

untitled1

March 16, 2025

```
[66]: pwd
```

```
[66]: 'C:\\Users\\Mohammed Al Malki'
```

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

```
[3]: import numpy as np
```

```
[5]: import seaborn as sns
```

```
[6]: import matplotlib.pyplot as plt
```

```
[7]: df = pd.read_csv('Titanic-Dataset.csv')
```

```
[8]: df.head()
```

```
[8]:
```

	PassengerId	Survived	Pclass	\
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	0	3	

	Name	Sex	Age	SibSp	\
0	Braund, Mr. Owen Harris	male	22.0	1	
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	
2	Heikkinen, Miss. Laina	female	26.0	0	
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	
4	Allen, Mr. William Henry	male	35.0	0	

	Parch	Ticket	Fare	Cabin	Embarked
0	0	A/5 21171	7.2500	NaN	S
1	0	PC 17599	71.2833	C85	C
2	0	STON/O2. 3101282	7.9250	NaN	S
3	0	113803	53.1000	C123	S
4	0	373450	8.0500	NaN	S

```
[9]: df.describe()
```

```
[9]:
```

	PassengerId	Survived	Pclass	Age	SibSp \
count	891.000000	891.000000	891.000000	714.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008
std	257.353842	0.486592	0.836071	14.526497	1.102743
min	1.000000	0.000000	1.000000	0.420000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000
50%	446.000000	0.000000	3.000000	28.000000	0.000000
75%	668.500000	1.000000	3.000000	38.000000	1.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000

	Parch	Fare
count	891.000000	891.000000
mean	0.381594	32.204208
std	0.806057	49.693429
min	0.000000	0.000000
25%	0.000000	7.910400
50%	0.000000	14.454200
75%	0.000000	31.000000
max	6.000000	512.329200

```
[10]: print(df.head())
```

	PassengerId	Survived	Pclass	\
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	0	3	

	Name	Sex	Age	SibSp \
0	Braund, Mr. Owen Harris	male	22.0	1
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1
2	Heikkinen, Miss. Laina	female	26.0	0
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1
4	Allen, Mr. William Henry	male	35.0	0

	Parch	Ticket	Fare	Cabin	Embarked
0	0	A/5 21171	7.2500	NaN	S
1	0	PC 17599	71.2833	C85	C
2	0	STON/O2. 3101282	7.9250	NaN	S
3	0	113803	53.1000	C123	S
4	0	373450	8.0500	NaN	S

```
[11]: print(df.isnull().sum())
```

```
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

```

```
[27]: median_age = df['Age'].median()
```

```
[28]: df['Age'] = df['Age'].fillna(median_age)
```

```
[29]: df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
```

```
[30]: df['Cabin'] = df['Cabin'].fillna('Unknown')
```

```
[31]: #Data
```

```
[32]: df.to_csv('Titanic-Cleaned-Dataset.csv', index=False)
```

```
[33]: 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          0
Embarked       0
dtype: int64

```

```
[34]: Q1 = df['Fare'].quantile(0.25)
```

```
[35]: Q3 = df['Fare'].quantile(0.75)
```

```
[36]: IQR = Q3 - Q1
```

```
[37]: lower_bound = Q1 - 1.5 * IQR
```

```
[38]: upper_bound = Q3 + 1.5 * IQR
```

```
[39]: df['Fare'] = np.where(df['Fare'] < lower_bound, lower_bound, df['Fare'])
```

```
[40]: df['Fare'] = np.where(df['Fare'] > upper_bound, upper_bound, df['Fare'])
```

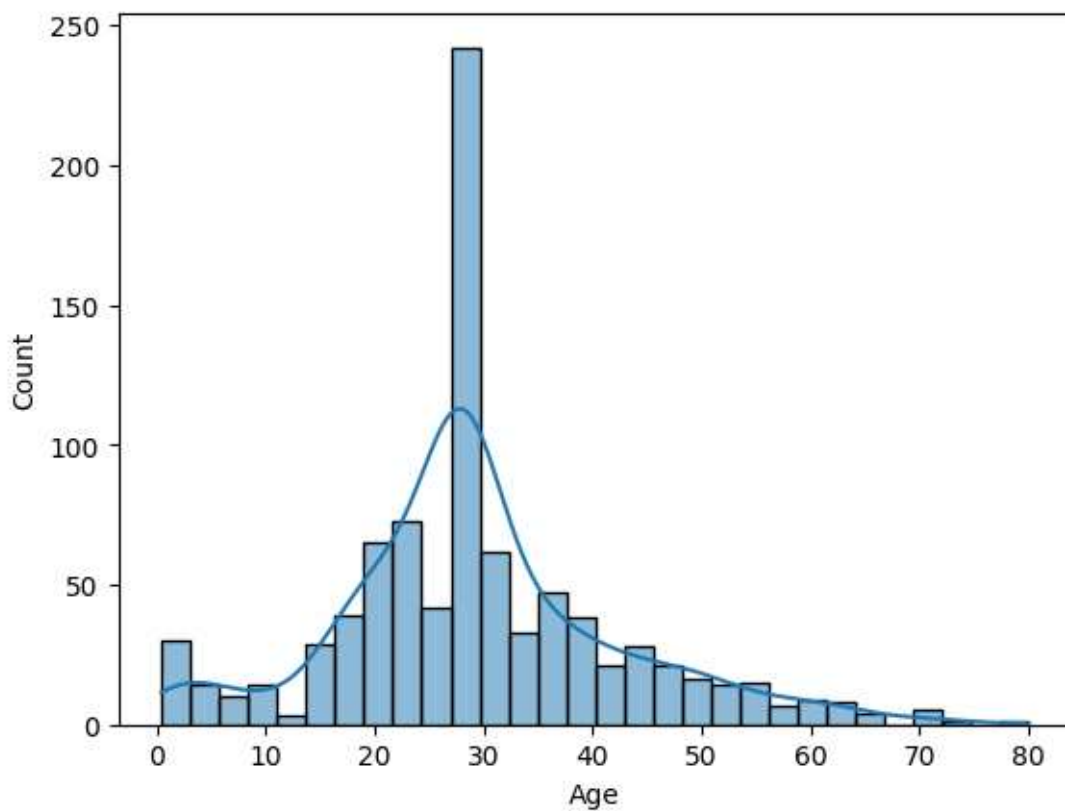
```
[41]: plt.figure(figsize=(10, 6))
```

```
[41]: <Figure size 1000x600 with 0 Axes>
```

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

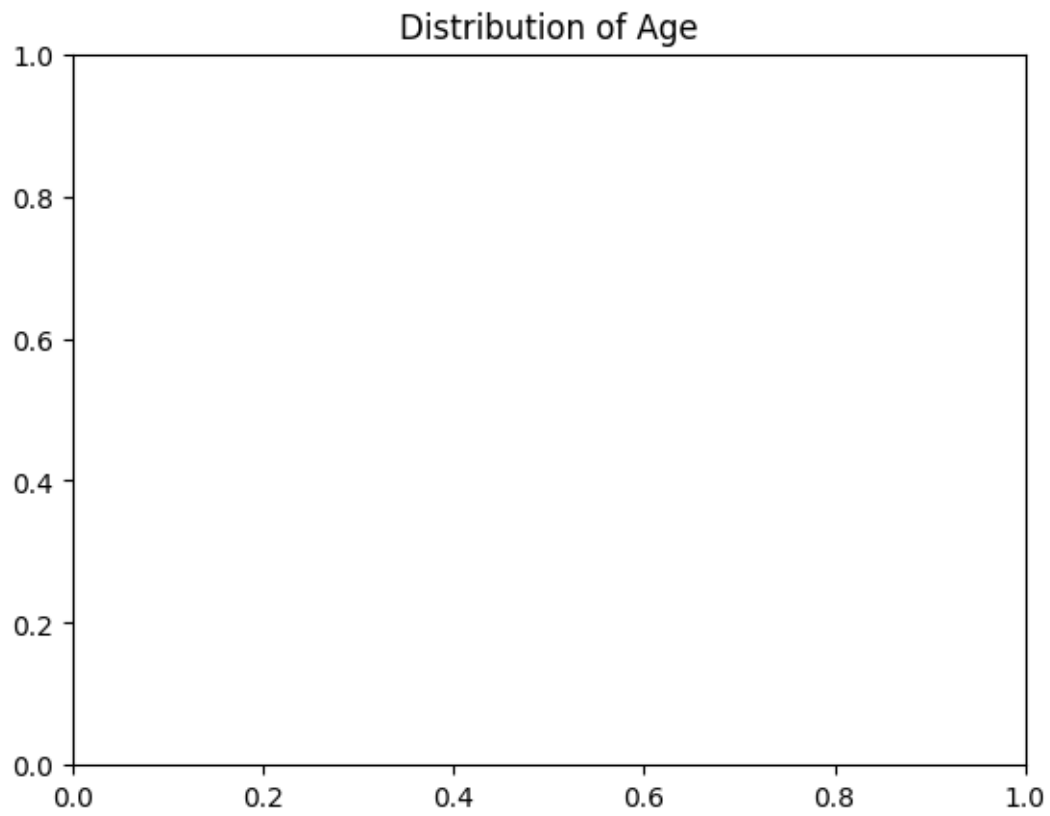
```
[42]: sns.histplot(df['Age'], bins=30, kde=True)
```

```
[42]: <Axes: xlabel='Age', ylabel='Count'>
```



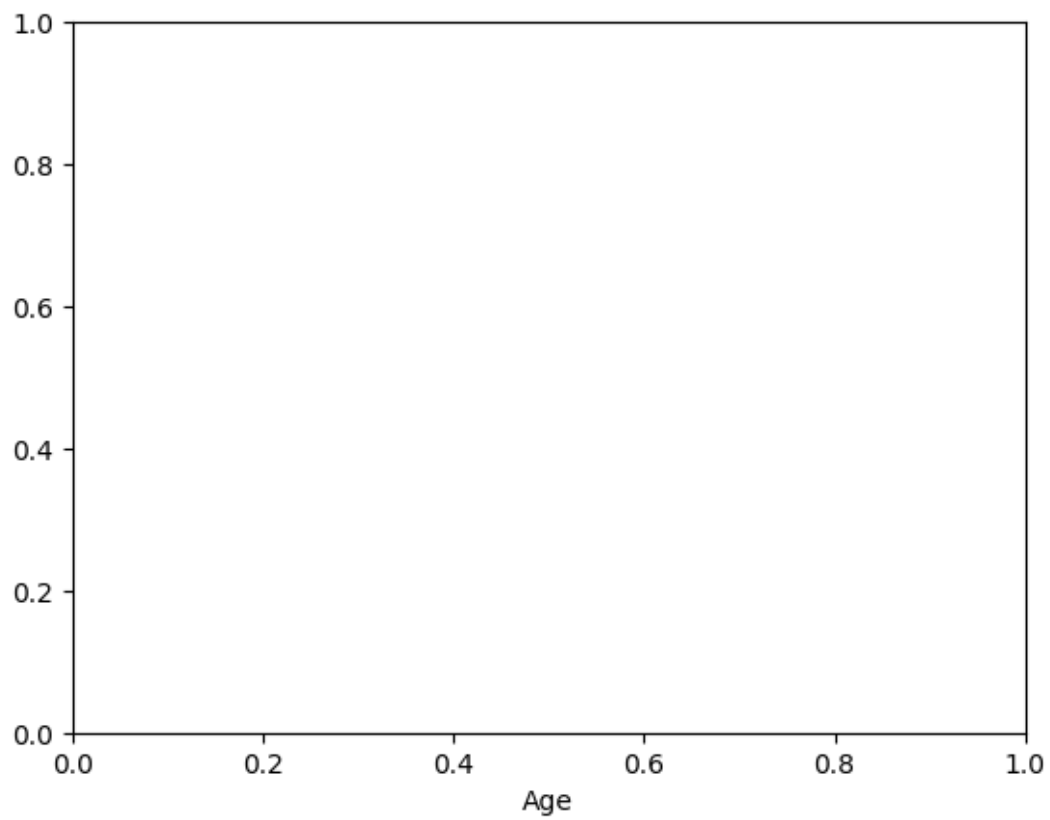
```
[43]: plt.title('Distribution of Age')
```

```
[43]: Text(0.5, 1.0, 'Distribution of Age')
```



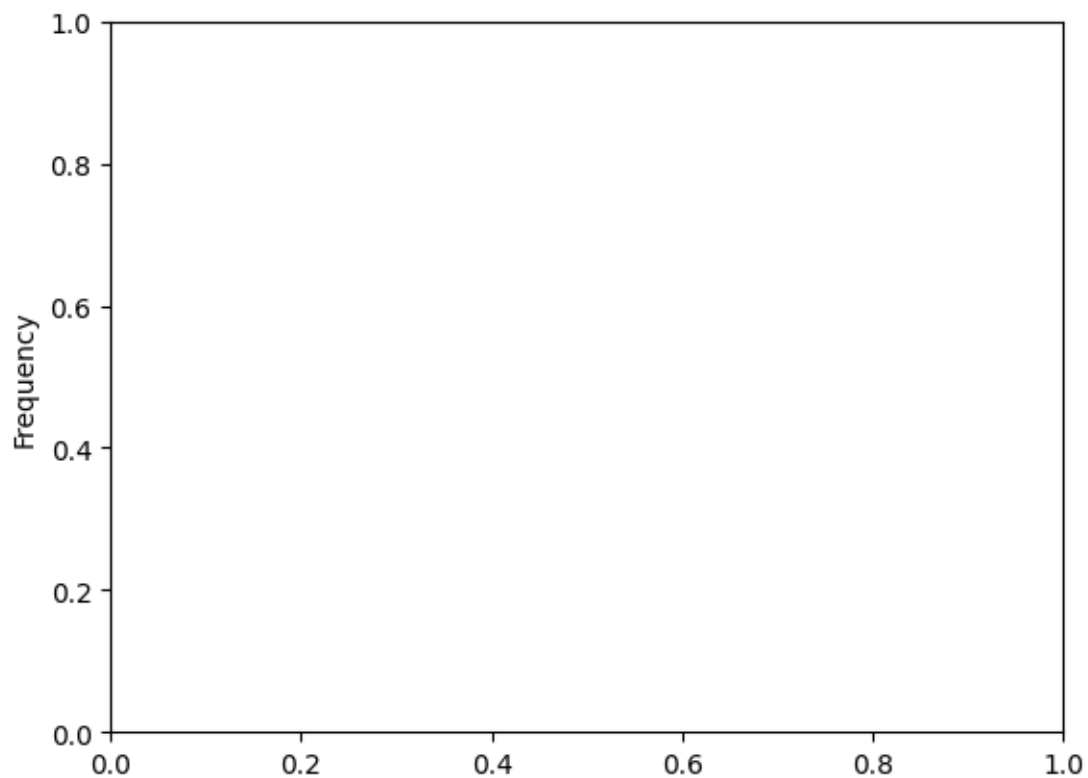
```
[44]: plt.xlabel('Age')
```

```
[44]: Text(0.5, 0, 'Age')
```



```
[45]: plt.ylabel('Frequency')
```

```
[45]: Text(0, 0.5, 'Frequency')
```



```
[46]: plt.show()
```

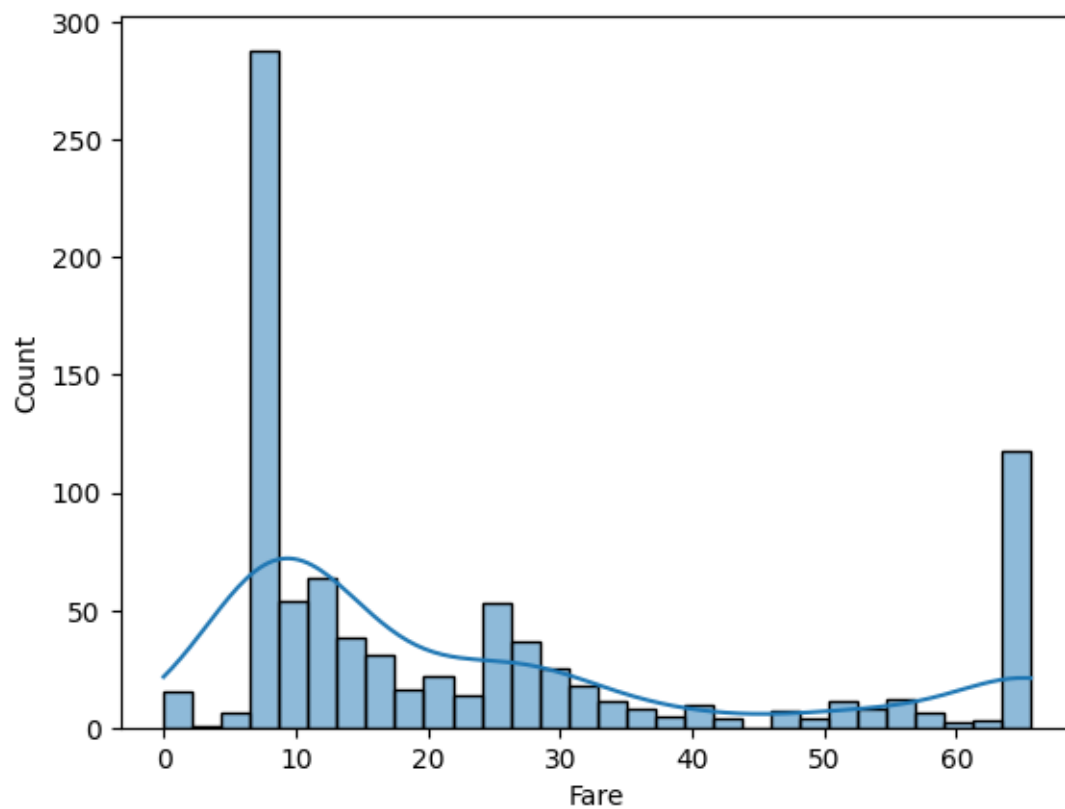
```
[47]: plt.figure(figsize=(10, 6))
```

```
[47]: <Figure size 1000x600 with 0 Axes>
```

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

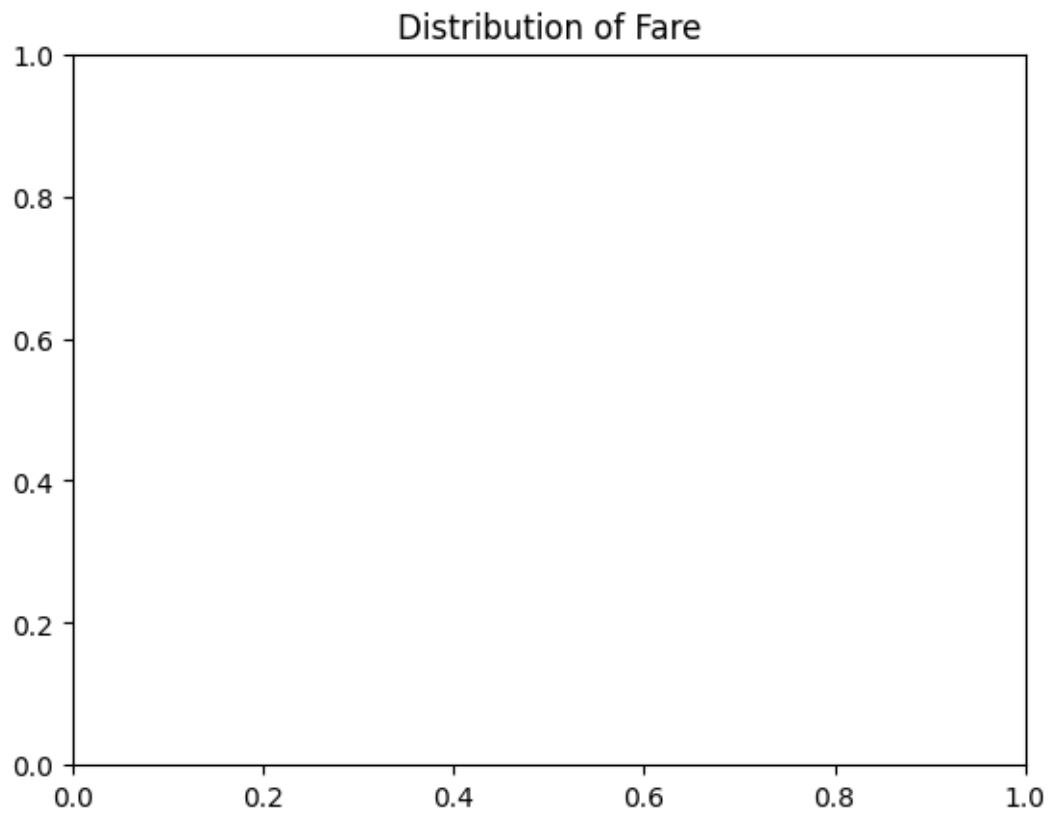
```
[48]: sns.histplot(df['Fare'], bins=30, kde=True)
```

```
[48]: <Axes: xlabel='Fare', ylabel='Count'>
```



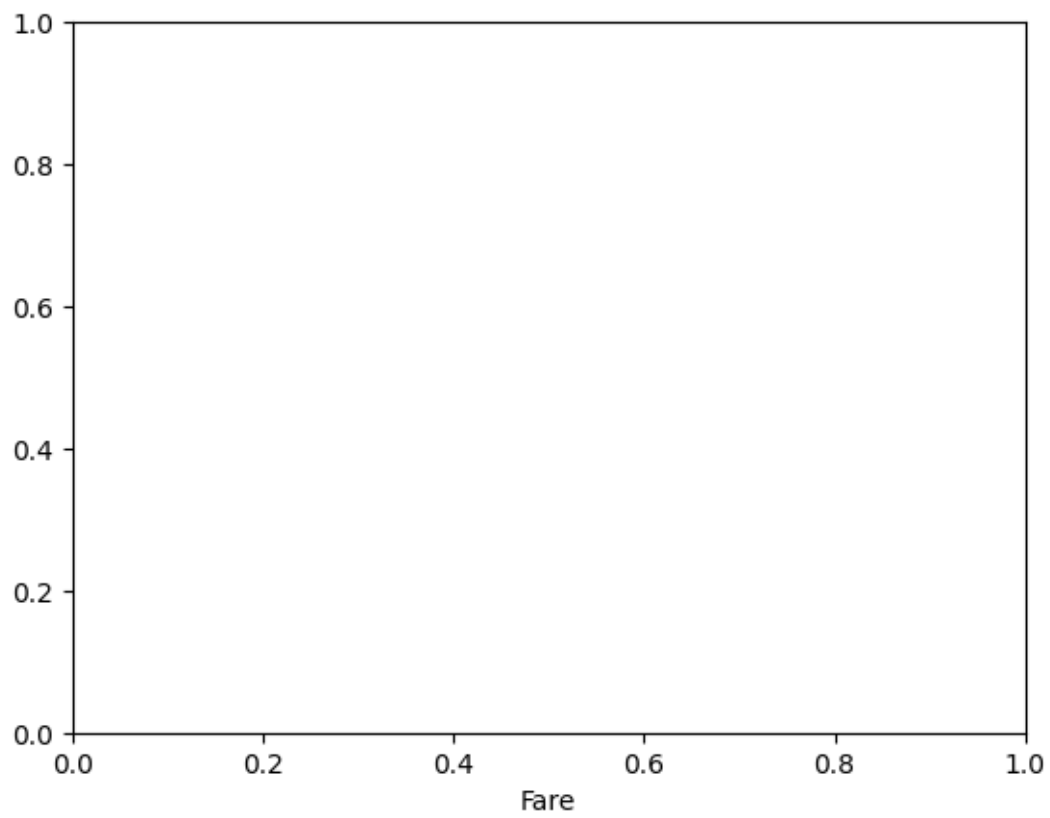
```
[49]: plt.title('Distribution of Fare')
```

```
[49]: Text(0.5, 1.0, 'Distribution of Fare')
```



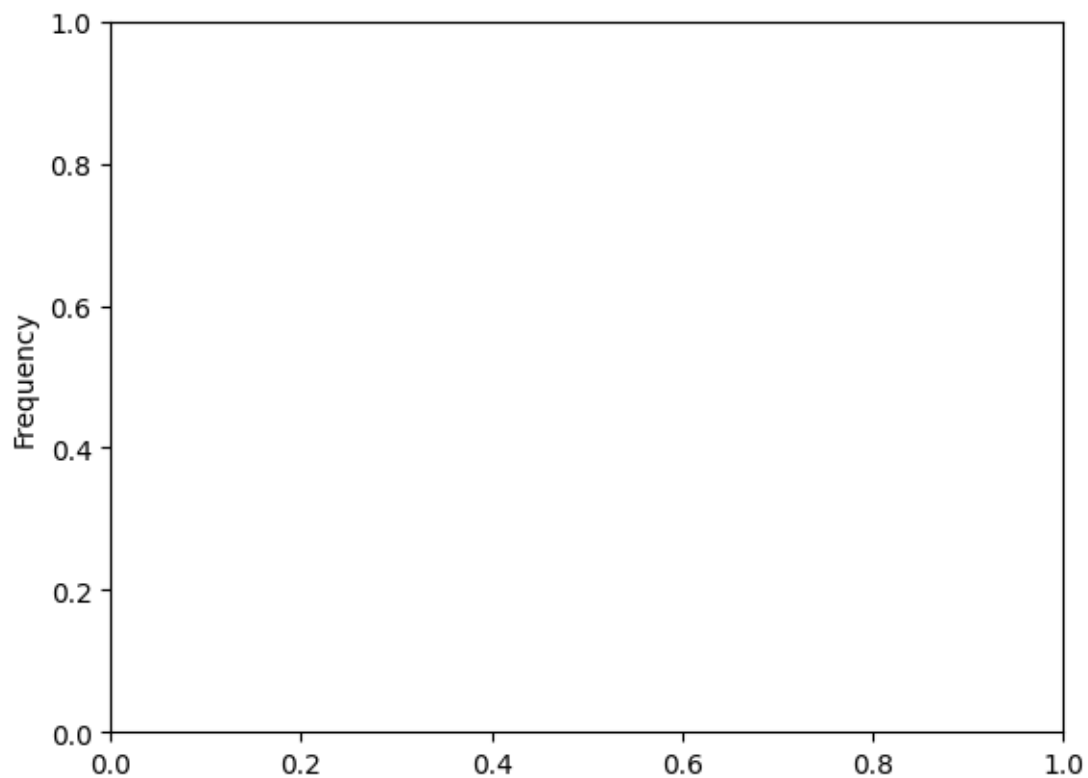
```
[50]: plt.xlabel('Fare')
```

```
[50]: Text(0.5, 0, 'Fare')
```



```
[51]: plt.ylabel('Frequency')
```

```
[51]: Text(0, 0.5, 'Frequency')
```



```
[52]: plt.show()
```

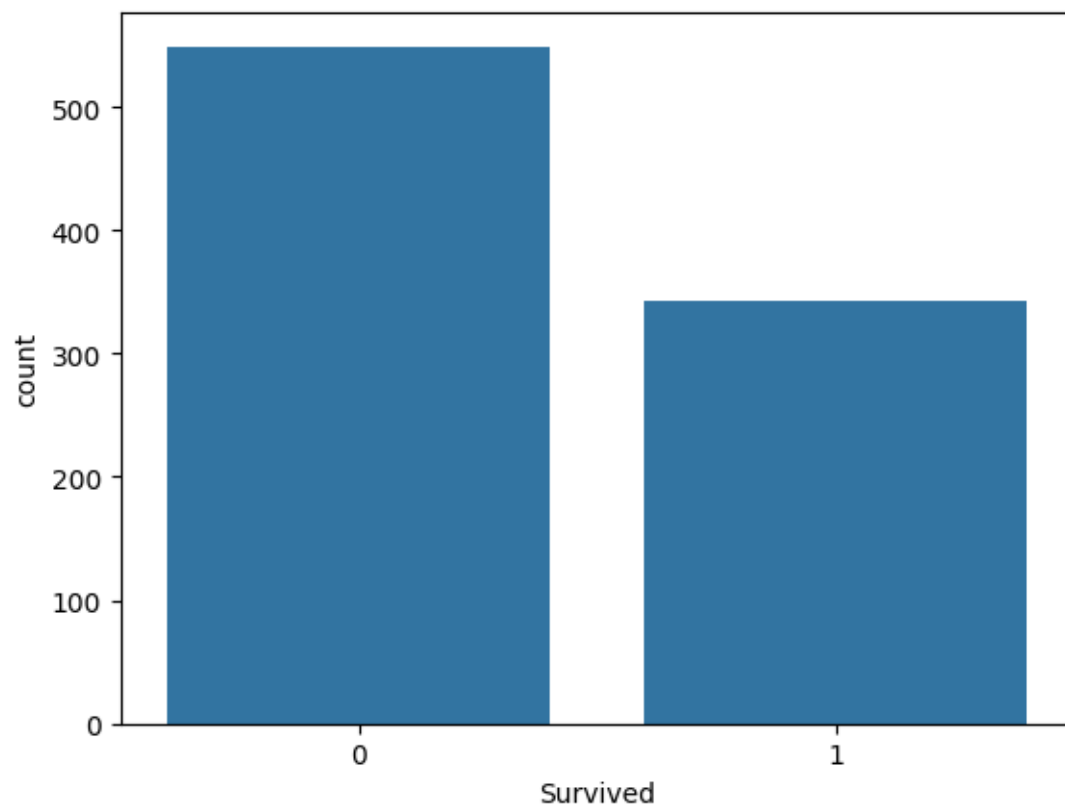
```
[53]: plt.figure(figsize=(6, 4))
```

```
[53]: <Figure size 600x400 with 0 Axes>
```

```
<Figure size 600x400 with 0 Axes>
```

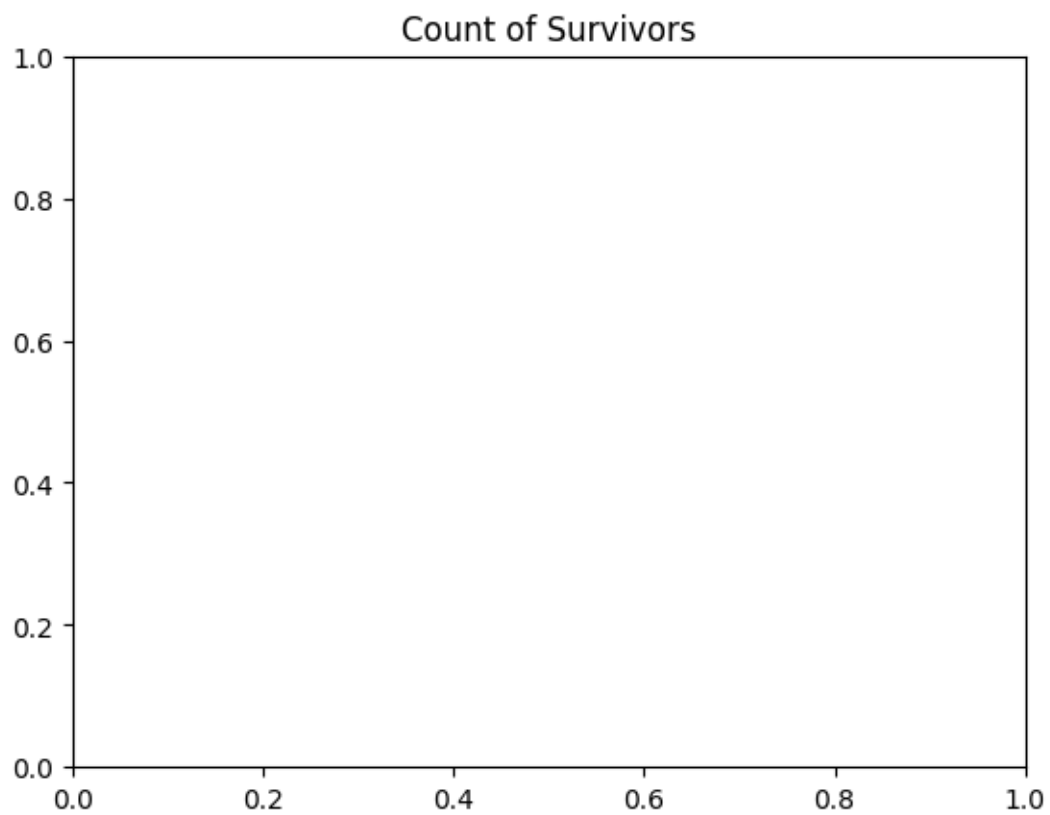
```
[54]: sns.countplot(x='Survived', data=df)
```

```
[54]: <Axes: xlabel='Survived', ylabel='count'>
```



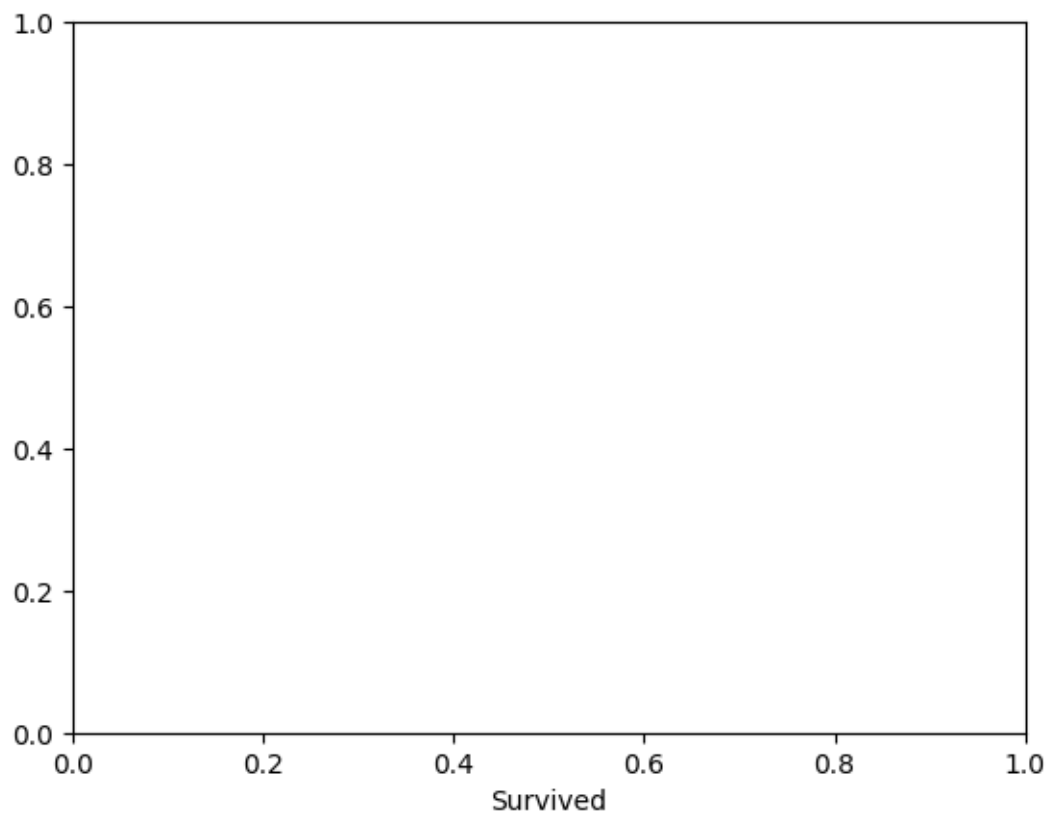
```
[55]: plt.title('Count of Survivors')
```

```
[55]: Text(0.5, 1.0, 'Count of Survivors')
```



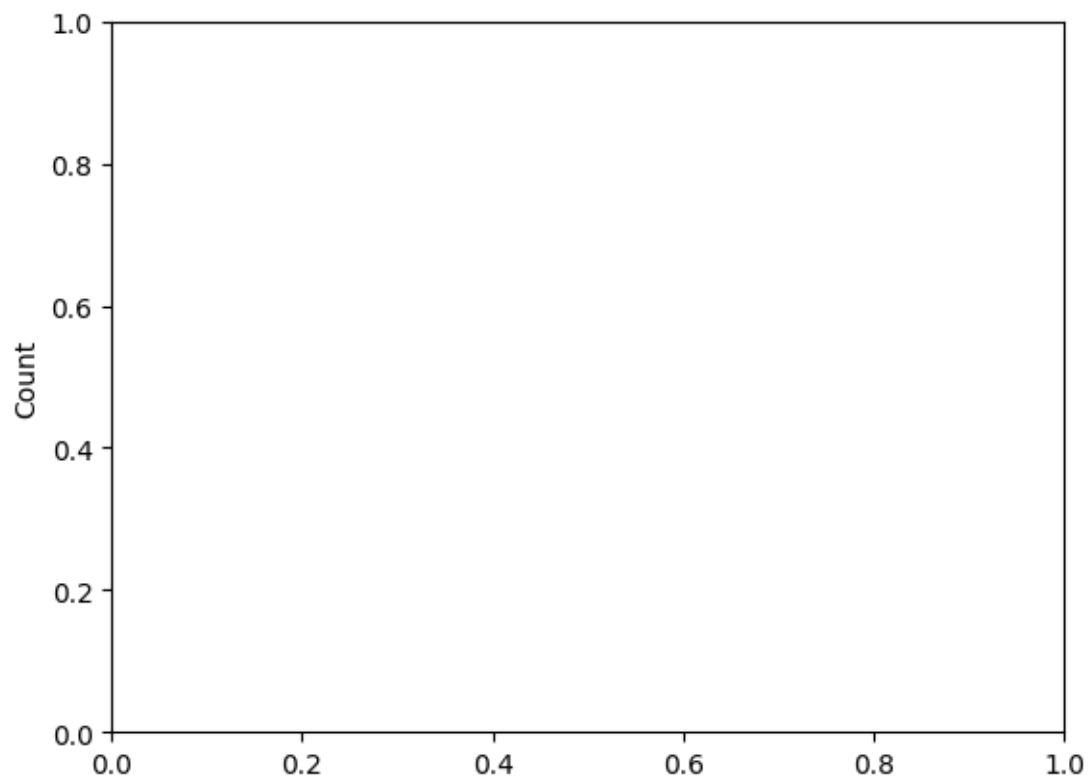
```
[56]: plt.xlabel('Survived')
```

```
[56]: Text(0.5, 0, 'Survived')
```



```
[57]: plt.ylabel('Count')
```

```
[57]: Text(0, 0.5, 'Count')
```



```
[58]: plt.show()
```

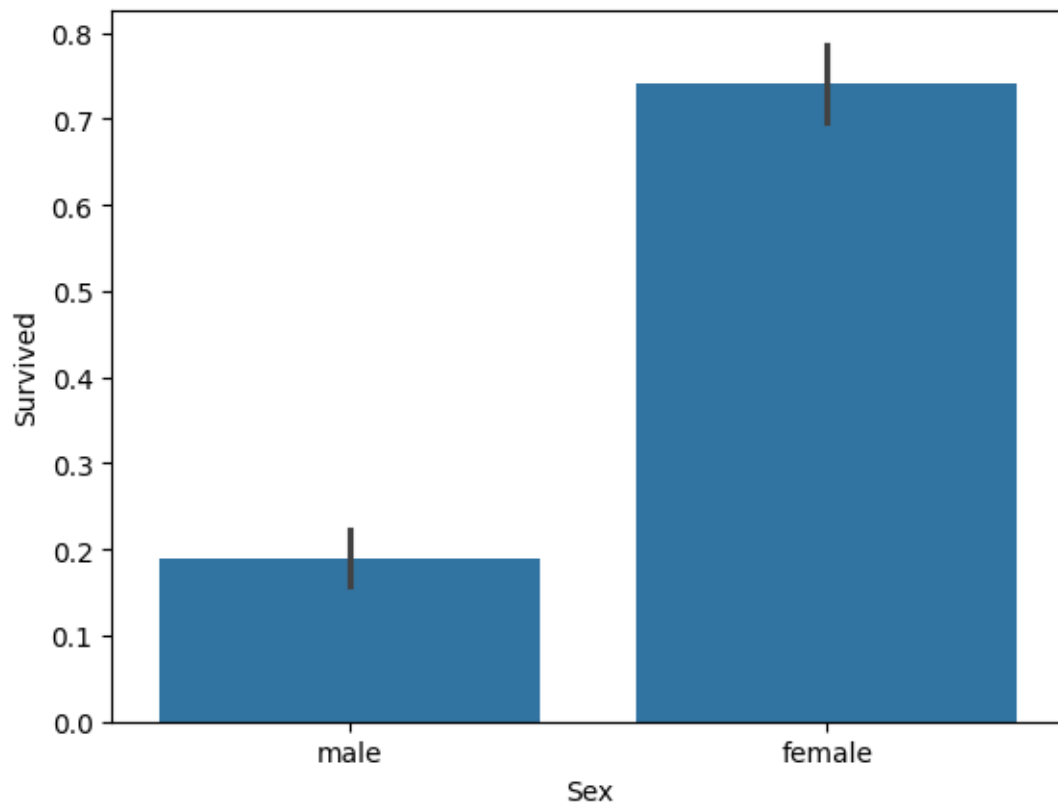
```
[59]: plt.figure(figsize=(6, 4))
```

```
[59]: <Figure size 600x400 with 0 Axes>
```

```
<Figure size 600x400 with 0 Axes>
```

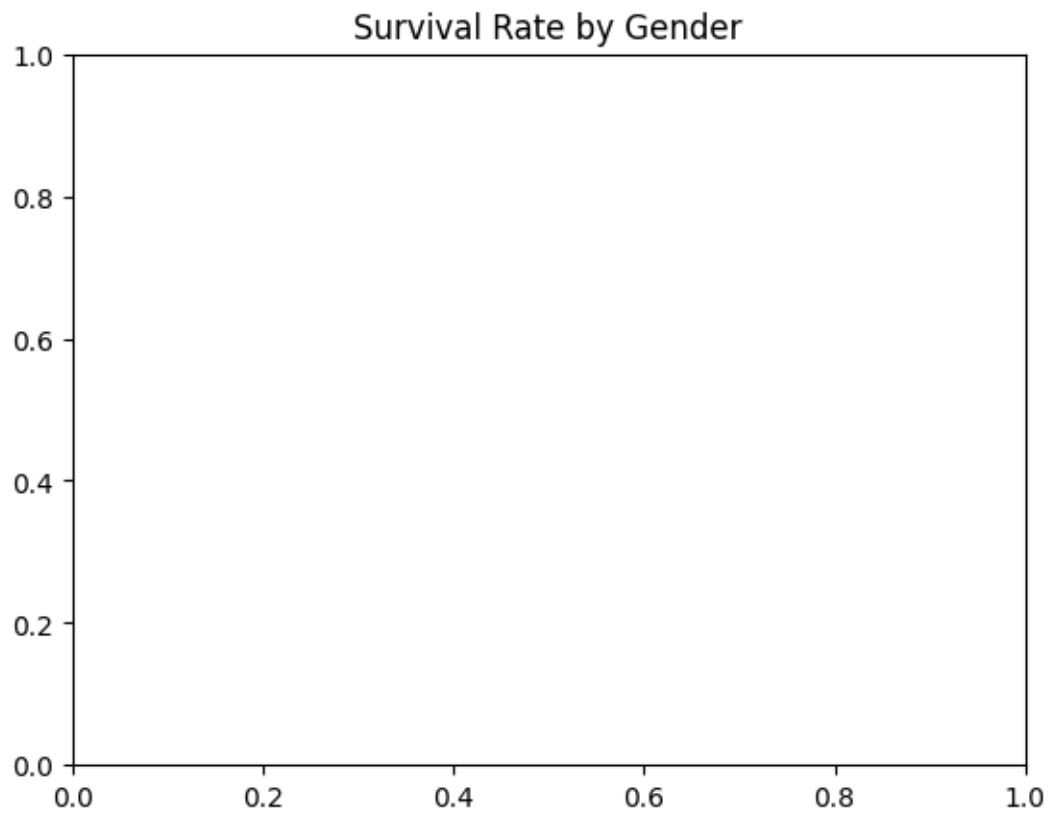
```
[60]: sns.barplot(x='Sex', y='Survived', data=df)
```

```
[60]: <Axes: xlabel='Sex', ylabel='Survived'>
```



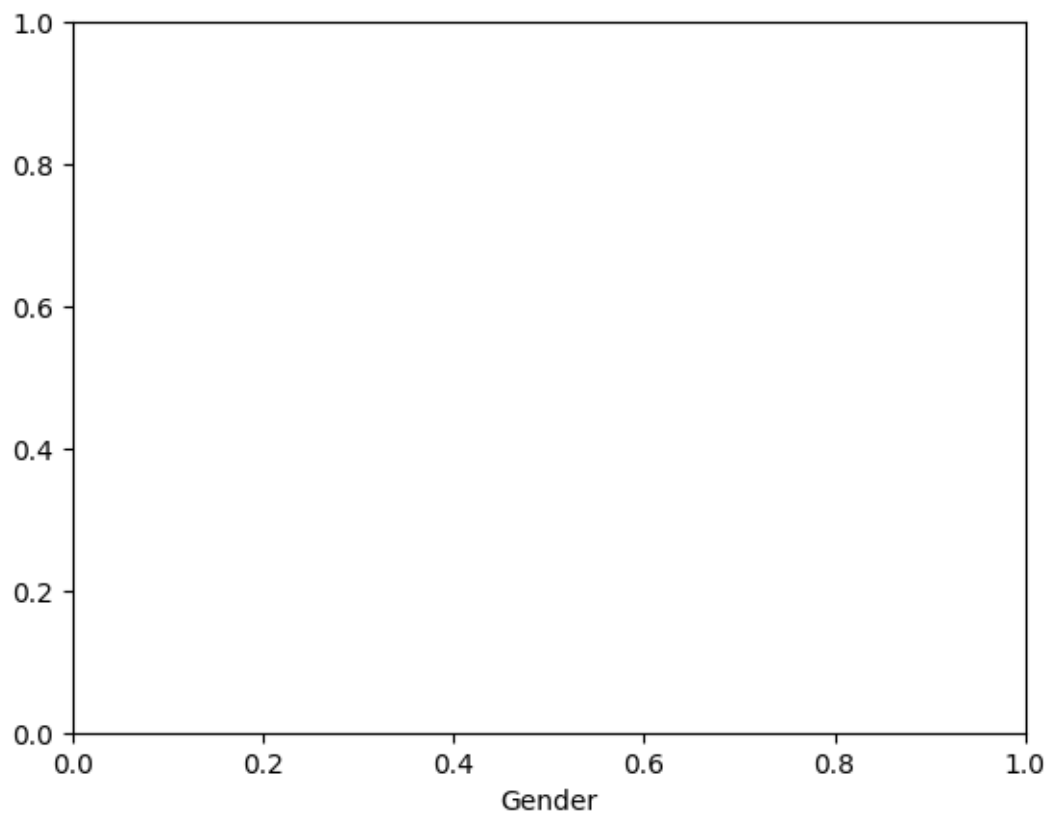
```
[61]: plt.title('Survival Rate by Gender')
```

```
[61]: Text(0.5, 1.0, 'Survival Rate by Gender')
```



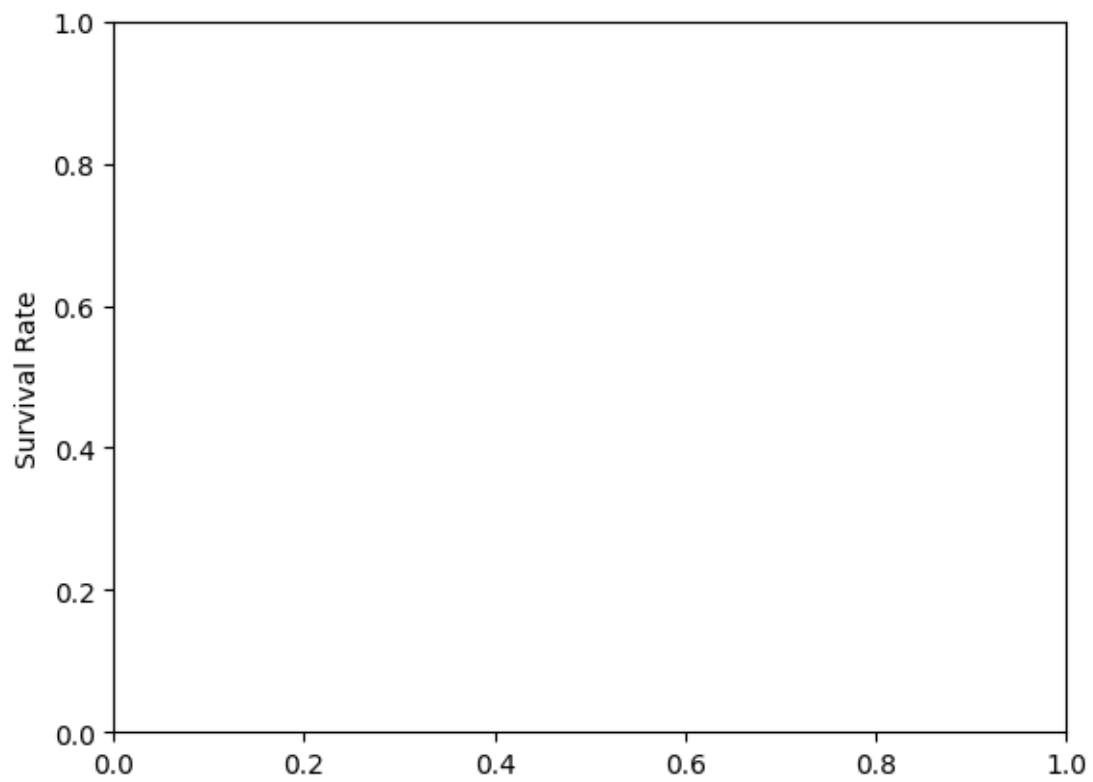
```
[62]: plt.xlabel('Gender')
```

```
[62]: Text(0.5, 0, 'Gender')
```



```
[63]: plt.ylabel('Survival Rate')
```

```
[63]: Text(0, 0.5, 'Survival Rate')
```



```
[64]: plt.show()
```

```
[65]: df.to_csv('Titanic-Cleaned-Dataset.csv', index=False)
```

```
[ ]:
```