Name : Umesh Mali Div :- A Roll No : 43

Practical No:- 1

Perform the following operations using Python on any open source dataset (e.g., data.csv) the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the

1. **Import all the required Python Libraries.**
2. **Locate an 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.**
4. **Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.**
5. **Data Formatting and Data Normalization: Summarize the types of variables by checking**

data set. If variables are not in the correct data type, apply proper type conversions. 6. Turn categorical variables into quantitative variables in Python.

---

1) Import all the required Python Libraries.

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

2) Locate an 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).

dataset URL:-https://www.kaggle.com/datasets/swatikhedekar/python-project-on-weather-dataset/data

3) Load the Dataset into pandas dataframe.

In [2]:
```
df = pd.read_csv("Titanic.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S

In [4]:
```
df.tail()
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
707	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.125	Q
708	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.000	S
709	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.000	S
710	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.000	C
711	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	Q

4) Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.

Check for missing values

In [5]:
```
df.isnull().sum()
```

Out[5]:

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	0
SibSp	0
Parch	0
Ticket	0
Fare	0
Embarked	0

dtype: int64

Geting describe/statistics :

In [6]:
```
df.describe()
```

Out[6]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	712.000000	712.000000	712.000000	712.000000	712.000000	712.000000	712.000000
mean	448.589888	0.404494	2.240169	29.642093	0.514045	0.432584	34.567251
std	258.683191	0.491139	0.836854	14.492933	0.930692	0.854181	52.938648
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	222.750000	0.000000	1.000000	20.000000	0.000000	0.000000	8.050000
50%	445.000000	0.000000	2.000000	28.000000	0.000000	0.000000	15.645850
75%	677.250000	1.000000	3.000000	38.000000	1.000000	1.000000	33.000000
max	891.000000	1.000000	3.000000	80.000000	5.000000	6.000000	512.329200

Check dimensions of the data frame

In [7]:
```
df.shape
```

Out[7]: (712, 11)

5) 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.

Convert 'Age' to int64 data type

In [8]:
```
df['Age'] = df['Age'].astype('int64')
```

In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 712 entries, 0 to 711 Data
columns (total 11 columns):
 #    Column          Non-Null Count  Dtype
-----  -------         --------------  ------
 0    PassengerId     712  non-null   int64
 1    Survived        712  non-null   int64
 2    Pclass          712  non-null   int64
 3    Name            712  non-null   object
 4    Sex             712  non-null   object
 5    Age             712  non-null   int64
 6    SibSp           712  non-null   int64
 7    Parch           712  non-null   int64
 8    Ticket          712  non-null   object
 9    Fare            712  non-null   float64
 10   Embarked        712  non-null   object
dtypes: float64(1), int64(6), object(4) memory
usage: 61.3+ KB
```

6) Turn categorical variables into quantitative variables in Python.

In [10]:
```
df = pd.get_dummies(df, columns=['Sex'], prefix='Sex')
```

In [11]:
```
df['Sex_female'] = df['Sex_female'].astype(int)
df['Sex_male'] = df['Sex_male'].astype(int) df.head()
```

Out[11]:

	PassengerId	Survived	Pclass	Name	Age	SibSp	Parch	Ticket	Fare	Embarked	Sex_female	Sex_male
0	1	0	3	Braund, Mr. Owen Harris	22	1	0	A/5 21171	7.2500	S	0	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38	1	0	PC 17599	71.2833	C	1	0
2	3	1	3	Heikkinen, Miss. Laina	26	0	0	STON/O2. 3101282	7.9250	S	1	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35	1	0	113803	53.1000	S	1	0
4	5	0	3	Allen, Mr. William Henry	35	0	0	373450	8.0500	S	0	1

In [12]:
```
df.drop(columns='Sex_male', inplace=True)
```

Out[12]:

	PassengerId	Survived	Pclass	Name	Age	SibSp	Parch	Ticket	Fare	Embarked	Sex_female
0	1	0	3	Braund, Mr. Owen Harris	22	1	0	A/5 21171	7.2500	S	0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38	1	0	PC 17599	71.2833	C	1
2	3	1	3	Heikkinen, Miss. Laina	26	0	0	STON/O2. 3101282	7.9250	S	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35	1	0	113803	53.1000	S	1
4	5	0	3	Allen, Mr. William Henry	35	0	0	373450	8.0500	S	0