DS5230 Final: Phase 2

Preprocessing Pipeline & EDA

Michael Massone and Nelson Farrell

Spring 2024

Data Overview:
Dry Bean Dataset

- Size: (13611 x 17) – 13611 rows, 17 columns
- This dataset is composed of data derived from 13,611 images of 7 species of beans. Bean images obtained by computer vision system were subjected to segmentation and feature extraction stages, and a total of 16 features; 12 dimensions and 4 shape forms, were generated.
- Koklu, M. and Ozkan, I.A., (2020), “Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques.” Computers and Electronics in Agriculture, 174, 105507.
- DOI: https://doi.org/10.1016/j.compag.2020.105507
- Link: https://archive.ics.uci.edu/dataset/602/dry+bean+dataset

Preprocessing

Missingness:

- Our first step will be to identify any attributes in the design matrix with a missingness proportion above **0.20**. These will be added to a ***missingness_drop_list*** and excluded from the machine learning pipeline.

Nominal & Numerical Attributes:

- Next we will identify the nominal and numerical attributes within the design matrix. This is necessary because the pipeline procedures differ between the types.

Unique Values:

- Next we will flag all attributes with a unique proportion equal to 1 for further evaluation. The motivation being that these may in fact be an identifier column.

Results

Missingness:

- *None*

Unique Value Attributes:

- *Unnamed: 0*
- *id*

Nominal Attributes:

- *None*

Numerical Attributes:

- *Unnamed: 0*
- *Id*
- *Area*
- *Perimeter*
- *MajorAxisLenght*
- *MinorAxisLenght*
- *AspectRation*
- *Eccentricity*
- *ConvexArea*
- *EquiDiameter*
- *Solidity*
- *Roundness*
- *Compactness*
- *ShapeFactor1*
- *ShapeFactor2*
- *ShapeFactor3*
- *ShapeFactor4*

With these results we generated refined lists and identified ML attributes

Results conti...

Drop List

- This list of attributes was dropped from the dataframe.
- It contained only 1 attribute
- **Unnamed: 0**
- This attribute appeared to be a duplicate identifier column

Non-ML Attributes

- *These are attributes that will not be used in the machine learning model*
- ***id***

Nominal Attributes:

- ***None***

Numerical Attributes:

- ***ConvexArea***
- ***EquiDiameter***
- ***Solidity***
- ***Roundness***
- ***Compactness***
- ***ShapeFactor1***
- ***ShapeFactor2***
- ***ShapeFactor3***
- ***ShapeFactor4***
- ***Area***
- ***Perimeter***
- ***MajorAxisLenght***
- ***MinorAxisLenght***
- ***AspectRation***
- ***Eccentricity***

Finalized Lists:

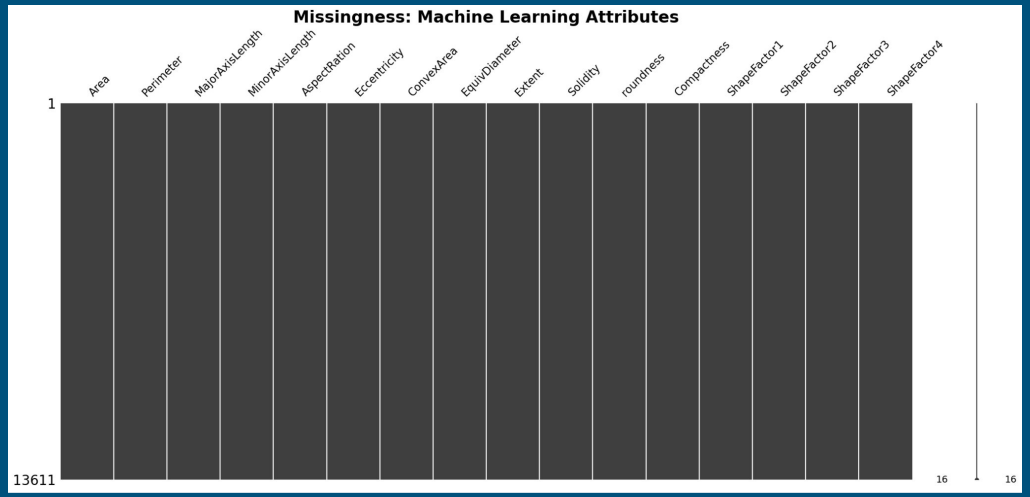
ML Ignore List:

- Attributes that will not be used in modeling
- **ml_ignore_list** = [**non_ml_attr** + **missingness_list**]

ML Attributes:

- These are attributes that will be used in the machine learning model
- Nominal attributes
  - ***nominal_cols***
- Numerical Attributes
  - ***numerical_cols***

Missingness Machine Learning Attributes



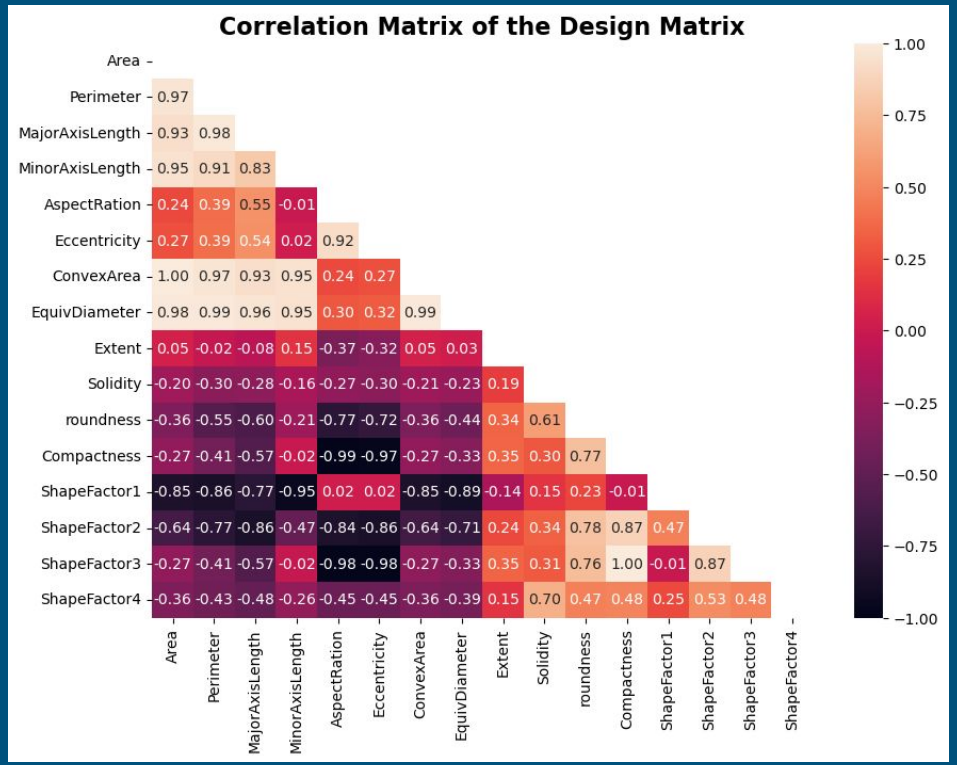
Having established the attributes that will be used in the machine learning model we examined their missingness. As this graphic demonstrates, we have zero missing values in the features that will be used in machine learning.

Feature Descriptions

- `Area` (A) - ***float64 - Ratio*** – The area of a bean zone and the number of pixels within its boundaries.
- `Perimeter` (P) - ***float64 - Ratio*** – Bean circumference is defined as the length of its border.
- `MajorAxisLength` (L) - ***float64 - Ratio*** – The distance between the ends of the longest line that can be drawn from a bean.
- `MinorAxisLength` (l) - ***float64 - Ratio*** – The longest line that can be drawn from the bean while standing perpendicular to the main axis.
- `AspectRatio` (K) - ***float64 - Ratio*** – Defines the relationship between L and l.
- `Eccentricity` (Ec) - ***float64 - Ratio*** – Eccentricity of the ellipse having the same moments as the region.
- `ConvexArea` (C) - ***float64 - Ratio*** – Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
- `EquivDiameter` (Ed) - ***float64 -*** Ratio – The diameter of a circle having the same area as a bean seed area.
- `Extent` (Ex) - ***float64 - Ratio -*** The ratio of the pixels in the bounding box to the bean area.
- `Solidity` (S) - ***float64 - Ratio -*** Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
- `Roundness` (R) - ***float64 - Ratio -*** Calculated with the following formula: $\frac{(4\pi A)}{(P^2)}$
- `Compactness` (CO) - ***float64 - Ratio*** – Measures the roundness of an object: $\frac{Ed}{L}$
- `ShapeFactor1` (SF1) - ***float64 - Ratio*** - $\frac{L}{A}$
- `ShapeFactor2` (SF2) - ***float64 - Ratio*** - $\frac{l}{A}$
- `ShapeFactor3` (SF3) - ***float64 - Ratio*** - $\frac{A}{\frac{L}{2} \cdot \frac{L}{2} \cdot \pi}$
- `ShapeFactor4` (SF4) - ***float64 - Ratio*** $\frac{A}{\frac{L}{2} \cdot \frac{l}{2} \cdot \pi}$

Collinearity

Computing the a correlation matrix for the numerical attributes shows a very high degree of both positive and negative correlation. This is to be expected based on the attribute descriptions from the previous slides. Many of the attributes are functions of the others or are related metrics.



Variance Inflation Factor

VIF provides a better measure of collinearity. The correlation matrix depicts bivariate relationships between attributes. The VIF is a measure of how each attribute impacts the regression coefficients of the other variables in the design matrix.

However, since our values are very highly correlated, this does not provide much additional information. We are not concerned with eliminating collinear attributes. If we were, using VIF would allow us to see the effect of removing the most collinear attributes.

Variance Inflation Factors Above Threshold(5):

	Variable	VIF
0	Area	81390.772240
1	Perimeter	3573.980197
2	MajorAxisLength	87936.513102
3	MinorAxisLength	77482.673330
4	AspectRation	13918.856718
5	Eccentricity	1183.004420
6	ConvexArea	78552.516774
7	EquivDiameter	314929.768895
9	Solidity	14.053550
10	roundness	104.028512
11	Compactness	276769.624251
12	ShapeFactor1	607.175671
13	ShapeFactor2	1245.330594
14	ShapeFactor3	200286.806522
15	ShapeFactor4	67.043776

Variance Inflation Factors Below Threshold(5):

	Variable	VIF
8	Extent	1.241536

Pipeline: Sklearn

- With our machine attributes identified and explored, we used **sklearn pipeline** to further process the attributes in preparation for modeling
- Nominal and numerical attributes require different processing, so we set two transformers and put both in the final pipeline.

```
# instantiate numerical transformer
numerical_transformer = Pipeline(
    steps=[
            ('imputer', SimpleImputer()),
            ('scaler', StandardScaler())
        ]
    )
```

```
# instantiate nominal transformer
nominal_transformer = Pipeline(
    steps=[
            ('imputer', SimpleImputer(strategy='most_frequent')),
            ('onehot_encoder', OneHotEncoder(sparse_output = False,
                                             min_frequency=min_frequency))
    ]
)
```

- Both transformers impute missing values, ***numerical*** with the ***mean***, and ***nominal*** with the ***most frequent***.
- Numerical transformer standardizes the values using: ***z = x - mean / standard deviation***
- Nominal transformer uses ***OneHotEncoder*** to expand the attribute into binary form (a column for each label) and true/false indicators.

Pipeline: Sklearn

```
# instantiate pipeline
preprocessor = ColumnTransformer(
        transformers= [
            ('numerical', numerical_transformer, numerical_cols),
            ('nominal', nominal_transformer, nominal_cols)
        ]
)
```

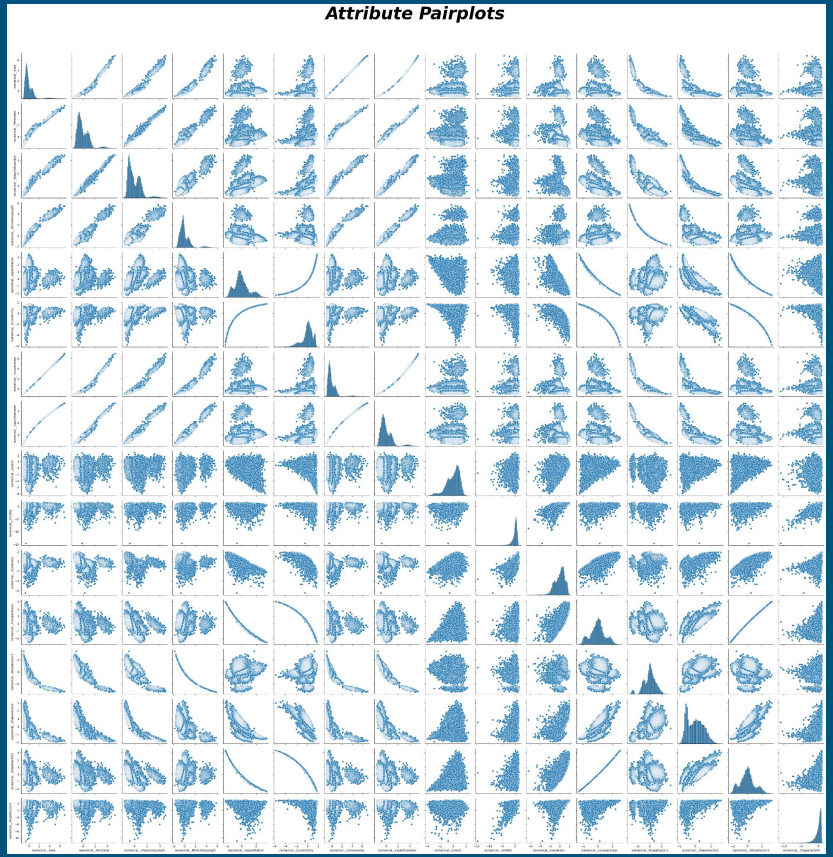
The finalized pipeline used to process and transform the machine learning attributes is seen above.

Numerical Attribute Pairplots

256 pairplots (including 16 histograms)

Findings:

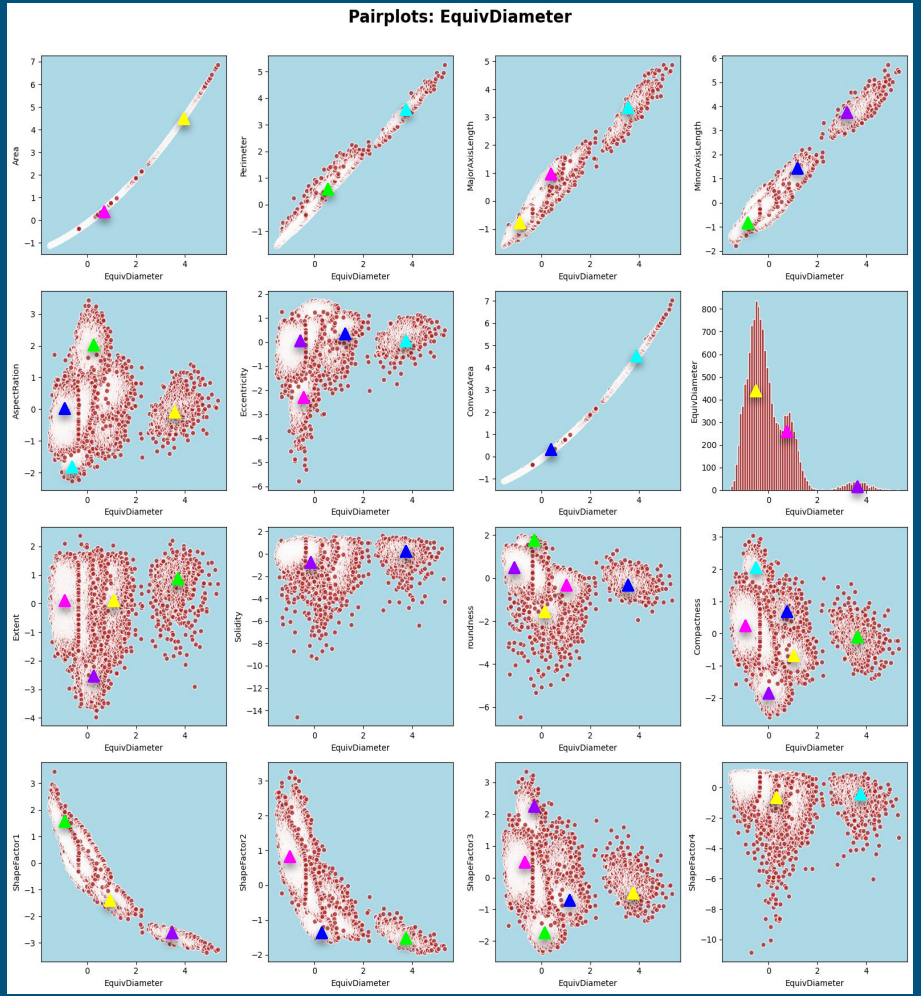
- Some attributes are functions of other attributes. We knew this from the attribute descriptions, and saw this effect in the correlation matrix and VIF scores. This results in linear or quadratic shaped distributions.
- There are many pairs plots that contain distinct cluster. These are easier to see on in the sub-pairplots. We have presented some of these on the next few slides to discuss potential clusters. However there were enough attributes with potential clusters that we will not be discussing all of them.



EquivDiamter

A closer look - example:

- Potential clusters have been labeled with the triangles - color in this case has no particular meaning. In most cases the labels represent best guesses, since some cluster are distinct while others appear to overlap, or may be a single continuous, irregular cluster.
- The top row of pairplots are all roughly linear. Diameter is directly related to perimeter, area, and the axis length of the images. We can however see two possible clusters in the perimeter, major axis and minor axis.
- The histogram also shows a trimodal distribution.
- The other pairplots all show some degree of clustering with both globular and what appear to be some overlapping cluster that might be discernible with additional processing. There does seem to be at least two very distinct custer in all the pairplots. The larger than the other with an irregular shape that seems to indicate the presence of overlapping clusters.



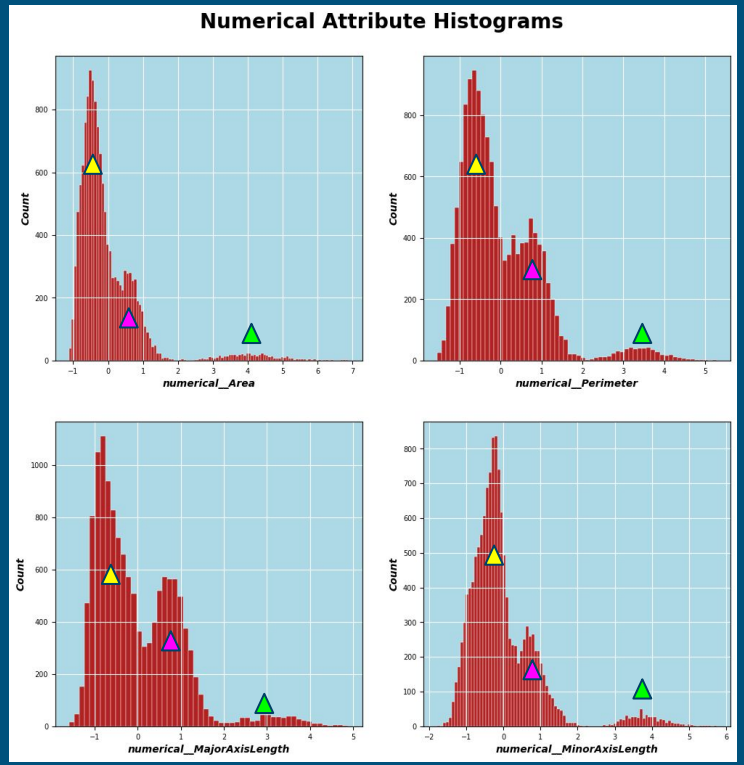
Attributes that exhibit clustering:

Numerous attributes exhibited some degree of clustering on the pairplots - we will not show or discuss all of the them. However, the attributes that showed the most potentials clusters are indicated in the table below.

Attribute	Best Clusters when plotted against:
Area	MinorAxisLength, AspectRatio, Eccentricity, Extent, Solidity, roundness, compactness, SF3
Perimeter	MinorAxisLength, AspectRatio, Eccentricity, Extent, Solidity, roundness, compactness, SF3
MajorAxis Length	MinorAxisLength, AspectRatio, Eccentricity, Compactness, SF3
MinorAxis Length	Area, Perimeter, Major Axis Length, AspectRatio, Eccentricity, Convex Area, EquivDistance, SF2, SF3
AspectRatio	Area, Perimeter, MajorAxisLength, MinorAxisLength, AspectRatio, ConvexArea, EquivDiameter, SF1, SF2
Eccentricity	Area, Perimeter, MajorAxisLength, MinorAxisLength, ConvexArea, EquivDistance, SF1
ConvexArea	MinorAxisLength, AspectRatio, Eccentricity, EquivDistance, Roundness, SF3
Equiv Diameter	Aspect Ratio, Eccentricity, Compactness, SF1, SF3

Attribute	Best Clusters when plotted against:
Extent	None
Solidity	Area, MinorAxisLength, ConvexArea, EquivDiameter
Roundness	Area, Perimeter, MinorAxisLength,Convex Area, EquivDistance
Compactness	Area, Perimeter, Major Axis Length, Minor Axis Length, Convex Area, EquivDistnace, Shapefacter 1
SF1	AspectRatio, Eccentricity, Compactness, SF2, SF3
SF2	Minor Axis Length, ConvexArea, EquivDiameter, SF1
SF3	Area, MajorAxisLength, MinorAxi Length, ConvexArea, EquivDiameter, SF1
SF4	None

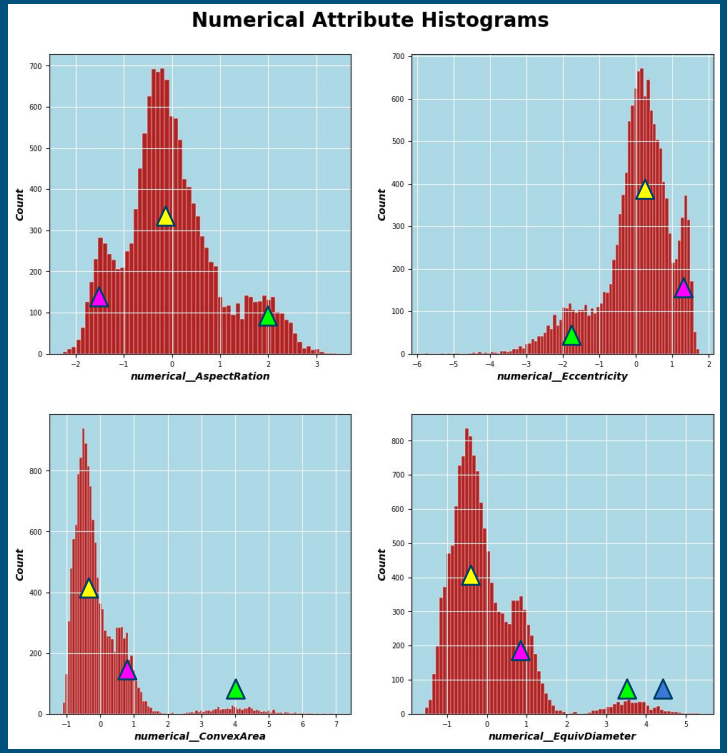
Histograms of Numerical Attributes:



Potential Clusters:

- All 4 of the histograms here appear to display 3 distinct clusters.
- The 3 clusters in the ***numerical_area*** histogram are perhaps less clear, but nevertheless present.

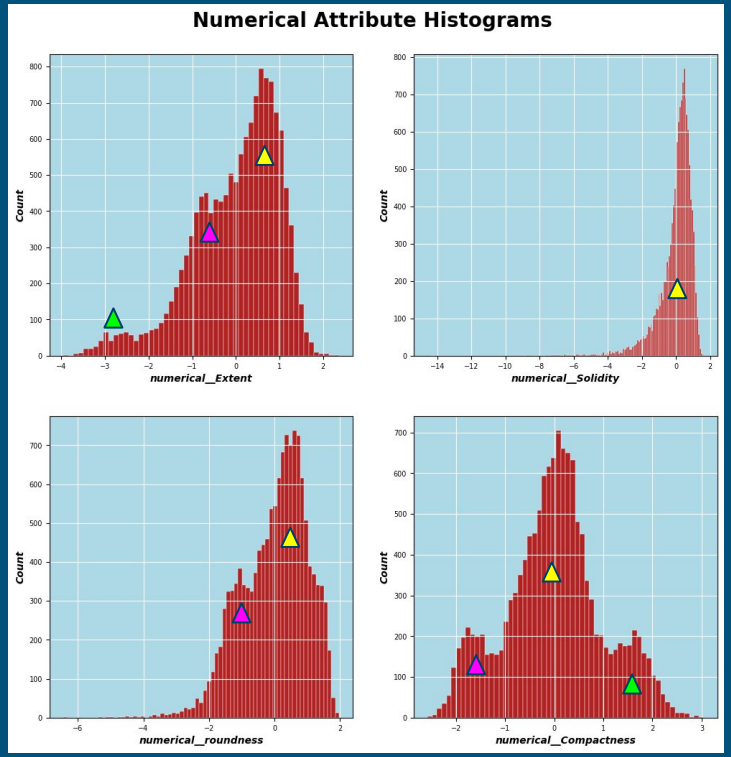
Histograms of Numerical Attributes:



Potential Clusters:

- All 4 of the histograms here appear to display 3 distinct clusters.
- There is perhaps even a fourth cluster in the ***numerical_EquivDiameter*** histogram. This is highlighted with the blue triangle and certainly less distinct.

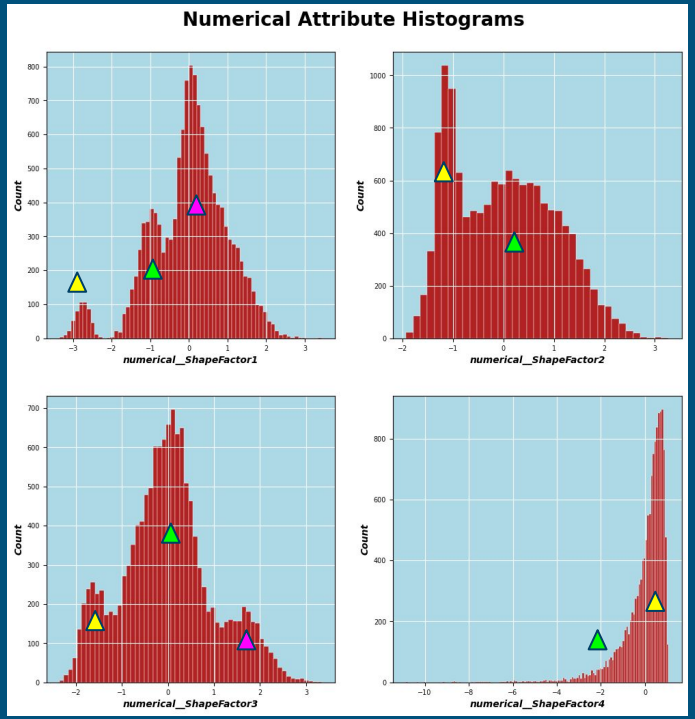
Histograms of Numerical Attributes:



Potential Clusters:

- ***numerical_Extent*** appears to show 3 relatively distinct clusters, although not as clearly defined as the previous histograms.
- ***numerical_roundess*** appears to only show two well-defined clusters.
- ***numerical_Solidity*** appears to show only a single cluster.
- ***numerical_Compactness*** appears to show 3 distinct and clearly defined clusters.

Histograms of Numerical Attributes:



Potential Clusters:

- ***numerical_ShapeFactor1*** appears to show 3 distinct and well-defined clusters.
- ***numerical_ShapeFactor2*** appears to only show two well-defined clusters.
- ***numerical_ShapeFactor3*** appears to show 3 distinct and well-defined clusters.
- ***numerical_ShapeFactor4*** appears to show 2 clusters. That is however, considering the tail elements a cluster, which may or may not be accurate.

Summery

- The histograms of the transformed data appear to display distinct clusters
  - Between 1 and 4
- The pairplots reveal numerous potential clusters that can be further explored. There are also two attributes with minimal clustering that could potentially be removed to aid in dimensionality reduction.
- Overall, there appears to be a high degree of structure in the data.
- Dimensionality reduction may help delineate these potential clusters.
- A successful outcome when performing clustering appears likely.

Acknowledgments

1. KOKLU, M. and OZKAN, I.A., (2020), “Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques.” Computers and Electronics in Agriculture, 174, 105507. DOI: https://doi.org/10.1016/j.compag.2020.105507
2. Dr. Steve Morin – Class slides and labs.
3. UC Irvine Machine Learning Repository
4. Sci-kit Learn Documentation, scikit-learn.org