EDA with Python onTitanic Data

In [204…
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
```

The Data

In [205…
```
df = pd.read_csv(r"C:\Users\ASUS\Downloads\titanic.csv")
```

In [206…
```
df
```

Out[206]:

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [207…
```
df.shape
```

Out[207]:
```
(891, 12)
```

In [208…
```
df.dtypes ==  "object"
```

Out[208]:
```
PassengerId    False
Survived       False
Pclass         False
Name            True
Sex             True
Age            False
SibSp          False
Parch          False
Ticket          True
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

In [209…
```
df.head(10)
```

Out[209]:

	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
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

In [210…
```
df.head()
```

Out[210]:

	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

In [211…
```
df.tail()
```

Out[211]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

In [212…
```
df.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
```

Exploratory Data Analysis

Let's begin some exploratory data analysis we'll start by checking out missing data!

Missing Data

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

In [213…
```
df.isnull()
```

Out[213]:

	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

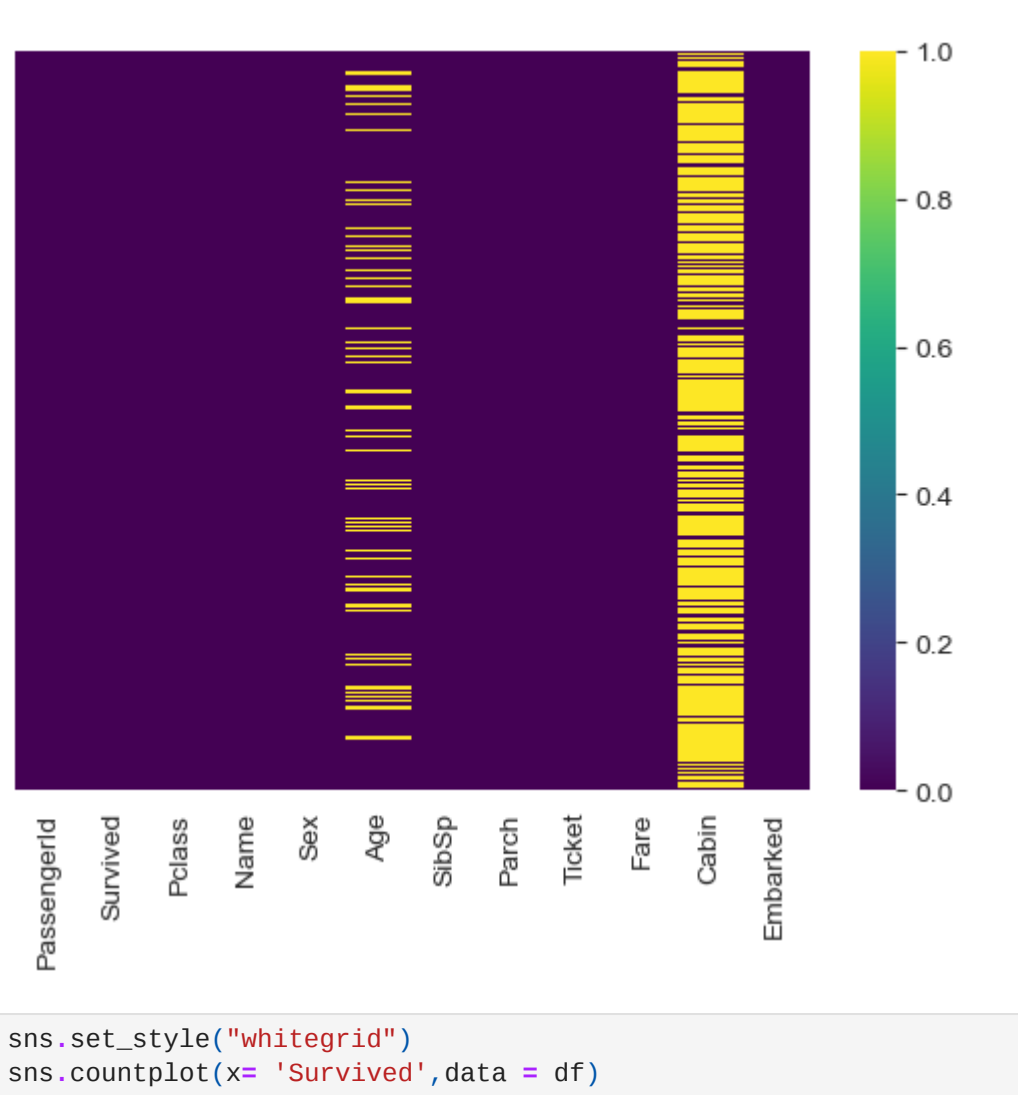
891 rows × 12 columns

In [214…
```
df.isnull().sum()
```

Out[214]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

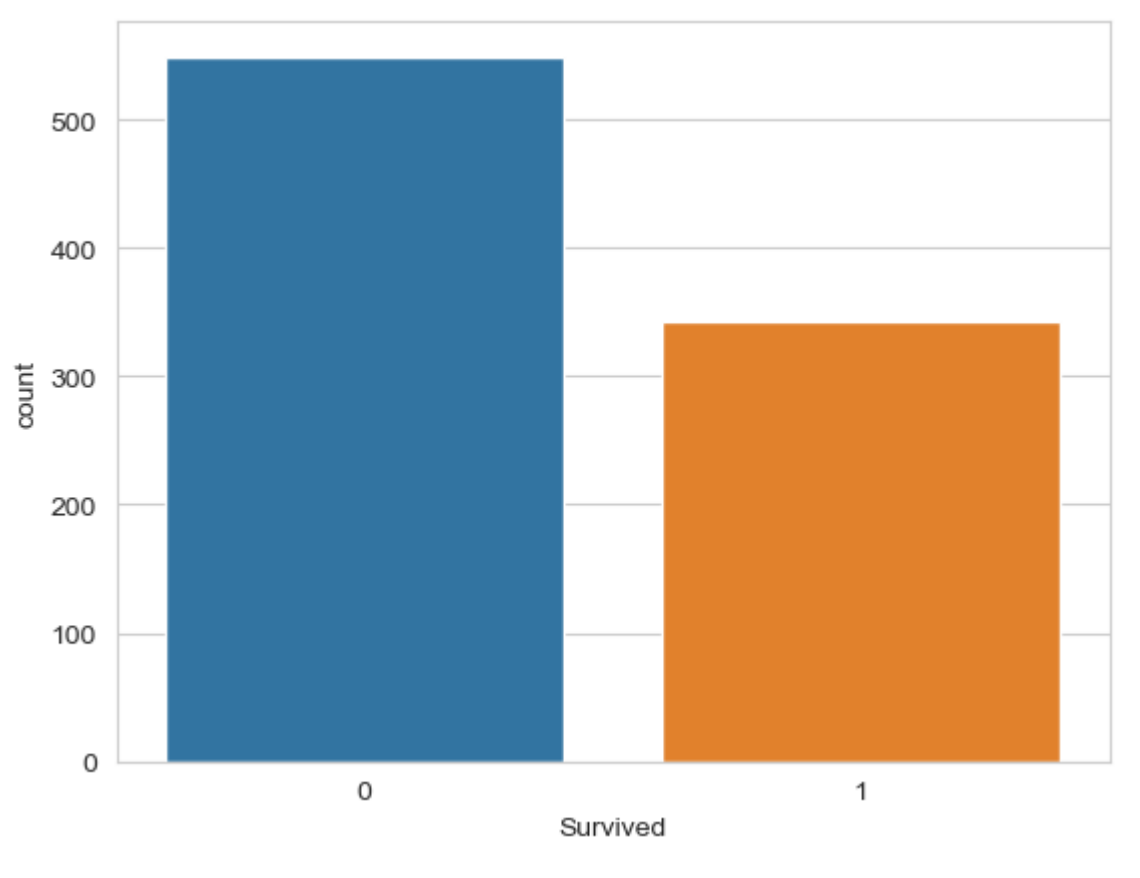
In [215…
```
sns.heatmap(df.isnull(),yticklabels=False,cmap ='viridis')
```

Out[215]:
```
<Axes: >
```



In [216…
```
sns.set_style("whitegrid")
sns.countplot(x= 'Survived',data = df)
```

Out[216]:
```
<Axes: xlabel='Survived', ylabel='count'>
```



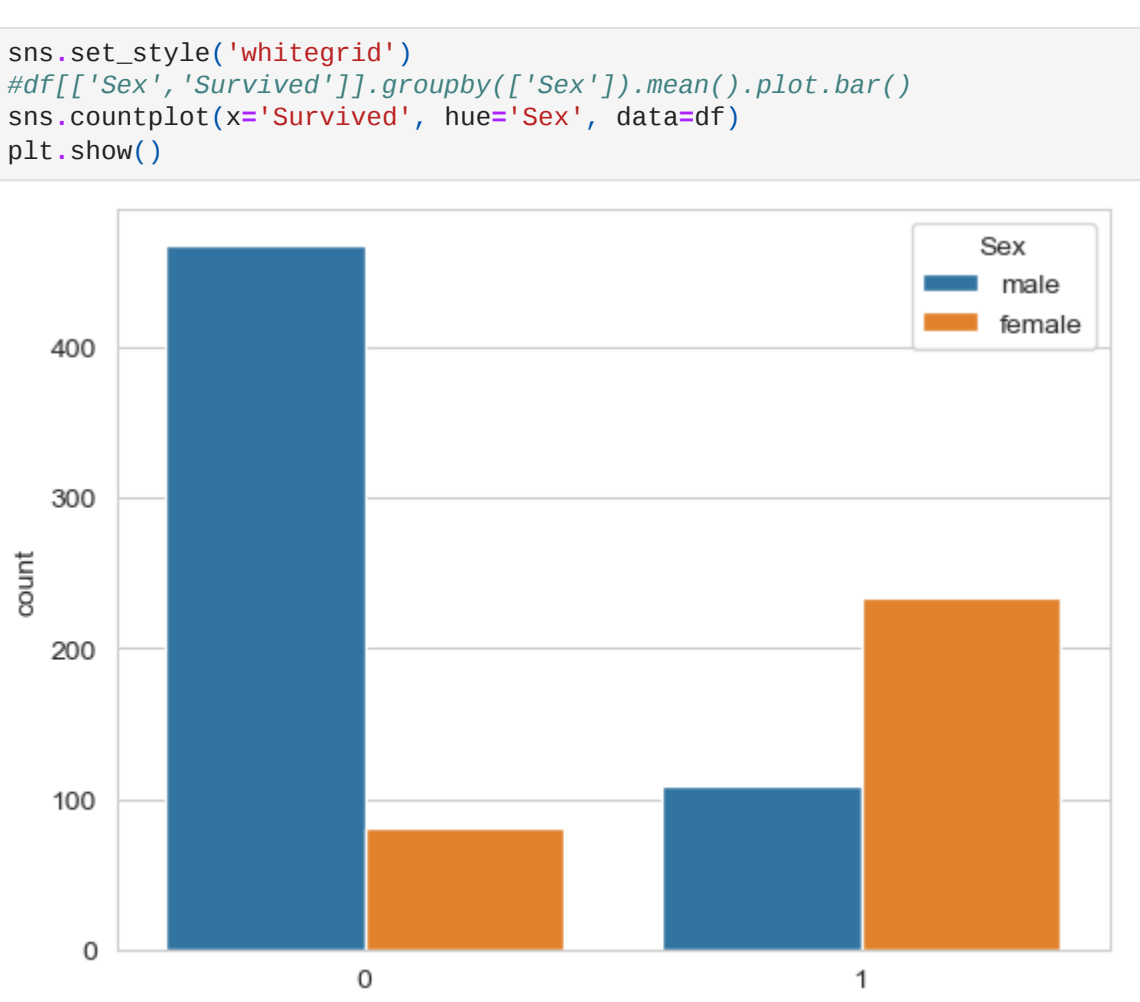
From the above graph it is clear that not many person survived. out of 891 persons in training dataset only 350,38.4% of total training dataset survived . we will get more insights of data by exploring more.

In [217…
```
df.groupby(['Sex','Survived'])['Survived'].count()
```

Out[217]:
```
Sex     Survived
female  0            81
        1           233
male    0           468
        1           109
Name: Survived, dtype: int64
```

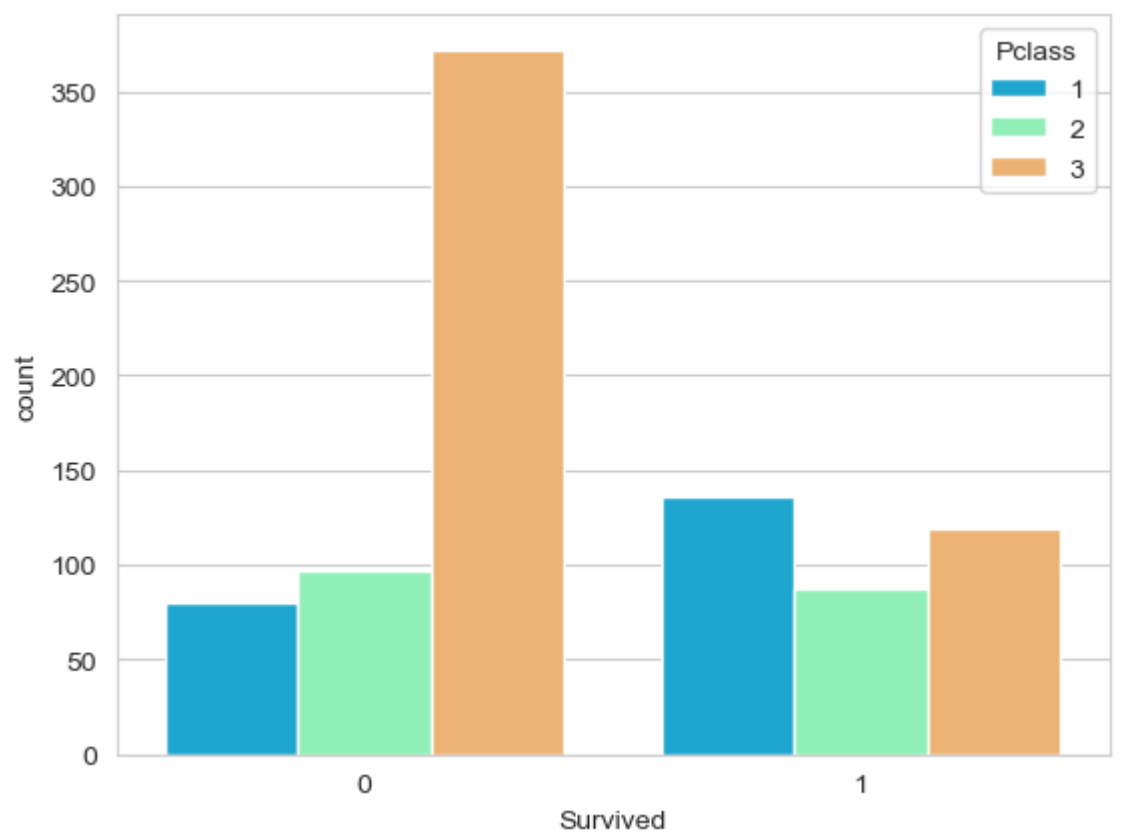
It is clear the 233 female survived out of 344.And out of the 577 male 109 survived. The survival ratio of female is much greater than of male.

In [218…
```
sns.set_style('whitegrid')
#df[['Sex','Survived']].groupby(['Sex']).mean().plot.bar()
sns.countplot(x='Survived', hue='Sex', data=df)
plt.show()
```



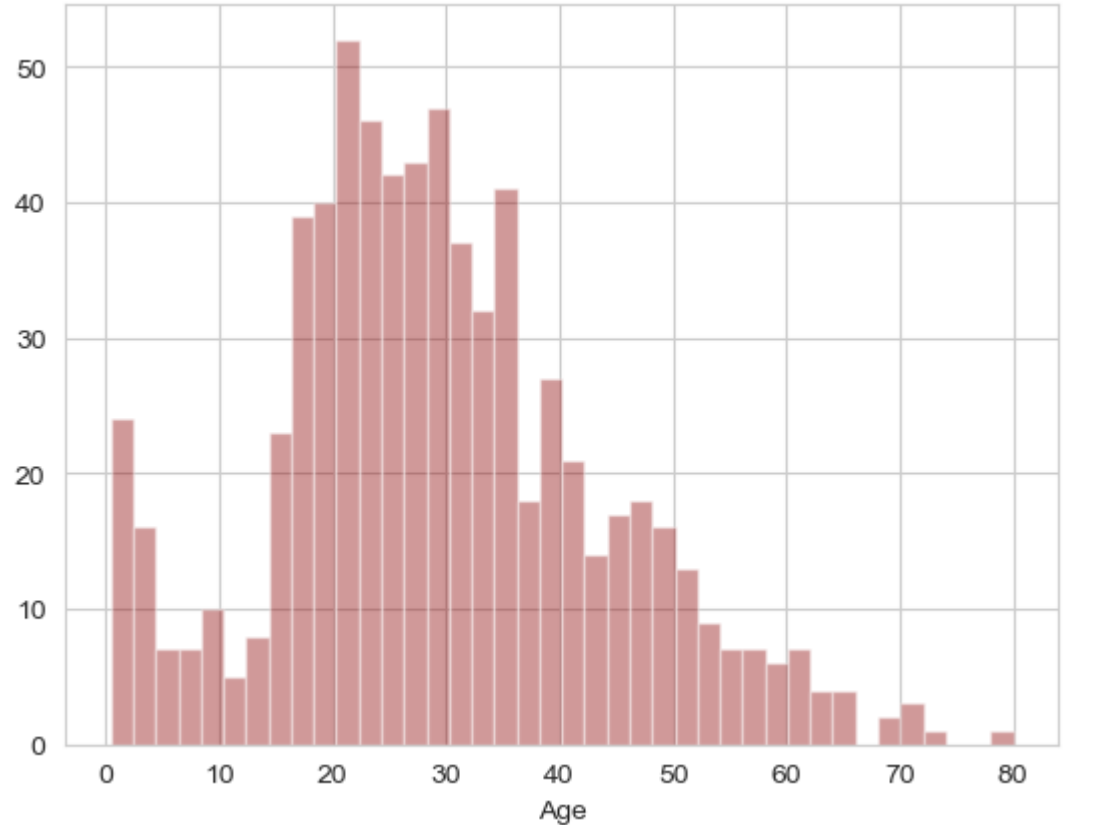
In [219…
```
sns.set_style('whitegrid')
sns.countplot(x= "Survived",hue ='Pclass', data = df, palette = 'rainbow')
```

Out[219]:
```
<Axes: xlabel='Survived', ylabel='count'>
```



In [220…
```
sns.distplot(df['Age'].dropna(),kde=False,color = 'darkred',bins = 40)
```

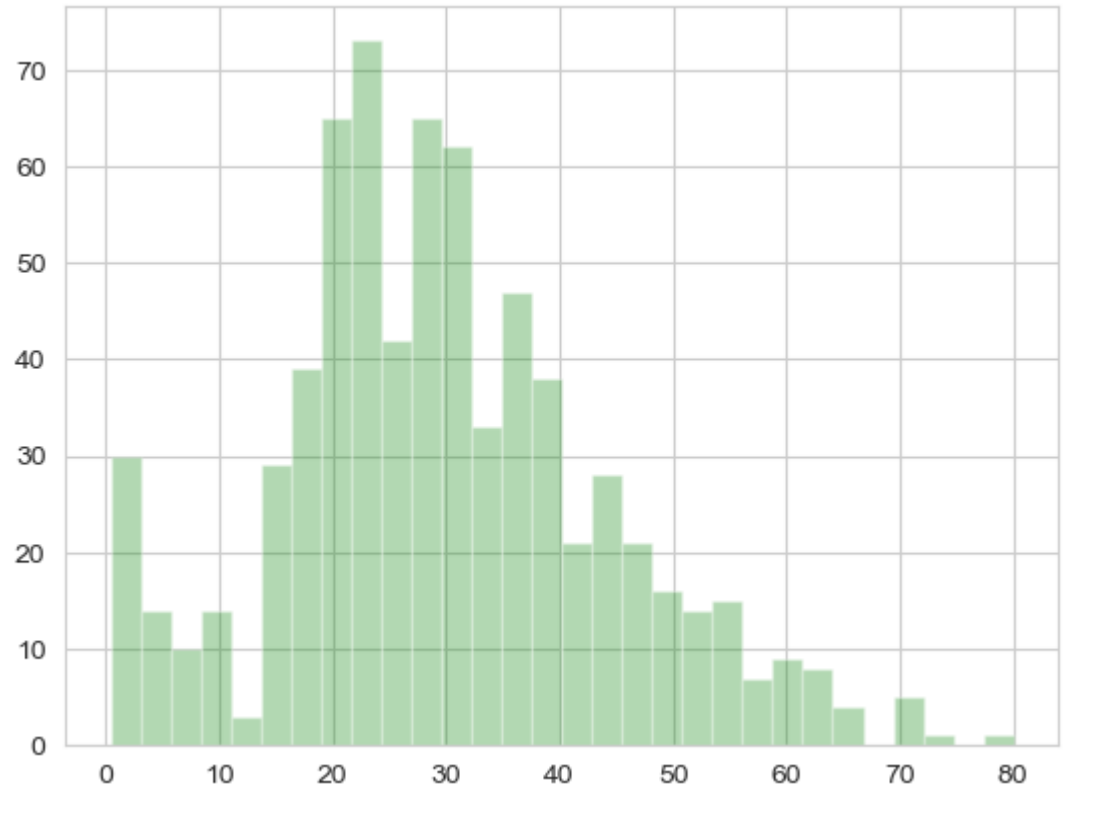
Out[220]:
```
<Axes: xlabel='Age'>
```



we can use distplot or histplot they give same result

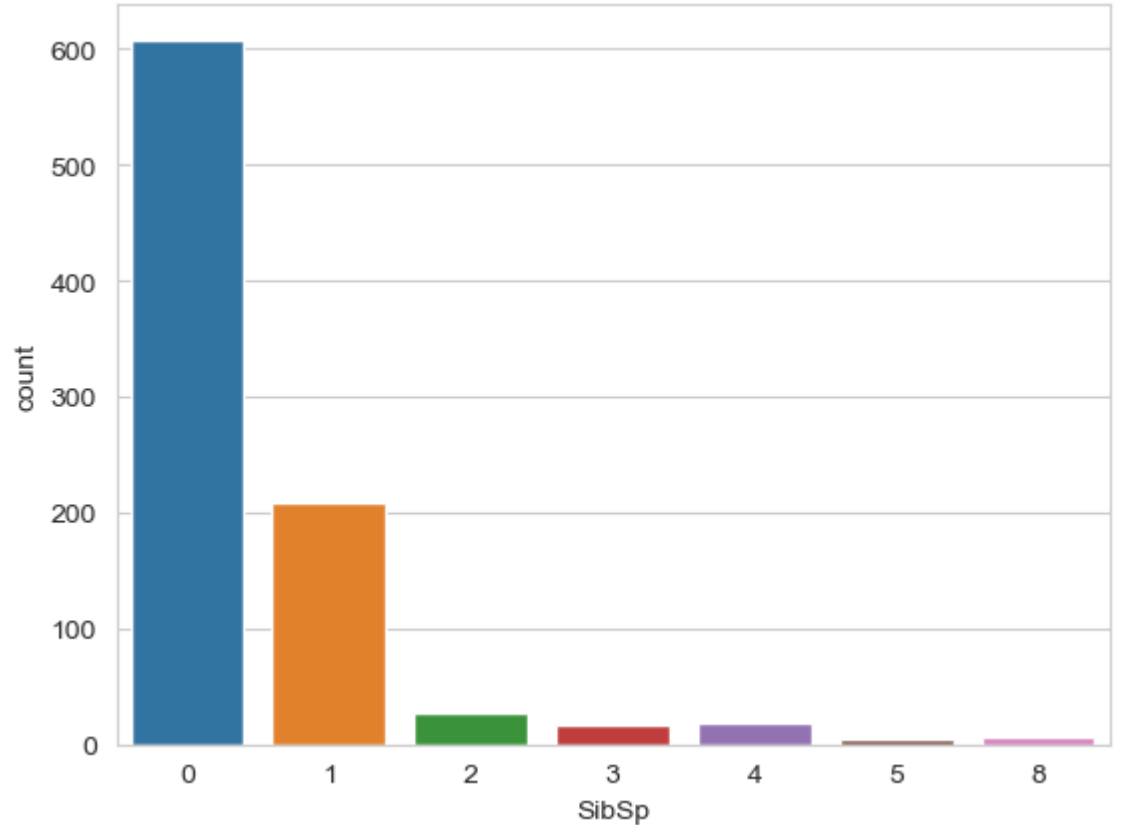
In [221…
```
df["Age"].hist(bins=30,color= "green",alpha=0.3)
```

Out[221]:
```
<Axes: >
```



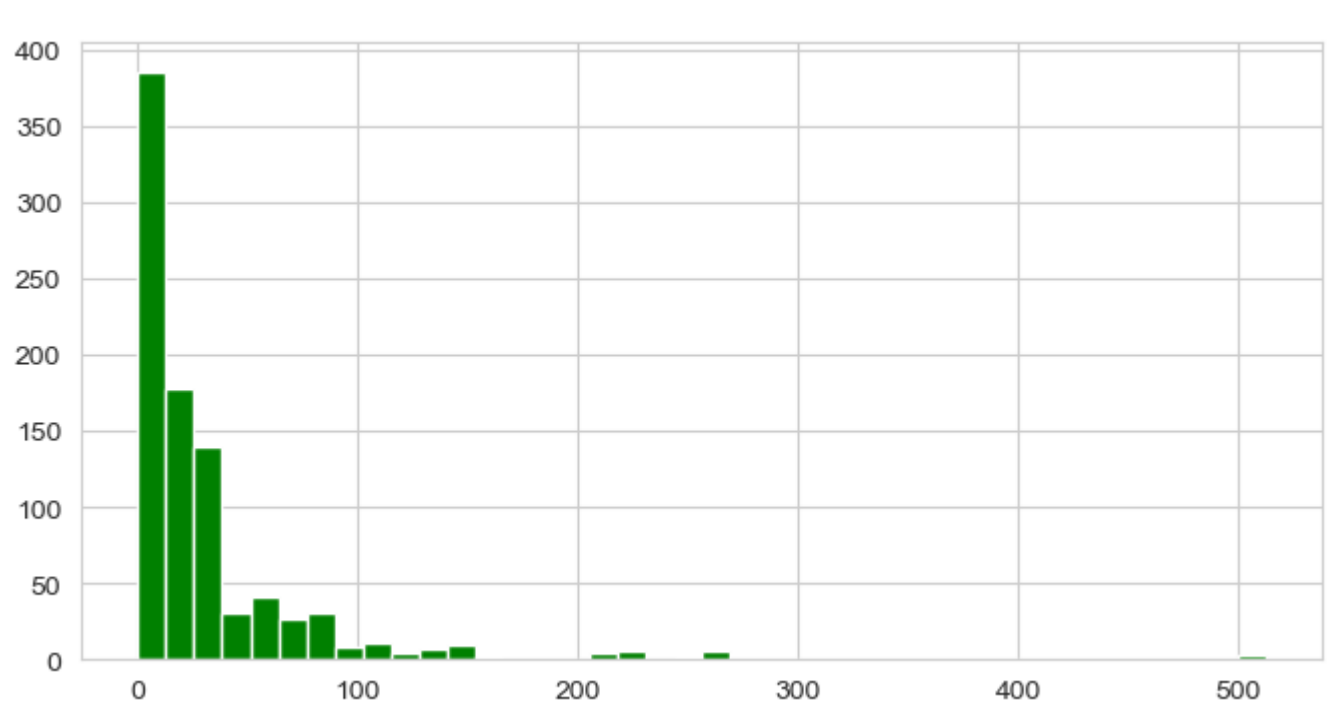
In [222…
```
sns.countplot(x='SibSp',data=df)
```

Out[222]:
```
<Axes: xlabel='SibSp', ylabel='count'>
```



In [223…
```
df['Fare'].hist(color='green',bins=40,figsize=(8,4))
```

Out[223]:
```
<Axes: >
```

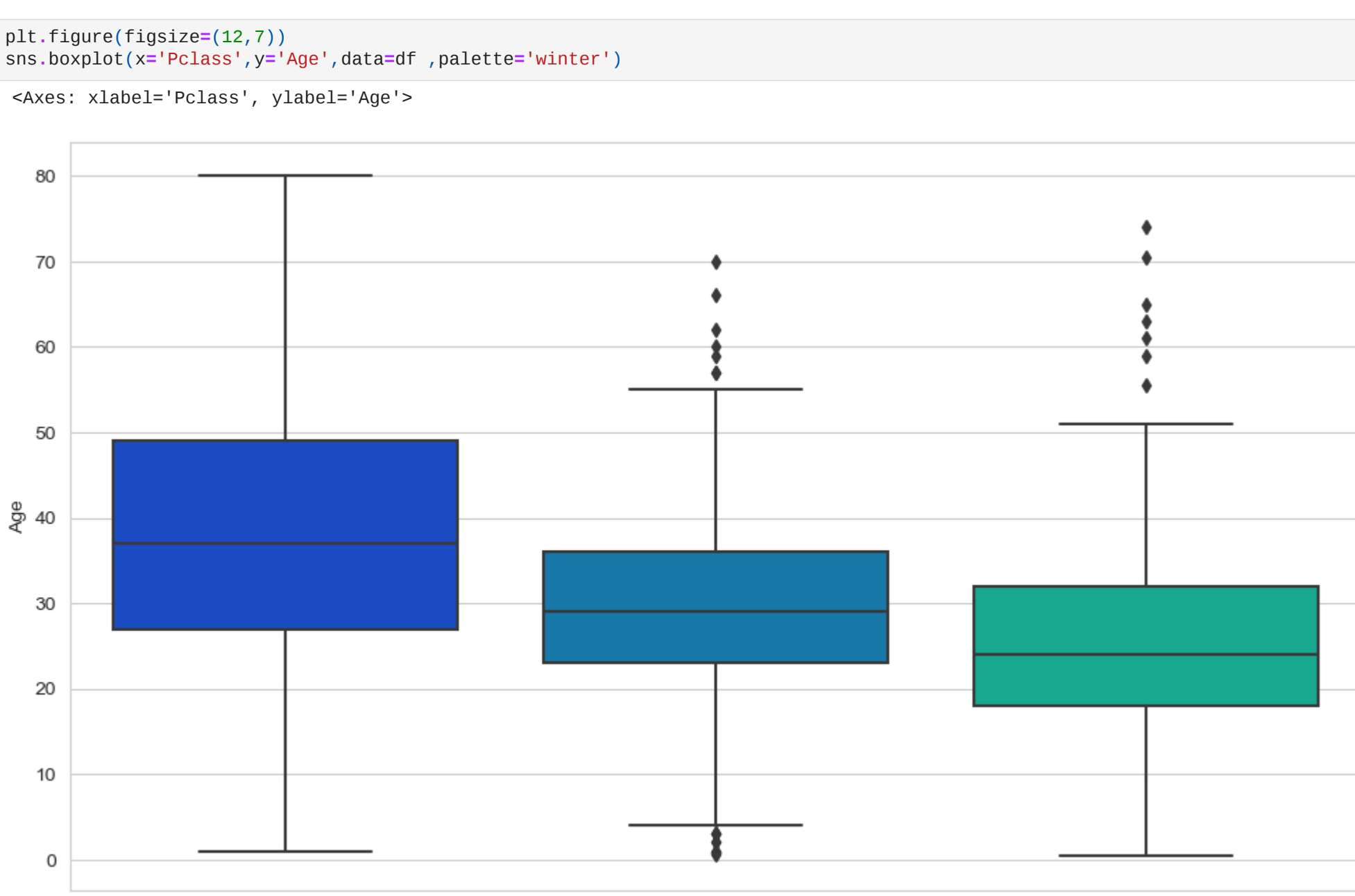


Data Cleaning

We want to fill in missing age data insted of just dropping the missing age data rows.

In [224…
```
plt.figure(figsize=(12,7))
sns.boxplot(x='Pclass',y='Age',data=df ,palette='winter')
```

Out[224]:
```
<Axes: xlabel='Pclass', ylabel='Age'>
```



We can see the wealthier passengers in the higher classes tend to be older.which make sense. We'll use these these average age values to impute based on Plcass for Age.

In [225…
```
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
```

here is the 37,29,24 are all avg value in box plot

In [226…
```
df['Age'] = df[['Age','Pclass']].apply(impute_age,axis = 1)
```

In [227…
```
df.drop('Age', axis = 1,inplace = True)
```

In [228…
```
column_list = list(df.columns)
print(column_list)
```

```
['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked']
```

Now lte's check that heat mao again!

In [229…
```
sns.heatmap(df.isnull(),yticklabels = False,cmap ='viridis')
```

Out[229]:
```
<Axes: >
```

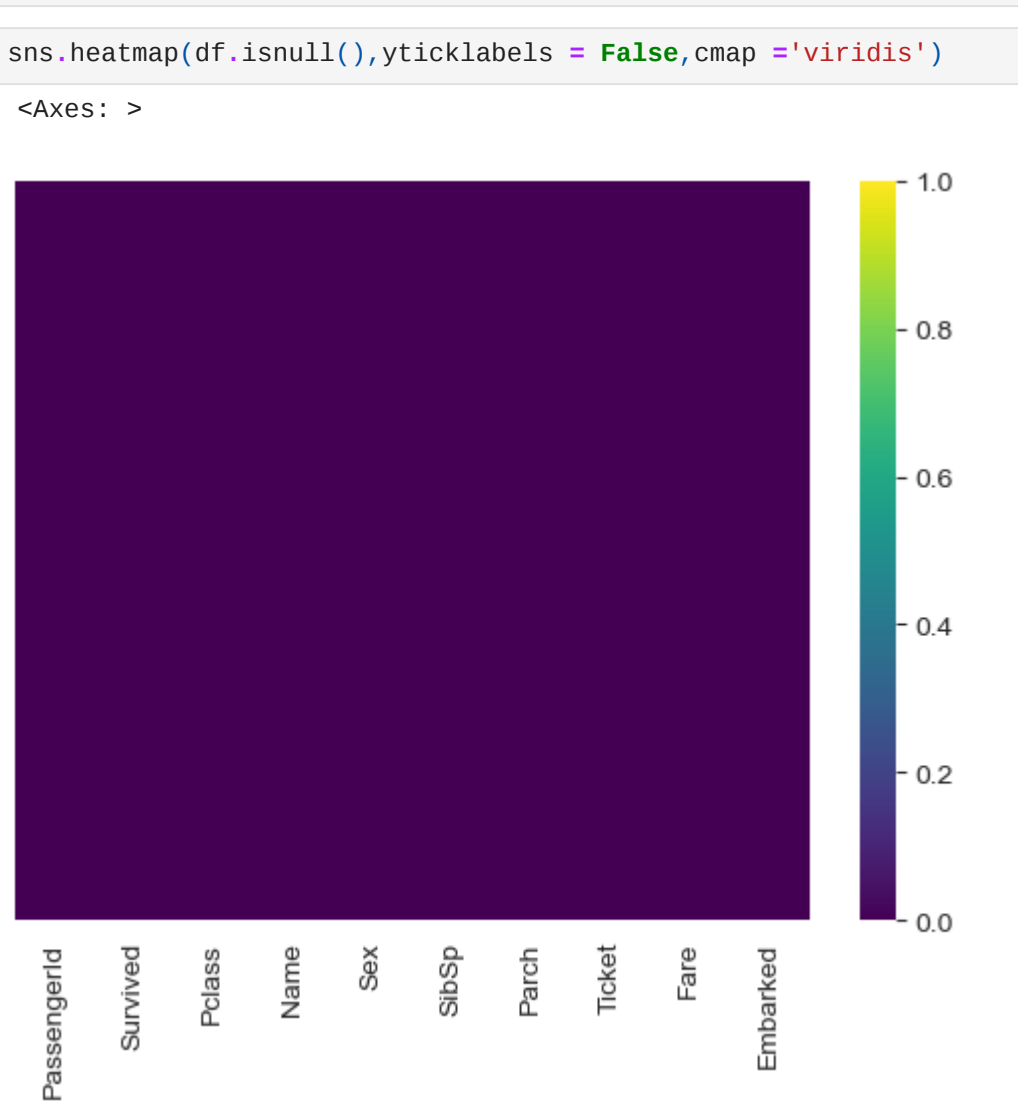
And Here's show the age column has replaced with respect to the Pclass

Drop the Cabin column and row in Emabarked that is NaN

In [230…
```
df.drop('Cabin',axis=1,inplace=True)
```

In [231…
```
sns.heatmap(df.isnull(),yticklabels = False,cmap ='viridis')
```

Out[231]:
```
<Axes: >
```



In [233…
```
df.head()
```

Out[233]:

	PassengerId	Survived	Pclass	Name	Sex	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	0	0	373450	8.0500	S

conclusion

In my analysis of the Titanic dataset, I found missing data regarding passengers' ages, cabin details, and embarkation points. Interestingly, more women and passengers from higher classes survived the disaster. Most passengers were young, but those in higher classes tended to be older. To tidy up the data, I estimated missing ages based on class and removed rows with missing cabin and embarkation details. Moving forward, I plan to explore relationships between different factors and possibly create new features to improve predictive modeling."

In [ ]:

In [ ]: