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 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 predict model to find out the mean absolute error and estimate how much outlier has been removed from each method.

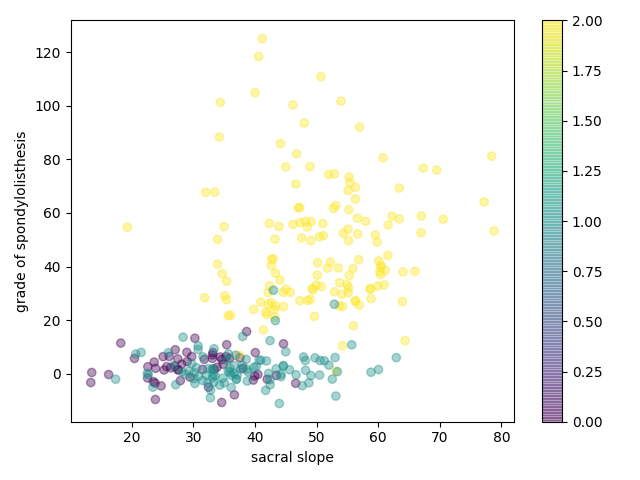
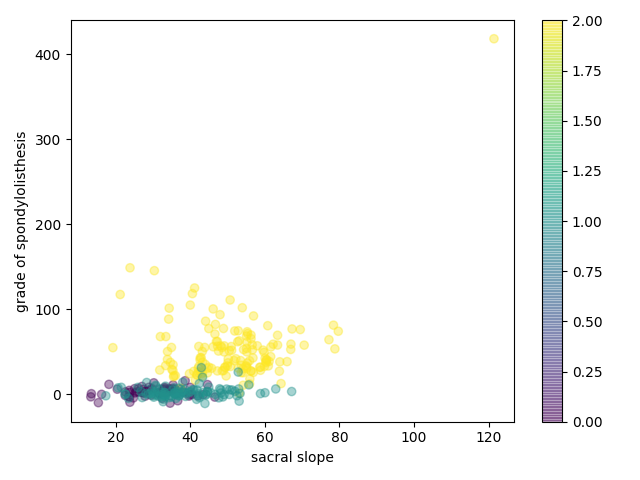
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method Name | | Final train shape | Removed outliers | | MAE(%.3f) |
| Base | 207 | | 0 | 0.445 | |
| Isolation Forest | 186 | | 21 | 0.415 | |
| MCD | 202 | | 5 | 0.413 | |
| LOF | 204 | | 3 | 0.408 | |

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 

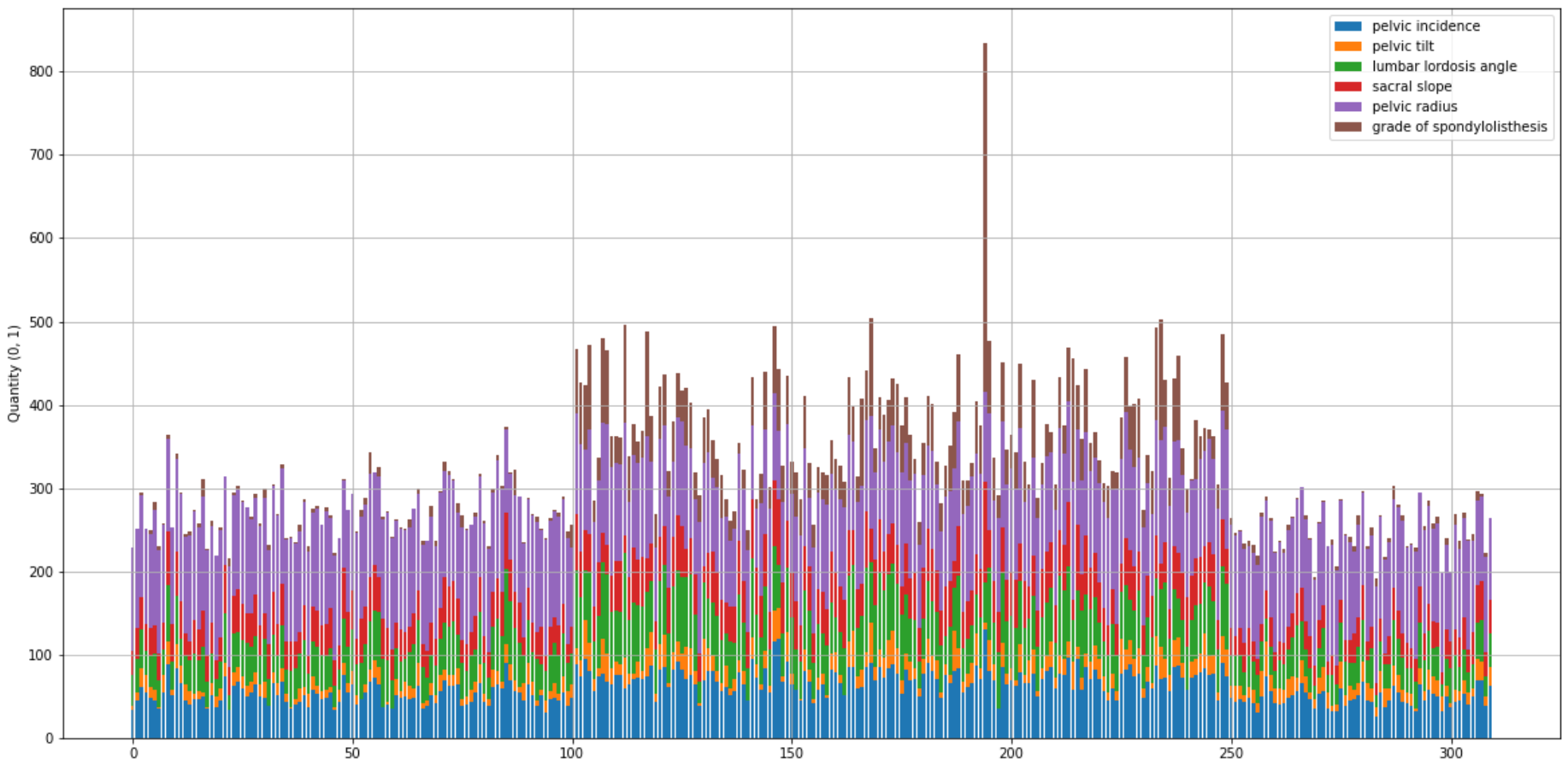
And I have scatter plot the two attributes that considered as the key attribute for doing decision tree classifying in assignment 2 ("sacral slope", "grade of spondylolisthesis") and showing out the before and after of removing outlier process.

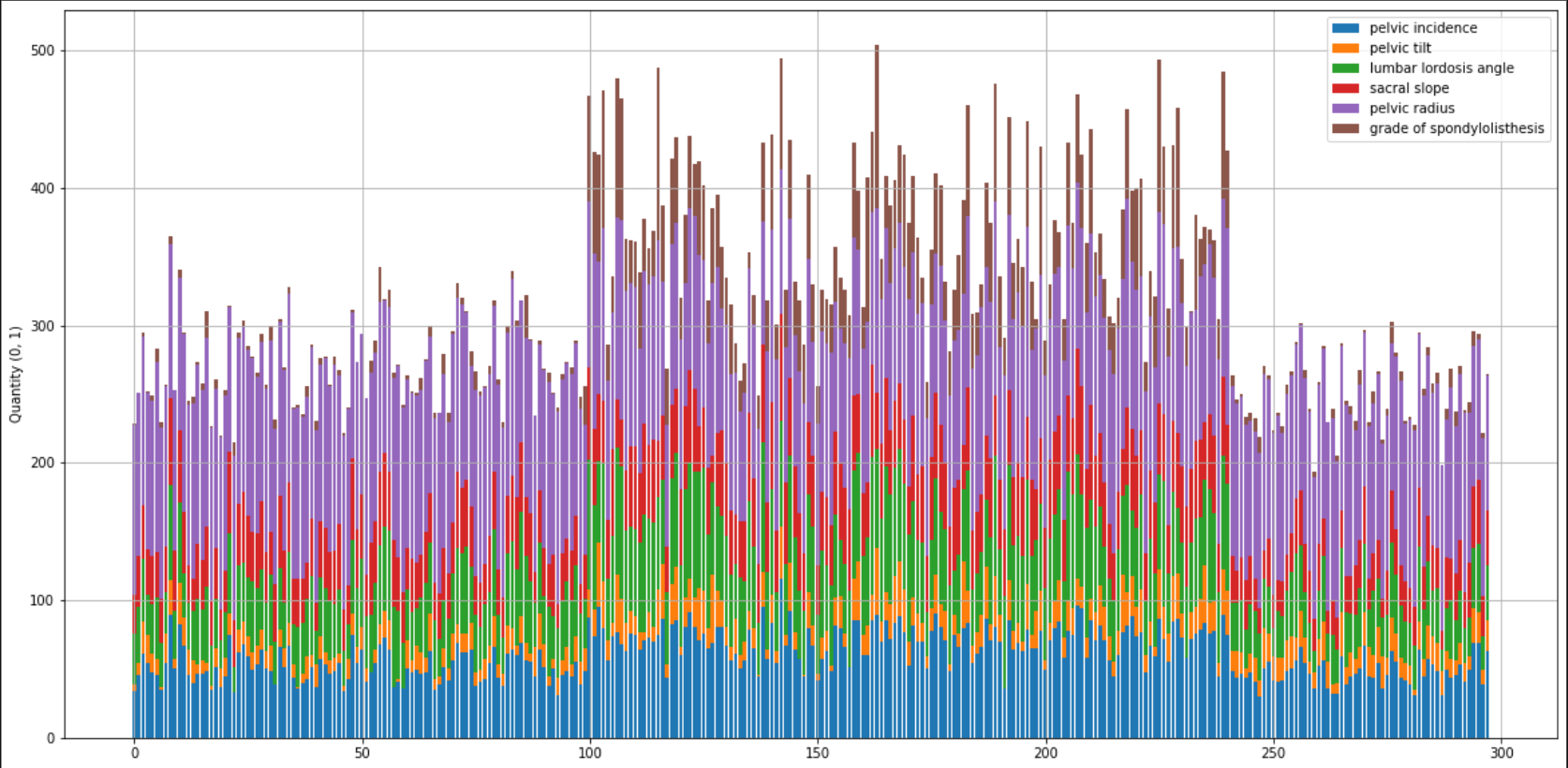




Outlier Found

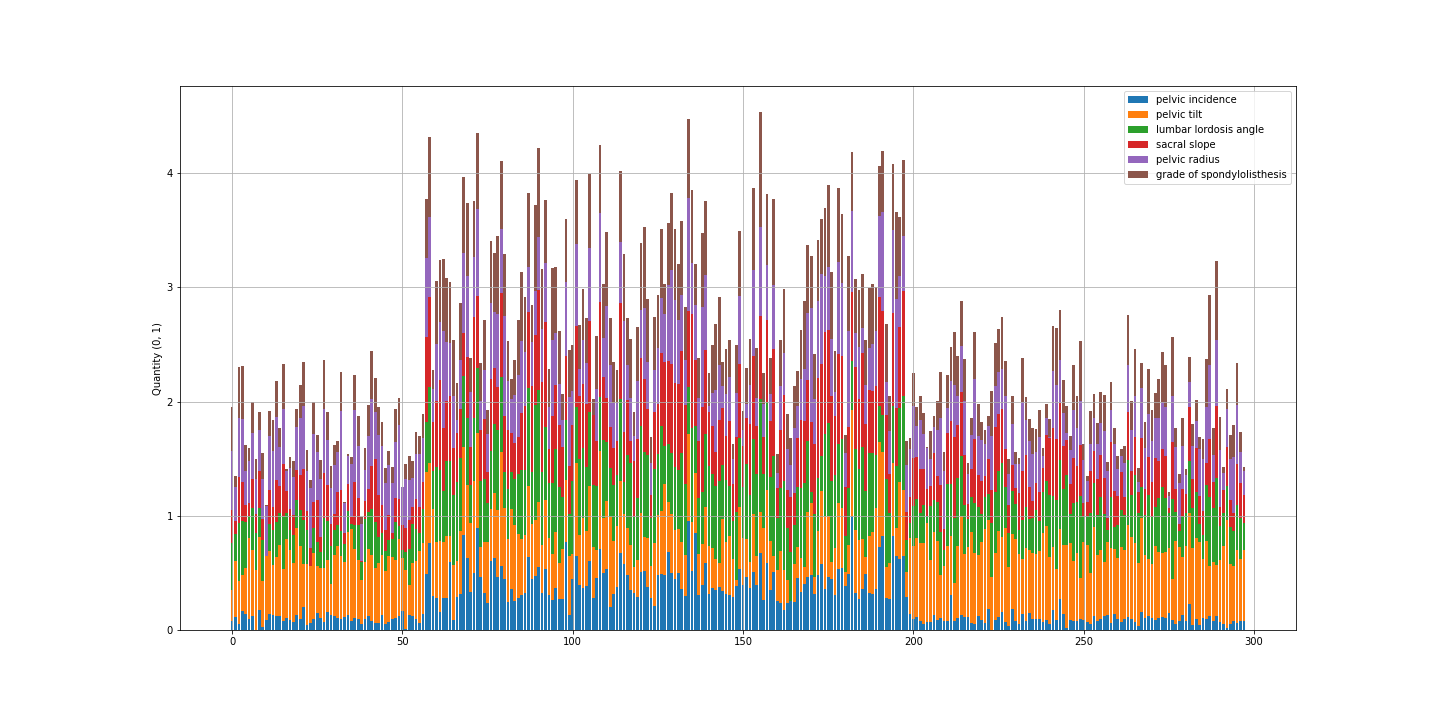
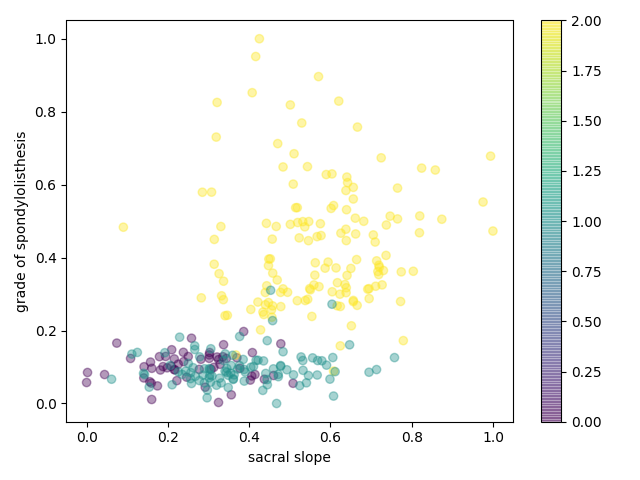






**Normalization**

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 Link**

