Classifying Titanic

December 26, 2022

[ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set_style('whitegrid')
import warnings
warnings.filterwarnings('ignore')
```

[ ]:
```
df = pd.read_csv('train.csv')
df
```

[ ]:
```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         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
..                                                 …     …   …    …
886                            Montvila, Rev. Juozas    male  27.0      0
887                     Graham, Miss. Margaret Edith  female  19.0      0
888         Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                            Behr, Mr. Karl Howell    male  26.0      0
890                              Dooley, Mr. Patrick    male  32.0      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
..     …               …      …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

[ ]:
```
df.columns
```

[ ]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

[ ]:
```
df.shape
```

[ ]:
```
(891, 12)
```

[ ]:
```
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
```

[ ]:
```
df.describe()
```

[ ]:
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

[ ]:
```
df.nunique ()
```

[ ]:
```
PassengerId    891
Survived         2
Pclass           3
Name           891
Sex              2
Age             88
SibSp            7
Parch            7
Ticket         681
Fare           248
Cabin          147
Embarked         3
dtype: int64
```

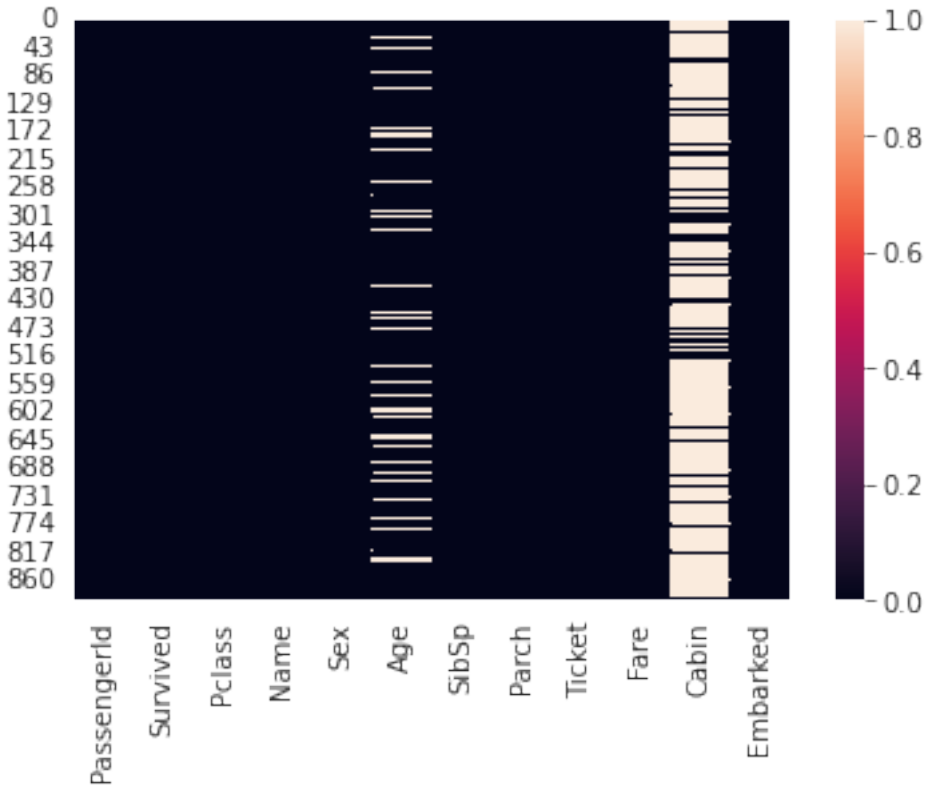
[ ]:
```
df.isnull().sum()
```

[ ]:
```
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
```

[ ]:
```
sns.heatmap(df.isnull())
```

[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5194fadd50>
```



[ ]:
```
display(df[['Sex','Survived']].groupby(df['Sex'],as_index=False).sum(),round(3))
```

```
   Survived
0       233
1       109


3
```

[ ]:
```
df.Sex.value_counts()
```

[ ]:
```
male      577
female    314
Name: Sex, dtype: int64
```

[ ]:
```
male = df.loc[df.Sex == 'male']
```

[ ]:
```
Survived_male = sum(male.Survived)
print('Total Survived male = ',Survived_male)
```

```
Total Survived male =  109
```

[ ]:
```
Per_male = sum(male.Survived)/len(male)
print('% male = ', Per_male)
```

```
% male =  0.18890814558058924
```

[ ]:
```
female = df.loc[df.Sex == 'female']
```

[ ]:
```
Survived_female = sum(female.Survived)
print('Total survived female = ', Survived_female)
```

```
Total survived female =  233
```

[ ]:
```
Per_female = sum(female.Survived)/len(female)
print('% female = ', Per_female)
```

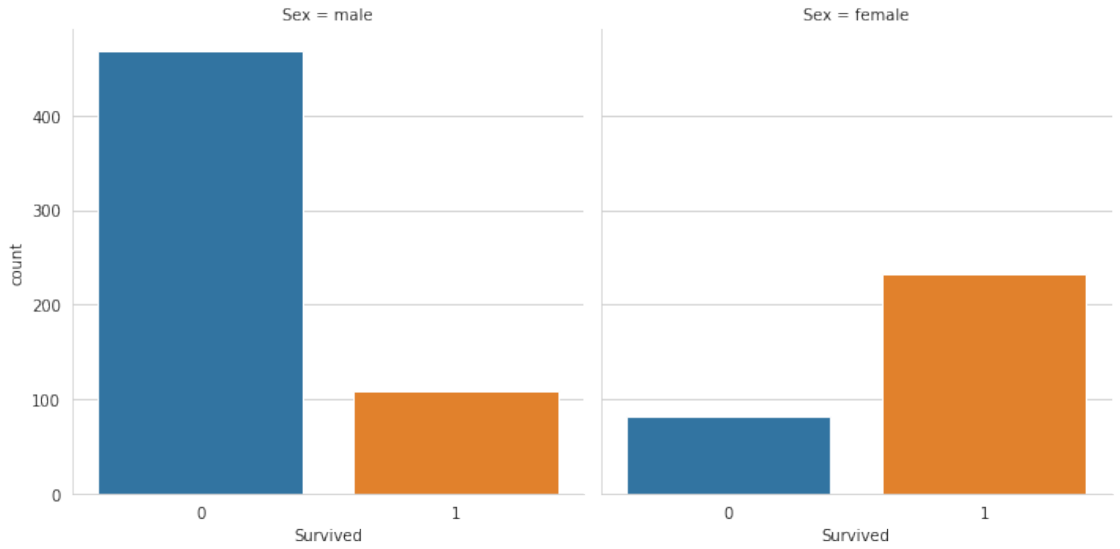
```
% female =  0.7420382165605095
```

[ ]:
```
display(df[['Pclass','Survived']].groupby(['Pclass'],as_index=False).
↪sum(),round(3))
```

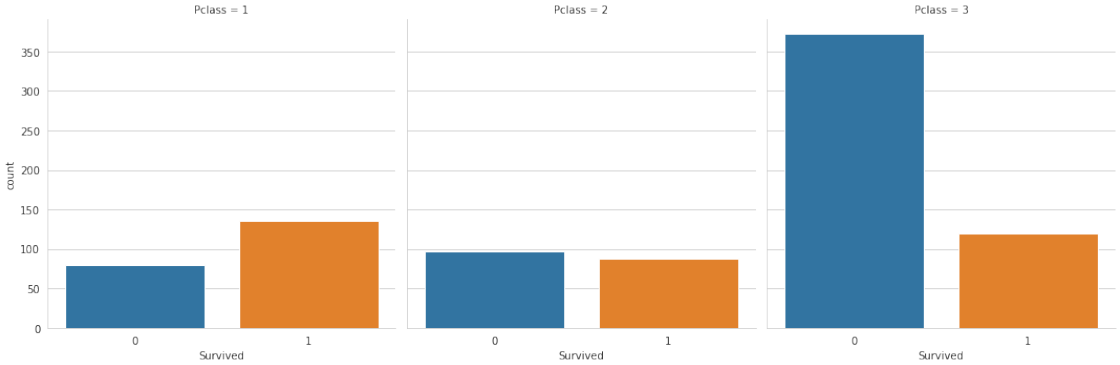
```
   Pclass  Survived
0       1       136
1       2        87
2       3       119


3
```

[ ]:
```
sns.factorplot(x='Survived',
    col='Sex',
    kind='count',
    data=df)
```

[ ]:
```
<seaborn.axisgrid.FacetGrid at 0x7f5194f82d10>
```

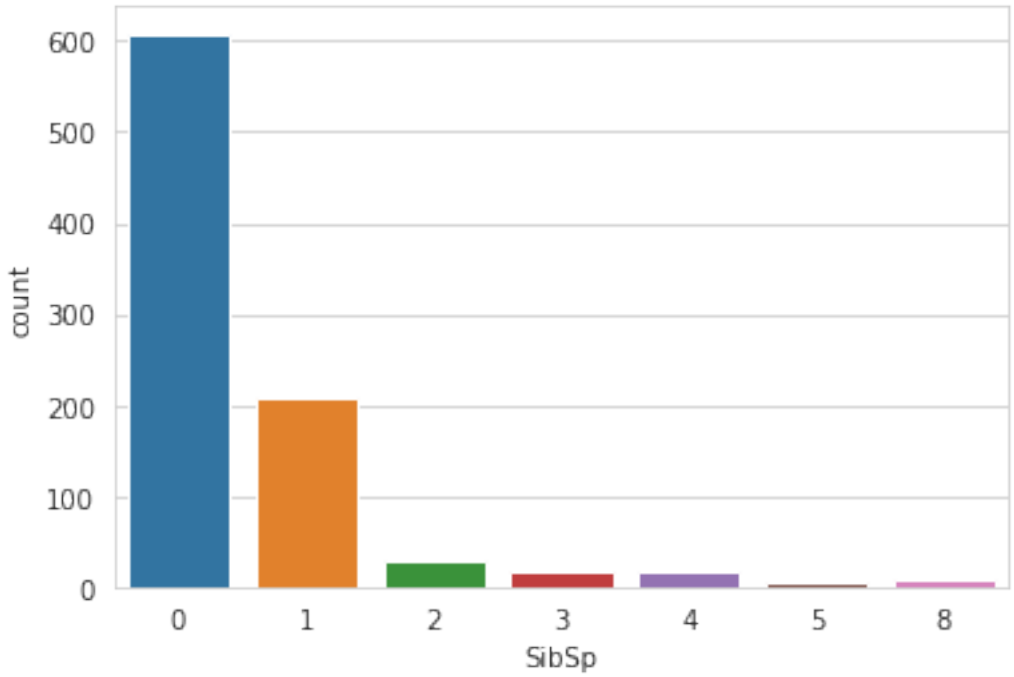


[ ]:
```
sns.factorplot(x='Survived',
    col='Pclass',
    kind='count',
    data=df)
plt.show()
```



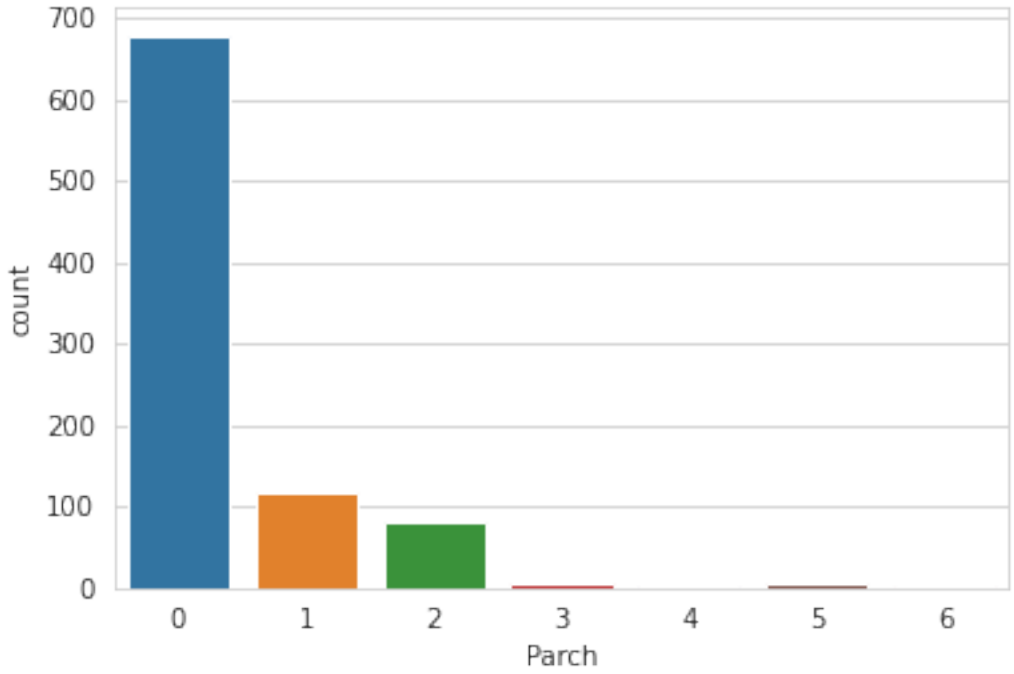
[ ]:
```
sns.countplot('SibSp',data=df)
```

[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5191758f50>
```



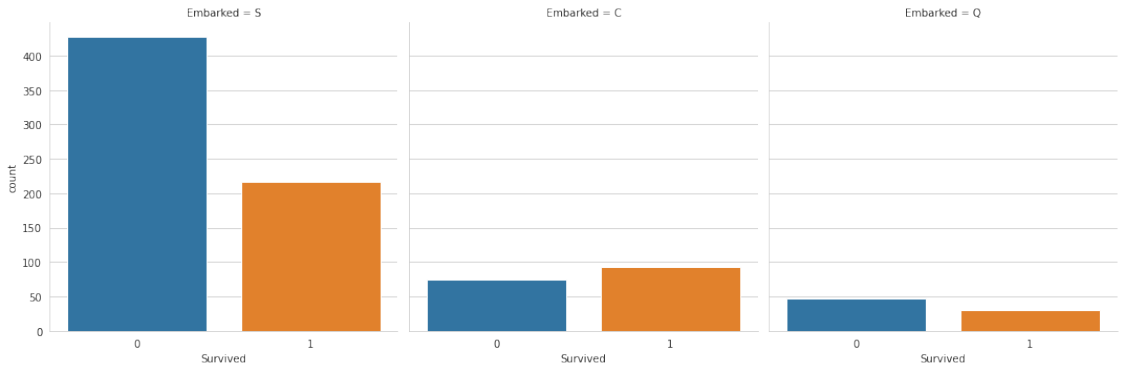
[ ]:
```
sns.countplot('Parch',data=df)
```

[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f51917399d0>
```



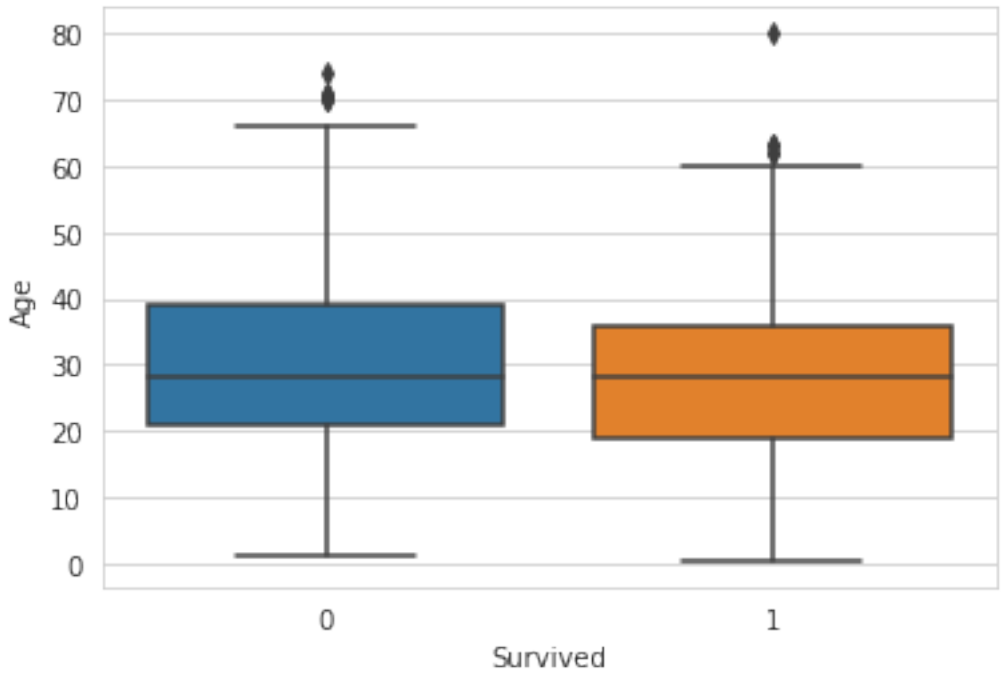
[ ]:
```
sns.factorplot(x='Survived',
    col='Embarked',
    kind='count',
    data=df)
```

[ ]:
```
<seaborn.axisgrid.FacetGrid at 0x7f51917a9b10>
```



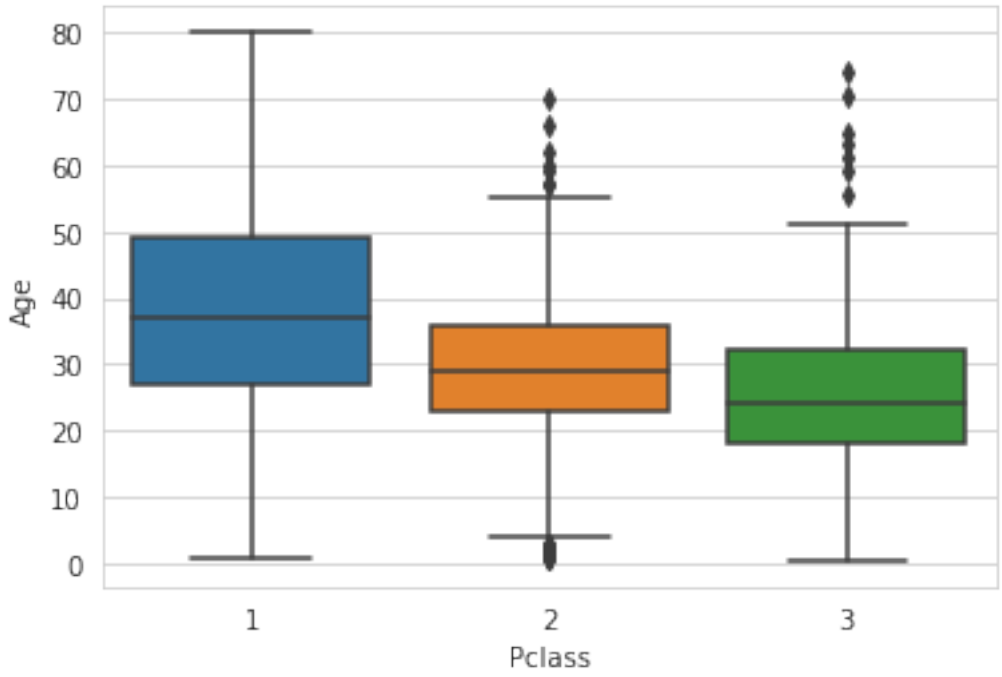
[ ]:
```
# Age and Survived
sns.boxplot(x='Survived',y='Age',data=df)
```

[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f51915ac050>
```



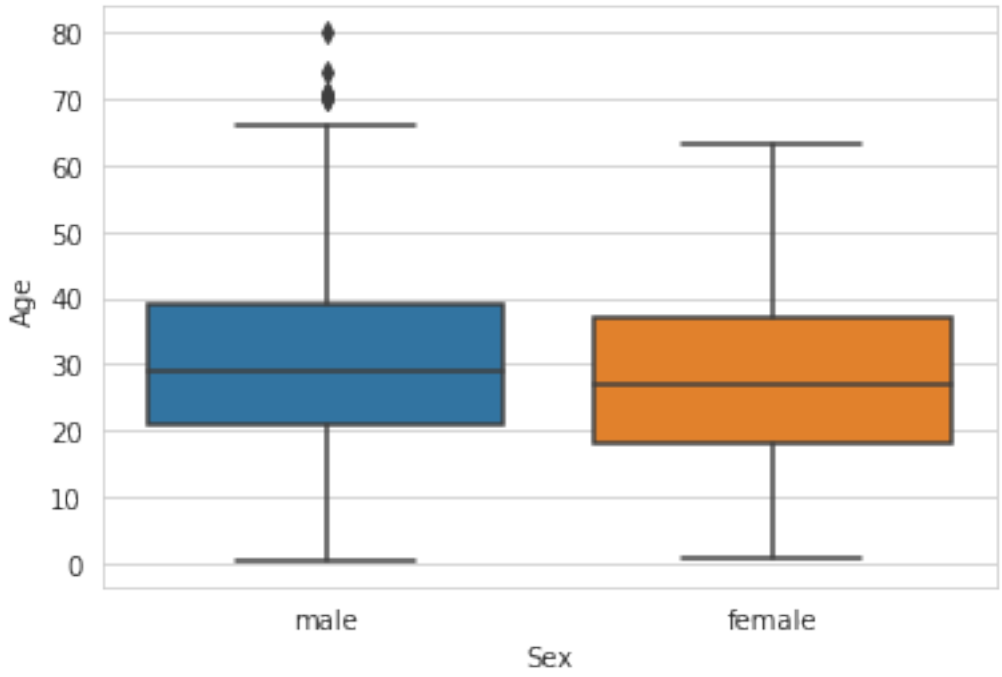
[ ]:
```
# Age and Pclass
sns.boxplot(x='Pclass',y='Age',data=df)
```

[ ]: <matplotlib.axes._subplots.AxesSubplot at 0x7f51915f0a50>



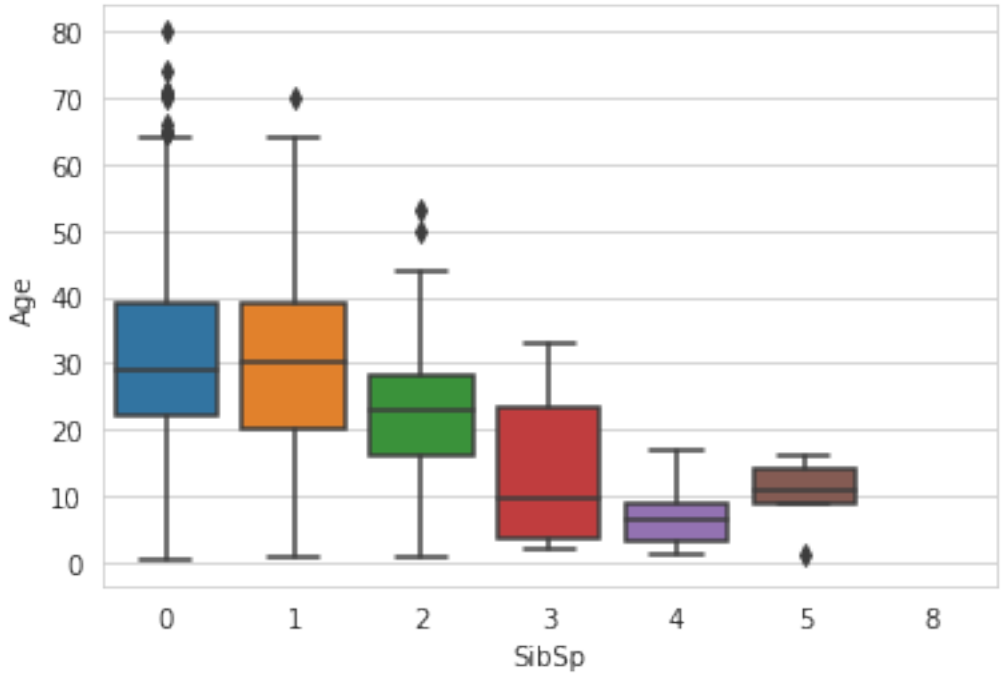
[ ]:
```
# Age and Sex
sns.boxplot(x='Sex',y='Age',data=df)
```

[ ]: <matplotlib.axes._subplots.AxesSubplot at 0x7f51914222d0>



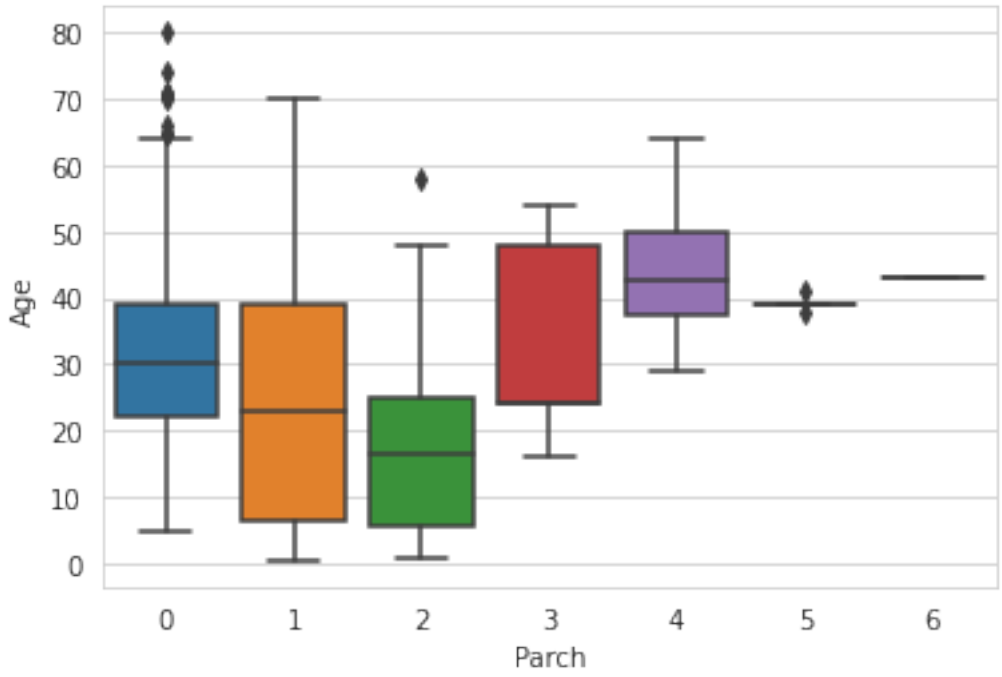
[ ]:
```
# SipSb and Age
sns.boxplot(x='SibSp',y='Age',data=df)
```

[ ]: <matplotlib.axes._subplots.AxesSubplot at 0x7f51913a3710>



[ ]:
```
# Parch and Age
sns.boxplot(x='Parch',y='Age',data=df)
```

[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5191295790>
```



[ ]:
```
df.columns
```

[ ]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

[ ]:
```
df.
↪drop(['PassengerId','Embarked','Name','Ticket','Fare','Cabin'],axis=1,inplace=True)
```

[ ]:
```
df
```

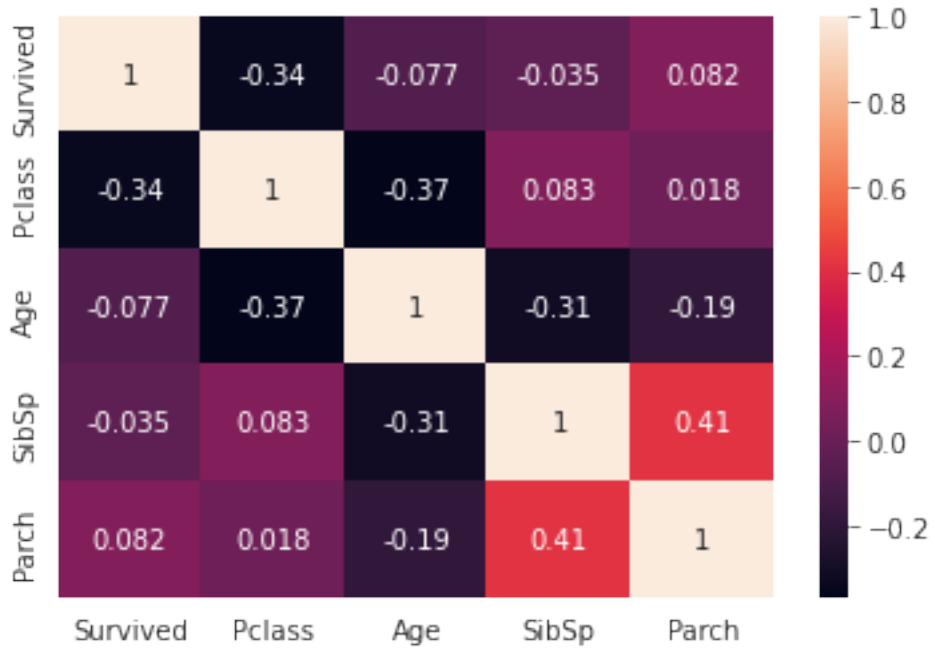
[ ]:

	Survived	Pclass	Sex	Age	SibSp	Parch
0	0	3	male	22.0	1	0
1	1	1	female	38.0	1	0
2	1	3	female	26.0	0	0
3	1	1	female	35.0	1	0
4	0	3	male	35.0	0	0

```
..          …      …      …    …     …     …
886         0      2      male  27.0     0     0
887         1      1    female  19.0     0     0
888         0      3    female   NaN     1     2
889         1      1      male  26.0     0     0
890         0      3      male  32.0     0     0

[891 rows x 6 columns]
```

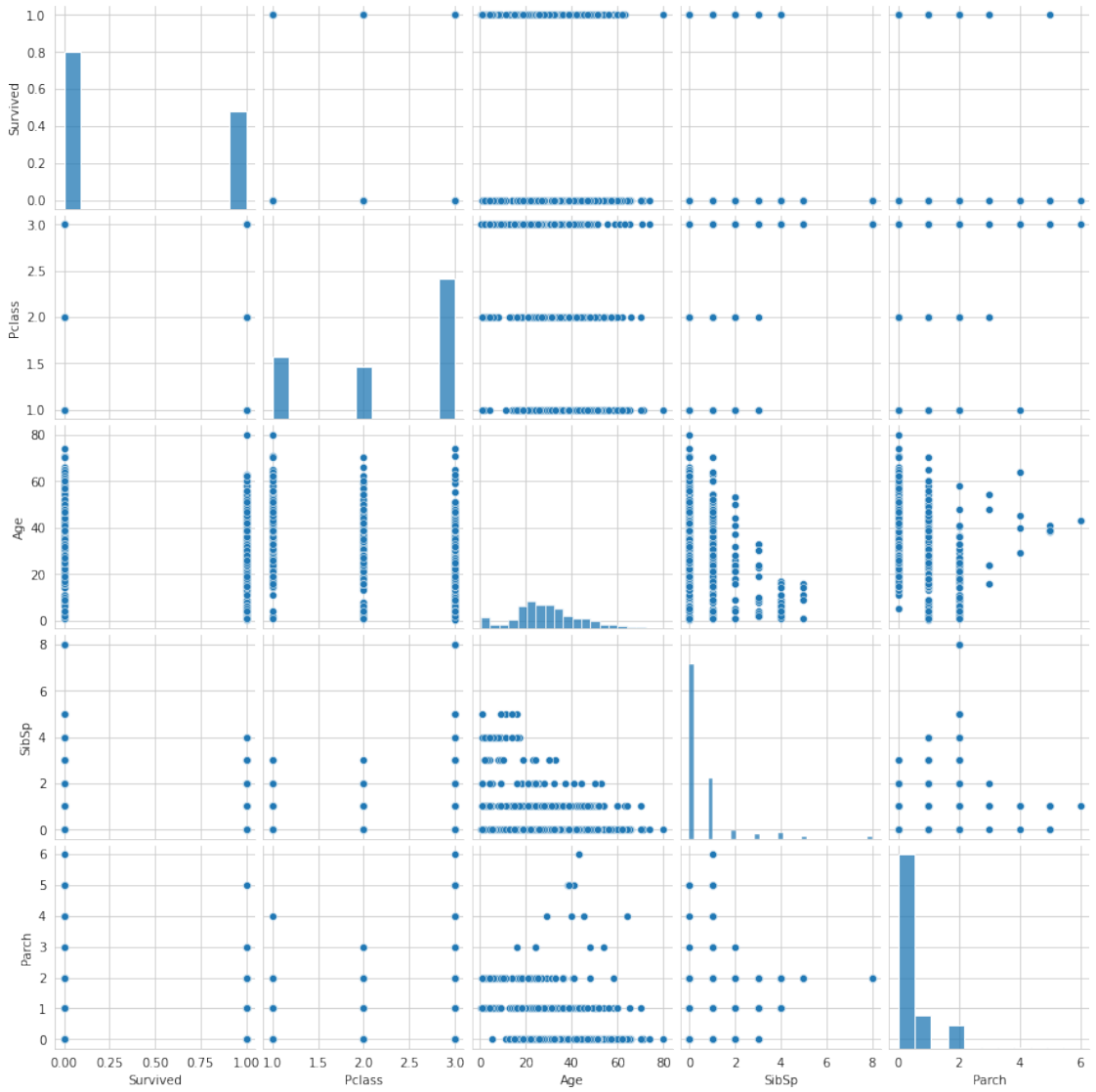
[ ]:
```
corr = df.corr()
sns.heatmap(corr, annot = True)
```

[ ]: <matplotlib.axes._subplots.AxesSubplot at 0x7f5191210bd0>



[ ]:
```
sns.pairplot(df)
```

[ ]: <seaborn.axisgrid.PairGrid at 0x7f5190fc6790>



[ ]:
```
!apt-get install texlive texlive-xetex texlive-latex-extra pandoc
!pip install pypandoc
```

[ ]:
```
!jupyter nbconvert --to PDF '
```