Data Preprocessing for beginners

In [1]:

```
#Importing Libraries
import numpy as np
import pandas as pd

import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```
#Importing Dataset
df= pd.read_csv('train.csv')
df.head(5)
```

Out[2]:

	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

Performing Data Cleaning and Analysis

In [3]:

```
df.describe()
```

Out[3]:

	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 [4]:

```
#Name column cam never decide survival of a person, hence we can delete it.
del df["Name"]
df.head()
```

Out[4]:

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

In [5]:

```
del df["Ticket"]
df.head()
```

Out[5]:

	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 [6]:

```
del df['Fare']
df.head()
```

Out[6]:

	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 [7]:

```
del df['Cabin']
df.head()
```

Out[7]:

	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 [8]:

```
# Changing Value for "Male, Female" string values to numeric values , male=1 and female=

def getNumber(str):
    if str=="male":
        return 1
    else:
        return 2
df["Gender"]= df['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
df.head()
```

Out[8]:

	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 [9]:

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

Out[9]:

	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 [10]:

```
df.isnull()
```

Out[10]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Embarked	Gender
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False
887	False	False	False	False	False	False	False	False
888	False	False	False	True	False	False	False	False
889	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False

891 rows × 8 columns

In [11]:

```
df.isnull().sum()
```

Out[11]:

```
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 [12]:

```
means = df[df.Survived==1].Age.mean()
means
```

Out[12]:

```
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 [13]:

```
df["age"]=np.where(pd.isnull(df.Age) & df["Survived"]==1  ,means, df["Age"])
df.head()
```

Out[13]:

	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 [14]:

```
df.isnull().sum()
```

Out[14]:

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

In [15]:

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

Out[15]:

```
30.62617924528302
```

In [16]:

```
df.age.fillna(meanNS,inplace=True)          #fillna---Fill NA/NaN values using the specifi
df.head()
```

Out[16]:

	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 [17]:

```
df.isnull().sum()
```

Out[17]:

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

In [18]:

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

Out[18]:

	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 [19]:

```
df.isnull().sum()
```

Out[19]:

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

In [20]:

```
pd.isnull(df).sum()
```

Out[20]:

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

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 [21]:

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

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

```
30
93
217
```

In [22]:

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

```
47
75
427
```

In [23]:

```
df.isnull().sum()
```

Out[23]:

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

In [24]:

```
df.dropna(inplace=True)      #remove all missing value rows
df.isnull().sum()
```

Out[24]:

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

In [25]:

```
#Renaming "age" and "gender" columns
df.rename(columns={'age':'Age'}, inplace=True)
df.head()
```

Out[25]:

	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 [26]:

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

Out[26]:

	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 [27]:

```
#Changing the categorical value of 'Embarked' attribute to numerical value, S=1, Q=2 and

def getEmbark(str):
    if str=="S":
        return 1
    elif str=="Q":
        return 2
    else:
        return 3

df["Embark"]=df["Embarked"].apply(getEmbark)
df.head()
```

Out[27]:

	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 [28]:

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

Out[28]:

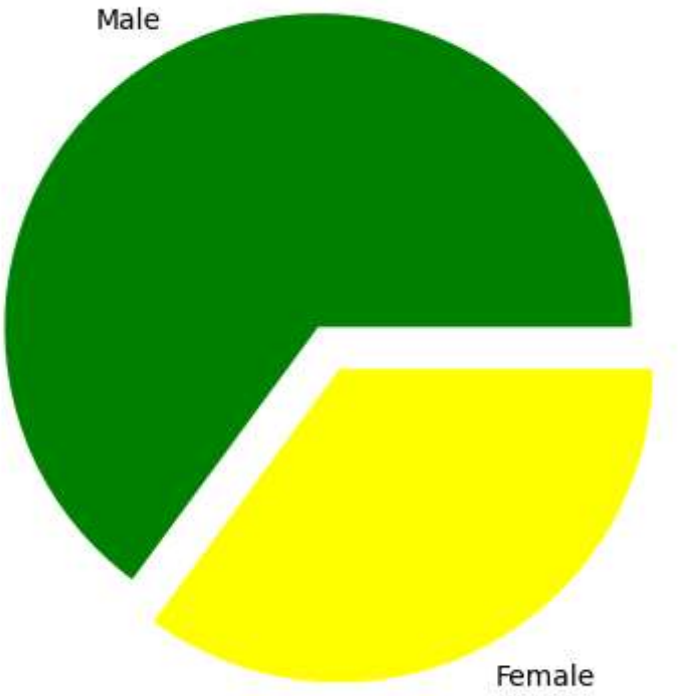
	PassengerId	Survived	Pclass	SibSp	Parch	Sex	Age	Embarked
0	1	0	3	1	0	1	22.0	1
1	2	1	1	1	0	2	38.0	3
2	3	1	3	0	0	2	26.0	1
3	4	1	1	1	0	2	35.0	1
4	5	0	3	0	0	1	35.0	1

In [29]:

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

males = (df['Sex'] == 1).sum()
#Summing up all the values of column gender with a
#condition for male and similary for females
females = (df['Sex'] == 2).sum()
print(males)
print(females)
p = [males, females]
plt.pie(p,    #giving array
       labels = ['Male', 'Female'], #Correspndingly 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
```



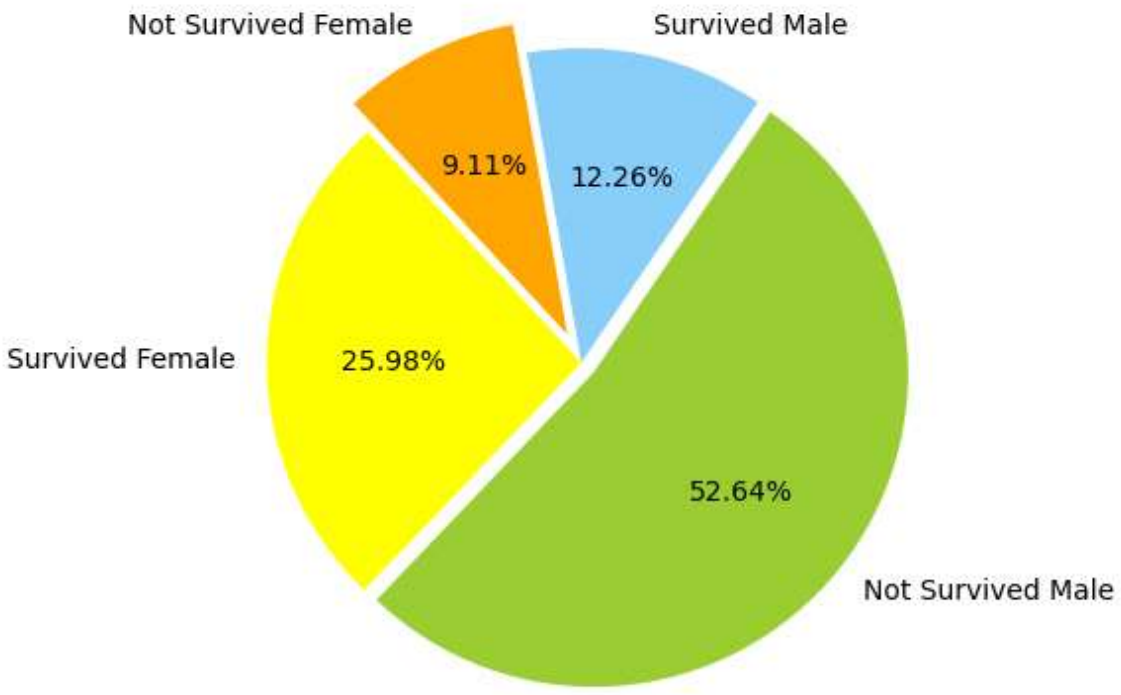
In [30]:

```
# More Precise Pie Chart
MaleS=df[df.Sex==1][df.Survived==1].shape[0]
print(MaleS)
MaleN=df[df.Sex==1][df.Survived==0].shape[0]
print(MaleN)
FemaleS=df[df.Sex==2][df.Survived==1].shape[0]
print(FemaleS)
FemaleN=df[df.Sex==2][df.Survived==0].shape[0]
print(FemaleN)
```

```
109
468
231
81
```

In [31]:

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



In [ ]:

In [ ]:

```
```