In [7]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
tat=pd.read_csv(r"C:\Users\user\Downloads\10th\10th\project - data preprocessing\trai
tat.describe()
```

Out[7]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Far
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 [8]:

```
tat.head()
```

Out[8]:

	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	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500

In [9]:

```
tat.tail()
```

Out[9]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	I
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	I
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	I

In [10]:

```
del tat["Name"]
del tat["Ticket"]
del tat['Cabin']
del tat['Fare']
```

In [11]:

```
tat
```

Out[11]:

	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
...	...	...	...	...	...	...	...	...
886	887	0	2	male	27.0	0	0	S
887	888	1	1	female	19.0	0	0	S
888	889	0	3	female	NaN	1	2	S
889	890	1	1	male	26.0	0	0	C
890	891	0	3	male	32.0	0	0	Q

891 rows × 8 columns

In [13]:

```
def getNumber(str):
    if str=='male':
        return 1
    else:
        return 0
tat['Gender'] =tat['Sex'].apply(getNumber)
tat.head()
```

Out[13]:

	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	0
2	3	1	3	female	26.0	0	0	S	0
3	4	1	1	female	35.0	1	0	S	0
4	5	0	3	male	35.0	0	0	S	1

In [14]:

```
del tat['Sex']
tat.head()
```

Out[14]:

	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	0
2	3	1	3	26.0	0	0	S	0
3	4	1	1	35.0	1	0	S	0
4	5	0	3	35.0	0	0	S	1

In [15]:

```
tat.isnull().sum()
```

Out[15]:

```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
dtype: int64
```

In [29]:

```
avgS=tat[tat.Survived==1].Age.mean()
avgS
```

Out[29]:

```
28.343689655172412
```

In [28]:

```
tat['age']=np.where(pd.isnull(tat.Age) & tat["Survived"]==1  ,avgS, tat["Age"])
tat.head()
```

Out[28]:

	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	0	38.0
2	3	1	3	26.0	0	0	S	0	26.0
3	4	1	1	35.0	1	0	S	0	35.0
4	5	0	3	35.0	0	0	S	1	35.0

In [21]:

```
tat.tail()
```

Out[21]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Embarked	Gender	age
886	887	0	2	27.0	0	0	S	1	27.0
887	888	1	1	19.0	0	0	S	0	19.0
888	889	0	3	NaN	1	2	S	0	NaN
889	890	1	1	26.0	0	0	C	1	26.0
890	891	0	3	32.0	0	0	Q	1	32.0

In [25]:

```
tat.head(18)
```

Out[25]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Embarked	Gender	age
0	1	0	3	22.00000	1	0	S	1	22.00000
1	2	1	1	38.00000	1	0	C	0	38.00000
2	3	1	3	26.00000	0	0	S	0	26.00000
3	4	1	1	35.00000	1	0	S	0	35.00000
4	5	0	3	35.00000	0	0	S	1	35.00000
5	6	0	3	NaN	0	0	Q	1	NaN
6	7	0	1	54.00000	0	0	S	1	54.00000
7	8	0	3	2.00000	3	1	S	1	2.00000
8	9	1	3	27.00000	0	2	S	0	27.00000
9	10	1	2	14.00000	1	0	C	0	14.00000
10	11	1	3	4.00000	1	1	S	0	4.00000
11	12	1	1	58.00000	0	0	S	0	58.00000
12	13	0	3	20.00000	0	0	S	1	20.00000
13	14	0	3	39.00000	1	5	S	1	39.00000
14	15	0	3	14.00000	0	0	S	0	14.00000
15	16	1	2	55.00000	0	0	S	0	55.00000
16	17	0	3	2.00000	4	1	Q	1	2.00000
17	18	1	2	28.34369	0	0	S	1	28.34369

In [26]:

```
tat.isnull().sum()
```

Out[26]:

```
PassengerId      0
Survived         0
Pclass           0
Age            125
SibSp            0
Parch            0
Embarked         2
Gender           0
age            125
dtype: int64
```

In [30]:

```
avgNS=tat[tat.Survived==0].Age.mean()
avgNS
```

Out[30]:

```
30.62617924528302
```

In [31]:

```
tat.age.fillna(avgNS,inplace=True)
tat.tail()
```

Out[31]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Embarked	Gender	age
886	887	0	2	27.0	0	0	S	1	27.000000
887	888	1	1	19.0	0	0	S	0	19.000000
888	889	0	3	NaN	1	2	S	0	30.626179
889	890	1	1	26.0	0	0	C	1	26.000000
890	891	0	3	32.0	0	0	Q	1	32.000000

In [32]:

```
tat.head()
```

Out[32]:

	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	0	38.0
2	3	1	3	26.0	0	0	S	0	26.0
3	4	1	1	35.0	1	0	S	0	35.0
4	5	0	3	35.0	0	0	S	1	35.0

In [33]:

```
tat.isnull().sum()
```

Out[33]:

```
PassengerId      0
Survived         0
Pclass           0
Age            125
SibSp            0
Parch            0
Embarked         2
Gender           0
age              0
dtype: int64
```

In [34]:

```
del tat['Age']
tat.head()
```

Out[34]:

	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	0	38.0
2	3	1	3	0	0	S	0	26.0
3	4	1	1	1	0	S	0	35.0
4	5	0	3	0	0	S	1	35.0

In [35]:

```
srvQ = tat[tat.Embarked == 'Q'][tat.Survived == 1].shape[0]
srvC = tat[tat.Embarked == 'C'][tat.Survived == 1].shape[0]
srvS = tat[tat.Embarked == 'S'][tat.Survived == 1].shape[0]
print(srvQ)
print(srvC)
print(srvS)
```

```
30
93
217

C:\Users\user\AppData\Local\Temp\ipykernel_29924\3375765492.py:1: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  srvQ = tat[tat.Embarked == 'Q'][tat.Survived == 1].shape[0]
C:\Users\user\AppData\Local\Temp\ipykernel_29924\3375765492.py:2: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  srvC = tat[tat.Embarked == 'C'][tat.Survived == 1].shape[0]
C:\Users\user\AppData\Local\Temp\ipykernel_29924\3375765492.py:3: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  srvS = tat[tat.Embarked == 'S'][tat.Survived == 1].shape[0]
```

In [36]:

```
srvQ = tat[tat.Embarked == 'Q'][tat.Survived == 0].shape[0]
srvC = tat[tat.Embarked == 'C'][tat.Survived == 0].shape[0]
srvS = tat[tat.Embarked == 'S'][tat.Survived == 0].shape[0]
print(srvQ)
print(srvC)
print(srvS)
```

```
47
75
427

C:\Users\user\AppData\Local\Temp\ipykernel_29924\1280627143.py:1: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  srvQ = tat[tat.Embarked == 'Q'][tat.Survived == 0].shape[0]
C:\Users\user\AppData\Local\Temp\ipykernel_29924\1280627143.py:2: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  srvC = tat[tat.Embarked == 'C'][tat.Survived == 0].shape[0]
C:\Users\user\AppData\Local\Temp\ipykernel_29924\1280627143.py:3: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  srvS = tat[tat.Embarked == 'S'][tat.Survived == 0].shape[0]
```

In [37]:

```
tat.dropna(inplace=True)
tat.head()
```

Out[37]:

	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	0	38.0
2	3	1	3	0	0	S	0	26.0
3	4	1	1	1	0	S	0	35.0
4	5	0	3	0	0	S	1	35.0

In [38]:

```
tat.isnull().sum()
```

Out[38]:

```
PassengerId    0
Survived       0
Pclass         0
SibSp          0
Parch          0
Embarked       0
Gender         0
age            0
dtype: int64
```

In [43]:

```
tat.rename(columns={'age':'Age'},inplace=True)
tat.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	0	38.0
2	3	1	3	0	0	S	0	26.0
3	4	1	1	1	0	S	0	35.0
4	5	0	3	0	0	S	1	35.0

In [44]:

```
tat.rename(columns={'Gender':'Sex'},inplace=True)
tat.head()
```

Out[44]:

	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	0	38.0
2	3	1	3	0	0	S	0	26.0
3	4	1	1	1	0	S	0	35.0
4	5	0	3	0	0	S	1	35.0

In [45]:

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

Out[45]:

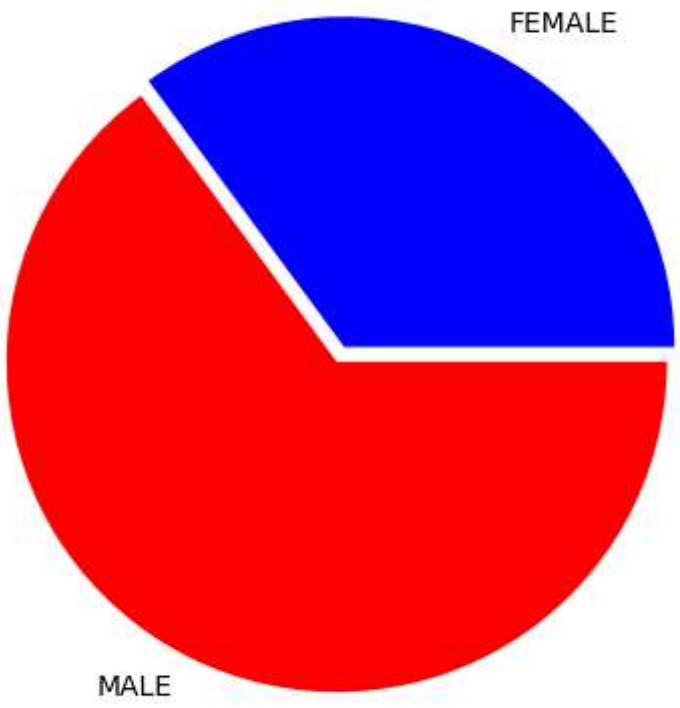
	PassengerId	Survived	Pclass	SibSp	Parch	Embarked	Sex	Age	Embark
0	1	0	3	1	0	S	1	22.0	1
1	2	1	1	1	0	C	0	38.0	3
2	3	1	3	0	0	S	0	26.0	1
3	4	1	1	1	0	S	0	35.0	1
4	5	0	3	0	0	S	1	35.0	1

In [56]:

```
from matplotlib import style

males = (tat['Sex'] == 1).sum()
females = (tat['Sex'] == 0).sum()
print(males)
print(females)
p = [females, males]
plt.pie(p,
        labels = ['FEMALE', 'MALE'],
        colors = ['blue', 'red'],
          explode = (0.05, 0),
         startangle = 0
        )
plt.axis('equal')
plt.show()
```

```
577
312
```



In [57]:

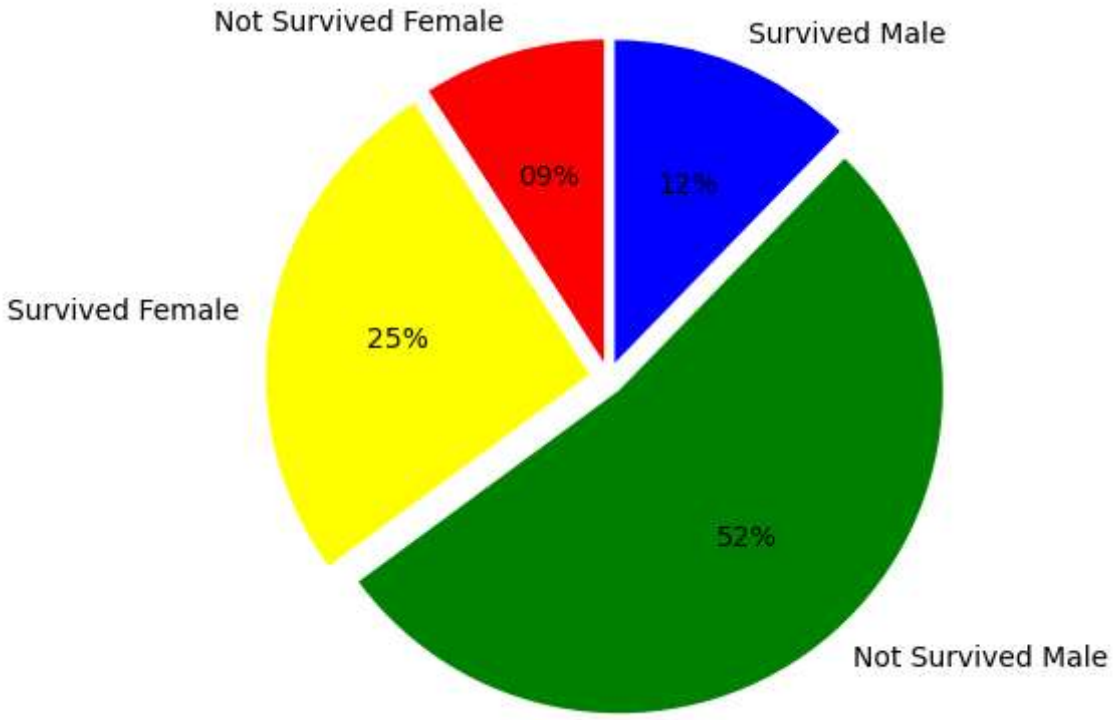
```
MaleS=tat[tat.Sex==1][tat.Survived==1].shape[0]
print(MaleS)
MaleN=tat[tat.Sex==1][tat.Survived==0].shape[0]
print(MaleN)
FemaleS=tat[tat.Sex==0][tat.Survived==1].shape[0]
print(FemaleS)
FemaleN=tat[tat.Sex==0][tat.Survived==0].shape[0]
print(FemaleN)
```

```
109
468
231
81

C:\Users\user\AppData\Local\Temp\ipykernel_29924\2924554640.py:1: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  MaleS=tat[tat.Sex==1][tat.Survived==1].shape[0]
C:\Users\user\AppData\Local\Temp\ipykernel_29924\2924554640.py:3: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  MaleN=tat[tat.Sex==1][tat.Survived==0].shape[0]
C:\Users\user\AppData\Local\Temp\ipykernel_29924\2924554640.py:5: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  FemaleS=tat[tat.Sex==0][tat.Survived==1].shape[0]
C:\Users\user\AppData\Local\Temp\ipykernel_29924\2924554640.py:7: UserWa
rning: Boolean Series key will be reindexed to match DataFrame index.
  FemaleN=tat[tat.Sex==0][tat.Survived==0].shape[0]
```

In [80]:

```
chart=[MaleS,MaleN,FemaleS,FemaleN]
colors=['blue','green','Yellow','Red']
labels=["Survived Male","Not Survived Male","Survived Female","Not Survived Female"]
explode=[0.05,0.05,0.06,0.05]
plt.pie(chart,labels=labels,colors=colors,explode=explode,startangle=90,counterclock=
plt.axis("equal")
plt.show()
```



In [ ]:

```

```