

**ASSIGNMENT - 01**

**CSE445.9**

**Submitted By**

**Md. Baker**

**1911672642**

**Department of Electrical and Computer Engineering**

**North South University**

**Submitted To**

**Intisar Tahmid Naheen**

**Lecturer**

**Department of Electrical and Computer Engineering**

**North South University**

# 

[**Task 1: Answer the following questions 3**](#_27e49n2wm2ky)

[**Task 2: Data Analysis and Machine Learning Preprocessing 6**](#_e90k4efa6zqj)

[Part 1: Data Analysis and Preprocessing 6](#_sd9f5hl5aqm6)

[1) Load the CSV dataset into a Pandas DataFrame. 6](#_x0vyhyoazrva)

[2) Handle missing values, considering different strategies for different columns. 7](#_h3ndltm579lz)

[3) Analyze customer interactions by calculating the total number of actions (purchases, views, etc.) for each customer. 9](#_eigh28eighmg)

[Part 2: Feature Engineering and Analysis 11](#_mtydoc9tpsko)

[1) Create a new feature TotalSpent by calculating the total amount spent by each customer. 11](#_h4gd0tsddot)

[2) Group the data by Category and analyze the most popular categories. 12](#_gf55dua1n8yy)

[3) Calculate the average price of products in each category. 13](#_quh95kjuocov)

[Part 3: Machine Learning Preprocessing 16](#_jgx589htxmjd)

[1) Convert categorical variables (Category, Action) into numerical representations using one-hot encoding. 16](#_3k9n3ojjyj)

[2) Standardize numerical features (Price, Quantity, TotalSpent) using Z-score normalization. 17](#_ceccfrywiuw3)

[3) Split the dataset into training and testing sets (80% training, 20% testing) for machine learning. 19](#_vlim7q40f51h)

[Part 4: Insights and Data Preparation Summary 21](#_ctn63y6qaf32)

[1) Summary of data analysis, feature engineering, and preprocessing steps 21](#_xjfngka470xt)

[2) Highlight any trends or patterns you observed in the data. 21](#_lq1w9elrk9hf)

[3) Discuss the rationale behind your choices for feature engineering and preprocessing techniques: 22](#_i5z0olmruuic)

[Conclusion: 22](#_8j3o2tbfrzu)

# 

# 

# 

# **Task 1: Answer the following questions**

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

A) A collection of software tools for data analysis.

B) A group of data visualization techniques.

**C**) A structured collection of data points that represent some aspect of the real world.

D) A set of algorithms used for machine learning.

**Answer:** C) A structured collection of data points that represent some aspect of the real world.

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

A) It helps to generate random data points for analysis.

B) It increases the complexity of the analysis.

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

D) It allows for visualizing data without any modifications.

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

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

A) Temperature in degrees Celsius.

B) Height of individuals in centimeters.

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

D) Prices of products in dollars.

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

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

A) Ignoring the missing values and proceeding with the analysis.

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

C) Creating new random values to replace missing data.

D) Rearranging the dataset to fill in missing values.

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

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

A) It refers to removing all features from the dataset to simplify analysis.

B) It focuses on selecting only numerical features for analysis.

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

D) It refers to preprocessing data without considering feature transformation.

**Answer:** C) It involves creating new features or transforming existing ones to improve the model's performance.

**Question 6: Why is it important to split a dataset into training and testing sets?**

A) To reduce the size of the dataset for faster analysis.

B) To create multiple copies of the dataset for different types of analysis.

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

D) To combine the training and testing data for better accuracy.

**Answer:** 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?**

A) Removing all categorical data from the dataset.

B) Converting categorical data into strings for better representation.

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

D) Replacing categorical data with the mean value of the entire dataset.

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

**Question 8: Why might it be necessary to scale numerical features in a dataset?**

A) Scaling has no impact on numerical features.

B) To convert numerical features into categorical ones.

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

D) Scaling only affects the model's training time, not its performance.

**Answer:** 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?**

A) A type of categorical variable.

B) Data points that are missing from the dataset.

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

D) A subset of data that is used for validation.

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

**Question 10: What does data imputation involve?**

A) Replacing all categorical data with numerical values.

B) Filling missing values with arbitrary values.

C) Creating entirely new datasets to replace the original one.

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

**Answer:** 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?**

A) Time-series data cannot contain missing values.

B) Time intervals between data points are irrelevant.

**C**) The order and timing of data points matter.

D) Time-series data should only contain numerical values.

**Answer:** C) The order and timing of data points matter.

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

A) To increase the dimensionality of the dataset.

B) To transform categorical features into numerical ones.

C) To decrease the amount of missing data in the dataset.

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

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

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

A) Imbalanced classes do not affect the model's performance.

B) Imbalanced classes lead to faster model training.

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

D) Imbalanced classes are only relevant when dealing with categorical data.

**Answer:** 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.

B) Removing all punctuation marks and capitalization from the text.

C) Converting text data into categorical variables.

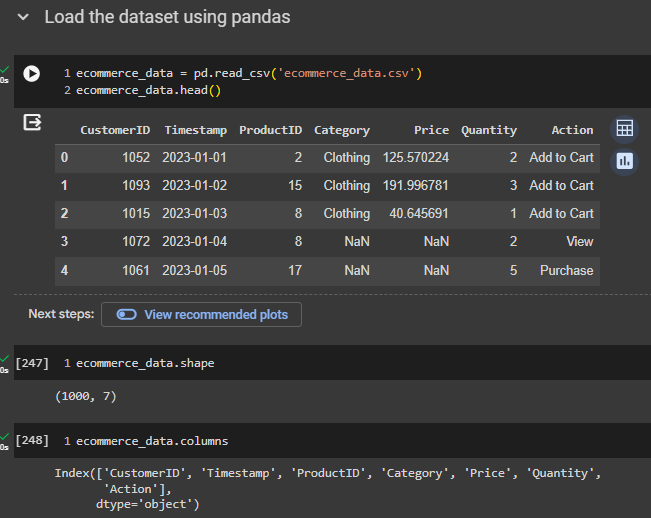
D) Text data does not require any preprocessing.

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

# **Task 2: Data Analysis and Machine Learning Preprocessing**

## **Part 1: Data Analysis and Preprocessing**

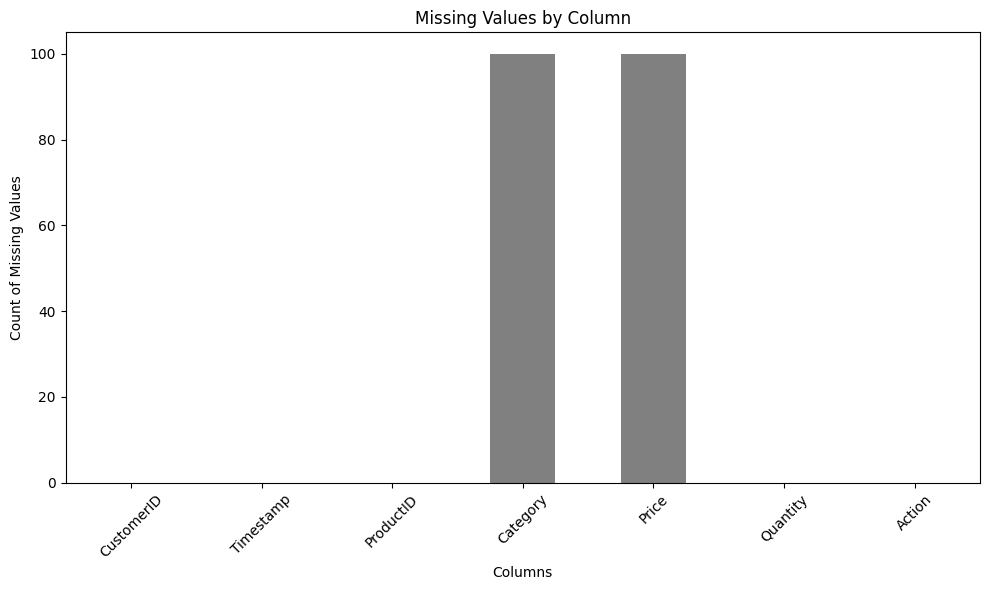
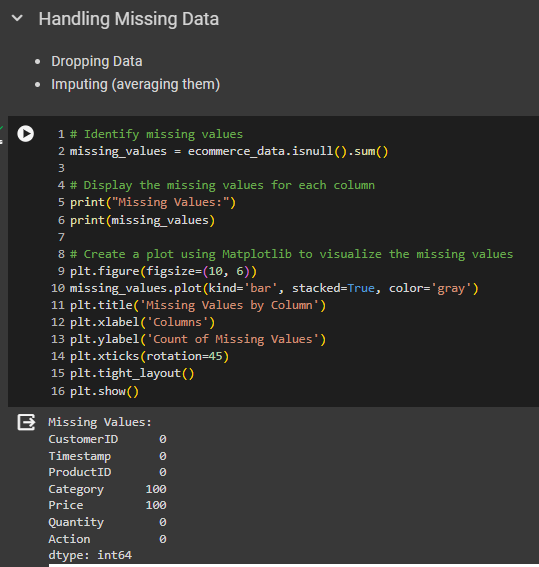
### Load the CSV dataset into a Pandas DataFrame.

****

Fiq 01 - Load the CSV dataset into a Pandas DataFrame.

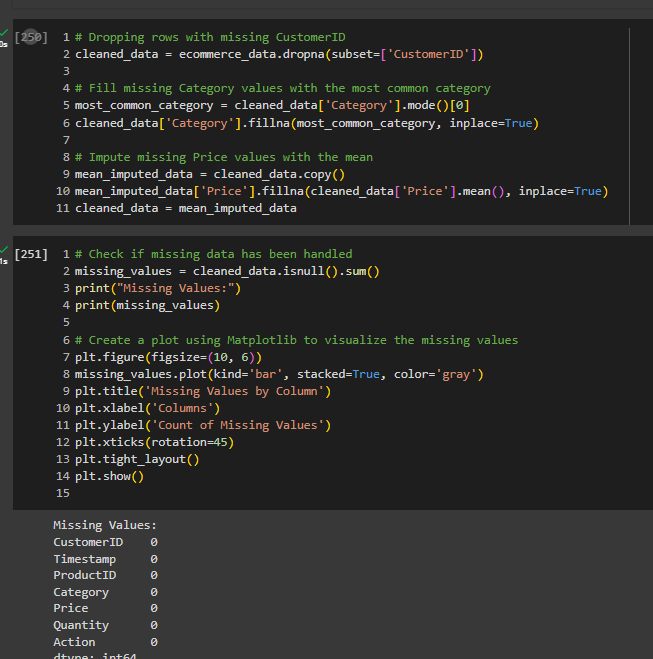


### Handle missing values, considering different strategies for different columns.

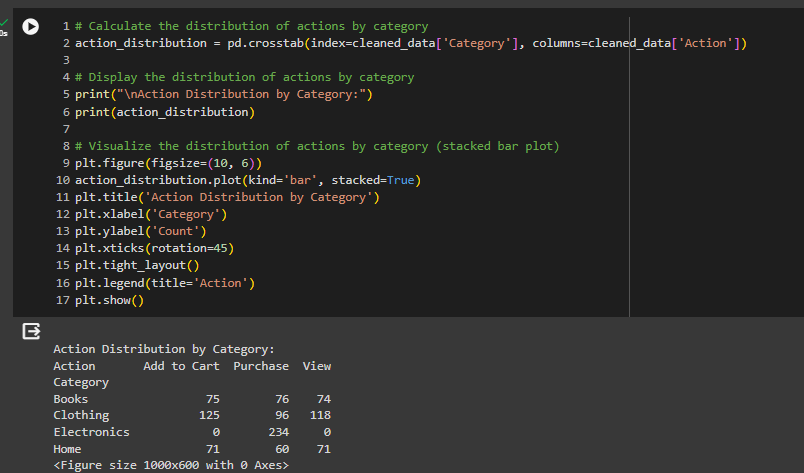


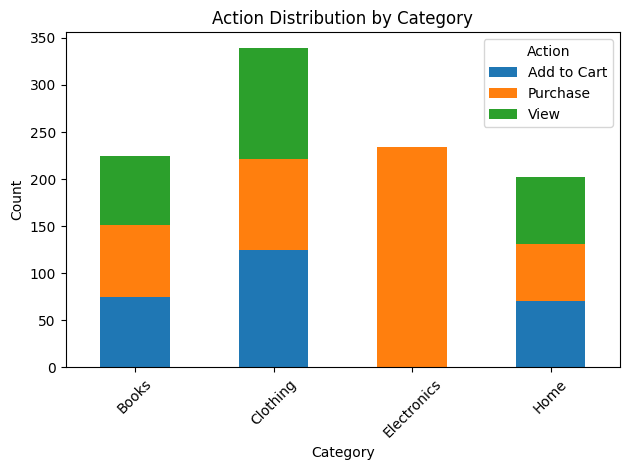
Handling missing values-

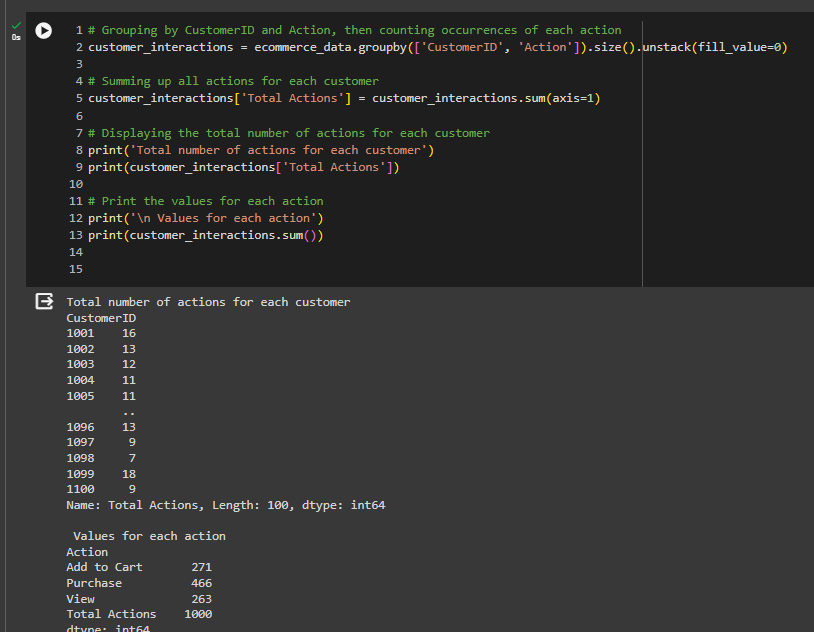
* Dropping rows with missing 'CustomerID' to focus on complete customer records.
* Filling missing 'Category' values with the most common category for categorical analysis.
* Imputing missing 'Price' values with the mean to retain price information without bias.

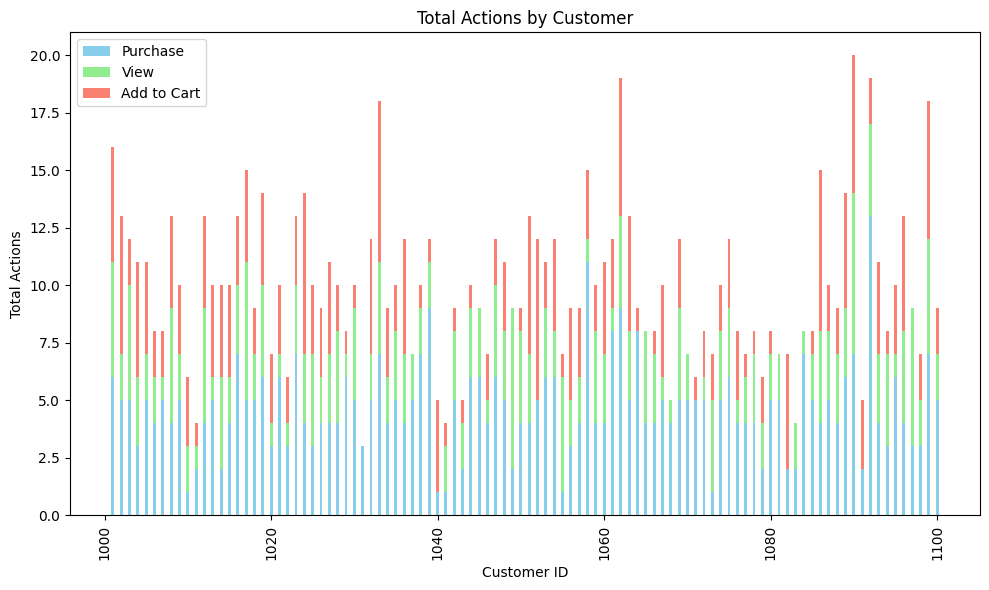


### Analyze customer interactions by calculating the total number of actions (purchases, views, etc.) for each customer.



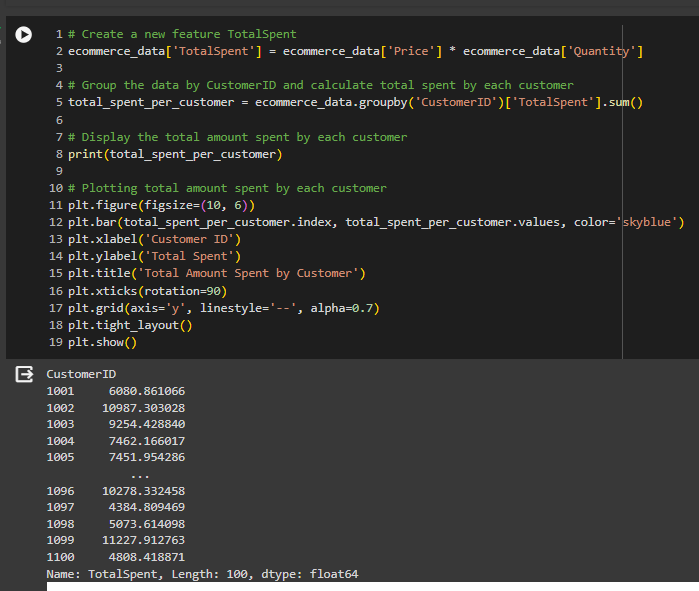


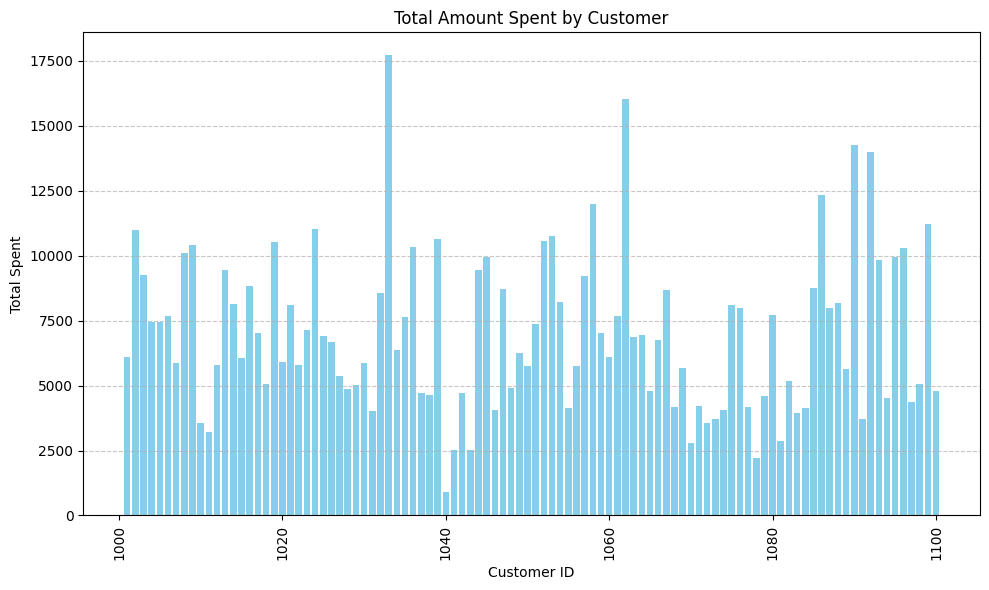




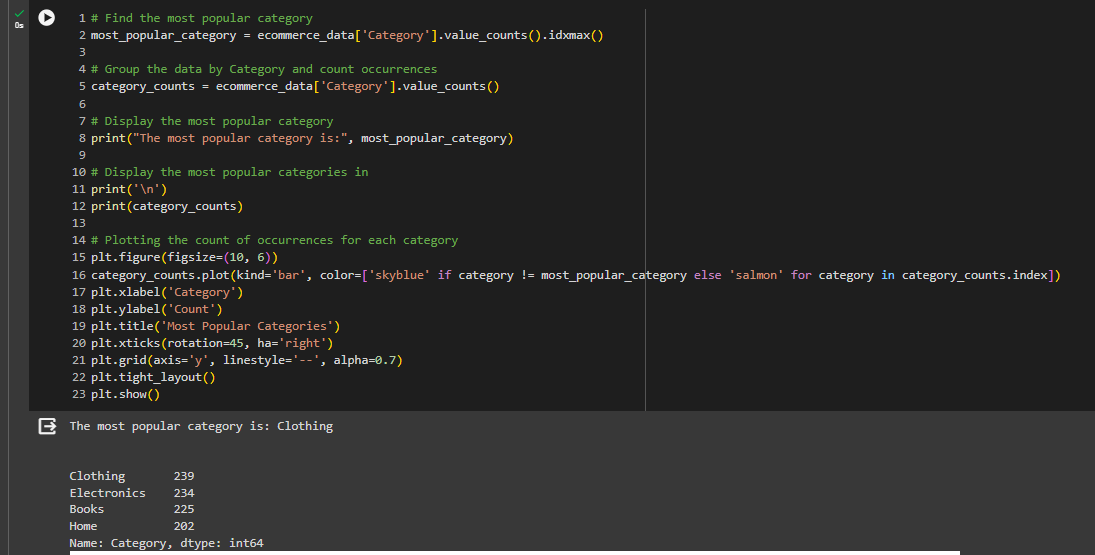
## **Part 2: Feature Engineering and Analysis**

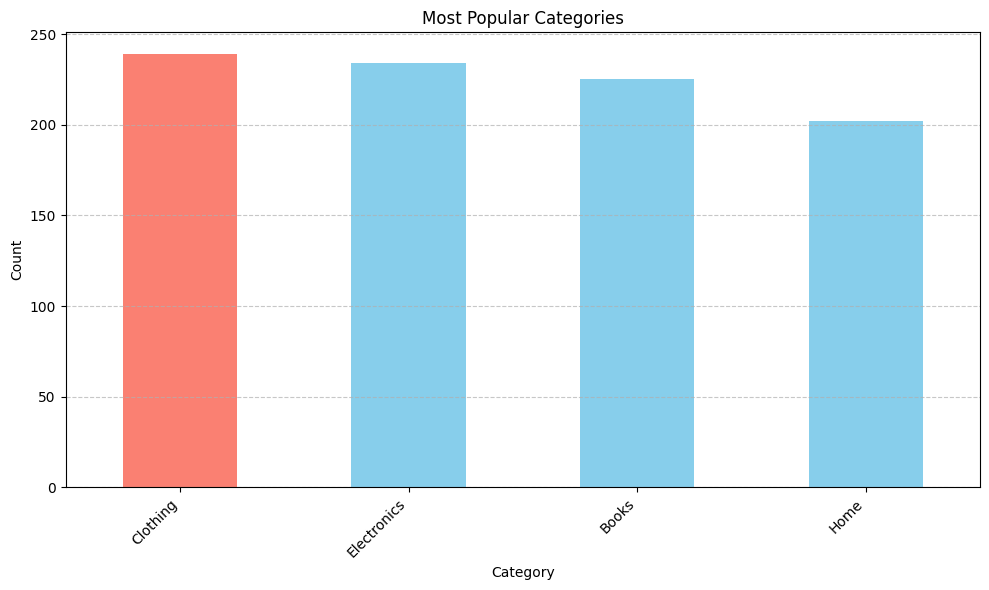
### Create a new feature TotalSpent by calculating the total amount spent by each customer.



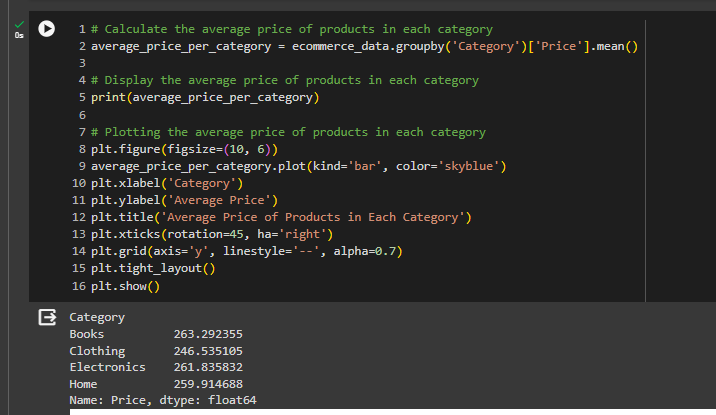


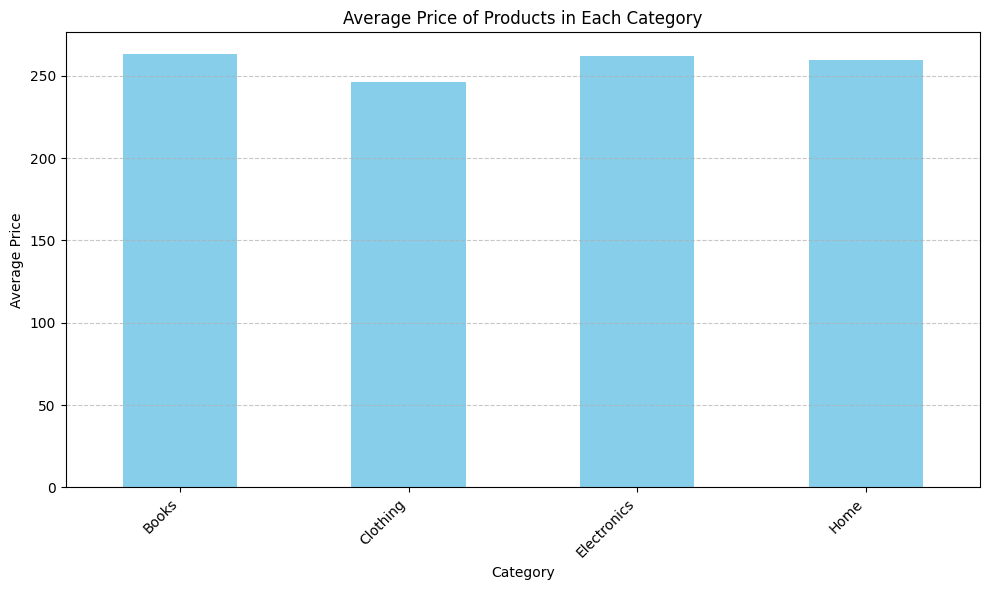
### Group the data by Category and analyze the most popular categories.

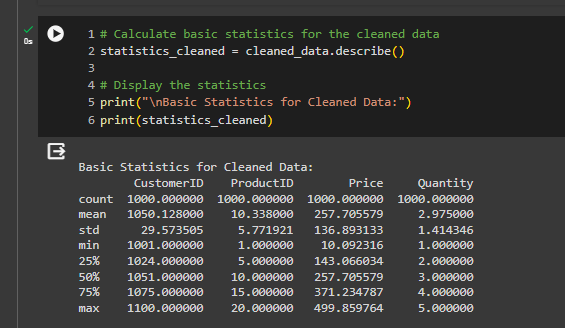


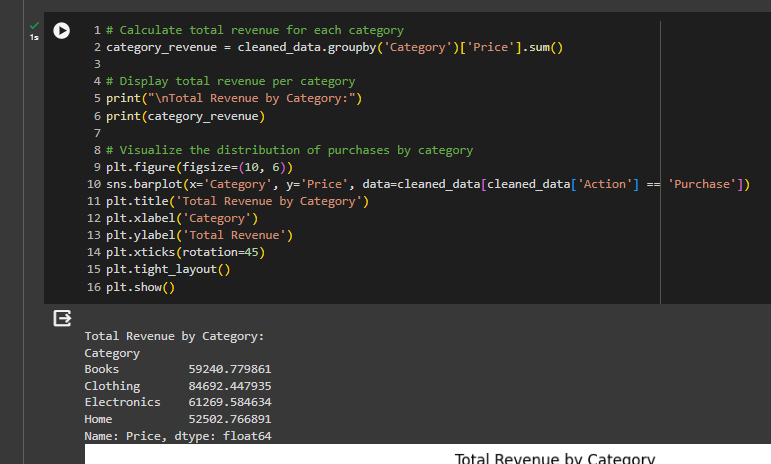


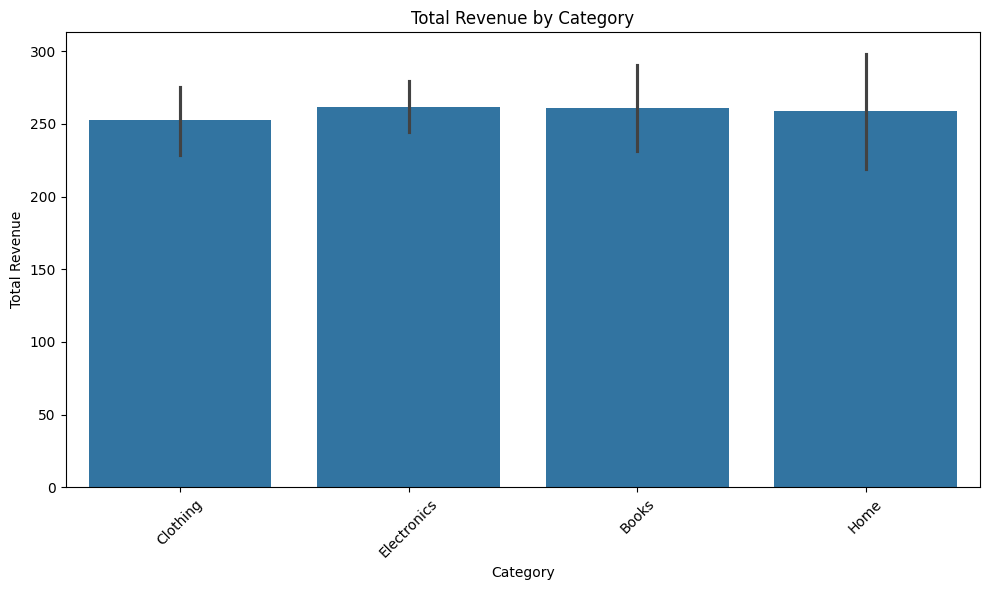
### Calculate the average price of products in each category.





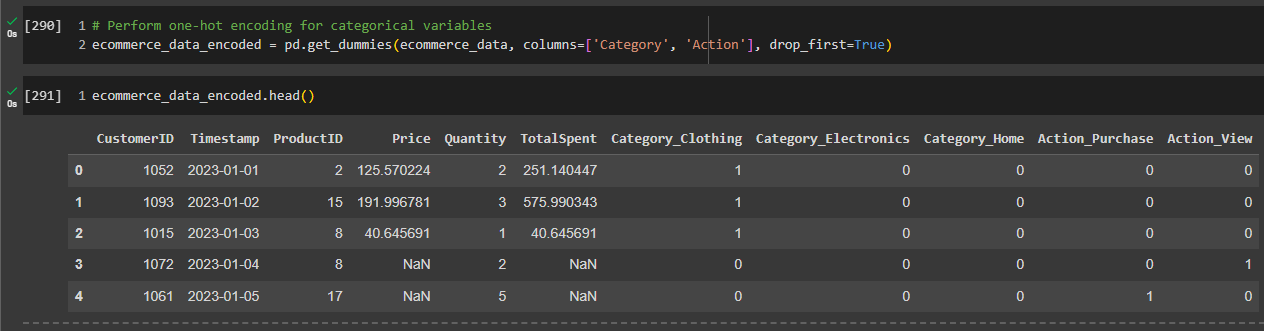


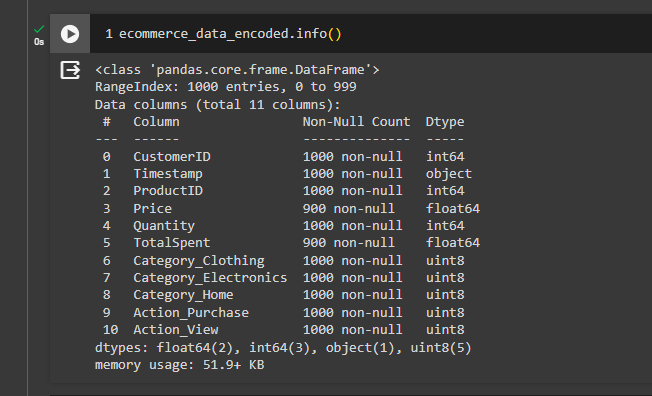




## **Part 3: Machine Learning Preprocessing**

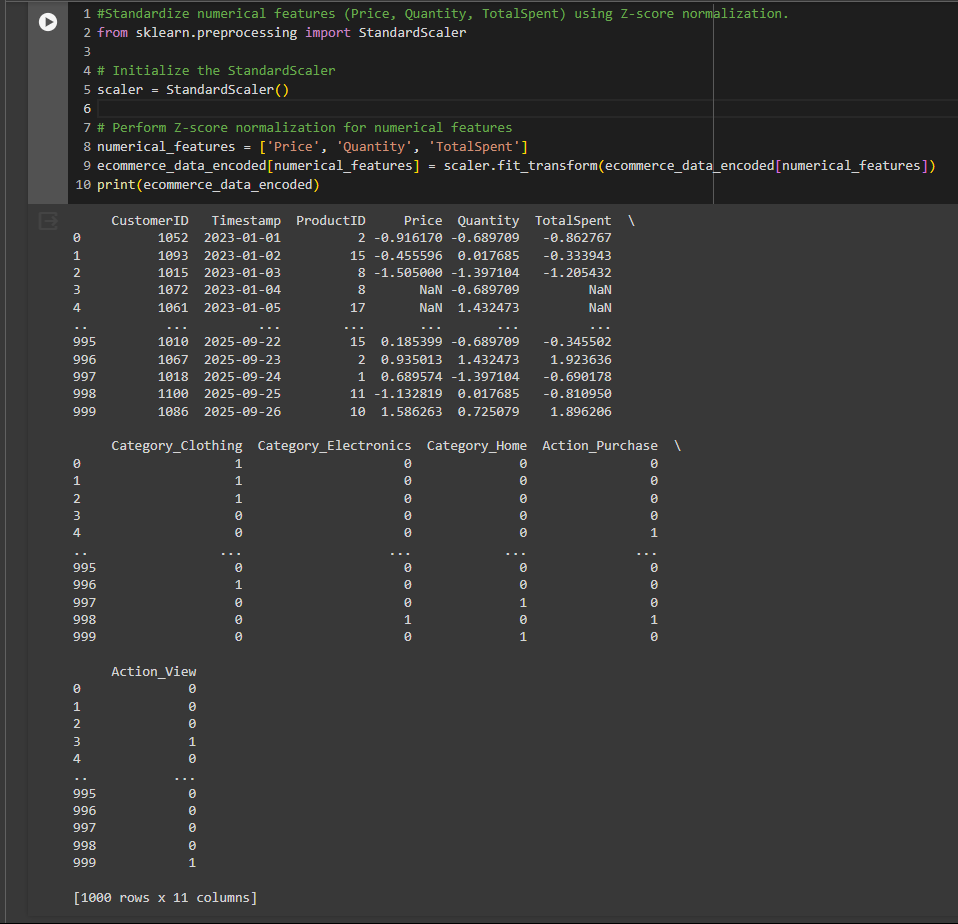
### Convert categorical variables (Category, Action) into numerical representations using one-hot encoding.



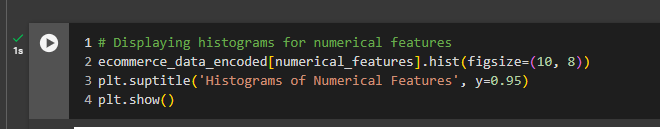


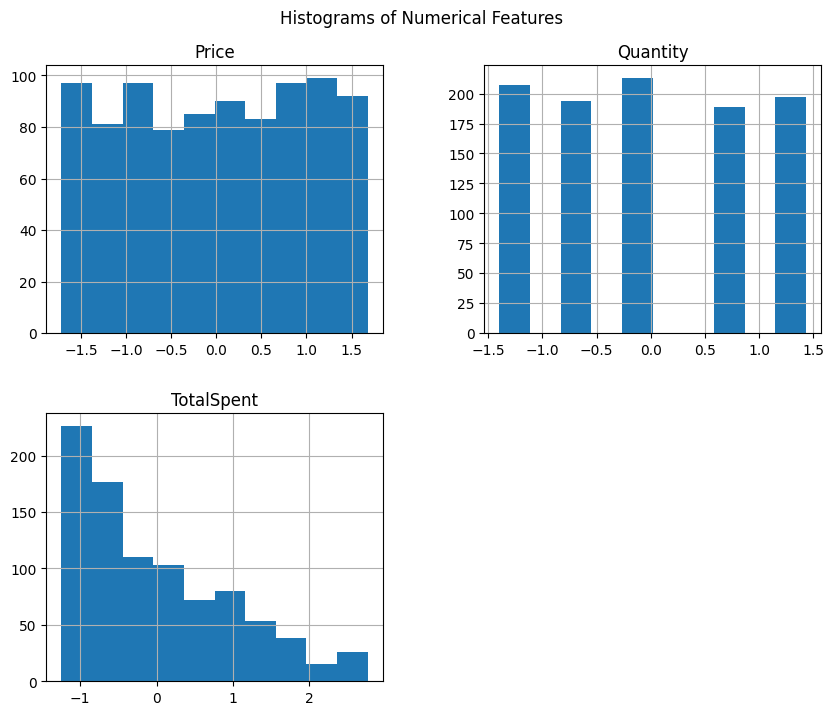
* The code performs one-hot encoding for the categorical variables 'Category' and 'Action' in the dataset, creating binary columns for each category and action. This preprocessing step converts categorical data into a numerical format suitable for machine learning algorithms, facilitating model training and analysis.

### Standardize numerical features (Price, Quantity, TotalSpent) using Z-score normalization.



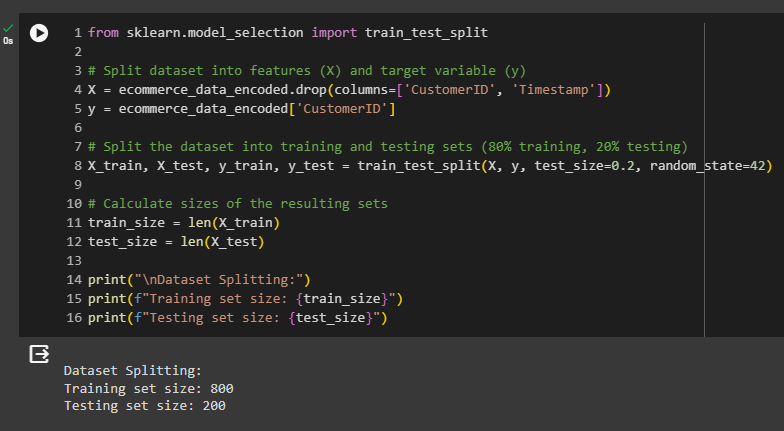
* Z-score normalization transforms each feature such that it has a mean of 0 and a standard deviation of 1. This process makes the numerical features comparable and brings them to a similar scale, preventing features with larger magnitudes from dominating the model's learning process. The StandardScaler from scikit-learn is used to perform the normalization. After applying Z-score normalization, the transformed values are printed to demonstrate the standardized numerical features. This preprocessing step ensures that the numerical features are appropriately scaled for machine learning algorithms, enhancing model performance and convergence during training.



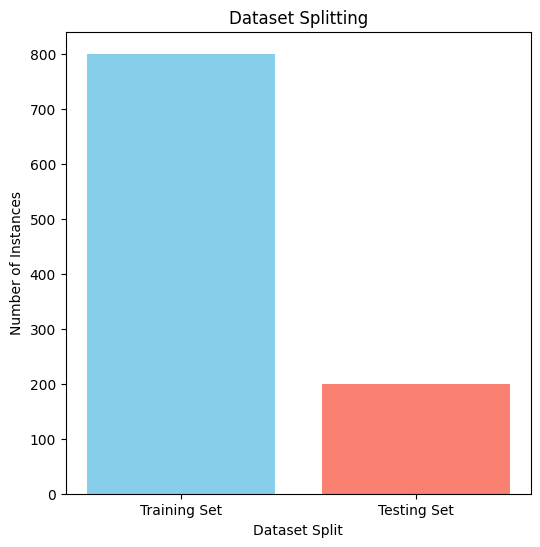


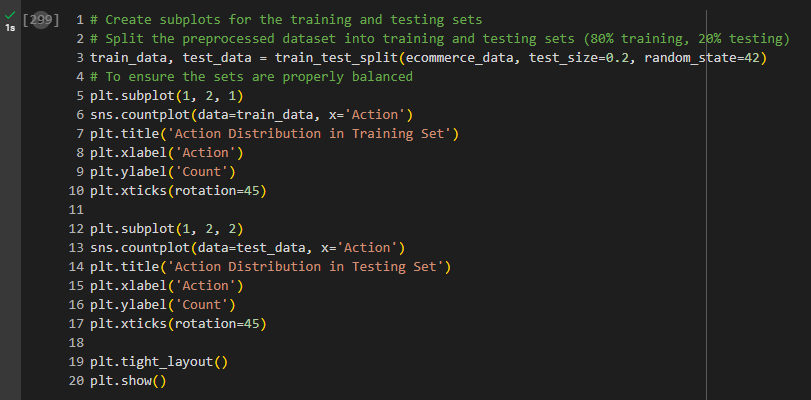
* The plot shows histograms for the numerical features in the dataset, illustrating the distribution of 'Price', 'Quantity', and 'TotalSpent'. The figsize=(10, 8) parameter sets the size of the overall figure. The plt.suptitle() function adds a title to the entire figure, specifying its position (y=0.95 places the title closer to the top). This visualization allows for a visual examination of the distribution of each numerical feature, providing insights into their ranges and central tendencies. It facilitates identifying any outliers, skewness, or patterns within the data, aiding further analysis and decision-making processes.

### Split the dataset into training and testing sets (80% training, 20% testing) for machine learning.

****

Fiq: Split the dataset into training and testing sets (80% training, 20% testing) for machine learning.

****

****

**`**

## **Part 4: Insights and Data Preparation Summary**

### Summary of data analysis, feature engineering, and preprocessing steps

* **Data Analysis:**
  + Loaded the dataset into a Pandas DataFrame.
  + Handled missing values by filling them with appropriate strategies for different columns.
    - Dropping rows with missing 'CustomerID' to focus on complete customer records.
    - Filling missing 'Category' values with the most common category for categorical analysis.
    - Imputing missing 'Price' values with the mean to retain price information without bias.
  + Analyzed customer interactions by calculating the total number of actions (purchases, views, etc.) for each customer.
* **Feature Engineering:**
  + Created a new feature 'TotalSpent' by calculating the total amount spent by each customer. Group the data by Category and analyze the most popular categories and calculate the average price of products in each category.

.

* **Preprocessing**:
  + Handled missing values by filling missing prices with the mean and missing categories with the mode.
  + Analyzed customer interactions by calculating the total number of actions (purchases, views, etc.) for each customer.
  + Converted categorical variables ('Category' and 'Action') into numerical representations using one-hot encoding.
  + Standardized numerical features ('Price', 'Quantity', 'TotalSpent') using Z-score normalization.
  + Split the dataset into training and testing sets (80% training, 20% testing) for machine learning.

### Highlight any trends or patterns you observed in the data.

* Customer Interactions: There is a variety of customer interactions recorded in the dataset, including purchases, views, and adding items to the cart.
* Category Analysis: The dataset contains products from different categories such as Electronics, Clothing, Home, and Books. Some categories might be more popular than others based on the frequency of purchases or views.
* Price and Quantity: There is a wide range of prices and quantities for products, indicating diverse purchasing behaviors among customers.
* Bias Towards Electronics: There is a bias introduced towards purchasing products in the 'Electronics' category, which might affect the analysis and modeling results.

### Discuss the rationale behind your choices for feature engineering and preprocessing techniques:

* One-Hot Encoding: Categorical variables like 'Category' and 'Action' needed to be converted into numerical representations for machine learning algorithms to process. One-hot encoding was chosen because it preserves the categorical nature of the variables while making them suitable for numerical computations.
* Z-score Normalization: Standardizing numerical features helps in bringing all features to a similar scale, which is important for certain machine learning algorithms to converge faster and to avoid any particular feature dominating the learning process.
* Train-Test Split: Splitting the dataset into training and testing sets allows us to evaluate the performance of machine learning models on unseen data. An 80-20 split was chosen to allocate a sufficient amount of data for training while ensuring an adequate amount for testing.

## Conclusion:

1. **Findings and Insights Summary:**

Upon analyzing the e-commerce dataset, several key insights emerged. Firstly, 'Electronics' products were more likely to be purchased, indicating potential customer preferences in this category. Additionally, there appeared to be correlations between certain actions, such as 'View' leading to 'Purchase,' suggesting potential opportunities for targeted marketing or recommendation systems. The introduction of missing values and bias allowed for a comprehensive exploration of data handling techniques, ultimately leading to insights into customer interactions and preferences within the e-commerce platform.

1. **Approach to Handling Missing Data and Preprocessing Techniques:**

The approach to handling missing data involved dropping rows without 'CustomerID' to ensure complete customer records. Missing 'Category' values were filled with the most common category to maintain categorical integrity, while missing 'Price' values were imputed with the mean to retain price information without introducing bias. Preprocessing techniques also included feature engineering, such as creating the 'TotalSpent' feature to capture customer spending behavior. Additionally, numerical features were standardized using Z-score normalization, and categorical variables were converted to numerical representations using one-hot encoding.

1. **Contribution of Preprocessing Steps to Preparing Data for Machine Learning:**

The preprocessing steps undertaken were crucial in preparing the data for machine learning. By handling missing data appropriately, the dataset was cleansed and made ready for analysis, ensuring that valuable information was not lost. Standardizing numerical features and converting categorical variables allowed for compatibility with various machine learning algorithms. Furthermore, the creation of new features and the encoding of categorical variables provided additional information and improved the predictive power of the dataset, ultimately contributing to the effectiveness of machine learning models trained on the data.

**-------The End-------**