In [51]:
```
import numpy as np
import pandas as pd
```

In [52]:
```
df = pd.read_csv("Titanic-Dataset.csv")
print(df.head())
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

In [53]:
```
print(df.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
None
```

In [54]:
```
df.isnull().sum()
```

Out[54]:
```
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 [55]:
```
df.Age = df.Age.fillna(np.mean(df['Age']))
```

In [56]:
```
df.Embarked = df.Embarked.fillna('S')    #S is the most repeated value
```

In [57]:
```
mode = df['Cabin'].mode().iloc[0]
df['Cabin'] = df['Cabin'].fillna(mode)
```

In [58]:
```
df.Cabin.value_counts()
```

Out[58]:
```
B96 B98        691
C23 C25 C27      4
G6               4
F33              3
E101             3
              ...
C95              1
C87              1
E46              1
B73              1
B30              1
Name: Cabin, Length: 147, dtype: int64
```

In [61]:
```
from sklearn.preprocessing import LabelEncoder

# Initialize the encoder
label_encoder = LabelEncoder()

# Fit and transform the data
df['Sex'] = label_encoder.fit_transform(df['Sex'])
```

In [62]:
```
df['Name'] = label_encoder.fit_transform(df['Name'])
```

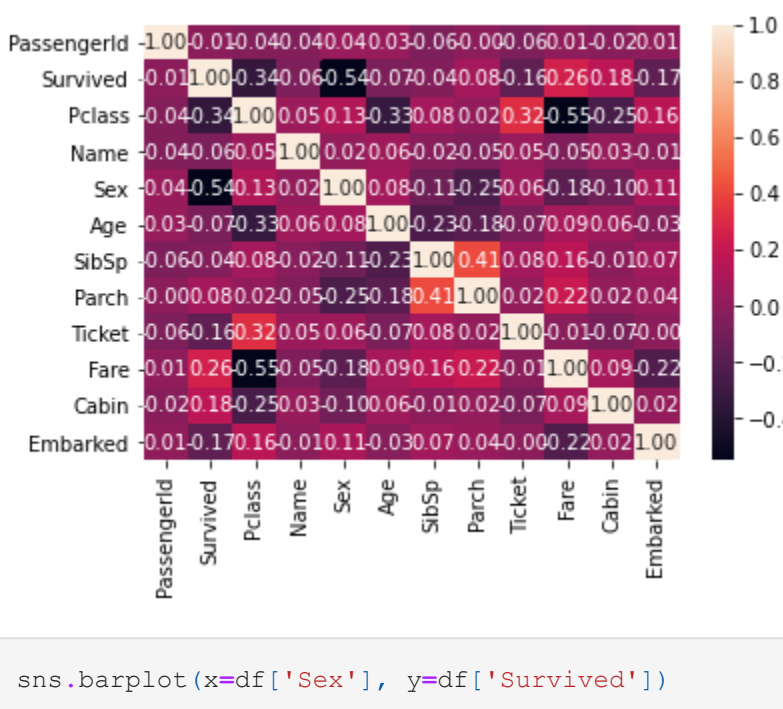
In [63]:
```
df['Ticket'] = label_encoder.fit_transform(df['Ticket'])
```

In [64]:
```
df['Cabin'] = label_encoder.fit_transform(df['Cabin'])
```

In [65]:
```
df['Embarked'] = label_encoder.fit_transform(df['Embarked'])
```

In [66]:
```
import seaborn as sns
sns.heatmap(df.corr(), annot=True, fmt = '.2f')
```

Out[66]:
```
<AxesSubplot:>
```

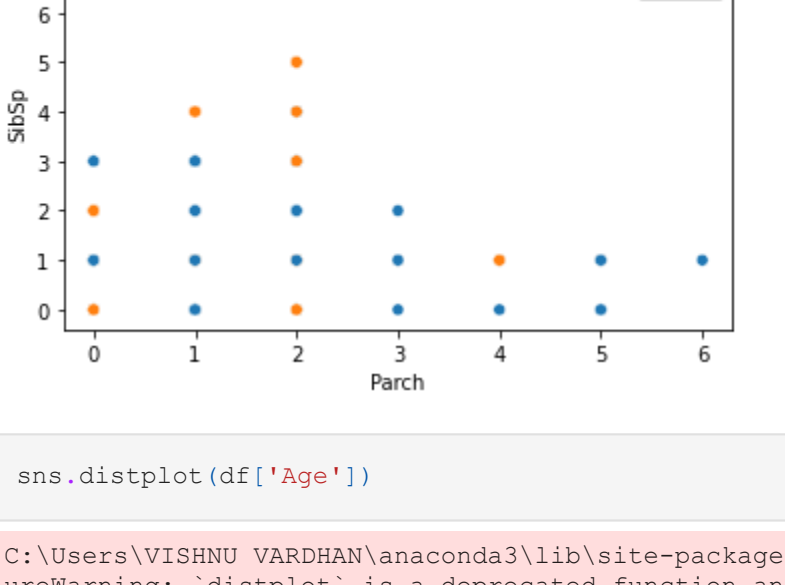


In [67]:
```
sns.barplot(x=df['Sex'], y=df['Survived'])
```

Out[67]:
```
<AxesSubplot:xlabel='Sex', ylabel='Survived'>
```

In [68]:
```
sns.scatterplot(x=df['Parch'], y=df['SibSp'], hue = df['Sex'])
```

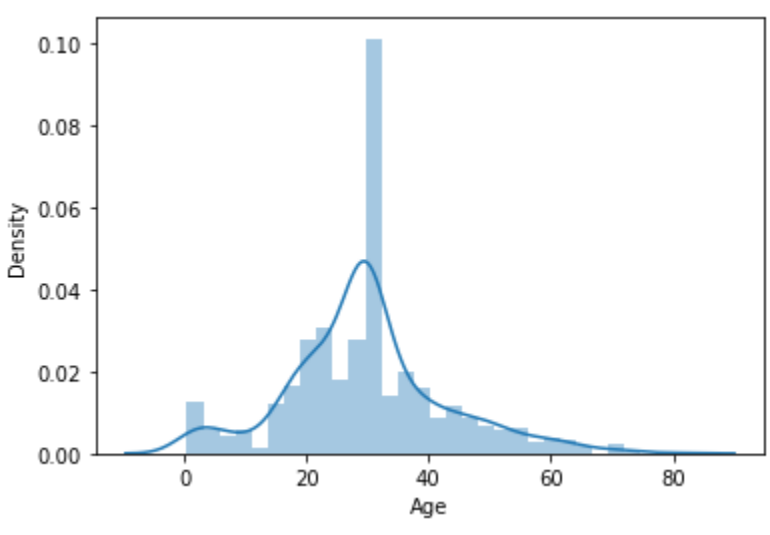
Out[68]:
```
<AxesSubplot:xlabel='Parch', ylabel='SibSp'>
```



In [69]:
```
sns.distplot(df['Age'])
```

```
C:\Users\VISHNU VARDHAN\anaconda3\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[69]:
```
<AxesSubplot:xlabel='Age', ylabel='Density'>
```



In [74]:
```
X = df.drop(['Age','Ticket','Name','Fare','Cabin'],axis=1)
y = df['Survived']
```

In [75]:
```
from sklearn.preprocessing import MinMaxScaler

# Assuming X is your feature matrix (a 2D array or DataFrame)
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(df)
```

In [76]:
```
from sklearn.model_selection import train_test_split
```

In [77]:
```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2, random_state=0)
```

In [78]:
```
X_train.shape
```

Out[78]:
```
(712, 7)
```

In [79]:
```
X_test.shape
```

Out[79]:
```
(179, 7)
```

In [80]:
```
y_train.shape
```

Out[80]:
```
(712,)
```

In [81]:
```
y_test.shape
```

Out[81]:
```
(179,)
```

In [ ]:
```

```