Data preprocessing with Titanic Train Data ¶

Import Libraries

In [13]:
```
import pandas as pd
import numpy as np
```

Read the Dataset

In [14]:
```
titanic = pd.read_csv(r'c:\Users\hp\OneDrive\Documents\Desktop\train.csv')

titanic .tail()
```

Out[14]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embark
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	

Performing Data Cleaning and Analysis

In [15]:
```
titanic.describe()
```

Out[15]:

	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 [16]:
```
del titanic ["Name"]
titanic.head()
```

Out[16]:

	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 [17]:
```
del titanic["Ticket"]
titanic.head()
```

Out[17]:

	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 [18]:
```
del titanic["Fare"]
titanic.head()
```

Out[18]:

	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 [19]:
```
del titanic["Cabin"]
titanic.head()
```

Out[19]:

	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 and female string value to numerical value , male=1,female=
```

In [21]:
```
def getNumber(str):
    if str=="male":
        return 1
    else:
        return 2
titanic["Gender"]=titanic["Sex"].apply(getNumber)
```

In [22]:
```
##We have created a new column called "Gender" and
#filling it with values 1,2 based on the values of sex column
```

In [23]:
```
titanic.head(1)
```

Out[23]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Embarked	Gender
0	1	0	3	male	22.0	1	0	S	1

In [24]:
```
#Deleting Sex column, since no use of it now
del titanic["Sex"]
titanic.head()
```

Out[24]:

	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 [25]:
```
titanic.isnull().sum()
```

Out[25]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
dtype: int64
```

ill 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 [27]:
```
meanS= titanic[titanic.Survived==1].Age.mean()
meanS
```

Out[27]:
```
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 [28]:
```
titanic["age"]=np.where(pd.isnull(titanic.Age) & titanic["Survived"]==1 ,meanS, titan
titanic.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	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 [29]:
```
titanic.isnull().sum()
```

Out[29]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
age            125
dtype: int64
```

In [30]:
```
# Finding the mean age of "Not Survived" people
```

In [31]:
```
meanNS=titanic[titanic.Survived==0].Age.mean()
meanNS
```

Out[31]:
```
30.62617924528302
```

In [32]:
```
titanic.age.fillna(meanNS,inplace=True)
titanic.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	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 [33]:
```
titanic.isnull().sum()
```

Out[33]:
```
PassengerId      0
Survived         0
Pclass           0
Age            177
SibSp            0
Parch            0
Embarked         2
Gender           0
age              0
dtype: int64
```

In [34]:
```
del titanic["Age"]
titanic.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	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 [35]:
```
#We want to check if "Embarked" column is is important for analysis or not, that is wh
# Finding the number of people who have survived
# given that they have embarked or boarded from a particular port
```

In [36]:
```
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\hp\AppData\Local\Temp\ipykernel_11888\2602345876.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\hp\AppData\Local\Temp\ipykernel_11888\2602345876.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\hp\AppData\Local\Temp\ipykernel_11888\2602345876.py:3: UserWarning: Boolean
Series key will be reindexed to match DataFrame index.
  survivedS = titanic[titanic.Embarked == 'S'][titanic.Survived == 1].shape[0]
```

In [37]:
```
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\hp\AppData\Local\Temp\ipykernel_11888\1056497950.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\hp\AppData\Local\Temp\ipykernel_11888\1056497950.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\hp\AppData\Local\Temp\ipykernel_11888\1056497950.py:3: UserWarning: Boolean
Series key will be reindexed to match DataFrame index.
  survivedS = titanic[titanic.Embarked == "S"][titanic.Survived == 0].shape[0]
```

In [38]:
```
#As there are significant changes in the survival rate based on which port the passeng

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

Out[38]:

	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 [39]:
```
titanic.isnull().sum()
```

Out[39]:
```
PassengerId    0
Survived       0
Pclass         0
SibSp          0
Parch          0
Embarked       0
Gender         0
age            0
dtype: int64
```

In [40]:
```
# Renameing "age" and "gander" columns
titanic.rename(columns={"age":"Age"}, inplace=True)
titanic.head()
```

Out[40]:

	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 [41]:
```
titanic.rename(columns={"Gender":"Sex"}, inplace=True)
titanic.head()
```

Out[41]:

	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 [42]:
```
del titanic["Embarked"]
titanic.rename(columns={"Embark":"Embarked"}, inplace=True)
titanic.head()
```

Out[42]:

	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

Data Visualization

In [43]:
```
#Drawing a pie chart for number of males and females aboard

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

In [44]:
```
males = (titanic["Sex"] == 1).sum()
```

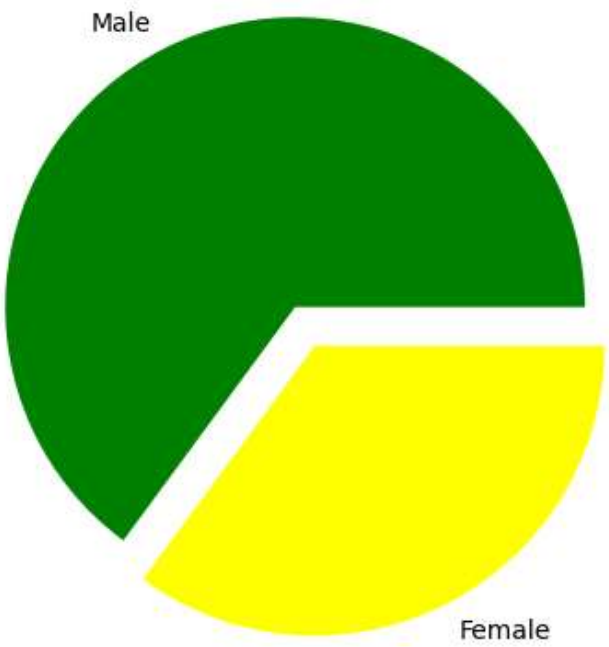
In [45]:
```
#summing up all the values of columns gender with a
#condition for male and similary for female
```

In [46]:
```
females = (titanic["Sex"] == 2).sum()
print(males)
print(females)
p = [males, females]
plt.pie(p, #givig array
        labels = ["Male", "Female"],# correspondingly giving labels

        colors = ["green", "yellow"], # corresponding colors

        explode = (0.15, 0), # how much the gap should me there between the pies
        startangle = 0) # what start angle should be given
plt.axis("equal")
plt.show()
```

```
577
312
```



More Precise Pie Chart

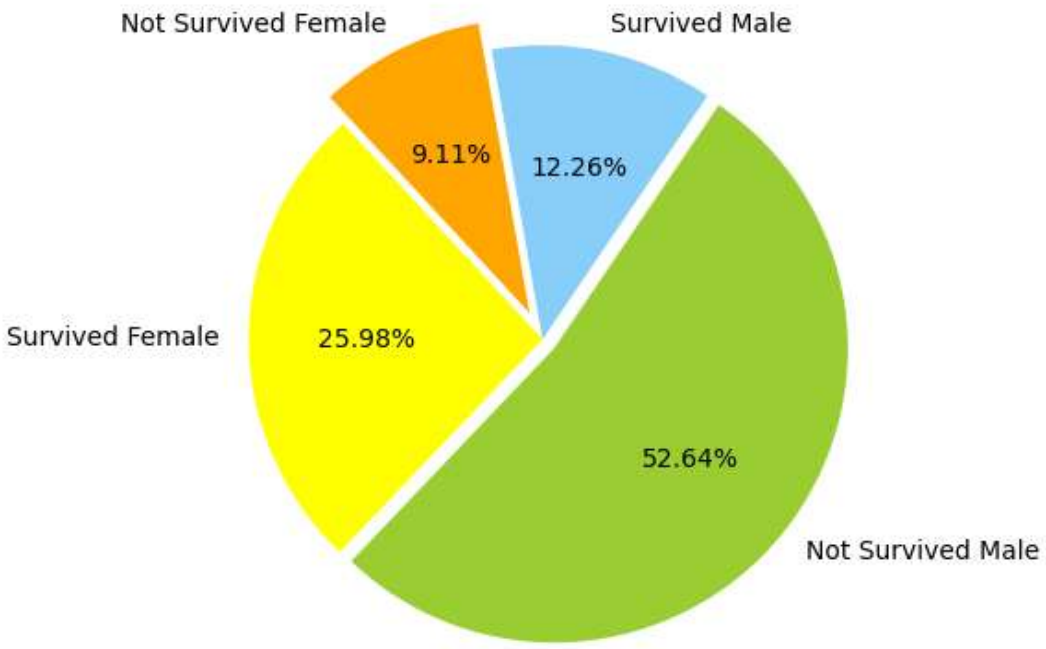
In [47]:
```
MaleS=titanic[titanic.Sex==1][titanic.Survived==1].shape[0]
print(MaleS)
MaleN=titanic[titanic.Sex==1][titanic.Survived==0].shape[0]
print(MaleN)
FemaleS=titanic[titanic.Sex==2][titanic.Survived==1].shape[0]
print(FemaleS)

FemaleN=titanic[titanic.Sex==2][titanic.Survived==0].shape[0]
print(FemaleN)

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
plt.axis("equal")
plt.show()
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_11888\1381038892.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\hp\AppData\Local\Temp\ipykernel_11888\1381038892.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\hp\AppData\Local\Temp\ipykernel_11888\1381038892.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\hp\AppData\Local\Temp\ipykernel_11888\1381038892.py:8: UserWarning: Boolean
Series key will be reindexed to match DataFrame index.
  FemaleN=titanic[titanic.Sex==2][titanic.Survived==0].shape[0]

109
468
231
81
```



In [ ]: