Crunching the Numbers: EDA of Chips Industry Data

A QUANTIUM Job Simulation

Presented by: Avik Sarkhel

Task 1

Data Analysis of Customer Purchase Behavior

Project Overview

- **Project Source:** QUANTIUM Job Simulation (Forage Website)
- **Objective:** Analyze customer purchase behavior in the chips industry
- **EDA Performed Using:** Google Colab (Python, Pandas, NumPy, Matplotlib, Seaborn)

Objective

Goal:

- Analyze customer purchase behavior using transaction and customer datasets.
- Identify trends in spending habits, customer segmentation, and product preferences.

Data Sources:

- **purchase_behaviour.csv** – Contains **72,637 records**, mapping customers to their purchasing category (Mainstream, Budget, Premium).
- **transaction_data.csv** – Contains **264,836 records**, detailing individual transactions, including product details and purchase amounts.

Datasets Used

Dataset 1: Purchase Behavior

- **Records:** 72,637
- **Columns:** Loyalty Card No, Life-stage, Premium Customer
- **Format:** CSV

Dataset 2: Transaction Data

- **Records:** 264,836
- **Columns:** Date, Store No, Loyalty Card Number, Transaction ID, Product No, Product Name, Product Quantity, Total Sales
- **Format:** Worksheet (converted to CSV for processing)

Data Processing

- **Imported Libraries:** Pandas, NumPy, Matplotlib, Seaborn
- **Uploaded Datasets to Google Drive & Mounted in Colab**
- **Merged Datasets on 'Loyalty Card No'**
- **Final Merged Dataset:** 264,836 records, 10 columns
- **Unique Loyalty Card Numbers:** 72,637
- **Checked for Null Values:** None Found

Data Cleaning

Company Name Standardization

- Found inconsistencies in company names (e.g., 'Red Rock Deli' vs. 'RRD')
- Created a dictionary to map all variations to standardized names
- Used ChatGPT for initial mapping, then manually refined it
- Created a new column: **'Company Name'**

Removed Incomplete Records:

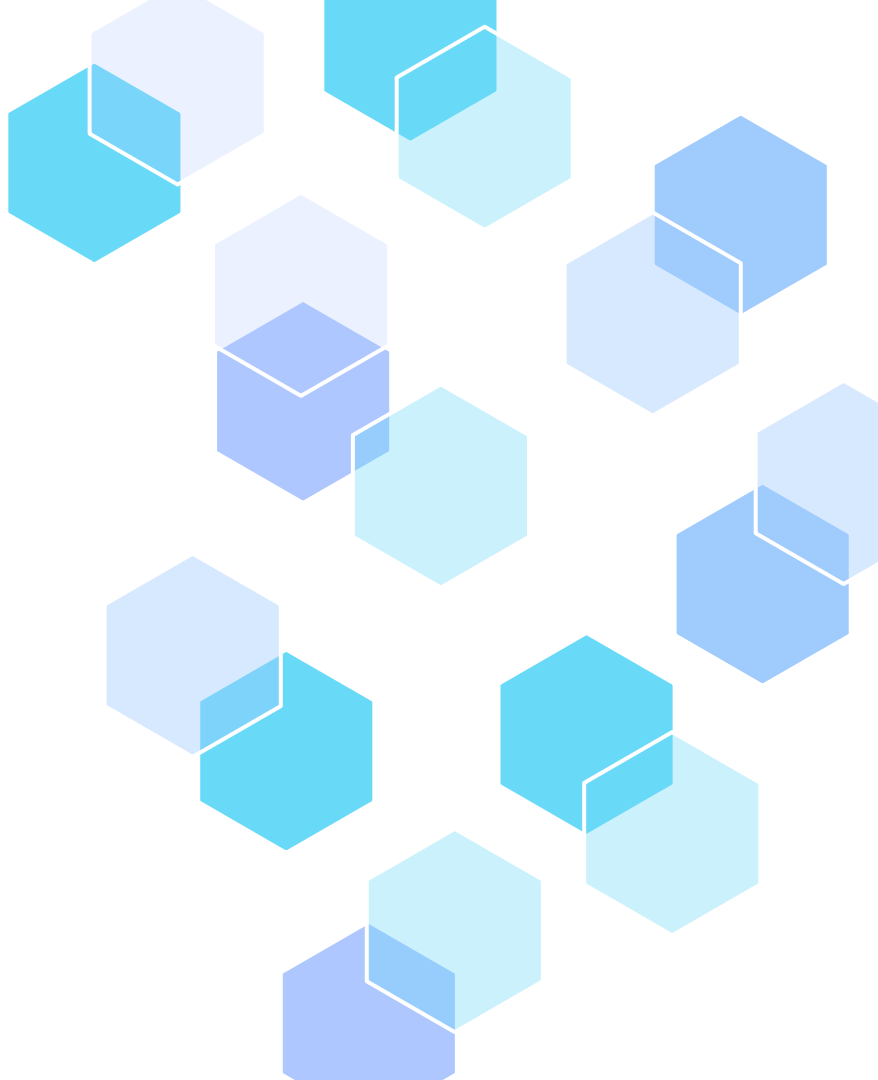
- 'French Fried Potato Chips' had no company name
- Removed **1,418 records** containing this product

Date Format Issue: Unable to fix, kept as-is

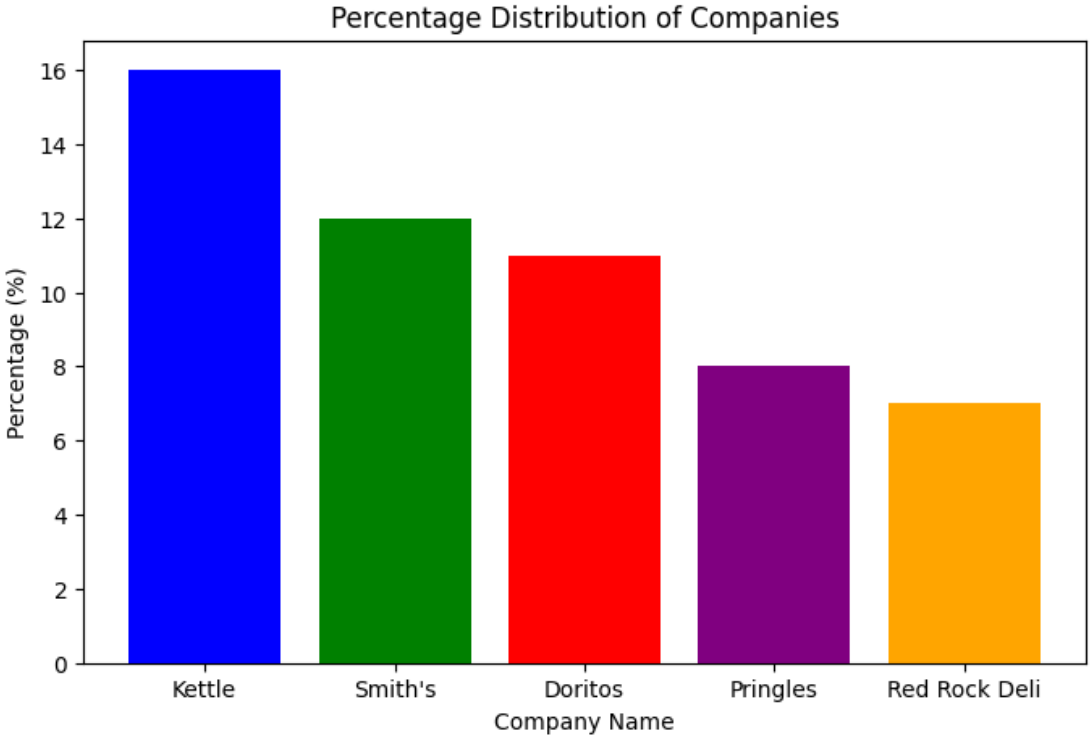
Exploratory Data Analysis (EDA)

Key Analysis Areas:

- Top 5 Chip Brands
- Customer Segmentation by Lifestage (Age Groups)
- Customer Segmentation by Premium/Budget/Mainstream (Premium Customer Column)

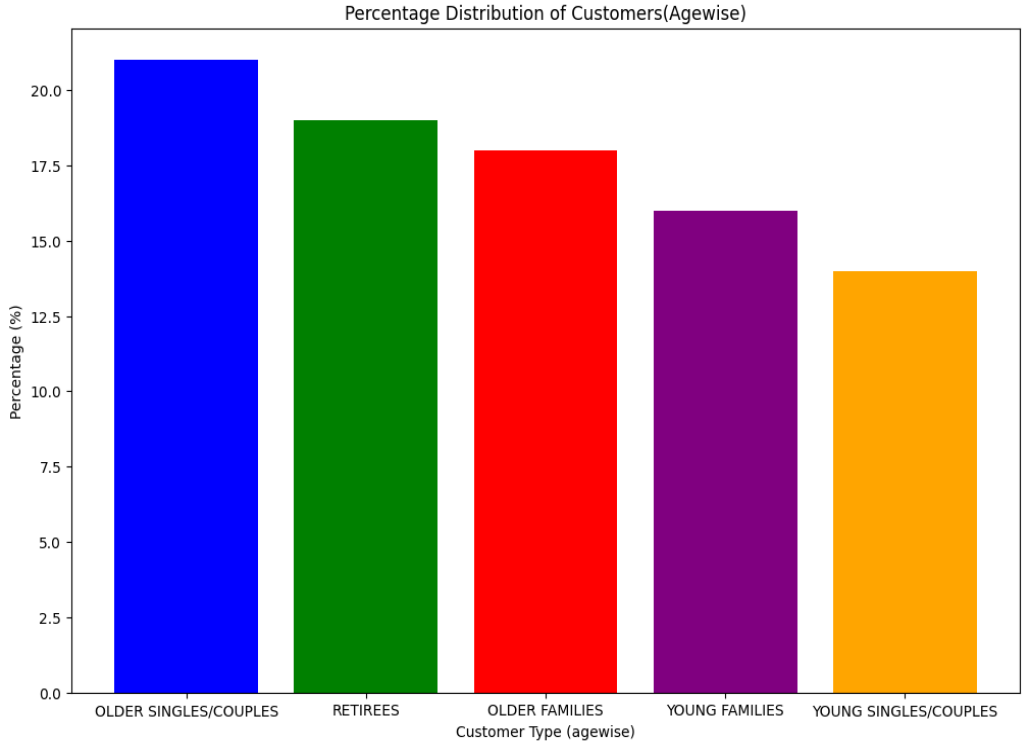


Visualisation 1: Top 5 Chip Brands



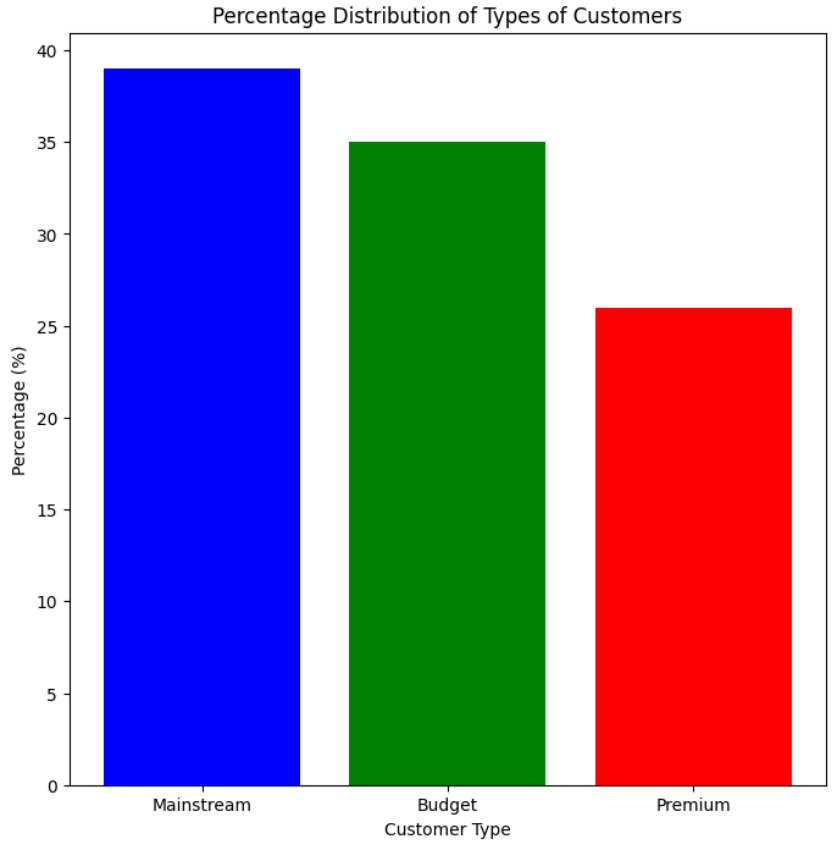
- **Kettle (16%)** is the most popular chip brand, leading in sales.
- **Smith's (12%) and Doritos (11%)** are the next most purchased, showing strong consumer preference.
- **Pringles (8%) and Red Rock Deli (7%)** have a smaller but significant market share.
- The trend suggests that **consumers prefer premium or well-established brands** in the chips market

