In [33]:
```
import pandas as pd
df=pd.read_csv(r"C:\Users\SRAVANI\Documents\datasets\test.csv")
```

In [34]:
```
df.head()
```

Out[34]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN

In [35]:
```
df.describe()
```

Out[35]:

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	3.000000	27.000000	0.000000	0.000000	14.454200
75%	1204.750000	3.000000	39.000000	1.000000	0.000000	31.500000
max	1309.000000	3.000000	76.000000	8.000000	9.000000	512.329200

In [36]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Pclass       418 non-null    int64  
 2   Name         418 non-null    object 
 3   Sex          418 non-null    object 
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64  
 6   Parch        418 non-null    int64  
 7   Ticket       418 non-null    object 
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object 
 10  Embarked     418 non-null    object 
dtypes: float64(2), int64(4), object(5)
memory usage: 36.1+ KB
```

In [37]:
```
df.isnull().sum()
```

Out[37]:
```
PassengerId      0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

In [38]:
```
df.duplicated().sum()
```

Out[38]:
```
0
```

In [39]:
```
df['Sex'].value_counts()
```

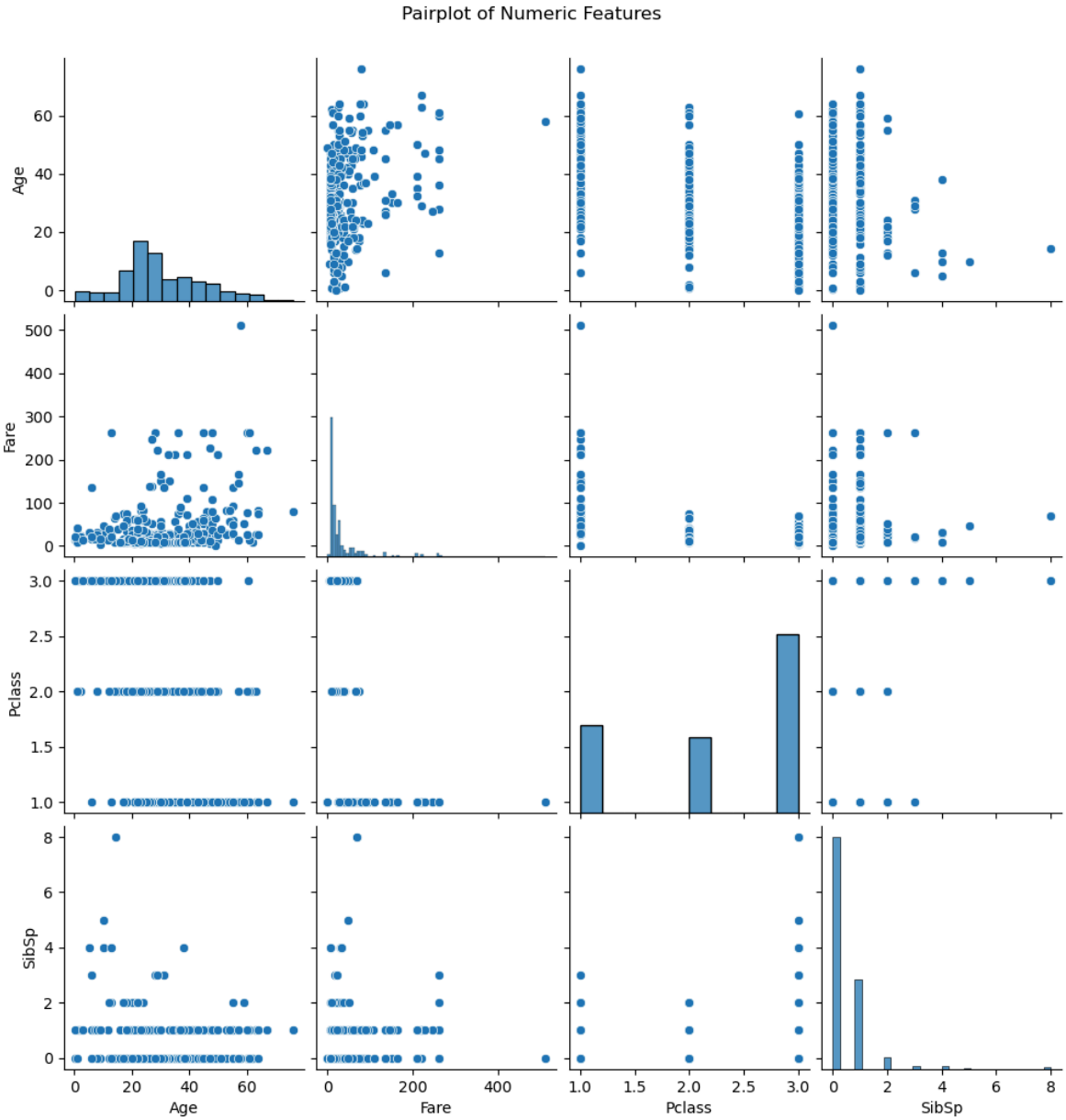
Out[39]:
```
Sex
male      266
female    152
Name: count, dtype: int64
```

In [40]:
```
df['Age'].value_counts()
```

Out[40]:
```
Age
21.0    17
24.0    17
22.0    16
30.0    15
18.0    13
        ..
76.0     1
28.5     1
22.5     1
62.0     1
38.5     1
Name: count, Length: 79, dtype: int64
```

Pairplot

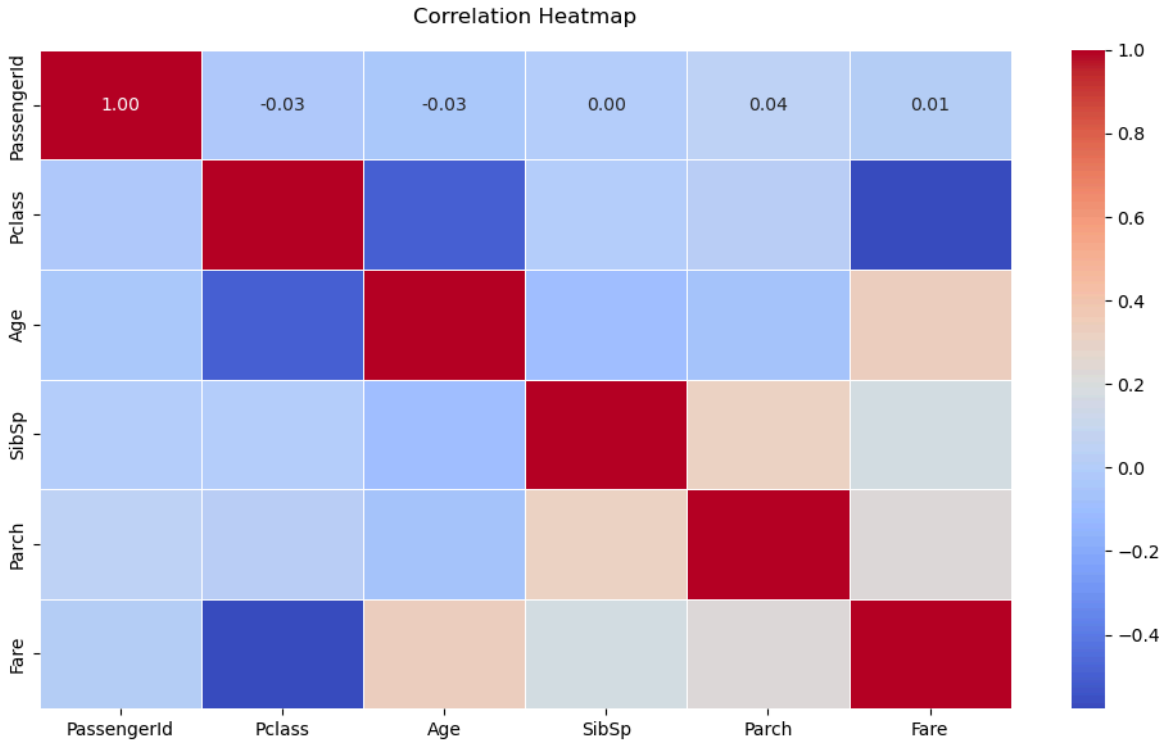
In [49]:
```
import matplotlib.pyplot as plt
import seaborn as sns
sns.pairplot(df[['Age', 'Fare', 'Pclass', 'SibSp']])
plt.suptitle("Pairplot of Numeric Features", y=1.03)
plt.show()
```



Heatmap

In [59]:
```
correlation_matrix = df.corr(numeric_only=True)

plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidt
plt.title('Correlation Heatmap', pad=15)
plt.tight_layout()
plt.show()
```



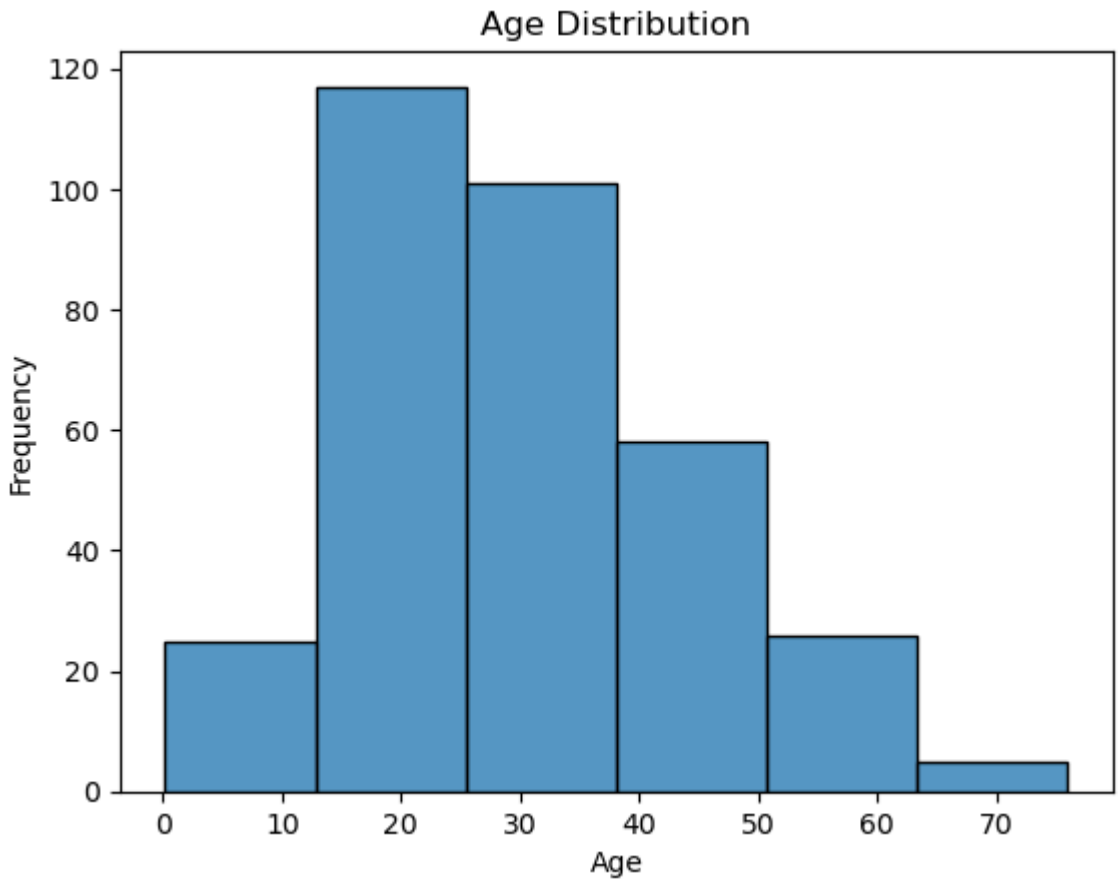
In [63]:
```
df.corr(numeric_only=True)
```

Out[63]:

	PassengerId	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.026751	-0.034102	0.003818	0.043080	0.008211
Pclass	-0.026751	1.000000	-0.492143	0.001087	0.018721	-0.577147
Age	-0.034102	-0.492143	1.000000	-0.091587	-0.061249	0.337932
SibSp	0.003818	0.001087	-0.091587	1.000000	0.306895	0.171539
Parch	0.043080	0.018721	-0.061249	0.306895	1.000000	0.230046
Fare	0.008211	-0.577147	0.337932	0.171539	0.230046	1.000000

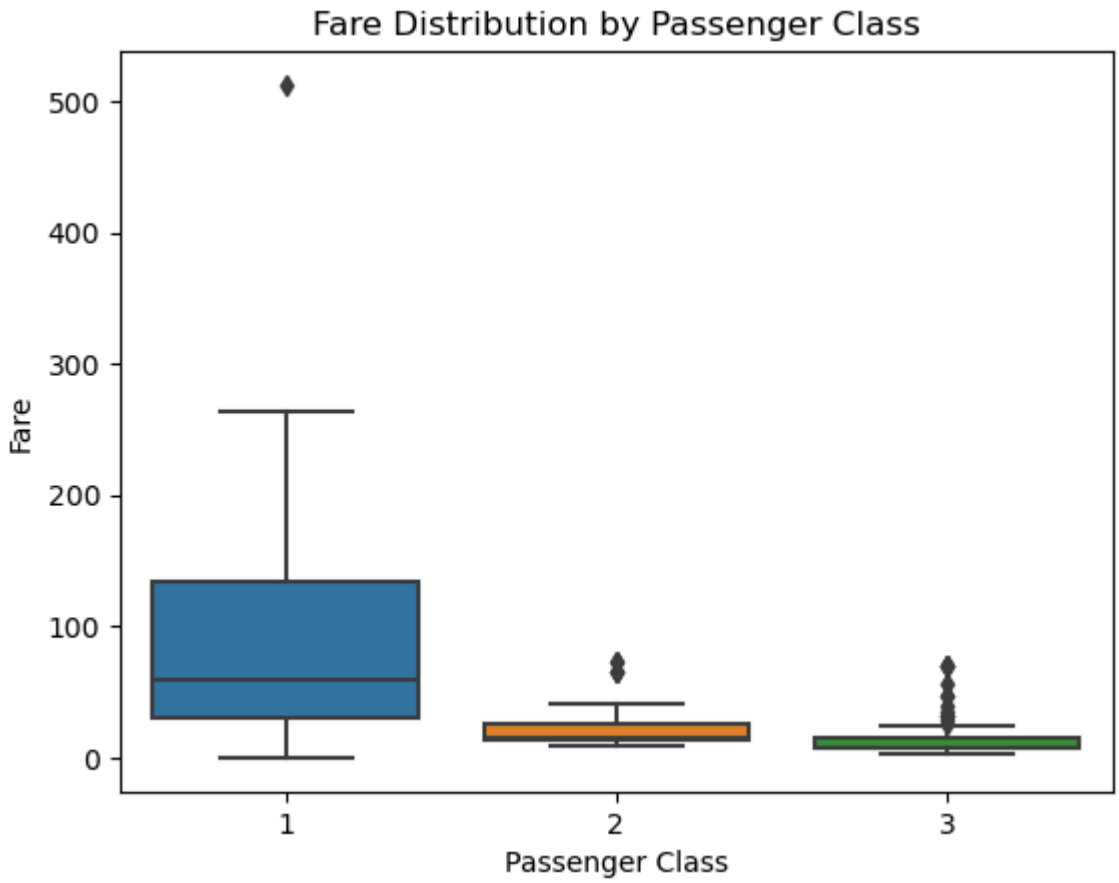
Histogram

In [75]:
```
sns.histplot(df['Age'], bins=6)
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()
```

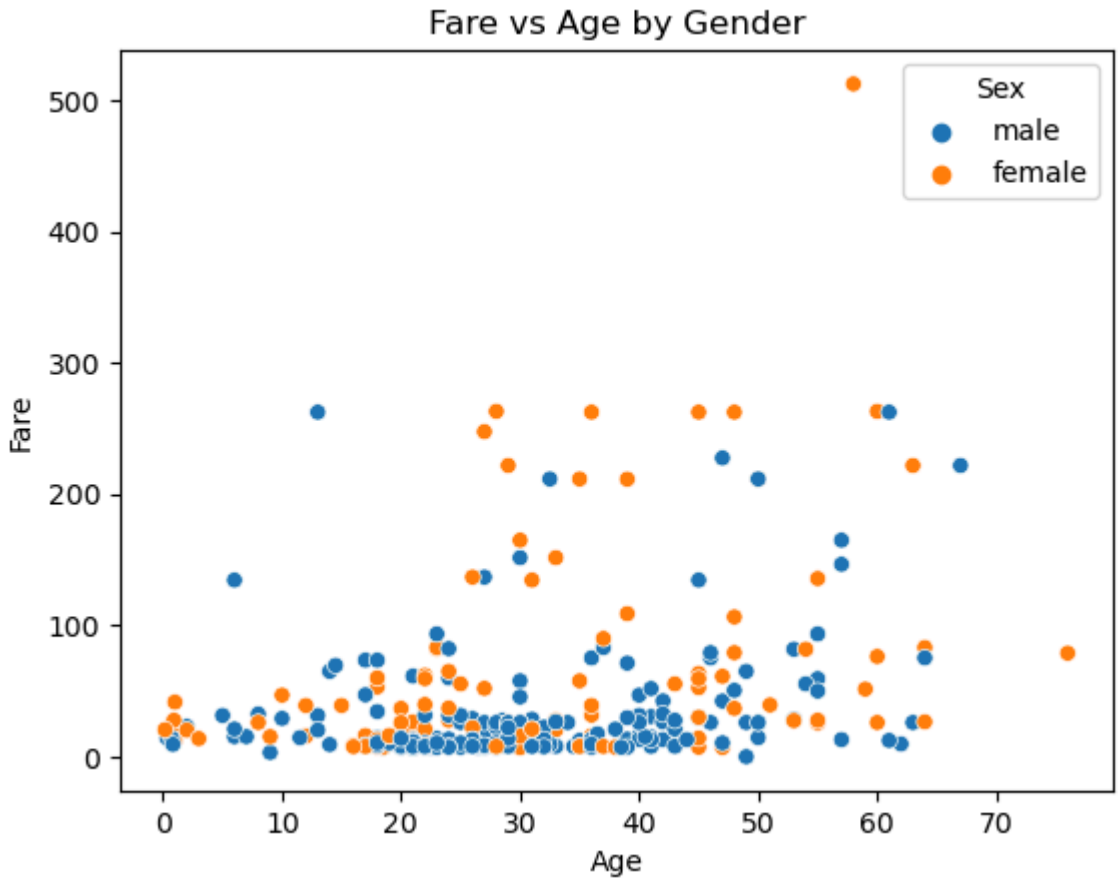


Boxplot

In [81]:
```
sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title("Fare Distribution by Passenger Class")
plt.xlabel("Passenger Class")
plt.ylabel("Fare")
plt.show()
```



In [89]:
```
sns.scatterplot(x='Age', y='Fare', hue='Sex', data=df)
plt.title("Fare vs Age by Gender")
plt.show()
```



Pairplot:

The pairplot provides scatterplots and distributions for multiple numerical features (e.g., Age, Fare, SibSp, Parch).

It helps reveal patterns and relationships between features.

Some clustering is visible in Fare vs Pclass—1st class passengers generally paid higher fares.

Heatplot:

The heatmap shows the correlation between numerical features in the dataset.

Strong positive correlation is observed between SibSp and Parch, indicating passengers with siblings likely had parents/children aboard too.

Pclass has a moderate negative correlation with Fare—higher classes paid higher fares.

This visualization helps quickly identify which features may be useful for predicting or understanding other features

Histogram:

Most passengers are in the 20–40 age group.

Boxplot:

1st class passengers paid significantly higher fares.

Scatterplot:

No clear trend between age and fare, but females appear to have slightly higher fares in some cases.

In [ ]: