Statistical Learning
Report
On
Home Assignment 3

HÖGSKOLAN
DALARNA

By
Manthri Bala Kiran
H19baman@du.se

Introduction:

For the task one we are been investigating the inferential properties of the linear regression model.

For the task two we have been given with Personality of 1,00,000 people and their personality is been tested on 5 main important things which consist of 50 statements. The participants could give a numerical response based on their agreement with the statement on scale of 1-5.

Dataset:

For the task 2, we have been given with the personality data of 1,000,000 individuals who have participated a test consisting of 50 statements which would be leading to 5 personality. For every statement, the participants would be giving a numerical response based on their personality and agreement with the statement on a scale of 1-5.

Initial Steps:

For the Task 2, I have loaded the Text file "five_personality.txt" to the R studio as a dataframe. In the dataset we have more than 1,000,000 observations with 50 variables and we do not have any null values in the dataset. As sampling I have taken nearly 20% of the data as my new dataframe for performing Kmeans clustering and 6% of my overall data for the testing the model for the Kmeans.

Methods:

Task 1:

1) A) $b_1 = \frac{S_{xy}}{S_{xx}}$ $b_2 = \frac{S_{uv}}{S_{uv}}$

$$b_1 = \frac{\sum (y_i - \bar{y}_i)(x - \bar{x}_i)}{\sum (x - \bar{x})^2}$$

$$b_2 = \frac{\sum (u - \bar{u})(v - \bar{v})}{\sum (u - \bar{u})^2}$$

when the numerator is considered we would be having $b_1 = b_2$

$$b_1 = \frac{\sum |y_i - \bar{y}_i||x - \bar{x}_i|}{\sum (x - \bar{x})^2}$$

$$b_2 = \frac{\sum |u - \bar{u}_i||v - \bar{v}_i|}{\sum (u - \bar{u})^2}$$

we can say that they are equal because they have the equal variances (x and u) & (y & v).

Task 2:

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them to clusters.

Kmeans Clustering:

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into kpre-defined distinct non-overlapping subgroups which are called as clusters. Where each data points belongs to only one group.

Hierarchical Clustering:

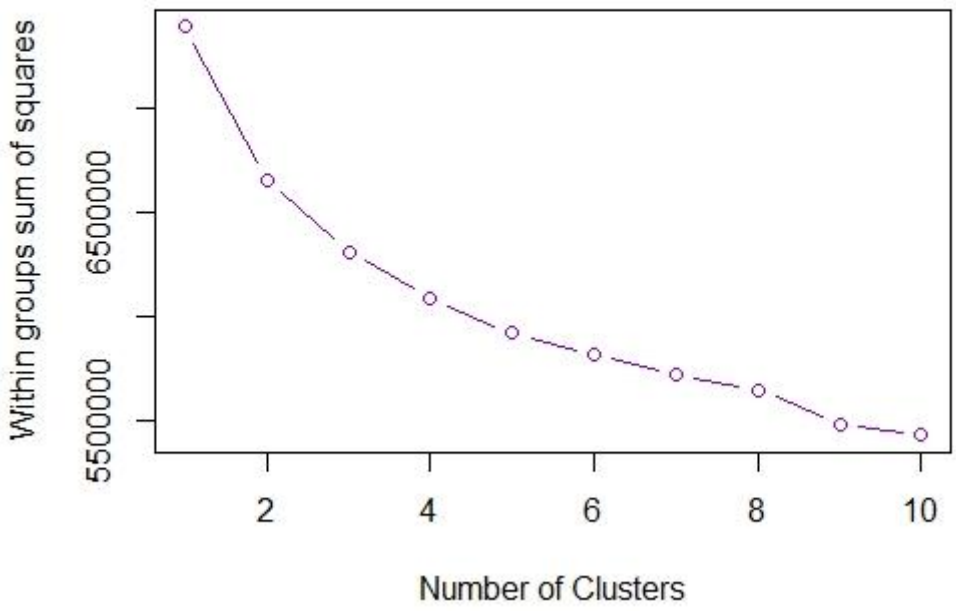
Hierarchical Clustering involves creating clusters that have a predetermined ordering from top to buttom.

Kmeans Clustering:

Kmeans is performed on the 50 statements for the five-personality data.

Clusters:

The number of the clusters have been chosen from the elbow plot below. We also know that we have five different personality in the data.



From the above plot, I have decided to choose the number of clusters K as 5 but k = 6 is also a best choice to choose.

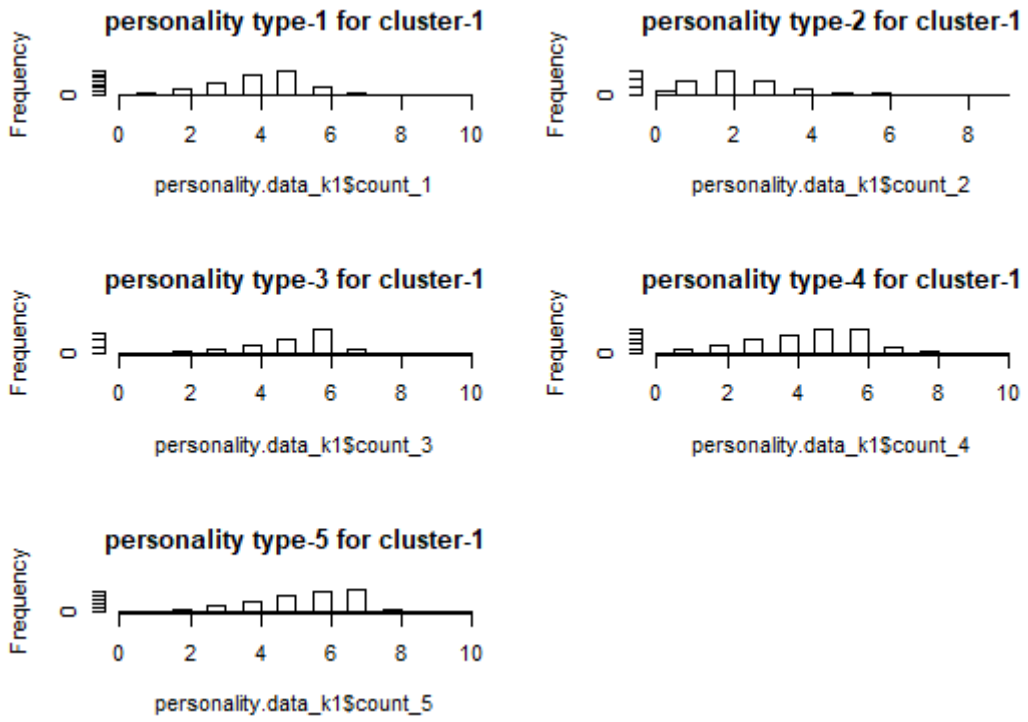
Hierarchical Clustering:

Hierarchical clustering analysis is performed for the data which consists of the personality of the participants. In Hierarchical clustering each data point is considered as a cluster and then are they are grouped as a single cluster based on the similarities of the features. The results are displayers in the form of dendrogram.

Results:

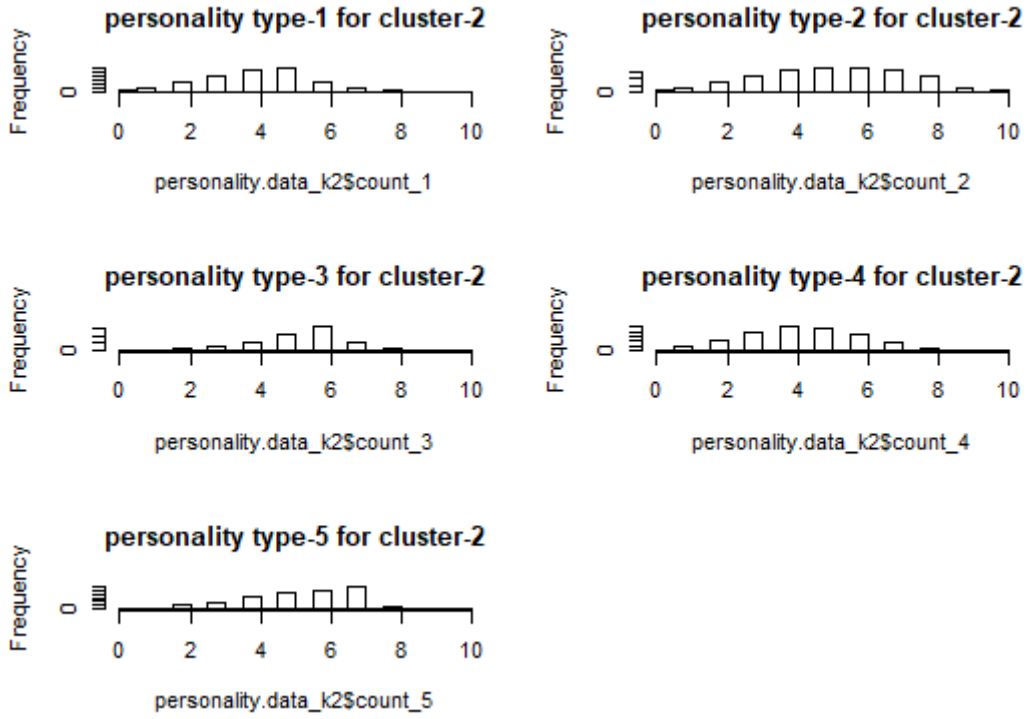
Kmeans Clustering:

For the cluster 1:



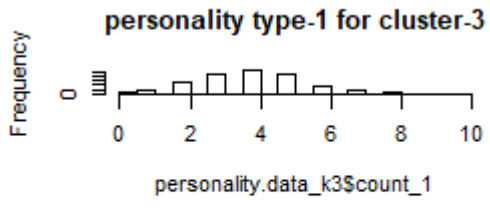
From the above plot we can see that the cluster 1 has been the mixture of type 4 participants as most people voted 4 and 5 for many statements.

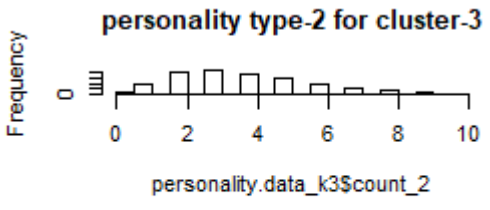
For the Cluster 2:

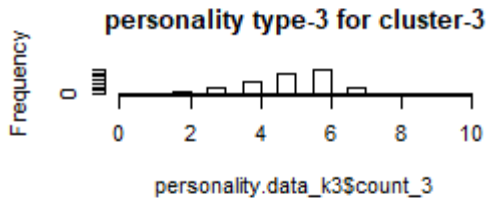


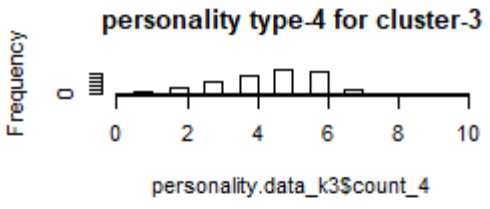
From the above plot we can see that the cluster 2 has been the mixture of type 2 participants as most people voted 4 and 5 for many statements.

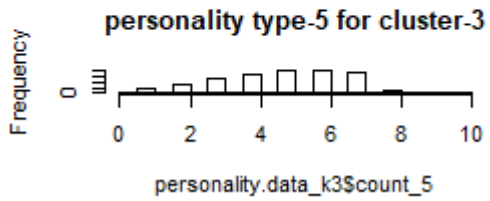
For the cluster 3:





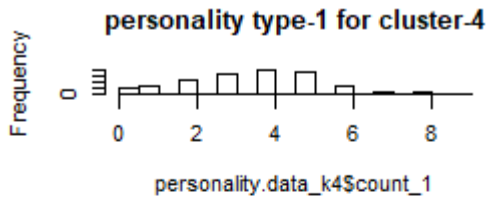


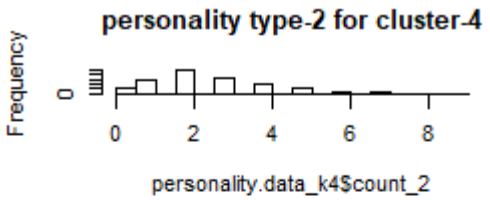


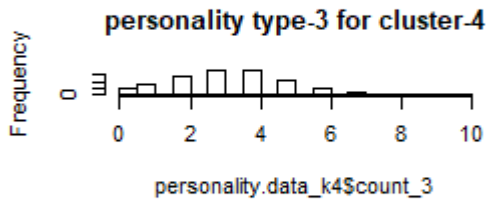


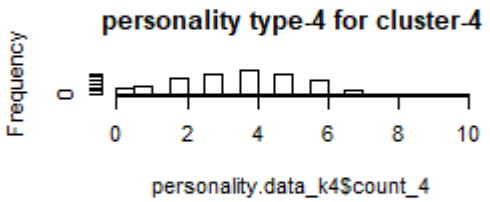
From the above plot we can see that the cluster 3 has been the mixture of type 5 participants as most people voted 4 and 5 for many statements.

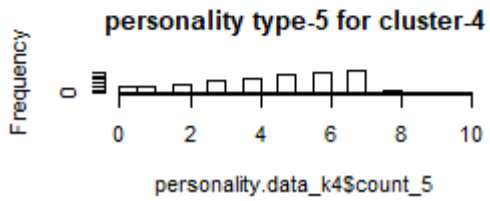
For the Cluster 4:











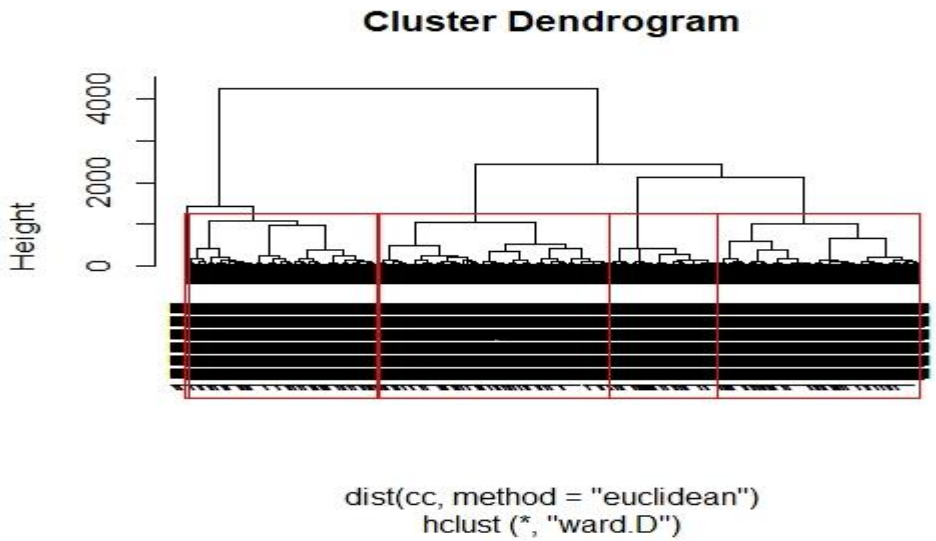
From the above plot we can see that the cluster 4 has been the mixture of type 1 participants as most people voted 4 and 5 for many statements.

Cluster 5:

From the above plot we can see that the cluster 5 has been the mixture of type 3 participants as most people voted 4 and 5 for many statements.

For the Prediction of kmeans, I have opted an accuracy of 15%.

Hierarchical:



I have decided with the optimal of clusters by visualizing the dendogram. The threshold is picked based up on the highest number of clusters after cutting the dendrogram. The distance metric used is “Euclidean” and the method is “ward D” through hclust.