*[8]:
```
# Reading the data:
train = pd.read_csv('titanic_train.csv')

train.head()
```

[8]:

	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

[9]:
```
# Exploratory Data Analysis: Basics of EDA include data cleaning and dealing with missing data and null values:
# We can use seaborn to create a simple heatmap to see where we are missing data.

train.isnull()
```

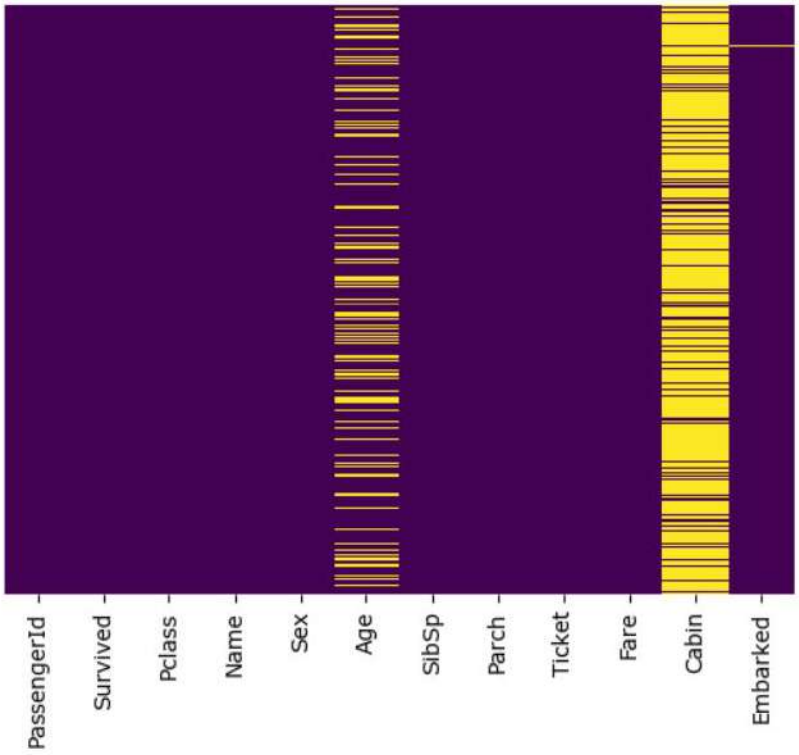
[9]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False	False	True	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	True	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	True	False

891 rows × 12 columns

[10]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

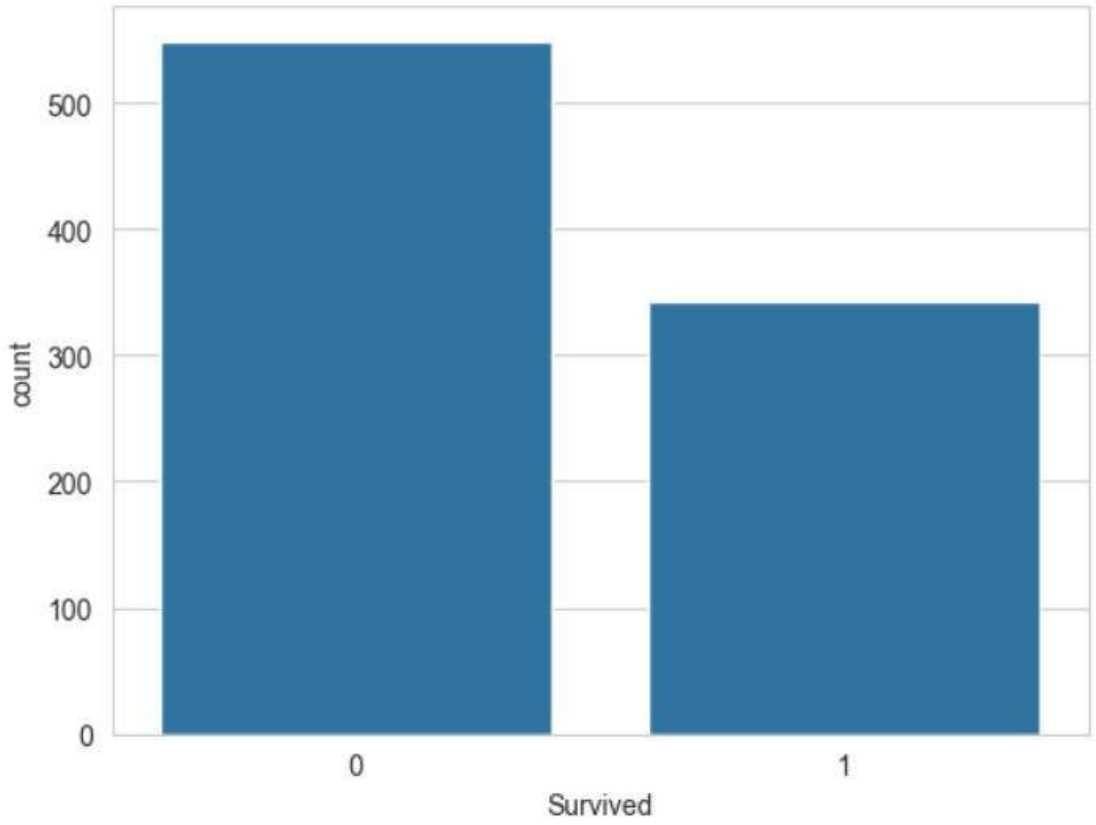
[10]:
```
<Axes: >
```



[11]:
```
''' Approximately 20% of the Age data is missing, which is a manageable amount for imputation.
In contrast, the Cabin column has a significant amount of missing data, making it unsuitable for direct use.
We'll likely either drop it or transform it into a binary feature indicating whether cabin information is present.
The following visualizations will provide further insights.'''
```

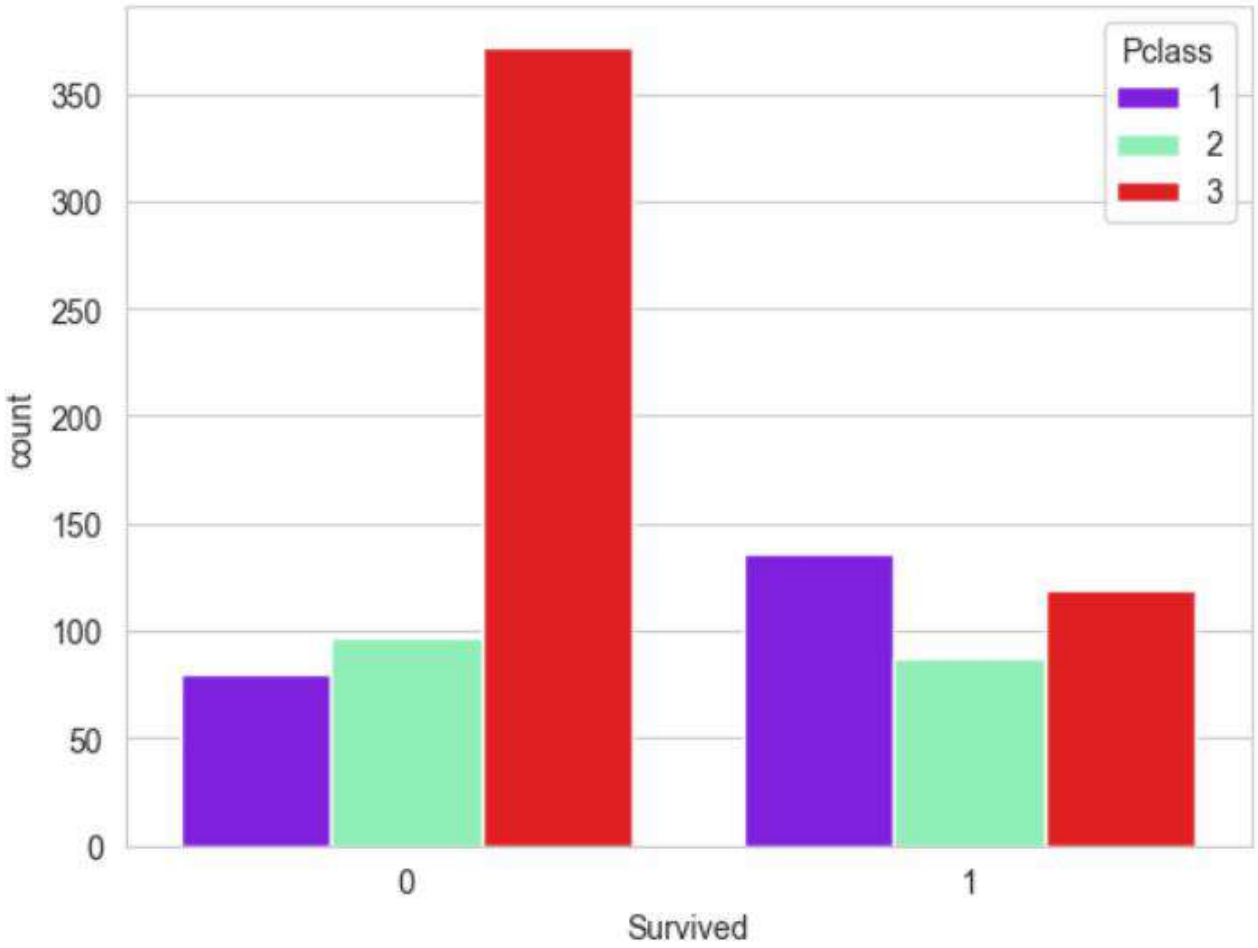
[17]:
```
sns.set_style('whitegrid')
sns.countplot(x='Survived',data=train)
```

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



[16]:
```
sns.set_style('whitegrid')
sns.countplot(x='Survived',hue='Pclass',data=train,palette='rainbow')
```

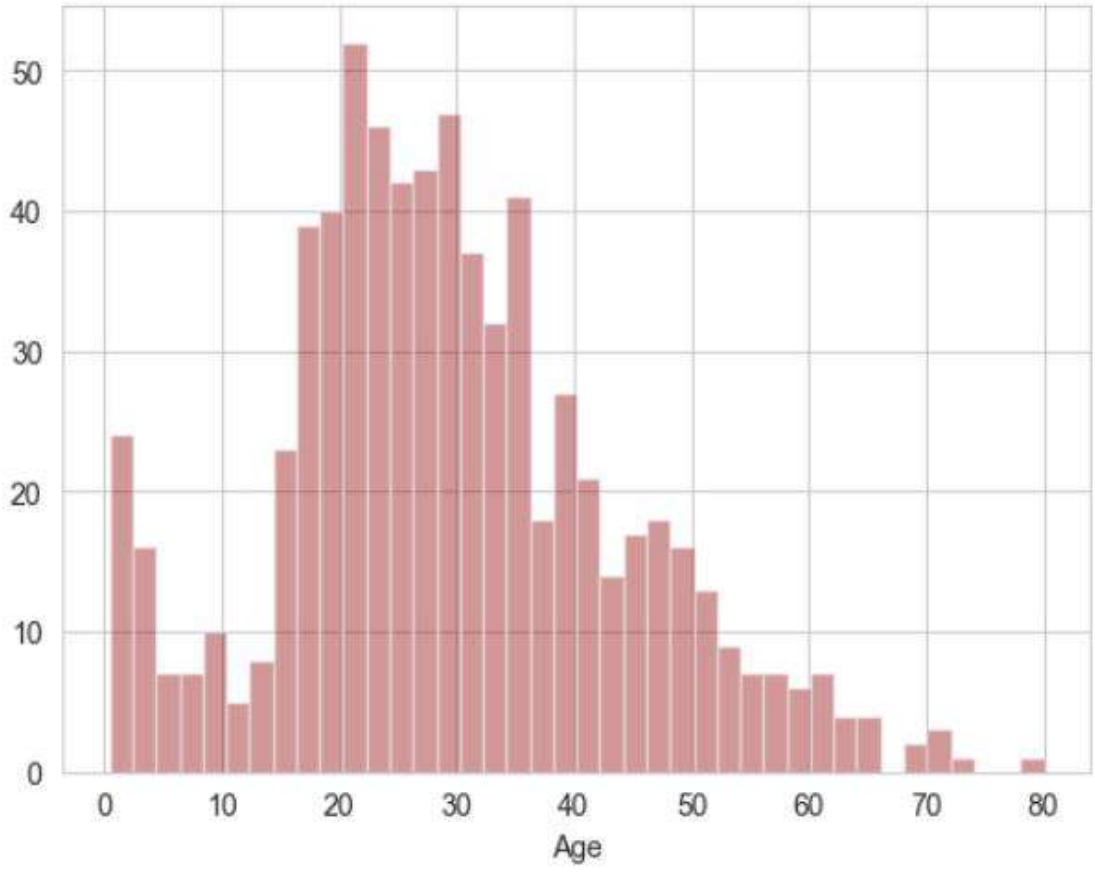
[16]:
```
<Axes: xlabel='Survived', ylabel='count'>
```



[18]:
```
sns.distplot(train['Age'].dropna(),kde=False,color='darkred',bins=40)
```

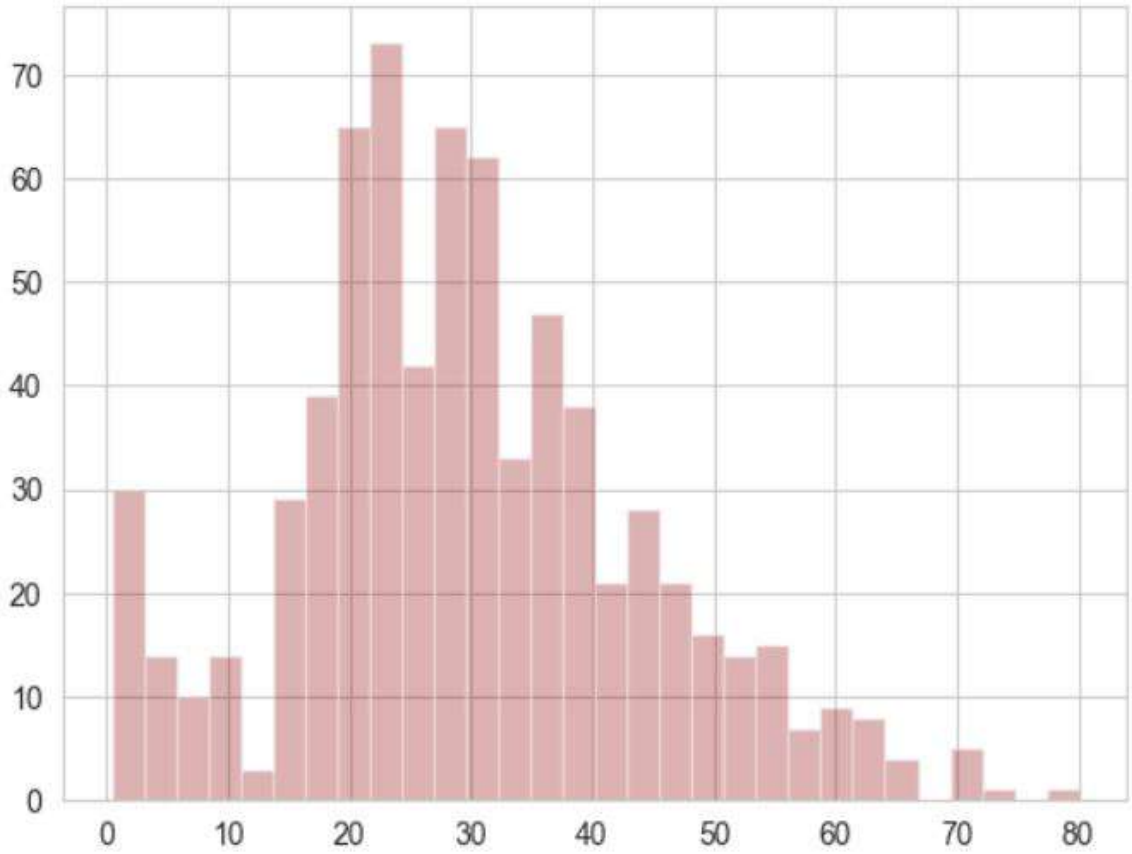
```
C:\Users\aksab\AppData\Local\Temp\ipykernel_5040\2002818437.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(train['Age'].dropna(),kde=False,color='darkred',bins=40)
```

[18]:
```
<Axes: xlabel='Age'>
```



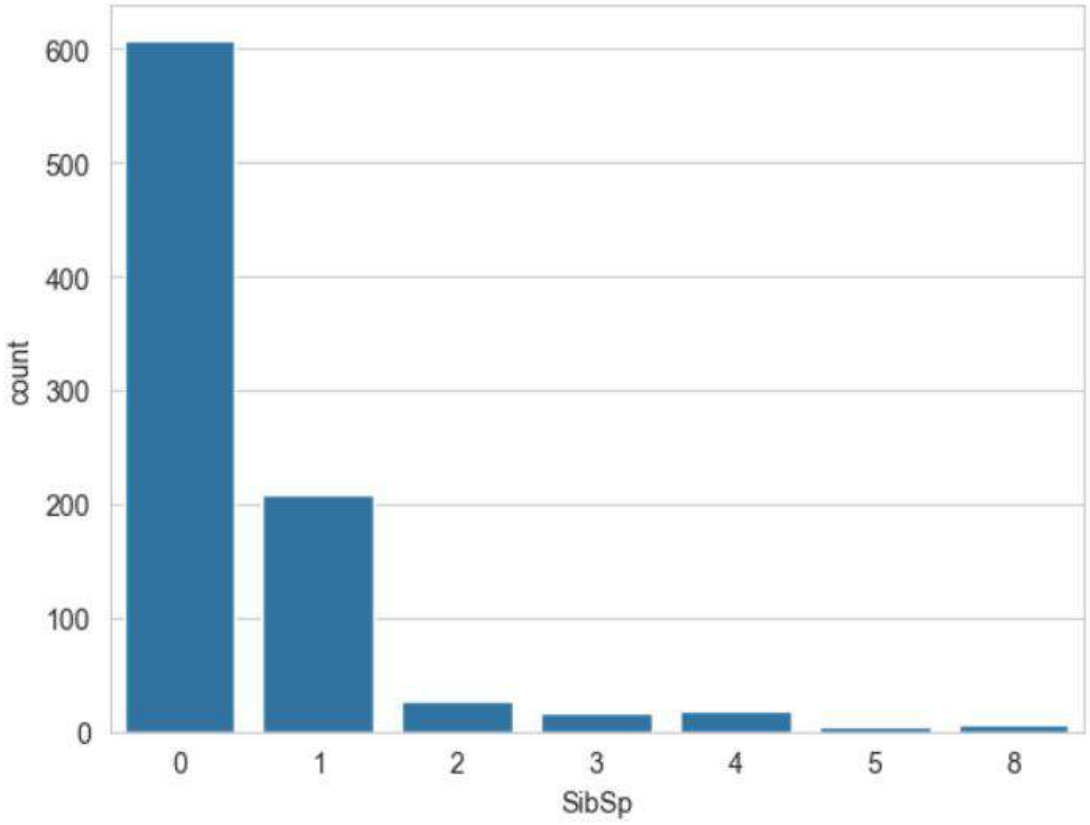
[19]:
```
train['Age'].hist(bins=30,color='darkred',alpha=0.3)
```

[19]: <Axes: >



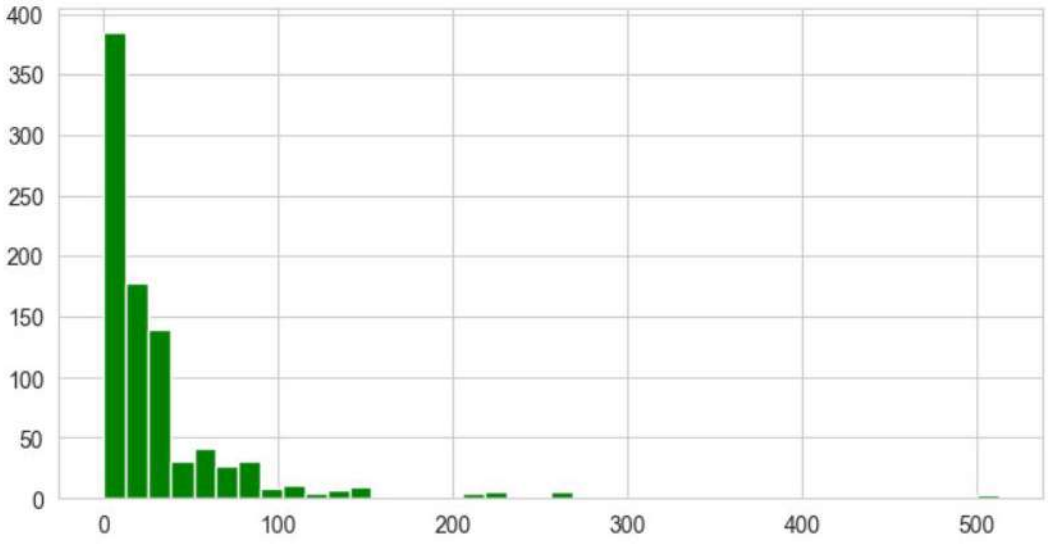
[20]:
```
sns.countplot(x='SibSp',data=train)
```

[20]:
```
<Axes: xlabel='SibSp', ylabel='count'>
```



[21]:
```
train['Fare'].hist(color='green',bins=40,figsize=(8,4))
```

[21]: <Axes: >

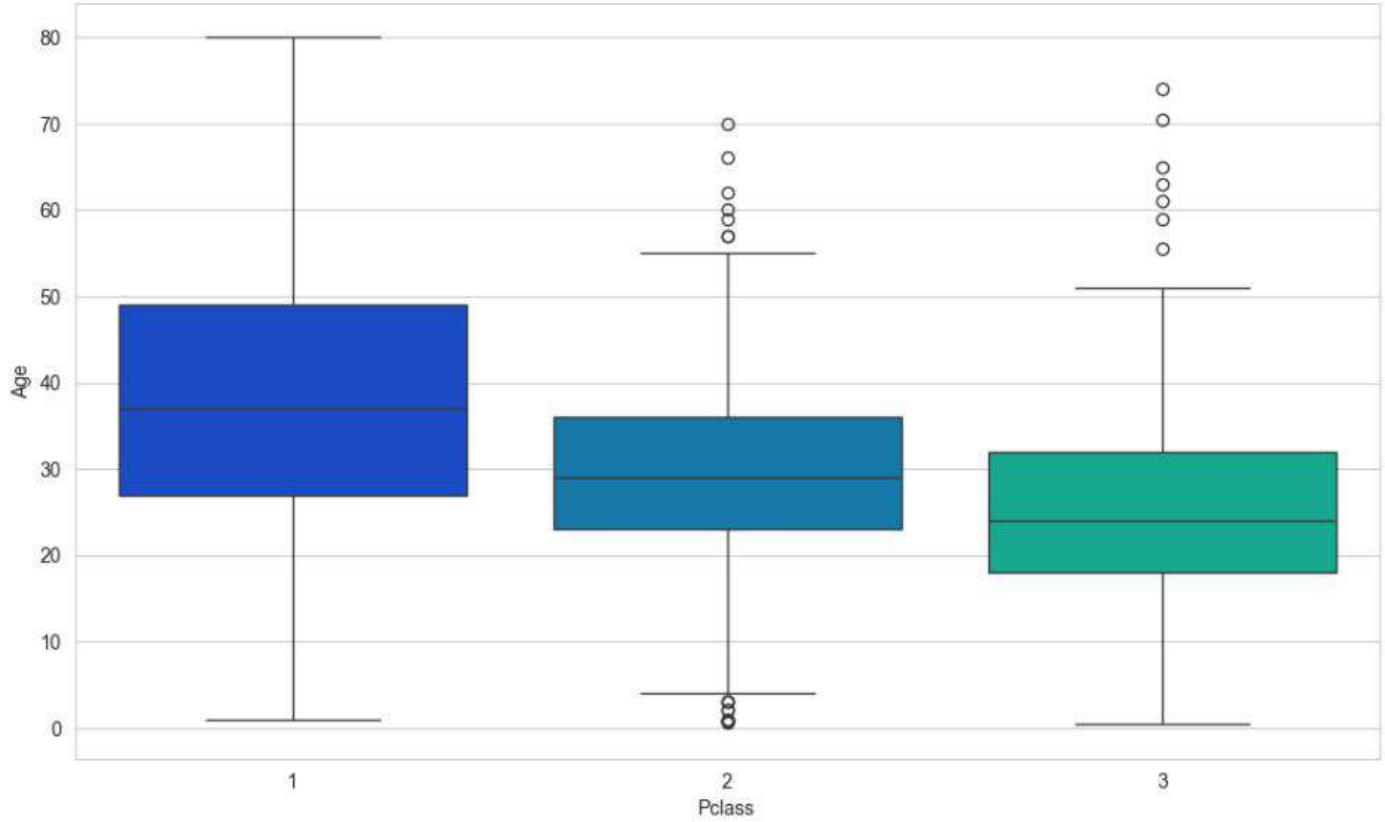


[34]:
```
' Data cleaning: There are two effective ways to handle the missing age data: we can either fill in the average age of all\npassengers or, for a mor
ccurate approach, we can impute based on the average age within each passenger class. For example: '
```

[26]:
```
plt.figure(figsize=(12, 7))
sns.boxplot(x='Pclass',y='Age',data=train,palette='winter')
```

```
C:\Users\aksab\AppData\Local\Temp\ipykernel_5040\1683851715.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for
same effect.

  sns.boxplot(x='Pclass',y='Age',data=train,palette='winter')
```

[26]:
```
<Axes: xlabel='Pclass', ylabel='Age'>
```



*[27]:
```
''' Wealthier passengers in the higher classes are generally older, which is a logical pattern.
We'll use these average age values, categorized by passenger class, to more accurately fill in the missing age data. '''
```

[27]:
```
''' Wealthier passengers in the higher classes are generally older, which is a logical pattern.
We'll use these average age values, categorized by passenger class, to more accurately fill in the missing age data. '''
```

[28]:
```
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
```

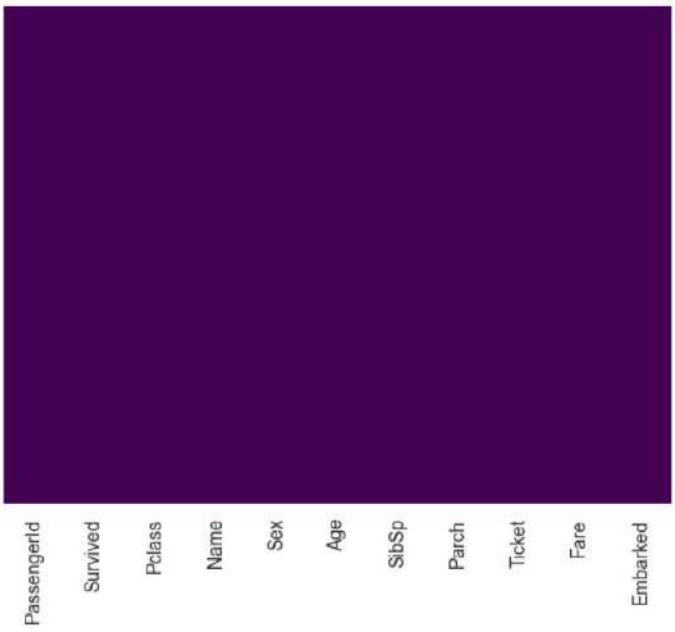
[29]:
```
train['Age'] = train[['Age','Pclass']].apply(impute_age,axis=1)
```

[ ]:
```
# After fixing the age column, we can drop the cabin column as it's not needed.
```

[31]:
```
train.drop('Cabin',axis=1,inplace=True)
```

[32]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

[32]:
```
<Axes: >
```



[33]:
```
# Now the data is ready and clean for further analysis

train.head()
```

[33]:

	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

[ ]: