DATA MANAGEMENT PROJECT REPORT

(Project Semester: August-December 2021)

LOVELY PROFESSIONAL UNIVERSITY

Online Business Sale 2017-2019

Submitted by

Ayisha Nourish

11902302

Programme and Section: B.Tech (CSE), K19AM

Course Code: INT217

Under the Guidance of

Maneet Kaur- 15709

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

Lovely Professional University, Phagwara

DECLARATION

I, Ayisha Nourish, student of B.Tech CSE under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine. I completed my project in Time.

Date: 20/12/2021

Ayisha Nourish

Registration No.: 11902302

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my teacher Mrs.Maneet Kaur who gave me the golden opportunity to do this wonderful project of analysis of the data of a superstore namely "Online Business Sale 2017-2019" which also helped me in doing a lot of research and I came to know about so many new things. I am thankful to them. Secondly, I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

TABLE OF CONTENT

INTRODUCTION

This project give the information about the online Business sale in a period of 2017-2019 .From the data we analyze about the product details that we sold with all the information regarding it . By taking monthly ways consideration in each year we will analyze its order details, total number of orders , discounts of each month , returns happened in each month. Likewise taking all such data we gone to find solution for a customer problem to find which month have high discount and minimum shipping fee and also find the details about the sale happened for each product in a particular month.

Data Analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decisionmaking. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science.

SCOPE OF ANALYSIS

This project on Online Business Sale Statistics of India provides the overall Statistics details of online purchase of customers in a period of 2017-2019.

For analysing Data Excel and Tableau Prep software is used here

Excel is a handy software that can be used to store and organize many data sets. Using its features and formulas, you can also use the tool to make sense of your data. For example, you could use a spreadsheet to track data and automatically see sums averages and totals.

Tableau Prep is a personal data preparation tool that empowers the user with the ability to cleanse, aggregate, merge or otherwise prepare their data for analysis in Tableau.

OBJECTIVES

The main objective of the project is to make questions from the data set and making visual insight then finding answer from it.

I pointed to some questions to answer to the overall analysis of dataset

Questions:

1. **Which month in each year has the most discount sale had happened?**
2. **Which month has the highest number of orders in each year?**
3. **Which product has highest discount?**
4. **What is the percentage of orders in each year?**
5. **Which Product has the most returns?**
6. **What is most sold product or customer most bought product?**
7. **What is the total shipping cost of all months in each year?**
8. **What is the net quantity of each item?**
9. **What is the sum of total sales of every month in each year?**

First cleaning data using Tableu prep then getting this data into excel finding answer for this question by making visual insights and dashboards.

This way help us to find hidden answers from it and can utilize the result in business which definitely help to grow business to next level.

SOURCE OF DATASET

The data is being taken from the Kaggle .Kaggle is an AirBnB for Data Scientists – this is where they spend their nights and weekends. It's a crowd-sourced platform to attract, nurture, train and challenge data scientists from all around the world to solve data science, machine learning and predictive analytics problems. It has over 536,000 active members from 194 countries and it receives close to 150,000 submissions per month. Started from Melbourne, Australia Kaggle moved to Silicon Valley in 2011, raised some 11 million dollars from the likes of Hal Varian (Chief Economist at Google), Max Levchin (Paypal), Index and Khosla Ventures and then ultimately been acquired by the Google in March of 2017. Kaggle is the number one stop for data science enthusiasts all around the world who compete for prizes and boost their Kaggle rankings. There are only 94 Kaggle Grandmasters in the world to this date

Kaggle Dataset Link: https://www.kaggle.com/tylermorse/retail-business-sales-20172019

ETL PROCESS

In computing, extract, transform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing. Data extraction involves extracting data from homogeneous or heterogeneous sources, while data transformation processes data by transforming them into a proper storage format/structure for the purposes of querying and analysis. finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, or a data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

- Extraction

  In this step data is extracted from the source system into the staging area. Transformation if any are done in staging area so that performance of source system is not degraded.Also,if corrupted data is copied directly from the source data warehouse database, rollback will be challenge.Staging area gives an opportunity to validate data before it moves into the data warehouse.

- Transformation

  Data extracted from source server is raw and not usable in it’s orginal form. Therefore it needs to be cleaned, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data.In this step,you apply a set of functions on extracted data.Data that does not required any transformation is called as direct move or pass through data.

- Load

The load phase loads the data into the end target, which may be a simple delimited flat file or a data warehouse. Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis. Other data warehouses (or

even other parts of the same data warehouse) may add new data in a historical form at regular intervals.for example, hourly. To understand this, consider a data warehouse that is required to maintain sales records of the last year. This data warehouse overwrites any data older than a year with newer data. However, the entry of data for any one-year window is made in a historical manner. The timing and scope to replace or append are strategic design choices dependent on the time available and the business needs. More complex systems can maintain a history and audit trail of all changes to the data loaded in the data warehouse. As the load phase interacts with a database, the constraints defined in the database schema as well as in triggers activated upon data load apply (for example, uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.

APPLICATIONS OF ETL PROCESS

1. A financial institution might have information on a customer in several departments and each department might have that customer's information listed in a different way. The membership department might list the customer by name, whereas the accounting department might list the customer by number. ETL can bundle all these data elements and consolidate them into a uniform presentation, such as for storing in a database or data warehouse.
2. Companies use ETL is to move information to another application permanently. For instance, the new application might use another database vendor and most likely a very different database schema. ETL can be used to transform the data into a format suitable for the new application to use.

In this Process there mainly involve collecting and Cleaning data

There is different type of data

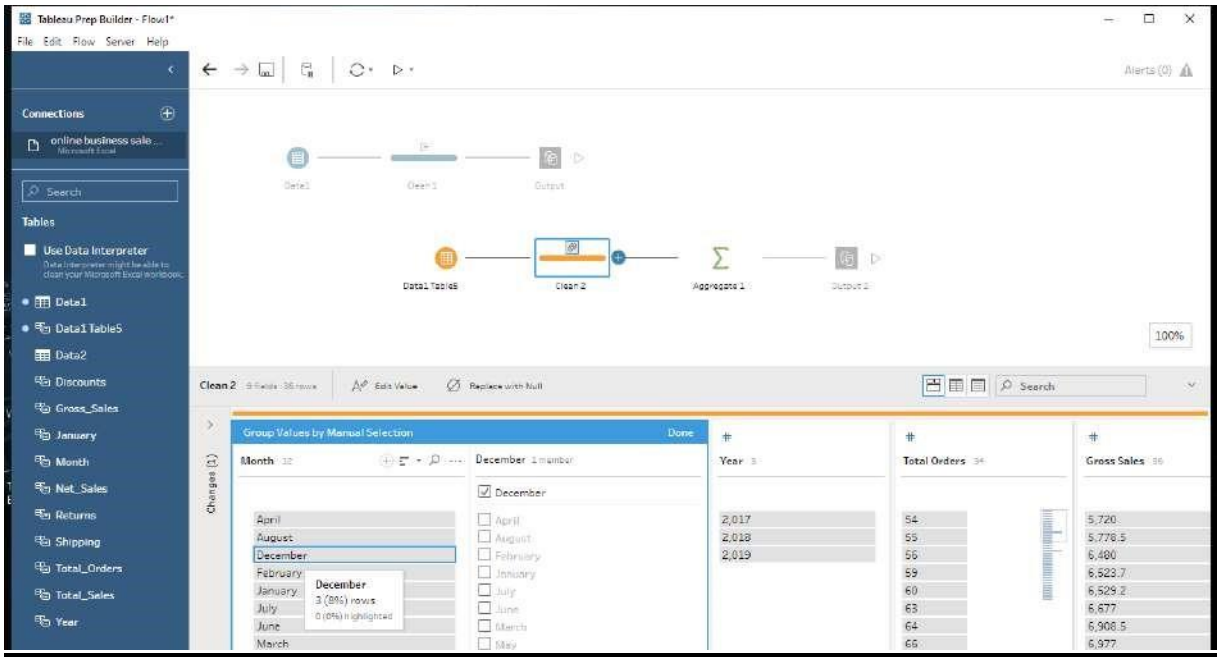
1. Structured data

2. Unstructured

In this project the data is taken from kaggle which is a Structured data

For Structured data we only needed simple cleanings like removing duplicates, grouping misspelled details, Removing null values to approximate values, Removing Unwanted details.

Step 1: Simple Cleaning in Tableau (Because it is a structured data)

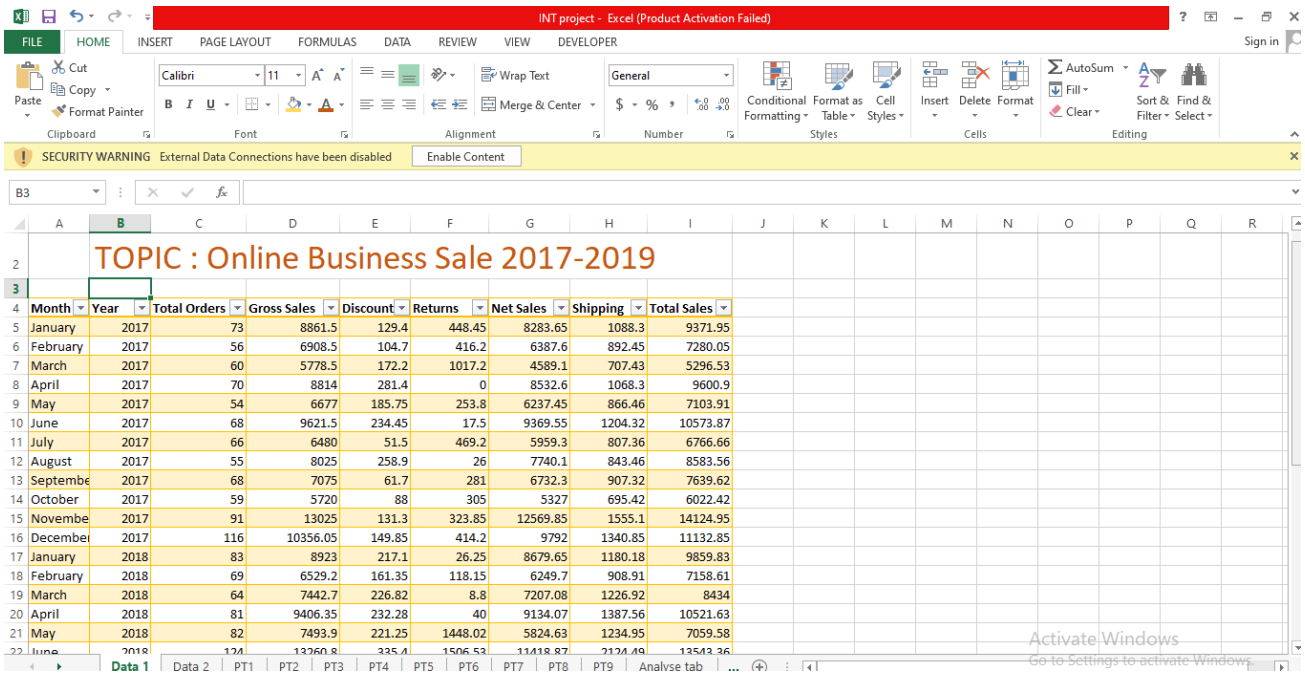


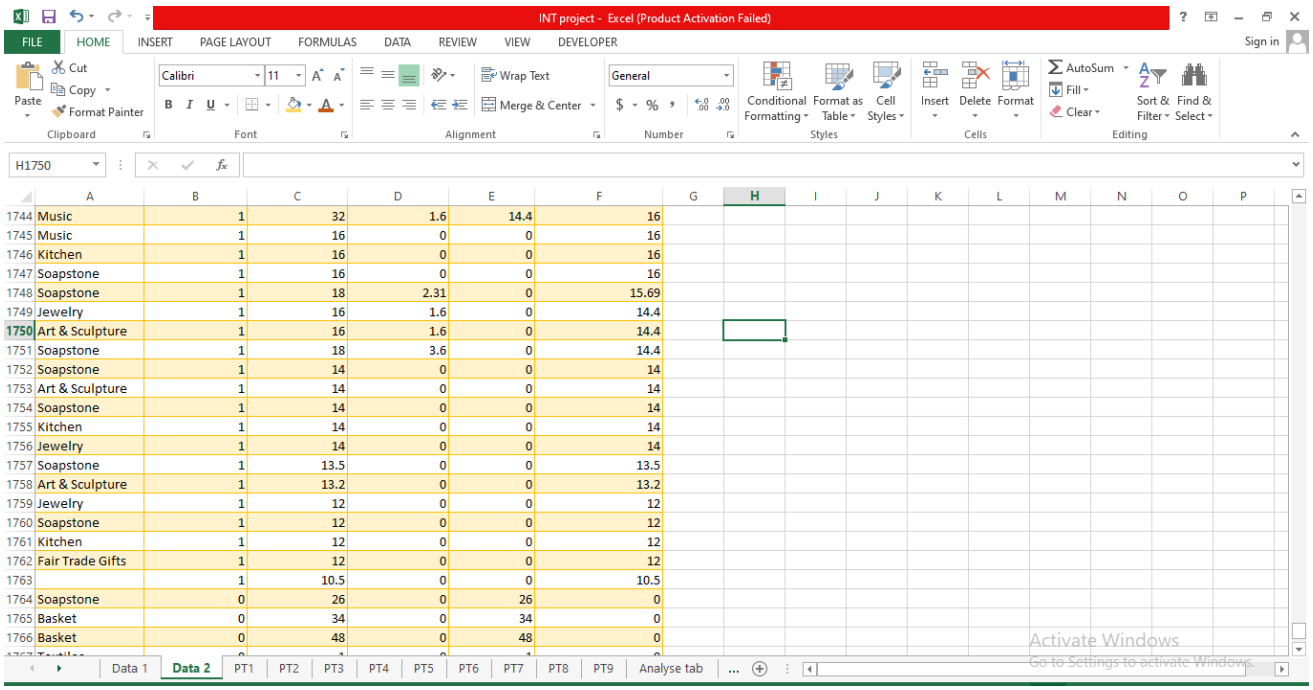
After Simple cleaning taking it output and merging in a excel sheet.

Excel

Step 2: Putting Filter to all data

Select Dataset → Home → Sort and Filter → Filter





	A	B	C	D	E	F
1744	Music	1	32	1.6	14.4	16
1745	Music	1	16	0	0	16
1746	Kitchen	1	16	0	0	16
1747	Soapstone	1	16	0	0	16
1748	Soapstone	1	18	2.31	0	15.69
1749	Jewelry	1	16	1.6	0	14.4
1750	Art & Sculpture	1	16	1.6	0	14.4
1751	Soapstone	1	18	3.6	0	14.4
1752	Soapstone	1	14	0	0	14
1753	Art & Sculpture	1	14	0	0	14
1754	Soapstone	1	14	0	0	14
1755	Kitchen	1	14	0	0	14
1756	Jewelry	1	14	0	0	14
1757	Soapstone	1	13.5	0	0	13.5
1758	Art & Sculpture	1	13.2	0	0	13.2
1759	Jewelry	1	12	0	0	12
1760	Soapstone	1	12	0	0	12
1761	Kitchen	1	12	0	0	12
1762	Fair Trade Gifts	1	12	0	0	12
1763		1	10.5	0	0	10.5
1764	Soapstone	0	26	0	26	0
1765	Basket	0	34	0	34	0
1766	Basket	0	48	0	48	0

Giving filter is the best way to get data that we wanted. we can filter out according to our analysis.

Step 3: Making Pivot Table

For analysis each data separately this is the best way To make pivot table : Select anywhere in the Dataset → insert → Pivot table → select output sheet .

Making Pivot Table according to our objectives

For data analysis first step is always making question then planning to How to get into answer? What are the things needed for it?

Making Pivot table include in planning.According to our each question we make a pivot table

After making pivot table use it to make Visual insights.

According to data we can make Bar graph, Pie graph, Scatter graph … to make it as easily analysing visual insights.

Step 4: Making pie chart/ Bar graph / Column chart

To make pie chart:

Select the pivot table →Insert → pie chart

To make Bar graph:

Select the pivot table →Insert → Bar graph

To make column chart:

Select the pivot table → insert → Column chart

Analysis of data set based on objective

OBJECTIVE 1

Which month in each year has the most discount sale had happened?

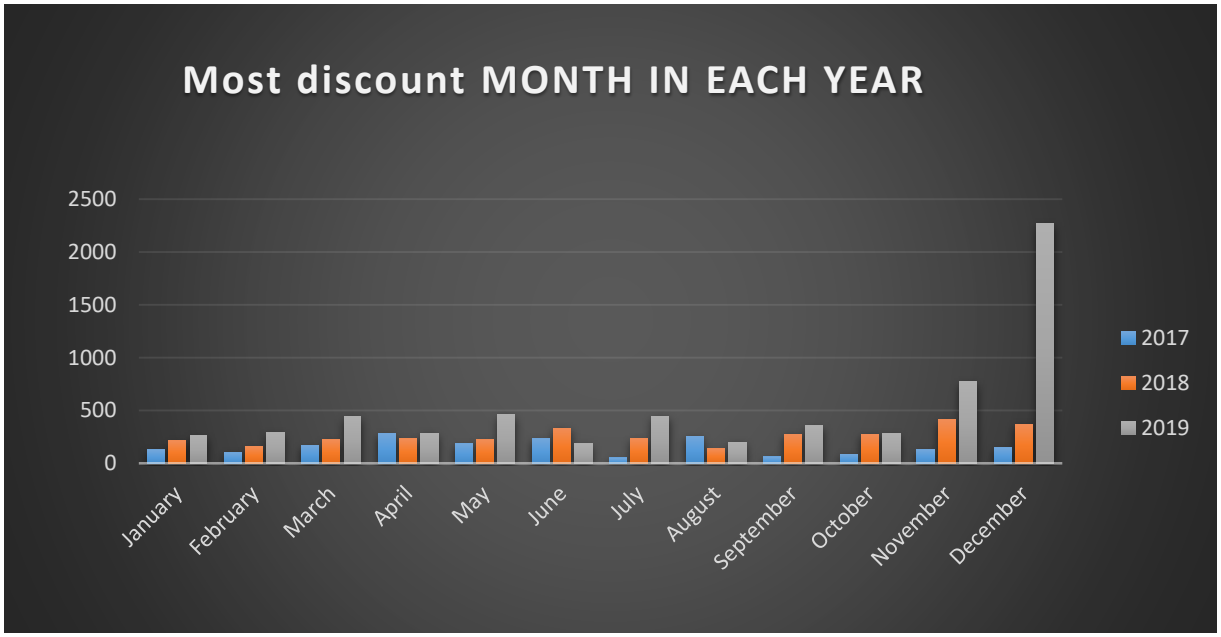
SPECIFIC REQUIREMENT/FUNCTIONS AND FORMULAS

- Pivot table of the month wise discount
- With the help of this plot 2D clustered column Chart.

DATA ANALYSIS

Sum of Discounts	Year			
Month	2017	2018	2019	Grand Total
January	129.4	217.1	261.97	608.47
February	104.7	161.35	288.7	554.75
March	172.2	226.82	439.85	838.87
April	281.4	232.28	285.4	799.08
May	185.75	221.25	460.9	867.9
June	234.45	335.4	186.02	755.87
July	51.5	237.87	447.07	736.44
August	258.9	140.57	201.67	601.14
September	61.7	276.15	354.89	692.74
October	88	277.95	279.42	645.37
November	131.3	414.45	776.84	1322.59
December	149.85	371.2	2269.51	2790.56
Grand Total	**1849.15**	**3112.39**	**6252.24**	**11213.78**

VISUALIZATION



RESULT ANALYSIS

From the chart it is clear that in 2017-April ,2018-November and in 2019-December Had most discount sale happened.

OBJECTIVE 2

Which month has the highest number of orders in each year?

SPECIFIC REQUIREMENT/FUNCTIONS AND FORMULAS

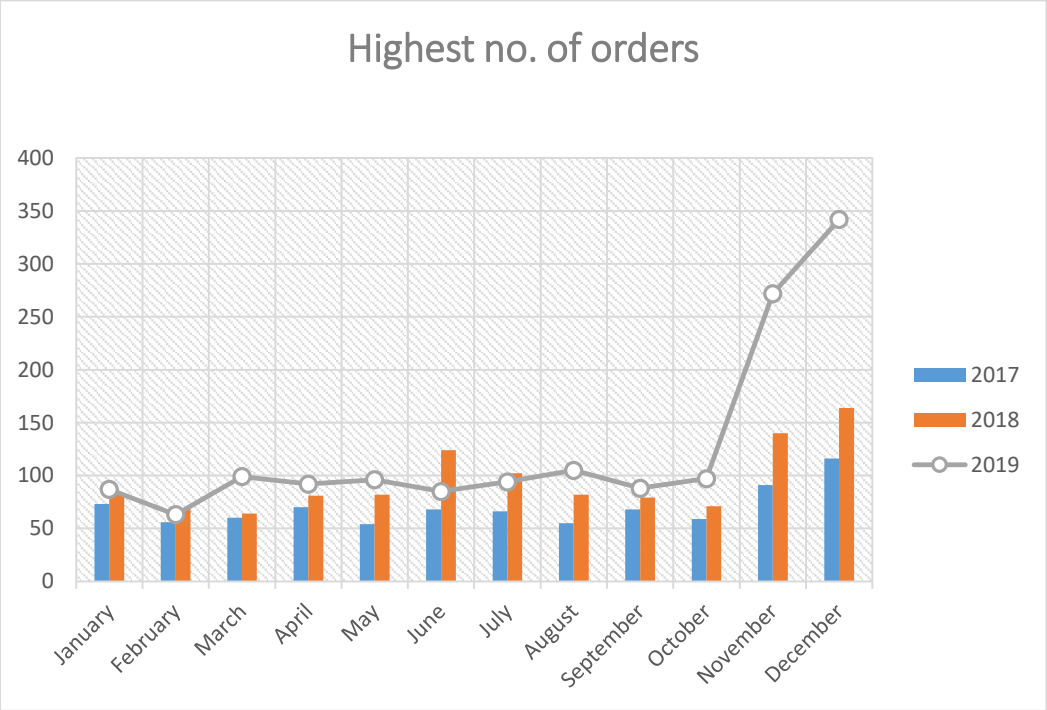
- Pivot table of number of orders in each month of years
- With the help of 2D clustered column line chart

DATA ANALYSIS

Sum of Total Orders	Year			
Month	2017	2018	2019	Grand Total
January	73	83	87	243
February	56	69	63	188
March	60	64	99	223
April	70	81	92	243

May	54	82	96	232
June	68	124	85	277
July	66	102	94	262
August	55	82	105	242
September	68	79	88	235
October	59	71	97	227
November	91	140	272	503
December	116	164	342	622
Grand Total	**836**	**1141**	**1520**	**3497**

VISUALIZATION



RESULT ANALYSIS

From 2017-19,Dec has the highest number of orders.

OBJECTIVE 3

Which product has highest discount?

SPECIFIC REQUIREMENT/FUNCTIONS AND FORMULA

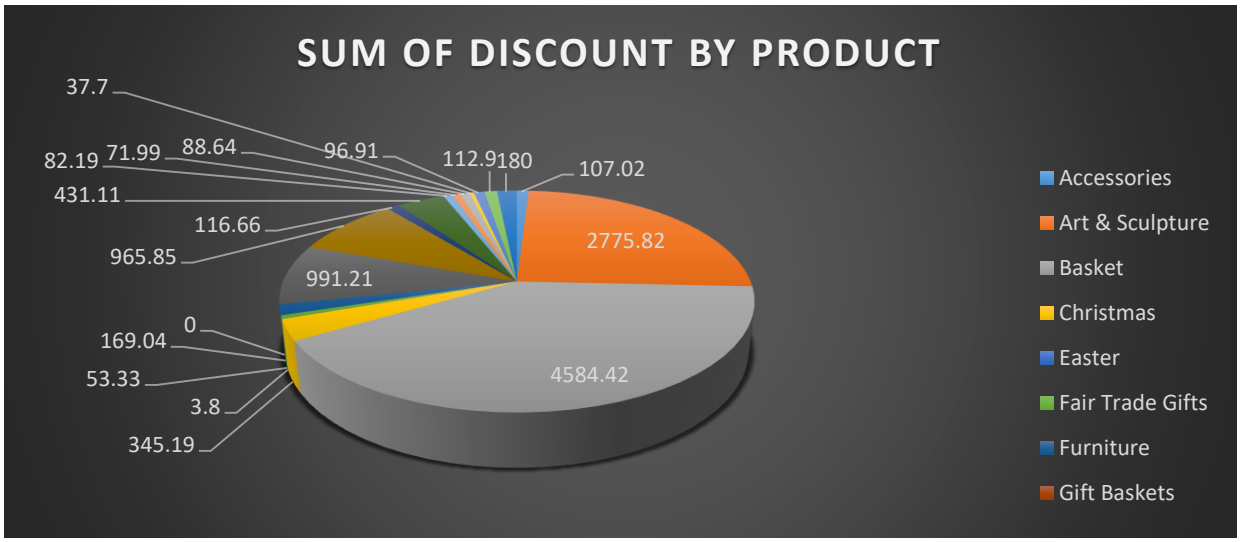
- Pivot table of sum of discounts by product type.

- With the help of this plot 3D pie Chart.

DATA ANALYSIS

Product Type	Sum of Discounts
Accessories	107.02
Art & Sculpture	2775.82
Basket	4584.42
Christmas	345.19
Easter	3.8
Fair Trade Gifts	53.33
Furniture	169.04
Gift Baskets	0
Home Decor	991.21
Jewelry	965.85
Kids	116.66
Kitchen	431.11
Music	82.19
One-of-a-Kind	71.99
Recycled Art	88.64
Skin Care	37.7
Soapstone	96.91
Textiles	112.9
(blank)	180
Grand Total	**11213.78**

VISUALIZATON



RESULT ANALYSIS

Basket and Arts & Sculpture has the max discount occur.

OBJECTIVE 4

What is the percentage of orders in each year?

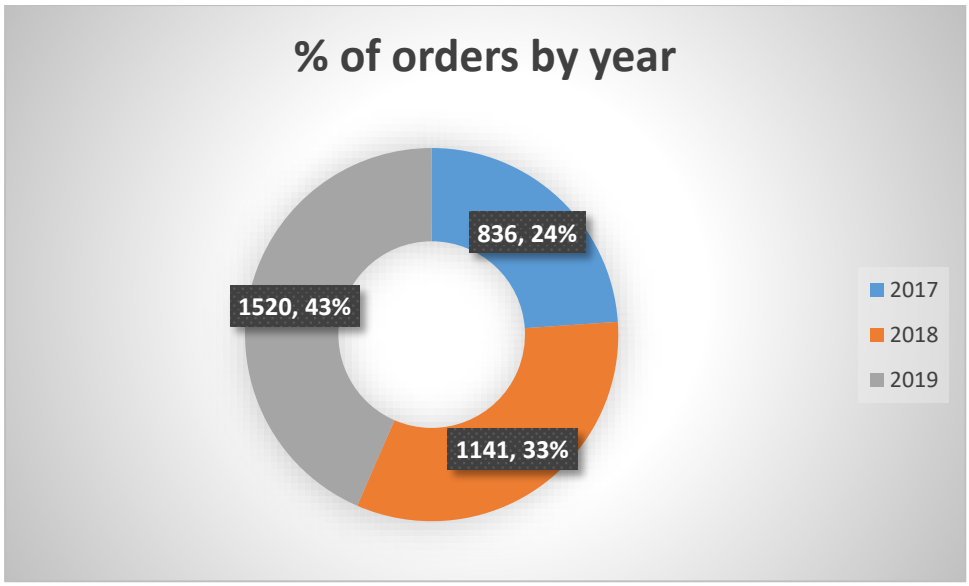
SPECIFIC REQUIREMENT/FUNCTIONS AND FORMULA

- Pivot table of sum of total orders by year.
- With the help of this Doughnut pie chart is plotted

DATA ANALYSIS

Year	Sum of Total Orders
2017	836
2018	1141
2019	1520
Grand Total	**3497**

VISUALIZATION



RESULT ANALYSIS

In 2019 has happened most no. of orders.

OBJECTIVE 5

Which product has the most returns?

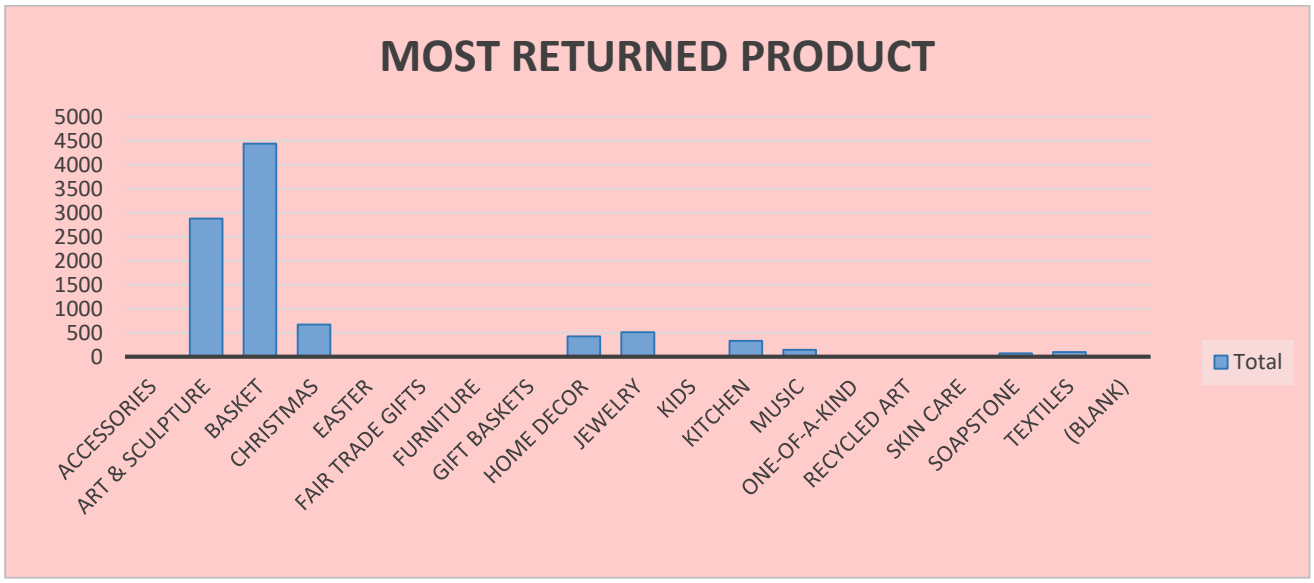
SPECIFIC REQUIREMENT/FUNCTIONS AND FORMULA

- Plot the pivot table of sum of returns of each product.
- With the help of this 2D clustered column chart is plotted

DATA ANALYSIS

Product Type	Sum of Returns
Accessories	0
Art & Sculpture	2879.93
Basket	4439.69
Christmas	670
Easter	0
Fair Trade Gifts	0
Furniture	0
Gift Baskets	0
Home Decor	423.35
Jewelry	509.2
Kids	0
Kitchen	328.07
Music	142.41
One-of-a-Kind	0
Recycled Art	0
Skin Care	0
Soapstone	69.5
Textiles	97
(blank)	0
Grand Total	**9559.15**

VISUALIZATION



RESULT ANALYSIS

Basket is the most returned product.

OBJECTIVE 6

What is most sold product or customer most bought product?

SPECIFIC REQUIREMENT/FUNCTIONS AND FORMULA

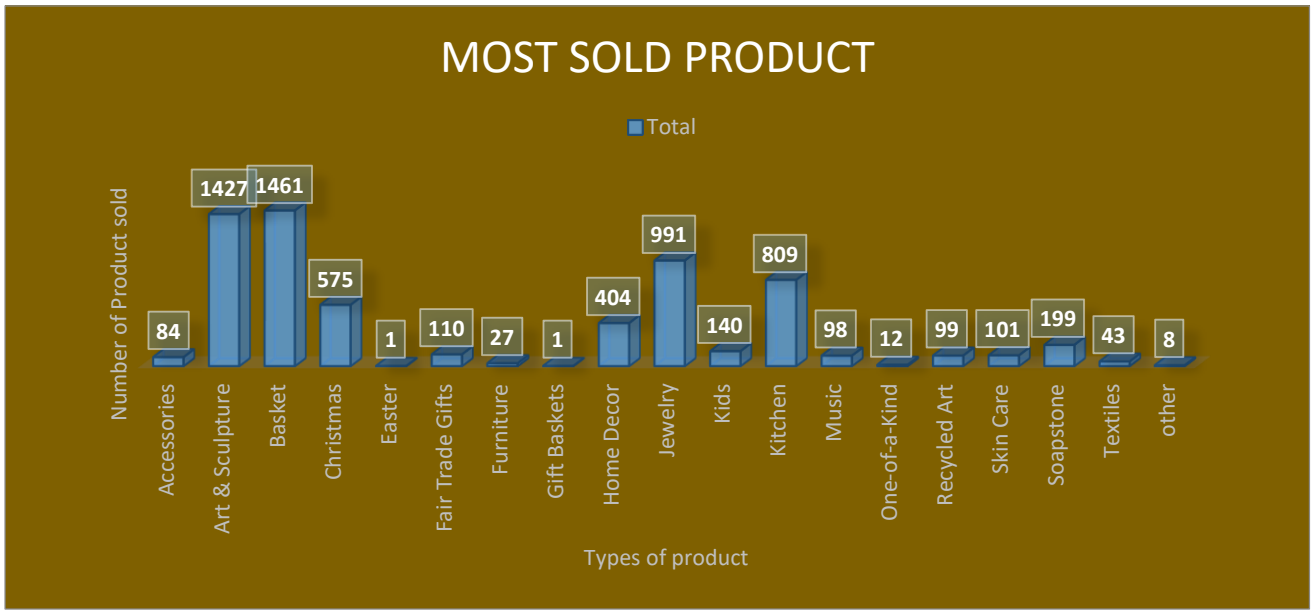
- Pivot table of sum of net quantity by product
- With the help of this 3D clustered column chart is plotted.

DATA ANALYSIS

Product	Sum of Net Quantity
Accessories	84
Art & Sculpture	1417
Basket	1461
Christmas	575

Easter	1
Fair Trade Gifts	110
Furniture	27
Gift Baskets	1
Home Decor	404
Jewelry	991
Kids	140
Kitchen	809
Music	98
One-of-a-Kind	12
Recycled Art	99
Skin Care	101
Soapstone	199
Textiles	43
(blank)	18
Grand Total	**6590**

VISUALIZATION



RESULT ANALYSIS

Basket and Arts & sculpture are the most sold product.

OBJECTIVE 7

What is the total shipping cost of all months in each year?

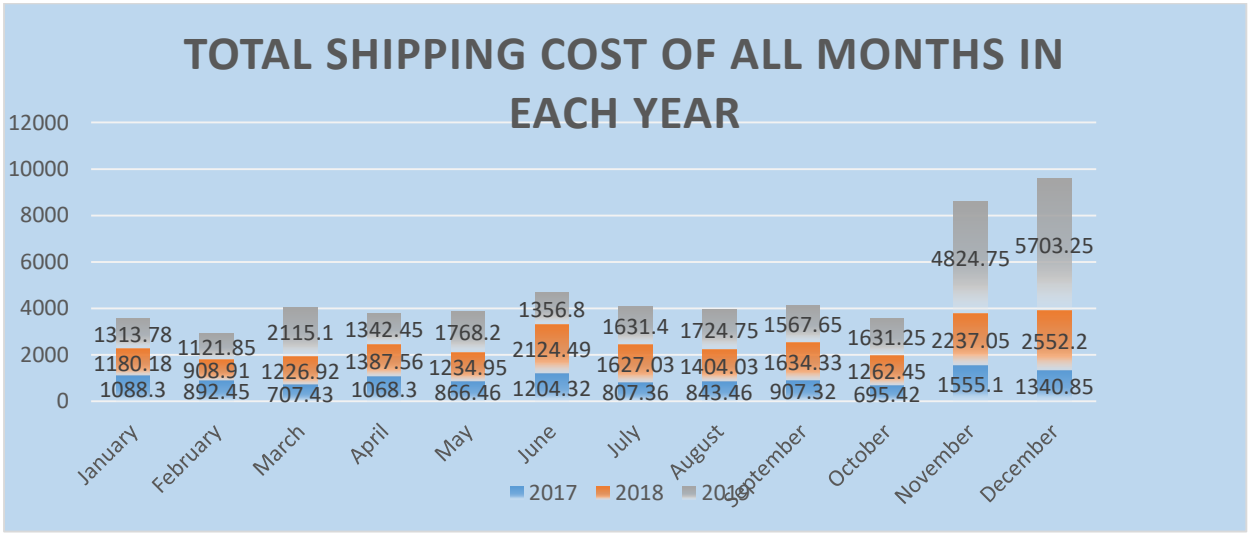
SPECIFIC REQUIREMENT/FUNCTIONS AND FORMULA

- Pivot table of total shipping cost in each month of 2017 ,2018 and 2019
- With the help of this 2D stacked column chart is plotted

DATA ANALYSIS

Sum of Shipping	Year			
Month	2017	2018	2019	Grand Total
January	1088.3	1180.18	1313.78	3582.26
February	892.45	908.91	1121.85	2923.21
March	707.43	1226.92	2115.1	4049.45
April	1068.3	1387.56	1342.45	3798.31
May	866.46	1234.95	1768.2	3869.61
June	1204.32	2124.49	1356.8	4685.61
July	807.36	1627.03	1631.4	4065.79
August	843.46	1404.03	1724.75	3972.24
September	907.32	1634.33	1567.65	4109.3
October	695.42	1262.45	1631.25	3589.12
November	1555.1	2237.05	4824.75	8616.9
December	1340.85	2552.2	5703.25	9596.3
Grand Total	**11976.77**	**18780.1**	**26101.23**	**56858.1**

VISUALIZATION



RESULT ANALYSIS

Total shipping cost is Maximum in December.

OBJECTIVE 8

What is the net quantity of each item?

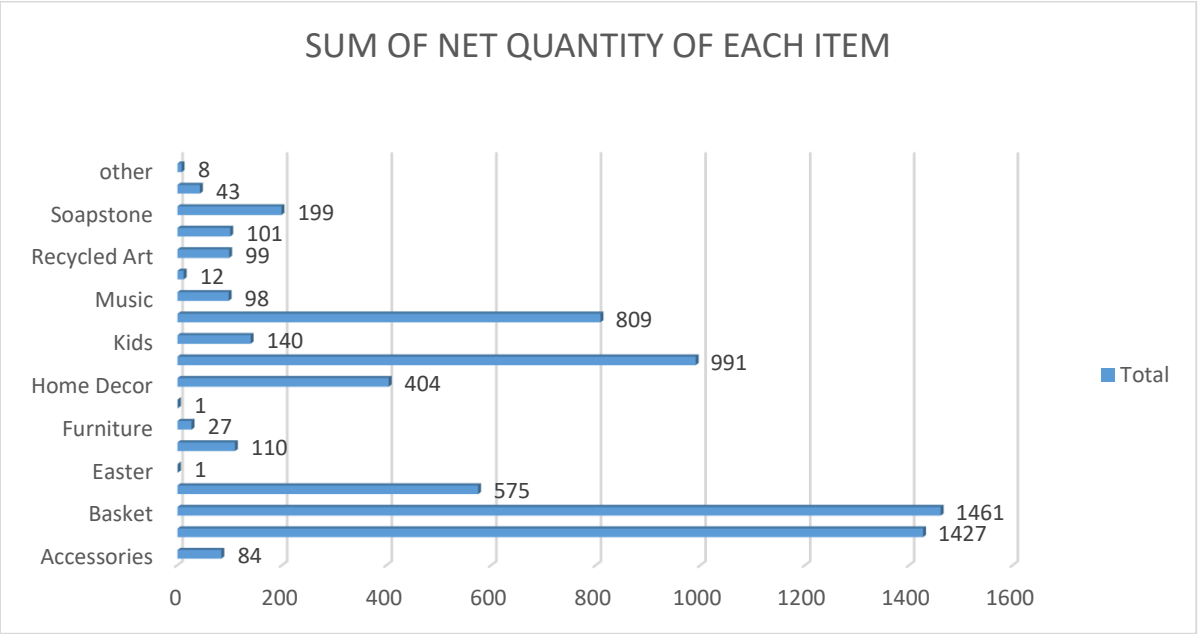
SPECIFIC REQUIREMENT/FUNCTIONS AND FORMULA

- Pivot table of net quantity of each product
- With the help of this clustered 3-D bar graph is plotted.

DATA ANALYSIS

Product Type	Sum of Net Quantity
Accessories	84
Art & Sculpture	1427
Basket	1461
Christmas	575
Easter	1
Fair Trade Gifts	110
Furniture	27
Gift Baskets	1
Home Decor	404
Jewelry	991
Kids	140
Kitchen	809
Music	98
One-of-a-Kind	12
Recycled Art	99
Skin Care	101
Soapstone	199
Textiles	43
other	8
Grand Total	**6590**

VISUALIZATION



RESULT ANALYSIS

Basket has the most net quantity.

OBJECTIVE 9

What is the Sum of total sales of every month in each year?

SPECIFIC REQUIREMENT/FUNCTIONS AND FORMULA

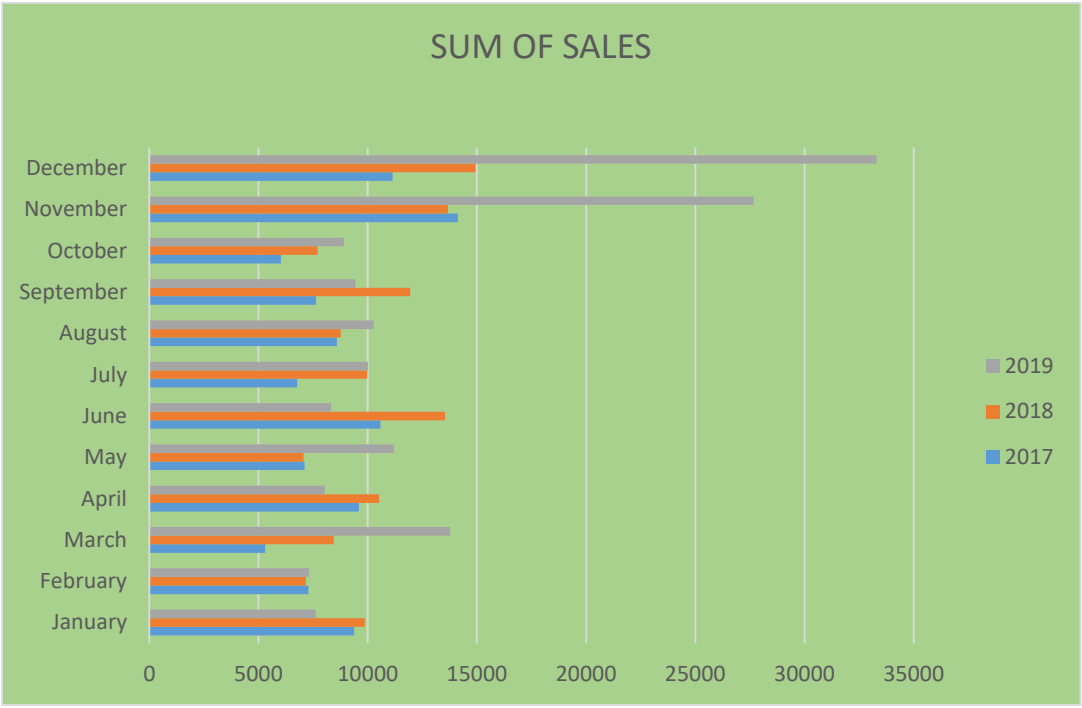
- Pivot charge of sum of total sales of every months in each year
- With the help of this 2-D bar graph is plotted.

DATA ANALYSIS

Sum of Total Sales	Year			
Month	2017	2018	2019	Grand Total
January	9371.95	9859.83	7615.91	26847.69
February	7280.05	7158.61	7318.15	21756.81

March	5296.53	8434	13769.75	27500.28
April	9600.9	10521.63	8024.05	28146.58
May	7103.91	7059.58	11216.2	25379.69
June	10573.87	13543.36	8327.13	32444.36
July	6766.66	9974.13	10014.78	26755.57
August	8583.56	8764.81	10278.21	27626.58
September	7639.62	11941.03	9436.86	29017.51
October	6022.42	7704.25	8911.53	22638.2
November	14124.95	13670.9	27681.3	55477.15
December	11132.85	14936.15	33306.46	59375.46
Grand Total	**103497.3**	**123568.3**	**155900.3**	**382965.88**

VISUALIZATION

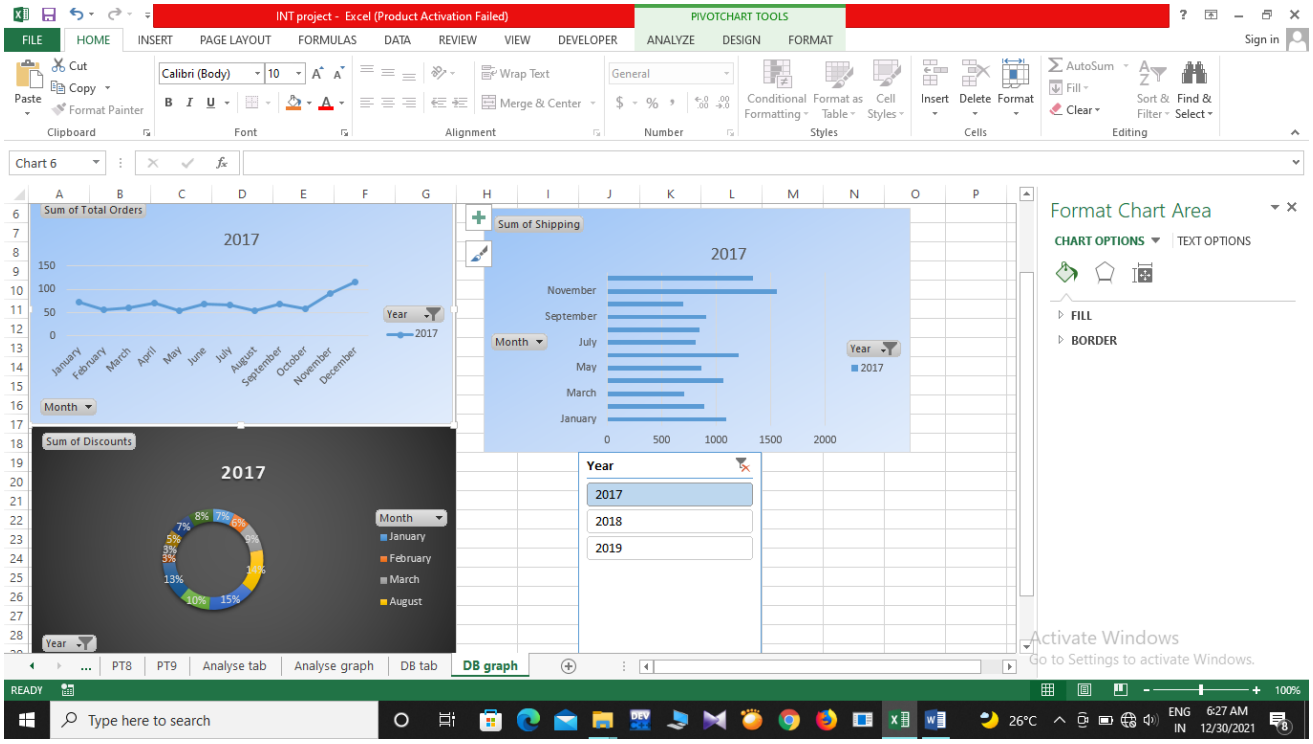


RESULT ANALYSIS

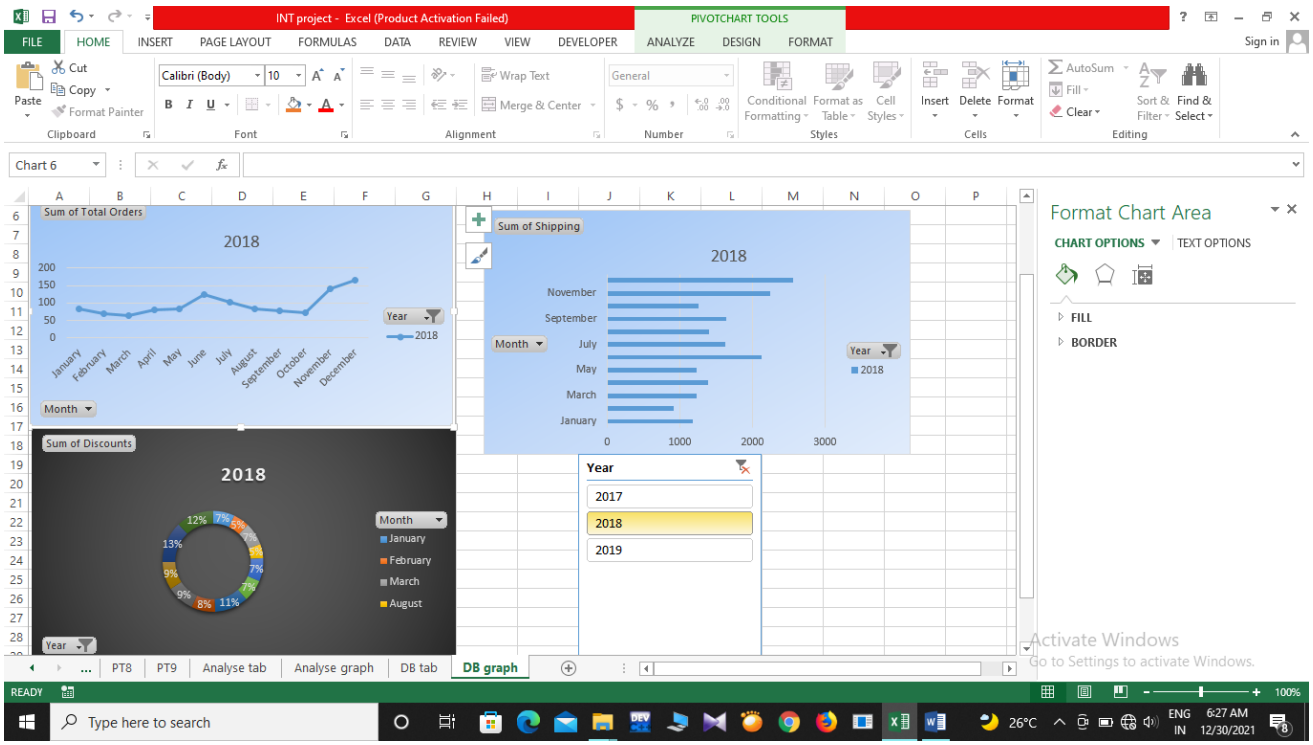
December has the max sum of total sales had happened.

DASHBOARD

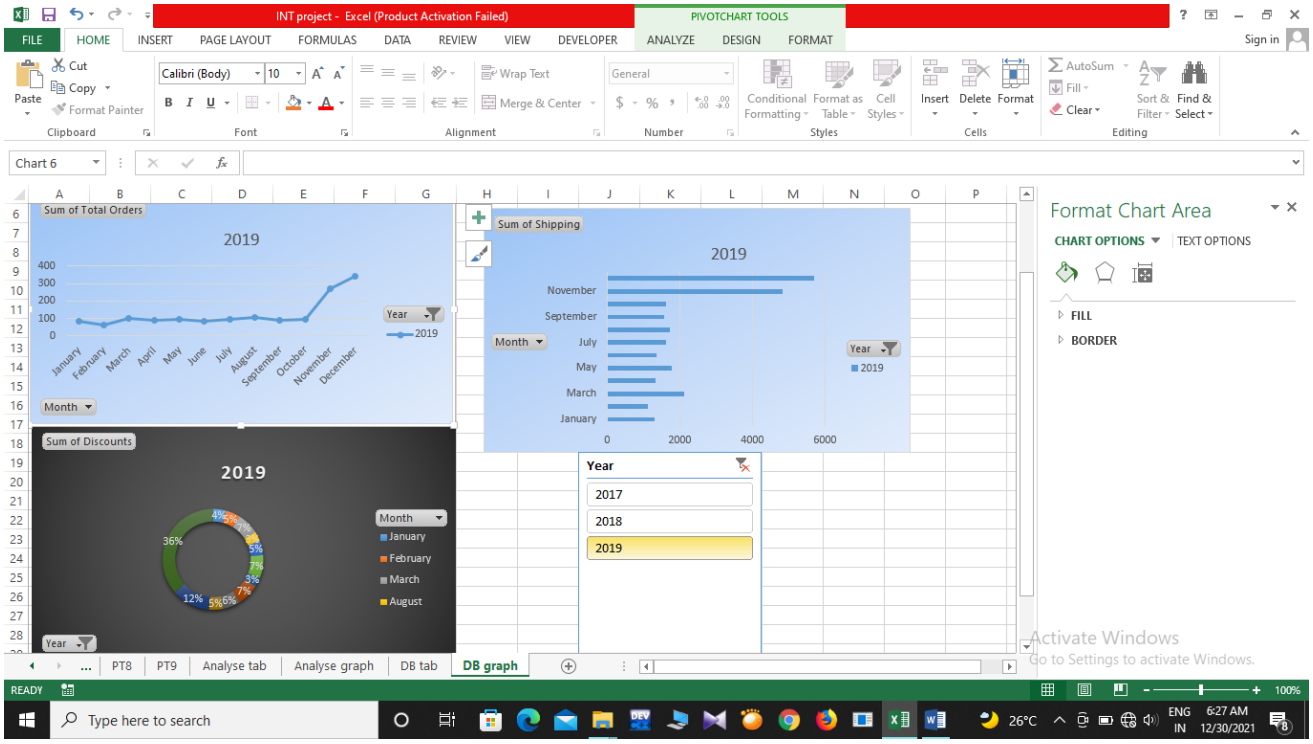
We can select each year to analyse data easily.



Sum of shipping,Sum of discounts and sum of total orders in **2017**



Sum of shipping,Sum of discounts and sum of total orders in **2018**



Sum of shipping,Sum of discounts and sum of total orders in **2019**

Here I made the Dashboard with the help of 3 pivot table of

- Sum of shipping
- Sum of Discount
- Sum of total orders

Pivot Table of Sum of Total orders (table 1)

Sum of Total Orders	Year
Month	2018
January	83
February	69
March	64
April	81
May	82
June	124
July	102
August	82
September	79
October	71
November	140
December	164
Grand Total	**1141**

Pivot table of Sum of Total shipping(table 2)

Sum of Shipping	Year
Month	2018
January	1180.18
February	908.91
March	1226.92
April	1387.56
May	1234.95
June	2124.49
July	1627.03
August	1404.03
September	1634.33
October	1262.45
November	2237.05
December	2552.2
Grand Total	**18780.1**

Pivot table of Total discount(table 3)

Sum of Discounts	Year
Month	2018
January	217.1
February	161.35
March	226.82
April	232.28
May	221.25
June	335.4
July	237.87
August	140.57
September	276.15
October	277.95
November	414.45
December	371.2
Grand Total	**3112.39**

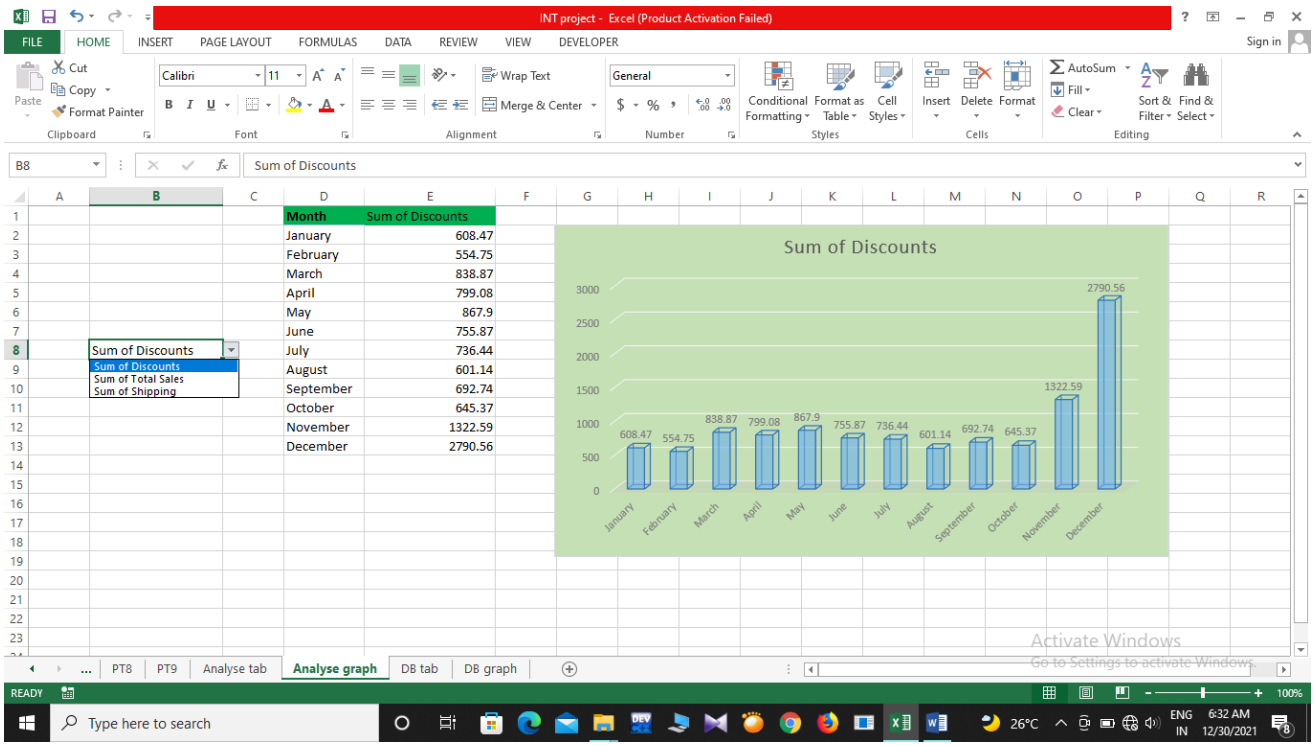
In these tables add slicer tool to filter the data according to the year

1. Click anywhere in the pivot table.
2. On the Home tab ,go to **Insert-→Slicer**
3. In the Insert slicers dialog box,select the **year** check boxes for the fields I want to Display, then select **OK**
4. A slicer will be created for every field that I selected.Clicking any of the the slicer buttons will automatically apply that filter to the linked pivot table.

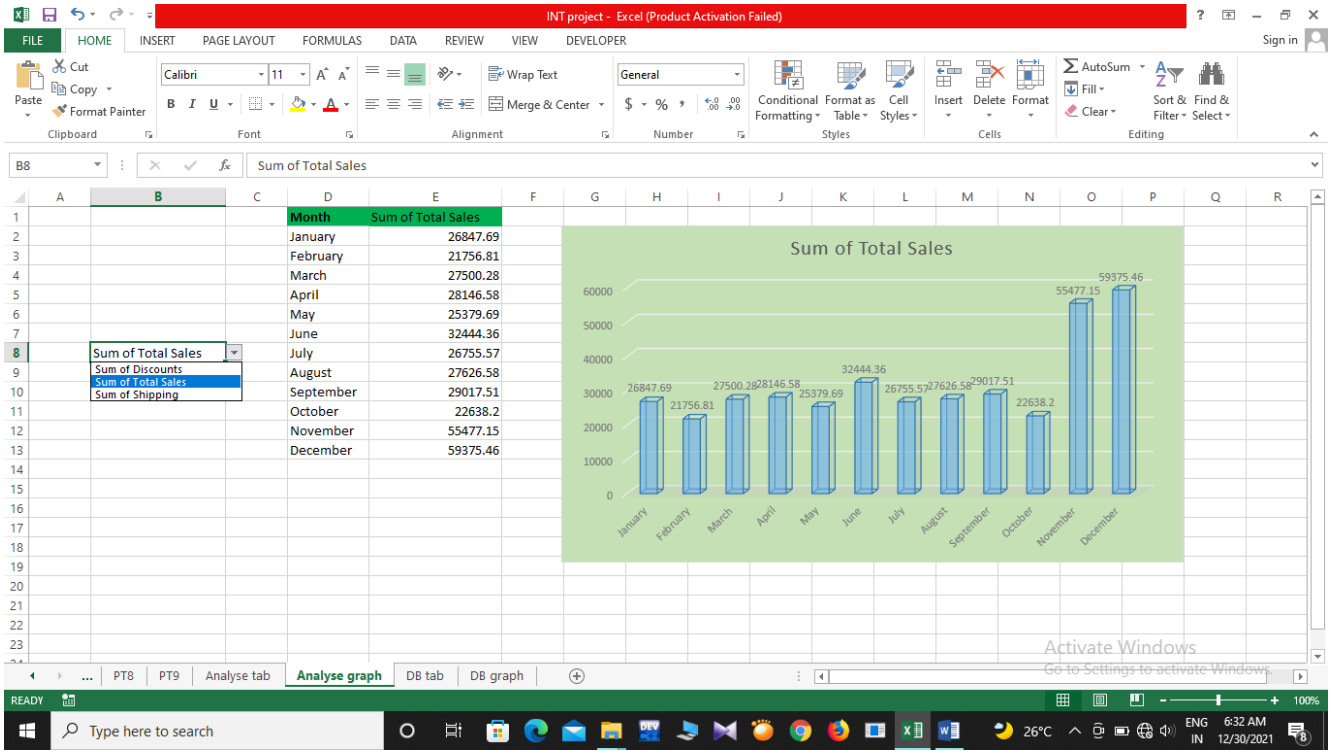
5. Here I connected 3 pivot tables like table 1,table 2 ,table 3 so to connect a slicer to more than one pivot table go to **Slicer →Report connections** →check the pivot table 1,2,3 to include then select **OK**

ANALYSE

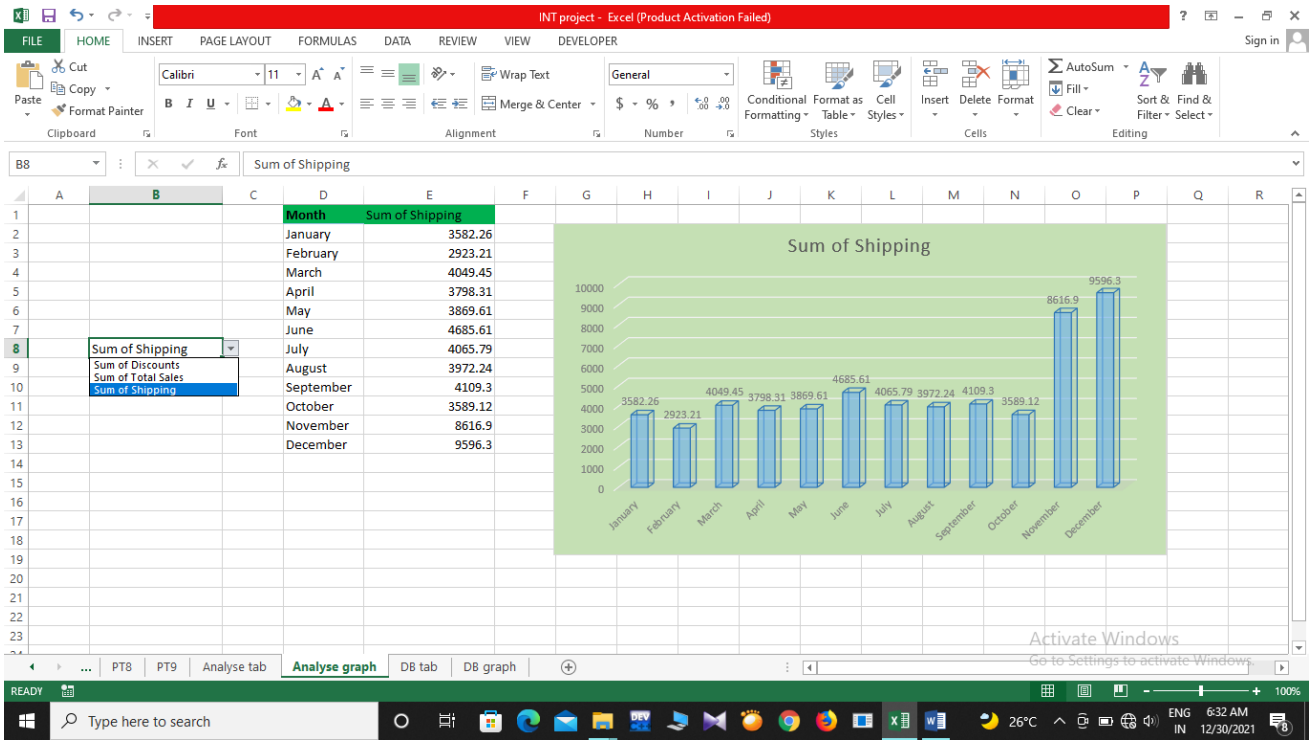
Here we can easily analyse the total sales, total disounts and total shipping charges of 2017, 2018, and 2019 by using **Data Validation Tool**



Most Discount had happened in the month of December in 2017 ,2018 and 2019



More no. of total sales had happened in the month of December in 2017 ,2018 and 2019



More shipping charge had happened in the month of December in 2017 ,2018 and 2019

For making this Analyse Chart we have to make the Pivot Table of following.

Month	Sum of Discounts	Sum of Total Sales	Sum of Shipping
January	608.47	26847.69	3582.26
February	554.75	21756.81	2923.21
March	838.87	27500.28	4049.45
April	799.08	28146.58	3798.31
May	867.9	25379.69	3869.61
June	755.87	32444.36	4685.61
July	736.44	26755.57	4065.79
August	601.14	27626.58	3972.24
September	692.74	29017.51	4109.3
October	645.37	22638.2	3589.12
November	1322.59	55477.15	8616.9
December	2790.56	59375.46	9596.3

For making List: **Data → Data Validation → Settings →** Select **list** from Validation Criteria →

Add **sheet location →** Select **OK**

For making Connection of List and Table:

=INDEX('Analyse tab'!$B$2:$D$13,MATCH('Analyse graph'!D2,'Analyse tab'!$A$2:$A$13,0),MATCH('Analyse graph'!$E$1,'Analyse tab'!$B$1:$D$1,0))

CONCLUSION

From the analysis we get that basket is most sold item as well as it have high discount rate. When checking for the better shipping rate we can see that different year there are different month. But from 2017-2019 we can say that December have minimum shipping rate and December have the most discount sale. Highest number of order is happen in December but we can see there is lot of returns also happened in December in this month most of the items were baskets that mean when we buying basket we want assure quality. From 2017-2019 we see that there is a drastic change in online purchasing . Every year the amount of customers is increasing which mean world is moving to platform of ecommerce. From the data we see that household item are more trending were customer trust is patched on it.

REFERENCES

Youtube

Kaggle

Wikipedia.com

Google.com