E-Commerce Sales and Customer Behavior Analysis in USA

Sireesha Pothumudi
College of Science and Engineering
University of Minnesota
Student ID: 5880639
email: pothu012@umn.edu

Sri Krishna Vamsi Koneru
College of Science and Engineering
University of Minnesota
Student ID: 5881358
email: koner033@umn.edu

Suchithra Moolinti
College of Science and Engineering
University of Minnesota
Student ID: 5909334
email: m0006012@umn.edu

Nithya Murikinati
College of Science and Engineering
University of Minnesota
Student ID: 5903224
email: murik002@umn.edu

***Abstract*—This report constitutes a detailed examination of E-commerce sales and customer behavior within the United States, framed within the context of database systems and principles. The study employs robust database concepts and methodologies to analyze key trends, patterns, and factors influencing the E-commerce landscape, focusing on the role of databases in shaping and optimizing online retail operations. The analysis encompasses a comprehensive evaluation of sales performance, tracking growth trajectories and market dynamics by leveraging data within a database framework. The report delves into the intricacies of database design and management to understand how effective data storage, retrieval, and indexing contribute to enhancing overall E-commerce efficiency.**

Customer behavior, a crucial aspect of the study, is scrutinized through the lens of database principles, emphasizing data normalization, data pre-processing and exploratory data analysis. The report investigates factors such as consumer preferences and purchasing patterns, utilizing relational databases to analyze key performance indicators impacting sales. Furthermore, the role of databases in supporting emerging technologies within the E-commerce sector is explored, focusing on how these technologies rely on sound database architecture for effective implementation. Visual representation of the results produced by the analysis has been reported for better understanding.

In conclusion, by integrating database concepts into the analysis of E-commerce sales and customer behavior, this project contributes to a deeper understanding of how effective database management is pivotal in optimizing online retail operations and ensuring a seamless customer experience.

Keywords— E-Commerce, Sales Analysis, customer behavior, key performance indicators

I. Introduction

The landscape of commerce has undergone a transformative shift with the rise of E-commerce, a dynamic arena where products are exchanged through online channels rather than traditional Mom & Pop or Brick & Mortar stores. In this rapidly evolving domain, E-commerce giants such as Amazon, Walmart, and Target vie for market dominance, amplifying the competition and necessitating a nuanced understanding of consumer behavior.

The motivation behind our project stems from the unique challenges and opportunities intrinsic to E-commerce. Unlike traditional commerce, where trust is often built over time, success in E-commerce hinges on providing customers with unparalleled value and convenience. This, in turn, necessitates a comprehensive understanding of customer behavioral patterns. Leveraging Online Analytical Processing (OLAP), our project delves into the intricacies of customer purchase patterns, with a specific focus on the pivotal role that discounts play in shaping market share and expanding customer bases.

Goals:

The overarching objectives of our analysis are to decipher the intricate patterns within the realm of E-commerce sales, with a keen eye on key performance indicators (KPIs) that significantly influence market dynamics. Our goals include:

1. Analyze 2022 Sales Patterns: Uncover insights into sales peaks, slumps, and monthly trends throughout the entire year, providing a comprehensive understanding of the market's ebb and flow.

2. Examine Discount Impact on Sales: Investigate how discounts impact the overall sales value, identifying the correlation between promotional strategies and consumer purchasing behavior.

3. Customer Segmentation Analysis: Explore KPIs related to customer categories, delineating distinct segments based on purchasing behavior and demographics.

4. Location-Based Analysis: Investigate the impact of location on sales, discerning patterns and variations in order behavior across different geographical areas.

5. Category-Specific Analysis: Examine KPIs associated with product categories, unveiling insights into the top-selling categories and products that drive market success.

Scope:

Within the purview of our analysis, we will delve into historical sales data, identifying popular product categories and top-selling products. Customer segmentation, considering both behavior and demographics, will be a focal point, alongside the evaluation of the impact of discounts on overall sales. The findings will be presented through a variety of visualization tools, including charts, graphs, and dashboards, providing a visually intuitive representation of the intricate dynamics within the E-commerce landscape.

In navigating the complexities of E-commerce, this report aims to unravel the patterns and factors that contribute to success, offering valuable insights for stakeholders seeking a competitive edge in this rapidly expanding digital marketplace.

II. Data Description

For the comprehensive analysis of E-commerce sales and customer behavior in the United States, we leveraged an E-Commerce sales dataset sourced from Kaggle, specifically curated for the year 2022. This dataset served as the foundational bedrock for our project, providing a wealth of information on online retail transactions.

The dataset contained 19 attributes and they are as follows:

1. Order Date
2. Row ID
3. Order ID
4. Ship Mode
5. Customer ID
6. Segment
7. Country
8. City
9. State
10. Postal Code
11. Region
12. Product ID
13. Category
14. Sub-category
15. Product Name
16. Sales
17. Quantity
18. Discount
19. Profit

This initial dataset, while rich in content, presented challenges that necessitated meticulous data preprocessing to render it analytically useful. Several issues were addressed during this phase:

Missing Order IDs:

The dataset exhibited instances of missing Order IDs, a critical component for tracking and identifying individual transactions. Rigorous efforts were undertaken to rectify and fill these gaps, ensuring the completeness and accuracy of the dataset.

Special Characters:

Special characters within the dataset posed potential impediments to effective data analysis. Through data preprocessing techniques, these characters were identified and either standardized or removed, mitigating potential anomalies in the dataset.

Inconsistent Date Formatting:

Date-related data is fundamental to understanding sales trends over time. However, the dataset contained inconsistencies in date formatting. By applying data preprocessing methodologies, we harmonized the date formats, enabling a seamless temporal analysis of sales patterns.

Data Standardization:

To maintain consistency and coherence in the dataset, various elements such as product names, categories, and customer information underwent standardization. This step was essential for accurate comparisons and classifications during subsequent analyses.

The meticulous data preprocessing efforts undertaken were facilitated primarily through SQL queries, ensuring that the dataset was cleansed, standardized, and prepared for in-depth exploration. The resulting refined dataset forms the basis of our analysis, providing a reliable foundation upon which we conduct our investigation into E-commerce sales and customer behavior in the United States during the year 2022.

III. Methodology

The methodology employed for this report involved a systematic approach to E-commerce sales and customer behavior analysis, integrating fundamental database concepts throughout the entire process. The key steps undertaken are outlined below:

1. Data Collection:

Datasets were sourced from Kaggle websites, specifically targeting United States E-commerce sales records. This initial step laid the foundation for subsequent analyses by providing raw data representative of the online retail landscape.

2. Architecture

The Extract, Load, Transform (ELT) process is a data integration method that involves three key stages in managing and preparing data for analysis. In the first step, data is extracted from various sources, such as databases, applications, or external systems. This raw data is then loaded into a centralized data storage system, such as a data warehouse, without undergoing significant transformation. Once the data is securely stored, the transformation phase takes place, where it is cleaned, enriched, and structured to meet the specific requirements of the analytical tasks at hand. ELT has gained popularity in the era of big data and cloud computing, as it takes advantage of the processing power and scalability of modern data platforms, enabling organizations to efficiently handle and analyze large volumes of data for business intelligence and decision-making purposes.

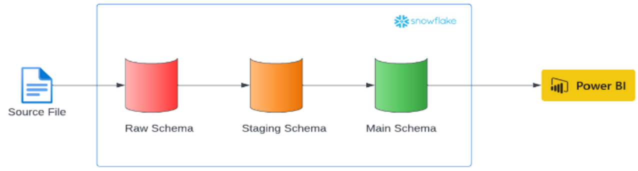


Fig1. Architecture

This fig1 depicts the high-level architecture and in this project, we utilized the ELT process to retrieve the dataset from Kaggle and initially loaded it to the Raw Schema in the snowflake data warehouse. In this raw data each data record isn't cleansed so there are chances that data record will have null values, there will be different formats for similar attributes, the data isn't normalized, and each attribute of the data record is of type "VARCHAR". So, this data is further scanned through, and each data record is attached with an error_flag as "Y" in case there is error data like data format mismatch etc. Data type for each attribute is updated accordingly. From this the data is stored in a staging schema. At this stage all the data is cleaned for null values and duplicates. Records with error flag enabled are updated and the data is then stored in the Main Schema. This Main Schema is used for exploratory data analysis and the data related to the defined KPI's is retrieved. Further after the analysis the reported data is imported into PowerBI, and the dashboard is created to visualize the data.

3. Data Warehousing:

A cloud data warehouse, Snowflake, was chosen as the platform for housing and managing the datasets. Raw data files were loaded into the Snowflake data warehouse, ensuring a centralized and scalable infrastructure for subsequent analysis.

4. Data Normalization and Schema Design

In this phase, the data underwent normalization processes to eliminate redundancy and improve efficiency. A well-defined schema was designed to organize and structure the data logically, facilitating seamless retrieval and analysis.

5. Data Preprocessing:

The raw data underwent thorough preprocessing to enhance its quality and usability. SQL queries were employed to address missing values, handle outliers, and standardize formats for consistency. This step was crucial in ensuring the integrity of the data for subsequent analyses.

6. Exploratory Data Analysis (EDA):

In-depth exploratory data analysis was conducted to uncover underlying sales trends, correlations, and patterns. Visualizations and statistical techniques were applied to gain insights into the dynamics of E-commerce sales within the United States. The emphasis on database principles facilitated efficient querying and analysis of large datasets.

7. Identify Key Factors:

The analysis focused on identifying key factors influencing E-commerce sales, with a particular emphasis on discounts, product categories, and seasonality impacts. SQL queries were instrumental in dissecting the dataset to reveal the relationships and dependencies among these key factors.

8. Visualization with Power BI

The analyzed and processed data from Snowflake was exported to Power BI for the creation of interactive and informative dashboards. Power BI was utilized to visually represent sales trends over time and to highlight the influence of discounts, product categories, and seasonality on sales. This step ensured that the findings were presented in a clear and accessible manner for stakeholders.

By seamlessly integrating database concepts into each stage of the methodology, this approach aimed to provide a robust foundation for analyzing E-commerce sales and customer behavior in the USA, facilitating meaningful insights for decision-makers in the online retail sector.

IV. Data Warehousing

Our project revolves around the transformative Snowflake data warehousing platform, a pivotal choice for decoding e-commerce intricacies. Snowflake's unique architecture, embracing a Software as a Service (SaaS) model, strategically separates storage, compute, and services, providing adaptability and flexibility crucial for our analytical pursuits.

Snowflake's versatility shines as it handles diverse data types, enriching our understanding of customer demographics and product intricacies. The platform's on-demand scalability meets the dynamic flow of e-commerce data, ensuring efficiency and cost-effectiveness by allowing independent scaling of resources. The storage-compute separation simplifies management, creating synergy for seamless data analysis. Efficient query performance, a hallmark of Snowflake, becomes pivotal in unraveling sales trends and customer behavior in the intricate web of e-commerce data. The Extract, Load, Transform (ELT) process within Snowflake initiates a dance of data migration. Diverse data sources, including historical sales, customer info, and products, converge in Snowflake's tables like orders and customers.

The transformative phase ensures data homogeneity by addressing special characters. Snowflake's support for various data formats turns special characters from roadblocks into steppingstones in our analytical journey. Its innate capabilities preserve data integrity, paving the way for comprehensive insights.

As we usher data into Snowflake's embrace, anticipation rises for unparalleled insights. Snowflake's prowess, combined with meticulous ELT processes, promises a deep exploration of e-commerce trends. This journey propels us towards refined strategies and heightened customer satisfaction, where technology and commerce converge seamlessly in the dance of data warehousing with Snowflake.

V. Data Normalization

In the realm of e-commerce, where vast amounts of data flow through various channels, the importance of data normalization cannot be overstated. This report delves into the meticulous process of normalizing data during the loading phase into the landing/raw schema of our database for the E-commerce Sales and Customer Behavior Analysis project.

Normalization of Orders Fact Table:

The core of our dataset resides in the "orders" table, housing crucial transactional data. To optimize this table, we employed normalization techniques to minimize redundancy and enhance data integrity. In this normalization process, we transformed the original orders table into a more structured form, introducing foreign keys that serve as references to the dimension tables. Instead of replicating detailed information in every order entry, we substituted this redundancy with foreign keys, such as customer_id and product_id, which point to the respective dimension tables—customers and products. This normalization not only conserves storage space but also establishes a logical and interconnected database model, paving the way for efficient querying and analysis.

The below part of the code performs a bulk data load into the TB_LND_ORDERS table in the landing schema.

```
COPY INTO
DB_US_ECOM.LANDING.TB_LND_ORDERs(
ORDER_DATE,ROW_ID,ORDER_ID,SHIP_MODE,CUSTO
MER_ID,PRODUCT_ID,SALES,QUANTITY,DISCOUNT,
PROFIT,LOAD_TIMESTAMP,FILENAME,ROW_NUM)
FROM ( SELECT
$1,$2,$3,$4,$5,$12,$16,$17,$18,$19,
CURRENT_TIMESTAMP::TIMESTAMP_NTZ,METADATA$
FILENAME, METADATA$FILE_ROW_NUMBER FROM
@DB_US_ECOM.LANDING.ecom_data_stage/US_E_c
ommerce_records_2022.csv)
   FILE_FORMAT =
(TYPE=CSV,SKIP_HEADER=1,FIELD_DELIMITER=',
',TRIM_SPACE=FALSE,FIELD_OPTIONALLY_ENCLOS
ED_BY='"',REPLACE_INVALID_CHARACTERS=TRUE,
DATE_FORMAT=AUTO,TIME_FORMAT=AUTO,TIMESTAM
P_FORMAT=AUTO)ON_ERROR=ABORT_STATEMENT;
```

Dimension Tables and Detailed Information:

The dimension tables, namely "customers" and "products," are pivotal in preserving detailed information about customers and products, respectively. In these tables, we retained a normalized form, attributing each table with specific attributes related to the entity it represents. For instance, the "customers" table encapsulates details such as segment, city, country, postal code, and region, while the "products" table stores information like product name, category, and sub-category.

The below part of the code performs a bulk data load into the TB_LND_CUSTOMERS table in the landing schema.

```
COPY INTO
DB_US_ECOM.LANDING.TB_LND_PRODUCTS(
PRODUCT_ID,CATEGORY,SUB_CATEGORY,PRODUCT_N
AME,LOAD_TIMESTAMP,FILENAME,ROW_NUM) FROM
(SELECT $12,$13,$14,$15,
CURRENT_TIMESTAMP::TIMESTAMP_NTZ,METADATA$
FILENAME, METADATA$FILE_ROW_NUMBER FROM
@DB_US_ECOM.LANDING.ecom_data_stage/US_E_c
ommerce_records_2022.csv)FILE_FORMAT =
(TYPE=CSV,SKIP_HEADER=1,FIELD_DELIMITER=',
',TRIM_SPACE=FALSE,FIELD_OPTIONALLY_ENCLOS
ED_BY='"',REPLACE_INVALID_CHARACTERS=TRUE,
DATE_FORMAT=AUTO,TIME_FORMAT=AUTO,TIMESTAM
P_FORMAT=AUTO)ON_ERROR=ABORT_STATEMENT;
```

Like the previous section, this code segment loads data into the TB_LND_CUSTOMERS table, capturing customer-related information.

```
COPY INTO
DB_US_ECOM.LANDING.TB_LND_CUSTOMERS(
CUSTOMER_ID,SEGMENT,COUNTRY,CITY,STATE,POS
TAL_CODE,REGION,LOAD_TIMESTAMP,FILENAME,RO
W_NUM) FROM (SELECT $5,$6,$7,$8,$9,$10,
$11,CURRENT_TIMESTAMP::TIMESTAMP_NTZ,
METADATA$FILENAME,METADATA$FILE_ROW_NUMBER
FROM
@DB_US_ECOM.LANDING.ecom_data_stage/US_E_c
ommerce_records_2022.csv)FILE_FORMAT =
(TYPE=CSV,SKIP_HEADER=1,FIELD_DELIMITER=',
',TRIM_SPACE=FALSE,FIELD_OPTIONALLY_ENCLOS
ED_BY='"',REPLACE_INVALID_CHARACTERS=TRUE,
DATE_FORMAT=AUTO,TIME_FORMAT=AUTO,TIMESTAM
P_FORMAT=AUTO)ON_ERROR=ABORT_STATEMENT;
```

The normalization of these dimension tables contributes significantly to the overall integrity and coherence of the database. By structuring the data in this manner, we ensure that each entity is described comprehensively and accurately, avoiding data redundancy, and maintaining a streamlined database model.

The normalization process brings forth several advantages for our E-commerce Sales and Customer Behavior Analysis project. Firstly, it facilitates consistency by ensuring that each piece of information is stored in one place and updated uniformly. Secondly, it reduces redundancy, as detailed information about customers and products is stored only in the respective dimension tables, preventing unnecessary duplication. Furthermore, data normalization enhances data integrity by minimizing the risk of inconsistencies and errors. The structured and interconnected model allows for efficient querying and analysis across multiple dimensions, providing a solid foundation for the exploration of e-commerce sales trends and customer behavior.

Entity-Relationship (ER) Diagram:

To provide a comprehensive view of the database structure for the E-commerce Sales and Customer Behavior Analysis project, we have crafted an Entity-Relationship (ER) diagram. This visual representation delineates the entities, their attributes, and the relationships between them. Below is a brief description of the main components depicted in the ER diagram:

Orders Entity:

Attributes: ORDER_DATE, ROW_ID, ORDER_ID, SHIP_MODE, SALES, QUANTITY, DISCOUNT, PROFIT.

Relationships: Connected to the CUSTOMERS and PRODUCTS entities through foreign keys (CUSTOMER_ID and PRODUCT_ID).

Customers Entity:

Attributes: CUSTOMER_ID, SEGMENT, COUNTRY, CITY, STATE, POSTAL_CODE, REGION, and others.

Relationships: Connected to the ORDERS entity through the foreign key CUSTOMER_ID.

Products Entity:

Attributes: PRODUCT_ID, CATEGORY, SUB_CATEGORY, PRODUCT_NAME, and others.

Relationships: Connected to the ORDERS entity through the foreign key PRODUCT_ID.

This ER diagram visually represents the logical connections between different entities, facilitating a clear understanding of the database structure. It serves as a valuable reference for developers, analysts, and stakeholders involved in the E-commerce Sales and Customer Behavior Analysis project.

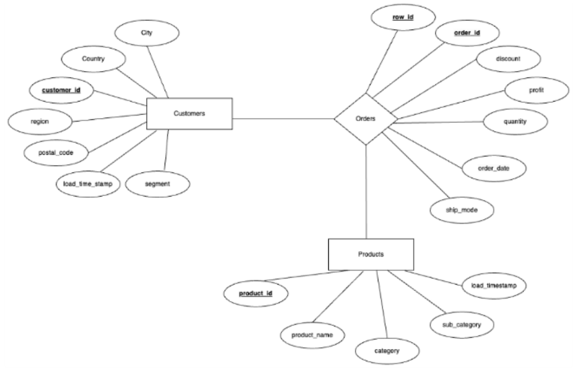


Fig2. ER Diagram

VI. Data Preprocessing

The raw data underwent thorough preprocessing to enhance its quality and usability. SQL queries were employed to address missing values, handle outliers, and standardize formats for consistency. This step was crucial in ensuring the integrity of the data for subsequent analyses.

Removing the Null Values:

In this operation updates the `error_flag` and `error_message` columns in the orders, customers, and products tables within the `STAGING` schema of the `DB_US_ECOM` database. These queries mark records where the `order_id`,‘ Row_id’ , ‘Customer_id’ and ‘Product_id’ are null by setting `error_flag` to 'Y' and assigning the error message 'OrderID is Null.' in respective tables.

```
update DB_US_ECOM.STAGING.TB_STG_ORDERS
set error_flag = 'Y' , error_message =
'OrderID is Null' where order_id is null;


update DB_US_ECOM.STAGING.TB_STG_ORDERS
set error_flag = 'Y' , error_message =
'RowID is Null' where row_id is null;


update DB_US_ECOM.STAGING.TB_STG_CUSTOMERS
set error_flag = 'Y' , error_message =
'CUSTOMER_ID is Null' where CUSTOMER_id is
null;


update DB_US_ECOM.STAGING.TB_STG_PRODUCTS
set error_flag = 'Y' , error_message =
'PRODUCT_id is Null' where PRODUCT_id is
null;
```

Standardizing the data by identifying the duplicates:

The query updates the `error_flag` and `error_message` columns in the `TB_STG_ORDERS` table within the `STAGING` schema of the `DB_US_ECOM` database. It marks records as duplicates based on a combination of the `row_id` and `order_id` columns by setting `error_flag` to 'Y' and assigning the error message 'Duplicate record based on PK Combination RowID, OrderID.' The duplication is identified by comparing the count of occurrences of unique combinations of `row_id` and `order_id` in the `TB_STG_ORDERS` table, flagging records where the count is greater than one. This query is instrumental in maintaining data integrity by highlighting instances of duplicate records in the specified primary key combination.

```
update DB_US_ECOM.STAGING.TB_STG_ORDERS
set error_flag = 'Y' , error_message =
'Duplicate record based on PK Combination
RowID, OrderID' where (row_id, order_id)
in (select row_id, order_id from (select
row_id, order_id , count(*) cnt from
DB_US_ECOM.STAGING.TB_STG_ORDERS group by
1, 2 having cnt>1));
```

Once the data is identified the duplicates are removed using the query:

```
delete from DB_US_ECOM.MAIN.ORDERS where
(order_date,row_id,order_id) in (select
distinct order_date,row_id,order_id from
DB_US_ECOM.STAGING.TB_STG_ORDERS);
```

Utilize Error Flagging and load data into main table:

The provided SQL query inserts data into the `ORDERS` table within the `MAIN` schema of the `DB_US_ECOM` database. The data is sourced from the `TB_STG_ORDERS` table in the `STAGING` schema, with the condition `error_flag = 'N'` ensuring that only records without flagged errors are selected. The selected columns (`ORDER_DATE`, `ROW_ID`, `ORDER_ID`, etc.) from the staging table are inserted into corresponding columns in the main `ORDERS` table. This operation effectively transfers clean and error-free data from the staging area to the main database, ensuring the integrity and reliability of the inserted records.

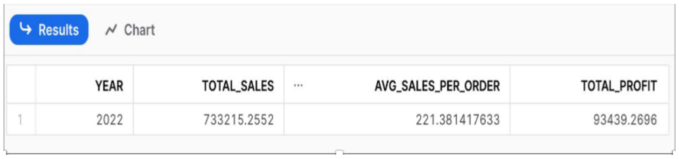
```
insert into
DB_US_ECOM.MAIN.ORDERS(ORDER_DATE, ROW_ID,
ORDER_ID, SHIP_MODE, CUSTOMER_ID,
PRODUCT_ID, SALES, QUANTITY, DISCOUNT,
PROFIT, UPDATED_TIMESTAMP, LOAD_TIMESTAMP,
FILENAME)

select ORDER_DATE, ROW_ID, ORDER_ID,
SHIP_MODE, CUSTOMER_ID, PRODUCT_ID, SALES,
QUANTITY, DISCOUNT, PROFIT,
UPDATED_TIMESTAMP, LOAD_TIMESTAMP,
FILENAME from
DB_US_ECOM.STAGING.TB_STG_ORDERS where
error_flag = 'N';


delete from DB_US_ECOM.MAIN.PRODUCTS where
PRODUCT_ID in (select distinct PRODUCT_ID
from DB_US_ECOM.STAGING.TB_STG_PRODUCTS );

insert into
DB_US_ECOM.MAIN.PRODUCTS(PRODUCT_ID,
CATEGORY, SUB_CATEGORY, PRODUCT_NAME,
UPDATED_TIMESTAMP, LOAD_TIMESTAMP,
FILENAME)

select PRODUCT_ID, CATEGORY, SUB_CATEGORY,
PRODUCT_NAME,  UPDATED_TIMESTAMP,
LOAD_TIMESTAMP, FILENAME from
DB_US_ECOM.STAGING.TB_STG_PRODUCTS where
error_flag = 'N';


delete from DB_US_ECOM.MAIN.CUSTOMERS
where CUSTOMER_ID in (select distinct
CUSTOMER_ID from
DB_US_ECOM.STAGING.TB_STG_CUSTOMERS );

insert into
DB_US_ECOM.MAIN.CUSTOMERS(CUSTOMER_ID,
SEGMENT, COUNTRY, CITY, STATE,
POSTAL_CODE, REGION, UPDATED_TIMESTAMP,
LOAD_TIMESTAMP, FILENAME)

select CUSTOMER_ID, SEGMENT, COUNTRY,
CITY, STATE, POSTAL_CODE, REGION,
UPDATED_TIMESTAMP, LOAD_TIMESTAMP,
FILENAME from
DB_US_ECOM.STAGING.TB_STG_CUSTOMERS where
error_flag = 'N';
```

VII. Exploratory Data Analysis

On top of the preprocessed data, exploratory data analysis has been performed on certain crucial KPIs. Exploratory Data Analysis can be defined as an analytical approach in which initial investigation is conducted on the data to discover general trends and uncover patterns. This serves as a foundation upon which further in-depth analysis can be conducted to obtain specific information about the data and identify additional patterns. It also enables the execution of further analytics, such as Predictive and Prescriptive Analytics.

