In [2]:
```
import pandas as pd
import numpy as np
```

In [5]:
```
df=pd.read_csv("train.csv")
```

In [7]:
```
print(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
```

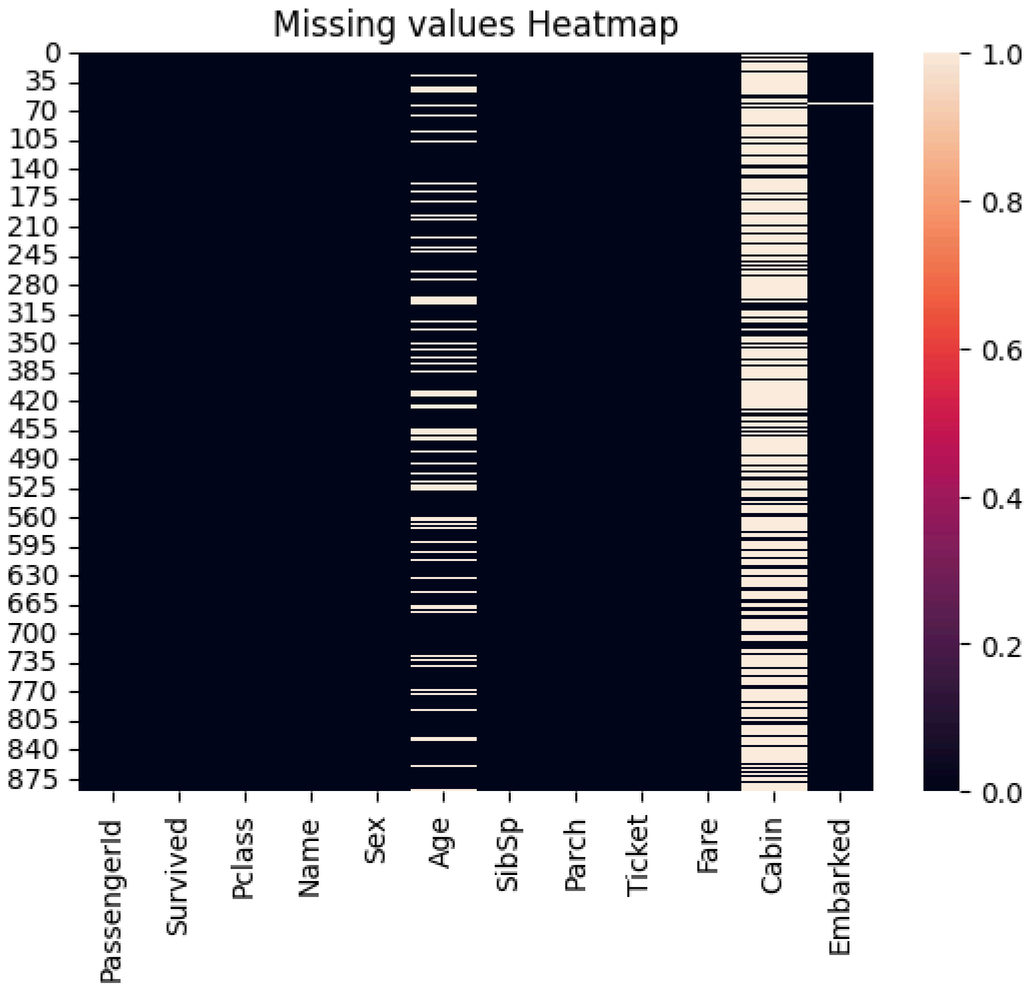
In [8]:
```
print(df.describe(include='all'))
```

```
        PassengerId    Survived      Pclass                 Name   Sex  \
count    891.000000  891.000000  891.000000                  891   891   
unique          NaN         NaN         NaN                  891     2   
top             NaN         NaN         NaN  Dooley, Mr. Patrick  male   
freq            NaN         NaN         NaN                    1   577   
mean     446.000000    0.383838    2.308642                  NaN   NaN   
std      257.353842    0.486592    0.836071                  NaN   NaN   
min        1.000000    0.000000    1.000000                  NaN   NaN   
25%      223.500000    0.000000    2.000000                  NaN   NaN   
50%      446.000000    0.000000    3.000000                  NaN   NaN   
75%      668.500000    1.000000    3.000000                  NaN   NaN   
max      891.000000    1.000000    3.000000                  NaN   NaN   

               Age       SibSp       Parch  Ticket        Fare Cabin Embarked  
count   714.000000  891.000000  891.000000     891  891.000000   204      889  
unique         NaN         NaN         NaN     681         NaN   147        3  
top            NaN         NaN         NaN  347082         NaN    G6        S  
freq           NaN         NaN         NaN       7         NaN     4      644  
mean     29.699118    0.523008    0.381594     NaN   32.204208   NaN      NaN  
std      14.526497    1.102743    0.806057     NaN   49.693429   NaN      NaN  
min       0.420000    0.000000    0.000000     NaN    0.000000   NaN      NaN  
25%      20.125000    0.000000    0.000000     NaN    7.910400   NaN      NaN  
50%      28.000000    0.000000    0.000000     NaN   14.454200   NaN      NaN  
75%      38.000000    1.000000    0.000000     NaN   31.000000   NaN      NaN  
max      80.000000    8.000000    6.000000     NaN  512.329200   NaN      NaN  
```

In [26]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [27]:
```
sns.heatmap(df.isnull())
plt.title("Missing values Heatmap")
plt.show()
```



In [38]:
```
df.fillna({'Age':df['Age'].mean()},inplace=True)
```

In [40]:
```
df.isnull().sum()
```

Out[40]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [46]:
```
df['Pclass'] = df['Pclass'].astype('category')
```

In [48]:
```
df.dtypes
```

Out[48]:
```
PassengerId       int64
Survived          int64
Pclass         category
Name             object
Sex              object
Age             float64
SibSp             int64
Parch             int64
Ticket           object
Fare            float64
Cabin            object
Embarked         object
dtype: object
```

In [52]:
```
df = pd.get_dummies(df, columns=['Sex', 'Embarked'], drop_first=True)
```

In [54]:
```
df.describe(include='all')
```

Out[54]:

	PassengerId	Survived	Pclass	Name	Age	SibSp	Parch	'
count	891.000000	891.000000	891.0	891	891.000000	891.000000	891.000000	
unique	NaN	NaN	3.0	891	NaN	NaN	NaN	
top	NaN	NaN	3.0	Dooley, Mr. Patrick	NaN	NaN	NaN	3
freq	NaN	NaN	491.0	1	NaN	NaN	NaN	
mean	446.000000	0.383838	NaN	NaN	29.699118	0.523008	0.381594	
std	257.353842	0.486592	NaN	NaN	13.002015	1.102743	0.806057	
min	1.000000	0.000000	NaN	NaN	0.420000	0.000000	0.000000	
25%	223.500000	0.000000	NaN	NaN	22.000000	0.000000	0.000000	
50%	446.000000	0.000000	NaN	NaN	29.699118	0.000000	0.000000	
75%	668.500000	1.000000	NaN	NaN	35.000000	1.000000	0.000000	
max	891.000000	1.000000	NaN	NaN	80.000000	8.000000	6.000000	

In [56]:
```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
df[['Age', 'Fare']] = scaler.fit_transform(df[['Age', 'Fare']])
```

In [58]:
```
df.describe(include='all')
```

Out[58]:

	PassengerId	Survived	Pclass	Name	Age	SibSp	Parch	
count	891.000000	891.000000	891.0	891	891.000000	891.000000	891.000000	
unique	NaN	NaN	3.0	891	NaN	NaN	NaN	
top	NaN	NaN	3.0	Dooley, Mr. Patrick	NaN	NaN	NaN	3
freq	NaN	NaN	491.0	1	NaN	NaN	NaN	
mean	446.000000	0.383838	NaN	NaN	0.367921	0.523008	0.381594	
std	257.353842	0.486592	NaN	NaN	0.163383	1.102743	0.806057	
min	1.000000	0.000000	NaN	NaN	0.000000	0.000000	0.000000	
25%	223.500000	0.000000	NaN	NaN	0.271174	0.000000	0.000000	
50%	446.000000	0.000000	NaN	NaN	0.367921	0.000000	0.000000	
75%	668.500000	1.000000	NaN	NaN	0.434531	1.000000	0.000000	
max	891.000000	1.000000	NaN	NaN	1.000000	8.000000	6.000000	

In [60]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    category
 3   Name         891 non-null    object
 4   Age          891 non-null    float64
 5   SibSp        891 non-null    int64
 6   Parch        891 non-null    int64
 7   Ticket       891 non-null    object
 8   Fare         891 non-null    float64
 9   Cabin        204 non-null    object
 10  Sex_male     891 non-null    bool
 11  Embarked_Q   891 non-null    bool
 12  Embarked_S   891 non-null    bool
dtypes: bool(3), category(1), float64(2), int64(4), object(3)
memory usage: 66.4+ KB
```

In [62]:

```
df.head()
```

Out[62]:

	PassengerId	Survived	Pclass	Name	Age	SibSp	Parch	Ticket	Far
0	1	0	3	Braund, Mr. Owen Harris	0.271174	1	0	A/5 21171	0.01415
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0.472229	1	0	PC 17599	0.13913
2	3	1	3	Heikkinen, Miss. Laina	0.321438	0	0	STON/O2. 3101282	0.01546
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0.434531	1	0	113803	0.10364
4	5	0	3	Allen, Mr. William Henry	0.434531	0	0	373450	0.01571

In [114…:

```
df.to_csv('TitanicCleaned.csv',index=False)
```

In [ ]: