DEBRE BERHAN UNIVERSITY

INSTITUTION Of TECHNOLOGY

COLLAGE Of COMPUTING

DEPARTMENT Of SOFTWARE ENGINEERING

FUNDAMENTAL OF BIG DATA ANALYTICS AND BUSINESS INTELLIGENCY(SEng5112)

Prepared by: - Name	**ID No**
Name: Dabi Haile	DBUR/2054/13

Submitted to: Mr. Derbew Felasman(MSc)

February, 2025

Debre Berhan, Ethiopia

Table of Contents

List of Figures

1. Introduction

This project builds an ETL (Extract, Transform, Load) pipeline to integrate Online Retail data from UCI Machine Learning Repository with Telegram messages from multiple channels(3 channels). The goal is to automate data collection, clean and store it in PostgreSQL, and visualize insights using Power BI. The pipeline is designed to enhance decision-making by combining structured Online Retail transactions with unstructured customer interactions from Telegram.

2. Data Sources

- **Online RetailData (UCI Machine Learning Repository)**

  Format: CSV

  Link: https://archive.ics.uci.edu/dataset/352/online+retail

  Fields: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

- **Telegram Data**

  Channels: easybuyethiopia, Ecommerceaddis, jiji_shop_ethiopia

  Extracted Fields: Date, Message, User ID

3. Data Extraction

- **Online Retail Data Extraction**

  Loaded CSV using pandas (read_csv())

  Validated and cleaned before transformation

- **Telegram Data Extraction**

Step 1: Go to https://my.telegram.org

Step 2: Navigate to API Development Tools

Step 3: Create a new application to get api_id and api_hash

Step 4: Use these credentials in the Telethon client setup

elethon to Extracted date, message text, and user ID

4. Data Cleaning

- **Online RetailData**

  Dropped missing values (dropna())

  Removed duplicates (drop_duplicates())

  Converted InvoiceDate to datetime, UnitPrice to float, and CustomerID to int

- **Telegram Data**

Preserved negative User IDs for groups

Standardized date format (datetime)

Removed duplicate messages

Ensured user_id values fit within PostgreSQL BIGINT limits

5. Database Schema

Rental_Dataset Table

```
CREATE TABLE Rental_Dataset (
    order_id SERIAL PRIMARY KEY,
    InvoiceNo VARCHAR(50),
    StockCode VARCHAR(20),
    Description TEXT,
    Quantity INTEGER,
    InvoiceDate TIMESTAMP,
    UnitPrice DECIMAL(10,2),
    CustomerID BIGINT,
    Country VARCHAR(100)
);
```

telegram_messages Table

```
CREATE TABLE telegram_messages (
    message_id SERIAL PRIMARY KEY,
    date TIMESTAMP,
    message TEXT,
    user_id BIGINT
);
```

6. Data Loading

- **Database Connection**

  Connected using psycopg2

  Created tables with appropriate constraints

- **Online RetailData Insertion**

```
INSERT INTO Rental_Dataset (InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country)
VALUES (%s, %s, %s, %s, %s, %s, %s, %s);
```

- **Telegram Data Insertion**

```
INSERT INTO telegram_messages (date, message, user_id)
```

VALUES (%s, %s, %s);

7. Development Tools

VS Code - Used for writing and testing Python scripts

Jupyter Notebook - Used for interactive data analysis and debugging

PostgreSQL - Database for storing structured data

Power BI - Visualization and reporting

GitHub - Version control and collaboration

8. Data Visualization

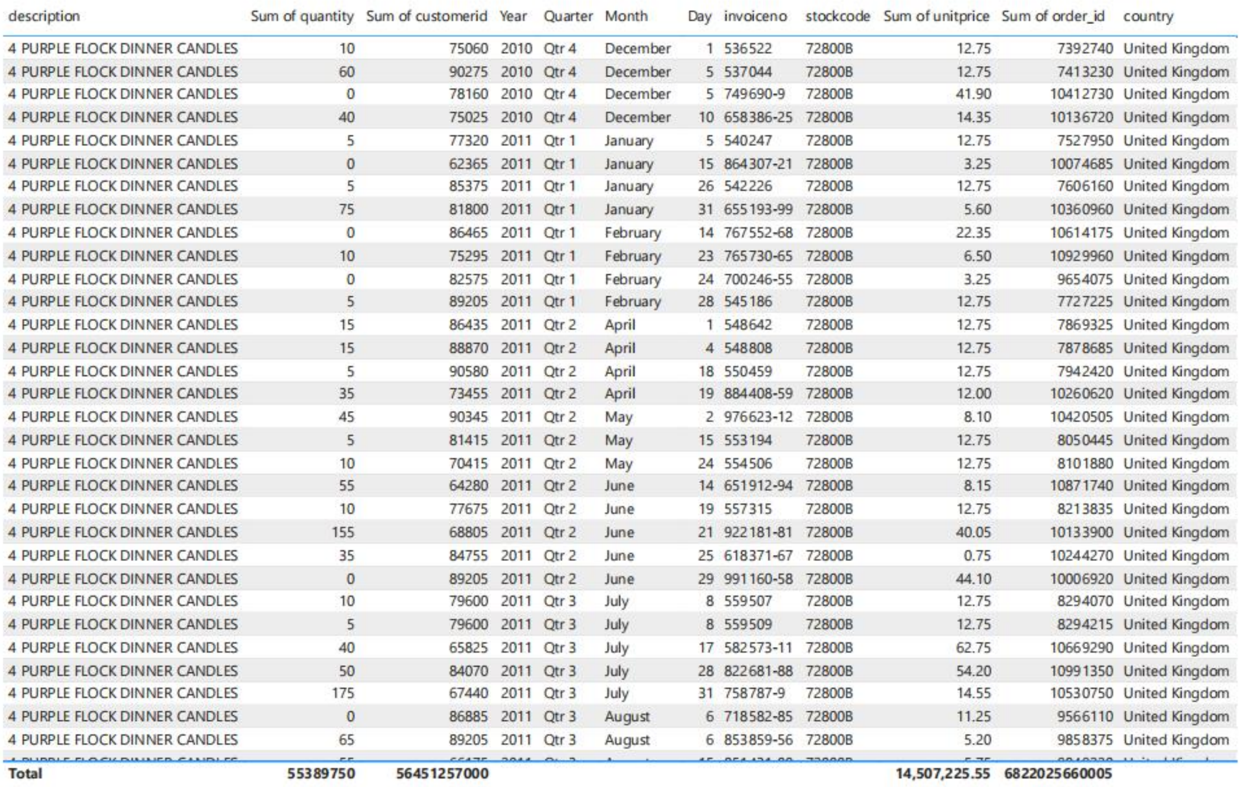
Power BI Dashboards

Online RetailSales Trends (monthly and yearly patterns)

Customer Segmentation (purchase behavior analysis)

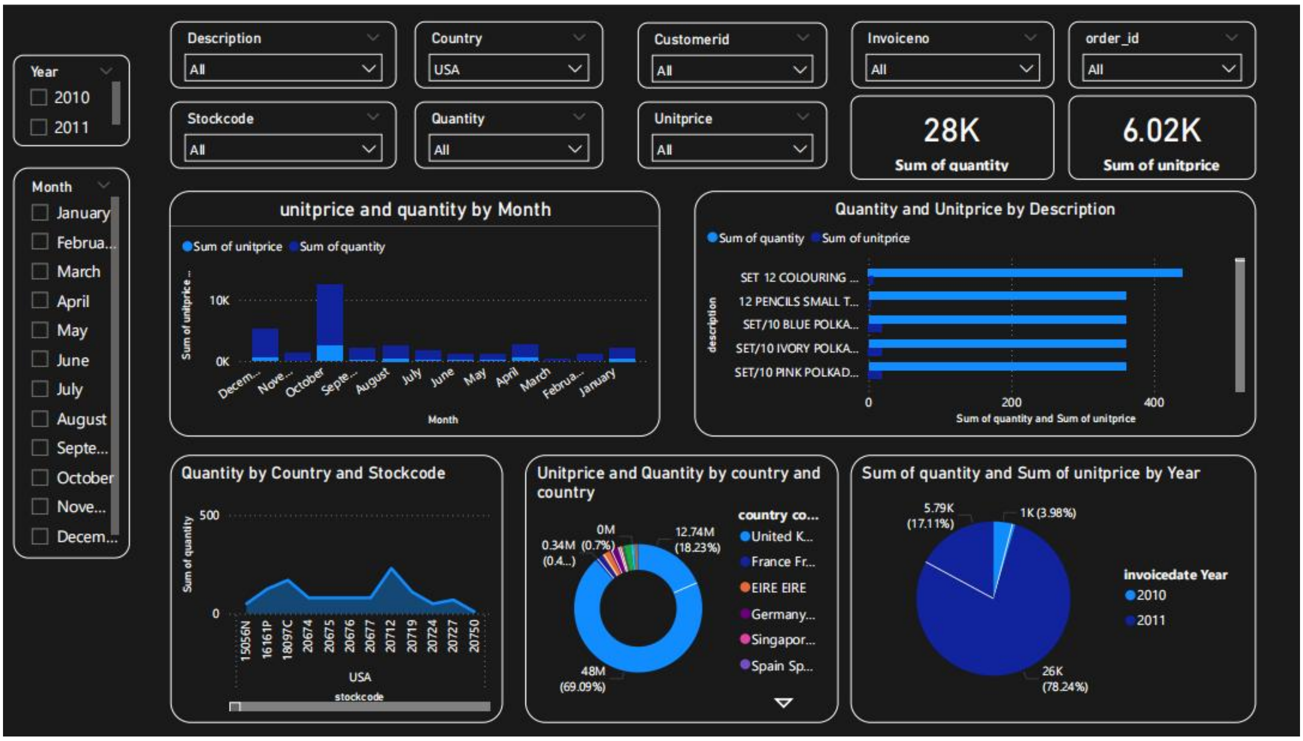
Telegram Message Analysis (sentiment & trending topics)

9. Reports on Rental Dataset Using Power BI



description	Sum of quantity	Sum of customerid	Year	Quarter	Month	Day	invoiceno	stockcode	Sum of unitprice	Sum of order_id	country
4 PURPLE FLOCK DINNER CANDLES	10	75060	2010	Qtr 4	December	1	536522	72800B	12.75	7392740	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	60	90275	2010	Qtr 4	December	5	537044	72800B	12.75	7413230	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	0	78160	2010	Qtr 4	December	5	749690-9	72800B	41.90	10412730	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	40	75025	2010	Qtr 4	December	10	658386-25	72800B	14.35	10136720	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	5	77320	2011	Qtr 1	January	5	540247	72800B	12.75	7527950	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	0	62365	2011	Qtr 1	January	15	864307-21	72800B	3.25	10074685	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	5	85375	2011	Qtr 1	January	26	542226	72800B	12.75	7606160	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	75	81800	2011	Qtr 1	January	31	655193-99	72800B	5.60	10360960	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	0	86465	2011	Qtr 1	February	14	767552-68	72800B	22.35	10614175	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	10	75295	2011	Qtr 1	February	23	765730-65	72800B	6.50	10929960	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	0	82575	2011	Qtr 1	February	24	700246-55	72800B	3.25	9654075	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	5	89205	2011	Qtr 1	February	28	545186	72800B	12.75	7727225	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	15	86435	2011	Qtr 2	April	1	548642	72800B	12.75	7869325	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	15	88870	2011	Qtr 2	April	4	548808	72800B	12.75	7878685	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	5	90580	2011	Qtr 2	April	18	550459	72800B	12.75	7942420	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	35	73455	2011	Qtr 2	April	19	884408-59	72800B	12.00	10260620	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	45	90345	2011	Qtr 2	May	2	976623-12	72800B	8.10	10420505	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	5	81415	2011	Qtr 2	May	15	553194	72800B	12.75	8050445	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	10	70415	2011	Qtr 2	May	24	554506	72800B	12.75	8101880	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	55	64280	2011	Qtr 2	June	14	651912-94	72800B	8.15	10871740	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	10	77675	2011	Qtr 2	June	19	557315	72800B	12.75	8213835	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	155	68805	2011	Qtr 2	June	21	922181-81	72800B	40.05	10133900	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	35	84755	2011	Qtr 2	June	25	618371-67	72800B	0.75	10244270	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	0	89205	2011	Qtr 2	June	29	991160-58	72800B	44.10	10006920	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	10	79600	2011	Qtr 3	July	8	559507	72800B	12.75	8294070	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	5	79600	2011	Qtr 3	July	8	559509	72800B	12.75	8294215	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	40	65825	2011	Qtr 3	July	17	582573-11	72800B	62.75	10669290	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	50	84070	2011	Qtr 3	July	28	822681-88	72800B	54.20	10991350	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	175	67440	2011	Qtr 3	July	31	758787-9	72800B	14.55	10530750	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	0	86885	2011	Qtr 3	August	6	718582-85	72800B	11.25	9566110	United Kingdom
4 PURPLE FLOCK DINNER CANDLES	65	89205	2011	Qtr 3	August	6	853859-56	72800B	5.20	9858375	United Kingdom
Total	**55389750**	**56451257000**							**14,507,225.55**	**6822025660005**	

Figure 1: Visualizing all Columns of Rental datasets able



Year
□ 2010
□ 2011

Month
□ January
□ Februa...
□ March
□ April
□ May
□ June
□ July
□ August
□ Septe...
□ October
□ Nove...
□ Decem...

Description: All

Country: USA

Customerid: All

Invoiceno: All

order_id: All

Stockcode: All

Quantity: All

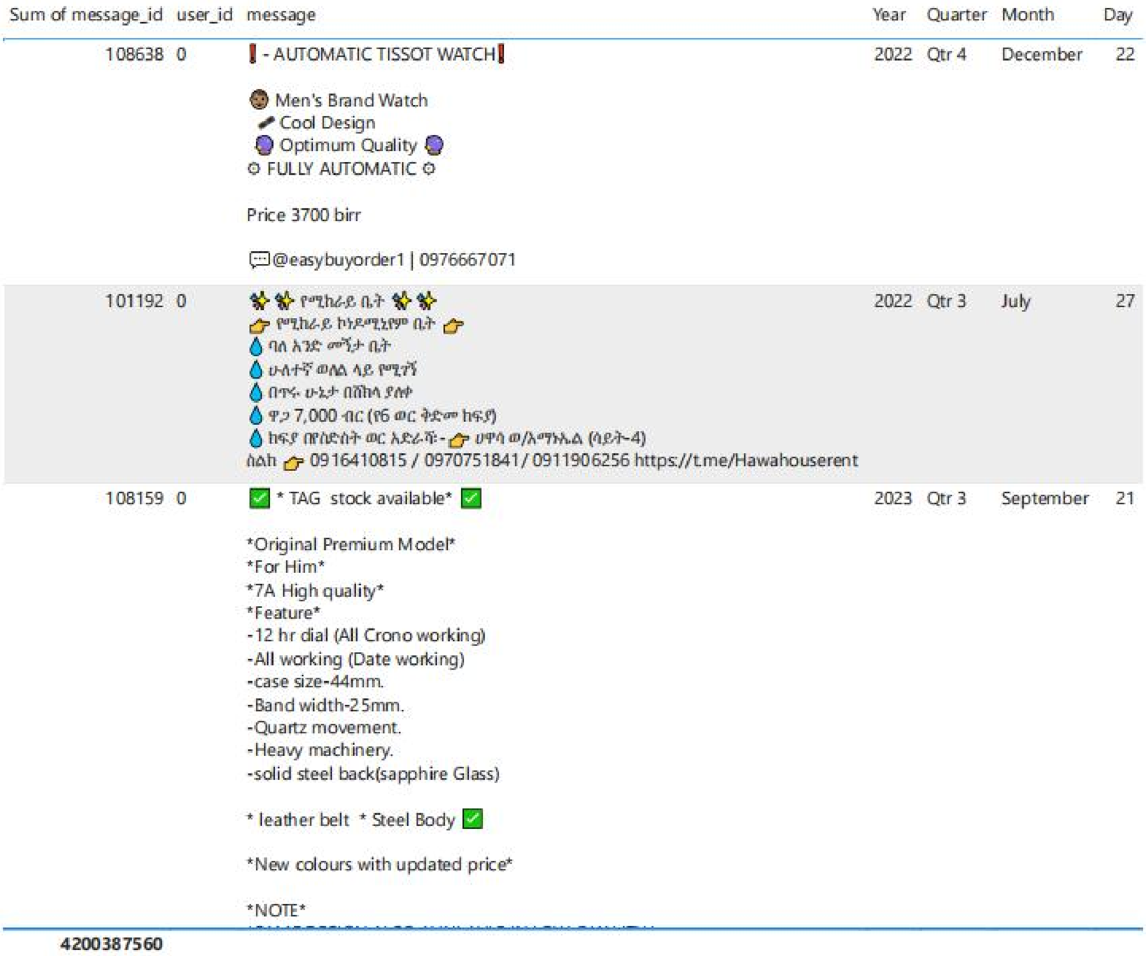
Unitprice: All

28K
Sum of quantity

6.02K
Sum of unitprice

Figure 2: Dashboard for Rental datasets

10. Reports on Telegram Message Dataset Using Power BI



Sum of message_id	user_id	message	Year	Quarter	Month	Day
108638	0	❗ - AUTOMATIC TISSOT WATCH❗ 👨🏾 Men's Brand Watch ✒ Cool Design 🔮 Optimum Quality 🔮 ⚙ FULLY AUTOMATIC ⚙ Price 3700 birr 💬@easybuyorder1 | 0976667071	2022	Qtr 4	December	22
101192	0	✨✨ የሚከራይ ቤት ✨✨ 👉 የሚከራይ ኮንዶሚኒየም ቤት 👉 💧 ባለ አንድ መኝታ ቤት 💧 ሁለተኛ ወለል ላይ የሚገኝ 💧 በጥሩ ሁኔታ በሸክላ ያለቀ 💧 ዋጋ 7,000 ብር (የ6 ወር ቅድመ ክፍያ) 💧 ከፍያ በየስድስት ወር አድራሻ፡- 👉 ሀዋሳ ወ/አማኑኤል (ሳይት-4) ስልክ 👉 0916410815 / 0970751841/ 0911906256 https://t.me/Hawahouserent	2022	Qtr 3	July	27
108159	0	✅ * TAG stock available* ✅ *Original Premium Model* *For Him* *7A High quality* *Feature* -12 hr dial (All Crono working) -All working (Date working) -case size-44mm. -Band width-25mm. -Quartz movement. -Heavy machinery. -solid steel back(sapphire Glass) * leather belt * Steel Body ✅ *New colours with updated price* *NOTE*	2023	Qtr 3	September	21
4200387560						

Figure 3: Visualizing ALL columns of Telegram Message datasets

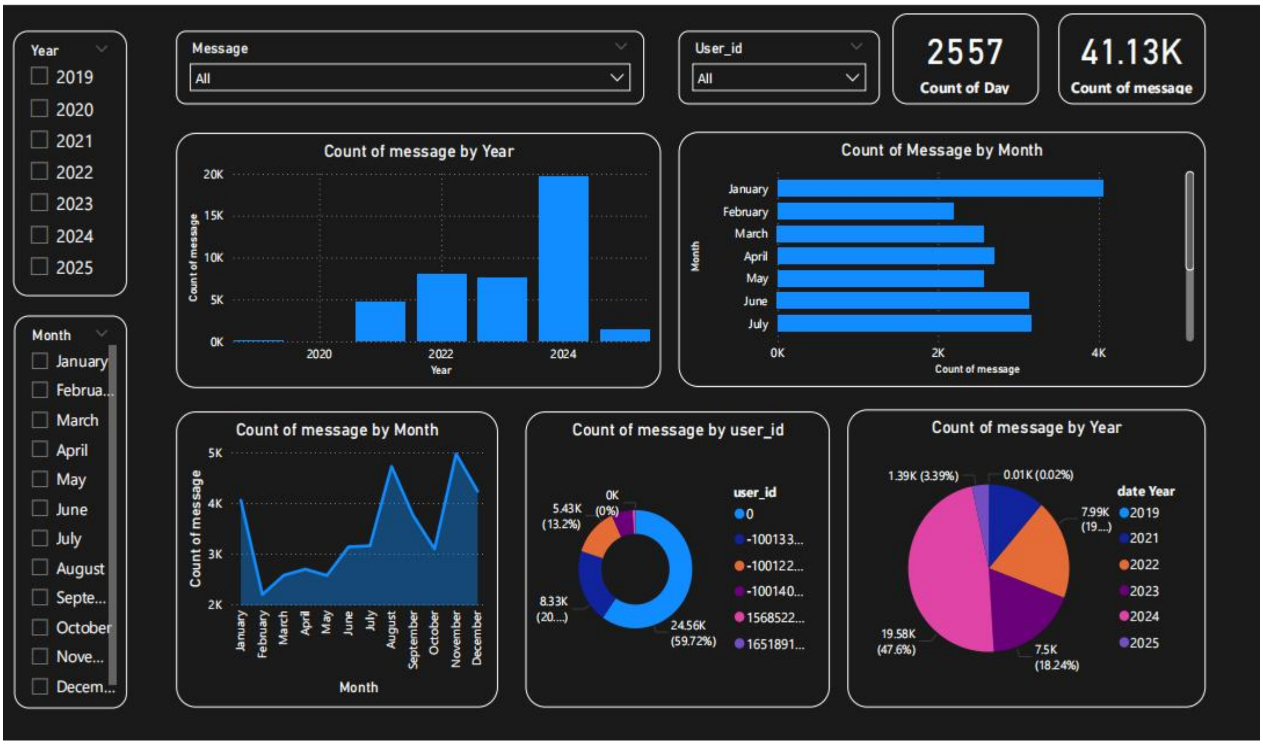


Figure 4: Dashboard for Telegram Message datasets

GitHub Repository Link

https://github.com/dabi938/End-to-end-DP