In [41]:
```
import pandas as pd                         #Importing Libraries
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
```

In [42]:
```
df = pd.read_csv('train.csv')          #Load the Dataset
```

In [43]:
```
df.head()          #To Check first few rows
```

Out[43]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [44]:
```
df.info()                #This Checks the data types and non-null counts
```

```
<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
```

In [45]:
```
df.describe()           #Get summary statistics
```

Out[45]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [46]:
```
df['Survived'].value_counts()          #It Checks the distribution of target
```

Out[46]:
```
Survived
0    549
1    342
Name: count, dtype: int64
```

In [47]:
```
df.isnull()
```

Out[47]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False	False	True	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	True	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	True	False

891 rows × 12 columns

In [48]:
```
df.isnull().sum()          #Checking for missing values
```

Out[48]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [49]:
```
df['Age']=df['Age'].fillna(df['Age'].median())                    #Fill the missing values
df['Embarked']=df['Embarked'].fillna(df['Embarked'].mode()[0])
```

In [50]:
```
df
```

Out[50]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	28.0	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [51]:

```
df.isnull().sum()
```

Out[51]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         0
dtype: int64
```

In [52]:
```
df.drop('Cabin',axis=1,inplace= True)       #Drop Column
df
```

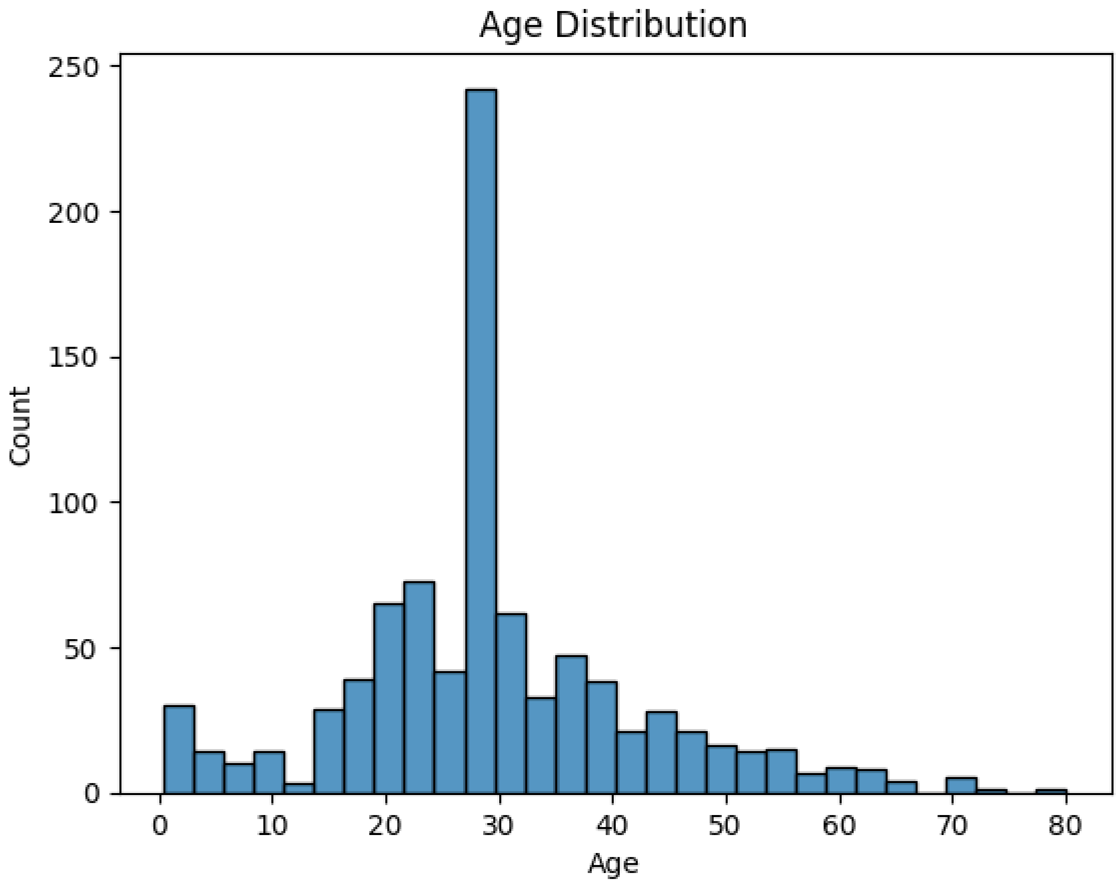
Out[52]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	28.0	1	2	W./C. 6607	23.4500	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Q

891 rows × 11 columns

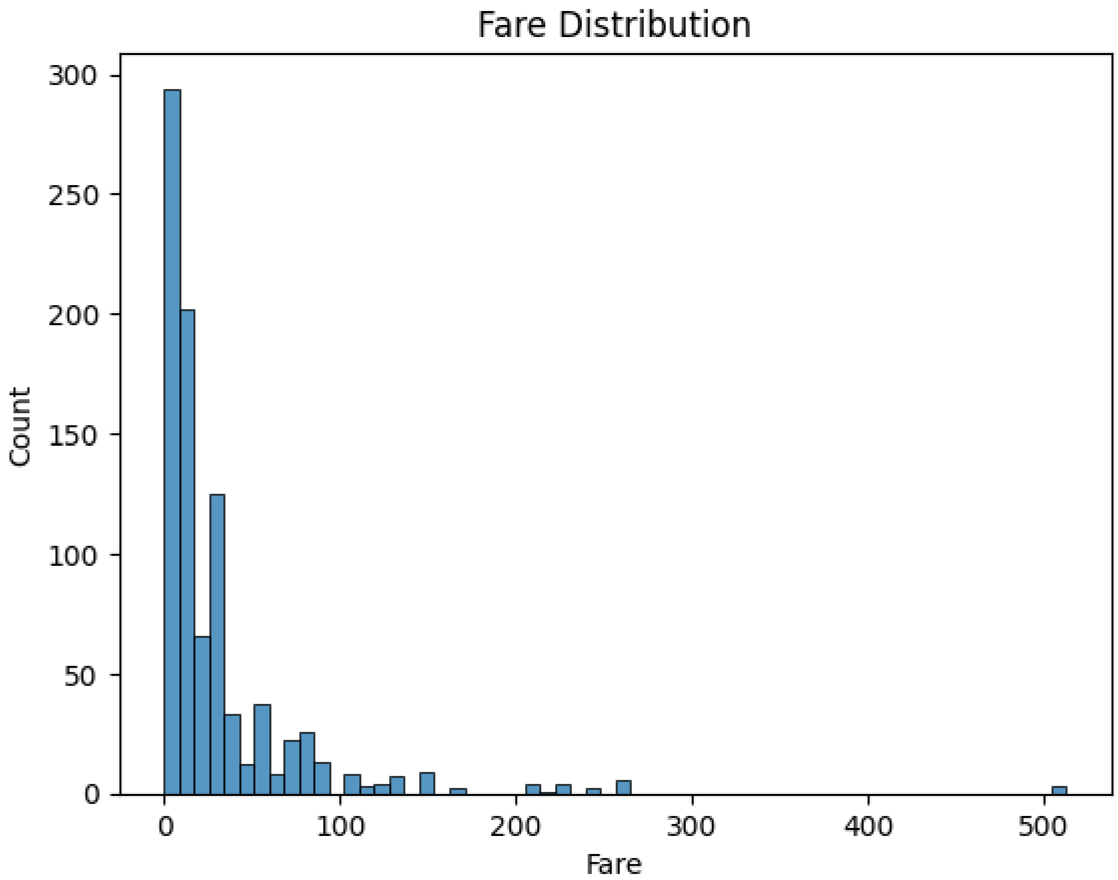
In [53]:

```
sns.histplot(x='Age',data=df)
plt.title('Age Distribution')
plt.show()
```



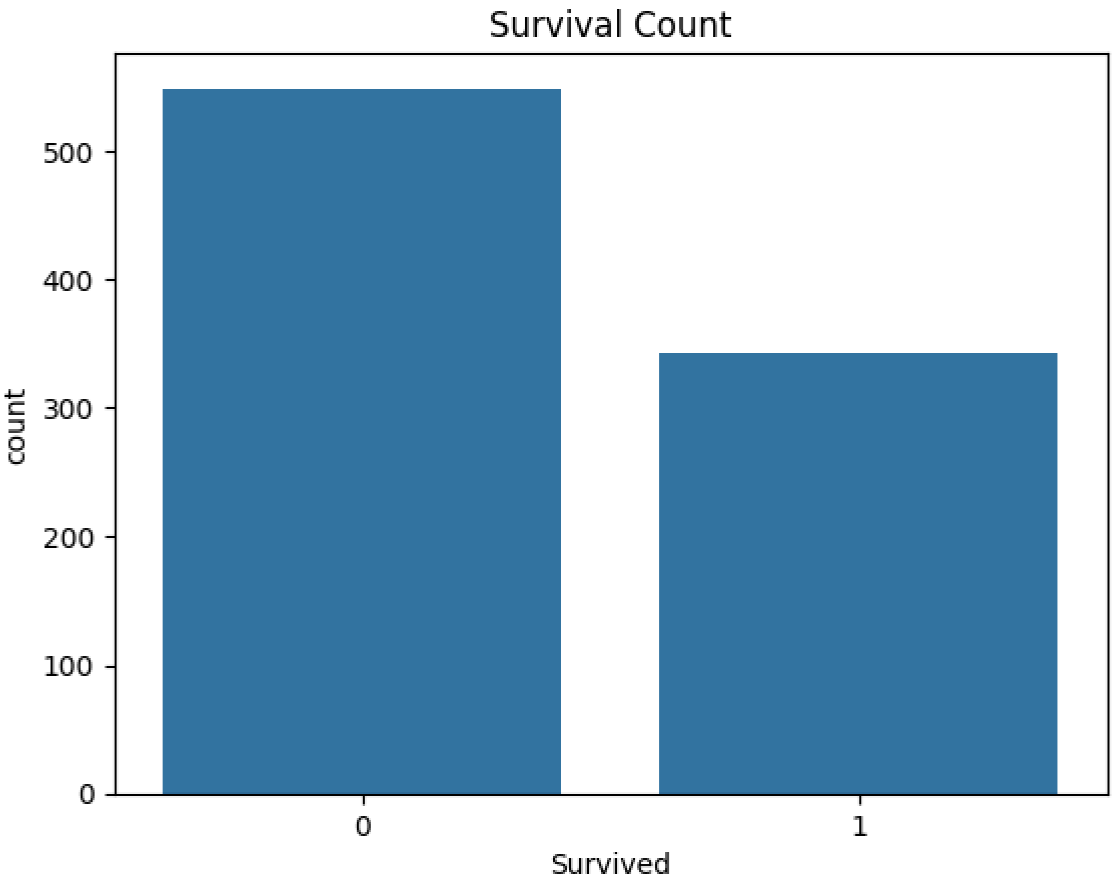
Observation: Most passengers are between '20–40' years old, indicating a young passenger distribution.

In [54]:
```
sns.histplot(x='Fare',data=df)
plt.title('Fare Distribution')
plt.show()
```



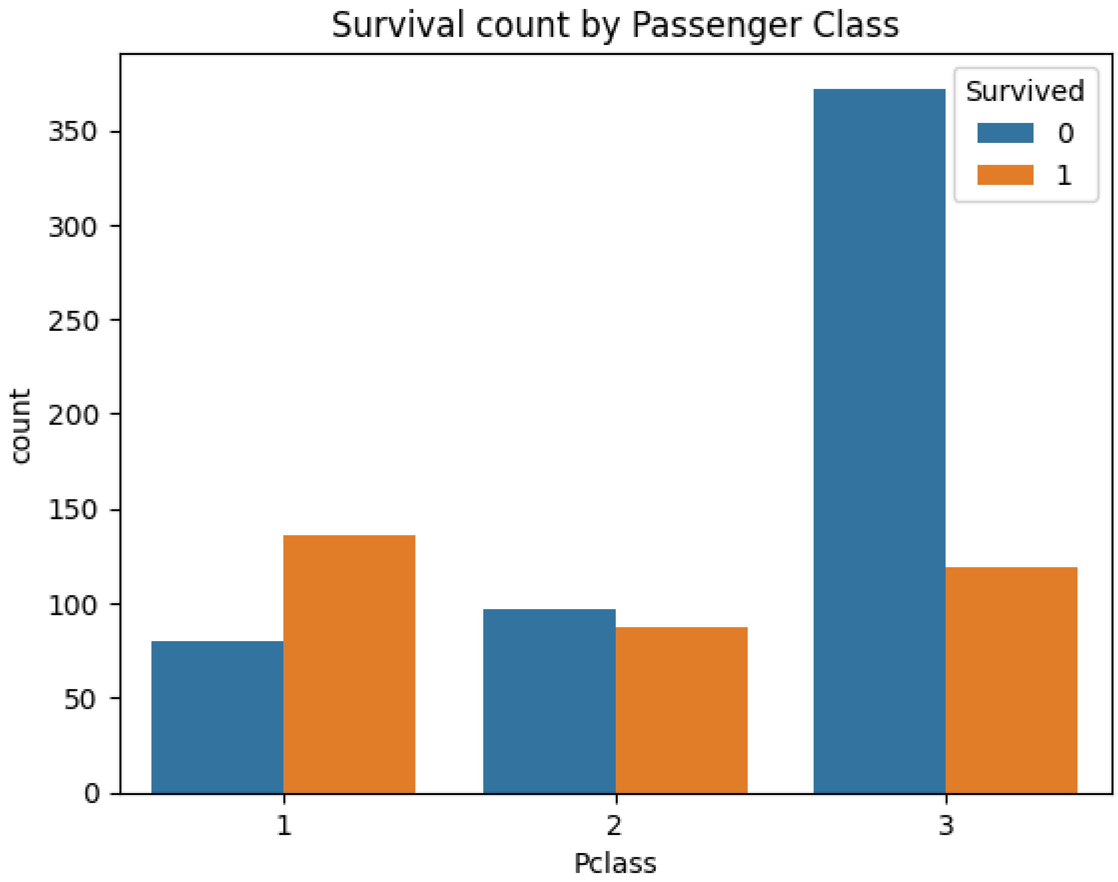
Observation: Fare distribution is right-skewed; most fares are low, with a few high-fare outliers.

In [55]:
```
sns.countplot(x='Survived',data=df)
plt.title('Survival Count')
plt.show()
```



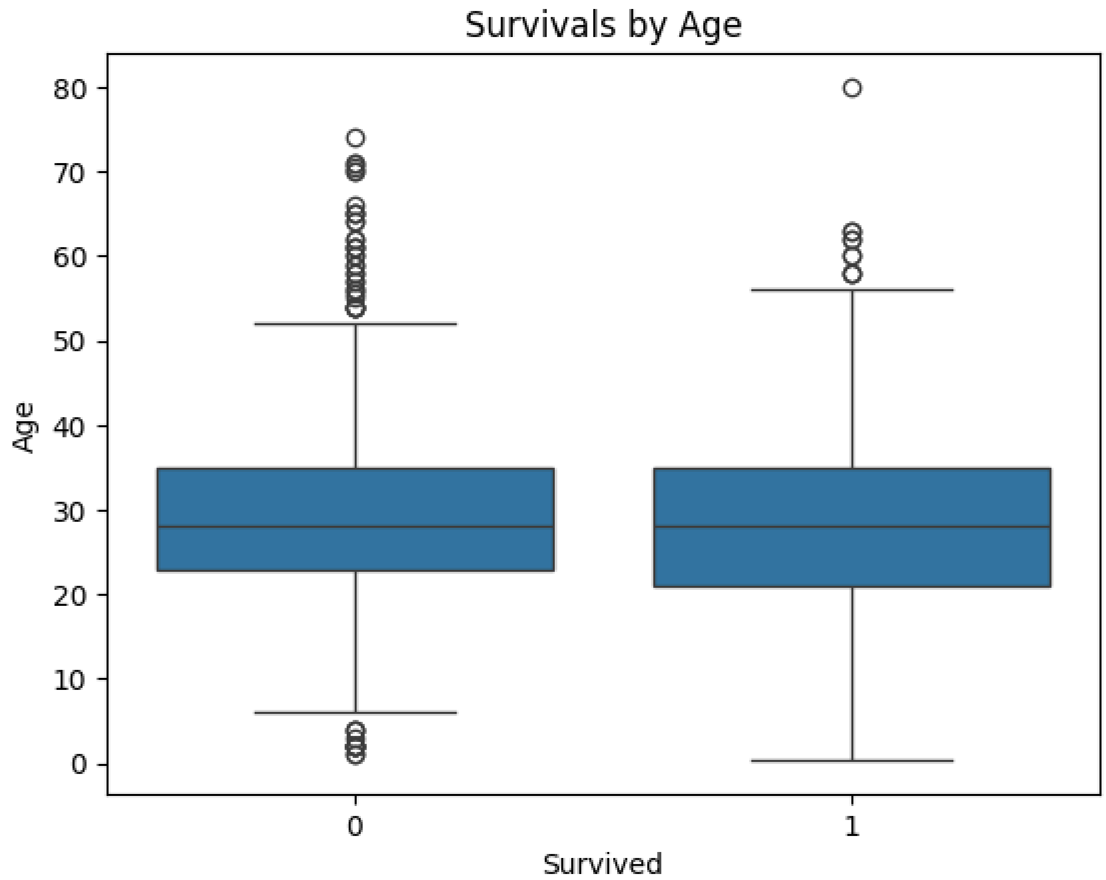
Observation: Around 550 passengers did not survive, while about 340 survived, indicating a survival rate of ~38%.

In [56]:
```
sns.countplot(x='Pclass',hue='Survived',data=df)
plt.title('Survival count by Passenger Class')
plt.show()
```



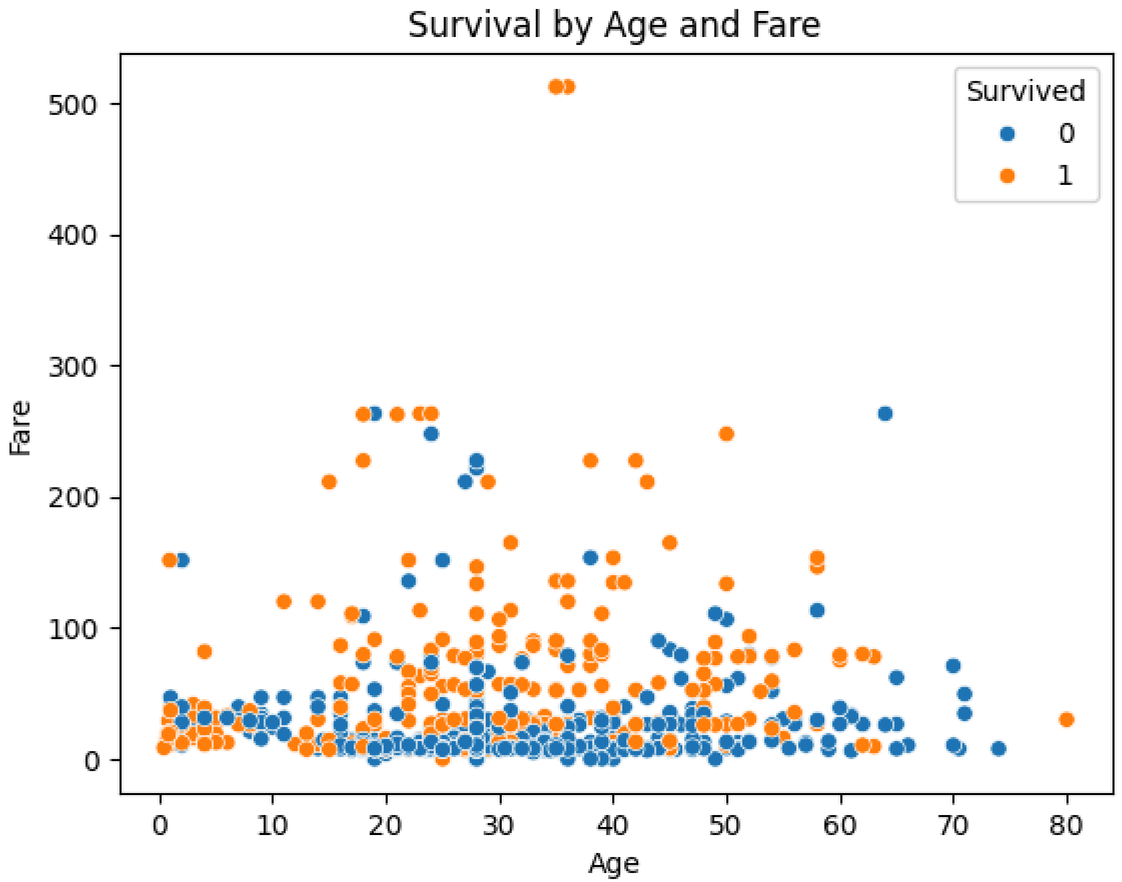
Observation: Higher survival rates are seen in 1st class, with the lowest in 3rd class, indicating class impacts survival.

In [57]:
```
sns.boxplot(x='Survived',y='Age',data=df)
plt.title('Survivals by Age')
plt.show()
```



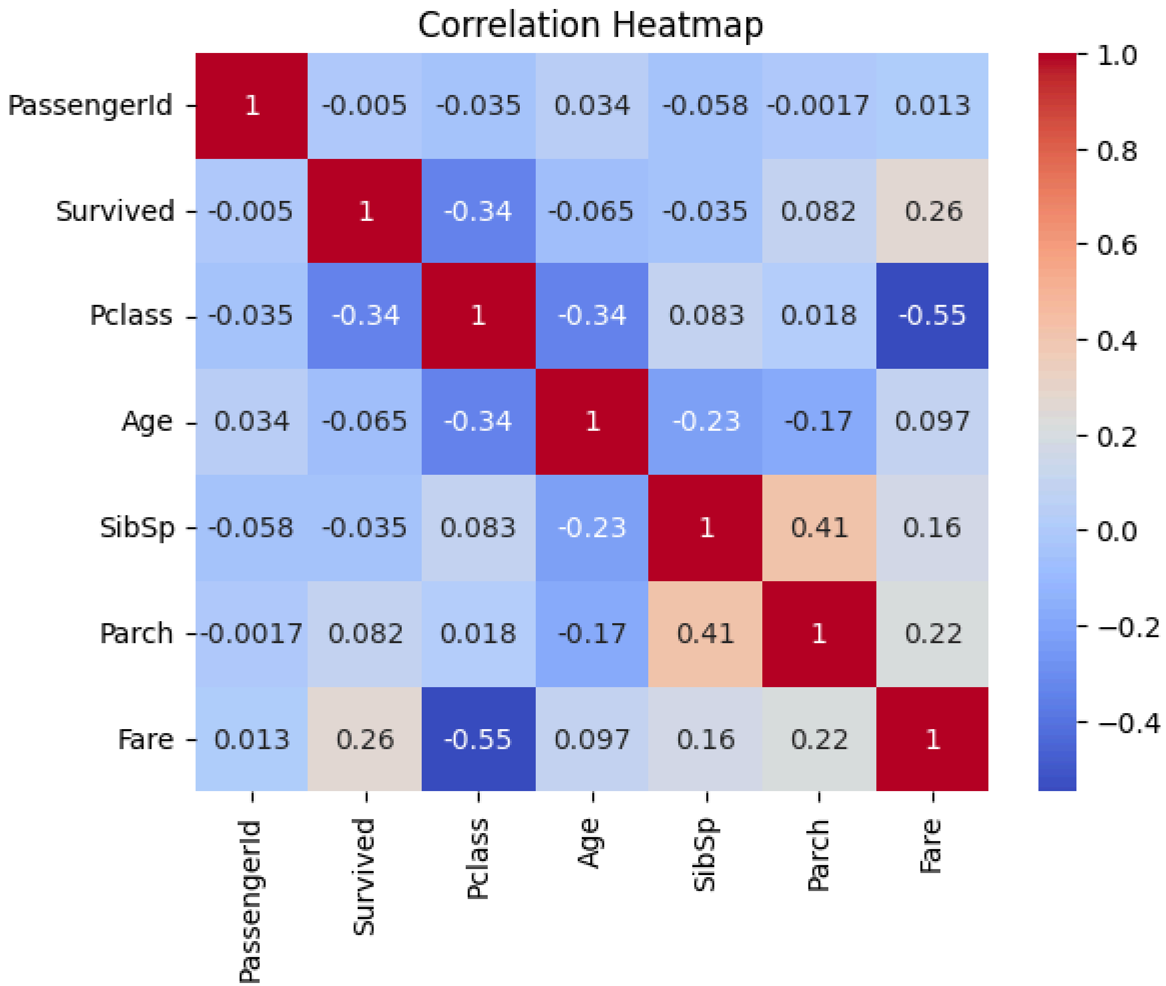
Observation: Median age of survivors and non-survivors is similar, but young children have higher survival rates.

In [58]:
```
sns.scatterplot(x='Age',y='Fare',hue='Survived',data=df)
plt.title('Survival by Age and Fare')
plt.show()
```



Observation: Younger passengers are spread across all fare ranges, while high fares are associated with older passengers in some cases.

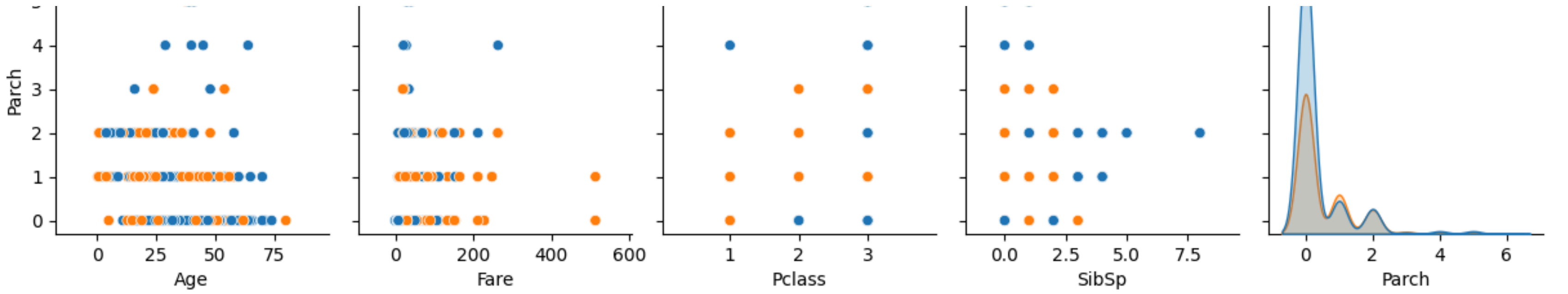
In [59]:
```
numeric_df=df.select_dtypes(include='number')
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



Observation: Fare and Pclass show moderate correlation. Other numeric features do not show strong correlations, reducing multicollinearity concerns.

In [60]:
```
sns.pairplot(df[['Survived','Age','Fare','Pclass','SibSp','Parch']],hue='Survived')
plt.show()
```


Age
Fare
Pclass
SibSp
Survived
0
1
80
60
40
20
0
500
400
300
200
100
3.0
2.5
2.0
1.5
1.0
8
6
4
2
5



Observation: Survivors are more concentrated in higher fare ranges and lower Pclass values, showing clear relationships with survival.

Summary of Findings

- Dataset: Titanic, 891 records, 12 columns.
- Missing values handled for 'Age' (median) and 'Embarked' (mode).
- Most passengers are young adults, and fares are right-skewed.
- Higher survival rates observed among females and first-class passengers.
- Fare and Pclass correlate with survival, making them important features.
- Multicollinearity is not a concern as numeric features show low correlations.
- Insights gained can guide predictive modeling and feature selection.