**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: Vishvash C**

**Batch Id: 23012024**

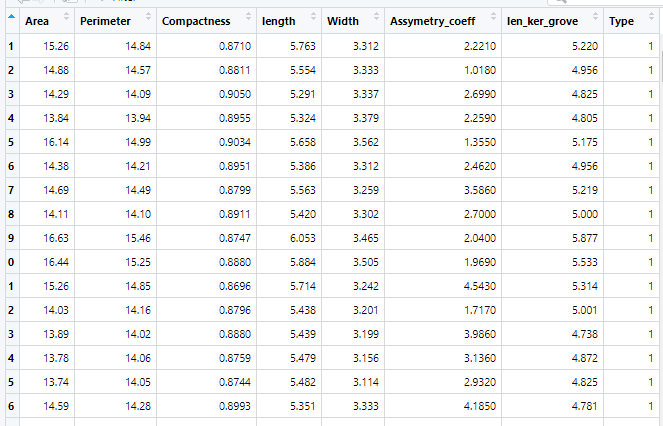
**Topic: Data Pre-Processing**

**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.



**Hints:**

For each assignment, the solution should be submitted in the below format

1. Work on each feature to create a data dictionary as displayed in the image displayed below:
2. Refer to the Seeds\_data.csv file.
3. Research and perform all possible steps for obtaining the solution.
4. All the codes (executable programs) should execute without errors.
5. Code modularization should be followed.
6. Each line of code should have comments explaining the logic and why you are using that function.

import pandas as pd

import numpy as np

import seaborn as sns

data = pd.read\_csv(r"C:/Users/Lenovo/Downloads/Study material/EDA/InClass\_DataPreprocessing\_datasets/Seeds\_data.csv")

a = data.describe()

print(a)

### Standardization

from sklearn.preprocessing import StandardScaler

# Initialise the Scaler

scaler = StandardScaler()

# To scale data

df = scaler.fit\_transform(data)

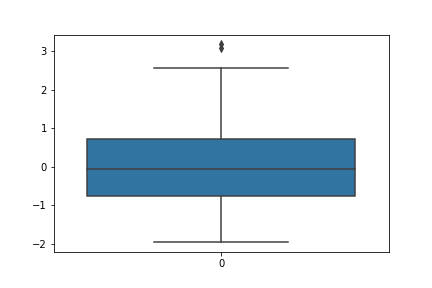
# Convert the array back to a dataframe

dataset = pd.DataFrame(df, columns=data.columns)

res = dataset.describe()

print(res)

sns.boxplot(dataset.Assymetry\_coeff)



'''Robust Scaling- robust to outliers'''

from sklearn.preprocessing import RobustScaler

robust\_model = RobustScaler()

df\_robust = robust\_model.fit\_transform(data)

dataset\_robust = pd.DataFrame(df\_robust, columns=data.columns)

res\_robust = dataset\_robust.describe()

print(res\_robust)

# Normalization

from sklearn.preprocessing import MinMaxScaler

minmaxscale = MinMaxScaler()

df\_n = minmaxscale.fit\_transform(df)

dataset1 = pd.DataFrame(df\_n)

res1 = dataset1.describe()

print(res1)

**Output:**

Area Perimeter Compactness length Width \

count 210.000000 210.000000 210.000000 210.000000 210.000000

mean 14.847524 14.559286 0.870999 5.628533 3.259081

std 2.909699 1.305959 0.023629 0.443063 0.377088

min 10.590000 12.410000 0.808100 4.899000 2.630000

25% 12.270000 13.450000 0.856900 5.262250 2.944000

50% 14.355000 14.320000 0.873450 5.523500 3.237000

75% 17.305000 15.715000 0.887775 5.979750 3.561750

max 21.180000 17.250000 0.918300 6.675000 4.033000

Assymetry\_coeff len\_ker\_grove Type

count 210.000000 210.000000 210.000000

mean 3.700201 5.408071 2.000000

std 1.503557 0.491480 0.818448

min 0.765100 4.519000 1.000000

25% 2.561500 5.045000 1.000000

50% 3.599000 5.223000 2.000000

75% 4.768750 5.877000 3.000000

max 8.456000 6.550000 3.000000

Area Perimeter Compactness length Width \

count 2.100000e+02 2.100000e+02 2.100000e+02 2.100000e+02 2.100000e+02

mean -3.383537e-16 1.184238e-16 1.285744e-15 -2.097793e-15 1.319579e-15

std 1.002389e+00 1.002389e+00 1.002389e+00 1.002389e+00 1.002389e+00

min -1.466714e+00 -1.649686e+00 -2.668236e+00 -1.650501e+00 -1.672245e+00

25% -8.879552e-01 -8.514330e-01 -5.980791e-01 -8.286816e-01 -8.375592e-01

50% -1.696741e-01 -1.836639e-01 1.039927e-01 -2.376280e-01 -5.869636e-02

75% 8.465989e-01 8.870693e-01 7.116771e-01 7.945947e-01 8.045655e-01

max 2.181534e+00 2.065260e+00 2.006586e+00 2.367533e+00 2.057259e+00

Assymetry\_coeff len\_ker\_grove Type

count 2.100000e+02 2.100000e+02 210.000000

mean -3.383537e-16 -8.987520e-17 0.000000

std 1.002389e+00 1.002389e+00 1.002389

min -1.956769e+00 -1.813288e+00 -1.224745

25% -7.591477e-01 -7.404953e-01 -1.224745

50% -6.746852e-02 -3.774588e-01 0.000000

75% 7.123789e-01 9.563941e-01 1.224745

max 3.170590e+00 2.328998e+00 1.224745

Area Perimeter Compactness length Width \

count 2.100000e+02 2.100000e+02 2.100000e+02 2.100000e+02 210.000000

mean 9.782002e-02 1.056449e-01 -7.939850e-02 1.463879e-01 0.035744

std 5.778946e-01 5.765822e-01 7.653252e-01 6.175101e-01 0.610422

min -7.477656e-01 -8.432671e-01 -2.116599e+00 -8.703833e-01 -0.982598

25% -4.141013e-01 -3.841060e-01 -5.360324e-01 -3.641115e-01 -0.474302

50% -1.765081e-16 -3.920475e-16 -1.797824e-15 -6.192963e-16 0.000000

75% 5.858987e-01 6.158940e-01 4.639676e-01 6.358885e-01 0.525698

max 1.355511e+00 1.293598e+00 1.452632e+00 1.604878e+00 1.288547

Assymetry\_coeff len\_ker\_grove Type

count 2.100000e+02 2.100000e+02 210.000000

mean 4.584934e-02 2.224416e-01 0.000000

std 6.811902e-01 5.907218e-01 0.409224

min -1.283905e+00 -8.461538e-01 -0.500000

25% -4.700419e-01 -2.139423e-01 -0.500000

50% -1.005598e-16 -5.336443e-16 0.000000

75% 5.299581e-01 7.860577e-01 0.500000

max 2.200476e+00 1.594952e+00 0.500000

0 1 2 3 4 5 \

count 210.000000 210.000000 210.000000 210.000000 210.000000 210.000000

mean 0.402032 0.444067 0.570767 0.410773 0.448383 0.381633

std 0.274759 0.269826 0.214423 0.249473 0.268773 0.195498

min 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000

25% 0.158640 0.214876 0.442831 0.204533 0.223806 0.233575

50% 0.355524 0.394628 0.593013 0.351633 0.432644 0.368474

75% 0.634089 0.682851 0.723004 0.608530 0.664113 0.520570

max 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000

6 7

count 210.000000 210.000000

mean 0.437751 0.500000

std 0.241989 0.409224

min 0.000000 0.000000

25% 0.258986 0.000000

50% 0.346627 0.500000

75% 0.668636 1.000000

max 1.000000 1.000000

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| **ID** |  | **Quantitative/ Nominal** | **Irrelevant (ID does not provide useful information)** |
| Area | area of the kernel | Quantitative | Relevant |
| Perimeter | perimeter of the kernel | Quantitative | Relevant |
| Compactness | compactness of the kernel | Quantitative | Relevant |
| Length | length of the kernel | Quantitative | Relevant |
| Width | width of the kernel | Quantitative | Relevant |
| Assymetry\_coeff | asymmetry coefficient of the kernel | Quantitative | Relevant |
| len\_ker\_grove | length of kernel groove | Quantitative | Relevant |
| Type | type of wheat | Nominal | Relevant |