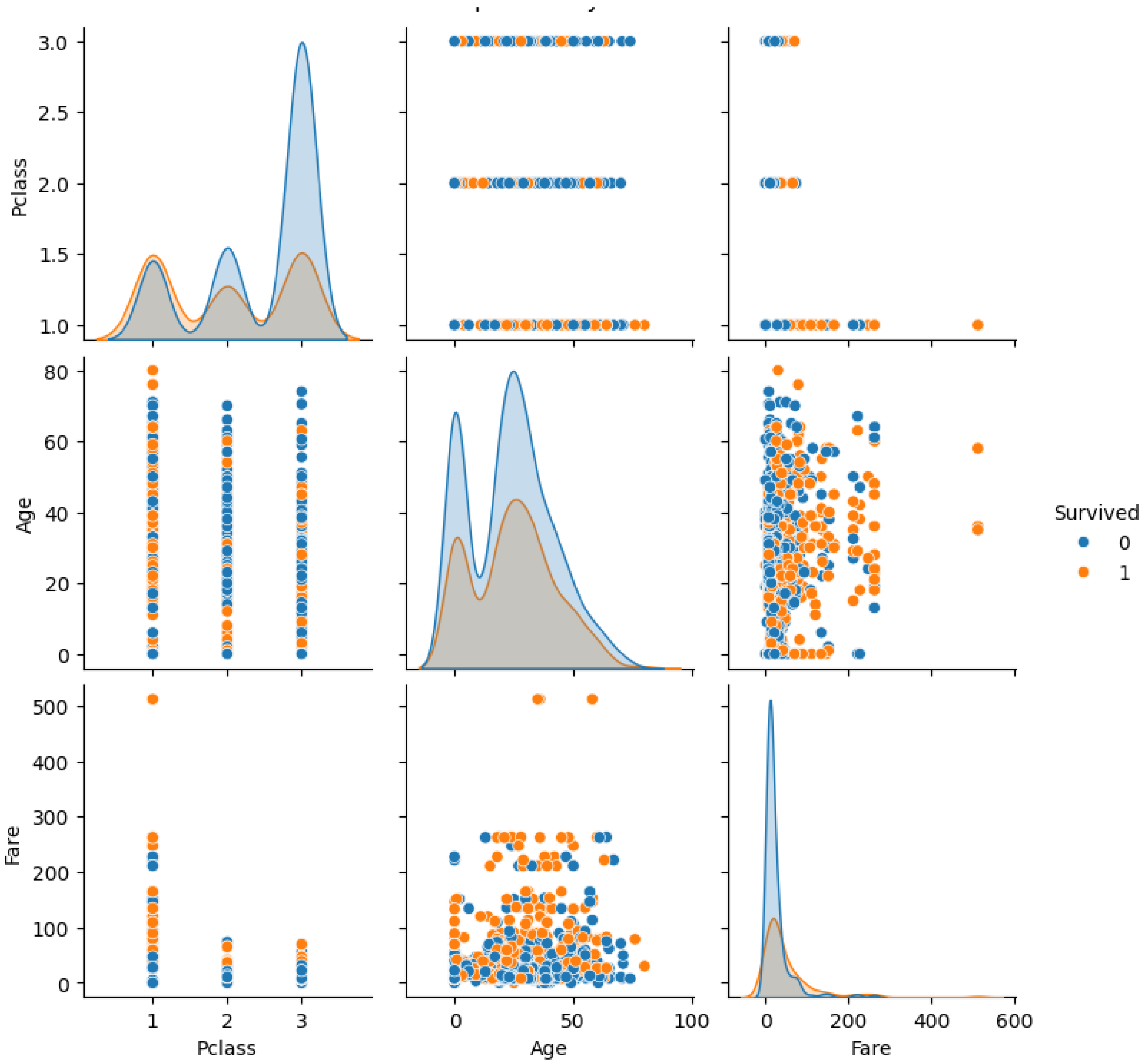
Titanic Data Analysis

```
# Step 1: Import Libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Step 2: Load the Data
df = pd.read_csv(r'C:\Users\Arun Kumar (AJ)\OneDrive\Desktop\Internship\Titanic.csv')

# Step 3: Explore the Data
print(df.info())
print(df.describe())
print(df['Pclass'].value_counts())
print(df['Survived'].value_counts())
print(df['Sex'].value_counts())
print(df['Embarked'].value_counts())

# Step 4: Visualizations
# Histograms
plt.figure(figsize=(10,5))
df['Age'].hist(bins=30)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```

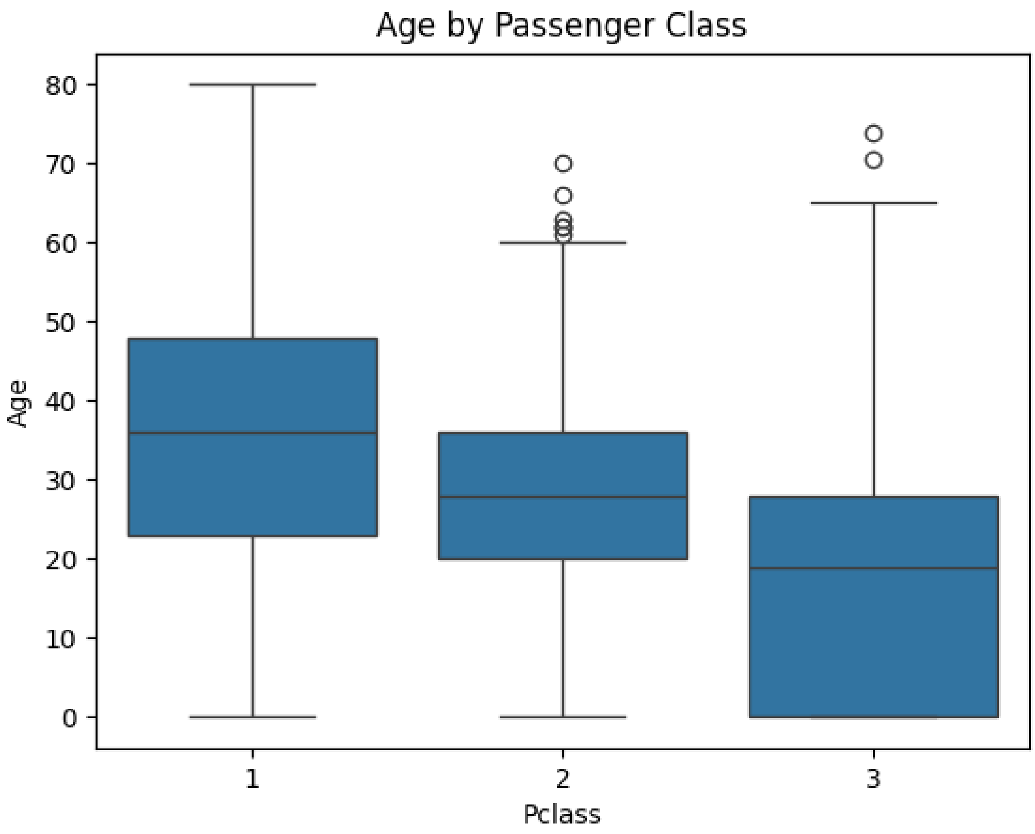
```
# Boxplots
plt.figure(figsize=(10,5))
sns.boxplot(x='Pclass', y='Age', data=df)
plt.title('Age by Passenger Class')
plt.show()

plt.figure(figsize=(10,5))
sns.boxplot(x='Survived', y='Fare', data=df)
plt.title('Fare by Survival Status')
plt.show()

# Scatterplot
plt.figure(figsize=(10,5))
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.title('Fare vs Age (colored by Survived)')
plt.show()

# Heatmap
plt.figure(figsize=(10,8))
corr = df[['Pclass', 'Age', 'SibSp', 'Parch', 'Fare', 'Survived']].corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()

# Pairplot
sns.pairplot(df[['Pclass', 'Age', 'Fare', 'Survived']], hue='Survived')
plt.suptitle('Pairplot of Key Features', y=1.02)
plt.show()
```

PairPlot



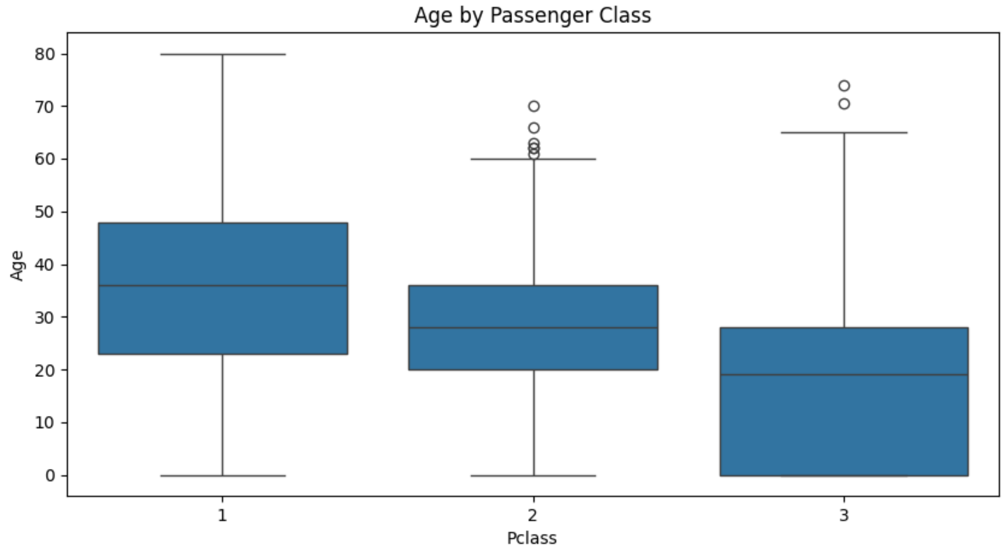
Relationships Between Pclass, Age, Fare, and Survival: The pairplot gives a multi-dimensional view of the relationships between Pclass, Age, Fare, and Survived. It shows clear clustering of Pclass and Fare, with first-class passengers generally paying higher fares. The age distribution shows that younger passengers are spread across all classes, and survival rates appear slightly higher for certain age groups and fare amounts. The color coding (orange for survived, blue for not survived) highlights survival patterns across the variables.

Barplot 1



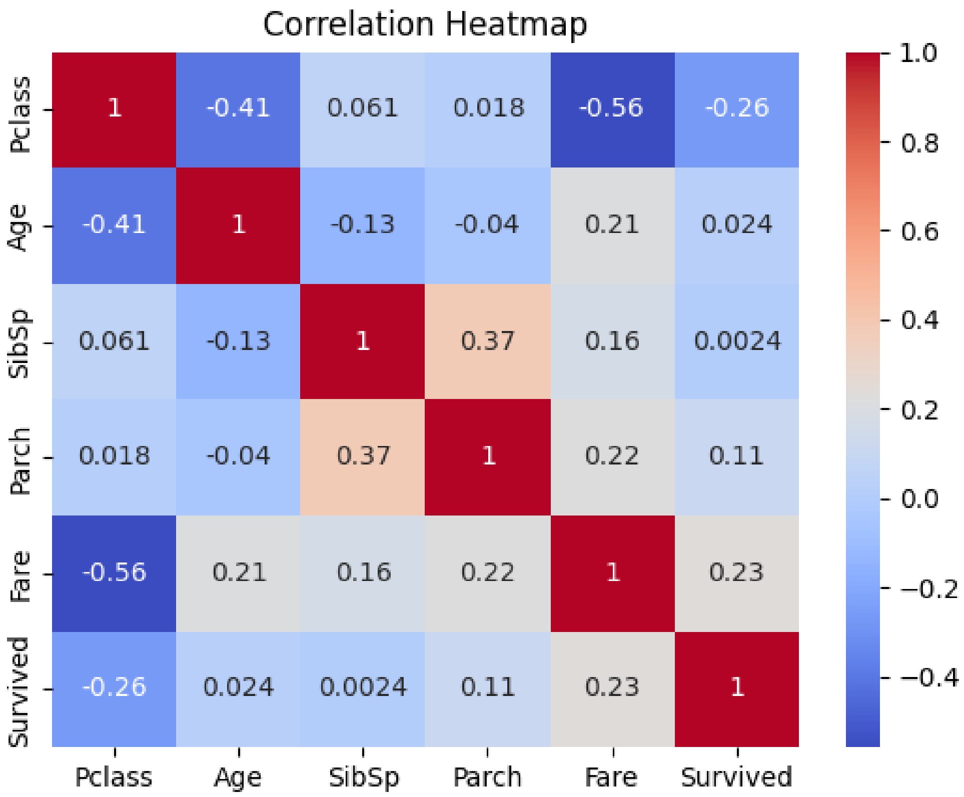
Age by Passenger Class (Wide Format): This boxplot visualizes the distribution of passengers' ages across the three passenger classes (Pclass). It shows that first-class passengers tend to be older on average compared to those in second and third classes. The median age is highest for first-class, followed by second-class and then third-class. We can observe a wider spread in age among first-class passengers, with several outliers across all classes, especially among younger ages.

Barplot 2



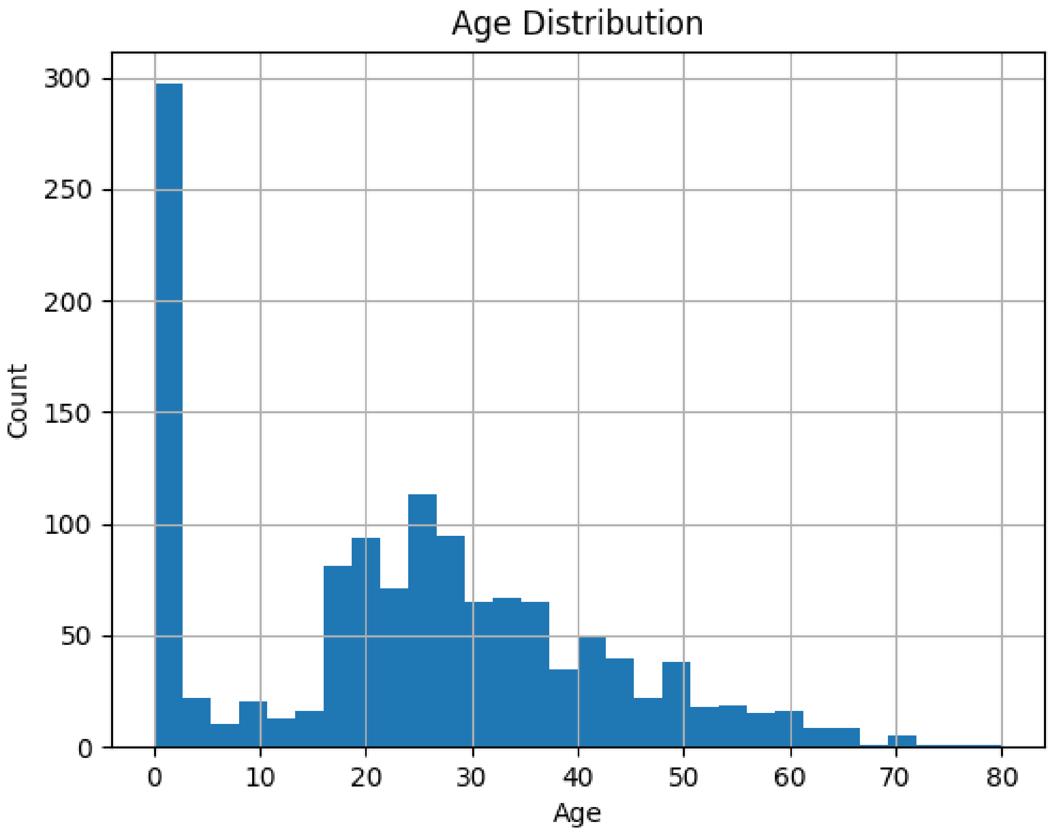
Age by Passenger Class (Compact Format): This boxplot is similar to the first but presented in a more compact format. It reinforces the same observations: first-class passengers are generally older, and third-class passengers tend to be younger. The data again shows multiple outliers, particularly at younger ages across all classes, suggesting a significant presence of children and infants among passengers.

Heatmap



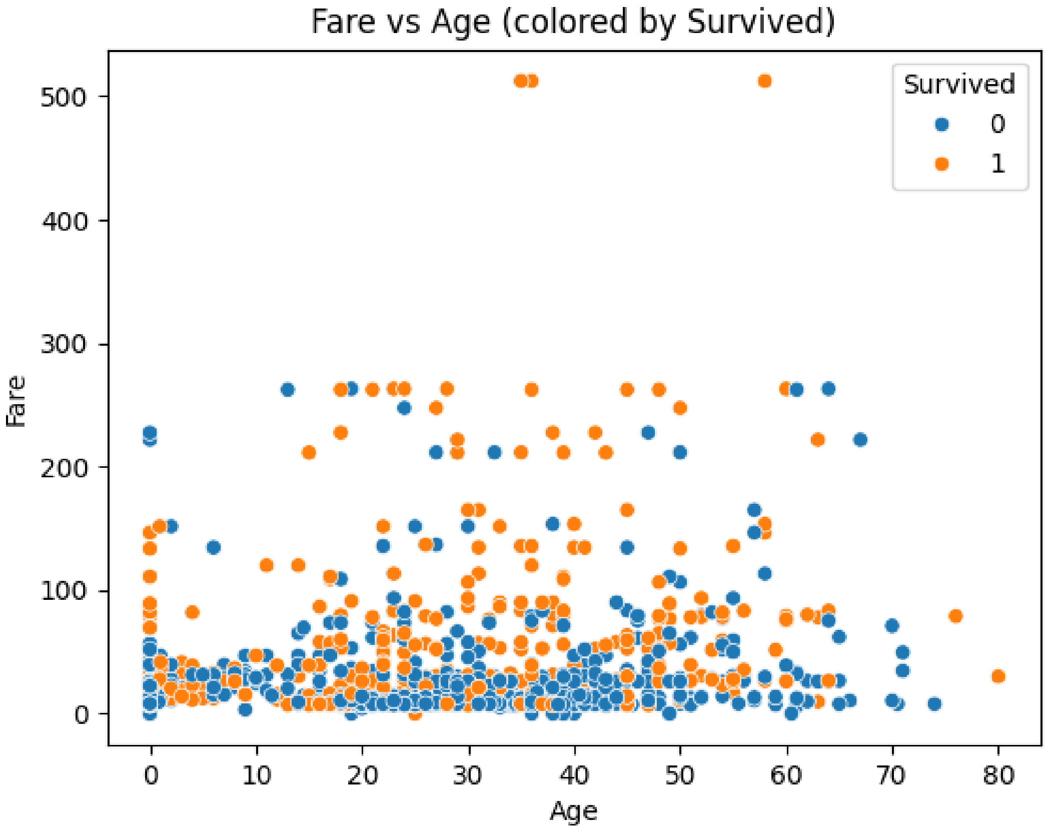
Correlation Matrix: The heatmap shows the correlation coefficients between different features such as Pclass, Age, SibSp (siblings/spouses aboard), Parch (parents/children aboard), Fare, and Survived. Notable relationships include a strong negative correlation between Pclass and Fare, meaning passengers in higher classes (lower Pclass numbers) paid more fares. Age shows a weak correlation with other variables, and survival is moderately positively correlated with Fare and slightly negatively correlated with Pclass.

Histogram



Age Distribution: This histogram shows the overall age distribution of the passengers. It is right-skewed, indicating that there were more younger passengers on board. There's a notable spike at very young ages (0–5 years), suggesting many infants and toddlers were present. After this initial spike, the number of passengers gradually decreases with age.

Scatterplot



Fare vs Age (Colored by Survival): This scatterplot displays the relationship between Fare and Age, colored by survival status. There is no strong trend between age and fare directly, but we notice that higher fares generally have a greater proportion of survivors (more orange points at higher fares). Lower fare passengers, who dominate the data, have a mix of survival and non-survival, suggesting socioeconomic status (indicated by fare) influenced survival chances.

Observations for Each Visual

Visual	Observation
Boxplot	Older passengers mostly traveled first class. Younger passengers in third class.
Heatmap	Strong negative correlation between Pclass and Fare. Positive correlation between Fare and Survival.
Histogram	Peak ages are around 20–40 years, significant number of infants (~0 years old).
Pairplot	High fare + first class = higher survival rate.
Scatterplot	Survivors generally paid higher fares, regardless of age.

Summary of Findings

- Class and Fare are major factors influencing survival.
  - First-class passengers had the highest survival rates.
  - Higher ticket prices are associated with better survival chances.
- Age is not a strong determinant of survival, but infants and very young children were a significant group.
- Family members traveling together slightly increased survival odds.
- Passengers from third class (cheapest fares) had the lowest chance of survival.