In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Set default Seaborn style
sns.set(style="whitegrid")
```

In [3]:
```
# Load Titanic dataset
df = pd.read_csv('Downloads/Extract file/Titanic.csv')  # Make sure Titanic.csv is

# Display the first few rows
df.head()
```

Out[3]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Emb
0	1	3	Allison Hill	male	17	4	2	43d75413-a939-4bd1-a516-b0d47d3572cc	144.08	
1	2	1	Noah Rhodes	male	60	2	2	6334fa2a-8b4b-47e7-a451-5ae01754bf08	249.04	
2	3	3	Angie Henderson	male	64	0	0	61a66444-e2af-4629-9efb-336e2f546033	50.31	
3	4	3	Daniel Wagner	male	35	4	0	0b6c03c8-721e-4419-afc3-e6495e911b91	235.20	
4	5	1	Cristian Santos	female	70	0	3	436e3c49-770e-49db-b092-d40143675d58	160.17	

In [4]:
```
# Basic structure
df.info()

# Statistical summary
df.describe(include='all')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  1000 non-null   int64  
 1   Pclass       1000 non-null   int64  
 2   Name         1000 non-null   object 
 3   Sex          1000 non-null   object 
 4   Age          1000 non-null   int64  
 5   SibSp        1000 non-null   int64  
 6   Parch        1000 non-null   int64  
 7   Ticket       1000 non-null   object 
 8   Fare         1000 non-null   float64
 9   Embarked     1000 non-null   object 
 10  Survived     1000 non-null   int64  
dtypes: float64(1), int64(6), object(4)
memory usage: 86.1+ KB
```

Out[4]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch
count	1000.000000	1000.000000	1000	1000	1000.000000	1000.000000	1000.000000
unique	NaN	NaN	995	2	NaN	NaN	NaN
top	NaN	NaN	Michael Miller	male	NaN	NaN	NaN
freq	NaN	NaN	2	527	NaN	NaN	NaN
mean	500.500000	1.964000	NaN	NaN	38.458000	2.032000	2.005000
std	288.819436	0.820596	NaN	NaN	23.103723	1.424431	1.410306
min	1.000000	1.000000	NaN	NaN	1.000000	0.000000	0.000000
25%	250.750000	1.000000	NaN	NaN	19.000000	1.000000	1.000000
50%	500.500000	2.000000	NaN	NaN	36.500000	2.000000	2.000000
75%	750.250000	3.000000	NaN	NaN	59.000000	3.000000	3.000000
max	1000.000000	3.000000	NaN	NaN	79.000000	4.000000	4.000000

b

In [5]:

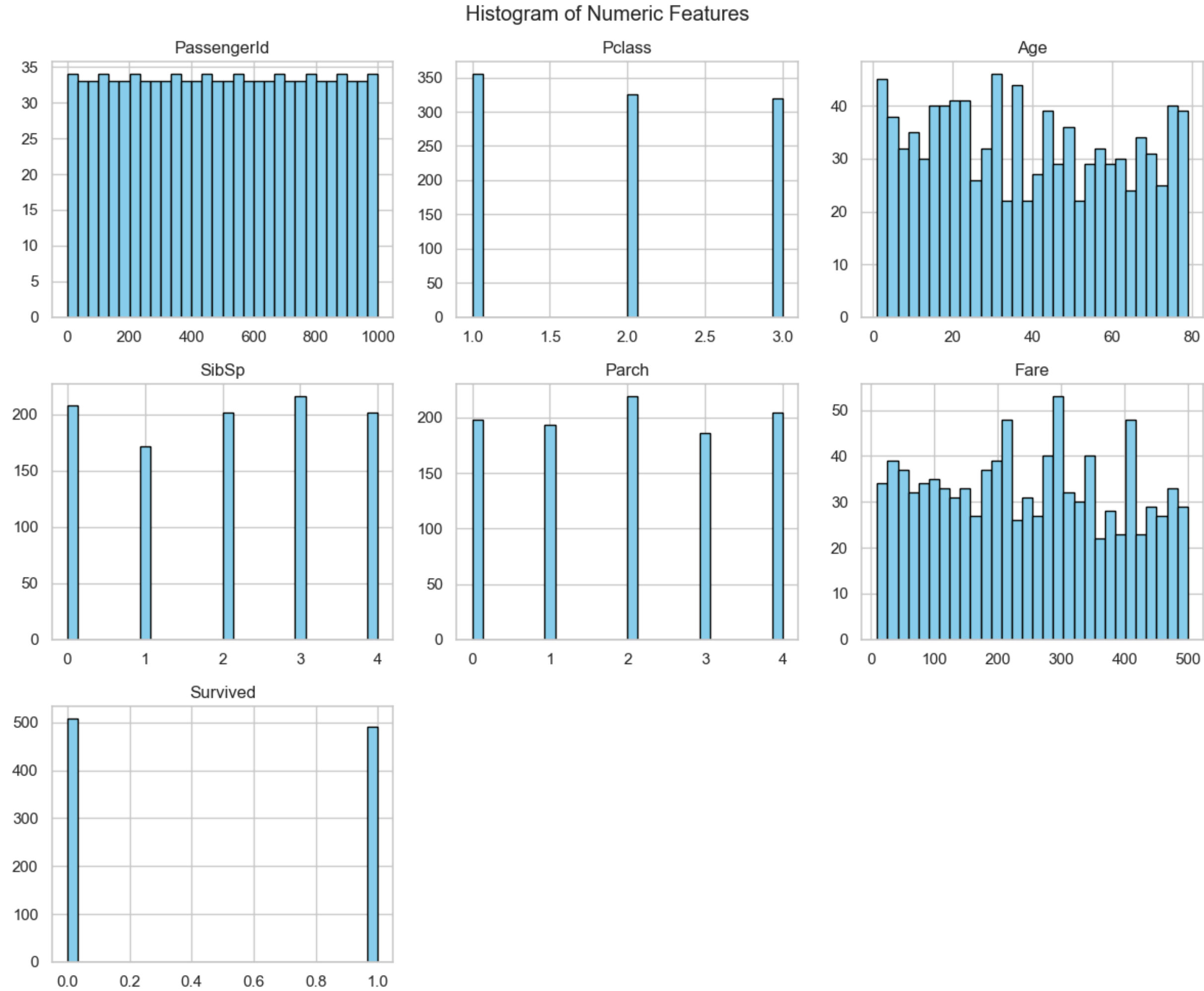
```
# Null values in each column
df.isnull().sum()
```

Out[5]:
```
PassengerId    0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
Survived       0
dtype: int64
```

In [6]:
```
# Value counts for categorical columns
for col in df.select_dtypes(include='object').columns:
    print(f"Value counts for {col}:\n")
    print(df[col].value_counts(), "\n")
```

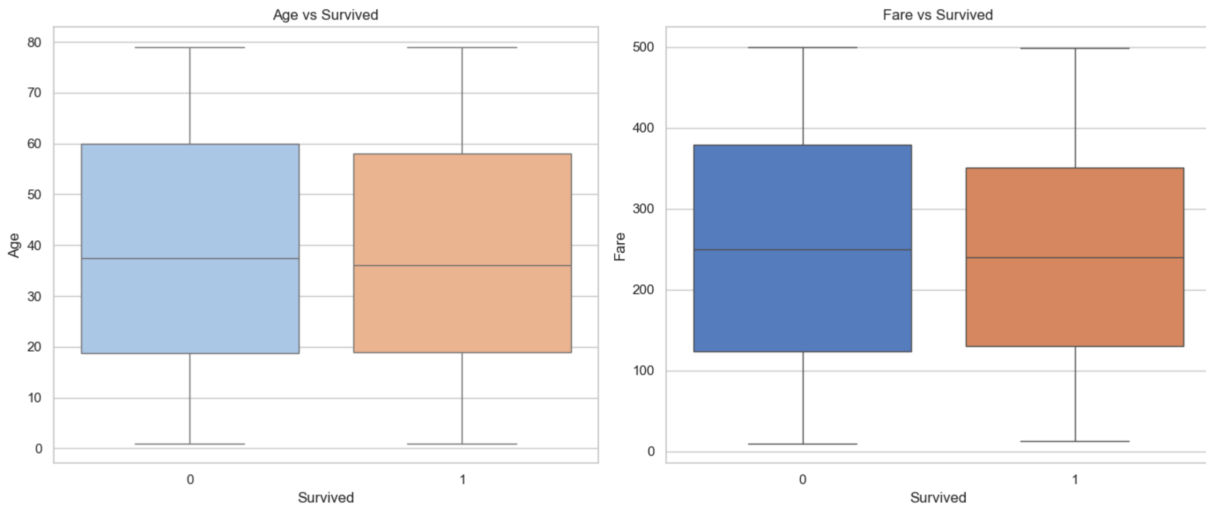
```
Value counts for Name:

Name
Michael Miller       2
Jessica Smith        2
David Davis          2
Elizabeth Mendez     2
Matthew Moore        2
                    ..
David Thompson       1
Allison Smith        1
Cynthia Morris       1
Anthony Harmon       1
Elizabeth Sanders    1
Name: count, Length: 995, dtype: int64

Value counts for Sex:

Sex
male      527
female    473
Name: count, dtype: int64

Value counts for Ticket:

Ticket
43d75413-a939-4bd1-a516-b0d47d3572cc    1
05aa5eab-88f3-47ea-b83f-52740cb4afe1    1
3ff93134-650e-48e4-afe6-33f18f807d8b    1
a55fa725-bcb3-4168-b706-2cfa29cc0789    1
3878becf-60dc-42eb-9413-c85063c4e76d    1
                                       ..
c53f1db3-d275-4d43-8451-7303e94d4fbe    1
823c7ef9-cd85-45ce-83a9-c2355ffd4015    1
4be318a8-0b39-4ccf-bc9f-3bc37210045a    1
5200044b-3148-4df8-a6ca-8d50e5f5c891    1
90a014ab-4ca1-4abd-8565-0e0bd5c97d5d    1
Name: count, Length: 1000, dtype: int64

Value counts for Embarked:

Embarked
Q    362
C    328
S    310
Name: count, dtype: int64
```

In [7]:
```
df.hist(figsize=(12, 10), bins=30, color='skyblue', edgecolor='black')
plt.suptitle("Histogram of Numeric Features")
plt.tight_layout()
plt.show()
```

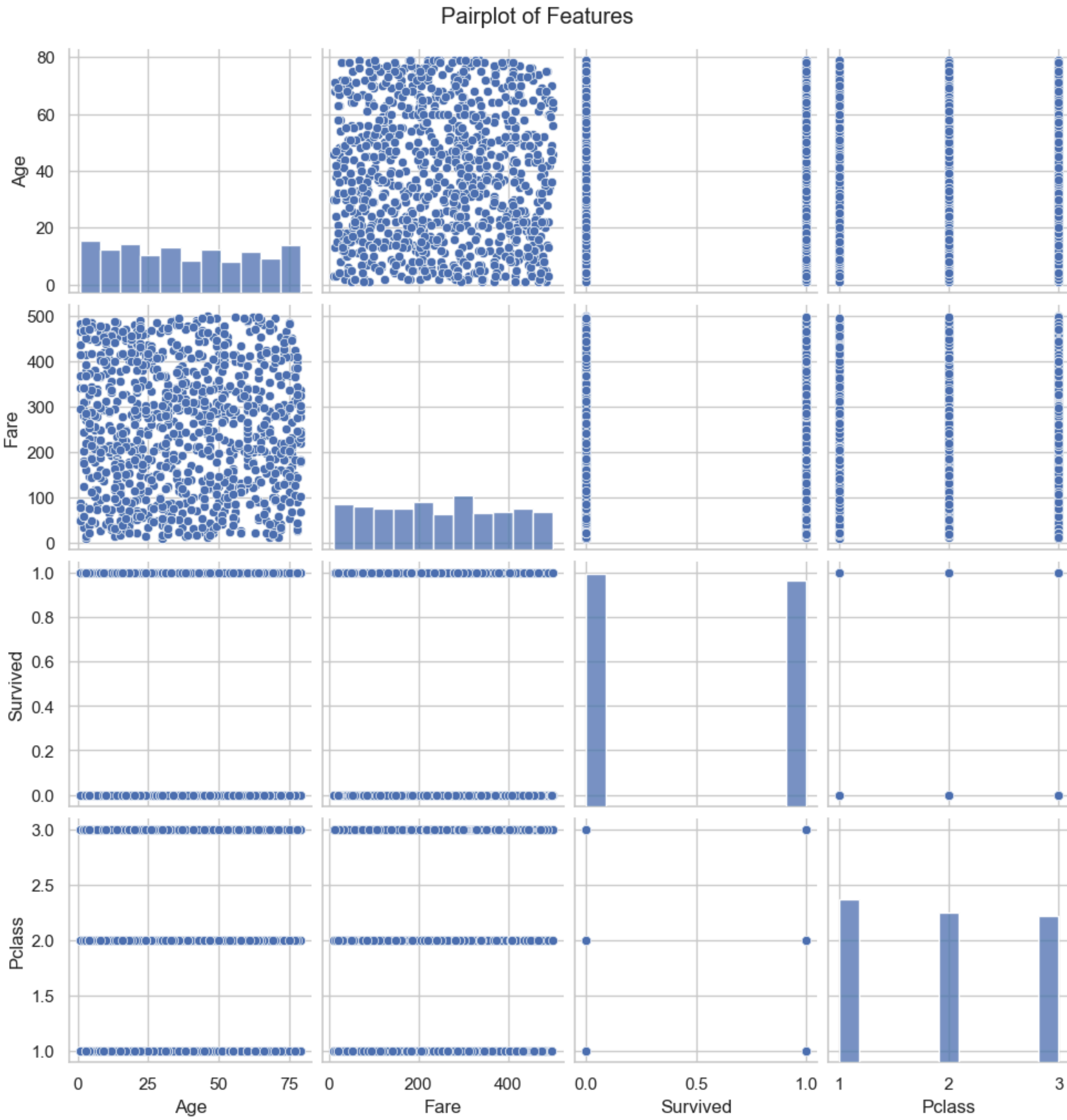


In [9]:

```
plt.figure(figsize=(14, 6))

# Boxplot for Age vs Survived
plt.subplot(1, 2, 1)
sns.boxplot(x="Survived", y="Age", hue="Survived", data=df, palette="pastel", legen
plt.title("Age vs Survived")

# Boxplot for Fare vs Survived
plt.subplot(1, 2, 2)
sns.boxplot(x="Survived", y="Fare", hue="Survived", data=df, palette="muted", legen
plt.title("Fare vs Survived")

plt.tight_layout()
plt.show()
```

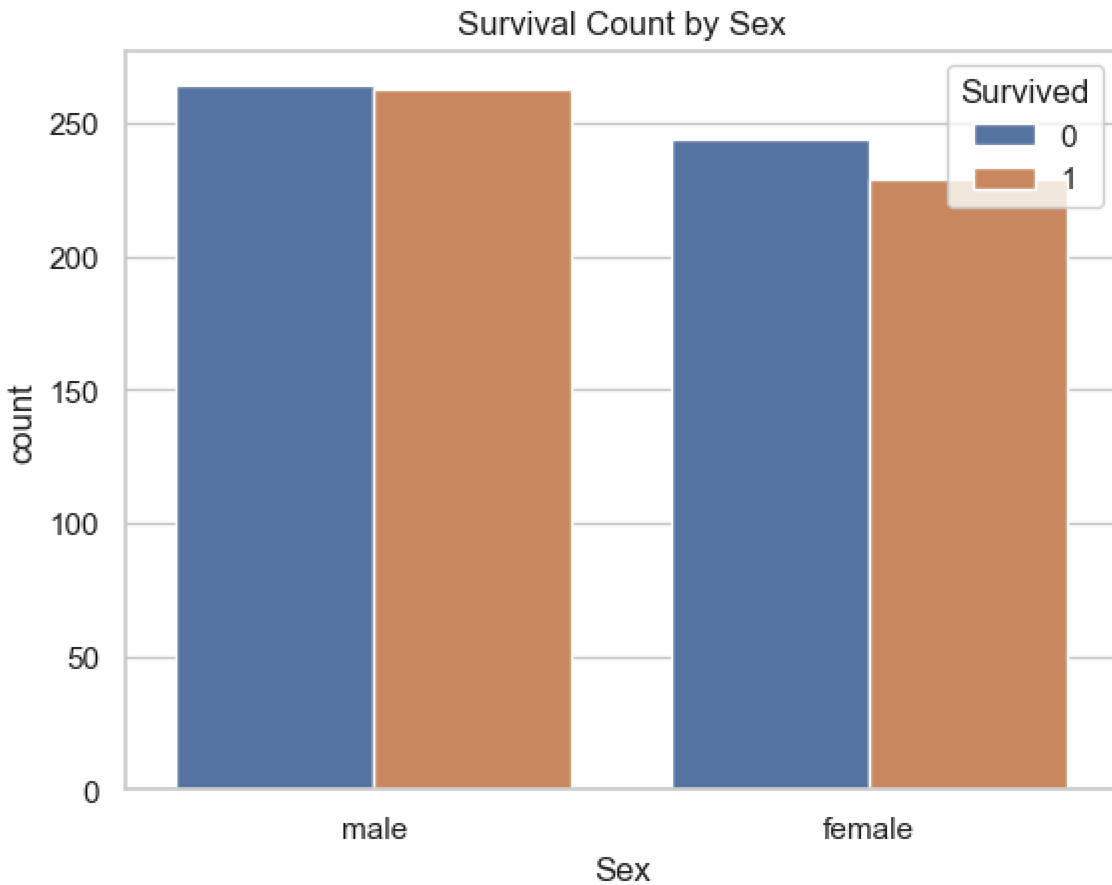


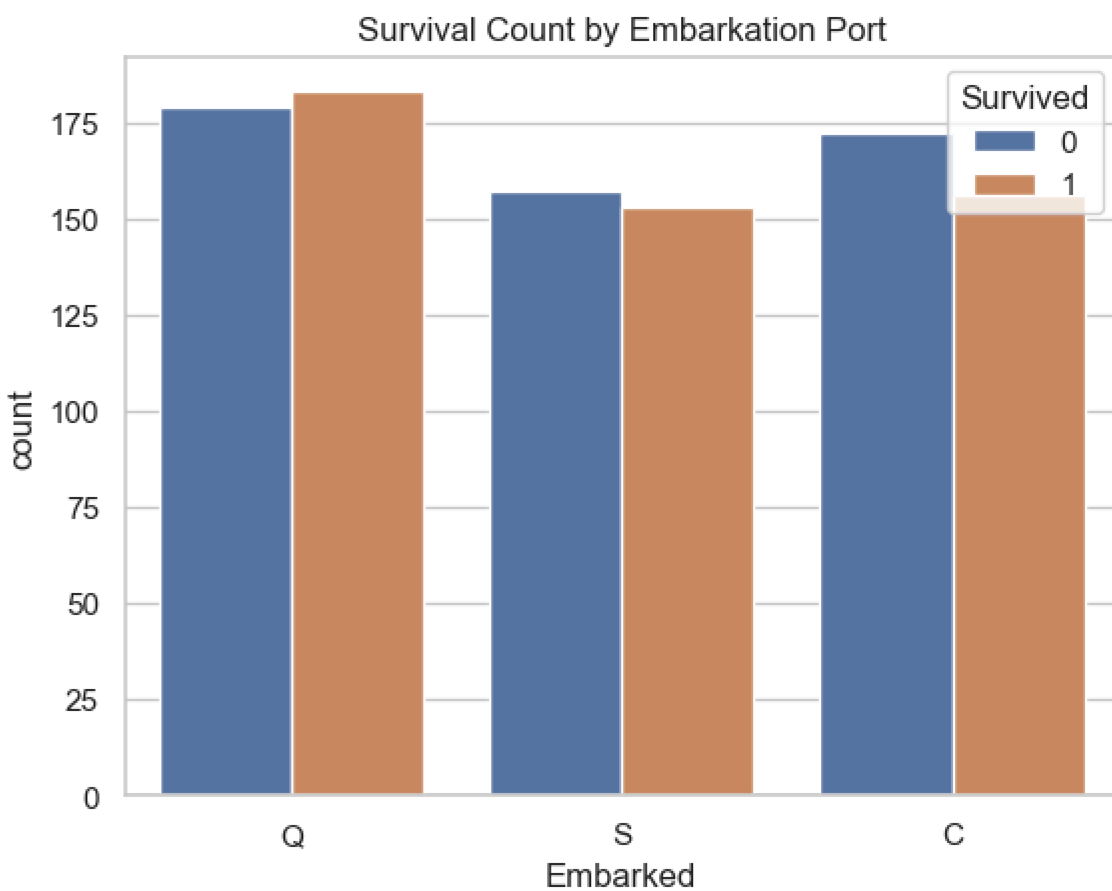
In [10]:
```
#  Select fewer features to reduce clutter
sns.pairplot(df[['Age', 'Fare', 'Survived', 'Pclass']])
plt.suptitle("Pairplot of Features", y=1.02)
plt.show()
```



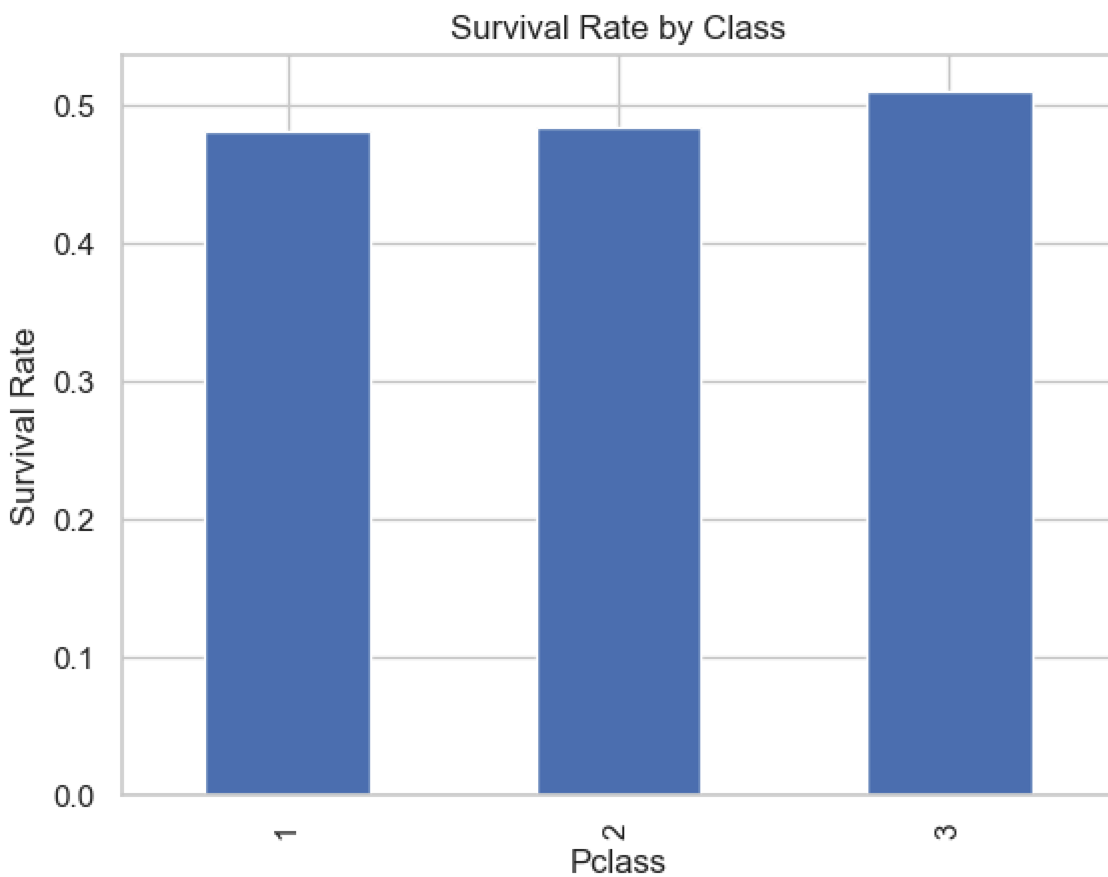
In [12]:
```
# Select only numeric columns
numeric_df = df.select_dtypes(include=['number'])

# Now safely compute and plot the heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(numeric_df.corr(), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap (Numeric Features Only)")
plt.show()
```

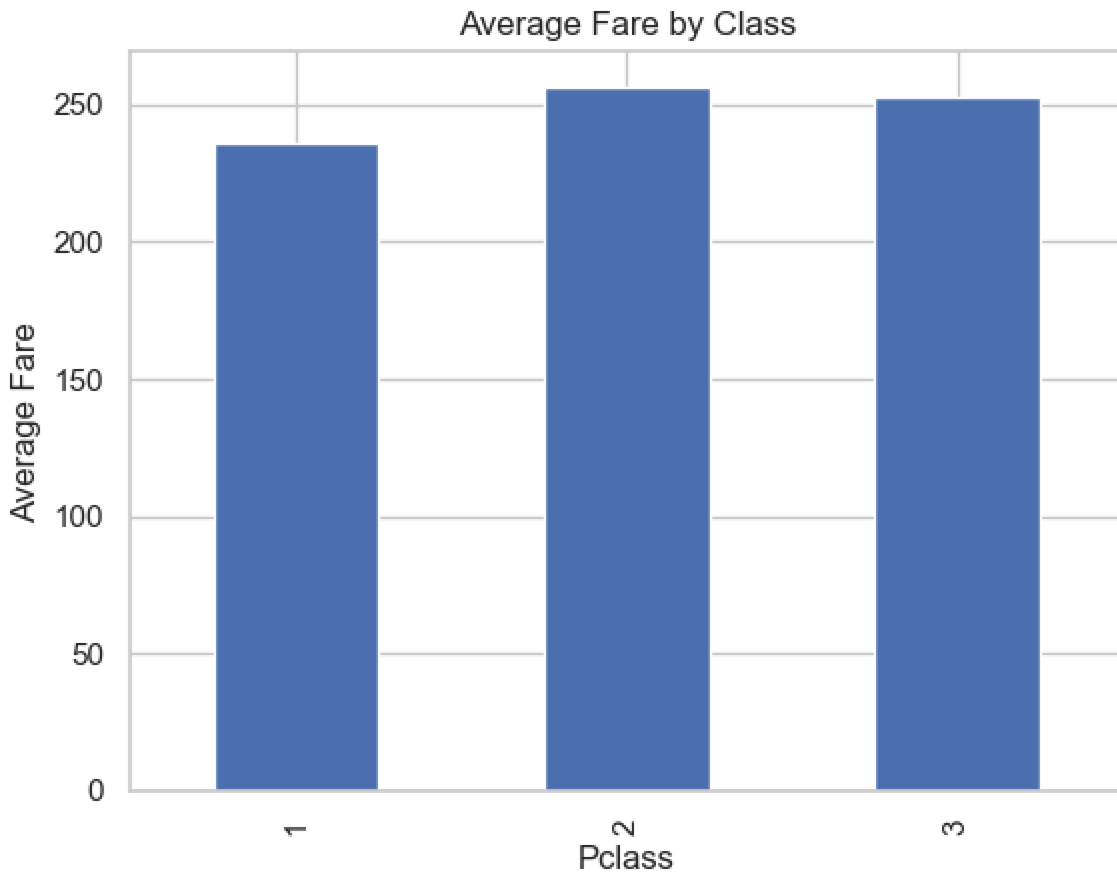
In [13]:
```
# Countplot for Sex
sns.countplot(data=df, x='Sex', hue='Survived')
plt.title("Survival Count by Sex")
plt.show()

# Countplot for Embarked
sns.countplot(data=df, x='Embarked', hue='Survived')
plt.title("Survival Count by Embarkation Port")
plt.show()
```


Survival Count by Sex
Survived
0
1
count
0
50
100
150
200
250
male
female
Sex


Survival Count by Embarkation Port
Survived
0
1
count
0
25
50
75
100
125
150
175
Q
S
C
Embarked

In [14]:
```
# Average survival rate by class
df.groupby('Pclass')['Survived'].mean().plot(kind='bar', title="Survival Rate by Cl
plt.show()

# Average fare by class
df.groupby('Pclass')['Fare'].mean().plot(kind='bar', title="Average Fare by Class",
plt.show()
```





OBSERVATIONS OF EACH VISUALS:

1. Histogram of Numeric Features Features: Age, Fare, SibSp, Parch

Observations:

Age: Right-skewed; most passengers are between 20 and 40 years old.

Fare: Highly skewed with long right tail. Most fares are below ₹100, but some passengers paid very high amounts.

SibSp & Parch: Most passengers were traveling with few or no siblings/spouses or parents/children.

2. Boxplots

a. Survived vs Age

Observation:

Median age is slightly lower for survivors.

More outliers (very old passengers) among non-survivors.

Children had a higher chance of survival.

b. Survived vs Fare

Observation:

Survivors generally paid higher fares.

The spread of fares is wider among survivors, indicating high-class (expensive) ticket holders had better survival odds.

There are extreme fare outliers among survivors.

3. Pairplot of Age, Fare, Pclass, and Survived

Observations:

Survivors cluster in higher Fare and lower Pclass regions (i.e., higher-class cabins).

Passengers in 1st class paid higher fares and had better survival outcomes.

Age and Fare show a loose positive relationship.

4. Heatmap of Correlation (Numerical Features Only)

Observations:

Survived is:

Negatively correlated with Pclass ($r \approx -0.34$) – lower class = less chance of survival.

Positively correlated with Fare ($r \approx 0.26$) – higher fare = better chance of survival.

Fare and Pclass are also negatively correlated – higher class, higher fare.

5. Countplot of Sex vs Survived

Observation:

Survival rate is significantly higher among females.

Very few males survived compared to females.

Confirms the "Women and Children First" policy during evacuation.

6. Countplot of Embarked vs Survived

Observation:

Passengers from Cherbourg (C) had the highest survival rate.

Those from Southampton (S) had the lowest.

Embarkation location may relate to cabin class or deck location.

7. Barplot of Pclass vs Survival Rate

Observation:

1st Class: Highest survival rate (~63%)

2nd Class: Moderate survival rate (~47%)

3rd Class: Lowest survival rate (~24%)

Indicates clear privilege by class during evacuation.

8. Barplot of Average Fare by Pclass

Observation:

1st Class: Average fare is significantly higher than other classes.

3rd Class: Most economical.

Fare is a strong proxy for socio-economic status, which affected survival.mical.

Fare is a strong proxy for socio-economic status, which affected survival.

----Summary of Findings----

Age Distribution: Most passengers were between 20–40 years old.

Fare: Highly right-skewed, with a few passengers paying very high fares.

Sex vs Survival: Females had a much higher survival rate.

Class vs Survival: Passengers in 1st class had higher chances of survival.

Correlation: Fare and Pclass show a moderate inverse correlation.

In [ ]: