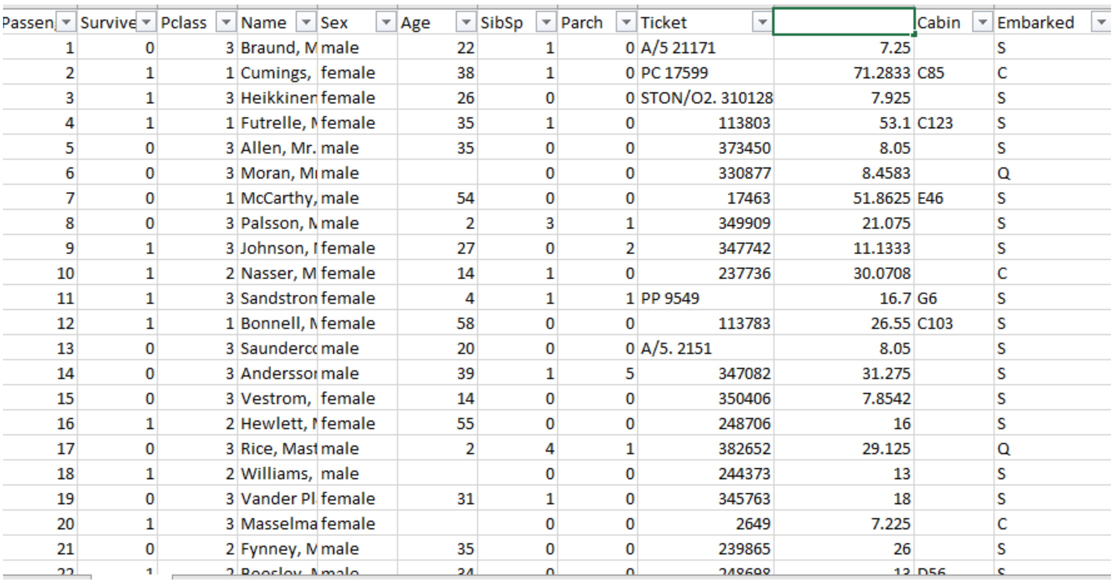
DATA ANALYST INTERNSHIP

Task 4: Exploratory Data Analysis (EDA)

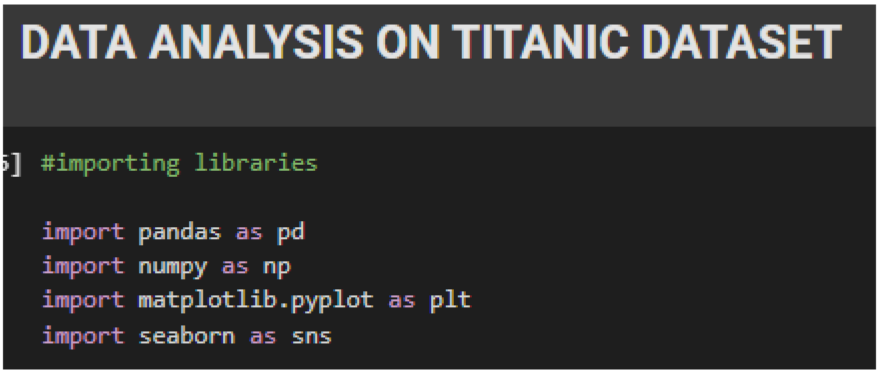
Selected Dataset: Titanic Dataset

Raw Data:



Passen	Survive	Pclass	Name	Sex	Age	SibSp	Parch	Ticket		Cabin	Embarked
1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings,	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkiner	female	26	0	0	STON/O2. 310128	7.925		S
4	1	1	Futrelle, N	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mı	male		0	0	330877	8.4583		Q
7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, N	male	2	3	1	349909	21.075		S
9	1	3	Johnson, I	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
11	1	3	Sandstron	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, N	female	58	0	0	113783	26.55	C103	S
13	0	3	Saunderco	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Anderssoı	male	39	1	5	347082	31.275		S
15	0	3	Vestrom,	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, I	female	55	0	0	248706	16		S
17	0	3	Rice, Mast	male	2	4	1	382652	29.125		Q
18	1	2	Williams,	male		0	0	244373	13		S
19	0	3	Vander Pl	female	31	1	0	345763	18		S
20	1	3	Masselma	female		0	0	2649	7.225		C
21	0	2	Fynney, M	male	35	0	0	239865	26		S
22	1	2	Beesley, M	male	34	0	0	248698	13	D56	S

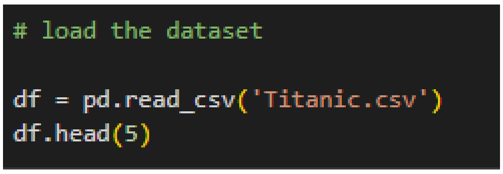
Tools used: Python (Pandas, Matplotlib, Seaborn)



DATA ANALYSIS ON TITANIC DATASET

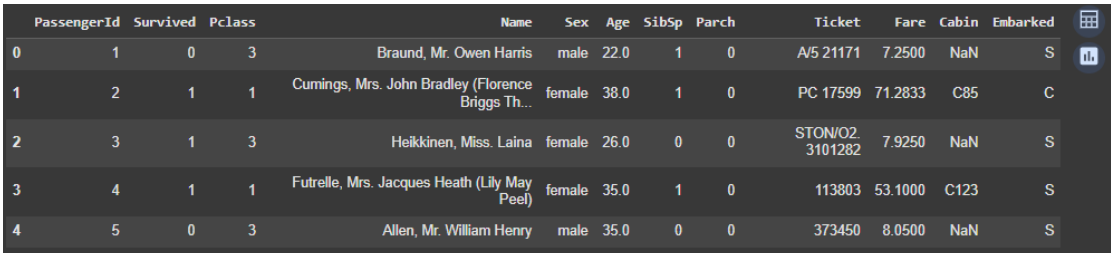
```
5] #importing libraries

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

Load the dataset



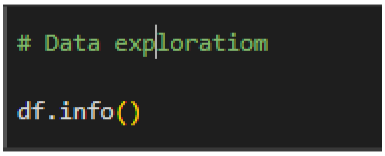
```
# load the dataset

df = pd.read_csv('Titanic.csv')
df.head(5)
```

Overview:



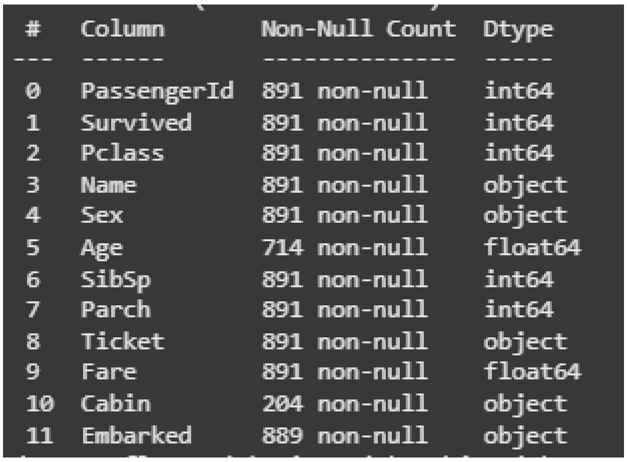
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 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

Data Exploration:

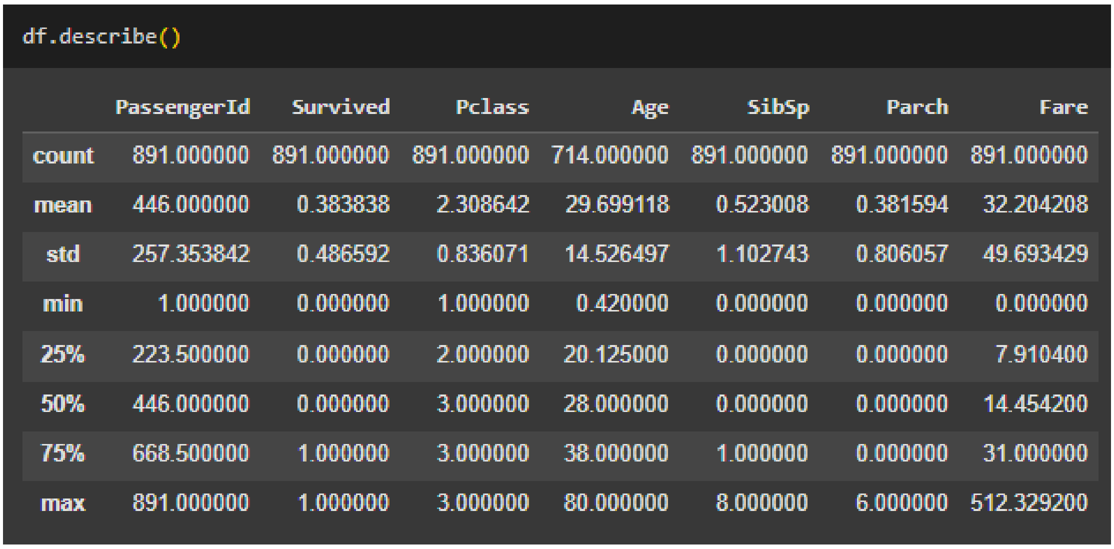


```
# Data exploratiom

df.info()
```

Output:



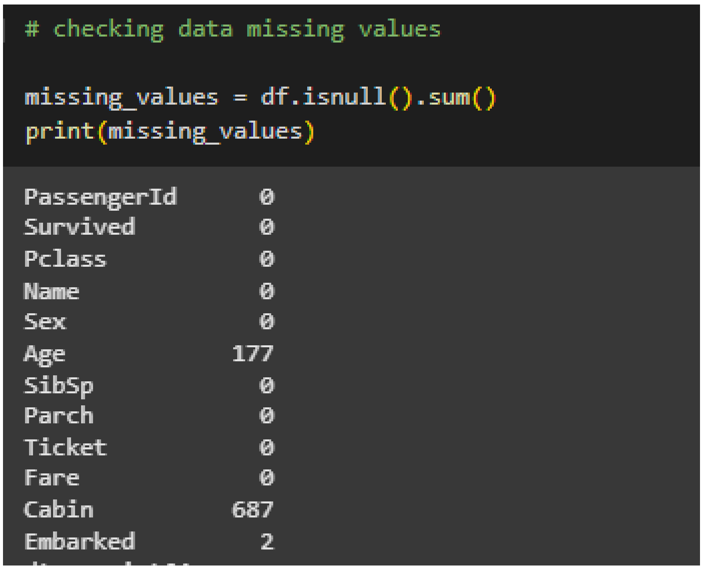
```
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          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
```



```
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

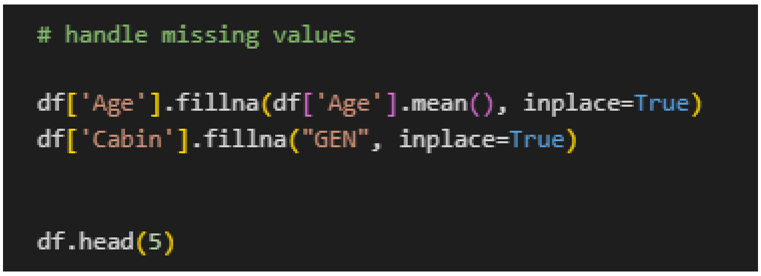
Missing Values count:



```
# checking data missing values

missing_values = df.isnull().sum()
print(missing_values)
```

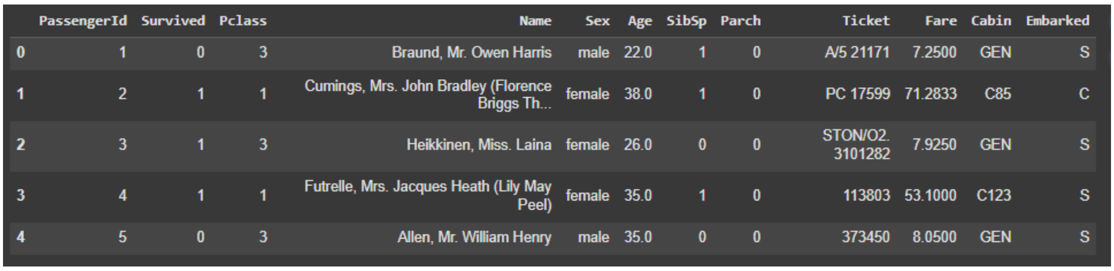
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
```

Handle Missing values:



```
# handle missing values

df['Age'].fillna(df['Age'].mean(), inplace=True)
df['Cabin'].fillna("GEN", inplace=True)


df.head(5)
```

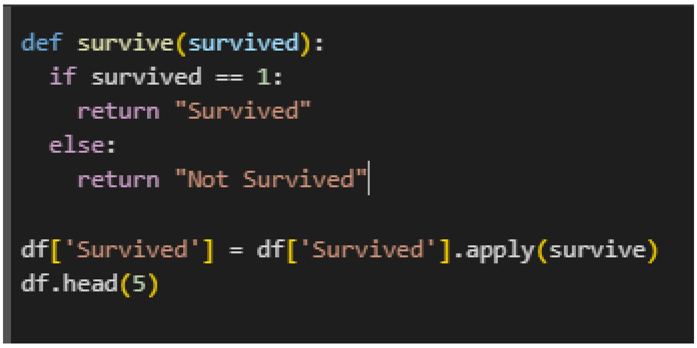
Overview of Modified Data:



	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	GEN	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	GEN	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	GEN	S

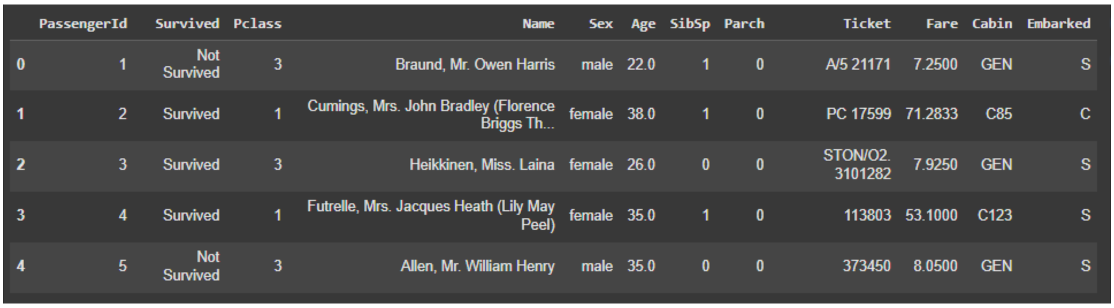
Exploratory Data Analysis:

Data Modification:



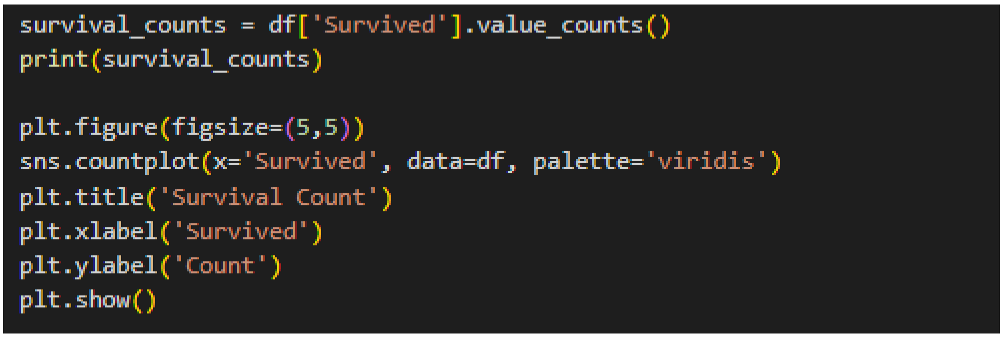
```python
def survive(survived):
  if survived == 1:
    return "Survived"
  else:
    return "Not Survived"

df['Survived'] = df['Survived'].apply(survive)
df.head(5)
```

Output:

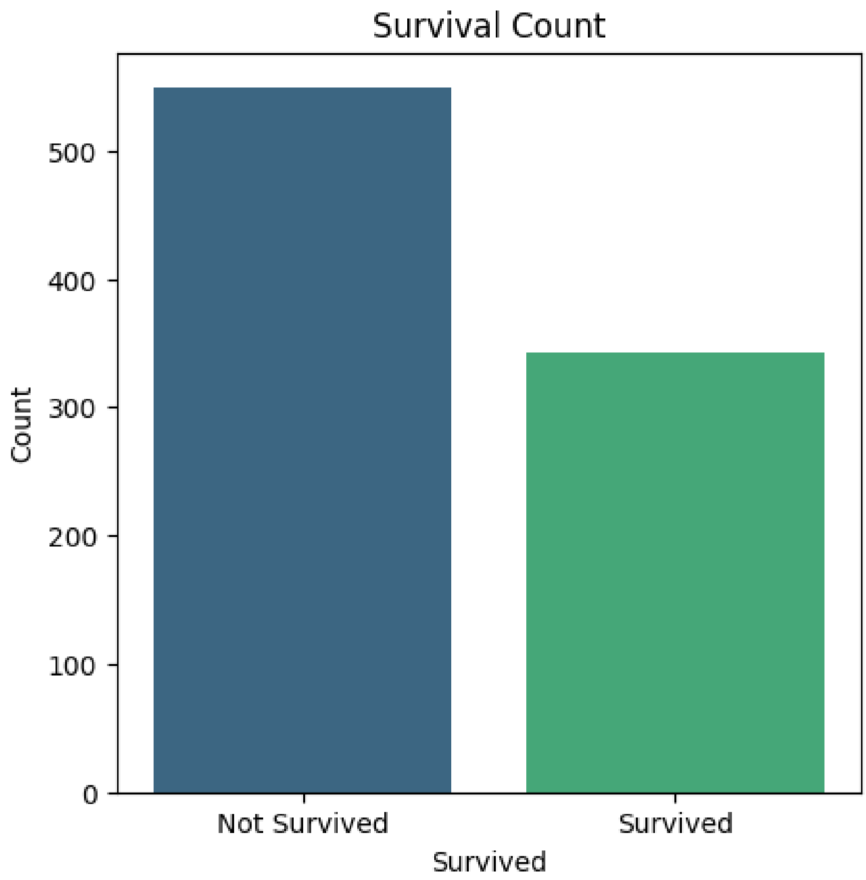


	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	Not Survived	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	GEN	S
1	2	Survived	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	Survived	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	GEN	S
3	4	Survived	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	Not Survived	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	GEN	S

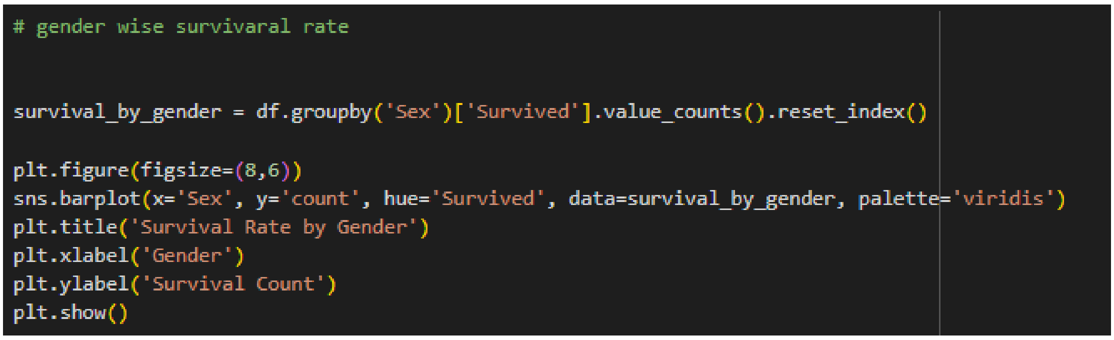
Survival Counts:



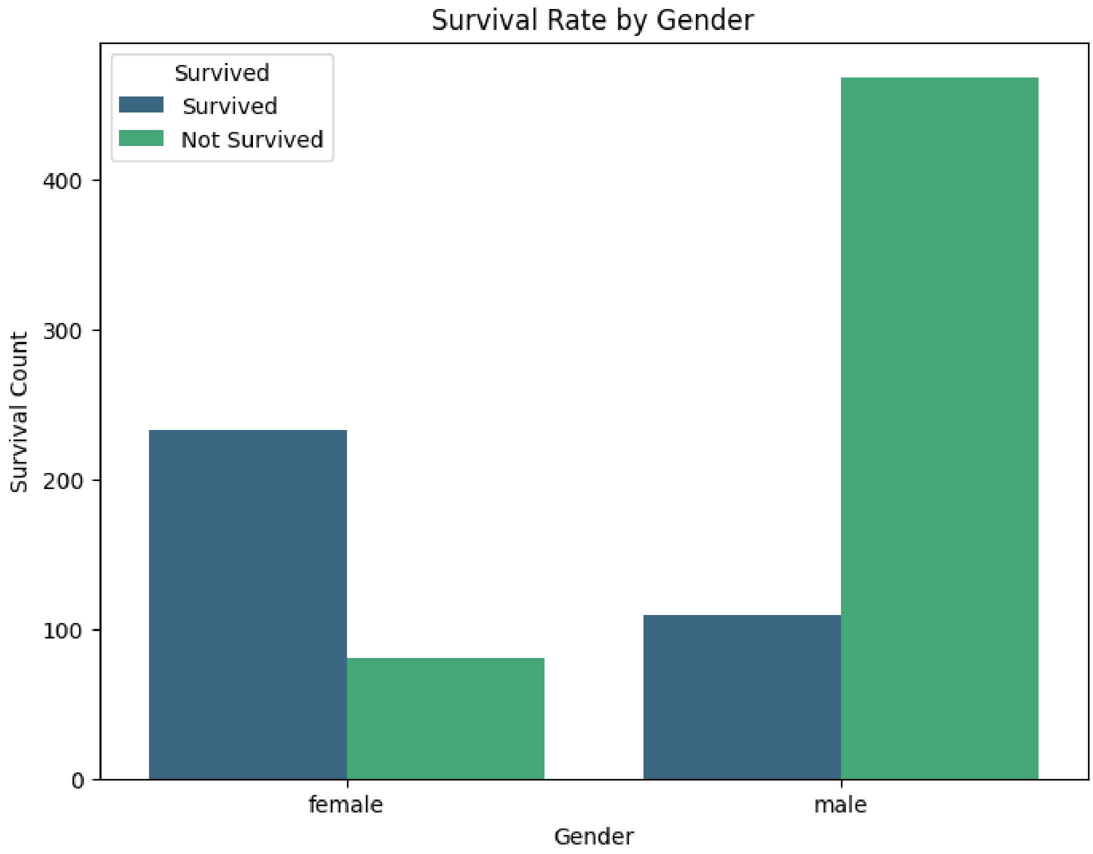
```python
survival_counts = df['Survived'].value_counts()
print(survival_counts)

plt.figure(figsize=(5,5))
sns.countplot(x='Survived', data=df, palette='viridis')
plt.title('Survival Count')
plt.xlabel('Survived')
plt.ylabel('Count')
plt.show()
```



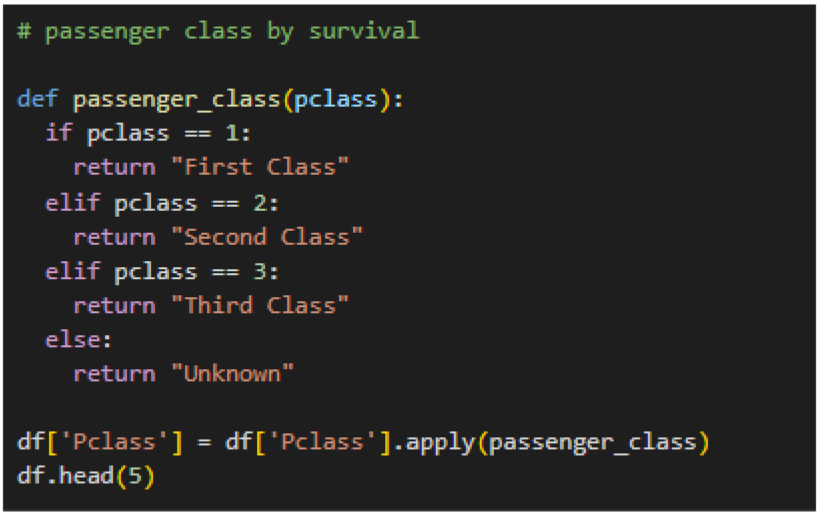
Gender wise survival rate:



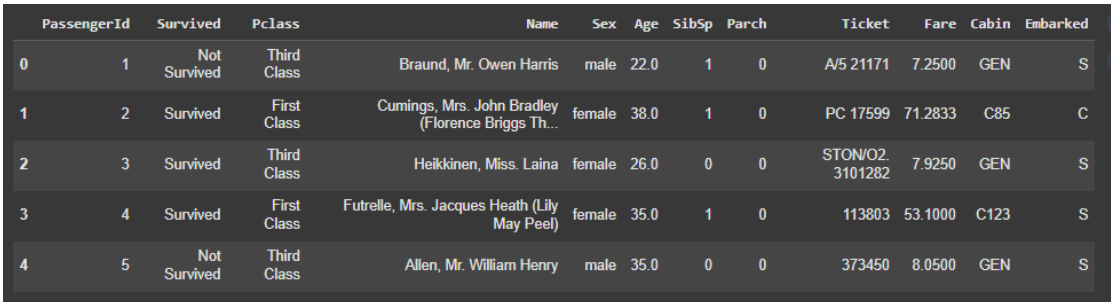
```
# gender wise survivaral rate


survival_by_gender = df.groupby('Sex')['Survived'].value_counts().reset_index()

plt.figure(figsize=(8,6))
sns.barplot(x='Sex', y='count', hue='Survived', data=survival_by_gender, palette='viridis')
plt.title('Survival Rate by Gender')
plt.xlabel('Gender')
plt.ylabel('Survival Count')
plt.show()
```



Passenger Class and Survival:

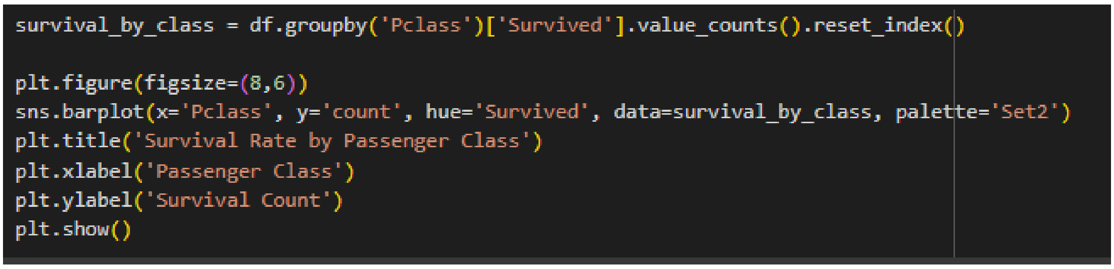


```
# passenger class by survival

def passenger_class(pclass):
  if pclass == 1:
    return "First Class"
  elif pclass == 2:
    return "Second Class"
  elif pclass == 3:
    return "Third Class"
  else:
    return "Unknown"

df['Pclass'] = df['Pclass'].apply(passenger_class)
df.head(5)
```

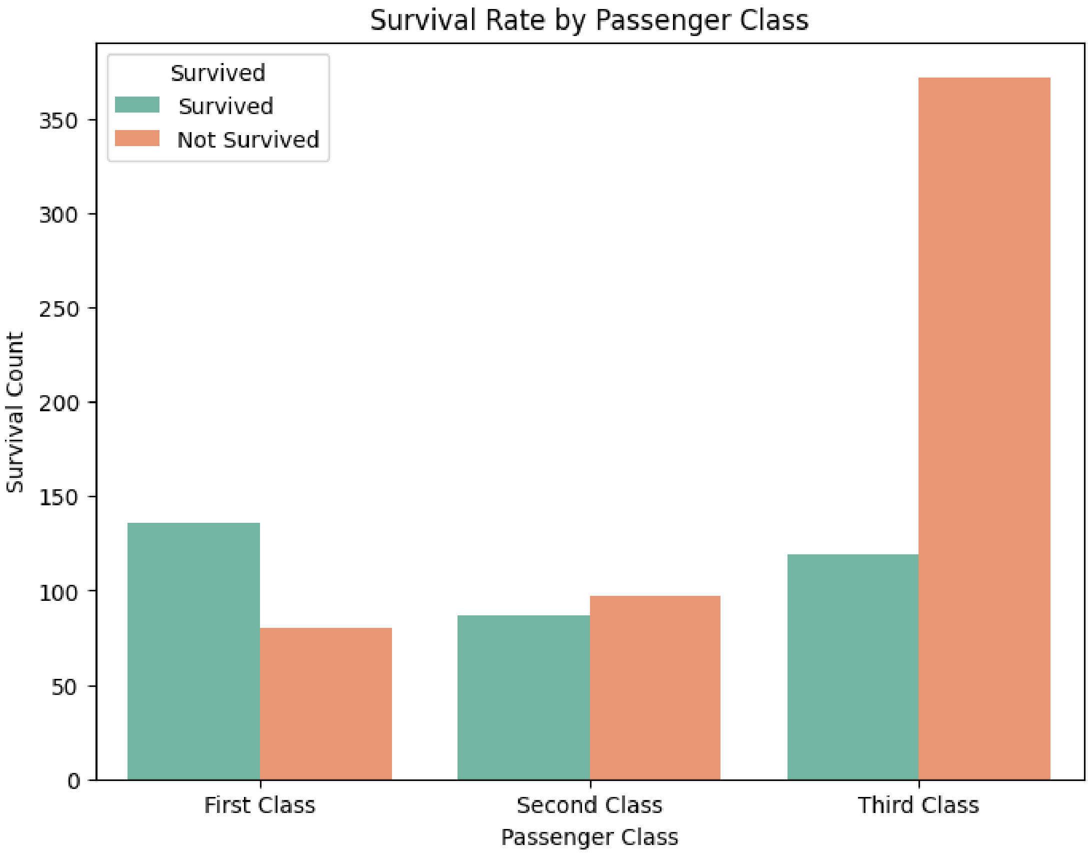


	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	Not Survived	Third Class	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	GEN	S
1	2	Survived	First Class	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	Survived	Third Class	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	GEN	S
3	4	Survived	First Class	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	Not Survived	Third Class	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	GEN	S

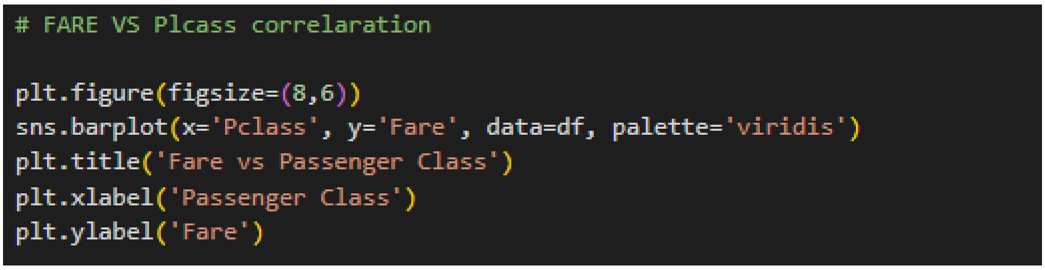
EDA:



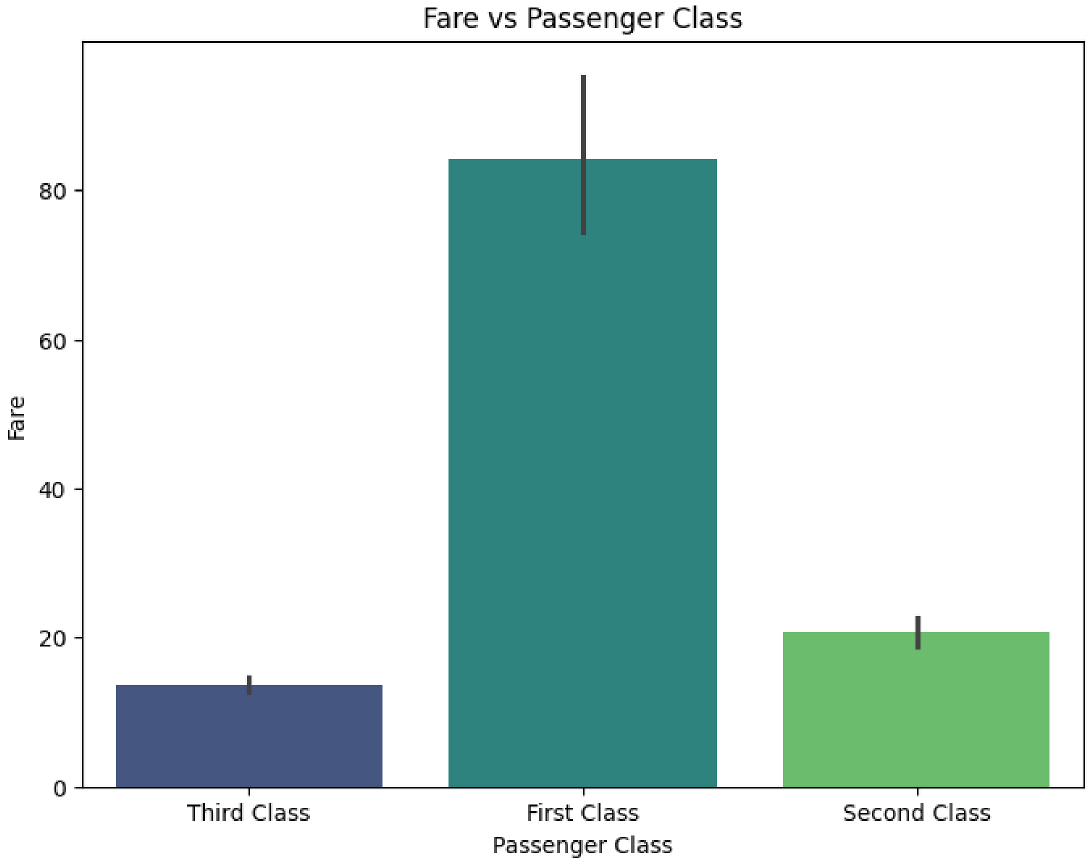
```
survival_by_class = df.groupby('Pclass')['Survived'].value_counts().reset_index()

plt.figure(figsize=(8,6))
sns.barplot(x='Pclass', y='count', hue='Survived', data=survival_by_class, palette='Set2')
plt.title('Survival Rate by Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Survival Count')
plt.show()
```

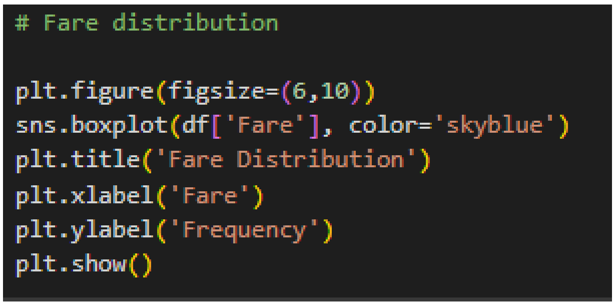


Fare vs Passenger Class Relation:

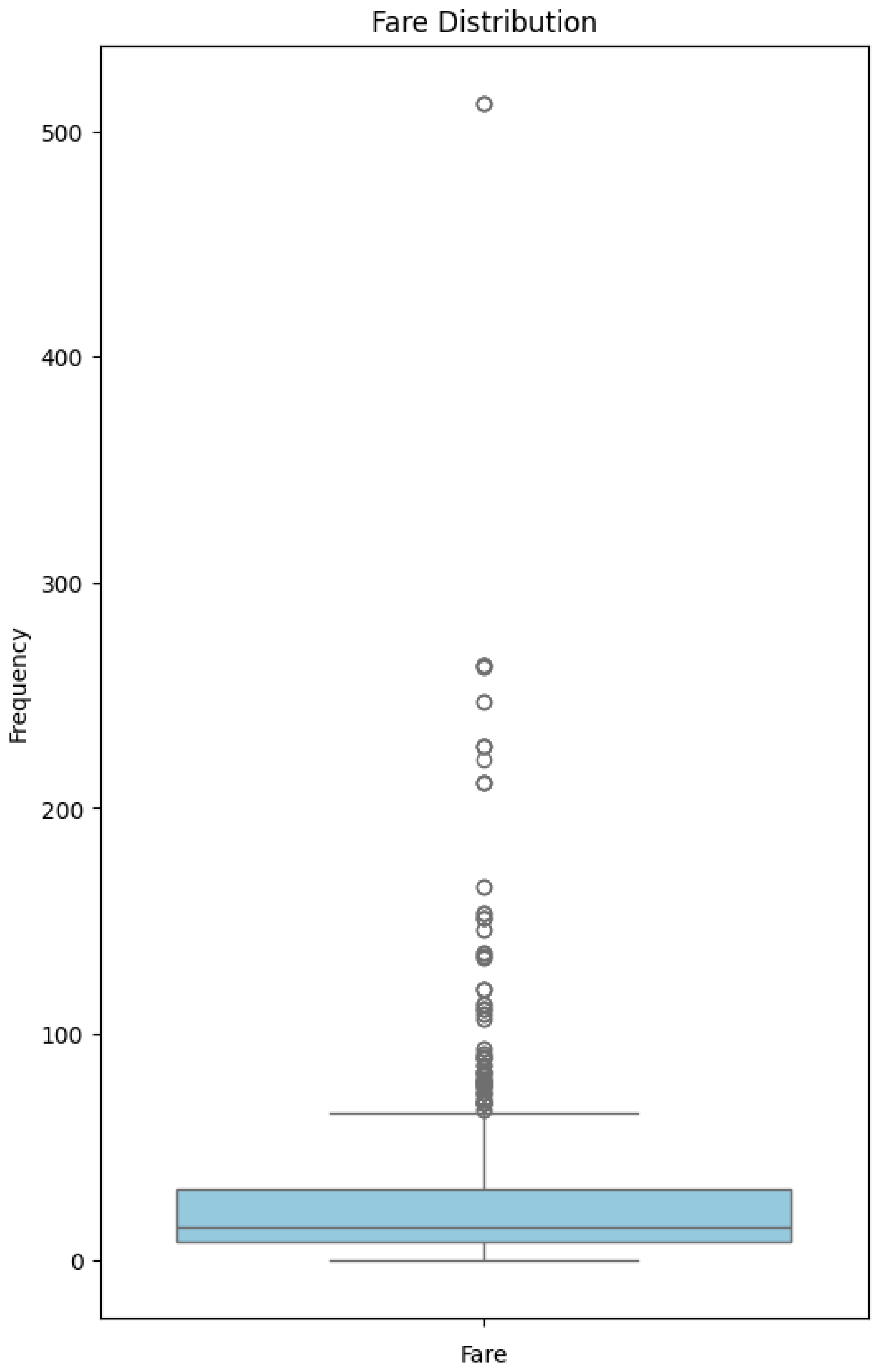


```
# FARE VS Plcass correlaration

plt.figure(figsize=(8,6))
sns.barplot(x='Pclass', y='Fare', data=df, palette='viridis')
plt.title('Fare vs Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Fare')
```





```
# Fare distribution

plt.figure(figsize=(6,10))
sns.boxplot(df['Fare'], color='skyblue')
plt.title('Fare Distribution')
plt.xlabel('Fare')
plt.ylabel('Frequency')
plt.show()
```



Insights:

Visual Insights

1. Survival Count
    - About 38% survived, 62% did not.
    - Dataset is slightly imbalanced.
2. Gender vs Survival
    - Females had a much higher survival rate than males.
    - Majority of passengers were male.
3. Passenger Class vs Survival
    - 1st class passengers survived the most.
    - Survival dropped significantly in 3rd class.
4. Fare Distribution by Class
    - 1st class had higher fares and more variation.
    - Some outliers in fare among all classes.

Thank You