

SIT719 Task 2.1 Sanket Thakur

December 5, 2020

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

```
[4]: iris_data = pd.read_csv("https://archive.ics.uci.edu/ml/
↳machine-learning-databases/iris/iris.data", header=None)
```

```
[5]: iris_data
```

```
[5]:      0      1      2      3      4
0    5.1  3.5  1.4  0.2    Iris-setosa
1    4.9  3.0  1.4  0.2    Iris-setosa
2    4.7  3.2  1.3  0.2    Iris-setosa
3    4.6  3.1  1.5  0.2    Iris-setosa
4    5.0  3.6  1.4  0.2    Iris-setosa
..    ...    ...    ...    ...    ...
145   6.7  3.0  5.2  2.3  Iris-virginica
146   6.3  2.5  5.0  1.9  Iris-virginica
147   6.5  3.0  5.2  2.0  Iris-virginica
148   6.2  3.4  5.4  2.3  Iris-virginica
149   5.9  3.0  5.1  1.8  Iris-virginica
```

[150 rows x 5 columns]

```
[8]: iris_data.
↳columns=['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class']
```

```
[9]: iris_data
```

```
[9]:      sepal_length  sepal_width  petal_length  petal_width      class
0              5.1           3.5           1.4           0.2  Iris-setosa
1              4.9           3.0           1.4           0.2  Iris-setosa
2              4.7           3.2           1.3           0.2  Iris-setosa
3              4.6           3.1           1.5           0.2  Iris-setosa
4              5.0           3.6           1.4           0.2  Iris-setosa
..              ...           ...           ...           ...    ...
145             6.7           3.0           5.2           2.3  Iris-virginica
146             6.3           2.5           5.0           1.9  Iris-virginica
```

147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

[150 rows x 5 columns]

```
[10]: iris_data.isnull().tail()
```

```
[10]:      sepal_length  sepal_width  petal_length  petal_width  class
145          False          False          False          False  False
146          False          False          False          False  False
147          False          False          False          False  False
148          False          False          False          False  False
149          False          False          False          False  False
```

```
[11]: iris_data.loc[9:29, 'petal_length'] = pd.np.NaN
```

<ipython-input-11-9f54047bc343>:1: FutureWarning: The pandas.np module is deprecated and will be removed from pandas in a future version. Import numpy directly instead

```
iris_data.loc[9:29, 'petal_length'] = pd.np.NaN
```

```
[13]: iris_data.isnull()
```

```
[13]:      sepal_length  sepal_width  petal_length  petal_width  class
0          False          False          False          False  False
1          False          False          False          False  False
2          False          False          False          False  False
3          False          False          False          False  False
4          False          False          False          False  False
..          ...          ...          ...          ...          ...
145         False          False          False          False  False
146         False          False          False          False  False
147         False          False          False          False  False
148         False          False          False          False  False
149         False          False          False          False  False
```

[150 rows x 5 columns]

```
[14]: pd.isnull(iris_data).sum().sum()
```

```
[14]: 21
```

```
[15]: iris_data.fillna(10.0)
```

```
[15]:      sepal_length  sepal_width  petal_length  petal_width  class
0          5.1          3.5          1.4          0.2  Iris-setosa
```

1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
..
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

[150 rows x 5 columns]

[]:

SIT719 Security and Privacy Issues in Analytics

Pass Task 2.1: Basic scripting with python

Section 1

Instructions: In this task, you will be asked to perform some basic python operations using pandas and numpy libraries. Please write the code, execute and take a screenshot of the results of the completed outputs.

Step 1. Import the pandas and numpy libraries

Answer1: (This one has been done for you)

```
In [140]: import pandas as pd
...: import numpy as np
```

Step 2. Import the popular 'iris' dataset from the below address. And then check the header of the dataset.

<https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>

Answer2: (This one has also been done for you)

```
In [141]: url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'

In [142]: iris = pd.read_csv(url)

In [143]: iris.head()
Out[143]:
5.1 3.5 1.4 0.2 Iris-setosa
0 4.9 3.0 1.4 0.2 Iris-setosa
1 4.7 3.2 1.3 0.2 Iris-setosa
2 4.6 3.1 1.5 0.2 Iris-setosa
3 5.0 3.6 1.4 0.2 Iris-setosa
4 5.4 3.9 1.7 0.4 Iris-setosa
```

Step 3. You can see that the column headers are missing in the above case. Therefore this step is related to the creation of column heads for the dataset. [Write code to create 5 column heads. Next write a code to display or show the headers.](#)

1. sepal_length
2. sepal_width
3. petal_length
4. petal_width
5. class

Answer3: (write your code)

```
iris_data.columns=['sepal_length','sepal_width','petal_length','petal_width','class']
```

Step 4. [Write a code to check if there are any missing values in the dataframe?](#)

Answer4: (write your code)

```
iris_data.isnull().tail()
```

Hints: there is no missing values but check it thorough the code

Step 5. Write a code to set the values of the rows 10 to 29 of the column 'petal length' to NaN.

Answer5: (write your code)

```
iris_data.loc[9:29, 'petal_length'] = pd.np.NaN
```

Step 6. Now again, check if there is any missing values (NaN) in the dataframe? Count, how many missing values.

Answer6: (write your code)

```
iris_data.isnull()  
pd.isnull(iris_data).sum().sum()
```

Hints: this time you will have missing values.

Step 7. [Substitute the NaN values to 10.0](#)

Answer7: (write your code)

```
iris_data.fillna(10.0)
```

Section 2

Numpy is an open source library written in C++, with Python being the basic package for scientific computing.

It includes a range of methods for most machine learning tasks, algorithms

Pandas is an open source data analysis and manipulation tool that is quick, efficient, versatile and easy to use, built on top of the programming language of Python

Matplotlib is a Python programming language plotting library and its NumPy numeric al mathematics extension

We can use slicing and indexing in python for creating a subset of the given data frame. It can be also used to operate on a particular section in the given dataset.

Visualising data gives us a very intuitive insight of the data we want to work with. Matplotlib is one of the best visualization tools available in Python.

```
In [4]: from matplotlib import pyplot as plt
```

```
In [9]: x = [1,5,6]
y = [2,3,4]
plt.plot(x,y)
plt.plot(y,x)
plt.title('test plot')
plt.xlabel('x')
plt.ylabel('y')
plt.show()
```

