ST3240 Project Group 3

Done By:
1. Guan Soon Quan
2. N Selva Kumar
3. Ng Wei Lun
4. Png Wei Song Jonathan
5. Richard Wong Ho Chuan
6. Wong Tau Yew, Mark

Introduction

This report will cover the accuracy of clustering methods applied on Fisher's Iris data set (1936). We will first identify the most appropriate number of clusters, after which we will identify if there is an overlying significant method that outperforms other clustering techniques. Various clustering techniques and a range of K values will then be used to choose the “best” clustering method as our final model.

Summary of Data

	Sepal Length	Sepal Width	Petal Length	Petal Width
Minimum	4.300	2.000	1.000	0.100
1st Quantile	5.100	2.800	1.600	0.300
Median	5.800	3.000	4.350	1.300
Mean	5.843	3.057	3.758	1.199
3rd Quantile	6.400	3.300	5.100	1.800
Maximum	7.900	4.400	6.900	2.500

Table 1.1

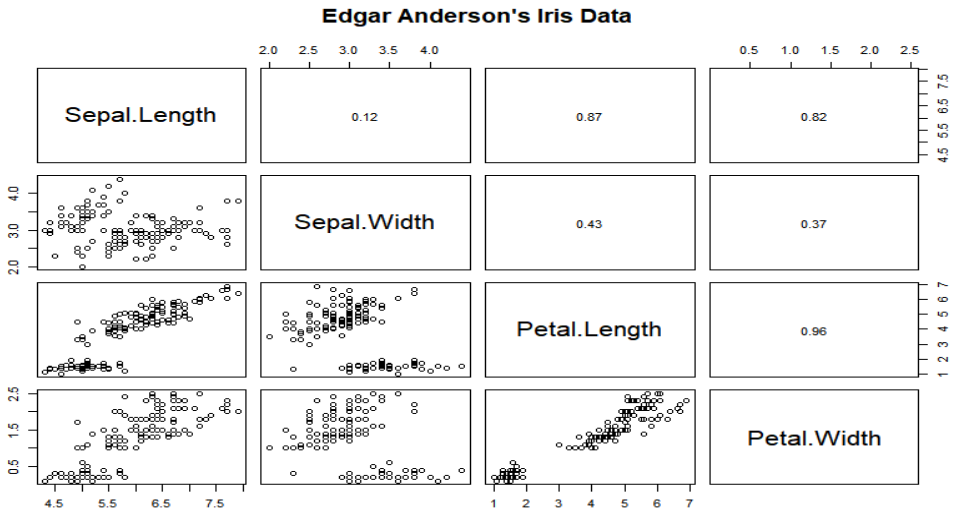

Figure 1.1

The summary table (*Table 1.1*) shows that data values for petal width is much lower than the other variables. Hence, we will assume that each variable has an equal effect in determining the final model and the data points will be scaled. This is especially important for models that are depending on distance as an indicator of dissimilarity.

The correlation plot (*Figure 1.1*) shows that there seems to be a correlation between variables. Petal length and petal width, sepal length and petal length, sepal length and petal

width, are strongly positively correlated with values of 0.96, 0.87, 0.82 respectively. Hence, it may hint for the use of a simpler model through the use of variable selection.

Distribution

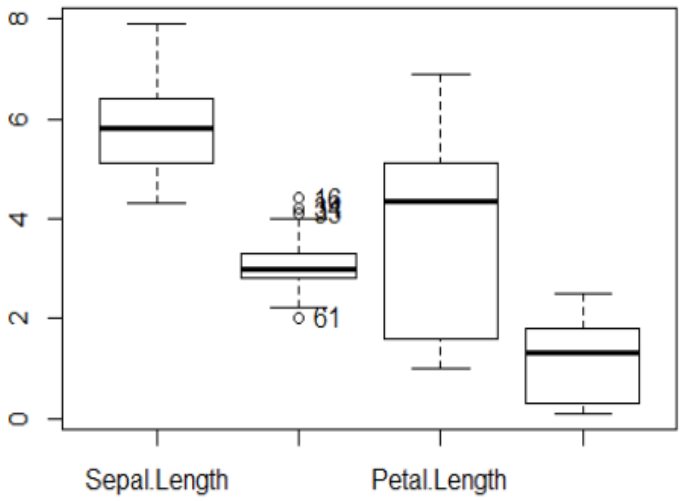

Figure 2.1

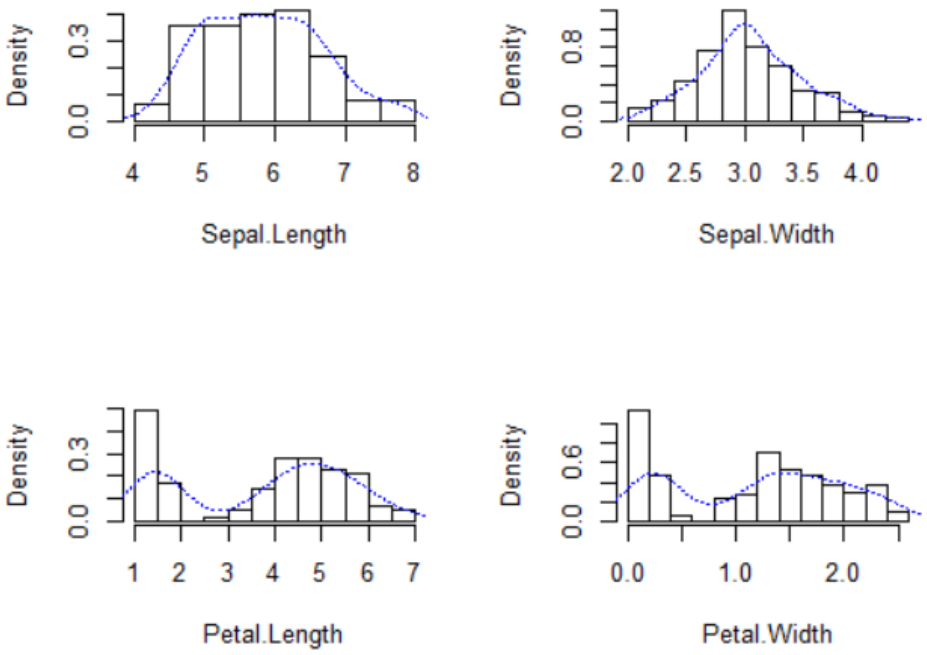

Figure 2.2

The sepal width boxplot (*Figure 2.1)* shows that there seems to be 4 outliers present. These points will not be removed when we determine our initial model, they will be re-examined after finding the best possible number of clusters . To determine if they are influential points, we will examine the model with and without these points.

From the histogram plots (*Figure 2.2)*, we notice that sepal length and sepal width are unimodal distributions whereas petal length and petal width are bimodal distributions. The distribution has little effect in determining our model as we do not need multivariate normality

assumption for most of the clustering methods used. The assumption of multivariate normality will be evaluated when our clustering methods require this assumption.

Outliers:
16, 33, 34, 61

Variable Selection (PCA)

For the iris dataset, there are 4 variables. Hence, principal component analysis is done for dimensionality reduction.

	PC1	PC2	PC3	PC4
Standard Deviation	1.7084	0.9560	0.38309	0.14393
Cumulative Proportion	0.7296	0.9581	0.9984	1.000
Sepal Length	0.5211	-0.3774	0.7196	0.2613
Sepal Width	-0.2693	-0.9233	-0.2444	-0.1235
Petal Length	0.5804	-0.02449	-0.1421	-0.8014
Petal Width	0.5649	-0.066942	-0.6343	0.5236

Table 3.1

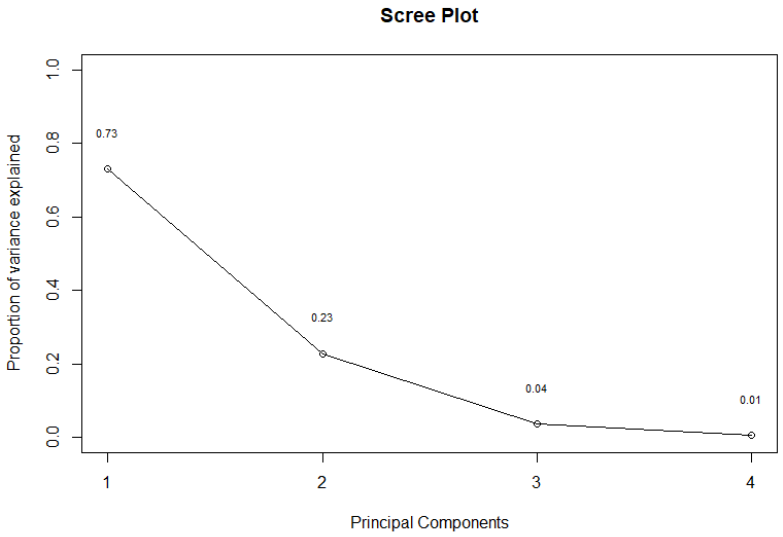

Figure 3.1

From the scree plot (*Figure 3.1*), the elbow is roughly at p=2, which means that the dimension of the variables can be reduced from 4 to 2 while retaining roughly 96% of the variance explained. For the first principal component, the magnitude of each variable is almost equal to one another except petal width, which has a negative and a lower magnitude. For the second principal component, the magnitude of sepal width dominates the other variables, where petal length and petal width has a smaller magnitude compared to sepal length and sepal width.

Used metrics

We will be using 2 forms of measure metric to evaluate performance of a clustering method.

1. Between Clusters Sum of Squares (BSS)
   Measures the inter-cluster distance by the sum of squared distances between the centres of clusters and the grand mean. The higher the metric, the better the clustering since it implies greater separation of clusters.
   Since Total Sum of Squares is a sum of Within Clusters Sum of Squares (WSS) and the Between Cluster Sum of Squares, we only need to pick one from the two since the two metrics measurements are related.

2. Average silhouette width[1]
   Measures how well an observation is clustered and how close each point in one cluster is to other points in neighbouring clusters.

   Denote A be the cluster an observation i belongs to.
   d(i,j) is the dissimilarity between observation i and observation j.

   $$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i,j)$$
   $$= \text{average dissimilarity of i to all other objects of A}$$

   Denote C be another cluster i.e. i does not belong to C

   $$d(i,C) = \frac{1}{|C|} \sum_{j \in C} d(i,j)$$
   $$= \text{average dissimilarity of i to all other objects of C}$$

   $$b(i) = min\, d(i,C)$$

   Then the silhouette width of observation i is defined as:

   $$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

   $$\text{It can be proven that } -1 \leq s(i) \leq (1)$$

   When s(i) is close to 1, a(i) is small relative to b(i) => object i is well classified
   When s(i) is close to -1, a(i) is large relative to b(i) => object i is badly classified
   Average silhouette width is then the total silhouette width of all observations divided by number of observations.

[1] Cluster Validation Statistics: Must Know Methods. (2018, October 22). Retrieved October 31, 2020, from https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/

Hierarchical clustering

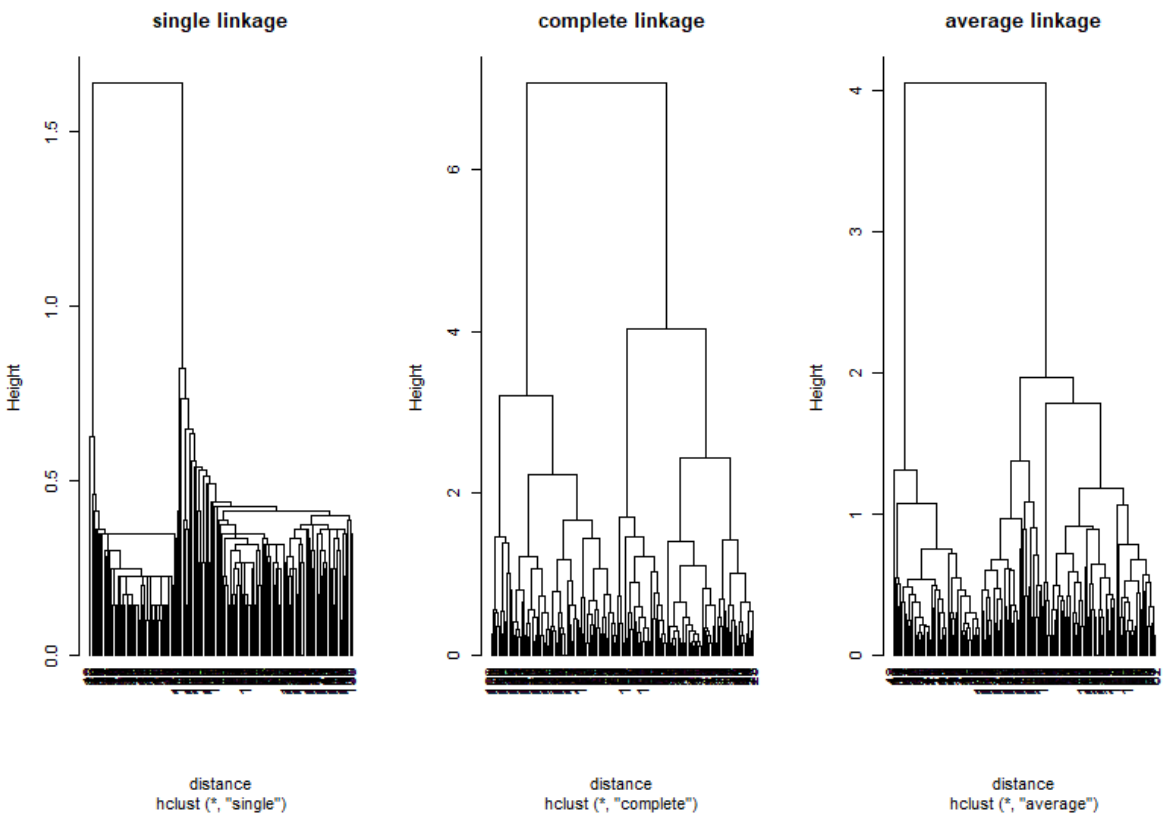

Figure 4.1

Since the vertical lines/ height represent the distance or dissimilarity between clusters, all 3 dendrograms have at least 2 clusters that are far from one another (clusters merge together only at a greater height). However, in complete and average linkages we can observe a possibility of more than 2 distinct clusters, as the distance between smaller clusters is small.

Finding K

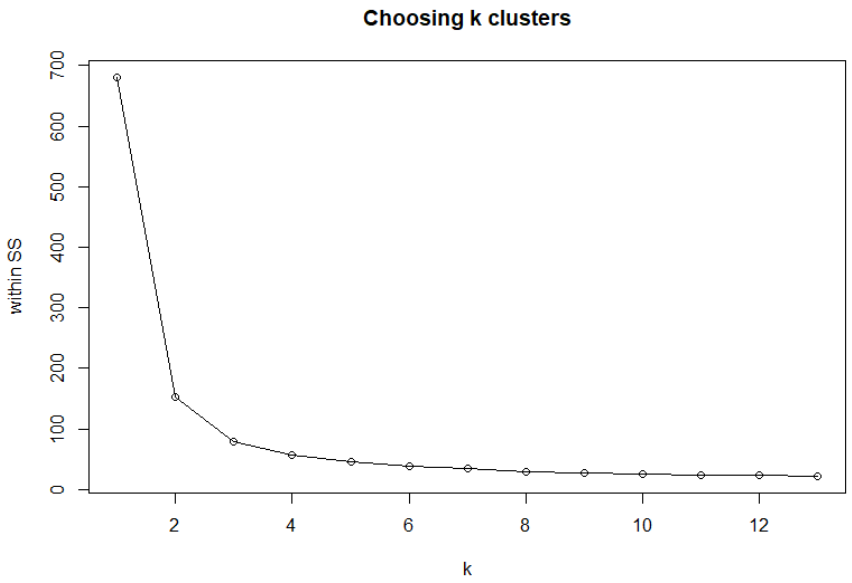

Figure 5.1

'Elbow' method was used to determine the appropriate number of K clusters for K-means clustering. As seen from the graph (*Figure 5.1*), K = 2 or K = 3 clusters might be appropriate for the clustering method. Since this is subjective, gap statistics will be used to help us determine the most appropriate K value.

Gap statistics[2]

Gap statistics is a technique used for approximating the "correct" number of K clusters. This is achieved by assessing the metric, Within cluster Sum of Squares(WSS) with regards to different K values. This technique enables us to obtain a 'gap' value and the maximum gap value applied with the '1 standard error rule' (choosing the optimal K value over a range of gap value of ± 1se because gap statistics has a random nature) would indicate to us the most appropriate K.
The optimal number of clusters is determined to be 3 since it has the highest frequency.

Optimal no. of clusters	3	4	5
Frequency	83	1	16

Table 5.1

[2] Counting Clusters. (n.d.). Retrieved October 22, 2020, from http://blog.echen.me/2011/03/14/counting-clusters/

K	1	2	3	4	5
K-means	0	529.0226	602.5192	624.1421	634.9244
Single linkage	0	526.4236	538.8912	540.9472	561.1172
Complete linkage	0	446.2175	591.8456	620.3976	626.5607
Average linkage	0	526.4236	601.9252	612.7443	624.5518

Table 5.2

The metric used here is the BSS. When the number of clusters is 0, the BSS is also 0 since there is only 1 centroid which is equivalent to the grand mean.

For k=2 to 5, the k-means model has the highest BSS indicating it yields the “best” clustering compared to the other hierarchical clustering methods.

Clustering Types and Methods

1. **Partitioning** algorithms

A medoid (point where dissimilarity to all observations in the cluster is the minimal)/centroid (mean of all points in a cluster) is constructed to represent each cluster. These methods seek to derive the optimal set of medoids/centroids based on certain dissimilarity criterions.

Algorithms covered in the module include: **K-means**
Additional algorithms experimented: **Spectral Clustering, Clara, Fanny/ Fuzzy analysis**

Spectral clustering[3]
Observations are represented by eigenvectors of the Laplacian graph that can be computed using the connections generated by the k-nearest neighbors (KNN) algorithm. Subsequently applying K-means algorithm to cluster these eigenvectors (except the constant eigenvector) allows for better clustering of non-convex shapes (e.g. clusters are in a ring-shaped, each cluster is in a smaller ring in one another).
Parameters: Number of nearest neighbours = 10, number of eigenvectors that can sufficiently represent the observations (up to number of observations-1) = 2, number of clusters = 3.

Clara[4]
The partition around medoids (PAM) algorithm is performed on a small sample of observations to obtain a medoid to represent each cluster. This is repeated with samples of the same size (sampling with replacement similar to bootstrapping) with the average

[3] Nura, K. (2018, February 24). Spectral Clustering. Retrieved October 31, 2020, from https://rpubs.com/nurakawa/spectral-clustering
[4] Alboukadel, K. (n.d.). CLARA in R : Clustering Large Applications. Retrieved October 31, 2020, from https://www.datanovia.com/en/lessons/clara-in-r-clustering-large-applications/

dissimilarity between every observation and its closest medoid calculated each round. The output of the method is the set of medoids that corresponds with the lowest dissimilarity which corresponds to the medoids with minimal sampling bias.
Parameters: Number of samples = 50, number of clusters = 3, size of each sample=min(number of observations, 40 + 2 * number of clusters)
Uses the R function “clara”.

Fanny/Fuzzy analysis[5] (soft clustering)
Each observation has a probability of belonging to each cluster. The centroid of a cluster is then the mean of all points in the cluster weighted by the probability of the points belonging to the cluster. Based on each centroid, the closer an observation is to the centroid, the higher the probability it belongs to that cluster.
Parameters: Number of clusters = 3
Uses the R function “fanny”.

2. **Hierarchical/Agglomerative** clustering

Clustering is done by building up a full clustering tree (each cluster contains a single observation)
Clustering trees are then pruned to give a smaller number of larger clusters.

Algorithms covered in module: **Single, Complete, Average linkages**
Additional algorithms experimented: **Ward**

Ward[6]
This is an agglomerative clustering method where we start with n clusters and slowly combine these clusters to form 1 final cluster by combining clusters that give us the smallest increase in WSS.
Uses the R package “FactoClass” for the “ward” function.

3. **Density based** clustering

Clusters are formed based on some density-based function/criterion.
Advantages of these methods include being able to determine outliers as well as dynamic determination of the number of clusters.

Affinity propagation (AP)[7]
A “responsibility” matrix R indicates the suitability of a point to act as an exemplar (or medoid) for another point relative to other possible exemplars. A “availability” matrix A indicates the appropriateness of a point to act as an exemplar for another point taking into account the preferences of all other points for the exemplar. During the AP algorithm,

---

[5] Kaufman, L. and Rousseeuw, P.J. (2008). Fuzzy Analysis (Program FANNY). In Finding Groups in Data (eds L. Kaufman and P.J. Rousseeuw). doi:10.1002/9780470316801.ch4
[6] 14.7 - Ward's Method: STAT 505. (n.d.). Retrieved October 31, 2020, from https://online.stat.psu.edu/stat505/lesson/14/14.7
[7] AP_affinity_propagation: Affinity propagation clustering. (2020, July 02). Retrieved October 31, 2020, from https://rdrr.io/cran/ClusterR/man/AP_affinity_propagation.html

messages are passed between the R and A matrices and these are updated iteratively until convergence or when a predefined no. of iterations is reached.

Density-based spatial clustering of applications with noise (**DBSCAN**)[8]
The algorithm clusters points based on density. Density is defined to be how "reachable" a point is from the other; if it is within a set distance of a point. It works by repeatedly finding points within a set distance from the point, then finding other neighboring points in these new points. We will only be required to choose how close 2 points are and the number of points required to form a cluster using this clustering method.
Parameters: epsilon (distance) = 0.9, minimum points = 5
Uses the R package "dbscan" for the "dbscan" function.

4. **Model based** clustering

Assumptions made on distributions of the clusters (number of clusters specified before clustering)
Based on these assumptions, the Expectation Maximisation (EM) algorithm is used to estimate the parameters of these distributions by maximising the likelihood fit of the observations.

Gaussian Mixture[9]
The observations are assumed to be derived from a mixture of gaussian distributions. The means, variances and the probability of each distribution in the mixture is estimated by the EM algorithm.
Parameters: number of gaussian distributions/clusters=3
Uses the R package "mclust" for the "Mclust" function.

Checking normality assumption
Q-Qplots are constructed between the quantiles of each variable and the normal quantiles based on the clusters derived from the 4 variable best subset.

And the corresponding correlation coefficients are as such:

	1(50 observations)	2(45 observations)	3(55 observations)
Sepal length	0.99048	0.98811	0.98350
Sepal width	0.98304	0.98725	0.98460
Petal length	0.97544	0.97495	0.98012
Petal width	0.89285	0.96418	0.98275

Table 6.1

[8]Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. Article 19 (August 2017).
[9] Model-based clustering and Gaussian mixture model in R. (n.d.). Retrieved October 31, 2020, from https://en.proft.me/2017/02/1/model-based-clustering-r/

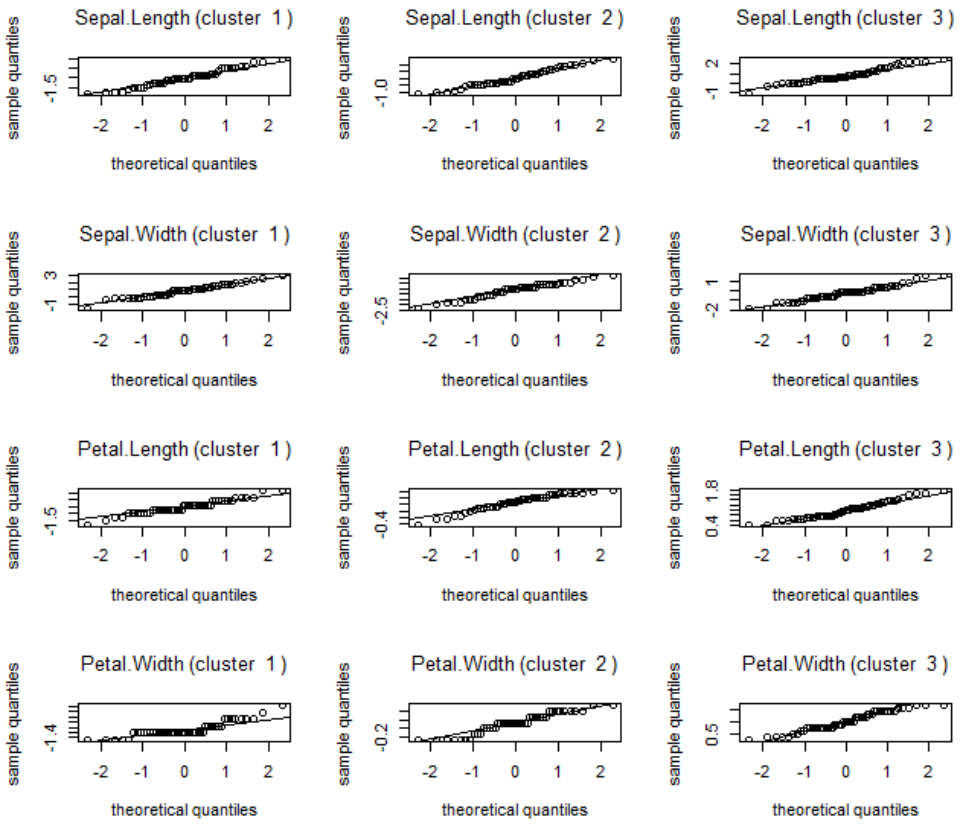

Figure 6.1

Comparing Q-Q plot correlation test for normality (lecture 2 page 10), at significance level α = 0.01, normality is rejected only for petal width in cluster 1 and 2. There is some evidence of normality present in every other cluster.

Methodology

There are many clustering methods and we chose some methods from the aforementioned to evaluate and illustrate possible differences between each technique.
For each clustering method, the best subset of variables is selected based on maximising for the 2 specified metrics. Since the number of variables p=4 is small, the best subset is easily computed for combinations of 1-4 variables.
For density based models, clusters are fixed at 3 (stopping condition) even though they can dynamically determine clusters. This is done so as to allow a fairer comparison between all models since a larger number of clusters will artificially improve a metric.

All metric values will be compared within the same metric and across all cluster methods and their best subset variable combinations.

Interpretation of results

1. Importance of variables
    - Importance of variables are determined by its inclusion/exclusion in the best subset selection for each method

Results

- For best subset selection of 1 variable, petal length is selected across all methods with the exception of DBSCAN which is unable to produce multiple clusters with a single variable & average and complete linkages which select petal width for its best subset.
- For best subset selection of 2 variables, petal length and petal width are selected across all methods.
- For best subset selection of 3 variables, petal length, petal width and sepal length are selected across all methods.

- Since 1 variable best subset $\subset$ 2 variables best subset $\subset$ 3 variables best subset, the importance of each variable is likely to be ordered as such: petal length>petal width>sepal length>sepal width.

2. Result Metric Output

Between Cluster Sum of Square	Number of Variables			
	1	2	3	4
Single linkage	44679.05	89346.48	133976.7	178586.3
Complete linkage	44690.78	89371.30	134027.1	178647.1
Average linkage	44691.53	89381.52	134014.9	178603.1
K-means	**44692.13**	**89382.09**	**134037.4**	**178661.1**
Affinity propagation	44691.88	**89382.09**	134036.9	178660.2
Spectral clustering	44678.15	89381.73	134027.8	178658.0
Ward	44692.02	89374.24	134035.5	178652.1
DBSCAN	44686.84	89347.65	133993.2	178580.7
Clara	44691.93	**89382.09**	134037.0	178659.6
Gaussian Mixture	**44692.13**	89381.41	134023.9	178632.3
Fanny	**44692.13**	**89382.09**	134037.3	178660.9

Table 7.1

Columns 1-4:
Column i contains the BSS of the i variables best subset selected by the various methods. Based solely on BSS, K-means achieves the “best” classification across all subsets of variables.

Average silhouette width	Number of Variables			
	1	2	3	4
Single linkage	0.5800009	0.5209724	**0.5434935**	0.5046456

Complete linkage	0.7077405	0.5548602	0.5014259	0.4496185
Average linkage	0.7246130	0.6654765	**0.5434935**	0.4802669
K-means	0.7268315	**0.6741313**	0.5386510	0.4599482
Affinity propagation	0.7246130	**0.6741313**	0.5375781	0.4590416
Spectral clustering	0.6529337	0.6685785	0.5252169	0.4604880
Ward	0.6872515	0.6099441	0.5387955	0.4466890
DBSCAN	**0.8050104**	0.5209724	**0.5434935**	**0.5216965**
Clara	0.7268315	**0.6741313**	0.5384269	0.4507912
Gaussian Mixture	0.7162689	0.6625331	0.4836671	0.3741649
Fanny	0.7246130	**0.6741313**	0.5390464	0.4566338

Table 7.2

Based solely on average silhouette width, DBSCAN achieves the “best” classification across all subsets of variables except for its 2 variable subset model. K-means, AP, Clara and Fanny perform the “best” with their 2 variables subset models.

Note: Gaussian mixture produces the best classification accuracy. However, this project is concerned about unsupervised clustering and to assume that each cluster is normally distributed requires backing (some evidence presented under the Gaussian mixture section) and there is no sure way of quantification when comparing with other methods. Hence, the “best” model can only be selected solely based on a metric.

Outlier Analysis

From the earlier boxplots, particularly from the Sepal Width boxplot, we have identified points 16,33,34,61 as outliers.

Between Cluster Sum of Square	Number of Variables			
	1	2	3	4
Single linkage	42319.54	84627.59	126912.0	169151.8
Complete linkage	42332.13	84651.28	126949.2	169226.1
Average linkage	42331.80	84661.84	126939.4	169169.8
K-means	**42332.28**	**84662.43**	**126960.6**	**169226.6**
Affinity propagation	42332.10	**84662.43**	126960.5	169225.6
Spectral clustering	42327.37	84656.09	126950.0	169225.0
Ward	42332.13	84654.93	126959.0	169213.3
DBSCAN	42326.74	84628.73	126913.9	169146.4

Clara	42332.11	**84662.43**	126960.5	169223.8
Gaussian Mixture	**42332.28**	84661.65	126947.5	169165.2
Fanny	42332.26	**84662.43**	126960.5	169225.8

Table 7.3

Omission of these points results in a reduction in the BSS values across all methods and variable combinations. K-means again achieves the “best” classification across all subsets of variables.

Average silhouette width	Number of Variables			
	1	2	3	4
Single linkage	0.5822832	0.5192317	**0.5485343**	0.5156993
Complete linkage	0.7236339	0.5432392	0.4965234	0.4606468
Average linkage	0.7236339	0.6612378	**0.5485343**	0.4903466
K-means	0.7236339	**0.6702200**	0.5408349	0.4670405
Affinity propagation	0.7236339	**0.6702200**	0.5419356	0.4605084
Spectral clustering	0.6576412	0.6552629	0.5356878	0.4660127
Ward	0.7054481	0.6069021	0.5419664	0.4294317
DBSCAN	**0.8455750**	0.5192317	**0.5485343**	**0.5287166**
Clara	0.7256054	0.6702200	0.5419356	0.4527938
Gaussian Mixture	0.7150281	0.6594695	0.4847686	0.3153903
Fanny	0.7236339	**0.6702200**	0.5408734	0.4575908

Table 7.4

Based solely on average silhouette width, DBSCAN again achieves the “best” classification across all subsets of variables except for its 2 variable subset model. K-means, AP, Clara and Fanny perform the “best” with their 2 variable subset models.

The omission of points 16,33,34,61 does not have a significant impact on changes to the efficacy of the methods.

Cluster Prediction

From the results and outlier analysis, since K-means with 2 variables performed consistently well across both metrics in comparison to the other clustering methods with or without outliers, we decided to use K-Means as our model with K=3 and using only the variables Petal.Width and Petal.Length.

Cluster 1	Cluster 2	Cluster 3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50	51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 72 73 74 75 76 77 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 107 120 134 135	71 78 101 102 103 104 105 106 108 109 110 111 112 113 114 115 116 117 118 119 121 122 123 124 125 126 127 128 129 130 131 132 133 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150

Table 8.1

The previously mentioned outliers (16, 33, 34, 61) are not misclassified according to the K-means clustering method used. 16, 33, 34 was classified with the other labels that are between 1 and 50, while 61 is classified with the other labels between 51 and 100.

On the contrary, the points 107, 120, 134, 135 were misclassified under cluster 2 (labels between 51 and 100), while labels 71, 78 were misclassified under cluster 3 (labels between 101 and 150). This is mainly due to the points being too close to the boundary between Clusters 2 and 3. This is illustrated in Fig. 8.2 below.

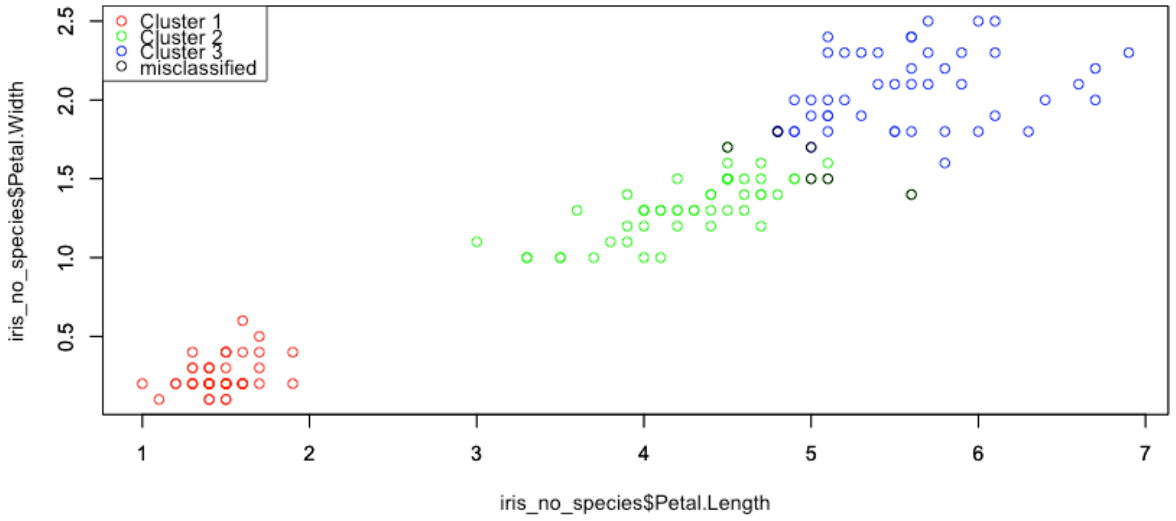

Fig. 8.2