Exploratory Data Analysis (EDA) on Titanic Survival Dataset

Analyzing Factors Affecting Passenger Survival Using Visual and Statistical Techniques

Installing the Tools Required

```
pip install pandas matplotlib seaborn
```

```
Requirement already satisfied: pandas in d:\anaconda\lib\site-packages
(1.3.5)
Requirement already satisfied: matplotlib in d:\anaconda\lib\site-
packages (3.1.1)
Requirement already satisfied: seaborn in d:\anaconda\lib\site-
packages (0.9.0)
Requirement already satisfied: python-dateutil>=2.7.3 in d:\anaconda\
lib\site-packages (from pandas) (2.8.0)
Requirement already satisfied: pytz>=2017.3 in d:\anaconda\lib\site-
packages (from pandas) (2019.3)
Requirement already satisfied: numpy>=1.17.3 in d:\anaconda\lib\site-
packages (from pandas) (1.21.6)
Requirement already satisfied: cycler>=0.10 in d:\anaconda\lib\site-
packages (from matplotlib) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in d:\anaconda\lib\
site-packages (from matplotlib) (1.1.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!
=2.1.6,>=2.0.1 in d:\anaconda\lib\site-packages (from matplotlib)
(2.4.2)
Requirement already satisfied: scipy>=0.14.0 in d:\anaconda\lib\site-
packages (from seaborn) (1.3.1)
Requirement already satisfied: six in d:\anaconda\lib\site-packages
(from cycler>=0.10->matplotlib) (1.12.0)
Requirement already satisfied: setuptools in d:\anaconda\lib\site-
packages (from kiwisolver>=1.0.1->matplotlib) (41.4.0)
Note: you may need to restart the kernel to use updated packages.

WARNING: Ignoring invalid distribution -andas (d:\anaconda\lib\site-
packages)
WARNING: Ignoring invalid distribution -andas (d:\anaconda\lib\site-
packages)
```

Loading the Data and Basic Information & Summary Stats

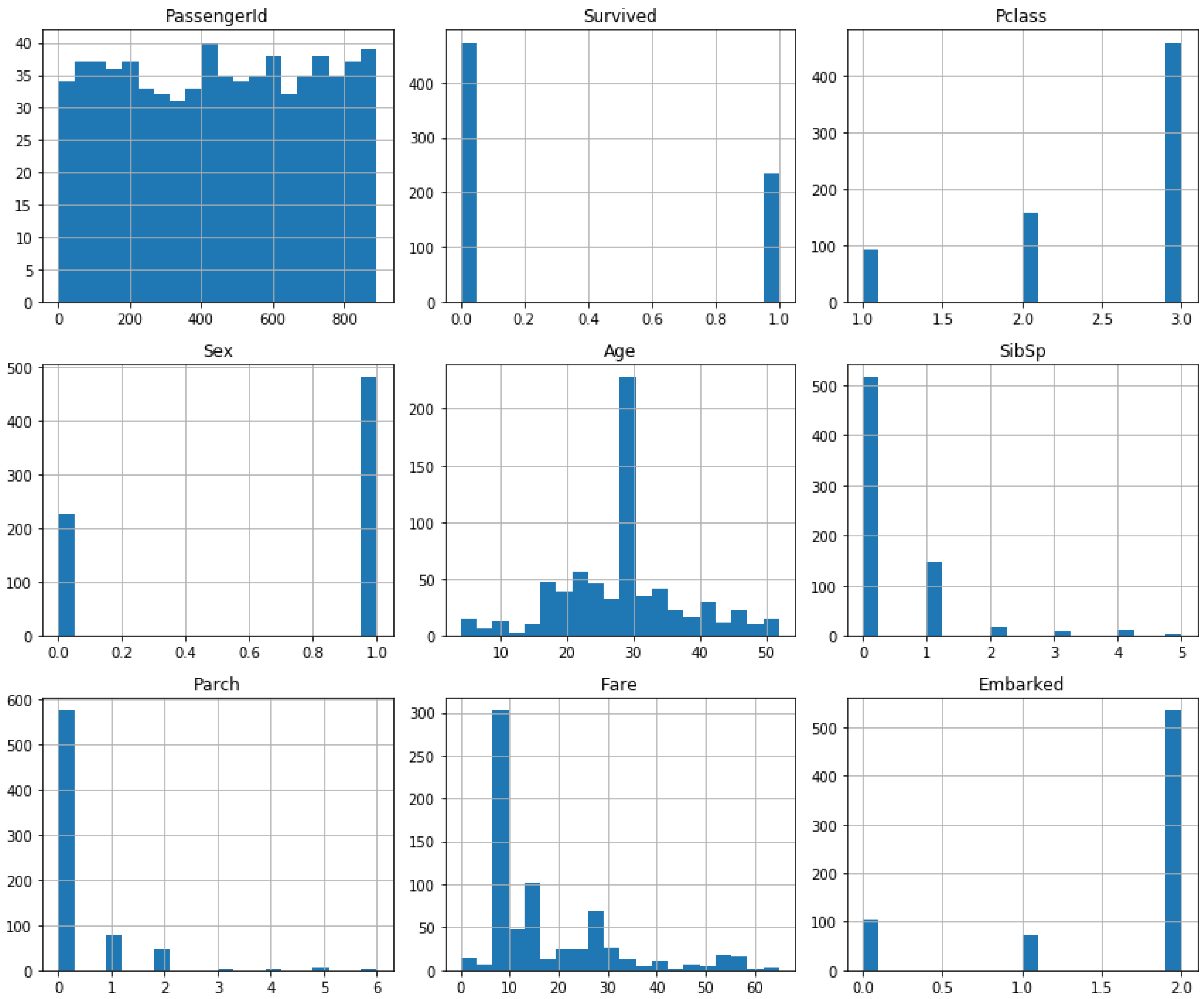
The dataset contains 708 rows and 11 columns after cleaning. There are no missing values. Key features include Pclass, Sex, Age, Fare, and Survived.

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

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

# Overview of dataset
#df.info()

# Statistical summary
#df.describe()

# Value counts for categorical features
print(df['Sex'].value_counts())
print(df['Embarked'].value_counts())
print(df['Pclass'].value_counts())
```

```
1    482
0    226
Name: Sex, dtype: int64
2    534
0    103
1     71
Name: Embarked, dtype: int64
3    459
2    157
1     92
Name: Pclass, dtype: int64
```

Univariate Analysis

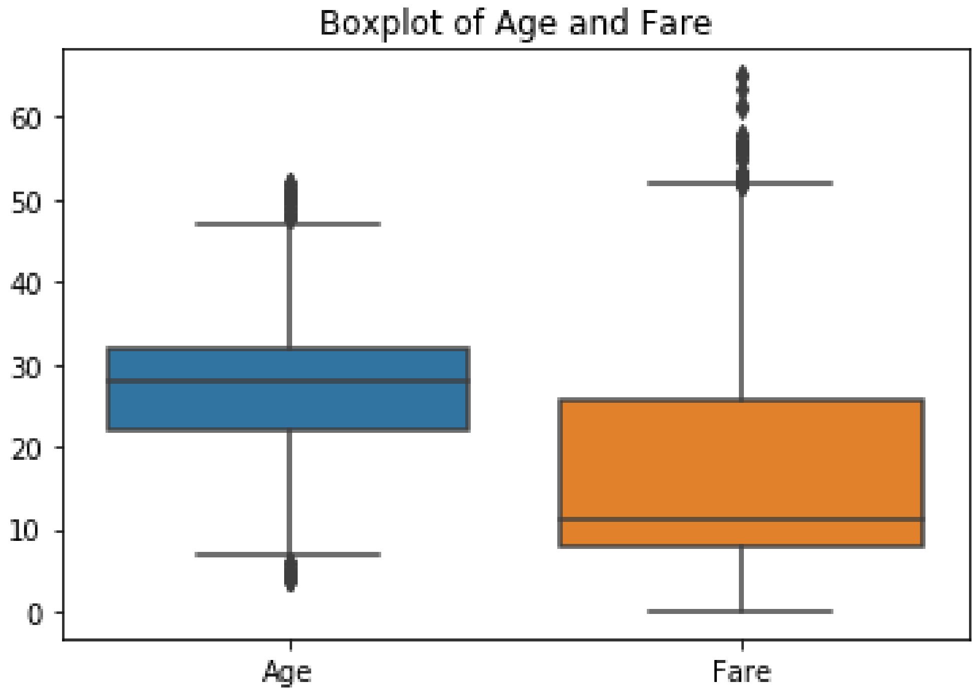
```
df.hist(figsize=(12, 10), bins=20)
plt.tight_layout()
plt.show()
```



Age is right-skewed with many passengers in their 20s and 30s.

Fare has a long tail; many paid lower fares, few paid very high fares.

```
sns.boxplot(data=df[['Age', 'Fare']])
plt.title("Boxplot of Age and Fare")
plt.show()
```

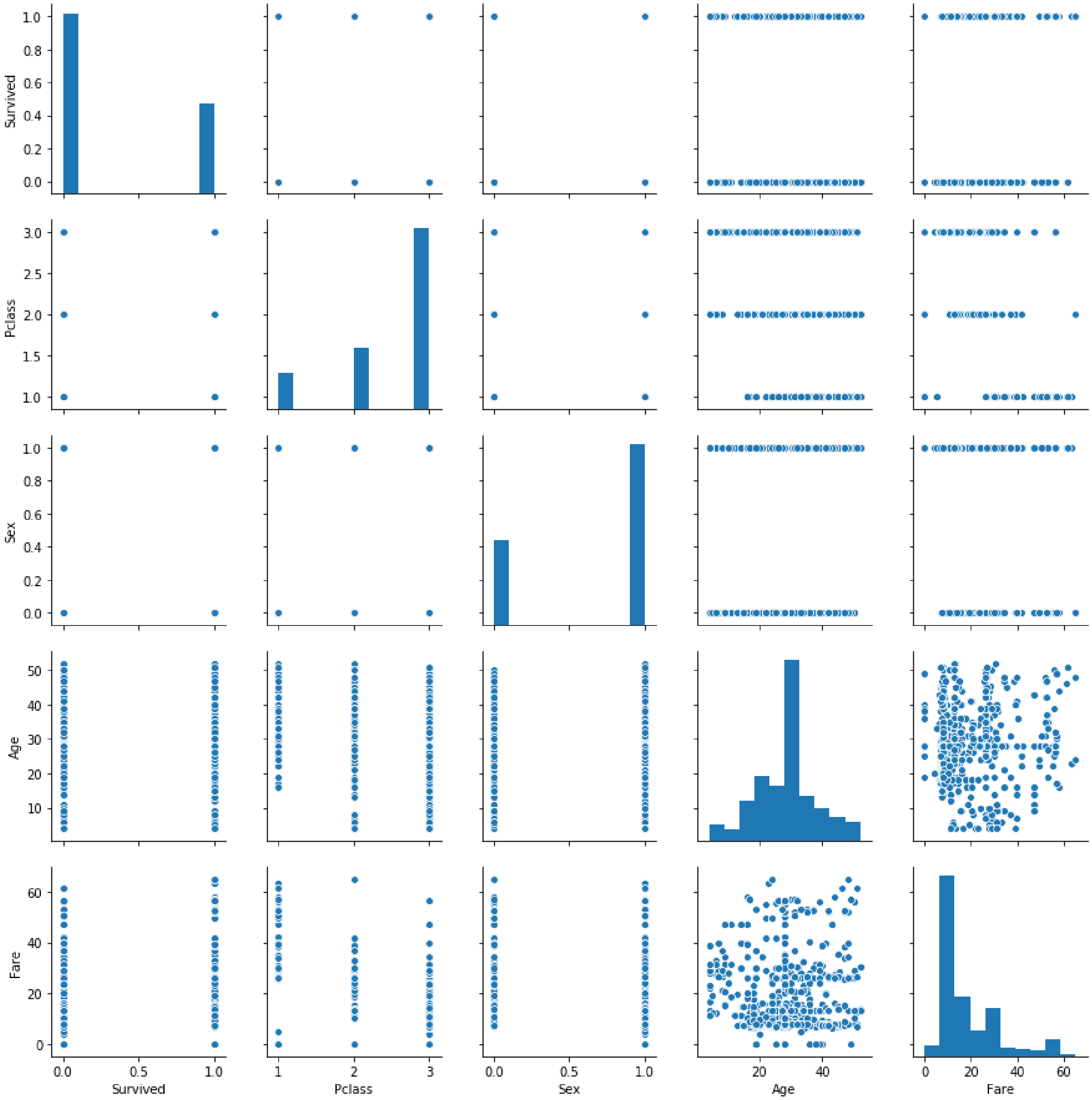


Outliers are present in both Age and Fare.

Fare shows high variability, especially among upper classes.

Bivariate Analysis

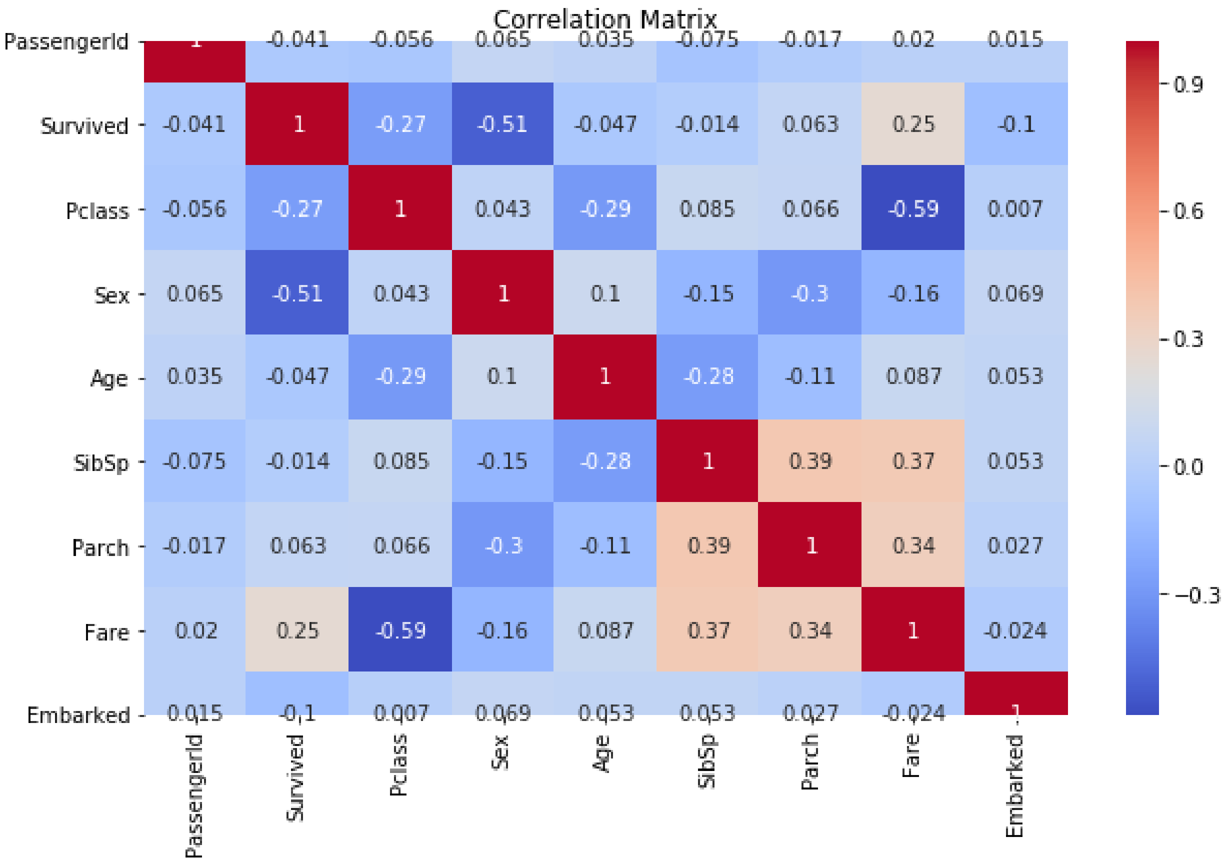
```
sns.pairplot(df[['Survived', 'Pclass', 'Sex', 'Age', 'Fare']])
plt.show()
```



Survivors tend to cluster around younger ages and lower Pclass (1st class).

Gender appears to have a strong visual impact on survival.

```
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Matrix")
plt.show()
```



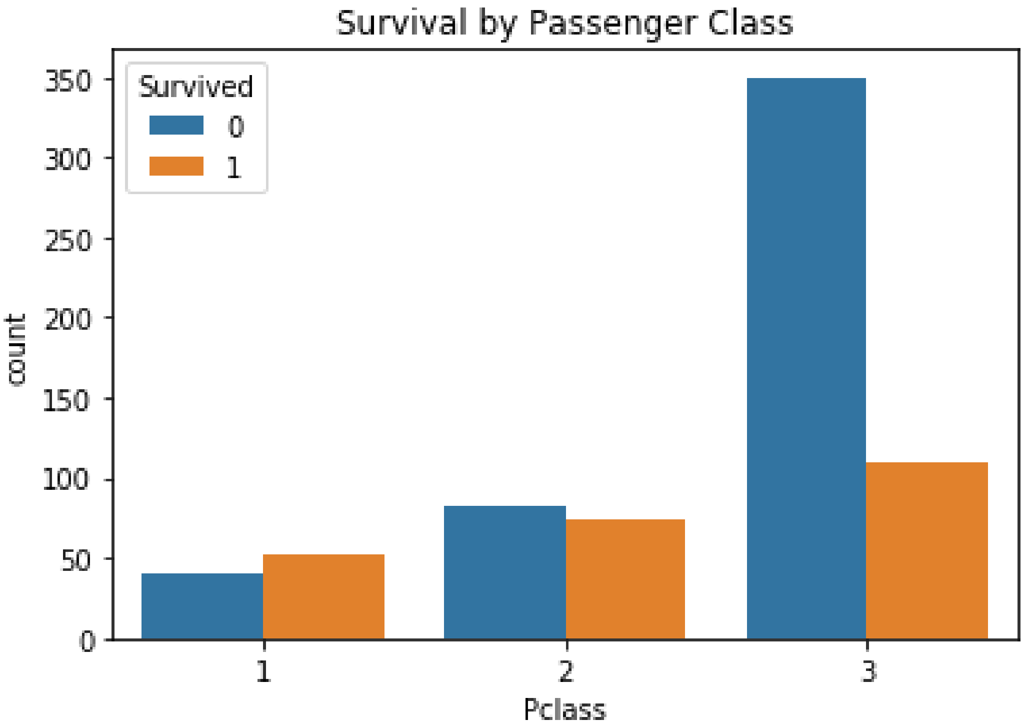
Pclass, Sex, and Fare show correlation with Survived.

Age has a weaker correlation.

Strong negative correlation between Pclass and Fare (higher class → higher fare).

Categorical vs Target Variable

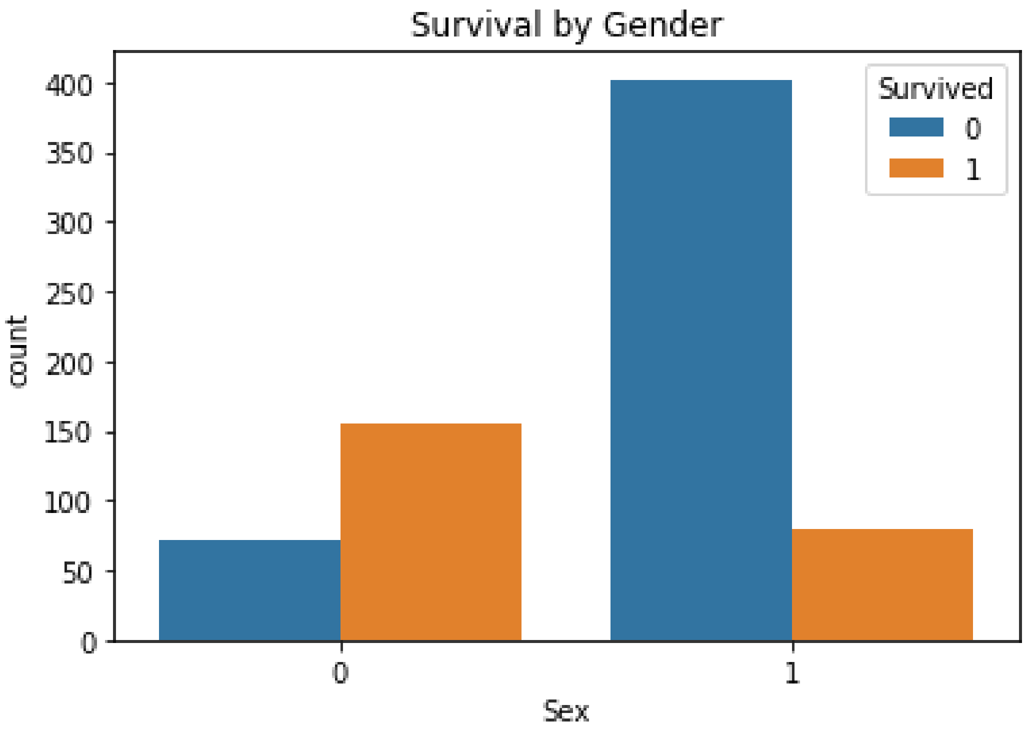
```
sns.countplot(data=df, x='Pclass', hue='Survived')
plt.title("Survival by Passenger Class")
plt.show()
```



Passengers in 1st class had the highest survival rate.

Survival rates decrease with lower class.

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



A significantly higher proportion of females (0) survived compared to males (1).

This supports the "women and children first" rescue policy.

Key Findings

Gender and Pclass were the most influential factors for survival.

Females and 1st class passengers were more likely to survive.

Fare also correlated with survival, suggesting socio-economic status mattered.

Minimal impact was observed from Age, though children had slightly higher survival.