DR. D. Y. PATIL SCHOOL OF SCIENCE & TECHNOLOGY
DR. D. Y. PATIL VIDYAPEETH, PUNE
(Deemed to be University)
(Accredited (3rd cycle) by NAAC with a CGPA of 3.64 on four-point scale at 'A++' Grade)
(Declared as Category - I University by UGC under Graded Autonomy Regulations, 2018)
(An ISO 9001: 2015 and 14001:2015 Certified University and Green Education Campus)

2023 -2024

A PROJECT REPORT ON

Instacart Market Basket Analysis

SUBMITTED BY

Team Members	Roll No.
Aditya Rawat	BTAI - 01
Afthab Backer	BTAI - 02
Shivani Panicker	BTAI - 35
Lalit Kumar	BTAI - 24
Vansh Lakhwani	BTAI - 57

DEPARTMENT OF

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

CERTIFICATE

This is to certify that the project report entitles

Instacart Market Basket Analysis

Submitted By

Lalit Kumar	**BTAI – 24**
Aditya Pratap Singh Rawat	**BTAI - 01**
Afthab Backer	**BTAI – 02**
Shivani Panicker	**BTAI – 35**
Vansh Lakhwani	**BTAI - 57**

is a bonafide student of this institute and the work has been carried out by him/her under the supervision of **Ms. Anagha Kulkarni** and it is approved for the partial fulfillment of the S. Y. B. Tech requirement of Dr. D. Y. Patil University.

(Prof. Anagha Kulkarni)
Subject coordinator

(**Dr. Manisha Bhende)**
Head of the department

Place: Pune
Date:

ACKNOWLEDGEMENT

We take this opportunity with great pleasure to express our deep sense of gratitude towards our guide **Prof. Anagha Kulkarni** for her valuable guidance and incessant encouragement and co-operation extended to us during this project work.

We are also thankful to **Prof. Dr. Manisha Bhende**, Head, Computer Science & Design Department, for her valuable guidance and providing all departmental facilities for this work.

ABSTRACT

Understanding customer behavior has become more difficult as a result of the rise in online grocery shopping, especially when it comes to market basket analysis. This abstract describes a study that uses an analysis of Instacart data to reveal user preferences and purchasing trends.

Retailers looking to improve overall customer satisfaction, personalized recommendations, and maximize inventory management must grasp market basket dynamics in the context of emerging e-commerce trends. Although previous research offers valuable insights into customer behavior in general, there is a dearth of studies that concentrate on particular platforms such as Instacart.

This research intends to close this gap by performing an extensive study of the data from Instacart. The study uses statistical methods and machine learning algorithms to address important topics like item affinities, seasonality impacts, and customer division according to past purchases.

Data gathering entails gaining access to Instacart's anonymized transactional data, which includes details about orders, items, users, and timestamps. Findings are guaranteed to be robust and broadly applicable when a varied dataset covering many locations and time periods is used.

A sizable sample of Instacart transactions, consisting of millions of orders and thousands of products, are analyzed as part of the experimental design. To identify the underlying consumer behavior, patterns and trends are retrieved using machine learning algorithms and rigorous statistical analysis.

Interesting consumer preferences are revealed by preliminary study, including the frequency of specific product combinations, time fluctuations in purchase behavior, and different customer groupings with different shopping patterns. Cross-validation and hypothesis testing methods are used to validate these results.

The findings of this study have ramifications for data scientists, marketers, and merchants providing practical advice on customer segmentation, advertising tactics, and product placement. Stakeholders can adapt their strategies to meet changing customer demands and improve the entire shopping experience by understanding consumer behavior in the context of online grocery shopping.

Going forward, further research is necessary to examine aspects like demographics and marketing tactics that have a greater impact on purchasing decisions. Furthermore, incorporating cutting-edge machine learning methods and real-time data streams could improve the research and offer more detailed insights into customer behavior.

Essentially, this initiative serves parties interested in improving online shopping experiences while also adding to a larger understanding of consumer behavior in the digital age. In the end, it emphasizes how important data-driven decision-making is for negotiating the complexity of contemporary consumer marketplaces.

TABLE OF CONTENTS

CHAPTER 01
INTRODUCTION

1.1 Overview

Instacart Market Basket Analysis is used extensively in retail and e-commerce to identify relationships between products that are frequently purchased together. The primary goal of this project is to uncover patterns in customer purchasing behavior, specifically looking for items that tend to be bought simultaneously.

1.2 Motivation

Conducting a Market Basket Analysis for Instacart offers a chance to understand customer purchasing behavior deeply. By identifying product associations, Instacart can offer personalized recommendations, enhancing the shopping experience and satisfaction. This analysis streamlines inventory management, logistics, and reduces costs while ensuring product availability. Leveraging insights for targeted marketing campaigns boosts engagement and revenue. Ultimately, this approach enables Instacart to continuously improve and ensure long-term success.

1.3 Problem Definition and Objectives

Problem Definition :

This analysis aims to extract insights from a retail dataset to enhance store performance and customer satisfaction. Objectives include Market Basket Analysis for optimizing store layout and marketing, Customer Segmentation for tailored marketing efforts, Seasonal Trends Analysis for inventory optimization, Customer Churn Prediction for proactive retention strategies, and identifying Product Association Rules for cross-selling opportunities. Ultimately, the goal is to drive business growth and maximize profitability in the retail sector.

Objectives :

- Analyze the anonymized data of 3 million grocery orders from more than 200,000 Instacart users open sourced by Instacart
- Find out hidden association between products for better cross-selling and upselling
- Perform customer segmentation for targeted marketing and anticipate customer behavior
- Build a Machine Learning model to predict which previously purchased product will be in user’s next order.

1.4 Project Scope & Limitations

Project Scope :

The project scope includes analyzing a retail dataset to answer several key questions :

- Identify the top 50 products most frequently purchased.
- Determine the percentage of products re-ordered.
- Explore the highest re-ordered products.
- Categorize time and day of week on which people order the most.
- Segment highest re-ordered department and aisles.
- Calculate the number of orders placed by each customer.
- Analyze the distribution of orders across the day.
- Determine the maximum number of variety in aisles and department.

Limitations :

- **Data Quality** : The analysis heavily relies on the quality and completeness of the dataset. Inaccurate or incomplete data could lead to biased conclusions and inaccurate insights.
- **Sample Bias** : The dataset may not represent the entire customer base or may be biased towards certain demographics, regions, or purchasing channels. This could limit the generalizability of findings.
- **Privacy and Ethics** : Analyzing customer data raises privacy concerns, particularly regarding sensitive information such as personally identifiable data. Adhering to ethical guidelines and data privacy regulations is crucial to safeguarding customer privacy and maintaining trust.

CHAPTER 02
LITERATURE SURVEY

Sr. No.	Title of the paper	Name of authors	Publication Details	Research Methodology / Algorithm / Tool / Techniqu	Results/ Conclusion / Product	Future Scope
1)	Market Basket Analysis for Next Basket Item Prediction Using Data Mining and Machine Learning.	Javeria Altaf, Sana Jamshaid, Maryem Ismail, Hamid Ghous	01 February 2024	French Retail Store Dataset (FRSD), Bread Basket Dataset (BBD), the model predicts future item baskets. ML classifiers including Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbours (KNN), Decision Tree (DT)	Show promising results, with RF achieving accuracies of 92.2% (FRSD) and 93.0% (BBD). This predictive model has the potential to significantly boost sales for organizations.	This model will help organization to increase their sales.
2)	Personalized Cadence Awareness for Next Basket Recommendation	Ori Katz, Oren Barkan, Noam Koenigstein	21 March 2024	Review the common characteristics of users' repurchase patterns, which characterize the NBRR problem. Building on these insights, we introduce a novel hyper convolutional model tailored to capture behavioural patterns associated with repeated purchases. To evaluate its effectiveness, we conduct experiments on three publicly available datasets, offering a comprehensive analysis.	This predictive model holds promise for substantially increasing sales for organizations and extending their reach.	This research contributes valuable insights into enhancing repurchase recommendation systems and advancing the understandin gof user purchase behavior in general
3)	A scalable and flexible basket analysis system for big transaction data in Spark	This paper proposes a scalable distributed frequent itemset mining (ScaDistFIM) algorithm for basket analysis on big transaction data to solve these two problems. ScaDistFIM is	March 2024	Spark framework FIM algorithm Distributed algorithm named ScaDistFIM business requirements	The primary benefit of this approach lies in its scalability, capable of handling transaction data containing billions of records, and its flexibility to perform a wide	Our forthcoming research will extend this approach by introducing a new FIM algorithm capable of computing approximate sets of frequent

		performed in two stages. The first stage uses the FP-Growth algorithm to compute the local frequent item sets from each random subset of the distributed transaction dataset, and all random subsets are computed in parallel. The second stage uses an approximation method to aggregate all local frequent itemsets to the final approximate set of frequent itemsets where the support values of the frequent itemsets are estimated.			range of basket analysis tasks to fulfil diverse terabyte-scale transaction datasets. Additionally, we will explore novel methods for conducting basket analysis on geographicall y distributed transaction datasets stored across multiple data centres.	itemsets from extensive random samples of
4)	Modelling Personalized Item Frequence Information for Next-basket Recommendation	This paper focuses on reproducing and extending the results of the paper:"Modeling Personalized Item Frequency Information for Next-basket Recommendation" which introduced the TIFU-KNN model and proposed to utilize Personalized Item Frequency (PIF) for Next Basket Recommendation (NBR).	27 Feb 2024	We utilized publicly available grocery shopping datasets used in the original paper and incorporated additional datasets to assess the generalizability of the findings. We evaluated the performance of the models using metrics such as Recall@K, NDCG@K, personalized-hit ratio (PHR), and Mean Reciprocal Rank (MRR).	The experimental results confirmed that the reproduced model, TIFU-KNN, outperforms the baseline model, Personal Top Frequency, on various datasets and metrics.	The findings Highlight both the potential of the predictive model to boost sales and the challenges posed by smaller basket sizes in certain datasets. Future research should focus on improving NBR Performance.

CHAPTER 03
SOFTWARE AND HARDWARE REQUIREMENTS SPECIFICATION

3.1Assumptions and Dependencies

1. **Data Quality** : Assumption of accurate and complete transactional data.
2. **Data Availability** : Dependency on having sufficient historical transaction data.
3. **Domain Knowledge** : Assumption of analysts possessing domain knowledge of the grocery retail industry.
4. **Privacy and Compliance** : Dependency on compliance with data privacy regulations.
5. **IT Infrastructure** : Dependency on robust IT infrastructure for data handling and analysis.
6. **Stakeholder Collaboration** : Assumption of collaboration with stakeholders from various departments.
7. **Actionable Insights** : Dependency on translating insights into actionable strategies and decisions.

3.2 System Requirements

3.2.1 Database Requirements

Csv files

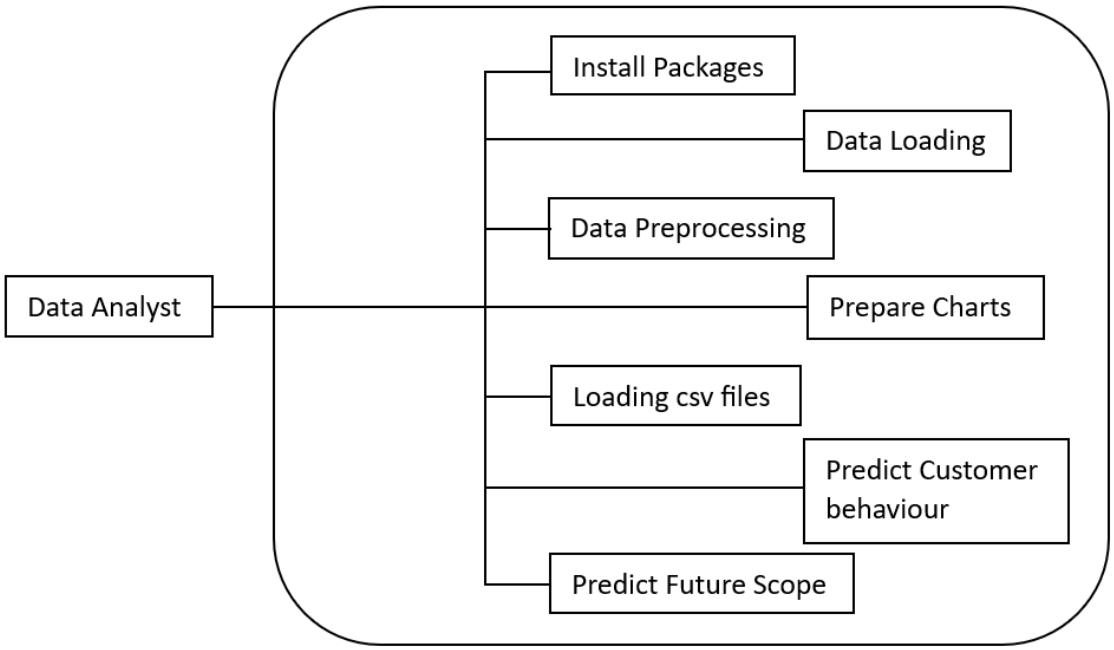
- aisles.csv
- orders.csv
- products.csv
- departments.csv
- order_products_prior.csv
- order_products_train.csv

3.2.2 Software Requirements (Platform Choice)

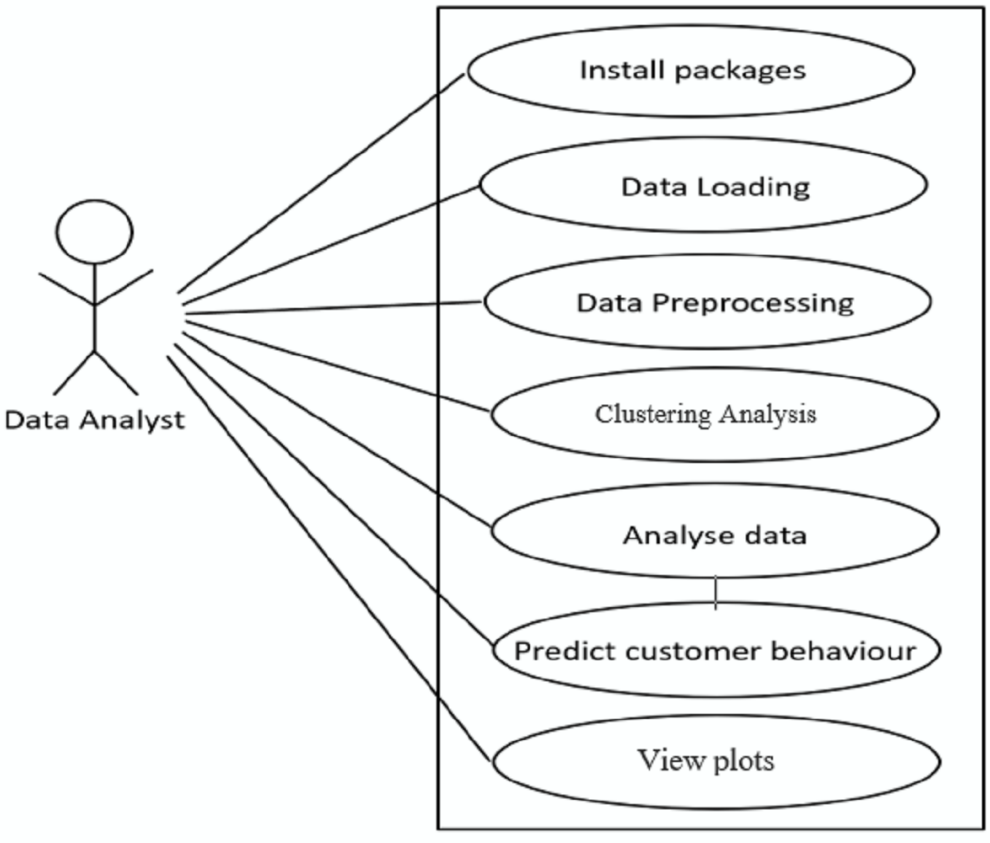
Programming Language	**Description**
R	For data analysis
Libraries	**Description**
ggplot2	to create graphics declaratively
dplyr	perform data wrangling and data analysis.
stringr	provides simplicity and consistency to use wrappers for the 'stringi' package
scales	provides the internal scaling infrastructure used by ggplot2
knitr	engine for dynamic report generation

CHAPTER 04
SYSTEM DESIGN

4.1 System Architecture



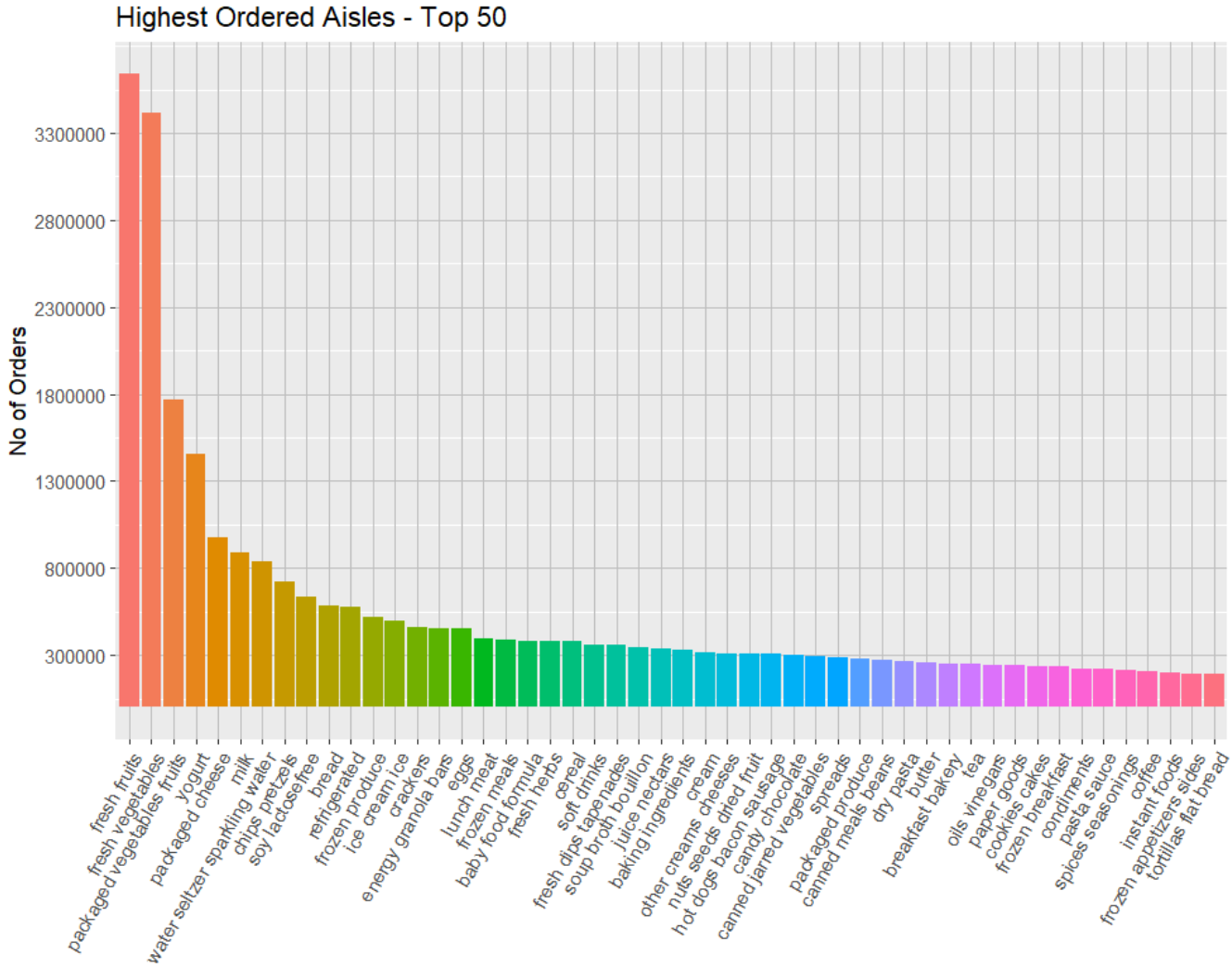
4.2 Use case Diagram

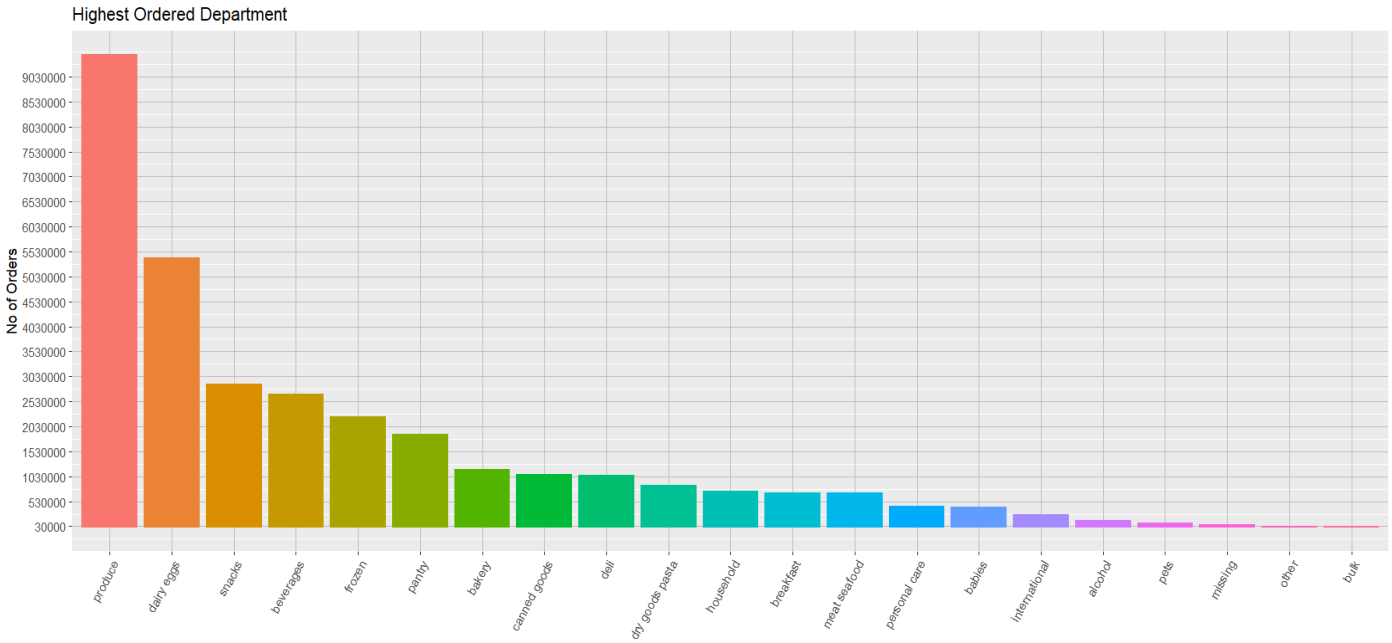


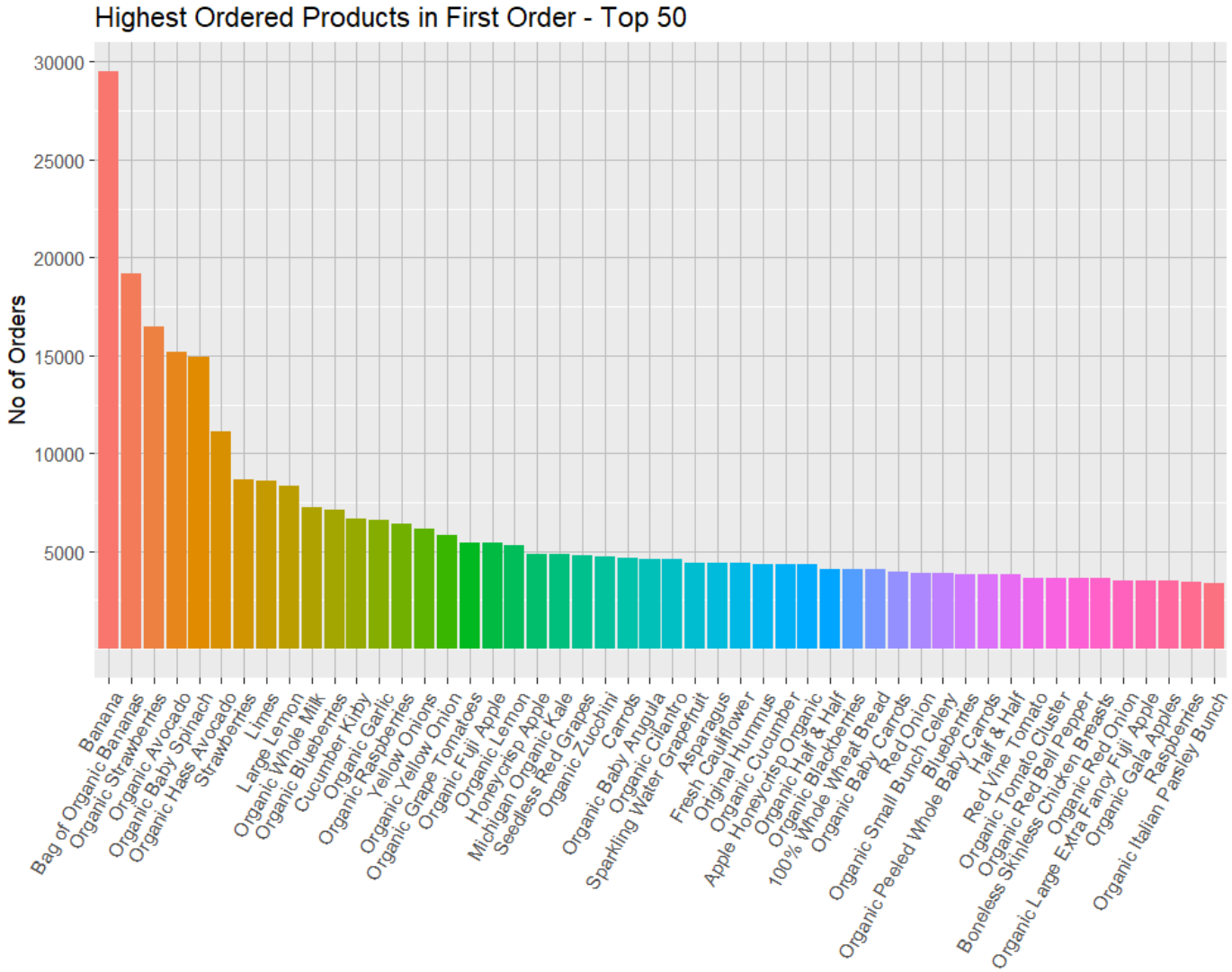
CHAPTER 05
PROJECT IMPLEMENTATION

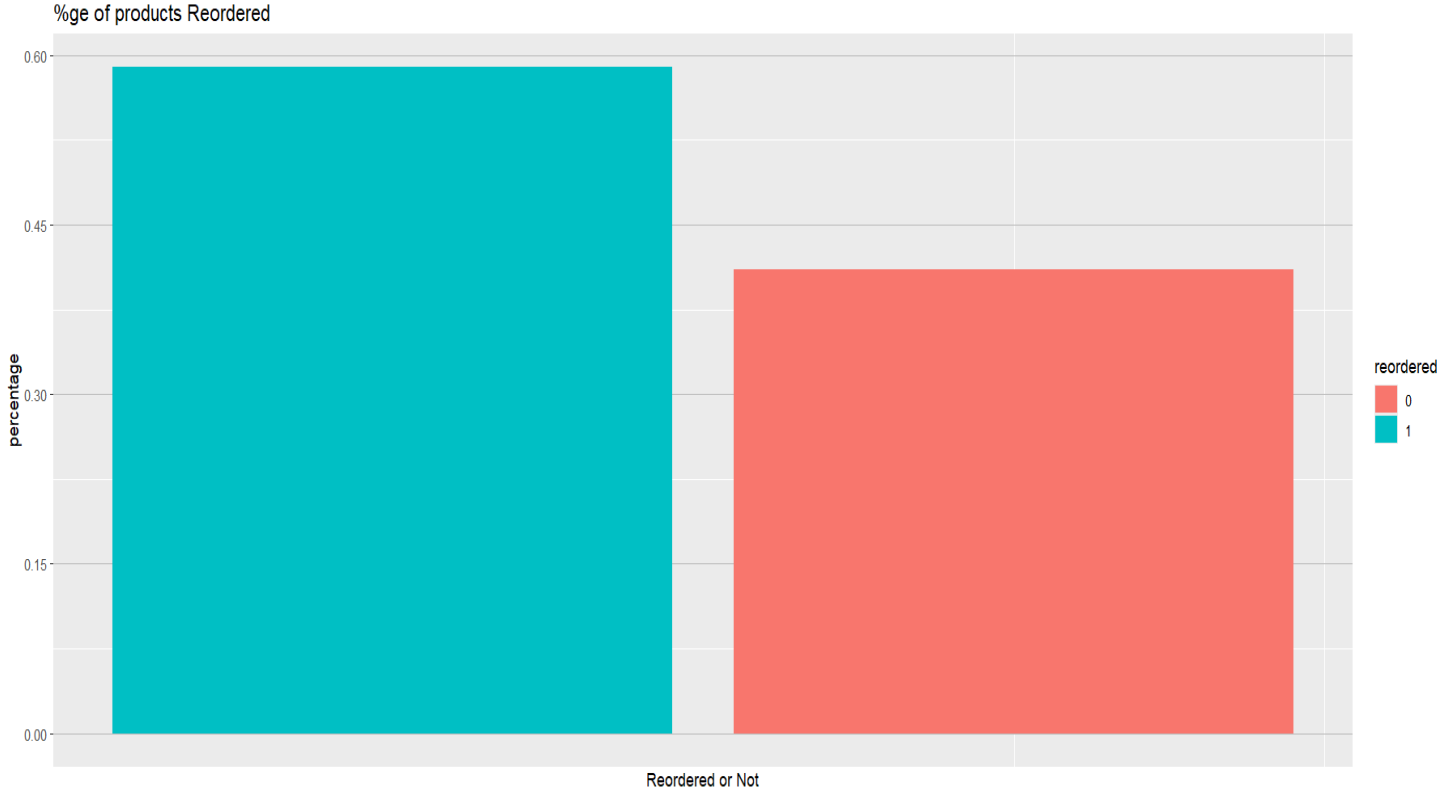
1. **Data Preparation** : Read the necessary CSV files (orders, products, departments, aisles, order_products_prior, and order_products_train) into your R environment.
2. **Data Cleaning** : Convert data types, such as hour of the day and day of the week, to the appropriate formats. Ensure consistency and accuracy in the data.
3. **Data Merging** : Merge relevant datasets, such as products with aisles and departments, to enrich the analysis with additional information.
4. **Exploratory Data Analysis (EDA)** : Conduct EDA to gain insights into the data, such as identifying top products, departments, and aisles based on order frequency.
5. **Visualization** : Create visualizations, such as bar charts and histograms, to represent key findings effectively.
6. **Market Basket Analysis** : Use association rule mining techniques to identify frequently co-purchased products and aisles.
7. **Customer Behavior Analysis** : Analyze customer behavior, such as the time of day and day of the week with the highest order frequency.
8. **Product Reordering Analysis** : Investigate product reordering patterns to understand customer preferences and trends.
9. **Department Analysis** : Explore department-level insights, such as the most ordered departments and changes in department ranking during reorder.
10. **Documentation** : Document your analysis process, findings, and insights for future reference and communication.

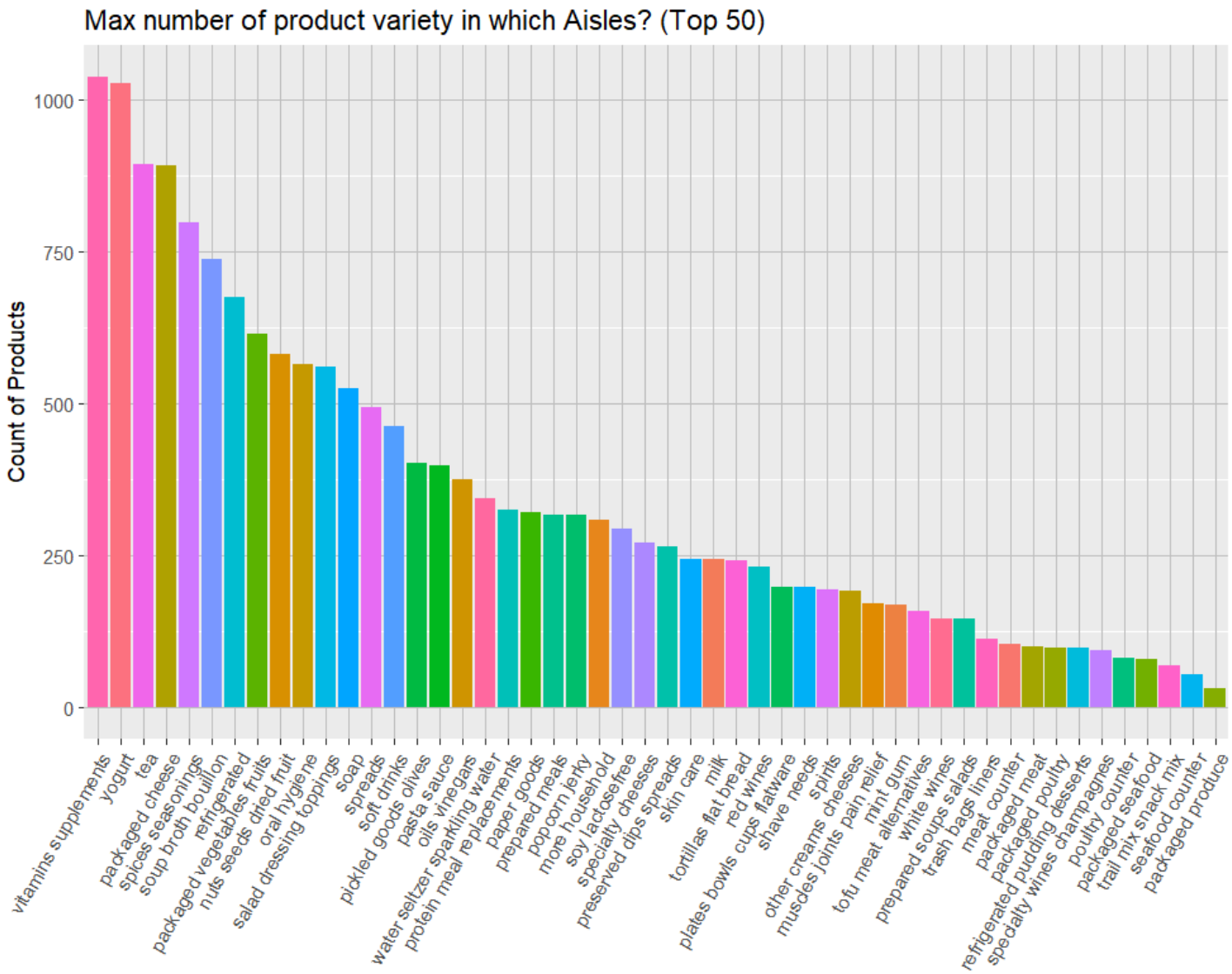
CHAPTER 06
RESULTS

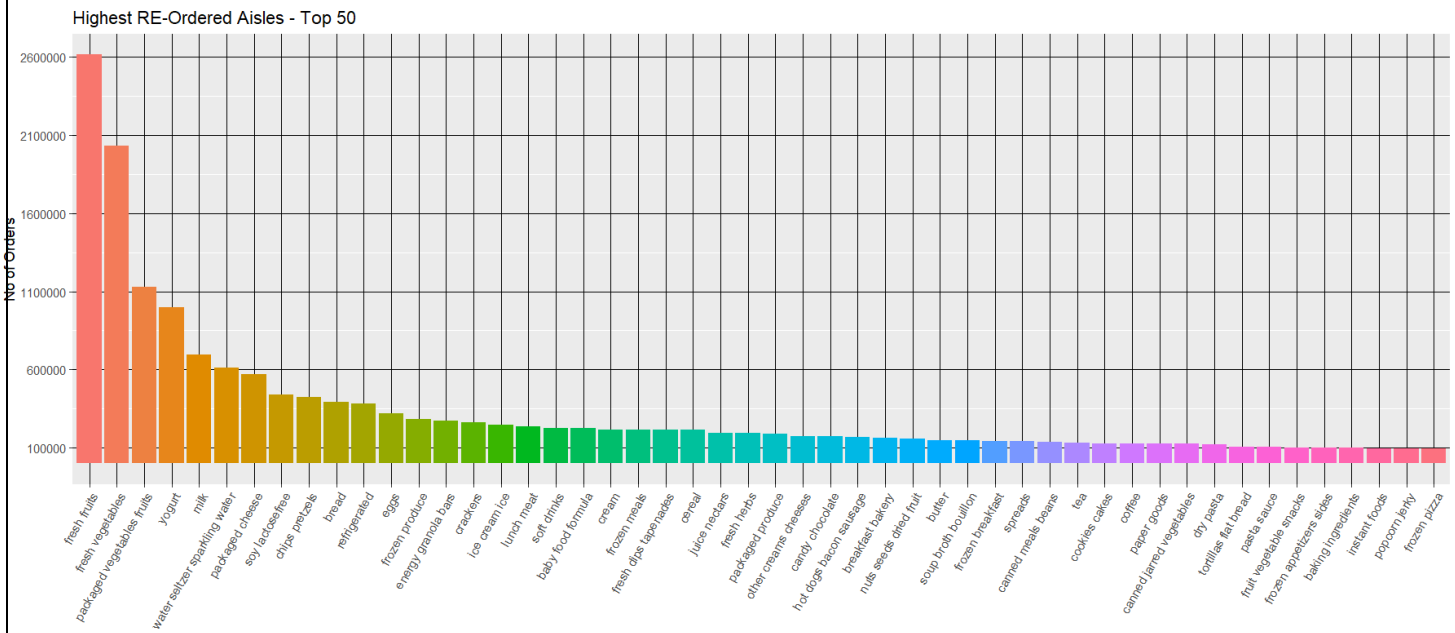


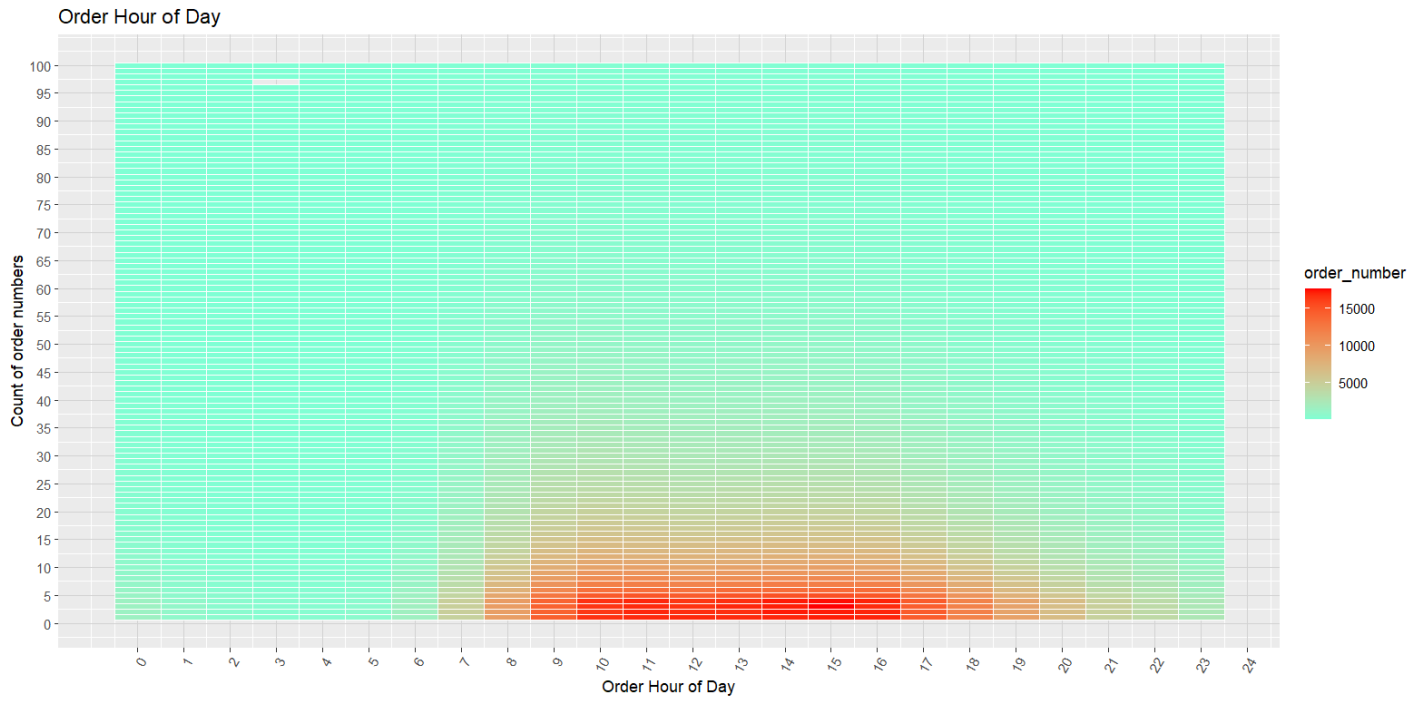


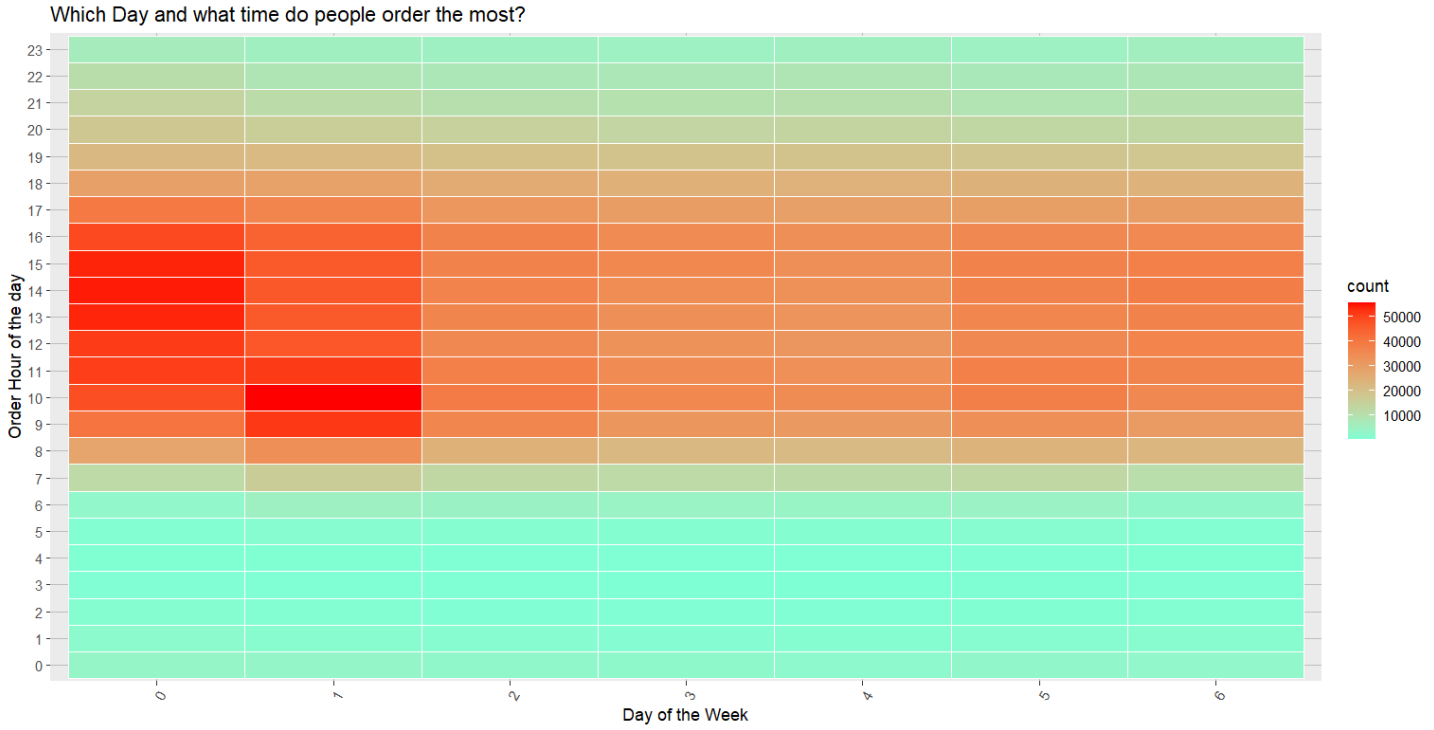

Highest Ordered Products in First Order - Top 50
No of Orders
30000
25000
20000
15000
10000
5000
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Avocado
Organic Baby Spinach
Organic Hass Avocado
Strawberries
Limes
Large Lemon
Organic Whole Milk
Organic Blueberries
Cucumber Kirby
Organic Garlic
Organic Raspberries
Yellow Onions
Organic Yellow Onion
Organic Grape Tomatoes
Organic Fuji Apple
Organic Lemon
Honeycrisp Apple
Michigan Organic Kale
Seedless Red Grapes
Organic Zucchini
Carrots
Organic Baby Arugula
Organic Cilantro
Sparkling Water Grapefruit
Asparagus
Fresh Cauliflower
Original Hummus
Organic Cucumber
Apple Honeycrisp Organic
Organic Half & Half
100% Whole Wheat Bread
Organic Blackberries
Organic Baby Carrots
Red Onion
Organic Small Bunch Celery
Blueberries
Organic Peeled Whole Baby Carrots
Half & Half
Red Vine Tomato
Organic Tomato Cluster
Organic Red Bell Pepper
Boneless Skinless Chicken Breasts
Organic Large Extra Fancy Fuji Apple
Organic Red Onion
Organic Gala Apples
Raspberries
Organic Italian Parsley Bunch

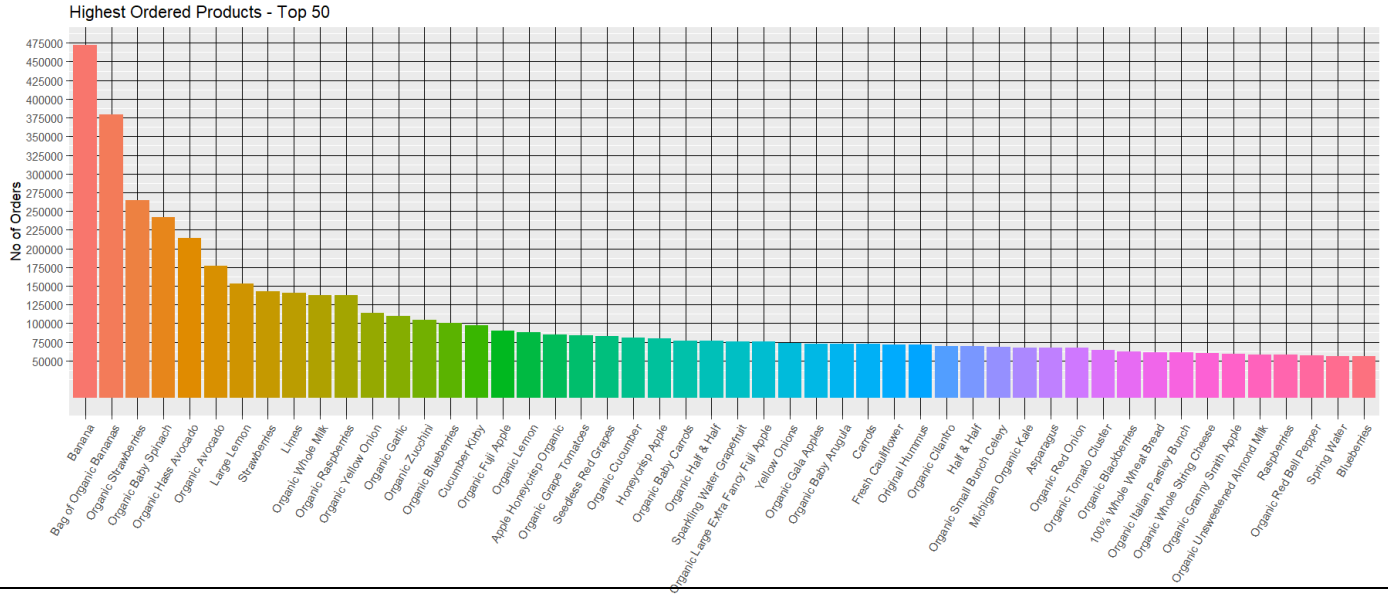

%ge of products Reordered
percentage
0.60
0.45
0.30
0.15
0.00
Reordered or Not
reordered
0
1

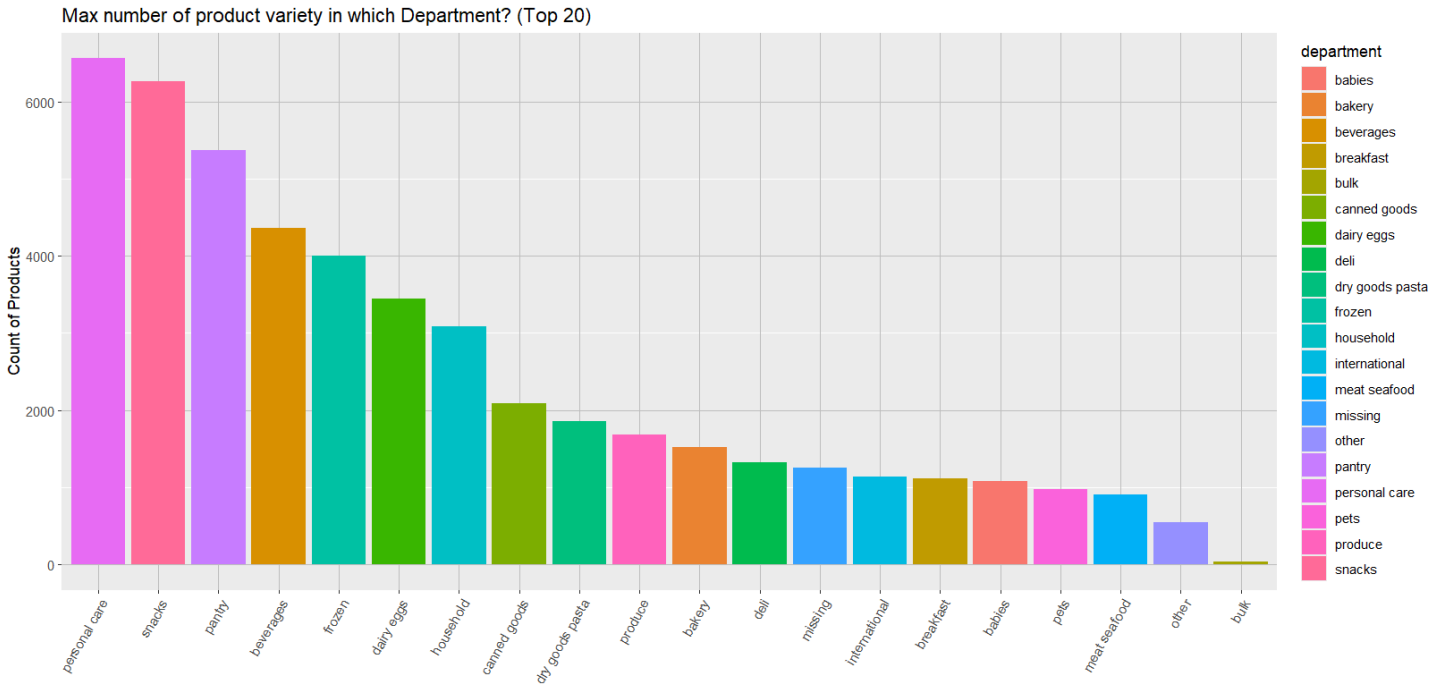

Max number of product variety in which Aisles? (Top 50)
Count of Products
1000
750
500
250
0
vitamins supplements
yogurt
tea
packaged cheese
spices seasonings
soup broth bouillon
refrigerated
packaged vegetables fruits
nuts seeds dried fruit
oral hygiene
salad dressing toppings
soap
spreads
soft drinks
pickled goods olives
pasta sauce
oils vinegars
water seltzer sparkling water
protein meal replacements
paper goods
prepared meals
popcorn jerky
more household
soy lactosefree
specialty cheeses
preserved dips spreads
skin care
milk
tortillas flat bread
red wines
plates bowls cups flatware
shave needs
spirits
other creams cheeses
musdes joints pain relief
tofu meat alternatives
mint gum
white wines
prepared soups salads
trash bags liners
meat counter
packaged meat
packaged poultry
refrigerated pudding desserts
specialty wines champagnes
poultry counter
packaged seafood
trail mix snack mix
seafood counter
packaged produce

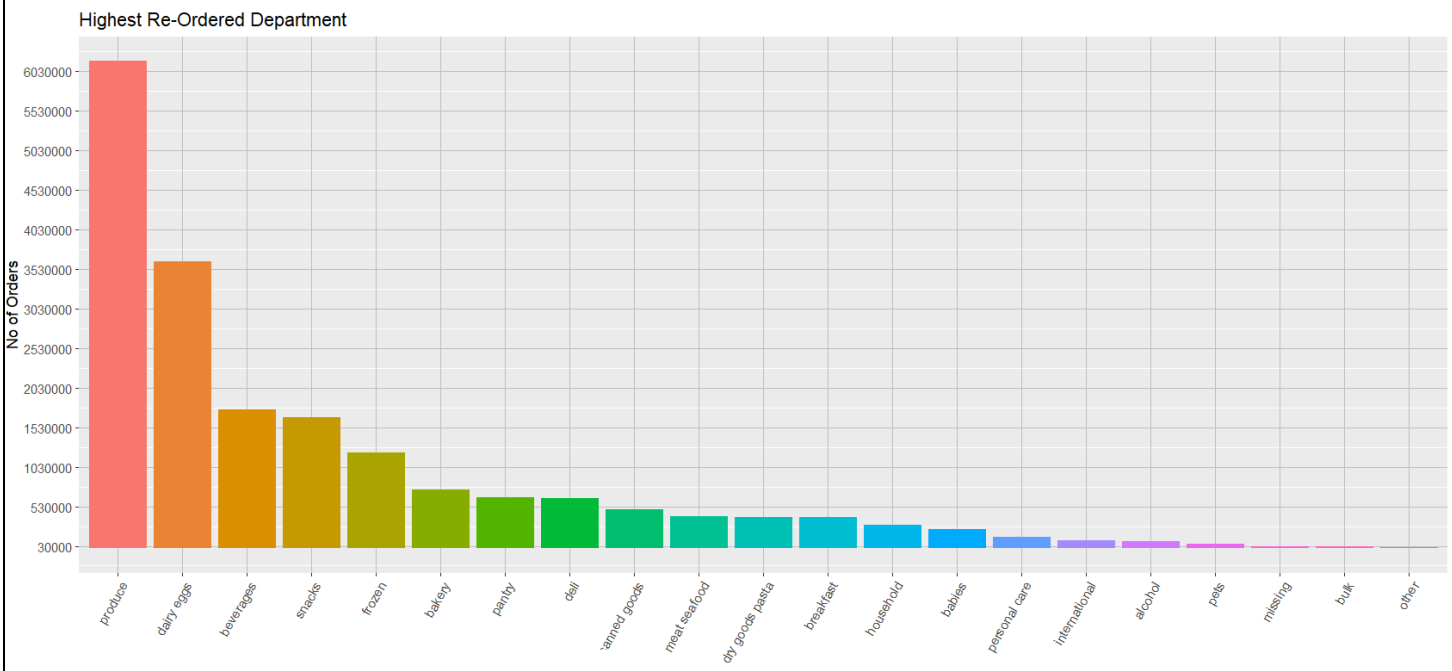

Highest RE-Ordered Aisles - Top 50
No of Orders
2600000
2100000
1600000
1100000
600000
100000
fresh fruits
fresh vegetables
packaged vegetables fruits
yogurt
milk
water seltzer sparkling water
packaged cheese
soy lactosefree
chips pretzels
bread
refrigerated
eggs
frozen produce
energy granola bars
crackers
ice cream ice
lunch meat
soft drinks
baby food formula
cream
frozen meals
fresh dips tapenades
cereal
juice nectars
fresh herbs
packaged produce
other creams cheeses
candy chocolate
hot dogs bacon sausage
breakfast bakery
nuts seeds dried fruit
butter
soup broth bouillon
frozen breakfast
spreads
canned meals beans
tea
cookies cakes
coffee
paper goods
canned jarred vegetables
dry pasta
tortillas flat bread
pasta sauce
fruit vegetable snacks
frozen appetizers sides
baking ingredients
instant foods
popcorn jerky
frozen pizza


Order Hour of Day
Count of order numbers
Order Hour of Day
order_number
15000
10000
5000


Which Day and what time do people order the most?
Order Hour of the day
Day of the Week
count
50000
40000
30000
20000
10000

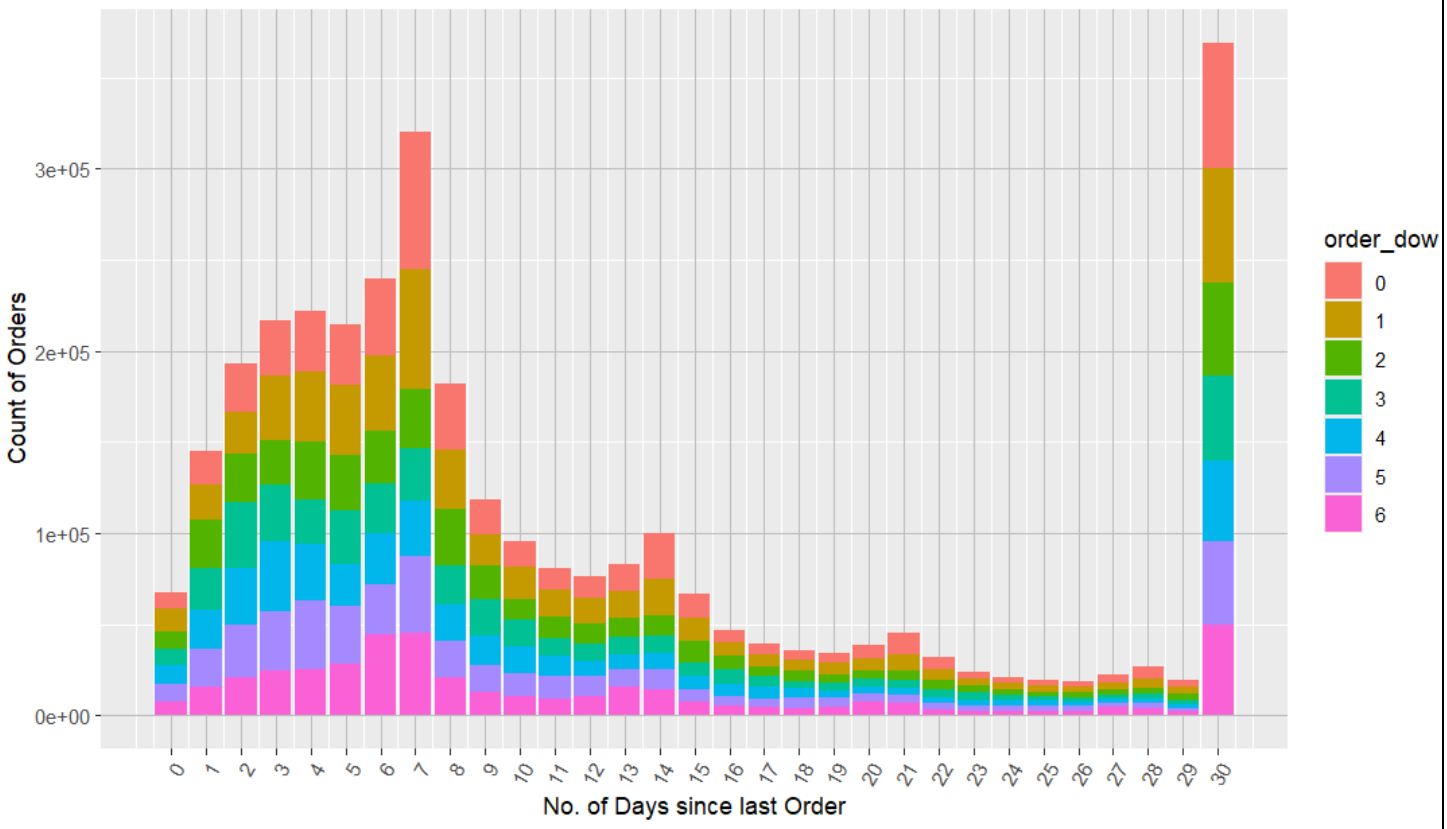

Highest Ordered Products - Top 50
No of Orders
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Organic Hass Avocado
Organic Avocado
Large Lemon
Strawberries
Limes
Organic Whole Milk
Organic Raspberries
Organic Yellow Onion
Organic Garlic
Organic Zucchini
Organic Blueberries
Cucumber Kirby
Organic Fuji Apple
Organic Lemon
Apple Honeycrisp Organic
Organic Grape Tomatoes
Seedless Red Grapes
Organic Cucumber
Honeycrisp Apple
Organic Baby Carrots
Organic Half & Half
Sparkling Water Grapefruit
Organic Large Extra Fancy Fuji Apple
Yellow Onions
Organic Gala Apples
Organic Baby Arugula
Carrots
Fresh Cauliflower
Original Hummus
Organic Cilantro
Half & Half
Organic Small Bunch Celery
Michigan Organic Kale
Asparagus
Organic Red Onion
Organic Tomato Cluster
Organic Blackberries
100% Whole Wheat Bread
Organic Italian Parsley Bunch
Organic Whole String Cheese
Organic Granny Smith Apple
Organic Unsweetened Almond Milk
Raspberries
Organic Red Bell Pepper
Spring Water
Blueberries





user_id	order_number	cnt
1	1	5
2	1	13
3	1	10
4	1	4
5	1	11
6	1	4
7	1	12
8	1	21
9	1	30
10	1	5
11	1	13
12	1	3
13	1	5
14	1	5
15	1	4

user_id	order_number	cnt
Min. : 1	Min. :1	Min. : 1.00
1st Qu.: 51553	1st Qu.:1	1st Qu.: 5.00
Median :103105	Median :1	Median : 8.00
Mean :103105	Mean :1	Mean :10.08
3rd Qu.:154657	3rd Qu.:1	3rd Qu.:14.00
Max. :206209	Max. :1	Max. :98.00



	Var1	Freq.x	Freq.y
1	produce	9479291	6160710
2	dairy eggs	5414016	3627221
3	snacks	2887550	1657973
4	beverages	2690129	1757892
5	frozen	2236432	1211890
6	pantry	1875577	650301
7	bakery	1176787	739188
8	canned goods	1068058	488535
9	deli	1051249	638864
10	dry goods pasta	866627	399581
11	household	738666	297075
12	breakfast	709569	398013
13	meat seafood	708931	402442
14	personal care	447123	143584
15	babies	423802	245369
16	international	269253	99416
17	alcohol	153696	87595
18	pets	97724	58760
19	missing	69145	27371
20	other	36291	14806
21	bulk	34573	19950

CHAPTER 07
CONCLUSION

7.1 Conclusion :

In conclusion, the analysis of customer behavior and purchasing patterns offers invaluable insights for refining business strategies. Through an understanding of product associations, customer segmentation, and seasonal trends, businesses can elevate marketing endeavors, refine product placement strategies, and craft promotions tailored to distinct customer segments. Furthermore, insights into customer churn rates and recent activity pinpoint opportunities for bolstering customer retention and engagement efforts. Ultimately, leveraging these findings can propel business growth, enhance customer satisfaction, and establish a stronger, more competitive presence in the market.

7.2 Future Work :

- **Promotion of Related Products** : Deploy strategies to strategically position complementary items together both online and in-store, leveraging insights from frequently bought product pairs to enhance the shopping experience and drive additional purchases.
- **Tailored Marketing for Customer Segments** : Customize marketing efforts based on customer segmentation derived from spending habits and purchase frequency. Implement loyalty programs or exclusive deals for high-spenders, while offering targeted promotions to occasional buyers to increase engagement.
- **Seasonal Marketing Strategies** : Capitalize on seasonal trends by launching targeted marketing campaigns during peak months. Offer special promotions and discounts during holidays or seasonal events to attract customers and boost sales.
- **Preventing Customer Attrition** : Implement personalized outreach strategies to prevent customer attrition. Identify customers who haven't made a purchase in the last 30 days and engage them with targeted offers or reminders to re-engage with your products and services.
- **Offering Product Bundles and Deals** : Create bundled offers for frequently purchased product combinations based on insights from product association analysis. By providing customers with attractive bundle options, you can increase the overall value of orders and drive sales.

REFERENCES

[1] Gurudath, Shruthi. (2020). Market Basket Analysis & Recommendation System Using Association Rules.

[2] M. Dhanabhakyam and M. Punithavalli, “A survey on data mining algorithm for market basket analysis,” Global Journal of Computer Science and Technology, 2011.

[3] The Instacart Online Grocery Shopping Dataset 2017, Accessed from

https://www.instacart.com/datasets/grocery-shopping-2017

[4] https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_612