# ANALYSIS ON MILK SAMPLES

Presented by Anisha Mittal

# CONTENTS

OUR MAIN
TOPICS TODAY

# AIM

Data Visualisation and exploration for the protein and technological traits of milk samples.

Clustering the cow breeds based on the MIR spectra of their milk samples.

Classifying milk samples having heat stability less than 10mins based on their MIR spectra.

# DATA SET

## ABOUT THE DATA

**Initial Data set**

582 variables and 431 observations.

**After removing NA columns**

574 variables and 430 observations.

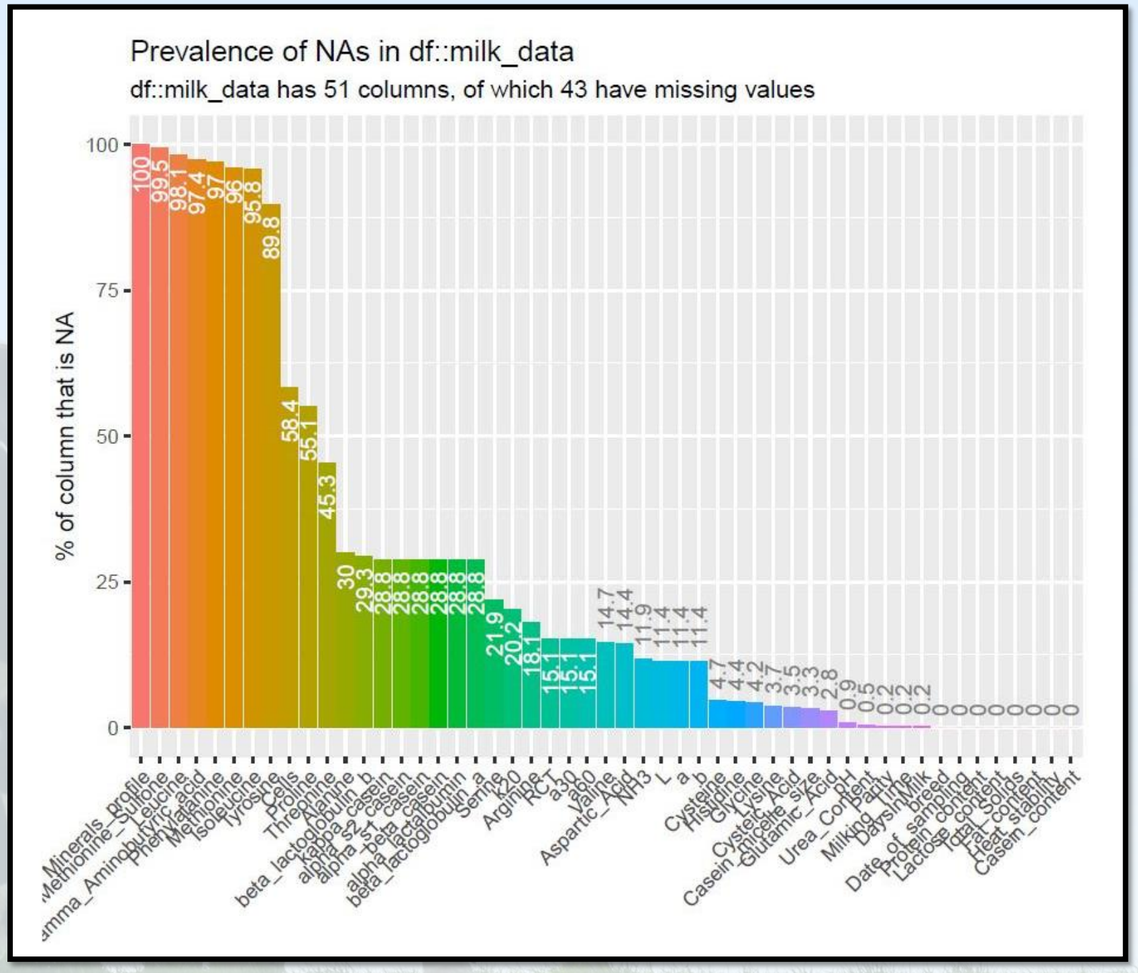
**After removing outliers**

574 variables and 326 observations.

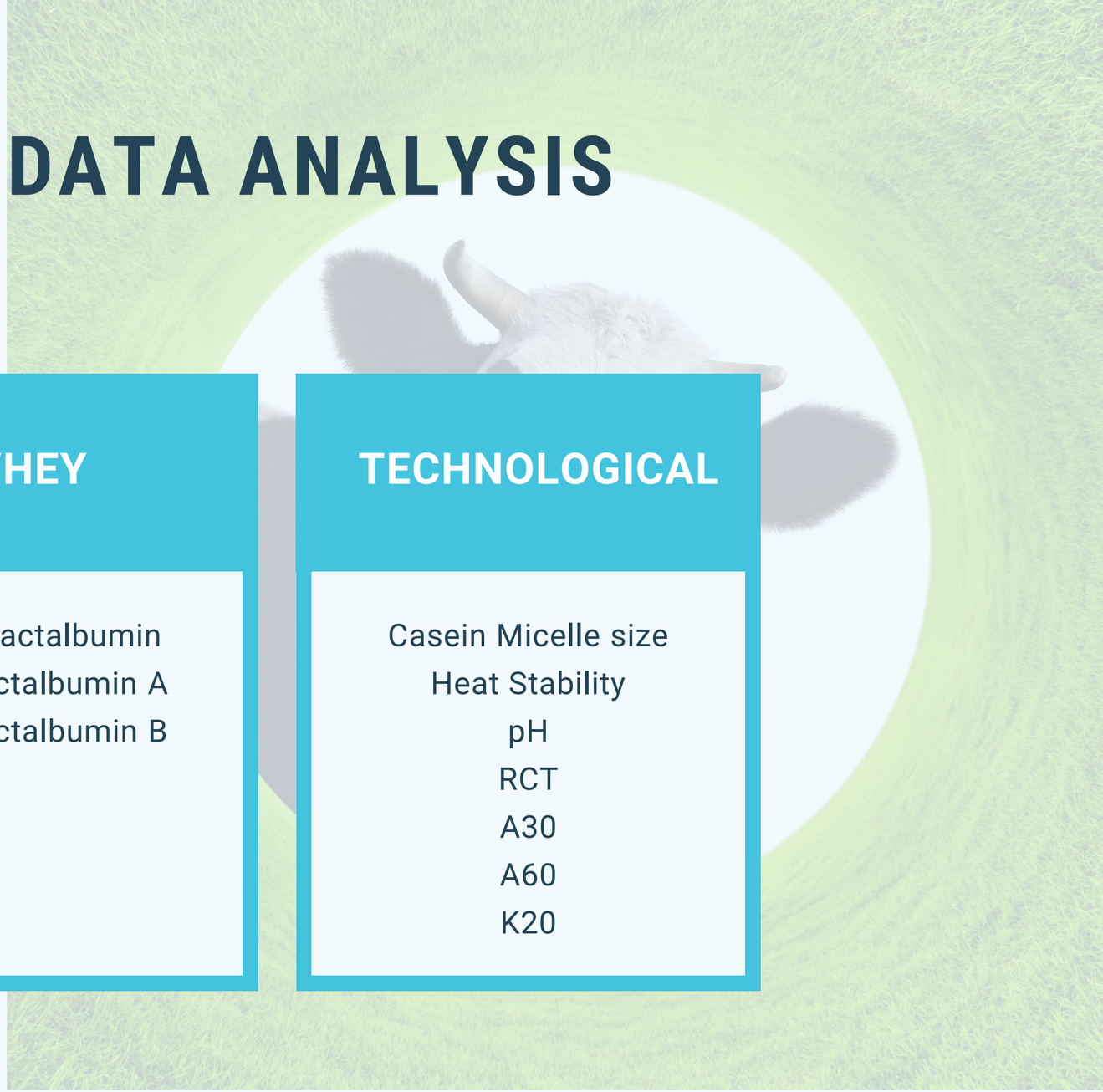
**Final Data set**

1:43 trait variables

44:531 MIR spectra variables



# EXPLORATORY DATA ANALYSIS



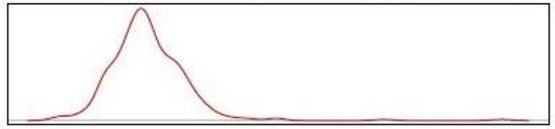
| CASEIN | WHEY | TECHNOLOGICAL |
| --- | --- | --- |
| Kappa Casein | Alpha Lactalbumin | Casein Micelle size |
| Beta Casein | Beta Lactalbumin A | Heat Stability |
| Alpha s1 Casein | Beta Lactalbumin B | pH |
| Alpha s2 Casein | | RCT |
| | | A30 |
| | | A60 |
| | | K20 |

# CASEIN TRAITS

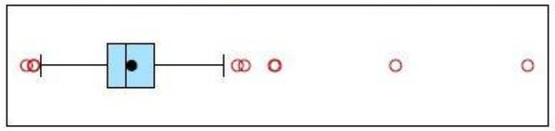
## EXPLORATORY DATA ANALYSIS

Histogram of milk_data$kappa_casein

Density of milk_data$kappa_casein



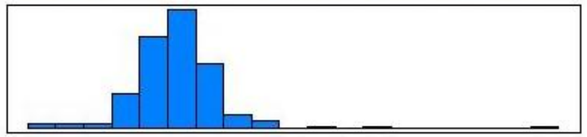
Boxplot of milk_data$kappa_casein



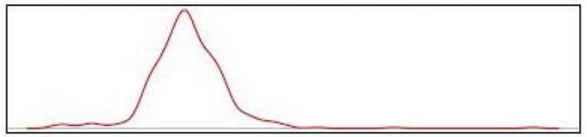
Q−Q Plot of milk_data$kappa_casein

## EXPLORATORY DATA ANALYSIS

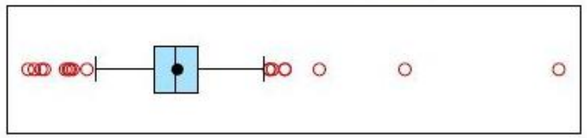
Histogram of milk_data$beta_casein



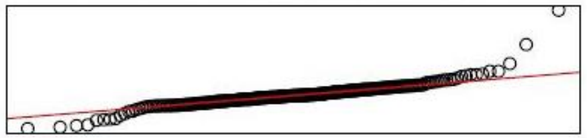
Density of milk_data$beta_casein



Boxplot of milk_data$beta_casein

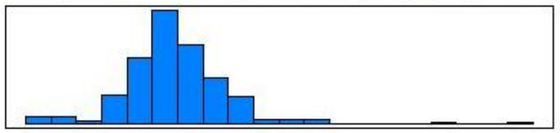


Q−Q Plot of milk_data$beta_casein

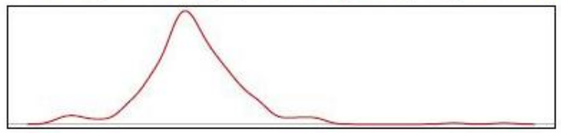


## EXPLORATORY DATA ANALYSIS

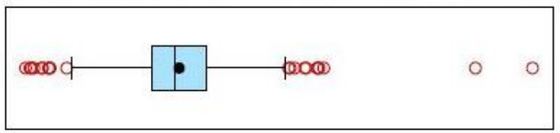
Histogram of milk_data$alpha_s2_casein



Density of milk_data$alpha_s2_casein



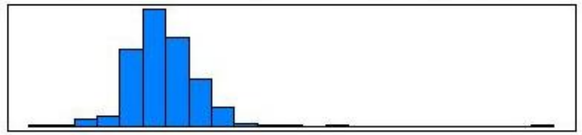
Boxplot of milk_data$alpha_s2_casein



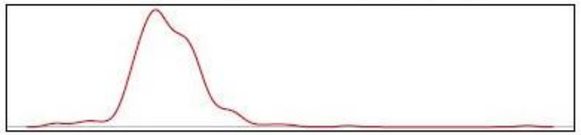
Q−Q Plot of milk_data$alpha_s2_casein

## EXPLORATORY DATA ANALYSIS

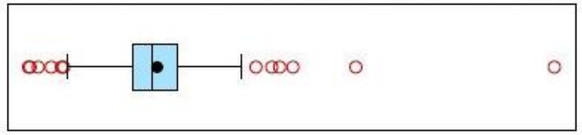
Histogram of milk_data$alpha_s1_casein



Density of milk_data$alpha_s1_casein



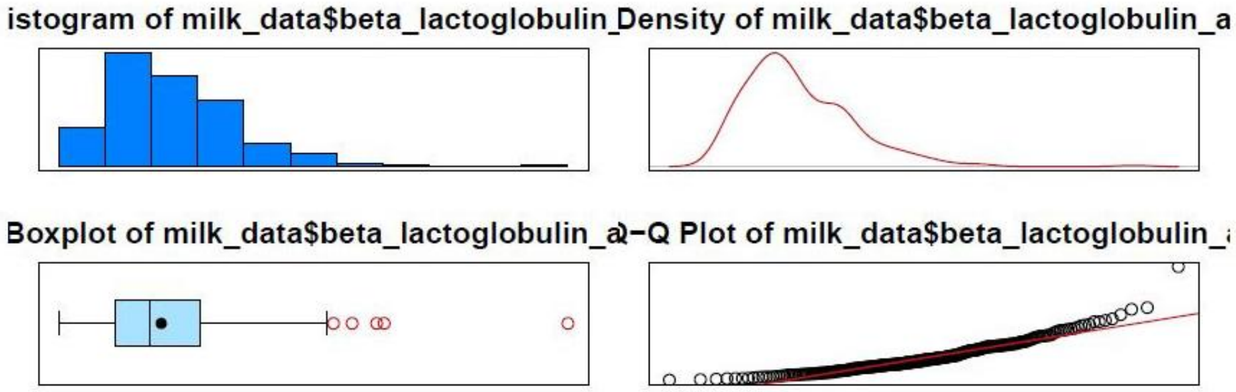
Boxplot of milk_data$alpha_s1_casein



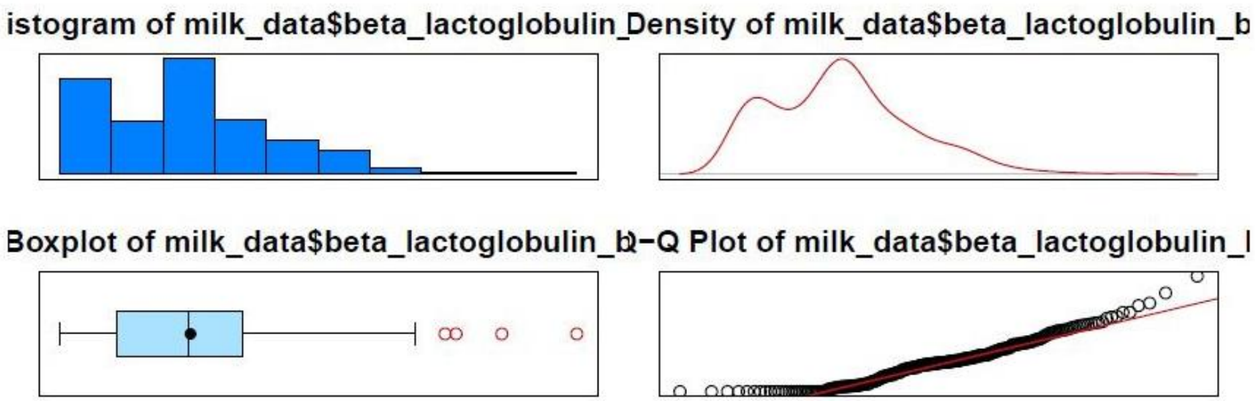
Q−Q Plot of milk_data$alpha_s1_casein

# WHEY TRAITS

EXPLORATORY DATA ANALYSIS

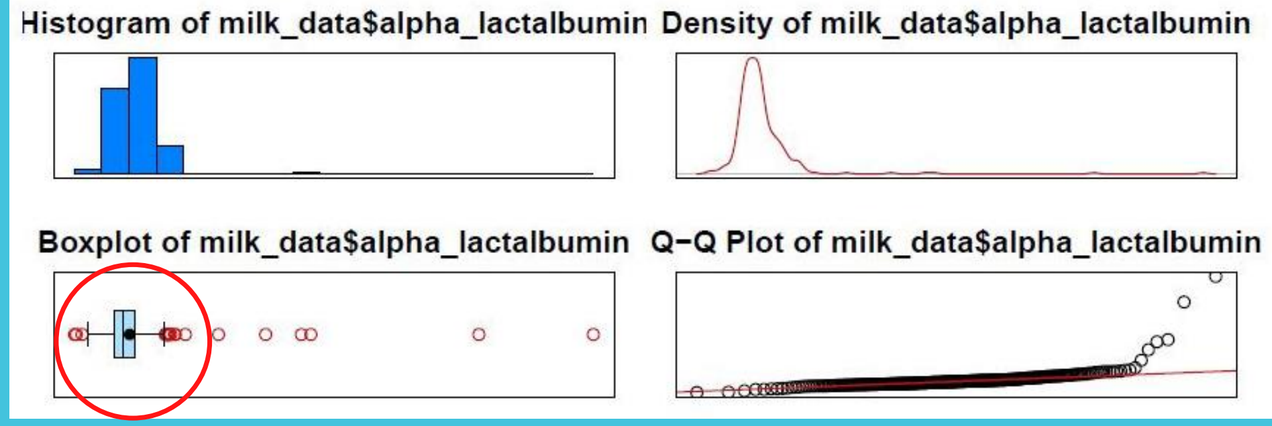



EXPLORATORY DATA ANALYSIS



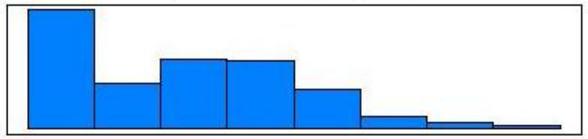

**Variable Alpha Lactalbumin shows some variability and is right-skewed but mean and median are equal.**

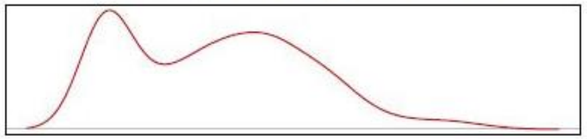
EXPLORATORY DATA ANALYSIS

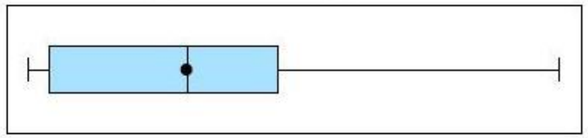


# TECHNOLOGICAL TRAITS

## EXPLORATORY DATA ANALYSIS

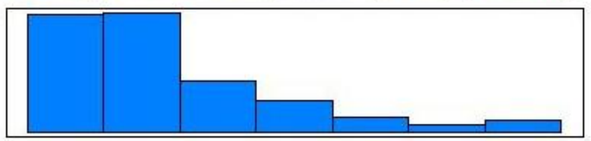


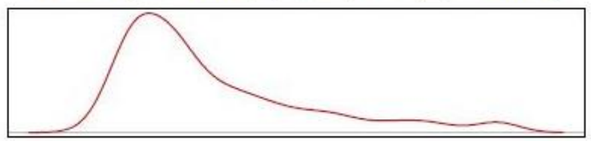



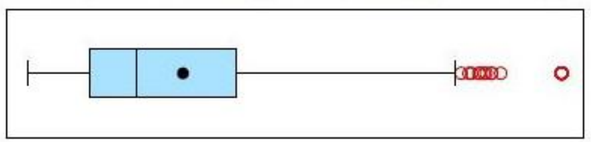





## EXPLORATORY DATA ANALYSIS

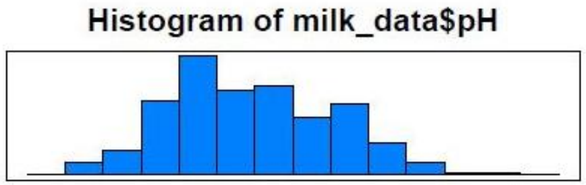


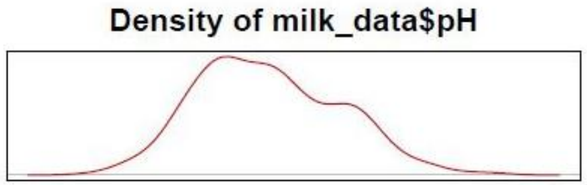



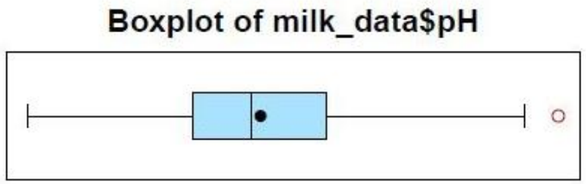



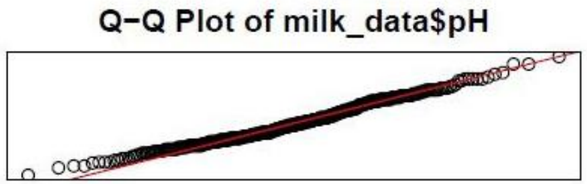


## EXPLORATORY DATA ANALYSIS

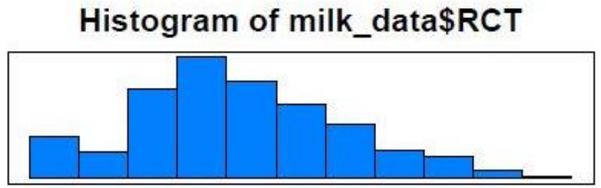


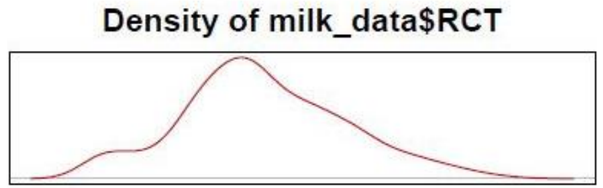



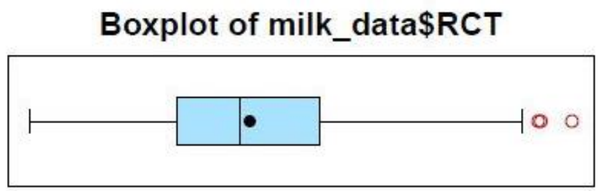



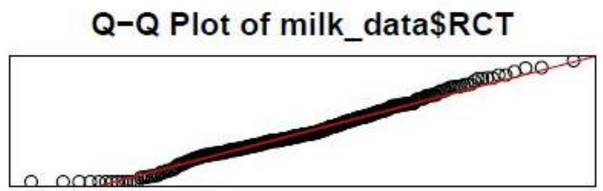




## EXPLORATORY DATA ANALYSIS



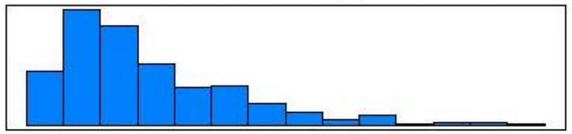



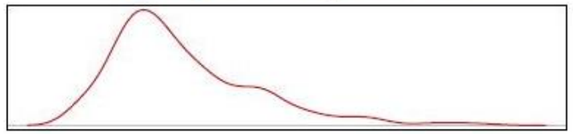



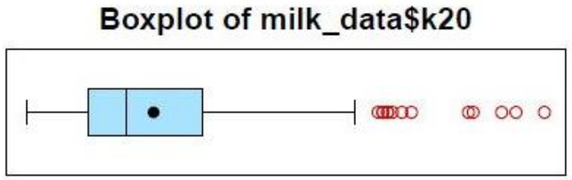



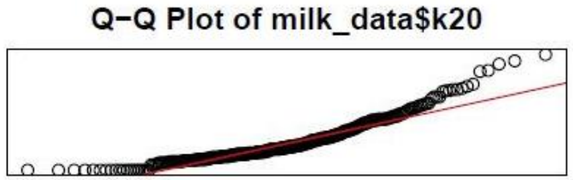
# TECHNOLOGICAL TRAITS

EXPLORATORY DATA ANALYSIS

