EDA for Titanic Dataset

1)Importing Libraries and Data

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [2]:

```
#from tabula import convert_into
```

In [3]:

```
# df = tabula.read_pdf("PassengerId-200611-000941.pdf", pages='1-16')
```

In [4]:

```
# df
```

In [5]:

```
# convert_into("PassengerId-200611-000941.pdf", "test_s.csv", output_format="csv",pages =
'1-16')
```

In [6]:

```
dataset = pd.read_csv("test_s.csv")
```

In [7]:

```
dataset
```

Out[7]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emb
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Cumings, Mrs. John Bradley\r(Florence Briggs T...	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2.\r3101282	7.9250	NaN	
3	4	1	1	Futrelle, Mrs. Jacques Heath\r(Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	
871	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	
872	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	
				Johnston, Miss								

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emb
873	889	0	3	Johnston, Miss. Catherine Helen\r"Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	
874	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	
875	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	

876 rows × 12 columns

In [8]:

```
dataset.head()
```

Out[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarl
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Cumings, Mrs. John Bradley\r(Florence Briggs T...	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2.\r3101282	7.9250	NaN	
3	4	1	1	Futrelle, Mrs. Jacques Heath\r(Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	

2)Variable Identification

Dependent Variable:

Survived

Independent Variables/Predictor Variables:

1.PassengerId 2.Pclass 3.Sex 4.Age 5.SibSp 6.Parch 7.Fare 8.Embarked

In [9]:

```
dataset.head()
```

Out[9]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarl
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Cumings, Mrs. John Bradley\r(Florence Briggs T...	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2.\r3101282	7.9250	NaN	
3	4	1	1	Futrelle, Mrs. Jacques Heath\r(Lily May	female	35.0	1	0	113803	53.1000	C123	

Peel)

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarl
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	

In [10]:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 876 entries, 0 to 875
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  876 non-null    int64  
 1   Survived     876 non-null    int64  
 2   Pclass       876 non-null    int64  
 3   Name         876 non-null    object 
 4   Sex          876 non-null    object 
 5   Age          701 non-null    float64
 6   SibSp        876 non-null    int64  
 7   Parch        876 non-null    int64  
 8   Ticket       876 non-null    object 
 9   Fare         876 non-null    float64
 10  Cabin        202 non-null    object 
 11  Embarked     874 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 82.2+ KB
```

In [11]:

```
dataset.describe()
```

Out[11]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	876.000000	876.000000	876.000000	701.000000	876.000000	876.000000	876.000000
mean	445.929224	0.384703	2.304795	29.719215	0.528539	0.385845	32.391794
std	257.600137	0.486803	0.836059	14.583577	1.110102	0.809645	50.020501
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	222.750000	0.000000	2.000000	20.000000	0.000000	0.000000	7.917700
50%	446.500000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.250000	1.000000	3.000000	38.000000	1.000000	0.000000	31.068750
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [12]:

```
dataset
```

Out[12]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emb
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Cumings, Mrs. John Bradley\r(Florence Briggs T...	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2.\r3101282	7.9250	NaN	
3	4	1	1	Futrelle, Mrs. Jacques Heath\r(Lily May	female	35.0	1	0	113803	53.1000	C123	

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emb
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	
871	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	
872	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	
873	889	0	3	Johnston, Miss. Catherine Helen\r"Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	
874	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	
875	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	

876 rows × 12 columns

In [13]:

```
dataset.shape
```

Out[13]:

```
(876, 12)
```

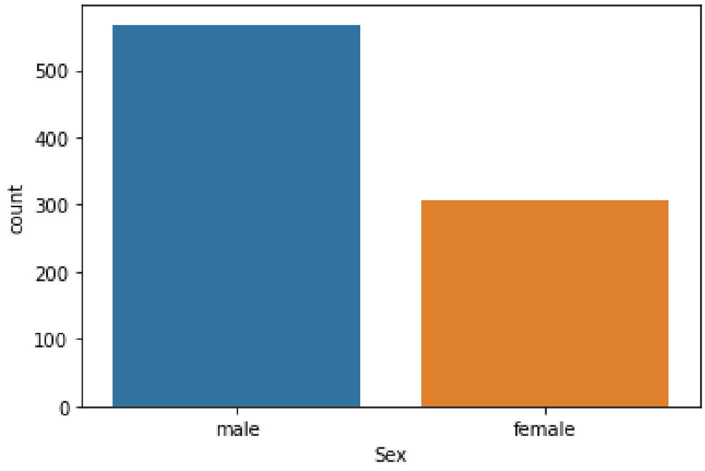
3)Univariate Analysis

In [14]:

```
sns.countplot(x='Sex', data=dataset)
```

Out[14]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f3b4388>
```

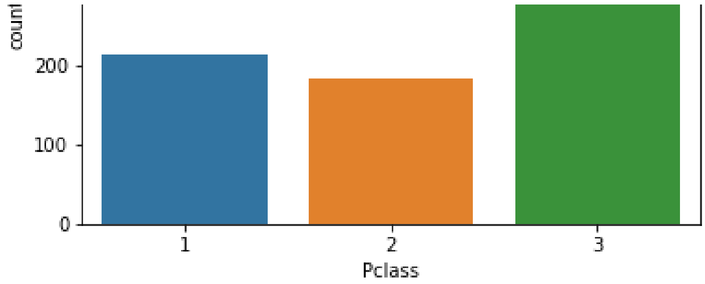


In [15]:

```
sns.countplot(x='Pclass', data=dataset)
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f6b4a48>
```

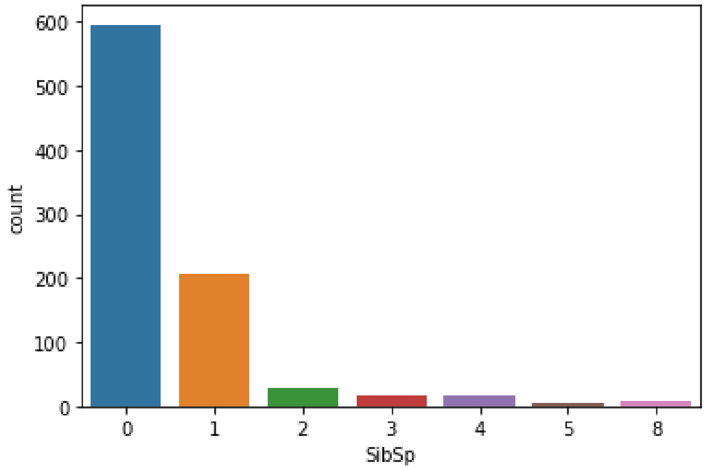


In [16]:

```
sns.countplot(x='SibSp', data=dataset)
```

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f724cc8>
```

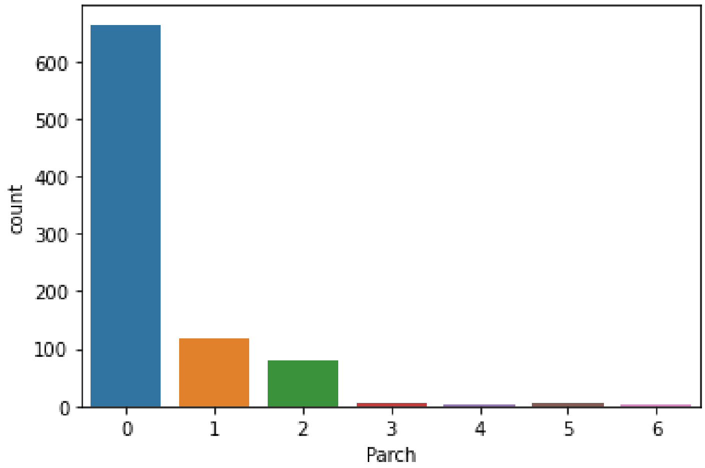


In [17]:

```
sns.countplot(x='Parch', data=dataset)
```

Out[17]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f7a2188>
```

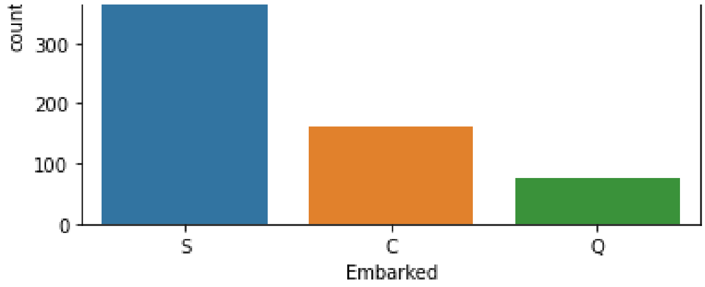


In [18]:

```
sns.countplot(x='Embarked', data=dataset)
```

Out[18]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f823688>
```

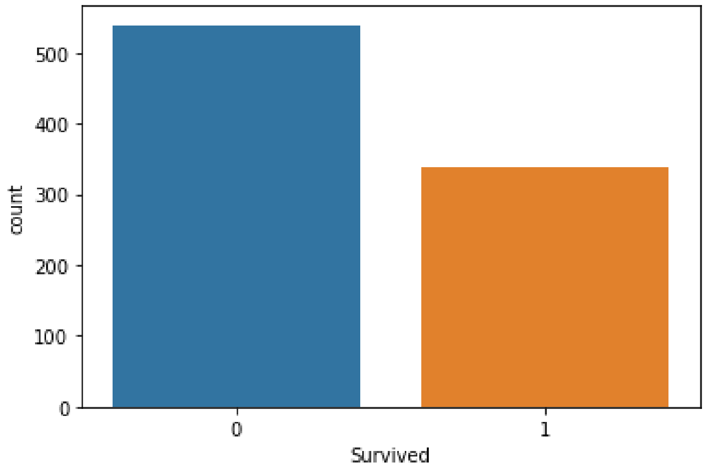


In [19]:

```
sns.countplot(x='Survived', data=dataset)
```

Out[19]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f89c848>
```

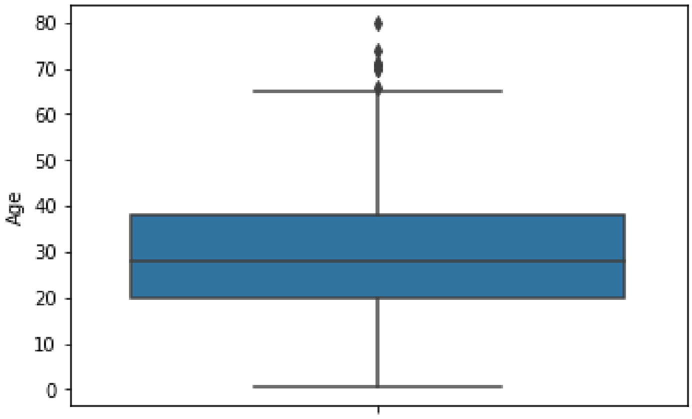


In [20]:

```
sns.boxplot(y = 'Age', data = dataset)
```

Out[20]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f82f308>
```

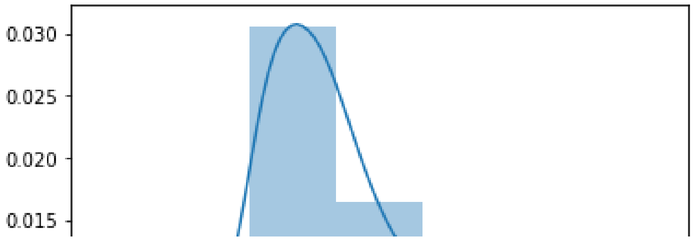


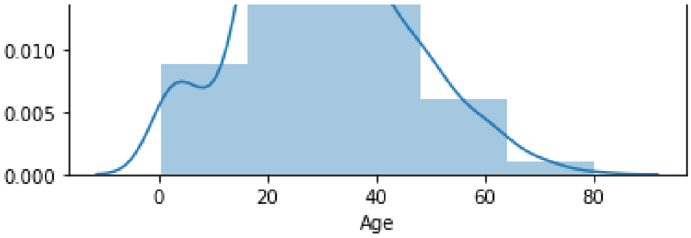
In [21]:

```
sns.distplot(dataset['Age'], bins=5)
```

Out[21]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f93fc48>
```



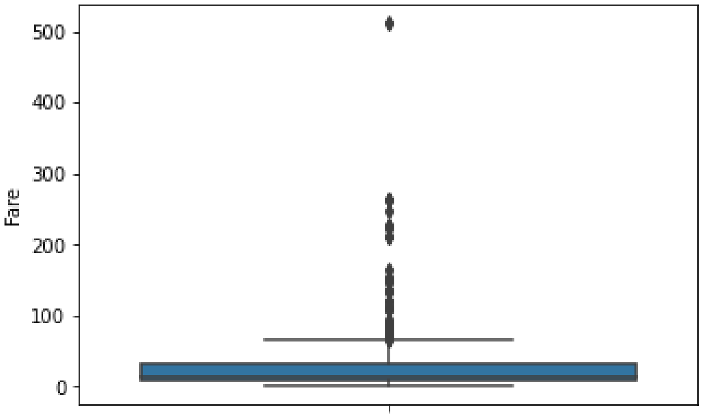


In [22]:

```
sns.boxplot(y='Fare', data=dataset)
```

Out[22]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f997888>
```

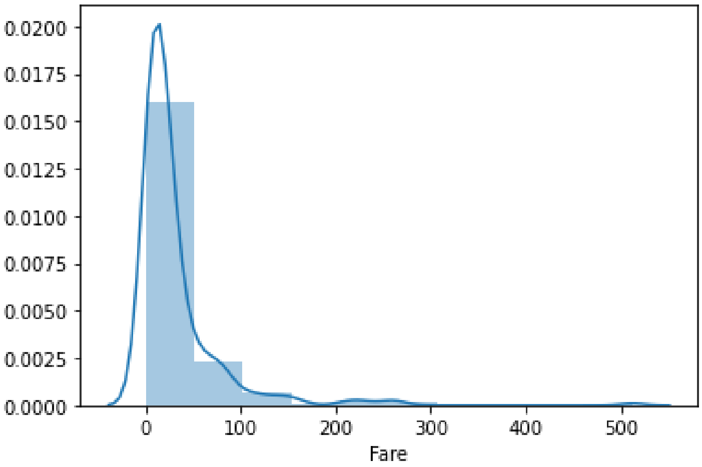


In [23]:

```
sns.distplot(dataset['Fare'], bins=10)
```

Out[23]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77f9d6e08>
```



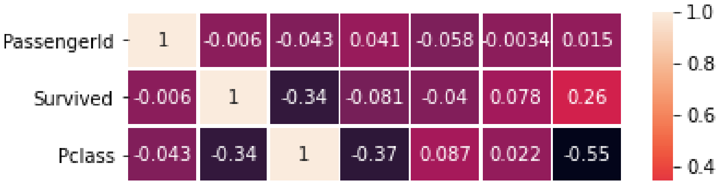
4)Bivariate Analysis

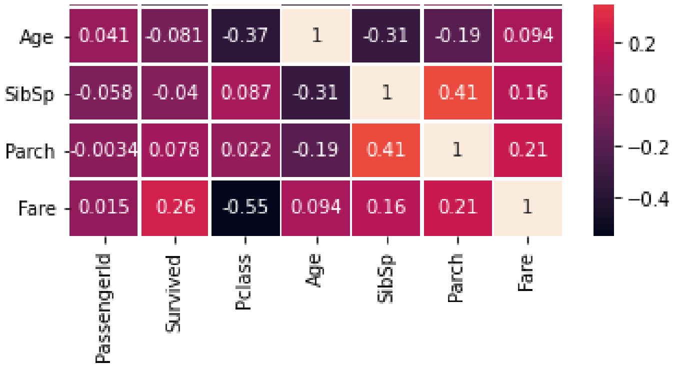
In [24]:

```
sns.heatmap(dataset.corr(), annot=True, linewidth = 0.5)
```

Out[24]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e77fadb2c8>
```



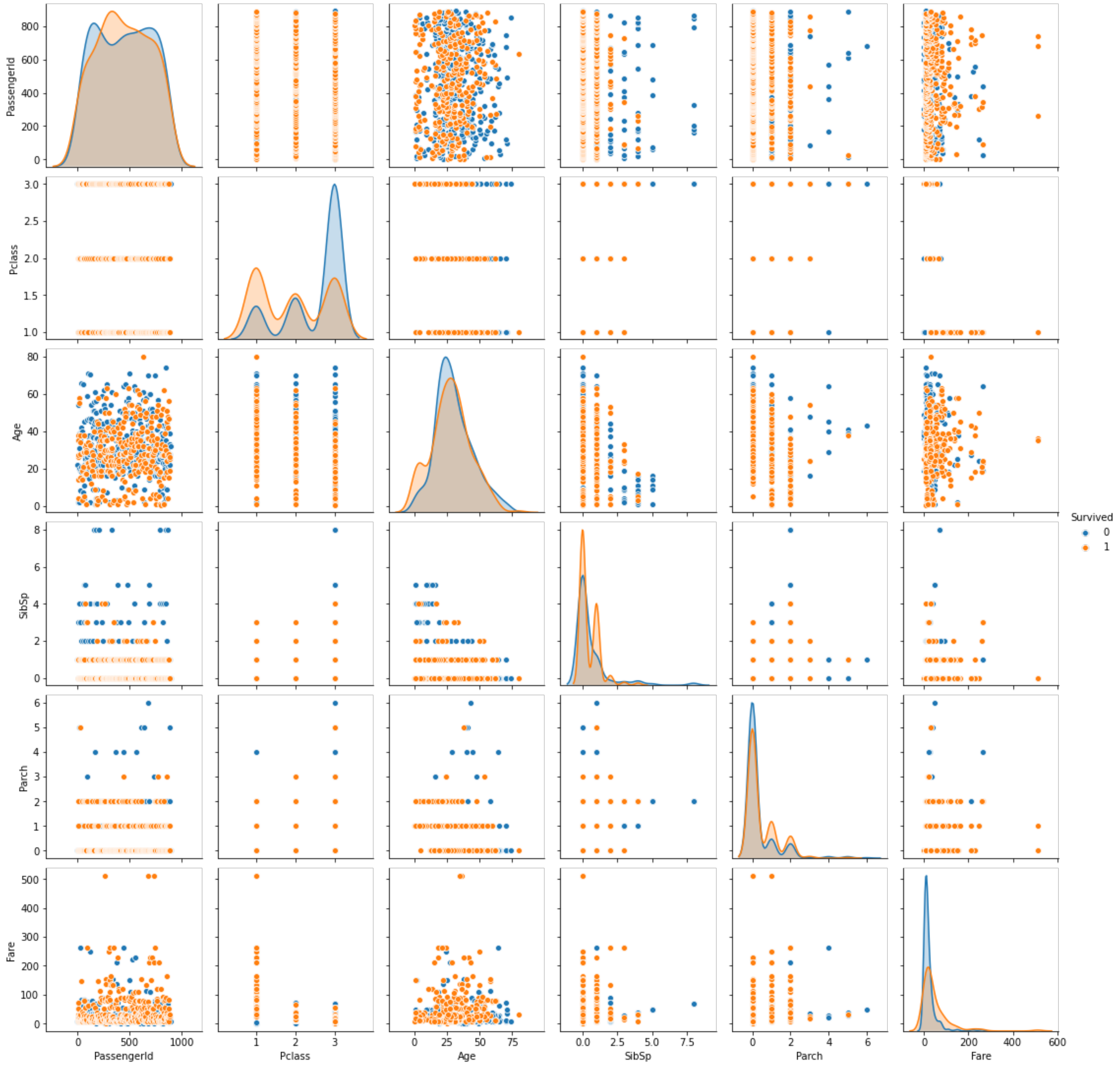


In [25]:

```
sns.pairplot(dataset,hue = 'Survived',dropna = True)
```

Out[25]:

```
<seaborn.axisgrid.PairGrid at 0x1e77fbf3588>
```



In [26]:

```
counts = dataset.groupby(['Survived', 'Sex'], axis= 0)
counts.size()
```

Out[26]:

```
Survived  Sex
0         female     79
          male      460
1         female    229
          male      108
dtype: int64
```

In [27]:

```
counts = dataset.groupby(['Survived', 'Pclass'], axis= 0)
counts.size()
```

Out[27]:

```
Survived  Pclass
0         1          78
          2          97
          3         364
1         1         135
          2          86
          3         116
dtype: int64
```

In [28]:

```
counts = dataset.groupby(['Survived', 'Parch'], axis= 0)
counts.size()
```

Out[28]:

```
Survived  Parch
0         0        435
          1         53
          2         40
          3          2
          4          4
          5          4
          6          1
1         0        229
          1         65
          2         39
          3          3
          5          1
dtype: int64
```

In [29]:

```
counts = dataset.groupby(['Survived', 'SibSp'], axis= 0)
counts.size()
```

Out[29]:

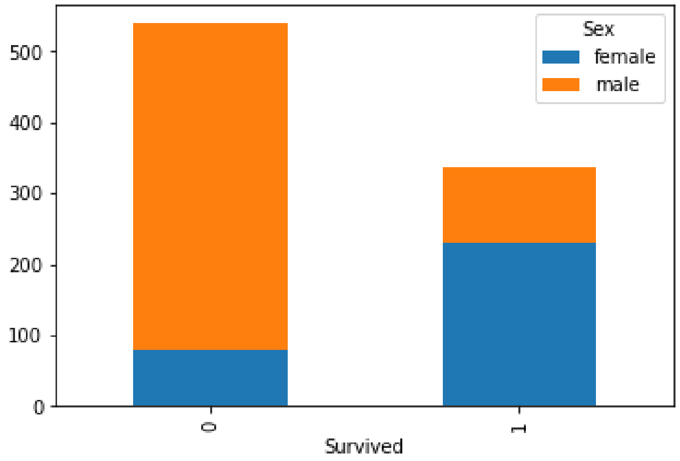
```
Survived  SibSp
0         0        388
          1         97
          2         15
          3         12
          4         15
          5          5
          8          7
1         0        208
          1        109
          2         13
          3          4
          4          3
dtype: int64
```

In [30]:

```
pd.crosstab(dataset['Survived'],dataset['Sex']).plot(kind='bar',stacked=True)
```

Out[30]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e701e20148>
```

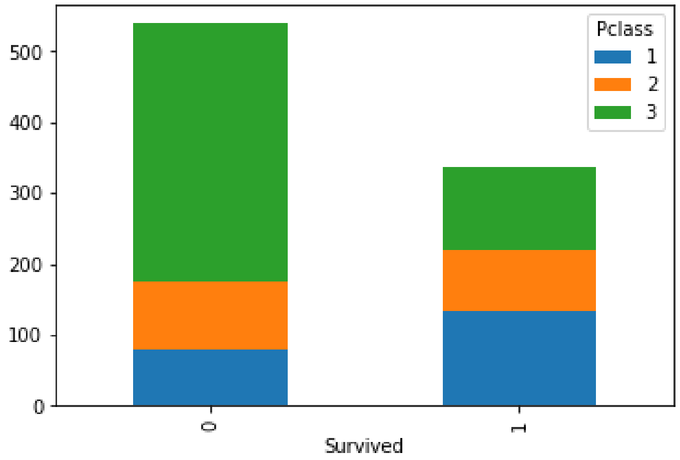


In [31]:

```
pd.crosstab(dataset['Survived'],dataset['Pclass']).plot(kind='bar',stacked=True)
```

Out[31]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e701df06c8>
```



5)Missing Values Treatment

In [32]:

```
dataset.isnull()
```

Out[32]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
871	False	False	False	False	False	False	False	False	False	False	True	False
872	False	False	False	False	False	False	False	False	False	False	False	False
873	False	False	False	False	False	True	False	False	False	False	True	False
874	False	False	False	False	False	False	False	False	False	False	False	False
875	False	False	False	False	False	False	False	False	False	False	True	False

876 rows × 12 columns

In [33]:

```
dataset.isnull().values.any()
```

Out[33]:

```
True
```

In [34]:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 876 entries, 0 to 875
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  876 non-null    int64  
 1   Survived     876 non-null    int64  
 2   Pclass       876 non-null    int64  
 3   Name         876 non-null    object 
 4   Sex          876 non-null    object 
 5   Age          701 non-null    float64
 6   SibSp        876 non-null    int64  
 7   Parch        876 non-null    int64  
 8   Ticket       876 non-null    object 
 9   Fare         876 non-null    float64
 10  Cabin        202 non-null    object 
 11  Embarked     874 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 82.2+ KB
```

In [35]:

```
dataset.describe()
```

Out[35]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	876.000000	876.000000	876.000000	701.000000	876.000000	876.000000	876.000000
mean	445.929224	0.384703	2.304795	29.719215	0.528539	0.385845	32.391794
std	257.600137	0.486803	0.836059	14.583577	1.110102	0.809645	50.020501
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	222.750000	0.000000	2.000000	20.000000	0.000000	0.000000	7.917700
50%	446.500000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.250000	1.000000	3.000000	38.000000	1.000000	0.000000	31.068750
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [36]:

```
dataset.isnull().sum()
```

Out[36]:

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            175
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          674
```

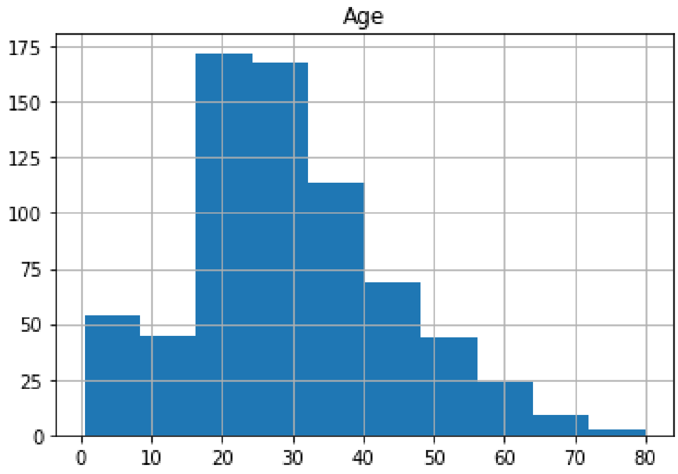
```
Embarked         2
dtype: int64
```

In [37]:

```
dataset.hist(column=['Age'], bins=10)
```

Out[37]:

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001E7029ECA48>]],
      dtype=object)
```



In [38]:

```
dataset['Age'].fillna(value=dataset['Age'].median(),inplace = True)
```

In [39]:

```
dataset
```

Out[39]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emba
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Cumings, Mrs. John Bradley\r(Florence Briggs T...	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2.\r3101282	7.9250	NaN	
3	4	1	1	Futrelle, Mrs. Jacques Heath\r(Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	
871	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	
872	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	
873	889	0	3	Johnston, Miss. Catherine Helen\r"Carrie"	female	28.0	1	2	W./C. 6607	23.4500	NaN	
874	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	
875	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	

876 rows × 12 columns

In [40]:

```
# dataset['Embarked'].fillna(value=dataset['Embarked'].mode(),inplace = True)
dataset.Embarked.fillna(dataset.Embarked.mode()[0], inplace = True)
```

In [41]:

```
dataset['Embarked'][60]
```

Out[41]:

```
'S'
```

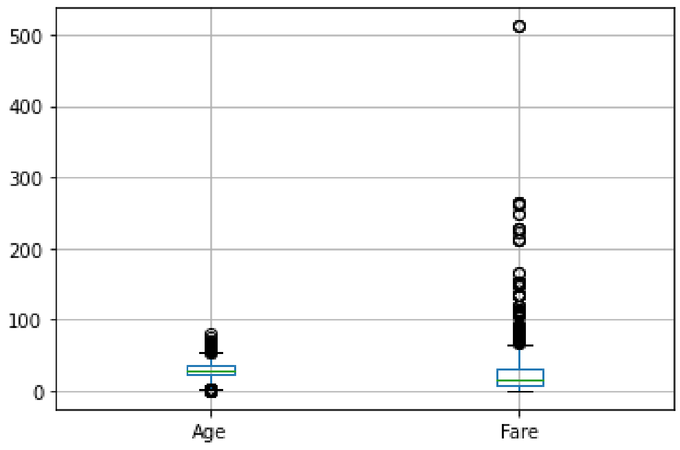
6)Outliers

In [42]:

```
dataset.boxplot(column=['Age','Fare'])
```

Out[42]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e702a18208>
```

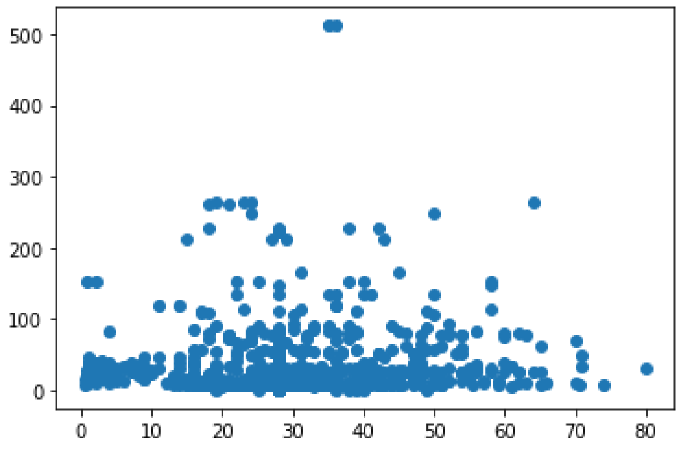


In [43]:

```
plt.scatter(dataset['Age'],dataset['Fare'])
```

Out[43]:

```
<matplotlib.collections.PathCollection at 0x1e702b0f988>
```



In [44]:

```
dataset['Age'].describe()
```

Out[44]:

```
count    876.000000
mean      29.375765
std       13.062068
min        0.420000
25%       22.000000
50%       28.000000
75%       35.000000
max       80.000000
Name: Age, dtype: float64
```

In [45]:

```
IQR_Age = dataset['Age'].quantile(0.75) - dataset['Age'].quantile(0.25)
print(IQR_Age)
```

```
13.0
```

In [46]:

```
Upper_OutlierLimit_Age = dataset['Age'].quantile(0.75) + 1.5*IQR_Age
Lower_OutlierLimit_Age = dataset['Age'].quantile(0.25) - 1.5*IQR_Age
print(Upper_OutlierLimit_Age)
print(Lower_OutlierLimit_Age)
```

```
54.5
2.5
```

In [47]:

```
OutlierValues_Age = dataset[(dataset['Age']>=Upper_OutlierLimit_Age) | (dataset[
'Age']<=Lower_OutlierLimit_Age)]
```

In [48]:

```
OutlierValues_Age
```

Out[48]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750	NaN	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.5500	C103	S
15	16	1	2	Hewlett, Mrs. (Mary D\rKingcome)	female	55.00	0	0	248706	16.0000	NaN	S
16	17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	29.1250	NaN	Q
33	34	0	2	Wheadon, Mr. Edward H	male	66.00	0	0	C.A. 24579	10.5000	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
813	828	1	2	Mallet, Master. Andre	male	1.00	0	2	S.C./PARIS 2079	37.0042	NaN	C
815	830	1	1	Stone, Mrs. George Nelson\r(Martha Evelyn)	female	62.00	0	0	113572	80.0000	B28	S
817	832	1	2	Richards, Master. George Sibley	male	0.83	1	1	29106	18.7500	NaN	S
836	852	0	3	Svensson, Mr. Johan	male	74.00	0	0	347060	7.7750	NaN	S
864	880	1	1	Potter, Mrs. Thomas Jr (Lily\rAlexenia Wilson)	female	56.00	0	1	11767	83.1583	C50	C

66 rows × 12 columns

In [49]:

```
dataset.loc[dataset.Age > 54.5,'Age'] =  dataset['Age'].quantile(0.95)
dataset.loc[dataset.Age < 2.5,'Age'] =  dataset['Age'].quantile(0.05)
```

In [50]:

```
dataset['Age']
```

Out[50]:

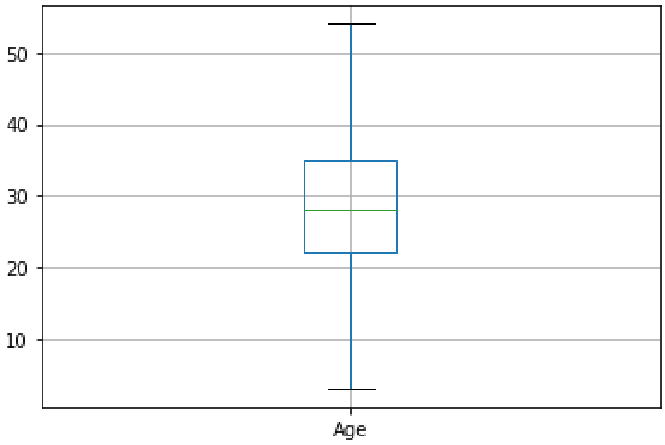
```
0      22.0
1      38.0
2      26.0
3      35.0
4      35.0
       ...
871    27.0
872    19.0
873    28.0
874    26.0
875    32.0
Name: Age, Length: 876, dtype: float64
```

In [51]:

```
dataset.boxplot(column=['Age'])
```

Out[51]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e702b8da48>
```



In [52]:

```
dataset['Fare'].describe()
```

Out[52]:

```
count    876.000000
mean      32.391794
std       50.020501
min        0.000000
25%        7.917700
50%       14.454200
75%       31.068750
max      512.329200
Name: Fare, dtype: float64
```

In [53]:

```
IQR_Fare = dataset['Fare'].quantile(0.75) - dataset['Fare'].quantile(0.25)
print(IQR_Fare)
```

```
23.15105
```

In [54]:

```
Upper_OutlierLimit_Fare = dataset['Fare'].quantile(0.75) + 1.5*IQR_Fare
Lower_OutlierLimit_Fare = dataset['Fare'].quantile(0.25) - 1.5*IQR_Fare
print(Upper_OutlierLimit_Fare)
print(Lower_OutlierLimit_Fare)
```

```
65.795325
-26.808875000000004
```

In [55]:

```
OutlierValues_Fare = dataset[(dataset['Fare']>=Upper_OutlierLimit_Fare) | (dataset[
'Fare']<=Lower_OutlierLimit_Fare)]
```

In [56]:

```
OutlierValues_Fare
```

Out[56]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embark
1	2	1	1	Cumings, Mrs. John Bradley\r(Florence Briggs T...	female	38.0	1	0	PC 17599	71.2833	C85	
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	263.0000	C23\rC25\rC27	
31	32	1	1	Spencer, Mrs. William Augustus\r(Marie Eugenie)	female	28.0	1	0	PC 17569	146.5208	B78	
34	35	0	1	Meyer, Mr. Edgar Joseph	male	28.0	1	0	PC 17604	82.1708	NaN	
60	62	1	1	Icard, Miss. Amelie	female	38.0	0	0	113572	80.0000	B28	
...	...	...	...	...	...	...	...	...	...	...	...	
831	847	0	3	Sage, Mr. Douglas Bullen	male	28.0	8	2	CA. 2343	69.5500	NaN	
834	850	1	1	Goldenberg, Mrs. Samuel L\r(Edwiga Grabowska)	female	28.0	1	0	17453	89.1042	C92	
841	857	1	1	Wick, Mrs. George Dennick\r(Mary Hitchcock)	female	45.0	1	1	36928	164.8667	NaN	
848	864	0	3	Sage, Miss. Dorothy Edith\r"Dolly"	female	28.0	8	2	CA. 2343	69.5500	NaN	
864	880	1	1	Potter, Mrs. Thomas Jr (Lily\rAlexenia Wilson)	female	54.0	0	1	11767	83.1583	C50	

115 rows × 12 columns

In [57]:

```
dataset.loc[dataset.Fare > 65.795325,'Fare'] =  dataset['Fare'].quantile(0.95)
dataset.loc[dataset.Fare < -26.808875000000004,'Fare'] =  dataset['Fare'].quantile(0.05)
```

In [58]:

```
dataset['Fare']
```

Out[58]:

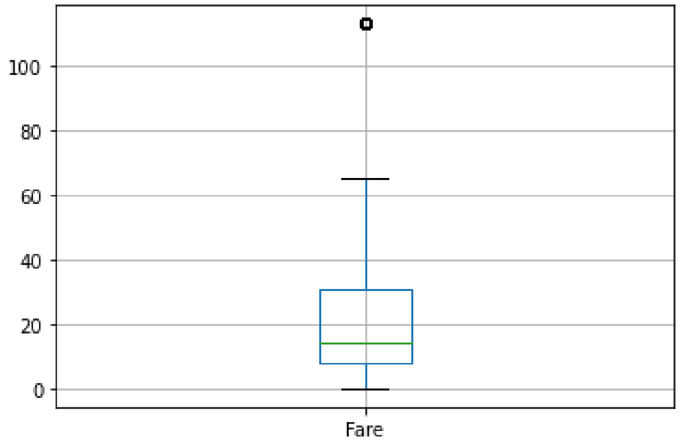
```
0        7.250
1      113.275
2        7.925
3       53.100
4        8.050
        ...   
871     13.000
872     30.000
873     23.450
874     30.000
875      7.750
Name: Fare, Length: 876, dtype: float64
```

In [59]:

```
dataset.boxplot(column=['Fare'])
```

Out[59]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e702beaac8>
```



7)Feature Engineering - Variable and Dummy Variable Creation

In [60]:

```
obj = dataset.dtypes == np.object
print(obj)
```

```
PassengerId    False
Survived       False
Pclass         False
Name            True
Sex             True
Age            False
SibSp          False
Parch          False
Ticket          True
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

In [61]:

```
dataset.columns[obj]
```

Out[61]:

```
Index(['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'], dtype='object')
```

In [62]:

```
dataset.drop(['PassengerId', 'Name', 'Ticket', 'Cabin'],axis = 1,inplace = True)
```

In [63]:

```
obj = dataset.dtypes == np.object
print(obj)
```

```
Survived     False
Pclass       False
Sex           True
Age          False
SibSp        False
Parch        False
Fare         False
Embarked      True
dtype: bool
```

In [64]:

```
dataset.columns[obj]
```

Out[64]:

```
Index(['Sex', 'Embarked'], dtype='object')
```

In [65]:

```
dummydf = pd.DataFrame()
for i in dataset.columns[obj]:
    dummy= pd.get_dummies(dataset[i], drop_first=True)
    dummydf = pd.concat([dummydf, dummy], axis=1)


print(dummydf)
```

```
     male  Q  S
0       1  0  1
1       0  0  0
2       0  0  1
3       0  0  1
4       1  0  1
..    ... .. ..
871     1  0  1
872     0  0  1
873     0  0  1
874     1  0  0
875     1  1  0

[876 rows x 3 columns]
```

In [66]:

```
final_dataset = pd.concat([dataset,dummydf], axis=1)
final_dataset
```

Out[66]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	male	Q	S
0	0	3	male	22.0	1	0	7.250	S	1	0	1
1	1	1	female	38.0	1	0	113.275	C	0	0	0
2	1	3	female	26.0	0	0	7.925	S	0	0	1
3	1	1	female	35.0	1	0	53.100	S	0	0	1
4	0	3	male	35.0	0	0	8.050	S	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...
871	0	2	male	27.0	0	0	13.000	S	1	0	1
872	1	1	female	19.0	0	0	30.000	S	0	0	1
873	0	3	female	28.0	1	2	23.450	S	0	0	1

874	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	male	Q	S
875	0	3	male	32.0	0	0	7.750	Q	1	1	0

876 rows × 11 columns

In [67]:

```
final_dataset = final_dataset[['Pclass','Age','SibSp','Parch','Fare','male','Q','S','Surv
ived']]
```

In [68]:

```
final_dataset
```

Out[68]:

	Pclass	Age	SibSp	Parch	Fare	male	Q	S	Survived
0	3	22.0	1	0	7.250	1	0	1	0
1	1	38.0	1	0	113.275	0	0	0	1
2	3	26.0	0	0	7.925	0	0	1	1
3	1	35.0	1	0	53.100	0	0	1	1
4	3	35.0	0	0	8.050	1	0	1	0
...	...	...	...	...	...	...	...	...	...
871	2	27.0	0	0	13.000	1	0	1	0
872	1	19.0	0	0	30.000	0	0	1	1
873	3	28.0	1	2	23.450	0	0	1	0
874	1	26.0	0	0	30.000	1	0	0	1
875	3	32.0	0	0	7.750	1	1	0	0

876 rows × 9 columns

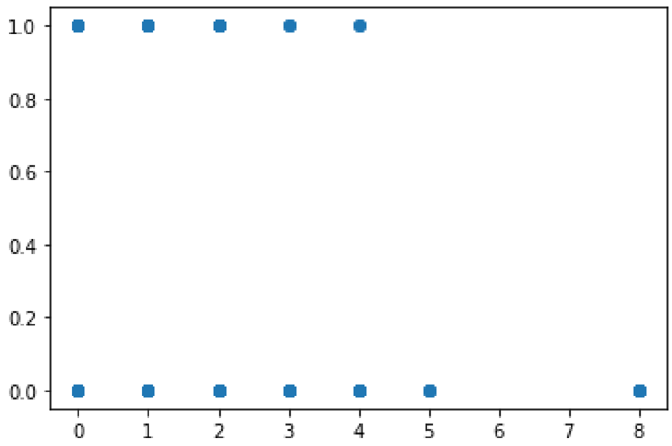
7)Feature Engineering - Variable Transformation

In [69]:

```
plt.scatter(final_dataset['SibSp'],final_dataset['Survived'])
```

Out[69]:

```
<matplotlib.collections.PathCollection at 0x1e702c6b308>
```

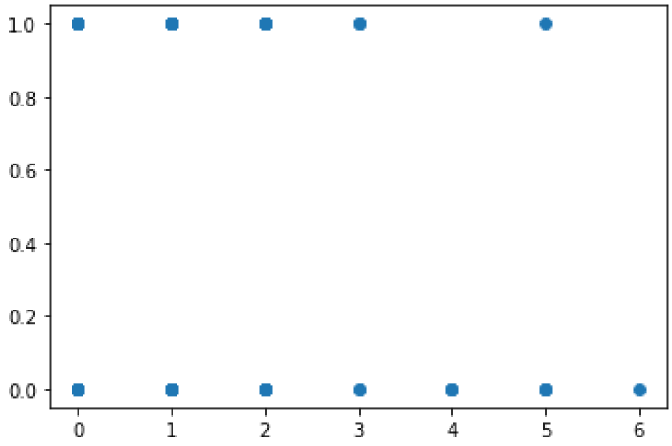


In [70]:

```
plt.scatter(final_dataset['Parch'],final_dataset['Survived'])
```

Out[70]:

```
<matplotlib.collections.PathCollection at 0x1e702ce8a48>
```



In [71]:

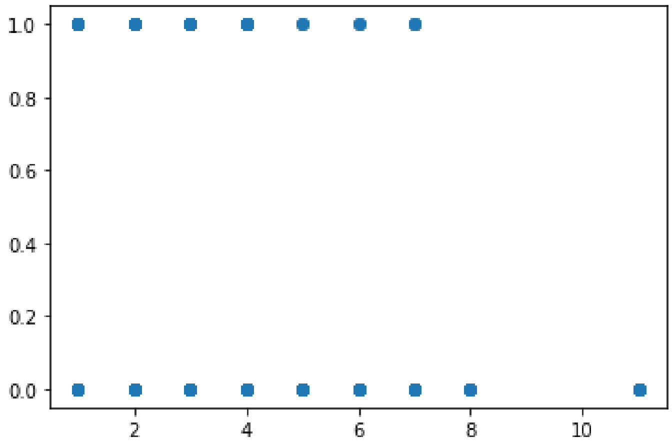
```
New_Feature = final_dataset.SibSp + final_dataset.Parch + 1
```

In [72]:

```
plt.scatter(New_Feature,final_dataset['Survived'])
```

Out[72]:

```
<matplotlib.collections.PathCollection at 0x1e702d4dc48>
```



1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10