Heart Disease Cluster Analysis

By Brandon Wong

Introduction

The data being used is about heart disease (coronary artery disease and cardiomyopathy) and has 1035 rows with information about patients like their age, resting blood pressure (in mm Hg), cholesterol level (in mg/dL), maximum heart rate, ST depression induced by exercise, The slope of the ST segment during peak exercise (0, 1, or 2), presence or absence of exercise-induced chest pain, Number of major vessels (0-3) colored by fluoroscopy, and Thalassemia type (1: normal, 2: fixed defect, 3: reversible defect). The dataset, sourced from Kaggle, was compiled by Mahsa Sanaei, a Kaggle Datasets Grandmaster. The data set has no missing or null values and has been updated as recently as 4 months before the creation of this report.

Question: When using hierarchical clustering on the continuous features, how do the dendrogram cut-off points affect the number and nature of clusters formed?

Methods

The variables age, resting_BP, cholesterol, thalach (maximum heart rate achieved), old peak (ST depression induced by exercise relative to rest) are chosen as the predictors. A Hierarchical Agglomerative Clustering (HAC) model is made from the data, and a dendrogram is made off of that to find the best cut-off point for the clusters, and then each cluster is compared to each other to find any trends.

Results

As can be seen in Figure 1 the data is very suited for the HAC model as the data in the clusters are very cohesive (as the lines are small at the bottom), but there are cutoff points to make the clusters very separable (as the lines are large at the top). Personally, I choose to cut the data into 2 clusters as they are the most separable in the dendrogram. In Figures 1, 4, and 5 we can see that the people in Group 1 are most likely to be ~5 years older, and have ~80 more mg/dL of cholesterol but ~15 less maximum heart rate than Group 2

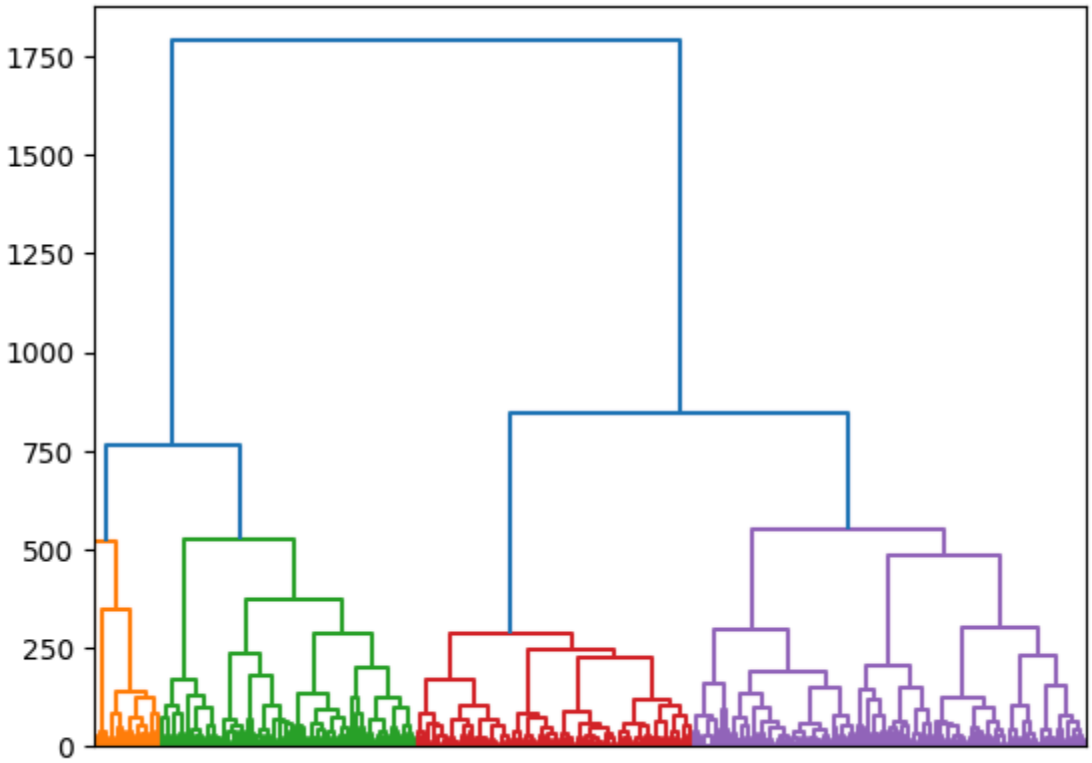

Figure 1: Dendrogram of the data

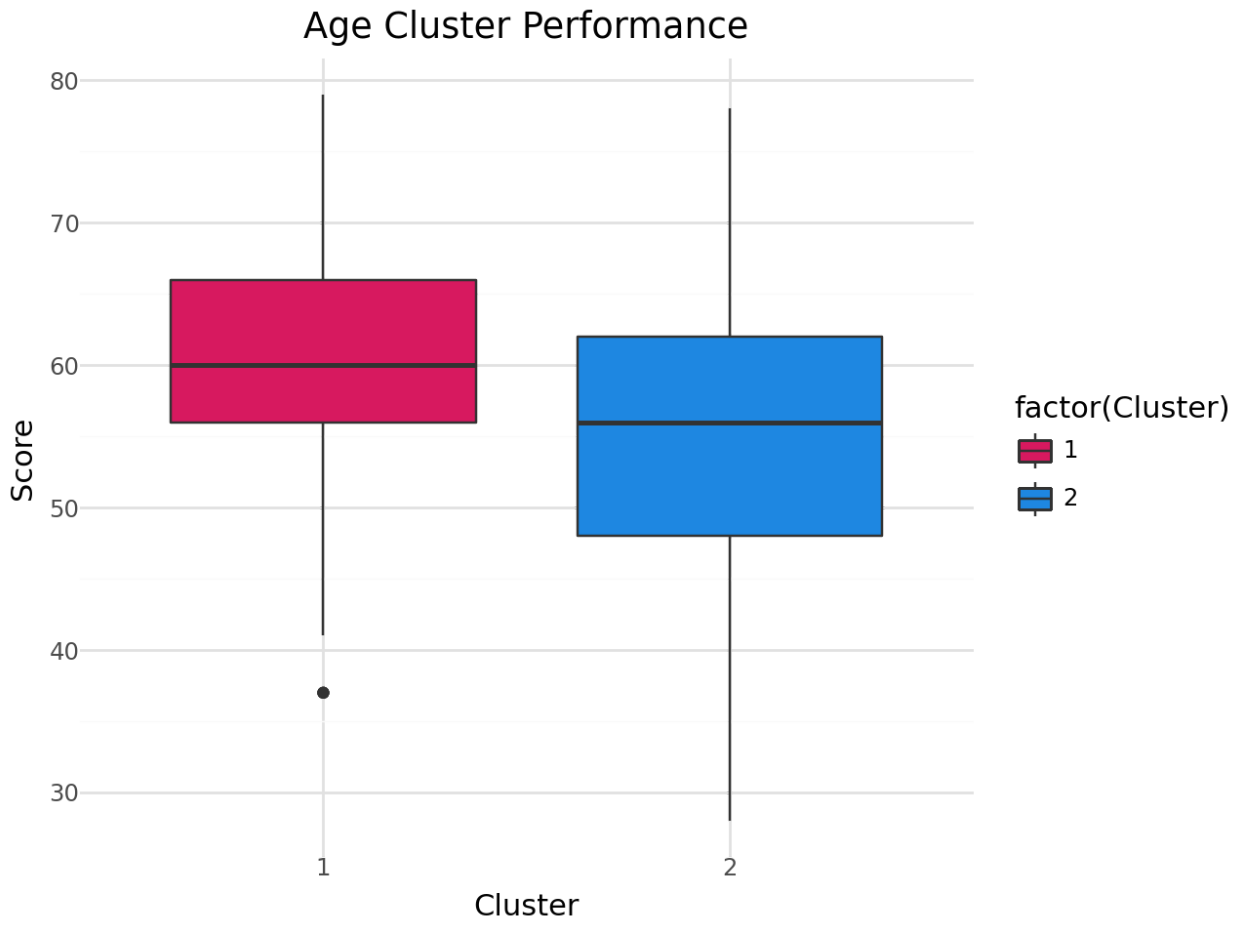

Figure 2: Cluster performance based on age

Figure 3: Cluster performance based on resting blood pressure

Figure 4: Cluster performance based on cholesterol level

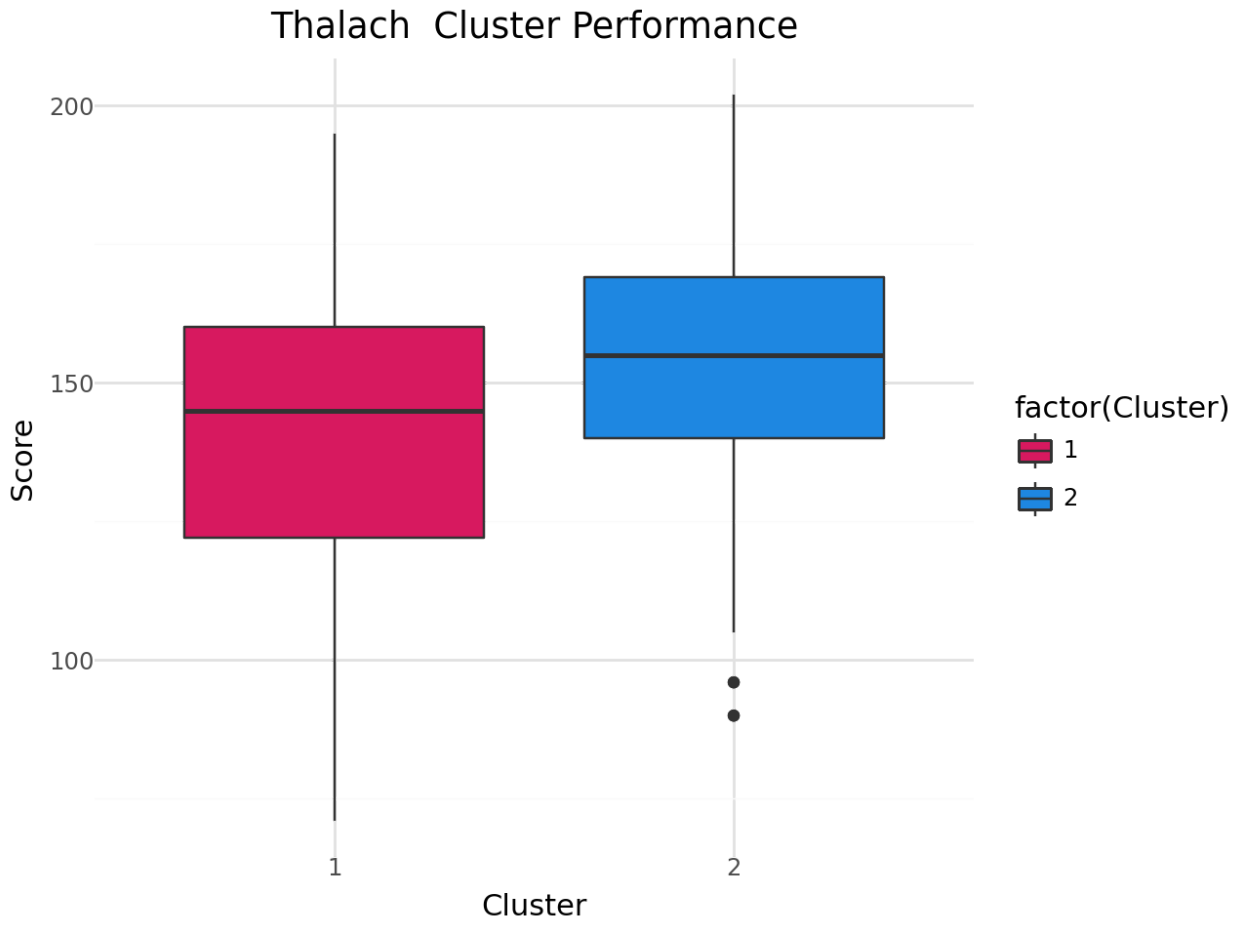

Figure 5: Cluster performance based on maximum heart rate

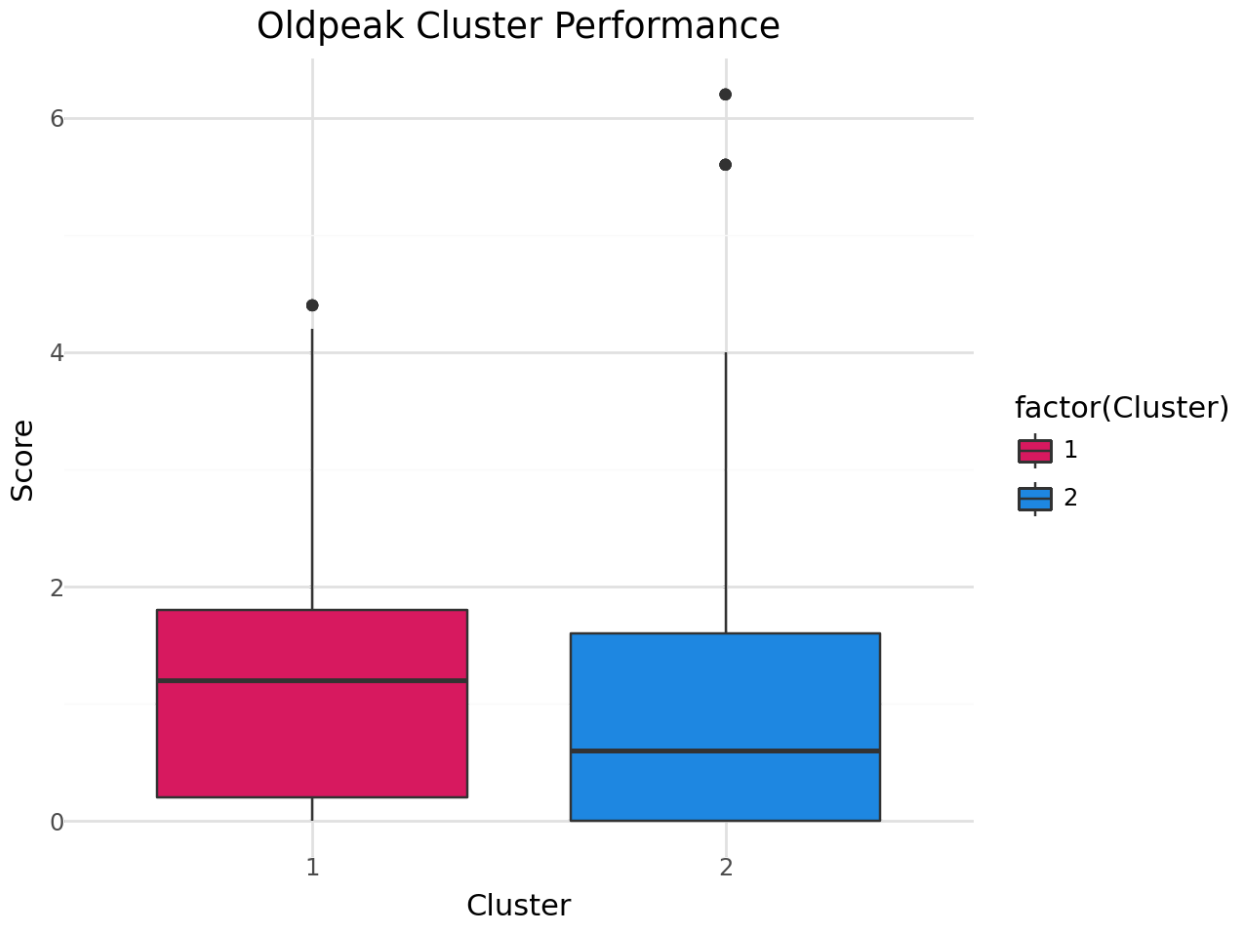

Figure 6: Cluster performance based on ST depression induced by exercise

Discussion

This model could be used to group people and determine the severity/likelihood of heart disease. If group 2 is found to be more likely to have heart disease, then people who do not have heart disease in group 2 could be given preventive treatment and increased monitoring.