CPSC 531-02

Advanced Database Management

E-commerce Sales Analysis

Report by

Anisha Singh (885159103)
Sonal Mogra (884426719)

Table of contents

Introduction

The Evolving Landscape of E-commerce and the Importance of Data Analytics

The e-commerce industry is experiencing explosive growth, with global sales expected to reach $7.4 trillion by 2025. This dynamic landscape demands a constant evolution of strategies and tactics to remain competitive. To navigate this ever-changing environment, businesses must leverage data analytics to gain valuable insights into their customers, operations, and market trends. Conducting e-commerce analysis offers several significant benefits:

Data-Driven Decision Making: Move beyond intuition and make informed decisions based on concrete evidence. Optimize pricing, tailor marketing, personalize customer experiences, and develop targeted offerings.

Enhanced Operational Efficiency: Identify bottlenecks and inefficiencies in inventory, logistics, and fulfillment processes. Streamline operations, reduce costs, and improve margins and customer service.

Increased Customer Engagement: Understand customer preferences and purchase behavior. Personalize product recommendations, offer relevant promotions, and provide a more engaging shopping experience. Foster customer loyalty and brand advocacy.

Reduced Risk and Improved Profitability: Identify trends and anticipate future demand. Optimize inventory levels, prevent stockouts, avoid overstocking, and maximize profitability.

Competitive Advantage: Stay ahead of the curve with data-driven insights. Offer differentiated products and services, attract new customers, and retain existing ones.

Project Overview: E-commerce Sales Analysis with PySpark on Databricks

This project demonstrates the development and deployment of a scalable big data solution for e-commerce sales analysis using PySpark on Databricks. We leverage the power of cloud computing with AWS and Databricks to efficiently process and analyze 5 GB of sales data, revealing valuable insights for strategic business decisions.

Project Objectives:

- **Robust Big Data Architecture:** Utilize AWS, Databricks, and Jupyter Notebooks for a scalable and efficient solution for large datasets.
- **Cloud-Based Data Processing:** Demonstrate PySpark job deployment and execution on Databricks, highlighting cloud-based scalability and resource management.
- **Actionable Insights:** Analyze sales data with PySpark and Matplotlib to uncover insights into product performance, customer behavior, and market trends.
- **Business Value of Data Analytics:** Showcase how data-driven insights can inform strategic decisions, optimize operations, and improve customer engagement.

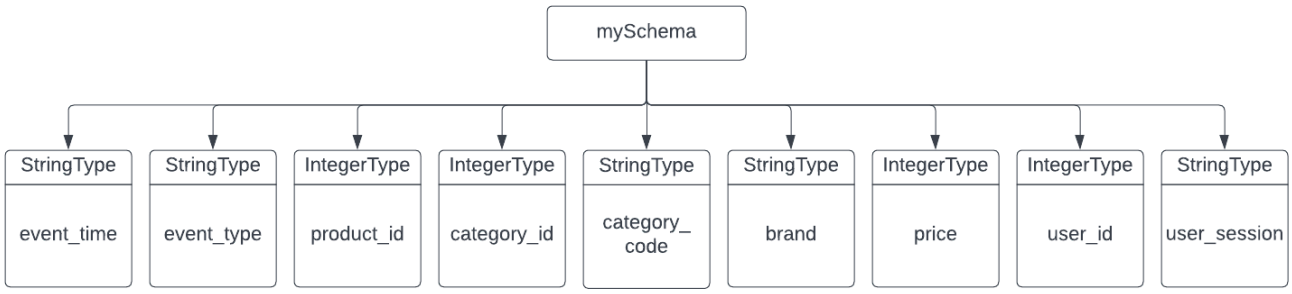
Project Benefits:

This project showcases the power of cloud-based big data analytics in revolutionizing e-commerce operations. By harnessing PySpark on Databricks, businesses can extract actionable insights from vast amounts of data, enabling informed decisions, optimized operations, and a competitive edge.

Functionalities

1. **Dataset:** The foundation of our analysis is a comprehensive dataset from a multi-category e-commerce store, sourced from Kaggle. This dataset, sized at 5GB, encompasses over 42 million rows, providing a rich canvas for our big data endeavors.

   Schema:



2. **Processing Tool:**
   - PySpark: Our project harnesses PySpark for data processing, utilizing its ability to perform complex algorithms across clusters.
   - Python: Python's versatility is employed for scripting and automation, enabling sophisticated data manipulation and analysis.

3. **Data Storage:**
   - Amazon S3: We utilize Amazon S3's robust infrastructure to securely house our extensive dataset.

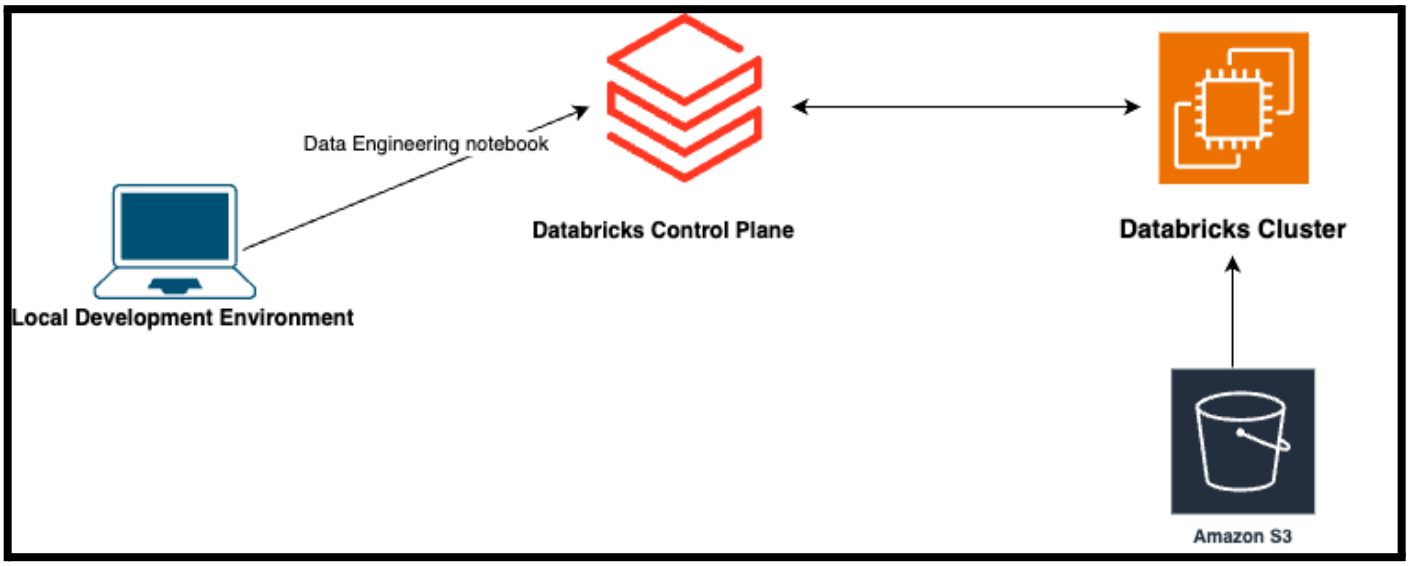
4. **Development Environment:**
   - Jupyter Notebooks: We ensure seamless local development through the use of Jupyter Notebooks, which offers an intuitive interface for writing and testing PySpark code before deployment.

5. **Data Processing and Analysis:**
   - Databricks on AWS EC2: We leverage Databricks hosted on AWS EC2 for its powerful data processing capabilities.
   - Real-Time Analytics: Our architecture enables the real-time processing of massive datasets, offering immediate insights.

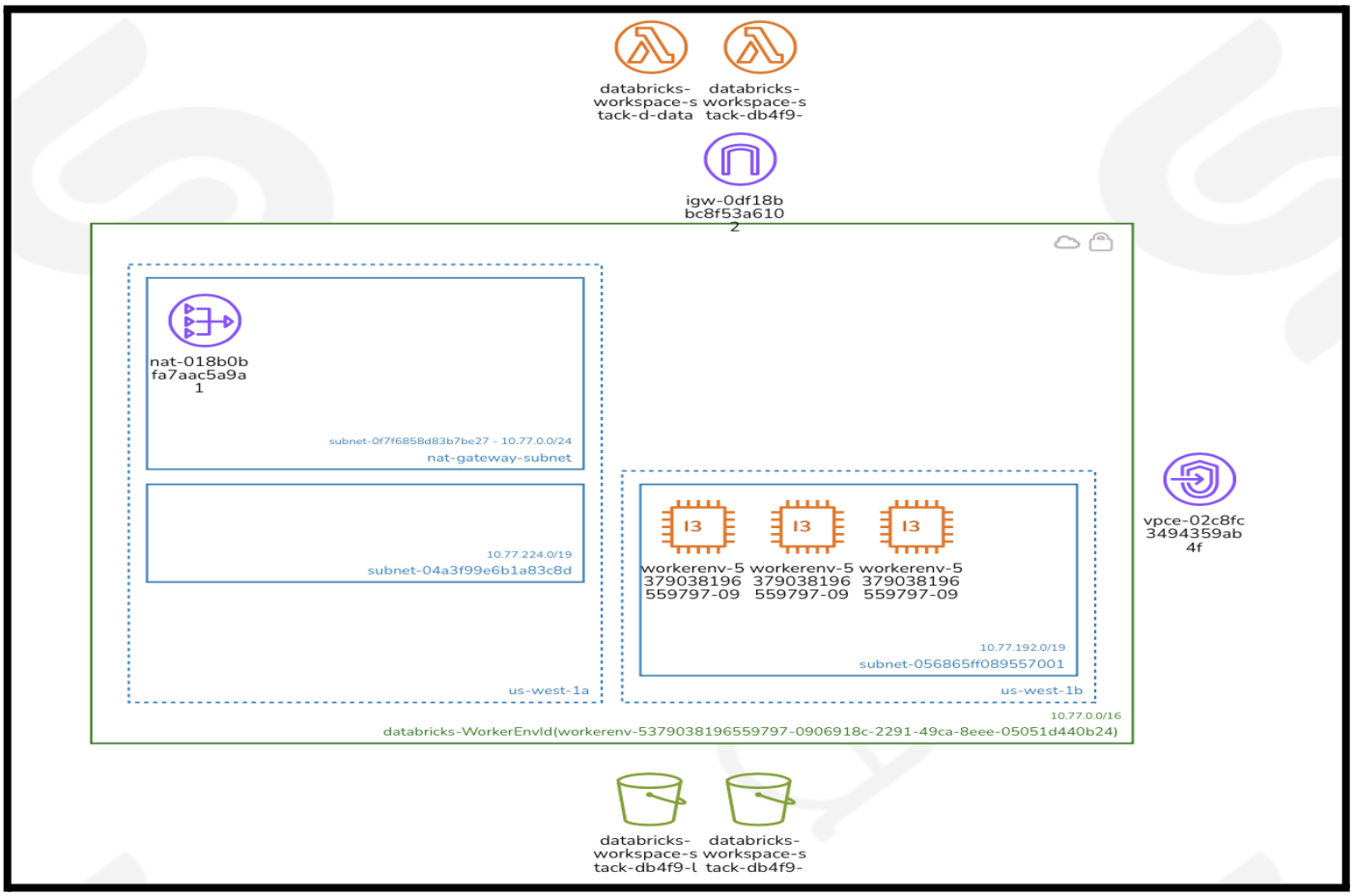
6. **Job Scheduling:** Manages and automates job scheduling within Databricks, seamlessly syncing with local Jupyter notebooks for efficient cloud-based execution.

7. **Visualization:**
   - Matplotlib: Data visualization is accomplished using Matplotlib, enhancing our ability to communicate findings and patterns in the data through graphical representations.

Architecture & Design

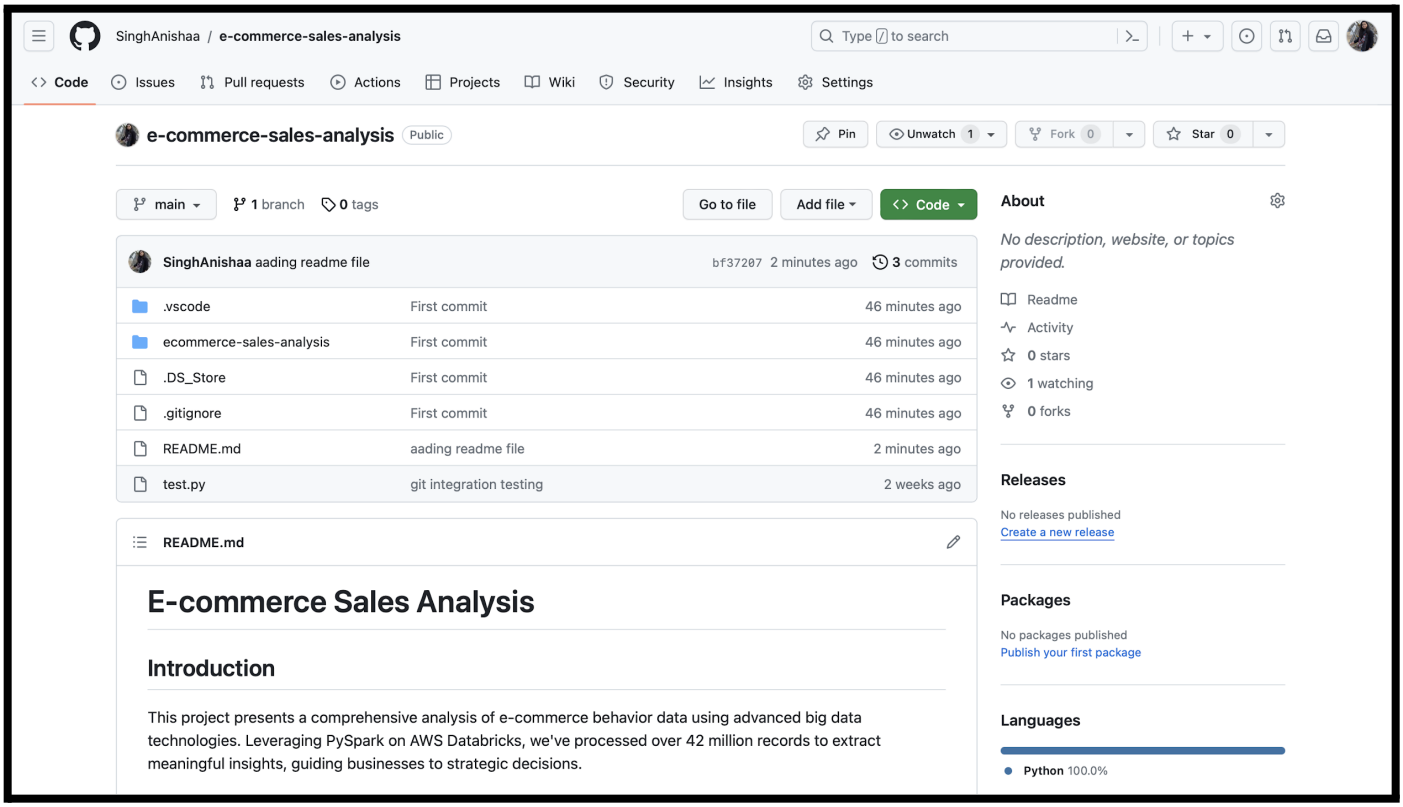


- **Amazon VPC architecture:**



GitHub Location of Code

https://github.com/SinghAnishaa/e-commerce-sales-analysis.git

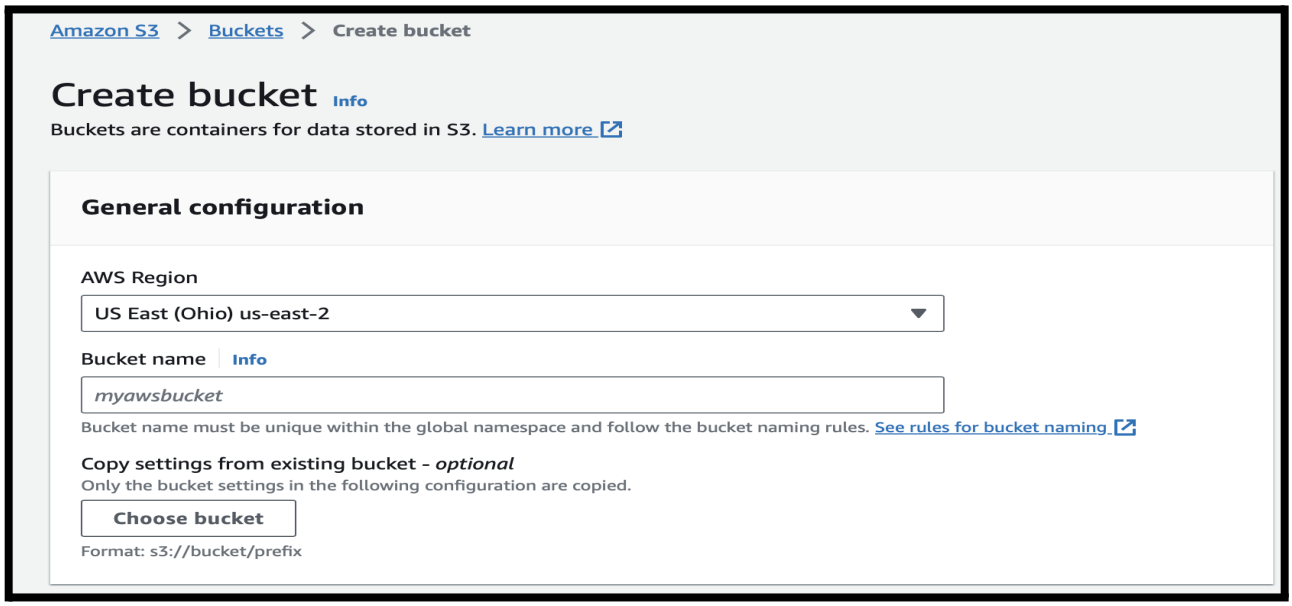


Deployment Instructions

1. **AWS S3 Bucket Setup:**

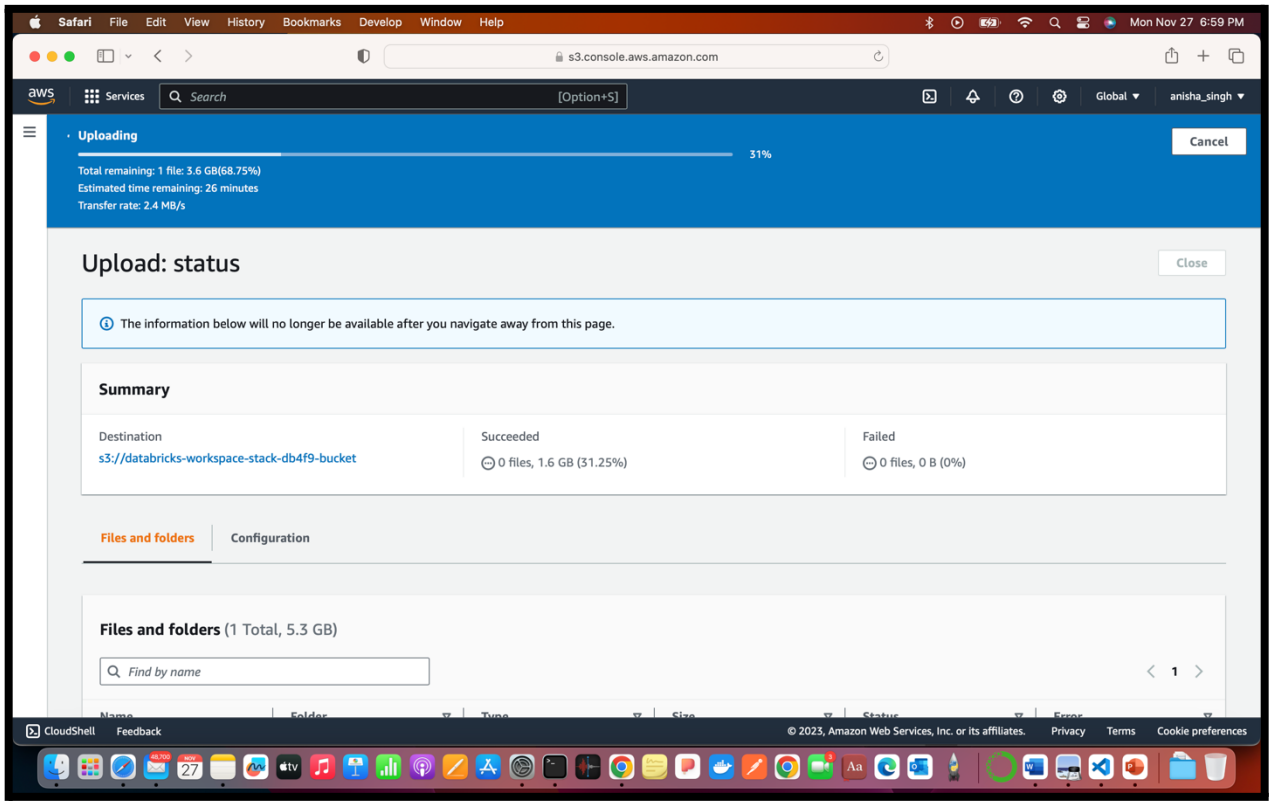
   Step a: Log in to AWS: Create an account and login to AWS services console.

   Step b: Create a new S3 bucket:



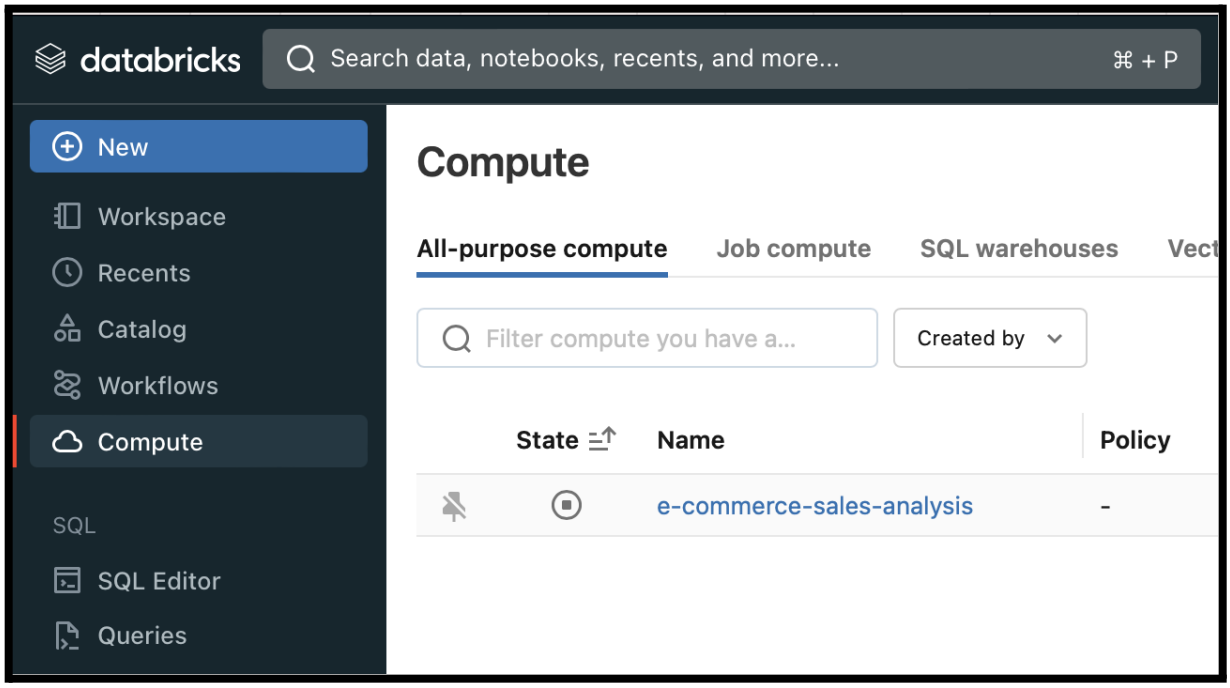
Step c:

- Click "Upload" in the bucket's interface
- Select your dataset files from your local computer.
- Review the settings and permissions for the files, then click "Upload".

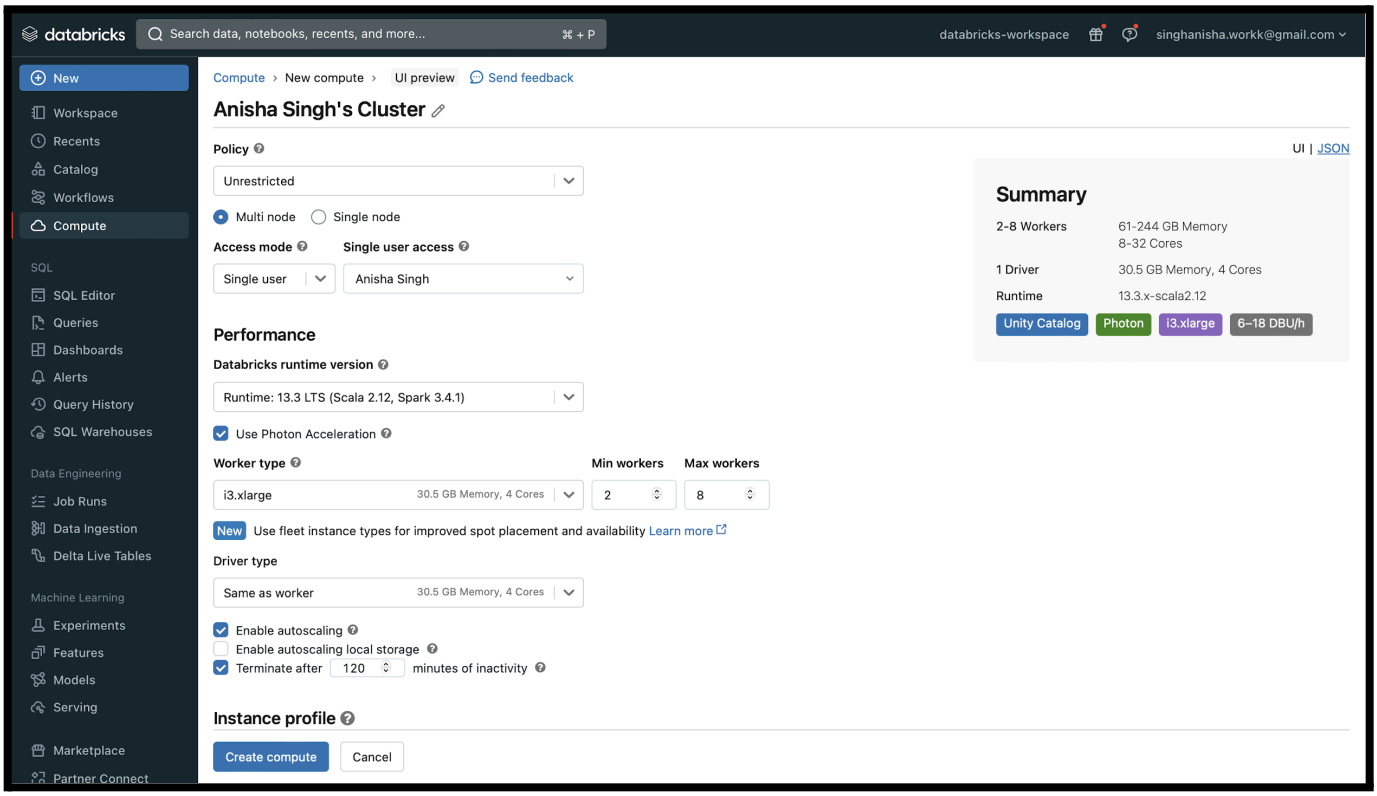


2. **Create and configure a Databricks Cluster:**

Step a: Log in Databricks workspace and navigate to Compute section.



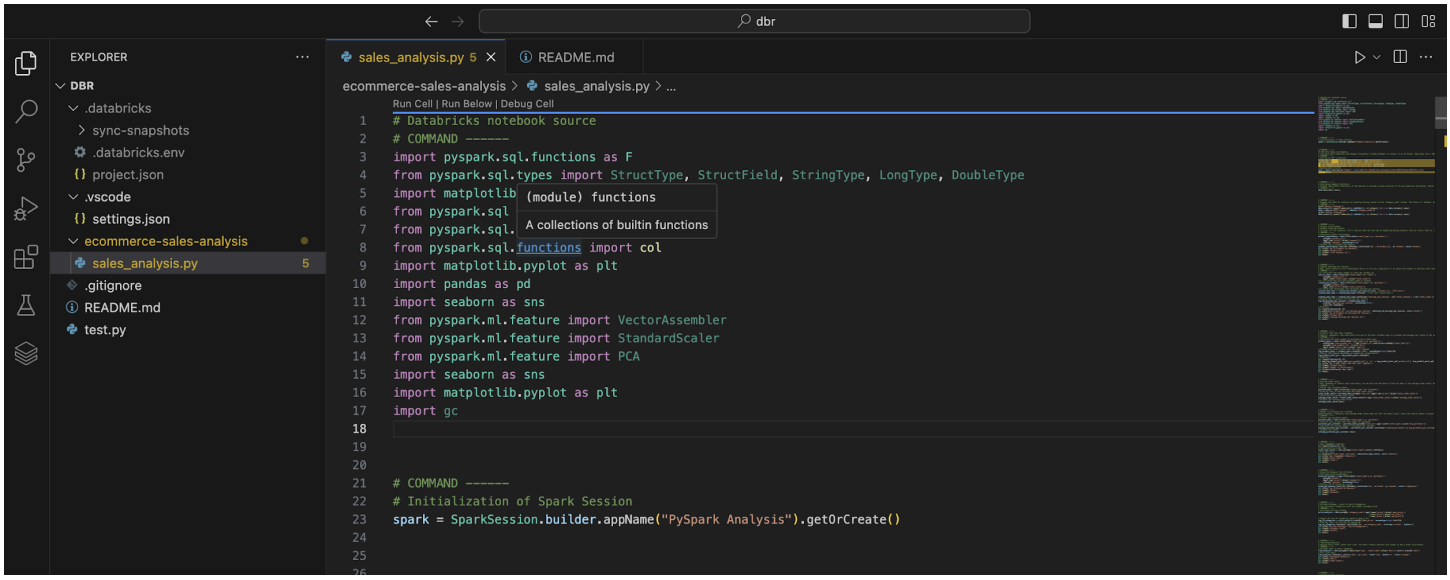
Step b: Click on “Create Compute”, fill the desired configuration and click on “Create Compute".



Step c: Upload the converted Jupyter Notebooks to Databricks and attach them to your cluster for execution.

Steps to Run the Application

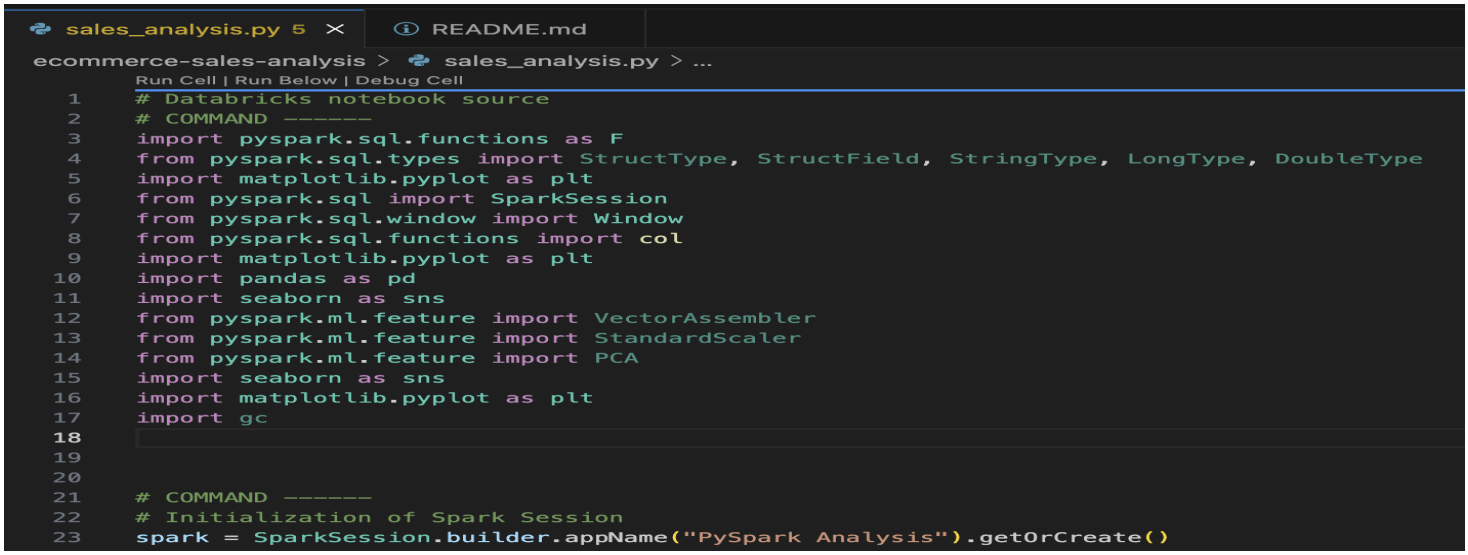
Step 1: Develop your code in your local in a .py file:



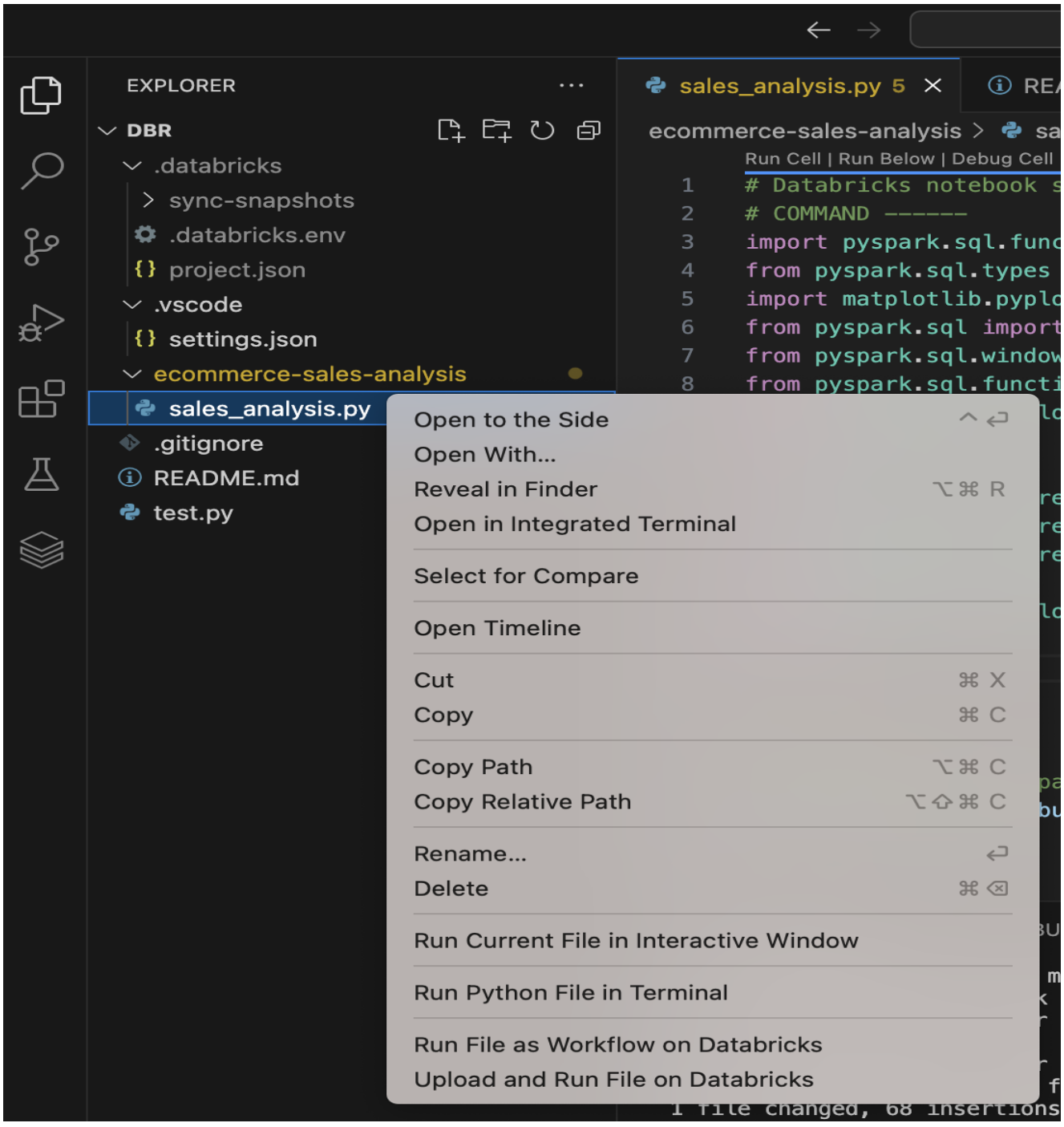
Step 2: Convert your .py script to a Databricks-compatible Jupyter notebook:

Use the command **# Databricks notebook source** at the start of the .py file to convert it into notebook.

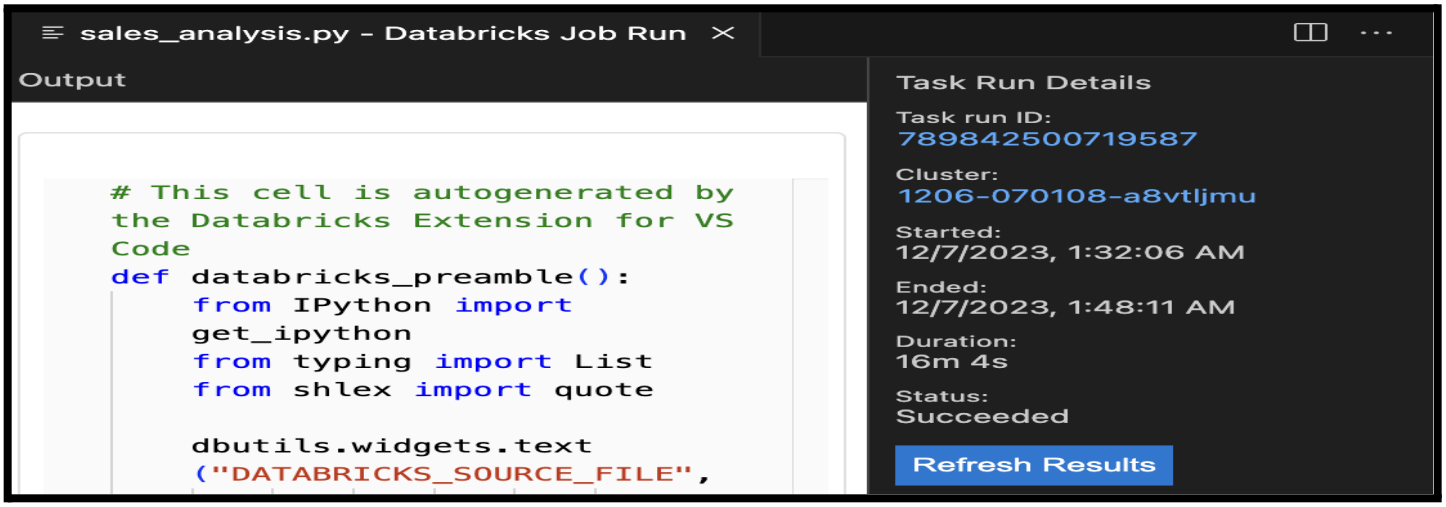
Step 3: Split your .py script into cells, preceded by **"# COMMAND ----------"** to denote separate executable cells.



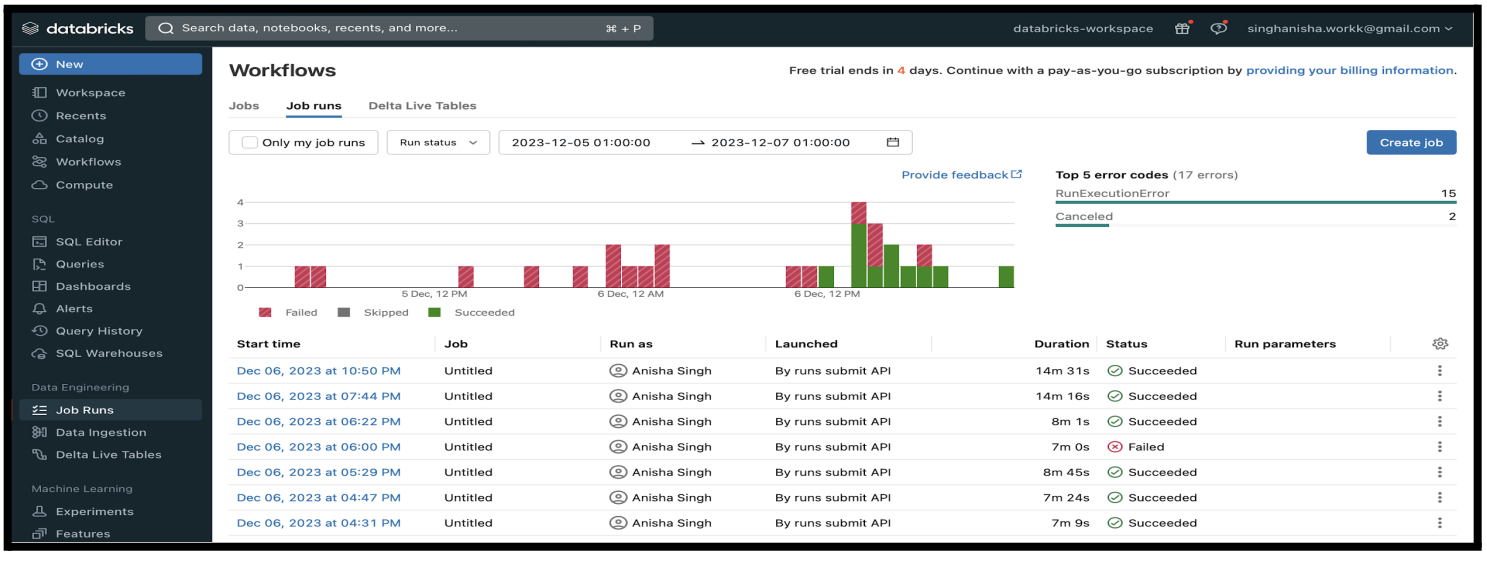
Step 4: Run the notebook as a job on the Databricks cluster:



Step 5: View real-time execution of code on databricks cluster by clicking on the “Task run ID” which gets generated for each run:

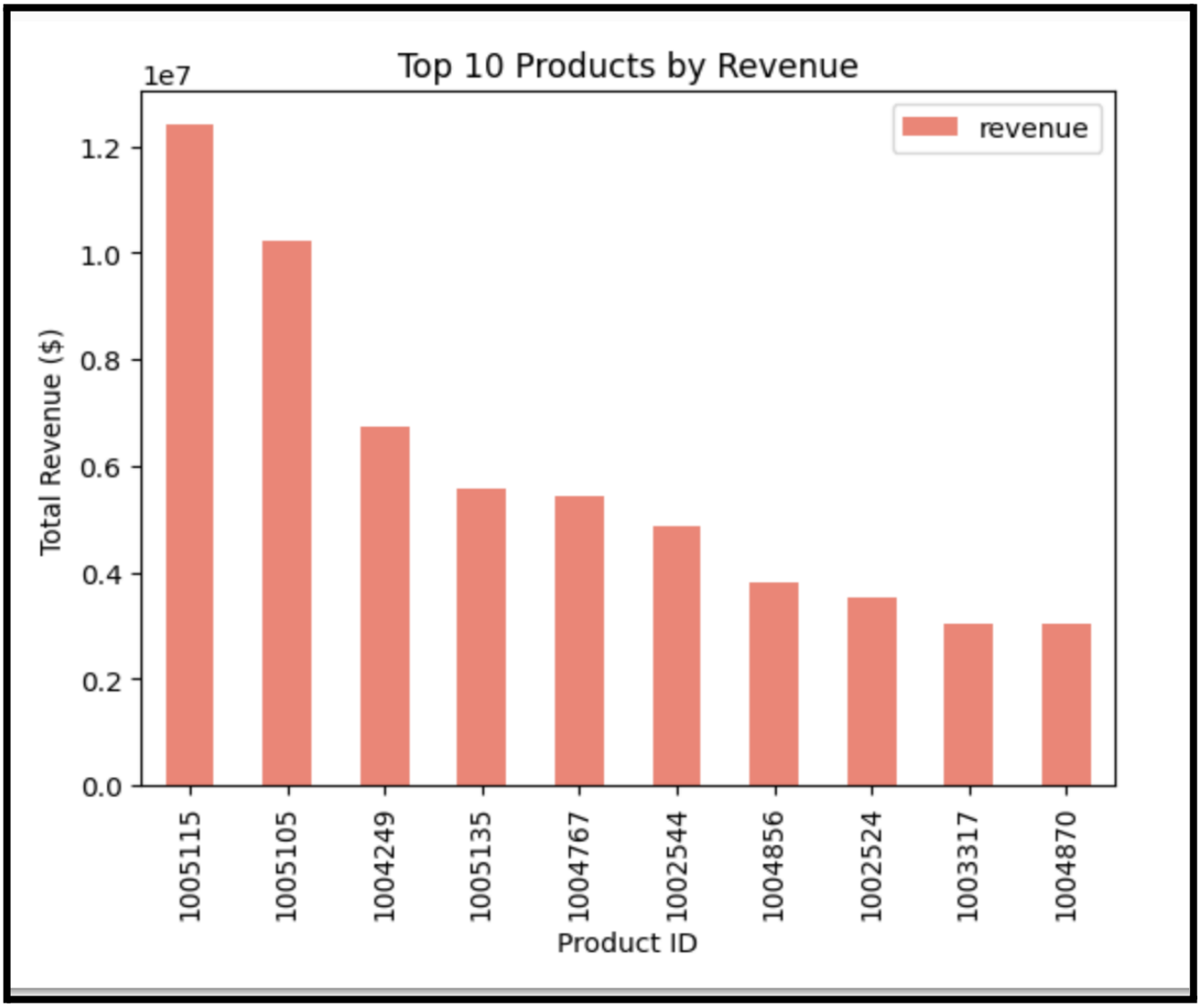


Step 6: Refer to “Jobs Runs” dashboard for information on each run.

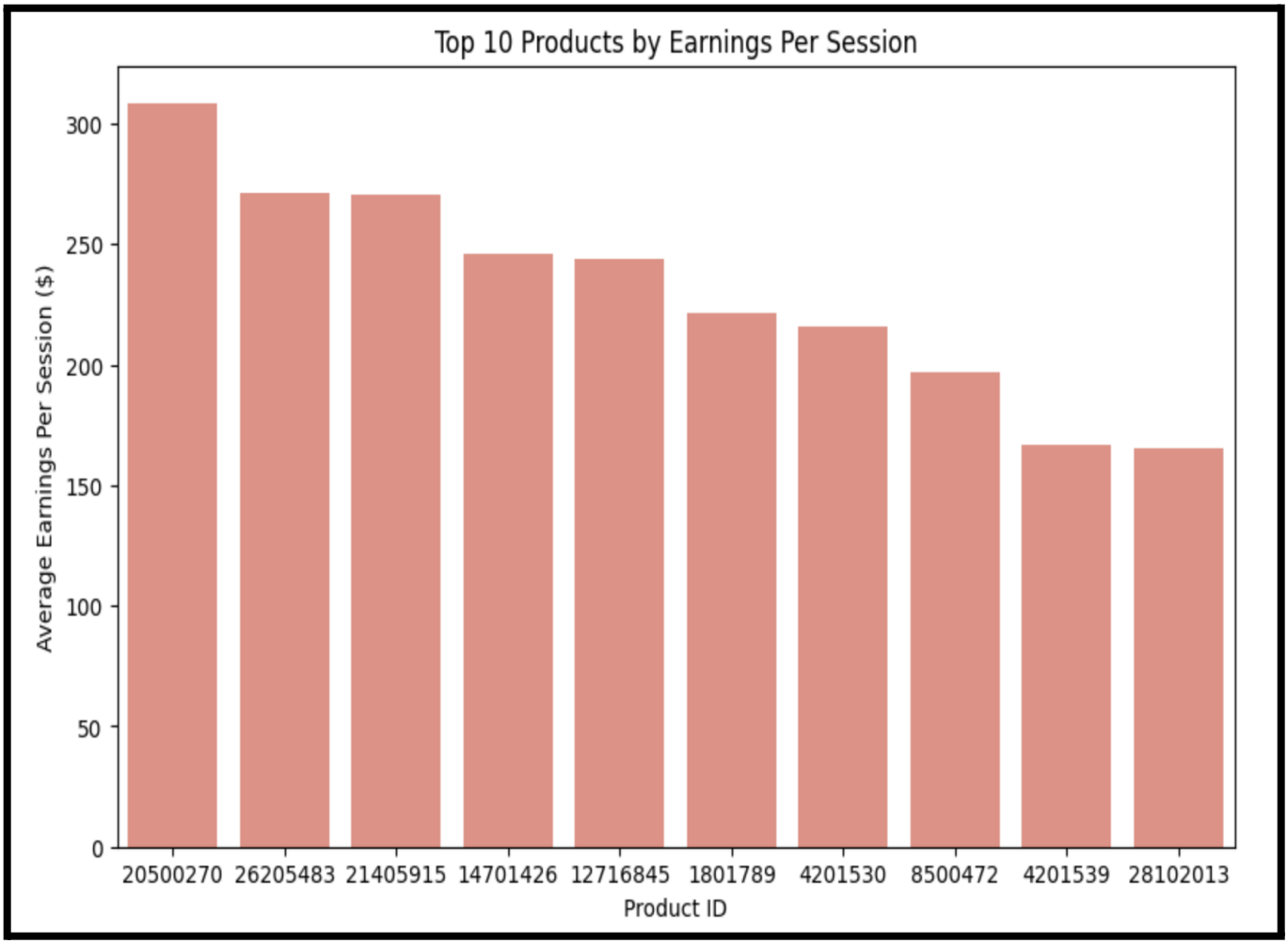


Test Results

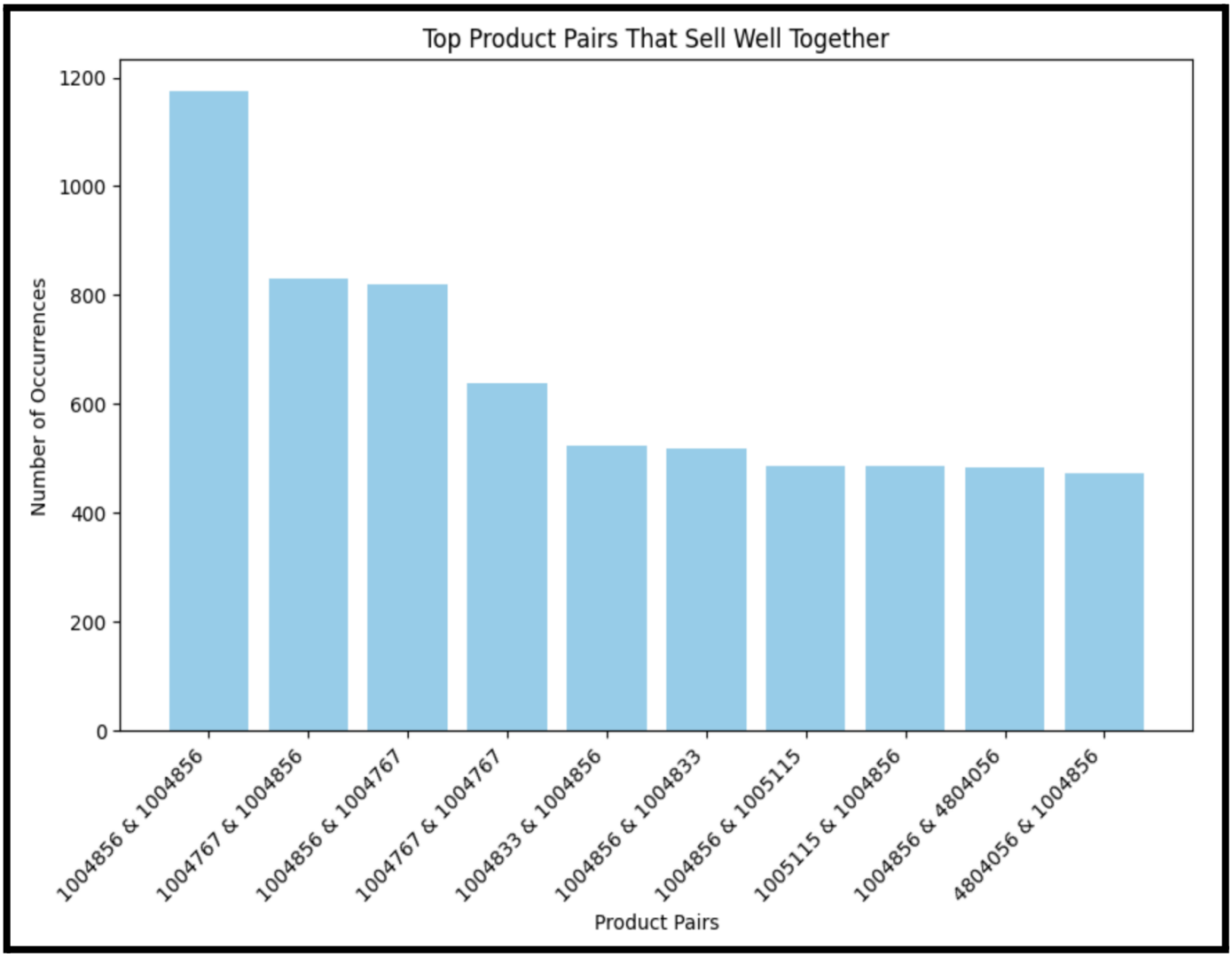
- **Highest Grossing Product:** To start off our analysis, we found out what are the top 10 highest-grossing products from our store, that is, the products that make the most amount of money in terms of revenue.



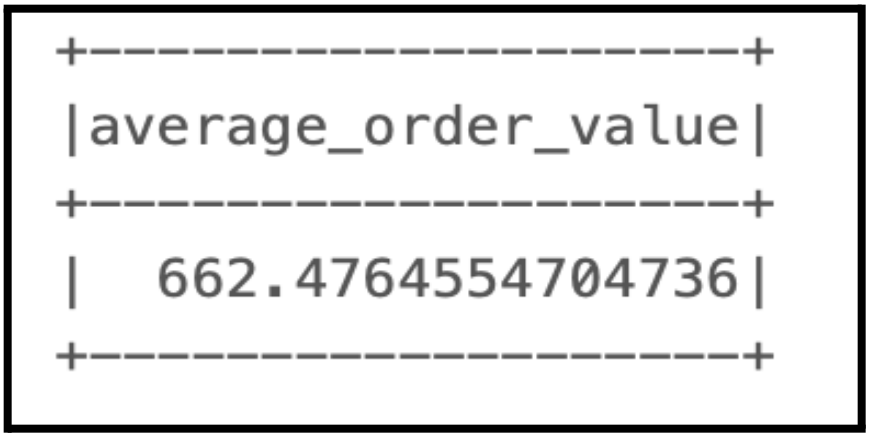
- **Highest Earnings Per Session:** Now, total revenue is not a meaningful metric on its own, especially if we ignore the number of sessions that landed on the product page. After all, if you are sending a lot of traffic to a page that is doing a poor job at converting users, you could easily be leaving money on the table. So, to account for the fact that different products have different conversion rates, we found out what are the top 10 products by earnings per session.



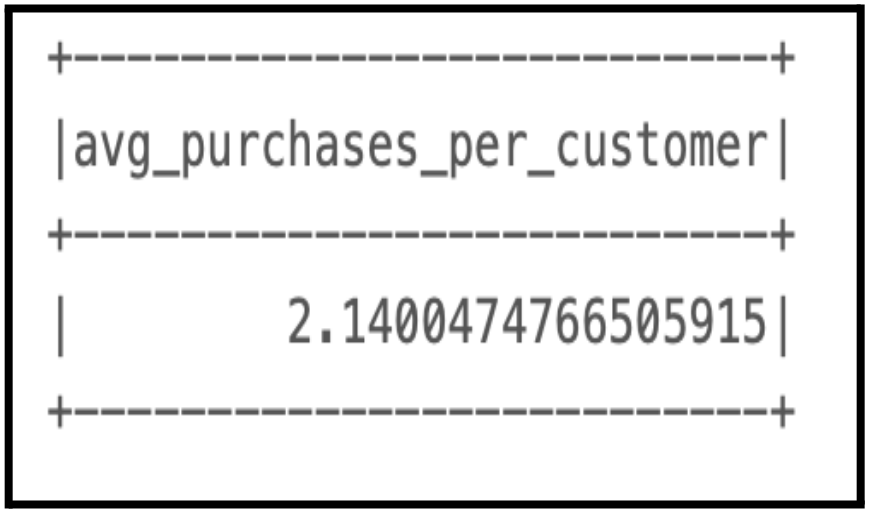
- **Products That Sell Well Together:** Upsells, down-sells, and cross-sells are one of the most reliable ways to increase the average cart value of the customer. But to optimize them, the products offered cannot be picked at random. After all, each and every product caters to different customer avatars, and trying to use a one size fits all approach on the upsell path is a guaranteed way to leave money on the table. So, we dug into the data and find out what products users tend to buy in the same session.



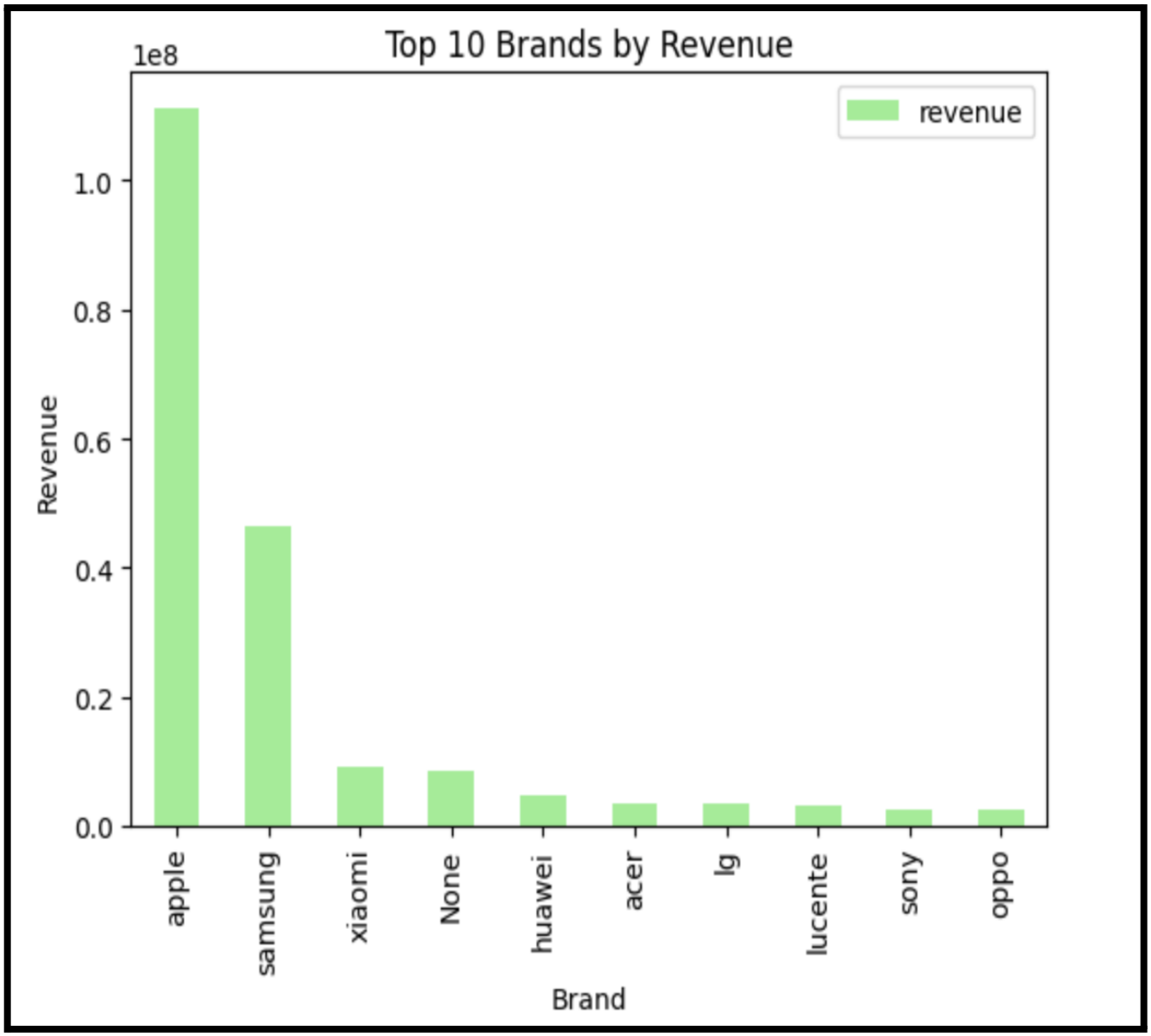
- **Average Order Value:** Now, speaking of upsells and cross-sells, we also used the data to find out what is the average order value:



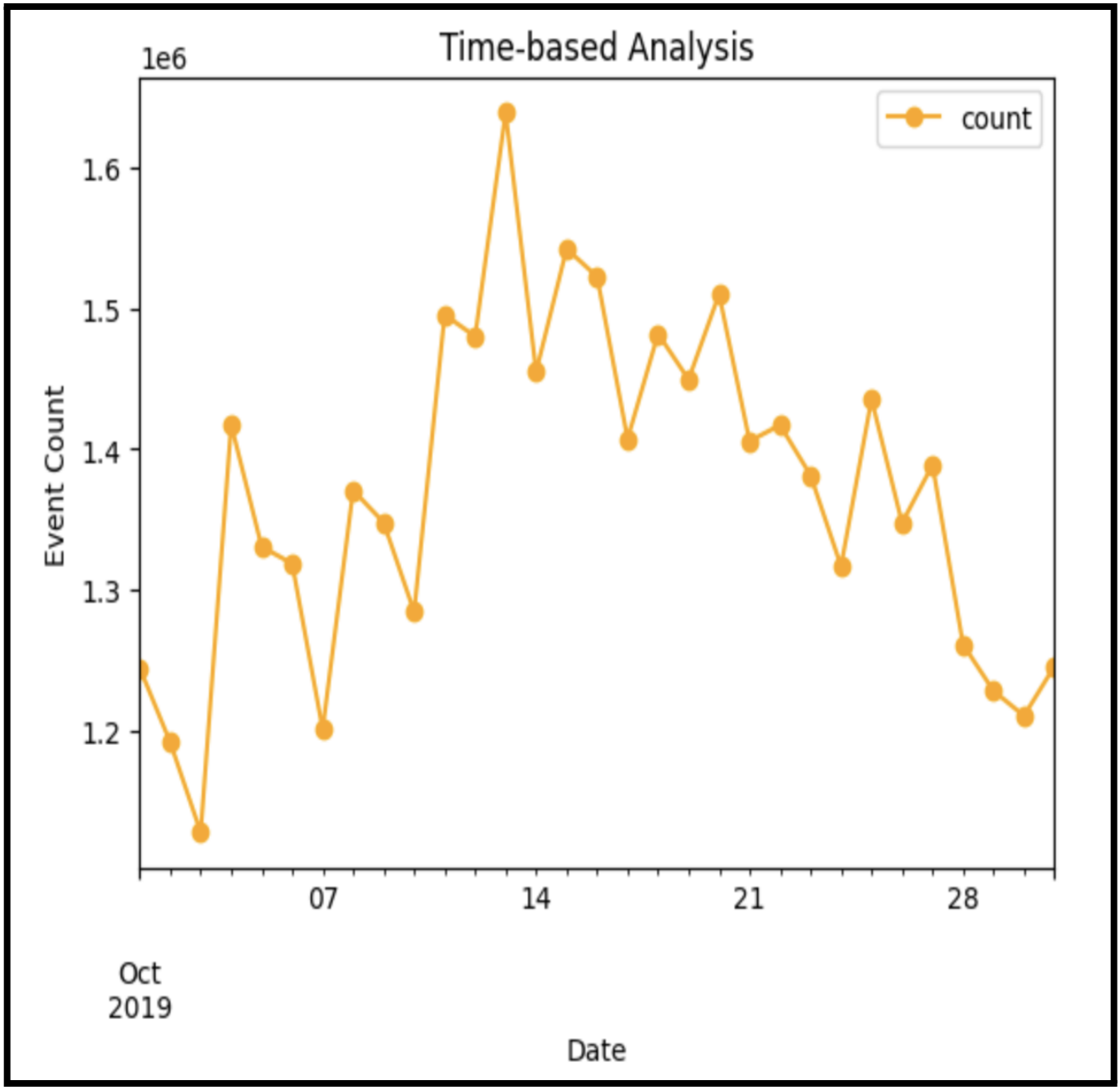
- **Number of purchases per customer:** Unfortunately, looking at the Average Order Value does not tell the whole story, since this metric doesn't account for the fact that some customers make more than one purchase and keep coming back. So, we looked at an important metrics of customer retention: The average number of purchases per customer.



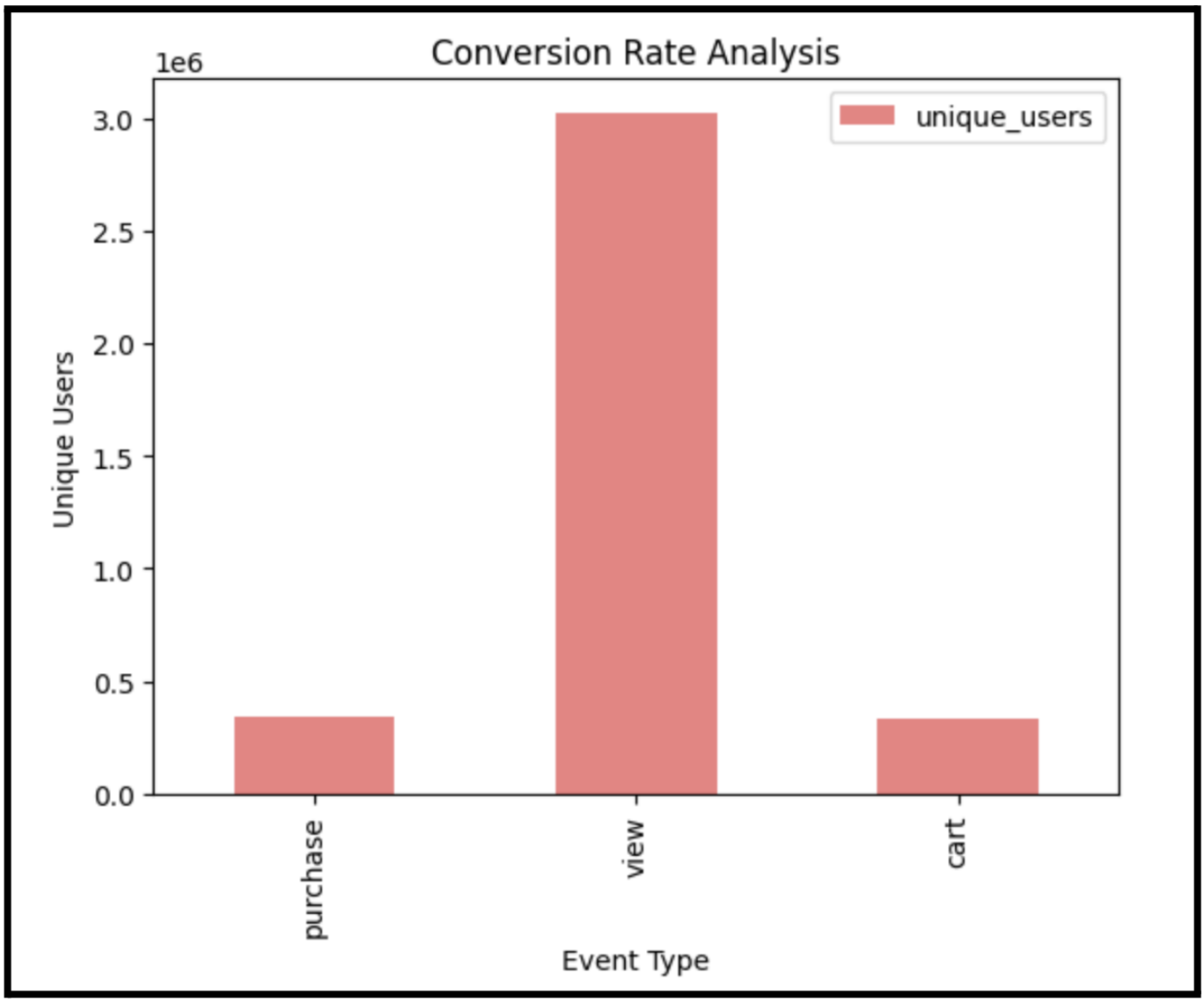
- **Brand Performance (Top 10 Brand):** Delving into brand dynamics, we identified the top 10 brands that are excelling in the market. By measuring their sales volumes, revenue generation, and customer preferences, we gained insight into which brands are leading the pack and why.



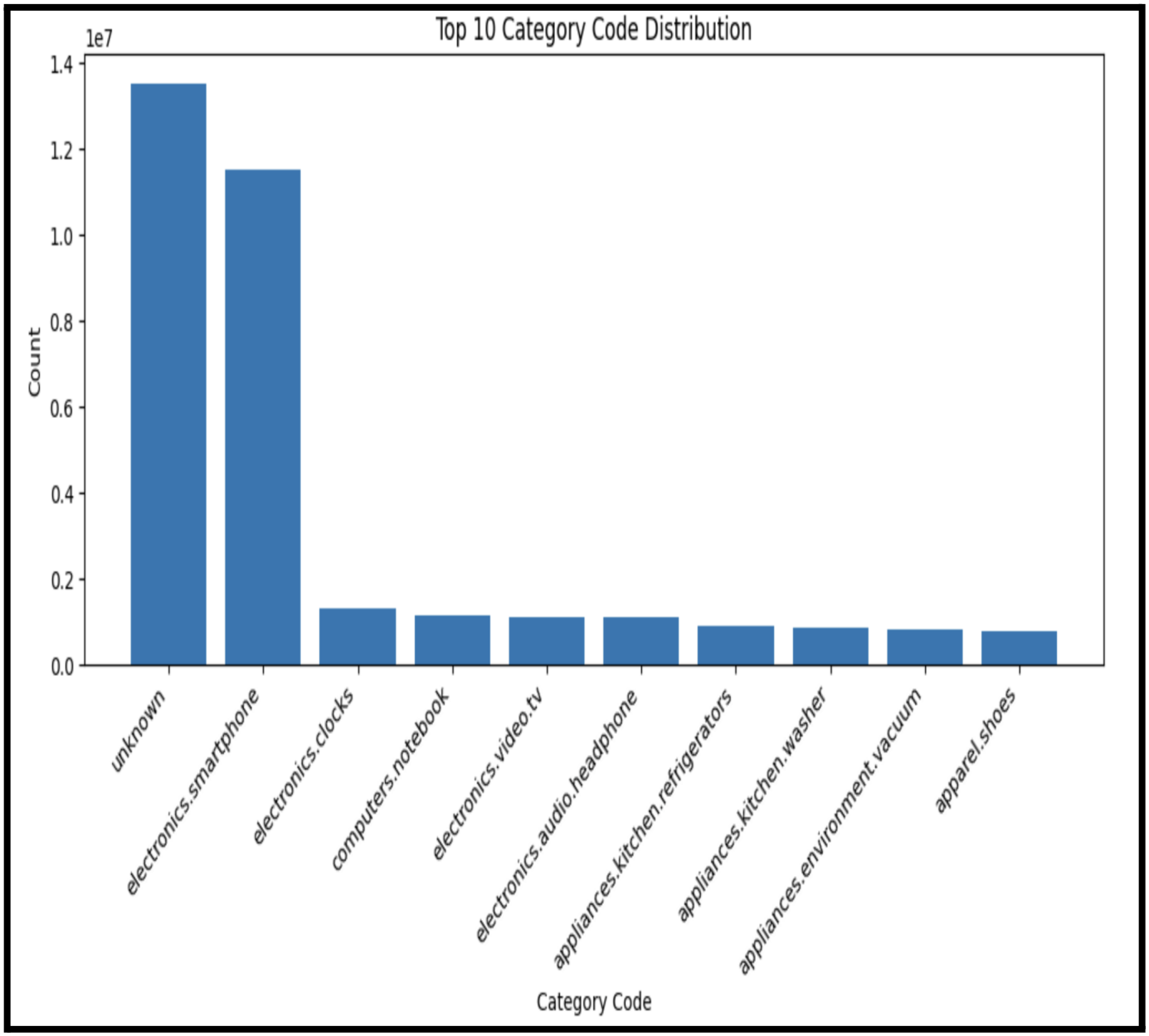
- **Time-based Analysis:** We conducted a thorough examination of purchasing patterns over time, identifying any seasonal trends, fluctuations, and consistencies in consumer behavior. This temporal analysis was crucial in forecasting demand and optimizing stock levels.



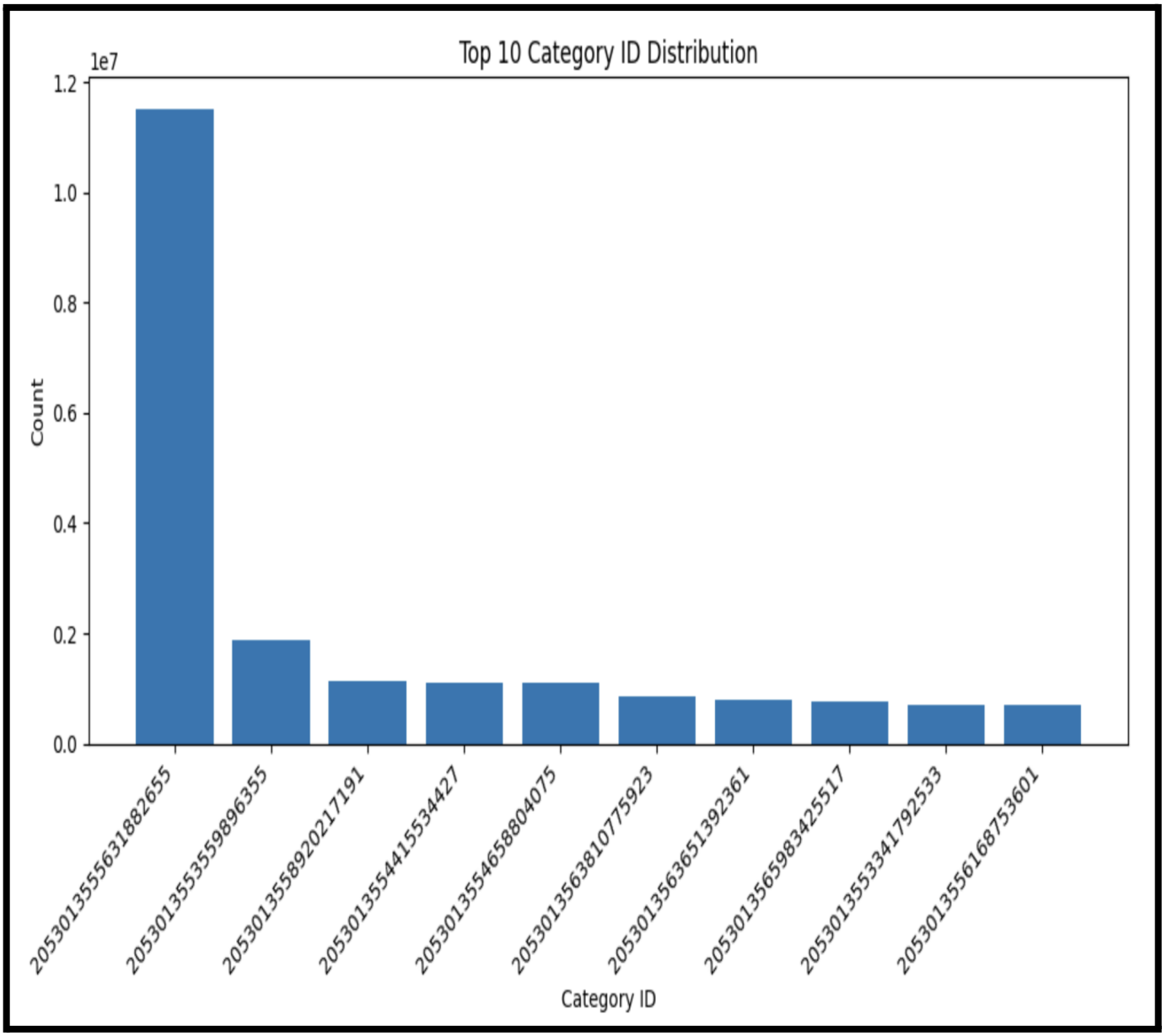
- **Conversion Rate Analysis:** Understanding the funnel from visitor to customer is critical. We computed the conversion rates for each step in the customer journey, shedding light on where we're winning customers and where there's room for improvement.



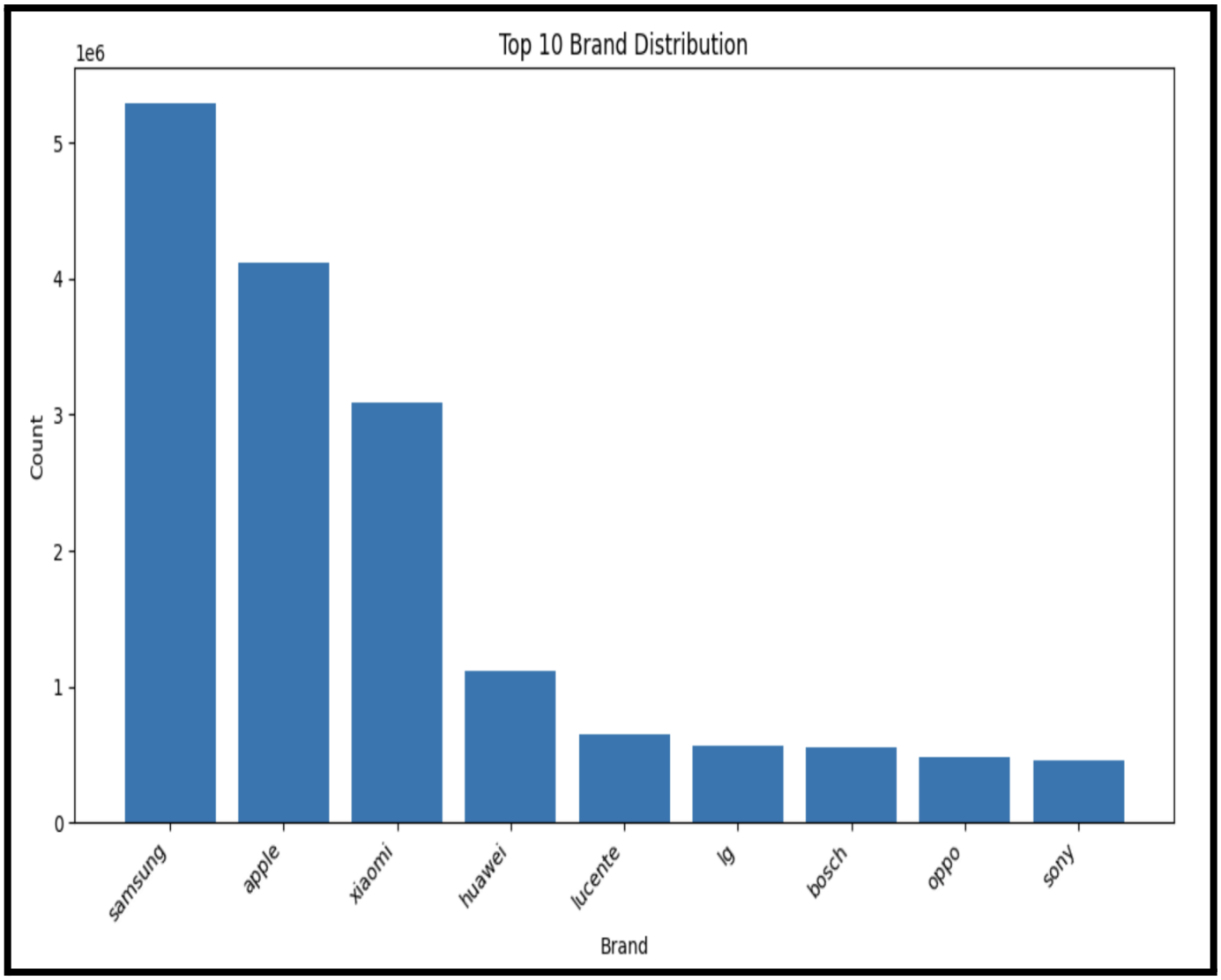
- **Plot the category code distribution:** By mapping out the frequency of purchases across different product categories, we uncovered which categories are the most popular and drive the most engagement, enabling targeted marketing strategies.



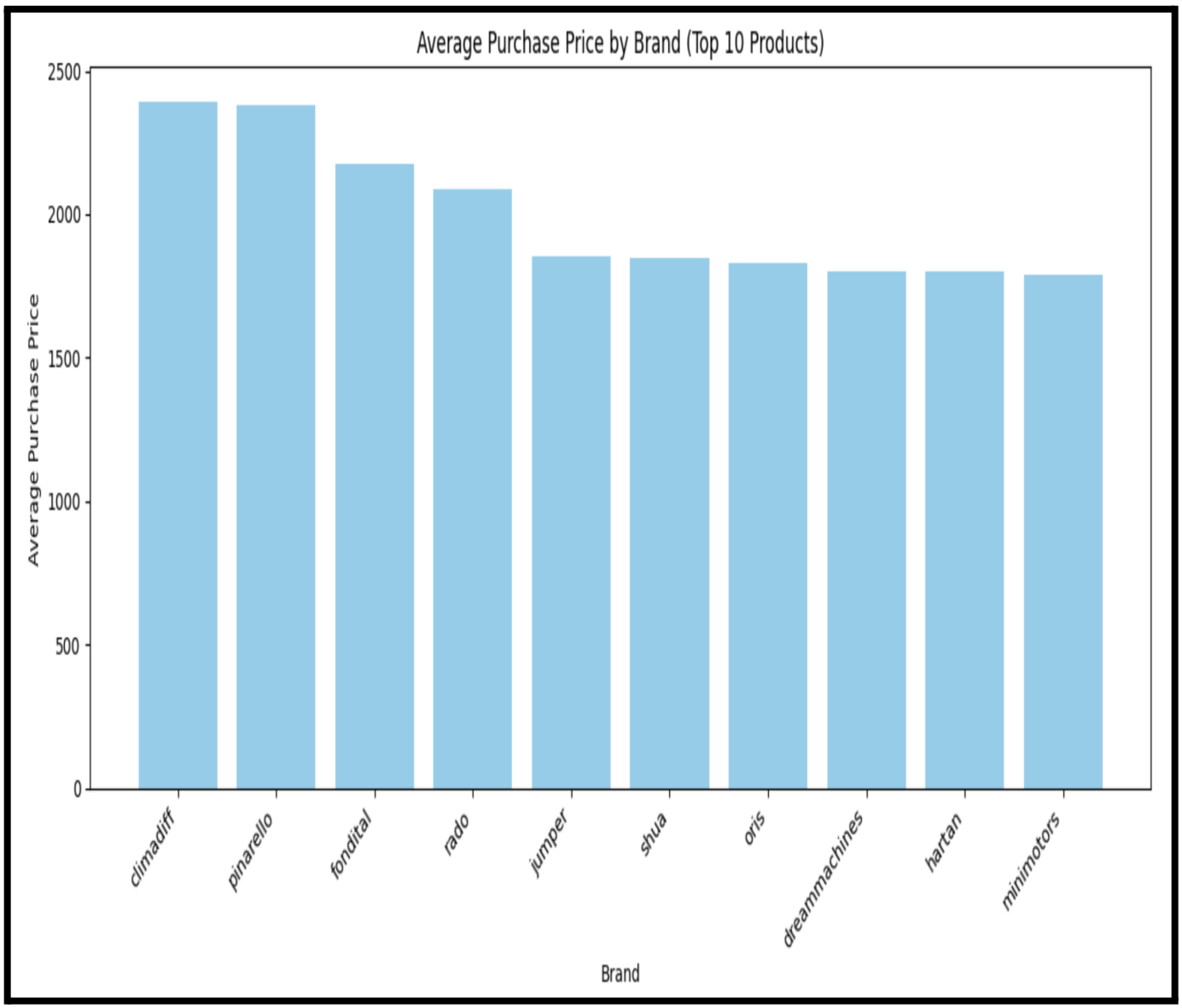
- **Plot the category ID distribution:** Like category codes, we also analyzed the distribution of sales by category ID, focusing on identifying the top-performing categories for strategic stock and promotion planning.



- **Plot the brand distribution:** We visualized how different brands stack up against each other in terms of sales events, providing a clear picture of brand market share and popularity.



- **Average purchase price by brand:** By calculating the average purchase price for each of the top brands, we understood their positioning in the market—are they premium or budget-friendly? This helps in aligning marketing messages.



- **Total revenue:** We quantified the total revenue generated, presenting a clear and concise overview of the financial health and success of e-commerce venture.