DEBRE BERHAN UNIVERSITY

COLLEGE OF COMPUTING

DEPARTMENT OF SOFTWARE ENGINEERING

FUNDAMENTALS OF BIG DATA ANALYTICS AND BUSINESS INTELLIGENCE (SEng5112)

INDIVIDUAL ASSIGNMENT

Eyob chiksa

0489/13

Submitted to: Derbew Felasman(MSc)

February, 2025

Table of Contents

Introduction

This report presents a comprehensive analysis of an e-commerce dataset containing over 1.3 million transactions. The dataset includes fields such as event_time, order_id, product_id, category_code, brand, price, and user_id. The primary objective of this project is to construct a complete data pipeline—from data extraction to visualization—and derive actionable insights into sales trends, customer behavior, and product performance.

event_time	order_id	product_id	category_id	category_code	brand	price	user_id
2020-04-24 11:50:	22943599320545	151596622350908	22681054266481	electronics.tablet	samsung	162.01	1515915625441993984
2020-04-24 11:50:	22943599320545	151596622350908	22681054266481	electronics.tablet	samsung	162.01	1515915625441993984
2020-04-24 14:37:	22944440240580	227394831905718	22681054301629	electronics.audio.h	huawei	77.52	1515915625447879434
2020-04-24 14:37:	22944440240580	227394831905718	22681054301629	electronics.audio.h	huawei	77.52	1515915625447879434
2020-04-24 19:16:	22945842631540	227394831681742	2268105471367840086		karcher	217.57	1515915625443148002
2020-04-26 08:45:	22957165214496	151596622350926	22681054426368	furniture.kitchen.ta	maestro	39.33	1515915625450382722
2020-04-26 09:33:	22957405947497	151596622350910	22681054281665	electronics.smartpl	apple	1387.01	1515915625448766480
2020-04-26 09:33:	22957405947497	151596622350910	22681054281665	electronics.smartpl	apple	1387.01	1515915625448766480
2020-04-26 09:33:	22957405947497	151596622350910	22681054281665	electronics.smartpl	apple	1387.01	1515915625448766480
2020-04-26 09:33:	22957405947497	151596622350910	22681054281665	electronics.smartpl	apple	1387.01	1515915625448766480
2020-04-26 14:55:	22959024902032	227394831174231	22681053938487	appliances.kitchen.	lg	462.94	1515915625450561165
2020-04-26 23:35:	22961643244874	151596622350925	22681054024470	appliances.persona	polaris	30.07	1515915625446798439
2020-04-27 07:24:	22964004809909	227394830866369	23744989140005	electronics.video.t	samsung	416.64	1515915625450899340
2020-04-27 14:57:	22966282379308	151596622350908	22681054100219	computers.compo	intel	91.41	1515915625451131565
2020-04-27 14:57:	22966282379308	151596622350908	22681054100219	computers.compo	intel	91.41	1515915625451131565
2020-04-27 14:57:	22966282379308	151596622350908	22681054100219	computers.compo	intel	91.41	1515915625451131565
2020-04-28 02:21:	22969727010608	151596622350910	2268105402774193030		philips	23.13	1515915625451212869
2020-04-28 03:47:	22970160082310	151596622350908	22681054072201	computers.notebo	asus	509.24	1515915625443158850
2020-04-28 04:25:	22970347371993	151596622350971	2268105635507732512			6.94	1515915625447779982
2020-04-28 04:25:	22970347371993	151596622350971	2268105635507732512			6.94	1515915625447779982
2020-04-28 09:01:	22971740445558	227394822295729	22681054092250	computers.periphe	samsung	254.61	1515915625442675260
2020-04-28 09:01:	22971740445558	227394822295729	22681054092250	computers.periphe	samsung	254.61	1515915625442675260
2020-04-28 11:36:	22972520544075	227394830317754	22681054079331	computers.periphe	epson	164.33	1515915625450916989
2020-04-28 11:36:	22972520544075	227394830317754	22681054079331	computers.periphe	epson	164.33	1515915625450916989
2020-04-28 11:36:	22972520544075	227394830317754	22681054079331	computers.periphe	epson	164.33	1515915625450916989
2020-04-29 03:25:1	22977294079109	151596622350910	2268105427528974760		sbs	0.02	1515915625441708399
2020-04-29 03:25:1	22977294079109	151596622350910	2268105427528974760		sbs	0.02	1515915625441708399
2020-04-29 04:46:	22977704050598	151596622350908	22681054281665	electronics.smartpl	samsung	300.9	1515915625451641617
2020-04-29 06:20:	22978177167586	151596622351017	2268105442242593506		geyzer	6.23	1515915625451580783

Data Extraction

Data extraction is the initial phase of the ETL process. The dataset, obtained from Kaggle, was downloaded in CSV format. The following steps were carried out:

- **Loading the Dataset:** The CSV file was imported using Python's pandas library for an initial review. It was then loaded into a pandas DataFrame, facilitating seamless manipulation and analysis.
- **Handling Missing Data:** The dataset contained missing values and inconsistencies, requiring a comprehensive data cleaning process. Missing values in key columns, such as price, were identified and addressed.

Data Transforming Process

The dataset was refined for accuracy and consistency through:

- Converting event time to datetime format
- Removing duplicates
- Handling missing values and outliers
- Converting IDs to integers

These transformations ensured data integrity, making it more reliable for analysis. Converting timestamps enabled proper time-based analysis, while removing duplicates prevented redundancy. Addressing missing values and outliers improved data quality, and standardizing IDs ensured consistency across records.

Data Storage (PostgreSQL)

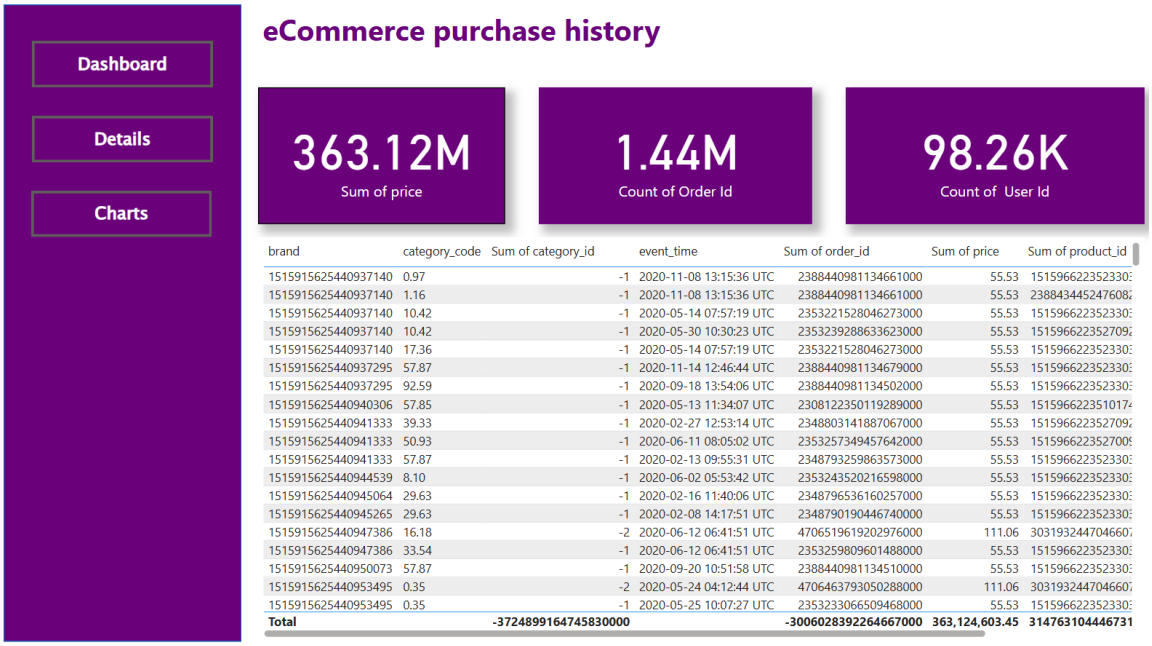
The cleaned dataset was stored in a PostgreSQL relational database. The following schema was used:

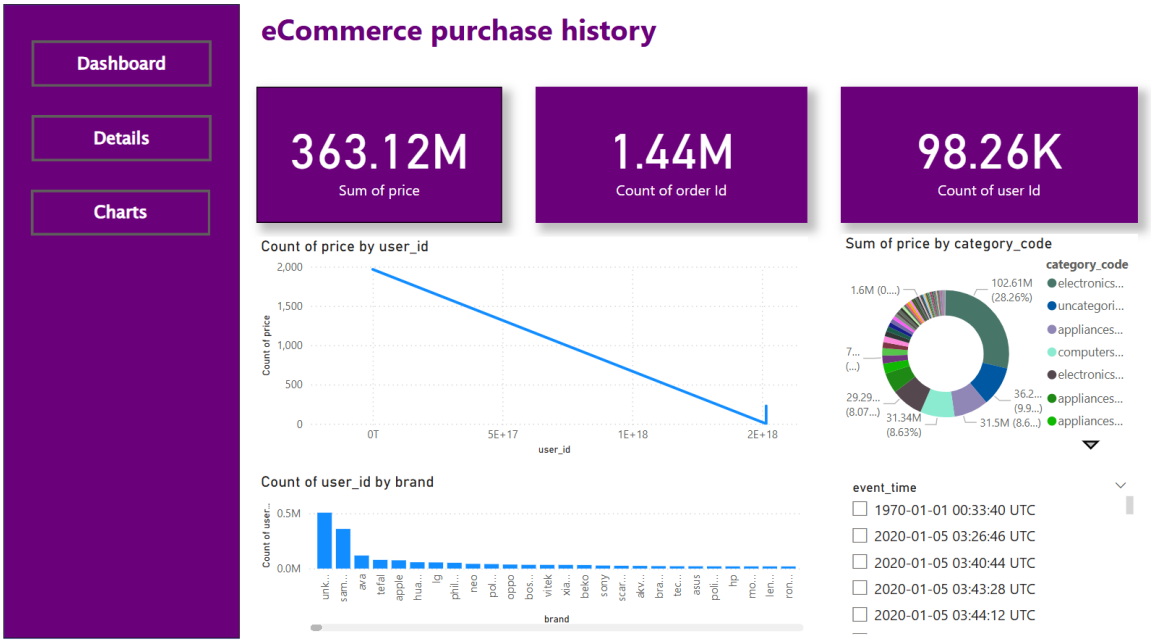
Column Name	Data Type	Description
event_time	TIMESTAMP	Timestamp of the transaction
order_id	BIGINT	Unique identifier for the order
product_id	BIGINT	Unique identifier for the product
category_id	BIGINT	Unique identifier for the category
category_code	TEXT	Product category (e.g., electronics)
brand	TEXT	Brand of the product
price	FLOAT	Price of the product
user_id	BIGINT	Unique identifier for the customer

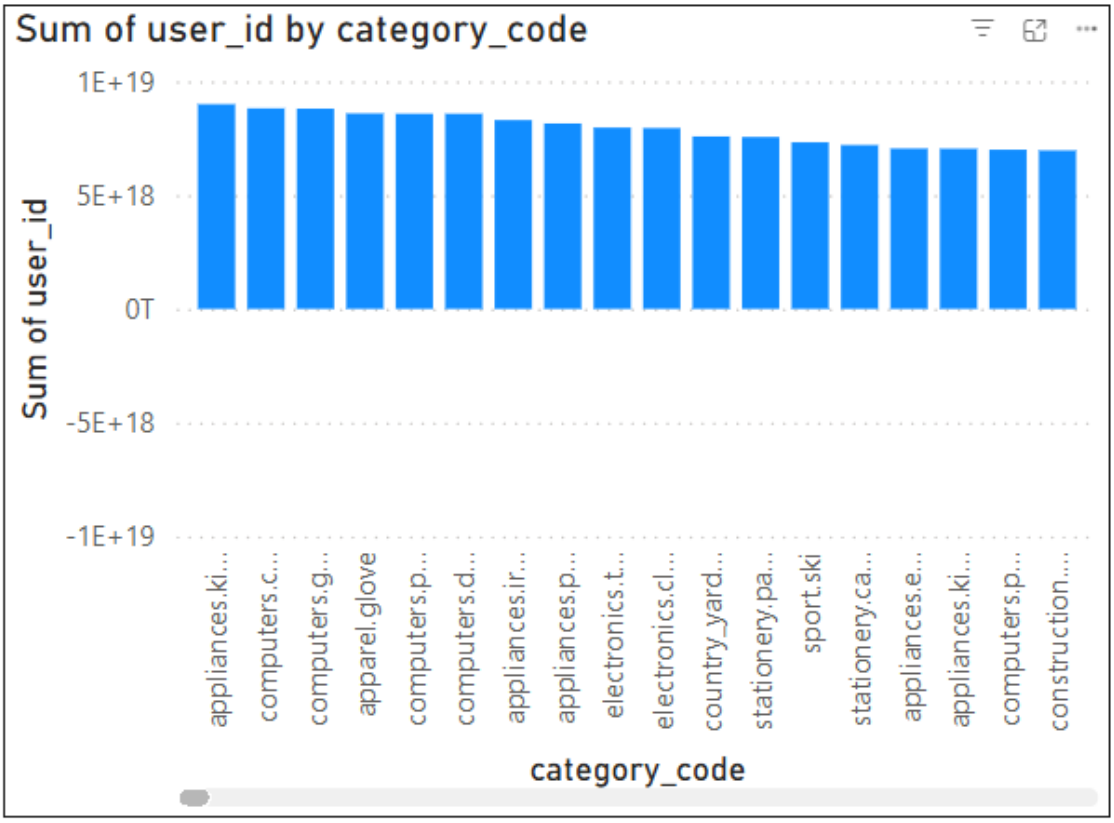
Data Visualization and Insights

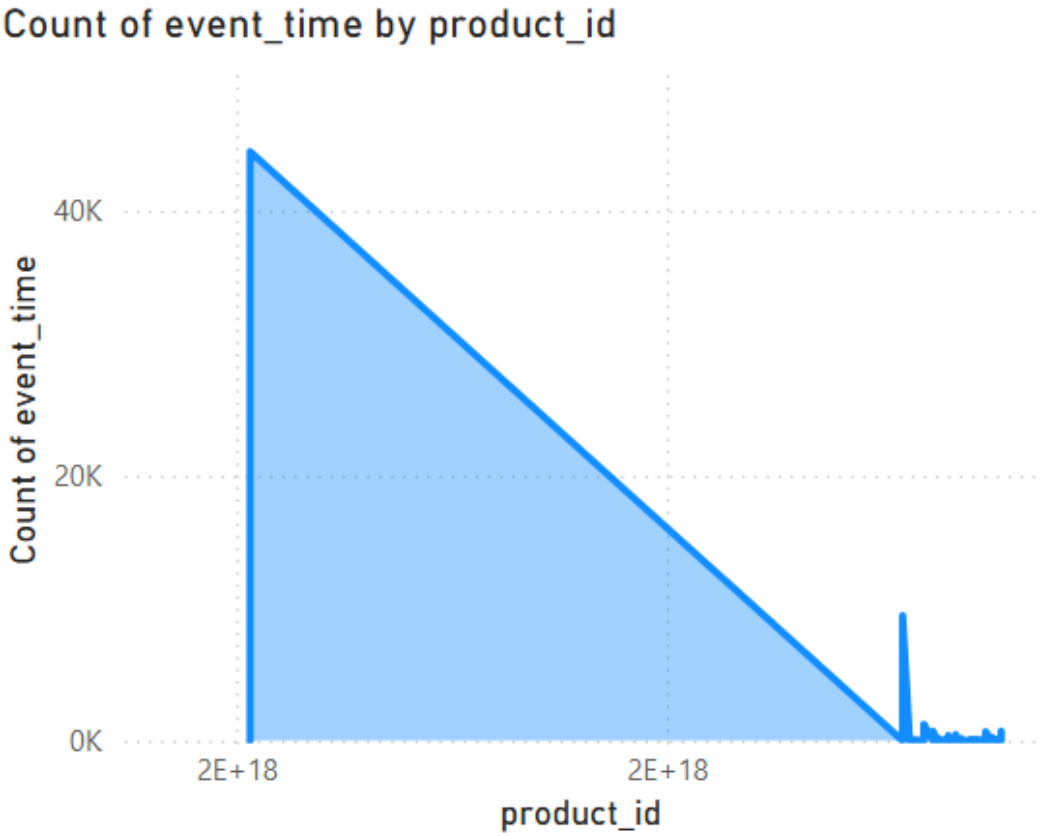
Microsoft Power BI was used to create interactive dashboards to visualize key insights. The following visualizations were generated:

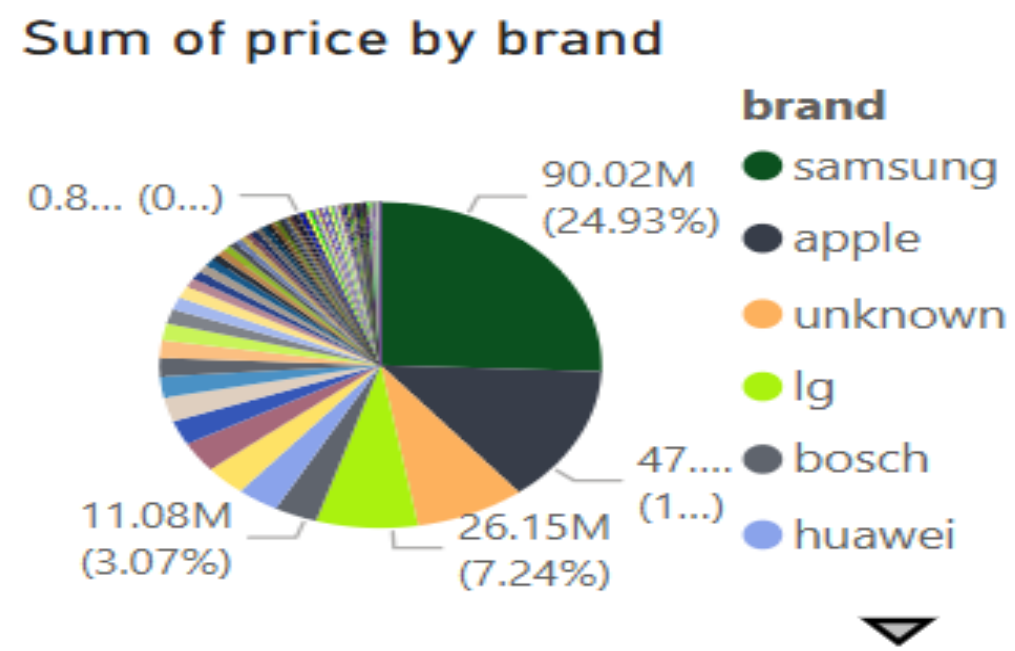
https://app.powerbi.com/links/gUxh-oB1xa?ctid=1695066a-e388-40d1-8ed5-5d0b28ba9f80&pbi_source=linkShare for more visualization pictures

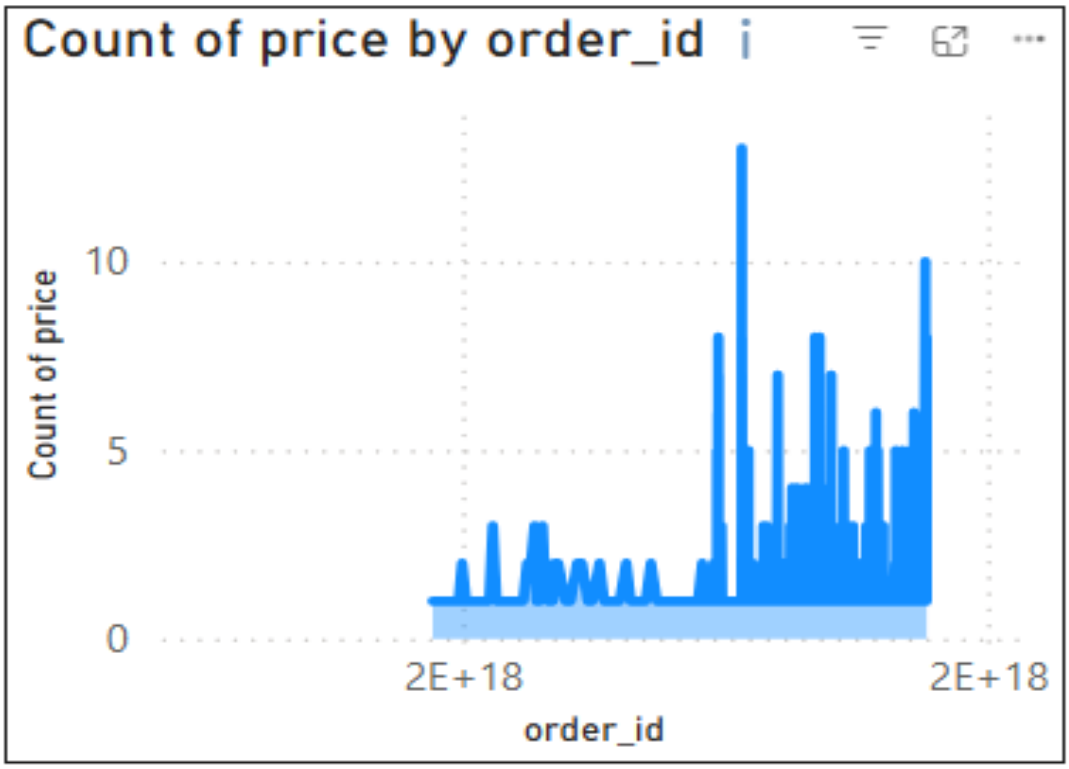


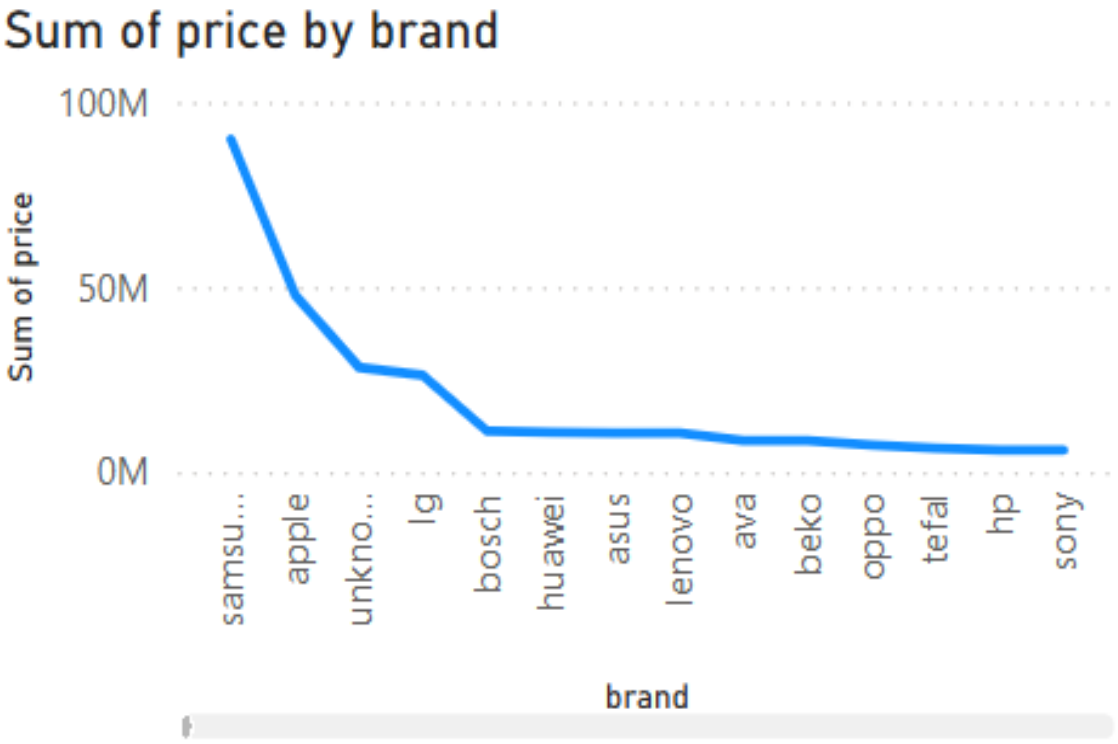



Sum of user_id by category_code
1E+19
5E+18
0T
-5E+18
-1E+19
Sum of user_id
appliances.ki...
computers.c...
computers.g...
apparel.glove
computers.p...
computers.d...
appliances.ir...
appliances.p...
electronics.t...
electronics.cl...
country_yard...
stationery.pa...
sport.ski
stationery.ca...
appliances.e...
appliances.ki...
computers.p...
construction....
category_code


Count of event_time by product_id
40K
20K
0K
Count of event_time
2E+18
2E+18
product_id


Sum of price by brand
brand
samsung
apple
unknown
lg
bosch
huawei
90.02M
(24.93%)
0.8... (0...)
47....
(1...)
11.08M
(3.07%)
26.15M
(7.24%)

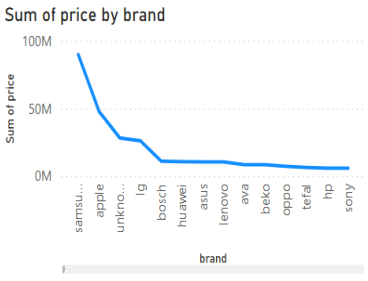

Count of price by order_id
Count of price
10
5
0
2E+18
2E+18
order_id

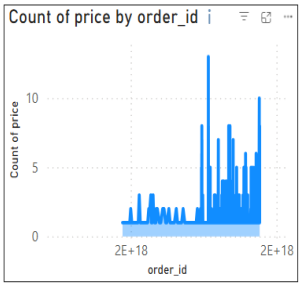


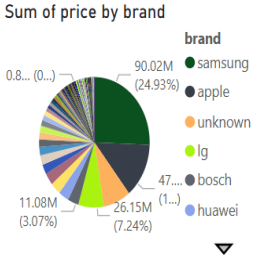
Dashboard

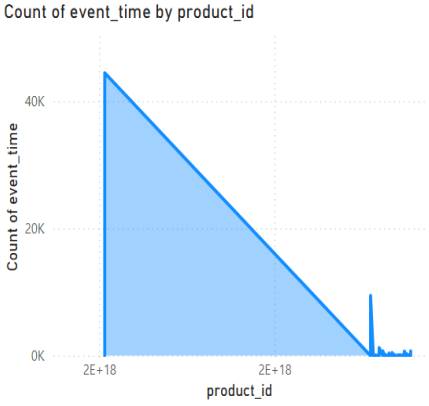
Details

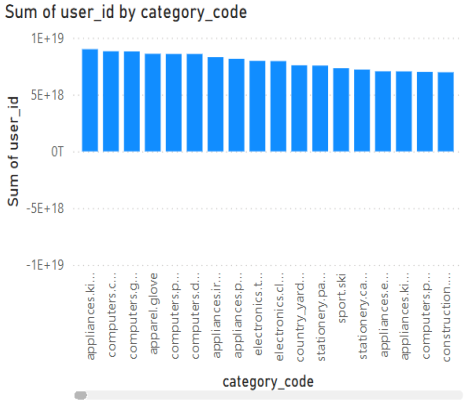
eCommerce purchase history











Conclusion

This project successfully implemented an end-to-end data pipeline for processing and analyzing e-commerce data. The insights derived from the dataset can help businesses optimize pricing strategies, improve inventory management, and enhance customer engagement.