```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
print("done")
```

```
done
```

```
data1= pd.read_csv("C:/Users/hp/Desktop/Skills/EDA/Titanic1.csv")
data2= pd.read_csv("C:/Users/hp/Desktop/Skills/EDA/Titanic2.csv")
print("successfully imported")
```

```
successfully imported
```

```
df1 = pd.DataFrame(data1)
df2 = pd.DataFrame(data2)
df= pd.concat([df1,df2], ignore_index = True)
print("successfully concatenated")
```

```
successfully concatenated
```

+ Code + Text

```
df.head()
```

| | 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 |

```
df.tail()
```

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1304** | 1305 | 0 | 3 | Spector, Mr. Woolf | male | NaN | 0 | 0 | A.5. 3236 | 8.0500 | NaN | S |
| **1305** | 1306 | 1 | 1 | Oliva y Ocana, Dona. Fermina | female | 39.0 | 0 | 0 | PC 17758 | 108.9000 | C105 | C |
| **1306** | 1307 | 0 | 3 | Saether, Mr. Simon Sivertsen | male | 38.5 | 0 | 0 | SOTON/O.Q. 3101262 | 7.2500 | NaN | S |
| **1307** | 1308 | 0 | 3 | Ware, Mr. Frederick | male | NaN | 0 | 0 | 359309 | 8.0500 | NaN | S |
| **1308** | 1309 | 0 | 3 | Peter, Master. Michael J | male | NaN | 1 | 1 | 2668 | 22.3583 | NaN | C |

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  1309 non-null   int64
 1   Survived     1309 non-null   int64
 2   Pclass       1309 non-null   int64
 3   Name         1309 non-null   object
 4   Sex          1309 non-null   object
 5   Age          1046 non-null   float64
 6   SibSp        1309 non-null   int64
 7   Parch        1309 non-null   int64
 8   Ticket       1309 non-null   object
 9   Fare         1308 non-null   float64
 10  Cabin        295 non-null    object
 11  Embarked     1307 non-null   object
dtypes: float64(2), int64(5), object(5)
memory usage: 122.8+ KB
```

```
df.describe()
```

| | PassengerId | Survived | Pclass | Age | SibSp | Parch | Fare |
|---|---|---|---|---|---|---|---|
| **count** | 1309.000000 | 1309.000000 | 1309.000000 | 1046.000000 | 1309.000000 | 1309.000000 | 1308.000000 |
| **mean** | 655.000000 | 0.377387 | 2.294882 | 29.881138 | 0.498854 | 0.385027 | 33.295479 |
| **std** | 378.020061 | 0.484918 | 0.837836 | 14.413493 | 1.041658 | 0.865560 | 51.758668 |
| **min** | 1.000000 | 0.000000 | 1.000000 | 0.170000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 328.000000 | 0.000000 | 2.000000 | 21.000000 | 0.000000 | 0.000000 | 7.895800 |
| **50%** | 655.000000 | 0.000000 | 3.000000 | 28.000000 | 0.000000 | 0.000000 | 14.454200 |
| **75%** | 982.000000 | 1.000000 | 3.000000 | 39.000000 | 1.000000 | 0.000000 | 31.275000 |
| **max** | 1309.000000 | 1.000000 | 3.000000 | 80.000000 | 8.000000 | 9.000000 | 512.329200 |

```
df.isnull().sum()
```

```
PassengerId       0
Survived          0
Pclass            0
Name              0
Sex               0
Age             263
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin          1014
Embarked          2
dtype: int64
```

```
#dealing with missing values in Age column.
mean_age = df['Age'].mean()
rounded_mean_age = round(mean_age, 2)
print(rounded_mean_age)
df['Age'] = df['Age'].fillna(rounded_mean_age)
```

```
29.88
```

```
df.tail()
```

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1304** | 1305 | 0 | 3 | Spector, Mr. Woolf | male | 29.88 | 0 | 0 | A.5. 3236 | 8.0500 | NaN | S |
| **1305** | 1306 | 1 | 1 | Oliva y Ocana, Dona. Fermina | female | 39.00 | 0 | 0 | PC 17758 | 108.9000 | C105 | C |
| **1306** | 1307 | 0 | 3 | Saether, Mr. Simon Sivertsen | male | 38.50 | 0 | 0 | SOTON/O.Q. 3101262 | 7.2500 | NaN | S |
| **1307** | 1308 | 0 | 3 | Ware, Mr. Frederick | male | 29.88 | 0 | 0 | 359309 | 8.0500 | NaN | S |
| **1308** | 1309 | 0 | 3 | Peter, Master. Michael J | male | 29.88 | 1 | 1 | 2668 | 22.3583 | NaN | C |

```
df.isnull().sum()
```

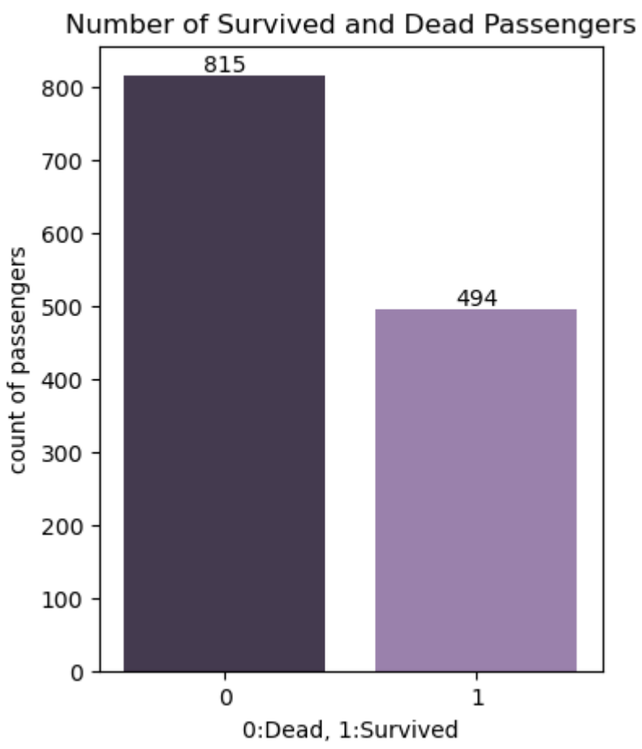
```
PassengerId       0
Survived          0
Pclass            0
Name              0
Sex               0
Age               0
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin          1014
Embarked          2
dtype: int64
```

```
df["Embarked"] = df["Embarked"].replace(np.nan, "unknown")


df.isnull().sum()
```

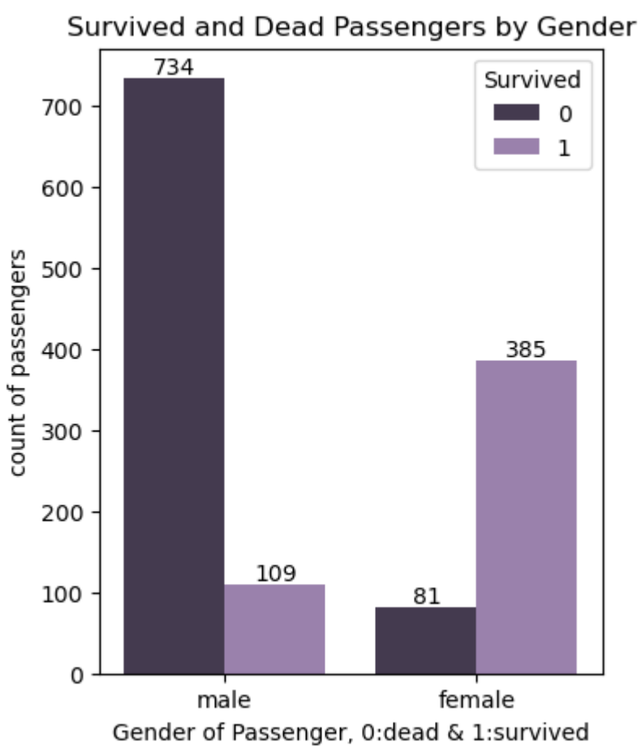
```
PassengerId       0
Survived          0
Pclass            0
Name              0
Sex               0
Age               0
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin          1014
Embarked          0
dtype: int64
```

```
plt.figure(figsize = (4,5))
plot1= sns.countplot(x= "Survived", data = df, palette=["#443752","#9b7ab3"])
plot1.bar_label(plot1.containers[0])
plot1.bar_label(plot1.containers[1])
plt.title("Number of Survived and Dead Passengers")
plt.xlabel("0:Dead, 1:Survived")
plt.ylabel("count of passengers")
plt.show()
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_2720\1167081184.py:2: FutureWarning:
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  plot1= sns.countplot(x= "Survived", data = df, palette=["#443752","#9b7ab3"])
```

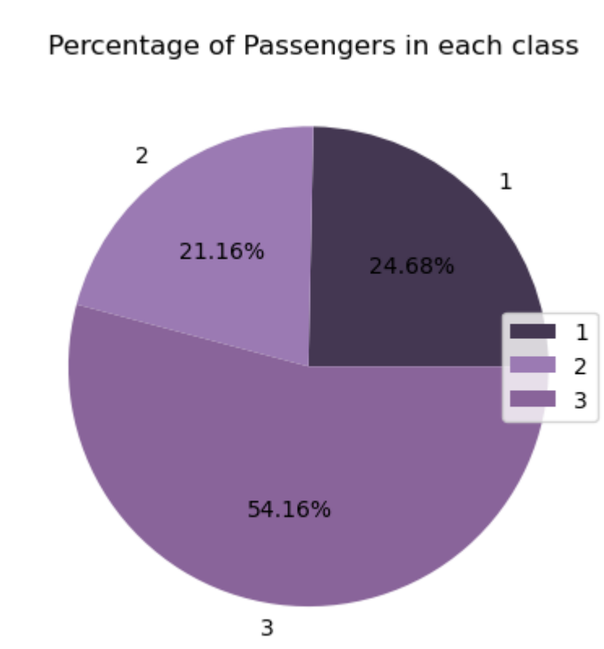



```
plt.figure(figsize = (4,5))
plot1= sns.countplot(x= "Sex", data = df, palette=["#443752","#9b7ab3"], hue="Survived")
plot1.bar_label(plot1.containers[0])
plot1.bar_label(plot1.containers[1])
plt.title("Survived and Dead Passengers by Gender")
plt.xlabel("Gender of Passenger, 0:dead & 1:survived")
plt.ylabel("count of passengers")
plt.show()
```

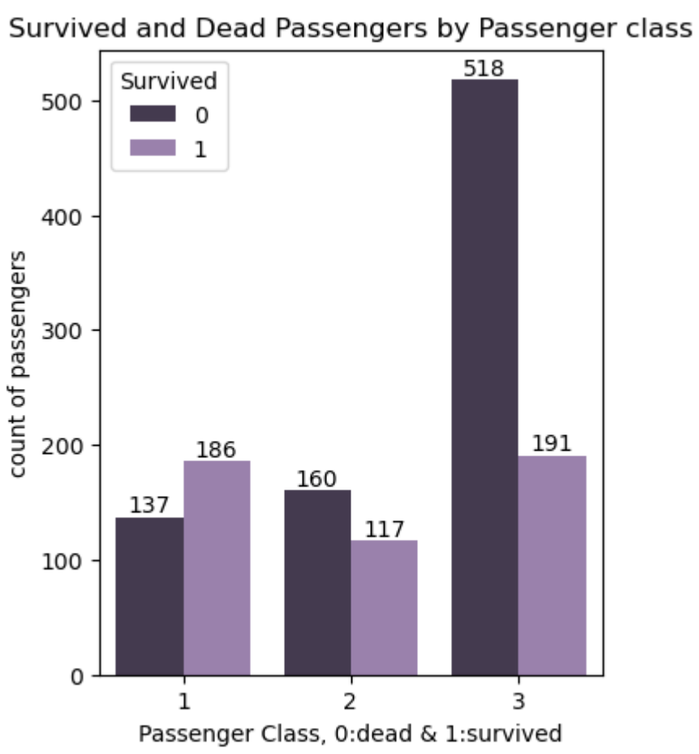



Here, we conclude that female passengers have a higher survival rate than male passengers.

```
colour =["#443752","#9b7ab3","#89649a"]
gb= df.groupby("Pclass").agg({"Pclass": "count"})
plot2= plt.pie(gb["Pclass"],autopct="%1.2f%%",colors= colour,labels= gb.index)
plt.title(" Percentage of Passengers in each class")
plt.legend(loc= "center right")
plt.show()
```

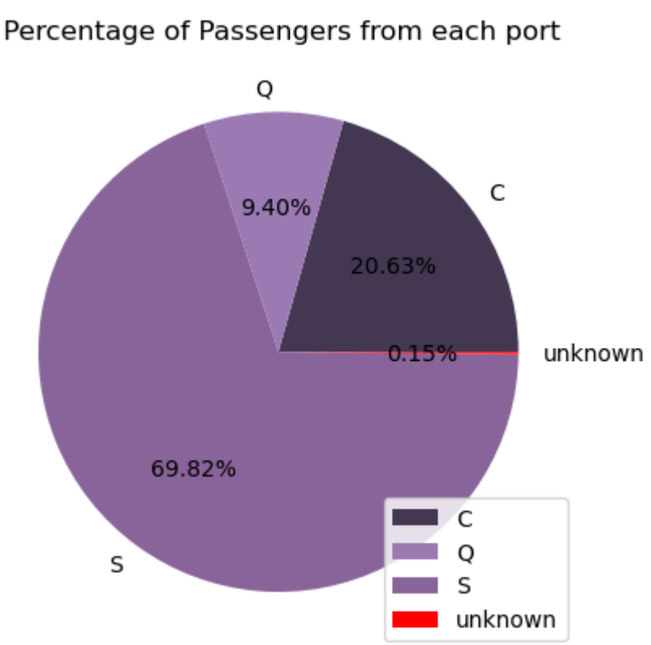



```
plt.figure(figsize = (4,5))
plot1= sns.countplot(x= "Pclass", data = df, palette=["#443752","#9b7ab3"], hue="Survived")
plot1.bar_label(plot1.containers[0])
plot1.bar_label(plot1.containers[1])
plt.title("Survived and Dead Passengers by Passenger class")
plt.xlabel("Passenger Class, 0:dead & 1:survived")
plt.ylabel("count of passengers")
plt.show()
```




Through the above two graphs, we can conclude that the largest population of passengers belongs to Pclass 3, at the same time the maximum number of passengers died belongs also belongs to Pclass3 which implies that passengers from Pclass 1 and 2 were prioritized at the time of rescue.

```
colour =["#443752","#9b7ab3","#89649a","red"]
gb1= df.groupby("Embarked").agg({"Embarked": "count"})
plot2= plt.pie(gb1["Embarked"],autopct="%1.2f%%",colors= colour,labels= gb1.index)
plt.title(" Percentage of Passengers from each port")
plt.legend(loc= "lower right")
plt.show()
```




The name of the ports are: C = Cherbourg, Q = Queenstown, S = Southampton.From this pie chart we can conclude that the maximum passengers embarked from Southampton port.

```
plt.figure(figsize = (4,5))
plot1= sns.countplot(x= "Survived", data = df, palette=["#443752","#9b7ab3","#89649a","#574571"], hue="Embarked")
plot1.bar_label(plot1.containers[0])
plot1.bar_label(plot1.containers[1])
plot1.bar_label(plot1.containers[2])
plot1.bar_label(plot1.containers[3])
plt.title("Survived and Dead Passengers by each Embarked port")
plt.xlabel("Embarked port, 0:dead & 1:survived")
plt.ylabel("count of passengers")
plt.show()
```