Exploratory Data Analysis (EDA) - Titanic Dataset

This report summarizes the results of an Exploratory Data Analysis (EDA) on the Kaggle Titanic dataset. The objective was to identify patterns, relationships, and potential predictors of passenger survival.

Dataset Overview:
- Rows & Columns: 891 × 12
- Types: Numerical (Age, Fare), Categorical (Sex, Pclass, Embarked), Mixed
- Missing values: Age (177), Cabin (~687), Embarked (2)

Univariate Analysis:
- Survival: ~38% survived, ~62% did not.
- Passenger Class: Majority in 3rd class.
- Gender: More males than females.
- Age: Most between 20–40 years.
- Fare: Right-skewed distribution, most fares low with some high outliers.

Bivariate Analysis:
- Pclass vs Survival: Higher class → higher survival rate.
- Sex vs Survival: Females had much higher survival rate than males.
- Age vs Survival: Children had slightly higher survival in some groups.
- Fare vs Survival: Higher fares correlated with survival.

Correlation Analysis:
- Survival correlated positively with being female and higher fare.
- Pclass had a negative correlation with survival (lower Pclass number = higher class = higher survival).

Key Insights:
1. 1st class passengers and females had the highest survival rates.
2. Higher fares indicated better survival odds.
3. Age had a moderate impact, with some advantage for younger passengers.
4. Cabin data mostly missing but could be engineered into a binary feature.

Conclusion:
Passenger class, gender, and fare are the strongest indicators of survival, with age also playing a role. These findings can guide feature selection in predictive modeling.

Code Used for EDA:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load dataset
df = pd.read_csv('train.csv')

# Info & summary
print(df.info())
print(df.describe())
print(df.describe(include=['O']))

# Missing values
```

```
missing = df.isnull().sum().sort_values(ascending=False)
missing_pct = (df.isnull().mean()*100).sort_values(ascending=False)

# Univariate plots
sns.countplot(x='Survived', data=df)
sns.countplot(x='Pclass', data=df, order=[1,2,3])
sns.countplot(x='Sex', data=df)
plt.hist(df['Age'].dropna(), bins=30)
plt.hist(df['Fare'].dropna(), bins=40)

# Bivariate plots
sns.barplot(x='Pclass', y='Survived', data=df, estimator=np.mean)
sns.barplot(x='Sex', y='Survived', data=df, estimator=np.mean)
sns.boxplot(x='Survived', y='Age', data=df)
sns.boxplot(x='Survived', y='Fare', data=df)

# Correlation
encoded_df = df.copy()
encoded_df['Sex'] = encoded_df['Sex'].map({'male':0, 'female':1})
encoded_df['Embarked'] = encoded_df['Embarked'].map({'S':0, 'C':1, 'Q':2})
numeric_cols = encoded_df.select_dtypes(include=[np.number])

sns.heatmap(numeric_cols.corr(), annot=True, fmt='.2f', cmap='coolwarm')
sns.pairplot(numeric_cols[['Survived','Pclass','Sex','Age','Fare']].dropna(),
             hue='Survived', diag_kind='hist', corner=True)
```