PROJECT REPORT ON

A Distributorship Analysis Of FMCG (Unibic) Products In Vadodara

A REPORT SUBMITTED TO

DEPARTMENT OF STATISTICS
FACULTY OF SCIENCE
THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA

GUIDE:
Mr. Shrey Pandya

PROJECT MEMBERS
Shah Mona Deepak
Meenakshi Porwal
Srushti Sanjay Sali
Mohammad Saeed
Ravikumar Rathva
Sujit Raval

ACKNOWLEDGEMENT

First and foremost, we would like to thank our guide, Mr. Shrey Pandya who always helped us, welcomed our questions, keep us motivated and gave us a lot of suggestions. We would not have reached this phase, if it were not for his permanent support and guidance.

We also would like to extend our special thanks to Head Dr. Vipul Kalamkar for their support, guidance and helpful feedback.

Our sincere thanks to our parents for guiding us decently and supporting at every stage in our life also their wishes for successful completion of this project.

Last but not the least, we would like to express our eternal gratitude to friends for their support, appreciation and patience. We would like to dedicate this report to them all.

DEPARTMENT OF STATISTICS
FACULTY OF SCIENCE
THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA

CERTIFICATE

This is to certify that Shah Mona Deepak, Meenakshi Porwal, Srushti Sanjay Sali, Mohammad Saeed, Ravikumar Rathva, and Sujit Raval have satisfactorily completed the project entitled:

" A Distributorship Analysis Of FMCG (Unibic) Products In Vadodara "

As a team in the academic year 2021-22, this work is submitted to the Department of Statistics as a fulfilment for degree of Master of Science in Statistics.

Throughout the semester they carried out work with sincerity & has presented on the time and with enthusiasm.

Mr. Shrey Pandya
(Guide)

Dr. Vipul Kalamkar
(Head, Department of Statistics)

Date: 13th May 2022

INDEX

ABSTRACT

This paper concentrates on retailers demographics and purchasing behavior of Cookies(Biscuits) category of FMCG(Unibic brand) in Vadodara city of Gujarat. This study is specifically conducted for a Vadodara based distributor-Prem Agencies, and their information for retailers.

Fast moving consumer goods are generally lower profit margin products and therefore it is selling in large quantities in the regional market. Thus, it is very important for a distributor to understand how to improve brand value for the customers (retailers) as many brands are available for the same categories of products in the market.

The fast-moving consumer goods (FMCG) sector is an indispensable contributor to India"s
GDP . It constitutes a large part of consumers" budget in all countries. FMCG Industry is featured by an established distribution of network, an intense competition between the organized and unorganized segments, lower penetration levels, lower operating cost and also a lower per capita consumption.

Some of the top FMCG companies in India are - Dabur (60%), Colgate (54.7%), Hindustan Unilever (54%).

India"s FMCG sector creates an employment opportunity. It is currently growing at double-digit rate and is expected to maintain a high growth rate. Indian buyers were conservative partly due to low disposable income and few competitive products. Inflation in food products could restrict the demand among consumers and pricing flexibility for FMCG while lowering purchasing power of the consumer that diverts purchases.

INTRODUCTION

This study focuses on retail market of Unibic products in Vadodara based on a Distributorship company Prem Agencies. The report examines sales distrbution of Unibic cookies among the retail market in Vadodara , retailers demographics and their purchasing behavior since July 2019 till Feb 2022. This study brings insights useful for a distributor to understand its business potential among retailers, specifically for Unibic cookies and their categories.

DETAILS ABOUT CLIENT

CLIENT: PREM AGENCIES
(UDYAM-GJ-24-0009250)

Our client sells and supplies Food products to the retailers in Vadodara.Along with being a food supplier it is also a member of Vadodara Distributor Association(VDA). Our Client has all Good and Services Tax (GST) Audited reports (GSTIN 24AHDPR7310N1Z5).

INFORMATION ABOUT BRAND (UNIBIC PRODUCTS)

UNIBIC brand was structured in India by former Britannia COO Nikhil Sen, Unibic Australia's Michael Quinn and entrepreneur Dhruv Deepak Saxena in the year of 2004 in Bengaluru. It has regional offices at Delhi NCR, Mumbai and Kolkata. The company is growing at 45% of Compound Annual Growth Rate (CAGR). There are more than 800 strong teams spread across India. It is found that they have 30 different variants of premium cookies with new innovative products launching regularly.

SUPPLY CHAIN

Supply chain management is the practice of coordinating the various activities necessary to produce and deliver goods and services to a business's customers. Examples of supply chain activities can include designing, farming, manufacturing, packaging, or transporting.

Producers:

Producers are the organizations that make products or services available for their customers. Producers of raw materials and producers of finished goods. Producers buy goods and services and transform them into a sellable product, which they sell to their customers for the purpose of making a profit.

Examples of producers are farmers, manufacturers and construction companies.

Distributors:

A distributor is an intermediary entity between the producer of a product and another entity in the distribution channel or supply chain, such as a retailer, a value-added reseller (VAR) or a system integrator (SI). They sell their products in larger quantities.

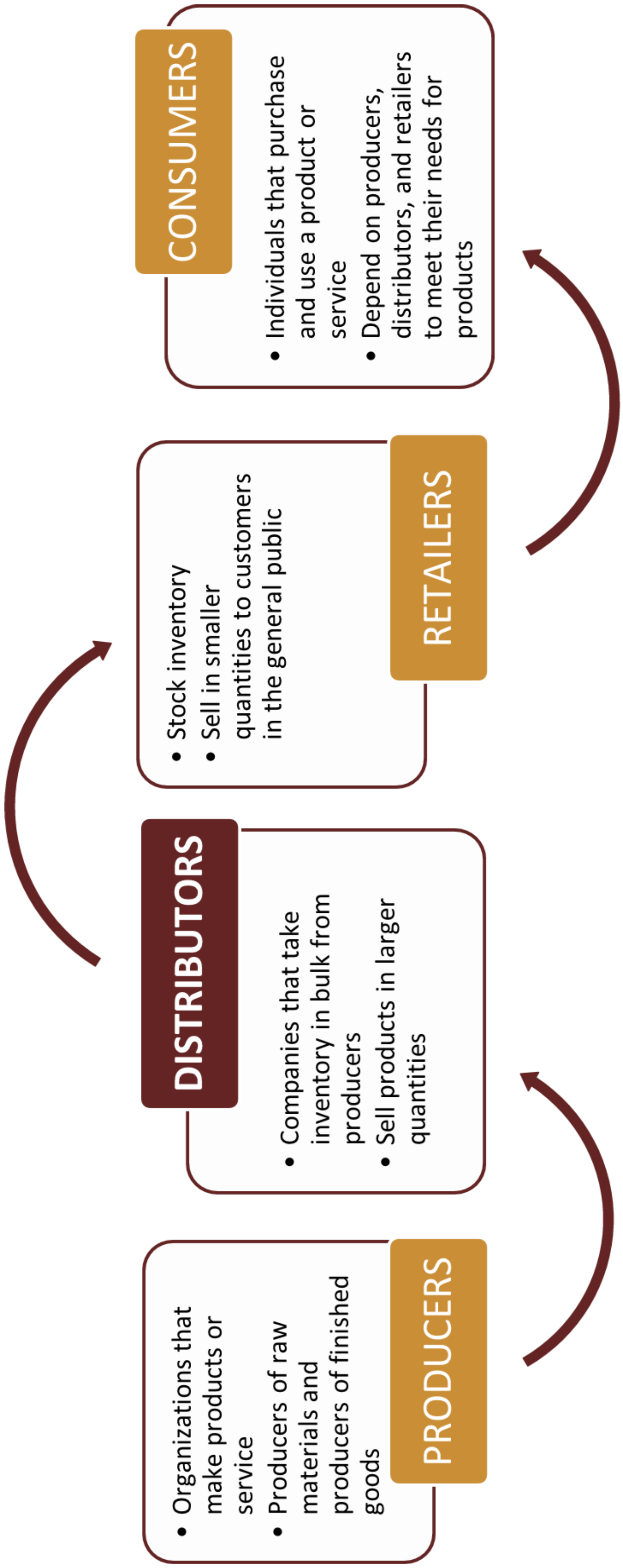
Retailers:

A retailer performs the dual functions of buying and assembling of goods. The responsibility of a retailer is to identify the most economical source for obtaining the goods from the suppliers and passing on the advantages to the consumer. The retailers perform the functions of warehousing and storing.

Consumers:

Consumers are the individuals that purchase and use a product or a service. These Depend on the producers, distributors, and retailers to meet their needs for products.

From the beginning of an order until order delivery, customers are involved in the process. The customer not only pays for the product or service, but they also decide whether or not to do business with your company again based on their experience.


PRODUCERS
Organizations that make products or service
Producers of raw materials and producers of finished goods
DISTRIBUTORS
Companies that take inventory in bulk from producers
Sell products in larger quantities
RETAILERS
Stock inventory
Sell in smaller quantities to customers in the general public
CONSUMERS
Individuals that purchase and use a product or service
Depend on producers, distributors, and retailers to meet their needs for products

WHO IS A DISTRIBUTOR?

Distributors are wholesale agents who connect manufacturers and retailers. They eliminate the need for manufacturers to contact a large number of retailers one by one. Their main concerns are:

To provide fast delivery

To Assist buying decision

To Buy and hold stocks

To Forecast market needs

An example of a distributor is the part in a gas lawnmower that controls the flow of electrical currents to spark plugs. One that markets or sells merchandise, especially a wholesaler.

A distributor is any one individual, or group working under one person. They act as a middleman between the manufacturer and retailers. They act as a link between the main organization and the consumer.

Amazon sells its products via its web application and mobile apps to its customers according to their needs. Customers can be either retailers or wholesalers. Amazon's purpose is to earn profit. So, Amazon is an online retailer, it's not a wholesaler.

In the last decade increasing attention has been focused upon the newer and creative areas of marketing such as consumer behavior, new product development and the communications interchanges between producers and consumers. More recently, the subject of distribution has been receiving a wider conceptual treatment as “marketing logistics” than had been afforded to it earlier by operational researchers searching for optimum rather than acceptable solutions. In the case of consumer goods these customers are the channel intermediaries such as wholesalers, cash and carry operators, independent and multiple retailers as well as the voluntary chains.

It has been inferred from the observations that non -leading FMCG distributors do offer relatively higher margins to retailers so that retailers accept their products and promote them to end customer as well.

OBJECTIVES

1) **To utilize the past years' data to identify trend and envision the sales for the upcoming year.**

We have data for three financial years, however, the data for the last financial is not complete (two months). We utilize the past data to envision an increase in sales for the same.
The study also includes a quarter-wise comparison of sales for these financial years.

2) **To classify SKUs of Unibic products into different categories and analyze them.**

Identifying the Top performing Product, Categories along with products in particular categories.

3) **To compare supplies of Unibic products in different stores diversified over several areas in Vadodara.**
Identify to Top performing Areas based on the purchasing behavior of customers in those particular areas

4) **To compare the sales of two years and identify the factors contributing to the difference.**

Observing how secondary variables affect the sales of the distributor, and studying their influence.

5) **To segment retailers based on their purchasing behavior.**

To identify Loyal Customers, At risk of leaving customers, and Lost customers based on after analysing their purchasing behaviour.

DATA PROCESSING METHODOLOGY

1) Raw Data

   Sales data
   SKU Mapping
   Stores Mapping

2) Secondary Data

   Temperature
   Rainfall
   Festivals

3) Data Cleaning

   Dealing with Missing Values
   Eliminating unrequired data

4) Transformed Data

   Merging raw and secondary data
   Transforming data into different data files

RAW DATA AND SECONDARY DATA:

Details about the data:

- Type of data : Secondary data
- Source : Prem Agencies

Primary Data Variables:

1. Year-Month:
   Time frame of three financial years those are Sales_19-20, Sales_20-21 and Sales_21-22 that is from Jul 19 to Jan 22. We have four number of observations and levels as 2019/ 2020/ 2021/ 2022.

2. Stores:
   There were 481 unique stores.

3. Channel Types:
   Two distinct channel types those are General and Institutional Trade.

4. Area:
   45 different Areas under Vadodara i.e. Varasiya, OP Road, Karelibaug, Manjalpur, Akota, Alkapuri, Maneja, Sayajigunj, Nizampura, Gorwa, Vasna, Gotri, Makarpura, Sama, Sama, Savli Road, Waghodiya, Tarsali, Subhanpura, Vadsar, Vasna Bhayli, ELLORA PARK, Harni, Jetalpur Road, Chhani, Fatehgunj, Pratapnagar,Atladra, Bill, Bhayli, City, Ajwa Road, Raopura, Panigate, Tandalja, Vemali, Sunpharma, Salatwada, Navapura, VIP Road, Chokhandi, Khodiyar Nagar, Karjan, Hathikhana,, Bapod, RC Dutt Road.

5. Route:
   Two different routes to reach the retailer in Vadodara i.e. Old city and New city routes.
6. Category:
   Six different Categories of Different SKUs i.e. Combo/ Display box/ Family Pack/ Super Saver/ Sugar free/ Mass.

7. Quantity:
   Quantity of each SKU purchased by particular store.

8. MRP:
   MRP of each SKU is provided.

Derived Variable:

Sales: Using MRP and Quantity of SKUs we derived the variable sales.

Secondary Variablees:

1. Festivals: Taking dummy values for the months having big festivals or vacation for long duration.

2. Temperature: We considered average temperature of different months.

3. Rainfall: We considered average rainfall in the months of distinct years.

DATA CLEANING

Data cleansing ensures you only have the most recent files and important documents, so when you need to, you can find them with ease.

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

We have performed below mentioned tasks for cleaning our data:

- Replacing Negative quantity values by zero.
- Removing Data points corresponding to SKUs belonging to the Bar category.
- Removing Data points corresponding to Stores located in areas outside Vadodara.

DATA TRANSFORMATION:

Data transformation is the process of converting data from one format or structure into another.

Pre-processing:

1. Three files were supplied by Prem Agencies:

   i. Sales data:
   Month-wise quantities of different UNIBIC products purchased by different stores for three years in respective files :- Sales_19-20, Sales_20-21 and Sales_21-22.

PREM AGENCIES

122/123, Sunrise Comp- B4, Opp Ambe School, Manjalpur, VADODARA - Mob 9586882828

152, Parshram Ind Estate, Nr Janta Nagar, Ramol, AHEMDABAD - Mob 9586883311

MONTH Partywise Issue Summary

From Date 01/04/2021 To 31/03/2022

Product Name	Qty	Month
April		
UN Combo75 OM/HN/FN(24x3x100g)	15	April
Total	**15**	
***RAJA TRADERS**		
UN FP20 CASHW/BDM (72x120g)	72	April
UN FP20 CH CHIPS (144x75g)	432	April
UN FP25 BUTTER (144x75g)	144	April
UN FP25 CASHEW (144x75g)	288	April
UN FP25 CHOCO NUT (144x75g)	288	April
UN FP25 FRUIT & NUT (144x75g)	288	April
UN FP25 HONEY OTMIL (144x75g)	144	April
UN FP25 PISTA BDM (144x75g)	288	April
UN SS 100 CSHW/BDM (24x300g)	24	April
UN SS 165 CHOCONUT (18x500g)	18	April
Total	**1986**	

ii. <u>SKU mapping:</u>

Mapping of different products to their respective categories along with MRP.

Product Name	Mapping	Category	MRP
U CH Chips (100gx5x28) MRP 165	SS Choco Chips (100gx5)	Super Saver	165
U Choc/Rpl (100gx2x48) MRP 38	1+1 Choco Ripple (100gx2)	Combo	40
U Fruit & Nut(100gx5x28) MRP 155	SS Fruit & Nut (100gx5)	Super Saver	155
U Orng/Milk(100gx2x48) MRP 38	1+1 Orange Milk (100gx2)	Combo	40
U Choco/HzlNut (200gx2x36) 60/-	1+1 Choco Hazelnut (200gx2)	Combo	60
U Orng/Milk (200gx2x36) 60/-	1+1 Orange Milk (200gx2)	Combo	60
U Asrtd Pack of 6 (24) MRP 125	Assorted Pack of 6	Combo	125
U Butter FP (100gx2x18) MRP 50	FP50 Butter (100gx2)	Family Pack	50
U Choc Chips (24) MRP 90	Assorted Pack of 3	Combo	90
U Combo OM/HO/FN (24) MRP 75	Assorted Pack of 3	Combo	75
U Crom Combo (32) MRP 55	Assorted Pack of 3	Combo	55
U Cshw/Bdm (18) MRP 120	SS Cashew Badam (100gx5)	Super Saver	120

iii. <u>Store mapping:</u>

Mapping of different stores to their respective channel type, area and route.

Store Name	Store Name	Channel Type	Area	Route
.ARIHANT MEDICAL GEN STORE	.ARIHANT MEDICAL GEN STORE	General Trade	Harni	Old City
.ARYA SNACKS	.ARYA SNACKS	General Trade	Alkapuri	New City
.ASHA STORE	.ASHA STORE	General Trade	Ajwa Road	Old City
.ASHAPURI PROV (GORWA)	.ASHAPURI PROV	General Trade	Gorwa	New City
.ASHAPURI STORE	.ASHAPURI STORE	General Trade	Harni	Old City
.ASHIRWAD CHEMIST	.ASHIRWAD CHEMIST	General Trade	Karelibaug	Old City
.BABA SUPER STORE	.BABA SUPER STORE	General Trade	Ajwa Road	Old City
.BARODA DAIRY PARLOUR	.BARODA DAIRY PARLOR	General Trade	Vadsar	New City
.BHAGAWATI MEDICAL AND GENERAL STORE	.BHAGAWATI MEDICAL AND GENERAL STORE	General Trade	Sama Savli Road	Old City
.BHAGVATI ENTERPRISE	.BHAGVATI ENTERPRISE	General Trade	Harni	Old City
.BHAGVATI GIFT SHOP	.BHAGVATI GIFT SHOP	General Trade	Varasiya	Old City
.BHAVANI SUPER MARKET	.BHAVANI SUPER MARKET	General Trade	Gorwa	New City
.BIG MART	.BIG MART	General Trade	Nizampura	New City
.BRAVO BAZZAR	.BRAVO BAZZAR	General Trade	Sama Savli Road	Old City
.CENTRAL SUPER MARKET	.CENTRAL SUPER MARKET	General Trade	Nizampura	New City
.CHARMY FARSAN	.CHARMY FARSAN	General Trade	Chhani	New City

2. Secondary variables taken under consideration:

   i. Temperature: Monthly Average temperature at 2 meters in °C
   ii. Rainfall: Monthly Average precipitation at 2 meters in (mm/day)
   iii. Festivals: Dummy variable

$$Festival = \begin{cases} 1, & if\ the\ month\ has\ any\ big\ festival\ or\ vacation \\ 0, & Otherwise \end{cases}$$

3. Transformed data files are divided into three files:

   a. **Data with Mass category**

      i. <u>Combined Data:</u>

Different SKUs purchased by particular store in a month of a specific year. For three years each month, quantities of distinct SKUs belonging to various categories were purchased by several stores of different channel types, located in different areas of a specific route. We have also added a few secondary variables like temperature, rainfall, and festival to know their influence on the distributorship.

Year	Month	Store Names	Channel Type	Area	Route	Product Names	Category	Qty	MRP	Sales	Festival	Temperature	Rainfall
2019	July	.MOHAN STORE	General Trade	Varasiya	Old City	SS Choco Chips (100gx5)	Super Saver	3	165	495	0	29.77322581	8.7
2019	July	.MOHAN STORE	General Trade	Varasiya	Old City	1+1 Choco Ripple (100gx2)	Combo	2	40	80	0	29.77322581	8.7
2019	July	.MOHAN STORE	General Trade	Varasiya	Old City	SS Fruit & Nut (100gx5)	Super Saver	3	155	465	0	29.77322581	8.7
2019	July	.MOHAN STORE	General Trade	Varasiya	Old City	1+1 Orange Milk (100gx2)	Combo	2	40	80	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	SS Choco Chips (100gx5)	Super Saver	3	165	495	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	1+1 Choco Ripple (100gx2)	Combo	1	40	40	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	1+1 Choco Hazelnut (200gx2)	Combo	1	60	60	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	SS Fruit & Nut (100gx5)	Super Saver	1	155	155	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	1+1 Orange Milk (100gx2)	Combo	1	40	40	0	29.77322581	8.7
2019	July	XGANESH STORE	Institutional Trade	-	-	SS Choco Chips (100gx5)	Super Saver	1	165	165	0	29.77322581	8.7
2019	July	XGANESH STORE	Institutional Trade	-	-	1+1 Choco Ripple (100gx2)	Combo	1	40	40	0	29.77322581	8.7
2019	July	XGANESH STORE	Institutional Trade	-	-	SS Fruit & Nut (100gx5)	Super Saver	1	155	155	0	29.77322581	8.7
2019	July	XGANESH STORE	Institutional Trade	-	-	1+1 Orange Milk (200gx2)	Combo	1	60	60	0	29.77322581	8.7
2019	July	XGANESH STORE	Institutional Trade	-	-	1+1 Orange Milk (100gx2)	Combo	1	40	40	0	29.77322581	8.7

      ii. <u>By Stores:</u>

Month wise the total sales of unique stores correspond to their respective channel type and areas of different levels of routes. Includes Store wise sales per month along with the channel type, area and route of the store and average temperature(°C) and average rainfall(mm/day) for every month.

Ye	Month	Store Names	Channel Type	Area	Route	Sales	Festiv	Temperatu	Rainf
2019	July	.MOHAN STORE	General Trade	Varasiya	Old City	1120	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	790	0	29.77322581	8.7
2019	July	XGANESH STORE	Institutional Trade	-	-	460	0	29.77322581	8.7
2019	July	XPRAHLADBHAI - VARSIA	Institutional Trade	-	-	495	0	29.77322581	8.7
2019	August	.MOHAN STORE	General Trade	Varasiya	Old City	1740	1	27.21096774	15.938
2019	August	.RAJA TRADERS	Institutional Trade	-	-	245	1	27.21096774	15.938
2019	September	.APEX DRY FRUIT STORE (O P RD)	General Trade	OP Road	New City	1860	0	27.266	10.712
2019	September	.JAGDISH FARSAN PVT LTD	Institutional Trade	Karelibaug	-	32840	0	27.266	10.712

iii. By Products:

Month-wise total sales of different quantities of unique products are represented with respect to the categories in which they belong. Includes Product wise sales per month along with the category and MRP of that product, total quantity of it purchased in a month and average temperature(°C) and average rainfall(mm/day) for every month.

Year	Month	Product Names	Category	Qty	MRP	Sales	Festival	Temperature	Rainfall
2019	July	1+1 Choco Hazelnut (200gx2)	Combo	1	60	60	0	29.77322581	8.7
2019	July	1+1 Choco Ripple (100gx2)	Combo	4	40	160	0	29.77322581	8.7
2019	July	1+1 Orange Milk (100gx2)	Combo	4	40	160	0	29.77322581	8.7
2019	July	1+1 Orange Milk (200gx2)	Combo	1	60	60	0	29.77322581	8.7
2019	July	SS Choco Chips (100gx5)	Super Saver	10	165	1650	0	29.77322581	8.7
2019	July	SS Fruit & Nut (100gx5)	Super Saver	5	155	775	0	29.77322581	8.7
2019	August	1+1 Choco Hazelnut (200gx2)	Combo	3	60	180	1	27.21096774	15.938

iv. Mass Products:

Month-wise total sales of different quantities of unique products are represented with respect to the category MASS. Includes Product wise sales per month along with the category Mass and MRP of that product, total quantity of it purchased in a month and average temperature(°C) and average rainfall(mm/day) for every month.

Year	Month	Store Names	Channel Type	Area	Route	Product Names	Category	Qty	MRP	Sales	Festival	Temperature	Rainfall
2020	January	.Alembic Empl C	General Trade	Subhanpura	New City	BD Cashew Badam (12x16x60g)	Mass	2	120	240	0	18.04645161	0
2020	January	.Alembic Empl C	General Trade	Subhanpura	New City	BD Choco Hazelnut (12x16x60g)	Mass	2	120	240	0	18.04645161	0
2020	January	.Alembic Empl C	General Trade	Subhanpura	New City	BD Choco Ripple (12x16x60g)	Mass	2	120	240	0	18.04645161	0
2020	January	.ASHOK PROVISI	General Trade	Vadsar	New City	BD Cashew Badam (12x16x60g)	Mass	1	120	120	0	18.04645161	0
2020	January	.ASHOK PROVISI	General Trade	Vadsar	New City	BD Choco Hazelnut (12x16x60g)	Mass	1	120	120	0	18.04645161	0
2020	January	.ASHOK PROVISI	General Trade	Vadsar	New City	BD Choco Ripple (12x16x60g)	Mass	1	120	120	0	18.04645161	0
2020	January	.CENTRAL SUPER	General Trade	Nizampura	New City	BD Cashew Badam (12x16x60g)	Mass	3	120	360	0	18.04645161	0

v. Store vs Categories:

(Qty): For three years of each month, the quantity of unique stores is noted corresponding to each **category** purchased by that particular store. This particular sheet would represent the quantities of different categories purchased by each store in a month.

Year	Month	Store Names	Coml	Display B	Family Pa	Mas	Sugarfr	Super Sav	Grand To
2019	July	.MOHAN STORE	4					6	10
2019	July	.RAJA TRADERS	3					4	7
2019	July	XGANESH STORE	3					2	5
2019	July	XPRAHLADBHAI - VARSIA						3	3
2019	August	.MOHAN STORE	16		3			3	22
2019	August	.RAJA TRADERS	2					1	3
2019	September	.APEX DRY FRUIT STORE (O P RD)					42		42

(Sales): For three years of each month, sales of unique stores are noted corresponding to each **category** purchased by that particular store.

Year	Month	Store Names	Coml	Display B	Family Pa	Mass	Sugarfr	Super Sav	Grand To
2019	July	.MOHAN STORE	160					960	1120
2019	July	.RAJA TRADERS	140					650	790
2019	July	XGANESH STORE	140					320	460
2019	July	XPRAHLADBHAI - VARSIA						495	495
2019	August	.MOHAN STORE	1230		150			360	1740
2019	August	.RAJA TRADERS	80					165	245
2019	September	.APEX DRY FRUIT STORE (O P RD)					1860		1860

vi. Store vs Products:

(Qty): For three years each month, quantities purchased by unique stores are noted corresponding to each **SKUs**. This particular sheet would represent the quantities of different SKUs purchased by each store in a month.

Year	Month	Store Names	SF DD Multigrain (72x75g)	SF DD Oatmeal (72x75g)	SS Cashew Badam (100gx5)	SS Choco Chips (100gx5)	Grand Total
2019	July	.MOHAN STORE				3	10
2019	July	.RAJA TRADERS				3	7
2019	July	XGANESH STORE				1	5
2019	July	XPRAHLADBHAI - VARSIA				3	3
2019	August	.MOHAN STORE			3		22
2019	August	.RAJA TRADERS				1	3
2019	September	.APEX DRY FRUIT STORE (O P RD)	6	6			42
2019	September	.JAGDISH FARSAN PVT LTD	24	24	52	52	384
2019	September	.JAGDISH FOODS PVT LTD	12	12			72

(Sales): For three years of each month, sales of unique stores are noted corresponding to each **SKUs** purchased by that particular store.

Year	Month	Store Names	SF DD Multigrain (72x75g)	SF DD Oatmeal (72x75g)	SS Cashew Badam (100gx5)	SS Choco Chips (100gx5)	Grand Total
2019	July	.MOHAN STORE				495	1120
2019	July	.RAJA TRADERS				495	790
2019	July	XGANESH STORE				165	460
2019	July	XPRAHLADBHAI - VARSIA				495	495
2019	August	.MOHAN STORE			360		1740
2019	August	.RAJA TRADERS				165	245
2019	September	.APEX DRY FRUIT STORE (O P RD)	300	240			1860
2019	September	.JAGDISH FARSAN PVT LTD	1200	960	6240	8580	32840
2019	September	.JAGDISH FOODS PVT LTD	600	480			3120

b. Data without Mass category

i. Combined Data:

Different kinds of SKUs purchased by a particular store in some month of a specific year. For three years each month, quantities of distinct SKUs belonging to various categories EXCLUDING MASS were purchased by several stores of different channel types, located in different areas of a specific route. We have also added a few secondary variables like temperature, rainfall, and festival to know their influence on the distributorship.

Year	Month	Store Names	Channel Type	Area	Route	Product Names	Category	Qty	MRP	Sales	Festival	Temperature	Rainfall
2019	July	.MOHAN STORE	General Trade	Varasiya	Old City	SS Choco Chips (100gx5)	Super Saver	3	165	495	0	29.77322581	8.7
2019	July	.MOHAN STORE	General Trade	Varasiya	Old City	1+1 Choco Ripple (100gx2)	Combo	2	40	80	0	29.77322581	8.7
2019	July	.MOHAN STORE	General Trade	Varasiya	Old City	SS Fruit & Nut (100gx5)	Super Saver	3	155	465	0	29.77322581	8.7
2019	July	.MOHAN STORE	General Trade	Varasiya	Old City	1+1 Orange Milk (100gx2)	Combo	2	40	80	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	SS Choco Chips (100gx5)	Super Saver	3	165	495	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	1+1 Choco Ripple (100gx2)	Combo	1	40	40	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	1+1 Choco Hazelnut (200gx2)	Combo	1	60	60	0	29.77322581	8.7
2019	July	.RAJA TRADERS	Institutional Trade	-	-	SS Fruit & Nut (100gx5)	Super Saver	1	155	155	0	29.77322581	8.7

ii. By Stores:

Month wise the total sales of unique stores correspond to their respective channel types and areas of different levels of routes. Includes Store wise sales per month along with the channel types, area and route of the store and average temperature(°C) and average rainfall(mm/day) for every month.

Year	Month	M	Store Names	Channel Type	Area	Route	Qty	MRP	Sales	Festival	Temperature	Rainfall
2019	July	July_2019	.MOHAN STORE	General Trade	Varasiya	Old City	10	40	1120	0	29.77322581	8.7
2019	July	July_2019	.RAJA TRADERS	Institutional Trade	-	-	7	40	790	0	29.77322581	8.7
2019	July	July_2019	XGANESH STORE	Institutional Trade	-	-	5	40	460	0	29.77322581	8.7
2019	July	July_2019	XPRAHLADBHAI - VARSIA	Institutional Trade	-	-	3	165	495	0	29.77322581	8.7
2019	August	August_20	.MOHAN STORE	General Trade	Varasiya	Old City	22	60	1740	1	27.21096774	15.938
2019	August	August_20	.RAJA TRADERS	Institutional Trade	-	-	3	40	245	1	27.21096774	15.938
2019	Septembe	Septembe	.APEX DRY FRUIT STORE (O P RD)	General Trade	OP Road	New City	42	50	1860	0	27.266	10.712

iii. By Products:

Month-wise total sales of different quantities of unique products are represented with respect to their respective categories EXCLUDING MASS in which they belong. Includes Product wise sales per month along with the category and MRP of that product, total quantity of it purchased in a month and average temperature(°C) and average rain-fall(mm/day) for every month.

Year	Month	M	Product Names	Category	Qty	MRP	Sales	Festival	Temperature	Rainfall
2019	July	July_2019	1+1 Choco Hazelnut (200gx2)	Combo	1	60	60	0	29.77322581	8.7
2019	July	July_2019	1+1 Choco Ripple (100gx2)	Combo	4	40	160	0	29.77322581	8.7
2019	July	July_2019	1+1 Orange Milk (100gx2)	Combo	4	40	160	0	29.77322581	8.7
2019	July	July_2019	1+1 Orange Milk (200gx2)	Combo	1	60	60	0	29.77322581	8.7
2019	July	July_2019	SS Choco Chips (100gx5)	Super Saver	10	165	1650	0	29.77322581	8.7
2019	July	July_2019	SS Fruit & Nut (100gx5)	Super Saver	5	155	775	0	29.77322581	8.7
2019	August	August_20	1+1 Choco Hazelnut (200gx2)	Combo	3	60	180	1	27.21096774	15.938

iv. Store vs Products:

(Qty): For three years each month, quantities purchased by unique stores are noted corresponding to each **SKUs**. This particular sheet would represent the quantities of different SKUs purchased by each store in a month.

Year	Month	Store Names	1+1 Choco Hazelnut (200gx2)	1+1 Choco Ripple (100gx2	1+1 DD Oatmeal (150gx2)	1+1 Orange Milk (100gx2)	1+1 Orange Milk (200gx2)	Grand Total
2019	July	.MOHAN STORE		2		2		10
2019	July	.RAJA TRADERS	1	1		1		7
2019	July	XGANESH STORE		1		1	1	5
2019	July	XPRAHLADBHAI - VARSIA						3
2019	August	.MOHAN STORE	3	2			3	22
2019	August	.RAJA TRADERS		1		1		3
2019	Septembei	.APEX DRY FRUIT STORE (O P RD)						42
2019	Septembei	.JAGDISH FARSAN PVT LTD		24		24		384
2019	Septembei	.JAGDISH FOODS PVT LTD						72

(Sales): For three years of each month, sales of unique stores are noted corresponding to each **SKUs** purchased by that particular store.

Year	Month	Store Names	1+1 Choco Hazelnut (200gx2)	1+1 Choco Ripple (100gx2	SS Choco Chips (100gx5)	SS Fruit & Nut (100gx5	Grand Total
2019	July	.MOHAN STORE		80	495	465	1120
2019	July	.RAJA TRADERS	60	40	495	155	790
2019	July	XGANESH STORE		40	165	155	460
2019	July	XPRAHLADBHAI - VARSIA			495		495
2019	August	.MOHAN STORE	180	80			1740
2019	August	.RAJA TRADERS		40	165		245
2019	Septembei	.APEX DRY FRUIT STORE (O P RD)					1860
2019	Septembei	.JAGDISH FARSAN PVT LTD		960	8580	8060	32840

v. Store vs Categories:

(Qty): For three years of each month, the quantity of unique stores is noted corresponding to each **category** EXCEPT MASS purchased by that particular store. This particular sheet would represent the quantities of different categories purchased by each store in a month.

Year	Month	Store Names	Combo	Display Box	Family Pack	Sugarfree	Super Saver	Grand Total
2019	July	.MOHAN STORE	4				6	10
2019	July	.RAJA TRADERS	3				4	7
2019	July	XGANESH STORE	3				2	5
2019	July	XPRAHLADBHAI - VARSIA					3	3
2019	August	.MOHAN STORE	16		3		3	22
2019	August	.RAJA TRADERS	2				1	3
2019	Septembe	.APEX DRY FRUIT STORE (O P RD)				42		42

(Sales): For three years of each month, sales of unique stores are noted corresponding to each **category** EXCEPT MASS purchased by that particular store.

Year	Month	Store Names	Combo	Display Box	Family Pack	Sugarfree	Super Saver
2019	July	.MOHAN STORE	160				960
2019	July	.RAJA TRADERS	140				650
2019	July	XGANESH STORE	140				320
2019	July	XPRAHLADBHAI - VARSIA					495
2019	August	.MOHAN STORE	1230		150		360
2019	August	.RAJA TRADERS	80				165
2019	Septembe	.APEX DRY FRUIT STORE (O P RD)				1860	
2019	Septembe	.JAGDISH FARSAN PVT LTD	1920			8040	22880

c. Quaterwise Data of Sales

i. Quarterwise data:

Here we have shown along with the Combined data, the distribution of the data into different Quarters of that particular year.

Year	Month	Quarter	Store Names	Channel Type	Area	Route	Product Names	Category	Qty	MRP	Sales	Festival	Temperature	Rainfall
2019	July	Q3_2019	.MOHAN STORE	General Trade	Varasiya	Old City	SS Choco Chips (100gx5)	Super Saver	3	165	495	0	29.77322581	8.7
2019	July	Q3_2019	.MOHAN STORE	General Trade	Varasiya	Old City	1+1 Choco Ripple (100gx2)	Combo	2	40	80	0	29.77322581	8.7
2019	July	Q3_2019	.MOHAN STORE	General Trade	Varasiya	Old City	SS Fruit & Nut (100gx5)	Super Saver	3	155	465	0	29.77322581	8.7
2019	July	Q3_2019	.MOHAN STORE	General Trade	Varasiya	Old City	1+1 Orange Milk (100gx2)	Combo	2	40	80	0	29.77322581	8.7
2019	July	Q3_2019	.RAJA TRADERS	Institutional Trade	-	-	SS Choco Chips (100gx5)	Super Saver	3	165	495	0	29.77322581	8.7
2019	July	Q3_2019	.RAJA TRADERS	Institutional Trade	-	-	1+1 Choco Ripple (100gx2)	Combo	1	40	40	0	29.77322581	8.7
2019	July	Q3_2019	.RAJA TRADERS	Institutional Trade	-	-	1+1 Choco Hazelnut (200gx2)	Combo	1	60	60	0	29.77322581	8.7
2019	July	Q3_2019	.RAJA TRADERS	Institutional Trade	-	-	SS Fruit & Nut (100gx5)	Super Saver	1	155	155	0	29.77322581	8.7

ii. Categorywise Quarterly Quantities:

The Quantities of different categories falling under specific Quarter of that particular year.

Categories	Q3_2019	Q4_2019	Q1_2020	Q2_2020	Q3_2020	Q4_2020	Q1_2021	Q2_2021	Q3_2021	Q4_2021	Q1_2022	Grand Total
Combo	93	1781	3541	1573	2599	1373	873	717	1528	887		14965
Display Box		1692	2213	1626	3101	2604	1611	1766	3535	2953	10	21111
Family Pack	3	1268	5497	8986	16684	12497	11318	12154	16034	7161		91602
Mass			245	83	579	2165	9311	11052	2794	1372		27601
Sugarfree	696	1323	1916	1332	3027	3591	2451	2374	2520	1090		20320
Super Saver	287	663	220	383	1265	2500	2960	2824	5161	2181		18444
Grand Total	1079	6727	13632	13983	27255	24730	28524	30887	31572	15644	10	194043

iii. Categorywise Quarterly Sales:

The Sales of different categories falling under specific Quarter of that particular year.

Categories	Q3_2019	Q4_2019	Q1_2020	Q2_2020	Q3_2020	Q4_2020	Q1_2021	Q2_2021	Q3_2021	Q4_2021	Q1_2022	Grand Total
Combo	4500	106700	185665	79505	112850	61380	42810	32285	67970	39200		732865
Display Box		82200	108040	81300	152830	129880	80490	88230	158530	127950	480	1009930
Family Pack	150	36685	141020	207920	386345	314715	279115	294395	390505	167255		2218105
Mass			29400	9960	69480	245100	1105620	1321200	325560	163320		3269640
Sugarfree	30720	57820	83420	60800	127405	140145	95080	95315	98695	43020		832420
Super Saver	42330	97120	31715	54325	138760	324440	404740	352445	654590	279475		2379940
Grand Total	77700	380525	579260	493810	987670	1215660	2007855	2183870	1695850	820220	480	10442900

iv. Productwise Quarterly Quantities:

The Quantities of different SKUs falling under specific Quarter of that particular year.

Products	Q3_2019	Q4_2019	Q1_2020	Q2_2020	Q3_2020	Q4_2020	Q1_2021	Q2_2021	Q3_2021	Q4_2021	Q1_2022	Grand Total
1+1 Choco Hazelnut (200gx2)	4											4
1+1 Choco Ripple (100gx2)	32	404	1078	387	671	301	117	67	237	143		3437
1+1 DD Ajwain (150gx2)		68	227	136			61	9				501
1+1 DD Oatmeal (150gx2)	15	512	716	676	697	367	171	70	101			3325
1+1 Milk (100gx2)			128	17								145
1+1 Orange Milk (100gx2)	30	300	882	245	847	147	96	26	22			2595
1+1 Orange Milk (200gx2)	4											4
1+1 Wafer Cheese (30x2x75g)							68	186	438	248		940

v. Productwise Quarterly Sales:

The Sales of different SKUs falling under specific Quarter of that particular year.

Products	Q3_2019	Q4_2019	Q1_2020	Q2_2020	Q3_2020	Q4_2020	Q1_2021	Q2_2021	Q3_2021	Q4_2021	Q1_2022	Grand Total
1+1 Choco Hazelnut (200gx2)	240											240
1+1 Choco Ripple (100gx2)	1280	16160	43120	15480	26840	12040	4680	2680	9480	5720		137480
1+1 DD Ajwain (150gx2)		4080	13620	8160			3660	540				30060
1+1 DD Oatmeal (150gx2)	750	25600	35800	33800	34850	18350	8550	3500	5050			166250
1+1 Milk (100gx2)			5120	680								5800
1+1 Orange Milk (100gx2)	1200	12000	35280	9800	33880	5880	3840	1040	880			103800
1+1 Orange Milk (200gx2)	240											240
1+1 Wafer Cheese (30x2x75g)							3060	8370	19710	11160		42300

vi. Storewise Quarterly Quantities:

The Quantities of different Stores falling under specific Quarter of that particular year.

Stores	Q3_2019	Q4_2019	Q1_2020	Q2_2020	Q3_2020	Q4_2020	Q1_2021	Q2_2021	Q3_2021	Q4_2021	Q1_2022	Grand Total
.24x7 CAFETERIA			2		81	35	192	192	312	120		934
.A B SUPER STORE							1	16	2			19
.AARDEEP STORE									9			9
.ADINATH GRAIN AND PROV STORE									2			2
.AGRAWAL FARM FRESH									60	30		90
.AGRAWAL JUICE CENTRE					4							4
.AGRAWAL PRO STORE					10	16						26
.AGRAWAL SUPER STORE								2	2			4

vii. Storewise Quarterly Sales:

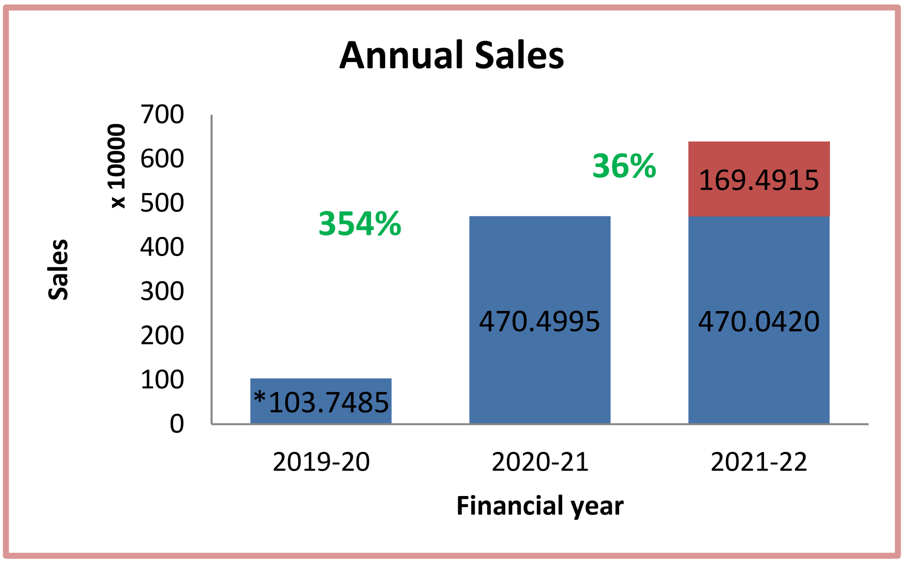
The Sales of different Stores falling under specific Quarter of that particular year.

Stores	Q3_2019	Q4_2019	Q1_2020	Q2_2020	Q3_2020	Q4_2020	Q1_2021	Q2_2021	Q3_2021	Q4_2021	Q1_2022	Grand Total
.24x7 CAFETERIA			240		2210	1005	19200	19200	26880	7680		76415
.A B SUPER STORE							120	1920	240			2280
.AARDEEP STORE									510			510
.ADINATH GRAIN AND PROV STORE									120			120
.AGRAWAL FARM FRESH									2700	1400		4100
.AGRAWAL JUICE CENTRE					180							180
.AGRAWAL PRO STORE					600	1920						2520
.AGRAWAL SUPER STORE								240	240			480

GRAPHICAL VISUALIZATIONS

Overall Sales :-

1)

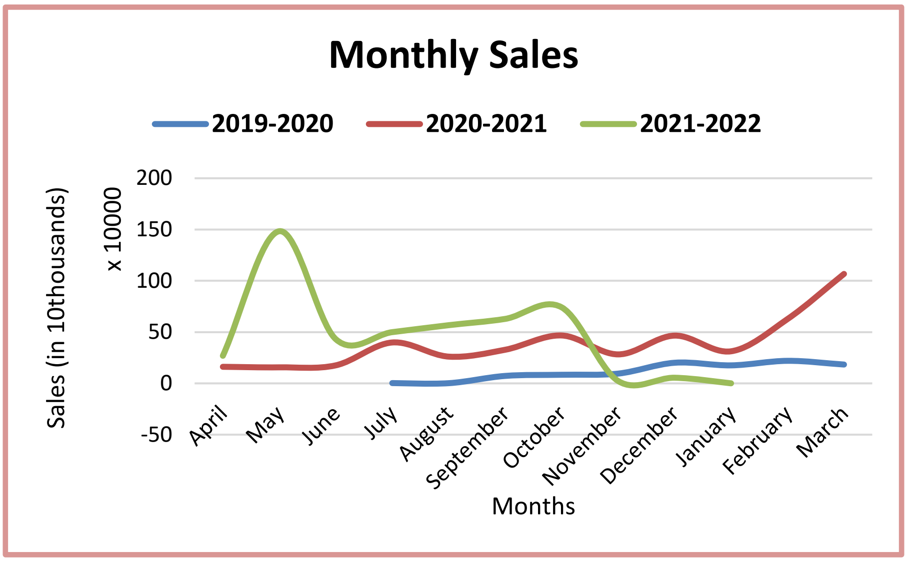


*For the year 2019 data was received from July onwards.

Interpretation: Here,

1. We can observe that there is 354% increase in sales from financial year 2019-20 (July – March) to financial year 2020-21
2. We predicted that there would be 36% increase in sales from financial year 2020-21 to 2021-22 by assuming the sales of February and March to be the same as previous year.

2)



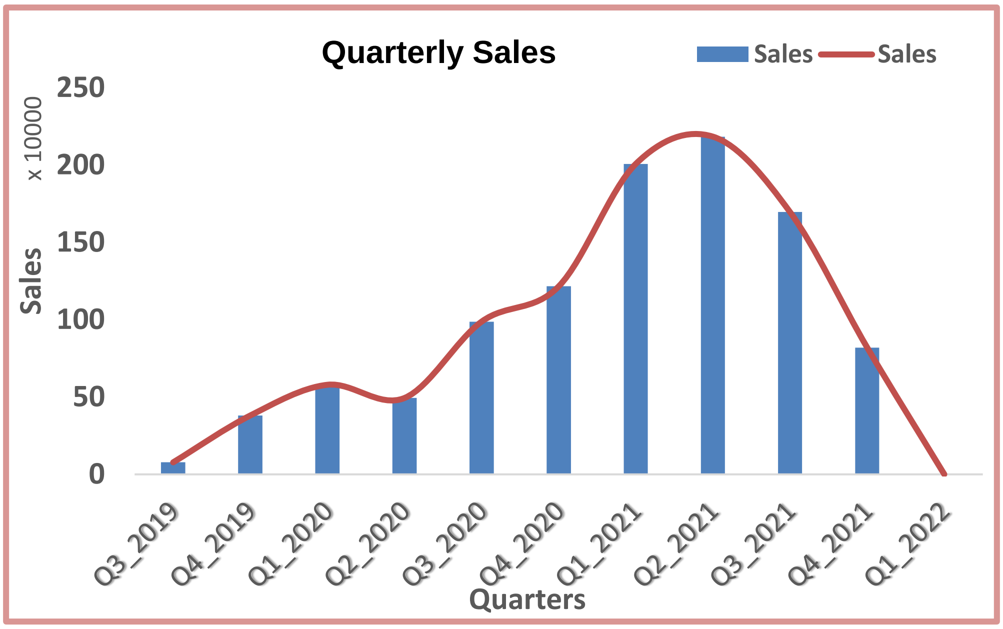
Interpretation:-

Here , from the figure, we can observe that there is steady pace of sales in both 2019-20 and 2020-21 while in 2021-22 , the sales has a peak in the month of May and then gradually decreases.

From the financial year 2020-21 and 2021-22 , we can observe a similar trend that there is a peak of sales in the month of October, while it falls for November.

Quarterwise Sales:-

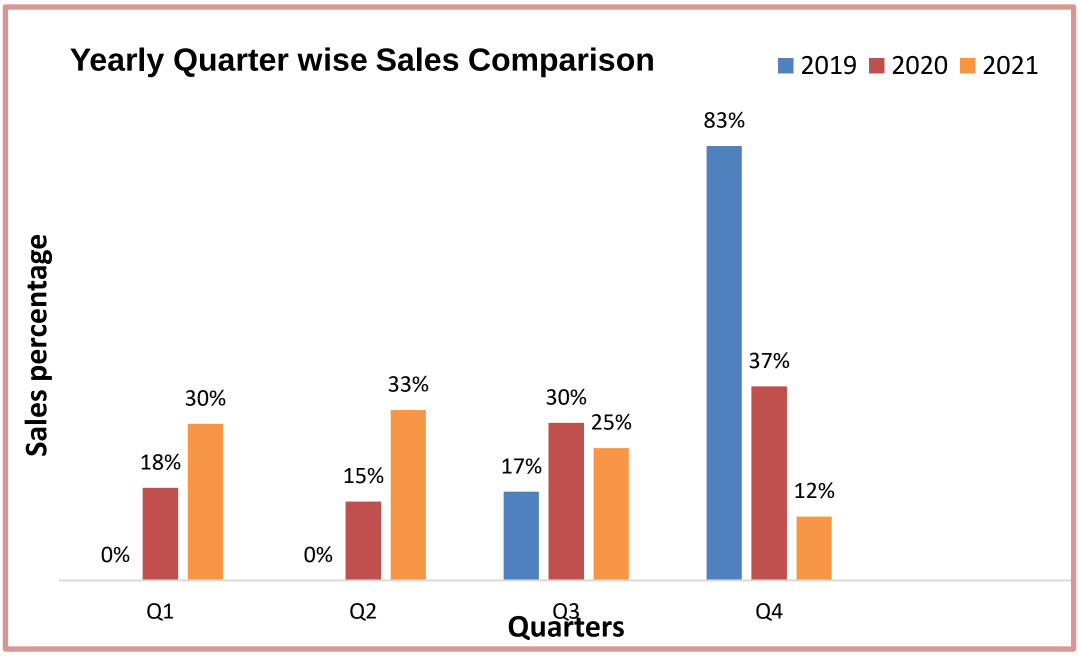
1)



Interpretation:

1. Quarter 1 sales: Increases from 2020 to 2021
2. Quarter 2 sales: Increases from 2020 to 2021
3. Quarter 3 sales: Increases from 2019, 2020 to 2021
4. Quarter 4 sales: Increases from 2019 to 2020, decreases from 2020 to 2021

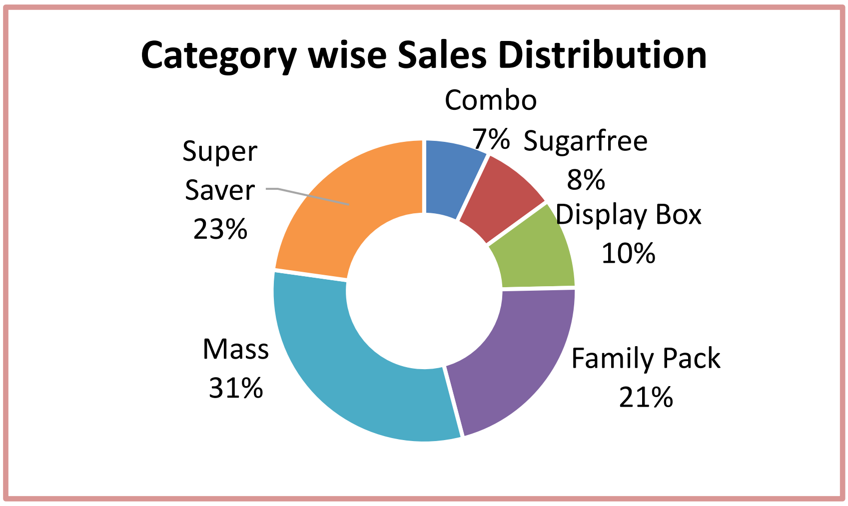
2)



Interpretation:
Sales from Q1 to Q4 increases for the year 2020 while Sales from Q1 to Q4 decreases in the year 2021

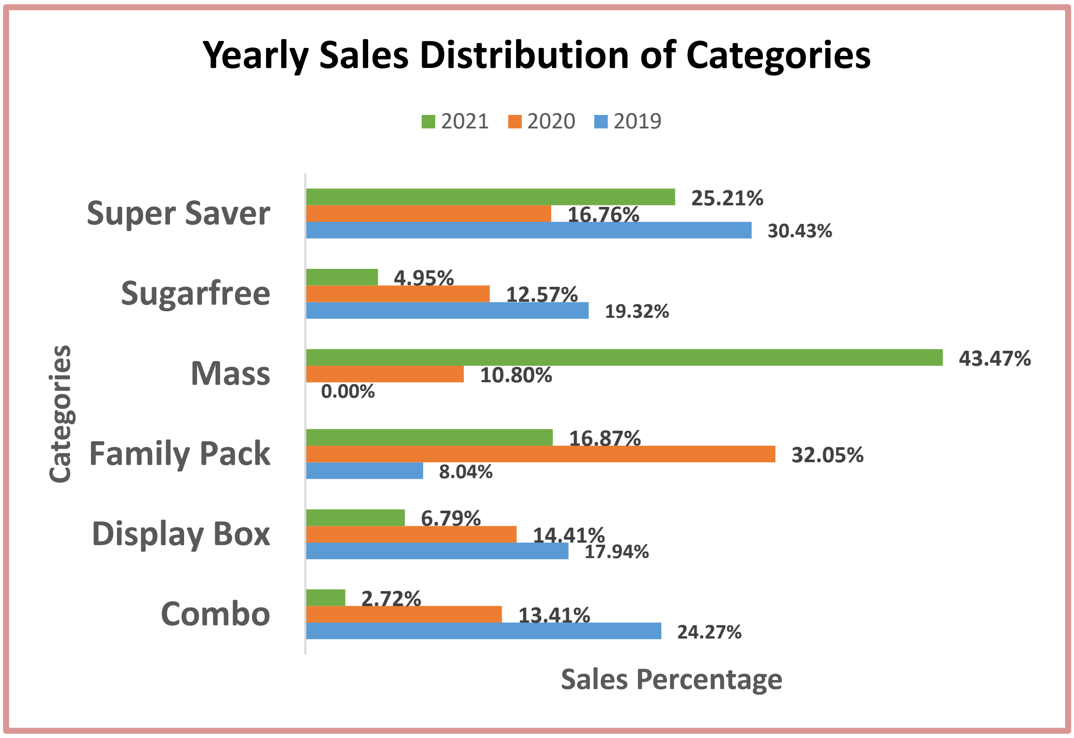
Categorywise sales

1)



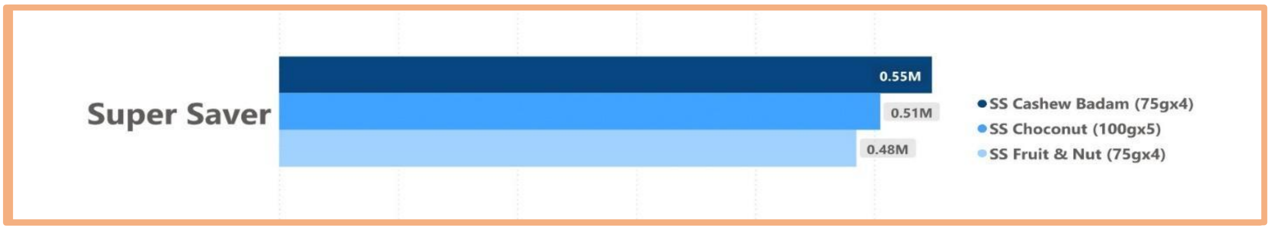
Interpretation:
Mass Category (31%) is most sold category followed by Super Saver (23%) and Family Pack (21%)

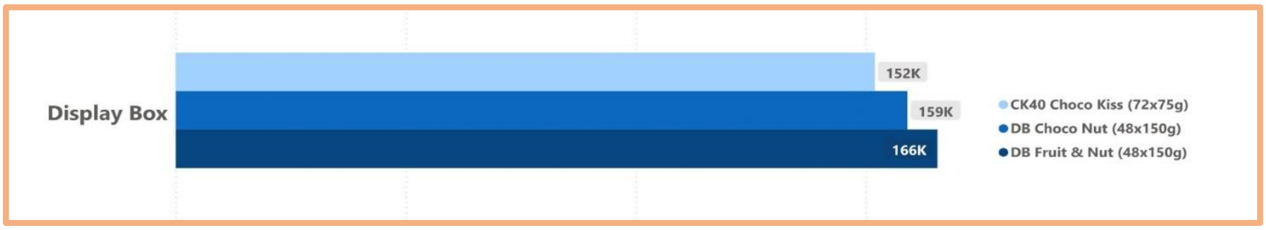
2)

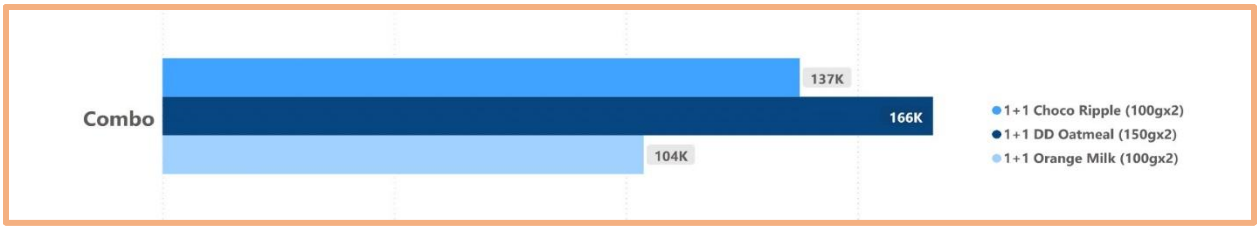


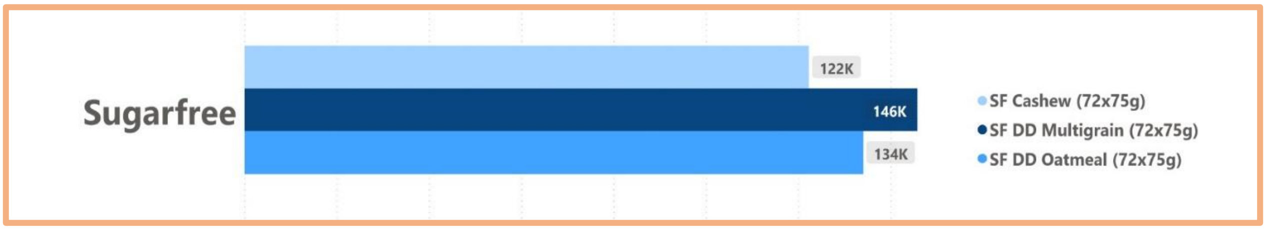
Interpretation:
Combo Category (24.27%) has the highest Sales in 2019, whereas in 2020 Family pack (32.05%) the has highest sales andin 2021 , Mass Category (43.47%) has the highest Sales.

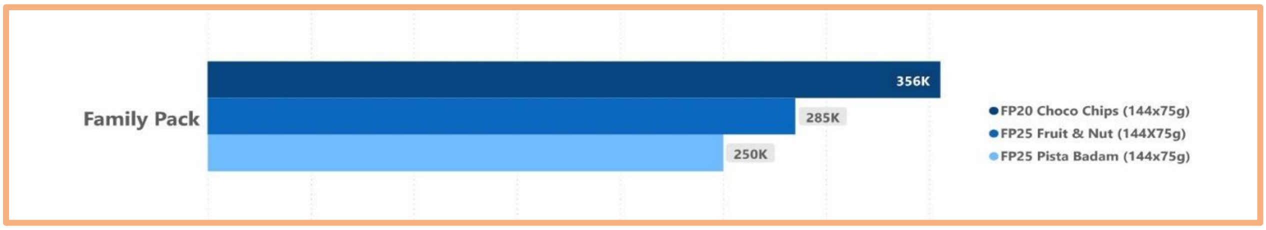
3) Category-wise sales of top 3 products









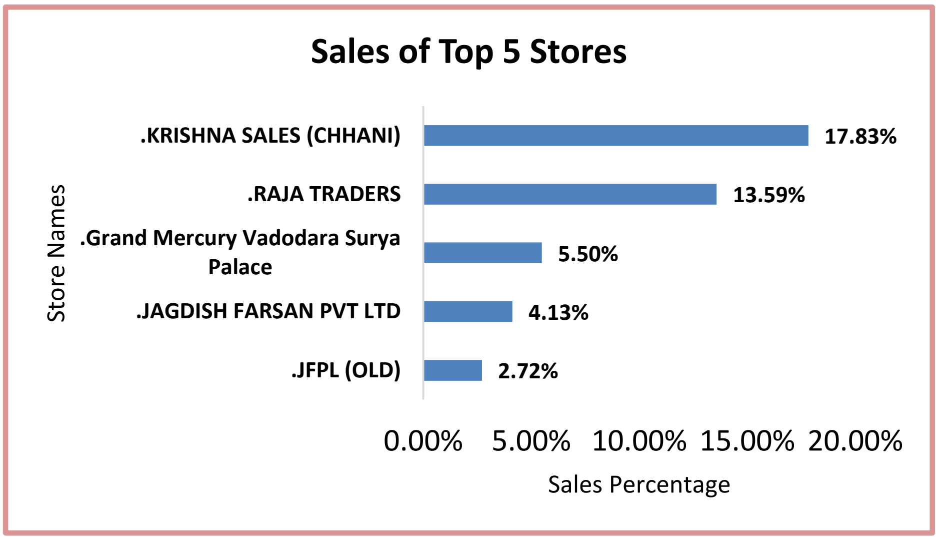


Storewise:

1) Yearwise distribution of Stores

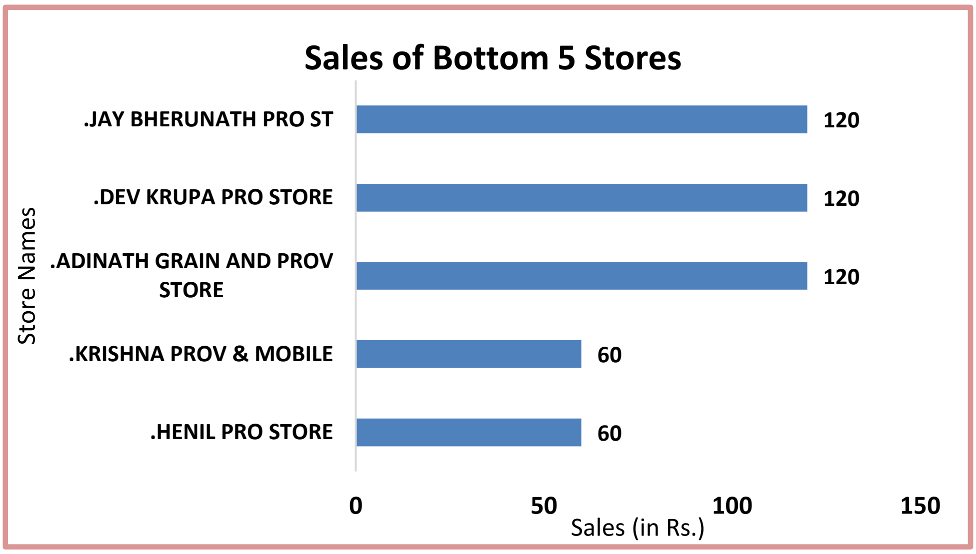
Interpretation:
In year 2019 , existing store were 37. Out of this stores 27 continued in year 2020 while 183 new stores were added and 10 stores discontinued. Out of 210 stores in 2020, 115 continued in 2021 and 262 were newly added while 95 stores discontinued.

2) Sales of Top 5 and Bottom 5 Stores:



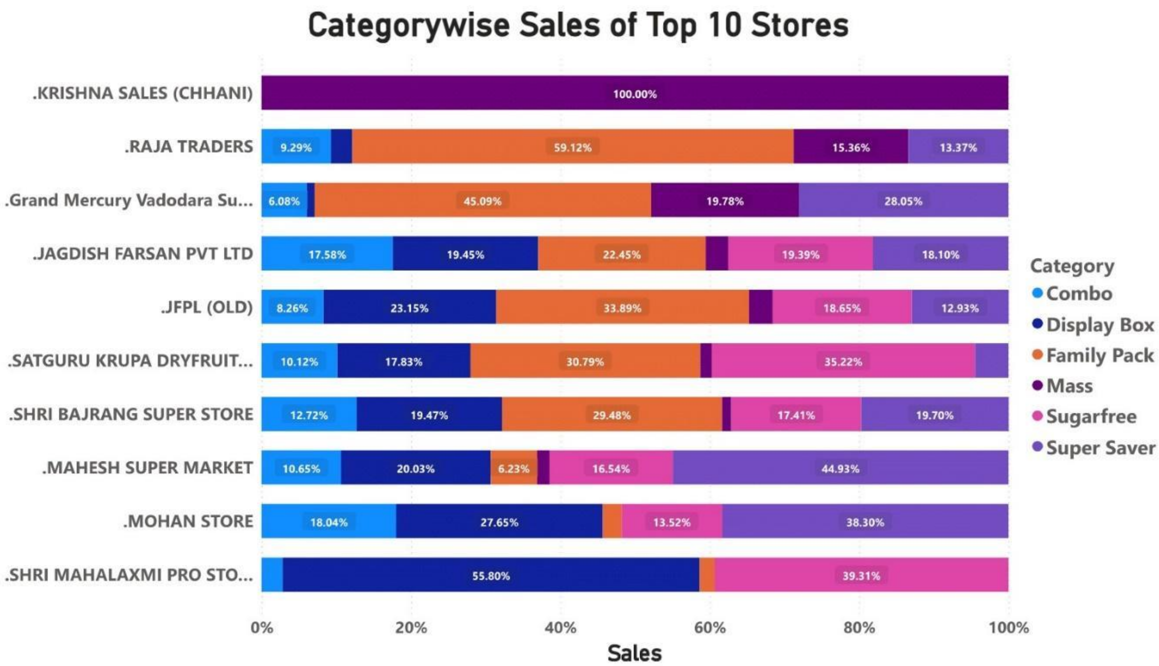
Interpretation:
KRISHNA SALES (17.83%) has the most sales out of all stores followed by RAJA TRADERS (13.59%) and Surya Palace(5.50%).

3)



Interpretation :
These are the stores which have the least sales among all the stores.

4) Sales of Top Ten Stores Sub-divided in Categories



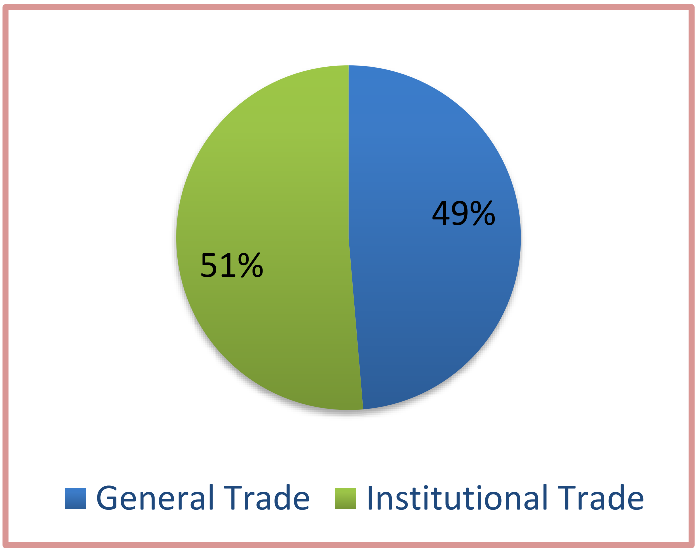
Interpretation :

a.We can observe that Krishna Sales (Chhani) ,which is also the TOP 1 store with highest sales, sells products of Mass Category.

b. We can also observe funnel shapes for Family Pack and Display Box Category. As the rank for Top selling stores decreases (Top to bottom), the distribution of Family pack decreases, suggesting that higher the sales of Family pack, higher would be the Sales for the store.

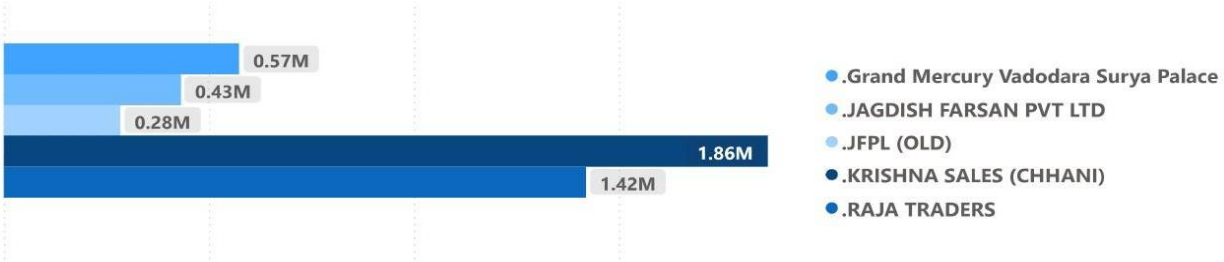
c. Similarly, we can observe as the rank for Top selling stores increases (Bottom to Top) , the distribution of Display box decreases, suggesting that less the distribution of Display box, higher would be the sales for the store.

5) Sales By Channel Type



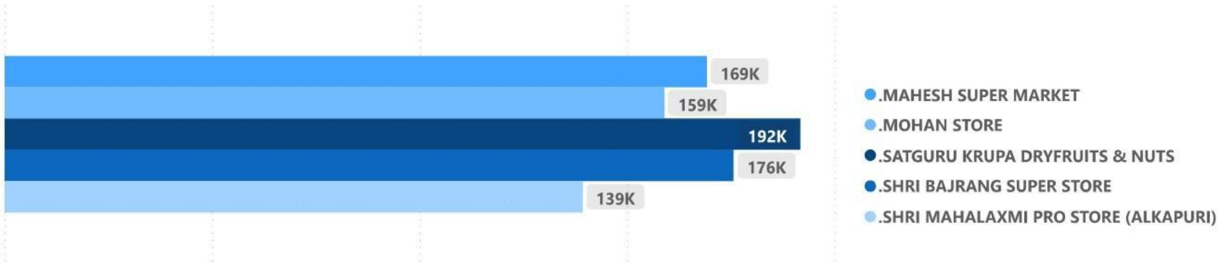
Interpretation :
General Trade accounts for 51% of total Sales while Institutional Trade covers 49% of total Sales.

6) Sales of Top 5 Stores of Institutional trade



Interpretation :
KRISHNA SALES is the highest selling store in Institutional trade

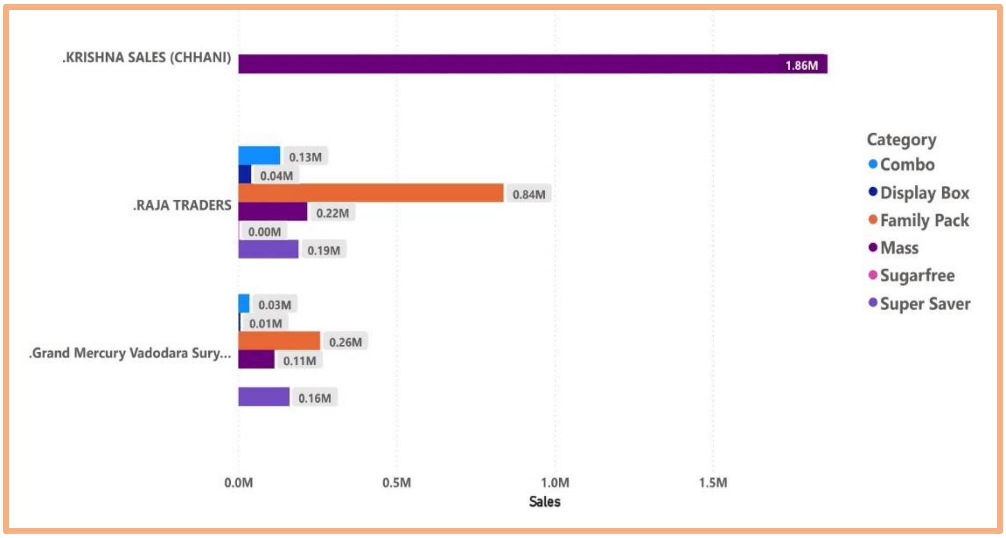
7) Sales of Top 5 stores of General Trade



Interpretation :
SATGURU KRUPA DRYFRUITS & NUTS is the highest selling store in General trade.

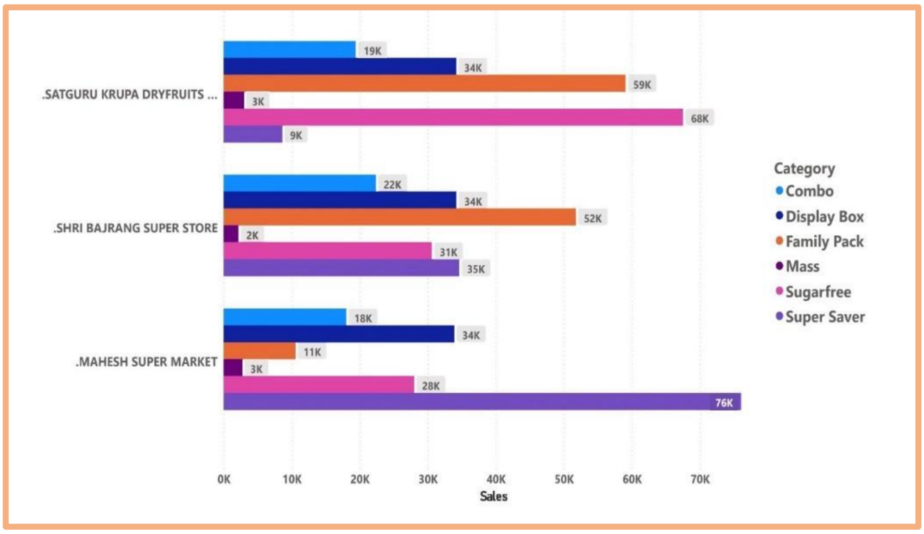
Category wise - Sales of Top Three Stores (Channels)

1) Institutional Trade



Interpretation :
Products of Mass Category and Family Pack category are the most popular in the Top 3 stores of Institutional trade.

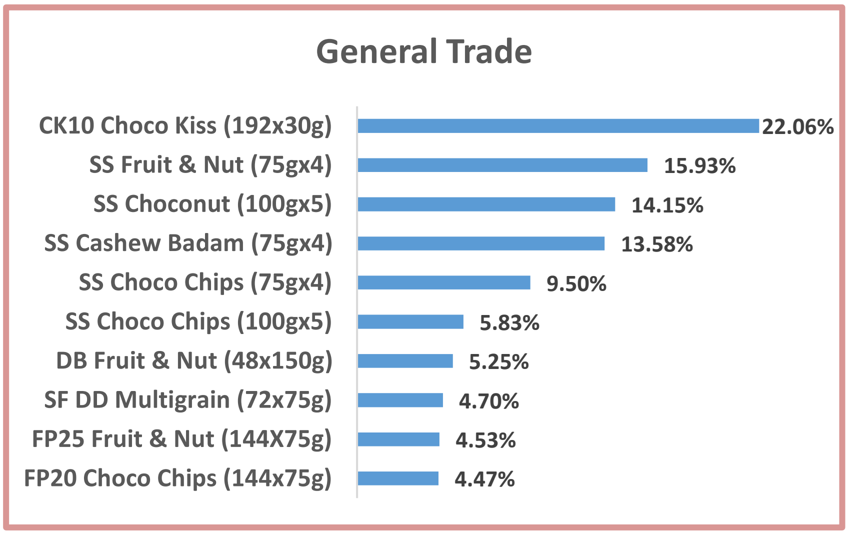
2) General Trade



Interpretation :
Products of Sugarfree , Super Saver and Family Pack are the most popular in the TOP 3 stores of General Trade.

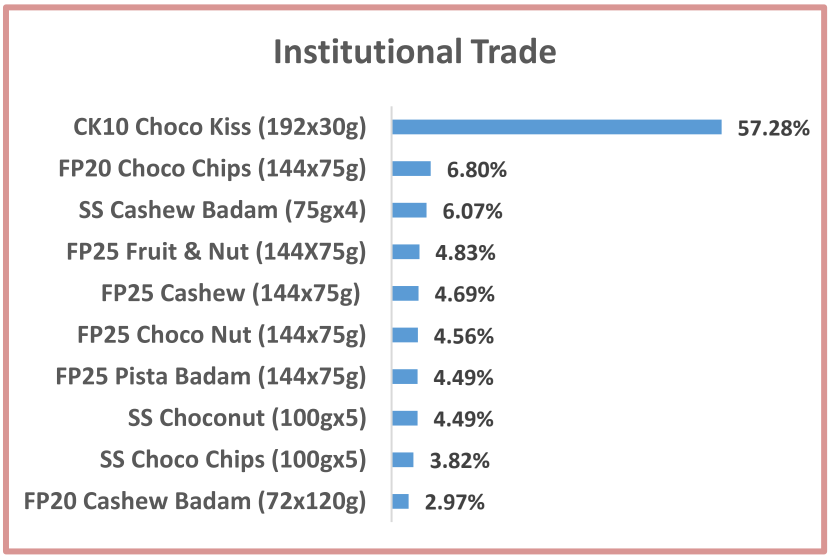
Productwise

Sales of Top Ten Products in Channels
1)



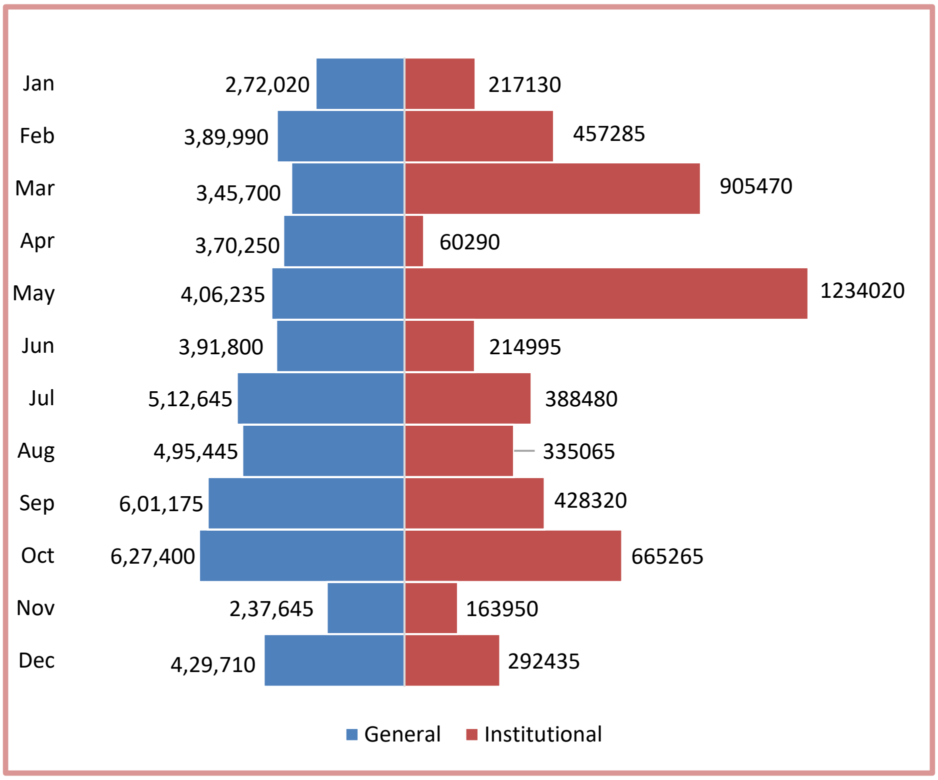
Interpretation :
Choco Kiss (22%) is the highest selling product in General Trade followed by Fruit & Nut (15.93%) and Choconut (14%).

2)

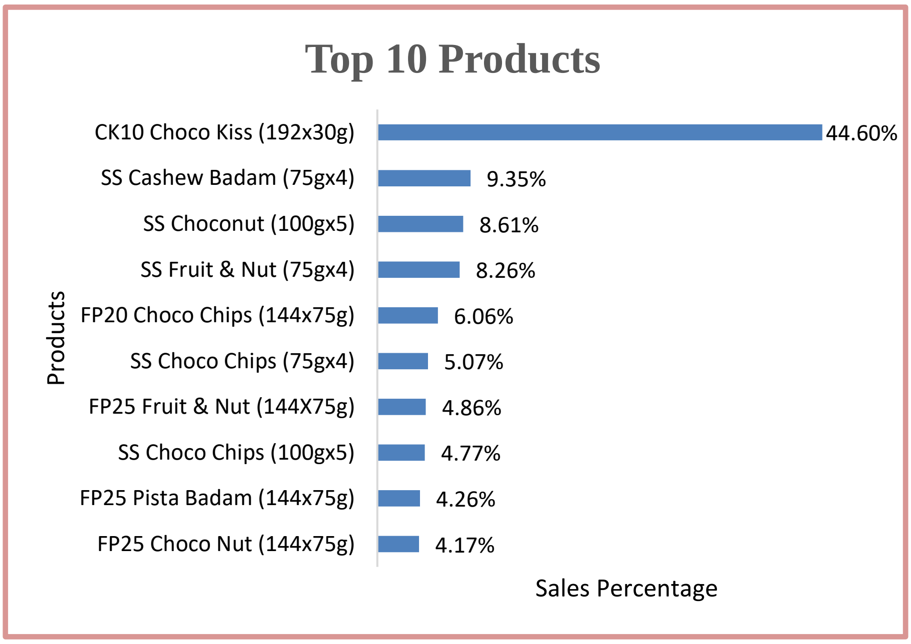


Interpretation :
Choco Kiss (57%) is also the highest sold product in Institutional Trade followed by Choco chips (6.8%) and Cashew Badam (6%).

3) Overall Monthwise Sales of channels

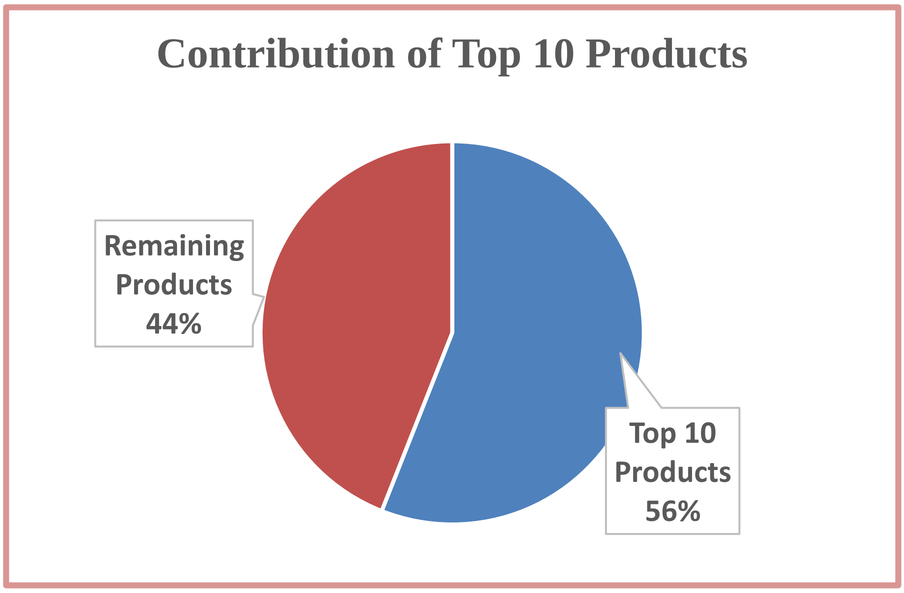


4) Sales of Top 10 Products



Interpretation :
Choco Kiss accounts for 44.6% of total sales followed by Cashew Badam (9%) and Choconut (8.6%).

5)

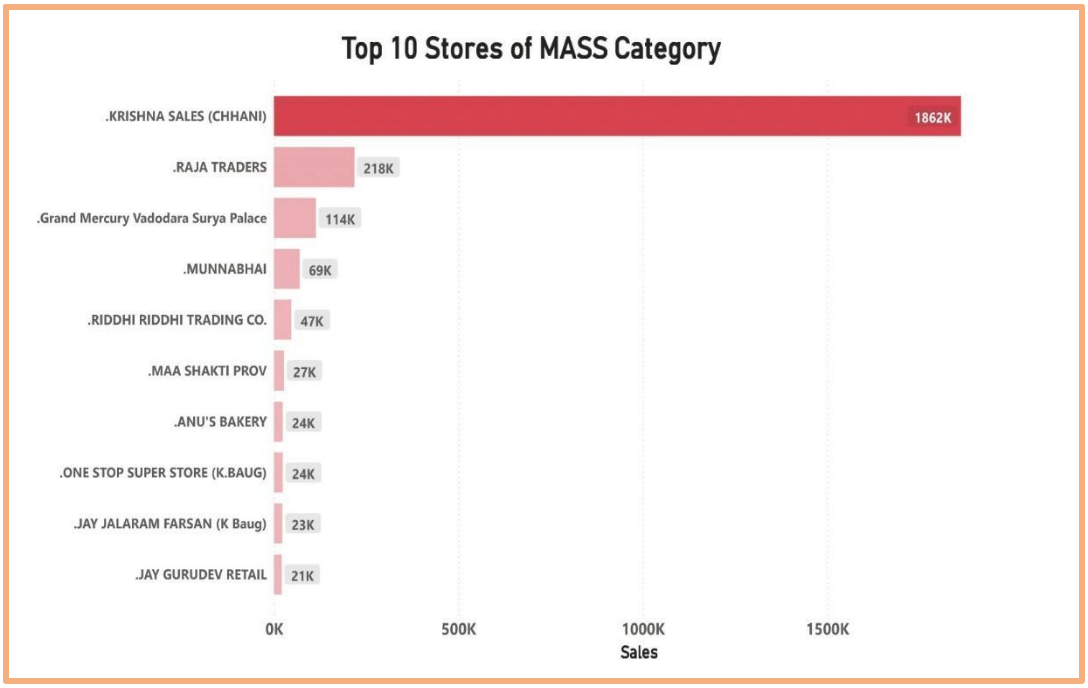


Interpretation :

Top 10 products contribute 56% of total Sales while the remaining products account for 44% of total Sales.

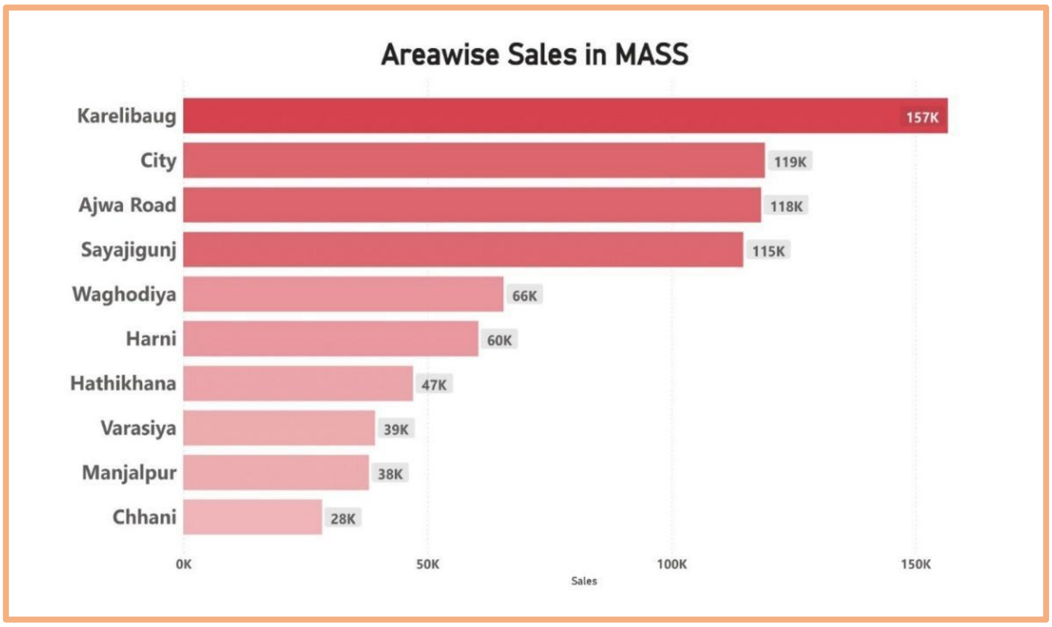
Mass Category

1)



Interpretation :
Krishna Sales is the highest selling store of products of MASS category.

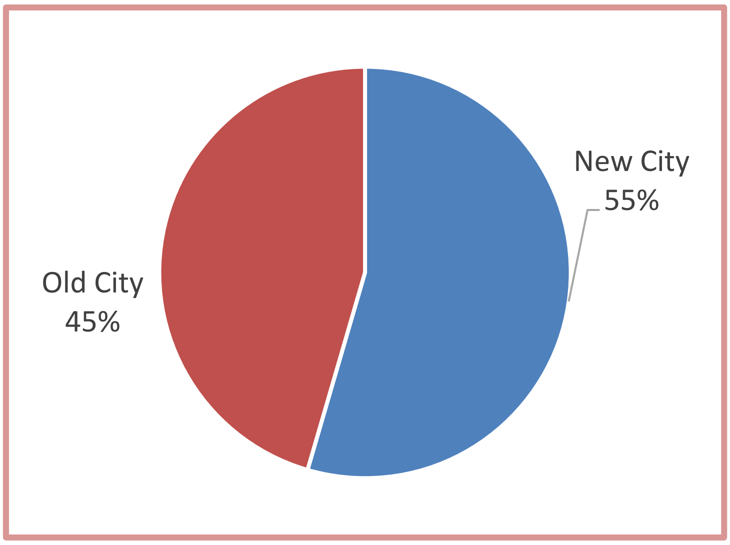
2)



Interpretation :
Karelibaug is the location with highest sales of products of MASS category followed by City area and Ajwa Road.

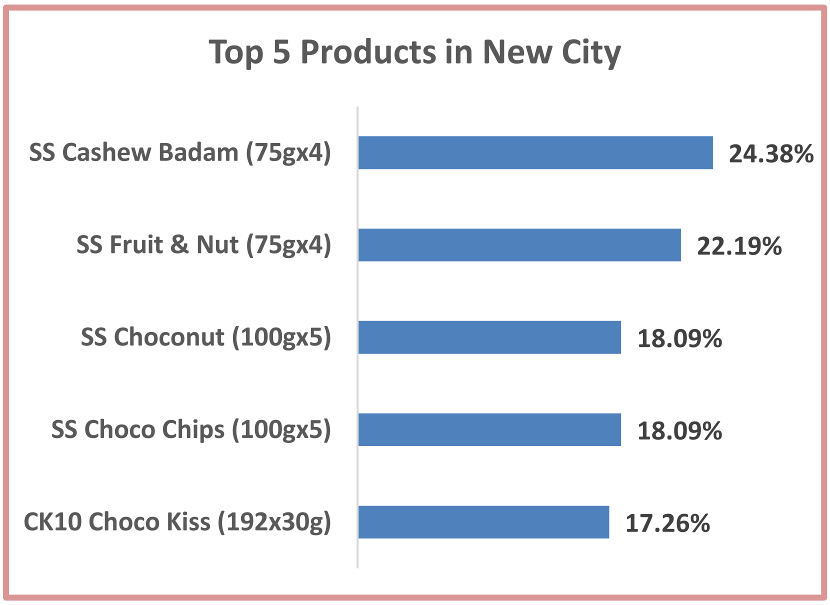
Routewise Sales

1)



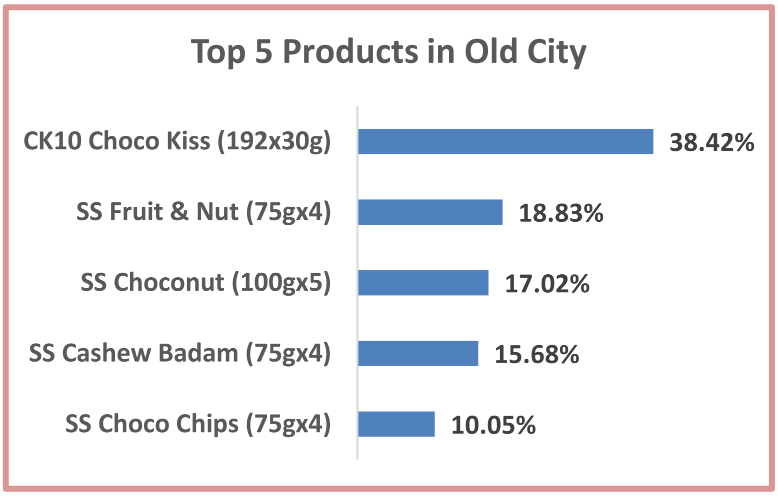
Interpretation :
The route of New city accounts for 55% of total sales while Old city route accounts for 45% of total sales.

2) Route - Sales of Top Five Product



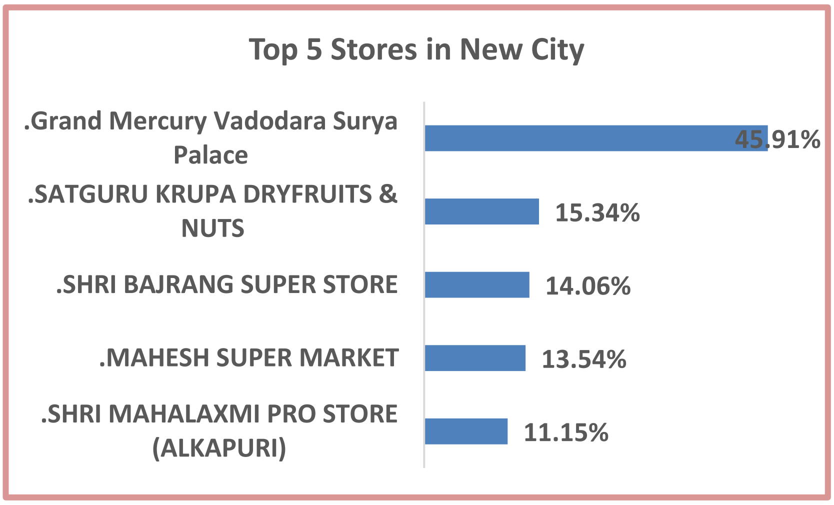
Interpretation :
Cashew Badam (24%) is the highest selling product in New City Route followed by Fruit & Nut (22%) and Choconut (18%).

3)



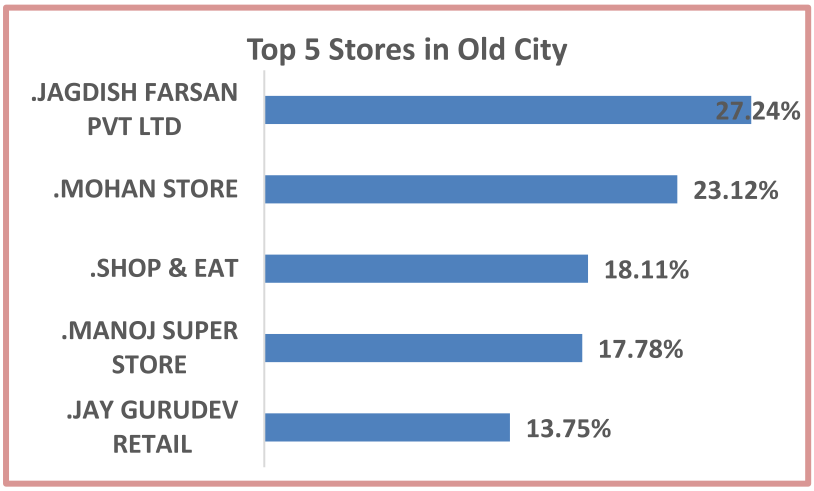
Interpretation :
Choco Kiss (38.42%) is the highest selling product in Old City Route followed by Fruit & Nut (18.8%) and Choconut (17%)

4) Route - Sales of Top Five Stores



Interpretation :
Grand Surya Palace (46%) is the highest selling store in New city Route followed by SATGURU KRUPA (15.3%) and SHRI BAJRANG SUPER STORE (14%).

5)

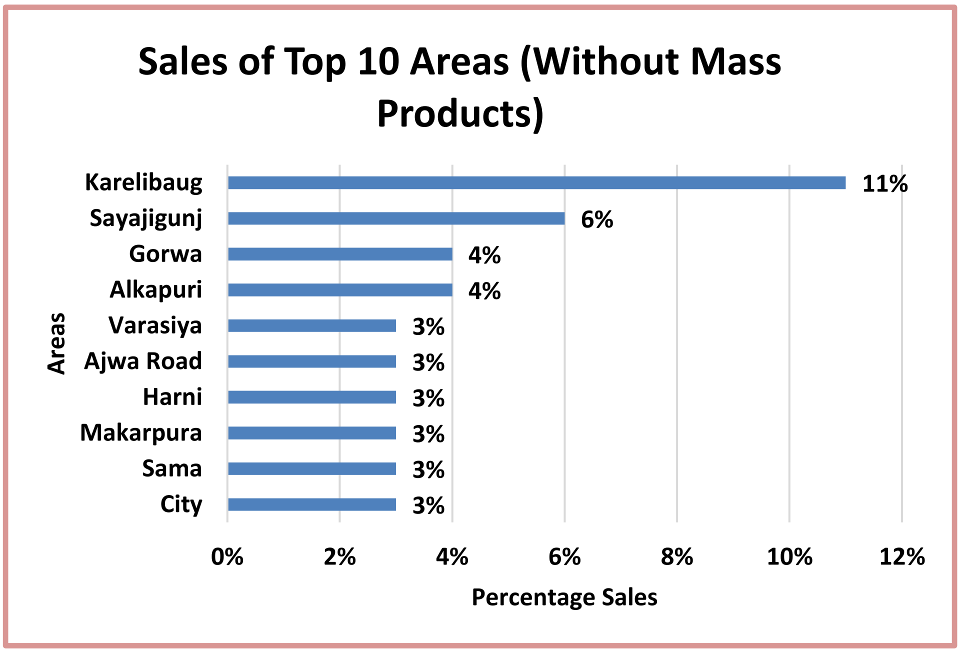


Interpretation:

JAGDISH FARSAN PVT LTD (27%) is the highest selling store in Old city Route followed by MOHAN STORE (23%) and SHOP & EAT (18%).

Areawise

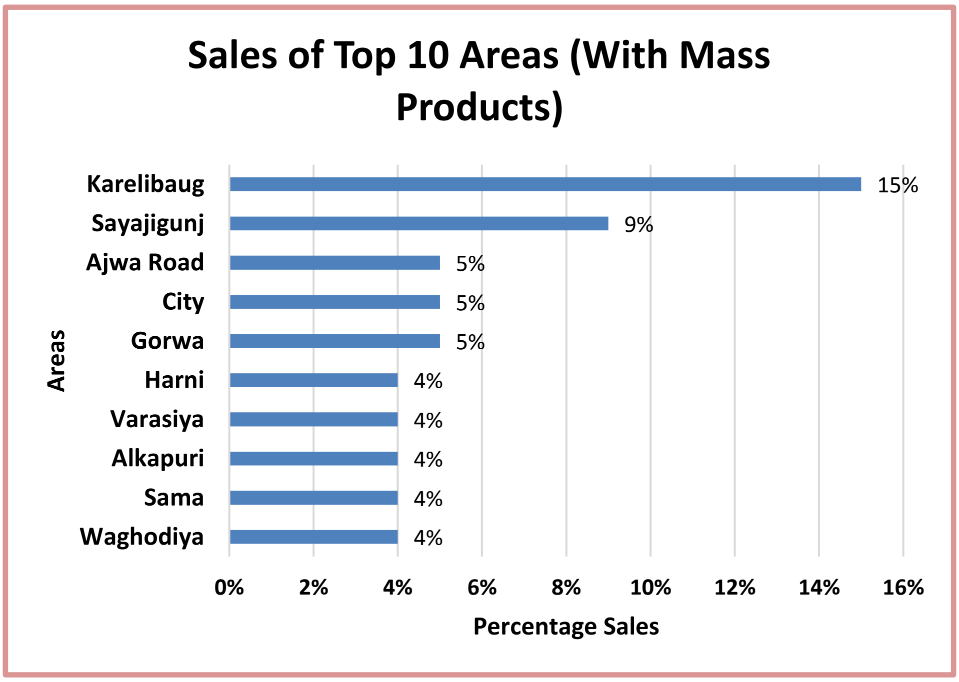
1)



Interpretation :

Karelibaug (11%) is the area with highest sales (without MASS products) followed by Sayajigunj (6%) and Gorwa (4%).

2)



Interpretation :

Karelibaug (15%) is the area with highest sales (with MASS products) followed by Sayajigunj (9%) and Ajwa Road (5%).

UNPAIRED TWO SAMPLE T-TEST IN R

1) Significance difference between Sales in New city and Old city routes

	Route	Sales
1	New City	210945
2	New City	215320
3	New City	175680
4	New City	245245
5	New City	242035
6	New City	206690
7	New City	326835
8	New City	253590
9	New City	395760
10	New City	472455
11	New City	128155
12	New City	366145
13	Old City	121355
14	Old City	307815
15	Old City	207185
16	Old City	129045
17	Old City	211390
18	Old City	223985
19	Old City	242960
20	Old City	314705
21	Old City	307840
22	Old City	356130
23	Old City	131900
24	Old City	159035

Firstly, we need to check our data is normally distributed or not.

Hypothesis : - :
H0 :- Data is normally distributed vs.
H1 :- Data is not normally distributed.

# Shapiro-Wilk normality test for New city's Sales
with(data, shapiro.test(Sales[Route == "New City"]))

```
        Shapiro-Wilk normality test

data:  Sales[Route == "New City"]
W = 0.93366, p-value = 0.4205
```

# Shapiro-Wilk normality test for Old city's Sales
with(data, shapiro.test(Sales[Route == "Old City"]))

```
        Shapiro-Wilk normality test

data:  Sales[Route == "Old City"]
W = 0.92116, p-value = 0.2957
```

<u>Interpretation</u>:
From the output, the two p-values are greater than the significance level 0.05, implying that the distribution of the data is normally distributed.
In other words, we can assume the normality.

Now to check if variances are equal or not.

Using F-test to test for homogeneity in variances

Hypothesis :
H0 : there is no significant difference between the variances of the two sets of data.
Vs.
H1 : there is significant difference between the variances of the two sets of data.

```
res.ftest <- var.test(Sales ~ Route, data = data)
res.ftest
```

```
        F test to compare two variances

data:  Sales by Route
F = 1.5183, num df = 11, denom df = 11, p-value = 0.5
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4370768 5.2740247
sample estimates:
ratio of variances
          1.518273
```

Interpretation :

The p-value of F-test is p = 0.1713596 which is greater than the significance level alpha = 0.05.

Hence we can conclude that, there is no significant difference between the variances of the two sets of data.

Therefore, we can use the classic t-test which assume equality of the two variances.

```
# Compute t-test
res <- t.test(Sales ~ Route, data = data, var.equal = TRUE)
res
```

```
        Two Sample t-test

data:  Sales by Route
t = 1.1755, df = 22, p-value = 0.2523
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -33465.2 121050.2
sample estimates:
mean in group New City mean in group Old City
              269904.6               226112.1
```

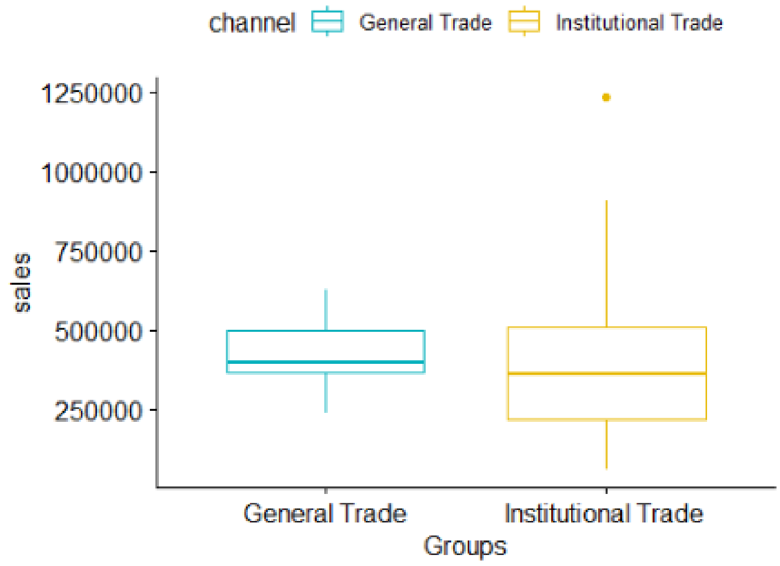
Interpretation :

The p-value of the test is 0.2523, which is greater than the significance level alpha = 0.05.

We can conclude that average sales through New city's route is not significantly different from Sales through old city's route at 5% level of confidence.

2) Significance difference between Institutional Trade and General Trade Sales

	channel	sales
1	General Trade	272020
2	General Trade	389990
3	General Trade	345700
4	General Trade	370250
5	General Trade	406235
6	General Trade	391800
7	General Trade	512645
8	General Trade	495445
9	General Trade	601175
10	General Trade	627400
11	General Trade	237645
12	General Trade	429710
13	Institutional Trade	217130
14	Institutional Trade	457285
15	Institutional Trade	905470
16	Institutional Trade	60290
17	Institutional Trade	1234020
18	Institutional Trade	214995
19	Institutional Trade	388480
20	Institutional Trade	335065
21	Institutional Trade	428320
22	Institutional Trade	665265
23	Institutional Trade	163950
24	Institutional Trade	292435



Firstly, we need to check our data is normally distributed or not.

Hypothesis : - :
H0 :- Data is normally distributed
vs.
H1 :- Data is not normally distributed.

# Shapiro-Wilk normality test for Institutional Trade 's Sales
with(data, shapiro.test(sales[channel == "Institutional Trade"]))

```
        Shapiro-Wilk normality test

data:  sales[channel == "Institutional Trade"]
W = 0.87442, p-value = 0.07437
```

# Shapiro-Wilk normality test for General Trade 's Sales
with(data, shapiro.test(sales[channel == "General Trade"]))

```
        Shapiro-Wilk normality test

data:  sales[channel == "General Trade"]
W = 0.96054, p-value = 0.7915
```

Interpretation:

From the output, the two p-values are greater than the significance level 0.05, implying that the distribution of the data is normally distributed.

In other words, we can assume the normality.

Now to check if variances are equal or not.

Using F-test to test for homogeneity in variances

Hypothesis :
H0 : there is no significant difference between the variances of the two sets of data.
Vs.
H1 : there is significant difference between the variances of the two sets of data.

```
res.ftest <- var.test(sales ~ channel, data = data)
res.ftest
```

```
        F test to compare two variances

data:  sales by channel
F = 0.1237, num df = 11, denom df = 11, p-value = 0.001667
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.0356115 0.4297092
sample estimates:
ratio of variances
         0.1237036
```

Interpretation :

The p-value of F-test is p = 0.0016 which is less than the significance level alpha = 0.05.

Hence we conclude that there is significant difference between the variances of the two sets of data.

Therefore, we can use the welch t-test which assume inequality of the two variances.

Compute t-test

```
res <- t.test(sales ~ channel, data = data, var.equal = FALSE)
res
```

```
        Welch Two Sample t-test

data:  sales by channel
t = -0.22802, df = 13.68, p-value = 0.823
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -245632.9  198517.9
sample estimates:
      mean in group General Trade mean in group Institutional Trade
                        423334.6                          446892.1
```

Interpretation :

The p-value of the test is 0.823, which is greater than the significance level alpha = 0.05.

We can conclude that Institutional Trade's Sales is not significantly different from General Trade's Sales with 5% level of confidence.

Linear Mixed Effects models

Linear Mixed Effects models are used for regression analyses involving dependent data. Such data arise when working with longitudinal and other study designs in which multiple observations are made on each subject. Some specific linear mixed effects models are

- *Random intercepts models*, where all responses in a group are additively shifted by a value that is specific to the group.
- *Random slopes models*, where the responses in a group follow a (conditional) mean trajectory that is linear in the observed covariates, with the slopes (and possibly intercepts) varying by group.
- *Variance components models*, where the levels of one or more categorical covariates are associated with draws from distributions. These random terms additively determine the conditional mean of each observation based on its covariate values.

The statsmodels implementation of LME is primarily group-based, meaning that random effects must be independently-realized for responses in different groups. There are two types of random effects in our implementation of mixed models: (i) random coefficients (possibly vectors) that have an unknown covariance matrix, and (ii) random coefficients that are independent draws from a common univariate distribution. For both (i) and (ii), the random effects influence the conditional mean of a group through their matrix/vector product with a group-specific design matrix.

A simple example of random coefficients, as in (i) above, is:

$$Yij = \beta 0 + \beta 1Xij + \gamma 0i + \gamma 1iXij + \varepsilon ij$$

Here, Yij is the jth measured response for subject i, and Xij is a covariate for this response. The “fixed effects parameters” β0 and β1 are shared by all subjects, and the errors ϵij are independent of everything else, and identically distributed (with mean zero). The “random effects parameters” γ0i and γ1i follow a bivariate distribution with mean zero, described by three parameters: var(γ0i), var(γ1i), and cov(γ0i,γ1i).

```
import pandas as pd

#Importing the data
data=pd.read_excel('GLM Data.xlsx')
data.head()
```

	Year	Month	Store Names	Area	Route	Sales	Festival	Temperature	Rainfall	ward	Population
0	2019	July	.MOHAN STORE	Varasiya	Old City	1120	0	29.773226	8.700000	ward 2	4.801369e+07
1	2019	August	.MOHAN STORE	Varasiya	Old City	1740	1	27.210968	15.937742	ward 2	4.801369e+07
2	2019	September	.APEX DRY FRUIT STORE (O P RD)	OP Road	New City	1860	0	27.266000	10.712000	ward 6	5.113295e+07
3	2019	September	.MAHESH SUPER MARKET	Akota	New City	11640	0	27.266000	10.712000	ward 6	5.113295e+07
4	2019	September	.MOHAN STORE	Varasiya	Old City	4470	0	27.266000	10.712000	ward 2	4.801369e+07

```
import statsmodels.api as sm
import statsmodels.formula.api as smf

#Defining the model
md = smf.mixedlm("Sales ~Area+Festival+Temperature+Rainfall+Population ", data, groups=data["Store Names"])
mdf = md.fit()
print(mdf.summary())
```

```
                Mixed Linear Model Regression Results
==============================================================================
=
Model:              MixedLM      Dependent Variable:      Sales
No. Observations:      1591        Method:                 REML
No. Groups:           394         Scale:              10924083.9900
Min. group size:        1         Log-Likelihood:        -14927.5055
Max. group size:       25          Converged:            Yes
Mean group size:       4.0
-----------------------------------------------------------------------------
                    Coef.    Std.Err.   z    P>|z|   [0.025     0.975]
-----------------------------------------------------------------------------
Intercept              3555.891  941.118  3.778 0.000   1711.334  5400.448
Area[T.Akota]           -1648.174 1329.557 -1.240 0.215  -4254.058   957.711
Area[T.Alkapuri]         -623.356 1145.756 -0.544 0.586  -2868.997  1622.285
Area[T.Atladra]         -3230.570 1246.240 -2.592 0.010  -5673.156  -787.984
Area[T.Bapod]           -4831.625 4014.421 -1.204 0.229 -12699.745  3036.495
Area[T.Bhayli]          -3803.349 2474.194 -1.537 0.124  -8652.681  1045.983
Area[T.Chhani]          -2868.484 1333.316 -2.151 0.031  -5481.736  -255.232
Area[T.Chokhandi]       -3785.298 2675.257 -1.415 0.157  -9028.704  1458.109
Area[T.ELLORA PARK]     -1754.674 1868.152 -0.939 0.348  -5416.186  1906.837
Area[T.Ellora Park]     -1296.899 3982.315 -0.326 0.745  -9102.092  6508.295
Area[T.Fatehgunj]         825.126 1251.811  0.659 0.510  -1628.378  3278.631
Area[T.GOTRI]            -182.293 1697.861 -0.107 0.914  -3510.040  3145.454
Area[T.Gorwa]           -1508.257  955.751 -1.578 0.115  -3381.494   364.981
Area[T.Gotri]           -1991.000  840.227 -2.370 0.018  -3637.814  -344.186
Area[T.Harni]           -1587.739  917.549 -1.730 0.084  -3386.103   210.625
Area[T.Hathikhana]      19895.584 3229.540  6.161 0.000  13565.802 26225.366
Area[T.Jetalpur Road]   -2415.711 2786.567 -0.867 0.386  -7877.281  3045.859
Area[T.Karelibaug]      -2038.147  812.527 -2.508 0.012  -3630.670  -445.624
Area[T.Khodiyar Nagar]   -641.200 2402.652 -0.267 0.790  -5350.311  4067.911
Area[T.Makarpura]        -572.777 1285.811 -0.445 0.656  -3092.920  1947.366
Area[T.Maneja]           1872.886 1491.404  1.256 0.209  -1050.212  4795.985
Area[T.Manjalpur]       -1508.281  948.540 -1.590 0.112  -3367.385   350.823
Area[T.Navapura]        -3558.644 1801.371 -1.976 0.048  -7089.265   -28.022
Area[T.Nizampura]       -1891.877 1072.731 -1.764 0.078  -3994.392   210.637
```

Area[T.OP Road]	-2101.156	1239.567	-1.695	0.090	-4530.663	328.351
Area[T.Panigate]	-1161.280	1553.028	-0.748	0.455	-4205.160	1882.600
Area[T.Pratapnagar]	184.381	1591.261	0.116	0.908	-2934.433	3303.195
Area[T.RC Dutt Road]	-3340.693	3989.110	-0.837	0.402	-11159.206	4477.819
Area[T.Raopura]	-3114.708	1747.915	-1.782	0.075	-6540.558	311.141
Area[T.SAMA]	-3786.749	4024.648	-0.941	0.347	-11674.914	4101.416
Area[T.Salatwada]	-2606.831	3998.460	-0.652	0.514	-10443.668	5230.006
Area[T.Sama]	-2608.844	982.470	-2.655	0.008	-4534.450	-683.238
Area[T.Sayajigunj]	-2847.307	2533.171	-1.124	0.261	-7812.232	2117.617
Area[T.Subhanpura]	-1729.334	1392.793	-1.242	0.214	-4459.157	1000.490
Area[T.Sunpharma]	-1985.963	2544.182	-0.781	0.435	-6972.468	3000.541
Area[T.Tandalja]	-4088.162	4002.852	-1.021	0.307	-11933.608	3757.285
Area[T.Tarsali]	-1911.037	1099.948	-1.737	0.082	-4066.895	244.822
Area[T.VIP Road]	-1610.813	1921.609	-0.838	0.402	-5377.097	2155.472
Area[T.Vadsar]	-1883.353	973.030	-1.936	0.053	-3790.456	23.751
Area[T.Varasiya]	-2196.049	913.305	-2.405	0.016	-3986.094	-406.005
Area[T.Vasna]	-2778.168	1023.644	-2.714	0.007	-4784.474	-771.862
Area[T.Vasna Bhayli]	-3437.437	1456.020	-2.361	0.018	-6291.183	-583.690
Area[T.Vemali]	-3331.906	1267.095	-2.630	0.009	-5815.366	-848.446
Area[T.Waghodiya]	-1466.389	833.565	-1.759	0.079	-3100.146	167.369
Festival	46.199	187.713	0.246	0.806	-321.712	414.110
Temperature	78.382	24.323	3.223	0.001	30.709	126.054
Rainfall	-30.457	21.018	-1.449	0.147	-71.652	10.737
Population	-0.000	0.000	-4.585	0.000	-0.000	-0.000
Group Var	4615499.007	183.418				

===
=

RFM ANALYSIS

RFM Analysis History

RFM Model was introduced by Hughes in 1994 for customer value analysis and effective customer segmentation. This model has been used for more than 30 years now and still remains a useful method for optimizing sales and building campaigns to engage customers.
The simplicity and grounded analysis of RFM Model makes it a worthy analytical method for direct marketing. RFM Model Analytics primarily assists in effective Customer Segmentation.

Importance of RFM Analysis

RFM is the abbreviated form of Recency, Frequency, and Monetary Value. In which, Recency (R) refers to the number of days or months since the last purchase was made by a customer. Frequency (F) denotes the number of purchases in a certain time period. While Monetary Value (M) refers to the total amount of money spent by a customer during a specific period of time.

Using RFM indices helps in the formation of customer segments. Customer segmentation is the process of identifying a group of customers who share similar characteristics. By creating customer segments, a store provides customized product promotions to those who are interested in them.

RFM analysis works upon the marketing axiom (The Pareto Principle) (80:20 Rule) that “80% of a company’s business comes from 20% of its customers”.

When is the Best Time to Try RFM Analysis?

- You’re selling fast-moving consumer goods in a market filled with competitors.
- You’ve launched your loyalty program and your customer base is growing fast.
- You know your audience is diverse, and you’re looking for an easy way to segment it.
- You’ve never personalized your offers in your emails/advertising/promo materials, but you want to.
- The effectiveness of your advertising campaigns is low because you don’t have proper targeting.
- You’re ready to build a remarketing strategy, but you don’t know where to start.

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K distinct clusters. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as far as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

Elbow Curve:

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.

The **customers** are **different Stores** to which the UNIBIC products are distributed by Prem Agencies.

For the RFM analysis, the variables Recency, Frequency, and Monetary value are defined as:-

Recency: When was the last purchase of a store?

Number of months since the last purchase

Frequency: How frequently does the store purchase from the distributor?

Frequency of purchase of a store, i.e., how many times in the period of 31 months the store has purchased products.

Monetary Value: How much has the store spent?

Total sales of a store

Steps followed:

Step 1: Importing necessary libraries

```
In [1]: # Importing Libraries

        %matplotlib inline
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
```

The basic libraries NumPy, Pandas, and matplotlib (for visualization) used are imported for the analysis.

Step 2: Importing data from CSV file

The data Includes Store wise sales per month.

CSV File:

	A	B	C	D	E
1	Year	Month	M	Store Names	Sales
2	2019	July	31-07-2019	.MOHAN STORE	1120
3	2019	July	31-07-2019	.RAJA TRADERS	790
4	2019	July	31-07-2019	XGANESH STORE	460
5	2019	July	31-07-2019	XPRAHLADBHAI - VARSIA	495
6	2019	August	31-08-2019	.MOHAN STORE	1740
7	2019	August	31-08-2019	.RAJA TRADERS	245
8	2019	Septembe	30-09-2019	.APEX DRY FRUIT STORE (O P RD)	1860
9	2019	Septembe	30-09-2019	.JAGDISH FARSAN PVT LTD	32840

The columns included in the CSV data file are:

1. Year: Year of purchase
2. Month: Month of purchase
3. M: Last day of the Month and year of purchase
   (As daily data was not available, we had only the month and year for every purchase. For the purpose of calculating Recency, the last day of a particular month is considered.)
4. Store Names: (Customers of the distributor i.e., Prem Agencies)
5. Sales: Monthly Sales of the store

This CSV file is imported to python and stored in a Data frame object named '**Stores**'

```
In [2]: # Importing Dataset

        Stores = pd.read_csv("D:\\Srushti\\MSc\\Sem 4\\Project\\Analysis\\Stores.csv" )
        Stores.head()
```

Out[2]:

	Year	Month	M	Store Names	Sales
0	2019	July	31-07-2019	.MOHAN STORE	1120
1	2019	July	31-07-2019	.RAJA TRADERS	790
2	2019	July	31-07-2019	XGANESH STORE	460
3	2019	July	31-07-2019	XPRAHLADBHAI - VARSIA	495
4	2019	August	31-08-2019	.MOHAN STORE	1740

```
In [3]: Stores.shape

Out[3]: (1947, 5)

In [4]: Stores.isnull().sum(axis=0)

Out[4]: Year           0
        Month          0
        M              0
        Store Names    0
        Sales          0
        dtype: int64
```

Output 4 shows that null values are not there in the data.

The column needs to be converted to a datetime format for the ease of performing operations on dates.

```
In [5]: Stores['M'] = pd.to_datetime(Stores['M'])
        Stores.head()
```

Out[5]:

	Year	Month	M	Store Names	Sales
0	2019	July	2019-07-31	.MOHAN STORE	1120
1	2019	July	2019-07-31	.RAJA TRADERS	790
2	2019	July	2019-07-31	XGANESH STORE	460
3	2019	July	2019-07-31	XPRAHLADBHAI - VARSIA	495
4	2019	August	2019-08-31	.MOHAN STORE	1740

Step 3: Calculating Recency, Frequency, and Monetary value for all the stores.

1. As mentioned earlier, the Recency of a store is considered as the no. of months since the last purchase.
   To calculate this, we considered the last day of February 2022 as the latest date (i.e., 28/02/2022), since we had the data up to January 2022.
   e.g. - Suppose we have a store for which the last purchase was in Dec 2021, then the no. of months from the last purchase should be 2.

   So, to calculate the Recency in this way, a function month is defined, where the arguments are two dates. This function calculates the number of months between the provided two dates.

```
In [6]: import datetime as dt
        from dateutil import relativedelta
        latest_date = dt.datetime(2022, 2, 28)
        def months(ld, sd):
            a = relativedelta.relativedelta(ld,sd)
            m = a.months + (a.years * 12)
            return m
```

2. Using the 'Stores' data frame, a new data frame **'RFMScores'** including store-wise Recency, Frequency, and Monetary value is created. For this, the lambda function is used.

   Recency is calculated using the month function defined.

   Frequency is calculated using the count of how many times a particular store is repeated in the Store Names column of the Stores data frame.

   Monetary value is calculated using the sum of sales of the corresponding store.

```
In [7]: RFMScores = Stores.groupby('Store Names').agg({'M' : lambda x: months(latest_date, x.max()),
                                                    'Store Names' : lambda x: len(x),
                                                    'Sales' : lambda x: x.sum() })
        RFMScores['M'] = RFMScores['M'].astype(int)
        RFMScores.rename(columns = {'M' : 'Recency',
                                   'Store Names': 'Frequency',
                                   'Sales': 'Monetary'} , inplace = True )
        RFMScores.reset_index().head()
```

Out[7]:

	Store Names	Recency	Frequency	Monetary
0	.24x7 CAFETERIA	4	14	76415
1	.A B SUPER STORE	6	4	2280
2	.AARDEEP STORE	6	1	510
3	.ADINATH GRAIN AND PROV STORE	6	1	120
4	.AGRAWAL FARM FRESH	4	2	4100

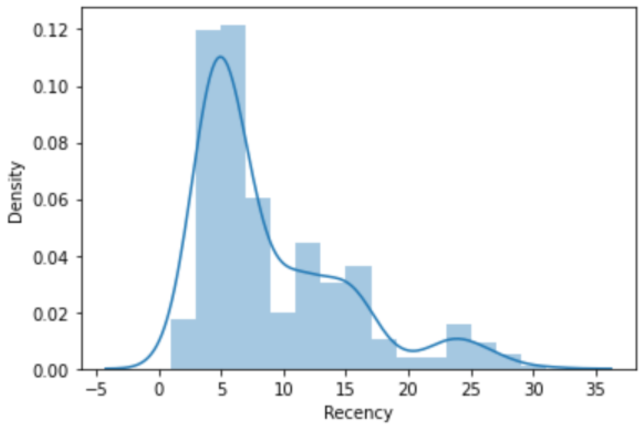
Descriptive Statistics of Recency, Frequency, and Monetary value are obtained to get a better idea of these three variables.

a. Recency-

```
In [8]: RFMScores.Recency.describe()

Out[8]: count    481.000000
        mean       8.835759
        std        5.984011
        min        1.000000
        25%        4.000000
        50%        6.000000
        75%       12.000000
        max       31.000000
        Name: Recency, dtype: float64
```

```
In [9]: import seaborn as sns
        x = RFMScores['Recency']
        ax1 = sns.distplot(x)
```

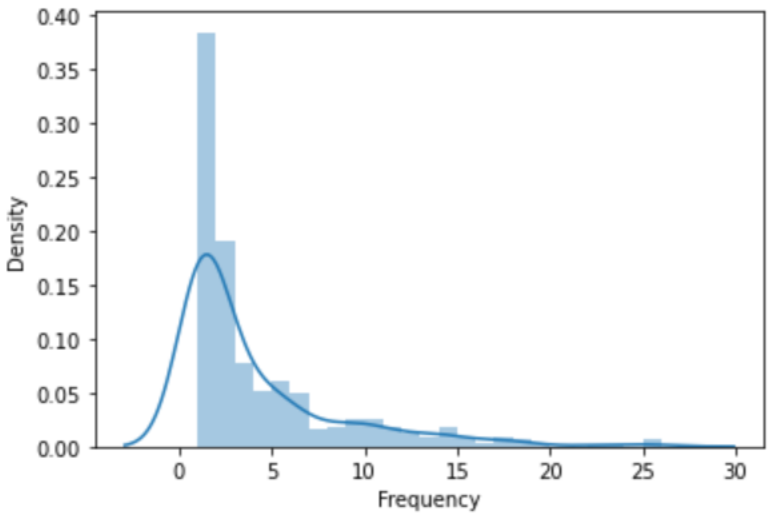


b. Frequency-

In [10]:
```
RFMScores.Frequency.describe()
```

Out[10]:
```
count    481.000000
mean       4.047817
std        4.501828
min        1.000000
25%        1.000000
50%        2.000000
75%        5.000000
max       26.000000
Name: Frequency, dtype: float64
```

In [11]:
```
y = RFMScores['Frequency']
ax2 = sns.distplot(y)
```

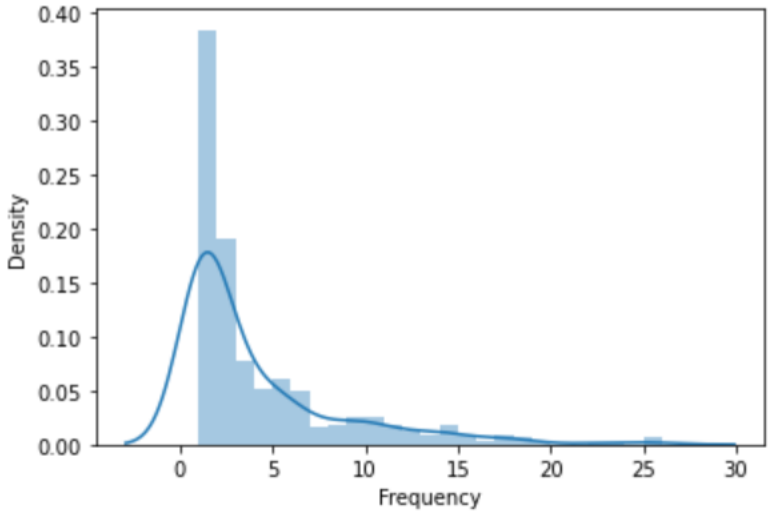


c. Monetary Value-

```
In [10]: RFMScores.Frequency.describe()

Out[10]: count    481.000000
         mean       4.047817
         std        4.501828
         min        1.000000
         25%        1.000000
         50%        2.000000
         75%        5.000000
         max       26.000000
         Name: Frequency, dtype: float64
```

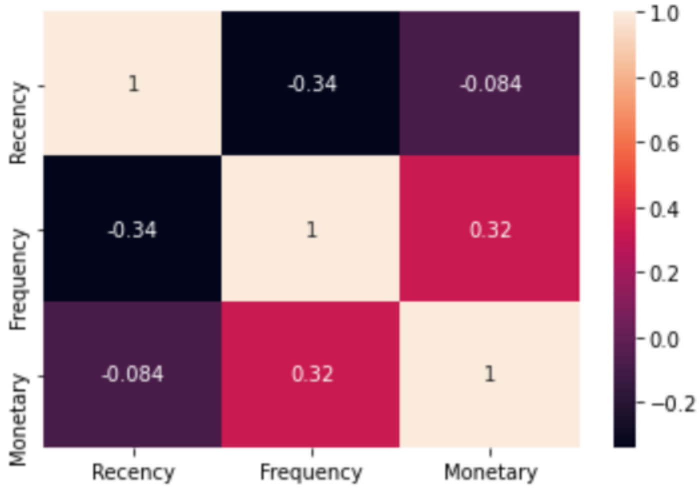
```
In [11]: y = RFMScores['Frequency']
         ax2 = sns.distplot(y)
```



3. Heatmap for Recency, Frequency, and Monetary Value is plotted using the seaborn module.

```
In [14]: sns.heatmap(RFMScores.corr(), annot= True)

Out[14]: <AxesSubplot:>
```



From this heatmap,

a. Weak negative (-0.34) correlation between Recency and frequency. (Higher the recency, the lower the frequency.
b. Weak negative correlation (-0.084) between Recency and monetary value. (Higher the recency, the lower the monetary value.)
c. Weak positive correlation between frequency and monetary value. (Higher the frequency, the higher the monetary value.

Step 3: Calculating the Recency (R), Frequency (F), and Monetary Value (M) score

For each and every store the R, F, and M< scores are calculated using quantiles. Here, the scoring is on a scale of 1-5. Thus, 4 quantiles are considered for scoring.

```
In [15]: #Split into five segments using quantiles
         quantiles = RFMScores.quantile(q=[0.2, 0.4, 0.6, 0.8])
         quantiles = quantiles.to_dict()
         quantiles

Out[15]: {'Recency': {0.2: 4.0, 0.4: 6.0, 0.6: 8.0, 0.8: 14.0},
          'Frequency': {0.2: 1.0, 0.4: 2.0, 0.6: 3.0, 0.8: 6.0},
          'Monetary': {0.2: 600.0, 0.4: 1920.0, 0.6: 4380.0, 0.8: 14870.0}}
```

Using these quantiles, two functions RScore and FnMScore are defined to get the R, F and M scores.

Lower the Recency i.e., the lesser the no. of months since last purchase, the higher the R score.

Lower the frequency, higher the monetary value.

```
In [16]: #Functions to create R, F and M segments

         def RScore(x, p, d):
             if x <= d[p][0.2]:
                 return 5
             elif x <= d[p][0.4]:
                 return 4
             elif x <= d[p][0.6]:
                 return 3
             elif x <= d[p][0.8]:
                 return 2
             else:
                 return 1

         def FnMScore(x,p,d):
             if x <= d[p][0.2]:
                 return 1
             elif x <= d[p][0.4]:
                 return 2
             elif x <= d[p][0.6]:
                 return 3
             elif x <= d[p][0.8]:
                 return 4
             else:
                 return 5
```

In [17]:
```
RFMScores['R'] = RFMScores['Recency'].apply(RScore, args = ('Recency', quantiles,))
RFMScores['F'] = RFMScores['Frequency'].apply(FnMScore, args = ('Frequency', quantiles,))
RFMScores['M'] = RFMScores['Monetary'].apply(FnMScore, args = ('Monetary', quantiles,))
RFMScores.head()
```

Out[17]:

Store Names	Recency	Frequency	Monetary	R	F	M
.24x7 CAFETERIA	4	14	76415	5	5	5
.A B SUPER STORE	6	4	2280	4	4	3
.AARDEEP STORE	6	1	510	4	1	1
.ADINATH GRAIN AND PROV STORE	6	1	120	4	1	1
.AGRAWAL FARM FRESH	4	2	4100	5	2	3

These three scores R, F, and M are further grouped in the RFMGroup, and the sum of these score is counted in the RFMScore column.

In [18]:
```
# Calculating RFMGroup value column showing combined concatenated score of RFM
RFMScores['RFMGroup'] = RFMScores.R.map(str) + RFMScores.F.map(str) + RFMScores.F.map(str)

# Calculating RFMScore value column showing total sum of RFMGroup values
RFMScores['RFMScore'] = RFMScores[['R', 'F', 'M']].sum(axis=1)
RFMScores.head()
```

Out[18]:

Store Names	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore
.24x7 CAFETERIA	4	14	76415	5	5	5	555	15
.A B SUPER STORE	6	4	2280	4	4	3	444	11
.AARDEEP STORE	6	1	510	4	1	1	411	6
.ADINATH GRAIN AND PROV STORE	6	1	120	4	1	1	411	6
.AGRAWAL FARM FRESH	4	2	4100	5	2	3	522	10

We can group the customers based on the above 3 Factors(RFM).

- **Best customers**: High 'R', High 'F' and High 'M'

  Customers who bought most recently, most often, and are heavy spenders.

- **At-Risk Customers**: Low 'R', 'High F' and High 'M'

  Customers who purchased often and spent big amounts, but haven't purchased recently.

- **Lost Customers:** Low 'R' , Low 'R' and Low 'M'

  Customers with Low frequency and spending amount. Not placing an order recently.

The Data Frame RFMScores is exported in Excel file to proceed with further analysis.

```
In [24]: RFMScores.to_excel("RFM.xls")
```

To understand the grouping in a better way Clustering is performed on these RFM scores.

K MEANS CLUSTERING

RFM scoring + K-Means Clustering analysis is performed to get a better idea of how the customers are segmented.

The Clustering was performed in R Studio.

Step 1: Checking and preprocessing data:

As mentioned earlier, the RFM scores i.e., 'R' , 'F', and 'M' scores are to be considered as the feature variables for clustering. So these scores calculated in RFM excel file are stored in a CSV file RFMAnalysis which is imported in R for further analysis.

Step 2: Importing required packages:

The packages ggplot2, factoextra, NbClust, and cluster are installed and called using library() function

.

Step 3: Importing Dataset

The CSV file is imported and stored in a data frame object rfm.

```
rfm <- read.csv("D:\\Srushti\\MSc\\Sem 4\\Project\\Analysis\\RFMAnalysis.csv")
str(rfm)
```

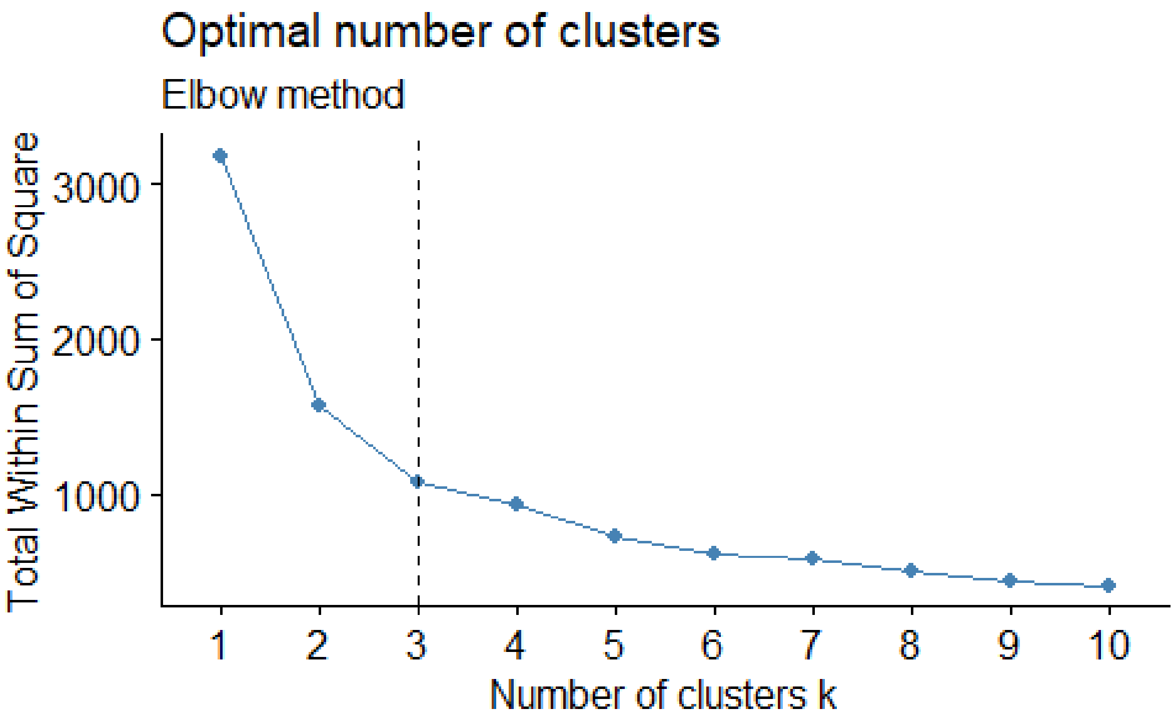
```
## 'data.frame':    481 obs. of  3 variables:
##  $ R: int  5 4 4 4 5 1 1 3 2 5 ...
##  $ F: int  5 4 1 1 2 1 3 2 5 1 ...
##  $ M: int  5 3 1 1 3 1 3 1 5 2 ...
```

Step 4: Selecting the optimal value of k

The k for k Means clustering is calculated using elbow method.

```
fviz_nbclust(rfm, kmeans, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2)+
  labs(subtitle = "Elbow method")
```

fviz_nbclust(): Determines and visualize the optimal number of clusters using different methods: within cluster sums of squares (WSS), average silhouette and gap statistics.



From this plot, Elbow is obtained at k=3, where the decrement in Total Within Sum of Squares after k=3 is insignificant, it does not worth to further complicate the model.
So, we select k=3

Step 5: Fitting the clustering model:

```
kmf = kmeans(rfm,3)
```

```
kmf$size
```

```
## [1] 247  74 160
```

```
kmf$cluster
```

```
##   [1] 3 3 1 1 3 1 2 1 2 1 1 3 3 1 1 3 3 3 3 3 3 1 1 1 3 3 1 1 3 1 1 3 1 3 1 3 3
##  [38] 3 1 3 1 1 1 3 3 3 3 1 2 2 1 3 1 3 1 2 3 1 1 3 2 1 3 1 2 1 3 2 3 1 2 1 1 1
##  [75] 1 3 1 3 3 1 1 1 3 3 1 3 3 3 3 3 2 2 1 1 1 3 3 1 1 3 1 3 3 3 1 1 1 3 3 2 3
## [112] 3 3 3 3 2 3 3 1 1 1 3 3 1 3 1 3 3 3 3 1 3 3 2 2 3 3 2 1 3 2 1 2 1 1 3 1 1
## [149] 1 1 1 1 1 2 3 3 2 3 1 1 1 1 1 2 1 1 2 1 1 3 2 1 1 1 1 3 1 1 3 3 1 2 1 1 3
## [186] 3 2 3 3 1 1 1 3 1 1 3 1 2 3 1 1 3 1 1 2 2 1 1 3 1 1 3 1 3 1 1 1 2 1 2 3 1
## [223] 1 1 1 3 1 1 1 2 1 3 1 1 3 3 1 1 1 3 1 3 3 1 2 1 1 1 3 1 1 1 3 3 3 3 2 1 1
## [260] 1 1 1 1 2 3 1 3 1 1 3 1 3 3 3 3 1 1 1 1 1 1 3 1 1 3 1 2 1 1 2 3 1 1 3 1 1
## [297] 3 2 1 3 3 1 3 1 1 1 3 3 3 1 1 3 3 2 1 2 3 3 3 1 3 3 1 1 1 1 2 1 1 1 3 2 1
## [334] 1 1 3 3 2 1 2 3 1 1 1 1 1 1 1 3 1 1 1 3 1 1 2 3 1 2 2 2 1 1 1 1 2 1 1 3 3
## [371] 3 1 2 1 2 1 3 1 1 3 1 2 1 2 3 1 3 3 2 1 2 3 1 1 2 1 1 1 2 3 3 2 1 1 1 3 1
## [408] 2 1 3 3 2 1 1 1 3 1 2 3 1 2 3 3 2 1 1 1 2 1 1 1 3 3 1 1 1 2 2 2 1 2 1 2 1
## [445] 2 1 2 1 1 1 2 1 3 1 3 1 1 1 3 1 1 1 2 3 3 2 1 3 3 1 1 1 1 3 2 1 1 1 1 1 1
```

The fitted k means clustering model (k=3) is stored in kmf object.

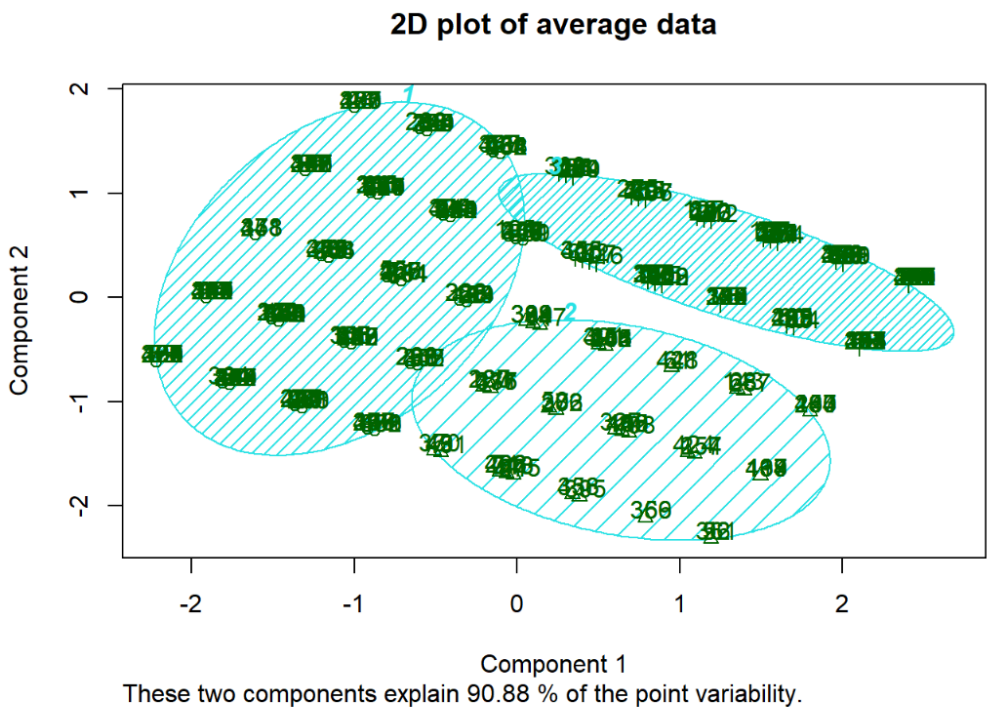
The size of kmf determines the no. of observations falling in each cluster. So from the above output, it can be concluded that 247 stores are grouped in cluster 1, 74 stores in cluster 2, and 160 stores in cluster 3.

Kmf$cluster gives the cluster number corresponding to each store. These values are further stored in a column.

```
cd=  cbind(kmf$cluster)
```

Step 6: Visualizing the clusters using clustplot

```
clusplot(rfm,kmf$cluster,main = "2D plot of average data", labels = 2, shade = TRUE, lines = 0)
```



In this clustplot, PCA works on the backend. To understand the components PCA is run on rfm data.

```
mypca = prcomp(rfm)
summary(mypca)
```

```
## Importance of components:
##                           PC1    PC2     PC3
## Standard deviation     2.0895 1.2853 0.77615
## Proportion of Variance 0.6595 0.2495 0.09099
## Cumulative Proportion  0.6595 0.9090 1.00000
```

The summary function on the result object gives us the standard deviation, the proportion of variance explained by each principal component, and the cumulative proportion of variance explained.

Here, it can be concluded that 65.95% of the variation is explained by PC1 and 24.95% variation is explained by PC2. These 2 components are used to plot the clustplot. Thus, in the clustplot obtained, it is given that these 2 components explain 90.9% of the point variability.

```
mypca
```

```
## Standard deviations (1, .., p=3):
## [1] 2.089528 1.285255 0.776150
##
## Rotation (n x k) = (3 x 3):
##          PC1        PC2         PC3
## R -0.4321746 -0.9015622 -0.02026421
## F -0.6780170  0.3396672 -0.65185821
## M -0.5945738  0.2679771  0.75807007
```

In this way, the PCA works on the backend.

Clusters obtained in kmf$cluster are appended to the Data frame rfm.

```
rfm = cbind(rfm, kmf$cluster)
head(rfm)
```

```
##   R F M kmf$cluster
## 1 5 5 5           3
## 2 4 4 3           3
## 3 4 1 1           1
## 4 4 1 1           1
## 5 5 2 3           3
## 6 1 1 1           1
```

To understand the nature of recency, frequency, and monetary value in a clearer way, this rfm data frame is exported in an excel file. And that file is further used in python.

Before importing the data in python, the Store Names column is again added to the excel file exported from R output.

Step 1: Importing Dataset in python

The imported data is sored in RFMScores data frame.

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
```

```
In [2]: RFMScores = pd.read_csv("D:\\Srushti\\MSc\\Sem 4\\Project\\Analysis\\RFM Clustering.csv")
        RFMScores.head()
```

Out[2]:

	Store Names	R	F	M	Cluster
0	.24x7 CAFETERIA	5	5	5	3
1	.A B SUPER STORE	4	4	3	3
2	.AARDEEP STORE	4	1	1	1
3	.ADINATH GRAIN AND PROV STORE	4	1	1	1
4	.AGRAWAL FARM FRESH	5	2	3	3

Step 2: Plotting a lineplot

Pandas.melt() unpivots a DataFrame from wide format to long format. melt() function is useful to message a DataFrame into a format where one or more columns are identifier variables, while all other columns, considered measured variables, are unpivoted to the row axis, leaving just two non-identifier columns, variable and value.

In [3]:
```
melted_rfm = pd.melt(RFMScores.reset_index(),
                     id_vars=['Cluster'],
                     value_vars = ['R', 'F', 'M'],
                     var_name = 'Features',
                     value_name = 'Value')

melted_rfm.head()
```

Out[3]:

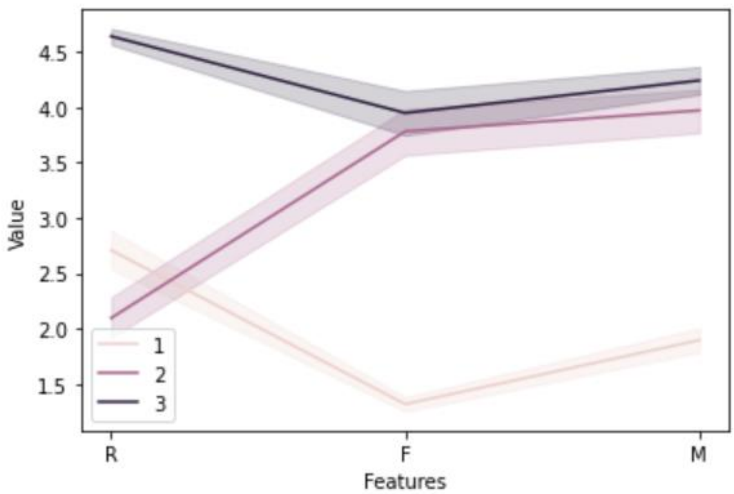
	Cluster	Features	Value
0	3	R	5
1	3	R	4
2	1	R	4
3	1	R	4
4	3	R	5

Using this melted data frame, a lineplot is plotted for each cluster.

In [4]:
```
sns.lineplot('Features', 'Value', hue = 'Cluster', data = melted_rfm)

plt.legend()
```

Out[4]: <matplotlib.legend.Legend at 0x2acacc7da30>



From this lineplot, it can be observed that,

Stores in cluster 1 are having low to moderate ‘R’ scores, low ‘F’ scores, and low ‘M’ scores.

Stores in cluster 2 are having low to moderate ‘R’ scores, moderate to high ‘F’ scores, and moderate to high ‘M’ scores.

Stores in cluster 3 are having high values for all three scores i.e., ‘R’, ‘F’, and ‘M’ scores.

Cluster 3:
Loyal customers: High ‘R’, High ‘F’ and High ‘M’
Customers who bought most recently, most often, and are heavy spenders.

Cluster 2:
At-Risk Customers: Low ‘R’, ‘High F’ and High ‘M’
Customers who purchased often and spent big amounts, but haven’t purchased recently.

Cluster 1:
Lost Customers: Low ‘R’, Low ‘R’, and Low ‘M’
Customers with Low frequency and spending amount. Not placed an order recently.

A statistics summary for RFM scores of each cluster is obtained. It gives the result about RFM scores for each cluster is visualized in above lineplot

```
In [5]: RFMScores.groupby('Cluster').agg({
            'R': ['mean', 'min', 'max'],
            'F': ['mean', 'min', 'max'],
            'M': ['mean', 'min', 'max', 'count']
        })
```

Out[5]:

	R			F			M			
	mean	min	max	mean	min	max	mean	min	max	count
Cluster										
1	2.708502	1	5	1.315789	1	3	1.894737	1	4	247
2	2.094595	1	3	3.783784	2	5	3.972973	2	5	74
3	4.643750	4	5	3.950000	1	5	4.243750	2	5	160

From this table, it can be interpreted as-

There are 247 stores in cluster 1. For this cluster, the ‘R’ score lies between 1-5 with a mean of 2.708502, the ‘F’ score lies between 1-3 with a mean of

1.315789, and the ‘M’ score lies between 1-4 with a mean 1.894737.

Similarly for the other 2 clusters range and mean of R, F, and M.

Conclusions:

Cluster	Customer Type	RFM Characteristics	Action
3	Best/ Loyal Customers	Frequent and recent shoppers. Heavy spending	Potential to be target customers for newly launched products
2	At risk of leaving Customers	Frequent and heavy-spent shoppers. It has been some time since the last purchase	Figuring out the reasons for leaving. Customized plans encouraging purchase again
1	Lost Customers	Low frequency and spending amount and not placing an order recently	The business might have lost them. Finding out the reason for being churned. Improving the services to avoid further losing

CONCLUSIONS

1) From the study, we found that FY 2021-22 has the highest sales.
2) Mass Category (31%) is most sold category followed by Super Saver (23%) and Family Pack (21%).
3) Combo Category (24.27%) has highest Sales in 2019 , whereas in 2020 Family pack (32.05%) and in 2021 , Mass Category (43.47%) has highest Sales.
4) KRISHNA SALES (17.83%) has the most sales out of all stores followed by RAJA TRADERS (13.59%) and Surya Palace(5.50%).
5) General Trade accounts for 51% of total Sales while Institutional Trade covers 49% of total Sales.
6) Choco Kiss is the highest selling product in General Trade and Institutional Trade.
7) Based on T-test , we concluded that there is no significant difference in sales of Old city and New City Route.
8) Similarly, we concluded that there is no significant difference in sales of Institutional Trade and General Trade.

9) Acccording to RFM analysis, we can group the customers into three segments.

Best customers: Customers who bought most recently, most often, and are heavy spenders.

At-Risk Customers: Customers who purchased often and spent big amounts, but haven't purchased recently.

Lost Customers: Customers with Low frequency and spending amount. Not placing an order recently.

REFERENCES

1. The_analysis_of_fmcg_product_availability_in_retail
2. Consumer behavior towards the new packaging of fmcg products
3. Getting started in fixed/random effects models using r
4. Beverage purchases from stores in Mexico under the excise tax on sugar sweetened beverages: observational study
5. A study on fast moving consumer goods FMCG - with special reference to Kancheepuram district
6. https://www.extendoffice.com/documents/excel/2027-excel-bi-directional-bar-chart.html
7. RFM analysis tutorial: simple segmentation analysis- https://www.youtube.com/watch?v=i-hnjzeoomy&feature=youtu.be
8. Temperature and precipitation data online download by google earth- https://www.youtube.com/watch?v=ngtmojivnus
9. https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/
10. Historical climate data- https://www.youtube.com/watch?v=zjpggflmois&feature=youtu.be
11. A review of the application of rfm model- https://www.researchgate.net/publication/228399859_a_review_of_the_application_of_rfm_model
12. Customer segmentation by using RFM model and clustering methods: a case study in retail industry - https://core.ac.uk/download/pdf/296912799.pdf

13.https://www.researchgate.net/publication/338645176_retail_marketing_a_critical_analysis

14.https://www.researchgate.net/publication/328079302_the_consumer_production_journey_marketing_to_consumers_as_co-producers_in_the_sharing_economy

15.https://www.researchgate.net/publication/320613784_sales_and_distribution

16.https://www.emerald.com/insight/content/doi/10.1108/09576059810218982/full/html

17.GLM - https://www.statsmodels.org/stable/mixed_linear.html

18.T-TEST-http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r

19.https://medium.com/swlh/exploring-customers-segmentation-with-rfm-analysis-and-k-means-clustering-93aa4c79f7a7

20.https://clevertap.com/blog/rfm-analysis/

21.https://www.analyticsvidhya.com/blog/2021/07/customer-segmentation-using-rfm-analysis/

22.https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp

23.India Population Growth Rate 1950-2022 | MacroTrends

24. India Population (2022) - Worldometer

25.Population growth (annual %) - India | Data

26. https://power.larc.nasa.gov/data-access-viewer/

27. https://vmc.gov.in/pdf/administation%20ward%20Final.pdf