Project Report

Instacart Market Basket Analysis

Prof. Dr. Thomas Lavastida

BUAN 6340 – Programming for Data Science

Analysis Done By

- Kushagra Rastogi (kxr220031)
- Varun Vinod (vxv220005)
- Aditya Kiran Anasane (aka220002)

Executive Summary

This project presents a comprehensive analysis of customer purchase patterns and behaviors using Instacart Market Basket Analysis. Through exploratory data analysis, popular days and times for orders were identified, along with insights into reorder frequencies. Utilizing clustering techniques, customer segments were identified based on purchasing habits. Additionally, a predictive model was developed to forecast product reorders, enabling Instacart to enhance user experience and increase revenue. Recommendations include tailored marketing strategies and product recommendations to optimize customer satisfaction and sales.

Market Basket Analysis is a data mining technique used in retail settings to identify associations between items. The method involves analyzing combinations of items that appear together frequently in transactions, allowing retailers to identify relationships between the items that customers purchase. In simpler terms, it helps retailers find patterns in customer behavior with respect to item purchases.

The project provides an analysis of customer purchase patterns, frequently purchased items and units from Instacart to maintain sufficient product inventory and facilitate reordering. The team aims to identify customer clusters and subgroups with comparable purchasing behaviors to provide useful recommendations and improve revenue and customer experience via segmentation and prediction models.

Furthermore, we'll try to predict whether a product will be reordered or not by using order-related and time-related features from the dataset. The results and insights from the analysis will be utilized to enhance user experience by recommending likely products for customers to purchase during the order process. Additionally, a marketing strategy for Instacart and other similar retailers will be proposed, including customized communications, that highlight anticipated products to remind customers to reorder. In conclusion, this project offers Instacart valuable insights to enhance customer experience and increase revenue by employing efficient segmentation and prediction models.

DATASET AND PREPROCESSING

The Instacart Market Basket Analysis Kaggle dataset consists of 6 different datasets that provide transactional and purchasing details of customers. The first two datasets, Aisles and Departments, provide information on the names and IDs of the aisles and departments where products were organized. The third and fourth datasets, Order_Products_prior and Order_Products_train, provide information on the orders, products, and reordered products. The fifth dataset, Orders, provides information about customer orders such as order ID, order number, week day of the order, hour of the order, user ID, and days since prior order. The sixth dataset, Products, gives information on the products such as product name, product ID, aisle and departments where they were sold. The objective of this project is to predict the next products customers tend to purchase and whether a product is reordered or not in their next purchase.

A few details about the dataset are as follows:

No. of Rows: 33819106
No. of Columns: 15

Some of the key data points required for the analysis in this project are: Transactional details of the products, order number, date and time of the transactions, aisles and departments where the product belongs, Reordered details, time and day details of the products ordered, transactional details, departments and aisles details

Table: Aisles

Variables	Description
Aisle ID	Labels the ID of the aisles
Aisle Name	Mentions the aisle name in the retail stores

Table: Department

Variables	Description
Department ID	Labels the ID of the departments
Department name	Mentions the department name in the retail stores

Table: Order_Products_prior

Variables	Description
Order_id	Id of the order
Product_id	Id of the products
add_to_cart_order	Order which add to cart
Reordered	The product which re-order

Table: Order_Products_train

Variables	Description
Order_id	Id of the order
Product_id	Id of the products
add_to_cart_order	Order which add to cart
Reordered	The product which re-order

Table: Orders

Variables	Description
Order ID	Labels the ID of the order made by customers
User ID	Labels the ID of the users who made the purchase
Order number	Denotes the order number made by the customer
Order_dow	Denotes the day of the week, the order made by the customer

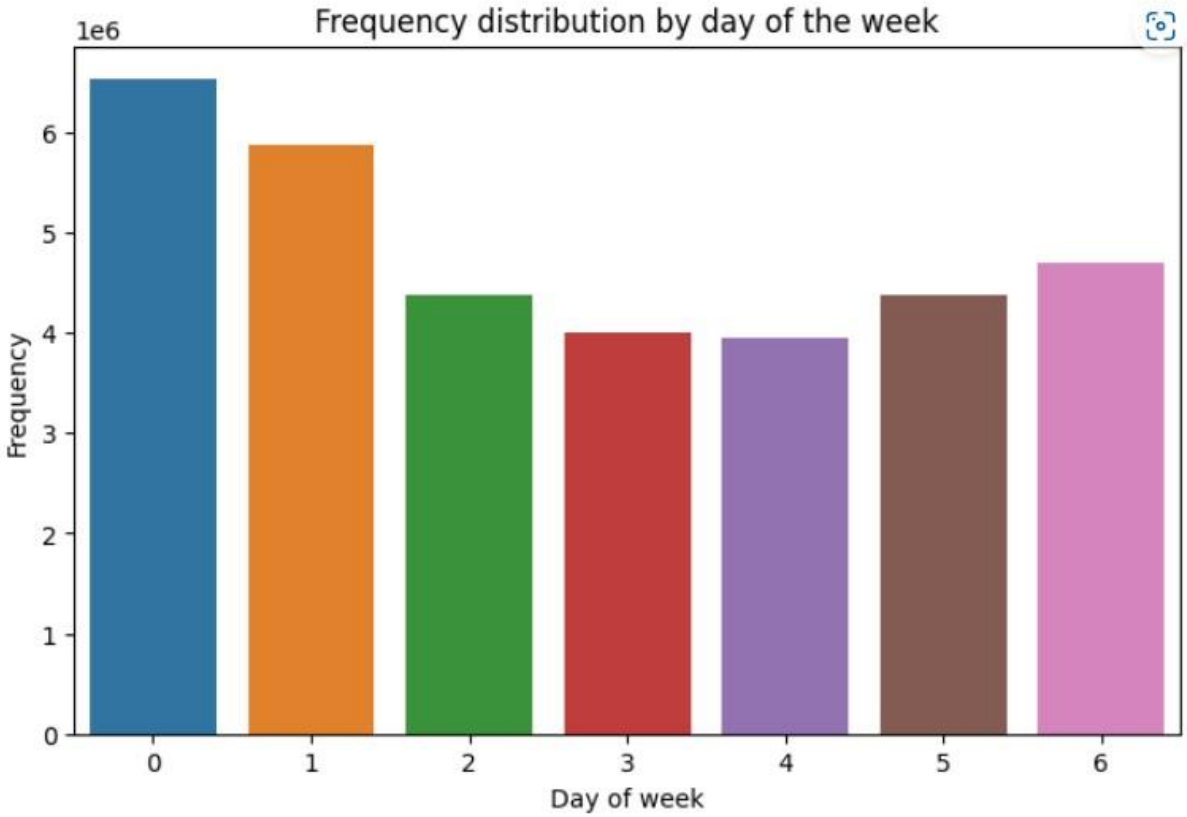
Order hour of day	Denotes the hour of the day, the order made by the customer
Days since prior order	Denotes the number of days since last order

Table: Products

Variables	Description
Product ID	Labels the ID of the products purchased by customers
Product Name	Denotes the product name purchased by the customer
Aisle ID	Labels the ID of the aisles
Departments ID	Labels the ID of the departments

EXPLORATORY DATA ANALYSIS

1] What is the most popular day of the week for orders?

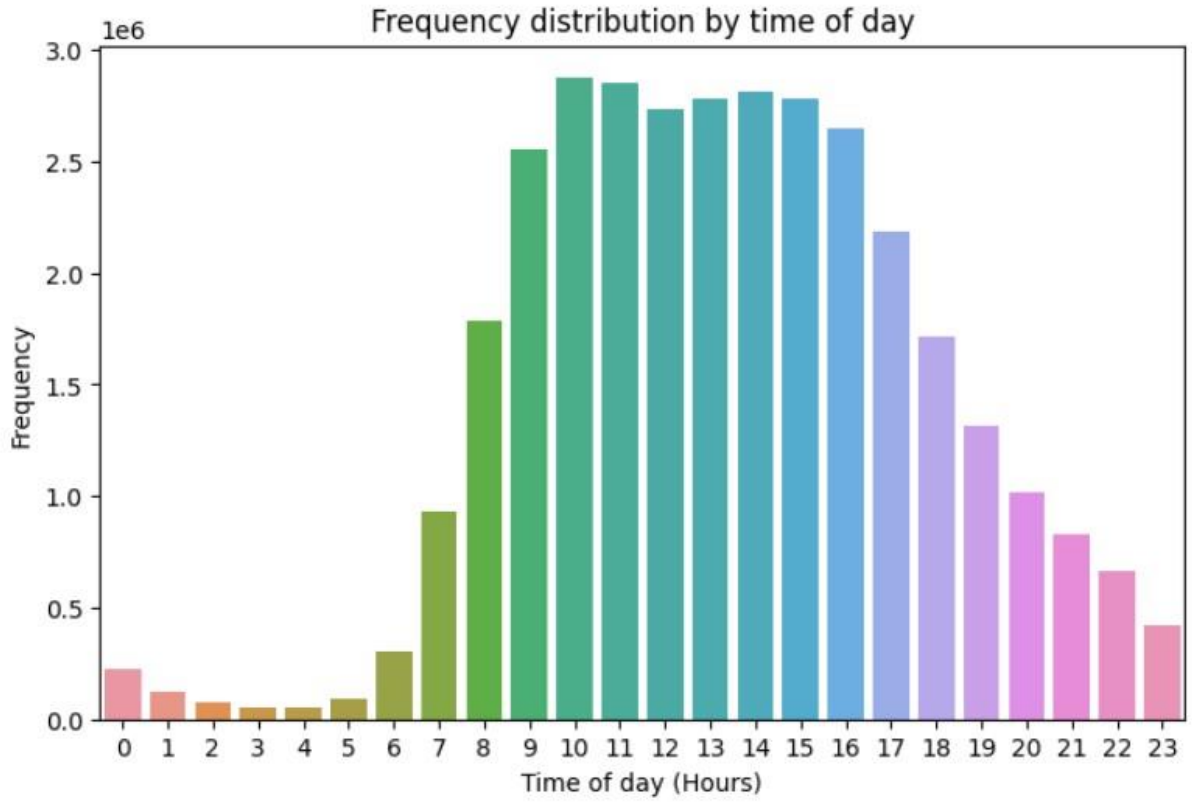


A logical and intuitive deduction from the frequency distribution of orders by day of week, would be:

- Days 0, 1 -> Weekends
- Days 2, 3, 4, 5, 6 -> Weekdays Assuming Day 0 as Saturday:

- Weekends followed by Friday seem to be more spend heavy days, i.e, more no. of orders are made during weekends
- Weekdays exhibit lower user activity ,i.e, fewer no. of orders are made during weekdays

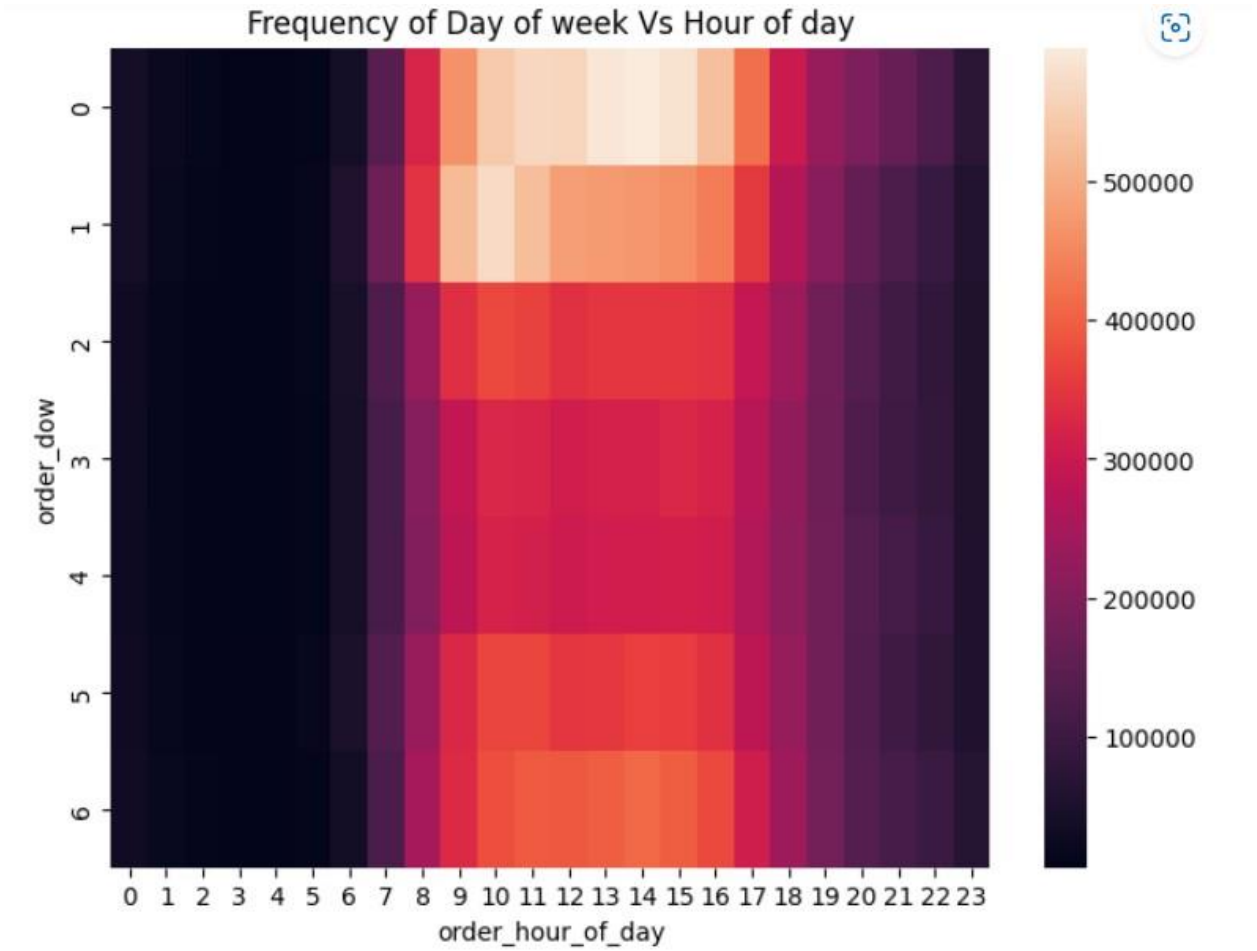
2] What is the most popular time of the day for order



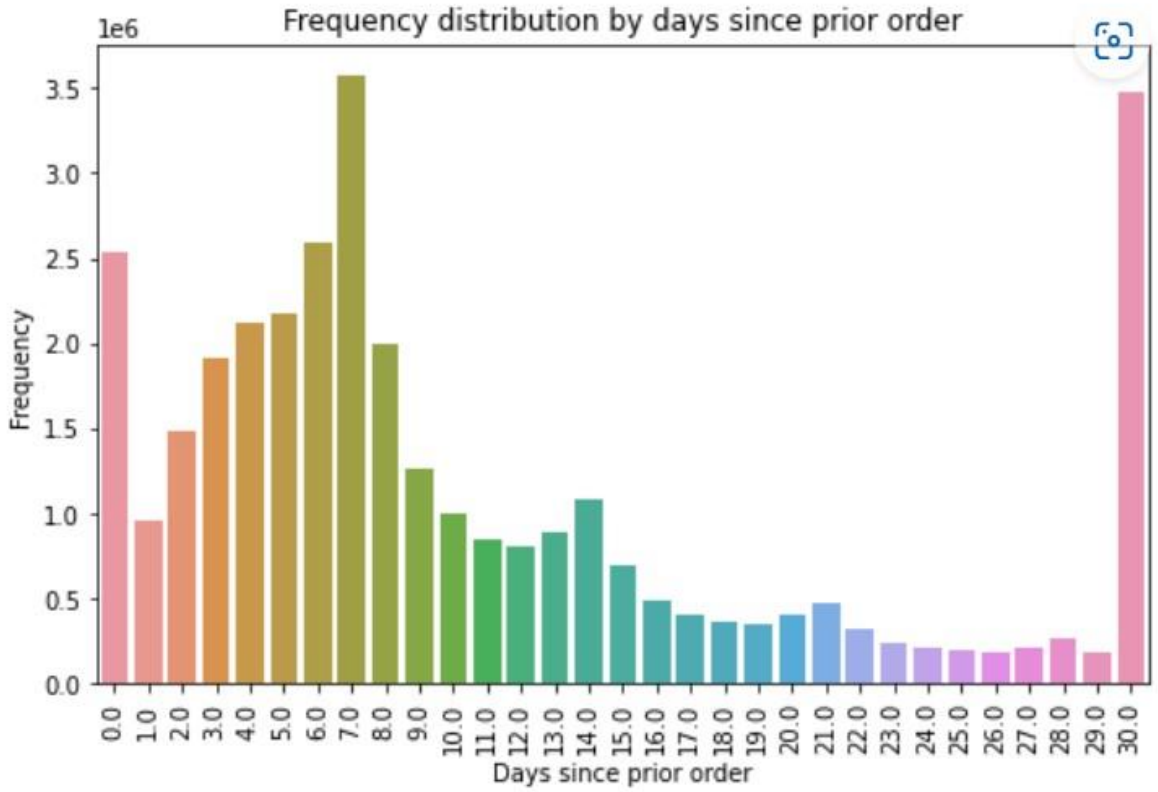
Visualizing the frequency distribution of orders by time of day (Hours), we can deduce the following :

- 10 am seems to be the most popular hour of day in terms of user activity, i.e., highest spend hour
- Highest user activity can be observed between 10 am and 4 pm, indicating higher frequency of orders made within this window
- Post 4 pm, we see a decline in user activity
- 12 am to 5 am exhibits the least user activity and consequently the least preferred time window to make purchases

3] What time of the day and day-time is the most popular for placing orders?



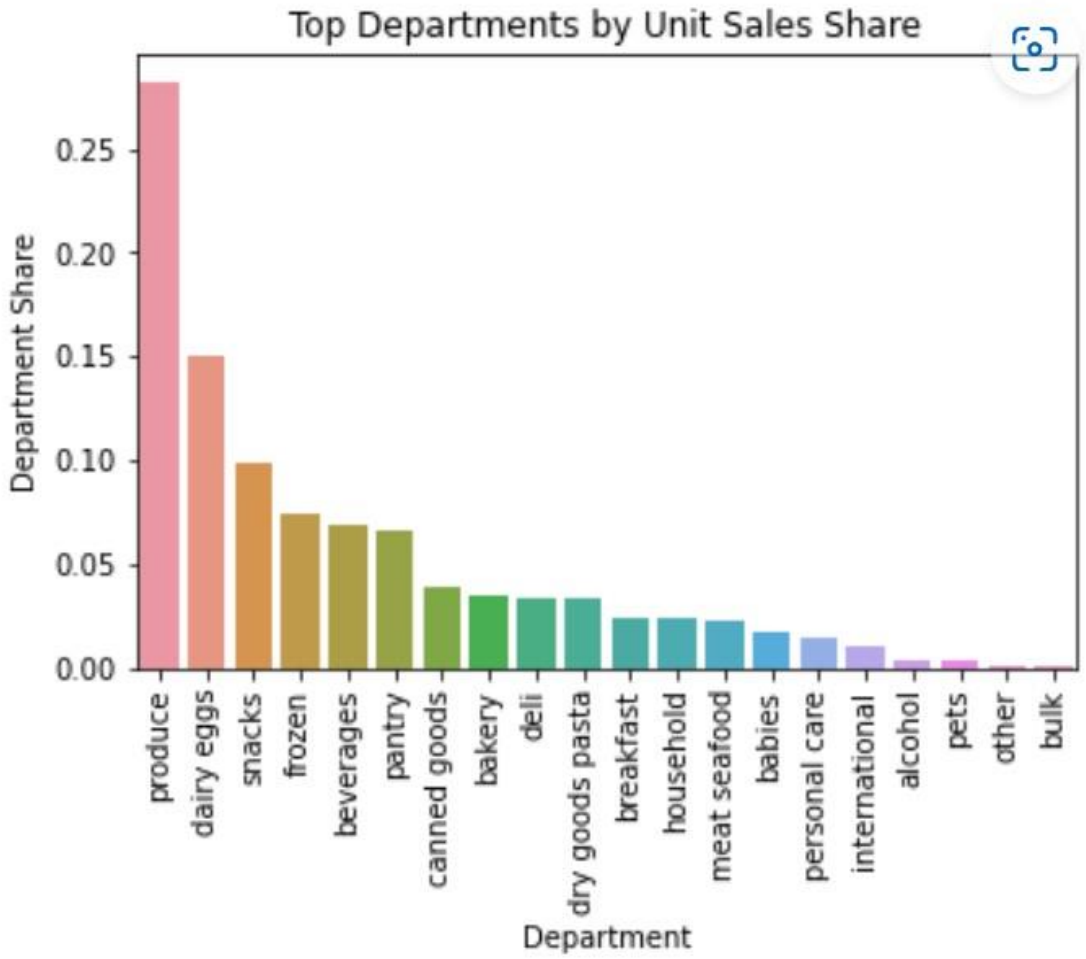
4] How many days since the previous order was placed do the customers go for reordering again?



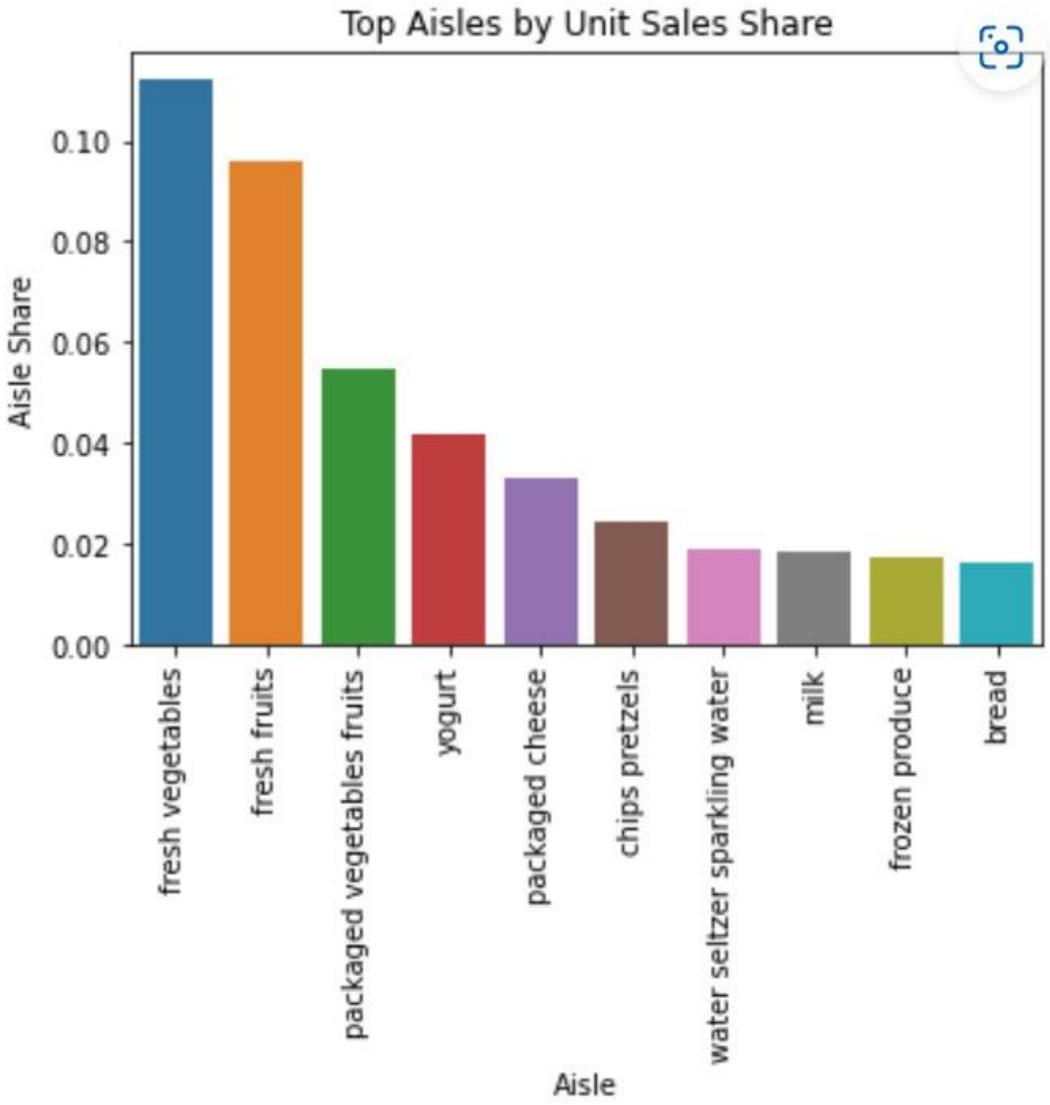
Visualizing the frequency distribution of day since prior order, we can deduce the following :

- The customers have the highest likelihood of placing their next order within a period of 7 days or 30 days
- Subsequent orders are generally placed on a weekly or monthly basis
- The spikes observed on the 7th, 14th, 21st, and 28th days signify a weekly order placement trend

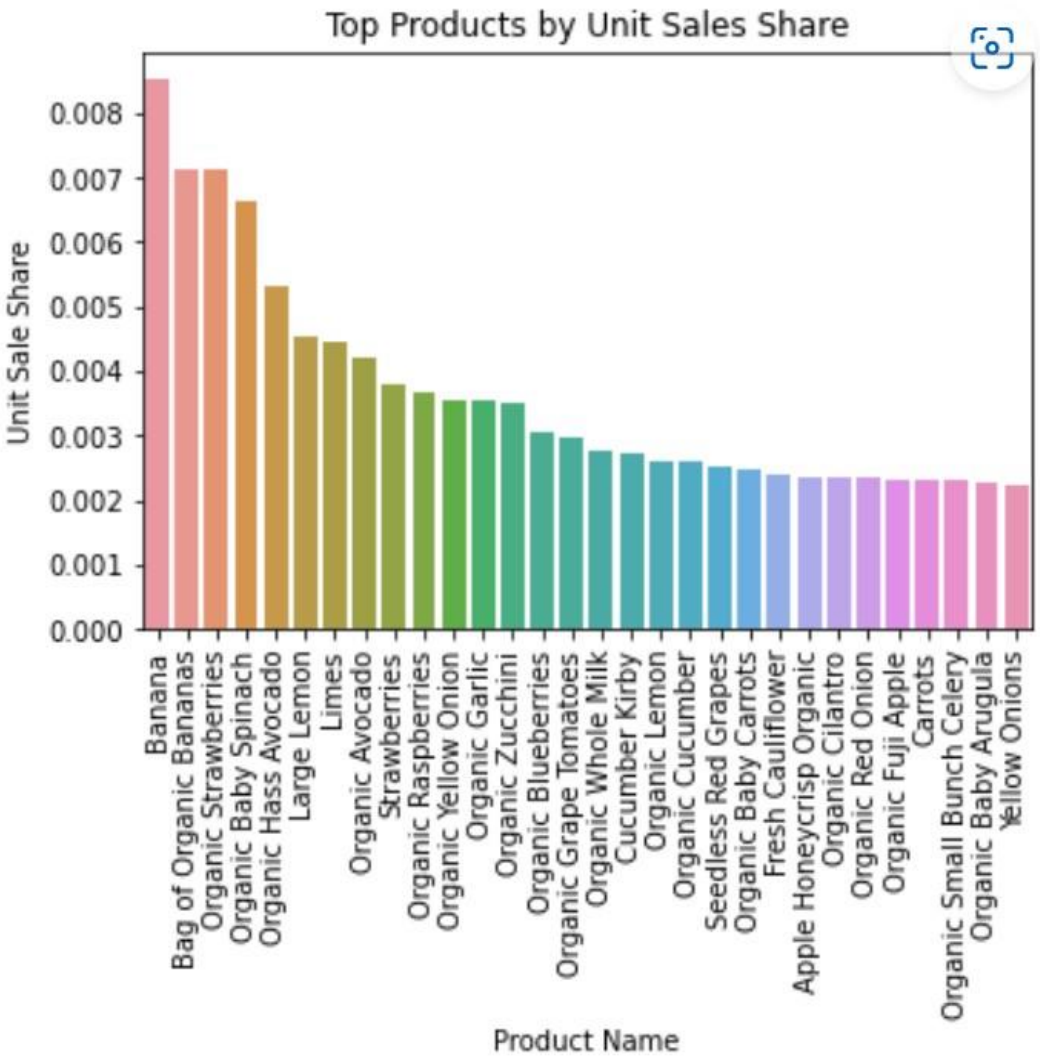
5] Which department has the most sales?



6] Which aisles contribute to the most number of sales?



7] Which products are sold the most?



MODELS

1. **To create clusters of users based on common purchase habits.**

   The Instacart dataset provides us with User Ids and their respective order IDs which correspond to a certain product. Products are also grouped as per the aisles.

   Based on this data we can create groups/clusters of userIds (users) who have similar product purchasing habits to identify the different customer segments , analyze the clusters and unearth consumer behavioral patterns and basis these patterns provide suggestions and recommendations to Instacart in order to drive sales, revenue, and customer traffic.

Approach used:- We performed dimension reduction and PCA. After these we used K-means clustering to achieve the classification. The detailed code is present in the iPython Notebook. Here are some of the insights of our model:-

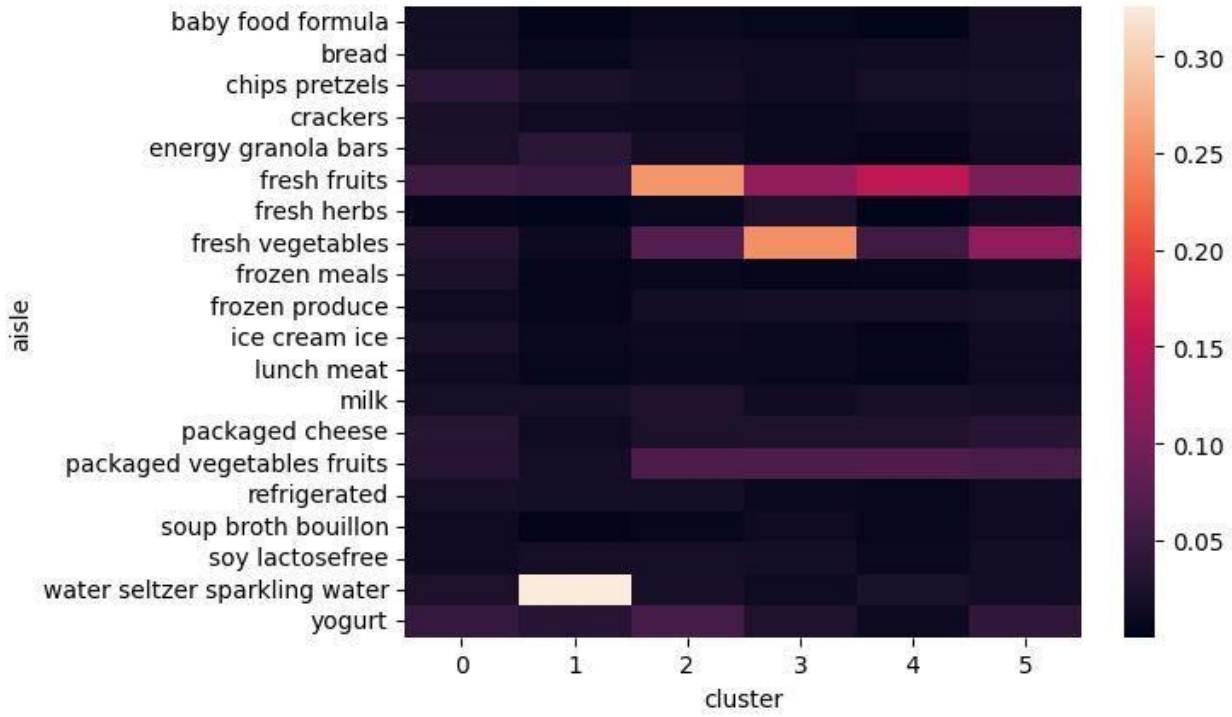
The below table shows an insight of how the userId's are belonging to which cluster

```
1 cluster_df
```

	user_id	0	1	2	3	4	5	cluster
0	1	-0.103536	0.048406	-0.031841	-0.005104	-0.052794	-0.055039	1
1	2	-0.078234	0.077940	-0.076190	-0.102570	0.080712	-0.040004	1
2	3	0.057084	0.085735	0.001869	-0.012541	-0.028985	0.113021	3
3	4	-0.055881	0.134633	-0.002237	-0.036645	-0.088069	-0.027561	2
4	5	0.133678	-0.019483	-0.021091	-0.004071	0.057342	0.139165	3
...	...	...	...	...	...	...	...	...
206204	206205	-0.017316	0.088300	-0.075846	-0.095354	0.164143	-0.023124	3
206205	206206	-0.040591	-0.036797	-0.024151	0.010472	-0.051143	0.000681	1
206206	206207	-0.022375	-0.032704	-0.009552	-0.015954	0.021188	0.020505	3
206207	206208	-0.002028	-0.010538	-0.037729	-0.008029	0.028668	0.045144	3
206208	206209	-0.072817	0.017375	-0.043608	-0.002974	-0.010667	0.027696	1

206209 rows × 8 columns

Heatmap illustrating share of purchases by aisle for the top 20 Instacart aisles


baby food formula
bread
chips pretzels
crackers
energy granola bars
fresh fruits
fresh herbs
fresh vegetables
frozen meals
frozen produce
ice cream ice
lunch meat
milk
packaged cheese
packaged vegetables fruits
refrigerated
soup broth bouillon
soy lactosefree
water seltzer sparkling water
yogurt
aisle
0
1
2
3
4
5
cluster
0.30
0.25
0.20
0.15
0.10
0.05

The model performance can be judged from the below WCSS matrix

	Clusters	WSS
0	2	5172.134349
1	3	4370.253418
2	4	3841.941541
3	5	3396.683867
4	6	3043.733954
5	7	2767.923924
6	8	2569.519547
7	9	2421.631614
8	10	2283.418701
9	11	2174.948974
10	12	2080.590775
11	13	2001.803736
12	14	1941.868104

2. **To predict if a product will be reordered or not**

Instacart provided data on the past orders of 200,000 users, which are classified as prior, train, or test orders. The goal is to predict which products will be in a user's future order, and this is a classification problem. Predictor variables (X) will be calculated based on the characteristics of the product and the user's behavior. The XGBoost algorithm is used to create a predictive model, which is applied to the test dataset to predict the "reordered" variable. The method includes importing features, creating test and train DataFrames, creating the predictive model, and applying the model.

Model specifics:

The independent/predictors we chose were as follows:

'user_id', 'product_id', 'order_dow', 'order_hour_of_day', 'days_since_prior_order'

The dependent variable we tried to predict was:- **'reordered'**

The model performance is summarized below:

Accuracy: 0.6634976497016036
Precision: 0.650421813296836
Recall: 0.9286153991143158
F1 Score: 0.7650126380809197

Conclusions derived from the model:

These metrics indicate that the model is able to correctly classify 66.35% of the instances in the test dataset. The precision of 0.6504 suggests that when the model predicts a product will be reordered, it is correct about 65.04% of the time. The recall of 0.9286 indicates that the model is able to identify 92.86% of the products that will be reordered. The F1 score of 0.7650 provides a balanced view of the model's overall performance by taking into account both precision and recall.

References

Kaggle: https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis?select=de partments.csv