INTRODUCTION TO DATA MANAGEMENT

PROJECT REPORT

(Project Semester August-December 2020)

U.S. SUPERSTORE DATA ANALYSIS

Submitted by

Ishika Aggarwal

Registration No: 11904047

Programme and Section: P192-ND & KM007

Course Code: INT 217

Under the Guidance of

Sandeep Kaur : 23614 and Assistant Professor

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Ishika Aggarwal bearing Registration no. 11904047 has completed INT 217 project titled, **“U.S. Superstore Data Analysis”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor: Ms. Sandeep Kaur
Designation of the Supervisor: Assistant Professor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 10-12-2021

DECLARATION

I, Ishika Aggarwal, student of Integrated B.Tech – M.Tech 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.

Date: 10-12-21 Signature: Ishika

Registration No. : 11904047 Name of the student: Ishika Aggarwal

ACKNOWLEDGEMENT

A project work is a combination of views, ideas, suggestions and contribution of many people. Thus, I feel obliged as well as pleasure to extend my gratitude towards all those contributors who helped me during my endeavor of making this project.

I would like to thank my Professor as well as supervisor Ms. Sandeep Kaur for providing me with this opportunity to make a minor project as well as implement my practical knowledge by making a dashboard on the project titled, **“U.S. Superstore Data Analysis”** by providing me constant support as well as constructive criticism which helped me to understand my mistakes and improve my project thoroughly.

I would like to extend my gratitude to all those as well who responded to my queries in a well-defined manner and helped me acquiring knowledge without whom my project would not have been such a great learning experience.

Regards

Ishika Aggarwal

Section: KM007

Registration No: 11904047

TABLE OF CONTENTS

1. INTRODUCTION

1.1 *What is Excel?*

Microsoft Excel is a software program included in the Microsoft Office suite. It is used to create spreadsheets, which are documents in which data is laid out in rows and columns — like a big table.

Due to its extreme versatility and power, Excel has become one of the most-used software programs in the business world since its launch in 1985. Indeed, the personal computing renaissance of the 1980s and 1990s was largely driven by the many uses of Excel and other spreadsheet software.

1.2 *Some Important Terms*

1.2.1 Dataset: A data set is an ordered collection of data. While handling the data, the data set can be a bunch of tables, schema and other objects. The data are essentially organized to a certain model that helps to process the needed information. The set of data is any permanently saved collection of information that usually contains either case-level, gathered data, or statistical guidance level data.

1.2.2 Spreadsheet: A spreadsheet is a special way of organizing data into rows and columns to make it simpler to read and manipulate. A default spreadsheet is often named as Sheet1, Sheet2, Sheet3 and so on.

1.2.3 Column: A column is a vertical series of cells in a chart, table, or spreadsheet.

1.2.4 Row: A row is a series of data banks laid out horizontally in a table or spreadsheet.

1.2.5 Workbook: A workbook is a collection of one or more spreadsheets also called worksheets in a single file. A default workbook is often named as Book1, Book2, Book3 and son on.

1.2.6 Formula Bar: The formula bar is a section in Microsoft Excel and other spreadsheet applications in which it allows the contents of the current cell and also allows us to create and view formulas.

1.2.7 Pivot Table: Pivot table in excel is used to categorize, sort, filter, and summarize any length of data table which we want to get count, sum, values either in tabular form or in the form of 2 column sets.

1.2.8 Pivot Chart: A pivot chart is especially useful for user when dealing with tremendous amounts of data by quickly reorganizing and visualizing data in an understandable manner and facilitate the entire process.

1.2.9 Slicers: Slicers in Excel are software filters used along with excel tables or pivot tables over a large amount of data. Not just filtering out the data, but slicers also help you with an easy understanding of the information being extracted and displayed on the screen.

1.2.10 Timeline: Timeline in Excel actually represents the time span from the start to end on a bar. So for this, we should have any time frame such as Dates, Months, Minutes, Hours, etc

1.2.11 Hyperlinks: Hyperlinks in Excel allow users to create a shortcut way to reach any certain worksheet, file, folder or webpage. It helps us to reach to any specific folder or link quickly.

1.2.12 Data Analytics: Data analytics is the science of analysing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.

1.2.13 Data Visualization: Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

1.2.14 Dashboard: Excel dashboards make it easy to perform quick overviews of data reports rather than going through large volumes of data. Overviews help in making quick and urgent decisions since one can skim through a lot of information at once and within a short time.

2. SCOPE OF THE ANALYSIS

2.1 *Project Description*

The report gives us an overview on the sales activity of the store over a period of 4 years (2014-2018). For the company's better sales performance and steady growth, it is important to know what are the factors that are contributing to the growth and which factor is leading to constant losses if any. Sales reporting is important to derive insights and examine whether the sales operations are successful and the strategies that were build as the foundation are proving to be efficient. We have chosen this dataset to analyse and generate some insights on the different categories of products sold by the Superstore. The data collected provides us with the basic information on the sales of Superstore, the orders received, and the returns made.

