Exploratory Data Analysis and Time Series Insights for Retail Store Sales

Arkadyuti Paul

Master of Science in Data Science

St. Xavier's College (Autonomous), Kolkata

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project takes a deep dive into a Retail Store Sales datasets, using time-series analysis to uncover meaningful patterns in customer behaviour. Starting with raw transaction data, the records were carefully cleaned, checked for gaps, and enhanced with extra features like week and month indicators, as well as basket-level details.

Exploratory Data Analysis (EDA) highlighted the best-selling products, brands, and stores, while also taking a closer look at customer habits - such as basket size, overall spending, and preferences for different pack sizes. The time-based trends revealed strong weekly and seasonal rhythms, with sales peaking during holiday periods (no surprise - retail therapy is a holiday tradition). Further analysis looked at how sales are tied to price, quantity, and customer segments. We saw clear evidence of price sensitivity and differences in brand performance, both of which play a major role in shaping overall sales.

2. Introduction

Retailers generate huge amounts of transaction data every single day - and tucked inside that data are valuable clues about customer behaviour, sales patterns, and opportunities for smarter decision- making. This project takes on the challenge of turning that raw information into meaningful insights using time- series analysis and visualization. The ultimate goal: to help improve inventory planning, promotions, and overall strategy.

The dataset included detailed records on product sales, quantities, prices, stores, and customer segments. Like most raw data, it needed some care before analysis - so the cleaning process involved fixing missing values, adjusting data types, removing outliers, and creating new features such as week and month indicators, plus basket-level metrics. Once tidied up, the dataset was in good shape for exploration.

On the technical side, the project leaned on Python and its data science toolkit: pandas and numpy for data handling, matplotlib and seaborn for visualizations. Together, these tools made it possible to handle messy real-world data while creating clear, engaging visuals.

A quick look at related work showed a common theme: before retailers dive into forecasting models, they usually start with exploratory analysis and visualization. Seasonality, price sensitivity, and brand performance consistently appear as key drivers of sales, which aligns well with the approach in this project. Here, the emphasis was on generating actionable insights through accessible exploration rather than jumping straight into advanced prediction models.

The process unfolded step by step - starting with data ingestion and cleaning, moving to exploratory data analysis, and then pulling the pieces together through charts and narratives. The results shed light on which products and brands perform best, how customers shop over time, and when sales naturally spike throughout the year.

In the end, the project shows how raw transactional data can be transformed into clear, business-ready insights. Retailers can use these findings to spot high-performing products, anticipate seasonal demand, understand customer preferences, and make decisions that are both data-driven and practical.

The list of topics that I have received training on during the first two weeks of internship:

- Python programming fundamentals (data types, loops, data structures, OOP, Numpy, Pandas).
- Core machine learning concepts, with practical sessions on regression and classification.
- An introduction to Large Language Models (LLMs).
- Professional communication skills.

3. Project Objective

The main goal of this project is to explore the Retail Sales dataset using Exploratory Data Analysis (EDA) to uncover useful insights about customer purchasing habits and overall transaction patterns. The dataset includes details such as customer demographics, product categories, purchase dates, and transaction values. By examining these elements step by step, the project aims to highlight key trends, relationships, and even unusual patterns hidden within the data.

The analysis focuses on learning more about different customer groups, understanding how products are performing, and studying how purchase behaviour changes over time. Another objective is to see whether factors like age, location, or product category have a noticeable effect on how much customers spend.

To make this exploration clear and structured, the project is guided by the following objectives:

- Identifying store-wise and product-level sales trends
- Analysing customer purchase behaviour based on demographics
- Examining temporal patterns in transactions
- Exploring the relationship between price, quantity purchased, and total sales
- Studying brand-level performance and customer preferences

4. Methodology

This project followed a step-by-step approach to make sure the retail sales data was properly prepared, explored, and visualized for meaningful insights. The workflow can be broken down into the following stages:

4.1. Data Collection

The datasets consisted a year's worth of potato chip transactions and customer data for a region with details such as purchase date, store ID, product name, quantity, selling price, total sales and customer segment.

The data came from a retail shop of San Francisco and is imported into Python using the pandas library.

Since the project relied on existing (secondary) data, no surveys were conducted, meaning questionnaires or sampling methods were not needed here.

4.2. Tools and Technologies

Language: Python was the primary tool for analysis.

Libraries:

- pandas and numpy – Data manipulation, preprocessing, and handling structured/tabular datasets efficiently
- matplotlib and seaborn – Static (traditional) visualizations for clear charts, distribution plots, and summaries
- datetime – Date handling and feature engineering, especially for time-series analysis
- re (Regular Expressions) – Text cleaning and pattern matching for handling messy string data
- collections.Counter – Quick frequency counts, ideal for analyzing repeated elements such as most common words, products, or categories
- Environment: Jupyter Notebook – Combines code, analysis, and explanations in one place for a reproducible, interactive workflow

4.3. Data Preprocessing and Cleaning

To make the dataset reliable for analysis, several steps were taken:

Date Conversion: Transformed purchase dates into proper datetime format for time-series analysis.

Handling Missing Values: Checked for gaps in key fields (like product name, store, quantity). Missing entries were fixed via imputation or removal.

Outlier Detection: Looked at extreme values in price and quantity to avoid distorted insights.

Feature Engineering: Created new time-based features such as Year, Month, Quarter, Week, and Day to enrich the analysis.

4.4. Exploratory Data Analysis (EDA)

The heart of the project was exploring the dataset to uncover patterns and insights:

- **Store-wide Sales:** Looked at how each store was performing overall, spotting which ones were leading in sales and which might need extra support or focus.
- **Product-level Sales Trends:** Tracked how individual products were selling—finding the consistent best-sellers, seasonal favorites, and items that weren't doing so well.
- **Brand-level Performance:** Checked the strength of different brands, showing which ones dominate, where customer loyalty lies, and where there are opportunities to grow.
- **Demographics:** Broke down purchases by customer characteristics (like age, gender, and location) to see how different groups shop and spend.
- **Basket/Transaction Analysis:** Took a closer look at how many items customers usually buy in one trip and how much they spend—revealing typical basket size and value.
- **Product Quantity vs Customer Segment:** Compared how much different customer groups buy at a time—whether they're bulk buyers or prefer smaller packs.

- **Monthly/Quarterly Trends:** Explored how sales move across months and quarters, helping spot seasonal spikes and broader demand cycles.
- **Week-over-Week Sales:** Looked at shorter-term ups and downs in sales, making it easier to see how weekly performance changes and how promotions or events affect results.
- **Temporal Segment Analysis:** Studied how different customer groups behave across time—like whether certain segments shop more heavily during holidays or weekends.
- **Correlation Matrix:** Analyzed how different factors (like price, quantity, and sales) are linked with each other, making relationships clearer through a matrix and heatmap.
- **Price Sensitivity by Demographics:** Found out which customer groups are more price-conscious and which ones stay loyal to premium options, no matter the cost.
- **Pack Size Preference:** Looked at what product pack sizes (small, medium, bulk) customers lean toward, guiding smarter stocking and promotions.
- **Store & Brand Interaction:** Mapped how different brands perform in different stores, showing that brand strength can vary a lot depending on location and customer preferences.

4.5. Interpretation of Results

Every visualization and summary table was explained with clear takeaways.

Key observations included spikes during holiday weeks, certain dominant brands, and noticeable differences in how customer segments behaved.

4.6. Documentation and Deliverables

All steps were thoroughly documented in the Jupyter Notebook for transparency and reproducibility.

Final deliverables included annotated charts, summary tables, and a set of actionable insights.

The focus stayed on exploration and visualization, so no machine learning models were implemented at this stage.

4.7. Workflow Summary (Flowchart)

The overall flow of work can be summarized as:

Data Collection → Data Cleaning → Feature Engineering → Exploratory Data Analysis → Visualization & Interpretation → Conclusions & Recommendations

For further information, the official documentation can be accessed at the following link –

https://github.com/ArkadyutiPaul/IDEAS-AUTUMN-INTERNSHIP-PROGRAM-ON-DATA-SCIENCE-2025/blob/main/Visualizing_Time_Series_Dataset_Retail_Sales_Data.ipynb

5. Data Analysis and Results

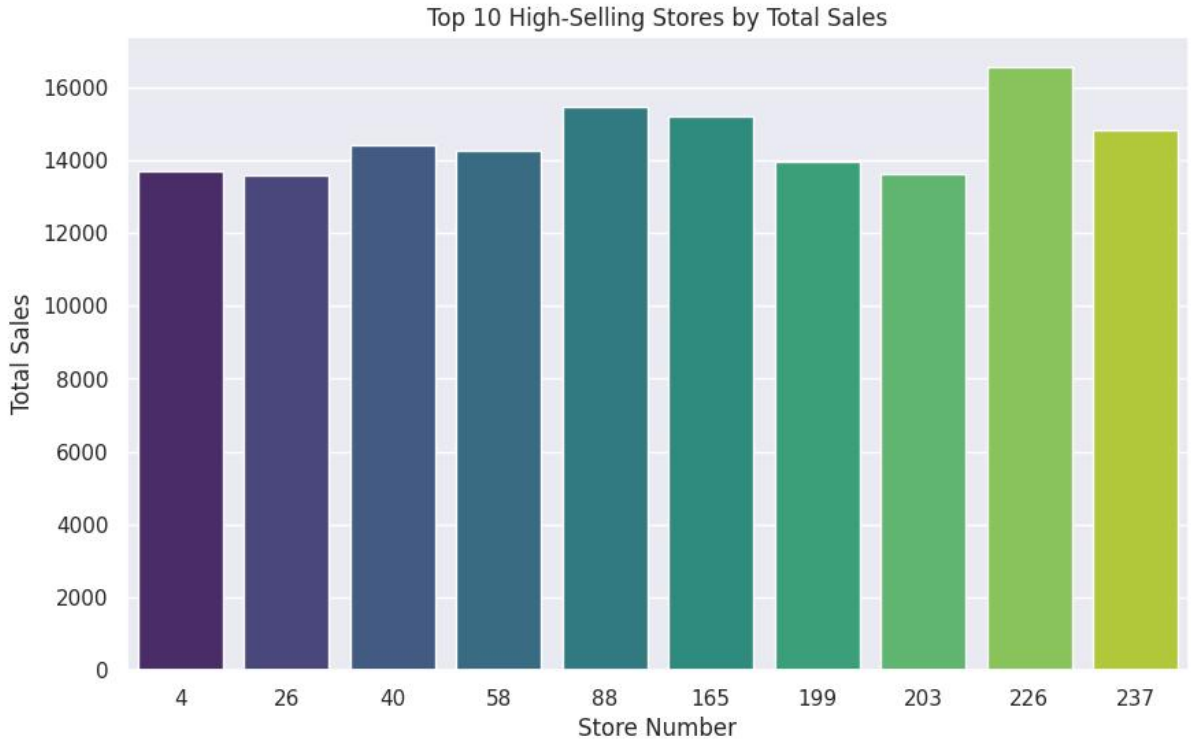
This section explores the dataset through clear analysis and visuals. The results bring out the main patterns and trends, helping us understand the key insights hidden in the data.

5.1. Store-wide Sales:

Top 10 High-Selling Stores:

STORE_NBR	TOT_SALES
226	16544.65
88	15445.85
165	15188.35
237	14830.60
40	14427.30
58	14256.95
199	13975.90
4	13709.25
203	13623.40
26	13597.20

- **Store 226** has the highest overall sales, followed closely by Stores **88** and **165**.
- Stores such as **40, 58, and 237** also show consistently strong sales performance, placing them in the higher tier.
- In contrast, while Stores **4, 26, 199, and 203** make it to the top 10 list, their cumulative sales are relatively lower than the leading stores.



As depicted in the bar graph above, we can conclude that certain stores significantly outperform others in terms of their total sales. To study this, the dataset was grouped using

the groupby() function in Python on the STORE_NBR column, and the corresponding total sales for each store were aggregated using .sum().

The results were then sorted in descending order, and the **top 10 stores by total sales** were visualized using the **Seaborn library.**

The results show that **a small number of stores are responsible for most of the revenue in the retail business**.

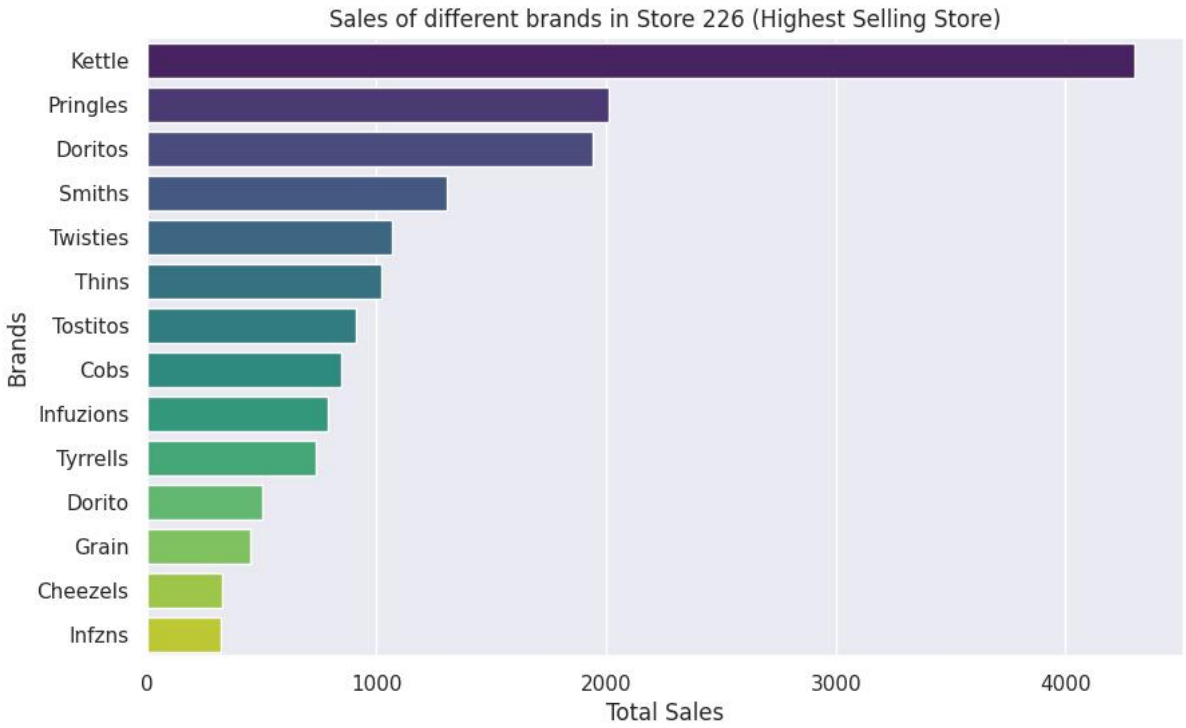
This may be due to factors like the **size of the store**, **its location**, and the **range of products** it offers.
By understanding these factors, **businesses can make better choices about resource allocation, marketing activities, and store expansion decisions**.

Brand Sales within Store 226 (Highest-Selling Store):

Seles of Brands in top selling Store:

BRANDS	TOT_SALES
Kettle	4300.60
Pringles	2012.80
Doritos	1941.60
Smiths	1309.40
Twisties	1070.10
Thins	1023.00
Tostitos	910.80
Cobs	843.60
Infuzions	790.40
Tyrrells	735.00
Dorito	503.75
Grain	450.00
Cheezels	330.60
Infzns	323.00

- **Kettle** is the dominant brand in Store 226, with sales figures far exceeding those of other brands, crossing the **4,000** mark.
- **Pringles** and **Doritos** follow as the next highest-selling brands, although their sales are roughly half of Kettle's total, highlighting a significant performance gap.
- Other brands such as **Smiths, Twisties, and Thins** also show strong contributions but remain in the mid-tier range.
- Meanwhile, brands like **Cheezels, Grain, and Infzns** account for comparatively lower sales, making up the lower tier in this store's brand performance.



As depicted in the bar graph above, the sales contribution of different brands within **Store 226 (the highest-selling store)** can be observed.

The dataset was filtered specifically for this store, and total sales were aggregated across each brand using the groupby() function in Python.

The results were then sorted in descending order and visualized using the **Seaborn library.**

This analysis suggests that **a few key brands drive the majority of sales within the top-performing store**.

This implies that **stocking strategies, promotional offers, and marketing campaigns** for these dominant brands could have a significant impact on overall store revenue. Conversely, understanding why lower-performing brands contribute less may help in **optimizing shelf space, rethinking product placement, or tailoring promotions** to boost their sales.

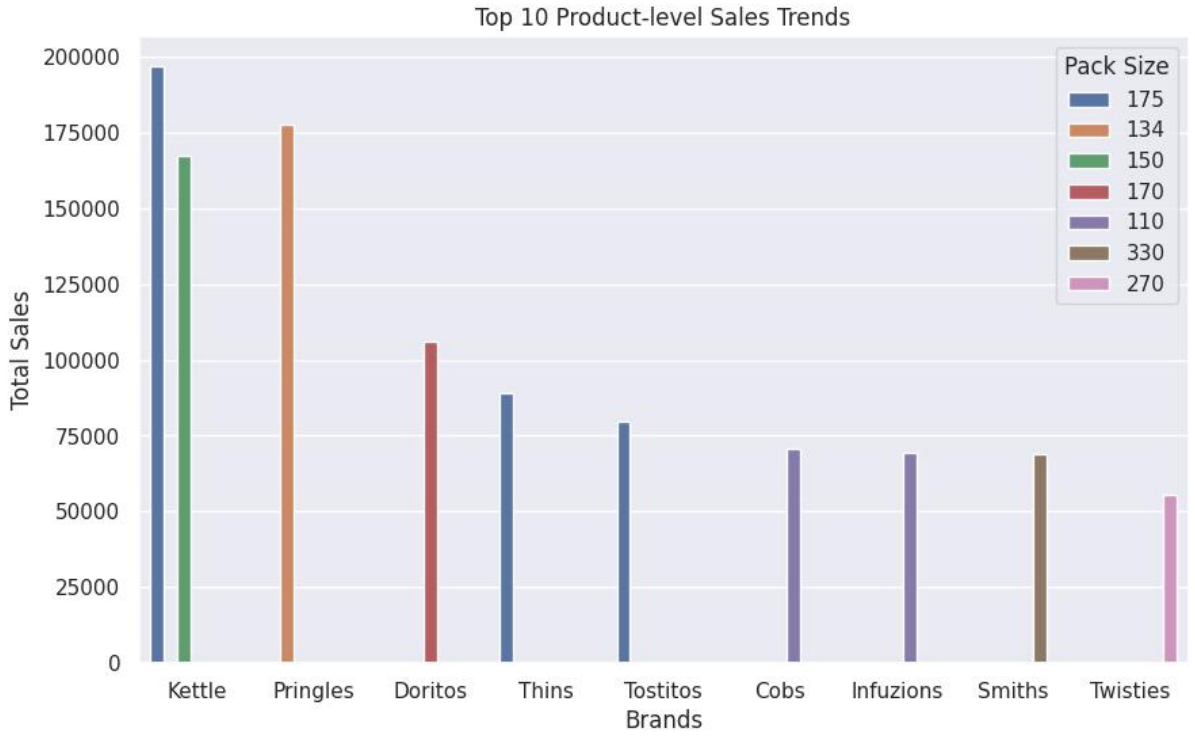
5.2. Product-level Sales Trends:

Top 10 High-Selling Products:

BRANDS	PACK_SIZE	TOT_SALES
Kettle	175	196668.0
Pringles	134	177655.5
Kettle	150	167481.4
Doritos	170	106264.4
Thins	175	88852.5
Tostitos	175	79789.6
Cobs	110	70569.8

Infuzions	110	69395.6
Smiths	330	69106.8
Twisties	270	55425.4

- **Kettle (175g pack)** is the leading product, achieving close to **200,000** in total sales.
- This is followed by **Pringles (134g pack)** and **Kettle (150g pack)**, both of which show strong performance with sales figures exceeding **170,000**.
- **Doritos (170g pack)** ranks next with sales above **100,000**, while other products such as **Thins (175g), Tostitos (175g), Cobs (110g), Infuzions (110g), Smiths (110g), and Twisties (270g)** make up the remaining positions in the top 10.



As shown in the bar graph above, the **top 10 product-level sales trends** are illustrated by aggregating sales across both **brand** and **pack size**.

The dataset was grouped using the groupby() function on the BRANDS and PACK_SIZE columns, and the total sales (TOT_SALES) were computed. The results were then sorted in descending order, and the highest-selling brand–pack combinations were visualized using the **Seaborn library.**

- **Mid-sized packs (110g–175g)** dominate sales compared to smaller or very large pack sizes, suggesting that customers prefer moderate pack sizes that balance price and quantity.
- The dominance of brands such as **Kettle, Pringles, and Doritos** highlights strong **brand loyalty** and **consumer preference** in this retail dataset.

This analysis suggests that both **brand strength** and **optimal pack size** are critical drivers of product-level sales.

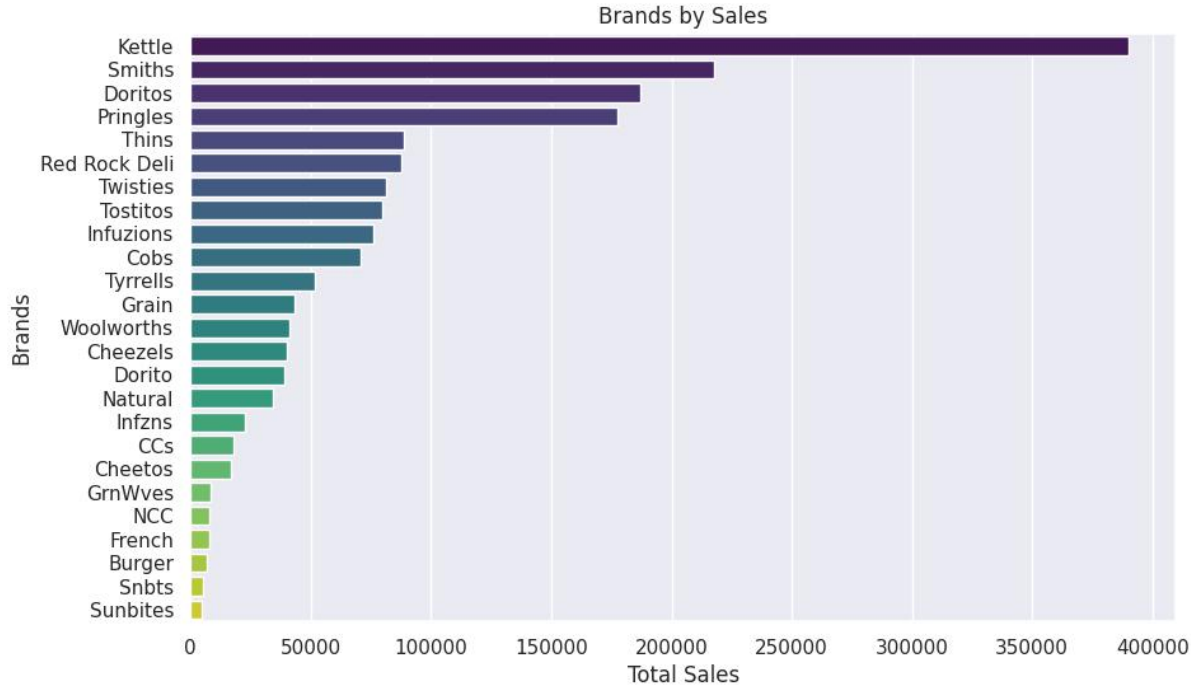
Understanding these patterns can guide **inventory planning, product assortment strategies, and promotional campaigns**, with a focus on high-performing product-pack combinations, while also reevaluating the role of lower-selling sizes.

5.3. Brand-level Performance:

Top 10 High-Selling brands:

BRANDS	TOT_SALES
Kettle	390239.8
Smiths	217492.0
Doritos	187277.9
Pringles	177655.5
Thins	88852.5
Red Rock Deli	87607.5
Twisties	81522.1
Tostitos	79789.6
Infuzions	76247.6
Cobs	70569.8

- **Kettle** is the leading brand, generating sales close to **400,000**, which is significantly higher than all other brands.
- **Smiths, Doritos, and Pringles** follow as other strong performers, with sales ranging between **150,000 and 200,000**.
- Mid-tier brands such as **Thins, Red Rock Deli, Twisties, and Tostitos** contribute steadily but at much lower levels compared to the top four.
- Brands such as **Infuzions, CCs, Cheetos, and Sunbites** represent the lower end of the sales spectrum with comparatively minimal contributions.



As illustrated in the bar chart above, the sales contribution of different **brands** has been computed by aggregating the TOT_SALES column across all transactions.
The dataset was grouped by the BRANDS field using the groupby() function in Python, followed by summation of total sales.
The results were then sorted in descending order and visualized using the **Seaborn library**.

This analysis suggests that a handful of brands dominate the overall sales performance, with **Kettle alone contributing disproportionately** to total revenue.
Such dominance reflects both **brand loyalty and customer preference**, and it highlights the importance of prioritizing **inventory management, marketing strategies, and promotional campaigns** for the top-selling brands.
Lower-performing brands may require targeted actions such as **discounts, better shelf placement, or repositioning** to improve their sales.

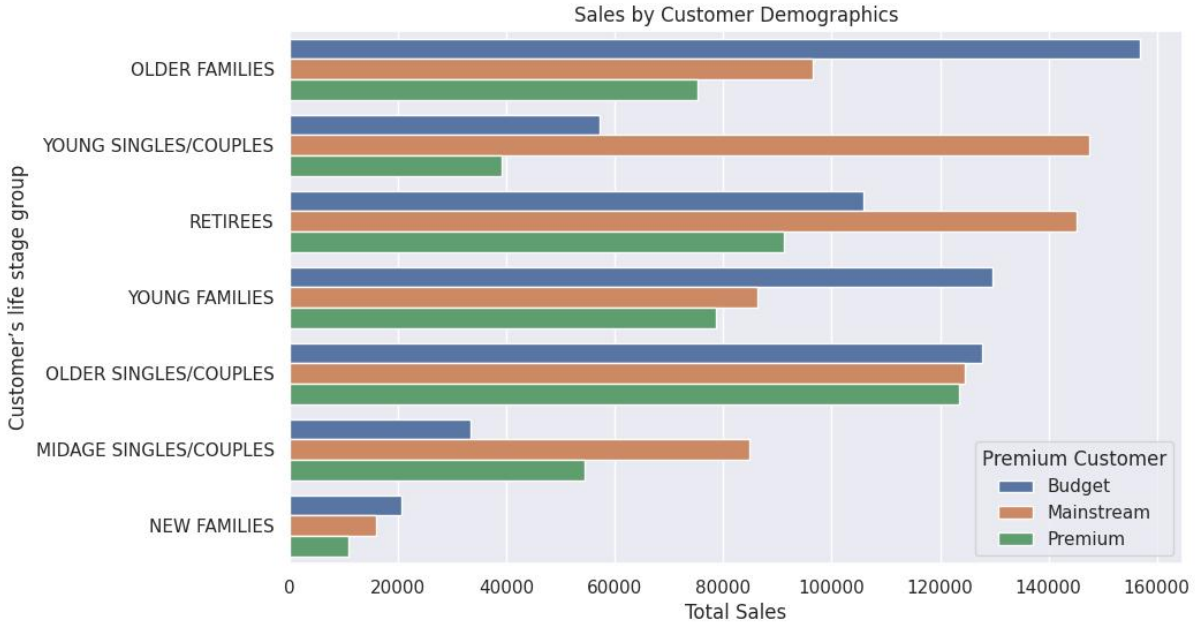
5.4. Demographics:

The total sales by LIFESTAGE and PREMIUM_CUSTOMER:

LIFESTAGE	PREMIUM_CUSTOMER	TOT_SALES
OLDER FAMILIES	Budget	156863.75
YOUNG SINGLES/COUPLES	Mainstream	147582.20
RETIREES	Mainstream	145168.95
YOUNG FAMILIES	Budget	129717.95
OLDER SINGLES/COUPLES	Budget	127833.60
OLDER SINGLES/COUPLES	Mainstream	124648.50
OLDER SINGLES/COUPLES	Premium	123537.55
RETIREES	Budget	105916.30
OLDER FAMILIES	Mainstream	96413.55
RETIREES	Premium	91296.65
YOUNG FAMILIES	Mainstream	86338.25

MIDAGE SINGLES/COUPLES	Mainstream	84734.25
YOUNG FAMILIES	Premium	78571.70
OLDER FAMILIES	Premium	75242.60
YOUNG SINGLES/COUPLES	Budget	57122.10
MIDAGE SINGLES/COUPLES	Premium	54443.85
YOUNG SINGLES/COUPLES	Premium	39052.30
MIDAGE SINGLES/COUPLES	Budget	33345.70
NEW FAMILIES	Budget	20607.45
NEW FAMILIES	Mainstream	15979.70
NEW FAMILIES	Premium	10760.80

- **Older Families** generate the highest total sales, particularly within the **Budget segment**, indicating they are a dominant customer group for this product category.
- **Young Singles/Couples** and **Retirees** are also major contributors, with **Mainstream customers** leading in both groups.
- **Young Families** show strong sales, mostly driven by **Budget customers**, followed closely by **Mainstream customers**.
- **Older Singles/Couples** also contribute significantly across all three premium categories, showing a fairly **balanced distribution**.
- **Midage Singles/Couples** sales are moderate, with **Mainstream being the largest contributor**.
- **New Families** contribute the least overall, suggesting they are a relatively smaller consumer base in this dataset.



The chart shows **total sales segmented by customer life stage group and premium status** (Budget, Mainstream, Premium).

- **Life stage matters:** Sales are concentrated **in Older Families, Young Singles/Couples, Retirees, and Young Families** — these groups should be the focus of marketing and promotions.
- **Premium segmentation:**
  - **Budget customers dominate** in *Older Families and Young Families.*
  - **Mainstream customers dominate** in *Young Singles/Couples and Retirees.*
  - **Premium customers** are strongest in *Older Singles/Couples*, showing they may be more **brand/quality-oriented** in that stage.
- **Smaller groups** (e.g., *New Families, Midage Singles/Couples*) represent niche markets but may be growth opportunities with targeted campaigns.

This analysis suggests that **family life stage** and **spending profile** (budget vs. mainstream vs. premium) are critical factors in sales performance.

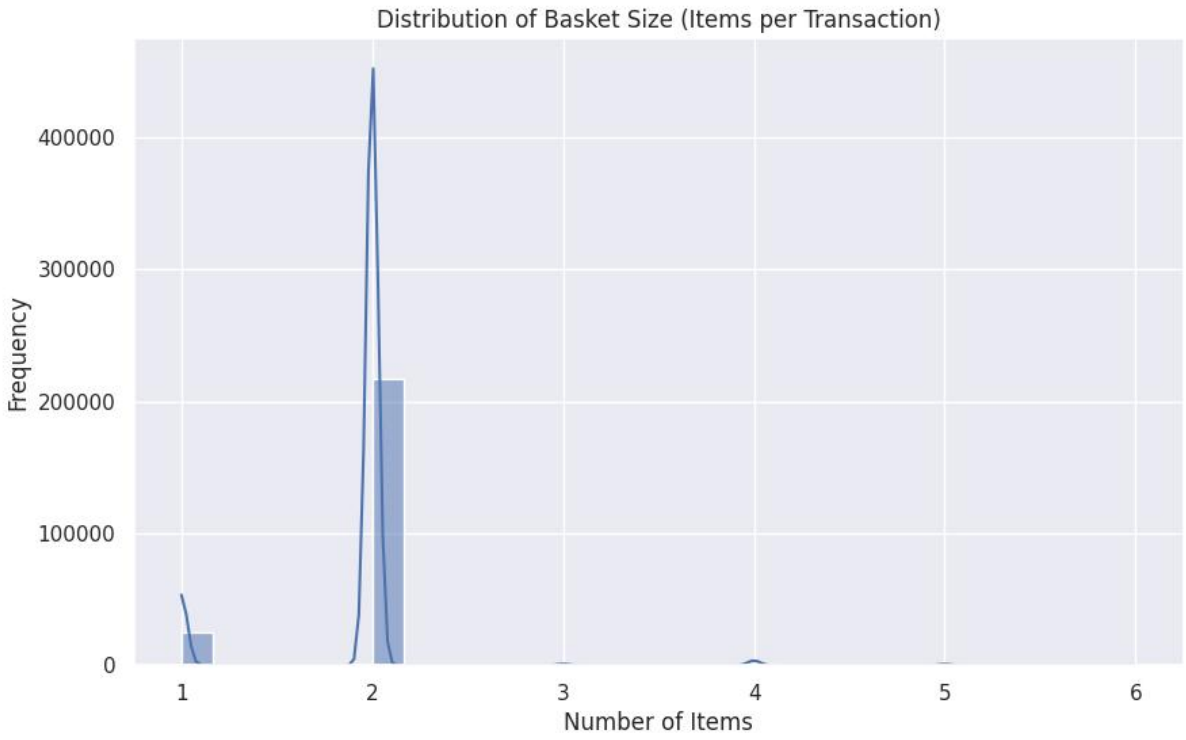
- To maximize revenue, businesses should prioritize **Older Families, Young Singles/Couples, and Retirees**, tailoring promotions to their premium status (*budget vs. mainstream*).
- **Premium positioning** works well with **Older Singles/Couples**, while **budget-friendly offers** resonate with **family groups**.
- For **underrepresented groups** (e.g., *New Families*), specialized campaigns could unlock **new sales potential**.

5.5. Basket/Transaction Analysis:

Distribution of Basket Size (Items per Transaction)

The basket size:

PROD_QTY	count
2	217316
1	25319
4	1724
3	468
5	420
6	8



- The **basket size distribution** is highly concentrated around **2 items per transaction**, which represents the most common shopping behaviour.
- A small number of transactions consist of only **1 item**, and very few transactions exceed **3 or more items**.
- This indicates that the majority of customers purchase **snack products in small quantities**, treating them as impulse or supplementary buys rather than bulk purchases.

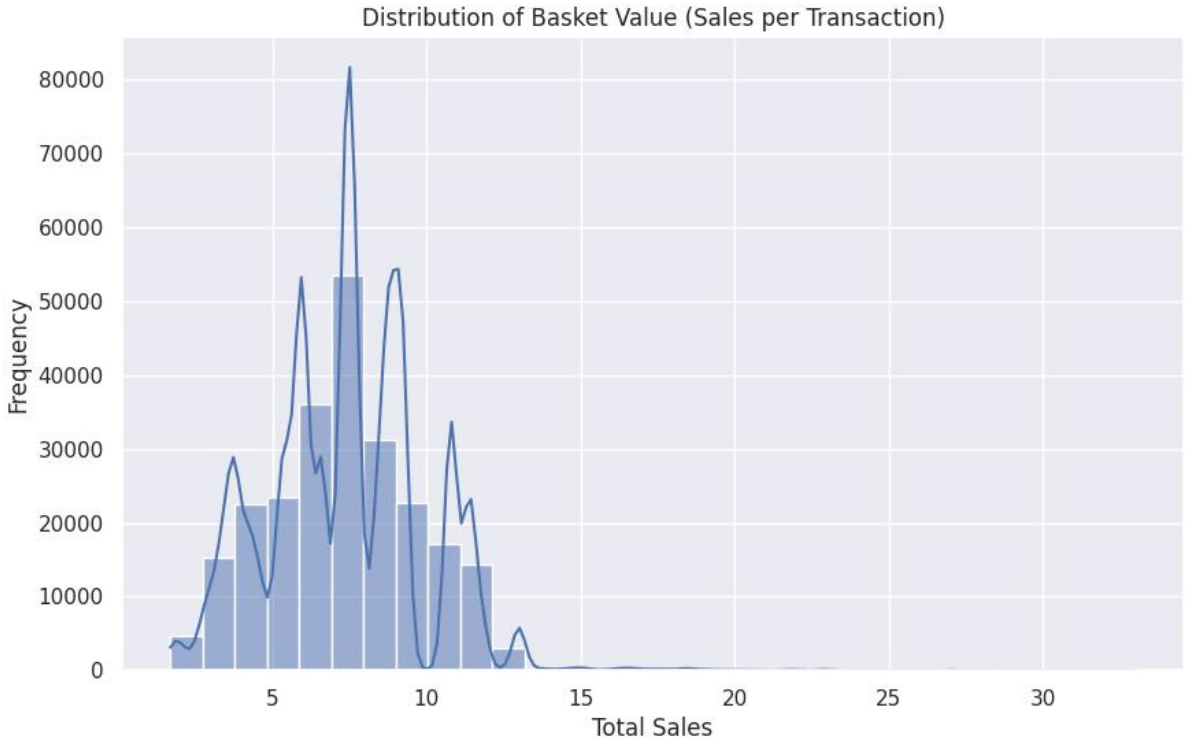
Most transactions are light baskets (1–2 items), so promotions encouraging multi-buy offers (e.g., *"Buy 2, Get 1 Free"*) could potentially increase basket size.

Distribution of Basket Value (Sales per Transaction)

The basket value:

TOT_SALES	count
9.2	22548
7.4	22247
6.0	20478
7.6	19976
8.8	19672
...	...
23.6	1
4.9	1
10.7	1
6.3	1
18.9	1

208 rows × 1 columns



- The **basket value distribution** is right-skewed, with most transactions falling between **5 and 10 $** in sales value.
- A peak is observed around (7–9), showing this is the typical spend per transaction.
- Higher-value transactions (>15) are relatively rare, indicating that bulk or premium purchases are exceptions rather than the norm.

Customers usually spend within a small range, reinforcing the idea of snack purchases being **low-ticket items**. Retailers could introduce **bundle pricing** or **upsell strategies** to push average spend slightly higher. The dataset reflects a **convenience/snack-driven purchasing pattern** with small baskets and modest spending.

- Business opportunities lie in **nudging customers towards slightly larger basket sizes and higher-value transactions** through promotions, product bundling, or targeted offers.

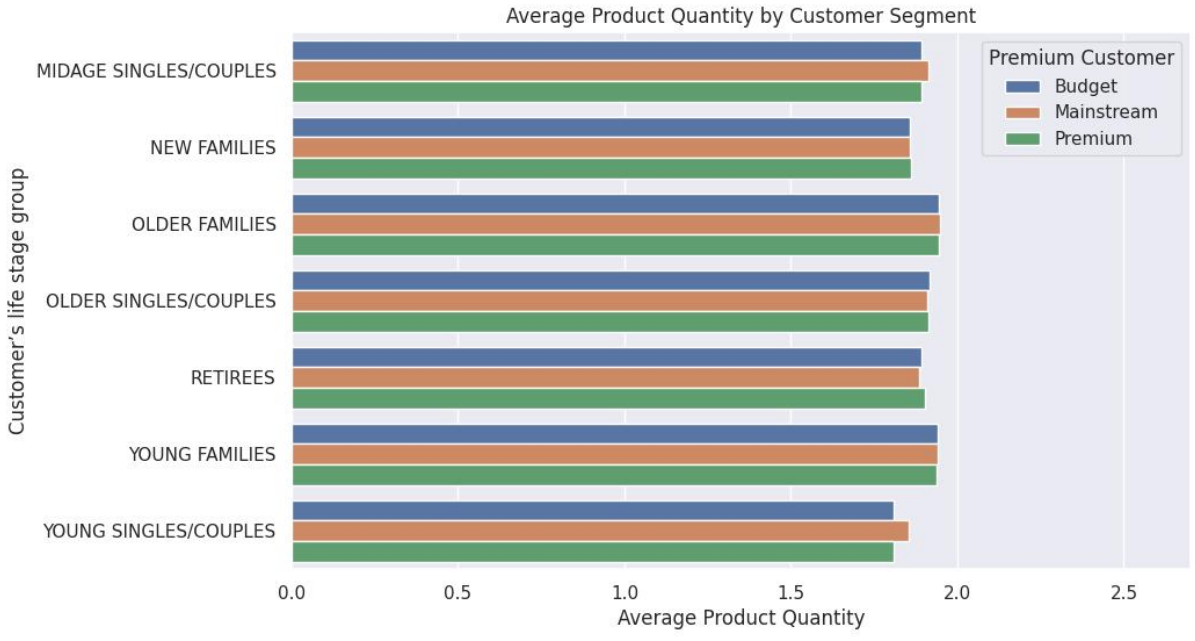
5.6. Product Quantity vs Customer Segment:

The average product quantity by customer segment:

LIFESTAGE	PREMIUM_CUSTOMER	PROD_QTY
OLDER FAMILIES	Mainstream	1.948795
OLDER FAMILIES	Premium	1.945496
OLDER FAMILIES	Budget	1.945384
YOUNG FAMILIES	Mainstream	1.941408
YOUNG FAMILIES	Budget	1.941226
YOUNG FAMILIES	Premium	1.938149

OLDER SINGLES/COUPLES	Budget	1.914920
OLDER SINGLES/COUPLES	Premium	1.913949
MIDAGE SINGLES/COUPLES	Mainstream	1.911942
OLDER SINGLES/COUPLES	Mainstream	1.911201
RETIREES	Premium	1.901438
MIDAGE SINGLES/COUPLES	Budget	1.893626
RETIREES	Budget	1.893286
MIDAGE SINGLES/COUPLES	Premium	1.891750
RETIREES	Mainstream	1.886680
NEW FAMILIES	Premium	1.860887
NEW FAMILIES	Mainstream	1.858124
NEW FAMILIES	Budget	1.855878
YOUNG SINGLES/COUPLES	Mainstream	1.853510
YOUNG SINGLES/COUPLES	Budget	1.808002
YOUNG SINGLES/COUPLES	Premium	1.807075

- Across **all life stages and premium customer groups**, the **average product quantity per transaction is close to 2 items**.
- **Variation is minimal** across different segments — indicating a consistent purchasing pattern regardless of life stage or premium classification.
- Slight differences are observed:
    - **Young Families** and **Older Families** show marginally higher averages, suggesting family groups may purchase slightly larger quantities.
    - **Young Singles/Couples** tend to have a slightly lower average basket size compared to other groups.
    - Differences between **Budget, Mainstream, and Premium** customers are very small, showing that spending tier has little effect on average item quantity.



- **Customer life stage** has a modest impact: Families (both young and older) lean toward buying marginally more products per transaction than singles/couples.
- **Premium segmentation** (Budget vs. Mainstream vs. Premium) is not a strong driver of product quantity — customer type does not significantly change basket size.
- The overall **stability of basket quantity (~2 items)** aligns with earlier findings that transactions are generally small, reinforcing the snack/convenience nature of purchases.

This analysis suggests that, while different demographic groups (life stages) influence **who buys more**, the differences are relatively small.

- **Promotional campaigns** may need to focus less on increasing *quantity per basket* (since it's already stable) and more on **increasing basket value** through upselling or premium product positioning.
- For **family segments**, bundle or multi-pack deals could resonate better, while **singles/couples** may be more responsive to value-driven promotions.

5.7. Monthly/Quarterly Trends:

The monthly sales trends:

Period	TOT SALES
2018-7	154514.50
2018-8	147422.05
2018-9	150058.90
2018-10	153454.10
2018-11	149609.50
2018-12	156461.60
2019-1	151593.80
2019-2	140480.00
2019-3	155772.80
2019-4	149374.20
2019-5	146300.35
2019-6	150135.90

- Overall sales remain **fairly stable**, fluctuating around **148K – 156K** across the months.
- **Highest peak:** December 2018 approximately (156K) and March 2019 (~156K), suggesting strong seasonal or promotional sales.
- **Lowest point:** February 2019 (~141K), showing a sharp dip compared to surrounding months.
- Sales **show repeated surges around year-end and early Quatre2**, hinting at periodic demand spikes.

- **Seasonality effect:**
  - December peak likely tied to holiday shopping.
  - March rebound could reflect post-holiday restocking or seasonal promotions.
- **Sales dip in February:**
  - Consistent with shorter month length and possibly reduced consumer spending after holiday season.
- **Steady baseline:** Outside of peaks and dips, sales hover in a narrow range (~148K–153K), showing stable demand.

The data indicates that while the business enjoys stable monthly sales, there are clear seasonal spikes and dips.

- **December** is a major sales driver -- leverage with targeted promotions, gift bundles, and holiday campaigns.
- **February slump** -- opportunity to counteract with loyalty rewards or off-season discounts.
- Maintaining strong performance in March suggests potential for quarterly promotions timed around this rebound.

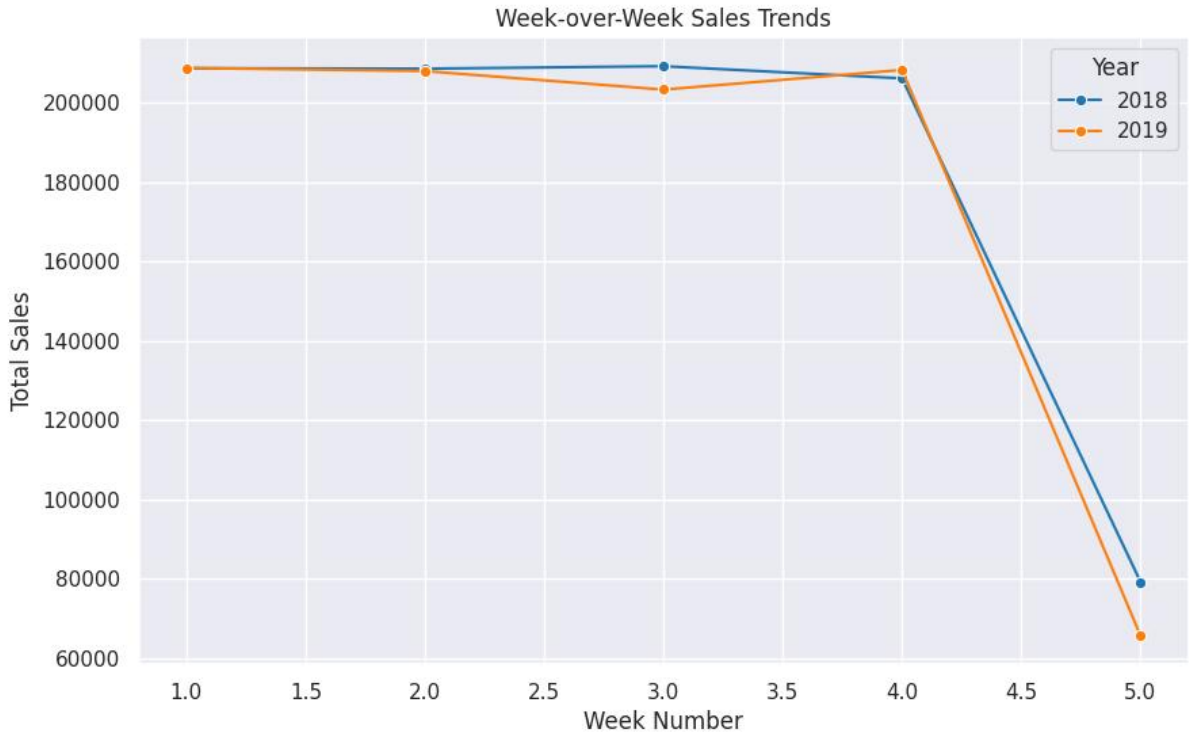
5.8. Week-over-Week Sales:

The weekly sales trends:

YEAR	WEEK	TOT_SALES
2018	3	209151.70
2019	1	208711.90
2018	1	208621.90

2018	2	208529.25
2019	4	208208.40
2019	2	207921.65
2018	4	206085.40
2019	3	203255.60
2018	5	79132.40
2019	5	65559.50

- **High sales consistency:** Weeks 1–4 show very stable sales levels, staying around **205K–210K** in both years (2018 and 2019).
- **Sharp drop in Week 5:** Sales plummet drastically -- ~80K in 2018 and ~65K in 2019, marking the lowest point in the period.
- **Year-over-year comparison:**
    - Sales in 2018 and 2019 track very closely across all weeks.
    - The Week 5 decline is more severe in 2019 than 2018.



- **Strong start of the month:** Consistent high sales in the first four weeks suggest stable consumer demand.
- **End-of-month drop:** The steep fall in Week 5 indicates a structural trend -- likely due to shorter purchase cycles, pay check timing, or fewer shopping days in those weeks.
- **Year-over-year similarity:** Since both years follow the same weekly pattern, the trend seems to be systemic rather than random.

This analysis highlights that sales are **front-loaded within each month**, with the bulk occurring in the first four weeks.

- Businesses can **capitalize early in the month** with promotions when sales are naturally strong.
- **Week 5 is a weak sales period** -- targeted discounts or marketing campaigns could help offset the slowdown.
- Since the trend repeats across years, planning inventory and marketing around this **monthly sales cycle** will optimize performance.

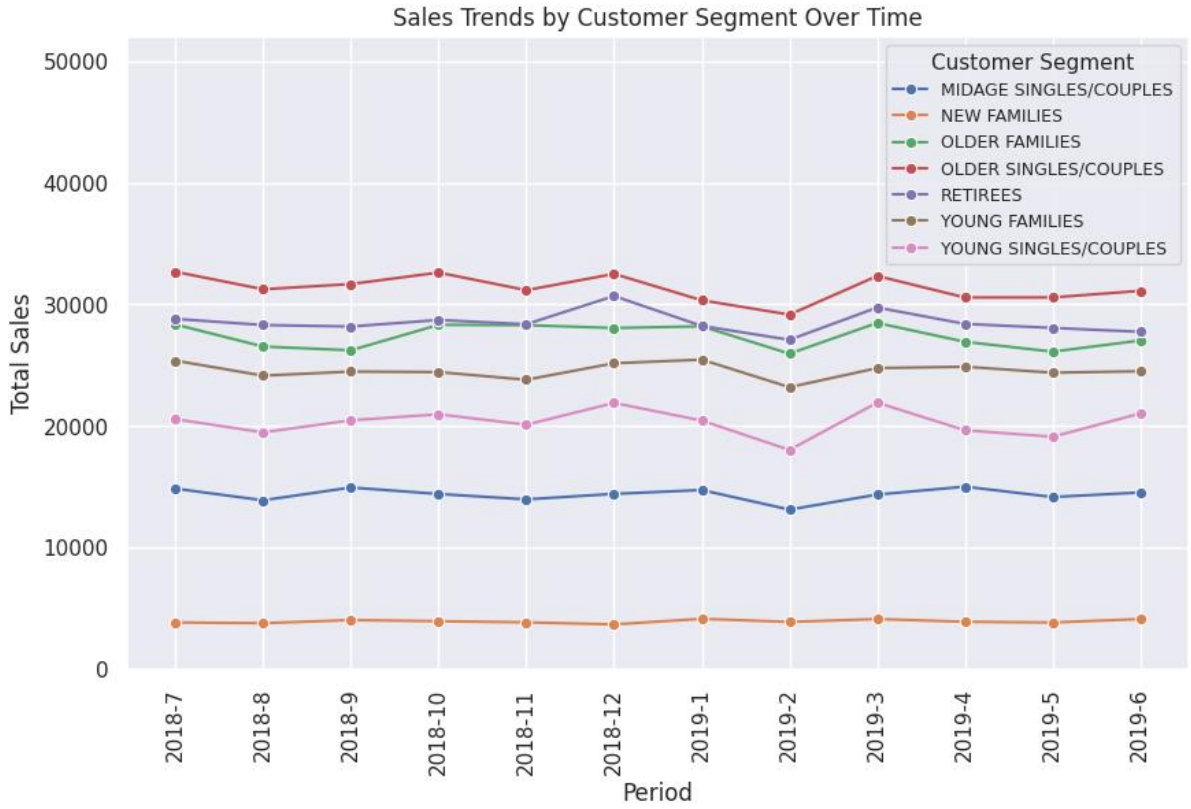
5.9. Temporal Segment Analysis:

The sales trends by customer segment:

Period	LIFESTAGE	TOT_SALES
2018-7	OLDER SINGLES/COUPLES	32683.70
2018-10	OLDER SINGLES/COUPLES	32618.30
2018-12	OLDER SINGLES/COUPLES	32511.30
2019-3	OLDER SINGLES/COUPLES	32341.30
2018-9	OLDER SINGLES/COUPLES	31676.70
...	...	...
2018-11	NEW FAMILIES	3869.20
2018-7	NEW FAMILIES	3859.60
2019-5	NEW FAMILIES	3855.80
2018-8	NEW FAMILIES	3799.25
2018-12	NEW FAMILIES	3696.70

84 rows × 3 columns

- **Older Singles/Couples** consistently lead sales, averaging around 30K - 33K per month, showing their dominance as the largest customer group.
- **Retirees** and **Older Families** follow closely, maintaining sales in the 27K - 30K range, indicating strong and stable purchasing power.
- **Young Families** and **Young Singles/Couples** contribute moderate sales (~20K - 25K), with slight month-to-month fluctuations.
- **Midage Singles/Couples** show lower sales (~13K - 15K), remaining steady but relatively smaller contributors.
- **New Families** are consistently the smallest group, with sales below 5K, highlighting their limited role in total sales.



- **Stability across segments:** Most segments follow a stable pattern over time, with only minor seasonal variations.
- **Core customer base:** Older demographics (Singles/Couples, Families, Retirees) drive the majority of sales, showing brand loyalty and steady demand.
- **Growth opportunities:** Younger and midage groups underperform compared to older segments, suggesting potential for targeted promotions to capture more share.
- **New Families as niche:** Their consistently low sales indicate either limited demand or misalignment with the product category — marketing campaigns may need to be re-evaluated.

This trend analysis reveals that **sales are heavily concentrated in older life-stage groups**, making them the key drivers of revenue.
For sustained growth, the business should **continue catering to older customer segments** while **designing targeted strategies for younger and family-oriented groups** to expand their sales contribution over time.

5.10. Correlation Matrix:

Observations:

- **TOT_SALES vs SELLING_PRICE → Strong positive correlation (0.84)**
    - Revenue is heavily influenced by product price.
- **TOT_SALES vs PROD_QTY → Moderate positive correlation (0.54)**
    - Higher sales volume contributes to revenue but less strongly than price.

- **SELLING_PRICE vs PROD_QTY → Very weak correlation (0.026)**
    - Price and quantity sold are nearly independent → price changes don't strongly affect sales volume.

- **Pricing is the main revenue driver** -- Optimizing price has the greatest effect on total sales.
- **Quantity has secondary impact** -- Increasing units sold helps, but less than price adjustments.
- **Low price–quantity link** -- Suggests demand may be relatively **price inelastic**, giving flexibility for pricing strategies.

The analysis highlights that **selling price is the most critical factor** for driving revenue, while sales quantity plays a supporting role.

- Use **price optimization** (e.g., discounts, premium pricing, bundling) to maximize revenue.
- Support with **quantity-based promotions** (e.g., "buy more, save more") to boost units sold.
- Since demand volume shows little dependence on price, the business can experiment with pricing without risking large drops in sales quantity.

5.11. Price Sensitivity by Demographics:

The average price paid by customer segment:

LIFESTAGE	PREMIUM_CUSTOMER	SELLING_PRICE
YOUNG SINGLES/COUPLES	Mainstream	4.065642
MIDAGE SINGLES/COUPLES	Mainstream	3.994241
RETIREES	Budget	3.924404
RETIREES	Premium	3.920942
NEW FAMILIES	Budget	3.917688
NEW FAMILIES	Mainstream	3.916133
OLDER SINGLES/COUPLES	Premium	3.893182
OLDER SINGLES/COUPLES	Budget	3.882096
NEW FAMILIES	Premium	3.872110
RETIREES	Mainstream	3.844294
OLDER SINGLES/COUPLES	Mainstream	3.814665
MIDAGE SINGLES/COUPLES	Premium	3.770698
YOUNG FAMILIES	Premium	3.762150
YOUNG FAMILIES	Budget	3.760737
OLDER FAMILIES	Budget	3.745340
MIDAGE SINGLES/COUPLES	Budget	3.743328
OLDER FAMILIES	Mainstream	3.737077
YOUNG FAMILIES	Mainstream	3.724533
OLDER FAMILIES	Premium	3.717000
YOUNG SINGLES/COUPLES	Premium	3.665414
YOUNG SINGLES/COUPLES	Budget	3.657366

- Across all **lifestages**, the **average selling price** is fairly consistent, ranging around **3.5 – 4.2**.
- **Mainstream customers** tend to pay slightly more than **Budget** and **Premium** segments in several groups (e.g., *Young Singles/Couples*, *Midage Singles/Couples*).
- Differences across **LIFESTAGE groups** are minor, showing no extreme price variation by age or family status.
- **Retirees, Older Singles/Couples, and Young Families** consistently fall in the **3.8 – 4.0** price range across all segments.

- **Price stability:** Pricing strategy appears uniform across customer groups, suggesting little price discrimination between lifestages.
- **Mainstream customers paying more:** Indicates possible willingness among the middle-market segment to accept higher prices.
- **Premium segment not always highest:** Surprising finding—Premium customers are not consistently paying the most, which could suggest effective discounting or loyalty benefits targeted at them.

The visualization suggests that **average price sensitivity is low across demographic (lifestage) groups**, and instead, **purchase behaviour by segment type (Budget, Mainstream, Premium)** has more influence.

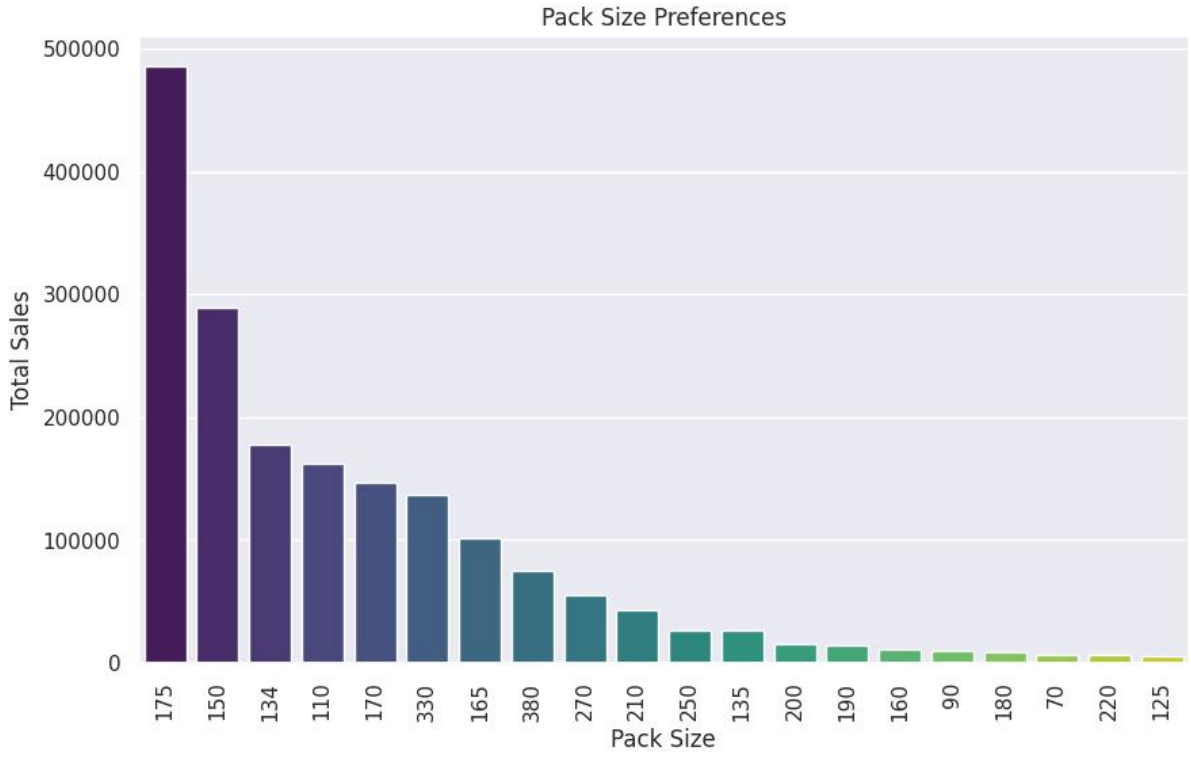
- **Opportunity with mainstream customers** → They already accept slightly higher prices, making them ideal for upselling or bundling strategies.
- **Premium customers** → May be receiving discounts or targeting deals; business could revisit if this aligns with profitability goals.
- **Budget customers** → Their spending is close to Premium, meaning pricing tiers are not sharply differentiated—potential risk of **cannibalization** between segments.

5.12. Pack Size Preference:

The Pack size preferences:

PACK_SIZE	TOT_SALES
175	485437.4
150	289681.8
134	177655.5
110	162765.4
170	146673.0
330	136794.3
165	101360.6
380	75419.6
270	55425.4
210	43048.8
250	26096.7
135	26090.4
200	16007.5
190	14412.9
160	10647.6
90	9676.4
180	8568.4
70	6852.0
220	6831.0
125	5733.0

- **175 pack size** dominates sales by a huge margin (~480K), far ahead of all other pack sizes.
- The **next highest sellers** are **150 (~285K)** and **134 (~175K)** pack sizes, but they are still significantly lower than the 175 size.
- Mid-range pack sizes like **110, 170, 330, and 165** contribute moderately (~140K–180K).
- Larger pack sizes (e.g., **380, 270, 210**) show noticeably lower sales compared to smaller sizes.
- Very small and niche sizes (e.g., **70, 90, 125, 160, 180, 200, 220**) contribute minimally to total sales.



- **Strong consumer preference for smaller pack sizes:** 175, 150, and 134 dominate, suggesting convenience or affordability drives purchase decisions.
- **Mid-size packs (110 - 330 range)** still perform reasonably well, appealing to a broad customer base.
- **Large packs have weaker demand:** Possibly due to higher upfront cost or storage constraints.
- **Skewed sales distribution:** A few pack sizes account for the majority of revenue, while many others contribute marginally.

The analysis shows that sales are highly concentrated in a few pack sizes, especially the 175 pack, which could be the company's flagship offering.

- **Focus marketing and availability** on top-performing pack sizes (175, 150, 134) to sustain strong sales.

- **Evaluate underperforming pack sizes** -- consider reducing SKUs or repositioning them (e.g., bundling large packs with discounts).
- **Mid-range packs provide a balanced opportunity** -- maintain supply for varied customer preferences.
- Insights suggest customers value **affordability, convenience, and portion control** when choosing pack sizes.

5.13. Store & Brand Interaction:

Observations:

- **Kettle** is the strongest-performing brand across all top stores, with sales consistently above **3,600–4,300**, making it the clear leader.
- **Doritos** is the second-best performer, with sales ranging **~1,700–2,060**, showing strong popularity.
- **Pringles** and **Smiths** sit in the mid-range, with sales between **1,500–2,000** and **900–1,300**, respectively.
- Smaller brands like **Twisties, Tostitos, Thins, Cobs, Infuzions** generate lower but steady sales (**600–1,000** per store).
- Store **226** leads in total sales across multiple brands, especially **Kettle (4,301)** and **Pringles (2,013)**.
- Stores **88, 165, and 226** consistently perform well across most brands, indicating strong customer traffic or brand affinity.

- **Brand dominance** -- Kettle is the primary sales driver, with Doritos as the next key contributor.

- **Cross-store consistency** -- Kettle and Doritos are strong across all stores, showing broad consumer appeal.
- **Store-specific strengths** -- Stores like **226 and 88** outperform others, making them ideal for promotions or launches.
- **Smaller brands' role** -- Though weaker individually, niche brands (Cobs, Infuzions, Thins, Twisties) add variety and maintain steady sales.

The heatmap reveals that **a few brands dominate overall sales (Kettle and Doritos)**, while secondary brands provide product diversity.

- **Focus on top brands:** Ensure stock availability and targeted promotions for Kettle and Doritos.
- **Leverage high-performing stores:** Use stores like **226 and 88** as pilots for campaigns or new product launches.
- **Revisit weaker brands:** Consider repositioning, bundling, or promotions for consistently underperforming brands.
- **Balance assortment:** Maintain variety to serve niche customer preferences while prioritizing leading brands.

6. Conclusion

After performing a detailed exploratory data analysis and visualization of the retail sales dataset, several important conclusions were drawn. These findings summarize the key patterns observed across stores, brands, products, customer segments, and time-based sales behaviour. The conclusions are structured below for clarity:

Store-level Insights

- A small number of stores, such as Store 226, 88, and 165, account for the majority of sales.
- This suggests that factors like store size, location, and customer base play a major role in driving revenue.
- These top-performing stores can be used as strategic outlets for piloting new campaigns, product launches, or targeted promotions.

Brand-level Insights

- **Kettle** is the leading brand, consistently dominating both overall and store-specific sales.
- Other strong performers include Doritos, Pringles, and Smiths, while smaller brands make only a minor contribution.

- This reflects strong brand loyalty and concentrated consumer preference around a few key brands.

Product & Pack Size Preferences

- **175g packs** are by far the most preferred, well ahead of other sizes.
- Mid-sized packs (150g–175g) dominate sales, while larger packs sell poorly.
- Customers clearly favour convenient, moderately priced pack sizes, suggesting purchases are often impulse-driven or intended for household consumption.

Customer Demographics

- Older Families, Retirees, and Young Singles/Couples contribute the most to overall sales.
- Budget and Mainstream customers dominate across most life stages, while Premium customers contribute selectively.
- Demographics strongly influence what people buy and how much they spend, but the average number of items per basket remains consistent across groups.

Basket Analysis

- Most baskets contain only 1–2 items, with an average spend of **$5–$10** per transaction.
- This underlines the **snack/convenience nature** of purchases rather than bulk buying.
- Upselling strategies such as "Buy 2, Get 1 Free" could be effective in increasing basket size.

Temporal Sales Trends

- Sales remain fairly stable overall, but **seasonality patterns** are clear:
    - Peaks in December (holiday season) and March (quarterly rebound).
    - A slump in February, likely due to the short month and post-holiday slowdown.
- Week-over-week analysis shows strong front-loading of sales in Weeks 1–4, followed by sharp declines in Week 5.
- These cyclical patterns can guide promotion timing and inventory planning.

Correlation Analysis

- Selling price has the strongest impact on total sales (**correlation = 0.84**) compared to product quantity.
- Price and quantity are largely independent, showing that demand is relatively **price inelastic**.
- This gives businesses flexibility to adjust pricing without drastically affecting volumes.

Store–Brand Interaction

- Kettle and Doritos consistently dominate across top stores.
- Stores such as 226 and 88 perform strongly across multiple brands, making them strategic hubs.
- Secondary brands add variety but contribute far less to overall revenue.

Final Interpretation

- Revenue is heavily concentrated in a few stores, brands, and pack sizes.
- Purchases are small and frequent, fitting the snack/convenience category.
- Demographics and seasonality significantly shape sales patterns, while pricing strategy emerges as the most powerful lever for maximizing revenue.

Overall Conclusion

- Success in this retail business is driven by a small number of dominant stores, brands, and pack sizes, combined with seasonal but stable demand patterns.
- By fine-tuning pricing, focusing on high-value customer groups, and tailoring inventory and marketing strategies, businesses can substantially improve sales performance and profitability.

Based on the findings from this analysis, several recommendations are proposed for improving retail performance and guiding future research. These recommendations focus on promotions, inventory planning, customer segmentation, pricing, and advanced analytics:

Targeted Promotions & Campaigns

- Use high-performing stores (e.g., Store 226, 88) to test new product launches and pilot marketing campaigns.
- Run **multi-buy promotions** to encourage larger basket sizes.
- Plan seasonal campaigns around December and March to maximize sales, and use discounts or loyalty offers to soften the February slump.

Inventory & Assortment Planning

- Prioritize stocking top-selling pack sizes (175g, 150g, 134g).
- Reevaluate underperforming large packs — consider repositioning, bundling, or removing them from the assortment.
- Ensure steady availability of top brands like Kettle, Doritos, and Pringles across all outlets.

Customer Segmentation Strategy

- Focus on Older Families, Retirees, and Young Singles/Couples as the main contributors to sales.
- Tailor promotions differently for Budget and Premium customers, since they both contribute meaningfully but show distinct buying behaviours.
- Explore opportunities among New Families and Mid-age groups with specialized marketing strategies.

Pricing Strategy

- Since demand is relatively price inelastic, experiment with premium pricing for high-demand products.
- Offer bundle discounts to encourage customers to buy more per transaction.
- Track price sensitivity across customer groups to refine targeted offers.

Future Analytical Enhancements

- Apply **time-series forecasting** methods (ARIMA, Prophet, etc.) to predict future sales and guide inventory planning.
- Conduct **market basket analysis** (association rules) to find product combinations and cross-sell opportunities.
- Use **clustering techniques** for more advanced customer segmentation beyond simple demographics.
- Build **predictive models** (e.g., regression, XGBoost) to measure the effect of price, promotions, and demographics on sales.

7. APPENDICES

You may create separate Appendix for the following:

1. References

- https://www.ibm.com/think/topics/exploratory-data-analysis
- https://www.datascienceportfol.io/TharakhGeorgeChacko/projects/2,
- Szabo, B. (2020). How to create a seaborn correlation heatmap in python? ,
- https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e,
- https://studyonline.unsw.edu.au/blog/descriptive-predictive-prescriptive-analytics,
- https://jakevdp.github.io/PythonDataScienceHandbook/,
- https://machinelearningmastery.com/start-here/#timeseries ,
- https://www.rit.edu/ischoolprojects/sites/rit.edu.ischoolprojects/files/document_library/Rashmi_Jeswani_Capstone.pdf

- https://github.com/ArkadyutiPaul/IDEAS-AUTUMN-INTERNSHIP-PROGRAM-ON-DATA-SCIENCE-2025 for the codes developed and other Documents (a copy of this report, data sheet, presentation)