In [1]:
```
#From https://www.kaggle.com/aqurilla/the-titanic-an-analysis
import numpy as np
import pandas as pd
```

In [2]:
```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

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

In [4]:
```
train.head()
```

Out[4]:

	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 [5]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

In [6]:
```
train.describe()
```

Out[6]:

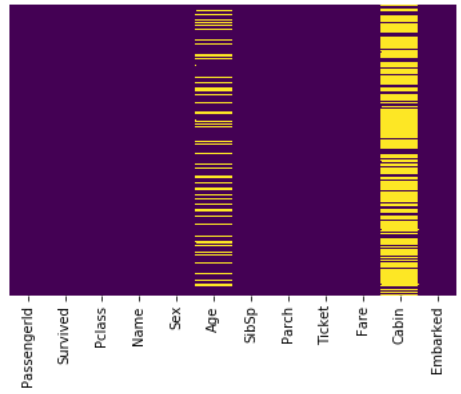
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	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 [7]:

```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

Out[7]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x16aaf57e358>
```



In [8]:

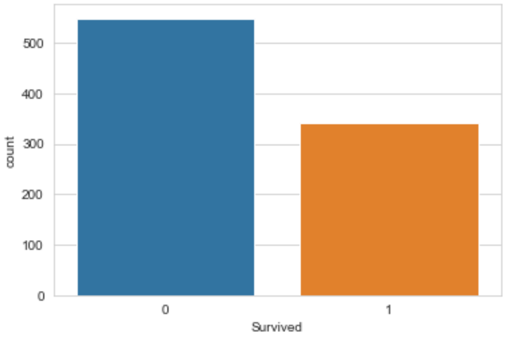
```
sns.set_style('whitegrid')
```

In [9]:

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

Out[9]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x16aaf885c88>
```

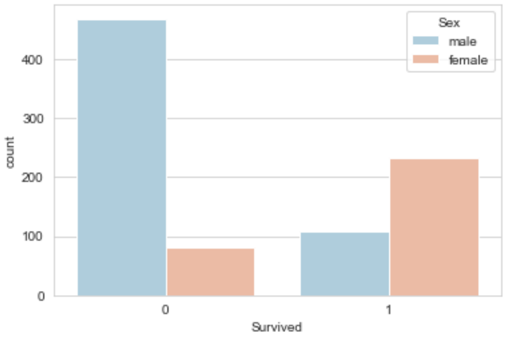


In [10]:

```
#EDA2
sns.countplot(x='Survived',data=train,hue='Sex',palette='RdBu_r')
```

Out[10]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x16aaf903d68>
```

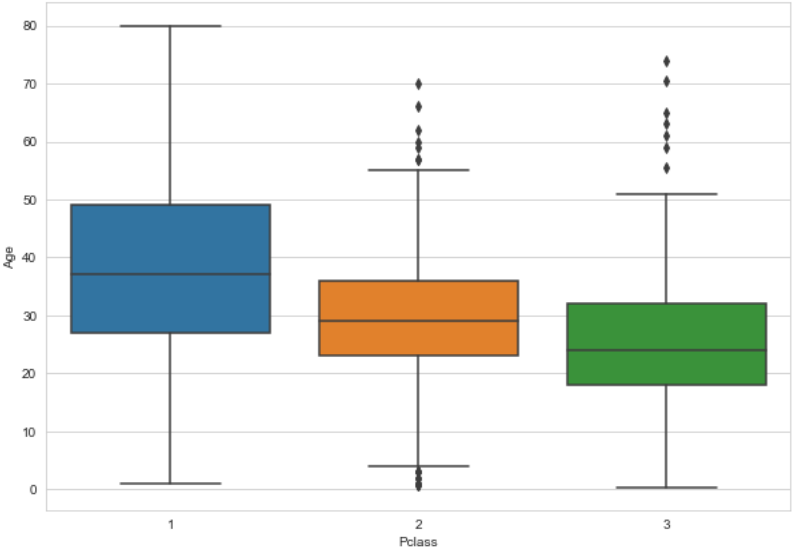


In [11]:

```
#Data analysis
plt.figure(figsize=(10,7))
sns.boxplot(x='Pclass',y='Age',data=train)
```

Out[11]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x16aaf92cc50>
```



In [12]:

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

In [13]:

```
#Data fill
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 [14]:

```
train['Age'] = train[['Age','Pclass']].apply(impute_age, axis=1)
```

In [15]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            891 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

In [16]:
```
train.drop('Cabin', inplace=True, axis=1)
train.drop('Ticket', inplace=True, axis=1)
```

In [17]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 10 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            891 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Fare           891 non-null float64
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(3)
memory usage: 69.7+ KB
```

In [18]:
```
train.dropna(inplace=True)
```

In [19]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 889 entries, 0 to 890
Data columns (total 10 columns):
PassengerId    889 non-null int64
Survived       889 non-null int64
Pclass         889 non-null int64
Name           889 non-null object
Sex            889 non-null object
Age            889 non-null float64
SibSp          889 non-null int64
Parch          889 non-null int64
Fare           889 non-null float64
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(3)
memory usage: 76.4+ KB
```

In [20]:
```
train['Sex'] = pd.get_dummies(train['Sex'],drop_first=True)
```

In [21]:
```
train.head()
```

Out[21]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	1	22.0	1	0	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	0	26.0	0	0	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	53.1000	S
4	5	0	3	Allen, Mr. William Henry	1	35.0	0	0	8.0500	S

In [22]:
```
train.drop('Embarked', inplace=True, axis=1)
```

In [23]:

```
train.head()
```

Out[23]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Fare
0	1	0	3	Braund, Mr. Owen Harris	1	22.0	1	0	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	71.2833
2	3	1	3	Heikkinen, Miss. Laina	0	26.0	0	0	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	53.1000
4	5	0	3	Allen, Mr. William Henry	1	35.0	0	0	8.0500

In [24]:

```
train.drop('Name', inplace=True, axis=1)
train.head()
```

Out[24]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
0	1	0	3	1	22.0	1	0	7.2500
1	2	1	1	0	38.0	1	0	71.2833
2	3	1	3	0	26.0	0	0	7.9250
3	4	1	1	0	35.0	1	0	53.1000
4	5	0	3	1	35.0	0	0	8.0500

In [25]:

```
train.drop('PassengerId', inplace=True, axis=1)
train.head()
```

Out[25]:

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

In [26]:
```
pclass = pd.get_dummies(train['Pclass'], drop_first=True)
pclass.columns=['Class=2','Class=3']
pclass
```

Out[26]:

	Class=2	Class=3
0	0	1
1	0	0
2	0	1
3	0	0
4	0	1
5	0	1
6	0	0
7	0	1
8	0	1
9	1	0
10	0	1
11	0	0
12	0	1
13	0	1
14	0	1
15	1	0
16	0	1
17	1	0
18	0	1
19	0	1
20	1	0
21	1	0
22	0	1
23	0	0
24	0	1
25	0	1
26	0	1
27	0	0
28	0	1
29	0	1
...	...	...
861	1	0
862	0	0
863	0	1
864	1	0
865	1	0
866	1	0
867	0	0
868	0	1
869	0	1
870	0	1
871	0	0
872	0	0
873	0	1
874	1	0
875	0	1
876	0	1
877	0	1
878	0	1
879	0	0
880	1	0
881	0	1
882	0	1
883	1	0

	Class=2	Class=3
884	0	1
885	0	1
886	1	0
887	0	0
888	0	1
889	0	0
890	0	1

889 rows × 2 columns

In [27]:

```
train.head()
```

Out[27]:

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

In [28]:

```
train=pd.concat([train,pclass],axis=1)
train.head()
```

Out[28]:

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

In [29]:

```
train.drop('Pclass',inplace=True,axis=1)
train.head()
```

Out[29]:

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

In [30]:

```
train.columns
```

Out[30]:

```
Index(['Survived', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Class=2',
       'Class=3'],
      dtype='object')
```

In [ ]: