DEBRE BERHAN UNIVERSITY

INSTITUTION OF TECHNOLOGY

COLLEGE OF COMPUTING

DEPARTMENT OF SOFTWARE ENGINEERING

FUNDAMENTALS OF BIG DATA ANALYTICS AND BUSINESS INTELLIGENCE (SEng5112)

Name: Natnael Yonas
ID: DBU/T/1307331

Submitted to: Mr. Derbew Felasman (MSc)
Date: February 2025, Debre Berhan, Ethiopia

Table of Content

1. Overview

This document outlines the development of a robust Extract, Transform, and Load (ETL) pipeline designed to process an Amazon e-commerce dataset. The pipeline begins by extracting raw data from CSV files, followed by a cleaning and transformation process to ensure consistency and structure. The transformed data is then loaded into a PostgreSQL database, enabling efficient querying and storage. Key insights, including sales trends, product performance, and customer engagement, are visualized using Power BI. The ultimate objective is to provide data-driven insights to optimize business strategies, enhance customer experience, and improve decision-making processes.

2. Dataset Description

The dataset used in this ETL pipeline contains **Amazon product data**, including product details, pricing, customer ratings, and sales trends. The dataset consists of the following key columns:

- ❖ **asin**: A unique identifier assigned to each product.
- ❖ **title**: The name of the product.
- ❖ **imgUrl**: The URL linking to the product's image.
- ❖ **productURL**: The direct link to the product's page on Amazon.
- ❖ **stars**: The average customer rating of the product.
- ❖ **reviews**: The total number of reviews the product has received.
- ❖ **price**: The current price of the product.
- ❖ **listPrice**: The original listed price of the product before discounts.
- ❖ **categoryName**: The category or department under which the product falls.
- ❖ **isBestSeller**: A Boolean value indicating whether the product is a bestseller.
- ❖ **boughtInLastMonth**: The number of units sold within the last month.

3. Tools and Technologies Used

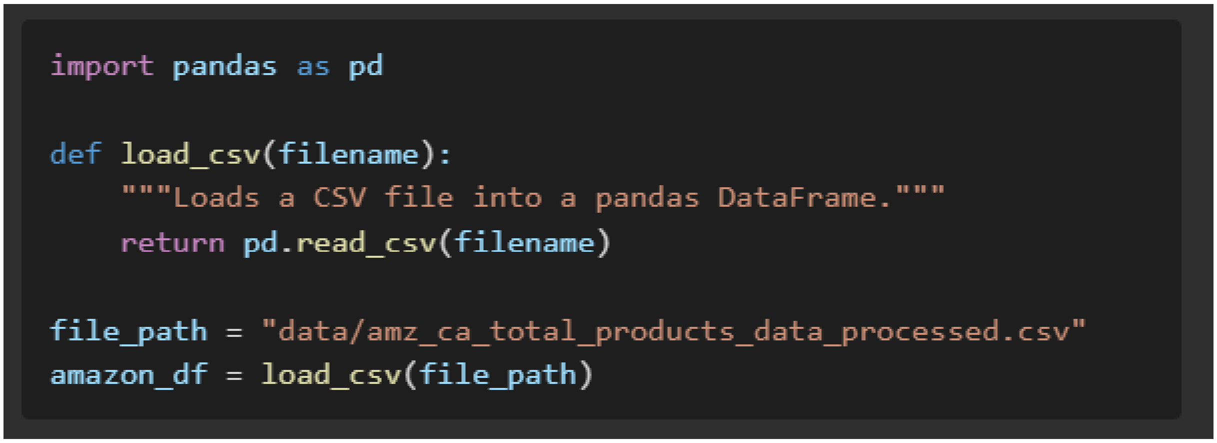
- ➢ **Python:** Utilized for scripting and automating the ETL (Extract, Transform, Load) process, handling data manipulation, and integrating with other tools.
- ➢ **Pandas:** Used for efficient data extraction, transformation, and cleaning, including operations like handling missing values, data type conversion, and filtering.
- ➢ **Matplotlib** : Employed for creating static, animated, and interactive visualizations, allowing detailed exploration of data trends, distributions, and relationships.

- **psycopg2:** A PostgreSQL adapter for Python, used to connect to and interact with the PostgreSQL database, enabling data insertion, updates, and queries.
- **PostgreSQL:** A powerful relational database management system used to securely store processed data, with robust support for complex queries and data integrity.
- **Power BI:** Utilized for advanced data visualization and generating actionable insights through interactive dashboards, allowing decision-makers to explore data and trends efficiently.

4. ETL Pipeline Steps

4.1 Extraction

The extraction phase involves loading raw data from a **CSV file** into a **Pandas DataFrame**. Below is the Python code snippet for extracting the data:

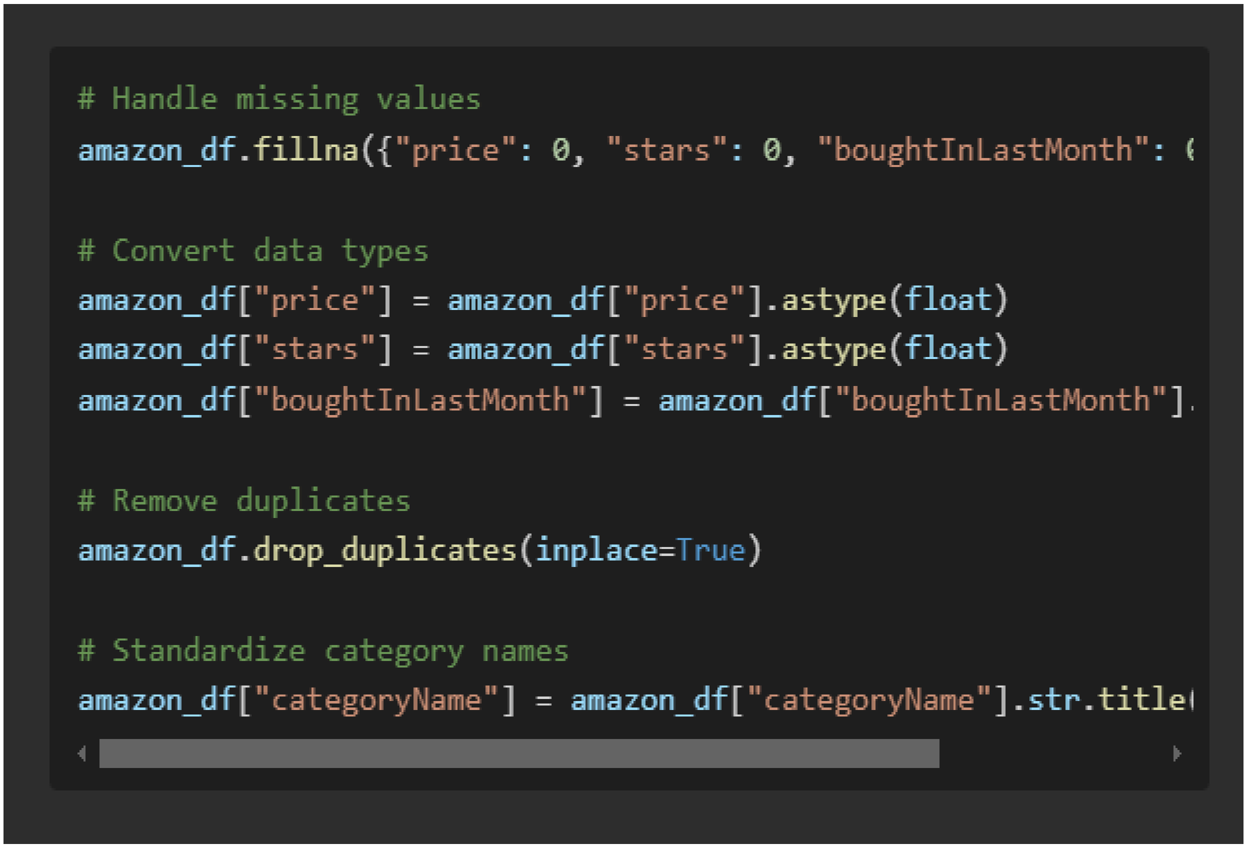


```python
import pandas as pd

def load_csv(filename):
    """Loads a CSV file into a pandas DataFrame."""
    return pd.read_csv(filename)

file_path = "data/amz_ca_total_products_data_processed.csv"
amazon_df = load_csv(file_path)
```

4.2 Transformation

The transformation phase ensures data consistency and integrity by performing the following steps:

- **Handling missing values**: Missing or NaN values are replaced with meaningful default values, ensuring that incomplete data does not hinder analysis or processing.
- **Ensuring correct data types**: Fields are converted to appropriate data types, such as transforming prices and ratings into float values for precision, and converting sales into integer values to maintain numerical integrity.
- **Removing duplicates**: Duplicate entries are identified and removed to prevent redundancy, ensuring that the dataset remains unique and accurate.
- **Standardizing text fields**: Text fields like category names are standardized to title case, ensuring consistent formatting and improving readability.



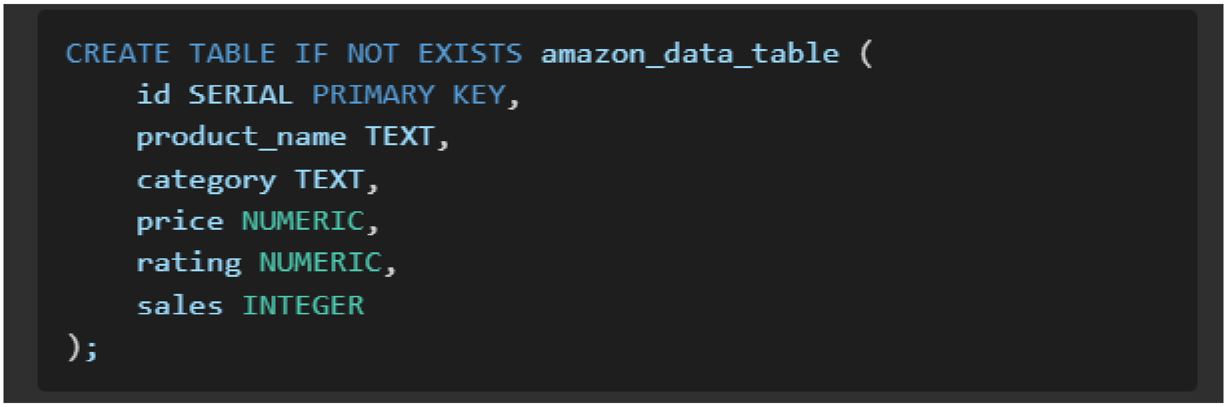
```python
# Handle missing values
amazon_df.fillna({"price": 0, "stars": 0, "boughtInLastMonth": 

# Convert data types
amazon_df["price"] = amazon_df["price"].astype(float)
amazon_df["stars"] = amazon_df["stars"].astype(float)
amazon_df["boughtInLastMonth"] = amazon_df["boughtInLastMonth"].

# Remove duplicates
amazon_df.drop_duplicates(inplace=True)

# Standardize category names
amazon_df["categoryName"] = amazon_df["categoryName"].str.title(
```

4.3 Loading

The transformed data is subsequently stored in a PostgreSQL database for efficient and structured data management.

PostgreSQL Table Creation:

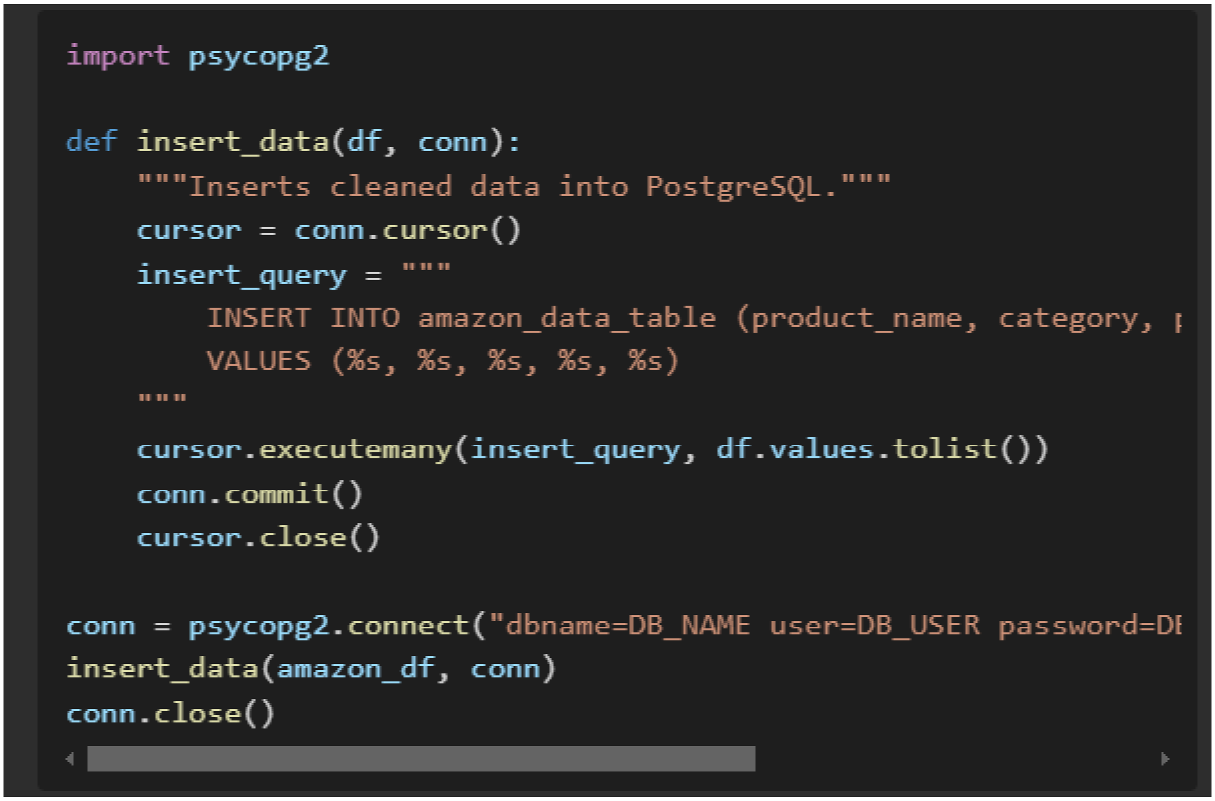
To create a table in PostgreSQL, the CREATE TABLE statement is used, followed by defining the table name and its columns, data types, and constraints.



```sql
CREATE TABLE IF NOT EXISTS amazon_data_table (
    id SERIAL PRIMARY KEY,
    product_name TEXT,
    category TEXT,
    price NUMERIC,
    rating NUMERIC,
    sales INTEGER
);
```

Python Script to Insert Data:

To insert data into a PostgreSQL table, you can use the psycopg2 library, which is a PostgreSQL adapter for Python.



```
import psycopg2

def insert_data(df, conn):
    """Inserts cleaned data into PostgreSQL."""
    cursor = conn.cursor()
    insert_query = """
        INSERT INTO amazon_data_table (product_name, category,
        VALUES (%s, %s, %s, %s, %s)
    """
    cursor.executemany(insert_query, df.values.tolist())
    conn.commit()
    cursor.close()

conn = psycopg2.connect("dbname=DB_NAME user=DB_USER password=DB
insert_data(amazon_df, conn)
conn.close()
```

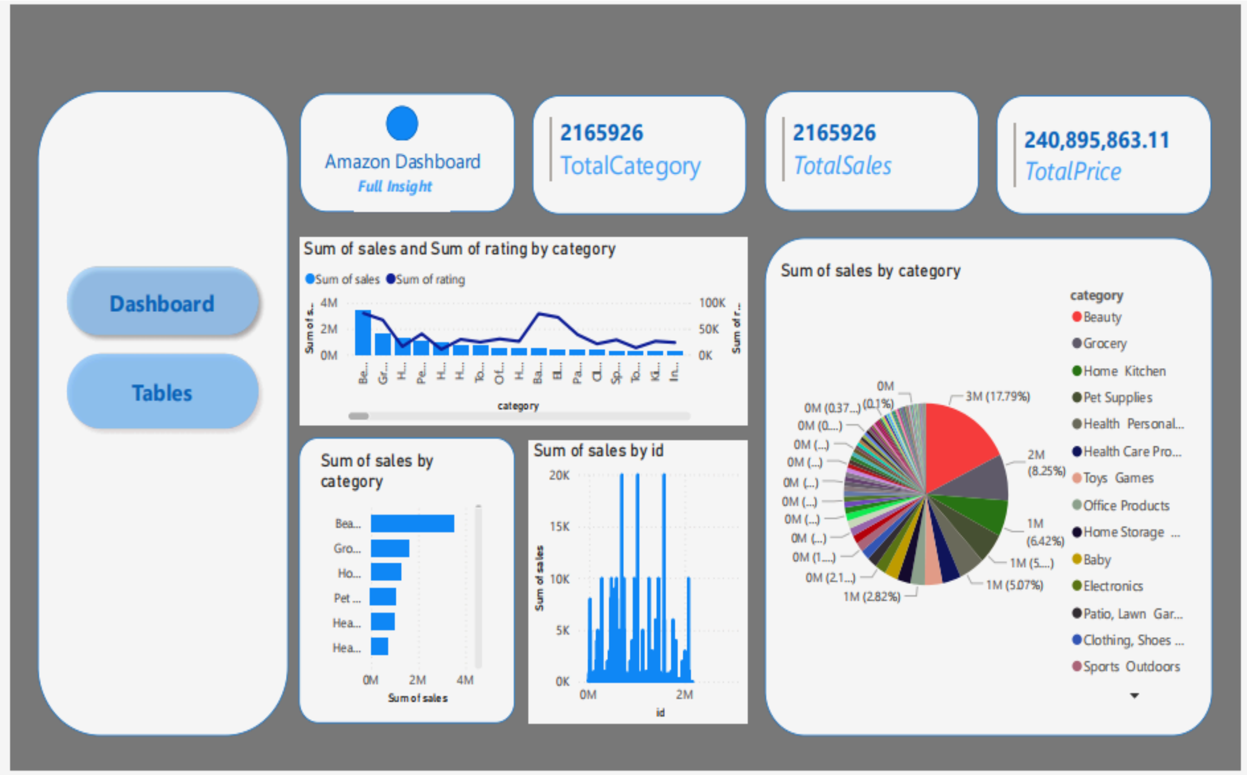
5. Data Cleaning Processes

To ensure accuracy and reliability, the following steps were applied:

1. **Handling missing values** by replacing NaNs with meaningful defaults.
2. **Data type conversion** to ensure numerical columns are properly formatted.
3. **Removing duplicates** to maintain data integrity.
4. **Ensuring category names are properly formatted** for consistency.

6. Visualization and Insights

Dashboard:



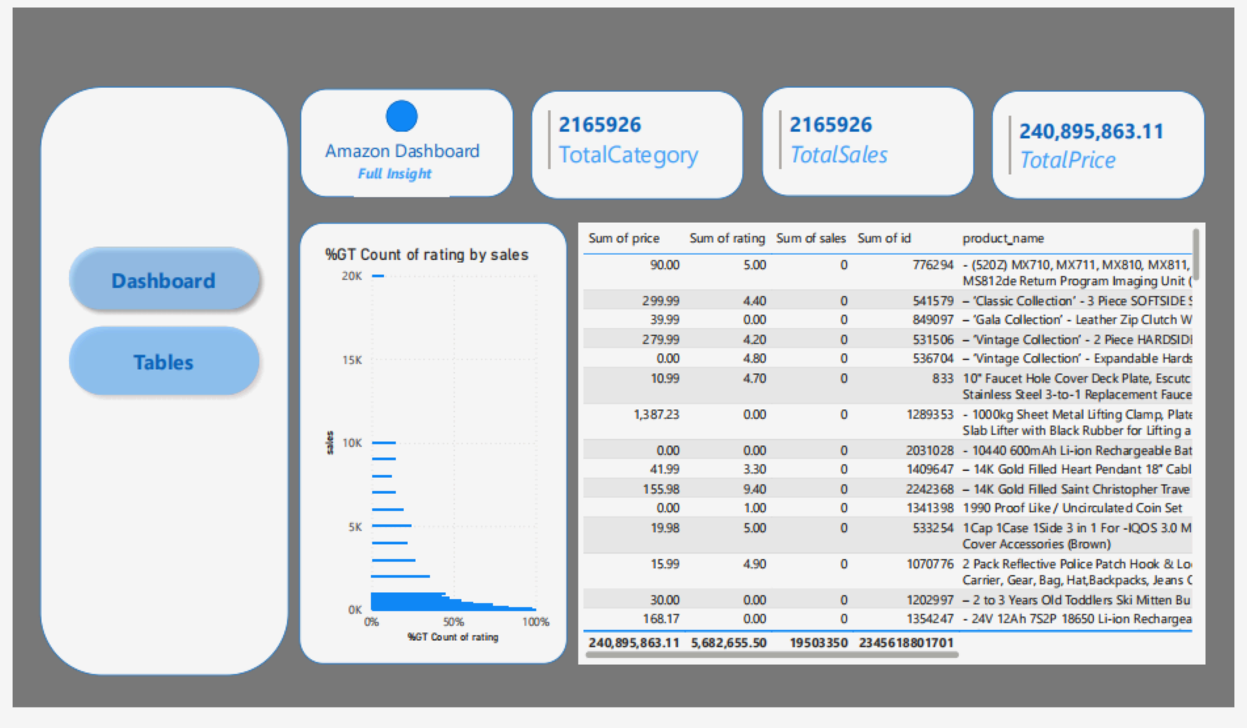
Amazon Dashboard
Full Insight

2165926
TotalCategory

2165926
TotalSales

240,895,863.11
TotalPrice

Table:



Amazon Dashboard
Full Insight

2165926
TotalCategory

2165926
TotalSales

240,895,863.11
TotalPrice

Sum of price	Sum of rating	Sum of sales	Sum of id	product_name
90.00	5.00	0	776294	- (520Z) MX710, MX711, MX810, MX811, MS812de Return Program Imaging Unit (
299.99	4.40	0	541579	– 'Classic Collection' - 3 Piece SOFTSIDE S
39.99	0.00	0	849097	– 'Gala Collection' - Leather Zip Clutch W
279.99	4.20	0	531506	– 'Vintage Collection' - 2 Piece HARDSIDI
0.00	4.80	0	536704	– 'Vintage Collection' - Expandable Hards
10.99	4.70	0	833	10" Faucet Hole Cover Deck Plate, Escutc Stainless Steel 3-to-1 Replacement Fauce
1,387.23	0.00	0	1289353	- 1000kg Sheet Metal Lifting Clamp, Plate Slab Lifter with Black Rubber for Lifting a
0.00	0.00	0	2031028	- 10440 600mAh Li-ion Rechargeable Bat
41.99	3.30	0	1409647	– 14K Gold Filled Heart Pendant 18" Cabl
155.98	9.40	0	2242368	– 14K Gold Filled Saint Christopher Trave
0.00	1.00	0	1341398	1990 Proof Like / Uncirculated Coin Set
19.98	5.00	0	533254	1Cap 1Case 1Side 3 in 1 For -IQOS 3.0 M Cover Accessories (Brown)
15.99	4.90	0	1070776	2 Pack Reflective Police Patch Hook & Lo Carrier, Gear, Bag, Hat,Backpacks, Jeans C
30.00	0.00	0	1202997	– 2 to 3 Years Old Toddlers Ski Mitten Bu
168.17	0.00	0	1354247	- 24V 12Ah 7S2P 18650 Li-ion Rechargea
240,895,863.11	**5,682,655.50**	**19503350**	**2345618801701**	

6.1 Sales Trends Over Time

- **Visualization**: Line Chart
- **Insight**: Sales exhibited significant fluctuations, suggesting potential seasonal patterns or cyclical demand. Identifying these trends can help businesses optimize inventory management and promotional strategies during peak periods.

6.2 Product Performance Analysis

- **Visualization**: Bar Chart
- **Insight**: Several products showed high view counts but low conversion rates, implying potential issues with pricing, product descriptions, or customer perception. This highlights the need for optimization in these areas to improve sales.

6.3 Customer Engagement Analysis

- **Visualization**: Pie Chart (Rating Distribution)
- **Insight**: A large proportion of products received ratings above 4.0, reflecting high levels of customer satisfaction. This suggests that product quality and customer experiences are strong, potentially driving repeat business and positive reviews.

7. Conclusion

This project successfully implemented a comprehensive end-to-end ETL pipeline for an Amazon e-commerce dataset, encompassing data extraction, transformation, and loading. The data was extracted from raw sources, cleansed and transformed to ensure consistency, and loaded into a PostgreSQL database for efficient storage and retrieval. Advanced data visualizations were created using Power BI to derive actionable insights, enabling businesses to optimize pricing strategies, identify high-demand products, and enhance customer engagement. This pipeline ensures the seamless flow of data from raw input to actionable business intelligence, providing a solid foundation for data-driven decision-making.

8. References

- **Python Documentation**: https://www.python.org/
- **PostgreSQL Documentation**: https://www.postgresql.org/
- **Microsoft Power BI Documentation**: https://powerbi.microsoft.com/