BrainStation

Capstone Project

Prediction of Next Purchase Time and Quantity by Customer and Product

2024

Project Overview

Objective of the project:

The objective of this project is to **predict the time until the next purchase by customers and the quantities to be bought**, using their purchase history, quantities purchased, and product prices.

By applying predictive models, **the goal is to anticipate when customers will make a transaction again**. With this information, the aim is to optimize business and inventory decisions.

Motivation:

This is a real work case:

A customer is requesting a report (query) **to calculate the time and quantity products to be purchase for each client.**

Considering that a simple calculation of average purchase times and quantities is not sufficient, the idea is to **apply a more efficient model to predict this information.**

About the Company

Company Details:

For security reasons, the company's name is omitted.

The company is located in the USA and **specializes in the manufacturing and printing of foodservice products.**

Its main lines of business include: napkins, cups, bags, and TechnoLiners.

Napkins

Cups

Paper Bags

Techno Liners

Project Overview

Proposed Solution:

This project aims to answer the following two questions:

1. **What is the estimate purchase time for each customer?** With this information, we can estimate when the customer will make their next purchase.
2. **What is the estimate purchase quantity of the items?** This information allows us to anticipate the inventory quantities needed to cover future customer orders.

Customer	Purchase Time	Time to next Purchase
C001	Jan	0
C001	Feb	30
C001	Apr	60
C001	Jul	90
C001	**When is the next Purchase?**	

Dataset:

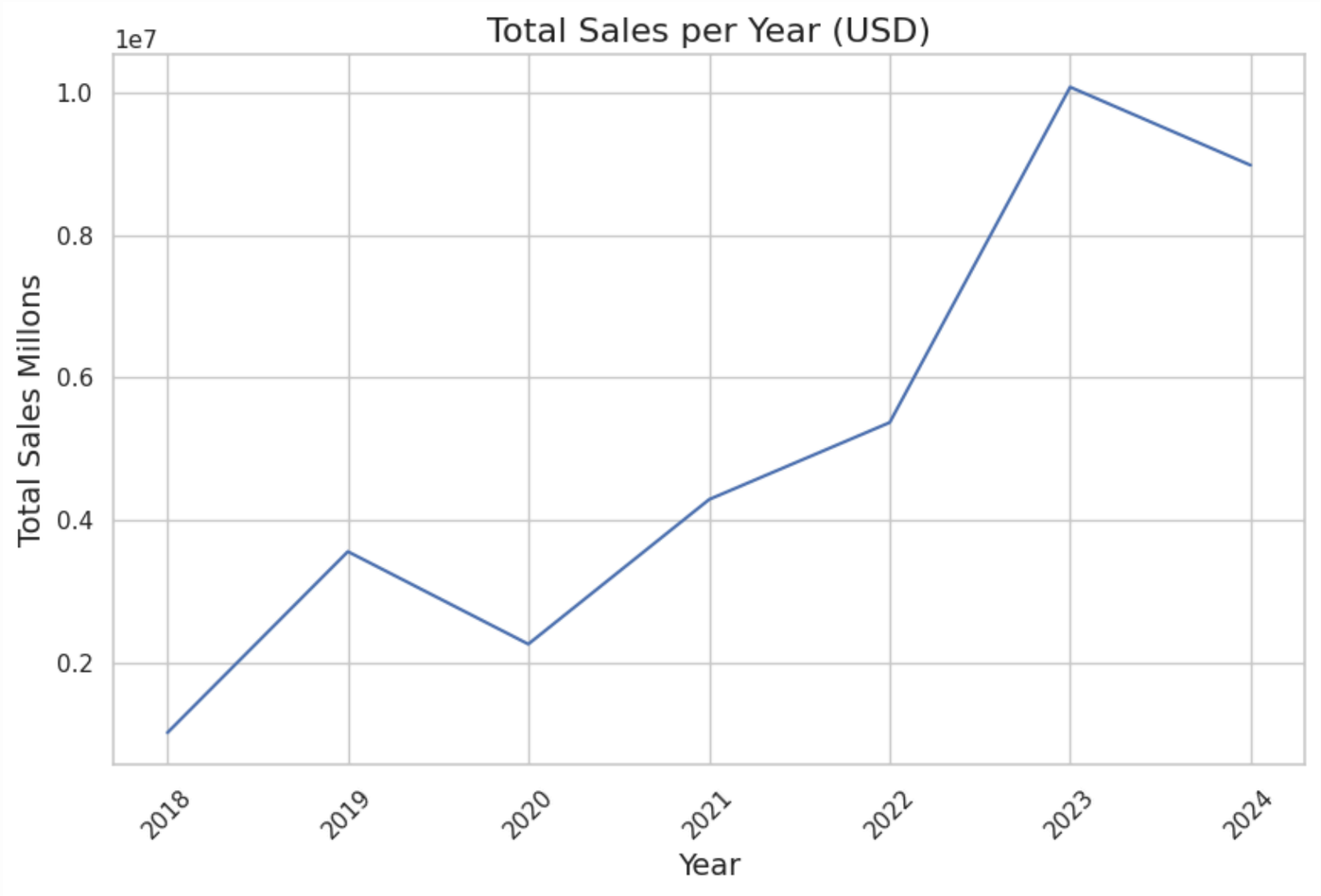
The dataset has been obtained directly from the company's ERP through a query, extracting historical information of sales, such as customers, invoices, items, quantities, totals, and other fields.
The records are since 2018 to 2024 Oct.
Rows of sales 16724 and 28 columns.

For security reasons, the dataset will not be public.

Project Execution

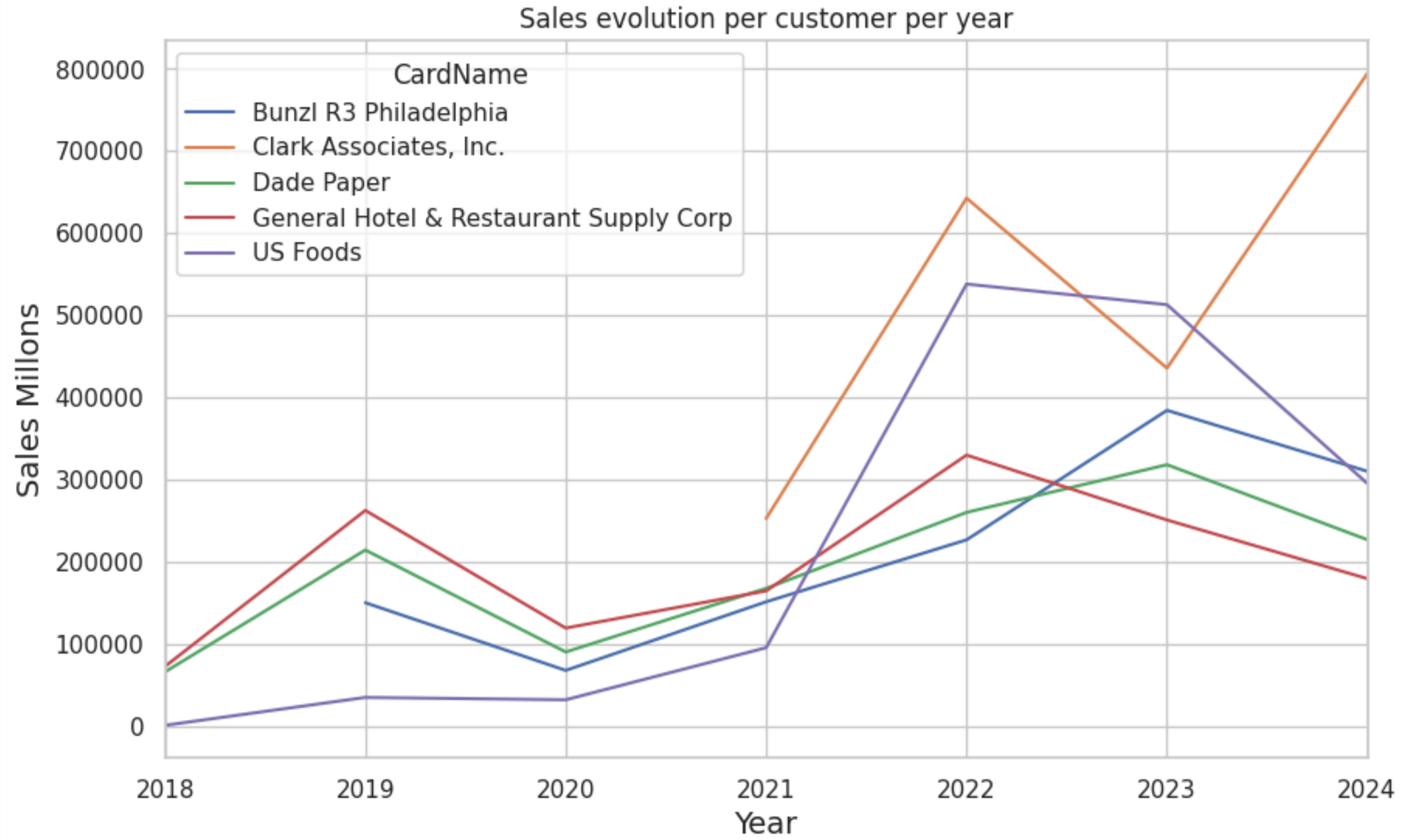
Preliminay EDA findings:

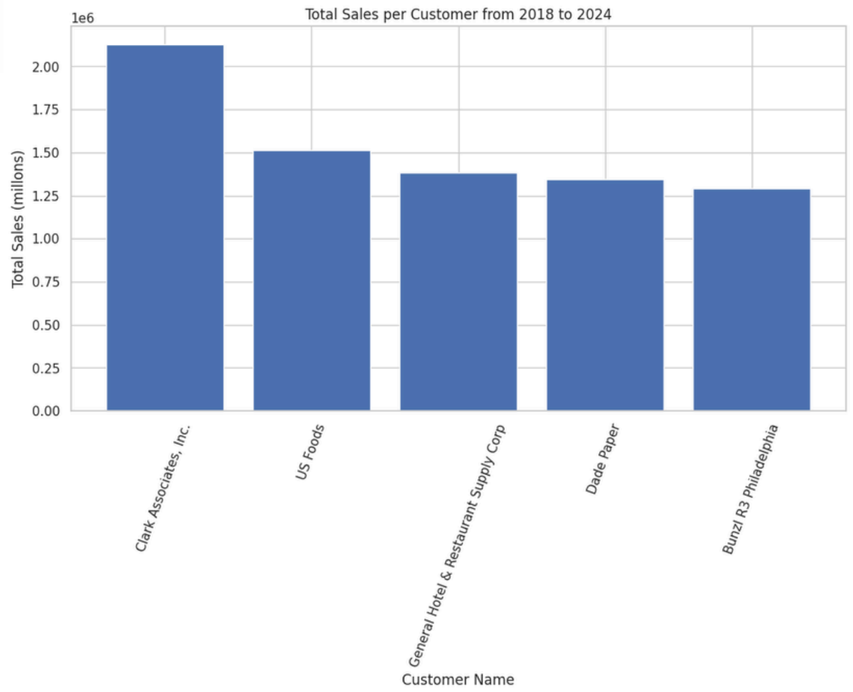
A global analysis of overall sales, sales by customers, and sales by products is conducted to identify the best customers and the best-selling products for each customer.



- The sales evolution shows a steady increase since the beginning of operations in 2018.
- The trend from 2020 to 2023 is very clear, and despite 2022 being a pandemic year, there is no noticeable impact on the trend's evolution.

Top 5 Customers by Sales Amount

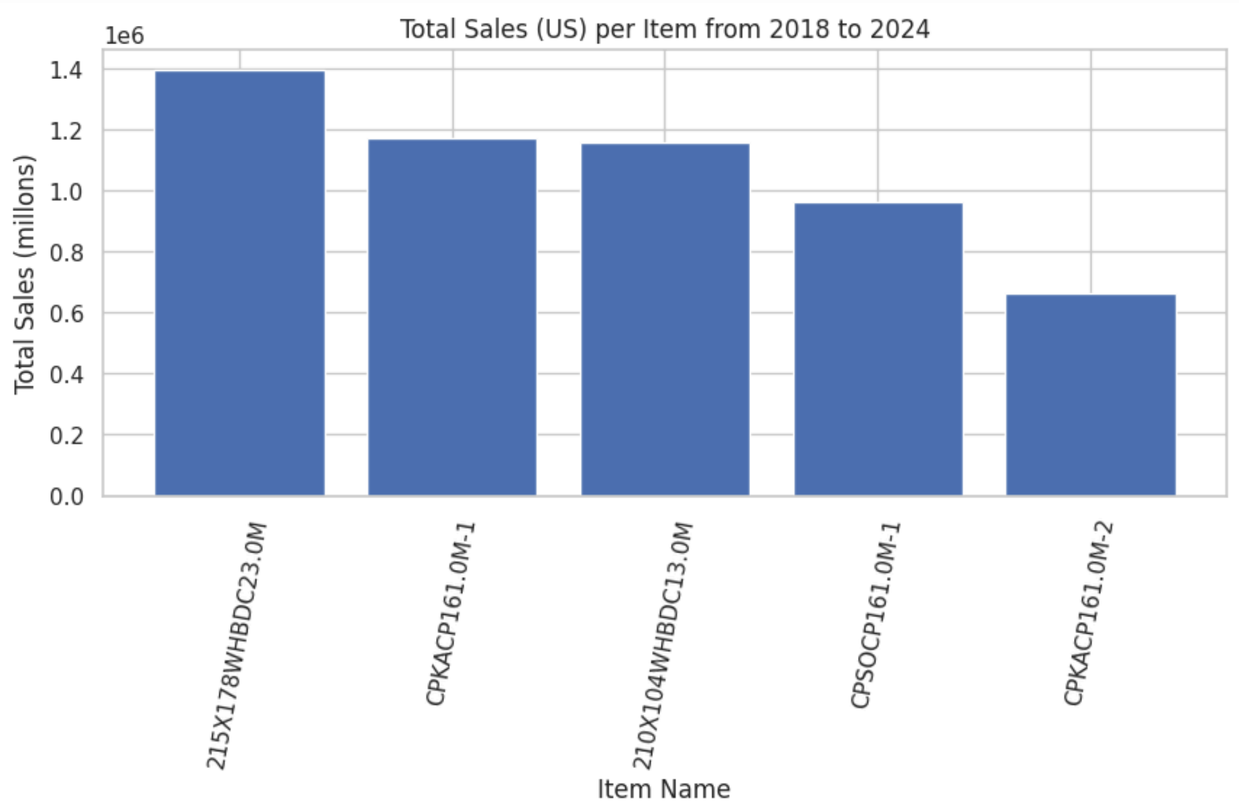


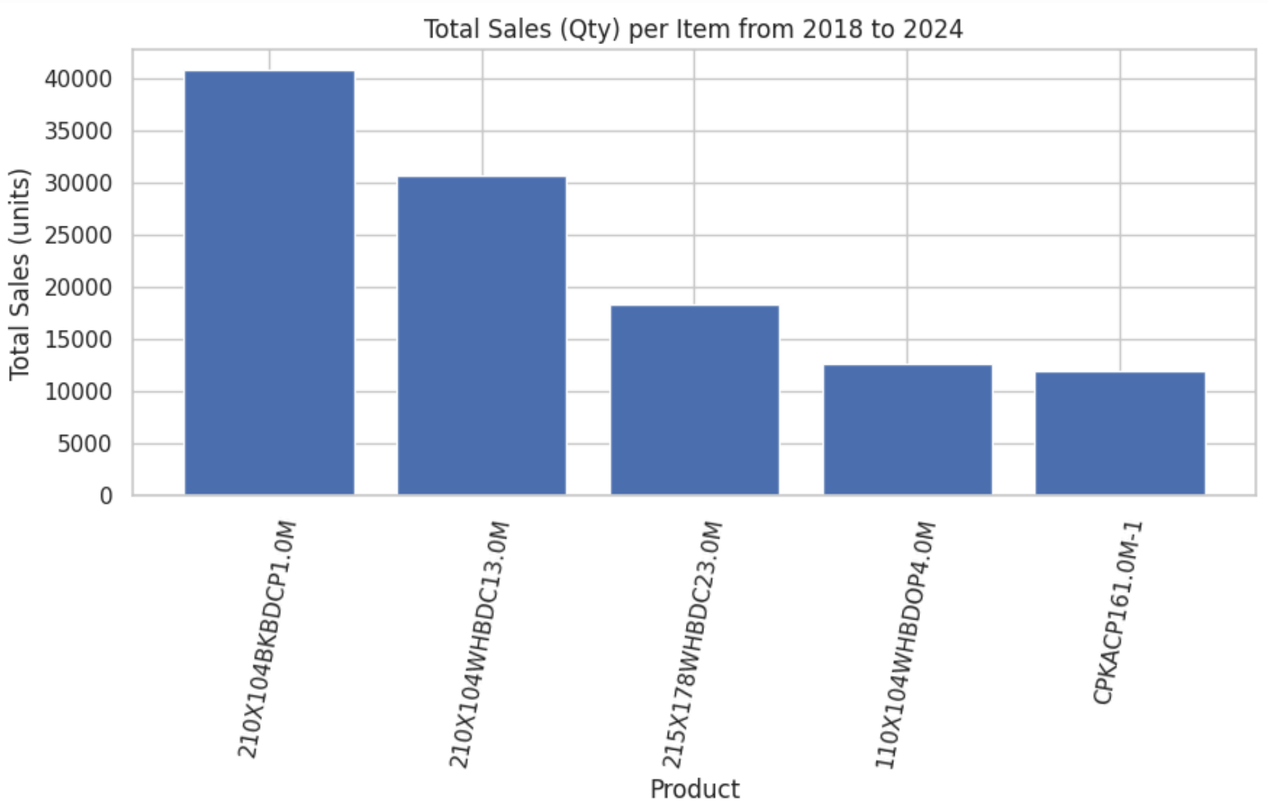


- Customer Clark Ass., is the primary client based on purchase amount.
- US Foods, the second largest client, has been making purchases since 2018 but shows a decline since 2022.
- The rest exhibit very similar purchasing patterns.

Project Execution

Top 5 Products by Sales Amount and Quantity

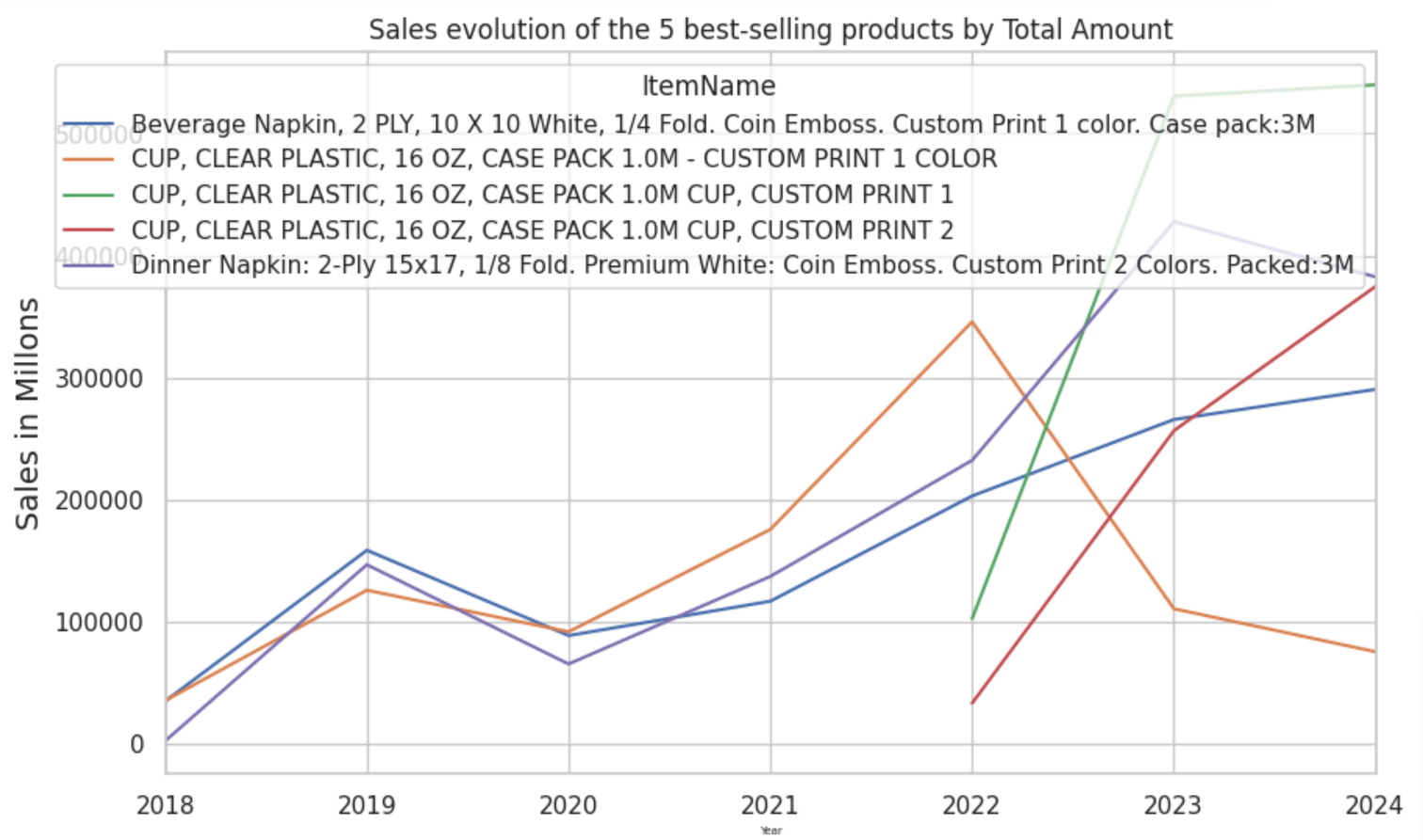


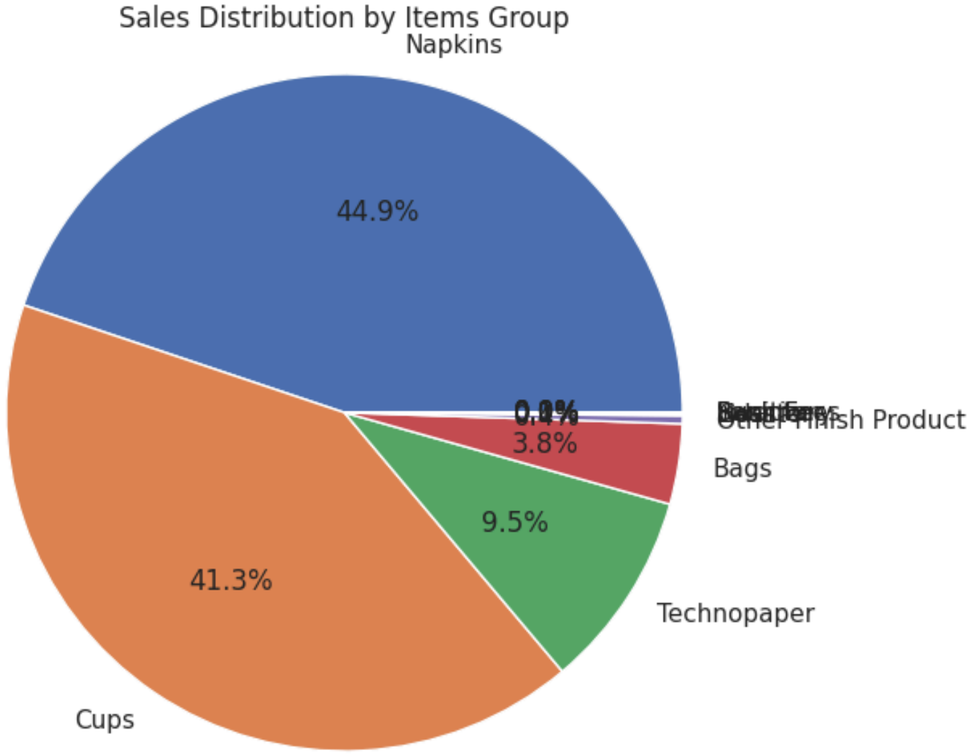


Some of the top products appear in both charts, indicating that they not only lead in quantities sold but also have a significant impact on the company's revenue.

Project Execution

Top 5 Products by Sales Amount





Napkins and Cups family items represent the high concentration of income for the company.

Project Execution

Modeling Approach

Calculation of Estimated Purchase Time and Quantities per Customer

- Simple Average Method
- Sales Frequency of Products by Customer
- Sales average - time and quantity purchased per item and per customer
- Linear Regression for Estimated Purchasing Time

Simple Average Method

We will begin by calculating a simple **average** of the **time between each purchase** made by each customer.

	CardName	avg_days_between_purchase
0	Bunzl R3 Philadelphia	42.704545
1	Clark Associates, Inc.	0.708383
2	Dade Paper	4.422619
3	General Hotel & Restaurant Supply Corp	4.067890
4	US Foods	5.584184

'avg_days_between_purchase': estimated days of purchase

If customer 'Bunzl R3', buy today, next purchase would be in approximately 42 days.

Disadvantages

- **Variability in purchase patterns:** It assumes regular purchases, which does not reflect well for customers with irregular or seasonal buying habits.
- **Does not detect trends:** It does not capture if a customer is changing their purchase frequency or if external factors are influencing their behavior.

Possible Improvements

- **Median instead of average:** it is less affected by unusually long or short intervals between purchases.
- **Weighted calculation:** Weight the time between purchases based on the quantity or value of transactions.
- **More advanced models:** Linear Regression, Time series analysis techniques.

Sales Frequency of Products by Customer

	CardCode	ItemCode	sales_count
0	C00086	Beverage Napkin, 2 PLY, 10 X 10 White, 1/4 Fol...	141
1	C00129	Beverage Napkin: 1-Ply 10x10, 1/4 Fold. Regula...	91
2	C00335	Beverage Napkin, 2 PLY, 10 X 10 White, 1/4 Fol...	28
3	C00421	Dinner Napkin: 2-Ply 15x17, 1/8 Fold. Premium ...	28
4	C00706	CUP, CHOICE, CLEAR PLASTIC, 16 OZ, CASE PACK 1...	374

Including Item and the combination of customer and their most purchased product to calculate estimated purchase time not only by customer but also by item.

	CardCode	ItemCode	avg_days_between_purchases	avg_quantity_purchased
17	C00086	210X104WHBDC13.0M	15.557143	47.262411
108	C00129	110X104WHBDC16.0M	24.188889	81.329670
176	C00335	210X104WHBDC13.0M	67.407407	21.107143
287	C00421	215X178WHBDC23.0M	69.592593	530.928571
311	C00706	CPCHCP161.0M-1	1.134048	7.048128

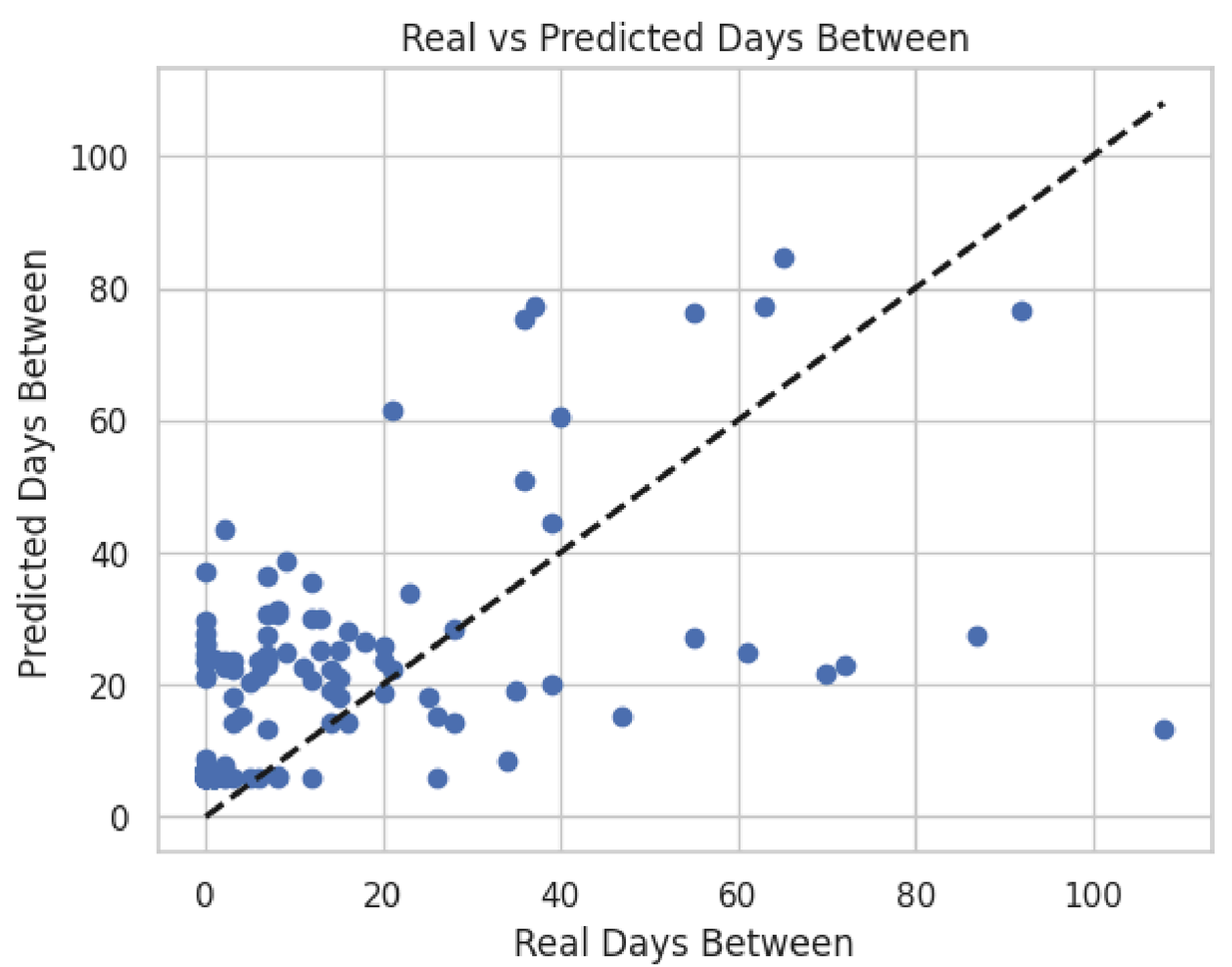
Now we have 2 new calculations:

- Average Days between Purchase
- Average Quantity Purchased

With this first approach we have answered the 2 questions of the project:

1. What is the estimated purchase time for each customer?
2. What is the estimated purchase quantity of the items?

Linear Regression for Estimated Purchasing Time



Linear regression to predict the **days between purchases (y)** for each customer based on **quantity and price (X)** for each customer-product.

- **Mean Squared Error (MSE): 350**

MSE of 350.41 is relatively high, indicating that the model's predictions tend to deviate significantly from the actual values.

- **R-squared (R^2): 0.13**

Only 13% of the variability in purchase intervals is being explained by the model. This means the model is not capturing most of the variation in purchase intervals.

The relatively high MSE and low R^2 suggest that the current Linear Regression **model is not performing optimally in predicting the time between purchases**. The model struggles to fit the data and explain the variability in purchase intervals.

Cristhian Lima

THANKS