Final Project	Husam Aldeen Khaled Odat
Course Name	**SQL & Data Analytics**
Instructor	**Bassam Ksasbeh**

1. Abstract*:*

In the field of data analytics, significant patterns, and insights are drawn from massive datasets to support decision-making. It includes a range of techniques and tools to analyze data effectively. In real-life applications, data analytics is essential in various industries, empowering businesses

to learn crucial information and make informed decisions based on the insights they take from the data.

Table of Contents:

1.1. Data Analytics Methods:

There are 2 methods for Data Analytics:

Descriptive Analytics: it focuses on understanding and summarizing historical data to gain insight into what has happened in the past. It makes use of methods including EDA, data visualization, and data aggregation.

Predictive Analytics: it uses statistical modeling techniques to make predictions about upcoming events or trends. To find patterns and connections in the data and extrapolate them into reliable forecasts, it makes use of algorithms and machine learning models. Organizations can predict outcomes and trends thanks to predictive analytics, which makes proactive decision-making and planning easier.

1.2. Uses of data analytics in real life:

Business decision-making (Descriptive): it helps organizations to make informed decisions, optimize strategies, and identify new opportunities that can benefit the business.

Healthcare and Medicine improve patient care (Descriptive and Predictive), personalized medicine, and health management by analyzing the patient's data to identify patterns of disease as a type of descriptive analytics and predicting patient outcomes as a type of predictive analytics.

Security and Fraud Detection (Descriptive and Predictive): identifies anomalies and prevents fraudulent activities by analyzing historical data to find patterns to predict any future robbery activity.

From the above example, we can see how data analytics enables businesses in a variety of sectors to improve operations and gain a competitive edge.

2. Descriptive Analytics:

2.1. Techniques & Examples:

In descriptive analytics, many techniques are employed to summarize an interpreter’s data to gain important insights into past and present trends. Here are some techniques that I worked on

Descriptive Statistics: calculating and analyzing basic statistical measures like:

- Mean: Average value.

- Standard deviation: Data variability
- Min: Smallest Value
- the value below which 25% of the data falls.
- Midpoint (median): A value in the middle.
- a value below which 75% of the data falls.
- Max: The maximum value.

These methods provide a concise summary of the central tendency, variability, range, and distribution shape of the data.

Data Visualization: visual representation like charts, graphs, and plots which I use to see how the data is distributed and to see if a specific pattern in the data is there.

2.1.1 Case Study description: A network traffic dataset

the data set describes a network traffic dataset. It contains information about individual network flows and how attacks affect them, such as the source and destination IP addresses, ports, protocols, timestamp, duration, packet and byte counts, packet lengths, inter-arrival times, flags, header lengths, and other metrics.

The data set can be used for a variety of purposes, such as:

- Network anomaly detection: By identifying unusual patterns in network traffic, the data set can be used to detect malicious activity.

This visualization can be used by security researchers to identify potential attack types based on network traffic patterns. For example, if a security researcher notices a flow with a very high backward inter-arrival time, they may suspect that it is a denial-of-service attack.

2.2 descriptive description:

Exploratory Data Analysis (EDA):

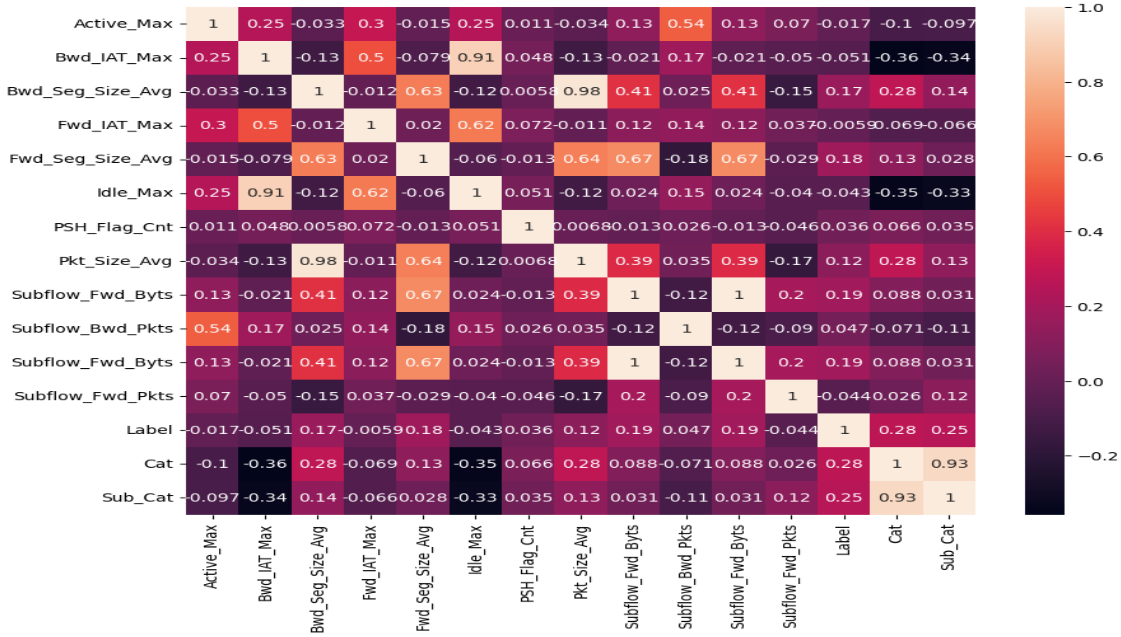
Firstly. I started doing data analysis (EDA) tasks:

- creates a copy of the original Data Frame named tem_df1.
- generates summary statistics using describe for all columns in tem_df1.
- checks for missing values with null().sum().
- identifies and removes duplicated rows with drop duplicates ().
- provides information about the Data Frame’s columns with info().

- calculates the number of unique values for each column with unique().
- selects specific columns of interest and stores them in the Column Data Frame.

Label Encoding: Categorical columns in Column are label encoded using Label Encoder from sklearn preprocessing. This is typically done to convert categorical data into numerical format for machine learning.

Correlation Heatmap: To compute the correlation matrix between columns in Column and creates a heatmap using seaborn and matplotlib to visualize the correlations.

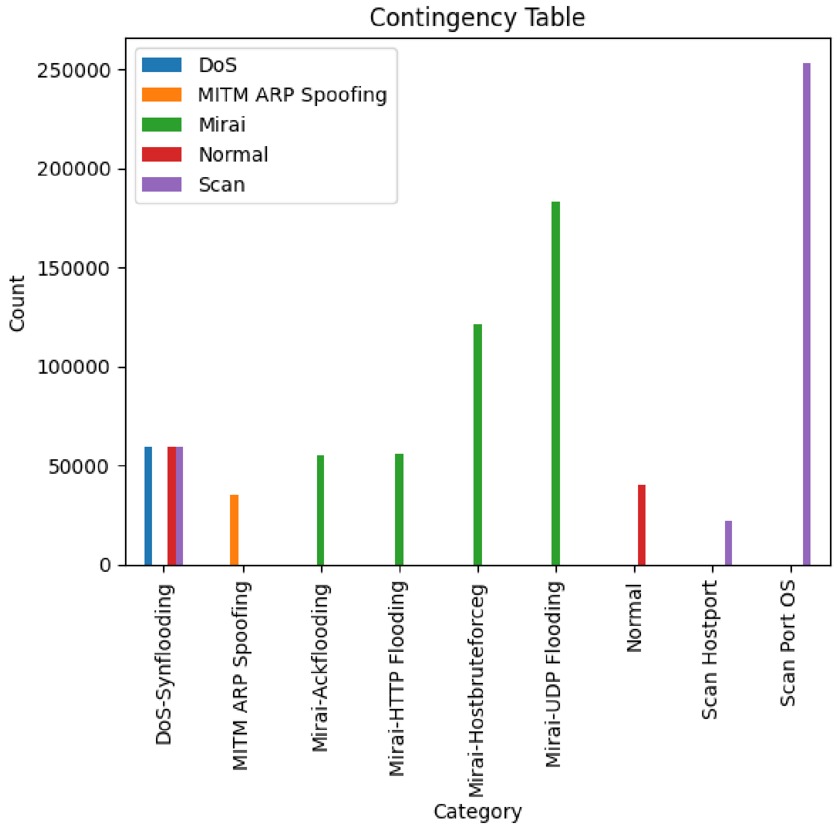


Contingency Table: A contingency table is created using crosstab to show the relationship between two categorical columns, 'Sub Cat' and 'Cat'. The row and column percentages are also calculated and displayed.

Cat	**DoS**	**MITM ARP Spoofing**	**Mirai**	**Normal**	**Scan**
Sub_Cat					
DoS-Synflooding	59391	0	0	0	0
MITM ARP Spoofing	0	35377	0	0	0
Mirai-Ackflooding	0	0	55124	0	0
Mirai-HTTP Flooding	0	0	55818	0	0

Cat	DoS	MITM ARP Spoofing	Mirai	Normal	Scan
Sub_Cat					
Mirai-Hostbruteforceg	0	0	121181	0	0

Data Visualization: To various data visualizations, to explore the distribution and relationships within the data. These visualizations are specific to different subcategories ('Sub_Cat').

