Introduction to Data Science Supervised learning (Working with data)

Q1) Download data set form kaggel or Google datasets.

Data set from Kaggle website.

Seven different types of dry beans were used in this research, taking into account the features such as form, shape, type, and structure by the market situation. A computer vision system was developed to distinguish seven different registered varieties of dry beans with similar features in order to obtain uniform seed classification. For the classification model, images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera. 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 obtained from the grains.

Attribute Information:

1.) Area (A): The area of a bean zone and the number of pixels within its boundaries.

2.) Perimeter (P): Bean circumference is defined as the length of its border.

3.) Major axis length (L): The distance between the ends of the longest line that can be drawn from a bean.

4.) Minor axis length (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.

5.) Aspect ratio (K): Defines the relationship between L and l.

6.) Eccentricity (Ec): Eccentricity of the ellipse having the same moments as the region.

7.) Convex area (C): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.

8.) Equivalent diameter (Ed): The diameter of a circle having the same area as a bean seed area.

9.) Extent (Ex): The ratio of the pixels in the bounding box to the bean area.

10.)Solidity (S): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.

11.)Roundness (R): Calculated with the following formula: (4piA)/(P^2)

12.)Compactness (CO): Measures the roundness of an object: Ed/L

13.)ShapeFactor1 (SF1)

14.)ShapeFactor2 (SF2)

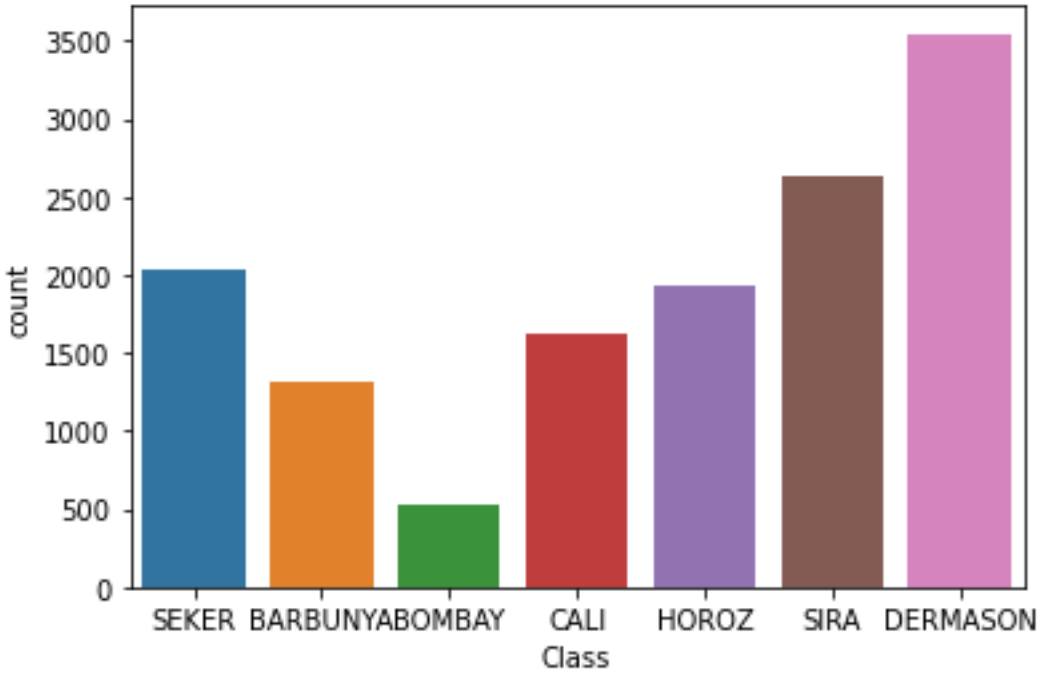
15.)ShapeFactor3 (SF3)

16.)ShapeFactor4 (SF4)

17.)Class (Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira)

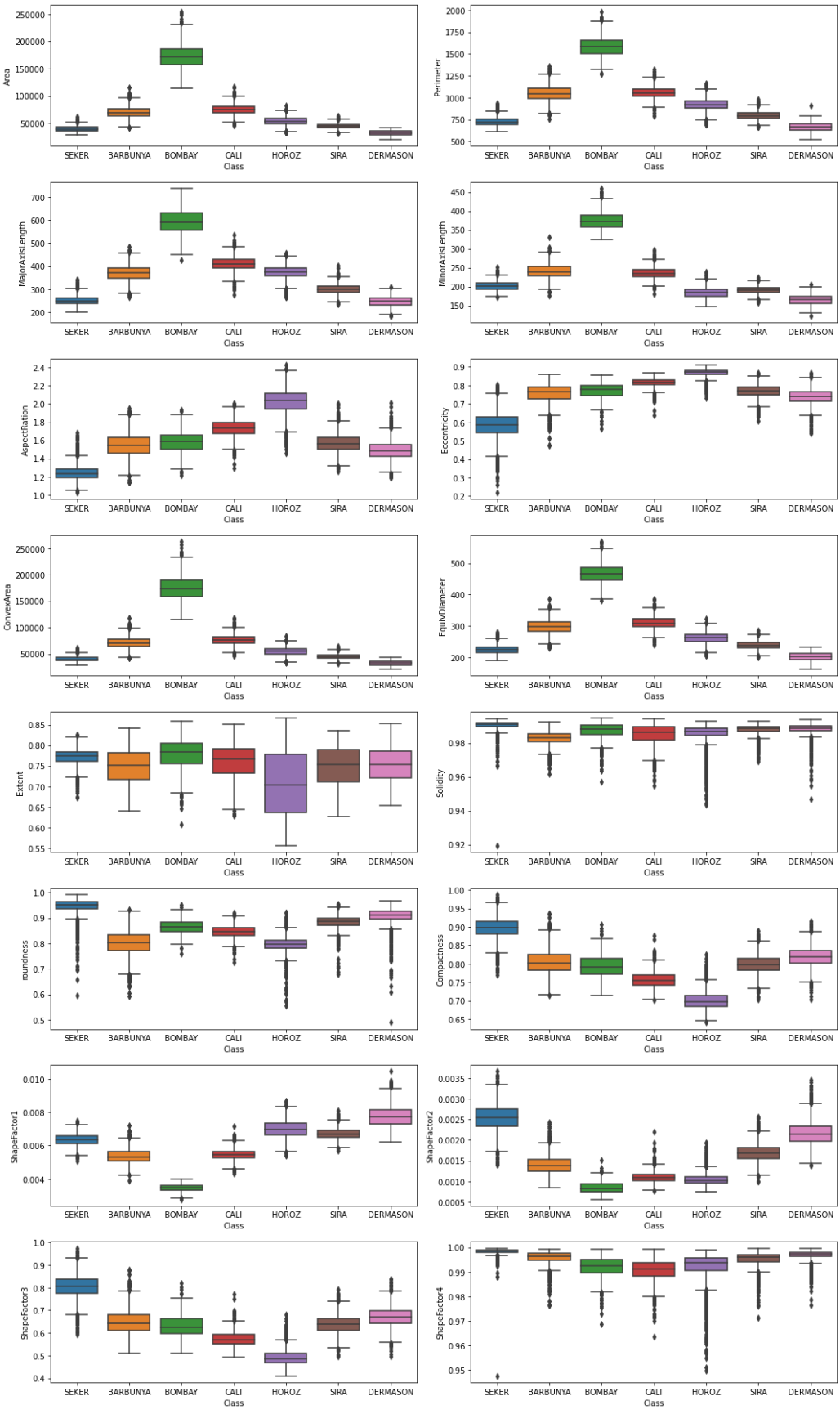
--visualize data using chart:

1)count plot: count the number of record of each class:



From the plot above, we can see that there is variation between the size of classes, where dermason is dominant class.

2) box plot for numeric features:


Area
Perimeter
MajorAxisLength
MinorAxisLength
AspectRation
Eccentricity
ConvexArea
EquivDiameter
Extent
Solidity
roundness
Compactness
ShapeFactor1
ShapeFactor2
ShapeFactor3
ShapeFactor4
SEKER
BARBUNYA
BOMBAY
CALI
HOROZ
SIRA
DERMASON
Class

-- aggregate and summarize dataset:

Using pandas.describe()

	Area	Perimeter	...	ShapeFactor3	ShapeFactor4
count	13611.000000	13611.000000	...	13611.000000	13611.000000
mean	53048.284549	855.283459	...	0.643590	0.995063
std	29324.095717	214.289696	...	0.098996	0.004366
min	20420.000000	524.736000	...	0.410339	0.947687
25%	36328.000000	703.523500	...	0.581359	0.993703
50%	44652.000000	794.941000	...	0.642044	0.996386
75%	61332.000000	977.213000	...	0.696006	0.997883
max	254616.000000	1985.370000	...	0.974767	0.999733

[8 rows x 16 columns]

Q2) Create a list of 1000 random integers between 1 and 100000, then calculate the Z-Score to check for the outliers. Consider values Z-Score > 2 as outliers.

Comment: the range of z-score is between -1 and 1, so there is no outliers.

Q3) Create a list of 10 random integers between 1 and 100.

Comment: the range of min-max is between 0 and 1, where, the range of z-score is between -1 and 1.

Q4) Create a list of 100 random pair of integers (x,y) between 1 and 100. Draw visualization of the data using different.

Scatter plot of x and y: