

Implementing several techniques for Data Analytics and Visualisation

Student's ID:

Student's Name:

Date:

Table of Contents

Introduction.....	3
Business Requirement.....	3
Vision and Goals	4
Star Schema	5
Draw.IO.....	5
Implementation of Data Warehouse in SQL	6
ETL for Populating Data Warehouse	7
Tools used for Data Warehouse.....	8
SQL Server Reporting Services	9
Visualization	11
Conclusion	11
References.....	Error! Bookmark not defined.
Appendix.....	12
Dataset Link	12
Code	12

Introduction

The chosen topic is related to comprehensive dataset of an e-commerce sales, checking whether the product meeting customer demand, product category, customer reviews, pricing. The dataset will be capable of answering queries like whether the product is on par with current market trend, what the customer rating was, whether the customer queries were answered, and the importance of the product. Furthermore, it will be able to explore into the data to uncover insights that can optimize product listings, pricing strategies, and marketing campaigns. The dataset used for model building contains 1000 observations of 18 variables. The study will provide a star schema and visualization reports of the dataset through the implementation of Tableau and SQL Server (Kaggle, 2024).

Business Requirement

The chosen dataset will include details on the product's rating, the number of customer service calls related to it, the product quality, and the other sales report regarding the product. These elements will present chances to recognize patterns and trends in the current market. The company will benefit greatly from this model developed upon sales report in terms of business intelligence development, which will enable it to obtain insights from its client database and make more effective decisions (Kaggle, 2024). One of the most important things is to make sure that product IDs, dates, times, and category IDs are recorded for both customers and products. Comprehending the choices and preferences of customers is an additional crucial requirement. Their order pattern, IDs should all be taken into consideration while classifying them. Improving sales related to the product, is an important consideration. Analysis of product ratings, warehouse details, and product importance is another requirement.

Key Stakeholders

The key stakeholders of this supermarket sales data are shown below:

- **Management:** The management team is responsible for making strategic decisions through business analytics tools.
- **Business Owners/Entrepreneurs:** Founders or owners who establish and operate the e-commerce platform are one of the major stakeholders of an e-commerce market.
- **Customer service Department:** This department can get insight information regarding the customer behavior and also solves queries regarding the products and they also can evaluate the current market trends for better results.
- **Sales team:** The sales can determine the most valuable product of the supermarket for the e-commerce company.
- **Finance Department:** This department is responsible for tracking the transactional data such as total sales, tax, gross margin, etc. They are also responsible for handling online transactions and process payments of the market.

Vision and Goals

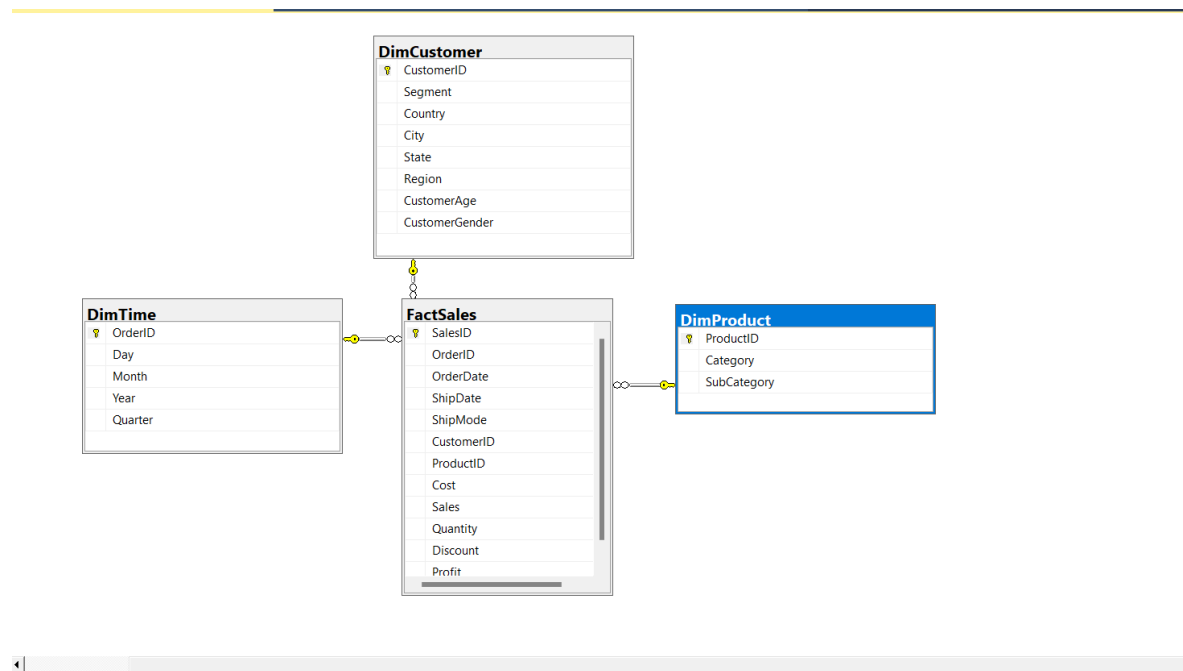
The primary vision of this work is to create an efficient warehouse that will help the e-commerce company analyze sales data of the e-commerce market. These analytic factors can provide actionable insights regarding the market trend analysis and decision-making process for an optimized and structured business model. The primary goal of this data analytics is to integrate the sales data of various products of the e-commerce market into the developed model, which helps to analyze business operations.

Star Schema

The star schema for e-commerce serves as the central point of analysis which is used for organizing of the data which are sales dataset of the market. The star schema modelling can be done using various types of data storage methods such as data warehousing.

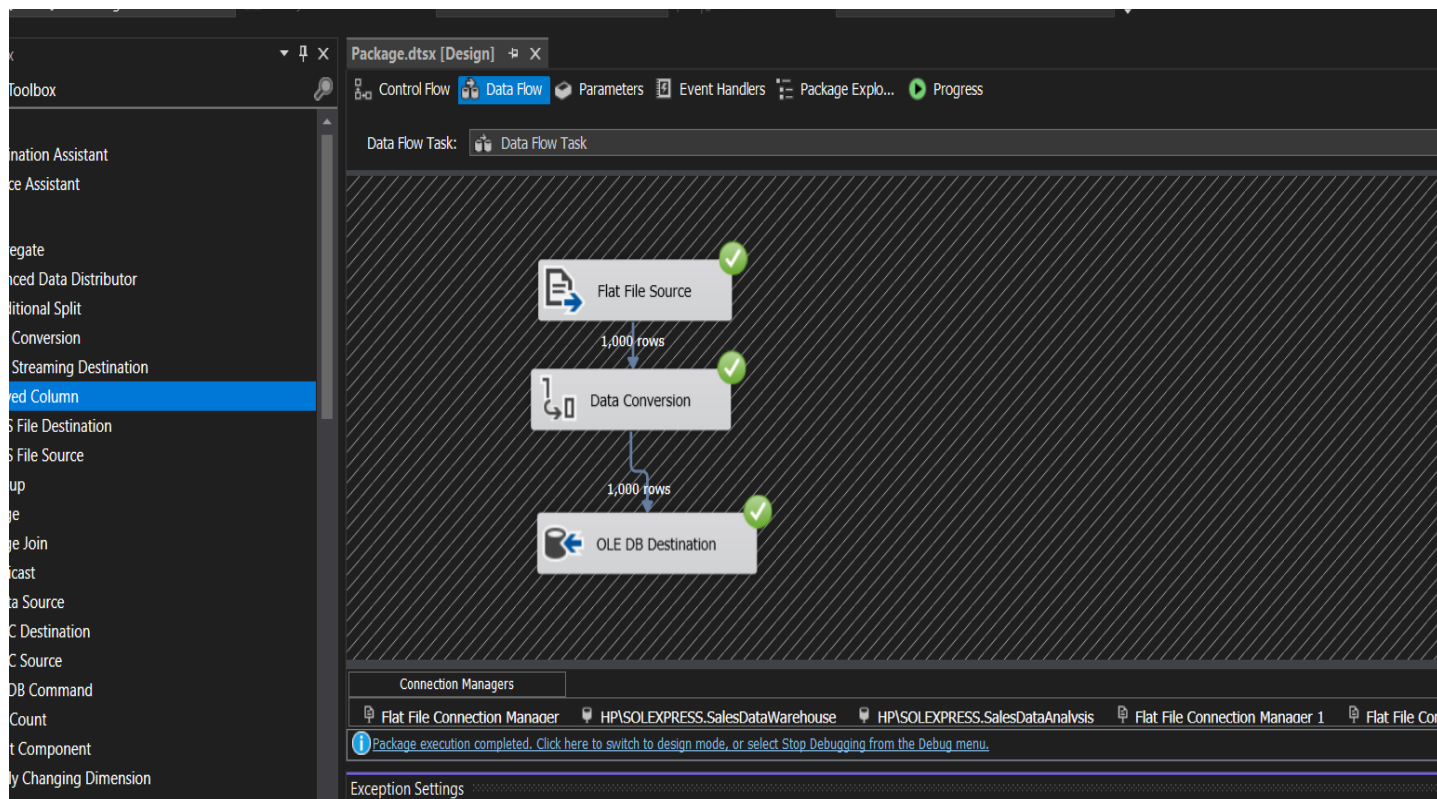
Draw.IO

In the schema there are total three dimensions and one fact table which consists of all the measures and primary keys of all the dimension that are tracking the details of all the dimensions. These dimensions track details like product details, product category, customer details like name and address details, and lastly time details of the product delivery. The fact sales table contains primary key to all the dimensions. Below a detailed explanation of all the tables will be provided.



Implementation of Data Warehouse in SQL

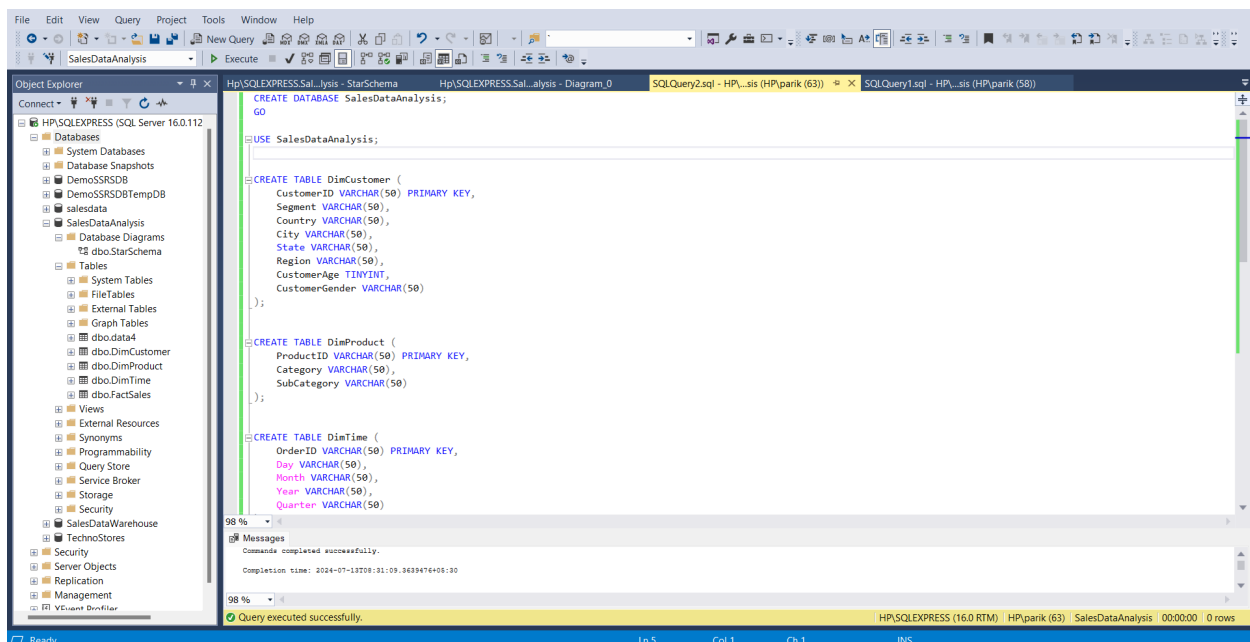
The first step includes steps like analysis of the source system from which the data will be extracted from the data set and a conceptual model, representing the relationship between each entity will be designed, and tables have been designed. The source data plays a pivotal role into identifying data elements and determining their mining. Dimension models are designed based on the source data and implemented based on the business requirements of the e-commerce company. Dimensional modeling techniques involve structuring data into fact tables and dimension tables using techniques such as star schema. These models are built in order to optimize query performance and offer efficient analysis of data for the understanding of the market trends for the e-commerce market. After that the dimension models are deployed into a data warehouse environment for storage and easier accessibility. After the deployment phase is completed, data is extracted from source systems and loaded into the data warehouse. Data integration and ETL (Extract, Transform, Load) processes are implemented to cleanse the source data before loading it into the warehouse. Furthermore, data quality and validation are also checked. Finally, automation through various processes has been done for better efficiency, reliability, and scalability of the model.



ETL for Populating Data Warehouse

I have created my dataset of 1000 rows by using Python and after creating the dataset I had created tables in new database so that I can insert data in those tables afterwards by using SSIS. I have used SSIS for ETL purpose to populate data in SQL Server's database but there are different ways by which we can insert data in our datasets Likee this, the creation of the creation of the data model is the first step of the implementation process. Secondly, the dimension models have been deployed. The SQL model consists of tables like Dimension product, which has product name as its primary key, and product name, category id, price, review score, review count as its other entities. The dimension table is connected to Fact sales, which has sales ID as its primary key, and it consist of the primary keys of the other dimensions which are dimension customer and dimension

time. The dimension customer has customer id as its primary key, other entities are customer name, segment, state and country. The dimension time has time id as its primary key, other entities are year, month, day and quarter. Lastly, dimension category is another dimension connected to dimension product which has category id as its primary key, and category name as it's another entity.

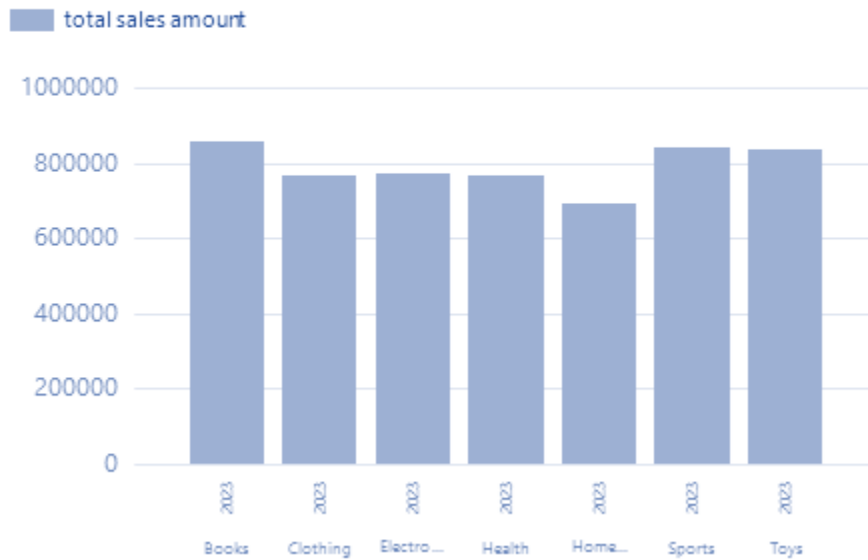


Tools used for Data Warehouse

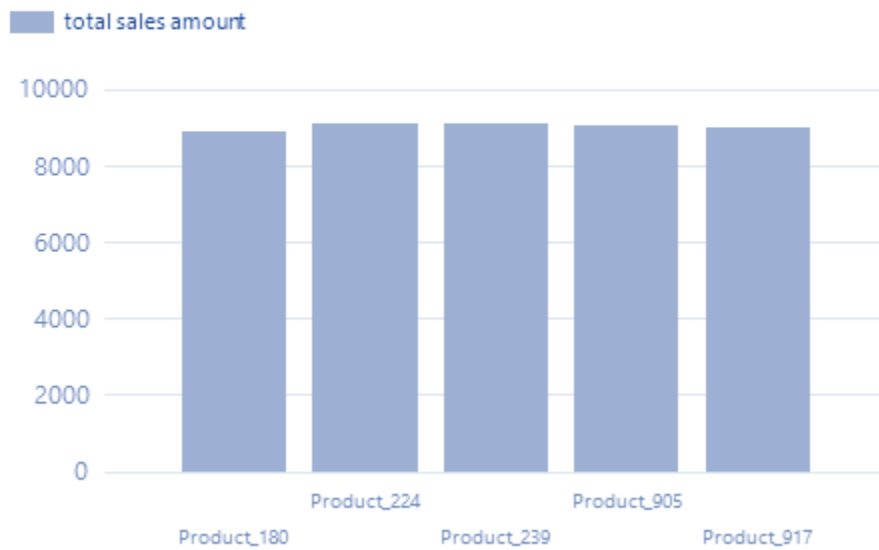
A wide range of tools have been used for warehousing of the data. Some of the tools which have been implemented for the purpose of this project are SQL Server, Tableau, and Visual Studio. For star schema, Draw.IO is used.

SQL Server Reporting Services

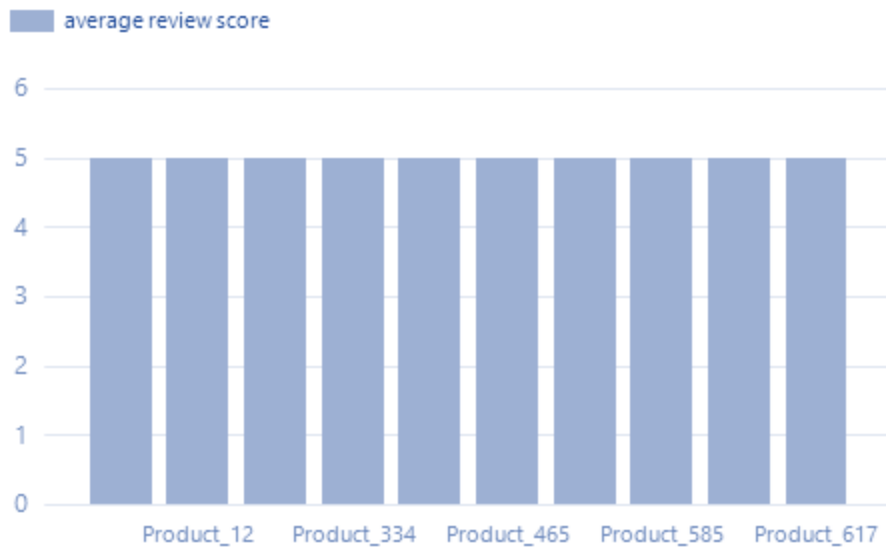
Total Sales Amount by Product Category for a Specific Year



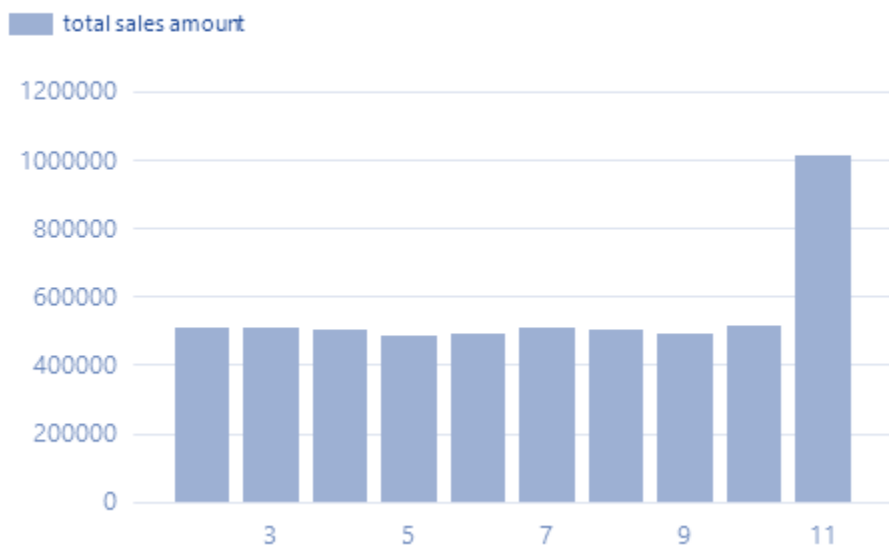
Top 5 Best-Selling Products by Sales Amount



Top 10 Products by Average Review Score



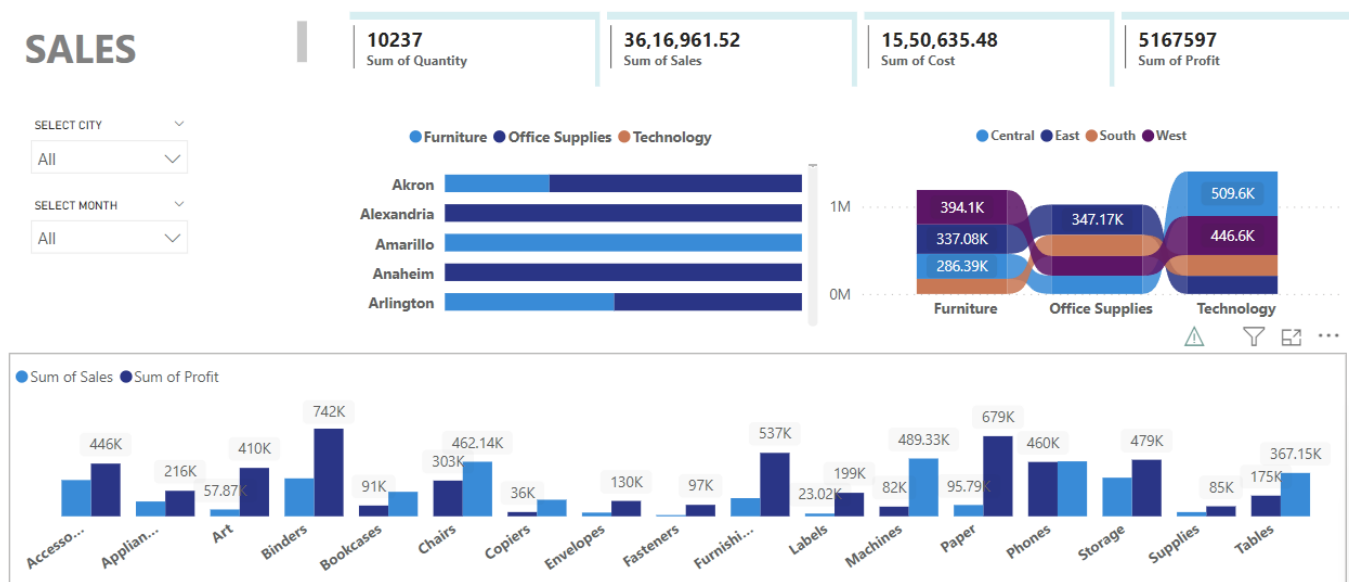
Total Sales Amount by Month



Visualization

Various types of data visualization techniques have been used for the purpose of this project.

Below a visual representation of all the techniques will be provided.



Conclusion

In this study, the sales data of and e-commerce market has been analyzed. From the analytics, several insights, such as product quality, customer preference, current market trends and product importance, have been gained. Furthermore, a decision-making strategy has also been determined through this analysis, which results in more sales, more effective analysis of market trends and better decision-making.

Appendix

Dataset Link

https://github.com/parikshitcarpenter/SalesDataset_Random_1000_Rows/blob/master/SalesDataMain.csv

Code

```
# Generate 1000 unique Sales IDs
sales_ids = set()
while len(sales_ids) < 1000:
    sales_ids.add(generate_unique_ids(sales_ids, 8))

order_ids = set()
while len(order_ids) < 1000:
    order_ids.add(generate_unique_ids(order_ids, 10))

start_date = datetime(2015, 1, 1)
end_date = datetime(2023, 12, 31)
date_ids = [(start_date + timedelta(days=random.randint(0, (end_date - start_date).days))).strftime('%Y%m%d') for _ in range(1000)]

customer_ids = set()
while len(customer_ids) < 1000:
    customer_ids.add(generate_unique_ids(customer_ids, 6))

df = pd.DataFrame({
    'SalesID': list(sales_ids),
    'OrderID': list(order_ids),
    'DateID': date_ids,
    'CustomerID': list(customer_ids)
})

df.sort_values(by=['SalesID', 'OrderID', 'DateID', 'CustomerID'], inplace=True)

print(df.head())

df.to_csv('random_data.csv', index=False)
```

