Multivariate Analysis on Wheat Seed Data

DEEPAK G KUTTY

INTRODUCTION

MULTIVARIATE ANALYSIS

- **'Multivariate'** refers to the presence of **multiple random variables.**
- When a phenomenon under study is **multivariate** in nature and cannot be explained by single independent variable, here it becomes necessary to analyze these **multiple variables.**
- Analysis of multiple variables is known as **Multivariate analysis.**
- **'Multivariate analysis'** is a statistical technique which simultaneously analyzes more than two variables on a set of observations.

ABOUT THE DATA

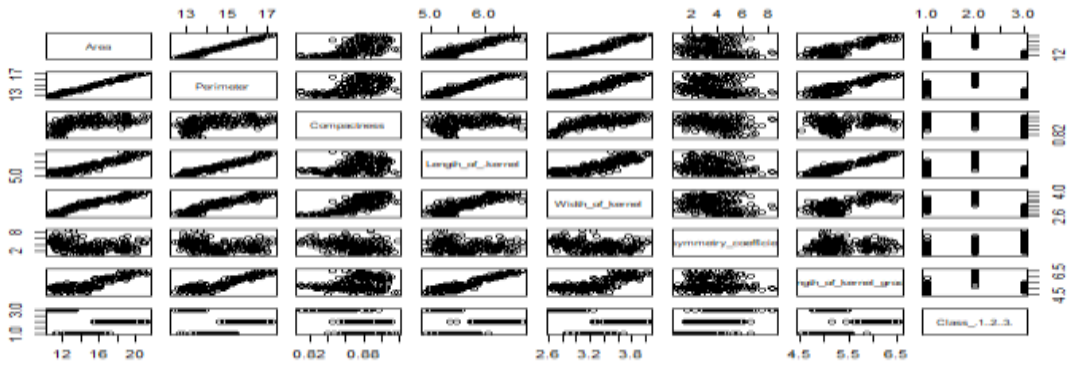
- The data used for the analysis is a secondary data from kaggle
- It contains measurement of various attributes of wheat seeds, classified into three different classes :- Kama(1),Rose(2),Canadian(3)
- The data set provide a total seven attributes or feautres that describes each wheat seed samples.

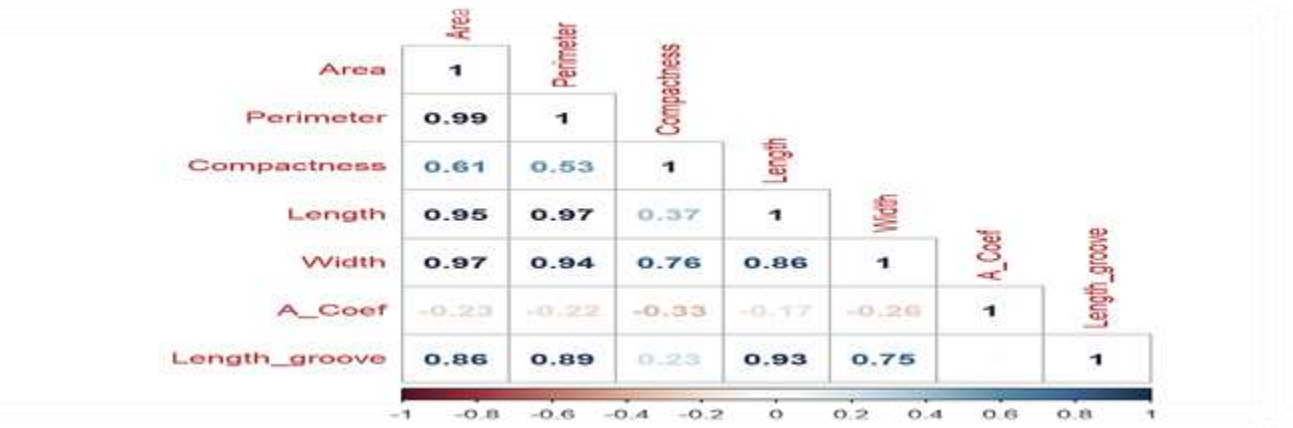
ATTRIBUTES ARE

- **AREA (X1)**
- **PERIMETER (X2)**
- **COMPACTNESS (X3)**
- **LENGTH OF KERNAL (X4)**
- **WIDTH OF KERNAL (X5)**
- **ASYMMETRY COEFFICIENT (X6)**
- **LENGTH OF KERNAL GROVE (X7)**

EXPLORATORY DATA ANALYSIS

- EDA has been promoted by “John Turkey”. In statistics, exploratory data analysis is an approach of analyzing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods. Also make sure that there is no missing values in dataset.





- Based on the scatter pot and correlation matrix , we can observe that the variables are positively as well as negatively correlated.
- Area and Perimeter have high positive correlation likewise Compactness and Asymmetery Coefficent have high negative correlation

- So we can apply dimension reduction analysis to reduce the dimension and represent these highly correlated variables with fewer new transformed variables.

OBJECTIVE

- The primary objective of this project is to conduct a multivariate analysis on the dataset using R and SPSS
- This analysis aim to uncover the relationships , dependencies and to enchance our understanding of wheat seed characterstics and their potential implications for agriculture.

MULTIVARIATE TECHNIQUES SUCH AS:-

1). Principal Component Analysis(pca)

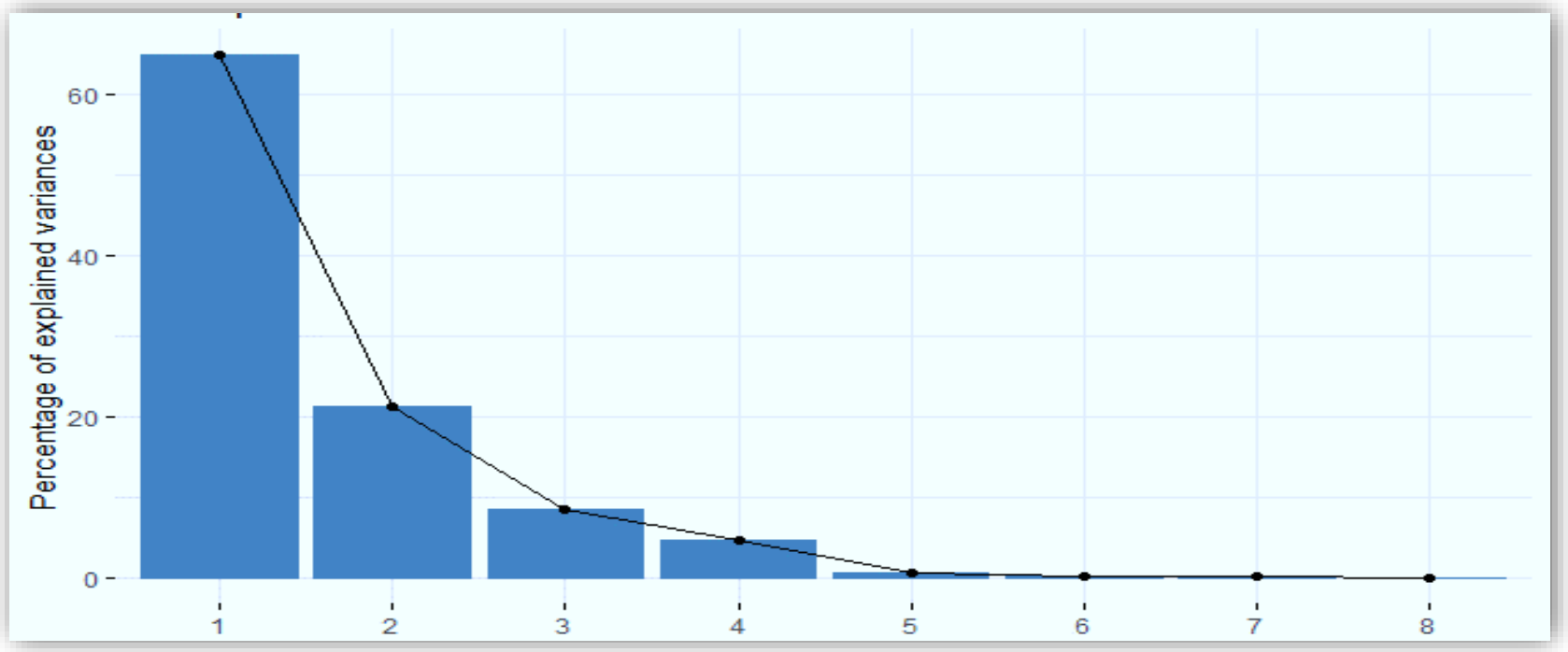
2).Cluster Analysis

3).Factor Analysis

PRINCIPAL COMPONENT ANALYSIS

- **Principle Component Analysis (PCA)** emerges as fundamental technique that can be used to simplify a dataset.
- PCA is used to extract the important information from multivariate data and express this information as a set of few new variables called **components.**
- It is a linear transformation that chooses a new coordinate system for the data set such that greatest variance by any projection of the dataset comes to lie on the first axis (first principal component), second greatest variance on the second axis and so on.
- PCA can be used for reducing dimensionality by eliminating the later principal components.

SCREEPLOT



From the scree plot, we can extract two principal components.

Variance Explained		
	PC 1	PC 2
Sum of square loading	5.19	1.70
Proportion variance	0.65	0.21
Cumulative Variance	0.65	0.86

From the variance table we can see that, **65%** of variance in the data is explained by PC1, 21% **by PC2**. That is, first two components explain **86%** of variation of data set.

COMPONENT MATRIX		
	PC1	PC2
AREA (X1)	0.992	0.105
PERIMETER (X2)	0.983	0.153
COMPACTNESS (X3)	0.654	-0.469
LENGTH OF KERNAL (X4)	0.936	-0.262
WIDTH OF KERNAL(X5)	0.975	-0.39
ASYMMETRY OF COEFFICIENT (X6)	-0.318	0.724
LENGTH OF KERNAL GROVE (X7)	0.831	0.498

The above table shows the correlation between the variables and the components. The first component is highly influenced by the variables X_1 followed by X2, X5 and X4,second component is highly influenced by the variables X6 followed by X7,X4 and X2

FACTOR ANALYSIS

- It is also a dimension reduction technique requires a large sample size
- Is based on the correlation matrix of the variables involved
- There are many different methods that can be used to conduct a factor analysis
- Also different types of rotations that can be done after the initial extraction of factors

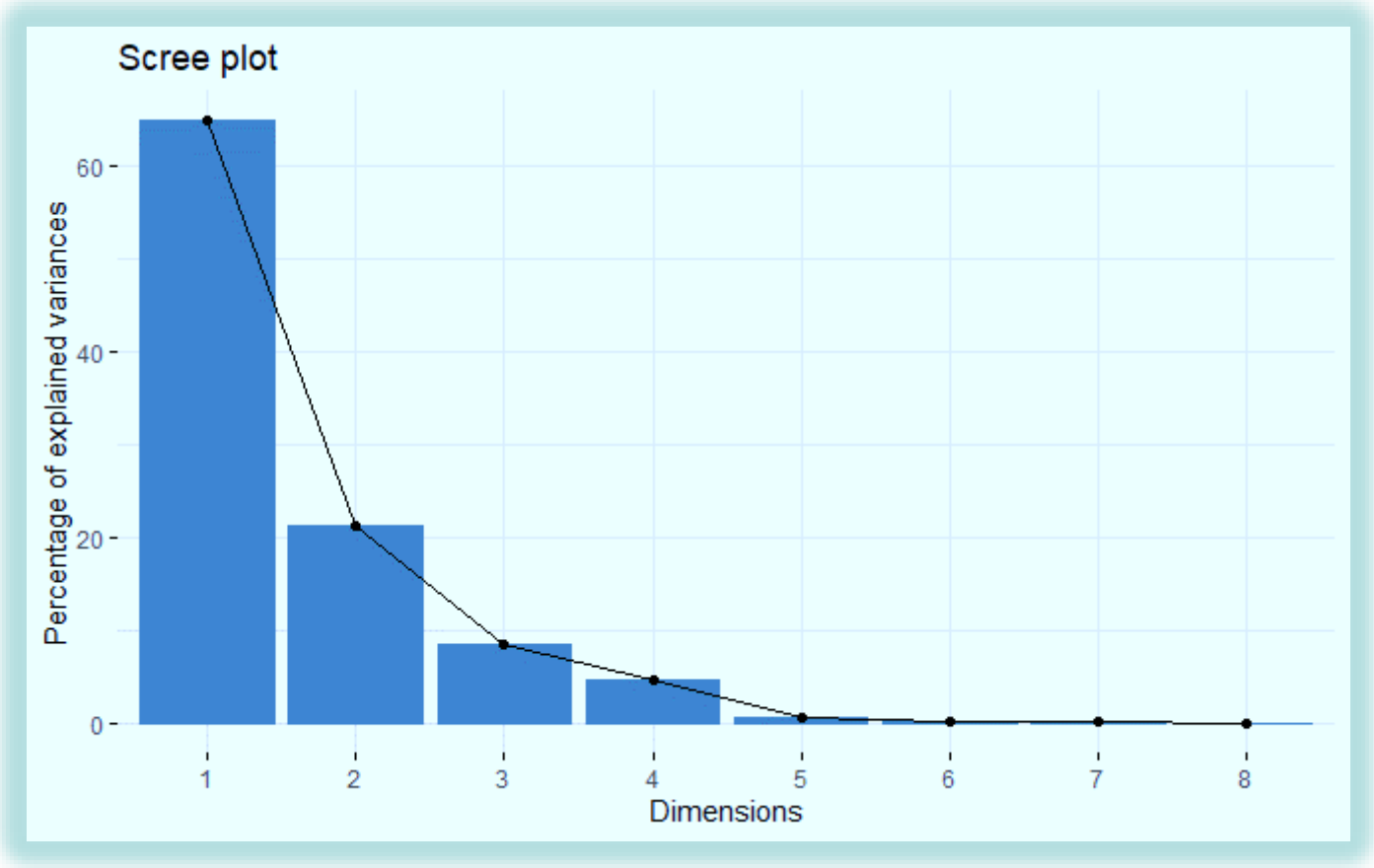
Kaiser Meyer Olkin and Bartlett's Test

Measures the strength of relationship among the variables .It is a measure of how suited the data for factor analysis, it varies between 0 and 1, and values closer to 1 are better

In the table KMO measure is 0.788, which is acceptable and therefore factor analysis can be done

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.788
Bartlett's Test of Sphericity	Approx. Chi-Square	3623.407
	Df	21
	Sig.	.000

SCREEPLOT



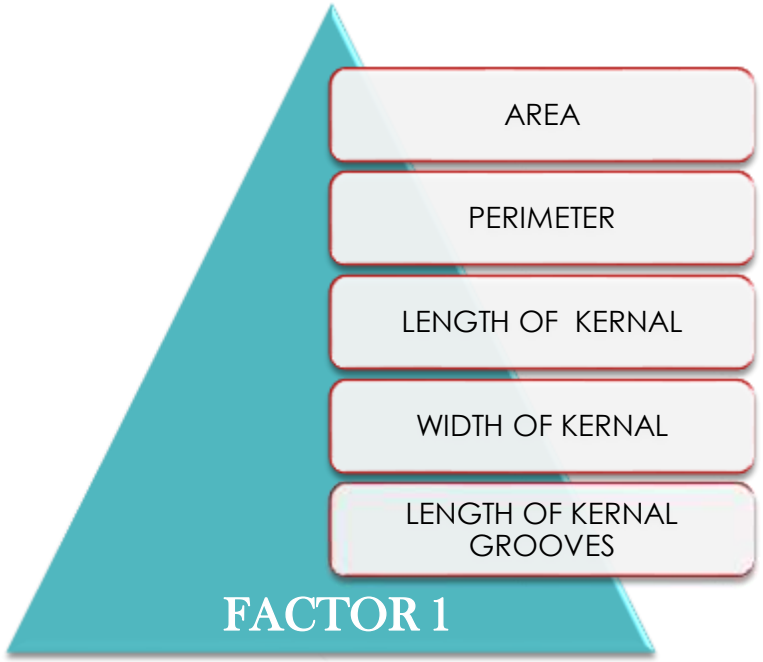
From the scree plot, we can extract two factors.

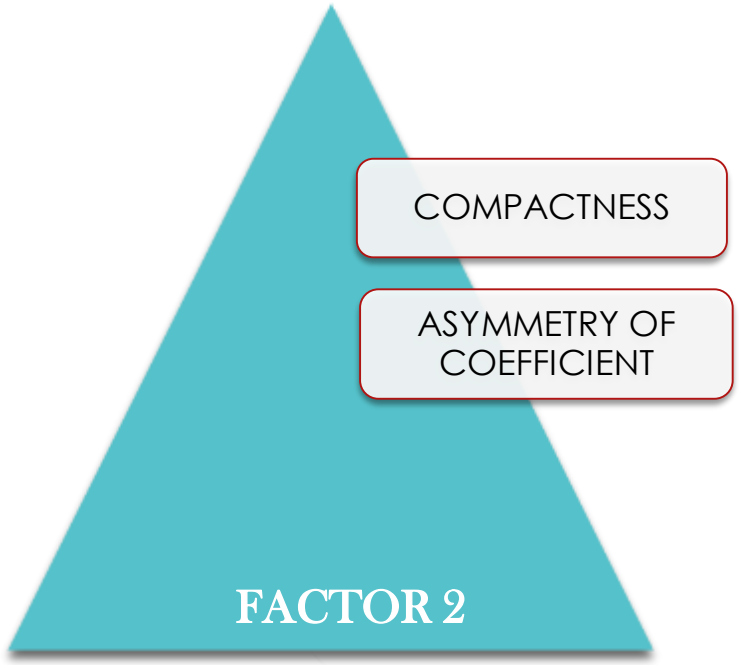
Variance Explained		
	RC1	**RC2**
Sum of square loadings	4.67	2.22
Proportion variance	0.58	0.28
Cumulative variance	0.58	0.86

From the variance table we can see that, **46.7**% of variance in the data is explained by **RC1**, **22.2**% **by RC2** That is, first two factors explain **86**% of variation of data set.

Rotated Component Matrix		
	Component	
	RC1	RC2
AREA	0.955	-0.287
PERIMETER	0.966	-0.240
COMPACTNESS	0.421	0.685
LENGTH OF KERNAL	0.965	-0.121
WIDTH OF KERNAL	0.884	-0.414
ASYMMETRY OF COEFFICIENT	-0.13	.791
LENGTH OF KERNAL GROVE	0.958	0.137
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.		

From the table of factor loadings, we can see which of the variables contribute more loadings corresponds to the two factors. The higher the absolute value of the loading, the more the factor contribute to the variable





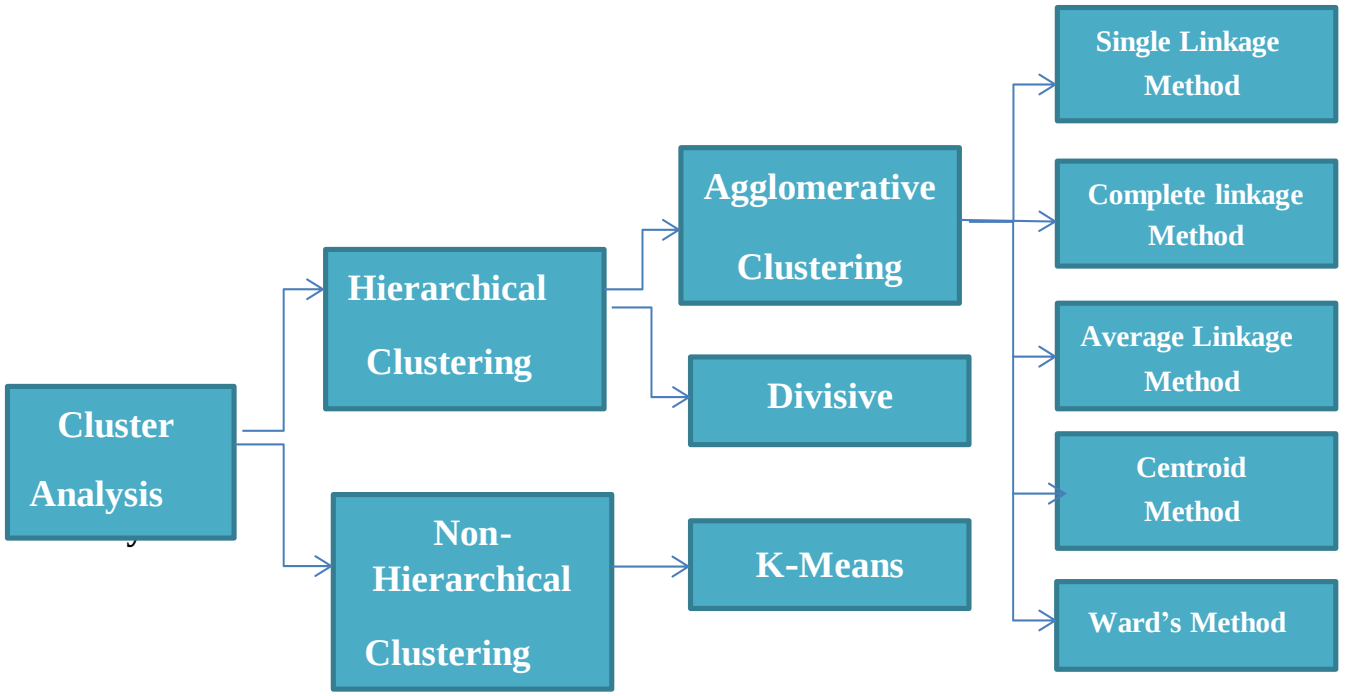
1) The first factor offers a point of attributes related to the overall size and dimensions of the seeds. This factor could represent a measure of seed size ,capturing the spatial dimensions that influence attributes like growth capacity. Seeds with higher factor score on this factor might exhibit larger physical dimension , potentially indicating adaptation for certain ecological niches.

2) The second factor offers a different lens through which to view seed attributes . This factor seems to capture attributes linked to the shape and symmetry of the seeds . High factor score on this factor might imply seeds with greater compactness and symmetry , which could relate to optimized packing efficiency , protection against environmental stress

CLUSTER ANALYSIS

- Clustering is a process of partitioning a set of data (or objects) in a set of meaningful subclasses, called clusters.
- It attempts to maximise the homogeneity of objects within the clusters while also maximise the heterogeneity between the clusters
- The basic criterion for a any clustering is distance. Objects that are near each other should belong to the same cluster, and objects that are far from each other should belong to different clusters.
- **By using this procedure ,we categorize the wheat seed data on the basis of attributes to make an inference , understanding for ultimately supporting the various aspects of wheat cultivation**

CLUSTERING PROCEDURES

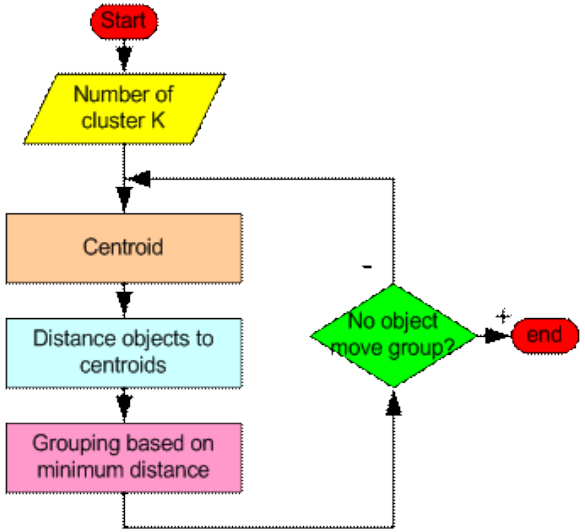


HOPKINS STATISTIC

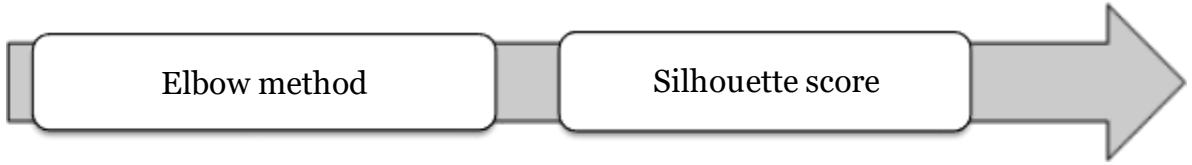
Before starting the analysis ensure that the dataset is suitable for clustering , for this we use Hopkins statistic which states that the value close to 1 indicates the data is highly clustered. In this case value is *0.75* which claims that dataset is significantly clusterable

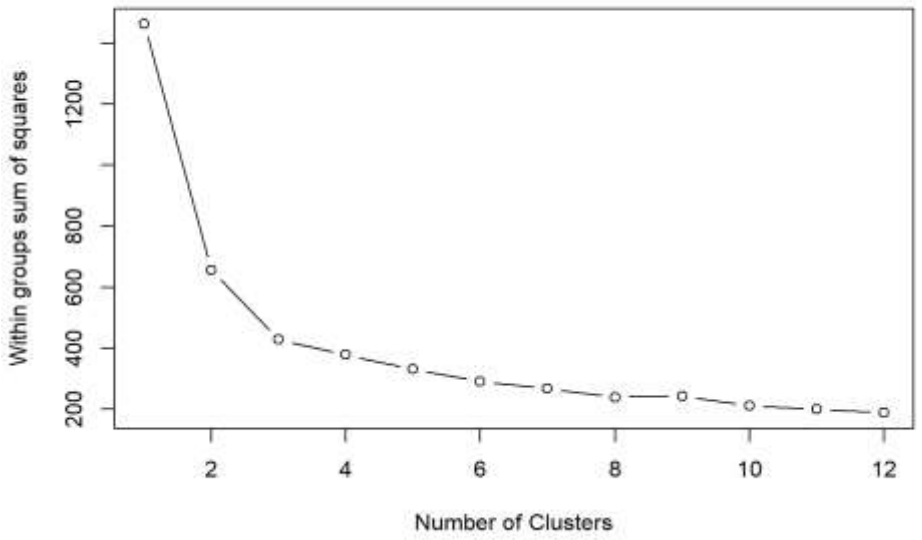
K-MEANS CLUSTERING

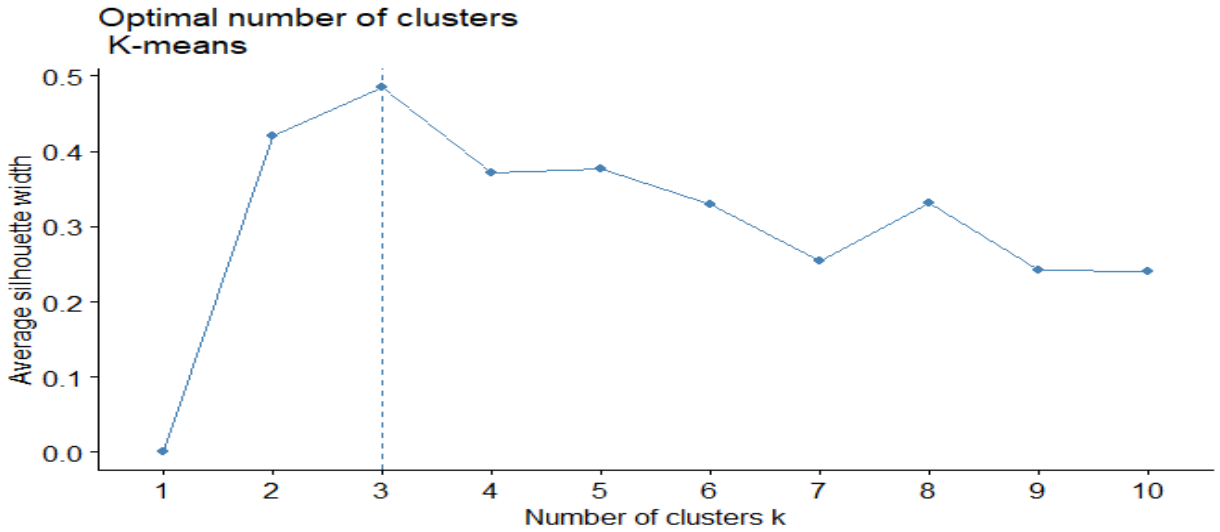
The k-means method assigns each item to the cluster having the nearest centroid. In its simplest version, the process is compared of the following flow chart



- The important step in K-means clustering technique is to decide the number of cluster



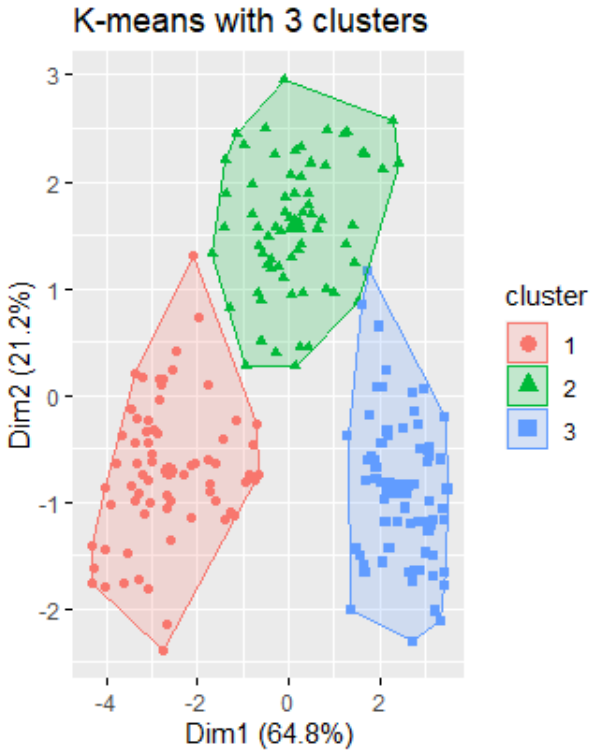


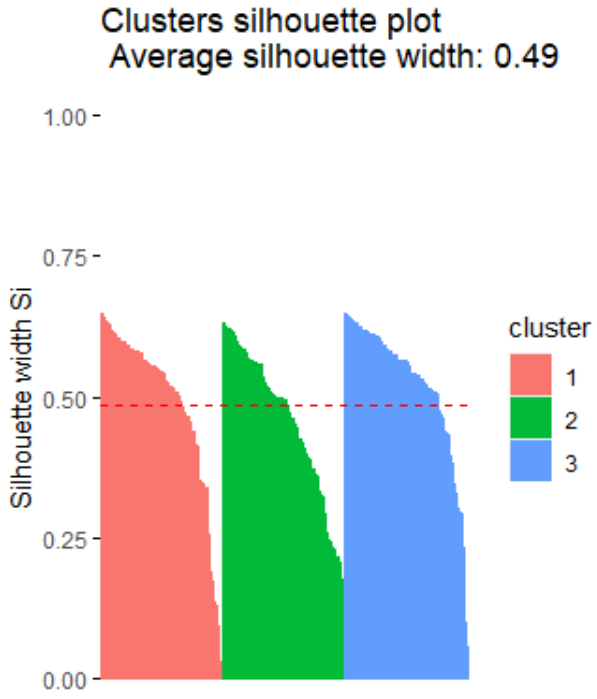


We see a bend (elbow) in the graph at k=3, therefore, 3 is the optimal number of clusters.

Another method for finding the optimal number of the clusters for means is **Silhouette statistic**
From silhouette statistic it is clear that the optimal number of cluster is k=3.

Hence , we are proceeding our k means with taking k=3





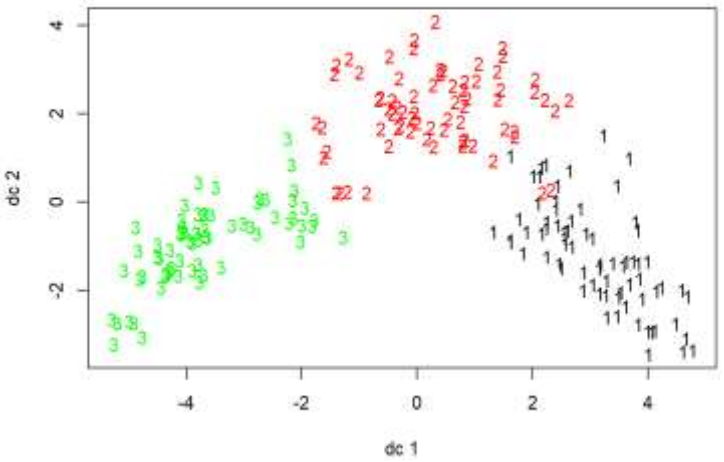
Clusters	1	2	3
Size	70	70	70
Avg. silhouette. Width	0.48	0.45	0.52

CLUSTER MEMBERSHIP

When k=3 the result looks better since we got a higher average silhouette width with no negative values, and any of the clusters are not overlapping

Average value of K-means Cluster

Area	Perimeter	Compactness	Length of kernel	Width of kernel	Asymmetry coefficient	Length of Grooves
1.20981	1.2212	0.5160033	1.2062015	1.1126	-0.05527	1.2689
-0.1878	-0.21702	0.397688	-0.3056103	-0.040821	-0.6684	-0.67545
-1.0219	-1.00418	-0.91369	-0.90059	-1.071797	0.72327	-0.5934



- Looking at the variables in the above table we can get the idea about which group of seeds are better.so according to average value table , the cluster one is the one with characteristics of seed have higher values, so that cluster one represents an excellent choice for wheat cultivation.

- After conducting average value and cluster plot analysis and validations,it becomes evident that the Canadian seed variety belongs to cluster one and that cluster one indeed offer highly favourable condition for wheat cultivation,or we can say confidently state that the Canadian seed variety is well suited for wheat cultivation

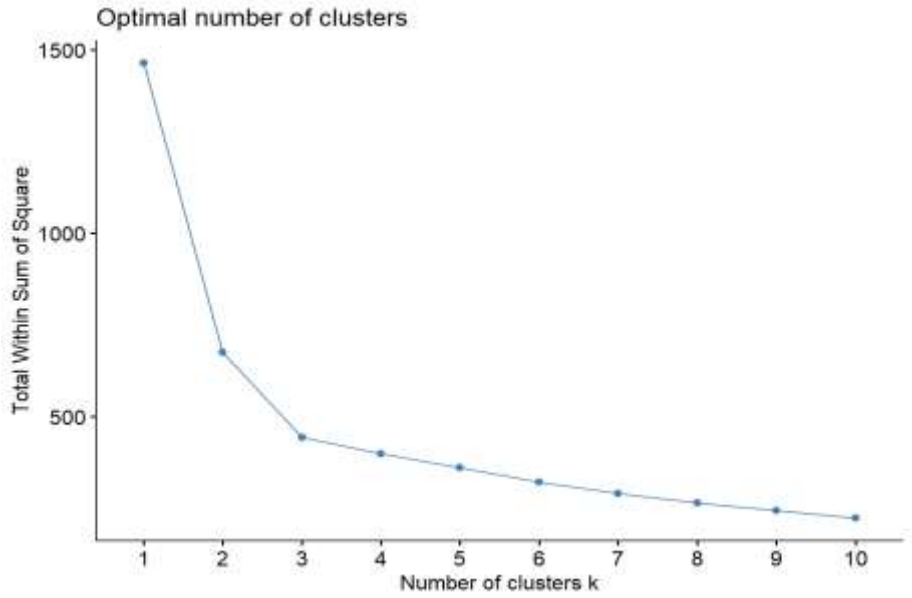
HIERARCHICAL CLUSTERING

- **Hierarchical clustering** is a method of cluster analysis that builds a hierarchy of clusters by either merging or dividing data points.
- Types of hierarchical clustering **:-** *Agglomerative* *Divisive*
- In this data , we are using agglomerative method from hierarchical clustering
- **Agglomerative method** :- it starts with many clusters and converges to fewer clusters. The proximity matrix can be formulated by using different methods.
- Types of Agglomerative Method :- Single linkage , Complete linkage , Average linkage , Ward linkage

There are several methods for clustering the data, to decide which method to use by calculating the agglomerative coefficient for each method which measures the amount of clustering structure found, value closer to 1 suggest strong clustering structure

Average	Single	Complete	Ward
0.8654	0.607	0.92332	0.98929

Choosing complete and ward's method which have value closer to 1.
Next to decide the optimal no. of cluster, using Elbow method we find the optimal number of clusters.

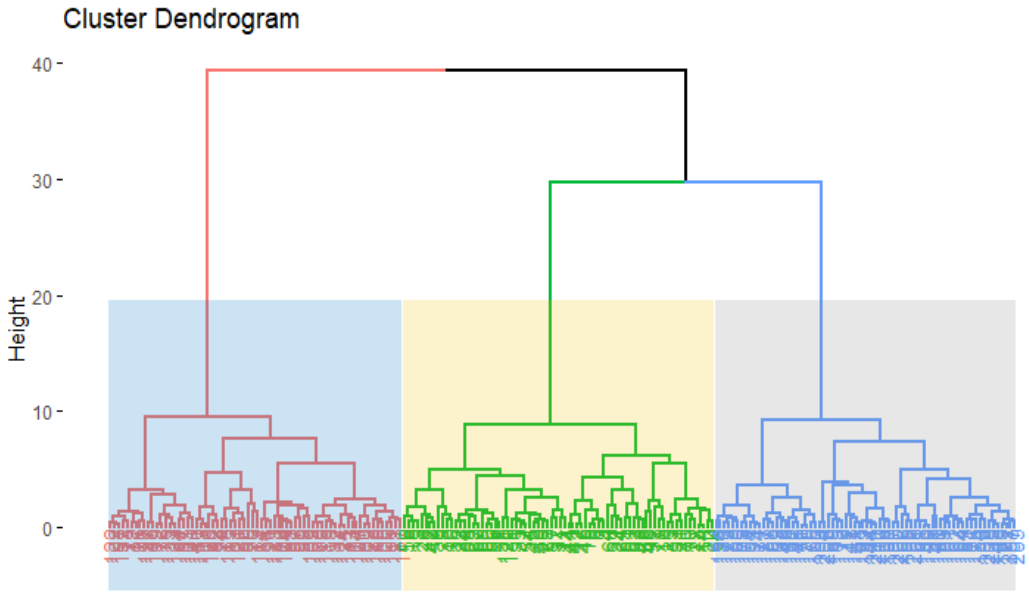


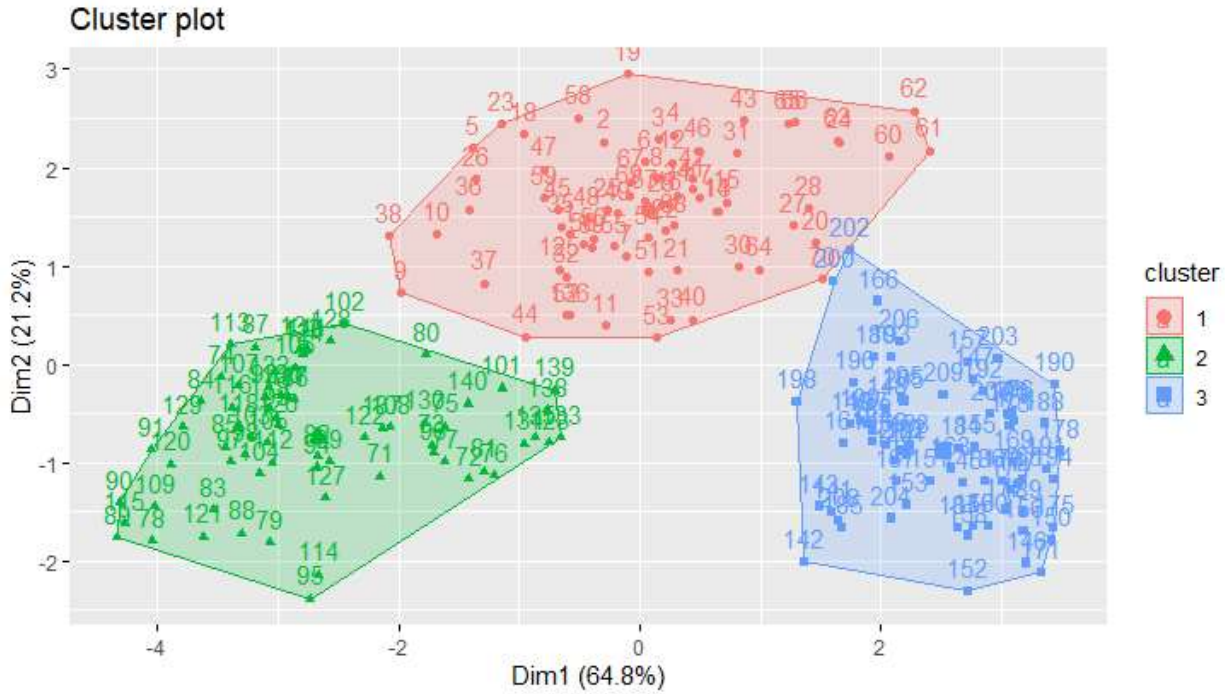
We see a bend (elbow) at k=3, therefore, 3 is the optimal number of clusters.

Hence , we are proceeding our hierarachical method with taking k=3

Clustering using wards and complete method as,

Wards Method

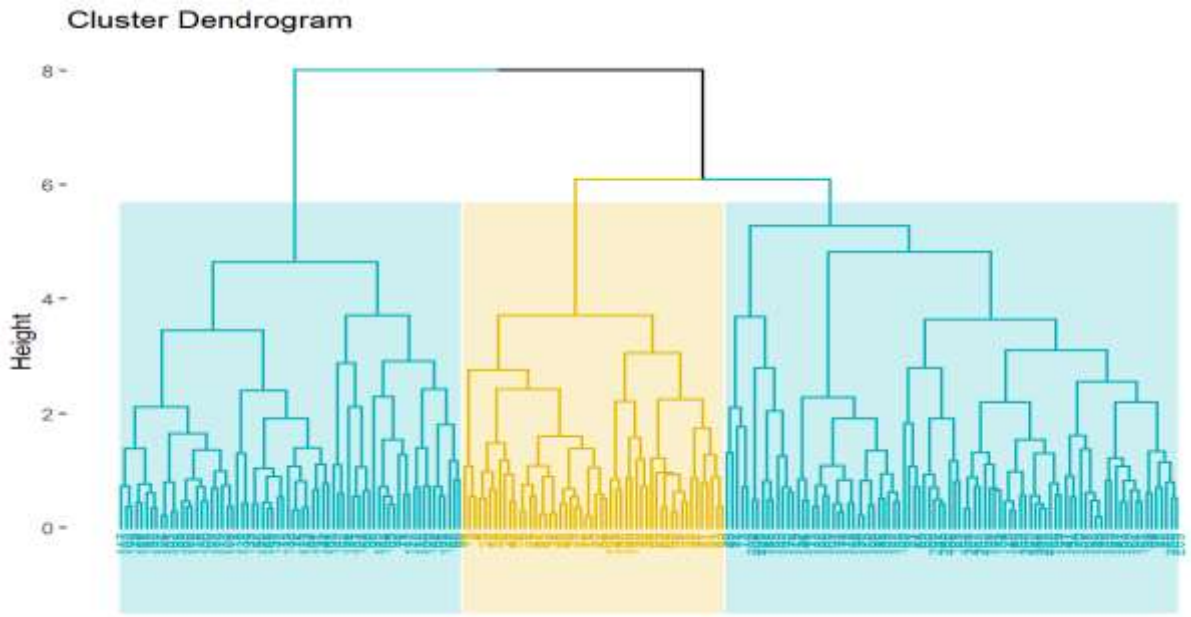


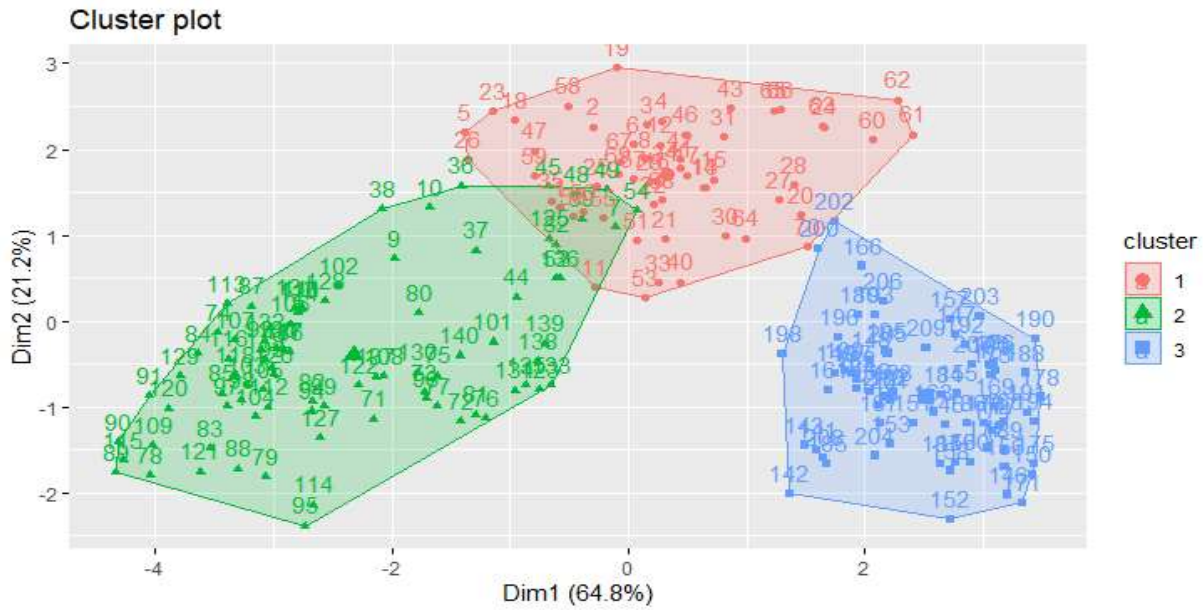


Here cluster 1 , 2 and 3 does not overlap, so it represents a distinct set of data points.

Therefore clustering is good in wards method

Complete Linkage Method





Here cluster 1 and cluster 2 overlaps each other and cluster 2 and cluster 3overlaps too.so it means that there are data points that belong to both clusters.

Silhouette Statistics

The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters . The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

On comparing silhouette statistic of both method Wards method has value `0.4836766` and complete linkage method has **0.4310212.** Wards method with a higher value of the Average Silhouette Width so it is the good method

FINAL CLUSTERS				
Member Complete Linkage	Member Wards Linkage			
		1	2	3
	1	64	4	2
	2	4	66	0
	3	5	0	65

CLUSTER MEMBERSHIP

AVERAGE VALUE OF WARD's METHOD

Area	Perimeter	Compactness	Length of kernel	Width of kernel	Asymmetry coefficient	Length of Grooves
-0.222	-0.2228	0.34667	-0.339230	-0.08512	-0.723630	-0.6549
1.211	1.2145	0.56712	1.195400	1.127899	-0.040599	1.23972
-1.022	-0.9971	-0.97027	-0.879316	-0.87931	0.83085	-0.58163

Looking at the variables in the above table we can get the idea about which group of seeds are better.so according to average value table of Ward's method , the cluster 2 is the one with characteristics of seed have higher values

In this case, Cluster 2, contains a combination of multiple classes ("kama" and "rose"), it can indeed make the interpretation and analysis of the clusters more challenging. When a cluster contains a mixture of different classes or patterns, it suggests that the clustering algorithm may not be effectively capturing the underlying structure in the data, or that the data itself may not exhibit clear and distinct clusters.

So ,by using the Adjusted Rand Index (ARI) to compare clustering methods and selecting K-means over Ward's method because it has a higher ARI value is a reasonable approach to make a choice between the two methods for your analysis.

The ARI is a commonly used external evaluation metric that measures the similarity between the true class labels (or ground truth) and the clustering results. A higher ARI value indicates better agreement between the clustering and the ground truth. In this case, K-means(0.8678) has a larger ARI compared to Ward's method(0.7136) .

THEREFORE K-MEANS IS PROVIDING BETTER CLUSTERING RESULTS.


THANK
YOU

MULTIVARIATE ANALYSIS ON WHEAT SEED DATA

Project submitted to

MAHATMA GANDHI UNIVERSITY, KOTTAYAM

In partial fulfillment of the requirement for awarding the degree of

MASTER OF SCIENCE IN STATISTICS

Submitted by

DEEPAK.G.KUTTY

Reg.no:210011014360

2021-2023

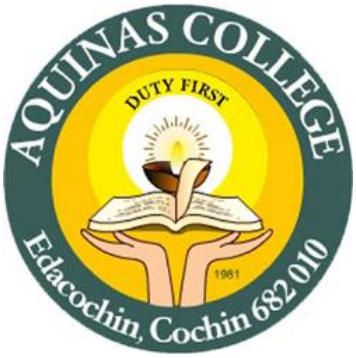
Guided by

Dr. JOSEPH JUSTIN REBELLO

Co-guided by

SHAFNA NOUSHAD

Guest Lecturer

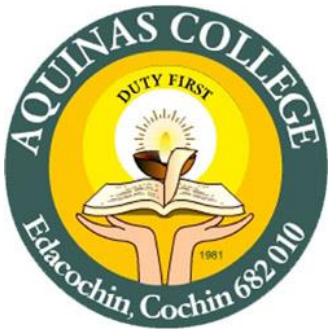


DEPARTMENT OF MATHEMATICS AND STATISTICS

AQUINAS COLLEGE, EDACOCHIN

MARCH 2023

AQUINAS COLLEGE, EDACOCHIN



CERTIFICATE

This is to certify that the project report entitled **MULTIVARIATE ANALYSIS ON WHEAT SEED DATA** is a bonafide work carried out by **DEEPAK.G.KUTTY (Reg. No. 210011014360)** during the academic year 2021-2023. This report is submitted to the Department of Mathematics and Statistics, Aquinas College, Edacochin, towards the partial fulfillment of the requirements for the award of degree of **Master of Science in Statistics** from Mahatma Gandhi University, Kottayam.

Head of the Department

Dr. Joseph Justin Rebello

Department of Mathematics and Statistics

External Examiner:

Place: Edacochin

Date: 14/09/2023

DECLARATION

I DEEPAK.G.KUTTY (Reg. No. 210011014360) hereby declare that the project entitled **MULTIVARIATE ANALYSIS ON WHEAT SEED DATA** is an original work submitted by me under the guidance of **Dr. JOSEPH JUSTIN REBELLO** and co-guidance of **Ms. SHAFNA NOUSHAD** Department of Mathematics and Statistics, Aquinas College, Edacochin in partial fulfillment of the requirements for the award of degree of Master of Science in Statistics.

Place**:** Edacochin **DEEPAK.G.KUTTY**

Date: 14/09/2023

ACKNOWLEDGEMENT

For any accomplishment or achievement, the prime requisite is the blessing of the almighty and it's the same that made this world possible. I bow to the lord with a grateful heart and prayerful mind.

It is with great pleasure that I express my sincere gratitude to Dr. Joseph Justin Rebello, Head of the Department of Mathematics and Statistics, for his guidance, overwhelming support, motivation and encouragement.

I also express my sincere gratitude to my beloved teachers Ms. Shafna Noushad and Ms. Mary Mrudhula P L, Guest Lecturer Department of Mathematics and Statistics, for their constant encouragement, advice and inspiring guidance throughout the course of this work.

I would like to acknowledge my deep sense of gratitude to all the faculty members of the department and my friends who helped me directly and indirectly through their valuable suggestions and self-criticisms, which came a long way in ensuring that this project becomes a success.

Finally, I am deeply indebted to me for not giving up in any situation and also thankful to my family for their support and prayers.

DEEPAK.G.KUTTY

MULTIVARIATE ANALYSIS ON WHEAT SEED DATA

Contents

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Multivariate analysis is a statistical procedure for the analysis of data involving more than one type of measurement or observation. It may also mean solving problems where more than one dependent variable is analyzed simultaneously with other variables.

Multivariate analysis is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time. In design and analysis, the technique is used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest.

Uses of multivariate analysis includes,

- Design for capability
- Inverse design, where any variable can be treated as an independent variable
- Analysis of alternatives, the selection of concepts to fulfill a customer need.
- Analysis of concepts with respect to changing scenarios.
- Identification of critical design-drivers and correlations across hierarchical levels.

The measurements and analysis of dependence between variables, between sets of variables and between variables are fundamental to multivariate analysis. Multivariate analysis can be complicated by the desire to include physics based analysis to calculate the effect of variables for a hierarchical “system of systems”. Often studied that wishes to use multivariate analysis is called by the dimensionality of the problem.

Multivariate analysis relies on several key assumptions to ensure the validity and reliability of the results. The specific assumptions may vary depending on the multivariate technique used, but here are some common assumptions in multivariate analysis:

- Linearity: Many multivariate techniques assume a linear relationship between variables. That is, the relationships between variables are assumed to be linear, and any non-linear relationships may not be accurately captured by the analysis. Some techniques, such as linear regression or principal component analysis, explicitly assume linearity.

- Normality: Multivariate analysis often assumes that the variables follow a multivariate normal distribution. This assumption means that the data should be normally distributed within each variable and that the joint distribution of the variables is multivariate normal. Violations of this assumption can impact the accuracy and interpretability of results.

- Homoscedasticity: Homoscedasticity assumes that the variance of the variables is constant across different levels or combinations of variables. In other words, the spread of the data is consistent across the range of values. Violations of this assumption, such as heteroscedasticity, may affect the validity of statistical tests or estimation procedures.

- Independence: Multivariate analysis assumes that the observations are independent of each other. Each observation is assumed to be unrelated to other observations in the dataset. Violations of independence, such as autocorrelation or clustered data, may require specialized techniques or adjustments to account for the dependence structure.

- Absence of Outliers: Multivariate analysis assumes the absence of influential outliers that can disproportionately affect the results. Outliers can distort parameter estimates, affect the distributional assumptions, and influence the relationships among variables.

- Equal Covariance Matrix: Some multivariate techniques, such as discriminant analysis, assume that the covariance matrices of the variables are equal across groups or categories. This assumption is necessary to make accurate comparisons and classifications based on the covariance structure.

- Sample Size: The sample size in multivariate analysis should be large enough to ensure the validity of the results. Adequate sample size helps ensure the stability and precision of estimates, reduces the risk of overfitting, and provides robustness to violations of assumptions.

- Multivariate techniques are complex and involve high level mathematics that require a statistical program to analyze the data. These statistical programs can be expensive for an individual to obtain. One of the biggest limitations of multivariate analysis is that statistical modelling outputs are not always easy for students to interpret. For multivariate techniques to give meaningful results, they need a large sample of data; otherwise, the results are meaningless due to high standard errors.

1.2 **ABOUT THE DATA**

The data used for the analysis is a secondary data from Institute of Agrophysics of the Polish Academy of Sciences in Lublin .The datasets are uploaded in Kaggle also.

It contains measurements of various attributes of wheat seeds, classified into three different classes: Kama(1), Rose(2), and Canadian(3). The dataset is often used for classification algorithms to distinguish between these three classes based on their characteristics.

The dataset provides a total of seven attributes or features that describe each wheat seed sample,and classified into 3 classes. The attributes include:

X1-Area: The area of the seed, measured in square millimeters.

X2-Perimeter: The perimeter of the seed, measured in millimeters.

X3-Compactness: A measure of the seed's compactness, calculated as the ratio of perimeter to area.

X4-Length of Kernel: The length of the kernel of the seed, measured in millimeters.

X5-Width of Kernel: The width of the kernel of the seed, measured in millimeters.

X6-Asymmetry Coefficient: A measure of the seed's asymmetry.

X7-Length of Kernel Groove: The length of the kernel groove, measured in millimeters.

1.3 **OBJECTIVE OF THE STUDY**

In this project,we aim to uncover the relationships , dependencies and to enchance our understanding of wheat seed characterstics and their potential implications for agriculture by employing various multivariate analysis technique;

Exploratory Data Analysis(EDA): Performing EDA on the wheat seeed dataset to gain insights into the data distribution ,summary statistics , missing values and potential outlier.This step will help in understanding the dataset and identifying any preprocessing steps required

Principal Component Analysis(PCA):-Conducting PCA on the dataset to transform the orginal variables into a new set of uncorrelated variables called principal component.PCA can help to identify the most important components that capture the maximum variance in the data.This can aid in dimensionality reduction and visulalising the dataset in a lower-dimensional space

Clustering:-Utilising the clustering alogorithms such as k-means ,hierarchical clustering,to group the wheat seed samples based on their similarities in attribute values.clustering can help identifying distinct patterns or groups within the dataset,potentially revealing different varieties or subclasses of wheat seeds.This can aid in understanding the inherent structure of the data and discovering meaningful clusters.

Factor Analysis:-Applying Factor analysis techiniques to identify latent factor or underlying dimensions in the dataset.Factor analysis can help reduce the dimensionality of the dataset by extracting meaningful factors that explain the observed variation in the wheat seed attributes.This can provide a deeper understanding of the relationships between the attributes and potentially simplify subsequent analysis

CHAPTER 2

DATA CLEANING AND VISUALIZATION

2.1 EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is the first step in the data analysis pro-cess. In statistics, exploratory data analysis is an approach of analyzing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods. EDA relies on data visualizations that enable researchers to identify and define patterns and characteristics in the dataset that they otherwise would not have known to look for. EDA was originally developed by John Tukey, an American mathematician, in the 1970s. It's often thought of as more of a philosophical approach to data analysis than a statistical method. EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. The main purpose of EDA is to help look at data before making any assumptions.

Categorical variable: Weed sead dataset

Numerical variable: Area, Perimeter, Compactness ,Length of kernel ,Width of kernel , Asymmetry Coefficient, Length of Kernel Groove

Area	Perimeter	Compactness	Length	Width	A_Coef	Length_groove	target
15.26	14.84	0.871	5.763	3.312	2.221	5.22	1
14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
14.29	14.09	0.905	5.291	3.337	2.699	4.825	1
13.84	13.94	0.8955	5.324	3.379	2.259	4.805	1
16.14	14.99	0.9034	5.658	3.562	1.355	5.175	1
14.38	14.21	0.8951	5.386	3.312	2.462	4.956	1
14.69	14.49	0.8799	5.563	3.259	3.586	5.219	1
14.11	14.1	0.8911	5.42	3.302	2.7	5	1
16.63	15.46	0.8747	6.053	3.465	2.04	5.877	1
16.44	15.25	0.888	5.884	3.505	1.969	5.533	1

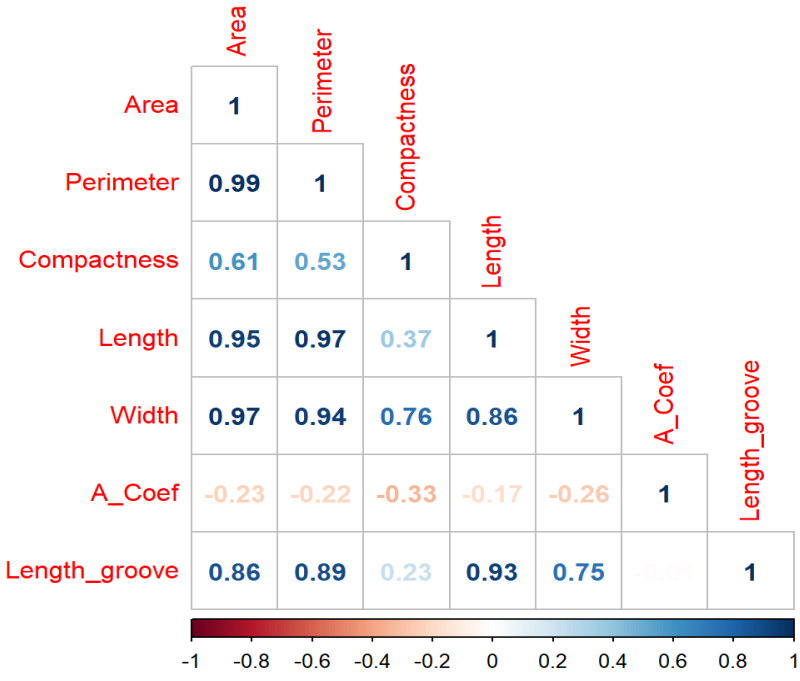
Missing values

For the purpose of this study zero values have been classified as "missing" values and have been replaced with the mean of the column. There are no missing values so we do not have, to modify anything related with that matter in our dataset. It is also possible to appreciate that some features tend to have larger values than others, so to avoid a problem in which some features come to dominate solely.

Area	Perimeter	Compactness	Length	width	Asymmetry Coefficient	Length of Kernel Groove	Target
0	0	0	0	0	0	0	0

2.2 CORRELATION MATRIX

A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. The matrix is a table in which every cell contains a correlation coefficient, where 1 is considered a strong relationship between variables, 0 a neutral relationship and -1 a not strong relationship.

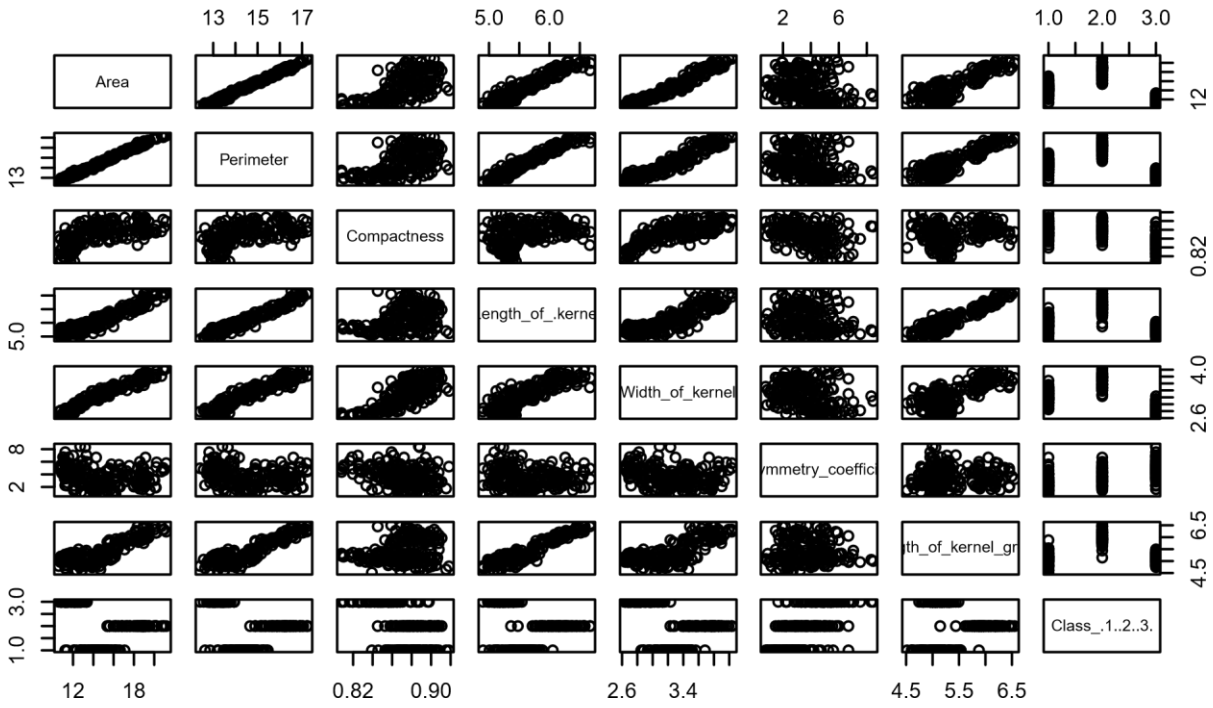


Based on the correlation matrix, it is obvious that the variables are positively as well as negatively correlated.

Area and Perimeter are highly positive correlated, as well as Area and Width are highly positive correlated as Length and Perimeter, Length and Area, Width and Perimeter, Length and Length of groove are all highly positively correlated. The goal of dimension reduction analysis is to reduce dimension and represent these highly correlated variables with fewer new transformed variables.

2.3 SCATTER PLOT

It is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables – one plotted along the x-axis and the other plotted along y-axis. Scatter plots are used when you want to show the relationship between two variables. Scatter plots are sometimes called correlation plots because they show how to two variables are correlated



CHAPTER 3

MULTIVARIATE TECHNIQUES

3.1 **PRINCIPAL COMPONENT ANALYSIS**

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. If there are n observations with p variables, then the number of distinct principal components is min(n-1,p). This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. A principle component analysis is concerned with explaining the variance-covariance structure of a set of variables through a *few linear* combinations of these variables. Its general objectives are

(1) Data reduction and

(2) Interpretation

Data reduction techniques are applied to aggregate the information attained in large sets of data into manageable information nuggets. In a multivariate study most oftenly the variables under study are highly correlated. So that some variables may give the same information as that by some others. In such cases one may think about Principle Component analysis (PCA), Factor analysis (FA) etc.

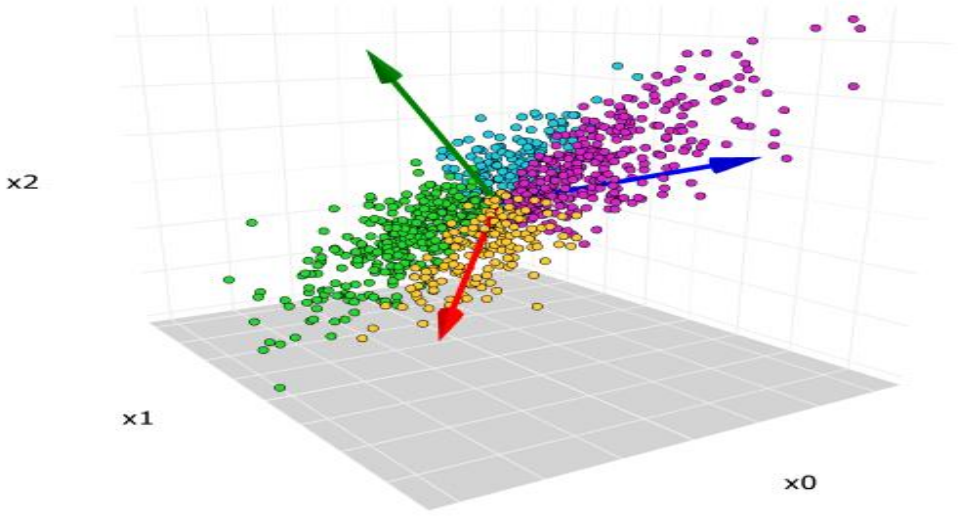
Principle component may be useful to transform the original set of variables to a new set of uncorrelated variables. These new variables are called Principle components which are the normalized linear combination of the original variables and are derived in the decreasing order of importance. So that the first principle component (PC) accounts for the maximum variation of the

original data. The is the maximum variation is explained by the principle component and it is variance turns out to be the largest characteristic root or eigen value of Σ. Similarly, the second principle component is the normalized linear combination of original variables having the second maximum variance. That is the second largest characteristic root and the second principle component should be uncorrelated with first principle component. Finally, one may discard the linear combination with the smallest variance so that the data may be projected into a lower dimensionality space variance. (Number of principle components finally selected will be less than the number of original variables).

PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It's often used to visualize genetic distance and relatedness between populations. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z scores) the data matrix for each attribute. The results of a PCA are usually discussed in terms of *component scores, sometimes called factor scores* (the transformed variable values corresponding to a particular data point), and *loadings* (the weight by which each standardized original variable should be multiplied to get the component score).

PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a projection of this object when viewed from its most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

PCA is closely related to factor analysis. Factor analysis typically incorporates more domain specific assumptions about the underlying structure and solves eigenvectors of a slightly different matrix.



PCA is also related to canonical correlation analysis (CCA). CCA defines coordinate systems that optimally describe the cross-covariance between two datasets while PCA defines a new orthogonal coordinate system that optimally describes variance in a single dataset.

3.1.1 BASIC DEFINITIONS

Characteristic equations

Let A be a $P \times P$ matrix. The scalars $\lambda_1, \lambda_2, \ldots \ldots \lambda_P$ which satisfies the polynomial equation, $|A - \lambda \mathrm{I}| = 0$ is called the characteristic equation and scalars $\lambda_1, \lambda_2, \ldots \ldots \lambda_P$ are called the characteristic root or eigen values of A.

Eigen vector

Let A be a $P \times P$ matrix and λ be a an eigen value of A. If $X \neq 0$ is a $P \times 1$ vector satisfying the characteristic equation $\mathrm{AX} = \lambda \mathrm{X}$, then X is called the eigen vector associated with eigen value λ.

Normalised Eigen value

If we take $\beta = \frac{X}{\sqrt{X'X}}$ so that $\beta'\beta = 1$, then β is called normalised eigen vectors of A.

Thus A has p-pairs of eigen value and eigen vectors namely

$(\llbracket \lambda \rrbracket \_i \ , \beta\hat{}((i)\ )); i = 1,2,3, \ldots . . P$

If in addition the eigen values are so choosen such that $\beta^{i'}\beta = 1$ and $\beta^{(i)'}\beta^{(j)} = 0,$ for all

$i \neq j$ then, β's are called normalised orthogonal eigen vectors

Spectral Decomposition

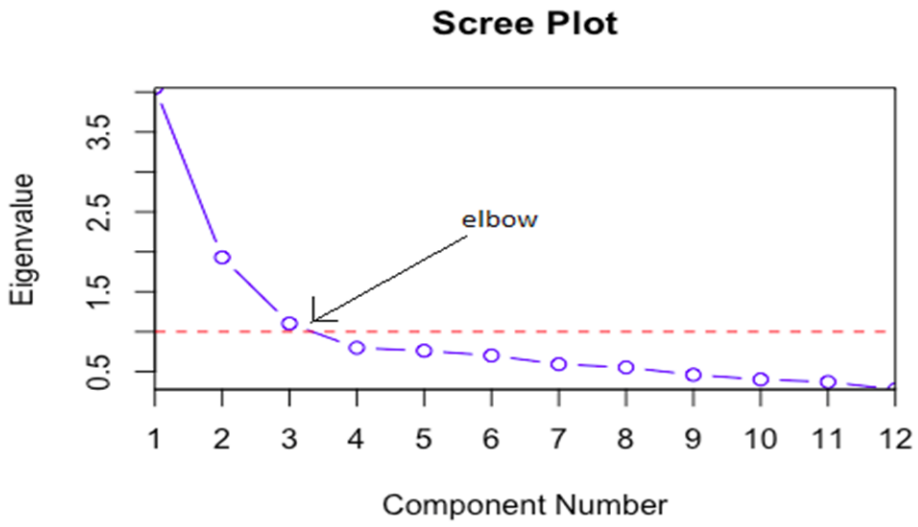
Let $B = \left(\beta^{(1)}\ \beta^{(2)}\ \ldots \ldots \ldots \beta^{(P)}\right)$ be a $P \times P$ matrix of the normalised orthogonal eigen vectors of A then if, $A = B\lambda B'$, where $\lambda = \text{diag}(\lambda_1, \lambda_2, \ldots \ldots \lambda_P)$ is called the spectral decomposition of A.

$$A=\sum_{i=1}^{p} \lambda_i \beta^i \beta^{(i)'}$$

That is, it allows any quadratic function into a weighted sum of squared linear functions involving vectors that are mutually orthogonal, the weights are the eigen values.

3.1.2 SCREE PLOT

Decision regarding the number of principle components to be taken in any data analysis is decided graphically by a scree plot. It is a two dimensional graph with principle component of the X-axis and eigen -values on the Y-axis. The principle components are typically arranged in descending order of their variance in a scree plot. That is, the first variable account for the maximum variance the second variance is second maximum and so on. So at the end the decision is based on the subjectivity of the person who do the analysis and evident *bent* or *elbow* in the scree plot decides the number of principle component. The term '*scree*' is taken from the word for the rubble at the bottom of the mountain.



3.1.3 PROCEDURES OF EXTRACTION OF THE PRINCIPAL COMPONENT

Let X be a P component random variable whose mean is assumed to be μ and dispersion matrix Σ, where Σ is a real positive matrix. The equation for the characteristic root and the corresponding characteristic vector is given by

$$\Sigma X = \Lambda X \qquad \ldots\ldots\ldots (1)$$

According to Hotelling's iterative procedure we start with an initial $P \times 1$ vector X_0 which is not orthogonal to e_1 , the characteristic vector corresponding to the largest characteristic root λ_1 of Σ .

Define $X_i = \Sigma Z_{i-1}\ ; i = 1,2,3, \ldots \ldots \ldots P$

$$Z_i = \frac{\text{Xi}}{\sqrt{XiXi}}\ ; i = 1,2,3, \ldots \ldots \ldots P \qquad \ldots\ldots\ldots (2)$$

It can be shown that,

$\lim_{i\to\infty} z_i = \pm e_i$

$\lim_{i\to\infty} X_i' X_i = \lambda_1^{\,2} \qquad \ldots\ldots\ldots (3)$

To find the second characteristic root and the corresponding characteristic vector we define,

$$\Sigma_2 = \Sigma - \lambda_1 e_1 e_1'$$

Now to find λ_1 and e_2 we use the same iterative procedure toΣ_2. Repeat the steps (2) and (3)

Thus the non-zero Eigen values of Σ are $\lambda_1, \lambda_2, \lambda_3, \ldots \ldots \lambda_P$ with Eigen vectors $e_1, e_2, e_3, \ldots \ldots e_P$

So we get the linear combination $Z_i = e_i' X$

That is,

$$Z_1 = e_{11}X_1 + e_{12}X_2 \ldots \ldots \ldots e_{1P}X_P$$

$$Z_2 = e_{21}X_{1+}e_{22}X_2 \ldots \ldots \ldots e_{2P}X_P$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$Z_P = e_{P1}X_{1+}e_{P2}X_2 \ldots \ldots \ldots e_{PP}X_p \qquad \ldots\ldots\ldots\ldots (4)$$

With the condition $V(Z_1) \geq V(Z_2) \ldots \ldots V(Z_p) \geq 0$ …… (5)

The linear combination (4) is called principal component satisfying (5)

3.1.4 METHOD OF EXTRACTION OF PRINCIPAL COMPONENT

Let X be P×1 random vector with covariance matrix Σ. Since, we are interested in variances and covariances no need to assume the statistical model or the distribution of the original variables.

Step 1: First Principal component is the normalised linear combination of the original variables X. Let $u_1 = l_1'X$; $l = (l_{11}, l_{12}, \ldots l_{1p})'$ which maximises $V(u_1) = l_1'\sum l_1$ subject to $l_1'l_1 = 1$.

Step 2: The second principal component is $u_2 = l_2' X$ such that $V(u_1) = l_2'\sum l_2$ is the second maximum subject to $l_2' l_2 = 1$ and $cov(u_1, u_2) = l_1'\sum l_2 = 0$.

Step 3: The i^{th} principal component is $u_i = l_i'X$ which maximises that $V(u_i) = l_i'\sum l_i$ subject to $l_i'l_i = 1$ and $cov(u_i, u_k) = l_i'\sum k = 0; i \neq k$ and $k = 1,2, \ldots i - 1$.

3.1.5 . ADVANTAGES AND DISADVANTAGES OF PRINCIPAL COMPONENT ANALYSIS

ADVANTAGES:

- **Easy to compute:** PCA is based on linear algebra , which is computationally easy to solve by computers

- **Counteracts the issue of high dimensional data:** High dimensional data causes regression based algorithms to overfit easily. By using PCA beforehand to lower the dimension of the dataset , we prevent the predictive algorithms from overfitting.

- **Speeds up other machine learning algorithms:** Machine learning algorithms converge faster when trained on principal components instead of the original dataset.

DISADVANTAGES:

- **Low interpretability of principal components:** They are linear combinations of the features from the original data , but they are not as easy to interpret. For example , it is difficult to tell which are the most important features in the dataset after computing principal components.

- **The trade-off between information loss and dimensionality reduction:** Although dimensionality reduction is useful , it comes at a cost. Information loss is a necessary part of PCA. Balancing the trade-off between dimentionality reduction and information loss is unfortunately a necessary compromise that we have to make when using PCA.

3.2 FACTOR ANALYSIS

In Multivariate analysis to study the complex relationship among the variables, we need a technique that is used to reduce a large number of variables into fewer numbers of factors; such an effort is called Factor Analysis. This is a procedure for summarization and reduction of Data. Factors are the random quantities, if possible the covariance relationship among many variables in terms of a few underlying but unobservable. This technique extracts maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis. It also used by researchers to investigate concepts that are not easily measured directly by collapsing a large number of variables into a few interpretable underlying factors.

Factor analysis was developed originally for the analysis of scores on mental tests; however, the methods are useful in a much wider range of situations such as analyzing sets of tests of attitudes, sets of physical measurements and sets of economic quantities. Factor can be considered as an extension of Principal component analysis. Both can be viewed as attempts to approximate the covariance matrix $\sum$. The approximations based on Factor analysis are more elaborate.

There are two major types of Factor analysis are

1. Exploratory Factor analysis(EFA)
2. Confirmatory Factor analysis(CFA)

Also, there are many methods for the estimation of parameters namely, The principal component method, Method of Least Squares, the Maximum likelihood method and Alpha factoring. In the present study, the method of Principal components is used.

The solution obtained can be rotated in order to simplify the interpretation of factors. Rotation is used to improve the interpretability and scientific utility of the solution. For any type of rotation, a Factor score coefficient matrix, used to estimate scores on factors from scores on observed variables for each individual. All produce factor scores that are correlated but not perfectly, with the factors.

3.2.1 DISTINGUISH BETWEEN EFA AND CFA

EFA	CFA
❖ Actual theoretical model is used to explore the dimentionality of the data ❖ Number of factors are decided by Principle component Analysis ❖ Allows all the variables/items to load on all the factors ❖ Use of MLE's of factor loading is allowed	❖ Hypothesized model is test against the actual data ❖ Researcher have to specify the number of factors in advance ❖ A particular factor structure is to be specified in which the researchers indicate which items to load on which factors ❖ Use of MLE's of factor loading is not allowed

Note :

1) In any Factor Analysis there are same number of factors as the number of variables. Each factor captures a certain amount of overall variance in the observed variables and the factors are listed in the order of the variations they explained.
2) Eigen value is a means of showing how much of variance of the observed variable that a factor explains. Any factor with eigen value explains more variance than a single observed variable.

3.2.2 GENERAL PURPOSE AND DESCRIPTION

Factor analysis has provoked rather turbulent controversy throughout its history. Its modern beginnings lie in the early 20 th century attempts of Karl Pearson,Charles Spearman and others to define and measure intelligence. A major use of Factor analysis is in the development of objective tests for measurements of personality and intelligence.

The essential purpose of Factor analysis is to describe, if possible the covariance relationship among many variables in terms of a few underlying, but unobservable, random quantities called Factors. Suppose variables can be grouped by their correlations, ie, if variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. Then it is conceivable that each group of variables represents a single underlying construct or factor that is responsible for the correlations. Thus, the Factor analysis gives a description or explanation of the interdependence of a set of variables in terms of factors without regard to observed variability.

The correlation matrix produced by the observed variables is called observed correlation matrix. The correlation matrix produced from factors is called the reproduced correlation matrix. The difference between the observed and reproduced correlation matrices is the residual correlation matrix. In a good Factor analysis, correlations in the residual matrix are small.

3.2.3 THE ORTHOGONAL FACTOR MODEL

The observable random vector **X** with **p** components has mean **μ** and covariance matrix **∑.** The factor model postulates that **X** is linearly dependent upon a few unobservable random variables $F_1, F_2, \ldots, F_m$ called common factors and p additional sources of variation $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \ldots, \mathcal{E}_p$ called errors or specific factors.

In particular, the Factor analysis model is

$$
\begin{aligned}
X_1-\mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \mathcal{E}_1 \\
X_2-\mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \mathcal{E}_2 \\
&\vdots \\
X_p-\mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \mathcal{E}_p
\end{aligned}
\qquad (6)
$$

Or in matrix notation,

$$\mathbf{X} - \boldsymbol{\mu} = LF + \boldsymbol{\varepsilon} \quad \text{.....(7)}$$

Where $X-\mu$ is a p x 1 vector, L is a p x m matrix , F is a m x 1 vector and $\mathcal{E}$ is a p x 1 vector.

The coefficient l_{ij} is called the loading of the i^{th}variable on the j^{th} factor, so that the matrix L is the matrix of factor loadings. The i^{th} specific factor ε_i is associated only with the i^{th} response X_i. The p deviations $X_1-\mu_1$, $X_2-\mu_2$,….$X_p-\mu_p$, are expressed in terms of $p+m$ random variables $F_1,F_2,\ldots.F_m,\varepsilon_1,\varepsilon_2\ldots..\varepsilon_p$ which are unobservable. This distinguises the factor model expressed in equation (7) from the regression model in the independent variables observed.

With so many unobservable quantities, a direct verification of the factor model from observations on $X_1,X_2,\ldots.X_P$ is hopless.however with some additional assumptions about the random vectors F and ε, the model

$$\mathbf{X}-\mu = LF+\varepsilon ,$$

Implies certain covariance relationships can be checked.

The assumptions are:

i) $\mathrm{E}(F) = 0_{m\times 1}$ $\mathrm{Cov}(F) = \mathrm{E}(FF') = I_{m\times m}$

ii) $\mathrm{E}(\varepsilon) = 0_{p\times 1}$ $\mathrm{Cov}(\varepsilon) = \varphi_{p\times p} = \begin{bmatrix}\varphi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi_p\end{bmatrix}$ .....(8)

iii) F and ε are independent.

ie, $\mathrm{Cov}(F,\varepsilon) = \mathrm{E}(\varepsilon F') = 0_{p\times m}$

these assumptions and the relation $\mathbf{X}-\mu = LF+\varepsilon$ constitute the orthogonal factor model. The factor analysis model with the above assumptions is called orthogonal Factor analysis model.

3.2.4 METHODS OF ESTIMATION

Given observations $X_1,X_2,\ldots X_n$ on p generally correlated variables. The sample covariance matrix S is an estimator of the unknown population covariance matrix $\sum$. If the off-diagonal elements of **S** are small or those of the sample correlation matrix R essentially zero, the variables are not related, and a factor analysis will not be useful. In such circumstances, the specific factors

play a dominant role, but the aim of factor analysis is to determine a few important common factors.

If $\sum$ appears to deviate significantly from a diagonal matrix, then a factor model can be entertained and the initial problem is one of estimating the factor loadings I_{ij} and specific variance φ_i.

The two methods of parameter estimation in factor analysis are:

i) Maximum likelihood estimation

ii) Principal component analysis.

3.2.5 MAXIMUM LIKELIHOOD ESTIMATION

If the common factors F and the specific factors ε can be assumed to be normally distributed, then the maximum likelihood estimates of the factor loadings and the specific variances may be obtained. When F_j and ε_j are jointly normal, the observation

$X_j - \mu = LF_j + \varepsilon_j$ are then normal. Then the Likelihood function is given by,

$$L(\mu, \sum) = \frac{1}{(2\prod)^{(np)/2}|\sum|^{\frac{n}{2}}} e^{-1/2\{\sum_{j=1}^{n} (x_j-\mu)'\sum^{-1}(x_j-\mu)\}}$$

$$= \frac{1}{(2\pi)^{(np/2)}}|\sum|^{-n/2} \exp\{-\frac{1}{2}Tr\sum^{-1}\}$$

This likelihood will depend on L and Ψ through the relation $Cov(X)=\sum=LL'+\Psi$. Because of the multiplicity of choices for L made possible by the orthogonal transformations this model is not well defined. Hence, to make L well defined, it is desirable to impose the uniqueness condition, which is computationally convenient.

i.e. $L' \varphi^{-1}L=\Delta$, a diagonal matrix.

Hence, the MLE's $\hat{L}$ and $\hat{\varphi}$ can be obtained from numerical maximization.

3.2.6 THE PRINCIPAL COMPONENT METHOD

Principal components are linear combinations of random or statistical variables, which have special properties in terms of variances. Thus, the principal component analysis is concerned with explaining the variance-covariance structure of a set variable through a few linear

combinations of these variables. The set of principal components yields a convenient set of coordinate, and the accompanying variances of the components characterize their statistical properties. Hence, this is a way of reducing the number of variables and discarding the linear combinations, which have small variances and study those with large variances.

One factoring of the covariance matrix $\sum$ can be obtained by spectral decomposition. If $\sum$ have Eigen value-Eigen vector pair $(\lambda_i, e_i)$with$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, then

$$\sum = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p'$$

$$=[\sqrt{\lambda_1} e_1 \sqrt{\lambda_2} e_2 \ldots\ldots\ldots \sqrt{\lambda_p} e_p] \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \vdots \\ \sqrt{\lambda_p} e_p' \end{bmatrix}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$are the Eigen values and $e_1, e_2, \ldots . e_p$ are the associated normal Eigen vectors.

When the number of factors and number of variables are same (ie, when m = p), the above equation fits the prescribed covariance structure for the Factor analysis model. In this case the specific variances $\psi_i = 0$.The loading matrix has j^{th} column given by $\sqrt{\lambda_j} e_j$

Hence,

$$\sum_{p\times p} = LL'_{P\times P} + 0_{P\times P} = LL'.$$

Apart from the scale factor $\sqrt{\lambda_j}$, the factor loadings on the j^{th} factor are the coefficients for the j^{th}principal component of population.

Although ,the Factor analysis representation given by $\sum = LL'$is exact , it is not particularly useful since it employs as many common factors as there are variables and does not allow for any variation in specific factors ε in $X = \mu + LF + \varepsilon$.

We prefer models that explain the covariance structure in terms of just a few common factors. One approach when the last p-m Eigen values are small is to neglect the contribution of

$\lambda_{m+1} e_{m+1} e'_{m+1} + \lambda_{m+2} e_{m+2} e'_{m+2} + \cdots + \lambda_p e_p e'_p$ to

$\sum = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p'$.

Hence

$$\sum - LL' = \lambda_{m+1} e_{m+1} e'_{m+1} + \lambda_{m+2} e_{m+2} e'_{m+2} + \cdots + \lambda_p e_p e'_p.$$

If specific factors are included in the model, their variances may be taken to be the diagonal elements of $\sum - LL'$, where LL' is defined as follows:

$$\sum = [\sqrt{\lambda_1}e_1 \sqrt{\lambda_2}e_2 \ldots\ldots\ldots \sqrt{\lambda_m}e_m] \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \sqrt{\lambda_2}e_2' \\ \vdots \\ \sqrt{\lambda_m}e_m' \end{bmatrix} = L_{p\times m}L'_{m\times p}.$$

Allowing for specific factors, find that, approximation becomes $\sum = LL' + \psi$

$$= [\sqrt{\lambda_1}e_1 \sqrt{\lambda_2}e_2 \ldots\ldots\ldots \sqrt{\lambda_m}e_m] \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \sqrt{\lambda_2}e_2' \\ \vdots \\ \sqrt{\lambda_m}e_m' \end{bmatrix} + \begin{pmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{pmatrix} \text{---- (A)}$$

Where $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2$, for i=1,2,…p

To apply this approach to a data set $X_1, X_2, \ldots\ldots X_n$ it is customary first to center the observations by subtracting the sample mean $\underline{X}$. Thecentered observations

$$X_j - \overline{X} = [x_{j1}\, x_{j2} \ldots\ x_{jp}] - [\overline{x_1}, \overline{x_2} \ldots\ \overline{x_p}] = [x_{j1} - \overline{x_1}\, x_{j2} - \overline{x_2} \ldots\ x_{jp} - \overline{x_p}], j = 1,2, \ldots\ldots n$$

have the same sample covariance matrix **S** as the original observations .

In case when the units of the variables are not commensurate, it is usually desirable to work with the standardized variables

$$Z_j = \left[\frac{x_{j1}-\overline{x_1}}{\sqrt{s_{11}}}\ \frac{x_{j2}-\overline{x_2}}{\sqrt{s_{22}}} \ldots\ \vdots\ \frac{x_{jp}-\overline{x_p}}{\sqrt{s_{pp}}}\right], j=1, 2, \ldots\ldots\ldots, n$$

whose sample covariance matrix is the sample correlation matrix **R** of the observations $X_1, X_2, \ldots\ldots X_n$.

The matrix representation given in equation (A) when applied to sample covariance matrix **S** or sample correlation matrix **R** is known as *principal component solution.* The name follows that the factor loadings are the scaled coefficients of the first few sample principal components.

3.2.7 FACTOR ROTATIONS

The results of factor extraction, unaccompanied by rotation are likely to be hard to interpret regardless, of which method of extraction is used. After extraction, rotation is used to improve the interpretability and scientific utility o solution. It is not used to improve the quality of mathematical fit between observed and reproduced correlation matrices because all orthogonally rotated solutions are equivalent to one another and to the solution before rotation.

All factor loadings obtained from the initial loadings by an orthogonal transformation have the same ability to reproduce the covariance (or correlation) matrix. We know that, an orthogonal transformation correspond to a rigid rotation of the co-ordinate axes. For this reason, an orthogonal transformation of the factor loadings, as well as the implied orthogonal transformations of the factors, is called *factor rotation*. Rotations are ordinarily used after extraction to maximize high correlations and minimize low ones. There are two general classes of rotation – orthogonal and oblique. In orthogonal rotation, the factors are uncorrelated. Orthogonal solution offers ease of interpreting, describing and reporting results. In oblique rotation, the factors may be correlated, with conceptual advantage but practical disadvantages in interpreting, describing and reporting results.

Numerous methods of rotations are available, namely Varimax, Quartimax and Equimax in which all are orthogonal techniques.

If $\hat{L}$ is the $p \times m$ matrix of estimated factor loadings obtained by any method (principal component or maximum likelihood) then

$\hat{L^*} = \hat{L}T$, where $TT' = T'T = I$ (orthogonal).

Hence $\hat{L^*} = \hat{L}T$ is a $p \times m$ matrix of "rotated" loadings. Moreover, the estimated covariance (or correlation) matrix remains unchanged since

$\hat{L}\hat{L}' + \hat{\Psi} = \hat{L}TT'\hat{L}' + \hat{\Psi} = \hat{L^*}\hat{L^*}' + \hat{\Psi}$.

Hence the residual matrix

$$S_n - \hat{L}\hat{L}' - \hat{\Psi} = S_n - \hat{L^*}\hat{L^*}' + \hat{\Psi}$$

remains unchanged. Moreover, the specific variances $\hat{\Psi_i}$ and hence the communalities $\hat{h_i^2}$ are unaltered.

At times, the original loadings may not be readily interpretable. In such situations, the usual practice is to rotate them until a "simpler structure" is achieved. Graphical method and analytical method are usually adopted to see the pattern of loading such that each variable loads highly on a single factor and has small to moderate loadings on the remaining factors.

3.2.8 FACTOR SCORES

Usually in Factor analysis, the interest in centered on the parameters in the factor model. We may also require the estimated values of the common factors called factor scores. These quantities are often used for diagnostic purposes, as well as inputs to a subsequent analysis.

Factor scores are not estimates of unknown parameters in the usual sense. Rather, they are estimates of the values for the unobserved random vectors $F_j, j = 1,2, \ldots\ldots, n$. That is, factor scores

$f_j^{\wedge}$ =Estimates of the values f_j attained by F_j (j^{th}case).

The estimation of factor scores is done using weighted least squares method as follows:

Suppose the mean vector μ, the factor loading L and specific variance $\boldsymbol{\Psi}$ are known for the factor model, $\mathrm{X} - \mu = LF + \varepsilon$.

Further, regard the specific factor $\varepsilon' = (\varepsilon_1, \varepsilon_2, \ldots\ldots, \varepsilon_p)$ as errors. Since

$\mathrm{V}(\varepsilon_i) = \boldsymbol{\Psi}_i, i = 1,2, \ldots p$ need not be equal. Bartlett suggested that weighted least squares can be used to estimate the common factor values.

The sum of squares of the errors weighted, by the reciprocal of their variance is

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\boldsymbol{\Psi}_i} = \varepsilon'^{\boldsymbol{\Psi}^{-1}}\varepsilon = (\mathrm{X} - LF - \mu)'\boldsymbol{\Psi}^{-1}(\mathrm{X} - LF - \mu)$$

If we take $L^{\wedge}, \boldsymbol{\Psi}^{\wedge}$and $\mu^{\wedge} = \underline{\mathrm{X}}$, the estimates of $L, \boldsymbol{\Psi}$ and μ as the true values, then the factor scores for the j^{th} case is obtained by minimizing $(\mathrm{X_j} - LF_j - \hat{\mu})'\,\boldsymbol{\Psi}^{-1}(\mathrm{X_j} - LF_j - \hat{\mu})$.

The solution is given by

$$F_j^{\wedge} = (L^{\wedge\prime}\boldsymbol{\Psi}^{-1}L^{\wedge})^{-1}L^{\wedge\prime}\boldsymbol{\Psi}^{\wedge-1}(\mathrm{X_j} - \overline{\mathrm{X}}), \quad j=1,2.....n$$

The factor scores generated have sample mean vector $\mathbf{0}$ and zero sample covariance matrix. There are several sophisticated statistical approaches for estimating factors. All produce factor scores that are correlated but not perfectly, with factors. The correlations between factors and factor scores are higher when communalities are higher and when the ratio of variables to factors is

higher. But as long as communalities are estimated, factor score suffer from indeterminacy because there is an infinite number of possible factor scores that all have the same mathematical characteristics and there is no way to decide among them.

3.2.9 ADVANTAGES AND DISADVANTAGES OF FACTOR ANALYSIS

ADVANTAGES:

- Both objective and subjective attributes can be used provided the subjective attributes can be converted to scores.
- Factor analysis can be used to identify hidden dimensions or constructs which may not be apparent from direct analysis.
- It is easy and inexpensive to do.

DISADVANTAGES:

- Usefulness depends on the researcher's ability to collect a sufficient set of product attributes. If important attributes are missing, the value of the procedure is reduced.
- If sets of observed variables are highly similar to each other but distinct from other items. Factor analysis will assign a single factor to them. This may make it harder to identify factors that capture more relationships that are interesting
- Naming the factors may require background knowledge of theory because multiple attributes can be highly correlated for no apparent reason.

3.2.10 SCREE PLOT

It is a graphical procedure in which we decide upon the number of factors to be retained. It is a two-dimensional graph with factor along and loadings along . Loadings are arranged in a descending order. The first two factors account for most of variations. The factors towards the end of the figure/plot may be discarded and evident bend / elbow in the scree plot decides the number of factors.

3.3 CLUSTER ANALYSIS

"Cluster analysis" is a statistical method used to group similar objects or observations together based on their attributes or characteristics. The goal of cluster analysis is to identify natural groupings, or clusters, within a dataset, in which members of each cluster are more similar to one another than they are to members of other clusters. Broadly speaking, cluster analysis involves categorization: dividing a large group of observations into smaller groups so that the observations within each group are relatively similar and the observations in different groups are relatively dissimilar. Cluster analysis is a more primitive technique in that no assumptions are made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities).

Cluster analysis differs fundamentally from classification analysis. In classification analysis, we allocate the observations to a known number of predefined groups or populations. In cluster analysis, neither the number of groups nor the groups themselves are known in advance.

There are several methods of cluster analysis, including hierarchical clustering and k-means clustering. In hierarchical clustering, the algorithm starts with each observation as a separate cluster and then recursively merges clusters based on similarity, until all observations belong to a single cluster. In k-means clustering, the algorithm assigns each observation to one of a predetermined number of clusters, based on the similarity of their attributes, and then iteratively refines the assignments to minimize the distance between the observations and theirclustercentroids

Cluster analysis has wide applications in biology, medicine, agriculture, marketing, social network, image segmentation, bioinformatics etc. In taxonomy the various families of the plant species are clustered according to their similarities in psychiatry for the correct diagnosis. The cluster symptoms is essential. In archaeology researchers have attempted to establish taxonomies of stone tools, funeral objects, weapons etc by applying cluster analysis. The prices of land showing similar fertility for a particular crop are clustered in group farming. In marketing people having the consumptions-buying habits are clustered. In general whenever we need to classify a large data to a manageable and meaningful pieces.

It is of great utility. It is a powerful tool for identifying patterns and relationships in complex datasets and can help to guide decision-making and strategic planning in many fields

LIMITATIONS OF CLUSTER ANALYSIS

There are several things to be aware of when conducting cluster analysis.

❖ The different methods of clustering usually give very different results. This occurs because of the different criterion for merging clusters(including cases). It is important to think carefully about which method is best for what you are interested in looking at.

❖ With the exception of simple linkage, the results will be affected by the way in which the variables are ordered.

❖ The analysis is not stable when cases are dropped: this occurs because selection of a cases(or merger of clusters) depends on similarity of one case to the cluster. Dropping one case candrastically affect the course in which the analysis progresses. .

❖ The hierarchical nature of the analysis means that early 'bad judgements' cannot be rectified.

3.3.1 Similarity Measures

Similarity measures are used in cluster analysis to determine how close or far apart two data points are from each other. Usually this is indicated by two types of measures.

(a) Distance type measure

- **Minkowski's Distance :** Let the p observation variables be denoted by the random variables $X_i = (X_{i1} , X_{i2}.....X_{ip})'$; i = 1,2…..n. Thus the distance d_{ij} between i^{th} and j^{th} objects.

$$d_{ij} = \left[\sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|^{r} \right]^{1/r}$$ is called the minkowski's matrix.

- **Squared Eucledian Distance:** It is a commonly used distance metric to measure the similarity or dissimilarity between two observations or data points. It is defined as the sum of the squared differences between the corresponding coordinates of two points. i.e when r = 2 on the minkowski's matrix we get squared eucledian distance.

$$d_{ij} = \left[\sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|^{2} \right]^{1/2}$$

However,one potential drawback of the squared eucledian diatance is that it places more emphasis on the larger differences between the coordinates ,since they are squared. This can problematic if the data has outliers or if the scales of the variables are not standardized.In such cases, alternative distance metrics, such as Manhattan distance or Mahalanobis distance , may be more appropriate.

- **Manhattan Distance:** It is also called taxicab distance or city-block distance .It calculates the distance between two data points by adding up the absolute differences between the corresponding coordinates of the two points.

$$d_{ij} = \sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|$$

It is the distance that one would travel if the i^{th} and j^{th} points are located at opposite corner of a city.

- **Mahalanobis Distance**: It is used to assess the similarity between two data points. Specifically, it measures the distance between a point and the centre of a cluster, taking into account the covariance matrix of the data with the cluster. This can help to ensure that the clusters are formed based on the most relevant and informative variables

$$d_{ij} = \left(X_{ik} - X_{jk} \right)' S^{-1} \left(X_{ik} - X_{jk} \right)$$

- **Cosine similarity:** It is used to measure the similarity between two vectors of an inner product space. It measures the cosine of the angle between the two vectors, and ranges from -1 to 1. A value of 1 indicates that the two vectors are identical, while a value of -1 indicates that they are completely dissimilar.

- **Jaccard similarity:** It is used to measure the similarity between two sets. It calculates the ratio of the size of the intersection of the two sets to the size of the union of the two sets.

- **Pearson correlation coefficient**: It is used to measure the linear relationship between two variables. It measures the correlation between two variables by dividing the covariance of the two variables by the product of their standard deviations.

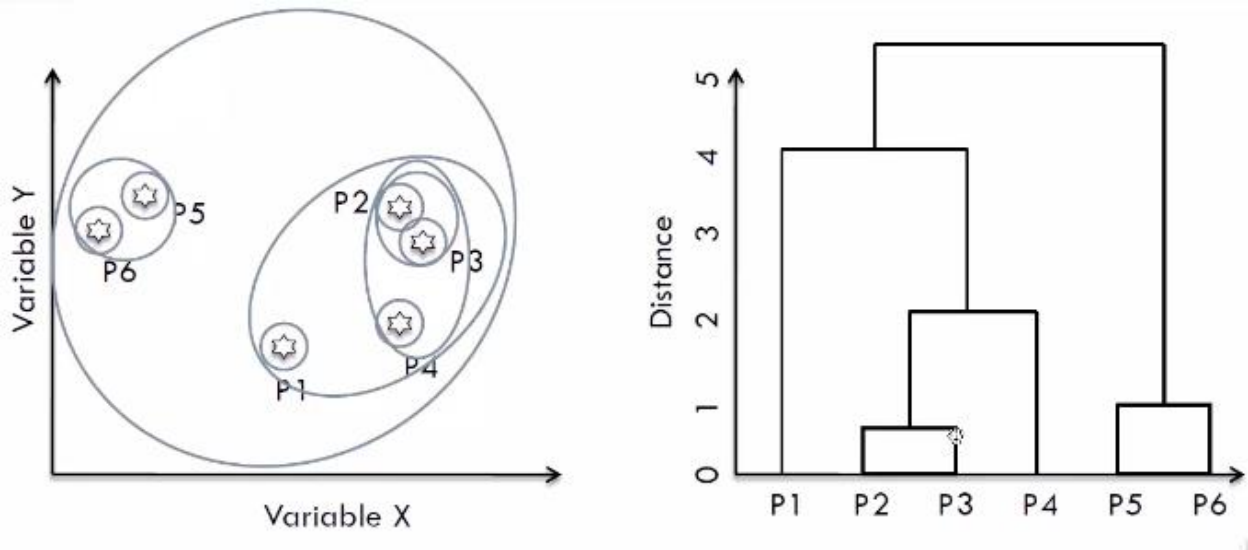
The choice of similarity measure depends on the type of data being analyzed and the specific research question being asked.

(b) Matching type measures

Such a measure is used when the data is qualitative. There is possibility that two or more sample objects are similar in the values of the variables. Since the objects are associated in respective of the values of the variables, the matching type measures are also called association co-efficients. The variables either categorical or numerical. These are converted into numerical information by frequencies and total frequency turns out to

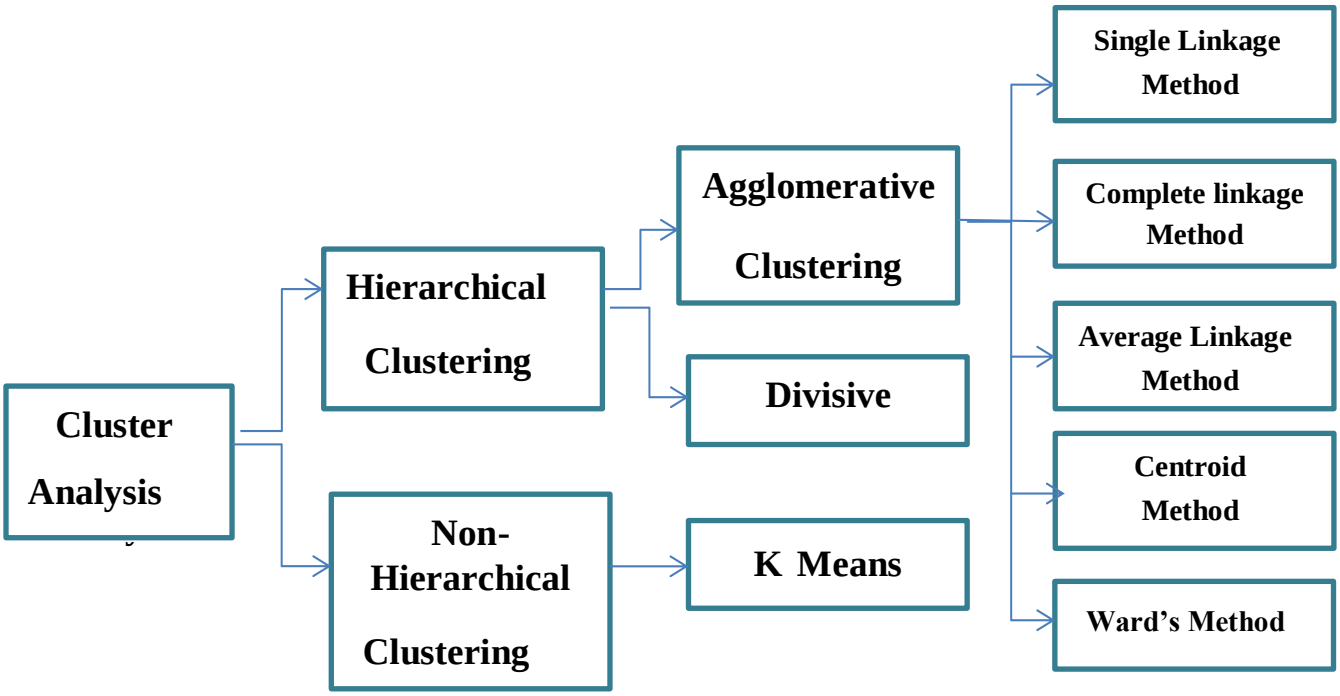
be the similarity or dissimilarity. Usually contigency tables are used for the cluster analysis using matching type measures.

3.3.2 Dendogram



When carrying out a hierarchical cluster analysis, the process can be represented on a diagram known as dendogram. This diagram illustrates which clusters have been joined at each stage of the analysis and distance between clusters at the time of joining. If there is a large jump in the distance between clusters from one stage to another then this suggests that at one stage clusters are relatively close together were joined whereas , at the following stage, the clusters that were joined were relatively far apart.

3.3.3 CLUSTERING PROCEDURES



There are a number of different methods that can be used to carry out a cluster analysis; these can be classified as Hierarchical and Non- Hierarchical methods.

1. Hierarchical methods
2. Non-Hierarchical methods

Hierarchical Clustering Methods

Hierarchical clustering techniques proceed by either a series of successive merges or a series of successive divisions .It is of two types ; Agglomerative hierarchical method and Divisive hierarchical method.

i. Agglomerative Hierarchical Methods

It starts with the individual objects.Thus, there are initially as many clusters as objects.The most similar objects are first grouped and these initial groups are merged according to their similarities.The following are the steps in the agglomerative hierarchical clustering algorithm for grouping N objects :

1. Start with N clusters, each containing a single entity and NxN symmetric matrix of distances(or similarities)

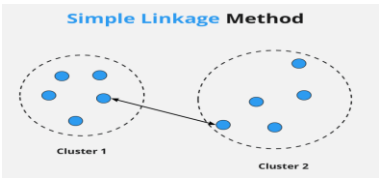
$$D = [\ d_{ij}\ ]$$

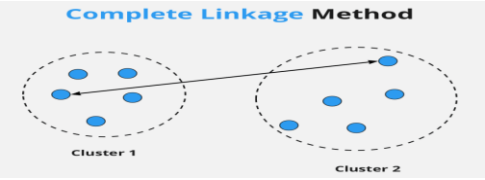
2. Search the distance matrix for the nearest pair of clusters. Let the distance between "most similar" clusters u and v be d_{uv}.
3. Merge clusters u and v . Label the newly formed clusters (uv) update the entries in the distance matrix by (.) deleting the rows and columns corresponding to u and v and (*)adding a row and columns giving the distances between cluster (uv) and the the remaining clusters
4. Repeat step 2 and 3 a total of N-1 times. All objects will be in a single cluster after the algorithm terminates. Record the identity of clusters thet are merged and the levels at which the merges takes place.

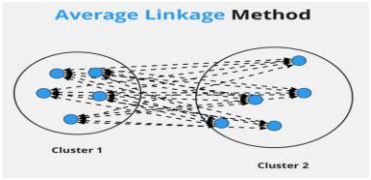
In step (*) there are different possibilities of defining distances :

(a) Single Linkage (minimum distance): Groups are formed from the individual entities by merging nearest neighbours, where the term nearest neighbor means the smallest distance or largest similarity. Initially, we must find the smallest distance in $D = [\ d_{ij}]$ and merge the corresponding objects, say u and v , to get the cluster (uv). For step (*) of the general algorithm, the distance between (uv) and other cluster w are computed by

$$d_{(uv)w} = \min \{\ d_{uw}\ ,\ d_{vw}\ \}$$







(b) Complete linkage (maximum distance): It proceeds in much the same manner as single linkage clustering, with one important exception; At each stage, the distance between clusters id determined by the distance between two elements, onr from each clusters, that are more distant. Initially, we must find the smallest distance in

D = [d_{ij}] and merge the corresponding objects say u and v to get the cluster (uv).

For step (*) of general algorithm, the distances between (uv) and other cluster w are computed by

$$d_{(uv)w} = \max \{ d_{uw} , d_{vw} \}$$

(c) Average linkage method: It treats the distance between two clusters as the average distance between all pairs of items where one number of a pair belongs to each other.

For (*) of the general algorithm, the distance between (uv) and any other clusters w are computed by

$$d_{(uv)w} = \frac{\sum_i \sum_j d_{ij}}{N_{(uv)} N_w}$$

Where d_{ij} is the distance between object i in the cluster (uv) and object j in the cluster w, respectively.

(d) Ward's method: It is based on minimising "the loss of information" from joining two groups**.** In this method all possible pairs of cluster are combined and the sum of squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen.

ii. Divisive Hierarchical Methods

It works in a opposite direction of agglomerative hierarchical methods. An initial single group of objects is divided into two subgroups such that the objects in one subgroup are "far from" the objects in the other. These subgroups are th en further divided into dissimilar subgroups. The process continuous until there are many subgroups as objects. i.e. until each object forms a group.

Divisive algorithms are generally of two classes: Monothetic and Polythetic .In a mononthetic approach, the division of a group into two subgroups is based on a single variable, whrereas, the polythetic approach uses all p variables to make the split. For a polythetic approach, the following are the steps in the divisive hierarchical clustering algorithm.

1. Find the objects ha ving highest average dissimilarity. This is obtained from the proximity matrix. These objects form a new group called the splinter group.

2. All other objects having another cluster called the main group.

3. Find the average distance of other objects in the main group from the objects in the splinter group.

4. Also, find the average distance of objects in the main group.

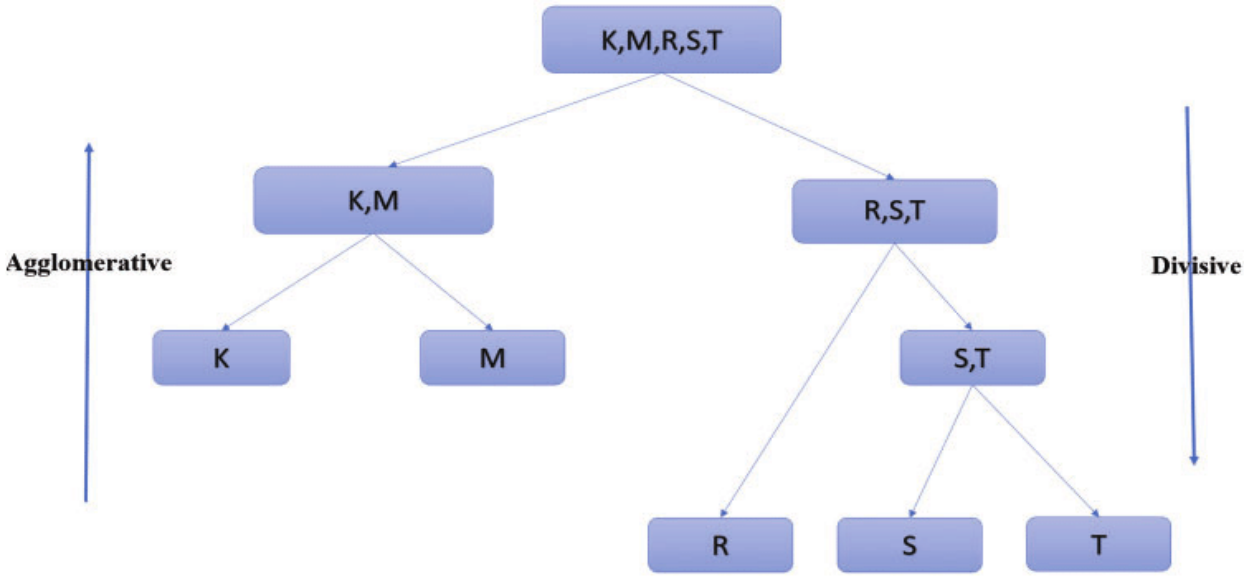
5. Compute D_i = { average d_{ij} , j ϵ main group – average d_{ij} , j ϵ splinter group} for all i ϵ main group

6. Merge the objects that have smallest positive Di to the splinter group.

7. Repeat the steps until all the D_i's are negative. The data is then split into two clusters.

8. Repeat these steps steps until both the clusters are further divided into some clusters with single observation.

Diagramatic representation of agglomertive and divisive method



iii) Non-Hierarchical methods

Non-Hierarchical clustering techniques are designed to group items, rather than variables into a collection of k clusters. The number of clusters, k may either be specified in advance or determined and the basic data do not have to be stored during the computer run, non-hierarchical methods can be applied to much larger sets than hierarchical techniques. The k-means clustering method is a commonly used non-hierarchical clustering technique.

3.3.4 K-means method

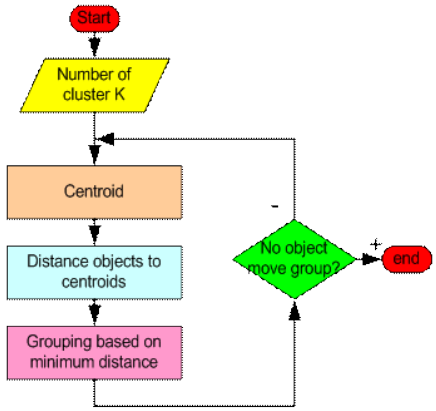
The k-means method assigns each item to the cluster having the nearest centroid. In its simplest version, the process is compared of the following steps:

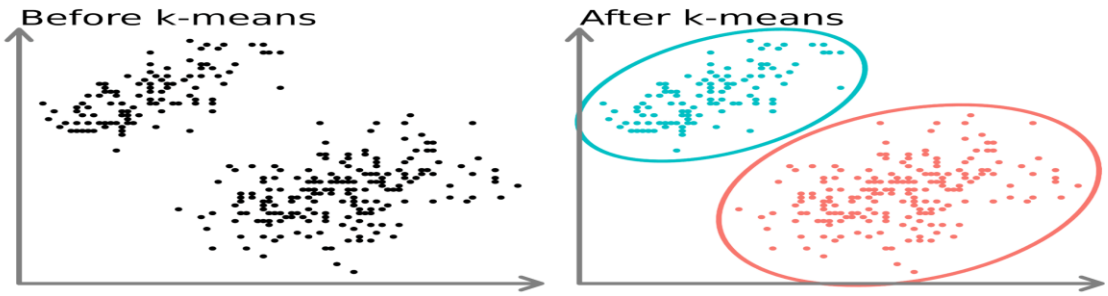
1. Partition the cluster into k initial or specify k initial centroids.

2. Calculate the cluster centroid of each cluster. Proceed the list of items, assigning an item to the cluster whose centroid is nearest. Distance is usually computed using Eucledian distance.

3. Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

4. Repeat the step 2 - 3 until no reassignments take place.





Assumptions made by K Means

1. All clusters are the same size.
2. Clusters have the same extent in every direction.
3. Clusters have similar numbers of points assigned to them

Final comments as K-means Methods

Fixing K in advance we faces the following problem

❖ If two or more seed points lie initially within a single cluster, their resultant clusters cannot be differentiated.

❖ The presence of outlier may produce at least one group with a very dispersed/scattered observations.

❖ Even if the population consists of, say K groups, the data need not consists of units from these K groups(some groups may not have representation in the sample) such grouping may lead to clusters without any sense.

3.3.5 Hierarchical Methods Vs Non-Hierarchical Methods

Hierarchical Methods	Non-Hierarchical Methods
❖ No decision about the number of clusters. ❖ Problems when data contain a high level of error ❖ Can be very slow, preferable with small data-sets. ❖ Initial decisions are more influential (one step only) ❖ At each step they require computation of the full proximity matrix	❖ Faster, more reliable, works with large data Sets ❖ Need to specify the number of clusters ❖ Need to set the initial seeds ❖ Only cluster distances to seeds need to becomputed in each iteration.

CHAPTER 4

ANALYSIS OF DATA

4.1 PRINCIPAL COMPONENT ANALYSIS

The result obtained by using the principal components analysis as the extraction method is given below.

COMMUNALITIES

	Initial	Extraction
AREA	1.000	0.995
PERIMETER	1.000	0.990
COMPACTNESS	1.000	0.647
LENGTH OF KERNAL	1.000	0.945
WIDTH OF KERNAL	1.000	0.953
ASYMMETRY OF COEFFICIENT	1.000	0.625
LENGTH OF KERNAL GROVE	1.000	0.937

Communalities

This is the proportion of each variable that can be explained by the Principle Components.

Initial

Initial values of the communality in a principle component analysis are one.

Extraction

It indicates that proportion of variance that can be explained by the principal components.

Now, a scree plot displays the eigen values associated with a component or factor in descending order versus the number of the components or factor. We use scree plots in principal components analysis and the factor analysis to visually assess which components or factors explain most of the variability in the data

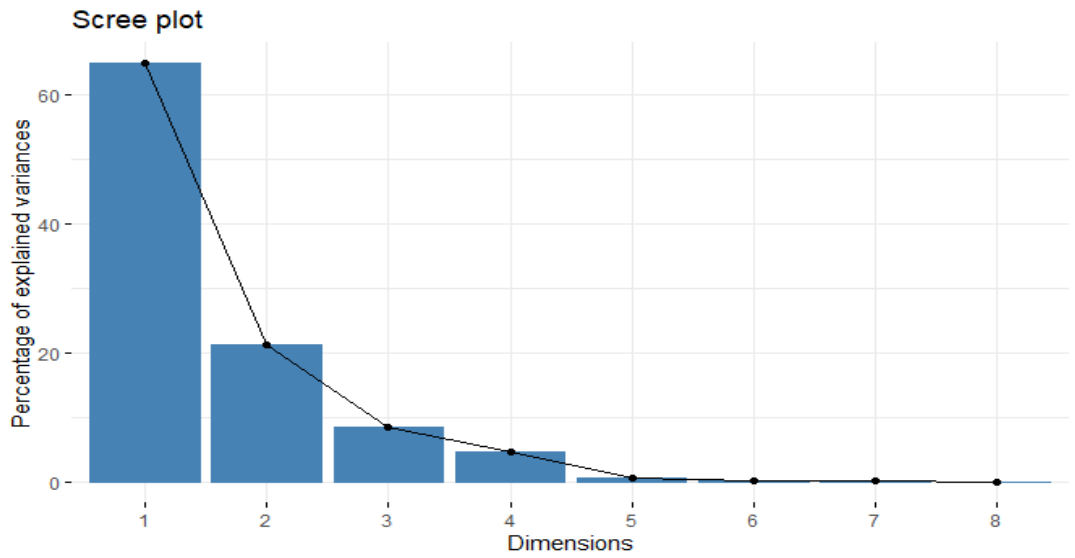
Variation Explained

Variance Explained		
	PC 1	PC 2
Sum of square loading	5.19	1.70
Proportion variance	0.65	0.21
Cumulative Variance	0.65	0.86

For each of the principal components we have three values: a standard deviation, proportion variance, cumulative variance

The standard deviation is the standard deviation of the data along a single principal components. The proportion of variance is the proportion of all the variability in the original data ,i.e., in our case the 65% of variance in the data is explained by PC1,21% of variance by PC2…and so on. The cumulative proportion is the same idea but cumulative.so if you take PC1 we will be able to explain 65%,if we take PC1 and PC2 we explain 86%

We use scree plot in principal component analysis to visually assess which component explain most of the variability in the data



This plot shows the percentage of variance across its principal component. Of course ,the variance is the square of the standard deviation and so that you can see that when it has been taken into account the first and second principal components big part of the variability in our data has been accounted for.

Another statistical way of interpreting the relationship between each variable in our orginal dataset and the principal components is by looking at the correlations between our orginal variable and PCs

COMPONENT MATRIX		
	PC1	PC2
AREA	0.992	0.105
PERIMETER	0.983	0.153
COMPACTNESS	0.654	-0.469
LENGTH OF KERNAL	0.936	-0.262
WIDTH OF KERNAL	0.975	-0.39
ASYMMETRY OF COEFFICIENT	-0.318	0.724
LENGTH OF KERNAL GROVE	0.831	0.498

The above are Pearson correlation Coefficients, so for example, the correlation between the variable 'Area' and **PC1** is large and positive which means that the value of the observation **PC1** along with 'Area'. Alternatively for the variable 'Asymmetry coefficent' we have moderate negative correlation between them, so that , as 'Asymmetry coefficient' decrease **PC1** increases Similarly,the correlation between **PC2** and 'Asymmetry coefficient' and is negative with respect 'compactness'

Also we get, the first component is highly influenced by the variables X_1 followed by X2, X5 and X4,second component is highly influenced by the variables X6 followed by X7,X4 and X2

From the component matrix

Components	
1	2
$\mathbf{X_1}$ (Area)	$\mathbf{X_6}$ (Asymmetry of coefficent)
$\mathbf{X_2}$ (Perimeter)	$\mathbf{X_7}$ (Length of groove)
$\mathbf{X_5}$ (Width of the kernel)	$\mathbf{X_4}$ (Length of the kernel)
$\mathbf{X_4}$ (Length of the kernel)	$\mathbf{X_2}$ (Perimeter)

Let the principal components be Y_1 ,Y_2

From the score coefficient matrix we get,

$$\mathbf{Y_1 = 0.992X_1+0.983X_2+0.654X_3+0.936X_4+0.975X_5+-0.318X_6+0.831X_7+-0.437X_8}$$

$$\mathbf{Y_2 = 0.105X_1+0.153X_2+-0.469X_3+0.262X_4+-0.39X_5+0.724X_6+0.498X_7+0.776X_8}$$

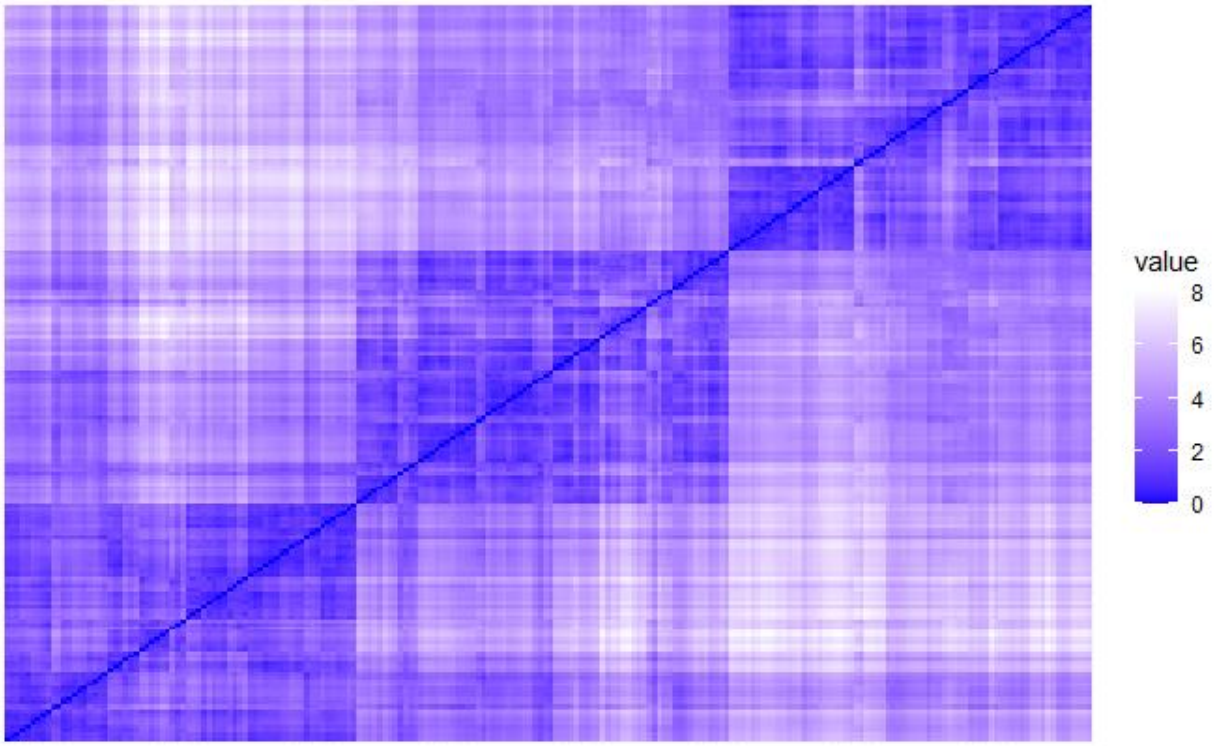
It is seen that the first component explains 65% of variation of the data set. Second component explains 21% of data set That is, first two components explain 86.063% of variation of data set. It is also from the component matrix that the first two components are highly influenced by all the eight factors.

4.2 Cluster Analysis

Hopkins Statistic

The Hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) is a way of measuring the cluster tendency of a data set. It belongs to the family of sparse sampling tests. It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by

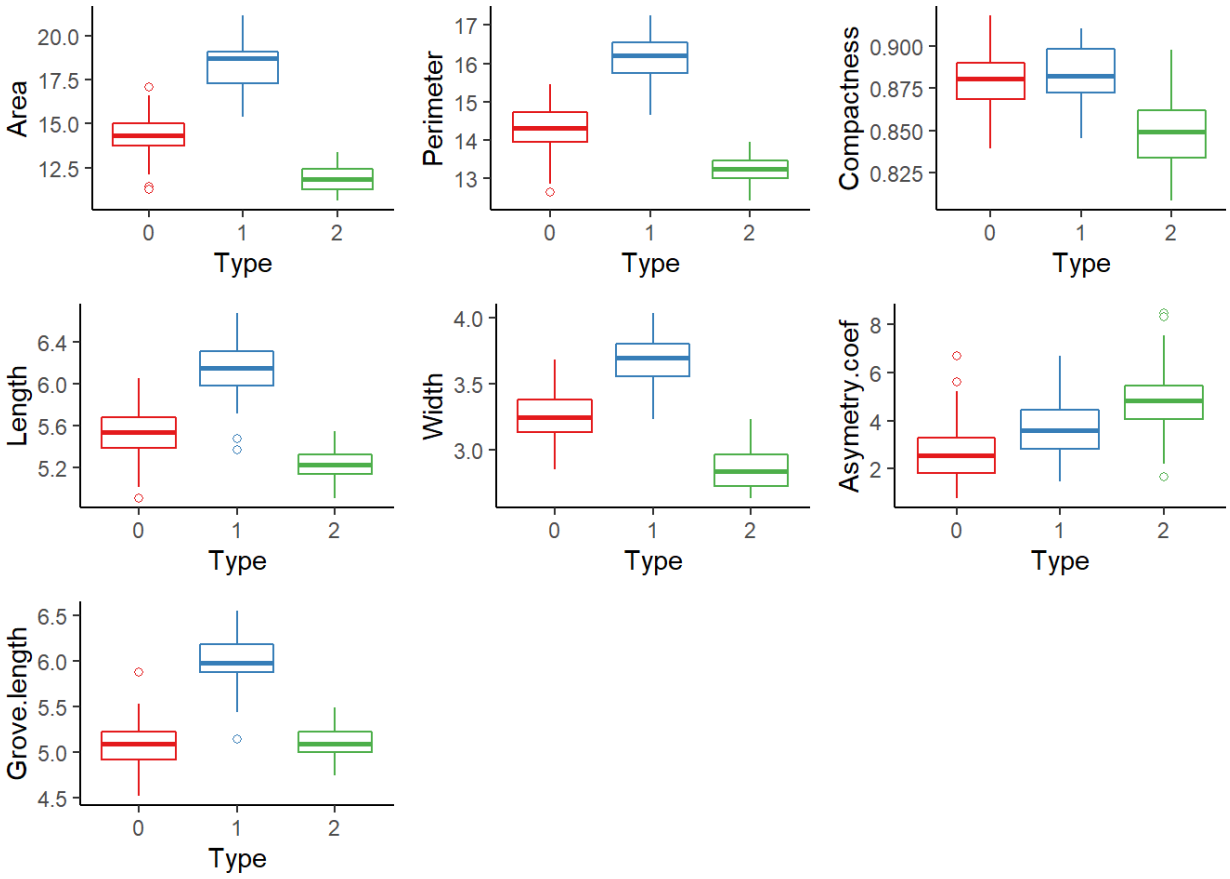
a Poisson point process and are thus uniformly randomly distributed. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.



In the case of our dataset the value of the Hopkins statistic is 0.75 which claims that the dataset is significantly clusterable. Also the darker rectangular blocks shown in the graph are another proof of the clusterability of the dataset.

Clustering Opportunity

Clustering opportunity is important in clustering analysis because it enables pattern discovery, data exploration, anomaly detection, targeted marketing, data compression, preprocessing, and knowledge discovery. It helps in extracting meaningful information, gaining insights, and making informed decisions based on the underlying structure and relationships present in the data.



From the barplot, we can see that in general, Kernel type 1 has the highest value on most of the variables. Even so, there was a slight difference when we use Asymmetry of coefficient whereas the value of Kernel 0 < 1 < 2. Compactness also reveals a different result whereas the value of Kernel 0 almost equal to Kernel 1 and then followed by Kernel 2.

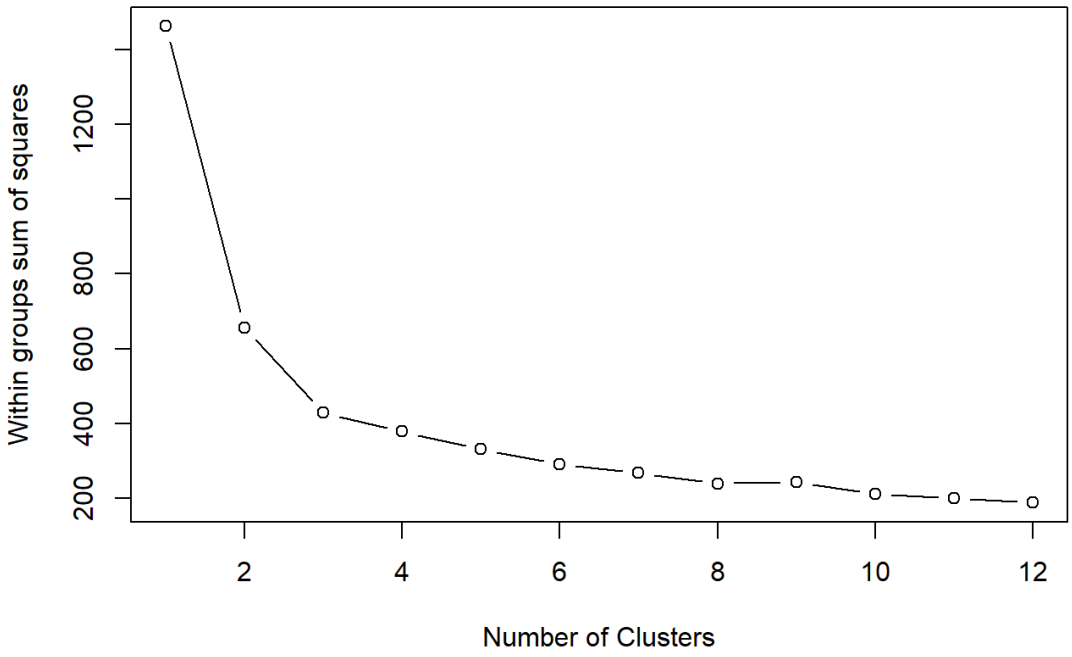
By having a distinct separation of values between each type of Kernel, this data might be suitable for **clustering using K-means**.

4.2.1 K-Means

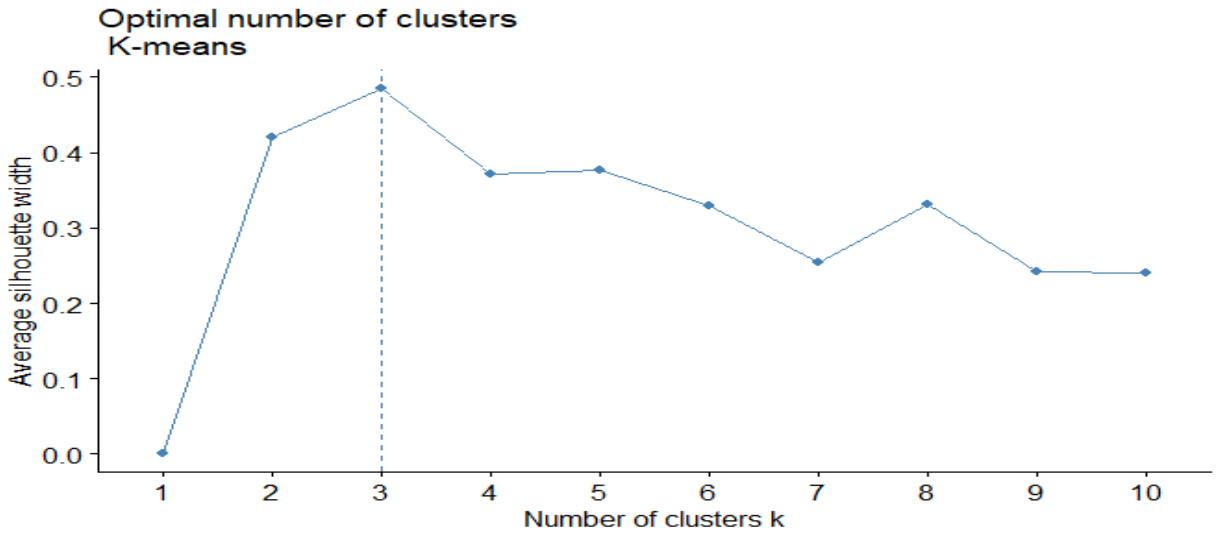
The important step in K-means clustering technique is to decide the number of cluster. So by Elbow method.

Elbow Method

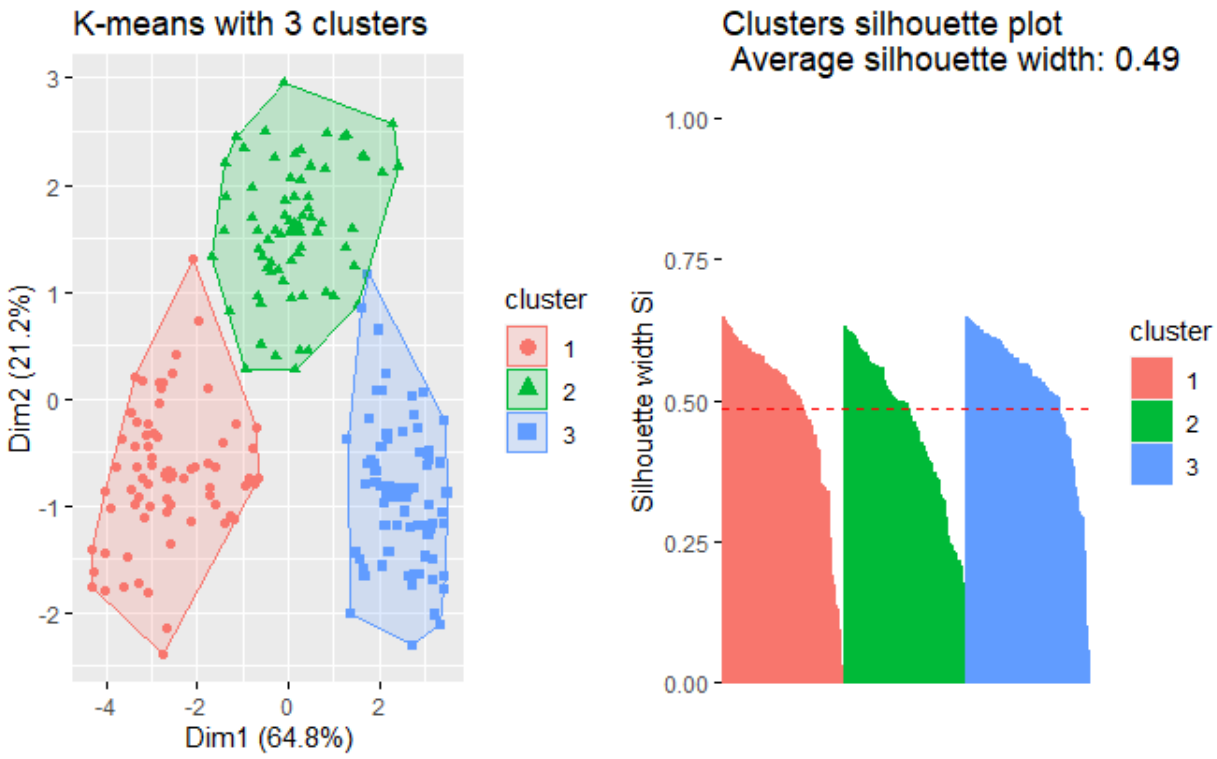
The elbow method is a popular technique used in K-means clustering to determine the optimal number of clusters to use for a given dataset. It helps to find the "elbow" or the point of inflection on a plot of the within-cluster sum of squares (WCSS) against the number of clusters.



We see a bend (elbow) in the graph at k=3, therefore, 3 is the optimal number of clusters.
Another method for finding the optimal number of the clusters for means is **Silhouette statistic**
From silhouette statistic it is clear that the optimal number of cluster is k=3.so we are proceeding our k means with taking k=3

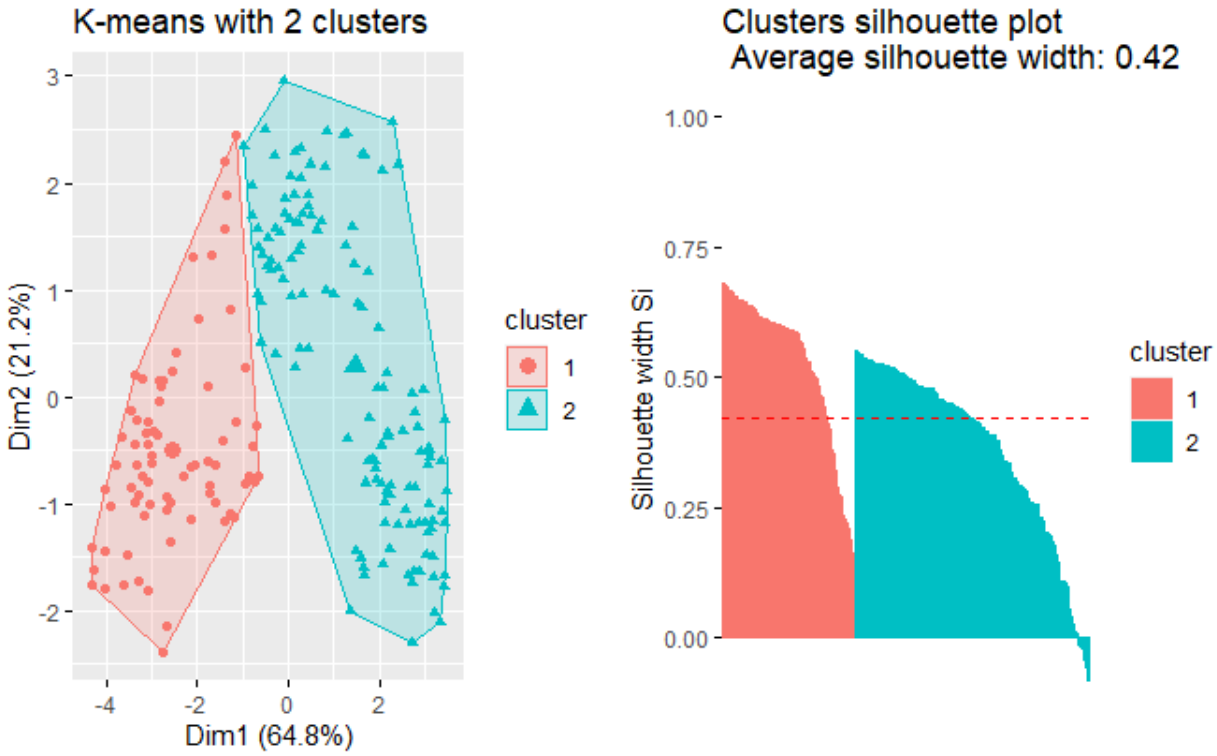


When K=3



When k=3 the result looks better since we got a higher average silhouette width with no negative values, and any of the clusters are not overlapping

When K=2



When k=2 the result looks not better since we got a higher average silhouette width with one negative values, and any of the clusters are not overlapping

So K-means with 3 clusters is the best option according with the silhouette statistics

Clusters	1	2	3
Size	70	70	70
Avg. silhouette. Width	0.41	0.42	0.49

Average value of K-means Cluster

Area	Perimeter	Compactness	Length of kernel	Width of kernel	Asymmetry coefficient	Length of Grooves	Class
1.20981	**1.2212**	**0.5160033**	**1.2062015**	**1.1126**	**-0.05527**	**1.2689**	*-0.0349*
-0.1878	*-0.21702*	*0.397688*	*-0.3056103*	*-0.040821*	*-0.6684*	*-0.67545*	*-1.1869*
-1.0219	*-1.00418*	*-0.91369*	*-0.90059*	*-1.071797*	*0.72327*	*-0.5934*	*1.22182*

Looking at the variables in the above table we can get the idea about which group of seeds are better.so according to average value table , the cluster one is the one with characteristics of seed have higher values. With respect to the first cluster both cluster 2 and 3 possess similar inferior characteristics.

HIERARCHIAL CLUSTERING

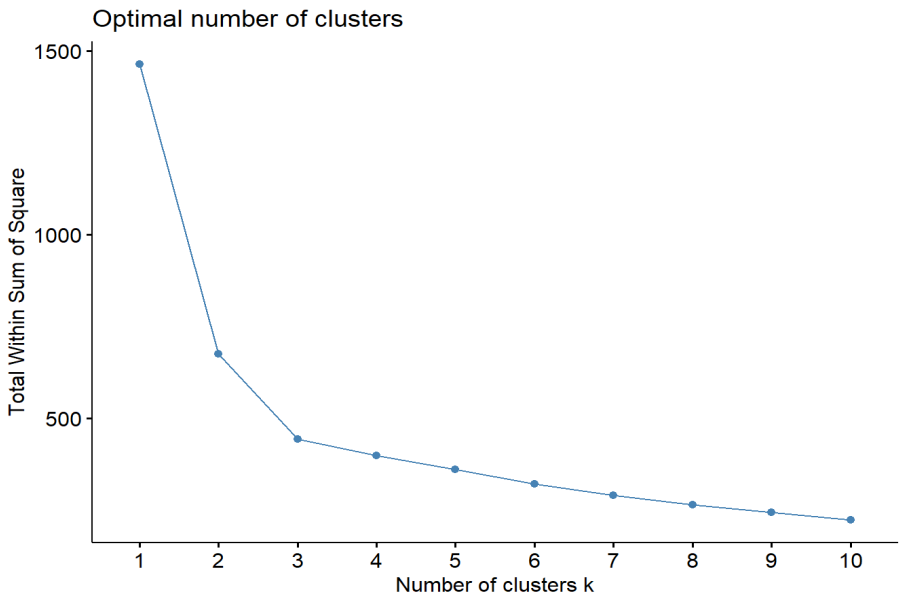
It is good to use extra methods for clustering our data to check if our results are consistent. That is why a Hierarchical clustering analysis has been done.

To decide which method to use for the Hierarchical clustering let's calculate the agglomerative coefficient for each method. This coefficient measures the amount of clustering structure found, values closer to 1 suggest strong clustering structure.

Average	Single	Complete	Ward
0.8654	0.607	*0.92332*	**0.98929**

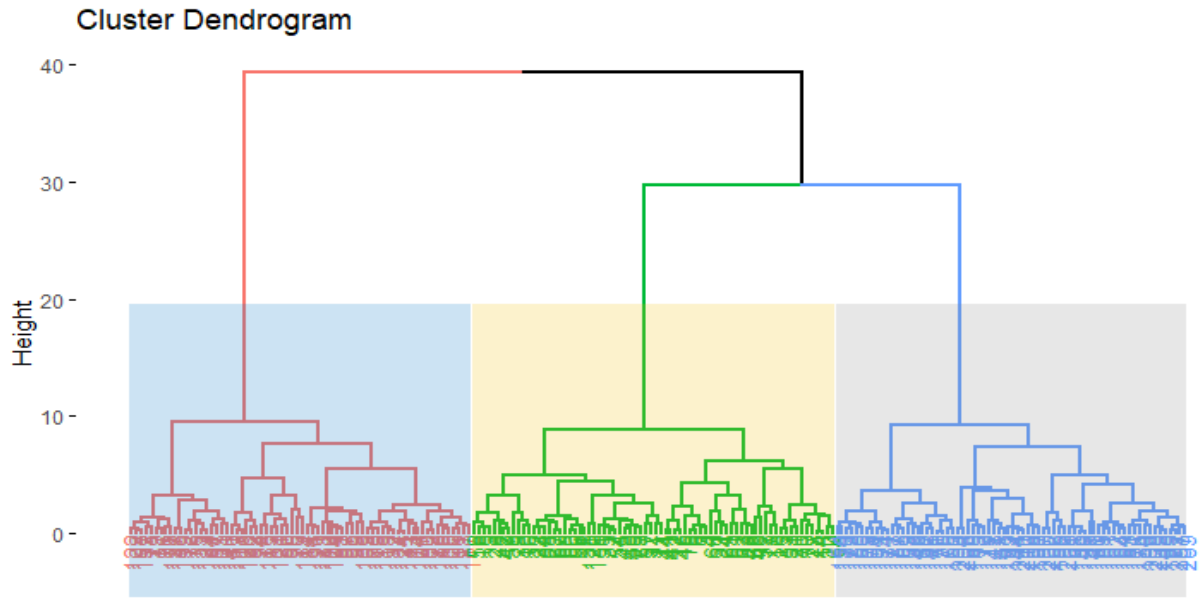
Choosing Ward's and Complete Linkage's methods since they are the ones with closest values to 1.

Now to choose the proper number of clusters,using Elbow method we find the optimal number of clusters.

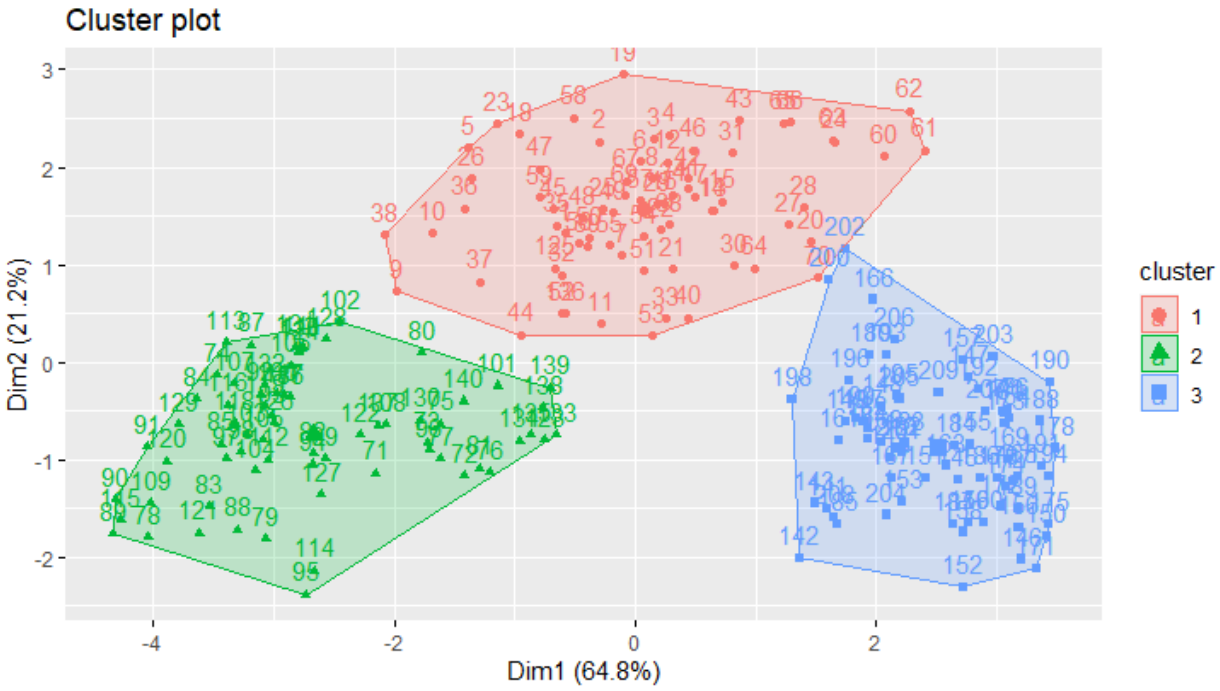


Clustering using wards and complete method as,

4.2.2 Wards Method



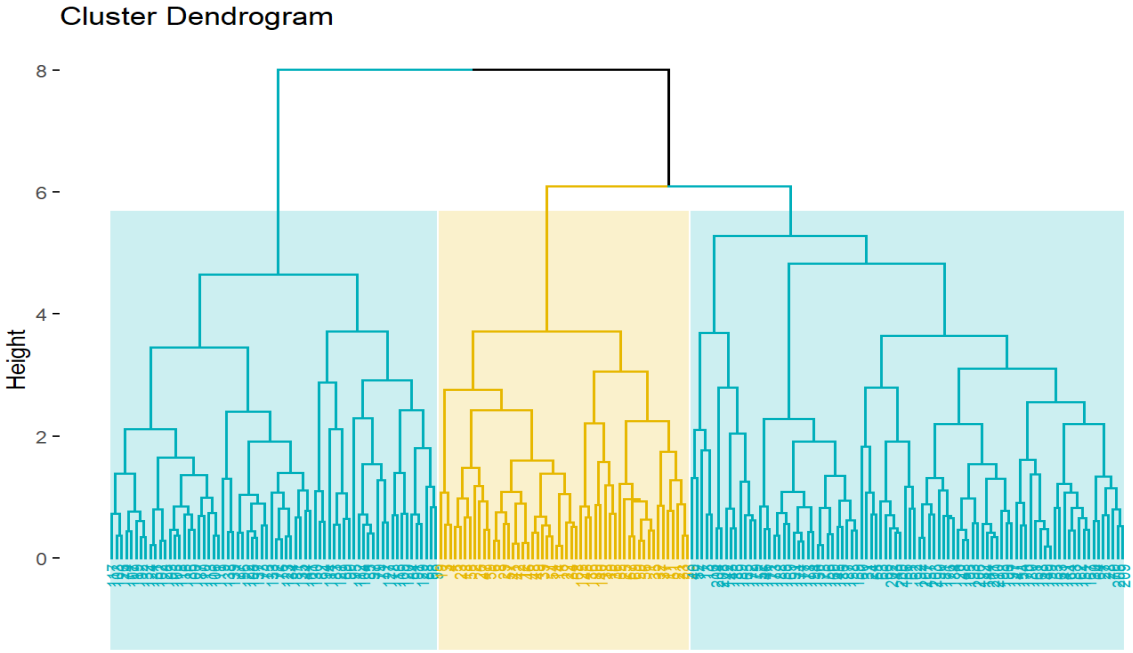
The cluster plot is given in Figure

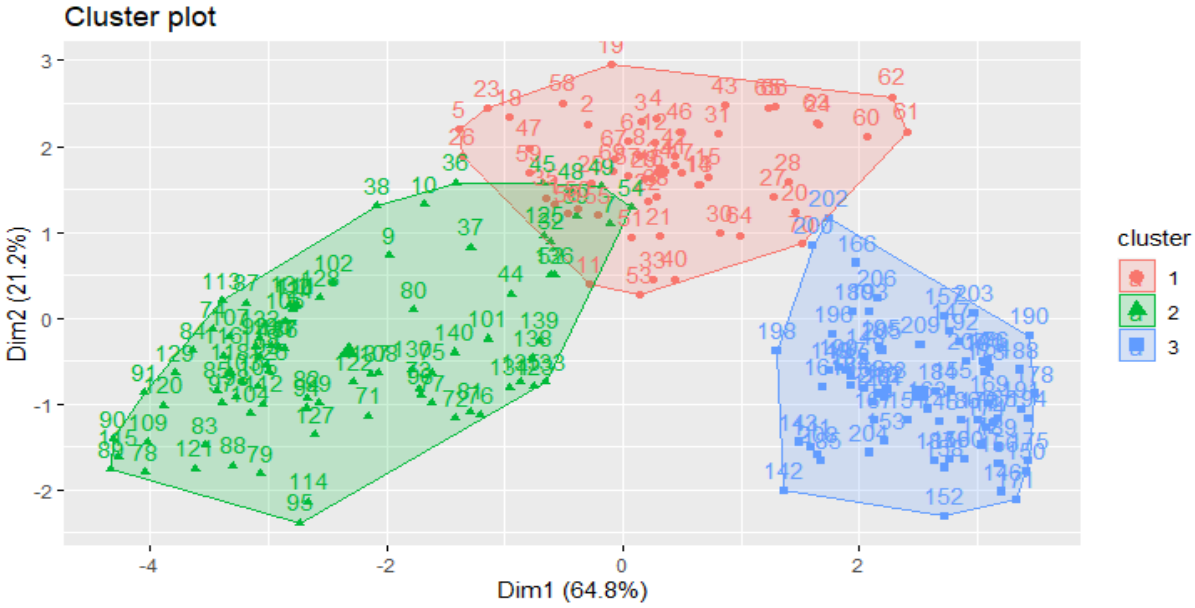


Here cluster 1 , 2 and 3 does not overlap, so it represents a distinct set of data points.

Therefore clustering is good in wards method

4.2.3 Complete Linkage Method





Here cluster 1 and cluster 2 overlaps each other and cluster 2 and cluster 3overlaps too.so it means that there are data points that belong to both clusters.

Final Clusters

Final Clusters				
Member Complete Linkage	Member Wards Linkage			
		1	**2**	**3**
	1	64	4	2
	2	4	66	0
	3	5	0	65

Silhouette Statistics

The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters . The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

On comparing silhouette statistic of both method Wards method has value **0.4836766** and complete linkage method has **0.4310212**. Wards method with a higher value of the Average Silhouette Width so it is the good method

Average value of Ward's Method

Area	Perimeter	Compactness	Length of kernel	Width of kernel	Asymmetry coefficient	Length of Grooves
-0.222	*-0.2228*	*0.34667*	*-0.339230*	*-0.08512*	*-0.723630*	*-0.6549*
1.211	**1.2145**	**0.56712**	**1.195400**	**1.127899**	**-0.040599**	*1.23972*
-1.022	*-0.9971*	*-0.97027*	*-0.879316*	*-0.87931*	*0.83085*	*-0.58163*

Looking at the variables in the above table we can get the idea about which group of seeds are better.so according to average value table of Ward's method , the cluster 2 is the one with characteristics of seed have higher values. With respect to the second cluster, both cluster 1 and 3 possess similar inferior characteristics.

4.3 Factor Analysis

Kaiser Meyer Olkin(KMO) and Bartlet's Test

The Kaiser–Meyer–Olkin (KMO) **test** is a statistical measure to determine how suited data is for factor analysis. The test measures sampling adequacy for each variable in the model and the complete model. The statistic is a measure of the proportion of variance among variables that might be common variance. The higher the proportion, the higher the KMO-value, the more suited the data is to factor analysis.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**0.788**
Bartlett's Test of Sphericity	Approx. Chi-Square	3623.407
	Df	21
	Sig.	.000

In general, KMO values between 0.8 and 1 indicate the sampling is adequate. KMO values less than 0.6 indicate the sampling is not adequate and that remedial action should be taken. In contrast, others set this cutoff value at 0.5 A KMO value close to zero means that there are large partial correlations compared to the sum of correlations. In other words, there are widespread correlations which would be a large problem for factor analysis. In the table KMO measure is 0.788, which is acceptable and therefore factor analysis can be done

Bartlett's Test of Sphericity compares an observed correlation matrix to the identity matrix. Essentially it checks to see if there is a certain redundancy between the variables that we can summarize with a few number of factors**.**An identity matrix is a matrix in which all of the values along the diagonal are 1 and all of the other values are 0. We want to reject the null hypothesis. From the same table we can see that significant value is 0.000 which is lessthan 0.05.Thus the significant level is small enough to reject the null hypothesis . This means that the correlation matrix is not an identity matrix.

COMMUNALITIES

	Initial	Extraction
AREA	1.000	0.995
PERIMETER	1.000	0.990
COMPACTNESS	1.000	0.647
LENGTH OF KERNAL	1.000	0.945
WIDTH OF KERNAL	1.000	0.953
ASYMMETRY OF COEFFICIENT	1.000	0.625
LENGTH OF KERNAL GROVE	1.000	0.937

Communalities indicate the amount of variance in each variable that is accounted for. Initial communalities are estimates of the variance in each variable accounted for by all components or factors. For principal components extraction, this is always equal to 1.0 for correlation analysis.

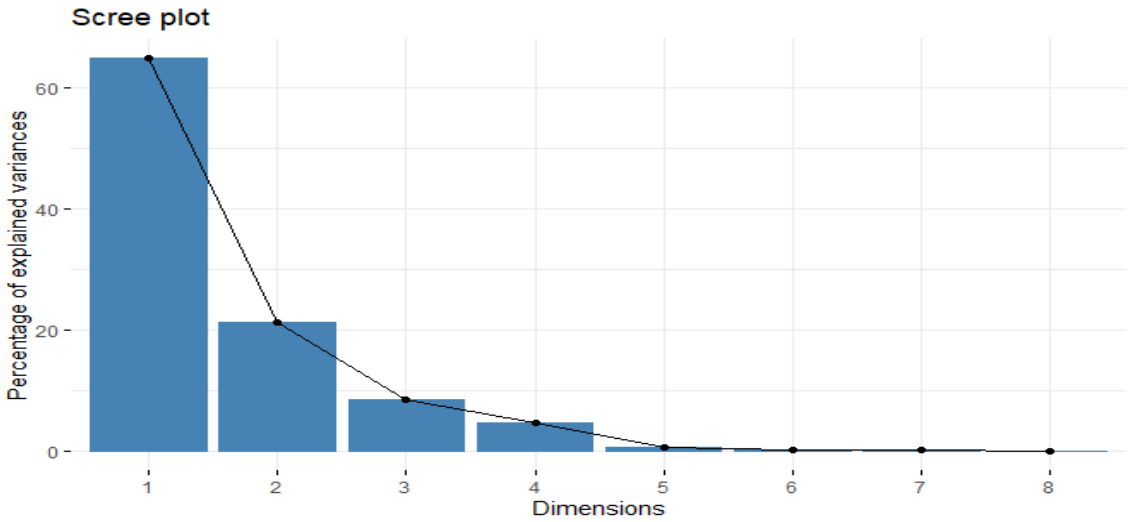
Extraction communalities are estimates of the variance in each variable accounted for by the components. The communalities in this table are the high values and low values . The high value indicates that the extracted components represent the variables are well.

Variance Explained

	RC 1	RC 2
S S Loadings	4.67	2.22
Proportion Variance	0.58	0.28
Cumulative Variance	0.58	0.86

The proportion of variance means the proportion of all the variability in the data,explained away by the rotated comoponent. In our case 58% of variance in the data is explained by RC1,28% by RC2 ,that , is the first two component explains 86% of variation of the data.

Now we use scree plot in factor analysis to visually assess which factor explain most of the variability of the data and also used determine the number of factors that can be extracted



This plot shows the percentage of variances across its Factors.Of course, the variance is the square of the standard deviation and so that you can see that when it has been taken into account the first and second factors big part of the variability in our data has been accounted for

Component Matrix

	Component	
	1	2
Area	.992	.105
Perimeter	.983	.153
Compactness	.654	-.469
Length_of_ kernel	.936	.262
Width_of_kernel	.975	-.039
Asymmetry_coefficient	-.318	.724
Length_of_kernel_groove	.831	.498
Class_(1, 2, 3)	-.437	.776

Extraction Method: Principal Component Analysis.

. 2 components extracted.

The component Matrix shows the correlations between the ites and the components. For some dumb reason, these correlations are called factor loadings.The table contains component loadings,which are the correlations between variable and component .because these are correlations,possible values range from -1 to +1

From the above table shows the loadings of 8 variables on the two factors extracted.The value of the loadings lie between -1 and +1.The higher the absolute value of the loading,the more the factors contributed are variable

Rotated Component Matrix

	Component	
	1	2
Area	.955	-.287
Perimeter	.966	-.240
Compactness	.421	-.685
Length_of_ kernel	.965	-.121
Width_of_kernel	.884	-.414
Asymmetry_coefficient	-.013	.791
Length_of_kernel_groove	.958	.137
Class_(1, 2, 3)	-.103	.884

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Rotation converged in 3 iterations.

Factor 1 :

1) Area
2) Perimeter
3) Length of kernel
4) Width of kernel
5) Length of kernel Groove

Factor 2:

1) Compactness
2) Asymmetry coefficient

On analyzing the rotated component matrix, as a precaution suppressing the small coefficients by giving the absolute value 0.5 so that the rotated component matrix have value greater than 0.5 will be displayed by this cross loading can be avoided .

The rotated component matrix, helps you to determine what the components represent. Sometimes referred to as the loadings, is the key output of principal component analysis. It contains estimates of the correlations between each of the variables and the estimated components. Rotation does not actually change anything but makes the interpretation of the analysis easier

From the table we can see that, the first factor is highly correlated with Perimeter and length of the kernel whereas in the second factor is correlated with asymmetry of coefficient

CHAPTER 5
CONCLUSION

5.1. PRINCIPAL COMPONENT ANALYSIS

Using scree plot, it can be concluded that we can extract two principle components

The component are

$$\mathbf{Y_1 = 0.992X_1 + 0.983X_2 + 0.654X_3 + 0.936X_4 + 0.975X_5 + -0.318X_6 + 0.831X_7 + -0.437X_8}$$

$$\mathbf{Y_2 = 0.105X_1 + 0.153X_2 + -0.469X_3 + 0.262X_4 + -0.39X_5 + 0.724X_6 + 0.498X_7 + 0.776X_8}$$

It is seen that the first component explains 65% of variation of the data set. Second component explains 21% of data set That is, first two components explain 86.063% of variation of data set. It is also from the component matrix that the first two components are highly influenced by all the eight factors.

Also we get, the first component is highly influenced by the variables $\mathbf{X_1}$ followed by **X2, X5** and **X4**,second component is highly influenced by the variables **X6** followed **by X7,X4** and **X2**

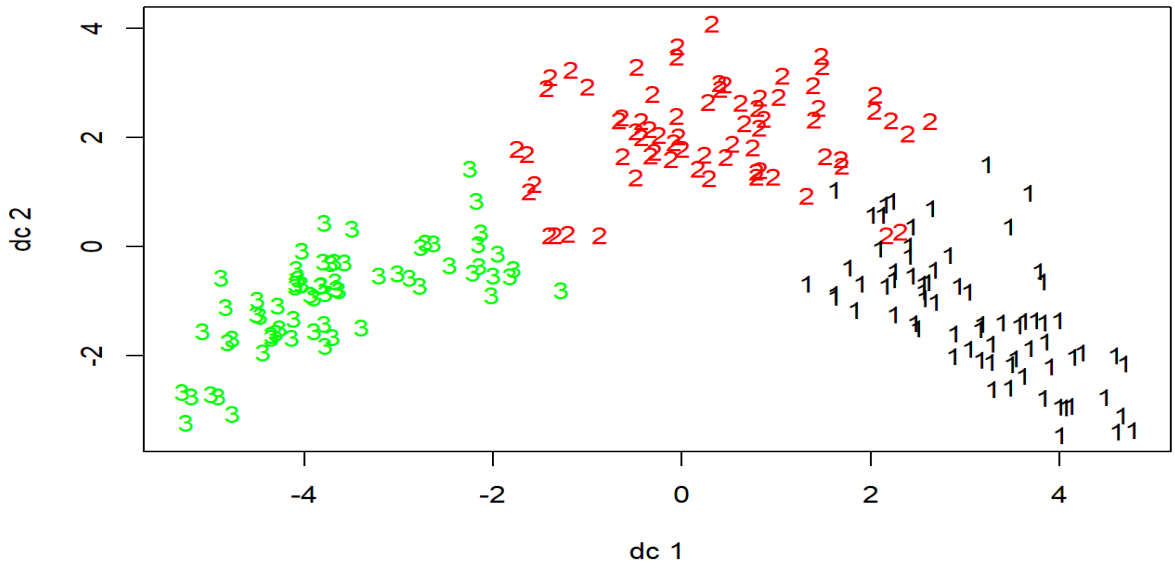
From the component matrix

5.2 CLUSTERING ANALYSIS

K MEANS

Area	Perimeter	Compactness	Length of kernel	Width of kernel	Asymmetry coefficient	Length of Grooves	Class
1.20981	**1.2212**	**0.5160033**	**1.2062015**	**1.1126**	**-0.05527**	**1.2689**	*-0.0349*
-0.1878	*-0.21702*	*0.397688*	*-0.3056103*	*-0.040821*	*-0.6684*	*-0.67545*	*-1.1869*
-1.0219	*-1.00418*	*-0.91369*	*-0.90059*	*-1.071797*	*0.72327*	*-0.5934*	*1.22182*

Looking at the variables in the above table we can get the idea about which group of seeds are better.so according to average value table , the **cluster** one is the one with characteristics of seed have higher values, so that cluster one represents an excellent choice for wheat cultivation.
After conducting average value and cluster plot analysis and validations,it becomes evident that the Canadian seed variety belongs to cluster one and that cluster one indeed offer highly favourable condition for wheat cultivation,or we can say confidently state that the **CANADIAN SEED** variety is well suited for wheat cultivation



WARD'S METHOD AND COMPLETE LINKAGE METHOD

Area	Perimeter	Compactness	Length of kernel	Width of kernel	Asymmetry coefficient	Length of Grooves
-0.222	*-0.2228*	*0.34667*	*-0.339230*	*-0.08512*	*-0.723630*	*-0.6549*
1.211	**1.2145**	**0.56712**	**1.195400**	**1.127899**	**-0.040599**	**1.23972**
-1.022	*-0.9971*	*-0.97027*	*-0.879316*	*-0.87931*	*0.83085*	*-0.58163*

Looking at the variables in the above table we can get the idea about which group of seeds are better.so according to average value table of Ward's method , the cluster 2 is the one with characteristics of seed have higher values. With respect to the second cluster, both cluster 1 and 3 possess similar inferior characteristics.

In this case, Cluster 2, contains a combination of multiple classes ("kama" and "rose"), it can indeed make the interpretation and analysis of the clusters more challenging. When a cluster contains a mixture of different classes or patterns, it suggests that the clustering algorithm may not be effectively capturing the underlying structure in the data, or that the data itself may not exhibit clear and distinct clusters.

So ,by using the Adjusted Rand Index (ARI) to compare clustering methods and selecting K-means over Ward's method because it has a higher ARI value is a reasonable approach to make a choice between the two methods for your analysis.

The ARI is a commonly used external evaluation metric that measures the similarity between the true class labels (or ground truth) and the clustering results. A higher ARI value indicates better agreement between the clustering and the ground truth. In this case, K-means(0.8678) has a larger ARI compared to Ward's method(0.7136), it suggests that K-means is providing better clustering results.

5.3. FACTOR ANALYSIS

Using the scree plot, it can be concluded that we can extract 2 factors

1) The first factor offers a point of attributes related to the overall size and dimensions of the seeds. This factor could represent a measure of seed size ,capturing the spatial dimensions that influence attributes like growth capacity. Seeds with higher factor score on this factor might exhibit larger physical dimension , potentially indicating adaptation for certain ecological niches.
2) The second factor offers a different lens through which to view seed attributes . This factor seems to capture attributes linked to the shape and symmetry of the seeds . High factor score on this factor might imply seeds with greater compactness and symmetry , which could relate to optimized packing efficiency , protection against environmental stress

The factors can be written :-

$$Y_1=.995X_1+.966X_2+.421X_3+.965X_4+.884X_5-.013X_6+.958X_7-103X_8$$

$$Y_2=-.287X_1-.240X_2-.685X_3-.121X_4-.414X_5+.791X_6+.137X_7+.884X_8$$

It is seen that the 1 st component explains 58% of variation of the data set, 2nd component explains 28%.That is, first two component explain 86% of variation of data set. It is also from the component matrix that the first two components are highly influenced by all the 7 factors.

REFERENCES

- ANDERSON.T.W (2003): An Introduction to Multivariate Statistical Analysis, Third Edition, New York, published by *John Wiley & sons.*
- K. BHUYAN (2008), 'Multivariate Analysis and its Applications', *New Central Book Agency (p) Ltd.*
- RICHARD A. JOHNSON & DEAN W. WICHERN (2002),'Applied Multivariate Statistical Analysis', Fifth Edition, *Pearson Education Asia*.

- https://www.kaggle.com/datasets/sushilyeotiwad/wheat-seed-dataset

- https://www.statology.org/elbow-method-in-r/