KPI 1: Total sales, Average sales, and Total profit per Year

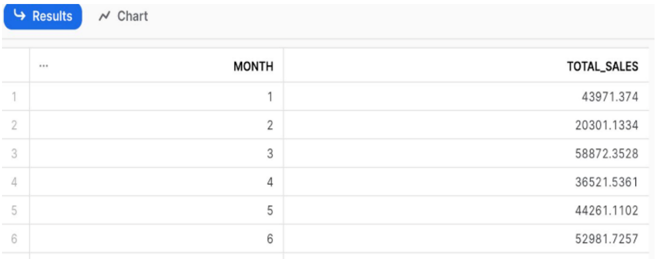
In our E-commerce Sales and Customer Behavior Analysis project in the USA, we used SQL to extract key performance indicators (KPIs) for Total Sales, Average Sales per Order, and Total Profit per Year from our database. The code selects the year, calculates the total sales, average sales per order, and total profit by grouping the data based on the order date. This information helps us understand the overall performance and profitability of our e-commerce business on an annual basis. Additionally, we explored product popularity by analyzing sales in different categories. Through a SQL query involving a left join between orders and products tables, grouping by category, and sorting by total sales, we determined that the Technology category had the highest total sales, making it the most popular product category in our dataset. This analysis provides valuable insights for strategic decision-making in our e-commerce operations.



	YEAR	TOTAL_SALES	···	AVG_SALES_PER_ORDER	TOTAL_PROFIT
1	2022	733215.2552		221.381417633	93439.2696

KPI 2: Understand sales peak, slump, monthly trends for the complete year.

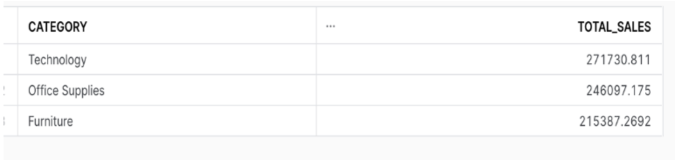
Our further analysis centers around another crucial Key Performance Indicator (KPI) that offers valuable insights into the dynamics of sales trends throughout the entire year. This KPI helps us pinpoint when sales reach their highest points, identify periods of lower sales (slumps), and understand how they fluctuate on a monthly basis. To accomplish this, we employ a specific piece of code written in SQL, which efficiently extracts relevant information from the ORDERS table in our database. This code intelligently groups the total sales data by month, providing us with a detailed overview of sales performance over time. This systematic approach enables us to create visualizations and charts that reveal patterns, highlight peak periods, and shed light on any downturns in sales throughout the annual cycle. The strategic insights gained from this analysis contribute significantly to our understanding of the nuanced landscape of E-commerce sales in the USA.



	MONTH	TOTAL_SALES
1	1	43971.374
2	2	20301.1334
3	3	58872.3528
4	4	36521.5361
5	5	44261.1102
6	6	52981.7257

KP1 3: Identifying popular product categories

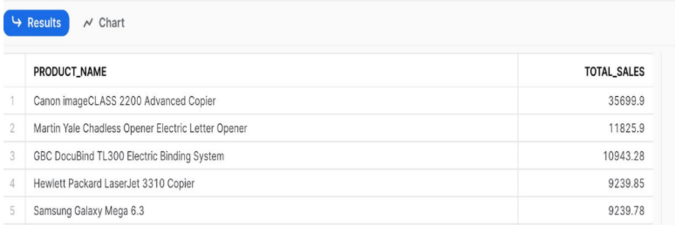
We identified which product category was the most popular in the dataset by identifying the number of products sold from various categories and then categorizing them into categories and comparing the total sales of each category thus deciding which category came out on top. We did this through SQL where we performed a left join to the orders and products table over product_ID. Then grouped them by category to sort them into categories and finally ordered them by total sales to get the category wise sales in an ascending order. From our analysis, we were able to conclude that the Technology category had the highest total sales and thus was the most popular product category.



CATEGORY	TOTAL_SALES
Technology	271730.811
Office Supplies	246097.175
Furniture	215387.2692

KPI 4: Identifying TOP 5 products based on sales

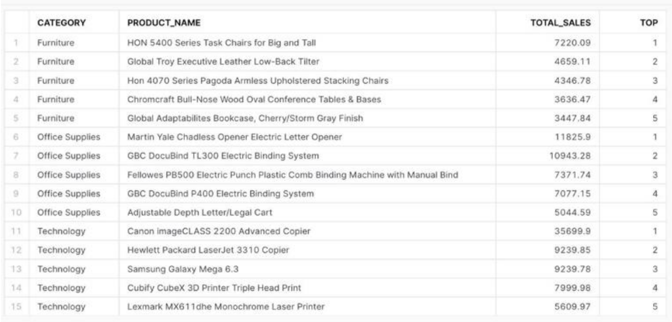
We also identified which specific products were the most popular in the dataset, we identified the total sales of each product and then based on our observations we were able to conclude the popularity of the products in the dataset. We achieved this through SQL by again performing a left join of the tables orders and products on product ID but here we grouped them by product names as opposed to earlier and then ordered by total sales to get the sorted order of the various products.



	PRODUCT_NAME	TOTAL_SALES
1	Canon imageCLASS 2200 Advanced Copier	35699.9
2	Martin Yale Chadless Opener Electric Letter Opener	11825.9
3	GBC DocuBind TL300 Electric Binding System	10943.28
4	Hewlett Packard LaserJet 3310 Copier	9239.85
5	Samsung Galaxy Mega 6.3	9239.78

KPI 4.5: Identifying TOP 5 category wise products based on sales

We then found out which five products were the most popular in their respective categories. To do this, in SQL we first performed a left join like before on the orders and products tables over product ID and then we grouped them by category and product name but we used qualify command to ensure only top five products from each category was in the output. We then ordered the results by category and in ascending order in each category by sales of the products.

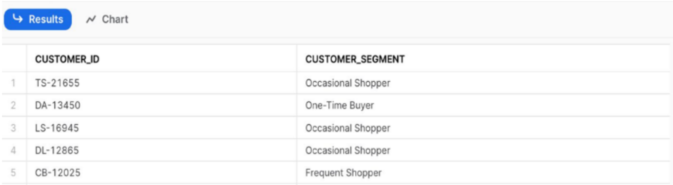


	CATEGORY	PRODUCT_NAME	TOTAL_SALES	TOP
1	Furniture	HON 5400 Series Task Chairs for Big and Tall	7220.09	1
2	Furniture	Global Troy Executive Leather Low-Back Tilter	4659.11	2
3	Furniture	Hon 4070 Series Pagoda Armless Upholstered Stacking Chairs	4346.78	3
4	Furniture	Chromcraft Bull-Nose Wood Oval Conference Tables & Bases	3636.47	4
5	Furniture	Global Adaptabilites Bookcase, Cherry/Storm Gray Finish	3447.84	5
6	Office Supplies	Martin Yale Chadless Opener Electric Letter Opener	11825.9	1
7	Office Supplies	GBC DocuBind TL300 Electric Binding System	10943.28	2
8	Office Supplies	Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind	7371.74	3
9	Office Supplies	GBC DocuBind P400 Electric Binding System	7077.15	4
10	Office Supplies	Adjustable Depth Letter/Legal Cart	5044.59	5
11	Technology	Canon imageCLASS 2200 Advanced Copier	35699.9	1
12	Technology	Hewlett Packard LaserJet 3310 Copier	9239.85	2
13	Technology	Samsung Galaxy Mega 6.3	9239.78	3
14	Technology	Cubify CubeX 3D Printer Triple Head Print	7999.98	4
15	Technology	Lexmark MX611dhe Monochrome Laser Printer	5609.97	5

We can observe from the results we obtained the specific products that were the most popular in the various categories as in the image below.

KPI 5: Identifying the customer categories based on number of orders

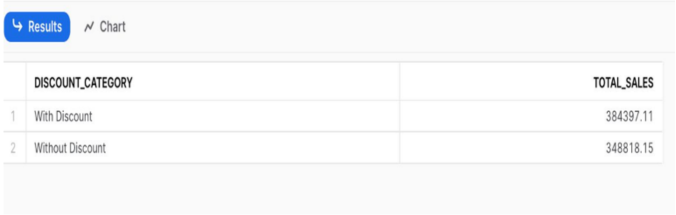
We then proceeded to find trends in the customers. We classified the customers into different categories like frequent, occasional buyers based on the number of orders they had in the dataset over the course of the year using a support threshold that we have set . To achieve this in SQL, we made use of the DISTINCT qualifier to get this. First, we performed a left join of orders and customers tables over customer id and then based on the count of distinct order ids we segmented the customer into various categories.



	CUSTOMER_ID	CUSTOMER_SEGMENT
1	TS-21655	Occasional Shopper
2	DA-13450	One-Time Buyer
3	LS-16945	Occasional Shopper
4	DL-12865	Occasional Shopper
5	CB-12025	Frequent Shopper

KPI 6: Analyze Impact of Discounts based on Total Sales

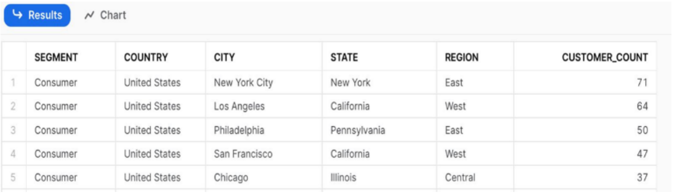
We then observed the impact of discounts on sales. In the dataset in the records where the discounts column was not zero and had a value we interpreted that as with discount and those with zero values as without discount and then aggregated the sales of such records to identify the trend in sales with respect to discounts. In SQL, from the orders table we aggregated the records where discount was greater than zero into With Discount category and other records into Without Discount category after aggregating them into the two categories by grouping the records based on this discount category.



	DISCOUNT_CATEGORY	TOTAL_SALES
1	With Discount	384397.11
2	Without Discount	348818.15

KPI 7: The distribution of customers across different demographics

We then shifted our attention towards the regional distribution of the customers in the dataset and on a higher level to which states the the customers belonged to. We focused on the individual customers rather than the number of orders thus considering only unique customer ids in the analysis and ignore the duplicates. In SQL, we again used distinct to ensure no duplicate records of customer IDs were considered in the analysis. We performed a left join of the orders and customers tables over customer ID and then did a count of the distinct customer ids. We selected the regional information like state,city along with the customer count and then grouped them by each specific category in top down approach like first by region then state and then city.

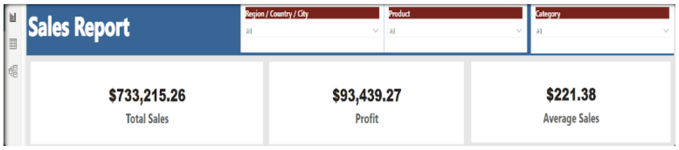


	SEGMENT	COUNTRY	CITY	STATE	REGION	CUSTOMER_COUNT
1	Consumer	United States	New York City	New York	East	71
2	Consumer	United States	Los Angeles	California	West	64
3	Consumer	United States	Philadelphia	Pennsylvania	East	50
4	Consumer	United States	San Francisco	California	West	47
5	Consumer	United States	Chicago	Illinois	Central	37

VIII. Results

KPI 1: Total sales, Average sales, and Total profit per Year

In our USA E-commerce project, our analytical focus for the year 2022 was directed through a dedicated Power BI dashboard. This tool provided a comprehensive overview of critical financial metrics, specifically Total Sales, Average Sales, and Total Profit. The dashboard offered a visual representation of the aggregated revenue, typical transaction values, and net financial gains post-cost considerations. This data-driven approach using Power BI for the year 2022 equipped our team with valuable insights, facilitating strategic decision-making and a nuanced understanding of the financial landscape within our E-commerce operations.

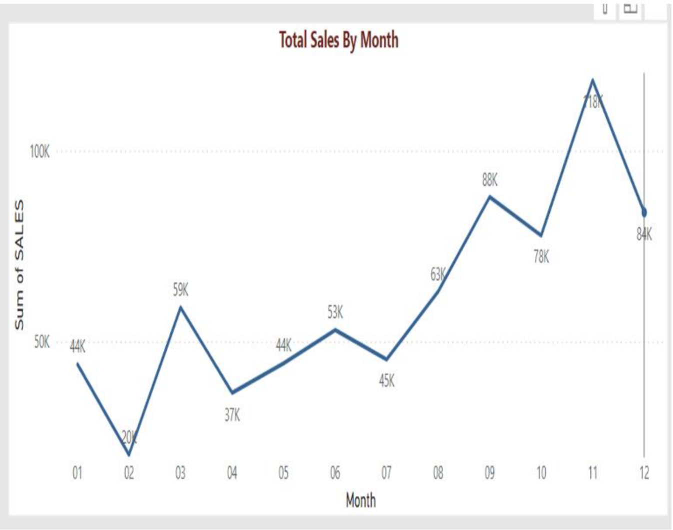


Sales Report

$733,215.26	$93,439.27	$221.38
Total Sales	Profit	Average Sales

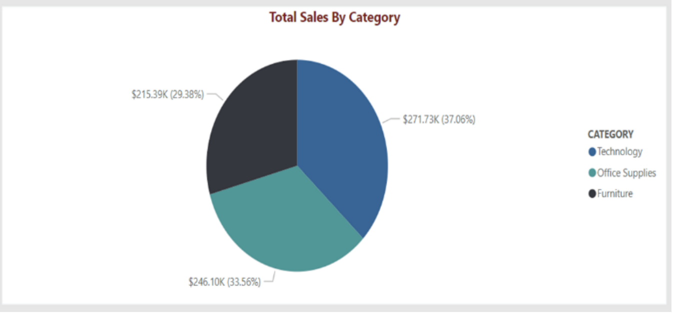
KPI 2: Understand sales peak, slump, monthly trends for the complete year

Commencing our analytical journey, we utilized a line graph for visualization, meticulously tracking sales peaks, slumps, and monthly trends throughout the entire year. This visual representation allowed us to easily identify when sales were at their highest and lowest, along with any recurring patterns each month. Notably, we observed a significant peak in e-commerce sales during the Thanksgiving month of November in the USA. This insight adds a valuable layer to our understanding, emphasizing the influence of holiday seasons on sales trends. Analyzing these patterns provides us with actionable insights for strategic decision-making and optimizing our approach to meet customer demands effectively.



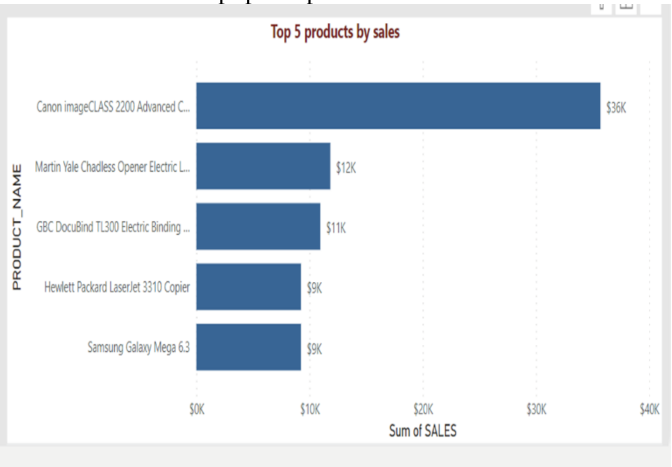
KP1 3: Identifying popular product categories

We visualized the categories in Power BI in the form of a pie chart where each sector was representing the different categories and from the size of each sector we can conclude which categories had higher sales among the three. A closer examination of the visualization would give us a clearer overview of the percentages accounted for by each category and the sales amounts.



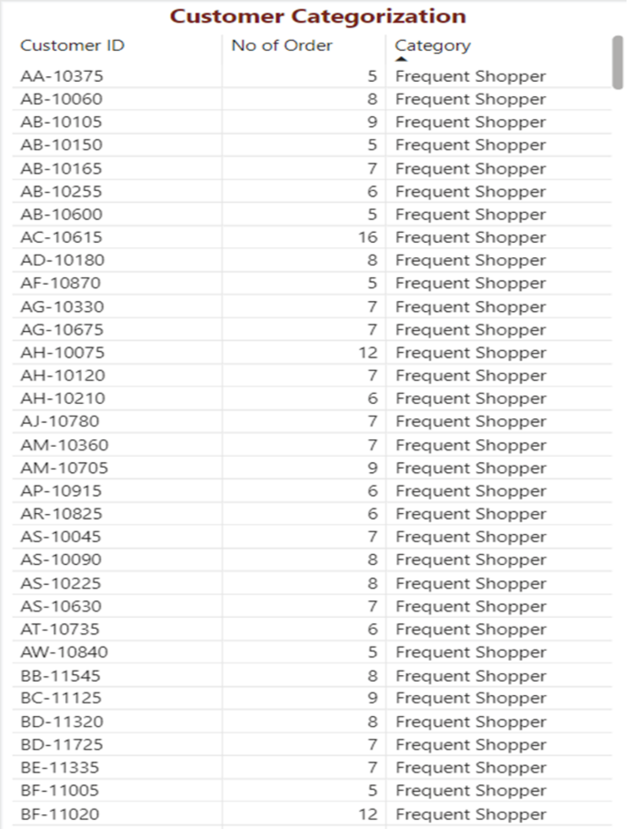
KPI 4: Identifying TOP 5 products based on sales

We visualized the most popular products it in PowerBI in the form of horizontal bars graph with product names on y-axis and the aggregated product sales on the x-axis. From the length of each bar, it is easy to conclude which product was popular. We plotted the most popular five products in the graph and an interesting observation was that in our dataset the Canon image copier which was the most popular product had almost triple the sales of the next popular product.



KPI 5: Identifying the customer categories based on number of orders

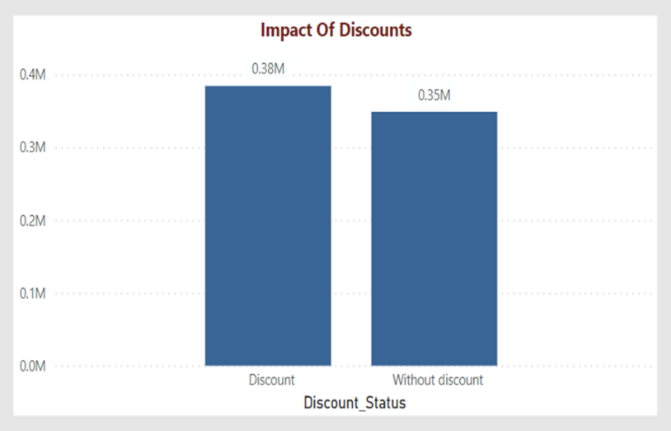
We can observe from the results obtained based on the number of orders how the various customer IDs thus denoting the different customers are classified into the various categories. We visualized it in table form to present a readily observable format and draw conclusions from the dataset.



Customer Categorization

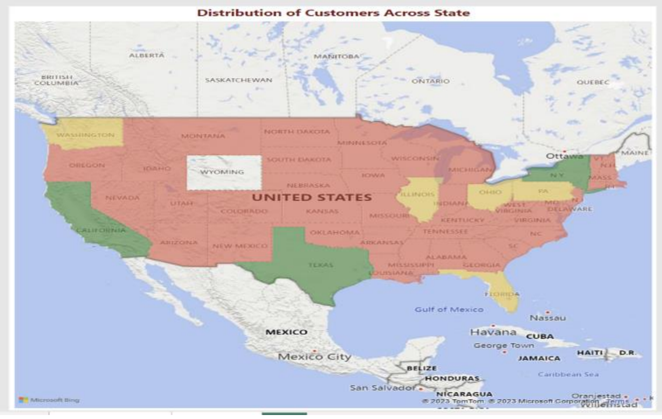
Customer ID	No of Order	Category
AA-10375	5	Frequent Shopper
AB-10060	8	Frequent Shopper
AB-10105	9	Frequent Shopper
AB-10150	5	Frequent Shopper
AB-10165	7	Frequent Shopper
AB-10255	6	Frequent Shopper
AB-10600	5	Frequent Shopper
AC-10615	16	Frequent Shopper
AD-10180	8	Frequent Shopper
AF-10870	5	Frequent Shopper
AG-10330	7	Frequent Shopper
AG-10675	7	Frequent Shopper
AH-10075	12	Frequent Shopper
AH-10120	7	Frequent Shopper
AH-10210	6	Frequent Shopper
AJ-10780	7	Frequent Shopper
AM-10360	7	Frequent Shopper
AM-10705	9	Frequent Shopper
AP-10915	6	Frequent Shopper
AR-10825	6	Frequent Shopper
AS-10045	7	Frequent Shopper
AS-10090	8	Frequent Shopper
AS-10225	8	Frequent Shopper
AS-10630	7	Frequent Shopper
AT-10735	6	Frequent Shopper
AW-10840	5	Frequent Shopper
BB-11545	8	Frequent Shopper
BC-11125	9	Frequent Shopper
BD-11320	8	Frequent Shopper
BD-11725	7	Frequent Shopper
BE-11335	7	Frequent Shopper
BF-11005	5	Frequent Shopper
BF-11020	12	Frequent Shopper

In Power BI we used bar graphs to visualize the impact of the discount on sales. We have the two bars in the graph where y-axis represented the total amount of sales while x-axis denoted discount category or status denoting the two categories and then from the graph it is clear that the bar of the discount category is higher. While by itself this analysis might not provide much information regarding the patterns in the dataset, concluding that there is indeed higher sales when discounts were given provides a foundational basis for further analysis where we can then focus on specific trends and observed the trends specific to the discount amounts and the products which benefitted the most.



KPI 7: The distribution of customers across different demographics

In our visualization in Power BI, we used a map of the country, and we displayed the state wise findings of the customer counts. We chose four colors based on the customer count we had from each state. We had three states of color green which represented that those states have the highest customer count among the states which was greater than the defined threshold for highest distribution. Then the color yellow represented moderate customer distribution and most of the states had the color red which denoted that while they had enough customer count they did not have as much as the count required to be over the moderate category threshold. One interesting observation from the dataset was that the state Wyoming has no representation in the dataset indicating that no customer residing in that state had ordered in the year 2022 which was the year we considered in the analysis.



IX. CONCLUSION

The project has resulted in a profound transformation of our data management and reporting capabilities. The implementation of a streamlined data architecture, utilizing Snowflake capabilities. The integration of an Extract, Load, Transform (ELT) process within the whole process of extracting data to the schema and then utilizing normalization of the staging schema has significantly elevated data quality, ensuring a robust foundation for reporting. The application of normalization techniques to structure the core tables and related dimensions reflects a strategic approach, minimizing redundancy and optimizing both data integrity and query performance. The meticulous execution of data cleaning steps, including handling missing values, removing duplicates, and standardizing formats, underscores our commitment to unparalleled data quality and reliability. The establishment of a reliable OLAP process along the reporting layer within the main schema has delivered a clean dataset for reporting the KPI's and the seamless Power BI integration, enabling accurate and insightful reporting and analysis.In summary, the project has successfully orchestrated a well-structured, efficient, and reliable data pipeline from raw data ingestion to refined reporting.

- Streamlined data architecture with Snowflake capabilities.
- ELT process ensures high-quality data for Power BI reporting.
- Utilization of normalization techniques for improved data structure.
- Rigorous data cleaning steps for enhanced data quality and reliability.
- Implementation of a data visualization dashboard in Power BI for actionable insights.

X. FUTURE SCOPE

By thoroughly analyzing historical sales data, customer trends, and product preferences, our E-commerce project has created a strong baseline understanding of consumer behavior. However, ample opportunity remains to expand the impact of our analysis through predictive analytics, personalized recommendations, competitive pricing strategies, and beyond. Specifically, we aim to leverage the rich dataset from this project to power advanced solutions for demand forecasting, calculating customer lifetime value, price optimization, and inventory planning. The ultimate goal is data-driven automation that revolutionizes the customer experience while maximizing sales growth and profitability into the future.

Predictive Analysis:
The data we have shows the types of products which will be in demand at any time of the year.
We can use this data to predict future demand of products and keep them in stock and suggest variations to customers.

Customer Lifetime Value(CLV) prediction:
Calculate Customer Lifetime Value (CLV) for different customer segments. This can be done using historical data and predictive analysis on future purchases. This is an important metric to predict business growth and trends.

Price competitiveness:
Scrape data from competitor sites and map them to the data in our warehouse. This can be used in studying price competitiveness of our products. This can also be extended as a price matching solution in future.

ACKNOWLEDGEMENTS

We would like to extend our heartfelt appreciation to all those who contributed to the completion of this project. First and foremost, we are grateful for Prof. Jaideep Srivastava unwavering support and mentorship throughout the project. Their valuable insights, constructive feedback, and dedication to our academic growth have been truly inspiring.

In the course of the project, we utilized various resources, including datasets and software packages. The dataset, sourced from Kaggle, served as a foundational component for analysis. Additionally, codes and programming resources from open-source libraries or online repositories contributed to the implementation of specific functionalities. The use of specific software packages, such as Snowflake for data warehousing and POWERBI for visualization, played pivotal roles in achieving project goals.

REFERENCES

1. https://pages.cs.wisc.edu/~dbbook/ (Textbook)
2. Deming, Chunhua, Sreekanth Dekkati, and Harshith Desamsetti. "Exploratory Data Analysis and Visualization for Business Analytics." Asian Journal of Applied Science and Engineering 7, no. 1 (2018): 93-100.
3. Dageville, Benoit, Thierry Cruanes, Marcin Zukowski, Vadim Antonov, Artin Avanes, Jon Bock, Jonathan Claybaugh et al. "The snowflake elastic data warehouse." In Proceedings of the 2016 International Conference on Management of Data, pp. 215-226. 2016.
4. Widjaja, Surlisa, and Tuga Mauritsius. "The development of performance dashboard visualization with power BI as platform." Int. J. Mech. Eng. Technol 10, no. 5 (2019): 235-249.
5. https://www.temida.si/~bojan/MPS/materials/Data%20Analysis%20Using%20SQL%20and%20Excel.pdf