```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

Data Loading

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
d=pd.read_csv("/content/drive/MyDrive/Colab Notebooks/train.csv")
```

Data cleaning

Checking for missing values:

```
missing_values=d.isnull().sum()
print(missing_values)
```

```
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
```

Handling missing values:

```
d['Age'].fillna(d['Age'].median(), inplace=True)
d['Embarked'].fillna(d['Embarked'].mode()[0], inplace=True)
d['Cabin'].fillna('Unknown', inplace=True)
print("Missing values are handled")
```

```
Missing values are handled
```

Exploratory Data Analysis (EDA)

Summary statistics:

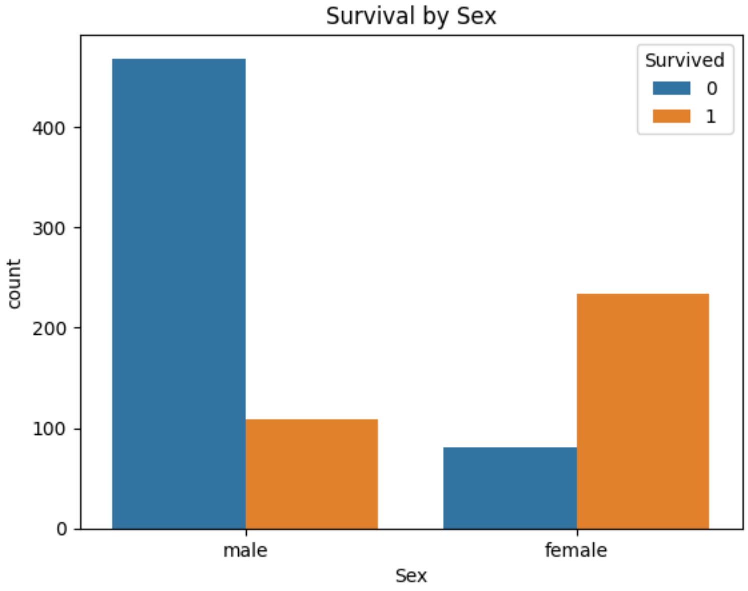
```
print(d.describe())
```

```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  891.000000  891.000000
mean    446.000000    0.383838    2.308642   29.361582    0.523008
std     257.353842    0.486592    0.836071   13.019697    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   22.000000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   35.000000    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
```

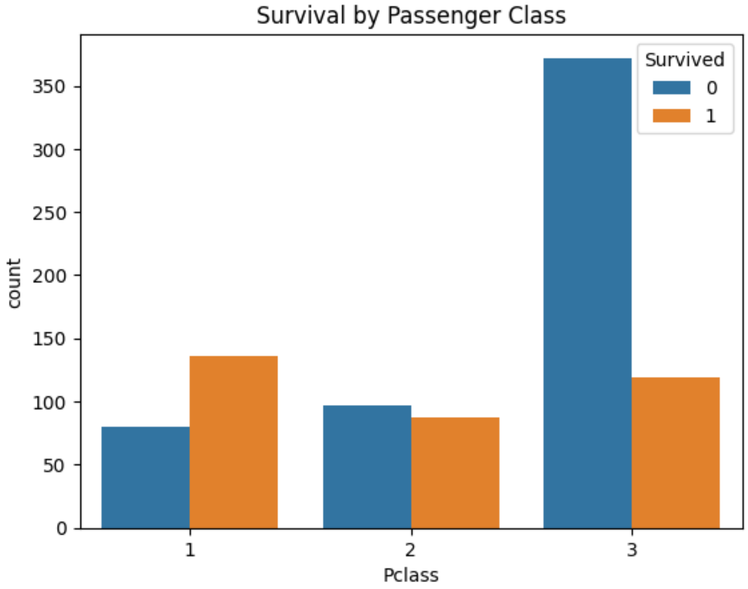
Exploring relationships:

```
sns.countplot(x='Sex', hue='Survived', data=d)
plt.title('Survival by Sex')
plt.show()
```



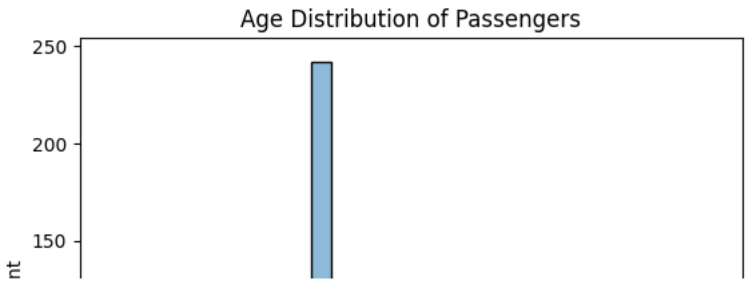
Survival by Pclass (Passenger Class):

```
sns.countplot(x='Pclass', hue='Survived', data=d)
plt.title('Survival by Passenger Class')
plt.show()
```



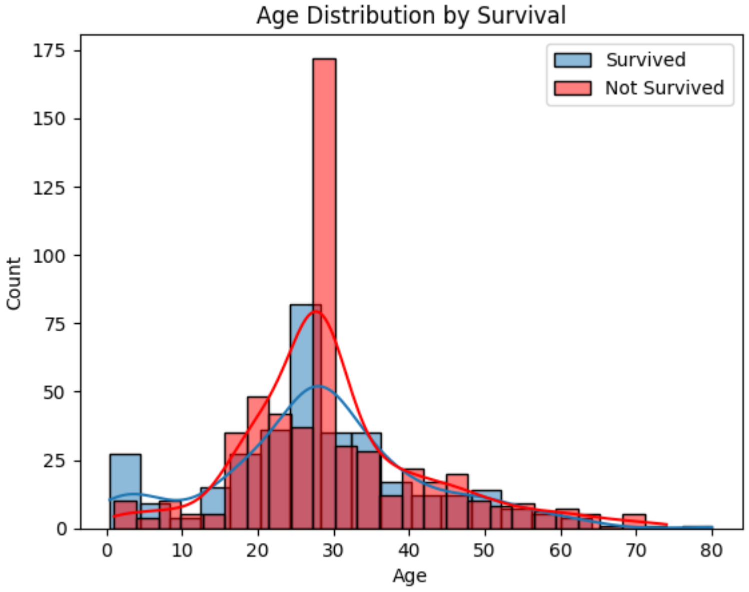
Age distribution of passengers: New Section

```
sns.histplot(d['Age'], kde=True)
plt.title('Age Distribution of Passengers')
plt.show()
```



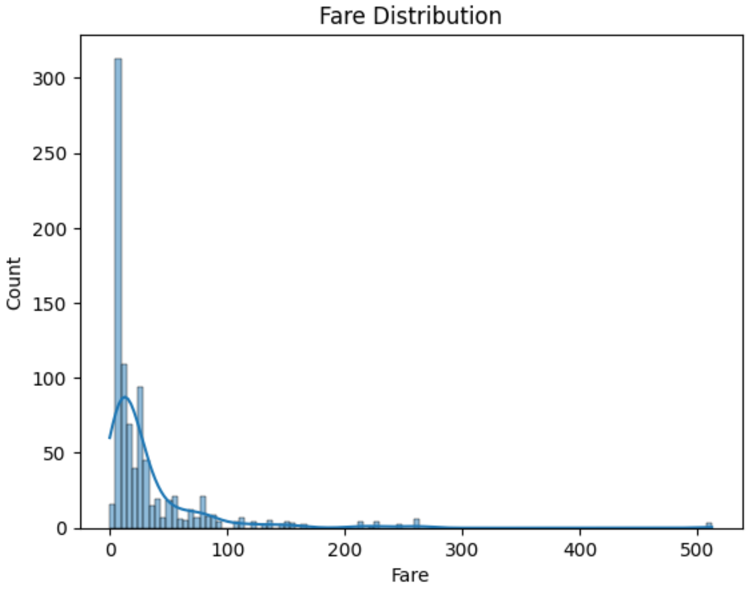
Age distribution of passengers by Survival:

```
sns.histplot(d[d['Survived'] == 1]['Age'], kde=True, label='Survived')
sns.histplot(d[d['Survived'] == 0]['Age'], kde=True, label='Not Survived', color='red')
plt.title('Age Distribution by Survival')
plt.legend()
plt.show()
```



Fare distribution:

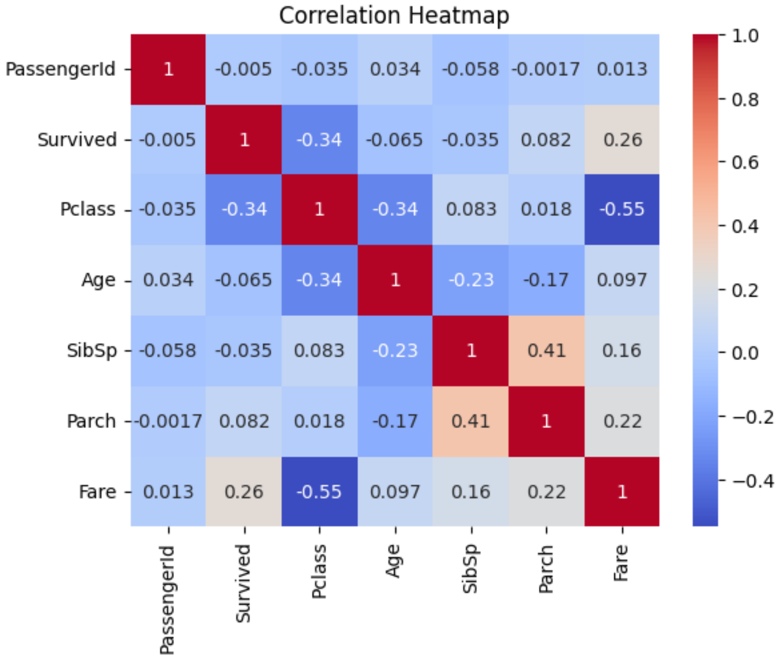
```
sns.histplot(d['Fare'], kde=True)
plt.title('Fare Distribution')
plt.show()
```



Correlations

```
# Heatmap to visualize correlations
correlation_matrix = d.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

```
<ipython-input-20-e3a265499a9b>:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future ver
  correlation_matrix = d.corr()
```


Correlation Heatmap
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare