

```
from google.colab import files
uploaded = files.upload()
```

Choose Files 3 files
gender_submission.csv(text/csv) - 3258 bytes, last modified: 9/29/2025 - 100% done
test.csv(text/csv) - 28629 bytes, last modified: 9/29/2025 - 100% done
train.csv(text/csv) - 61194 bytes, last modified: 9/29/2025 - 100% done
Saving gender_submission.csv to gender_submission (1).csv
Saving test.csv to test (1).csv
Saving train.csv to train (1).csv

```
import pandas as pd

train = pd.read_csv("train.csv")
test = pd.read_csv("test.csv")
gender_submission = pd.read_csv("gender_submission.csv")

train.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
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Next steps: [Generate code with train](#) [New interactive sheet](#)

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

sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (10,6)
```

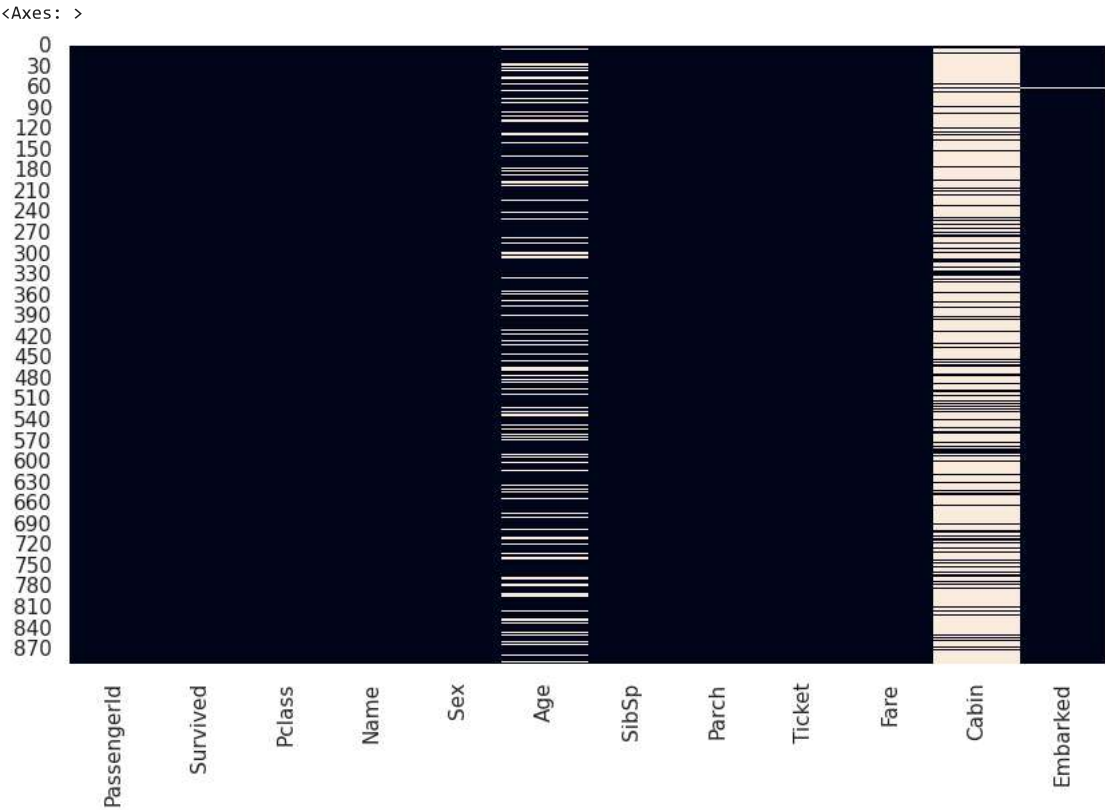
```
train.info()
train.describe(include='all').T
```



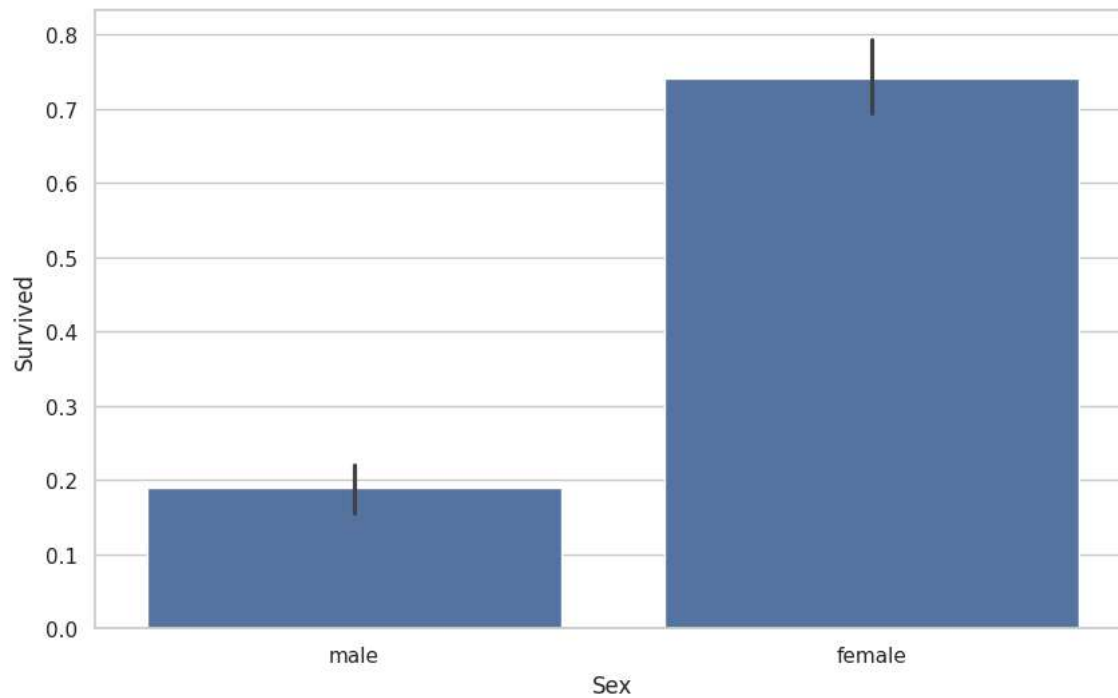
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   PassengerId  891 non-null    int64
1   Survived     891 non-null    int64
2   Pclass       891 non-null    int64
3   Name         891 non-null    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
```

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
PassengerId	891.0	NaN	NaN	NaN	446.0	257.353842	1.0	223.5	446.0	668.5	891.0
Survived	891.0	NaN	NaN	NaN	0.383838	0.486592	0.0	0.0	0.0	1.0	1.0
Pclass	891.0	NaN	NaN	NaN	2.308642	0.836071	1.0	2.0	3.0	3.0	3.0
Name	891	891	Dooley, Mr. Patrick	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Sex	891	2	male	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	714.0	NaN	NaN	NaN	29.699118	14.526497	0.42	20.125	28.0	38.0	80.0
SibSp	891.0	NaN	NaN	NaN	0.523008	1.102743	0.0	0.0	0.0	1.0	8.0
Parch	891.0	NaN	NaN	NaN	0.381594	0.806057	0.0	0.0	0.0	0.0	6.0
Ticket	891	681	347082	7	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Fare	891.0	NaN	NaN	NaN	32.204208	49.693429	0.0	7.9104	14.4542	31.0	512.3292
Cabin	204	147	G6	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Embarked	889	3	S	644	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
train.isnull().sum()
sns.heatmap(train.isnull(), cbar=False)
```



```
fig = sns.barplot(data=train, x='Sex', y='Survived')  
fig.figure.savefig("survival_by_sex.png")
```



```
from google.colab import files  
files.download("survival_by_sex.png")
```

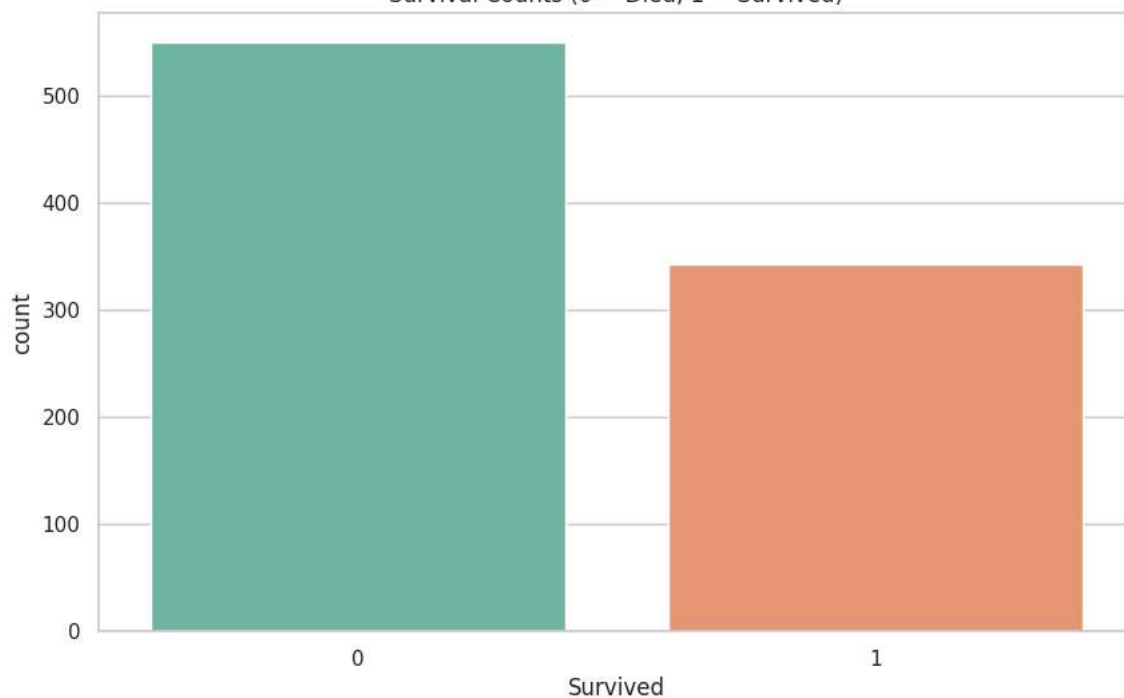
```
sns.countplot(data=train, x='Survived', palette="Set2")  
plt.title("Survival Counts (0 = Died, 1 = Survived)")  
plt.show()
```

```
/tmp/ipython-input-720095889.py:1: FutureWarning:
```

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `1`

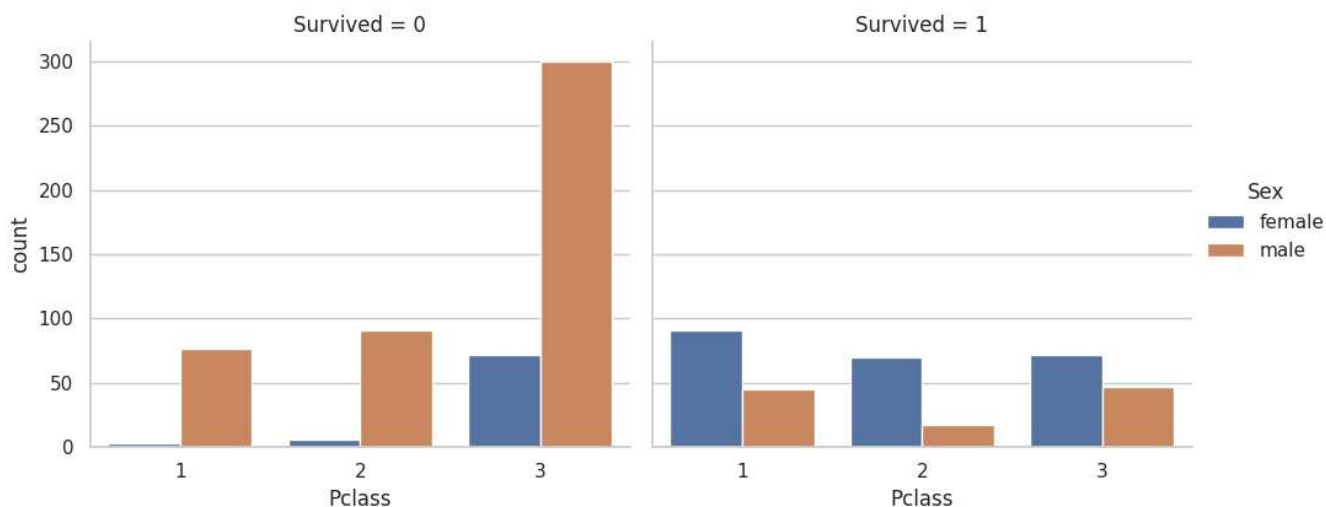
```
sns.countplot(data=train, x='Survived', palette="Set2")
```

Survival Counts (0 = Died, 1 = Survived)



```
sns.catplot(data=train, x="Pclass", hue="Sex", col="Survived",
            kind="count", height=5, aspect=1)
plt.subplots_adjust(top=0.8)
plt.suptitle("Passenger Class & Sex vs Survival")
plt.show()
```

Passenger Class & Sex vs Survival



```
sns.kdeplot(data=train, x="Age", hue="Survived", shade=True)
plt.title("Age Distribution: Survived vs Died")
plt.show()
```

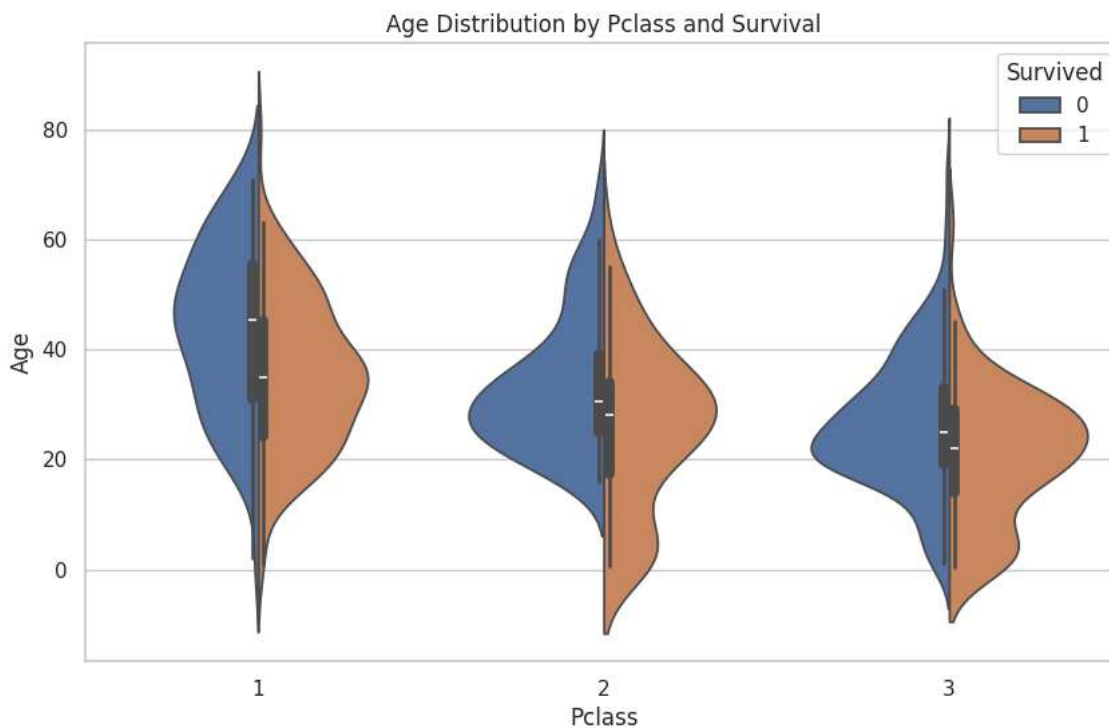
```
/tmp/ipython-input-3370827614.py:1: FutureWarning:
```

```
`shade` is now deprecated in favor of `fill`; setting `fill=True`.
This will become an error in seaborn v0.14.0; please update your code.
```

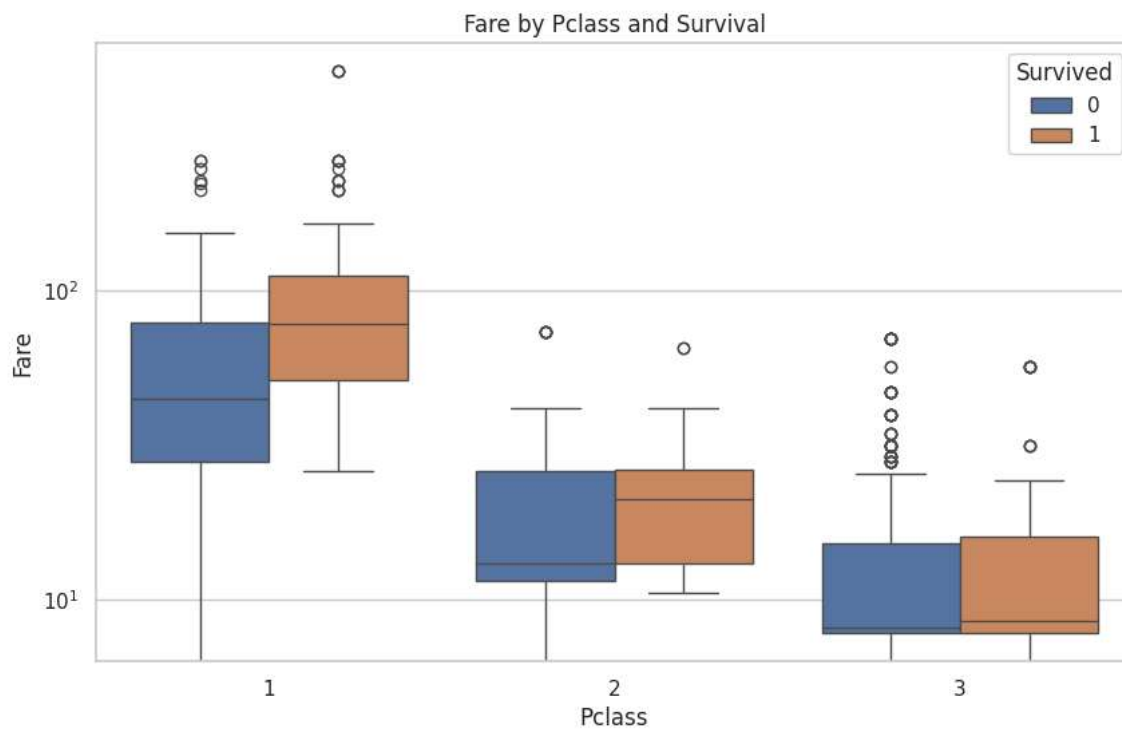
```
sns.kdeplot(data=train, x="Age", hue="Survived", shade=True)
```



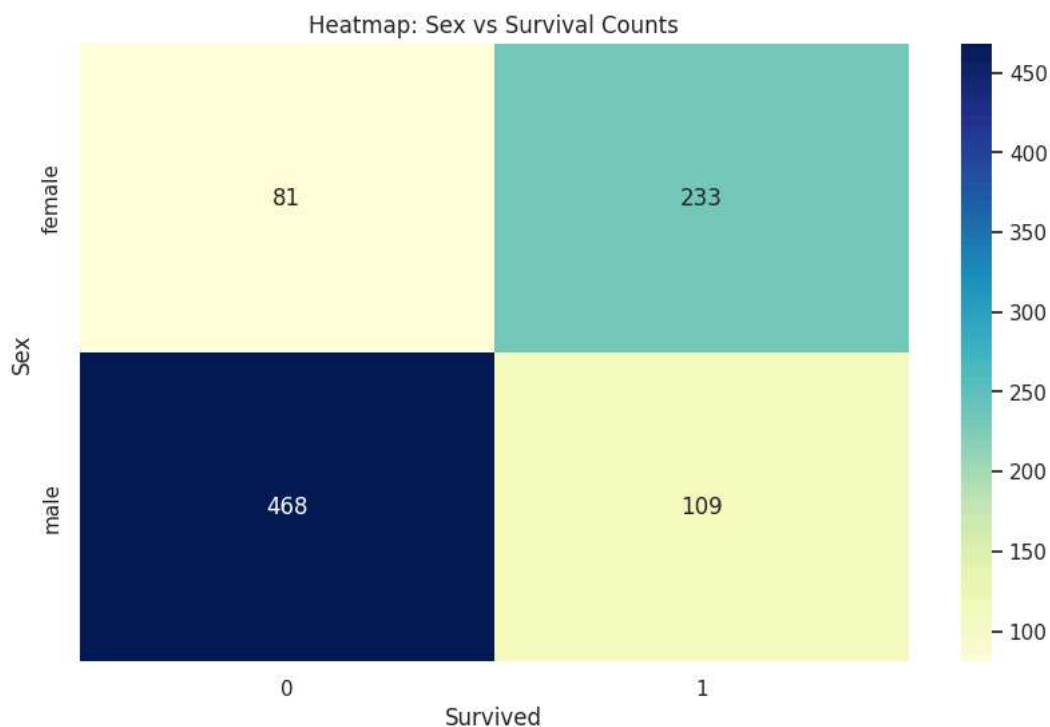
```
sns.violinplot(data=train, x="Pclass", y="Age", hue="Survived", split=True)
plt.title("Age Distribution by Pclass and Survival")
plt.show()
```



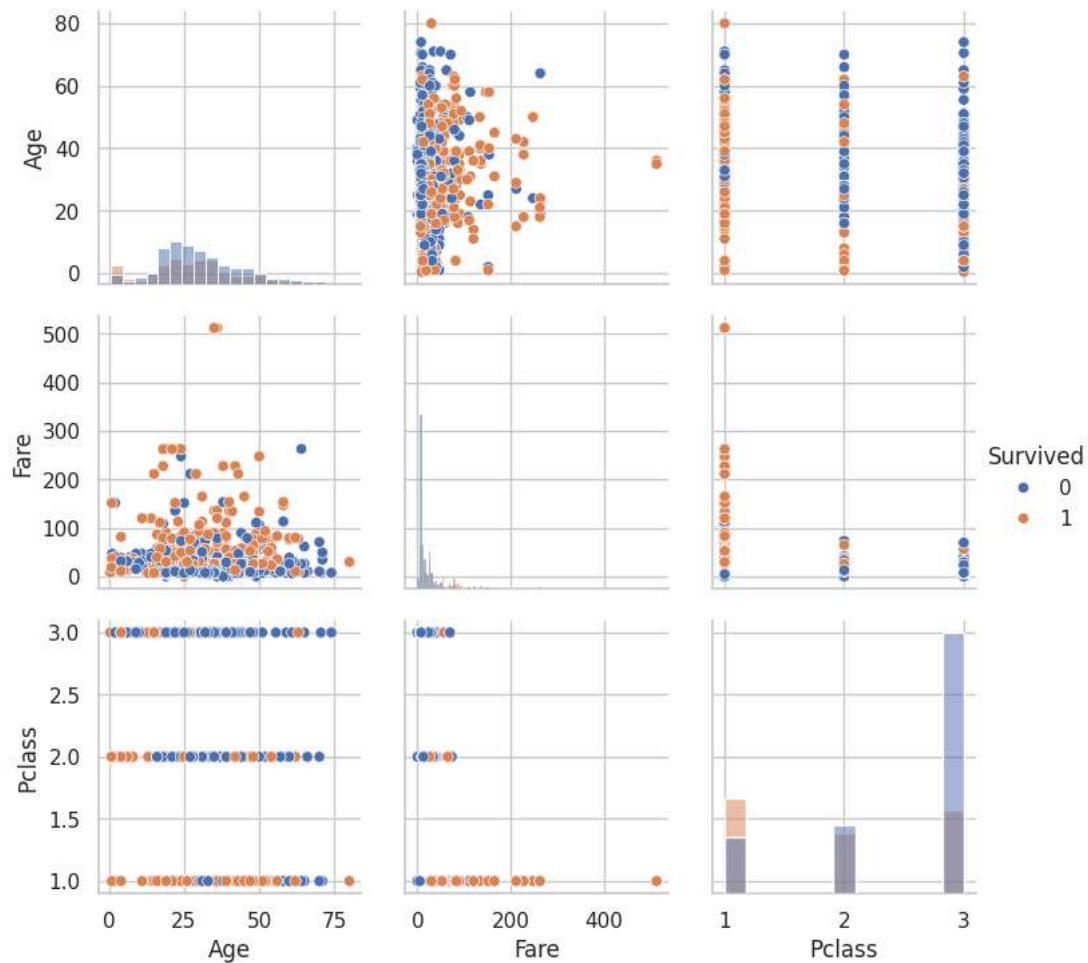
```
sns.boxplot(data=train, x="Pclass", y="Fare", hue="Survived")
plt.yscale('log')
plt.title("Fare by Pclass and Survival")
plt.show()
```



```
cat_features = ['Sex', 'Pclass', 'Embarked', 'Title']
survival_matrix = pd.crosstab(index=train[cat_features[0]], columns=train['Survived'])
sns.heatmap(survival_matrix, annot=True, fmt="d", cmap="YlGnBu")
plt.title("Heatmap: Sex vs Survival Counts")
plt.show()
```

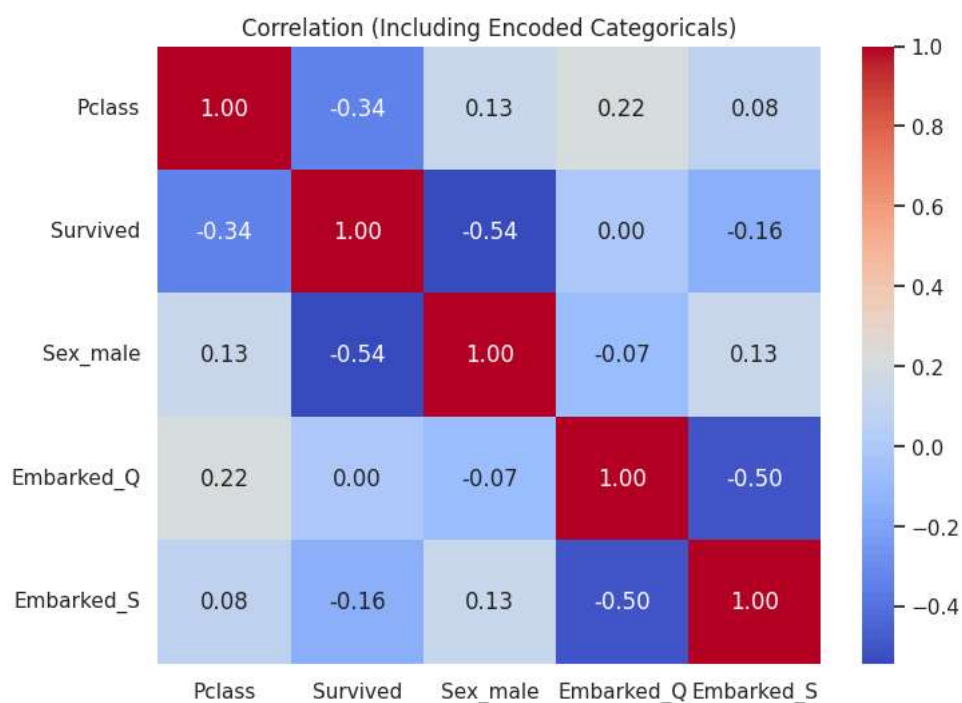


```
g = sns.PairGrid(train[['Age', 'Fare', 'Pclass', 'Survived']], hue="Survived")
g.map_diag(sns.histplot)
g.map_offdiag(sns.scatterplot)
g.add_legend()
plt.show()
```



```
encoded = pd.get_dummies(train[['Sex', 'Embarked', 'Pclass', 'Survived']], drop_first=True)
corr = encoded.corr()
```

```
plt.figure(figsize=(8,6))
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation (Including Encoded Categoricals)")
plt.show()
```



```
train['FamilySize'] = train['SibSp'] + train['Parch'] + 1
```

```
sns.barplot(data=train, x="FamilySize", y="Survived", palette="viridis")  
plt.title("Survival Rate by Family Size")  
plt.show()
```

/tmp/ipython-input-3725237907.py:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `1

```
sns.barplot(data=train, x="FamilySize", y="Survived", palette="viridis")
```