Visualization 2 - Customer Segmentation by Life-stage



- **Older Singles/Couples (21%)** are the largest segment of chip buyers, indicating that this group has a strong preference for chips.
- **Retirees (19%)** and **Older Families (18%)** also have a significant share, suggesting that older age groups are key consumers in this market.
- **Young Families (16%)** and **Young Singles/Couples (14%)** have a lower but still considerable presence, indicating that younger demographics also engage in chip purchases, though to a lesser extent.
- This trend suggests that **chips are popular across all age groups**, but older individuals and families tend to purchase them more frequently.

Visualization 3 - Customer Segmentation by Premium/Budget/Mainstream



- **Mainstream customers (39%)** form the largest segment, indicating that the majority of customers prefer standard-priced chips rather than high-end or budget options.
- **Budget customers (35%)** are also a significant group, suggesting that a large portion of customers are price-conscious and opt for more affordable chips.
- **Premium customers (26%)** represent the smallest segment, implying that fewer consumers are willing to pay extra for high-end chip brands.
- This data suggests that chip companies should primarily focus on mainstream and budget product lines while maintaining premium options for niche consumers.

Task 2

Finding Control Stores for Trial Stores

Project Overview

- **Objective:** Identify control stores for trial stores based on sales performance metrics.
- **Dataset:** Cleaned dataset with 264,834 records and 12 columns.
- **Trial Stores:** 77, 86, 88
- **Trial Period:** February 2019 – April 2019
- **Data Range:** July 2018 – June 2019

Control Store Selection Methodology

Performance Similarity Measures:

- Total Sales Revenue
- Total Number of Customers
- Average Transactions per Customer

Methods Used:

- Magnitude Distance Formula
- Pearson Correlation

Pre-trial & Trial Period Division:

- Pre-trial: July 2018 – January 2019
- Trial: February 2019 – April 2019

Control Store Selection Process

Data Preprocessing:

- Converted DATE column to datetime format
- Filtered dataset for the required date range
- Removed 'French Fried Potato Chips' records

Aggregated Store Metrics:

- Computed total sales, unique customers, and avg. transactions/customer

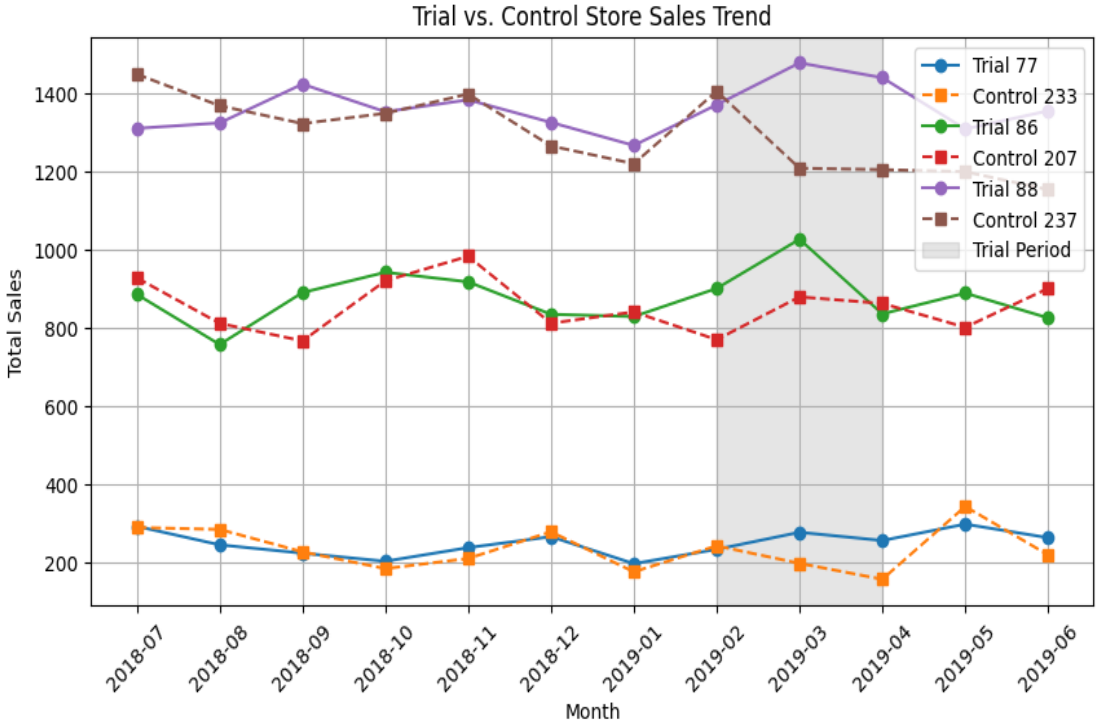
Similarity Calculation:

- Calculated Euclidean Distance and Pearson Correlation

Selected Control Stores:

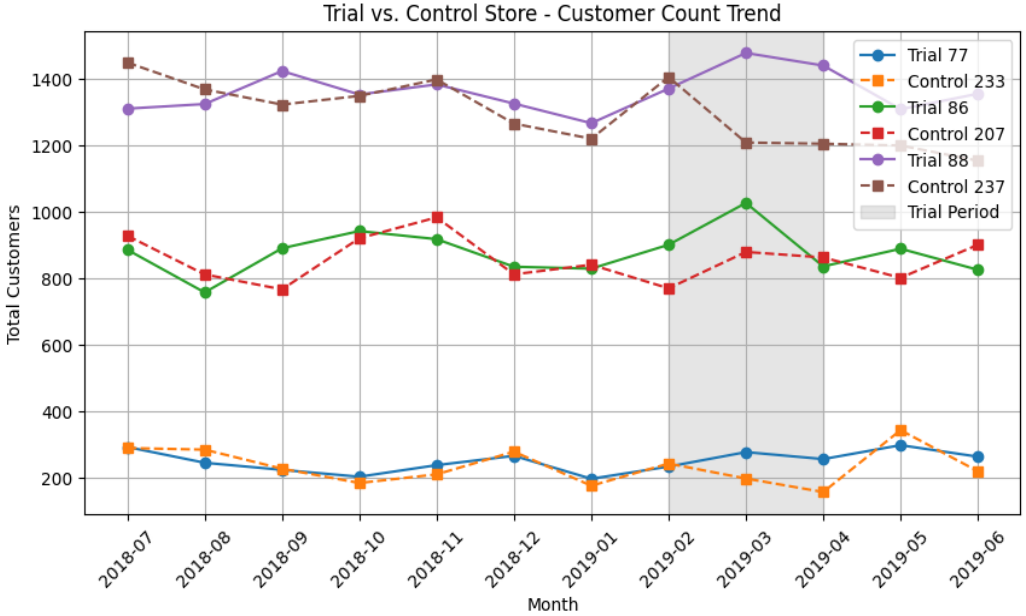
- Identified best-matching stores for each trial store

Visualisation 1: Total Sales Comparison



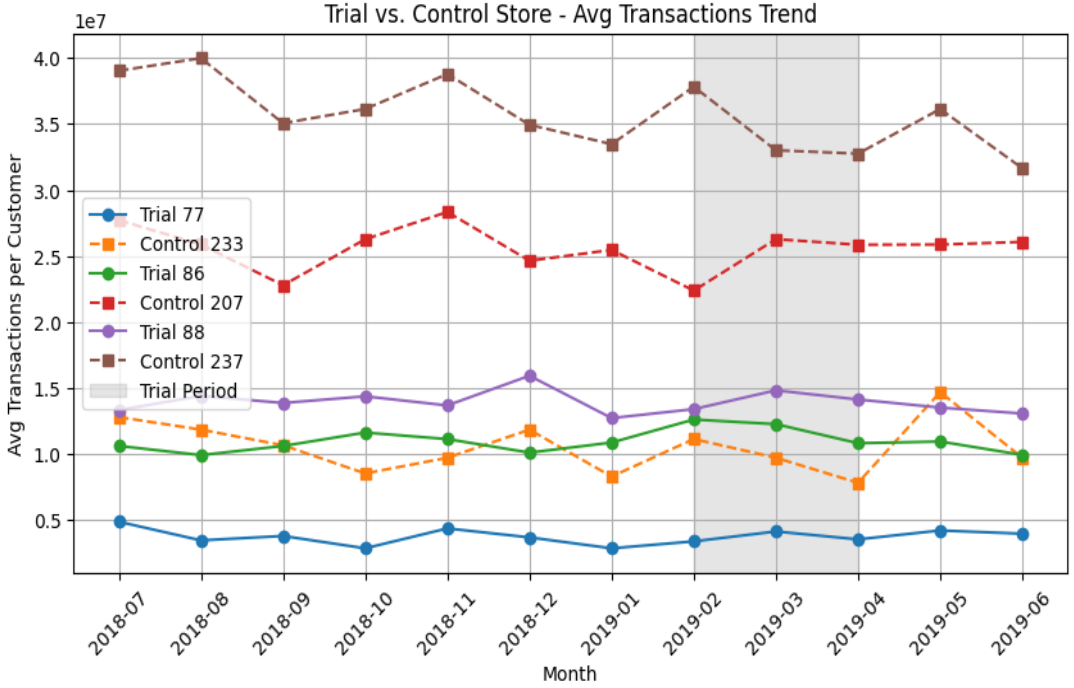
- Trial's impact on sales varied across stores. Some trial stores saw increased sales, while others didn't.
- **Trial 88** and **Control 237** consistently had the highest sales.
- **Trial 77** and **Control 233** consistently had the lowest sales.
- Sales fluctuated throughout the year, suggesting seasonality or other factors.
- The trial period seemed to have a more noticeable impact on customer count than on sales.
- **Store 86 saw the most significant increase** in customers during the trial but **also had a declining** trend before and after.

Visualization 2: Total Customers Comparison



- **Trial stores (77, 86, 88) generally saw an increase in customer count during the trial period** (Feb–Apr 2019), while control stores remained relatively flat.
- **Trial store 86 showed the most significant increase during the trial**, but also had a declining trend before and after.
- **Control store 233 had the highest overall customer count.**
- **Performance of trial stores varied**, suggesting the trial had different impacts depending on the store.

Visualization 3: Avg Transactions Per Customer



- **Control 237** consistently had the **highest** average transactions per customer.
- **Trial 77** consistently had the **lowest** average transactions per customer.
- **Trial 86** showed **a slight increase** in average transactions during the trial period, but the effect is not dramatic.
- The trial period (shaded gray) had a limited and varied impact on average transactions across different stores.
- There are fluctuations in average transactions across all stores throughout the year, suggesting potential seasonality or external factors.
- The average transactions per customer for most stores remained relatively stable throughout the year, with no significant upward or downward trends.

Conclusion

Key Findings from Task 1

- Identified top chip brands, customer segments, and spending behaviors.
- Older Singles/Couples and Retirees dominate chip purchases.
- Mainstream and budget customer segments form the majority of the market.

Key Findings from Task 2

- Successfully identified control stores using similarity measures.
- Sales, customer count, and transaction patterns closely match trial stores.
- Control stores provide a reliable benchmark for trial store performance evaluation.

Next Steps

- Further analysis on trial store performance post-trial period.
- Evaluate marketing intervention effectiveness

Thank You!

Have any questions? Feel free to reach out:

📧 Email: avik305sarkhel@gmail.com

🔗 Job Simulation Link [Click Here]

📁 GitHub Repository [Click Here]