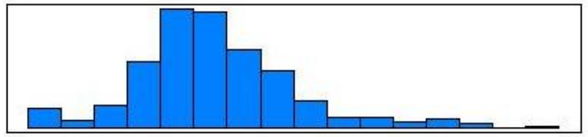


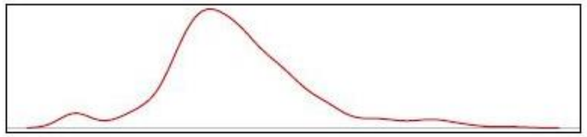



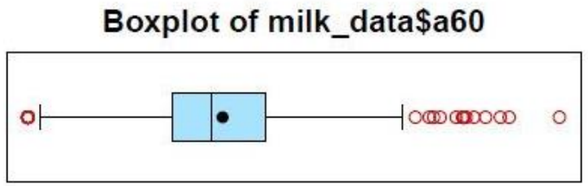



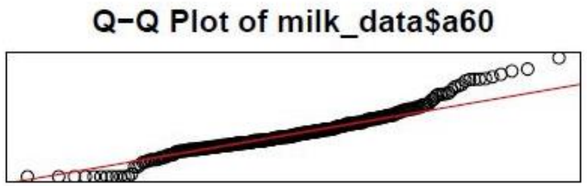




EXPLORATORY DATA ANALYSIS



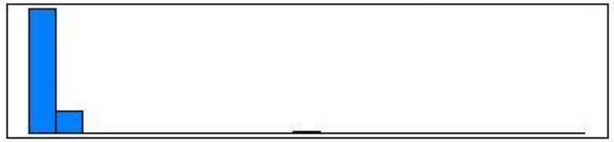



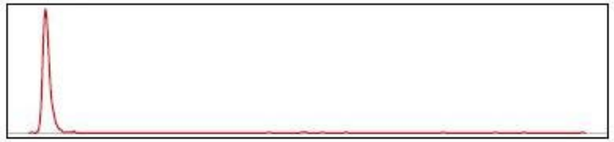



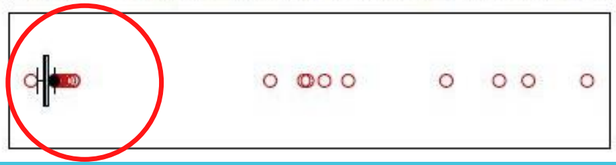




**Variable Casein Micelle size shows a lot of variability and is right-skewed.**

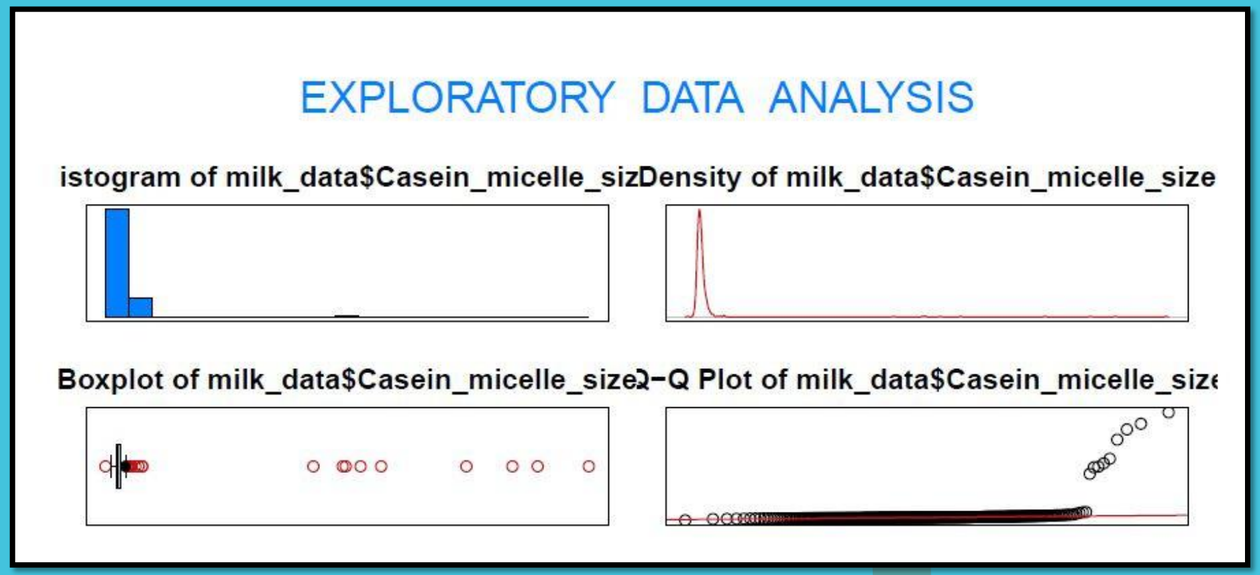
EXPLORATORY DATA ANALYSIS



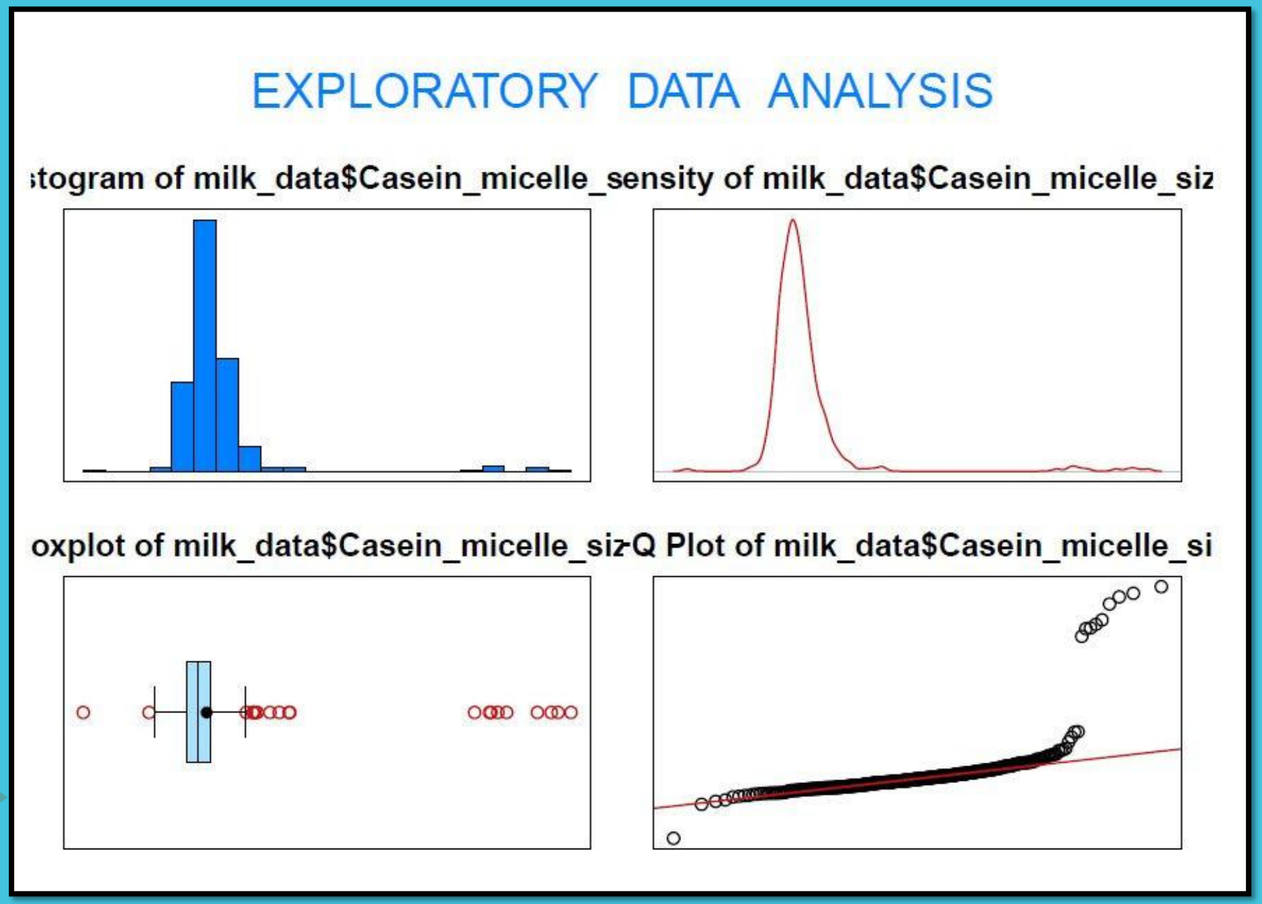




# Casein Micelle Size

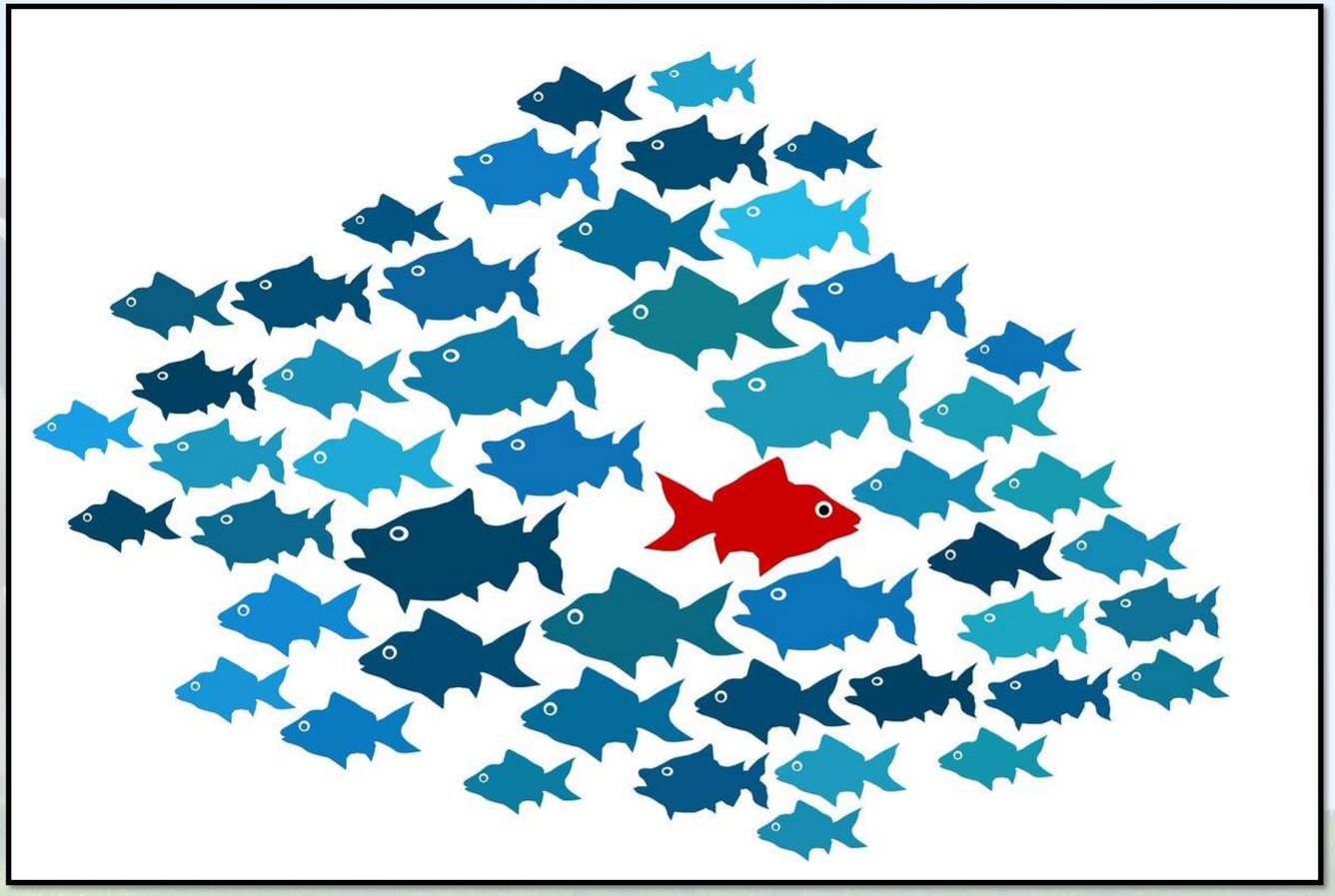



**EDA of variable Casein Micelle size after log Transformation.**

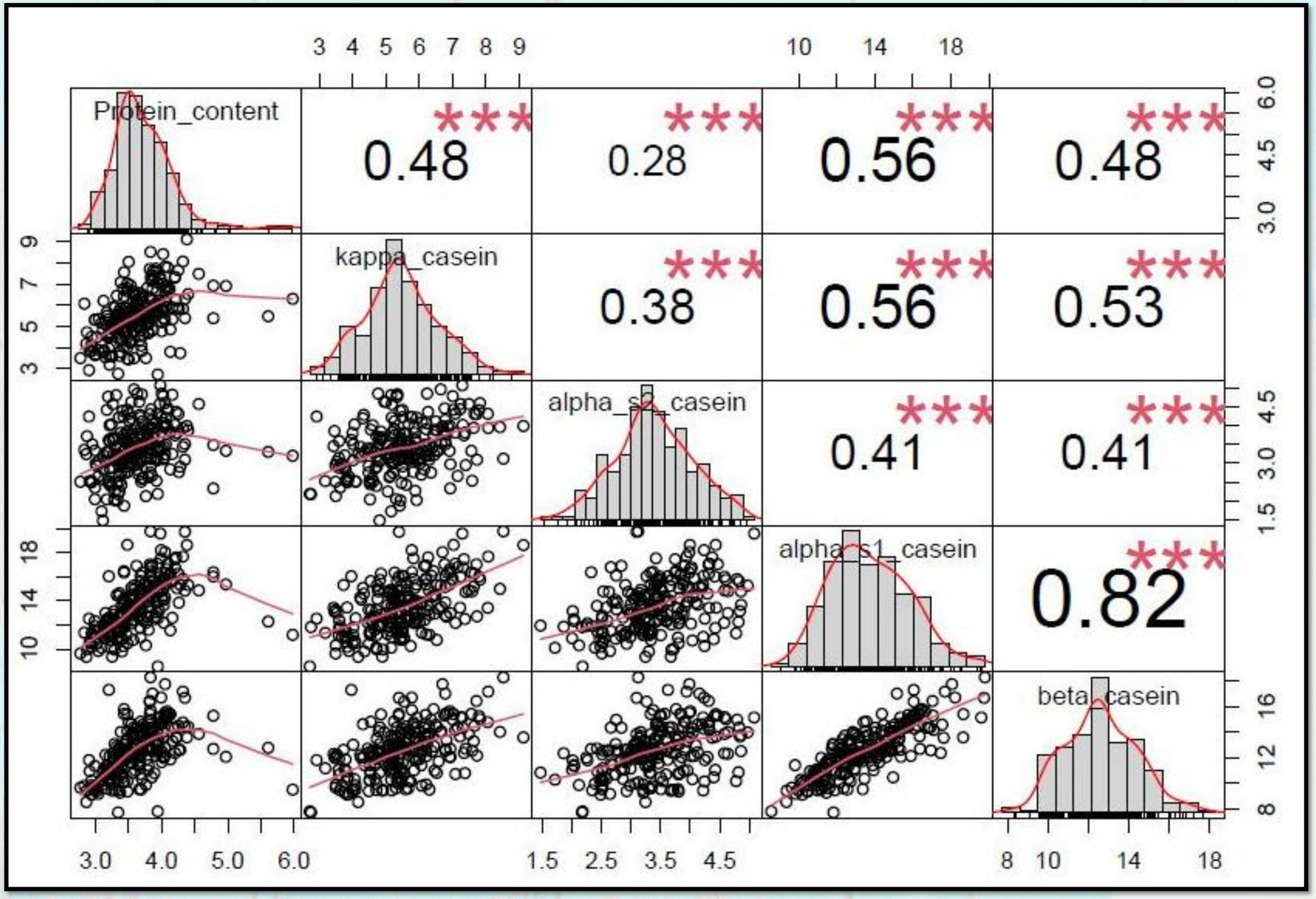


# OUTLIER REMOVAL

Outliers are unusual values in a dataset, and they can distort statistical analyses and violate their assumptions.



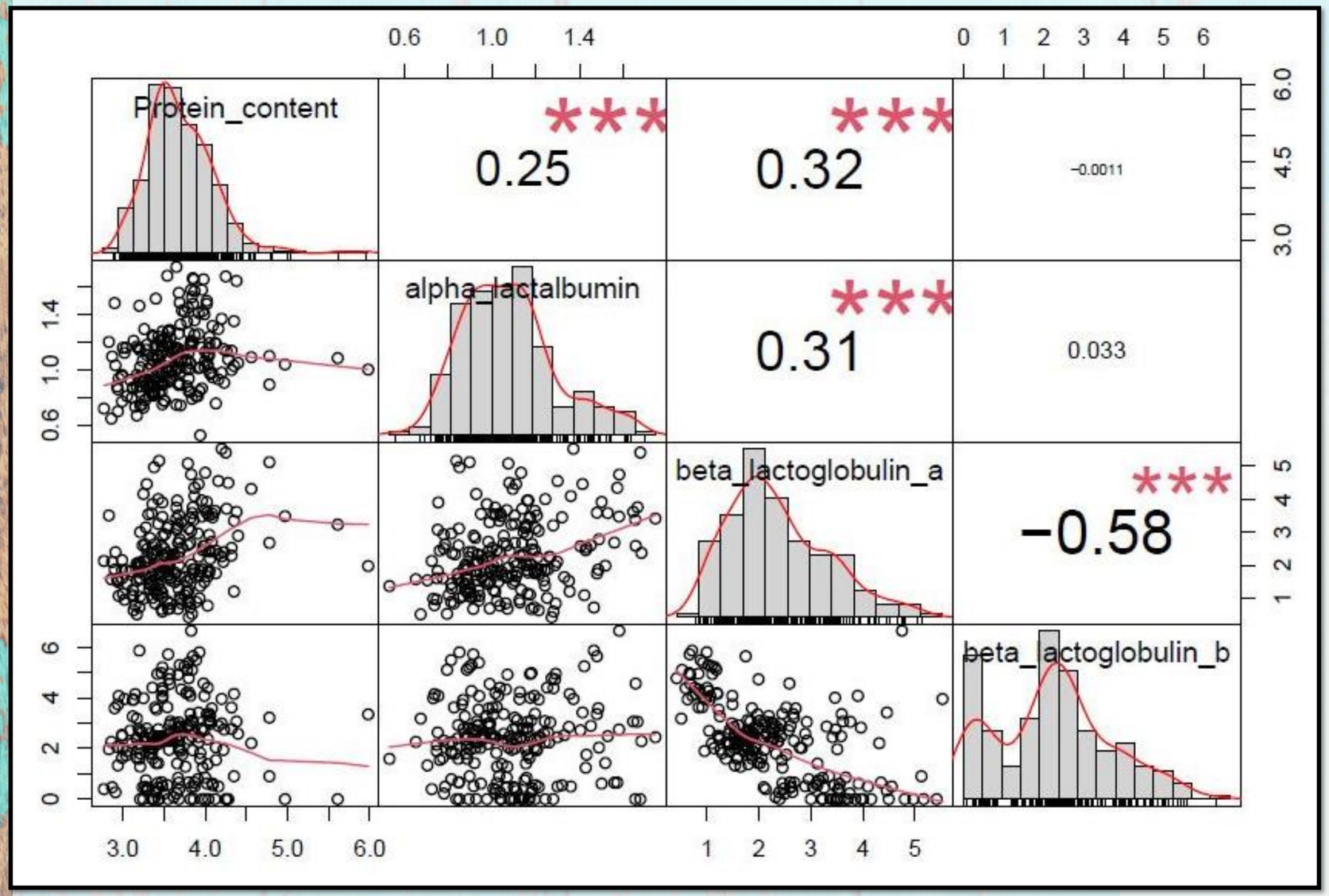
# CORRELATION CHART




## CASEIN PROTEIN

We can see that all the Casein protein traits are highly positively correlated to each other.

# CORRELATION CHART

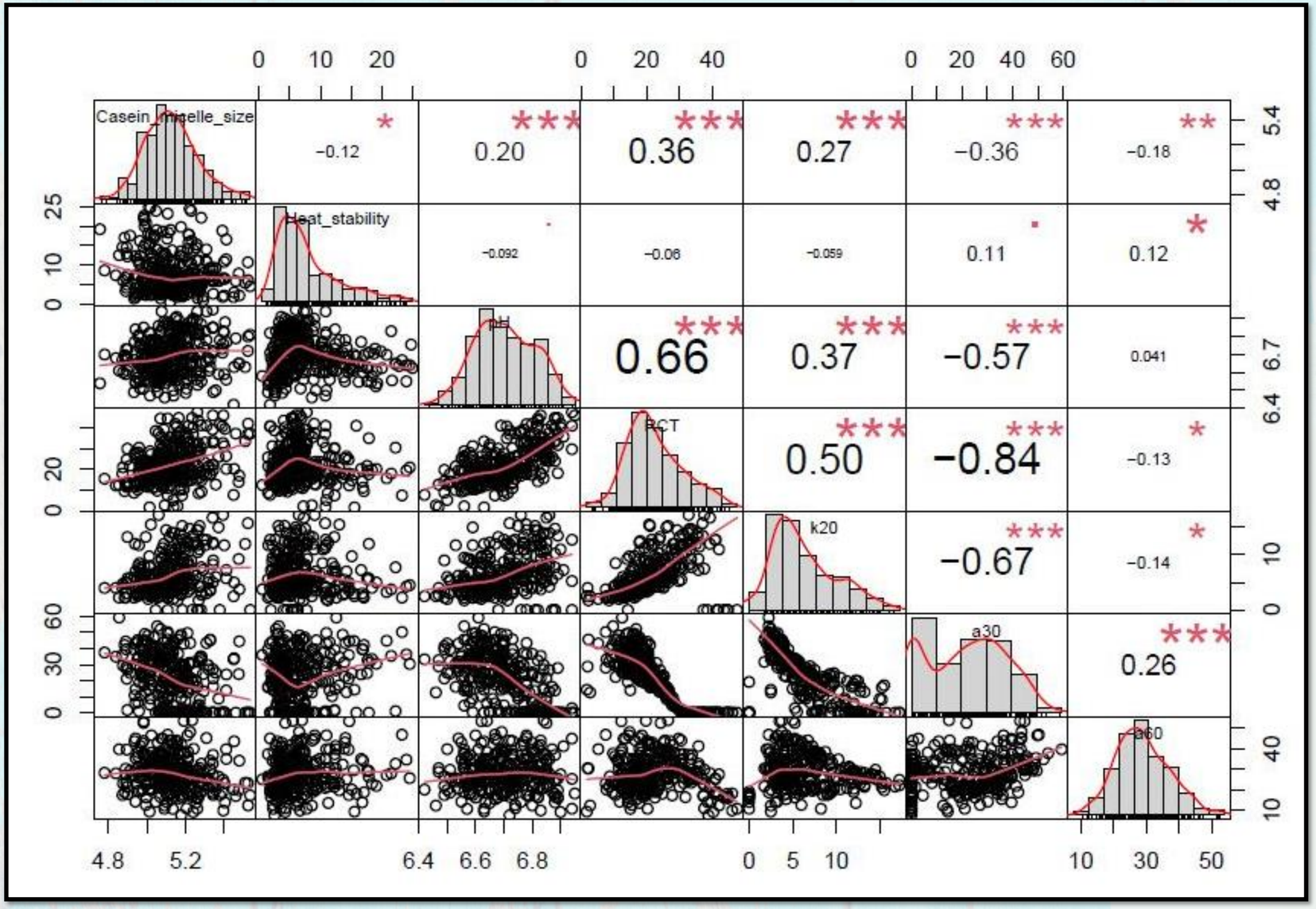



## WHEY PROTEIN

Beta-lactalbumin B and Alpha-lactalbumin traits are not significantly correlated.

Beta-lactalbumin B is negatively correlated to protein content in a milk sample.

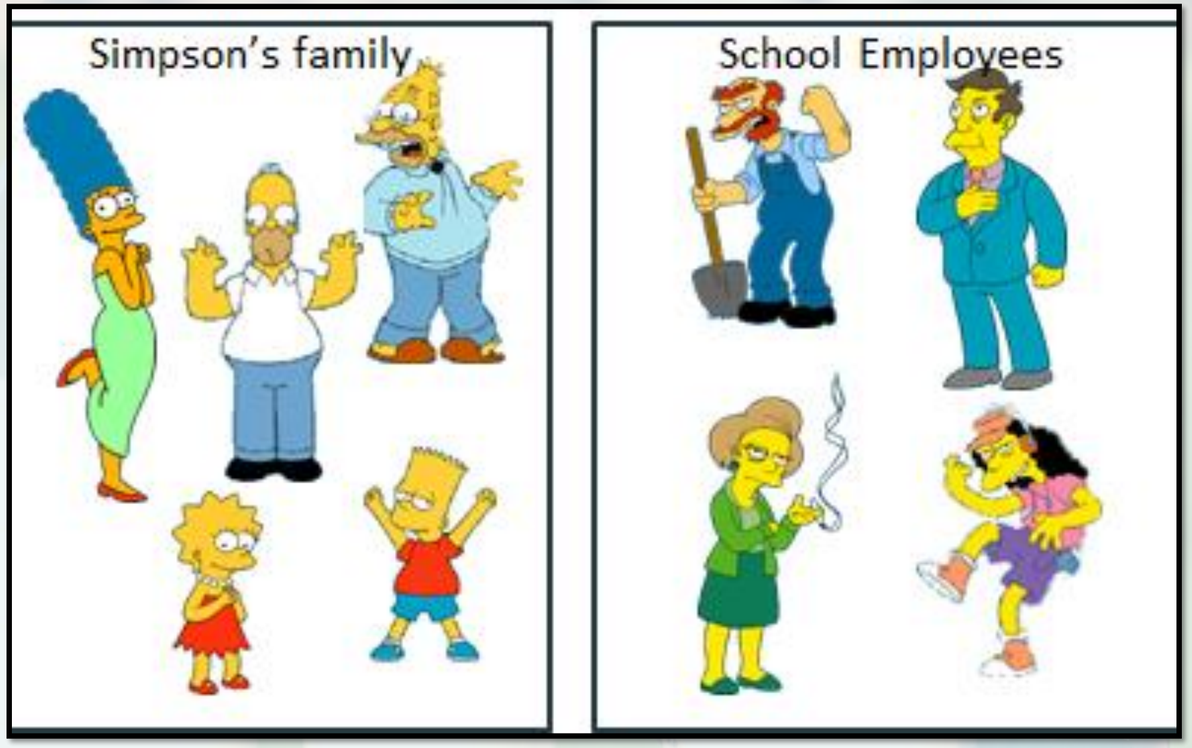
# CORRELATION CHART



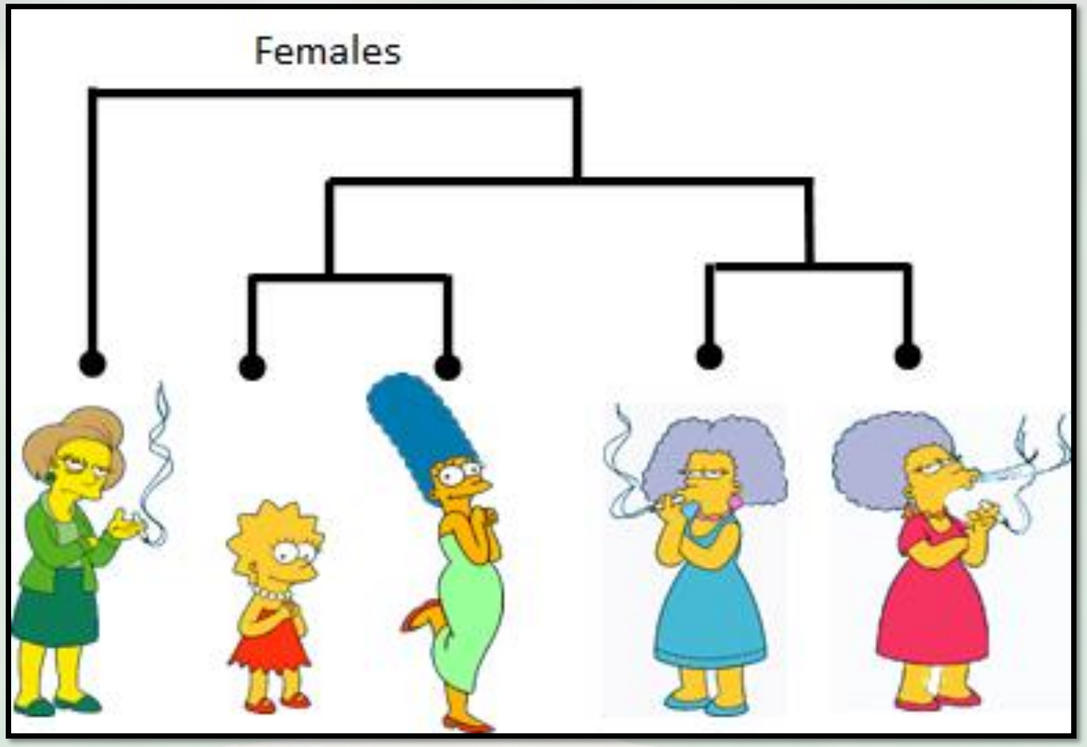

## TECHNOLOGICAL

We can see that there are very few technological traits that are correlated to each other.

# CLUSTERING




PARTITIONING

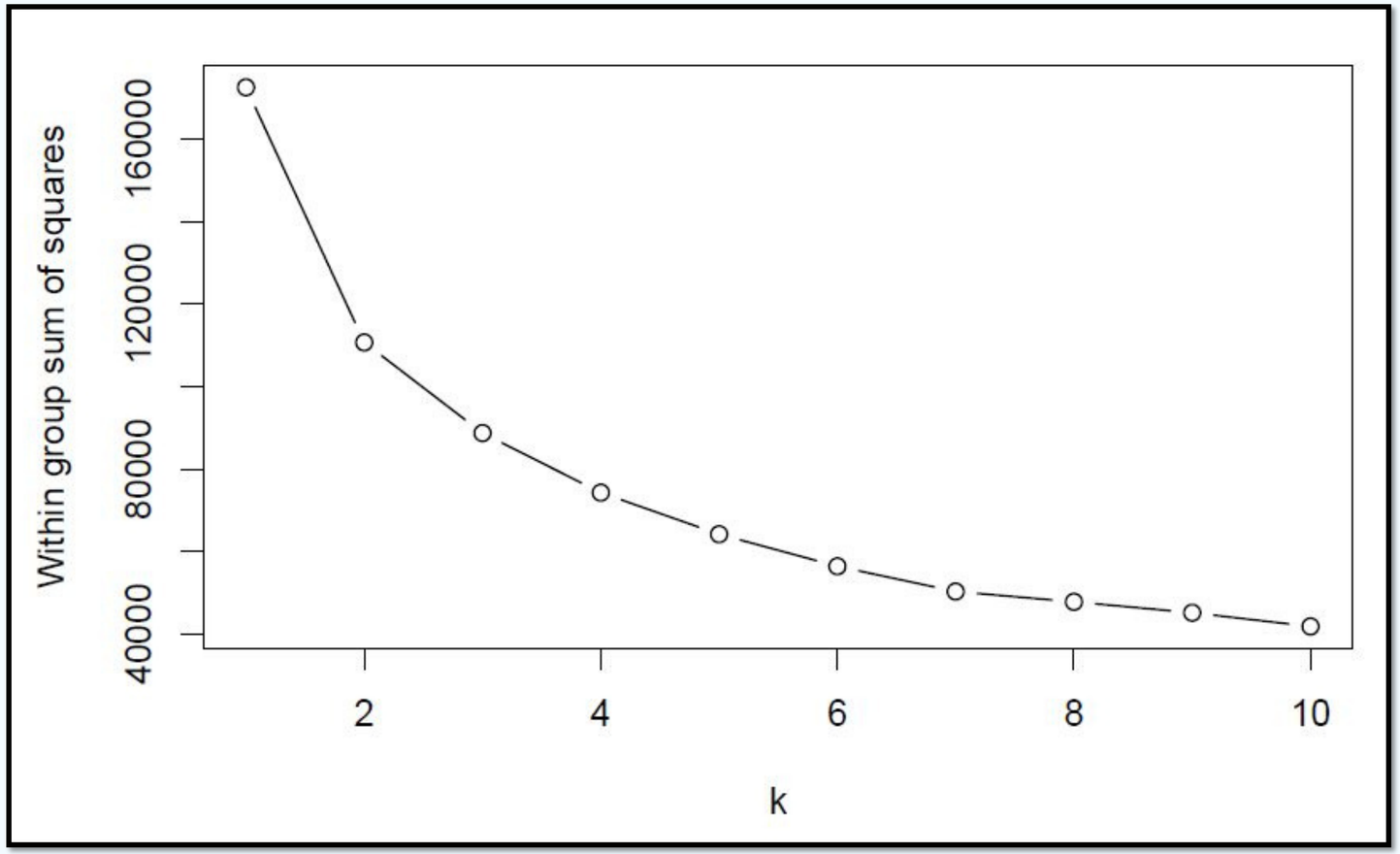



HIERARCHAL

# WHAT K TO SELECT?

## WITHIN SUM OF SQUARE

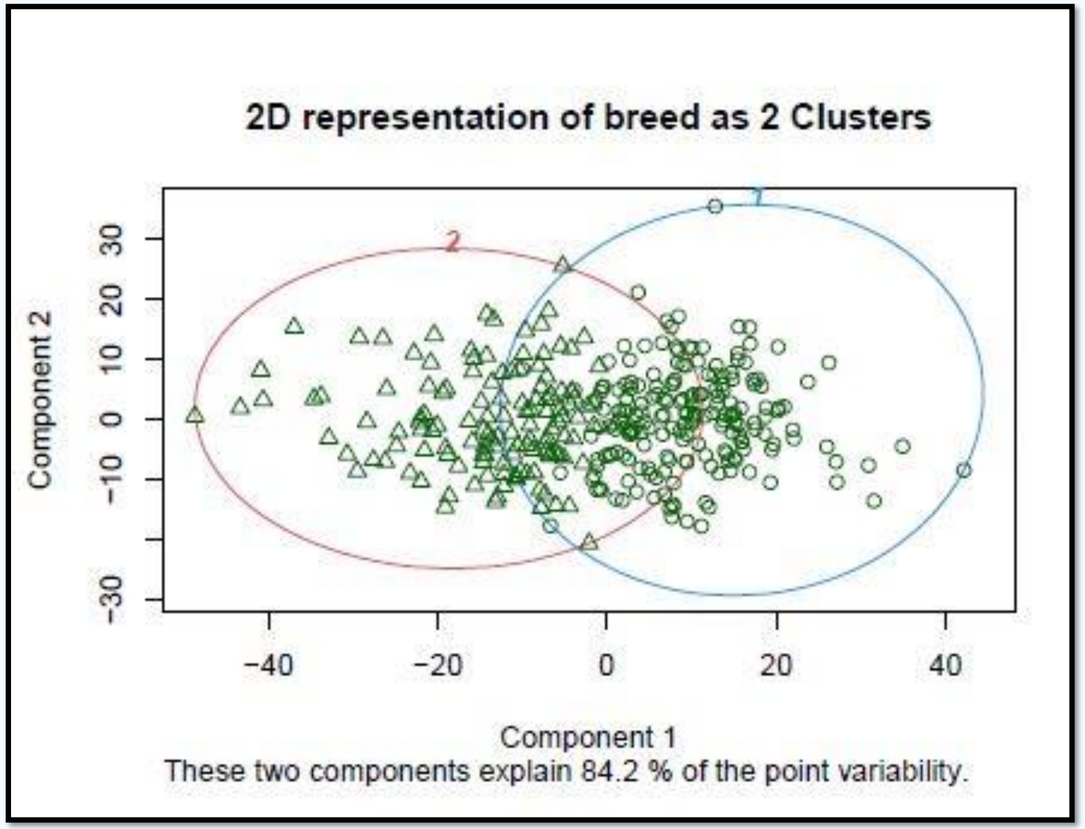
Plot a range of k values against within sum of squares and select the K with the least sum of squares value.

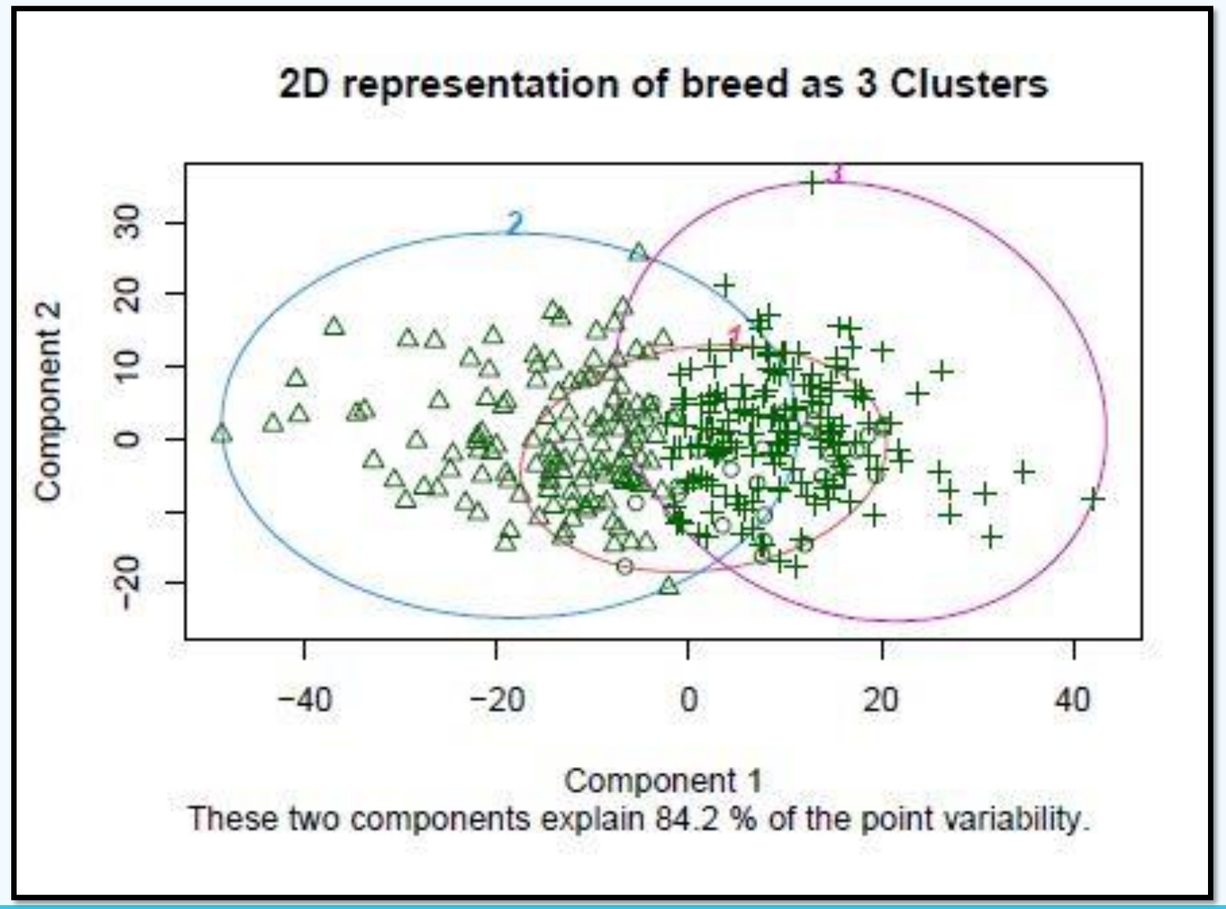


# K=2

Cluster 1 = 195 observations

Custer 2 = 131 observations




# K=3

Cluster 1 = 36 observations

Custer 2 = 126 observations

Cluster 3 = 164 observations

# VALIDATION

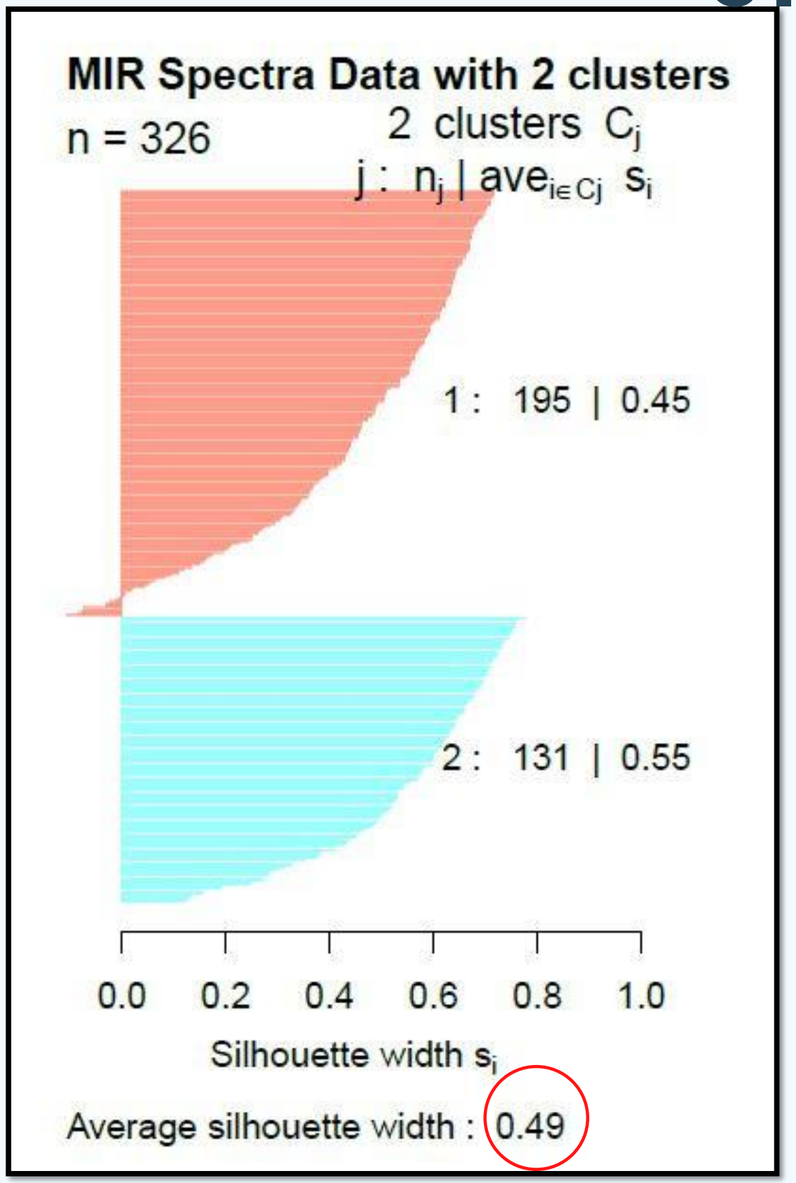
## INTERNAL

Useful to check whether the appropriate K is used or not. eg: Average Silhouette width, and Calinski-Harabasz Index

## EXTERNAL

Useful to assess the quality of the given clustering in comparison to a reference grouping of the observations.
eg: Rand Index, and Adjusted Rand Index

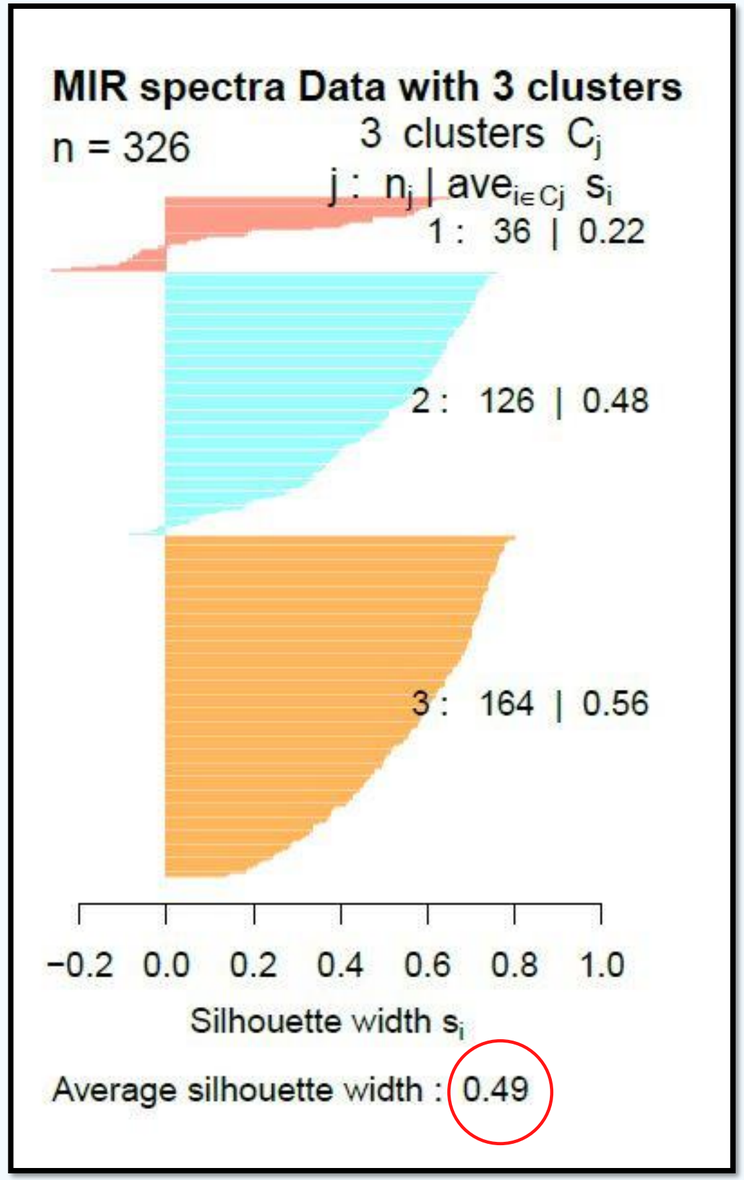
# SILHOUETTE PLOTS




## EUCLIDEAN

Formula : $\sqrt{\sum_{i=1}^{p}(x_{ij} - x_{ji})^2}$

Calculates the square root of sum of squared differences between two observations one by one.

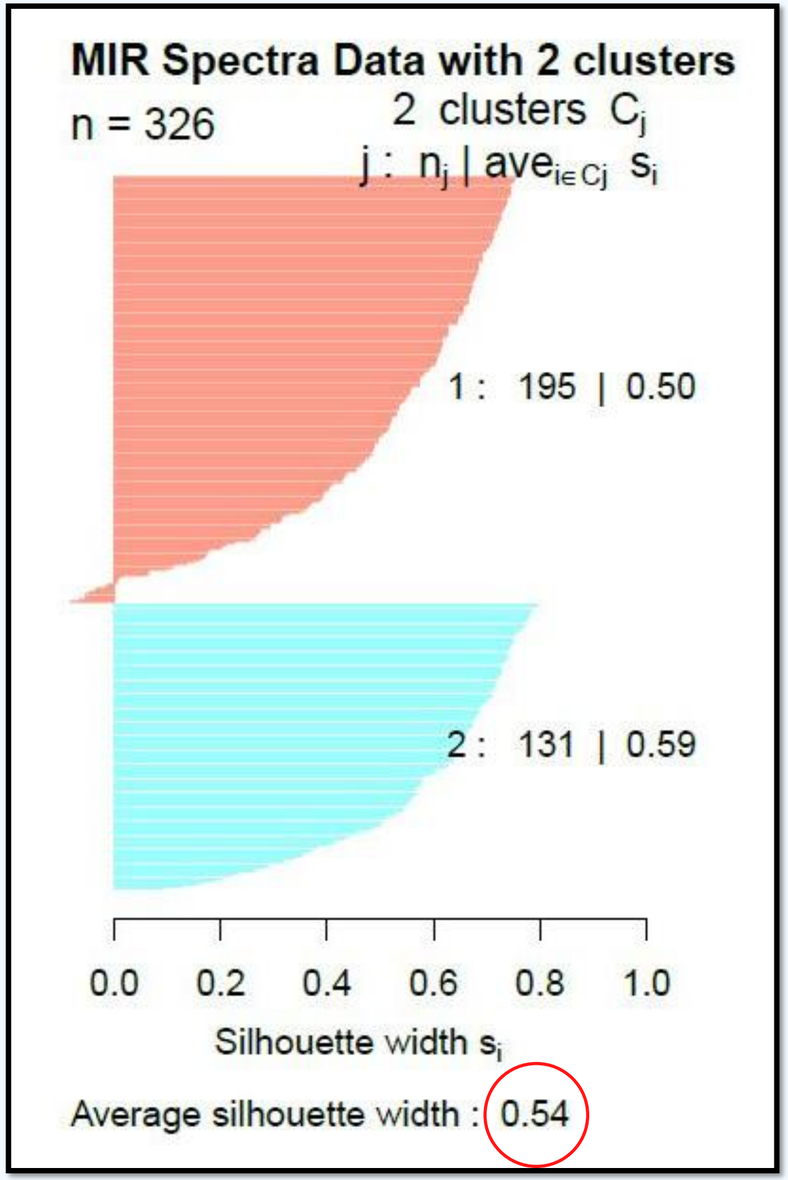


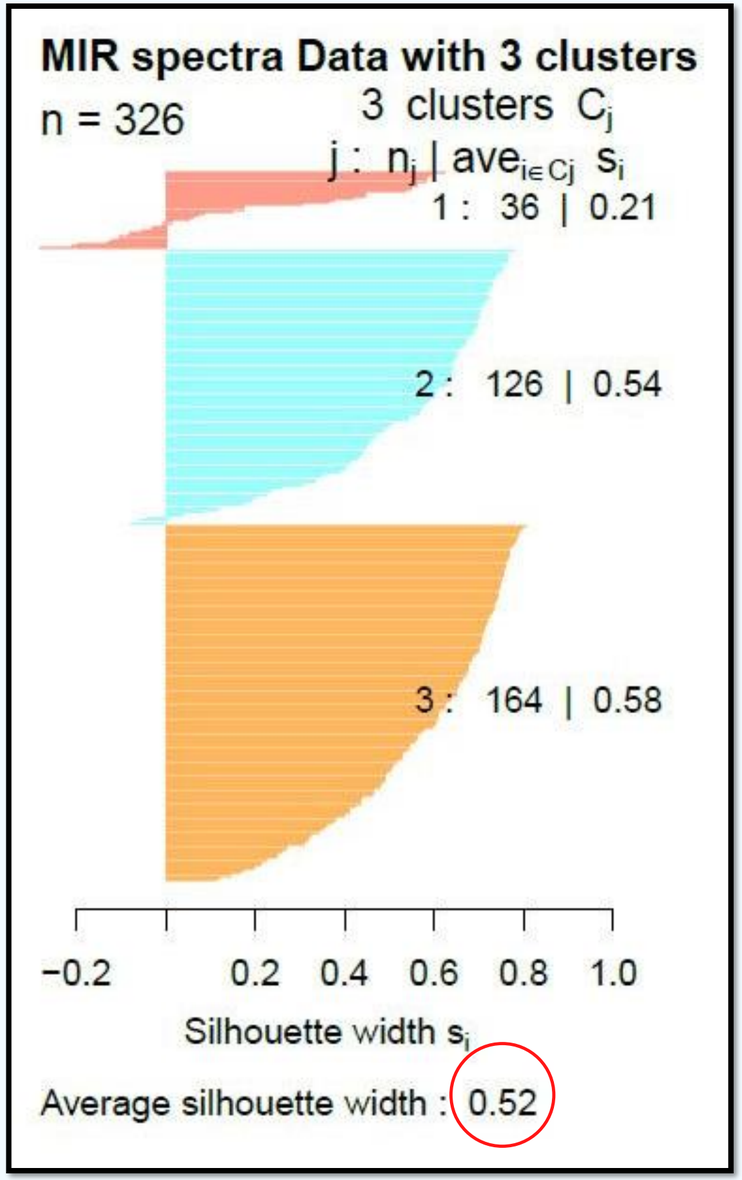
# SILHOUETTE PLOTS

## MANHATTAN

Formula: $\sum_{i=1}^{p} |x_{ij} - x_{ji}|$

Calculates the sum of absolute differences between two observations one by one.



# CLASS AGREEMENT

A higher percentage means a better fit.

| | Rand Index | Adjusted Rand Index |
|---|---|---|
| k = 2 | 52% | 5% |
| k = 3 | 51% | 1% |

# RESULT

No. of cows of different breeds in similar clusters.

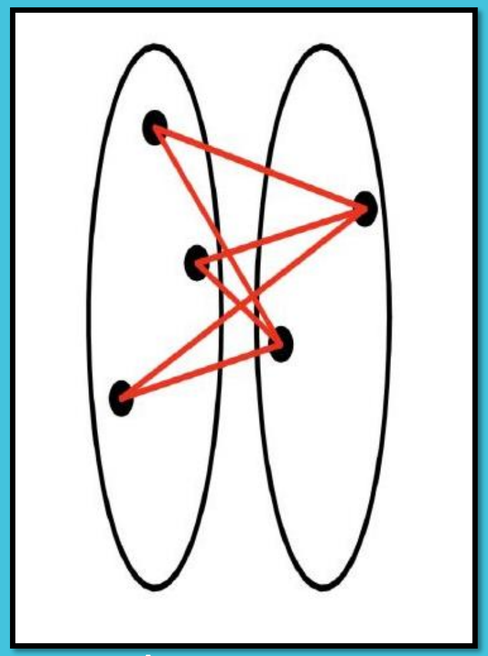
| | FRX- | FRX- | HOX | HOL FRI | HOX | HOX- | JE | JEX- | MO | NR |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 9 | 1 | 144 | 10 | 1 | 18 | 4 | 1 | 6 |
| 2 | 1 | 4 | 0 | 76 | 1 | 2 | 28 | 14 | 0 | 6 |

1 observation did not have a breed name and was put in cluster 1.

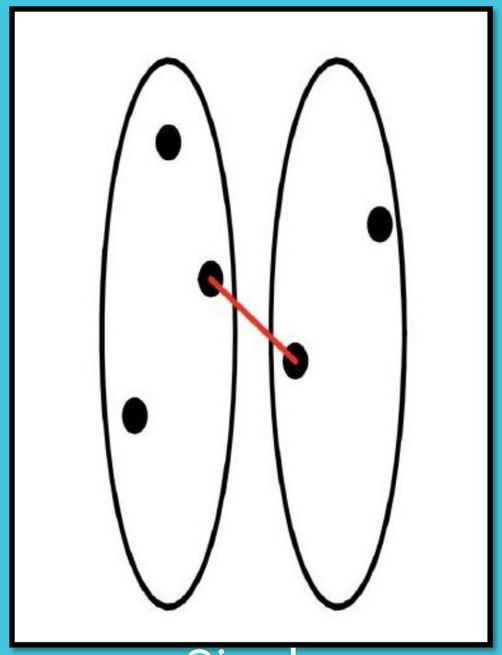
# HIERARCHICAL CLUSTERING

## LINKAGE

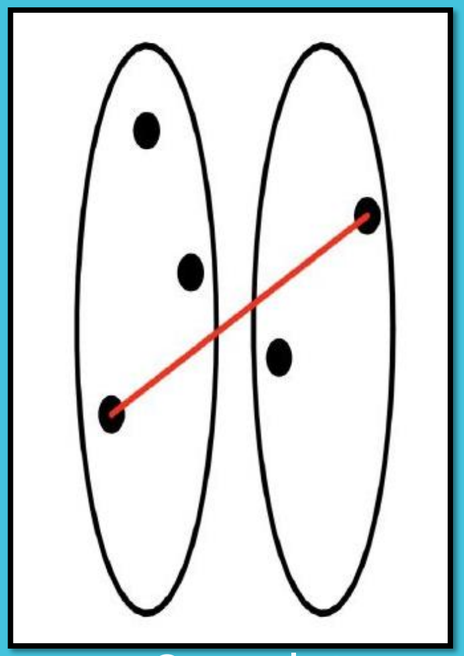
MEASURE OF DISSIMILARITY



Average

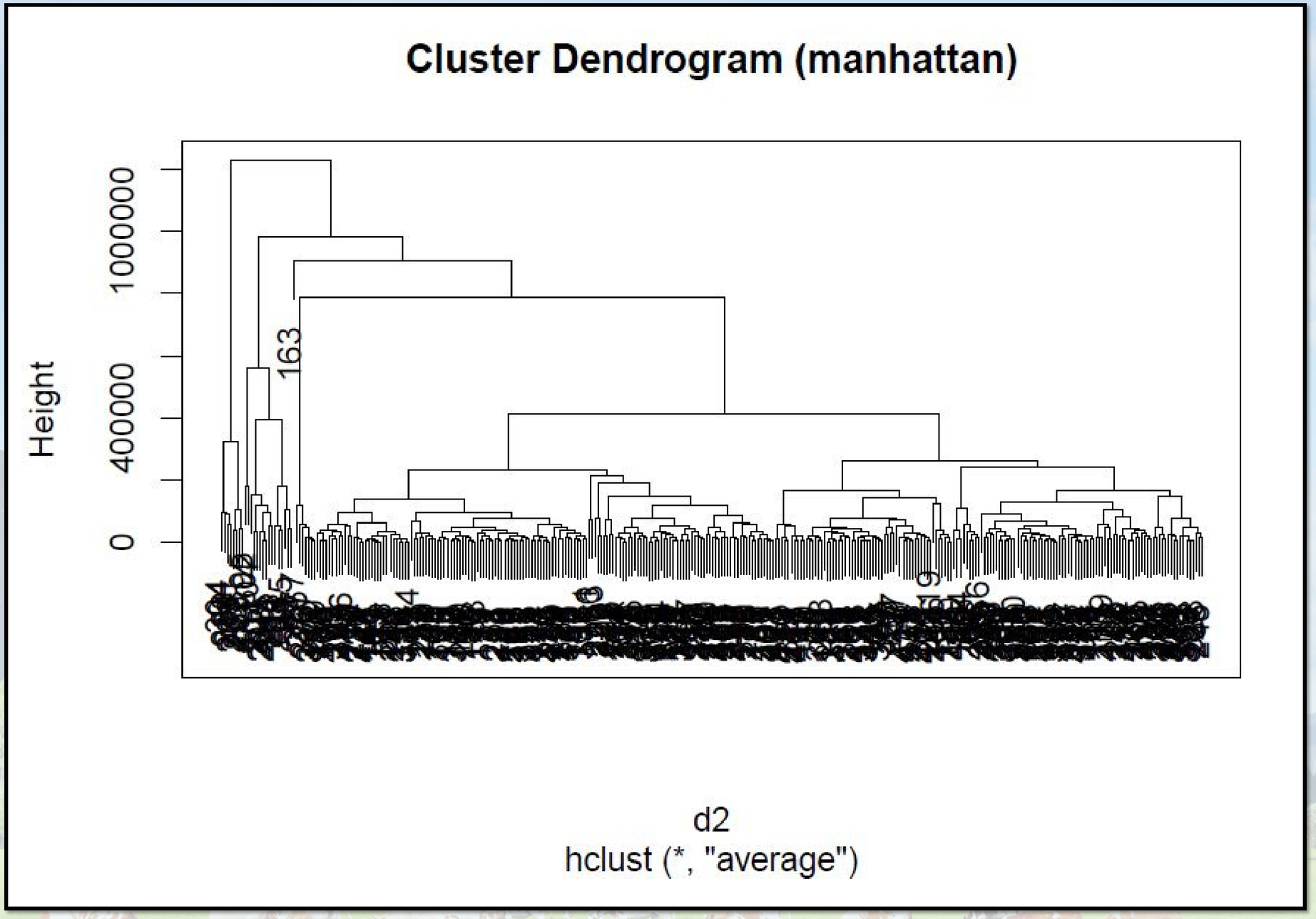


Single

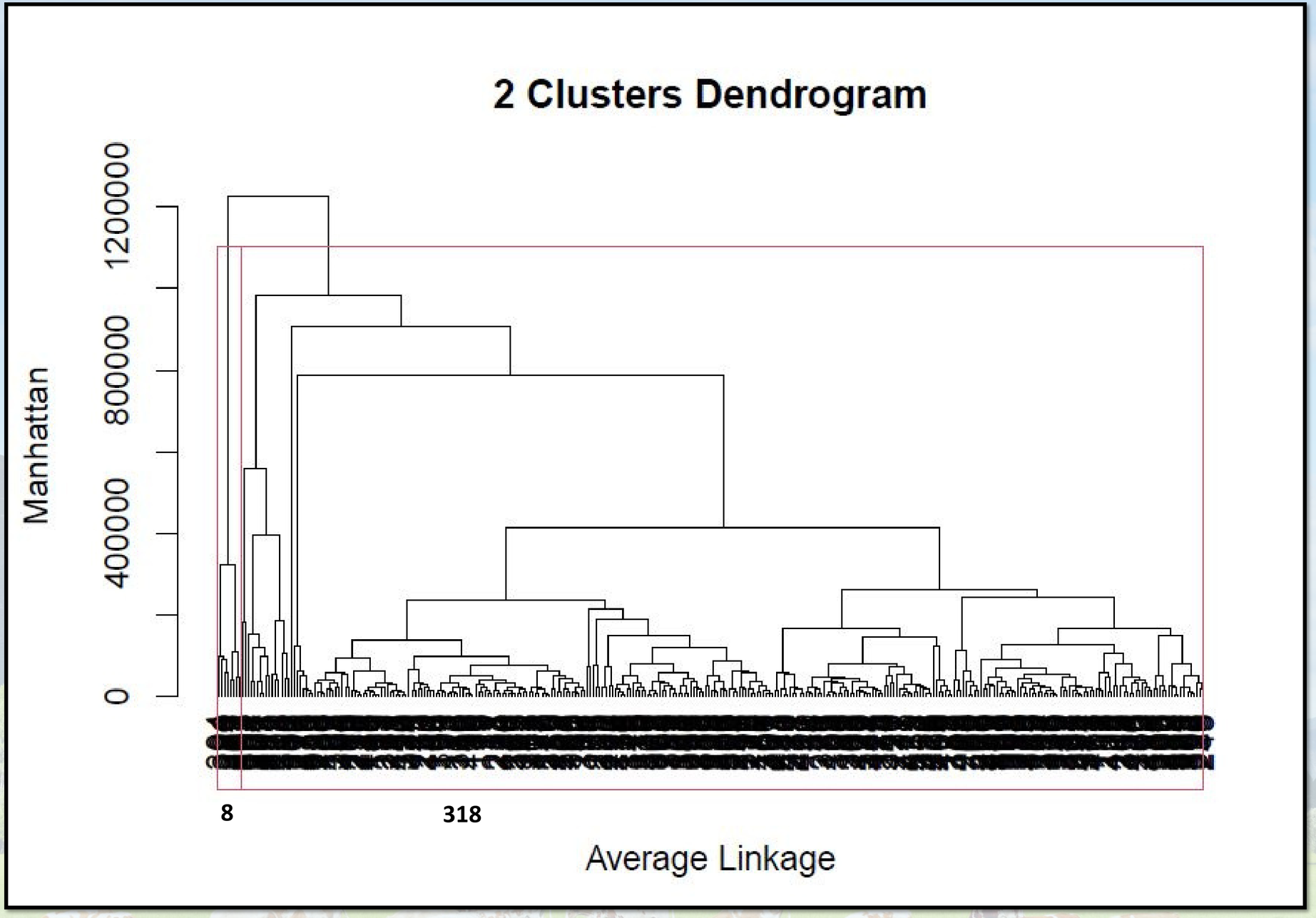


Complete

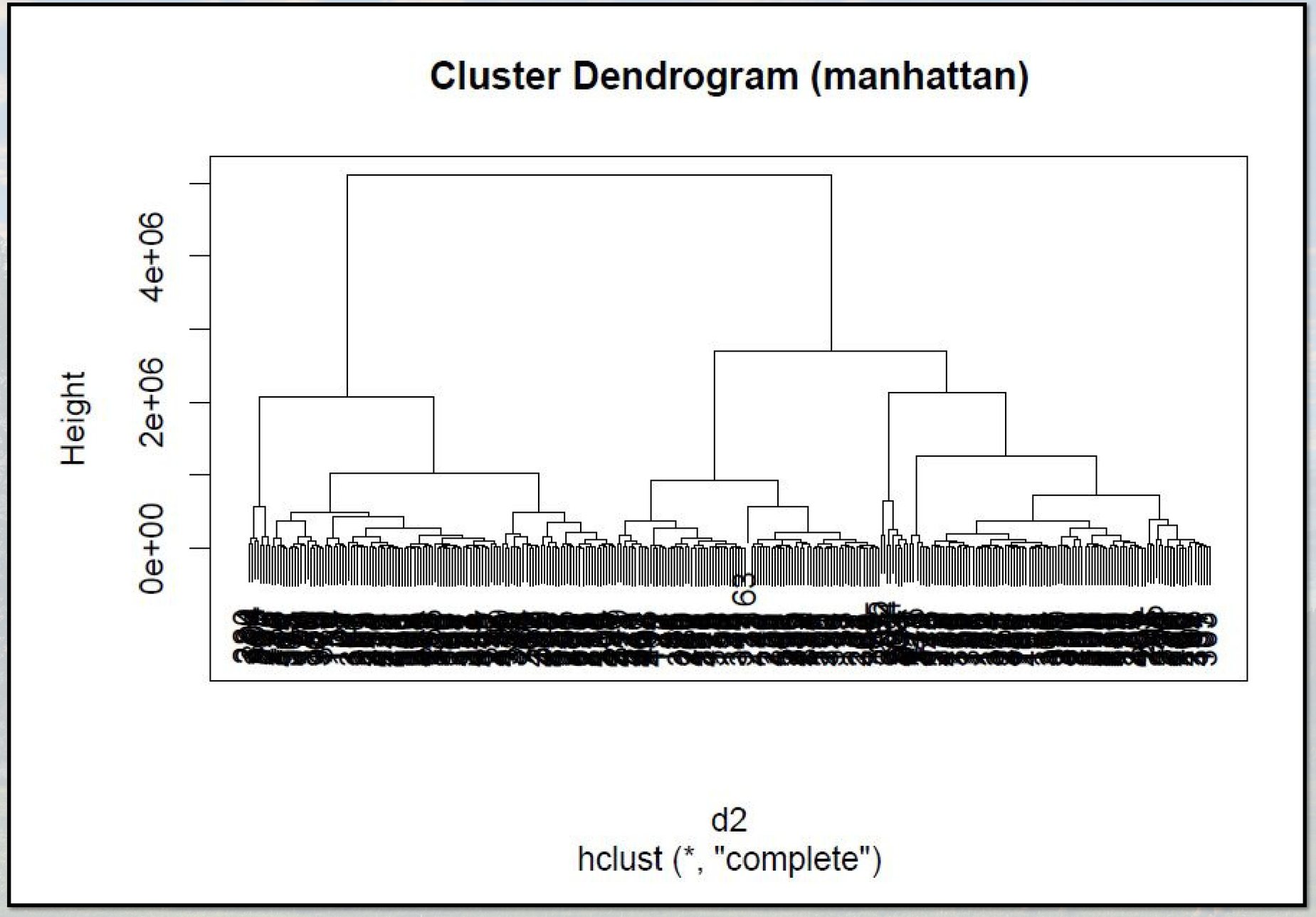
# AVERAGE LINKAGE



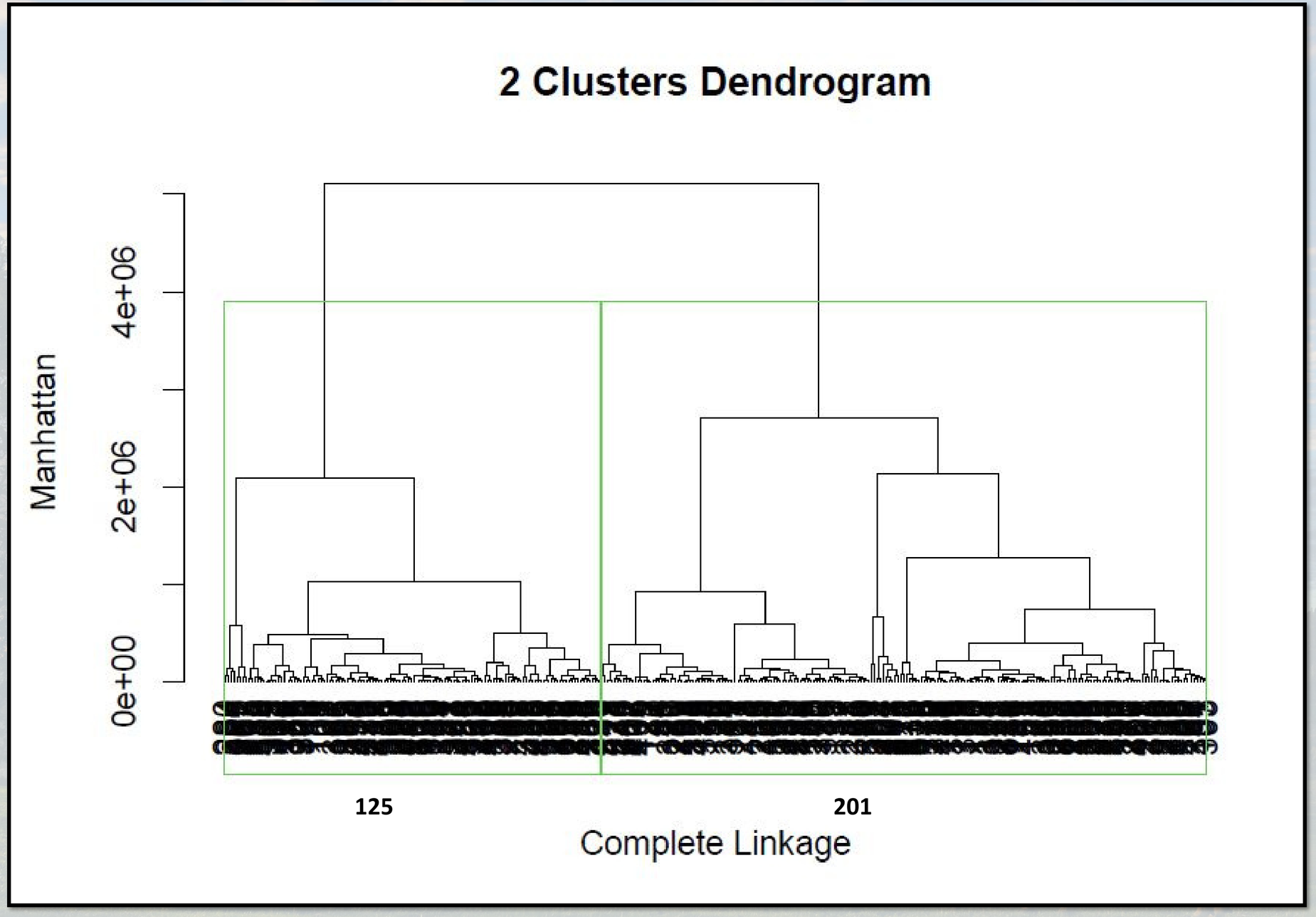
# AVERAGE LINKAGE



# COMPLETE LINKAGE



# COMPLETE LINKAGE



# CLASSIFICATION

## K-NEAREST NEIGHBORS

Non - Parametric classifier

## LINEAR DISCRIMINANT ANALYSIS

Parametric classifier

## QUADRATIC DISCRIMINANT ANALYSIS

Parametric classifier

# K-NN CLASSIFICATION

## TRAINING DATA

Data set used to create the model
We have taken 80% of the whole data as the training data.

## TESTING DATA

Data set used to validate the performance of the model created. We have taken the remaining 20% of the whole data as the testing data.

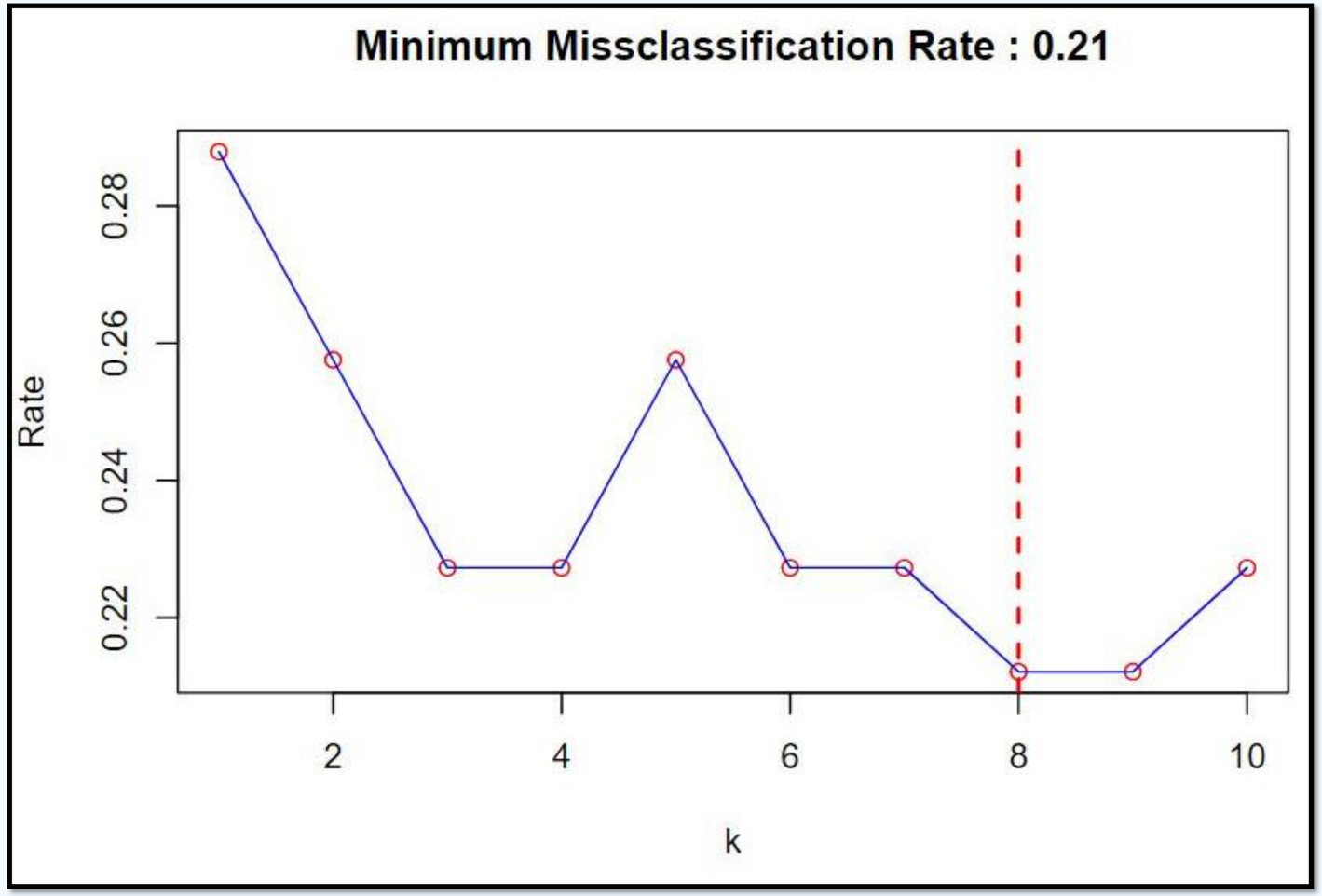
## CLASSIFICATION

k-NN looks at the k points that are closest to the point of unknown class.
The new observation is classified into the group to which the majority of these k points belong.

# WHAT K TO SELECT?

## MISCLASSIFICATION RATE

Plot a range of k values against the misclassification rate and select the K with the least rate.



# RESULT

0 represents: heat stability > 10 mins

1 represents: heat stability < 10 mins

| Result | 0 | 1 |
|---|---|---|
| 0 | 4 | 4 |
| 1 | 10 | 48 |

Accuracy rate = 78.79%

# CONCLUSION

All the casein protein traits are positively correlated to protein content. Beta Lactalbumin B is negatively correlated to protein content but is not too significant. Very few technological traits are correlated to each other.

Cow breeds can be divided into two clusters based on the MIR spectra of the milk samples. The two clusters are most populated by Hol fri and JE breed.

We can classify approximately 75% of the observations have a heat stability of less than 10 mins based on the MIR spectra of their milk samples.

# THANK YOU

## QUESTIONS?