Handle missing values

August 17, 2024

```
[13]: ## import the required library
      import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt
      import seaborn as sns
```

```
[3]: data = pd.read_csv("titanic.csv")
```

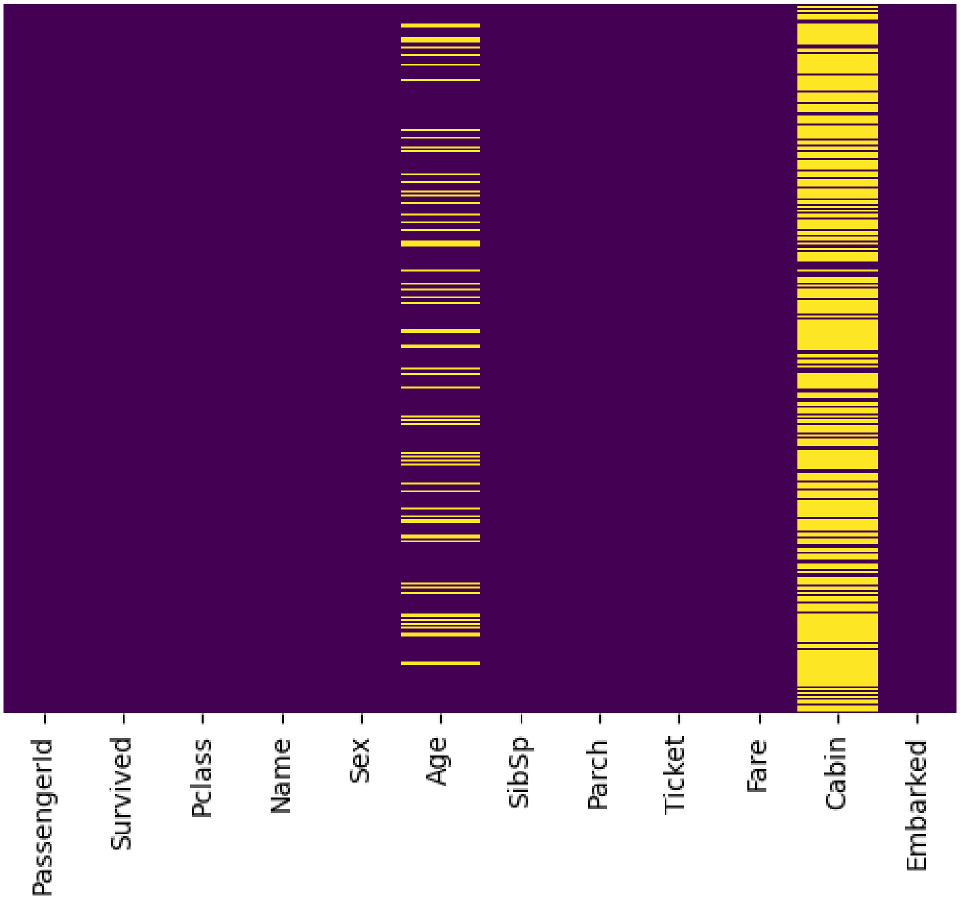
```
[4]: data.head()
```

```
[4]:    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 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
```

1 Exploratory data analysis

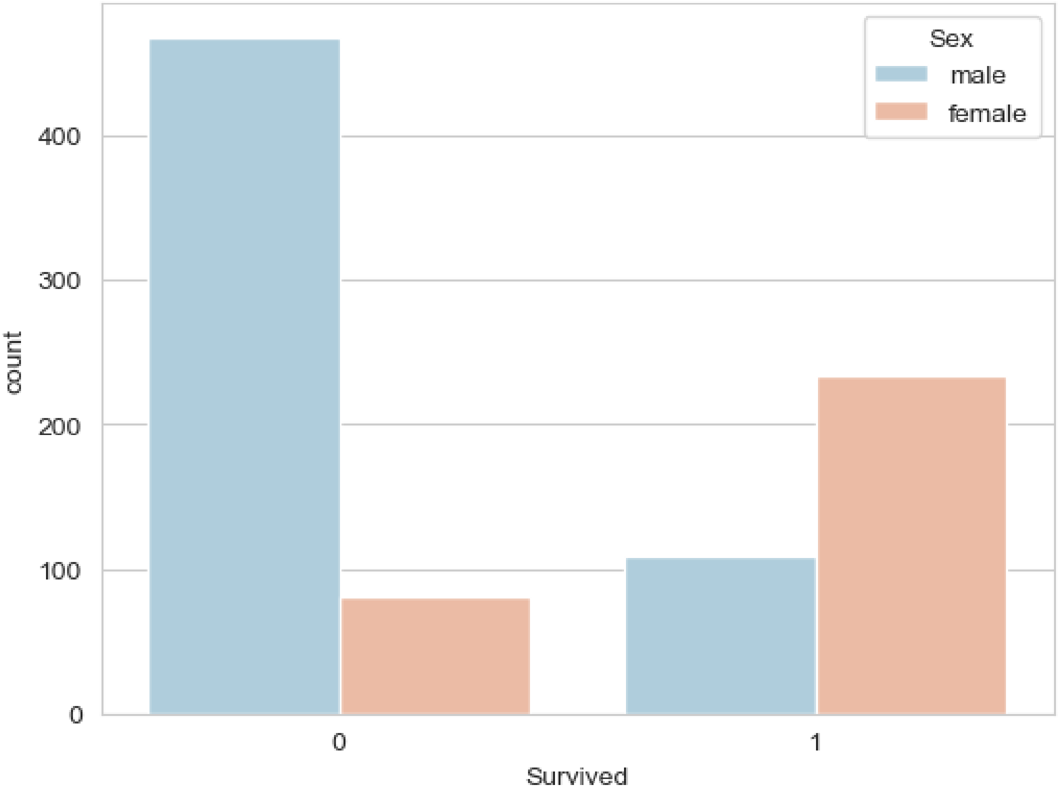
```
[6]: sns.heatmap(data.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

```
[6]: <Axes: >
```



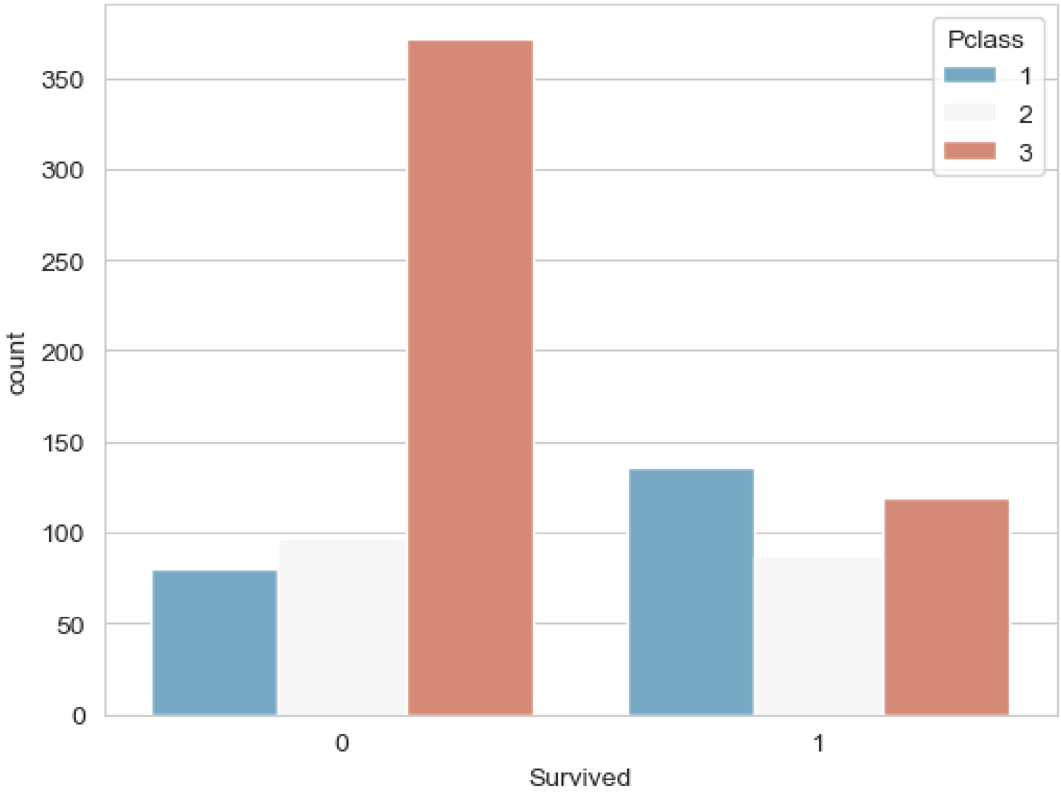
```
[9]: sns.set_style('whitegrid')
     sns.countplot(x='Survived',hue='Sex',data=data,palette='RdBu_r')
```

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



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

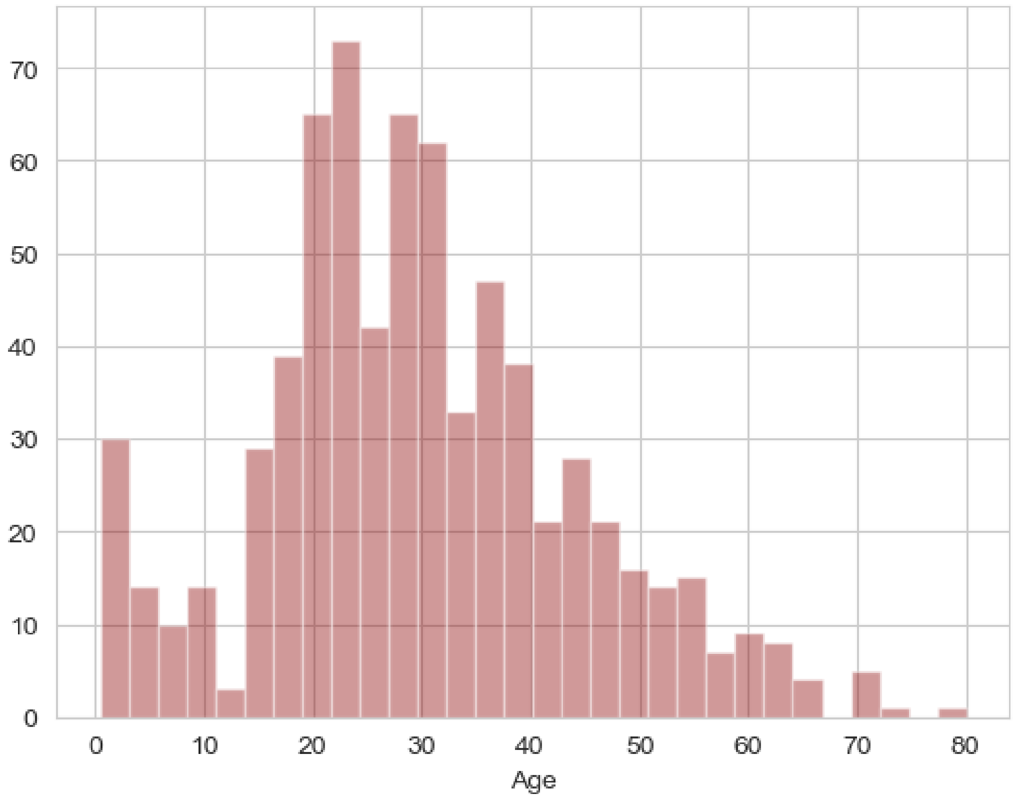
```
[10]: <Axes: xlabel='Survived', ylabel='count'>
```



[14]:

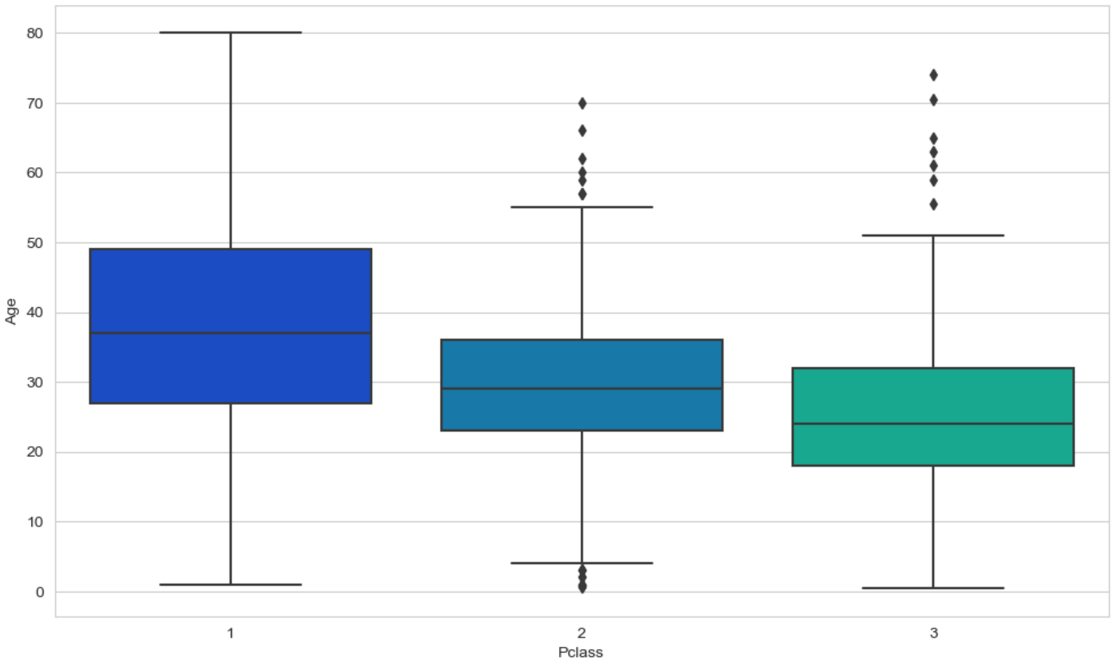
```
sns.distplot(data['Age'].dropna(),kde=False ,color='darkred',bins=30)
plt.show()
```

```
C:\Users\Vikas\AppData\Local\Temp\ipykernel_26228\640333185.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(data['Age'].dropna(),kde=False ,color='darkred',bins=30)
```



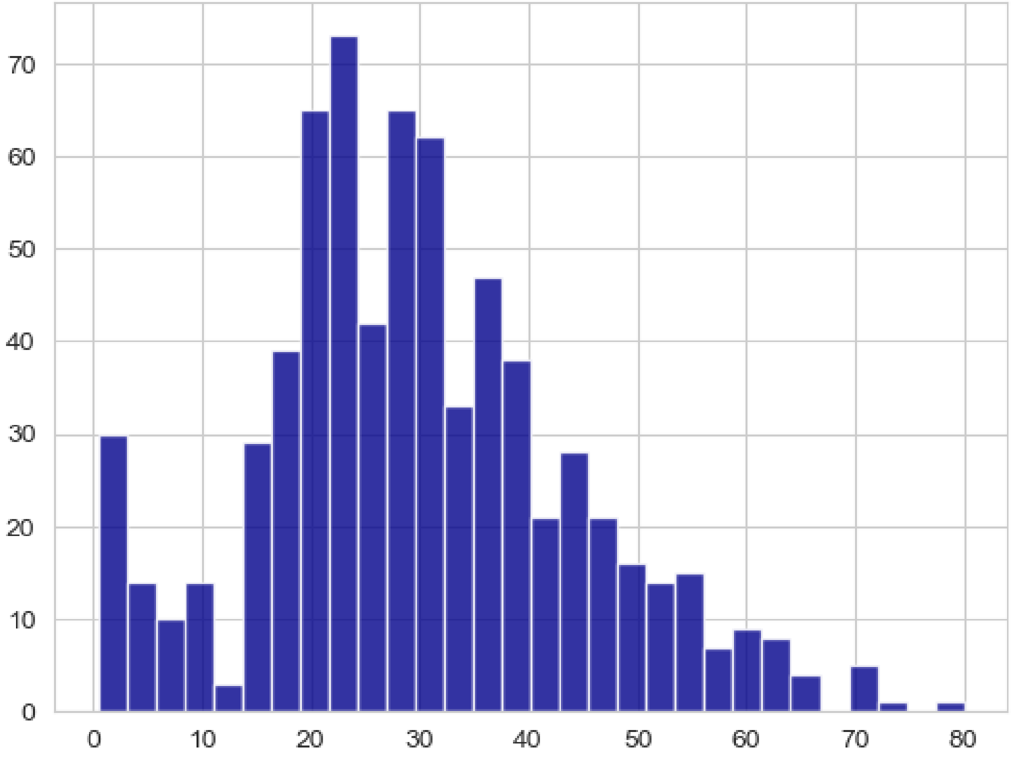
[15]:
```
plt.figure(figsize=(12,7))
sns.boxplot(x='Pclass',y='Age',data=data,palette='winter')
```

[15]:
```
<Axes: xlabel='Pclass', ylabel='Age'>
```



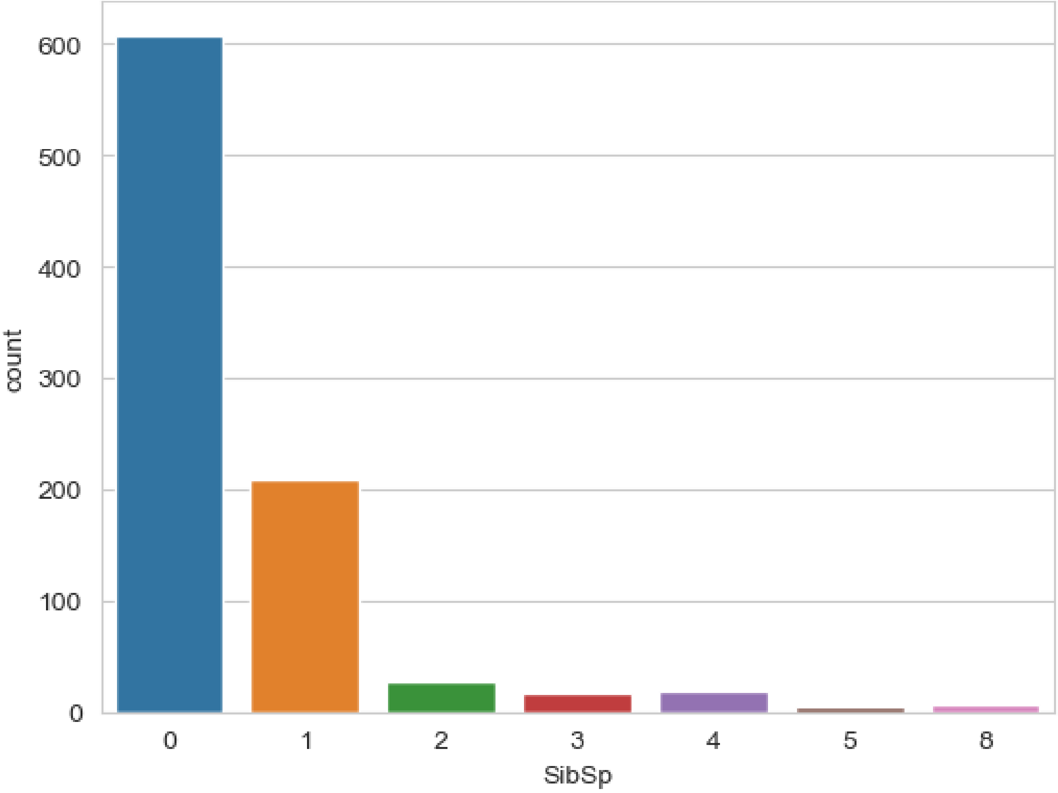
```
[26]: data['Age'].hist(bins=30,color='darkblue',alpha=0.8)
```

```
[26]: <Axes: >
```



[30]:
```
sns.countplot(x='SibSp',data=data)
```

[30]:
```
<Axes: xlabel='SibSp', ylabel='count'>
```



[38]:
```
# Define the impute_age function
def impute_age(row):
    Age = row['Age']
    Pclass = row['Pclass']

    # Impute Age based on Pclass
    if pd.isnull(Age):
        if Pclass == 1:
            return 37  # Example value for Pclass 1
        elif Pclass == 2:
            return 29  # Example value for Pclass 2
        else:
            return 24  # Example value for Pclass 3
    else:
        return Age

# Apply the impute_age function row by row
data['Age'] = data.apply(impute_age, axis=1)

print(data)
```

```
     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  24.0      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]
```

[39]:

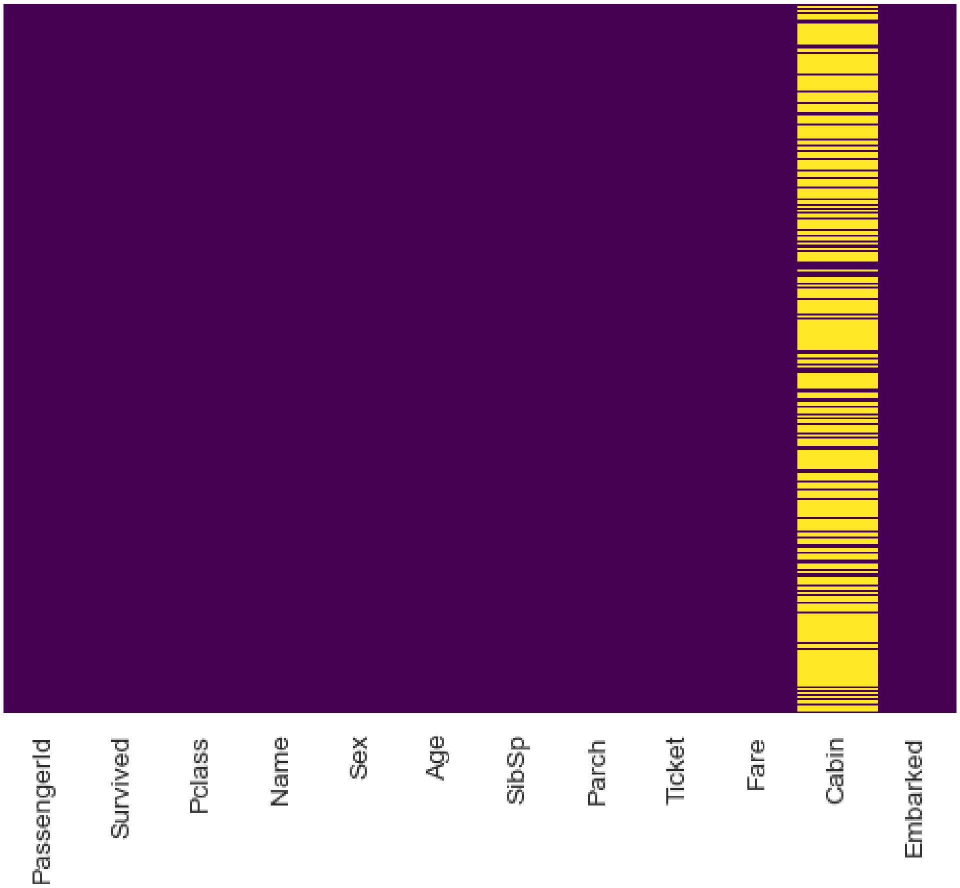
```
data['Age']=data[['Age','Pclass']].apply(impute_age,axis=1)
```

[40]:

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

[40]:

```
<Axes: >
```



[43]:

[44]:

[51]:

[52]:

[56]:

```
tuple()
```

[56]:

```
()
```

[55]:

```
tuple
```

[55]:

```
tuple
```

[ ]:

```
```