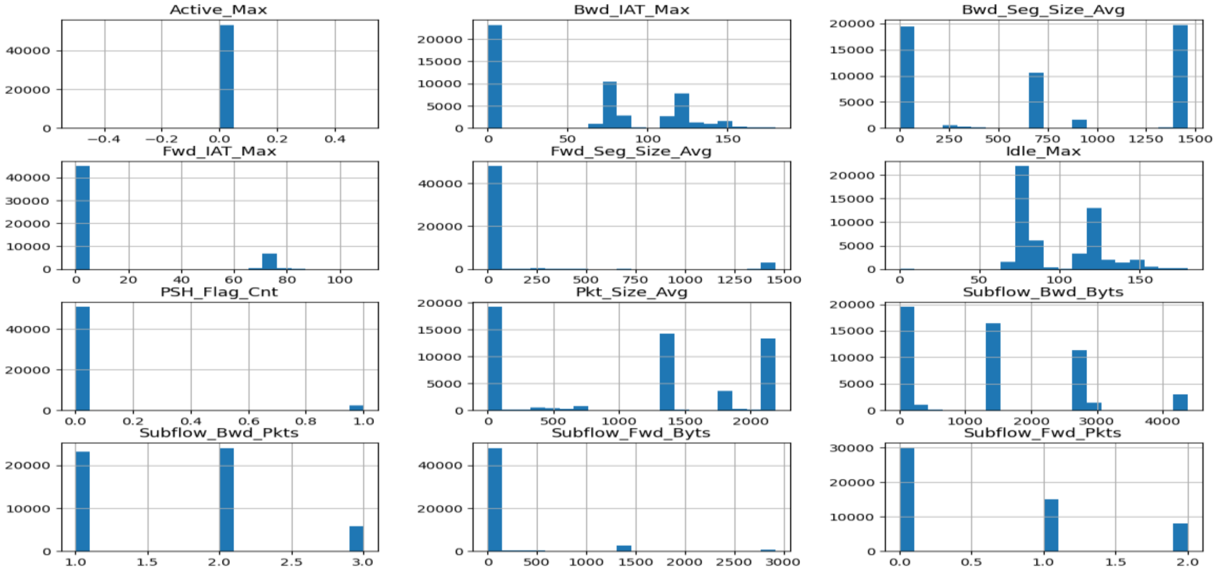


The diagram shows the number of network attacks by attack type and Sub category. The rows of the table show the attack type, and the columns show the category. The values at the intersection of each row and column show the number of attacks of that type in that category.The table shows that the most common attack type is (scan port OS), with over 250,000 attacks.

Finally, Thats performs a comprehensive exploratory data analysis on the IoT network intrusion dataset, including data cleaning, visualization, and statistical analysis by subcategory. That aims to provide insights into the dataset's characteristics and relationships between variables.

2.2.1 Filter Data for "Risk of cancellation level" Cases

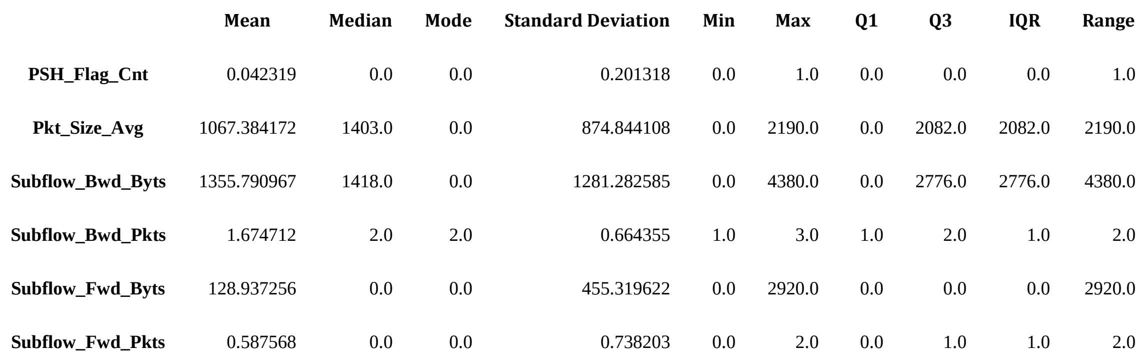
To filter the data for the "Scan Port OS" cases and then generate histograms for those features using Matplotlib.



This provided computes and displays descriptive statistics for the numeric columns of the "Scan Port OS" risk level or category. Here's a breakdown of the displayed statistics:

```
Scan Port OS
```

	Mean	Median	Mode	Standard Deviation	Min	Max	Q1	Q3	IQR	Range
Active_Max	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0
Bwd_IAT_Max	57.522488	73.0	0.0	54.393341	0.0	180.0	0.0	116.0	116.0	180.0
Bwd_Seg_Size_Avg	704.326600	709.0	0.0	604.228784	0.0	1460.0	0.0	1388.0	1388.0	1460.0
Fwd_IAT_Max	11.190926	0.0	0.0	26.527102	0.0	109.0	0.0	0.0	0.0	109.0
Fwd_Seg_Size_Avg	100.805833	0.0	0.0	340.240941	0.0	1460.0	0.0	0.0	0.0	1460.0
Idle_Max	97.905394	84.0	74.0	26.325240	0.0	180.0	75.0	120.0	45.0	180.0



	Mean	Median	Mode	Standard Deviation	Min	Max	Q1	Q3	IQR	Range
PSH_Flag_Cnt	0.042319	0.0	0.0	0.201318	0.0	1.0	0.0	0.0	0.0	1.0
Pkt_Size_Avg	1067.384172	1403.0	0.0	874.844108	0.0	2190.0	0.0	2082.0	2082.0	2190.0
Subflow_Bwd_Byts	1355.790967	1418.0	0.0	1281.282585	0.0	4380.0	0.0	2776.0	2776.0	4380.0
Subflow_Bwd_Pkts	1.674712	2.0	2.0	0.664355	1.0	3.0	1.0	2.0	1.0	2.0
Subflow_Fwd_Byts	128.937256	0.0	0.0	455.319622	0.0	2920.0	0.0	0.0	0.0	2920.0
Subflow_Fwd_Pkts	0.587568	0.0	0.0	0.738203	0.0	2.0	0.0	1.0	1.0	2.0

These statistics for each numeric column within the "Scan Port OS" risk level. These statistics provide insights into the distribution and characteristics of the data within this specific category.

Additionally, it appears that you are using the statistics dictionary to store statistics Data Frames for different risk levels, making it easy to retrieve and analyze statistics for various subsets of the data. However, in the provided code snippet, only the statistics for the "Scan Port OS" risk level are displayed.

4 .predictive Description:

4.1 machine learning tasks:

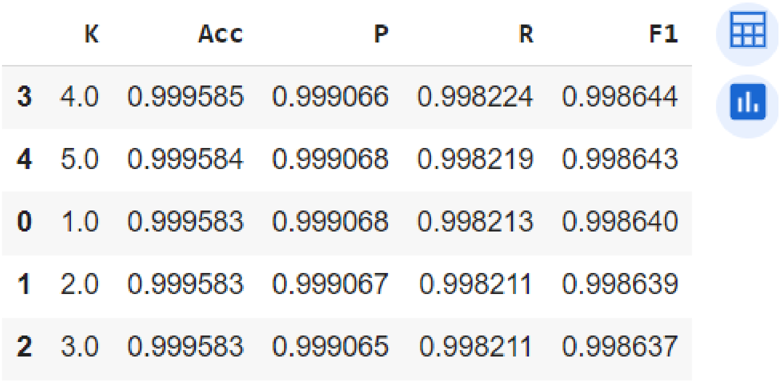
Firstly, To perform machine learning tasks, specifically classification, on a dataset. Here's a description of the main steps and operations performed in the code:

4.2 Data Exploration:

Encoding categorical variables using label encoding. Preprocessing the data for modeling by handling missing values and scaling numerical features.

4.4 Feature Selection with Select K Best:

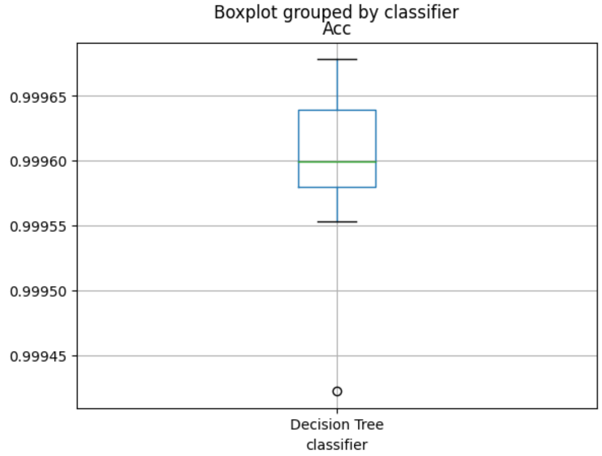
The value of k is determined through an iterative process where different values of k are evaluated to find the optimal number of features.



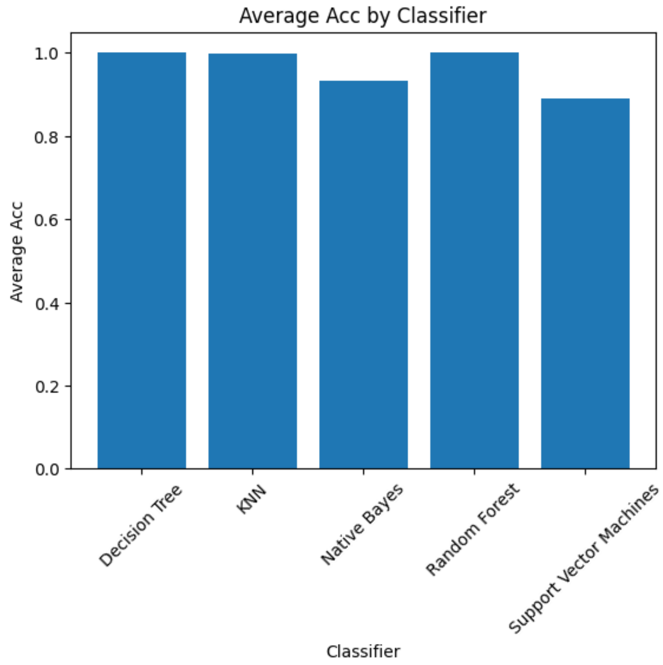
	K	Acc	P	R	F1
3	4.0	0.999585	0.999066	0.998224	0.998644
4	5.0	0.999584	0.999068	0.998219	0.998643
0	1.0	0.999583	0.999068	0.998213	0.998640
1	2.0	0.999583	0.999067	0.998211	0.998639
2	3.0	0.999583	0.999065	0.998211	0.998637

Data is split into training and testing sets. Missing values are handled with imputation.

Feature selection is applied based on the optimal number of features. The selected classifier is trained on the training data. Predictions are made on the testing data. Evaluation metrics such as accuracy, precision, recall, and F1-score are computed and stored. Model Comparison: The code compares the performance of different classifiers based on the average values of accuracy, precision, recall, and F1-score across multiple runs.



Visualization: To generate various plots to visualize the performance of different classifiers and metrics. This includes boxplots and bar charts to compare the classifiers based on accuracy and other metrics.



Overall, the best value is **Decision Tree**. This comprehensive machine learning pipeline for classification tasks on the given dataset. It explores the data, preprocesses it, selects relevant features, trains multiple classifiers, and evaluates their performance, allowing for a comparison of different machine learning algorithms in terms of their predictive accuracy and other metrics.

5 .Power BI

5.1 Database description

There are three tables of data, each with its own set of columns. That break down the description of each table for better understanding.

Customer Information - A unique identifier for each customer, their geographical location, name, birthdate, marital status, occupation, house ownership status, number of cars owned, address, phone number, and date of first purchase.

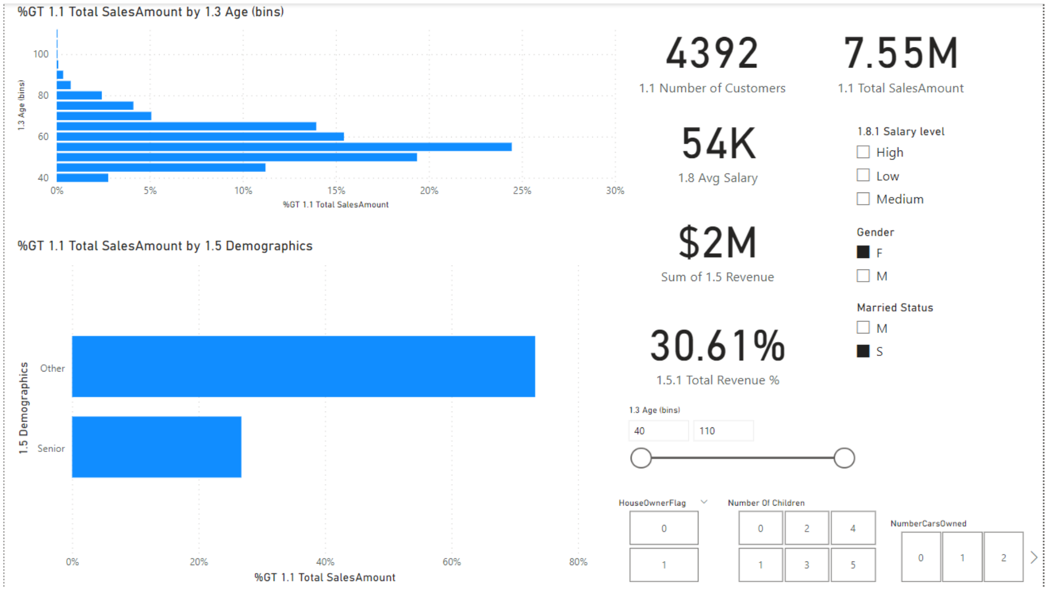
Product Information - A unique identifier for each product, its name, standard cost, color, safety stock level, list price, size, size range, weight, number of days required to manufacture, product line, dealer price, class, model name, description, start and end dates for availability, status, subcategory, and category.

Sales Order Information - A reference to the product being sold, the order date, a key associated with the order date, a reference to the customer making the purchas.

These tables appear to contain data related to customer information, product information, and sales order details, respectively.

5.2 Dashboard 1 :

The graph shows the total sales amount and revenue by customer segmentation criteria. Each customer is assigned to a segment based on their demographic or behavioral characteristics.


%GT 1.1 Total SalesAmount by 1.3 Age (bins)
1.3 Age (bins)
100
80
60
40
0%
5%
10%
15%
20%
25%
30%
%GT 1.1 Total SalesAmount
%GT 1.1 Total SalesAmount by 1.5 Demographics
1.5 Demographics
Other
Senior
0%
20%
40%
60%
80%
%GT 1.1 Total SalesAmount

4392
1.1 Number of Customers

7.55M
1.1 Total SalesAmount

54K
1.8 Avg Salary

1.8.1 Salary level
☐ High
☐ Low
☐ Medium

$2M
Sum of 1.5 Revenue

Gender
■ F
☐ M

Married Status
☐ M
■ S

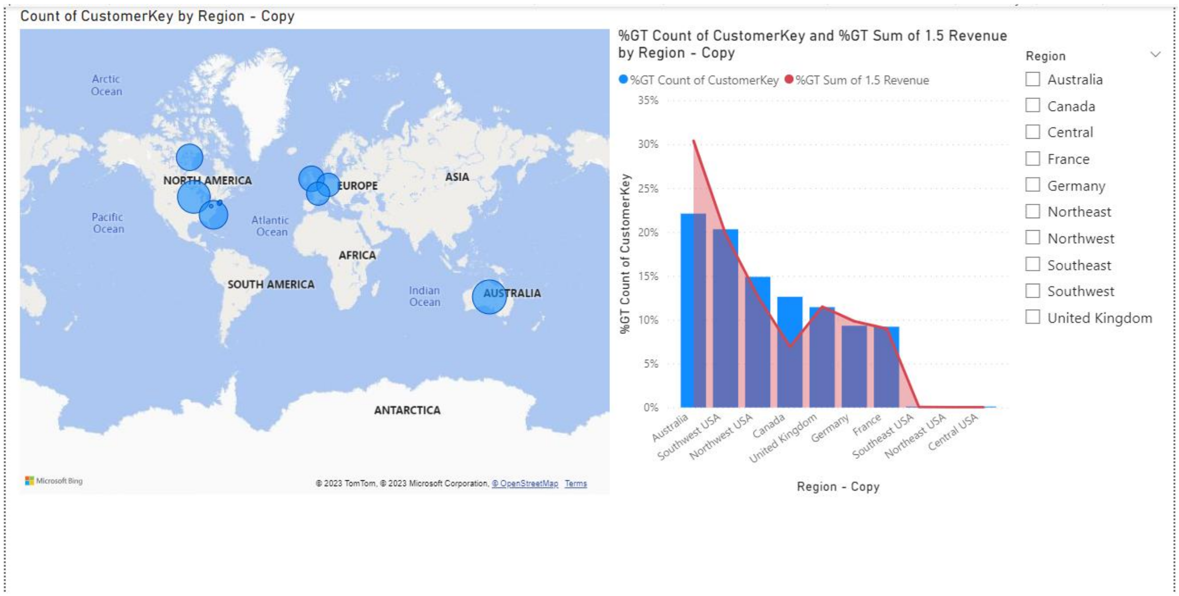
30.61%
1.5.1 Total Revenue %

1.3 Age (bins)
40 110

HouseOwnerFlag
0
1

Number Of Children
0 2 4
1 3 5

NumberCarsOwned
0 1 2



The segments in terms of total sales amount are:

High salary level ,Medium salary level ,Low salary level ,Male ,Female , Age , Married status How many children have, Number of cars have, Own a house and Customer location.

It is also important to consider the specific needs and preferences of each customer segment. For like customers with a high salary level may be more interested in luxury products, while customers with a low salary level may be more interested in budget-friendly products.

By understanding the different customer segments and their needs, businesses can develop targeted marketing strategies and product offerings to maximize sales and profitability.

5.2.1 Creative response:
By studying this map, businesses can identify the areas where they can focus their efforts to grow their sales. For example, a business might decide to invest in marketing campaigns targeting the high salary level and youth level segments.

The map can also be used to identify potential opportunities for new products and services. For example, a business might notice that there is a gap in the market for budget-friendly luxury products. This could be an opportunity for the business to develop a new product line that meets the needs of this underserved segment.

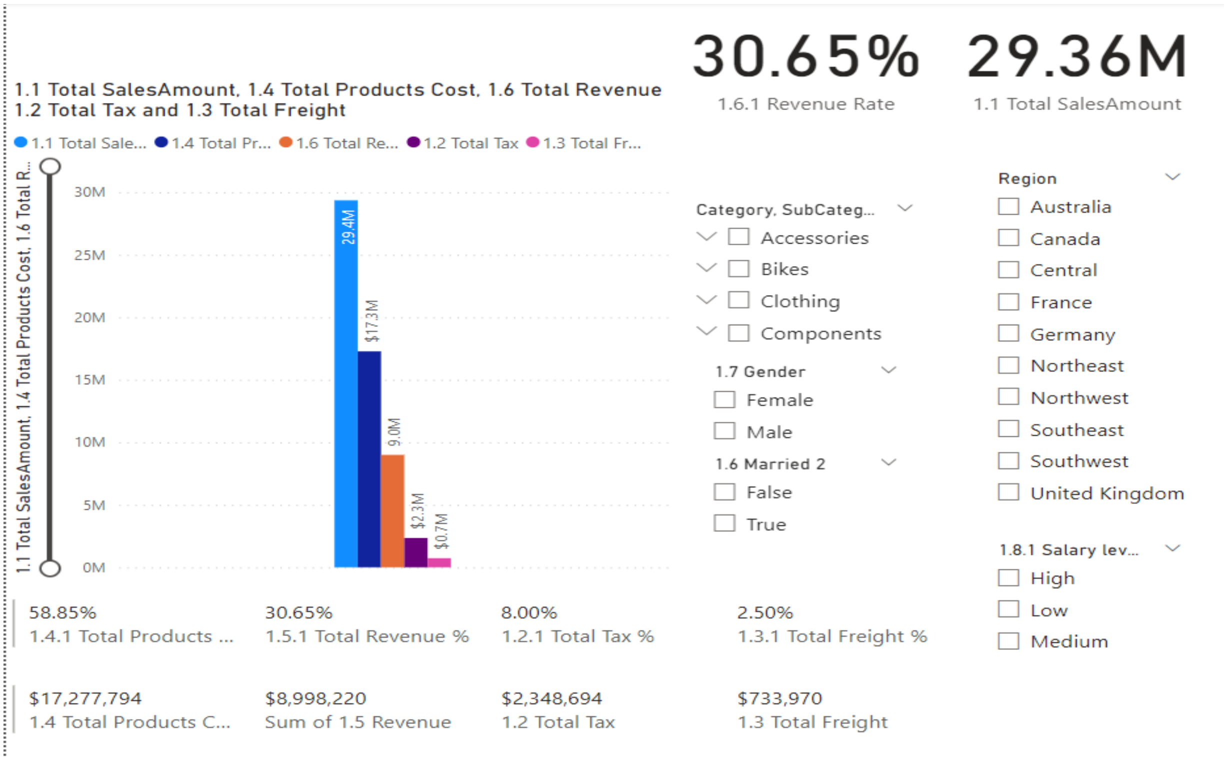
Overall, the graph you sent is a valuable tool for businesses that want to understand their customers and grow their sales.

5.3 Dashboard 2 :

The graph shows the total sales amount, revenue, tax and shipment cost with their percentage to total product price cost by product category, with criteria and characteristics of each category.

The criteria and characteristics for each category include characters like, Product type ,Brand ,Price ,Quality ,Features , and Target audience.

Businesses can use the information in the graph to make informed decisions about product development, marketing, and sales. A business might decide to focus its efforts on the specific category because it is the most popular category among consumers and have the highest revenue rate


1.1 Total SalesAmount, 1.4 Total Products Cost, 1.6 Total Revenue
1.2 Total Tax and 1.3 Total Freight
● 1.1 Total Sale... ● 1.4 Total Pr... ● 1.6 Total Re... ● 1.2 Total Tax ● 1.3 Total Fr...
1.1 Total SalesAmount, 1.4 Total Products Cost, 1.6 Total R...
30M
25M
20M
15M
10M
5M
0M
29.4M
$17.3M
9.0M
$2.3M
$0.7M

30.65%
1.6.1 Revenue Rate

29.36M
1.1 Total SalesAmount

Category, SubCateg...
□ Accessories
□ Bikes
□ Clothing
□ Components

1.7 Gender
□ Female
□ Male

1.6 Married 2
□ False
□ True

Region
□ Australia
□ Canada
□ Central
□ France
□ Germany
□ Northeast
□ Northwest
□ Southeast
□ Southwest
□ United Kingdom

1.8.1 Salary lev...
□ High
□ Low
□ Medium

58.85%
1.4.1 Total Products ...

30.65%
1.5.1 Total Revenue %

8.00%
1.2.1 Total Tax %

2.50%
1.3.1 Total Freight %

$17,277,794
1.4 Total Products C...

$8,998,220
Sum of 1.5 Revenue

$2,348,694
1.2 Total Tax

$733,970
1.3 Total Freight

.

5.3.1 Creative Response:

The graph you sent is like a map of the different product categories, with the total sales amount and revenue representing the elevation of each category. The highest peaks are the categories that are most popular among consumers.

By studying this map, businesses can identify the areas where they can focus their efforts to grow their sales. A business might decide to invest in product development and marketing campaigns targeting the electronics category.

The map can also be used to identify potential opportunities for new products and services. A business might notice that there is a gap in the market for affordable electronics products. This could be an opportunity for the business to develop a new product line that meets the needs of this underserved segment.

Overall, the graph you sent is a valuable tool for businesses that want to understand their customers and grow their sales.

6. Ecommerce database project:

6.1 Description

An e-commerce database is designed to store and manage data related to an online retail business. It includes various tables and relationships to efficiently handle products, customers, orders, payments, and more. Here's a description of the key components and tables commonly found in an e-commerce database.

6.1.1 Products Table:

Stores information about the products available for sale.

6.1.2 Product Categories Table:

Contains categories or departments to organize products.

6.1.3 Products Inventory Table:

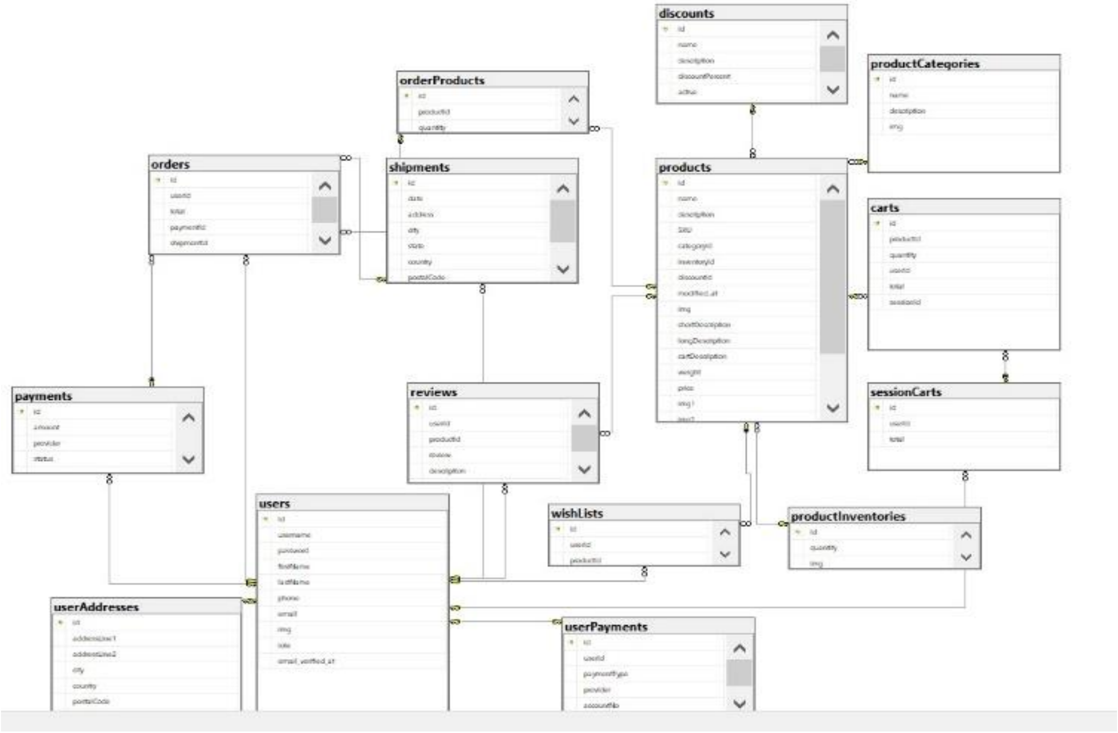
Tracks the quantity of each product in stock.
May include fields like product ID and available quantity.

6.1.4 Discounts Table:

Stores information about discounts or promotions offered on products.
Fields can include discount name, description, discount percentage, and active status.

6.1.5 Customers/Users Table:

Holds data about registered customers or users.



Fields may include username, password, name, contact details, email, role (customer, admin), and profile image. includes foreign key relationships with user addresses and payment methods.

6.1.6 Carts Table:

Represents shopping carts where customers store items before making a purchase.
Contains information about products in the cart, including product ID, quantity, and total cost.
Linked to user profiles or sessions.

6.1.7 Orders Table:

Records completed purchases made by customers.
Includes order details like order number, customer ID, total cost, payment ID, and shipping details. May have relationships with order items and payment records.

6.1.8 Order Items/Line Items Table:

Stores information about individual products within an order. Includes product ID, quantity, and associated order ID.Links products to orders.

6.1.9 Payments Table:

Records payment transactions for orders.
Contains data such as payment amount, payment provider, status (pending, completed), and user ID.Often linked to orders.

6.1.11 Session Carts Table:

Like regular carts but tied to user sessions for guest users. Contains user ID, total cost, and product details. Used to temporarily store items for unregistered users.

6.2 Conclusion

This description provides an overview of the essential tables and their relationships in an e-commerce database. The specific structure and fields may vary depending on the needs of the online retail business and the complexity of the system. Properly designed databases enable efficient management of products, orders, payments, and customer interactions in an e-commerce platform.

These relationships enable the database to maintain data integrity and provide efficient ways to retrieve and manipulate data for various e-commerce operations such as placing orders, managing user profiles, and handling product catalog information.

References

Plotting Histogram in Python using Matplotlib - GeeksforGeeks

Histogram of an image using matplotlib in Python - CodeSpeedy

Using matplotlib how could I plot a histogram with given data in python - Stack Overflow

pylab_examples example code: histogram_demo.py — Matplotlib 1.2.1 documentation

Matplotlib Histogram Plot - Tutorial and Examples (stackabuse.com)