PRODIGY INFOTECH - TASK-02

Perform data cleaning and exploratory data analysis (EDA) on a dataset and explore relations and identify patterns and trends in data

In [1]:

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

In [2]:

```
data= pd.read_csv(r'C:\Users\HP\Documents\Datasets\titanic_train.csv')
data
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	7
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	5
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	1
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	3
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	2
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	3
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	

891 rows × 12 columns

In [83]:

```
data.info()
```

```
<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
```

In [84]:

```
data.describe()
```

Out[84]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.00
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.20
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.69
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.00
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.91
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.45
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.00
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.32

In [86]:

```
data.isna().sum()
```

Out[86]:

```
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
```

In [87]:

```
data['Age'].mean()
```

Out[87]:

```
29.69911764705882
```

In [88]:
```
data['Age'].fillna(int(data['Age'].mean()), inplace= True)
```

In [90]:
```
data.drop(columns= ['Cabin'], axis=1, inplace= True)
```

In [91]:
```
data.dropna(inplace= True)
```

In [93]:
```
data.isna().sum()
```

Out[93]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

In [95]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 889 entries, 0 to 890
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  889 non-null    int64  
 1   Survived     889 non-null    int64  
 2   Pclass       889 non-null    int64  
 3   Name         889 non-null    object 
 4   Sex          889 non-null    object 
 5   Age          889 non-null    float64
 6   SibSp        889 non-null    int64  
 7   Parch        889 non-null    int64  
 8   Ticket       889 non-null    object 
 9   Fare         889 non-null    float64
 10  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(4)
memory usage: 83.3+ KB
```

In [96]:
```
data
```

Out[96]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	7
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	5
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	1
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	3
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.0	1	2	W./C. 6607	2
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	3
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	

889 rows × 11 columns

In [98]:

```
data['Sex']= data['Sex'].map({'male': 0, 'female': 1})
data['Embarked']= data['Embarked'].map({'S': 0, 'C': 1, 'Q': 2})
data
```

Out[98]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	0	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	1	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	0	35.0	0	0	373450	8.0
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	0	27.0	0	0	211536	13.0
887	888	1	1	Graham, Miss. Margaret Edith	1	19.0	0	0	112053	30.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	1	29.0	1	2	W./C. 6607	23.4
889	890	1	1	Behr, Mr. Karl Howell	0	26.0	0	0	111369	30.0
890	891	0	3	Dooley, Mr. Patrick	0	32.0	0	0	370376	7.7

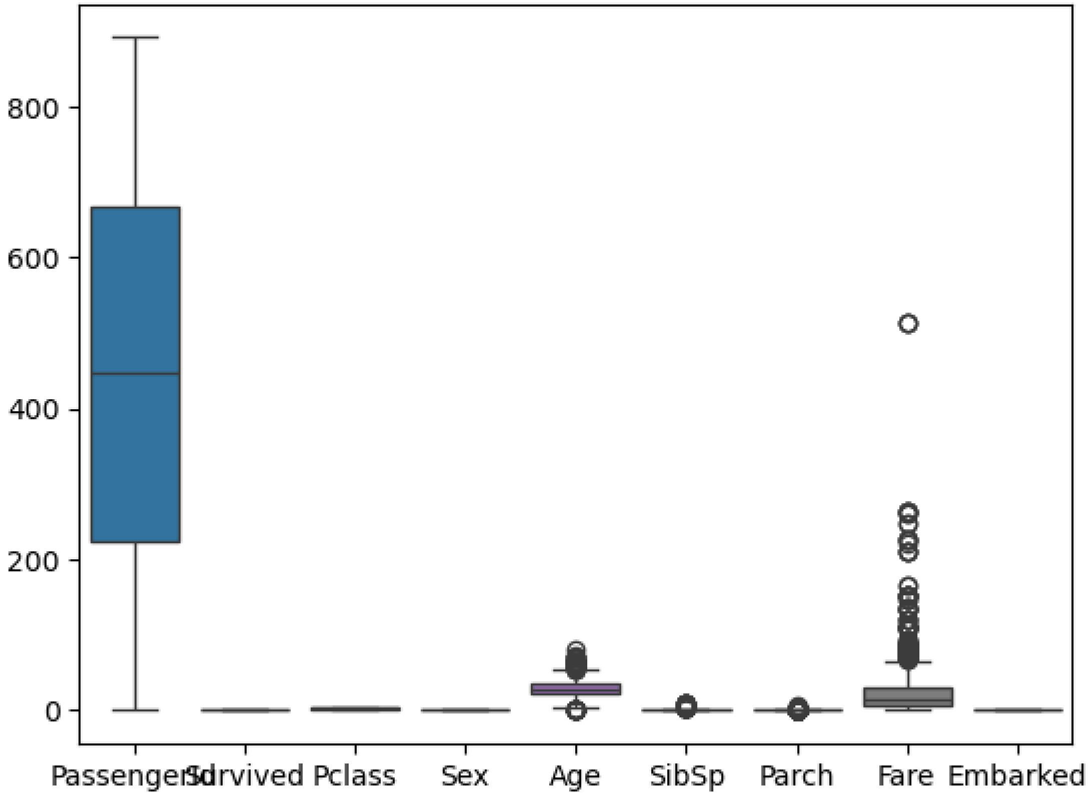
889 rows × 11 columns

In [100…

```
sns.boxplot(data)
```

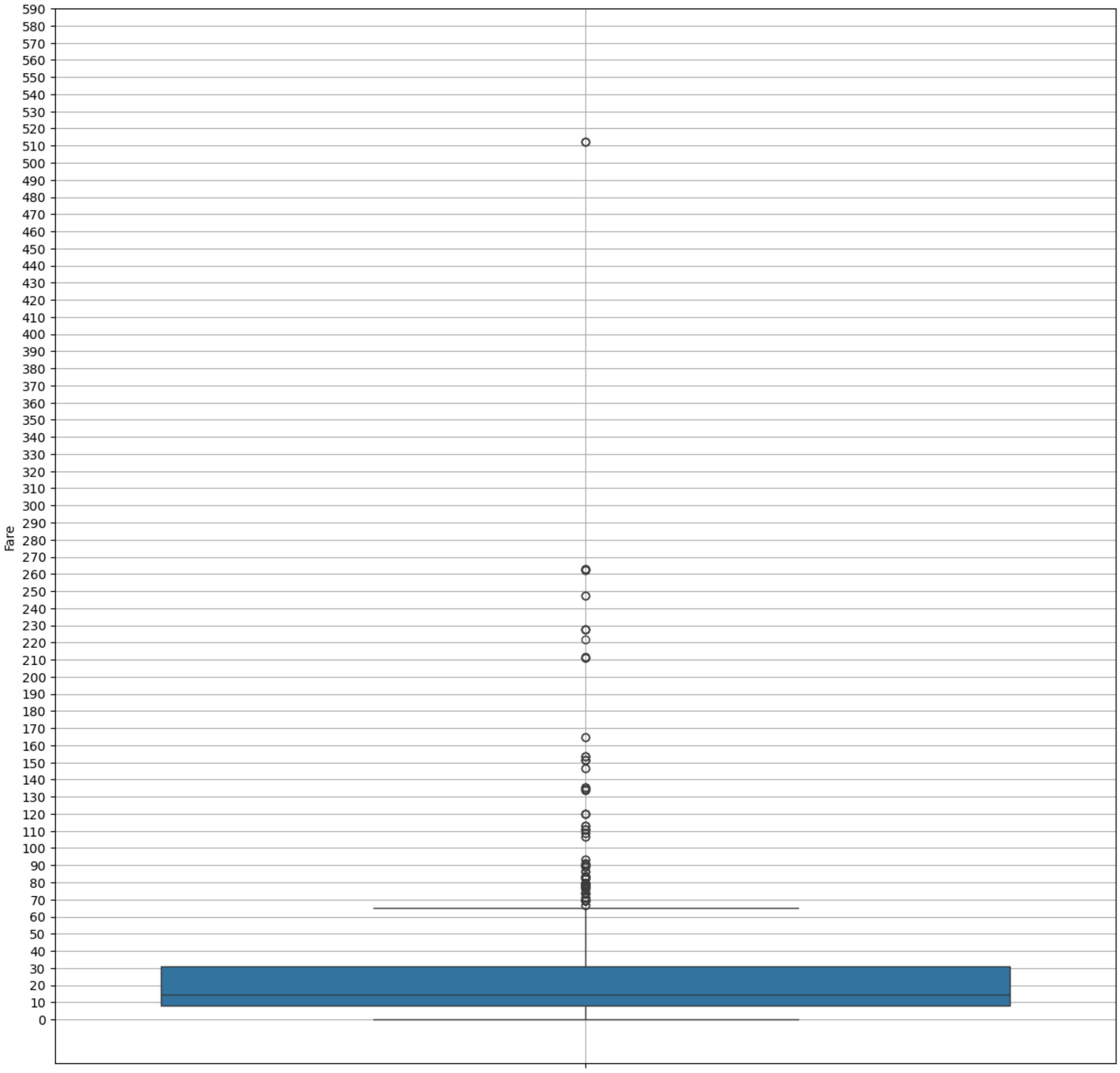
Out[100…

```
<Axes: >
```



In [102…

```
plt.figure(figsize= (15,15))
sns.boxplot(data['Fare'])
plt.yticks(np.arange(0,600,10))
plt.grid()
plt.show()
```



In [103…

```
data[data['Fare'] > 80].index
```

Out[103…

```
Index([ 27,  31,  34,  62,  88, 118, 195, 215, 224, 230, 245, 257, 258, 268,
       269, 291, 297, 299, 305, 306, 307, 310, 311, 318, 319, 325, 332, 334,
       337, 341, 373, 375, 377, 380, 390, 393, 412, 435, 438, 445, 453, 484,
       486, 498, 504, 505, 520, 527, 537, 544, 550, 557, 581, 609, 659, 660,
       679, 689, 698, 700, 708, 716, 730, 737, 742, 759, 763, 779, 802, 820,
       835, 849, 856, 879],
      dtype='int64')
```

In [106…

```
data.drop(index= [27,  31,  34,  62,  88, 118, 195, 215, 224, 230, 245, 257, 258
       269, 291, 297, 299, 305, 306, 307, 310, 311, 318, 319, 325, 332, 334,
       337, 341, 373, 375, 377, 380, 390, 393, 412, 435, 438, 445, 453, 484,
       486, 498, 504, 505, 520, 527, 537, 544, 550, 557, 581, 609, 659, 660,
       679, 689, 698, 700, 708, 716, 730, 737, 742, 759, 763, 779, 802, 820,
       835, 849, 856, 879], axis=0, inplace= True)
```

In [109…

```
data[data['Fare'] > 60].index
```

Out[109…

```
Index([  1,  52,  54,  72,  92,  97, 102, 120, 124, 139, 151, 155, 159, 180,
       201, 218, 256, 262, 275, 290, 324, 336, 366, 369, 385, 496, 540, 558,
       585, 587, 591, 615, 627, 641, 645, 655, 665, 681, 741, 745, 754, 765,
       789, 792, 846, 863],
      dtype='int64')
```

In [112…

```
data.drop(index= [1,  52,  54,  72,  92,  97, 102, 120, 124, 139, 151, 155, 159,
       201, 218, 256, 262, 275, 290, 324, 336, 366, 369, 385, 496, 540, 558,
       585, 587, 591, 615, 627, 641, 645, 655, 665, 681, 741, 745, 754, 765,
       789, 792, 846, 863], axis= 0, inplace= True)
```

In [113…

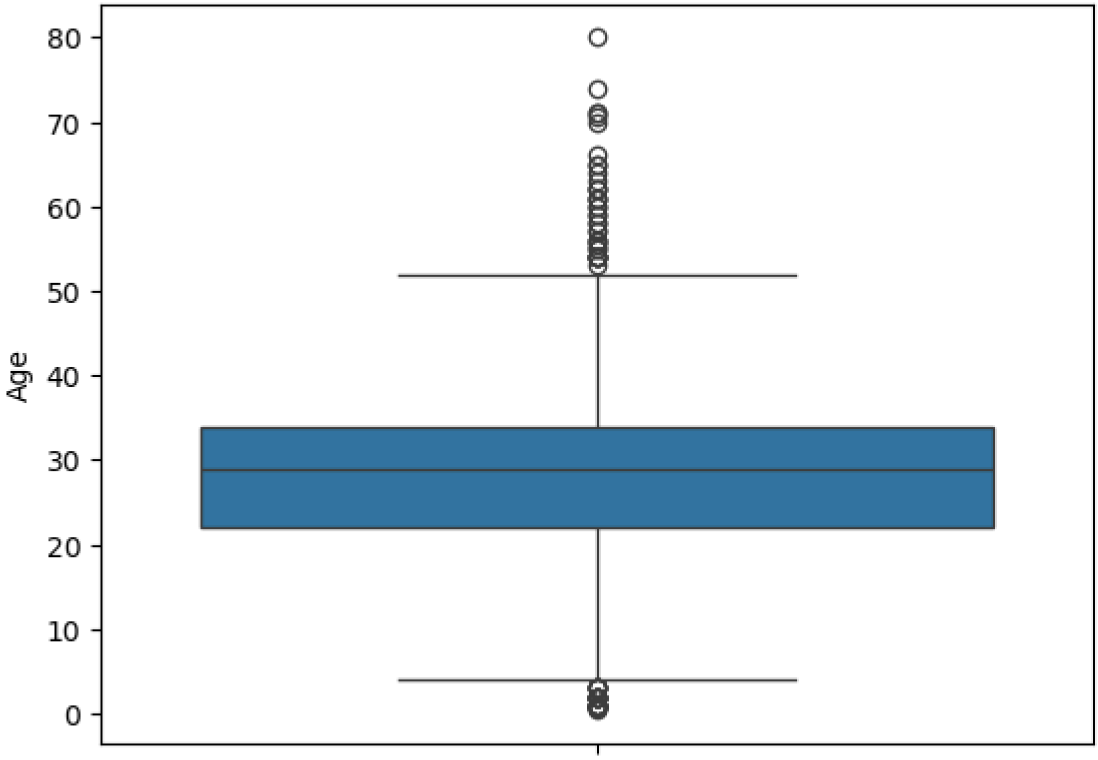
```
data.drop_duplicates(keep= 'first', inplace= True)
```

In [114…

```
sns.boxplot(data['Age'])
```

Out[114…

```
<Axes: ylabel='Age'>
```



In [116…

```
data[data['Age'] > 45].index
```

Out[116…

```
Index([  6,  11,  15,  33,  94,  96, 110, 116, 132, 150, 152, 170, 174, 177,
       203, 222, 232, 249, 252, 259, 280, 317, 326, 331, 397, 406, 434, 449,
       456, 458, 460, 462, 463, 467, 482, 483, 487, 492, 493, 513, 515, 526,
       545, 555, 556, 570, 571, 582, 586, 592, 597, 599, 625, 626, 630, 631,
       647, 662, 672, 684, 694, 695, 712, 714, 723, 736, 771, 772, 774, 796,
       851, 857, 862, 871, 873],
      dtype='int64')
```

In [117…

```
data.drop(index= [6,  11,  15,  33,  94,  96, 110, 116, 132, 150, 152, 170, 174,
       203, 222, 232, 249, 252, 259, 280, 317, 326, 331, 397, 406, 434, 449,
       456, 458, 460, 462, 463, 467, 482, 483, 487, 492, 493, 513, 515, 526,
       545, 555, 556, 570, 571, 582, 586, 592, 597, 599, 625, 626, 630, 631,
       647, 662, 672, 684, 694, 695, 712, 714, 723, 736, 771, 772, 774, 796,
       851, 857, 862, 871, 873], axis=0, inplace= True)
```

In [118…

```
data[data['Age'] < 10].index
```

Out[118…

```
Index([  7,  10,  16,  24,  43,  50,  58,  63,  78, 119, 147, 164, 165, 171,
       172, 182, 183, 184, 193, 205, 233, 237, 261, 278, 340, 348, 374, 381,
       386, 407, 448, 469, 479, 480, 489, 530, 535, 541, 549, 618, 634, 642,
       644, 691, 720, 750, 751, 755, 777, 787, 788, 803, 813, 824, 827, 831,
       850, 852, 869],
      dtype='int64')
```

In [122…

```
data.drop(index= [7,  10,  16,  24,  50,  58,  63,  78, 119, 147, 164, 165, 171,
       182, 183, 184, 193, 205, 233, 237, 261, 278, 340, 348, 374, 381, 407,
       448, 469, 479, 489, 530, 535, 541, 549, 618, 634, 642, 644, 691, 720,
       750, 751, 755, 777, 787, 788, 803, 813, 824, 827, 831, 850, 852, 869], ax
```

In [126…

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 638 entries, 0 to 890
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  638 non-null    int64  
 1   Survived     638 non-null    int64  
 2   Pclass       638 non-null    int64  
 3   Name         638 non-null    object 
 4   Sex          638 non-null    int64  
 5   Age          638 non-null    float64
 6   SibSp        638 non-null    int64  
 7   Parch        638 non-null    int64  
 8   Ticket       638 non-null    object 
 9   Fare         638 non-null    float64
 10  Embarked     638 non-null    int64  
dtypes: float64(2), int64(7), object(2)
memory usage: 59.8+ KB
```

In [128…

```
data
```

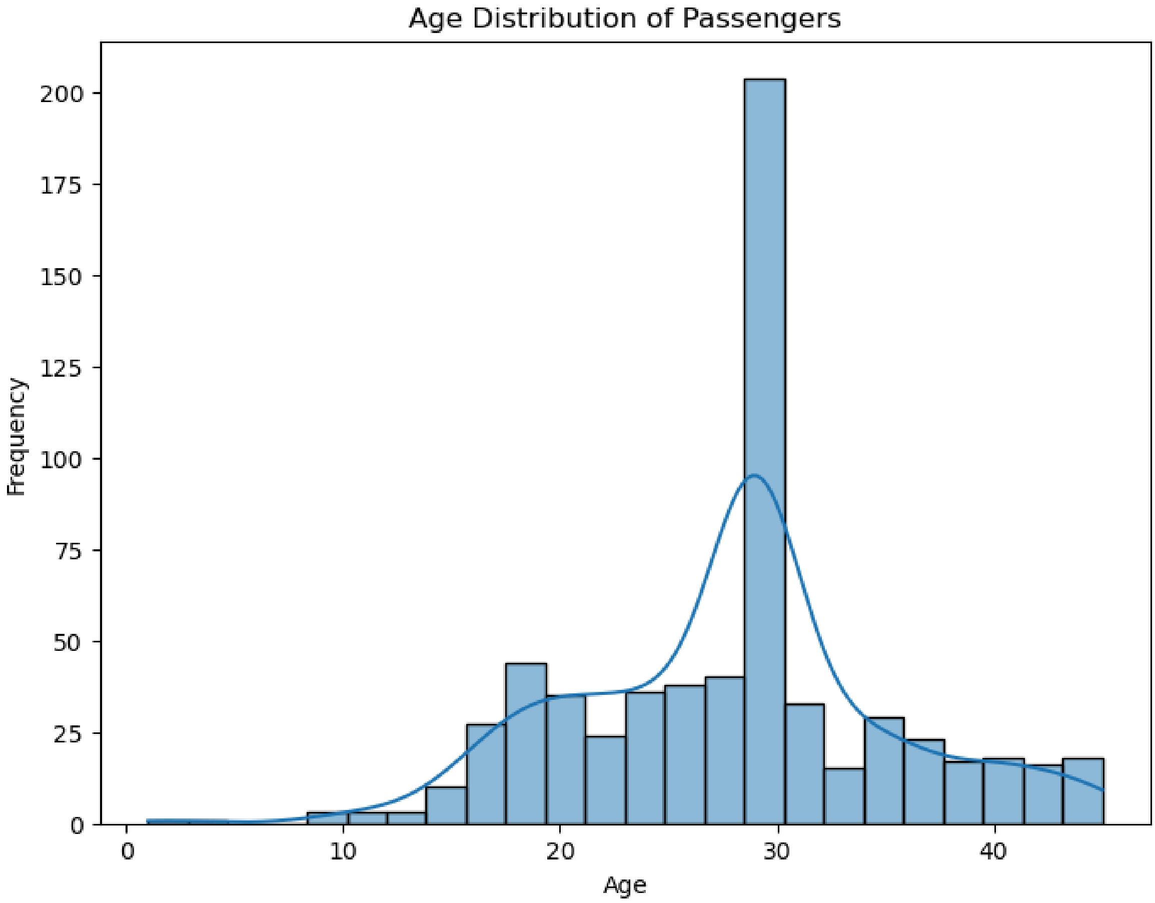
Out[128…

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	0	22.0	1	0	A/5 21171	7.2
2	3	1	3	Heikkinen, Miss. Laina	1	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	0	35.0	0	0	373450	8.0
5	6	0	3	Moran, Mr. James	0	29.0	0	0	330877	8.4
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	0	27.0	0	0	211536	13.0
887	888	1	1	Graham, Miss. Margaret Edith	1	19.0	0	0	112053	30.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	1	29.0	1	2	W./C. 6607	23.4
889	890	1	1	Behr, Mr. Karl Howell	0	26.0	0	0	111369	30.0
890	891	0	3	Dooley, Mr. Patrick	0	32.0	0	0	370376	7.7

638 rows × 11 columns

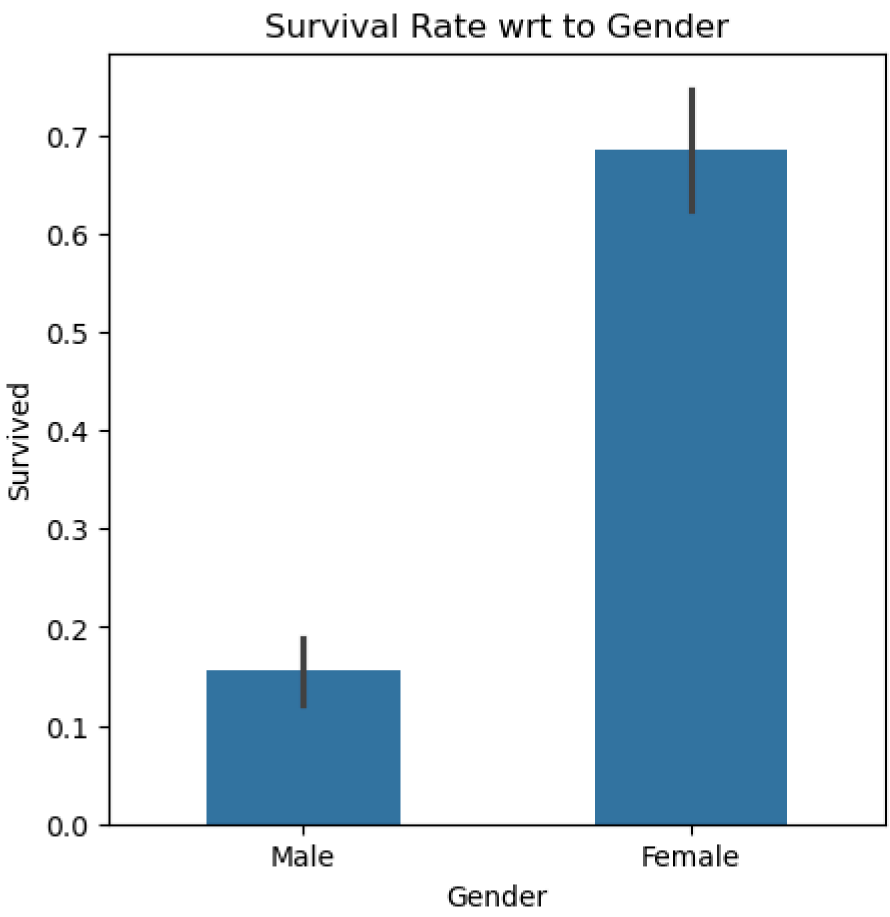
In [135…

```
plt.figure(figsize= (8,6))
sns.histplot(data['Age'], kde= True)
plt.title('Age Distribution of Passengers')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



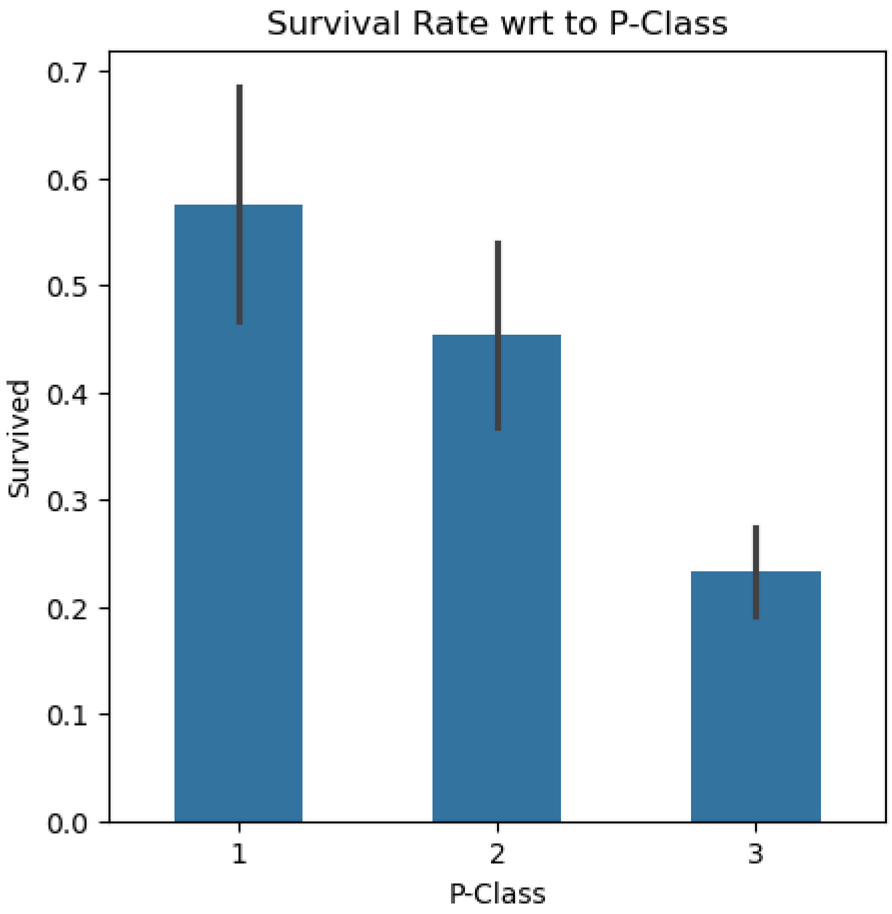
In [136…

```
plt.figure(figsize= (5,5))
sns.barplot(x='Sex', y='Survived', data= data, width= 0.5)
plt.title('Survival Rate wrt to Gender')
plt.xlabel('Gender')
plt.xticks(ticks= [0,1], labels= ['Male','Female'])
plt.show()
```



In [139…

```
plt.figure(figsize= (5,5))
sns.barplot(x= 'Pclass', y= 'Survived', data= data, width= 0.5)
plt.title('Survival Rate wrt to P-Class')
plt.xlabel('P-Class')
plt.show()
```

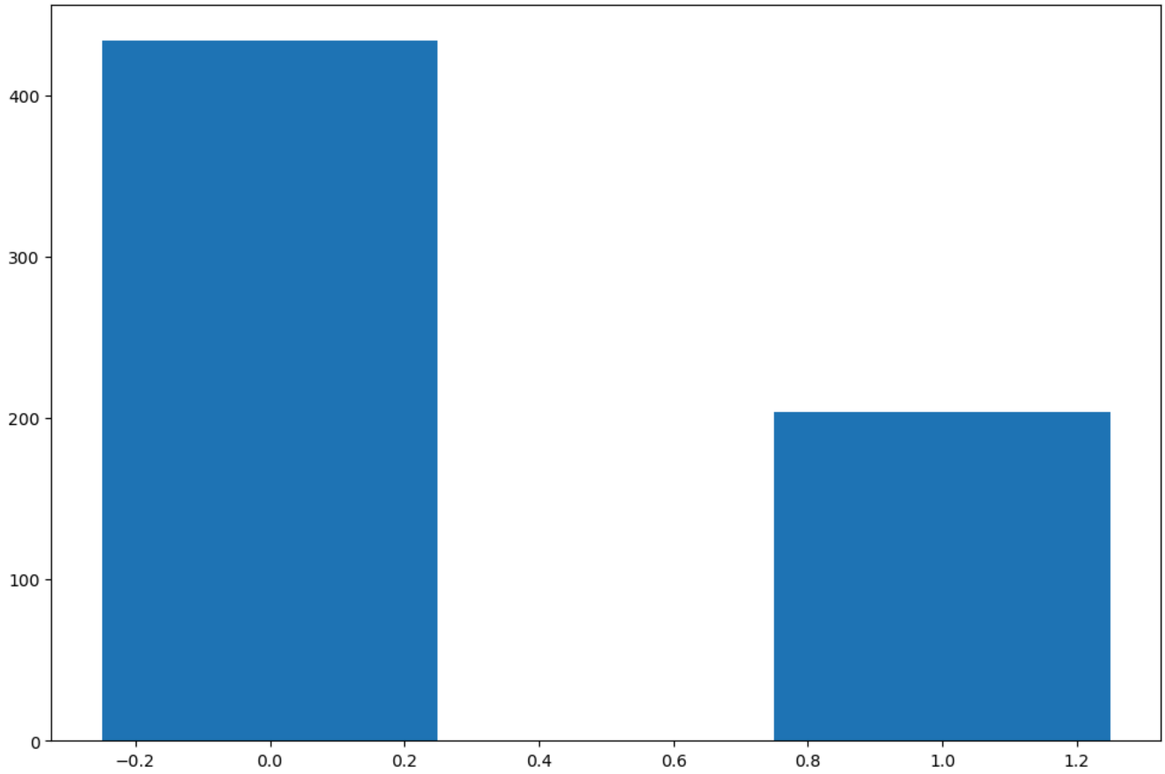


In [98]:

```
plt.figure(figsize= (12,8))
width= 0.5
x= data['Sex'].unique()
y= list(data['Survived'].value_counts())
survived= [len(data)-y[0], y[1]]
nonsurvived= [y[0], len(data)-y[1]]
values= np.arange(len(x))
plt.bar(values, y, width, label='Survived')
```

Out[98]:

```
<BarContainer object of 2 artists>
```



In [145…

```
data.groupby('Sex')['Survived'].value_counts()
```

Out[145…

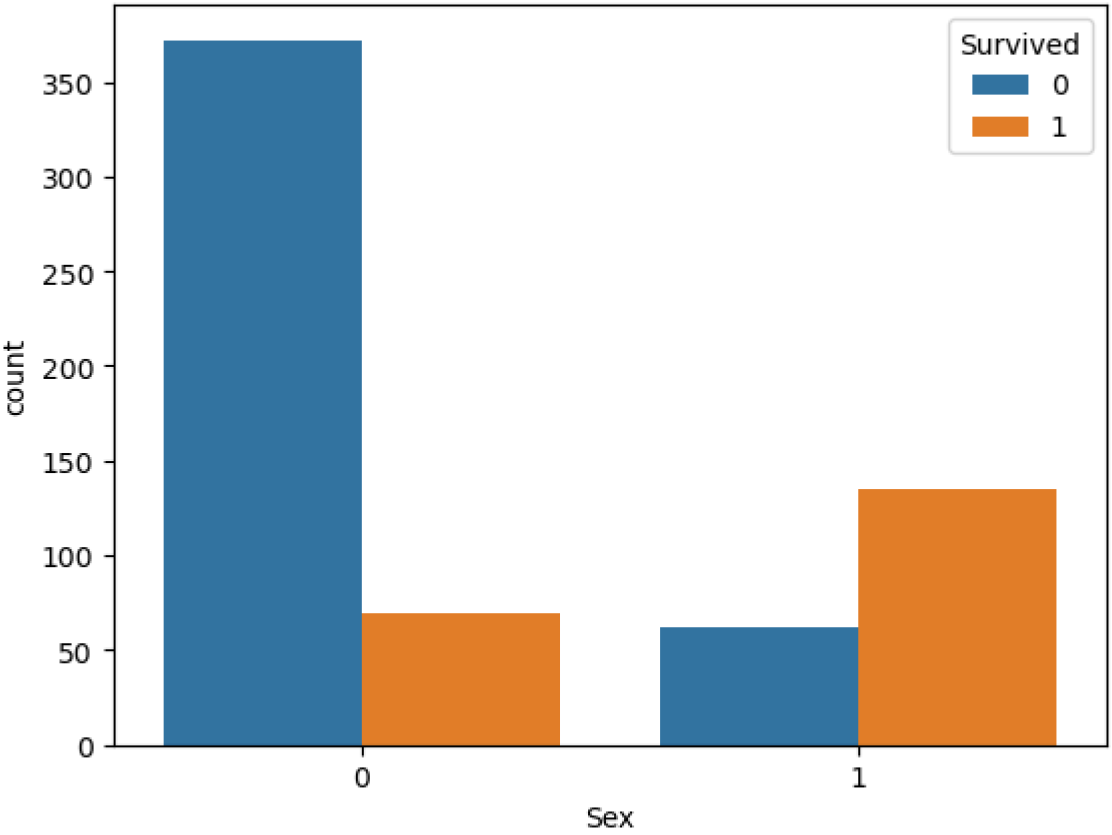
```
Sex  Survived
0    0           372
     1            69
1    1           135
     0            62
Name: count, dtype: int64
```

In [141…

```
sns.countplot(x= "Sex", data= data, hue= 'Survived')
```

Out[141…

```
<Axes: xlabel='Sex', ylabel='count'>
```



In [96]:
```
data.drop(columns= ['Name','Ticket'], axis= 1, inplace= True)
```

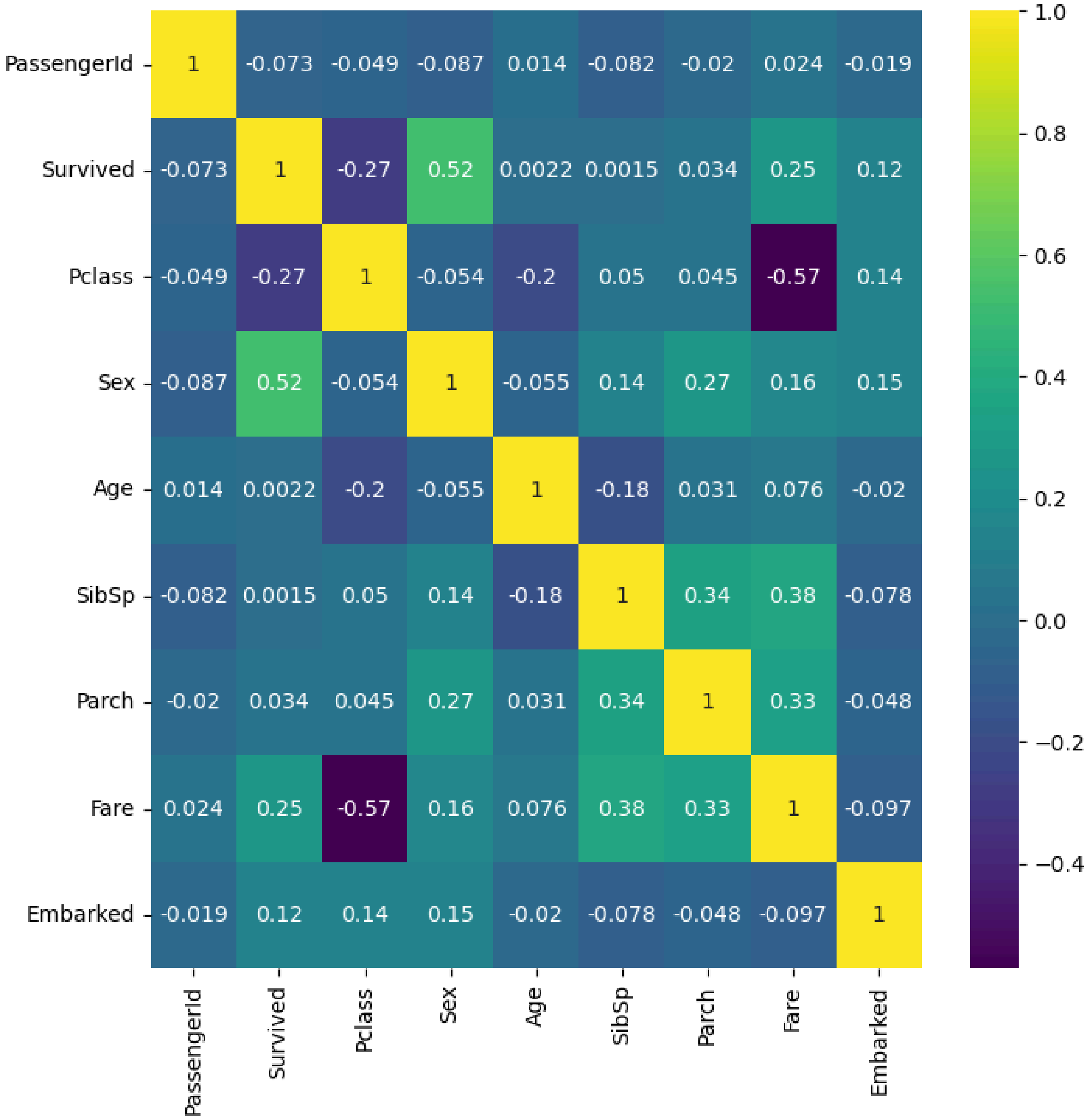
In [98]:
```
data.corr()
```

Out[98]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parc
PassengerId	1.000000	-0.073069	-0.048729	-0.086720	0.014349	-0.081591	-0.01965
Survived	-0.073069	1.000000	-0.272161	0.523839	0.002240	0.001496	0.03372
Pclass	-0.048729	-0.272161	1.000000	-0.054085	-0.203933	0.050404	0.04533
Sex	-0.086720	0.523839	-0.054085	1.000000	-0.054688	0.142961	0.26841
Age	0.014349	0.002240	-0.203933	-0.054688	1.000000	-0.177038	0.03070
SibSp	-0.081591	0.001496	0.050404	0.142961	-0.177038	1.000000	0.33966
Parch	-0.019653	0.033722	0.045335	0.268417	0.030701	0.339669	1.00000
Fare	0.024096	0.252518	-0.572309	0.157645	0.075522	0.378919	0.32878
Embarked	-0.019002	0.121725	0.140412	0.151057	-0.020077	-0.078063	-0.04803

In [132…
```
plt.figure(figsize= (8,8))
sns.heatmap(data.corr(), annot= True, cmap= 'viridis')
```

Out[132…
```
<Axes: >
```



End Of Project

In [ ]: