Titanic Data EDA

December 20, 2022

[1]:
```
print("Hello World.This is my first EDA in Titanic Data")
```

```
Hello World.This is my first EDA in Titanic Data
```

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

[3]:
```
titanic_data = pd.read_csv("https://raw.githubusercontent.com/mohittomar2008/
↪Titanic-Project/main/titanic_train.csv")
```

[4]:
```
titanic_data
```

[4]:

	PassengerId	Survived	Pclass	\
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	0	3	
..	…	…	…	
886	887	0	2	
887	888	1	1	
888	889	0	3	
889	890	1	1	
890	891	0	3	

	Name	Sex	Age	SibSp	\
0	Braund, Mr. Owen Harris	male	22.0	1	
1	Cumings, Mrs. John Bradley (Florence Briggs Th…	female	38.0	1	
2	Heikkinen, Miss. Laina	female	26.0	0	
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	
4	Allen, Mr. William Henry	male	35.0	0	
..	…	…	…	…	
886	Montvila, Rev. Juozas	male	27.0	0	
887	Graham, Miss. Margaret Edith	female	19.0	0	
888	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	

```
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     …               …       …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

Exploratory Data Analysis

Let's begin some exptoratory data analysis! We'll start by checking out missing data1

[5]:
```
## Seeing columns names
titanic_data.columns
```

[5]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

[6]:
```
## Geting information about data
titanic_data.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
```

[7]:
```
titanic_data.shape
```

[7]:
```
(891, 12)
```

Missing data

We can use seaborn to create a simple heatmap to see where we are missing data!

[8]:
```
# finding where valeu is null
titanic_data.isnull()
```

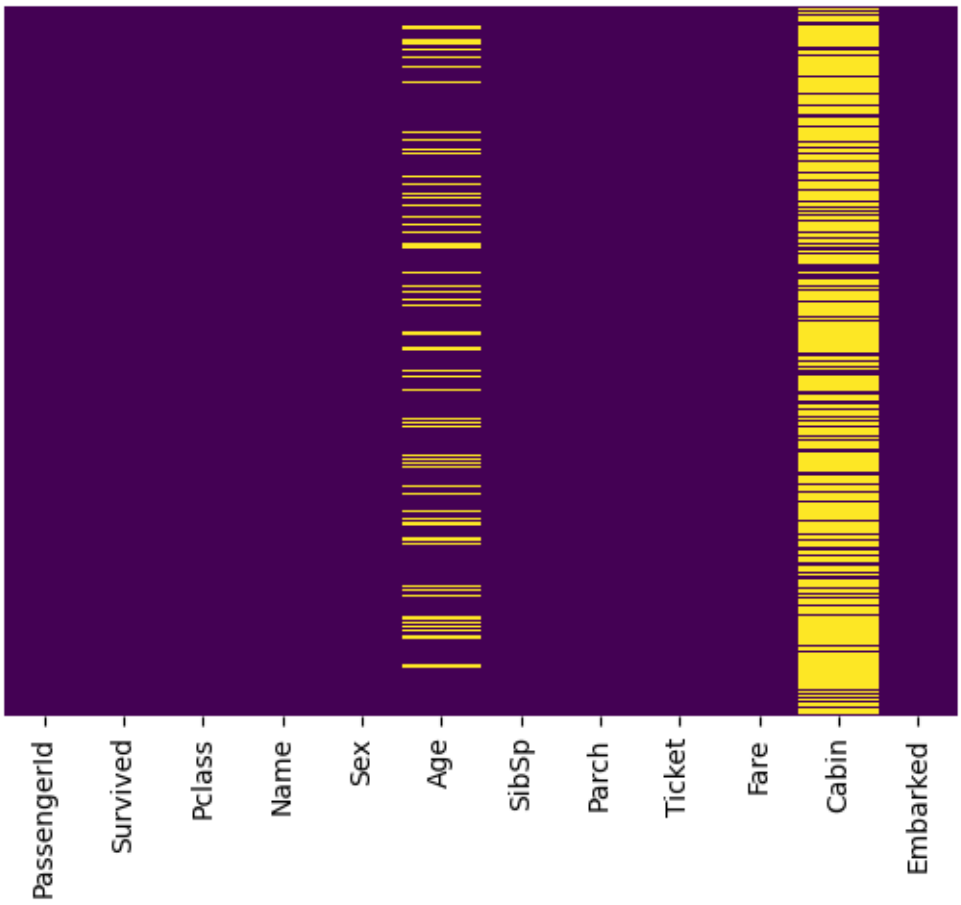
[8]:
```
     PassengerId  Survived  Pclass   Name    Sex    Age  SibSp  Parch  Ticket  \
0          False     False   False  False  False  False  False  False   False
1          False     False   False  False  False  False  False  False   False
2          False     False   False  False  False  False  False  False   False
3          False     False   False  False  False  False  False  False   False
4          False     False   False  False  False  False  False  False   False
..           …         …     …      …      …      …      …      …       …
886        False     False   False  False  False  False  False  False   False
887        False     False   False  False  False  False  False  False   False
888        False     False   False  False  False   True  False  False   False
889        False     False   False  False  False  False  False  False   False
890        False     False   False  False  False  False  False  False   False

      Fare  Cabin  Embarked
0    False   True     False
1    False  False     False
2    False   True     False
3    False  False     False
4    False   True     False
..     …    …       …
886  False   True     False
887  False  False     False
888  False   True     False
889  False  False     False
890  False   True     False

[891 rows x 12 columns]
```

[9]:
```
## viewing null value using heatmap
sns.heatmap(titanic_data.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

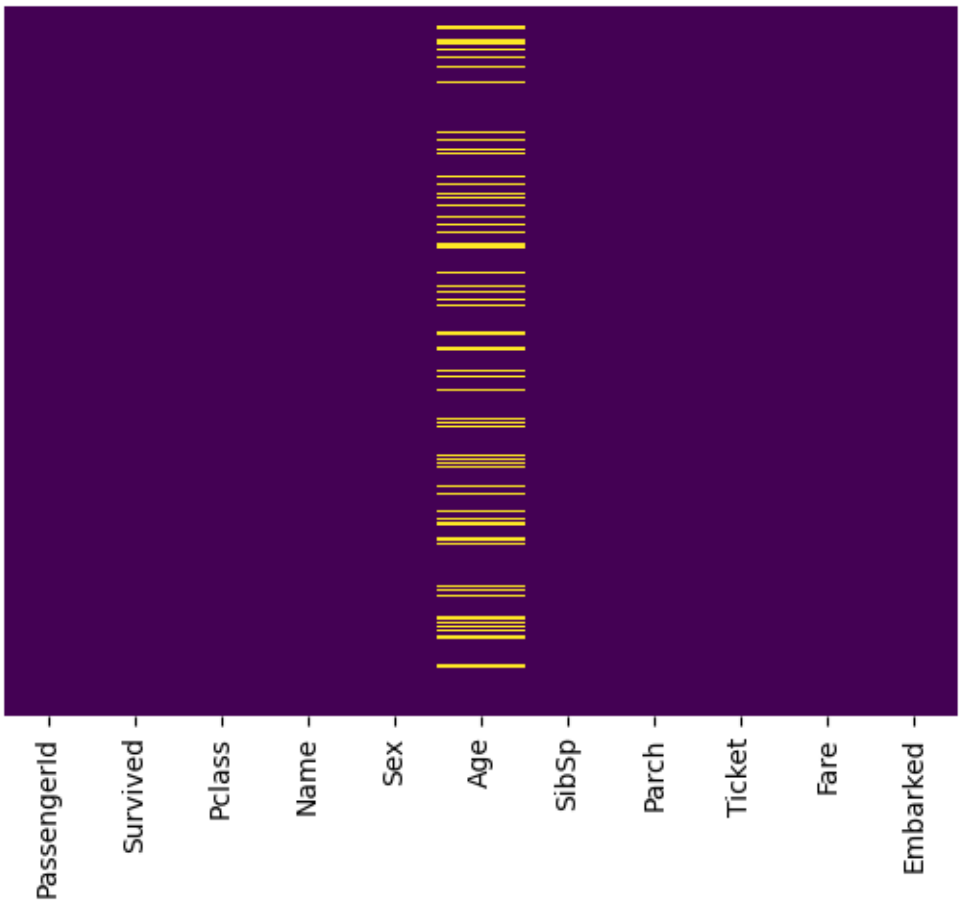
[9]:
```
<AxesSubplot: >
```



[10]:
```
# Droping column cabin
titanic_data.drop('Cabin',axis=1,inplace=True)
```

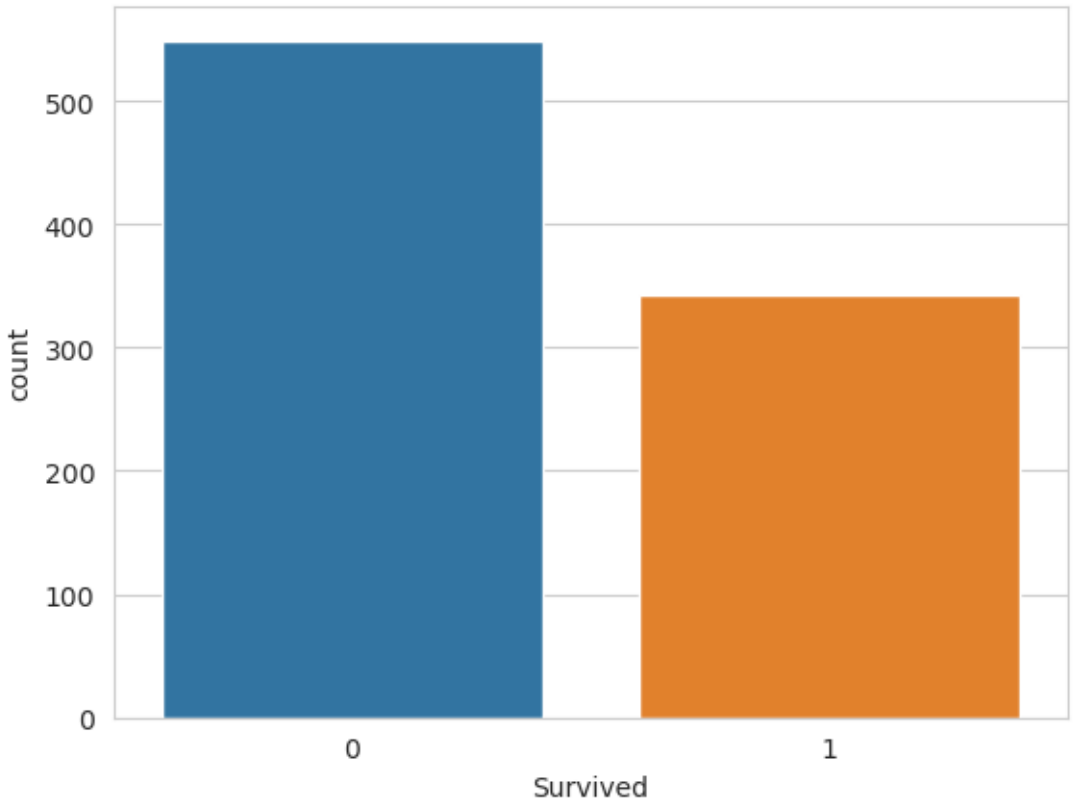
[11]:
```
## viewing null value using heatmap
sns.heatmap(titanic_data.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

[11]: <AxesSubplot: >



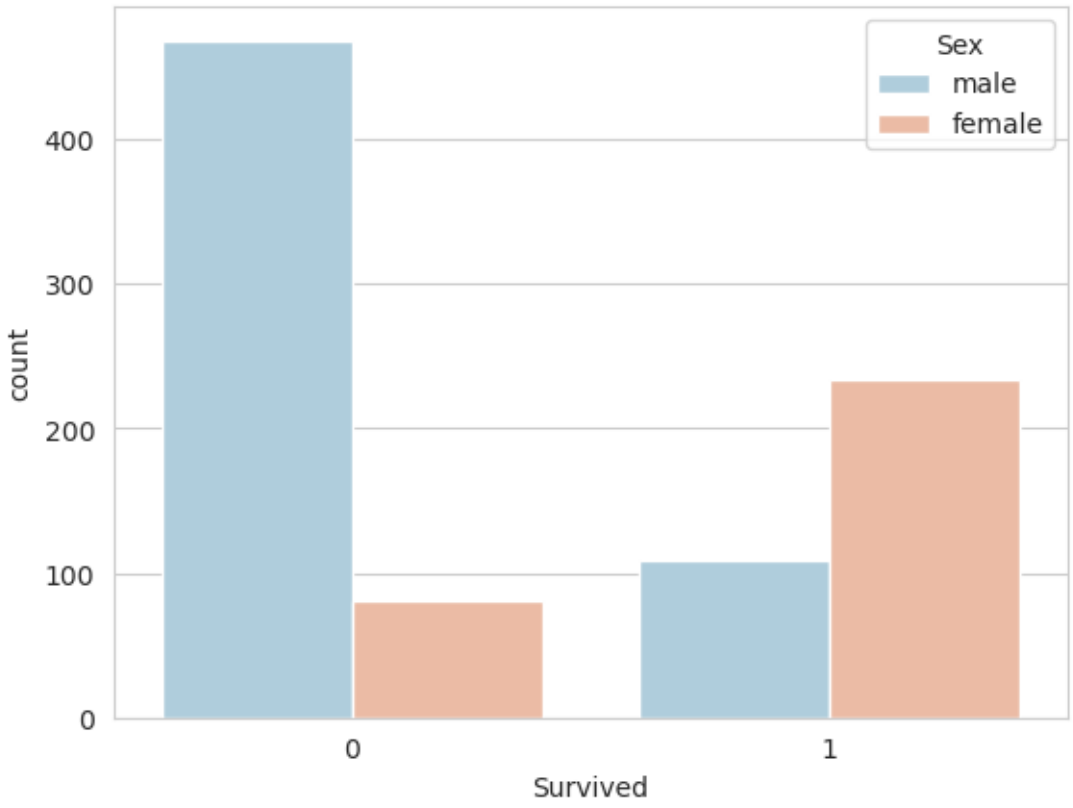
[12]:
```
##Viewing survived  data using countplot
sns.set_style('whitegrid')
sns.countplot(x='Survived',data=titanic_data)
```

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



[13]:
```
##Viewing sex and survived  data using countplot
sns.set_style('whitegrid')
sns.countplot(x='Survived',hue='Sex',data=titanic_data,palette='RdBu_r')
```

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



[14]:
```
titanic_data.head()
```

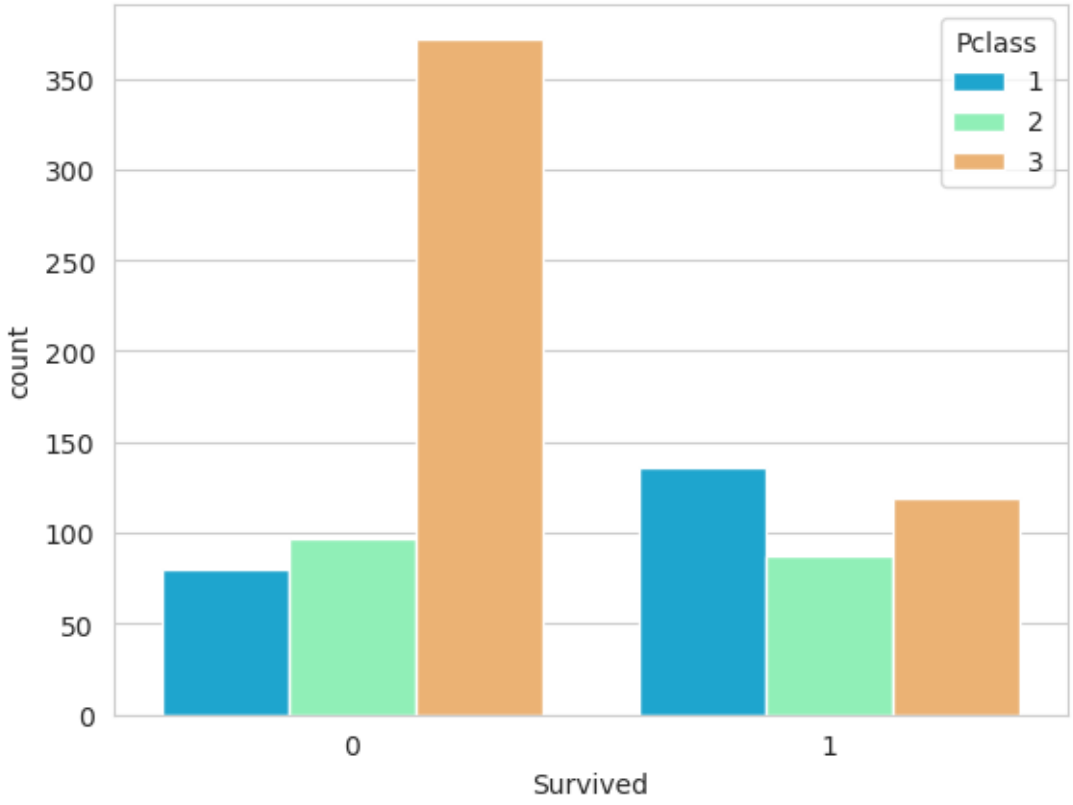
[14]:

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

```
4       0             373450   8.0500        S
```

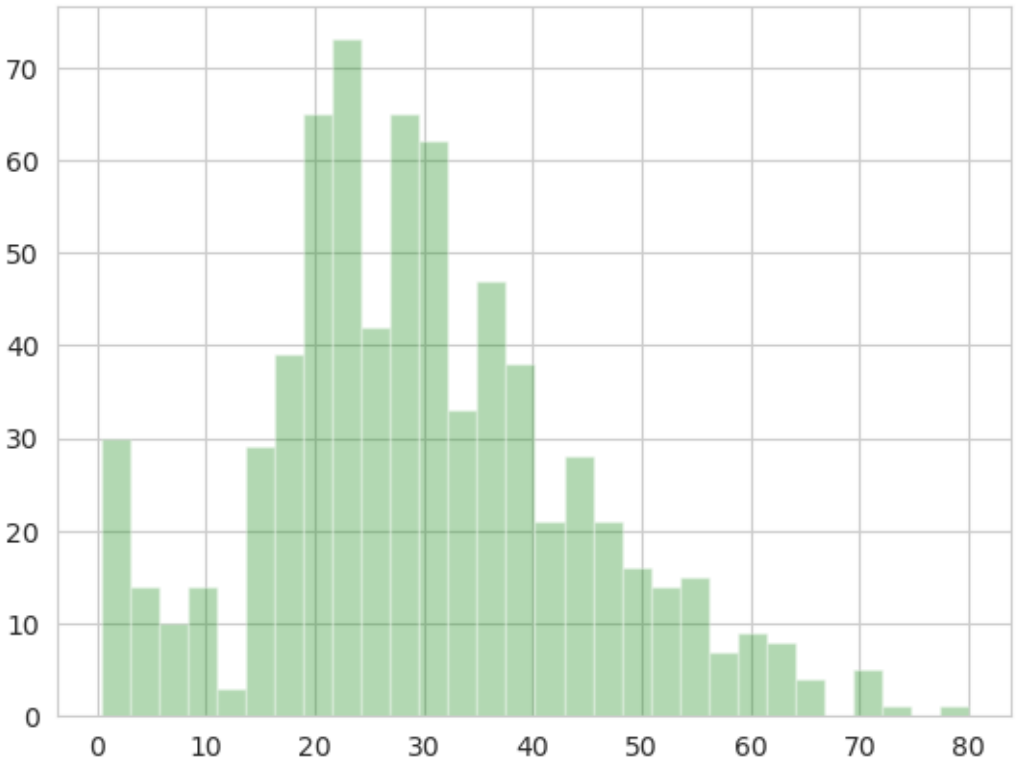
[17]:
```
##Viewing Passengerclass and survived  data using countplot
sns.set_style('whitegrid')
sns.countplot(x='Survived',hue='Pclass',data=titanic_data,palette='rainbow')
```

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



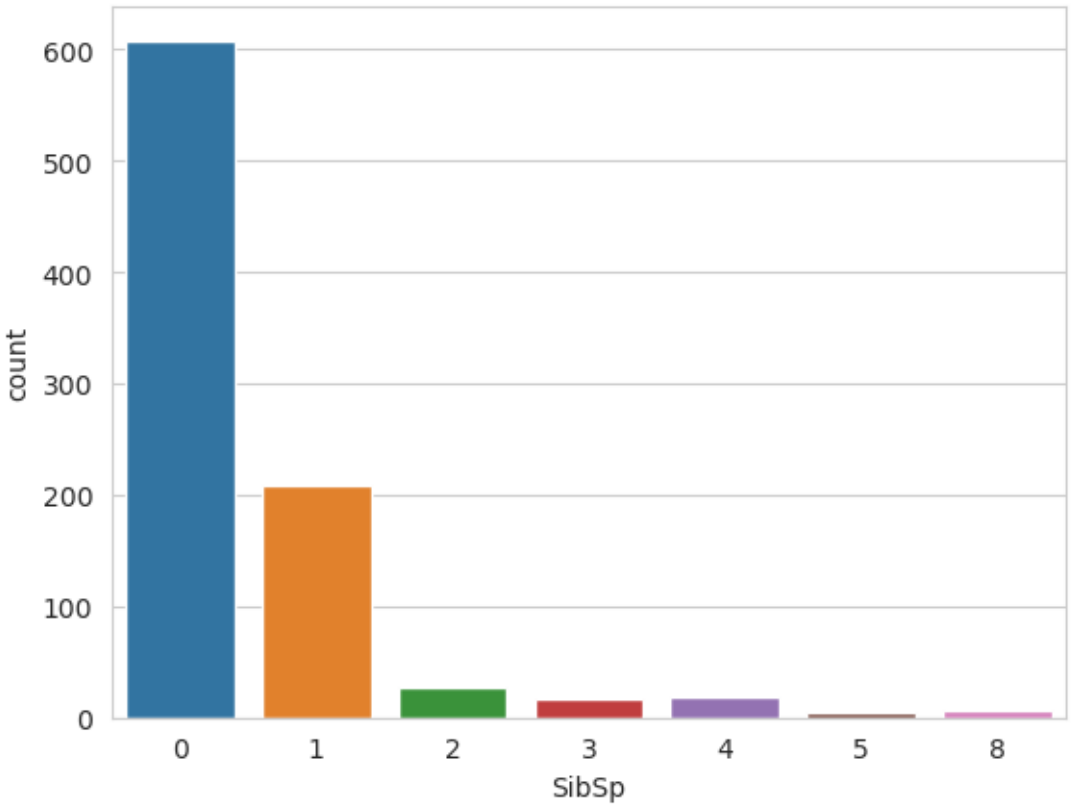
[18]:
```
titanic_data['Age'].hist(bins=30,color='g',alpha=0.3)
```

[18]: <AxesSubplot: >



[19]:
```
sns.countplot(x='SibSp',data=titanic_data)
```

[19]:
```
<AxesSubplot: xlabel='SibSp', ylabel='count'>
```

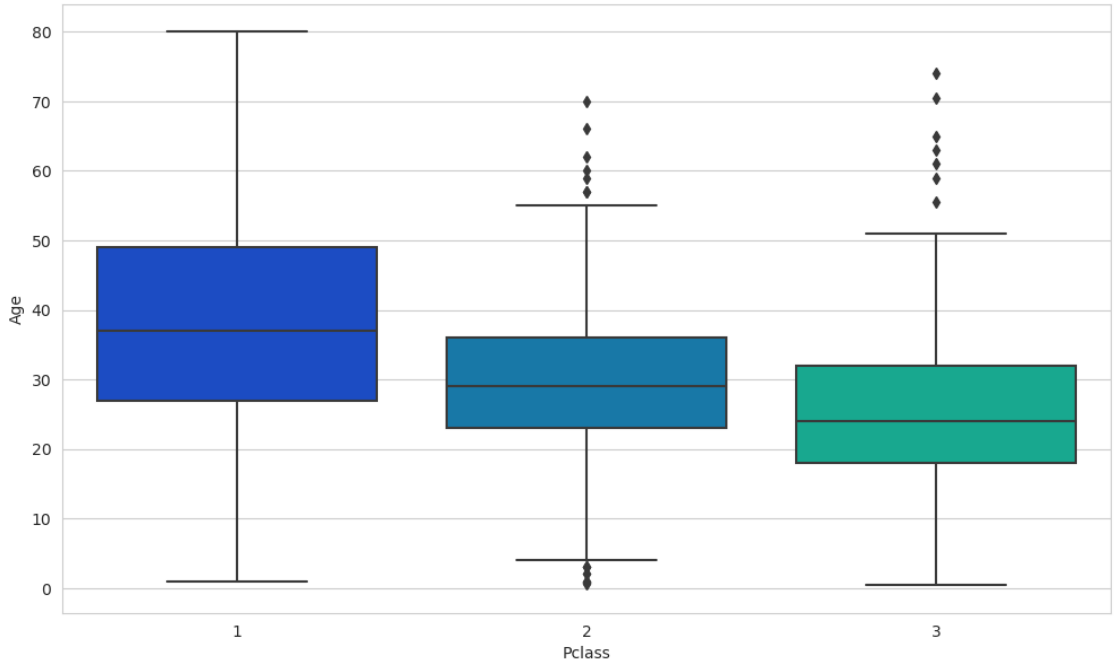


Data Cleaning

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). However we can be smarter about this and check the average age by passenger class.For example

[20]:
```
plt.figure(figsize=(12, 7))
sns.boxplot(x='Pclass',y='Age',data=titanic_data,palette='winter')
```

[20]:
```
<AxesSubplot: xlabel='Pclass', ylabel='Age'>
```



[21]:
```
titanic_data[['Age','Embarked']].isnull().sum()
```

[21]:
```
Age         177
Embarked      2
dtype: int64
```

[22]:
```
# sum of null value from dataset
titanic_data.isnull().sum()
```

[22]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Embarked         2
dtype: int64
```

[23]:
```
titanic_data['Age'].value_counts()
```

[23]:
```
24.00    30
22.00    27
18.00    26
19.00    25
28.00    25
         ..
36.50     1
55.50     1
0.92      1
23.50     1
74.00     1
Name: Age, Length: 88, dtype: int64
```

[24]:
```
titanic_data['Age'].unique()
```

[24]:
```
array([22.  , 38.  , 26.  , 35.  ,   nan, 54.  ,  2.  , 27.  , 14.  ,
        4.  , 58.  , 20.  , 39.  , 55.  , 31.  , 34.  , 15.  , 28.  ,
        8.  , 19.  , 40.  , 66.  , 42.  , 21.  , 18.  ,  3.  ,  7.  ,
       49.  , 29.  , 65.  , 28.5 ,  5.  , 11.  , 45.  , 17.  , 32.  ,
       16.  , 25.  ,  0.83, 30.  , 33.  , 23.  , 24.  , 46.  , 59.  ,
       71.  , 37.  , 47.  , 14.5 , 70.5 , 32.5 , 12.  ,  9.  , 36.5 ,
       51.  , 55.5 , 40.5 , 44.  ,  1.  , 61.  , 56.  , 50.  , 36.  ,
       45.5 , 20.5 , 62.  , 41.  , 52.  , 63.  , 23.5 ,  0.92, 43.  ,
       60.  , 10.  , 64.  , 13.  , 48.  ,  0.75, 53.  , 57.  , 80.  ,
       70.  , 24.5 ,  6.  ,  0.67, 30.5 ,  0.42, 34.5 , 74.  ])
```

[25]:
```
titanic = titanic_data.fillna(method="pad")
```

[26]:
```
titanic['Age'].fillna(method="ffill")
```

[26]:
```
0      22.0
1      38.0
2      26.0
3      35.0
4      35.0
       …
886    27.0
887    19.0
888    19.0
889    26.0
890    32.0
Name: Age, Length: 891, dtype: float64
```

[27]:
```
titanic=titanic.fillna(method='ffill')
```

[28]:
```
titanic.isnull().sum()
```

[28]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

[29]:
```
titanic.fillna(method='bfill')
```

[29]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th…	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S
..	…	…	…	…	…	…	…	…	…	…	…
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	19.0	1	2	W./C. 6607	23.4500	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Q

```
..     …                …       …      …
886      0            211536  13.0000        S
887      0            112053  30.0000        S
888      2        W./C. 6607  23.4500        S
889      0            111369  30.0000        C
890      0            370376   7.7500        Q

[891 rows x 11 columns]
```

[30]:
```
titanic.isna()
```

[30]:
```
     PassengerId  Survived  Pclass   Name    Sex    Age  SibSp  Parch  Ticket  \
0          False     False   False  False  False  False  False  False   False
1          False     False   False  False  False  False  False  False   False
2          False     False   False  False  False  False  False  False   False
3          False     False   False  False  False  False  False  False   False
4          False     False   False  False  False  False  False  False   False
..           …         …     …      …      …      …      …      …
886        False     False   False  False  False  False  False  False   False
887        False     False   False  False  False  False  False  False   False
888        False     False   False  False  False  False  False  False   False
889        False     False   False  False  False  False  False  False   False
890        False     False   False  False  False  False  False  False   False

      Fare  Embarked
0    False     False
1    False     False
2    False     False
3    False     False
4    False     False
..     …         …
886  False     False
887  False     False
888  False     False
889  False     False
890  False     False

[891 rows x 11 columns]
```

[31]:
```
titanic.isna().sum()
```

[31]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
```

```
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

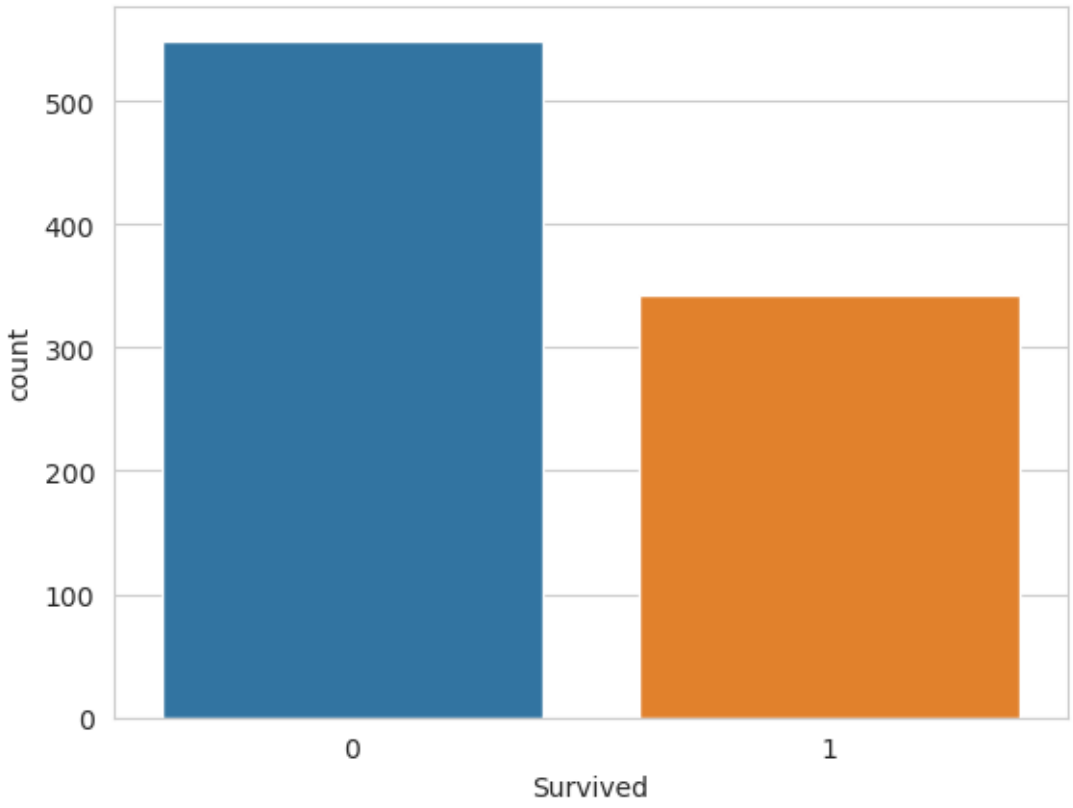
[32]:
```
titanic.tail()
```

[32]:
```
     PassengerId  Survived  Pclass                                      Name  \
886          887         0       2                     Montvila, Rev. Juozas
887          888         1       1              Graham, Miss. Margaret Edith
888          889         0       3  Johnston, Miss. Catherine Helen "Carrie"
889          890         1       1                     Behr, Mr. Karl Howell
890          891         0       3                       Dooley, Mr. Patrick

        Sex   Age  SibSp  Parch      Ticket   Fare Embarked
886    male  27.0      0      0      211536  13.00        S
887  female  19.0      0      0      112053  30.00        S
888  female  19.0      1      2  W./C. 6607  23.45        S
889    male  26.0      0      0      111369  30.00        C
890    male  32.0      0      0      370376   7.75        Q
```

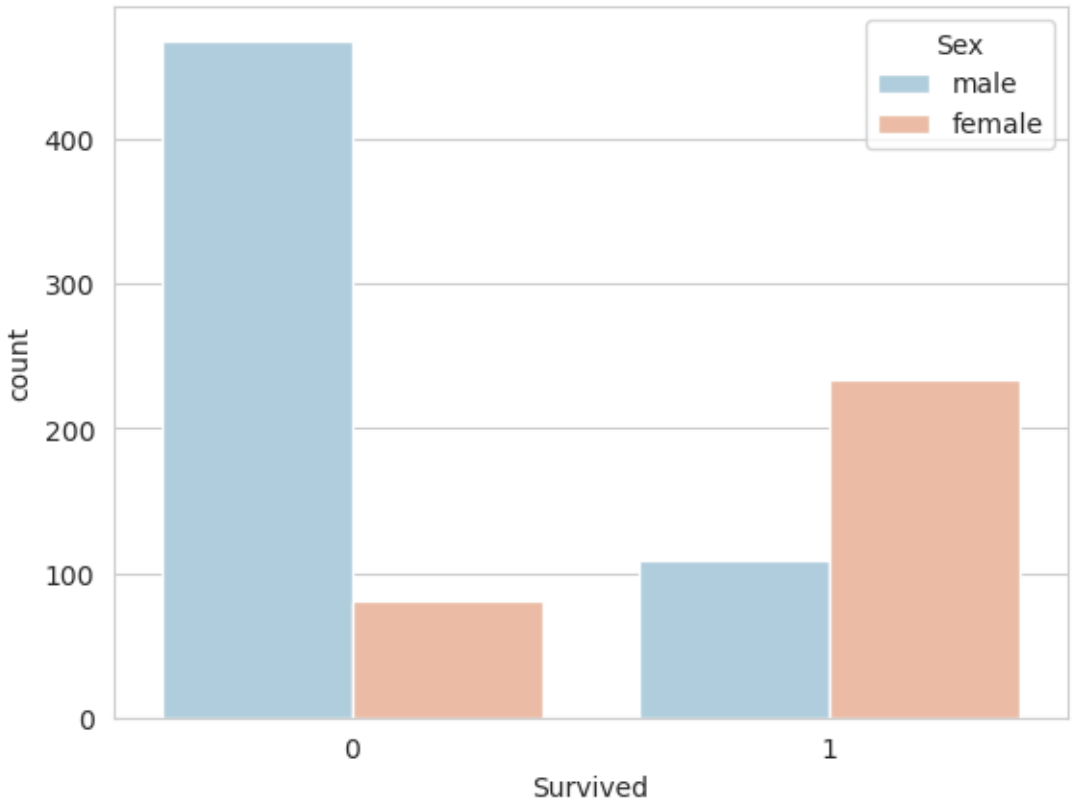
[33]:
```
##Viewing survived  data using countplot
sns.set_style('whitegrid')
sns.countplot(x='Survived',data=titanic)
```

[33]:
```
<AxesSubplot: xlabel='Survived', ylabel='count'>
```



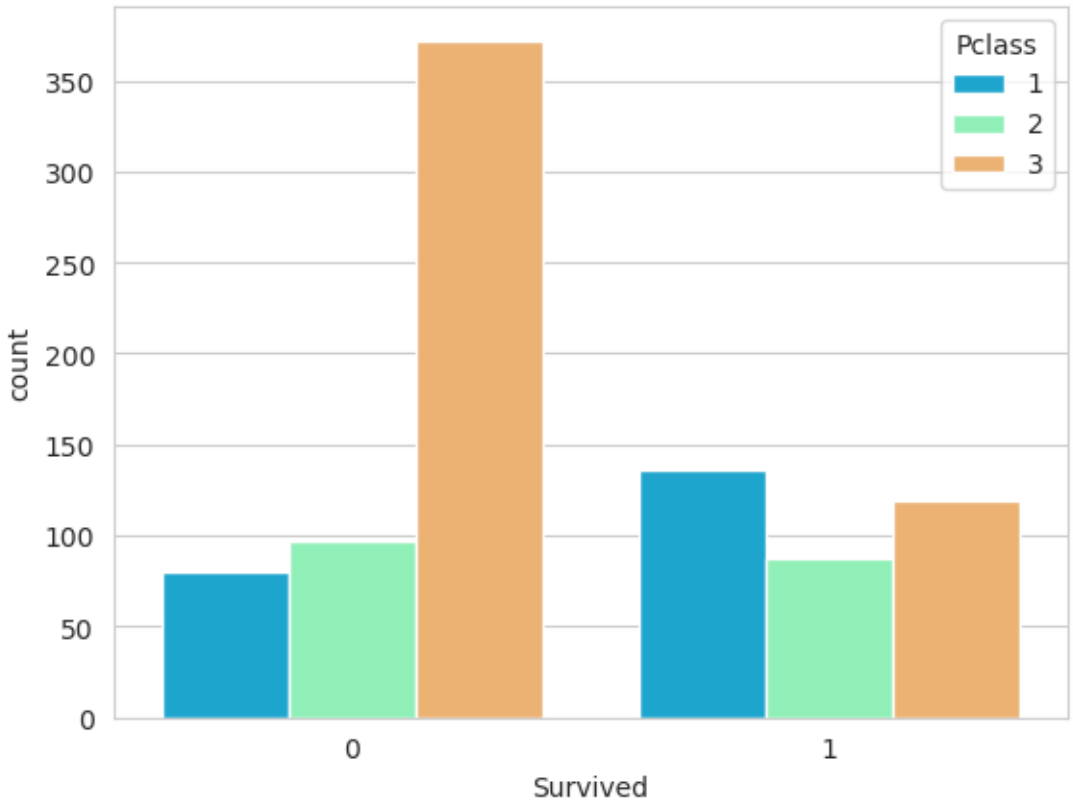
[34]:
```
##Viewing sex and survived  data using countplot
sns.set_style('whitegrid')
sns.countplot(x='Survived',hue='Sex',data=titanic,palette='RdBu_r')
```

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



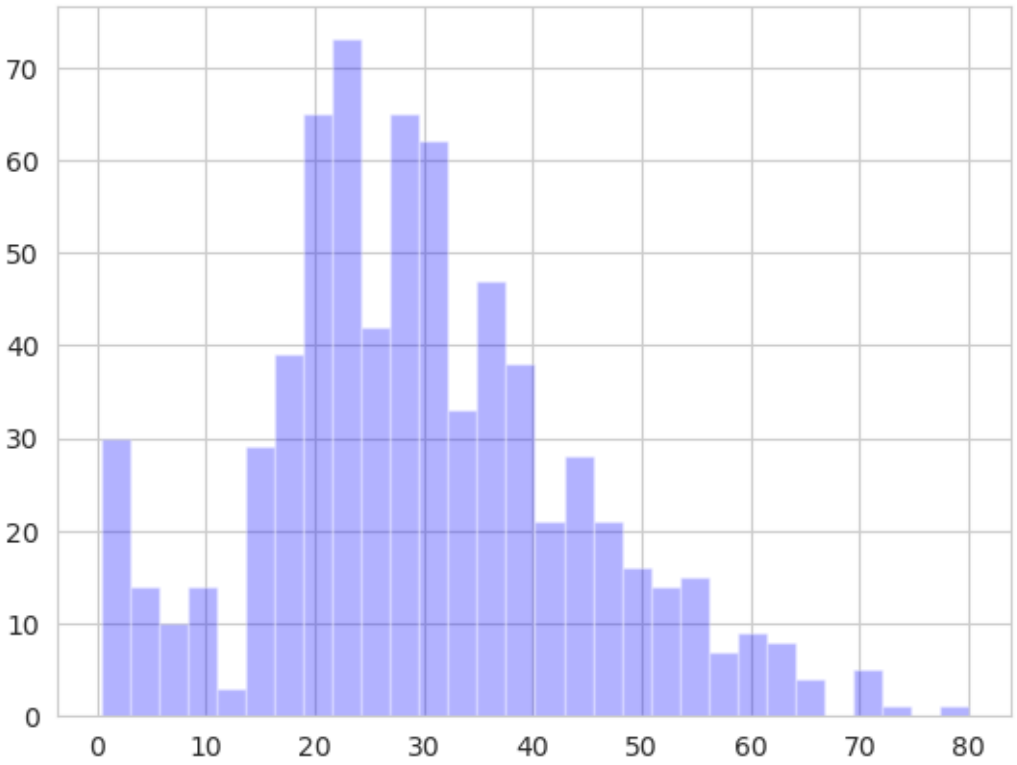
[35]:
```
##Viewing Passengerclass and survived  data using countplot
sns.set_style('whitegrid')
sns.countplot(x='Survived',hue='Pclass',data=titanic,palette='rainbow')
```

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



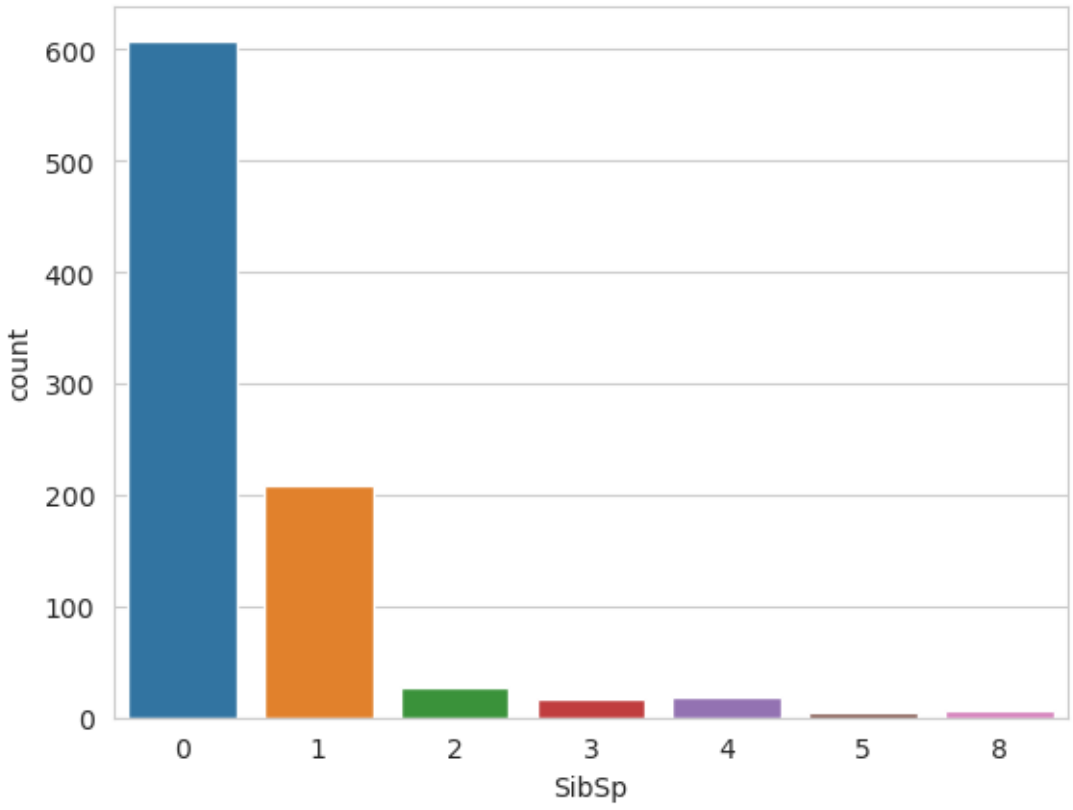
[36]:
```
titanic_data['Age'].hist(bins=30,color='b',alpha=0.3)
```

[36]: <AxesSubplot: >



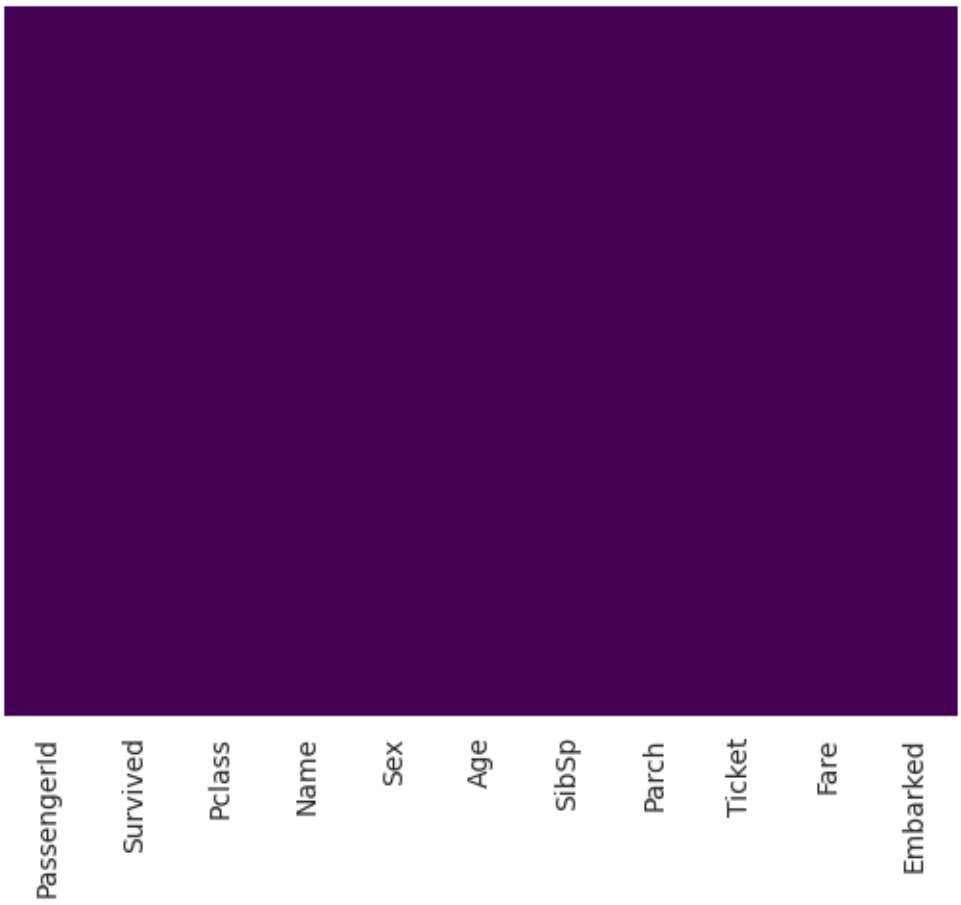
[37]:
```
sns.countplot(x='SibSp',data=titanic)
```

[37]:
```
<AxesSubplot: xlabel='SibSp', ylabel='count'>
```



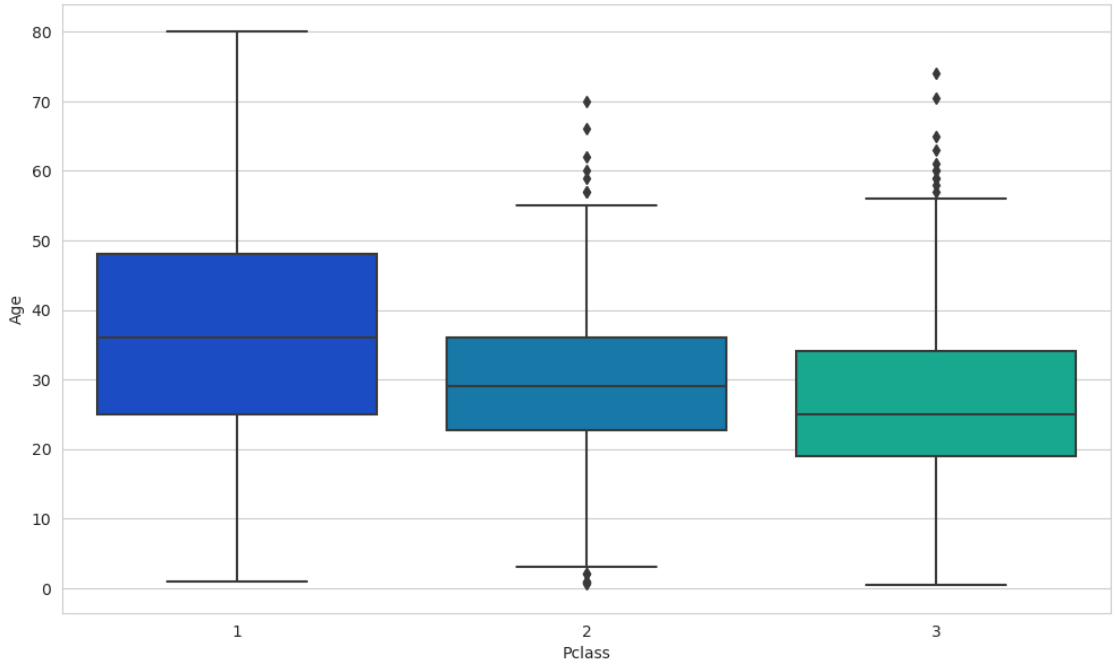
[38]:
```
## viewing null value using heatmap
sns.heatmap(titanic.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

[38]: <AxesSubplot: >



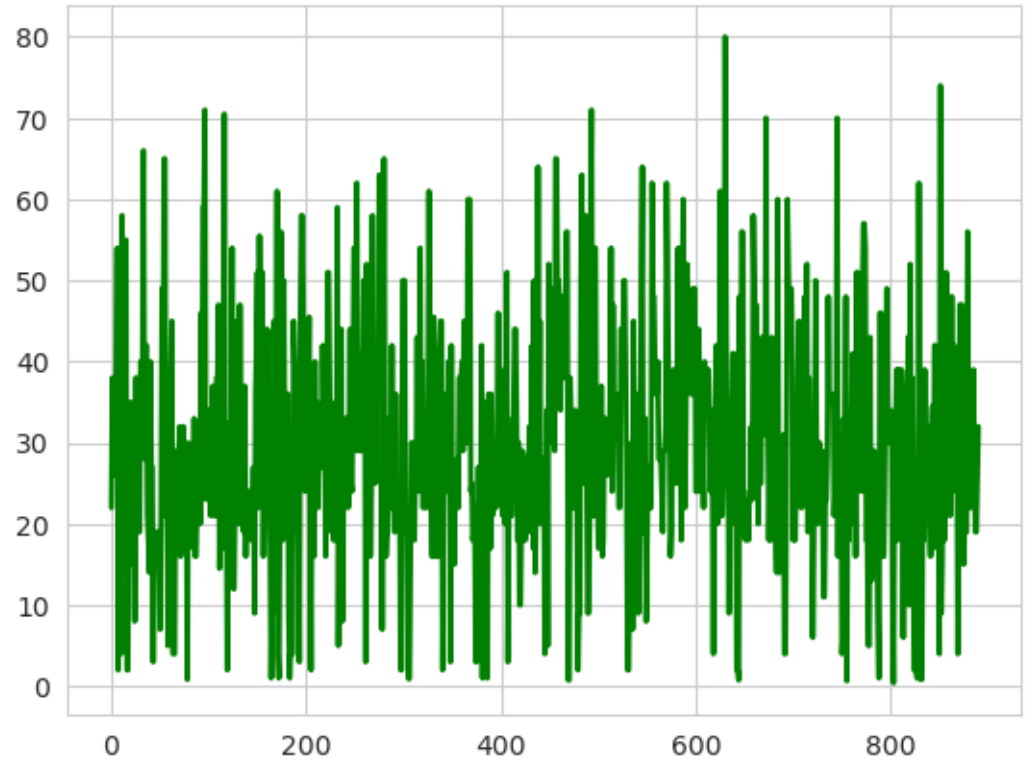
[39]:
```
plt.figure(figsize=(12, 7))
sns.boxplot(x='Pclass',y='Age',data=titanic,palette='winter')
```

[39]:
```
<AxesSubplot: xlabel='Pclass', ylabel='Age'>
```



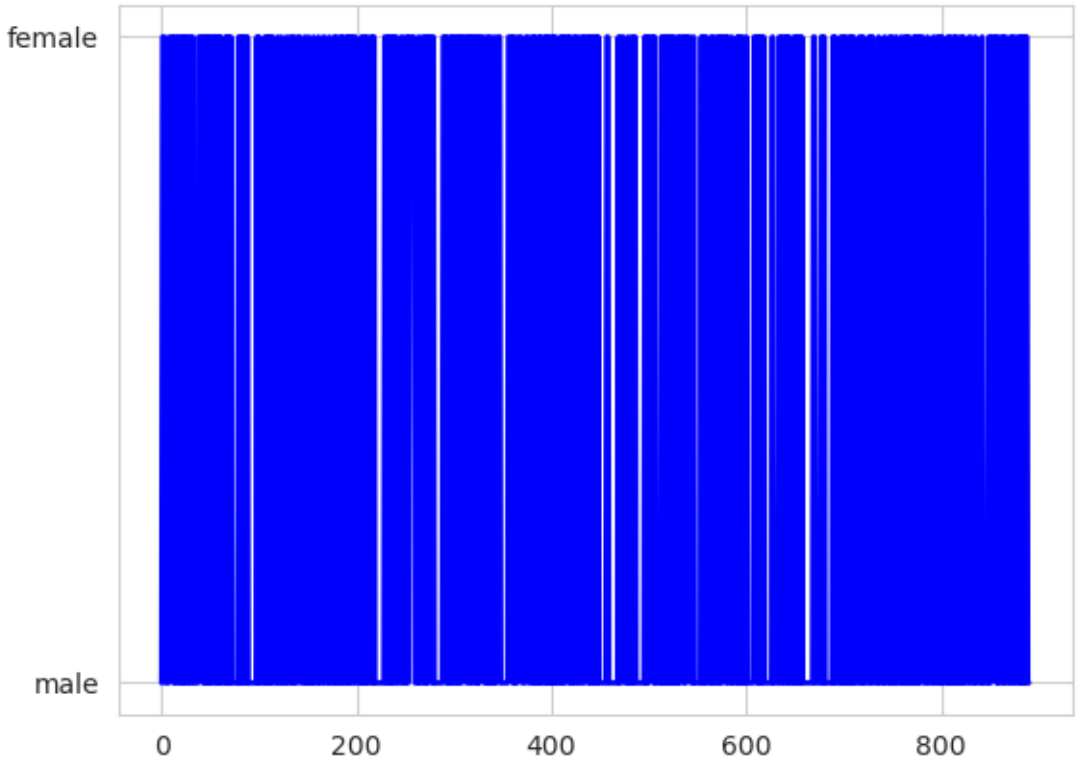
[40]:
```
plt.plot(titanic['Age'],color = 'g',label = 'Survived', linewidth = 2)
```

[40]: [<matplotlib.lines.Line2D at 0x7f9b96d24190>]



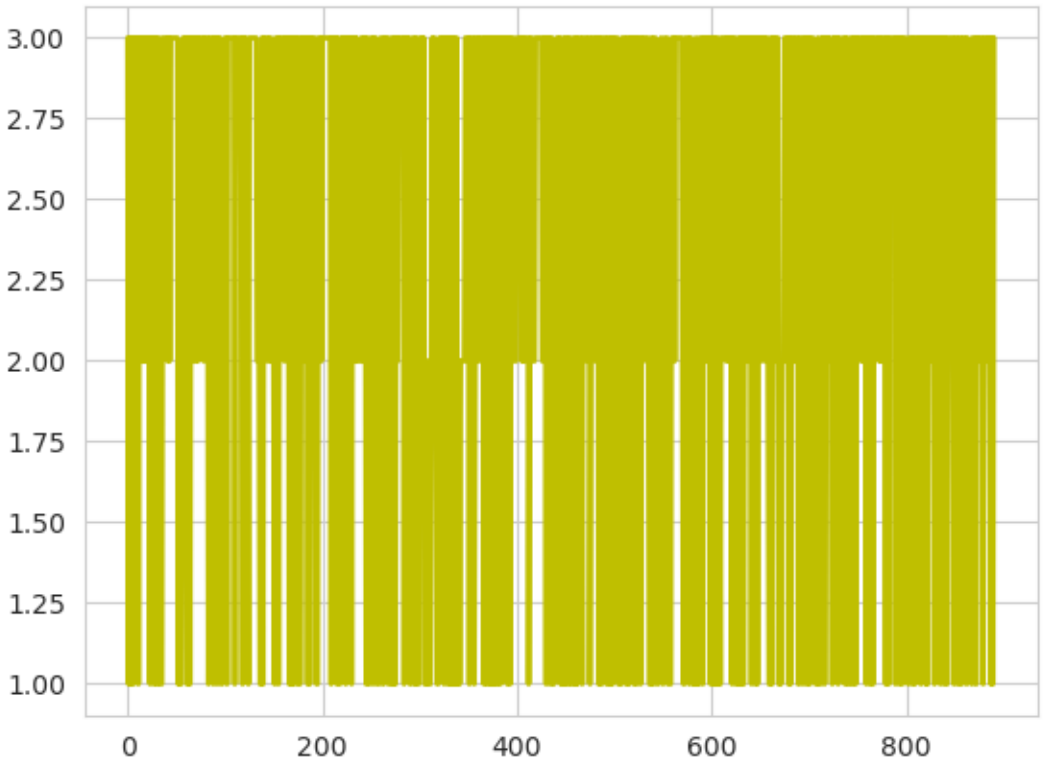
[41]:
```
plt.plot(titanic['Sex'],color = 'b',label = 'Survived', linewidth = 2)
```

[41]: [<matplotlib.lines.Line2D at 0x7f9b96d8a020>]



[42]:
```
plt.plot(titanic['Pclass'],color = 'y',label = 'Parch', linewidth = 2)
```

[42]: [<matplotlib.lines.Line2D at 0x7f9b96bd2fb0>]


3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0
200
400
600
800

[ ]:

[ ]: