

PYTHON

DATA SCIENCE LANJUT

A. Mengolah Data Tabel Dengan Library Pandas.

Untuk mengolah data tabel library pandas pertama kita perlu mengimport terlebih dahulu dataset iris file csv yang telah dimasukkan di VSCode dengan nama iris.csv.

1. Import Pandas DataFrame

DataFrame adalah struktur data dua dimensi yang berbentuk tabular (mempunyai baris dan kolom). Hampir semua data memiliki banyak kolom, sehingga lebih cocok menggunakan Pandas DataFrame untuk mengolahnya. Penggunaan variabel dataframe pada Python biasanya menggunakan syntax: df.

a. Untuk mengimport pandas menggunakan sintak dibawah :

```
import pandas as pd
```

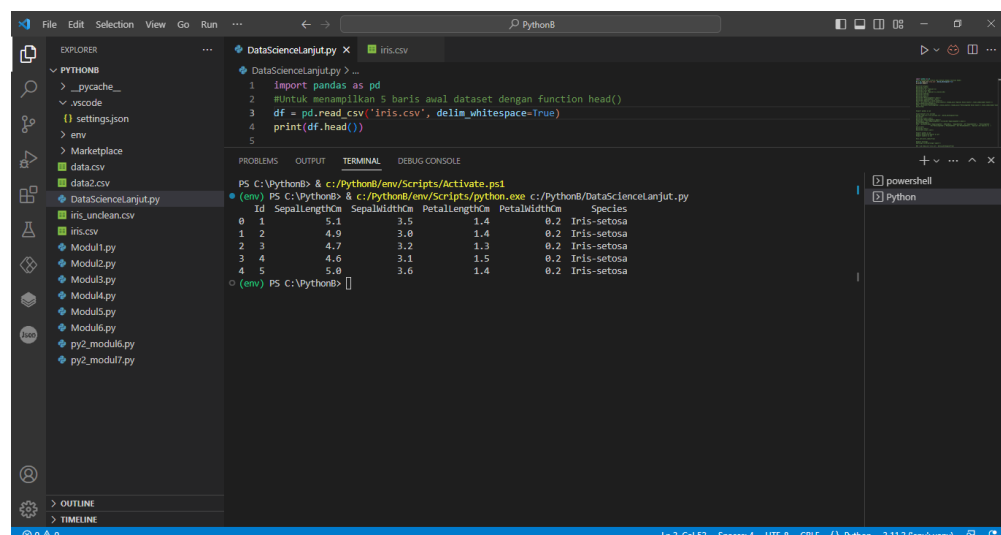
2. Mengolah DataFrame

Dengan memanggil file (iris.csv) yang sudah ada.

b. Untuk menampilkan 5 baris awal dataset menggunakan kode program berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df.head())
```

Dengan output :



The screenshot shows the VS Code interface with a file explorer on the left, a code editor in the center, and a terminal at the bottom. The code editor contains a Python script named `DataScienceLanjut.py` with the following content:

```
1 import pandas as pd
2 #Untuk menampilkan 5 baris awal dataset dengan function head()
3 df = pd.read_csv('iris.csv', delim_whitespace=True)
4 print(df.head())
5
```

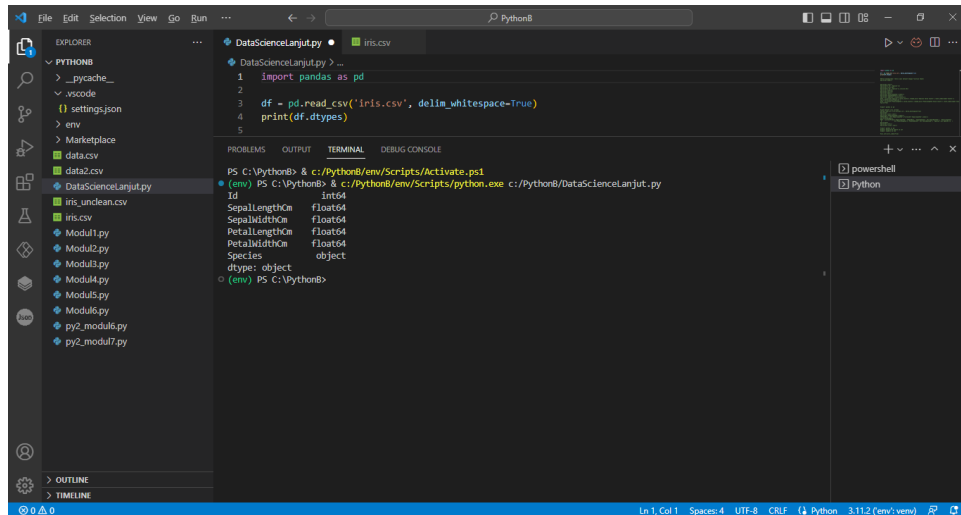
The terminal output shows the execution of the script, displaying the first 5 rows of the `iris.csv` dataset:

```
PS C:\Python8> & c:/Python8/env/Scripts/Activate.ps1
(env) PS C:\Python8> & c:/Python8/env/Scripts/python.exe c:/Python8/DataScienceLanjut.py
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm  Species
0  1         5.1         3.5         1.4         0.2  Iris-setosa
1  2         4.9         3.0         1.4         0.2  Iris-setosa
2  3         4.7         3.2         1.3         0.2  Iris-setosa
3  4         4.6         3.1         1.5         0.2  Iris-setosa
4  5         5.0         3.6         1.4         0.2  Iris-setosa
(env) PS C:\Python8>
```

- c. Untuk menampilkan tipe data dari kolom data yang ada pada dataset dengan menggunakan types, code program sebagai berikut:

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df.dtypes)
```

Dengan output :



The screenshot shows a VS Code editor with a Python file named 'DataScienceLanjut.py'. The code in the file is:

```
1 import pandas as pd
2
3 df = pd.read_csv('iris.csv', delim_whitespace=True)
4 print(df.dtypes)
```

The terminal output shows the following dtypes:

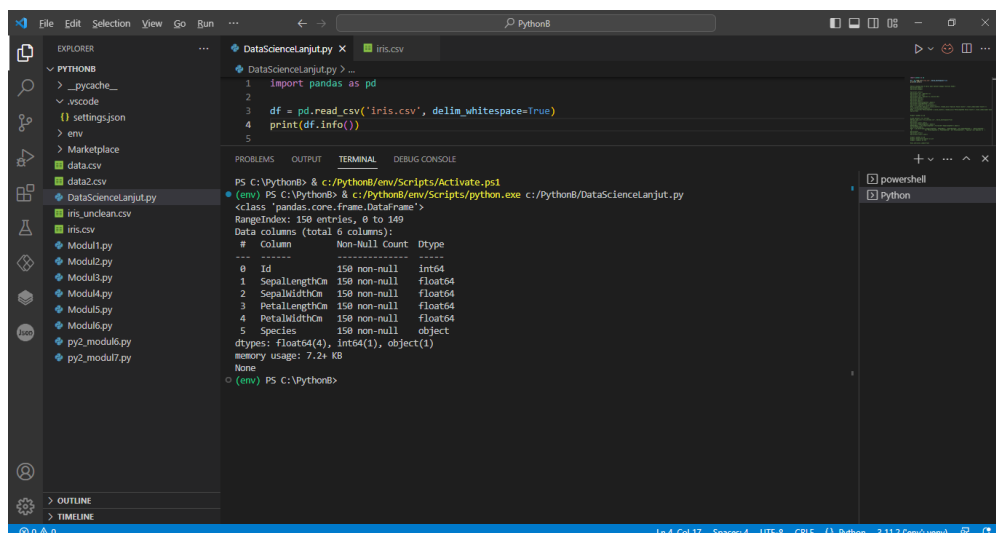
```
Id          int64
SepalLengthCm  float64
SepalWidthCm   float64
PetalLengthCm  float64
PetalWidthCm   float64
Species       object
dtype: object
```

Tipe Data dari kolom yang ada di dataset:

1. Kolom "Id" memiliki tipe data Integer (int64).
 2. Kolom "SepalLengthCm" memiliki tipe data = Float (float64).
 3. Kolom "Species" memiliki tipe data = Object (object).
- d. Hitung ukuran (jumlah baris dan kolom) dari dataset dapat menggunakan df.info(), code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df.info())
```

Output :



The screenshot shows a VS Code editor with a Python file named 'DataScienceLanjut.py'. The code in the file is:

```
1 import pandas as pd
2
3 df = pd.read_csv('iris.csv', delim_whitespace=True)
4 print(df.info())
```

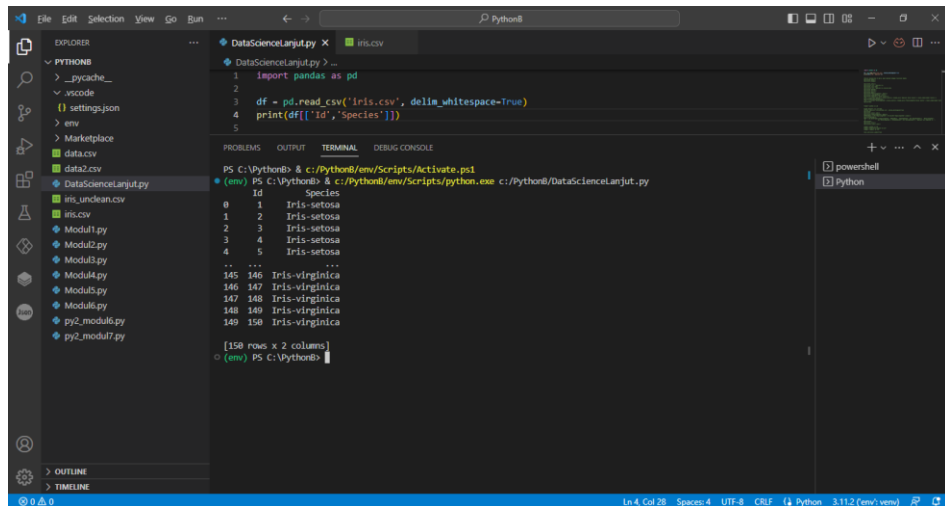
The terminal output shows the following information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
#   column      Non-Null count  Dtype
---  -
0    Id          150 non-null      int64
1    SepalLengthCm  150 non-null      float64
2    SepalWidthCm   150 non-null      float64
3    PetalLengthCm  150 non-null      float64
4    PetalWidthCm   150 non-null      float64
5    Species       150 non-null      object
dtypes: float64(4), int64(1), object(1)
memory usage: 7.2+ KB
None
```

- e. Tampilkan data untuk kolom "Id" kolom dan 'Species' dalam bentuk dataframe dengan `df[["Id", "Species"]]`, code program sebagai berikut:

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df[['Id', 'Species']])
```

Output :



```
PS C:\Python8 & c:/Python8/env/Scripts/activate.ps1
(env) PS C:\Python8 & c:/Python8/env/Scripts/python.exe c:/Python8/DataScienceLanjut.py

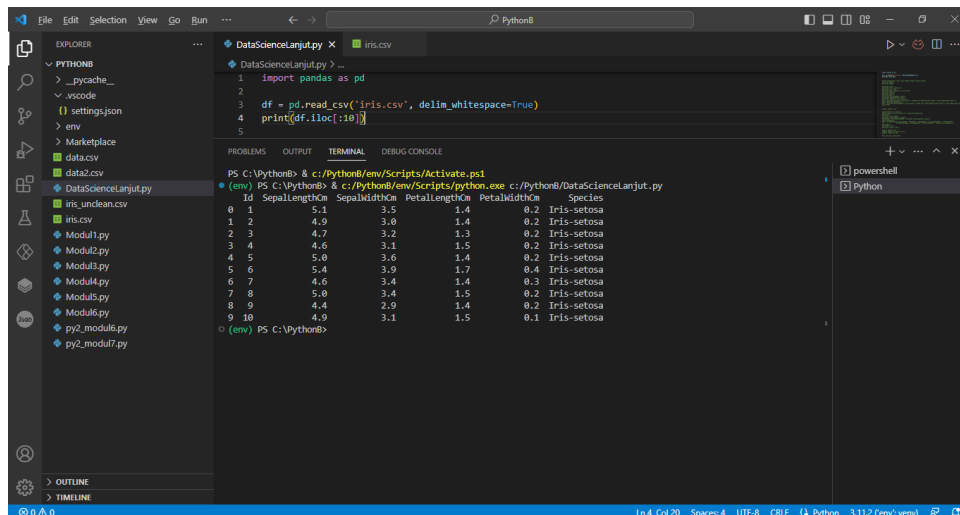
Id      Species
0      1  Iris-setosa
1      2  Iris-setosa
2      3  Iris-setosa
3      4  Iris-setosa
4      5  Iris-setosa
...
145    146 Iris-virginica
146    147 Iris-virginica
147    148 Iris-virginica
148    149 Iris-virginica
149    150 Iris-virginica

[150 rows x 2 columns]
(env) PS C:\Python8 >
```

- f. Tampilkan data baris indeks ke-0 (nol) sampai dengan indeks ke-9 (sembilan) dengan `df.iloc[0:10]`, code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df.iloc[0:10])
```

Output :



```
PS C:\Python8 & c:/Python8/env/Scripts/activate.ps1
(env) PS C:\Python8 & c:/Python8/env/Scripts/python.exe c:/Python8/DataScienceLanjut.py

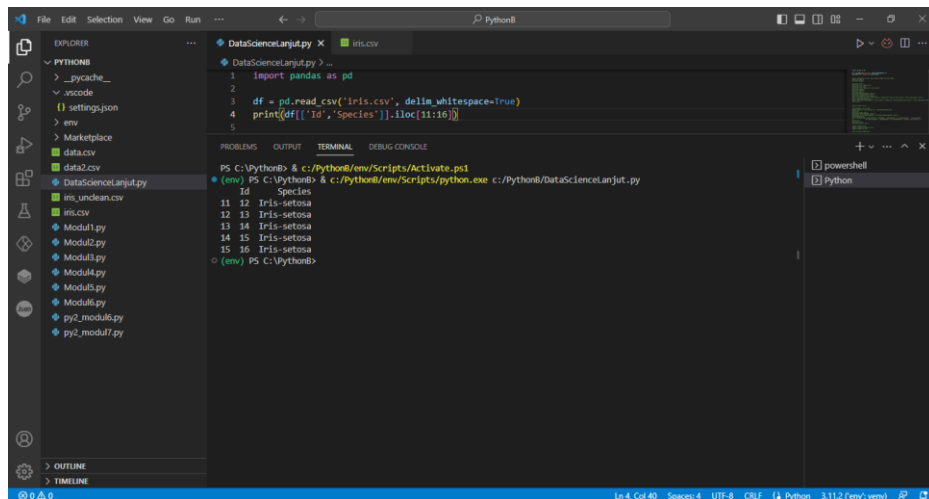
Id      SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm  Species
0      1          5.1         3.5          1.4         0.2  Iris-setosa
1      2          4.9         3.0          1.4         0.2  Iris-setosa
2      3          4.7         3.2          1.3         0.2  Iris-setosa
3      4          4.6         3.1          1.5         0.2  Iris-setosa
4      5          5.0         3.6          1.4         0.2  Iris-setosa
5      6          5.4         3.9          1.7         0.4  Iris-setosa
6      7          4.6         3.4          1.4         0.3  Iris-setosa
7      8          5.0         3.4          1.5         0.2  Iris-setosa
8      9          4.4         2.9          1.4         0.2  Iris-setosa
9     10          4.9         3.1          1.5         0.1  Iris-setosa

(env) PS C:\Python8 >
```

- g. Tampilkan data hanya kolom "Id" dan kolom "Species", dan yang ditampilkan adalah data indeks ke-11 (sebelas) sampai dengan indeks ke-15 (limabelas) dengan `df[['Id','Species']].iloc[11:16]`, code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df[['Id','Species']].iloc[11:16])
```

Output :

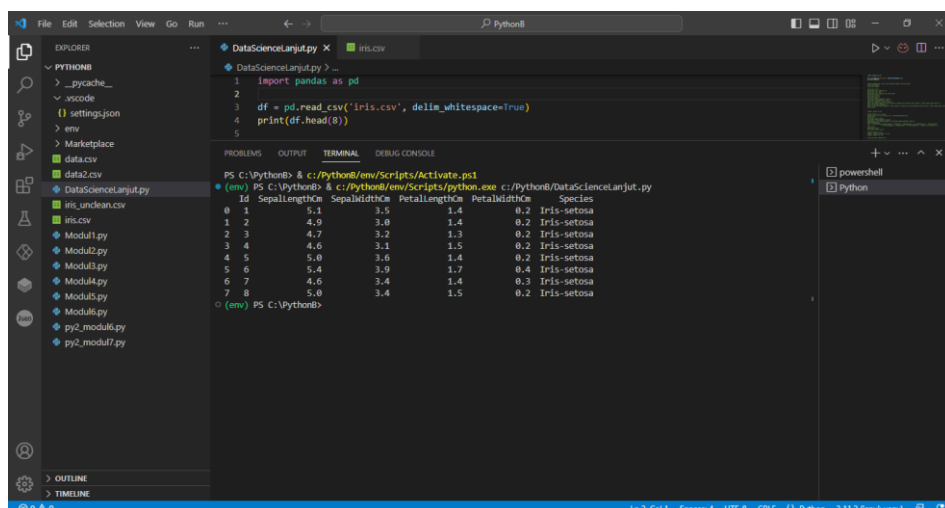


```
PS C:\Python8> & c:/Python8/env/Scripts/Activate.ps1
(env) PS C:\Python8> & c:/Python8/env/Scripts/python.exe c:/Python8/DataScienceLanjut.py
Id Species
11 12 Iris-setosa
12 13 Iris-setosa
13 14 Iris-setosa
14 15 Iris-setosa
15 16 Iris-setosa
```

- h. Tampilkan data 8 baris pertama dengan `df.head(8)`, code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df.head(8))
```

Output :

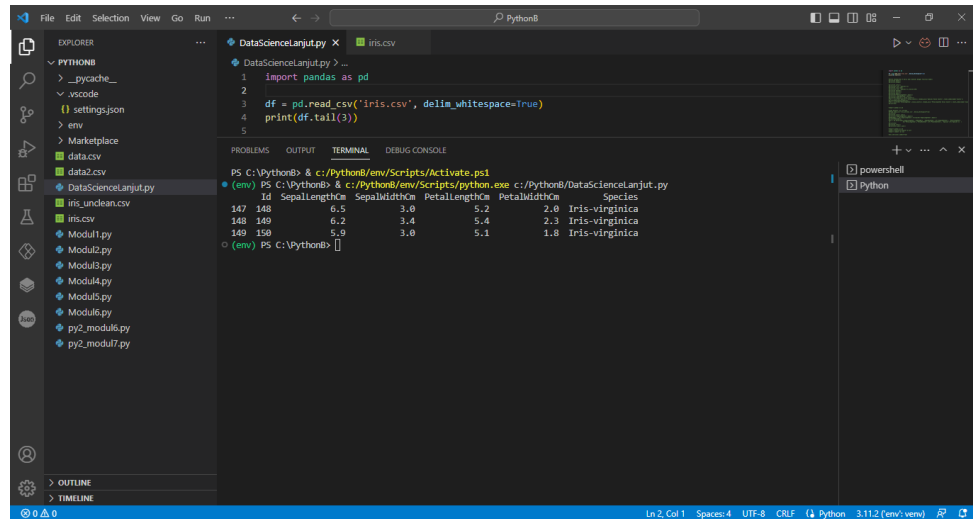


```
PS C:\Python8> & c:/Python8/env/Scripts/Activate.ps1
(env) PS C:\Python8> & c:/Python8/env/Scripts/python.exe c:/Python8/DataScienceLanjut.py
Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species
0 1 5.1 3.5 1.4 0.2 Iris-setosa
1 2 4.9 3.0 1.4 0.2 Iris-setosa
2 3 4.7 3.2 1.3 0.2 Iris-setosa
3 4 4.6 3.1 1.5 0.2 Iris-setosa
4 5 5.0 3.6 1.4 0.2 Iris-setosa
5 6 5.4 3.9 1.7 0.4 Iris-setosa
6 7 4.6 3.4 1.4 0.3 Iris-setosa
7 8 5.0 3.4 1.5 0.2 Iris-setosa
```

- i. Tampilkan data 3 baris terakhir dengan `df.tail(3)`, code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df.tail(3))
```

Output :



The screenshot shows a VS Code editor with a file named `iris.csv` open. The code in the editor is:

```
1 import pandas as pd
2
3 df = pd.read_csv('iris.csv', delim_whitespace=True)
4 print(df.tail(3))
5
```

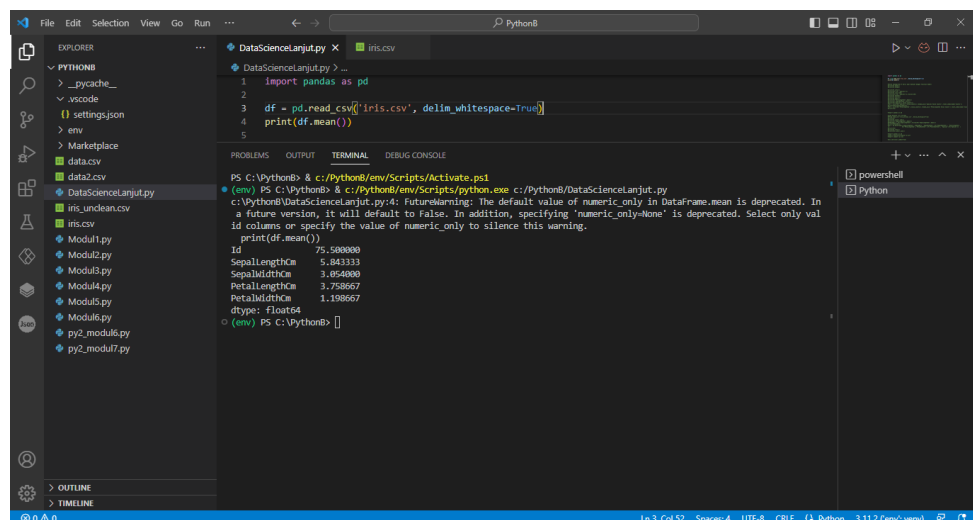
The terminal output shows the following data:

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
147	14.8	6.5	3.0	5.2	Iris-virginica
148	14.9	6.2	3.4	5.4	Iris-virginica
149	15.0	5.9	3.0	5.1	Iris-virginica

- j. Hitung nilai mean dari dataset dengan `mean()`, code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df.mean())
```

Output :



The screenshot shows a VS Code editor with a file named `iris.csv` open. The code in the editor is:

```
1 import pandas as pd
2
3 df = pd.read_csv('iris.csv', delim_whitespace=True)
4 print(df.mean())
5
```

The terminal output shows the following data:

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
75.500000	5.843333	3.854000	3.758667	1.198667

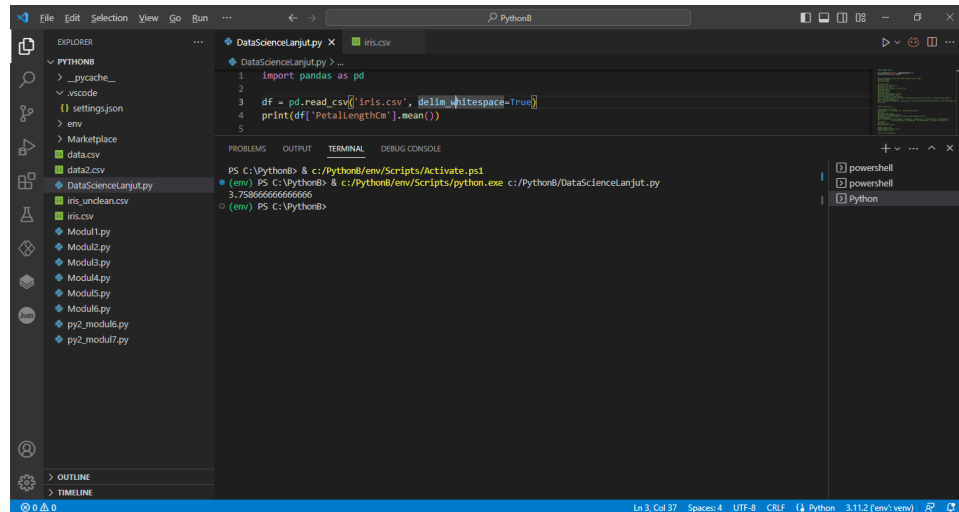
A warning message is also displayed in the terminal:

```
FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is deprecated. Select only valid id columns or specify the value of numeric_only to silence this warning.
```

- k. Hitung nilai mean untuk kolom PetalLengthCm dengan `df['PetalLengthCm'].mean()`, code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df['PetalLengthCm'].mean())
```

Output :



The screenshot shows the Visual Studio Code interface with a Python file named `DataScienceLanjut.py` open. The code in the editor is:

```
1 import pandas as pd
2
3 df = pd.read_csv('iris.csv', delim_whitespace=True)
4 print(df['PetalLengthCm'].mean())
5
```

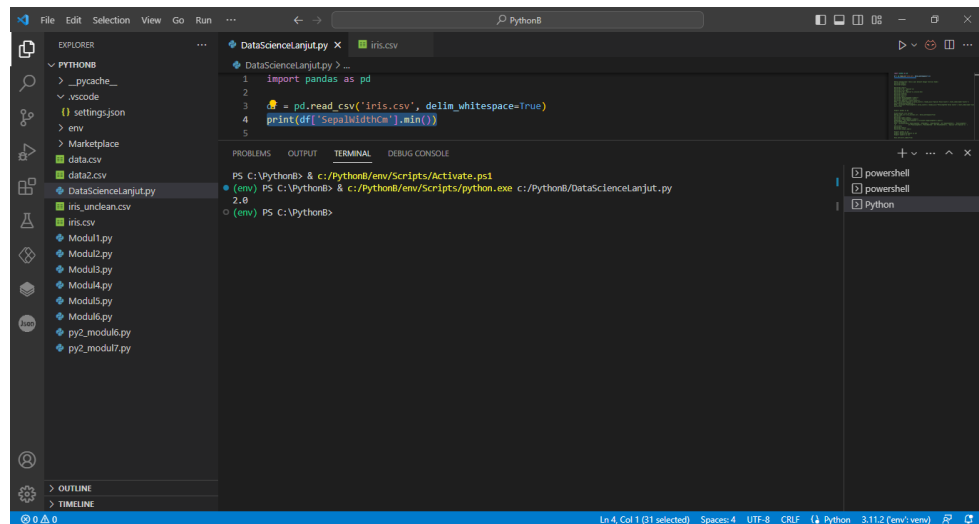
The terminal output shows the command prompt running the script and displaying the result:

```
PS C:\Python8> & c:/Python8/env/Scripts/Activate.ps1
(env) PS C:\Python8> & c:/Python8/env/Scripts/python.exe c:/Python8/DataScienceLanjut.py
3.7586666666666666
(env) PS C:\Python8>
```

- l. Cari nilai minimal untuk kolom SepalWidth this code program sebagai berikut :

```
print(df['SepalWidthCm'].min())
```

Output :



The screenshot shows the Visual Studio Code interface with the same Python file `DataScienceLanjut.py`. The code in the editor is:

```
1 import pandas as pd
2
3 df = pd.read_csv('iris.csv', delim_whitespace=True)
4 print(df['SepalWidthCm'].min())
5
```

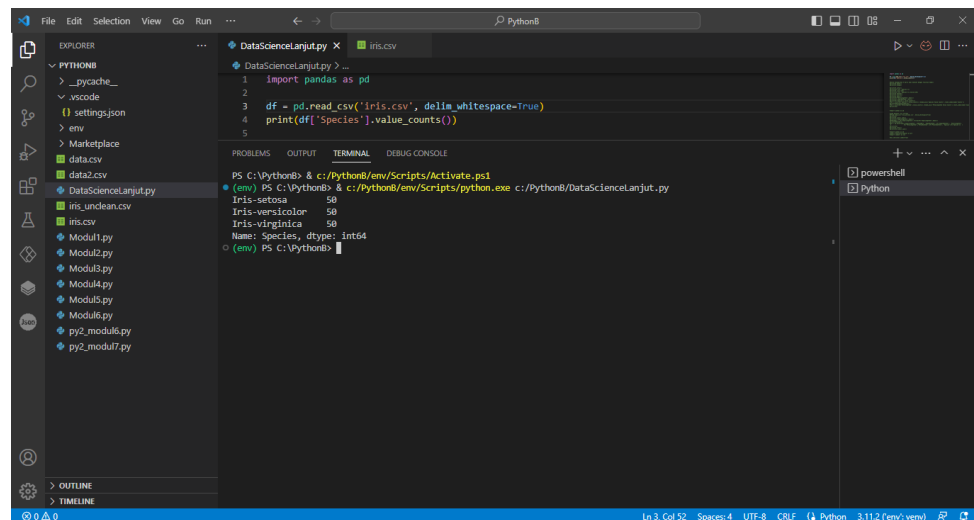
The terminal output shows the command prompt running the script and displaying the result:

```
PS C:\Python8> & c:/Python8/env/Scripts/Activate.ps1
(env) PS C:\Python8> & c:/Python8/env/Scripts/python.exe c:/Python8/DataScienceLanjut.py
2.0
(env) PS C:\Python8>
```

- m. Hitung frekuensi pada kolom Species dengan menggunakan metode `value_counts()` dengan `df["Species"].value_counts()`, code program sebagai berikut :

```
print(df['Species'].value_counts())
```

Output :



```
1 import pandas as pd
2
3 df = pd.read_csv('iris.csv', delim_whitespace=True)
4 print(df['Species'].value_counts())
5
```

```
PS C:\Python8> & c:/Python8/Scripts/Activate.ps1
(env) PS C:\Python8> c:/Python8/Scripts/python.exe c:/Python8/DataScienceLanjut.py
Iris-setosa      50
Iris-versicolour 50
Iris-virginica   50
Name: Species, dtype: object
(env) PS C:\Python8>
```

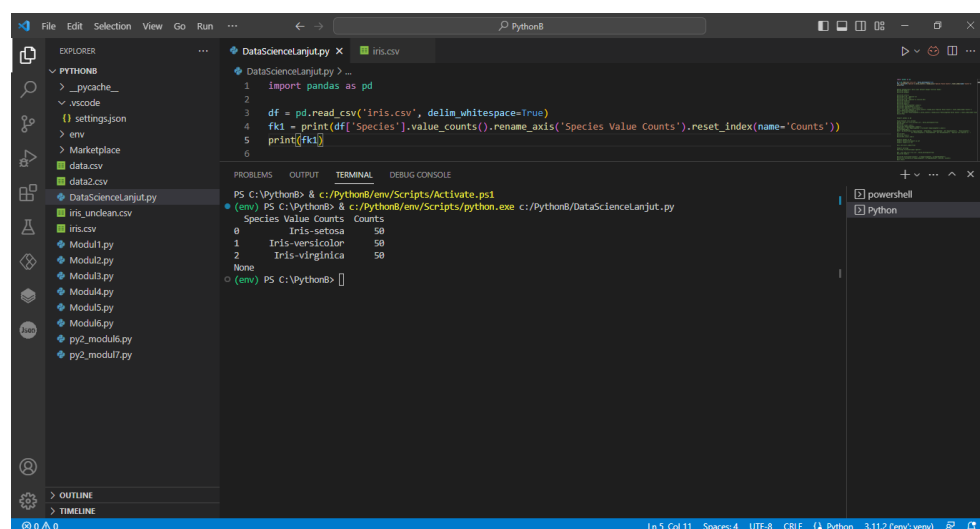
- n. Tampilkan perhitungan frekuensi pada kolom Speciem dengan menggunakan `value_counts()` dalam bentuk dataframe, code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
```

```
fk1 = print(df['Species'].value_counts().rename_axis('Species Value Counts').reset_index(name='Counts'))
```

```
print(fk1)
```

Output :



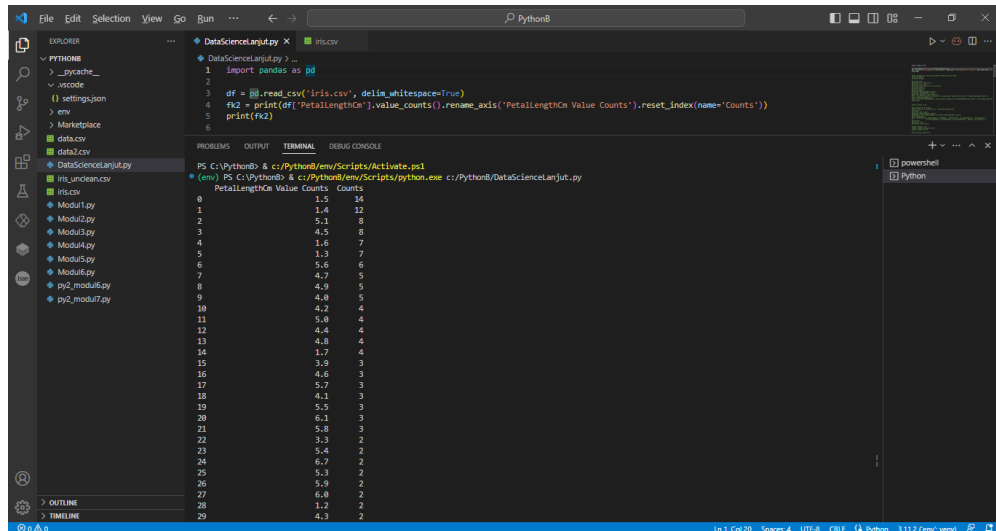
```
1 import pandas as pd
2
3 df = pd.read_csv('iris.csv', delim_whitespace=True)
4 fk1 = print(df['Species'].value_counts().rename_axis('Species Value Counts').reset_index(name='Counts'))
5 print(fk1)
6
```

```
PS C:\Python8> & c:/Python8/Scripts/Activate.ps1
(env) PS C:\Python8> c:/Python8/Scripts/python.exe c:/Python8/DataScienceLanjut.py
Species Value Counts  Counts
0      Iris-setosa      50
1  Iris-versicolour      50
2      Iris-virginica      50
None
(env) PS C:\Python8>
```

- o. Hitung frekuensi pada kolom PetalLengthCm dengan menggunakan `value_counts()` dalam bentuk datafram, code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
fk2=df['PetalLengthCm'].value_counts().rename_axis('PetalLengthCm
Value Counts').reset_index(name='Counts')
print(fk2)
```

Output :



B. Mengolah Data Iris Yang Tidak Lengkap

1. Download data iris unclean kemudian masukkan ke dalam vscode sebagai dataset file csv.
2. Import Pandas DataFrame
 - a. Import data pandas dengan menggunakan kode program sebagai berikut :
3. Load Dataset dan Cek Data
 - b. Load dataset iris unclean dapat menggunakan code program berikut:

```
df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
```

- c. Tampil data atau cetak dataset iris unclean dapat menggunakan code program berikut:

```
df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
print(df)
```


[illegible]

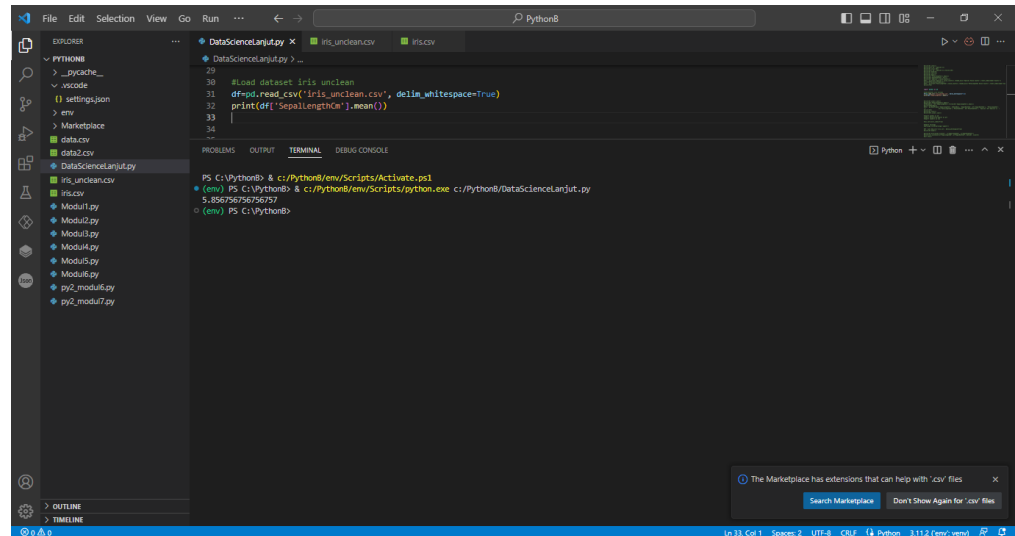
```
df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
print(df.isna().sum())
```

Inputasi adalah pilihan penanganan missing data yang paling bijak daripada membuang sebagian observasi atau variabel yang mengandung missing value, mengingat bahwa data sangat mahal dan berharga. Sebelumnya terlihat bahwa ada 2 data yang hilang pada SepalLengthCm.

- e. Cari nilai mean dari SepalLengthCm, dengan code program sebagai berikut :

```
df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
print(df['SepalLengthCm'].mean())
```

Output :



The screenshot shows the Visual Studio Code interface with a Python file named 'DataScienceLatut.py'. The code in the editor is as follows:

```
29
30 #load dataset iris unclean
31 df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
32 print(df['SepalLengthCm'].mean())
33
```

The terminal output shows the command prompt running the script and displaying the mean value:

```
PS C:\Python8> & c:/Python8/env/Scripts/Activate.ps1
(env) PS C:\Python8> & c:/Python8/env/Scripts/python.exe c:/Python8/DataScienceLatut.py
5.8506756756756757
(env) PS C:\Python8>
```

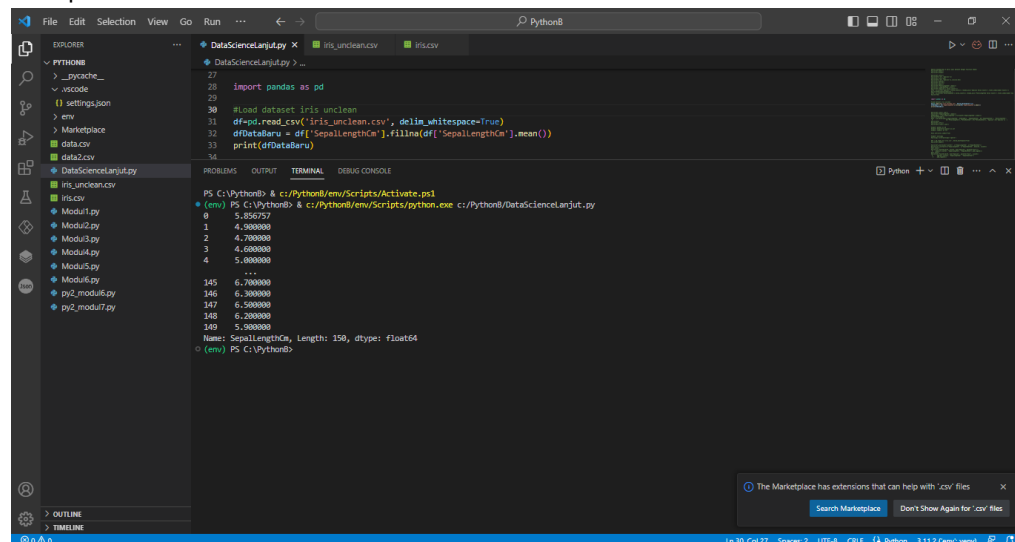
- f. Mengganti missing value dengan mean(), kemudian masukkan kedalam variabel, dengan menggunakan code program sebagai berikut :

```
df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
dfDataBaru=df['SepalLengthCm'].fillna(df['SepalLengthCm'].mean())
```

- g. Cetak data baru dengan code program :

```
print(dfDataBaru)
```

Output :



The screenshot shows the Visual Studio Code interface with the same Python file. The code in the editor is as follows:

```
27
28 import pandas as pd
29
30 #load dataset iris unclean
31 df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
32 dfDataBaru = df['SepalLengthCm'].fillna(df['SepalLengthCm'].mean())
33 print(dfDataBaru)
34
```

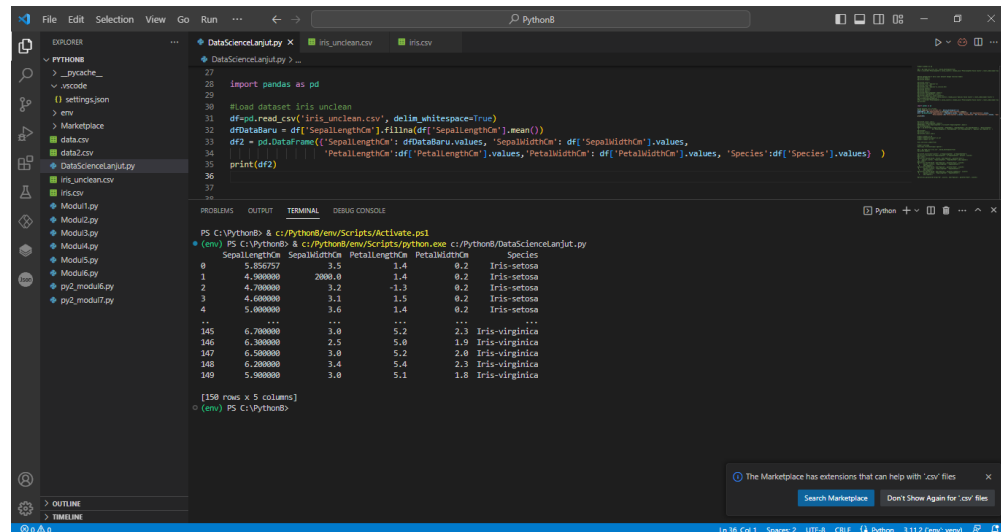
The terminal output shows the command prompt running the script and displaying the resulting DataFrame:

```
PS C:\Python8> & c:/Python8/env/Scripts/Activate.ps1
(env) PS C:\Python8> & c:/Python8/env/Scripts/python.exe c:/Python8/DataScienceLatut.py
0      5.850675
1      4.900000
2      4.700000
3      4.600000
4      5.000000
...
145     6.700000
146     6.300000
147     6.500000
148     6.200000
149     5.900000
Name: SepalLengthCm, Length: 150, dtype: float64
(env) PS C:\Python8>
```

- h. Gabung data baru menjadi DataFrame, dengan code program berikut:

```
df2=pd.DataFrame({'SepalLengthCm':dfDataBaru.values,'SepalWidthCm':df['SepalWidthCm'].values,'PetalLengthCm':df['PetalLengthCm'].values,'PetalWidthCm':df['PetalWidthCm'].values,'Species':df['Species'].values})
print(df2)
```

Output :



The screenshot shows a VS Code editor with a Python file named 'DataScienceLanjut.py'. The code in the file is as follows:

```
27
28
29 import pandas as pd
30
31 #load dataset iris unclean
32 df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
33 dfDataBaru = df['SepalLengthCm'].fillna(df['SepalLengthCm'].mean())
34 df2 = pd.DataFrame({'SepalLengthCm': dfDataBaru.values, 'SepalWidthCm': df['SepalWidthCm'].values,
35                    'PetalLengthCm': df['PetalLengthCm'].values, 'PetalWidthCm': df['PetalWidthCm'].values, 'Species': df['Species'].values})
36 print(df2)
37
```

The terminal output shows the first 5 rows of the resulting DataFrame:

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	5.856737	3.5	1.4	0.2	Iris-setosa
1	4.900000	2.800000	1.4	0.2	Iris-setosa
2	4.700000	3.2	1.3	0.2	Iris-setosa
3	4.600000	3.1	1.5	0.2	Iris-setosa
4	5.000000	3.6	1.4	0.2	Iris-setosa

The terminal also shows the last 5 rows of the DataFrame:

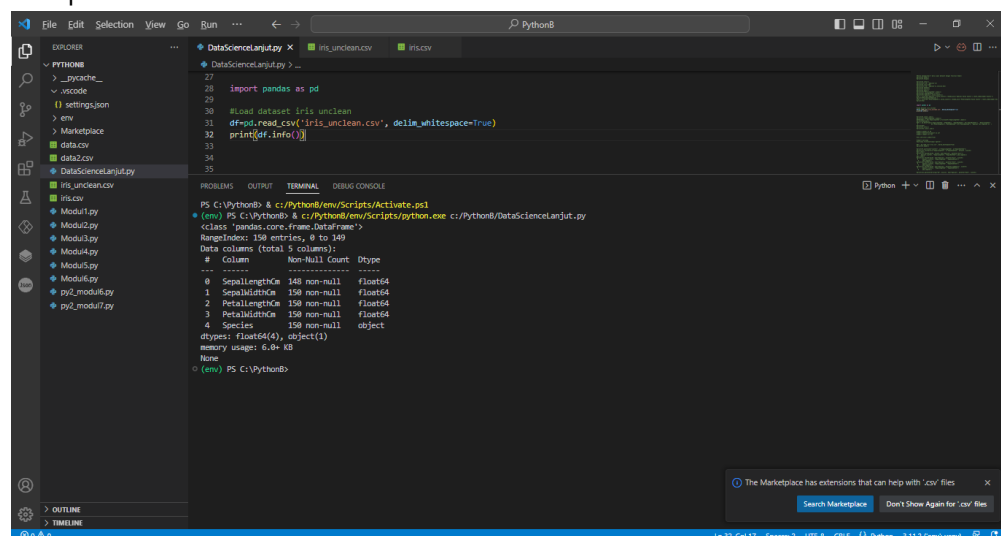
	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
145	6.700000	3.0	5.2	2.3	Iris-virginica
146	6.300000	2.5	5.0	1.9	Iris-virginica
147	6.500000	3.0	5.2	2.0	Iris-virginica
148	6.200000	3.4	5.4	2.3	Iris-virginica
149	5.900000	3.0	5.1	1.8	Iris-virginica

The terminal output concludes with: [150 rows x 5 columns] (env) PS C:\Python8\.

- i. Cek jumlah baris dan kolom, dengan code program berikut :

```
df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
print(df.info())
```

Output :



The screenshot shows a VS Code editor with a Python file named 'DataScienceLanjut.py'. The code in the file is as follows:

```
27
28
29 import pandas as pd
30
31 #load dataset iris unclean
32 df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
33 print(df.info())
34
```

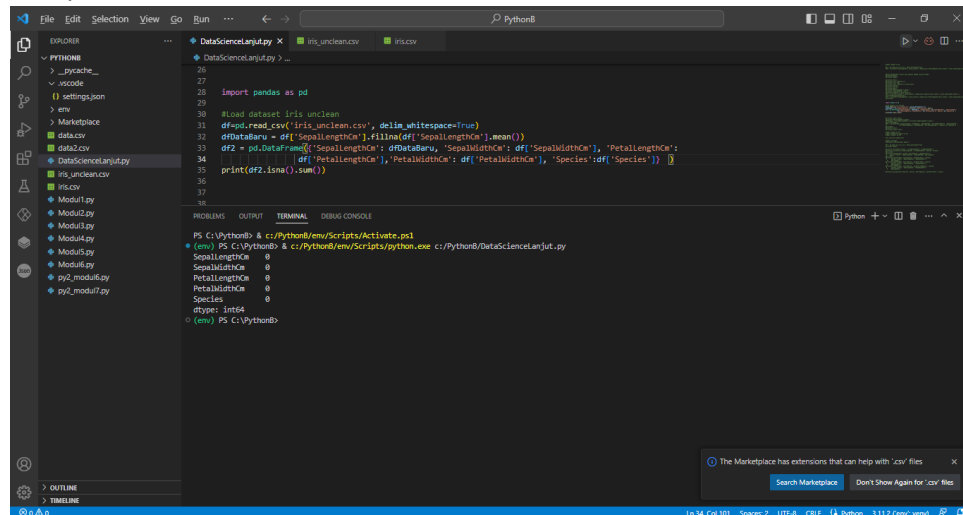
The terminal output shows the information for the DataFrame:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  --
 0   SepalLengthCm  148 non-null   float64
 1   SepalWidthCm   150 non-null   float64
 2   PetalLengthCm  150 non-null   float64
 3   PetalWidthCm   150 non-null   float64
 4   Species        150 non-null   object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
None
(env) PS C:\Python8\.
```

- j. Hitung jumlah nilai null pada DataFrame baru, dengan code program berikut :

```
df=pd.read_csv('iris_unclean.csv', delim_whitespace=True)
dfDataBaru=df['SepalLengthCm'].fillna(df['SepalLengthCm'].mean())
df2 = pd.DataFrame({'SepalLengthCm': dfDataBaru, 'SepalWidthCm':
df['SepalWidthCm'], 'PetalLengthCm':df['PetalLengthCm'], 'PetalWidh
Cm': df['PetalWidthCm'], 'Species':df['Species']})
print(df2.isna().sum())
```

Output :



C. Visualisasi DataSet Iris

1. Download data iris kemudian masukkan ke dalam vscode sebagai dataset file csv.
2. Import Library Pandas, Metplotlib, dan seaborn
 - a. kode program import library pandas, metplotlib, dan seaborn sebagai berikut :

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

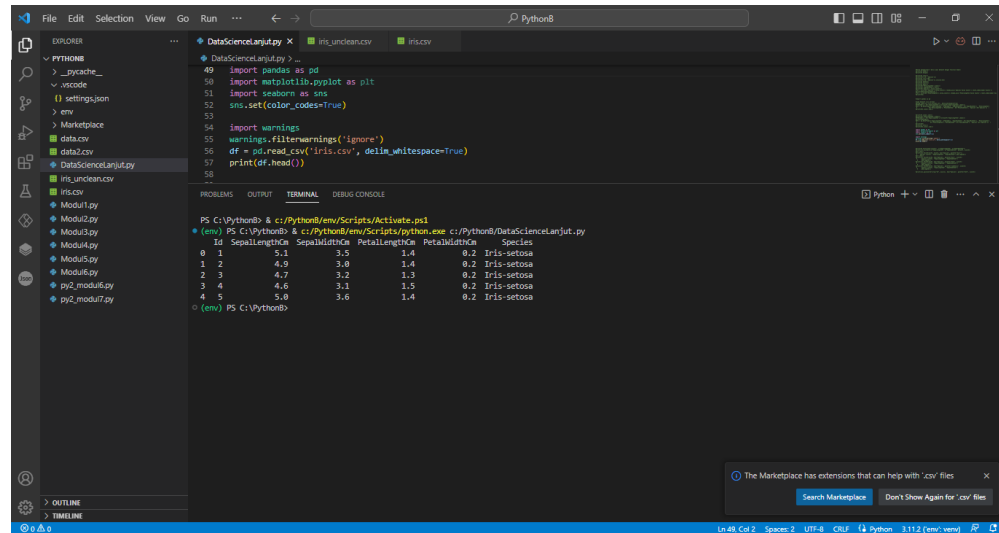
sns.set(color_codes=True)
import warnings

warnings.filterwarnings('ignore')
```

Untuk menampilkan data menggunakan code program berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
print(df.head())
```

Output :



The screenshot shows a VS Code editor with a file named 'DataScienceLanjut.py' open. The code in the file is as follows:

```
40 import pandas as pd
41 import matplotlib.pyplot as plt
42 import seaborn as sns
43 sns.set(color_codes=True)
44
45 import warnings
46 warnings.filterwarnings('ignore')
47 df = pd.read_csv('iris.csv', delim_whitespace=True)
48 print(df.head())
```

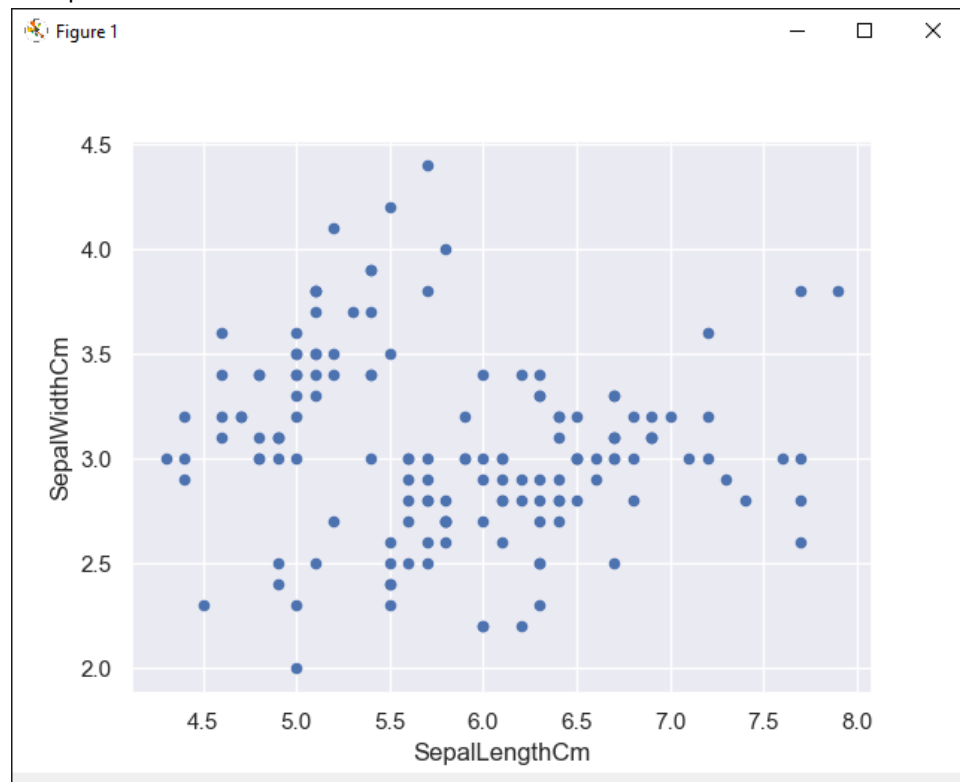
The terminal output shows the first five rows of the 'iris.csv' file:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

b. Membuat scatter plot dari fitur SepalLengthCm dan SepalWidthCm, menggunakan code program berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
df.plot(kind="scatter",x="SepalLengthCm",y="SepalWidthCm")
plt.show()
```

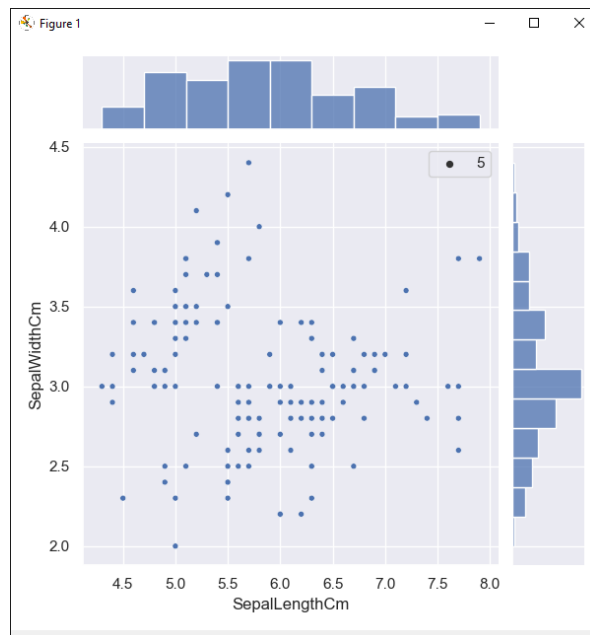
Output :



- c. Membuat scatter plot dengan library seaborn, code program sebagai berikut :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
sns.jointplot(x="SepalLengthCm", y="SepalWidthCm", data=df,
size=5)
plt.show()
```

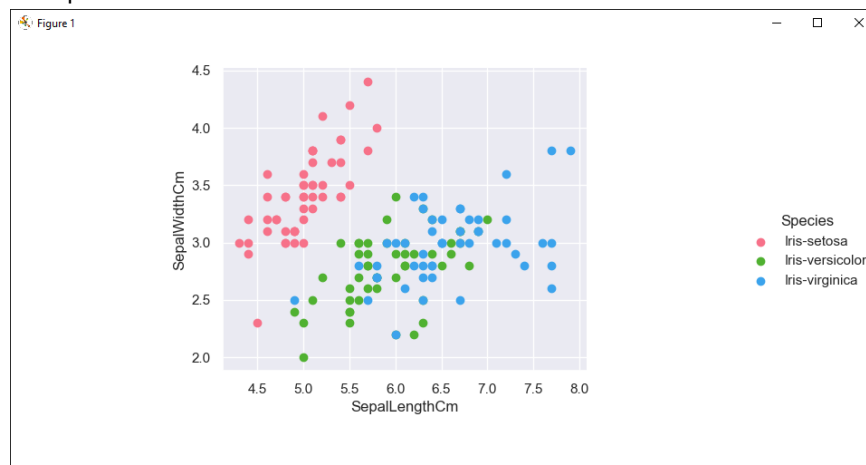
Output :



- d. Salah satu informasi yang hilang dalam plot di atas adalah jenis tanaman (Species), gunakan FacetGrid Seaborn untuk mewarnai sebaran Species, code program sebagai berikut :

```
sns.FacetGrid(df, hue="Species", palette="husl").map(plt.scatter,
"SepalLengthCm", "SepalWidthCm").add_legend()
plt.show()
```

Output :



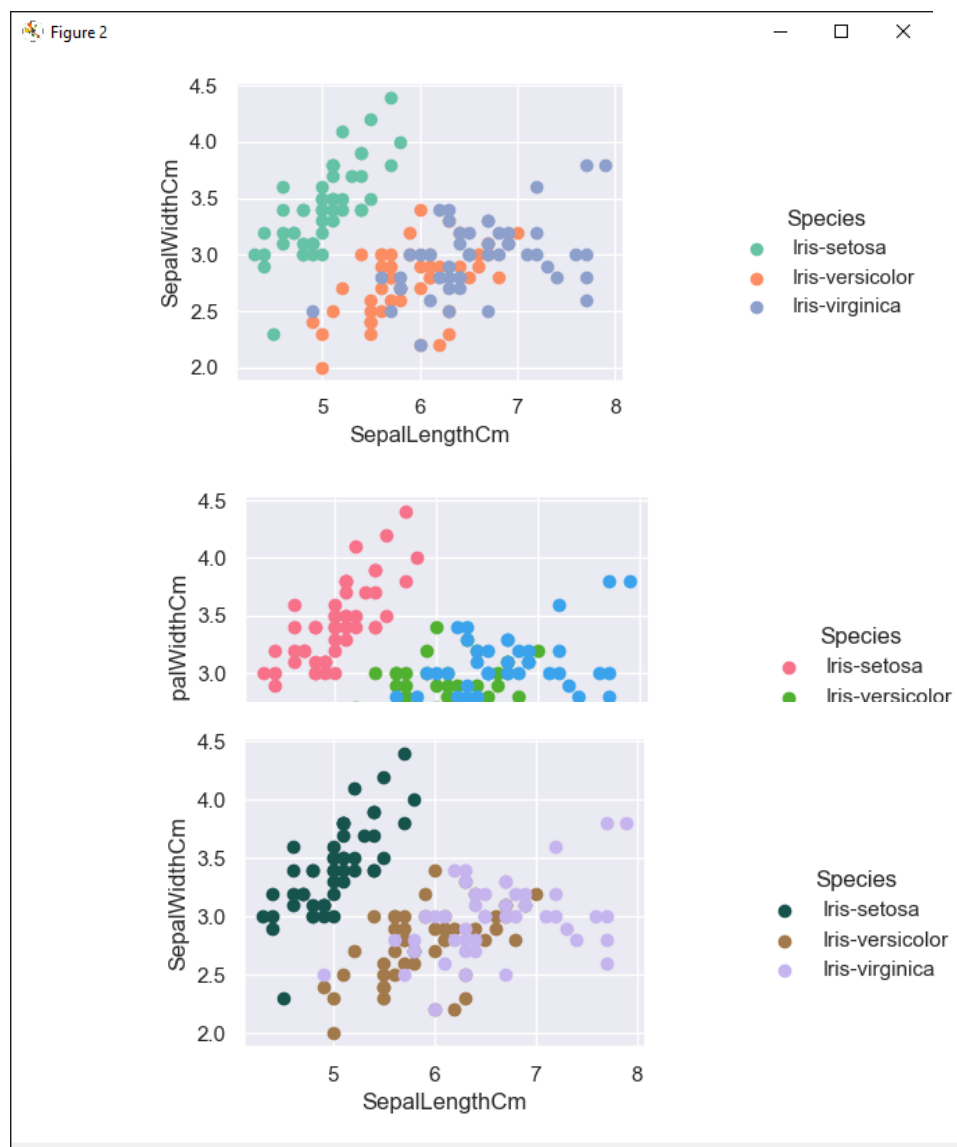
- e. Tiga plot dengan jenis palette atau warna yang berbeda, code program :

```
sns.FacetGrid(df, hue="Species", palette="husl").map(plt.scatter,  
"SepalLengthCm", "SepalWidthCm").add_legend()  
plt.show()
```

```
sns.FacetGrid(df, hue="Species", palette="Set2").map(plt.scatter,  
"SepalLengthCm", "SepalWidthCm").add_legend()  
plt.show()
```

```
sns.FacetGrid(df,hue="Species",palette="cubehelix").map(plt.scatter,  
"SepalLengthCm", "SepalWidthCm").add_legend()  
plt.show()
```

Output :



- f. Pairplot, gambar dibawah ada kolom id yang dihapus karena memiliki korelasi dengan variabel lain, code program :

```
df = pd.read_csv('iris.csv', delim_whitespace=True)
sns.pairplot(df.drop("Id", axis=1), hue="Species", palette="Set2",
size=3)
plt.show()
```

Output :

