STANDARDIZATION & NORMALIZATION

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

Name:ULLI VENKATA SAI KUMAR

Batch Id: 04072024HYD10AM

Topic: Data Pre-Processing

Problem Statement:

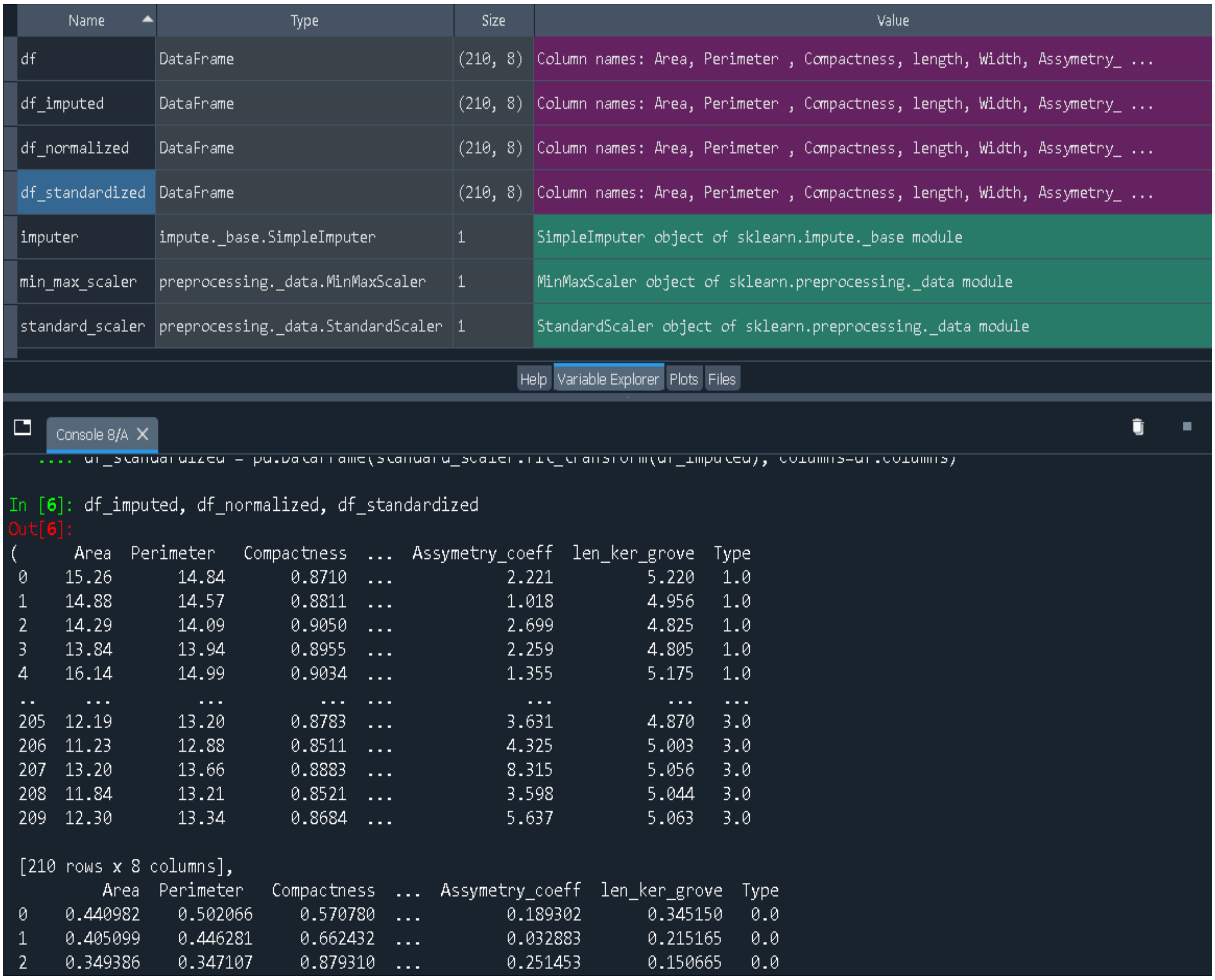
Data is one of the most important assets. Often the data are stored in distinct systems with different formats and scales. These seemingly small differences in how the data is stored can result in misinterpretations and inconsistencies in your analytics. Inconsistency can make it impossible to deliver reliable information to management for good decision-making. We have the preprocessing techniques to make the data uniform. To explore the various techniques to have reliable uniform standard data, you can go through this link:

https://360digitmg.com/mindmap-data-science

1) Prepare the dataset by performing the preprocessing techniques, to have the standard scale to data.

```
import pandas as pd

from sklearn.preprocessing import MinMaxScaler, StandardScaler

from sklearn.impute import SimpleImputer


# Assuming the dataset is already loaded in df

df = pd.read_csv(r"Seeds_data.csv")  # Replace with your actual dataset file


# 1. Identify and Handle Missing Data (if any)

imputer = SimpleImputer(strategy='mean')

df_imputed = pd.DataFrame(imputer.fit_transform(df), columns=df.columns)

# 2. Normalization (Min-Max Scaling)

min_max_scaler = MinMaxScaler()
```

```
df_normalized = pd.DataFrame(min_max_scaler.fit_transform(df_imputed), columns=df.columns)

# 3. Standardization (Z-Score Normalization)

standard_scaler = StandardScaler()

df_standardized = pd.DataFrame(standard_scaler.fit_transform(df_imputed), columns=df.columns)

# Output the preprocessed datasets

df_imputed, df_normalized, df_standardized
```



Name	Type	Size	Value
df	DataFrame	(210, 8)	Column names: Area, Perimeter , Compactness, length, Width, Assymetry_ ...
df_imputed	DataFrame	(210, 8)	Column names: Area, Perimeter , Compactness, length, Width, Assymetry_ ...
df_normalized	DataFrame	(210, 8)	Column names: Area, Perimeter , Compactness, length, Width, Assymetry_ ...
df_standardized	DataFrame	(210, 8)	Column names: Area, Perimeter , Compactness, length, Width, Assymetry_ ...
imputer	impute._base.SimpleImputer	1	SimpleImputer object of sklearn.impute._base module
min_max_scaler	preprocessing._data.MinMaxScaler	1	MinMaxScaler object of sklearn.preprocessing._data module
standard_scaler	preprocessing._data.StandardScaler	1	StandardScaler object of sklearn.preprocessing._data module

```
   ....: df_standardized = pd.DataFrame(standard_scaler.fit_transform(df_imputed), columns=df.columns)

In [6]: df_imputed, df_normalized, df_standardized
Out[6]:
(      Area  Perimeter  Compactness  ...  Assymetry_coeff  len_ker_grove  Type
 0    15.26      14.84       0.8710  ...            2.221          5.220   1.0
 1    14.88      14.57       0.8811  ...            1.018          4.956   1.0
 2    14.29      14.09       0.9050  ...            2.699          4.825   1.0
 3    13.84      13.94       0.8955  ...            2.259          4.805   1.0
 4    16.14      14.99       0.9034  ...            1.355          5.175   1.0
 ..     ...        ...          ...  ...              ...            ...   ...
 205  12.19      13.20       0.8783  ...            3.631          4.870   3.0
 206  11.23      12.88       0.8511  ...            4.325          5.003   3.0
 207  13.20      13.66       0.8883  ...            8.315          5.056   3.0
 208  11.84      13.21       0.8521  ...            3.598          5.044   3.0
 209  12.30      13.34       0.8684  ...            5.637          5.063   3.0

 [210 rows x 8 columns],
           Area  Perimeter  Compactness  ...  Assymetry_coeff  len_ker_grove  Type
 0    0.440982   0.502066     0.570780  ...         0.189302       0.345150   0.0
 1    0.405099   0.446281     0.662432  ...         0.032883       0.215165   0.0
 2    0.349386   0.347107     0.879310  ...         0.251453       0.150665   0.0
```