**Assignment 3 Report**

1. The hierarchical structures generated using single link, complete link and group average for the Vertebral Column data set are different. From the dendrogram, we can see that firstly, for the single link, the merging order is defined as the minimum of the distance between any two points. Therefore, the space between every clusters is small. It is very different to see if we can divide it into 3 clusters. Therefore, the degree of balance is low. It is more likely the data set is just in one big cluster.

Secondly, for the average link, since the merging order is defined as the average distance across all pairs of points. The space between every clusters is wider than the above single link. From the dendrogram, we can observe that the data set can be divided into 3 clusters. However, the degree of balance is relatively low too. Because the number of points in each cluster has huge difference. One cluster has more points but the other has very less points.

Thirdly, for the complete link, as the merging order is defined as the maximum of the distance between any two points. The space between every clusters is wider than the above structures. From the dendrogram, we can clearly observe that the data set can be divided into 3 clusters. Therefore, the degree of balance is relatively higher than the above structures. Although some clusters have more data points, the overall number of points in each cluster is much balanced.

To conclude, using the complete link structures results in better clustering solution that we can clearly divide the data set into three clusters.

1. Between the single link and complete link, we can observe a large difference of the set of distance values at which cluster merge occurs. Because for the complete link, the distance between two clusters is defines as the maximum of the distance between any two points, with one of the points in the first cluster, and the other point in the second cluster. Therefore, from the dendrogram, we can see that the distance values at which the cluster merge occurs with two points is the maximum. The clusters are wider to each other.

On the other hand, for the single link, because the distance between two clusters is defined as the minimum of the distance between any two points, with one of the points in the first cluster, and the other point in the second cluster. Therefore, from the dendrogram, we can observe that the distance values at which the cluster merge occurs with two points is the minimum. The clusters are closer to each other.

Finally, for the group average structure, the set of distance values is the medium at which cluster merge occurs between the above two structures. Because the distance between two clusters is defined as the average distance across all pairs of points and this is an intermediate approach between the single and complete link approaches. Therefore, from the dendrogram, we can observe that the distance values at which the cluster merge occurs with two points is the medium. The clusters are closer to each other. The clusters are not very wide or very close to each other.

1. Since from the above analysis, we know that complete link has the best performance. Thus, we will compare the complete link structure with 3 number of clusters with the corresponding K-means clustering solution. The complete link structure has a higher accuracy than the K-means solution. The complete link structure has around 32% accuracy comparing with the actual data points. However, for the k-means clustering solution, the accuracy is only around 20% ~ 32%. We can see that the complete link structure is more stable and accurate than K-means.
2. If only the first three attributes which are pelvic incidence, pelvic tilt and lumbar lordosis angle are selected to perform the hierarchical clustering, we can see that the complete link structure generated from the dendrogram, the data set is much more clearly than the original hierarchical structure. The number of data points in each cluster are closer to each other than the original one. Thus, it is much balanced. However, the accuracy of the structure is close to the original one which is also around 31%.

When only the first four attributes are selected to perform the hierarchical clustering, this time we can see that the complete link structure generated from the dendrogram, it is not really balanced than the above and the original one. One cluster obviously has more data points, and one cluster has very few. However, the accuracy of the structure is around 61% which is higher than the above and the original one.