

## Titanic Data Analysis

This is to show the EDA(Exploratory Data Analysis) of the Titanic dataset along with visualization. The objective is to analyse the factors that influenced passenger survival. Through compelling visualizations and statistical techniques, we'll unearth hidden insights within the dataset.

EDA stands for Exploratory Data Analysis. In Python, it refers to the process of using Python libraries and techniques to investigate, analyze, and visualize data to understand its characteristics, patterns, and relationships. It will have Observational summary, statistical summary, missing values and it's potential impact on the data.

After my analysis, we would able to observe and find certain answer to the questions such as "what sorts of people were more likely to survive?" using passenger data (ie name, age, gender, socio-economic class, etc).

By the end, we will have a detailed visual and quantitative understanding of the Titanic's passenger data.

```
#import data and required packages
```

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

```
#Loading the file
titanic_df = pd.read_csv("/content/train.csv")
```

```
#Displaying first five data using head()
titanic_df.head()
```

```
{"summary":{"\n  \"name\": \"titanic_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      \"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      \"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\": \"\"
```



```
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
```

here, the info() is displaying information regarding the data set in which the shape of the titanic\_df DataFrame is 891 rows,12 columns(numeric col:7 and non numeric col:5) with their respective datatype. Also we can observe that dataframe has few null values in Age,Cabin and Embarked col.

```
titanic_df.describe()

{"summary":{"name": "titanic_df", "rows": 8, "fields": [{"column": "PassengerId", "properties": {"dtype": "number", "std": 320.8159711429855, "min": 1.0, "max": 891.0, "num_unique_values": 6, "samples": [891.0, 446.0, 668.5]}, {"column": "Survived", "properties": {"dtype": "number", "std": 314.8713661874558, "min": 0.0, "max": 891.0, "num_unique_values": 5, "samples": [0.3838383838383838, 1.0, 0.4865924542648575]}, {"column": "Pclass", "properties": {"dtype": "number", "std": 314.2523437079694, "min": 0.836071240977049, "max": 891.0, "num_unique_values": 6, "samples": [891.0, 2.308641975308642, 3.0]}], "description": ""}, {"name": "Age", "rows": 714, "fields": [{"column": "Age", "properties": {"dtype": "number", "std": 242.9056731818781, "min": 0.42, "max": 714.0, "num_unique_values": 8, "samples": [29.69911764705882, 28.0, 714.0]}], "description": ""}]}
```

```

n    },\n    {\n        \"column\": \"SibSp\", \n        \"properties\": {\n        \"dtype\": \"number\", \n        \"std\": 314.4908277465442, \n        \"min\": 0.0, \n        \"max\": 891.0, \n        \"num_unique_values\":\n        6, \n        \"samples\": [\n            891.0, \n            0.5230078563411896, \n            8.0\n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\"\n    }\n    },\n    {\n        \"column\": \"Parch\", \n        \"properties\": {\n        \"dtype\": \"number\", \n        \"std\": 314.65971717879, \n        \"min\": 0.0, \n        \"max\": 891.0, \n        \"num_unique_values\":\n        5, \n        \"samples\": [\n            0.38159371492704824, \n            6.0, \n            0.8060572211299483\n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\"\n    }\n    },\n    {\n        \"column\": \"Fare\", \n        \"properties\": {\n        \"dtype\": \"number\", \n        \"std\": 330.6256632228578, \n        \"min\": 0.0, \n        \"max\": 891.0, \n        \"num_unique_values\":\n        8, \n        \"samples\": [\n            32.204207968574636, \n            14.4542, \n            891.0\n        ], \n        \"semantic_type\":\n        \"\", \n        \"description\": \"\"\n    }\n    }\n    ],\n    \"type\": \"dataframe\"}

```

The describe() is used to check the statistical summary of the dataframe. Survived col is an indicator where 1 represents survival and 0 represents non-survival. It is showing clear that around 38.4% of passengers (mean survival rate of 0.38) managed to survive, suggesting that less than half of the passengers survived.

The average passenger class (Pclass) sits at 2.3, indicating that most passengers belonged to the second or third class, with fewer in the more luxurious first class.

The youngest passenger was a mere infant (0.42 years old), while the oldest was a seasoned traveler of 80 years.

*#Looking for null values in the dataset*

```
titanic_df.isnull()
```

```

{"summary": "{\n    \"name\": \"titanic_df\", \n    \"rows\": 891, \n    \"fields\": [\n        {\n            \"column\": \"PassengerId\", \n            \"properties\": {\n                \"dtype\": \"boolean\", \n                \"num_unique_values\": 1, \n                \"samples\": [\n                    false\n                ], \n                \"semantic_type\": \"\", \n                \"description\": \"\"\n            }\n        }, \n        {\n            \"column\": \"Survived\", \n            \"properties\": {\n                \"dtype\": \"boolean\", \n                \"num_unique_values\": 1, \n                \"samples\": [\n                    false\n                ], \n                \"semantic_type\": \"\", \n                \"description\": \"\"\n            }\n        }, \n        {\n            \"column\": \"Pclass\", \n            \"properties\": {\n                \"dtype\": \"boolean\", \n                \"num_unique_values\": 1, \n                \"samples\": [\n                    false\n                ], \n                \"semantic_type\": \"\", \n                \"description\": \"\"\n            }\n        }, \n        {\n            \"column\": \"Name\", \n            \"properties\": {\n                \"dtype\": \"boolean\", \n                \"num_unique_values\": 1, \n                \"samples\": [\n                    false\n                ], \n                \"semantic_type\": \"\", \n                \"description\": \"\"\n            }\n        }\n    ], \n    \"type\": \"dataframe\"}

```

```

{"semantic_type": "\n",\n      "description": "\n",\n      },\n      {\n        "column": "Sex",\n        "properties": {\n          "dtype": "boolean",\n          "num_unique_values": 1,\n          "samples": [\n            false\n          ],\n          "semantic_type": "\n",\n          "description": "\n",\n          },\n          {\n            "column": "Age",\n            "properties": {\n              "dtype": "boolean",\n              "num_unique_values": 2,\n              "samples": [\n                true\n              ],\n              "semantic_type": "\n",\n              "description": "\n",\n              },\n              {\n                "column": "SibSp",\n                "properties": {\n                  "dtype": "boolean",\n                  "num_unique_values": 1,\n                  "samples": [\n                    false\n                  ],\n                  "semantic_type": "\n",\n                  "description": "\n",\n                  },\n                  {\n                    "column": "Parch",\n                    "properties": {\n                      "dtype": "boolean",\n                      "num_unique_values": 1,\n                      "samples": [\n                        false\n                      ],\n                      "semantic_type": "\n",\n                      "description": "\n",\n                      },\n                      {\n                        "column": "Ticket",\n                        "properties": {\n                          "dtype": "boolean",\n                          "num_unique_values": 1,\n                          "samples": [\n                            false\n                          ],\n                          "semantic_type": "\n",\n                          "description": "\n",\n                          },\n                          {\n                            "column": "Fare",\n                            "properties": {\n                              "dtype": "boolean",\n                              "num_unique_values": 1,\n                              "samples": [\n                                false\n                              ],\n                              "semantic_type": "\n",\n                              "description": "\n",\n                              },\n                              {\n                                "column": "Cabin",\n                                "properties": {\n                                  "dtype": "boolean",\n                                  "num_unique_values": 2,\n                                  "samples": [\n                                    false\n                                  ],\n                                  "semantic_type": "\n",\n                                  "description": "\n",\n                                  },\n                                  {\n                                    "column": "Embarked",\n                                    "properties": {\n                                      "dtype": "boolean",\n                                      "num_unique_values": 2,\n                                      "samples": [\n                                        true\n                                      ],\n                                      "semantic_type": "\n",\n                                      "description": "\n",\n                                      }\n                                    }\n                                  }\n                                }\n                              }\n                            }\n                          }\n                        }\n                      }\n                    }\n                  }\n                }\n              }\n            }\n          }\n        }\n      }\n    ],\n    "type": "dataframe"}

```

```
titanic_df.isnull().sum()
```

```

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

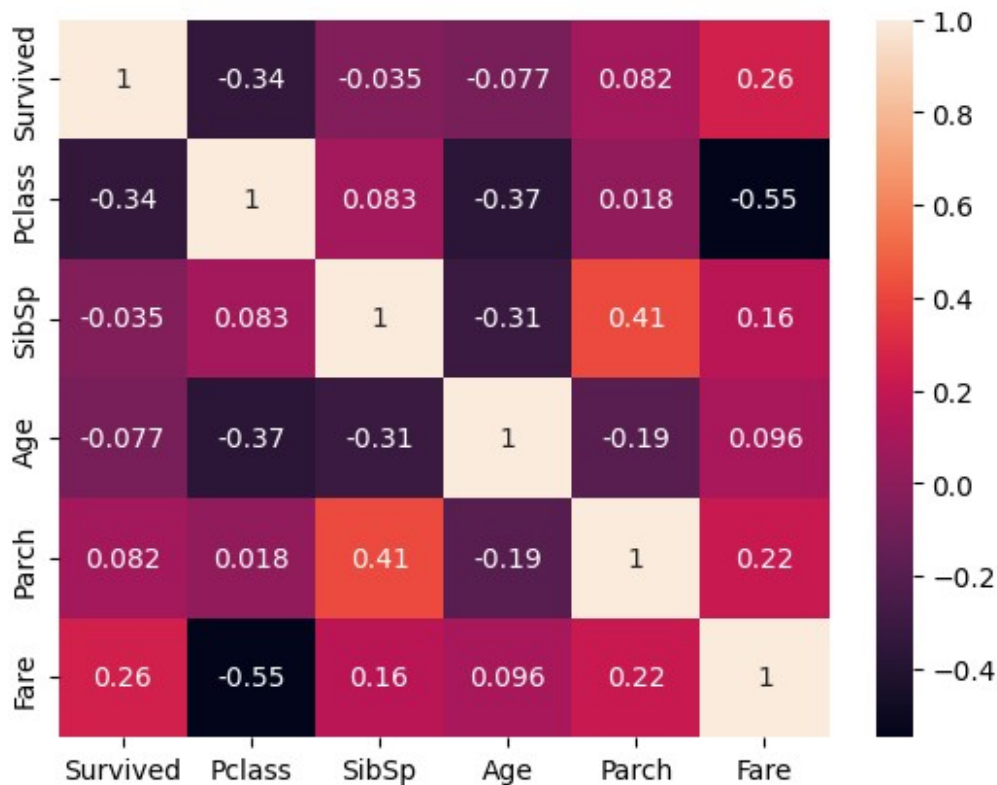
```

```
titanic_df.isnull().sum().sum()
```

866

here, we have calculated total number of null values in the dataframe. Age has 177 missing entries. Cabin has 687 missing entries. Embarked has 2 missing entries.

```
heatmap = sns.heatmap(titanic_df[['Survived',  
'Pclass', 'SibSp', 'Age', 'Parch', 'Fare']].corr(), annot = True)
```



```
titanic_df['Survived'].unique()
```

```
array([0, 1])
```

```
titanic_df['Pclass'].unique()
```

```
array([3, 1, 2])
```

pclass col is the Ticket class, 1 = 1st class of ticket, 2 = 2nd, 3 = 3rd

```
titanic_df['SibSp'].unique()
```

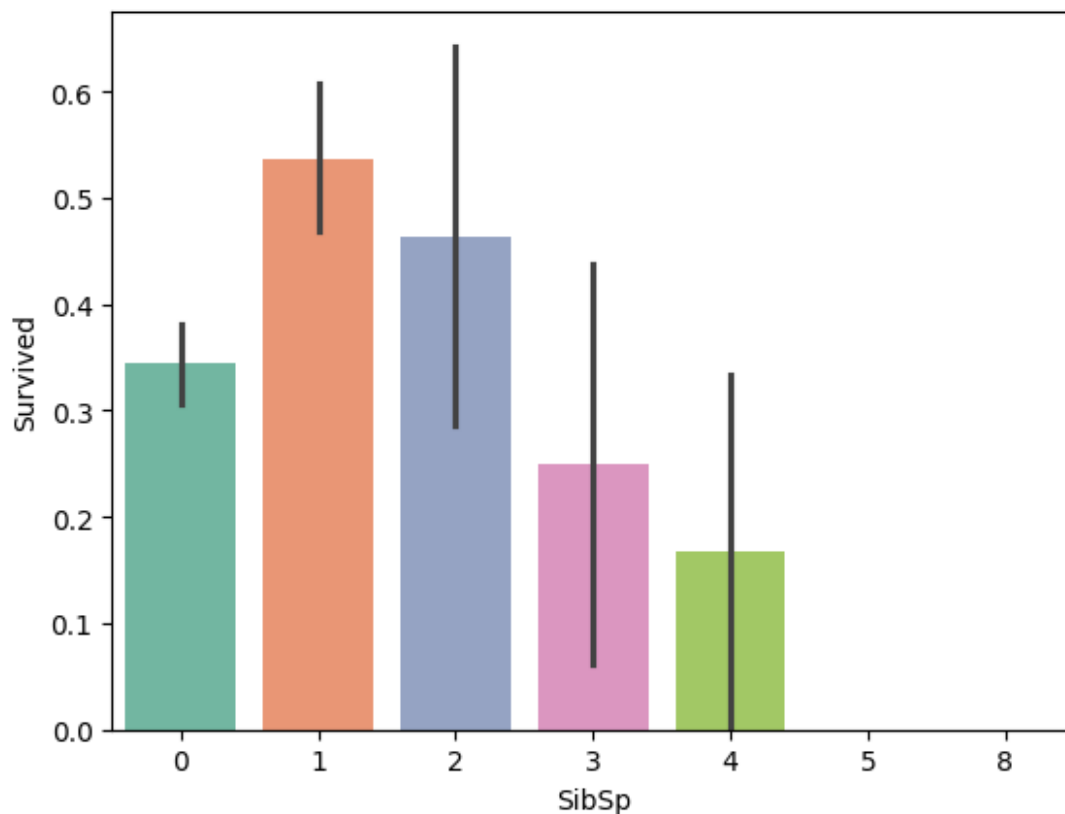
```
array([1, 0, 3, 4, 2, 5, 8])
```

```
titanic_df['Parch'].unique()
```

```
array([0, 1, 2, 5, 3, 4, 6])
```

Heatmap is used to visualize the correlation coefficients between the variables in a dataset. Positive correlations (variables tend to move together) while while negative correlations (variables tend to move in opposite directions). The intensity of the color would indicate the strength of the correlation, with darker colors representing stronger correlations. Additionally, numerical values within each square would likely represent the actual correlation coefficient between the corresponding features. therefore,we analysed the variables: Survived column has relation with Fare column, the relation is positive (0.26) meaning, more the fare, higher is the chances of survival. likewise, we may compare variables and generate more insights.

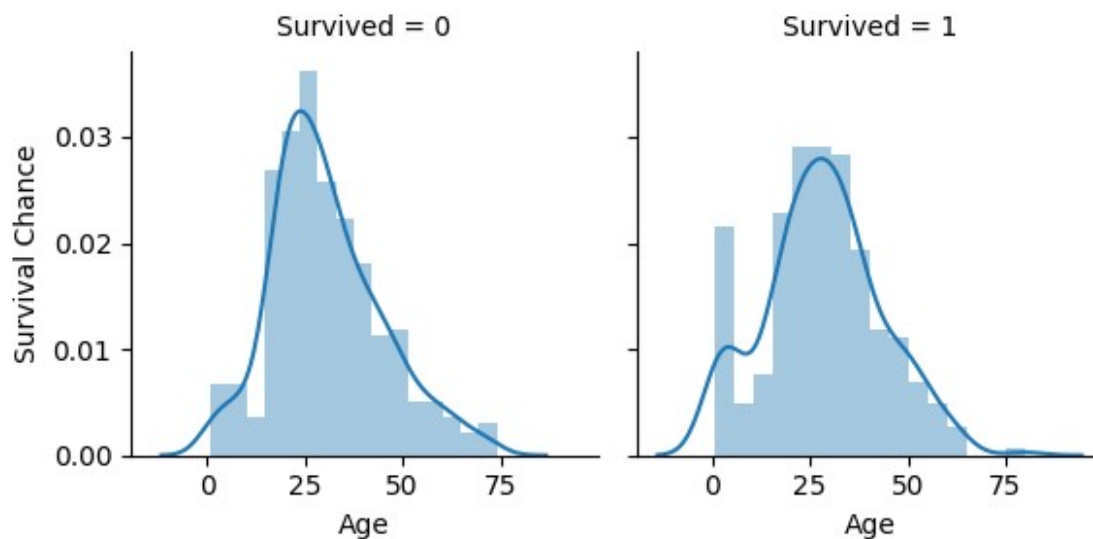
```
bar_sibsp = sns.barplot(data = titanic_df, x = 'SibSp', y =  
'Survived',palette="Set2")  
plt.show()
```



Passengers having more number of siblings are less likely to survive.

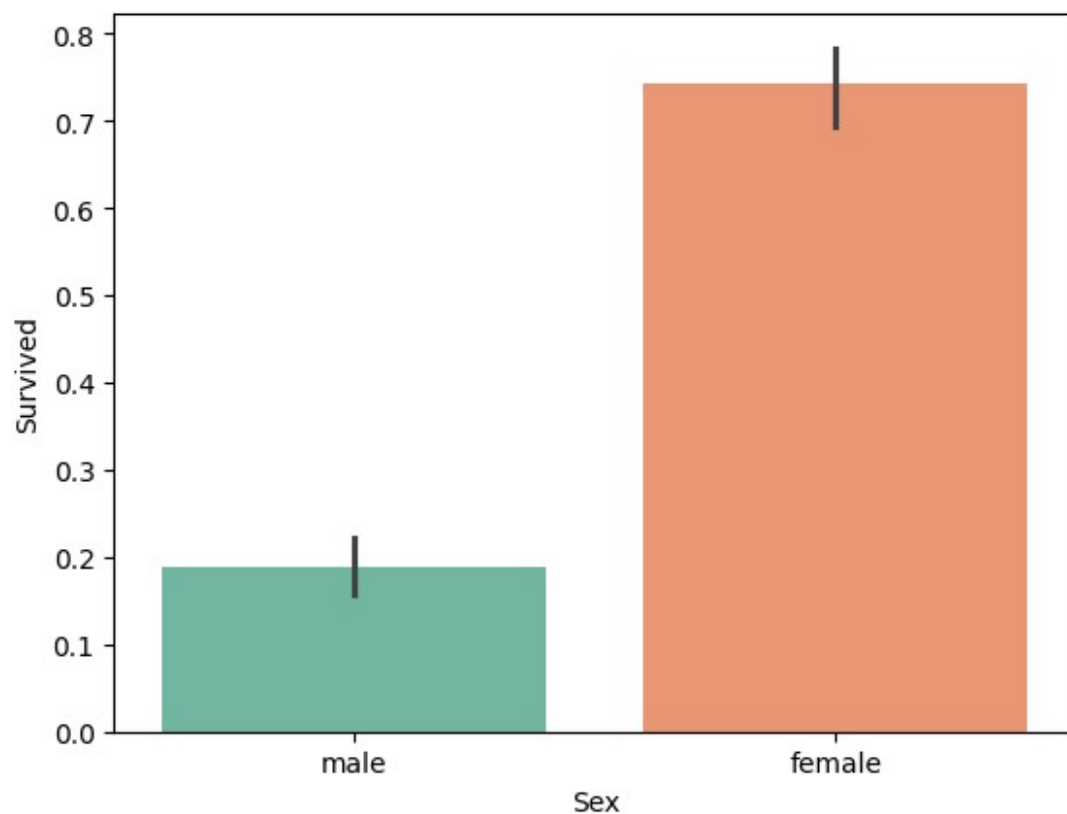
Passengers who are single or have 1 or 2 siblings are more likely to survive.

```
ageplot = sns.FacetGrid(titanic_df, col = 'Survived')  
plt.figure(figsize = (10,8))  
ageplot = ageplot.map(sns.distplot,'Age')  
ageplot = ageplot.set_ylabels('Survival Chance')
```



<Figure size 1000x800 with 0 Axes>

```
genderplot = sns.barplot(x = 'Sex', y = 'Survived', data =
titanic_df, palette="Set2")
plt.figure(figsize = (10,2))
plt.show()
```





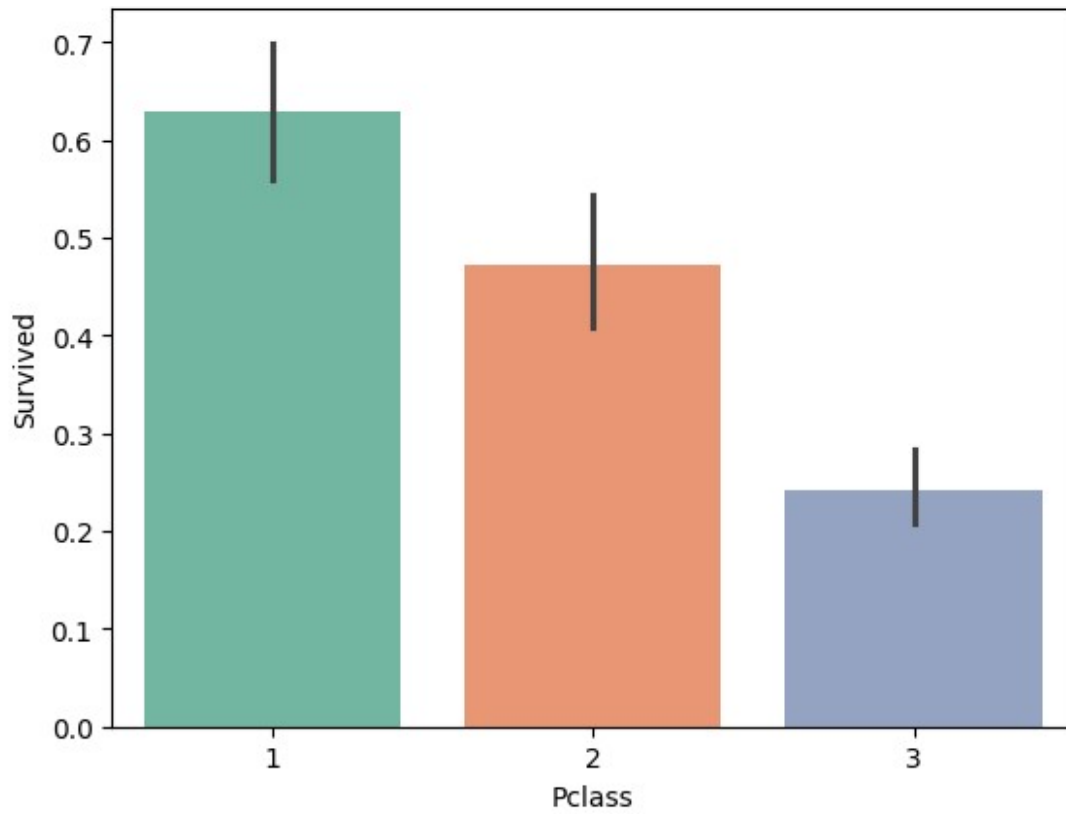
<Figure size 1000x200 with 0 Axes>

```
titanic_df[["Sex", "Survived"]].groupby("Sex").mean()
```

```
{"summary": "{\n  \"name\": \"titanic_df[[\\\"Sex\\\",\\\"Survived\\\"]]\",\n  \"rows\": 2,\n  \"fields\": [\n    {\n      \"column\": \"Sex\",\n      \"properties\": {\n        \"dtype\": \"string\",\n        \"num_unique_values\": 2,\n        \"samples\": [\n          \"male\",\n          \"female\"\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Survived\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 0.391122024068098,\n        \"min\": 0.18890814558058924,\n        \"max\": 0.7420382165605095,\n        \"num_unique_values\": 2,\n        \"samples\": [\n          0.18890814558058924,\n          0.7420382165605095\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    }\n  ],\n  \"type\": \"dataframe\"}
```

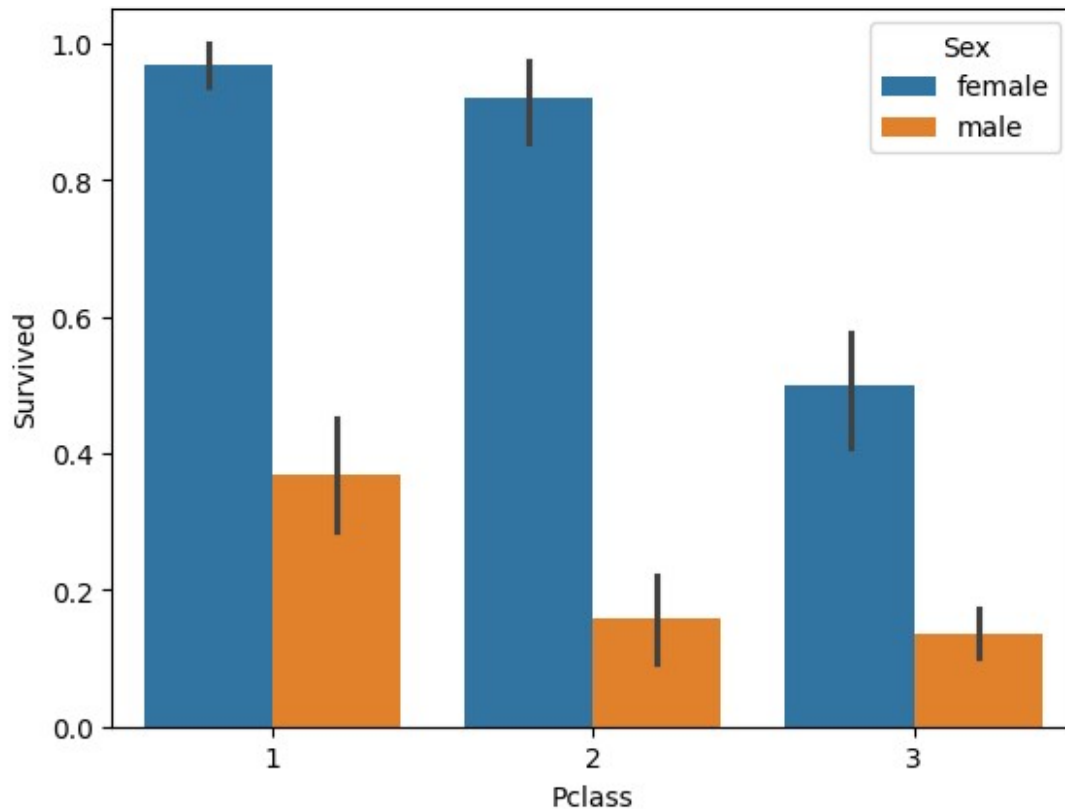
we can see that Female has Higher Survival Rates than Male, as females were rescued first in the lifeboats. So Sex played an important role during the evacuation process and hence women had more chances of survival.

```
Pclassplot = sns.barplot(x = "Pclass", y = "Survived", data =  
titanic_df, palette="Set2")  
plt.show()
```



The Passenger class increases, hence the chances of survival increases.

```
x = sns.barplot(x = "Pclass",y = "Survived",data = titanic_df, hue =  
"Sex")  
plt.show()
```



Passenger having high class ticket(1st class) and also if gender is female , then most chances of survival has increased.

### Conclusion:

Unequal Fate: Social Class and Demographics Shaped Survival on the Titanic dataset.

Our analysis reveals that survival on the RMS Titanic was far from random. Socio-economic factors, particularly passenger class, played a significant role alongside demographics like age, gender, and family ties.

### Key Findings:

Gender and Age: Women and children exhibited higher survival rates, suggesting potential prioritization during the evacuation.

Social Class Divide: A clear correlation emerged between passenger class and survival. Passengers in the higher classes (first class) had a significantly better chance of surviving the disaster.

By highlighting these disparities, our investigation underscores the profound impact of social status, class, and demographics on survival outcomes during maritime tragedies like the Titanic disaster.