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)

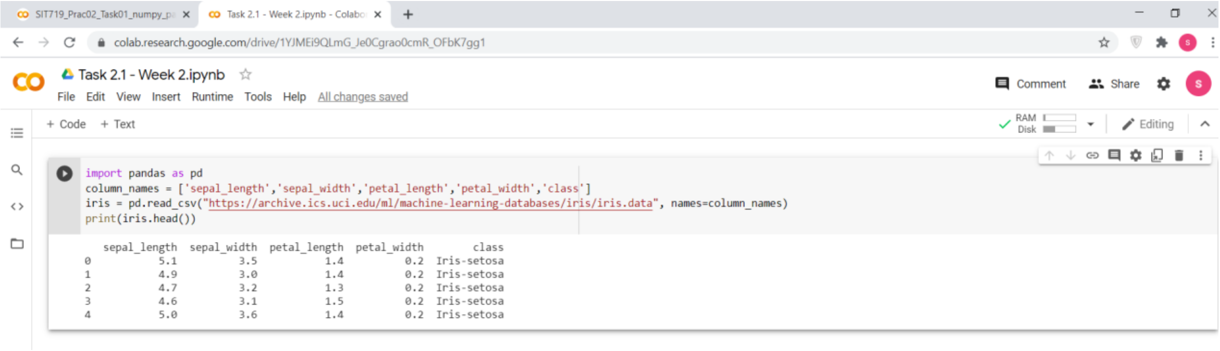
```
In [141]: url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'

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
column_names = ['sepal_length','sepal_width','petal_length','petal_width
','class']
iris = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-
databases/iris/iris.data", names=column_names)
print(iris.head())
```


SIT719_Prac02_Task01_numpy_p
Task 2.1 - Week 2.ipynb - Colabo
colab.research.google.com/drive/1YJMEi9QLmG_Je0Cgrao0cmR_OFbK7gg1
Task 2.1 - Week 2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing

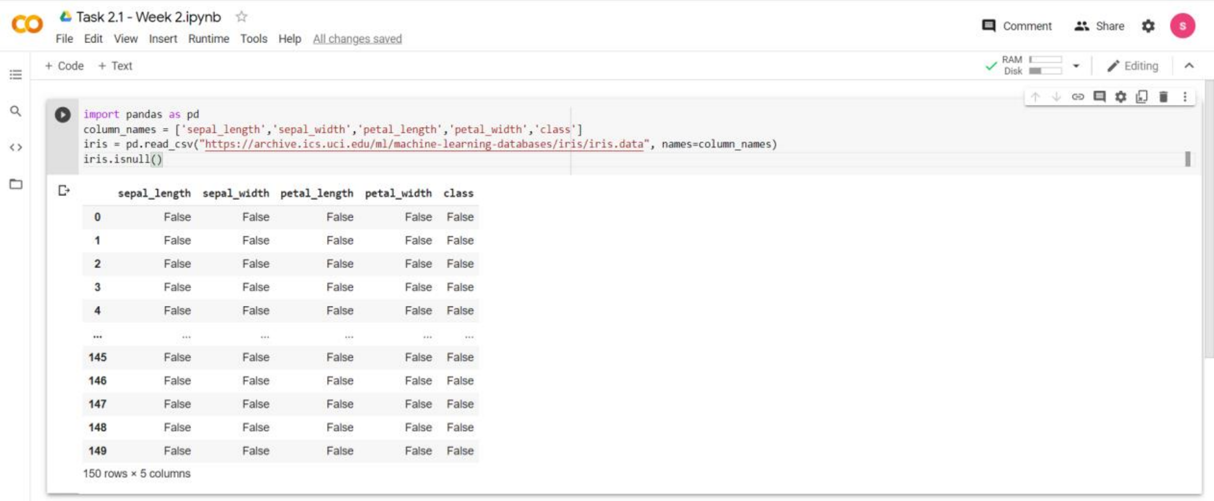
```
import pandas as pd
column_names = ['sepal_length','sepal_width','petal_length','petal_width','class']
iris = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data", names=column_names)
print(iris.head())
```

```
   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
```

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

Answer4: (write your code)

```
import pandas as pd
column_names = ['sepal_length','sepal_width','petal_length','petal_width','class']
iris = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data", names=column_names)
iris.isnull()
```



```
import pandas as pd
column_names = ['sepal_length','sepal_width','petal_length','petal_width','class']
iris = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data", names=column_names)
iris.isnull()
```

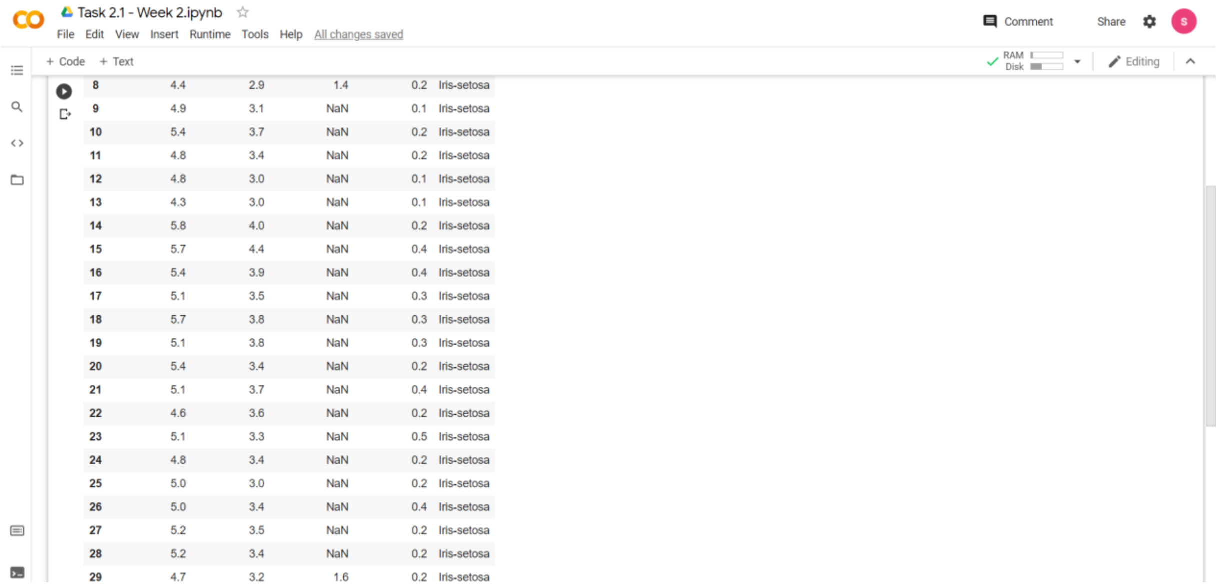
```
     sepal_length  sepal_width  petal_length  petal_width  class
0           False        False         False        False  False
1           False        False         False        False  False
2           False        False         False        False  False
3           False        False         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 × 5 columns
```

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
from numpy import nan as na
column_names = ['sepal_length','sepal_width','petal_length','petal_width','class']
iris = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data", names=column_names)
iris.loc[9:28]['petal_length']=na
iris.loc[8:30]
```

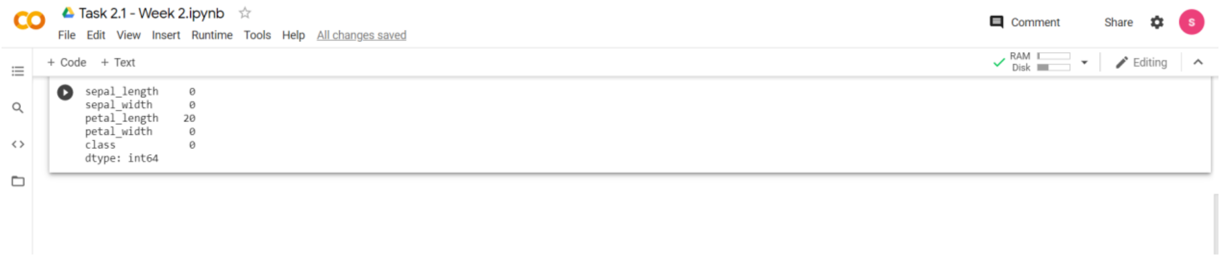


8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	NaN	0.1	Iris-setosa
10	5.4	3.7	NaN	0.2	Iris-setosa
11	4.8	3.4	NaN	0.2	Iris-setosa
12	4.8	3.0	NaN	0.1	Iris-setosa
13	4.3	3.0	NaN	0.1	Iris-setosa
14	5.8	4.0	NaN	0.2	Iris-setosa
15	5.7	4.4	NaN	0.4	Iris-setosa
16	5.4	3.9	NaN	0.4	Iris-setosa
17	5.1	3.5	NaN	0.3	Iris-setosa
18	5.7	3.8	NaN	0.3	Iris-setosa
19	5.1	3.8	NaN	0.3	Iris-setosa
20	5.4	3.4	NaN	0.2	Iris-setosa
21	5.1	3.7	NaN	0.4	Iris-setosa
22	4.6	3.6	NaN	0.2	Iris-setosa
23	5.1	3.3	NaN	0.5	Iris-setosa
24	4.8	3.4	NaN	0.2	Iris-setosa
25	5.0	3.0	NaN	0.2	Iris-setosa
26	5.0	3.4	NaN	0.4	Iris-setosa
27	5.2	3.5	NaN	0.2	Iris-setosa
28	5.2	3.4	NaN	0.2	Iris-setosa
29	4.7	3.2	1.6	0.2	Iris-setosa

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
from numpy import nan as na
column_names = ['sepal_length','sepal_width','petal_length','petal_width
','class']
iris = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-
databases/iris/iris.data", names=column_names)
iris.loc[9:28]['petal_length']=na
iris.isnull()
iris.isna().sum()
```



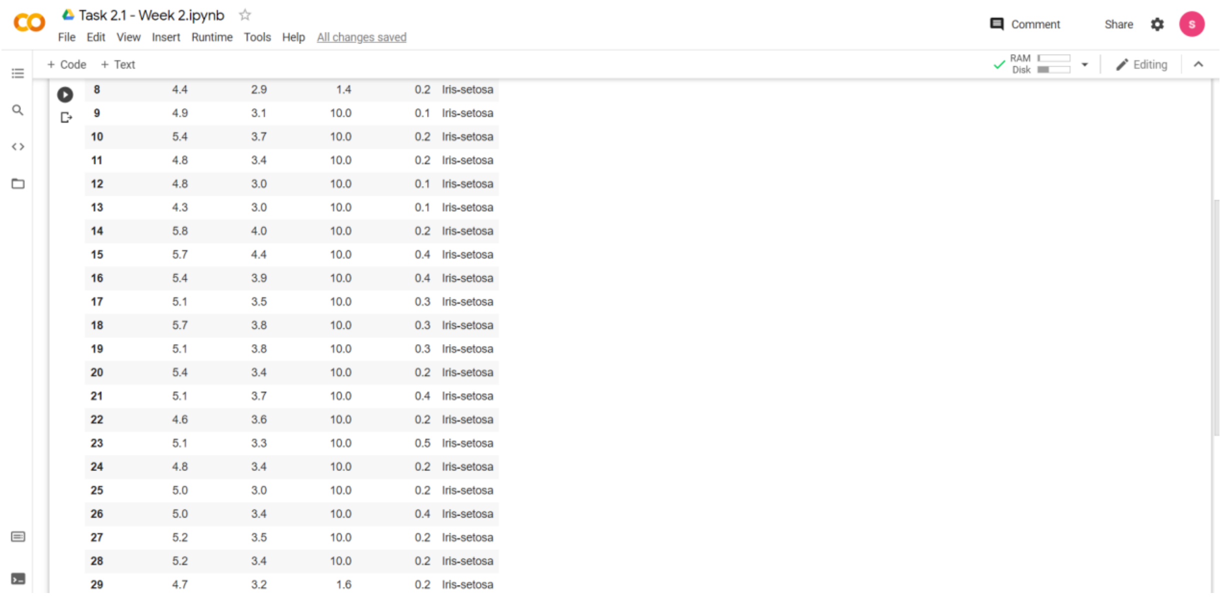
```
sepal_length     0
sepal_width      0
petal_length    20
petal_width      0
class            0
dtype: int64
```

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
from numpy import nan as na
column_names = ['sepal_length','sepal_width','petal_length','petal_width
','class']
iris = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-
databases/iris/iris.data", names=column_names)
iris.loc[9:28]['petal_length']=na
iris.isnull()
iris['petal_length'] = iris['petal_length'].replace(na,10)
iris.loc[8:30]
```



```
8    4.4   2.9    1.4   0.2   Iris-setosa
9    4.9   3.1   10.0   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    1.6   0.2   Iris-setosa
```

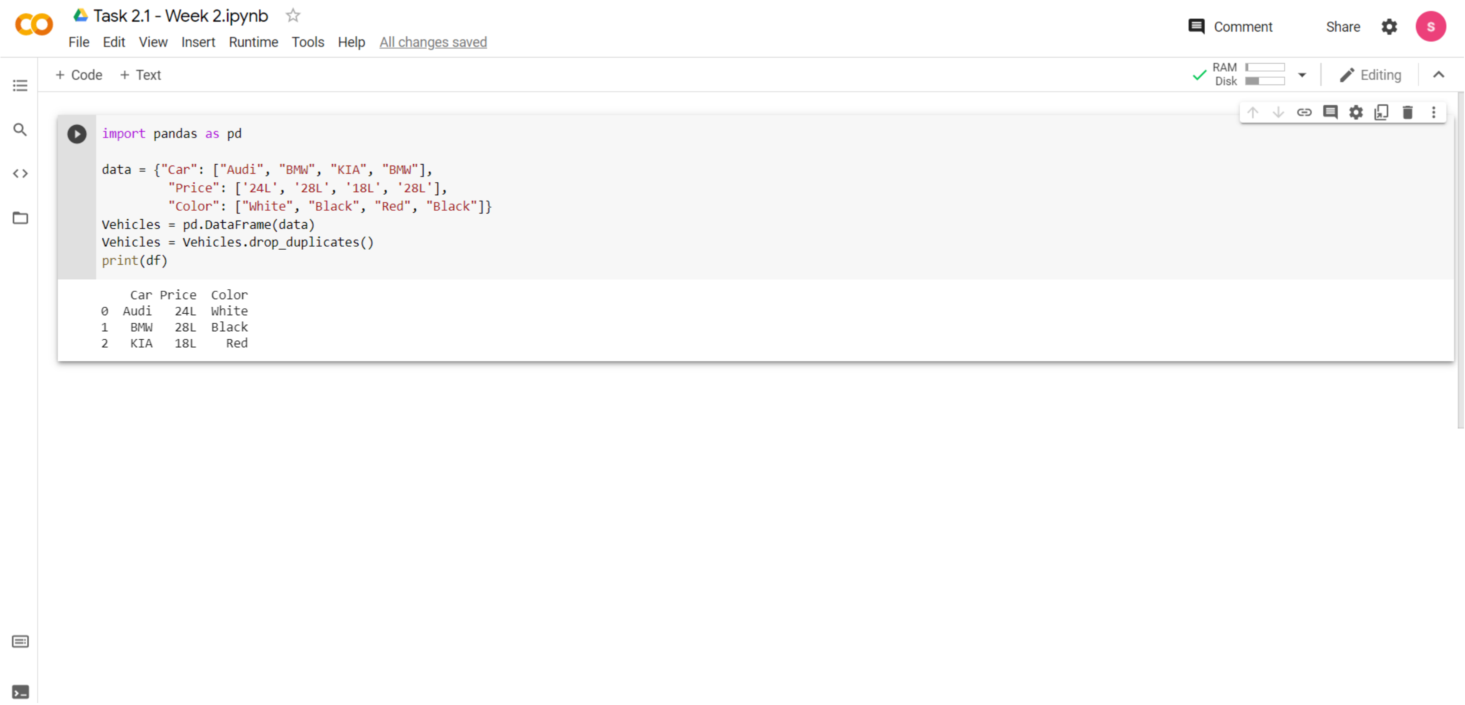
Section 2.

Importance of python libraries for data analysis

Python is one of the general-purpose languages used by data scientists. It is been popular among data scientists and programmers because of its simple commands. Python simplifies data processing and saves time. Some of the important Python libraries for data science are Numpy, Pandas, Matplotlib, and Scikit. These python libraries help in performing basic and advanced array operations - Numpy. Pandas help the developers to work with labeled and relational data. Matplotlib is a standard library for data visualization in two-dimensional diagrams and graphs.

Some common functionalities and usages related to dataframe manipulation (for example, NaN check, slicing dataset using iloc). Just show 2 examples.

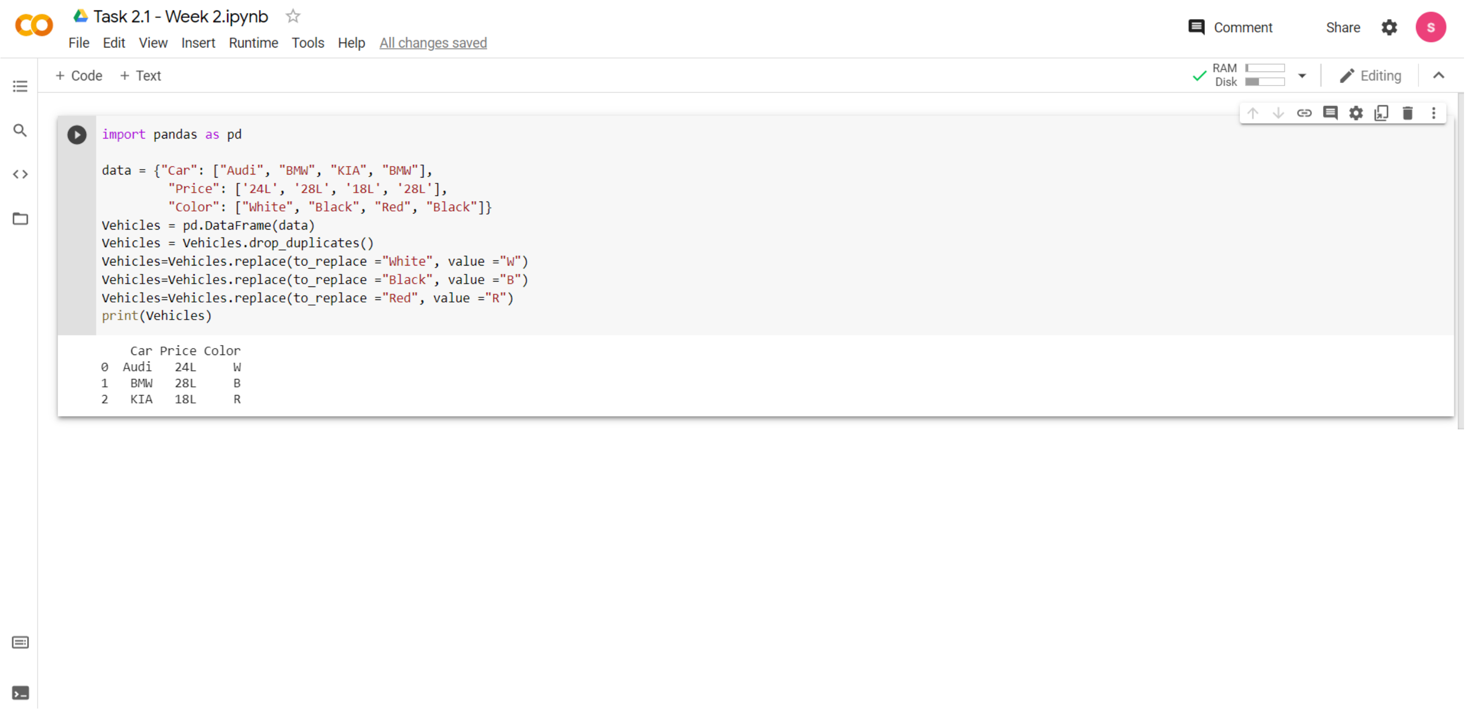
Dropping Duplicates



```
import pandas as pd

data = {"Car": ["Audi", "BMW", "KIA", "BMW"],
         "Price": ['24L', '28L', '18L', '28L'],
         "Color": ["White", "Black", "Red", "Black"]}
Vehicles = pd.DataFrame(data)
Vehicles = Vehicles.drop_duplicates()
print(df)
```

```
    Car Price  Color
0  Audi   24L  White
1   BMW   28L  Black
2   KIA   18L    Red
```

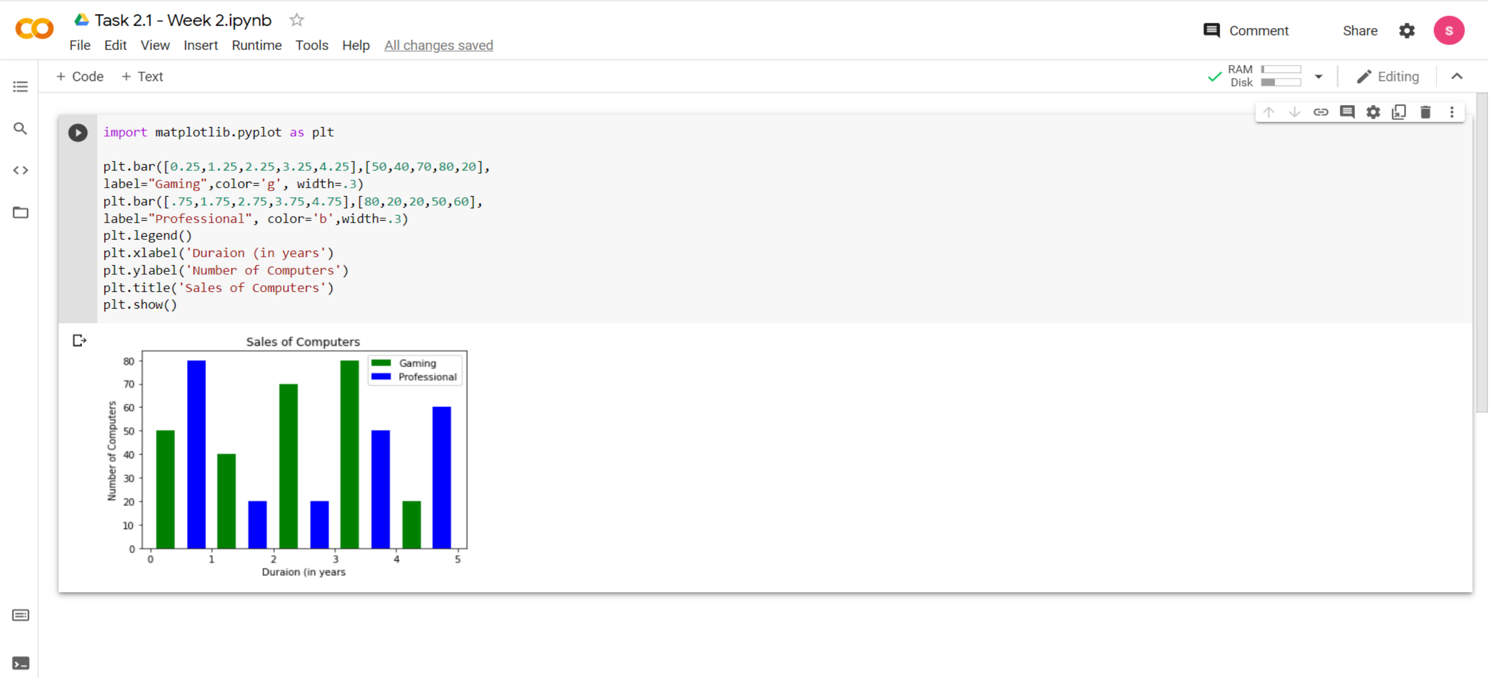
Replacing Values



```
import pandas as pd

data = {"Car": ["Audi", "BMW", "KIA", "BMW"],
         "Price": ['24L', '28L', '18L', '28L'],
         "Color": ["White", "Black", "Red", "Black"]}
Vehicles = pd.DataFrame(data)
Vehicles = Vehicles.drop_duplicates()
Vehicles=Vehicles.replace(to_replace ="White", value ="W")
Vehicles=Vehicles.replace(to_replace ="Black", value ="B")
Vehicles=Vehicles.replace(to_replace ="Red", value ="R")
print(Vehicles)
```

```
    Car Price Color
0  Audi   24L     W
1   BMW   28L     B
2   KIA   18L     R
```

Demonstration of a sample visualization example using matplotlib library.



```
import matplotlib.pyplot as plt

plt.bar([0.25,1.25,2.25,3.25,4.25],[50,40,70,80,20],
label="Gaming",color='g', width=.3)
plt.bar([.75,1.75,2.75,3.75,4.75],[80,20,20,50,60],
label="Professional", color='b',width=.3)
plt.legend()
plt.xlabel('Duraion (in years')
plt.ylabel('Number of Computers')
plt.title('Sales of Computers')
plt.show()
```