TITANIC

November 13, 2024

Import Libraries

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib_inline
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

THE DATA

[4]:
```
train = pd.read_csv('train.csv')
```

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

[5]:

	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

Exploratory Data Analysis

[7]:
```
train.isnull()
```

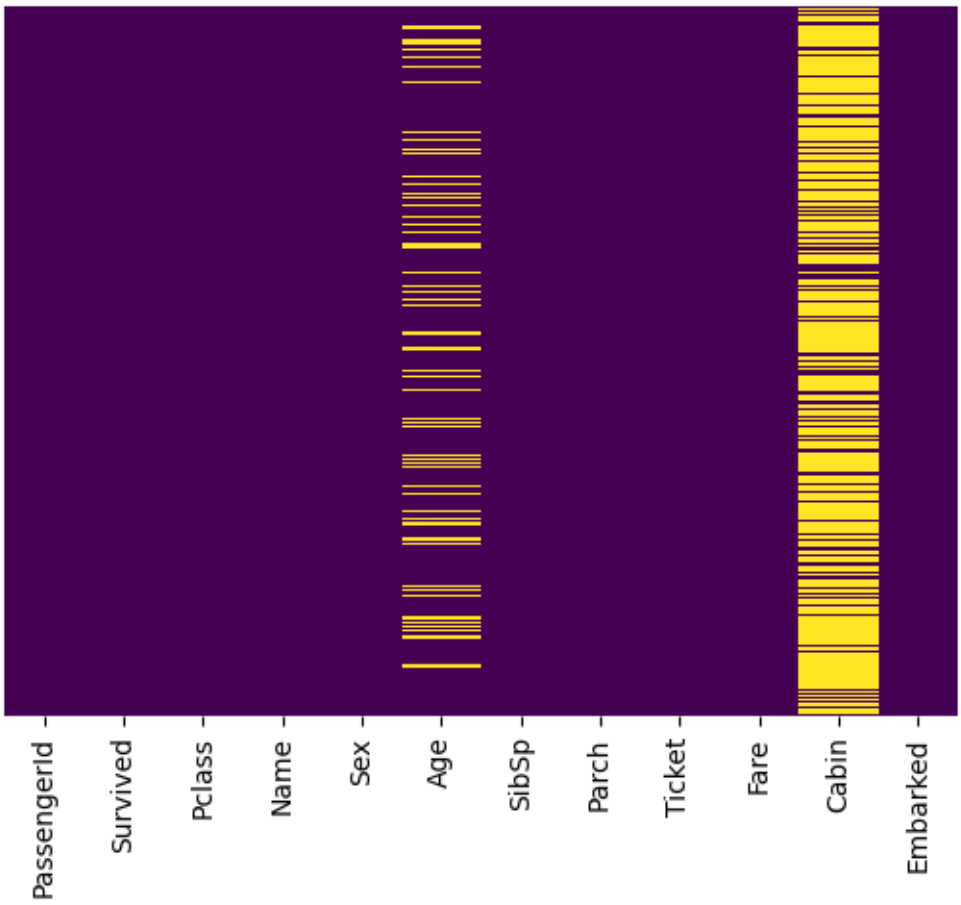
[7]:

	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 x 12 columns]

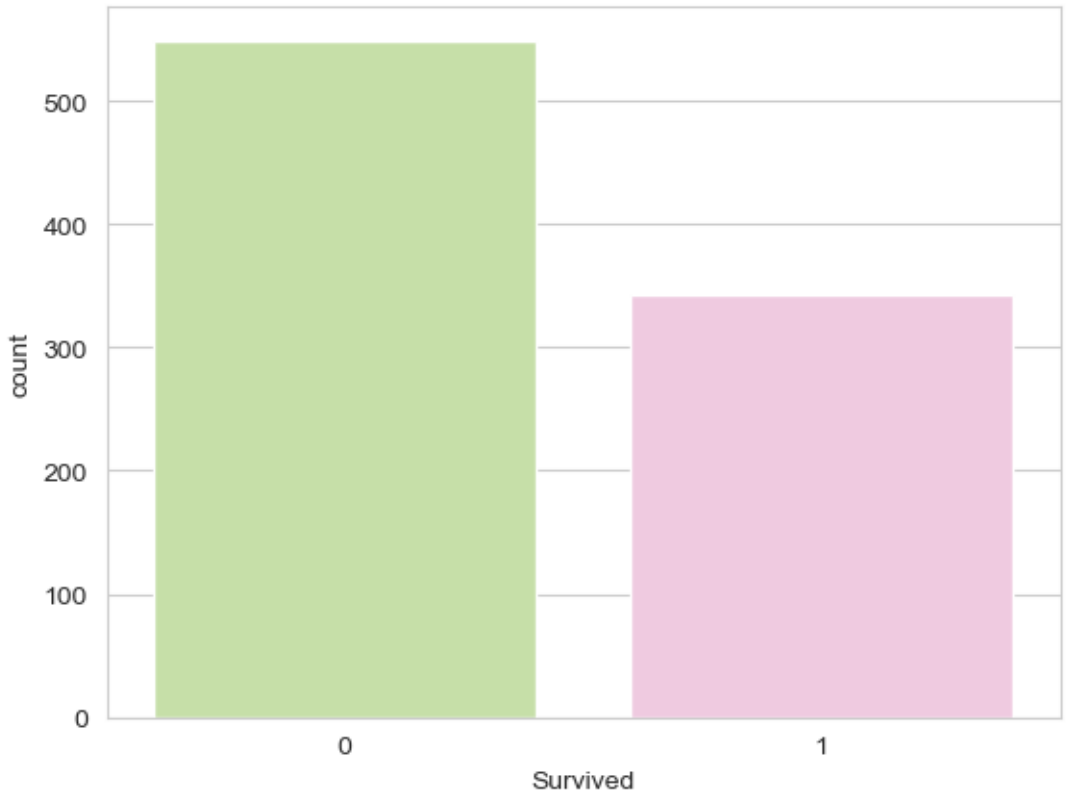
[8]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

[8]:
```
<Axes: >
```



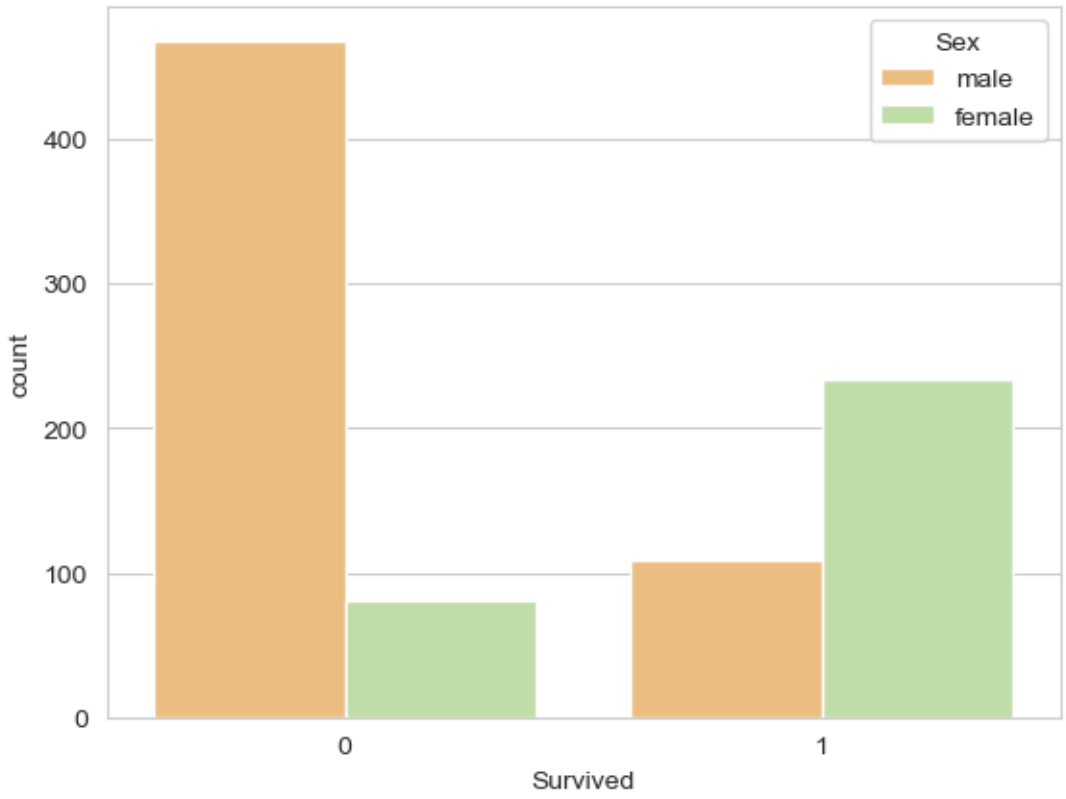
[9]:
```
sns.set_style('whitegrid')
sns.countplot(x='Survived', data=train, palette='PiYG_r')
```

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



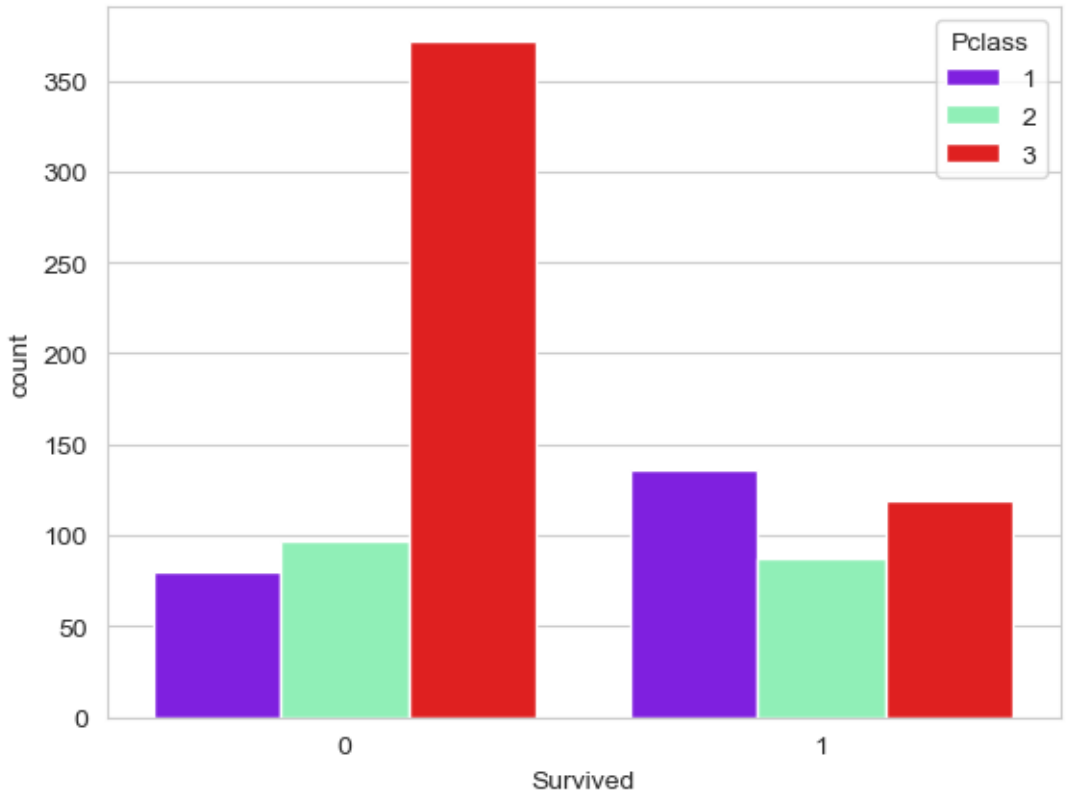
[10]:
```
sns.set_style('whitegrid')
sns.countplot(x='Survived', hue='Sex',data=train , palette='Spectral')
```

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



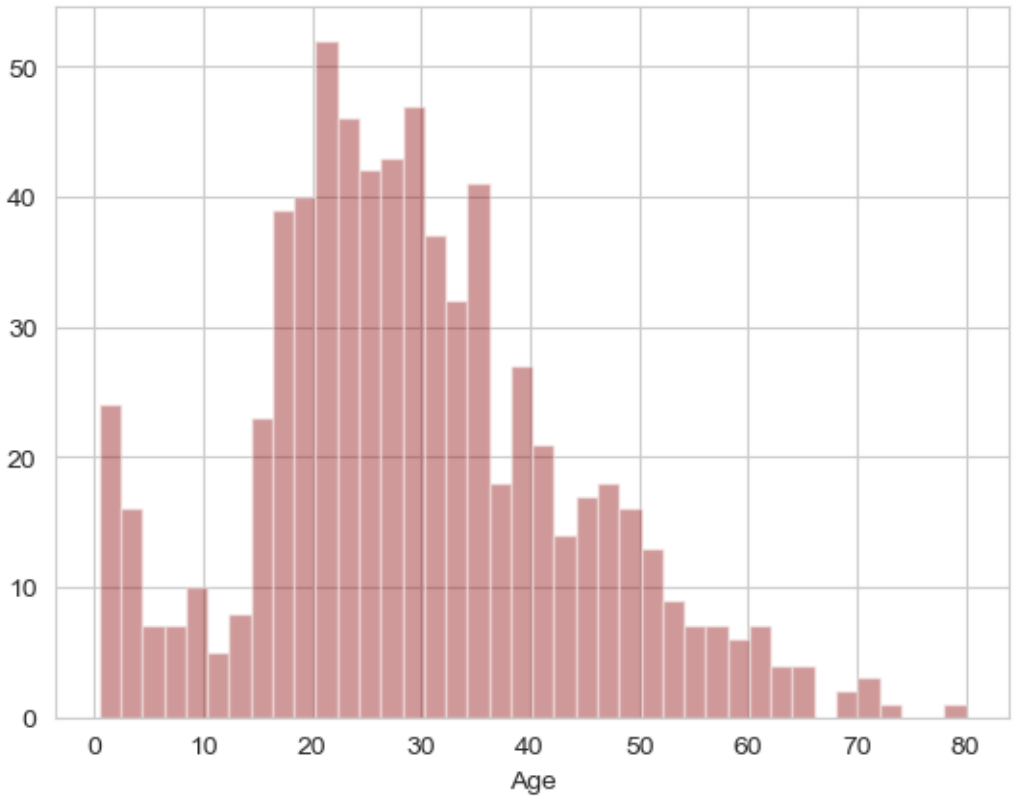
[11]:
```
sns.set_style('whitegrid')
sns.countplot(x='Survived', hue='Pclass',data=train , palette='rainbow')
```

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



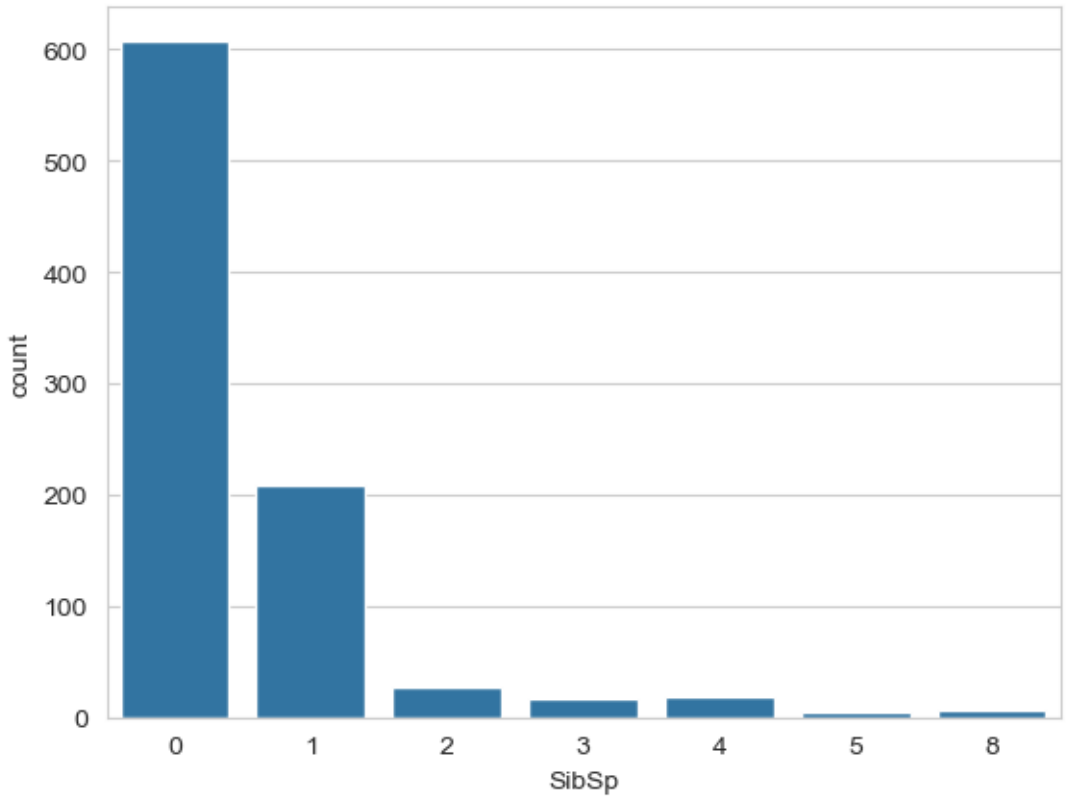
[12]:
```
sns.distplot(train['Age'].dropna(),kde=False, color='darkred', bins=40)
```

[12]: <Axes: xlabel='Age'>



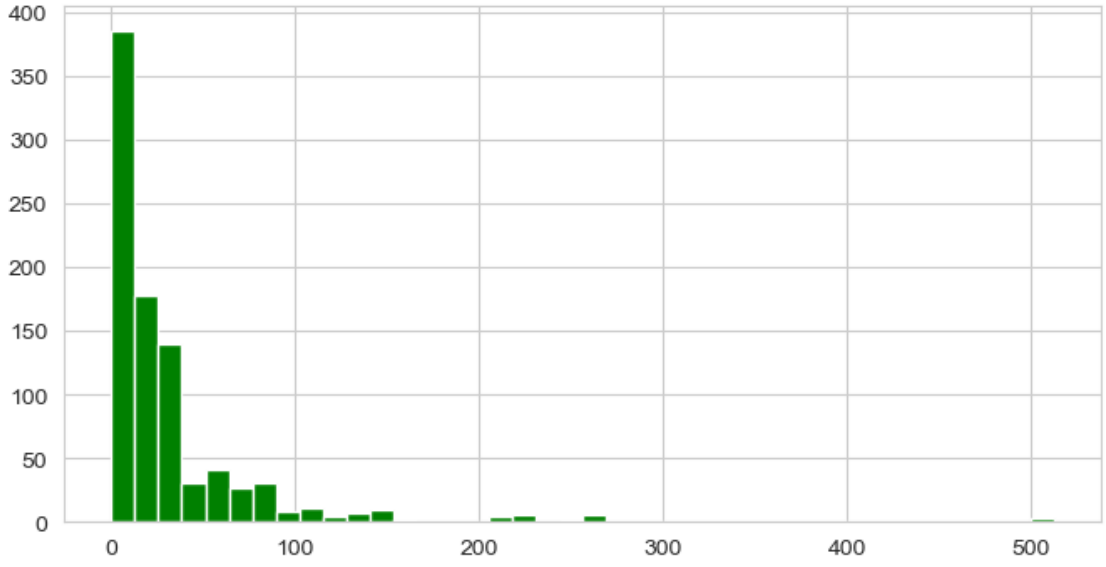
[13]:
```
sns.countplot(x='SibSp',data=train)
```

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



[14]:
```
train['Fare'].hist(color='green',bins=40,figsize=(8,4))
```

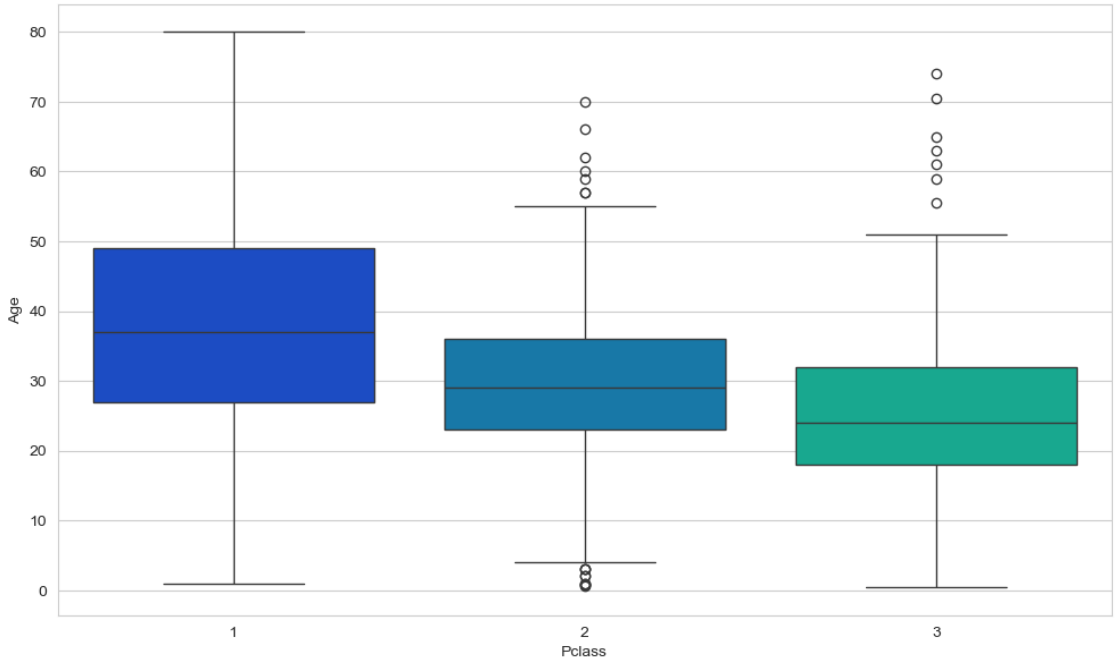
[14]: <Axes: >



DATA CLEANING

[16]:
```
plt.figure(figsize=(12,7))
sns.boxplot(x='Pclass',y='Age',data=train,palette='winter')
```

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

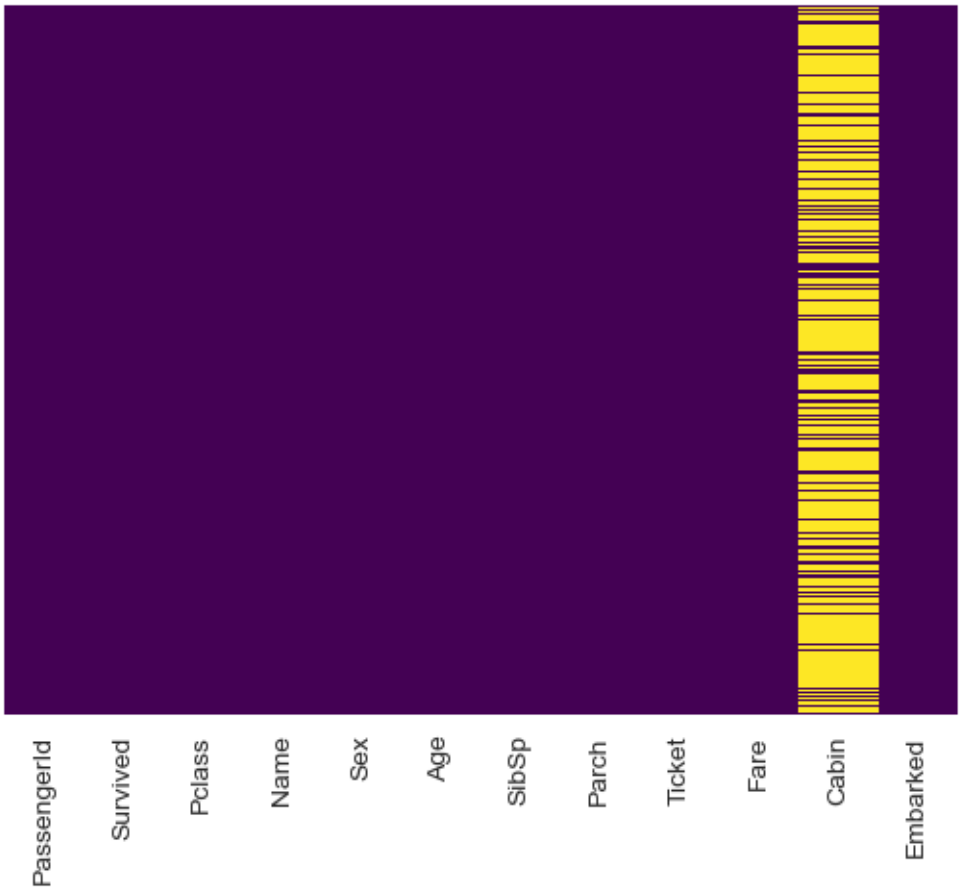


[17]:
```
def impute_age(cols):
    Age = cols[0]
    Pclass = cols[1]

    if pd.isnull(Age):

        if Pclass == 1:
            return 37

        elif Pclass == 2:
            return 29

        else:
            return 24

    else:
        return Age
```

[18]:
```
train['Age'] = train[['Age','Pclass']].apply(impute_age,axis=1)
```

[19]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

[19]: <Axes: >



[20]:
```
train.drop('Cabin',axis=1,inplace=True)
```

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

[21]:

	PassengerId	Survived	Pclass	\
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	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

   Parch            Ticket     Fare Embarked
0      0         A/5 21171   7.2500        S
1      0          PC 17599  71.2833        C
2      0  STON/O2. 3101282   7.9250        S
3      0            113803  53.1000        S
4      0            373450   8.0500        S
```

[22]:
```
train.dropna(inplace=True)
```

CONVERTING CATEGORICAL FEATURES

[24]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 889 entries, 0 to 890
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  889 non-null    int64
 1   Survived     889 non-null    int64
 2   Pclass       889 non-null    int64
 3   Name         889 non-null    object
 4   Sex          889 non-null    object
 5   Age          889 non-null    float64
 6   SibSp        889 non-null    int64
 7   Parch        889 non-null    int64
 8   Ticket       889 non-null    object
 9   Fare         889 non-null    float64
 10  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(4)
memory usage: 83.3+ KB
```

[25]:
```
pd.get_dummies(train['Embarked'],drop_first=True).head()
```

[25]:
```
       Q      S
0  False   True
1  False  False
2  False   True
3  False   True
4  False   True
```

[26]:
```
sex = pd.get_dummies(train['Sex'],drop_first=True)
embark = pd.get_dummies(train['Embarked'],drop_first=True)
```

[27]:
```
train.drop(['Sex','Embarked','Name','Ticket'],axis=1,inplace=True)
```

[28]:
```
train.head()
```

[28]:

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

[29]:
```
train = pd.concat([train,sex,embark],axis=1)
```

[30]:
```
train.head()
```

[30]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	male	Q	\
0	1	0	3	22.0	1	0	7.2500	True	False	
1	2	1	1	38.0	1	0	71.2833	False	False	
2	3	1	3	26.0	0	0	7.9250	False	False	
3	4	1	1	35.0	1	0	53.1000	False	False	
4	5	0	3	35.0	0	0	8.0500	True	False	

	S
0	True
1	False
2	True
3	True
4	True

Building a Logistic Regression model

[32]:
```
train.drop('Survived',axis=1).head()
```

[32]:

	PassengerId	Pclass	Age	SibSp	Parch	Fare	male	Q	S
0	1	3	22.0	1	0	7.2500	True	False	True
1	2	1	38.0	1	0	71.2833	False	False	False
2	3	3	26.0	0	0	7.9250	False	False	True
3	4	1	35.0	1	0	53.1000	False	False	True
4	5	3	35.0	0	0	8.0500	True	False	True

[33]:
```
train['Survived'].head()
```

[33]:
```
0    0
1    1
2    1
3    1
4    0
```

```
Name: Survived, dtype: int64
```

Training and Predicting

[35]:
```
from sklearn.model_selection import train_test_split
```

[36]:
```
X_train, X_test, y_train, y_test = train_test_split(train.
↪drop('Survived',axis=1),
                                                    train['Survived'],␣
↪test_size=0.30,
                                                    random_state=101)
```

[37]:
```
from sklearn.linear_model import LogisticRegression
```

[38]:
```
logmodel = LogisticRegression()
logmodel.fit(X_train,y_train)
```

[38]:
```
LogisticRegression()
```

[39]:
```
predictions = logmodel.predict(X_test)
```

[40]:
```
from sklearn.metrics import confusion_matrix
```

[41]:
```
accuracy=confusion_matrix(y_test,predictions)
```

[42]:
```
accuracy
```

[42]:
```
array([[149,  14],
       [ 39,  65]], dtype=int64)
```

[43]:
```
from sklearn.metrics import accuracy_score
```

[44]:
```
accuracy=accuracy_score(y_test,predictions)
accuracy
```

[44]:
```
0.8014981273408239
```

[45]:
```
predictions
```

[45]:
```
array([0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1,
       1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,
       0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0,
       0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1,
       0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0,
```

```
0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0,
0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0,
1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0,
0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0,
0, 1, 1], dtype=int64)
```

Evaluation

[47]:
```
from sklearn.metrics import classification_report
```

[48]:
```
print(classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

           0       0.79      0.91      0.85       163
           1       0.82      0.62      0.71       104

    accuracy                           0.80       267
   macro avg       0.81      0.77      0.78       267
weighted avg       0.80      0.80      0.80       267
```

[ ]:

[ ]:

[ ]:

[ ]: