In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
import warnings
warnings.simplefilter("ignore")
```

In [2]:
```
dataset=pd.read_csv("C:\\Users\\SRI KAAVYA\\OneDrive\\Desktop\\Internship proj
```

In [7]:
```
dataset.head().style.set_properties(**{'background-color':'pink'})
```

Out[7]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
0	892	0	3	Kelly, Mr. James	male	34.500000	0	0	330911	7.8292(
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.000000	1	0	363272	7.0000(
2	894	0	2	Myles, Mr. Thomas Francis	male	62.000000	0	0	240276	9.6875(
3	895	0	3	Wirz, Mr. Albert	male	27.000000	0	0	315154	8.6625(
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.000000	1	1	3101298	12.2875(

In [10]:
```
dataset.tail().style.set_properties(**{'background-color':'skyblue'})
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
413	1305	0	3	Spector, Mr. Woolf	male	nan	0	0	A.5. 3236	
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.000000	0	0	PC 17758	10
415	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.500000	0	0	SOTON/O.Q. 3101262	
416	1308	0	3	Ware, Mr. Frederick	male	nan	0	0	359309	
417	1309	0	3	Peter, Master. Michael J	male	nan	1	1	2668	2

In [11]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Survived     418 non-null    int64  
 2   Pclass       418 non-null    int64  
 3   Name         418 non-null    object 
 4   Sex          418 non-null    object 
 5   Age          332 non-null    float64
 6   SibSp        418 non-null    int64  
 7   Parch        418 non-null    int64  
 8   Ticket       418 non-null    object 
 9   Fare         417 non-null    float64
 10  Cabin        91 non-null     object 
 11  Embarked     418 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 39.3+ KB
```

In [12]:
```
dataset.describe()
```

Out[12]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	0.363636	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.481622	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	0.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	0.000000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	0.000000	3.000000	27.000000	0.000000	0.000000	14.454200
75%	1204.750000	1.000000	3.000000	39.000000	1.000000	0.000000	31.500000
max	1309.000000	1.000000	3.000000	76.000000	8.000000	9.000000	512.329200

In [13]:
```
dataset.isnull().sum()
```

Out[13]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

In [14]:
```
dataset.drop(columns =['Cabin', 'Name','Ticket','PassengerId'],inplace=True)
```

In [17]:
```
dataset['Survived'].value_counts()
```

Out[17]:
```
0    266
1    152
Name: Survived, dtype: int64
```

In [18]:
```
dataset['Sex'].value_counts()
```

Out[18]:
```
male      266
female    152
Name: Sex, dtype: int64
```

In [19]:
```
dataset
```

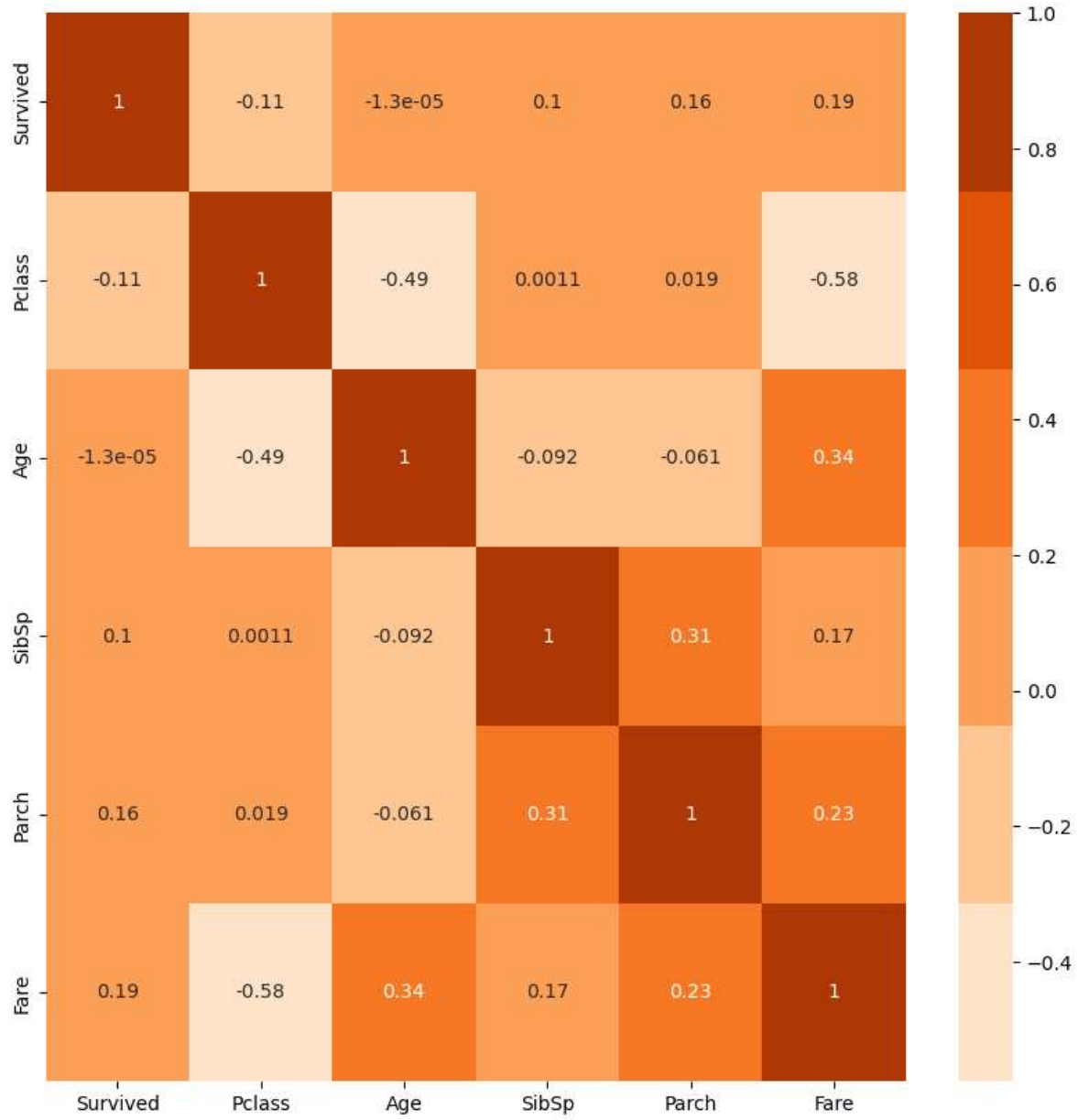
Out[19]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	34.5	0	0	7.8292	Q
1	1	3	female	47.0	1	0	7.0000	S
2	0	2	male	62.0	0	0	9.6875	Q
3	0	3	male	27.0	0	0	8.6625	S
4	1	3	female	22.0	1	1	12.2875	S
...	...	...	...	...	...	...	...	...
413	0	3	male	NaN	0	0	8.0500	S
414	1	1	female	39.0	0	0	108.9000	C
415	0	3	male	38.5	0	0	7.2500	S
416	0	3	male	NaN	0	0	8.0500	S
417	0	3	male	NaN	1	1	22.3583	C

418 rows × 8 columns

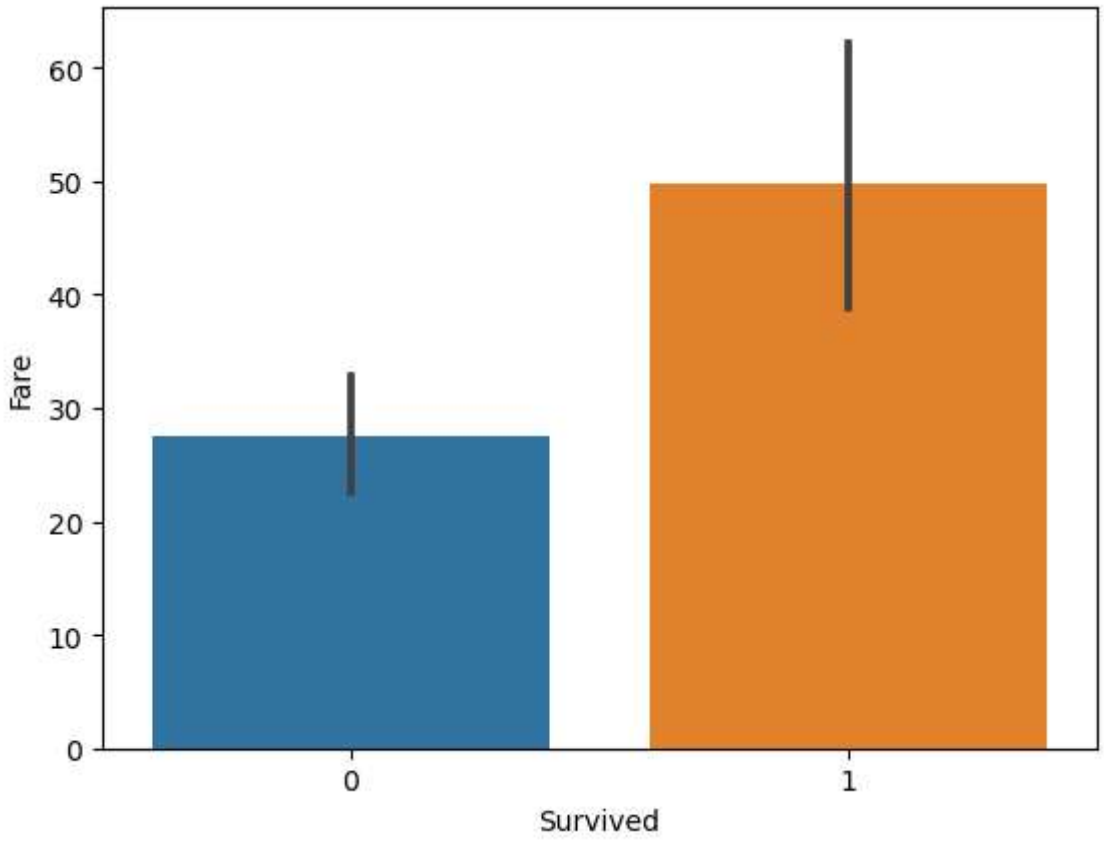
In [22]:
```
plt.figure(figsize=(10,10))
sns.heatmap(data=dataset.corr(),annot=True,cmap=sns.color_palette("Oranges"))
```

Out[22]: <Axes: >



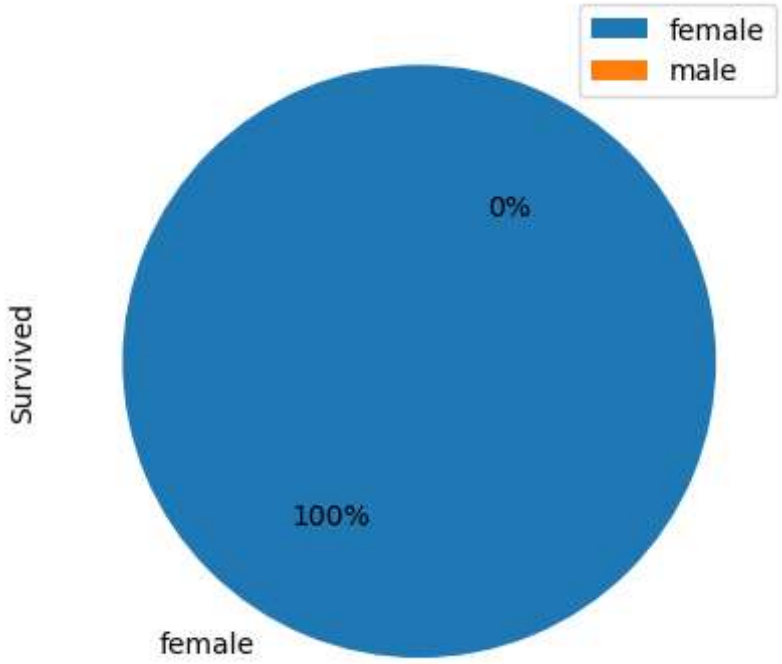
In [23]:
```
sns.barplot(y=dataset['Fare'],x=dataset["Survived"])
```

Out[23]:
```
<Axes: xlabel='Survived', ylabel='Fare'>
```



In [30]:
```
dataset.groupby(["Sex"]).sum().plot(kind='pie',y='Survived', autopct='%1.0f%%'
```

Out[30]:
```
<Axes: ylabel='Survived'>
```



In [25]:
```
dataset['Age']=dataset['Age'].fillna(dataset['Age'].mean())
dataset['Embarked']=dataset['Embarked'].fillna(dataset['Embarked'].mode()[0])
```

In [31]:
```
dataset.isnull().sum()
```

Out[31]:
```
Survived    0
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        1
Embarked    0
dtype: int64
```

In [32]:
```
dataset
```

Out[32]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	34.50000	0	0	7.8292	Q
1	1	3	female	47.00000	1	0	7.0000	S
2	0	2	male	62.00000	0	0	9.6875	Q
3	0	3	male	27.00000	0	0	8.6625	S
4	1	3	female	22.00000	1	1	12.2875	S
...	...	...	...	...	...	...	...	...
413	0	3	male	30.27259	0	0	8.0500	S
414	1	1	female	39.00000	0	0	108.9000	C
415	0	3	male	38.50000	0	0	7.2500	S
416	0	3	male	30.27259	0	0	8.0500	S
417	0	3	male	30.27259	1	1	22.3583	C

418 rows × 8 columns

In [33]:
```
from sklearn.preprocessing import LabelEncoder
```

In [34]:
```
le=LabelEncoder()
```

In [35]:
```
dataset['Sex']=le.fit_transform(dataset['Sex'])
dataset['Embarked']=le.fit_transform(dataset['Embarked'])
```

In [36]:
```
dataset['Sex'].value_counts()
```

Out[36]:
```
1    266
0    152
Name: Sex, dtype: int64
```

In [37]:
```
dataset['Sex'].value_counts()
```

Out[37]:
```
1    266
0    152
Name: Sex, dtype: int64
```

In [38]:
```
dataset['Embarked'].value_counts()
```

Out[38]:
```
2    270
0    102
1     46
Name: Embarked, dtype: int64
```

In [39]:
```
dataset
```

Out[39]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1	34.50000	0	0	7.8292	1
1	1	3	0	47.00000	1	0	7.0000	2
2	0	2	1	62.00000	0	0	9.6875	1
3	0	3	1	27.00000	0	0	8.6625	2
4	1	3	0	22.00000	1	1	12.2875	2
...	...	...	...	...	...	...	...	...
413	0	3	1	30.27259	0	0	8.0500	2
414	1	1	0	39.00000	0	0	108.9000	0
415	0	3	1	38.50000	0	0	7.2500	2
416	0	3	1	30.27259	0	0	8.0500	2
417	0	3	1	30.27259	1	1	22.3583	0

418 rows × 8 columns

In [40]:
```
x=dataset.iloc[:,1:]
```

In [41]:
```
y=dataset['Survived']
```

In [42]:
```
x
```

Out[42]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	1	34.50000	0	0	7.8292	1
1	3	0	47.00000	1	0	7.0000	2
2	2	1	62.00000	0	0	9.6875	1
3	3	1	27.00000	0	0	8.6625	2
4	3	0	22.00000	1	1	12.2875	2
...	...	...	...	...	...	...	...
413	3	1	30.27259	0	0	8.0500	2
414	1	0	39.00000	0	0	108.9000	0
415	3	1	38.50000	0	0	7.2500	2
416	3	1	30.27259	0	0	8.0500	2
417	3	1	30.27259	1	1	22.3583	0

418 rows × 7 columns

In [43]:
```
y
```

Out[43]:
```
0      0
1      1
2      0
3      0
4      1
      ..
413    0
414    1
415    0
416    0
417    0
Name: Survived, Length: 418, dtype: int64
```

In [44]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25)
x_train
```

Out[44]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
139	3	1	40.0	1	6	46.9000	2
343	1	0	58.0	0	1	512.3292	0
155	3	1	24.0	0	0	7.5500	2
109	2	1	18.5	0	0	13.0000	2
165	3	0	26.0	1	1	22.0250	2
...	...	...	...	...	...	...	...
374	1	0	54.0	1	1	81.8583	2
69	1	0	60.0	1	4	263.0000	2
212	2	1	17.0	0	0	73.5000	2
43	2	0	30.0	0	0	13.0000	2
295	3	1	26.0	0	0	7.8958	2

313 rows × 7 columns

In [45]:
```
y_train
```

Out[45]:
```
139    0
343    1
155    0
109    0
165    1
      ..
374    1
69     1
212    0
43     1
295    0
Name: Survived, Length: 313, dtype: int64
```

In [ ]: