Data and Artificial Intelligence
Cyber Shujaa Program

Week 3 Assignment
Exploratory Data Analysis - Titanic Dataset

Student Name: Rodney Roy Gitonga

Student ID: CS-DA03-26025

Introduction

This week's assignment focused on Exploratory Data Analysis (EDA), a fundamental step in the data science lifecycle used to summarize and visualize the main characteristics of a dataset. The goal was to analyse the Titanic Machine Learning from Disaster dataset to identify key factors that influenced passenger survival.

The objectives of the assignment were:

1. Perform initial data discovery to understand data types, missing values, and structure.
2. Clean the data by handling missing values (imputation) and managing outliers.
3. Conduct Univariate Analysis to visualize distributions of single variables (e.g., Age, Fare).
4. Conduct Bivariate Analysis to explore relationships between two variables (e.g., Sex vs. Survival).
5. Conduct Multivariate Analysis to examine complex interactions (e.g., Age, Class, and Survival combined).
6. Use Python libraries (Pandas, Matplotlib, Seaborn) on Kaggle to automate this analysis.

Tasks Completed

Below is the step-by-step sequence of tasks completed to achieve the assignment objectives, supported by code snippets and visual evidence.

Task 1: Initial Data Exploration & Setup

I began by loading the dataset into the Kaggle environment. To ensure robustness, I implemented a script that checks multiple standard file paths for the Titanic dataset. I then inspected the data structure using .head(), .info(), and .describe() to identify data types and missing values.

```
print("\n Data Snapshot (First 5 Rows) ")
display(df.head())

print(f"\n- Dataset Shape -\nRows: {df.shape[0]}, Columns: {df.shape[1]}")

print("\n Data Info ")
df.info()

print("\n Summary Statistics ")
display(df.describe())

print("\n Duplicate Check ")
print(f"Duplicate Rows: {df.duplicated().sum()}")

print("\n Missing Values Check ")
print(df.isnull().sum())
```

Task 2: Handling Missing Values and Outliers

My analysis revealed that 'Age' had significant missing values (~19%), and 'Cabin' was missing over 75% of its data.

- Imputation: I filled missing 'Age' values with the median (28.0) because it is robust against outliers. Missing 'Embarked' values were filled with the mode ('S').
- Dropping: I dropped the 'Cabin' column as it had too much missing data to be useful.
- Outliers: I detected extreme outliers in the 'Fare' column (tickets costing over $500). I capped these outliers at the upper Interquartile Range (IQR) bound to normalize the distribution without deleting valuable data points.

```
# Outlier Detection (Fare)
Q1 = df['Fare'].quantile(0.25)
Q3 = df['Fare'].quantile(0.75)
IQR = Q3 - Q1
upper_bound = Q3 + 1.5 * IQR

df['Fare'] = np.where(df['Fare'] > upper_bound, upper_bound, df['Fare'])
print(f"Fare outliers capped at: {upper_bound:.2f}")
```

Task 3: Univariate Analysis

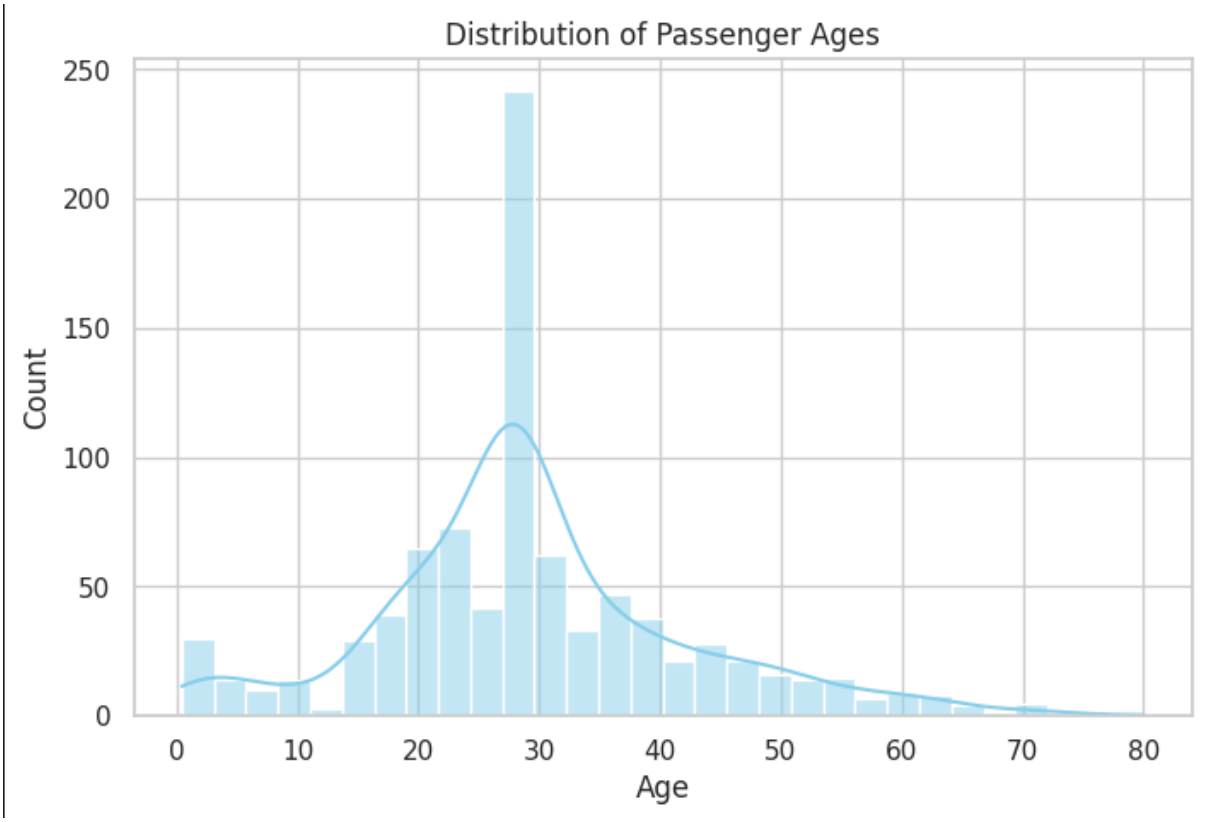
I analysed individual features to understand their distributions:

- Age: The distribution was slightly right skewed, indicating a younger passenger demographic (20-30 years).
- Target (Survived): The dataset is imbalanced, with significantly more casualties (0) than survivors (1).

```
print("\n 4. UNIVARIATE ANALYSIS ")

# A. Age
plt.figure(figsize=(8, 5))
sns.histplot(df['Age'], kde=True, bins=30, color='skyblue')
plt.title('Distribution of Passenger Ages')
plt.show()

# B. Embarked [FIXED WARNING 1]
plt.figure(figsize=(6, 4))
# Added hue='Embarked' and legend=False
sns.countplot(x='Embarked', data=df, hue='Embarked', legend=False, palette='viridis')
plt.title('Count of Passengers by Embarkation Port')
plt.show()
```



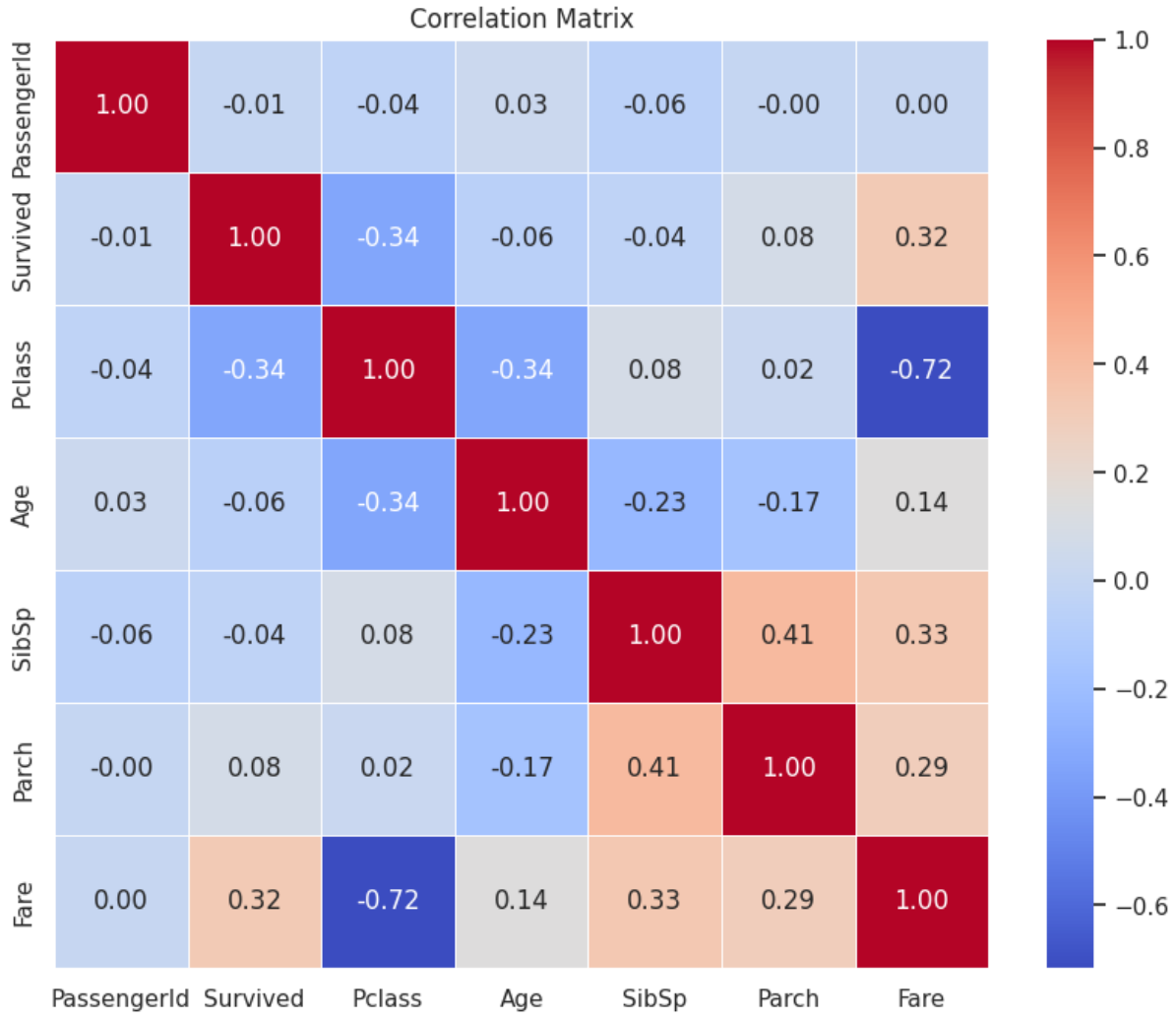
Task 4: Bivariate & Multivariate Analysis

I explored relationships between variables to find survival factors:

- Sex vs. Survival: Female passengers had a much higher survival rate than males.
- Class vs. Survival: First-class passengers were more likely to survive than those in lower classes.
- Multivariate (Class + Age + Survival): Using violin plots, I observed that children in 1st and 2nd class had very high survival rates, whereas survival for adult males in 3rd class was extremely low.

```
print("\n 7. MULTIVARIATE ANALYSIS ")

# A. Pclass, Age, Survival
plt.figure(figsize=(12, 6))
# This one was already fine because it had a hue
sns.violinplot(x='Pclass', y='Age', hue='Survived', data=df, split=True, inner='quart', palette
='Set2')
plt.title('Survival by Pclass and Age')
plt.show()
```



Link to Code:

https://www.kaggle.com/code/fytroy/rodney-roy-gitonga-exploratory-data-analysis

Conclusion

This week I gained a solid understanding of how to systematically explore a dataset. I learned that survival on the Titanic was not random; socio-economic status (Ticket Class) and demographics (Gender and Age) played critical roles. Women and children in First Class had the highest probability of survival, while adult males in Third Class had the lowest.

Technically, I improved my skills in handling missing data through imputation and managing outliers using statistical methods like IQR. I am now more confident in using libraries like Seaborn to create meaningful visualizations that tell a story about the data. I look forward to applying these EDA techniques to more complex datasets in future assignments.