Foundations of Data Science using Python

Session 3: Handling Missing Data in Pandas

Why we get missing data?

Missing data is a very big problem in real life scenario. Missing data can occur when information is not provided for one or more items or for a whole unit. In real-time, while obtaining the datasets, there is a high probability where some values might be missing for various reasons. For example, a customer might not share his / her salary details, contact, address etc., in this manner, some of the attributes will have missing values. When we load the dataset in to a dataframe, missing data arrive with missing data, because the data exists but was not collected or it never existed. Missing data is a common scenario in datasets.

Missing Data representation in Pandas

In Pandas, the missing values are represented by two values:

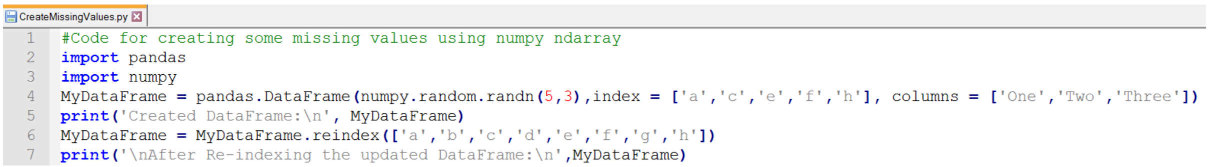
- **None:** None is a Python singleton object that is often used for missing data in Python code.
- **NaN:** NaN (an acronym for Not a Number), is a special floating-point value recognized by all systems that use the standard IEEE floating-point representation

Functions available for handling missing data in Pandas

In Pandas, the following functions facilitate us to work with missing values in the dataset:

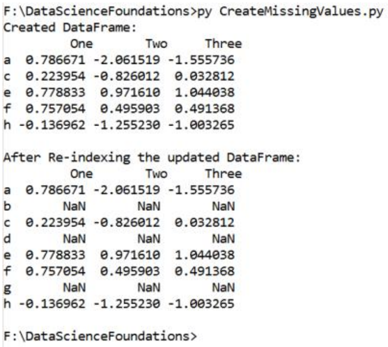
- isnull()
- notnull()
- dropna()
- fillna()
- replace()
- interpolate()

Creating a dataframe with missing values:



```
#Code for creating some missing values using numpy ndarray
import pandas
import numpy
MyDataFrame = pandas.DataFrame(numpy.random.randn(5,3),index = ['a','c','e','f','h'], columns = ['One','Two','Three'])
print('Created DataFrame:\n', MyDataFrame)
MyDataFrame = MyDataFrame.reindex(['a','b','c','d','e','f','g','h'])
print('\nAfter Re-indexing the updated DataFrame:\n',MyDataFrame)
```

Output:

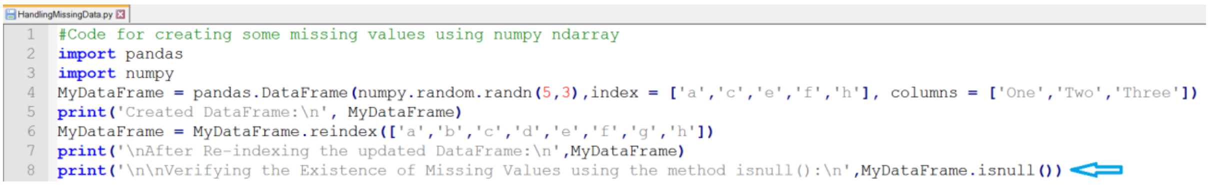


```
F:\DataScienceFoundations>py CreateMissingValues.py
Created DataFrame:
        One       Two     Three
a  0.786671 -2.061519 -1.555736
c  0.223954 -0.826012  0.032812
e  0.778833  0.971610  1.044038
f  0.757054  0.495903  0.491368
h -0.136962 -1.255230 -1.003265

After Re-indexing the updated DataFrame:
        One       Two     Three
a  0.786671 -2.061519 -1.555736
b       NaN       NaN       NaN
c  0.223954 -0.826012  0.032812
d       NaN       NaN       NaN
e  0.778833  0.971610  1.044038
f  0.757054  0.495903  0.491368
g       NaN       NaN       NaN
h -0.136962 -1.255230 -1.003265

F:\DataScienceFoundations>
```

Checking the existence of missing values:

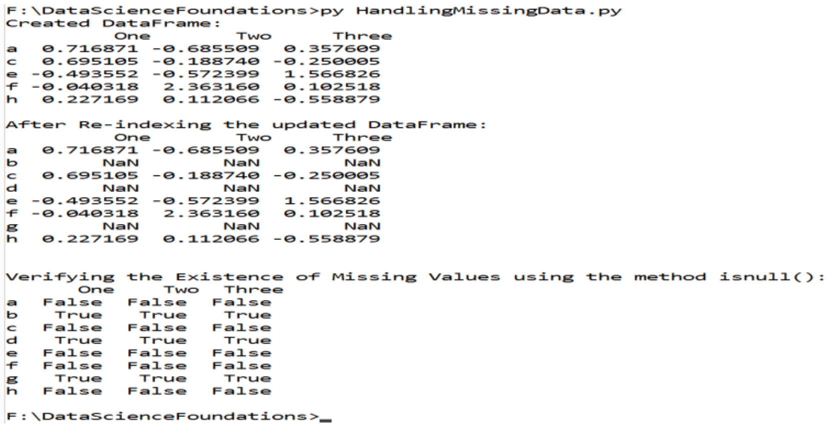
To detect the missing values, in Pandas, isnull() and notnull() methods are used. These methods return the Boolean value i.e., True or False depending on the instances in the dataframe. isnull() returns True in case of missing value and returns False in case if a value exists. notnull() returns False in case of missing value and returns True if the value exists.

Using isnull()



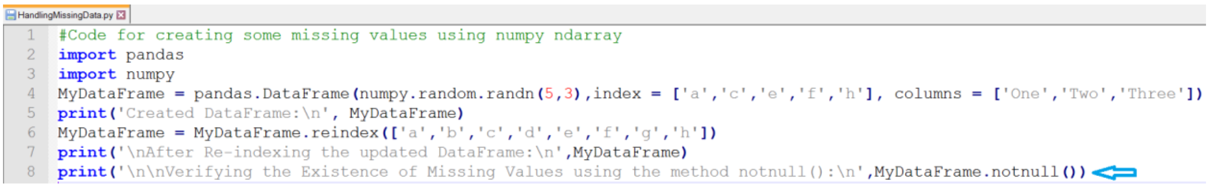
```
#Code for creating some missing values using numpy ndarray
import pandas
import numpy
MyDataFrame = pandas.DataFrame(numpy.random.randn(5,3),index = ['a','c','e','f','h'], columns = ['One','Two','Three'])
print('Created DataFrame:\n', MyDataFrame)
MyDataFrame = MyDataFrame.reindex(['a','b','c','d','e','f','g','h'])
print('\nAfter Re-indexing the updated DataFrame:\n',MyDataFrame)
print('\n\nVerifying the Existence of Missing Values using the method isnull():\n',MyDataFrame.isnull())
```

Output:



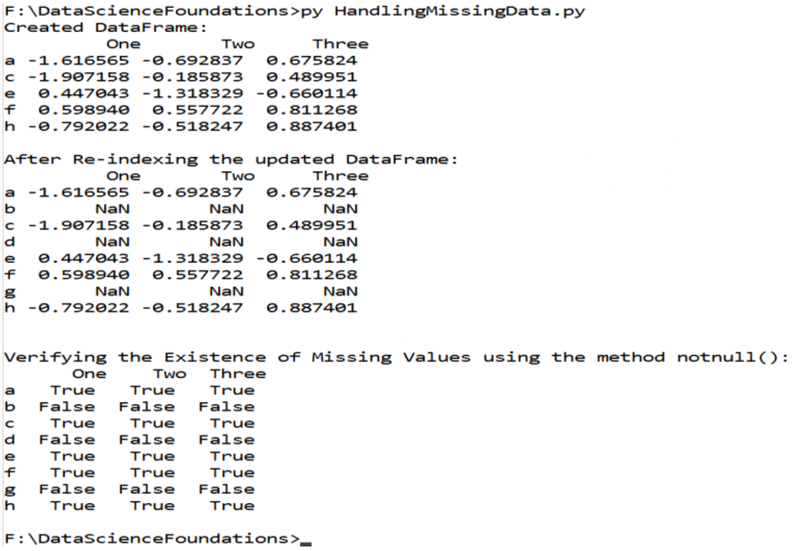
```
F:\DataScienceFoundations>py HandlingMissingData.py
Created DataFrame:
        One       Two     Three
a  0.716871 -0.685509  0.357609
c  0.695105 -0.188740 -0.250005
e -0.493552 -0.572399  1.566826
f -0.040318  2.363160  0.102518
h  0.227169  0.112066 -0.558879

After Re-indexing the updated DataFrame:
        One       Two     Three
a  0.716871 -0.685509  0.357609
b       NaN       NaN       NaN
c  0.695105 -0.188740 -0.250005
d       NaN       NaN       NaN
e -0.493552 -0.572399  1.566826
f -0.040318  2.363160  0.102518
g       NaN       NaN       NaN
h  0.227169  0.112066 -0.558879


Verifying the Existence of Missing Values using the method isnull():
     One    Two  Three
a  False  False  False
b   True   True   True
c  False  False  False
d   True   True   True
e  False  False  False
f  False  False  False
g   True   True   True
h  False  False  False

F:\DataScienceFoundations>_
```

Using notnull()



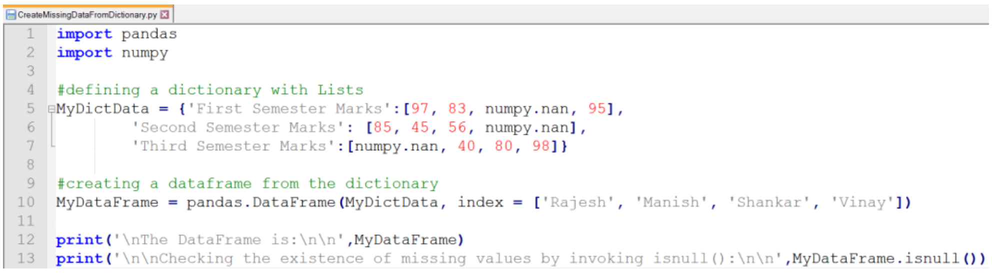
```
#Code for creating some missing values using numpy ndarray
import pandas
import numpy
MyDataFrame = pandas.DataFrame(numpy.random.randn(5,3),index = ['a','c','e','f','h'], columns = ['One','Two','Three'])
print('Created DataFrame:\n', MyDataFrame)
MyDataFrame = MyDataFrame.reindex(['a','b','c','d','e','f','g','h'])
print('\nAfter Re-indexing the updated DataFrame:\n',MyDataFrame)
print('\n\nVerifying the Existence of Missing Values using the method notnull():\n',MyDataFrame.notnull())
```

Output:



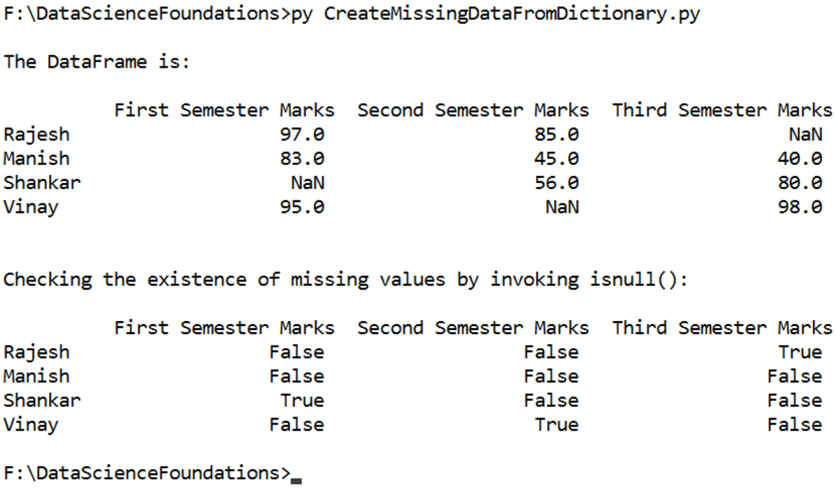
```
F:\DataScienceFoundations>py HandlingMissingData.py
Created DataFrame:
        One       Two     Three
a -1.616565 -0.692837  0.675824
c -1.907158 -0.185873  0.489951
e  0.447043 -1.318329 -0.660114
f  0.598940  0.557722  0.811268
h -0.792022 -0.518247  0.887401

After Re-indexing the updated DataFrame:
        One       Two     Three
a -1.616565 -0.692837  0.675824
b       NaN       NaN       NaN
c -1.907158 -0.185873  0.489951
d       NaN       NaN       NaN
e  0.447043 -1.318329 -0.660114
f  0.598940  0.557722  0.811268
g       NaN       NaN       NaN
h -0.792022 -0.518247  0.887401


Verifying the Existence of Missing Values using the method notnull():
     One    Two  Three
a   True   True   True
b  False  False  False
c   True   True   True
d  False  False  False
e   True   True   True
f   True   True   True
g  False  False  False
h   True   True   True

F:\DataScienceFoundations>_
```

Creating and Handling missing values from a Dictionary



```
import pandas
import numpy

#defining a dictionary with Lists
MyDictData = {'First Semester Marks':[97, 83, numpy.nan, 95],
        'Second Semester Marks': [85, 45, 56, numpy.nan],
        'Third Semester Marks':[numpy.nan, 40, 80, 98]}

#creating a dataframe from the dictionary
MyDataFrame = pandas.DataFrame(MyDictData, index = ['Rajesh', 'Manish', 'Shankar', 'Vinay'])

print('\nThe DataFrame is:\n\n',MyDataFrame)
print('\n\nChecking the existence of missing values by invoking isnull():\n\n',MyDataFrame.isnull())
```

Output:



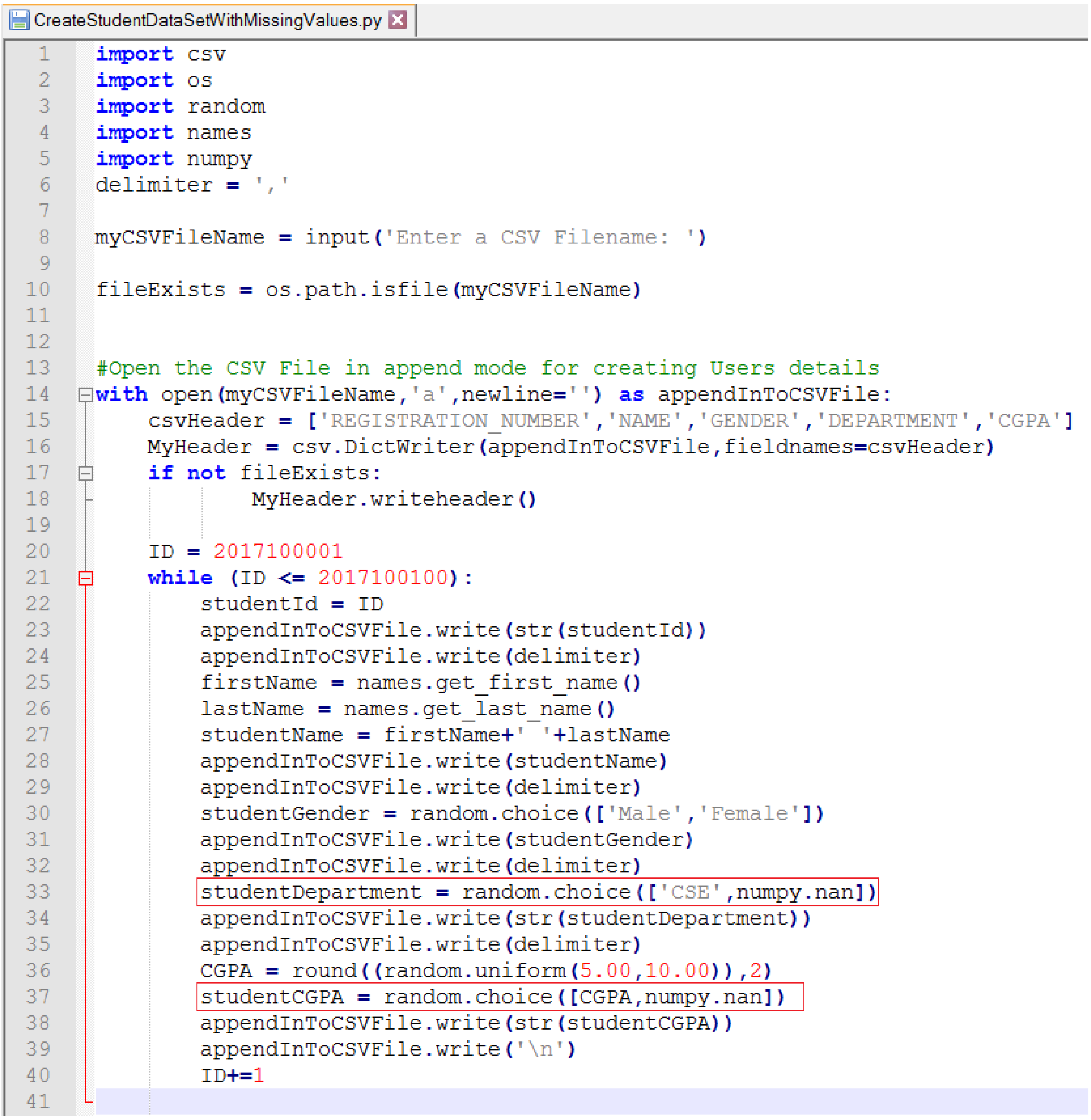
```
F:\DataScienceFoundations>py CreateMissingDataFromDictionary.py

The DataFrame is:

          First Semester Marks  Second Semester Marks  Third Semester Marks
Rajesh                   97.0                   85.0                   NaN
Manish                   83.0                   45.0                  40.0
Shankar                   NaN                   56.0                  80.0
Vinay                    95.0                    NaN                  98.0


Checking the existence of missing values by invoking isnull():

          First Semester Marks  Second Semester Marks  Third Semester Marks
Rajesh                  False                  False                  True
Manish                  False                  False                 False
Shankar                  True                  False                 False
Vinay                   False                   True                 False

F:\DataScienceFoundations>_
```

Handling missing values from a CSV File

Missing values is a common scenario, which we observe in the datasets obtained from various sources. After loading the CSV data into the dataframe, we can also identify and handle the missing values in the dataset.

Creating a CSV File with missing values:

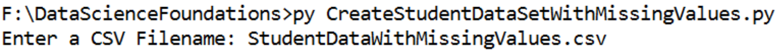
Let us consider a sample student data with 100 instances comprising of attributes – Registration Number, Name, Gender, Department and CGPA. With a specific purpose, in the python code shown below, randomly Department and CGPA values are taken as 'Nan' as highlighted in lines 33 and 37.



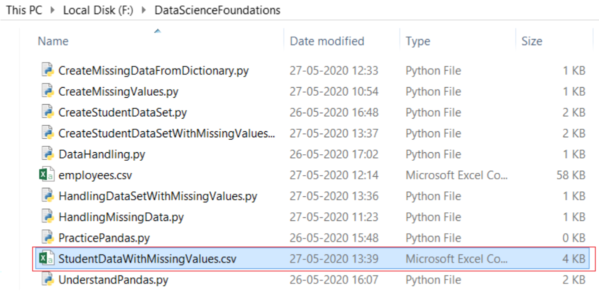
```
 1  import csv
 2  import os
 3  import random
 4  import names
 5  import numpy
 6  delimiter = ','
 7
 8  myCSVFileName = input('Enter a CSV Filename: ')
 9
10  fileExists = os.path.isfile(myCSVFileName)
11
12
13  #Open the CSV File in append mode for creating Users details
14  with open(myCSVFileName,'a',newline='') as appendInToCSVFile:
15      csvHeader = ['REGISTRATION_NUMBER','NAME','GENDER','DEPARTMENT','CGPA']
16      MyHeader = csv.DictWriter(appendInToCSVFile,fieldnames=csvHeader)
17      if not fileExists:
18              MyHeader.writeheader()
19
20      ID = 2017100001
21      while (ID <= 2017100100):
22          studentId = ID
23          appendInToCSVFile.write(str(studentId))
24          appendInToCSVFile.write(delimiter)
25          firstName = names.get_first_name()
26          lastName = names.get_last_name()
27          studentName = firstName+' '+lastName
28          appendInToCSVFile.write(studentName)
29          appendInToCSVFile.write(delimiter)
30          studentGender = random.choice(['Male','Female'])
31          appendInToCSVFile.write(studentGender)
32          appendInToCSVFile.write(delimiter)
33          studentDepartment = random.choice(['CSE',numpy.nan])
34          appendInToCSVFile.write(str(studentDepartment))
35          appendInToCSVFile.write(delimiter)
36          CGPA = round((random.uniform(5.00,10.00)),2)
37          studentCGPA = random.choice([CGPA,numpy.nan])
38          appendInToCSVFile.write(str(studentCGPA))
39          appendInToCSVFile.write('\n')
40          ID+=1
41
```

Output:

A CSV file ‘StudentDataWithMissingValues’ is created in the current directory with sample students’ data with missing values. Few records are shown below with marked missing values. The packages used in creating the sample students dataset are csv, os, random, names and numpy.

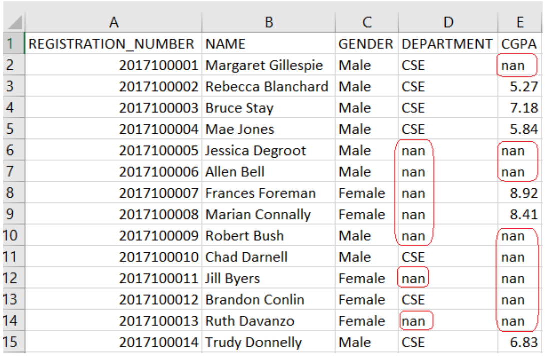


```
F:\DataScienceFoundations>py CreateStudentDataSetWithMissingValues.py
Enter a CSV Filename: StudentDataWithMissingValues.csv
```



This PC ▸ Local Disk (F:) ▸ DataScienceFoundations

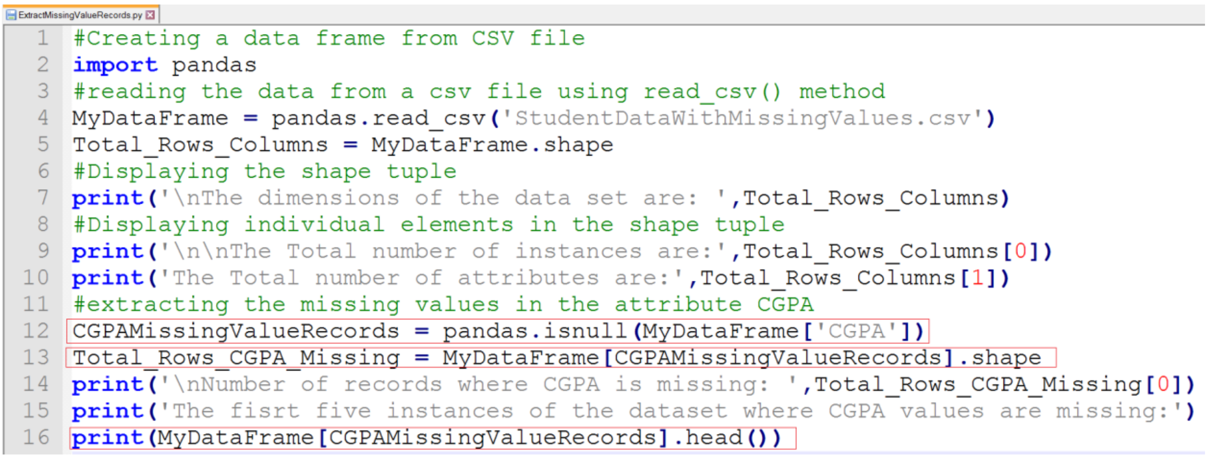
Name	Date modified	Type	Size
CreateMissingDataFromDictionary.py	27-05-2020 12:33	Python File	1 KB
CreateMissingValues.py	27-05-2020 10:54	Python File	1 KB
CreateStudentDataSet.py	26-05-2020 16:48	Python File	2 KB
CreateStudentDataSetWithMissingValues...	27-05-2020 13:37	Python File	2 KB
DataHandling.py	26-05-2020 17:02	Python File	1 KB
employees.csv	27-05-2020 12:14	Microsoft Excel Co...	58 KB
HandlingDataSetWithMissingValues.py	27-05-2020 13:36	Python File	1 KB
HandlingMissingData.py	27-05-2020 11:23	Python File	1 KB
PracticePandas.py	26-05-2020 15:48	Python File	0 KB
StudentDataWithMissingValues.csv	27-05-2020 13:39	Microsoft Excel Co...	4 KB
UnderstandPandas.py	26-05-2020 16:07	Python File	2 KB



	A	B	C	D	E
1	REGISTRATION_NUMBER	NAME	GENDER	DEPARTMENT	CGPA
2	2017100001	Margaret Gillespie	Male	CSE	nan
3	2017100002	Rebecca Blanchard	Male	CSE	5.27
4	2017100003	Bruce Stay	Male	CSE	7.18
5	2017100004	Mae Jones	Male	CSE	5.84
6	2017100005	Jessica Degroot	Male	nan	nan
7	2017100006	Allen Bell	Male	nan	nan
8	2017100007	Frances Foreman	Female	nan	8.92
9	2017100008	Marian Connally	Female	nan	8.41
10	2017100009	Robert Bush	Male	nan	nan
11	2017100010	Chad Darnell	Male	CSE	nan
12	2017100011	Jill Byers	Female	nan	nan
13	2017100012	Brandon Conlin	Female	CSE	nan
14	2017100013	Ruth Davanzo	Female	nan	nan
15	2017100014	Trudy Donnelly	Male	CSE	6.83

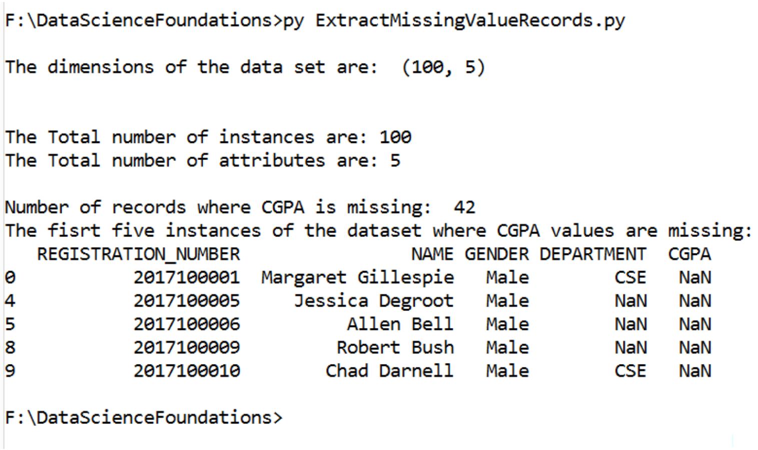
Extracting the data from the datasets with missing values

Let us find out the records where CGPA values are missing. Line 12 provides a Boolean series of records where CGPA is Nan. Line 13 provides the total number of instances available in the dataset where CGPA is missing. Line 14 displays the complete record information for the first five instances where CGPA values are missing.



```
 1  #Creating a data frame from CSV file
 2  import pandas
 3  #reading the data from a csv file using read_csv() method
 4  MyDataFrame = pandas.read_csv('StudentDataWithMissingValues.csv')
 5  Total_Rows_Columns = MyDataFrame.shape
 6  #Displaying the shape tuple
 7  print('\nThe dimensions of the data set are: ',Total_Rows_Columns)
 8  #Displaying individual elements in the shape tuple
 9  print('\n\nThe Total number of instances are:',Total_Rows_Columns[0])
10  print('The Total number of attributes are:',Total_Rows_Columns[1])
11  #extracting the missing values in the attribute CGPA
12  CGPAMissingValueRecords = pandas.isnull(MyDataFrame['CGPA'])
13  Total_Rows_CGPA_Missing = MyDataFrame[CGPAMissingValueRecords].shape
14  print('\nNumber of records where CGPA is missing: ',Total_Rows_CGPA_Missing[0])
15  print('The fisrt five instances of the dataset where CGPA values are missing:')
16  print(MyDataFrame[CGPAMissingValueRecords].head())
```

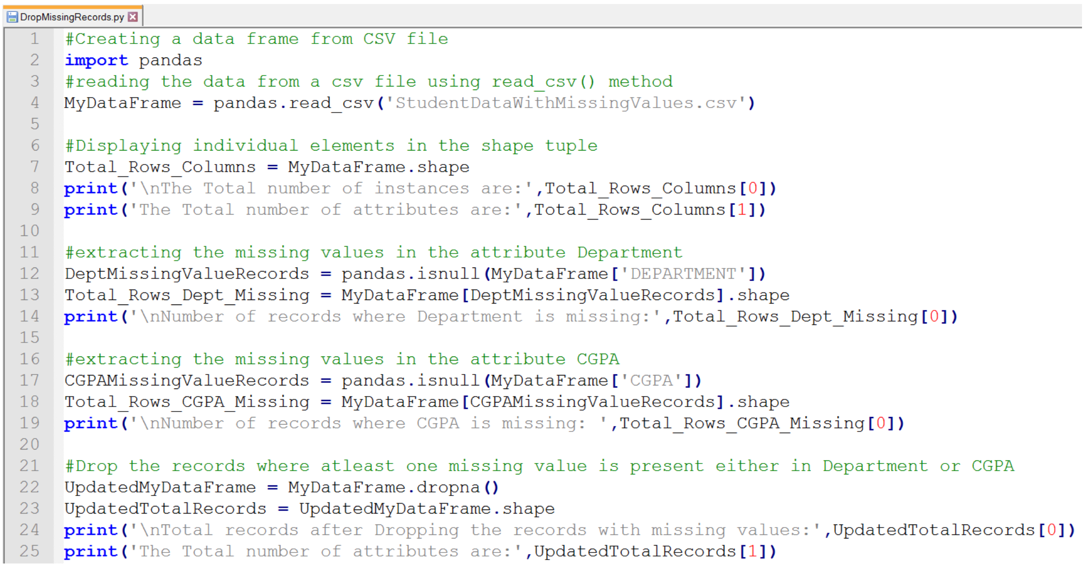
Output:



```
F:\DataScienceFoundations>py ExtractMissingValueRecords.py

The dimensions of the data set are:  (100, 5)


The Total number of instances are: 100
The Total number of attributes are: 5

Number of records where CGPA is missing:  42
The fisrt five instances of the dataset where CGPA values are missing:
   REGISTRATION_NUMBER                NAME GENDER DEPARTMENT  CGPA
0           2017100001  Margaret Gillespie   Male        CSE   NaN
4           2017100005     Jessica Degroot   Male        NaN   NaN
5           2017100006          Allen Bell   Male        NaN   NaN
8           2017100009         Robert Bush   Male        NaN   NaN
9           2017100010        Chad Darnell   Male        CSE   NaN

F:\DataScienceFoundations>
```

Dropping the records with at least one missing value

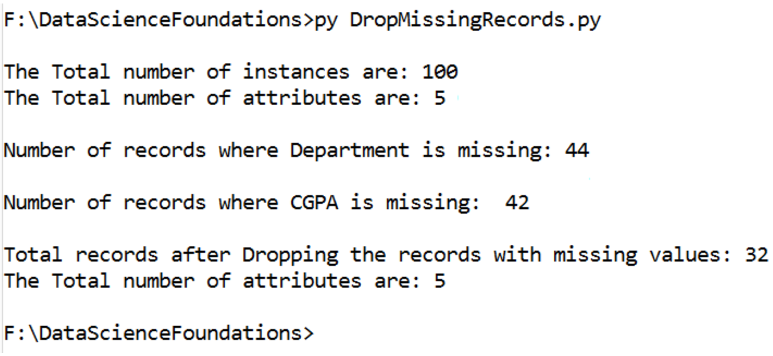
dropna() method identifies the missing values and drops the entire record in case if missing value exists.



```python
#Creating a data frame from CSV file
import pandas
#reading the data from a csv file using read_csv() method
MyDataFrame = pandas.read_csv('StudentDataWithMissingValues.csv')

#Displaying individual elements in the shape tuple
Total_Rows_Columns = MyDataFrame.shape
print('\nThe Total number of instances are:',Total_Rows_Columns[0])
print('The Total number of attributes are:',Total_Rows_Columns[1])

#extracting the missing values in the attribute Department
DeptMissingValueRecords = pandas.isnull(MyDataFrame['DEPARTMENT'])
Total_Rows_Dept_Missing = MyDataFrame[DeptMissingValueRecords].shape
print('\nNumber of records where Department is missing:',Total_Rows_Dept_Missing[0])

#extracting the missing values in the attribute CGPA
CGPAMissingValueRecords = pandas.isnull(MyDataFrame['CGPA'])
Total_Rows_CGPA_Missing = MyDataFrame[CGPAMissingValueRecords].shape
print('\nNumber of records where CGPA is missing: ',Total_Rows_CGPA_Missing[0])

#Drop the records where atleast one missing value is present either in Department or CGPA
UpdatedMyDataFrame = MyDataFrame.dropna()
UpdatedTotalRecords = UpdatedMyDataFrame.shape
print('\nTotal records after Dropping the records with missing values:',UpdatedTotalRecords[0])
print('The Total number of attributes are:',UpdatedTotalRecords[1])
```

Output: 44 and 42 records with missing values in Department and CGPA respectively.

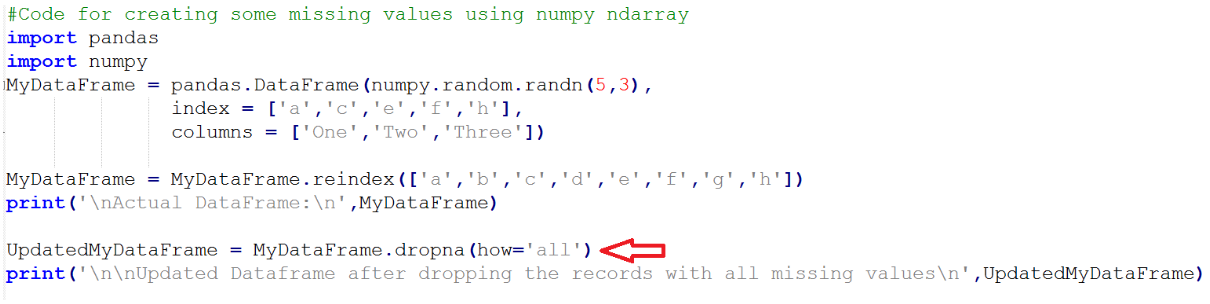
Total records with missing values in either Department or CGPA are 68.



```
F:\DataScienceFoundations>py DropMissingRecords.py

The Total number of instances are: 100
The Total number of attributes are: 5

Number of records where Department is missing: 44

Number of records where CGPA is missing:  42

Total records after Dropping the records with missing values: 32
The Total number of attributes are: 5

F:\DataScienceFoundations>
```

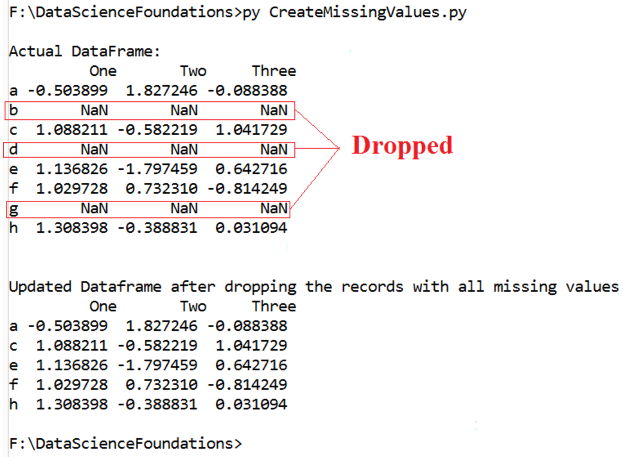
Dropping the records when all values are missing

dropna(how = ‘all) method identifies the records where all the values are missing and drops the entire record.



```
#Code for creating some missing values using numpy ndarray
import pandas
import numpy
MyDataFrame = pandas.DataFrame(numpy.random.randn(5,3),
                 index = ['a','c','e','f','h'],
                 columns = ['One','Two','Three'])

MyDataFrame = MyDataFrame.reindex(['a','b','c','d','e','f','g','h'])
print('\nActual DataFrame:\n',MyDataFrame)

UpdatedMyDataFrame = MyDataFrame.dropna(how='all')
print('\n\nUpdated Dataframe after dropping the records with all missing values\n',UpdatedMyDataFrame)
```

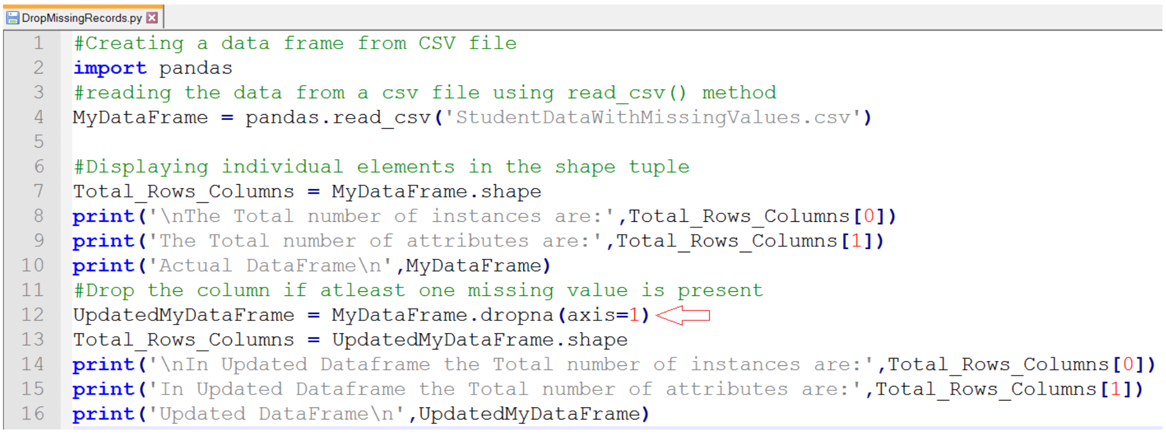
Output:



```
F:\DataScienceFoundations>py CreateMissingValues.py

Actual DataFrame:
         One       Two     Three
a -0.503899  1.827246 -0.088388
b       NaN       NaN       NaN
c  1.088211 -0.582219  1.041729
d       NaN       NaN       NaN
e  1.136826 -1.797459  0.642716
f  1.029728  0.732310 -0.814249
g       NaN       NaN       NaN
h  1.308398 -0.388831  0.031094


Updated Dataframe after dropping the records with all missing values
         One       Two     Three
a -0.503899  1.827246 -0.088388
c  1.088211 -0.582219  1.041729
e  1.136826 -1.797459  0.642716
f  1.029728  0.732310 -0.814249
h  1.308398 -0.388831  0.031094

F:\DataScienceFoundations>
```

Dropping columns that have at least one missing value

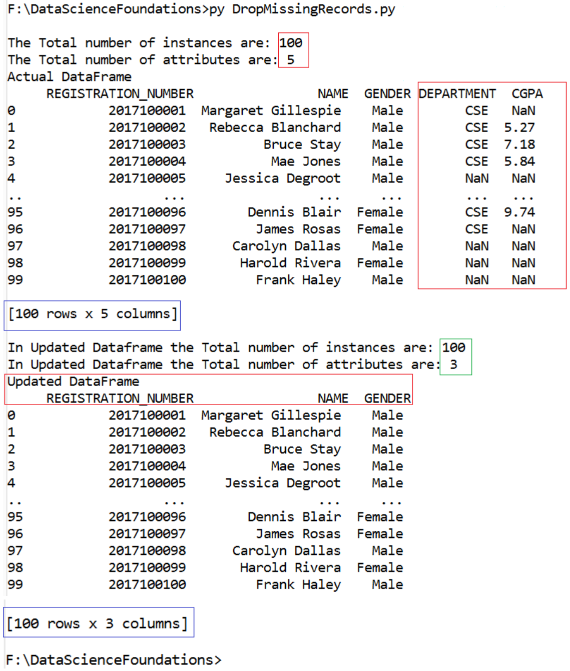
dropna(axis=1) method drops the entire column in case if at least one missing value is found in any column.



```
#Creating a data frame from CSV file
import pandas
#reading the data from a csv file using read_csv() method
MyDataFrame = pandas.read_csv('StudentDataWithMissingValues.csv')

#Displaying individual elements in the shape tuple
Total_Rows_Columns = MyDataFrame.shape
print('\nThe Total number of instances are:',Total_Rows_Columns[0])
print('The Total number of attributes are:',Total_Rows_Columns[1])
print('Actual DataFrame\n',MyDataFrame)
#Drop the column if atleast one missing value is present
UpdatedMyDataFrame = MyDataFrame.dropna(axis=1)
Total_Rows_Columns = UpdatedMyDataFrame.shape
print('\nIn Updated Dataframe the Total number of instances are:',Total_Rows_Columns[0])
print('In Updated Dataframe the Total number of attributes are:',Total_Rows_Columns[1])
print('Updated DataFrame\n',UpdatedMyDataFrame)
```

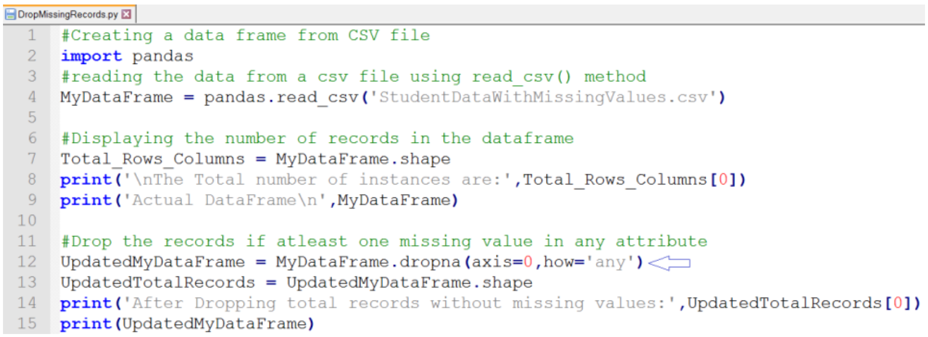
In our dataset, the columns DEPARTMENT and CGPA are having missing values while creating the dataset, so it is expected that both the columns will be dropped that results to only three columns i.e., REGISTRATION_NUMBER, NAME,GENDER

Output:



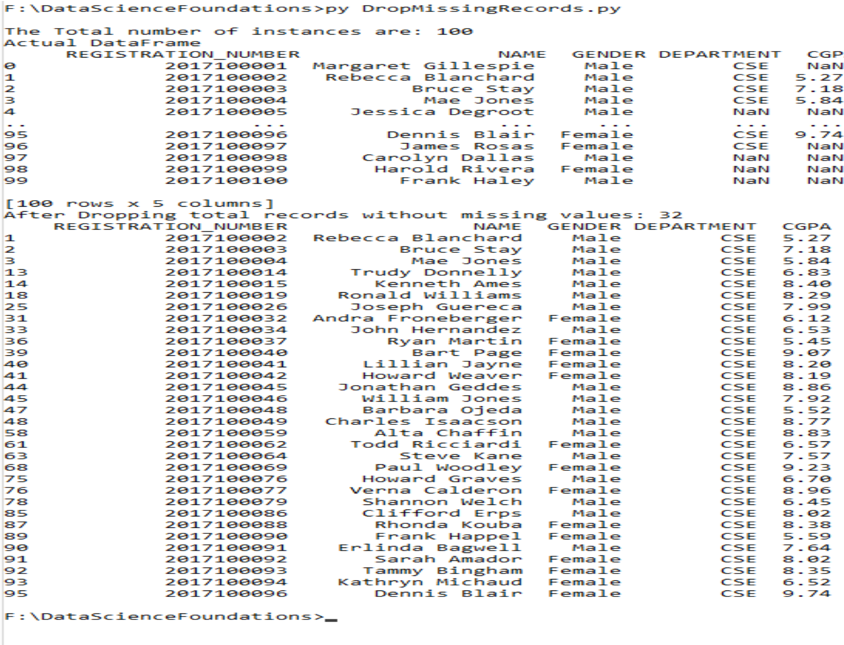
```
F:\DataScienceFoundations>py DropMissingRecords.py

The Total number of instances are: 100
The Total number of attributes are: 5
Actual DataFrame
     REGISTRATION_NUMBER                NAME  GENDER DEPARTMENT  CGPA
0             2017100001  Margaret Gillespie    Male        CSE   NaN
1             2017100002  Rebecca Blanchard    Male        CSE  5.27
2             2017100003          Bruce Stay    Male        CSE  7.18
3             2017100004           Mae Jones    Male        CSE  5.84
4             2017100005    Jessica Degroot    Male        NaN   NaN
..                   ...                 ...     ...        ...   ...
95            2017100096        Dennis Blair  Female        CSE  9.74
96            2017100097         James Rosas  Female        CSE   NaN
97            2017100098      Carolyn Dallas    Male        NaN   NaN
98            2017100099       Harold Rivera  Female        NaN   NaN
99            2017100100         Frank Haley    Male        NaN   NaN

[100 rows x 5 columns]

In Updated Dataframe the Total number of instances are: 100
In Updated Dataframe the Total number of attributes are: 3
Updated DataFrame
     REGISTRATION_NUMBER                NAME  GENDER
0             2017100001  Margaret Gillespie    Male
1             2017100002  Rebecca Blanchard    Male
2             2017100003          Bruce Stay    Male
3             2017100004           Mae Jones    Male
4             2017100005    Jessica Degroot    Male
..                   ...                 ...     ...
95            2017100096        Dennis Blair  Female
96            2017100097         James Rosas  Female
97            2017100098      Carolyn Dallas    Male
98            2017100099       Harold Rivera  Female
99            2017100100         Frank Haley    Male

[100 rows x 3 columns]

F:\DataScienceFoundations>
```

Dropping the record from a CSV file when at least one missing value is found



```python
#Creating a data frame from CSV file
import pandas
#reading the data from a csv file using read_csv() method
MyDataFrame = pandas.read_csv('StudentDataWithMissingValues.csv')

#Displaying the number of records in the dataframe
Total_Rows_Columns = MyDataFrame.shape
print('\nThe Total number of instances are:',Total_Rows_Columns[0])
print('Actual DataFrame\n',MyDataFrame)

#Drop the records if atleast one missing value in any attribute
UpdatedMyDataFrame = MyDataFrame.dropna(axis=0,how='any')
UpdatedTotalRecords = UpdatedMyDataFrame.shape
print('After Dropping total records without missing values:',UpdatedTotalRecords[0])
print(UpdatedMyDataFrame)
```

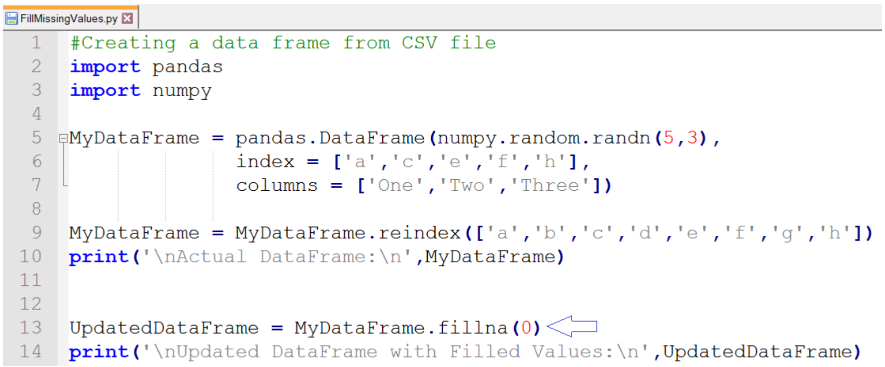
Output:



```
F:\DataScienceFoundations>py DropMissingRecords.py

The Total number of instances are: 100
Actual DataFrame
     REGISTRATION_NUMBER                NAME  GENDER DEPARTMENT  CGPA
0             2017100001  Margaret Gillespie    Male        CSE   NaN
1             2017100002   Rebecca Blanchard    Male        CSE  5.27
2             2017100003          Bruce Stay    Male        CSE  7.18
3             2017100004           Mae Jones    Male        CSE  5.84
4             2017100005     Jessica Degroot    Male        NaN   NaN
..                   ...                 ...     ...        ...   ...
95            2017100096        Dennis Blair  Female        CSE  9.74
96            2017100097         James Rosas  Female        CSE   NaN
97            2017100098      Carolyn Dallas    Male        NaN   NaN
98            2017100099       Harold Rivera  Female        NaN   NaN
99            2017100100         Frank Haley    Male        NaN   NaN

[100 rows x 5 columns]
After Dropping total records without missing values: 32
    REGISTRATION_NUMBER                NAME  GENDER DEPARTMENT  CGPA
1            2017100002   Rebecca Blanchard    Male        CSE  5.27
2            2017100003          Bruce Stay    Male        CSE  7.18
3            2017100004           Mae Jones    Male        CSE  5.84
13           2017100014     Trudy Donnelly    Male        CSE  6.83
14           2017100015       Kenneth Ames    Male        CSE  8.40
18           2017100019     Ronald Williams    Male        CSE  8.29
25           2017100026     Joseph Guereca    Male        CSE  7.99
31           2017100032   Andra Froneberger  Female        CSE  6.12
33           2017100034     John Hernandez    Male        CSE  6.53
36           2017100037         Ryan Martin  Female        CSE  5.45
39           2017100040           Bart Page  Female        CSE  9.07
40           2017100041       Lillian Jayne  Female        CSE  8.20
41           2017100042       Howard Weaver  Female        CSE  8.19
44           2017100045    Jonathan Geddes    Male        CSE  8.86
45           2017100046       William Jones    Male        CSE  7.92
47           2017100048      Barbara Ojeda    Male        CSE  5.52
48           2017100049   Charles Isaacson    Male        CSE  8.77
58           2017100059       Alta Chaffin    Male        CSE  8.83
61           2017100062     Todd Ricciardi  Female        CSE  6.57
63           2017100064          Steve Kane    Male        CSE  7.57
68           2017100069       Paul Woodley  Female        CSE  9.23
75           2017100076      Howard Graves    Male        CSE  6.70
76           2017100077     Verna Calderon  Female        CSE  8.96
78           2017100079      Shannon Welch    Male        CSE  6.45
85           2017100086      Clifford Erps    Male        CSE  8.02
87           2017100088      Rhonda Kouba  Female        CSE  8.38
89           2017100090       Frank Happel  Female        CSE  5.59
90           2017100091    Erlinda Bagwell    Male        CSE  7.64
91           2017100092       Sarah Amador  Female        CSE  8.02
92           2017100093      Tammy Bingham  Female        CSE  8.35
93           2017100094   Kathryn Michaud  Female        CSE  6.52
95           2017100096       Dennis Blair  Female        CSE  9.74

F:\DataScienceFoundations>_
```

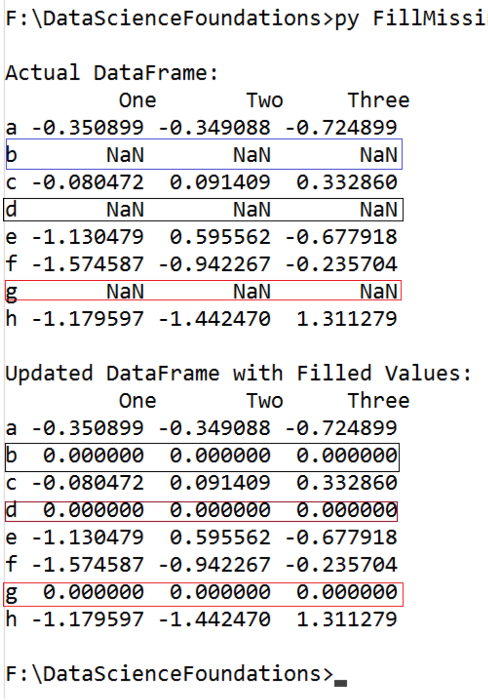
Filling the missing values with fillna(), replace() and interpolate()

The values that are missing can be filled with scalar values, random numbers. The values can be replaced with any specific value and the missing values can be filled using linear method using interpolate method.



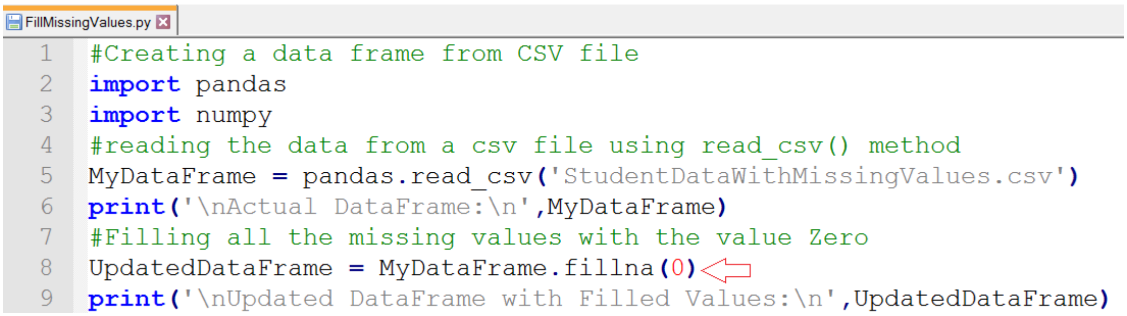
```python
#Creating a data frame from CSV file
import pandas
import numpy

MyDataFrame = pandas.DataFrame(numpy.random.randn(5,3),
                  index = ['a','c','e','f','h'],
                  columns = ['One','Two','Three'])

MyDataFrame = MyDataFrame.reindex(['a','b','c','d','e','f','g','h'])
print('\nActual DataFrame:\n',MyDataFrame)


UpdatedDataFrame = MyDataFrame.fillna(0)
print('\nUpdated DataFrame with Filled Values:\n',UpdatedDataFrame)
```

Output:



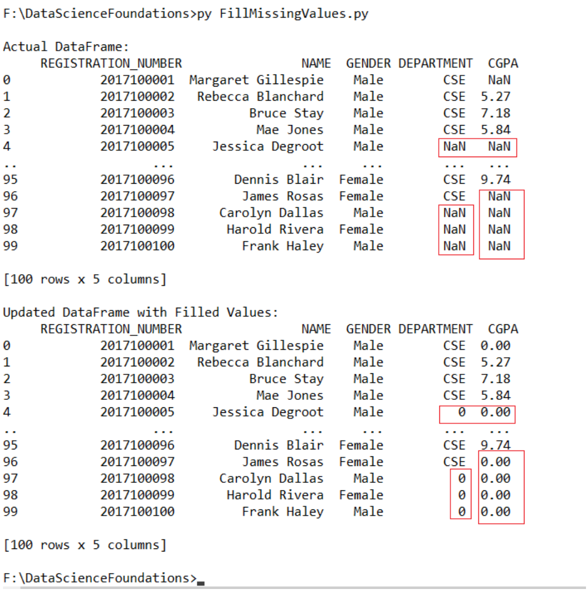
```
F:\DataScienceFoundations>py FillMissingValues.py

Actual DataFrame:
         One       Two     Three
a -0.350899 -0.349088 -0.724899
b       NaN       NaN       NaN
c -0.080472  0.091409  0.332860
d       NaN       NaN       NaN
e -1.130479  0.595562 -0.677918
f -1.574587 -0.942267 -0.235704
g       NaN       NaN       NaN
h -1.179597 -1.442470  1.311279

Updated DataFrame with Filled Values:
         One       Two     Three
a -0.350899 -0.349088 -0.724899
b  0.000000  0.000000  0.000000
c -0.080472  0.091409  0.332860
d  0.000000  0.000000  0.000000
e -1.130479  0.595562 -0.677918
f -1.574587 -0.942267 -0.235704
g  0.000000  0.000000  0.000000
h -1.179597 -1.442470  1.311279

F:\DataScienceFoundations>_
```

Another example with filling the missing values in the dataset



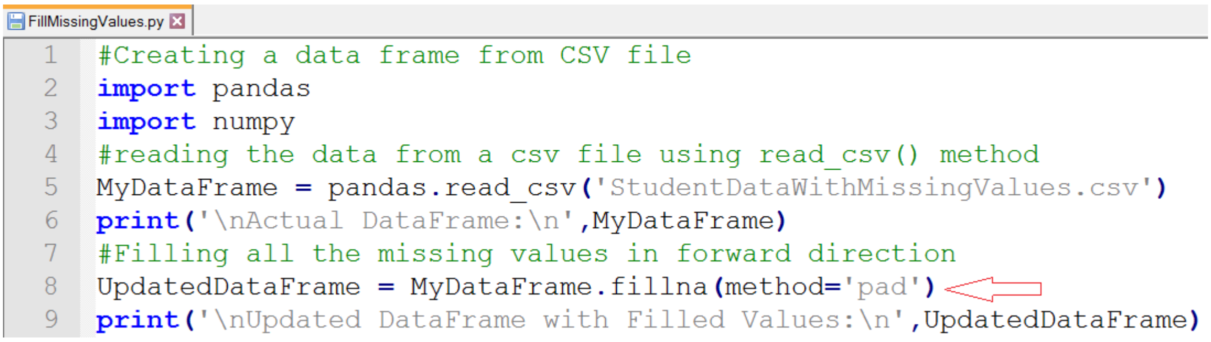
```python
#Creating a data frame from CSV file
import pandas
import numpy
#reading the data from a csv file using read_csv() method
MyDataFrame = pandas.read_csv('StudentDataWithMissingValues.csv')
print('\nActual DataFrame:\n',MyDataFrame)
#Filling all the missing values with the value Zero
UpdatedDataFrame = MyDataFrame.fillna(0)
print('\nUpdated DataFrame with Filled Values:\n',UpdatedDataFrame)
```

Output:



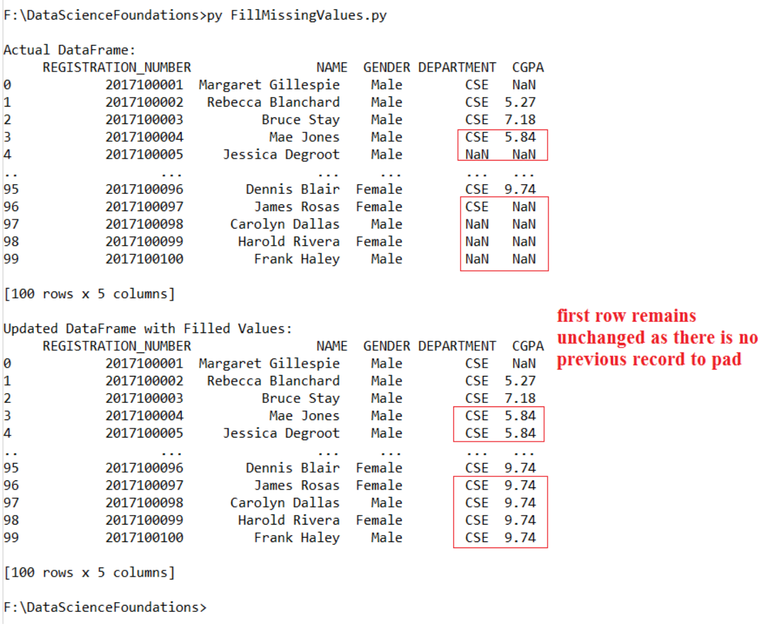
```
F:\DataScienceFoundations>py FillMissingValues.py

Actual DataFrame:
    REGISTRATION_NUMBER                NAME  GENDER DEPARTMENT  CGPA
0            2017100001  Margaret Gillespie    Male        CSE   NaN
1            2017100002   Rebecca Blanchard    Male        CSE  5.27
2            2017100003          Bruce Stay    Male        CSE  7.18
3            2017100004           Mae Jones    Male        CSE  5.84
4            2017100005     Jessica Degroot    Male        NaN   NaN
..                  ...                 ...     ...        ...   ...
95           2017100096        Dennis Blair  Female        CSE  9.74
96           2017100097         James Rosas  Female        CSE   NaN
97           2017100098      Carolyn Dallas    Male        NaN   NaN
98           2017100099       Harold Rivera  Female        NaN   NaN
99           2017100100         Frank Haley    Male        NaN   NaN

[100 rows x 5 columns]

Updated DataFrame with Filled Values:
    REGISTRATION_NUMBER                NAME  GENDER DEPARTMENT  CGPA
0            2017100001  Margaret Gillespie    Male        CSE  0.00
1            2017100002   Rebecca Blanchard    Male        CSE  5.27
2            2017100003          Bruce Stay    Male        CSE  7.18
3            2017100004           Mae Jones    Male        CSE  5.84
4            2017100005     Jessica Degroot    Male          0  0.00
..                  ...                 ...     ...        ...   ...
95           2017100096        Dennis Blair  Female        CSE  9.74
96           2017100097         James Rosas  Female        CSE  0.00
97           2017100098      Carolyn Dallas    Male          0  0.00
98           2017100099       Harold Rivera  Female          0  0.00
99           2017100100         Frank Haley    Male          0  0.00

[100 rows x 5 columns]

F:\DataScienceFoundations>_
```

Filling the values in forward direction using method, fillna(method = ‘pad’)



```python
#Creating a data frame from CSV file
import pandas
import numpy
#reading the data from a csv file using read_csv() method
MyDataFrame = pandas.read_csv('StudentDataWithMissingValues.csv')
print('\nActual DataFrame:\n',MyDataFrame)
#Filling all the missing values in forward direction
UpdatedDataFrame = MyDataFrame.fillna(method='pad')
print('\nUpdated DataFrame with Filled Values:\n',UpdatedDataFrame)
```

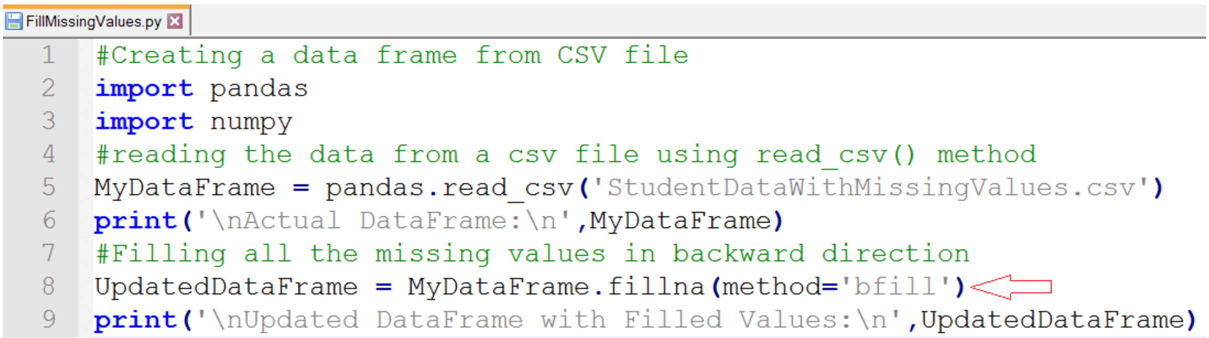
Output:



```
F:\DataScienceFoundations>py FillMissingValues.py

Actual DataFrame:
    REGISTRATION_NUMBER                NAME  GENDER DEPARTMENT  CGPA
0            2017100001  Margaret Gillespie    Male        CSE   NaN
1            2017100002   Rebecca Blanchard    Male        CSE  5.27
2            2017100003          Bruce Stay    Male        CSE  7.18
3            2017100004           Mae Jones    Male        CSE  5.84
4            2017100005     Jessica Degroot    Male        NaN   NaN
..                  ...                 ...     ...        ...   ...
95           2017100096        Dennis Blair  Female        CSE  9.74
96           2017100097         James Rosas  Female        CSE   NaN
97           2017100098      Carolyn Dallas    Male        NaN   NaN
98           2017100099       Harold Rivera  Female        NaN   NaN
99           2017100100         Frank Haley    Male        NaN   NaN

[100 rows x 5 columns]

Updated DataFrame with Filled Values:
    REGISTRATION_NUMBER                NAME  GENDER DEPARTMENT  CGPA
0            2017100001  Margaret Gillespie    Male        CSE   NaN
1            2017100002   Rebecca Blanchard    Male        CSE  5.27
2            2017100003          Bruce Stay    Male        CSE  7.18
3            2017100004           Mae Jones    Male        CSE  5.84
4            2017100005     Jessica Degroot    Male        CSE  5.84
..                  ...                 ...     ...        ...   ...
95           2017100096        Dennis Blair  Female        CSE  9.74
96           2017100097         James Rosas  Female        CSE  9.74
97           2017100098      Carolyn Dallas    Male        CSE  9.74
98           2017100099       Harold Rivera  Female        CSE  9.74
99           2017100100         Frank Haley    Male        CSE  9.74

[100 rows x 5 columns]

F:\DataScienceFoundations>
```

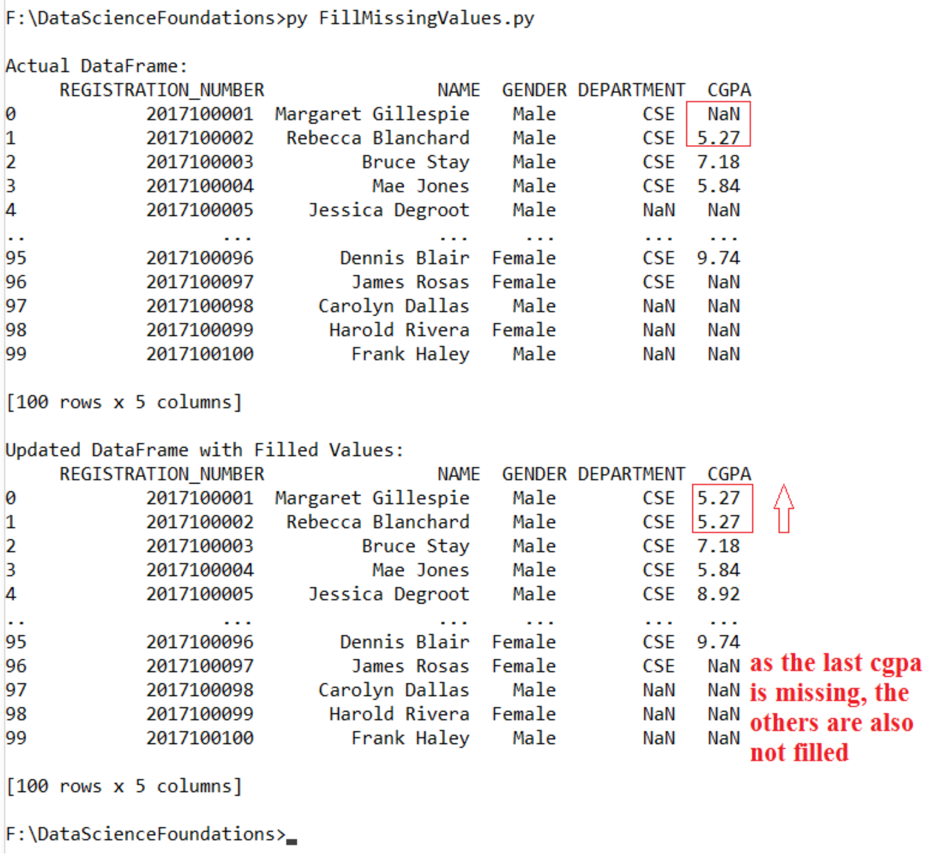
first row remains unchanged as there is no previous record to pad

Filling the missing values in backward direction using fillna(method = ‘bfill’)



```python
#Creating a data frame from CSV file
import pandas
import numpy
#reading the data from a csv file using read_csv() method
MyDataFrame = pandas.read_csv('StudentDataWithMissingValues.csv')
print('\nActual DataFrame:\n',MyDataFrame)
#Filling all the missing values in backward direction
UpdatedDataFrame = MyDataFrame.fillna(method='bfill')
print('\nUpdated DataFrame with Filled Values:\n',UpdatedDataFrame)
```

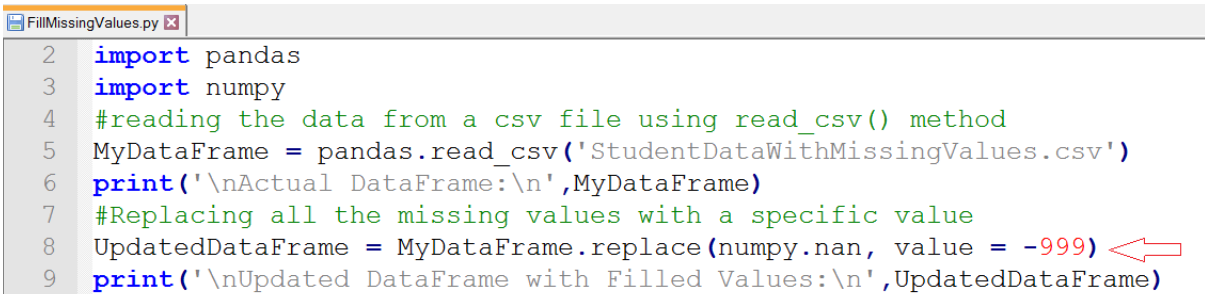
Output:



```
F:\DataScienceFoundations>py FillMissingValues.py

Actual DataFrame:
     REGISTRATION_NUMBER               NAME  GENDER DEPARTMENT  CGPA
0             2017100001  Margaret Gillespie    Male        CSE   NaN
1             2017100002   Rebecca Blanchard    Male        CSE  5.27
2             2017100003          Bruce Stay    Male        CSE  7.18
3             2017100004           Mae Jones    Male        CSE  5.84
4             2017100005     Jessica Degroot    Male        NaN   NaN
..                   ...                 ...     ...        ...   ...
95            2017100096        Dennis Blair  Female        CSE  9.74
96            2017100097         James Rosas  Female        CSE   NaN
97            2017100098      Carolyn Dallas    Male        NaN   NaN
98            2017100099       Harold Rivera  Female        NaN   NaN
99            2017100100         Frank Haley    Male        NaN   NaN

[100 rows x 5 columns]

Updated DataFrame with Filled Values:
     REGISTRATION_NUMBER               NAME  GENDER DEPARTMENT  CGPA
0             2017100001  Margaret Gillespie    Male        CSE  5.27
1             2017100002   Rebecca Blanchard    Male        CSE  5.27
2             2017100003          Bruce Stay    Male        CSE  7.18
3             2017100004           Mae Jones    Male        CSE  5.84
4             2017100005     Jessica Degroot    Male        CSE  8.92
..                   ...                 ...     ...        ...   ...
95            2017100096        Dennis Blair  Female        CSE  9.74
96            2017100097         James Rosas  Female        CSE   NaN
97            2017100098      Carolyn Dallas    Male        NaN   NaN
98            2017100099       Harold Rivera  Female        NaN   NaN
99            2017100100         Frank Haley    Male        NaN   NaN

[100 rows x 5 columns]

F:\DataScienceFoundations>
```

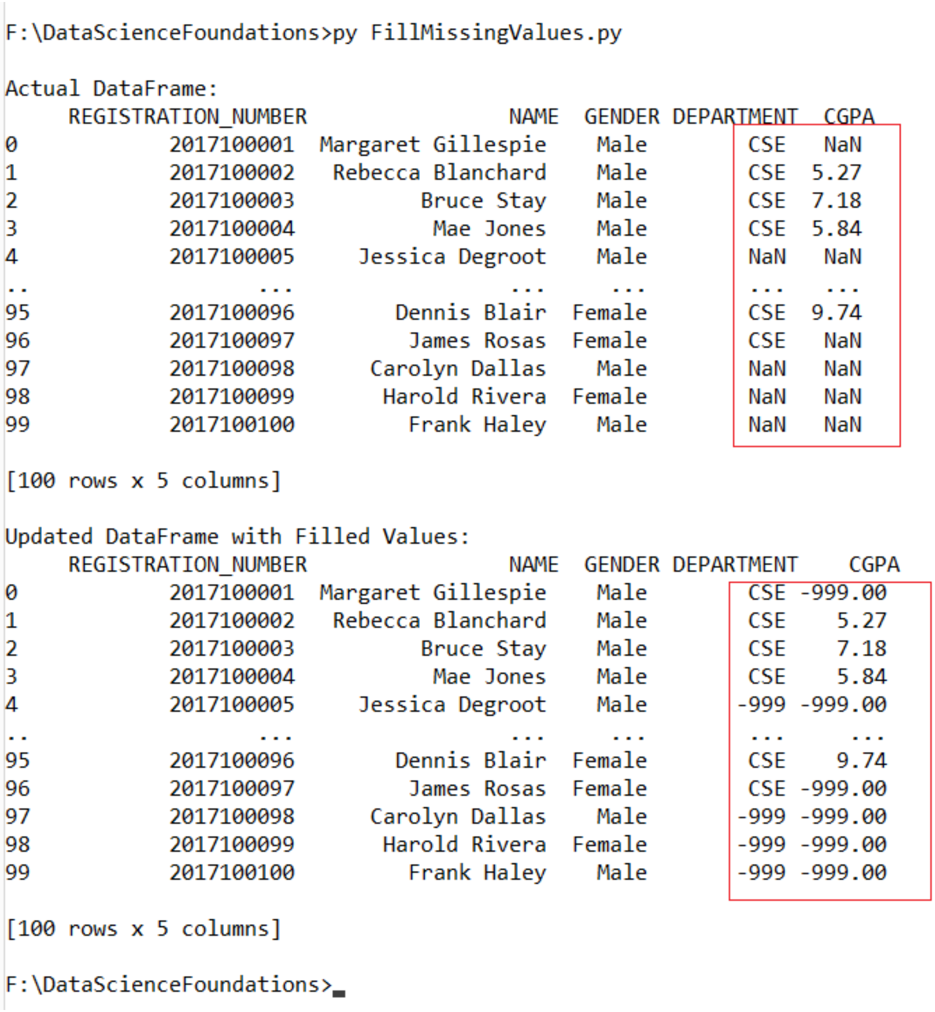
as the last cgpa is missing, the others are also not filled

Replacing the missing values with replace()



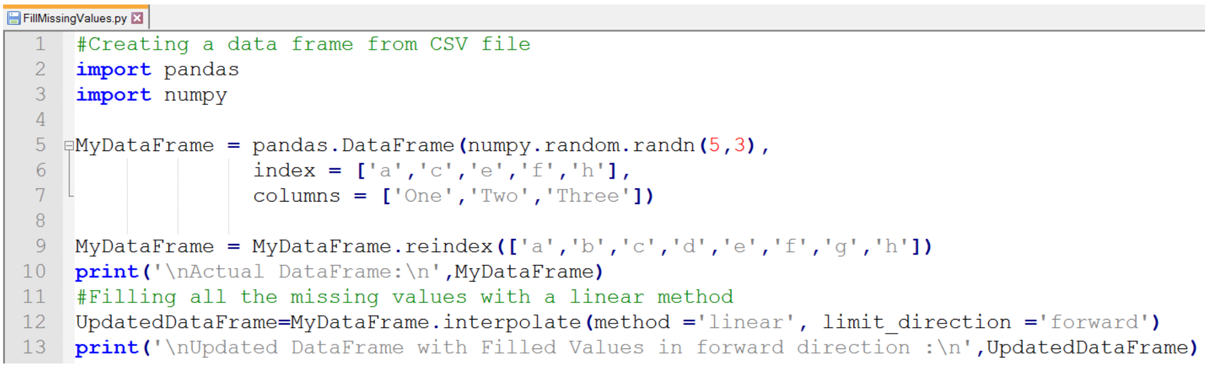
```python
import pandas
import numpy
#reading the data from a csv file using read_csv() method
MyDataFrame = pandas.read_csv('StudentDataWithMissingValues.csv')
print('\nActual DataFrame:\n',MyDataFrame)
#Replacing all the missing values with a specific value
UpdatedDataFrame = MyDataFrame.replace(numpy.nan, value = -999)
print('\nUpdated DataFrame with Filled Values:\n',UpdatedDataFrame)
```

Output:



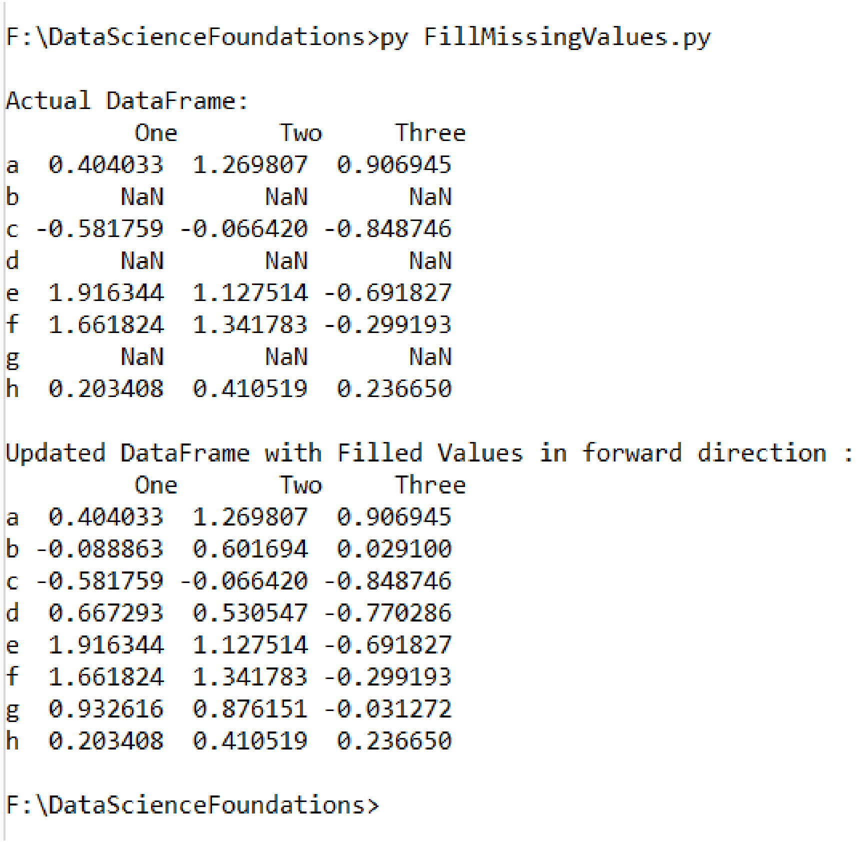
```
F:\DataScienceFoundations>py FillMissingValues.py

Actual DataFrame:
    REGISTRATION_NUMBER                NAME  GENDER DEPARTMENT  CGPA
0            2017100001  Margaret Gillespie    Male        CSE   NaN
1            2017100002   Rebecca Blanchard    Male        CSE  5.27
2            2017100003          Bruce Stay    Male        CSE  7.18
3            2017100004           Mae Jones    Male        CSE  5.84
4            2017100005     Jessica Degroot    Male        NaN   NaN
..                  ...                 ...     ...        ...   ...
95           2017100096        Dennis Blair  Female        CSE  9.74
96           2017100097         James Rosas  Female        CSE   NaN
97           2017100098      Carolyn Dallas    Male        NaN   NaN
98           2017100099       Harold Rivera  Female        NaN   NaN
99           2017100100         Frank Haley    Male        NaN   NaN

[100 rows x 5 columns]

Updated DataFrame with Filled Values:
    REGISTRATION_NUMBER                NAME  GENDER DEPARTMENT    CGPA
0            2017100001  Margaret Gillespie    Male        CSE -999.00
1            2017100002   Rebecca Blanchard    Male        CSE    5.27
2            2017100003          Bruce Stay    Male        CSE    7.18
3            2017100004           Mae Jones    Male        CSE    5.84
4            2017100005     Jessica Degroot    Male       -999 -999.00
..                  ...                 ...     ...        ...     ...
95           2017100096        Dennis Blair  Female        CSE    9.74
96           2017100097         James Rosas  Female        CSE -999.00
97           2017100098      Carolyn Dallas    Male       -999 -999.00
98           2017100099       Harold Rivera  Female       -999 -999.00
99           2017100100         Frank Haley    Male       -999 -999.00

[100 rows x 5 columns]

F:\DataScienceFoundations>
```

Using interpolate() function to fill the missing values using linear method



```python
#Creating a data frame from CSV file
import pandas
import numpy

MyDataFrame = pandas.DataFrame(numpy.random.randn(5,3),
              index = ['a','c','e','f','h'],
              columns = ['One','Two','Three'])

MyDataFrame = MyDataFrame.reindex(['a','b','c','d','e','f','g','h'])
print('\nActual DataFrame:\n',MyDataFrame)
#Filling all the missing values with a linear method
UpdatedDataFrame=MyDataFrame.interpolate(method ='linear', limit_direction ='forward')
print('\nUpdated DataFrame with Filled Values in forward direction :\n',UpdatedDataFrame)
```

Output:



```
F:\DataScienceFoundations>py FillMissingValues.py

Actual DataFrame:
        One       Two     Three
a  0.404033  1.269807  0.906945
b       NaN       NaN       NaN
c -0.581759 -0.066420 -0.848746
d       NaN       NaN       NaN
e  1.916344  1.127514 -0.691827
f  1.661824  1.341783 -0.299193
g       NaN       NaN       NaN
h  0.203408  0.410519  0.236650

Updated DataFrame with Filled Values in forward direction :
        One       Two     Three
a  0.404033  1.269807  0.906945
b -0.088863  0.601694  0.029100
c -0.581759 -0.066420 -0.848746
d  0.667293  0.530547 -0.770286
e  1.916344  1.127514 -0.691827
f  1.661824  1.341783 -0.299193
g  0.932616  0.876151 -0.031272
h  0.203408  0.410519  0.236650

F:\DataScienceFoundations>
```