In [184]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as pairplot
import warnings
warnings.filterwarnings('ignore')
```

In [185]:

```
df=pd.read_csv('full.csv')
df
```

Out[185]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	
...	...	...	...	...	...	...	...	...	...	
1304	1305	NaN	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	
1305	1306	NaN	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	1
1306	1307	NaN	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	
1307	1308	NaN	3	Ware, Mr. Frederick	male	NaN	0	0	359309	
1308	1309	NaN	3	Peter, Master. Michael J	male	NaN	1	1	2668	

1309 rows × 21 columns

In [186]:

```
df.head()
```

Out[186]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500

5 rows × 21 columns

In [187]:

```
df.columns
```

Out[187]:

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked', 'WikiId', 'Name_wik
i',
       'Age_wiki', 'Hometown', 'Boarded', 'Destination', 'Lifeboat', 'Bod
y',
       'Class'],
      dtype='object')
```

In [188]:

```
df.columns.values
```

Out[188]:

```
array(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked', 'WikiId',
       'Name_wiki', 'Age_wiki', 'Hometown', 'Boarded', 'Destination',
       'Lifeboat', 'Body', 'Class'], dtype=object)
```

In [189]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 21 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  1309 non-null   int64  
 1   Survived     891 non-null    float64
 2   Pclass       1309 non-null   int64  
 3   Name         1309 non-null   object 
 4   Sex          1309 non-null   object 
 5   Age          1046 non-null   float64
 6   SibSp        1309 non-null   int64  
 7   Parch        1309 non-null   int64  
 8   Ticket       1309 non-null   object 
 9   Fare         1308 non-null   float64
 10  Cabin        295 non-null    object 
 11  Embarked     1307 non-null   object 
 12  WikiId       1304 non-null   float64
 13  Name_wiki    1304 non-null   object 
 14  Age_wiki     1302 non-null   float64
 15  Hometown     1304 non-null   object 
 16  Boarded      1304 non-null   object 
 17  Destination  1304 non-null   object 
 18  Lifeboat     502 non-null    object 
 19  Body         130 non-null    object 
 20  Class        1304 non-null   float64
dtypes: float64(6), int64(4), object(11)
memory usage: 214.9+ KB
```

In [190]:

```
df.isnull().sum()
```

Out[190]:

```
PassengerId       0
Survived        418
Pclass            0
Name              0
Sex               0
Age             263
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin          1014
Embarked          2
WikiId            5
Name_wiki         5
Age_wiki          7
Hometown          5
Boarded           5
Destination       5
Lifeboat        807
Body           1179
Class             5
dtype: int64
```

In [191]:

```
df.drop(columns=['Cabin'],inplace=True)
```

In [192]:

```
df['Age'].fillna(df['Age'].mean(),inplace=True)
```

In [193]:

```
df.drop(columns=['Body'],inplace=True)
```

In [194]:

```
df.drop(columns=['Lifeboat'],inplace=True)
```

In [195]:

```
df['Embarked'].fillna('S',inplace=True)
```

In [196]:

```
df.describe()
```

Out[196]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	
count	1309.000000	891.000000	1309.000000	1309.000000	1309.000000	1309.000000	1308.0
mean	655.000000	0.383838	2.294882	29.881138	0.498854	0.385027	33.2
std	378.020061	0.486592	0.837836	12.883193	1.041658	0.865560	51.7
min	1.000000	0.000000	1.000000	0.170000	0.000000	0.000000	0.0
25%	328.000000	0.000000	2.000000	22.000000	0.000000	0.000000	7.8
50%	655.000000	0.000000	3.000000	29.881138	0.000000	0.000000	14.4
75%	982.000000	1.000000	3.000000	35.000000	1.000000	0.000000	31.2
max	1309.000000	1.000000	3.000000	80.000000	8.000000	9.000000	512.3

In [197]:

```
df['WikiId'].fillna(df['WikiId'].mean(),inplace=True)
```

In [198]:

```
df['Age_wiki'].fillna(df['Age_wiki'].median(),inplace=True)
```

In [199]:

```
df['Class'].fillna(df['Class'].mean(),inplace=True)
```

In [200]:

```
df['Fare'].fillna(df['Fare'].mean(),inplace=True)
```

In [201]:

```
df.isnull().sum()
```

Out[201]:

```
PassengerId      0
Survived       418
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Embarked         0
WikiId           0
Name_wiki        5
Age_wiki         0
Hometown         5
Boarded          5
Destination      5
Class            0
dtype: int64
```

In [202]:

```
df.drop(columns=['Boarded'],inplace=True)
```

In [203]:

```
df.drop(columns=['Hometown'],inplace=True)
```

In [204]:

```
df.isnull().sum()
```

Out[204]:

```
PassengerId      0
Survived       418
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Embarked         0
WikiId           0
Name_wiki        5
Age_wiki         0
Destination      5
Class            0
dtype: int64
```

In [205]:

```
df['Parch'].value_counts()
```

Out[205]:

```
0    1002
1     170
2     113
3       8
5       6
4       6
9       2
6       2
Name: Parch, dtype: int64
```

In [206]:

```
df['Survived']=df['Survived'].astype('category')
df['Sex']=df['Sex'].astype('category')
df['Pclass']=df['Pclass'].astype('category')
df['Age']=df['Age'].astype('category')
df['Embarked']=df['Embarked'].astype('category')
```

In [207]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 16 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  1309 non-null   int64
 1   Survived     891 non-null    category
 2   Pclass       1309 non-null   category
 3   Name         1309 non-null   object
 4   Sex          1309 non-null   category
 5   Age          1309 non-null   category
 6   SibSp        1309 non-null   int64
 7   Parch        1309 non-null   int64
 8   Ticket       1309 non-null   object
 9   Fare         1309 non-null   float64
 10  Embarked     1309 non-null   category
 11  WikiId       1309 non-null   float64
 12  Name_wiki    1304 non-null   object
 13  Age_wiki     1309 non-null   float64
 14  Destination  1304 non-null   object
 15  Class        1309 non-null   float64
dtypes: category(5), float64(4), int64(3), object(4)
memory usage: 122.7+ KB
```

In [208]:

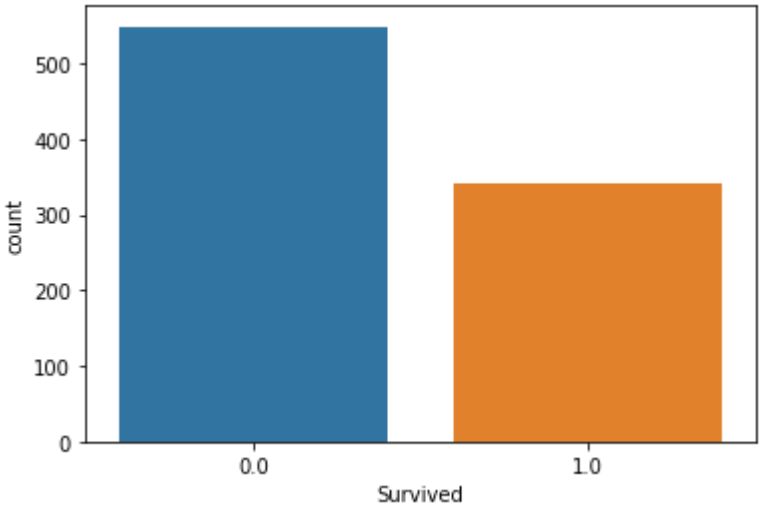
```
df.describe()
```

Out[208]:

	PassengerId	SibSp	Parch	Fare	WikiId	Age_wiki	
count	1309.000000	1309.000000	1309.000000	1309.000000	1309.000000	1309.000000	1309.
mean	655.000000	0.498854	0.385027	33.295479	658.534509	29.408258	2.
std	378.020061	1.041658	0.865560	51.738879	379.649656	13.722477	0.
min	1.000000	0.000000	0.000000	0.000000	1.000000	0.170000	1.
25%	328.000000	0.000000	0.000000	7.895800	328.000000	21.000000	2.
50%	655.000000	0.000000	0.000000	14.454200	659.000000	28.000000	3.
75%	982.000000	1.000000	0.000000	31.275000	986.000000	37.000000	3.
max	1309.000000	8.000000	9.000000	512.329200	1314.000000	74.000000	3.

In [209]:

```
sns.countplot(df['Survived'])
death_percent=round((df['Survived'].value_counts().values[0]/1309*100))
```



In [210]:

```
print('out of 1309 {} people died in the accident'.format(death_percent))
```

```
out of 1309 42.0 people died in the accident
```

In [211]:

```
print(df['Pclass'].value_counts()/1309*100)
```

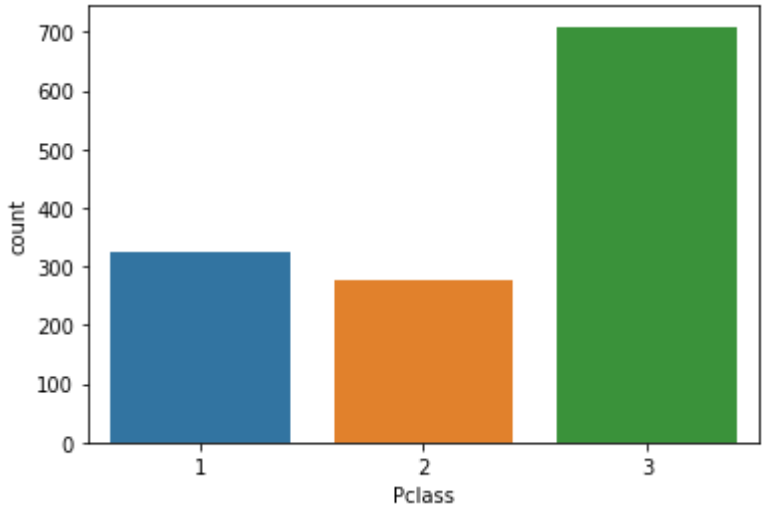
```
3    54.163484
1    24.675325
2    21.161192
Name: Pclass, dtype: float64
```

In [212]:

```
sns.countplot(df['Pclass'])
```

Out[212]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x284611aee08>
```



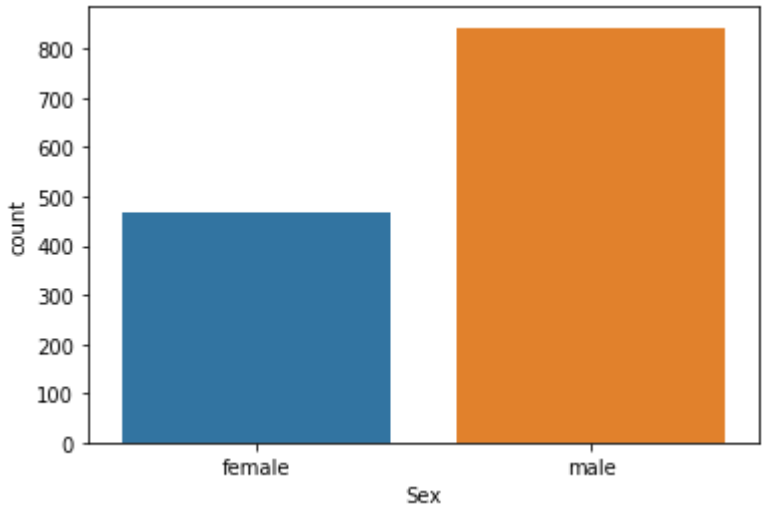
In [213]:

```
print(df['Sex'].value_counts()/1309*100)
sns.countplot(df['Sex'])
```

```
male      64.400306
female    35.599694
Name: Sex, dtype: float64
```

Out[213]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x2846121af88>
```



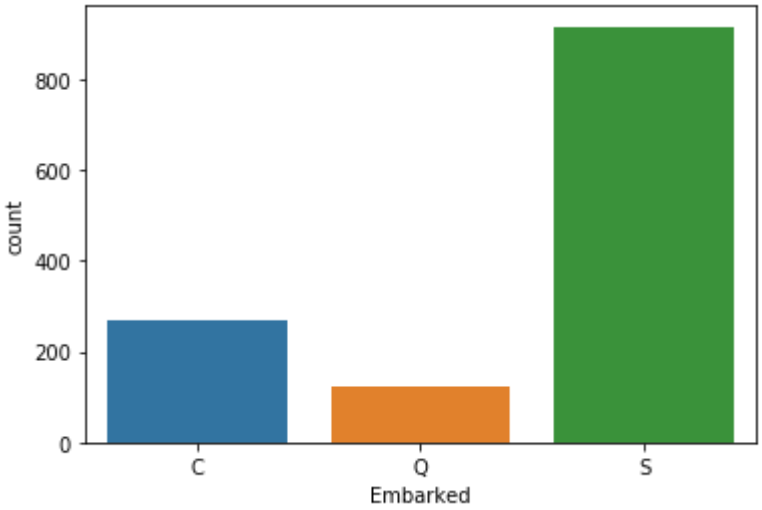
In [214]:

```
print(df['Embarked'].value_counts()/1309*100)
sns.countplot(df['Embarked'])
```

```
S    69.977082
C    20.626432
Q     9.396486
Name: Embarked, dtype: float64
```

Out[214]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x28461279388>
```

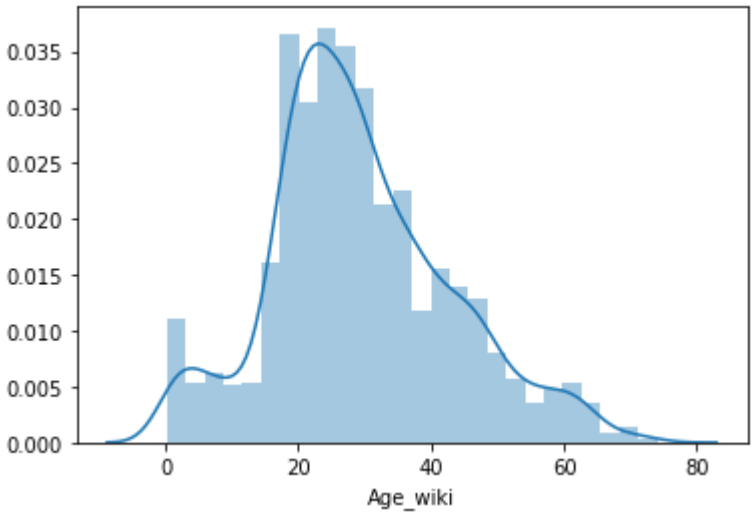


In [215]:

```
sns.distplot(df['Age_wiki'])
```

Out[215]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x2846120d308>
```

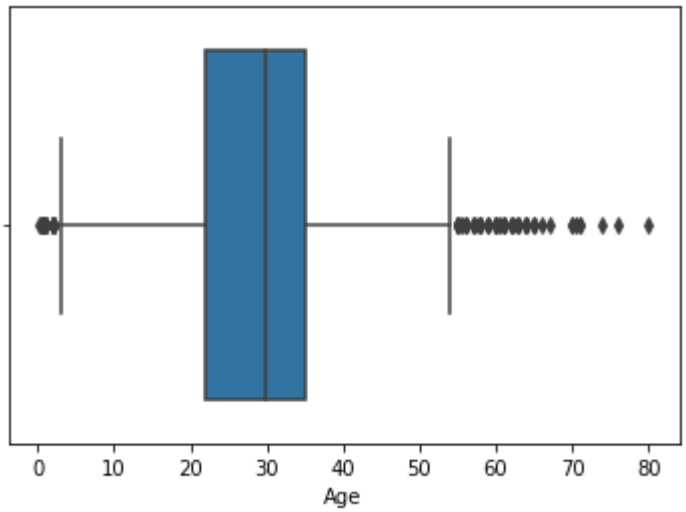


In [216]:

```
sns.boxplot(df['Age'])
```

Out[216]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x284611aed88>
```

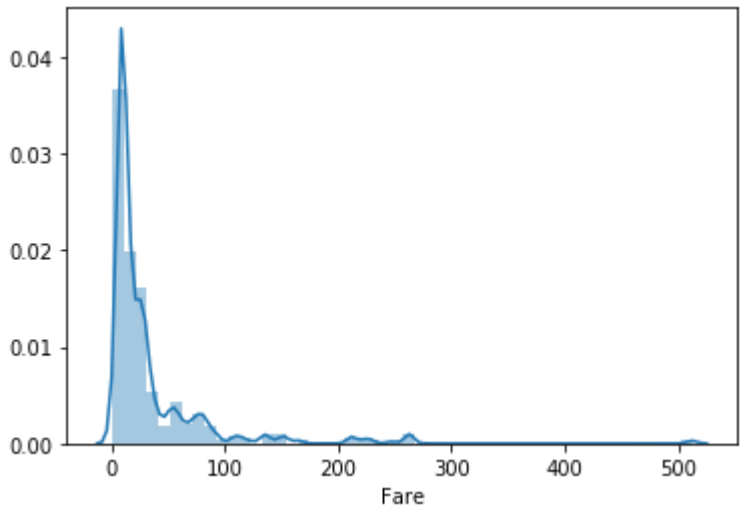


In [217]:

```
sns.distplot(df['Fare'])
```

Out[217]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x284613e5a88>
```



In [218]:

```
print(df['Fare'].skew())
```

```
4.369374593951007
```

In [219]:

```
print(df['Fare'].kurt())
```

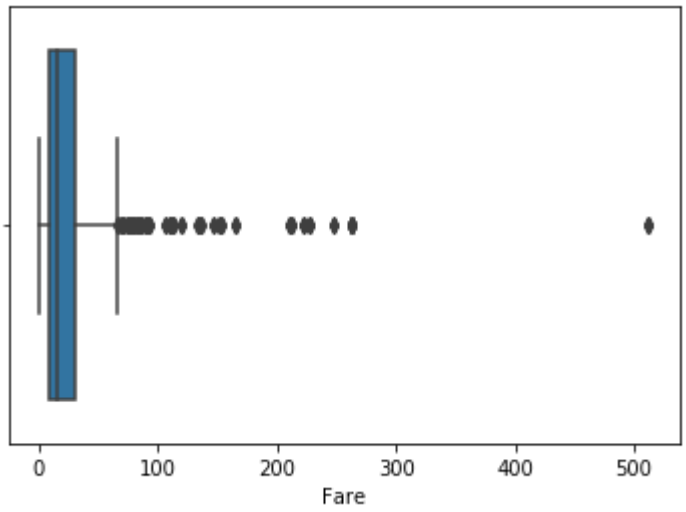
```
27.05086615808882
```

In [220]:

```
sns.boxplot(df['Fare'])
```

Out[220]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x284614e1cc8>
```

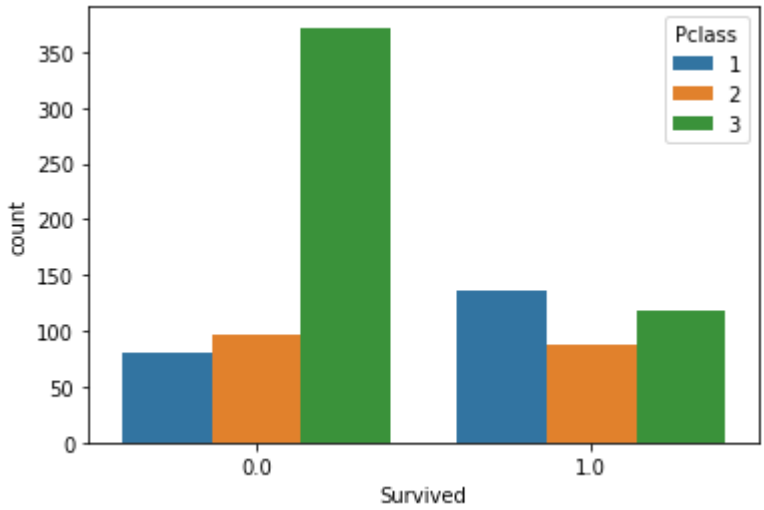


In [221]:

```
sns.countplot(df['Survived'],hue=df['Pclass'])
pd.crosstab(df['Pclass'],df['Survived']).apply(lambda r:round((r/r.sum())*100,1),axis=1
)
```

Out[221]:

Survived	0.0	1.0
Pclass		
1	37.0	63.0
2	52.7	47.3
3	75.8	24.2

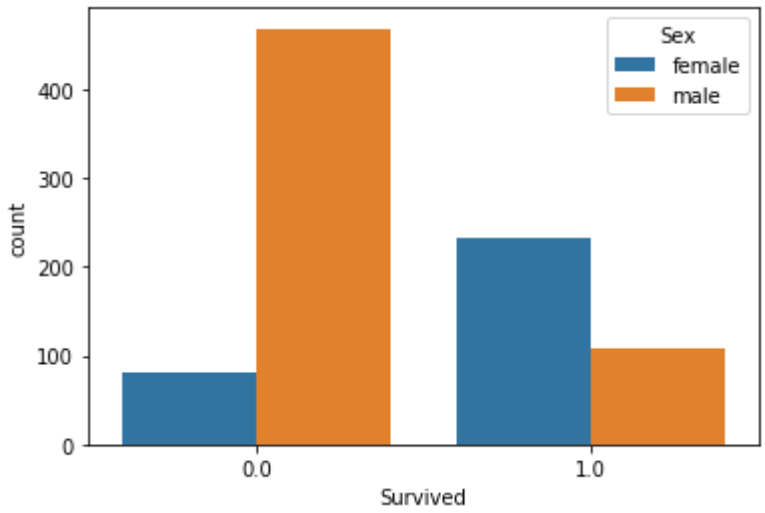


In [222]:

```
sns.countplot(df['Survived'],hue=df['Sex'])
pd.crosstab(df['Sex'],df['Survived']).apply(lambda r:round((r/r.sum())*100,1),axis=1)
```

Out[222]:

Survived	0.0	1.0
Sex		
female	25.8	74.2
male	81.1	18.9

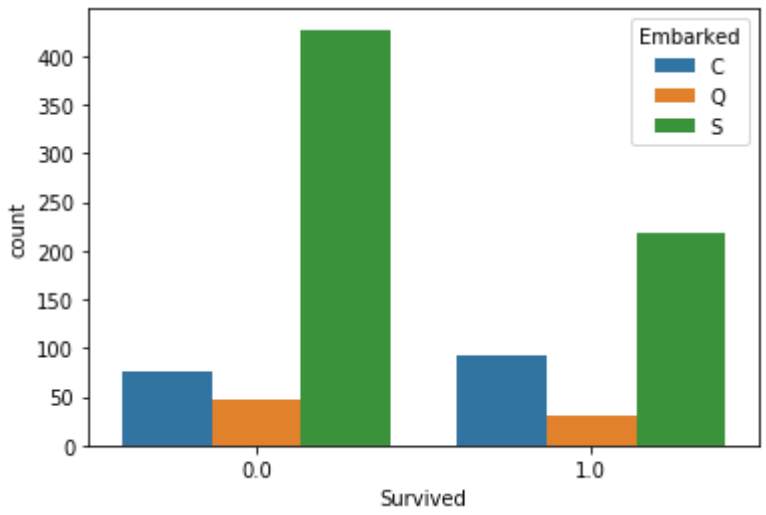


In [223]:

```
sns.countplot(df['Survived'],hue=df['Embarked'])
pd.crosstab(df['Embarked'],df['Survived']).apply(lambda r:round((r/r.sum())*100,1),axis
=1)
```

Out[223]:

Survived	0.0	1.0
Embarked		
C	44.6	55.4
Q	61.0	39.0
S	66.1	33.9

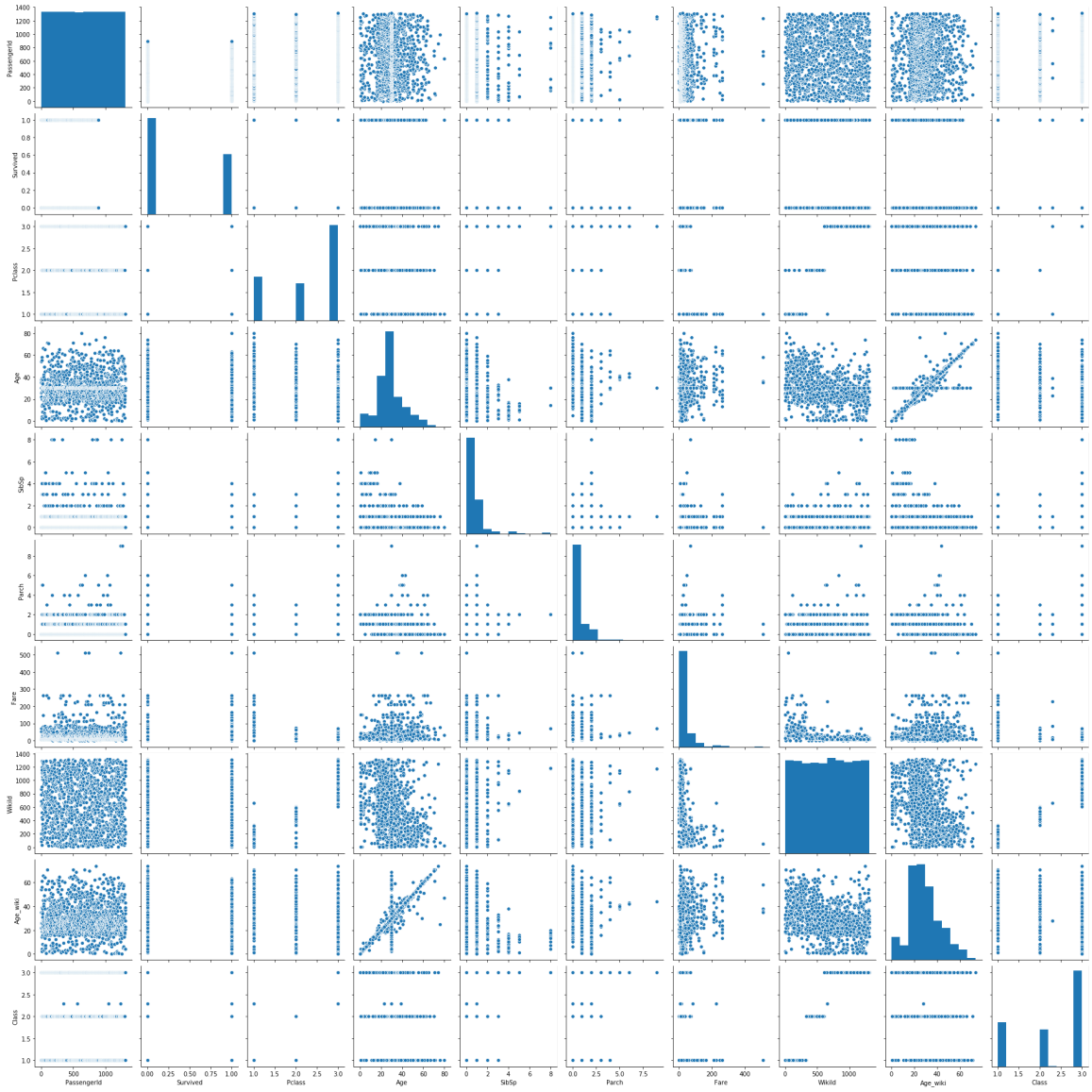


In [224]:

```
sns.pairplot(df)
```

Out[224]:

```
<seaborn.axisgrid.PairGrid at 0x284615b0448>
```

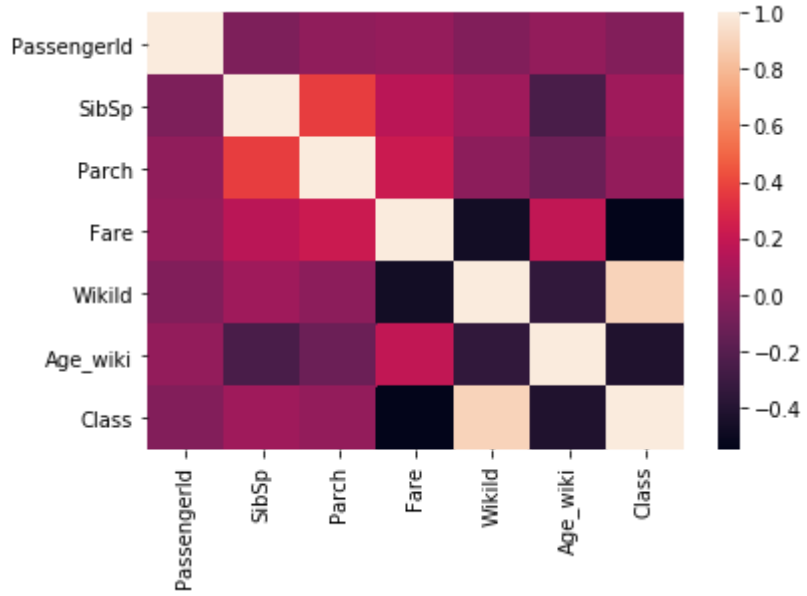


In [225]:

```
sns.heatmap(df.corr())
```

Out[225]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x2846494fac8>
```



In [226]:

```
df['family_size']=df['Parch']+df['SibSp']
```

In [227]:

```
df.sample(5)
```

Out[227]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
64	65	0.0	1	Stewart, Mr. Albert A	male	29.881138	0	0	PC 17605	27
1170	1171	NaN	2	Oxenham, Mr. Percy Thomas	male	22.000000	0	0	W./C. 14260	10
1094	1095	NaN	2	Quick, Miss. Winifred Vera	female	8.000000	1	1	26360	26
1144	1145	NaN	3	Salander, Mr. Karl Johan	male	24.000000	0	0	7266	9
482	483	0.0	3	Rouse, Mr. Richard Henry	male	50.000000	0	0	A/5 3594	8

In [228]:

```
def family_type(number):
    if number==0:
        return 'Alone'
    elif number>0 and number<=4:
        return 'Medium'
    else:
        return "Large"
```

In [229]:

```
df['family_type']=df['family_size'].apply(family_type)
```

In [230]:

```
df.sample(5)
```

Out[230]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1033	1034	NaN	1	Ryerson, Mr. Arthur Larned	male	61.000000	1	3	PC 17608
440	441	1.0	2	Hart, Mrs. Benjamin (Esther Ada Bloomfield)	female	45.000000	1	1	F.C.C. 13529
254	255	0.0	3	Rosblom, Mrs. Viktor (Helena Wilhelmina)	female	41.000000	0	2	370129
367	368	1.0	3	Moussa, Mrs. (Mantoura Boulos)	female	29.881138	0	0	2626
980	981	NaN	2	Wells, Master. Ralph Lester	male	2.000000	1	1	29103

In [231]:

```
df.drop(columns=['SibSp','Parch','family_size'],inplace=True)
```

In [232]:

```
df.sample(10)
```

Out[232]:

	PassengerId	Survived	Pclass	Name	Sex	Age	Ticket	Fare	Em
791	792	0.0	2	Gaskell, Mr. Alfred	male	16.000000	239865	26.0000	
657	658	0.0	3	Bourke, Mrs. John (Catherine)	female	32.000000	364849	15.5000	
975	976	NaN	2	Lamb, Mr. John Joseph	male	29.881138	240261	10.7083	
888	889	0.0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.881138	W./C. 6607	23.4500	
555	556	0.0	1	Wright, Mr. George	male	62.000000	113807	26.5500	
165	166	1.0	3	Goldsmith, Master. Frank John William "Frankie"	male	9.000000	363291	20.5250	
930	931	NaN	3	Hee, Mr. Ling	male	29.881138	1601	56.4958	
776	777	0.0	3	Tobin, Mr. Roger	male	29.881138	383121	7.7500	
1252	1253	NaN	2	Mallet, Mrs. Albert (Antoinette Magnin)	female	24.000000	S.C./PARIS 2079	37.0042	
664	665	1.0	3	Lindqvist, Mr. Eino William	male	20.000000	STON/O 2. 3101285	7.9250	

In [233]:

```
pd.crosstab(df['family_type'],df['Survived']).apply(lambda r:round((r/r.sum())*100,1),a
xis=1)
```

Out[233]:

Survived	0.0	1.0
family_type		
Alone	69.6	30.4
Large	85.1	14.9
Medium	44.0	56.0

In [235]:

```
q1=np.percentile(df['Fare'],0.25)
q3=np.percentile(df['Fare'],0.75)
```

In [236]:

```
q1
```

Out[236]:

```
0.0
```

In [237]:

```
q3
```

Out[237]:

```
0.0
```

In [243]:

```
plt.figure(figsize=(18,10))
sns.heatmap(df.corr(),cmap='summer')
```

Out[243]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x28467634308>
```

In [ ]: