Performing Data Cleaning and Analysis

In [1]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
Titanic = pd.read_csv(r"D:\Data Science\SEP\sep 17th - ML\TITANIC PROJECT\DATASET\titanic dataset.csv")
```

In [5]:
```
Titanic.tail()
```

Out[5]:

	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

Performing Data Cleaning and Analysis

1.Understanding meaning of each column:

- Data Dictionary: Variable Description
- Survived - Survived (1) or died (0)
- Pclass - Passenger's class (1 = 1st, 2 = 2nd, 3 = 3rd)
- Name - Passenger's name
- Sex - Passenger's gender (male/female)
- Age - Passenger's age
- SibSp - Number of siblings/spouses aboard

- Parch - Number of parents/children aboard (Some children travelled only with a nanny,therefore parch=0 for them.)
- Ticket - Ticket number
- Fare - Fare
- Cabin - Cabin
- Embarked - Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

2. Analysing which columns are completely useless in predicting the survival and deleting them

Note - Don't just delete the columns because you are not finding it useful. Or focus is not on deleting the columns. Our focus is on analysing how each column is affecting the result or the prediction and in accordance with that deciding whether to keep the column or to delete the column or fill the null values of the column by some values and if yes, then what values.

In [8]:
```
Titanic.describe()
```

Out[8]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [10]:
```
Titanic.columns
```

Out[10]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [12]:
```
#Name column can never decide survival of a person, hence we can safely delete it
del Titanic["Name"]
Titanic.head()
```

Out[12]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	male	35.0	0	0	373450	8.0500	NaN	S

In [14]:
```
del Titanic["Ticket"]
Titanic.head()
```

Out[14]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	1	0	3	male	22.0	1	0	7.2500	NaN	S
1	2	1	1	female	38.0	1	0	71.2833	C85	C
2	3	1	3	female	26.0	0	0	7.9250	NaN	S
3	4	1	1	female	35.0	1	0	53.1000	C123	S
4	5	0	3	male	35.0	0	0	8.0500	NaN	S

In [16]:
```
del Titanic["Fare"]
Titanic.head()
```

Out[16]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Cabin	Embarked
0	1	0	3	male	22.0	1	0	NaN	S
1	2	1	1	female	38.0	1	0	C85	C
2	3	1	3	female	26.0	0	0	NaN	S
3	4	1	1	female	35.0	1	0	C123	S
4	5	0	3	male	35.0	0	0	NaN	S

In [18]:
```
del Titanic['Cabin']
Titanic.head()
```

Out[18]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Embarked
0	1	0	3	male	22.0	1	0	S
1	2	1	1	female	38.0	1	0	C
2	3	1	3	female	26.0	0	0	S
3	4	1	1	female	35.0	1	0	S
4	5	0	3	male	35.0	0	0	S

In [20]:
```
# Changing Value for "Male, Female" string values to numeric values , male=1 and female=2
def getNumber(str):
    if str=="male":
        return 1
    else:
        return 2
Titanic["Gender"]=Titanic["Sex"].apply(getNumber)
#We have created a new column called "Gender" and
#filling it with values 1,2 based on the values of sex column
Titanic.head()
```

Out[20]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Embarked	Gender
0	1	0	3	male	22.0	1	0	S	1
1	2	1	1	female	38.0	1	0	C	2
2	3	1	3	female	26.0	0	0	S	2
3	4	1	1	female	35.0	1	0	S	2
4	5	0	3	male	35.0	0	0	S	1

In [22]:
```
#Deleting Sex column, since no use of it now
del Titanic["Sex"]
Titanic.head()
```

Out[22]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Embarked	Gender
0	1	0	3	22.0	1	0	S	1
1	2	1	1	38.0	1	0	C	2
2	3	1	3	26.0	0	0	S	2
3	4	1	1	35.0	1	0	S	2
4	5	0	3	35.0	0	0	S	1

In [24]:
```
Titanic.isnull().sum()
```

Out[24]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
dtype: int64
```

In [26]:
```
Titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 8 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   Age          714 non-null    float64
 4   SibSp        891 non-null    int64
 5   Parch        891 non-null    int64
 6   Embarked     889 non-null    object
 7   Gender       891 non-null    int64
dtypes: float64(1), int64(6), object(1)
memory usage: 55.8+ KB
```

Here we see age has some null values

Fill the null values of the Age column.

- Fill mean Survived age(mean age of the survived people) in the column where the person has survived and mean not Survived age (mean age of the people who have not survived) in the column where person has not survived

In [30]:
```
Mean= Titanic[Titanic.Survived==1].Age.mean()
Mean
```

Out[30]:
```
28.343689655172415
```

Creating a new "Age" column ,

- filling values in it with a condition if goes True then given values (here Mean) is put in place of last values else nothing happens, simply the values are copied from the "Age" column of the dataset###

In [33]:
```
Titanic["age"]=np.where(pd.isnull(Titanic.Age) & Titanic["Survived"]==1  ,Mean, Titanic["Age"])
Titanic.head()
```

Out[33]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Embarked	Gender	age
0	1	0	3	22.0	1	0	S	1	22.0
1	2	1	1	38.0	1	0	C	2	38.0
2	3	1	3	26.0	0	0	S	2	26.0
3	4	1	1	35.0	1	0	S	2	35.0
4	5	0	3	35.0	0	0	S	1	35.0

In [35]:
```
Titanic.isnull().sum()
```

Out[35]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
age            125
dtype: int64
```

In [37]:
```
# Finding the mean age of "Not Survived" people
meanNS=Titanic[Titanic.Survived==0].Age.mean()
meanNS
```

Out[37]:
```
30.62617924528302
```

In [39]:
```
Titanic.age.fillna(meanNS,inplace=True)
Titanic.head()
```

```
C:\Users\velug\AppData\Local\Temp\ipykernel_12500\525542227.py:1: FutureWarning: A value is trying to be set on a copy of a DataFr
ame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are settin
g values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[c
ol].method(value) instead, to perform the operation inplace on the original object.


  Titanic.age.fillna(meanNS,inplace=True)
```

Out[39]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Embarked	Gender	age
0	1	0	3	22.0	1	0	S	1	22.0
1	2	1	1	38.0	1	0	C	2	38.0
2	3	1	3	26.0	0	0	S	2	26.0
3	4	1	1	35.0	1	0	S	2	35.0
4	5	0	3	35.0	0	0	S	1	35.0

In [41]:
```
Titanic.isnull().sum()
```

Out[41]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
age              0
dtype: int64
```

In [43]:
```
del Titanic['Age']
Titanic.head()
```

Out[43]:

	PassengerId	Survived	Pclass	SibSp	Parch	Embarked	Gender	age
0	1	0	3	1	0	S	1	22.0
1	2	1	1	1	0	C	2	38.0
2	3	1	3	0	0	S	2	26.0
3	4	1	1	1	0	S	2	35.0
4	5	0	3	0	0	S	1	35.0

In [45]:
```
import warnings
warnings.filterwarnings('ignore')
```

We want to check if "Embarked" column is is important for analysis or not, that is whether survival of the person depends on the Embarked column value or not###

In [48]:
```
# Finding the number of people who have survived
# given that they have embarked or boarded from a particular port

survivedQ = Titanic[Titanic.Embarked == 'Q'][Titanic.Survived == 1].shape[0]
survivedC = Titanic[Titanic.Embarked == 'C'][Titanic.Survived == 1].shape[0]
survivedS = Titanic[Titanic.Embarked == 'S'][Titanic.Survived == 1].shape[0]
print(survivedQ)
print(survivedC)
print(survivedS)
```

```
30
93
217
```

In [50]:
```
survivedQ = Titanic[Titanic.Embarked == 'Q'][Titanic.Survived == 0].shape[0]
survivedC = Titanic[Titanic.Embarked == 'C'][Titanic.Survived == 0].shape[0]
survivedS = Titanic[Titanic.Embarked == 'S'][Titanic.Survived == 0].shape[0]
print(survivedQ)
print(survivedC)
print(survivedS)
```

```
47
75
427
```

As there are significant changes in the survival rate based on which port the passengers aboard the ship. We cannot delete the whole embarked column(It is useful). Now the Embarked column has some null values in it and hence we can safely say that deleting some rows from total rows will not affect the result. So rather than trying to fill those null values with some vales. We can simply remove them.

In [53]:
```
Titanic.dropna(inplace=True)
Titanic.head()
```

Out[53]:

	PassengerId	Survived	Pclass	SibSp	Parch	Embarked	Gender	age
0	1	0	3	1	0	S	1	22.0
1	2	1	1	1	0	C	2	38.0
2	3	1	3	0	0	S	2	26.0
3	4	1	1	1	0	S	2	35.0
4	5	0	3	0	0	S	1	35.0

In [55]:
```
Titanic.isnull().sum()
```

Out[55]:
```
PassengerId    0
Survived       0
Pclass         0
SibSp          0
Parch          0
Embarked       0
Gender         0
age            0
dtype: int64
```

In [57]:
```
#Renaming "age" and "gender" columns
Titanic.rename(columns={'age':'Age'}, inplace=True)
Titanic.head()
```

Out[57]:

	PassengerId	Survived	Pclass	SibSp	Parch	Embarked	Gender	Age
0	1	0	3	1	0	S	1	22.0
1	2	1	1	1	0	C	2	38.0
2	3	1	3	0	0	S	2	26.0
3	4	1	1	1	0	S	2	35.0
4	5	0	3	0	0	S	1	35.0

In [59]:
```
Titanic.rename(columns={'Gender':'Sex'}, inplace=True)
Titanic.head()
```

Out[59]:

	PassengerId	Survived	Pclass	SibSp	Parch	Embarked	Sex	Age
0	1	0	3	1	0	S	1	22.0
1	2	1	1	1	0	C	2	38.0
2	3	1	3	0	0	S	2	26.0
3	4	1	1	1	0	S	2	35.0
4	5	0	3	0	0	S	1	35.0

In [61]:
```
def getEmb(str):
    if str=="S":
        return 1
    elif str=='Q':
        return 2
    else:
        return 3
Titanic["Embark"]=Titanic["Embarked"].apply(getEmb)
Titanic.head()
```

Out[61]:

	PassengerId	Survived	Pclass	SibSp	Parch	Embarked	Sex	Age	Embark
0	1	0	3	1	0	S	1	22.0	1
1	2	1	1	1	0	C	2	38.0	3
2	3	1	3	0	0	S	2	26.0	1
3	4	1	1	1	0	S	2	35.0	1
4	5	0	3	0	0	S	1	35.0	1

In [63]:
```
del Titanic['Embarked']
Titanic.rename(columns={'Embark':'Embarked'}, inplace=True)
Titanic.head()
```

Out[63]:

	PassengerId	Survived	Pclass	SibSp	Parch	Sex	Age	Embarked
0	1	0	3	1	0	1	22.0	1
1	2	1	1	1	0	2	38.0	3
2	3	1	3	0	0	2	26.0	1
3	4	1	1	1	0	2	35.0	1
4	5	0	3	0	0	1	35.0	1

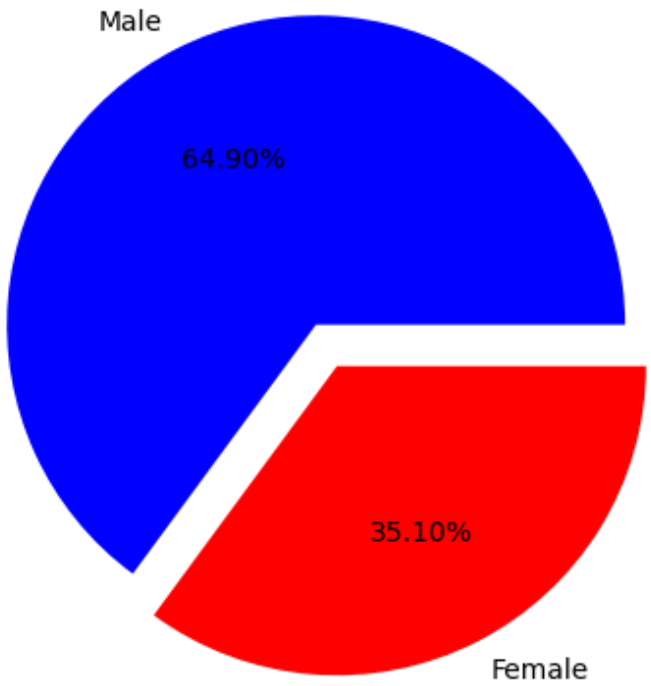
PIE CHART

In [73]:
```
#Drawing a pie chart for number of males and females aboard
import matplotlib.pyplot as plt
from matplotlib import style

males = (Titanic['Sex'] == 1).sum()
#Summing up all the values of column gender with a
#condition for male and similary for females
females = (Titanic['Sex'] == 2).sum()
print(males)
```

```
print(females)
p = [males, females]
plt.pie(p,    #giving array
       labels = ['Male', 'Female'], #Correspndingly giving labels
       colors = ["blue",'red'],   # Corresponding colors
       explode = (0.15, 0),    #How much the gap should me there between the pies
       startangle =0,
       autopct="%.2f%%")  #what start angle should be given
plt.axis('equal')
plt.show()
```

```
577
312
```

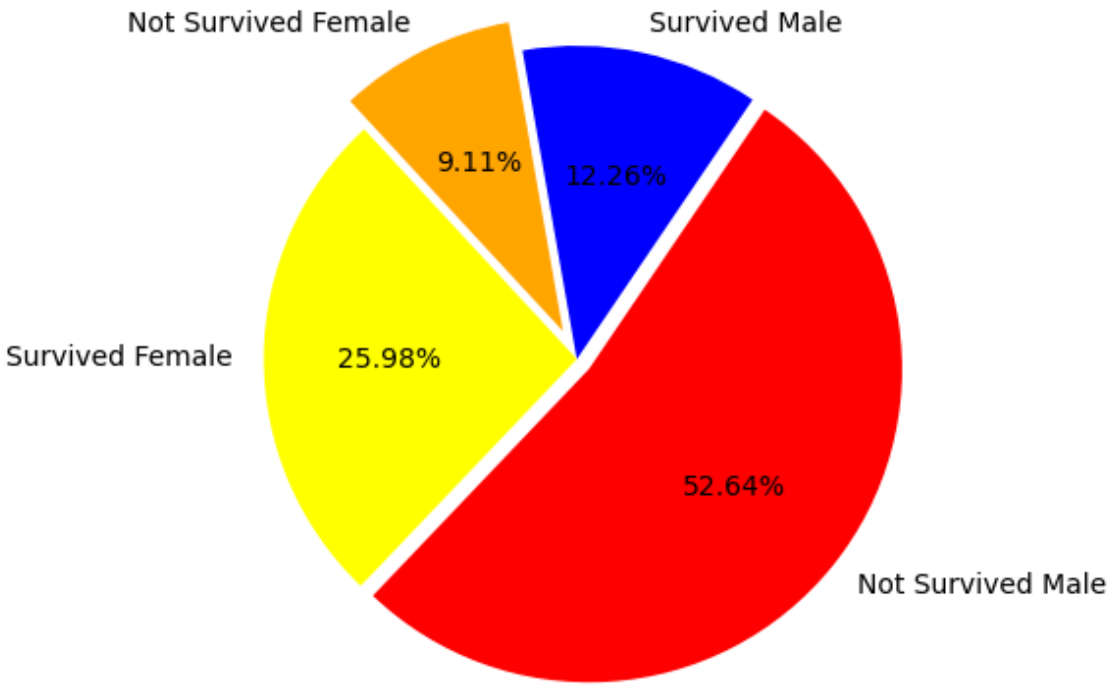


In [69]:
```
# More Precise Pie Chart
MaleS=Titanic[Titanic.Sex==1][Titanic.Survived==1].shape[0]
print(MaleS)
MaleN=Titanic[Titanic.Sex==1][Titanic.Survived==0].shape[0]
print(MaleN)
```

```
FemaleS=Titanic[Titanic.Sex==2][Titanic.Survived==1].shape[0]
print(FemaleS)
FemaleN=Titanic[Titanic.Sex==2][Titanic.Survived==0].shape[0]
print(FemaleN)
```

```
109
468
231
81
```

In [71]:
```
chart=[MaleS,MaleN,FemaleS,FemaleN]
colors=['blue','red','Yellow','Orange']
labels=["Survived Male","Not Survived Male","Survived Female","Not Survived Female"]
explode=[0,0.05,0,0.1]
plt.pie(chart,labels=labels,colors=colors,explode=explode,startangle=100,counterclock=False,autopct="%.2f%%")
plt.axis("equal")
plt.show()
```



In [ ]: