Exploratory Data Analysis of Titanic Dataset

---

1. Observations

- **Data Loading and Basic Info:**
  - The Titanic dataset was successfully loaded.
  - The dataset has **891 rows** and **12 columns**.
  - Missing values were found in the 'Age', 'Cabin', and 'Embarked' columns.

- **Statistical Description:**
  - The mean age of passengers is around **29.7 years**.
  - The average fare paid by passengers is approximately **32.2**.
  - The dataset includes continuous variables like 'Age', 'Fare', and discrete variables like 'Survived', 'Pclass', etc.

- **Checking for Missing Values:**
  - 'Age' has **177** missing values.
  - 'Cabin' has **687** missing values.
  - 'Embarked' has **2** missing values.

- **Value Counts:**
  - 'Survived': About **549 passengers** did not survive, while **342 passengers** survived.
  - 'Pclass': Most passengers were from the **3rd class.**
  - 'Sex': The number of male passengers is higher than female passengers.

- **Histograms:**
  - Majority of passengers are around **30 years** old.
  - Distribution is slightly right-skewed, with fewer older passengers.
  - Notable clusters appear near ages **20 and 25**.

- **Boxplots:**
  - First-class survivors paid the highest fares, with notable outliers.
  - Second-class survivors had slightly higher fares than non-survivors, with less variation.
  - Third-class passengers showed the lowest fares, with many low-fare outliers.

- **Correlation Heatmap:**
  - 'Fare' and 'Pclass' have a moderate negative correlation.
  - 'Survived' has a weak positive correlation with 'Fare' and a negative correlation with 'Pclass'.

- **Scatter Plots:**
  - Higher fares are linked to better survival chances.
  - Most survivors are in younger to mid-age ranges.
  - Some outliers, like high-fare older non-survivors, stand out.

- **Label Encoding Process:**
    - The label encoding transformed the Color column into numerical values (Red → 2, Blue → 0, Green → 1), making it machine-readable.
    - The encoding follows alphabetical order, which ensures consistency but may introduce unintended numerical hierarchy.
    - For non-ordinal categories, one-hot encoding could be considered as an alternative to avoid misinterpretation.

---

2. Summary of Findings

- The dataset contains missing values, particularly in the 'Cabin' and 'Age' columns.
- Majority of the passengers belonged to the 3rd class.
- Male passengers outnumber female passengers in the dataset.
- Survival rate was lower than non-survival rate among passengers.
- Younger passengers (ages 20-40) formed the bulk of the population.
- 1st class passengers paid much higher fares and also had higher median age compared to others.
- Fare shows a right-skewed distribution, indicating few passengers paid a very high fare.
- Survival is positively associated with paying a higher fare and negatively associated with lower passenger class.