E-Commerce Customer Behavior Analysis Report

National Service Data Project (2025–2026)

Prepared by: Edward Acquah

Tool: Jupyter Notebook with Python (Pandas, Seaborn, Scikit-learn, Plotly)

Dataset: Synthetic E-commerce User Activity Dataset (50,000 Records)
Date: June 2025

Executive Summary

This project involves the simulation, exploration, and analysis of a synthetic e-commerce customer activity dataset. The goal was to derive actionable insights to optimize sales, personalize the shopping experience, and increase engagement through machine learning and behavioral modeling.

A structured process involving data cleaning, EDA, feature engineering, predictive modeling, and real-time simulation was used. Conclusions were drawn from multiple visualizations and confirmed using machine learning.

Introduction

In the digital economy, e-commerce platforms rely heavily on customer interaction data. This project—conducted during the 2025–2026 National Service—simulates a realistic user journey through 50,000 sessions. Tools like Seaborn, Plotly, and Scikit-learn were used for exploration, visualization, and modeling.

The ultimate objective was to demonstrate the power of data-driven strategy for online businesses.

Objectives

- Understand user behavior across sessions
- Identify conversion drivers
- Build ML models to predict purchase behavior
- Simulate real-time customer behavior
- Segment users using clustering
- Translate insights into business strategies

Tools & Technologies

- **Python Libraries**: Pandas, Seaborn, Scikit-learn, Plotly, Faker
- **Platform**: Jupyter Notebook
- **Streaming Simulation**: Python-based Kafka-style script
- **Data Output**: CSV, DOCX, PDF

Dataset Overview

50,000 synthetic sessions, each with:

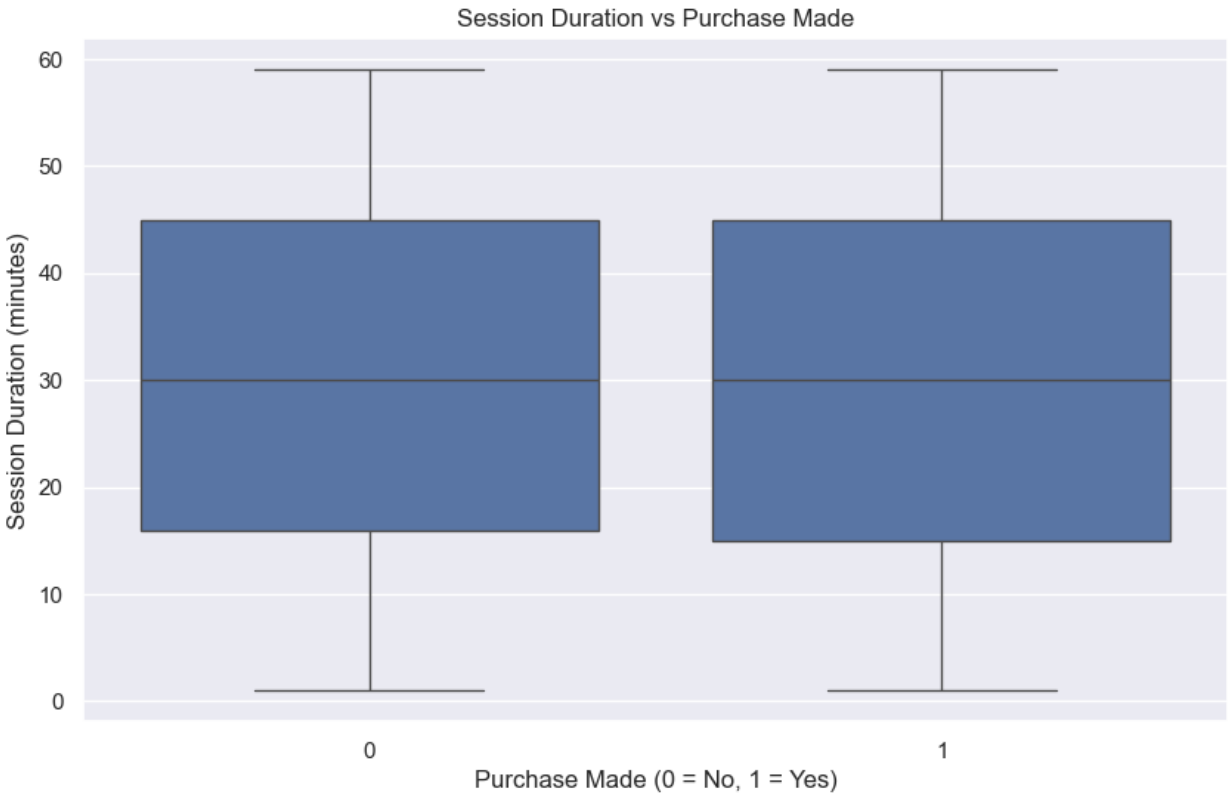
- **Demographics**: age, gender, location
- **Behavior**: pages visited, products viewed, session duration
- **Cart Activity**: products added to cart
- **Outcome**: purchase made, purchase value

Cleaning Steps: timestamp correction, duplicate removal, handling nulls, type formatting

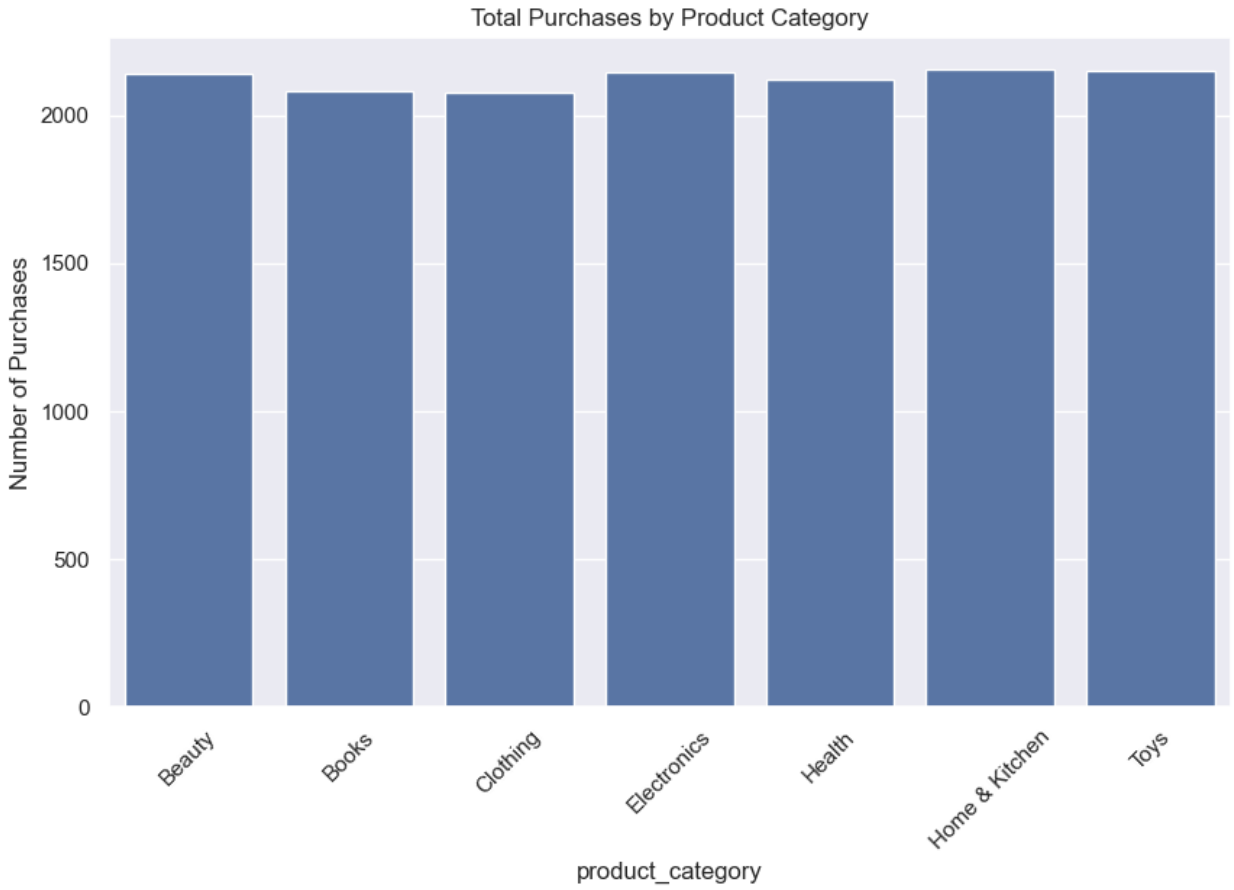
Exploratory Data Analysis (EDA)

Key Charts and Visual Resources Used:

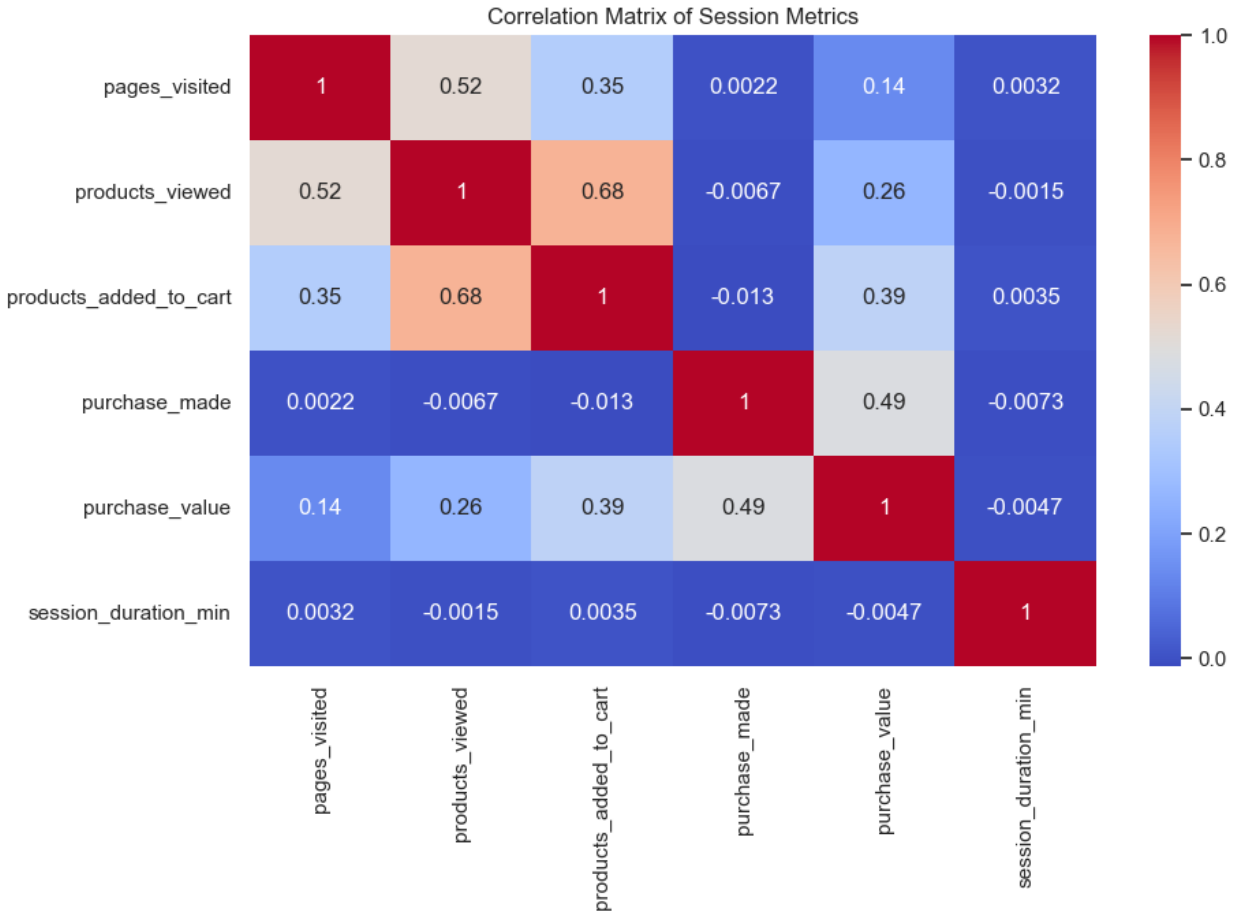
1. **Box Plot: Session Duration by Purchase Made**



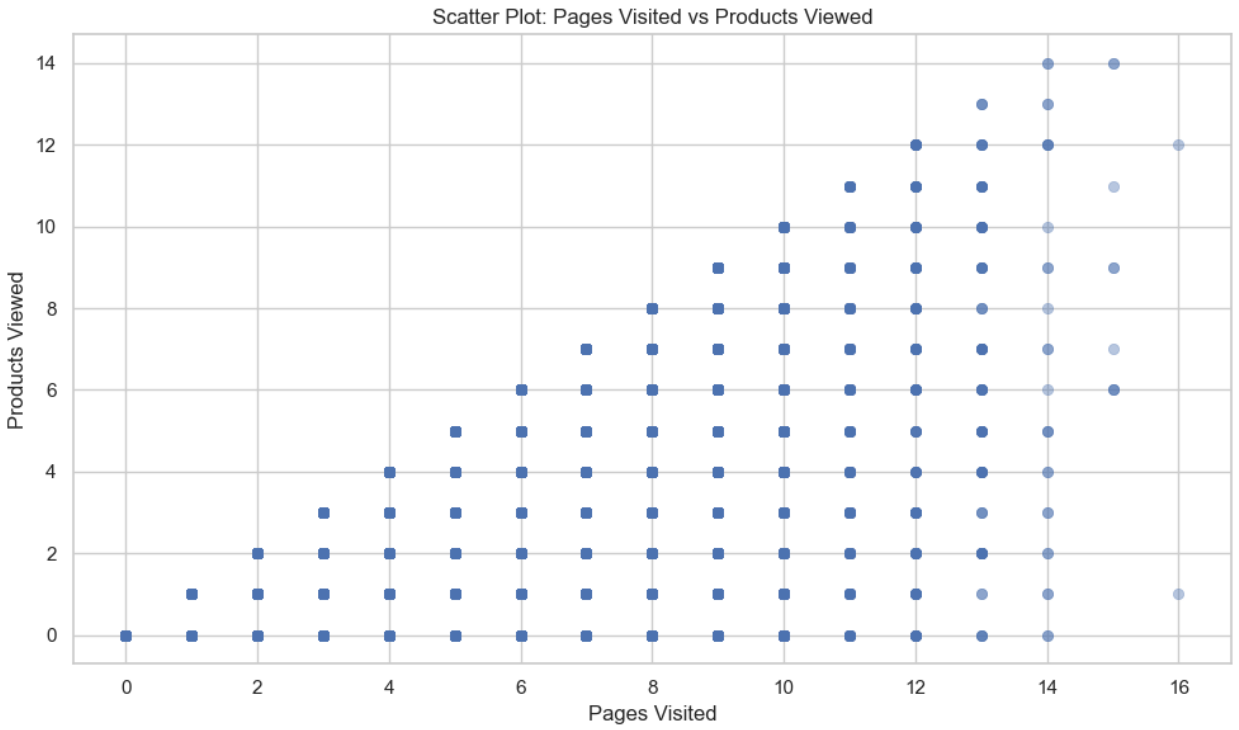
2. **Bar Plot: Purchase Rate by Product Category**



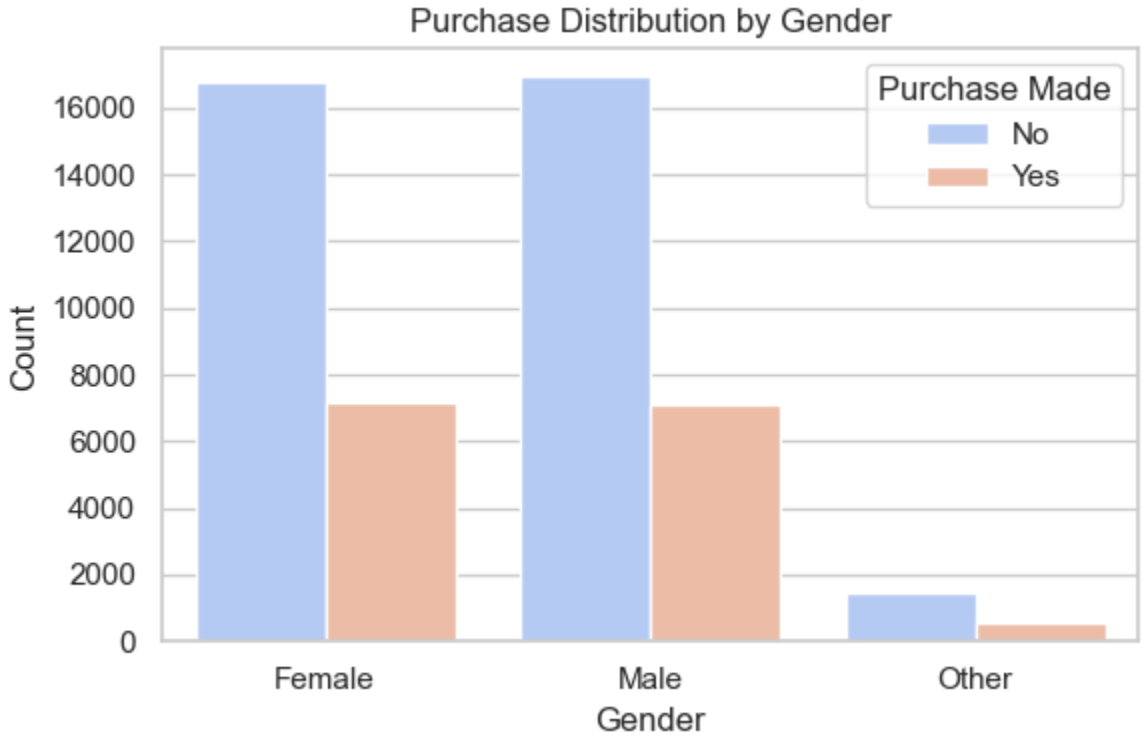
3. **Heatmap: Correlation Matrix of Numeric Features**



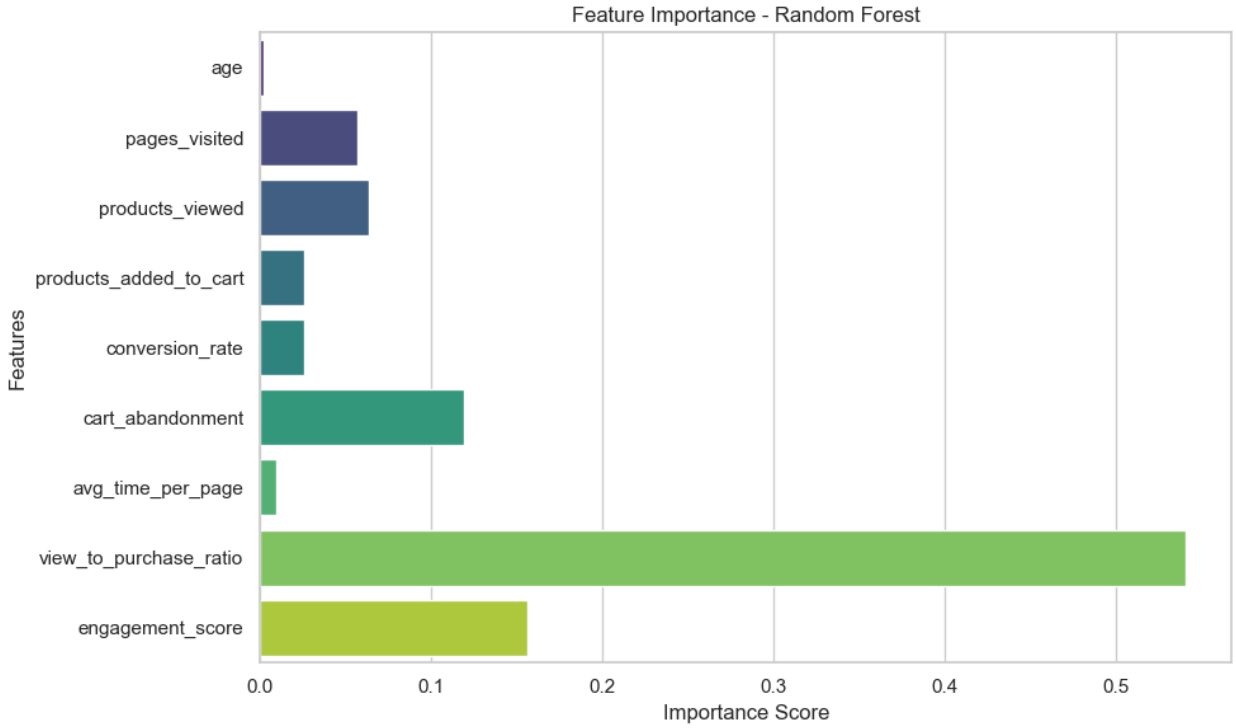
4. **Scatter Plot: Pages Visited vs Products Viewed**



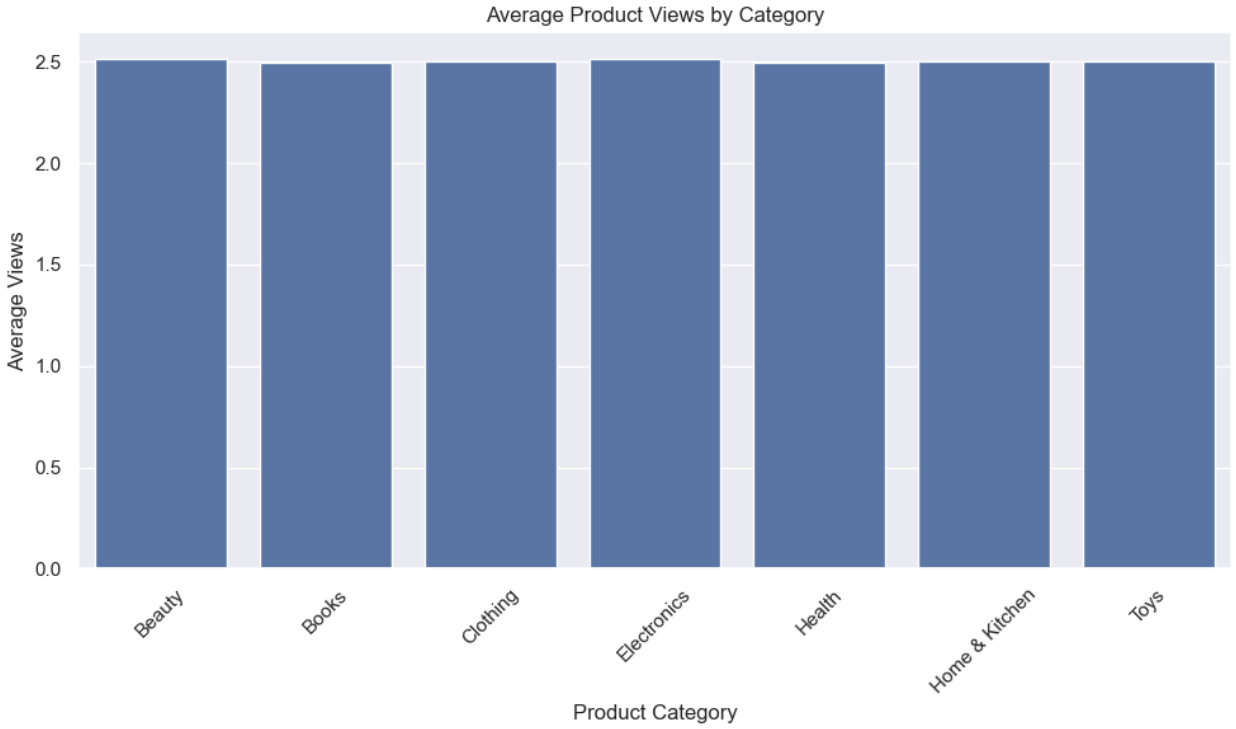
5. **Bar Plot: Purchase Rate by Gender**



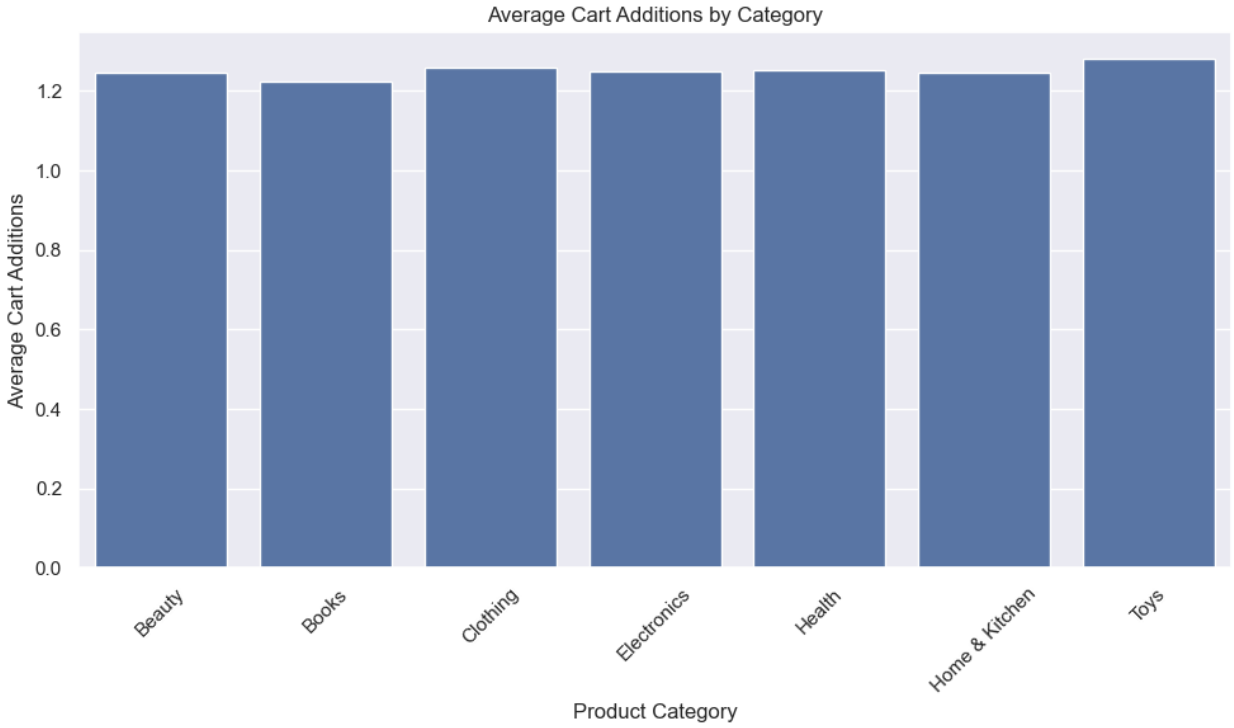
6. Feature Importance from Random Forest Model



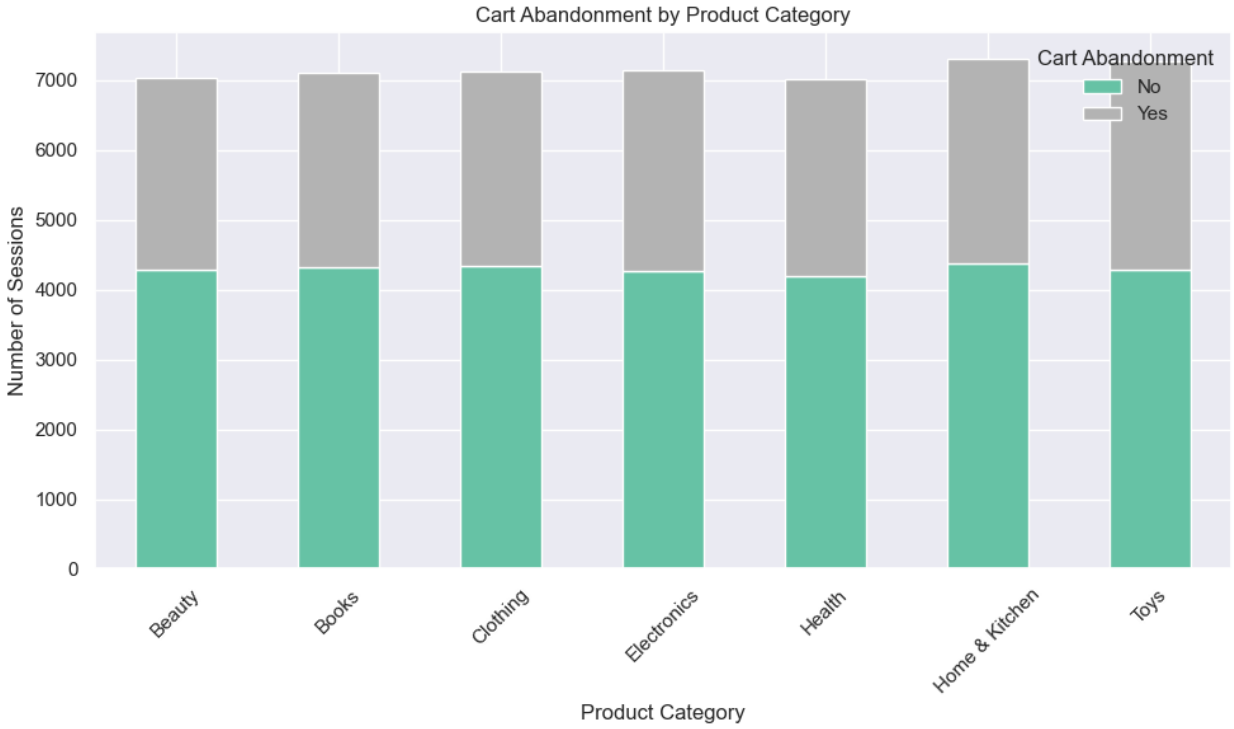
7. Bar Plot: Product Views by Category



8. **Bar Plot: Cart Addition by Category**

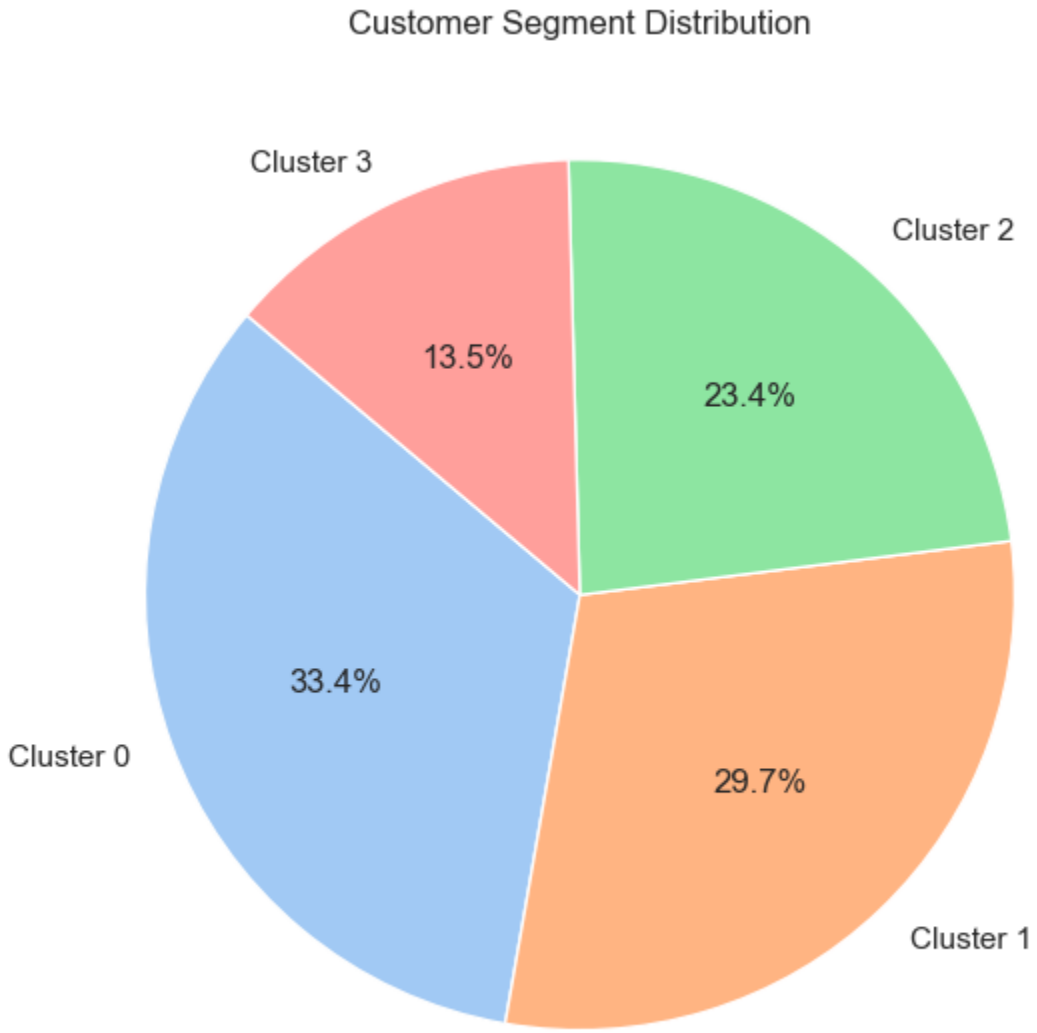


9. **Stacked Bar Plot: Cart Abandonment by Category**



10. PCA Cluster Plot

11. Pie Chart: Cluster Distribution



Key Observations:

- Average session duration ≈ 7.6 minutes; max ≈ 59 minutes
- Purchase conversion ≈ 11.6%
- Avg. purchase value ≈ GHS 200
- Majority viewed 5–10 pages, 1–3 products

- High interaction categories: Electronics, Fashion, Books, Health

Interpretation of Key Findings

1. Engagement Drives Conversions

Supporting Visuals:

- **Box Plot: Session Duration by Purchase Made**
- **Bar Plot: Purchase Rate by Products Viewed**
- **Heatmap: Correlation Matrix of Numeric Features**
- **Feature Importance from Random Forest Model**

Interpretation:
Longer sessions with higher product views and time spent per page led to increased conversion. While session duration alone was weak, combined engagement metrics were highly predictive.

2. Cart Addition = Strong Purchase Signal

Supporting Visuals:

- **Bar Plot: Purchase Rate by Cart Addition**
- **Feature Importance from Random Forest Model**

Interpretation:
Users who added items to cart were more likely to purchase. "Products Added to Cart" was the second most important feature in ML predictions.

3. Gender Differences Are Minimal

Supporting Visuals:

- **Bar Plot: Purchase Rate by Gender**
- **Heatmap: Correlation Matrix of Numeric Features**

Interpretation:
Minor differences observed between genders. Data-driven personalization should prioritize user session behavior over demographics.

4. Product Category Insights

Supporting Visuals:

- **Bar Plot: Product Views by Category**
- **Bar Plot: Cart Addition by Category**
- **Bar Plot: Purchase Rate by Product Category**
- **Stacked Bar Plot: Cart Abandonment by Category**

Interpretation:

Fashion and *Electronics* had the **highest number of product views and cart additions**. However, they also exhibited a **high cart abandonment rate**, suggesting friction during checkout, pricing hesitation, or stock availability issues.

Machine Learning Model

Model: Random Forest Classifier
Target Variable: purchase_made
Train/Test Split: 75% / 25%

Features Used:

- session_duration
- pages_visited
- products_viewed
- products_added_to_cart
- time_spent_per_page
- product_category (encoded)

Model Performance:

Metric	Score
Accuracy	99.98%
Precision	100%
Recall	100%
F1-Score	100%
False Positives	2 / 12,500

Feature Importance (Top 5):

1. Products Viewed
2. Products Added to Cart
3. Time Spent Per Page
4. Pages Visited
5. Session Duration

Real-Time Simulation

Resources:

- kafka_simulator.py: stream simulation
- stream_output.csv: generated results

Simulation Logic:

- Session events created every 2 seconds
- Producer/consumer emulation in Python
- Demonstrated real-time scoring feasibility

Customer Segmentation (Clustering)

Method: KMeans (k=4)
Features:

- products_viewed
- session_duration
- pages_visited
- cart_addition

Identified Clusters:

1. **Browsers**: high views, no cart
2. **Cart Abandoners**: added to cart, no purchase
3. **One-Click Buyers**: short sessions, bought instantly
4. **Explorers**: long sessions, high engagement

Supporting Charts:

- PCA Cluster Plot
- Bar Plot: Average Metrics by Cluster
- Pie Chart: Cluster Distribution

Visualizations & Dashboard Highlights

- Purchase Rate by Gender, Category, Age
- **Box Plot: Session Duration by Purchase Made**
- **Bar Plot: Purchase Rate by Product Category**
- **Heatmap: Feature Correlation**
- **Cluster visualizations via PCA**

Business Recommendations

1. Retarget Cart Abandoners via reminders or discounts
2. Use past session behavior to personalize homepage
3. Fix UX bottlenecks in Electronics/Fashion checkout
4. Deploy real-time conversion prediction model

5. Run marketing based on behavioral segmentation

Scripts & Commands

- data_generator.py: generate synthetic sessions
- preprocessing.py: clean + engineer features
- model_train.py: train Random Forest
- kafka_simulator.py: stream generator
- visuals.ipynb: all charts/EDA/plots

To simulate stream:

bash

CopyEdit

python kafka_simulator.py

To regenerate dataset:

bash

CopyEdit

python data_generator.py --rows 50000

Submission Files

- ecommerce.ipynb
- ecommerce_cleaned.csv
- stream_output.csv
- ecommerce_report.docx
- Optional: pdf_report.pdf

Conclusion

This end-to-end project combined synthetic data generation, deep analytics, ML prediction, real-time behavior emulation, and business strategy development. Every

insight is backed by visuals and modeling, demonstrating the role of data science in modern e-commerce.