Prepared by: seyfedin shukur

ID: dbu/046/14



**1. Data Source Identification & Understanding**

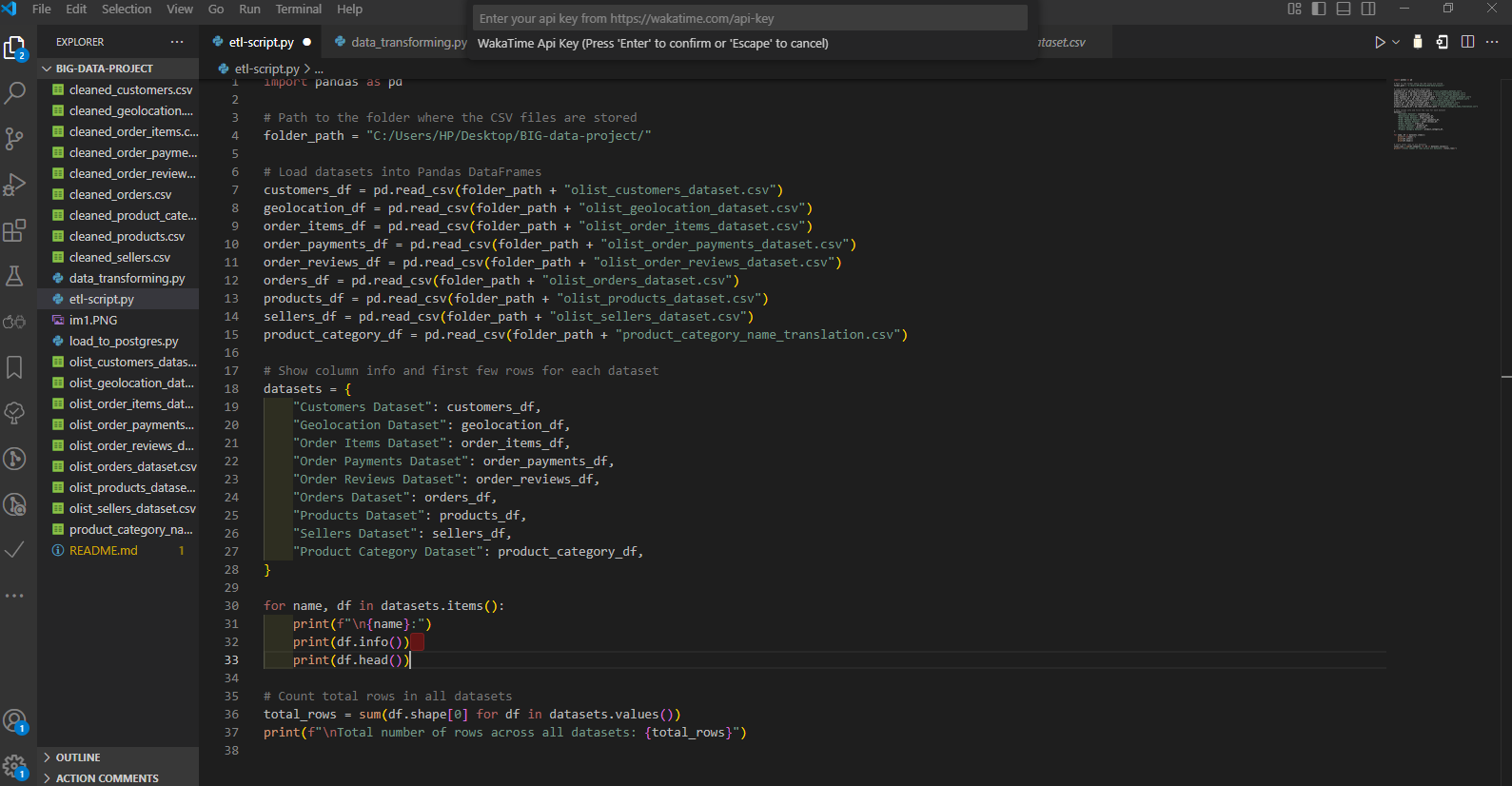
**Objective**: The goal of this step is to identify and select a suitable dataset for analysis and transformation, focusing on e-commerce-related data. The dataset selected for this project contains over 1.5 million rows and is used to gain insights into customer behavior, product trends, seller performance, and reviews. It will serve as the basis for various data analysis tasks such as cleaning, transformation, and visualization.

### ****1.1 Dataset Overview****

* **Dataset Name:** Brazilian E-commerce Data (Olist)
* **Source:** Kaggle - Brazilian E-commerce Data
* **Dataset Size**: Contains over 1.5 million rows across multiple CSV files.
* **Data Type:** Structured data in CSV format.
* **Data Fields:** The dataset includes multiple tables with different features such as:
* Orders data
* Product details
* Customer information
* Reviews and ratings
* Seller details
* Payment information

**Task:** Analyze the dataset structure by reading a CSV and examining its schema using pandas.

**Code to Load Data and Read**



**1.2 Description of the Dataset**

This dataset is a comprehensive record of transactions from a Brazilian e-commerce platform, Olist. It includes detailed information about customer orders, products, payment methods, sellers, and reviews. The data can be used to analyze various aspects of e-commerce performance, including product sales, customer feedback, seller activity, and order processing times.

**Key CSV Files:**

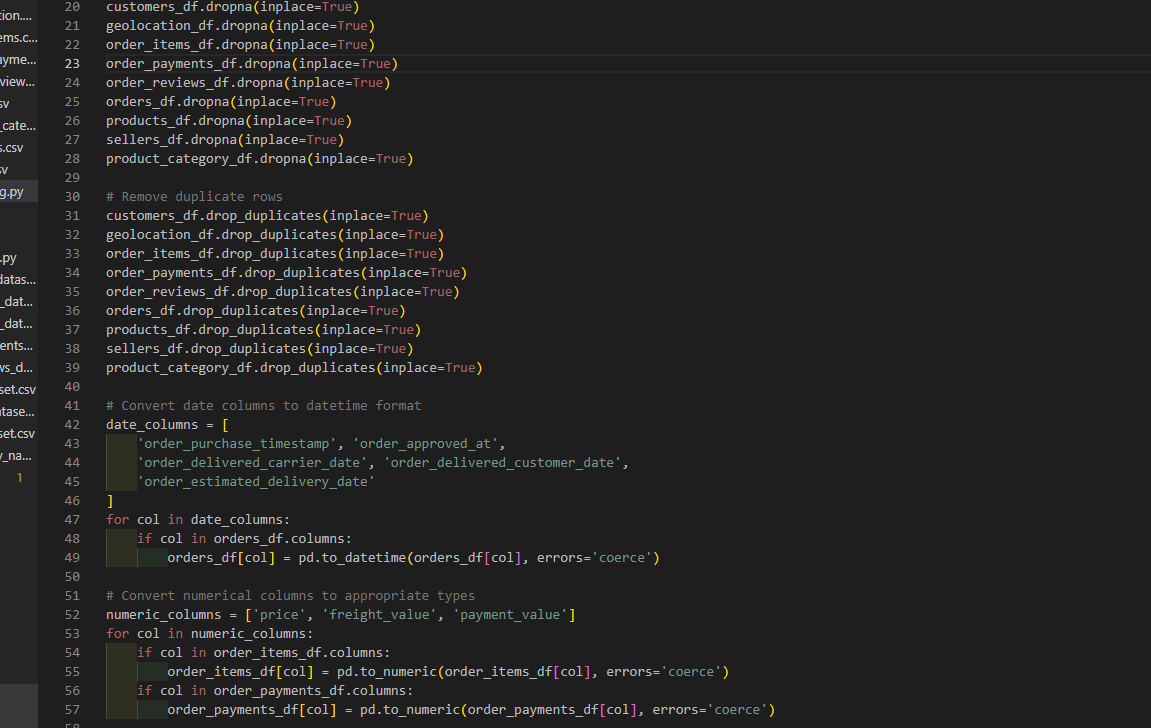
* **orders.csv:** Contains details about customer orders such as order IDs, payment values, and timestamps.
* **order\_items.csv:** Includes data on the individual items ordered, linked to the orders.csv file.
* **products.csv:** Contains product-related information such as product categories, descriptions, and pricing.
* **customers.csv:** Includes customer details such as their city, state, and unique customer ID.
* **reviews.csv:** Contains product reviews, ratings, and customer feedback.
* **sellers.csv**: Includes seller information such as seller ID, location, and sales.
* **product\_category\_name\_translation.csv:** Maps product categories to English translations.
* **payment.csv:** Contains payment details related to orders, including payment method and amount.

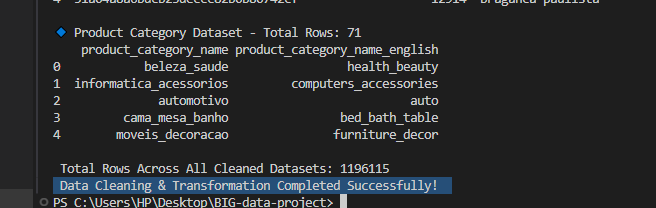
3. Data Transformation

**3.1 Cleaning:**

* Handle Missing Values: Apply techniques to handle missing data in the e-commerce dataset.
* Remove Duplicates: Identify and remove duplicates.
* Format Data Types: Convert dates to datetime and numerical data to appropriate types.
* Handle Inconsistencies and Errors: Standardize product names, categories, or payment methods.

**Code for Data Cleaning**:

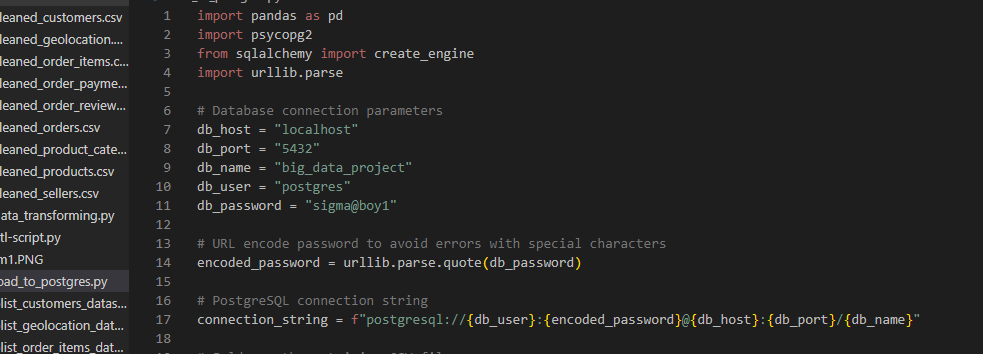




1. **Loading Processed Data into PostgreSQL**

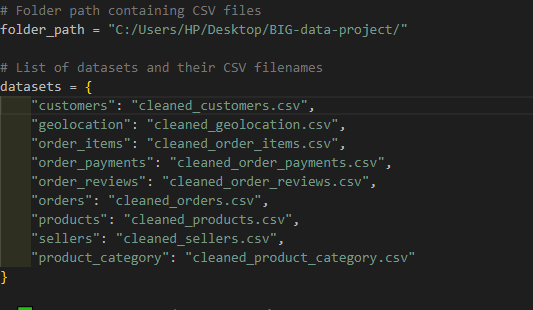
**Prerequisites**

* Ensure PostgreSQL is installed and running. You also need Pandas and psycopg2 installed in Python:
* pip install pandas psycopg2
* Data Loading Using Python (Pandas & psycopg2)
* A.Define Database Connection Parameters



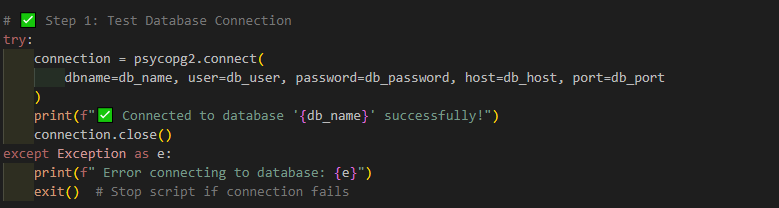
1. **Define Dataset Paths**

Specify the folder path containing the CSV files and list of datasets:



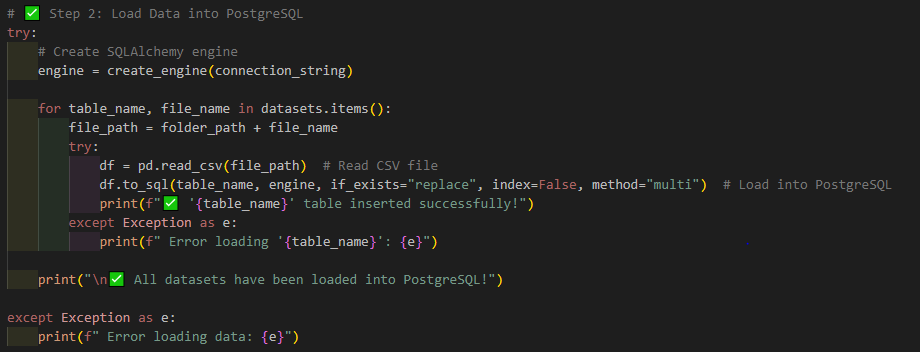
1. Test Database Connection

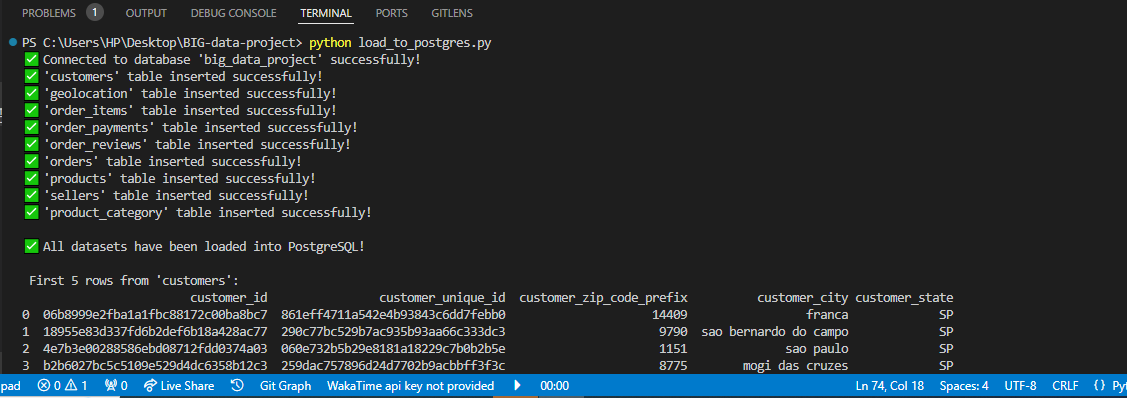
Verify the connection to PostgreSQL:



### ****Load CSV Data into PostgreSQL****

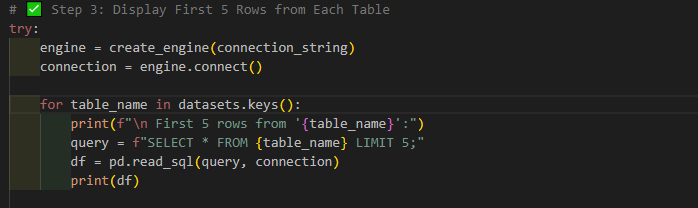
This script reads each dataset and loads it into PostgreSQL:

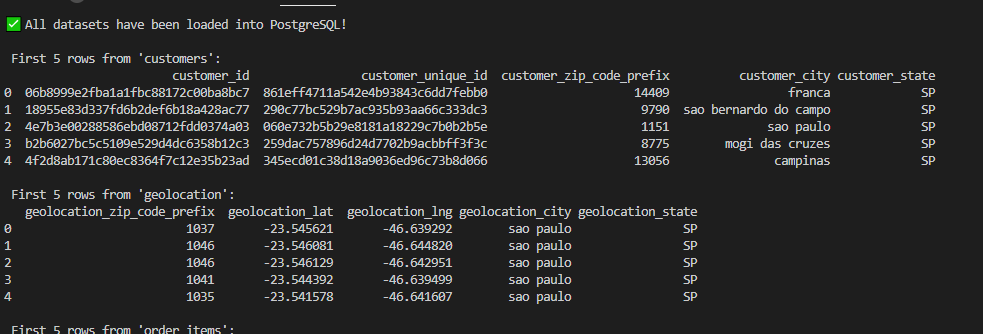




### ****Verify Data Load by Displaying First 5 Rows****

Check if the data was inserted correctly:





## ****Data Visualization and Insights****

### ****1. Connecting Power BI to PostgreSQL****

* Open Power BI Desktop.
* Click Home > Get Data > PostgreSQL Database.
* Enter server details (localhost).
* Click Import to load data.
* Select tables and click Load.

### ****Interactive Dashboards****

### ****Order ID and Customer City in a Bar Chart****

* The bar chart visualizes the number of orders per customer city, providing insights into regional sales distribution. The X-axis represents different customer cities, while the Y-axis shows the count of order IDs.
* This visualization helps identify high-demand regions, customer concentration, and potential market expansion opportunities. It can also reveal geographic patterns in purchasing behavior, aiding in targeted marketing and logistics optimization.

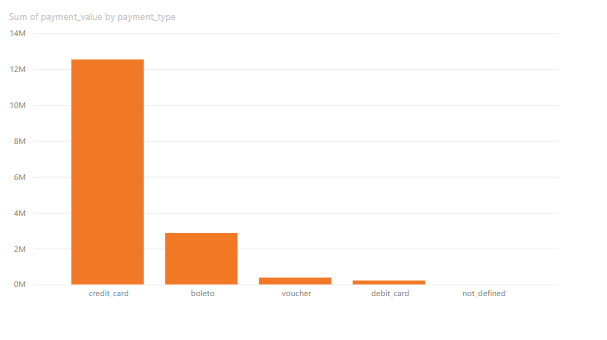


### ****B.Payment Value by Payment Type in a Column Chart****

The column chart visualizes the total payment value for each payment type, helping to understand customer preferences in transaction methods.

* **X-axis:** Different payment types (e.g., credit card, debit card, boleto, voucher).
* **Y-axis:** Sum of payment values representing total revenue from each method.

This visualization helps businesses identify the most popular payment methods, optimize payment processing strategies, and assess financial trends related to customer transactions.

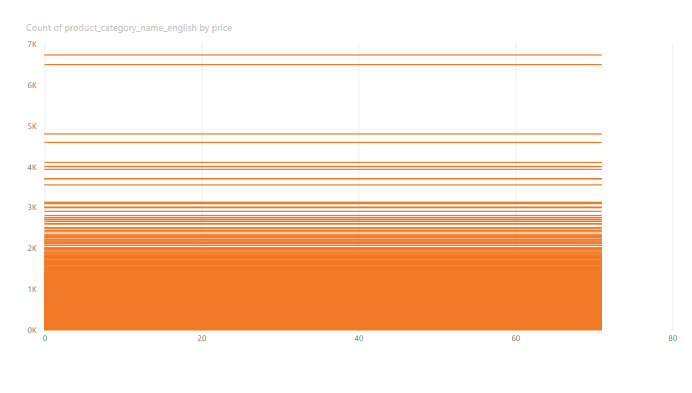


### ****C.Product Category Name (English) by Price****

This visualization displays the average product price across different product categories to analyze pricing trends.

* **X-axis:** Product Category Name (English) – Represents different product categories.
* **Y-axis**: Average Price – Shows the mean price of products within each category.

This analysis helps businesses understand price variations across categories, identify high-value product segments, and make data-driven pricing decisions.

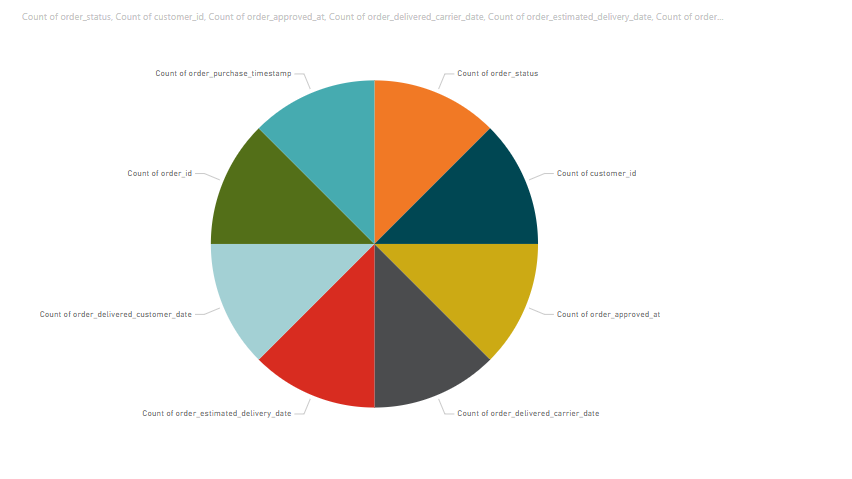


### ****Orders Distribution (Pie Chart)****

This pie chart visualizes the distribution of orders across different categories or segments. It helps in identifying the proportion of total orders contributed by various factors such as:

