```
import pandas as pd
import numpy as  np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
from google.colab import files
uploaded=files.upload()
```

Choose Files titanic.csv

- **titanic.csv**(text/csv) - 61192 bytes, last modified: 5/19/2024 - 100% done

```
Saving titanic.csv to titanic.csv
```

```
#load the titanic dataset
titanic_df = pd.read_csv('titanic.csv')
titanic_df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53

```
#display the first few rows of the dataset
print(titanic_df.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   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
```

```
#check the data types and missing values
print(titanic_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
None
```

```
#summary statistics
print(titanic_df.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
```

```
#drop unnecessary columns
titanic_df=titanic_df.drop(['PassengerId','Name','Ticket','Cabin'],axis=1)
titanic_df
```

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

891 rows × 8 columns

```
#fill missing values in the age column with the median age
titanic_df['Age'].fillna(titanic_df['Age'].median(),inplace=True)
titanic_df
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S
...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S
887	1	1	female	19.0	0	0	30.0000	S
888	0	3	female	28.0	1	2	23.4500	S
889	1	1	male	26.0	0	0	30.0000	C
890	0	3	male	32.0	0	0	7.7500	Q

891 rows × 8 columns

```
#fill missing values in the embarked column with the mode
mode_embarked=titanic_df['Embarked'].mode()[0]
titanic_df['Embarked'].fillna(mode_embarked,inplace=True)
titanic_df
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S
...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S
887	1	1	female	19.0	0	0	30.0000	S
888	0	3	female	28.0	1	2	23.4500	S
889	1	1	male	26.0	0	0	30.0000	C
890	0	3	male	32.0	0	0	7.7500	Q

891 rows × 8 columns

```
#Convert categorical variables into dummy\indication variables
titanic_df=pd.get_dummies(titanic_df,columns=['Sex','Embarked'],drop_first=True)
titanic_df
```

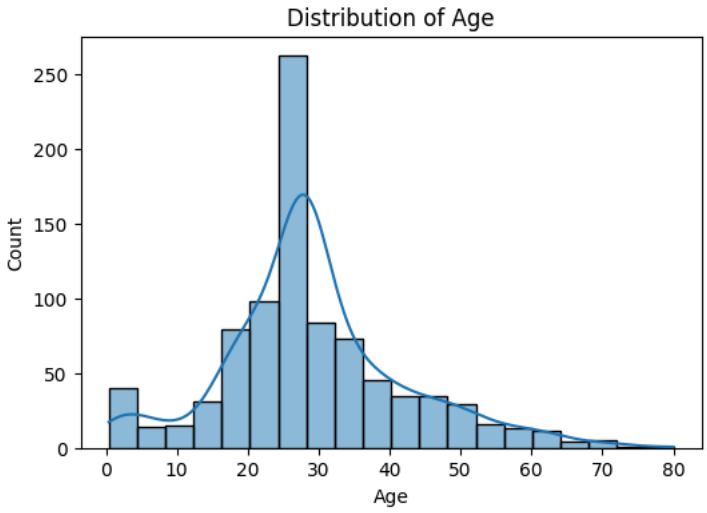
	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_male	Embarked_Q	Embarked_S
0	0	3	22.0	1	0	7.2500	True	False	True
1	1	1	38.0	1	0	71.2833	False	False	False
2	1	3	26.0	0	0	7.9250	False	False	True
3	1	1	35.0	1	0	53.1000	False	False	True
4	0	3	35.0	0	0	8.0500	True	False	True
...	...	...	...	...	...	...	...	...	...
886	0	2	27.0	0	0	13.0000	True	False	True
887	1	1	19.0	0	0	30.0000	False	False	True
888	0	3	28.0	1	2	23.4500	False	False	True
889	1	1	26.0	0	0	30.0000	True	False	False
890	0	3	32.0	0	0	7.7500	True	True	False

891 rows × 9 columns

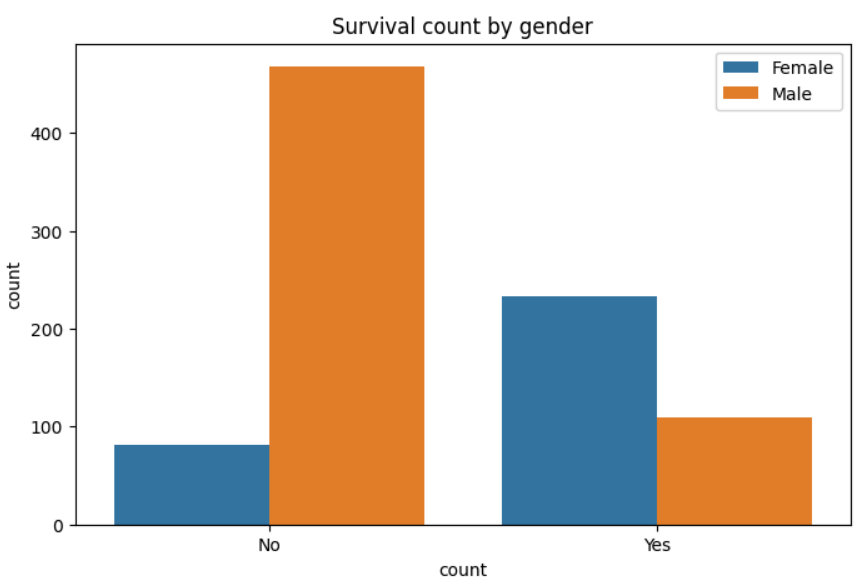
```
#check for any remaining missing values
print(titanic_df.isnull().sum())
```

```
Survived      0
Pclass        0
Age           0
SibSp         0
Parch         0
Fare          0
Sex_male      0
Embarked_Q    0
Embarked_S    0
dtype: int64
```

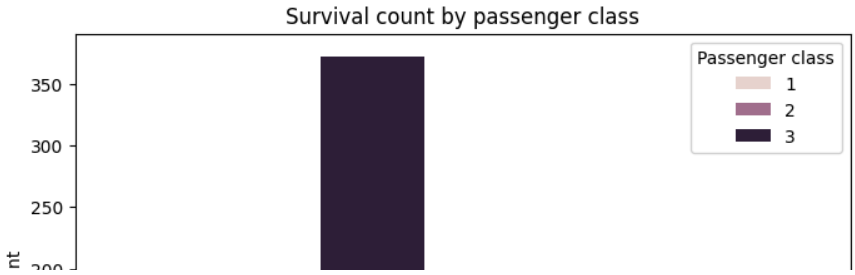
```
#visualize the distribution of Age
plt.figure(figsize=(6,4))
sns.histplot(titanic_df['Age'],bins=20,kde=True)
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```




```
#Explore the survival rate by gender
plt.figure(figsize=(8,5))
sns.countplot(x='Survived',hue='Sex_male',data=titanic_df)
plt.title('Survival count by gender')
plt.xlabel('count')
plt.xticks([0,1],['No','Yes'])
plt.legend(['Female','Male'])
plt.show()
```




```
#explore the survival rate by passenger class
plt.figure(figsize=(8,5))
sns.countplot(x='Survived',hue='Pclass',data=titanic_df)
plt.title('Survival count by passenger class')
plt.xlabel('Survived')
plt.ylabel('Count')
plt.xticks([0,1],['No','Yes'])
plt.legend(title='Passenger class')
plt.show()
```




```
#Explore the relationship between fare and survival
plt.figure(figsize=(10,6))
sns.boxplot(x='Survived',y='Fare',data=titanic_df)
plt.title('survival by fare')
plt.xlabel('survived')
plt.ylabel('fare')
plt.xticks([0,1],['no','yes'])
plt.show()
```