DES432
Statistics and Data Modeling

Online Buying and Selling

(E-commerce Transactions)

Report

8 February 2026

Members

6622781035 Mathanapan Sangwongkaro

6622781043 Nutnicha Masomboon

Sirindhorn International Institute of Technology

Thammasat University

Introduction

Online buying and selling platforms play a significant role in modern commerce. With increasing access to the internet and digital technologies, consumers rely heavily on e-commerce platforms to purchase goods and services. As a result, these platforms generate large amounts of transactional data related to customer behavior, spending patterns, and delivery performance. Such data can provide valuable insights for improving business operations and customer satisfaction.

Despite the potential value of e-commerce data, real-world datasets are often imperfect. Missing values, extreme observations, and inconsistencies frequently occur due to incomplete user input or system limitations. These data quality issues must be carefully addressed before meaningful analysis can be conducted.

The objective of this project is to apply exploratory data analysis (EDA) and basic statistical inference techniques to an online shopping dataset. Rather than focusing on prediction accuracy, the project emphasizes statistical reasoning, transparent data cleaning decisions, and interpretation of results with uncertainty.

Dataset Description

The dataset used in this project represents online shopping transactions from an e-commerce platform. Each observation corresponds to one completed order, which serves as the unit of analysis. The dataset contains 1,000 observations, meeting the project's minimum dataset size requirement.

The dataset includes both numerical and categorical variables. Numerical variables consist of customer age, order value, and delivery days. Categorical variables include gender, product category, payment method, and return status. This mixture of variable types allows for distribution analysis, group comparisons, and basic inferential analysis.

Overall, the dataset reflects a realistic e-commerce environment and is appropriate for practicing exploratory analysis and statistical reasoning.

Data Quality Issues

Initial inspection of the dataset revealed several data quality issues. Missing values were identified in the customer age, gender, and delivery days variables. These missing values may result from incomplete customer profiles or system recording errors, which are common in online platforms.

In addition, the order value variable contains extreme values that can be considered outliers. These observations represent unusually large purchases. Rather than treating them as errors, these values were retained because they may reflect genuine customer behavior and provide useful information about high-spending customers.

Data Cleaning and Preprocessing

Data cleaning decisions were made carefully to maintain statistical validity and transparency. Before any cleaning steps were applied, missing values were identified and reported. The original dataset was preserved to allow for clear comparison between the raw and cleaned data.

Missing numerical values in customer age and delivery days were handled using median imputation. The median was chosen because it is less sensitive to skewed distributions and extreme values than the mean. Missing categorical values in gender were replaced with an “Unknown” category to preserve all observations.

No observations were removed solely due to missing values or extreme values. This approach helps reduce the risk of sampling bias and ensures that the cleaned dataset remains representative of the original data.

Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted using visual and numerical techniques. A histogram of order value shows a right-skewed distribution, indicating that most customers spend moderate amounts, while a small number make very high-value purchases. This pattern is common in e-commerce data.

Boxplots comparing order values across product categories show that the electronics category has a higher median order value than other categories. This suggests that electronics products tend to be more expensive and contribute significantly to total sales.

Additional boxplots comparing delivery days between returned and non-returned orders indicate that returned orders generally have slightly longer delivery times. Although the difference is not large, this pattern suggests a possible relationship between delivery performance and return behavior.

Descriptive Statistics

Descriptive statistics were used to summarize the central tendency and variability of key numerical variables in the dataset. For order value, the mean is noticeably higher than the median, which confirms the right-skewed distribution observed during the exploratory data analysis. This indicates that while most customers spend moderate amounts, a smaller proportion of customers make very high-value purchases that raise the overall average.

The interquartile range (IQR) of order value highlights substantial variability in customer spending behavior. This variability reflects the diversity of products available on the platform, ranging from low-cost items to more expensive goods. As a result, measures of spread such as the IQR are particularly useful for understanding typical spending patterns beyond simple averages.

Customer age also shows a relatively wide distribution, suggesting that the e-commerce platform attracts users from different age groups. Delivery days, in contrast, exhibit less variation, with most orders delivered within a relatively short time frame. This relatively consistent delivery performance aligns with standard expectations for online shopping services and provides context for interpreting customer satisfaction and return behavior.

Statistical Inference

Basic statistical inference was applied to account for uncertainty in the analysis and to move beyond purely descriptive findings. A 95% confidence interval was constructed for the mean order value to estimate the average amount spent by customers on the platform. Rather than relying on a single point estimate, this confidence interval provides a range of plausible values for the true mean and emphasizes the uncertainty inherent in sample-based data.

The confidence interval suggests that the true average order value is likely to fall within a reasonably narrow range, given the sample size of 1,000 observations. This result increases confidence in the stability of the estimate, while still acknowledging that the true population mean cannot be known with certainty.

In addition to confidence interval estimation, a comparison of mean delivery days between returned and non-returned orders was conducted. The results show that returned orders have a higher average delivery time than non-returned orders. This finding suggests a potential association between delivery delays and product returns. However, it is important to emphasize that this analysis does not establish causality. Other unobserved factors, such as product quality or customer expectations, may also contribute to return behavior.

Conclusion and Limitations

This project demonstrates how exploratory data analysis and basic statistical inference can be used to gain insights from realistic and imperfect e-commerce data. By identifying and addressing data quality issues such as missing values and outliers, the analysis highlights the importance of transparent data cleaning decisions in statistical practice.

Several limitations should be acknowledged. First, the dataset is simulated and may not fully capture the complexity of real-world online shopping environments. Second, the analysis focuses on a limited set of variables and does not include factors such as promotions, customer reviews, or seasonal effects, which may influence purchasing and return behavior. These limitations suggest that the results should be interpreted with caution.

Despite these limitations, the project provides a clear example of how meaningful insights can be derived from messy data through careful exploration and basic inference. The findings reinforce the value of statistical reasoning and uncertainty awareness when working with real-world datasets and support the key message of the project: understanding data is often more important than finding a single “correct” answer.

Dataset

The dataset used in this project is provided in CSV format.

File name:

online_shopping_dataset.csv

Code

The analysis code used in this project is publicly available on GitHub at the following link:

https://github.com/6622781043-svg/Project1/blob/main/project1-DES432