# Wheat seeds analysis based on their geometrical properties

Mohit Mehndiratta, 20622275
Master of Data Science
Western Sydney University

***Abstract*****—** Cultivation of wheat grass for its seeds and cereal grains are done worldwide. It is the most important and widely used food grain across the globe. There are several types of wheat kernels but here we will discuss only three types of wheat mainly “Kama”, “Rosa” and “Canadian”. Accurate classification of wheat types is required for the companies working under this domain to store the seeds. Nowadays, identification of wheat seeds is done with the labor force and under continuous supervision. However, this is not the productive way to perform this kind of classification manually because there are always a concern of human error and operational workload for employees. Due to this and organization might have to bear some losses as well. For this, Machine Learning can be very helpful to automate this process of classifying wheat seeds based on their kernels. Machine learning or data analysis is about extracting meaningful properties from a data. In this report, we used classification technique to classify and accurately predict whether a particular wheat has a type “Kama” or not.

***Keywords*****—** Wheat analysis, classification, kernels, regression, data analysis, data explore, data preprocessing,

## I. Introduction

Wheat grass is cultivated worldwide for its seeds and cereal grains. It is the most widely used food grain in terms of cereal grains superseding maize and rice as a source of protein. Wheat is eaten in 89 countries across the world. Wheat has various types and ways to be cultivated and originated. Out of various types of wheat in this report we will talk about “Kama”, “Rosa” and “Canadian” wheat types.

In many factories and warehouses where wheat is stored after the production, accurate classification of wheat kernels is required so that they don’t mix with one another. This important for an industrial point of view because mistakes like this can cause serious business losses. In today’s time wheat are mainly classified with labor forces manually, resulting in increased time and manual errors. Therefore, it is important to have an autonomous system that can provide wheat classification based on their geometrical properties. This will result in reduced manual load, time, and mistakes.

Machine learning can be very helpful for automating the classification of wheat seeds as per their types. With the data we can perform various methods for classification and train the model to have maximum accuracy.

In this report, we have used wheat seeds data set that is available at UCI machine learning repository. The data is about wheat kernel types “Kama”, “Rosa” and “Canadian” having their respective geometrical properties. These properties are Area, Perimeter, Compactness, Length of Kernel, Width of Kernel, Asymmetry Coefficient and Length of Kernel Groove.

The main objective of this work is to use a classification technique to correctly classify whether a particular observation of wheat is of a type “Kama” or not. Here, we will use 2 methods of classification techniques under supervised learning and 1 method of unsupervised learning. Post which we will compare the results of each method and find the best model that can accurately classify whether a particular wheat is “Kama” or not. To do this, first data exploration and preprocessing is required to modify and select the appropriate data for this specific task.

## II. Data Description Explore And Pre Processing

Exploring the data is one of the first steps in data preparation. It is a way to get to know the data before working on it. Data exploration allows deep understanding of data and its properties.

### *1. Data Description Explore*

### *A. Describe data*

We have gathered data from UCI machine learning repository and is called “seeds data set”. The data set comprises measurements of structural properties of seeds belonging to three different types of wheat. These types are “Kama”, “Rosa” and “Canadian”. The data set is multivariate. There are a total of 210 observations with 8 variables.

### *B. Explore data*

We are exploring the data using R and R studio as a tool. Upon looking at firs few observations of the seeds data set we can find that there are a total of 210 observations for 8 variables. The variables are Area A, Perimeter P, Compactness $C = 4*\pi*A / P^2$, Length of kernel, Width of kernel, Asymmetry Coefficient, Length of kernel groove and Type.

**TABLE 1: SEEDS DATA SET (First 5 Rows)**

| Area | Perimeter | Compactness | Length | Width | Asymmetry | Length Groove | Type |
|---|---|---|---|---|---|---|---|
| 15.2 | 14.8 | 0.871 | 5.76 | 3.31 | 2.22 | 5.22 | 1 |
| 14.8 | 14.5 | 0.881 | 5.55 | 3.33 | 1.01 | 4.95 | 1 |
| 14.2 | 14.0 | 0.905 | 5.29 | 3.33 | 2.69 | 4.82 | 1 |
| 13.8 | 13.9 | 0.895 | 5.32 | 3.37 | 2.25 | 4.80 | 1 |
| 16.1 | 14.9 | 0.903 | 5.65 | 3.56 | 1.35 | 5.17 | 1 |

The observations are divided into 3 sets of 70 as per its types. All the variables are numeric in nature except “Type”. “Type” is an integer type variable of value 1,2 and 3 for “Kama”, “Rosa” and “Canadian” respectively.

The variable descriptions are as follows:

1. Area represents Area of a kernel/seed.
2. Perimeter is the perimeter of a kernel
3. Compactness is calculated with the help of area and perimeter, formula is C = 4*π*A / $P^2$.
4. Length of Kernel is the total height of the seed.
5. width of kernel represents how wider the seed is.
6. Length of kernel groove depicts the height of the groove inside the kernel.
7. Asymmetry coefficient represents how symmetric a kernel is if we divide the seeds into 2 parts and compare it with left and right area.

From the correlation matrix, it is evident that Area, Perimeter, Length of Kernel, Width of Kernel and Length of Kernel Groove are strongly correlated with one another. While Compactness is somewhat correlated with Width of Kernel. On the other hand, Asymmetry Coefficient shows weak negative correlation with each variable.

Box Plots in figure 1, shows us that according to geometric parameters of seeds, Rosa is the largest seed and Canadian type seeds are the lowest. Compactness of Kama and Rosa are similar while Asymmetry coefficient is the only variable where Canadian seeds has highest value meaning Canadian kernels are more symmetric than others.

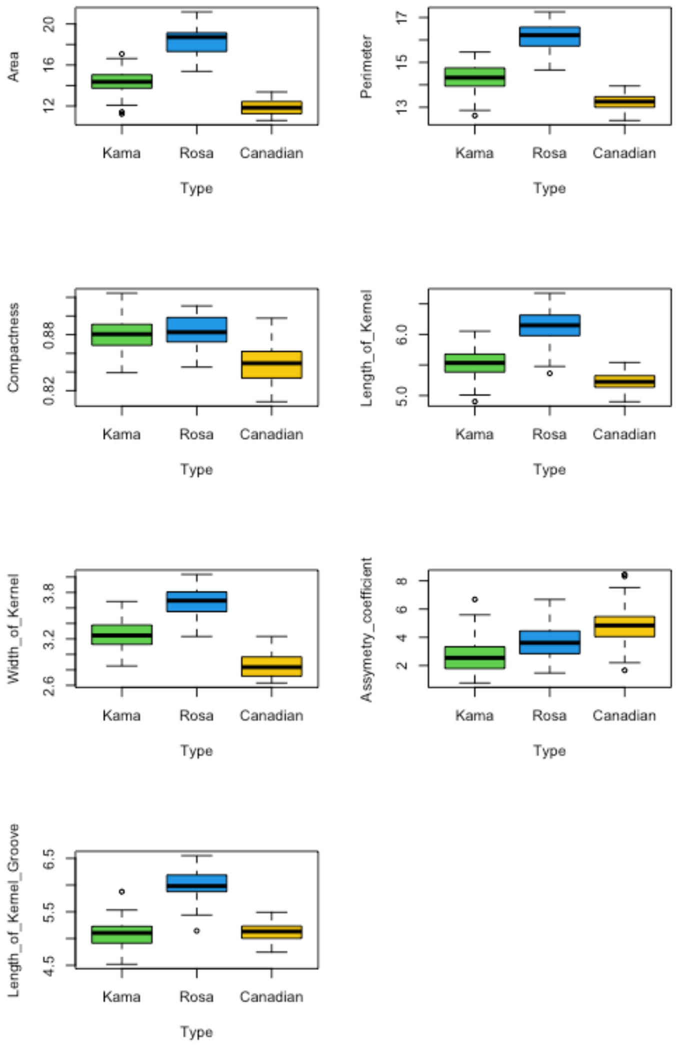


**FIG 1: WHEAT TYPES BASED ON THEIR GEOMETRICAL PROPERTIES**

### C. *Quality of the Data*

The data doesn’t have any null or NA values in there. The data set seems clean as of now.

## 2. *Data Pre-Processing*

Data Preprocessing is a data mining technique that involves transforming some raw data into meaningful format as per our goal. Data often consists of null values, some outliers, not as per our goal, not properly classified and that’s when preprocessing come in before we use the data for model building process.

As per seeds data set, there are few preprocessing of data that is required to finally use the data for our desired goal.

### *A. Construct data*

The seeds data set has a “Type” variable which consists of 3 values mainly 1,2 and 3 indicating “Kama”, “Rosa” and “Canadian” respectively. As per our aim, we need to classify a variable to see if it’s a “Kama” type or not. Therefore, with the use of “Type” variable, where we know 1 represents “Kama”, we will be constructing a new variable named “is_Kama” where 1 represents “Kama” and 0 represents “Not Kama”.

Post constructing this new variable, a new data is formed by sub-setting “Area”, “Perimeter”, “Compactness”, “Length of Kernel”, “Width of Kernel”, “Asymmetry Coefficient” and “Length of Kernel Groove” along with “is Kama” variable.

### *B. Conversion to categorical variable*

From the data exploration, we observed that R recorded “is Kama” variable as an integer type variable while we know that it is a factor variable of 2 levels mainly 0 and 1. “is Kama” being our response variable must be a factor/categorical variable to create the models based on classifications. Therefore, we use R Code to transfer the class of “is Kama” variable from integer to numeric by using “as.factor” function.

### *C. Dividing data into training and testing data set*

Another part of data preprocessing is dividing the data into training and testing data where model is trained on the training data set and then validated on testing data set. This is done so that we can predict model using unseen data which can give us the true accuracy of a model.

As far as our research is concerned, we will randomly divide data into training and testing data, as we will fit the model using training data and then later test the model accuracy using testing data. For this, using the seed value “39” and divided the whole data set into 70% training and 30% testing data set

## III. Aims And Objectives

A warehouse wants to store the seeds of wheat type “Kama” in a particular section, therefore based on seeds geometrical properties they want to classify and identify the correct “Kama” wheat type so that they can store and pack them without any human error.

The Aim of this project is as follows:

1. To correctly classify the “Kama” wheat type from the seeds data set.
2. To understand which of the geometric properties are more important to classify the wheat type “Kama”.
3. To experiment with different classification methods to which methods yields the highest accuracy for the above problem.

To start with, we are going to do Supervised learning now. Under supervised learning, we are using classification technique since our response variable is qualitative in nature. We will implement Logistic Regression and Decision trees as our two methods for classification.

## IV. Method 1 – Logistic Regression

Logistic regression is a statistical analysis method that allows us to study relationship between a qualitative and a quantitative variable. It is a classification technique in which the target variable is of qualitative or categorical type and predictor variables are of quantitative or numeric type. In R, we do logistic regression using “glm” function, which helps in fitting the model to desired formula.

The general equation for logistic regression in classification is:

$$P^{\wedge}(\mathrm{X}) = \frac{\mathrm{e}^{\alpha^{\wedge} + \beta^{\wedge}\mathrm{X}}}{1 + \mathrm{e}^{\alpha^{\wedge} + \beta^{\wedge}\mathrm{X}}}$$

### *A. Model Building*

Model 1:

Firstly, the logistic model is fit with the training data for “is_kama” variable in terms of all seven predictors, which are “Area”, “Perimeter”, “Compactness”, “Length of Kernel”, “Width of Kernel”, “Asymmetry Coefficient” and “Length of Kernel Groove”.

From the output of model 1, replacing coefficients and the variable from generic equation, the equation of model 1 becomes:

$$\mathrm{P(is\_kama)} = \frac{\mathrm{e}^{-921.0192-32.3928\mathrm{Area}+58.5974\mathrm{Perimeter}+506.2653\mathrm{Compactness}+47.3118\mathrm{LengthOfKernel}+7.4518\mathrm{WidthOfKernel}-2.1278\mathrm{AsymmetryCoefficient}-33.0578\mathrm{LengthOfKernelGroove}}}{1+\mathrm{e}^{-921.0192-32.3928\mathrm{Area}+58.5974\mathrm{Perimeter}+506.2653\mathrm{Compactness}+47.3118\mathrm{LengthOfKernel}+7.4518\mathrm{WidthOfKernel}-2.1278\mathrm{AsymmetryCoefficient}-33.0578\mathrm{LengthOfKernelGroove}}}$$

By using the above equation, we can get find the probability for is_kama on a given set of data.

As per our model 1, we find out that every geometrical property of seeds has some significance in determining the “Kama” seeds except the “Width of Kernel”. However, “Asymmetry Coefficient” and “Length of Kernel Groove” have more significance than others. Thus, they may place an important role in identifying “Kama” seeds.

Model 2:

Now, fitting the second model for is_kama where predictor variables will be the significant one from Model 1. In Model 1, we saw that except Width of Kernel each variable had some significance. Therefore, we are fitting is_kama with respect to all variables except “Width of Kernel”.

From the output of model 2, replacing coefficients and the variable from generic equation, the equation of model 2 becomes:

$$P(\text{is\_kama}) = \frac{e^{-928.1039-31.2810\text{Area}+58.3345\text{Perimeter}+533.3551\text{Compactness}+46.4126\text{LengthOfKernel}-2.1278\text{AsymmetryCoefficient}-33.0340\text{LengthOfKernelGroove}}}{1+e^{-928.1039-31.2810\text{Area}+58.3345\text{Perimeter}+533.3551\text{Compactness}+46.4126\text{LengthOfKernel}-2.1278\text{AsymmetryCoefficient}-33.0340\text{LengthOfKernelGroove}}}$$

By using the above equation, we can get find the probability for is_kama on a given set of data.

As per our model 2, we find out that every geometrical property of seeds used in model 2 has some significance in determining the "Kama" seeds. However, "Area", "Compactness", "Asymmetry Coefficient" and "Length of Kernel Groove" are more significant than others. Thus, they may place an important role in identifying "Kama" seeds.

Model 3:

Now, fitting the third model for is_kama where predictor variables will be the one which have high significance from Model 2. In Model 2, we saw that "Area", "Compactness", "Asymmetry Coefficient" and "Length of Kernel Groove" were highly significant than others. Therefore, we are fitting is_kama with respect to them.

From the output of model 3, replacing coefficients and the variable from generic equation, the equation of model 3 becomes:

$$P(\text{is\_kama}) = \frac{e^{90.1129+1.8548\text{Area}-48.0694\text{Compactness}-1.3145\text{AsymmetryCoefficient}-13.3516\text{LengthOfKernelGroove}}}{1+e^{90.1129+1.8548\text{Area}-48.0694\text{Compactness}-1.3145\text{AsymmetryCoefficient}-13.3516\text{LengthOfKernelGroove}}}$$

By using the above equation, we can get find the probability for is_kama on a given set of data.

As per our model 3, we find out that every geometrical property of seeds used in model 3 were highly significant in determining the "Kama" seeds except "Compactness". Therefore, except "Compactness" other properties may place an important role in identifying "Kama" seeds.

Model 4:

Now, fitting the fourth model for is_kama where predictor variables will be the one which have high significance from Model 3. In Model 3, we saw that "Area", "Asymmetry Coefficient" and "Length of Kernel Groove" were highly significant than others. Therefore, we are fitting is_kama with respect to them.

From the output of model 4, replacing coefficients and the variable from generic equation, the equation of model 4 becomes:

$$P(\text{is\_kama}) = \frac{e^{40.016+1.1309\text{Area}-1.153\text{AsymmetryCoefficient}-9.9484\text{LengthOfKernelGroove}}}{1+e^{40.016+1.1309\text{Area}-1.153\text{AsymmetryCoefficient}-9.9484\text{LengthOfKernelGroove}}}$$

By using the above equation, we can get find the probability for is_kama on a given set of data.

As per our model 4, we find out that every geometrical property of seeds used in model 3 were highly significant in determining the "Kama". Therefore, they may place an important role in identifying "Kama" seeds.

### *B. Model Validation*

Validating the model is essential, to ensure the quality of the model. There are several techniques to do cross validation, but we are here doing cross validation using K-fold Cross Validation.

For logistic regression, we will do cross validation using "cv.glm" function on the training data with each model and then comparing them to find out which one is performing better.

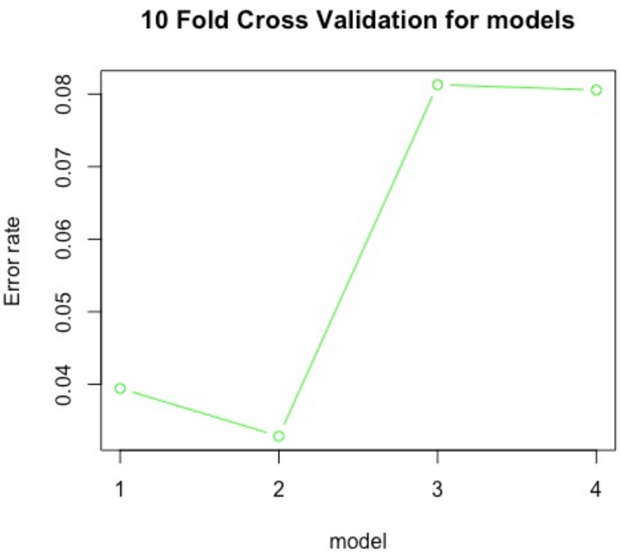


**FIG 2: Cross Validation error rates w.r.t 4 models.**

From Fig 2, we can see the cross-validation error rate with respect to models. We can observe that, error rates are 0.03943348, 0.03284388, 0.08129013, and 0.08056752 for model 1, model 2, model 3 and model 4 respectively.

As of now, it seems like model 2 is the better performing model which has the lowest error rate as compared to other models.

### *C. Model Evaluation*

Now, we need to evaluate which model is the best model and to select that model as our final model for logistic regression.

There are several ways to evaluate logistic regression's model. We will do one by one analysis of all those ways and then select the best model.

Upon looking at the output of cross validation plot in Fig 2, it is quite evident that model 2 has the lowest error rate among other models. Hence, if we consider cross validation for logistic regression model 2 is the best model.

From the summary of all 4 models, we got AIC (Akaike Information Criterion) values. The model which has the lowest AIC value among others considered to perform better than others. From all 4 models we got AIC values as 33.412, 31.479, 82.752, 84.939 for model 1,2,3 and 4 respectively. From this, it is evident that model 2 has lowest AIC value. Thus, if we consider AIC value for logistic regression model 2 is the best model.

From the above ways, we find out that model 2 is the best model. Therefore, model 2 will be the final model for logistic regression.

*D. Model Testing*

We fitted models using training data and identified the best model with the help of cross validation and AIC values. Now, to know how our model is performing we need to test our model accuracy. To do this, we will predict the model values using the testing data that we created earlier and with the help of predicted values we will find confusion matrix and misclassification rate of the model which will tell us how accurate our model is.

