IMPORTING LIBRARIES

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

READ THE DATASET

```
TRAIN
```

In [3]:
```
hii = pd.read_csv('train.csv')
```

EXAMINE THE DATASET

Finding Number of Rows and Columns

In [4]:
```
hii.shape
```

Out[4]:
```
(891, 12)
```

Representing first 10 rows with columns in DataFrame

In [11]:
```
hii.head(10)
```

Out[11]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabi
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	Na
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C8
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	Na
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C12
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	Na
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	Na
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E4
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	Na
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	Na
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	Na

Representing last 10 rows with columns in DataFrame

In [9]:
```
hii.tail(10)
```

Out[9]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
881	882	0	3	Markun, Mr. Johann	male	33.0	0	0	349257	7.8958
882	883	0	3	Dahlberg, Miss. Gerda Ulrika	female	22.0	0	0	7552	10.5167
883	884	0	2	Banfield, Mr. Frederick James	male	28.0	0	0	C.A./SOTON 34068	10.5000
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SOTON/OQ 392076	7.0500
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500

Analysing DATA in terms of statistics

In [10]:
```
hii.describe()
```

Out[10]:

	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

ANALYZING BY VISUALIZING DATE

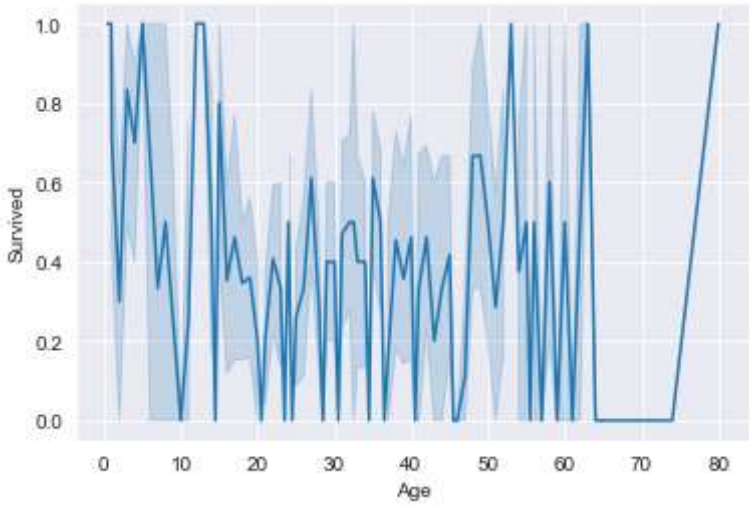
SCATTER PLOT

In [16]:
```
sns.pairplot(hii)
```

Out[16]: <seaborn.axisgrid.PairGrid at 0x1b34d3776d0>

LINE PLOT

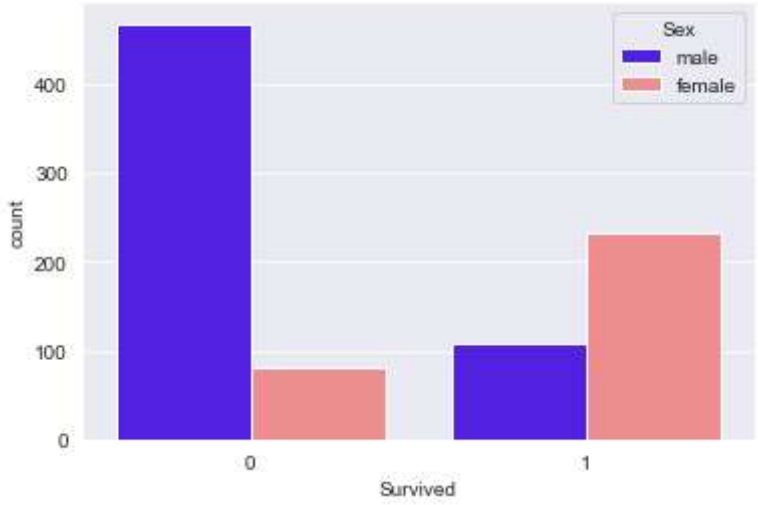
In [15]:
```
sns.lineplot(data=hii,x='Age',y='Survived')
plt.show()
```



COUNT PLOT

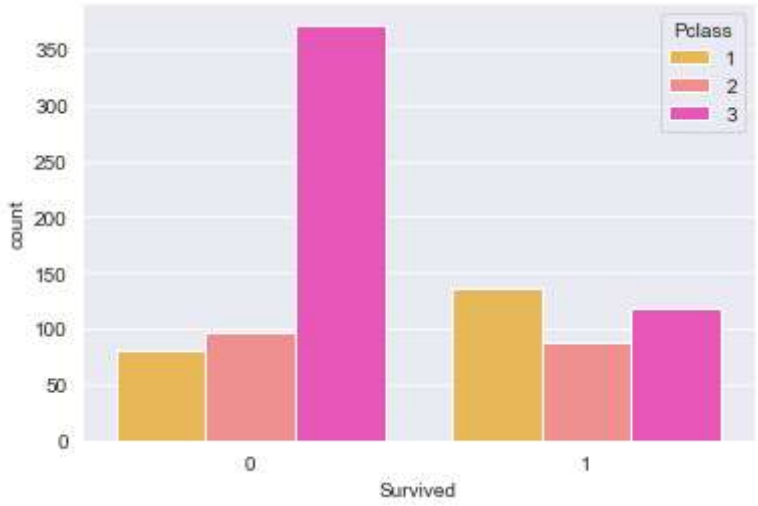
In [23]:
```
sns.set_style('darkgrid')
sns.countplot(x='Survived',hue='Sex',data=hii,palette='gnuplot2')
```

Out[23]: <AxesSubplot:xlabel='Survived', ylabel='count'>



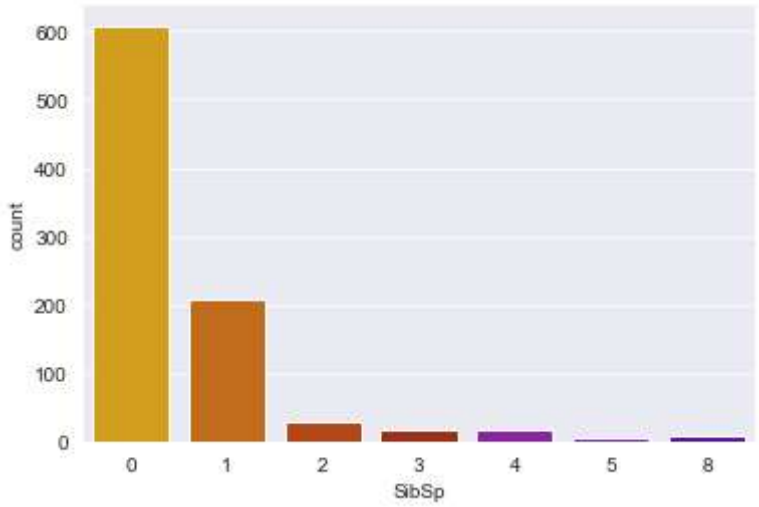
In [24]:
```
sns.set_style('darkgrid')
sns.countplot(x='Survived',hue='Pclass',data=hii,palette='spring_r')
```

Out[24]: <AxesSubplot:xlabel='Survived', ylabel='count'>



In [14]:
```
sns.countplot(x='SibSp',data=hii,palette='gnuplot_r')
```

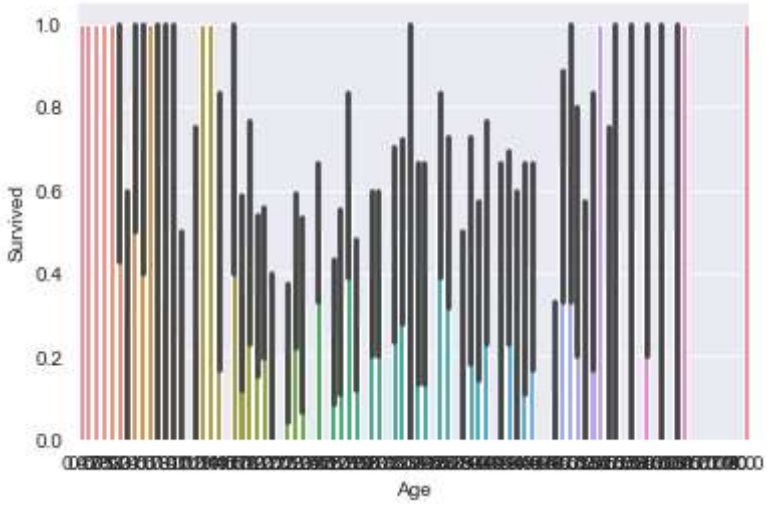
Out[14]: <AxesSubplot:xlabel='SibSp', ylabel='count'>



BAR PLOT

In [18]:
```
sns.barplot(data=hii,x="Age",y="Survived")
```

Out[18]: <AxesSubplot:xlabel='Age', ylabel='Survived'>

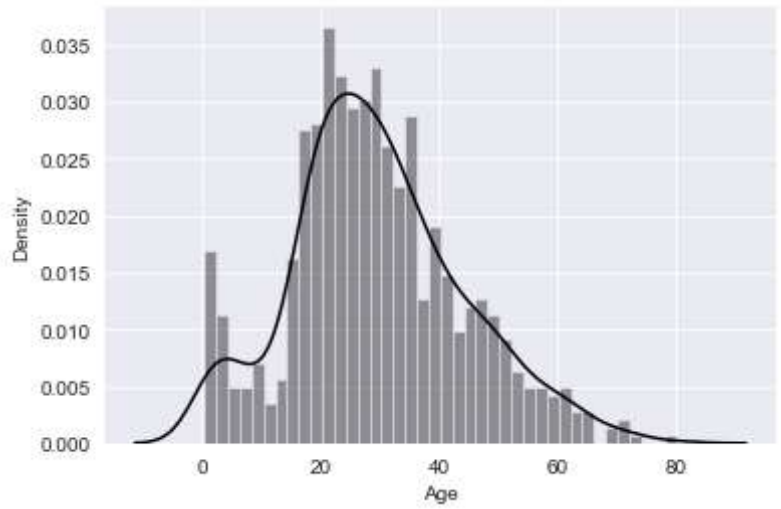


DISTPLOT

In [21]:
```
sns.distplot(hii['Age'].dropna(),color='black',bins=40)
```

```
C:\Users\pavanvamsi\anaconda3\lib\site-packages\seaborn\distributions.py:2551:
FutureWarning: `distplot` is a deprecated function and will be removed in a fut
ure version. Please adapt your code to use either `displot` (a figure-level fun
ction with similar flexibility) or `histplot` (an axes-level function for histo
grams).
  warnings.warn(msg, FutureWarning)
```

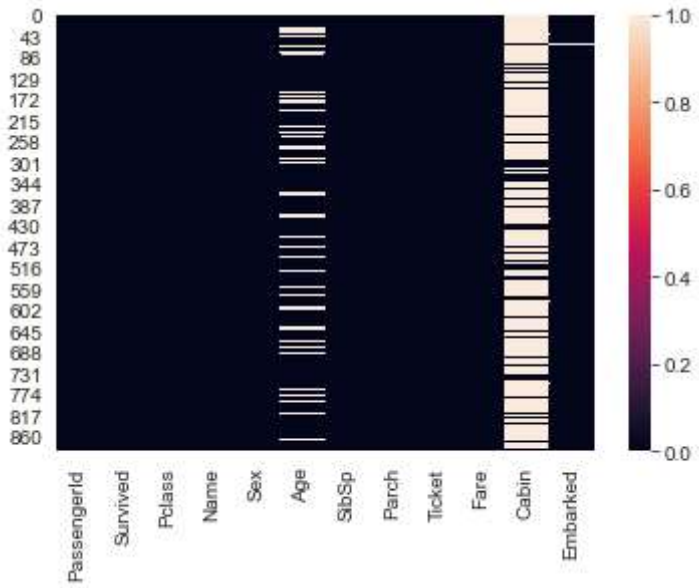
Out[21]:
```
<AxesSubplot:xlabel='Age', ylabel='Density'>
```



HEAT MAP FOR NaN

In [25]:
```
sns.heatmap(hii.isnull())
```

Out[25]: <AxesSubplot:>



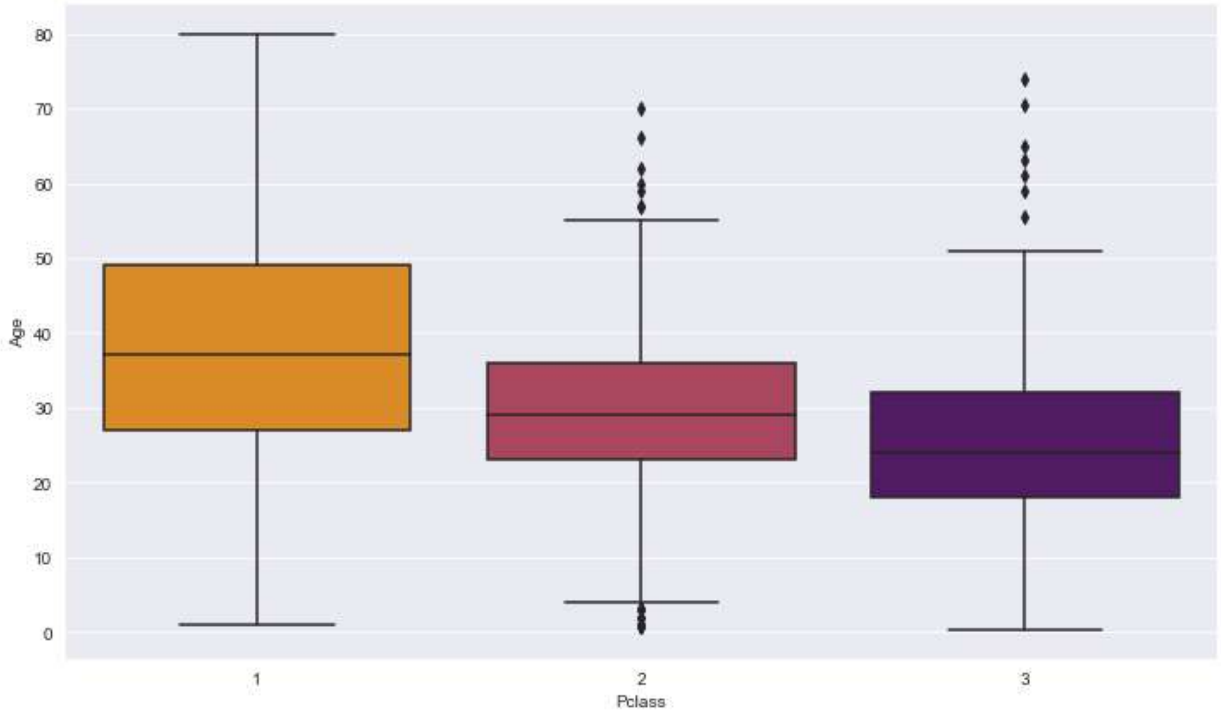
From the above visualization, it will be noticed that Age and Cabin contains NULL values. In order to use our data for Machine Learning or for any other purposes, DATA CLEANING is required.

DATA CLEANING

We want to fill in missing age data instead of just dropping the missing age data rows. One way to do this is by filling in the mean age of all the passengers (imputation).

In [29]:
```
plt.figure(figsize=(12, 7))
sns.boxplot(x='Pclass',y='Age',data=hii,palette='inferno_r')
```

Out[29]: <AxesSubplot:xlabel='Pclass', ylabel='Age'>



We'll use these average age values to impute based on Pclass for Age.

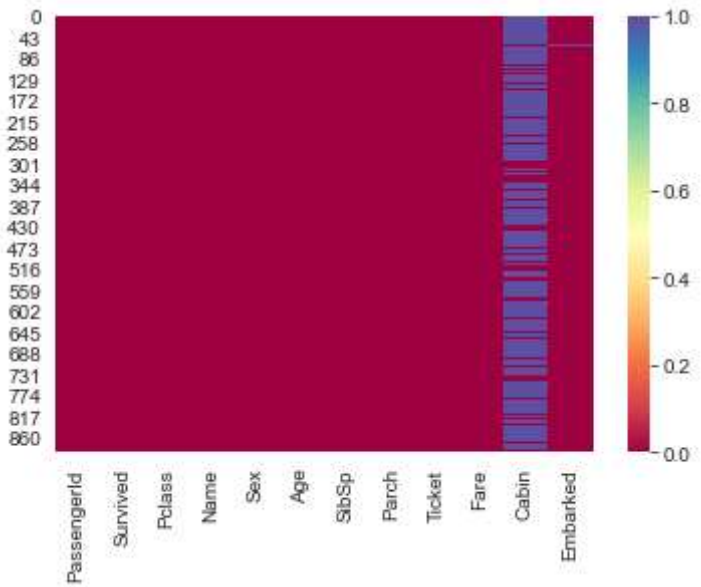
In [30]:
```
def nnull_age(cols):
    Age = cols[0]
    Pclass = cols[1]

    if pd.isnull(Age):

        if Pclass == 1:
            return 37

        elif Pclass == 2:
            return 29

        else:
            return 24

    else:
        return Age
```

In [32]:
```
hii['Age'] = hii[['Age','Pclass']].apply(nnull_age,axis=1)
```

Now check the heatmap again

In [39]:
```
sns.heatmap(hii.isnull(),cmap= ('Spectral'))
```

Out[39]: <AxesSubplot:>



From the above heat map it will be clear that values are assigned to NaN of AGE. Now we need to drop the Cabin column, because it is difficult to impute values to it.

In [40]:
```
hii.drop('Cabin',axis=1,inplace=True)
```

Let's check our dataset

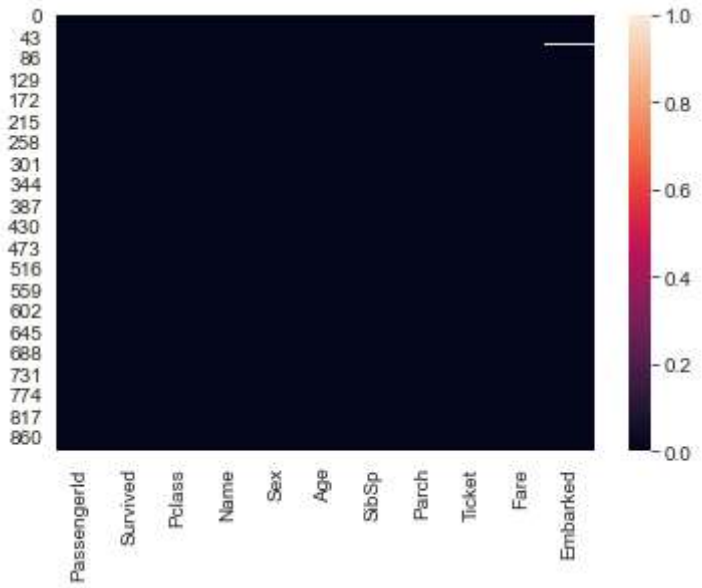
In [45]:
```
hii.head(10)
```

Out[45]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Emb
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	
5	6	0	3	Moran, Mr. James	male	24.0	0	0	330877	8.4583	
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	

In [44]:
```
sns.heatmap(hii.isnull())
```

Out[44]: <AxesSubplot:>



We noticed that null values are removed from the data by using imputation and dropna() methods.