Assignment 5

clustering: Clustering involves to identify a collection of data objects. It is an unsupervised machine learning technique which means we don’t know in advance, exactly what the group represents.

Clustering is also known as Segmentation.

There are two different types of clustering algorithms:

1. Partitional clustering: Partitional clustering takes a given dataset of objects and divides them into a pre-defined number of groups based on similarity
2. Hierarchical clustering: Hierarchical clustering acknowledges that natural hierarchies might exist within the data

K-means clustering: Kmeans clustering is an Iterative algorithm which divides a group of n datasets into k subgroups /clusters based on the similarity and their mean distance from the centroid of that particular subgroup.

Steps to achieve K-means:

* 1. Select the desired number of clusters k
  2. Select k initial seeds (often chosen at random)
  3. Calculate average cluster values (cluster centroids) over each variable (for the initial iteration, this will simply be the initial seeds)
  4. Assign each of the other observations to the cluster with the nearest centroid
  5. Recalculate cluster centroids (averages) based on the assignments from step 4
  6. Iterate between steps 4 and 5, stop when there are no more new assignments

About dataset:

Here I used  [ALS\_TrainingData\_2223.csv](https://unt.instructure.com/courses/58212/files/13298973/download?wrap=1) which contains 2223 rows and 101 columns and

[ALS\_TestingData\_78.csv](https://unt.instructure.com/courses/58212/files/13298972/download?wrap=1) which contains 78 rows and 131 columns.

Implementation of K- means clustering with the above data sets:

1) I have imported the required libraries to perform this analysis

2) Initially, I have imported the training data set

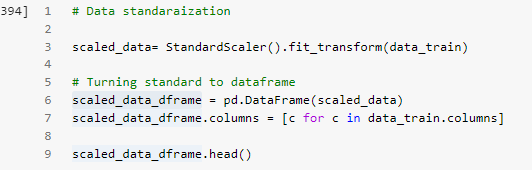
Data cleaning:

Checked for all the null values and removed them from the training dataset.

Data pre-processing:

* **Standardization** is an important consideration when performing cluster analysis
* Because similarity is measured in terms of distance, dimensions measured in large scales have a much larger effect.

I have standardized the data training data and converted the standardize data into a data frame



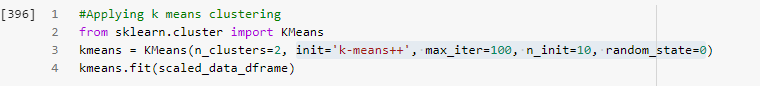
**preliminary visualization:**

**for preliminary visualization I have shown a scatter plot with** Age\_mean and ALSFRS\_slope

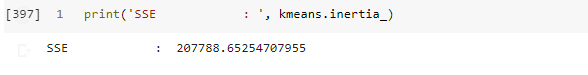


**Train a k-Means model on the data, experiment at least two different k values, and explain which k value is a better choice.**

I order to Train a K-means model, I am using sklearn.cluster library and taking K values as 2.



SSE: Sum of squared error is the difference between each observation and the mean of the cluster that is assigned to each object. According my learning increasing the number of clusters reduces the SSE.

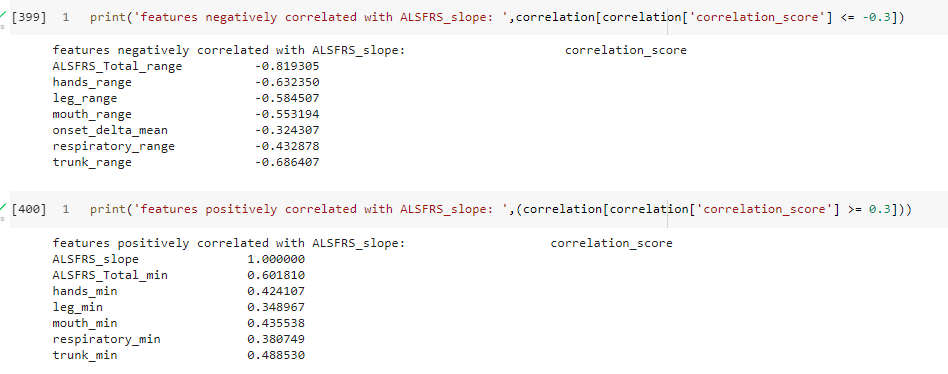


Aim: What patient phenotypes can be automatically and reliably determined and Predict the change of the ALSFRS slope?

For this I am comparing ALSFRS slope with other features in the data set

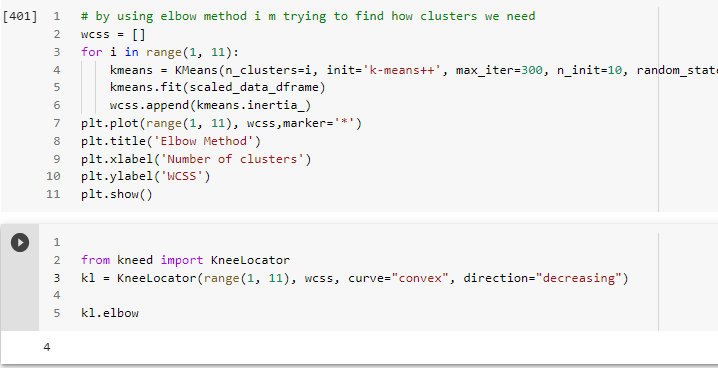


Based on a condition I am identifying the positive and negative correlation values id correlation score is >=,<= 0.3.



So, we can use this high correlated feature in this model.

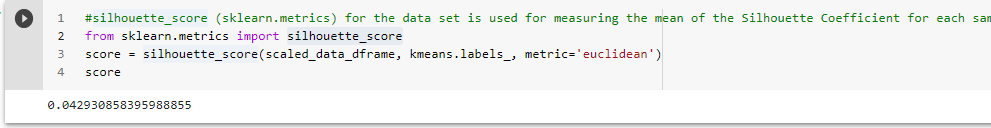
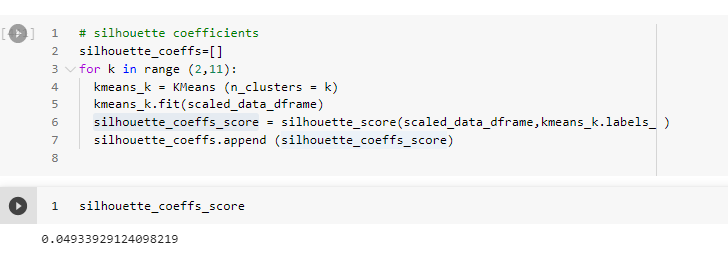
Before choosing another K value I want to know how many clusters are preferred for this data set. For that purpose, I am using Elbow method to determine the number of clusters.



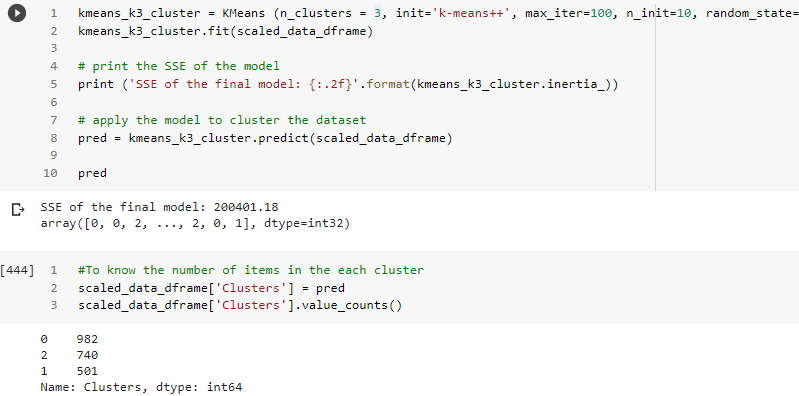
With in the range from 1 to 11 it shows 4 clusters are best choice, if the range is from 1 to 10 then it requires 3 clusters.

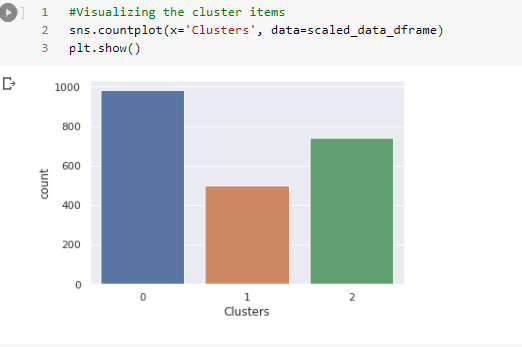
**Silhouette analysis**:

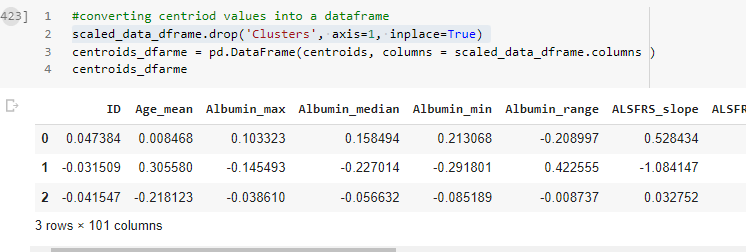
It is used to determine the degree of separation between clusters.

For each sample it determines the average distance from all the points in the same cluster and the average distance in the nearest cluster.  

I have checked my final model by taking the k value as 3 and number of items in the each cluster will be shown.



So, cluster 0 has the highest number of items



From the above screen shot , the centroid values for ALSFRS\_slope has the highest progress.

From the visualization graph we can see the three clusters are quite balanced.

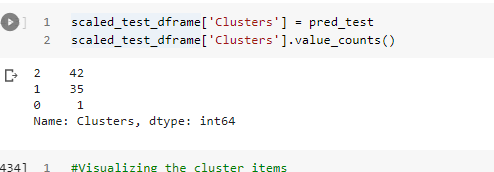
Now we can see the visualization for all the centroid values with respect the positive correlated features (I have taken 6):

Model evaluation:

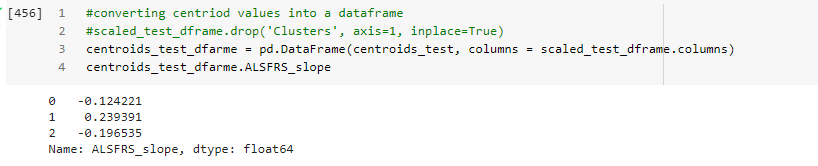
I am evaluting the model by using the testing data same as above tranining data methods.

For testing data , k=3



Visualization results as follows:





By observing above clustering grouping process, we can say that cluster 1 has fastest progress and cluster 0 and 2 has the slow progress.

Conclusion:

In order to predict the ALSFRS slope, it's difficult to draw conclusions about patient phenotypes. Based on our clusters, we can make some assumptions about three patient phenotypes.

We discovered that trunk range and albumin, as well as other physical movement abilities such as hands\_min, legs\_min, mouth\_minmay be the best candidates for predicting ALSFRS slope. Age does not appear to affect the ALSFRS slope.