**ACKNOWLEDGEMENT**

A project work is a combination of views, ideas, suggestions and contribution of many people. Thus, one of the pleasant parts of writing the report is to thank those who have contributed towards its fulfilment.

I consider it as great privilege to have esteemed Lecturer Ms. Savleen Kaur as my project guide. I take this opportunity to express my sincere gratitude to her through constant advice and constructive criticism nourished my interest in the subject and provided a free and pleasant atmosphere to work against all odd situations. I avail this opportunity to extend my heart full thanks and deep respect to faculty member for their able guidance during this project.

My gratitude to all those, who **responded to my questionnaire** in a well-defined manner and helped me acquiring knowledge.

I would like to communicate a deep sense of gratitude to all these people without whom my project would not have been such a great learning experience.

# Table of Contents

|  |  |
| --- | --- |
| Sr.No. | Particulars |
| 1. | Introduction |
| 2. | Scope of The Analysis |
| 3. | Existing System |
| 4. | ETL Process |
| 5. | Analysis on Dataset |
| 6. | List of Analysis with Results |
| 7. | Future Scope |
| 8. | References |
| 9. | Bibliography |

**Introduction**

Online shopping has grown in popularity over the years, mainly because people find it convenient and easy to bargain shop from the comfort of their home or office.

One of the most enticing factor about online shopping, particularly during a holiday season, is it alleviates the need to wait in long lines or search from store to store for a particular item.

# Scope of The Analysis

Superstores industry comprises of companies that operate by having large size spaces which store and supply large amounts of goods. The superstore industry is comprised of extensive stores that sell a typical product line of grocery items and merchandise products, such as food, pharmaceuticals, apparel, games and toys, hobby items, furniture and appliances.

The analysis of such industry is of great importance as it gives insights for the sales and profits of various products. Our analysis is based on a superstore dataset for US country where the products are ordered between 2015 and 2018.

Since such vast field of data present in the US Superstore Dataset there is wide range of scope of the analysis. For example:

1. Quantity per Sub-Category.
2. Sales and Proft per year.
3. Orders in each shipping mode,etc.

**Existing System**

Before existence of Data Science, analyzing data used to be hectic task and existing system didn’t used to analyses the data with perfection.

Without existence of current cutting-edge technology of data science, we can get actionable insights in the dataset.

Following are the benefits which were present in the existing system of data analyzing:

* Making Better Decision with The Help of Data
* Directing actions based on trends- which later defines the goals required for profit.
* Doing challenging stuffs with the help of prediction which is done by data.
* Identifying various opportunities to increase the profit,
* Making decision with Quantifiable, data driven evidence so that loss doesn’t happens.
* Testing the decisions taken by the data and watching and analyzing the trend.
* Source of The Dataset
  + The dataset is taken from the Kaggle with the name ‘US Superstore Data’.
  + <https://www.kaggle.com/juhi1994/superstore/>
* Author of the Dataset

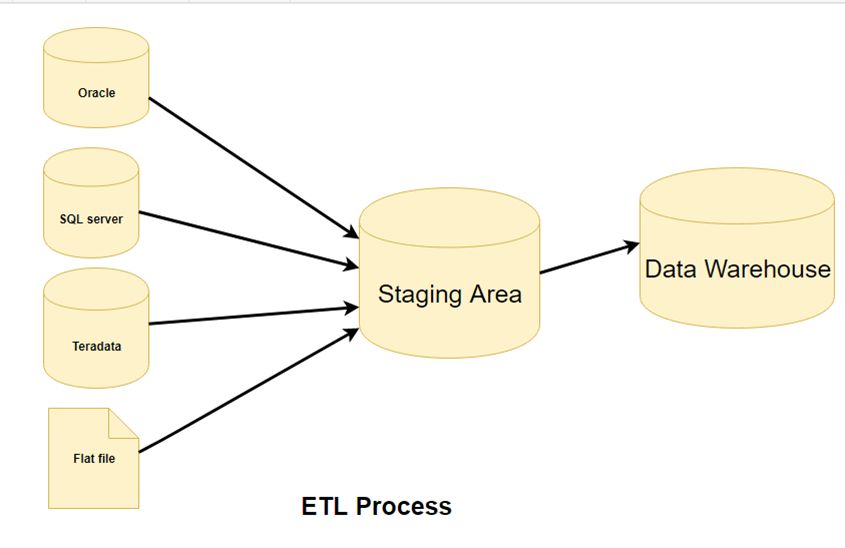
Juhi Badiyani

* Data last updated

December 2019

**ETL Process**

ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load.



Step 1) Extraction

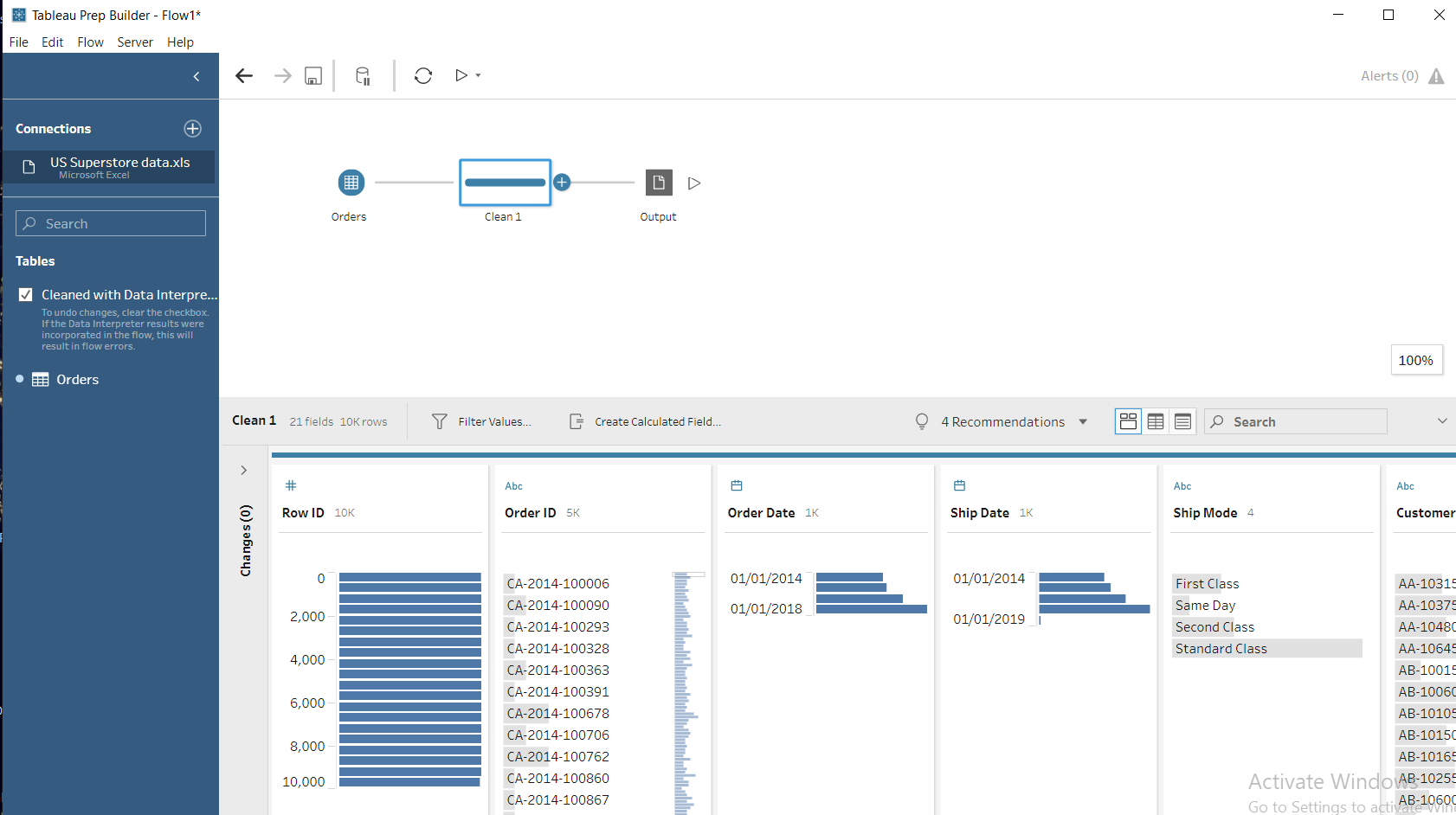
In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data warehouse database, rollback will be a challenge. Staging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

For this data, extraction is done directly from kaggle, since the data is already present in .csv format, we have converted the file to .xlsx format.

Irrespective of the method used, extraction should not affect performance and response time of the source systems. These source systems are live production databases. Any slow down or locking could effect company's bottom line.

Step 2) Transformation

Data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.



The data is cleaned using TableauPrep, the flow is shown above in the screenshot. Then it is taken out in csv format, and taken further for analysis.

Step 3) Loading

Loading data into the target datawarehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, load process should be optimized for performance.

Load verification

* Ensured that the key field data is neither missing nor null.
* Tested modeling views based on the target tables.
* Checked combined values and calculated measures.
* Data checks in dimension table as well as history table.

**Analysis of Dataset**

1. **Quantity per Sub-Category**

* Introduction: The analysis shows the amount of products available in each sub category.
* General Description:

From this graph, one can easily makeout which Category & Sub-Category to choose when they are looking to purchase a product.

* Specific requirements:

1. Power Pivot to manage dataset.
2. Pivot table for the dataset
3. Conditional Formatting in the Pivot table to make emphasis upon the data fluctuations.
4. Clustered pie chart

* Analysis Results:

The store has wide variety of Office Supplies especially in Binders and Paper department.

* Visualization:

1. **Sales and Profit per sub-category**

* Introduction: The analysis shows the sales and profit done in all sub category.
* General Description:

Through this analysis, the producers can compare the sales they have done in sub-categories and subsequently they can observe the profit gained, this can be used in strategic market analysis.

* Specific requirements:

1. Power Pivot to manage dataset.
2. Pivot table for the dataset
3. Conditional Formatting in the Pivot table to make emphasis upon the data fluctuations.
4. Clustered Bar chart.

* Analysis Results:

In tables sub-category, there is a huge sale unit to be seen but profit goes in negative scale, therefore the stores need to minimize the sales in this category.

Highest profit can be observed in Copiers sub category, whereas the sales amount in the same is more than twice the profit, which means the producer needs to verify with the quality of their product, so that subsequent profit can be observed.

Talking about chairs, sales tend to increase at rapid rate but the profit graph does not parallelly go well, they need to see the difference between the cost price and selling price.

* Visualization:

1. **Count of orders in each shipping mode**

* Introduction: The analysis shows the maximum and minimum orders placed in each shipping mode category.
* General Description:

Orders are placed via three shipping modes –

1. First Class
2. Second Class
3. Standard Class

Also, ‘Same Day’ shipping mode is also available to customers located in same locality.

* Specific requirements:

1. Power Pivot to manage dataset.
2. Pivot table for the dataset
3. Conditional Formatting in the Pivot table to make emphasis upon the data fluctuations.
4. Clustered Bar chart.

* Analysis Results:

Standard class has maximum sales in all the subcategories, followed by second class and then first class. Shipping on the same day has very less count as observed in the dataset.

This might be the reason that maximum customers can afford standard class easily, whereas first class and second class are less chosen by them as they exceed the shipping price threshold.

Also, to deliver on same day, not everyone prefers to use the option, might be they prefer waiting for the delivery and not investing in same day mode. The store manager should look into the shipping price for same day, as more customers in the locality might be attracted to their product, and it may result in more advertising of their store.

* Visualization:

1. **Category-wise Profit per Year**

* Introduction: The analysis shows the year wise profit in each category and sub category.
* General Description:

The products are classified into three categories-

1. Furniture- includes table, chair, bookcases, etc.
2. Office Supplies- includes binders, art supplies, envelopes, etc.
3. Technology- includes phones, machines, accessories, etc.

The sales in these categories are observed with respect to their profit and discount each year.

* Specific requirements:

1. Power Pivot to manage dataset.
2. Pivot table for the dataset
3. Conditional Formatting in the Pivot table to make emphasis upon the data fluctuations.
4. Clustered Bar chart.

* Analysis Results:

From the given analysis, it can be observed that maximum profit is achieved by Technology with maximum sale in all the years.

Talking about the sales in furniture, the sales is quite high as compared to the profit and quantity, producer need to see through the sales strategy.

In terms of office supplies, though profit is not observed much, but sales are also significantly low in comparison with Technology products.

Throughout the time period of all years, maximum sales and profit can be seen in the year 2017 in all three categories.

* Visualization:

1. **Profit in each segment with Quantity and Discount**

* Introduction: The analysis shows the profit in each segment with respect to discount provided to customers.
* General Description:

The segments described in dataset are –

1. Consumers
2. Corporate
3. Home Office

The products are sold on a particular discount in each category and the profit obtained is observed so that more profit can be made by visualizing maximum sales in each segment.

* Specific requirements:

1. Power Pivot to manage dataset.
2. Pivot table for the dataset
3. Conditional Formatting in the Pivot table to make emphasis upon the data fluctuations.
4. Clustered Bar chart.

* Analysis Results:

Maximum profit can be observed in Home office segment followed by Corporate and Consumer.

Sum of quantity is almost negligible as compared to their profit, this means they are in the right path

In furniture category, tables and book cases have their growth in negative scale which implies they are having loss in terms of profit as well in sales.

In technology category, maximum sales and profit is done by accessories and machines whereas supplies subcategory has pulled down the graph in office supplies category.

* Visualization:

**List of Analysis with Results**

From this data processing, we can get this following conclusion:

# Top products available in abundance are paper, binder, phone and art.

# Highest sales is see in chair subcategory but highest profit is seen in copier subcategory.

# Most customers prefer delivery through standard class because of low shipping rate, and least customers are interested in same day delivery mode.

# Maximum sales is done in year 2017 in technology category with maximum profit.

# Overall, home office has done maximum sales and profit followed by corporate and consumer.

# Maps –

# 

# Maximum sales is done in California.

# References

1. <https://www.youtube.com/watch?v=20zDV9MNE0s>
2. https://www.youtube.com/watch?v=kIZDb\_pHvX0
3. <https://www.excel-easy.com/examples/slicers.html>
4. https://www.kaggle.com/juhi1994/superstore

**Bibliography**

1. Microsoft Excel 2016 Bible: The Comprehensive Tutorial Resource by John Walkenbach, Wiley
2. Fundamentals of Business Analytics by R.N. Prasad, Seema Acharya, Wiley