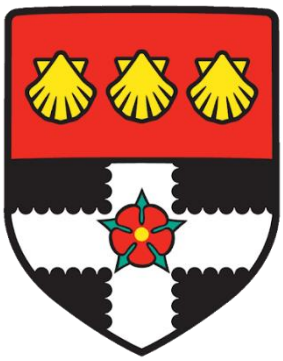


University of Reading

Department of Computer Science

Clustering Analysis on Wine Dataset

Shavin Croos

Module code: CS3DS19
Module title: Data Science Algorithms and Tools
Student Number: 27015244
Date of completion: 28/10/2021
Actual time spent on the assignment (hours): 10+ hours
Assignment evaluation (3 key points):
1. Sometimes the brief was a bit vague in some parts and could be slightly better written to make sure understanding is grasped better.
2. Gained an understanding of how nodes work and how the data can be manipulated.
3. Mostly a simple coursework to follow.

Task 1: Clustering without Normalisation

In the first task, clustering analysis of the wine dataset was carried out without normalisation of the data prior. For part 1 of the first task, the Principal Component Analysis (PCA) method was applied to the dataset to reduce it down into two-dimensional coordinates, where each data point of the dataset was assigned a colour according to their class. Red represented that the data point belonged to Class 1, purple represented that the data point belonged to Class 2 and blue represented that the data point belonged to Class 3. After being assigned the colours, the PCA values were plotted on a scatter graph (Figure 1) to show the distribution of the data points, where the x-axis represented PCA dimension 0 values, and the y-axis represented PCA dimension 1 values.

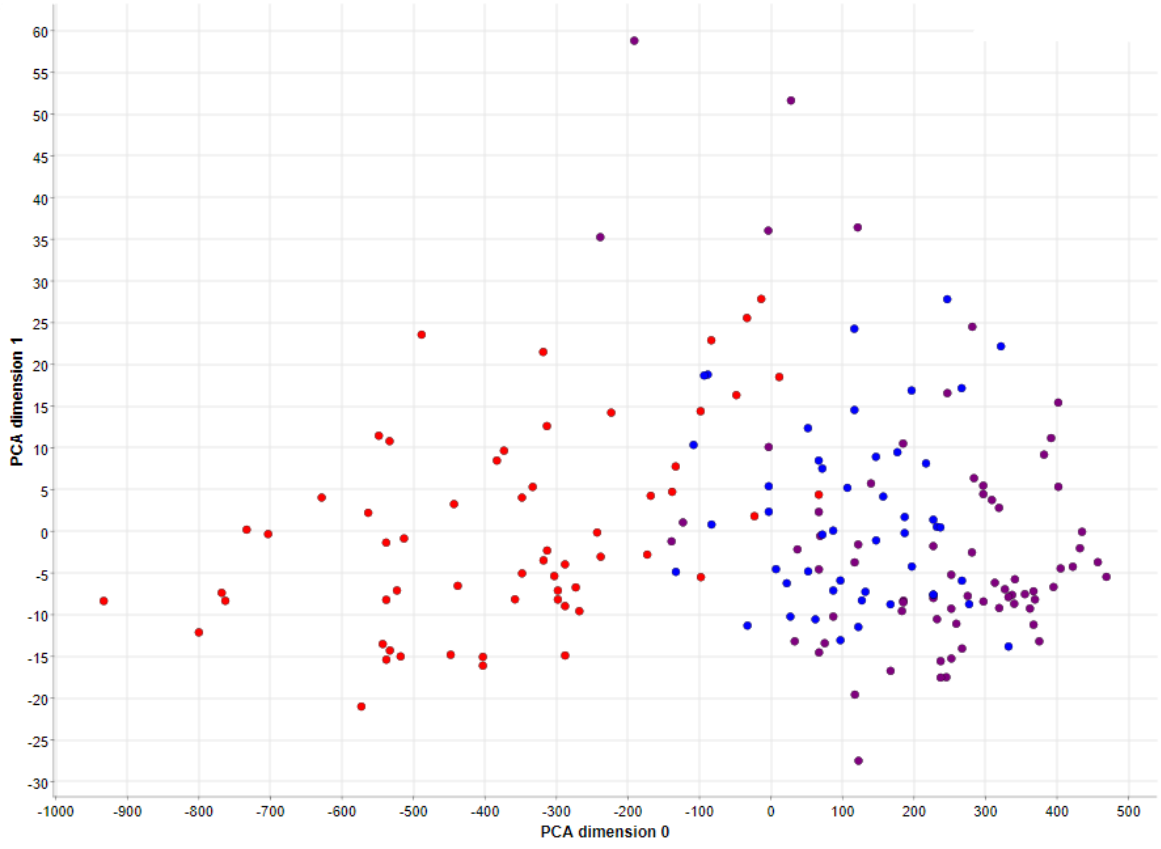


***Figure 1**. Scatter graph showing the distribution of the non-normalised PCA values of the wine dataset according to their class value.*

For part 2 of the first task, the K-Means algorithm was applied to the wine dataset to create 3 clusters (partitions). Once this was done, the PCA method was applied to all the 3 clusters to reduce them to two-dimensional coordinates. This time, each data point was assigned a colour according to their cluster ID. Green represented that the data point belonged to Cluster 0, orange represented that the data point belonged to Cluster 1 and brown represented that the data point belonged to Cluster 2. After being assigned the colours, the PCA values were plotted on another scatter graph (Figure 2) to show the distribution of these data points, where the x-axis represented PCA dimension 0 values, and the y-axis represented PCA dimension 1 values.

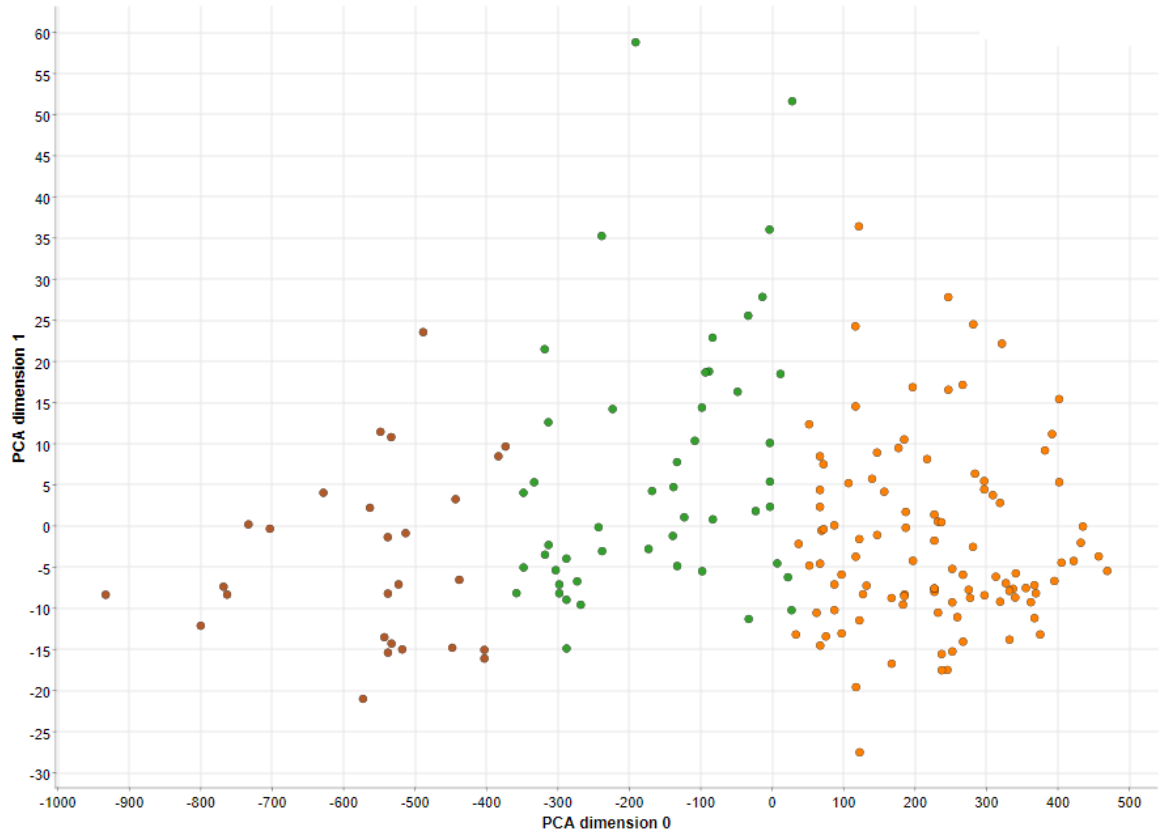


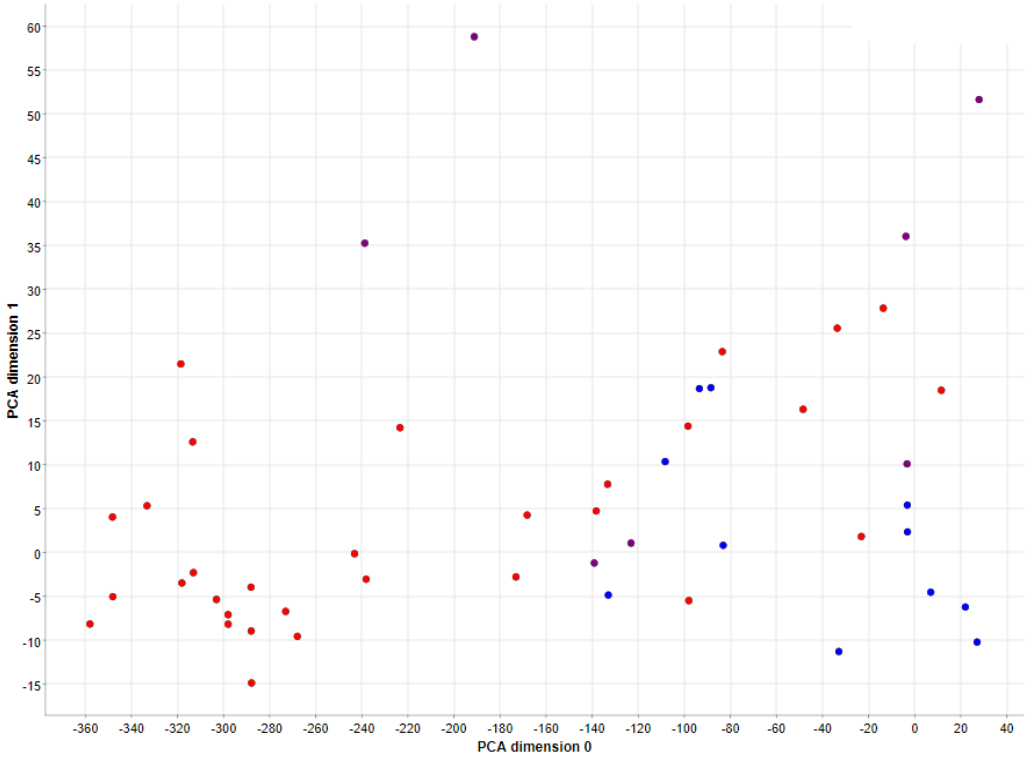
***Figure 2**. Scatter graph showing the distribution of the non-normalised PCA values of the wine dataset according to their cluster ID value.*

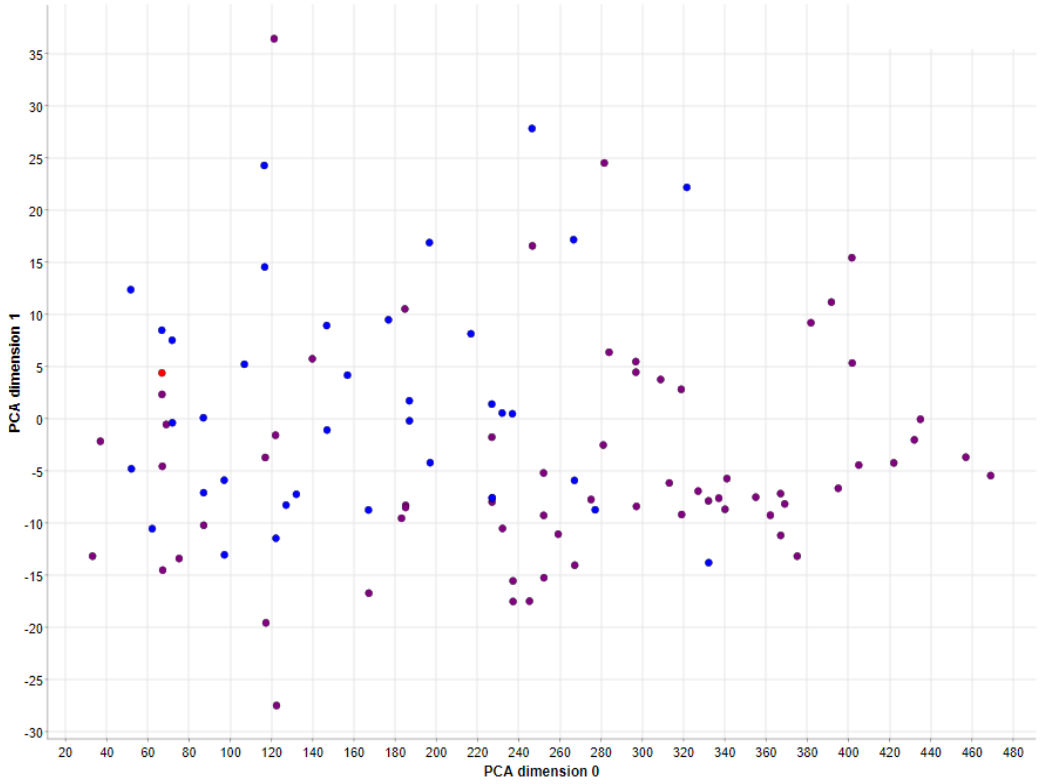
By observing the two scatter graphs generated, we can see that all the data points are in the same positions. Based off the colour coding used for the scatter graphs in figures 1 and 2, majority of the Class 1 data points fall into Clusters 0 and 2. Most these points also have PCA dimension 0 values that are less than 0 but have PCA dimension 1 values ranging from -21 to 57. It can also be seen that all the Class 3 data points straddle between Clusters 0 and 1. The PCA dimension 0 values of these points are greater than -150, but less than 350. On top of this, the PCA dimension 1 values of these points are greater than -15, but less than 30. Lastly, majority of the Class 2 data points can be seen falling into Cluster 1. The PCA dimension 0 values of these points range from being greater than -250 to being less than 480, whilst the PCA dimension 1 values of these points range from around -27 to less than 60.

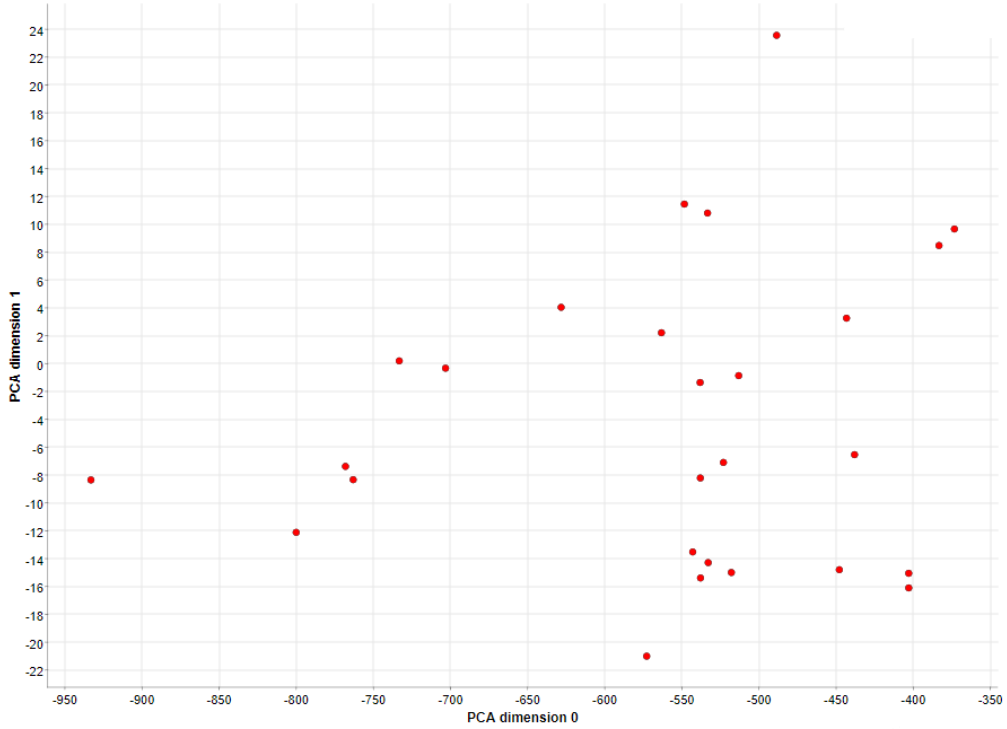
Based on the analysis of the points and their values of both graphs, it is fair to say that, although the wine dataset is shown to be partitioned into 3 clusters that are clearly well defined, doesn't mean that the class values are well grouped. The class values are more broadly spread out and mixed, which means that the data points plotted are unevenly distributed. This suggests that not all the class values had been assigned correctly into the created cluster groups, which shows that the data had not been normalised.

For part 4 of the first task, the process of applying K-means to split the dataset into 3 clusters and reducing those clusters into 2 dimensions through the PCA method was used again. After this, the data points associated with each cluster ID were plotted onto individual scatter

graphs, where like with part 1, each data point was assigned a colour according to their class (Figures 3a, 3b and 3c).







***Figures 3a, 3b and 3c**. Scatter graphs showing the distribution of the non-normalised PCA values of the wine dataset associated with their cluster ID but coloured according to their class value. 3a shows Cluster 0 datapoints, 3b shows Cluster 1 datapoints and 3c shows Cluster 2 datapoints.*

In the Cluster 0 graph (Figure 3a), it is seen that there is a mixture of data points from all 3 classes. The data points that are associated with Class 2 are very few and very sparse, ranging from greater than -240 to less than 30 on the PCA dimension 0 axis and ranging from greater than -5 to less than 60 on the PCA dimension 1 axis. Class 1 data points make up the majority of Cluster 0 and range from greater than -360 to less than 20 on the PCA dimension 0 axis and greater than -15 to less than 30 on the PCA dimension 1 axis. The data points of Class 3 make up the second majority of Cluster 0 and range from greater than -140 to less than 30 on the PCA dimension 0 axis and greater than -15 to less than 20 on the PCA dimension 1 axis.

In the Cluster 1 graph (Figure 3b), there are mostly a mixture of data points from 2 classes, which are Class 2 and Class 3. The only Class 1 data point is located at the coordinates (67, 4). Class 2 data points make up the majority of Cluster 1 and range from greater than 20 to less than 480 on the PCA dimension 0 axis and greater than -30 to less than 40 on the PCA dimension 1 axis. Class 3 data points make up the second majority of Cluster 1 and range from greater than -140 to less than 30 on the PCA dimension 0 axis and greater than -15 to less than 20 on the PCA dimension 1 axis.

In the Cluster 2 graph (Figure 3c), the only data points plotted are of Class 1. These data points range from greater than -950 to less than -350 on the PCA dimension 0 axis and greater than -22 to less than 24 on the PCA dimension 1 axis. Overall, even with applying the K-Means algorithm, the classes to do not get correctly sorted into the cluster groups that are made. This further proves that normalisation would fix this issue.

Two cluster validity measures (Figures 4a and 4b) were used to check the quality of the clustering that is returned after applying the K-Means algorithm to the dataset to generate the clusters. The first table (Figure 4a) checks the WSS (within cluster sum of squares) and BSS (between cluster sum of squares), which measures how closely related data points are in a cluster and how well-separated a cluster is from other clusters respectively. The results returned showed that the WSS and BSS values are incredibly high, which suggests that there is little to no relations between data points in a cluster and that the cluster-cluster separation is non-existent. The second table (Figure 4b) shows the entropy and quality of the clustering. The entropy and quality of the clustering returned was 0.942 and 0.4057 respectively, meaning that the clustering is overall bad. Both these findings add proof that data normalisation would help fix these issues.

Row ID	WSS	BSS
validity	2,633,614.463	14,958,788.24

Score	Value
Entropy:	0.942
Quality:	0.4057

Row ID	Size	Entropy	Normali...	Quality
cluster_2	27	0	0	?
cluster_1	102	1.018	0.642	?
cluster_0	49	1.303	0.822	?
Overall	178	0.942	0.594	0.406

***Figures 4a and 4b**. Cluster Validity Measures (unnormalised) using WSS & BSS scores and Entropy and Quality scores.*

Task 2: Clustering with Normalisation

In the second task, clustering analysis of the wine dataset was carried out with normalisation of the data prior. This was done by applying the Min-Max normalisation method (where the data was scaled from 0.0 to 1.0) to the entire dataset except for the Class column. For part 1 of the second task, the Principal Component Analysis (PCA) method was applied to the dataset to reduce it down into two-dimensional coordinates, where each data point of the dataset was assigned a colour according to their class. Red represented that the data point belonged to Class 1, purple represented that the data point belonged to Class 2 and blue represented that the data point belonged to Class 3. After being assigned the colours, the PCA values were plotted on a scatter graph (Figure 5) to show the distribution of the data points, where the x-axis represented PCA dimension 0 values, and the y-axis represented PCA dimension 1 values.

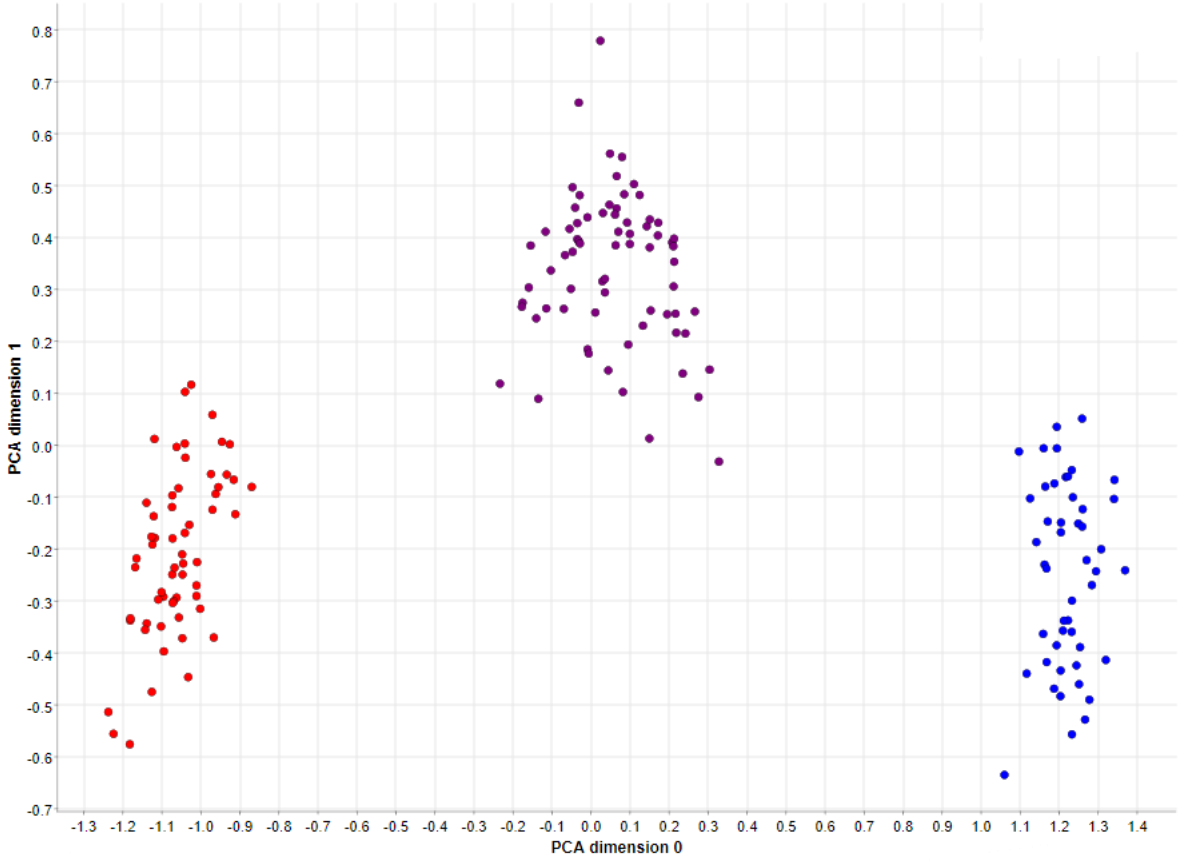


Figure 5. *Scatter graph showing the distribution of the normalised PCA values of the wine dataset according to their class value.*

For part 2 of the second task, the K-Means algorithm was applied to the wine dataset to create 3 clusters (partitions). Once this was done, the PCA method was applied to all the 3 clusters to reduce them to two-dimensional coordinates. This time, each data point was assigned a colour according to their cluster ID. Green represented that the data point belonged to Cluster 0, orange represented that the data point belonged to Cluster 1 and brown represented that the data point belonged to Cluster 2. After being assigned the colours, the PCA values were plotted on another scatter graph (Figure 6) to show the distribution of these data points, where the x-axis represented PCA dimension 0 values, and the y-axis represented PCA dimension 1 values.

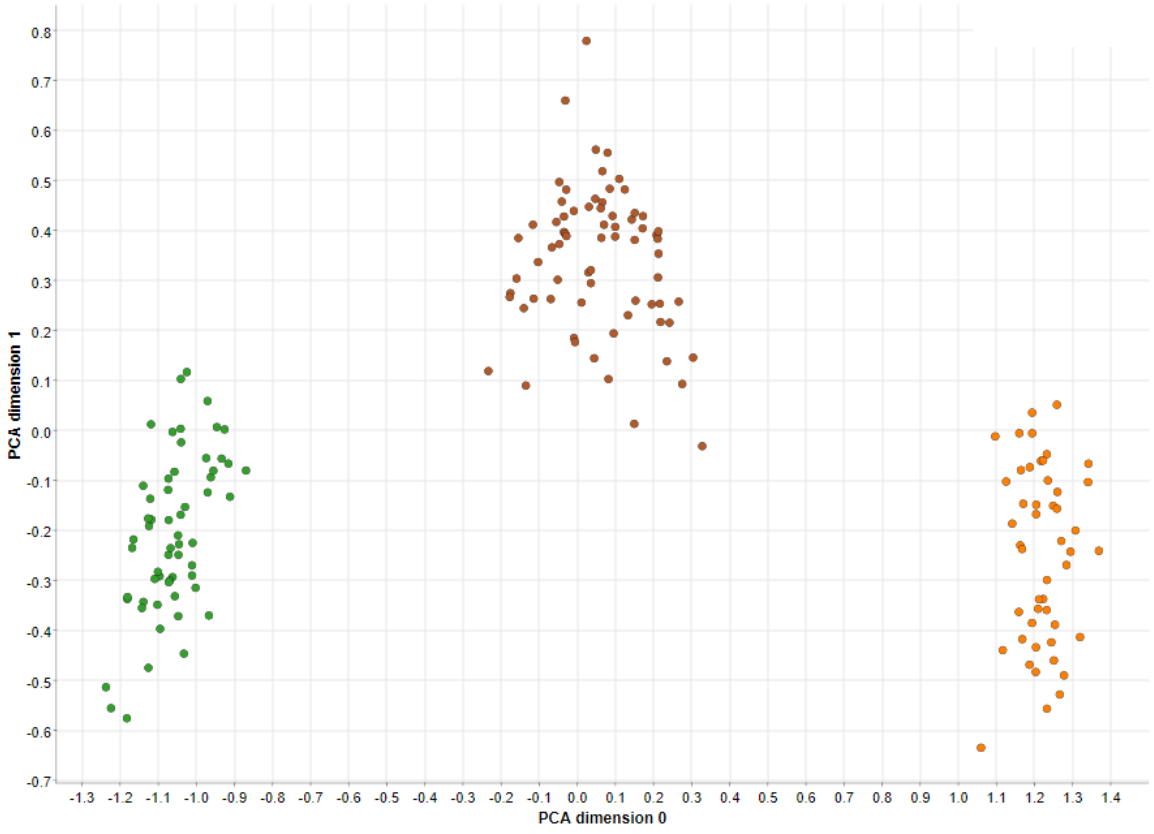


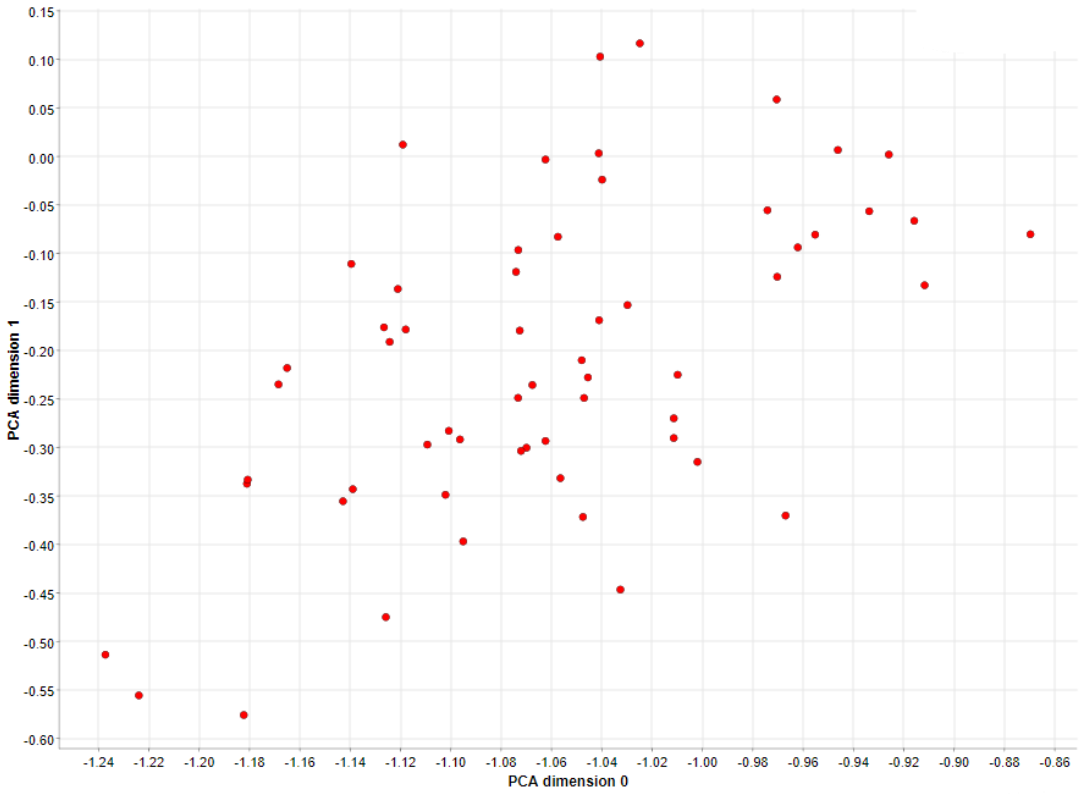
***Figure 6**. Scatter graph showing the distribution of the normalised PCA values of the wine dataset according to their cluster ID value.*

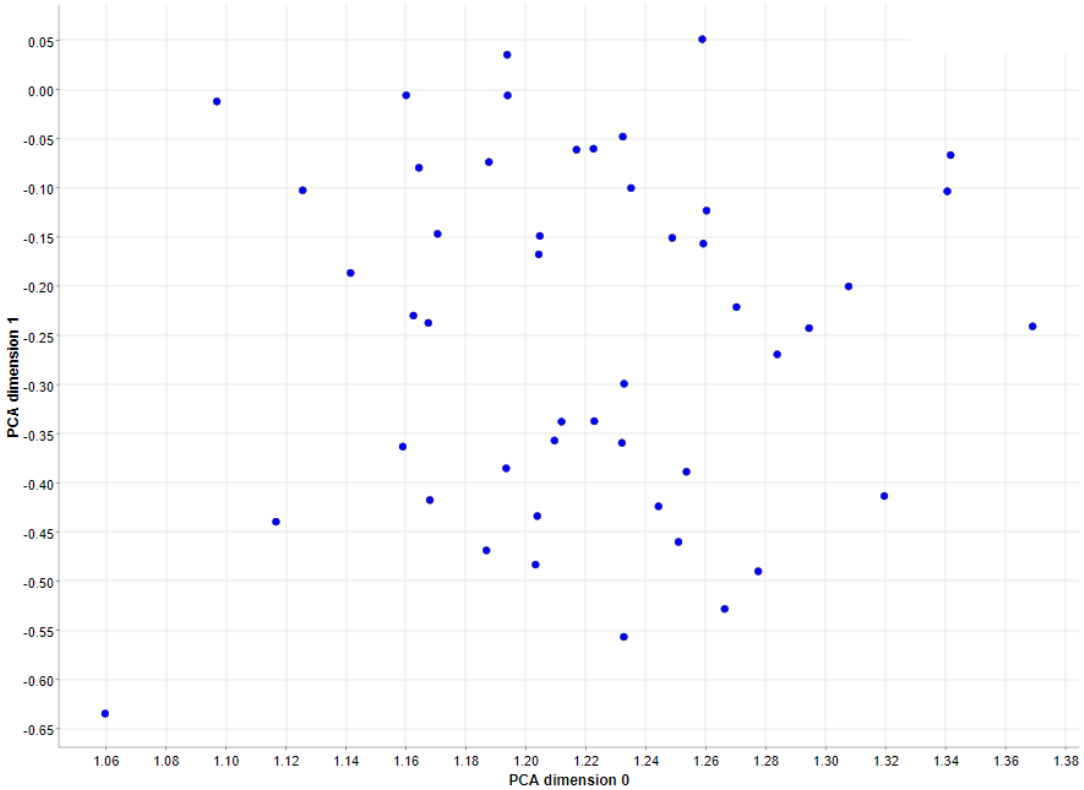
By observing the two scatter graphs generated, we can see that all the data points are now organised nicely into 3 separate groups. Based off the colour coding used for the scatter graphs in figures 6 and 7, it has been observed that all the Class 1 data points had grouped into Cluster 0, all the Class 2 data points had been grouped into Cluster 2 and all the Class 3 data points had been grouped into Cluster 1. The Class 1 data points range from greater than -1.3 to less than -0.8 on the PCA dimension 0 axis and greater than -0.6 to less than 0.2 on the PCA dimension 1 axis. The Class 2 data points range from greater than -0.3 to less than 0.4 on the PCA dimension 0 axis and greater than -0.1 to less than 0.8 on the PCA dimension 1 axis. Lastly, The Class 3 data points range from greater than 1.0 to less than 1.4 on the PCA dimension 0 axis and greater than -0.7 to less than 0.1 on the PCA dimension 1 axis.

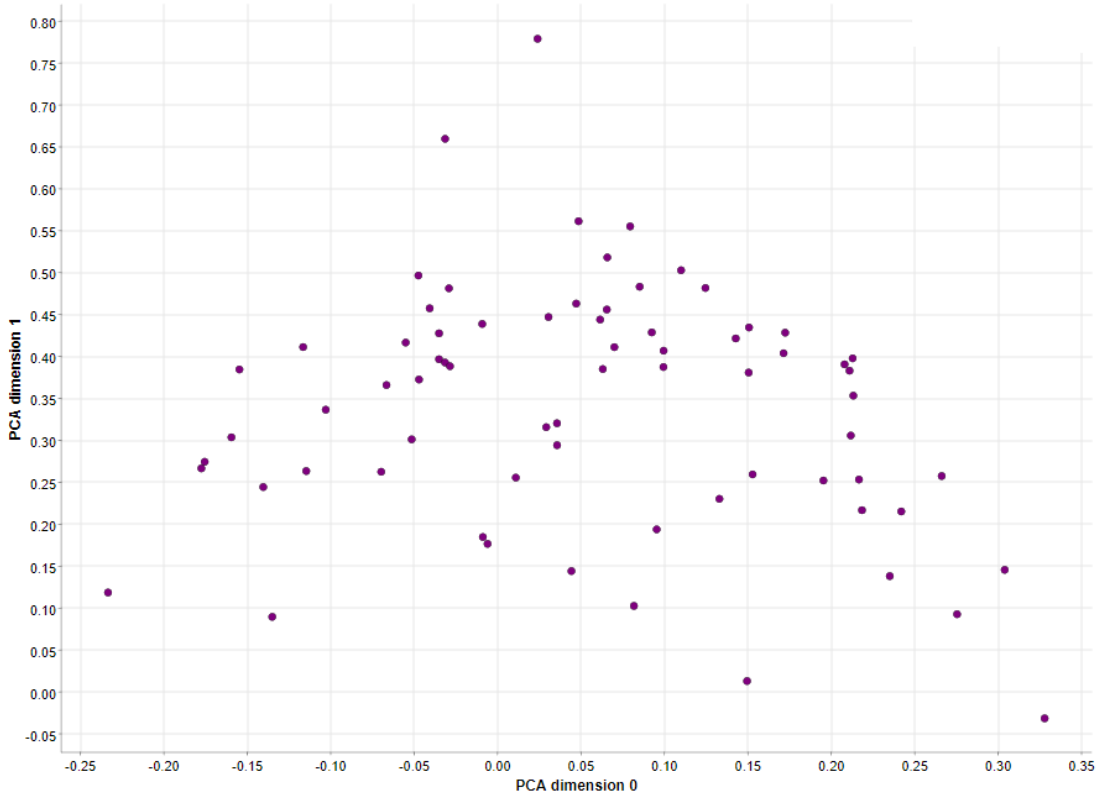
Based on the analysis of the points and their values of both graphs, it is fair to say that, thanks to normalisation carried out prior, the wine dataset has been partitioned into 3 clusters that are more defined than in the first task. The class values had been correctly assigned and organised into these clusters better than without normalisation. These values are more compact and not mixed, which means that the data points plotted are evenly distributed. This shows that the data had been normalised and that it had worked correctly.

For part 4 of the second task, the process of applying K-means to split the dataset into 3 clusters and reducing those clusters into 2 dimensions through the PCA method was used again. After this, the data points associated with each cluster ID were plotted onto individual

scatter graphs, where like with part 1, each data point was assigned a colour according to their class (Figures 7a, 7b and 7c).







Figures 7a, 7b and 7c*. Scatter graphs showing the distribution of the normalised PCA values of the wine dataset associated with their cluster ID but coloured according to their class value. 7a shows Cluster 0 datapoints, 7b shows Cluster 1 datapoints and 7c shows Cluster 2 datapoints.*

In the Cluster 0 graph (Figure 7a), there are only data points of Class 1 plotted. These data points range from greater than -1.24 to less than -0.86 on the PCA dimension 0 axis and greater than -0.60 to less than 0.15. In the Cluster 1 graph (Figure 7b), there are only data points of Class 3 plotted. These data points range from greater than 1.05 to less than 1.38 on the PCA dimension 0 axis and greater than -0.65 to less than 0.10. Lastly, in the Cluster 2 graph (Figure 7c), the only data points plotted are of Class 2. These data points range from greater than -0.25 to less than 0.35 on the PCA dimension 0 axis and greater than -0.05 to less than 0.80 on the PCA dimension 1 axis. This shows that the K-Means algorithm has worked more effectively in creating and organising the clusters better thanks to the normalisation step that had took place prior.

In the final part of the second task, two cluster validity measures (Figures 8a and 8b) were used to check the quality of the clustering that is returned after applying the K-Means algorithm to the dataset to generate the clusters. The first table (Figure 8a) checks the WSS (within cluster sum of squares) and BSS (between cluster sum of squares), which measures how closely related data points are in a cluster and how well-separated a cluster is from other clusters respectively. The results returned showed that the WSS and BSS values are way lower than what was obtained in the first task, which suggests that there are strong relations between data points in a cluster and that the cluster-cluster separation is present. The second table (Figure 8b) shows the entropy and quality of the clustering. The entropy and quality of

the clustering returned was 0.0 and 1 respectively, meaning that the clustering is perfect. Both these findings prove that the normalisation method had correctly assigned all the data points into their classes and that those classes had been properly assigned to their cluster ID.

Row ID	D WSS	D BSS
validity	49.999	151.921

Score	Value
Entropy:	0.0
Quality:	1

Row ID	I Size	D Entropy	D Normali...	D Quality
cluster_0	59	0	0	?
cluster_2	71	0	0	?
cluster_1	48	0	0	?
Overall	178	0	0	1

***Figures 8a and 8b**. Cluster Validity Measures (normalised) using WSS & BSS scores and Entropy and Quality scores.*