

## CHAPTER 1

### INTRODUCTION

---

#### **1.1 About Project: -**

1.1.1 Market Basket Analysis is a data mining technique that outputs correlations between various items in a customer's basket.

1.1.2 Market Basket Analysis reports are used to understand what sells with what and includes the probability and profitability of market baskets. Such a report can be used to plan promotions, optimize product placement, and support store planogram decisions. These reports help you understand the statistical relationship between sales for different merchandise.

1.1.3 Market Basket Analysis (MBA) helps you to find the relationship between items and groups of items in the basket of a customer. You can also use it to calculate a promotion-based historical baseline in order to provide insight into retail sale patterns and to improve your understanding of promotional effectiveness.

#### **1.2 Project Objectives :-**

1.2.1 To find frequently purchased item sets from large transactional database.

1.2.2 Which products to put on specials, promote, coupons.

1.2.3 Increasing selling of product.

## CHAPTER 2

### SOFTWARE & HARDWARE

---

#### 2.1 Software:-

Software: - Software used for Coding, Hosting and Database are given below:

For Developing:

2.1.1 Anaconda Navigator:-This is a free and open-source distribution of the Python Data Science Platform.

2.1.2 Python 3.72

2.1.3 matplotlib == 2.1.1

2.1.4 mlxtend == 0.10.0

2.1.5 numpy == 1.13.3

2.1.6 pandas == 0.21.1

2.1.7 scikit\_learn == 0.19.1

#### 2.2 Hardware:-

Hardware: -

- |              |   |                  |
|--------------|---|------------------|
| 1. Processor | : | 2.8 GHZ or above |
| 2. RAM       | : | 8GB or above     |
| 3. HDD       | : | 512GB or above   |

## CHAPTER 3

### PROBLEM DESCRIPTION

---

#### 3.1 Problem Statement :-

Nowadays people buy daily goods from supermarket nearby. There are many supermarkets that provide goods to their customer. The problem many retailers face is the placement of the items. They are unaware of the purchasing habits ,number of time customer visits and they don't know which items should be placed together in their store. With the help of this static website shop managers can determine the strong relationships between the items which ultimately helps them to put products that co-occur together close to one another. Also decisions like which item to stock more, cross selling, up selling, store good arrangement are determined the good customer behaviour.

## **CHAPTER-4**

### **LITERATURE AND SURVEY**

---

#### **4.1 Scope:-**

The scope of the application is limited to desktop application right now. The application is targeted towards a supermarket .

#### **4.2 Benefits:-**

There are lots of Benefits in Market Basket analysis some of these are:

- 4.2.1 Better target marketing.
- 4.2.2 Arrangement of items in retail stores.
- 4.2.3 Increasing growth of selling.
- 4.2.4 A client can compete with other retailers with more technical way.
- 4.2.4 Client can increase it's selling with using it's stored Data.

## CHAPTER-5

### SOFTWARE REQUIREMENT SPECIFICATION

---

#### 5.1 Functional Requirement:-

5.1.1 Login and Registration Panel.

5.1.2 Reports and Other Info panels for selling graphs.

#### 5.2 Non Functional Requirement:

A non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors. They are contrasted with functional requirements that define specific behavior or functions. A careful specification and adherence of non-functional requirements such as performance, security, privacy & availability are crucial to the success or failure of any software system. The correct specification and adherence of non-functional requirements similarly plays at least an equal, if not a greater role in the success of mobile applications.

##### 5.2.1 INTERFACE REQUIREMENT

How will the system interface work with its environment, users and other systems.

e.g. user-friendliness, simple and interactive.

##### 5.2.2 PERFORMANCE REQUIREMENT

Time/space bounds, such as workloads, response time, throughput and available storage space.

##### 5.2.3 SECURITY

Permissible access to data and operations. e.g user login.

## CHAPTER-6

### SOFTWARE DESIGN

---

#### 6.1 ER Diagram:-

An Entity Relationship (ER) Diagram illustrates how “entities” such as people, objects, or concepts relate to each other within a system. ER Diagrams are most often used to design or debug relational databases in the fields of software engineering, business information systems, education, and research. Also known as ERDs or ER Models, they use a defined set of symbols such as rectangles, diamonds, ovals, and connecting lines to depict the interconnectedness of entities, relationships and their attributes. They mirror grammatical structure, with entities as nouns and relationships as verbs.

ER Diagrams are composed of entities, relationships, and attributes. They also depict cardinality, which defines relationships in terms of numbers.

#### 6.2 Entity:-

A definable thing, such as a person, object, concept, or event that can have data stored about it. Examples: a customer, Client, student, car or product. They are represented as a rectangle.

**6.2.1 Entity type :-** A group of definable things, such as students or athletes, whereas the entity would be the specific student or athlete. Other examples: customers, cars or, products.

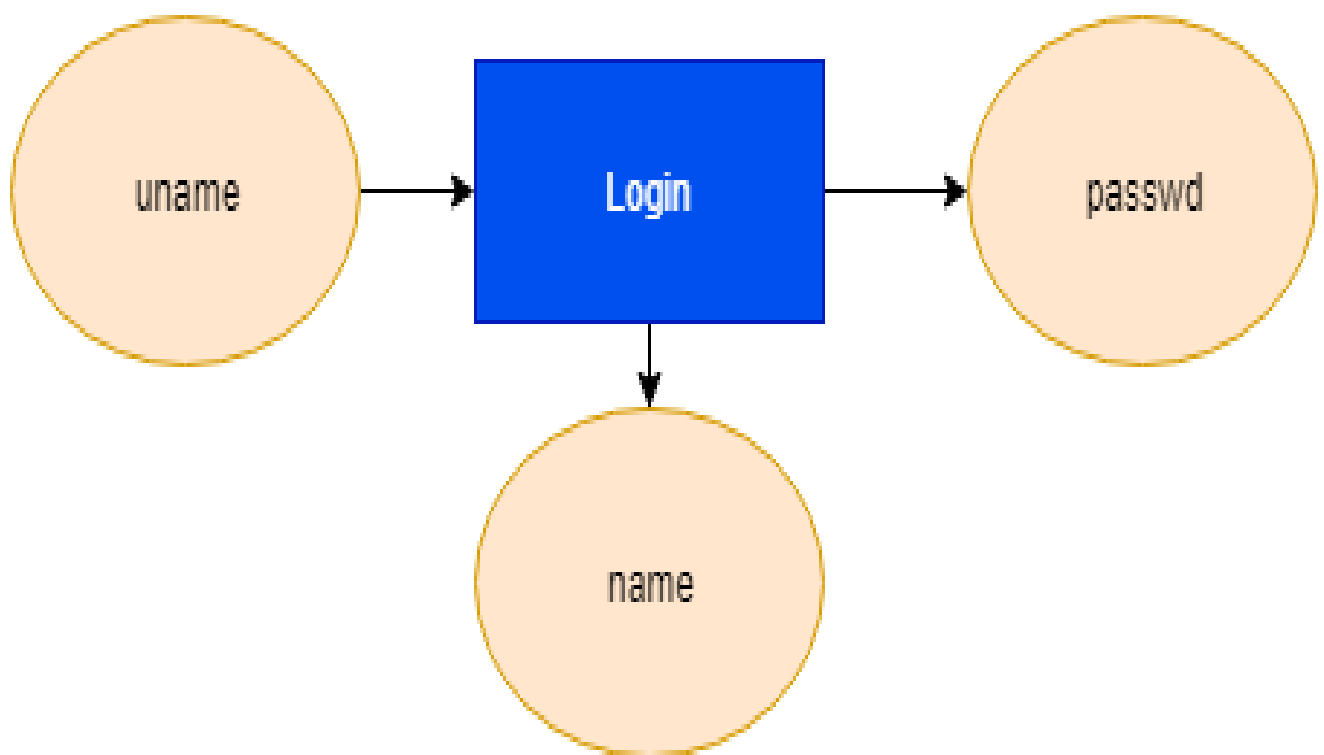
**6.2.2 Entity set :-** Same as an entity type, but defined at a particular point in time, such as students enrolled in a class on the first day. Other examples: Customers who purchased last month, cars currently registered in Florida. A related term is instance, in which the specific person or car would be an instance of the entity set.

**6.2.3 Entity categories :-** Entities are categorized as strong, weak, or associative. A strong entity can be defined solely by its own attributes, while a weak entity cannot. An associative entity associates entities (or elements) within an entity set.

**6.2.4 Entity keys:** Refers to an attribute that uniquely defines an entity in an entity set. Entity keys can be super, candidate, or primary. Super key: A set of attributes (one or more) that together define an entity in an entity set. Candidate key: A minimal super key, meaning it has

the least possible number of attributes to still be a super key. An entity set may have more than one candidate key. Primary key: A candidate key chosen by the database designer to uniquely identify the entity set. Foreign key: Identifies the relationship between entities.

**Fig. 6.1 – E-R Diagram**

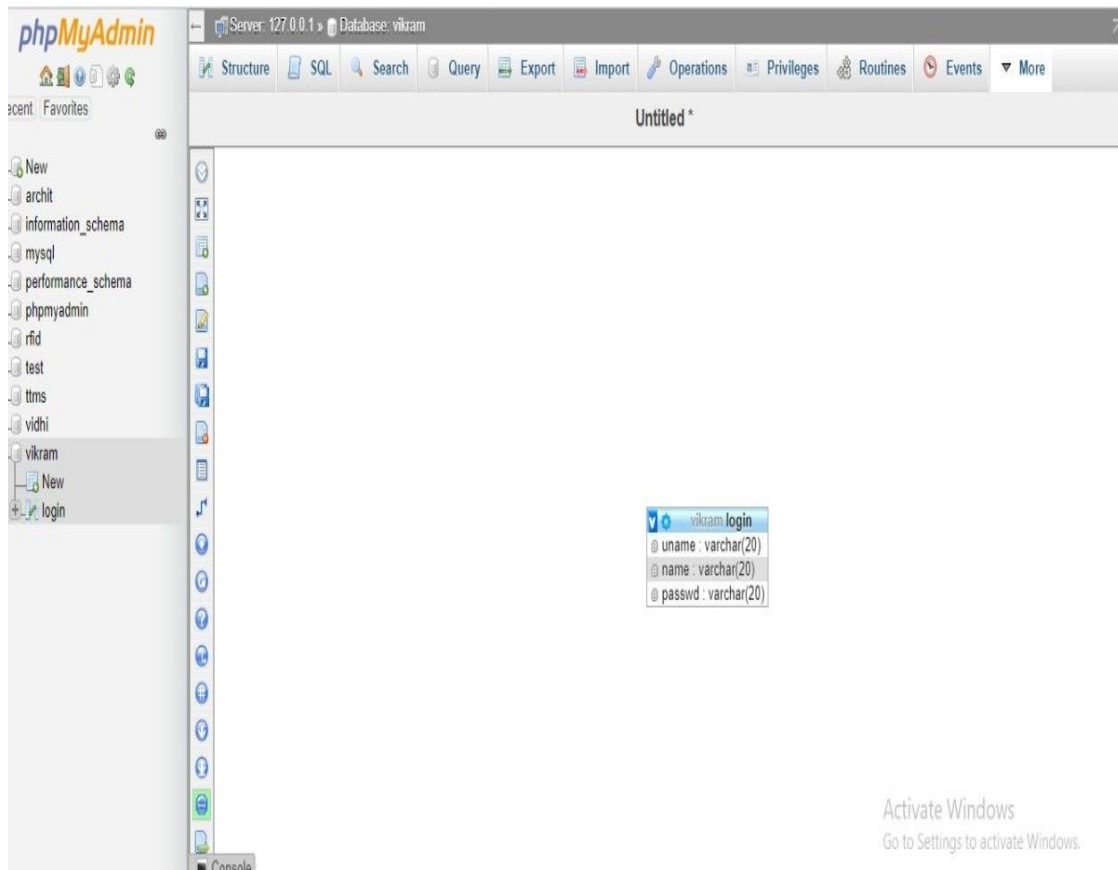


## 6.2 Table Structure

### Database Schema of our Table

#### 1. Member details .

Add member detail table look like this:



**Table 6.2.1– Members Details table**



## CHAPTER-7

### DEPLOYMENT

---

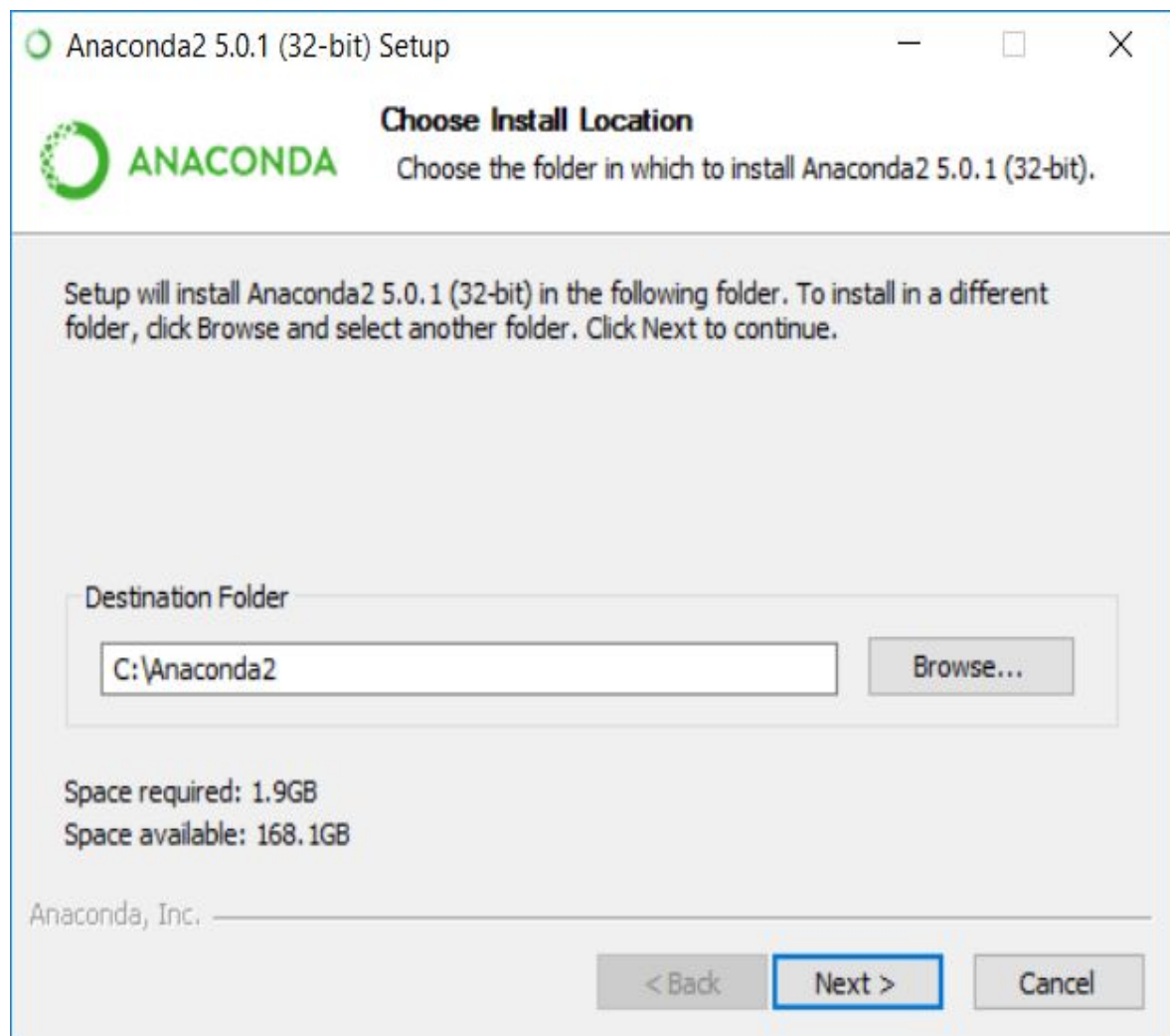
#### 7.1 ANACONDA

Anaconda Cloud is a package management service that makes it easy to find, access, store and share public notebooks, environments, and conda and PyPI packages.

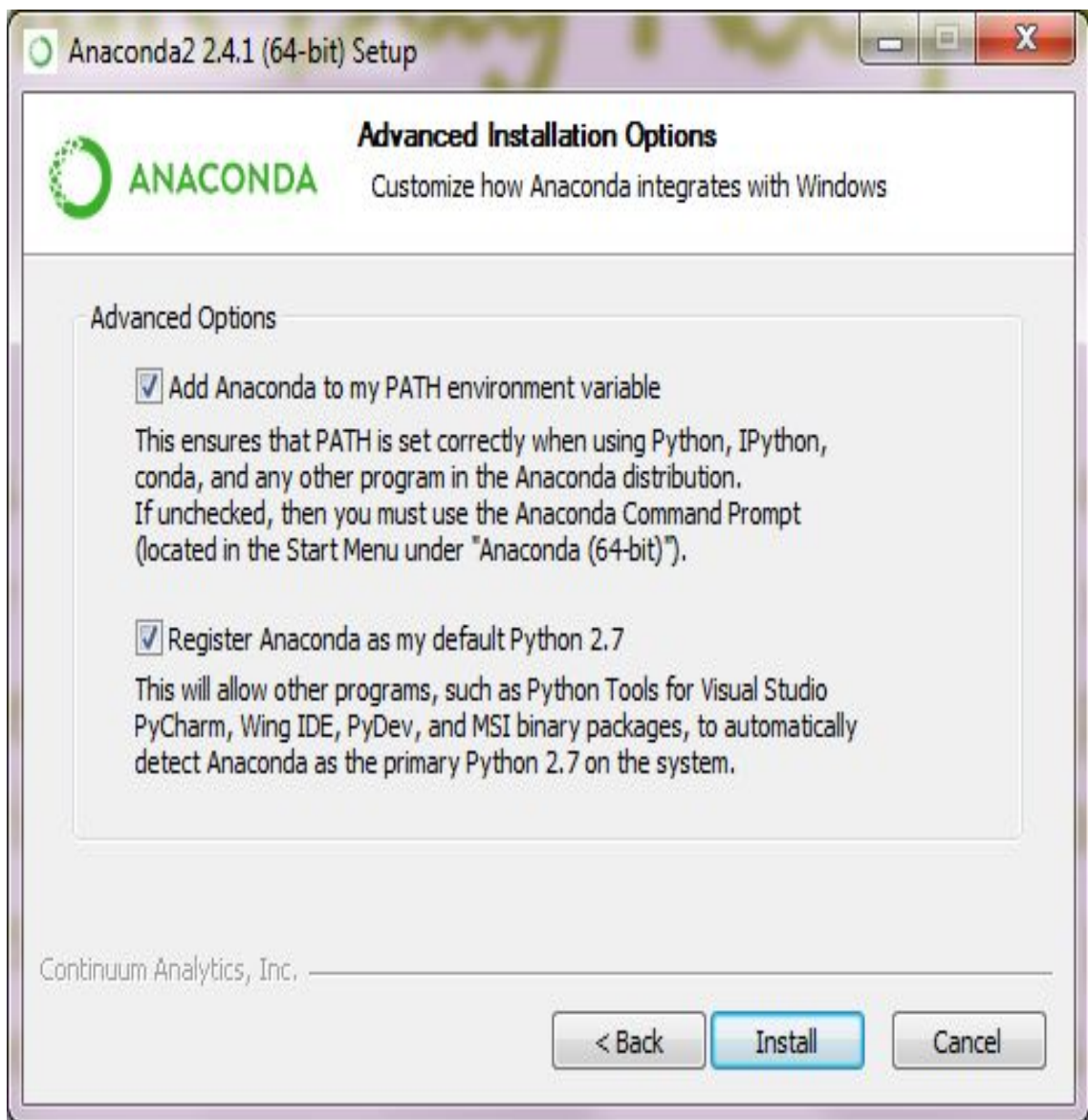
The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. For installation- Install Anaconda. Anaconda conveniently installs Python, the Jupyter Notebook, and other commonly used packages for scientific computing and data science.

Use the following for installation steps:

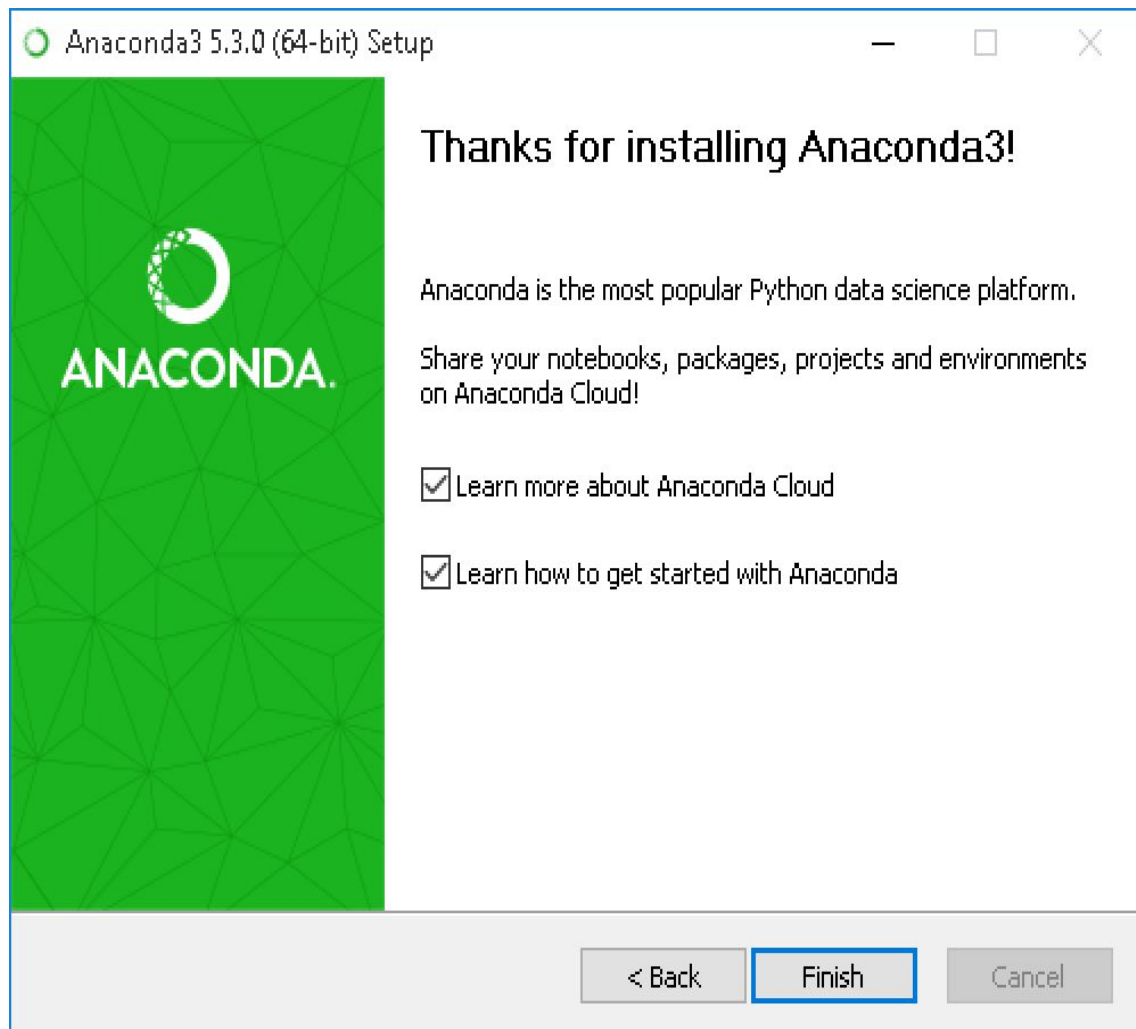
1. Download Anaconda. Install the version of Anaconda which you downloaded, following the instructions on the download page.
2. We have installed Jupyter Notebook. To run the notebook: Jupyter notebook.



**Fig 7.1 Anaconda Location**



**Fig 7.2 Anaconda Path**



(For Mac OS)

Skip this step. No need to do anything.

## CHAPTER-8

### OUTPUT SCREEN

---

In resource Module of shipyard there are many entities like:-

1. Login/Registration page.
2. Reports.
3. Other Info.

#### **Modules:-Login & Registration**

**Login** – In this project, Admin will be already registered. If admin is not registered in database, then Admin cannot access or use the system. So, there has to be a fix Login credential for Admin to access the System.

And user can login any time whenever they wants to check the reports.

**Registration:-**Users can registered themself.

If id and password is doesn't match with database, Nodal or HOD cannot access the system.

**Reports panel:** - User can check the reports in form of graphs .

**Other info panel :** User can check some more details in the form of tables.

## 8.1 Interface of Our Project



**fig 8.1** The Home Page Look like this



**Fig 8.2** Report panel (dropdown)



# Distribution of Number of Order in Percent



Fig. 8.3– Number of order in percent



## Total Number of Sales Per Item

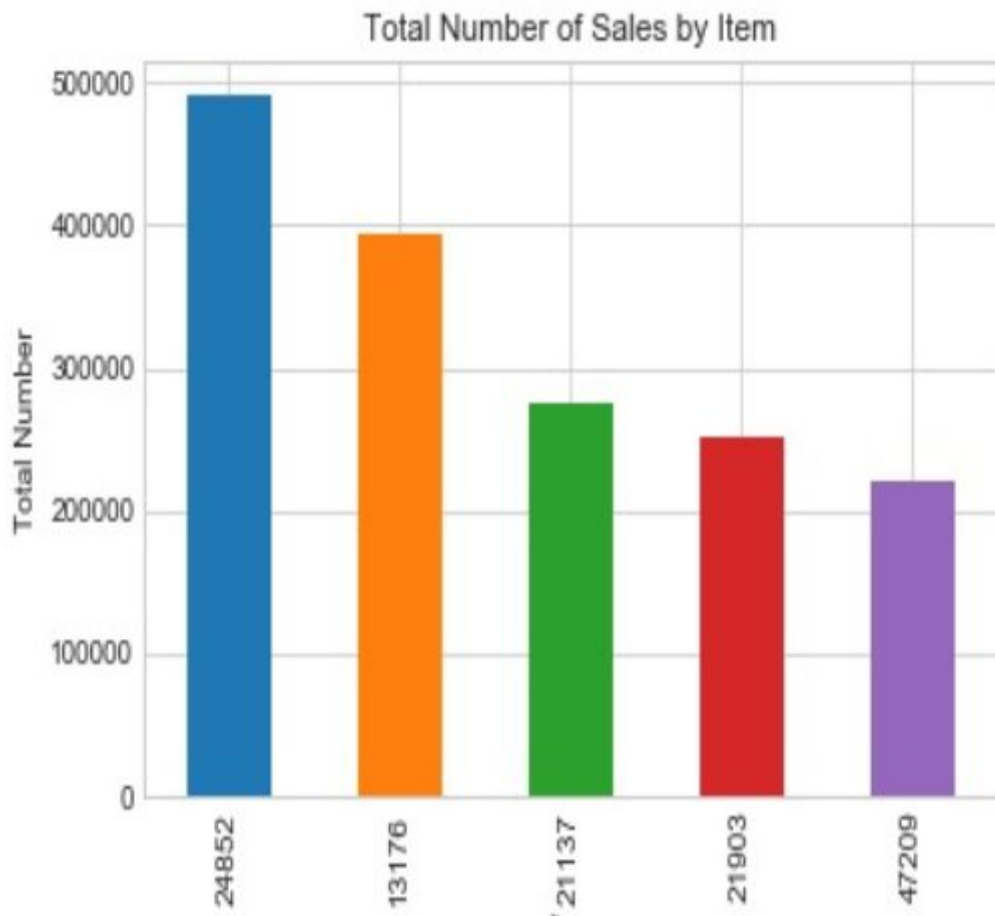
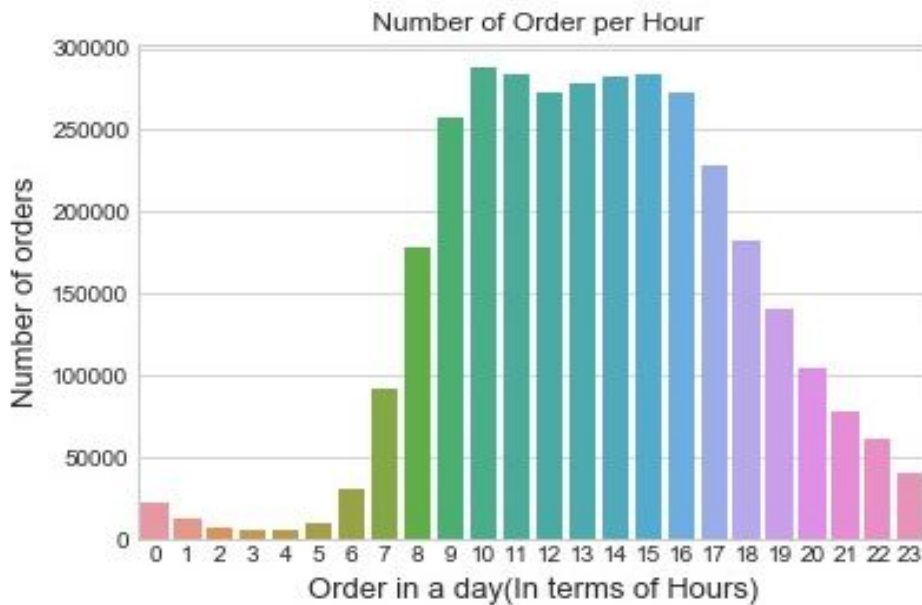


Fig. 8.4– Total number of sales per item



Fig. 8.5– Aisle available for shopping

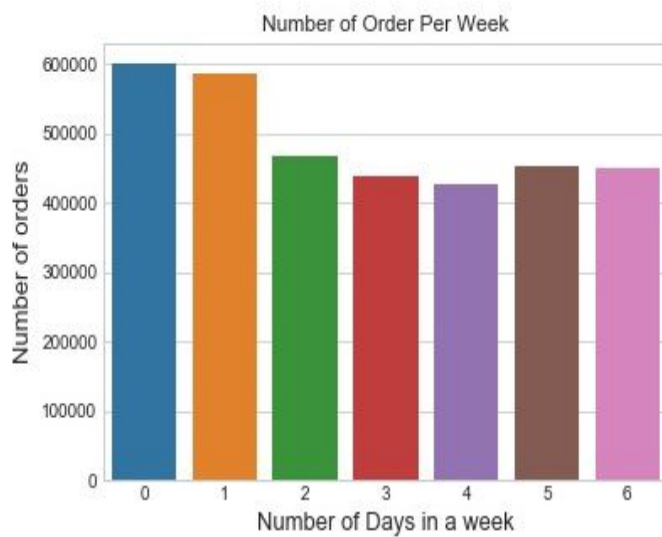
# Total Number of Order in a Day



Sales starts to pick up from 8am, till the busiest hour of the day at 10 am and 11am, then slowly drops till the late afternoon. It can be observed that most of the sales transactions took place during the lunch hours of the days

Fig. 8.6– Total Number of Order in a day

## Total number of Order in a week



Sunday and Monday is the busiest day of the week with the highest sales while Thursday is the quietest day with the lowest sales . This is an interesting insight, the owner of the Bakery should launch some promotion activities to boost up sales in the middle of the week when sales are slowest.

Fig. 8.7-Total number of orders per week



Fig 8.8 Other Info Panel

## BIG BASKET DATA SET

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	date
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom	2010-12-01
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	2010-12-01
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom	2010-12-01
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	2010-12-01
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	2010-12-01
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850.0	United Kingdom	2010-12-01
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850.0	United Kingdom	2010-12-01
536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850.0	United Kingdom	2010-12-01
536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850.0	United Kingdom	2010-12-01
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047.0	United Kingdom	2010-12-01

Table. 8.9 Big Basket Data Set



# Relationship Between Products

	product1	product2	Support	Confidence	Lift
3	fromage blanc	honey	0.003	0.245	5.193
0	chicken	light cream	0.005	0.283	4.749
2	pasta	escalope	0.006	0.373	4.727
8	pasta	shrimp	0.005	0.322	4.532
7	olive oil	whole wheat pasta	0.008	0.271	4.120
5	tomato sauce	ground beef	0.005	0.381	3.888
1	mushroom cream sauce	escalope	0.006	0.301	3.812
6	olive oil	light cream	0.004	0.225	3.415
4	herb & pepper	ground beef	0.016	0.323	3.301
9	avocado	spaghetti	0.003	0.417	3.109
10	cake	milk	0.004	0.280	3.097

Table. 8.10 Relationship Between Products

### Platinum Customers

	Recency	Frequency	Expenditure	cluster
CustomerID				
16446.0	0	2	168472.50	1
17450.0	8	46	194550.79	1
18102.0	0	60	259657.30	1

Total Platinum Custmor=3

### Gold Customer

	Recency	Frequency	Expenditure	cluster
CustomerID				
12346.0	325	1	77183.60	2
12748.0	0	209	33719.73	2
12931.0	21	15	42055.96	2
13081.0	11	11	28337.38	2
13089.0	2	97	58825.83	2

Total Gold Custmor=30

### Silver Customer

	Recency	Frequency	Expenditure	cluster
CustomerID				
12747.0	2	11	4196.01	0
12749.0	3	5	4090.88	0
12820.0	3	4	942.34	0
12821.0	214	1	92.72	0
12822.0	70	2	948.88	0

Total Sliver Custmor=3887

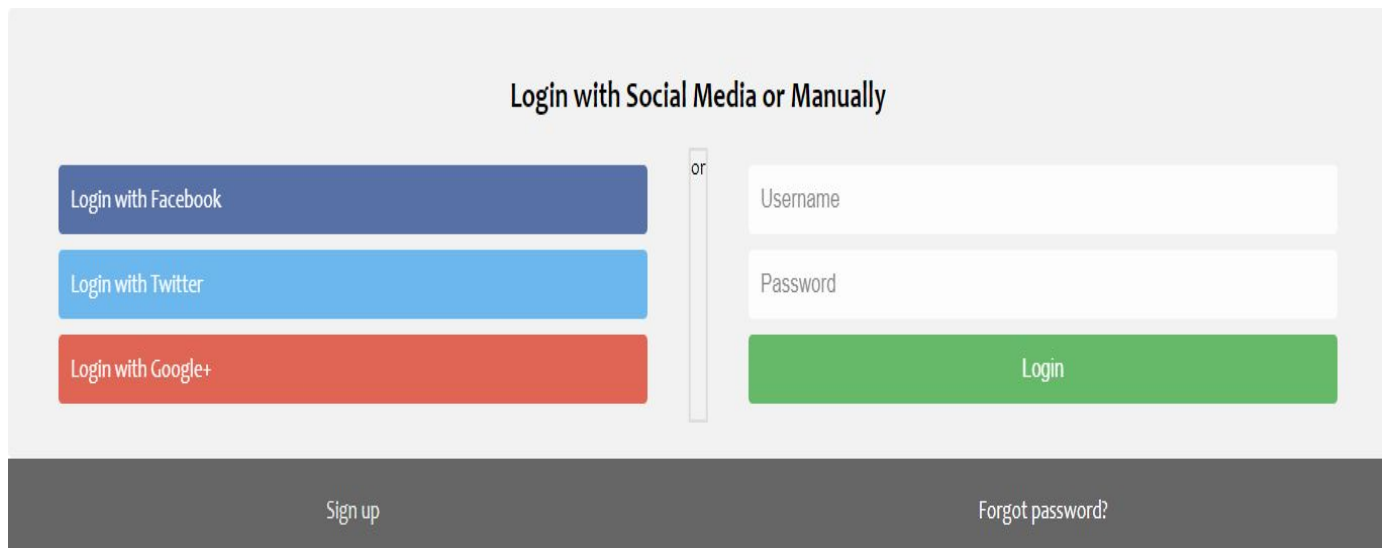
Table. 8.11 Classification of Customers



## TOP 20 PRODUCT DETAILS

	product_id	order_id	product_name	aisle_id	department_id
0	24852	963856	Banana	24	4
1	13176	774380	Bag of Organic Bananas	24	4
2	21137	540260	Organic Strawberries	24	4
3	21903	493626	Organic Baby Spinach	123	4
4	47209	434461	Organic Hass Avocado	24	4
5	47766	361039	Organic Avocado	24	4
6	47626	313449	Large Lemon	24	4
7	16797	292396	Strawberries	24	4
8	26209	287287	Limes	24	4
9	27845	280718	Organic Whole Milk	84	16
10	27966	279660	Organic Raspberries	123	4
11	22935	231142	Organic Yellow Onion	83	4
12	24964	223714	Organic Garlic	83	4
13	45007	214235	Organic Zucchini	83	4
14	39275	205086	Organic Blueberries	123	4
15	49683	197043	Cucumber Kirby	83	4
16	28204	182521	Organic Fuji Apple	24	4
17	5876	178997	Organic Lemon	24	4
18	40706	172333	Organic Grape Tomatoes	123	4
19	8277	172292	Apple Honeycrisp Organic	24	4

Table. 8.12 Top 20 Products Details



The image shows a login and registration interface. At the top, the text "Login with Social Media or Manually" is centered. Below this, on the left, are three stacked buttons: "Login with Facebook" (dark blue), "Login with Twitter" (light blue), and "Login with Google+" (red). To the right of these buttons is a vertical separator with the word "or" in the middle. Further right are two input fields: "Username" and "Password". Below these fields is a green "Login" button. At the bottom of the interface, there is a dark gray bar containing two links: "Sign up" on the left and "Forgot password?" on the right.

Login with Social Media or Manually

Login with Facebook

Login with Twitter

Login with Google+

or

Username

Password

Login

Sign up

Forgot password?

fig 8.13 Login/Registration Page

## CHAPTER-9

### METHODOLOGY

---

Currently the proposed approach is for a static data frame which work on in seven phases. In first phase we have done the preprocessing on the dataframe. In second phase we feature Engineering after the preprocessing of the dataframe. In third phase we use Recency Frequency Monetary. In fourth phase we use K-means algorithm for the classification of the customer on the basis of the Frequency, Recency, Expenditure which we get from the Recency Frequency Monetary. In fifth phase we have the visualization part where we generates some data and plot them into the graphs. And in final phase in machine learning section we use Apriori algorithm for finding the product relationship.

And with the help of the HTML, CSS and bootstrap we had made a static website show the reports (contain graphs) and other data which we have generated from the above phases this and connectivity for the admin user can be term as 7th phase also..

#### 9.1 Project Module:

9.1.1 The phase one ,phase two and the phase three is done by Sushen Patidar.

9.1.2 The fourth and fifth phase is don by Vikram Singh Rathore.

9.1.3 The last phase is done by Sima Kumari

9.1.4 The website designing part is done by Durgesh Nandini

##### 9.1.1 Phase -I:Data preprocessing

Data preprocessing, the more disciplined you handling of data, the more consistent the result you are like to achieve. The process for getting data ready for a machine learning algorithm can be summarized in three steps:

Step 1: Select Data :Consider what data is available, what data is missing and what data can be removed.

Step 2: Preprocess data :Organize your selected data by formatting, cleaning and sampling from it.

Step 3: Transform Data :Transform preprocessed data ready for other process by Features Engineering and convert data in different transformations as you work on your

problem and ML algorithms.

#### **9.1.1.2 Phase-II:Feature Engineering**

Organize Feature Engineering is the process of using Domain Knowledge of the data to create features that make machine learning algorithms work. If feature engineering is done correctly, it increases the predictive power of machine learning algorithms by creating features from raw data that help facilitate the machine learning process.

Feature Engineering is an art.

- 1)Create features
- 2)check how the features work with the model

#### **9.1.1.3 Phase III: RFM**

RFM is a method used for analyzing customer value. It is commonly used in database marketing and direct marketing and has received particular attention in retail and professional services industries

RFM stands for the three dimensions:

1. Recency – How recently did the customer purchase?
2. Frequency – How often do they purchase?
3. Monetary Value – How much do they spend?

#### **9.1.2.1 Phase IV: K-Means**

9.1.2.1 K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand.

9.1.2.2 The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached.

9.1.2.3 K-means clustering has uses in search engines, market segmentation, statistics and even astronomy.

#### **9.1.2.2 Phase V:Visualization**

In this we have plot some graphs and tables with the help of given data which

have been shown above.

9.1.2.2.1 Represent the number of order per customer.

9.1.2.2.2 Represent number of sales per product..

9.1.2.2.3 Represent the number of aisles.

9.1.2.2.4 Represent the number of order in a day.

9.1.2.2.5 Represent number of order per week.

9.1.2.2.6 Display the Big Basket dataset.

9.1.2.2.7 Display the Relationship between products.

9.1.2.2.8 Classification of customer.

9.1.2.2.9 Top 20 product details.

### **9.1.3 Phase VI:Apriori Algorithm**

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for association rules.

9.1.6.1 Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are tested against the data.

9.1.6.2 Apriori is designed to operate on database containing transactions (for example, collections of items bought by customers, or details of a website frequentation).

9.1.6..3 Association rules are used to find relationships between objects which are frequently used together. It implies that if an item A occurs,there is a probability of item B to occur as well.

### **9.1.4 Phase VII:Website Designing**

We have made a static website with the help of HTML,CSS,Bootstrap for the representation of the data in the form of graphs and tables and the connectivity for admin login with PHP

## REFERENCE

---

### YouTube

# Channel name :- Web dev

# [https://www.youtube.com/watch?v=uyaV\\_EWWRmo](https://www.youtube.com/watch?v=uyaV_EWWRmo)

### Website:-

<https://www.kaggle.com/c/instacart-market-basket-analysis/data>

<https://www.w3schools.com/css/default.asp>

<https://www.w3schools.com/bootstrap/default.asp>

## CONCLUSION

---

We have shown how market basket analysis using association rules work in determining the customer pattern. The really nice aspect of association analysis is that it is easy to run and relatively easy to interpret. If we did not have access to MLxtend, some other python libraries and this association rules, it would be exceedingly difficult to find these patterns by using basic Excel analysis. With python and MLxtend, the analysis process is relatively straightforward and since we are in python, we have access to all the additional visualization techniques and data analysis tools in the python ecosystem.