2.2 Columns Present In The Dataset

- Row ID
- Order ID
- Order Date
- Ship Date
- Ship Mode
- Customer ID
- Customer Name
- Segment
- Country
- City
- State
- Postal Code
- Region
- Product ID
- Category
- Sub-Category
- Product Name
- Sales
- Quantity
- Discount
- Profit

2.3 Scope

In order to have a better insight for future sales it is extremely important to analyze the data so that the superstores can run in an efficient manner without any or negligible loss as the main priority of a superstore to maintain profit margins and have better performance than their last one.

It is also important to have made strategies ready on the basis of sales report in order to have a better performance.

3. SOURCE OF THE DATASET

Source Of The Dataset: https://www.kaggle.com/juhi1994/superstore

Author: Juhi Badiyani

Last Updated: 2 years ago

4. ETL PROCESS

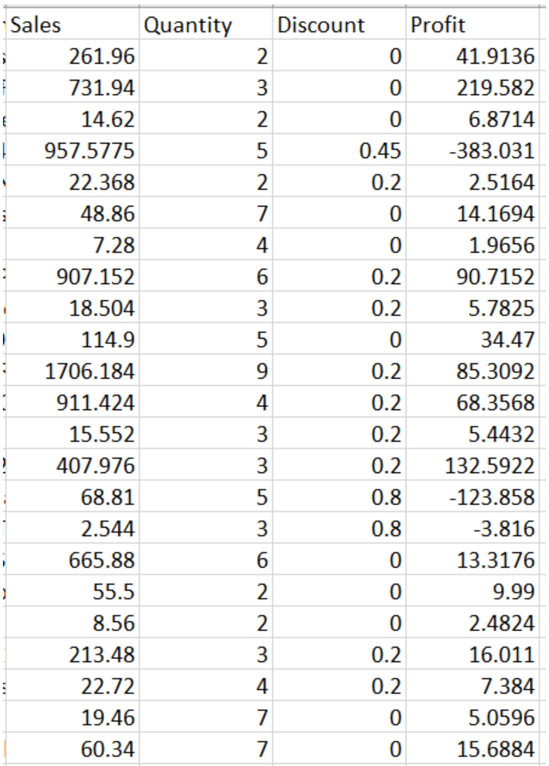
ETL is the process of extracting huge volumes of data from a variety of sources and formats and converting it to a single format before putting it into a database or destination file.

4.1 Extraction : Raw data is extracted from various sources like Kaggle, mydata.gov.in, etc. in the format of CSV files. So we are going to extract the file from the raw data and transform it into a target file and load it in the output.

4.2 Transformation: Dataset has been cleaned in order to make it into more readable format in excel and summarized into various tables according to our objectives in the form of pivot tables and charts.

Like Price column had ‘General’ format but it has been converted into ‘Accounting’ format with dollars ‘$’ as currency symbols.

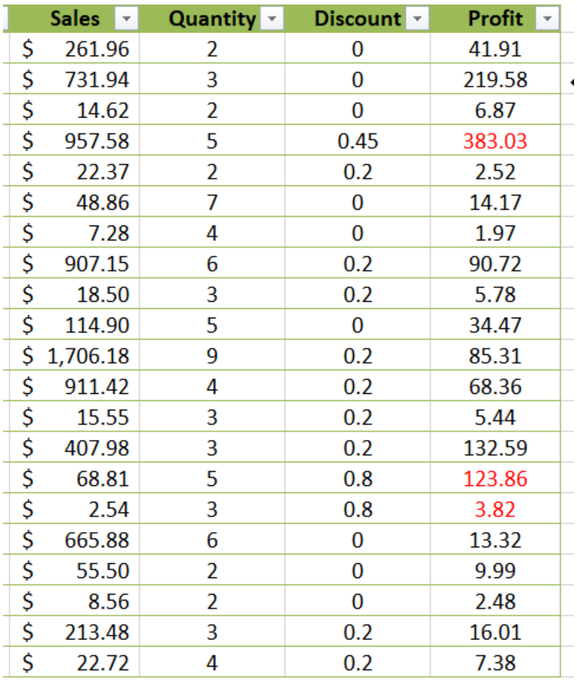
We also converted ‘General’ format to ‘Number’ format in Profit column and indicated all the negative values in red color for easy understanding.



Sales	Quantity	Discount	Profit
261.96	2	0	41.9136
731.94	3	0	219.582
14.62	2	0	6.8714
957.5775	5	0.45	-383.031
22.368	2	0.2	2.5164
48.86	7	0	14.1694
7.28	4	0	1.9656
907.152	6	0.2	90.7152
18.504	3	0.2	5.7825
114.9	5	0	34.47
1706.184	9	0.2	85.3092
911.424	4	0.2	68.3568
15.552	3	0.2	5.4432
407.976	3	0.2	132.5922
68.81	5	0.8	-123.858
2.544	3	0.8	-3.816
665.88	6	0	13.3176
55.5	2	0	9.99
8.56	2	0	2.4824
213.48	3	0.2	16.011
22.72	4	0.2	7.384
19.46	7	0	5.0596
60.34	7	0	15.6884

← Original CSV File (Raw Data)

Table 4.1



Sales	Quantity	Discount	Profit
$ 261.96	2	0	41.91
$ 731.94	3	0	219.58
$ 14.62	2	0	6.87
$ 957.58	5	0.45	383.03
$ 22.37	2	0.2	2.52
$ 48.86	7	0	14.17
$ 7.28	4	0	1.97
$ 907.15	6	0.2	90.72
$ 18.50	3	0.2	5.78
$ 114.90	5	0	34.47
$ 1,706.18	9	0.2	85.31
$ 911.42	4	0.2	68.36
$ 15.55	3	0.2	5.44
$ 407.98	3	0.2	132.59
$ 68.81	5	0.8	123.86
$ 2.54	3	0.8	3.82
$ 665.88	6	0	13.32
$ 55.50	2	0	9.99
$ 8.56	2	0	2.48
$ 213.48	3	0.2	16.01
$ 22.72	4	0.2	7.38

← Formatted Excel File

After cleaning the data and making it into readable format

Table 4.2

4.3 Loading: The load stage of the ETL process depends largely on what you intend to do with the data once it’s loaded into the data warehouse. Uses could include:

- Layering a business intelligence or analytics tool on top of the warehouse.
- Analyzing and visualizing the data in the form of Dashboard which is our intended target in this project.
- Creating a tool for site search.
- Building a machine learning algorithm to detect fraud

5. OBJECTIVES

Objective -1

a. Sales Based On States.
b. Sales Trend Over Time.
c. Sales & Profit Based On Region.

Objective – 2

a. Most Loyal Customers Based On Region (Top 5 customers Region-wise)

Objective -3

a. Top 5 products According To The Category & Region.
b. Sales & Profit Based On Category and Sub- Category.

Objective -4

a. Discount Based On Segment and Category.
b. Most Popular Region Based On Segment & Ship Mode.

Objective – 5

a. Most Preferred Ship Mode Based On Region & Ship Year.

6. ANALYSIS OF THE DATASET

6.1. Sales Based On States

a. Introduction: The analysis shows state wise sales of the U.S. Superstore.

b. Specific Requirements/Functions and Formulas:

- Pivot table of the dataset.
- GETPIVOTDATA function.
- Map of U.S. from Maps.

c. Analysis Results: We can easily analyse with the help of this Map that 'California' had the highest sum of sales.

d. Visualization: Red color represents highest sales; olive green color represent average sales and orange color represents less sales.

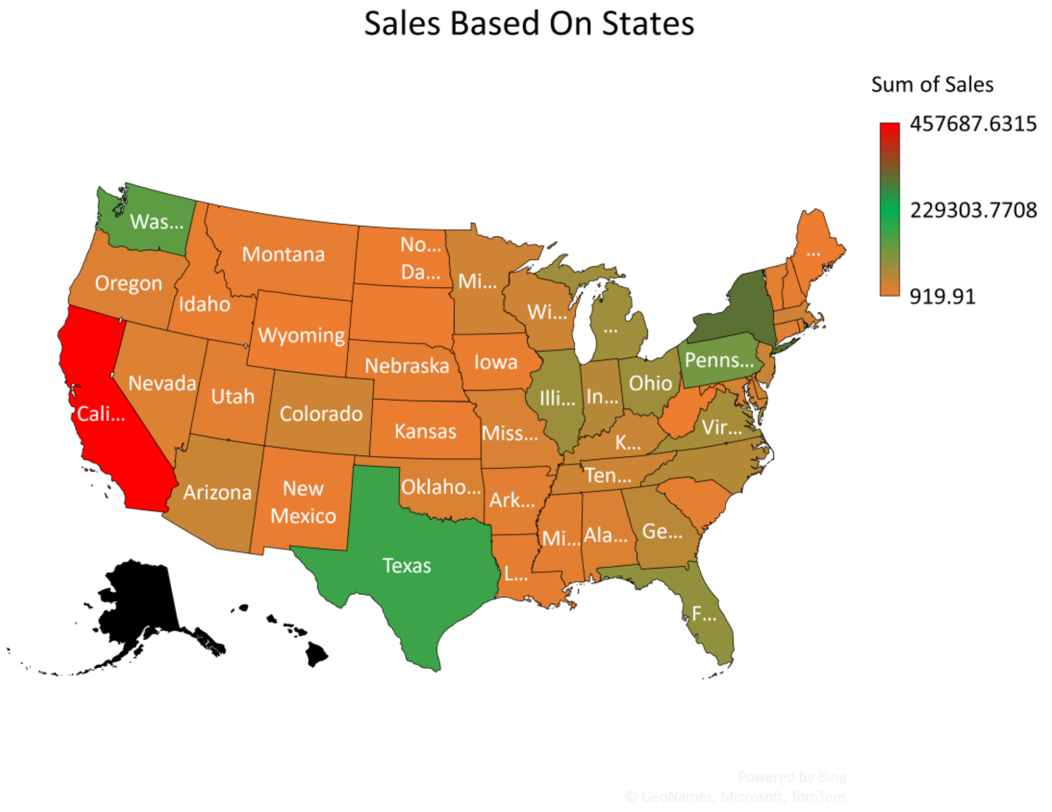


Fig 6.1

6.2 Sales Trend Over Time

a. Introduction: This analysis shows sum of sales as well as profit over time.

b. Specific Requirements/Functions and Formulas:

- Pivot table of the dataset.
- Pivot Combo Chart of area as well as line chart.

c. Analysis Results: We can analyze from this chart that sales is highest in ‘2017’ year and ‘Quarter 3’ while Profit is highest in ‘2016’ year and ‘Quarter 4’.

d. Visualization:

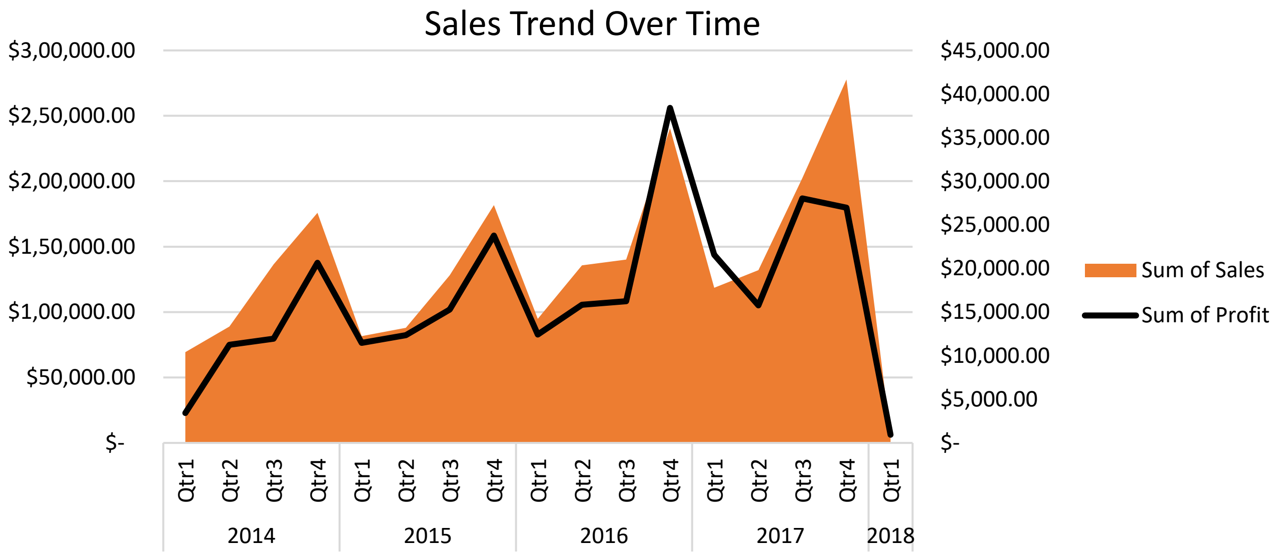


Fig 6.2 – For All Order Years

6.3 Sales & Profit Based On Region

a. Introduction: This analysis shows sum of sales and profit over region.

b. Specific Requirements/Functions and Formulas:

- Pivot table of the dataset.
- Stacked column chart.

c. Analysis Results: We can easily analyse through this chart that ‘West’ region has the highest sum of sales as well as profit out of all regions.

d. Visualization:

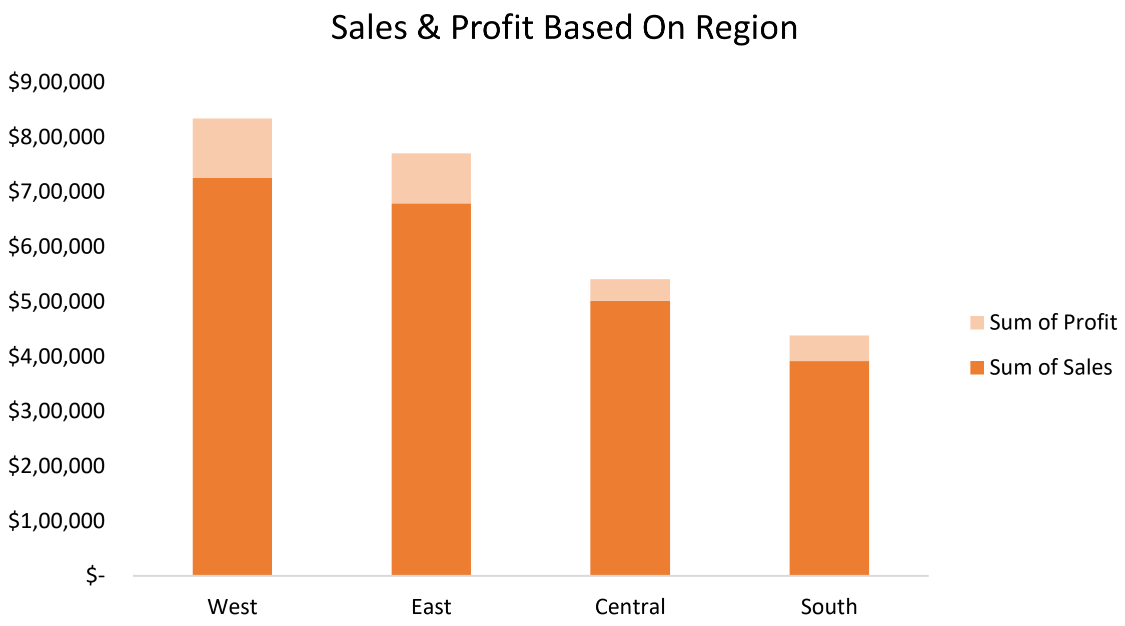


Fig 6.3 – For All Regions

6.4 Most Loyal Customers Based On Region

a. Introduction: This analysis how top 5 customers who have been loyal for quite a long time region wise.

b. Specific Requirements/Functions and Formulas:

- Pivot table of the dataset.
- Clustered column chart.

c. Analysis Results: We can analyze from this chart that 'Sean Miller' has been the most loyal customers of all the regions.

Central Region Top Customer: Tamara Chand

West Region Top Customer: Raymond Buch

East Region Top Customer: Tom Ashbrook

South Region Top Customer: Sean Miller

d. Visualization:

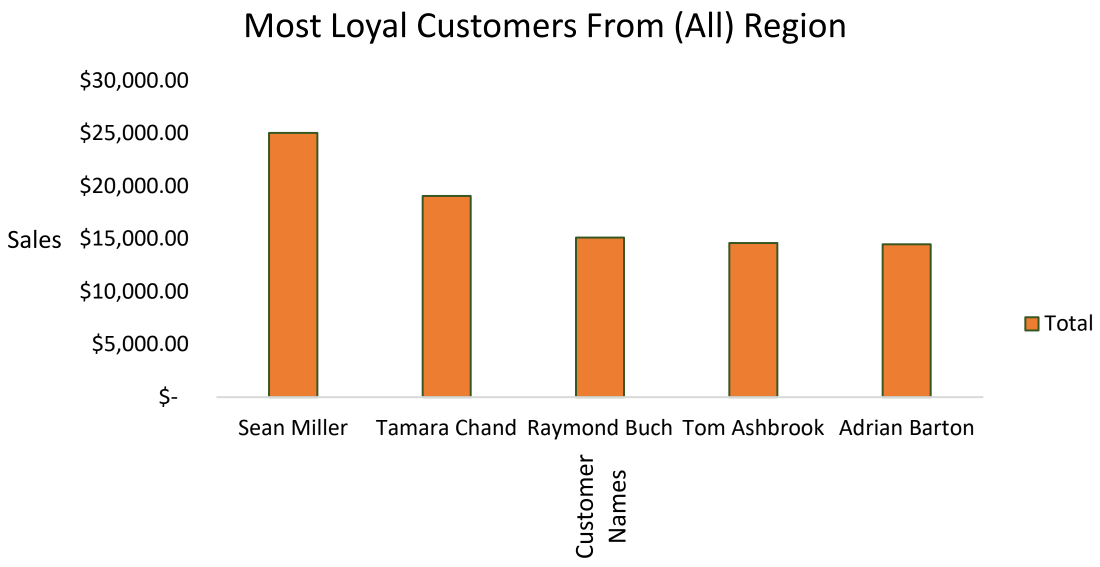


Fig 6.4

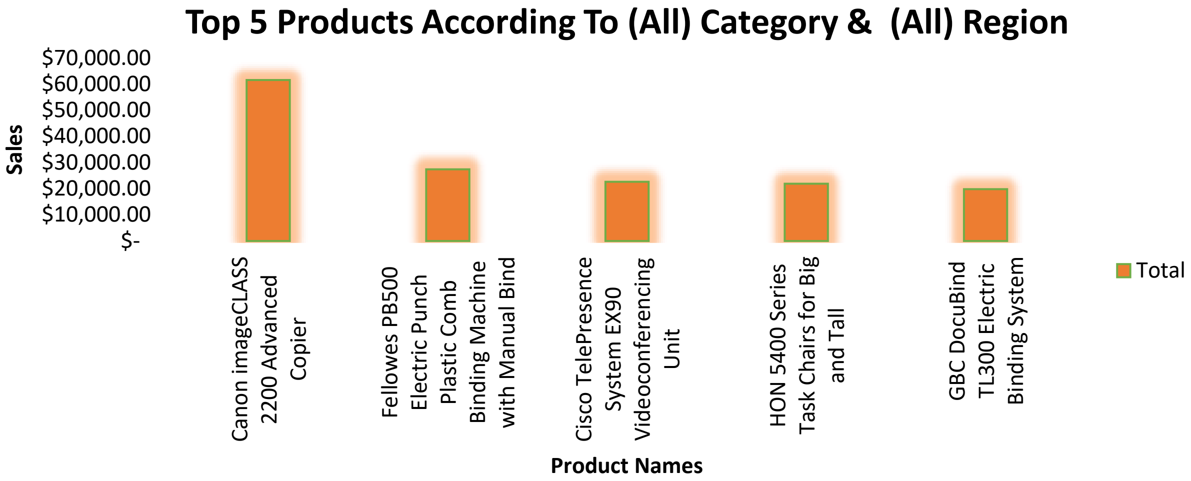
6.5 Top 5 products According To The Category & Region

a. Introduction: This analysis represents top 5 products according to the category as well as region.

b. Specific Requirements/Functions and Formulas:
   - Pivot table of the dataset.
   - Clustered column chart.

c. Analysis Results: We can easily analyze from this chart 'Canon imageCLASS 2200 Advanced Copier' is the topmost product among all the categories as well as region.
   Central Region Top Product: Canon imageCLASS 2200 Advanced Copier
   East Region Top Product: Canon imageCLASS 2200 Advanced Copier
   West Region Top Product: Canon imageCLASS 2200 Advanced Copier
   South Region Top Product: Cisco TelePresence System EX90 Videoconferencing Unit

d. Visualization:



6.6 Sales & Profit Based On Category & Sub- Category

a. Introduction: This analysis represents sum of sales and profit based on category as well as sub category.

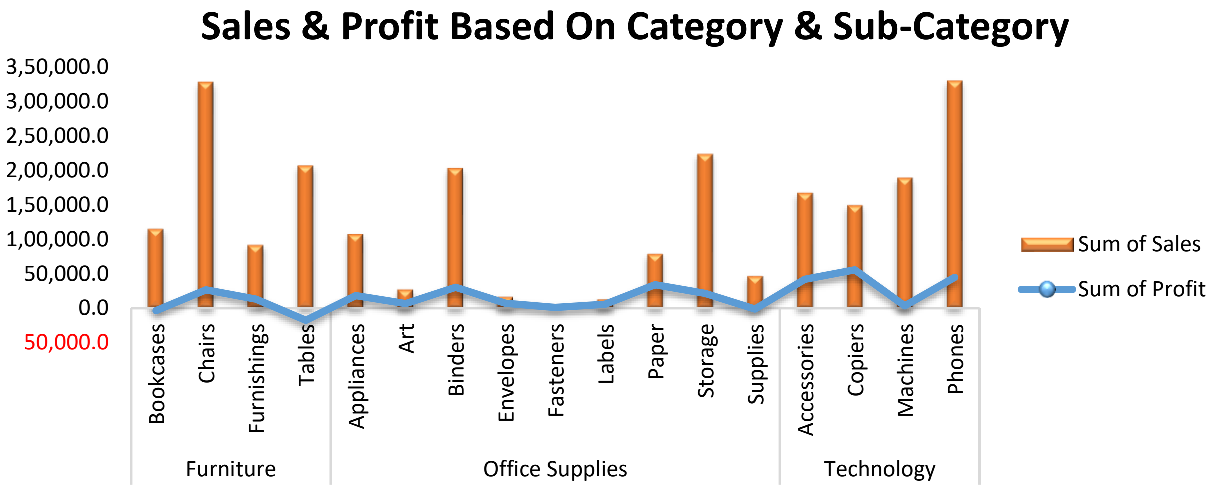
b. Specific Requirements/Functions and Formulas:

   - Pivot table of the dataset.
   - Pivot combo chart of clustered column as well as line chart.

c. Analysis Results: We can easily analyze that:

   - In ‘Furniture’ category, ‘Chairs’ sub-category has highest sales as well as profit.
   - In ‘Office Supplies’ category, ‘Storage’ sub-category has highest sales but ‘Paper’ sub-category has highest profit.
   - In ‘Technology’ category, ‘Phones’ sub-category has highest sales but ‘Copiers’ sub-category has made highest profit.

d. Visualization:



6.7 Discount Based On Segment & Category

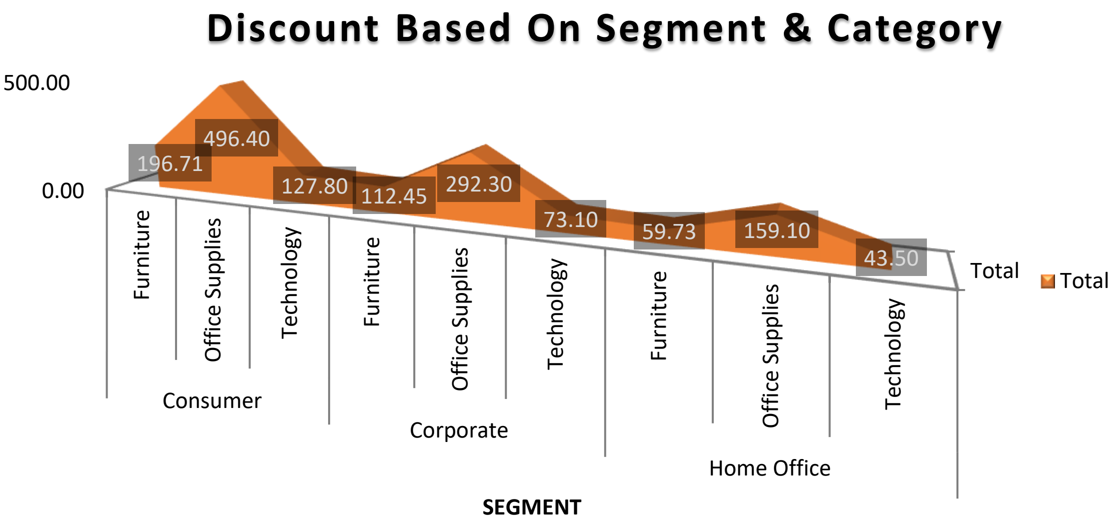
a. Introduction: This analysis represents discount on the basis of segment and category.

b. Specific Requirements/Functions and Formulas:

- Pivot table of the dataset.
- 3-D Area chart.

c. Analysis Results: We can infer from this chart that 'Office Supplies' category in the 'Consumer' segment had highest discounts on their products among all other segments as well as categories.

d. Visualization:



6.8 Most Popular Region Based On Segment & Ship Mode

a. Introduction: This analysis represents the most popular region based on segment and ship mode.

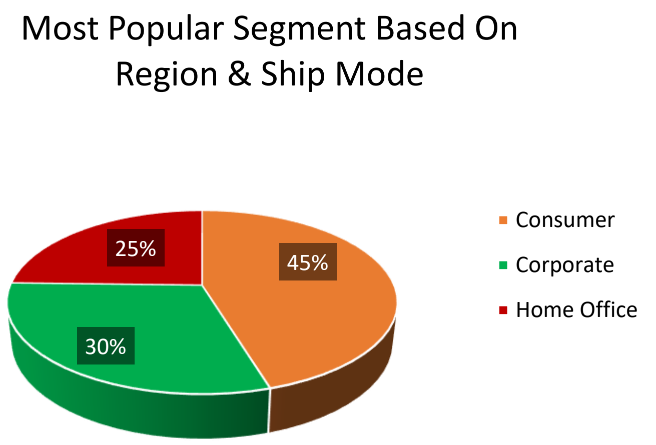
b. Specific Requirements/Functions and Formulas:

- Pivot table of the dataset.
- Pie chart.

c. Analysis Results: We can clearly analyse that 'Consumer' segment is the most popular segment among all regions as well as ship mode.

- Central region's most popular segment: Consumer
- East region's most popular segment: Consumer
- West region's most popular segment: Consumer
- South region's most popular segment: Consumer & Corporate

d. Visualization:



6.9 Most Preferred Ship Mode Based On Region & Ship Year

a. Introduction: This analysis represents most preferred ship mode on the basis of region and ship year.

b. Specific Requirements/Functions and Formulas:

