In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
train = pd.read_csv(r"C:\Users\laksh\OneDrive\Desktop\titanic_train.csv")
```

In [3]:
```
train.head()
```

Out[3]:

	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

EXPLORITARY DATA ANALYSIS

find MISSING DATA

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

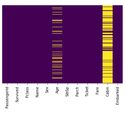
Out[4]:

	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

In [5]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

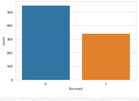
Out[5]:
<AxesSubplot:>



In [6]:
```
sns.set_style('whitegrid')
```

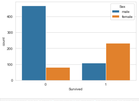
In [7]:
```
sns.countplot(x='Survived',data=train)
```

Out[7]:
<AxesSubplot:xlabel='Survived', ylabel='count'>



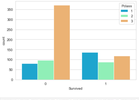
In [8]:
```
sns.countplot(x='Survived',hue='Sex',data=train)
```

Out[8]:
<AxesSubplot:xlabel='Survived', ylabel='count'>



In [9]:
```
sns.countplot(x='Survived',hue='Pclass',data=train,palette='rainbow')
```

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



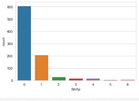
In [10]:
```
sns.distplot(train['Age'].dropna(),kde=False,color='darkblue',bins=40)
```

```
C:\Users\laksh\Anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[10]:
<AxesSubplot:xlabel='Age'>

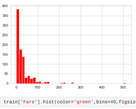
In [11]:
```
sns.countplot(x='SibSp',data=train)
```

Out[11]:
<AxesSubplot:xlabel='SibSp', ylabel='count'>



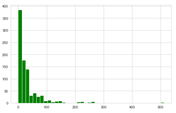
In [12]:
```
train['Fare'].hist(color='red',bins=40,)
```

Out[12]:
<AxesSubplot:>



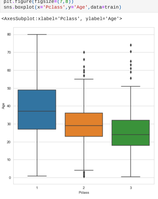
In [13]:
```
train['Fare'].hist(color='green',bins=40,figsize=(8,5))
```

Out[13]:
<AxesSubplot:>



In [14]:
```
plt.figure(figsize=(7,8))
sns.boxplot(x='Pclass',y='Age',data=train)
```

Out[14]:
<AxesSubplot:xlabel='Pclass', ylabel='Age'>



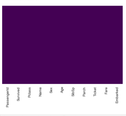
DATA CLEANING

In [45]:
```
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 [46]:
```
train['Age']=train[['Age','Pclass']].apply(impute_age,axis=1)
```

In [64]:
```
sns.heatmap(train.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

Out[64]:
<AxesSubplot:>



In [63]:
```
train.dropna(inplace=True)
```

In [66]:
```
train.head()
```

Out[66]:

	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

CATEGORIAL CALSSIFICATION

In [70]:
```
embark=pd.get_dummies(train['Embarked'],drop_first=True)
```

In [71]:
```
embark.head()
```

Out[71]:

	Q	S
0	0	1
1	0	0
2	0	1
3	0	1
4	0	1

In [72]:
```
sex=pd.get_dummies(train['Sex'],drop_first=True)
```

In [74]:
```
sex.head()
```

Out[74]:

	male
0	1
1	0
2	0
3	0
4	1

In [77]:
```
train.drop(['Sex','Embarked','Name','Ticket'],axis=1,inplace=True)
```

In [78]:
```
train.head()
```

Out[78]:

	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

In [82]:
```
pd.concat([train,sex,embark],axis=1)
```

Out[82]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	male	Q	S
0	1	0	3	22.0	1	0	7.2500	1	0	1
1	2	1	1	38.0	1	0	71.2833	0	0	0
2	3	1	3	26.0	0	0	7.9250	0	0	1
3	4	1	1	35.0	1	0	53.1000	0	0	1
4	5	0	3	35.0	0	0	8.0500	1	0	1
...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	27.0	0	0	13.0000	1	0	1
887	888	1	1	19.0	0	0	30.0000	0	0	1
888	889	0	3	24.0	1	2	23.4500	0	0	1
889	890	1	1	26.0	0	0	30.0000	1	0	0
890	891	0	3	32.0	0	0	7.7500	1	1	0

889 rows × 10 columns

In [1]:
```
jupyter nbconvert --to html notebook.ipynb
```

```
  File "<ipython-input-1-8f3604e61458c>", line 1
    jupyter nbconvert --to html notebook.ipynb
                  ^
SyntaxError: invalid syntax
```

In [ ]: