ST3240 Project Report

Group 3

Name
Carel Ong Shi Ting
Jayne Ng Su Hui
Lee Xian Wei Ivan
Ong Gigi
Ong Jun Jie
Vaskikari Valtteri Elias Aleksander

Table of Contents

1. Introduction

Johnson (1967), wrote his first paper which concreted the previous work of Ward (1963) and others about the concept of hierarchical clustering scheme. He provided two ways to construct hierarchical clustering schemes which eventually became probably the most fundamental procedures of clustering: single-link and complete-link hierarchical clustering methods.

MacQueen (1967) was the first one to use the term K-means for clustering process where items were allocated to K different clusters according to their distances to nearest cluster centroid (arithmetic mean). Essentially the MacQueen's K-means method recalculated cluster centroids after each allocation, so that the process usually relocated some items in the following stages if the distance between recalculated cluster centroid and a designated item became too large.

Over the years, extensive research has been done and Iris data has shown to be useful for cluster analysis. Hence, we will be using this data to investigate the effectiveness of different clustering methods.

2. Objective

The objective of the research is to find appropriate clusters for the Fisher's Iris data. This objective is guided through three research questions:

What are the most suitable clustering methods for the Fisher's Iris data?

Based on clustering with selected methods, what is the most optimal number of clusters?

Which method provides the best clustering structure/prediction accuracy?

3. Data Description and Exploratory Data Analysis

Data description
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. They are labelled as Petal Length, Petal Width, Sepal Length and Sepal Width. There is also a qualitative predictor, iris, which indicates the type of species for each observation.

The qualitative variable, Species, is removed from the data for this study. Thus, the type of species in which the observations are categorized is unknown and will be the main objective of this study. We first plot the scatter plot to find out about the relationship between the predictors and obtain the summary statistics to get a brief idea of the dataset.

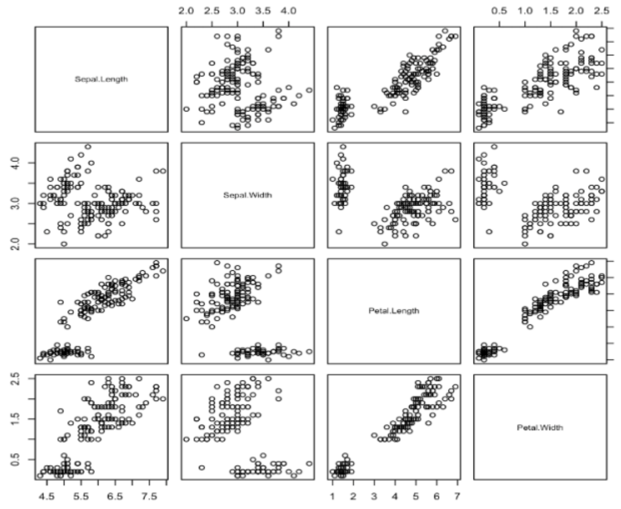
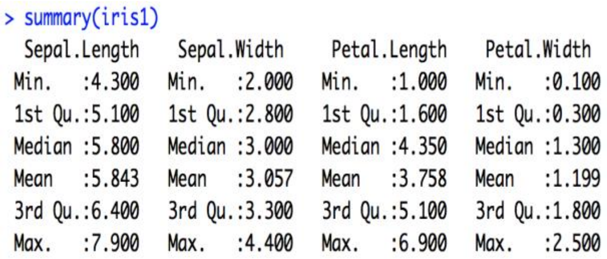


Figure 1: Scatter plot between predictors



```
> summary(iris1)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
```

Figure 2: Summary statistics for predictors

	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	1.00	-0.1175698	0.8717538	0.8179411
Sepal Width	-0.1175698	1.00	-0.4284401	-0.3661259
Petal Length	0.8717538	-0.4284401	1.00	0.9628654
Petal Width	0.8179411	-0.3661259	0.9628654	1.00

Table 1: Correlation matrix of predictors

From *Figure 1*, 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. Based on *Table 1,* these variables have generally high correlation coefficient with one another - 0.9628654, 0.8717538, 0.8179411 respectively. Additionally, Sepal Width and Sepal Length have a weak negative correlation of -0.1175698 while Petal Length and Sepal Width, Sepal Width and Petal Width both have a moderate negative linear relationship with one another with correlation coefficient of 0.4284401 and -0.3661259 respectively.

Exploratory Data Analysis

We further plot boxplots to help identify for potential outliers and if so, perform winsorization to the dataset.

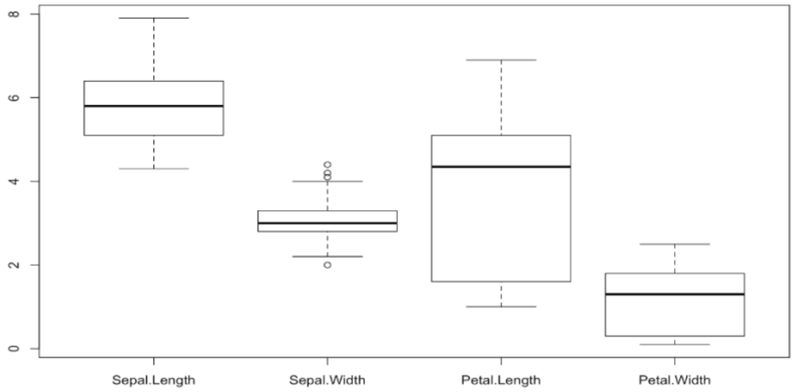


Figure 3: Box plots of predictors of original data

Looking at each of the predictors' box plot in *Figure 3*, for Sepal.Width, we can see that there are a few observations that are 1.5 IQR 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. The winsorized estimators obtained are generally more robust to outliers. The 5th percentile and 95th percentile of Sepal.Width variable is 2.345 and 3.800 respectively.

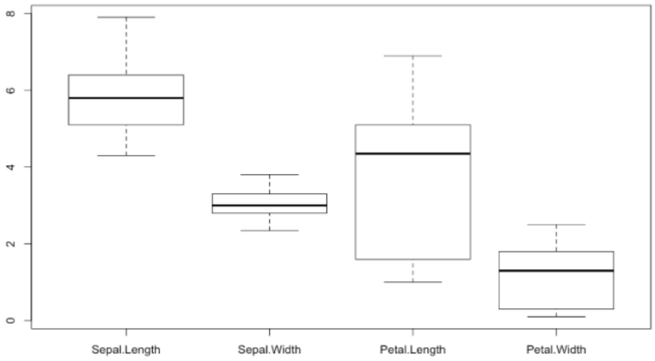


Figure 4: Box plots of predictors after 90% winsorization

Applying winsorization before performing any cluster analysis is important as some of the hierarchical clustering methods are sensitive to outliers and thus, the effect of outliers may affect our conclusion eventually. In *Figure 4*, we can see that after the 90% winsorization is done, there is no longer any outliers. Thus, we will be making use of both the original and winsorised data for our analysis in this report to help us find the optimal method which meets our goal.

4. Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering works by first treating all items as a cluster and computing all the pairwise inter-cluster dissimilarities. Dissimilarities (determined by the type of linkage used) between each cluster are computed and the two clusters that are least dissimilar will be merged. This will happen repeatedly with pairwise dissimilarities being recalculated each time until all observations are combined into one main cluster. In this section, we attempt to perform agglomerative hierarchical clustering on both the original and winsorized data using three linkage methods: Complete, Single and Average Linkage. The following steps are taken to select the best agglomerative hierarchical model.

i. Perform hierarchical clustering on the data and compare the dendrograms obtained to observe the differences in the structure of the clusters formed.

ii. Using the dendrograms obtained, we will cut it at different heights to obtain k number of clusters where $k \in (1,10)$. Within sum of squared error for each k is calculated and a scree plot is constructed to determine the optimal number of clusters.

iii. Convert the data into their principal components using principal component analysis to visualize and better decide on the optimal number of clusters

iv. Compute the agglomerative coefficient using the true clusters for each linkage method. Agglomerative coefficient (AC) is a measure of the strength of a clustering structure. The agglomerative coefficient measures the dissimilarity of an object to the first cluster it joins, divided by the dissimilarity of the final merger in the cluster analysis, averaged across all items. A value closer to 1 indicates stronger clustering structure. The model with the highest agglomerative coefficient is selected as the best clustering method within hierarchical clustering for the iris dataset.

4.1 Analysis on the original Iris data

Dendrograms produced under hierarchical clustering for each method are shown in Figure 1 below.

Dendrograms for Hierarchical Clustering

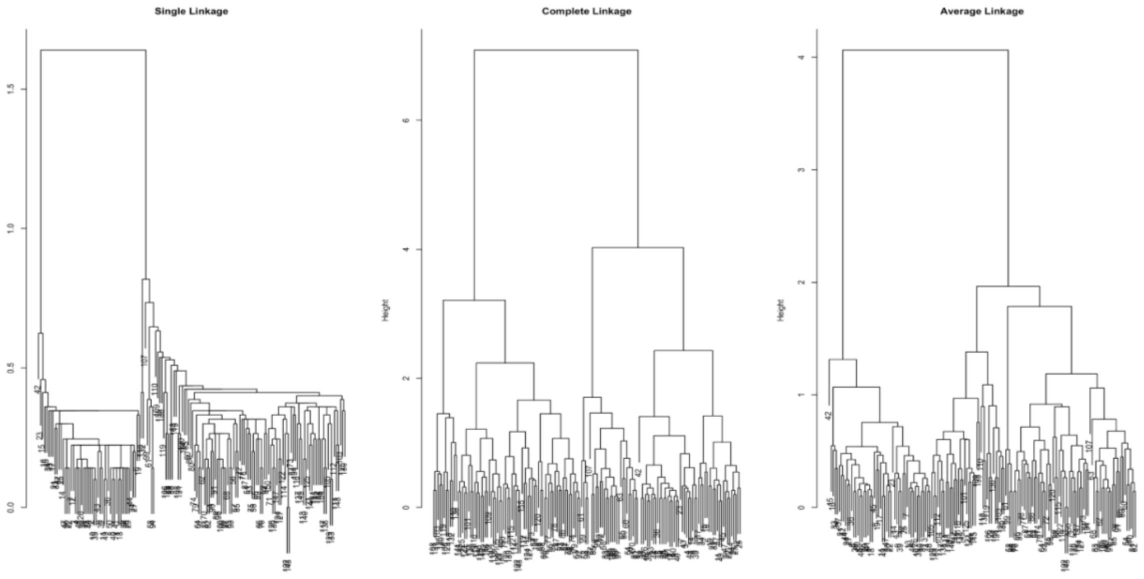


Figure 5: Dendrograms for single, complete and average linkage using original Iris data

Under the Single Linkage (SL) method, minimal inter-cluster dissimilarity is used where the distance between two clusters is the minimum value of the pairwise inter-cluster dissimilarities. Single linkage hierarchical clustering works by using the minimum nearest neighbor to combine two items into a single cluster. Since single-linkage merge clusters by using the shortest linkage between them, this method cannot discern poorly separated clusters. Moreover, single linkage also has the propensity to chain clusters together as we can see from the dendrogram for single linkage in *Figure 5*. These clusters may be inaccurate as items on opposing ends of the chain or cluster may be dissimilar. The scatterplot of the iris data seems to reveal two main clusters. However, further investigation is needed as there could potentially be more clusters within the scatterplot that are significant but close to each other and this would be difficult for single linkage method to identify.

Under the Complete Linkage (CL) method, dissimilarity is measured by the maximum pairwise distance between two clusters. Complete linkage method works by using the minimum furthest neighbor distance to combine two items into a single cluster. This form of linkage method ensures that all observations in a cluster are within a maximum distance and tends to produce clusters with similar diameters and will generally yield a balanced dendrogram as we can see in *Figure 5* above.

Under the Average Linkage (AL) method, the average Euclidean distance of every observation between clusters is calculated as a measure of dissimilarity. At every iteration, 2 clusters with the shortest average distance will be merged into one cluster until 1 big cluster containing every observation is formed. This algorithm is represented as a tree, known as a dendrogram. The dendrogram for average linkage is shown above. As the average Euclidean distance is considered as dissimilarity, it is known to produce trees that are more balanced as compared to Single Linkage as shown in *Figure 5*. The dendrogram for single Linkage shows more chaining compared to the dendrogram produced by average linkage.

Determining optimal number of clusters using within cluster sum of squares

The dendrogram obtained from hierarchical clustering cannot tell us the optimal number of clusters that would best cluster the iris dataset. To minimize within cluster variation, we will use total within cluster sum of squares to determine the optimal number of clusters among $k \in (1,10)$ clusters. Total within cluster sum of squares calculates the total squared distance between observations within a cluster and sums these values for all clusters. As total within sum of squares decreases with increasing clusters, the aim is to reduce total within cluster sum of squares, while choosing as little clusters as possible to prevent overfitting. Having a lower total within cluster sum of squares means that there is less variation within each cluster, hence signifying better clustering. Scree plot is shown below to determine the optimal number of clusters. To do so, we first plot a scree plot and look out for:

1. An 'elbow' in the plot. The elbow shows that increasing the number of clusters after a certain number of clusters will not decrease the total within cluster sum of squares significantly, indicating that clusters are well defined and well split.
2. If 'elbow' is not obvious, we look that the numbers at each point and find out where the total within cluster sum of squares decreases at a lower rate.

Figure 6 below shows the scree plot for all 3 methods of hierarchical clustering:

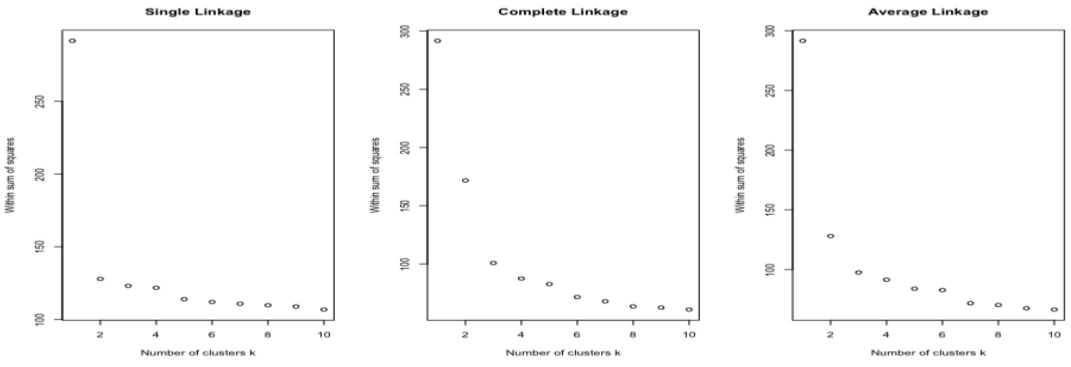


Figure 6: Plot of within sum of squares against no. of clusters for single, complete, average linkage

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

Table 2: Within sum of squares values for k clusters for single, complete, average linkage

From the scree plot in *Figure 6* for single linkage method, there is a significant decrease in within cluster sum of squares from k=1 to k=2. From *Table 2*, within cluster sum of squares decreases by more than 50% and as k increases from 2 to 10, the decrease in within cluster sum of squares is minimal. Hence, this indicates that the optimal number of clusters for single linkage method is 2.

From the scree plot in *Figure 6* for complete linkage and average linkage method, there is also a significant decrease in within cluster sum of squares from k=1 to k=2.From table 2, within cluster sum of squares decreases by more than 40%. Furthermore, as k increases from 2 to 3, within cluster sum of squares decreases by 41% for complete linkage method and 24% for average linkage method. The decrease is significantly lower as k increases from 3 to 10. Hence, this may indicate that the optimal number of clusters for complete and average linkage is 3. However, it is good to note that k=2 can also be regarded as an optimal number of clusters for average linkage because the elbow is more evident in k=2 than k=3 although within sum of squares did not stabilize after k=2 (difference in within sum of squares between k=2 and k=3 is still quite large). Moreover, k=2 splits the dataset quite clearly, as seen in the pairwise plot of the data in *Figure 1*, as well as the dendrogram for the original iris dataset for these 2 methods in *Figure 5*.

Principal Component Analysis (PCA) to visualize the clusters formed

Plotting the clusters can help us visualize and identify the optimal number of clusters and decide between 2 clusters or 3 clusters for some of the linkage methods. However, since the Iris data has 4 variables:” Sepal Length”, “Sepal Width”,” Petal Length” and” Petal Width”, we cannot obtain a 2-dimensional plot to visualize the clusters formed. Therefore, in order to get a 2-dimensional plot, we will perform PCA to reduce the dimensions of the data.

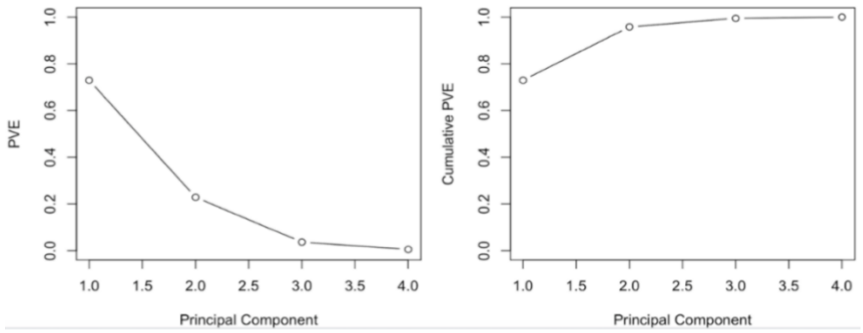


Figure 6: Scree plot

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

Table 3: Proportion of variance explained by each principal component

From the scree plot in *Figure 7*, we can see that the first 2 principal components explain a significantly higher proportion of variance compared to the 3rd and 4th principal components. Based on *Table 3*, the 1st and 2nd principal components combined explain about 96% of the total sample variance. On the other hand, the 3rd and 4th principal components combined explain less than 5% of the sample variance and hence they are probably redundant noise components. This implies that the contribution to the total variance by the 3rd and 4th principal components are negligible. Therefore, the first 2 principal components can effectively summarize the data and replace the 4 variables without losing much information.

Using the 2 sample principal components, we will now plot the k=2 and k=3 clusters for each of the linkage methods.

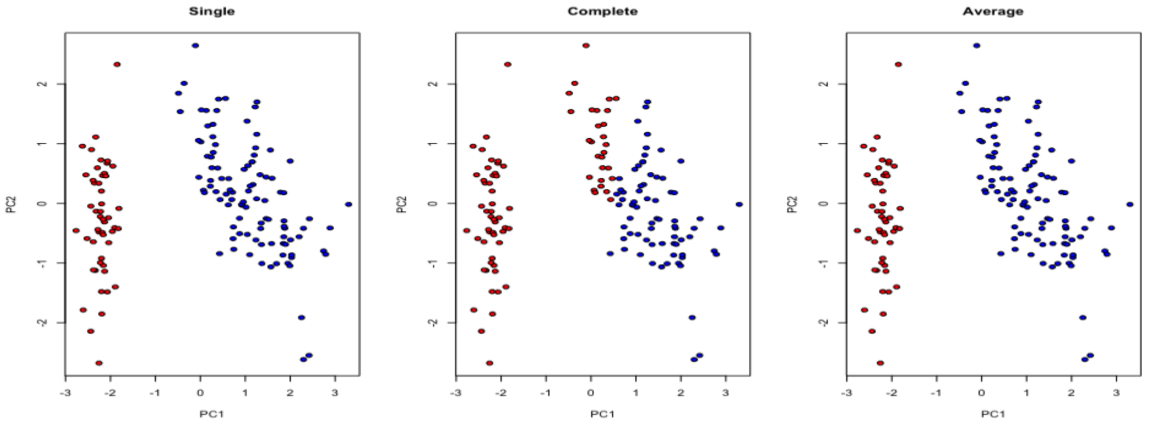


Figure 7: k=2 clusters formed using 2 principal components for each linkage method

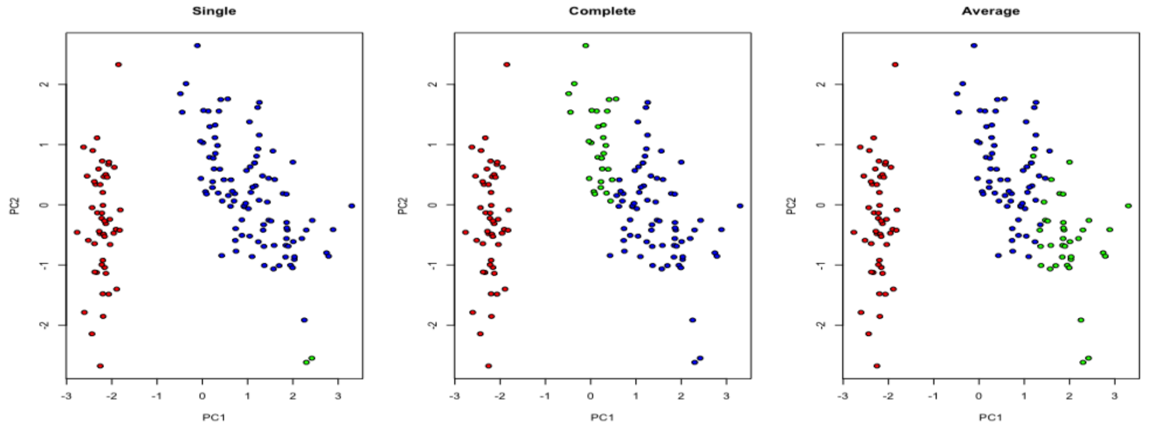


Figure 8: k=3 clusters formed using 2 principal components for each linkage method

From the plot in *Figure 8* and *Figure 9* for single linkage method, we can see that there is an obvious separation between both k=2 and k=3 clusters. However, for k=3 clusters, there are only 2 observations in the green cluster. Therefore, we conclude that k=2 cluster is the optimal number of clusters for the single linkage method.

From the plot in *Figure 8* for complete linkage method, though there is no obvious separation between 2 clusters, there is still a clear separation between them in the sense that there is no overlap between the 2 clusters. However, there is a group of points in the red cluster which are closer to the blue cluster instead. On the other hand, the plot for complete linkage in *Figure 9* shows a more obvious separation between the red and blue clusters. Although there is no obvious separation between the green and blue clusters, there are no overlap between these 2 clusters. Therefore, we conclude that k=3 is the optimal number of clusters for the complete linkage method.

As explained in the previous section, k=2 and k=3 are viable options for the final number of clusters using average linkage. However, we will choose k=3 as the optimal number of clusters as its total within sum of squares is significantly lower than when k=2. Moreover, there is a good split using average linkage for k = 3 as seen in *Figure 9,* where there is a roughly equal proportion of observations in each cluster and there is not much overlapping among the clusters as well.

Agglomerative coefficient and classification error rate from optimal number of clusters

Linkage methods	Single	Complete	Average
Agglomerative coefficient	0.849	0.957	0.930

Table 4: Agglomerative coefficient for each of the linkage method

From *Table 4*, we can conclude that for the original iris dataset, the dendrogram using complete linkage as the dissimilarity measure produces the best clustering model since it yields the highest agglomerative coefficient of 0.957.

4.2 Analysis on the 90% winsorized data

As explained in Section 4, agglomerative hierarchical clustering using complete linkage and single linkage are sensitive to outliers. For example, in the case of complete linkage, a single item far from centroid can inflate the within sum of squares of the cluster drastically and affect the final clustering. As such, we will also perform hierarchical clustering on the 90% winsorized data to investigate how the presence of potential outliers identified by the boxplots can affect our results.

Dendrograms for Hierarchical Clustering on 90% winsorized data

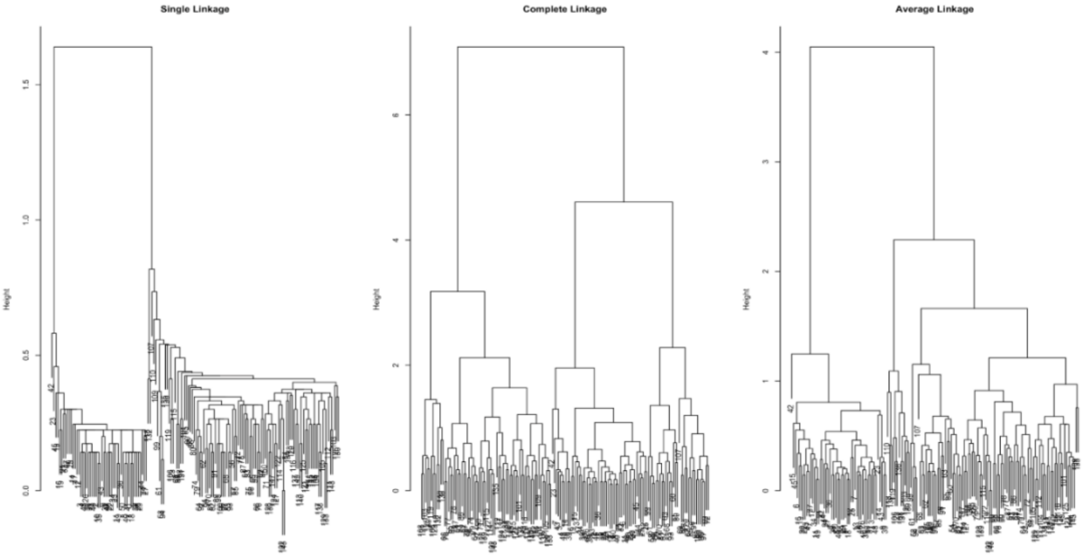


Figure 9: Dendrograms for single, complete and average linkage using 90% winsorized data

By analysing the dendrograms for single linkage in *Figure 5 and Figure 10*, we can see that the dendrograms are highly similar for all three methods with few differences. The dendrograms show that 2 main clusters are ideal in both the original and the winsorized data. For both the data, there seems to be two relatively large clusters within the rightmost cluster for single linkage, at height 0.4, which suggests that the rightmost cluster may be split into another 2 big clusters. This shows that while 2 clusters are the most obvious conclusion, there may be 3 main clusters within the data.

The dendrogram of 90% winsorized data using average linkage shows a clearer split of clusters as compared to the dendrogram of the original data. However, species number 42 still seems to stand out at the very end, being the 1-member cluster that merged with another cluster late in the clustering process which might indicate that the winsorized data might not affect the final number of clusters chosen

Determining optimal number of clusters using within cluster sum of squares

Similar to the analysis of the original data, we will also calculate the within cluster sum of squares to determine the optimal number of clusters for each of the linkage methods.

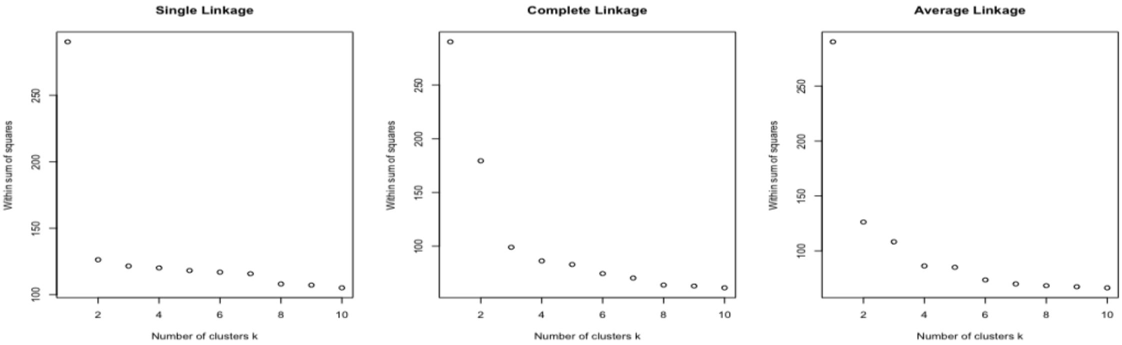


Figure 10: Plot of within sum of squares against no. of clusters for single, complete, average linkage

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

Table 5: Within sum of squares values for k clusters of single, complete, average linkage

The scree plot for single linkage method in *Figure 11* shows a significant decrease in within cluster sum of squares from k=1 to k=2. From *Table 5*, the within cluster sum of squares decreases by more than 50%. However, as k increases from 2 to 10, the decrease becomes very minimal - less than 5%. This indicates that the optimal number of clusters for single linkage method is 2.

The scree plot for complete linkage method in *Figure 11* shows a significant decrease in within cluster sum of squares from k=1 to k=2 and k=2 to k=3. From *Table 5*, within cluster sum of squares decreases by 38% and 44% respectively. However, as k increases from 3 to 10, the decrease becomes less significant. This may indicate that an optimal number of clusters for complete linkage method is 3.

The scree plot for average linkage method in *Figure 11* is slightly different from that in *Figure 6*. From *Figure 11* and *Table 5*, as k increases from 3 to 4, within cluster sum of squares still decreases significantly by 20%. The decrease only becomes less significant from after k=4. The elbow is most obvious at k=2 but only after k=4 does the total within sum of squared error decrease less significantly and somewhat stabilize. This may indicate that the optimal number of clusters for average linkage method can be either 2 or 4, different from the original dataset.

Principal Component Analysis (PCA) to visualize the clusters formed

Similar to our previous analysis on the original iris data, we will now perform PCA on the 90% winsorized data.

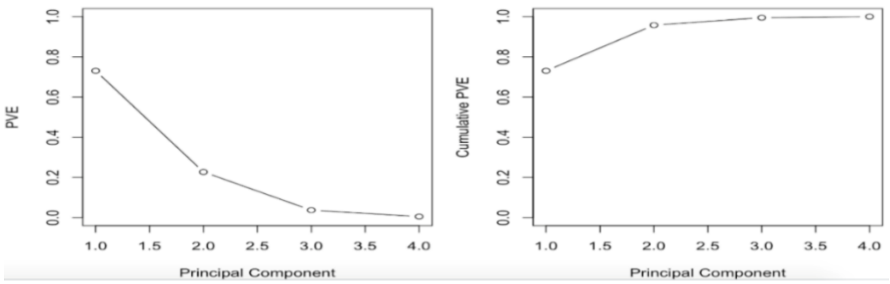


Figure 11: Scree plot

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

Table 6: Proportion of variance explained by each principal component

The scree plot in *Figure 12* shows that the first 2 principal components explain a significantly higher proportion of variance compared to the 3rd and 4th principal components. The 1st and 2nd principal components combined explain about 96% of the total sample variance, whereas the 3rd and 4th principal components combined explain less than 5% of the sample variance. This implies that the contribution to the total variance by the 3rd and 4th principal components are negligible, and they are probably redundant noise components. Hence, the first 2 principal components can effectively summarize the data and replace the 4 variables without losing much information.

Using the first 2 principal components, we now plot the k= 2, 3, 4 clusters for each linkage method:

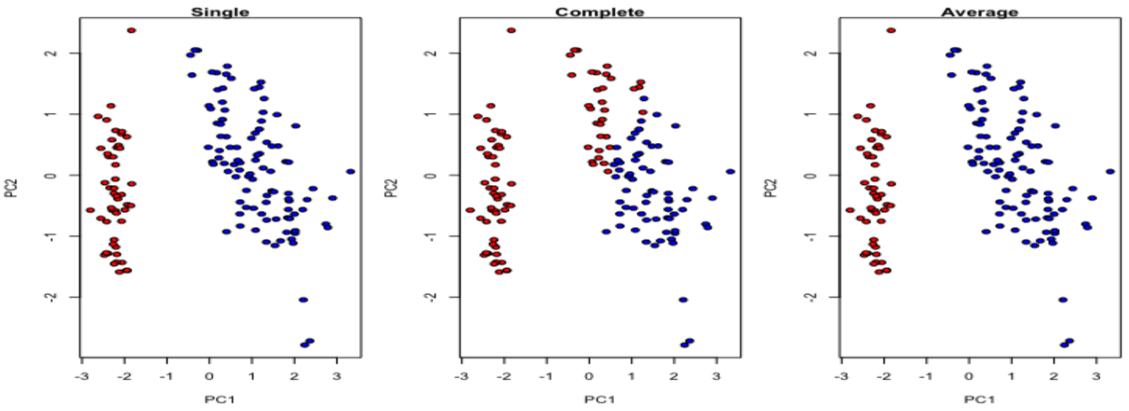


Figure 12: k=2 clusters formed using 2 principal components

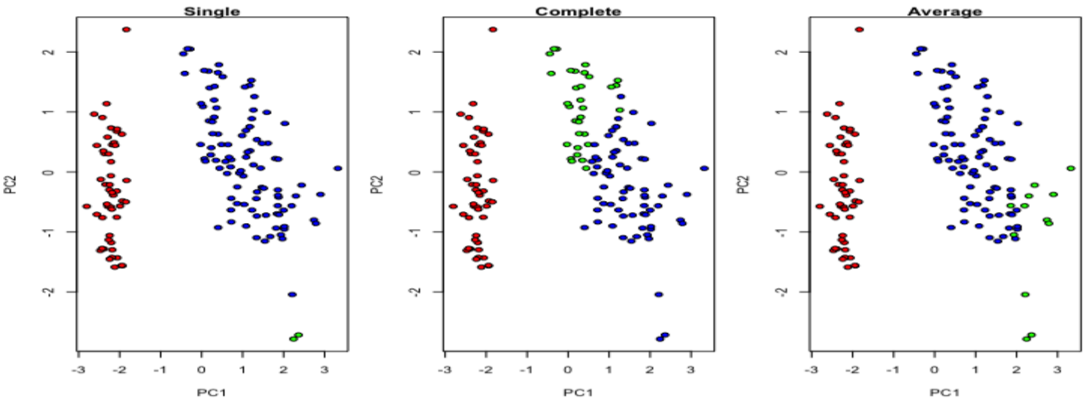


Figure 13: k=3 clusters formed using 2 principal components

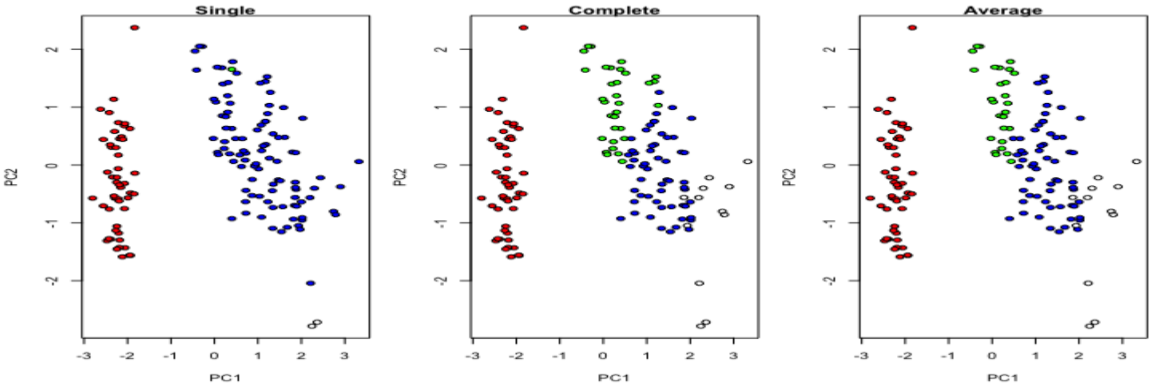


Figure 14: k=4 clusters formed using 2 principal components

From the plot in *Figure 13 and 14* for single linkage method, we can see that there is an obvious separation between both k=2 and k=3 clusters. However, for k=3 clusters, there are only 2 observations in the green cluster. As such, we conclude that k=2 cluster is the optimal number of clusters for the single linkage method.

From the plot in *Figure 13* for complete linkage method, though there is no obvious separation between 2 clusters for k =2, there is still a clear separation between them in the sense that there is no overlap between the 2 clusters. However, there is a group of points in the red cluster which are closer to the blue cluster instead. On the other hand, the plot for complete linkage in *Figure 14* shows a more obvious separation between the red and blue clusters. Although there is no obvious separation between the green and blue clusters, there is very little overlap between these 2 clusters. Comparing the plots in *Figure 14* and *Figure 15* for complete linkage method, the clusters are more well separated in k=3 as there is greater overlap between the clusters in *Figure 15*. Therefore, we conclude that k=3 cluster is the optimal number for complete linkage method.

As explained in the previous section, k=2, k=3 and k=4 are viable options for the final number of clusters using average linkage. From the plot in *Figure 14* for average linkage, the proportion of observations in the green cluster is significantly lesser. From the plot in Figure 15 for average linkage, the red cluster has more observations than the remaining 3 clusters.

Moreover, as total within sum of squares is significantly smaller than both within sum of squares for k=2 and k=3, we conclude that k=4 cluster is the optimal number for average linkage method.

Agglomerative coefficient and classification error rate from optimal number of clusters
Similar to the analysis on the original data, we will compute the agglomerative coefficient and misclassification rate using optimal number of clusters determined in the previous section for each of the linkage methods to determine the performance of each linkage method on the data:

Linkage methods	Single	Complete	Average
Agglomerative coefficient	0.855	0.960	0.933

Table 7: Agglomerative coefficients for each of the linkage methods

From *Table 7*, hierarchical clustering using complete linkage produces the highest agglomerative coefficient and hence complete linkage would be the best fit for the 90% winsorized iris dataset. Looking at the scatterplot of the dataset on the principal components, although there is still a presence of outliers, it is not as obvious as in the original data. Agglomerative coefficient increased by 0.960 - 0.957= 0.003, which is insignificant. This conclusion is still the same regardless of whether the dataset is winsorized and hence we conclude that winsorization does not change the model significantly.

However, it is good to note that after winsorization, hierarchical clustering clusters the dataset differently. As the observations are now closer to each other in general, agglomerative coefficient will naturally improve, indicating that winsorization might actually yield better clusters and help to put potential outliers into a cluster.

5. Non-Hierarchical Clustering: K- means clustering

We will now perform K-means clustering which is a form of non-hierarchical clustering so that we can compare the performance between hierarchical clustering and non-hierarchical clustering. Similar to non-hierarchical clustering, we will perform K-means on both the original and winsorized data to investigate if there are any significant differences between the clusters found.

5.1 Analysis on original Iris data

Like hierarchical clustering methods, K-means method aims to minimise the total within-cluster variation. However, one major difference is K-means method requires the number of clusters K, to be specified before the method can be performed. This may pose a problem since we do not know the true number of clusters in the data.

In order to determine the appropriate number of clusters K, we will perform K-means method from K=1 to K=10 and investigate if the increase in K will lead to a significant decrease in total within sum of squares.

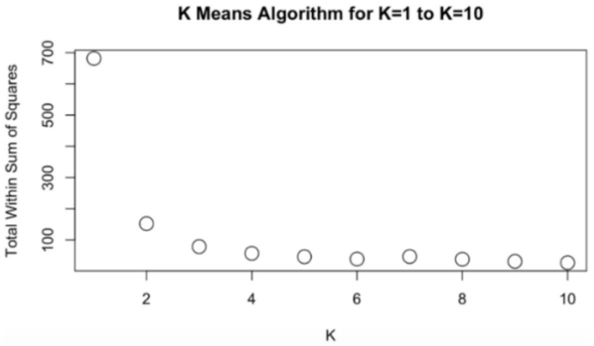


Figure 15: Plot of within sum of squares against no. of cluster

k	1	2	3	4	5	6	7	8	9	10
K-means	681.4	152.3	78.9	57.2	46.5	39.0	47.1	38.3	38.4	27.3

Table 8: Within sum of squares values for k clusters

From the *Figure 16 and Table 8*, we can see that there is a significantly huge decrease in total within sum of squares of more than 70% from 681 to 152 as K increases from 1 to 2. This suggests that there is some sort of clustering within the data.

Furthermore, as K increases from 2 to 3, the total within sum of squares decreased by about 50% from 152 to 78. As K increases from 4 to 149, there isn't any significantly decrease. Therefore, these hint that K=2, K=3 are the appropriate number of clusters to consider.

Similar to hierarchical clustering, we will plot our clusters on a 2-dimensional plot using PCA to visualize them.

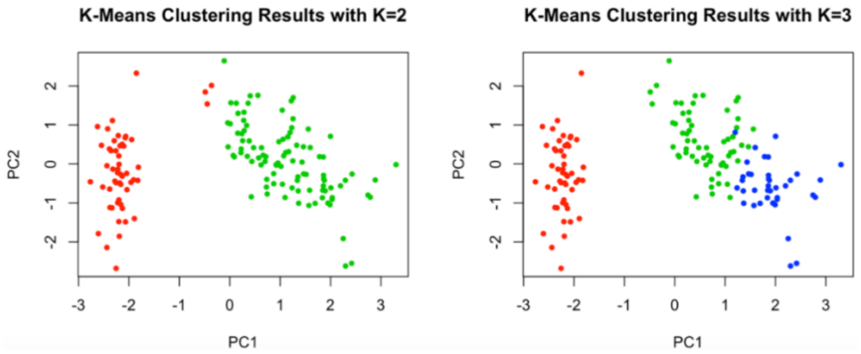


Figure 16: 2-dimensional plots using PCA for K Means Clustering (k=2 and k=3)

In the plot for K=2, there seem to be evidence of outliers in the red cluster. There are 3 points that are unusually far from the rest of the observation in the red cluster and are located much closer to the green cluster. However, there is still a clear separation between the red and green cluster as there is no overlap between the 2 clusters.

On the other hand, in the plot for K=3, we can see that the red and green clusters are more well-separated compared to that in K=2. There is an obvious separation between the red and green clusters now. Even though there is no obvious separation between the green and blue clusters, the green and blue clusters are still well-separated in the sense that there is very little overlap between the green and blue clusters. The green cluster is mainly concentrated around the top-right side while the blue cluster is mainly concentrated around the bottom-left side.

Therefore, we conclude that K=3 is the appropriate number of clusters in the original Iris data.

5.2 Analysis on the 90% winsorized data

We will repeat our analysis on the winsorized data.

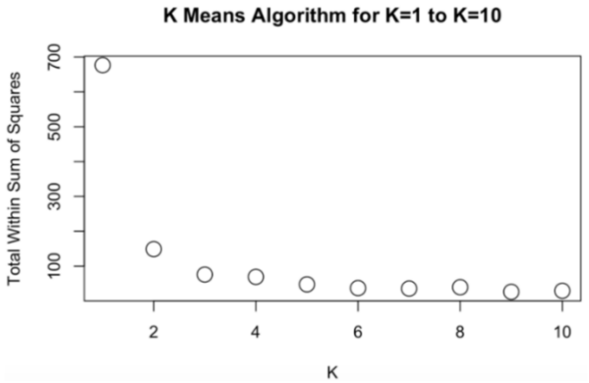


Figure 17: Plot of total within sum of squares

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

Table 9: Within sum of squares for k-clusters for 90% winsorized data

The plot in *Figure 18* is similar to the plot of the K-means on the original data, where there is significant decrease in total within sum of squares at K=2 and K=3. Hence, these also hint that K=2 and K=3 are the appropriate number of clusters to consider.

However, from the values of the total within sum of squares, we can see that the total within sum of squares are generally slightly lower than that when we perform K-means on the original data.

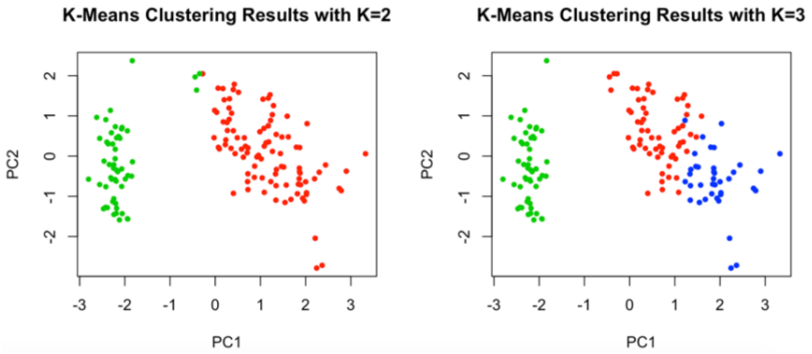


Figure 18: 2-dimensional plots using PCA for K Means Clustering (k=2 and k=3) for 90% winsorized data

From *Figure 19*, the K=2 plot looks different to the plot in K-means on the original data. There is now no obvious separation between the two clusters in the K=2 plot. This may be due to the effect of transforming the potential outliers, bringing the 2 clusters closer to one another. However, the K=3 plot look very similar to the plot in K-means on the original data.

Comparing the K=2 and K=3 plot, it seems that the clusters in K=3 are more well separated. There is an obvious separation between the green and red clusters in K=3. Furthermore, there is very little overlap between the red and blue clusters. The red clusters are concentrated in the top-left side, while the blue cluster is concentrated in the bottom-right side.

Therefore, similar to the K-means on original data, we will also conclude that K=3 is the appropriate number of clusters in the winsorized iris data.

6 Discussion and comparison of results

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

Table 10: Agglomerative Coefficients for Hierarchical clustering on Original and Winsorized data

In our analysis, we mentioned that hierarchical clustering using single and complete linkage are sensitive outliers. However, in *Table 10* we can see that the potential outliers did not affect our cluster significantly as there is no significant improve in the agglomerative coefficient for the winsorized data. Therefore, there is little evidence that winsorization is needed.

From *Table 10*, we can see that complete linkage has the highest agglomerative coefficient for both original and winsorized data. This means that the strength of the clusters using complete linkage is the strongest and we conclude that using complete linkage will give us the strongest clusters among the different types of hierarchical clustering.

In order to compare the strength of the clusters between hierarchical and non-hierarchical clustering, we will be comparing their total within sum of squares.

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

Table 11: Total Within Sum of Squares for the respective methods using K=3

Table 11 shows that K-means clustering has a lower total within sum of squares for both the original and winsorized data. This means that the clusters formed in K-means have lesser variation within each cluster and the clusters formed are stronger than those from hierarchical clustering using complete linkage. Therefore, we conclude that K-means clustering will give us a much stronger clusters compared to hierarchical clustering.

However, having the strongest cluster does not indicate that the model has the best classification accuracy. Since we have previously determined the optimal number of clusters for each method, we can further evaluate the methods by comparing their respective misclassification rates.

We assumed that the species in which majority of the observations from a specific cluster is being classified into, is the correct species for that cluster.

Original data

Cluster	Setosa	Versicolor	Virginica
1	50	0	0
2	0	50	50

Table 12: Misclassification table for single linkage method (k=2)

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)

Cluster	Setosa	Versicolor	Virginica
1	50	0	0
2	0	50	14
3	0	0	36

Table 14: Misclassification table for average linkage method (k=3)

Cluster	Setosa	Versicolor	Virginica
1	50	0	0
2	0	2	36
3	0	48	14

Table 15: Misclassification table for K Means Clustering method (k=3)

Winsorised data

Cluster	Setosa	Versicolor	Virginica
1	50	0	0
2	0	50	50

Table 16: Misclassification table for single linkage method (k=2)

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)

Cluster	Setosa	Versicolor	Virginica
1	50	0	0
2	0	24	37
3	0	26	1
4	0	0	12

Table 18: Misclassification table for average linkage method (k=4)

Cluster	Setosa	Versicolor	Virginica
1	50	0	0
2	0	2	36
3	0	48	14

Table 19: Misclassification table for K Means Clustering method (k=3)

Using *Table 12* to *Table 19*, error rates obtained for each method are summarised in *Table 20*.

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

Table 20: Misclassification rates for all methods

From *Table 20*, unlike the strength of clusters, performing winsorization does not necessarily improve the classification accuracy. There only improvement in error rate for complete linkage, a small decrease from 16% to 14.7%. On the other hand, there is a drastic increase in error rate for average linkage from 9.3% to 24.7%. This may hint that complete linkage may perform better on a data without outliers while average linkage performs better on a data with some outliers.

On the other hand, this may also mean that these “outliers” may just have rare distinctive features of a species. They still belong to one of the related species of Iris flowers. As such, one should not simply eliminate such observations when performing the cluster analysis. Therefore, transforming them through winsorization may not be appropriate and will lead to a higher error rate.

Table 20 shows that hierarchical clustering using average linkage have the best accuracy for the original data while k-means clustering has the best accuracy for winsorized data.

However, it must be highlighted that K-means clustering is much more consistent compared to average linkage. Even though average linkage has the lowest error rate of 9.3% for the original data, the error rate for K-means clustering for the original data is 10.7%, which is only slightly higher. On the other hand, for the winsorized data, average linkage has a high error rate of 24.7% while K-means clustering retains its error rate of 10.7%. Therefore, we conclude that K-means clustering is the best model for classification since it has a low and consistent misclassification rate.

7 Conclusion

The objective of the research was to find appropriate clusters for the classic statistical dataset of Iris flowers. Various grouping approaches have been proposed in the prior literature, but procedures presented by Ward (1963), Johnson (1967) and MacQueen (1967) was found most influential based on the number of citations and significant recognition of the followed multivariate research. Thus, three agglomerative hierarchical methods: single linkage, average linkage and complete linkage, as well as one non-hierarchical clustering procedure: K-Means, were applied.

Based our analysis, we found that majority of our clustering methods concluded that the optimal number of clusters is 3. As such, by comparing the strength and accuracy of the clusters obtained, we conclude that K-means clustering is the best model. K-means clustering will give us the strongest clusters while having a low misclassification rate. Future analysis can be done to investigate the occurrences of the outliers identified by the boxplots, appropriateness of winsorization and explore other ways to handle the outliers, which may lead to differing conclusions.

8 References

Johnson, Stephen C. "Hierarchical clustering schemes." *Psychometrika* 32.3 (1967): 241-254.

MacQueen, James. "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. No. 14. 1967.

Stephenson, William. "The inverted factor technique." *British Journal of Psychology* 26.4 (1936): 344.

Ward, Joe H. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American Statistical Association*, vol. 58, no. 301, 1963, pp. 236–244. *JSTOR*, www.jstor.org/stable/2282967.