In [59]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

In [60]:
```
titanic_dataset=pd.read_csv(r"D:\OneDrive\Desktop\Data sets for projects\titanic dataset.csv")
```

In [61]:
```
titanic_dataset
```

Out[61]:

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [62]:
```
titanic_dataset.tail()
```

Out[62]:

	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

In [63]:
```
titanic_dataset.head()
```

Out[63]:

	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

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 sex 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 [64]:
```
titanic_dataset.describe()
```

Out[64]:

	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 [65]:
```
del titanic_dataset['Name']
titanic_dataset.head()
```

Out[65]:

	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 [66]:
```
del titanic_dataset['Ticket']
titanic_dataset.head()
```

Out[66]:

	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 [67]:
```
del titanic_dataset['Fare']
titanic_dataset.head()
```

Out[67]:

	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 [68]:
```
del titanic_dataset['Cabin']
titanic_dataset.head()
```

Out[68]:

	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 [69]:
```
def getNumber(x):
    if x=='male':
        return 1
    else:
        return 2
titanic_dataset['Gender']=titanic_dataset['Sex'].apply(getNumber)
titanic_dataset.head()
```

Out[69]:

	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 [70]:
```
del titanic_dataset['Sex']
titanic_dataset.head()
```

Out[70]:

	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 [71]:
```
titanic_dataset.isnull().sum()
```

Out[71]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
dtype: int64
```

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 [72]:
```
meanS= titanic_dataset[titanic_dataset.Survived==1].Age.mean()
meanS
```

Out[72]: 28.343689655172415

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

In [73]:
```
titanic_dataset["age"]=np.where(pd.isnull(titanic_dataset.Age) & titanic_dataset["Survived"]==1  ,meanS, titanic_dataset["Age"
titanic_dataset.head()
```

Out[73]:

	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 [74]:
```
titanic_dataset.isnull().sum()
```

Out[74]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
age            125
dtype: int64
```

In [75]:
```
meanNS= titanic_dataset[titanic_dataset.Survived==0].Age.mean()
meanNS
```

Out[75]:
```
30.62617924528302
```

In [76]:
```
titanic_dataset.age.fillna(meanNS,inplace=True)
titanic_dataset.head()
```

Out[76]:

	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 [77]:
```
titanic_dataset.isnull().sum()
```

Out[77]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
age              0
dtype: int64
```

In [78]:
```
del titanic_dataset['Age']
titanic_dataset.head()
```

Out[78]:

	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

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 [79]:
```
survivedQ = titanic_dataset[titanic_dataset.Embarked == 'Q'][titanic_dataset.Survived == 1].shape[0]
survivedC = titanic_dataset[titanic_dataset.Embarked == 'C'][titanic_dataset.Survived == 1].shape[0]
survivedS = titanic_dataset[titanic_dataset.Embarked == 'S'][titanic_dataset.Survived == 1].shape[0]
print(survivedQ)
print(survivedC)
print(survivedS)
```

```
30
93
217
```

In [80]:
```
survivedQ = titanic_dataset[titanic_dataset.Embarked == 'Q'][titanic_dataset.Survived == 0].shape[0]
survivedC = titanic_dataset[titanic_dataset.Embarked == 'C'][titanic_dataset.Survived == 0].shape[0]
survivedS = titanic_dataset[titanic_dataset.Embarked == 'S'][titanic_dataset.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 [81]:
```
titanic_dataset.dropna(inplace=True)
titanic_dataset.head()
```

Out[81]:

	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 [82]:
```
titanic_dataset.isnull().sum()
```

Out[82]:
```
PassengerId    0
Survived       0
Pclass         0
SibSp          0
Parch          0
Embarked       0
Gender         0
age            0
dtype: int64
```

In [83]:
```
titanic_dataset.rename(columns={'age':'Age'},inplace=True)
titanic_dataset.head()
```

Out[83]:

	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 [84]:
```
titanic_dataset.rename(columns={'Gender':'Sex'},inplace=True)
titanic_dataset.head()
```

Out[84]:

	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 [85]:
```
def getEmb(str):
    if str=='S':
        return 1
    elif str=='Q':
        return 2
    else:
        return 3
titanic_dataset['Embarked']=titanic_dataset['Embarked'].apply(getEmb)
titanic_dataset.head()
```

Out[85]:

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

In [86]:
```
del titanic_dataset['Embarked']
titanic_dataset.head()
```

Out[86]:

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

In [87]:
```
titanic_dataset.rename(columns={'Embark':'Embarked'}, inplace=True)
titanic_dataset.head()
```

Out[87]:

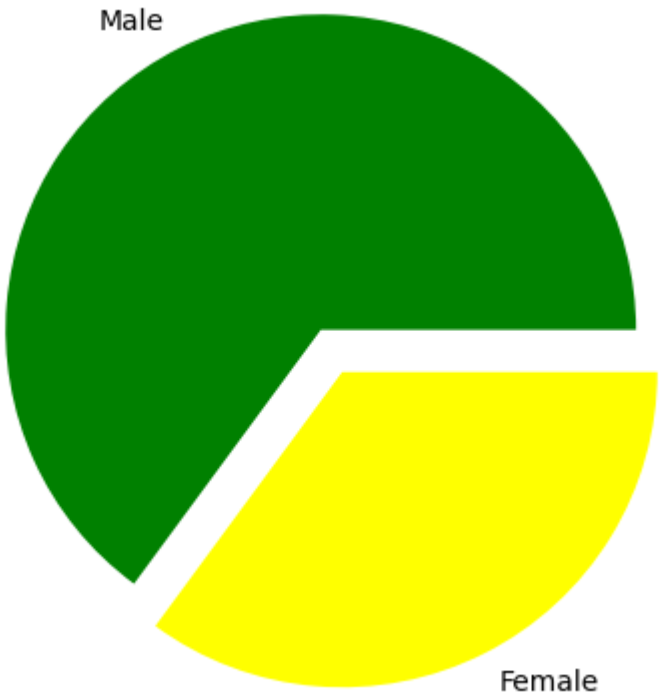
	PassengerId	Survived	Pclass	SibSp	Parch	Sex	Age
0	1	0	3	1	0	1	22.0
1	2	1	1	1	0	2	38.0
2	3	1	3	0	0	2	26.0
3	4	1	1	1	0	2	35.0
4	5	0	3	0	0	1	35.0

In [88]:

```
import matplotlib.pyplot as plt
from matplotlib import style

males=(titanic_dataset['Sex'] == 1).sum()
females=(titanic_dataset['Sex'] == 2).sum()
print(males)
print(females)
p= [males,females]
plt.pie(p,
       labels = ['Male','Female'],
       colors = ['green', 'yellow'],
       explode = (0.15, 0),
       startangle = 0)
plt.axis('equal')
plt.show()
```

```
577
312
```

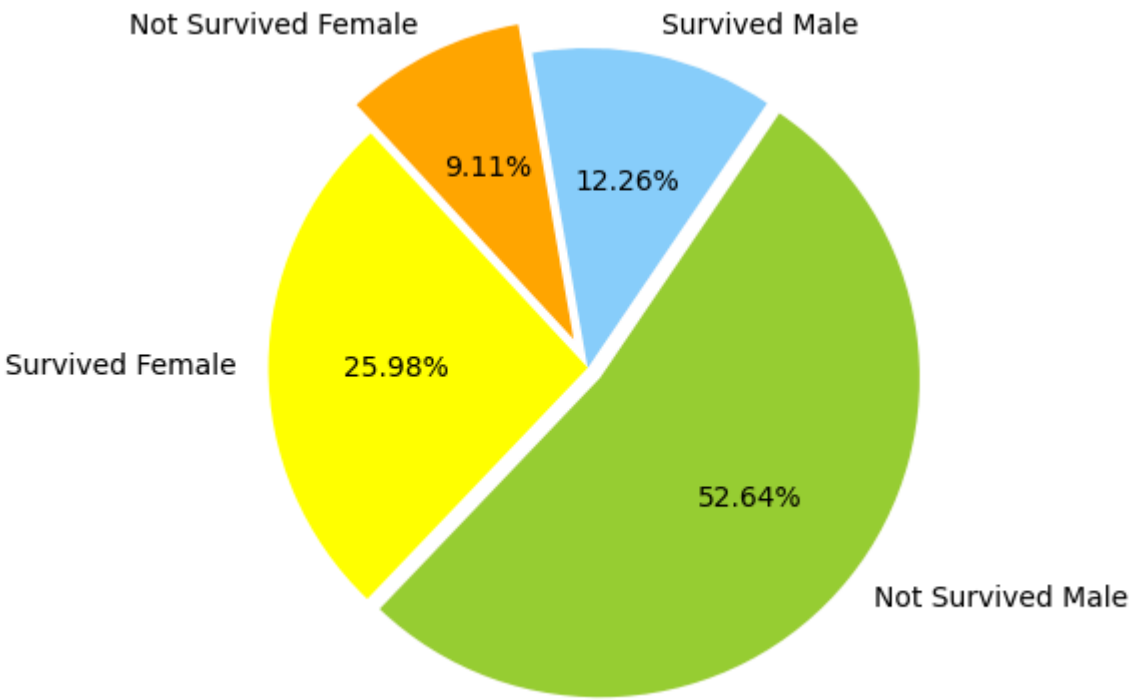


In [89]:
```
MaleS=titanic_dataset[titanic_dataset.Sex==1][titanic_dataset.Survived==1].shape[0]
print(MaleS)
MaleN=titanic_dataset[titanic_dataset.Sex==1][titanic_dataset.Survived==0].shape[0]
print(MaleN)
FemaleS=titanic_dataset[titanic_dataset.Sex==2][titanic_dataset.Survived==1].shape[0]
print(FemaleS)
FemaleN=titanic_dataset[titanic_dataset.Sex==2][titanic_dataset.Survived==0].shape[0]
print(FemaleN)
```

```
109
468
231
81
```

In [90]:
```
chart=[MaleS,MaleN,FemaleS,FemaleN]
colors=['lightskyblue','yellowgreen','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 [ ]: