```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os

os.chdir("D:\\DATASET")

titanic_data = pd.read_csv("train.csv")
```

Exploratory Data Analysis

```
titanic_data.head(10)
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3
5            6         0       3
6            7         0       1
7            8         0       3
8            9         1       3
9           10         1       2

                                                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
5                                   Moran, Mr. James    male   NaN
0
6                            McCarthy, Mr. Timothy J    male  54.0
0
7                     Palsson, Master. Gosta Leonard    male   2.0
3
8  Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)  female  27.0
0
9                Nasser, Mrs. Nicholas (Adele Achem)  female  14.0
1
```

```
   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
5      0            330877   8.4583   NaN        Q
6      0             17463  51.8625   E46        S
7      1            349909  21.0750   NaN        S
8      2            347742  11.1333   NaN        S
9      0            237736  30.0708   NaN        C
```

```
titanic_data.tail(10)
```

```
     PassengerId  Survived  Pclass
Name  \
881          882         0       3                        Markun, Mr.
Johann
882          883         0       3              Dahlberg, Miss. Gerda
Ulrika
883          884         0       2             Banfield, Mr. Frederick
James
884          885         0       3                    Sutehall, Mr.
Henry Jr
885          886         0       3      Rice, Mrs. William (Margaret
Norton)
886          887         0       2                     Montvila, Rev.
Juozas
887          888         1       1              Graham, Miss. Margaret
Edith
888          889         0       3  Johnston, Miss. Catherine Helen
"Carrie"
889          890         1       1                     Behr, Mr. Karl
Howell
890          891         0       3                       Dooley, Mr.
Patrick
```

```
        Sex   Age  SibSp  Parch            Ticket     Fare Cabin
Embarked
881    male  33.0      0      0            349257   7.8958   NaN
S
882  female  22.0      0      0              7552  10.5167   NaN
S
883    male  28.0      0      0  C.A./SOTON 34068  10.5000   NaN
S
884    male  25.0      0      0   SOTON/OQ 392076   7.0500   NaN
S
885  female  39.0      0      5            382652  29.1250   NaN
Q
886    male  27.0      0      0            211536  13.0000   NaN
```

```
S
887  female  19.0      0      0            112053  30.0000   B42
S
888  female   NaN      1      2        W./C. 6607  23.4500   NaN
S
889    male  26.0      0      0            111369  30.0000  C148
C
890    male  32.0      0      0            370376   7.7500   NaN
Q
```

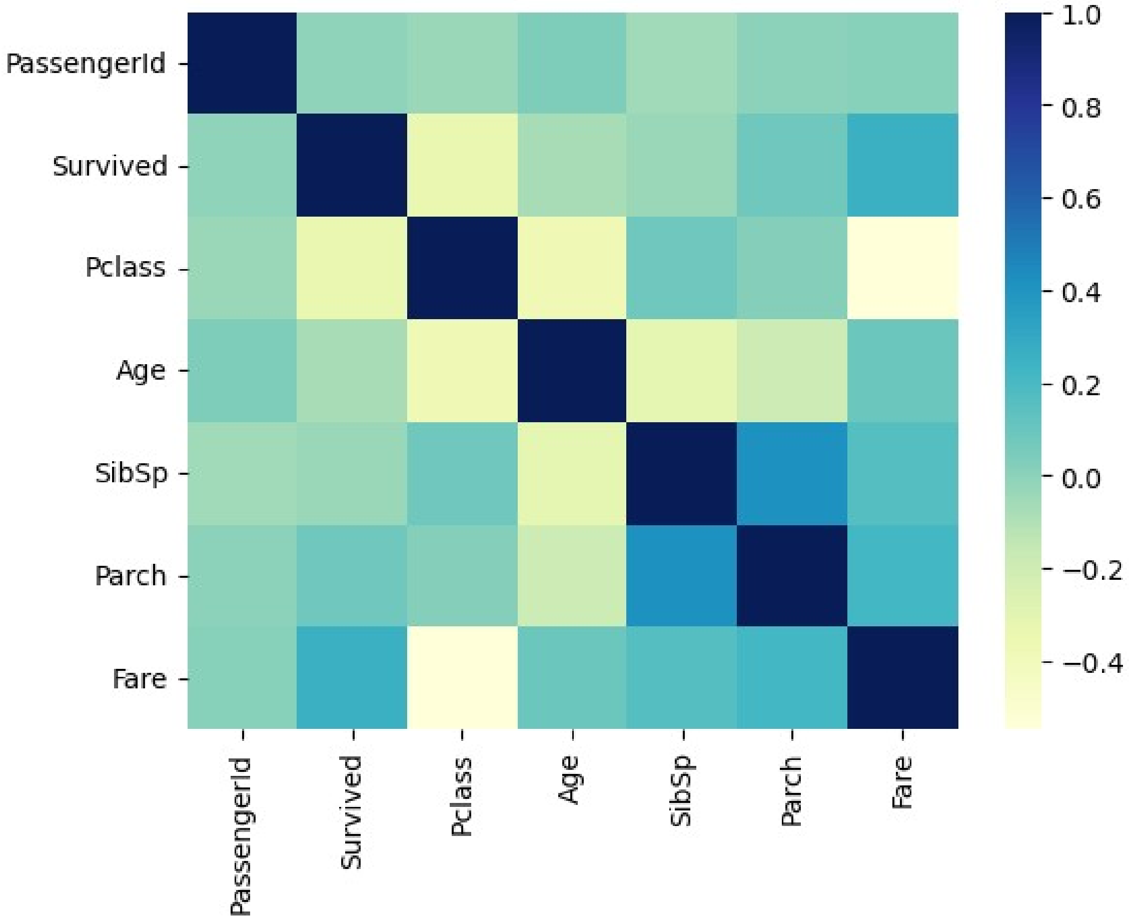
```
titanic_data.describe()
```

```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000   
mean    446.000000    0.383838    2.308642   29.699118    0.523008   
std     257.353842    0.486592    0.836071   14.526497    1.102743   
min       1.000000    0.000000    1.000000    0.420000    0.000000   
25%     223.500000    0.000000    2.000000   20.125000    0.000000   
50%     446.000000    0.000000    3.000000   28.000000    0.000000   
75%     668.500000    1.000000    3.000000   38.000000    1.000000   
max     891.000000    1.000000    3.000000   80.000000    8.000000   

            Parch        Fare  
count  891.000000  891.000000  
mean     0.381594   32.204208  
std      0.806057   49.693429  
min      0.000000    0.000000  
25%      0.000000    7.910400  
50%      0.000000   14.454200  
75%      0.000000   31.000000  
max      6.000000  512.329200  
```

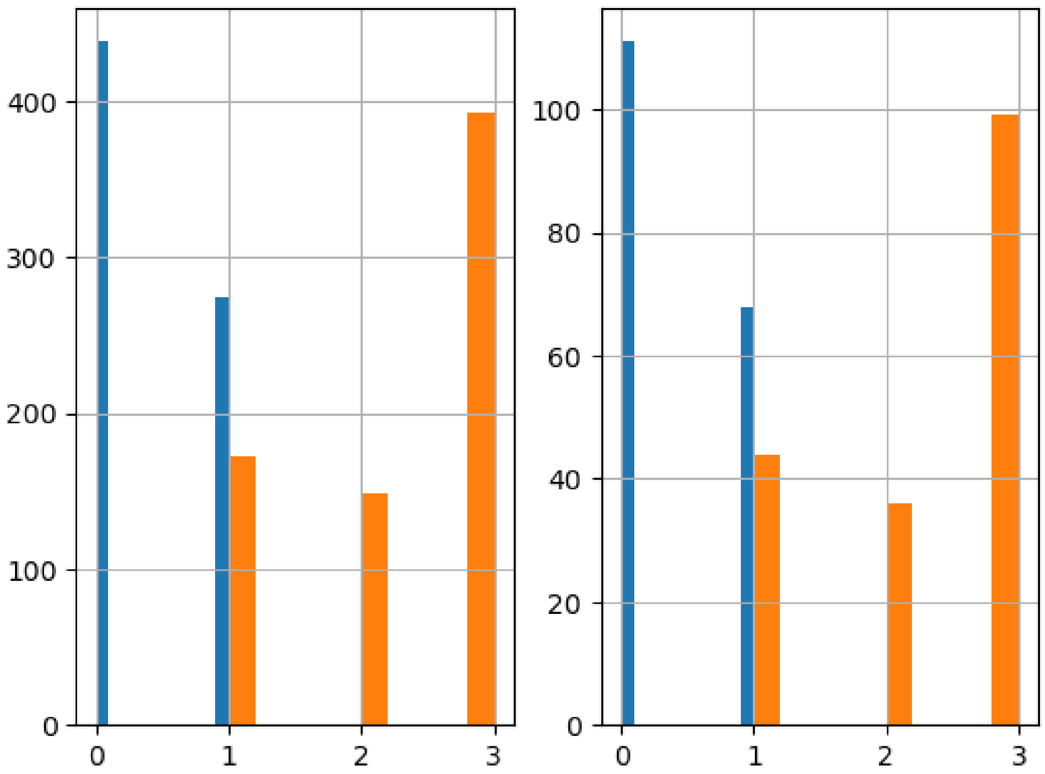
Visualization

```
import seaborn as sns

sns.heatmap(titanic_data.corr(), cmap="YlGnBu")
plt.show()
```

```
C:\Users\SUJIT KUMAR SAHOO\AppData\Local\Temp\
ipykernel_1616\1602845089.py:3: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  sns.heatmap(titanic_data.corr(), cmap="YlGnBu")
```



```python
from sklearn.model_selection import StratifiedShuffleSplit

split = StratifiedShuffleSplit(n_splits=1, test_size=0.2)
for train_indices, test_indices in split.split(titanic_data,
titanic_data[["Survived", "Pclass","Sex"]]):
    strat_train_set = titanic_data.loc[train_indices]
    strat_test_set = titanic_data.loc[test_indices]

plt.subplot(1,2,1)
strat_train_set['Survived'].hist()
strat_train_set['Pclass'].hist()

plt.subplot(1,2,2)
strat_test_set['Survived'].hist()
strat_test_set['Pclass'].hist()

plt.show()
```

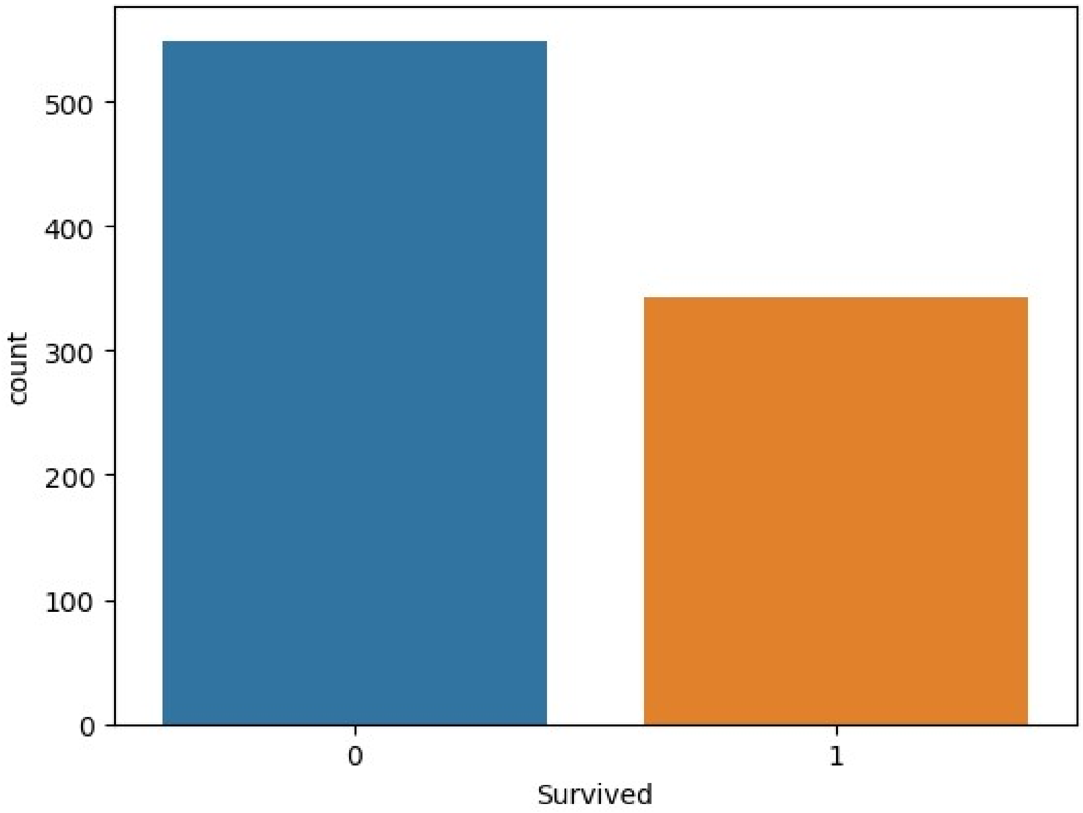


```
strat_train_set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 712 entries, 154 to 4
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  712 non-null    int64  
 1   Survived     712 non-null    int64  
 2   Pclass       712 non-null    int64  
 3   Name         712 non-null    object 
 4   Sex          712 non-null    object 
 5   Age          570 non-null    float64
 6   SibSp        712 non-null    int64  
 7   Parch        712 non-null    int64  
 8   Ticket       712 non-null    object 
 9   Fare         712 non-null    float64
 10  Cabin        166 non-null    object 
 11  Embarked     710 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 72.3+ KB
```

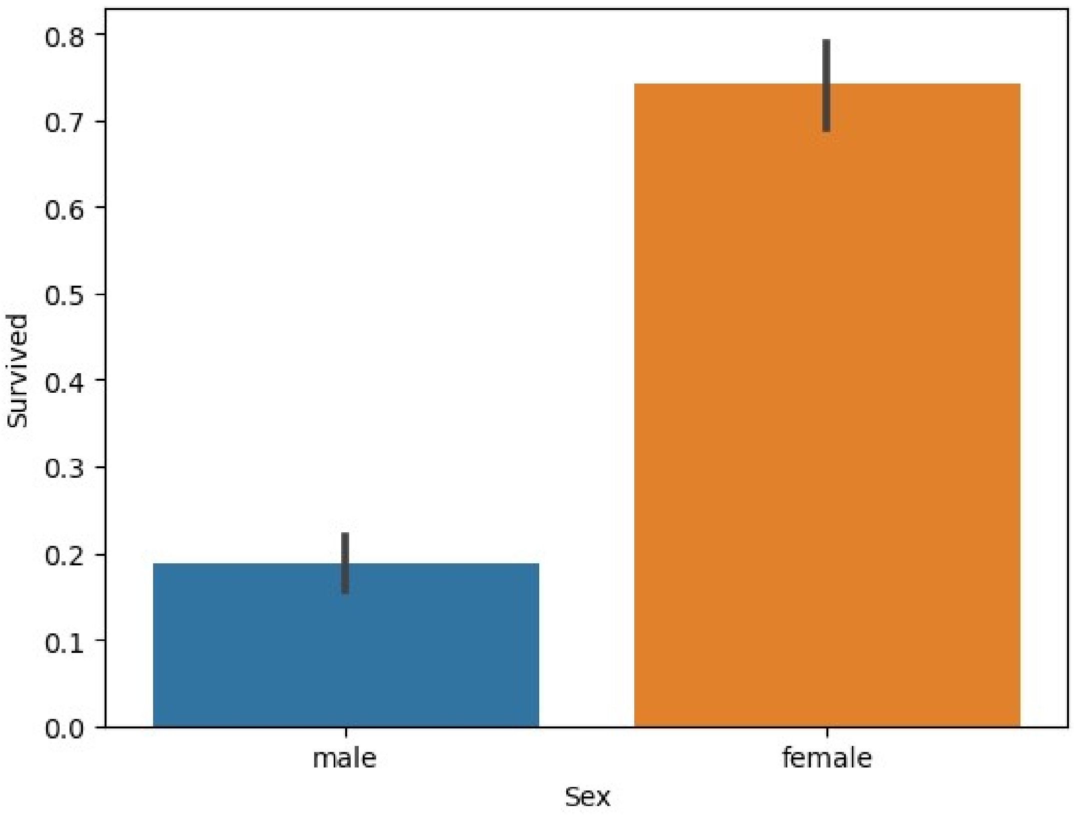
```
sns.countplot(data=titanic_data,x="Survived")
```

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



```
sns.barplot(x="Sex", y="Survived", data=titanic_data)
```

```
<Axes: xlabel='Sex', ylabel='Survived'>
```



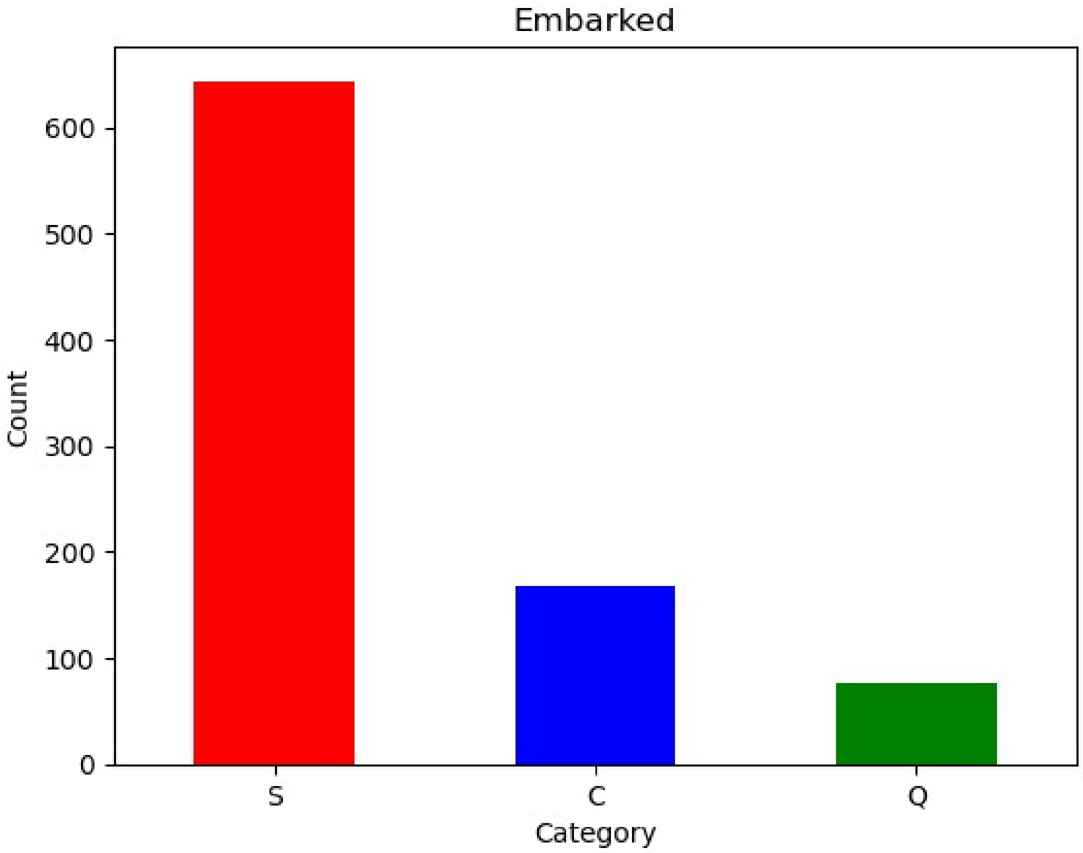
```
titanic_data[['Sex', 'Survived']].groupby(['Sex'],
as_index=False).mean().sort_values(by='Survived', ascending=False)
```

```
      Sex  Survived
0  female  0.742038
1    male  0.188908
```

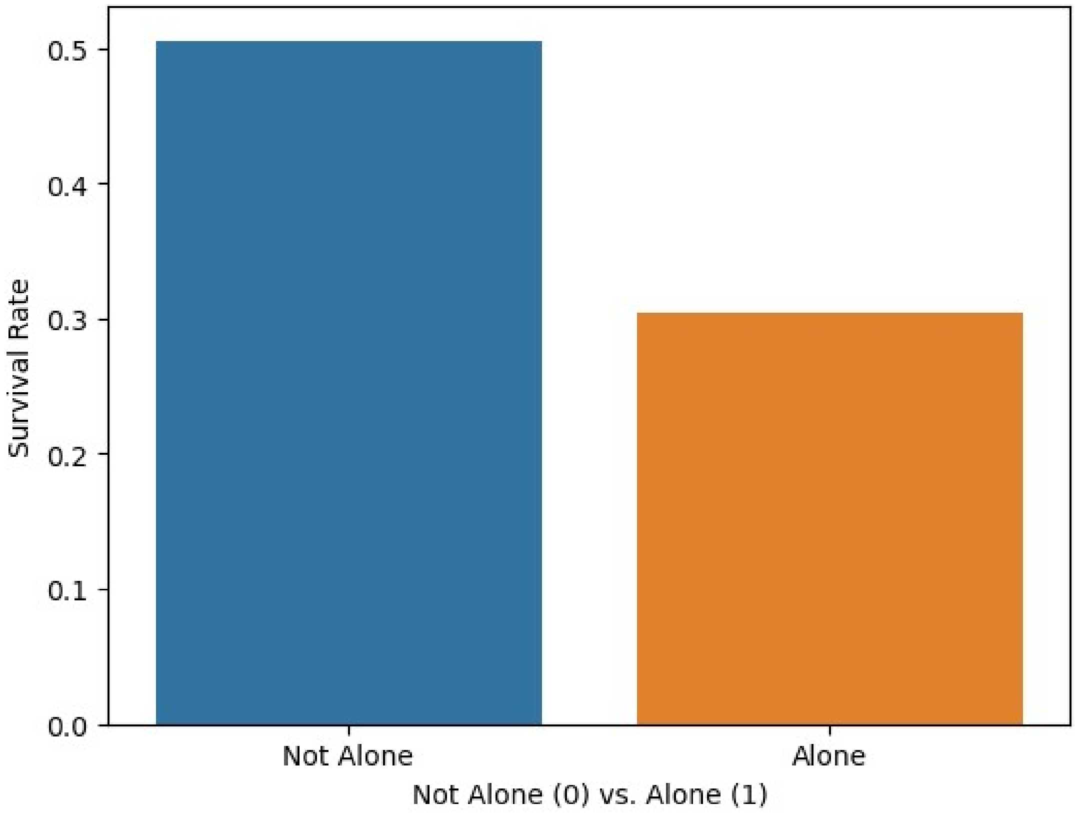
```
titanic_data[['Pclass', 'Survived']].groupby(['Pclass'],
as_index=False).mean().sort_values(by='Survived', ascending=False)
```

```
   Pclass  Survived
0       1  0.629630
1       2  0.472826
2       3  0.242363
```

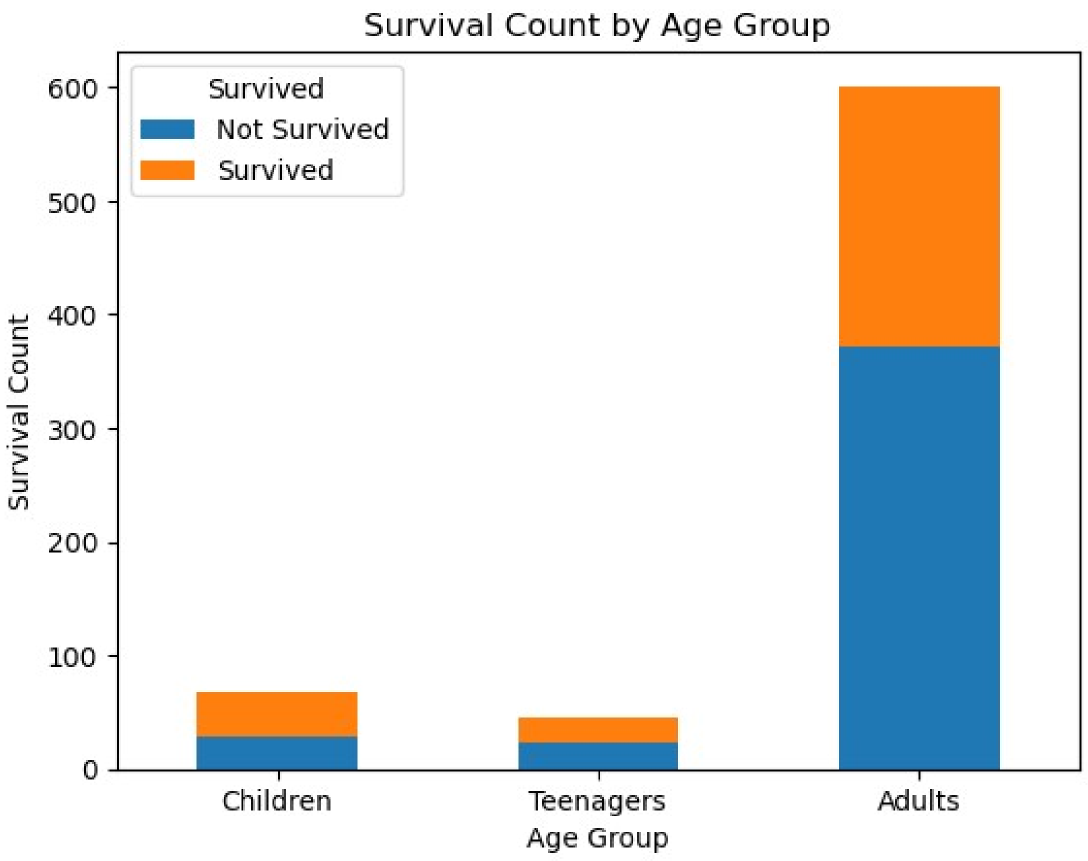
```
titanic_data['Embarked'].value_counts().plot(kind='bar', rot=0,
color=['red', 'blue', 'green'])
plt.title('Embarked')
plt.xlabel('Category')
plt.ylabel('Count')
plt.show()
```



```
travelling_partners = titanic_data['SibSp'] + titanic_data['Parch']
travelled_alone = np.where(travelling_partners > 0, 0, 1)

survival_rates = titanic_data.groupby(travelled_alone)
['Survived'].mean()

sns.barplot(x=survival_rates.index, y=survival_rates.values)
plt.xlabel('Not Alone (0) vs. Alone (1)')
plt.ylabel('Survival Rate')
plt.xticks([0, 1], ['Not Alone', 'Alone'])
plt.show()
```



```python
bins = [0, 12, 18, 100]
labels = ['Children', 'Teenagers', 'Adults']

age_groups = pd.cut(titanic_data['Age'], bins=bins, labels=labels,
right=False)
pivot_table = pd.crosstab(index=age_groups,
columns=titanic_data['Survived'])

# Plot the graph
ax = pivot_table.plot(kind='bar', stacked=True)
ax.set_xlabel('Age Group')
ax.set_ylabel('Survival Count')
ax.set_title('Survival Count by Age Group')
plt.xticks(rotation=0)
plt.legend(title='Survived', labels=['Not Survived', 'Survived'])
plt.show()
```



Data Cleaning

```
titanic_data['Age'].fillna(round(titanic_data['Age'].mean()),inplace=T
rue)
titanic_data['Embarked'].fillna('S',inplace=True)
titanic_data['Cabin'].fillna('C85', inplace=True)
titanic_data.head()

   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   C85        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   C85        S
3      0             113803  53.1000  C123        S
4      0             373450   8.0500   C85        S
```

```
cat_sex={"male":0,"female":1}
titanic_data["Sex"]=titanic_data["Sex"].map(cat_sex)

titanic_data.head()
```

```
   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
Parch  \
0                            Braund, Mr. Owen Harris    0  22.0      1
0
1  Cumings, Mrs. John Bradley (Florence Briggs Th...    1  38.0      1
0
2                             Heikkinen, Miss. Laina    1  26.0      0
0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)    1  35.0      1
0
4                           Allen, Mr. William Henry    0  35.0      0
0

             Ticket      Fare Cabin Embarked
0         A/5 21171   7.2500   C85        S
1          PC 17599  71.2833   C85        C
2  STON/O2. 3101282   7.9250   C85        S
3             113803  53.1000  C123        S
4             373450   8.0500   C85        S
```

```
cat_Embarked={"S":0,"C":1,"Q":2}
titanic_data["Embarked"]=titanic_data["Embarked"].map(cat_Embarked)
titanic_data.head()
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
```

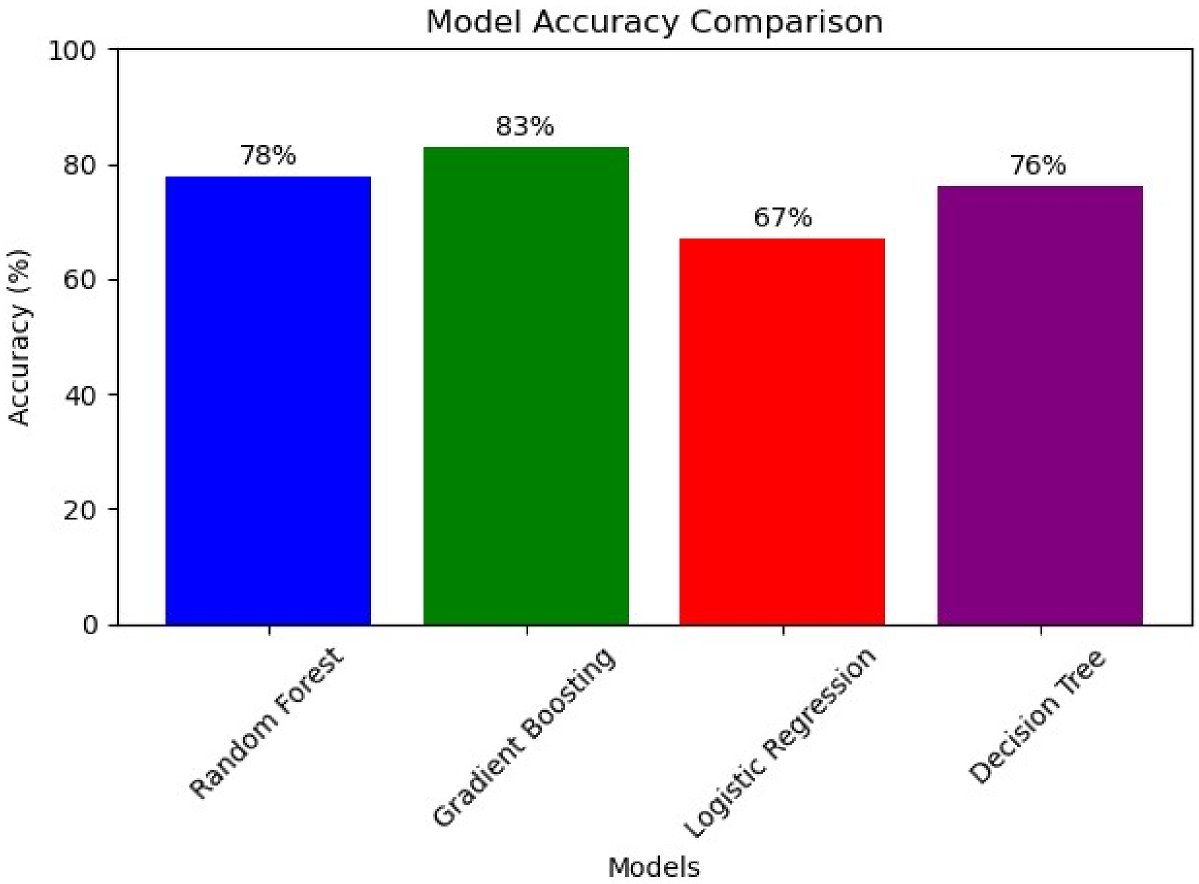
```
1            2          1       1
2            3          1       3
3            4          1       1
4            5          0       3

                                                Name  Sex   Age  SibSp  
Parch  \
0                            Braund, Mr. Owen Harris    0  22.0      1   
0   
1  Cumings, Mrs. John Bradley (Florence Briggs Th...    1  38.0      1   
0   
2                             Heikkinen, Miss. Laina    1  26.0      0   
0   
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)    1  35.0      1   
0   
4                           Allen, Mr. William Henry    0  35.0      0   
0   

             Ticket     Fare Cabin  Embarked  
0         A/5 21171   7.2500   C85         0  
1          PC 17599  71.2833   C85         1  
2  STON/O2. 3101282   7.9250   C85         0  
3            113803  53.1000  C123         0  
4            373450   8.0500   C85         0  
```

```
titanic_data.isnull().sum()
```

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

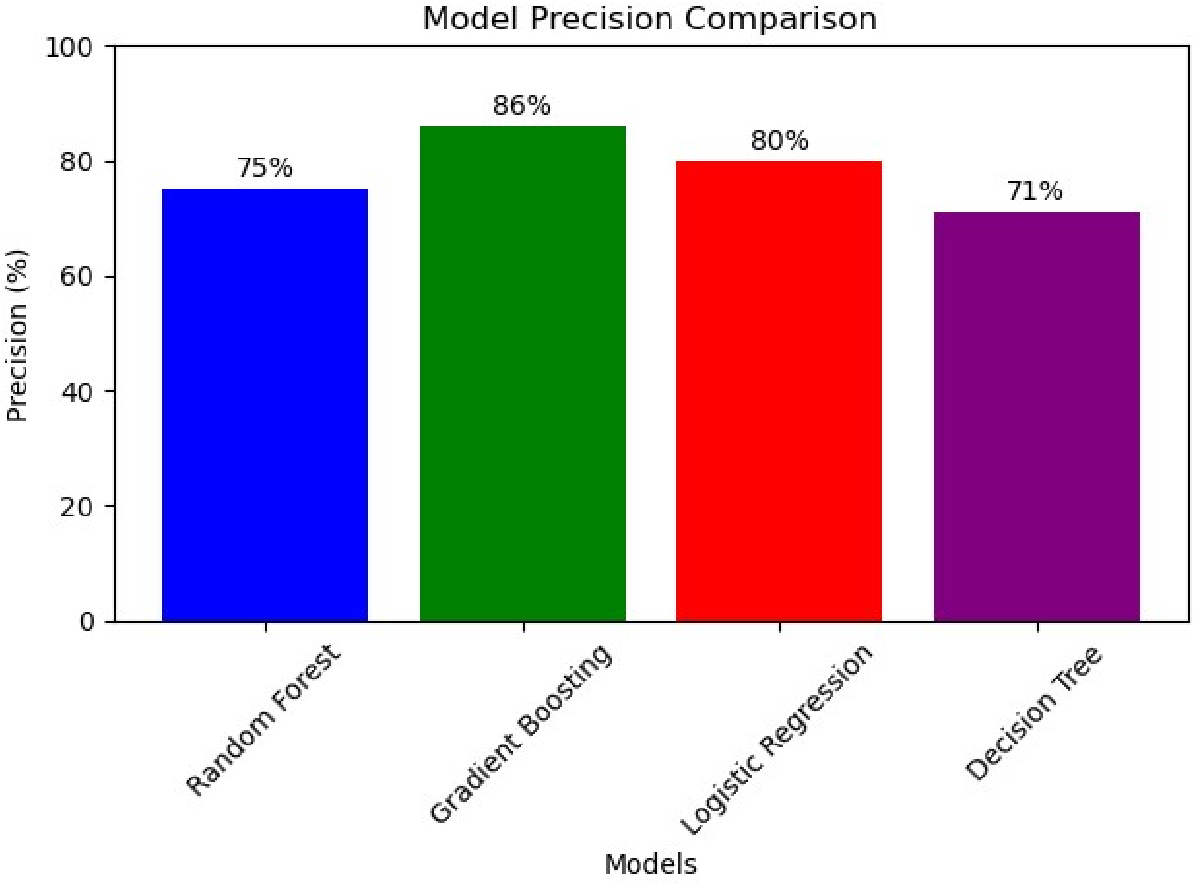
Model Comparision

```
models = ['Random Forest', 'Gradient Boosting', 'Logistic Regression', 
'Decision Tree']
accuracies = [78, 83, 67, 76]

plt.bar(models, accuracies, color=['blue', 'green', 'red', 'purple'])
plt.xlabel('Models')
plt.ylabel('Accuracy (%)')
plt.title('Model Accuracy Comparison')
plt.ylim(0, 100)
plt.xticks(rotation=45)
for i, accuracy in enumerate(accuracies):
    plt.text(i, accuracy + 1, f'{accuracy}%', ha='center',
va='bottom', fontsize=10)
plt.tight_layout()
plt.show()
```

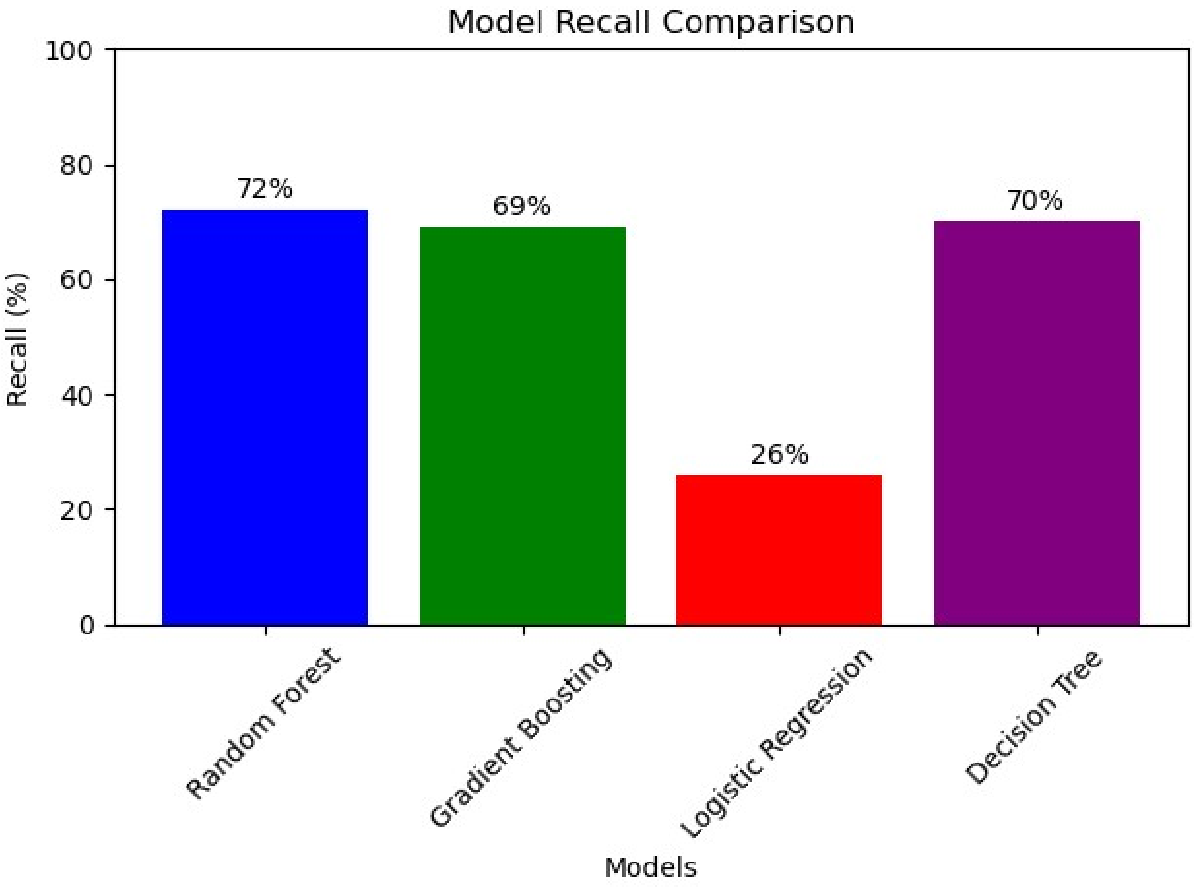


```
models = ['Random Forest', 'Gradient Boosting', 'Logistic Regression',
'Decision Tree']
Precisions = [75, 86, 80, 71]

plt.bar(models, Precisions, color=['blue', 'green', 'red', 'purple'])
plt.xlabel('Models')
plt.ylabel('Precision (%)')
plt.title('Model Precision Comparison')
plt.ylim(0, 100)
plt.xticks(rotation=45)
```

```
for i, Precision in enumerate(Precisions):
    plt.text(i, Precision + 1, f'{Precision}%', ha='center',
va='bottom', fontsize=10)
plt.tight_layout()
plt.show()
```



```
models = ['Random Forest', 'Gradient Boosting', 'Logistic Regression',
'Decision Tree']
Recall = [72, 69, 26, 70]

plt.bar(models, Recall, color=['blue', 'green', 'red', 'purple'])
plt.xlabel('Models')
plt.ylabel('Recall (%)')
plt.title('Model Recall Comparison')
plt.ylim(0, 100)
plt.xticks(rotation=45)
for i, R in enumerate(Recall):
    plt.text(i, R + 1, f'{R}%', ha='center', va='bottom', fontsize=10)
plt.tight_layout()
plt.show()
```


Model Recall Comparison
100
80
60
40
20
0
Recall (%)
72%
69%
26%
70%
Random Forest
Gradient Boosting
Logistic Regression
Decision Tree
Models