North South University
Department of Electrical & Computer Engineering

Assignment on EDA and Data Preprocessing (Individual assignment)

Course Code: CSE445
Section: 10
Faculty Initial: ITN
Student Name: MD Sadikul Haque Sadik
ID: 2031093642

Task 1: Answer the following questions

Question 1: Which of the following statements best describes a dataset?

C) A structured collection of data points representing some aspect of the real world.

Question 2: Why is data preprocessing an important step in data analysis?

C) It reduces noise and inconsistencies in the data, improving the quality of analysis.

Question 3: Which of the following is considered categorical data?

C) Colors of flowers (e.g., red, blue, yellow).

Question 4: What is one common method for handling missing data in a dataset?

B) Removing the entire row or column containing missing values.

Question 5: What does feature engineering involve in data analysis?

C) It involves creating or transforming new features to improve the model's performance.

Question 6: Why is splitting a dataset into training and testing sets important?

C) To ensure that the model's performance is evaluated on unseen data.

Question 7: What is a common technique to handle categorical data before feeding it into a machine learning model?

C) One-Hot Encoding, where each category becomes a binary column.

Question 8: Why might scaling numerical features in a dataset be necessary?

C) To ensure that all numerical features have the same unit of measurement.

Question 9: What is an outlier in the context of data analysis?

C) Unusual or extreme data points that significantly differ from the rest.

Question 10: What does data imputation involve?

D) Filling in missing values with estimated or calculated values.

Question 11: What is a consideration when dealing with time-series data in data analysis?

C) The order and timing of data points matter.

Question 12: What is the primary goal of dimensionality reduction techniques in data analysis?

D) To reduce the number of features while preserving relevant information.

Question 13: Why is addressing imbalanced classes important when building models?

C) Imbalanced classes can bias the model towards the majority class.

Question 14: Which preprocessing step is commonly used for text data before analysis?

A) Converting text data to numerical values using encoding techniques.

Task 2: Data Analysis and Machine Learning Preprocessing

Part 4: Insights and Data Preparation Summary

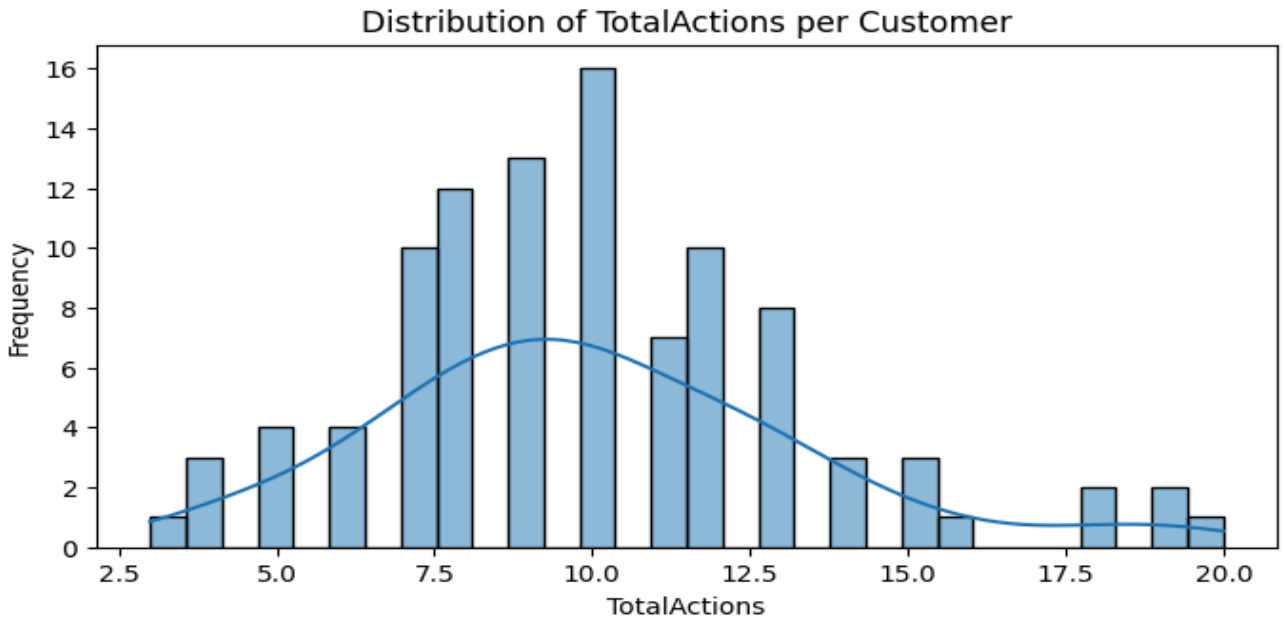
1. Summary of Data Analysis, Feature Engineering, and Preprocessing Steps

- **Loading and Inspecting Data**:
  The e-commerce dataset was loaded and previewed to assess its structure and the types of data in each column. This step helped identify missing values and potential issues, such as outliers and inconsistent data formats.
- **Handling Missing Values**:
  Missing values were addressed by filling **Price** with the mean, as it can handle data variability and does not distort central tendencies. Categorical columns **Category** was filled with "**N/A**" to avoid losing data during processing.
- **Feature Engineering**:
  A new feature, TotalSpent, was created by calculating the product of Price and Quantity, providing insights into the overall spending of each customer on individual transactions. Total spending per customer was also calculated to help identify high-value customers.
- **Grouping and Aggregation**:
  Data was grouped by Category to determine the most popular product categories, using frequency counts to evaluate popularity. Additionally, average Price for each category was calculated to provide insights into pricing strategies and variations across categories.
- **Encoding Categorical Variables**:
  One-hot encoding was applied to Category and Action, converting these categorical variables into a numerical format that machine learning algorithms can process without introducing arbitrary ordinal relationships.
- **Standardization of Numerical Features**:
  Price, Quantity, and TotalSpent were standardized using Z-score normalization. This ensures that all numerical features are on a similar scale, improving the performance and convergence of machine learning models.
- **Splitting the Data**:
  The data was split into training and testing sets (80% training, 20% testing) for machine learning. This setup allows for a reliable evaluation of model performance on unseen data.

2. Trends and Patterns Observed in the Data

Visualize the Distribution of Customer Actions:

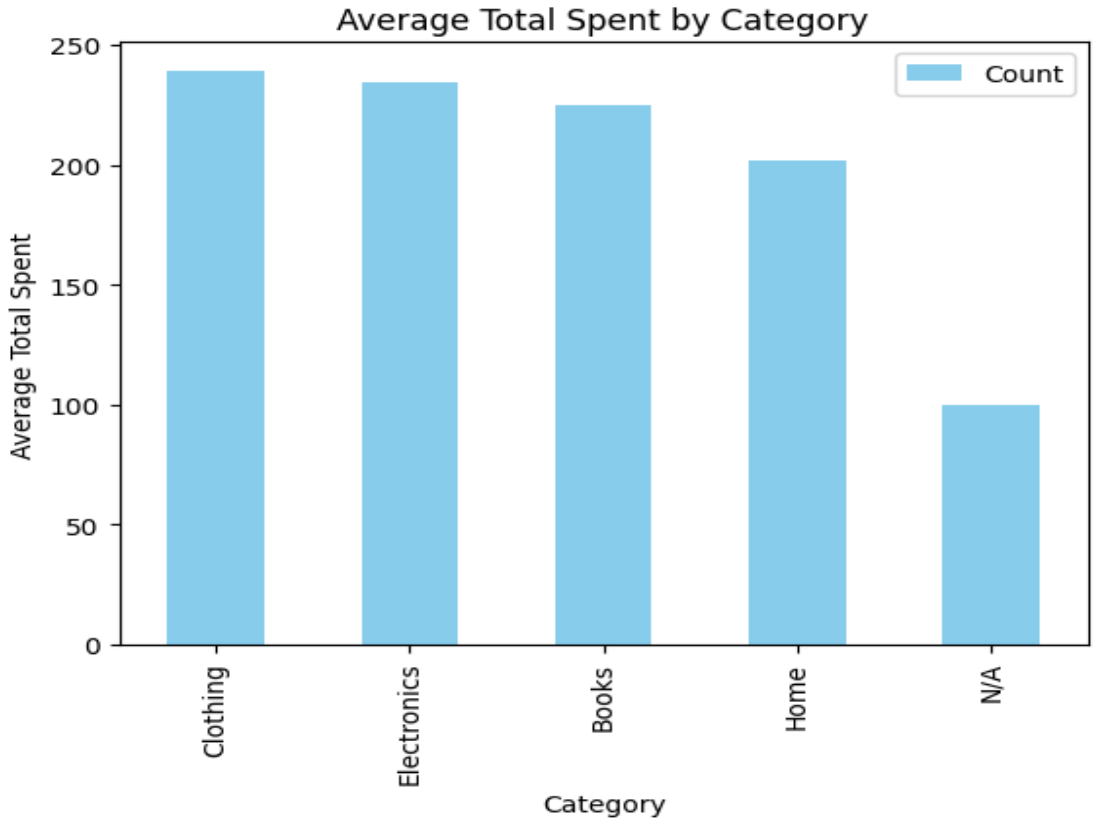
By analyzing TotalActions per customer, it was evident that certain customers had higher interaction frequencies, indicating a higher level of engagement. This could be used to segment customers and tailor marketing strategies to retain high-engagement users.



This histogram shows the distribution of total actions per customer, revealing clusters of high and low engagement.

Spending Patterns by Category:

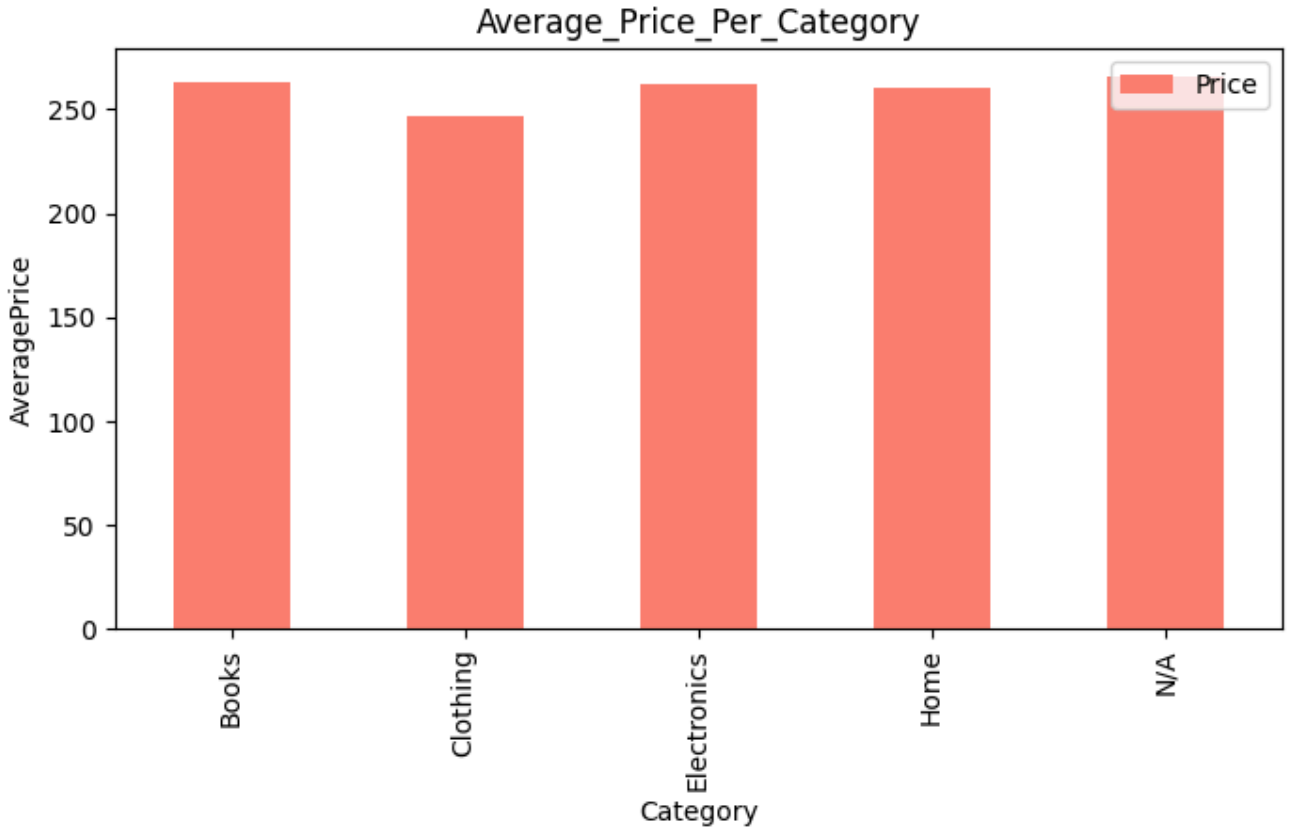
Analysis of TotalSpent revealed that certain categories generated higher spending, highlighting potentially popular product lines. This insight can help prioritize inventory and targeted marketing for specific categories.



This bar plot shows average spending across categories, which can guide inventory and sales strategies.

Average Price per Category:

Calculating the average price per category uncovered a range in pricing across product lines, indicating different price points associated with categories. Understanding this variation helps refine pricing strategies for each category.



This plot shows the average price distribution across categories, providing insight into pricing trends and possible premium product categories.

3. Rationale Behind Feature Engineering and Preprocessing Techniques

- **Handling Missing Values**:
  Missing values were filled based on data type and statistical appropriateness. For numerical columns, mean and median imputations were chosen as they provide reasonable central estimates. For categorical columns, "N/A" was used to maintain category diversity without discarding data.
- **One-Hot Encoding for Categorical Data**:
  One-hot encoding was applied to Category and Action to avoid imposing ordinal relationships

in non-ordinal data. This allows machine learning models to interpret the data accurately without bias from arbitrary category orderings.

- **Z-score Standardization**:
  Numerical features were standardized to prevent features with larger scales from disproportionately influencing model training. This technique improves the performance and convergence rate of models by ensuring all numerical data is on a comparable scale.
- **Train-Test Split**:
  The 80-20 split allows for an ample amount of data for training while maintaining a reliable test set for model evaluation. A random seed was used for consistency and reproducibility, ensuring that the same split is achieved each time.

Conclusion

The data analysis, preprocessing, and feature engineering steps taken have ensured that the dataset is clean, feature-rich, and ready for machine learning. The identified trends, such as high customer engagement in certain categories and spending patterns, provide actionable insights that can drive targeted marketing and inventory strategies. Additionally, by encoding and standardizing the data, it's now well-prepared for various machine learning models, laying a solid foundation for the next steps in predictive analysis and decision-making.