Titanic Dataset: EDA Report with Visual Insights

The Titanic dataset, provided by Kaggle, contains data on the passengers of the RMS Titanic. The goal is to predict survival based on features such as age, sex, ticket class, fare, and more. This dataset is widely used for binary classification problems and beginner-level machine learning practice.

1. Survival Distribution

A count plot shows the distribution of survivors and non-survivors. It reveals an imbalance, with more passengers not surviving.

```
sns.countplot(x='Survived', data=ship_data)
plt.title('Survival Distribution')
plt.show()
```

2. Pairplot of Numeric Features

Pairplots show relationships between numeric features like Age, Fare, Pclass, and Survived. Higher Fare and younger Age are weakly associated with survival.

```
sns.pairplot(encoded_data, hue='Survived', palette='coolwarm')
plt.show()
```

3. Histograms: Age and Fare

Histograms reveal that most passengers were in their 20s-30s. Fare distribution is right-skewed.

```
ship_data[['Age', 'Fare']].hist(bins=20, figsize=(10, 4))
plt.show()
```

4. Survival by Gender

Females had a significantly higher survival rate compared to males.

```
sns.countplot(x='Sex', hue='Survived', data=ship_data)
plt.title('Survival by Sex')
plt.show()
```

5. Correlation Heatmap

Heatmaps of numeric features show that Sex (encoded), Fare, and Pclass have notable correlations with survival.

```
sns.heatmap(encoded.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```