

**College of Computing**

**Department of Software Engineering**

**Big Data individual assignment**

**Course Code: SENG5112**

**Course Title: Fundamental of Big Data Analytics and Business Intelligence**

**Done by: Beimnet Wesley**

**Submitted to: Mr.Abebe**

Table of Contents

[**Introduction** 3](#_Toc190332800)

[**1. Data Extraction:** 3](#_Toc190332801)

[**2. Data Transformation:** 3](#_Toc190332802)

[**2.1 Handling Missing Values:** 3](#_Toc190332803)

[**2.2** **Removing Duplicates:** 3](#_Toc190332804)

[**2.3** **Replacing Dots with Spaces in Specific Columns:** 4](#_Toc190332805)

[**2.4** **Converting Timestamp Format:** 4](#_Toc190332806)

[3. Data Loading: 4](#_Toc190332807)

[**3.2 Data Insertion into DuckDB:** 5](#_Toc190332808)

[**4. Data Export:** 5](#_Toc190332809)

[**4.1 Export Data to Parquet:** 5](#_Toc190332810)

[**5. Data Analysis and Visualization with Plotly:** 5](#_Toc190332811)

[**5.1 Sales by Brand:** 5](#_Toc190332812)

[**5.2 Price vs. Sales by Category:** 6](#_Toc190332813)

[**5.3** **Product Category Share:** 7](#_Toc190332814)

[**5.4 Customer Segmentation by Spending:** 8](#_Toc190332815)

[**5.5** **Sales Trend Over Time:** 9](#_Toc190332816)

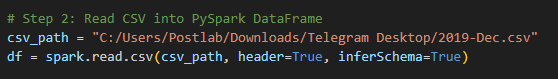
[**6. Power BI Integration:** 10](#_Toc190332817)

# **Introduction**

This documentation provides a comprehensive overview of the ETL pipeline used to transform, load, and analyze e-commerce data. The pipeline extracts raw data from a CSV file, processes it using Spark for transformation, and stores the results in a DuckDB database. Finally, it visualizes the data using Plotly and integrates it into Power BI for analysis.

# **1. Data Extraction:**

The raw e-commerce data is loaded from a CSV file into a PySpark DataFrame. This step initiates the extraction of information, including product details, user behavior, and transaction data.

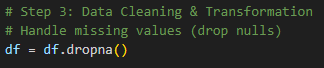


* **Path to CSV**: The location of the CSV file is specified (csv\_path).
* **Data Loading**: The spark.read.csv() function is used to read the CSV file into a DataFrame with inferred data types and headers.

# **2. Data Transformation:**

## **2.1 Handling Missing Values:**

To ensure data quality, rows with missing values are removed from the DataFrame.



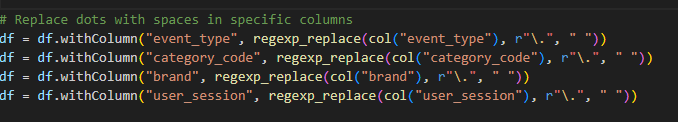
## **2.2** **Removing Duplicates:**

Any duplicate rows are removed to avoid redundant data processing and storage.



## **2.3** **Replacing Dots with Spaces in Specific Columns:**

To clean the data, dots in certain columns are replaced with spaces. This is done in columns like event\_type, category\_code, brand, and user\_session to standardize the text fields.



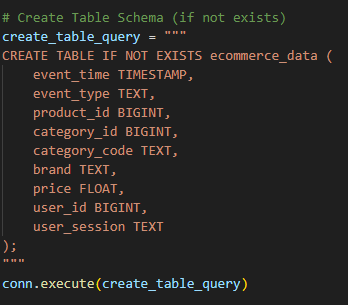
## **2.4** **Converting Timestamp Format:**

The **event\_time** column is converted from string format to a timestamp format, which allows for better time-series analysis and processing.



### ****3. Data Loading:****

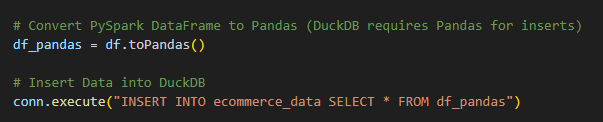
#### 3.1 ****Loading Data into DuckDB:****

The cleaned and transformed data is loaded into a DuckDB database for further analysis. 

* **Schema Design**: The schema design ensures that all relevant columns (e.g., product ID, category, price, user information) are properly defined.
* **Table Creation**: A query is executed to create a table if it doesn't already exist in the database.

## **3.2 Data Insertion into DuckDB:**

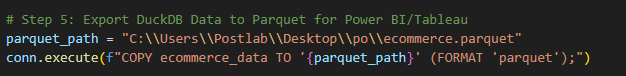
The transformed data (in Pandas format) is inserted into the DuckDB table for persistent storage.



# **4. Data Export:**

## **4.1 Export Data to Parquet:**

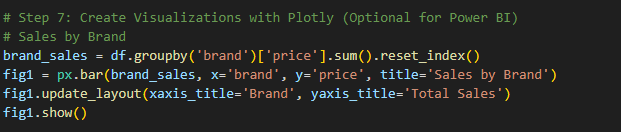
The data is exported to a Parquet file for further use in BI tools like Power BI and Tableau.

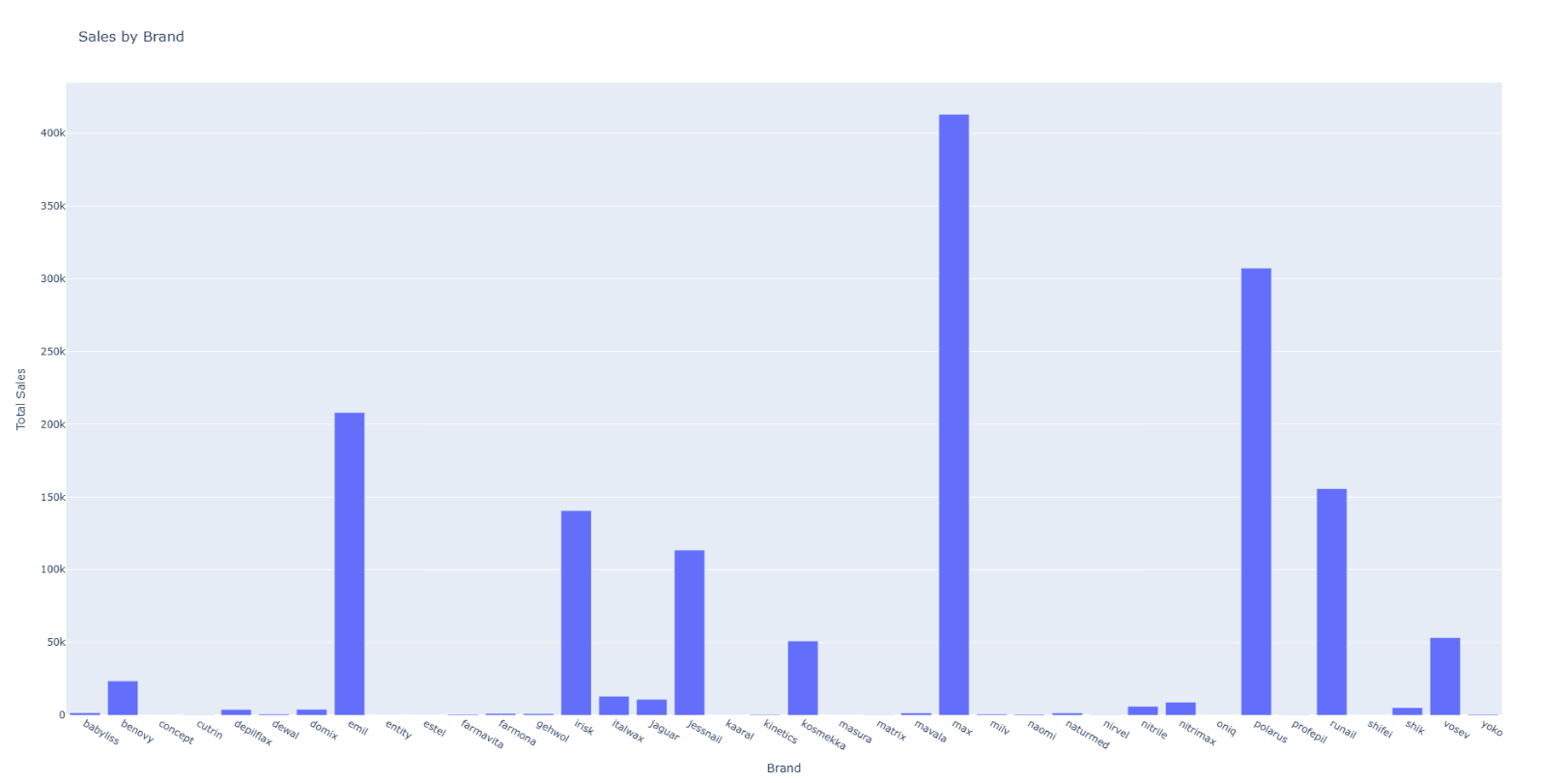


# **5. Data Analysis and Visualization with Plotly:**

## **5.1 Sales by Brand:**

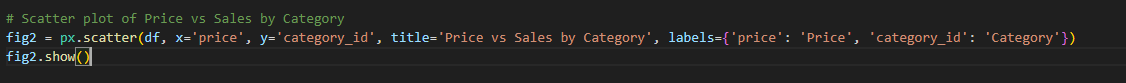
A bar chart is created to visualize the total sales by brand. This helps to identify top-performing brands.

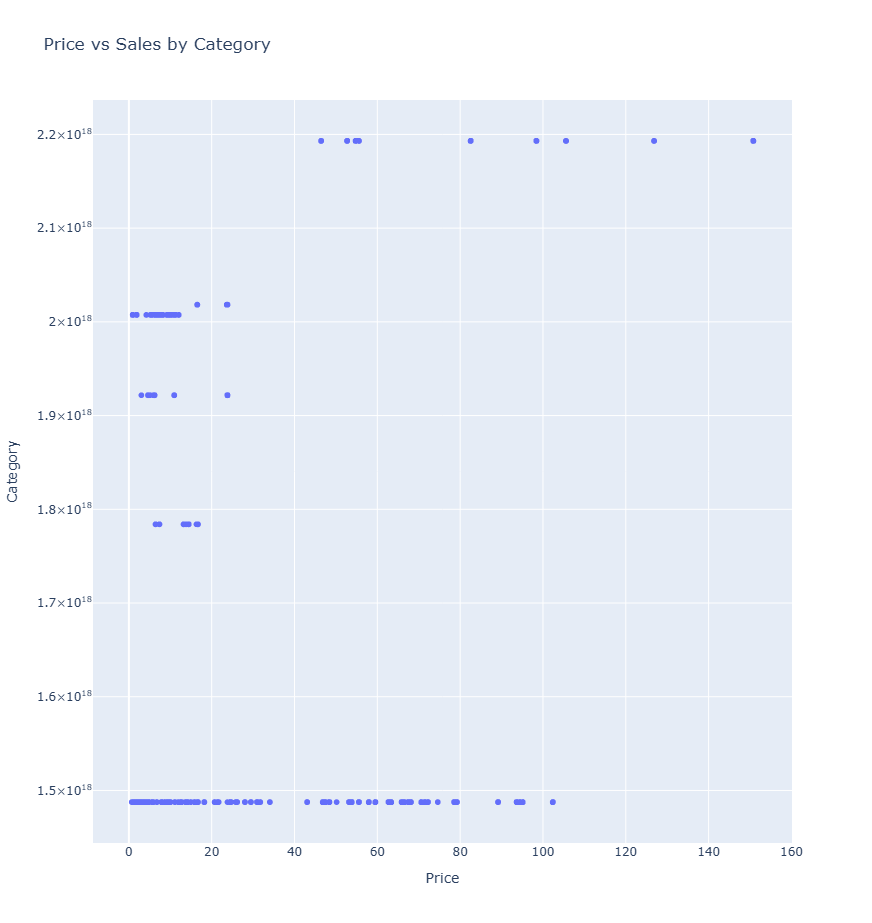




## **5.2 Price vs. Sales by Category:**

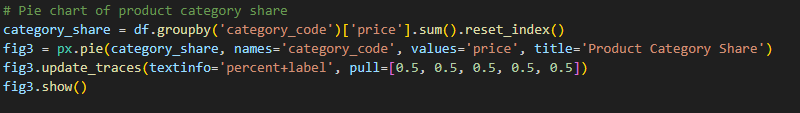
A scatter plot shows the relationship between price and sales for each category.

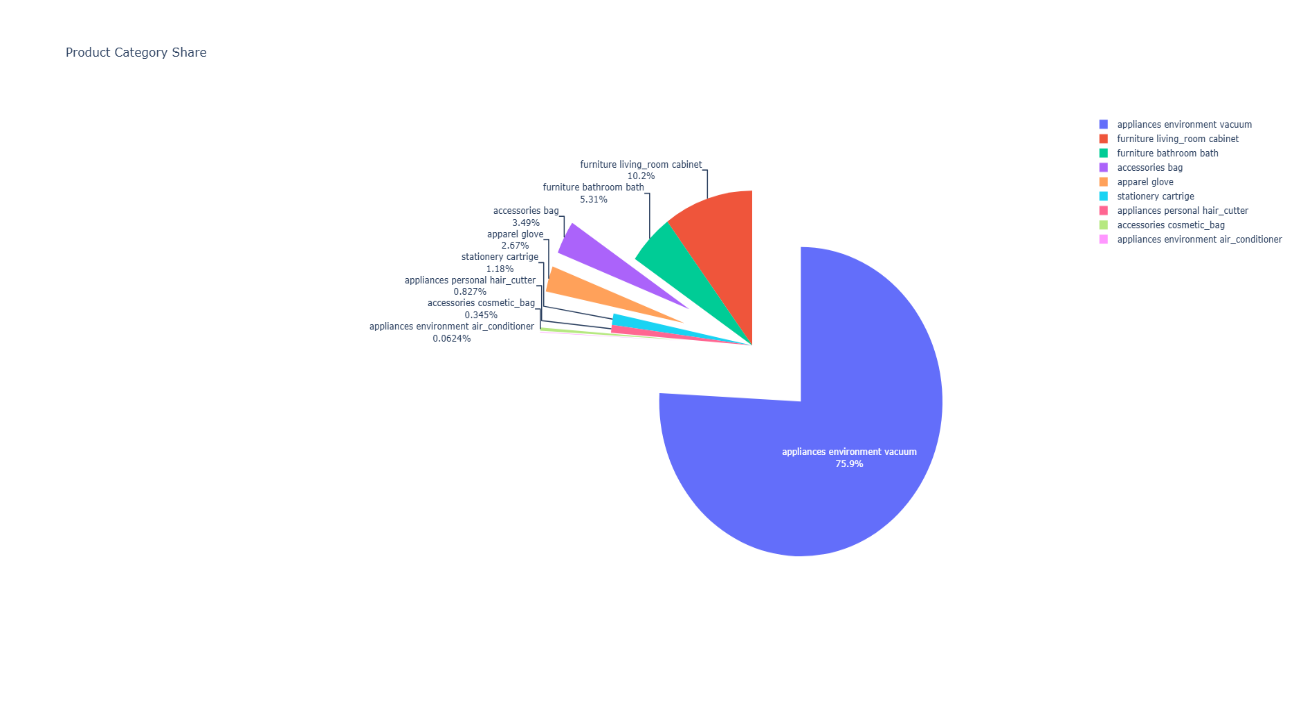




## **5.3** **Product Category Share:**

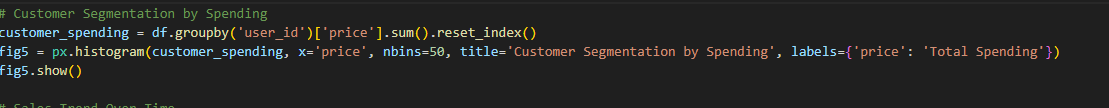
A pie chart is used to represent the share of total sales by product category.

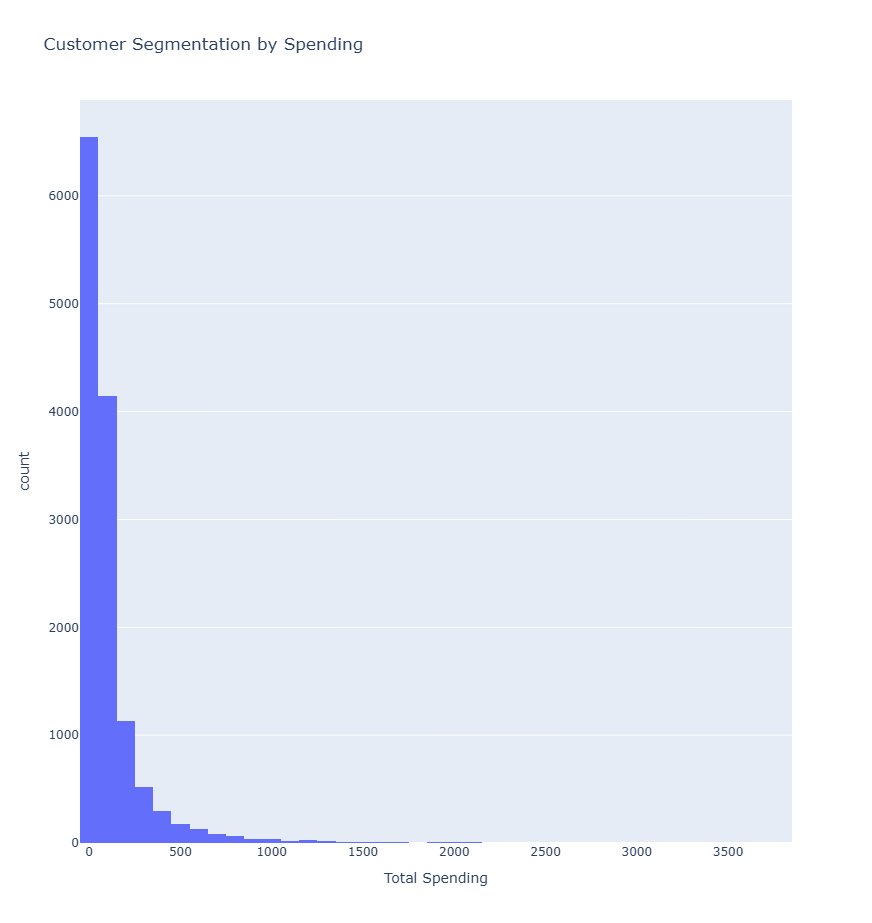




## **5.4 Customer Segmentation by Spending:**

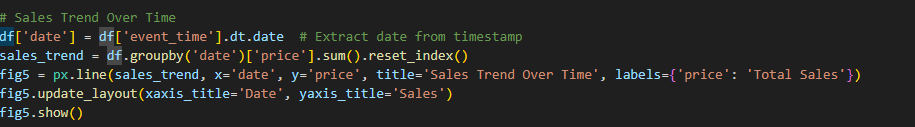
A histogram visualizes the distribution of customer spending.

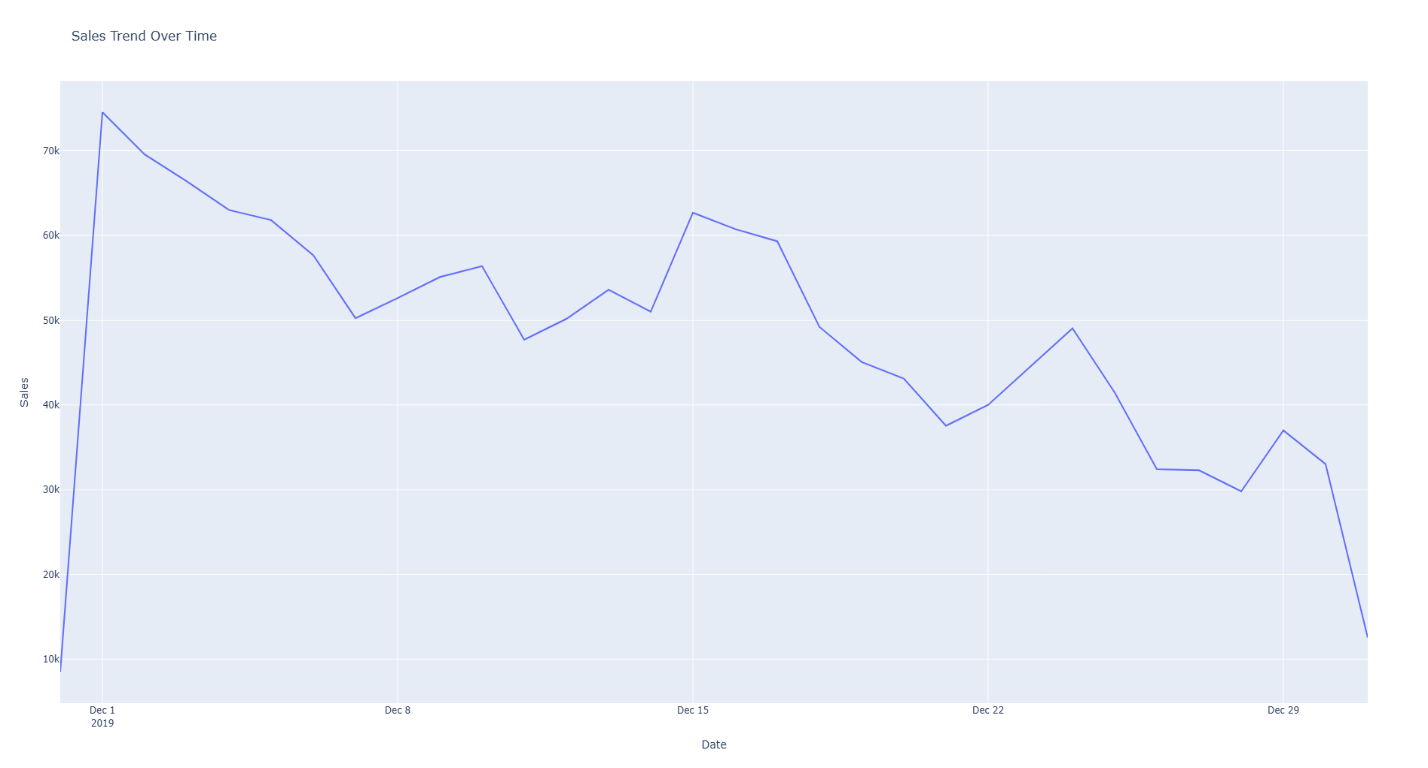




## **5.5** **Sales Trend Over Time:**

A line chart visualizes sales trends over time, helping to identify any seasonality or patterns.





# **6. Power BI Integration:**

#### 6.1 ****Load Data into Power BI using Python Script:****

To visualize data within Power BI, you can use a Python script to load data directly from DuckDB. This ensures that the latest data is always available.