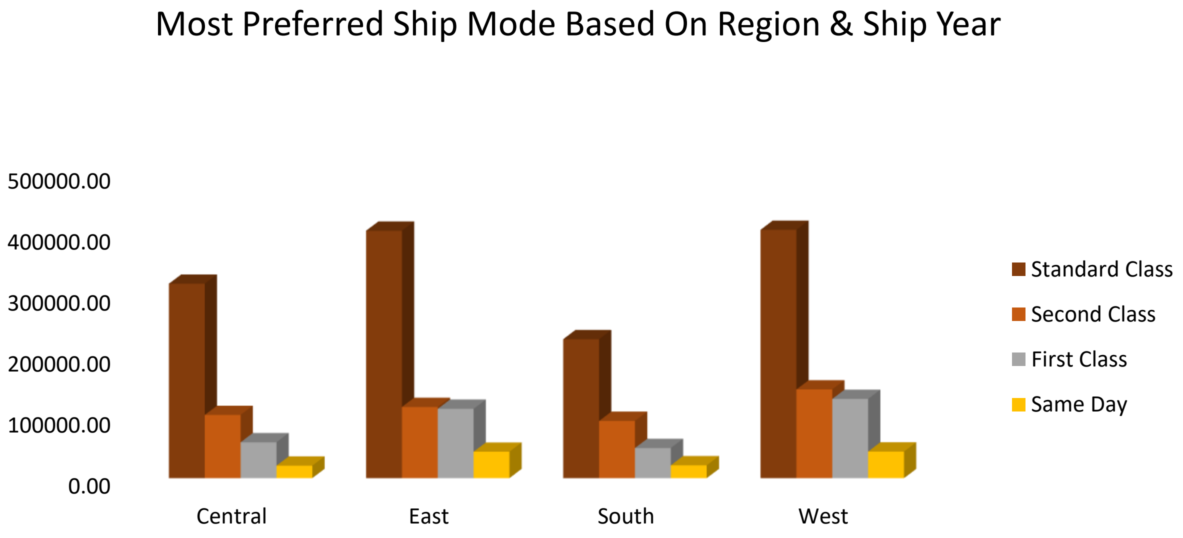
- Pivot table of the dataset.
- 3-D Clustered column

c. Analysis Results: We can infer from this chart 'Standard Class ' ship mode is the most preferred ship mode in all the regions.

- 2014: Standard Class
- 2015: Standard Class

- 2016: Standard Class
- 2017: Standard Class
- 2018: Standard Class

d. Visualization:



7. LIST OF ANALYSIS WITH RESULTS

In this section we can see that we are able to answer so many questions due to this analysis.

- *Which state had highest sales?*

  Ans: California
- *In which year and quarter, sales and profit were highest?*

  Ans: Sales was highest in '2017' year and 'Quarter 3' while Profit was highest in '2016' year and 'Quarter 4'.
- *Which Region had highest sum of sales and profit?*

  Ans: 'West' region had the highest sum of sales as well as profit out of all regions.
- *Which customer have been most loyal to the Superstore and also state region wise topmost loyal customers?*

  And: 'Sean Miller' has been the most loyal customers of all the regions.
  - Central Region Top Customer: Tamara Chand
  - West Region Top Customer: Raymond Buch
  - East Region Top Customer: Tom Ashbrook
  - South Region Top Customer: Sean Miller
- *Which of the product had highest sales and state region wise topmost product as well?*

  Ans: 'Canon imageCLASS 2200 Advanced Copier' is the topmost product among all the categories as well as region.
  - Central Region Top Product: Canon imageCLASS 2200 Advanced Copier
  - East Region Top Product: Canon imageCLASS 2200 Advanced Copier
  - West Region Top Product: Canon imageCLASS 2200 Advanced Copier
  - South Region Top Product: Cisco TelePresence System EX90 Videoconferencing Unit
- *State all the sub categories which had the highest sales as well as profit according to the category.*

  Ans:
  - In 'Furniture' category, 'Chairs' sub-category has highest sales as well as profit.

  - In ‘Office Supplies’ category, ‘Storage’ sub-category has highest sales but ‘Paper’ sub-category has highest profit.
  - In ‘Technology’ category, ‘Phones’ sub-category has highest sales but ‘Copiers’ sub-category has made highest profit.

- *Which segment had the highest discounts on their products among all segments and categories?*

  Ans: ‘Office Supplies’ category in the ‘Consumer’ segment had highest discounts on their products among all other segments as well as categories.

- *Which segment is the most popular segment among all regions and ship mode?*

  Ans: ‘Consumer’ segment is the most popular segment among all regions as well as ship mode.

- *Which ship mode is the most preferred one among all regions?*

  Ans: ‘Standard Class’ ship mode is the most preferred ship mode in all the regions

8. REFERENCES

- www.google.com
- www.kaggle.com
- www.youtube.com
- www.smartsheet.com
- www.github.com
- www.support.microsoft.com

9. BIBLIOGRAPHY

I. Microsoft Excel 2016 Bible: The Comprehensive Tutorial Resource by John Walkenbach, Wiley

II. Fundamentals of Business Analytics by R.N. Prasad, Seema Acharya, Wiley