```
!pip install pandas matplotlib seaborn
```

```
Defaulting to user installation because normal site-packages is not
writeable
Requirement already satisfied: pandas in c:\programdata\anaconda3\lib\
site-packages (2.2.3)
Requirement already satisfied: matplotlib in c:\programdata\anaconda3\
lib\site-packages (3.10.0)
Requirement already satisfied: seaborn in c:\programdata\anaconda3\
lib\site-packages (0.13.2)
Requirement already satisfied: numpy>=1.26.0 in c:\programdata\
anaconda3\lib\site-packages (from pandas) (2.1.3)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\
programdata\anaconda3\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\programdata\
anaconda3\lib\site-packages (from pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in c:\programdata\
anaconda3\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\programdata\
anaconda3\lib\site-packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in c:\programdata\
anaconda3\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\programdata\
anaconda3\lib\site-packages (from matplotlib) (4.55.3)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\programdata\
anaconda3\lib\site-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in c:\programdata\
anaconda3\lib\site-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in c:\programdata\anaconda3\
lib\site-packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\programdata\
anaconda3\lib\site-packages (from matplotlib) (3.2.0)
Requirement already satisfied: six>=1.5 in c:\programdata\anaconda3\
lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
```

```
df = pd.read_csv(r"C:\Users\munth\OneDrive\Desktop\train.csv")
```

```
print(os.path.isfile('train.csv'))
```

```
False
```

```
import os
print(os.getcwd())
```

```
C:\Users\munth\Desktop\hanushpython
```

```
print(os.path.isfile('train.csv'))
```

```
True
```

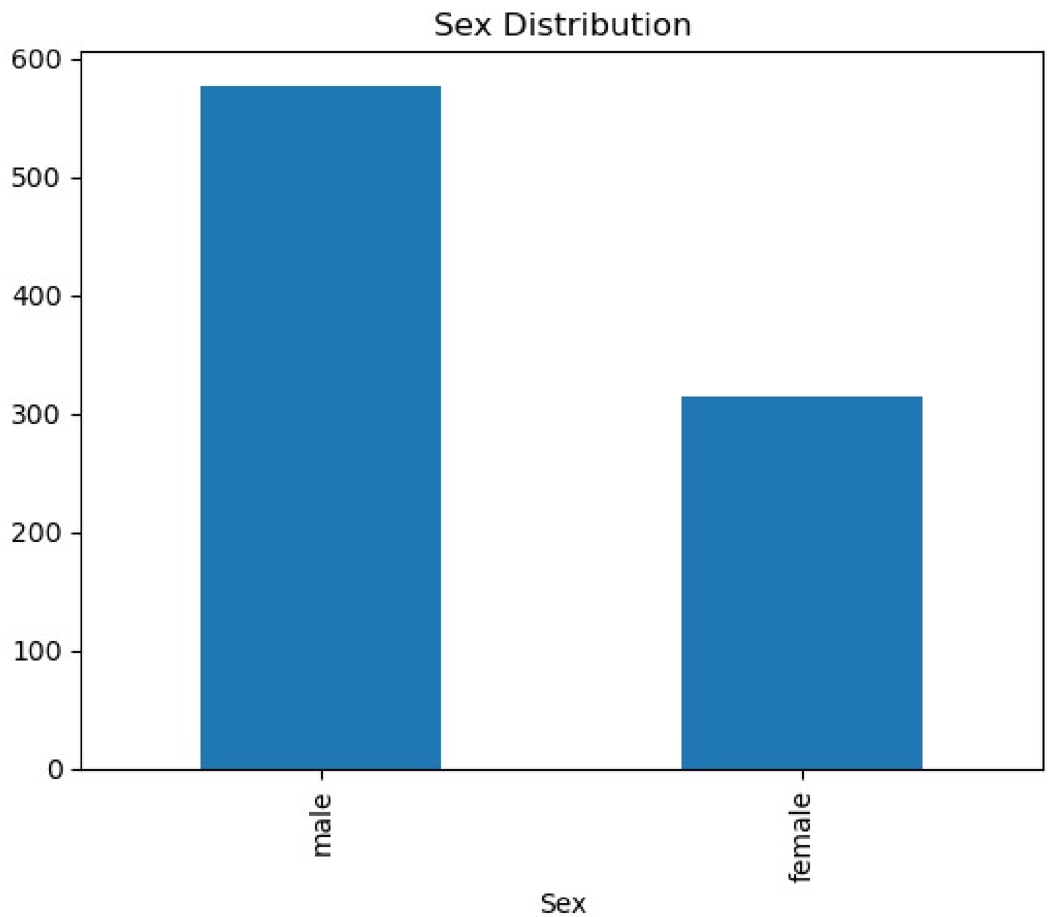
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load data
df = pd.read_csv("train.csv")

df.head()
df.info()
df.describe()
df.isnull().sum()
```

```
<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
```

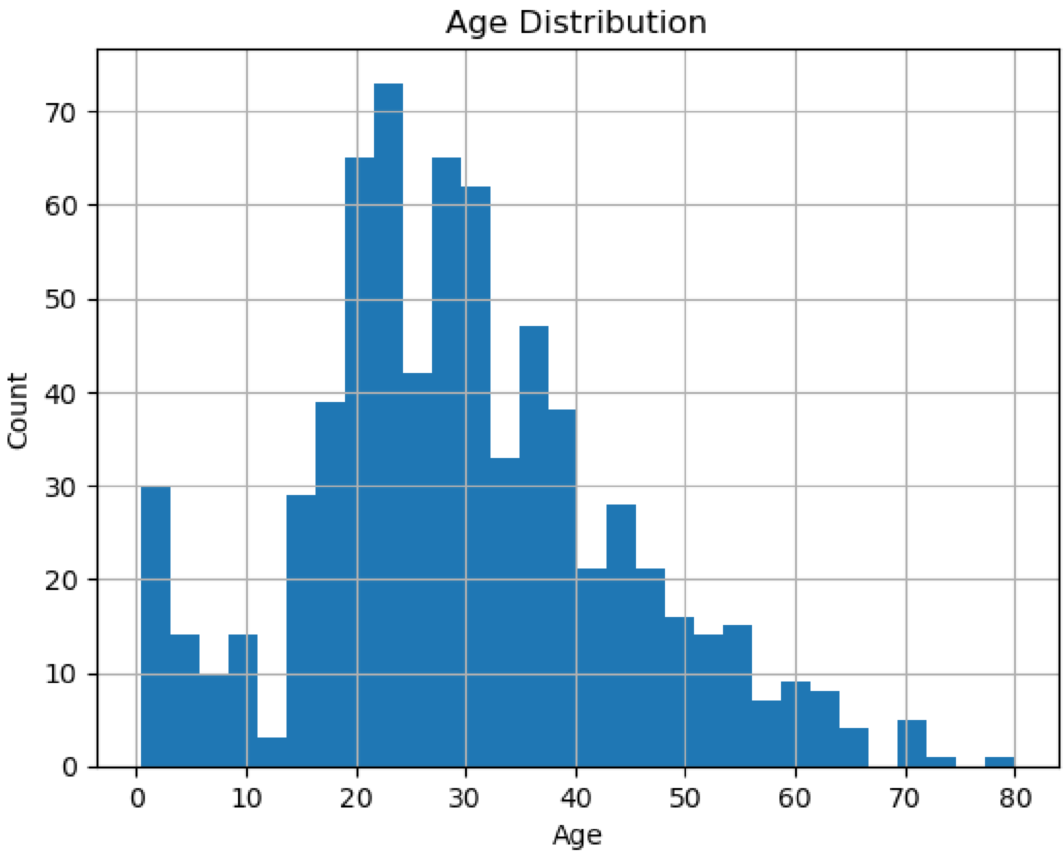
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

```python
#Observations and Notes:
# 891 passengers, 12 columns with numeric, categorical, and object
types.
```

```python
# Missing data mainly in Age (177 missing) and Cabin (687 missing).
# Most other columns are complete, ready for further analysis

df['Sex'].value_counts().plot(kind='bar')
plt.title('Sex Distribution')
plt.show()
```

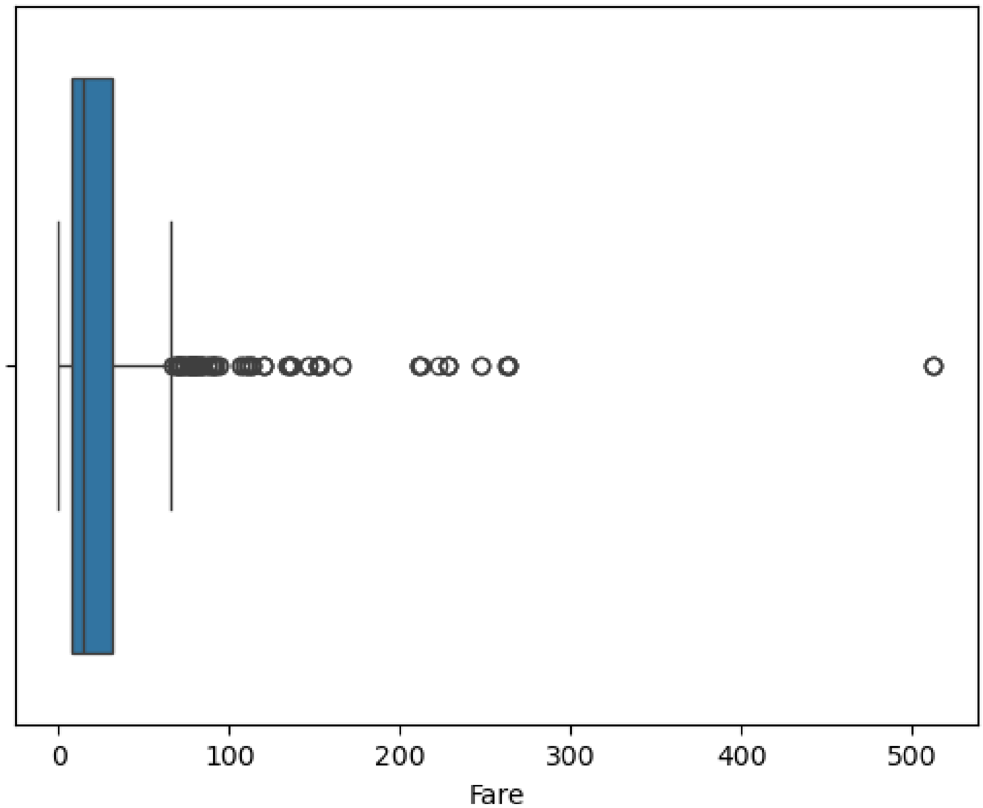



```python
#Observations and Notes:
#There are significantly more male passengers than female passengers
in the dataset.
#The sex distribution reveals a clear gender imbalance, with males
being the majority group.
#This imbalance may impact survival statistics and further analysis on
outcomes by gender.

df['Age'].hist(bins=30)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```

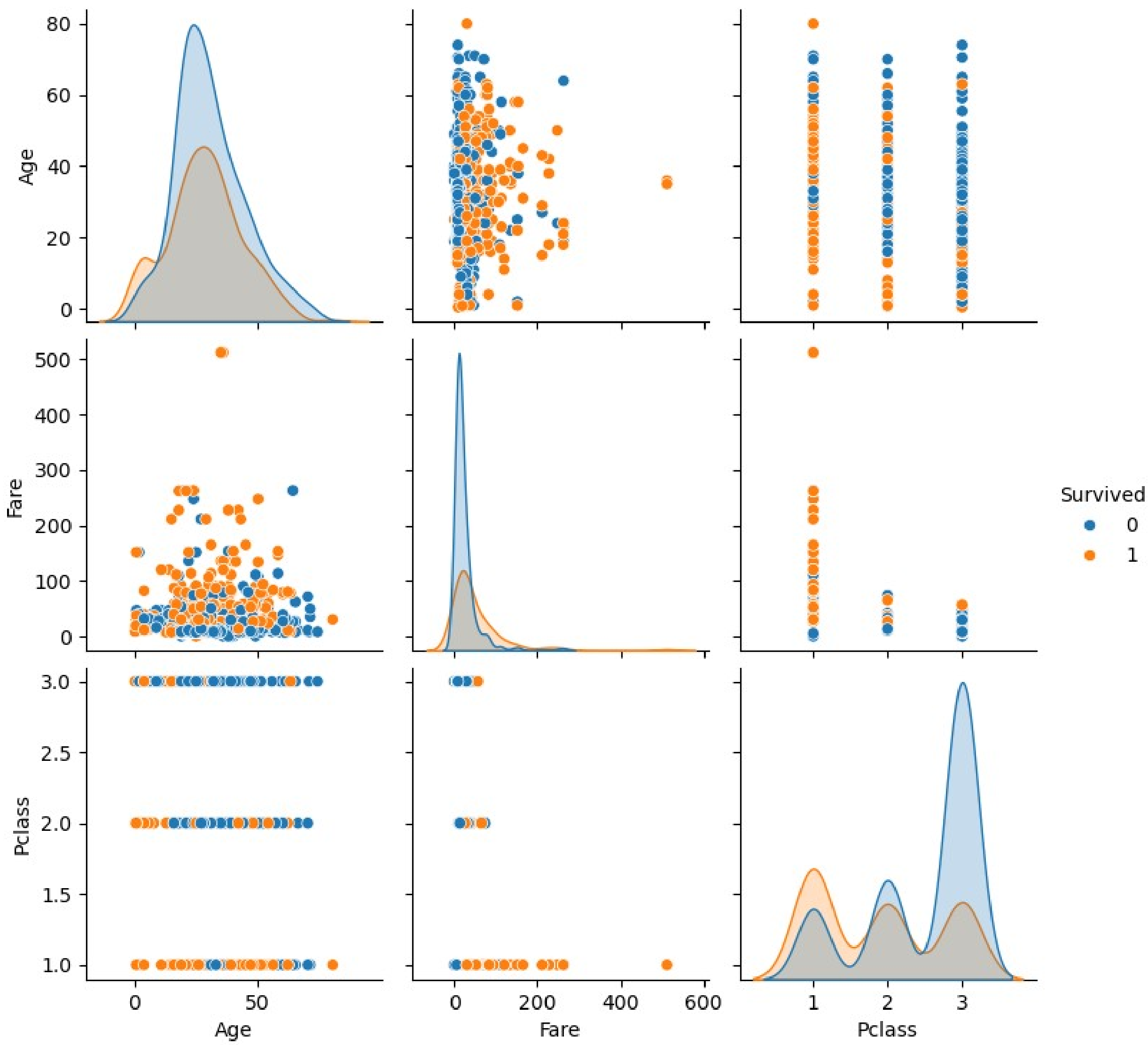



```
#Observations and Notes:
#Most passengers are between 20 and 40 years old, with a noticeable peak in this range.
#There are very few elderly passengers (above 70) and children (below 10).
#The distribution is right-skewed, with more younger adults than older individuals.

sns.boxplot(x=df['Fare'])
```

```
<Axes: xlabel='Fare'>
```

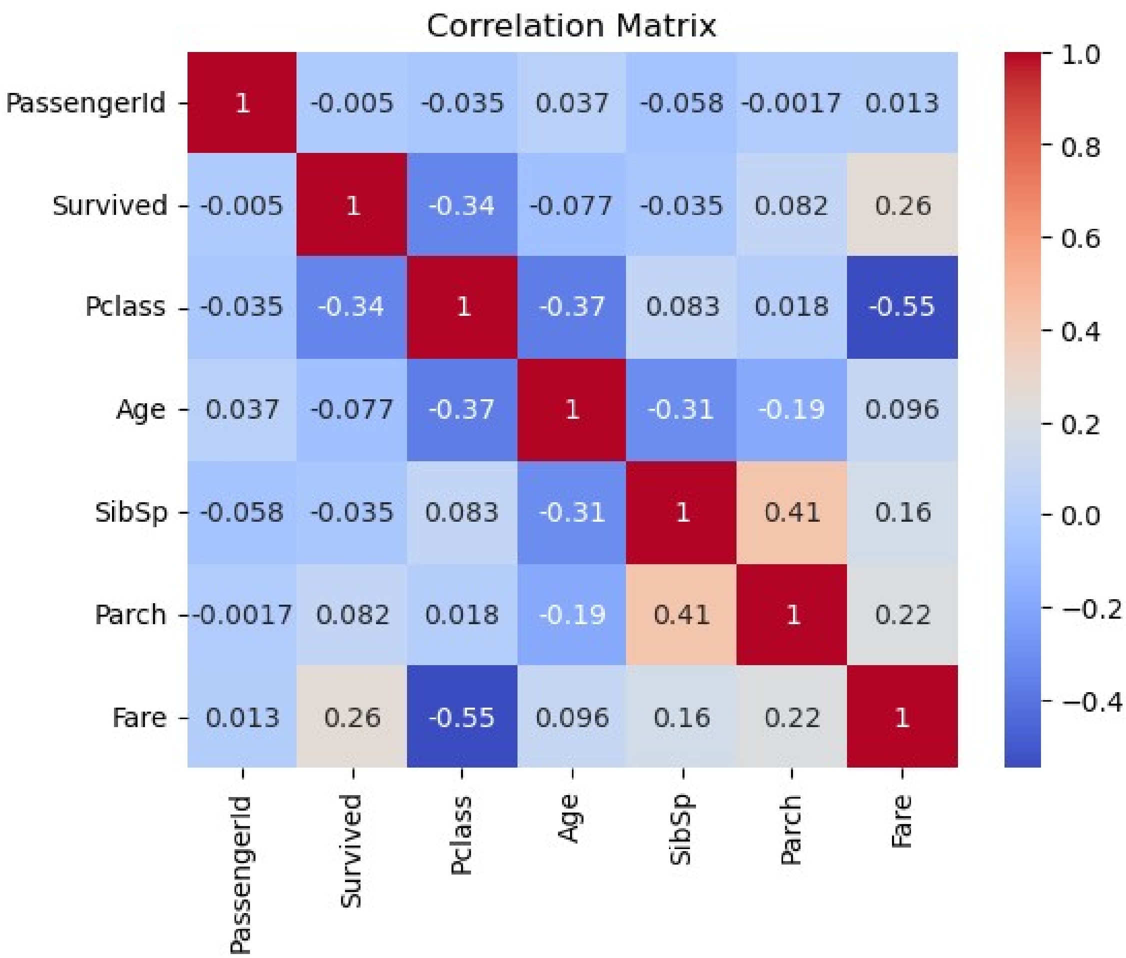



```
#Observations and Notes:
#Most fares are clustered at lower values, with only a few passengers
paying very high fares.
#There are several outliers, including some extremely high fare
amounts.
#The fare distribution is highly skewed to the right, indicating the
presence of a few expensive tickets.

sns.pairplot(df[['Survived', 'Age', 'Fare', 'Pclass']].dropna(),
hue='Survived')
plt.show()
```

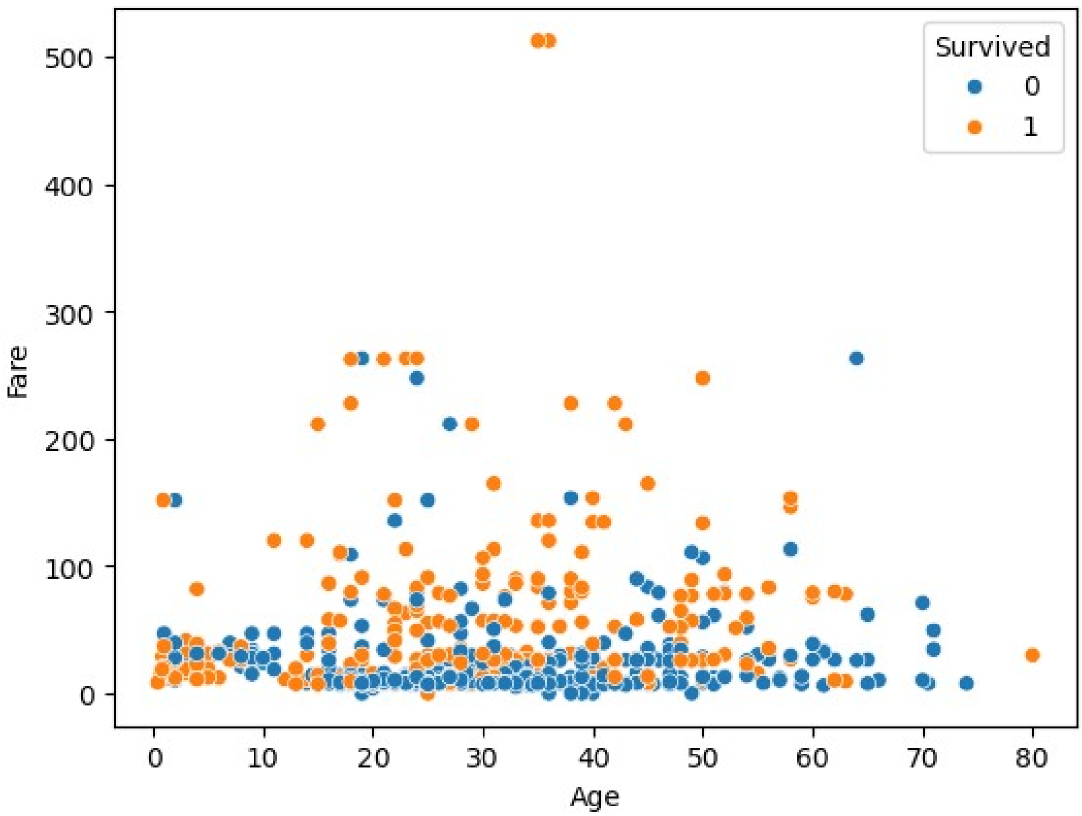



```
#Observations and Notes:
#Survivors tend to belong to higher classes (Pclass 1) and have paid
higher fares.
#Most non-survivors are concentrated in lower classes and paid lower
fares.
#Age distributions do not show a strong survival difference, but some
younger and higher-fare passengers had better survival rates.-
Survivors are more common among higher fare amounts and in higher
classes.
#Non-survivors cluster in lower Pclass and lower fare ranges.
#There is no strong visual separation by age, but younger passengers
appear across both survival groups.

numeric_df = df.select_dtypes(include=['number'])
corr_matrix = numeric_df.corr()
```

```
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm")
plt.title('Correlation Matrix')
plt.show()
```




```
#Observations and Notes:
#Survival has a negative correlation with Pclass, meaning higher
classes had better survival rates.
#Fare and Pclass are strongly negatively correlated, indicating higher
fares are associated with higher classes.
#Age and other variables show weak correlations with survival.

sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.show()
```




```
#Observations and Notes:
#Survivors (orange) appear more often among higher fares, regardless
of age.
#Most passengers paid low fares, and survival does not show a strong
trend with age.
#High-fare passengers are spread across different ages, but survival
is more likely for them.

#Summarize Key Insights:
#Males outnumber females in the dataset, but females have a higher
survival rate, especially in higher classes.
#Most passengers are aged between 20 and 40, with children and elderly
being fewer; younger passengers had slightly better survival chances.
#Fare paid is highly skewed, with most paying low fares and a few
outliers paying very high amounts; higher fares correlate with higher
survival.
#Passenger class strongly influences survival; first-class passengers
had the highest survival rates, followed by second and third classes.
#Correlation and scatterplots confirm survival is positively
associated with fare and negatively correlated with passenger class
number (lower class number = higher class).
```