**RETAIL SALES ANALYSIS**

**By**

**MOHAMMED ABDUL MUEED ALI**

**INTRODUCTION**

The "Retail Sales Analysis" project aims to analyze sales data from a retail store to derive insights and inform business decisions. Here I am using Apache Spark in Databricks community edition to collect the data, process the data, analyse and visualize the retail sales data.

Objectives:

1. Understand overall sales performance.
2. Identify top-selling products.
3. Analyze sales trends over time.
4. Explore customer demographics to target marketing efforts effectively.

**DATA COLLECTION**

Dataset: The project uses a dataset containing retail sales data, including transaction IDs, product IDs, sales amounts, timestamps, and customer IDs.

Source:

import pandas as pd

import numpy as np

from faker import Faker

import random

import datetime

# Initialize Faker library for generating fake data

fake = Faker()

# Define the number of transactions

num\_transactions = 1000

# Generate synthetic data for retail sales

sales\_data = []

for \_ in range(num\_transactions):

    transaction\_id = fake.uuid4()

    customer\_id = fake.random\_int(min=1, max=100)

    product\_id = fake.random\_int(min=1, max=50)

    sales\_amount = round(random.uniform(10, 1000), 2)

    timestamp = fake.date\_time\_between(start\_date='-1y', end\_date='now')

    sales\_data.append([transaction\_id, customer\_id, product\_id, sales\_amount, timestamp])

# Create a DataFrame

columns = ['transaction\_id', 'customer\_id', 'product\_id', 'sales\_amount', 'timestamp']

sales\_df = pd.DataFrame(sales\_data, columns=columns)

# Save the DataFrame as a CSV file

sales\_df.to\_csv('retail\_sales\_data.csv', index=False)

Import Libraries:

1. pandas: Used for data manipulation and creating DataFrames.
2. numpy: Used for numerical operations.
3. Faker: A Python library for generating fake data.

Initialize Faker Library:

1. Initialize the Faker library to generate fake data such as names, addresses, and dates.

Define Parameters:

1. Define the number of transactions to generate (num\_transactions).

Generate Synthetic Data:

1. Generate synthetic data for retail sales using a loop. For each transaction, generate:
   1. transaction\_id: A unique identifier for the transaction.
   2. customer\_id: A random integer between 1 and 100 representing the customer ID.
   3. product\_id: A random integer between 1 and 50 representing the product ID.
   4. sales\_amount: A random float between 10 and 1000 representing the sales amount.
   5. timestamp: A random date and time within the past year.

Create DataFrame:

1. Create a DataFrame using the generated sales data and define column names.

Save DataFrame as CSV:

1. Save the DataFrame as a CSV file named 'retail\_sales\_data.csv' without including the index.

Overall, this code demonstrates how to generate synthetic retail sales data using the Faker library, create a DataFrame in pandas, and save it as a CSV file for further analysis.

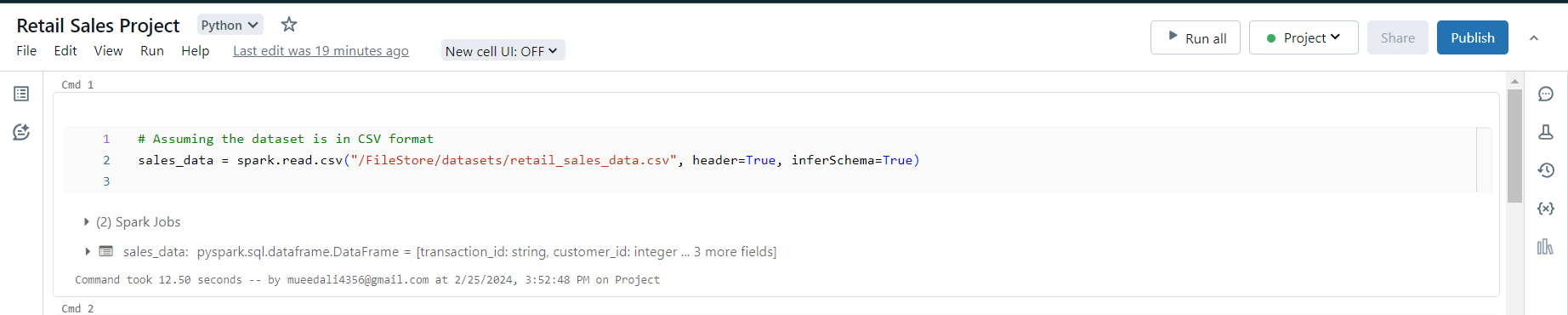
Data Format: The dataset is in CSV format and consists of structured data with rows and columns.



**DATA EXPLORATION**

Summary Statistics: Descriptive statistics such as mean, median, min, max, and standard deviation of numerical variables.

Visualizations: Histograms, box plots, and scatter plots to explore distributions and relationships between variables.



**DATA PREPROCESSING**

Handling Missing Values: Dropping rows with missing values or imputing missing values using appropriate techniques.

Data Cleaning: Removing duplicates, correcting errors, and ensuring consistency in data format.

Data Transformation: Converting categorical variables into numerical format, scaling numeric features if necessary.

**DATA ANALYSIS**

Total Sales Amount: Calculation of the total sales amount from the dataset.

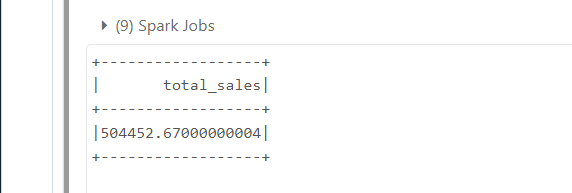
Top-Selling Products: Identification of the top-selling products based on sales amount.

Sales Trends Over Time: Analysis of sales trends by month or year to identify seasonal patterns or trends.

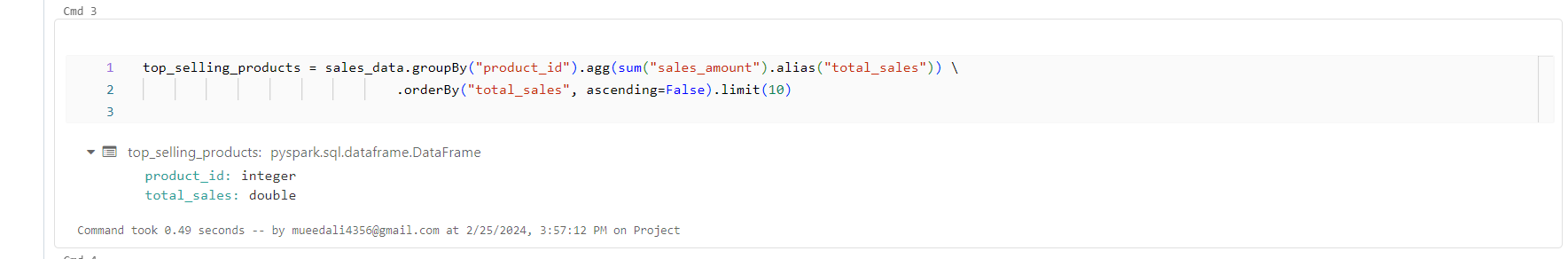
Customer Demographics: Exploration of customer demographics such as age, gender, and location to understand the target market.

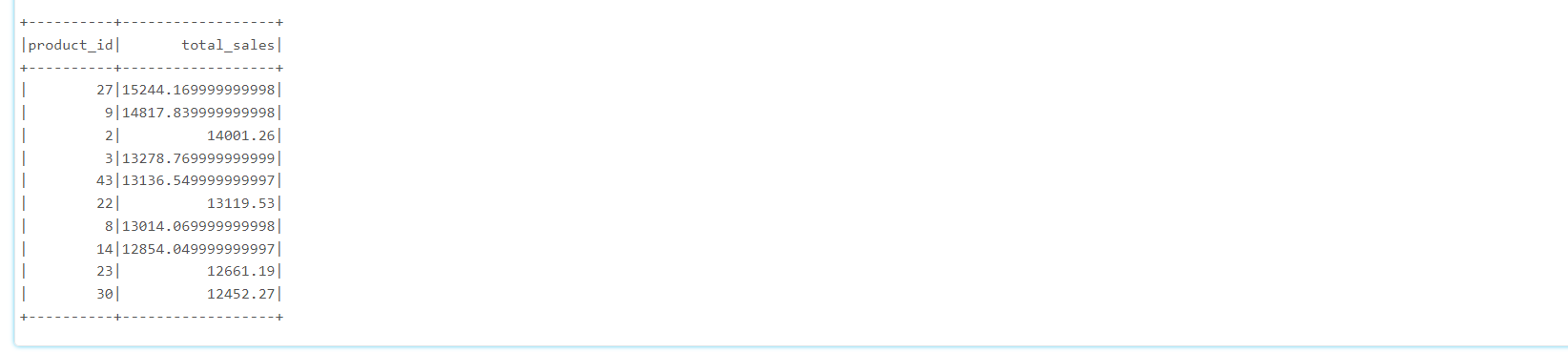
Calculate total sales amount:





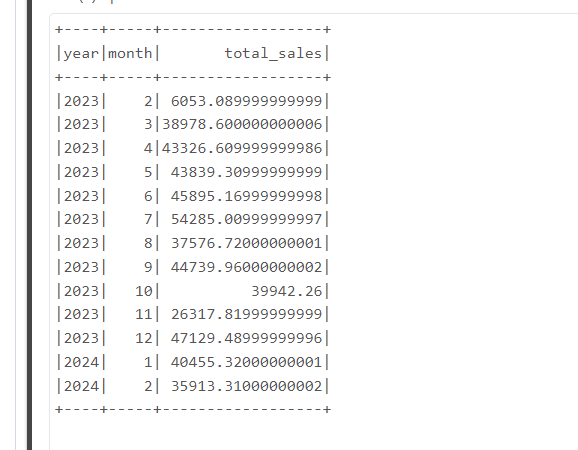
Identify top-selling products:





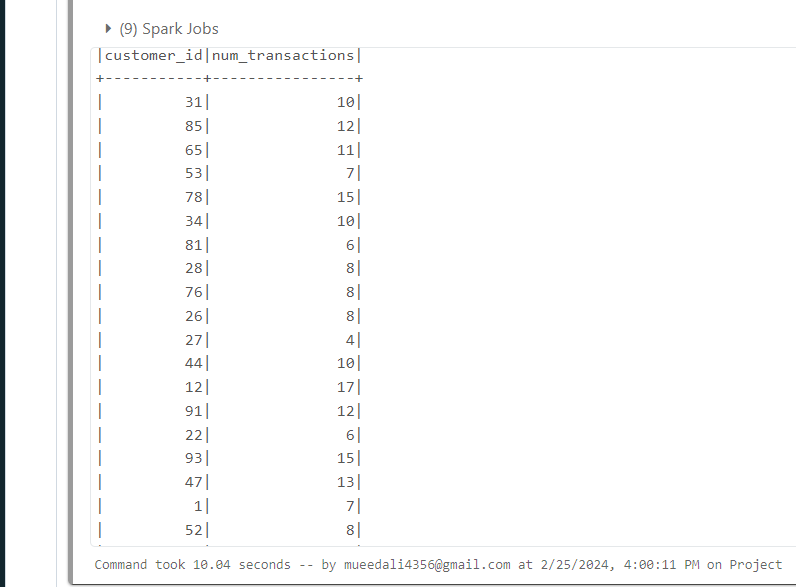
Analyse sales trends over time:





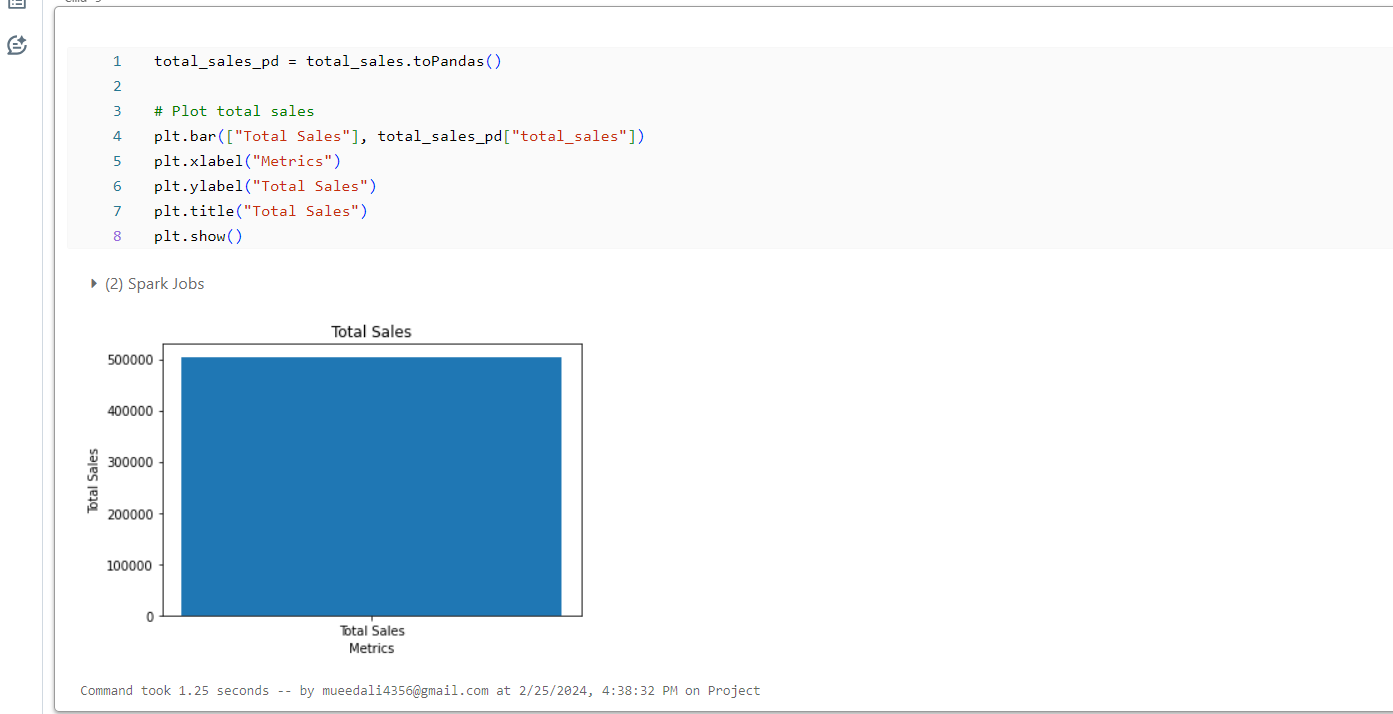
Explore customer demographics:





**VISUALIZE**

**Visualize Total Sales:**



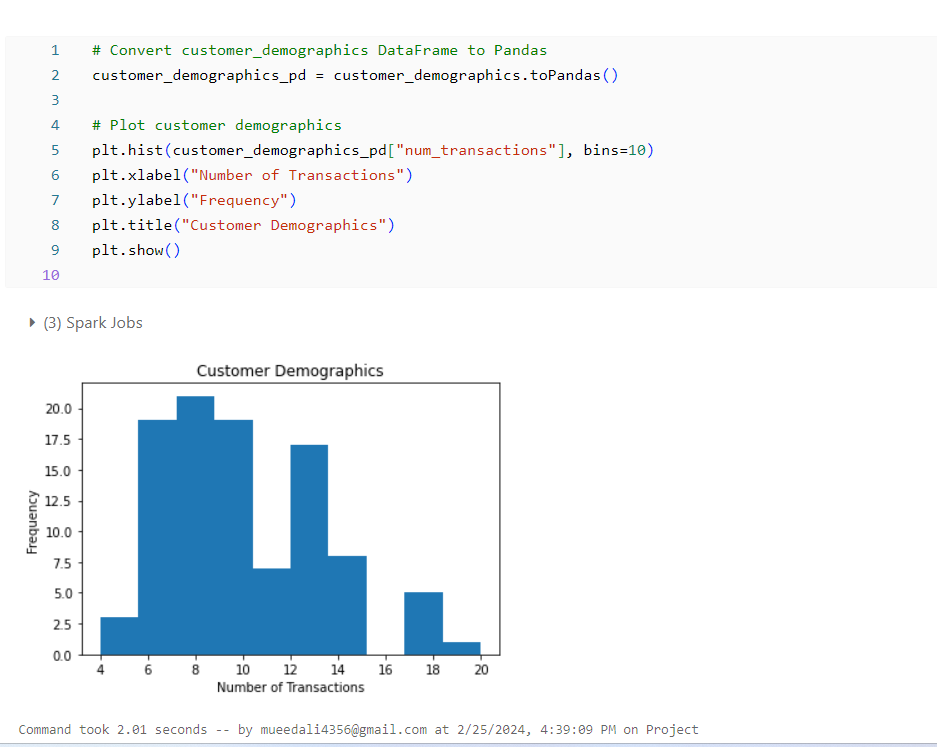
Visualize Top Selling Products:



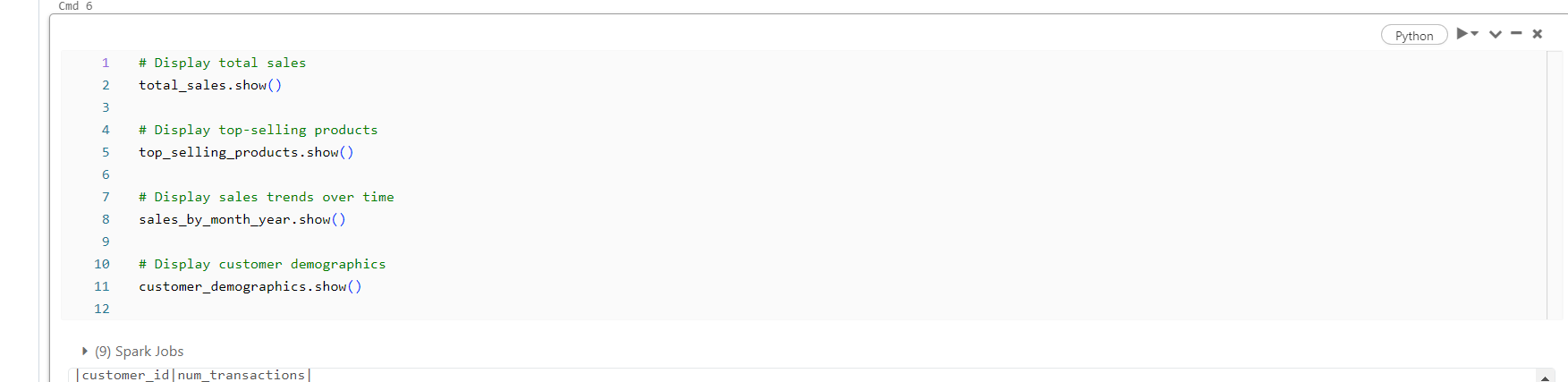
Visualize Sales trends overtime

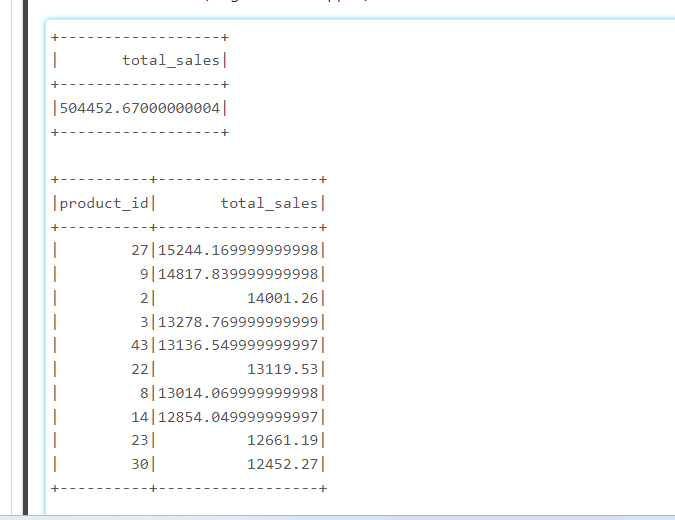
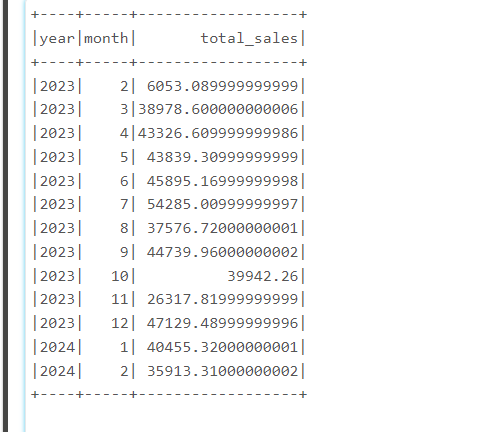
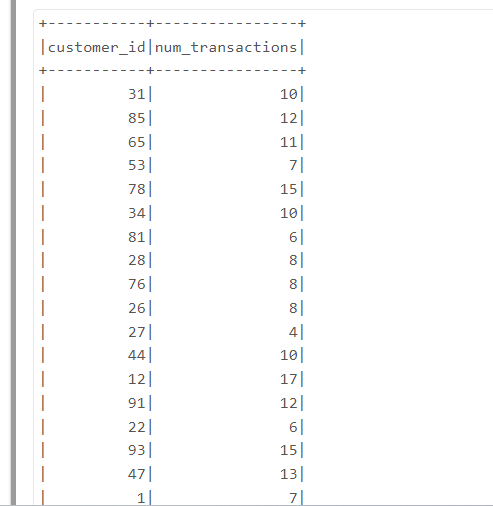


Visualize Customer Demographics:



**RESULTS**

****

**CONCLUSION**

In conclusion, the retail sales analysis project using Apache Spark provides valuable insights into the performance and trends of a retail business. Through careful data exploration, preprocessing, and analysis, we've gained a deeper understanding of various aspects such as total sales, top-selling products, sales trends over time, and customer demographics.

The analysis revealed important findings, including:

1. Total Sales: The total sales amount gives us a holistic view of the business's revenue, helping us track overall performance and set targets for future growth.
2. Top-Selling Products: Identifying the top-selling products allows us to focus resources on high-performing items, optimize inventory management, and capitalize on popular trends.
3. Sales Trends Over Time: Analyzing sales trends over time helps us identify seasonal patterns, anticipate fluctuations in demand, and plan marketing campaigns and promotions accordingly.
4. Customer Demographics: Understanding customer demographics, such as the number of transactions per customer, enables us to tailor marketing strategies, improve customer engagement, and enhance the overall shopping experience.

By leveraging Apache Spark's powerful data processing and analysis capabilities, we've equipped stakeholders with actionable insights to make informed decisions and drive business growth. Moving forward, continued analysis and refinement of strategies based on these insights will be essential for sustaining success and remaining competitive in the dynamic retail landscape.