Sales Data: Analyzing & Predicting

By: Amanda Jo Griffin (Aj)

Background

Project Overview: Junior Data Analyst

Data Processing:

- Cleaned and organized sales and customer data for analysis.
- Exploratory Data Analysis (EDA):
- Utilized statistical techniques to uncover patterns and trends.
- Employed visualization tools for data exploration.

Modeling:

- Developed and evaluated predictive models for insights.
- Iteratively refined models for optimal performance.

Streamlit App Development:

- Created user-friendly interface for showcasing analysis.
- Integrated interactive visualizations for enhanced understanding.

Findings Utilization:

- Leveraged insights to inform strategic decision-making.
- Identified factors influencing sales performance and customer behavior.

Purpose

Project Purpose & Ideal Outcomes:

Purpose:

- Drive business growth through data-driven insights.
- Optimize sales strategies and enhance customer experiences.

Why:

- In a dynamic e-commerce environment, tasked with analyzing sales and customer data.
- Uncover actionable insights to refine strategies and improve operations.
- Internal teams seek predictive models to forecast future sales based on demographic factors, such as gender.

Ideal Outcomes:

- Increased revenue and market share through optimized sales strategies.
- Enhanced customer satisfaction and loyalty.
- Empower stakeholders with data-driven decision-making capabilities.
- Streamlined operations for improved efficiency and profitability.

Data Overview

9

Columns

1000

Rows

Data On Company's Customers & Sales

About The Data

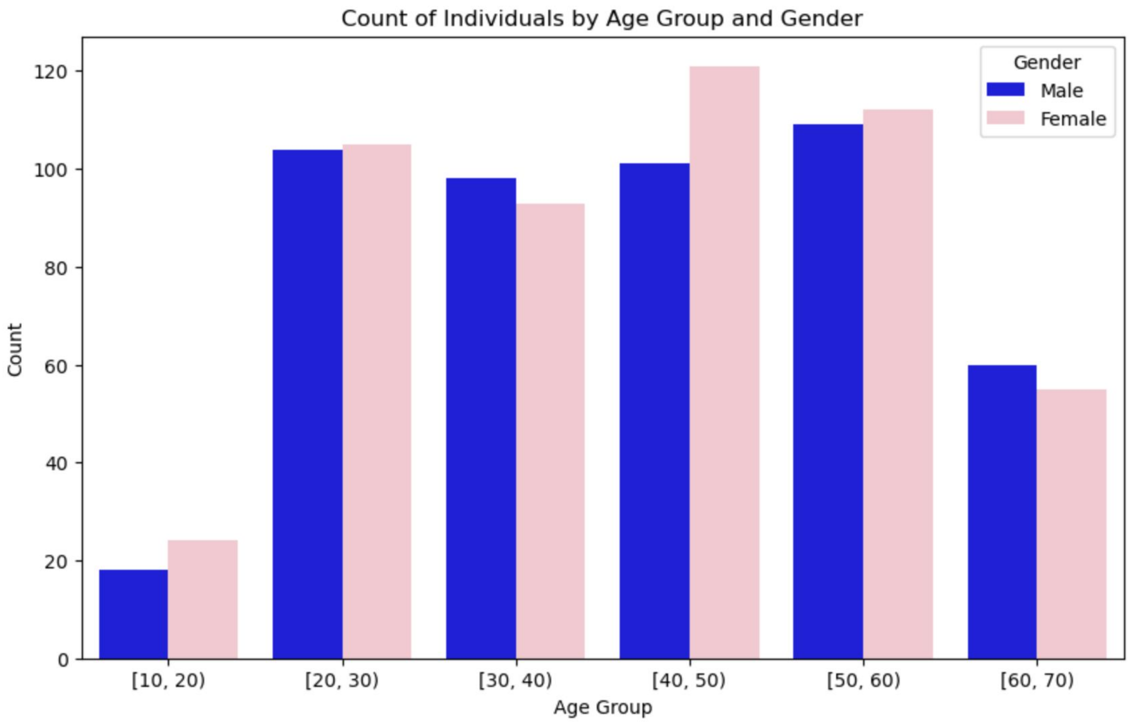
1

Current Trends

Count Of Buyers By Age Group

Purchasing patterns show a shift with age: males dominate in the 20-30 age group, while females lead in the 40-50 age bracket, suggesting targeted marketing opportunities.

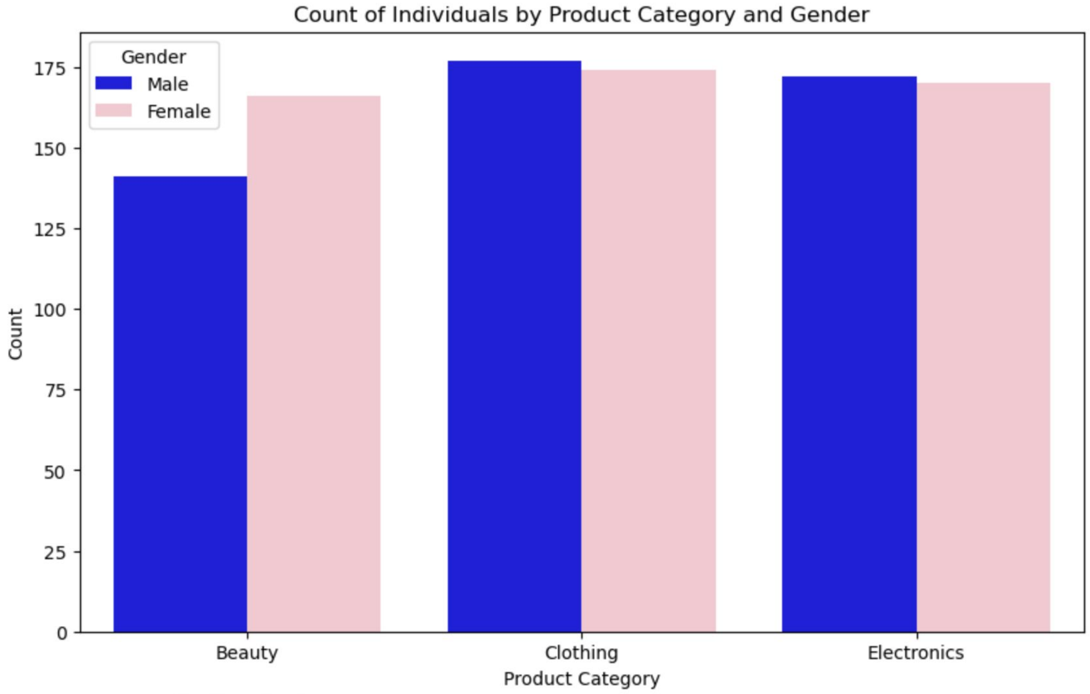
Across all age groups, females consistently represent slightly higher counts in retail purchases. Recognizing age-specific trends in gender-based buying behavior can enhance predictive models, potentially improving gender prediction accuracy and tapping into predictable gender preferences.



Count Of Buyers By Category & Gender

Beauty products see a clear gender disparity, with females leading in purchases, suggesting a potential focus for targeted marketing. Clothing purchases show balanced engagement between genders, indicating a less pronounced predictive power for gender based on this category alone.

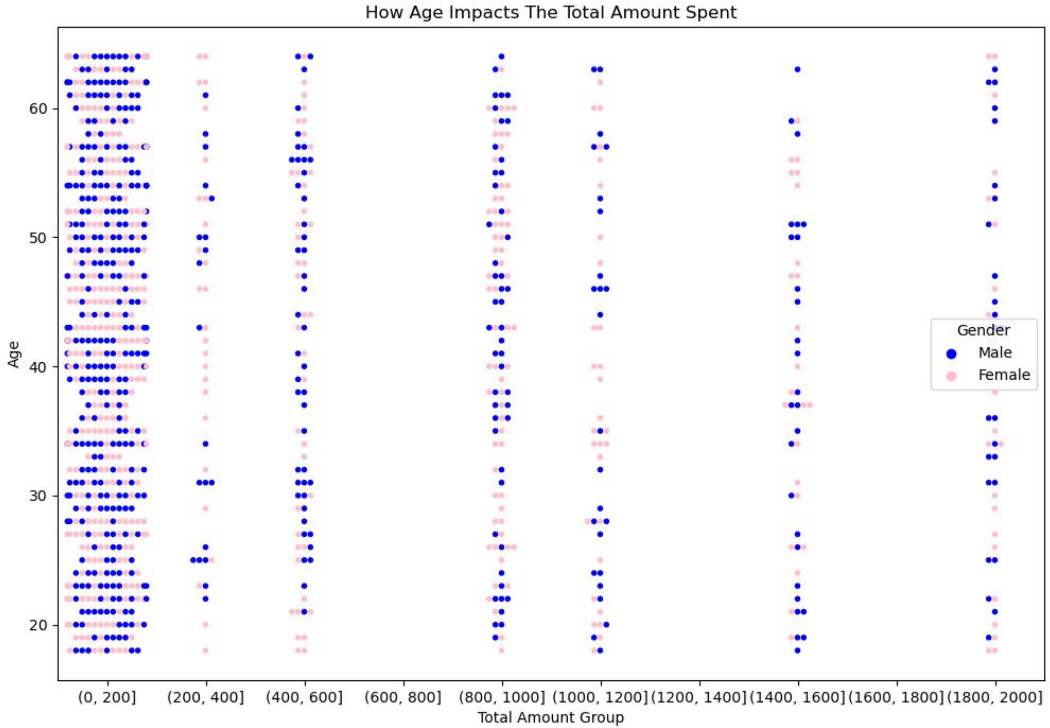
Electronics purchases exhibit comparable patterns between males and females, implying that this category might not be as informative for gender prediction. Other factors, such as age or specific product choices, may be more influential in accurate predictions.



How Age Impacts Total Amount Spent

Across all age groups, the majority of purchases fall within the 0 to 200 total amount spent range, with notable concentration observed. Age-specific peaks in higher spending ranges, particularly between ages 25-30 and 45-55, suggest distinct purchasing behaviors linked to certain age demographics.

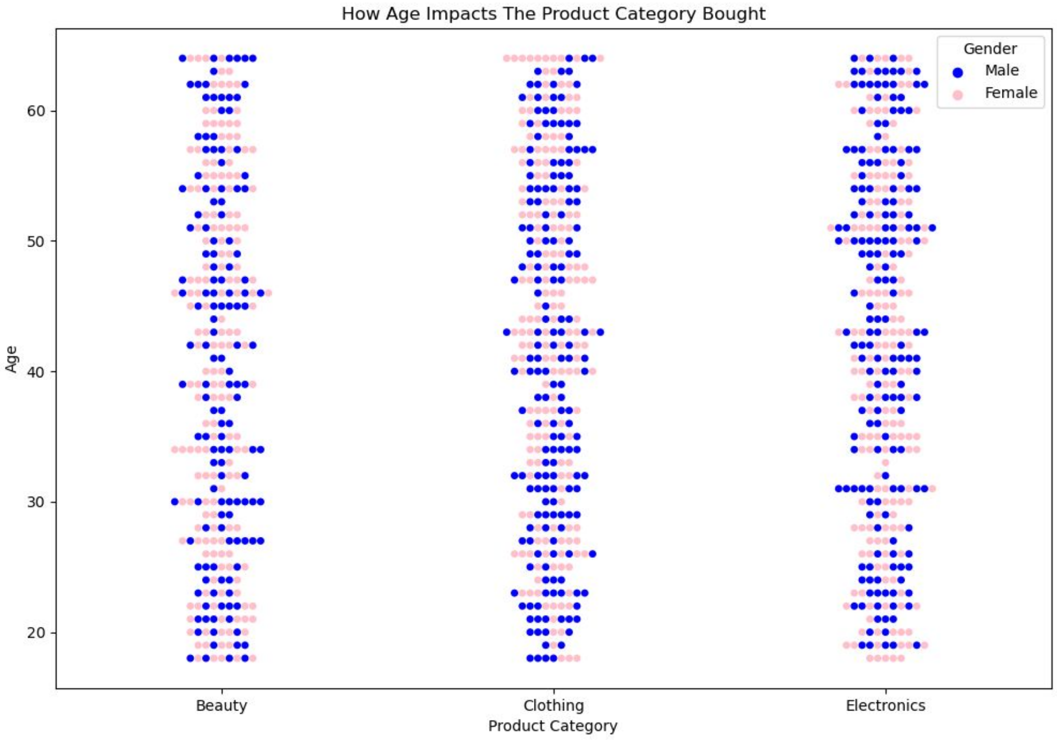
Understanding age-related spending patterns can enhance gender prediction models. Age-specific peaks and sparse spending brackets provide valuable insights into how age influences total spending behavior, aiding in the development of more accurate predictive models.



How Age Impacts Product Category Purchased

Clothing and Electronics emerge as dominant categories across age groups, with consistent interest indicated by steady dot clusters. However, the Beauty category exhibits more variability, suggesting age may play a nuanced role in influencing purchasing patterns.

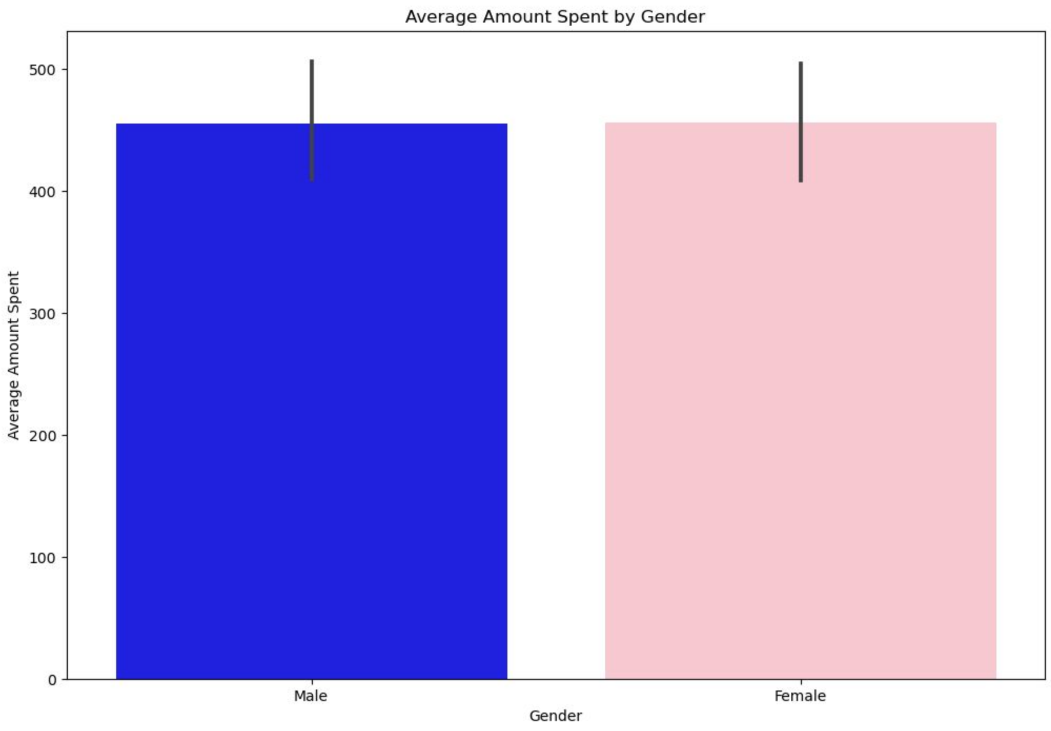
Targeted marketing opportunities arise from recognizing age-related preferences in product categories. While Clothing and Electronics show consistent interest, age-specific campaigns can leverage nuanced age-category relationships for more effective marketing strategies.



Average Amount Spent By Gender

The visualization indicates a minimal difference in average spending between males and females, with females potentially spending slightly more, albeit marginally.

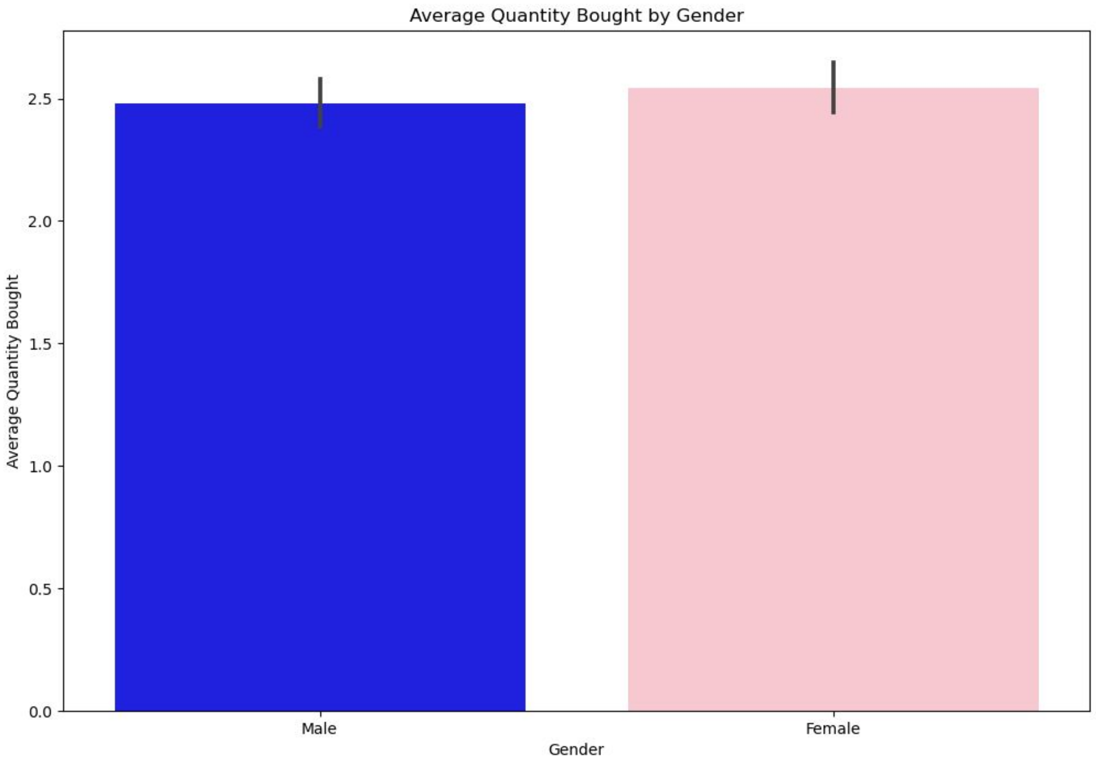
While this subtle distinction exists, the predictive power for gender classification in retail purchases may rely more on factors such as category, age, and price per unit. Emphasizing the need to explore additional features is crucial for building a robust predictive model.



Average Quantity Purchased By Gender

The visualization illustrates a consistent average quantity slightly over 2.5 for both males and females, with females holding a marginal lead, yet notably below 3.

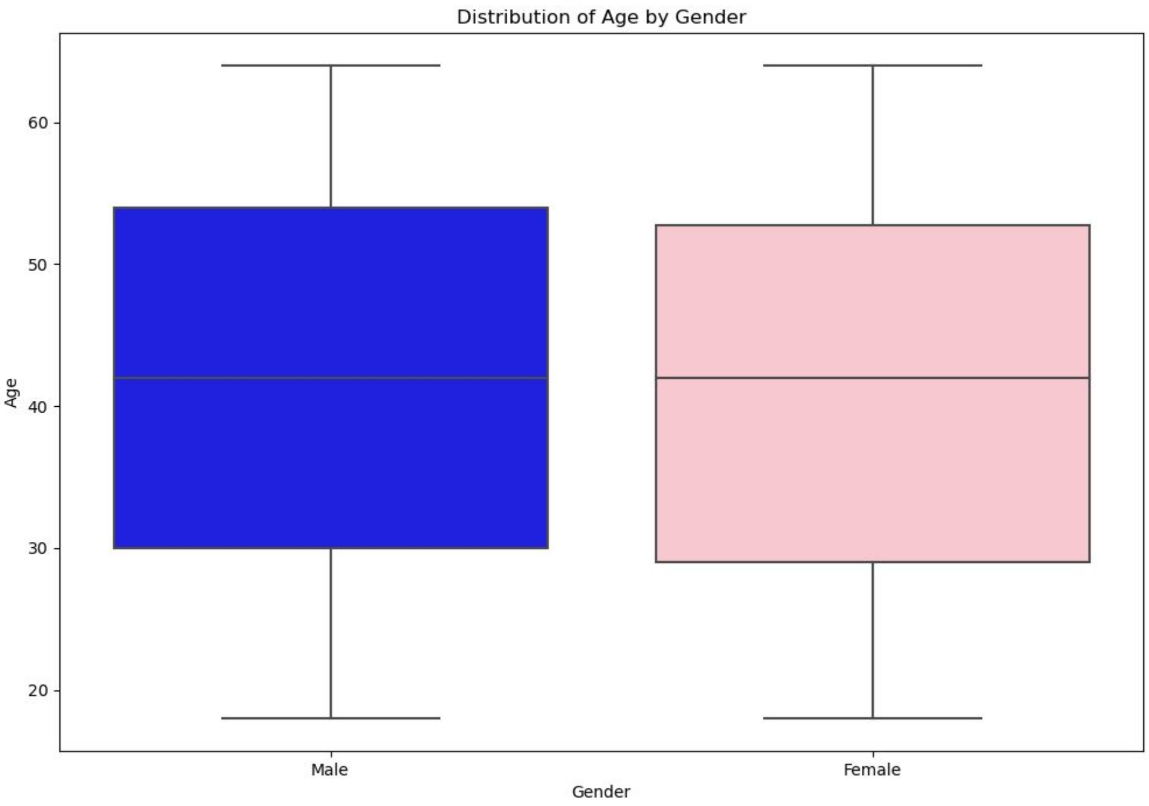
While there's a subtle difference in average quantity, this observation suggests that it may not be a strong predictor for distinguishing between male and female buyers. Exploring additional factors such as category, age, and price per unit is vital for developing a more accurate predictive model.



Distribution Of Age By Gender

Males typically range from 30 to 55 years old, with a median around age 42, while females start around age 28 and extend to approximately age 52, aligning with males at age 42.

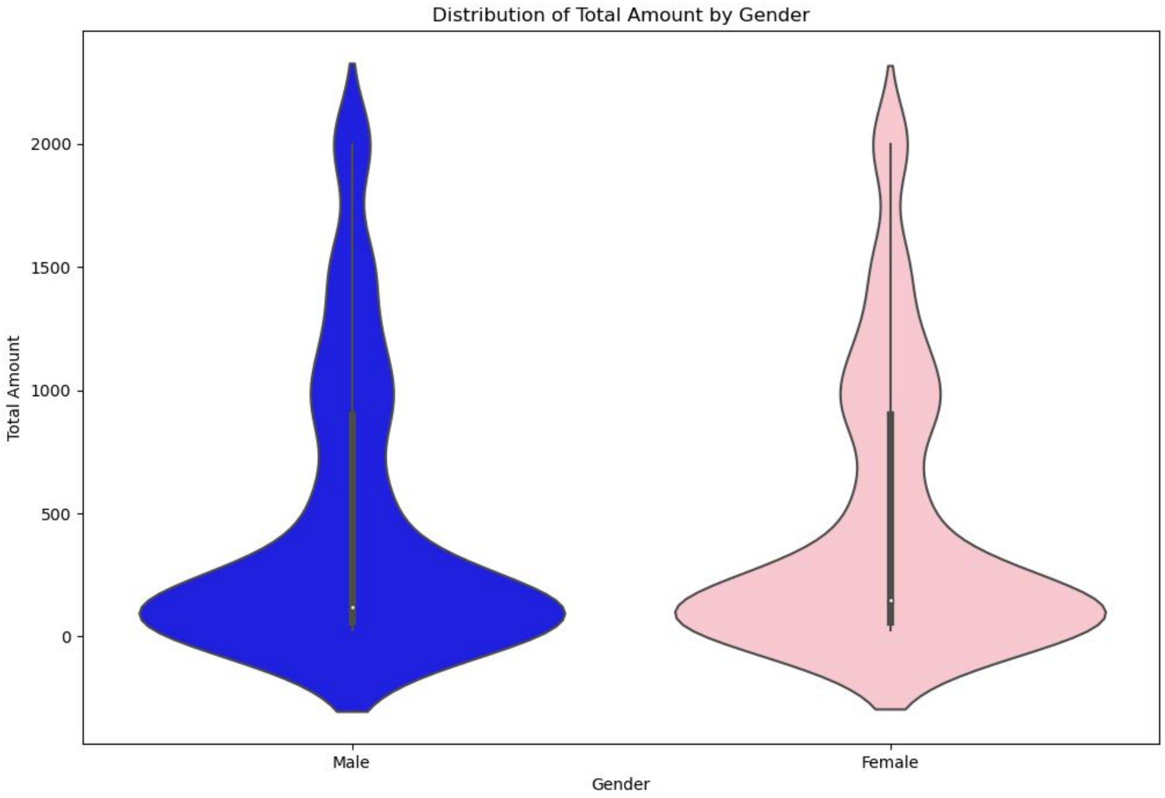
Age distributions across categories are consistent for males and females, suggesting age alone may not be highly discriminative for predicting gender. Recommending a comprehensive approach incorporating additional features like category and price per unit for a more robust gender prediction model is advisable.



Distribution Of Total Amount Spent By Gender

Both male and female violins exhibit similar distributions of total amounts, ranging from under 0 to over 2000, with a slight prominence for females around the 1000 mark.

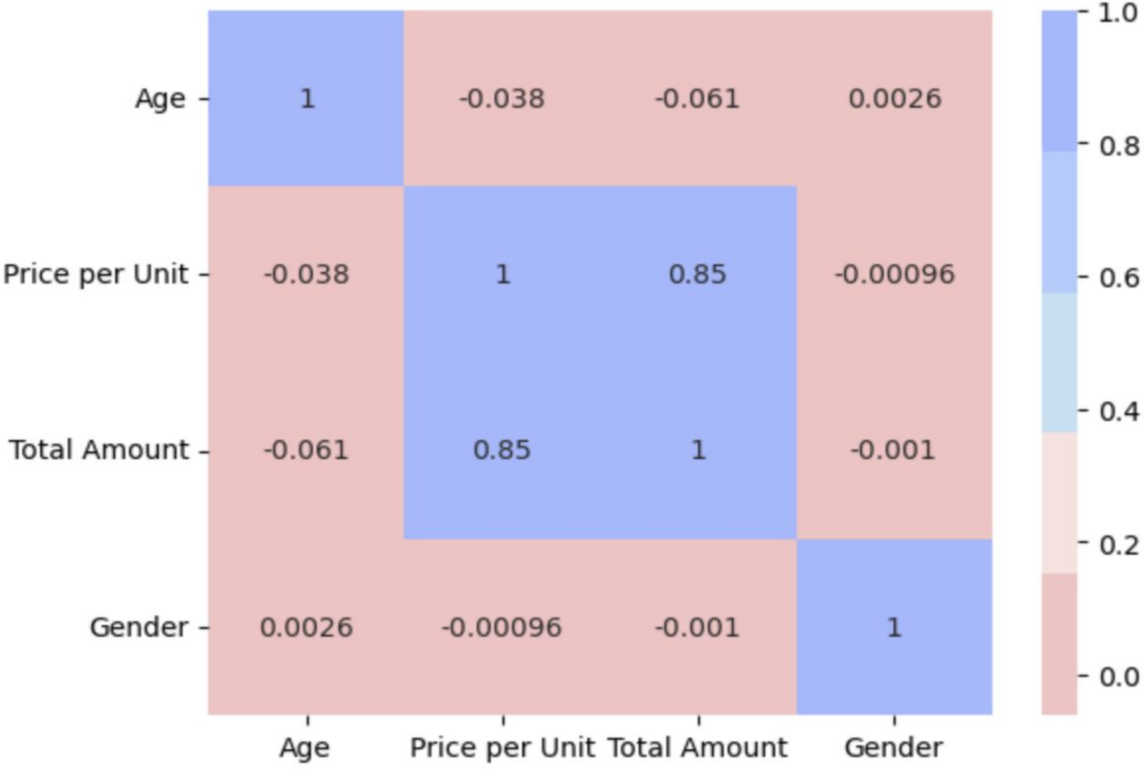
Despite some differences, both genders demonstrate comparable spending behavior across the entire range of total amounts. While there's a slightly higher concentration of female spending around the 1000 mark, total amount alone may not be a decisive factor in predicting gender.



HeatMap: How The Variables Impact Each Other

Interrelation of Variables: The heatmap illustrates how different variables correlate with each other. Gender exhibits a slight positive correlation with age (0.0026) and negative correlations with price per unit (-0.00096) and total amount spent (-0.001), indicating weak associations.

Impact on Prediction: While some variables show correlations, such as total amount spent having a strong positive correlation with price per unit (0.85), the heatmap suggests that gender alone may not strongly predict other variables or vice versa.



1

Modeling

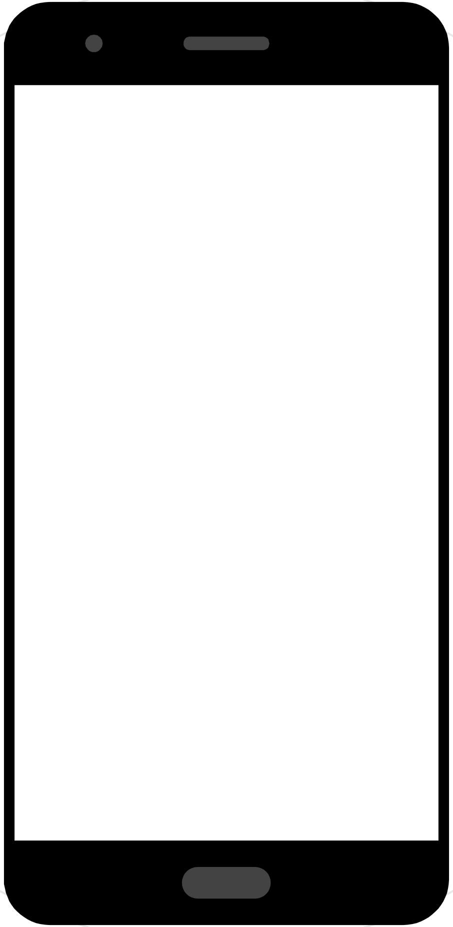
Performance

- **Model Performance Overview:**
  - Baseline Score: The baseline gender prediction accuracy stands at 51%.
  - K-Nearest Neighbors (KNN):
    - Best Testing Score: 56% (k=9).
    - Training Score for Best Performing Model: 64%.
  - Logistic Regression:
    - Testing Score: 48%.
    - Training Score: 52%.
  - Random Forest Classifier:
    - Best Testing Score: 53%.
    - Training Score: 100%.
- **Overall:** Best Performer: K-Nearest Neighbors with k=9 demonstrated the highest testing accuracy of 56%, outperforming Logistic Regression and Random Forest Classifier in terms of testing accuracy.

1

Application For Predicting

Streamlit App



Coming in next step

1

Synopsis

Conclusions:

Gender Prediction Insights:

The project successfully transformed and analyzed sales and customer data to predict gender based on demographic and transactional features, achieving a testing accuracy of up to 56% using K-Nearest Neighbors modeling.

Optimized Modeling Techniques:

By exploring various machine learning algorithms including K-Nearest Neighbors, Logistic Regression, and Random Forest Classifier, the project identified KNN as the most effective model for gender prediction, showcasing the importance of model selection and parameter tuning in achieving superior predictive performance.

Strategic Implications:

These insights offer actionable intelligence for marketing and sales teams to tailor strategies and promotions, catering to specific gender demographics. Furthermore, the project underscores the value of data-driven decision-making in enhancing business outcomes and driving competitive advantage in the e-commerce landscape.

Recommendations:

Continuous Data Refinement:

Sustained efforts in data collection and refinement, coupled with regular updates to the predictive model, can enhance its accuracy and relevance over time. This involves capturing additional relevant variables and refining existing features to better capture nuances in customer behavior and preferences.

Cross-Functional Collaboration:

Encouraging collaboration between data analysts, marketing specialists, and sales teams can foster a more holistic understanding of customer dynamics. By integrating insights from predictive models into strategic planning processes, organizations can better align their marketing efforts and product offerings with the diverse needs of their customer base, driving sustainable growth and competitive advantage.

Thanks!

Q&A