SIT719 Security and Privacy Issues in Analytics

Pass Task 2.1: Basic scripting with python

Overview

Python is an amazingly versatile programming language and extremely popular among the data science people. This powerful tool will give you access to a wide variety of data science libraries which will help you to develop your script easily. By the end of week 2, you will be familiar with basic python scripting. Please see the weekly resources for some basic operations.

If you are new to python scripting, you might follow the below references:

- Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython by Wes McKinney, O'Reilly Media, Inc.

Because of the evolving nature of the open-source tools like Python and its libraries, it is always wise to look for the updated learning material from the python library website tutorials, user guides and manuals. For example, the user guide of the pandas data frame can be obtained from the below link:
https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html
Similarly, numpy can be learned based on the material presented in the following links:
https://docs.scipy.org/doc/numpy/user/basics.html
https://docs.scipy.org/doc/numpy/user/quickstart.html

This is a Pass task, so you **MUST** complete the task and submit the evidence of your work to Ontrack.

Submit the following files to Ontrack:

- A screenshot of the output you obtained by executing the python program (in Section 1)
- Some reflections on what you got out of this experience of learning fundamental concepts of python scripting (see Section 2)

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
```

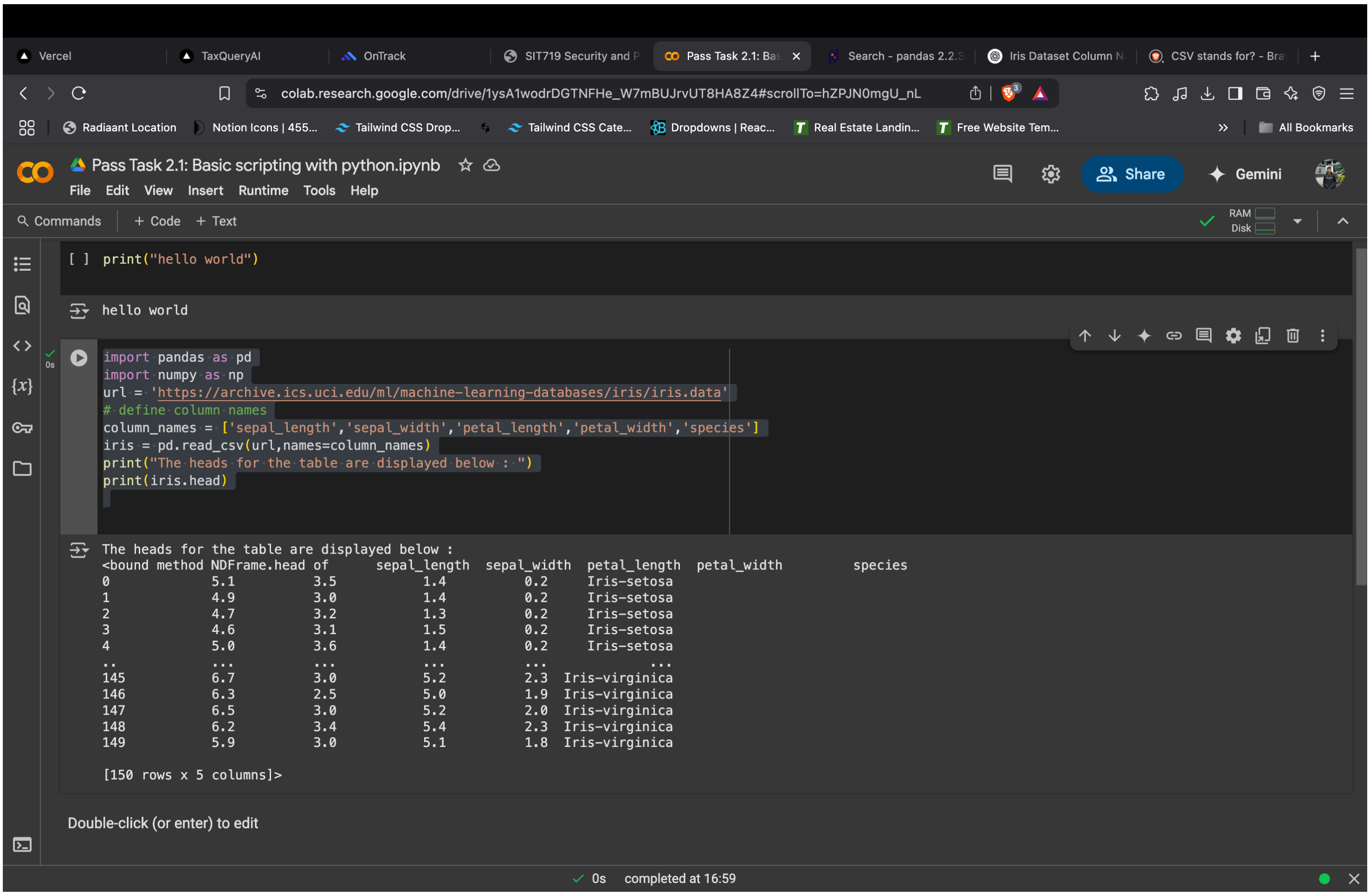
Step 3. You can see that the column headers are missing in the above case. Therefore this step is related to the creation of column heads for the dataset. Write code to create 5 column heads. Next write a code to display or show the headers.

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

Answer3: (write your code)

```
import pandas as pd
import numpy as np
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'
# define column names
column_names = ['sepal_length','sepal_width','petal_length','petal_width','species']
iris = pd.read_csv(url,names=column_names)
print("The heads for the table are displayed below : ")
print(iris.head)
```

Evidence



```
print("hello world")
```

```
hello world
```

```
import pandas as pd
import numpy as np
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'
# define column names
column_names = ['sepal_length','sepal_width','petal_length','petal_width','species']
iris = pd.read_csv(url,names=column_names)
print("The heads for the table are displayed below : ")
print(iris.head)
```

```
The heads for the table are displayed below :
<bound method NDFrame.head of      sepal_length  sepal_width  petal_length  petal_width         species
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]>
```

Double-click (or enter) to edit

✓ 0s completed at 16:59

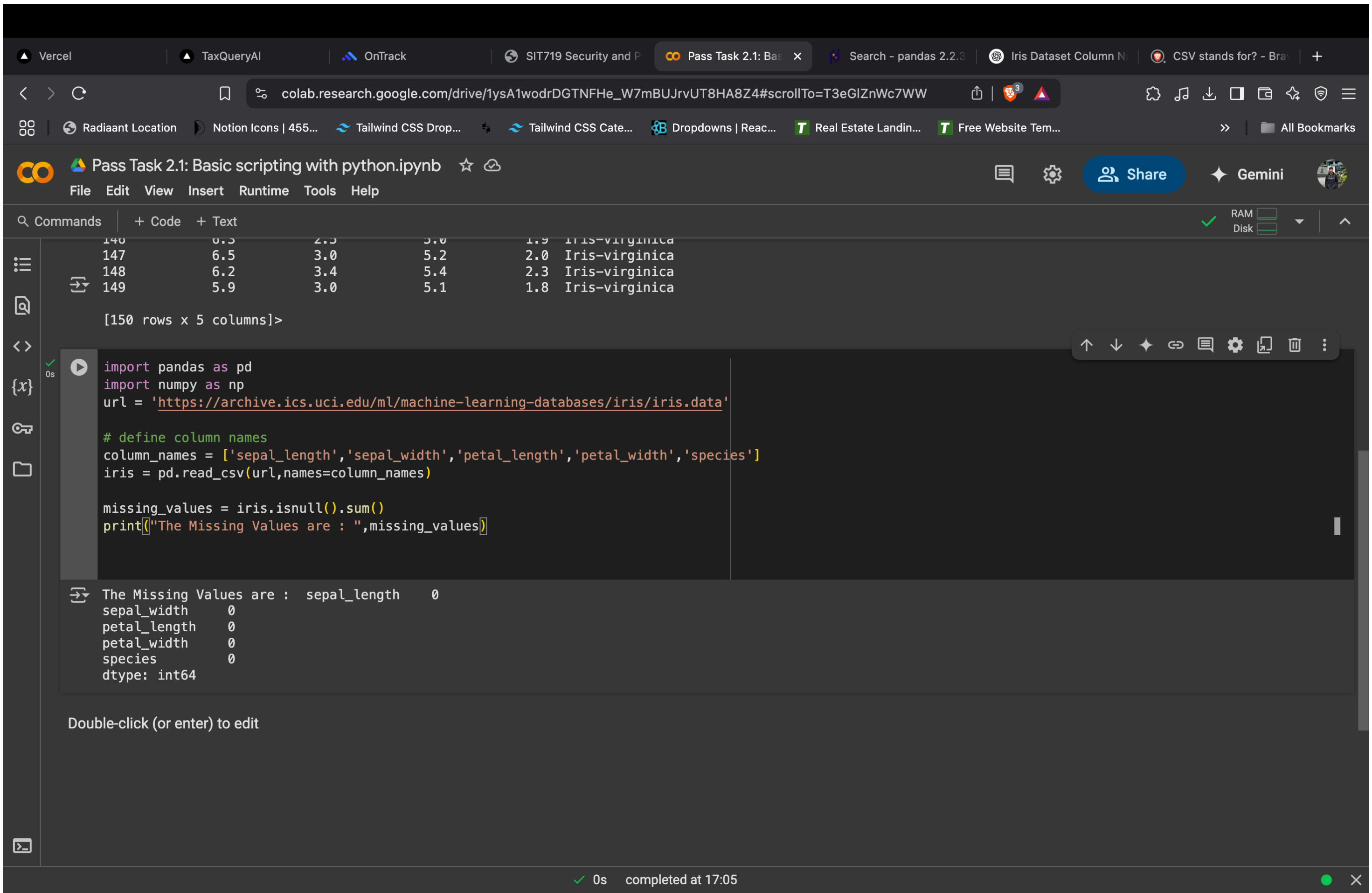
Step 4. Write a code to check if there are any missing values in the dataframe?

Answer4: (write your code)

```
import pandas as pd
import numpy as np
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'

# define column names
column_names = ['sepal_length','sepal_width','petal_length','petal_width','species']
iris = pd.read_csv(url,names=column_names)

missing_values = iris.isnull().sum()
print("The Missing Values are : ",missing_values)
```



```
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]>
```

```
import pandas as pd
import numpy as np
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'

# define column names
column_names = ['sepal_length','sepal_width','petal_length','petal_width','species']
iris = pd.read_csv(url,names=column_names)

missing_values = iris.isnull().sum()
print("The Missing Values are : ",missing_values)
```

```
The Missing Values are :  sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64
```

Double-click (or enter) to edit

Evidence :

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)

```
import pandas as pd
import numpy as np
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'

# define column names
column_names = ['sepal_length','sepal_width','petal_length','petal_width','species']
iris = pd.read_csv(url,names=column_names)
iris.loc[10:29,'petal_length'] = np.nan;
print(iris.head(30))
```

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

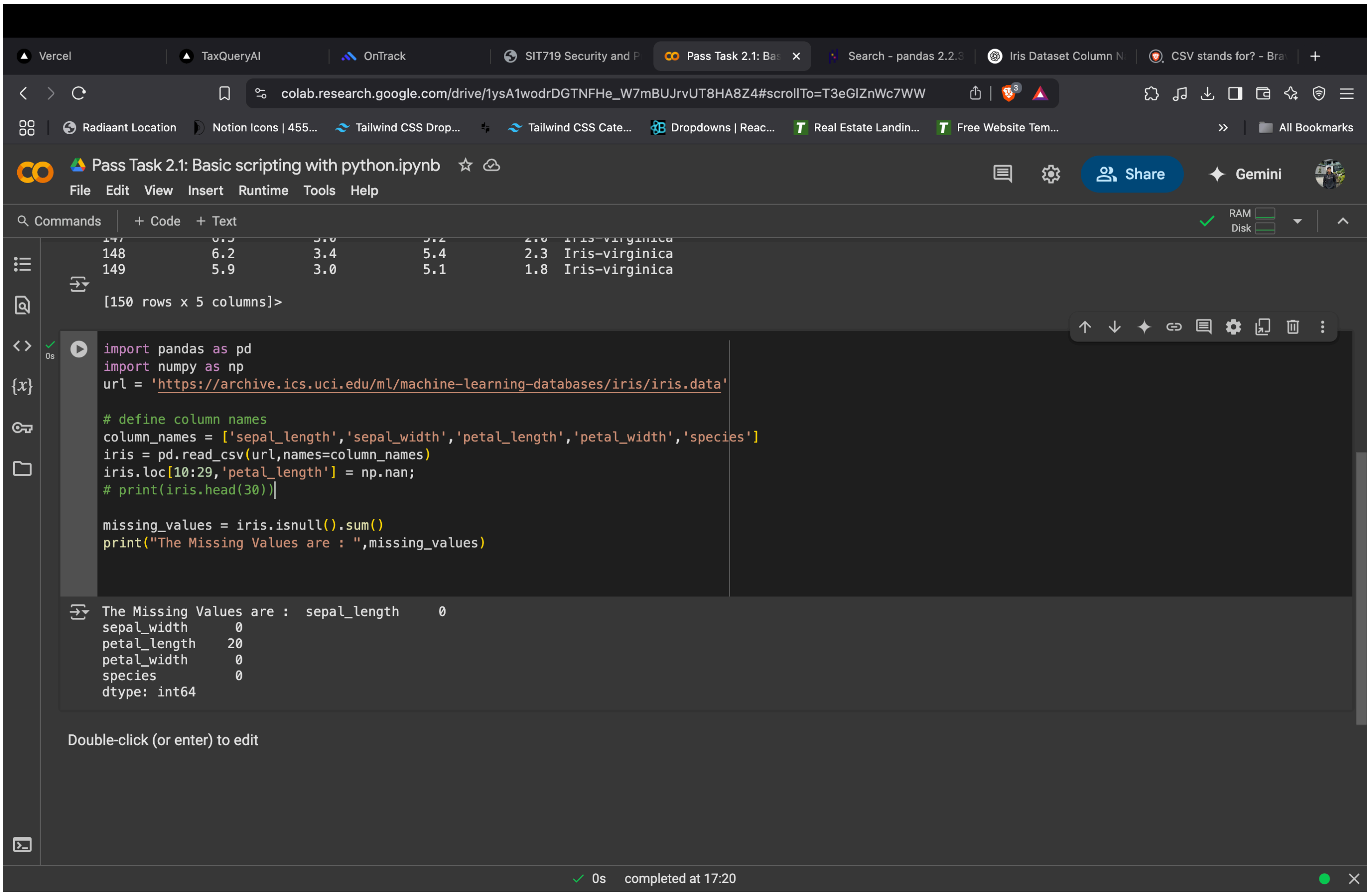
Answer6: (write your code)

```
import pandas as pd
import numpy as np
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'

# define column names
column_names = ['sepal_length','sepal_width','petal_length','petal_width','species']
iris = pd.read_csv(url,names=column_names)
iris.loc[10:29,'petal_length'] = np.nan;
# print(iris.head(30))

missing_values = iris.isnull().sum()
print("The Missing Values are : ",missing_values)
```

Evidence for answer 5 & 6 :



```
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]>
```

```
import pandas as pd
import numpy as np
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'

# define column names
column_names = ['sepal_length','sepal_width','petal_length','petal_width','species']
iris = pd.read_csv(url,names=column_names)
iris.loc[10:29,'petal_length'] = np.nan;
# print(iris.head(30))

missing_values = iris.isnull().sum()
print("The Missing Values are : ",missing_values)
```

```
The Missing Values are :  sepal_length     0
sepal_width      0
petal_length    20
petal_width      0
species          0
dtype: int64
```

Hints: this time you will have missing values.

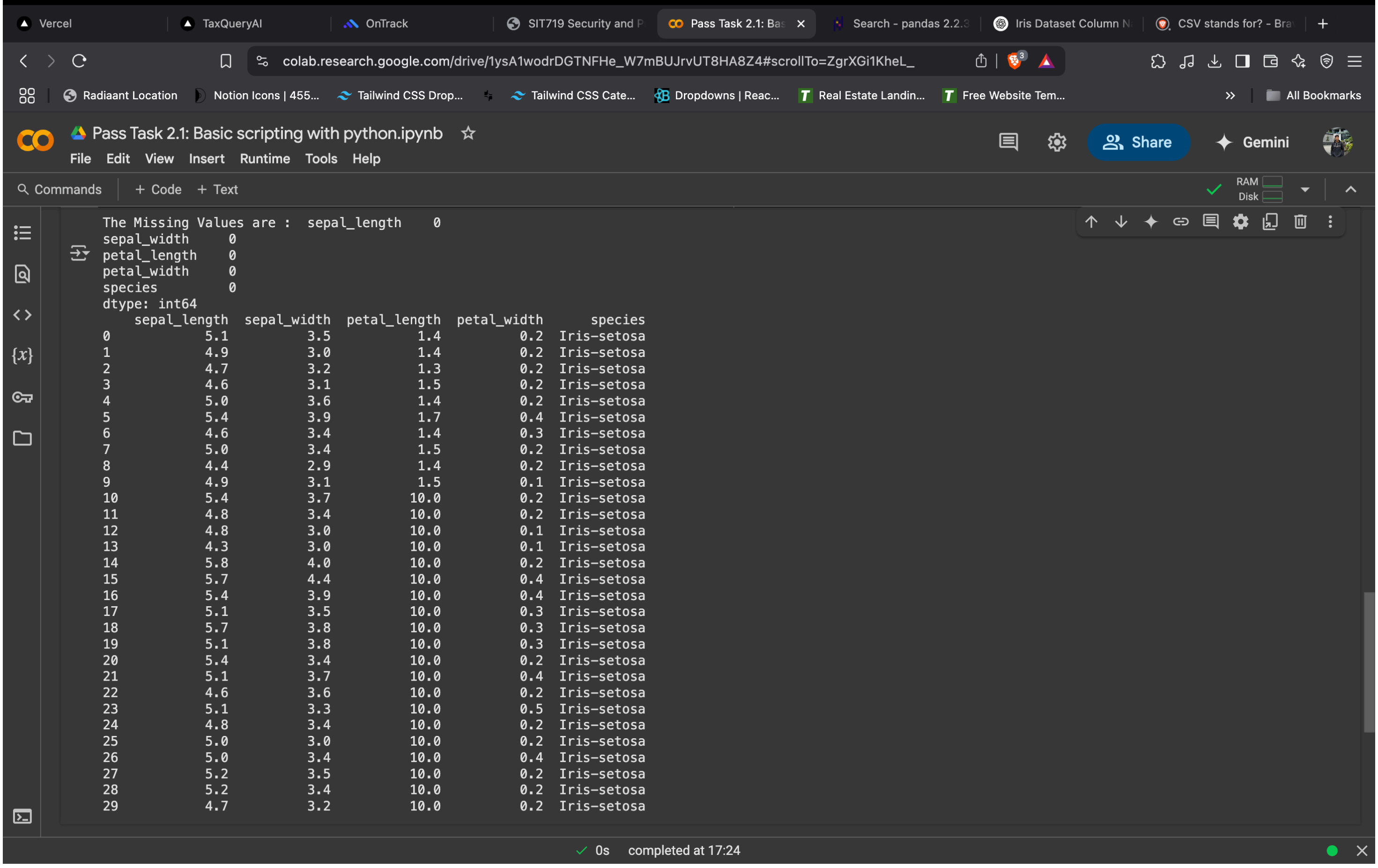
Step 7. Substitute the NaN values to 10.0

Answer7: (write your code)

```
import pandas as pd
import numpy as np
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'

# define column names
column_names = ['sepal_length','sepal_width','petal_length','petal_width','species']
iris = pd.read_csv(url,names=column_names)
iris.loc[10:29,'petal_length'] = np.nan;
iris.fillna(10.0, inplace=True)
missing_values = iris.isnull().sum()
print("The Missing Values are : ",missing_values)
print(iris.head(30))
```

Evidence :



```
The Missing Values are :  sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64
    sepal_length  sepal_width  petal_length  petal_width      species
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
5            5.4          3.9           1.7          0.4  Iris-setosa
6            4.6          3.4           1.4          0.3  Iris-setosa
7            5.0          3.4           1.5          0.2  Iris-setosa
8            4.4          2.9           1.4          0.2  Iris-setosa
9            4.9          3.1           1.5          0.1  Iris-setosa
10           5.4          3.7          10.0          0.2  Iris-setosa
11           4.8          3.4          10.0          0.2  Iris-setosa
12           4.8          3.0          10.0          0.1  Iris-setosa
13           4.3          3.0          10.0          0.1  Iris-setosa
14           5.8          4.0          10.0          0.2  Iris-setosa
15           5.7          4.4          10.0          0.4  Iris-setosa
16           5.4          3.9          10.0          0.4  Iris-setosa
17           5.1          3.5          10.0          0.3  Iris-setosa
18           5.7          3.8          10.0          0.3  Iris-setosa
19           5.1          3.8          10.0          0.3  Iris-setosa
20           5.4          3.4          10.0          0.2  Iris-setosa
21           5.1          3.7          10.0          0.4  Iris-setosa
22           4.6          3.6          10.0          0.2  Iris-setosa
23           5.1          3.3          10.0          0.5  Iris-setosa
24           4.8          3.4          10.0          0.2  Iris-setosa
25           5.0          3.0          10.0          0.2  Iris-setosa
26           5.0          3.4          10.0          0.4  Iris-setosa
27           5.2          3.5          10.0          0.2  Iris-setosa
28           5.2          3.4          10.0          0.2  Iris-setosa
29           4.7          3.2          10.0          0.2  Iris-setosa
```