Task Explanation: Exploratory Data Analysis (EDA)

The objective of this task was to explore and understand the Titanic dataset (tested.csv) using visual and statistical tools in Python. EDA is a critical step in any data science project that helps in identifying patterns, trends, anomalies, and relationships within the dataset before applying machine learning models. Tools used include Pandas for data handling, Matplotlib and Seaborn
for visualization.

Key Points Covered

- Data Loading and Preview: Used .read_csv() and .head() to view the initial structure of the dataset.

- Basic Structure: Used .info(), .describe(), .dtypes, .nunique() to understand data types, summary statistics, and unique value counts.

- Missing Values: Identified columns like Age, Cabin, and Embarked with missing data.

Univariate Analysis:

- Plotted histograms for distribution of numerical features.
- Used boxplots to detect outliers in Age and Fare.
- Countplots for categorical variables like Pclass, Sex, and Embarked.

- Bivariate Analysis:
  - Pairplots to analyze relationships between features, especially in relation to survival.
  - Heatmap to check correlation between numerical variables.
  - Scatterplot to visualize age vs. fare in relation to survival.

- Statistical Insights: Analyzed central tendency, spread, and correlations.

Summary of Findings

- The dataset contains 418 entries and 12 columns, a mix of numerical and categorical data.
- Significant missing values are present in Age, Cabin, and Embarked. These should be cleaned or imputed for accurate modeling.
- Most passengers are males and belong to Pclass 3, indicating many traveled in economy class.
- Fare has a high standard deviation and shows many outliers; some passengers paid significantly more than others.
- Age distribution is skewed toward younger passengers, mostly between 20 to 40 years.
- Survival is positively correlated with higher fare and negatively correlated with Pclass— passengers in 1st class had a better chance of survival.
- Visual analysis confirms survival is more common among women and high-paying passengers.

Conclusion

This EDA provided crucial insights into the Titanic dataset, helping us understand which features influence survival the most. Patterns like class, fare, and gender clearly stand out. The ana