Data Pre-processing

**Remove Outlier**

As we have discovered in the previous assignments, the dataset itself contains several outliers, which will affect the accuracy and outcome of the decision tree. However, for the hierarchical clustering method, the outlier will severely affect the outcome. Thus, I have applied several steps for the outlier detection (Isolation Forest, Minimum Covariance Determinant, Local Outlier Factor (LOF), One-Class SVM, Density-based spatial clustering of applications with noise(DBSCAN)), and selected LOF as the final candidate for this assignment.

To verify which outlier detection to use for doing clustering, I used the linear regression prediction model to find out the mean absolute error and estimate how much outlier has been removed from each method.

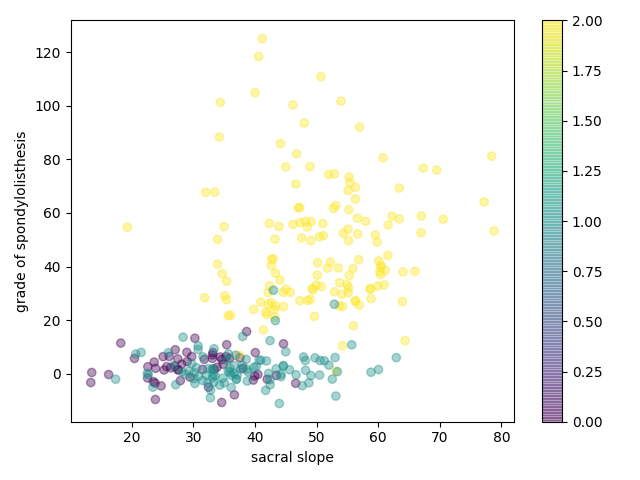
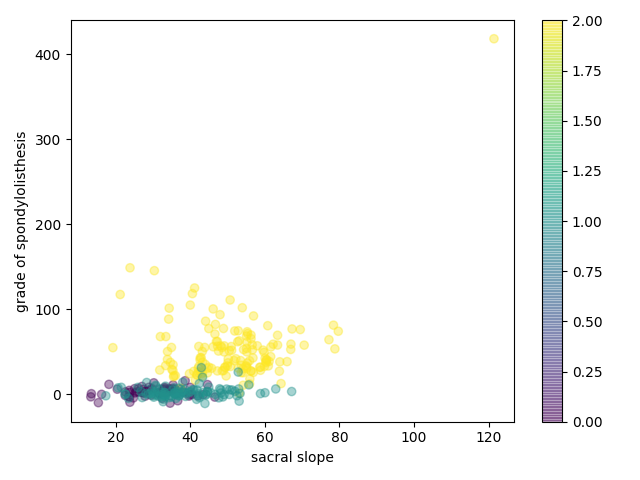
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method Name | | Final train shape | Removed outliers | | MAE(%.3f) |
| Base | 207 | | 0 | 0.445 | |
| Isolation Forest | 186 | | 21 | 0.415 | |
| MCD | 202 | | 5 | 0.413 | |
| LOF | 204 | | 3 | 0.408 | |

For the clustering methodology, it is much more accurate to use DBSCAN, LOF or OPTICS to remove the outlier based on the consideration of Euclidean distance. Besides, for the comparison between LOF and DBSCAN, DBSCAN computes a single level clustering, on the other hand, OPTICS and LOF represents the intrinsic, hierarchical structure of the data by ordering of the points. The resulting graph (called reachability plot) visualizes clusters of different densities as well as hierarchical clusters.

Here is the final outcome when passing all data (310) to the LOF outlier detection,

 It has removed the twelve outlier which is 

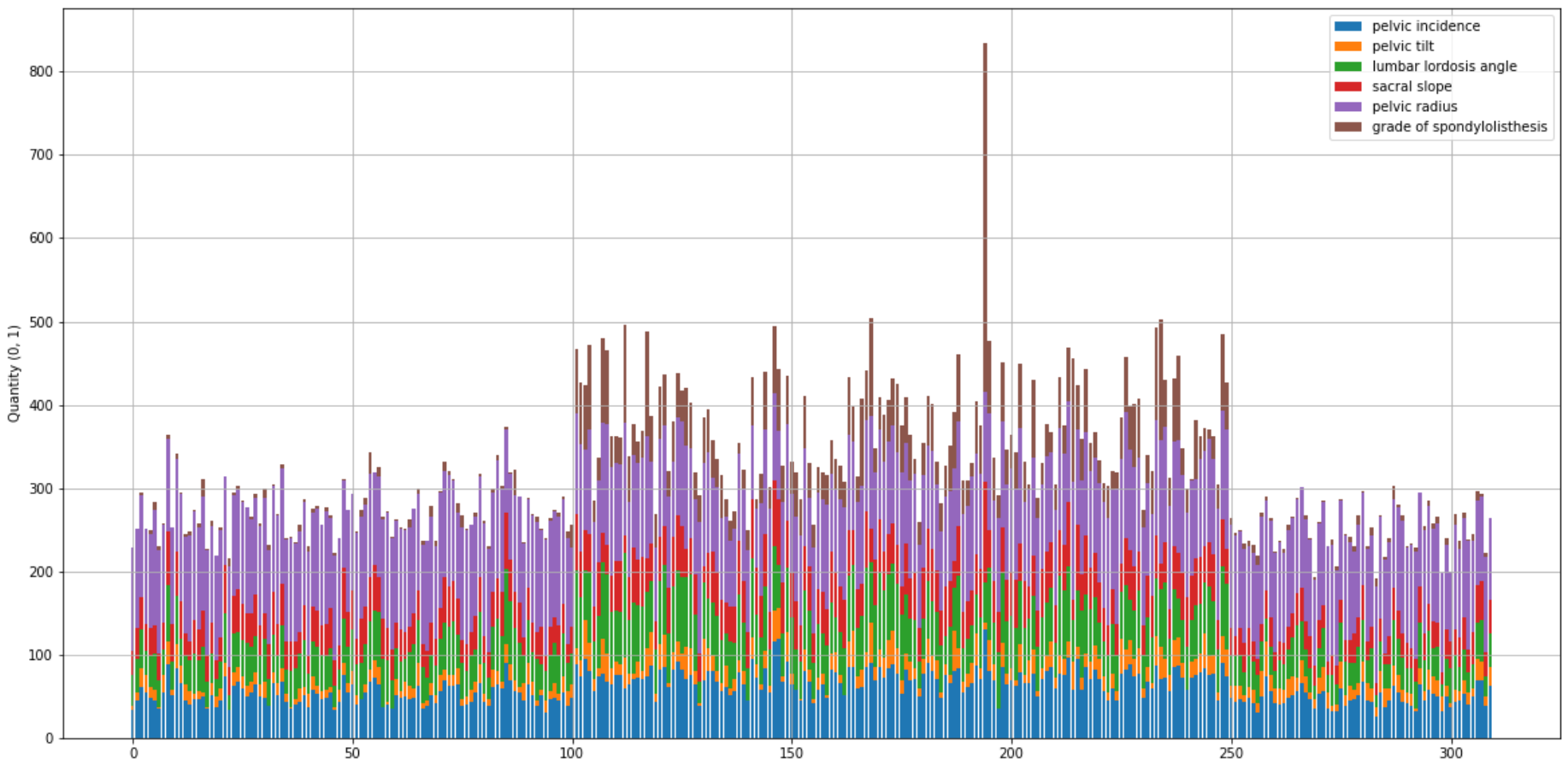
And I have scatter plot the two attributes that considered as the key attribute for doing decision tree classifying in assignment 2 ("sacral slope", "grade of spondylolisthesis") and showing out the before and after of removing outlier process.

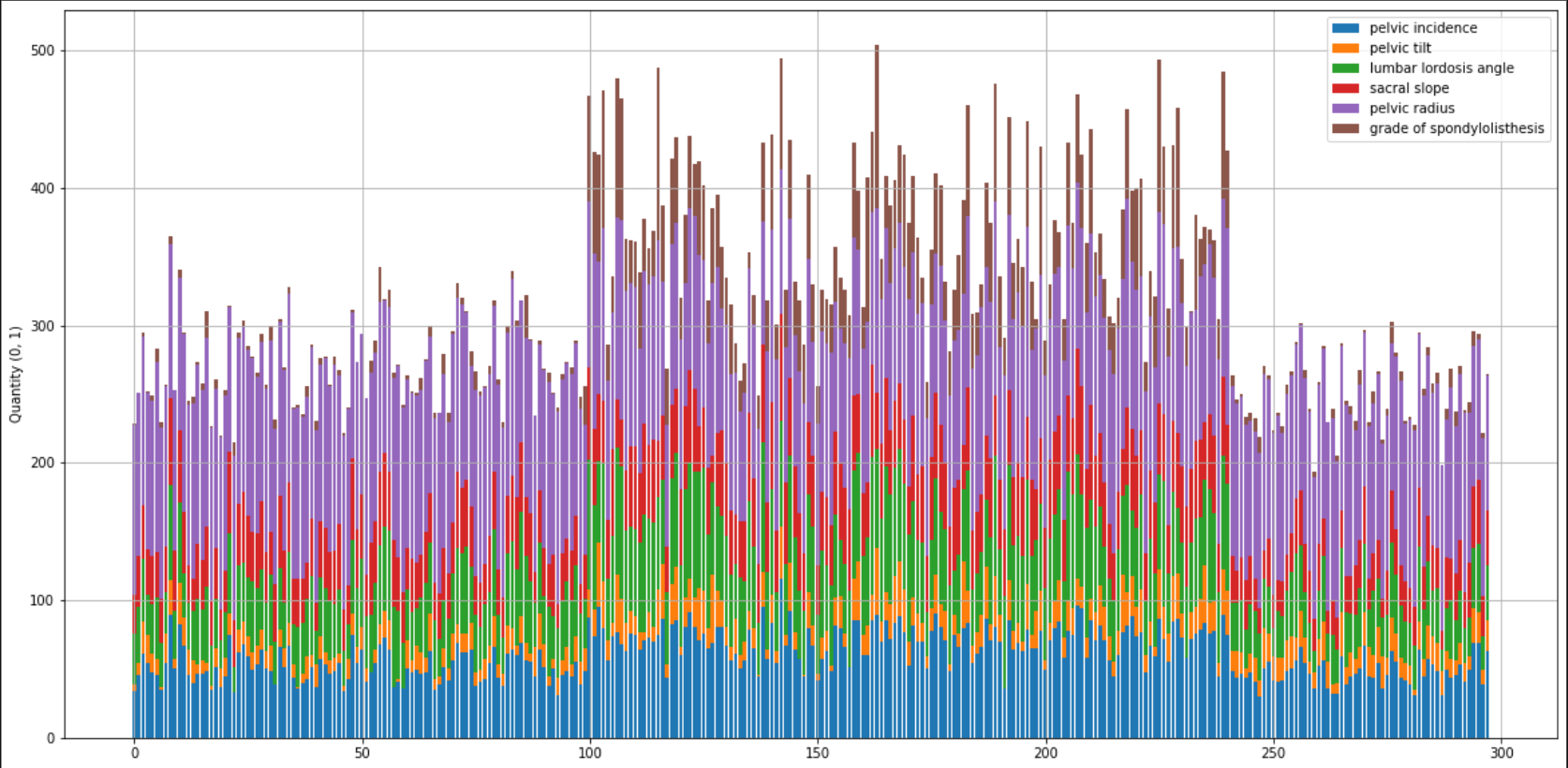




Outlier Found

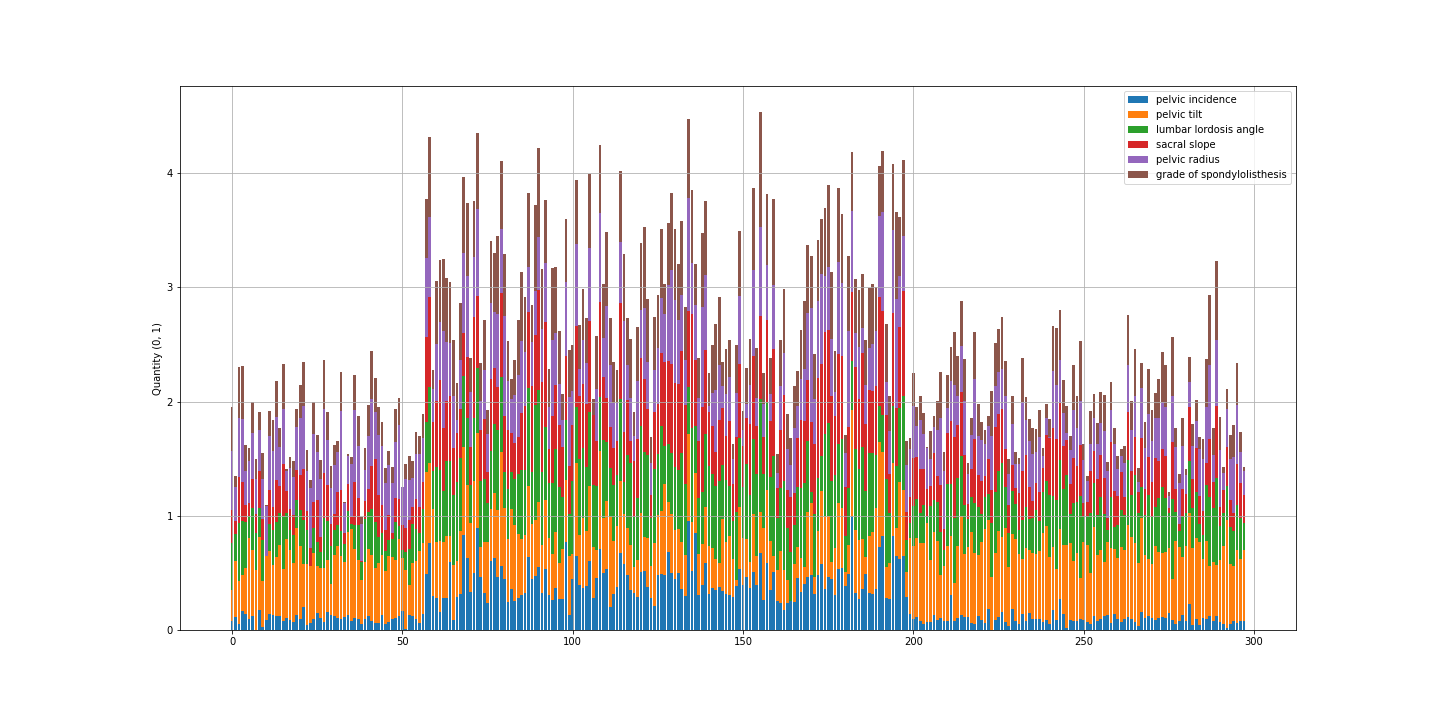
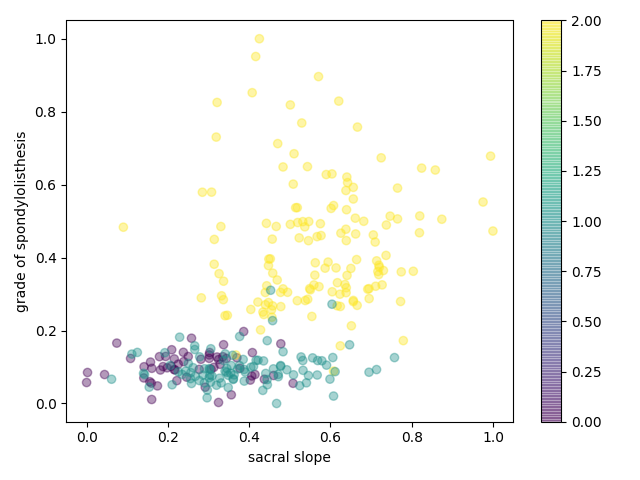






**Normalization**

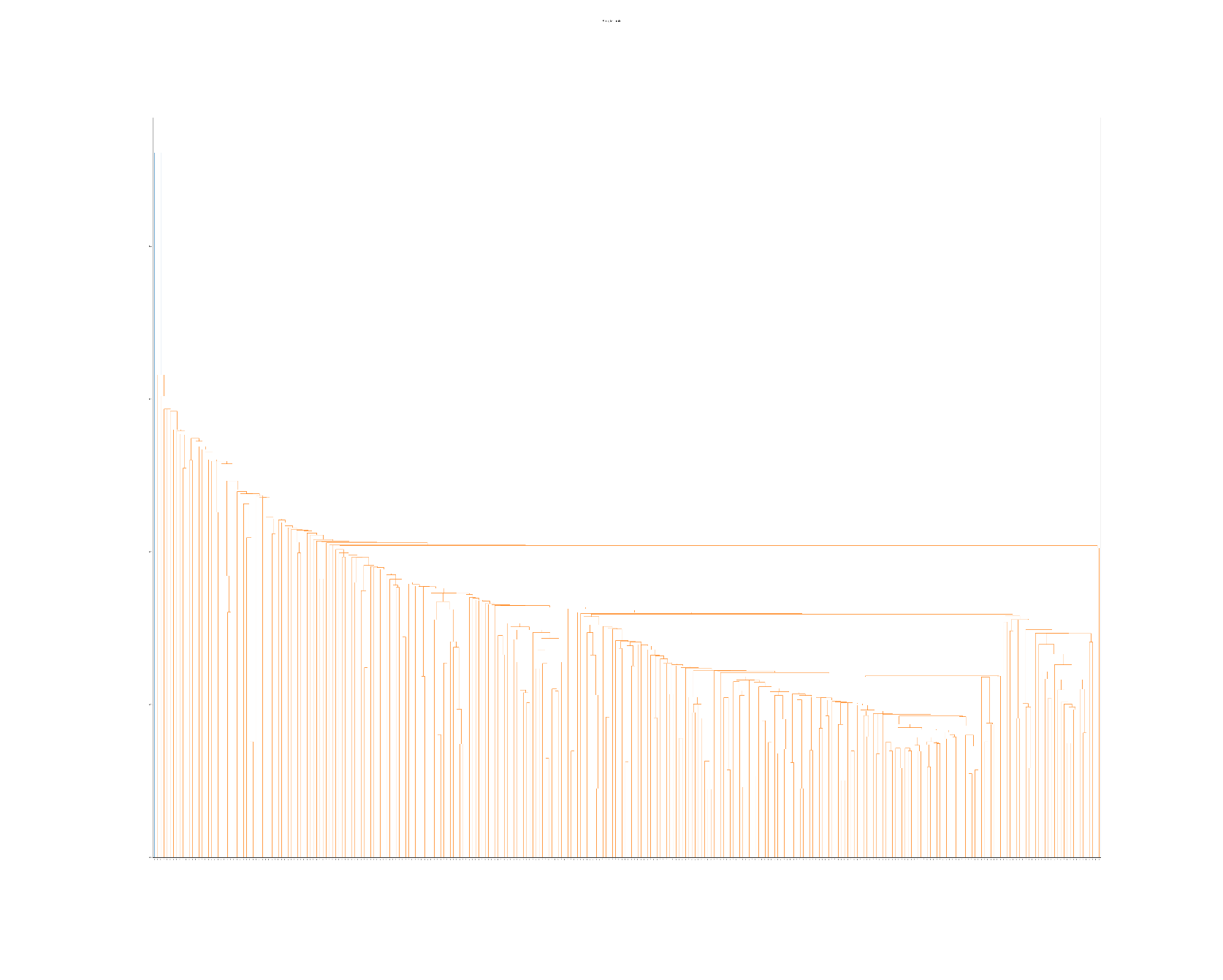
After removed all the outliers, now we can normalization all the rest data by the min\_max\_scaler function with the range from 0 ~ 1.

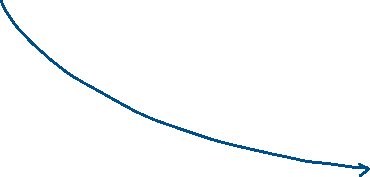


Question 1

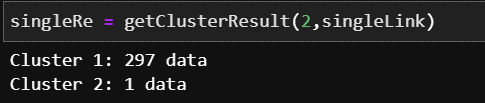
**Single Linkage**

By calculating the Euclidean distance compare with the point to the closest point in each cluster, we can visualize the result in dendrogram below.





From the figure, we can barely allocate the observation line to it since there is no apparent clusters on the top. We can also find out that most of the data is slowly clustered to each other and there is no good differentiator for these datasets. To better see the outcome of clustering, I have assumed that there are two clusters in total and see the clustering outcome. (Since the clustering effect isn’t good, two is enough…)



Most of the date were allocated to single main cluster.

There are two assumption which can illustrate this situation: 1) All the datasets are too close to each other, which means that distance between each point is too close to find out any “good” clusters. 2) From the exponential reduce trends from the left-hand side of the figure, we can hypothesize that the data tends to spread outwards from the single and main cluster. Whenever merging the data, the distance and difference become larger.

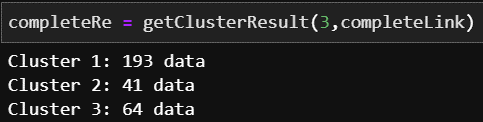
**Complete Linkage**

From the above dendrogram, we can find out that the data is too close to each other and we can use “Complete Linkage”, which compare the distance matrix with the furthest point of each cluster.

Although it’s better to classified it to two clusters when allocating the observations line, however we can appraise it to be three clusters representing DH, SL, NO.

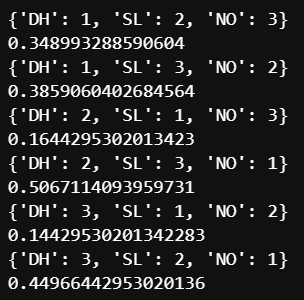
Chart, box and whisker chart

Description automatically generated



I made the assumption of three clusters in total.

Since the figure did not show me, which label is being categorized, I used another way which is to compare the accuracy score from the original dataset and analyze the maximum combination as the classification output of this methodology.



Picking the highest score of the combination, which is

{'DH': 2, 'SL': 3, 'NO': 1}, 0.5067114093959731

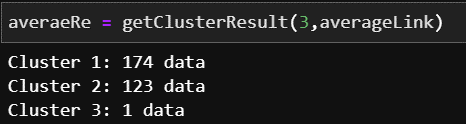
Even though the overall outcome is way much better than the single linkage, it is still apparent that the “NO” class has weighted much more than the other two, and it seems like to be only classify well of “NO” and “AB” = “DH”+”SL”.

**Group Average**

Since from the hypothesis which stated that the data is too similar to each other, it means that the group average linkage with taking the average distance between new spot and average of each clusters, will not accurately separate the DH, SL as well.

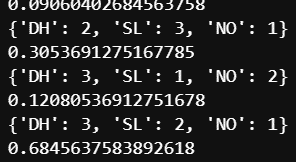
Chart

Description automatically generated



Data clustering into two clusters but not three labels as we expected.

And the accuracy outcome is similar to the complete linkage that it works well on classifying “NO” and “AB” but poorly perform on differentiating “DH”, “SL”.

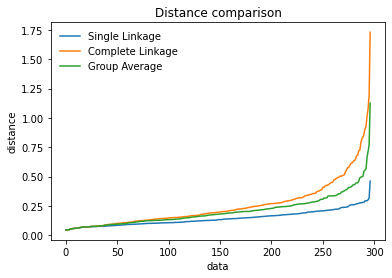


Compared to the complete linkage, the accuracy score is higher (0.685 > 0.507), which conducted that the average linkage has a much accurate separator between the NO and AB (DH+SL). However, among the three different method for comparing clusters, none of it can perform well on clustering DH and SL.

From the right-hand side of the figure, we can also see the length of branches is longer and exponential increasing. It indicated that this method can hardly find out the how should those data being clustered it to.

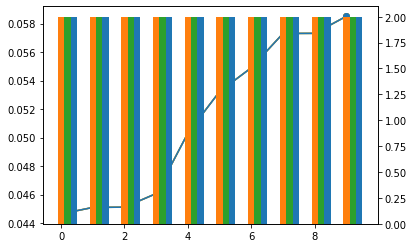
All in all, it reflects the outcome that we found on first and second assignment, the difference between data label of DH and SL is too small that it might confuse with each other.

Question 2

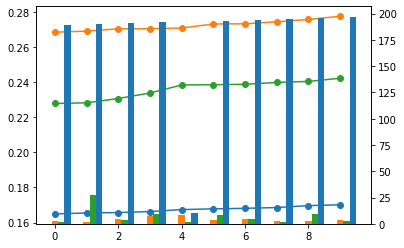
By extracting the [:, 2] of the linkage matrix, we can obtain the distance from data to clusters with its own methodology.

From the figure we can see that the distance increase exponentially, which illustrate that whenever the merging occurs, the distance (differentiation) to combine next cluster and data will become higher and higher. And the last few data or say merging has become extremely high since it’s trying to combine two big clusters that has not too much relationship.

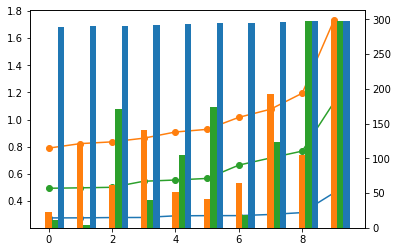
For better understanding of the period of merges, I extracted the first 10, 200-210, last 10 merges to observe. The figure shows the original data in new clusters on right y-axis and the normalized distance on the left y-axis with the merges index as the x-axis. The bar chart represents the number of original observations in the newly formed cluster. and the plot shows the distance.



From the first 10, it seems like all methods has the same speed on finding new clusters. And their growing rate on linkage distance are also the same, since it is just the beginning of the progress on clustering dataset.



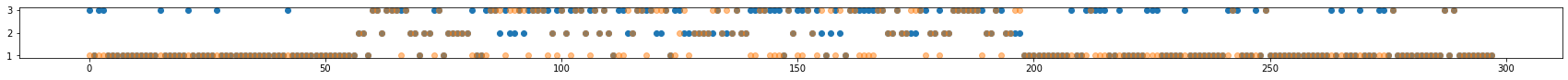
The right-hand-side figure shows the performance of index 200 ~ 210. It is obvious that the single linkage has already done the clustering job that the clusters remain to be the same height. And the other two methods seem like only has a little progress on clustering the dataset compare with the beginning. And we can find out that the difference of the distance shows the order from complete -> average -> single, which is the same as their definition.

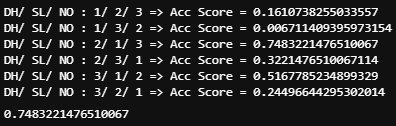
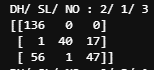
Further on observing the pattern of last 10 data, we can illustrate that the single linkage has done the clustering job early and the other two is still progressing on differentiate it which causing the valley and hill. The valley will be some small merges, and the hill will be the merge of two big clusters. Besides, we can find out that the distance of new data and nearest cluster data of single linkage seems to be the same and very close to each other, it has a high possibility that all data point is closed to each other and be classified into the same cluster. The other two methods seem like still perform well on the trend of increasing distance between new data and clusters. It is also interesting that the complete linkage seems to be parallel to the group average linkage which is below to it. And it has a high potential that the clustering methods conduct to the same clusters.

Question 3

3.a

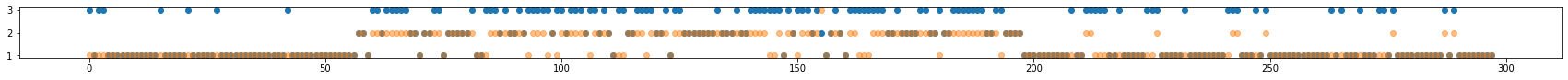
I have used the previous methods of premutation test, which will list out all the possible outcomes. And the one with the highest similarity between hierarchical clustering and K-means clustering will be selected as the observations.

With the comparison of Complete linkage and K-Means (n\_clusters = 3), I have obtained the following data: (Orange: Complete Linkage, Blue: K-Means)  


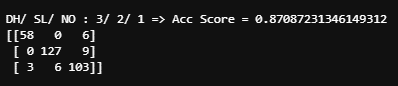
And as we can find out the similarity is pretty nice that the similarity has the highest at 74% with the confusion matrix:

And we can further verify again that the confusion between the classes has led to the variance results in clustering.

When we go on to the group average method, the accuracy score is much higher than what I have expected.



We can find out that the data overlapping is much higher than the complete linkage method. By using the same methods, we can get the highest scientific outcome below.

The similarity accuracy score is 13% higher than the complete linkage and it further explain that the definition of the K-Means is similar to the Group Average Linkage. From the K-Means method which finding and adjusting the centroid of the cluster, it is somehow similar to the group average linkage which always finding the distance between newcomer and the current centroid of each clusters. The following lines showing the centroid of the result of K-mean cluster.

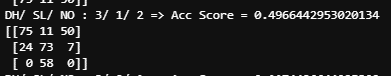
[[0.18758101 0.35277042 0.25446059 0.30069056 0.59477663 0.108691 ]

[0.60611982 0.65794434 0.70399197 0.58827576 0.58115094 0.58804206]

[0.4202395 0.39497966 0.49649293 0.56959469 0.42815523 0.27059773]]

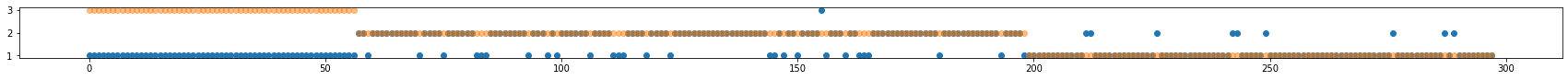
3.b

By using the same method explain on the previous question, we can compare the best result of the original dataset with the K-Means clusters. The accuracy score and confusion matrix are stated below.

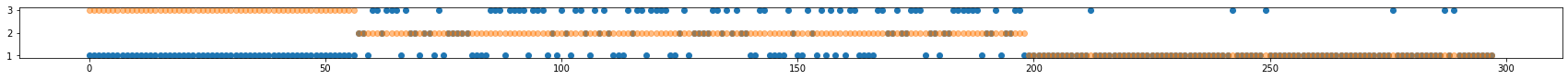


The outcome is similar to the result of group average linkage since as we found out that the similarity of group average and the K-means is very high.

When comparing the group average linkage to the dataset, the overlap of each data point is much higher than any other comparison.



And the below figure shows the comparison of complete linkage and original dataset, which showing the same trends as group average linkage on the range <50 & >200 but perform worse in the range from 50 ~ 200, which causing the worse accuracy score than group average.



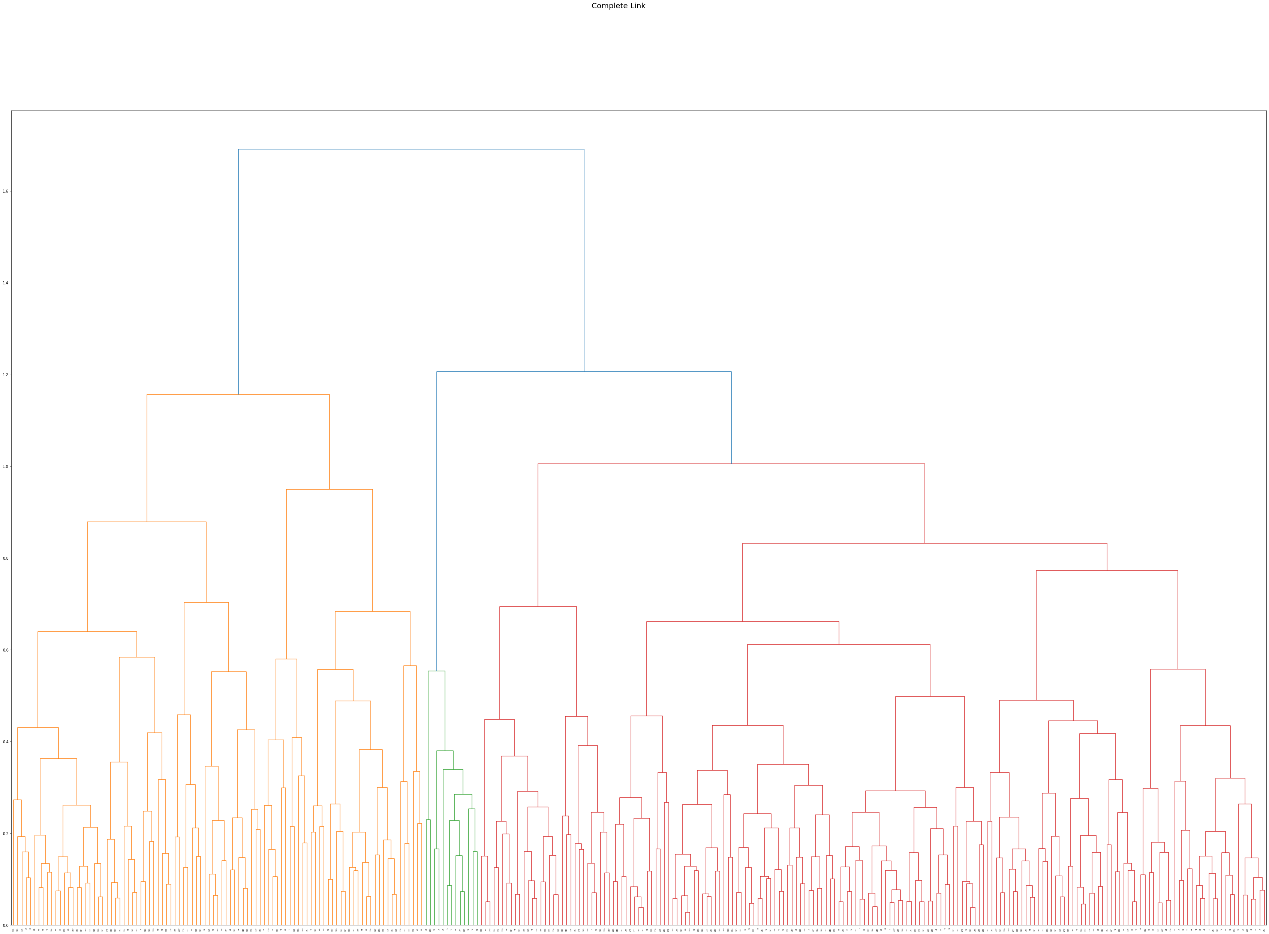
Question 4

For a better glimpse on the combinations, I have run through the 4 attributes’ combinations and 5 attributes’ combinations. Try to find out the “best” result and also figure out that classes that will mislead the clustering outcome. Firstly I run through the all possibilities of extracting 1 column out and the following table shows the highest accuracy complete linkage result. (the index refer to each attribute, 1 ~ 5)



|  |  |
| --- | --- |
| Original | 0.5067114093959731 |
| [1 ~ 5] | 0.6372881355932203 |
| [1 ~4, 6] | 0.5771186440677966 |
| [1~3, 5, 6] | 0.5771812080536913 |
| [1, 2, 4 ~ 6] | 0.523972602739726 |
| [1, 3 ~ 6] | 0.6216216216216216 |
| [2 ~ 6] | 0.5488215488215489 |

Surprisingly found out that all of the accuracy is higher than the result which considering every of each data column. It shows the “grade of spondylolisthesis” and “pelvic tilt” are the two main attributes dropping the accuracy of the complete linkage hierarchical clustering. When we print out the dendrogram without the “grade of spondylolisthesis”, it shows:



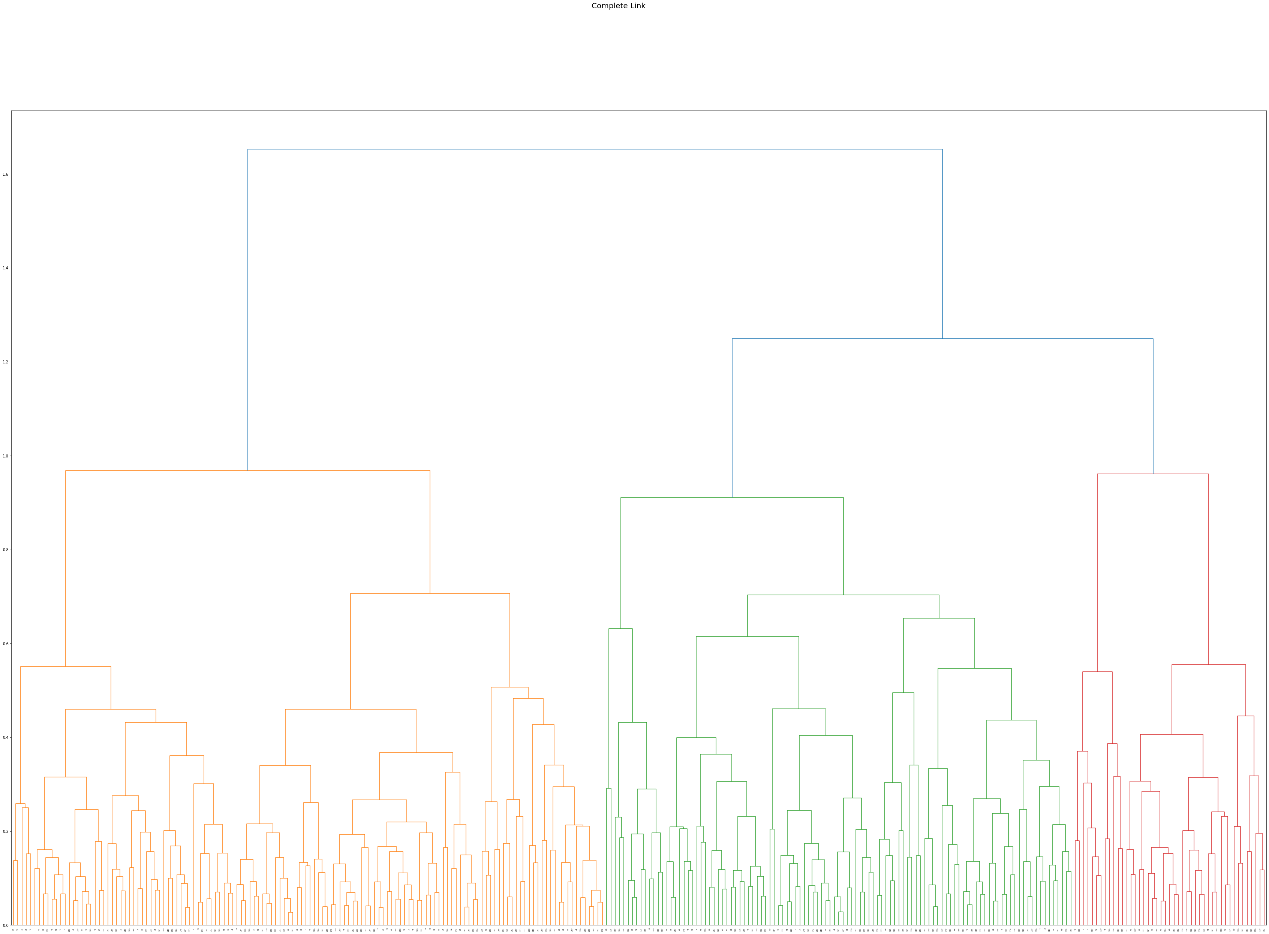
Cluster 1: 97 data

Cluster 2: 13 data

Cluster 3: 185 data

Which shows the better clustering outcome with 3 clusters than the original one which only find out the two clusters.

Further on reduce the two attributes “grade of spondylolisthesis” and “pelvic tilt” out of the dataset, and remove the outlier, normalize dataset again once. The re-perform dendrogram shows as below:

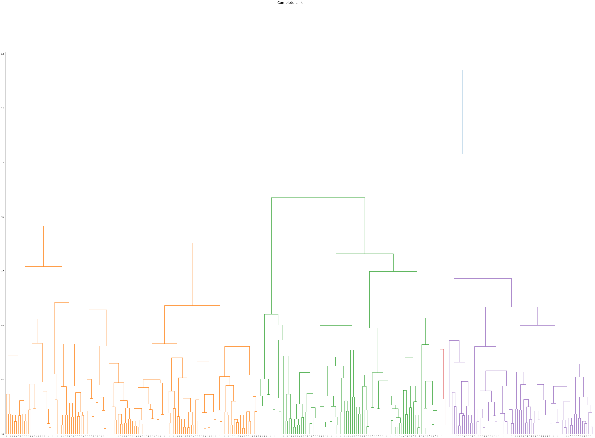
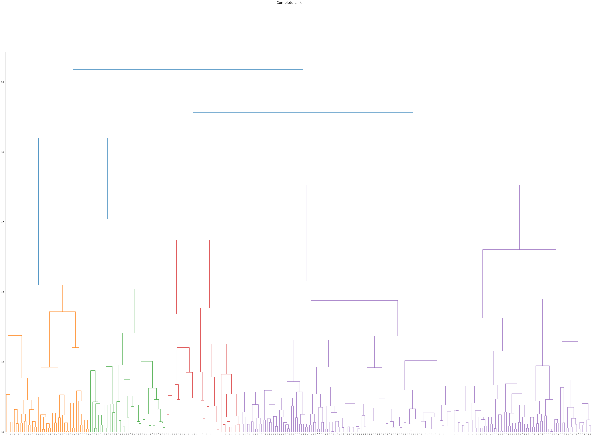


Cluster 1: 138 data

Cluster 2: 109 data

Cluster 3: 45 data

This is much more similar to the original and accurate label than considering all columns. This can further approve the result revealed from previous and the first assignments that there are some very confusable attributes which will lead to the classification error. Back to my previous assignment, I also pointed out that the “pelvic tilt” will be easily confused by the “pelvic radius”. Although the clustering method is very different from the decision tree, the result can still provide us the information that the dataset is uneven.

When we go on to take only three attributes, the clustering occasionally appears the new cluster and having the worse accuracy than the original dataset.