* **Customer City** – Orders grouped by customer locations.
* **Payment Type** – Orders categorized by payment methods used.
* **Order Status** – Distribution of completed, pending, and canceled orders.

This chart provides a quick insight into order trends, helping businesses optimize logistics, marketing strategies, and customer engagement

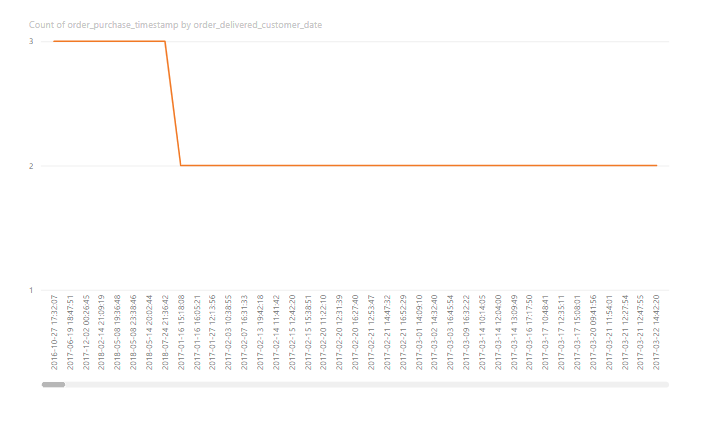


### ****E.Order Purchase Timestamp vs. Order Delivered Customer Date (Bar Chart)****

This bar chart compares the order purchase timestamp with the order delivery date to analyze delivery efficiency and processing times.

**Key Insights:**

* **Average Delivery Time:** Identifies the time taken from purchase to delivery.
* **Delays & Trends:** Highlights any seasonal trends or delays in order fulfillment.
* **Customer Experience**: Helps assess whether delivery times meet customer expectations.

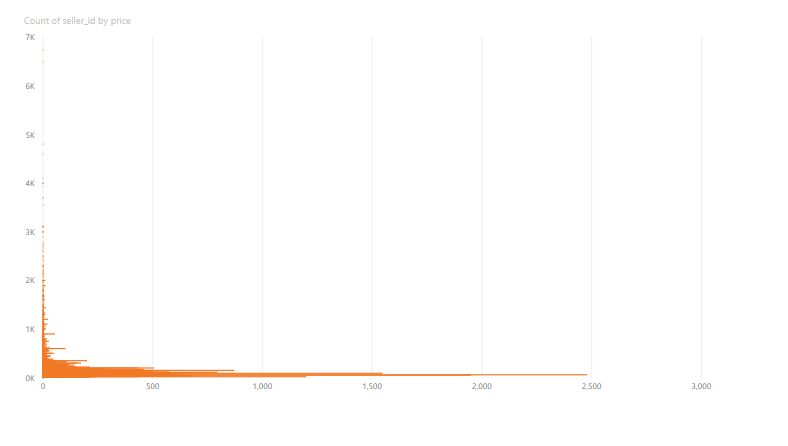


### ****F.Seller ID by Price (Bar Chart)****

This bar chart visualizes the total sales price associated with each seller ID, helping to identify top-performing sellers.

Key Insights:

* **Top Sellers**: Highlights which sellers generate the highest revenue.
* **Sales Distribution:** Shows variations in sales performance among sellers.
* **Business Strategy:** Helps optimize partnerships with high-revenue sellers.



## ****Tools & Technologies****

* **PostgreSQL:** Relational database for data storage
* **Pandas:** Data processing library for handling CSV files
* **SQLAlchemy:** Database toolkit for connecting and interacting with PostgreSQL
* **Power BI:** Business intelligence tool for visualization
* **VS Code:** Code editor for development