Data Collection and Preprocessing Phase

Date	03 June2024
Team ID	739676
Project Title	Harvesting Brilliance: A Taxanomic Tale of Pumpkin Seeds Varieties
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	This section provides an overview of the pumpkin seed varieties dataset. It includes basic statistics such as the number of varieties, dimensions of the dataset (e.g., number of rows and columns), and the general structure of the data (e.g., types of variables, data types)
Univariate Analysis	This section focuses on analyzing individual variables within the pumpkin seed varieties dataset. It involves calculating and interpreting descriptive statistics like mean, median, mode, and standard deviation for each variable.
Bivariate Analysis	This section examines the relationships between two variables in the pumpkin seed varieties dataset. It includes techniques like correlation analysis and scatter plots to understand how different variables interact with each other.
Multivariate Analysis	This section investigates patterns and relationships involving multiple variables simultaneously. It involves more complex statistical methods to understand how different variables collectively influence certain outcomes.
Outliers and Anomalies	This section focuses on identifying and treating outliers and anomalies within the pumpkin seed varieties dataset. Outliers are data points that deviate significantly from the rest of the data, which can affect the analysis.

Data Preprocessing Code Screenshots

| Loading Data | ```python
df=pd.read_csv("Pumpkin_Seeds_Dataset.xlsx - Pumpkin_Seeds_Dataset.csv",sep=",")
df
```<br><br>A dataframe showing columns: Area, Perimeter, Major_Axis_Length, Minor_Axis_Length, Convex_Area, Equiv_Diameter, Eccentricity, Solidity, Extent, Roundness, Aspect_Ration, Compactness, Class<br><br>0  56276  888.242  326.1485  220.2388  56831  267.6805  0.7376  0.9902  0.7453  0.8963  1.4809  0.8207  Çerçevelik<br>1  76631  1068.146  417.1932  234.2289  77280  312.3614  0.8275  0.9916  0.7151  0.8440  1.7811  0.7487  Çerçevelik<br>2  71623  1082.987  435.8328  211.0457  72663  301.9822  0.8749  0.9857  0.7400  0.7674  2.0651  0.6929  Çerçevelik<br>3  66458  992.051  381.5638  222.5322  67118  290.8899  0.8123  0.9902  0.7396  0.8486  1.7146  0.7624  Çerçevelik<br>4  66107  998.146  383.8883  220.4545  67117  290.1207  0.8187  0.9850  0.6752  0.8338  1.7413  0.7557  Çerçevelik<br>...<br>2495  79637  1224.710  533.1513  190.4367  80381  318.4289  0.9340  0.9907  0.4888  0.6672  2.7996  0.5973  Ürgüp Sivrisi<br>2496  69647  1084.318  462.9416  191.8210  70216  297.7874  0.9101  0.9919  0.6002  0.7444  2.4134  0.6433  Ürgüp Sivrisi<br>2497  87994  1210.314  507.2200  222.1872  88702  334.7199  0.8990  0.9920  0.7643  0.7549  2.2828  0.6599  Ürgüp Sivrisi<br>2498  80011  1182.947  501.9065  204.7531  80902  319.1758  0.9130  0.9890  0.7374  0.7185  2.4513  0.6359  Ürgüp Sivrisi<br>2499  84934  1159.933  462.8951  234.5597  85781  328.8485  0.8621  0.9901  0.7360  0.7933  1.9735  0.7104  Ürgüp Sivrisi<br><br>2500 rows × 13 columns |
| Handling Missing Data | ```python
df.info()
```<br><br>```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2500 entries, 0 to 2499
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Area               2500 non-null   int64
 1   Perimeter          2500 non-null   float64
 2   Major_Axis_Length  2500 non-null   float64
 3   Minor_Axis_Length  2500 non-null   float64
 4   Convex_Area        2500 non-null   int64
 5   Equiv_Diameter     2500 non-null   float64
 6   Eccentricity       2500 non-null   float64
 7   Solidity           2500 non-null   float64
 8   Extent             2500 non-null   float64
 9   Roundness          2500 non-null   float64
 10  Aspect_Ration      2500 non-null   float64
 11  Compactness        2500 non-null   float64
 12  Class              2500 non-null   object
dtypes: float64(10), int64(2), object(1)
memory usage: 254.0+ KB
``` |