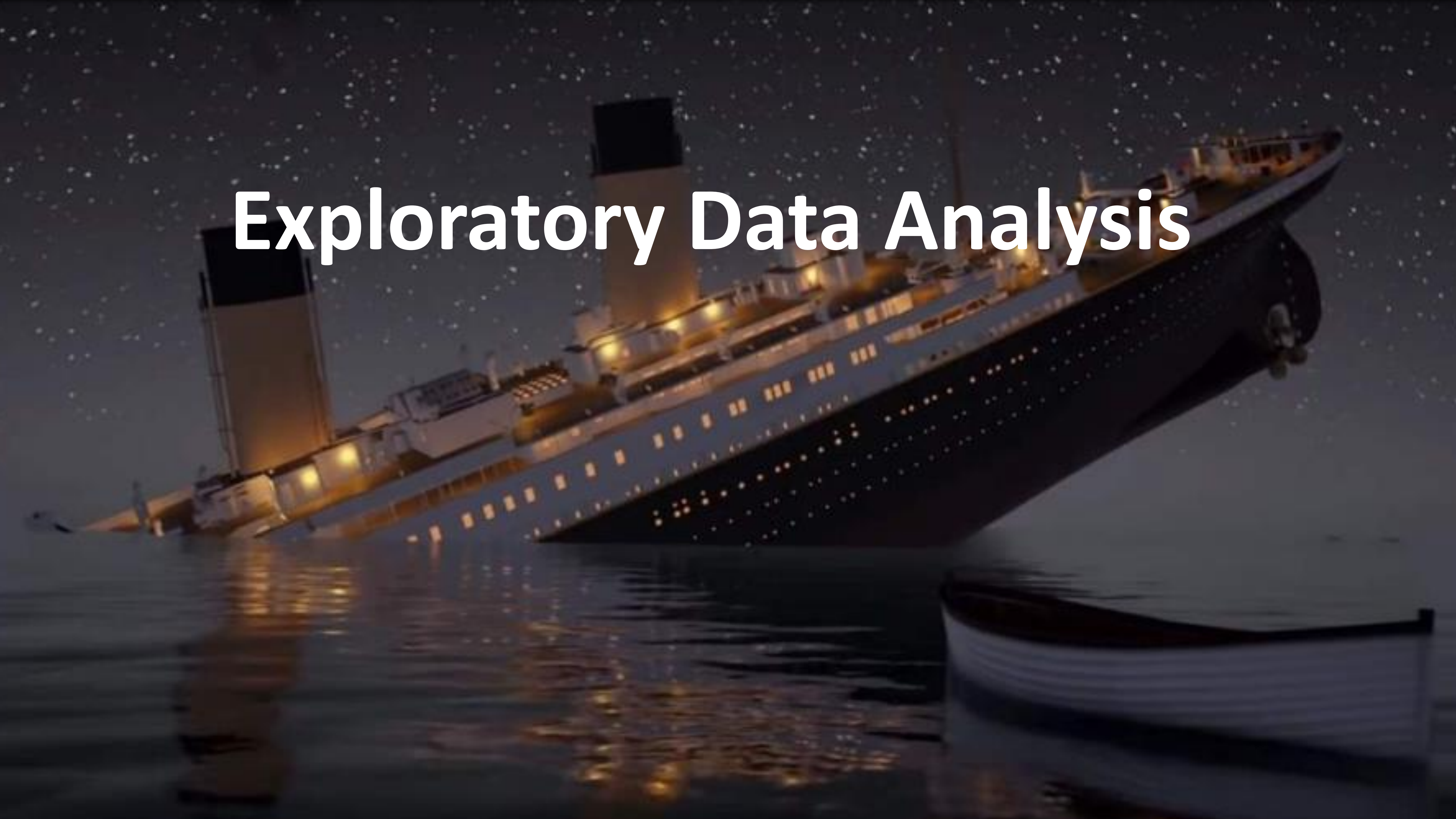
Exploratory Data Analysis

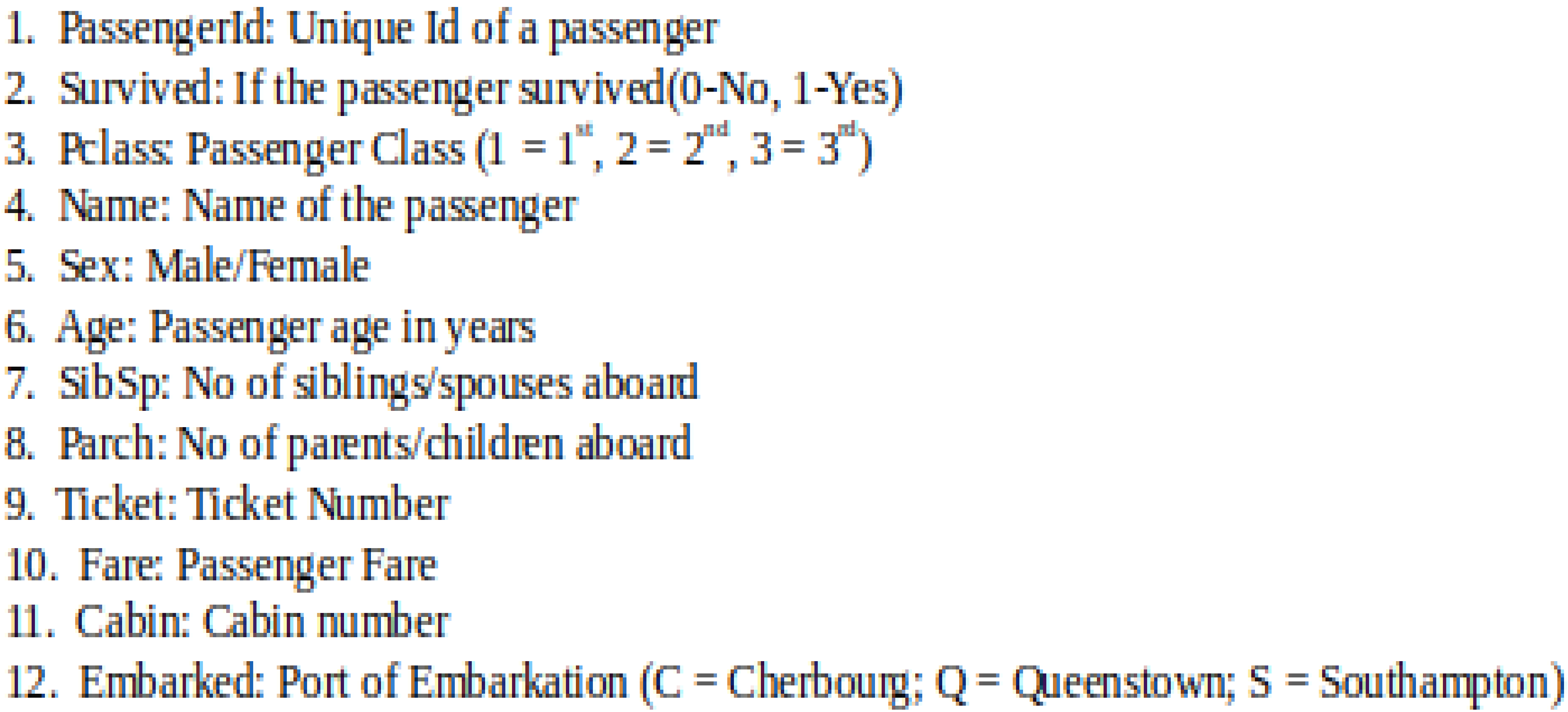


INTRODUCTION TO DATASET

Titanic Dataset –

It is one of the most popular datasets used for understanding machine learning basics. It contains information of all the passengers aboard the RMS Titanic, which unfortunately was shipwrecked. It is the purpose of this EDA to explain the impact of sex, passenger class, and age on a person's likelihood of surviving the shipwreck.

The csv file can be downloaded from Kaggle.



1. PassengerId: Unique Id of a passenger
2. Survived: If the passenger survived(0-No, 1-Yes)
3. Pclass: Passenger Class (1 = 1st, 2 = 2nd, 3 = 3rd)
4. Name: Name of the passenger
5. Sex: Male/Female
6. Age: Passenger age in years
7. SibSp: No of siblings/spouses aboard
8. Parch: No of parents/children aboard
9. Ticket: Ticket Number
10. Fare: Passenger Fare
11. Cabin: Cabin number
12. Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Data Set feature description:

1. PassengerId: An unique index for each passenger.
2. Survived: This shows if the passenger survived or not. '1' stands for survived and '0' stands for not survived.
3. Pclass: Ticket class:'1' stands for First class ticket,'2' stands for Second class ticket,'3' stands for Third class ticket.It is in a way proxy for socio-economic status (SES) :1st = Upper,2nd = Middle,3rd = Lower
4. Name: Passenger's name. The name also contain a title. "Mr." for a man. "Mrs." for a woman. "Miss" for a girl. "Master" for a boy.
5. Sex: Passenger's sex. It's either Male or Female.
6. Age: Passenger's age. "NaN" values in this column indicate that the age of that particular passenger has not been recorded.

7. SibSp: Number of siblings or spouses traveling with each passenger. The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored).

8. Parch: Number of parents of children traveling with each passenger. The dataset defines family relations in this way... Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children traveled only with a nanny, therefore parch=0 for them.

9. Ticket: Ticket number.

10. 10. Fare: The amount of money the passenger has paid for the travel journey(in dollars).

11. Cabin: Cabin number of the passenger. "NaN" values in this column indicate that the cabin number of that particular passenger has not been recorded.

12. Embarked: Port from where the particular passenger was embarked/boarded. Here: C = Cherbourg, Q = Queenstown, S = Southampton.

TOOLS USED FOR ANALYSIS AND VISUALIZATION

Python: An interpreted, object-oriented programming language with dynamic semantics. Its high level, built-in data structures, combined with dynamic binding, make it very attractive for rapid application development, also as to be used as a scripting or glue language to attach existing components together. Python and EDA are often used together to spot missing values in the data set, which is vital so you'll decide the way to handle missing values for machine learning.

Seaborn:
It is a python library used to statistically visualize data. Seaborn, built over Matplotlib, provides a better interface and ease of usage. It can be installed using the following command,

```
pip3 install seaborn
```

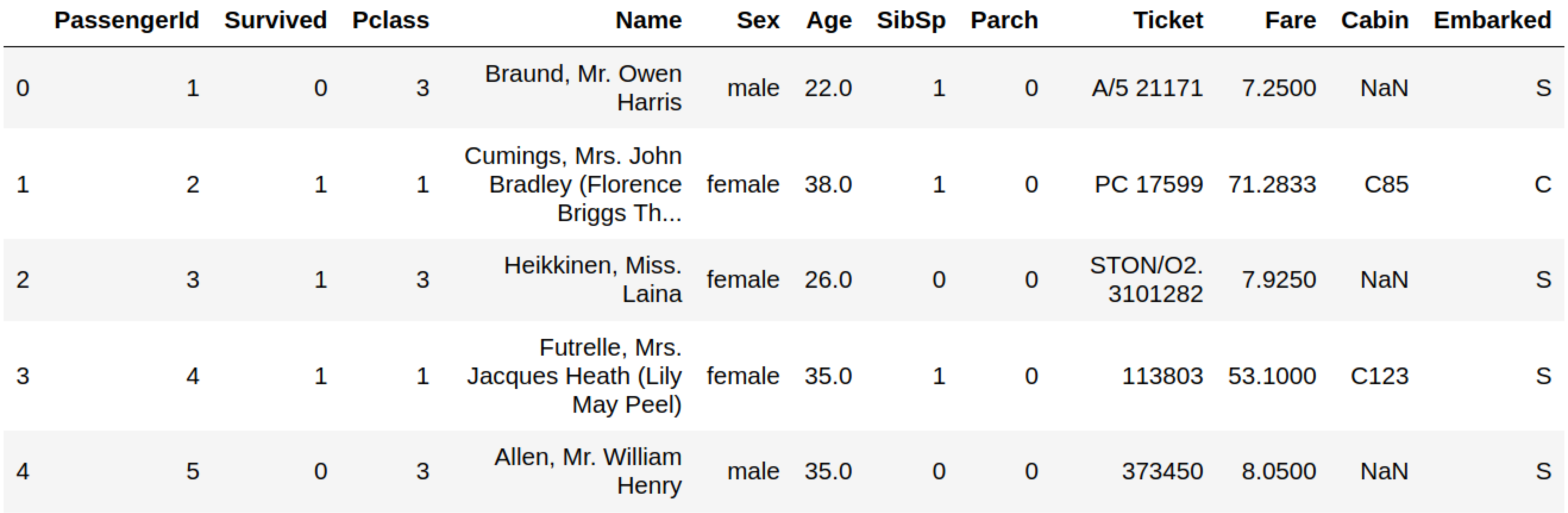
INITIAL HYPOTHESIS

The first hypothesis is that the passengers had higher chances of survival if:

- they had a high class ticket
- they were women
- they were young

LOADING THE DATASET

```
  import pandas as pd

  import numpy as np

  import matplotlib.pyplot as plt

  import seaborn as sns

  %matplotlib inline


df=pd.read_csv("C:\\Users\\shres\\OneDrive\\Documents\\titanic data set 2\\titanic_dataset.csv")

df.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

df.describe()

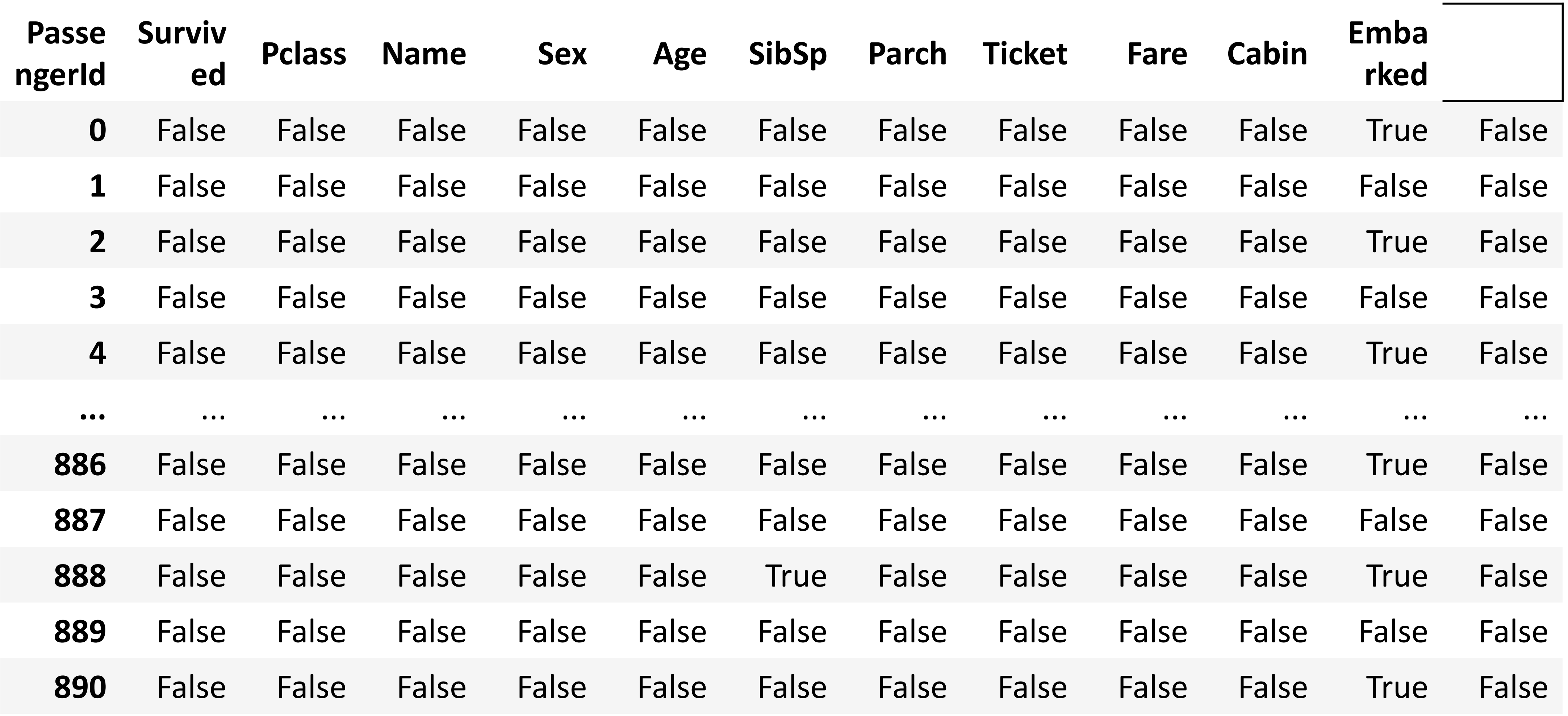
PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	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 CLEANING

1) Missing data

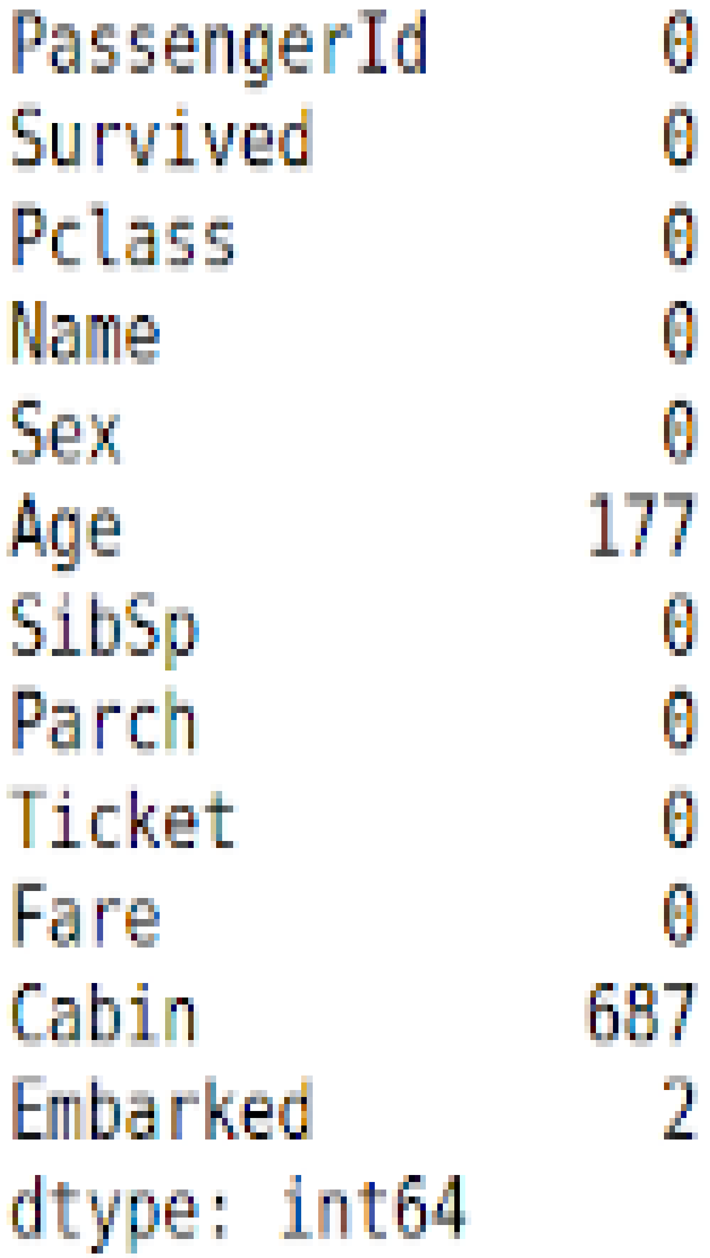
We can use seaborn to create a simple heatmap to see where we are missing data.

df.isnull()



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False	False	True	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	True	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	True	False

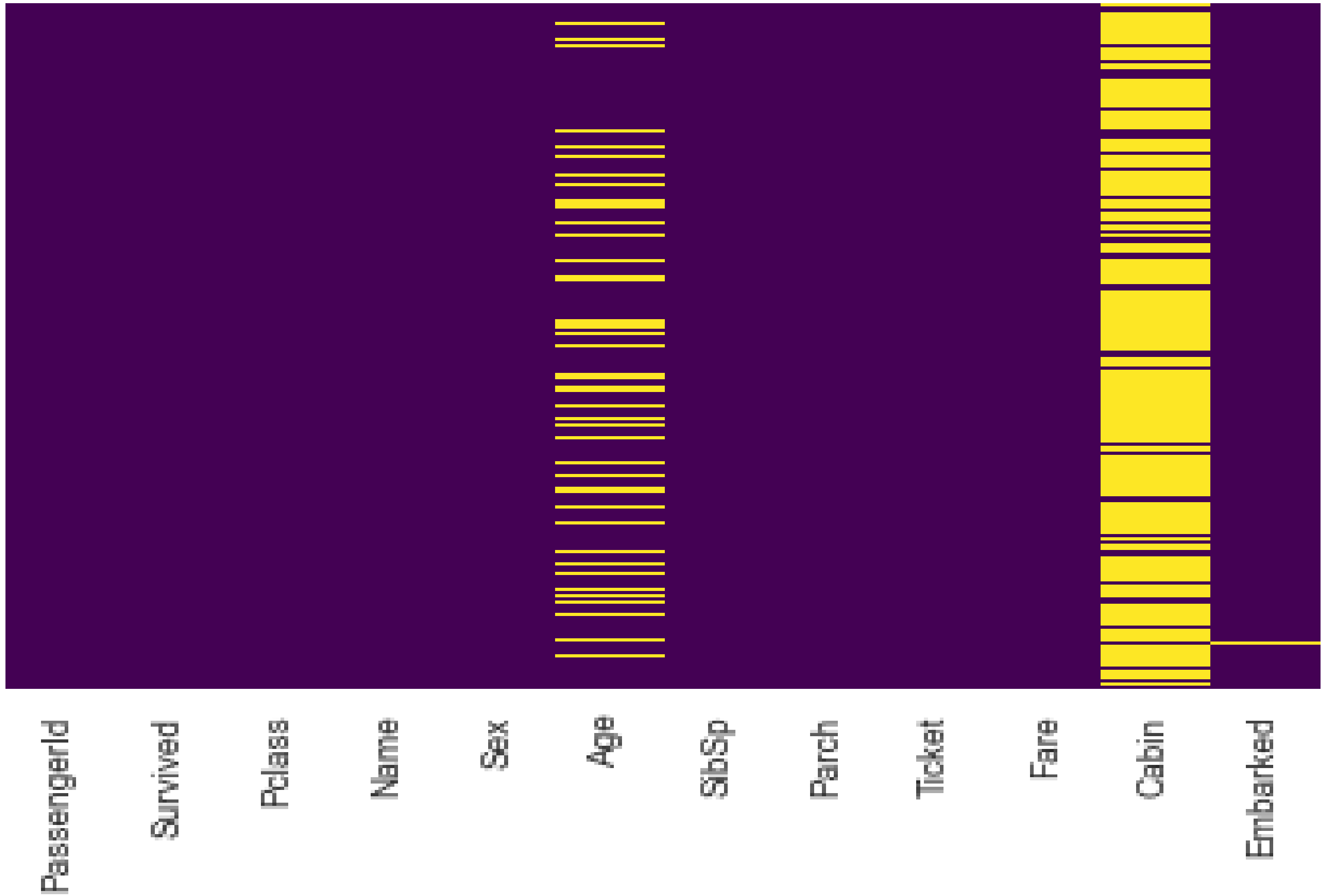
df.isnull().sum()



```
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
```

sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap="viridis“)

Roughly 20 percent of the Age data is missing. The proportion of Age missing is likely small enough for reasonable replacement with some form of imputation. Looking at the Cabin column, it looks like we are missing too much of that data to do something useful with at a basic level it can be dropped or it can be changed into another feature like "Cabin Known:1 or 0"



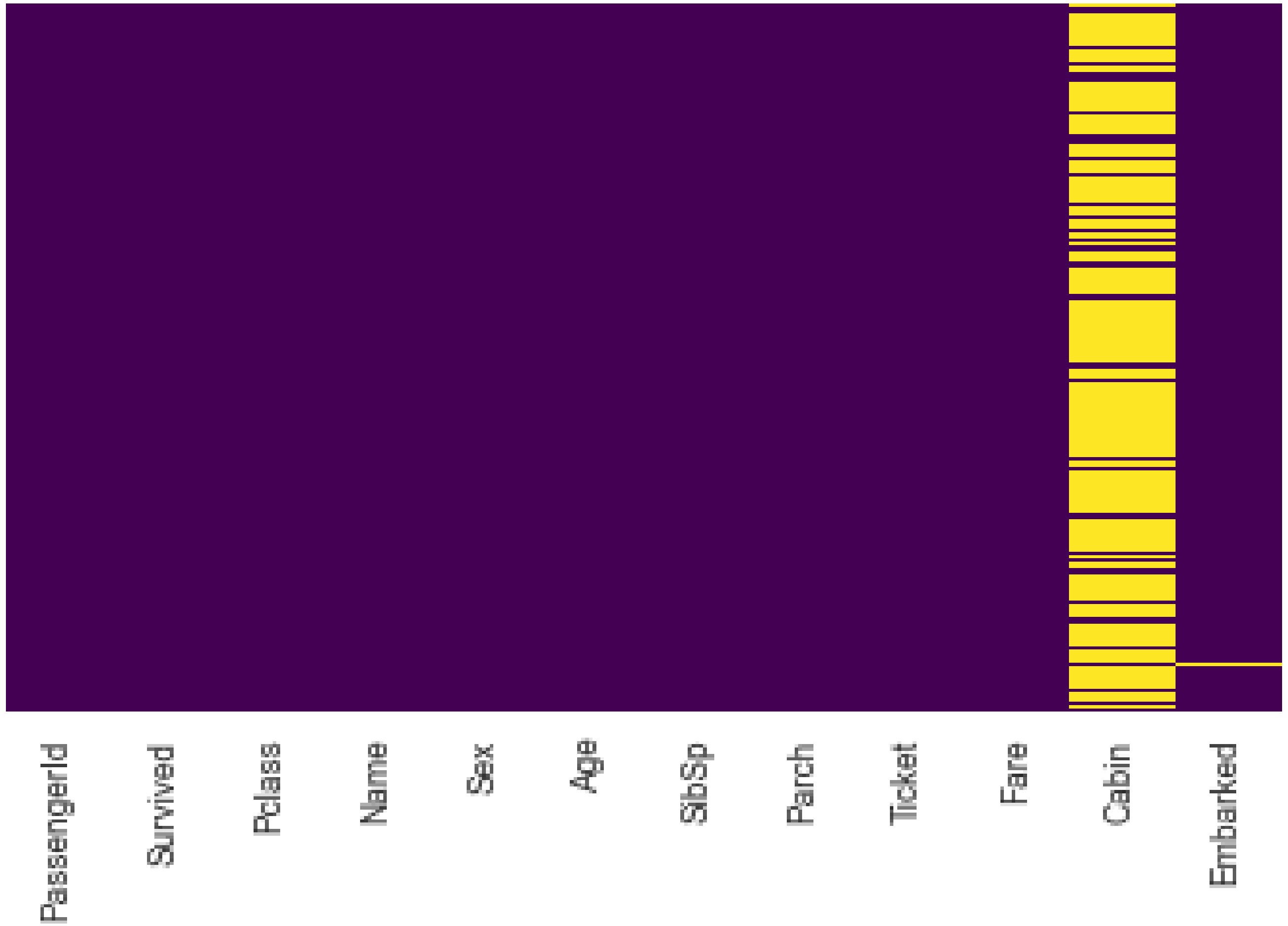
2) Filling Missing Values:

```
def impute_age(cols):
  Age = cols[0]
  Pclass = cols[1]
 if pd.isnull(Age):
      if Pclass == 1:
         return 37
      elif Pclass == 2:
         return 29
      else:
         return 24
 else:
     return Age
```

```
df["Age"]=df[["Age","Pclass"]].apply(impute_age,axis=1)
```

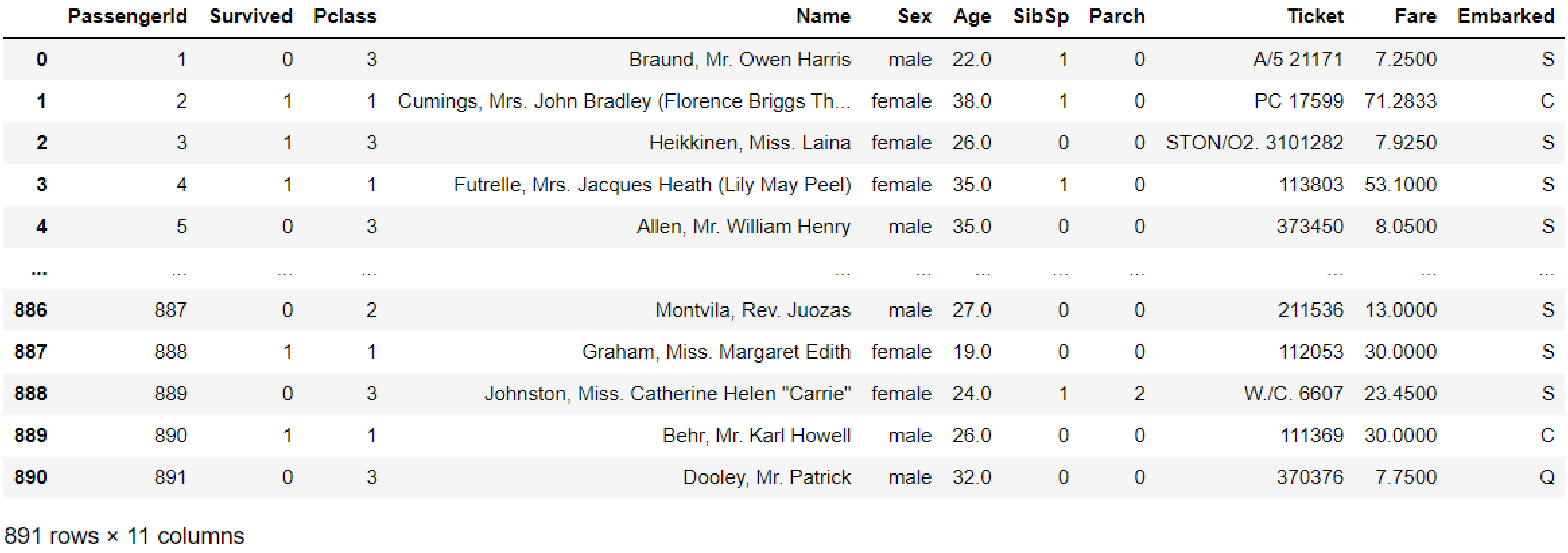
Now let's check heatmap to see if we are missing Age data.

```
sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap="viridis")
```



Now the Cabin cloumn is dropped and also the rows in Embarked that is NaN.

```
df.drop("Cabin",axis=1,inplace=True)
df
```



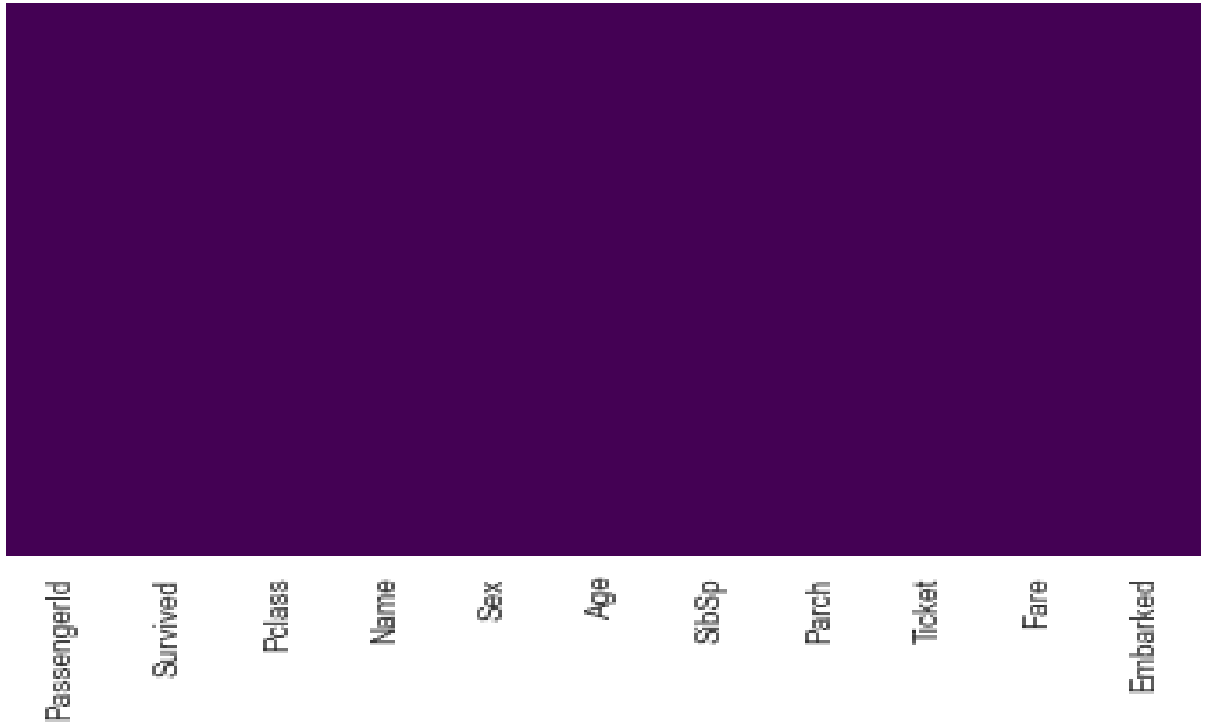
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	24.0	1	2	W./C. 6607	23.4500	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Q

891 rows × 11 columns

```
df.dropna(inplace=True)
```

Now let's check heatmap to see if we have missing data.

```
sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap="viridis")
```



There are no more NaN values in the df

PRELIMINARY QUESTIONS

1) Survived vs Not Survived

2) **Correlation between survival and sex**

3) **Correlation between survival and age**

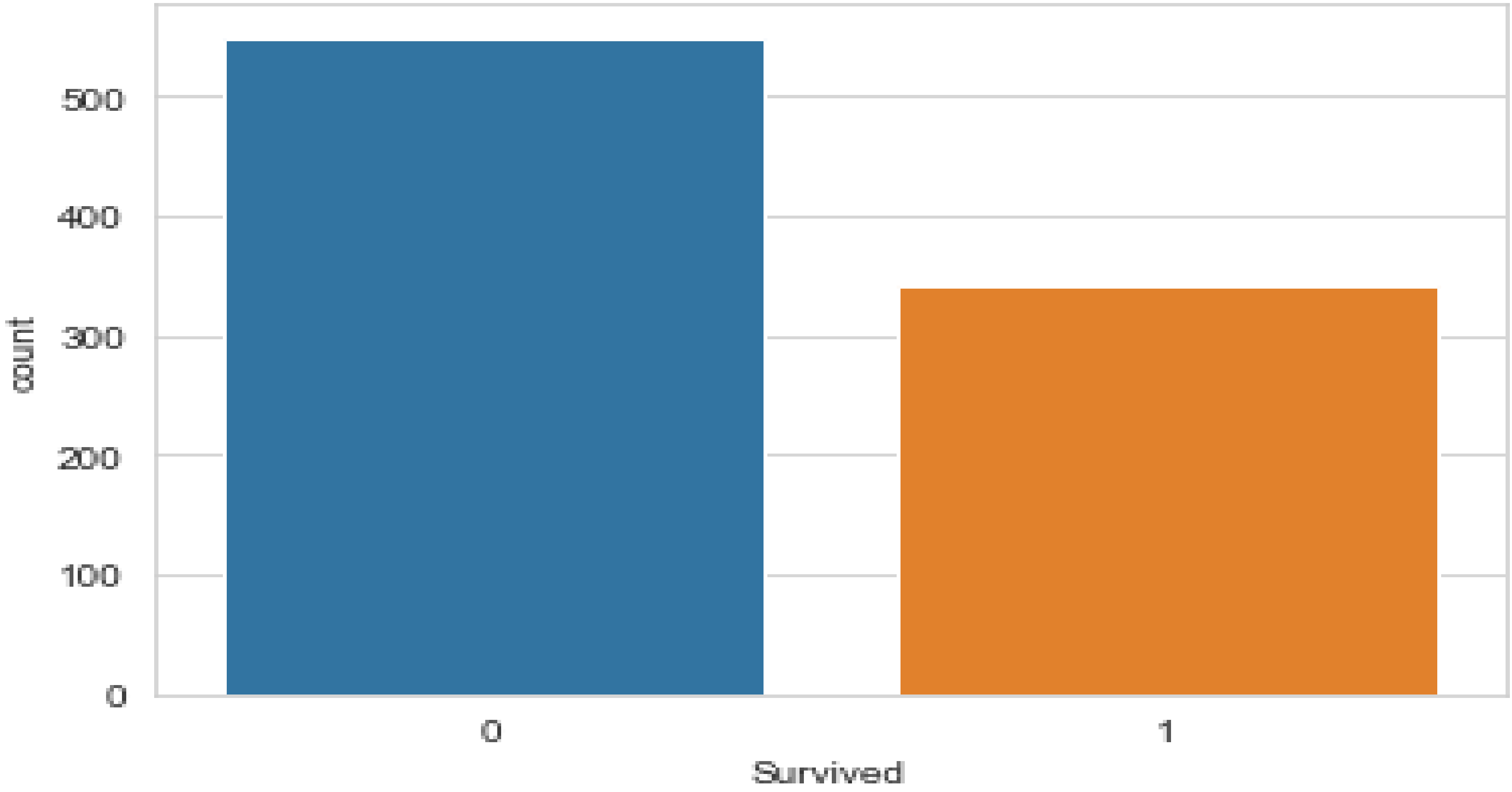
4) Average age by Passenger Class

5) Passenger class distribution –survived vs not survived

EXPLORATORY ANALYSIS AND VISUALIZATION

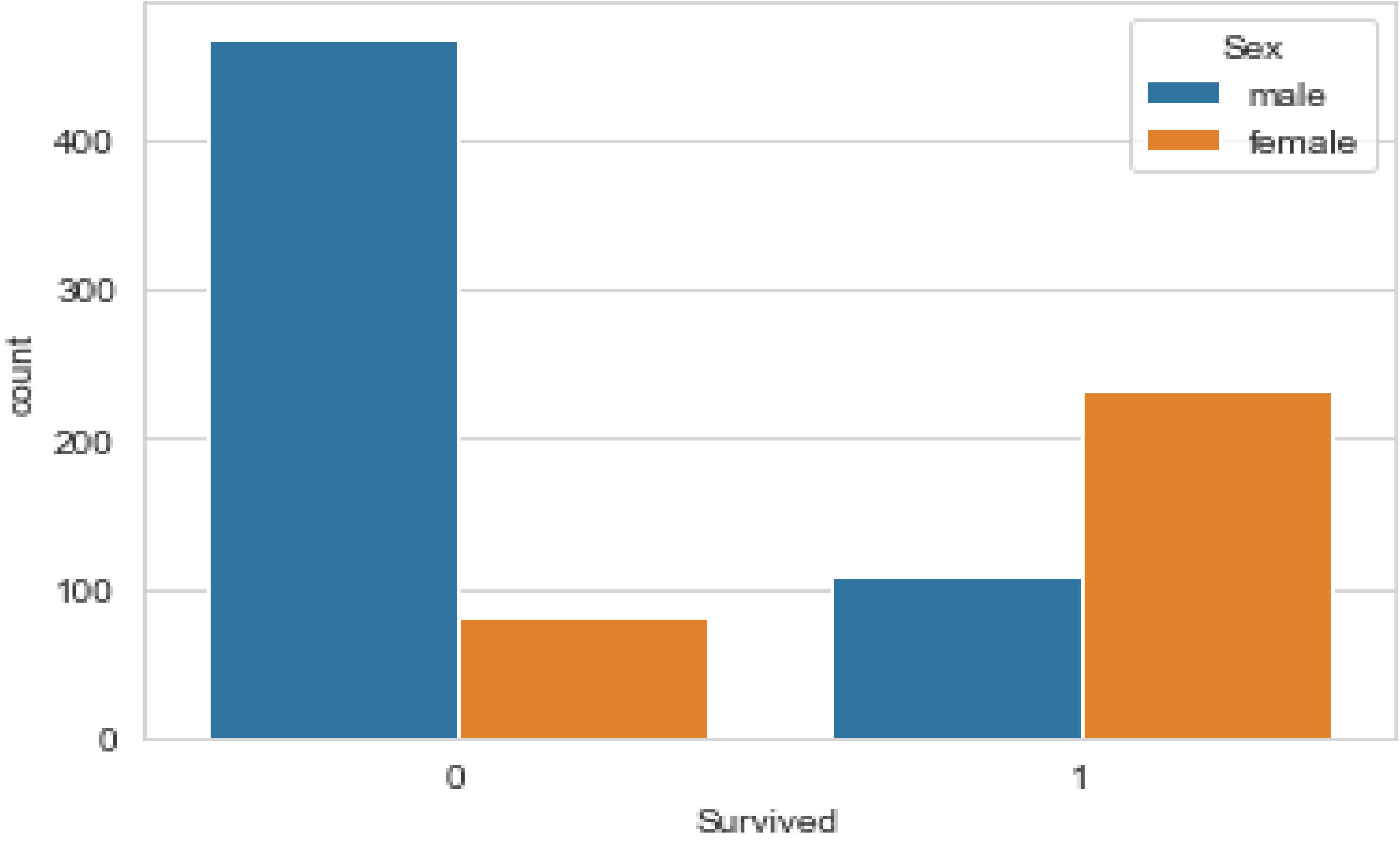
1) Survived vs Not Survived

```
sns.set_style("whitegrid")
sns.countplot(x="Survived",data=df)
```



2) Correlation between survival and sex

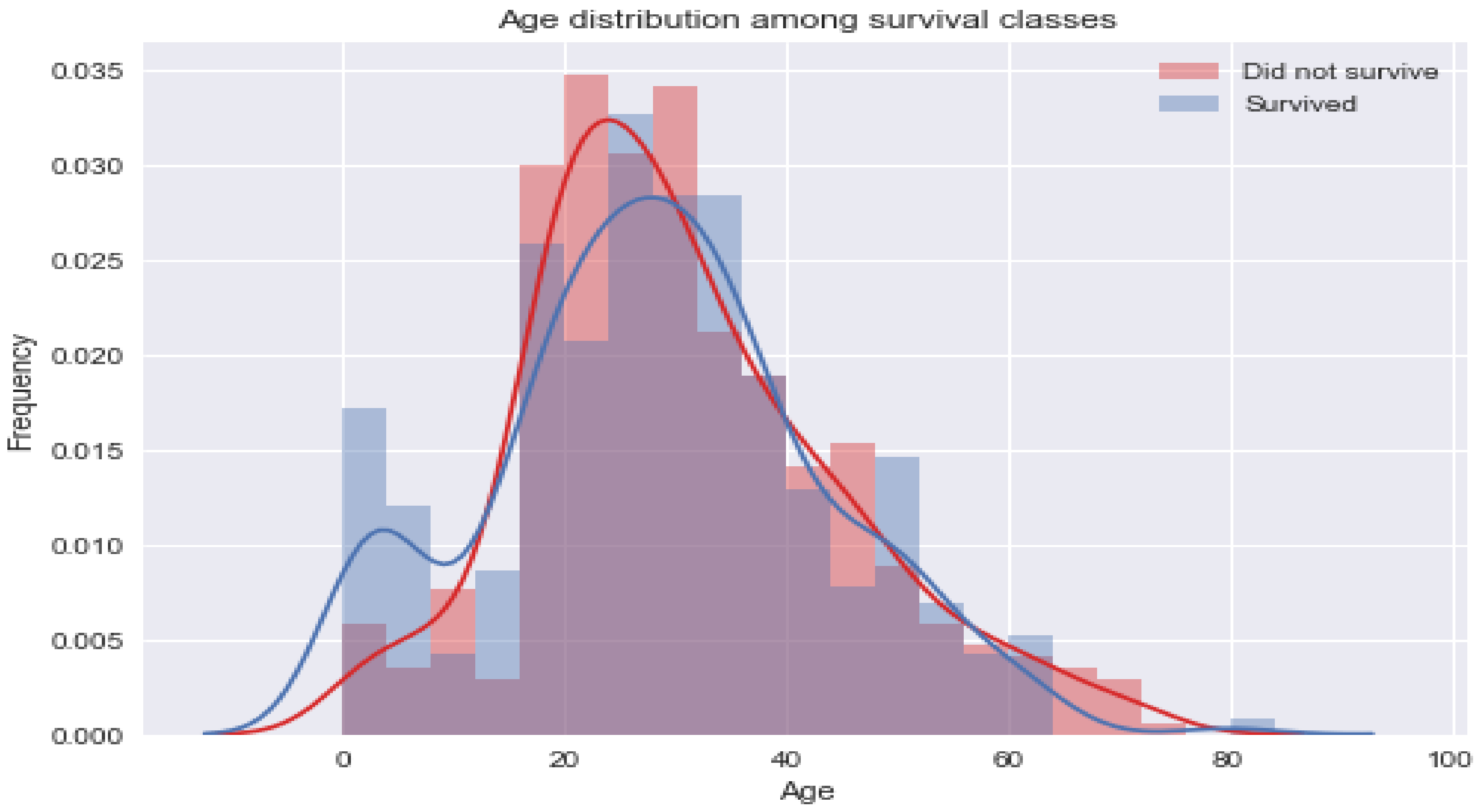
```
sns.set_style("whitegrid")
sns.countplot(x="Survived",hue="Sex",data=df)
```



As previously mentioned, women are much more likely to survive than men. **74% of the women survived, while only 18% of men survived.**

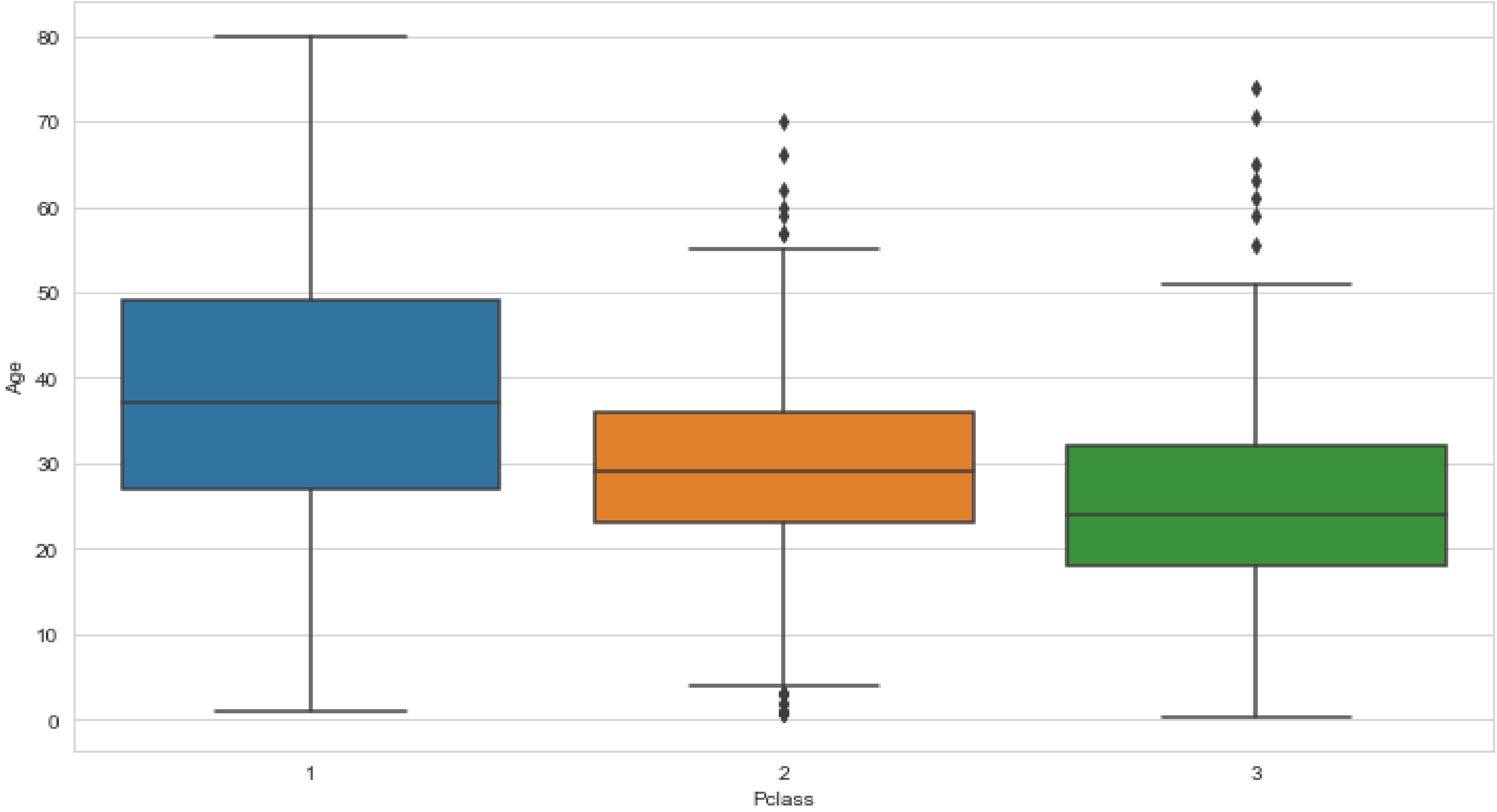
3) Correlation between survival and age

```
age_bins = np.arange(0, 100, 4)
sns.distplot(df.loc[(df['Survived']==0) & (~df['Age'].isnull()),'Age'], bins=age_bins, color='#d62728’)
sns.distplot(df.loc[(df['Survived']==1) & (~df['Age'].isnull()),'Age'], bins=age_bins)plt.title('Age distribution among survival classes')plt.ylabel('Frequency')plt.legend(['Did not survive', 'Survived'])plt.show()
```



4) Average age by Passenger Class

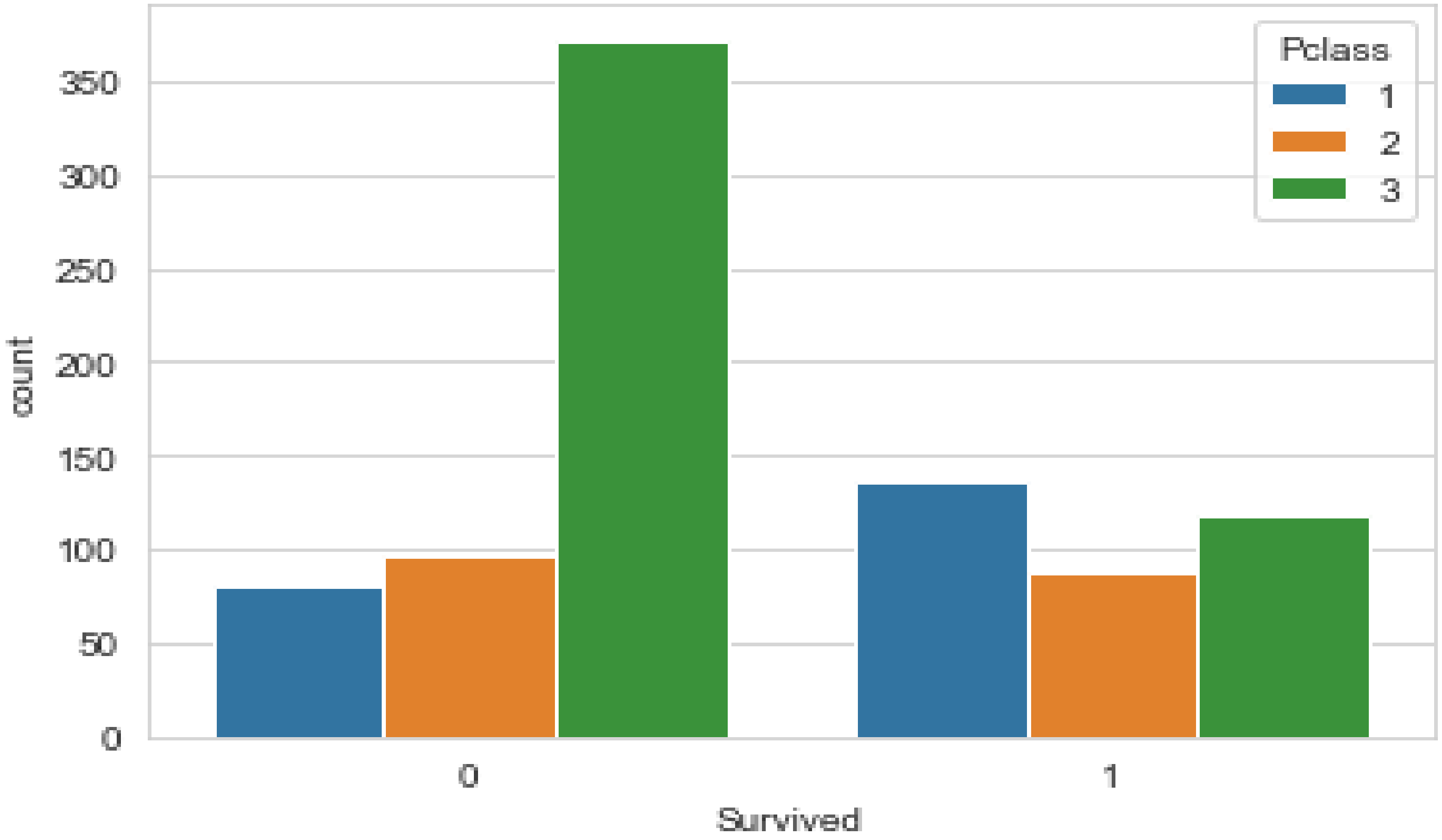
```
plt.figure(figsize=(12,7))
sns.boxplot(x="Pclass",y="Age",data=df)
```



We can see the wealthier passengers in the higher classes tend to be older

5) Passenger class distribution –survived vs not survived

```
sns.set_style("whitegrid")
sns.countplot(x="Survived",hue="Pclass",data=df)
```



The graphs above clearly shows that **economic status (Pclass)** played an important role regarding the potential survival of the Titanic passengers. First class passengers had a much higher chance of survival than passengers in the 3rd class.

- •63% of the 1st class passengers survived the Titanic wreck
- •48% of the 2nd class passenger survived
- •Only 24% of the 3rd class passengers survived

FURTHER INSIGHTS GAINED

- Whether a passenger is a male or a female plays an important role in determining if one is going to survive.
- Higher-class passengers had more survival rate than the lower class.
- Age has a high negative correlation with number of siblings.
- Children below 18 years of age have higher chances of surviving.
- **From Count plot of Sibsp we conclude that maximum people did not have siblings or spouse**

FINAL HYPOTHESIS

Overall, the hypothesis is proven.

- We have proved women have more survival rate sex wise
- First class passengers have more survival rate compared to second and third class
- Younger passengers have more survival rate age wise

Thank you