In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
train =pd.read_csv("titanic_train.csv")
```

In [3]:
```
train.head()
```

Out[3]:

	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

In [4]:
```
print(train.shape)
```

```
(891, 12)
```

In [5]:
```
Target = train['Survived']
train.drop('Survived', axis=1, inplace=True)
```

In [6]:
```
train.drop('PassengerId', axis=1, inplace=True)
train.head()
```

Out[6]:

	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [7]:
```
na_object_col = []
for col in train.columns:
    if train[col].dtype == 'object' and train[col].isnull().any() == True:
        na_object_col.append(col)

for col in na_object_col:
    print(col,':',train[col].isnull().sum())
```

```
Cabin : 687
Embarked : 2
```

In [8]:
```
na_numeric_col = []
for col in train.columns:
    if train[col].dtype != 'object' and train[col].isnull().any() == True:
        na_numeric_col.append(col)

for col in na_numeric_col:
    print(col ,':', train[col].isnull().sum())
```

```
Age : 177
```

In [9]:
```
train.isnull().sum()
```

Out[9]:
```
Pclass        0
Name          0
Sex           0
Age         177
SibSp         0
Parch         0
Ticket        0
Fare          0
Cabin       687
Embarked      2
dtype: int64
```

In [10]:
```
train['Cabin'].fillna('Not Available', inplace=True)

# Embarked only has 2 missing values, so replacing the NaN with mode
train['Embarked'].fillna(train['Embarked'].mode()[0], inplace = True)

print('Cabin:',train['Cabin'].isnull().sum())
print('Embarked:',train['Embarked'].isnull().sum())
```

```
Cabin: 0
Embarked: 0
```

In [11]:
```
train['Pclass'] = train['Pclass'].astype('object')
```

In [12]:
```
object_col = []
for col in train.columns:
    if train[col].dtype == 'object':
        object_col.append(col)
object_col
```

Out[12]:
```
['Pclass', 'Name', 'Sex', 'Ticket', 'Cabin', 'Embarked']
```

In [13]:
```
Sex_dummies = pd.get_dummies(train['Sex'], drop_first = True, prefix = 'Sex')
Embarked_dummies = pd.get_dummies(train['Embarked'], drop_first = True, prefix = 'Embarked')
Pclass_dummies = pd.get_dummies(train['Pclass'], drop_first = True, prefix = 'Pclass')
```

In [14]:
```
df2 = pd.concat([train, Sex_dummies, Embarked_dummies, Pclass_dummies], axis=1)
```

In [15]:
```
df2.drop(['Sex','Embarked','Pclass'], axis=1, inplace=True)
df2
```

Out[15]:

	Name	Age	SibSp	Parch	Ticket	Fare	Cabin	Sex_male	Embarked_Q	Embarked_S	Pclass_2	Pclass_3
0	Braund, Mr. Owen Harris	22.0	1	0	A/5 21171	7.2500	Not Available	1	0	1	0	1
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38.0	1	0	PC 17599	71.2833	C85	0	0	0	0	0
2	Heikkinen, Miss. Laina	26.0	0	0	STON/O2. 3101282	7.9250	Not Available	0	0	1	0	1
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35.0	1	0	113803	53.1000	C123	0	0	1	0	0
4	Allen, Mr. William Henry	35.0	0	0	373450	8.0500	Not Available	1	0	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
886	Montvila, Rev. Juozas	27.0	0	0	211536	13.0000	Not Available	1	0	1	1	0
887	Graham, Miss. Margaret Edith	19.0	0	0	112053	30.0000	B42	0	0	1	0	0
888	Johnston, Miss. Catherine Helen "Carrie"	NaN	1	2	W./C. 6607	23.4500	Not Available	0	0	1	0	1
889	Behr, Mr. Karl Howell	26.0	0	0	111369	30.0000	C148	1	0	0	0	0
890	Dooley, Mr. Patrick	32.0	0	0	370376	7.7500	Not Available	1	1	0	0	1

891 rows × 12 columns

In [16]:
```
df3 = df2.copy()
```

In [17]:
```
df3.drop(['Name','Ticket','Cabin'], axis=1, inplace = True)
df3.head()
```

Out[17]:

	Age	SibSp	Parch	Fare	Sex_male	Embarked_Q	Embarked_S	Pclass_2	Pclass_3
0	22.0	1	0	7.2500	1	0	1	0	1
1	38.0	1	0	71.2833	0	0	0	0	0
2	26.0	0	0	7.9250	0	0	1	0	1
3	35.0	1	0	53.1000	0	0	1	0	0
4	35.0	0	0	8.0500	1	0	1	0	1

In [18]:
```
#Imputing Numerical Columns

from sklearn.impute import KNNImputer
imputer = KNNImputer()
df3 = pd.DataFrame(imputer.fit_transform(df3), columns = df3.columns)
df3.isnull().sum()
```

Out[18]:
```
Age           0
SibSp         0
Parch         0
Fare          0
Sex_male      0
Embarked_Q    0
Embarked_S    0
Pclass_2      0
Pclass_3      0
dtype: int64
```

In [19]:
```
df3.describe()
```

Out[19]:

	Age	SibSp	Parch	Fare	Sex_male	Embarked_Q	Embarked_S	Pclass_2	Pclass_3
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	29.949201	0.523008	0.381594	32.204208	0.647587	0.086420	0.725028	0.206510	0.551066
std	13.501483	1.102743	0.806057	49.693429	0.477990	0.281141	0.446751	0.405028	0.497665
min	0.420000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	21.100000	0.000000	0.000000	7.910400	0.000000	0.000000	0.000000	0.000000	0.000000
50%	29.000000	0.000000	0.000000	14.454200	1.000000	0.000000	1.000000	0.000000	1.000000
75%	38.000000	1.000000	0.000000	31.000000	1.000000	0.000000	1.000000	0.000000	1.000000
max	80.000000	8.000000	6.000000	512.329200	1.000000	1.000000	1.000000	1.000000	1.000000

In [20]:
```
from sklearn.preprocessing import MinMaxScaler
```

In [21]:
```
scaler = MinMaxScaler()
df4 = pd.DataFrame(scaler.fit_transform(df3.iloc[:,:4]), columns= df3.iloc[:,:4].columns )
df4 = pd.concat([df4,df3.iloc[:,4:]], axis=1)
df4
```

Out[21]:

	Age	SibSp	Parch	Fare	Sex_male	Embarked_Q	Embarked_S	Pclass_2	Pclass_3
0	0.271174	0.125	0.000000	0.014151	1.0	0.0	1.0	0.0	1.0
1	0.472229	0.125	0.000000	0.139136	0.0	0.0	0.0	0.0	0.0
2	0.321438	0.000	0.000000	0.015469	0.0	0.0	1.0	0.0	1.0
3	0.434531	0.125	0.000000	0.103644	0.0	0.0	1.0	0.0	0.0
4	0.434531	0.000	0.000000	0.015713	1.0	0.0	1.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...
886	0.334004	0.000	0.000000	0.025374	1.0	0.0	1.0	1.0	0.0
887	0.233476	0.000	0.000000	0.058556	0.0	0.0	1.0	0.0	0.0
888	0.331490	0.125	0.333333	0.045771	0.0	0.0	1.0	0.0	1.0
889	0.321438	0.000	0.000000	0.058556	1.0	0.0	0.0	0.0	0.0
890	0.396833	0.000	0.000000	0.015127	1.0	1.0	0.0	0.0	1.0

891 rows × 9 columns

In [22]:
```
train_final = df4.loc[:train.index.max(), :].copy()
```

In [23]:
```
train_final.head()
```

Out[23]:

	Age	SibSp	Parch	Fare	Sex_male	Embarked_Q	Embarked_S	Pclass_2	Pclass_3
0	0.271174	0.125	0.0	0.014151	1.0	0.0	1.0	0.0	1.0
1	0.472229	0.125	0.0	0.139136	0.0	0.0	0.0	0.0	0.0
2	0.321438	0.000	0.0	0.015469	0.0	0.0	1.0	0.0	1.0
3	0.434531	0.125	0.0	0.103644	0.0	0.0	1.0	0.0	0.0
4	0.434531	0.000	0.0	0.015713	1.0	0.0	1.0	0.0	1.0

In [ ]:

In [ ]: