

## IMPORTING LIBRARIES

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

df = pd.read_csv('train.csv')
df.head()

{"summary": "{\n  \"name\": \"df\",\n  \"rows\": 891,\n  \"fields\": [\n    {\n      \"column\": \"PassengerId\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 257,\n        \"min\": 1,\n        \"max\": 891,\n        \"num_unique_values\": 891,\n        \"samples\": [\n          710,\n          440,\n          841\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Survived\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 0,\n        \"min\": 0,\n        \"max\": 1,\n        \"num_unique_values\": 2,\n        \"samples\": [\n          1,\n          0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Pclass\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 0,\n        \"min\": 1,\n        \"max\": 3,\n        \"num_unique_values\": 3,\n        \"samples\": [\n          3,\n          1\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Name\",\n      \"properties\": {\n        \"dtype\": \"string\",\n        \"num_unique_values\": 891,\n        \"samples\": [\n          \"Moubarak, Master. Halim Gonios (\\\\\\\\\"William George\\\\\\\\\")\",\n          \"Kvillner, Mr. Johan Henrik Johannesson\"\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Sex\",\n      \"properties\": {\n        \"dtype\": \"category\",\n        \"num_unique_values\": 2,\n        \"samples\": [\n          \"female\",\n          \"male\"\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Age\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 14.526497332334044,\n        \"min\": 0.42,\n        \"max\": 80.0,\n        \"num_unique_values\": 88,\n        \"samples\": [\n          0.75,\n          22.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"SibSp\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 1,\n        \"min\": 0,\n        \"max\": 8,\n        \"num_unique_values\": 7,\n        \"samples\": [\n          1,\n          0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Parch\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 0,\n        \"min\": 0,\n        \"max\": 6,\n        \"num_unique_values\": 7,\n        \"samples\": [\n          0,\n
```

```

1\n      ],\n      \"semantic_type\": \"\",\n      \"description\": \"\"\n    },\n    {\n      \"column\": \"Ticket\",\n      \"properties\": {\n        \"dtype\": \"string\",\n        \"num_unique_values\": 681,\n        \"samples\": [\n          \"11774\",\n          \"248740\"\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      },\n      {\n        \"column\": \"Fare\",\n        \"properties\": {\n          \"dtype\": \"number\",\n          \"std\": 49.693428597180905,\n          \"min\": 0.0,\n          \"max\": 512.3292,\n          \"num_unique_values\": 248,\n          \"samples\": [\n            11.2417,\n            51.8625\n          ],\n          \"semantic_type\": \"\",\n          \"description\": \"\"\n        },\n        {\n          \"column\": \"Cabin\",\n          \"properties\": {\n            \"dtype\": \"category\",\n            \"num_unique_values\": 147,\n            \"samples\": [\n              \"D45\",\n              \"B49\"\n            ],\n            \"semantic_type\": \"\",\n            \"description\": \"\"\n          },\n          {\n            \"column\": \"Embarked\",\n            \"properties\": {\n              \"dtype\": \"category\",\n              \"num_unique_values\": 3,\n              \"samples\": [\n                \"S\",\n                \"C\"\n              ],\n              \"semantic_type\": \"\",\n              \"description\": \"\"\n            }\n          }\n        }\n      ],\n      \"type\": \"dataframe\", \"variable_name\": \"df\"}

```

```

df.describe()
df.info()
df.isna().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

```
df.value_counts('Survived').reset_index()
```

```
{
  "summary": {
    "name": "df",
    "rows": 2,
    "fields": [
      {
        "column": "Survived",
        "properties": {
          "dtype": "number",
          "std": 0,
          "min": 0,
          "max": 1,
          "num_unique_values": 2,
          "samples": [1, 0]
        },
        "semantic_type": "description"
      },
      {
        "column": "count",
        "properties": {
          "dtype": "number",
          "std": 146,
          "min": 342,
          "max": 549,
          "num_unique_values": 2,
          "samples": [342, 549]
        },
        "semantic_type": "description"
      }
    ],
    "type": "dataframe"
  }
}
```

## REMOVING NULL VALUES FROM TABLE

```
df = df.dropna()
print(f"Original number of rows: {len(df)}")
print(f"Number of rows after dropping nulls: {len(df)}")
display(df.head())
```

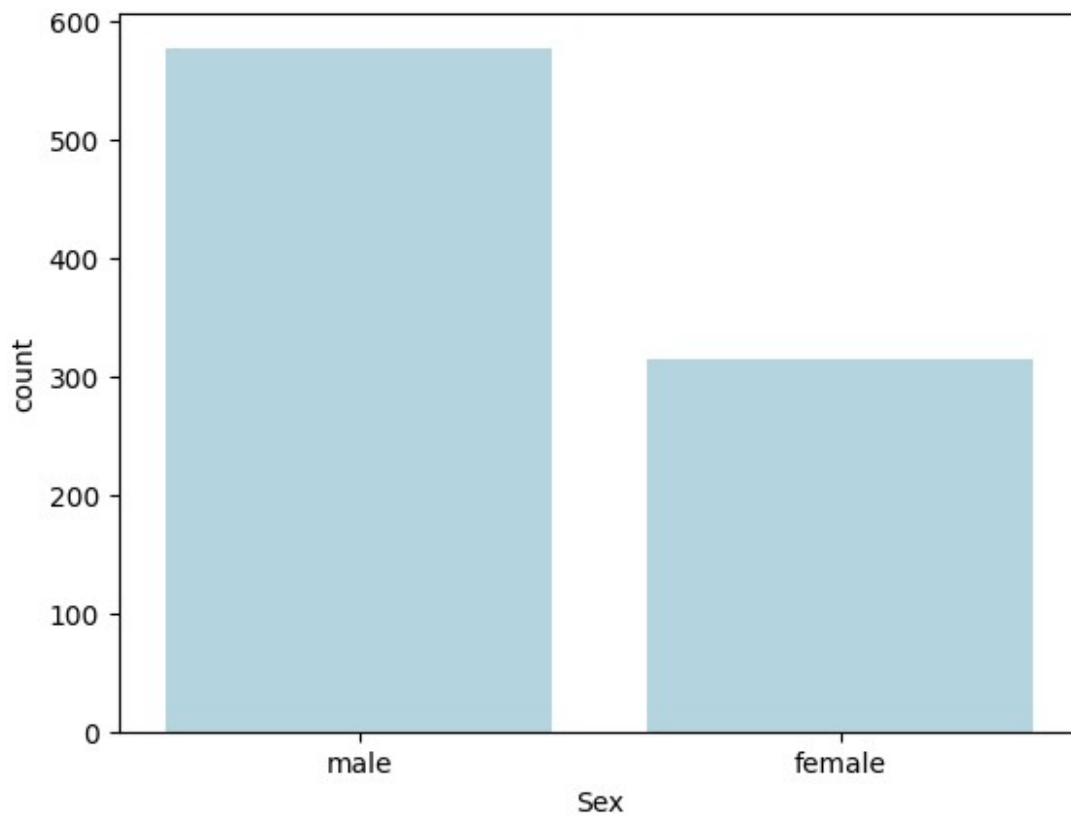
Original number of rows: 183  
Number of rows after dropping nulls: 183

```
{
  "summary": {
    "name": "display(df)",
    "rows": 5,
    "fields": [
      {
        "column": "PassengerId",
        "properties": {
          "dtype": "number",
          "std": 4,
          "min": 2,
          "max": 12,
          "num_unique_values": 5,
          "samples": [4, 12, 7]
        },
        "semantic_type": "description"
      },
      {
        "column": "Survived",
        "properties": {
          "dtype": "number",
          "std": 0,
          "min": 0,
          "max": 1,
          "num_unique_values": 2,
          "samples": [0, 1]
        },
        "semantic_type": "description"
      },
      {
        "column": "Pclass",
        "properties": {
          "dtype": "number",
          "std": 0,
          "min": 1,
          "max": 3,
          "num_unique_values": 2,
          "samples": [1, 3]
        },
        "semantic_type": "description"
      }
    ],
    "type": "dataframe"
  }
}
```

```
[
  3,
  1,
],
{
  "semantic_type":
  "\",
  "description": "\",
  },
  {
  "column": "Name",
  "properties": {
    "dtype":
    "string",
    "num_unique_values": 5,
    "samples":
    [
      "Futrelle, Mrs. Jacques Heath (Lily May Peel)",
      "Bonnell, Miss. Elizabeth"
    ],
    "semantic_type":
    "\",
    "description": "\",
    },
    {
    "column": "Sex",
    "properties": {
      "dtype":
      "category",
      "num_unique_values": 2,
      "samples":
      [
        "male",
        "female"
      ],
      "semantic_type": "\",
      "description": "\",
      },
      {
      "column": "Age",
      "properties": {
        "dtype": "number",
        "std": 21.33541656495134,
        "min": 4.0,
        "max": 58.0,
        "num_unique_values": 5,
        "samples": [
          35.0,
          58.0
        ],
        "semantic_type": "\",
        "description": "\",
        },
        {
        "column": "SibSp",
        "properties": {
          "dtype": "number",
          "std": 0,
          "min": 0,
          "max": 1,
          "num_unique_values": 2,
          "samples": [
            0,
            1
          ],
          "semantic_type": "\",
          "description": "\",
          },
          {
          "column": "Parch",
          "properties": {
            "dtype": "number",
            "std": 0,
            "min": 0,
            "max": 1,
            "num_unique_values": 2,
            "samples": [
              1,
              0
            ],
            "semantic_type": "\",
            "description": "\",
            },
            {
            "column": "Ticket",
            "properties": {
              "dtype": "string",
              "num_unique_values": 5,
              "samples": [
                "113803",
                "113783"
              ],
              "semantic_type": "\",
              "description": "\",
              },
              {
              "column": "Fare",
              "properties": {
                "dtype": "number",
                "std": 22.015613900207278,
                "min": 16.7,
                "max": 71.2833,
                "num_unique_values": 5,
                "samples": [
                  53.1,
                  26.55
                ],
                "semantic_type": "\",
                "description": "\",
                },
                {
                "column": "Cabin",
                "properties": {
                  "dtype": "string",
                  "num_unique_values": 5,
                  "samples": [
                    "C123",
                    "C103"
                  ],
                  "semantic_type": "\",
                  "description": "\",
                  },
                  {
                  "column": "Embarked",
                  "properties": {
                    "dtype":
                    "category",
                    "num_unique_values": 2,
                    "samples":
                    [
                      "S",
                      "C"
                    ],
                    "semantic_type": "\",
                    "description": "\",
                    }
                  }
                ]
              },
              "type": "dataframe"
            }
          ]
        }
      ]
    }
  ]
}
```

VISUALISATION OF DATASET

```
sns.countplot(x='Sex', data=df, color = 'lightblue')  
plt.show()
```



According to the above histogram, there were close to 600 males and 300 females on the Titanic altogether.

```
sns.pairplot(data = df, hue='Survived', palette = 'rainbow')  
plt.show()
```



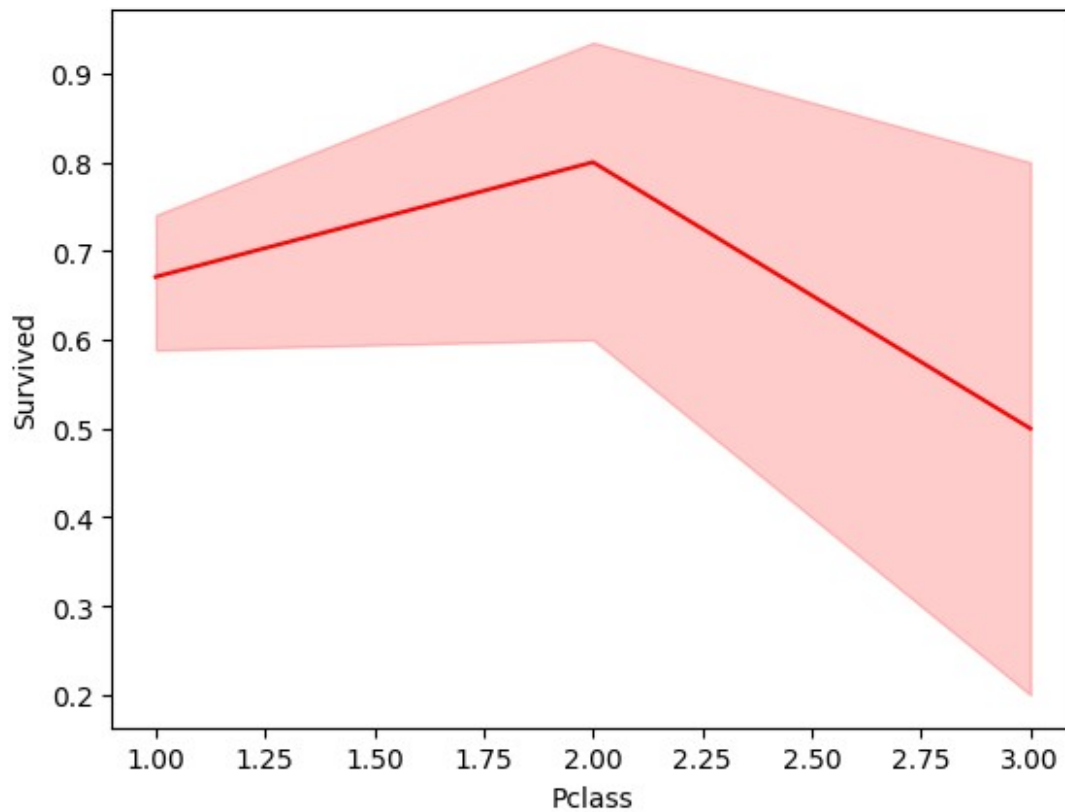
This Seaborn pairplot of the Titanic dataset compares passenger features and survival, with **yellow points representing survivors and cyan points representing non-survivors**.

**The diagonal plots show the distribution of each variable:** Pclass reveals more survivors in 1st class and more non-survivors in 3rd class, Age suggests higher survival among children, Fare indicates that passengers who paid higher fares had better survival rates, and SibSp and Parch show most people traveled alone, though small family groups had better survival odds.

**Off-diagonal scatterplots reveal relationships between variables:** Pclass and Fare are strongly related, with higher fares in 1st class and more survivors in that group; Age and Fare show older passengers often paid more in 1st class and survived more frequently; and SibSp and Parch suggest large family sizes had lower survival rates while small families fared better.

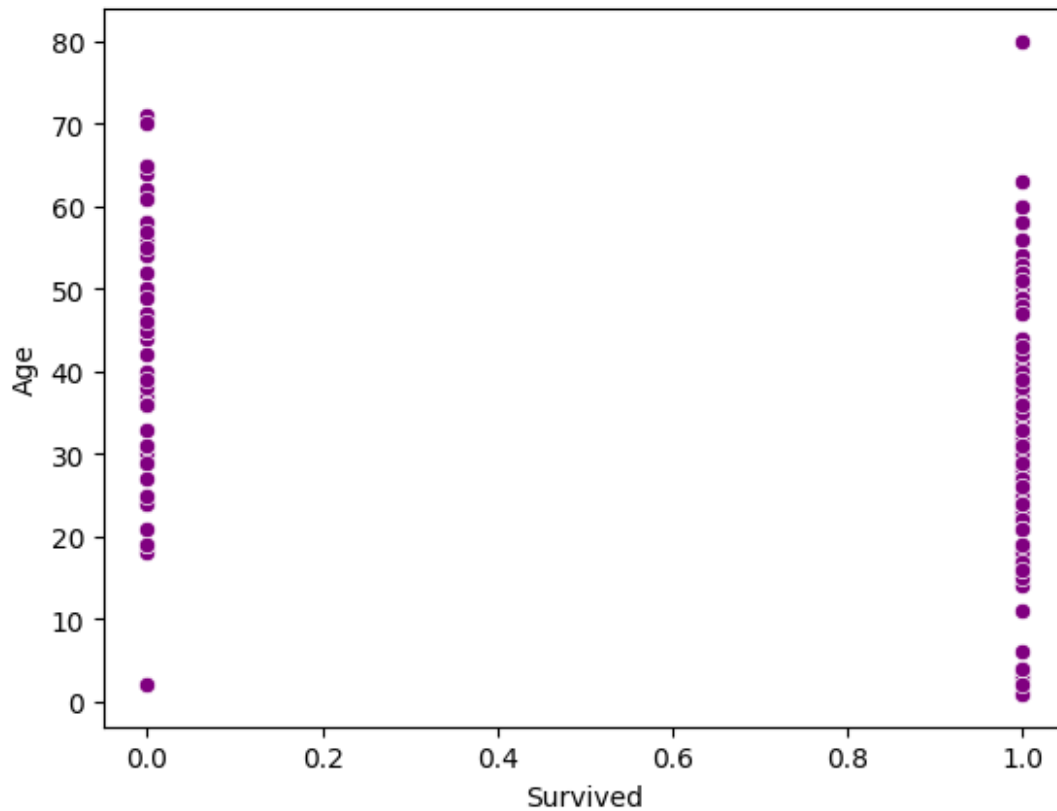
**Overall, higher class, higher fare, and being younger or part of a small family are visually associated with higher survival odds.**

```
sns.lineplot(x='Pclass', y='Survived', data=df, color = 'red')  
plt.show()
```



The above line plot shows the relationship between the class of a passenger and the rate of survival. It can be inferred that those in the second class had a higher rate of survival.

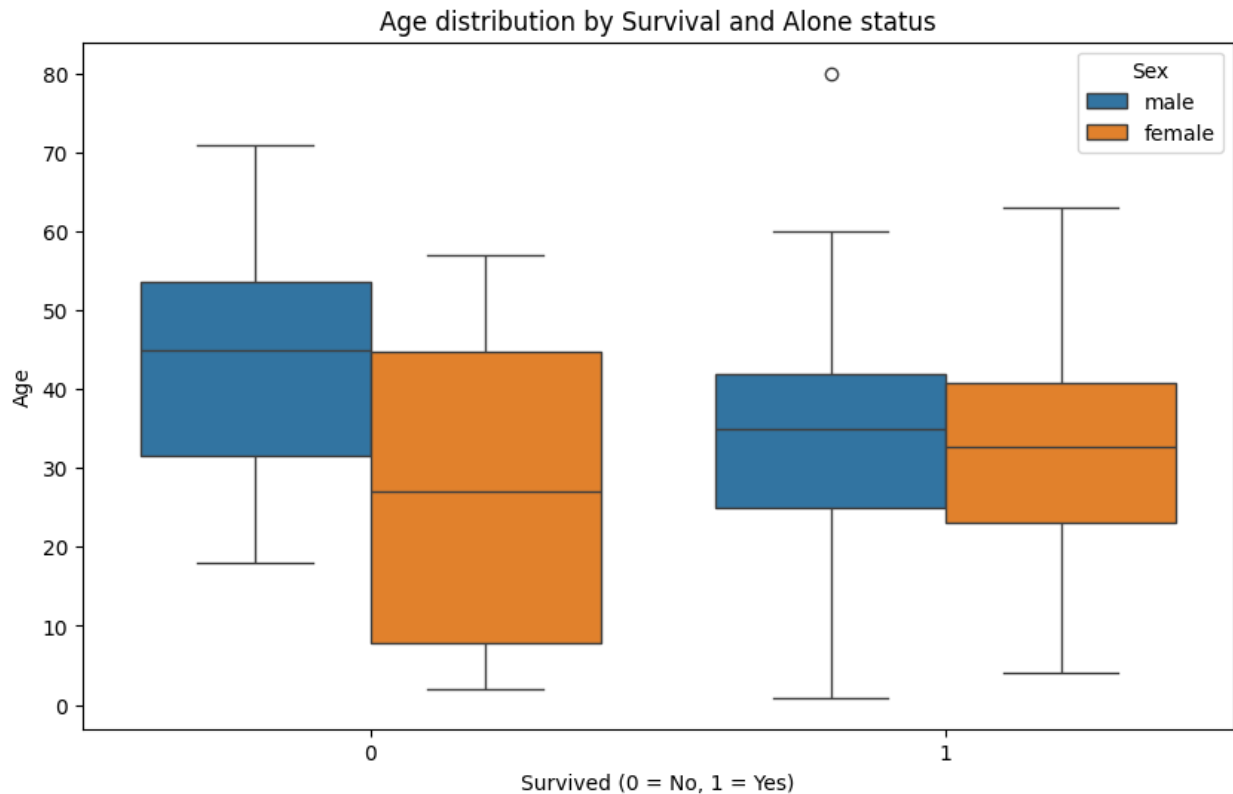
```
sns.scatterplot(x='Survived', y='Age', data=df, color = 'purple')  
plt.show()
```



The above scatterplot indicates that most survivors were in a lower age range.

```
plt.figure(figsize=(10,6))
sns.boxplot(x='Survived', y='Age', hue = 'Sex', data= df)
plt.xlabel('Survived (0 = No, 1 = Yes)')
plt.ylabel('Age')
plt.title('Age distribution by Survival and Alone status')
plt.show()
```





**This boxplot shows the distribution of passenger ages on the Titanic, split by survival status (0 = did not survive, 1 = survived) and sex, providing insights into how age and gender related to survival outcomes.**

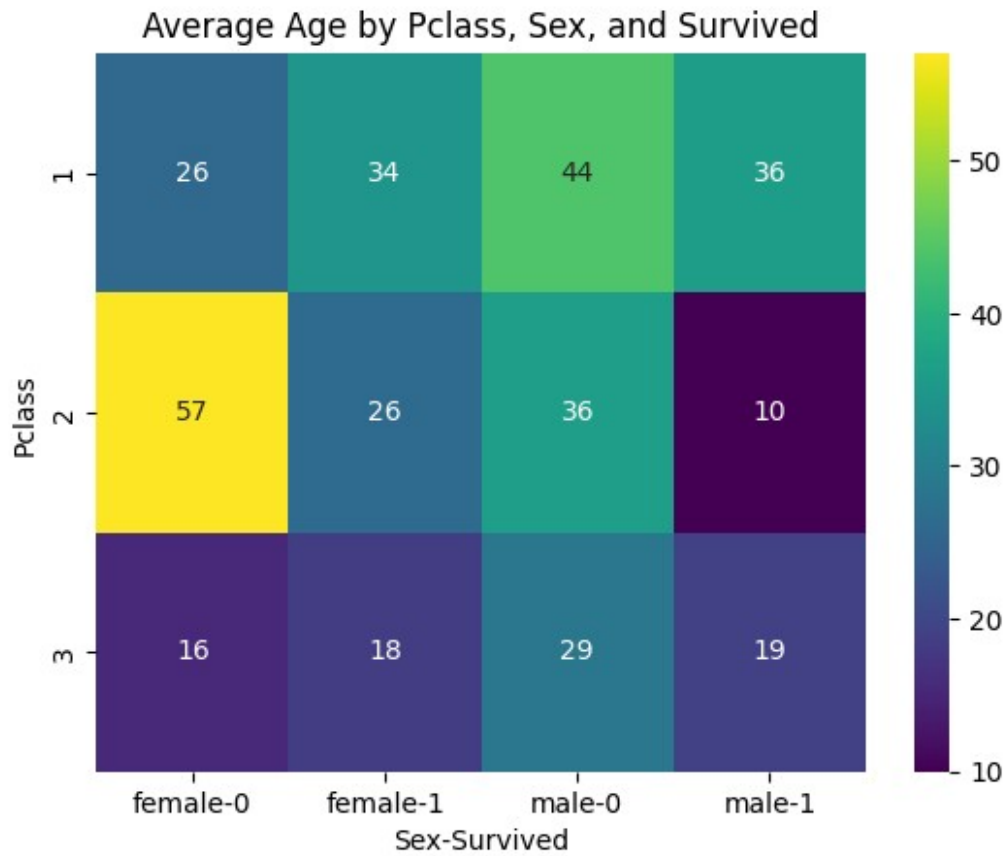
For non-survivors (0), males tended to be older on average, with a median age in the mid-40s, and had a broad spread from late teens to around 70, while females had a lower median age in the late 20s and a wider spread including many young children.

For survivors (1), both males and females generally had lower median ages than non-survivors, with male survivors' ages clustered in the mid-30s and female survivors in the early 30s, though female survivors showed a fairly balanced distribution across all ages and fewer extreme values.

The plot also shows outliers, such as an older male survivor around age 80, indicating exceptions to general trends.

**Overall, it suggests that survival chances were higher for women and younger individuals, aligning with the "women and children first" evacuation pattern.**

```
average_age = df.pivot_table(index='Pclass', columns=['Sex',  
'Survived'], values='Age', aggfunc='mean')  
sns.heatmap(data=average_age, annot=True, cmap='viridis')  
plt.title('Average Age by Pclass, Sex, and Survived')  
plt.show()
```



**This heatmap shows the average ages of Titanic passengers by class, sex, and survival, revealing that ages varied widely across groups.**

First-class passengers tended to be older, especially non-surviving males (44), while third-class passengers were generally younger, with many survivors in their teens or twenties.

Second-class non-surviving females had the highest average age (57), whereas surviving second-class males were the youngest (10), likely children.

**Overall, survival was often linked with being younger, female, and in a higher class.**