As we know our best model is model 2 therefore, we need to test the accuracy of model 2 now with the help of testing data. We calculated the misclassification rate for model 2 as "0.04761905" which is approx. 4.76%. That means our percentage of correctly classifying the "Kama" seeds is around 95.24%

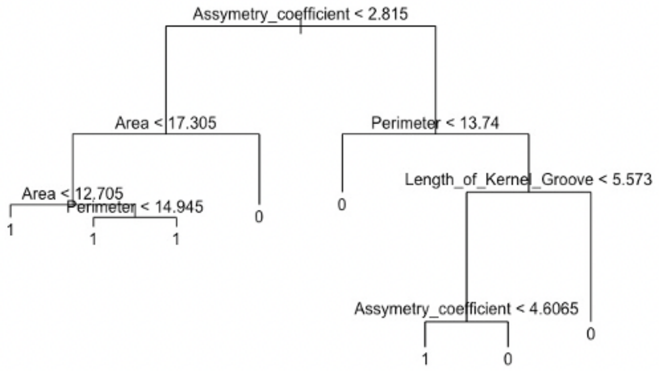
## V. Method 2 – Decision Trees

A decision tree is another form of supervised learning in which a model is represented in a flowchart like structure where a data set is split based on different conditions from the variables. In decision trees, each nodes has some specific conditions forming a tree like structure and the terminal nodes depicts the classification output of the target variable. for a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.

*A. Model Building*

Just like Logistic Regression, we will divide the data set into training and testing data set where we will use training data set to produce a decision trees and then later on test the accuracy using testing data set.

Upon creating the decision tree on the training data set, the summary of the tree shows us that we used 4 variables to form this tree with 8 terminal nodes for predicting "is_kama" and the misclassification error rate for tree is 0.03401.



Assymetry_coefficient < 2.815
Area < 17.305
Perimeter < 13.74
Area < 12.705
Perimeter < 14.945
Length_of_Kernel_Groove < 5.573
1
1
1
0
0
Assymetry_coefficient < 4.6065
0
1
0

**FIG 3: Decision tree on training data set.**

*B. Model Validation*

In decision trees, we used cross validation to validate the decision tree with training data. Cross validation is done using the full tree created with training data. In the full tree we noticed we have 8 terminal nodes. So, cross validation will try to find deviance by creating tree from 1 to 8 terminal nodes which is called tree size. Later we can determine best tree size by looking at the size which has least deviance (number of misclassifications).

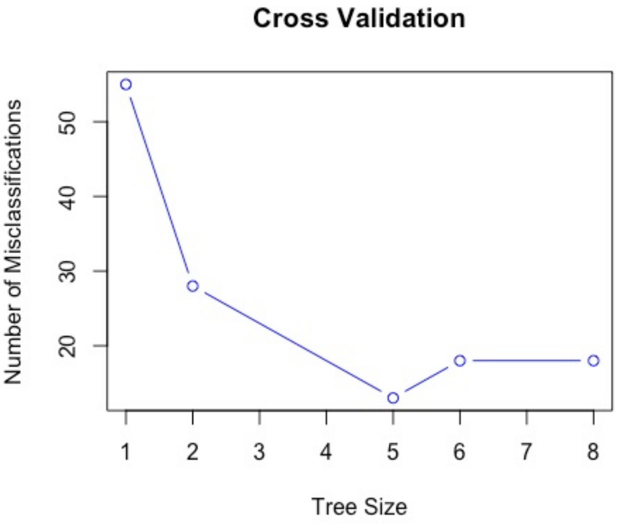


**FIG 4: Cross validations on different tree sizes.**

From fig 4, we can see that, deviance or number of misclassifications is lowest in the tree which has 5 terminal nodes. In other words, a tree size of 5 is the best tree which has lowest rate of misclassification and thus can perform better.

*C. Model Evaluation*

Now, we want to find out which size tree is the best tree which can be best suitable in correctly identifying the "Kama" seeds.

From the cross validations, we find out that out of 8 tree sizes, tree size 5 has the lowest deviance. Therefore, we can choose the tree with size as 5 as our best tree.

Now, to find out the specifications of the tree with size 5, we need to prune the original tree and obtain a new tree which will have 5 terminal nodes. Then, that newly obtained tree will be our best/resultant tree for the "Kama" seeds classification.

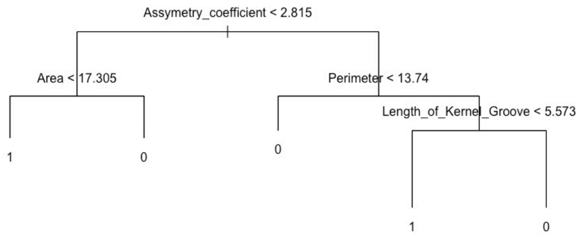

Assymetry_coefficient < 2.815
Area < 17.305
Perimeter < 13.74
Length_of_Kernel_Groove < 5.573
1
0
0
1
0

**FIG 5: Pruned Tree with 5 terminal nodes.**

Fig 5 is our pruned tree which is obtained from original tree. The pruned tree used 4 variables to form this tree with 5 terminal nodes for predicting "is_kama" and the misclassification error rate for the pruned tree is 0.04082.

Pruned tree explanation:

This tree states that, to correctly classify if a seed is "Kama" or not, we need to start with the geometrical property Asymmetry Coefficient.

Our tree representation is like this:

Terminal node 1: If "Asymmetry Coefficient" is less than 2.815 and "Area" is less than 17.305 then that seed can be classified as "Kama" seed.

Terminal node 2: If "Asymmetry Coefficient" is less than 2.815 and "Area" is greater than 17.305 then that seed cannot be classified as "Kama" seed.

Terminal node 3: If "Asymmetry Coefficient" is greater than 2.815 and "Perimeter" is less than 13.74 then that seed cannot be classified as "Kama" seed.

Terminal node 4: If "Asymmetry Coefficient" is greater than 2.815 and "Perimeter" is greater than 13.74 and "Length of Kernel Groove" is less than 5.573 then that seed can be classified as "Kama" seed.

Terminal node 5: If "Asymmetry Coefficient" is greater than 2.815 and "Perimeter" is greater than 13.74 and "Length of Kernel Groove" is greater than 5.573 then that seed cannot be classified as "Kama" seed.

### *D. Model Testing*

We created decision tree using training data and identified the best tree with the help of cross validation and then pruning the original to obtain the best tree. Now, to know how our pruned tree is performing we need to test its accuracy. To do this, we will make the predictions using the testing data that we created earlier and with the help of predicted values we will find confusion matrix and misclassification rate of the pruned tree which will tell us how accurate the tree is.

After predicting values on pruned tree, we calculated the misclassification rate of pruned tree as 0.1428571. This means, we incorrectly classified approx. 14.28% of the testing data. As of now, for our best tree the accuracy stands at 85.72%. approximately.

## VI. Model Comparison

We did supervised learning with Logistic Regression and Decision trees and gathered the best model out of them. Now, we want to understand which method is yielding the best result in terms of our objective.

For logistic regression, the best model was model 2 which was fitted for "is_kama" using the geometrical properties "Area", "Perimeter", "Compactness", "Length of Kernel", "Asymmetry Coefficient" and "Length of Kernel Groove". This means that to correctly classify if a seed is "Kama" or not we will need to use these variables as they can collectively identify the "Kama" seeds accurately.

For decision trees, the best tree was the tree with size 5. This tree used 4 geometrical properties of seeds to identify if a seed is "Kama" or not. The properties used are: "Area", "Perimeter", "Asymmetry Coefficient" and "Length of Kernel Groove". This means that we need to use the combination of these properties to identify the "Kama" seeds accurately.

Collectively we can see that both methods have "Area", "Perimeter", "Asymmetry Coefficient" and "Length of Kernel Groove" in common. This may provide the fact that these geometrical properties can place an important role in classifying the "Kama" seeds correctly.

To compare among these two methods and to find out which methods yields the highest accuracy for the problem, we need to compare their misclassification rate and accuracy that obtained using the same testing data. Misclassification rate for the best model of Logistic regression was approx. 4.76% and accuracy of the model was 95.24% while misclassification rate for the best tree of Decision tree method was approx. 14.28% and accuracy of the pruned tree was 85.72%. Clearly, there is a 9.52% difference between both method's accuracy. Therefore, we can safely say that Logistic Regression method yields the highest accuracy for the problem in which we need to correctly classify the "Kama" seeds using their geometrical properties.

At last, we can say that from our objective point of view if the warehouse implements a classification technique using the best model of Logistic regression, then they can classify if a particular seed is "Kama" or not with appropriate accuracy.

## VII. Unsupervised learning – Clustering

Till now we did a supervised learning to classify seeds based on their geometrical properties and received the best model out of it. Now, we want to find if there can be subgroups within the data by combining the variables together. This is done just to study the variables deeply and to know how they are linked.

Clustering is a broad set of techniques for finding subgroups, clusters, or patterns in a data set. In this we make the partitions of the data in which similar set of observations form a group together which can be a lot different from other created groups. To find the optimal number of clusters required in a data set we look for centroid values and total withinss of the cluster model. If cluster centroid values are getting almost consistent with the increase of cluster, then lowest of that will be preferred. Also, we should look for total withinss, as lower value of that gives more accurate results.

*A. Cluster Building*

Clustering with 2 clusters:

First, we did clustering using with 2 centers first. That means, data will be divided into two subgroups using the kmeans clustering algorithm. As per the outcome of cluster 2 model, 127 observations assigned to group 1 and 83 observations assigned to group 2.

**TABLE 2: Centroids of 2 cluster model**

| Clusters | Area | Perimeter | Compactness | Length of Kernel | Width of Kernel | Asymmetry Coefficient | Length of Kernel Groove |
|---|---|---|---|---|---|---|---|
| 1 | 12.8 | 13.64 | 0.86 | 5.33 | 3.01 | 3.87 | 5.08 |
| 2 | 17.97 | 15.97 | 0.88 | 6.09 | 3.65 | 3.44 | 5.91 |

Clustering with 3 clusters:

Now, finding the outcome of clusters with 3 centers. This means that data will be divided into 3 subgroups where observations are similar inside the groups. As per the outcome of cluster 3 model, 61 observations are assigned to group 1 whereas 77 observations are assigned to group 2 and 72 observations are in group 3.

**TABLE 3: Centroids of 3 cluster model**

| Clusters | Area | Perimeter | Compactness | Length of Kernel | Width of Kernel | Asymmetry Coefficient | Length of Kernel Groove |
|---|---|---|---|---|---|---|---|
| 1 | 11.96 | 13.27 | 0.85 | 5.23 | 2.87 | 4.76 | 5.09 |
| 2 | 14.65 | 14.46 | 0.88 | 5.56 | 3.28 | 2.65 | 5.20 |
| 3 | 18.72 | 16.30 | 0.89 | 6.21 | 3.72 | 3.6 | 6.07 |

Clustering with 4 clusters:

Similarly, outcome of 4 cluster groups is that 66 observations are associated with group 1, 44 observations are assigned to group 2, 46 observations are assigned to group 3 and lastly, 54 observations are assigned to group 4.

**TABLE 4: Centroids of 4 cluster model**

| Clusters | Area | Perimeter | Compactness | Length of Kernel | Width of Kernel | Asymmetry Coefficient | Length of Kernel Groove |
|---|---|---|---|---|---|---|---|
| 1 | 11.91 | 13.26 | 0.85 | 5.23 | 2.86 | 5.07 | 5.11 |
| 2 | 18.96 | 16.40 | 0.89 | 6.24 | 3.75 | 3.54 | 6.10 |
| 3 | 15.72 | 14.98 | 0.88 | 5.75 | 3.4 | 3.19 | 5.45 |
| 4 | 13.39 | 13.87 | 0.87 | 5.37 | 3.11 | 2.42 | 5 |

From Table 2,3 and 4 we can see that the cluster centroids are getting to similar level as we increase the cluster. We can see there isn't much difference in centroids for cluster 3 and 4. Hence we can estimate from here that cluster with 3 groups might be the best.

*B. Cluster Evaluation*

We can evaluate which group cluster is the best by looking at their total withinss values. The cluster with lower number of total withinss value is said to be the better.

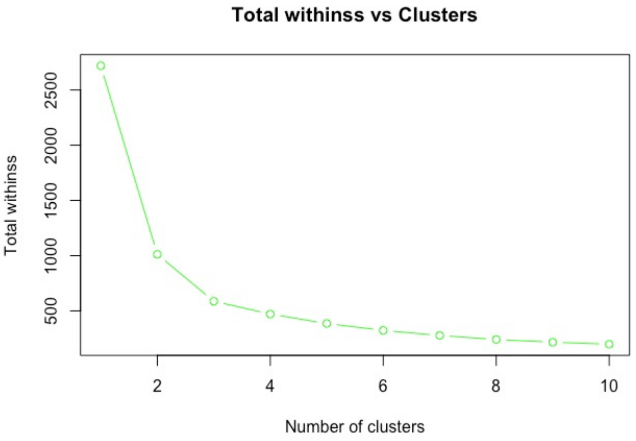


**FIG 6: Comparing total withinss of 10 clusters**

In fig 6, we compared total withinss of clusters from 1 group till 10 clusters group and obtained an elbow shaped graph. By looking at the plot we can see that as we increase the clusters their total withinss value is getting lesser, but their values do not vary much starting from cluster with 3 groups. After the clusters with 3 groups there isn't any significant drop in value. Therefore, from this we can safely say that cluster with group is the best cluster for this data set.

Next, we will visualize the plot of clusters with 3 groups with the help of PCA (Principal Component Analysis).

### *C. Cluster Visualization*

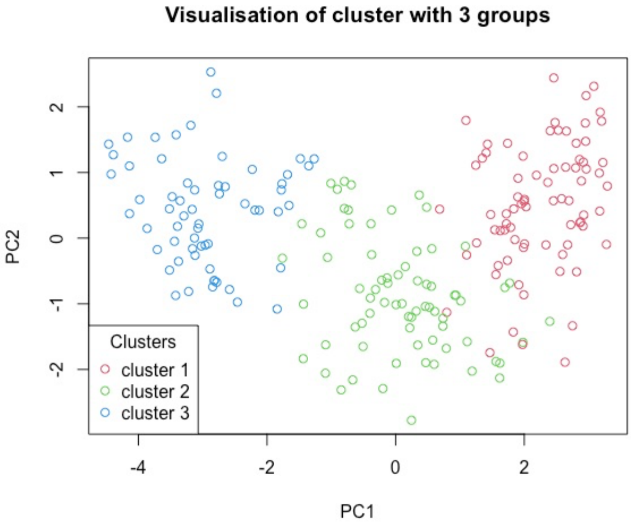


**FIG 7: Plot for 3 cluster groups**

Fig 7 represents the data set obtained within 3 cluster groups. Here, we can clearly see that observations that are similar are forming the cluster with a given centroid value.

## VIII. Results & Recommendations

Until now, we have explored the data set, analyzed it, and then performed some classification with Logistic Regression and Decision Trees from Supervised learning.

From our supervised learning methods, we concluded that Model 2 of Logistic Regression classified the “Kama” seeds with approx. 95.24% accuracy. Model 2 was fitted for “is_kama” variable against the geometrical properties “Area”, “Perimeter”, “Compactness”, “Length of Kernel”, “Asymmetry Coefficient” and “Length of Kernel Groove”. And out of these geometrical properties “Area”, “Compactness”, “Asymmetry Coefficients”, and “Length of Kernel Groove” are more significant and can be vital for identifying the “Kama” seeds. From this it is evident that these variables are more important in correctly classifying the Kama wheat seeds.

As per our objective, we wanted to identify “Kama” wheat type seeds correctly and to understand which of the geometrical properties of seeds are more important that can separate “Kama” wheat seeds from others. From this statistical analysis we found out that “Kama” wheat seed can be identified by using its geometrical properties. Out of which, strong properties are Area, Compactness, Asymmetry Coefficient and Length of Kernel Groove.

## IX. References

***1.*** ***https://wheat.org/wheat-in-the-world/***
***2.*** ***https://archive.ics.uci.edu/ml/datasets/seeds***
***3.*** ***https://towardsdatascience.com/***
***4.*** ***An introduction to Statistical Analysis with Applications in R textbook***
***5.*** ***Lecture Notes – Dr. Liwan (DS – WSU)***

# Appendix

### Loading the necessary libraries

```
library(boot)
library(tree)
```

### Loading the data set and setting the seed value.

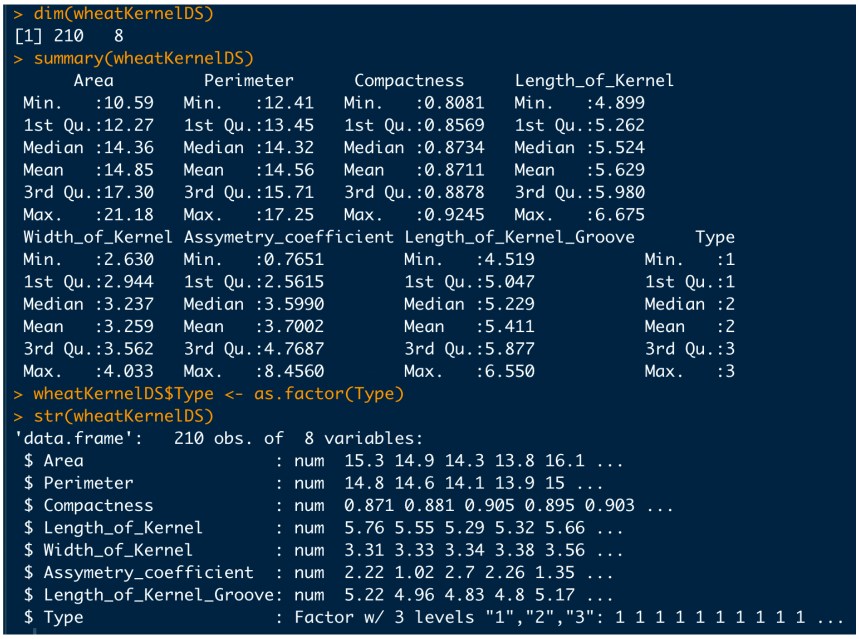
```
wheatKernelDS <- read.csv("wheat_kernels.csv")
wheat_modified <- wheatKernelDS ## created a copy of the data set for later use
attach(wheatKernelDS)
set.seed(39)
```

## Data Exploration

### Understanding the data set and exploring the variables.

```
head(wheatKernelDS)
dim(wheatKernelDS)
summary(wheatKernelDS)
wheatKernelDS$Type <- as.factor(Type)
str(wheatKernelDS)
```

**Output:**



```
> dim(wheatKernelDS)
[1] 210   8
> summary(wheatKernelDS)
      Area         Perimeter      Compactness     Length_of_Kernel
 Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899
 1st Qu.:12.27   1st Qu.:13.45   1st Qu.:0.8569   1st Qu.:5.262
 Median :14.36   Median :14.32   Median :0.8734   Median :5.524
 Mean   :14.85   Mean   :14.56   Mean   :0.8711   Mean   :5.629
 3rd Qu.:17.30   3rd Qu.:15.71   3rd Qu.:0.8878   3rd Qu.:5.980
 Max.   :21.18   Max.   :17.25   Max.   :0.9245   Max.   :6.675
 Width_of_Kernel Assymetry_coefficient Length_of_Kernel_Groove      Type
 Min.   :2.630   Min.   :0.7651        Min.   :4.519           Min.   :1
 1st Qu.:2.944   1st Qu.:2.5615        1st Qu.:5.047           1st Qu.:1
 Median :3.237   Median :3.5990        Median :5.229           Median :2
 Mean   :3.259   Mean   :3.7002        Mean   :5.411           Mean   :2
 3rd Qu.:3.562   3rd Qu.:4.7687        3rd Qu.:5.877           3rd Qu.:3
 Max.   :4.033   Max.   :8.4560        Max.   :6.550           Max.   :3
> wheatKernelDS$Type <- as.factor(Type)
> str(wheatKernelDS)
'data.frame':   210 obs. of  8 variables:
 $ Area                   : num  15.3 14.9 14.3 13.8 16.1 ...
 $ Perimeter              : num  14.8 14.6 14.1 13.9 15 ...
 $ Compactness            : num  0.871 0.881 0.905 0.895 0.903 ...
 $ Length_of_Kernel       : num  5.76 5.55 5.29 5.32 5.66 ...
 $ Width_of_Kernel        : num  3.31 3.33 3.34 3.38 3.56 ...
 $ Assymetry_coefficient  : num  2.22 1.02 2.7 2.26 1.35 ...
 $ Length_of_Kernel_Groove: num  5.22 4.96 4.83 4.8 5.17 ...
 $ Type                   : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
```

In this data set, there are 210 observations of 8 variables. The variables are the geometrical properties of wheat seeds. Except Type all the variables are numerical in nature. Type is a factor variable of 3 levels. Type 1 represents “Kama” wheat seeds, Type 2 represents “Rosa” wheat seeds and Type 3 represents “Canadian” wheat seeds.

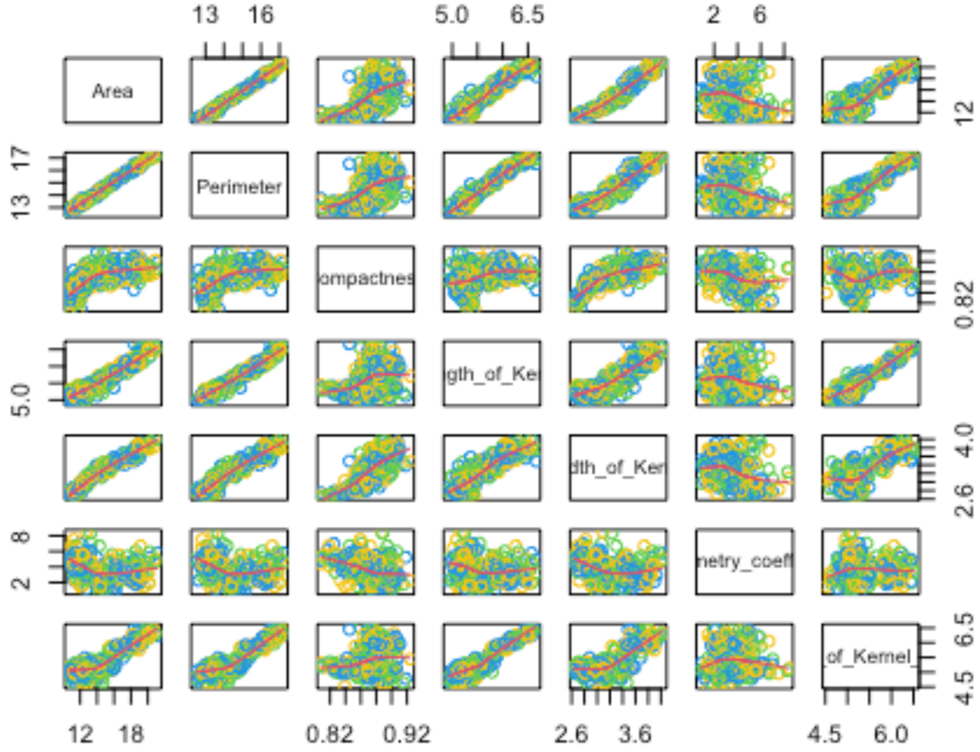
### check if there is any null value.

```
## checking if there is any null value in the data set
sum(is.na(wheatKernelDS))
```

From the output of this, we can see that the value returned is 0. Therefore, there isn’t any null values.

### explore the matrices of the variables with plot

```
## see the pairs
pairs(wheatKernelDS[,1:(length(wheatKernelDS) - 1)],
    panel = panel.smooth, col = c(3,4,7))
```



### understanding the relationship and variance within the variables

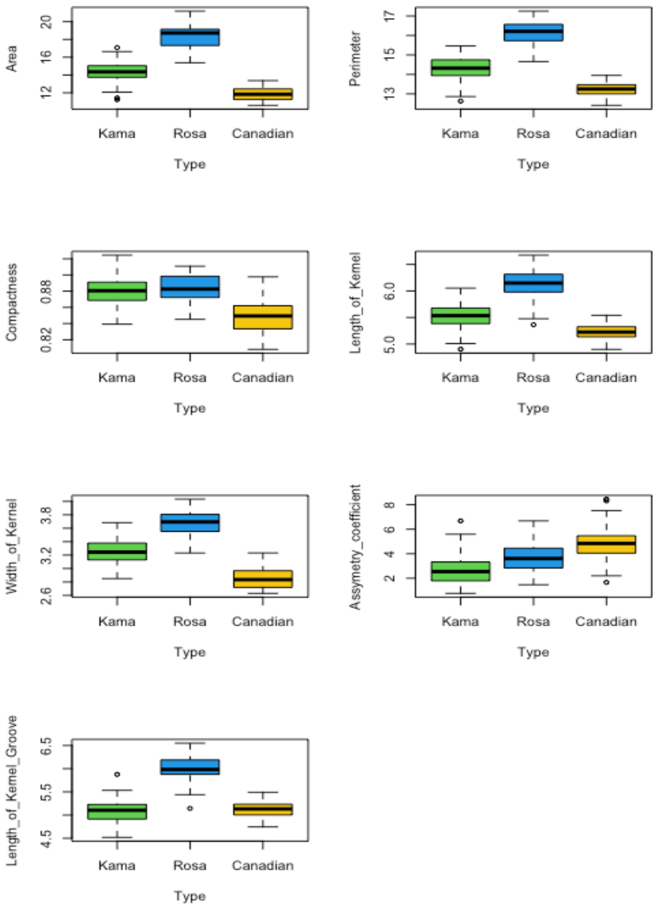
```
## Correlation matrix
cor(wheatKernelDS[,1:(length(wheatKernelDS) - 1)])
## Covariance matrix
cov(wheatKernelDS[,1:(length(wheatKernelDS) - 1)])
```

we can observe that each variable is positively correlated except the Asymmetry Coefficient which is negatively correlated with each other.

### creating the boxplot to visualise each variable with each Type

```
## Creating a box plot visualisation
labels <- c("Kama", "Rosa", "Canadian")
par(mfrow = c(4,2))
boxplot(Area~Type,names = labels, col = c(3,4,7))
boxplot(Perimeter~Type,names = labels, col = c(3,4,7))
boxplot(Compactness~Type,names = labels, col = c(3,4,7))
boxplot(Length_of_Kernel~Type,names = labels, col = c(3,4,7))
boxplot(Width_of_Kernel~Type,names = labels, col = c(3,4,7))
boxplot(Assymetry_coefficient~Type,names = labels, col = c(3,4,7))
boxplot(Length_of_Kernel_Groove~Type,names = labels, col = c(3,4,7))
```

**Output:**



From the boxplot Rosa wheat seeds are the largest among the mentioned categories while Canadian is the smallest in terms of geometrical properties. Kama wheat type looks neutral with respect to every variable.

Boxplot also showed that, besides being compact Canadian wheat seeds are highly symmetric in nature.

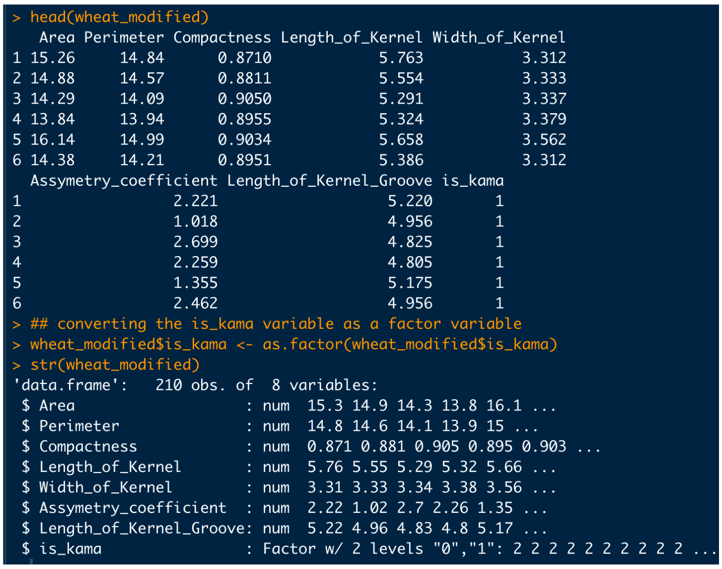
## Data Pre-processing

### create a variable is_kama and create a new data frame which have is_kama as the qualitative variable except Type.

```
# Forming a new variable is_Kama that will contain to values mainly 0 and 1.
is_kama <- wheat_modified$Type
is_kama[is_kama != 1] = 0
```

```
## adding a new column is_kama to the data frame.
wheat_modified <- cbind(wheat_modified, is_kama)

# sub setting the data set to remove Type Column because we don't need it
wheat_modified <- subset(wheat_modified, select = -Type)
attach(wheat_modified) head(wheat_modified)
## converting the is_kama variable as a factor variable
wheat_modified$is_kama <- as.factor(wheat_modified$is_kama)
str(wheat_modified)
```

**Output:**



```
> head(wheat_modified)
   Area Perimeter Compactness Length_of_Kernel Width_of_Kernel
1 15.26     14.84      0.8710            5.763           3.312
2 14.88     14.57      0.8811            5.554           3.333
3 14.29     14.09      0.9050            5.291           3.337
4 13.84     13.94      0.8955            5.324           3.379
5 16.14     14.99      0.9034            5.658           3.562
6 14.38     14.21      0.8951            5.386           3.312
  Assymetry_coefficient Length_of_Kernel_Groove is_kama
1                 2.221                   5.220       1
2                 1.018                   4.956       1
3                 2.699                   4.825       1
4                 2.259                   4.805       1
5                 1.355                   5.175       1
6                 2.462                   4.956       1
> ## converting the is_kama variable as a factor variable
> wheat_modified$is_kama <- as.factor(wheat_modified$is_kama)
> str(wheat_modified)
'data.frame':	210 obs. of  8 variables:
 $ Area                   : num  15.3 14.9 14.3 13.8 16.1 ...
 $ Perimeter              : num  14.8 14.6 14.1 13.9 15 ...
 $ Compactness            : num  0.871 0.881 0.905 0.895 0.903 ...
 $ Length_of_Kernel       : num  5.76 5.55 5.29 5.32 5.66 ...
 $ Width_of_Kernel        : num  3.31 3.33 3.34 3.38 3.56 ...
 $ Assymetry_coefficient  : num  2.22 1.02 2.7 2.26 1.35 ...
 $ Length_of_Kernel_Groove: num  5.22 4.96 4.83 4.8 5.17 ...
 $ is_kama                : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

### divide the data set into training and testing data set.

```
set.seed(39)
# Dividing into 70% training data and 30% testing data.
training_indexes <- sample(1:nrow(wheat_modified),
                    nrow(wheat_modified) * 0.7)

wheat_training <- wheat_modified[training_indexes,]
wheat_testing <- wheat_modified[-training_indexes,]
```

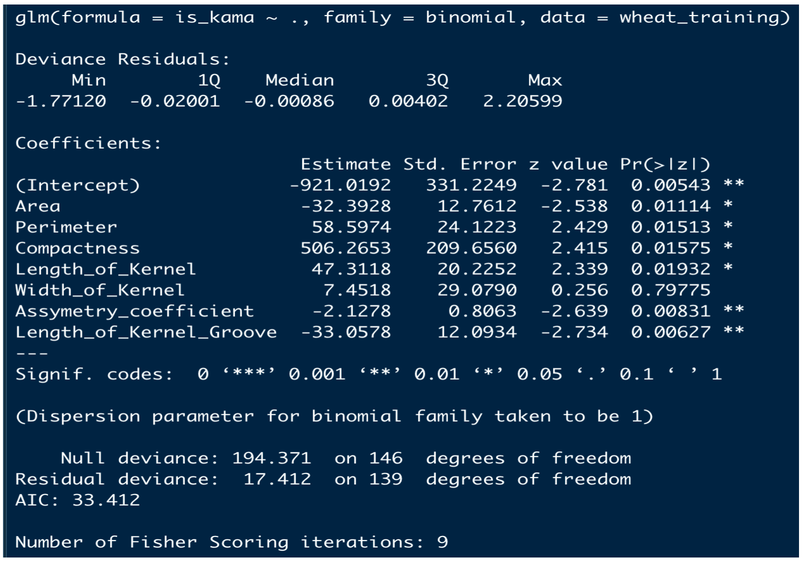
## Method 1 - Logistic Regression

### Model building

### fit a logistic regression model with training data set for is_kama variable

```
lr_model1 <- glm(is_kama~., data = wheat_training, family = binomial)
summary(lr_model1)
```

**Output:**



```
glm(formula = is_kama ~ ., family = binomial, data = wheat_training)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.77120  -0.02001  -0.00086   0.00402   2.20599  

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)   
(Intercept)             -921.0192   331.2249  -2.781  0.00543 **
Area                     -32.3928    12.7612  -2.538  0.01114 * 
Perimeter                 58.5974    24.1223   2.429  0.01513 * 
Compactness              506.2653   209.6560   2.415  0.01575 * 
Length_of_Kernel          47.3118    20.2252   2.339  0.01932 * 
Width_of_Kernel            7.4518    29.0790   0.256  0.79775   
Assymetry_coefficient     -2.1278     0.8063  -2.639  0.00831 **
Length_of_Kernel_Groove  -33.0578    12.0934  -2.734  0.00627 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.371  on 146  degrees of freedom
Residual deviance:  17.412  on 139  degrees of freedom
AIC: 33.412

Number of Fisher Scoring iterations: 9
```

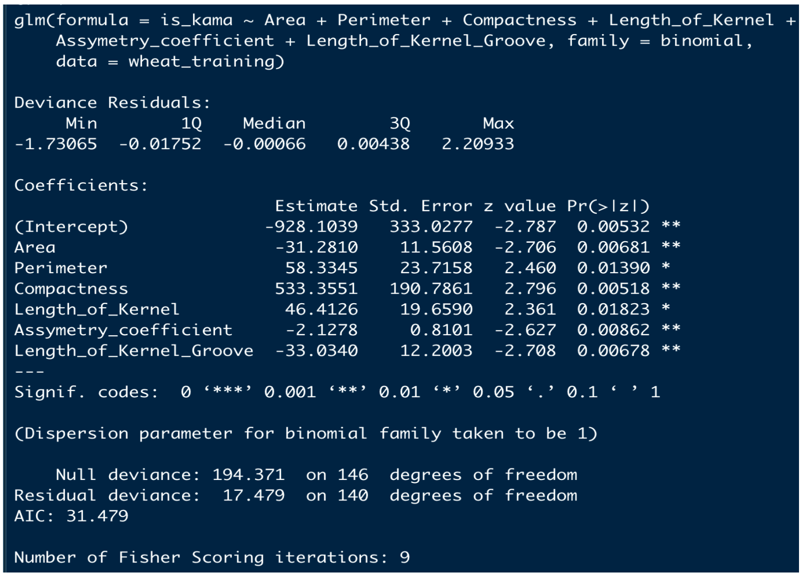
From the summary we can see that, except width of Kernel, every geometrical property has some significance. Asymmetry and length of kernel groove are highly significant.

The AIC value for this model is 33.412

fit a second model with significant variables from first model.

```
lr_model2 <- glm(is_kama~Area+Perimeter+Compactness+Length_of_Kernel+
                  Assymetry_coefficient+Length_of_Kernel_Groove,
                data = wheat_training, family = binomial)
summary(lr_model2)
```

**Output:**



```
glm(formula = is_kama ~ Area + Perimeter + Compactness + Length_of_Kernel + 
    Assymetry_coefficient + Length_of_Kernel_Groove, family = binomial, 
    data = wheat_training)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.73065  -0.01752  -0.00066   0.00438   2.20933  

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)   
(Intercept)             -928.1039   333.0277  -2.787  0.00532 **
Area                     -31.2810    11.5608  -2.706  0.00681 **
Perimeter                 58.3345    23.7158   2.460  0.01390 * 
Compactness              533.3551   190.7861   2.796  0.00518 **
Length_of_Kernel          46.4126    19.6590   2.361  0.01823 * 
Assymetry_coefficient     -2.1278     0.8101  -2.627  0.00862 **
Length_of_Kernel_Groove  -33.0340    12.2003  -2.708  0.00678 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.371  on 146  degrees of freedom
Residual deviance:  17.479  on 140  degrees of freedom
AIC: 31.479

Number of Fisher Scoring iterations: 9
```

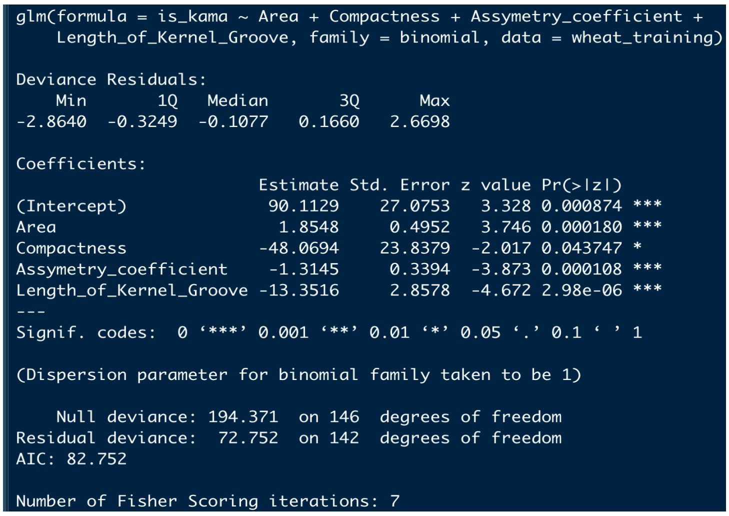
From the summary we can see that, every geometrical property in this model has some significance. Area, Compactness, Asymmetry and length of kernel groove are highly significant.

The AIC value for this model is 31.479

### fit a third model with highly significant variables from second model.

```
lr_model3 <- glm(is_kama~Area+Compactness+
               Assymetry_coefficient+Length_of_Kernel_Groove,
             data = wheat_training, family = binomial)
summary(lr_model3)
```

**Output:**



```
glm(formula = is_kama ~ Area + Compactness + Assymetry_coefficient +
    Length_of_Kernel_Groove, family = binomial, data = wheat_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8640  -0.3249  -0.1077   0.1660   2.6698

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              90.1129    27.0753   3.328 0.000874 ***
Area                      1.8548     0.4952   3.746 0.000180 ***
Compactness             -48.0694    23.8379  -2.017 0.043747 *
Assymetry_coefficient    -1.3145     0.3394  -3.873 0.000108 ***
Length_of_Kernel_Groove -13.3516     2.8578  -4.672 2.98e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.371  on 146  degrees of freedom
Residual deviance:  72.752  on 142  degrees of freedom
AIC: 82.752

Number of Fisher Scoring iterations: 7
```

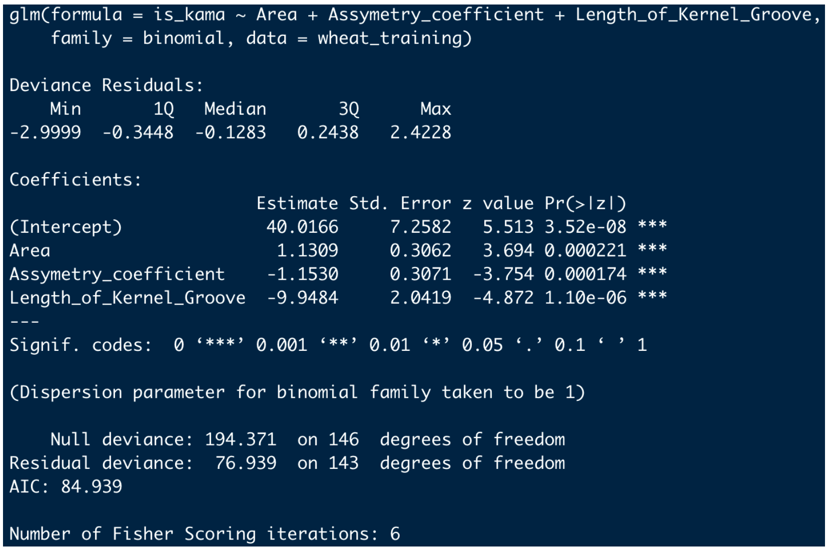
From the summary we can see that, every geometrical property in this model has some significance. Area, Asymmetry and length of kernel groove are highly significant.

The AIC value for this model is 82.752

### fit a fourth model with highly significant variables from third model.

```
lr_model4 <- glm(is_kama~Area+
               Assymetry_coefficient+Length_of_Kernel_Groove,
             data = wheat_training, family = binomial)
summary(lr_model4)
```

**Output:**



```
glm(formula = is_kama ~ Area + Assymetry_coefficient + Length_of_Kernel_Groove,
    family = binomial, data = wheat_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9999  -0.3448  -0.1283   0.2438   2.4228

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              40.0166     7.2582   5.513 3.52e-08 ***
Area                      1.1309     0.3062   3.694 0.000221 ***
Assymetry_coefficient    -1.1530     0.3071  -3.754 0.000174 ***
Length_of_Kernel_Groove  -9.9484     2.0419  -4.872 1.10e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.371  on 146  degrees of freedom
Residual deviance:  76.939  on 143  degrees of freedom
AIC: 84.939

Number of Fisher Scoring iterations: 6
```

From the summary we can see that, every geometrical property in this model has some significance. Area, Asymmetry and length of kernel groove are highly significant.

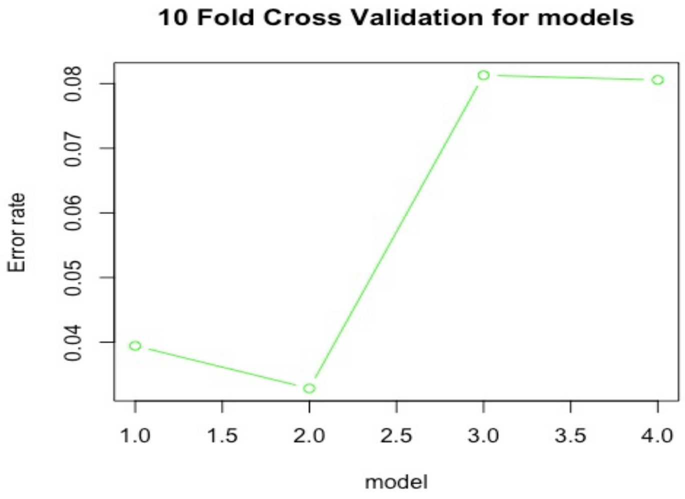
The AIC value for this model is 84.939

## Model validation

### Validate all those 4 models with cross validation by using K-Fold Cross validation technique.

```
##### Cross Validation
dev.off()
model <- c(1:4)
cv_errorKF= rep (0,4)
cv_errorKF[1] <- cv.glm(wheat_training,lr_model1, K=10)$delta[1]
cv_errorKF[2] <- cv.glm(wheat_training,lr_model2, K=10)$delta[1]
cv_errorKF[3] <- cv.glm(wheat_training,lr_model3, K=10)$delta[1]
cv_errorKF[4] <- cv.glm(wheat_training,lr_model4, K=10)$delta[1]

cv_errorKF
plot(cv_errorKF~model, type="b", col="green",
    ylab = "Error rate", main = "10 Fold Cross Validation for models",
    xaxt="n", xlim = c(1,4))
axis(1, at = 1:4)
```

**Output:**



## Model evaluation

From the cross-validation plot, we saw that model 2 has the least error rate and if we compare the AIC values, the lesser the AIC value better the model. So, with this also model has least AIC value. Therefore, Model 2 could be the best model for this logistic regression.

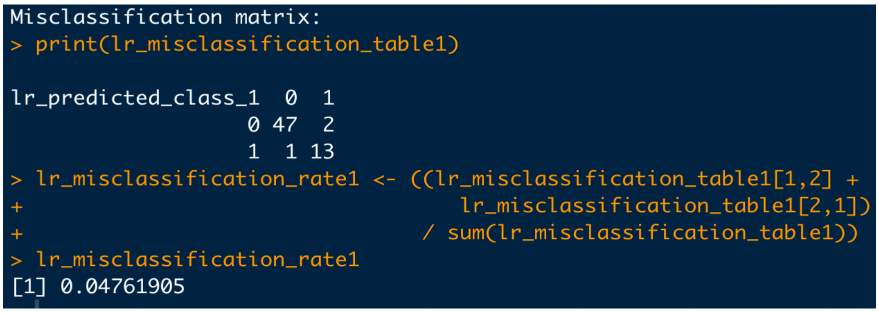
## Model testing

Now, after finding the best model, we need to check the accuracy of the best model to ensure that the results are true for every data. To do this we do testing on the testing data and find its confusion matrix and misclassification rate to know how well the model is performing.

```
lr_prediction1 <- predict(lr_model2, newdata = wheat_testing,
                        type = "response")
## creating a prediction class for misclassification rate
lr_predicted_class_1 <- rep(0, nrow(wheat_testing))

lr_predicted_class_1[lr_prediction1 > 0.5] = 1
```

```
## Calculating misclassification rate

lr_misclassification_table1 <- table(lr_predicted_class_1,
                                wheat_testing$is_kama)


cat("Misclassification matrix: ")
print(lr_misclassification_table1)

lr_misclassification_rate1 <- ((lr_misclassification_table1[1,2] +
                              lr_misclassification_table1[2,1])
                            / sum(lr_misclassification_table1))
lr_misclassification_rate1
```

**Output:**



```
Misclassification matrix:
> print(lr_misclassification_table1)

lr_predicted_class_1  0  1
                    0 47  2
                    1  1 13
> lr_misclassification_rate1 <- ((lr_misclassification_table1[1,2] +
+                                  lr_misclassification_table1[2,1])
+                               / sum(lr_misclassification_table1))
> lr_misclassification_rate1
[1] 0.04761905
```

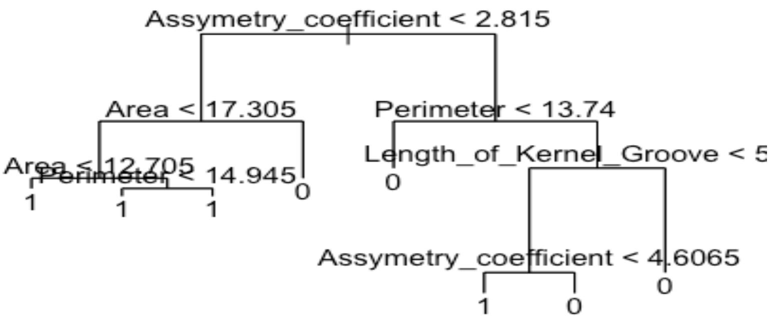
we can see that, the misclassification rate we got from model 2 is 0.04761905 which is approx. 4.76%. That means our model performed approx. 95.24% accurate result.

## Method 2 - Decision trees

### Model building

### create a full size decesion tree for the training data

```
class_tree_training <- tree(is_kama~., data = wheat_training)
summary(class_tree_training)
plot(class_tree_training,
    main = "Classification tree on training data")
text(class_tree_training, pretty = 0)
```


Assymetry_coefficient < 2.815
Area < 17.305
Perimeter < 13.74
Area < 12.705
Perimeter < 14.945
Length_of_Kernel_Groove < 5
0
1
1
1
0
Assymetry_coefficient < 4.6065
0
1
0

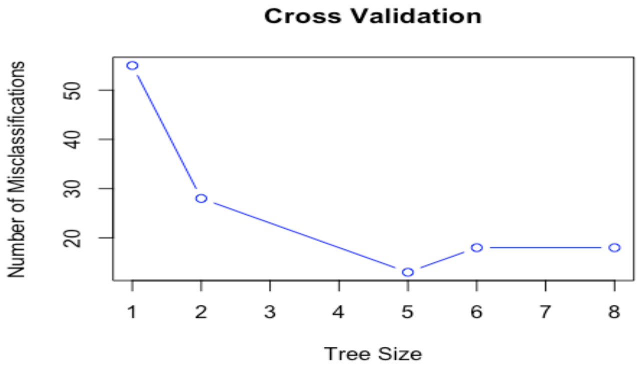
This decision tree is formed on training data set for wheat side to identify the outcome and path to classify the “Kama” wheat seeds.

This decesion tree used only 4 variables which are “Assymetry_coefficient”, “Area”, “Perimeter”, and “Length_of_Kernel_Groove”. This tree has 8 terminal nodes to classify the “Kama” wheat type. Misclassification rate for this tree is 0.03401 which is 3.4%.

## Model Validation

### validate the tree with the help of cross validation

```
cv_classtree <- cv.tree(class_tree_training, FUN = prune.misclass)
cv_classtree
plot(cv_classtree$size, cv_classtree$dev, type = "b", xlab = "Tree Size",
    ylab = "Number of Misclassifications",
    main = "Cross Validation", col = "blue")
```



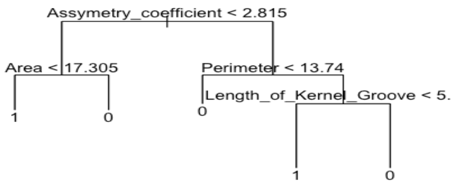
From the cross-validation plot, it is evident that tree with size 5 has the smaller number of misclassifications. That means decision tree with size 5 should be the best model.

## Model Evaluation

### find the model of best size by pruning it

Now we got to know that tree size 5 is optimal. To get the specifications of the tree we need to prune the tree to get the resultant tree with size 5.

```
pruned_classtree <- prune.misclass(class_tree_training, best = 5)
summary(pruned_classtree)
plot(pruned_classtree, main = "Pruned tree with 5 terminal nodes")
text(pruned_classtree, pretty = 0)
```



Here we got the pruned tree, which has 5 terminal nodes and used 4 variables.

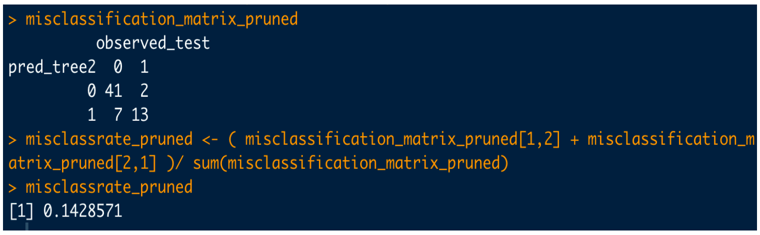
Our tree representation is like this: Terminal node 1: If “Asymmetry Coefficient” is less than 2.815 and “Area” is less than 17.305 then that seed can be classified as “Kama” seed. Terminal node 2: If “Asymmetry Coefficient” is less than 2.815 and “Area” is greater than 17.305 then that seed cannot be classified as “Kama” seed. Terminal node 3: If “Asymmetry Coefficient” is greater than 2.815 and “Perimeter” is less than 13.74 then that seed cannot be classified as “Kama” seed. Terminal node 4: If “Asymmetry Coefficient” is greater than 2.815 and “Perimeter” is greater than 13.74 and “Length of Kernel Groove” is less than 5.573 then that seed can be classified as “Kama” seed. Terminal node 5: If “Asymmetry Coefficient” is greater than 2.815 and “Perimeter” is greater than 13.74 and “Length of Kernel Groove” is greater than 5.573 then that seed cannot be classified as “Kama” seed.

## Model Testing

### predicting on testing data set to find the misclassification rate.

```
#### Testing the accuracy / Calculating misclassification rate

pred_tree2 <- predict(pruned_classtree, newdata = wheat_testing, type = "class")

observed_test <- wheat_testing[, "is_kama"]

misclassification_matrix_pruned <- table(pred_tree2, observed_test)
misclassification_matrix_pruned

misclassrate_pruned <- ( misclassification_matrix_pruned[1,2] + misclassification_matrix_pruned[2,1] )/ sum(misclassification_matrix_pruned)

misclassrate_pruned
```

**Output:**



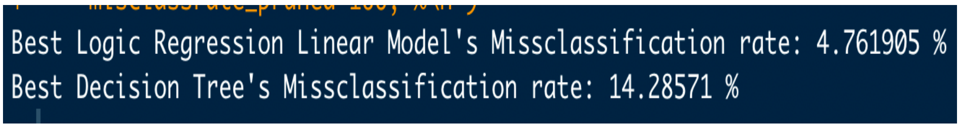
```
> misclassification_matrix_pruned
          observed_test
pred_tree2  0  1
         0 41  2
         1  7 13
> misclassrate_pruned <- ( misclassification_matrix_pruned[1,2] + misclassification_matrix_pruned[2,1] )/ sum(misclassification_matrix_pruned)
> misclassrate_pruned
[1] 0.1428571
```

from this, we can see that pruned tree obtained 14.28% misclassification rate. That means the accuracy from the final model of the decision tree is approx. 85.72%.

## Model Comparison

```
cat("Best Logic Regression Linear Model's Missclassification rate:"
   , lr_misclassification_rate1*100,"%\nBest Decision Tree's Missclassification rate:",
   misclassrate_pruned*100,"%\n")
```

**Output:**



```
+     misclassrate_pruned*100,"%\n")
Best Logic Regression Linear Model's Missclassification rate: 4.761905 %
Best Decision Tree's Missclassification rate: 14.28571 %
```

From this, we can see that logistic regression model is better because it correctly classified more observation than decision trees to identify if a seed is “Kama” or not.

## Clustering

### Cluster Building

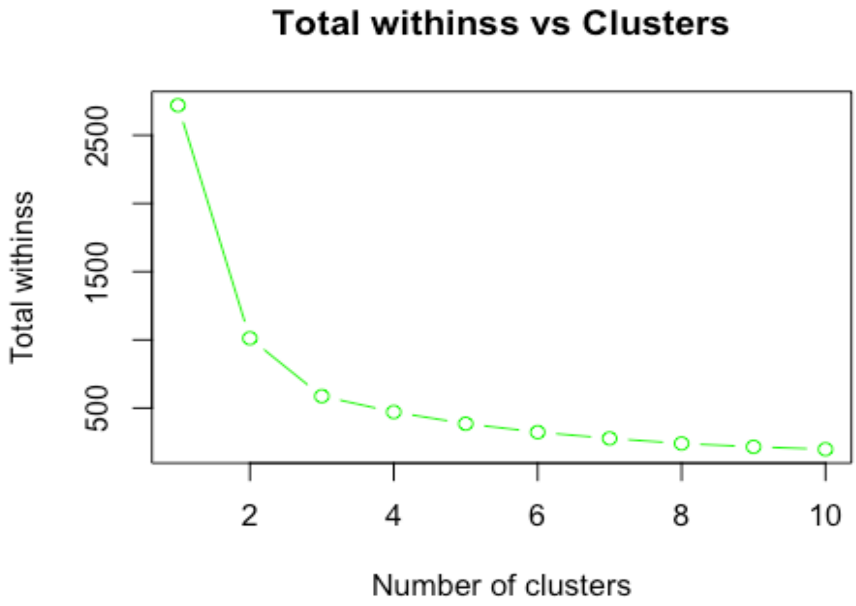
cluster the data with 2,3 and 4 centers.

```
set.seed(39)
wheat_clustering <- wheat_modified[,1:7]

km_cluster2 <- kmeans(wheat_clustering, centers = 2, nstart = 20)
km_cluster2

km_cluster3 <- kmeans(wheat_clustering, centers = 3, nstart = 20)
km_cluster3

km_cluster4 <- kmeans(wheat_clustering, centers = 4, nstart = 20)
km_cluster4
```

### Cluster Evaluation

find the best number of clusters by using their total.withinss value

```
clusters1 <- c(1:10)
withiness1 <- c()

for(i in clusters1){
  tot_with1 <- kmeans(wheat_clustering, centers = i, nstart = 20)$tot.withinss
  withiness1 <- c(withiness1, tot_with1)
}

plot(withiness1~clusters1, type="b", col="green",
     xlab = "Number of clusters", ylab = "Total withinss",
     main = "Total withinss vs Clusters")
```

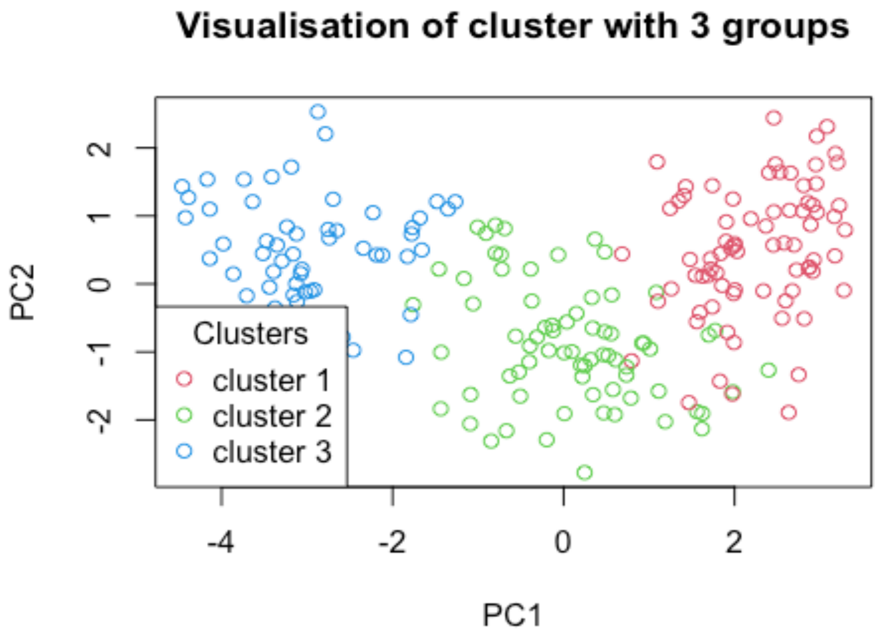


From this, we can see that clusters with 3 size is had a significant drop in total withinss and after that we can see the drop in the value but they are not that significant.

Therefore, selecting the best size of clusters as 3.

### visualise the clusters.

```
pca_wheat <- prcomp(wheat_modified[,1:7], scale. = TRUE)
plot(pca_wheat$x[,1:2], col = km_cluster3$cluster + 1,
    main = "Visualisation of cluster with 3 groups")
legend("bottomleft",
      legend=c("cluster 1","cluster 2", "cluster 3"), title="Clusters",
      col = c(2, 3, 4), pch = 1)
```


Visualisation of cluster with 3 groups
PC2
PC1
Clusters
○ cluster 1
○ cluster 2
○ cluster 3