SUPERVISED AND UNSUPERVISED
DATA ANALYSIS OF
MULTIVARIATE DATASET USING
LINEAR STATISTICAL METHODS

Arighna Roy

1. Introduction

Multivariate datasets are prevalent in today's world of predictive analysis. A Multivariate dataset can be defined as a collection of records where each record represents a collection values for a specific instance. The easiest way to represent such datasets is through matrices. Each row of a matrix is an instance or observation vector. Each column is a feature for which the data values are collected for all the instances.

The simplest methods to approach these datasets are the Generalized Linear Models (GLM). GLM uses specific assumptions such as independent distribution of variables, multivariate normal distribution and homogeneity of variances etc. However, these assumptions vary across models and scenarios.

We want to analyze the seeds data set from "Institute of Agrophysics of the Polish Academy of Sciences in Lublin". This data set contains the measurements of geometrical properties of kernels belonging to three different varieties of wheat. A soft X-ray technique and GRAINS package were used to construct all seven, real-valued attributes.

High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin.

To construct the data, seven geometric parameters of wheat kernels were measured:

1. area A,
2. perimeter P,
3. compactness C = 4*pi*A/P^2,
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.

All these parameters are real-valued continuous variable. As a contrast to the general multivariate dataset, each row is considered as an instance of a wheat seed and each column is a feature of the seed.

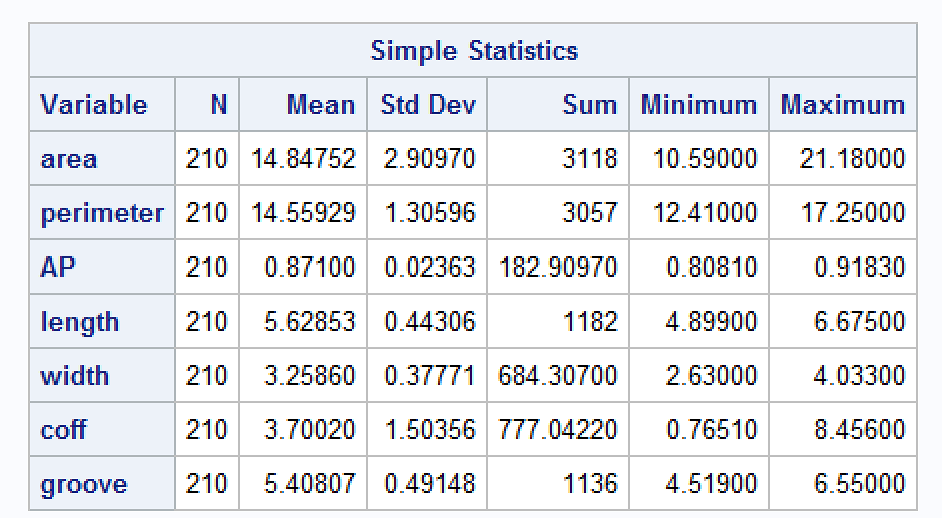
Our main purpose in this report is to examine the relationship among different variables (attributes), whether there exist significant differences among three different varieties of wheats, how to build a learning model with these features to predict a new instance, and reduce the number of dimensions (attributed).

We use several multivariate analysis techniques:

- MANOVA
- Discriminant Function Analysis
- Principal Component Analysis
- Classification
- Clustering

2. Data description and exploratory analysis

The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment (210 total observations). Table 1 contains descriptive statistics for each variable.



Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
area	210	14.84752	2.90970	3118	10.59000	21.18000
perimeter	210	14.55929	1.30596	3057	12.41000	17.25000
AP	210	0.87100	0.02363	182.90970	0.80810	0.91830
length	210	5.62853	0.44306	1182	4.89900	6.67500
width	210	3.25860	0.37771	684.30700	2.63000	4.03300
coff	210	3.70020	1.50356	777.04220	0.76510	8.45600
groove	210	5.40807	0.49148	1136	4.51900	6.55000

Table 1 : Summary

From table1, we can clearly see that the variances of the variables are quite different from each other. Therefore, the variables are not commensurate.

We have to keep in mind that the third variable AP (compactness) is a nonlinear derived variable from the first and the second variable.

In order to see if our data follows multivariate normal distribution, we investigate the normality of each variable separately and the correlation between each pair of variables. The normality plots for each of variables have shown in figure1. Based on the plots, we cannot reject the normality assumption of the variables. Moreover, we investigate the bivariate scatter plots of each pair of variables (figure2). As there are no nonlinear relationships between each pair of variables, we can conclude that there is not enough evidence to say the data distribution is departure from the MVN distribution.

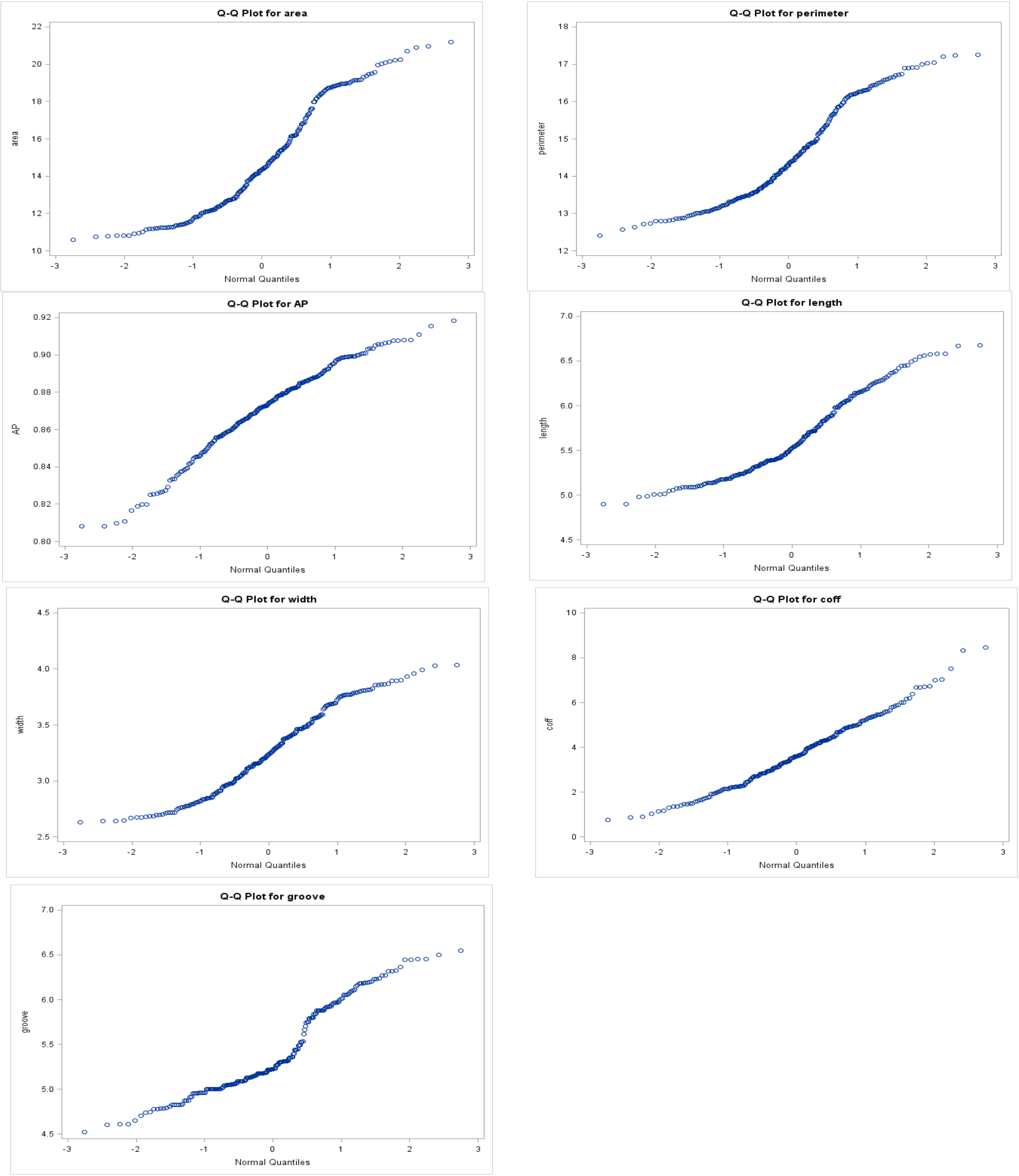


Figure 1 : Q-Q plot

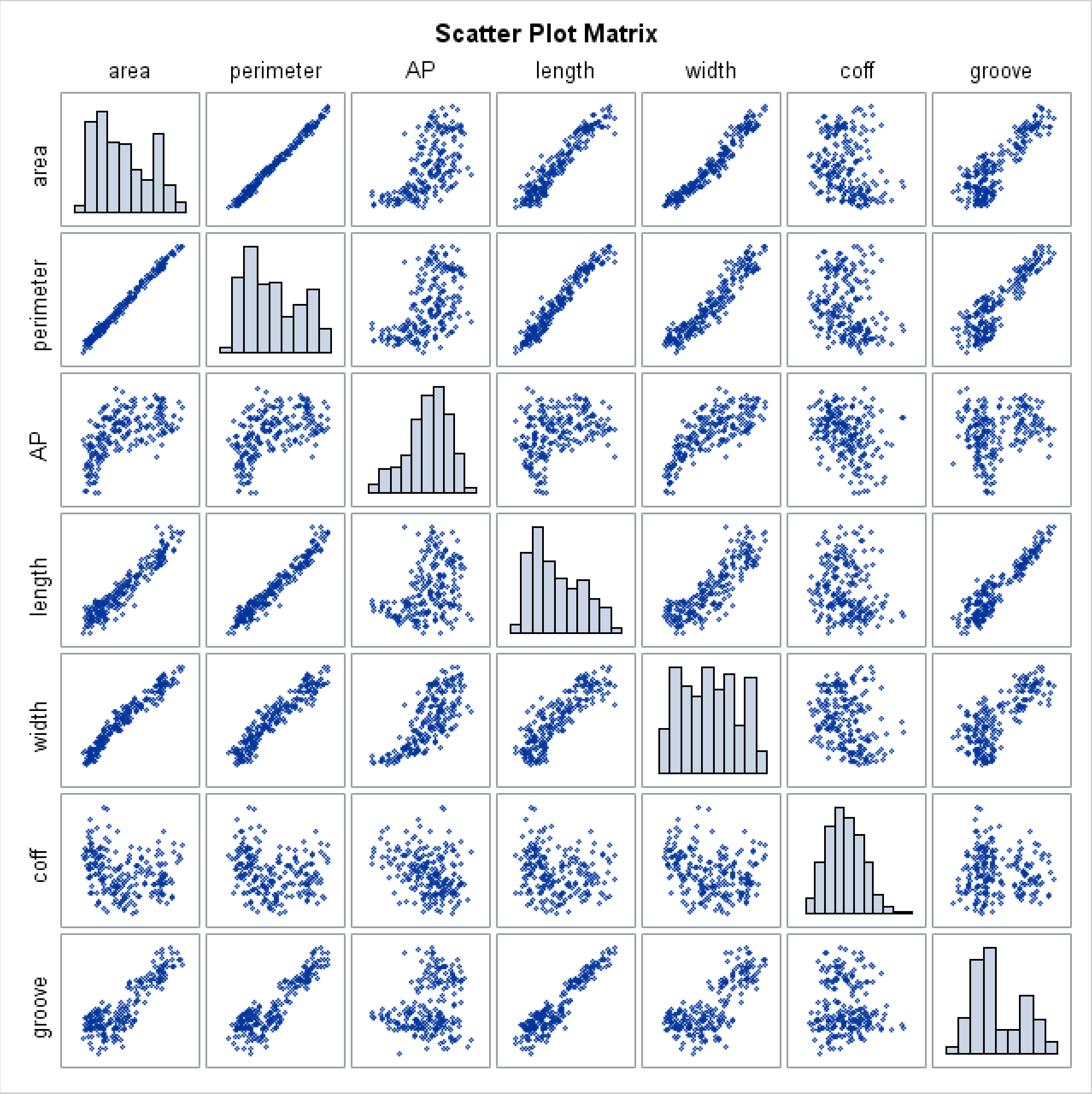
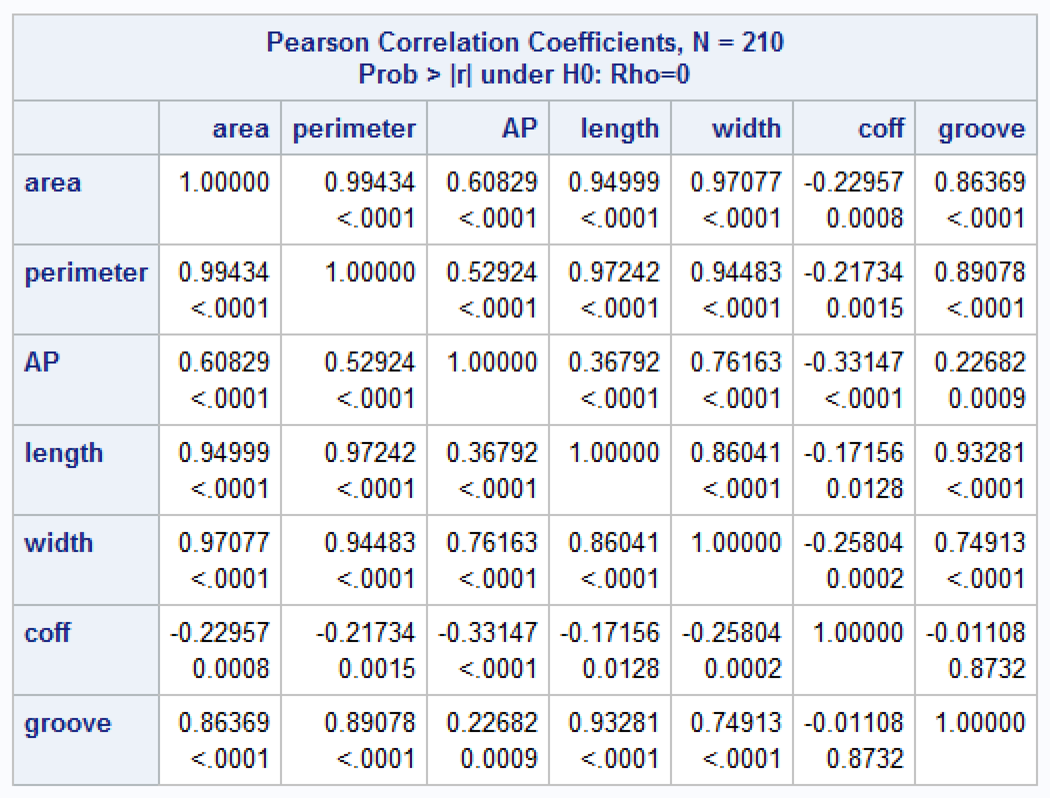


Figure 2: Scatterplot matrix of seed data

There are strong correlations between each pair of variables. The scatter plot (figure2) and also the Pearson correlation matric in table2 show this strong correlations.



Pearson Correlation Coefficients, N = 210
Prob > |r| under H0: Rho=0

	area	perimeter	AP	length	width	coff	groove
area	1.00000	0.99434 <.0001	0.60829 <.0001	0.94999 <.0001	0.97077 <.0001	-0.22957 0.0008	0.86369 <.0001
perimeter	0.99434 <.0001	1.00000	0.52924 <.0001	0.97242 <.0001	0.94483 <.0001	-0.21734 0.0015	0.89078 <.0001
AP	0.60829 <.0001	0.52924 <.0001	1.00000	0.36792 <.0001	0.76163 <.0001	-0.33147 <.0001	0.22682 0.0009
length	0.94999 <.0001	0.97242 <.0001	0.36792 <.0001	1.00000	0.86041 <.0001	-0.17156 0.0128	0.93281 <.0001
width	0.97077 <.0001	0.94483 <.0001	0.76163 <.0001	0.86041 <.0001	1.00000	-0.25804 0.0002	0.74913 <.0001
coff	-0.22957 0.0008	-0.21734 0.0015	-0.33147 <.0001	-0.17156 0.0128	-0.25804 0.0002	1.00000	-0.01108 0.8732
groove	0.86369 <.0001	0.89078 <.0001	0.22682 0.0009	0.93281 <.0001	0.74913 <.0001	-0.01108 0.8732	1.00000

Table 2 : correlation of different variables

3. Methods and Analysis

MANOVA

A one-way Multivariate Analysis of Variance (MANOVA) was carried out in order to investigate differences among three different wheat types. We ran a MANOVA testing the null hypothesis, H0: µ1= µ2= µ3.

MANOVA Assumptions Test

The assumptions of one way MANOVA: The 3 random samples of wheats are independent. The observation vectors come from multivariate normal populations. The three multivariate normal populations (from which the 3 samples of observation vectors are drawn) have a common population covariance matrix, Σ.

- To assess the multivariate normality of the data, first we look at the distribution of each variable separately for each group (class). As we see in the below figures, the Q-Q plots for each class show/follow nearly a straight line, so there is no indication of departure from normality for each variable.

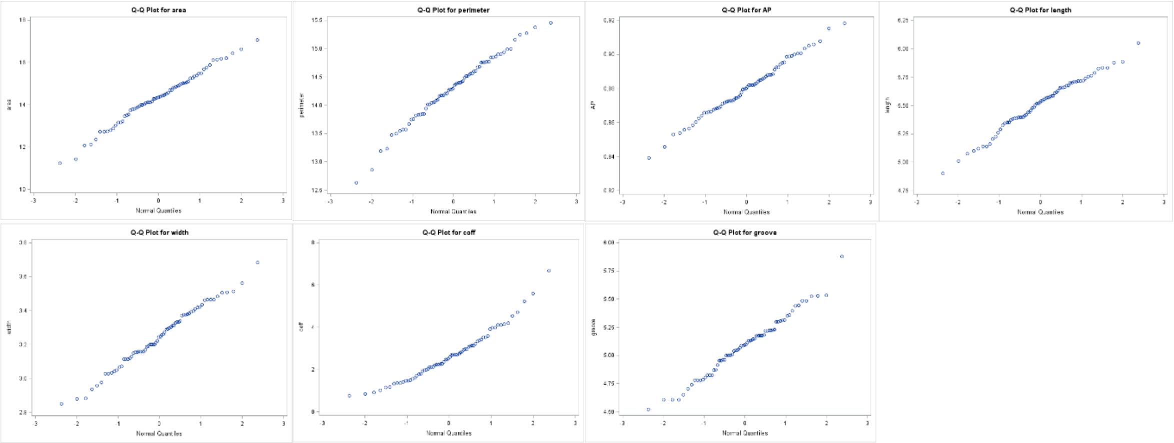


Figure 3: QQ plot for class 1

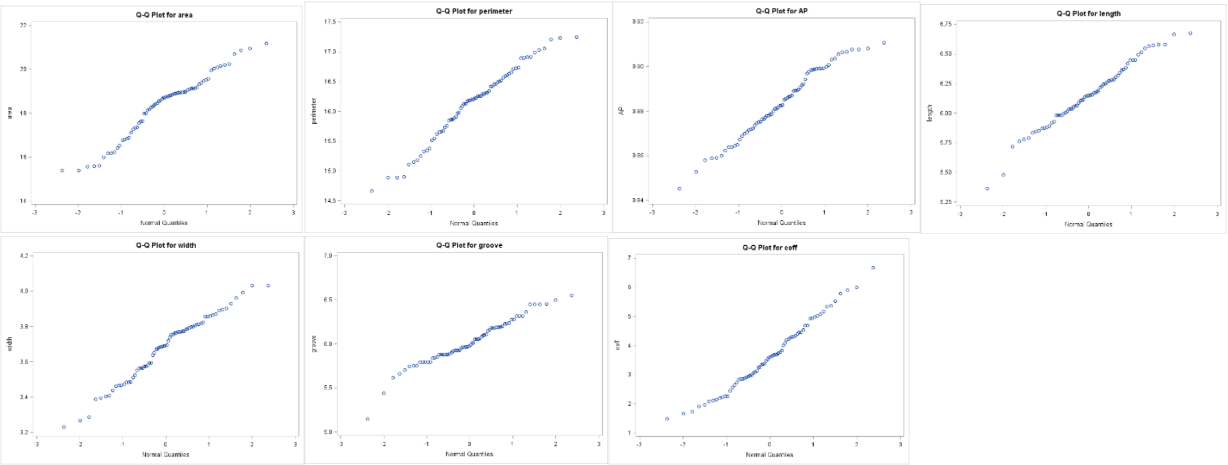


Figure 4: QQ plot for class 2

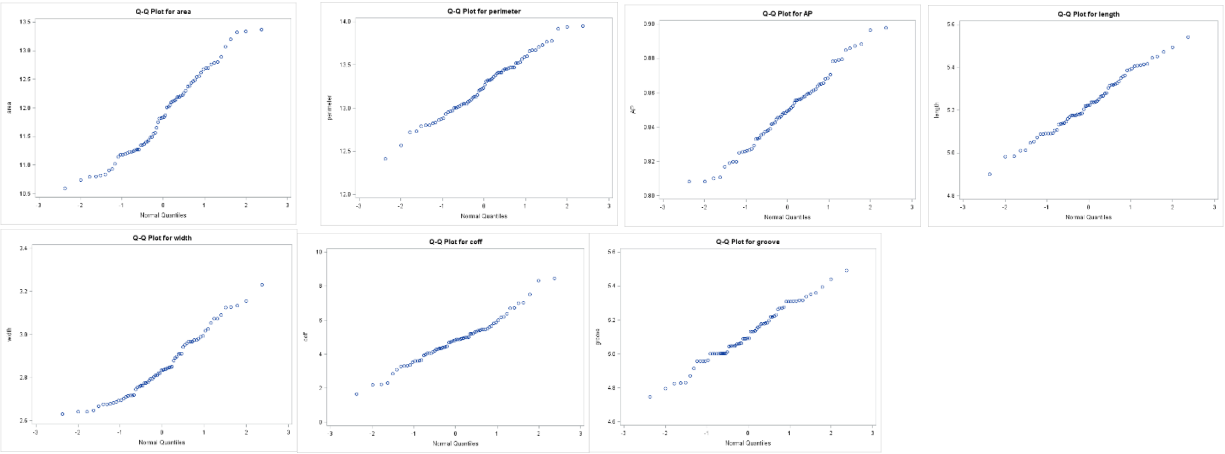
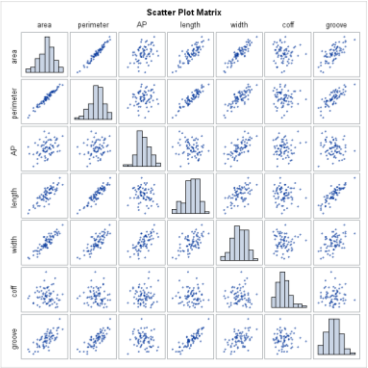
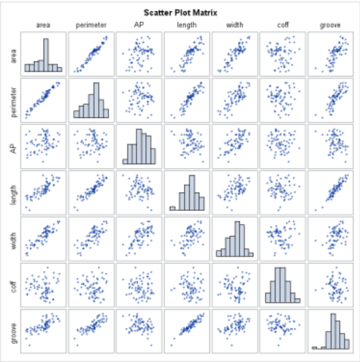


Figure 5: QQ plot for class 3

- Moreover, the bivariate scatter plots for each class can be used to assess the multivariate normality. As there is no nonlinear relationship between each pair, we conclude that there is no evidence of departure from MVN distribution.





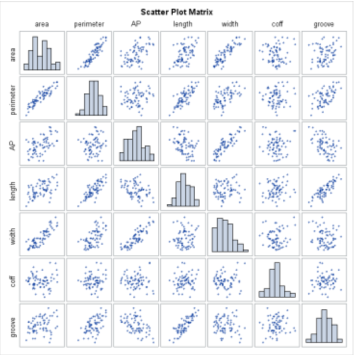
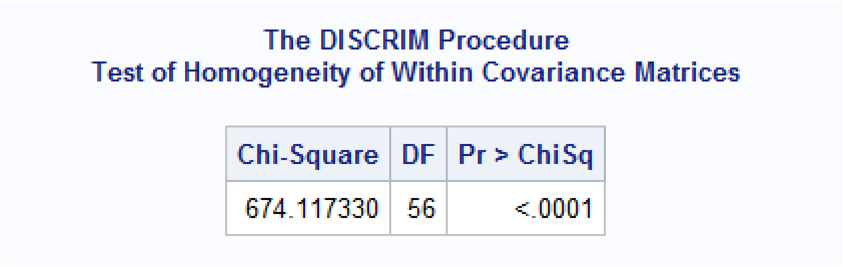


Figure 6: Scatter plot for class 1 *Figure 7: Scatter plot for class 2* *Figure 8: Scatter plot for class 3*

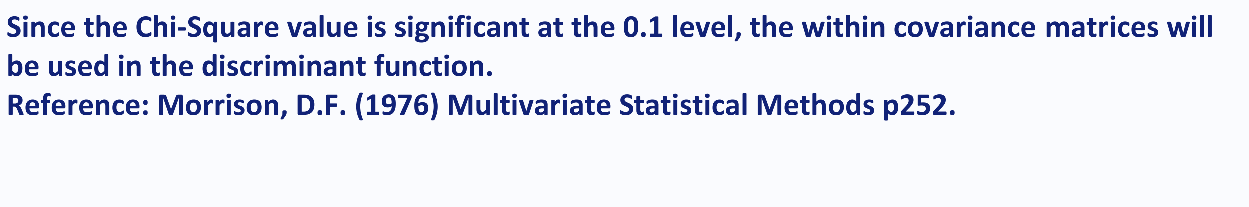
- In order to show that they have a common population covariance we do the test: H0: Σ1 = Σ2 = Σ3.



The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
674.117330	56	<.0001

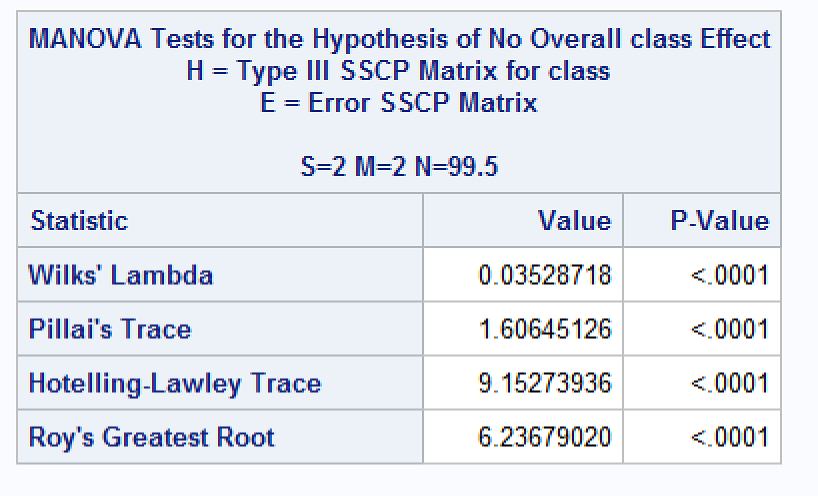
Table 3: Test of Homogeneity



Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

As we see we cannot reject the null hypnosis assumption and we conclude that there is not enough evidence to say the data do not have a common population covariance.

MANOVA results



MANOVA Tests for the Hypothesis of No Overall class Effect
H = Type III SSCP Matrix for class
E = Error SSCP Matrix

S=2 M=2 N=99.5

Statistic	Value	P-Value
Wilks' Lambda	0.03528718	<.0001
Pillai's Trace	1.60645126	<.0001
Hotelling-Lawley Trace	9.15273936	<.0001
Roy's Greatest Root	6.23679020	<.0001

Table 4 : MANOVA test Statistics

Wilk’s Lambda for this test has a p-value < 0.0001, we conclude that the population mean vectors differ for wheats in at least two wheat groups at α = 0.1 (1% confidence level).

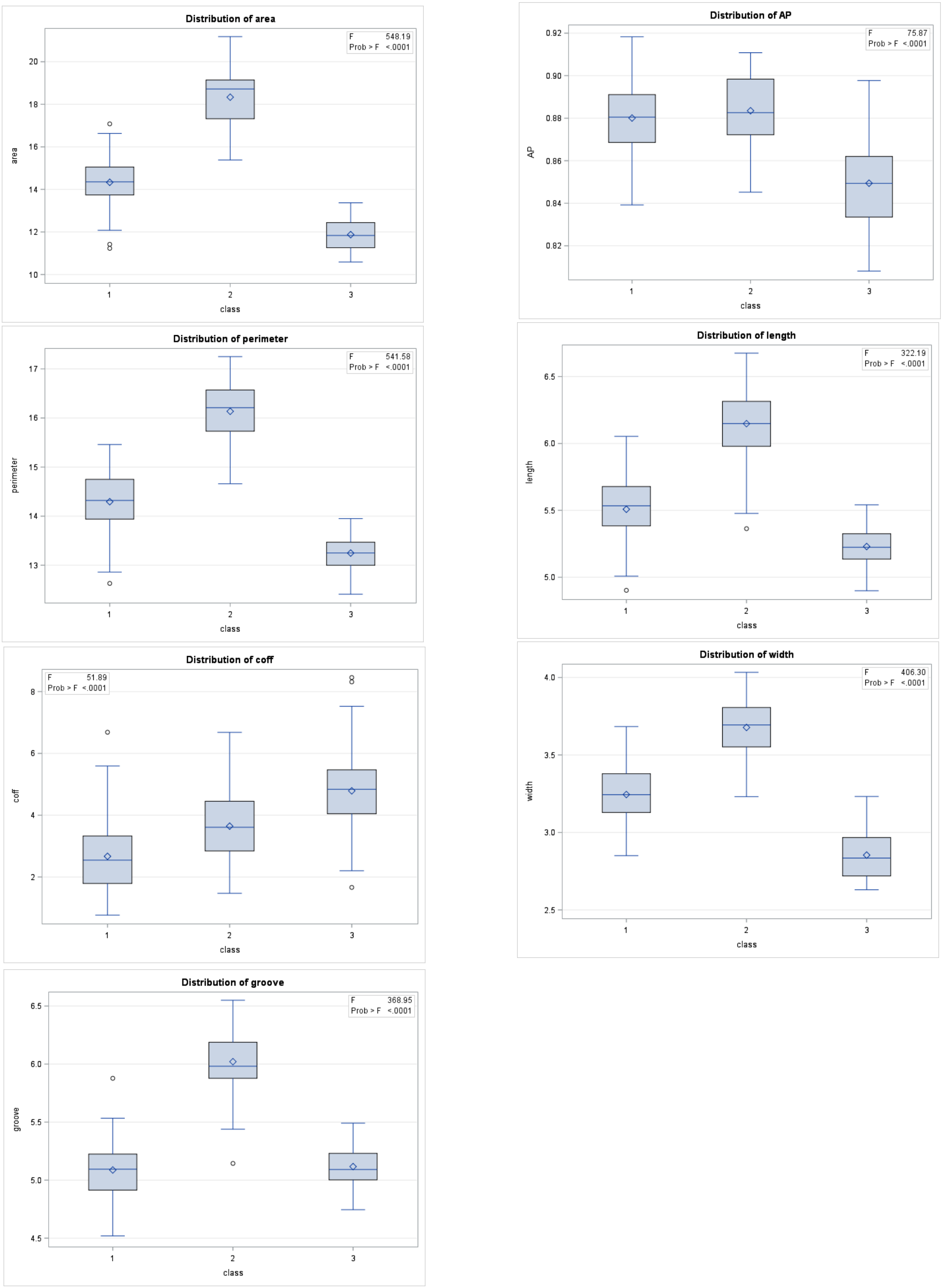
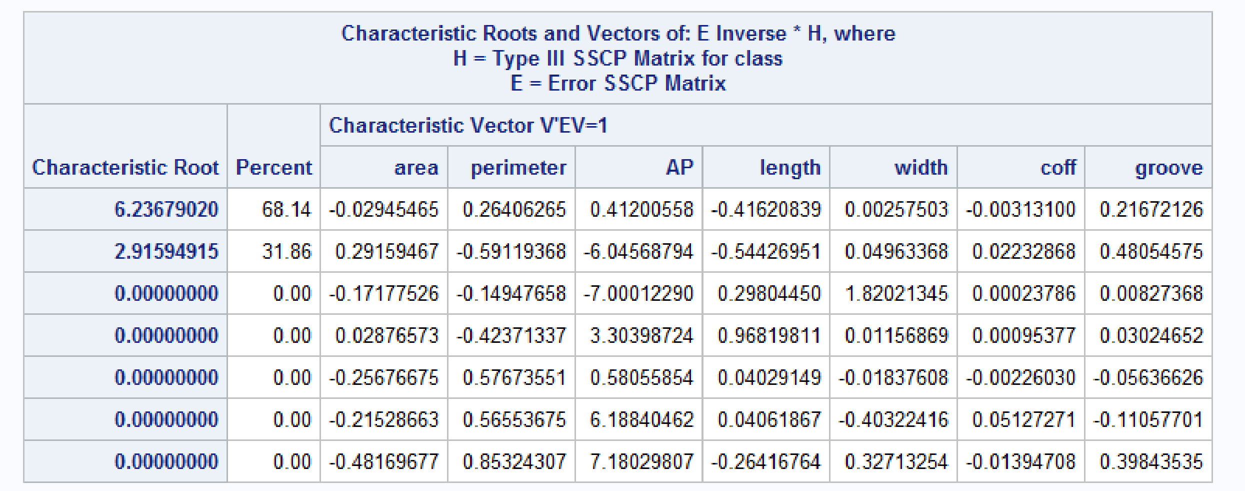


Figure 9 : Box plots for different variables

Discriminant Analysis

From the previous section, we got that there are differences between the mean vectors of three different wheats. In this section, we have performed the discriminant analysis in order to identify the relative contribution of the 7 variables to the separation of the three groups. Moreover, we wanted to find the optimal plane on which the points can be projected to best illustrate the configuration of the groups.

Form the MANOVA section we have the eigenvalues and eigenvectors of $E^{-1}H$ (Table5) that can be used in our analysis.



Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for class
E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1						
		area	perimeter	AP	length	width	coff	groove
6.23679020	68.14	-0.02945465	0.26406265	0.41200558	-0.41620839	0.00257503	-0.00313100	0.21672126
2.91594915	31.86	0.29159467	-0.59119368	-6.04568794	-0.54426951	0.04963368	0.02232868	0.48054575
0.00000000	0.00	-0.17177526	-0.14947658	-7.00012290	0.29804450	1.82021345	0.00023786	0.00827368
0.00000000	0.00	0.02876573	-0.42371337	3.30398724	0.96819811	0.01156869	0.00095377	0.03024652
0.00000000	0.00	-0.25676675	0.57673551	0.58055854	0.04029149	-0.01837608	-0.00226030	-0.05636626
0.00000000	0.00	-0.21528663	0.56553675	6.18840462	0.04061867	-0.40322416	0.05127271	-0.11057701
0.00000000	0.00	-0.48169677	0.85324307	7.18029807	-0.26416764	0.32713254	-0.01394708	0.39843535

*Table 5: The eigenvalues and eigenvectors of E invers*H*

There are only two non-zero eigenvalues and their corresponding eigenvectors are as follows:

Variables	a1	a2	Standardized a1	Standardized a1
Area	-0.0294547	0.2915947	-0.493772269	4.888231
Perimeter	0.2640627	-0.5911937	1.996977676	-4.47091
AP	0.4120056	-6.0456879	0.10691181	-1.5688
Length	-0.4162084	-0.5442695	-1.31454501	-1.71901
width	0.002575	0.0496337	0.006335536	0.122119
coff	-0.003131	0.0223287	-0.055544035	0.396112
groove	0.2167213	0.4805458	0.720731093	1.598109

Table 6

As the variables are not commensurate, so it is better to find the standard vectors to determine the contribution of each variable on the discriminant function. As we see, the variables perimeter, length, and groove have the most contribution in the separation of the group means based on the first discriminant function. In addition, the variables perimeter, area, and length have the most contribution in the separation of the group means based on the second discriminant function.

We reduced the dimension through the discriminant analysis; so we can plot our data in a two-dimension vector space. We got that two discriminate functions account for the total sum of eigenvalues. We showed the patterns in the mean vectors in two-dimensional scatterplot in figure 3. We see that the first discriminant function effectively separates group 2 from group 1 and 3, whereas the second discriminant function is less successful in separating group 2 from group 3.

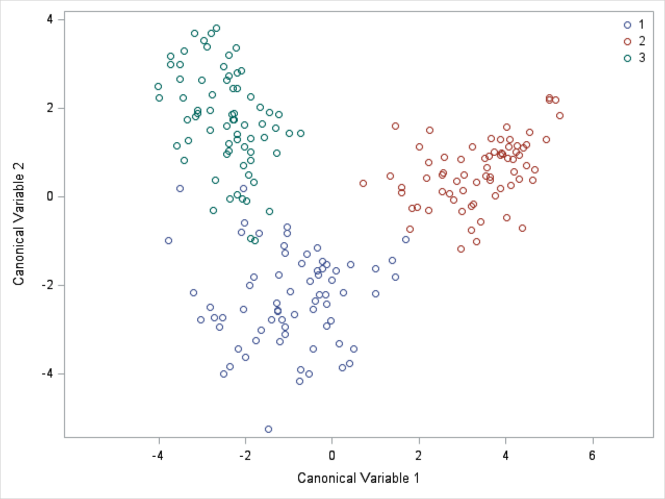
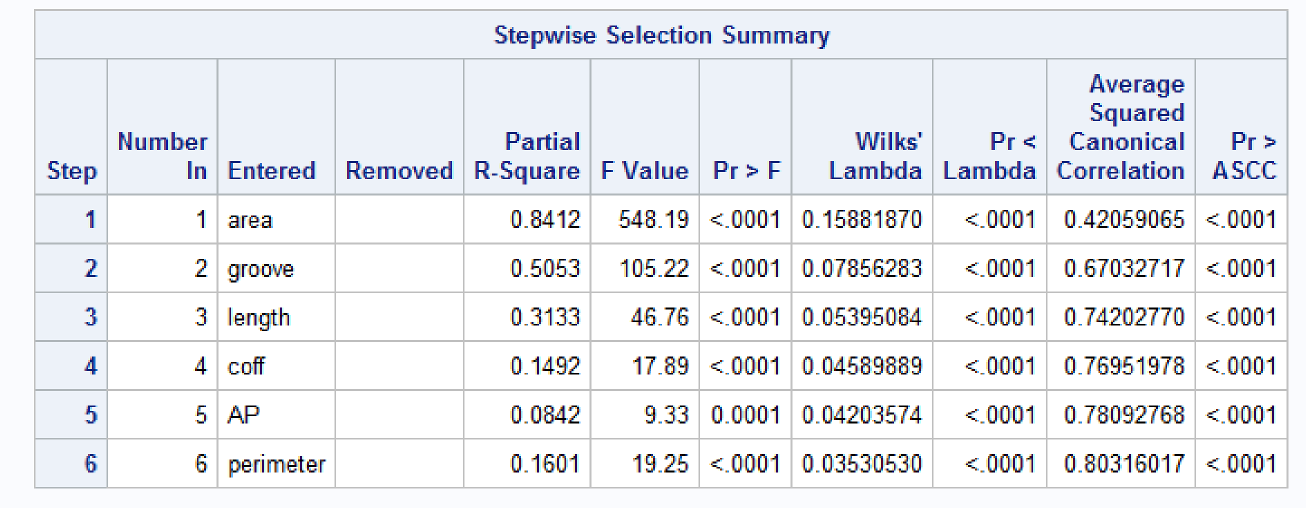


Figure 10

Stepwise

In the analysis procedure, we tried STEPWISE options in stepdisc procedure in SAS to select the "good" subsets of all variables. We found that the variables area, groove, length, coff, AP, and perimeter have significant discrimination power. As we see just one of the variables (width) is redundant. We will use this insight in classification section.



Stepwise Selection Summary

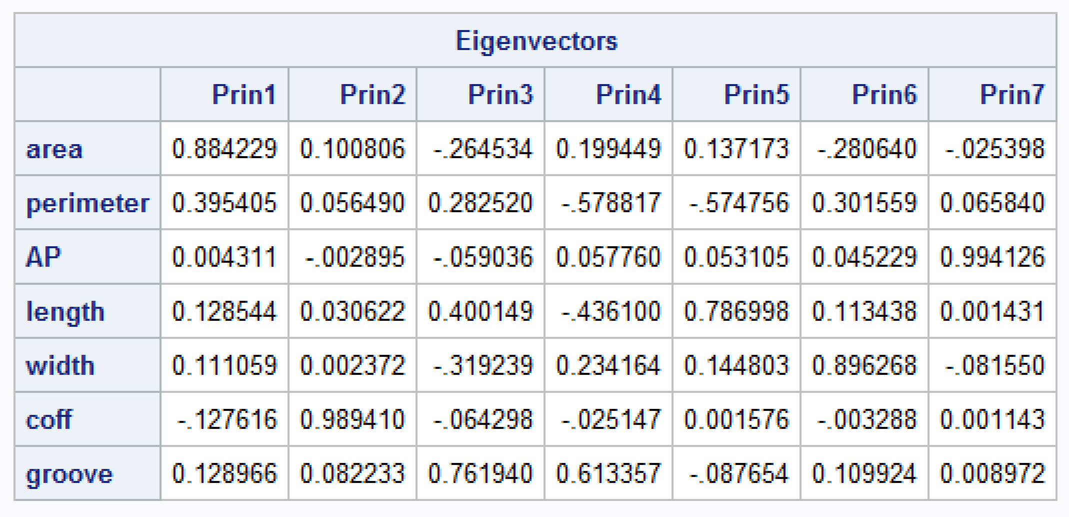
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	area		0.8412	548.19	<.0001	0.15881870	<.0001	0.42059065	<.0001
2	2	groove		0.5053	105.22	<.0001	0.07856283	<.0001	0.67032717	<.0001
3	3	length		0.3133	46.76	<.0001	0.05395084	<.0001	0.74202770	<.0001
4	4	coff		0.1492	17.89	<.0001	0.04589889	<.0001	0.76951978	<.0001
5	5	AP		0.0842	9.33	0.0001	0.04203574	<.0001	0.78092768	<.0001
6	6	perimeter		0.1601	19.25	<.0001	0.03530530	<.0001	0.80316017	<.0001

Table 7: Stepwise Selection Summary

Principal component analysis (PCA)

Principal Component Analysis is the most commonly used method for dimensionality reduction. Though we are not dealing with a huge dataset, but it's a nice idea to check whether we can reduce the number of dimensions while explaining the most part of the variance.

After we apply PCA, we get the below 7 Principal components.

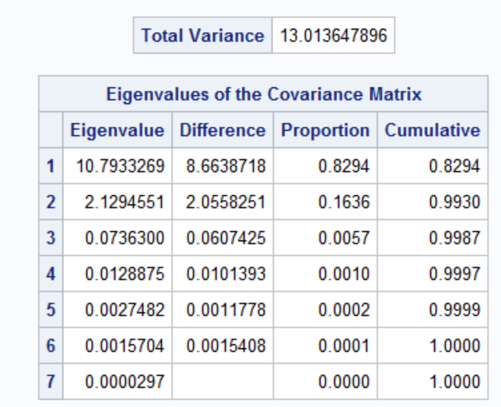


Eigenvectors							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
area	0.884229	0.100806	-.264534	0.199449	0.137173	-.280640	-.025398
perimeter	0.395405	0.056490	0.282520	-.578817	-.574756	0.301559	0.065840
AP	0.004311	-.002895	-.059036	0.057760	0.053105	0.045229	0.994126
length	0.128544	0.030622	0.400149	-.436100	0.786998	0.113438	0.001431
width	0.111059	0.002372	-.319239	0.234164	0.144803	0.896268	-.081550
coff	-.127616	0.989410	-.064298	-.025147	0.001576	-.003288	0.001143
groove	0.128966	0.082233	0.761940	0.613357	-.087654	0.109924	0.008972

Table 8

Now, these components can be used as features instead of the combination of area, perimeter, AP, length, width, coff, groove. We are not done with dimensionality reduction phase yet, as we still have 7 variables.

We have to find the percentage of variance explained by each principal component.



Total Variance	13.013647896

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	10.7933269	8.6638718	0.8294	0.8294
2	2.1294551	2.0558251	0.1636	0.9930
3	0.0736300	0.0607425	0.0057	0.9987
4	0.0128875	0.0101393	0.0010	0.9997
5	0.0027482	0.0011778	0.0002	0.9999
6	0.0015704	0.0015408	0.0001	1.0000
7	0.0000297		0.0000	1.0000

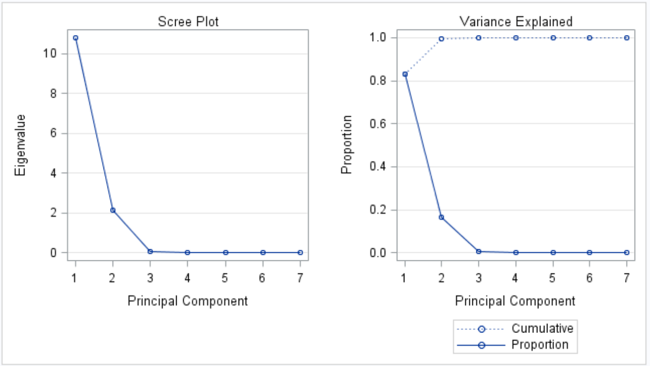
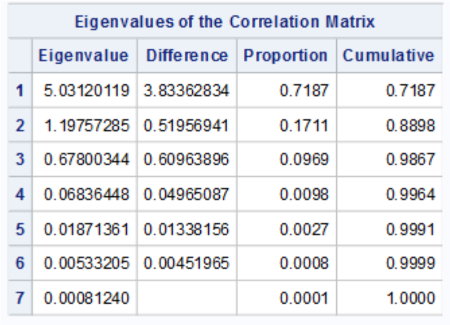


Table 9

Figure 11

Table9 shows that the first 2 principal components can explain 99% of the variance. Below is the chart which shows the percentage of explained variance.

However, we have to keep that in mind that variance of each variable is different. As the variables are not commensurate, we should use correlation instead of covariance. Table10 shows the principal components calculated from the correlation matrix. Now we need 4 principal components to explain 99% of the variance. But the first 2 principal components can explain 88% of the total variance which is quite good.



Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	5.03120119	3.83362834	0.7187	0.7187
2	1.19757285	0.51956941	0.1711	0.8898
3	0.67800344	0.60963896	0.0969	0.9867
4	0.06836448	0.04965087	0.0098	0.9964
5	0.01871361	0.01338156	0.0027	0.9991
6	0.00533205	0.00451965	0.0008	0.9999
7	0.00081240		0.0001	1.0000

Table 10

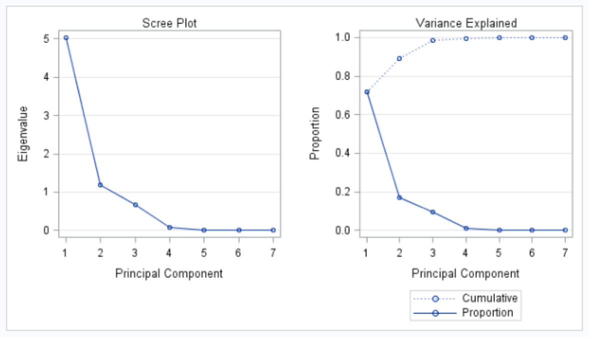
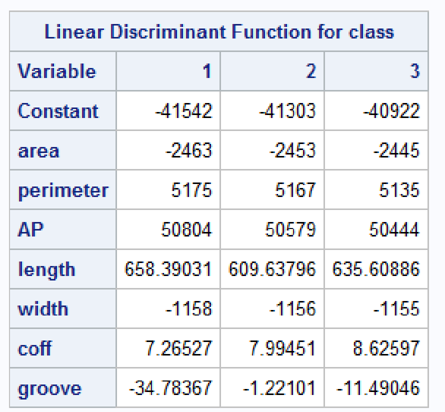


Figure 12

Classification

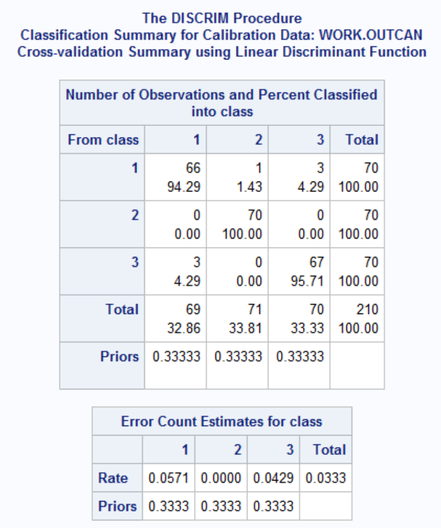
In this section, we classify the seeds into specific group of wheat it belongs to. We use SAS to identify the linear classification functions. So we collect new observation **Y**, we can use the determined discriminant functions in order to predict the class of new observation.



Linear Discriminant Function for class			
Variable	1	2	3
Constant	-41542	-41303	-40922
area	-2463	-2453	-2445
perimeter	5175	5167	5135
AP	50804	50579	50444
length	658.39031	609.63796	635.60886
width	-1158	-1156	-1155
coff	7.26527	7.99451	8.62597
groove	-34.78367	-1.22101	-11.49046

Table 11: Linear Discriminant Function for class

To measure the performance of the classification method, we use error rate. The methods for estimating the error rate are resubstituting and cross validation. Using these methods, the classification rule is applied to each observation vector and this observation is assigned to a group of wheat. Then, the number of misclassifications are counted. We prefer the cross validation, so we just mentioned the result of this method in table 8.



The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.OUTCAN
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into class

From class	1	2	3	Total
1	66 94.29	1 1.43	3 4.29	70 100.00
2	0 0.00	70 100.00	0 0.00	70 100.00
3	3 4.29	0 0.00	67 95.71	70 100.00
Total	69 32.86	71 33.81	70 33.33	210 100.00
Priors	0.33333	0.33333	0.33333	

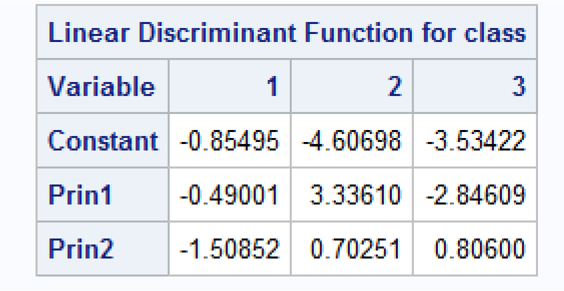
Error Count Estimates for class

	1	2	3	Total
Rate	0.0571	0.0000	0.0429	0.0333
Priors	0.3333	0.3333	0.3333	

Table 12 : Cross-validation Summary & Error rate

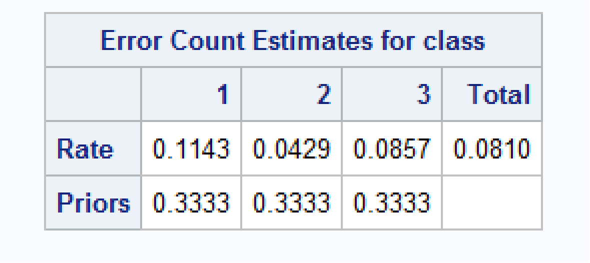
The error rate is the same when we skip the redundant variable (width) for the classification. So the stepwise discriminant function was helpful in reducing the dimension without losing the performance of the method.

Now, we reduce the dimesons of our data using the first and second principle components. If we use these two components in our classification instead of 7 variables we get different linear discriminant functions for each class (table 13).



Linear Discriminant Function for class

Variable	1	2	3
Constant	-0.85495	-4.60698	-3.53422
Prin1	-0.49001	3.33610	-2.84609
Prin2	-1.50852	0.70251	0.80600



Error Count Estimates for class

	1	2	3	Total
Rate	0.1143	0.0429	0.0857	0.0810
Priors	0.3333	0.3333	0.3333	

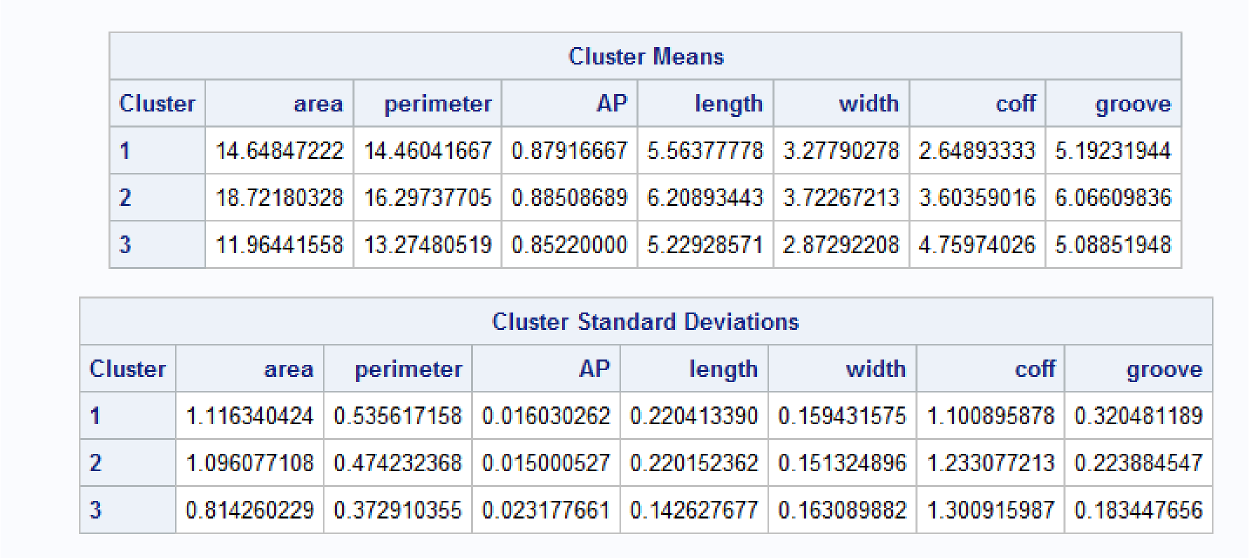
Table 13: Linear Discriminant function

Table 14: Error rate

The error goes higher a little bit from the previous classification process. However, we get a reduced dataset from the original one at the cost of some error.

Clustering

We have further used an unsupervised method to cluster (group) the data based on the available features. To cluster the data into groups, we have used k-means algorithm. Then we have matched the result with the initial group(class) labels. Below is the result of clustering.



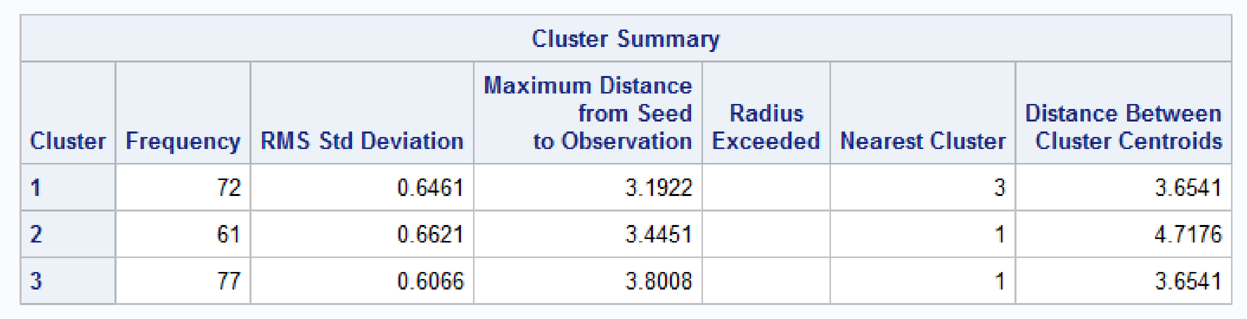
Cluster Means

Cluster	area	perimeter	AP	length	width	coff	groove
1	14.64847222	14.46041667	0.87916667	5.56377778	3.27790278	2.64893333	5.19231944
2	18.72180328	16.29737705	0.88508689	6.20893443	3.72267213	3.60359016	6.06609836
3	11.96441558	13.27480519	0.85220000	5.22928571	2.87292208	4.75974026	5.08851948

Cluster Standard Deviations

Cluster	area	perimeter	AP	length	width	coff	groove
1	1.116340424	0.535617158	0.016030262	0.220413390	0.159431575	1.100895878	0.320481189
2	1.096077108	0.474232368	0.015000527	0.220152362	0.151324896	1.233077213	0.223884547
3	0.814260229	0.372910355	0.023177661	0.142627677	0.163089882	1.300915987	0.183447656

Table 15



Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	72	0.6461	3.1922		3	3.6541
2	61	0.6621	3.4451		1	4.7176
3	77	0.6066	3.8008		1	3.6541

Table 16

From the above tables, we get the mean vectors of three clusters and the summary of each cluster.

To get a contrast with the assigned cluster label and provided class label, we have clustered the data into 3 groups. Below is the comparative result. The cells, which are highlighted into red, are the mis-clustered ones. We can see that the clustered data are quite similar to the provided class labels. Out of 210 instances, 22 are clustered into wrong group. But we still get almost 90% accuracy in clustering the seeds.

Class	ID	Cluster	Class	ID	Cluster	Class	ID	Cluster	Class	ID	Cluster	Class	ID	Cluster	Class	ID	Cluster
1	1	1	2	71	2	3	141	3	1	36	1	2	106	2	3	176	3
1	2	1	2	72	2	3	142	3	1	37	1	2	107	2	3	177	3
1	3	1	2	73	2	3	143	3	1	38	2	2	108	2	3	178	3
1	4	1	2	74	2	3	144	3	1	39	1	2	109	2	3	179	3
1	5	1	2	75	2	3	145	3	1	40	3	2	110	2	3	180	1
1	6	1	2	76	2	3	146	3	1	41	1	2	111	2	3	181	3
1	7	1	2	77	2	3	147	3	1	42	1	2	112	2	3	182	3
1	8	1	2	78	2	3	148	3	1	43	1	2	113	2	3	183	3
1	9	1	2	79	2	3	149	3	1	44	1	2	114	2	3	184	3
1	10	1	2	80	2	3	150	3	1	45	1	2	115	2	3	185	3
1	11	1	2	81	2	3	151	3	1	46	1	2	116	2	3	186	3
1	12	1	2	82	2	3	152	3	1	47	1	2	117	2	3	187	3
1	13	1	2	83	2	3	153	3	1	48	1	2	118	2	3	188	3
1	14	1	2	84	2	3	154	3	1	49	1	2	119	2	3	189	3
1	15	1	2	85	2	3	155	3	1	50	1	2	120	2	3	190	3
1	16	1	2	86	2	3	156	3	1	51	1	2	121	2	3	191	3
1	17	3	2	87	2	3	157	3	1	52	1	2	122	2	3	192	3
1	18	1	2	88	2	3	158	3	1	53	1	2	123	1	3	193	3
1	19	1	2	89	2	3	159	3	1	54	1	2	124	2	3	194	3
1	20	3	2	90	2	3	160	3	1	55	1	2	125	1	3	195	3
1	21	1	2	91	2	3	161	3	1	56	1	2	126	2	3	196	3
1	22	1	2	92	2	3	162	3	1	57	1	2	127	2	3	197	3
1	23	1	2	93	2	3	163	3	1	58	1	2	128	2	3	198	3
1	24	1	2	94	2	3	164	3	1	59	1	2	129	2	3	199	3
1	25	1	2	95	2	3	165	3	1	60	1	2	130	2	3	200	3
1	26	1	2	96	2	3	166	3	1	61	3	2	131	2	3	201	3
1	27	3	2	97	2	3	167	3	1	62	3	2	132	2	3	202	1
1	28	1	2	98	2	3	168	3	1	63	3	2	133	1	3	203	3
1	29	1	2	99	2	3	169	3	1	64	3	2	134	1	3	204	3
1	30	1	2	100	2	3	170	3	1	65	1	2	135	1	3	205	3
1	31	1	2	101	1	3	171	3	1	66	1	2	136	1	3	206	3
1	32	1	2	102	2	3	172	3	1	67	1	2	137	2	3	207	3
1	33	1	2	103	2	3	173	3	1	68	1	2	138	1	3	208	3
1	34	1	2	104	2	3	174	3	1	69	1	2	139	1	3	209	3
1	35	1	2	105	2	3	175	3	1	70	3	2	140	1	3	210	3

Table 17

4. Conclusion

The aim of this project is find some insight about the seeds data and find a way to distinguish the seeds for different classes. Throughout the analysis, we have only used the linear statistical models; which means any nonlinear components of the relationship would not be captured by these methods. However, our analysis was quite successful in finding some interesting insights about the relationship between the variables (also among the groups). In addition, we reduced the dimension of dataset. Thereafter, we performed a classification analysis with the whole dataset, filtered features and reduced features; we have done a comparative analysis of the performance of the classification for these methods. A clustering method has also been applied into the dataset and result was pretty good.