Introduction

🚢 Titanic Dataset Analysis

🎯 Objective

Perform Exploratory Data Analysis (EDA) to understand factors influencing passenger survival on the Titanic.

📂 Dataset Source

This dataset is from Kaggle - Titanic: Machine Learning from Disaster.

Step 1: Setup Environment:

In [1]:
```
pip install pandas matplotlib seaborn
```

```
Requirement already satisfied: pandas in c:\users\lohit\anaconda4\lib\site-packag
es (2.1.4)Note: you may need to restart the kernel to use updated packages.

Requirement already satisfied: matplotlib in c:\users\lohit\anaconda4\lib\site-pa
ckages (3.8.0)
Requirement already satisfied: seaborn in c:\users\lohit\anaconda4\lib\site-packa
ges (0.12.2)
Requirement already satisfied: numpy<2,>=1.23.2 in c:\users\lohit\anaconda4\lib\s
ite-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\lohit\anaconda4
\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\lohit\anaconda4\lib\site-
packages (from pandas) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\lohit\anaconda4\lib\sit
e-packages (from pandas) (2023.3)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\lohit\anaconda4\lib\s
ite-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\lohit\anaconda4\lib\site-
packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\lohit\anaconda4\lib
\site-packages (from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\lohit\anaconda4\lib
\site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\lohit\anaconda4\lib\si
te-packages (from matplotlib) (23.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\lohit\anaconda4\lib\site
-packages (from matplotlib) (10.2.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\lohit\anaconda4\lib\s
ite-packages (from matplotlib) (3.0.9)
Requirement already satisfied: six>=1.5 in c:\users\lohit\anaconda4\lib\site-pack
ages (from python-dateutil>=2.8.2->pandas) (1.16.0)
```

Step 2: Data Loading & Initial Exploration

In [2]:
```python
import pandas as pd

df = pd.read_csv("train.csv")  # Use correct path
df.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0

Observation:

The dataset contains 891 entries with 12 features including 'Survived', 'Pclass', 'Sex', 'Age', etc.

Step 3: Data Summary

In [3]:
```python
df.info()          # Data types & non-null values
df.describe()      # Summary statistics
df.isnull().sum() # Missing Values
df.nunique()       # Unique values per column
df['Sex'].value_counts()  # Example of categorical column check
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Out[3]:
```
Sex
male      577
female    314
Name: count, dtype: int64
```

Observation:

'Age' has missing values.

'Cabin' has many missing values, indicating it might not be useful for analysis.

Step 4: Visual Explorating using seaborn/Matplotlib

In [4]:
```python
import seaborn as sns
import matplotlib.pyplot as plt
```

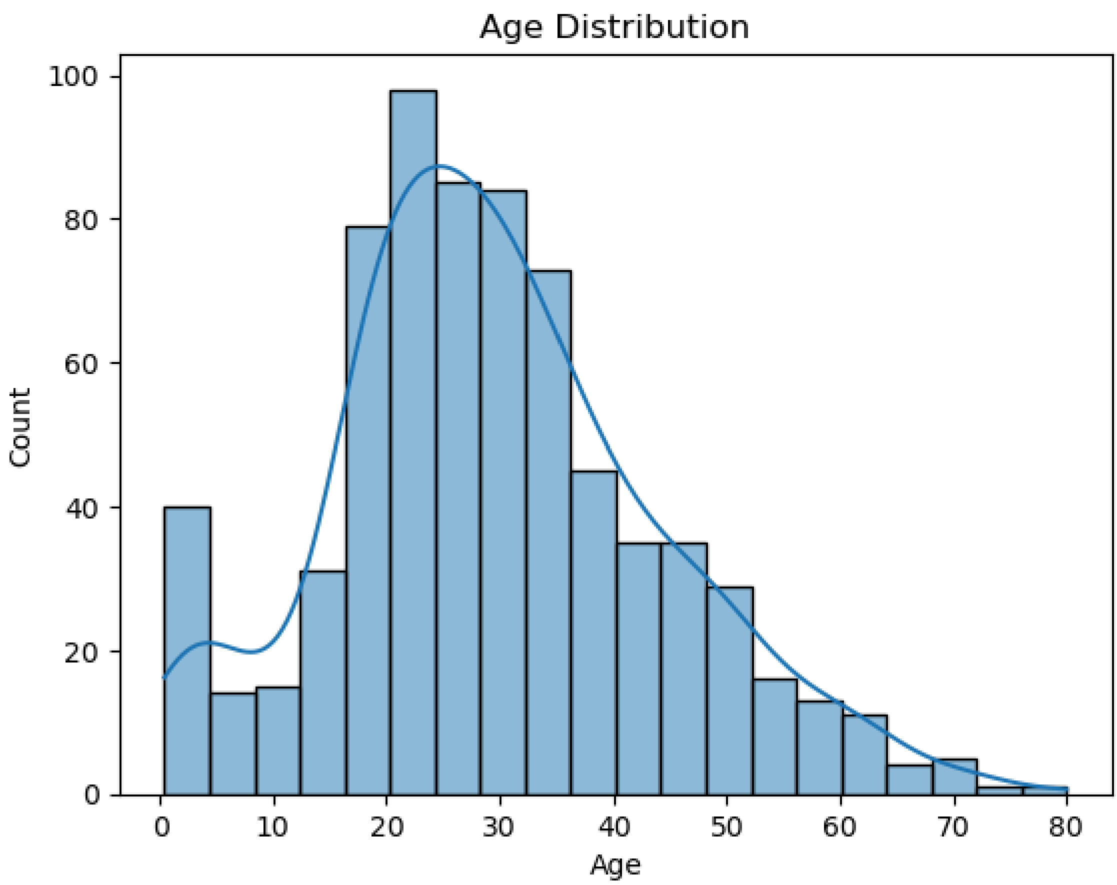
Univariate Analysis

Histogram for age:

In [5]:
```python
sns.histplot(df['Age'], kde=True)
plt.title("Age Distribution")
```

```
C:\Users\lohit\anaconda4\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

Out[5]:
```
Text(0.5, 1.0, 'Age Distribution')
```

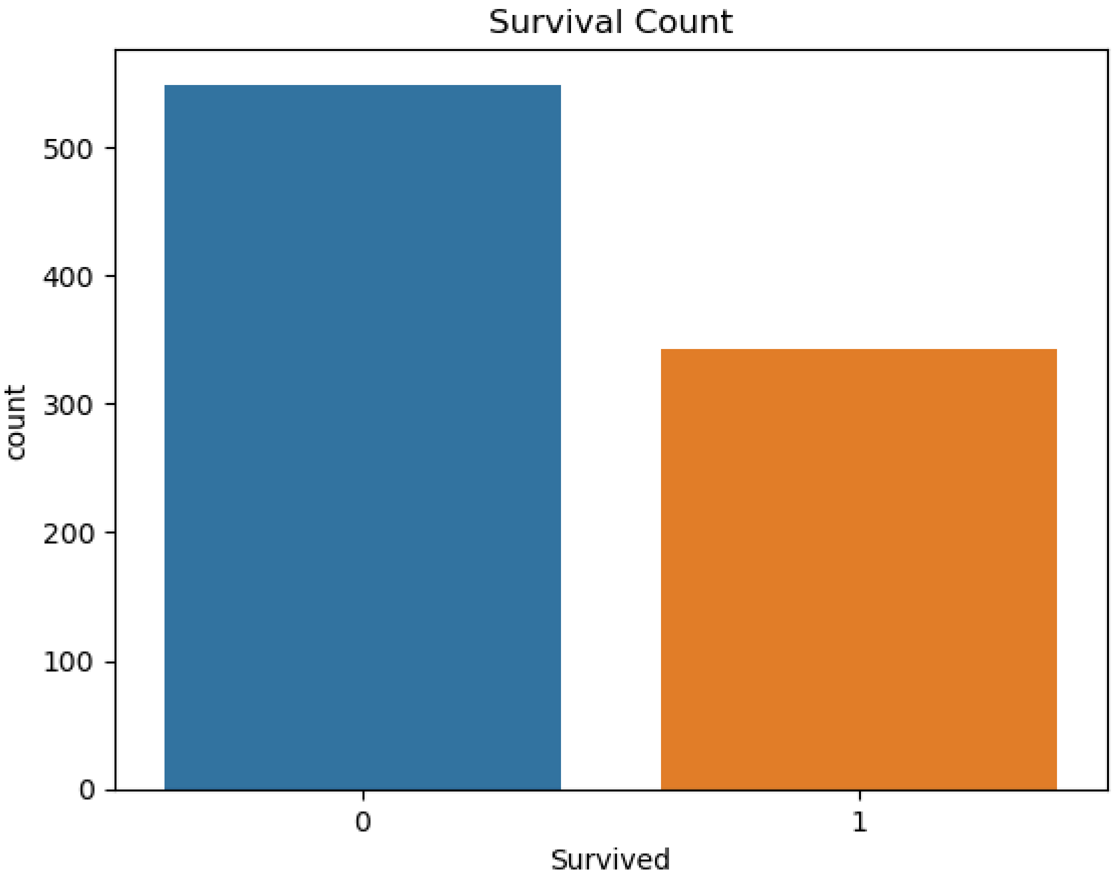


Observation:

- Most passengers were between 20 and 40 years old.
- The age distribution is slightly right-skewed.

Countplot for Survived:

In [15]:
```
sns.countplot(x='Survived', data=df)
plt.title("Survival Count")
plt.show()
```

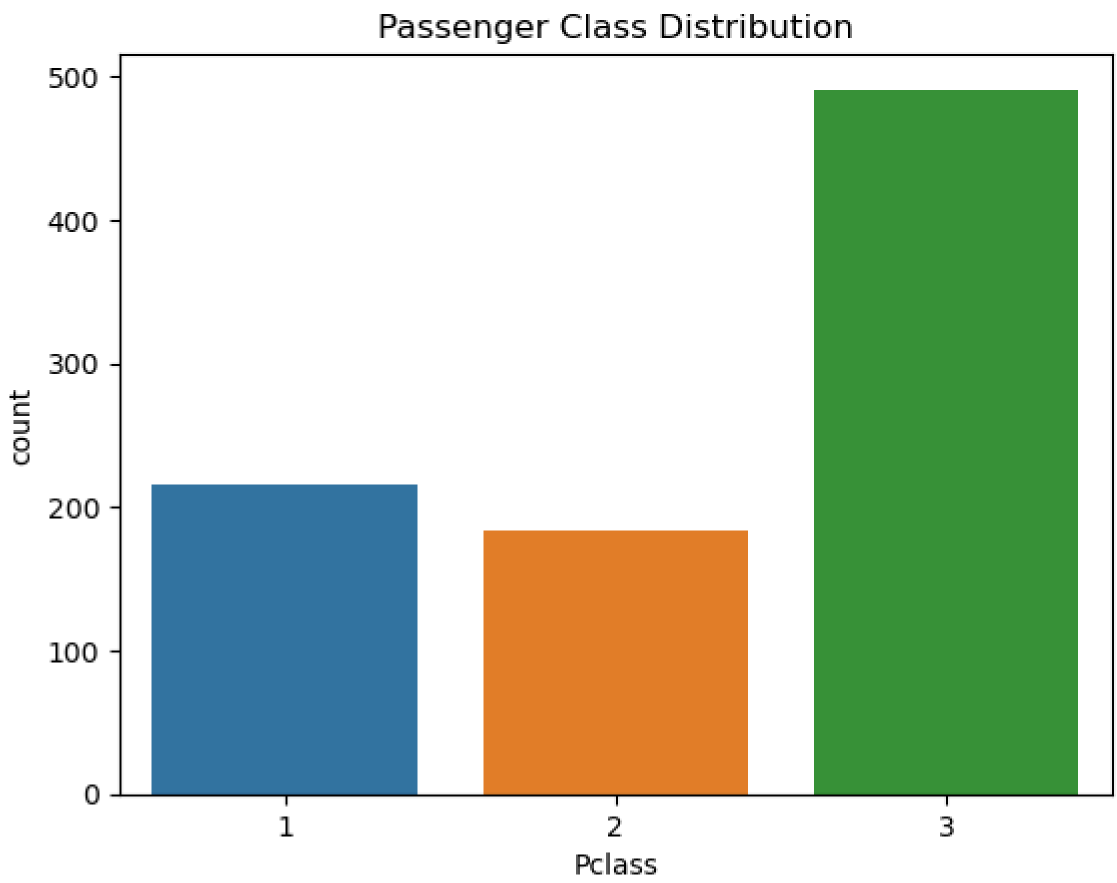


Observation:

More passengers did not survive (0) compared to those who did (1).

Passenger class Distribution

In [16]:
```
sns.countplot(x='Pclass', data=df)
plt.title('Passenger Class Distribution')
plt.show()
```



Observation:

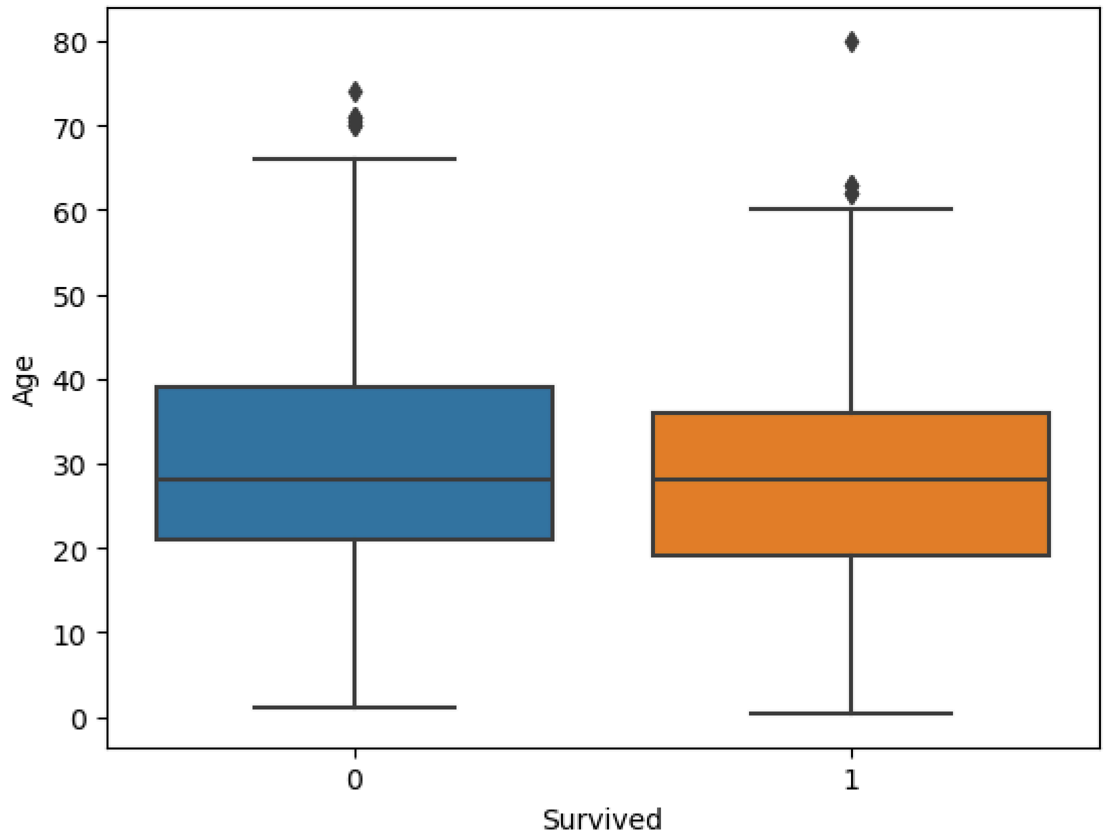
Most passengers were in the 3rd class.

Bivariate Analysis

Survival by sex

In [20]:
```
sns.boxplot(x='Survived', y='Age', data=df)
```

Out[20]:
```
<Axes: xlabel='Survived', ylabel='Age'>
```



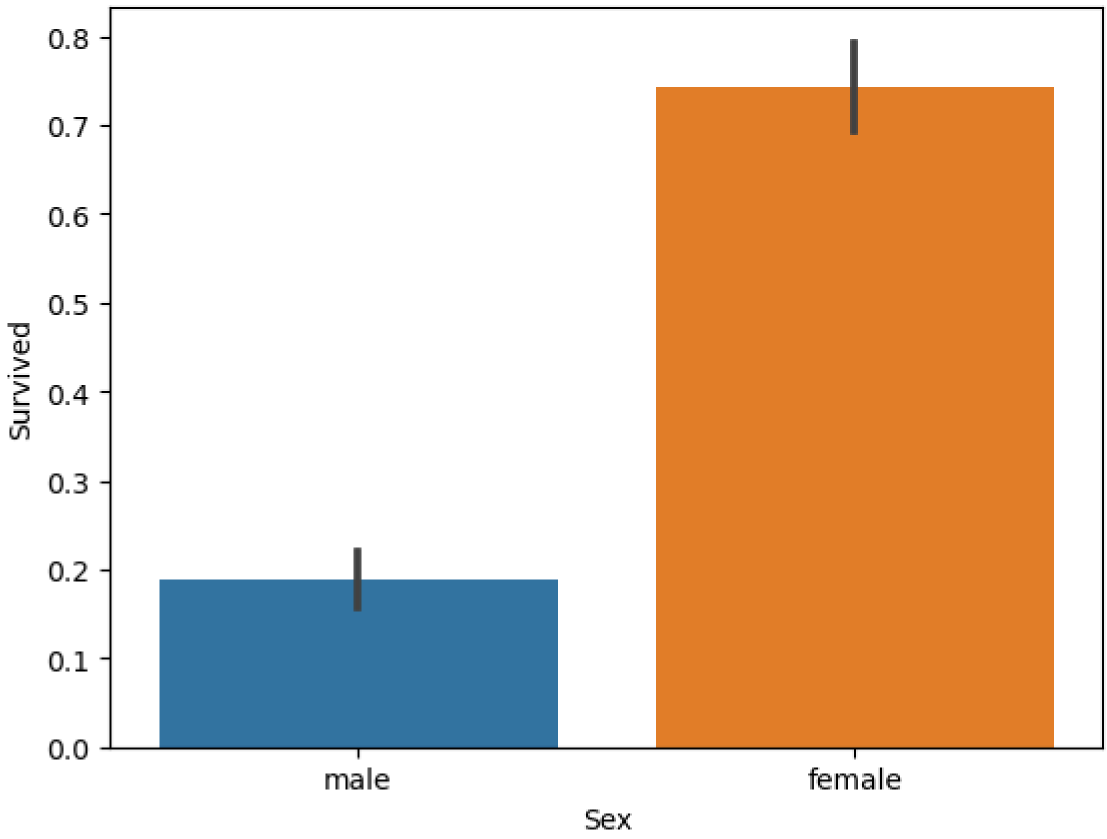
Observation:

Younger passengers with higher fares had better survival chances.

Barplot: Sex vs Survival

In [8]:
```
sns.barplot(x='Sex', y='Survived', data=df)
```

Out[8]: <Axes: xlabel='Sex', ylabel='Survived'>



Observation:

- Females had a significantly higher survival rate than males.

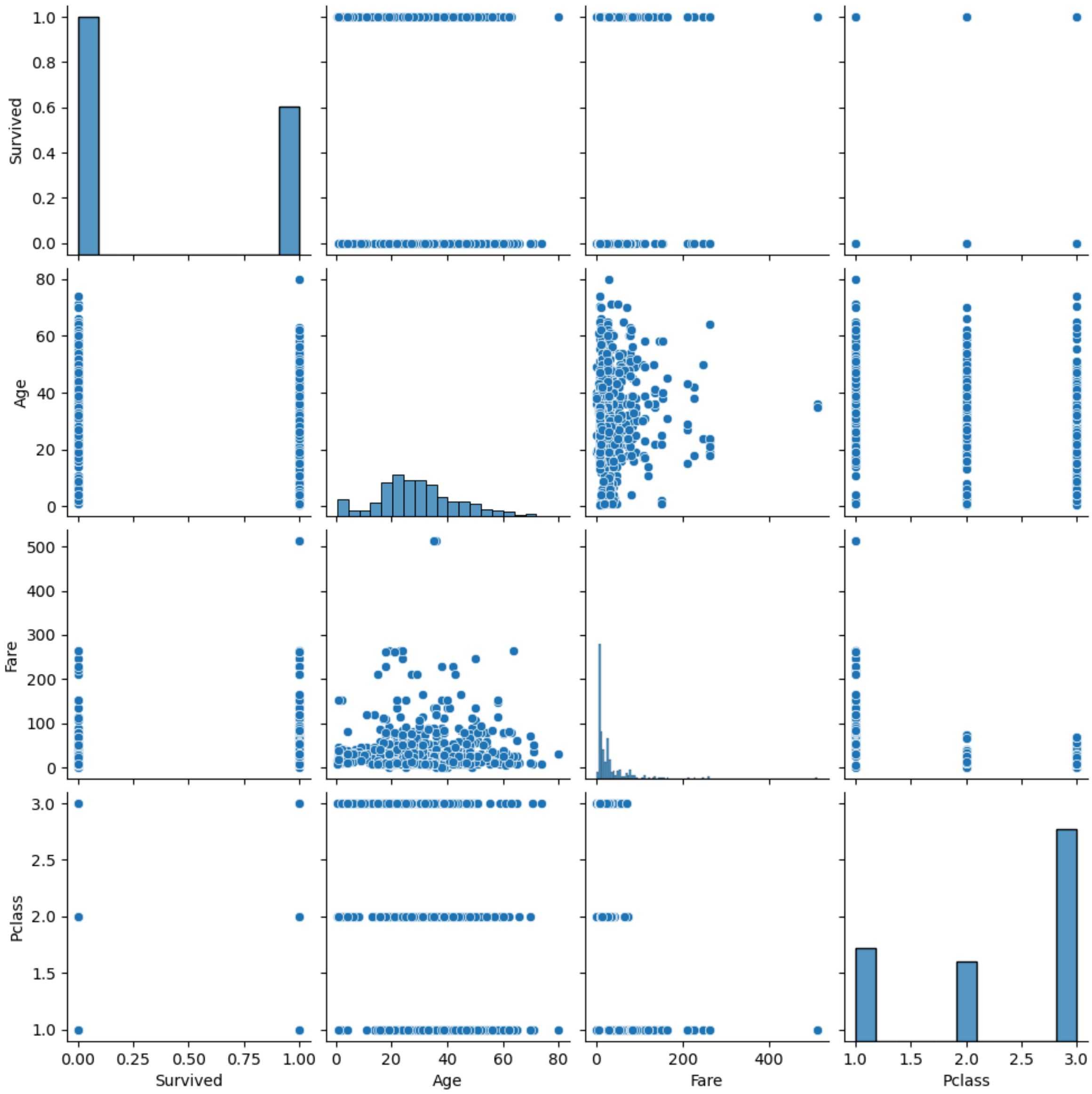
Multivariate Analysis

Pairplot

In [9]:
```
sns.pairplot(df[['Survived', 'Age', 'Fare', 'Pclass']])
```

```
C:\Users\lohit\anaconda4\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\lohit\anaconda4\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\lohit\anaconda4\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\lohit\anaconda4\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

Out[9]:
```
<seaborn.axisgrid.PairGrid at 0x1c69579f710>
```



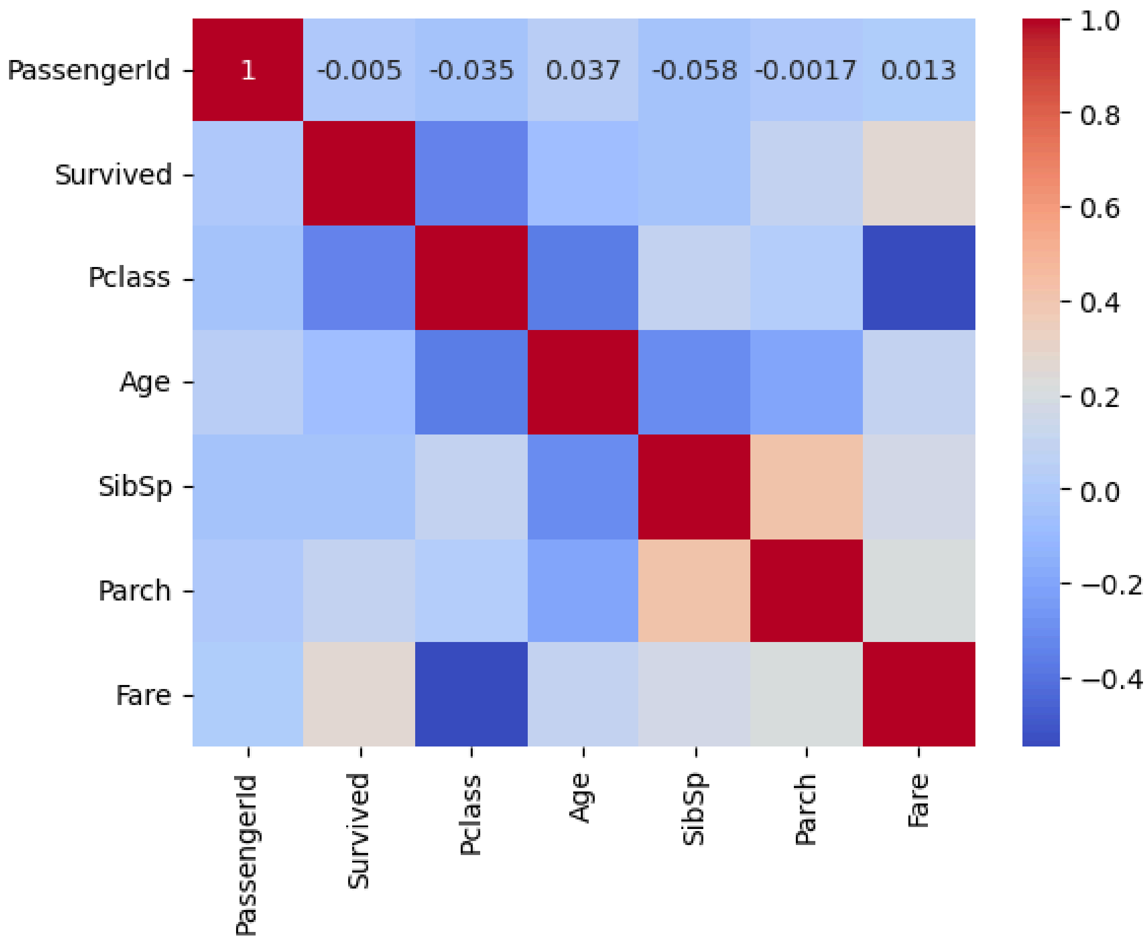
Observation:

- Higher fare and 1st class are linked with higher survival.
- Younger passengers also had better survival chances.
- Fare and Pclass show strong separation.

Heatmap (Correlation):

In [11]:
```
df_numeric = df.select_dtypes(include=['number'])
sns.heatmap(df_numeric.corr(), annot=True, cmap='coolwarm')
```

Out[11]: `<Axes: >`



Observation:

- Fare has a positive correlation with Survival.
- Pclass has a negative correlation with Survival.
- SibSp and Parch show weak correlation with Survival.

📊 Summary of Insights:

- Female passengers had a significantly higher survival rate than males.
- Passengers in 1st class had better survival chances than those in 2nd or 3rd class.
- Fare and survival rate are positively correlated - people who paid more had better chances of survival.
- Young children had relatively higher survival rates.
- The dataset contains missing values in 'Age' and 'Cabin' columns.

🎯 These insights can help understand what factors influenced survival in the Titanic disaster.

In [ ]: