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

In [2]:
```
df=pd.read_csv("https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv")
```

In [3]:
```
df
```

Out[3]:

	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	NaN	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 [4]:
```
df.shape
```

Out[4]:
```
(891, 12)
```

In [5]:
```
df.head()
```

Out[5]:

	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 [ ]:
```
```

In [6]:
```
df.info()
```

```
<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 [7]:
```
df.describe()
```

Out[7]:

	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 [8]:
```
df.dtypes
```

Out[8]:
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

In [9]:
```
#data cleaning

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

Out[9]:
```
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 [10]:
```
df.duplicated().sum()
```

Out[10]:
```
0
```

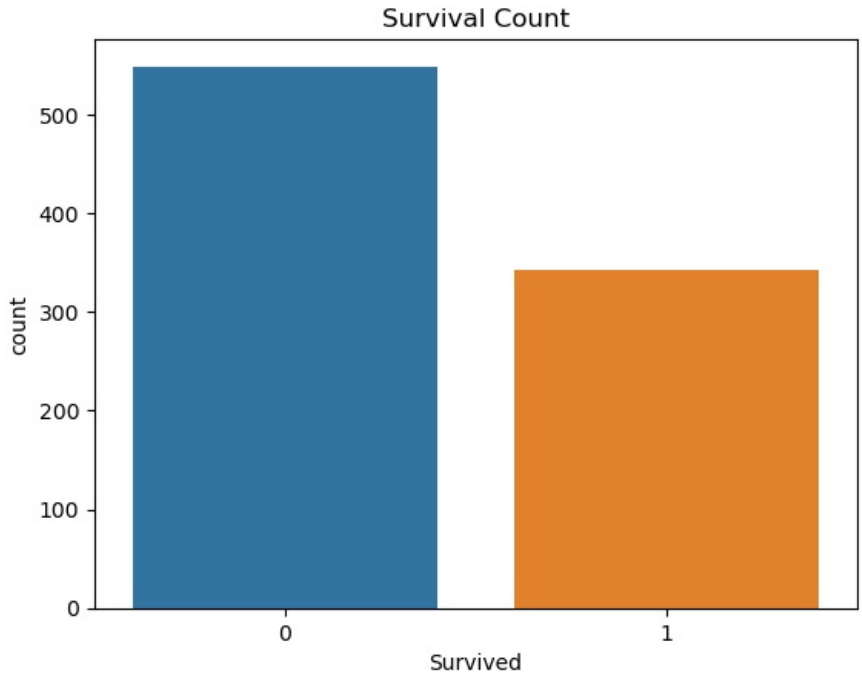
In [11]:
```
#converting all objects into numerical values

df.dtypes
```

Out[11]:
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

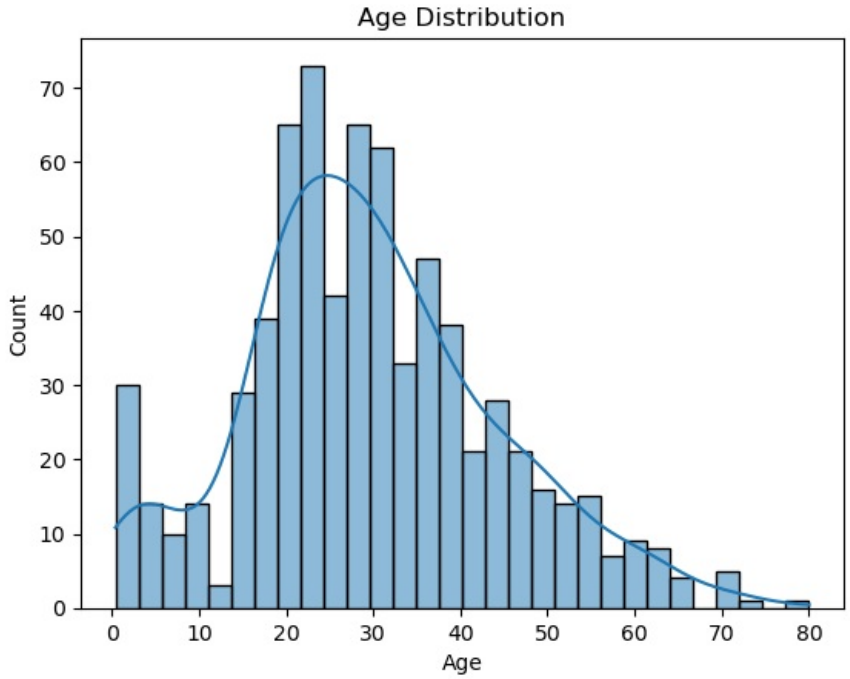
In [12]:
```
df["Survived"].value_counts()
```

Out[12]:
```
Survived
0    549
1    342
Name: count, dtype: int64
```

In [13]:
```
#EDA

sns.countplot(x=df["Survived"])
plt.title('Survival Count')
plt.show()
```

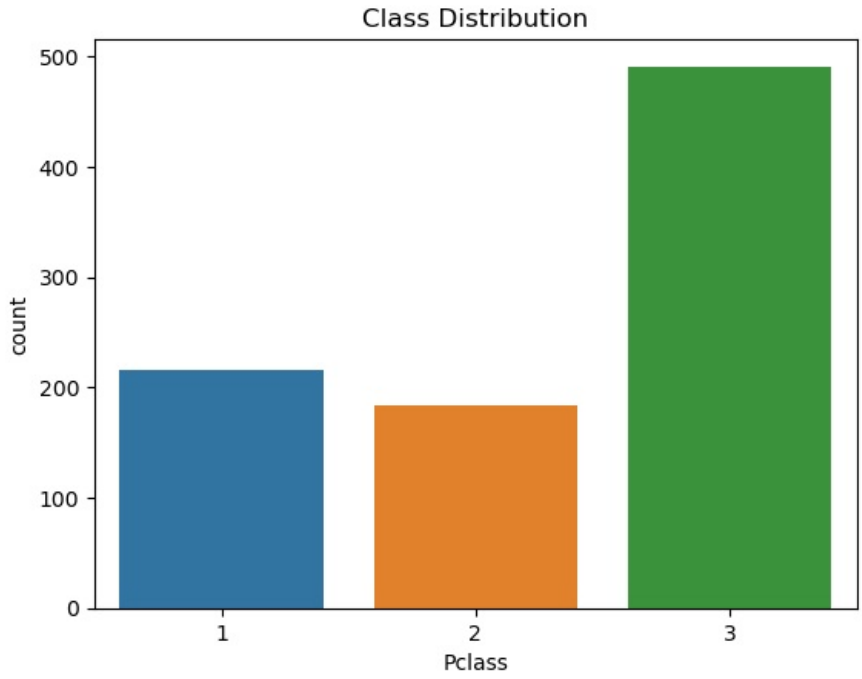


In [14]:
```
sns.histplot(df['Age'].dropna(), bins=30, kde=True)
plt.title('Age Distribution')
plt.show()
```



In [15]:
```
#From the above graph it is clear that not many persons survived.
#Out of 183 only 120 survived and 60 didn't survive
```

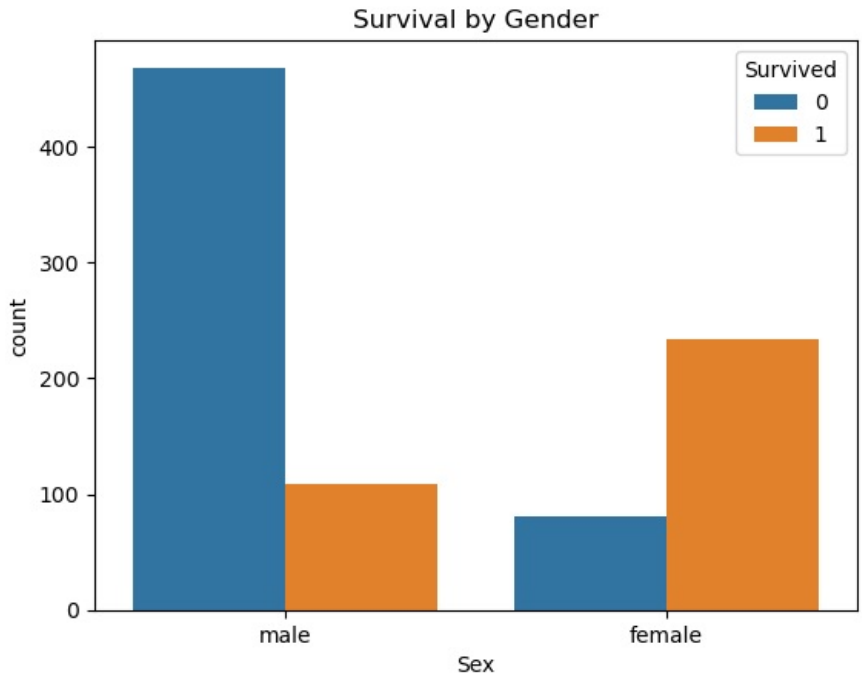
In [16]:
```
#Countplot of class (Pclass)
sns.countplot(x='Pclass', data=df)
plt.title('Class Distribution')
plt.show()
```



In [17]:
```
#Bivariate Analysis

df.groupby(['Sex','Survived'])['Survived'].count()
```

Out[17]:
```
Sex     Survived
female  0            81
        1           233
male    0           468
        1           109
Name: Survived, dtype: int64
```

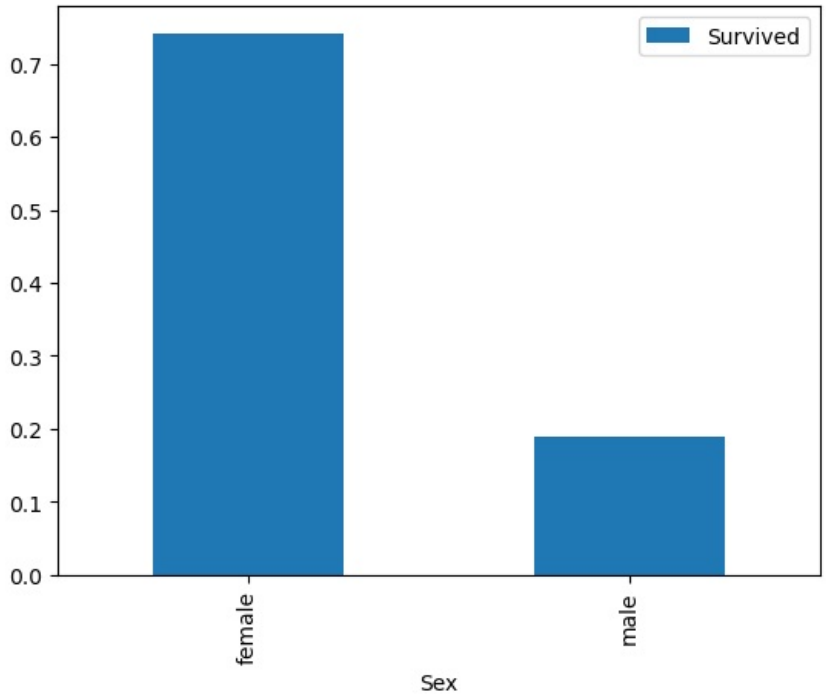
In [18]:
```
# Countplot of survival by gender
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival by Gender')
plt.show()
```



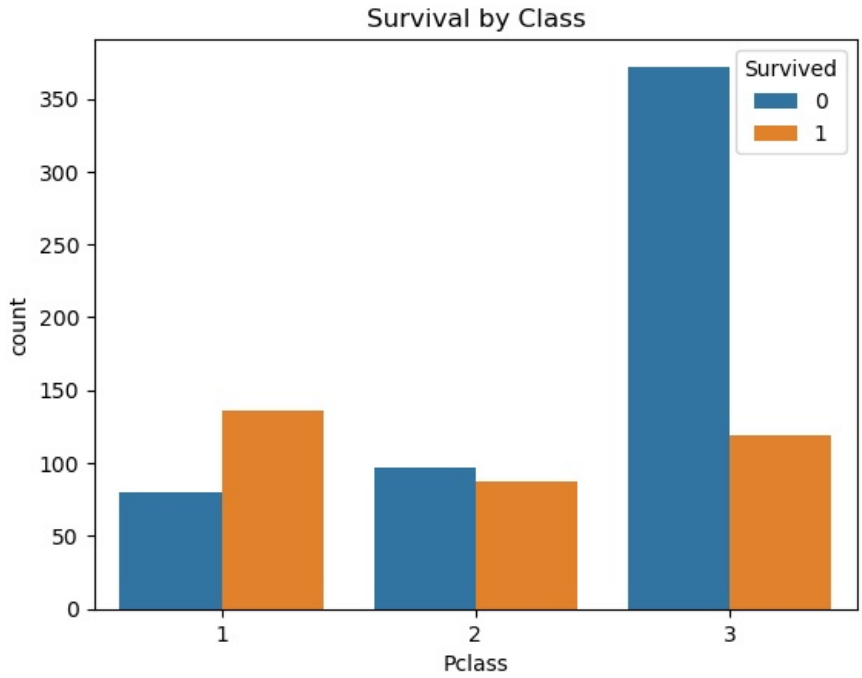
In [19]:
```
#From the above graph it is clear that not which sex people survived more.
#Out of 88 only 82 survived and 6 females didn't survive
#out of 95 males 54 could not survived and 41 survided.
#This shows female survivzal rate is more than male.
```

In [20]:
```
df[['Sex','Survived']].groupby(['Sex']).mean().plot.bar()
```

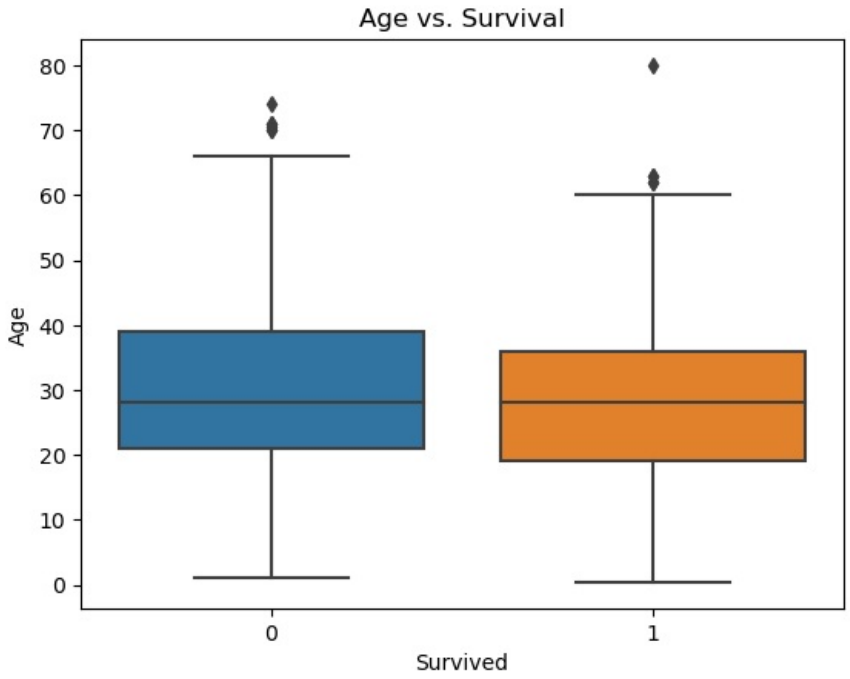
Out[20]:
```
<Axes: xlabel='Sex'>
```



In [21]:
```
# Countplot of survival by class
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title('Survival by Class')
plt.show()
```



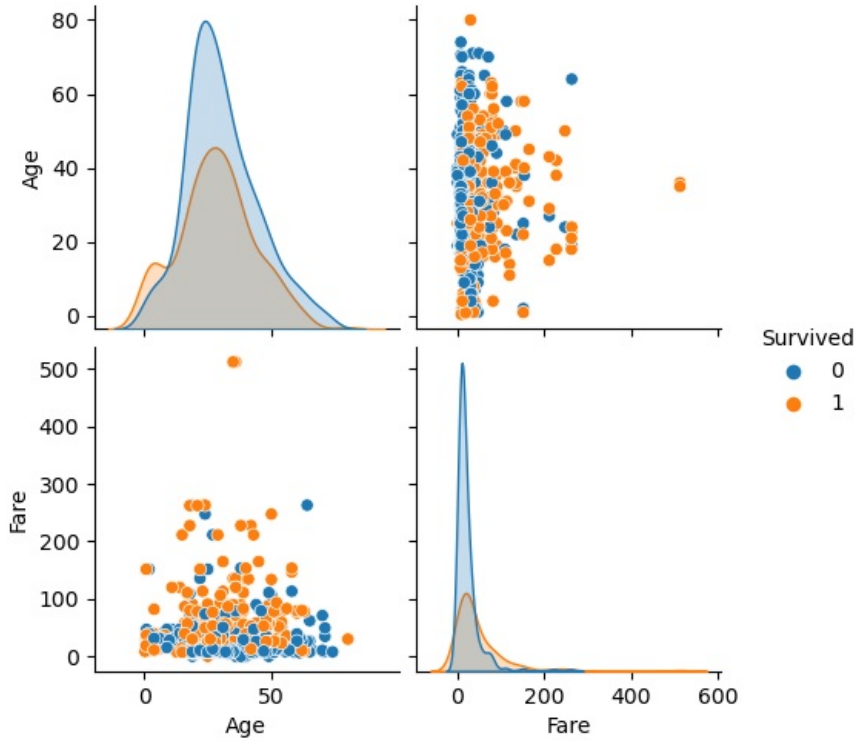
In [22]:
```
# Boxplot of age by survival
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age vs. Survival')
plt.show()
```



In [23]:
```
#Younger passengers had a higher survival rate on the Titanic, as indicated by a lower median age among survivo
# The survival rate was concentrated within a specific age range, suggesting age significantly influenced the l
#outier shows that one passenger of age around 80 also survived
```

In [24]:
```
#Multivariate Analysis
# Pairplot of selected features
sns.pairplot(df[['Age', 'Fare', 'Survived']], hue='Survived', diag_kind='kde')
plt.show()
```

```
C:\Users\om\Downloads\New folder\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The figure layout has
changed to tight
  self._figure.tight_layout(*args, **kwargs)
```



In [25]:
```
#insights
#Passengers who paid higher fares and were younger appear to have higher survival rates,
#as indicated by distinct clustering and distribution patterns in the pairplot.
```

In [26]:
```
df.columns
```

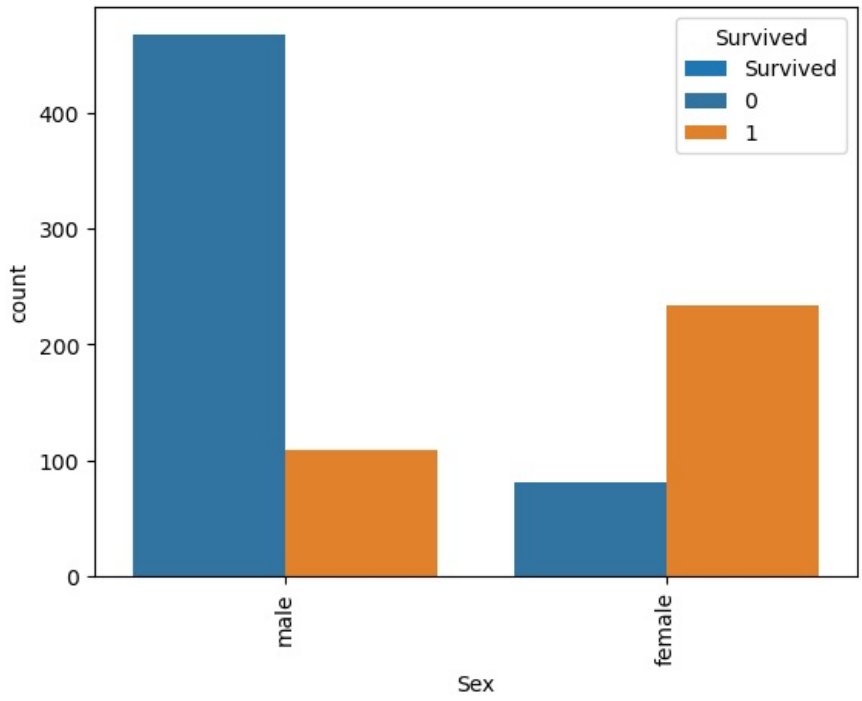
Out[26]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [27]:
```
# Fill missing age with the median age
df['Age'].fillna(df['Age'].median(), inplace=True)

# Fill missing embarked with the mode
```

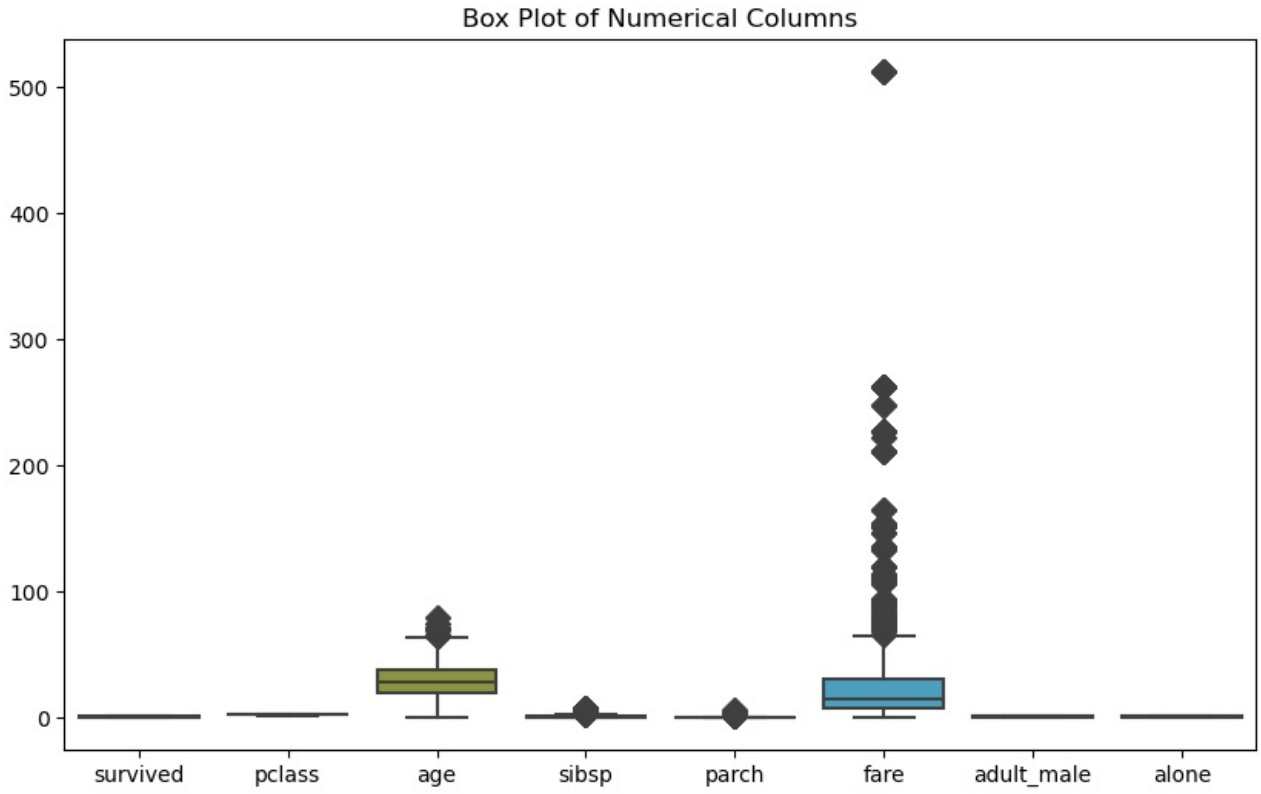
```
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

# Verify missing values after handling
print(df.isnull().sum())
```

```
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 [28]:
```
df[['Sex','Survived']].groupby(['Sex']).mean().plot.bar()
sns.countplot(x='Sex',hue='Survived',data=df)
plt.show()
```

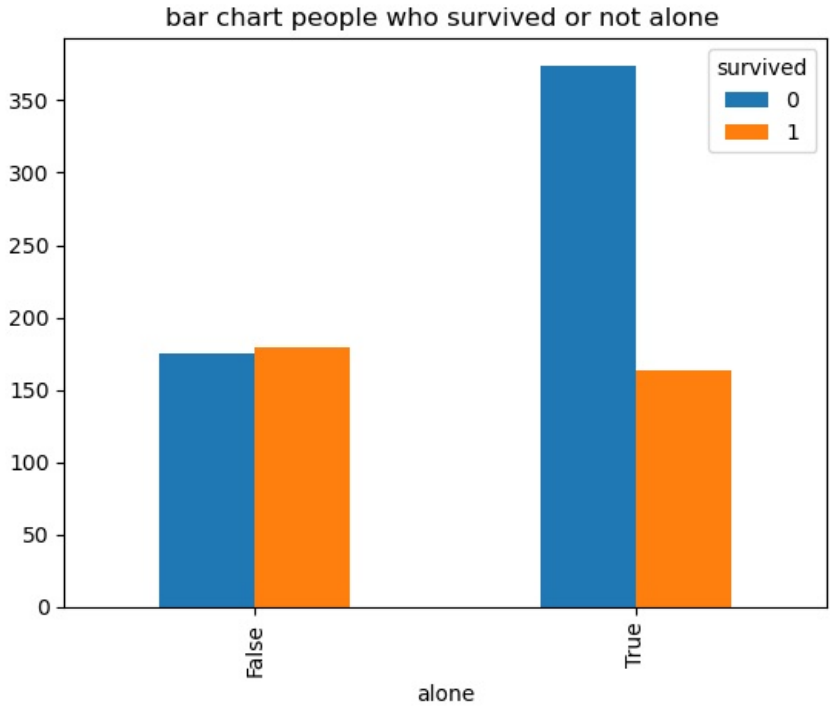


In [30]:
```
# pie chart for percentage of man ,women,and child
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the Titanic dataset
df = sns.load_dataset('titanic')

# Select numerical columns
num_col = df.select_dtypes(exclude="O").columns

flierprops = dict(marker='D', color='red', markersize=8)

# Create the box plot
plt.figure(figsize=(10, 6))
sns.boxplot(data=df[num_col], flierprops=flierprops,palette="husl")
plt.title("Box Plot of Numerical Columns")
plt.show()
```



In [31]:
```
plt.figure(figsize=(10,6))
pd.crosstab(df['alone'],df['survived']).plot(kind='bar')
plt.title("bar chart people who survived or not alone ")
```

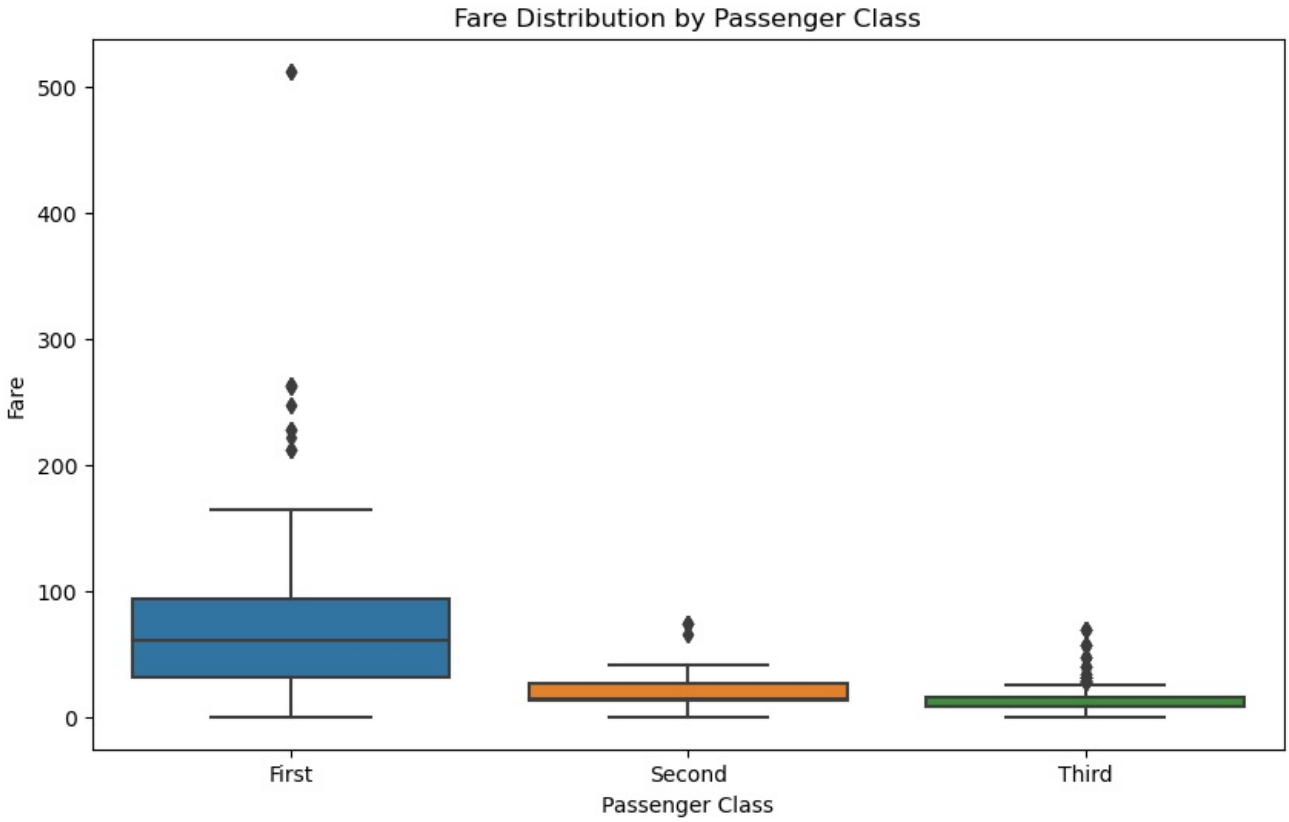
Out[31]:
```
Text(0.5, 1.0, 'bar chart people who survived or not alone ')

<Figure size 1000x600 with 0 Axes>
```



In [32]:
```
df.groupby('pclass')['age'].mean()  #average age of pansanges who is in p class 1 2 3
```

Out[32]:
```
pclass
1    38.233441
2    29.877630
3    25.140620
Name: age, dtype: float64
```

In [33]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='class', y='fare', data=df)
plt.title('Fare Distribution by Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Fare')
plt.show()
```



In [ ]: