Iris Data Cluster Analysis

Group 3

Carel Ong Shi Ting
Jayne Ng Su Hui
Lee Xian Wei Ivan
Ong Jun Jie
Ong Gi Gi
Valtteri Vaskikari

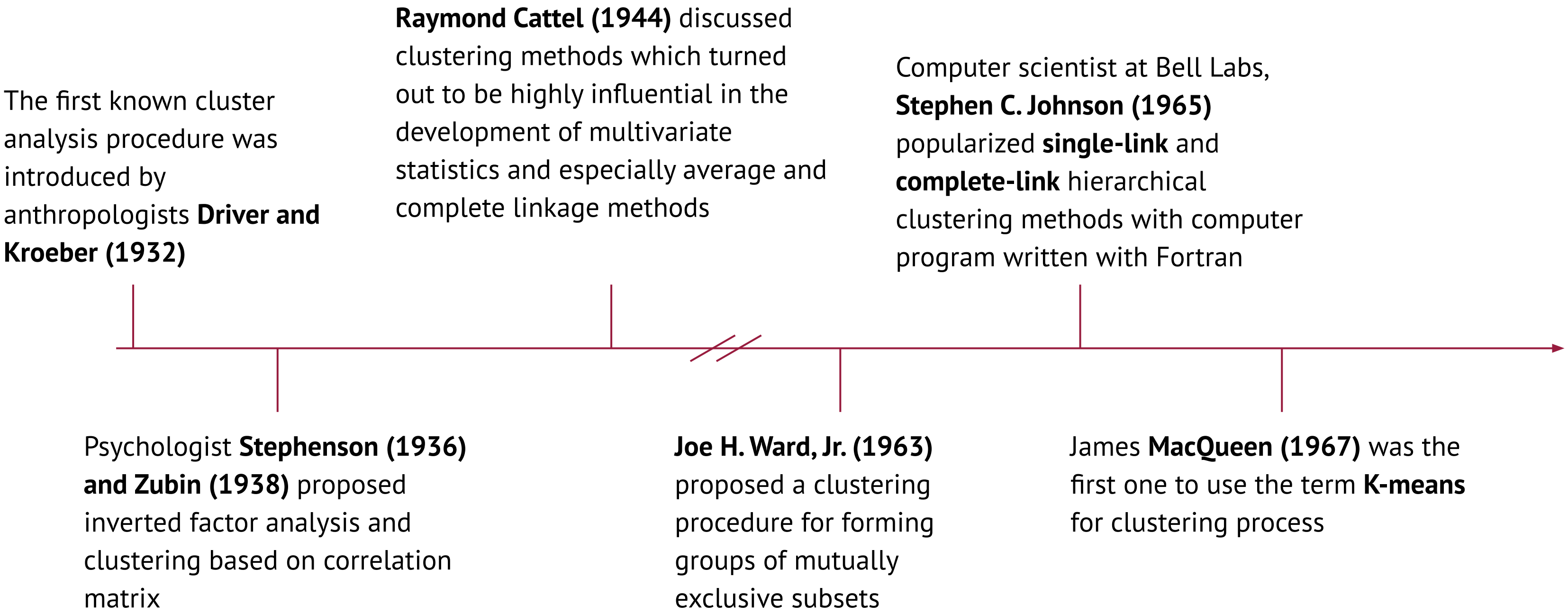
Introduction

Research objective

The objective of the research is to find appropriate clusters for the Fisher's Iris data set.

1. What are the most suitable clustering methods for the Fisher's Iris data set?
2. Based on clustering with selected methods, what is the most optimal number of clusters?
3. Which method provides the best clustering structure/prediction accuracy?

Previous literature in one minute



The first known cluster analysis procedure was introduced by anthropologists **Driver and Kroeber (1932)**

Psychologist **Stephenson (1936) and Zubin (1938)** proposed inverted factor analysis and clustering based on correlation matrix

Raymond Cattel (1944) discussed clustering methods which turned out to be highly influential in the development of multivariate statistics and especially average and complete linkage methods

Joe H. Ward, Jr. (1963) proposed a clustering procedure for forming groups of mutually exclusive subsets

Computer scientist at Bell Labs, **Stephen C. Johnson (1965)** popularized **single-link** and **complete-link** hierarchical clustering methods with computer program written with Fortran

James **MacQueen (1967)** was the first one to use the term **K-means** for clustering process

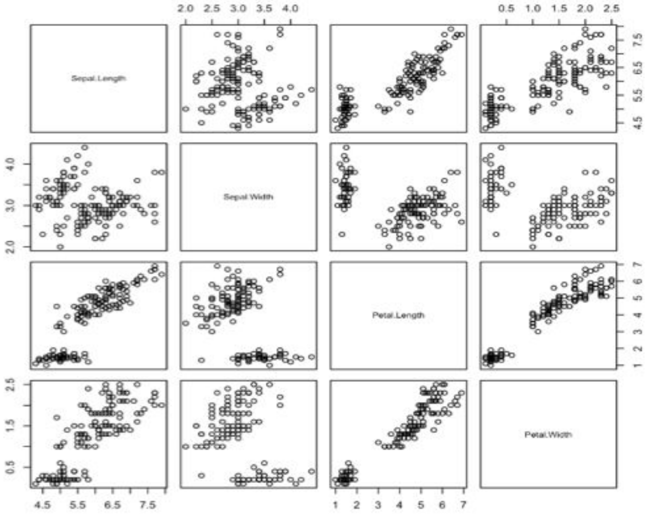
Preliminary Analysis

Iris data set

- The data consists of 150 observations of Iris flowers and 5 features. 4 quantitative predictors are being measured for each of the samples

- The predictors measured the length and width of the sepal and petal of each flower

- Petal Length and Petal Width, Sepal Length and Petal Length as well as Sepal Length and Petal Width have a strong positive linear relationship with one another

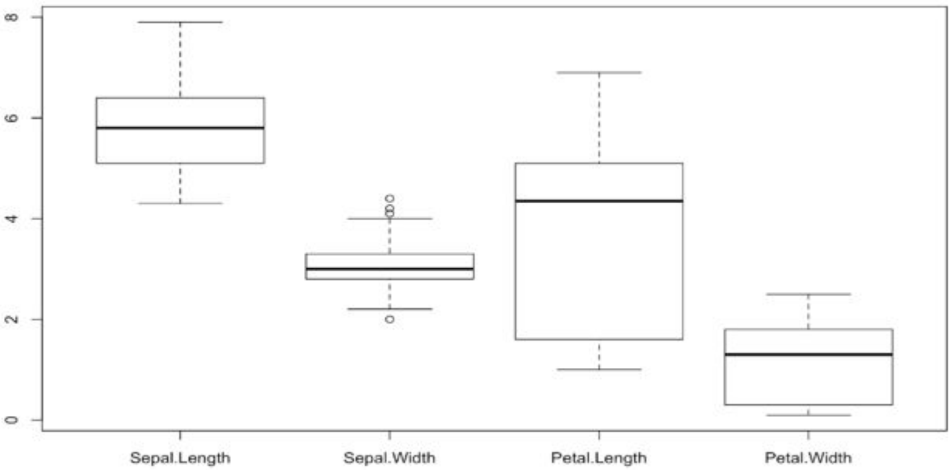


Exploratory data analysis: box and whisker plot

- Looking at each of the predictors' box plot, for Sepal.Width, we can see that there are a few observations that are **1.5 inter quartile ranges** below and above the first and third quantile respectively

- To treat these outliers, we will apply a 90% **winsorization** where data below the 5th percentile are set to the 5th percentile, and data above the 95th percentile are set to the 95th percentile

- After winzorization, no more outliers



Agglomerative Hierarchical Clustering

Methodology

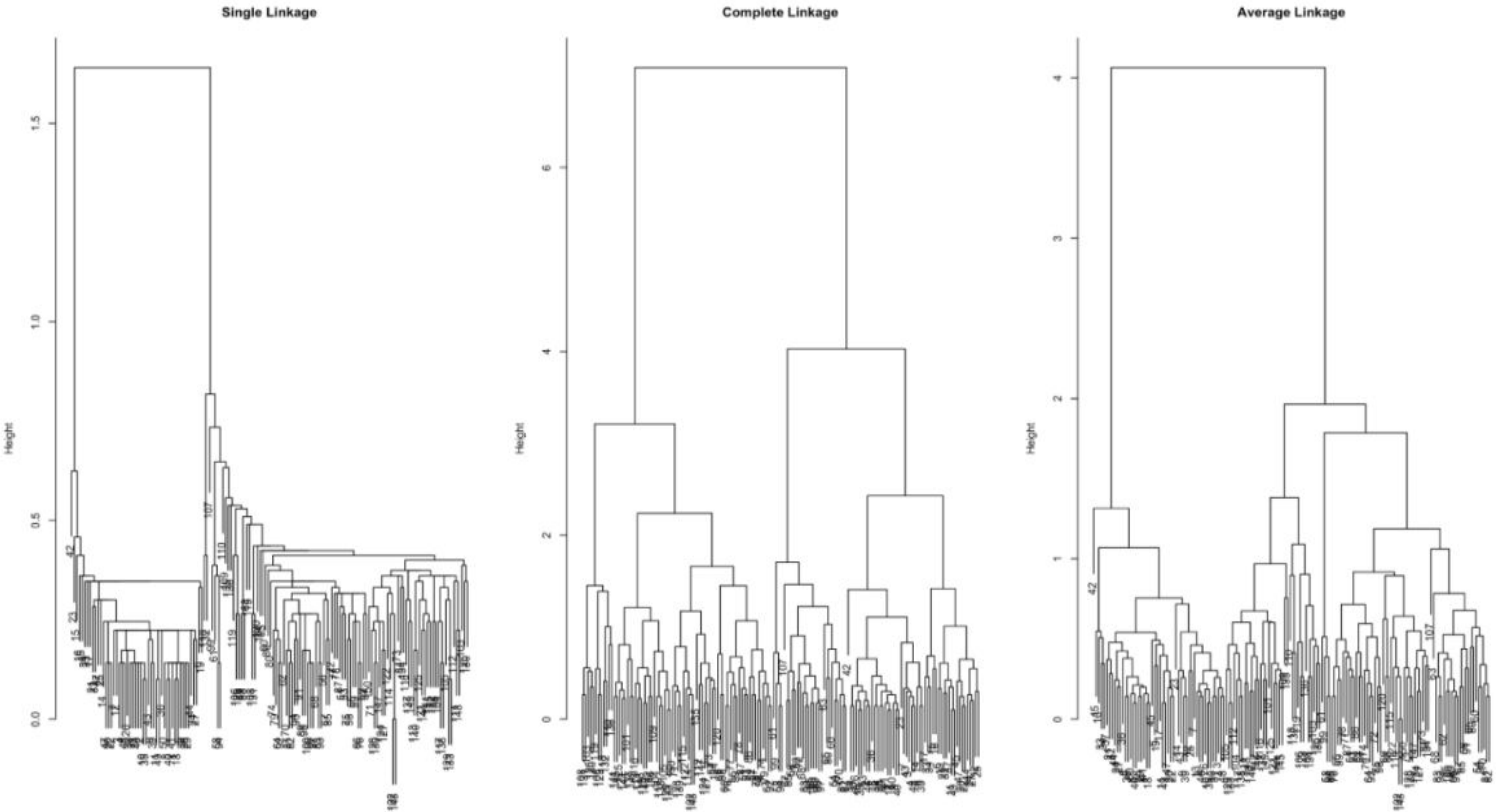
1. Hierarchical clustering → **Dendrograms**
    - ○ Single Linkage
    - ○ Complete Linkage
    - ○ Average Linkage

2. Cut resulting dendrograms at different heights to obtain k number of clusters where $k \in (1, \ldots, 10)$. Calculate **total within cluster sum of squares (WCSS)** and plot **scree plot** to determine the optimal number of clusters.

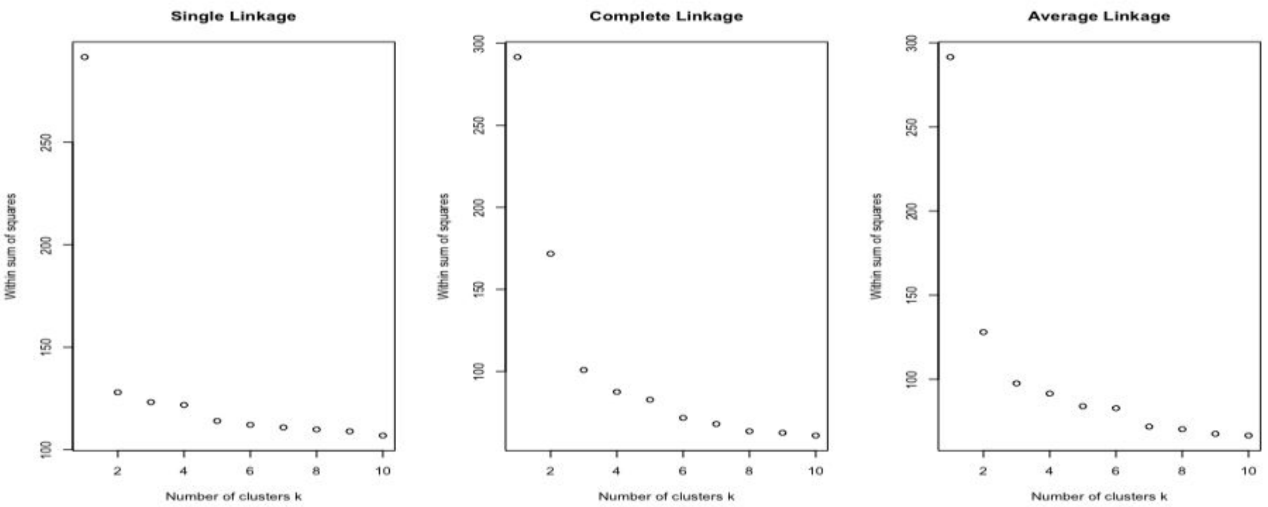
3. Visualize the optimal k's via Principal Component Analysis

4. Compute the agglomerative coefficient using the true clusters for each linkage method. The model with the highest agglomerative coefficient is selected as the best clustering method within hierarchical clustering

Analysis on original Iris data


Single Linkage
Height
0.0
0.5
1.0
1.5
Complete Linkage
Height
0
2
4
6
Average Linkage
Height
0
1
2
3
4

Scree plot: Determining k



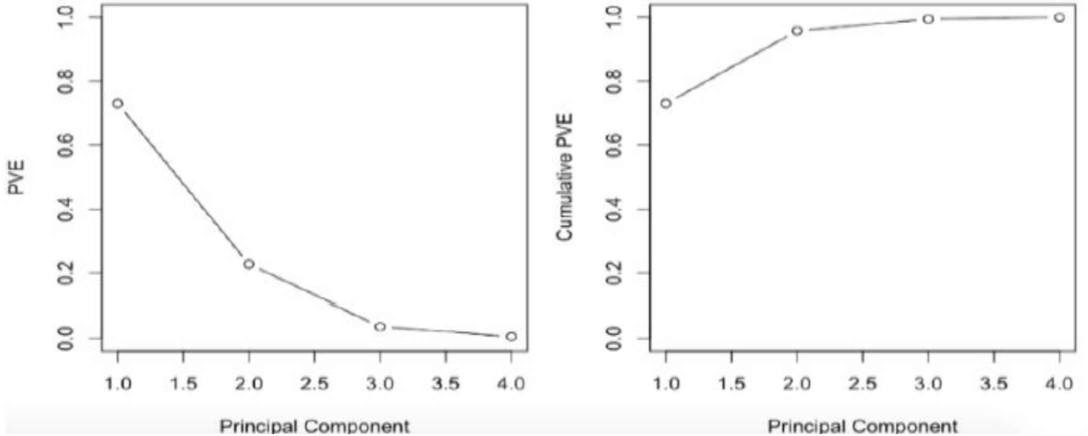
- Single Linkage: Decrease in WCSS minimal after k=2

- Complete Linkage: Decrease in WCSS > 40% before k=3, <14% after k=3

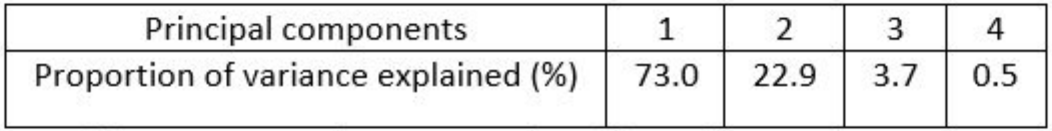
- Average Linkage: Decrease < 10% after k=3 k=2 viable as well

k	1	2	3	4	5	6	7	8	9	10
SL	291.6	128.0	123.3	121.9	114.1	112.2	110.9	109.8	109.0	106.9
CL	291.6	171.7	100.8	87.5	82.7	71.7	67.9	63.6	62.6	60.9
AL	291.6	128.0	97.5	91.5	84.0	82.7	71.8	70.28	67.5	66.5

PCA: To visualise clusters formed



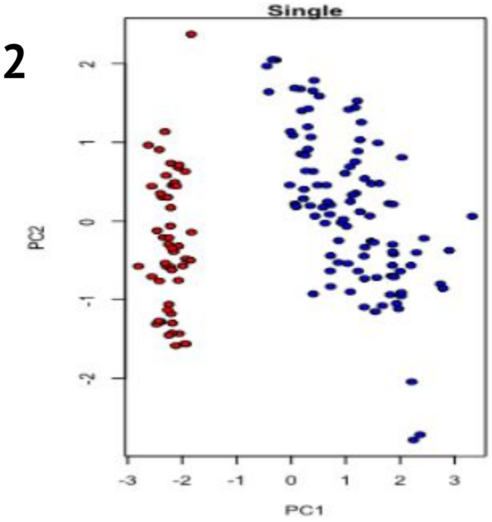
- First 2 principal components explain 96% of total variance
- Third and fourth principal components explain <5% of total variance
- First 2 principal components can replace the 4 Iris variables

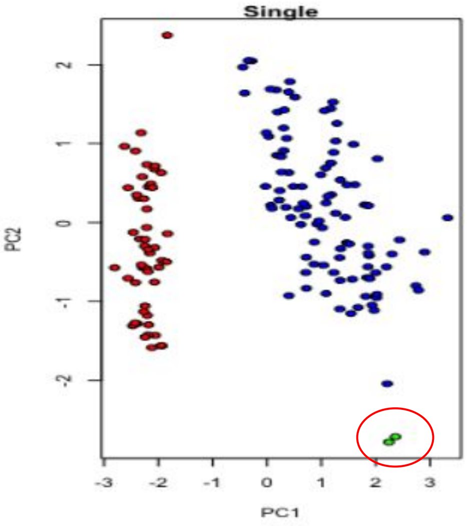


Principal components	1	2	3	4
Proportion of variance explained (%)	73.0	22.9	3.7	0.5

Original PCA Plot for Single Linkage

K=2



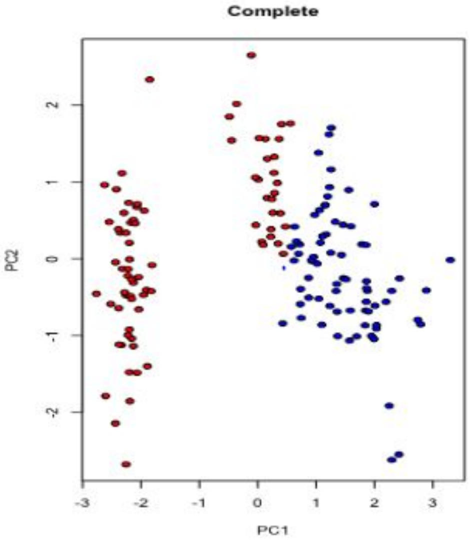


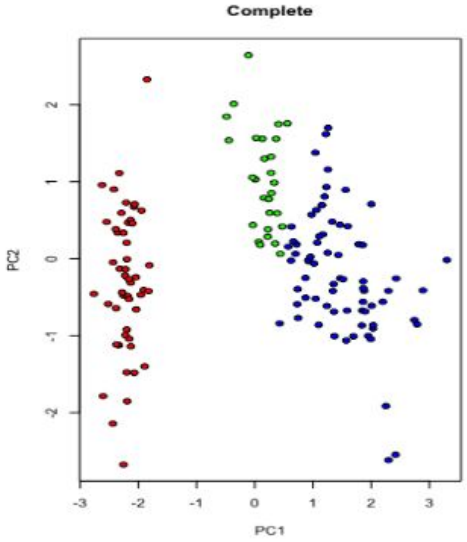
K=3

- K=2: 2 obvious clusters, with clear separation
- K=3: 3 separated clusters, but only 2 observations in the green cluster
- Optimal no. of clusters for single linkage method = 2

Original PCA Plot for Complete Linkage

K=2



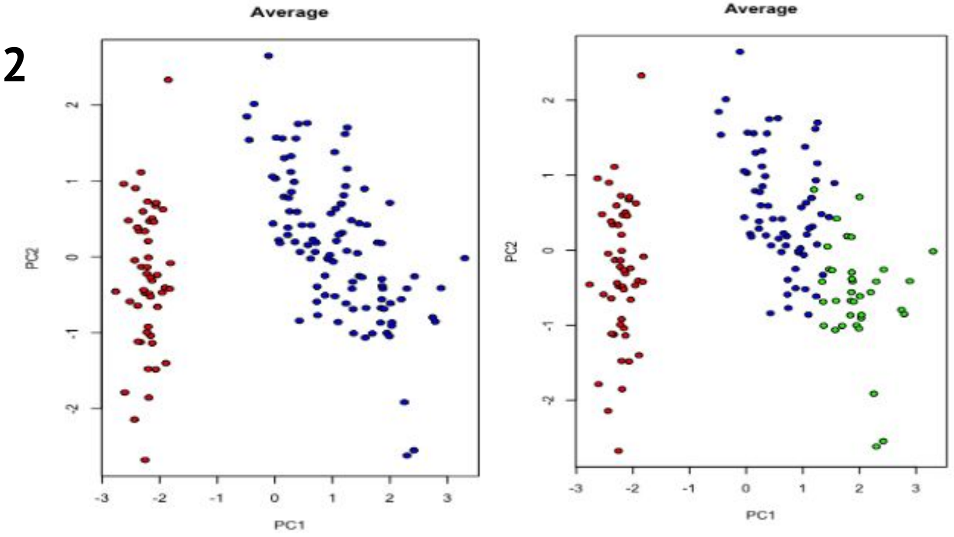


K=3

- K=2: Red cluster split into two, with a group of red points close to the blue cluster
- K=3: Green and blue clusters are close together, but with no overlap
- Optimal no. of clusters for complete linkage method = 3

Original PCA Plot for Average Linkage

K=2



K=3

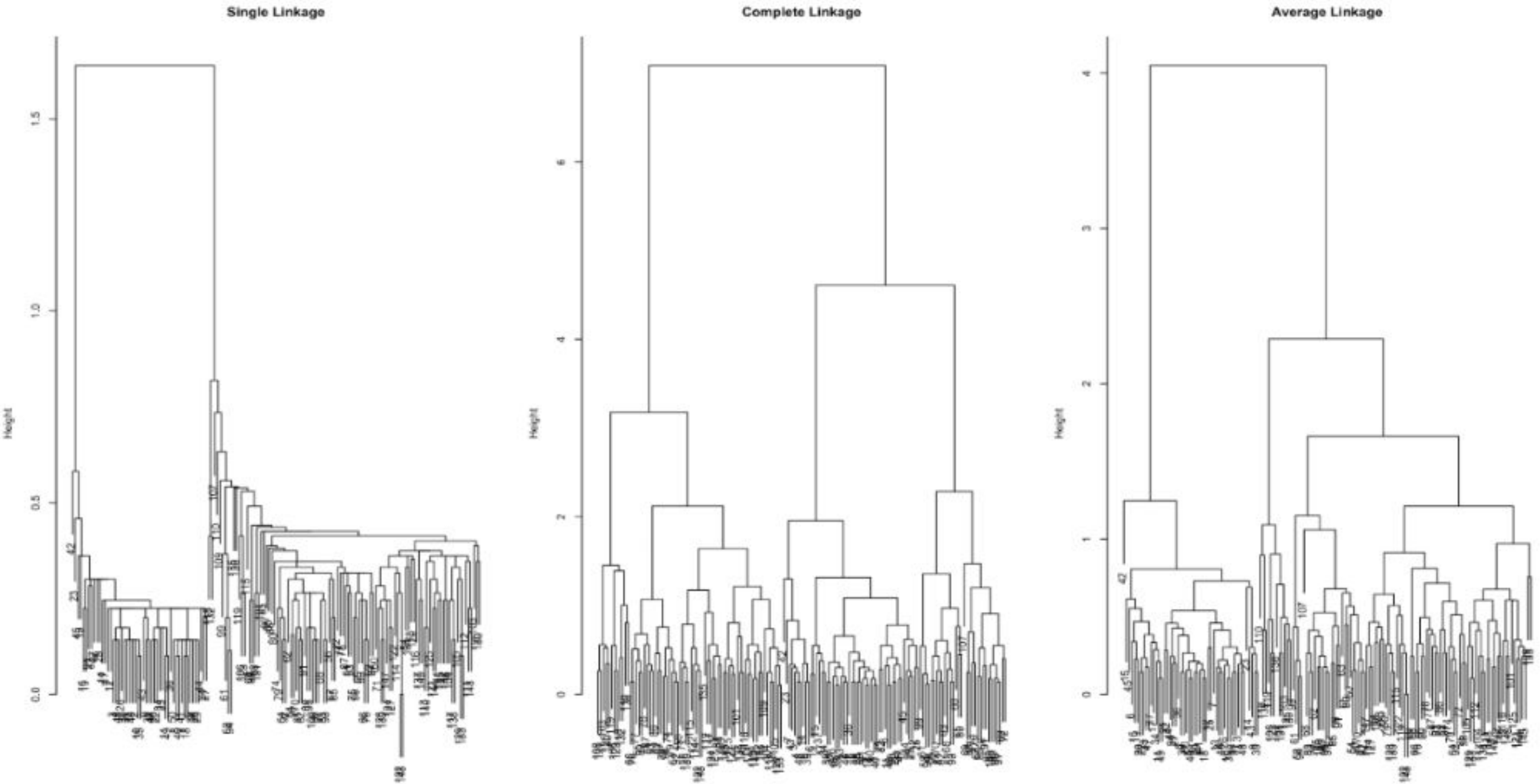
- K=2: 2 obvious clusters, with clear separation
- K=3: Green and blue clusters are close together, but with no overlap
- Optimal no. of clusters for average linkage method = 3

Agglomerative Coefficient

Linkage methods	Single	Complete	Average
Agglomerative coefficient	0.849	0.957	0.930

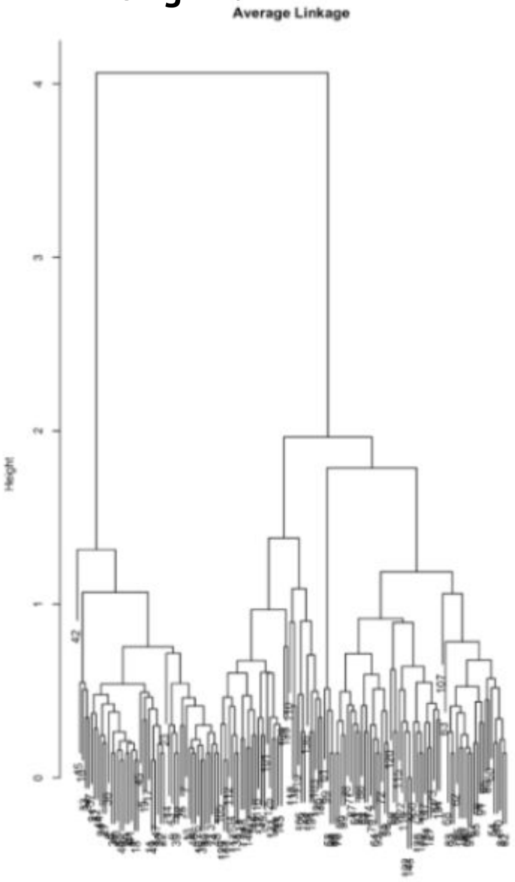
- Highest agglomerative coefficient: Complete Linkage
- Hierarchical clustering using complete linkage provides the best clustering structure

Analysis on 90% winsorized Iris data

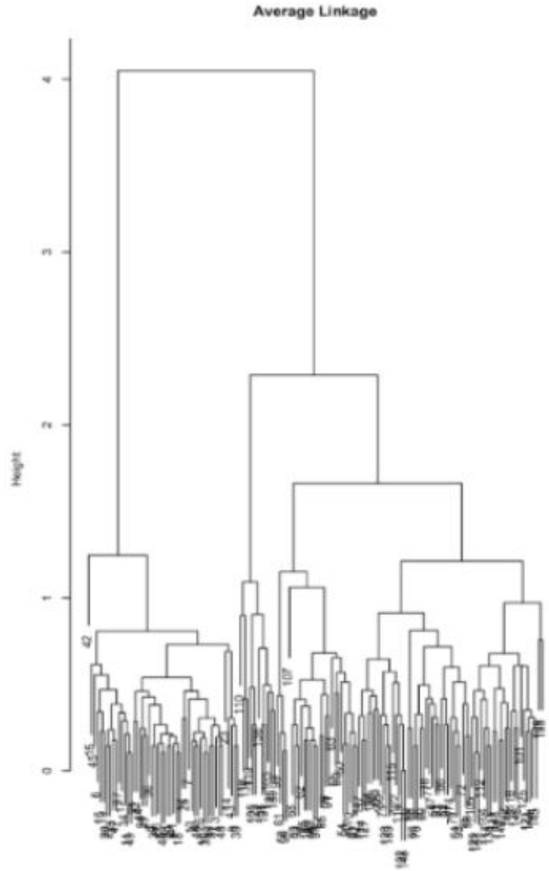

Single Linkage
Height
0.0
0.5
1.0
1.5
Complete Linkage
Height
0
2
4
6
Average Linkage
Height
0
1
2
3
4

Average Linkage

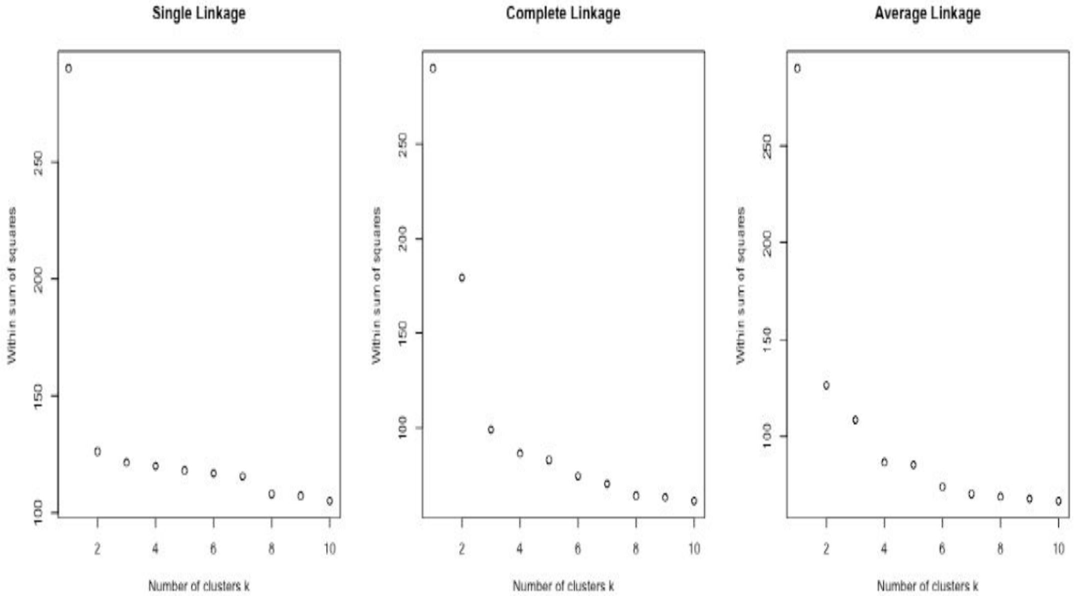
Original



90% Winsorized



Scree Plot: Determining k



- **Single Linkage**: ↓ in WSS is less significant from k=2 onwards
- **Complete Linkage**: ↓ in WSS is less significant from k=3 onwards
- **Average Linkage**: ↓in WSS from k=4 onwards,however elbow is more obvious at k=2

k	1	2	3	4	5	6	7	8	9	10
SL	290.3	126.3	121.5	120.1	118.2	117.0	115.8	108.0	107.2	105.1
CL	290.3	179.5	99.0	86.2	82.9	74.5	70.3	63.8	62.8	61.1
AL	290.3	126.3	108.3	86.3	85.1	73.6	70.0	68.4	67.4	66.4

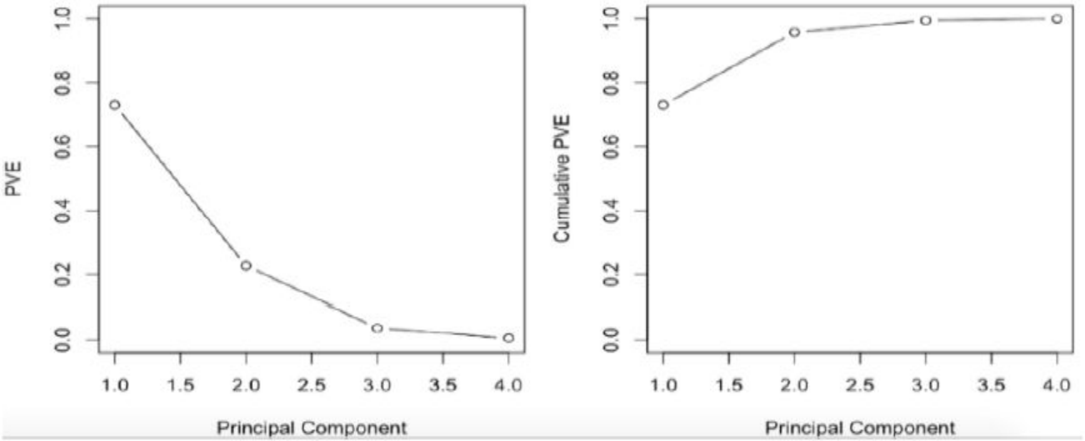
>50%

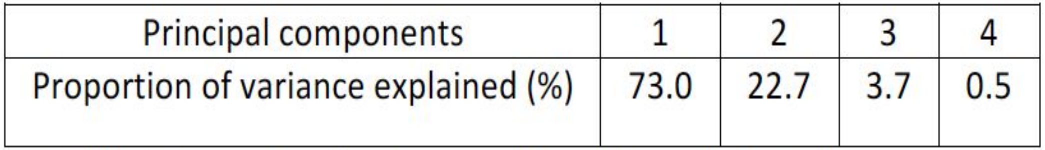
38%

44%

>50%

PCA: Visualizing clusters formed





Principal components	1	2	3	4
Proportion of variance explained (%)	73.0	22.7	3.7	0.5

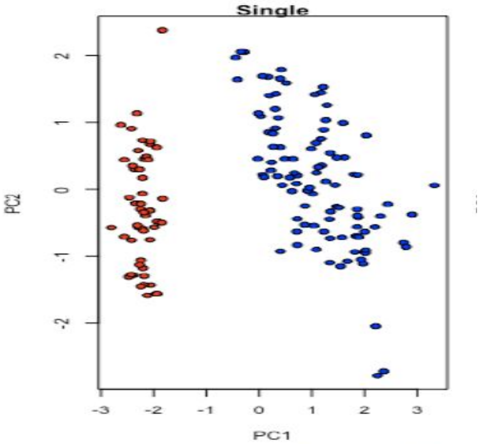
- First 2 principal components explains 96% of the total variance

- Third and fourth principal components explains <5% of the total variance

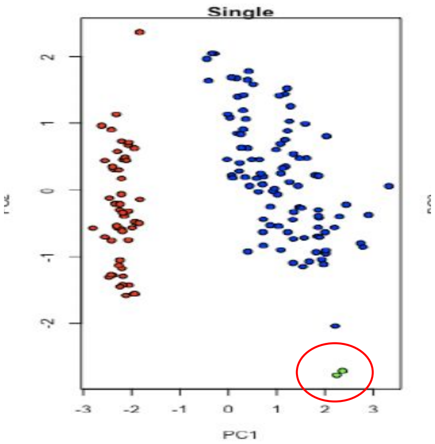
- First 2 principal components can replace the 4 variables

PCA plot for single linkage method

K = 2



K = 3



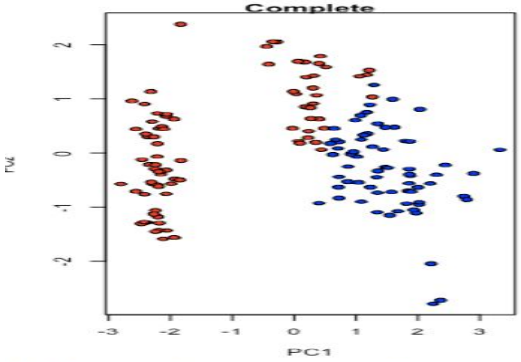
- Plot of k=2: Obvious separation between the 2 clusters.

- Plot of k=3: Obvious separation between the 3 clusters. However, only 2 observations in the green cluster.

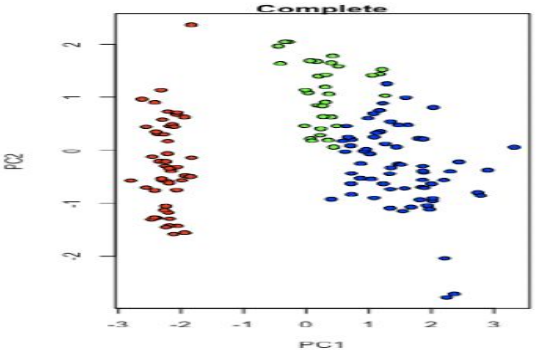
- Optimal number of clusters for single linkage: 2

PCA plot for complete linkage method

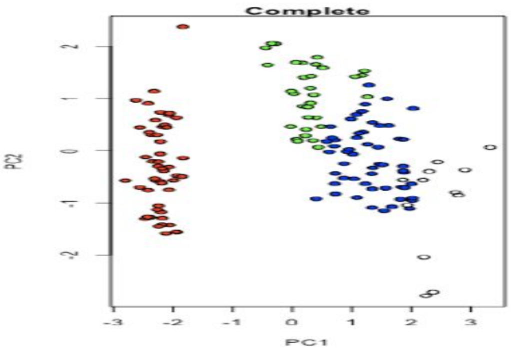
K = 2



K = 3



K = 4



- Plot of k=2: a group of points in the red cluster which are closer to the blue cluster instead

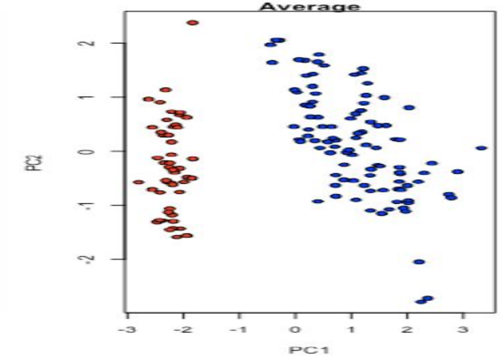
- Plot of k=3: No obvious separation between the green and blue clusters but very little overlap between the clusters.

- Comparing plot of k=3 and k=4, the clusters are more well separated in k=3 as there is lesser overlap among the clusters.

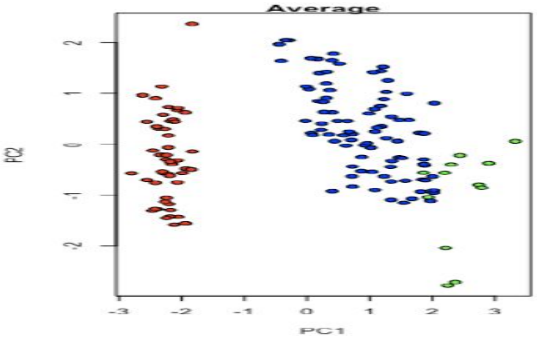
- Optimal number of clusters for complete linkage method: 3

PCA plot for average linkage method

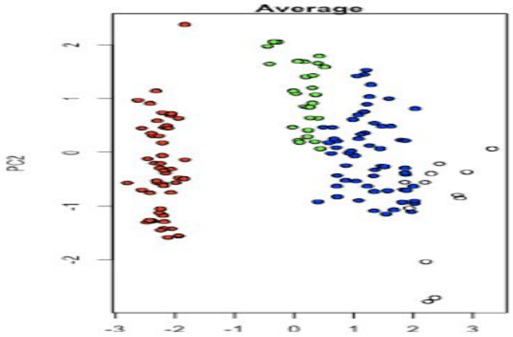
K = 2



K = 3



K = 4



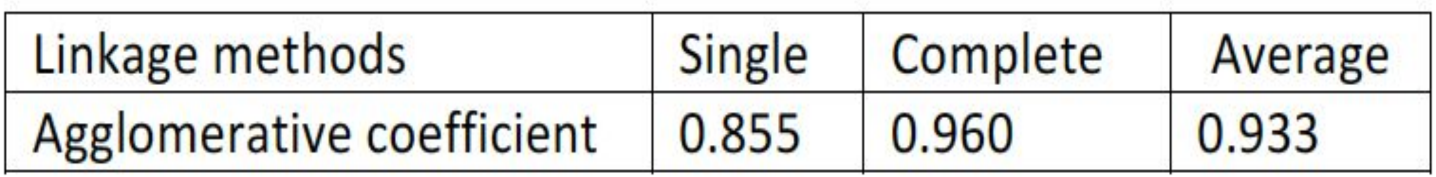
- Plot of k=2: Obvious separation between the 2 clusters

- Plot of k=3: No obvious separation between blue and green clusters but little overlap. Smaller proportion of observations in the green cluster as compared to the red and blue clusters

- Plot of k=4: Little overlap and more equal spread of observations among the 4 clusters

- Optimal number of clusters for average linkage method: 4

Agglomerative coefficient

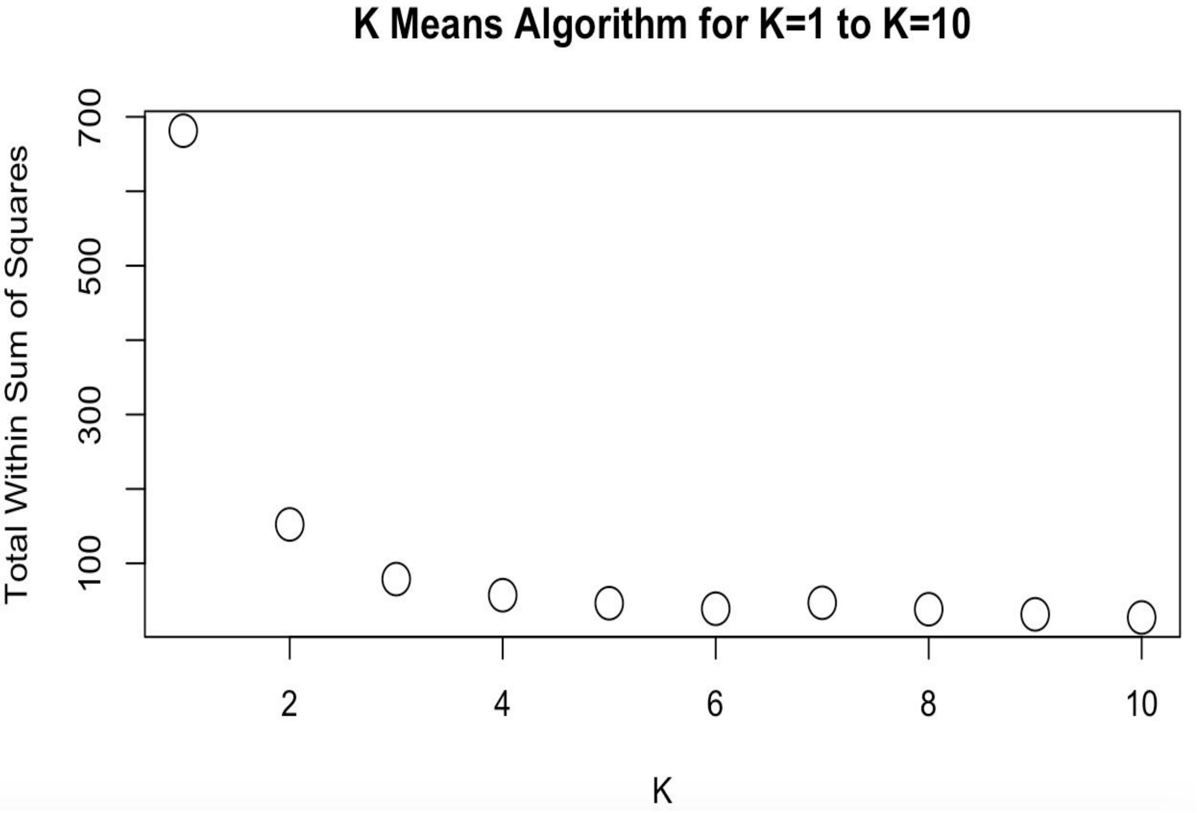


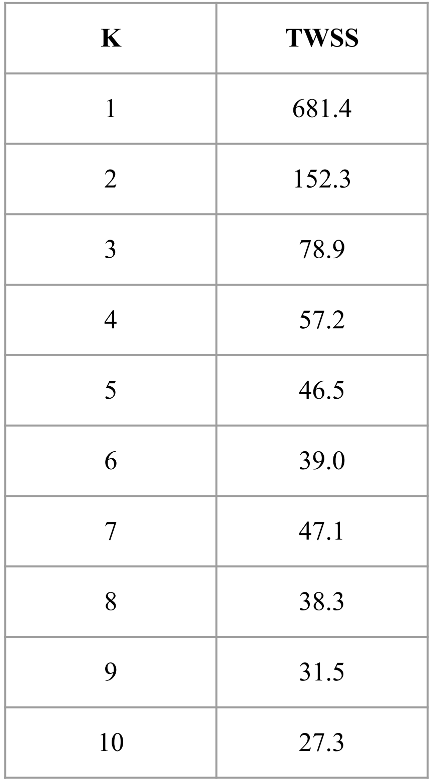
Linkage methods	Single	Complete	Average
Agglomerative coefficient	0.855	0.960	0.933

- Hierarchical clustering using complete linkage provides best clustering structure for the 90% winsorized Iris data.
- Complete linkage method still provides the best clustering structure even after winsorization

Non-Hierarchical Clustering: K-Means

Original data : Number of Clusters K


K Means Algorithm for K=1 to K=10

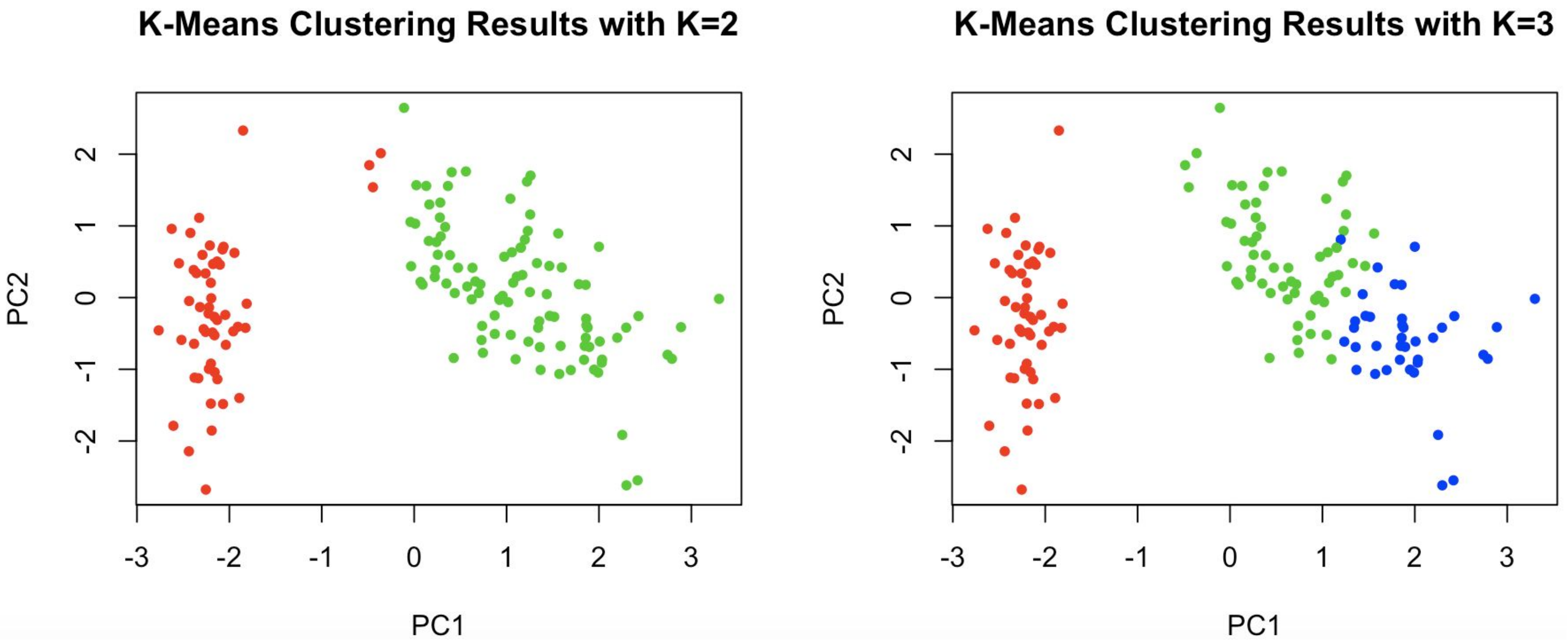


K	TWSS
1	681.4
2	152.3
3	78.9
4	57.2
5	46.5
6	39.0
7	47.1
8	38.3
9	31.5
10	27.3

70%

50%

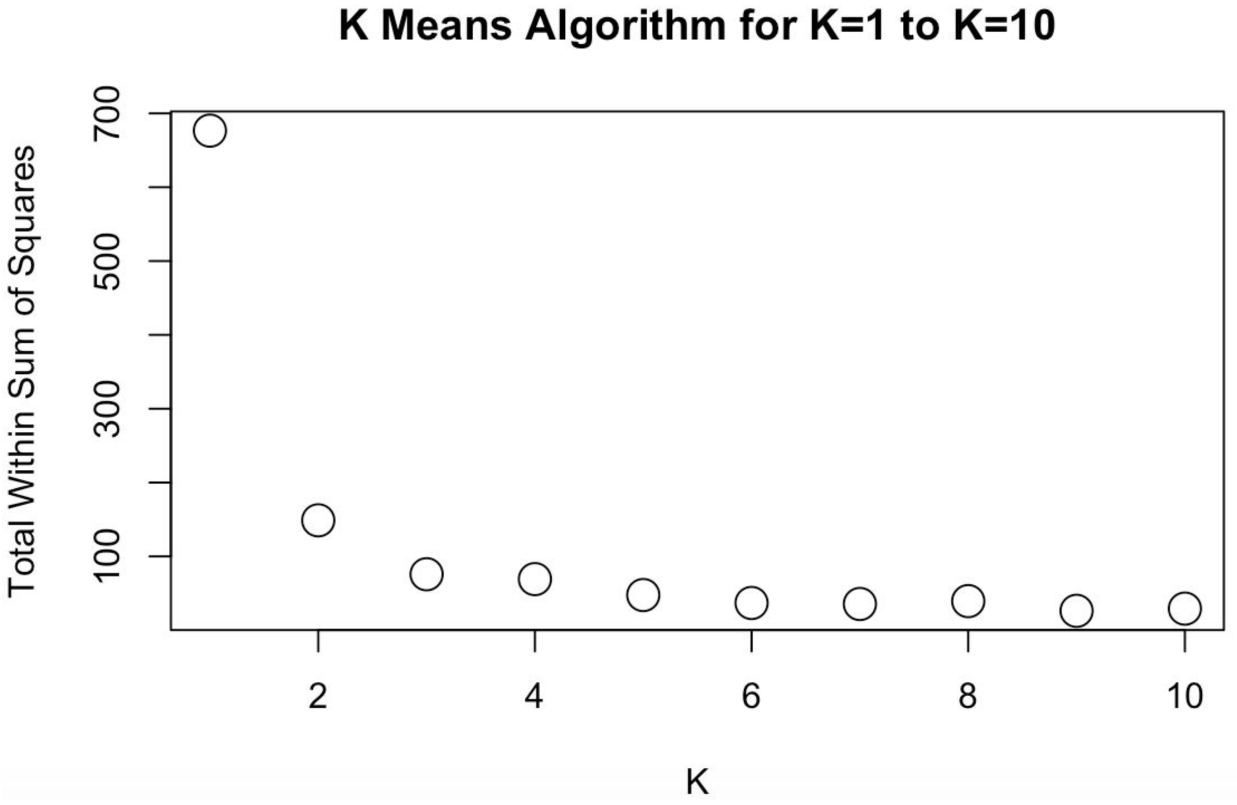
Original data: PCA Plots



K-Means Clustering Results with K=2

K-Means Clustering Results with K=3

Winsorized data : Number of Clusters K

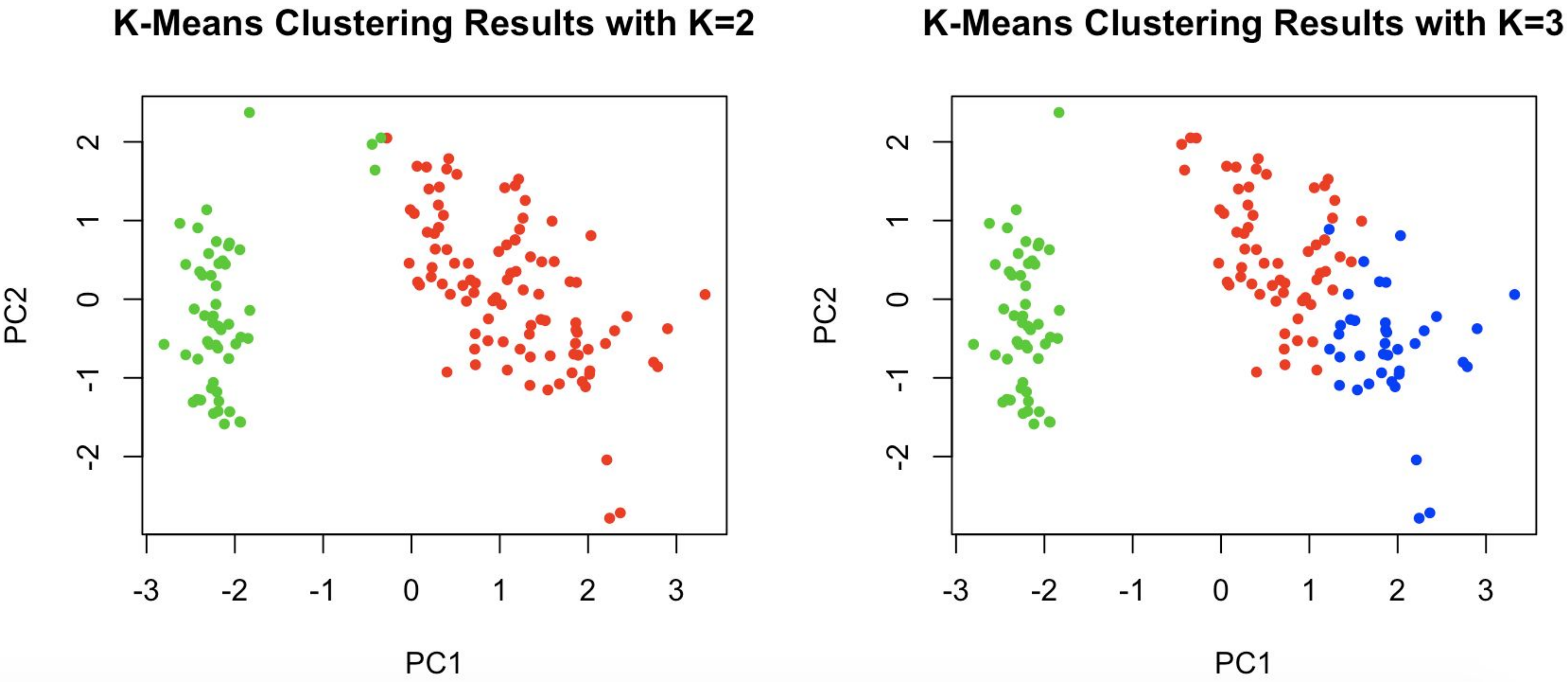


K	TWSS
1	676.6
2	148.8
3	75.8
4	69.2
5	47.6
6	36.9
7	35.5
8	39.4
9	26.3
10	29.2

80%

50%

Winsorized data : PCA Plots



Discussion

Comparison using agglomerative coefficients

Type of Linkage	Single	Complete	Average
AC for Original data	0.849	0.957	0.930
AC for Winsorized data	0.855	0.960	0.933

- Agglomerative coefficients measures the strength of the cluster.
- Complete linkage yields the best agglomerative coefficient (AC).
- Potential outliers did not the cluster analysis → Winsorization may not be need.

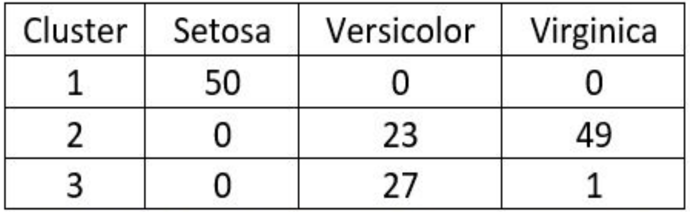
Comparison using Total Within Sum of Squares (TWSS)

	Complete Linkage (K=3)	K-means (K=3)
TWSS for Original data	100.8	78.9
TWSS for Winsorized data	99.0	75.8

- K Means has a generally lower Total Within Sum of Squares than Complete Linkage.
- Lesser variation between observations within each clusters and stronger cluster formed.
- K Means is a better method than hierarchical clustering methods.

Comparison using misclassification rates

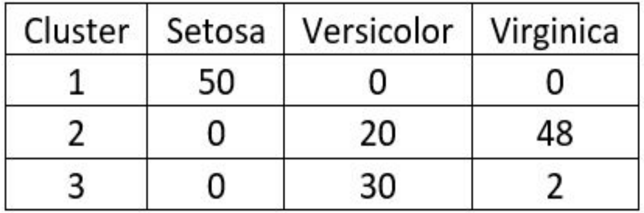
- Only for this section, we assumed that species are known and calculate the misclassification rates to see the accuracy of the methods done.
- Assumption: The true class of an observation is the one where majority of the observations are classified in.



Cluster	Setosa	Versicolor	Virginica
1	50	0	0
2	0	23	49
3	0	27	1

Table 13: Misclassification table for complete linkage method (k=3)

Original data



Cluster	Setosa	Versicolor	Virginica
1	50	0	0
2	0	20	48
3	0	30	2

Table 17: Misclassification table for complete linkage method (k=3)

Winsorised data

Comparison using misclassification rates

	Complete Linkage (K=3)	Single Linkage (K=2)	Average Linkage (K=3,4)	K-Means Clustering (K=3)
Original Data	24/150 = 16%	50/150 = 33.3%	14/150 = 9.3% (K=3)	16/150 = 10.7%
Winsorized Data	22/150 = 14.7%	50/150 = 33.3%	37/150 =24.7% (K=4)	16/150 = 10.7%

Comparison using misclassification rates

1

Some methods perform better on data with outliers (AL) and some methods perform otherwise. (CL)

2

AL has the best accuracy for original data (9.3%) while K Means performs the best for winsorised data. (10.7%)

K means is more consistent with relative low error rates of 10.7% for original data and retained its error rate when winsorised data was used.

Thus, K Means is chosen as the optimal method, with 3 clusters as the optimal number of clusters.

Thank you!

Fonts & colors used

This presentation has been made using the following fonts:

Poppins
(https://fonts.google.com/specimen/Poppins)

PT Sans
(https://fonts.google.com/specimen/PT+Sans)

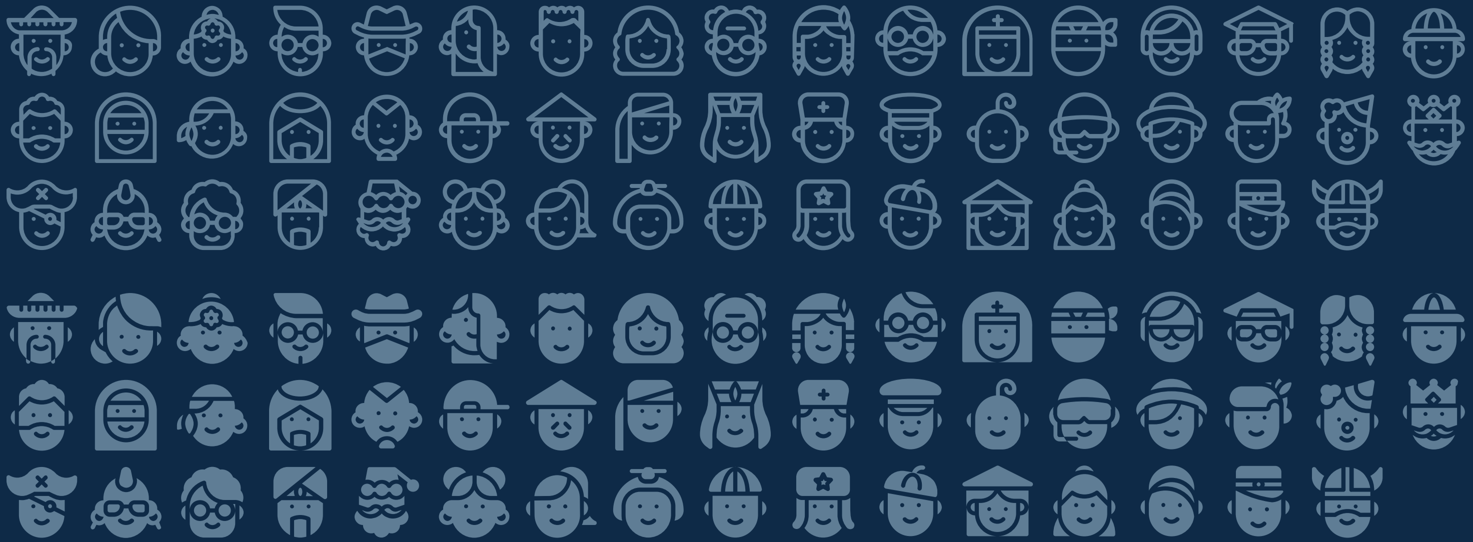
#981c3e	#a3d1aa	#26284e

...and our set of editable icons

You can resize these icons, keeping the quality.

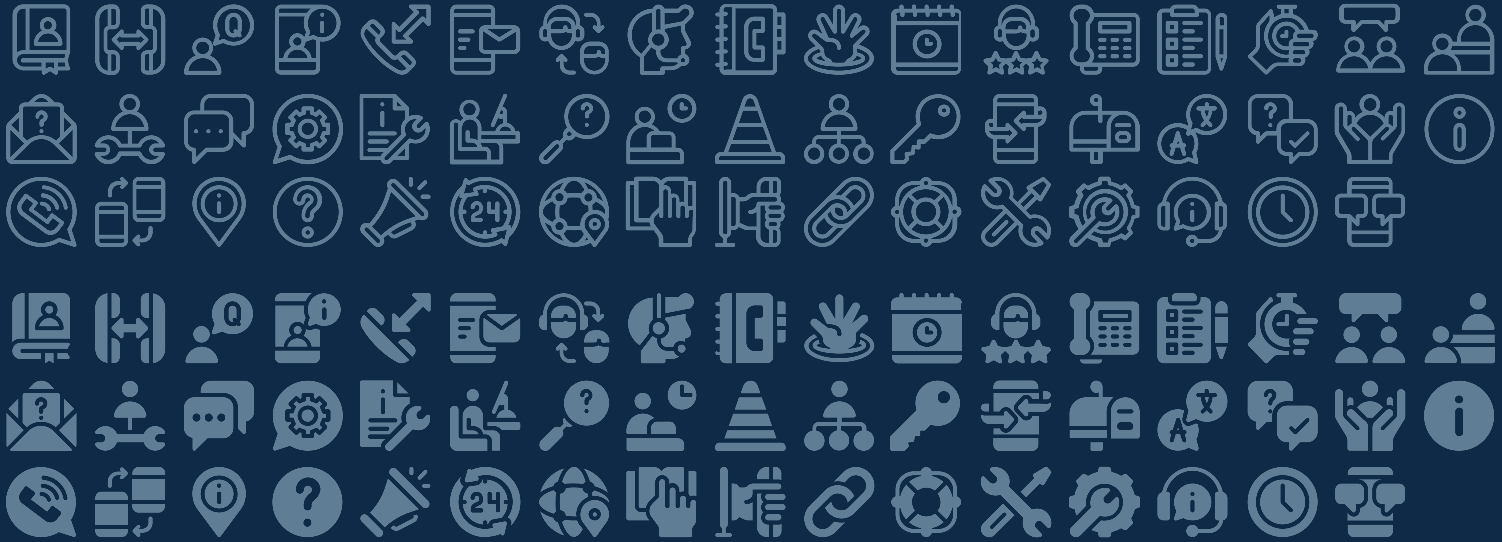
You can change the stroke and fill color; just select the icon and click on the paint-bucket/pen.

Avatar Icons



Educational Process Icons

Help & Support Icons



Nature Icons