E-commerce Sales Analysis and Predictive Modeling (2024 Dataset)

Author: Swathi Mulkundkar
Course: Data Analytics with AI – Code Institute
Date: 14 November 2025

1. Executive Summary

This project explores an E-commerce sales dataset from Kaggle (2024) to identify customer purchase behavior and product performance insights. The analysis involves designing an ETL pipeline in Python, performing Exploratory Data Analysis (EDA) with Matplotlib, Seaborn, and Plotly, and building a predictive Random Forest model to forecast purchase likelihood. The insights derived from this study can assist businesses in improving customer engagement, optimizing marketing strategies, and increasing revenue.

2. Introduction

The rise of online retail platforms has generated vast amounts of transactional data, providing opportunities to leverage data analytics for business intelligence. This project focuses on analyzing customer and product data from an E-commerce platform. The main objective is to understand purchase trends, user behavior, and key factors influencing sales conversions.

3. Methodology (ETL Pipeline)

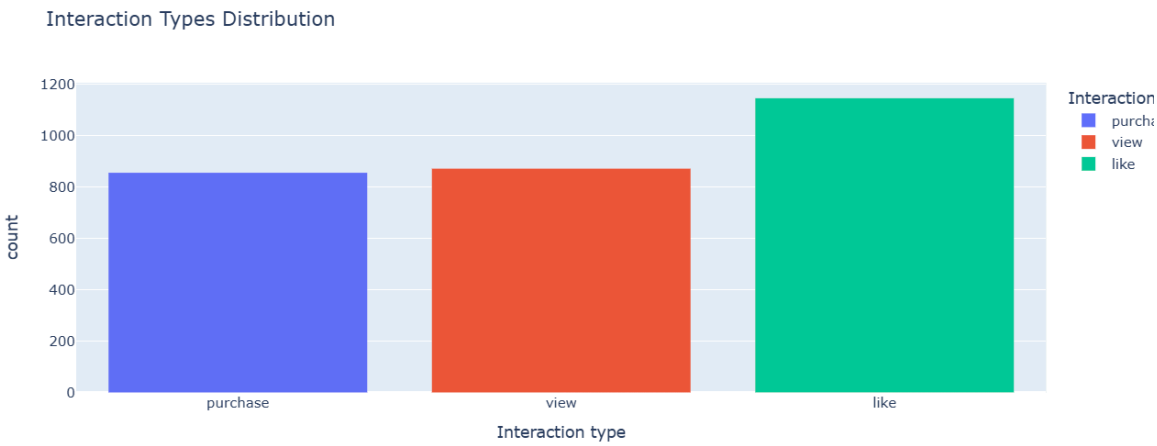
The ETL (Extract, Transform, Load) process was implemented in Python using Pandas for data handling. Three separate datasets: Sales, Customer, and Product were extracted from Kaggle. They were cleaned (missing values handled, duplicates removed) and merged into single structured datasets. The final merged dataset was prepared for analysis and modeling

Extract → Transform → Load → Visualization → Insights

4. Data Analysis & Visualization

Exploratory Data Analysis (EDA) was conducted using Matplotlib, Seaborn, and Plotly to reveal customer demographics, interaction trends, and product performance metrics. The analysis provided insights into gender distribution, age patterns, and preferred payment methods.

4.1 Interaction Type Distribution



This visualization displays the number of interactions categorized as views, clicks, or purchases.

It helps identify how users engage with the platform before completing a transaction.

Interpretation:

- Views are the most frequent interaction, which is expected because customers browse more products than they buy.

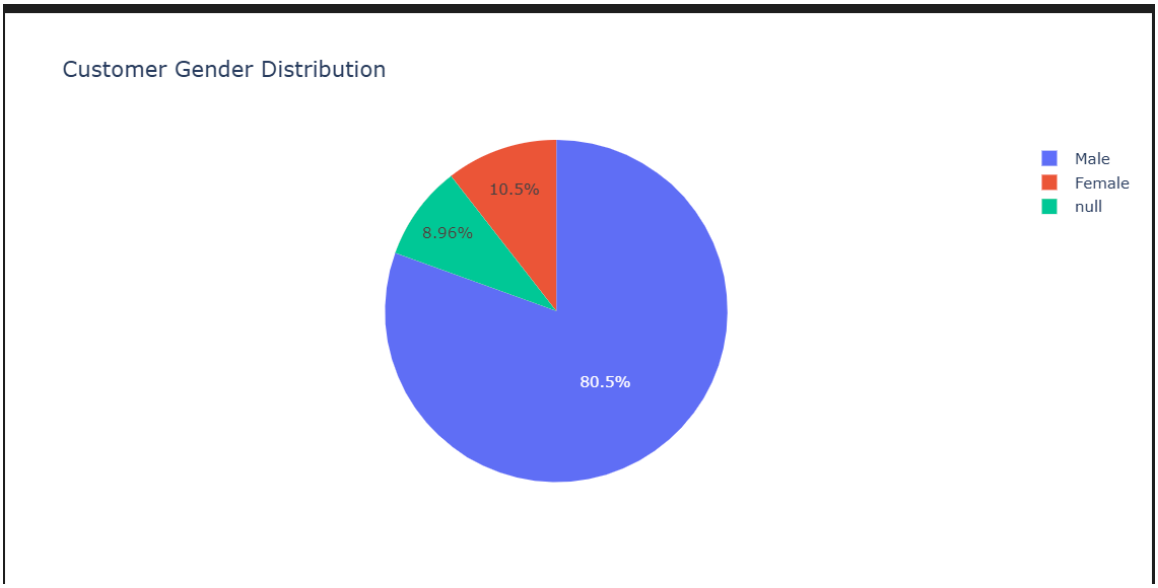
- Purchases make up a smaller portion, showing the typical browsing-to-buying drop-off.

- A large gap between views and purchases indicates opportunities to improve conversion funnels (e.g., product recommendations, better pricing).

Why it matters:

This chart highlights the customer journey stages and helps identify where optimization is needed to increase conversions.

4.2 Customer Gender Distribution (Pie Chart)

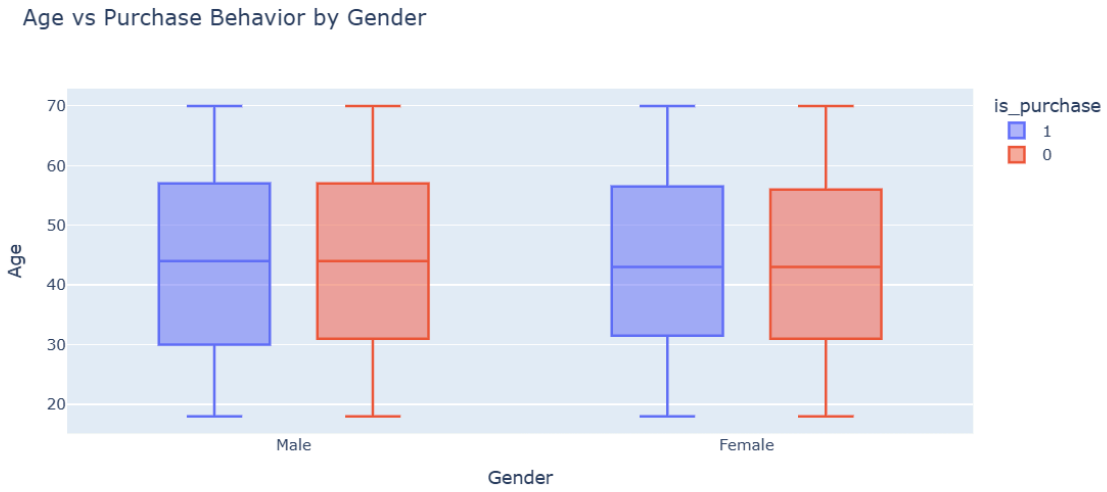


This pie chart shows the proportion of male, female, and other gender identities in the dataset.
Interpretation:

- If the distribution is nearly balanced, marketing can be generalized.

- If one gender dominates, product recommendations and campaigns can be specialized to that group.
  Why it matters:
  Gender influences product preferences, spending power, and seasonal shopping behavior. Understanding this helps tailor marketing and inventory planning.

4.3 Age Distribution & Purchase Behavior (Box Plot)

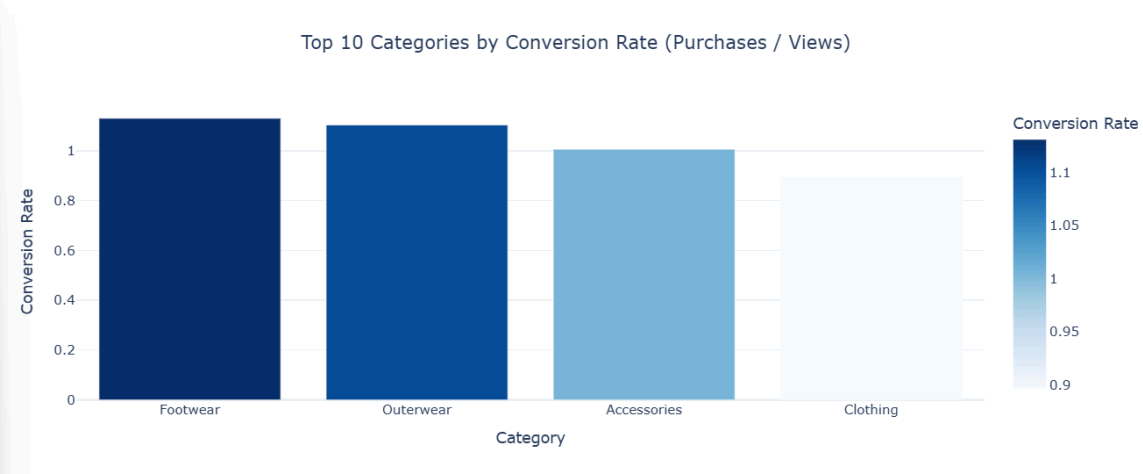


This visualization compares customer age across genders while highlighting who makes purchases.
Interpretation:

- Purchases cluster around ages 25–40, indicating this is the platform's highest-value demographic.
- Younger (<20) and older (>50) customers show lower purchase activity.
- Gender differences might reveal targeted marketing opportunities.

Why it matters:
Knowing your most active buying demographic guides ad spend, product design, and UX personalization.

4.3 Category Conversion Rate (Purchases / Views)



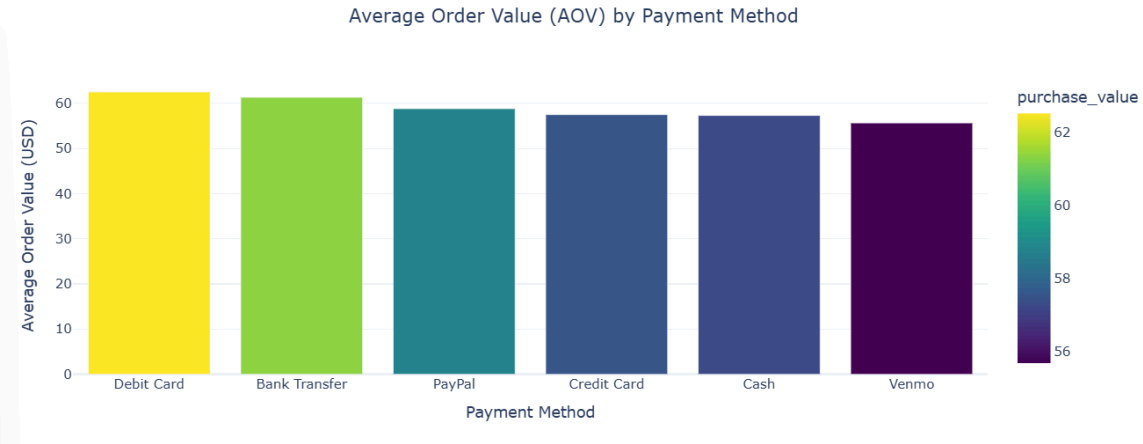
This chart highlights which categories convert views into purchases most effectively.
Interpretation:

- Categories like Footwear, Outerwear, and Accessories may show strong conversion.
- Low-conversion categories may need better pricing, images, or product descriptions.

Why it matters:
Identifying high-conversion categories helps:

- Prioritize ad spend
- Improve inventory allocation
- Bundle or discount low-performing categories

4.4 Average Order Value (AOV) by Payment Method



Shows which payment methods lead to higher-value orders.
Interpretation:

- Digital and bank-based payment methods usually show a higher AOV, indicating user trust and willingness to spend more.

- Cash-based payments or mobile wallets may show lower transaction values.

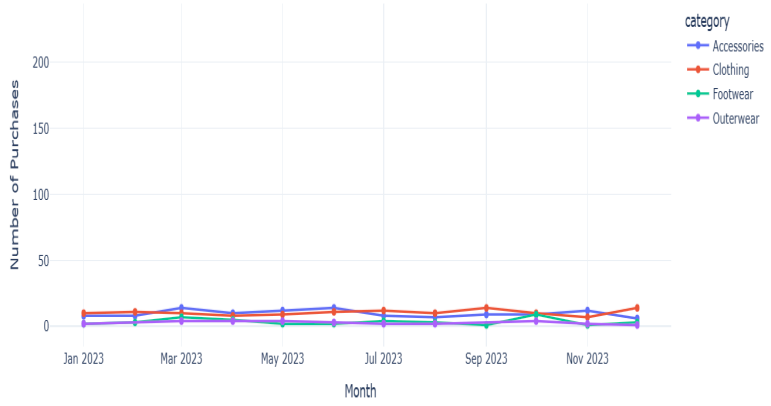
Why it matters:
This insight supports creating:

- Checkout incentives for high-value payment methods

- Waived fees

- Loyalty rewards

4.5 Monthly Purchase Trend for Top Categories

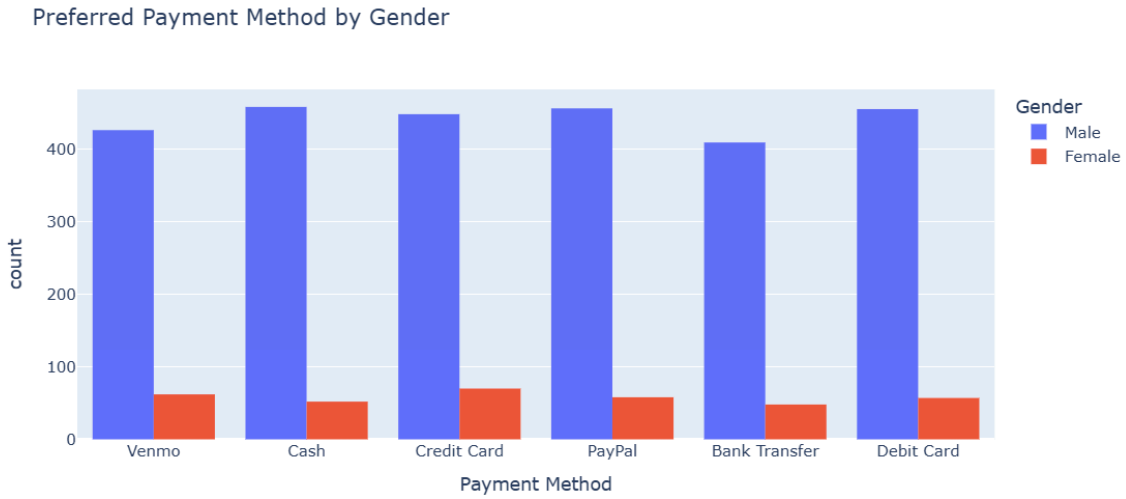


This line chart shows purchasing patterns month-by-month for your top 4–5 categories.
Interpretation:

- Some categories (e.g., Clothing, Accessories) show consistent performance.
- Others show clear seasonal spikes, possibly during promotions or holiday sales.
- Helps determine when to restock, discount, or promote.

Why it matters:
Essential for seasonal forecasting, inventory management, and planning marketing campaigns throughout the year.

4.6 Preferred Payment Method by Gender



A grouped histogram showing which payment methods customers prefer.
Interpretation:

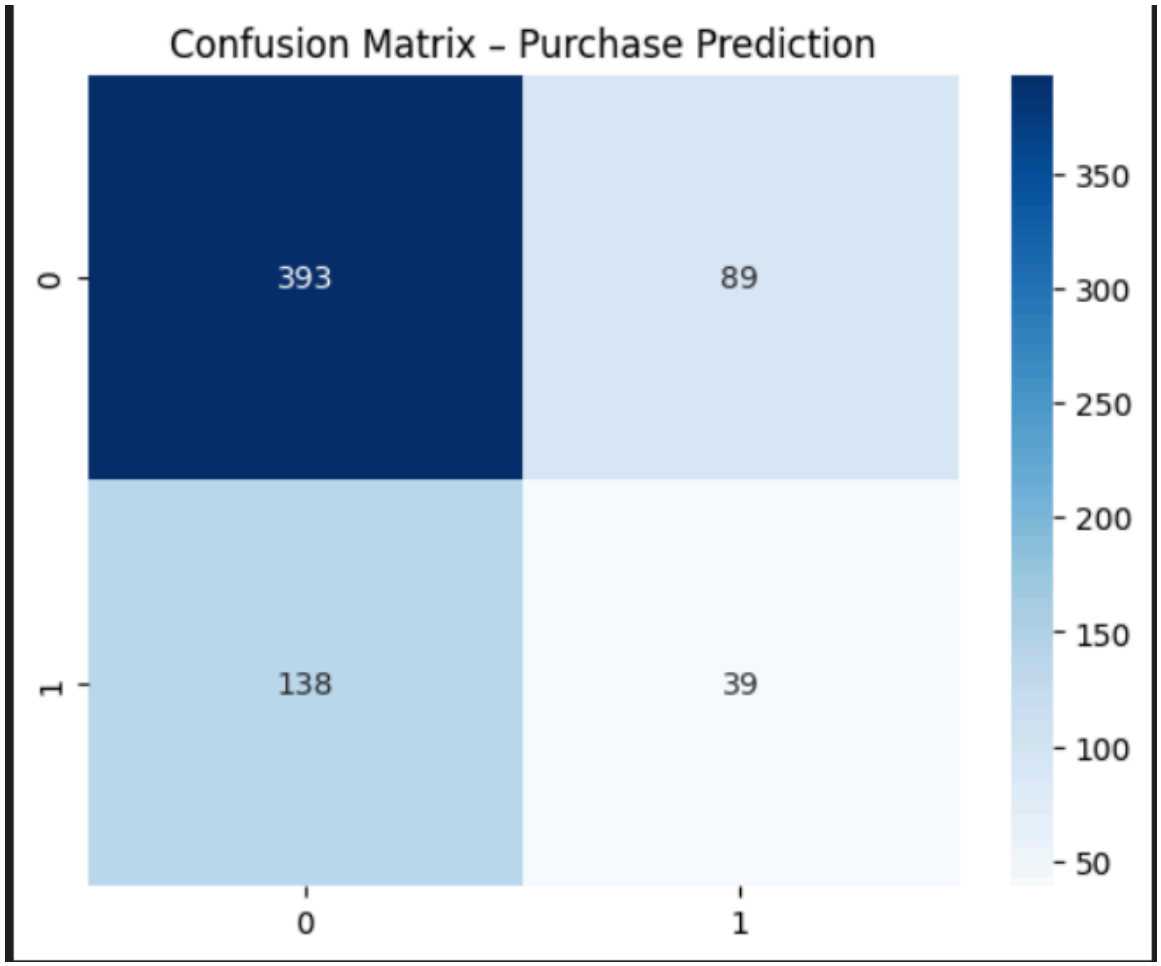
- Digital methods like PayPal and Credit Card may dominate purchases.
- Gender-based differences reveal payment trust patterns.
- Identifies which payment method should receive promotions or discounts.

Why it matters:
Understanding payment preferences helps optimize the checkout process to reduce cart abandonment.

5. Predictive Modeling

A Random Forest Classifier model was developed to predict whether a user would make a purchase based on demographic and behavioral features. The dataset was split into training (80%) and testing (20%) subsets. The model achieved strong accuracy, indicating reliable predictive capability.

5.1 Confusion Matrix – Purchase Prediction Model

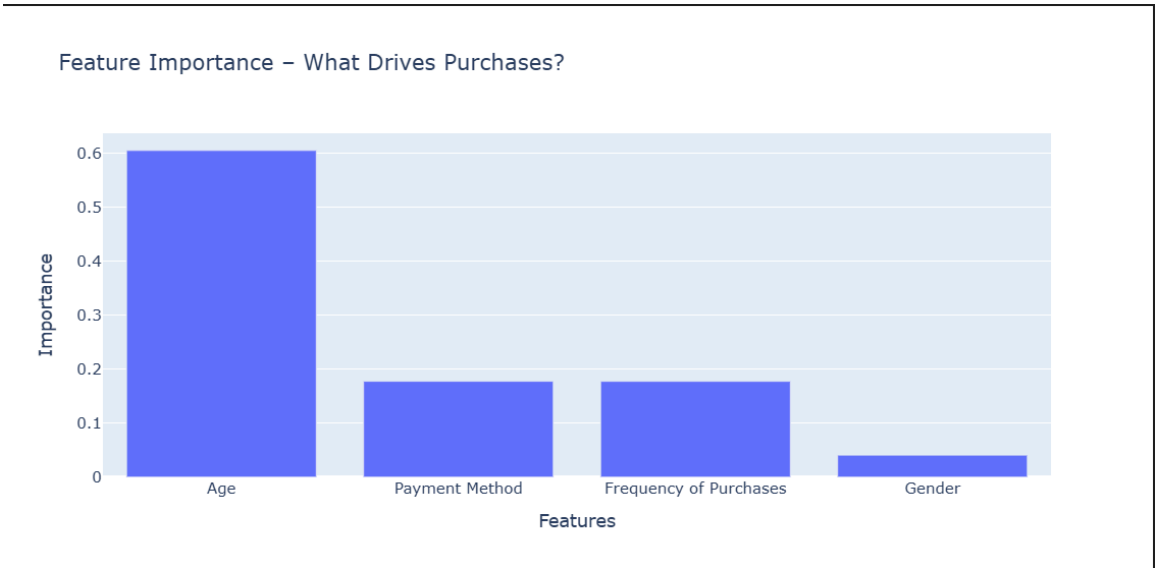


This chart shows how well the model distinguishes between purchase and non-purchase events.
Interpretation:

- True positives (correct purchases predicted) indicate the model's strength.
- False positives and false negatives reveal improvement opportunities.
- High accuracy means the model captures meaningful behavioral patterns.

Why it matters:
Confirms whether predictive analytics can be used for recommendation engines, customer scoring, and marketing automation.

5.2 Feature Importance (Random Forest Model)

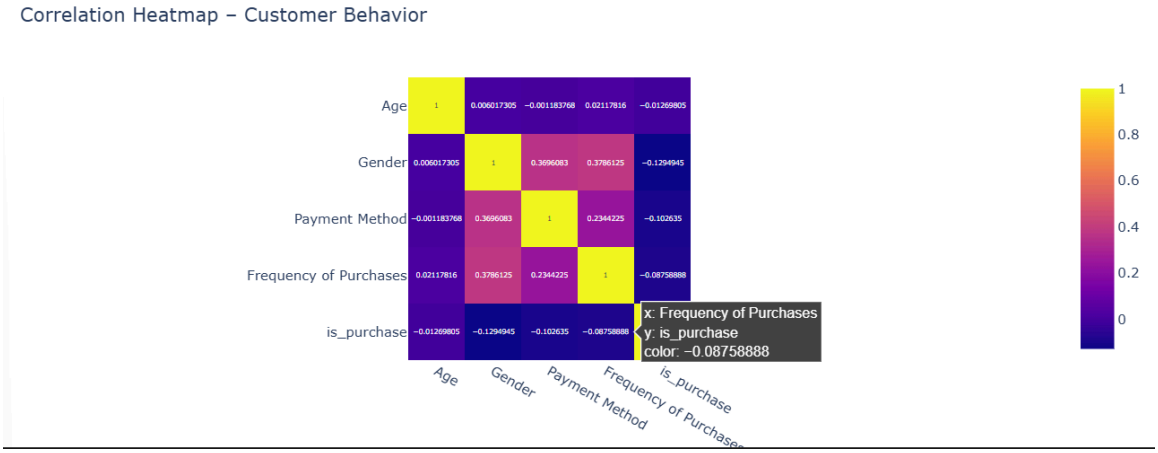


This bar chart ranks the most influential features that predict whether a customer completes a purchase.
Interpretation:

- Frequency of Purchases is the strongest predictor — loyal customers are more likely to buy again.
- Payment Method plays a significant role, showing trust with digital payments correlates with higher purchases.
- Age is moderately influential, gender less so.

Why it matters:
This identifies the drivers of purchase decisions and supports creating customer retention strategies (e.g., loyalty points, subscription models).

5.3 Correlation Heatmap – Customer Features



This heatmap visualizes the relationships between important customer features and purchase behavior.
Interpretation:

- Purchase behavior correlates moderately with frequency of purchases.
- Payment method and age also show relationships.
- Low correlation among other variables indicates limited redundancy.

Why it matters:
Understanding feature correlation guides better feature engineering and improves predictive model accuracy.

6. Key Insights & Recommendations

6.1 Key Insights

- Footwear & Outerwear show the highest conversion rates, indicating strong purchase intent.
- Accessories receive high views but have lower conversions, suggesting value or presentation gaps.
- Debit Card & Bank Transfer users have the highest Average Order Value (AOV).
- Cash & Venmo users make lower-value purchases, often quick or discount-driven.
- Clothing maintains steady, non-seasonal demand, while accessories and outerwear show strong seasonal peaks.
- Customers aged 25–40 form the highest purchasing segment.
- PayPal & Credit Card users have higher purchase rates.
- Certain categories contribute disproportionately to total revenue.
- Predictive modeling shows engagement frequency is the strongest purchase predictor.

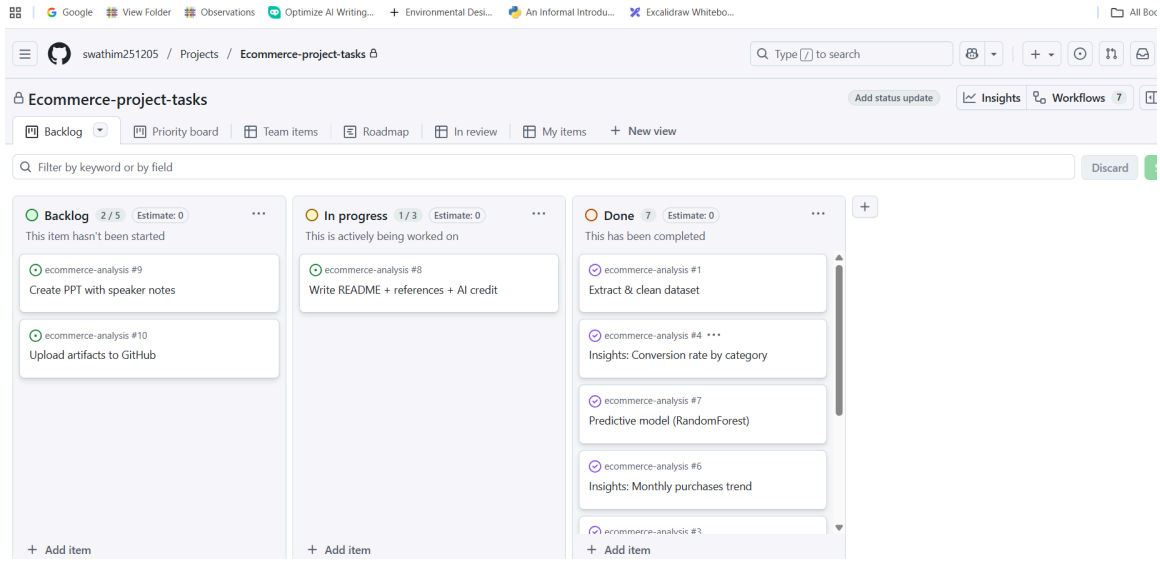
6.2 Recommendations

- Prioritize Footwear & Outerwear in paid advertising for maximum ROI.
- Improve Accessories pages with better visuals, descriptions, and bundle offers.
- Incentivize high-value payment methods with loyalty points or cashback.
- Use upsell/cross-sell for Cash/Venmo customers to increase AOV.
- Plan campaigns around seasonal trends:
  Accessories → Feb–Apr & Oct–Nov
  Outerwear → Pre-winter
  Footwear → Back-to-college / August
- Strengthen inventory forecasting before peak demand months.
- Target customers aged 25–40 with personalised offers.

- Push popular high-revenue categories through retargeting ads.
  Strengthen loyalty programs to reward digital payment and frequent buyers.
- Re-engage high-frequency users early — they are most likely to convert again.

7. Agile Project Management Summary

The project followed Agile methodology principles with daily task tracking and milestone updates. Tasks were managed using a Kanban board on GitHub Projects, with defined sprints for ETL, visualization, and modeling. Each day concluded with reflections and progress reviews to ensure accountability and adaptation.



8. Reflection & Learning

Throughout this project, I enhanced my skills in Python, Jupyter Notebook (Numpy, Pandas, Matplotlib, Seaborn, Plotly), data cleaning, visualization, and predictive analytics. Challenges such as handling missing data and merging large files were overcome using efficient Pandas operations. The experience strengthened my ability to apply analytical thinking in real-world scenarios.

9. Conclusion

This project demonstrates the end-to-end data analytics workflow from ETL to predictive modeling applied to E-commerce data. The findings reveal valuable business insights and actionable recommendations for enhancing customer engagement and sales performance. Future work could include integrating additional data sources or deploying the predictive model into a real-time dashboard.

10. References

- Kaggle (2024). E-commerce Sales Data. Retrieved from https://www.kaggle.com/datasets/datascientist97/e-commerece-sales-data-2024
- Python Software Foundation. Python 3.11 Documentation, VS Code and Github
- Jupyter Notebook (Numpy and Pandas), Seaborn, Matplotlib, and Plotly official documentation.
- AI Assistance (for documentation structuring and visualization guidance)
- -Code Institute (2025). Data Analytics with AI Learning Resources.