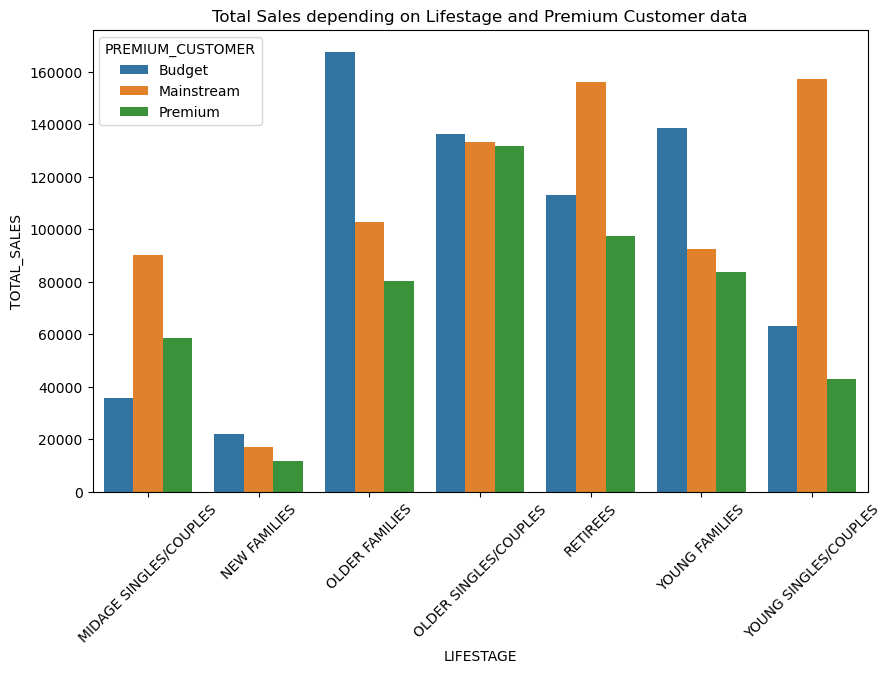
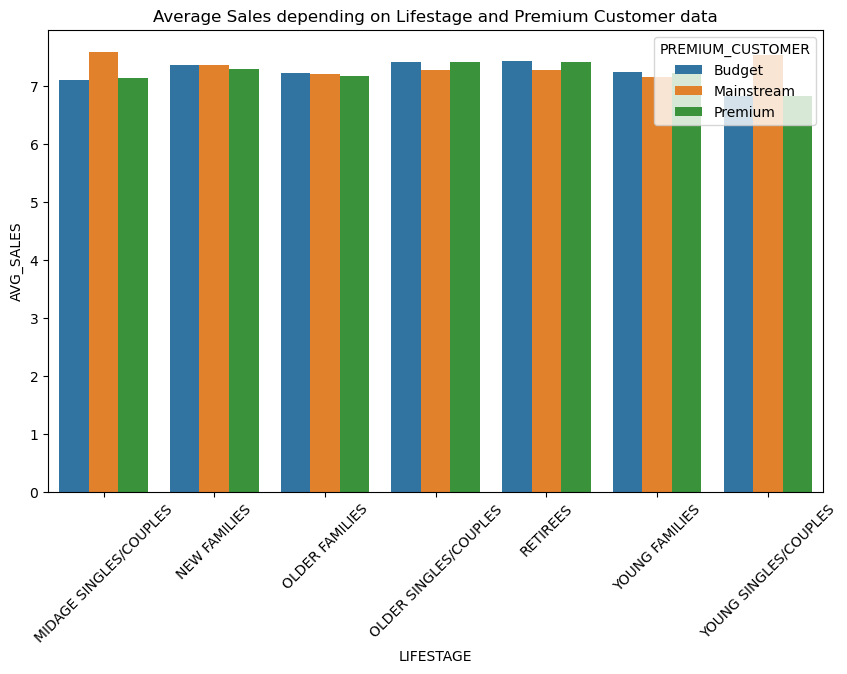
**5. Customer Data Analysis**











**Objective:**

Provide a data-driven strategic recommendation for an upcoming category review by analyzing chip purchasing behaviors across different customer

**Overview:**

We worked with two datasets:

1. Transaction Data: Contains sales transactions including product names, quantities, total sales, and other details.
2. Customer Data: Contains customer information like lifestage and premium customer status.

**Data Cleaning:**

1. Handled outliers using IQR method.
2. Addressed missing values by filling with medians where appropriate.
3. Extracted additional features such as PACK\_SIZE from product names and computed SALES\_PER\_UNIT.

**Feature Engineering:**

1. Created metrics like SALES\_PER\_UNIT to better understand per-product performance.
2. Merged transaction data with customer information for enhanced analysis.

**Customer Data Analysis:** Grouped customers by LIFESTAGE and PREMIUM\_CUSTOMER to examine total and average sales.

**Pack Size and Sales Correlation:** Analyzed the relationship between chip pack sizes and total/average sales.

**Visualization:**

1. Bar plots of total and average sales by customer segment and premium status.
2. Box plots highlighting price distribution across customer segments.
3. Scatter plots of packet sizes across the average sales.

**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.