Wine dataset

Business Objective:- Clustering wine samples according to their various attributes

Constaraints:- Precise data will be required for which a subject matter expert might be needed

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Description | Type | Relevance |
| Type | Type of wine | Categorical,Nominal | Provides type of wine |
| Alcohol | Alcohol content | Quantitative,Nominal | Alcohol content in sample |
| Malic | Malic acid | Quantitative,Nominal | Malic acid proportions in sample |
| Ash | Ash | Quantitative,Nominal | Ash proportions in sample |
| Alcalinity | Alcalinity | Quantitative,Nominal | Alcalinity proportions in sample |
| Magnesium | Magnesium | Quantitative,Nominal | Magnesium proportions in sample |
| Phenols | Phenols | Quantitative,Nominal | Phenols proportions in sample |
| Flavanoids | flavanoids | Quantitative,Nominal | Flavanoids proportions in sample |
| Nonflavanoids | nonflavanoids | Quantitative,Nominal | nonflavanoids  proportions in sample |
| Proanthocyanins | Proanthocyanis | Quantitative,Nominal | Proanthocyanis  proportions in sample |
| Color | color | Quantitative,Nominal | Color proportions in sample |
| Hue | Hue | Quantitative,Nominal | Hue proportions in sample |
| Dilution | Dilution | Quantitative,Nominal | Dilution proportions in sample |
| Proline | Proline | Quantitative,Nominal | Proline proportions in sample |

* Pre-processing,EDA and feature engineering

Data description and moments of business decision

* Mean

Type 1.938202

Alcohol 13.000618

Malic 2.336348

Ash 2.366517

Alcalinity 19.494944

Magnesium 99.741573

Phenols 2.295112

Flavanoids 2.029270

Nonflavanoids 0.361854

Proanthocyanins 1.590899

Color 5.058090

Hue 0.957449

Dilution 2.611685

Proline 746.893258

* Median

Type 2.000

Alcohol 13.050

Malic 1.865

Ash 2.360

Alcalinity 19.500

Magnesium 98.000

Phenols 2.355

Flavanoids 2.135

Nonflavanoids 0.340

Proanthocyanins 1.555

Color 4.690

Hue 0.965

Dilution 2.780

Proline 673.500

* Variance

Type 0.600679

Alcohol 0.659062

Malic 1.248015

Ash 0.075265

Alcalinity 11.152686

Magnesium 203.989335

Phenols 0.391690

Flavanoids 0.997719

Nonflavanoids 0.015489

Proanthocyanins 0.327595

Color 5.374449

Hue 0.052245

Dilution 0.504086

Proline 99166.717355

* Skewness

Type 0.107431

Alcohol -0.051482

Malic 1.039651

Ash -0.176699

Alcalinity 0.213047

Magnesium 1.098191

Phenols 0.086639

Flavanoids 0.025344

Nonflavanoids 0.450151

Proanthocyanins 0.517137

Color 0.868585

Hue 0.021091

Dilution -0.307285

Proline 0.767822

* Kurtosis

Type -1.322787

Alcohol -0.852500

Malic 0.299207

Ash 1.143978

Alcalinity 0.487942

Magnesium 2.104991

Phenols -0.835627

Flavanoids -0.880382

Nonflavanoids -0.637191

Proanthocyanins 0.554649

Color 0.381522

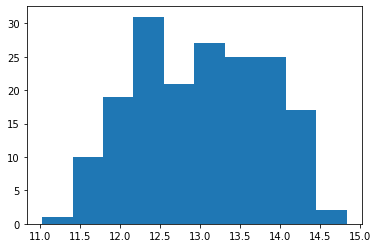
Hue -0.344096

Dilution -1.086435

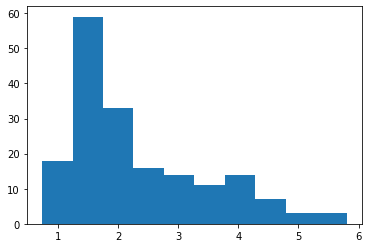
Proline -0.248403

* Data Distribution

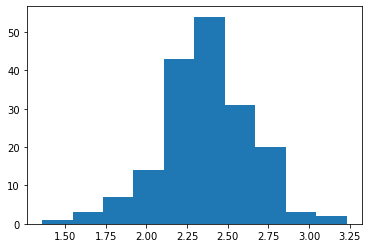
Alcohol



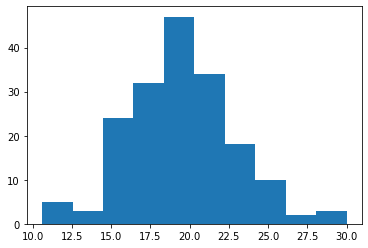
Malic



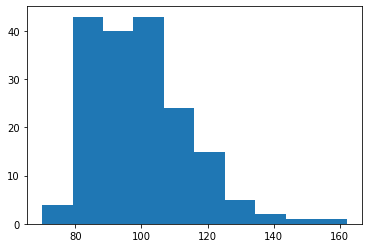
Ash



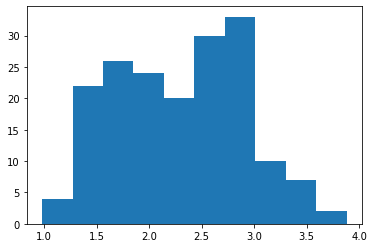
Alcalinity



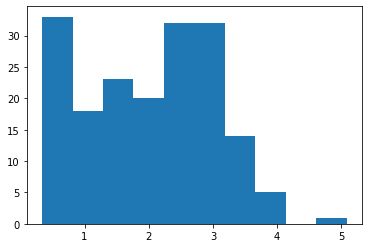
Magnesium



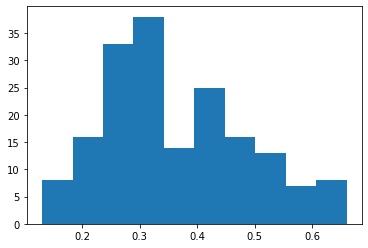
Phenols



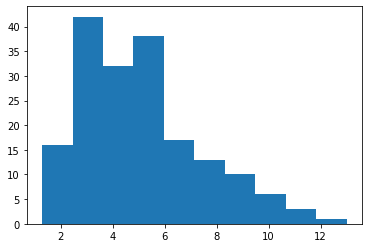
Flavanoids



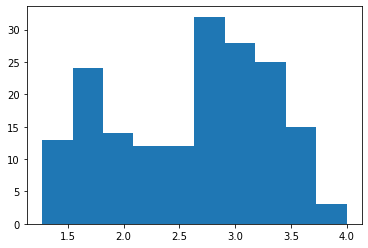
Nonflavanoids



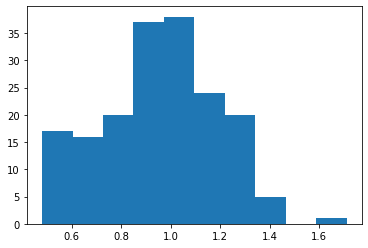
Color



Dilution



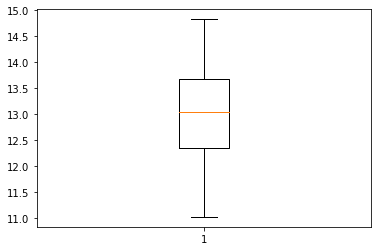
Hue



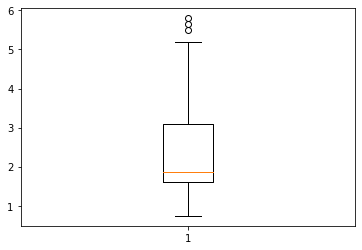
* Outliers treatment

Boxplots were used to identify outliers and features with outliers were winsorized

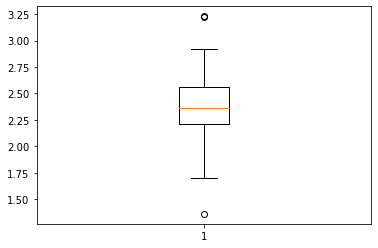
Alcohol



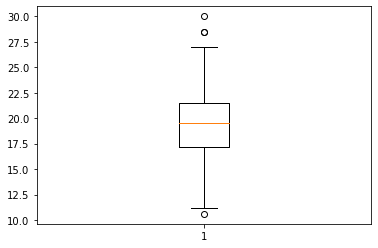
Malic



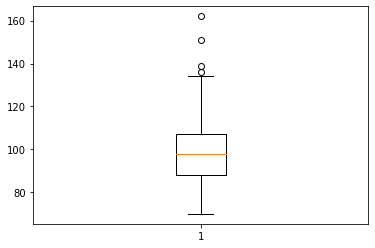
Ash



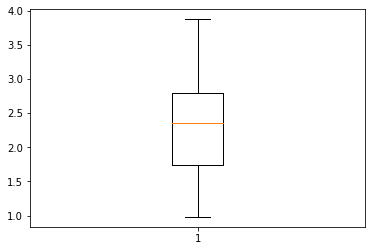
Alcalinity



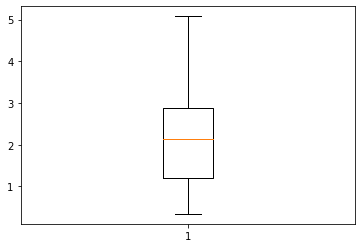
Magnesium



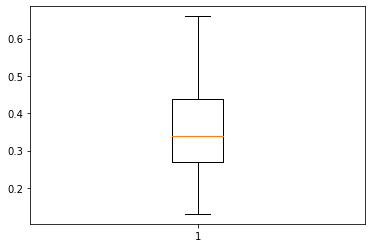
Phenols



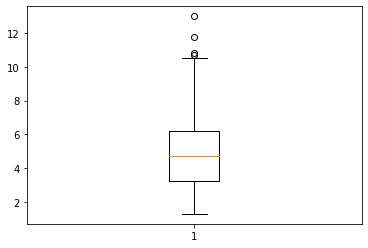
Flavanoids



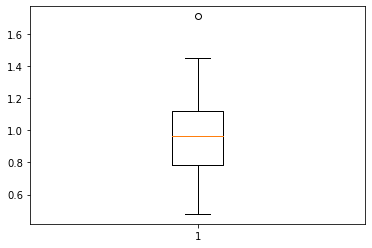
Nonflavanoids



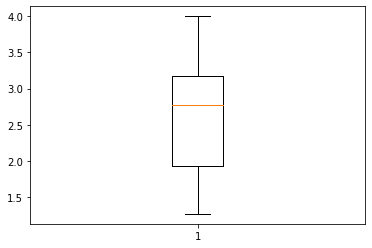
Color



Hue



Dilution



* Normalization

After winsorization data was normalized using custom made normalization fucntion

Description of normalized dataset is as follows

Alcohol Phenols ... Color Hue

count 178.000000 178.000000 ... 178.000000 178.000000

mean 0.518584 0.453487 ... 0.400229 0.416121

std 0.213639 0.215811 ... 0.240775 0.198120

min 0.000000 0.000000 ... 0.000000 0.000000

25% 0.350658 0.262931 ... 0.206603 0.263904

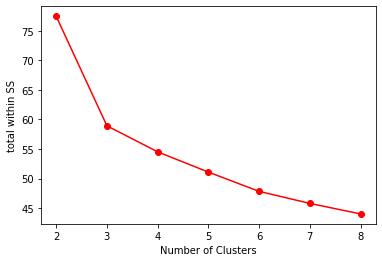
50% 0.531579 0.474138 ... 0.363152 0.423119

75% 0.696711 0.627586 ... 0.523962 0.558342

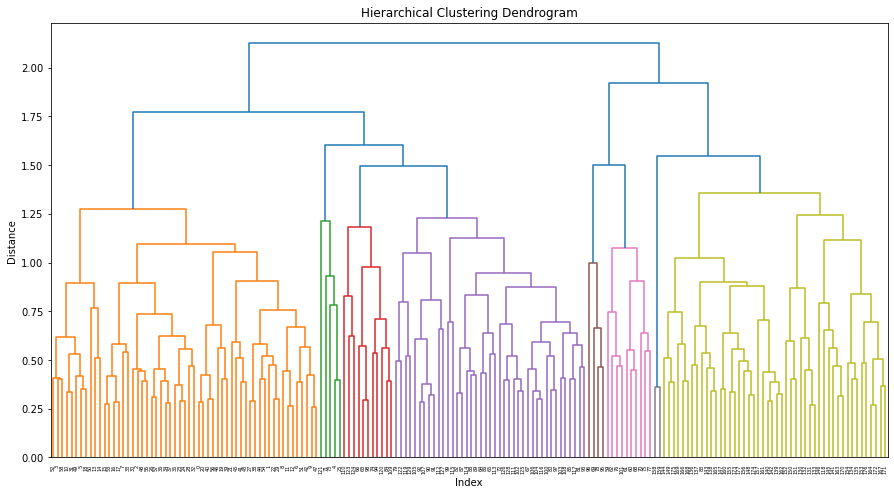
max 1.000000 1.000000 ... 1.000000 1.000000

* K means before clustering

K means was performed on dataset and clusters were stored at ‘clust1k’ feature in database



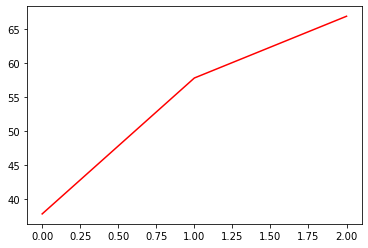
* Hierarchical clustering before PCA



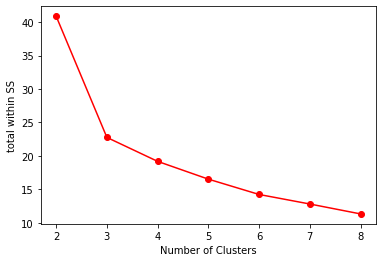
Hierarchical clustering was performed on non PCA dataset and clusters were stored under ‘clust1H’ feature

* PCA

PCA was done on dataset and as mentioned 3 PCs were derived, these PCs were stored in a different dataframa and Kmeans and hierarchical clustering was run on them



* K means clustering after PCA



To know about what difference clustering observed after PCA, head of dataset was taken

Type Alcohol Malic Ash ... clust1K clust1H clust2K clust2H

0 1 14.23 1.71 2.43 ... 1 0 1 2

1 1 13.20 1.78 2.14 ... 1 0 1 2

2 1 13.16 2.36 2.67 ... 1 0 1 2

3 1 14.37 1.95 2.50 ... 1 0 1 2

4 1 13.24 2.59 2.87 ... 1 0 1 2

25 1 13.05 2.05 3.22 ... 1 0 1 1

59 2 12.37 0.94 1.36 ... 2 2 0 0

60 2 12.33 1.10 2.28 ... 2 2 0 0

61 2 12.64 1.36 2.02 ... 0 2 2 0

62 2 13.67 1.25 1.92 ... 2 2 0 0

64 2 12.17 1.45 2.53 ... 2 0 0 1

67 2 12.37 1.17 1.92 ... 2 0 0 1

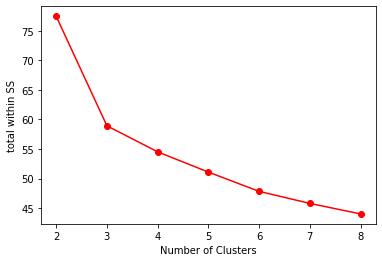
68 2 13.34 0.94 2.36 ... 2 2 0 0

71 2 13.86 1.51 2.67 ... 2 0 0 1

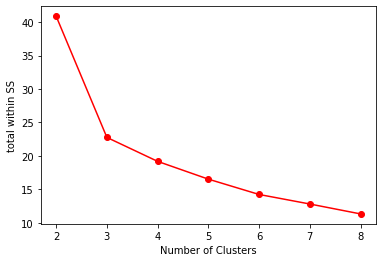
72 2 13.49 1.66 2.24 ... 2 0 0 1

As it can be observed above Clustering of Kmeans and agglomerative has happened and is almost same but not totally identical.

As per Scree plots of before and after K means clustering they are as follows



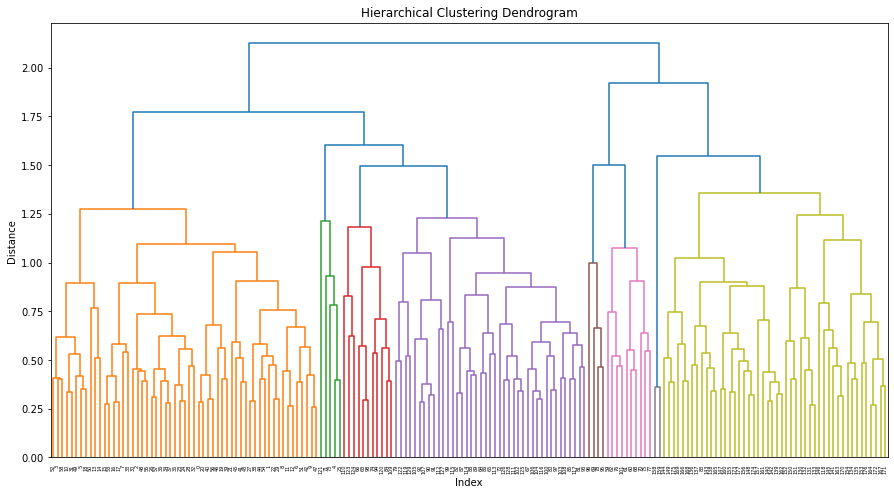
While after PCA scree plot is as follows



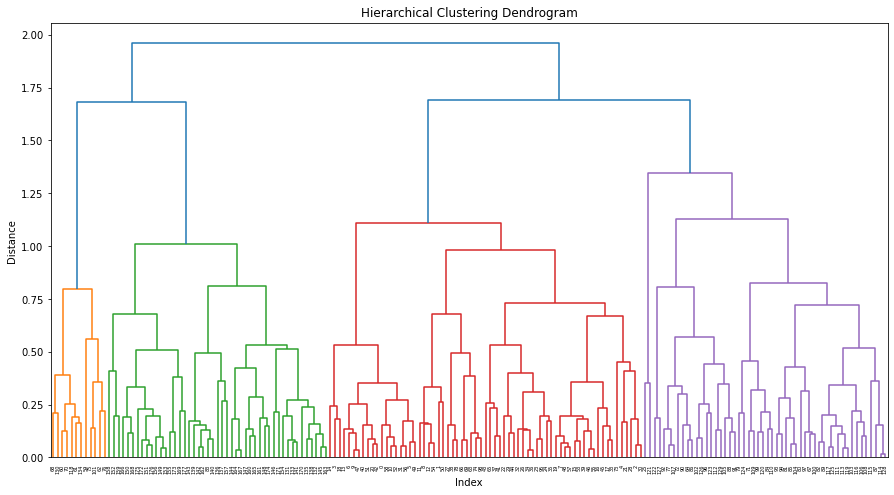
Both of these scree plots are almost identical

Now lets observe dendrograms

Before PCA dendrogram



After PCA dendrogram



As it can be clearly seen PCA has helped dendrogram to cluster better since after PCA dendrogram is better in terms of clarity and sophistication

* Benefits for client due to PCA

PCA helps performance of model at a slight acceptable cost of accuracy. But this cost is acceptable since as it can be observed above in terma of agglomerative clustering PCA makes clustering better.