

MS in Business Analytics

Project Report

Advanced Data Resource Management

# [Lecturer](https://faculty.rpi.edu/taxonomy/term/94) : Jonathan McKinney

# Student : Venkata saikiran Melam

Table of Contents

**Heading Pg No:**

**Executive Summary** 4

Problem statement 4

Data source 4

Approach followed 3

Findings 3

**Define** 4

Data description 4

**Collect** 4

Data Transformation 4

**Organize** 4

Dimension and Fact table 4

Dimension 1. dimlocation 5

Dimension 2. dimOrder 5

Dimension 3. dimShip 5

Dimension 4. dimShipMode 5

Dimension 5. dimCustomer 5

Dimension 6. dimCategory 5

Fact Table 7. FactStore 5

Star schema 5

**Visualisation** 5

**Analyse** 5

**Conclusion** 5

**References** 5

Table of Figures

**Figure Pg No:**

Table 1: Overview of data 5

Figure 1: Data transformation 6

Figure 2: Star schema 9

Figure 3: Yearly sales and profit gained 9

Figure 4: Profit by region 10

Figure 5: Profit by state 10

Figure 6: Shipping mode class wise 11

Figure 7: Yearly sales growth 11

Figure 8: Top 10 customers by sales and profit 12

Figure 9: Sales by category and sub category 12

Figure 10: Sales by sub category (Technology) 13

Figure 11: Sales by sub category (Office Supplies) 13

Figure 12: Sales by sub category (Furniture) 14

Figure 13: Profit by segment and category 14

Figure 14: Sales sub category wise 15

Figure 15: Profit by sub category 15

**Executive Summary:**

**Problem Statement:**

E-commerce is a modern method of promoting products and services with fact growth in the field of Internet. Online stores, such as Amazon and eBay, are now established, and thousands of individual, family and large businesses have incorporated e-commerce into their business activities. The speed and ease with which the buying and selling process can take place, rank e-commerce high in the preferences and activities of internet users. This method allows people to buy products from books, toys, clothes and shoes to food, furniture and other household. With that in mind, an analysis based on products from an online store is worth it to be done, because year by year e-commerce is replacing the purchases that can be done with being physical in a store. In addition, while doing such us analysis, a business like this can identify the products have sold most or what kind of products customers want, so that the store can be equipped to make them more loyal.

**Data Source:**

The dataset belongs to Kaggle.com. It refers to a superstore which is an online store based in United States. It contains data from 2014-2018 and describes transactions made in this online store during these years.

**Approach Followed:**

The DCOVA framework (Define, Collect, Organize, Visualize, Analyze) approach learnt in class and described in Levine, Stephan, and Szabat (2016) was followed to analyze the data.

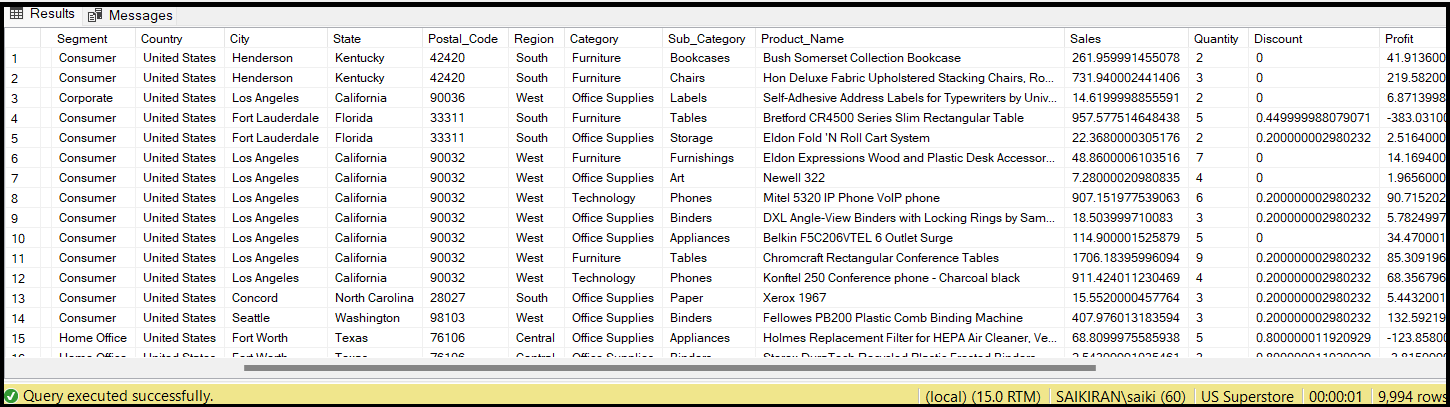
**Findings:**

* Sales of the store has increased every year resulting in high profit margin by the end of 2017.
* The most profitable region seems to be the Western with 38% overall sales registered. People residing in Western part of US tend to order more from the Superstore with California being the state with the most orders during the years.
* In contrast with California, Texas is the State with the lowest profit (negative). As regards to the ship mode preferences, the majority of people prefer standard class shipment (60%) while only 5% choose the same day shipping.
* There is an increasing tendency of sales by the years with peak season to be during autumn time.
* Although office supplies are the most selling category (60.3%) the profit is highest for the technology sector under which the profit has come more from the Consumers segment.
* Copiers is the least selling sub-category as we have already seen (only 17.88% of sales). However, copiers have given the most profit out of all the sub categories.

**Define:**

**Data description:**

Dataset contains information about **Date of order**, **Shipping** of products, **Customers** and their **Location**, the **Products** that they ordered, **Quantities**, **Total Amount** of each transaction, **Discounts** and the **Profit** that the business gained due to these orders. Products also belong to **Categories** and **Subcategories**. The initial dataset consists of 9,994 rows, each of which is a transaction, and twenty-one columns. The ETL (Extract, Transform, Load) process that have been made, reached to dataset of the same rows and sixteen columns. A small overview of our dataset before transformation and cleaning is shown below:



*Table 1. Overview of data*

**Collect:**

**Data Transformation:**

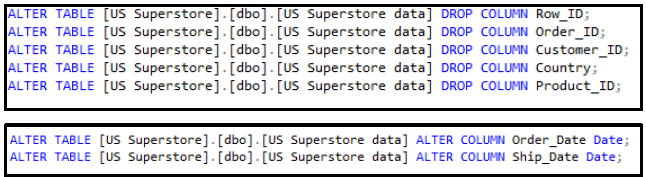
By taking the table 1 into consideration, identified that some changes must be done.

• Order\_Date and Ship\_Date variable should include only dates so time has to omitted.

• Row\_ID, Order\_ID, Customer\_ID and Product\_ID have to be removed as they do not give any additional information for our analysis (later created unique ID’s references for these variables).

• Column of Country takes only the value ‘United States’. We know that our superstore is located in United States, so this column is unnecessary.

created a query so as to delete and transform the data with the commands shown below:



*Figure 1: Data transformation*

**Organize:**

**Dimension and Fact Table:**

Created all the dimension tables, which refer to a collection of reference information about a measurable event. All dimension tables have been made in two stages. First, created a table for each of these only with the attributes that we wanted to include (e.g. Loc\_table has City, State, Region, Postal code). Next step was to fill these ‘first tables’ with values from ‘US Superstore’ table. In the example of loc\_table we have grouped by City, State, Region and Postal Code in order to have all combination of attributed listed inside table. Secondly, created the dimension tables by adding all the attributes again, plus with a unique ID (e.g. Location\_ID) that was added and includes numbers from 1 and so on. Finally, inserted values to dimension from the ‘first’ table’. ID’s are the primary keys of every dimension table.

Our final dimension table are as follows:

Dimension 1. dimLocation

Dimension 2. dimOrder

Dimension 3. dimShip

Dimension 4. dimShipMode

Dimension 5. dimCustomer

Dimension 6. dimCategory

Fact Table 7. FactStore

Dimension 1. dimLocation

|  |  |
| --- | --- |
| **dimLocation** | **Data type** |
| Location ID (PRIMARY KEY) | int |
| City | nvarchar |
| State | nvarchar |
| Region | nvarchar |
| Postal\_code | nvarchar |

Dimension 2. dimOrder

|  |  |
| --- | --- |
| **dimOrder** | **Data type** |
| Date\_id (PRIMARY KEY) | int |
| Order\_Date | date |
| Year\_Order | int |
| Month\_Order | int |
| Day\_Order | int |

Dimension 3. dimShip

|  |  |
| --- | --- |
| **dimShip** | **Data type** |
| DateShip\_id (PRIMARY KEY) | int |
| Ship\_Date | date |
| Year\_Ship | int |
| Month\_Ship | int |
| Day\_Ship | int |

Dimension 4. dimShipMode

|  |  |
| --- | --- |
| **dimShipMode** | **Data type** |
| Ship\_Mode\_ID (PRIMARY KEY) | int |
| Ship\_Mode | nvarchar |

Dimension 5. dimCustomer

|  |  |
| --- | --- |
| **dimCustomer** | **Data type** |
| Customer\_ID (PRIMARY KEY) | int |
| Customer\_name | nvarchar |
| Segment | nvarchar |

Dimension 6. dimCategory

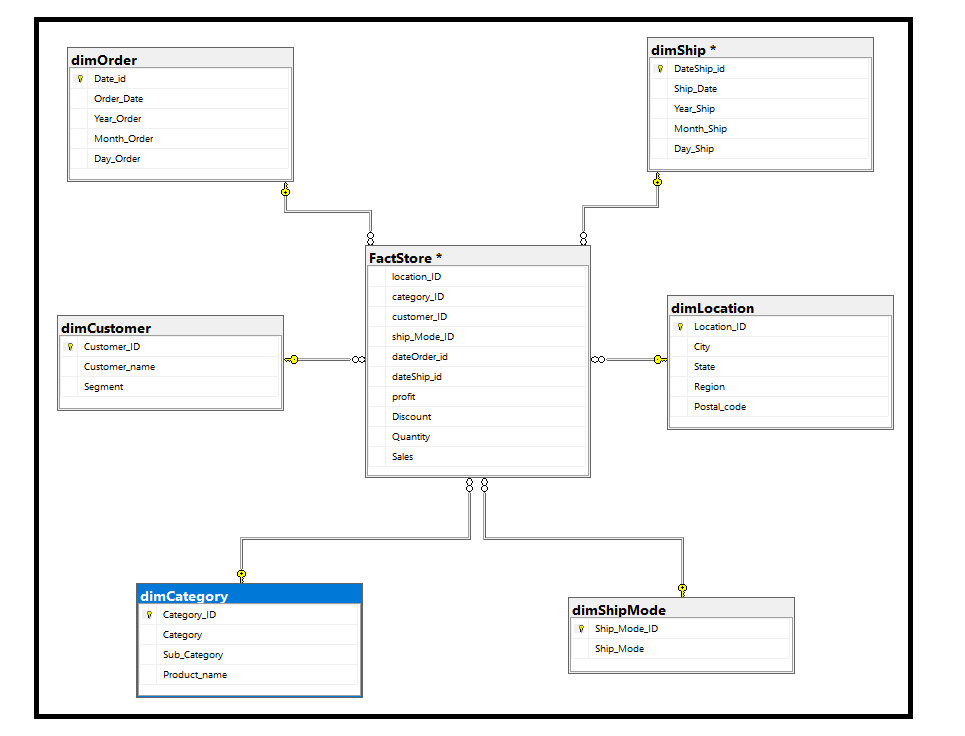
|  |  |
| --- | --- |
| **dimCustomer** | **Data type** |
| Category\_ID (PRIMARY KEY) | int |
| Category | nvarchar |
| Sub\_Category | nvarchar |
| Product\_name | nvarchar |

Fact Table 7. FactStore

|  |  |
| --- | --- |
| **FactStore** | **Data type** |
| location\_ID (FOREIGN KEY) | int |
| category\_ID(FOREIGN KEY) | int |
| customer\_ID(FOREIGN KEY) | int |
| ship\_Mode\_ID(FOREIGN KEY) | int |
| dateOrder\_id(FOREIGN KEY) | int |
| dateShip\_id | int |
| profit | float |
| Discount | float |
| Quantity | int |
| Sales | float |

Created the fact table which contains all numerical variables (profit, discount, quantity and sales), ID’s (location\_ID, category\_ID, customer\_ID, Ship\_Mode\_id, DateOrder\_id and dateShip\_id) and foreign keys that refer to every dimension table. Finally, inserted the values from the main table ‘US Superstore data’.

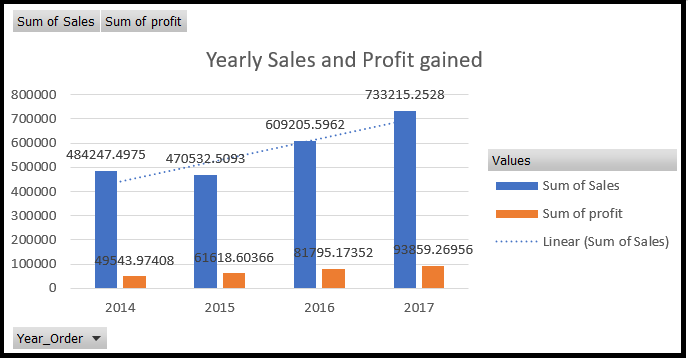
**Star Schema:**



*Figure 2: Star Schema*

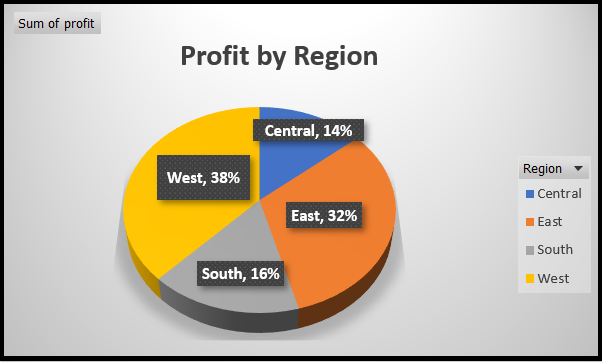
**Visualization:**

I used the MS Excel in order to make the visualization analysis of data. First of all, Let’s look at overview of the Superstore’s sales and profits. This could easily depict to a bar chart as shown below.



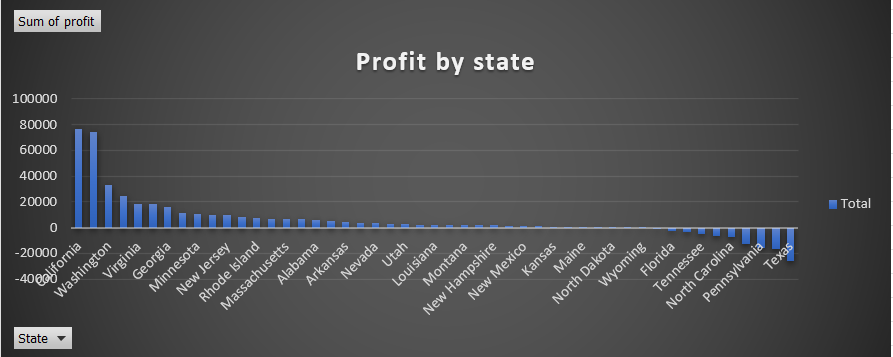
*Figure 3: Yearly Sales and profit gained*

* Sales of the store has increased every year resulting in high profit margin by the end of 2017.



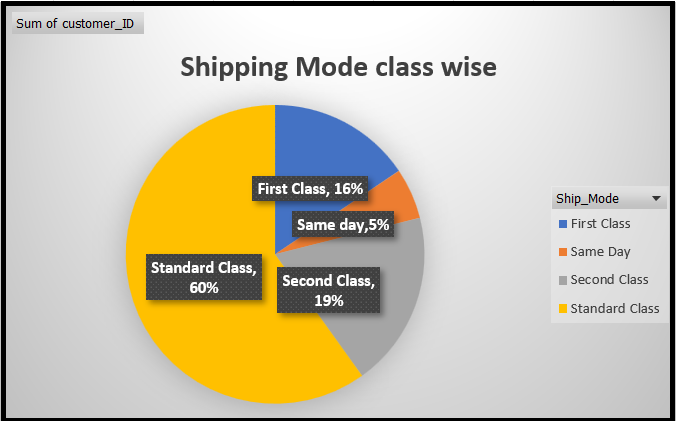
*Figure 4: Profit by Region*

* The most profitable region seems to be the Western with 38%.



*Figure 5: Profit by State*

* People residing in Western part of US tend to order more from the Superstore with California to be the state with the most orders during the years.
* In contrast with California, Texas is the State with the lowest profit (negative).

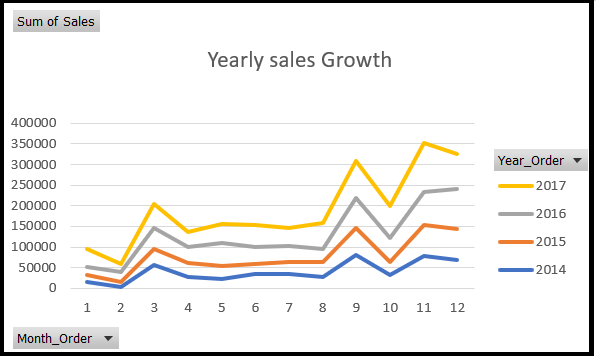


*Figure 6: Shipping Mode Class wise*

* As regards to the ship mode preferences, the majority of people prefer standard class shipment (60%) while only 5% choose the same day shipping.

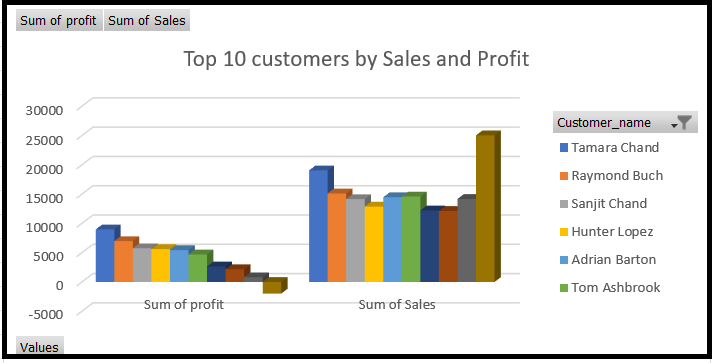
**Analyze:**

**Statistical Analysis:**



*Figure 7: Yearly Sales Growth*

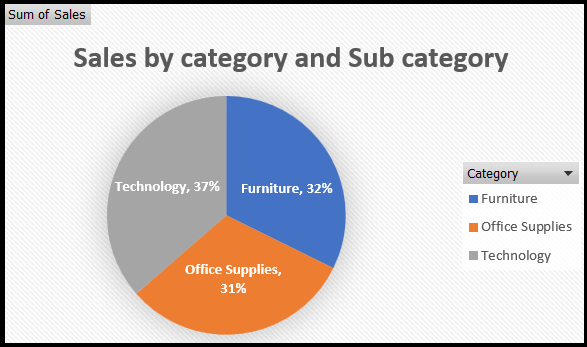
* There is an increasing tendency of sales by the years with peak season to be during autumn time.



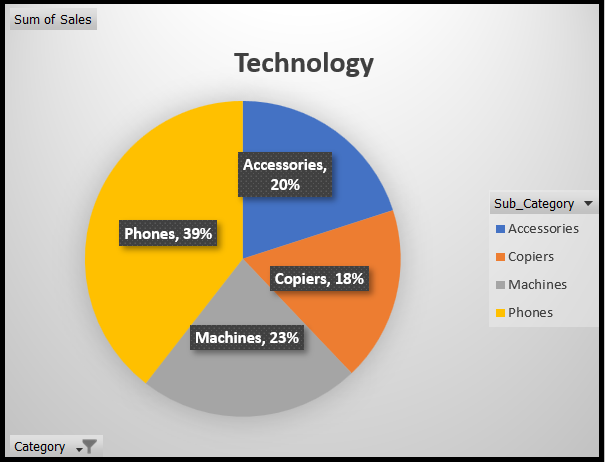
*Figure 8: Top 10 Customers by sales and profit*

* Tamara Chand seems to be the most loyal customer.

We would also like to analyze the products of the US Superstore. For this reason, drill down would be useful in order to come up with some interesting conclusions.

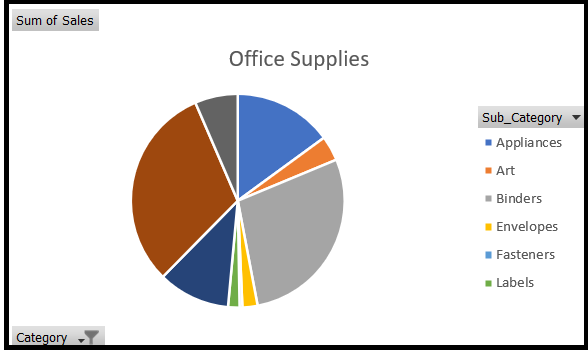
*Figure 9: Sales by category and sub category*

Technology category has the most sales (up to 36.40%) with phones to be the highest desired product. Chairs and storage from furniture and office supplies respectively are the products of the greatest sales.



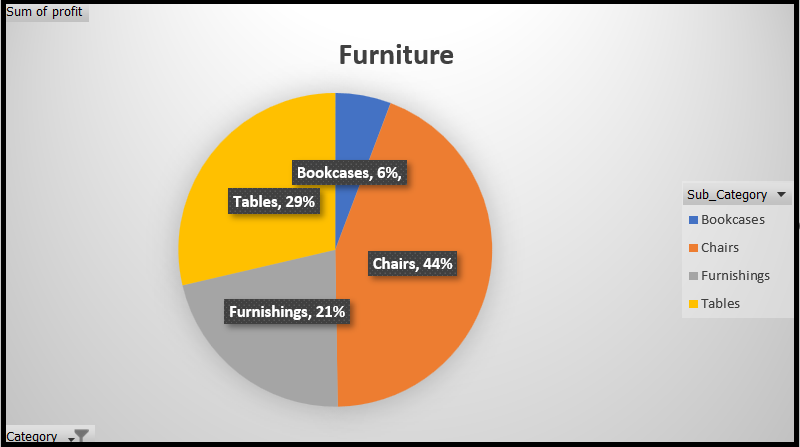
*Figure 10: Sales by sub category (Technology)*

* Mobile phones were the most sold in technology sub category



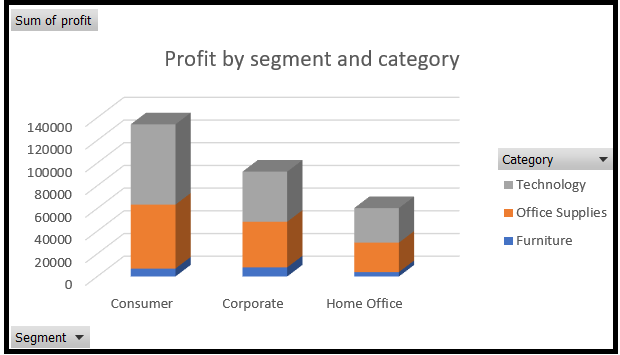
*Figure 11: Sales by sub category (Office Supplies)*

* Storage category got mostly sold in office supplies sub category



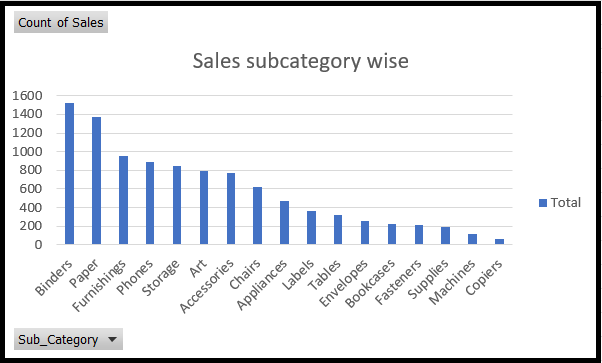
*Figure 12: Sales by sub category (Furniture)*

* Chairs were the most sold items in furniture sub category

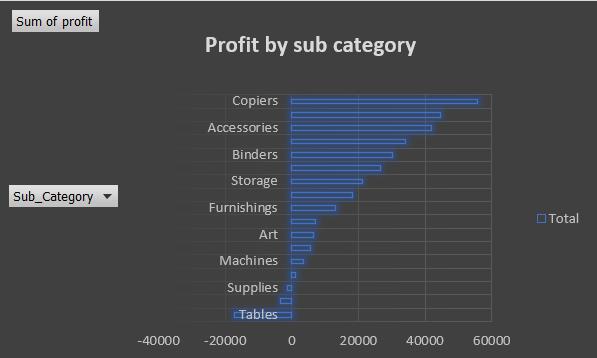


*Figure 13: Profit by segment and category*

* Although office supplies are the most selling category (60.3%) the profit is highest for the technology sector under which the profit has come more from the Consumers segment.



*Figure 14: Sales sub category wise*



*Figure 15: Profit by Sub category*

* Copiers is the least selling sub-category as we have already seen (only 17.88% of sales). However, taking into consideration this diagram above, copiers has given the most profit out of all the sub categories. Moreover, we could conclude that there is a huge loss from Tables.

**Conclusion:**

* If same day shipments could be given more discounts, it might encourage more sales/Profits.
* In order to avoid needless discounts for low sales, discounts should be dependent on sales
* and should not raise a specific range as needless discounts for low sales can witness huge
* losses.
* Binders and machines sector should receive greater attention. The furniture and Office supply sectors do not appear to be booming in the Central Region.
* Discount and Quantity doesn’t seem to have much correlation and same case with quantity and sales.
* Superstore should focus more on eastern and central region of the country as the sales numbers aren’t
* that good in these regions.
* Despite having fewer sales, Phone and Accessories generate higher earnings than Chairs, Appliances, and Machines.
* The highest sales and profit value is in copiers. This might be as a result of the low average discounts of 17.5% that it gives.

**References:**

Data Source:

https://www.kaggle.com/datasets/juhi1994/superstore