```
import pandas as pd
In [15]:
         #to use python libraries we import the required libraries using "import"
         # NOTE : Python is case Sensitive
              'Name':['Alice','Bob','Charlie'],
             'Age':[25,30,35]
         df = pd.DataFrame(data)
         #print(df)
         #print(df.head(8))
         #print(df.tail(5))
         df = pd.read_csv('data.csv')
In [16]:
         print(df)
         #To print the data set uploaded
            ID
                           Age Gender
                                               City
                                                        Salary Experience
                    Name
         0
             1
                    John 28.0
                                  Male
                                           New York
                                                       50000.0
                                                                         3
             2
                          NaN Female Los Angeles
                                                       52000.0
                                                                         4
         1
                    Sara
         2
             3
                   David 35.0
                                                                         8
                                  Male
                                            Chicago
                                                           NaN
                                           New york
         3
             4
                   Linda 29.0 Female
                                                       58000.0
                                                                         5
         4
             5
                 Michael 41.0
                                                       62000.0
                                  Male
                                              Miami
                                                                        10
         5
                          NaN Female Los Angeles
             6
                    Emma
                                                       54000.0
                                                                         4
                  Robert 23.0
         6
             7
                                            CHICAGO
                                                       49000.0
                                                                         2
                                  Male
         7
             8
                  Olivia 32.0 Female
                                              Miami 1000000.0
                                                                         7
         8
             9
                   James 27.0
                                  Male
                                           New York
                                                       51000.0
                                                                         3
         9
            10 Patricia 38.0 Female Los Angeles
                                                           NaN
         print(df.to_string())
In [5]:
         #Use df.to_string() to print with some start value and a end value when the data is too large
               Name Age
         0
                      25
              Alice
         1
                Bob
                      30
         2 Charlie
                      35
In [17]:
         print(df.head(8))
            ID
                   Name
                          Age
                               Gender
                                              City
                                                       Salary
                                                               Experience
         0
                   John
                        28.0
                                 Male
                                          New York
                                                      50000.0
                                                                        3
             1
                          NaN Female Los Angeles
         1
             2
                   Sara
                                                      52000.0
                                                                        4
         2
             3
                         35.0
                                                                        8
                  David
                                 Male
                                           Chicago
                                                          NaN
         3
             4
                         29.0 Female
                                                      58000.0
                                                                        5
                  Linda
                                          New york
         4
             5 Michael 41.0
                                                                       10
                                 Male
                                             Miami
                                                      62000.0
         5
             6
                   Emma
                         NaN Female Los Angeles
                                                      54000.0
                                                                        4
             7
                                                                        2
         6
                 Robert 23.0
                                 Male
                                           CHICAGO
                                                      49000.0
         7
                                                                        7
             8
                 Olivia 32.0 Female
                                             Miami 1000000.0
         print(df.tail(5))
In [18]:
            ID
                    Name
                           Age
                                Gender
                                               City
                                                        Salary
                                                                Experience
         5
             6
                    Emma
                           NaN
                                Female Los Angeles
                                                       54000.0
                                                                         4
         6
             7
                  Robert 23.0
                                  Male
                                            CHICAGO
                                                       49000.0
                                                                         2
         7
             8
                                                                         7
                  Olivia 32.0 Female
                                              Miami 1000000.0
             9
                                                                         3
         8
                   James 27.0
                                  Male
                                           New York
                                                       51000.0
            10 Patricia 38.0 Female Los Angeles
                                                           NaN
In [19]: print(df.info())
```

```
RangeIndex: 10 entries, 0 to 9
         Data columns (total 7 columns):
                          Non-Null Count Dtype
              Column
          0
              ID
                          10 non-null
                                           int64
                                           object
          1
                          10 non-null
              Name
          2
                          8 non-null
                                           float64
              Age
          3
              Gender
                          10 non-null
                                           object
              City
                          10 non-null
                                           object
          5
                          8 non-null
                                           float64
              Salary
              Experience 10 non-null
                                           int64
         dtypes: float64(2), int64(2), object(3)
         memory usage: 692.0+ bytes
         None
In [20]:
         print(df.describe())
                                                      Experience
                      ID
                                 Age
                                              Salary
         count 10.00000
                           8.000000
                                            8.000000
                                                       10.000000
                 5.50000 31.625000
                                       172000.000000
                                                        5.500000
         mean
         std
                 3.02765
                           6.045955
                                       334591.008333
                                                        2.798809
                 1.00000 23.000000
                                        49000.000000
                                                        2.000000
         min
         25%
                 3.25000 27.750000
                                                        3.250000
                                        50750.000000
         50%
                 5.50000 30.500000
                                        53000.000000
                                                        4.500000
         75%
                 7.75000 35.750000
                                        59000.000000
                                                        7.750000
                10.00000 41.000000 1000000.000000
         max
                                                       10.000000
         print(df.isnull().sum())
In [22]:
         #Number of all null values in the data set
         ID
                       0
                       0
         Name
         Age
                       2
         Gender
                       0
                       0
         City
                       2
         Salary
         Experience
         dtype: int64
         a =df['Age'].mean()#mean of age
In [23]:
         print(a)
         31.625
         b = df['Salary'].mean()
In [25]:
         print(b)
         #sum() Sum
         #std standard deviation
         172000.0
In [26]:
         #data types of data
         print(df.dtypes)
         ID
                         int64
         Name
                        object
         Age
                       float64
         Gender
                        object
         City
                        object
         Salary
                       float64
         Experience
                         int64
         dtype: object
In [27]:
         #print columns in data set
         print(df.columns)
         Index(['ID', 'Name', 'Age', 'Gender', 'City', 'Salary', 'Experience'], dtype='object')
```

<class 'pandas.core.frame.DataFrame'>

```
df.columns = df.columns.str.upper()
         print(df)
            ID
                    NAME
                           AGE GENDER
                                              CITY
                                                       SALARY EXPERIENCE
         0
             1
                    John 28.0
                                 Male
                                           New York
                                                      50000.0
                                                                        3
         1
             2
                    Sara
                          NaN Female Los Angeles
                                                      52000.0
                                                                        4
         2
                   David 35.0
                                           Chicago
             3
                                 Male
                                                          NaN
                                                                        8
                   Linda 29.0 Female
         3
             4
                                                                        5
                                           New york
                                                      58000.0
         4
            5
                                                                       10
                 Michael 41.0
                                 Male
                                             Miami
                                                      62000.0
         5
                    Emma
                          NaN Female Los Angeles
                                                      54000.0
                                                                        4
            6
            7
                  Robert 23.0
                                                                        2
         6
                                 Male
                                           CHICAGO
                                                      49000.0
         7
             8
                  Olivia 32.0 Female
                                             Miami 1000000.0
                                                                        7
         8
             9
                   James 27.0
                                 Male
                                          New York
                                                      51000.0
                                                                        3
                                                                        9
         9
           10 Patricia 38.0 Female Los Angeles
                                                          NaN
         #display the unique values in data set
In [39]:
         print(df.nunique())
         ID
                       10
         NAME
                       10
         AGE
                        8
         GENDER
                        2
         CITY
                        6
         SALARY
                        8
         EXPERIENCE
                        8
         dtype: int64
In [36]:
         #check null value in column in each
         print(df.isnull().sum()/len(df)*100)
         ID
                        0.0
         NAME
                        0.0
         AGE
                       20.0
         GENDER
                        0.0
                        0.0
         CITY
         SALARY
                       20.0
                        0.0
         EXPERIENCE
         dtype: float64
In [40]:
         #check the status of null values in data set
         print(df.isnull())
               ID
                            AGE GENDER
                                         CITY SALARY EXPERIENCE
                    NAME
           False
                  False False
                                 False False
                                                False
                                                            False
           False False
                         True
                                  False False
                                                 False
                                                            False
         2 False False False
                                 False False
                                                 True
                                                            False
         3
            False False False
                                 False False
                                                False
                                                            False
         4 False False False
                                 False False
                                                False
                                                            False
         5 False False
                         True
                                 False False
                                                False
                                                            False
         6 False False False
                                 False False
                                                False
                                                            False
         7 False False False
                                 False False
                                                False
                                                            False
         8 False False False
                                  False False
                                                            False
                                                False
         9 False False False
                                 False False
                                                 True
                                                            False
         #print total no of columns and rows in data set
In [41]:
         print(df.shape)
         #10 rows and 7 columns)
         (10, 7)
         #remove null values from the data set and print
In [44]:
         print(df)
         print()
         df1= df.dropna()#remove null vaues in columns
         print(df1)
```

In [37]:

#convert Column heading to capital letters

```
2
             3
                   David 35.0
                                  Male
                                            Chicago
                                                                          8
                                                            NaN
         3
                   Linda 29.0 Female
                                                                          5
             4
                                            New york
                                                        58000.0
         4
             5
                 Michael 41.0
                                  Male
                                              Miami
                                                        62000.0
                                                                         10
                           NaN Female Los Angeles
         5
                    Emma
                                                        54000.0
                                                                          4
             6
         6
             7
                  Robert 23.0
                                  Male
                                            CHICAGO
                                                        49000.0
                                                                          2
         7
             8
                  Olivia 32.0 Female
                                              Miami 1000000.0
                                                                          7
         8
             9
                   James 27.0
                                  Male
                                            New York
                                                        51000.0
                                                                          3
         9
                Patricia 38.0 Female Los Angeles
            10
                                                            NaN
                                                    SALARY EXPERIENCE
            ID
                   NAME
                          AGE GENDER
                                            CITY
         0
             1
                   John 28.0
                                 Male New York
                                                    50000.0
         3
                  Linda
                         29.0 Female
                                                    58000.0
                                                                      5
                                       New york
         4
                         41.0
                                                                     10
             5 Michael
                                 Male
                                           Miami
                                                    62000.0
         6
             7
                 Robert 23.0
                                         CHICAGO
                                                    49000.0
                                                                      2
                                 Male
         7
                 Olivia 32.0 Female
                                          Miami 1000000.0
                                                                      7
         8
                        27.0
                                 Male New York
                                                                      3
                  James
                                                    51000.0
         #copy dataframe to dataframe2
In [45]:
         df2 = df
         print(df)
         print()
         print(df2)
                                                        SALARY EXPERIENCE
            TD
                    NAME
                           AGE GENDER
                                                CITY
                    John 28.0
                                            New York
                                                        50000.0
         0
             1
                                   Male
                                                                          3
         1
             2
                    Sara
                           NaN Female Los Angeles
                                                        52000.0
                                                                          4
         2
                   David 35.0
                                                                          8
             3
                                  Male
                                            Chicago
                                                            NaN
         3
                   Linda 29.0
                               Female
                                            New york
                                                        58000.0
                                                                          5
         4
                 Michael 41.0
             5
                                  Male
                                              Miami
                                                        62000.0
                                                                         10
         5
             6
                    Fmma
                           NaN Female Los Angeles
                                                        54000.0
                                                                          4
         6
             7
                  Robert 23.0
                                  Male
                                            CHICAGO
                                                        49000.0
                                                                          2
         7
             8
                  Olivia 32.0 Female
                                               Miami 1000000.0
                                                                          7
             9
                   James 27.0
                                                                          3
         8
                                   Male
                                            New York
                                                        51000.0
         9
            10
                Patricia 38.0 Female Los Angeles
                                                            NaN
            ID
                    NAME
                           AGE
                                GENDER
                                                        SALARY
                                                                 EXPERIENCE
                                                CITY
         0
             1
                    John 28.0
                                  Male
                                            New York
                                                        50000.0
                                                                          3
                                                        52000.0
                                                                          4
         1
             2
                    Sara
                          NaN
                               Female Los Angeles
         2
                   David
                          35.0
                                   Male
                                            Chicago
                                                            NaN
                                                                          8
         3
                                                                          5
                   Linda 29.0
                               Female
                                            New york
                                                        58000.0
         4
                 Michael 41.0
                                                                         10
             5
                                  Male
                                               Miami
                                                        62000.0
         5
             6
                    Emma
                           NaN Female Los Angeles
                                                        54000.0
                                                                          4
         6
             7
                  Robert 23.0
                                  Male
                                            CHICAGO
                                                        49000.0
                                                                          2
         7
                                                                          7
             8
                  Olivia 32.0
                               Female
                                               Miami 1000000.0
         8
             9
                   James 27.0
                                                        51000.0
                                                                          3
                                   Male
                                            New York
                          38.0 Female Los Angeles
                                                                          9
            10
                Patricia
                                                            NaN
In [47]:
         #data Cleaning
         #Fill the missing values in the data set with 0
         df2.fillna(0,inplace=True)# opr perform to original dataset without creating or returning a n
         print(df2)
            ID
                    NAME
                                                CITY
                           AGE GENDER
                                                         SALARY EXPERIENCE
         0
             1
                    John
                          28.0
                                  Male
                                            New York
                                                        50000.0
                                                                          3
         1
                    Sara
                           0.0
                               Female
                                        Los Angeles
                                                        52000.0
                                                                          4
         2
                         35.0
                                            Chicago
                                                                          8
             3
                   David
                                  Male
                                                            0.0
         3
                   Linda 29.0 Female
                                                                          5
             4
                                            New york
                                                        58000.0
         4
             5
                 Michael 41.0
                                  Male
                                               Miami
                                                        62000.0
                                                                         10
         5
             6
                    Emma
                           0.0 Female Los Angeles
                                                        54000.0
                                                                          4
             7
         6
                  Robert 23.0
                                             CHICAGO
                                                                          2
                                  Male
                                                        49000.0
         7
                                                                          7
             8
                  Olivia
                          32.0
                                Female
                                               Miami 1000000.0
         8
             9
                   James
                          27.0
                                   Male
                                            New York
                                                        51000.0
                                                                          3
                                                                          9
            10
                Patricia 38.0 Female Los Angeles
                                                            0.0
         #Fill missing values with mean of the column
In [60]:
```

ID

1

0

1

NAME

Sara

#df.fillna(df.mean())

John 28.0

AGE

**GENDER** 

Male

NaN Female Los Angeles

CITY

New York

SALARY

50000.0

52000.0

**EXPERIENCE** 

3

```
TD
                    NAME
                          AGE GENDER
                                                       SALARY EXPERIENCE
                                              CITY
            5
                Michael 41.0
                                             Miami
                                                      62000.0
                                                                      10
         4
                                 Male
         9
            10 Patricia 38.0 Female Los Angeles
                                                         0.0
         2
                  David 35.0
                                                                       8
                                 Male
                                           Chicago
                                                          0.0
         7
             8
                  Olivia 32.0 Female
                                             Miami 1000000.0
                                                                       7
         3
                  Linda 29.0 Female
                                                                       5
            4
                                          New york
                                                      58000.0
         0
                   John 28.0
                                                                       3
            1
                                 Male
                                          New York
                                                      50000.0
         8
            9
                  James 27.0
                                 Male
                                          New York 51000.0
                                                                       3
            7
         6
                  Robert 23.0
                                 Male
                                           CHICAGO
                                                     49000.0
                                                                       2
         5
                   Emma
                         0.0 Female Los Angeles
                                                      54000.0
                                                                       4
            6
         1
             2
                   Sara
                          0.0 Female Los Angeles
                                                      52000.0
                                                                       4
In [55]:
         #Sort the values in the dataset based on age in ascending order
         #df.sort_values("Age",inplace=True)
         df.sort_values("AGE",inplace=True)
         print(df)
            ID
                                                      SALARY EXPERIENCE
                    NAME
                          AGE GENDER
                                              CITY
            2
                   Sara
                          0.0 Female Los Angeles
                                                      52000.0
         1
                          0.0 Female Los Angeles
         5
                                                      54000.0
                                                                       4
             6
                   Emma
                  Robert 23.0
         6
            7
                                 Male
                                           CHICAGO
                                                      49000.0
                                                                       2
         8
            9
                  James 27.0
                                 Male
                                          New York
                                                      51000.0
                                                                       3
         0
            1
                   John 28.0
                                 Male
                                          New York
                                                      50000.0
                                                                       3
         3
                  Linda 29.0 Female
                                                      58000.0
                                          New york
                                                                       7
         7
            8
                 Olivia 32.0 Female
                                             Miami 1000000.0
         2
                  David 35.0
            3
                                 Male
                                           Chicago
                                                                       8
                                                         0.0
         9
            10 Patricia 38.0 Female Los Angeles
                                                                       9
                                                         0.0
             5
                Michael 41.0
                                             Miami
                                                      62000.0
                                                                      10
                                 Male
In [58]:
         #Sort values in data set in desc order
         df.sort_values("AGE",ascending=False,inplace=True)
         print(df)
            ID
                    NAME AGE GENDER
                                              CITY
                                                       SALARY EXPERIENCE
                Michael 41.0
                                 Male
                                             Miami
                                                      62000.0
                                                                      10
                                                                       9
         9
            10 Patricia 38.0 Female Los Angeles
                                                          0.0
         2
            3
                  David 35.0
                                 Male
                                           Chicago
                                                         0.0
                                                                       8
         7
                 Olivia 32.0 Female
                                                                       7
            8
                                             Miami 1000000.0
         3
            4
                                                                       5
                  Linda 29.0 Female
                                          New york
                                                     58000.0
         0
            1
                   John 28.0
                                 Male
                                          New York
                                                      50000.0
                                                                       3
         8
            9
                  James 27.0
                                 Male
                                          New York
                                                      51000.0
                                                                       3
            7
                                                                       2
         6
                  Robert
                         23.0
                                 Male
                                           CHICAGO
                                                     49000.0
         5
                                                                       4
             6
                    Emma
                          0.0 Female Los Angeles
                                                      54000.0
             2
                    Sara
                          0.0 Female Los Angeles
                                                   52000.0
In [59]:
         #Find duplicated values in the data set
         print(df.duplicated())
         #Return ture for every row that is a duplicate , otherwise False
         4
              False
         9
              False
         2
              False
         7
             False
         3
             False
         0
             False
         8
              False
              False
         6
         5
              False
         1
              False
         dtype: bool
```

#print(df)

df.fillna(df.mean(numeric\_only=True), inplace=True)

#This will work onlu if all columns are numeric