- 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
               2
                                  1
1
                          1
2
               3
                          1
                                  3
3
                                  1
               4
                          1
4
               5
                          0
                                  3
886
             887
                          0
                                  2
                                  1
887
             888
                          1
                                  3
888
             889
                          0
889
             890
                          1
                                  1
                                  3
890
             891
                                                             Sex
                                                    Name
                                                                   Age
SibSp \
                                Braund, Mr. Owen Harris
0
                                                            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
                                  Montvila, Rev. Juozas
                                                            male 27.0
886
0
887
                           Graham, Miss. Margaret Edith
                                                          female 19.0
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
                                                         C
                      PC 17599
                                 71.2833
                                             C85
1
          0
2
                                                         S
          0
             STON/02. 3101282
                                  7.9250
                                             NaN
3
                                                         S
          0
                         113803
                                 53.1000
                                            C123
4
                                                         S
          0
                        373450
                                  8.0500
                                             NaN
. .
                                                         S
                        211536
                                 13.0000
886
          0
                                             NaN
887
                        112053
                                 30.0000
                                             B42
                                                         S
          0
                    W./C. 6607
                                                         S
888
          2
                                 23.4500
                                             NaN
                                                         C
889
          0
                        111369
                                 30.0000
                                            C148
890
                        370376
                                  7.7500
                                             NaN
                                                         0
[891 rows x 12 columns]
```

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

```
df.isnull()
df
                             Pclass \
     PassengerId
                  Survived
0
               1
                                  3
                          0
1
               2
                          1
                                  1
2
               3
                                  3
                          1
3
               4
                                  1
                          1
4
               5
                                  3
                          0
                                  2
886
             887
                          0
                                  1
887
             888
                          1
                                  3
888
             889
                          0
                                  1
                          1
889
             890
                                  3
890
             891
                                                    Name
                                                             Sex
                                                                   Age
SibSp \
                                Braund, Mr. Owen Harris
                                                            male 22.0
0
1
1
     Cumings, Mrs. John Bradley (Florence Briggs Th...
                                                          female 38.0
1
                                 Heikkinen, Miss. Laina
2
                                                          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
890
                                     Dooley, Mr. Patrick
                                                             male 32.0
0
                                   Fare Cabin Embarked
     Parch
                       Ticket
0
                    A/5 21171
                                          NaN
         0
                                7.2500
                                                      S
                                                      C
1
         0
                     PC 17599
                               71.2833
                                          C85
2
                                                      S
            STON/02. 3101282
                               7.9250
                                          NaN
         0
3
         0
                       113803
                               53.1000
                                         C123
                                                      S
4
                                                      S
         0
                       373450
                                8.0500
                                          NaN
                       211536
                               13.0000
                                                      S
886
         0
                                          NaN
                                                      S
                       112053
                                          B42
887
         0
                               30.0000
                                                      S
888
         2
                  W./C. 6607
                               23.4500
                                          NaN
889
                       111369
                               30.0000
                                         C148
                                                      C
         0
890
         0
                       370376
                               7.7500
                                                      0
                                          NaN
[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
               177
Age
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 714.000000 891.000000
```

```
446.000000
                       0.383838
                                    2.308642
                                                29.699118
                                                              0.523008
mean
                                                14.526497
std
        257.353842
                       0.486592
                                    0.836071
                                                              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
        891.000000
                       1.000000
                                    3.000000
                                                80.000000
                                                              8.000000
max
            Parch
                           Fare
       891.000000
                    891.000000
count
         0.381594
                     32.204208
mean
                     49.693429
std
         0.806057
min
         0.000000
                      0.000000
25%
         0.000000
                      7.910400
50%
         0.000000
                     14.454200
75%
         0.000000
                     31,000000
         6.000000
                    512.329200
max
df.describe(include=['object'])
                                          Ticket
                                                     Cabin Embarked
                             Name
                                    Sex
count
                              891
                                    891
                                             891
                                                       204
                                                                889
                                                       147
                                                                  3
unique
                              891
                                      2
                                             681
                                                                  S
top
        Braund, Mr. Owen Harris
                                   male
                                          347082
                                                  B96 B98
                                    577
                                                                644
freq
                                1
                                               7
                                                         4
```

Provide variable descriptions. Types of variables

```
df.dtypes
PassengerId
                  int64
Survived
                  int64
Pclass
                  int64
Name
                 object
Sex
                 object
                float64
Age
SibSp
                  int64
                  int64
Parch
Ticket
                 object
Fare
                float64
                 object
Cabin
Embarked
                 object
dtype: object
df.head(5)
   PassengerId
                 Survived
                            Pclass
0
              1
                         0
                                  3
              2
1
                         1
                                  1
2
              3
                         1
                                  3
```

```
3
                                1
                        0
4
             5
                                3
                                                  Name
                                                           Sex
                                                                  Age
SibSp \
                              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
                   PC 17599
1
       0
                             71.2833
                                       C85
                                                   C
2
       0
          STON/02. 3101282
                                       NaN
                                                   S
                              7.9250
                                                   S
3
       0
                     113803
                             53,1000
                                      C123
                                                   S
4
       0
                     373450
                              8.0500
                                       NaN
#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

```
2
     Pclass
                  891 non-null
                                   int64
 3
     Name
                  891 non-null
                                   object
 4
     Sex
                  891 non-null
                                   object
 5
                  714 non-null
                                   float64
     Age
 6
     SibSp
                  891 non-null
                                   int64
 7
     Parch
                  891 non-null
                                   int64
 8
     Ticket
                                   object
                  891 non-null
 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
               float64
Age
                 int64
SibSp
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)
                               Pclass \
     PassengerId
                    Survived
0
                                     3
                 1
                            0
                 2
1
                            1
                                     1
2
                 3
                                     3
                            1
3
                            1
                                     1
                 4
4
                 5
                            0
                                     3
                                   . . .
. .
               . . .
886
              887
                            0
                                     2
                                     1
887
              888
                            1
                                     3
888
              889
                            0
889
              890
                            1
                                     1
                                     3
890
              891
                            0
                                                        Name
                                                                   Sex
                                                                         Age
SibSp \
                                   Braund, Mr. Owen Harris
                                                                 male 22.0
```

```
1
     Cumings, Mrs. John Bradley (Florence Briggs Th... female 38.0
1
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
                           Graham, Miss. Margaret Edith
887
                                                          female 19.0
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
                   A/5 21171
0
                                7.2500
                                          NaN
                                                     S
         0
                                                     C
                               71.2833
1
         0
                     PC 17599
                                          C85
2
         0
            STON/02. 3101282
                               7.9250
                                                     S
                                          NaN
                                                     S
3
         0
                       113803
                               53.1000
                                         C123
4
         0
                       373450
                                8.0500
                                          NaN
                                                     S
                                                   . . .
                                                     S
         0
                       211536
                               13.0000
                                          NaN
886
                                                     S
887
         0
                       112053
                               30.0000
                                          B42
                                                     S
888
         2
                  W./C. 6607
                               23.4500
                                          NaN
                                                     C
                       111369
                               30.0000
                                         C148
889
         0
890
                       370376
                              7.7500
                                          NaN
                                                     0
[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
1
               2
                          1
                                  1
2
               3
                          1
                                  3
3
               4
                          1
                                  1
4
               5
                                  3
                          0
                                  2
                          0
886
             887
                                  1
887
             888
                          1
                                  3
888
             889
                          0
                          1
                                  1
889
             890
                                  3
890
             891
                                                   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
                                                                   35
                                                            male
0
                                  Montvila, Rev. Juozas
                                                            male
                                                                   27
886
0
887
                           Graham, Miss. Margaret Edith
                                                          female
                                                                 19
0
888
              Johnston, Miss. Catherine Helen "Carrie"
                                                                   20
                                                          female
1
889
                                  Behr, Mr. Karl Howell
                                                            male
                                                                   26
0
                                    Dooley, Mr. Patrick
890
                                                            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
            STON/02. 3101282
                                                      S
         0
                                7.9250
                                          NaN
                                                           True
                                                                 False
3
         0
                       113803
                               53.1000
                                         C123
                                                      S
                                                           True
                                                                  False
4
         0
                       373450
                                8.0500
                                          NaN
                                                      S
                                                          False
                                                                  True
                                          . . .
                                                            . . .
                       211536
                                13.0000
                                                      S
                                                          False
                                                                  True
886
                                          NaN
         0
887
         0
                       112053
                               30.0000
                                          B42
                                                      S
                                                          True
                                                                  False
         2
                                                      S
888
                   W./C. 6607
                               23.4500
                                          NaN
                                                           True
                                                                 False
                                                      C
889
         0
                       111369
                               30.0000
                                         C148
                                                          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		
Futrelle, Mrs. Jacques Heath (Lily May Peel)	35	1
0		
4 Allen, Mr. William Henry	35	0
Θ		
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
890
                                     Dooley, Mr. Patrick
                                                             32
                                                                      0
0
                            Fare Cabin Embarked
                Ticket
                                                  female
0
            A/5 21171
                         7.2500
                                   NaN
                                                   False
                                               S
                                               C
1
              PC 17599
                        71.2833
                                   C85
                                                    True
2
                                               S
     STON/02. 3101282
                         7.9250
                                   NaN
                                                    True
3
                113803
                        53.1000
                                  C123
                                               S
                                                    True
4
                                               S
                373450
                         8.0500
                                   NaN
                                                   False
                                   . . .
                211536
                        13.0000
                                   NaN
                                               S
                                                   False
886
                                               S
                                                    True
887
                112053
                        30.0000
                                   B42
888
           W./C. 6607
                        23.4500
                                   NaN
                                               S
                                                    True
889
                111369
                        30.0000
                                  C148
                                               C
                                                   False
890
                370376
                         7.7500
                                               0
                                                   False
                                   NaN
[891 rows x 12 columns]
df["Embarked_cat"] = df["Embarked"].astype('category')
df["Embarked num"] = df["Embarked cat"].cat.codes
df
                   Survived Pclass \
     PassengerId
0
                                   3
                1
                           0
1
                2
                           1
                                   1
2
                                   3
                3
                           1
3
                                   1
                4
                          1
4
                5
                                   3
                           0
886
              887
                           0
                                   2
                                   1
887
              888
                           1
                                   3
888
              889
                           0
                                   1
889
              890
                           1
                                   3
              891
890
                                                     Name
                                                               Sex Age
SibSp \
                                 Braund, Mr. Owen Harris
                                                                      22
0
                                                              male
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
                                                                     35
                                                              male
0
886
                                   Montvila, Rev. Juozas
                                                              male
                                                                     27
887
                            Graham, Miss. Margaret Edith
                                                                     19
                                                            female
0
888
               Johnston, Miss. Catherine Helen "Carrie" female
                                                                     20
1
                                   Behr, Mr. Karl Howell
889
                                                              male
                                                                     26
0
890
                                     Dooley, Mr. Patrick
                                                                     32
                                                              male
0
     Parch
                       Ticket
                                   Fare Cabin Embarked Embarked cat \
0
         0
                    A/5 21171
                                 7.2500
                                           NaN
                                                      S
                                                                    C
1
         0
                     PC 17599
                                71.2833
                                           C85
                                                      C
                                                                     S
2
         0
            STON/02. 3101282
                                7.9250
                                           NaN
                                                      S
                                                                    S
3
                                                      S
         0
                       113803
                                53.1000
                                          C123
                                                                    S
4
                                                      S
         0
                       373450
                                 8.0500
                                           NaN
                                           . . .
                                                      S
                                                                     S
886
                       211536
                                13.0000
         0
                                           NaN
                                                                    S
S
                                                      S
887
                       112053
                                30.0000
                                           B42
         0
                                                      S
         2
888
                   W./C. 6607
                                23.4500
                                           NaN
                                                                    C
         0
                       111369
                                                      C
889
                                30.0000
                                          C148
                                                                    0
890
         0
                       370376
                               7.7500
                                           NaN
                                                      0
     Embarked num
0
                 0
1
2
                 2
3
                 2
4
                 2
. .
               . . .
886
                 2
                 2
887
                 2
888
                 0
889
                 1
890
[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
```

0 1	Passen	gerId Su 1 2	rvived F 0 1	oclass 3 1	\			
2 3 4		3 4 5	1 1 0	3 1 3				
886 887 888		887 888 889	0 1 0	2 1 3				
889 890		890 891	1 0	1 3				
C : I C						Name	Sex	Age
SibS 0	p \			Braun	d, Mr. Ow	en Harris	male	22
1 1 1	Cuming	s, Mrs. Jo	ohn Bradl	ey (Flo	rence Bri	ggs Th	female	38
2				Heik	kinen, Mi	ss. Laina	female	26
0	F	utrelle, M	1rs. Jaco	ues Hea	th (Lily	May Peel)	female	35
1 4				Allen,	Mr. Will	iam Henry	male	35
0								
								27
886 0				Mon	tvila, Re	v. Juozas	male	27
887 0			Gra	nham, Mi	ss. Marga	ret Edith	female	19
888		Johnsto	on, Miss.	Cather	ine Helen	"Carrie"	female	20
1 889				Beh	r, Mr. Ka	rl Howell	male	26
0 890 0				D	ooley, Mr	. Patrick	male	32
	Parch		Ticket	Far	e Cahin F	mbarked Em	harked c	at \
0	0		/5 21171	7.250	0 NaN	S	bar kea_e	_
1 2	0 0	STON/02.	PC 17599 3101282	71.283 7.925		C S		S
2 3 4	0 0		113803 373450	53.100 8.050	0 C123	S S S		S C S S
 886			211536	13.000		 S	•	 S
887 888	0 2	W	112053 /C. 6607	30.000 23.450		S S C		S S S C
889	0	Wij	111369	30.000	0 C148			C
890	0		370376	7.750	0 NaN	Q		Q

```
Embarked num Labelencoding Embarked
0
                 2
1
                 0
                                          0
2
                 2
                                          2
                                          2
3
                 2
                                          2
4
                 2
                                          2
                 2
886
                                          2
                 2
887
                 2
                                          2
888
                                          0
889
                 0
890
                 1
[891 rows x 15 columns]
df['Labelencoding_Embarked'].value_counts()
Labelencoding_Embarked
2
     644
0
     168
1
      77
Name: count, dtype: int64
# replacing values
df['Sex'] = df['Sex'].replace(['male', 'female'], [0, 1])
df
                   Survived Pclass \
     PassengerId
0
                                   3
                1
1
                2
                                   1
                          1
2
                                   3
                3
                          1
3
                                   1
                4
                          1
4
                5
                          0
                                   3
886
                                   2
             887
                          0
                                   1
887
             888
                          1
                                   3
888
             889
                          0
                                   1
889
             890
                          1
                                   3
890
             891
                                                     Name
                                                           Sex
                                                                Age
SibSp \
                                 Braund, Mr. Owen Harris
                                                                 22
1
1
     Cumings, Mrs. John Bradley (Florence Briggs Th...
                                                                 38
1
                                  Heikkinen, Miss. Laina
2
0
```

```
3
           Futrelle, Mrs. Jacques Heath (Lily May Peel)
                                                                     35
1
4
                                 Allen, Mr. William Henry
                                                                     35
0
886
                                     Montvila, Rev. Juozas
                                                                     27
0
887
                             Graham, Miss. Margaret Edith
                                                                     19
               Johnston, Miss. Catherine Helen "Carrie"
888
                                                                     20
1
889
                                     Behr, Mr. Karl Howell
                                                                     26
0
890
                                       Dooley, Mr. Patrick
                                                                     32
0
     Parch
                        Ticket
                                     Fare Cabin Embarked Embarked cat
0
                     A/5 21171
                                  7.2500
                                             NaN
                                                         S
          0
                                                                       S
                                                         C
                                                                       C
1
                      PC 17599
                                 71.2833
                                             C85
          0
                                                                       S
2
                                                         S
             STON/02. 3101282
                                  7.9250
          0
                                             NaN
                                                                       S
3
                                                         S
                        113803
          0
                                 53.1000
                                           C123
                                                                       S
4
                                                         S
          0
                        373450
                                  8.0500
                                             NaN
                                                                       ..
S
S
                                                        . .
S
                                 13.0000
886
          0
                        211536
                                             NaN
                        112053
                                 30.0000
                                            B42
                                                         S
887
          0
                                                         S
888
          2
                    W./C. 6607
                                 23.4500
                                            NaN
                                                                       C
889
          0
                        111369
                                 30.0000
                                           C148
                                                         C
                                                                       Q
890
                        370376
                                  7.7500
                                             NaN
     Embarked_num
                     Labelencoding_Embarked
                  2
                                             2
1
                  0
                                             0
2
                  2
                                             2
3
                  2
                                             2
4
                  2
                                             2
886
                  2
                                             2
                  2
                                            2
887
                  2
                                             2
888
889
                  0
                                            0
890
                  1
[891 rows x 15 columns]
```