```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
import pandas as pd
data= pd.read_csv('/content/drive/MyDrive/train.csv')
data.head()
```

	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

Next steps: View recommended plots

```
data.tail()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

```
data.index
```

```
RangeIndex(start=0, stop=891, step=1)
```

```
data.shape
```

```
(891, 12)
```

```
data.size
```

```
10692
```

```
data.columns
```

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

```
data.describe()
```

	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

```
data.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
```

```
# Cleaning Data
missing_data = data.isnull().sum()
missing_data
```

```
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
```

```
data = data.drop_duplicates()
data.head(8)
```

	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
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthv. Mr. Timothv J	male	54.0	0	0	17463	51.8625	E46	S

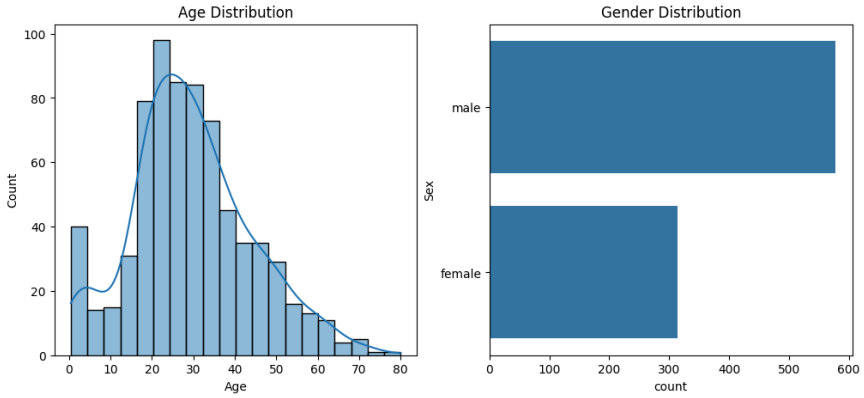
```
data['Age'] = pd.to_numeric(data['Age'], errors='coerce')
data['Fare'] = pd.to_numeric(data['Fare'], errors='coerce')
```

```
# Exploratory Data Analysis (EDA)
summary_stats = data.describe()
print(summary_stats)
```

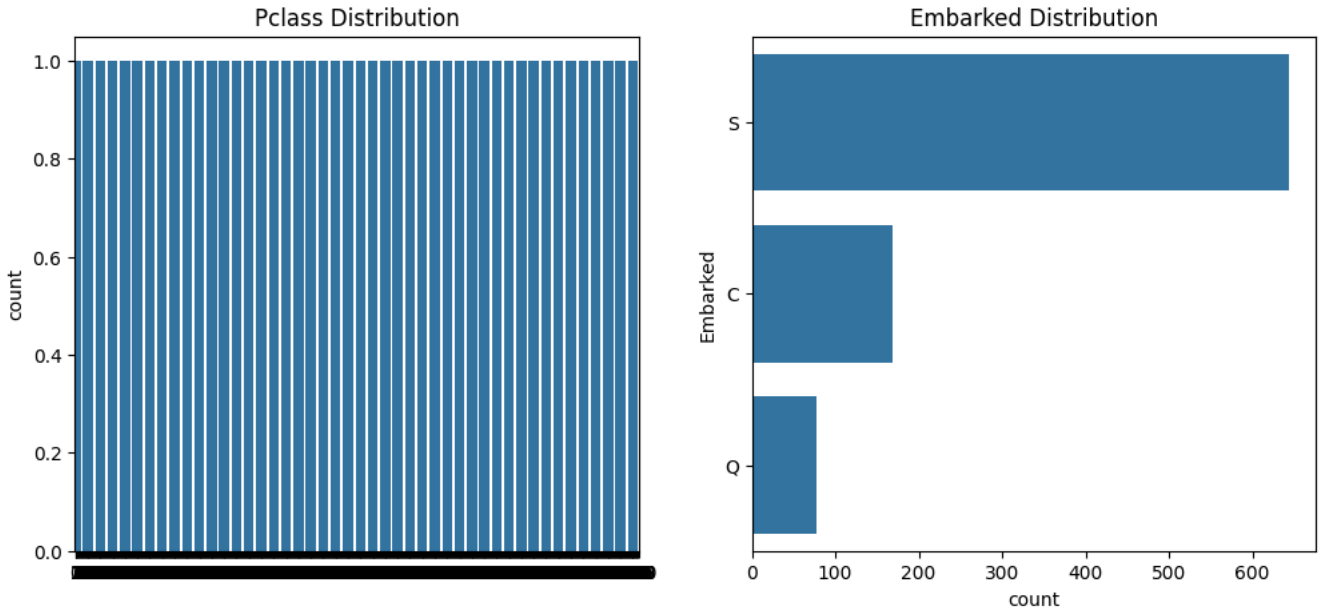
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    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
```

```
# Univariate Analysis:
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.histplot(data['Age'].dropna(), kde=True)
plt.title('Age Distribution')
plt.subplot(1, 2, 2)
sns.countplot(data['Sex'])
plt.title('Gender Distribution')
plt.show()
```

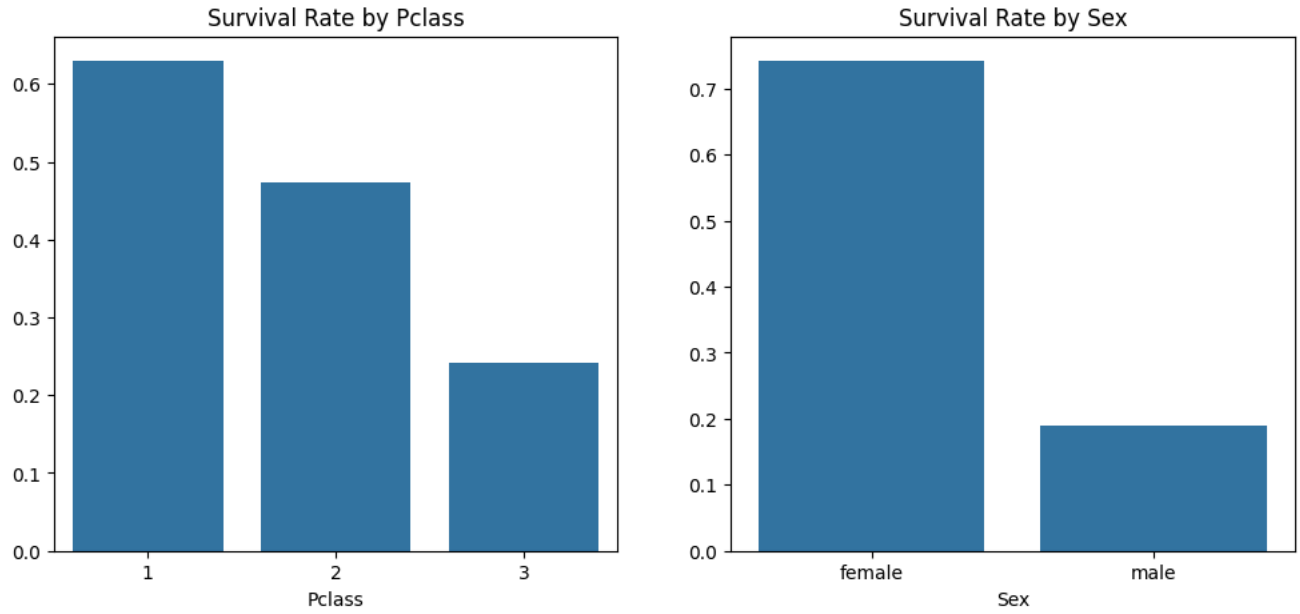



```
# Categorical Variables:
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.countplot(data['Pclass'])
plt.title('Pclass Distribution')
plt.subplot(1, 2, 2)
sns.countplot(data['Embarked'])
plt.title('Embarked Distribution')
plt.show()
```

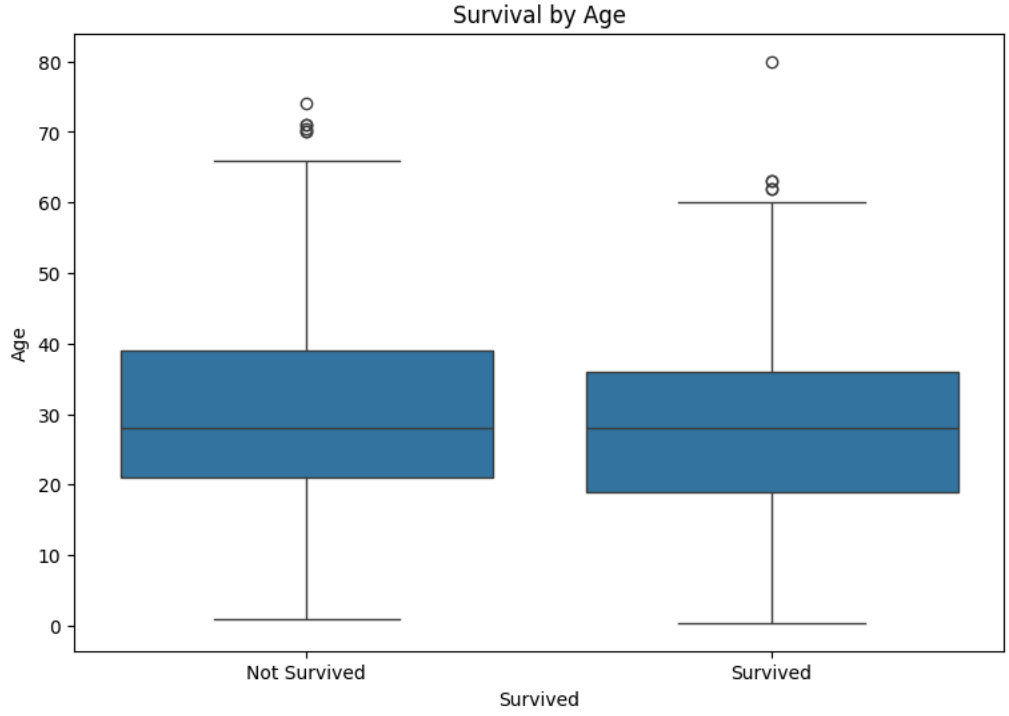


```
# Survival Analysis:
survival_rate_by_class = data.groupby('Pclass')['Survived'].mean()
survival_rate_by_sex = data.groupby('Sex')['Survived'].mean()

plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.barplot(x=survival_rate_by_class.index, y=survival_rate_by_class.values)
plt.title('Survival Rate by Pclass')
plt.subplot(1, 2, 2)
sns.barplot(x=survival_rate_by_sex.index, y=survival_rate_by_sex.values)
plt.title('Survival Rate by Sex')
plt.show()
```

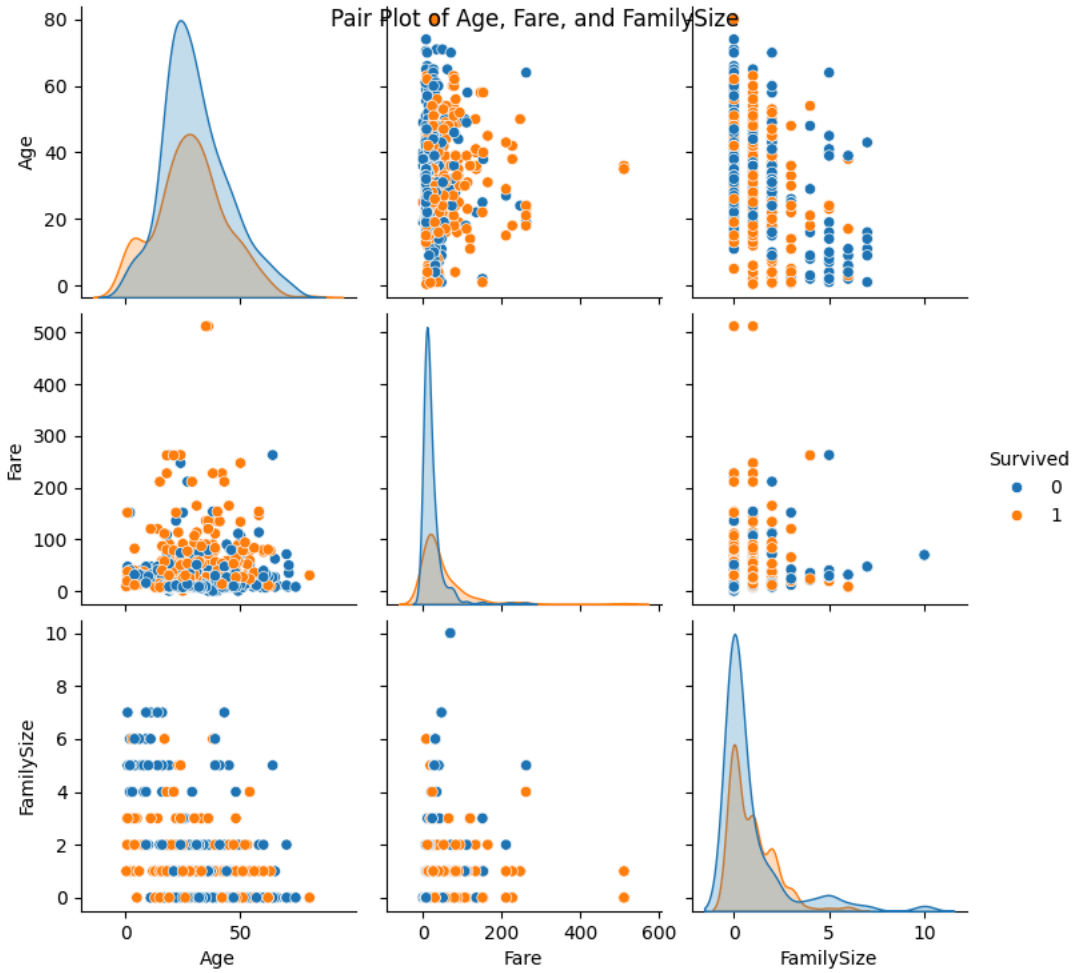



```
# Visualizing Survival by Age:
plt.figure(figsize=(9, 6))
sns.boxplot(x='Survived', y='Age', data=data)
plt.title('Survival by Age')
plt.xticks([0, 1], ['Not Survived', 'Survived'])
plt.show()
```

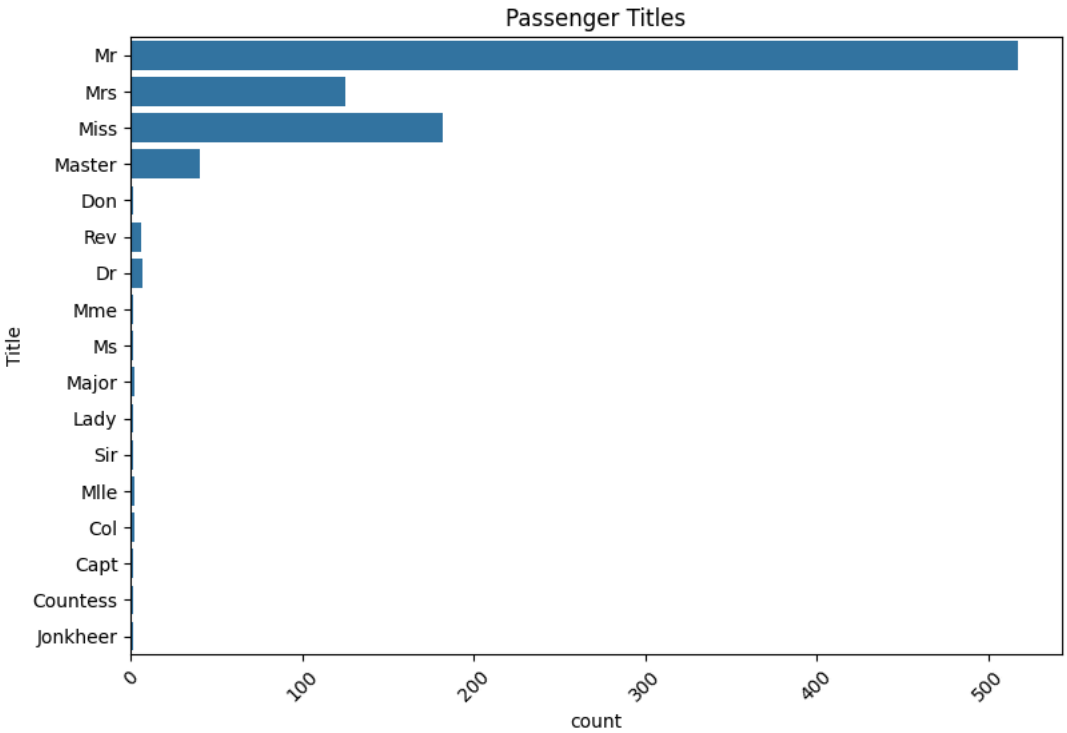



```
data['FamilySize'] = data['SibSp'] + data['Parch']
```

```
sns.pairplot(data, vars=['Age', 'Fare', 'FamilySize'], hue='Survived')
plt.suptitle('Pair Plot of Age, Fare, and FamilySize')
plt.show()
```




```
plt.figure(figsize=(9, 6))
data['Title'] = data['Name'].str.extract(' ([A-Za-z]+)\.')
sns.countplot(data['Title'])
plt.title('Passenger Titles')
plt.xticks(rotation=45)
plt.show()
```



```
# Survival by Title:
title_survival_rate = data.groupby('Title')['Survived'].mean()
plt.figure(figsize=(12, 5))
sns.barplot(x=title_survival_rate.index, y=title_survival_rate.values)
plt.title('Survival Rate by Title')
plt.xticks(rotation=45)
plt.show()
```