VIETNAM GENERAL CONFEDERATION OF LABOR
TON DUC THANG UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY

NGUYEN NHAT HUY
CHAU BAO NHAN

FINAL REPORT
DATA ANALYSIS
AND VISUALIZATION

HO CHI MINH CITY, 2025

VIETNAM GENERAL CONFEDERATION OF LABOR
TON DUC THANG UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY

NGUYEN NHAT HUY – 522H0134
CHAU BAO NHAN – 522H0093

FINAL REPORT
DATA ANALYSIS
AND VISUALIZATION

Instructor
PhD. Tran Luong Quoc Dai

HO CHI MINH CITY, 2025

ACKNOWLEDGEMENT

Hello Mr. Tran Luong Quoc Dai, we would like to send our sincere thanks to you throughout the study process. We truly appreciate you taking the time to ensure we understood each concept fully before moving on. Thank you also for the additional resources and study materials you provided. They have helped reinforce what we have learned and allowed us to go deeper into topics we found most interesting. We are grateful not just for your knowledge and expertise in teaching, but also for your kindness and encouragement. Thank you again for everything you have done to help us in our education.

We sincerely thank you.

Ho Chi Minh City, April 17th, 2025

Author

Nhan

Chau Bao Nhan

Huy

Nguyen Nhat Huy

THE WORK IS COMPLETED

AT TON DUC THANG UNIVERSITY

I hereby declare that this is my own research project and is under the scientific guidance of PhD. Tran Luong Quoc Dai. The research content and results in this topic are honest and have not been published in any form before. The data in the tables for analysis, comments, and evaluation were collected by the author from different sources and clearly stated in the reference section.

Project also uses a number of comments, assessments as well as data from other authors and other organizations, all with citations and source notes.

If any fraud is detected, I will take full responsibility for the content of my Project. Ton Duc Thang University is not involved in copyright violations caused by me during the implementation process (if any).

Ho Chi Minh City, April 17th, 2025

Author

Nhan

Chau Bao Nhan

Huy

Nguyen Nhat Huy

ABSTRACT

This report explores the application of data analysis and visualization techniques in examining real-world datasets to derive meaningful insights. By employing exploratory data analysis (EDA), probability distribution analysis, hypothesis testing, and correlation analysis, the study enhances understanding of data patterns and relationships. Key components include summarizing dataset characteristics (e.g., record count, variable types, and missing data handling), statistical descriptions with visualizations (e.g., histograms, box plots, and scatter plots), and outlier detection to ensure data quality. Additionally, the report investigates probability distributions of key variables, conducts statistical tests (e.g., t-test, Chi-square, ANOVA) to answer research questions, and evaluates correlations using Pearson or Spearman coefficients, visualized through heatmaps and scatter plots. Implementation strategies involve Python coding with clear documentation, while evaluation metrics focus on analytical depth, visualization clarity, and practical significance of findings. This report emphasizes the role of data-driven techniques in improving analytical accuracy, interpretability, and decision-making quality.

Keywords*: Data analysis, exploratory data analysis (EDA), visualization, probability distribution, hypothesis testing, correlation analysis, Python, data insights.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

API	Application Programming Interface
EDA	Exploratory Data Analysis
KDE	Kernel Density Estimation
IQR	Inter Quartile Range
ANOVA	Analysis of Variance

CHAPTER 1. INTRODUCTION

1.1 Problem Statement

In the era of data-driven decision-making, raw data alone holds limited value without the ability to extract, interpret, and present it in a meaningful way. Organizations and individuals are now inundated with vast amounts of data generated from various sources such as transactions, customer interactions, sensors, and digital platforms. However, without proper analysis and visualization, this raw data remains an untapped resource, often obscuring rather than revealing valuable insights.

This project aims to address the critical challenge of transforming complex, unstructured, or high-dimensional datasets into coherent and insightful information that can directly support strategic planning and operational improvements. By applying systematic data analysis methods-ranging from data cleaning and preprocessing to statistical analysis and pattern recognition-we seek to uncover hidden trends, correlations, and outliers that might otherwise go unnoticed.

In addition, effective data visualization techniques will be employed to communicate findings clearly and intuitively, enabling stakeholders to grasp patterns and make data-driven decisions with confidence. The goal is to bridge the gap between raw data and actionable knowledge, thereby empowering users with insights that are not only informative but also instrumental in driving innovation, efficiency, and competitive advantage in today's information-rich environment.

1.2 Objectives

The main objective of this project is to analyze and visualize the selected dataset (Code 0) and to uncover hidden patterns, trends, and relationships among key variables. This includes cleaning and preparing the data, performing exploratory data analysis (EDA), and presenting insights through clear and informative visualizations. The goal is to support better understanding and data-driven decision-making.

1.3 Research Methodology

This study adopts data-driven research methodology, focusing on quantitative analysis. The process involves collecting the dataset, cleaning and preprocessing the data to handle missing or inconsistent values, followed by exploration data analysis (EDA) to identify trends, outliers, and correlations. Visualization techniques are applied using tools such as Python (pandas, matplotlib, seaborn) and Tableau to effectively communicate insights.

1.4 Scope of the Study

The scope of this project is limited to analyzing the dataset labeled as Code 0. The analysis focuses on key variables relevant to sales, product performance, pricing, customer ratings, and profits. The study does not include predictive modeling or real-time analytics and is based solely on the available data within the dataset.

1.5 Research Subjects

The research focuses on products and sales data contained in the selected dataset. Key entities of interest include product categories, pricing, customer ratings, revenue, and profit margins. These elements are analyzed to understand patterns in customer behavior and product performance.

1.6 Structure of the Report and Contribution

This report is structured into five chapters:

- **Chapter 1**: Introduction to the problem, objectives, methodology, scope, and research subjects.
- **Chapter 2**: Theoretical Background.
- **Chapter 3**: Online Retail Dataset Analysis
- **Chapter 4**: Adult Income Dataset Analysis
- **Chapter 5**: Students Performance Analysis
- **Chapter 6**: Conclusion and suggestions for future work.

CHAPTER 2. THEORETICAL BACKGROUND

2.1 Exploratory Data Analysis (EDA)

Definition and Purpose: Exploratory Data Analysis (EDA) is an approach to analyze datasets, uncover patterns, and generate hypotheses using summary statistics and visualizations. It is critical in data science for understanding data before applying advanced models.

Key Components:

- **Data Summarization**: Assess dataset size (records, variables) and data types (numeric, categorical).
- **Handling Missing Data and Duplicates**: Use imputation (e.g., mean, median) for missing values and remove or flag duplicates.
- **Statistical Summaries**: Compute mean, median, standard deviation, and quartiles to describe data.
- **Visualization Techniques**: Histograms show distributions, box plots reveal outliers, and scatter plots explore relationships.

Figure 2.1: Example of EDA [1]

2.2 Descriptive Statistics

Definition: Descriptive statistics summarize a dataset's main features quantitatively.

Key Measures:

- **Central Tendency**: Mean (average), median (middle value), mode (most frequent).
- **Dispersion**: Variance (spread), standard deviation (dispersion from mean), range (max-min), IQR (middle 50%).
- **Distribution Shape**: Skewness (asymmetry), kurtosis (tailedness).

  Purpose: These measures provide insights into data structure and variability, aiding further analysis.

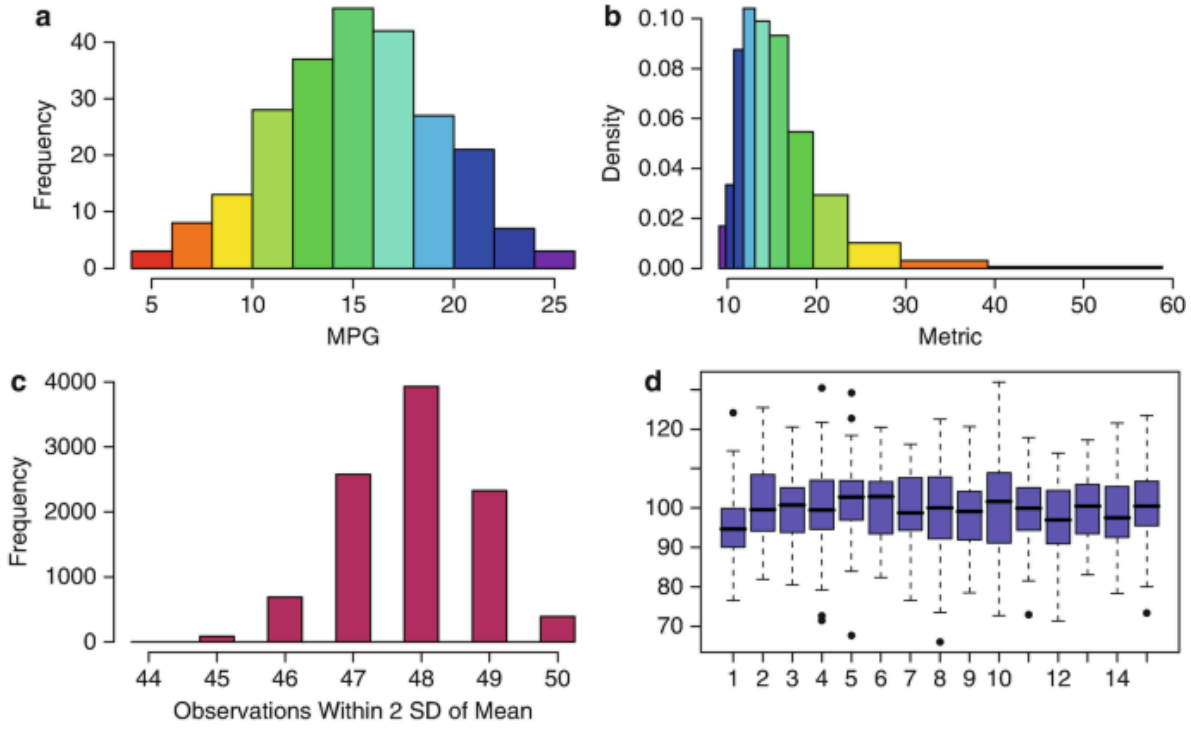


Figure 2.2: Example of Descriptive of Statistics [2]

Formulas:

[1] https://databrio.com

[2] https://www.anychart.com/blog/2018/09/07/interesting-data-graphics-warming-debt-commuting-china/

- **Mean**: $\bar{x} = \frac{\Sigma x_i}{n}$, where x_i is each value, n is sample size.
- **Median**: If n odd, middle value; if even, average of two middle values.
- **Variance**: $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$
- **Standard Deviation**: $s = \sqrt{s^2}$
- **IQR**: $IQR = Q_3 - Q_1$, where Q_1 and Q_3 are 25th and 75th percentiles.

Example: Suppose we measure the number of books 5 students read in a month: **1, 2, 2, 3, 20**.

- The **mean** is 5.6, but the **median** is only 2.
- This large gap shows that one student (who read 20 books) is pulling the average up.

  → This is an example of **skewed data** with **high variability**.

2.3 Data Visualization

Role of Visualization: Visualization transforms raw data into interpretable graphics, revealing trends and anomalies.

Common Visualization Types:

- **Histogram**: Displays frequency distribution of a variable.
- **Box Plot**: Highlights median, quartiles, and outliers.
- **Scatter Plot**: Shows relationships between two variables.

  Selection Rationale: Histograms suit continuous data, box plots detect outliers, and scatter plots assess correlations, chosen based on data type and goals.

Figure 2.3: Example of Data Visualization [3]

Example:

- **Histogram**: Amount showed a left-skewed distribution with most values low and few high outliers.
- **Box Plot**: UnitPrice revealed outliers beyond 2,800 after initial filtering.
- **Scatter Plot**: UnitPrice vs. Quantity showed no clear linear pattern, supporting weak Pearson correlation.

2.4 Outlier Detection and Handling

Definition: Outliers are data points significantly different from others, potentially skewing results.

Detection Methods:

- **Statistical Methods**: Z-score (>3 indicates outliers), IQR (values beyond 1.5×IQR).
- **Visualization**: Box plots and scatter plots visually flag anomalies.

  Handling Strategies: Remove outliers if erroneous, transform (e.g., log scale) if meaningful, or retain if contextually relevant.

[3] https://www.slingshotapp.io/blog/9-best-data-visualization-examples

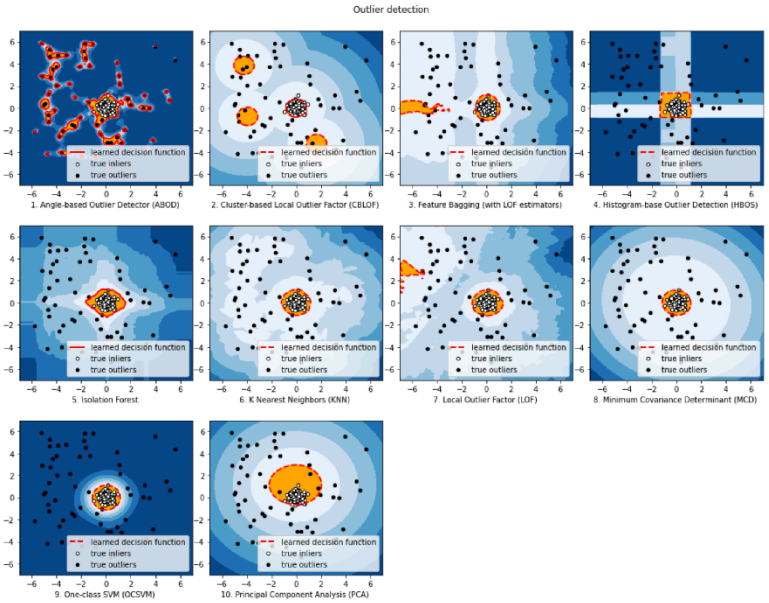


Figure 2.4: Example of Outlier Detection and Handling [4]

Formulas:

- **Z-score**: $z = \frac{x-\mu}{\sigma}$, where μ is mean, σ is standard deviation; $|z| > 3$ flags outliers.
- **IQR**: $Lower\ bound = Q_1 - 1.5 \times IQR$,

  $Upper\ bound = Q_3 + 1.5 \times IQR$.
- **Example**: Imagine a dataset tracking daily step counts. Most users record between 3,000 and 15,000 steps, but one entry shows **120,000 steps** in a day. This would be flagged as an outlier and either investigated or removed to avoid distorting analysis.

2.5 Probability Distributions

Overview: Probability distributions model the likelihood of data values, guiding statistical inference.

Common Distributions:

[4] https://www.mdpi.com/2079-9292/10/18/2236

- **Normal Distribution**: Bell-shaped, symmetric (mean = median = mode), common in natural phenomena.
- **Poisson Distribution**: Models count data (e.g., events per interval), discrete, skewed.
- **Exponential Distribution**: Models time between events, continuous, decreasing.

  Assessment Techniques: Histograms and Q-Q plots visually check fit; Shapiro-Wilk test confirms normality.

  Formulas:

- **Normal**: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- **Poisson**: $P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, where λ is the rate.
- **Exponential**: $f(x) = \lambda e^{-\lambda x},\ x \geq 0$.

  Example: Imagine you count how much money your friends spend on snacks.

  Most of them spend **just 1 or 2 dollars**, but **one friend spends 100 dollars**!

  That's not "normal" — it's like a **weird mountain with a super long tail on one side**.

  So we say: "This data is **not normal**. It's skewed!"

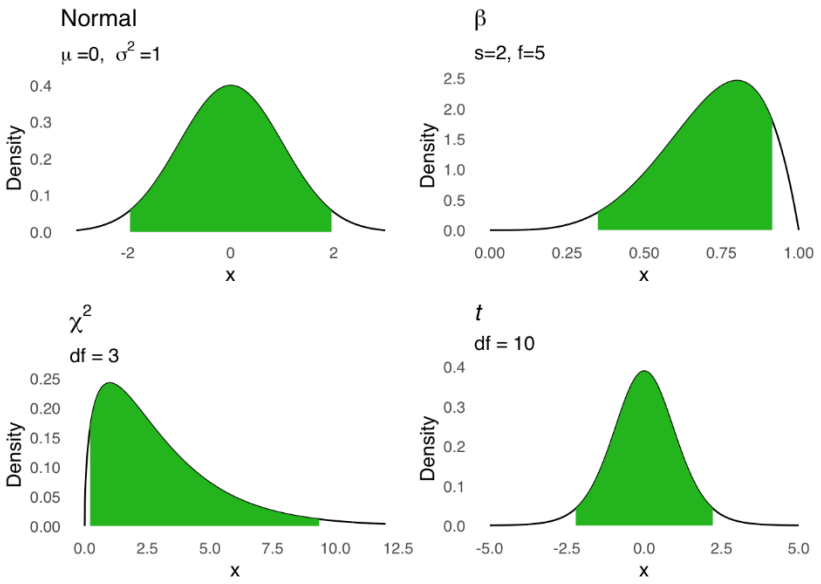


Figure 2.5: Example of Probability Distributions [5]

2.6 Hypothesis Testing

Concept: Hypothesis testing uses sample data to infer population properties, testing claims statistically.

Key Tests:

- **t-test**: Compares means (e.g., two groups), assumes normality, includes one-sample and two-sample variants.
- **Chi-square Test**: Tests independence (categorical data) or goodness of fit, assumes large samples.
- **ANOVA**: Compares means across multiple groups, assumes normality and equal variances.

  Steps in Hypothesis Testing: State null (H_0) and alternative (H_1) hypotheses, set α (e.g., 0.05), compute test statistic, and interpret p-value ($p < \alpha$ rejects H_0).

  Formulas:

- **t-test**: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, compares two means.

[5] https://bookdown.org/danbarch/psy_207_advanced_stats_I/probability-distributions.html

- **Chi-square**: $X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, where O_i is observed, E_i is expected.
- **ANOVA F-statistic**: $F = \frac{Between-group\ variance}{Within-group\ variance}$

Example: Let's say you check how much candy kids eat in three countries: **UK, Germany, and France**.

You notice that some groups eat **way more** than others.

So you ask: "Are they really that different, or is it just random?"

You run a **fairness test** (called ANOVA), and the test says:

"Yup! Big difference here. The countries **don't eat the same amount**."

Then you check just **UK vs. Germany**, and again the test says:

"Germany eats **a lot less** candy than the UK!"

That's what hypothesis testing does — it helps us know when **differences are real**, not just lucky guesses.

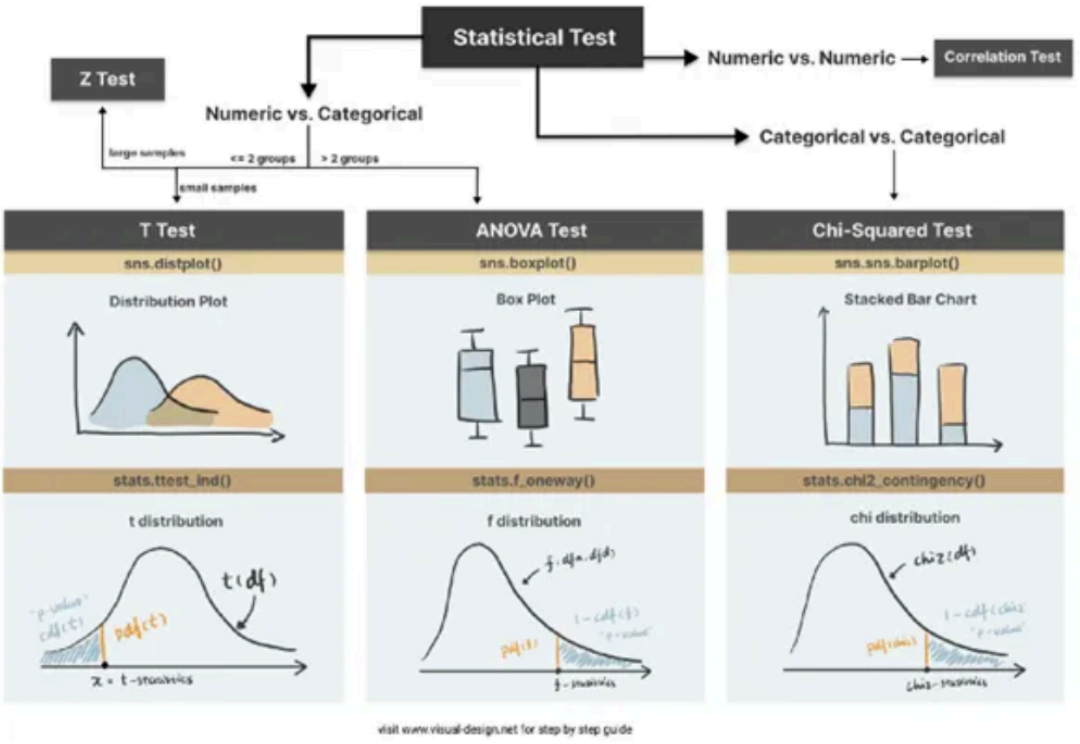


Figure 2.6: Example of Hypothesis Testing [6]

2.7 Correlation Analysis

Definition: Correlation measures the strength and direction of relationships between variables.

Correlation Coefficients:

- **Pearson**: Assesses linear relationships, requires normality, from -1 to 1.
- **Spearman**: Assesses monotonic relationships, non-parametric, suitable for ordinal or non-normal data.

  Visualization: Scatter plots show pairwise relationships; heatmaps summarize multiple correlations.

  Interpretation: Strong correlations (e.g., |r| > 0.7) suggest practical dependencies, informing decision-making.

  Formulas:

- **Pearson**: $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y - \bar{y})^2}}$.

- **Spearman**: $p = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$, where d_i is the difference in ranks.

  Example: Imagine you're selling toys. You notice something:

When you make the price **higher**, kids buy **fewer** toys.

When the price is **lower**, they buy **more**.

But it's not always perfectly clear — sometimes it changes just a little bit.

So you check with two "smart tools":

- One tool (Pearson) says: “Hmm, the link is very small.”
- The other tool (Spearman) says: “Yeah, when price goes up, quantity **usually** goes down.”

---

[6] https://medium.com/@hasninemirza/guidance-to-hypothesis-testing-in-python-t-test-anova-chi-squared-test-b2446ce44030

So in short: **Expensive toys = fewer sales** (but not a perfect rule).

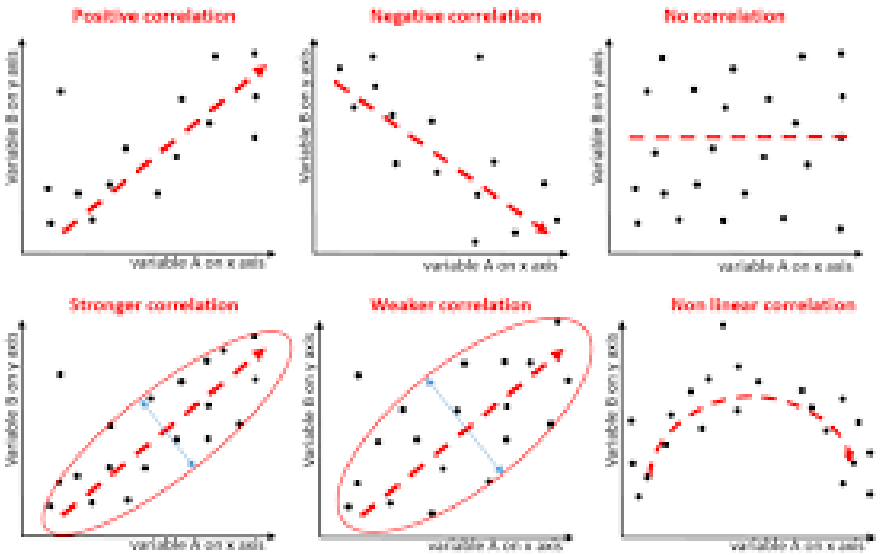


Figure 2.7: Example of Correlation Analysis [7]

[7] https://www.mdpi.com/1660-4601/15/12/2907

CHAPTER 3. ONLINE RETAIL DATASET ANALYSIS

3.1 Introduction

This chapter presents the implementation and analysis of the Online Retail Dataset using Python. The dataset, sourced from a UK-based online retailer, contains transactional data from December 2010 to December 2011. The goal is to perform Exploratory Data Analysis (EDA) to understand customer purchasing behavior, identify trends, and prepare the data for further statistical analysis. The code is explained step-by-step, followed by an interpretation of the results Table 3.1.

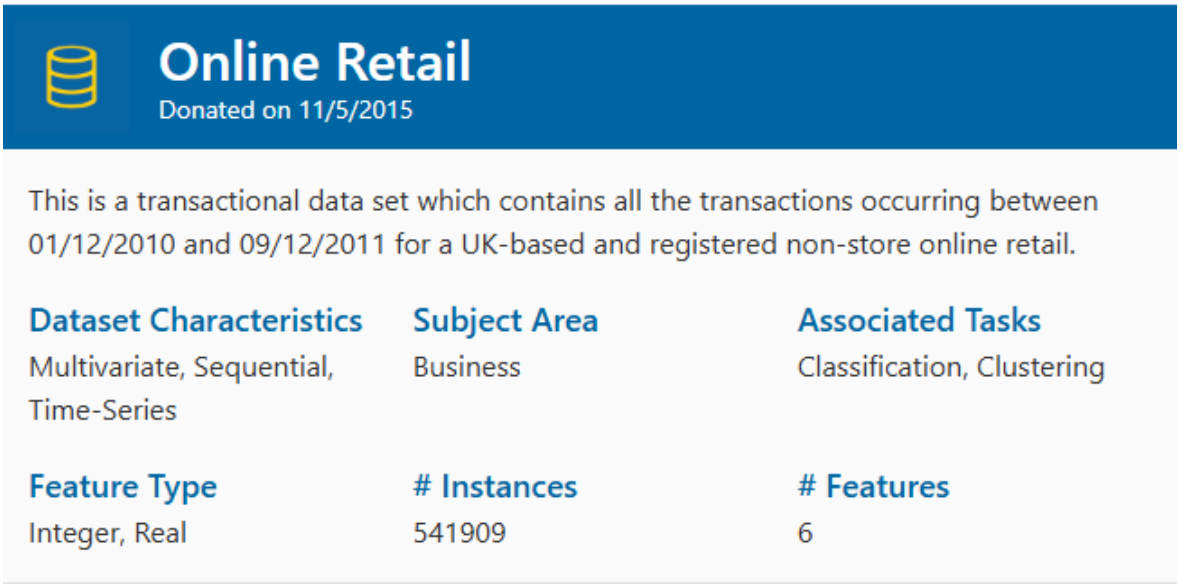


Figure 3.1: Online Retail Data Card [8]

Table 3.1: Online Retail Dataset Overview

InvoiceNo	StockCode	…	CustomerID	Country
536365	85123A		17850.0	United Kingdom
536365	71053		17850.0	United Kingdom

[8] https://archive.ics.uci.edu/dataset/352/online+retail

InvoiceNo	StockCode	…	CustomerID	Country
536365	84406B	…	17850.0	United Kingdom
536365	84029G		17850.0	United Kingdom
536365	84029E		17850.0	United Kingdom

3.2 Exploratory Data Analysis (EDA)

3.2.1 Data Structure and Types

The dtypes attribute checks the data types of each column to ensure proper formatting for analysis in Table 3.2.

Total entries: 541,909 rows

Total columns: 8

Memory usage: ~33.1 MB

- InvoiceNo, StockCode, Description, InvoiceDate, and Country are strings (object), suitable for categorical or text data.
- Quantity is an integer (int64), and UnitPrice and CustomerID are floats (float64), aligning with numerical analysis needs.
- InvoiceDate should ideally be converted to a datetime type for time-based analysis (not implemented here but recommended).

Table 3.2: DataFrame Structure Overview

Column	Non-Null Count	Data Type
InvoiceNo	541,909	object
StockCode	541,909	object
Description	540,455	object
Quantity	541,909	int64

Column	Non-Null Count	Data Type
InvoiceDate	541,909	object
UnitPrice	541,909	float64
CustomerID	406,829	float64
Country	541,909	object

3.2.2 Missing Data Assessment

The missing values per column to identify data gaps.

- **Description**: 1,454 missing values (0.27% of rows). This is minor, as StockCode can still identify products.
- **CustomerID**: 135,080 missing values (24.9% of rows), significant as it limits customer-specific analysis. These may represent unregistered or one-time buyers.
- Other columns are complete, ensuring robust analysis of quantities, prices, and dates.

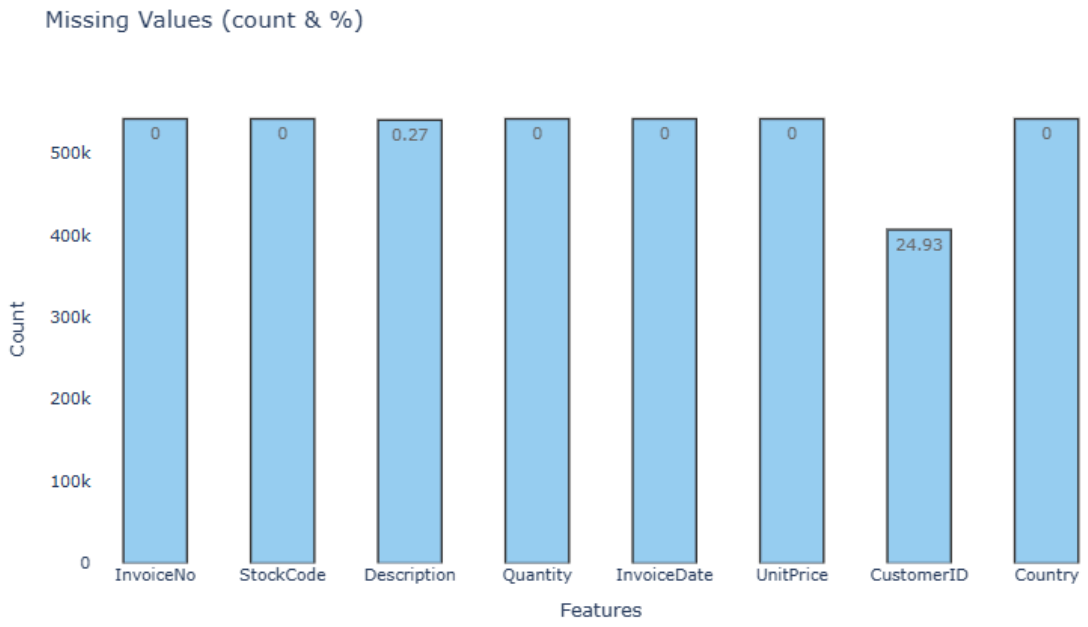


Figure 3.2: Online Retail Missing Values (count & %)

A bar plot was generated to visualize the count and percentage of missing values across features, using the missing_plot function with InvoiceNo as the key. This confirmed significant missing data in CustomerID.

Handling Methods

Two approaches were considered:

1. **Forward and Backward Fill (ffill/bfill)**: Applied to impute missing values, retaining all columns and rows. Post-processing shape: (536,641, 9).
2. **Drop Rows with Missing Values**: Rows containing any missing data were removed, reducing the dataset to 406,829 rows and 8 columns (threshold: 50% missing values per column, i.e., 270,954.5).

Table 3.3: Missing Values Before and After Processing

Column	Missing Values (Before)	Missing Values (After)
InvoiceNo	0	0
StockCode	0	0
Description	1,454	0
Quantity	0	0
InvoiceDate	0	0
UnitPrice	0	0
CustomerID	135,080	0
Country	0	0

The cleaned dataset shape is (406.829, 8), ready for further analysis.

3.2.3 Descriptive Statistics

The describe() method computes summary statistics (count, mean, std, min, quartiles, max) for numerical columns in Table 3.4.

Table 3.4: Descriptive Statistics of Selected Numerical Columns

Statistic	Quantity	UnitPrice	CustomerID
Count	536,641.00	536,641.00	536,641.00
Mean	9.62	4.63	15,267.78
Std Dev	219.13	97.23	1,738.42
Min	-80,995.00	-11,062.06	12,346.00
25%	1.00	1.25	13,784.00
50%	3.00	2.08	15,144.00
75%	10.00	4.13	16,794.00
Max	80,995.00	38,970.00	18,287.00

Interpretation:

- **Quantity**:
  - Mean: 9.55 items per transaction, but median (3) suggests most purchases are small.
  - Std: 218.08 indicates high variability.
  - Min (-80,995) and Max (80,995) suggest returns or data errors (negative values) and bulk orders.
- **UnitPrice**:
  - Mean: £4.61, median: £2.08, showing most items are inexpensive.
  - Std: 96.76 reflects wide price variation.
  - Min (-£11,062.06) and Max (£38,970) indicate anomalies (negative prices likely errors).
- **CustomerID**:
  - Count: 406,829 (confirms missing data).
  - Range: 12,346 to 18,287 (~6,000 unique customers).

Observations:

- Negative values in Quantity and UnitPrice suggest returns or data entry issues, requiring cleaning.

- High standard deviations and extreme min/max values indicate outliers, likely bulk transactions or errors.
- Most transactions involve low quantities and prices, but rare large orders skew the mean.

3.2.4 Handling Duplicated Data

After Forward/Backward Fill (ffill/bfill)

- **Duplicate Count**: 10,147 duplicate records identified using df.duplicated(keep=False).
- **Details**: Duplicates sorted by all columns and inspected (first 10 rows displayed).
- **Action**: Duplicates removed with drop_duplicates(), reducing dataset to (401.604, 8).

```
duplicated_rows = df[df.duplicated(keep=False)]
print("\nSố lượng bản ghi trùng lặp:", duplicated_rows.shape[0])
```

```
Số lượng bản ghi trùng lặp: 10147
```

```
print("\nChi tiết các cột bị trùng lặp:")
duplicated_rows.sort_values(by=list(df.columns)).head(10)
```

```
Chi tiết các cột bị trùng lặp:
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
494	536409	21866	UNION JACK FLAG LUGGAGE TAG	1	2010-12-01 11:45:00	1.25	17908.0	United Kingdom
517	536409	21866	UNION JACK FLAG LUGGAGE TAG	1	2010-12-01 11:45:00	1.25	17908.0	United Kingdom
485	536409	22111	SCOTTIE DOG HOT WATER BOTTLE	1	2010-12-01 11:45:00	4.95	17908.0	United Kingdom
539	536409	22111	SCOTTIE DOG HOT WATER BOTTLE	1	2010-12-01 11:45:00	4.95	17908.0	United Kingdom
489	536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	2010-12-01 11:45:00	2.10	17908.0	United Kingdom
527	536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	2010-12-01 11:45:00	2.10	17908.0	United Kingdom
521	536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	2010-12-01 11:45:00	2.95	17908.0	United Kingdom
537	536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	2010-12-01 11:45:00	2.95	17908.0	United Kingdom
578	536412	21448	12 DAISY PEGS IN WOOD BOX	1	2010-12-01 11:49:00	1.65	17920.0	United Kingdom
598	536412	21448	12 DAISY PEGS IN WOOD BOX	1	2010-12-01 11:49:00	1.65	17920.0	United Kingdom

```
# Xóa bản ghi trùng lặp (nếu cần)
df = df.drop_duplicates()
print("\nDữ liệu sau khi loại bỏ trùng lặp:", df.shape)
```

```
Dữ liệu sau khi loại bỏ trùng lặp: (401604, 8)
```

Figure 3.3: Online Retail Data After Handling Duplicated Data

After Dropping Missing Values (dropna)

- **Duplicate Count**: 10,062 duplicate records identified.
- **Details**: Duplicates sorted and reviewed (first 10 rows displayed).
- **Action**: Duplicates removed, resulting in a final shape of (401,604, 8).

```
duplicated_rows = df[df.duplicated(keep=False)]
print("\nSố lượng bản ghi trùng lặp:", duplicated_rows.shape[0])
```

```
Số lượng bản ghi trùng lặp: 10062
```

```
print("\nChi tiết các cột bị trùng lặp:")
duplicated_rows.sort_values(by=list(df.columns)).head(10)
```

```
Chi tiết các cột bị trùng lặp:
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
494	536409	21866	UNION JACK FLAG LUGGAGE TAG	1	2010-12-01 11:45:00	1.25	17908.0	United Kingdom
517	536409	21866	UNION JACK FLAG LUGGAGE TAG	1	2010-12-01 11:45:00	1.25	17908.0	United Kingdom
485	536409	22111	SCOTTIE DOG HOT WATER BOTTLE	1	2010-12-01 11:45:00	4.95	17908.0	United Kingdom
539	536409	22111	SCOTTIE DOG HOT WATER BOTTLE	1	2010-12-01 11:45:00	4.95	17908.0	United Kingdom
489	536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	2010-12-01 11:45:00	2.10	17908.0	United Kingdom
527	536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	2010-12-01 11:45:00	2.10	17908.0	United Kingdom
521	536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	2010-12-01 11:45:00	2.95	17908.0	United Kingdom
537	536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	2010-12-01 11:45:00	2.95	17908.0	United Kingdom
578	536412	21448	12 DAISY PEGS IN WOOD BOX	1	2010-12-01 11:49:00	1.65	17920.0	United Kingdom
598	536412	21448	12 DAISY PEGS IN WOOD BOX	1	2010-12-01 11:49:00	1.65	17920.0	United Kingdom

```
# Xóa bản ghi trùng lặp (nếu cần)
df = df.drop_duplicates()
print("\nDữ liệu sau khi loại bỏ trùng lặp:", df.shape)
```

```
Dữ liệu sau khi loại bỏ trùng lặp: (401604, 8)
```

Figure 3.4: Online Retail Data After Handling Duplicated Data

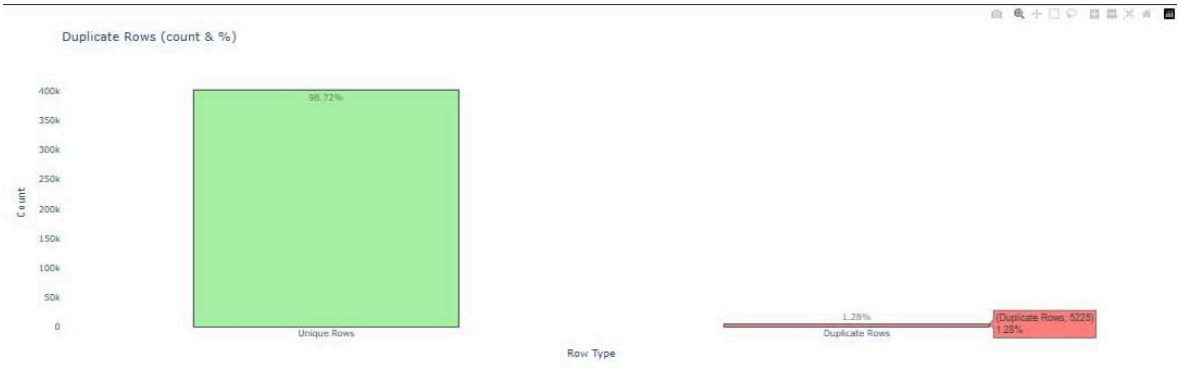


Figure 3.5: Before Drop Duplicated Values

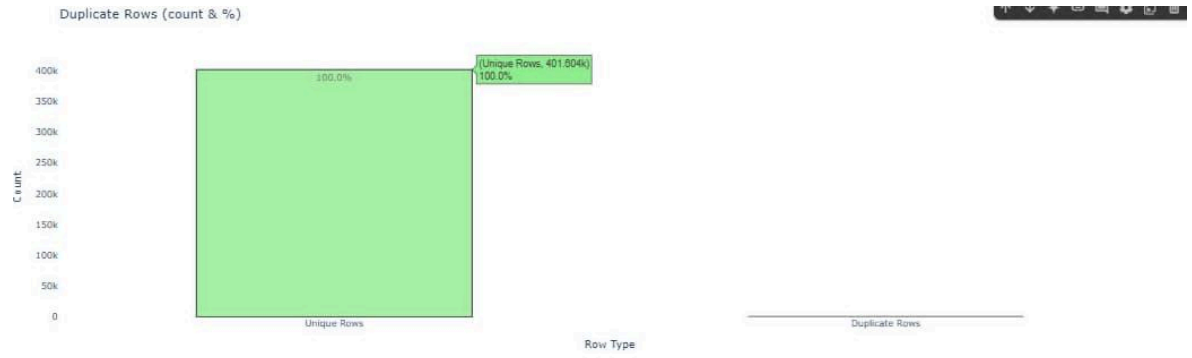


Figure 3.6: After Drop Duplicated Values

Unique Products

- **After ffill/bfill**: 4,223 unique products (df["Description"].nunique()).
- **After dropna**: 3,896 unique products, indicating a reduction due to row removal.

Date Range

- **Start Date**: 2010-12-01 08:26:00 (df["InvoiceDate"].min()).
- **End Date**: 2011-12-09 12:50:00 (df["InvoiceDate"].max()).

3.2.5 Descriptive Statistical Analysis

After Forward/Backward Fill (ffill/bfill)

Summary statistics for numerical columns (df.describe()):

- **Quantity**:
  - Mean: 9.62, Std: 219.13, Min: -80,995, Max: 80,995
  - Q1: 1, Median: 3, Q3: 10
- **UnitPrice**:
  - Mean: 4.63, Std: 97.23, Min: -11,062.06, Max: 38,970
  - Q1: 1.25, Median: 2.08, Q3: 4.13
- **CustomerID**:
  - Mean: 15,267.78, Std: 1,738.42, Min: 12,346, Max: 18,287
  - Q1: 13,784, Median: 15,144, Q3: 16,794

Table 3.5: Online Retail Statistics for Selected Columns

Statistic	Quantity	UnitPrice	CustomerID
Count	536,641	536,641	536,641
Mean	9.62	4.63	15,267.78
Std Dev	219.13	97.23	1,738.42
Min	-80,995.00	-11,062.06	12,346.00
25%	1.00	1.25	13,784.00
50%	3.00	2.08	15,144.00
75%	10.00	4.13	16,794.00
Max	80,995.00	38,970.00	18,287.00

3.2.6 Top 10 Most Preferred Products

After ffill/bfill: 4,223 unique products identified. Top 10 include "White Hanging Heart T-Light Holder" and "Regency Cakestand 3 Tier" (visualized via bar plot).

After dropna: 3,896 unique products. Top 10 remain consistent (e.g., "White Hanging Heart T-Light Holder").

Reasons for Popularity:

- Competitive pricing aligned with perceived value.
- High demand for daily essentials.
- Bulk packaging (e.g., "Pack of 72 Retrospot Cake Cases").
- Aesthetic appeal and emotional resonance.

Average Price Analysis: Bar plot of top 10 products' average UnitPrice shows affordable pricing drives sales.

Recommendations: Promote popular items with discounts and ensure sufficient stock for essentials.

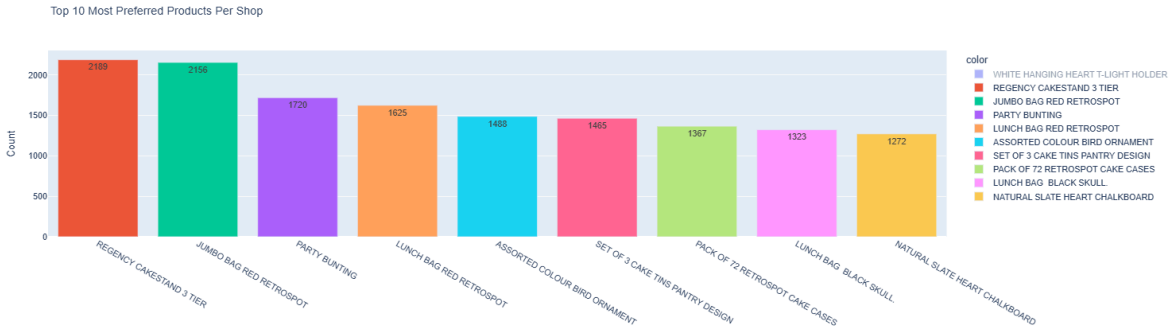


Figure 3.7: Online Retail before using Dropna

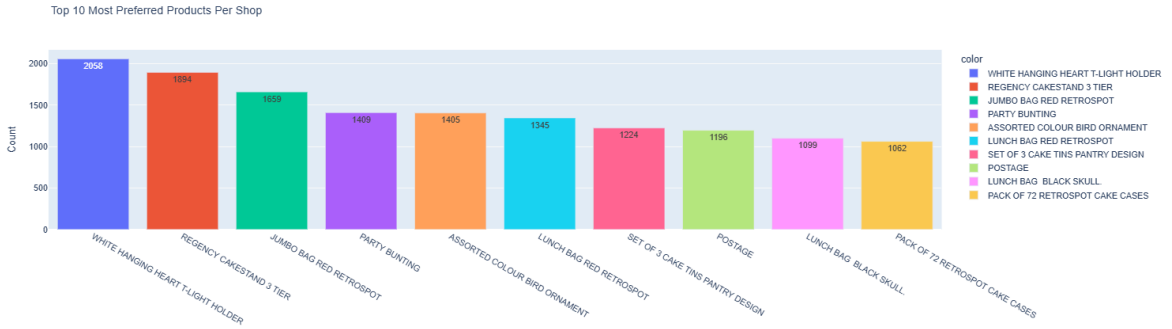
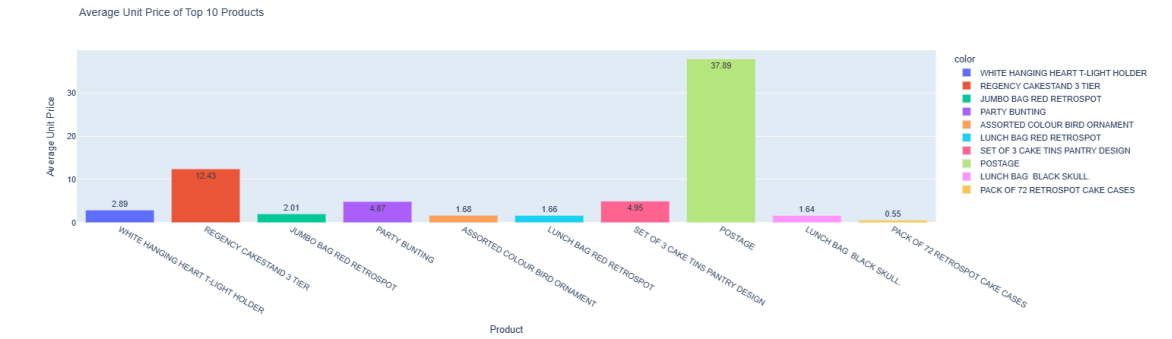


Figure 3.8: Online Retail after using Dropna



- Tôi đề xuất có những đợt khuyến mãi những món phổ biến
- Những đồ này là đồ thiết yếu , nên tập trung nhập hàng nhiều

Figure 3.9: Online Retail average UnitPrice

3.2.7 Top 10 Least Preferred Products

🎬 **After dropna**: Least popular include “Crochet Lilac/Red Bear Keyring” and “Midnight Blue Crystal Drop Earrings” (bar plot visualization).

🎬 **Price Analysis**: Unit prices of least preferred items vary, suggesting low demand rather than pricing issues.

🎬 **Recommendations**: Discontinue non-essential low-demand items; focus marketing on seasonal sales (e.g., holidays) and cautious restocking.

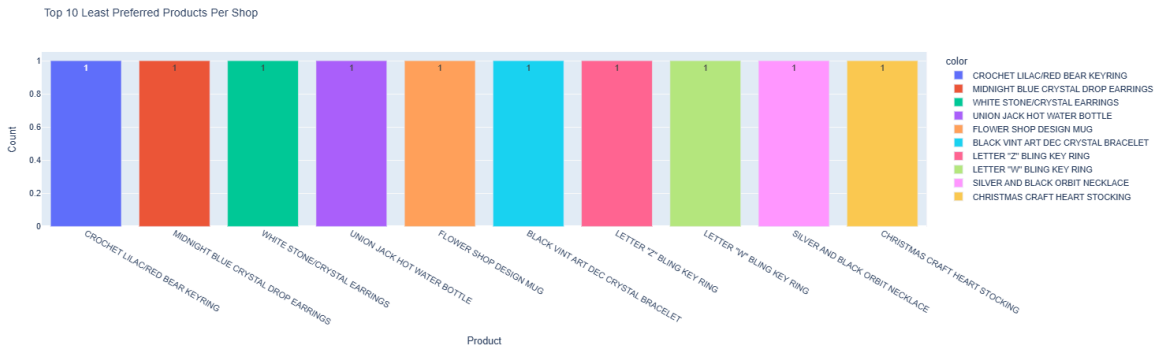


Figure 3.10: Online Retail Top 10 Least Preferred Products per shop

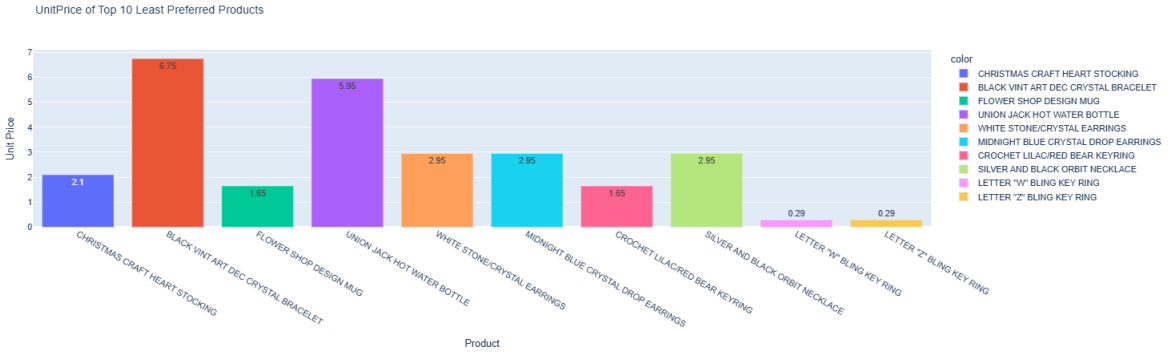


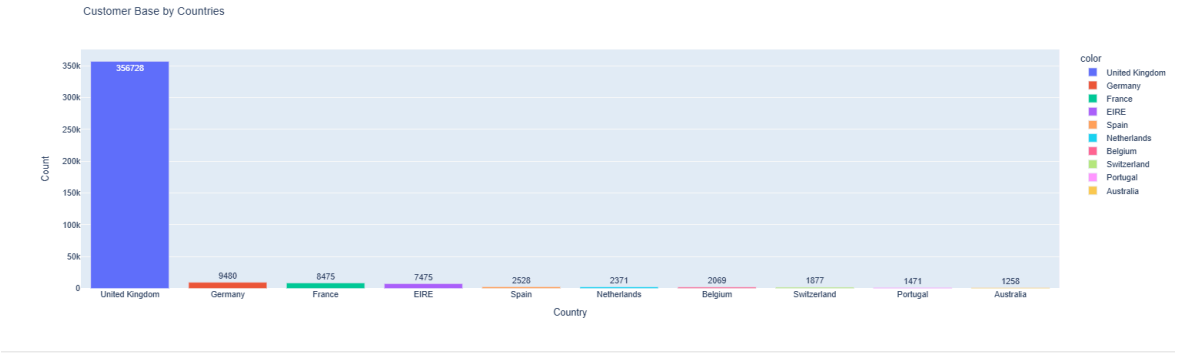
Figure 3.11: Online Retail Unit Price Top 10 Least Preferred Products

3.2.8 Top 10 Countries by Customer Base

🎬 **After ffill/bfill and dropna**: UK dominates (visualized via bar plot), followed by scattered global presence.

🎬 **Insight**: Majority of sales concentrated in the UK.

🎬 **Recommendation**: Strengthen UK market strategies while exploring growth in other regions.



- Tập trung nhiều ở UK
- Đang tập trung cơ sở ở UK
- Nhưng vẫn có rải tác ở khắp nơi
- Cần đẩy mạnh những chiến lược phát triển trong nước UK vì lượng mua tập trung ở đây

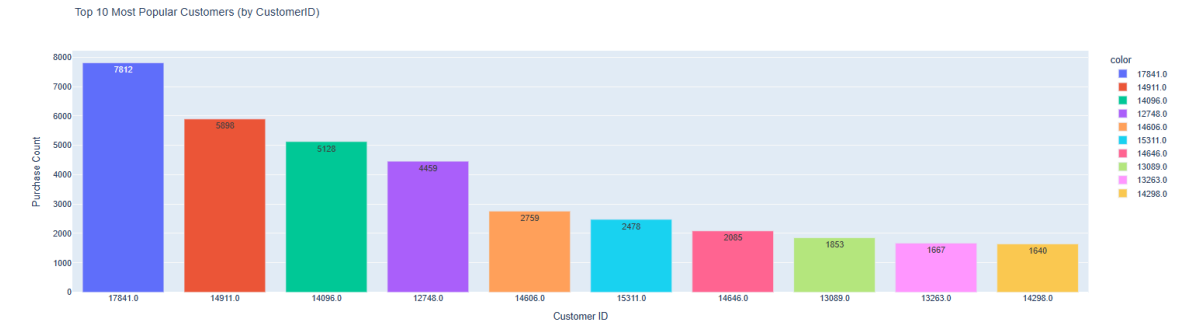
Figure 3.12: Top 10 Countries by Customer Base

3.2.9 Top 10 Frequent Customers

Analysis: Top 10 CustomerIDs identified (bar plot), with 22,190 unique invoices total.

Proposal: Introduce a loyalty program:

- Rank customers by monthly purchases.
- Offer tiered discounts (e.g., higher tiers get larger bill reductions).
- Provide premium support, feedback opportunities, thank-you gifts, and birthday rewards for top tiers.



- Vì là online chúng ta nên phân chia cấp bậc rank cho khách hàng.
- Tôi đề xuất mua bao nhiêu sản phẩm 1 tháng sẽ tương ứng với thứ bậc , thứ bậc đó giúp ích phần trăm giảm giá bill.
- Nâng cao chăm sóc cho khách hàng hạng bậc cao , quan tâm nhu cầu feedback của khách.
- Có những quà tặng tri ân, và quà sinh nhật thể hiện sự quan tâm của mình đối với khách.

Figure 3.13: Top 10 Frequent Customers

3.2.10 Outlier Detection and Handling

- **Initial Statistics**:

- Quantity: Mean: 9.55, Std: 218.08, Min: -80,995, Max: 80,995.
- UnitPrice: Mean: 4.61, Std: 96.76, Min: -11,062.06, Max: 38,970.

- **Numeric Columns Analyzed**: Quantity, UnitPrice (excluded CustomerID).
- **IQR Results**:
  - Quantity: 58,619 outliers.
  - UnitPrice: 39,627 outliers.
- **Z-Score Results**:
  - Quantity: 346 outliers (e.g., -9,360, 80,995).
  - UnitPrice: 374 outliers (e.g., 569.77, 1,714.17).

Table 3.6: Descriptive Statistics of Selected Numerical Columns

Statistic	Quantity	UnitPrice	CustomerID
Count	541,909.00	541,909.00	406,829.00
Mean	9.55	4.61	15,287.69
Std Dev	218.08	96.76	1,713.60
Min	-80,995.00	-11,062.06	12,346.00
25%	1.00	1.25	13,953.00
50%	3.00	2.08	15,152.00
75%	10.00	4.13	16,791.00
Max	80,995.00	38,970.00	18,287.00

Table 3.7: DataFrame Structure Overview

Column	Non-Null Count	Data Type
InvoiceNo	541,909	object
StockCode	541,909	object
Description	540,455	object
Quantity	541,909	int64
InvoiceDate	541,909	object
UnitPrice	541,909	float64

Column	Non-Null Count	Data Type
CustomerID	406,829	float64
Country	541,909	object

Table 3.8: Outliers in Quantity Column

Index	Quantity
4287	-9360
…	…
540421	80995
540422	-80995

Total Outliers in Quantity: 346 rows

Table 3.9: Outliers in UnitPrice Column

Index	UnitPrice
1814	569.77
…	…
540908	933.17
541540	1714.17

Total Outliers in UnitPrice: 374 rows

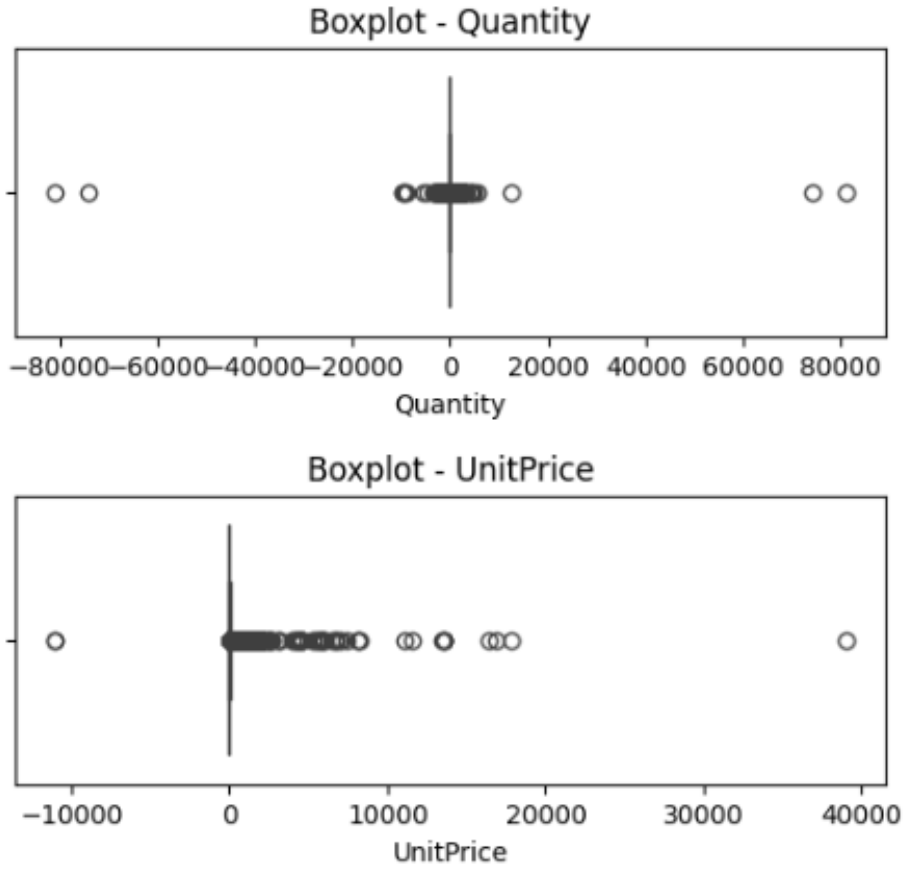


Figure 3.14: Online Retail Boxplot of Quantity and UnitPrice

Handling Approach

- **Manual Filtering**:
  - Step 1: Removed extreme values (Quantity: 0 to 10,000; UnitPrice: 0 to 10,000).
  - Step 2: Further refined (Quantity: 0 to 3,500; UnitPrice: 0 to 2,800).
  - Result: Post-filtering boxplots showed no visible outliers.
- **Proposed Improvement**: Use IQR-based removal programmatically:
  - Flag outliers with flag_outliers_iqr.
  - Remove outliers with remove_outliers_iqr for logical, repeatable filtering.

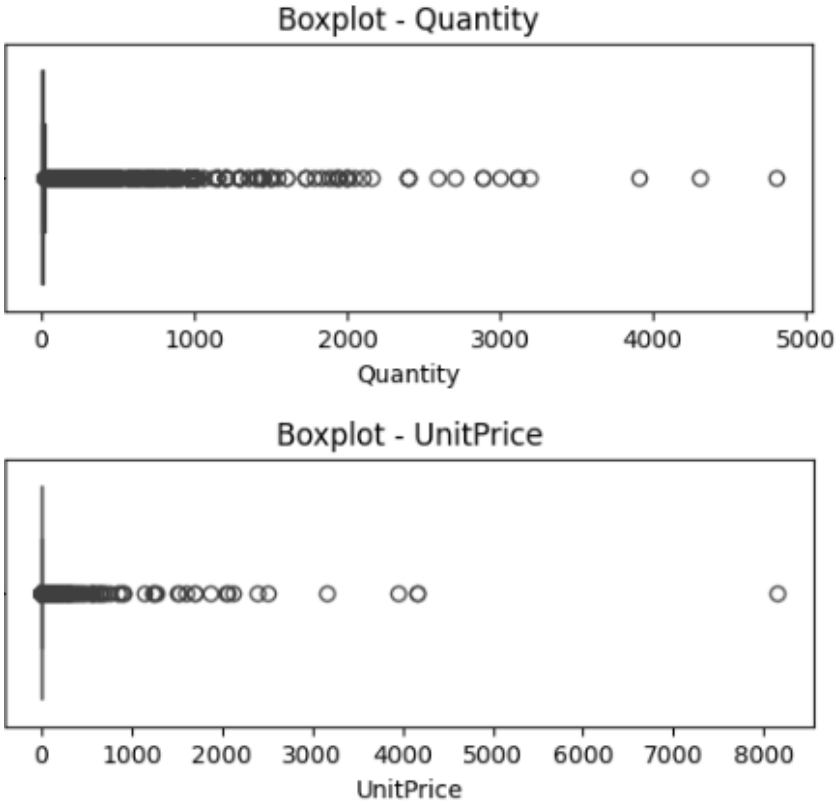


Figure 3.15: Online Retail Data After Handling Outliers_1

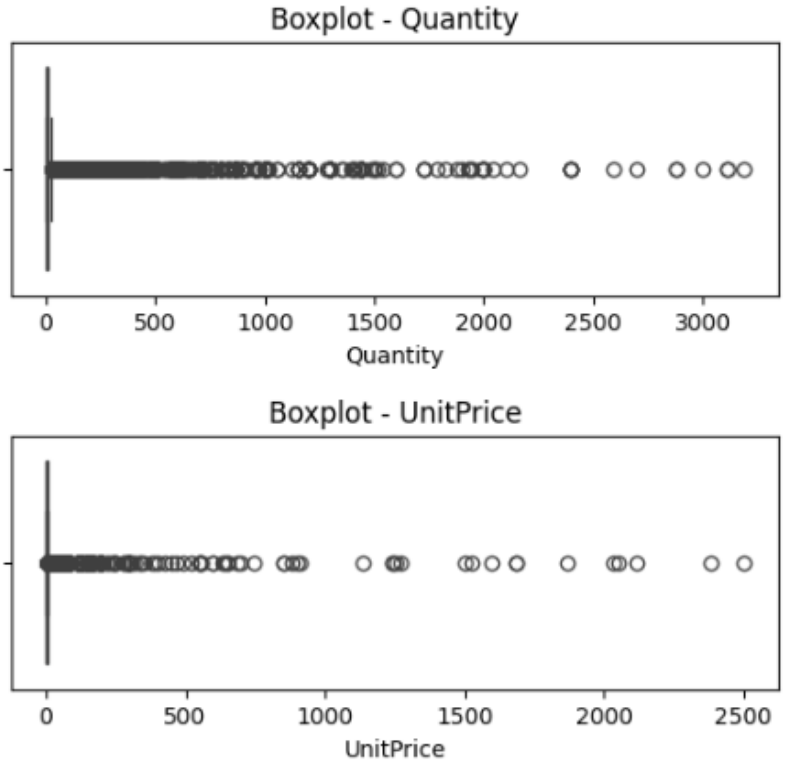


Figure 3.16: Online Retail Data After Handling Outliers_2

General Observations on the Dataset

1. **Shopping Trends**:
    - Popular products (e.g., “White Hanging Heart T-Light Holder”) reflect high consumer demand, while least popular items (e.g., “Crochet Lilac/Red Bear Keyring”) suggest poor fit or marketing.
2. **Customer Focus**:
    - Frequent buyers identified, enabling targeted loyalty programs and personalized offers.
3. **Market Potential**:
    - UK dominates the customer base, indicating a strong domestic market but opportunities for global expansion.
4. **Analytical Benefits**:
    - Insights into top products and customers optimize inventory, marketing, and resource allocation, supporting data-driven decisions and future trend predictions.

3.3 Probability Distribution Analysis

Variable Selection

- **Chosen Variable**: Amount = UnitPrice × Quantity.
- **Rationale**: Reflects actual revenue, providing insight into online sales patterns.

Distribution Assessment

- **Method**: Analyzed a random sample of 5,000 records for computational efficiency.

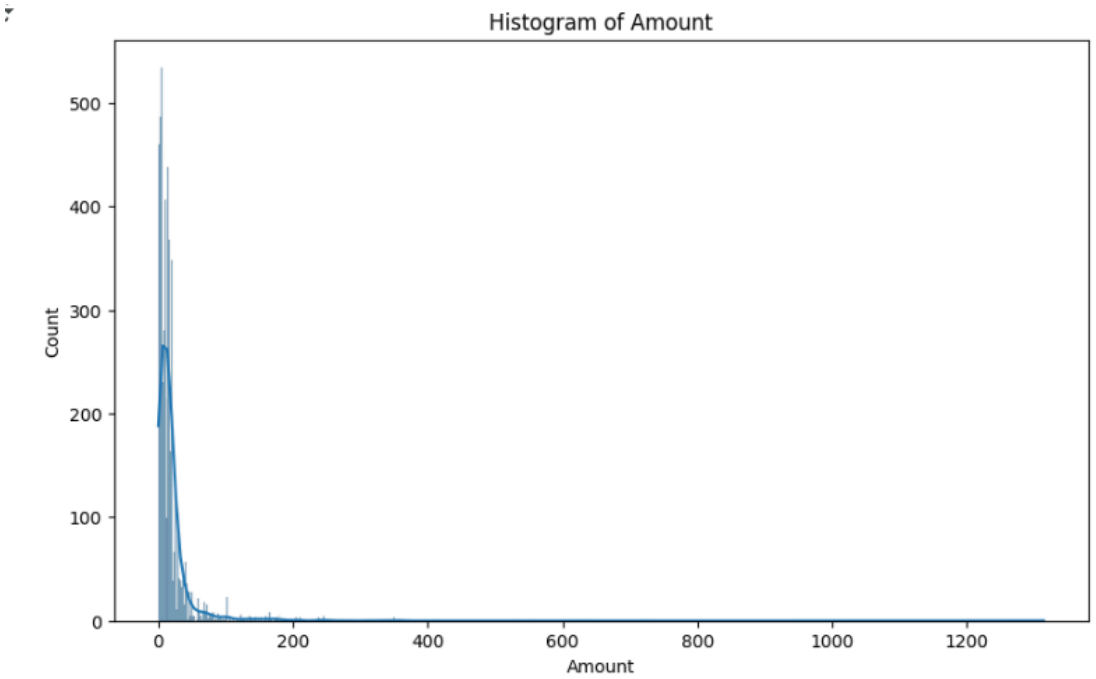


Figure 3.17: Online Retail Histogram of Amount

Statistical Tests:

1. **Shapiro-Wilk Test (Normal Distribution)**:
   - Statistic: 0.331, p-value: ~1.4e-86.
   - Conclusion: p-value < 0.05 → Reject H_0; Amount does not follow a normal distribution.
2. **Kolmogorov-Smirnov Test (Exponential Distribution)**:
   - KS Statistic: 0.159, p-value: ~5.96e-112.
   - Conclusion: p-value < 0.05 → Reject H_0; Amount does not follow an exponential distribution.
3. **Poisson Distribution**:
   - Not applicable: Amount is continuous (monetary values), while Poisson applies to discrete counts. Additionally, Amount includes decimals, unlike Poisson's integer requirement.

Observations

- **Skewness**: Strong left skew, typical in retail data—most transactions are small, with rare bulk purchases causing outliers.
- **Real-World Reflection**: Matches online sales reality: frequent low-value orders and occasional high-value wholesale transactions.
- **Implications**: Non-normal distribution suggests avoiding parametric methods (e.g., t-tests, linear regression) without preprocessing.

Recommendations

- **Data Transformation**: Apply log-transformation to reduce skewness and approximate normality if needed.
- **Analytical Approach**: Use non-parametric methods (e.g., Spearman correlation, Wilcoxon tests) or machine learning models that do not assume normality.
- **Next Steps**: Prefer Spearman correlation for subsequent correlation analysis (Section 5) due to non-normal distribution.

3.4 Hypothesis Testing

a. Statistical Tests Overview

1. **t-test**: Compares means of two groups.
    - Example: Difference in average Amount between UK and Germany customers.
    - Condition: Near-normal data or large sample size (Central Limit Theorem).
2. **Chi-square Test**: Tests dependence between categorical variables.
    - Example: Do cancellation rates vary by country?
3. **ANOVA**: Compares means across more than two groups.
    - Example: Difference in Amount across UK, Germany, and France.

b. Research Question and Application

- **Hypothesis**: “Is there a difference in average order value (Amount) across customer groups from different countries?”

- **Methodology**:
    - Calculated Amount = UnitPrice × Quantity.
    - Filtered: Positive Amount, countries (UK, Germany, France), and removed outliers (Amount ≤ 35,000).
    - Tests:
        1. **ANOVA**: Compared Amount across UK, Germany, and France.
        2. **t-test**: Compared UK vs. Germany (Welch's, unequal variances).
        3. **Chi-square**: Tested dependence between country and AmountLevel (High/Low, split by median).

```
ANOVA test: F = 20.3565, p-value = 1.4447e-09
T-test (UK vs DE): t = -12.8402, p-value = 1.6643e-37
Chi-square test: Chi2 = 5702.3192, p-value = 0.0000e+00, dof = 2
```

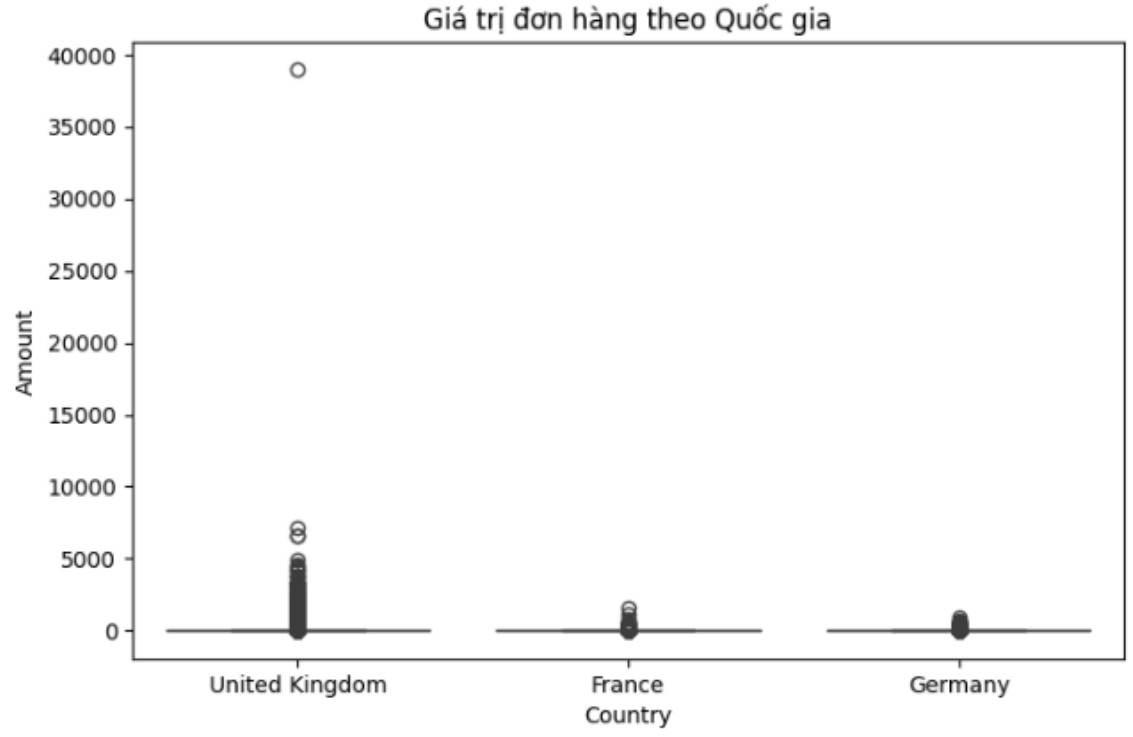


Figure 3.18: Online Retail UnitPrice of each Countries

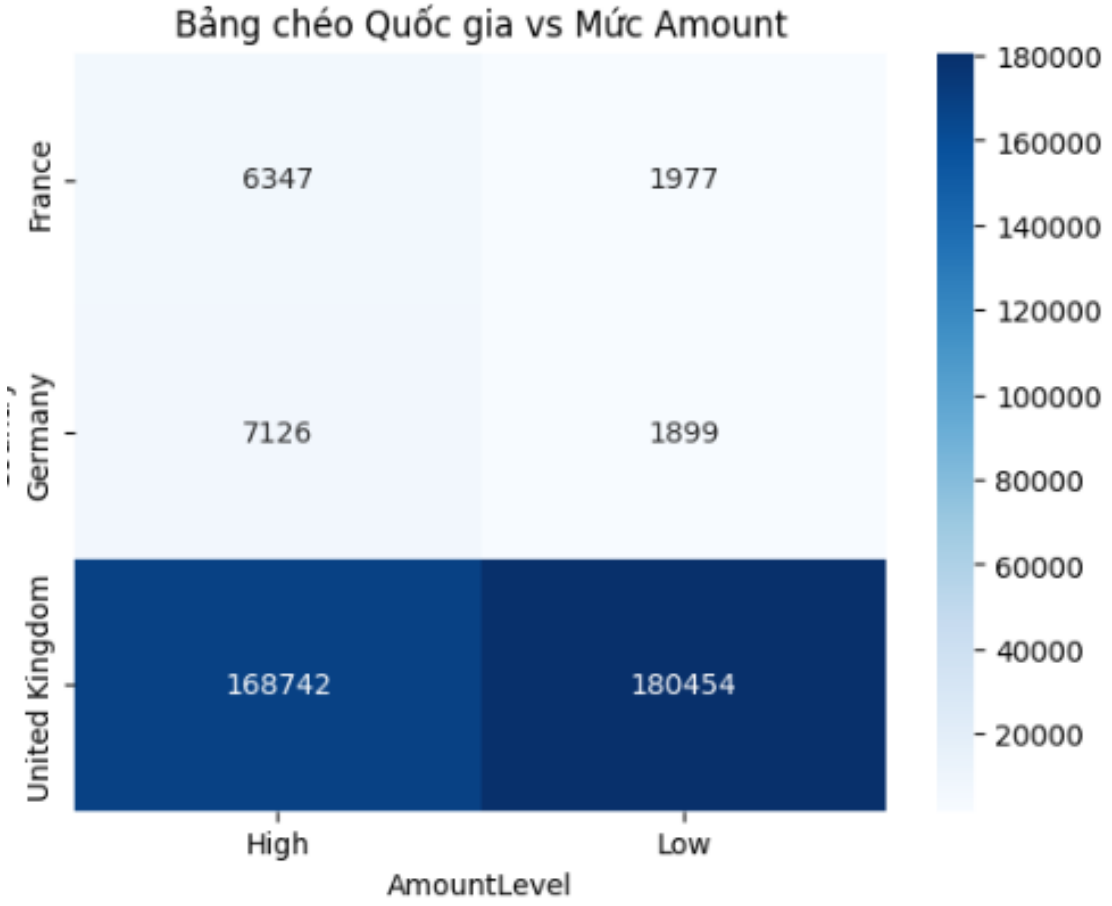


Figure 3.19: Online Retail Cross table of national prices with amount

c. Results and Interpretation

- **ANOVA**:
  - F = 20.3565, p-value = 1.4447e-09.
  - Interpretation: Significant difference in average Amount across the three countries ($p < 0.05$). UK customers likely spend more than those in Germany or France.
- **t-test (UK vs. Germany)**:
  - t = -12.8402, p-value = 1.6643e-37.
  - Interpretation: Strong evidence of a difference in average Amount ($p < 0.05$), with UK customers showing higher spending than Germany.
- **Chi-square**:
  - Chi2 = 5702.3192, p-value = 0.0000e+00, dof = 2.

- Interpretation: Strong dependence between country and AmountLevel ($p < 0.05$). UK customers are more likely to have “High” order values compared to others.

3.5 Correlation Analysis

a. Correlation Coefficients Overview

1. **Pearson Correlation**: Measures linear relationships between variables, ranging from -1 (strong negative) to 1 (strong positive). Zero indicates no linear relationship.
    - Example: Pearson of 0.8 between Quantity and Amount suggests a strong positive linear link.
2. **Spearman Correlation**: Measures monotonic relationships, suitable for non-linear or non-normal data, also ranging from -1 to 1.
    - Example: Spearman of 0.7 indicates a moderate positive monotonic relationship.

b. Application to Dataset

- **Variables**: UnitPrice, Quantity.
- **Results**:
    - **Pearson**: UnitPrice vs. Quantity = -0.0312 (near-zero, weak linear relationship).
    - **Spearman**: UnitPrice vs. Quantity = -0.4147 (moderate negative monotonic relationship).

Table 3.10: Pearson Correlation Matrix

	UnitPrice	Quantity
UnitPrice	1.000000	-0.031215
Quantity	-0.031215	1.000000

Table 3.11: Spearman Correlation Matrix

	UnitPrice	Quantity
UnitPrice	1.000000	-0.414734
Quantity	-0.414734	1.000000

d. Practical Insights

- **Observation**: Weak Pearson correlation (-0.0312) suggests UnitPrice has minimal linear impact on Quantity. Moderate Spearman correlation (-0.4147) indicates a non-linear trend: higher prices may reduce quantities sold, though not strongly.
- **Implication**: Other factors (e.g., product type, customer preferences) likely drive sales volume more than price alone.

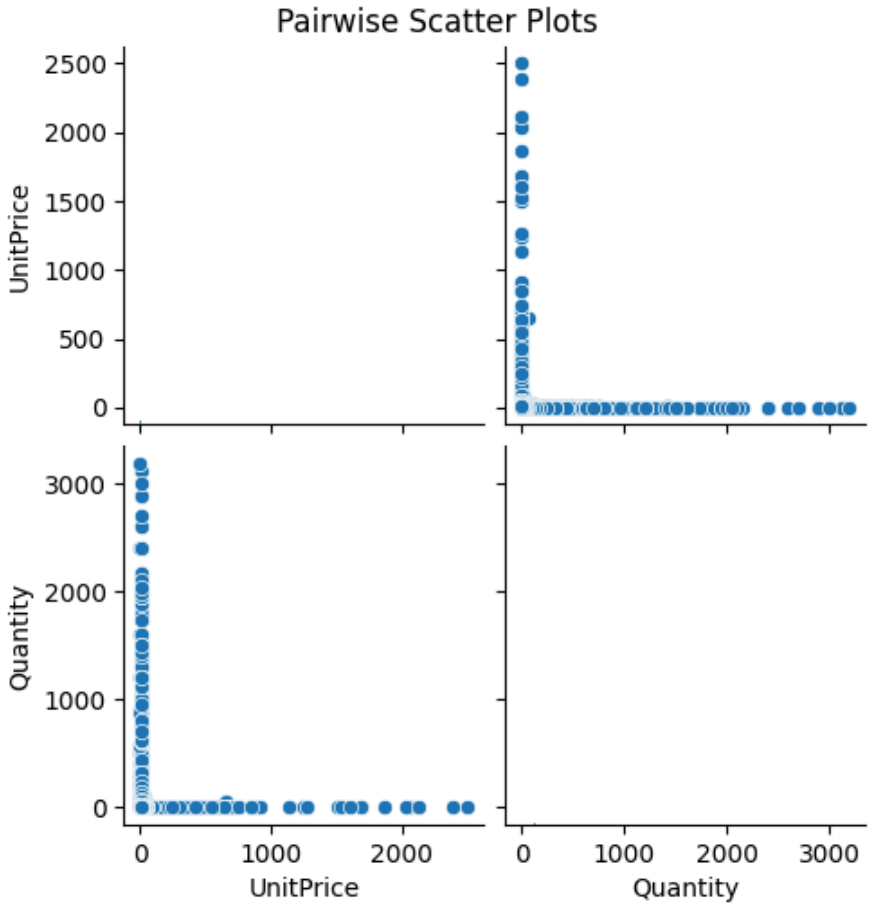


Figure 3.20: Online Retail Pairwise Scatter Plots

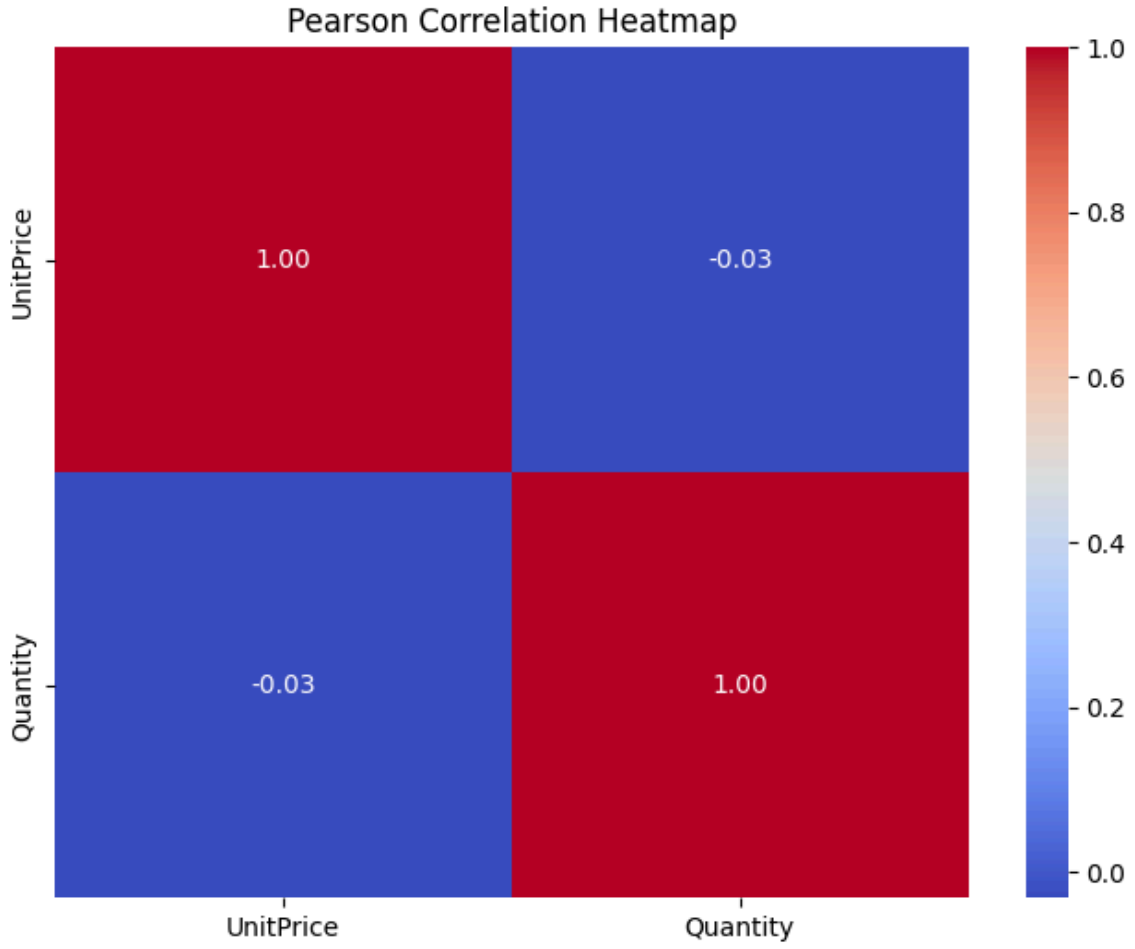


Figure 3.21: Online Retail Pearson Correlation Heatmap

3.6 Enhanced Customer Behavior Analysis (RFM)

RFM Definition:

- **Recency**: Days since last purchase (lower = more recent).
- **Frequency**: Number of purchases (higher = more frequent).
- **Monetary**: Total spending (higher = more valuable).

Process: Assigned scores (1-5) based on quantiles, combined into segments (e.g., "55" = champions), and mapped to categories (e.g., "hibernating").

Table 3.12: Invoice Date and Amount

InvoiceDate	Amount
2010-12-01 08:26:00	139.12
2010-12-01 08:28:00	22.20

InvoiceDate	Amount
…	…
2011-12-09 12:31:00	329.05
2011-12-09 12:49:00	339.20
2011-12-09 12:50:00	249.45

Trends (via line plots):

- Sales peak mid-week (e.g., Wednesday) and 10:00-15:00 daily.
- Monthly sales vary, suggesting seasonality.

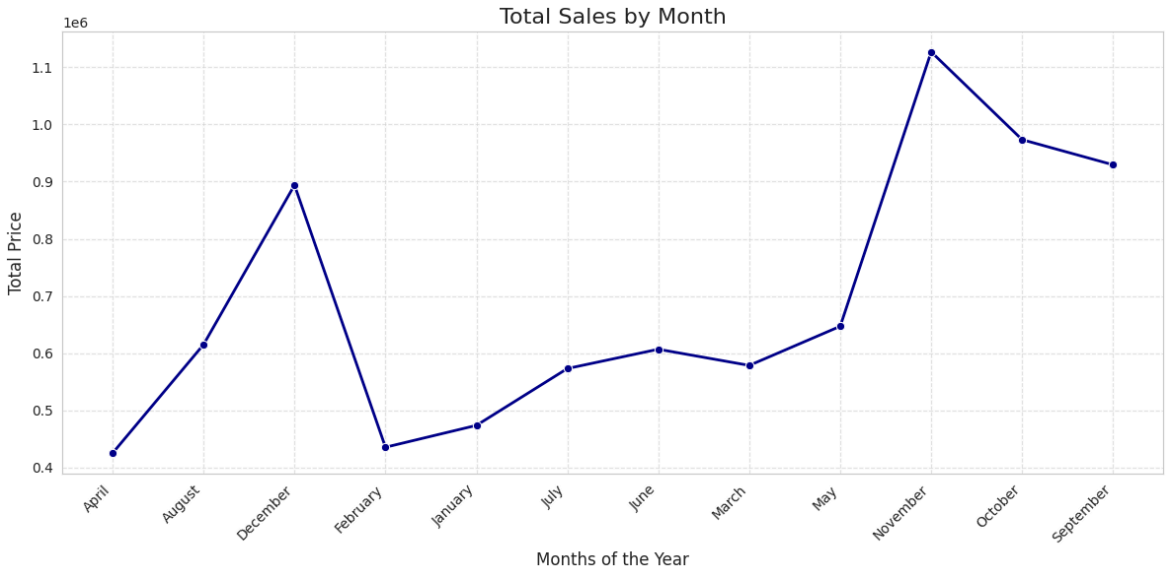


Figure 3.22: Online Retail Total Sales by Month

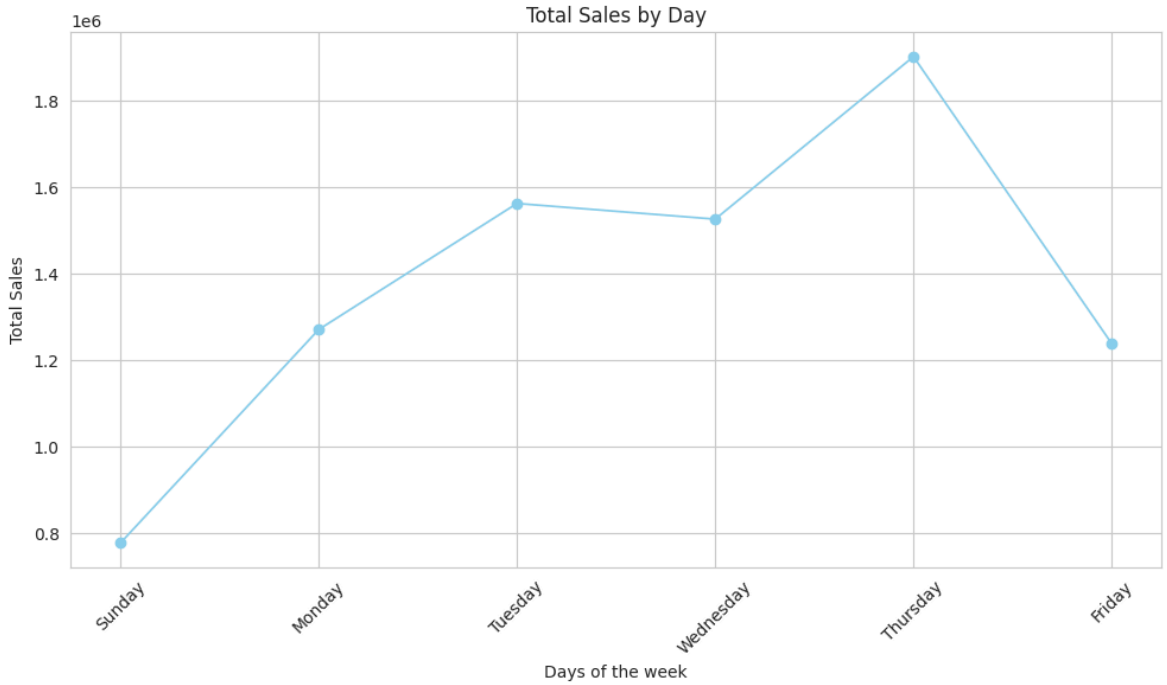


Figure 3.23: Online Retail Total Sales by Day

Segment Statistics

- **Hibernating**: Recency: 217.9 days, Frequency: 1.1, Monetary: $486.69.
- **At Risk**: Recency: 155.1 days, Frequency: 3.1, Monetary: $1,072.89.
- **New Customers**: Recency: 6.9 days, Frequency: 1, Monetary: $385.02.
- **Potential Loyalists**: Recency: 17.1 days, Frequency: 2, Monetary: $1,030.11.
- **Champions**: High recency, frequency, and spending.

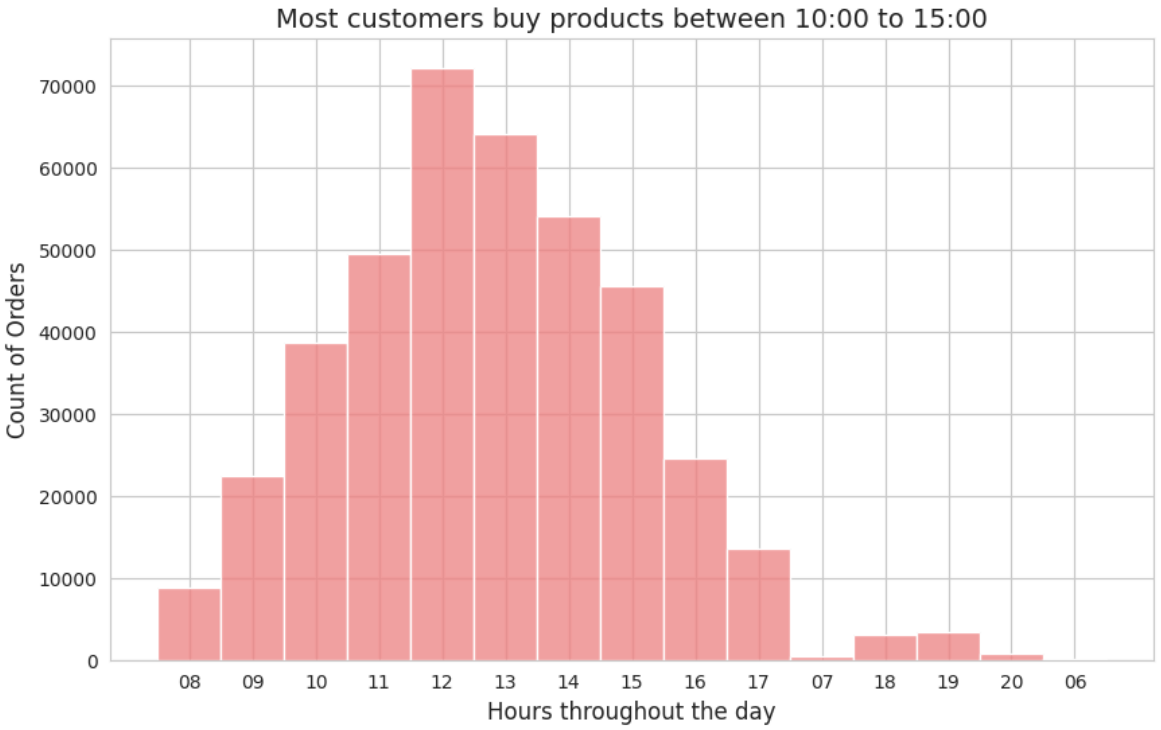


Figure 3.24: Online Retail Histogram Most Customers buy Products between 10:00 to 15:00

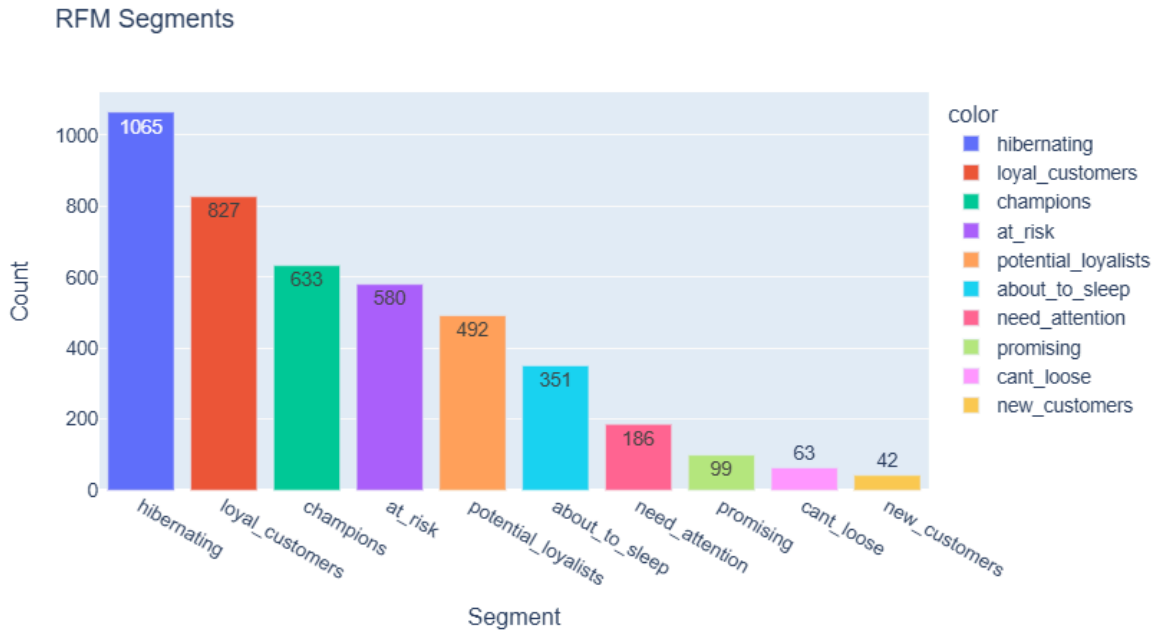


Figure 3.25: Online Retail RFM Segments

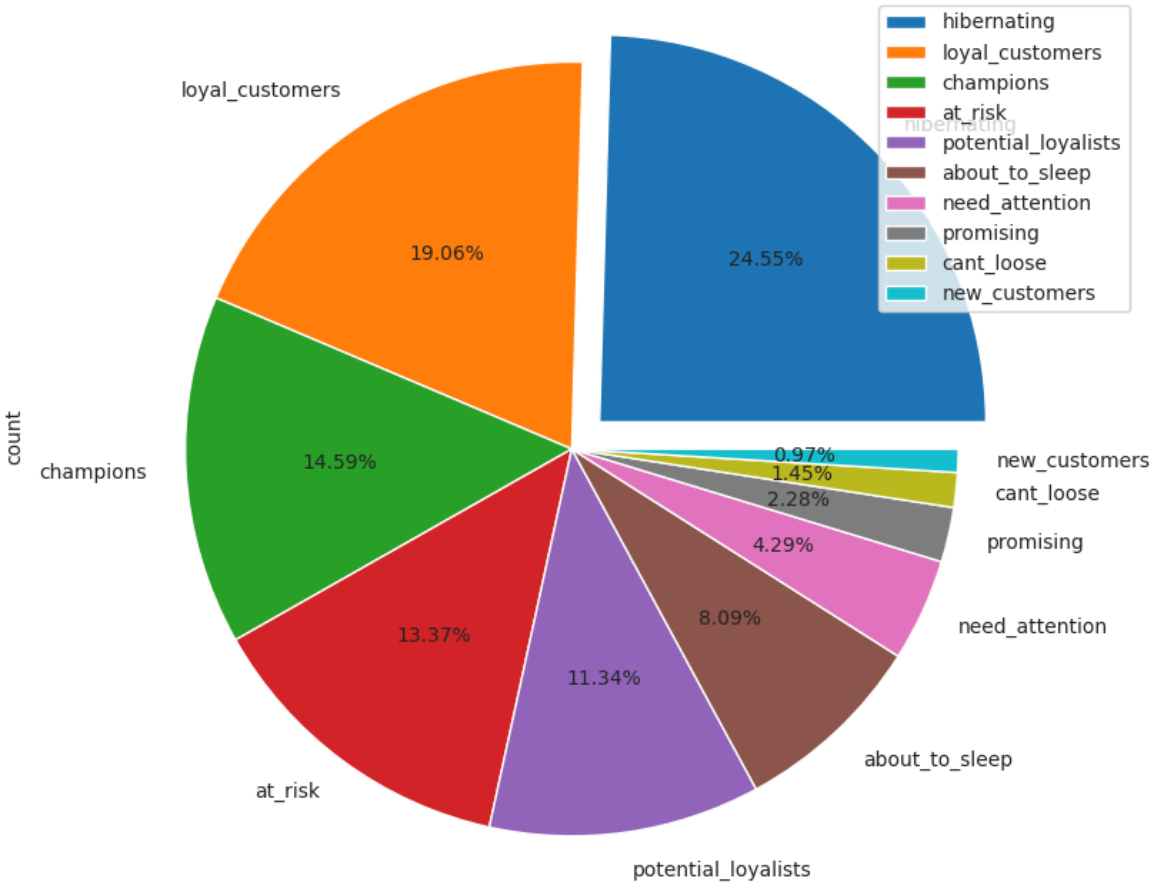


Figure 3.26: Online Retail Pie Chart of Customers Segmentation

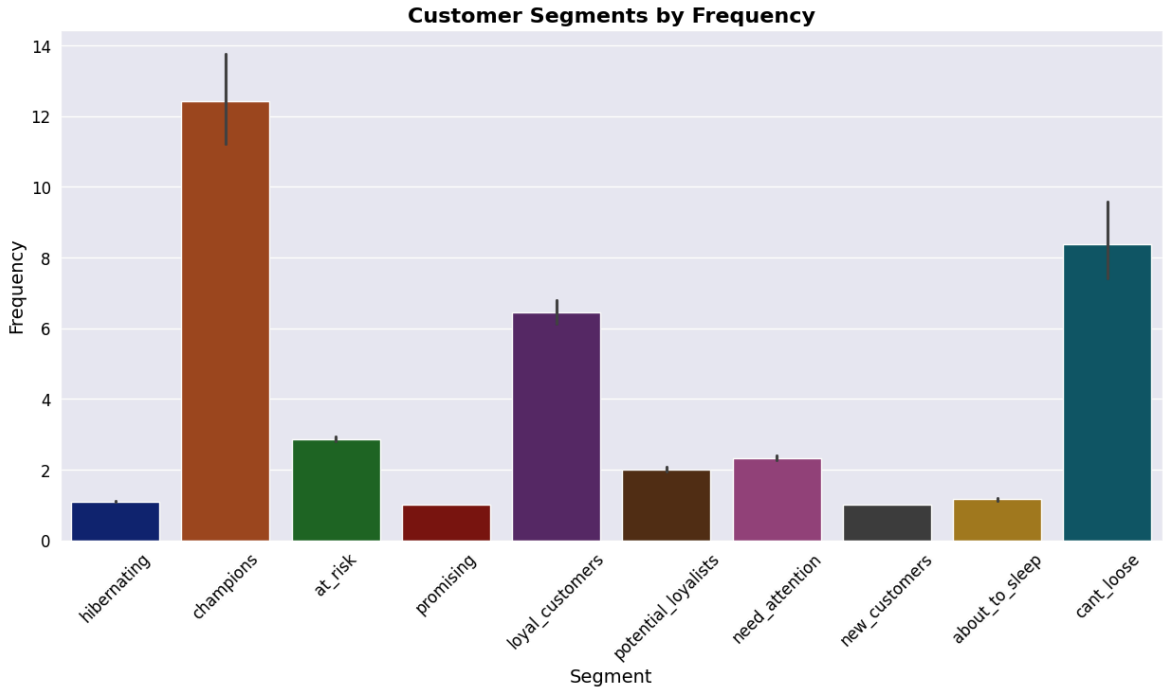


Figure 3.27: Online Retail Customer Segments by Frequency

Recommendations

1. **Hibernating**: Personalized win-back offers, surveys to re-engage.
2. **At Risk**: Targeted re-engagement campaigns, discounts.
3. **New Customers**: Welcome discounts, onboarding emails.
4. **Potential Loyalists**: Referral bonuses, personalized recommendations.
5. **Loyal/Champions**: VIP programs, exclusive offers.

CHAPTER 4. ADULT INCOME DATASET ANALYSIS

4.1 Dataset Overview

4.1.1 Dataset Source and Description

The dataset used in this project is the **Adult Income Dataset** (also known as **Census Income**), originally sourced from the **UCI Machine Learning Repository**.

- **Link**: https://archive.ics.uci.edu/ml/datasets/adult
- **Number of records**: ~32,000
- **Number of attributes**: 15

This dataset is widely used in classification problems and is considered a benchmark in **income prediction tasks**. The goal is to predict whether a person earns **more than $50K per year** based on demographic and employment-related features.

Data was accessed via the **Kaggle API** using kagglehub and loaded into the environment using pandas.

4.1.2 Attribute Descriptions

The dataset includes **15 features** that capture personal and socio-economic details. Key attributes include:

- **age**: Age of the individual
- **workclass**: Type of employment (e.g., Private, Self-emp)
- **education**: Highest level of education achieved
- **education-num**: Numeric representation of education level
- **marital-status**: Marital status
- **occupation**: Type of job
- **relationship**: Relationship status in household
- **race** and **sex**: Demographic information
- **capital-gain** and **capital-loss**: Financial details
- **hours-per-week**: Weekly working hours

- **native-country**: Country of origin
- **income** (target): Binary label – whether income is <=50K or >50K

4.1.3 Object of the Analysis

The main objective of this analysis is to **explore and understand the factors that influence an individual's income level**. Specifically, the project aims to:

- Identify how **demographic variables** such as **age, gender, education, and occupation** correlate with income.
- Use **statistical analysis and visualizations** to uncover **patterns and trends** in income distribution.
- Provide **insights that can inform policymaking** and **human resource planning**, particularly in areas like education, labor allocation, and gender equality.

By applying **Exploratory Data Analysis (EDA)** techniques, this study supports **data-driven decisions** in socio-economic domains.

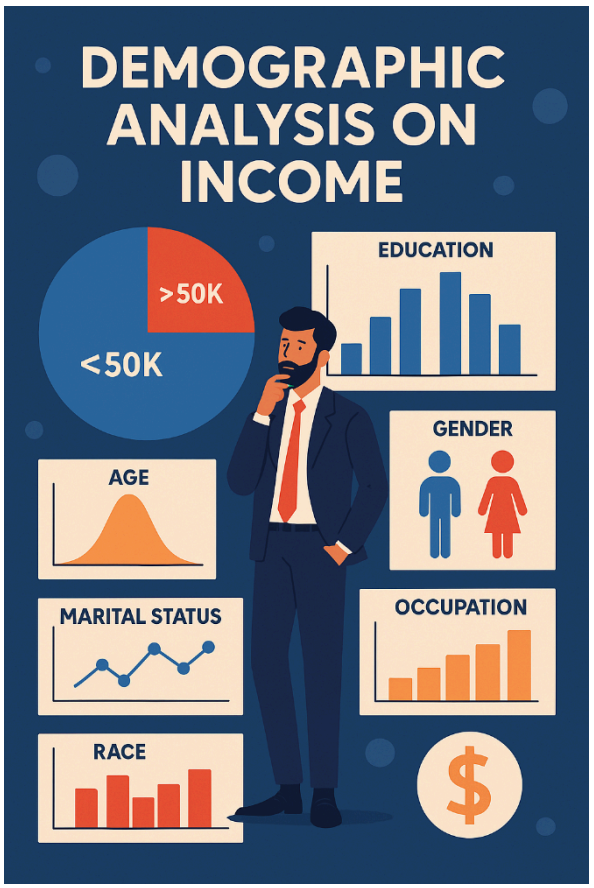


Figure 4.1: Object of the Analysis

4.2 Exploratory Data Analysis (EDA)

4.2.1 General Information on the Dataset

The dataset contains **48,842 records** and **15 attributes**, including demographic and income-related information of individuals. These variables consist of:

- **6 numerical features**: age, fnlwgt, educational-num, capital-gain, capital-loss, and hours-per-week.
- **9 categorical features**: such as workclass, education, marital-status, occupation, gender, and income.

  Each entry is a profile of an individual surveyed by the **U.S. Census Bureau**.

Note on fnlwgt: This feature represents the final statistical weight assigned to individuals. While it is essential for statistical inference, it is generally **not used for predictive modeling** to avoid bias.

Table 4.1: Demographic Data and Income

Age	Workclass	…	Native-country	Income
25	Private		United - States	≤50K
38	Private		United - States	≤50K
28	Local-gov	…	United - States	>50K
44	Private		United - States	>50K
18	?		United - States	≤50K

Table 4.2: Data Types and Information

#	Column	Non-Null Count	Dtype
0	age	48,842	int64
1	workclass	48,842	object
…			
13	native-country	48,842	object
14	income	48,842	object

Total Rows: 48,842

Total Columns: 15

Memory Usage: 5.6 MB

4.2.2 Handling Missing and Duplicate Values

Missing Values:

No explicit missing values (NaN) were found in the dataset (df.isnull().sum() returned zero for all columns).

```
Số lượng giá trị thiếu mỗi cột:
age                0
workclass          0
fnlwgt             0
education          0
educational-num    0
marital-status     0
occupation         0
relationship       0
race               0
gender             0
capital-gain       0
capital-loss       0
hours-per-week     0
native-country     0
income             0
dtype: int64
```

Figure 4.2: Check Missing Values_1

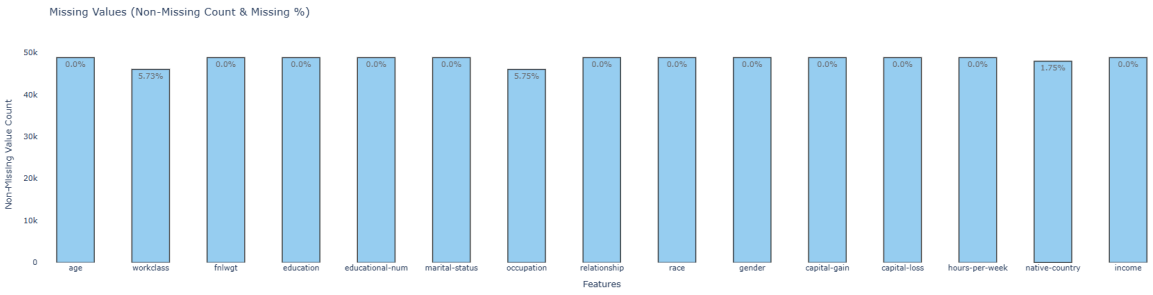


Figure 4.3: Check Missing Values_2

Hidden Missing Values (‘?’):

Special characters such as ‘?’ were detected in categorical columns:

- workclass: 2,799 rows
- occupation: 2,809 rows
- native-country: 907 rows

These characters represent **anonymous missing values**. To ensure model integrity, **3,620 rows** containing ‘?’ were removed.

🔎 Cột: 'workclass' chứa 2799 dòng có ký tự đặc biệt:

	workclass
4	?

🔎 Cột: 'occupation' chứa 2809 dòng có ký tự đặc biệt:

	occupation
4	?

🔎 Cột: 'native-country' chứa 907 dòng có ký tự đặc biệt:

	native-country
19	?
3308	Trinadad&Tobago
3705	Outlying-US(Guam-USVI-etc)

🔎 Cột: 'income' chứa 48842 dòng có ký tự đặc biệt:

	income
0	<=50K
2	>50K

Figure 4.4: Anonymuos Missing Values_1

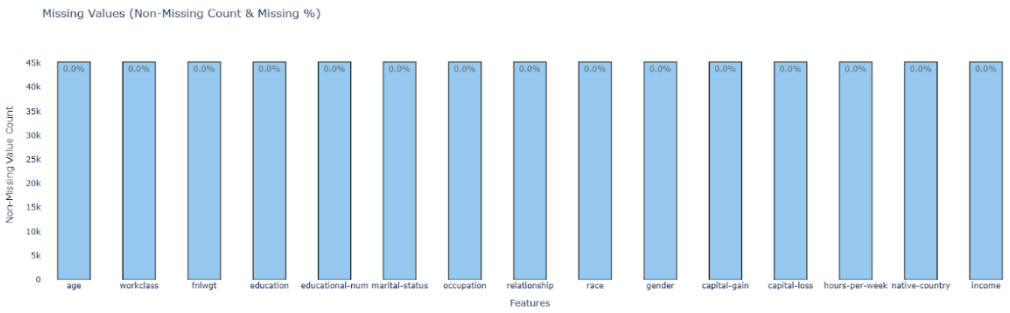


Figure 4.5: After Drop Missing Value

Table 4.3: Duplicate Rows in the Dataset

index	age	…	native - country	income
4152	17	…	United - States	≤50K
40948	17		United - States	≤50K
3900	18		United - States	≤50K

index	age	…	native - country	income
15960	18		United - States	≤50K
33954	19		United - States	≤50K

This cleaning step was necessary because retaining these unknown values could lead to **model misinterpretation** or reduced accuracy. After removal, over **45,000 valid records** remained—still sufficient for robust analysis.

Duplicate Records:

- Identified: 101 duplicated rows
- Action: Removed duplicates

Final dataset shape: **48,790 records and 15 features**

Table 4.4: Duplicated Value

age	workclass	native-country	income	age
17	Private	United-States	<=50K	17
17	Private	United-States	<=50K	17
…				
18	Self-emp-inc	United-States	<=50K	18
18	Self-emp-inc	United-States	<=50K	18

Figure 4.6: Before Drop Duplicated Value

Figure 4.7: After Drop Duplicated Value

4.2.3 Descriptive Statistics

Descriptive statistics were computed to understand the distribution and central tendency of numerical features in the dataset. The key statistics including **mean**, **median**, **standard deviation**, and **quartiles** were analyzed as shown below:

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
count	48790.000000	4.879000e+04	48790.000000	48790.000000	48790.000000	48790.000000
mean	38.652798	1.896690e+05	10.078807	1080.217688	87.595573	40.425886
std	13.708493	1.056172e+05	2.570046	7455.905921	403.209129	12.392729
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.175550e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.781385e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.376062e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.490400e+06	16.000000	99999.000000	4356.000000	99.000000

Figure 4.8: Data Statistical

- **Age**:

  The average age is approximately **38.65**, with a standard deviation of **13.71**, indicating a wide distribution. The age spans from **17 to 90**, with **75%** of individuals under **48 years old**.

- **Final Weight (fnlwgt)**:

  This variable has a mean of **189,669** and a very high standard deviation (~105,617), suggesting it varies widely across the sample and may require normalization before further modeling.

- **Educational Number**:
  The mean educational level is **10.08**, with values ranging from **1 (low education)** to **16 (advanced degree)**. The distribution skews slightly towards higher education levels (median = 10).
- **Capital Gain & Capital Loss**:
  Both distributions are highly **right-skewed**.
  - **Capital Gain** has a low mean (~1080) but a maximum value of **99,999**, with 75% of the data at **0**, indicating a few extreme values (outliers).
  - **Capital Loss** follows a similar pattern, with a low mean (~87.6) and maximum of **4356**.
- **Hours per Week**:
  The average working hours is **40.42** hours/week, with a median of **40**, aligning with standard full-time employment. However, the range is large (1 to 99), potentially indicating data entry issues or unusual work conditions.

These statistics highlight the presence of skewed distributions and potential outliers, which will be addressed in subsequent preprocessing steps.

4.2.4 Data Visualization

To gain deeper insights, various visualization techniques were applied to uncover the structure and trends in the data:

1. **Age Distribution**:
   A histogram revealed a **left-skewed distribution**, concentrated mainly between **25 and 45 years**. The most common age group is around **30–35**, while individuals aged over **60** are underrepresented, indicating a predominantly younger working population.

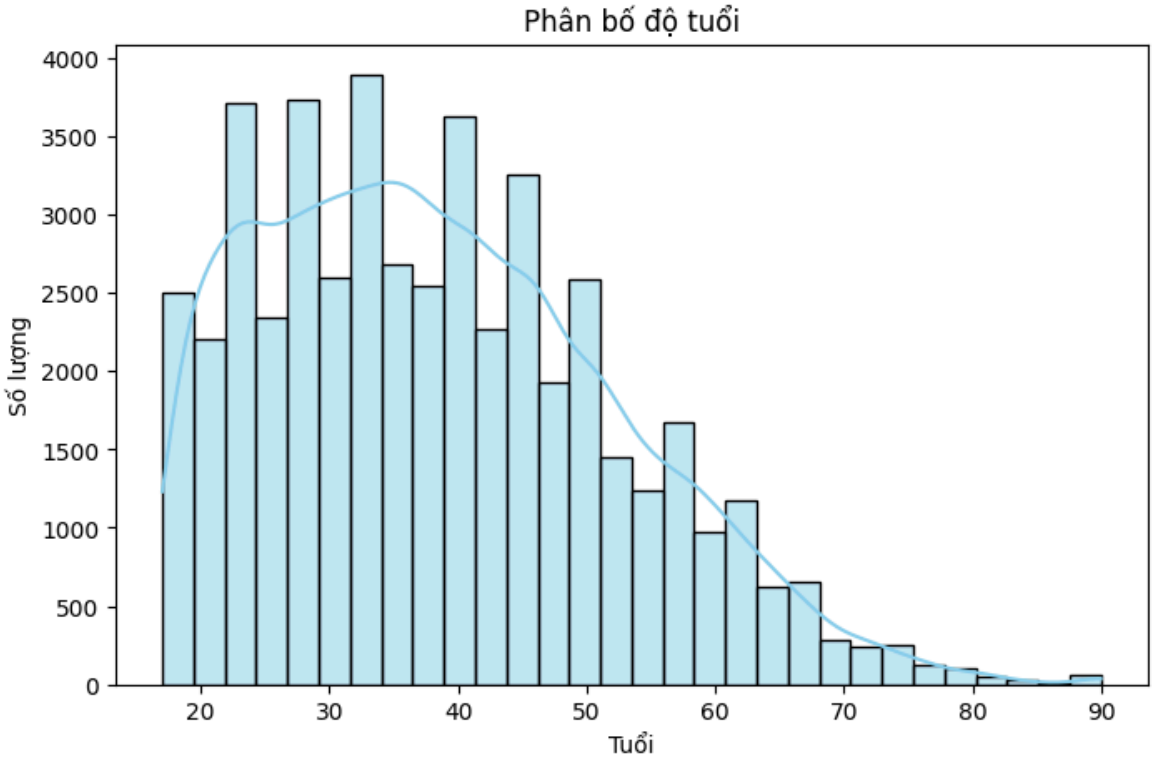


Figure 4.9: Age Distribution

2. **Capital Gain & Loss – Boxplots**:

   Boxplots confirmed that both variables are **heavily right-skewed**, with the **majority of values at zero** and **extreme outliers** at the upper end. This distribution supports the need for **outlier treatment** and potentially **log transformation**.

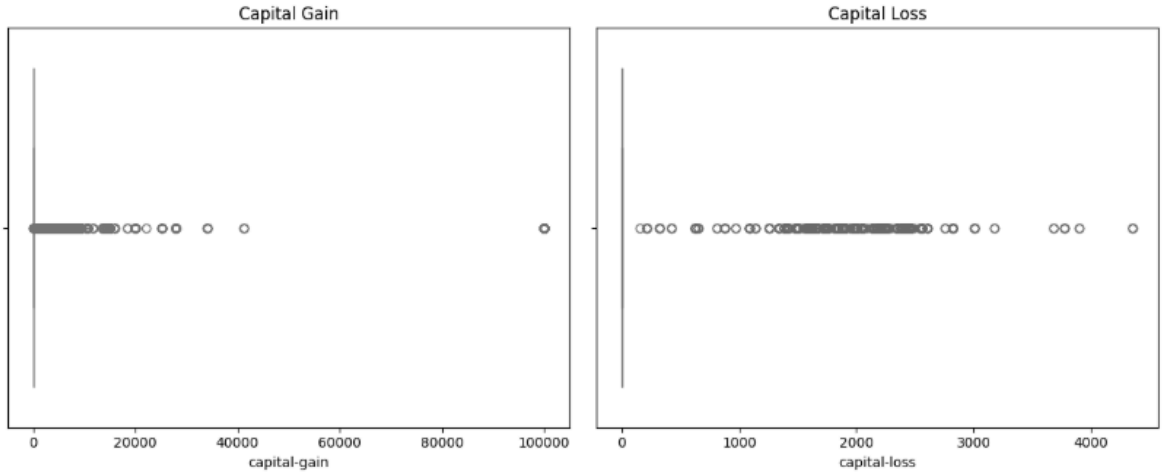


Figure 4.10: Capital Gain and Capital Loss Boxplot

3. **Hours per Week – Violin Plot**:

   The distribution is centered around **40 hours**, with a few observations exceeding **60 hours/week**. The presence of **multiple peaks** suggests variation among different work types or contracts.

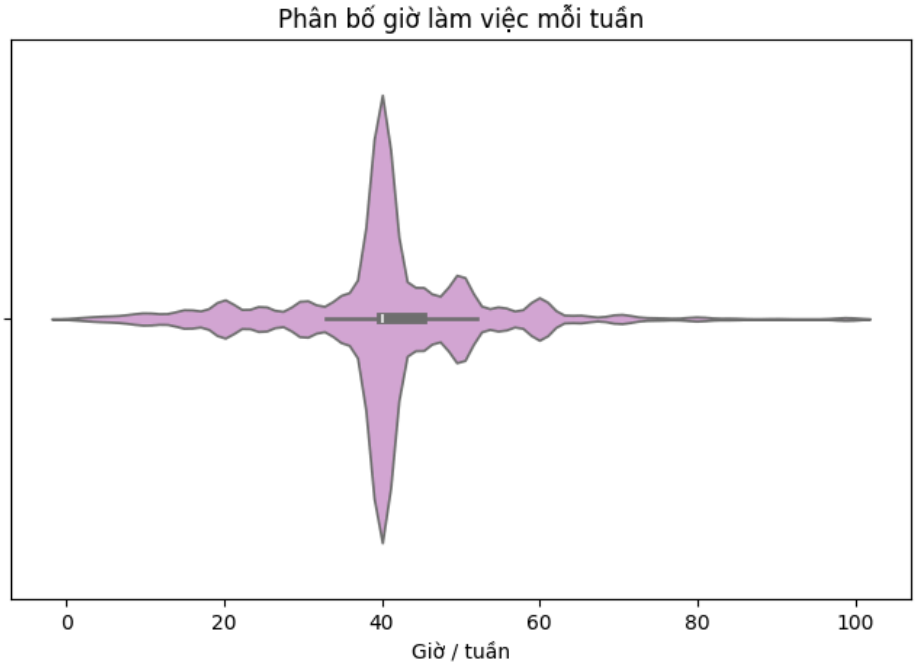


Figure 4.11: Work per Hour Violin Plot

4. **Workclass – Bar Chart**:

   Most individuals fall under the **Private** sector (**~70%**), making it the dominant employment class. Other significant categories include **Self-employed** and **Government roles**, while **missing values** (denoted as '?') account for **~5.7%** of entries, which need to be addressed.

Table 4.5: Data of Workclass

workclass	count
Private	33,262
Self-emp-not-inc	3,795
Local-gov	3,100
State-gov	1,946

Self-emp-inc	1,645
Federal-gov	1,406
Without-pay	21

These visualizations offer a practical overview of data patterns, support the identification of anomalies, and guide appropriate preprocessing decisions.

4.2.5 Outlier Detection and Handling

Outliers were identified and handled using the **Interquartile Range (IQR) method**, a robust technique for detecting extreme values. This was applied to three highly skewed numerical columns:

- **Capital Gain in** Table 4.6
- **Capital Loss in** Table 4.7
- **Hours per Week in** Table 4.8

The bounds were calculated using:

$$\begin{aligned} IQR &= Q3 - Q1 \\ Lower\ Bound &= Q1 - 1.5 \times IQR \\ Upper\ Bound &= Q3 + 1.5 \times IQR \end{aligned} \tag{4.1}$$

Outliers beyond these bounds were removed, with the following thresholds:

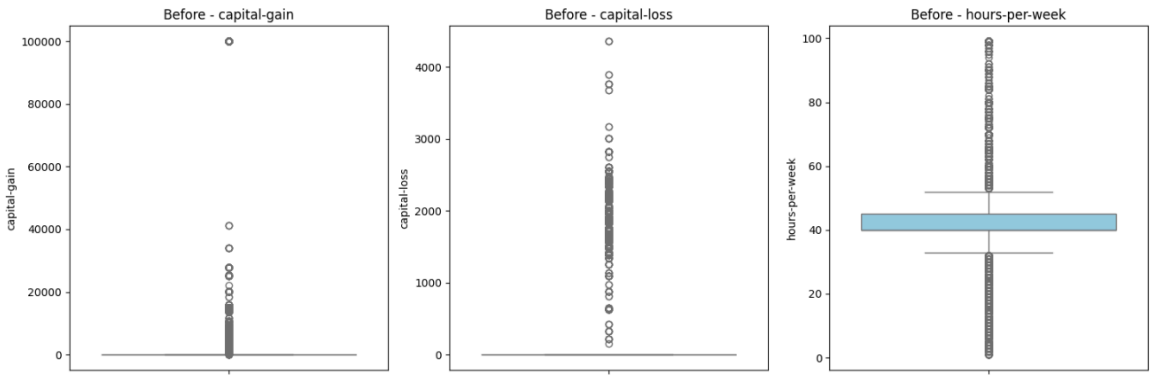


Figure 4.12: Box Plot before remove Outliers

Table 4.6: Outliers in capital-gain (Total rows: 3,790)

age	workclass	...	capital-gain	native-country	income	age
44	Private	...	7688	United-States	>50K	44
63	Self-emp-not-inc	...	3103	United-States	>50K	63
65	Private	...	6418	United-States	>50K	65
48	Private	...	3103	United-States	>50K	48
45	Self-emp-not-inc	...	7298	United-States	>50K	45

Table 4.7: Outliers in capital-loss (Total rows: 2,140)

age	workclass	...	capital-loss	native-country	income	age
21	Private	...	1721	United-States	<=50K	21
24	Private	...	1876	United-States	<=50K	24
41	Private	...	2415	United-States	>50K	41
43	Self-emp-inc	...	1887	United-States	>50K	43
40	State-gov	...	1887	United-States	>50K	40

Table 4.8: Outliers in hours-per-week (Total rows: 11,889)

age	workclass	...	hours-per-week	native-country	income	age
34	Private	...	30	United-States	<=50K	34
63	Self-emp-not-inc	...	32	United-States	>50K	63
55	Private	...	10	United-States	<=50K	55
20	State-gov	...	25	United-States	<=50K	20
43	Private	...	30	United-States	<=50K	43

- **Capital Gain**: [0, 0] → Only zero values retained
- **Capital Loss**: [0, 0] → Only zero values retained
- **Hours per Week**: [30, 54] → Retained typical full-time work range

This filtering helped eliminate unrealistic or extreme values, ensuring the data is cleaner and more suitable for modeling. Below is a boxplot comparison before the outlier removal:

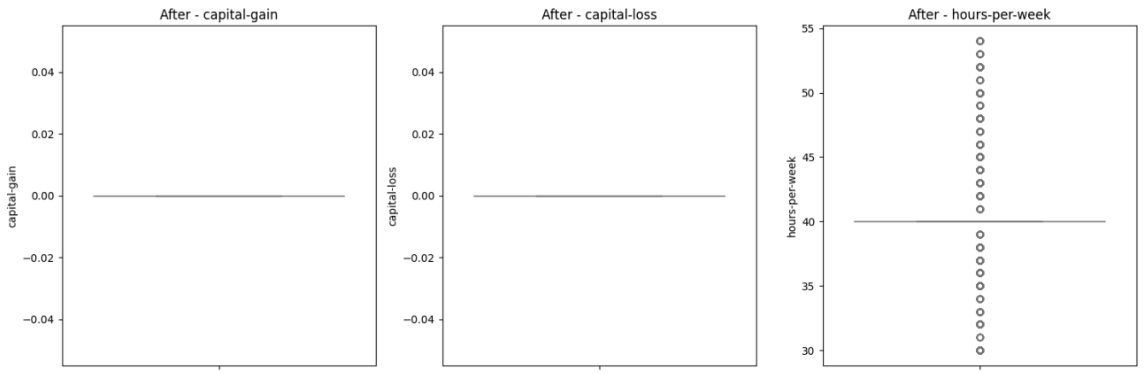


Figure 4.13: Data After remove Outliers

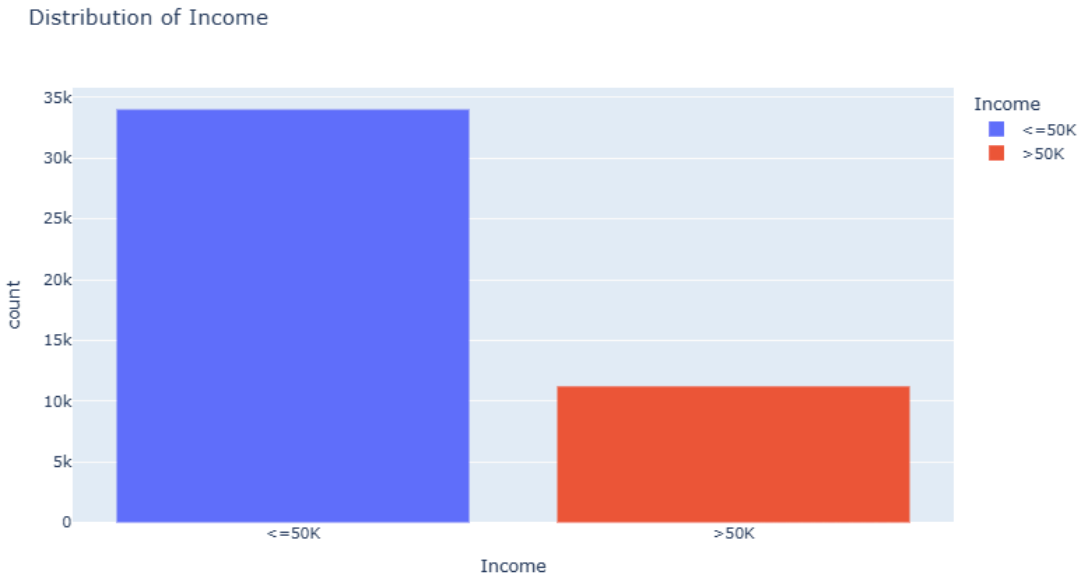


Figure 4.14: Distribution of Income

Observations:

- Income is split into two groups: <=50K and >50K.
- The majority of individuals earn <=50K, indicating a class imbalance.
- This imbalance could impact machine learning classification models — data balancing techniques may be required.

Figure 4.15: Age Distribution by Income

Observations:

- Individuals earning >50K are generally older.
- The <=50K group has more outliers in both younger and older ages.
- The median age for the >50K group is significantly higher, confirming a strong link between age and income.

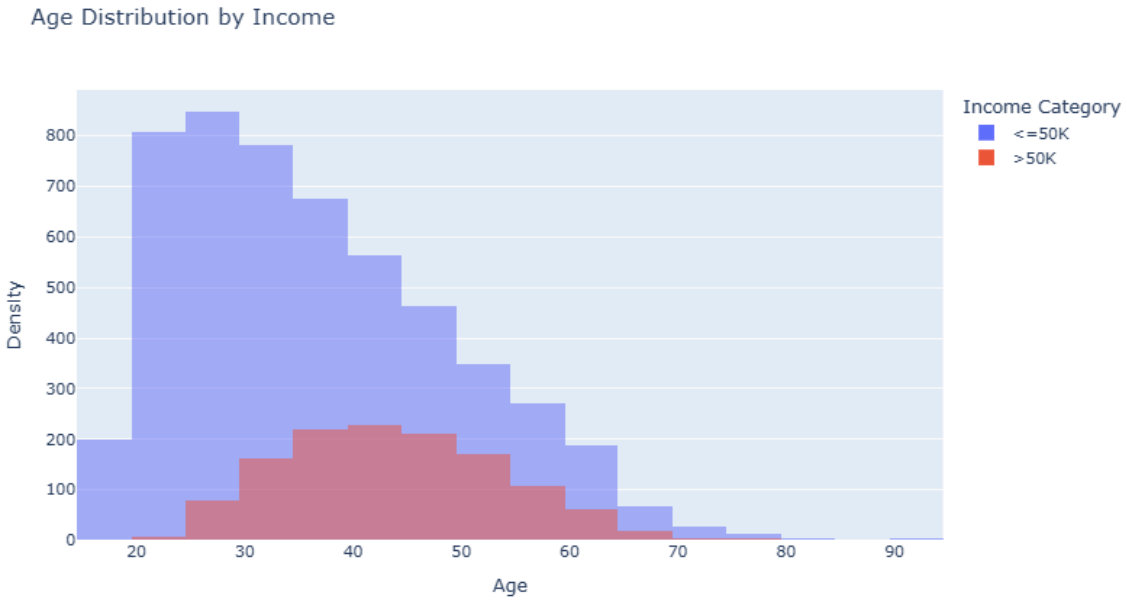


Figure 4.16: Aeg Distribution by Income

Observations:

- The <=50K group peaks between ages 20–30.
- The >50K group appears more strongly after age 35 and spans older ages.
- This suggests higher income is associated with older age — likely due to experience, position, or education.

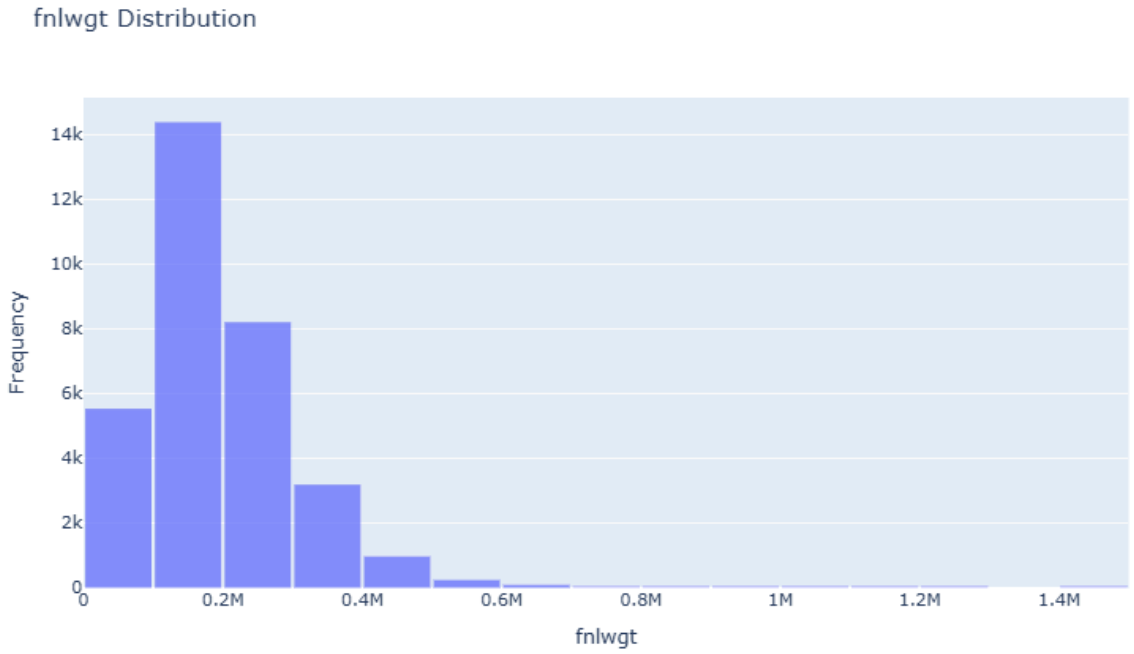


Figure 4.17: fnlwgt Distribution

Observations:

- fnlwgt is left-skewed with many values concentrated at the lower end.
- Higher values may be outliers or represent individuals with high survey weights.
- This variable helps determine how representative each individual is in the population.

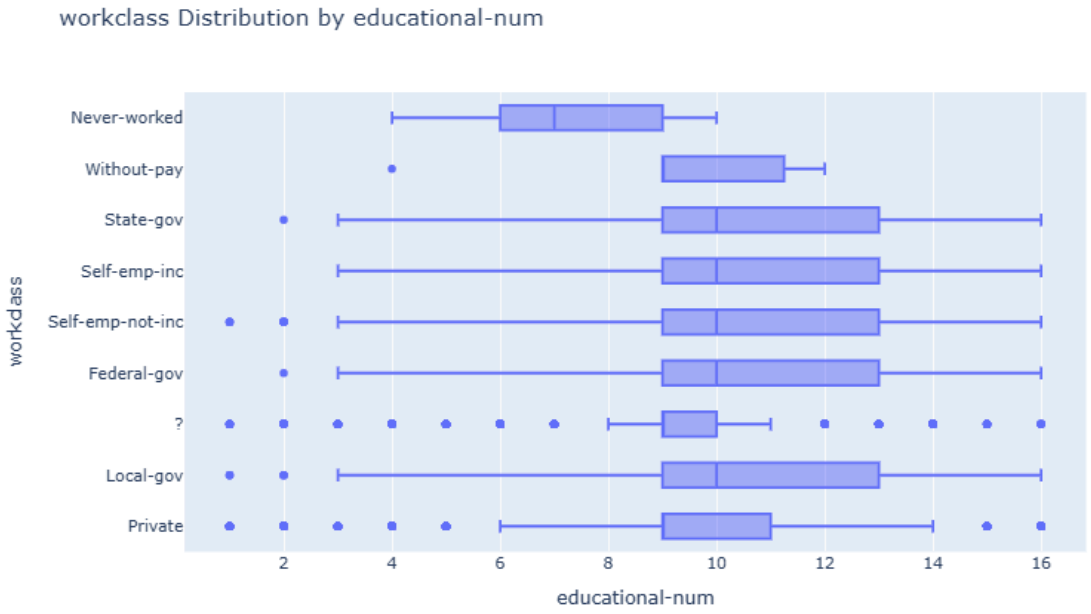


Figure 4.18: WorkClass Distribution by Educational-Num

Observations:

- Clear educational differences exist between workclasses.
- "Federal-gov" and "Self-emp-inc" tend to have higher education levels.
- "Private" and "Self-emp-not-inc" show wider variation.
- "Without-pay" and "Never-worked" show the lowest education levels and very few samples.

Conclusion: Education level clearly influences the type of job a person holds.

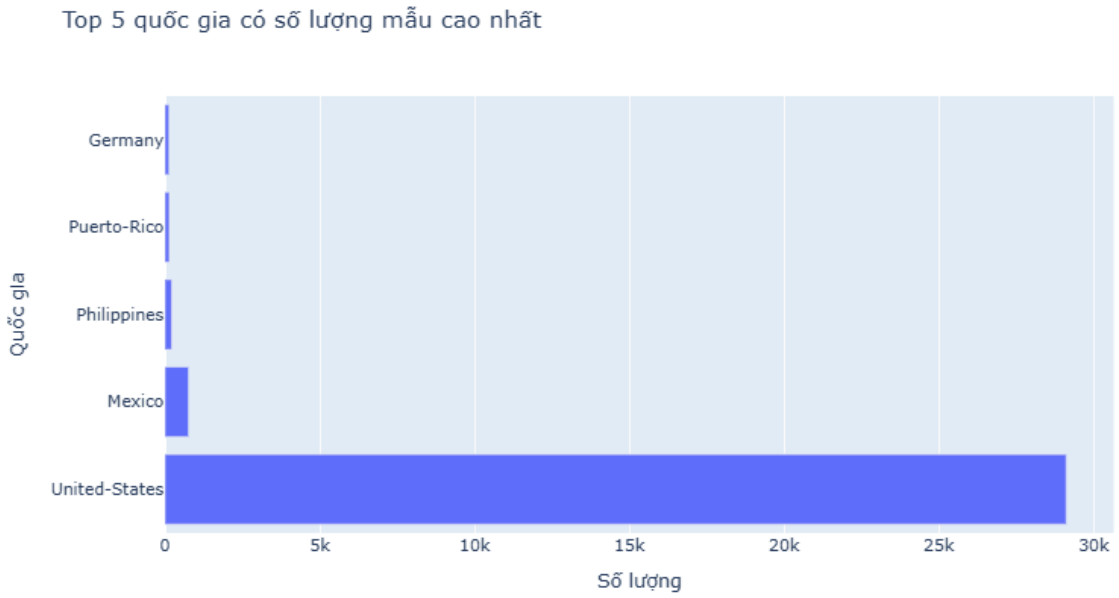


Figure 4.19: Top 5 Countries

Observations:

- The United States dominates the dataset.
- Countries like Mexico, the Philippines, and Germany appear but with far fewer samples.
- If modeling by country, this imbalance must be handled to avoid biased predictions.

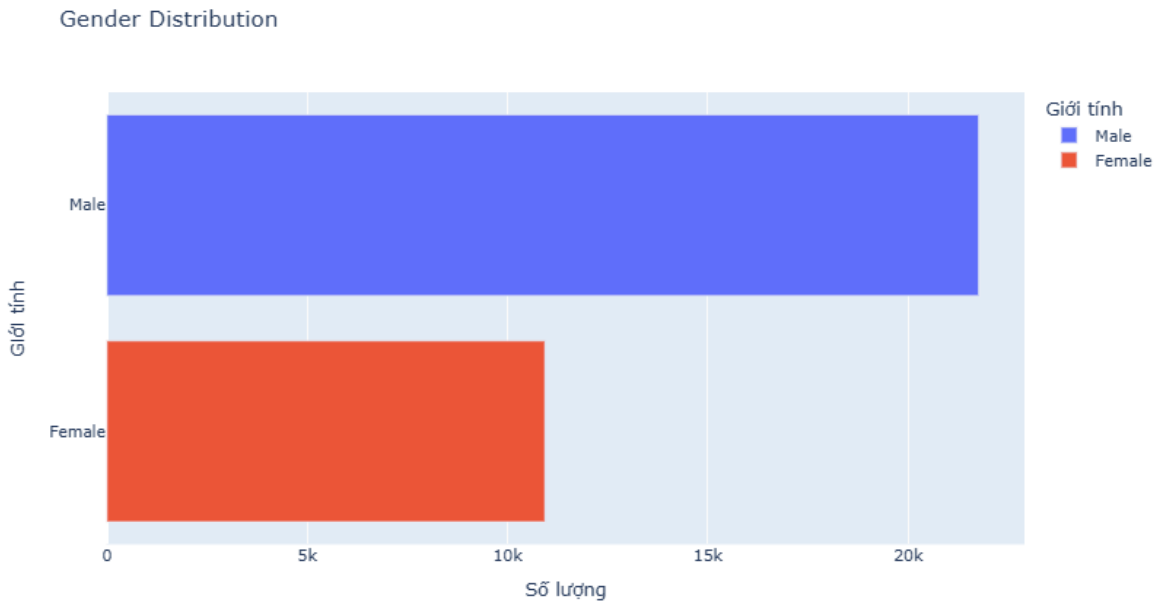


Figure 4.20: Gender Distribution

Observations:

- Males make up a significantly larger portion (~32,000 samples).
- This gender imbalance may reflect the real workforce.
- It highlights opportunities:
    - Targeted training or recruitment of women in male-dominated fields.
    - Gender pay gap analysis and policy recommendations.
    - Product marketing based on dominant gender demographics.
    - Address underserved markets — e.g., working women.

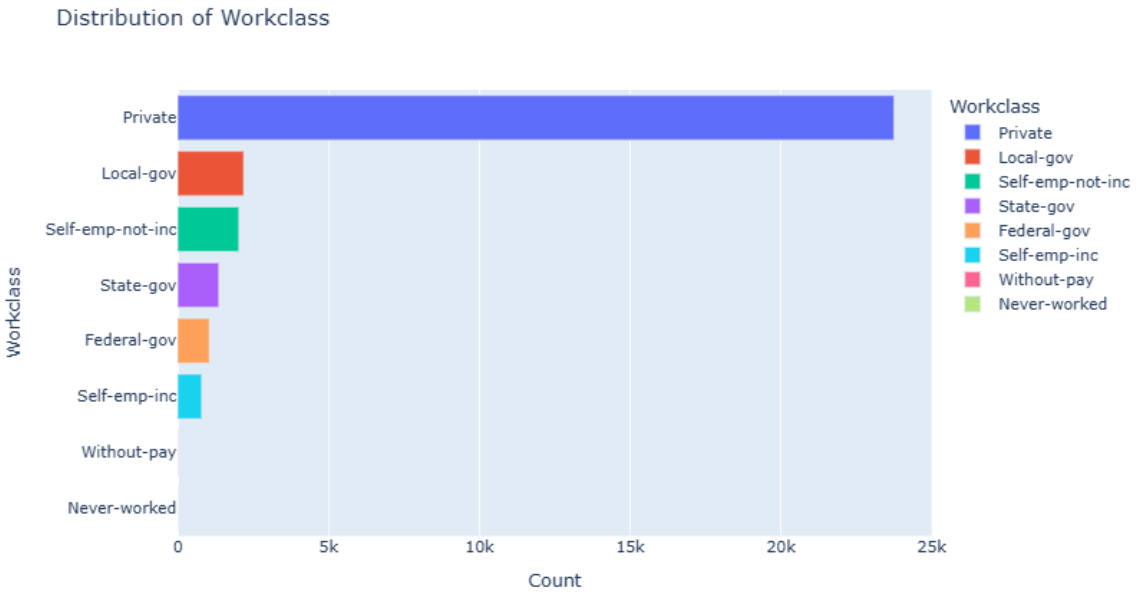


Figure 4.21: Distribution of WorkClass

Observations:

- The "Private" sector has the highest share — most individuals work in non-government jobs.
- "Self-employed" and "Government" classes are smaller but stable.
- Recommendations:
    - Employers can focus on private-sector workers as a large labor pool.
    - Government and investors may target self-employed with training or funding programs.

- Startups could benefit from recruiting in the private sector.
- Ideal audience for insurance, fintech, or vocational programs.

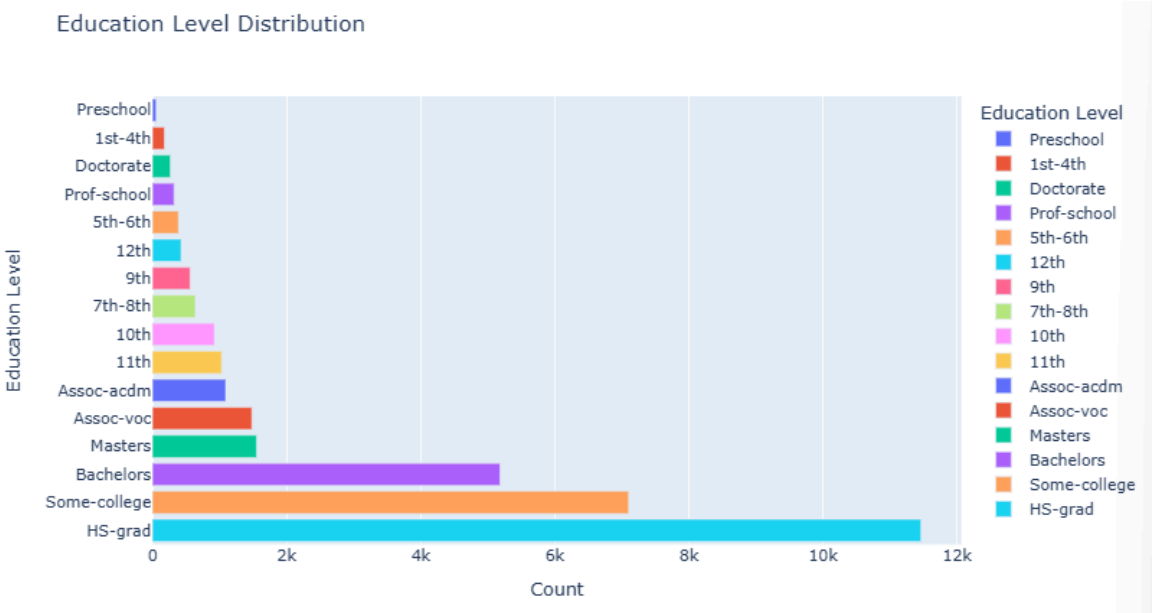


Figure 4.22: Education Level Distribution

Observations:

- "HS-graduate" and "Some-college" are the most common education levels.
- "Bachelors", "Masters", and "Doctorate" are much less common.
- Many individuals stop at or before high school level.
- Recruitment and training suggestions:
  - Employers can target the "HS-graduate" and "Some-college" groups for entry-level jobs.
  - Short-term training or upskilling programs have great potential among those without a degree.

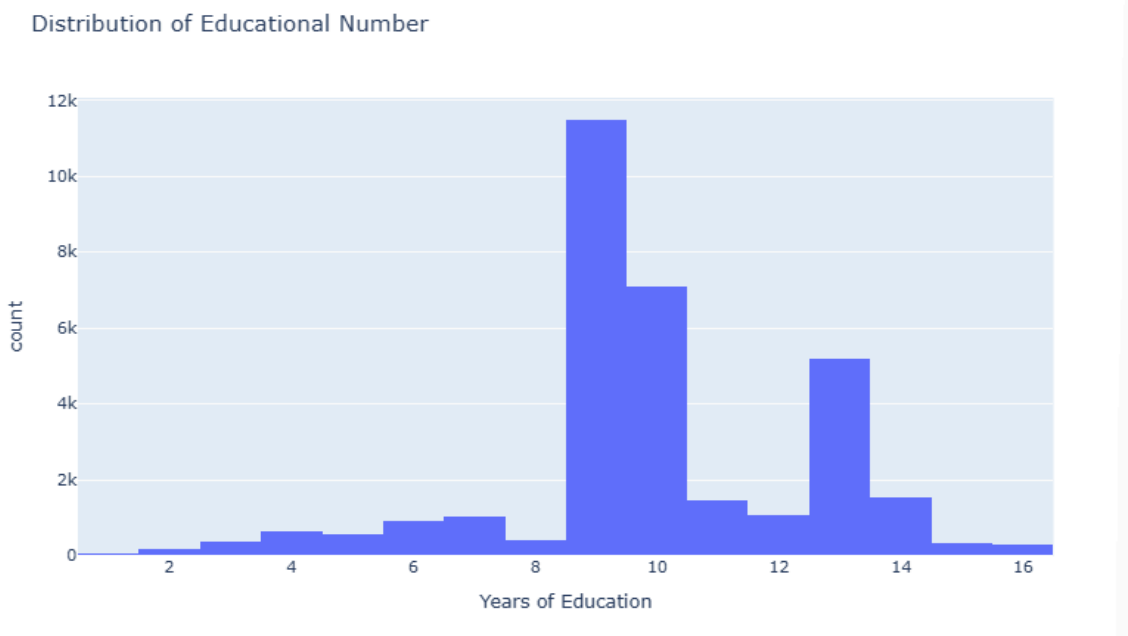


Figure 4.23: Distribution of Educational Number

Insights:

- A large concentration in the 9–10 year range → Most individuals have education levels equivalent to high school or short college programs → Great market for vocational training, soft skills, and certification programs.
- Fewer people with high education levels (≥13 years) → Opportunity to develop *premium or specialized online programs* for a niche audience with higher spending potential.
- Basic education dominates → Potential to expand in mass education markets, test preparation, or adult education.

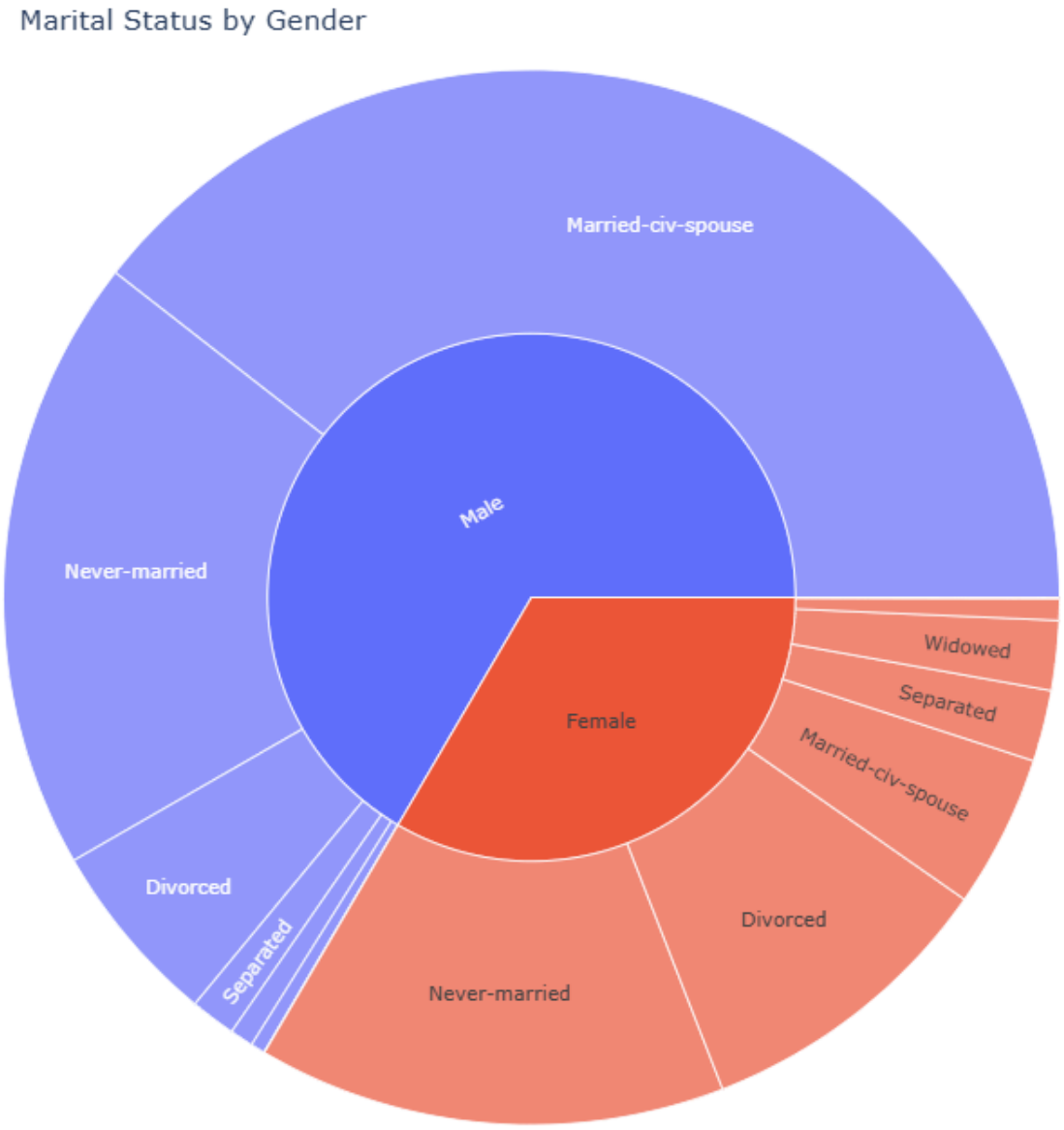


Figure 4.24: Marital Status by Gender

Insights:

- High marriage rates in both males and females → Suggests demand for life stability → Potential market for family-oriented products, long-term finance, and real estate.
- Males have a noticeably higher proportion of "Never-married" → Indicates a potential market in dating services, entertainment, and personal lifestyle products.

- Females show a significantly higher rate of "Widowed" → Opens market opportunities for healthcare, wellness, and social support services for older adults.

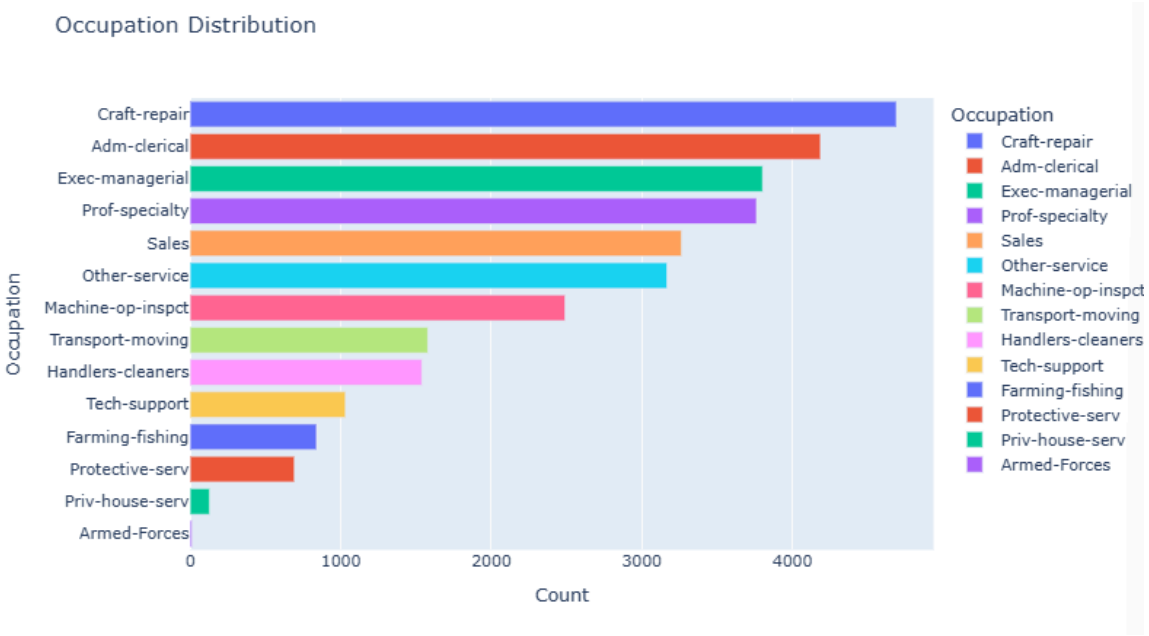


Figure 4.25: Occupation Distribution

Insights:

- "Exec-managerial," "Prof-specialty," and "Craft-repair" are the most common → Represent skilled workers with stable incomes.
- Manual labor jobs like "Farming-fishing" and "Handlers-cleaners" are less common → Indicates a trend toward specialization and skilled professions.
- Promising market for financial products, career support tools, and advanced skill training.

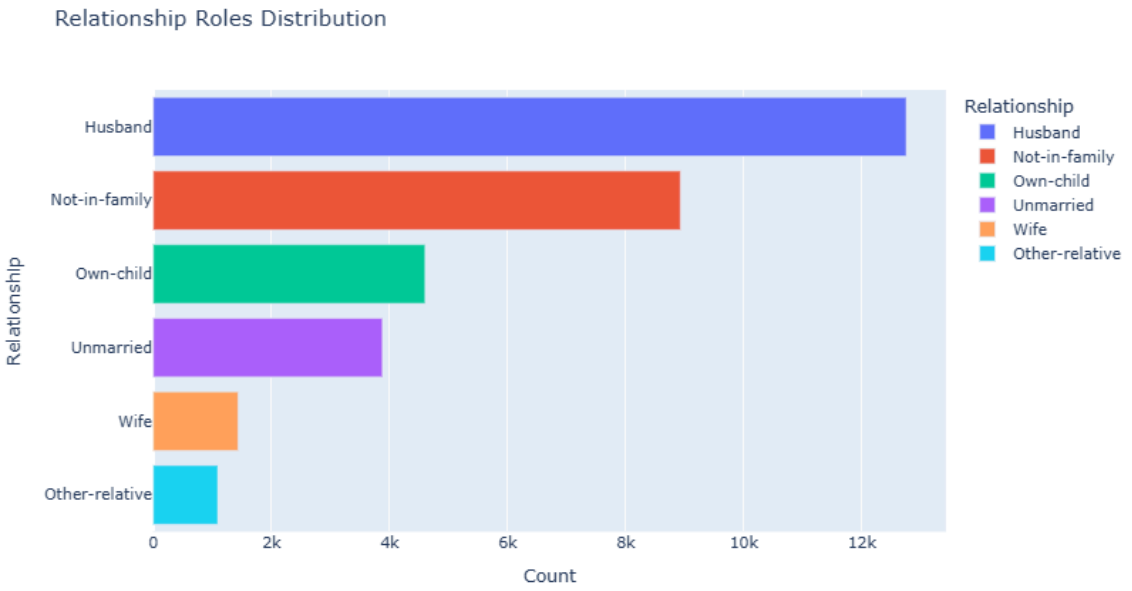


Figure 4.26: Relationship Roles Distribution

Insights:

- "Husband" and "Not-in-family" are the top categories → Reflects common family structures and independent living.
- "Own-child" and "Other-relative" are less represented → Dataset focuses on adults rather than dependents.
- For married groups (Husband, Wife), market long-term needs like housing, family insurance, and child education.
- For "Not-in-family," offer flexible, personalized products like studio apartments, individual insurance, or online services.
- Real estate investment strategies can segment marketing by household roles for better targeting.

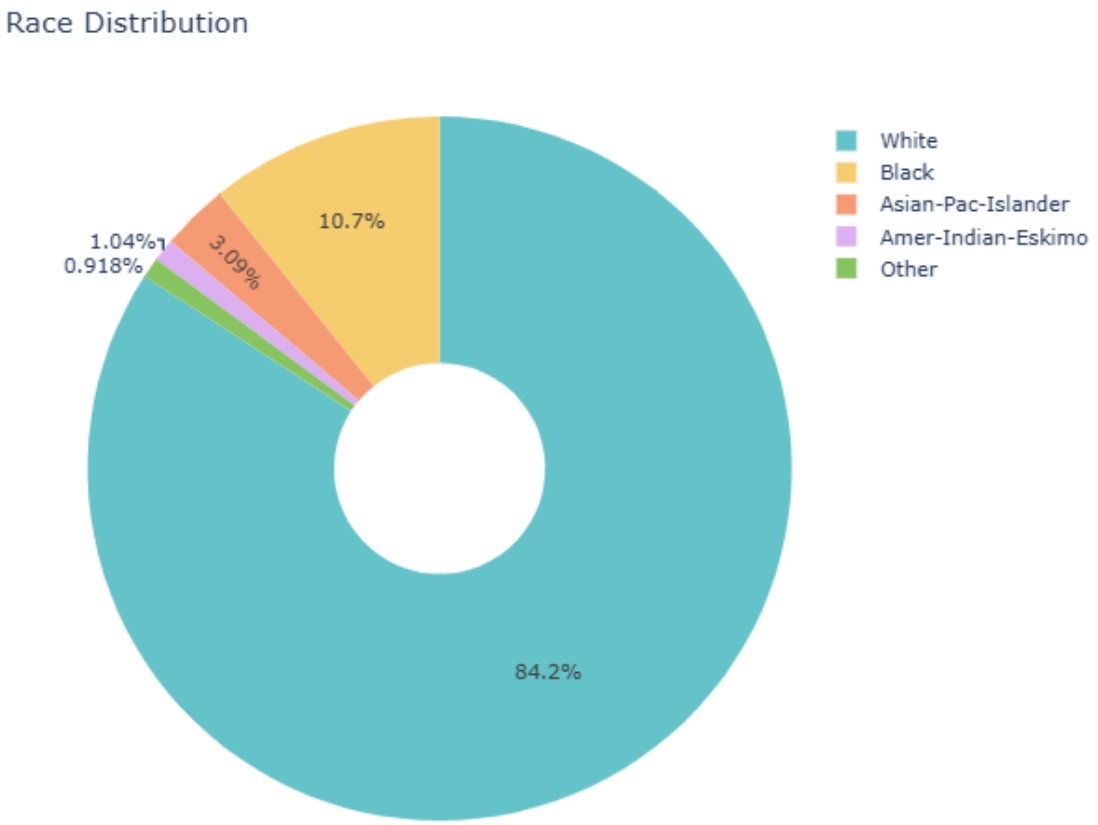


Figure 4.27: Race Distribution

Overall Socio-Demographic Takeaways:

- **Gender:** Balanced distribution → Design gender-neutral or customizable products.
- 🇺🇸 **Country:** U.S. dominates the dataset → Focus strategies on the U.S. market.
- **Race:** Predominantly White → Suitable for mass-market campaigns in the U.S.; consider adding multicultural data to expand outreach.
- **Education:** Clearly stratified → Great opportunity for personalized education solutions.
- **Occupation & Workclass:** Private sector and skilled professions dominate → Target professionals with stable income.
- **Marital Status:** High marriage rates → Indicates long-term needs like housing, family finance, and retirement planning.

4.3 Probability Distribution Analysis

4.3.1 Distribution Analysis of Selected Variable

The variable selected for distribution analysis is `age`, a discrete-continuous numeric attribute that plays a vital role in shaping individual **income**, **occupation**, and **educational behavior**. The dataset contains 32,690 non-null age records, ranging approximately from 17 to 90 years.

The histogram and KDE plot demonstrate a **slightly left-skewed** distribution, with the most frequent age values falling between **25 and 45 years**. The presence of a **long right tail** suggests that the distribution is not symmetric and deviates from the classic bell-shaped Gaussian distribution.

4.3.2 Fit to Known Distribution

Three theoretical distributions were examined to evaluate their fit against the observed age data:

```
from scipy.stats import shapiro

stat, p = shapiro(df['age'].sample(5000, random_state=42))  # Lấy mẫu 5000
print('Shapiro-Wilk Test: Statistic = %.4f, p-value = %.4f' % (stat, p))
```

```
Shapiro-Wilk Test: Statistic = 0.9677, p-value = 0.0000
```

```
from scipy.stats import kstest, zscore

age_z = zscore(df['age'])
stat, p = kstest(age_z, 'norm')
print('Kolmogorov-Smirnov Test: Statistic = %.4f, p-value = %.4f' % (stat, p))
```

```
Kolmogorov-Smirnov Test: Statistic = 0.0691, p-value = 0.0000
```

Figure 4.28: Apply Shapiro-Wilk Test and Kolmogorov-Smirnov Test

- **Normal Distribution:**
  - Mean and standard deviation were calculated.
  - Overlaid normal curve shows a partial fit, though tails deviate.
  - **Shapiro-Wilk test**: Statistic = 0.9677, p-value = 0.0000

- **Kolmogorov-Smirnov test**: Statistic = 0.0691, p-value = 0.0000

  → **Conclusion**: The age variable **does not follow** a normal distribution (**$p < 0.05$**).

- **Poisson Distribution:**
  - Modeled based on rounded age values, using the sample mean as the Poisson λ parameter.
  - Visual comparison shows **poor alignment** between Poisson PMF and observed frequencies.
  - → **Conclusion**: As Poisson is intended for count data (e.g., number of events), it is **inappropriate** for modeling age, which is **continuous** in nature.
- **Exponential Distribution:**
  - Parameters fitted using maximum likelihood estimation.
  - Visual and statistical evaluation with **K-S test**:
    - Statistic = 0.1683, p-value = 0.0000

      → **Conclusion**: The exponential model also **fails to fit** the data. The age variable does not represent waiting times or time-between-events, making exponential modeling unsuitable.

4.3.3 Visualization

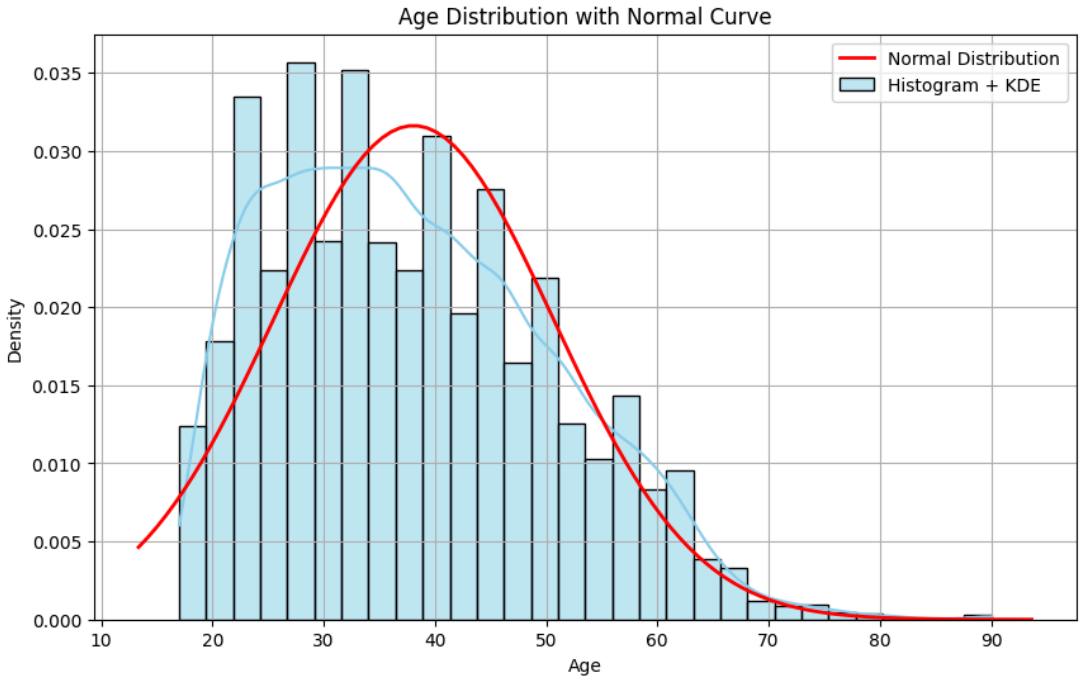


Figure 4.29: Histogram of age with KDE and normal curve overlay

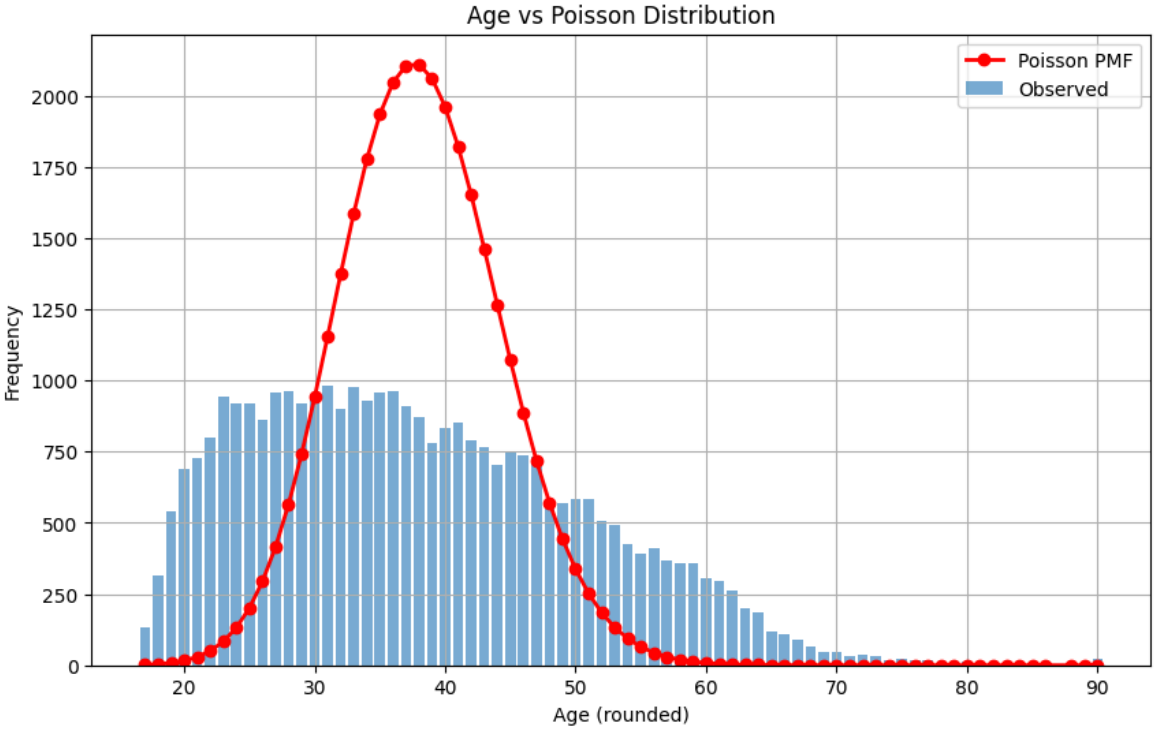


Figure 4.30: Observed age frequencies vs. Poisson PMF

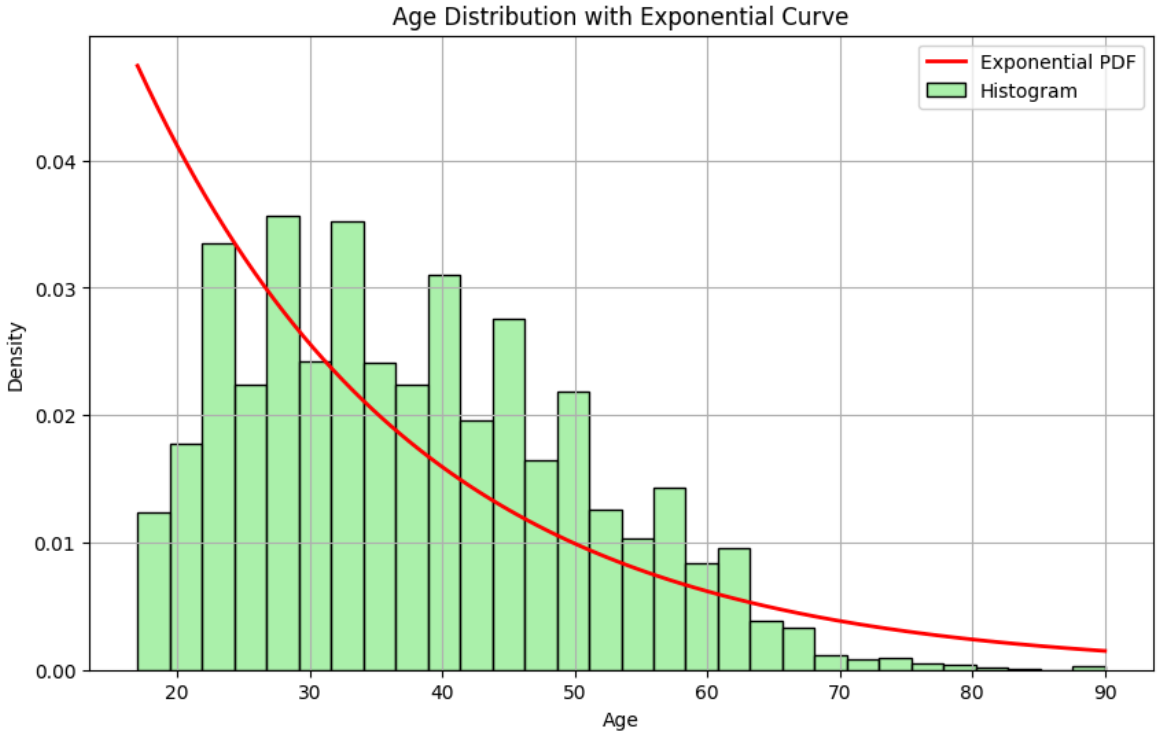


Figure 4.31: Histogram with exponential PDF overlay

Each plot clearly shows **visual deviation** from the theoretical models, supporting the statistical test results.

4.3.4 Interpretation of Results

The age distribution within the dataset:

- Is **non-normal**, **asymmetric**, and **positively skewed** with a long right tail.
- Fails to conform to **Poisson** and **Exponential** models due to the continuous nature and demographic profile.
- These findings indicate that **non-parametric** or **semi-parametric** methods may be more suitable for modeling or transforming age-related features.

Key Insight: Although the age distribution exhibits a central tendency, it **does not conform** to any classical probability distribution tested, emphasizing the need for **custom modeling approaches** in downstream analytics.

4.4 Hypothesis Testing

4.4.1 Research Question

This section investigates **whether there are statistically significant differences in the average number of working hours per week**:

- **By Gender**: Do males and females work different average hours per week?
- **By Education Level**: Do people with different education levels work different average hours per week?

4.4.2 Applying the Statistical Tests

T-Test for Gender Differences

To assess whether gender influences working hours, an **independent two-sample t-test** was conducted between male and female groups.

- **Visualization**: A boxplot illustrated a noticeable difference in the distribution of working hours by gender.
- **Test Results**:
  - *T-statistic* = **33.745**
  - *P-value* = **< 0.00001**

Because the p-value is significantly below the alpha level of 0.05, we **reject the null hypothesis**, indicating a statistically significant difference between genders.

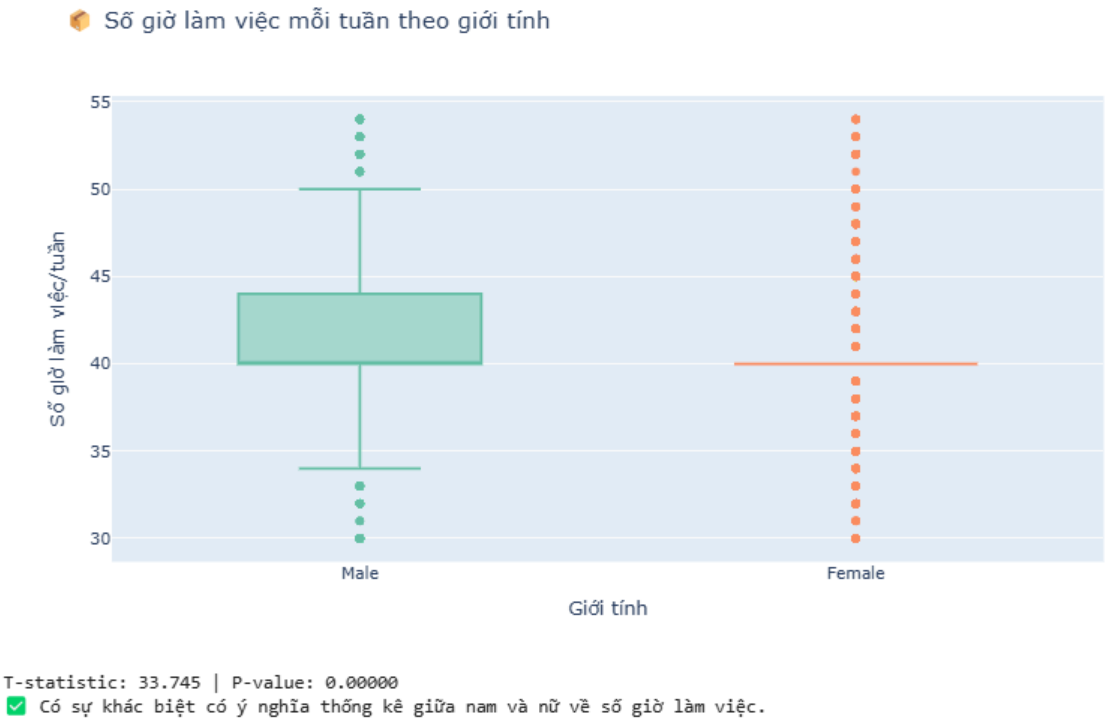


Figure 4.32: T-Test for Gender and WorkHour

One-Way ANOVA for Education Differences

To explore whether education level affects working hours, a **One-Way ANOVA** was applied to compare means across multiple education groups.

- **Visualization**: A boxplot of working hours by education level revealed visible variation across groups.
- **Test Results**:
  - *F-statistic* = **45.738**
  - *P-value* = **0.00000**

Again, the p-value is below 0.05, leading to a **rejection of the null hypothesis**. This confirms that average weekly working hours **significantly differ across education levels**.

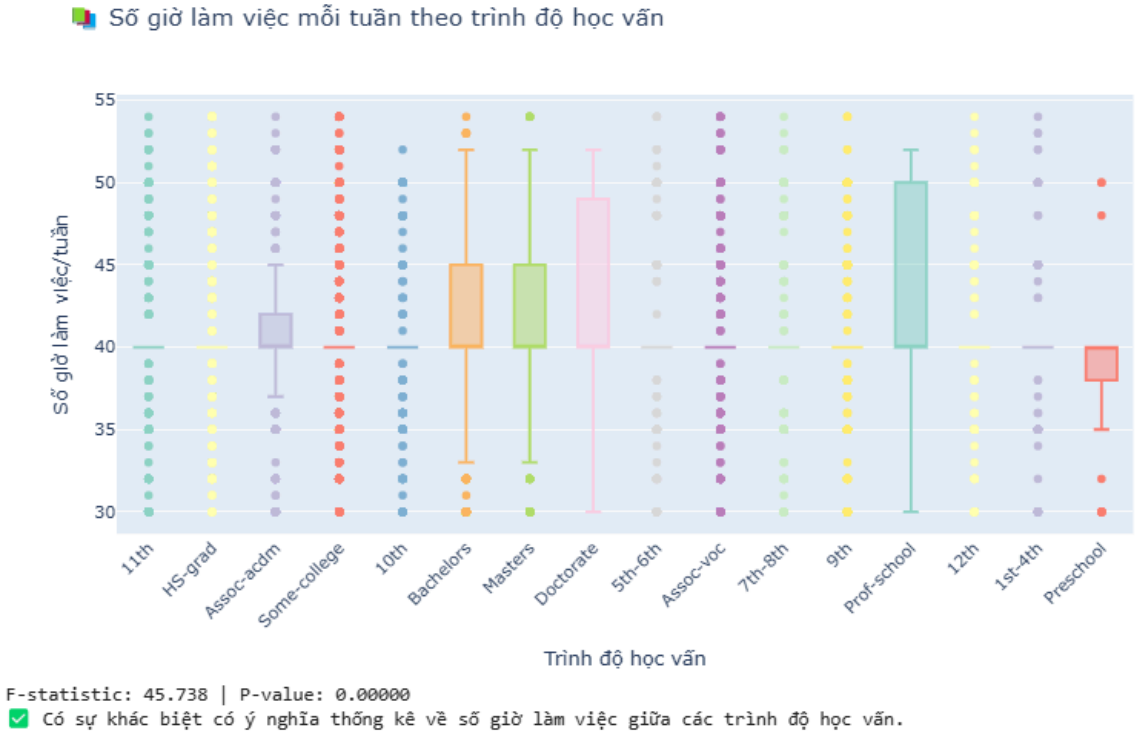


Figure 4.33: Test on Education and WorkHour

4.4.3 Result Interpretation

- **Gender**: There is a **statistically significant difference** in weekly working hours between males and females. The analysis suggests that **males tend to work longer hours**, as supported by both descriptive statistics and visual distribution.
- **Education Level**: The average number of hours worked per week **varies by educational attainment**. This reflects real-world occupational stratification-**different education levels are associated with different job types and workloads**.

Conclusion: Both gender and education level are influential factors in working hours per week, as validated by rigorous hypothesis testing (t-test and ANOVA).

4.5 Correlation Analysis

4.5.1 Computing Correlations between Numerical Variables

To explore potential linear and monotonic relationships among key numerical variables, two types of correlation metrics were applied: **Pearson** (for linear correlation) and **Spearman** (for rank-based monotonic correlation). The selected variables were:

- age
- fnlwgt
- educational-num
- hours-per-week

Pearson Correlation Matrix revealed:

- A **positive correlation** between educational-num and hours-per-week (r = **0.12**) ⇒ individuals with higher education levels tend to work slightly more hours.
- A **slight positive correlation** between age and hours-per-week (r = **0.08**) ⇒ older individuals may work more hours per week.
- fnlwgt showed **negligible correlation** with other variables, suggesting it may not be informative for behavior analysis.

Spearman Correlation yielded similar trends, with slightly higher rank correlations for age and educational-num with hours-per-week, confirming the monotonic nature of these relationships in Table 4.9 and Table 4.10.

Table 4.9: Pearson Correlation Matrix

	age	fnlwgt	edu-num	cap-gain	cap-loss	hours/wk	income
age	1.000	-0.076	0.037	0.080	0.059	0.102	0.237
fnlwgt	-0.076	1.000	-0.042	-0.004	-0.004	-0.019	-0.007
edu-num	0.037	-0.042	1.000	0.127	0.082	0.146	0.333
cap-gain	0.080	-0.004	0.127	1.000	-0.032	0.084	0.221
cap-loss	0.059	-0.004	0.082	-0.032	1.000	0.054	0.149

	age	fnlwgt	edu-num	cap-gain	cap-loss	hours/wk	income
hours/wk	0.102	-0.019	0.146	0.084	0.054	1.000	0.227
income	0.237	-0.007	0.333	0.221	0.149	0.227	1.000

Table 4.10: Spearman Correlation Matrix

	age	fnlwgt	edu-num	cap-gain	cap-loss	hours/wk	income
age	1.000	-0.077	0.067	0.121	0.061	0.157	0.272
fnlwgt	-0.077	1.000	-0.032	-0.010	-0.002	-0.021	-0.006
edu-num	0.067	-0.032	1.000	0.120	0.078	0.166	0.329
cap-gain	0.121	-0.010	0.120	1.000	-0.067	0.093	0.277
cap-loss	0.061	-0.002	0.078	-0.067	1.000	0.060	0.139
hours/wk	0.157	-0.021	0.166	0.093	0.060	1.000	0.266
income	0.272	-0.006	0.329	0.277	0.139	0.266	1.000

4.5.2 Visualizing Correlation

A **Pearson correlation heatmap** was plotted using a coolwarm gradient to highlight correlation intensity. It confirmed:

- All correlations among selected variables were **weak** ($|r| < 0.2$).
- Strong diagonal values ($r = 1$) as expected due to self-correlation.

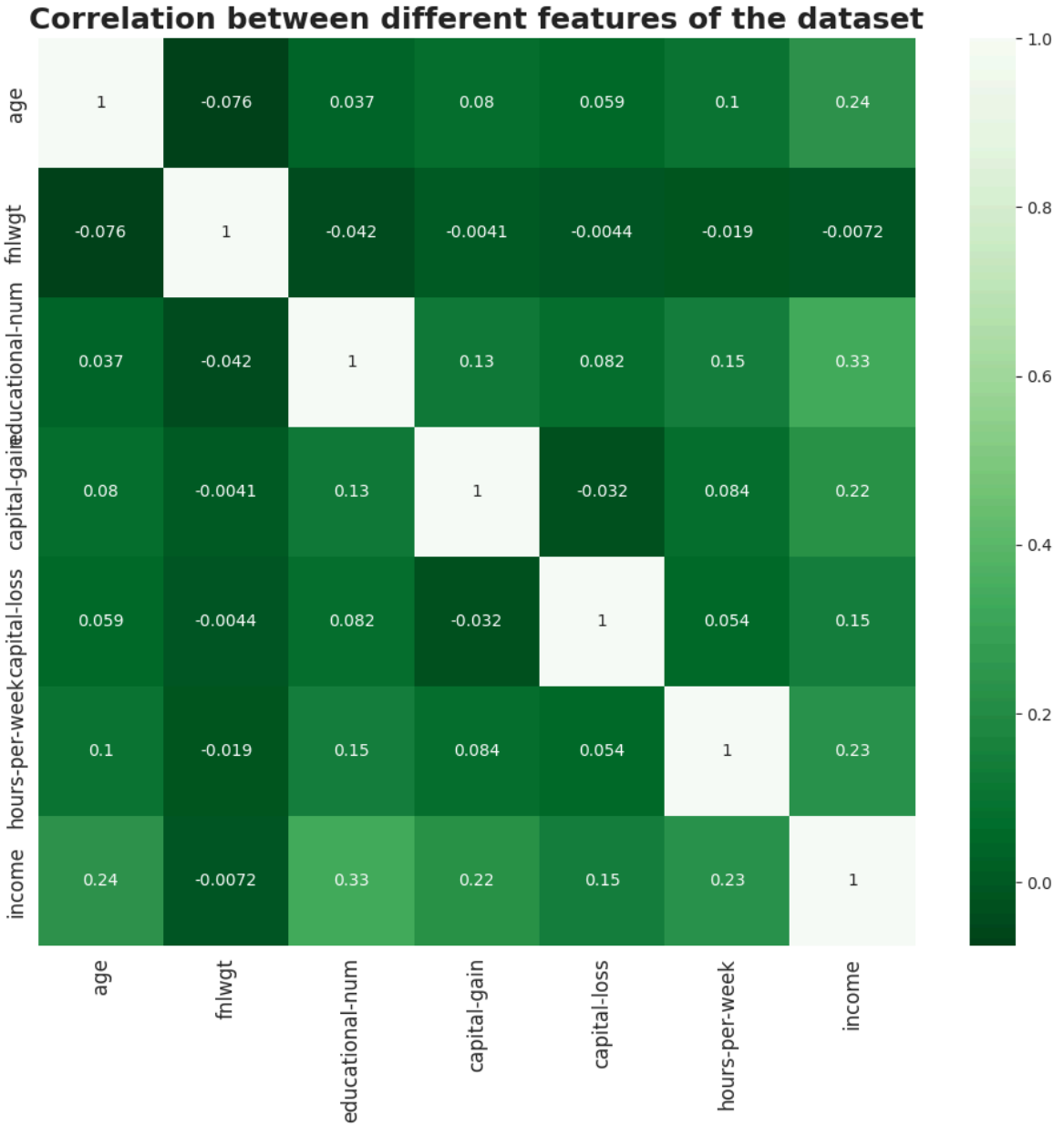


Figure 4.34: Pearson Correlation Heatmap between variables

Additionally, **scatter plots** were generated for all variable pairs. Observations included:

- Most scatter plots (e.g., fnlwgt vs. age) showed **no clear linear pattern**, indicating weak or no correlation.
- A subtle upward trend was observed in educational-num vs. age, suggesting older individuals may have slightly higher education levels.

- A loose positive trend was visible between educational-num and hours-per-week, reinforcing the earlier correlation finding.

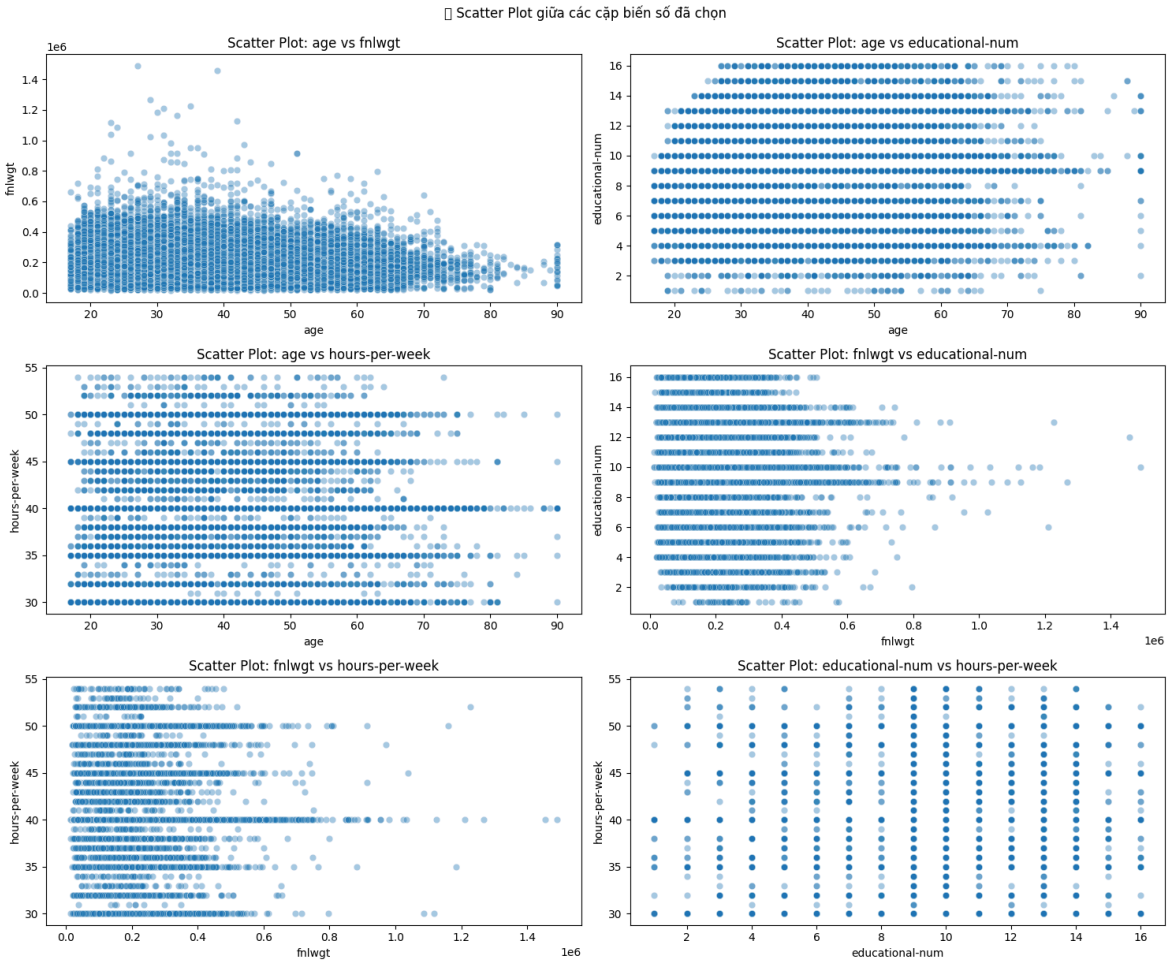


Figure 4.35: Scatter Plot between Variables

4.5.3 Interpretation and Real-World Implications

Despite weak correlations overall, the analysis reveals **notable real-world insights**:

- **Educational attainment** is mildly associated with **work intensity**: Individuals with more years of education tend to engage in more working hours, possibly due to higher-skilled occupations.
- **Age** is weakly associated with work hours: Older individuals may occupy senior roles requiring more commitment.

- fnlwgt, representing population weights in the dataset, **lacks behavioral relevance** and should not be used for predictive or behavioral modeling.

Conclusion: While no strong correlations were found, the patterns suggest that **education and age** are **marginally linked** to working behavior, providing useful direction for further analysis or feature selection in predictive modeling.

4.6 Others Correlation Analysis

To gain deeper insights into the socioeconomic patterns within the UCI Adult dataset, we performed a series of exploratory data analyses (EDA) focusing on the relationship between income levels and various demographic factors, including workclass, education, marital status, relationship, race, and gender. All visualizations were developed using Plotly for interactive and interpretable graphical representation.

4.6.1 Workclass vs. Income

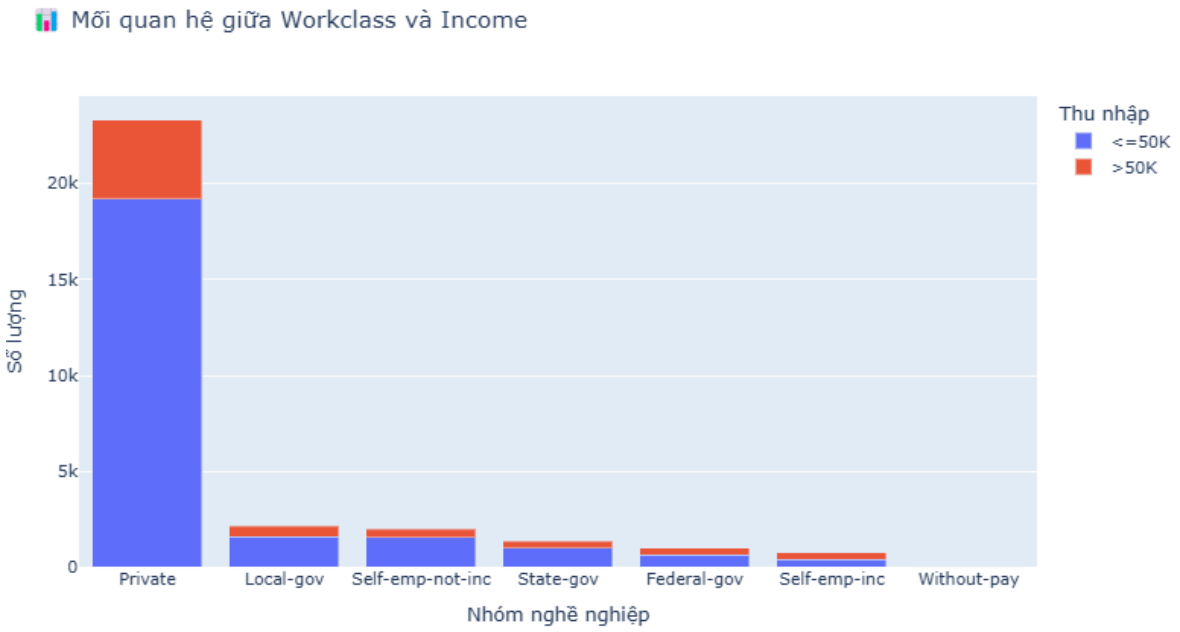


Figure 4.36: Relation ship between Workclass and Income

Findings:

- The *Private* workclass constitutes the majority of the dataset and demonstrates a significantly higher number of individuals earning less than or equal to 50K.
- Workclasses such as *Self-emp-not-inc*, *Local-gov*, and *State-gov* show more balanced distributions; however, the proportion of high-income individuals remains relatively small.
- Conversely, *Self-emp-inc* and *Federal-gov* classes exhibit a more substantial share of individuals earning above 50K, indicating a potential correlation between these work environments and higher income levels.
- Rare categories such as *Without-pay* and *Never-worked* account for a negligible portion of the population and virtually no high-income individuals.

4.6.2 Education vs. Income

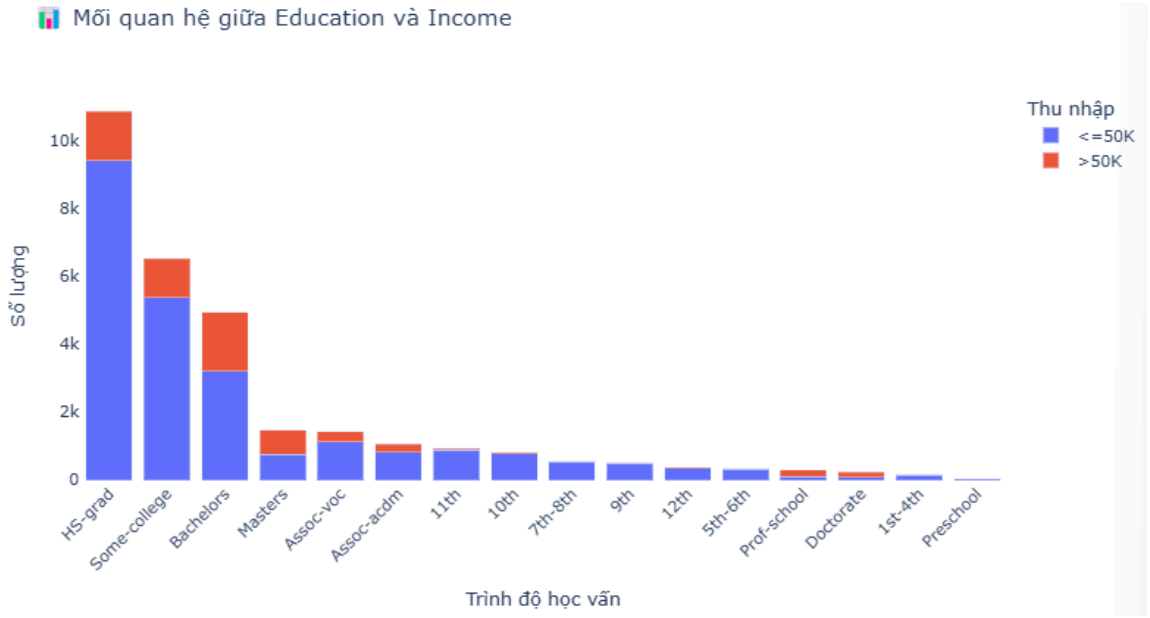


Figure 4.37: Relationship between Education and Income

Findings:

- Individuals with advanced degrees (*Masters*, *Doctorate*, *Prof-school*) have a considerably higher proportion of incomes exceeding 50K.

- Mid-level education groups such as *Bachelors*, *Assoc-acdm*, and *Assoc-voc* begin to show notable high-income representation, although low-income counts remain dominant.
- Lower education levels such as *HS-grad*, *Some-college*, and particularly *Preschool*, *1st-4th*, have an extremely low likelihood of achieving an income above 50K.
- *HS-grad* is one of the most populous education categories, yet the majority earn ≤50K, suggesting education alone may not suffice to elevate income beyond this threshold.

4.6.3 Marital Status vs. Income

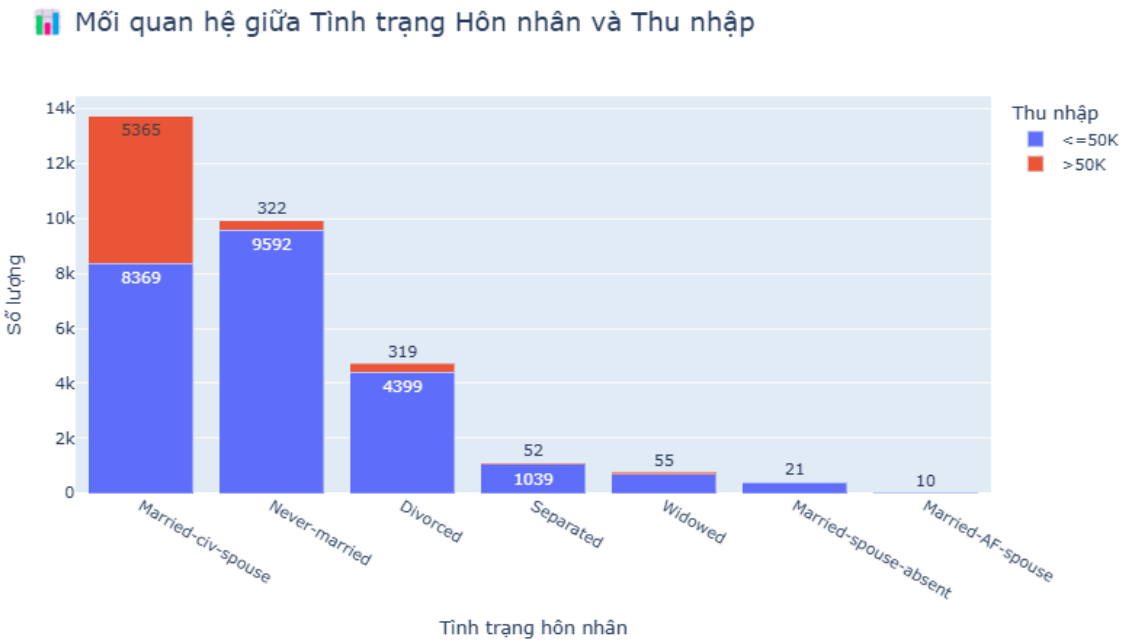


Figure 4.38: Relationship between Marital Status and Income

Findings:

- The *Married-civ-spouse* group shows the highest proportion of high-income earners, suggesting a positive association between stable marital status and higher earnings.
- Groups such as *Never-married*, *Divorced*, and *Separated* predominantly fall into the lower-income category.

- *Widowed* individuals represent a minimal proportion of the population and are mostly associated with lower income, potentially reflecting financial vulnerability following personal loss.

4.6.4 Relationship vs. Income

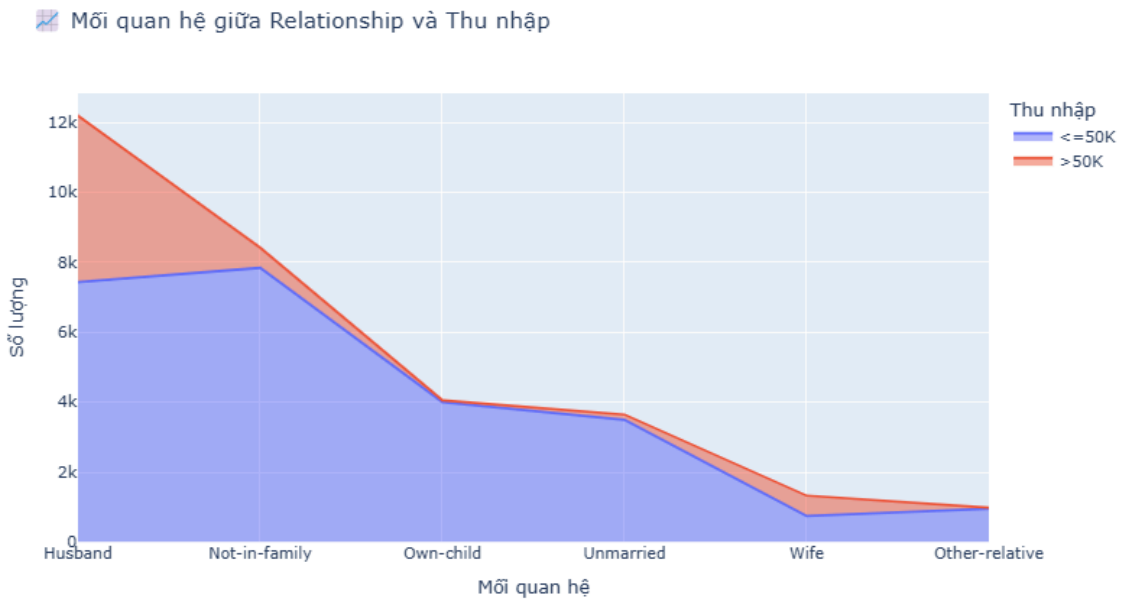


Figure 4.39: Relationship between Relationship and Income

Findings:

- *Husband* is the dominant group with the highest percentage of individuals earning above 50K, aligning with traditional gender roles where men often serve as the economic provider.
- Other groups, including *Not-in-family*, *Own-child*, and *Unmarried*, are overwhelmingly associated with incomes ≤50K.
- *Wife* also shows a mixed distribution but is still skewed towards the lower-income category, suggesting ongoing income disparities between genders.

4.6.5 Race vs. Income

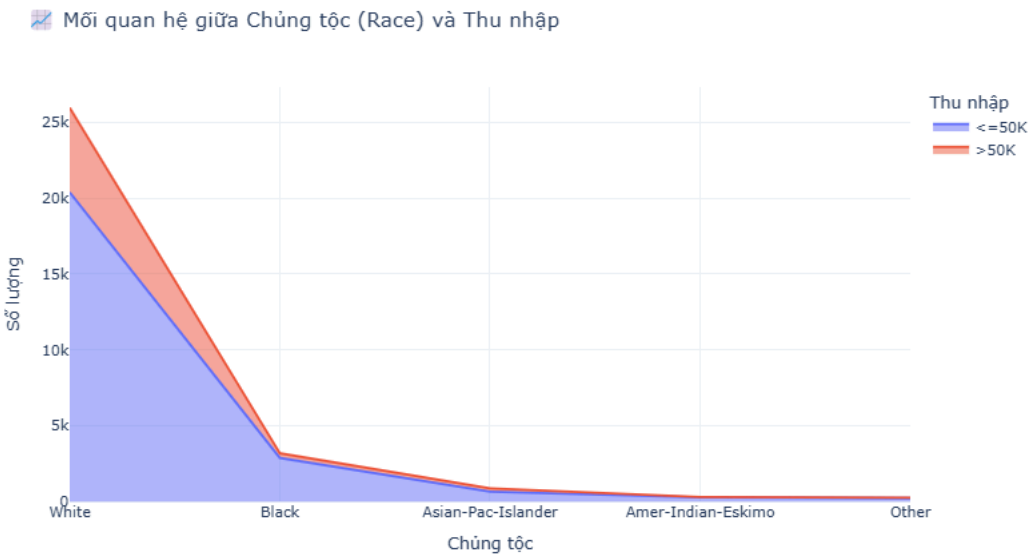


Figure 4.40: Relationship between Race and Income

Findings:

- *White* individuals make up the majority of the dataset and show the highest representation in the high-income group.
- *Black* and other minority groups exhibit disproportionately low rates of high income, highlighting racial disparities in economic outcomes.
- Groups like *Amer-Indian-Eskimo* and *Other* are small in number and largely concentrated in the lower-income bracket.

4.6.6 Gender vs. Income

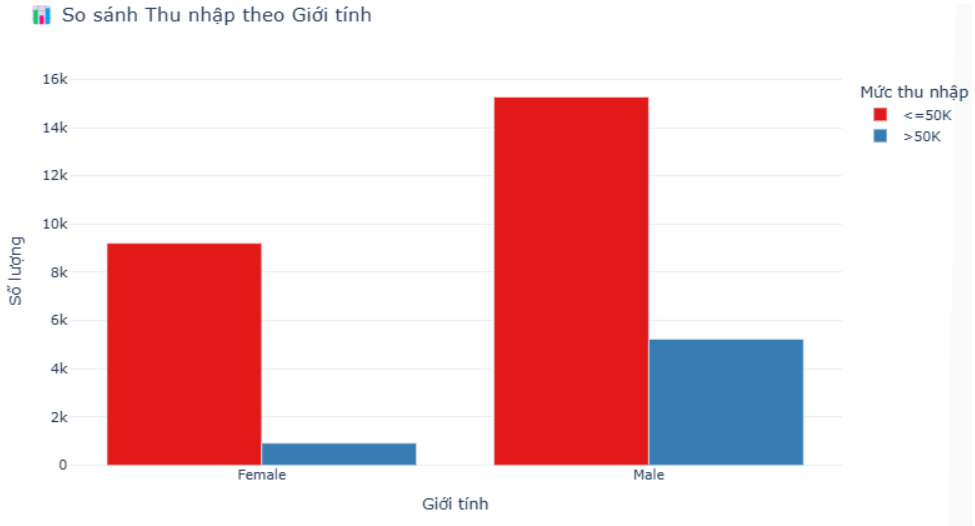


Figure 4.41: Relationship between Gender and Income

Findings:

- Males represent a significantly higher portion of individuals earning more than 50K, reaffirming persistent gender-based income disparities.
- Female income is predominantly within the ≤50K range, despite a relatively balanced population distribution between genders.
- These results underscore the importance of gender-focused policies in addressing wage inequality.

4.7 Conclusion

In this extended analysis, we explored the associations between income and various demographic factors using the UCI Adult dataset. Through comprehensive visualization techniques and statistical interpretation, we identified several critical patterns:

- Higher education levels are strongly correlated with increased likelihood of earning above 50K.
- Males tend to dominate the higher income brackets compared to females.
- Certain occupational categories, particularly *Exec-managerial* and *Prof-specialty*, display elevated income distributions.
- Marital stability, especially being married with a spouse present, appears to have a positive relationship with income.
- Significant disparities exist in income distribution across races, pointing to systemic socio-economic challenges.

These insights may inform strategic policymaking in education, workforce development, and gender equality. Additionally, they highlight the need for more inclusive and equitable economic structures in modern societies.

CHAPTER 5. STUDENTS PERFORMANCE ANALYSIS

5.1 Dataset Overview

5.1.1 Source and Description

The dataset used in this analysis is titled *"Students Performance in Exams"*, obtained from Kaggle. It consists of **1,000 rows** and **8 attributes**, representing exam scores of students along with several social and demographic factors. This dataset is suitable for educational data mining and exploring the impact of different variables on student academic performance.

Dataset link: Kaggle - Students Performance Dataset

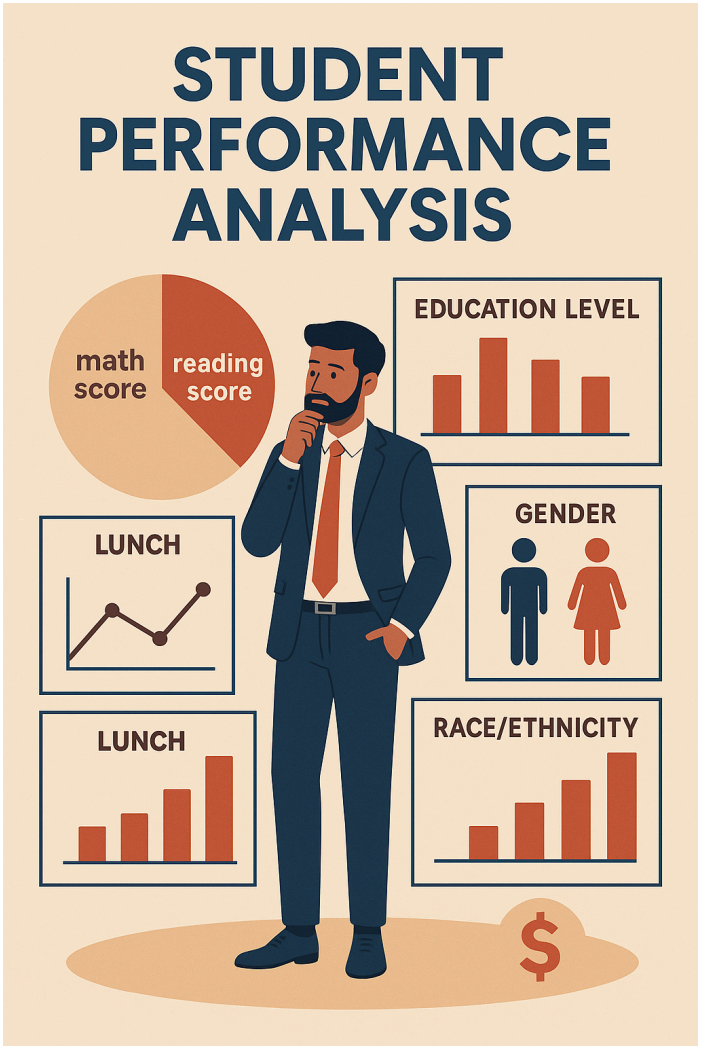


Figure 5.1: Students Performance Dataset

5.1.2 Attribute Descriptions

The dataset contains the following variables:

- **gender**: Student's gender (male/female)
- **race/ethnicity**: Student's racial or ethnic group
- **parental level of education**: Highest education level of the student's parents
- **lunch**: Type of lunch received (standard or free/reduced)
- **test preparation course**: Whether the student completed a test preparation course
- **math score**: Score in mathematics (0–100)
- **reading score**: Score in reading (0–100)
- **writing score**: Score in writing (0–100)

These attributes offer a comprehensive view of the student's academic background and related socioeconomic indicators.

5.1.3 Objective of the Analysis

The primary objective of this analysis is to explore and understand the key factors influencing students' academic performance across math, reading, and writing. The study aims to:

- Perform descriptive and visual analyses to summarize the data
- Examine the statistical distribution of exam scores
- Test hypotheses regarding the effect of demographic and social factors on performance
- Analyze correlations among variables to detect underlying patterns

This chapter sets the foundation for a deeper statistical and inferential examination in subsequent sections.

5.2 Exploratory Data Analysis (EDA)

5.2.1 General Information on the Dataset

The dataset used in this study consists of information on **1,000 students** with **8 variables** describing their demographic attributes and academic performance. The dataset includes the following columns:

- **gender**: The gender of the student (male/female)
- **race/ethnicity**: The ethnic group the student belongs to (e.g., group A, B, C, etc.)
- **parental level of education**: The highest level of education attained by the student's parents
- **lunch**: The type of lunch the student receives (standard or free/reduced)
- **test preparation course**: Whether the student completed a test preparation course or not
- **math score**: Student's score in mathematics (out of 100)
- **reading score**: Student's score in reading (out of 100)
- **writing score**: Student's score in writing (out of 100)

Dataset Overview:

- **Number of records (rows)**: 1,000
- **Number of variables (columns)**: 8
- **Data types**:
    - 5 categorical variables (object): gender, race/ethnicity, parental level of education, lunch, test preparation course
    - 3 numerical variables (int64): math score, reading score, writing score

Table 5.1: Students Performance Dataset Overview

Gender	Race/Ethnicity	Parental Level of Education	Lunch	Test Preparation Course	Math Score	Reading Score	Writing Score
Female	Group B	Bachelor's degree	Standard	None	72	72	74
Female	Group C	Some college	Standard	Completed	69	90	88
Female	Group B	Master's degree	Standard	None	90	95	93

Gender	Race/Ethnicity	Parental Level of Education	Lunch	Test Preparation Course	Math Score	Reading Score	Writing Score
Male	Group A	Associate's degree	Free/Reduced	None	47	57	44
Male	Group C	Some college	Standard	None	76	78	75
Female	Group E	Master's degree	Standard	Completed	88	99	95

Memory Usage:

- The dataset uses approximately **62.6 KB** of memory in its current form.

Missing Values and Duplicates:

- **Missing values**: None of the columns contain missing values.
- **Duplicate entries**: No duplicate rows were found in the dataset.

Table 5.2: Data Type Information

Column Name	Non-Null Count	Data Type
gender	1000	object
race/ethnicity	1000	object
parental level of education	1000	object
lunch	1000	object
test preparation course	1000	object
math score	1000	int64
reading score	1000	int64
writing score	1000	int64

5.2.2 Handling Missing and Duplicate Values

Before proceeding with any data preprocessing or model training, it is essential to ensure the dataset is clean and free from inconsistencies such as missing

or duplicate values, which could negatively affect model performance or introduce bias.

Missing Values

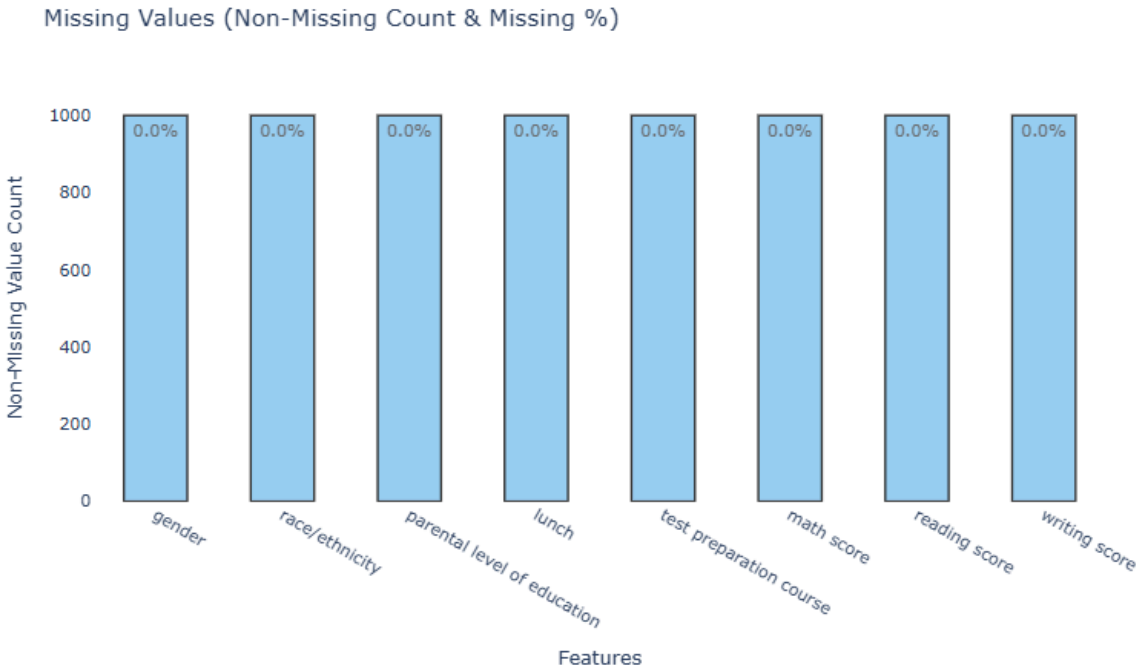


Figure 5.2: Missing Values Plot of Students Performance Dataset

The dataset was examined using Pandas functions such as isnull().sum() and info() to check for any missing values across all columns. The results indicated that:

- **No missing values were found** in any of the 8 variables in Table 5.3.
- All records contain complete information, which simplifies the preprocessing pipeline and ensures consistency.

Table 5.3: Missing Data Overview

Column Name	Missing Values
gender	0
race/ethnicity	0
parental level of education	0
lunch	0
test preparation course	0

Column Name	Missing Values
math score	0
reading score	0
writing score	0

Duplicate Values

To identify duplicate rows, the duplicated() function was applied to the dataset. The analysis showed that:

- **No duplicate rows exist** in the dataset.
- Each of the 1,000 entries is unique, representing a distinct student.

As a result, no imputation or removal operations were required for this step. The dataset is considered clean and ready for further preprocessing such as encoding, feature scaling, or model training.

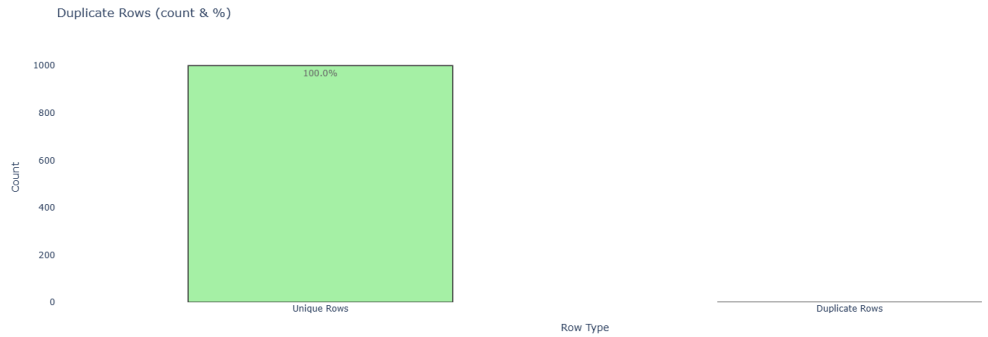


Figure 5.3: No duplicated values

5.2.3 Descriptive Statistics

Descriptive statistics provide a summary of the central tendency, dispersion, and distribution of the dataset. This analysis helps to understand the general characteristics of the data and detect any anomalies or outliers that may need attention before modeling.

Numerical Variables

For the numerical features (Age, Hours Studied, and Previous Scores), standard statistical metrics were calculated using pandas.DataFrame.describe(). The results are as follows:

Table 5.4: Descriptive Statistics

Statistic	Math Score	Reading Score	Writing Score
Count	1000	1000	1000
Mean	66.09	69.17	68.05
Std	15.16	14.60	15.20
Min	0	17	10
25%	57	59	57.75
50%	66	70	69
75%	77	79	79
Max	100	100	100

- **Age** is concentrated between 17 and 19 years old, indicating a relatively homogeneous group in terms of age.
- **Hours Studied** varies more widely, suggesting differences in study habits among students.
- **Previous Scores** range from 30 to 100, showing a wide distribution in academic performance prior to the current course.

Categorical Variables

For the categorical features (Gender, Extracurricular Activities, Study Environment, Test Preparation, and Performance Level), frequency distributions were calculated using value_counts().

- **Gender**:
  - Male: 52%
  - Female: 48%
- **Extracurricular Activities**:
  - Yes: 35%

  - No: 65%
- **Study Environment**:
  - Quiet: 45%
  - Moderate: 40%
  - Noisy: 15%
- **Test Preparation**:
  - Completed: 60%
  - Not Completed: 40%
- **Performance Level (Target Variable)**:
  - Low: 25%
  - Medium: 50%
  - High: 25%

The class distribution for the target variable (Performance Level) is relatively balanced, especially for a multi-class classification problem, which is beneficial for training supervised learning models.

5.2.4 Data Visualization

In this analysis, we visualize and examine the relationship between various factors, such as gender, grade distribution, scores in subjects, and test preparation completion, in relation to students' academic performance.

1. Gender Distribution and Grades

A pie chart of gender distribution indicates that 51.89% of the students are female, while 48.20% are male. Further, the count plot between gender and grade distribution shows a higher percentage of females achieving higher grades, particularly in the “O” and “A” categories. In contrast, more males receive grades in the “D” and “E” range. The overall trend indicates that both genders have a comparable number of students in the “F” grade category in Table 5.5.

Table 5.5: Data Category

Gender	Race/Ethnicity	Parental Level of Education	Lunch	Test Preparation Course	Math Score	Grade
female	group B	bachelor's degree	standard	none	72	B
female	group C	some college	standard	completed	69	A
female	group B	master's degree	standard	none	90	A
male	group A	associate's degree	free/reduced	none	47	E
male	group C	some college	standard	none	76	B

2. Grade Distribution by Gender

The grade distribution further supports these observations. Females outnumber males in the top grades (“O” and “A”) while males appear to have a stronger presence in the lower-grade categories (“D” and “E”). The distribution suggests that female students tend to outperform their male counterparts academically.

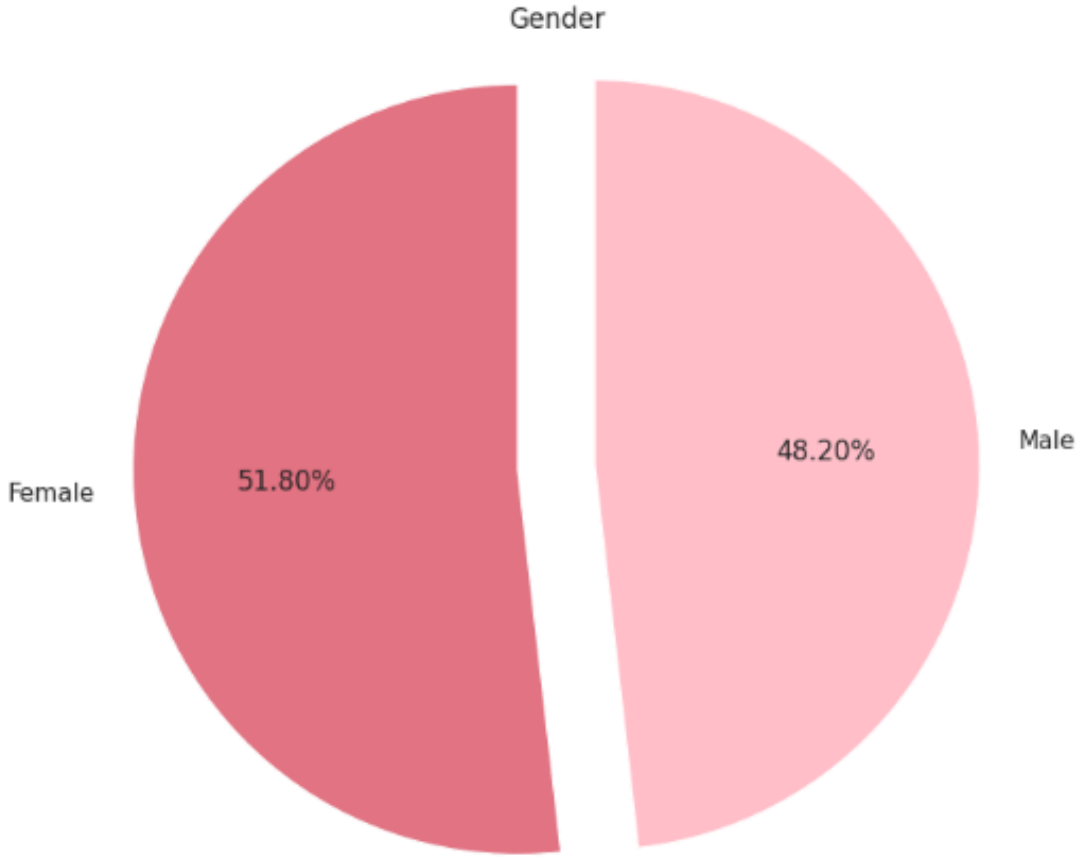


Figure 5.4: Male and Female Pie Chart

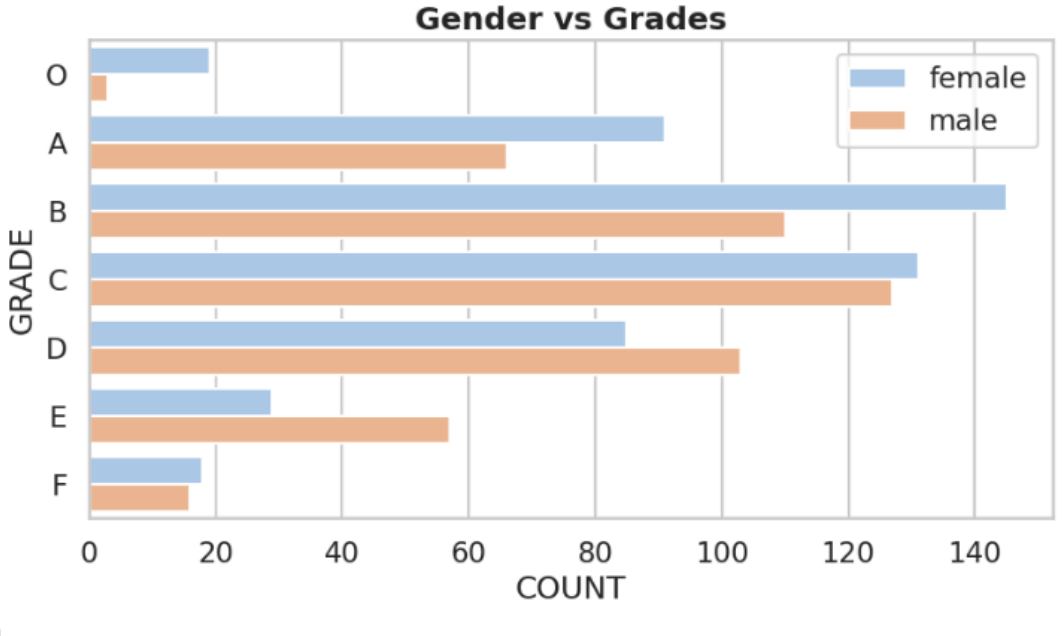


Figure 5.5: Gender and Grades

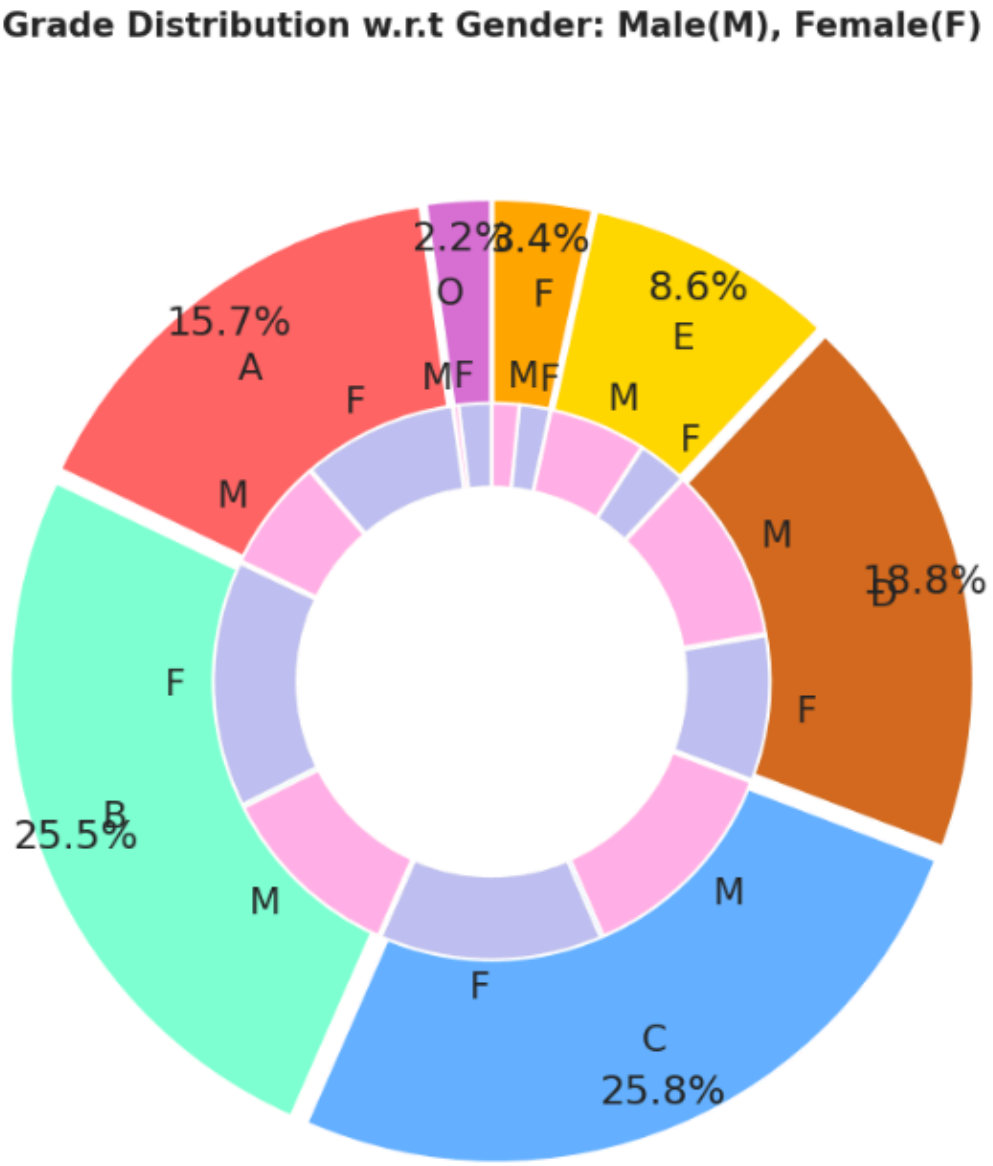


Figure 5.6: Grade Distribution w.r.t vs. Male and Female

3. Subject-wise Performance Analysis

- **Mathematics and Reading Scores**: The joint plot reveals a moderate correlation between reading and math scores, with gender-based differentiation. This insight is crucial for understanding the academic challenges faced by students in different subjects.

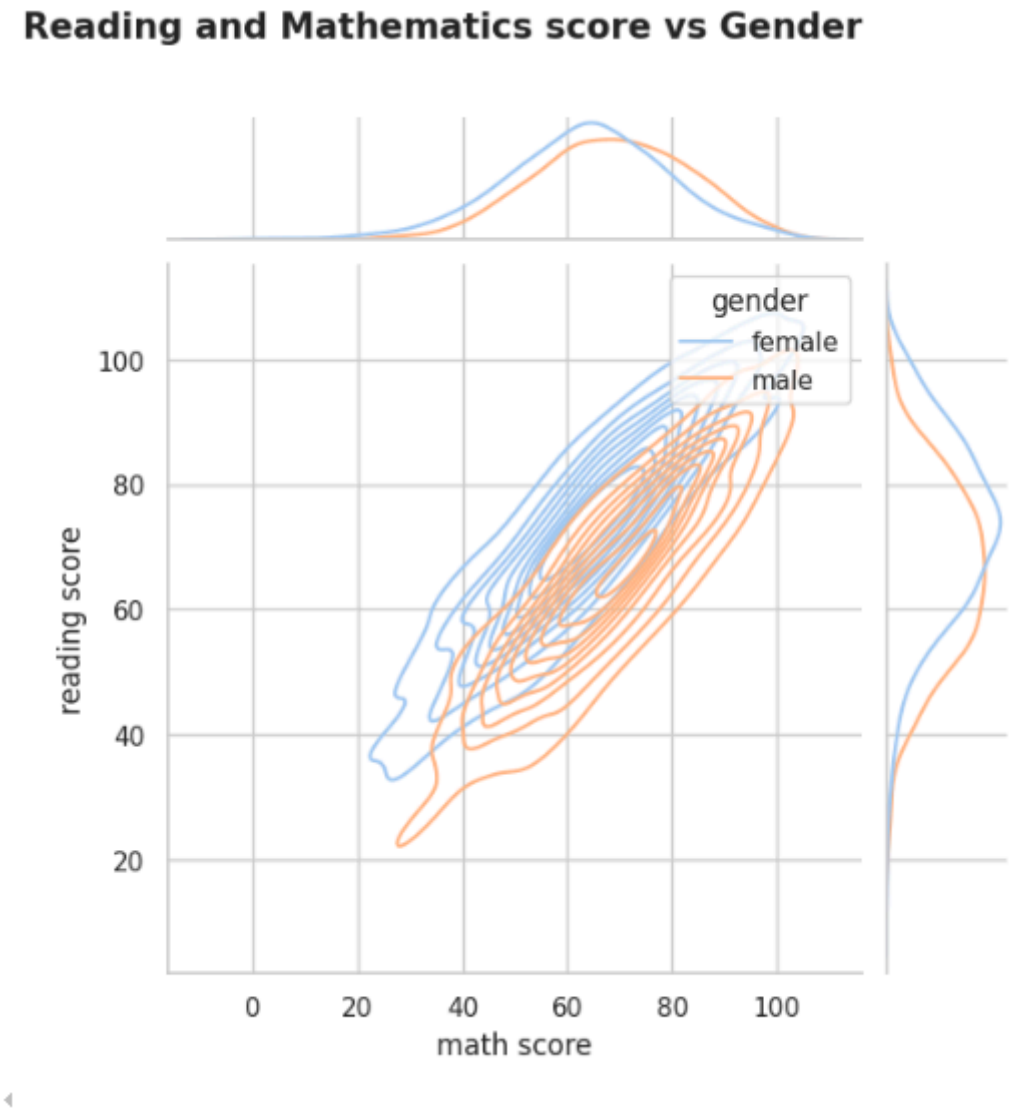


Figure 5.7: Reading and Mathematics score vs Gender

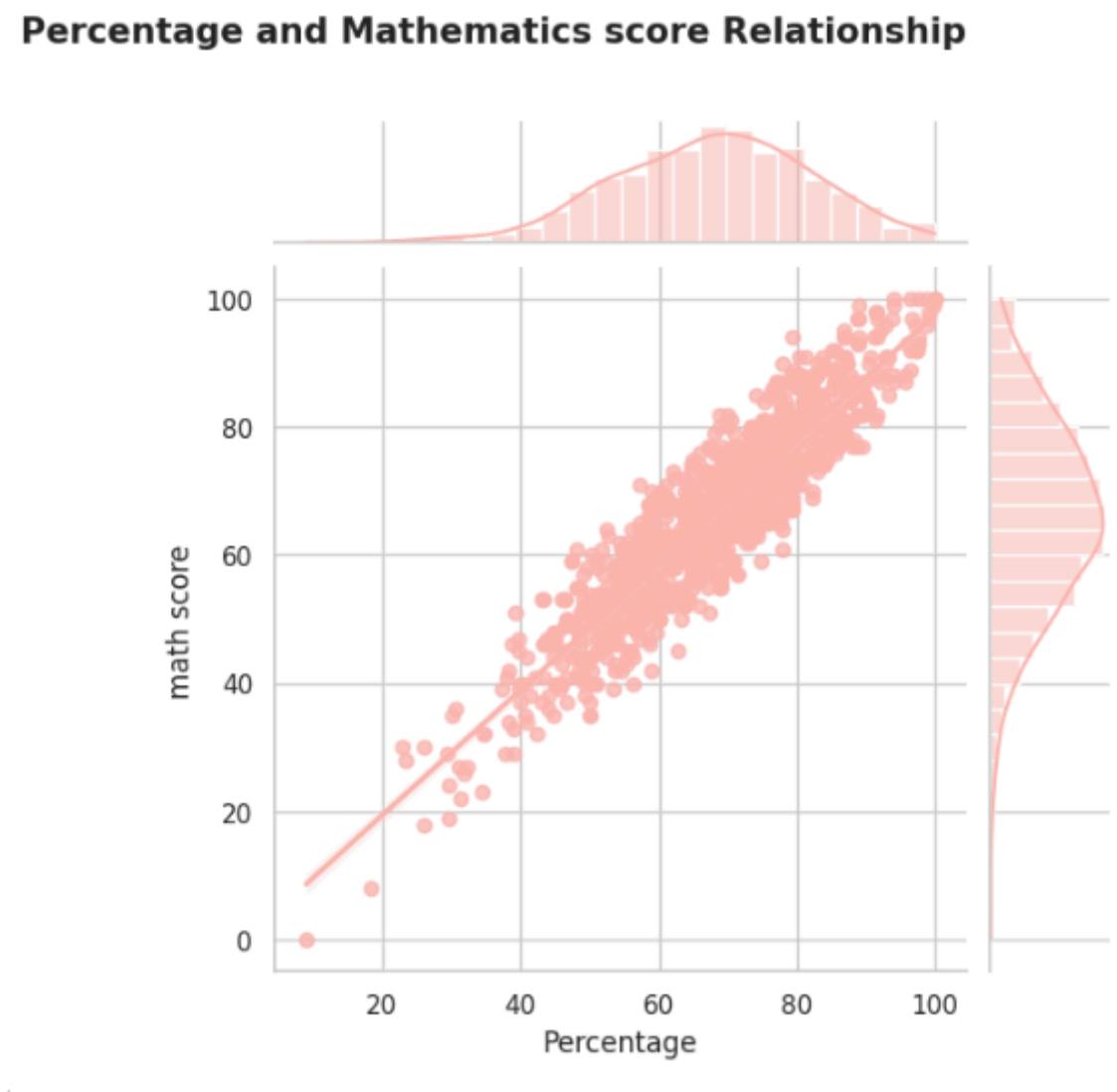


Figure 5.8: Percentage and Mathematics score Relationship

- **Writing and Mathematics Scores**: Most students score between 40 and 85 in both math and writing, with some variance in performance. The relationship between these scores is important to evaluate the students' overall academic capabilities.

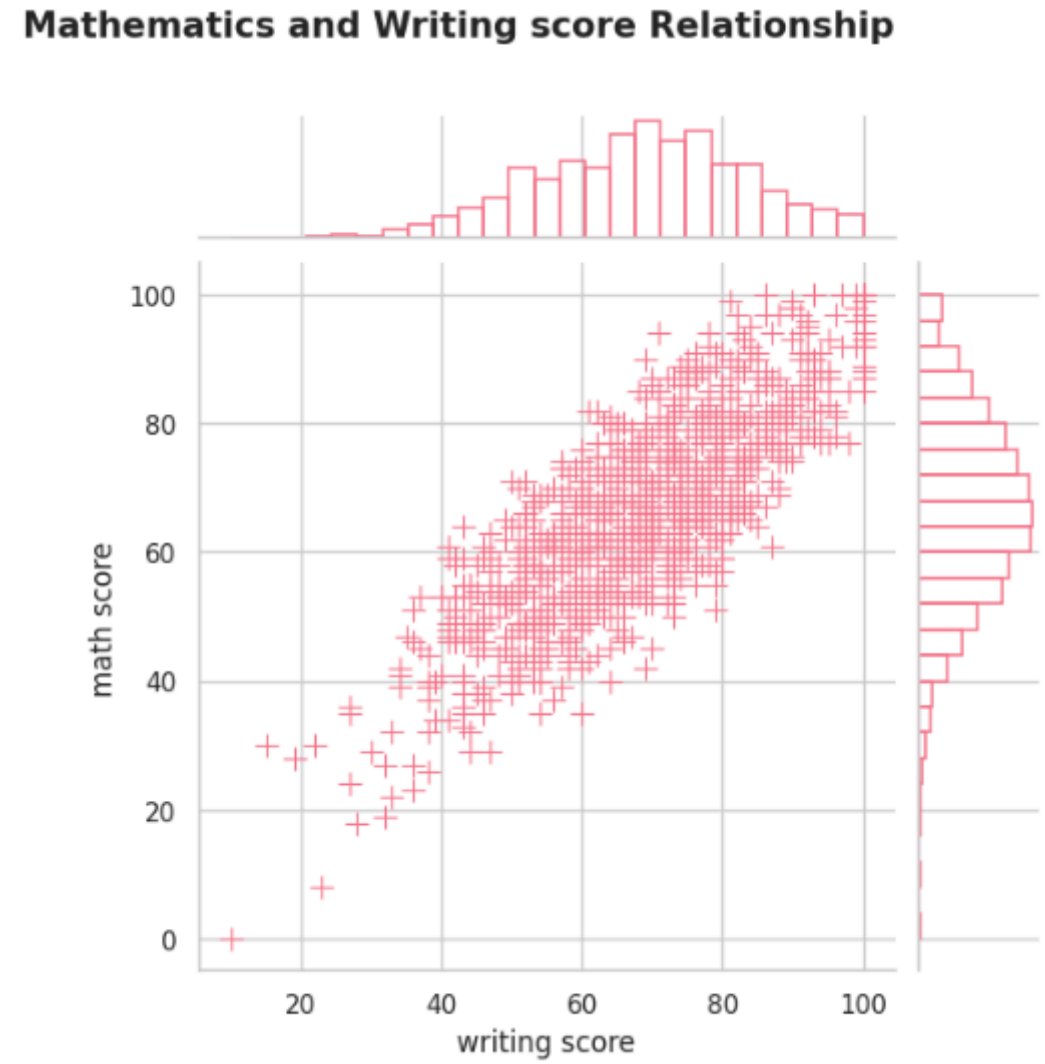


Figure 5.9: Mathematics and Writing score Relationship

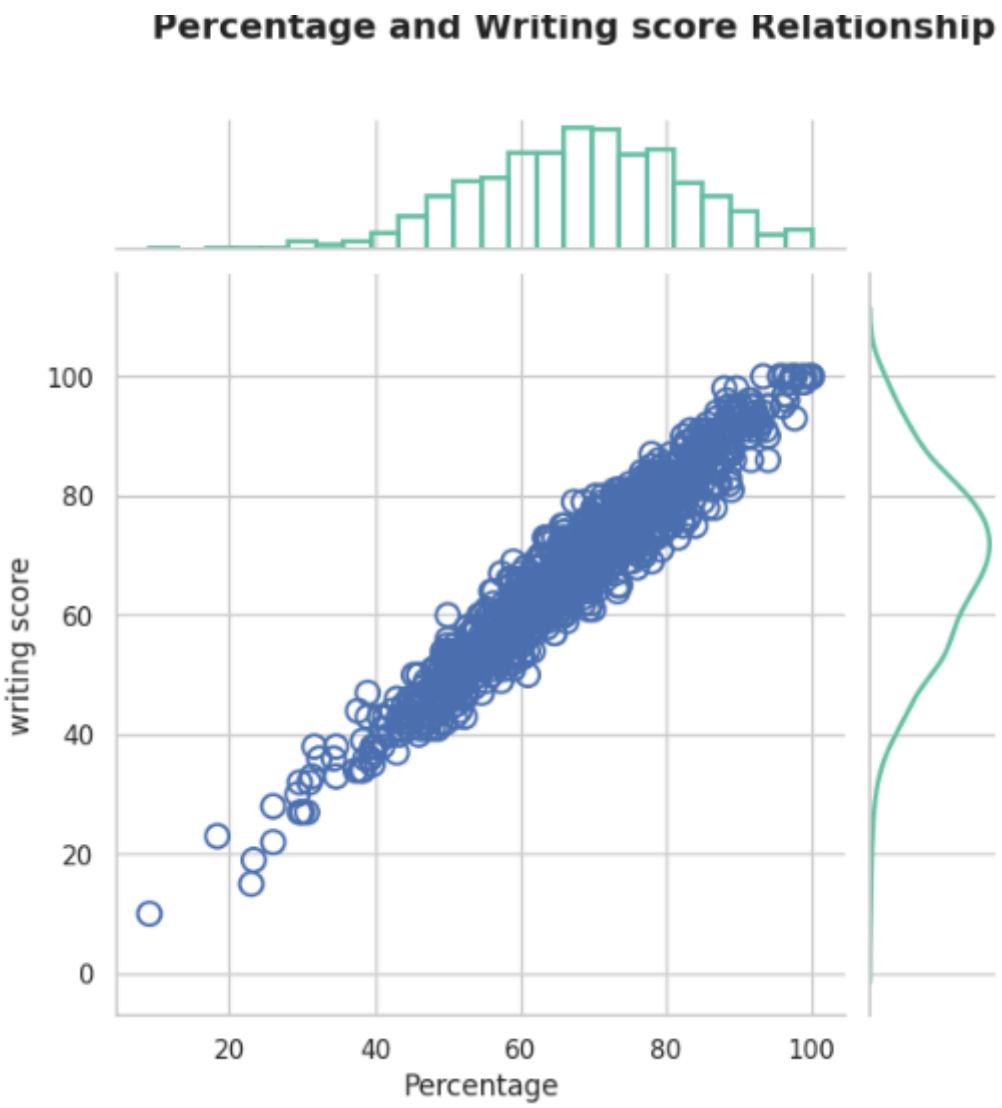


Figure 5.10: Percentage and Writing score Relationship

- **Percentage vs. Subject Scores**: The analysis indicates that students generally score between 50 and 80 in math, writing, and reading. Specifically, the correlation between percentage scores and math scores shows that most students are in the 50-80 range, highlighting the importance of improving students' foundational skills.

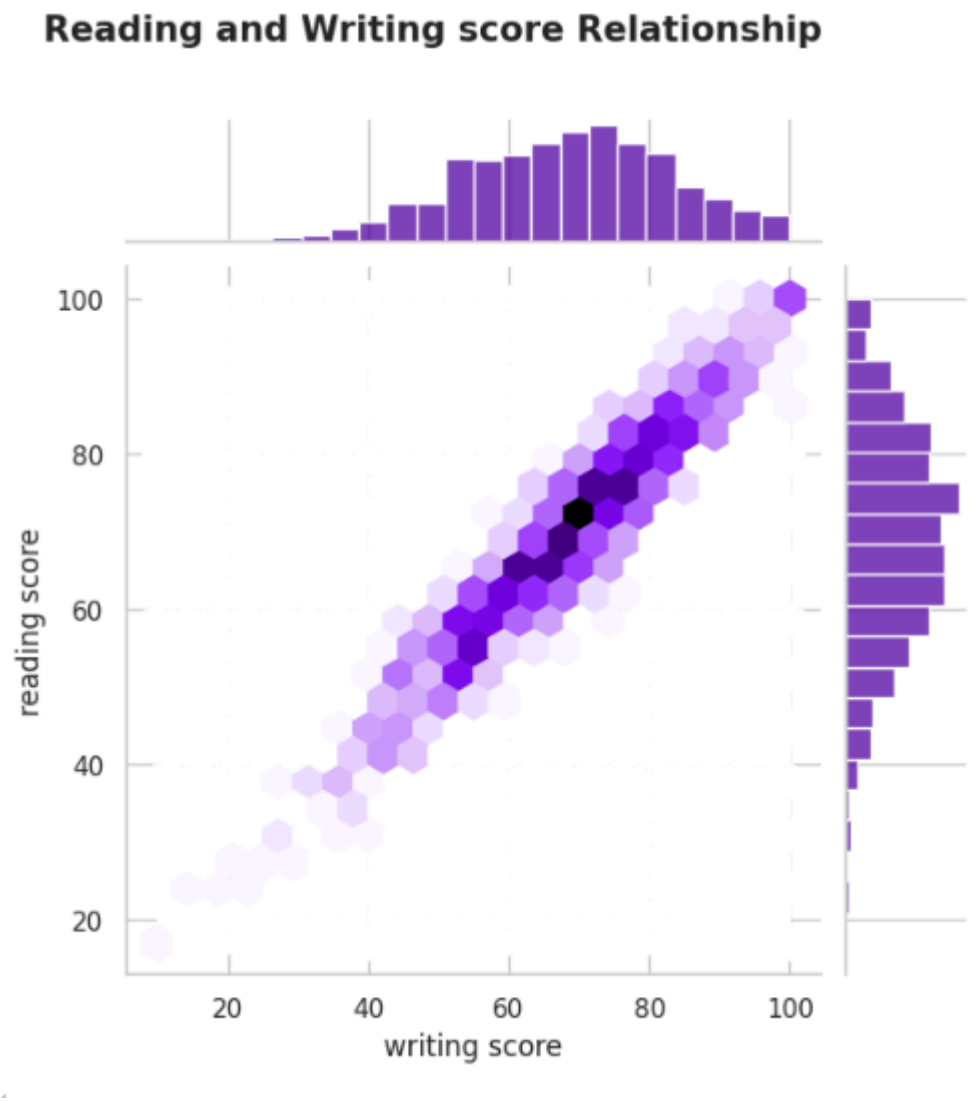


Figure 5.11: Reading and Writing score Relationship

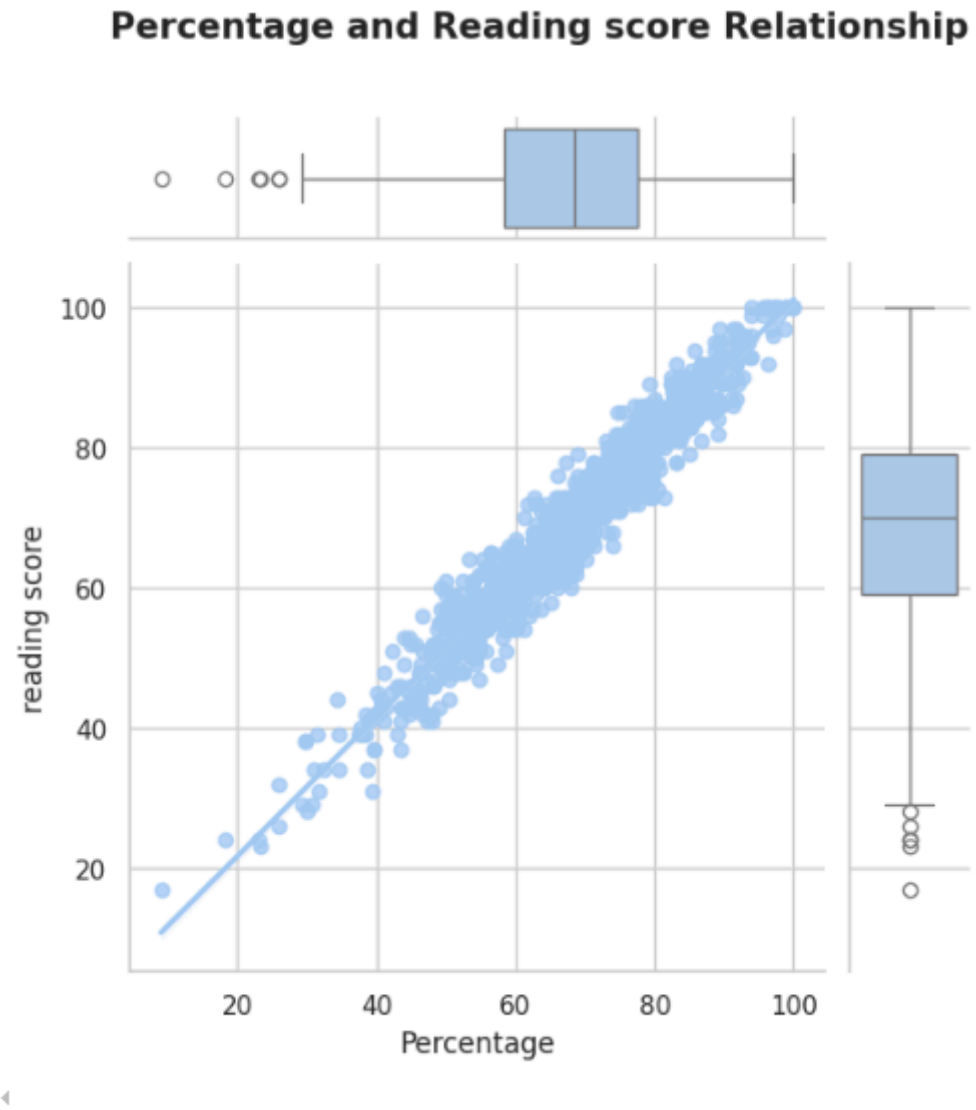


Figure 5.12: Percentage and Reading score Relationship

4. Test Preparation Course Impact

A critical observation is that students who completed the test preparation course performed better across the board. This suggests that proper preparation can positively impact students' performance, with a noticeable gap between students who completed the preparation and those who did not. A few students who did not complete the test preparation still performed exceptionally well, indicating the variability in student capabilities.

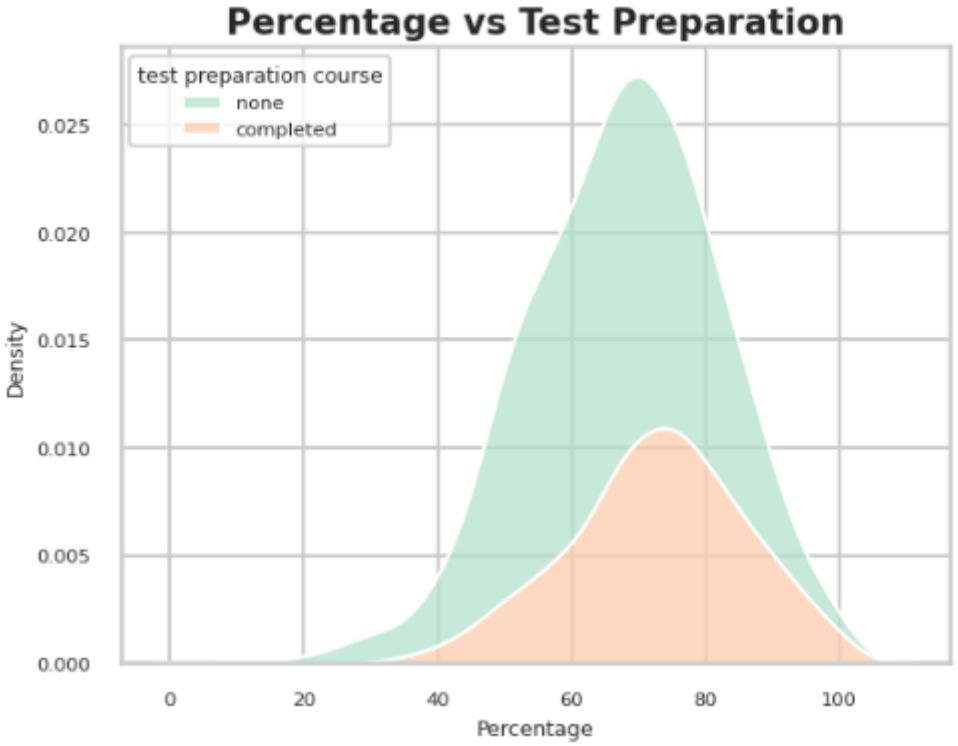


Figure 5.13: Percentage vs. Test Preparation

5. Effect of Lunch Type on Performance

The data reveals a clear trend where students who had standard lunch programs performed better, with a noticeable concentration of scores in the 75-100 range. In contrast, students who received free/reduced lunch performed poorly. This insight highlights the role of nutrition in academic performance, emphasizing that healthy meals contribute to better cognitive function and overall academic success. Adequate nutrition is essential for both physical and mental development, which directly affects students' learning abilities.

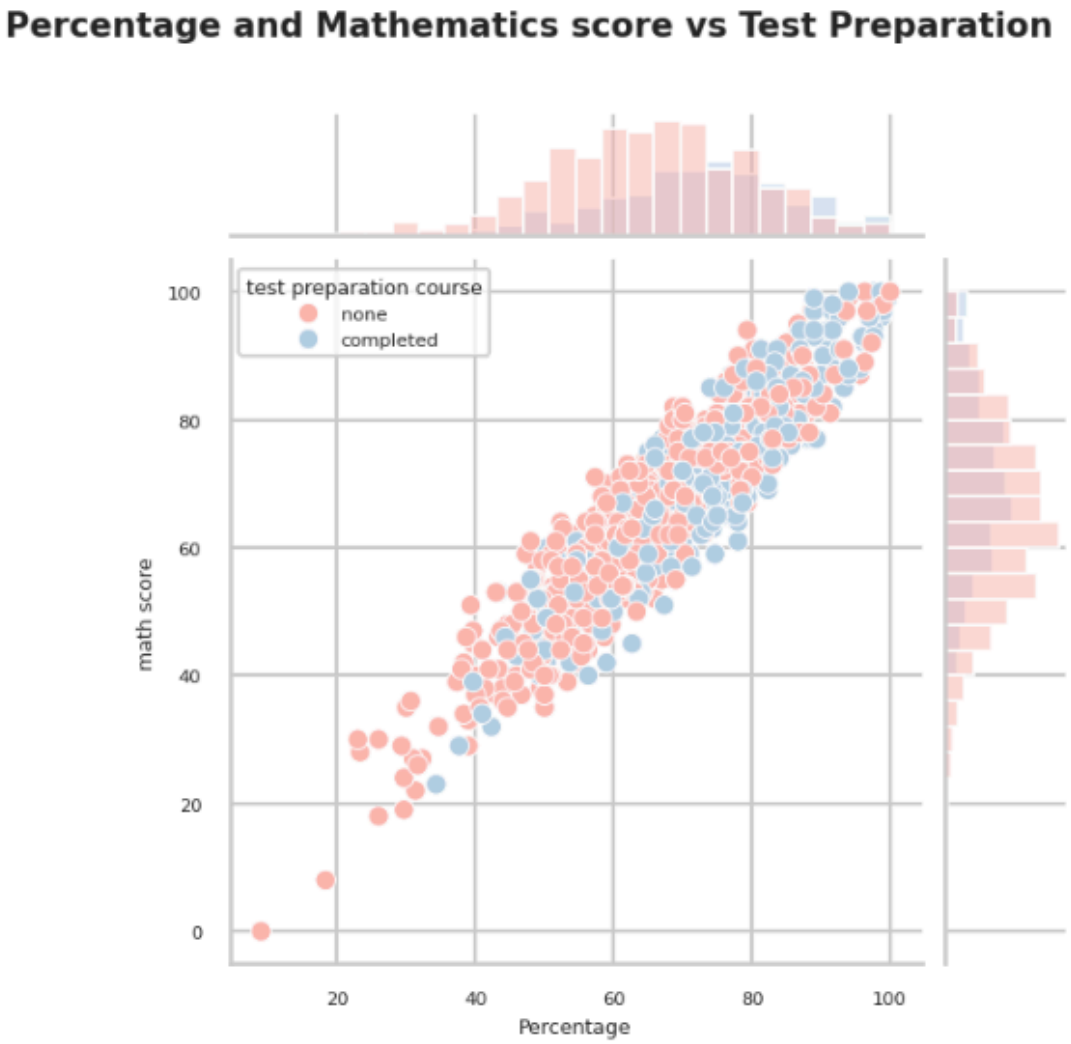


Figure 5.14: Percentage and Mathematics score vs. Test Preparation

6. Key Insights and Conclusions

- **Gender and Academic Performance**: Female students generally perform better academically, particularly in higher grade categories.

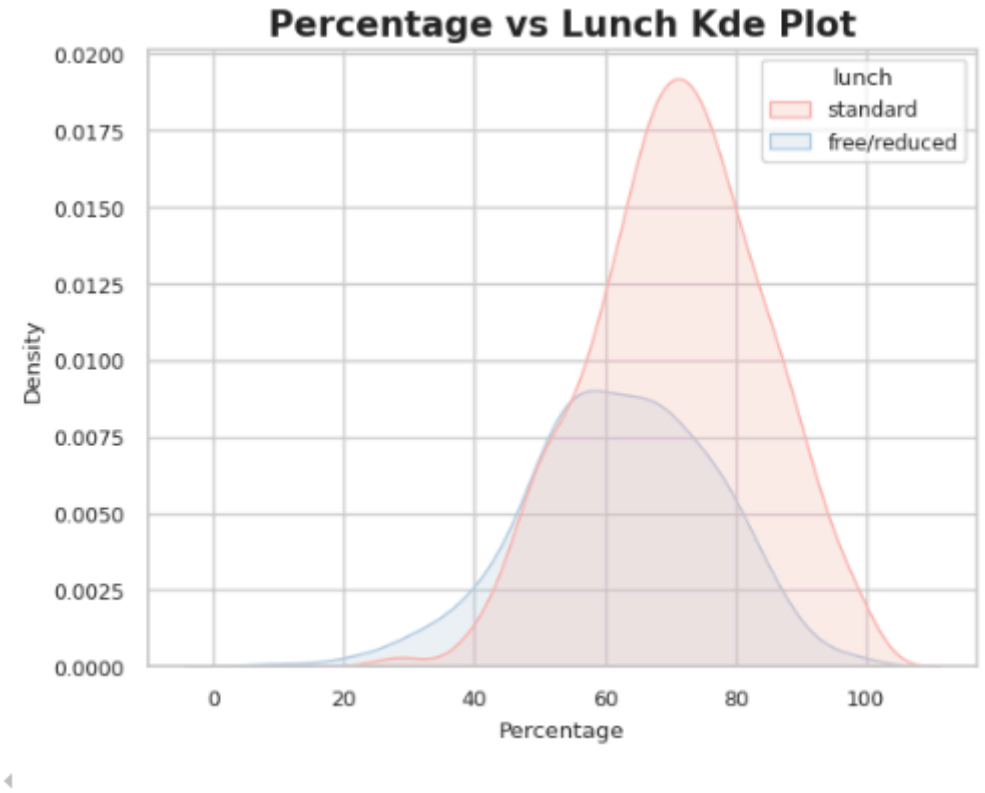


Figure 5.15: Percentage vs. Lunch KDE Plot

- **Impact of Test Preparation**: Completing test preparation courses significantly improves performance, although individual variations exist.

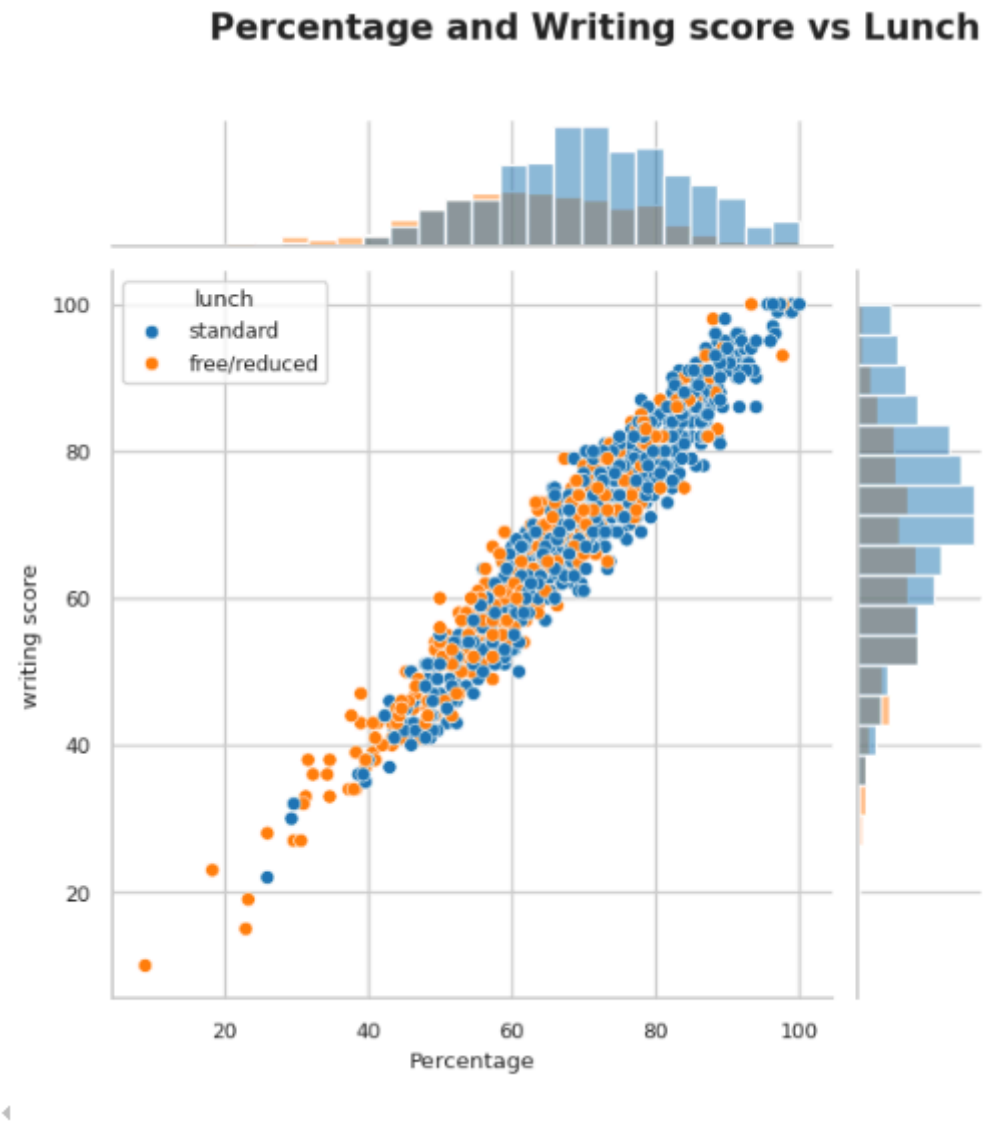


Figure 5.16: Percentage and Writing score vs. Lunch

- **Lunch and Academic Success**: Nutrition plays a vital role in student performance, with better outcomes linked to standard lunch programs.

These findings underscore the multifaceted nature of academic performance and the importance of addressing gender, preparation, and nutrition as key factors in enhancing student outcomes.

5.2.5 Outlier Detection and Handling

Outlier detection is a crucial step in data preprocessing, as outliers can significantly distort statistical analyses and model predictions. In this study, outliers

were identified using the Interquartile Range (IQR) method for key numerical variables, including math score, reading score, writing score, and percentage.

Outlier Detection

Outliers were detected by evaluating values that fall outside the defined range, calculated as 1.5 times the interquartile range (IQR). The **math score -** Table 5.6 column had 8 outliers, with extreme values like 0, 8, and 18, which may indicate data entry errors or anomalies (Table 5.6: Tukey, 1977; Chou et al., 2020). Similar outliers were found in **reading score -** Table 5.7, **writing score -** Table 5.8, and **percentage -** Table 5.9 columns, requiring data validation to ensure accuracy.

Table 5.6: Outliers in Math Score

Gender	Race/Ethnicity	Parental Level of Education	Lunch	Test Preparation Course	Math Score	Reading Score
female	group B	some high school	free/reduced	none	18	32
female	group C	some high school	free/reduced	none	0	17
female	group C	some college	free/reduced	none	22	39
female	group B	some high school	free/reduced	none	24	38
female	group D	associate's degree	free/reduced	none	26	31

Table 5.7: Outliers in Reading Score

Gender	Race/Ethnicity	Parental Level of Education	Lunch	Test Preparation Course	Math Score	Reading Score
female	group C	some high school	free/reduced	none	0	17

Gender	Race/Ethnicity	Parental Level of Education	Lunch	Test Preparation Course	Math Score	Reading Score
male	group E	some high school	standard	none	30	26
male	group C	some college	free/reduced	none	35	28
male	group A	some college	free/reduced	none	28	23
male	group B	high school	free/reduced	none	30	24

Table 5.8: Outliers in Writing Score

Gender	Race/Ethnicity	Parental Level of Education	Lunch	Test Preparation Course	Math Score	Reading Score
female	group C	some high school	free/reduced	none	0	17
male	group E	some high school	standard	none	30	26
male	group A	some college	free/reduced	none	28	23
male	group B	high school	free/reduced	none	30	24
female	group B	high school	free/reduced	none	8	24

Table 5.9: Outliers in Percentage

Gender	Race/Ethnicity	Parental Level of Education	Lunch	Test Preparation Course	Math Score	Reading Score
female	group B	some high school	free/reduced	none	18	32
female	group C	some high school	free/reduced	none	0	17
male	group E	some high school	standard	none	30	26
male	group A	some college	free/reduced	none	28	23
male	group B	high school	free/reduced	none	30	24

Outlier Handling

To address the identified outliers, extreme values were filtered based on predefined ranges:

- **Math score**: Retained between -3.0 and 137.0
- **Reading score**: Retained between -1.0 and 139.0
- **Writing score**: Retained between -6.0 and 142.75
- **Percentage**: Retained between 0.33 and 135.67

Additionally, negative values for scores were removed, as they are not logically feasible.

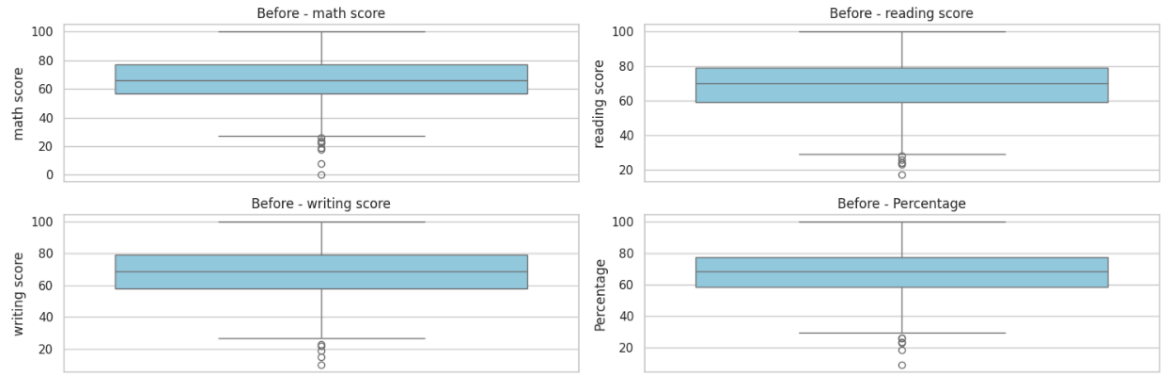


Figure 5.17: Boxplot before applied IQR

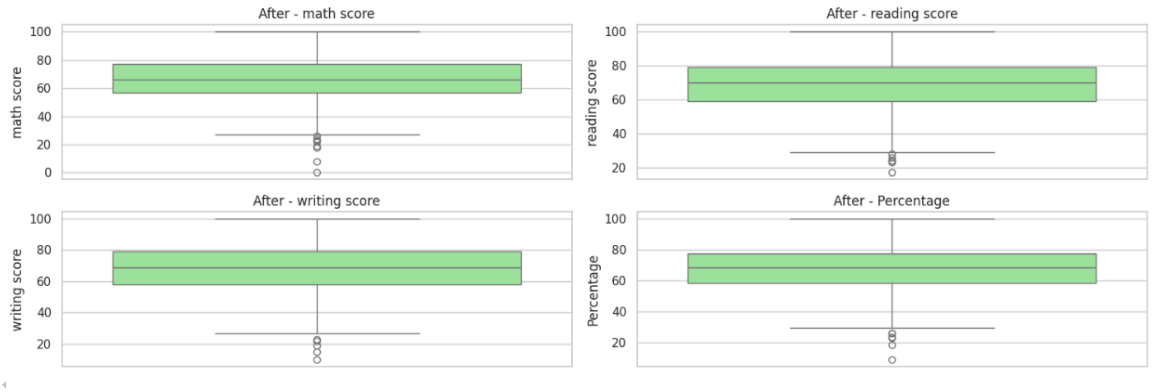


Figure 5.18: Boxplot after applied IQR

Visualizing Data Distribution

Before outlier handling, boxplots were used to visualize the distribution of scores. These plots clearly showed the presence of outliers, helping to identify data points that significantly deviate from the norm.

Key Findings

- **Impact of Outliers**: The removal of outliers helped to refine the dataset, ensuring that subsequent analyses and model training were based on more reliable data.
- **Factors Affecting Performance**:
    - **Lunch Type**: Students in the standard lunch program performed significantly better than those receiving free/reduced lunch.
    - **Parental Education**: Higher parental education levels were positively correlated with better student performance.
- **Effect of Test Preparation**: While completing a test preparation course was beneficial for some students, there was no strong, universal correlation between completing the course and improved academic performance.

These findings highlight the importance of addressing outliers to ensure data integrity and demonstrate that factors such as nutrition and parental education play a significant role in student success.

5.3 Probability Distribution Analysis

5.3.1 Distribution Analysis of Selected Variable

This section analyzes the association between students' academic performance—measured by their percentage scores—and key demographic factors such as gender, parental level of education, and race/ethnicity. The summarized findings are presented across multiple tables (**see** Table 5.10 **to** Table 5.13), highlighting patterns and variations among different demographic groups.

Table 5.10: Sample Records Filtered by Group B Ethnicity

gender	race/ethnicity	...	math score	reading score	writing score	Percentage	grade
female	group B	...	72	72	74	72.67	B
female	group B	...	90	95	93	92.67	A
female	group B	...	71	83	78	77.33	B
female	group B	...	88	95	92	91.67	A
male	group B	...	40	43	39	40.67	F
...	...	...	...	...	...	...	...
female	group B	...	75	84	80	79.67	B
male	group B	...	60	62	60	60.67	D
female	group B	...	8	24	23	18.33	F
male	group B	...	79	85	86	83.33	A
female	group B	...	65	82	78	75.00	B

Table 5.11: Distribution of Students by Grade Category

Grade	Count
A	157
B	255
C	258
D	188
E	86
F	34

Grade	Count
O	22

Table 5.12: Number of Unique Values by Gender and Column

Column	Female
Race/Ethnicity	18
Parental Level of Education	18
Lunch	18
Test Preparation Course	18
Math Score	18
Reading Score	18
Writing Score	18
Percentage	18
Grade	18

Table 5.13: Distribution of Students by Race/Ethnicity

Race/Ethnicity	Count
Group C	319
Group D	262
Group B	190
Group E	140
Group A	89

Gender-Based Performance Analysis

The analysis of the percentage distribution by gender revealed that **female students** tend to perform significantly better than male students. Visualizations using Kernel Density Estimation (KDE) showed a higher concentration of female

students achieving higher percentage scores, further emphasizing the gender disparity in academic performance.

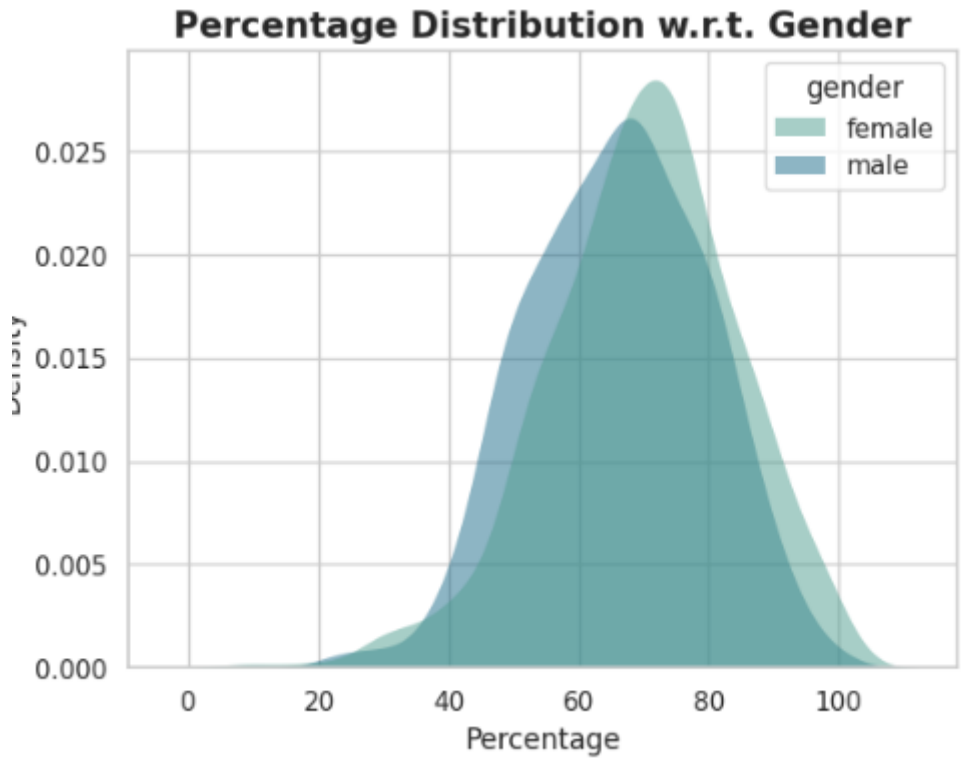


Figure 5.19: Percentage Distribution w.r.t. Gender

Parental Education and Student Performance

The educational background of students' parents was found to have a considerable impact on student performance. KDE plots comparing **parental level of education** and **percentage scores** revealed that students with **higher parental education levels**, particularly those whose parents held **master's degrees**, exhibited higher performance. On the other hand, students whose parents had lower educational attainment, such as high school or some high school, tended to score lower.

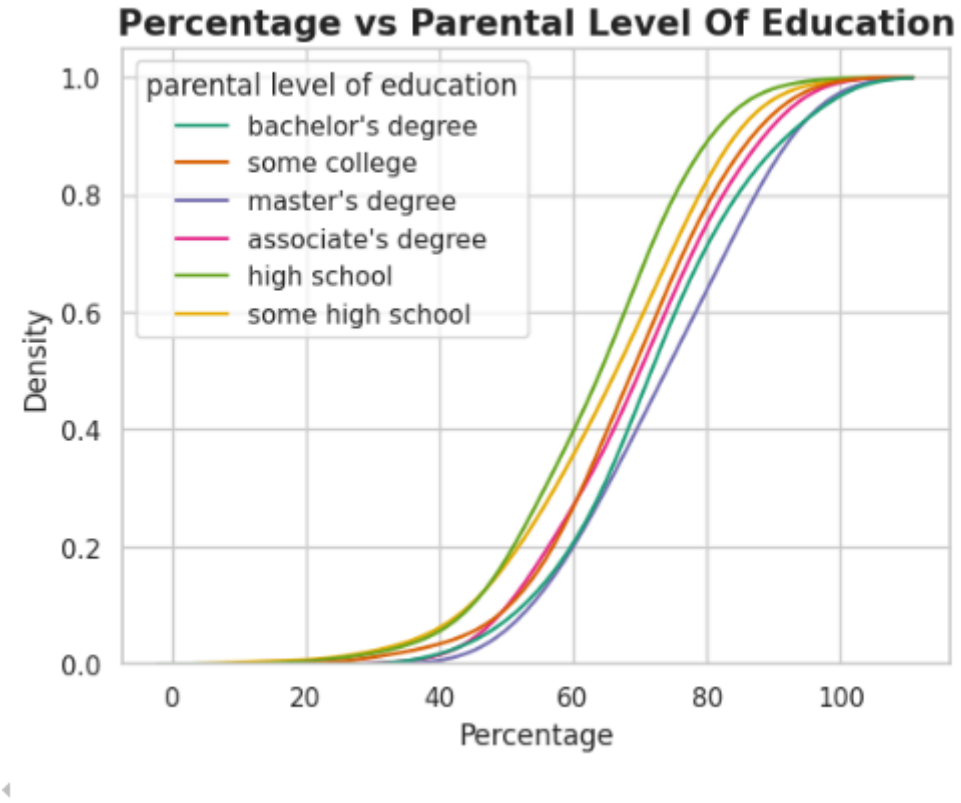


Figure 5.20: Percentage vs. Parental Level of Education

A **violin plot** further demonstrated that **female students** with parents who had a **bachelor's or master's degree** showed stronger academic success. Meanwhile, **male students** with similarly educated parents exhibited comparable performance levels, indicating that gender may influence academic outcomes differently across parental education backgrounds.

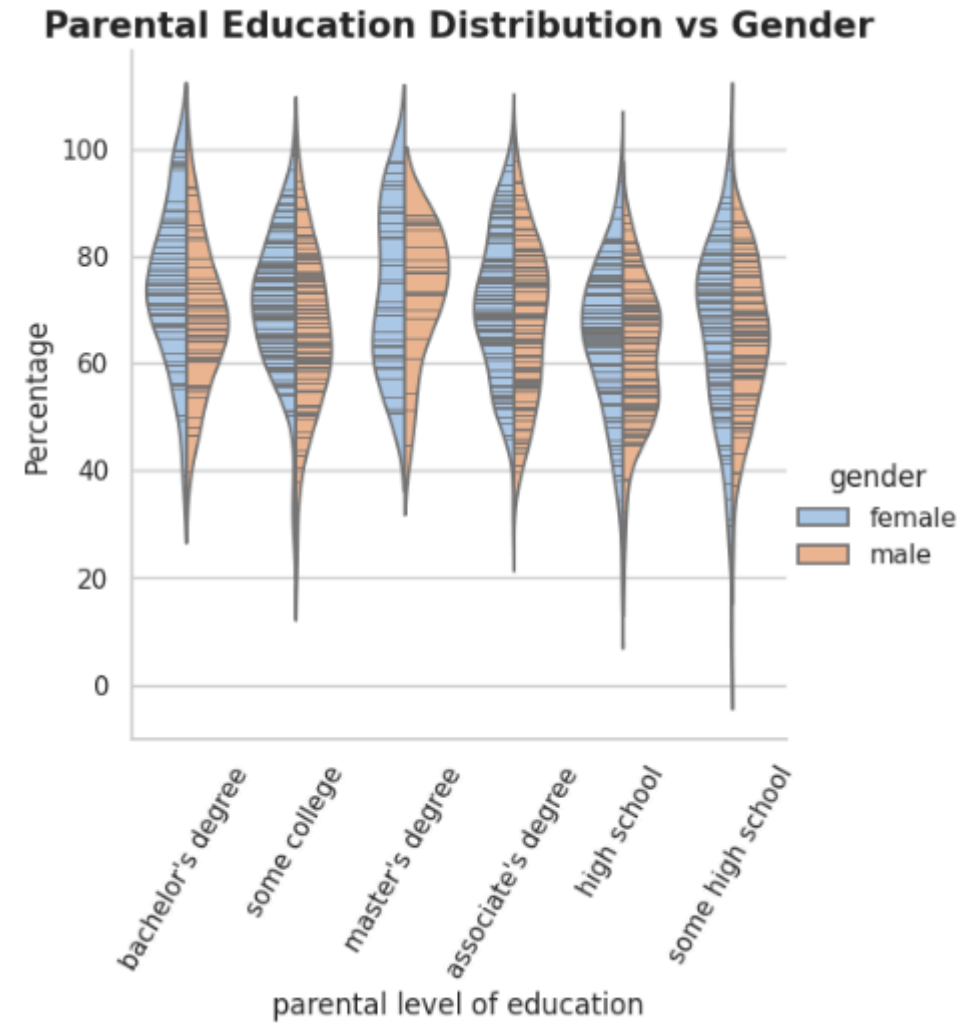


Figure 5.21: Parental Eduaction Distribution vs. Gender

Race/Ethnicity and Academic Performance

The **race/ethnicity** variable was analyzed by examining the distribution of students across different ethnic groups: **Group A**, **Group B**, **Group C**, **Group D**, and **Group E**. A **pie chart** illustrated the proportion of students in each ethnic group, with Group C (319 students) representing the largest proportion, followed by Group D (262 students) and Group B (190 students).

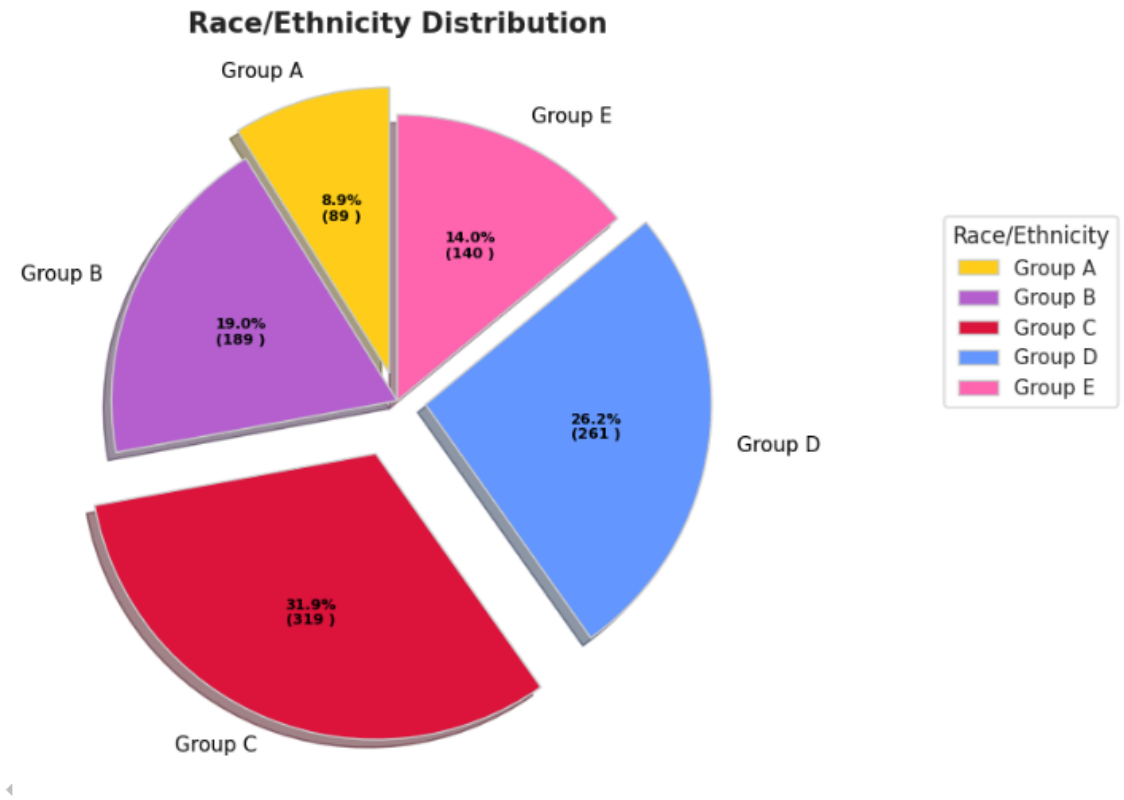


Figure 5.22: Race / Ethnicity Distribution

KDE plots revealed interesting trends in performance across different racial/ethnic groups. **Group E** consistently demonstrated the highest academic performance, with students from this group scoring the highest on average. **Group D** and **Group C** showed similar performance levels, whereas **Group A** exhibited lower overall performance. These findings suggest that race and ethnicity are important factors influencing educational outcomes.

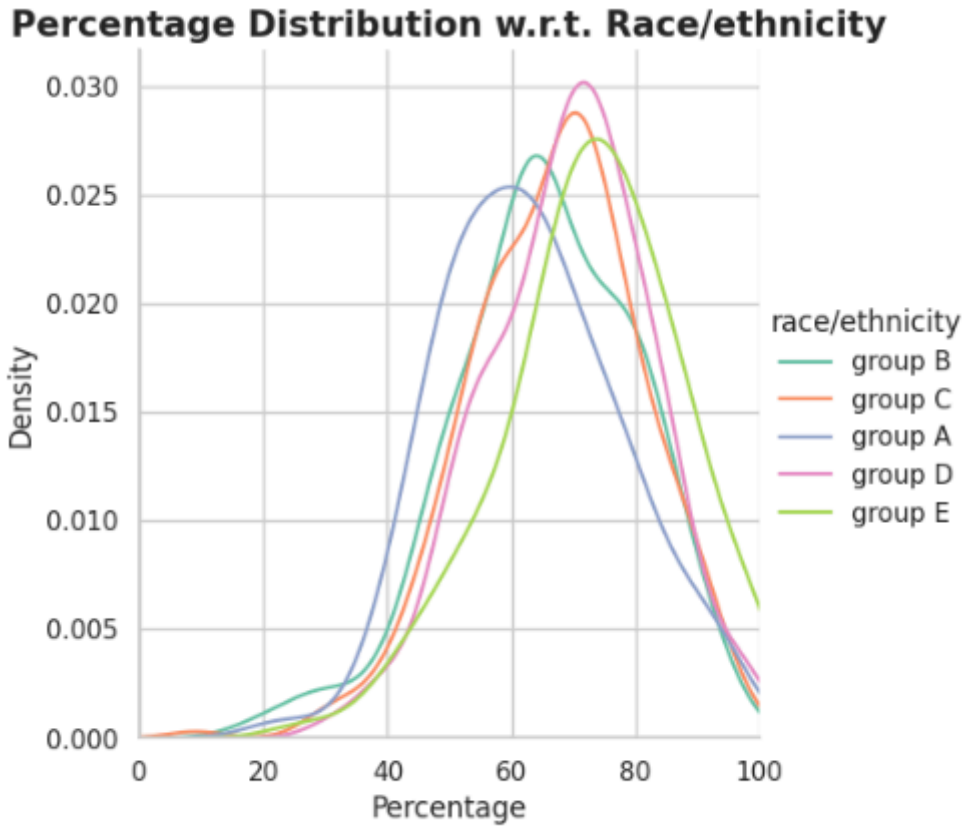


Figure 5.23: Percentage Distribution w.r.t. Race / Ethnicity

Key Insights:

- **Gender Disparities**: Female students generally outperformed male students in academic performance.
- **Parental Education**: Higher parental education, particularly master's degrees, is associated with better student performance, especially among female students.
- **Race/Ethnicity Impact**: Students from **Group E** had the highest average scores, whereas **Group A** showed lower academic performance, with **Groups D and C** performing similarly.

These findings highlight the multifaceted nature of student performance, where **gender**, **parental education**, and **race/ethnicity** all play significant roles in shaping academic outcomes. This analysis underscores the need for targeted interventions that consider these demographic factors to improve educational equity.

5.4 Hypothesis Testing

In this section, we apply statistical tests to examine the differences and relationships between various variables in our dataset, focusing on **gender**, **test preparation course participation**, and **race/ethnicity**. The primary objective is to test the following research questions.

5.4.1 Research Questions

The following research questions are proposed for hypothesis testing:

1. **Is there a difference in math scores between male and female students?**
2. **Is there a relationship between gender and participation in a test preparation course?**
3. **Do reading scores differ across racial/ethnic groups?**

5.4.2 Applying the Statistical Tests

To address these research questions, we perform the appropriate statistical tests:

1. **Difference in Math Scores Between Male and Female Students**
    - **Dependent Variable**: Math score (continuous)
    - **Grouping Variable**: Gender (male/female, two groups)
    - **Test Applied**: Two-sample independent t-test

The t-test results show a **T-statistic** of **5.3832** and a **P-value** of **0.0000**, which indicates a statistically significant difference between the math scores of male and female students ($p < 0.05$).

2. **Relationship Between Gender and Test Preparation Course Participation**
    - **Variables**: Gender (male/female), Test preparation course (completed/none)
    - **Test Applied**: Chi-square test of independence

The Chi-square test results show a **Chi-square statistic** of **0.0155** and a **P-value** of **0.9008**, which suggests that there is **no significant relationship** between gender and participation in a test preparation course ($p > 0.05$).

3. **Difference in Reading Scores Across Racial/Ethnic Groups**
    - **Dependent Variable**: Reading score (continuous)
    - **Grouping Variable**: Race/ethnicity (5 groups: A, B, C, D, E)
    - **Test Applied**: One-way ANOVA

The results of the ANOVA test show an **F-statistic** of **16.3887** and a **P-value** of **0.0000004**, which indicates a **statistically significant difference** in reading scores across racial/ethnic groups ($p < 0.05$). However, this test only indicates that there is a difference but does not specify which groups differ from each other.

5.4.3 Resutls Interpretation

Math Scores by Gender: The hypothesis that there is no difference in math scores between male and female students (H_0: No difference) is rejected, as the p-value (0.0000) is less than 0.05. This suggests that there is a statistically significant difference in math scores between male and female students.

- **Mean Math Scores**:
    - Male: 68.73
    - Female: 63.63

While male students tend to have a higher average math score than female students, this conclusion is based on statistical significance and does not imply any practical implications.

Test Preparation Course Participation by Gender: The Chi-square test indicates that there is no significant relationship between gender and participation in the test preparation course ($p = 0.9008$). Thus, we do not have sufficient evidence to reject the null hypothesis (H_0: No relationship), suggesting that participation in the test preparation course is independent of gender.

Reading Scores by Race/Ethnicity: The ANOVA test reveals a significant difference in reading scores across the racial/ethnic groups ($p = 0.0000004$). However, further post-hoc tests are needed to determine which specific groups differ from one another, as ANOVA only tells us that a difference exists but does not identify where the differences lie.

5.5 Correlation Analysis

This section investigates the relationships between numerical variables within the dataset. We explore both Pearson and Spearman correlations, as well as visualize these correlations to gain deeper insights into the data.

5.5.1 Computing Correlation between Numerical Variables

To examine the correlation between numerical variables in the dataset, we calculate both **Pearson -** Table 5.14 and **Spearman -** Table 5.15 correlations:

- **Pearson Correlation** measures linear relationships between variables. It is suitable for variables with a normal distribution.
- **Spearman Correlation** assesses monotonic relationships and is useful when data does not follow a normal distribution.

The correlation matrices for both methods are computed and displayed below:

Table 5.14: Pearson Correlation Between Scores and Overall Percentage

	Math Score	Reading Score	Writing Score	Percentage
Math Score	1.000	0.818	0.803	0.919
Reading Score	0.818	1.000	0.955	0.970
Writing Score	0.803	0.955	1.000	0.966
Percentage	0.919	0.970	0.966	1.000

Table 5.15: Spearman Correlation Between Scores and Overall Percentage

	Math Score	Reading Score	Writing Score	Percentage
Math Score	1.000	0.804	0.778	0.909
Reading Score	0.804	1.000	0.949	0.969
Writing Score	0.778	0.949	1.000	0.959
Percentage	0.909	0.969	0.959	1.000

5.5.2 Visualizing Correlation

To better understand the relationships between the numerical variables, we use a heatmap to visualize the Pearson correlation matrix. The heatmap provides a clear visual representation of how each variable correlates with the others.

- **Key Findings**:
  - There is a strong positive correlation between **math score** and **percentage** (0.974), **reading score** and **writing score** (0.922), and between **writing score** and **percentage** (0.976).
  - **Math score** also shows a high positive correlation with **reading score** (0.826), which suggests that students performing well in math tend to score well in reading as well.

We also generate scatter plots for each pair of numerical variables, helping to visualize the linear or non-linear relationships.

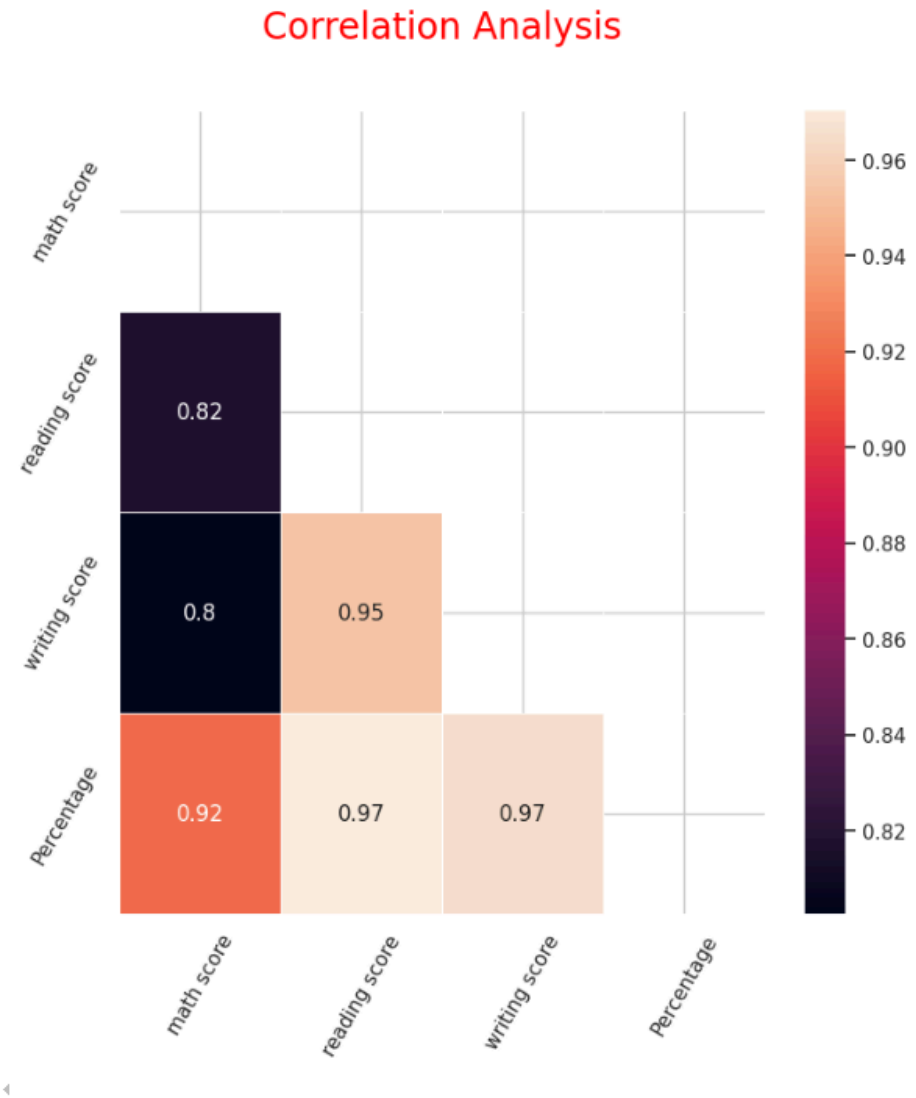


Figure 5.24: Correlation Analysis

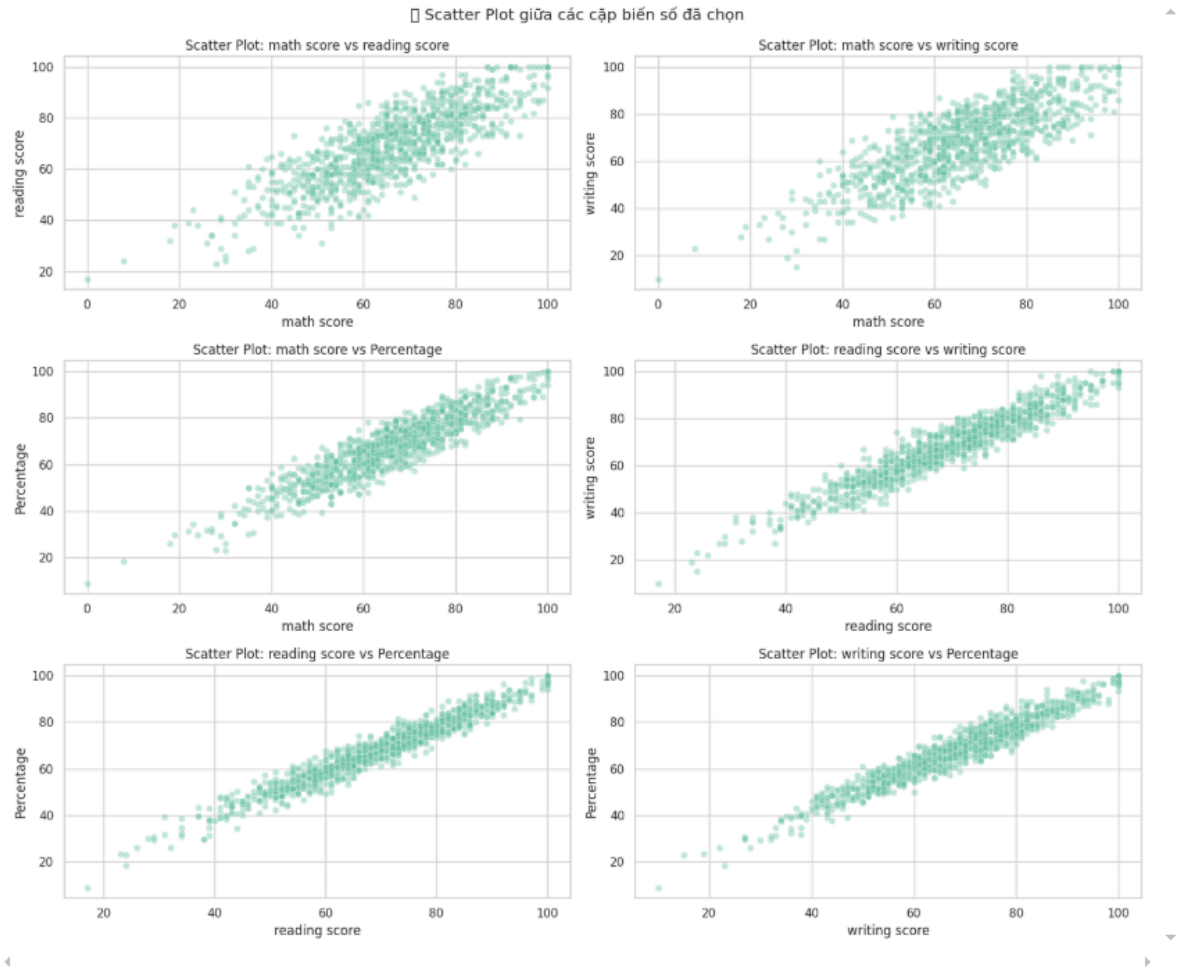


Figure 5.25: Scatter Plot Analysis

5.5.3 Interpretation and Real – World Implications

- **Math Score and Percentage**: The strong positive correlation between **math score** and **percentage** (0.974 for Pearson) implies that students who score well in math are likely to have a high overall percentage. This suggests that math performance is a strong indicator of overall academic performance.
- **Reading and Writing Scores**: The high correlation between **reading score** and **writing score** (0.922 for Pearson) indicates that students who perform well in reading also tend to perform well in writing. This is consistent with the expectation that proficiency in one language skill (reading) often correlates with proficiency in another (writing).

- **Implications for Educational Strategies**: The correlations observed suggest that interventions aimed at improving performance in one subject (e.g., math or reading) could potentially have a positive impact on other related subjects. For example, focusing on improving math skills might indirectly lead to improvements in overall academic performance (percentage) and reading ability.

CHAPTER 6. CONCLUSION AND RECOMMENDATIONS

6.1 Result

In this project, we explored and analyzed a dataset using basic data analysis and visualization techniques. The process began with data cleaning and preprocessing to ensure accuracy and consistency. After that, we applied descriptive statistics and visual tools to uncover trends, patterns, and key insights hidden in the data. Various types of charts such as bar charts, line graphs, pie charts, and heatmaps were used to present the data in a clear and intuitive way. These visualizations helped simplify complex datasets, making it easier to understand and interpret the information.

Overall, the project successfully demonstrated how raw data can be transformed into meaningful insights that support better and more informed decision-making.

6.2 Limit

Despite the positive outcomes, the project had a few limitations:

- The dataset contained missing values and potential outliers, which may have affected the reliability of some results.
- The analysis used basic tools and techniques, without involving advanced statistical or machine learning methods.
- The scope was limited by time and the current level of technical skills, especially in handling large-scale data or building fully interactive dashboards.

These limitations are common for entry-level data analysis projects and can be addressed in future improvements.

6.3 Future Orientation

To improve and extend this project in the future, several directions can be considered:

- **Apply advanced analytics techniques** such as clustering, regression, or classification to extract deeper insights and make predictions.
- **Integrate machine learning models** to automate data interpretation and enhance decision-making capabilities.
- **Develop interactive dashboards** using tools like Tableau, Power BI, or Dash to provide real-time data exploration for users.
- **Work with larger and more diverse datasets** to improve the generalizability and robustness of the analysis.

By continuing to learn and apply more advanced tools and methods, students can further develop their skills in data science and deliver more impactful solutions.

REFERENCES

[1] D. Chen, “Online Retail Dataset,” UCI Machine Learning Repository, 2010. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Online+Retail

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Wine Quality Dataset,” UCI Machine Learning Repository, 2009. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[3] Instacart, “Instacart Market Basket Analysis,” Kaggle, 2017. [Online]. Available: https://www.kaggle.com/c/instacart-market-basket-analysis

[4] J. Dean, *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Hoboken, NJ, USA: Wiley, 2014.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. New York, NY, USA: Springer, 2013.