EDA ON TITANIC DATASET

In [96]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

In [6]:
```
dataset=pd.read_csv(r"C:\Users\S SHYAMILI\OneDrive\Desktop\data science\machine lea
```

In [8]:
```
dataset
```

Out[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7

891 rows × 12 columns

In [46]:
```
dataset.shape
```

Out[46]: (891, 12)

In [30]:
```
dataset.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
```

In [12]:
```
dataset.head(10)
```

Out[12]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250(
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.283:
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925(
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100(
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050(
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.458:
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.862:
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.075(
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.133:
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.070:

In [14]:
```
dataset.nunique()
```

Out[14]:
```
PassengerId    891
Survived         2
Pclass           3
Name           891
Sex              2
Age             88
SibSp            7
Parch            7
Ticket         681
Fare           248
Cabin          147
Embarked         3
dtype: int64
```

In [ ]:
```
#FINDING AND FILLING THE MISSING VALUES
```

In [24]:
```
dataset.isnull().any()
```

Out[24]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age             True
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

In [34]:
```
dataset['Age']=dataset['Age'].fillna(np.mean(pd.to_numeric(dataset['Age'])))
```

In [38]:
```
dataset['Cabin']=dataset['Cabin'].fillna((dataset['Cabin'])).mode()[0]
```

In [42]:
```
dataset['Embarked']=dataset['Embarked'].fillna((dataset['Embarked'])).mode()[0]
```

In [44]:
```
dataset.isnull().any()
```

Out[44]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age            False
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin          False
Embarked       False
dtype: bool
```

In [48]:
```
#removing the irrelevant attributes
```

In [50]:
```
del dataset['Ticket']
```

In [52]:
```
del dataset['SibSp']
```

In [54]:
```
del dataset['Name']
```

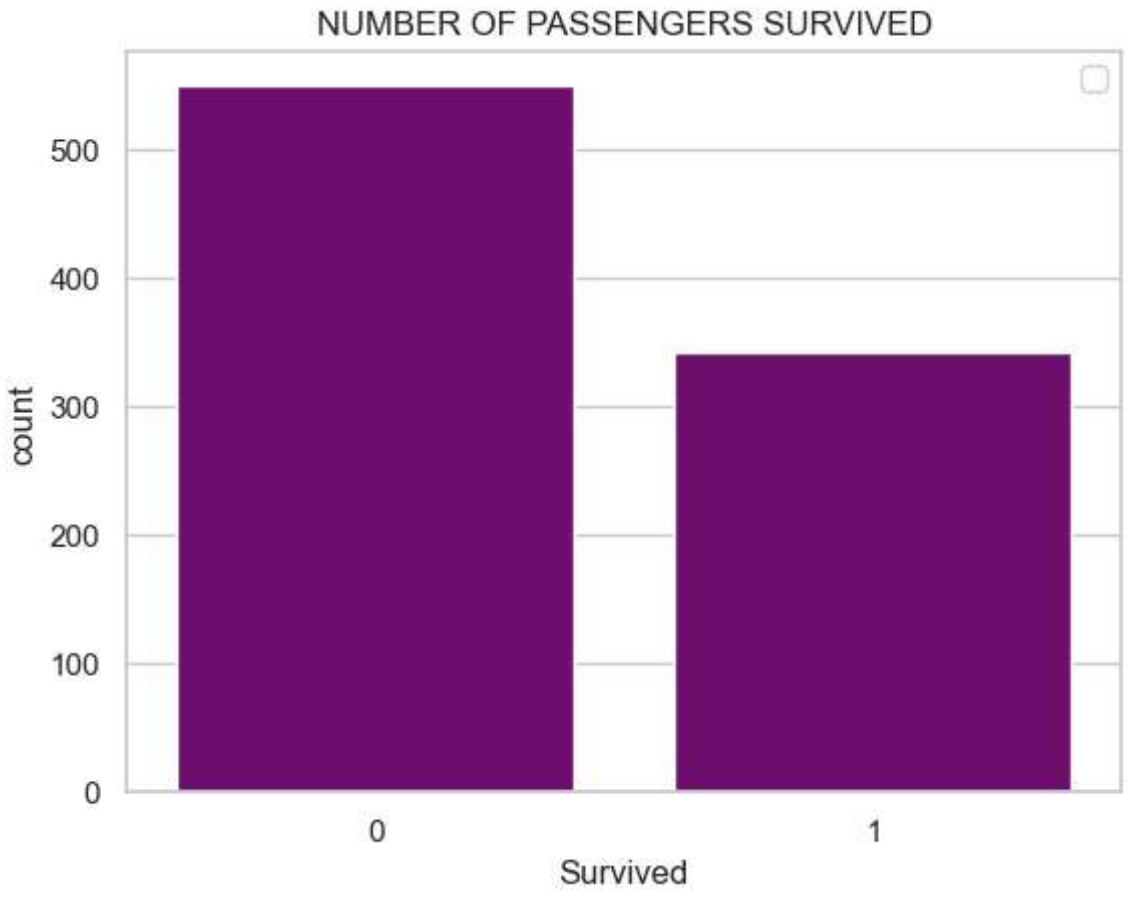
In [56]:
```
del dataset['Cabin']
```

In [58]:
```
#IDENTIFYING THE TRENDS AND PATTERNS
```

In [68]:
```
sns.set_theme(style="whitegrid")
```

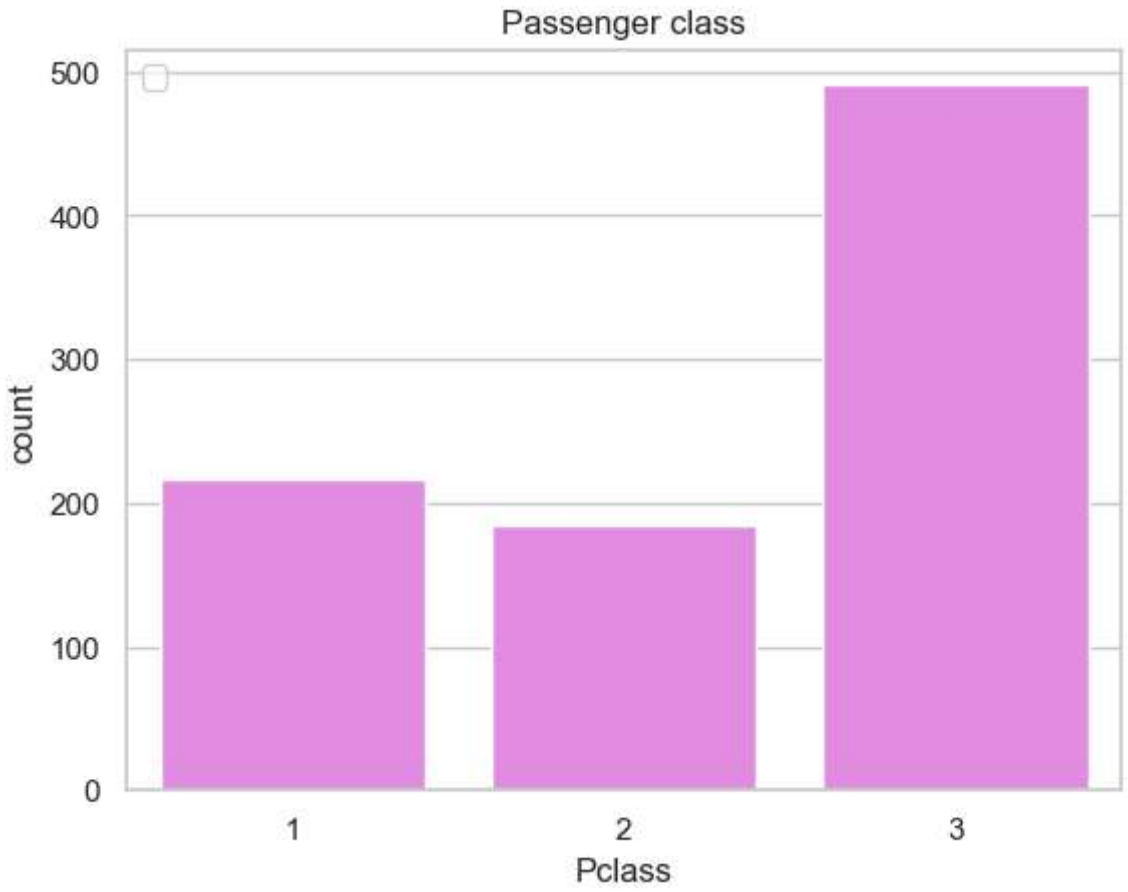
In [146…
```
sns.countplot(data=dataset,x='Survived',color='purple')
plt.title('NUMBER OF PASSENGERS SURVIVED')
plt.ylabel('count')
plt.legend()
```

Out[146… <matplotlib.legend.Legend at 0x21de4a652e0>

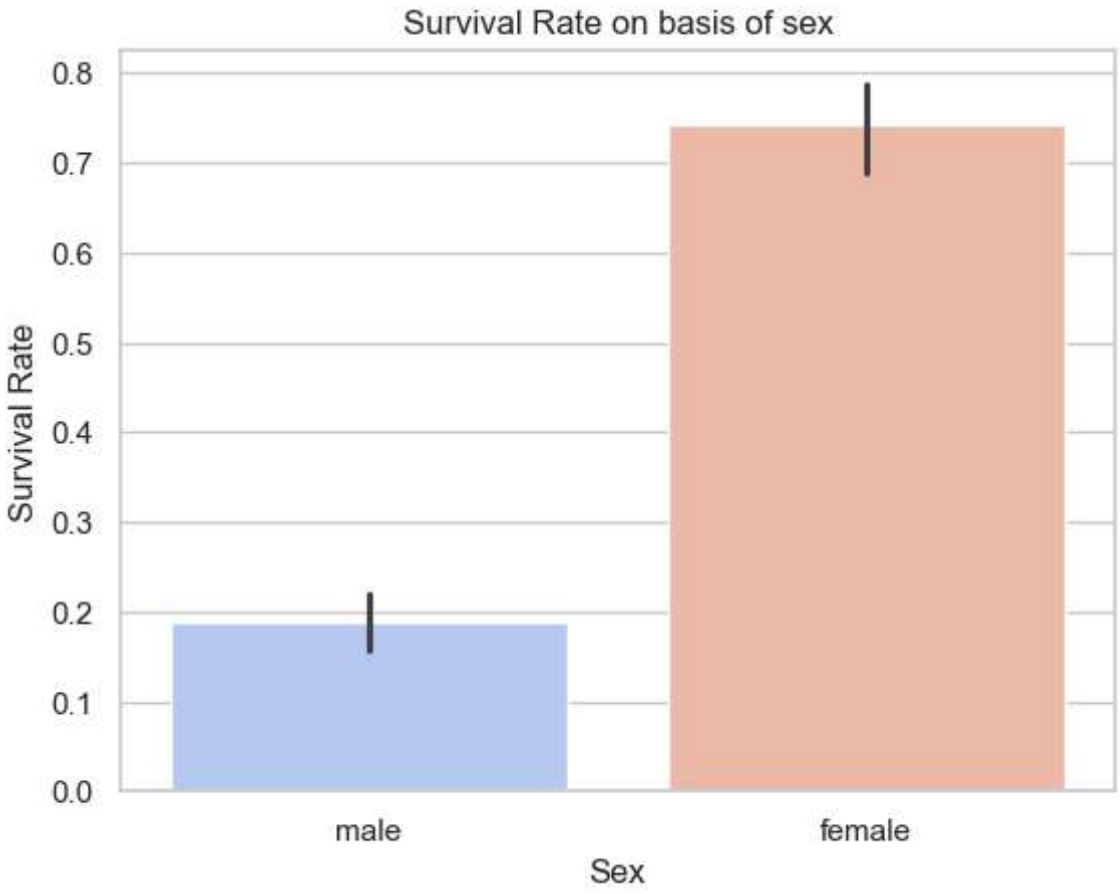


In [148…
```
sns.countplot(data=dataset,x='Pclass',color='violet')
plt.title('Passenger class')
plt.ylabel('count')
plt.legend()
```

Out[148… <matplotlib.legend.Legend at 0x21de4939490>

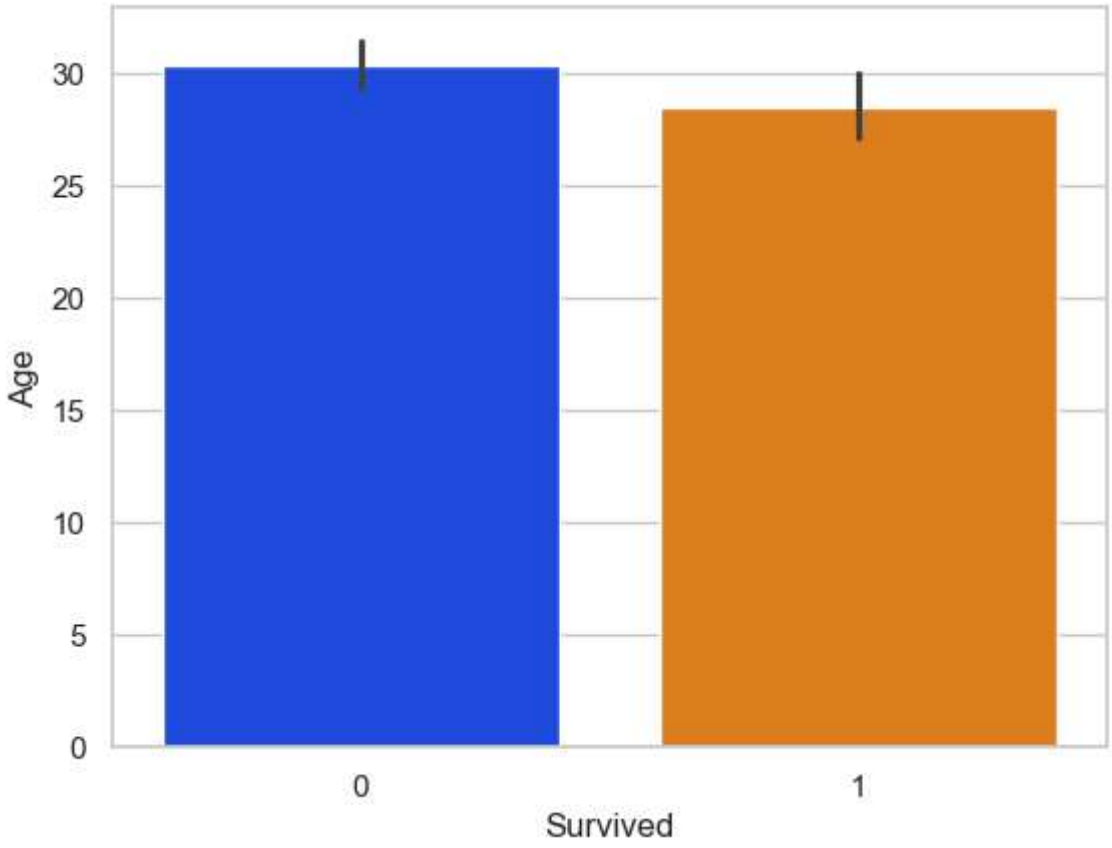


In [168…
```
sns.barplot(data=dataset,x='Sex',y='Survived',palette="coolwarm")
plt.title('NUMBER OF PASSENGERS SURVIVED')
plt.title("Survival Rate on basis of sex")
plt.ylabel("Survival Rate")
plt.show()
```



In [128…
```
sns.barplot(data=dataset,x='Survived',y='Age',palette='bright')
```

Out[128…
```
<Axes: xlabel='Survived', ylabel='Age'>
```

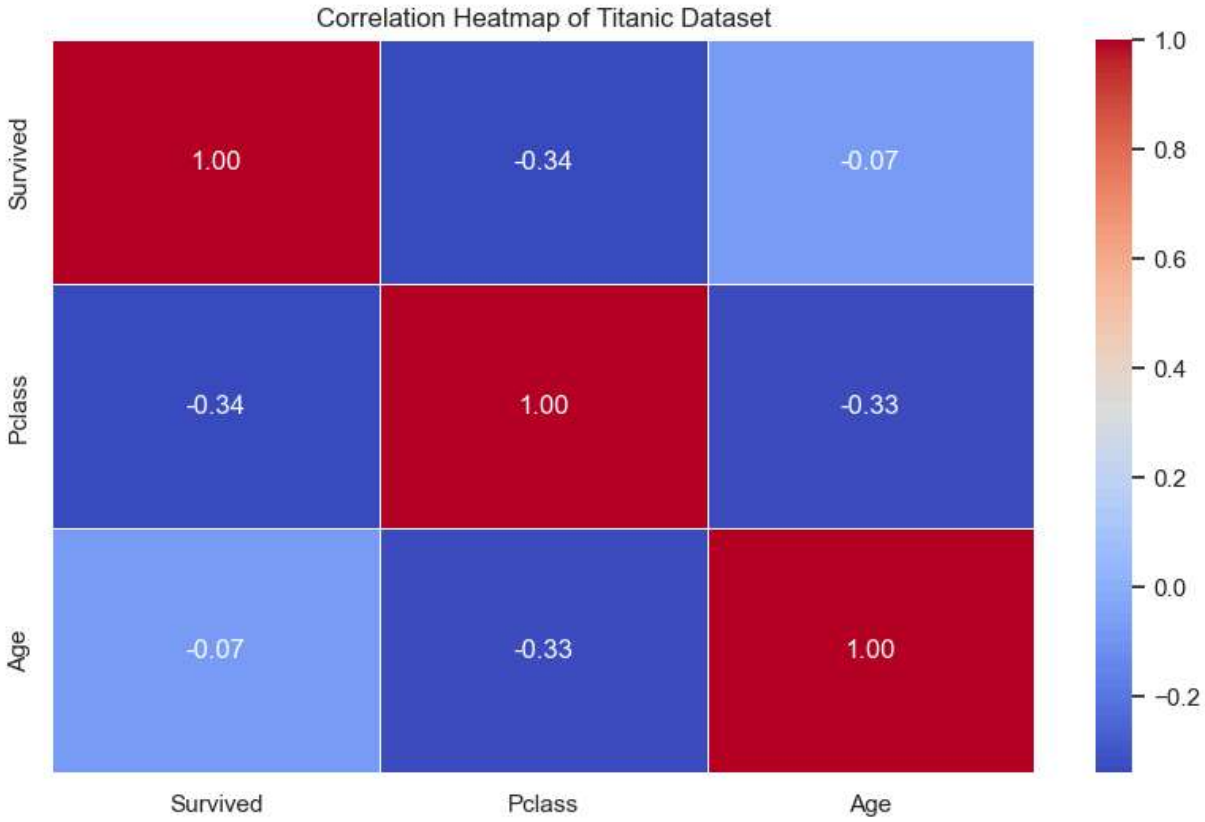


In [160…
```
#FINDING RELATIONS BETWEEN ATTRIBUTES
```

In [166…
```
# Selecting numerical columns for correlation analysis
num_cols = ["Survived", "Pclass", "Age"]

# Compute correlation matrix
corr_matrix = dataset[num_cols].corr()

# Plot heatmap
plt.figure(figsize=(10,6))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Correlation Heatmap of Titanic Dataset")
plt.show()
```



The conclusions drawn from this dataset are: 1.Less number of passengers survived. 2.More number of passengers were in class 3. 3.The survival rate of female is more than the male.