1. Import all the required Python Libraries.
2. Locate open source data from the web (e.g., https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site).
3. Load the Dataset into pandas dataframe

```
#Import Tabular Data from CSV Files into Pandas Dataframes
import pandas as pd
df= pd.read_csv(r"C:\Users\Jayditya\Downloads\DSBDA LAB\Lab\
Experiments\Datasets\13data.csv")
print(df)

     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name     Sex   Age
SibSp  \
0                              Braund, Mr. Owen Harris    male  22.0
1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                               Heikkinen, Miss. Laina  female  26.0
0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                             Allen, Mr. William Henry    male  35.0
0
..                                                 ...     ...   ...
...
886                              Montvila, Rev. Juozas    male  27.0
0
887                       Graham, Miss. Margaret Edith  female  19.0
0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN
1
889                              Behr, Mr. Karl Howell    male  26.0
0
890                                Dooley, Mr. Patrick    male  32.0
0

     Parch            Ticket     Fare Cabin Embarked
```

```
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     ...               ...      ...   ...      ...
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

4.Data Preprocessing: check for missing values in the data using pandas isnull()

```
df.isnull()
df

     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name     Sex   Age
SibSp  \
0                              Braund, Mr. Owen Harris    male  22.0
1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                               Heikkinen, Miss. Laina  female  26.0
0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                             Allen, Mr. William Henry    male  35.0
0
..                                                 ...     ...   ...
...
886                              Montvila, Rev. Juozas    male  27.0
0
887                       Graham, Miss. Margaret Edith  female  19.0
```

```
0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN
1
889                              Behr, Mr. Karl Howell    male  26.0
0
890                                Dooley, Mr. Patrick    male  32.0
0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     ...               ...      ...   ...      ...
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

```
df.isnull().sum().sum()
#returns the number of missing values in the dataset.
```

```
866
```

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

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

#describe() function to get some initial statistics

```
df.describe()
```

```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
```

```
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

```
df.describe(include=['object'])
```

```
                           Name   Sex  Ticket    Cabin Embarked
count                       891   891     891      204      889
unique                      891     2     681      147        3
top     Braund, Mr. Owen Harris  male  347082  B96 B98        S
freq                          1   577       7        4      644
```

Provide variable descriptions. Types of variables

```
df.dtypes
```

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

```
df.head(5)
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
```

```
3            4         1       1
4            5         0       3

                                                 Name     Sex   Age
SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0
1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                              Heikkinen, Miss. Laina  female  26.0
0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                            Allen, Mr. William Henry    male  35.0
0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

```
#Check the dimensions of the data frame
df.shape
```

```
(891, 12)
```

```
#number of rows of a DataFrame
len(df)
```

```
891
```

```
#total number of elements in the DataFrame
df.size
```

```
10692
```

Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions

```
df.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
```

```
df.dtypes
```

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

If variables are not in the correct data type, apply proper type conversions......Age--float64--int

```
df['Age']=df['Age'].fillna(20)
print(df)
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3   
1              2         1       1   
2              3         1       3   
3              4         1       1   
4              5         0       3   
..           ...       ...     ...   
886          887         0       2   
887          888         1       1   
888          889         0       3   
889          890         1       1   
890          891         0       3   

                                                  Name     Sex   Age  
SibSp  \
0                              Braund, Mr. Owen Harris    male  22.0   
```

```
1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                               Heikkinen, Miss. Laina  female  26.0
0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                             Allen, Mr. William Henry    male  35.0
0
..                                                 ...     ...   ...
...
886                              Montvila, Rev. Juozas    male  27.0
0
887                       Graham, Miss. Margaret Edith  female  19.0
0
888           Johnston, Miss. Catherine Helen "Carrie"  female  20.0
1
889                              Behr, Mr. Karl Howell    male  26.0
0
890                                Dooley, Mr. Patrick    male  32.0
0

     Parch             Ticket     Fare Cabin Embarked
0        0          A/5 21171   7.2500   NaN        S
1        0           PC 17599  71.2833   C85        C
2        0   STON/O2. 3101282   7.9250   NaN        S
3        0             113803  53.1000  C123        S
4        0             373450   8.0500   NaN        S
..     ...                ...      ...   ...      ...
886      0             211536  13.0000   NaN        S
887      0             112053  30.0000   B42        S
888      2         W./C. 6607  23.4500   NaN        S
889      0             111369  30.0000  C148        C
890      0             370376   7.7500   NaN        Q

[891 rows x 12 columns]

df['Age'] = df['Age'].astype('int64')
print(df.dtypes)

PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age              int64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
```

```
Cabin           object
Embarked        object
dtype: object
```

Turn categorical variables into quantitative variables in Python

```

dummies = pd.get_dummies(df.Sex)
merged = pd.concat([df, dummies], axis='columns')
merged.drop(['Sex', 'male'], axis='columns')
print(merged)

     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name     Sex  Age
SibSp  \
0                              Braund, Mr. Owen Harris    male   22
1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female   38
1
2                               Heikkinen, Miss. Laina  female   26
0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female   35
1
4                             Allen, Mr. William Henry    male   35
0
..                                                 ...     ...  ...
...
886                              Montvila, Rev. Juozas    male   27
0
887                       Graham, Miss. Margaret Edith  female   19
0
888           Johnston, Miss. Catherine Helen "Carrie"  female   20
1
889                              Behr, Mr. Karl Howell    male   26
0
890                                Dooley, Mr. Patrick    male   32
0
```

```
     Parch            Ticket     Fare Cabin Embarked  female   male
0        0         A/5 21171   7.2500   NaN        S   False   True
1        0          PC 17599  71.2833   C85        C    True  False
2        0  STON/O2. 3101282   7.9250   NaN        S    True  False
3        0            113803  53.1000  C123        S    True  False
4        0            373450   8.0500   NaN        S   False   True
..     ...               ...      ...   ...      ...     ...    ...
886      0            211536  13.0000   NaN        S   False   True
887      0            112053  30.0000   B42        S    True  False
888      2        W./C. 6607  23.4500   NaN        S    True  False
889      0            111369  30.0000  C148        C   False   True
890      0            370376   7.7500   NaN        Q   False   True

[891 rows x 14 columns]
```

```
dummies = pd.get_dummies(df.Sex)
merged = pd.concat([df, dummies], axis='columns')
merged.drop(['Sex', 'male'], axis='columns')
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name  Age  SibSp
Parch  \
0                              Braund, Mr. Owen Harris   22      1
0
1    Cumings, Mrs. John Bradley (Florence Briggs Th...   38      1
0
2                               Heikkinen, Miss. Laina   26      0
0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)   35      1
0
4                             Allen, Mr. William Henry   35      0
0
..                                                 ...  ...    ...
...
886                              Montvila, Rev. Juozas   27      0
0
887                       Graham, Miss. Margaret Edith   19      0
```

```
0
888           Johnston, Miss. Catherine Helen "Carrie"   20      1
2
889                              Behr, Mr. Karl Howell   26      0
0
890                                Dooley, Mr. Patrick   32      0
0

               Ticket     Fare Cabin Embarked  female
0           A/5 21171   7.2500   NaN        S   False
1            PC 17599  71.2833   C85        C    True
2    STON/O2. 3101282   7.9250   NaN        S    True
3              113803  53.1000  C123        S    True
4              373450   8.0500   NaN        S   False
..                ...      ...   ...      ...     ...
886            211536  13.0000   NaN        S   False
887            112053  30.0000   B42        S    True
888        W./C. 6607  23.4500   NaN        S    True
889            111369  30.0000  C148        C   False
890            370376   7.7500   NaN        Q   False

[891 rows x 12 columns]
```

```python
df["Embarked_cat"] = df["Embarked"].astype('category')
df["Embarked_num"] = df["Embarked_cat"].cat.codes
df
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name     Sex  Age
SibSp  \
0                              Braund, Mr. Owen Harris    male   22
1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female   38
1
2                               Heikkinen, Miss. Laina  female   26
0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female   35
1
```

```
4                             Allen, Mr. William Henry    male   35
0
..                                                 ...     ...  ...
...
886                              Montvila, Rev. Juozas    male   27
0
887                       Graham, Miss. Margaret Edith  female   19
0
888           Johnston, Miss. Catherine Helen "Carrie"  female   20
1
889                              Behr, Mr. Karl Howell    male   26
0
890                                Dooley, Mr. Patrick    male   32
0

     Parch            Ticket     Fare Cabin Embarked Embarked_cat  \
0        0         A/5 21171   7.2500   NaN        S            S
1        0          PC 17599  71.2833   C85        C            C
2        0  STON/O2. 3101282   7.9250   NaN        S            S
3        0            113803  53.1000  C123        S            S
4        0            373450   8.0500   NaN        S            S
..     ...               ...      ...   ...      ...          ...
886      0            211536  13.0000   NaN        S            S
887      0            112053  30.0000   B42        S            S
888      2        W./C. 6607  23.4500   NaN        S            S
889      0            111369  30.0000  C148        C            C
890      0            370376   7.7500   NaN        Q            Q

     Embarked_num
0               2
1               0
2               2
3               2
4               2
..            ...
886             2
887             2
888             2
889             0
890             1

[891 rows x 14 columns]
```

```
from sklearn.preprocessing import LabelEncoder
# creating instance of labelencoder
labelencoder = LabelEncoder()
# Assigning numerical values and storing in another column
df['Labelencoding_Embarked'] =
labelencoder.fit_transform(df["Embarked"])
df
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name     Sex  Age
SibSp  \
0                              Braund, Mr. Owen Harris    male   22
1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female   38
1
2                               Heikkinen, Miss. Laina  female   26
0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female   35
1
4                             Allen, Mr. William Henry    male   35
0
..                                                 ...     ...  ...
...
886                              Montvila, Rev. Juozas    male   27
0
887                       Graham, Miss. Margaret Edith  female   19
0
888           Johnston, Miss. Catherine Helen "Carrie"  female   20
1
889                              Behr, Mr. Karl Howell    male   26
0
890                                Dooley, Mr. Patrick    male   32
0

     Parch            Ticket     Fare Cabin Embarked Embarked_cat  \
0        0         A/5 21171   7.2500   NaN        S            S
1        0          PC 17599  71.2833   C85        C            C
2        0  STON/O2. 3101282   7.9250   NaN        S            S
3        0            113803  53.1000  C123        S            S
4        0            373450   8.0500   NaN        S            S
..     ...               ...      ...   ...      ...          ...
886      0            211536  13.0000   NaN        S            S
887      0            112053  30.0000   B42        S            S
888      2        W./C. 6607  23.4500   NaN        S            S
889      0            111369  30.0000  C148        C            C
890      0            370376   7.7500   NaN        Q            Q
```

```
     Embarked_num  Labelencoding_Embarked
0               2                       2
1               0                       0
2               2                       2
3               2                       2
4               2                       2
..            ...                     ...
886             2                       2
887             2                       2
888             2                       2
889             0                       0
890             1                       1

[891 rows x 15 columns]
```

```
df['Labelencoding_Embarked'].value_counts()
```

```
Labelencoding_Embarked
2    644
0    168
1     77
3      2
Name: count, dtype: int64
```

```
# replacing values
df['Sex'] = df['Sex'].replace(['male', 'female'], [0, 1])
df
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name  Sex   Age
SibSp  \
0                              Braund, Mr. Owen Harris    0   22
1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...    1   38
1
2                               Heikkinen, Miss. Laina    1   26
0
```

```
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)    1   35
1
4                             Allen, Mr. William Henry    0   35
0
..                                                 ...  ...  ...    ..
.
886                              Montvila, Rev. Juozas    0   27
0
887                       Graham, Miss. Margaret Edith    1   19
0
888           Johnston, Miss. Catherine Helen "Carrie"    1   20
1
889                              Behr, Mr. Karl Howell    0   26
0
890                                Dooley, Mr. Patrick    0   32
0

     Parch            Ticket     Fare Cabin Embarked Embarked_cat  \
0        0         A/5 21171   7.2500   NaN        S            S
1        0          PC 17599  71.2833   C85        C            C
2        0  STON/O2. 3101282   7.9250   NaN        S            S
3        0            113803  53.1000  C123        S            S
4        0            373450   8.0500   NaN        S            S
..     ...               ...      ...   ...      ...          ...
886      0            211536  13.0000   NaN        S            S
887      0            112053  30.0000   B42        S            S
888      2        W./C. 6607  23.4500   NaN        S            S
889      0            111369  30.0000  C148        C            C
890      0            370376   7.7500   NaN        Q            Q

     Embarked_num  Labelencoding_Embarked
0               2                       2
1               0                       0
2               2                       2
3               2                       2
4               2                       2
..            ...                     ...
886             2                       2
887             2                       2
888             2                       2
889             0                       0
890             1                       1

[891 rows x 15 columns]
```