NAME - Debashis Saha

INTERN OF THE INTERN ACADEMY

Exploratory Data Analysis on Titanic Dataset

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df = pd.read_csv("/content/Titanic.csv")
df.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [3]:
```
df.isnull().sum()
```

Out[3]:
```
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 [4]:
```
print("Shape of the training set", df.shape)
```

```
Shape of the training set (891, 12)
```

In [5]:
```
df.describe()
```

Out[5]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [6]:
```
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
```

In [7]:
```
df.columns
```

Out[7]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [8]:
```
df.drop(['Cabin','PassengerId','Name', 'Ticket'],axis=1,inplace=True)
df.head()
```

Out[8]:

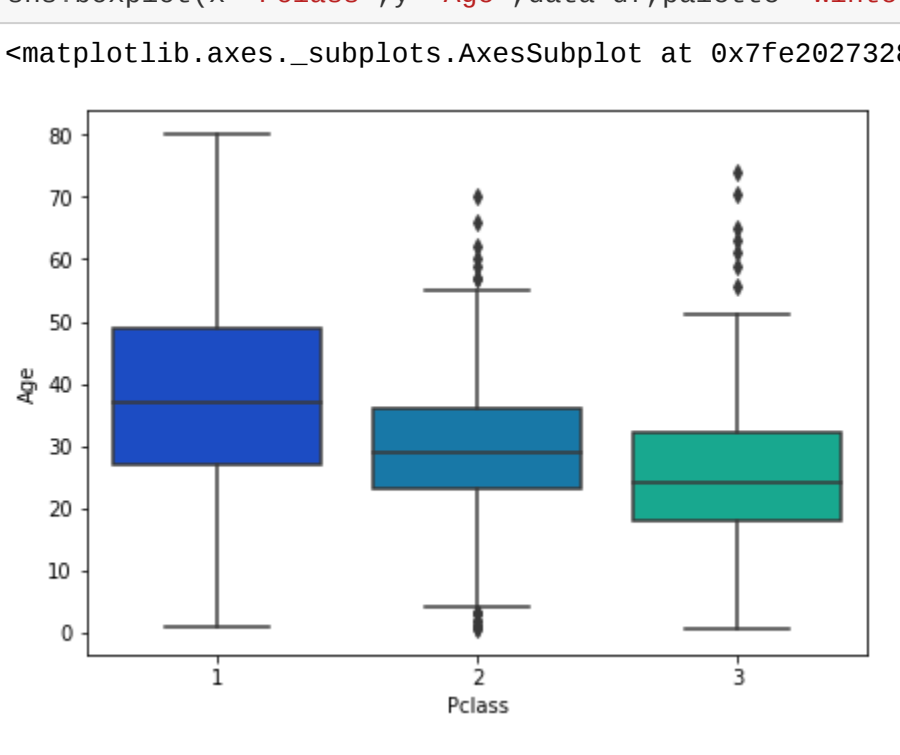
	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
Survived      0
Pclass        0
Sex           0
Age         177
SibSp         0
Parch         0
Fare          0
Embarked      2
dtype: int64
```

In [10]:
```
plt.figure(figsize=(7, 5))
sns.boxplot(x='Pclass',y='Age',data=df,palette='winter')
```

Out[10]: `<matplotlib.axes._subplots.AxesSubplot at 0x7fe202732810>`



In [11]:
```
def impute_age(cols):
    Age = cols[0]
    Pclass = cols[1]

    if pd.isnull(Age):

        if Pclass == 1:
            return 37

        elif Pclass == 2:
            return 29

        else:
            return 24

    else:
        return Age
```

In [12]:
```
df['Age'] = df[['Age','Pclass']].apply(impute_age,axis=1)
df.head()
```

Out[12]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S

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

Out[13]:
```
Survived    0
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        0
Embarked    2
dtype: int64
```

In [14]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df.Sex = le.fit_transform(df.Sex)
df.head()
```

Out[14]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1	22.0	1	0	7.2500	S
1	1	1	0	38.0	1	0	71.2833	C
2	1	3	0	26.0	0	0	7.9250	S
3	1	1	0	35.0	1	0	53.1000	S
4	0	3	1	35.0	0	0	8.0500	S

In [15]:
```
df[["Sex","Age"]].groupby("Sex").mean()
```

Out[15]:

Sex	Age
0	27.659236
1	29.832184

In [16]:
```
df[["Age","Pclass"]].groupby("Pclass").mean()
```

Out[16]:

Pclass	Age
1	38.062130
2	29.825163
3	24.824684

In [17]:
```
df.agg({
                "Age": ["min", "max", "median", "skew"],
                "Fare": ["min", "max", "median", "mean"],
    } )
```

Out[17]:

	Age	Fare
max	80.000000	512.329200
mean	NaN	32.204208
median	26.000000	14.454200
min	0.420000	0.000000
skew	0.548256	NaN

In [18]:
```
df.groupby(["Sex", "Pclass"])["Fare"].mean()
```

Out[18]:
```
Sex  Pclass
0    1         106.125798
     2          21.970121
     3          16.118810
1    1          67.226127
     2          19.741782
     3          12.661633
Name: Fare, dtype: float64
```

In [19]:
```
df.groupby(["Pclass"])["Sex"].value_counts()
```

Out[19]:
```
Pclass  Sex
1       1      122
        0       94
2       1      108
        0       76
3       1      347
        0      144
Name: Sex, dtype: int64
```

In [20]:
```
df = df.ffill(axis = 0)
df.head(4)
```

Out[20]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1	22.0	1	0	7.2500	S
1	1	1	0	38.0	1	0	71.2833	C
2	1	3	0	26.0	0	0	7.9250	S
3	1	1	0	35.0	1	0	53.1000	S

In [21]:
```
df.Embarked = le.fit_transform(df.Embarked)
df.head()
```

Out[21]:

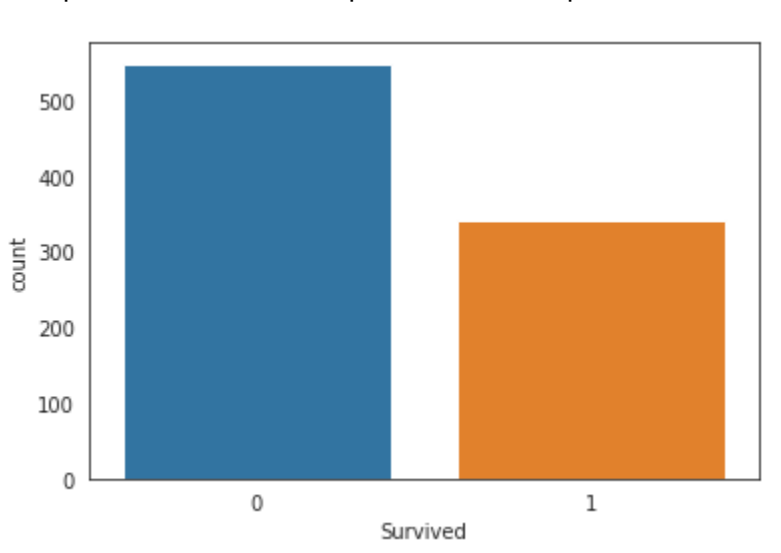
	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1	22.0	1	0	7.2500	2
1	1	1	0	38.0	1	0	71.2833	0
2	1	3	0	26.0	0	0	7.9250	2
3	1	1	0	35.0	1	0	53.1000	2
4	0	3	1	35.0	0	0	8.0500	2

In [22]:
```
df['Embarked'].unique()
```

Out[22]: `array([2, 0, 1])`

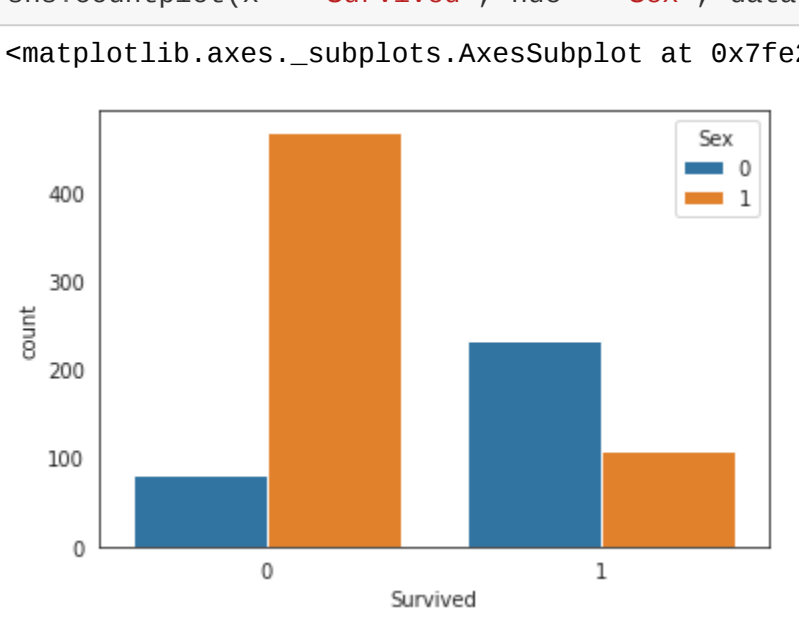
In [23]:
```
sns.set_style("white")
sns.countplot(x = "Survived", data = df)
```

Out[23]: `<matplotlib.axes._subplots.AxesSubplot at 0x7fe200651290>`



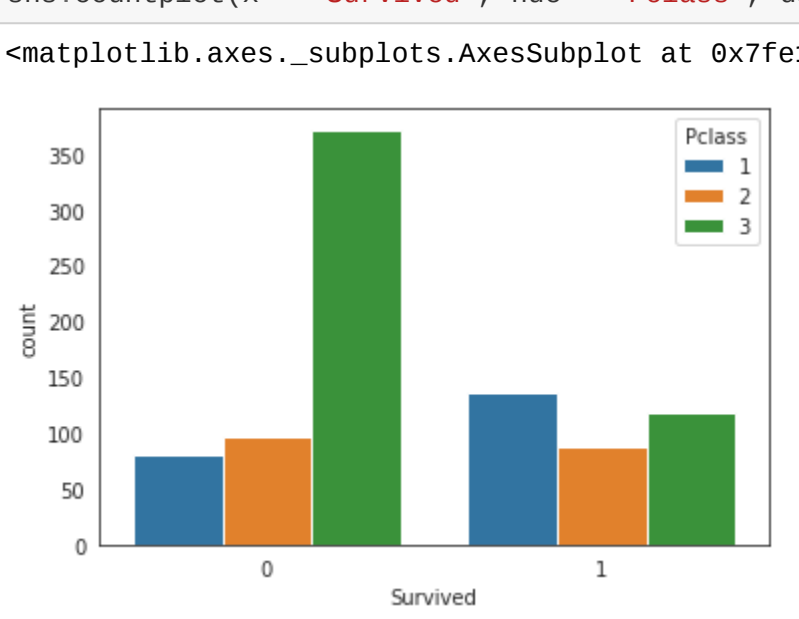
In [24]:
```
sns.set_style("white")
sns.countplot(x = "Survived", hue = "Sex", data = df)
```

Out[24]: `<matplotlib.axes._subplots.AxesSubplot at 0x7fe200396150>`

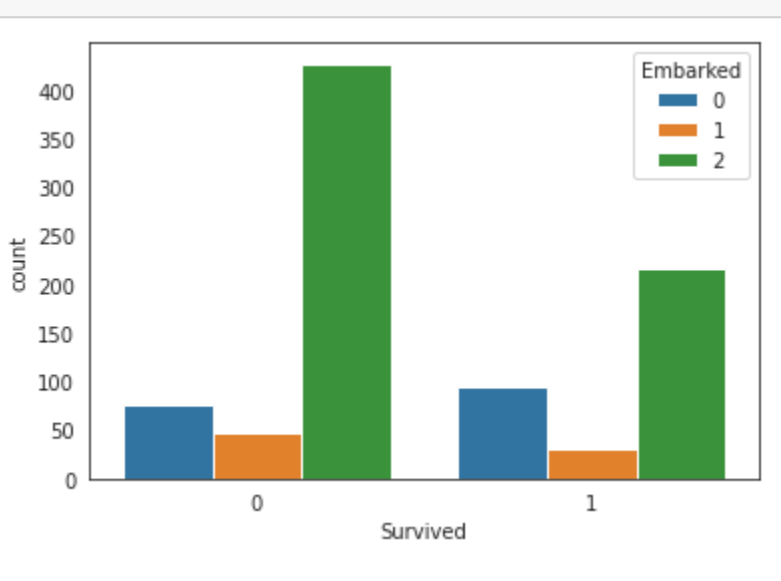


In [25]:
```
sns.countplot(x = "Survived", hue = "Pclass", data = df)
```

Out[25]: `<matplotlib.axes._subplots.AxesSubplot at 0x7fe1ffb07090>`



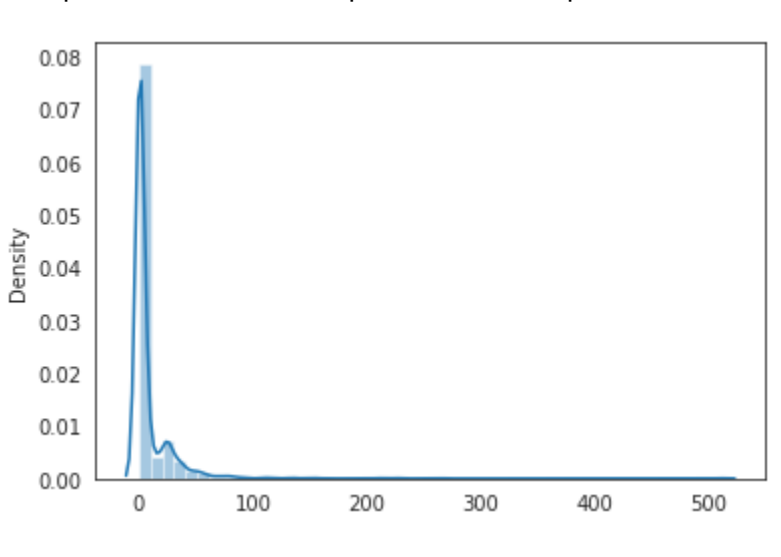
In [26]:
```
sns.countplot(x = "Survived", hue = "Embarked", data = df);
```



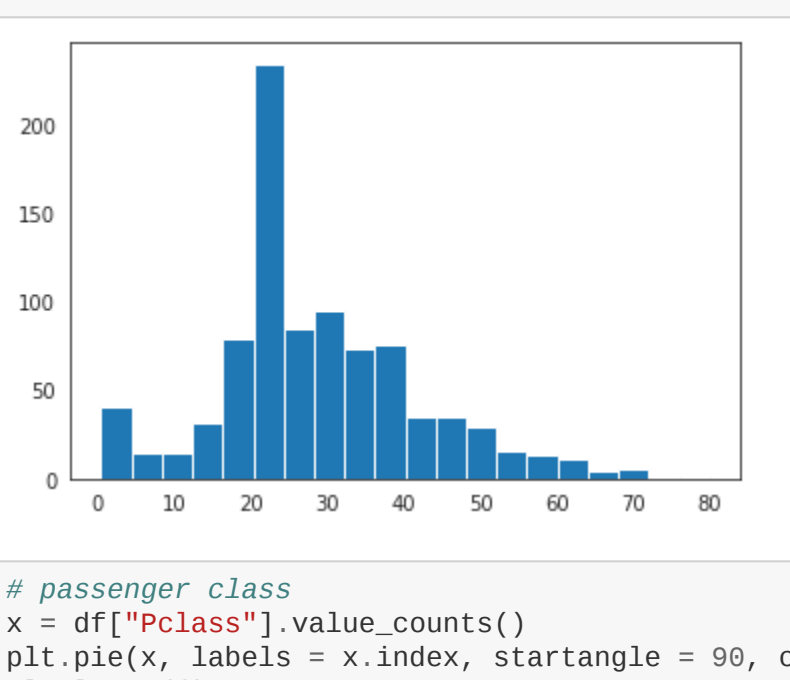
In [27]:
```
sns.distplot(df)
```

```
/usr/local/lib/python3.7/dist-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[27]: `<matplotlib.axes._subplots.AxesSubplot at 0x7fe1ff9fb190>`

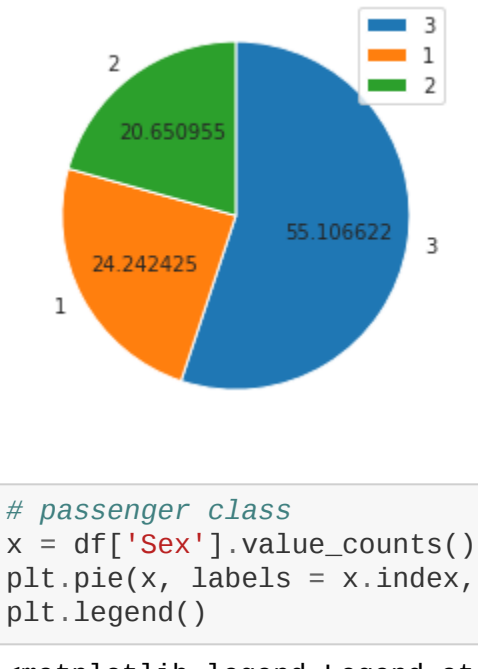


In [28]:
```
plt.hist(x = df["Age"],bins = 20);
```



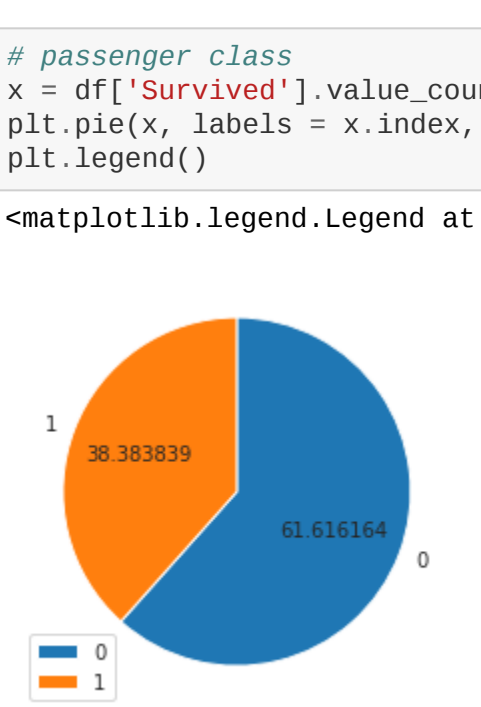
In [29]:
```
# passenger class
x = df["Pclass"].value_counts()
plt.pie(x, labels = x.index, startangle = 90, counterclock = False, autopct = "%2f");
plt.legend()
```

Out[29]: `<matplotlib.legend.Legend at 0x7fe1fd84f9d0>`



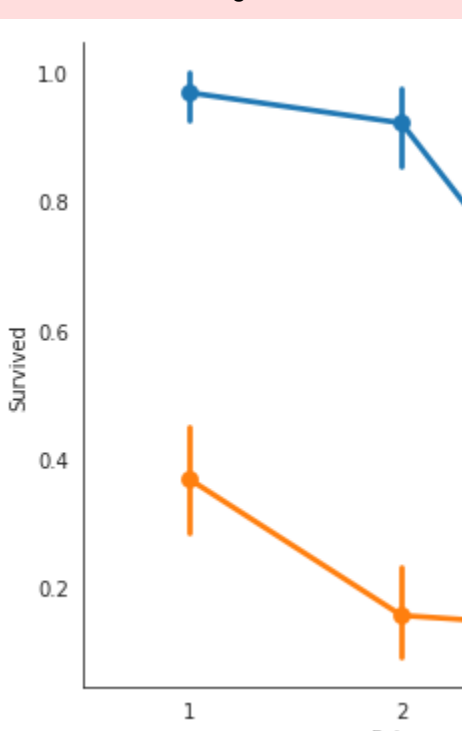
In [30]:
```
# passenger class
x = df['Sex'].value_counts()
plt.pie(x, labels = x.index, startangle = 90, counterclock = False, autopct = "%2f");
plt.legend()
```

Out[30]: `<matplotlib.legend.Legend at 0x7fe1fd81c610>`



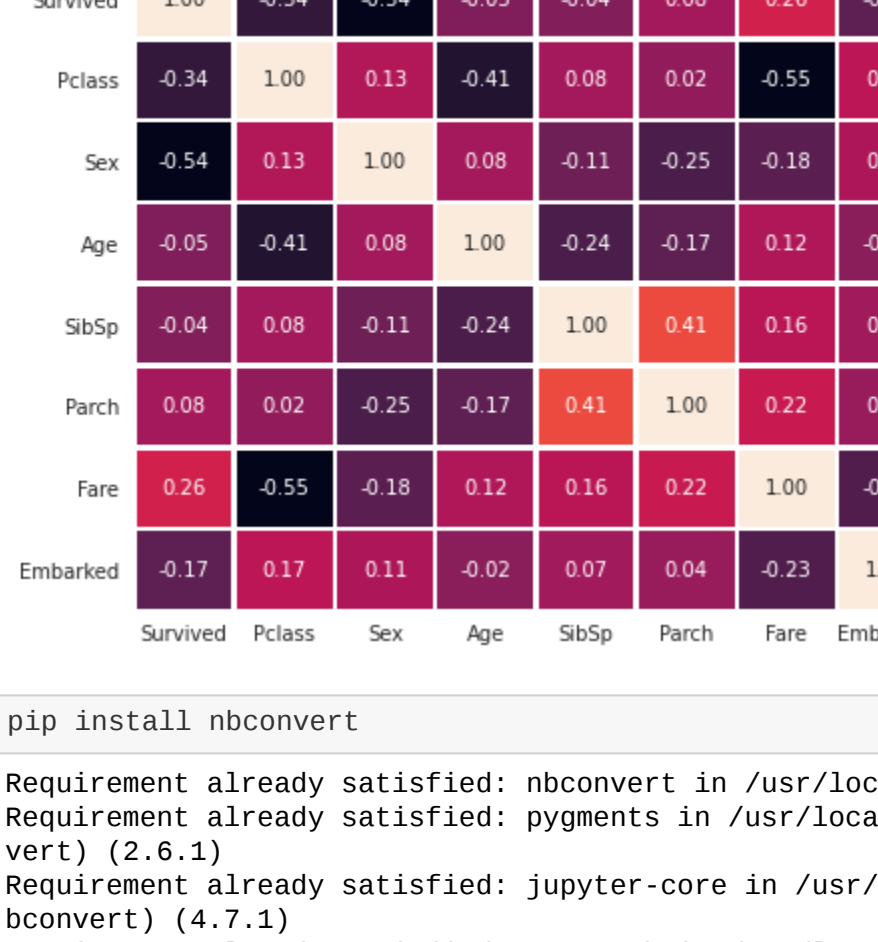
In [31]:
```
# passenger class
x = df['Survived'].value_counts()
plt.pie(x, labels = x.index, startangle = 90, counterclock = False, autopct = "%2f");
plt.legend()
```

Out[31]: `<matplotlib.legend.Legend at 0x7fe1fd779e10>`



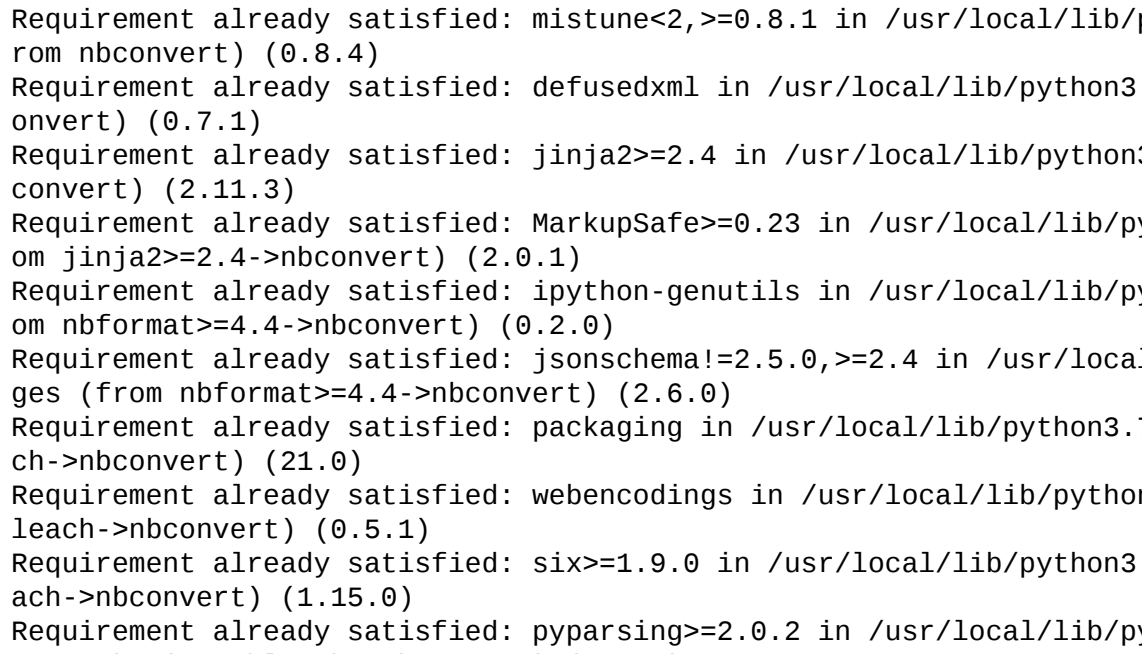
In [32]:
```
sns.catplot('Pclass','Survived',hue='Sex', kind='point', data=df);
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```



In [33]:
```
corr = df.corr()

f,ax = plt.subplots(figsize=(9,6))
sns.heatmap(corr, annot = True, linewidths=1.5 , fmt = '.2f',ax=ax)
plt.show()
```



In [34]:
```
pip install nbconvert
```

```
Requirement already satisfied: nbconvert in /usr/local/lib/python3.7/dist-packages (5.6.1)
Requirement already satisfied: pygments in /usr/local/lib/python3.7/dist-packages (from nbconvert) (2.6.1)
Requirement already satisfied: jupyter-core in /usr/local/lib/python3.7/dist-packages (from nbconvert) (4.7.1)
Requirement already satisfied: testpath in /usr/local/lib/python3.7/dist-packages (from nbconvert) (0.5.0)
Requirement already satisfied: nbformat>=4.4 in /usr/local/lib/python3.7/dist-packages (from nbconvert) (5.1.3)
Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.7/dist-packages (from nbconvert) (5.0.5)
Requirement already satisfied: entrypoints>=0.2.2 in /usr/local/lib/python3.7/dist-packages (from nbconvert) (0.3)
Requirement already satisfied: pandocfilters>=1.4.1 in /usr/local/lib/python3.7/dist-packages (from nbconvert) (1.4.3)
Requirement already satisfied: bleach in /usr/local/lib/python3.7/dist-packages (from nbconvert) (4.0.0)
Requirement already satisfied: mistune<2,>=0.8.1 in /usr/local/lib/python3.7/dist-packages (from nbconvert) (0.8.4)
Requirement already satisfied: defusedxml in /usr/local/lib/python3.7/dist-packages (from nbconvert) (0.7.1)
Requirement already satisfied: jinja2>=2.4 in /usr/local/lib/python3.7/dist-packages (from nbconvert) (2.11.3)
Requirement already satisfied: MarkupSafe>=0.23 in /usr/local/lib/python3.7/dist-packages (from jinja2>=2.4->nbconvert) (2.0.1)
Requirement already satisfied: ipython-genutils in /usr/local/lib/python3.7/dist-packages (from nbformat>=4.4->nbconvert) (0.2.0)
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in /usr/local/lib/python3.7/dist-packages (from nbformat>=4.4->nbconvert) (2.6.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-packages (from bleach->nbconvert) (21.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.7/dist-packages (from bleach->nbconvert) (0.5.1)
Requirement already satisfied: six>=1.9.0 in /usr/local/lib/python3.7/dist-packages (from bleach->nbconvert) (1.15.0)
Requirement already satisfied: pyparsing>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packaging->bleach->nbconvert) (2.4.7)
```

In [ ]: