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

In [2]:
```python
Titanic=pd.read_csv(r'C:\Users\DELL\Downloads\26th\project - data preprocessing\train.csv')
```

In [3]:
```python
Titanic
```

Out[3]:

	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 [4]:
```python
Titanic.shape
```

Out[4]:
```
(891, 12)
```

In [5]:
```python
Titanic.head()
```

Out[5]:

	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

In [6]:
```python
Titanic.tail()
```

Out[6]:

	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 [7]:
```python
Titanic.describe()
```

Out[7]:

	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 [8]:
```python
Titanic.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
```

In [9]:
```python
del Titanic["Name"]
Titanic.head()
```

Out[9]:

	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 [10]:
```python
del Titanic["Ticket"]
Titanic.head()
```

Out[10]:

	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 [12]:
```python
del Titanic["Fare"]
Titanic.head()
```

Out[12]:

	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 [13]:
```python
del Titanic["Cabin"]
Titanic.head()
```

Out[13]:

	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 [16]:
```python
def getNumber(str):
    if str=="male":
        return 1
    else:
        return 2
Titanic["Gender"]=Titanic['Sex'].apply(getNumber)
Titanic.head()
```

Out[16]:

	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 [18]:
```python
del Titanic["Sex"]
Titanic.head()
```

Out[18]:

	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 [19]:
```python
Titanic.isnull().sum()
```

Out[19]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
dtype: int64
```

In [20]:
```python
meanS= Titanic[Titanic.Survived==1].Age.mean()
meanS
```

Out[20]:
```
28.343689655172415
```

In [24]:
```python
Titanic['Age'].isnull()
```

Out[24]:
```
0      False
1      False
2      False
3      False
4      False
       ...  
886    False
887    False
888     True
889    False
890    False
Name: Age, Length: 891, dtype: bool
```

In [25]:
```python
Titanic["age"]=np.where(pd.isnull(Titanic.Age) & Titanic["Survived"]==1  ,meanS, Titanic["Age"])
Titanic.head()
```

Out[25]:

	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 [26]:
```python
Titanic.isnull().sum()
```

Out[26]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
age            125
dtype: int64
```

In [28]:
```python
meanNS=Titanic[Titanic.Survived==0].Age.mean()
meanNS
```

Out[28]:
```
30.62617924528302
```

In [29]:
```python
Titanic.age.fillna(meanNS,inplace=True)
Titanic.head()
```

Out[29]:

	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 [31]:
```python
Titanic.isnull().sum()
```

Out[31]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
age              0
dtype: int64
```

In [32]:
```python
del Titanic['Age']
Titanic.head()
```

Out[32]:

	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 [33]:
```python
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
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\1961588169.py:1: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  survivedQ = Titanic[Titanic.Embarked == 'Q'][Titanic.Survived == 1].shape[0]
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\1961588169.py:2: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  survivedC = Titanic[Titanic.Embarked == 'C'][Titanic.Survived == 1].shape[0]
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\1961588169.py:3: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  survivedS = Titanic[Titanic.Embarked == 'S'][Titanic.Survived == 1].shape[0]
```

In [34]:
```python
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
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\2340563221.py:1: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  survivedQ = Titanic[Titanic.Embarked == 'Q'][Titanic.Survived == 0].shape[0]
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\2340563221.py:2: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  survivedC = Titanic[Titanic.Embarked == 'C'][Titanic.Survived == 0].shape[0]
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\2340563221.py:3: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  survivedS = Titanic[Titanic.Embarked == 'S'][Titanic.Survived == 0].shape[0]
```

In [35]:
```python
Titanic.dropna(inplace=True)
Titanic.head()
```

Out[35]:

	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 [36]:
```python
Titanic.isnull().sum()
```

Out[36]:
```
PassengerId    0
Survived       0
Pclass         0
SibSp          0
Parch          0
Embarked       0
Gender         0
age            0
dtype: int64
```

In [37]:
```python
Titanic.rename(columns={'age':'Age'}, inplace=True)
Titanic.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	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 [38]:
```python
Titanic.rename(columns={'Gender':'Sex'}, inplace=True)
Titanic.head()
```

Out[38]:

	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 [39]:
```python
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[39]:

	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 [40]:
```python
del Titanic['Embarked']
Titanic.rename(columns={'Embark':'Embarked'}, inplace=True)
Titanic.head()
```

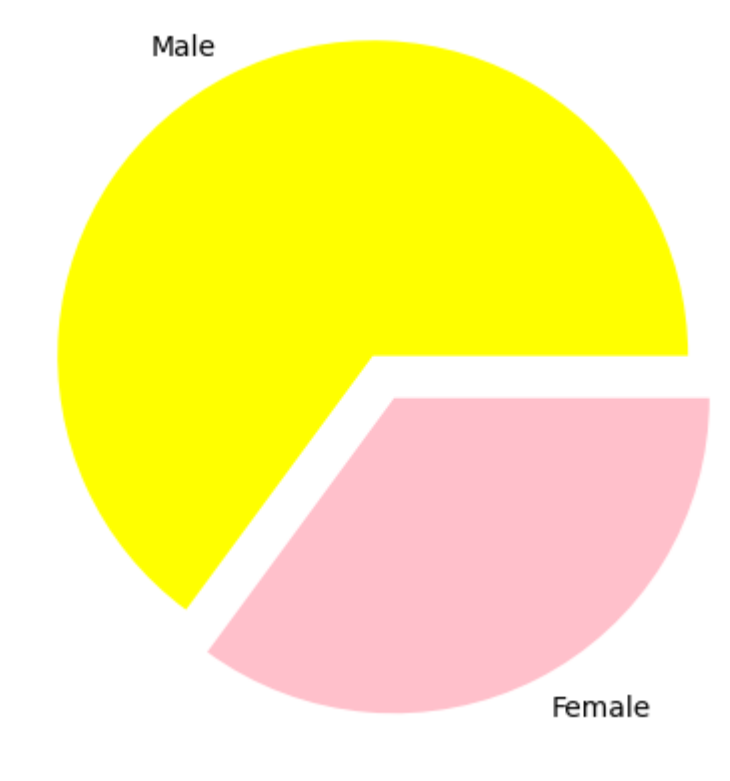
Out[40]:

	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

In [44]:
```python
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'],
       colors = ['yellow', 'pink'],
       explode = (0.15, 0),
       startangle = 0)
plt.axis('equal')
plt.show()
```

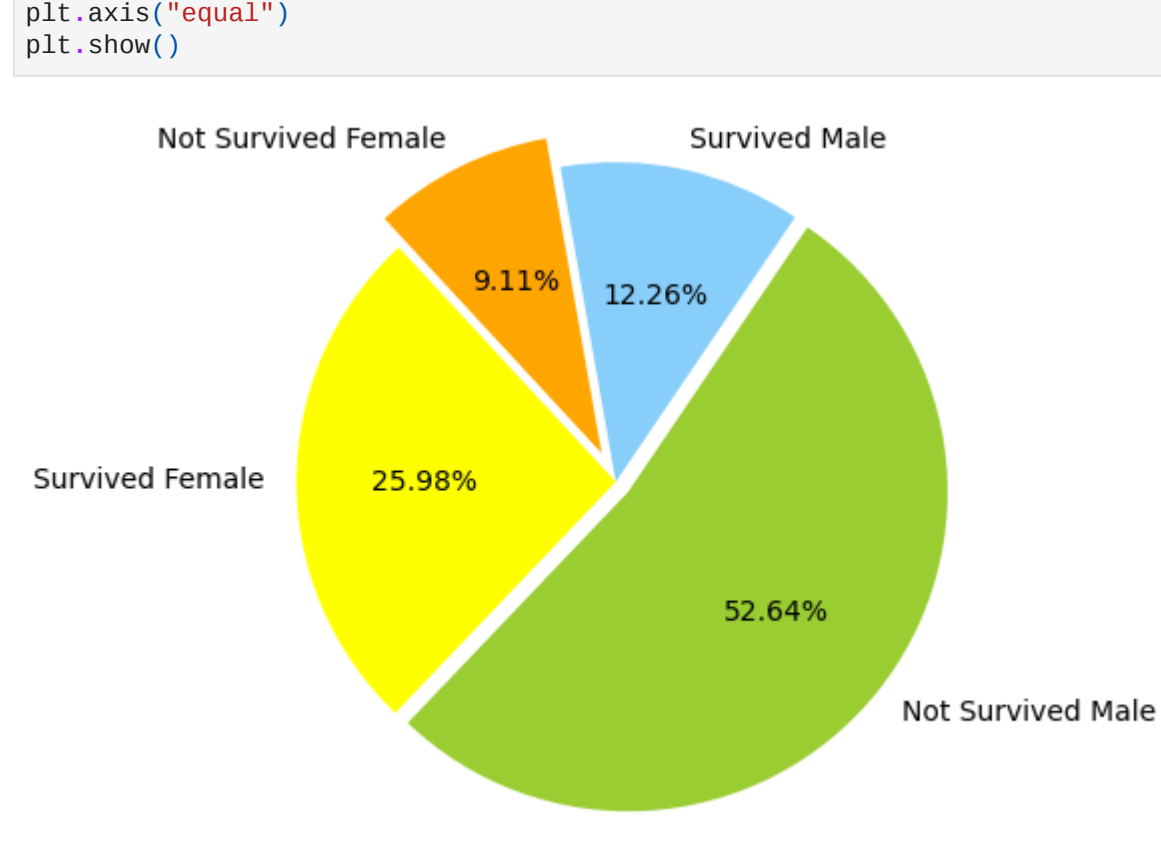
```
577
312
```



In [45]:
```python
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
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\3395068644.py:1: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  MaleS=Titanic[Titanic.Sex==1][Titanic.Survived==1].shape[0]
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\3395068644.py:3: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  MaleN=Titanic[Titanic.Sex==1][Titanic.Survived==0].shape[0]
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\3395068644.py:5: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  FemaleS=Titanic[Titanic.Sex==2][Titanic.Survived==1].shape[0]
C:\Users\DELL\AppData\Local\Temp\ipykernel_11868\3395068644.py:7: UserWarning: Boolean Series key will be reindexed to match DataFrame index.
  FemaleN=Titanic[Titanic.Sex==2][Titanic.Survived==0].shape[0]
```

In [46]:
```python
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 [ ]: