This case-study examines the patterns, symmetries, associations, and causality in a rare but devastating disease, amyotrophic lateral sclerosis (ALS). A major clinically relevant question in this biomedical study is: What patient phenotypes can be automatically and reliably identified and used to predict the change of the ALSFRS slope over time? This problem aims to explore the data set by unsupervised learning.

* Load and prepare the data.
* Perform summary and preliminary visualization.
* Train a k-Means model on the data, experiment at least two different k values, and explain which k value is a better choice.
* Evaluating the model performance by report the center of clusters.
* Visualize the final clustering result.

**SOLUTION:**

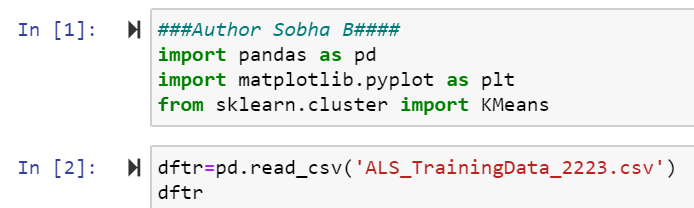
Attached is the Python code generated in Jupyter NB.

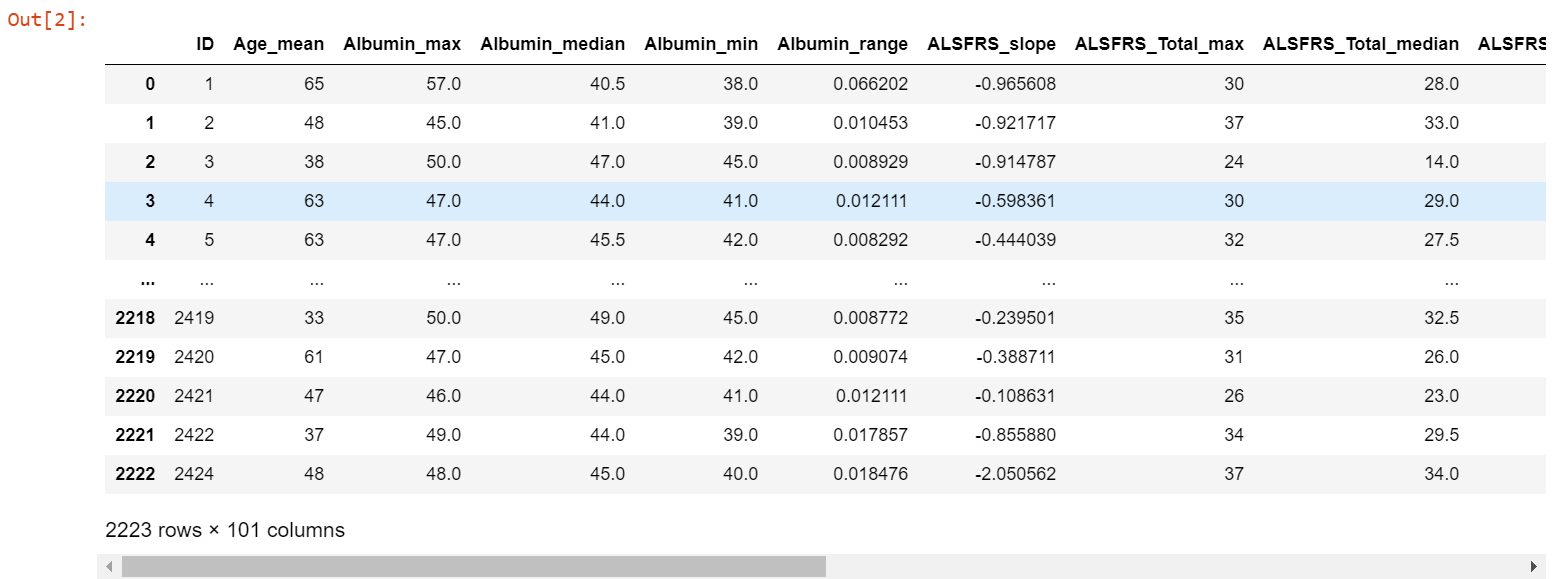
**K-Means (n=3) [KNN(k=3).ipynb]:**

k-Means is the unsupervised machine learning algorithm which is used to identify the pattern of the data without referring to a known outcome.

**Read the data from CSV file and IMPORT the libraries:**

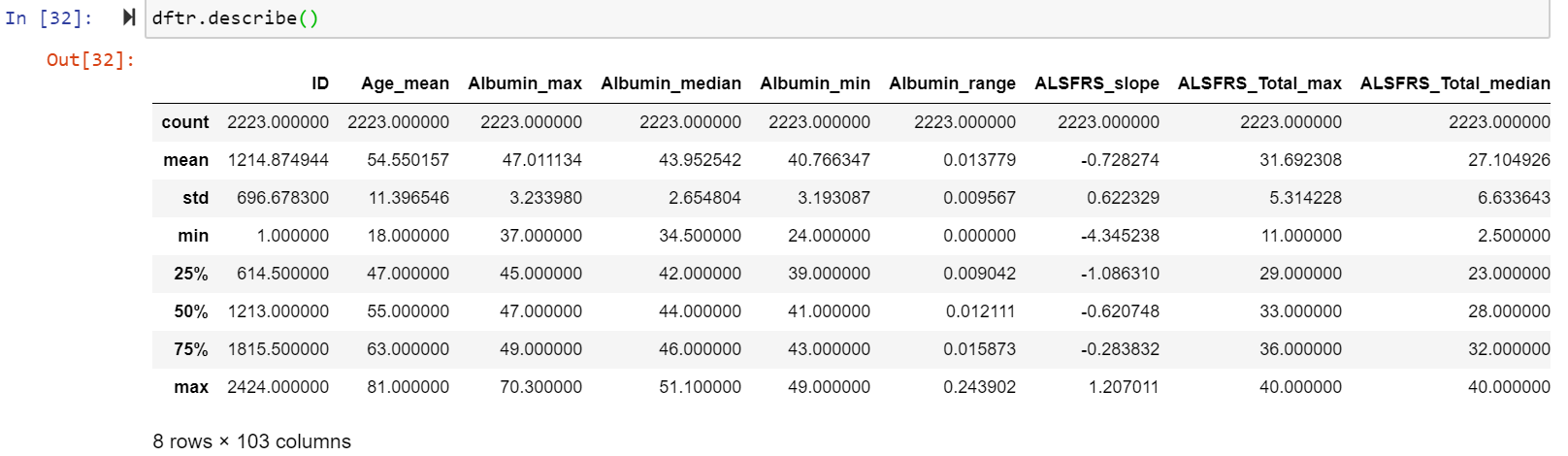
* As per the task requirement, I have imported the required libraries and read the ALS\_TrainingData\_2223.csv file. From which, I have read all the variables.



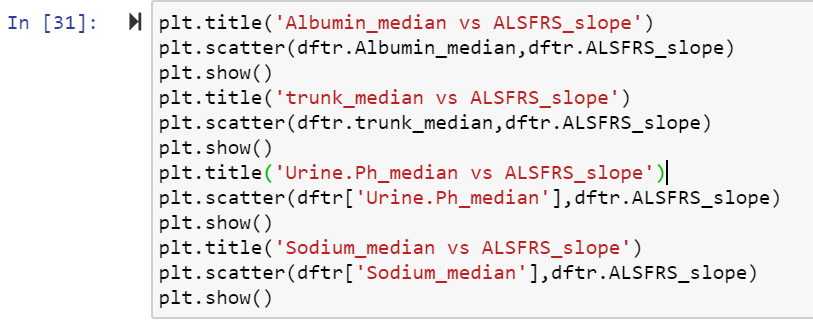


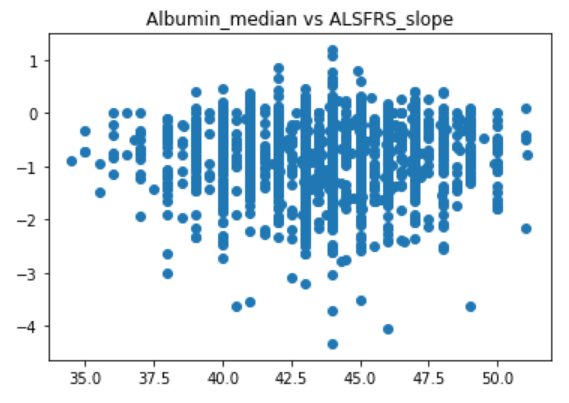
**Summary and visualizations:**

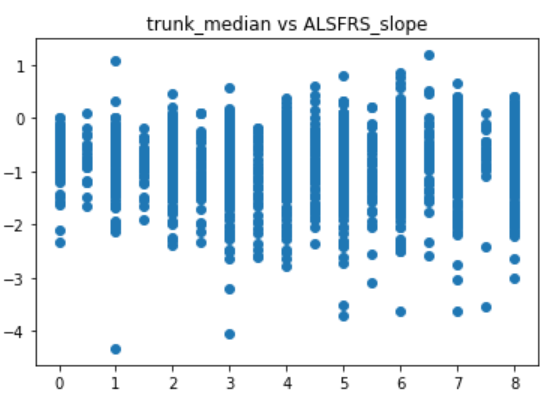
* I have created summary statistics using describe () function for the dataset.
* Below is an example of how program and output look.

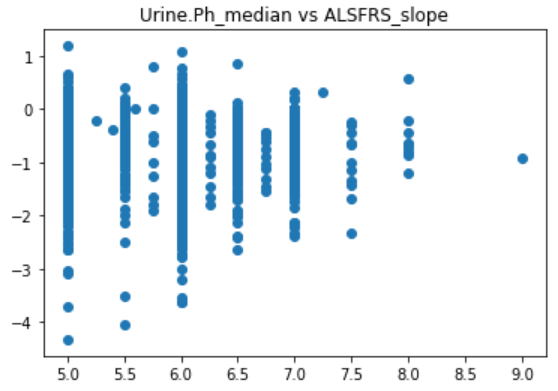


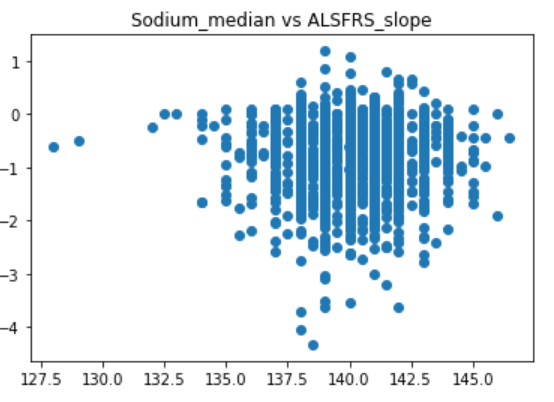
* I have also created scatter plot visualizations for 4 of the variables vs ALSFRS\_slope, to know the change in ALSFRS\_slope with the other phenotypes.







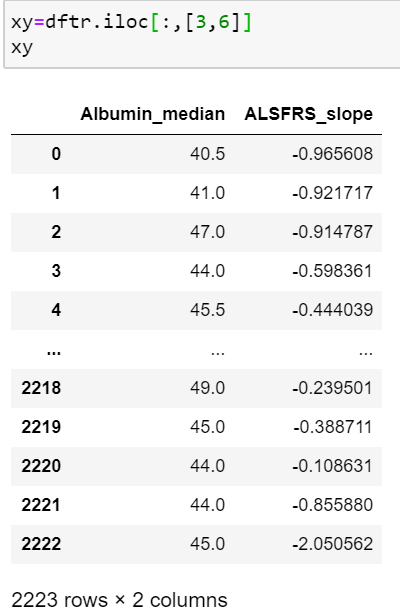




* Based on the above analysis, I have decided to consider the Albumin\_median vs ALSFRS\_slope as the scatter plot shows some kind if relation between the two variables than can be evaluated.

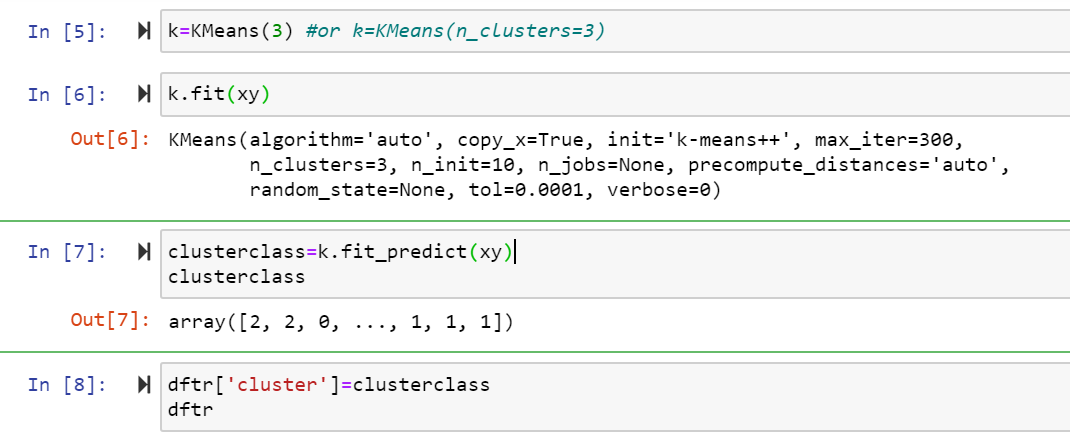
**Consider K=3:**

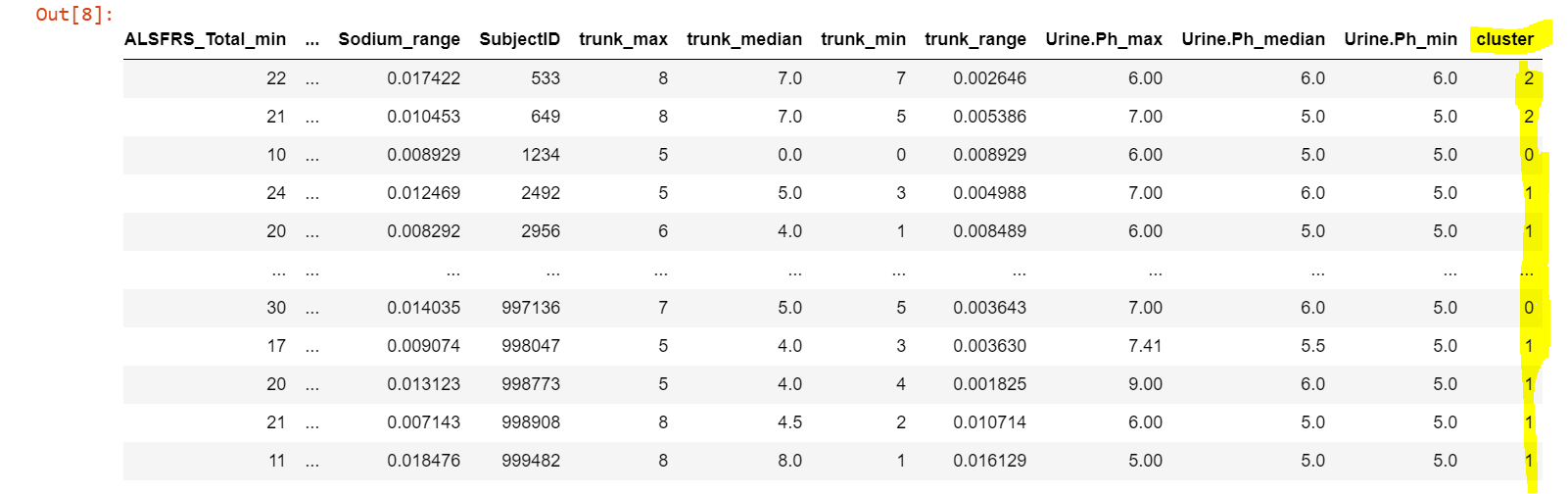
**Step1: Consider the two feature variables (**Albumin\_median and ALSFRS\_slope).



**Step2: Select K:**

* I used scikit learn, to create K means clusters.
* First step is to use KMeans function to mention no.of clusters.
* Then fit the model and predict the cluster class. Using this cluster class we are going to divide the clusters.

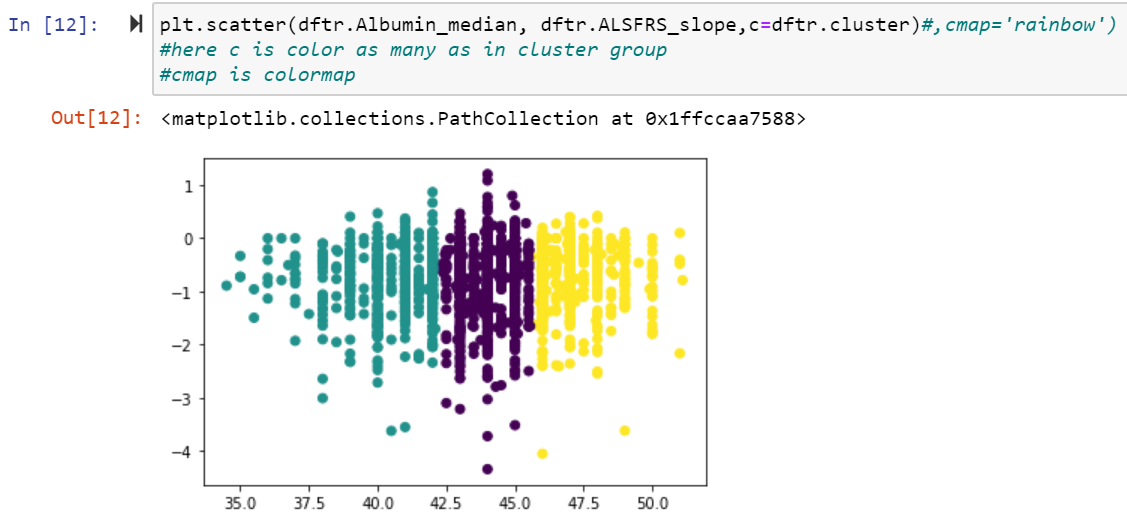




* Now, I have created a variable “cluster” which has the numeric values ranges from 0 to 2 as I have selected k=3.
* Once created, I have added this variable to the DFTR data using the code mentioned under ‘In [8]’.

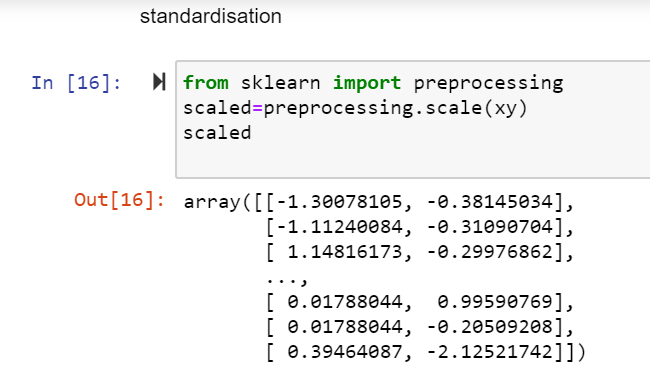
**Step3: Create clusters using scatter plot:**

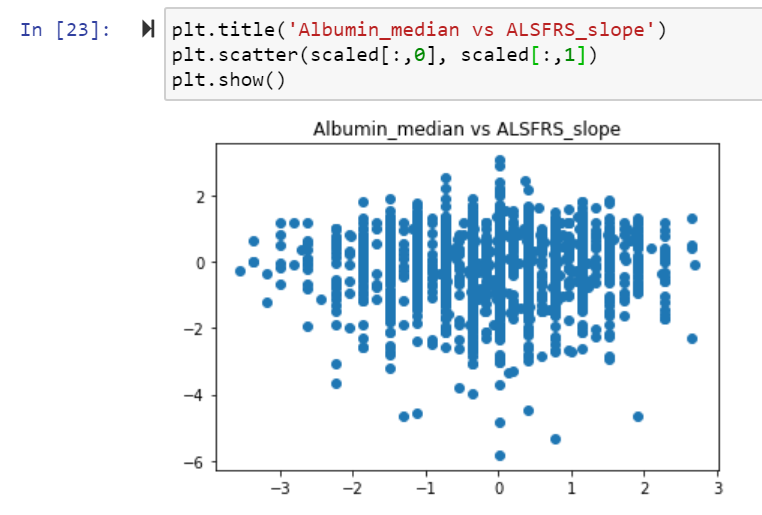
* Using the cluster variable, I have created a scatter plot that has clusters.



**Step4: Standardisation:**

* If you see in the above scatter plot, the y-axis values are very small when compared to the x-axis values (scale), hence the clusters were formed based on x-axis as much priority given to the Albumin\_median as it has huge values when compared to the ALSFRS\_slope.
* Hence, to nullify this, we need to standardise the variables. I have used the sklearn to standardise the variables.

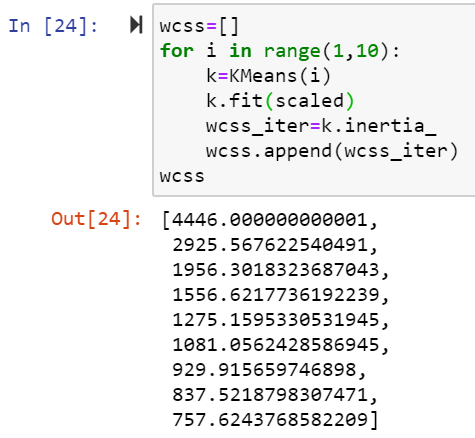


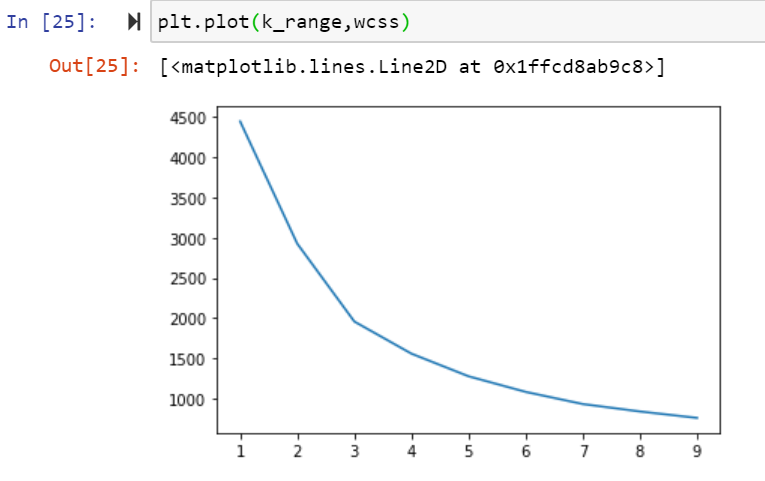


* So, the new variables are in the scaled data frame now.

**Step5: Find the optimum K value using elbow technique:**

* As the data is standardised now, find the optimum k-value.
* Inorder to find the optimum k values, we need to find the distance between the points, i.e., wcss (within cluster sum of squares).
* I am going to use “inertia\_” function to find the wcss. Then plot the scatter plot using wcss and k-value.

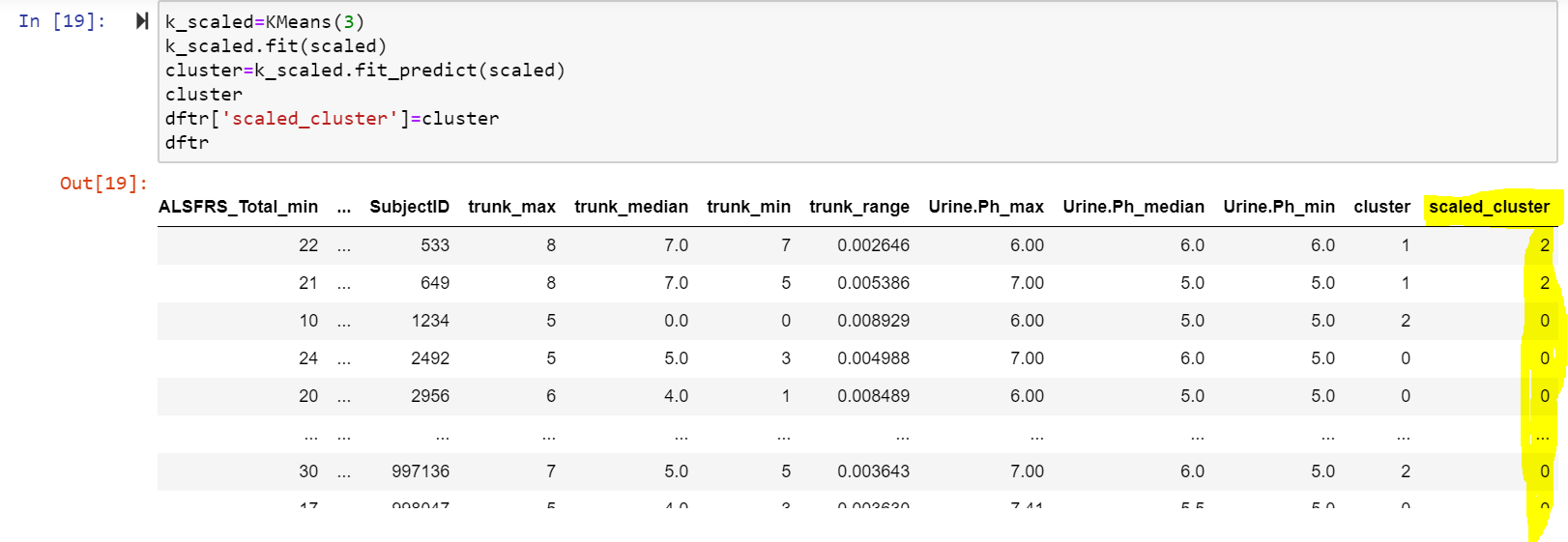




* Based on the above plot, we can say that after k=3, there is no much change on the wcss.
* We can consider 3 as optimum k value.

**Step6: Create cluster based on standardised data and do scatter plot:**

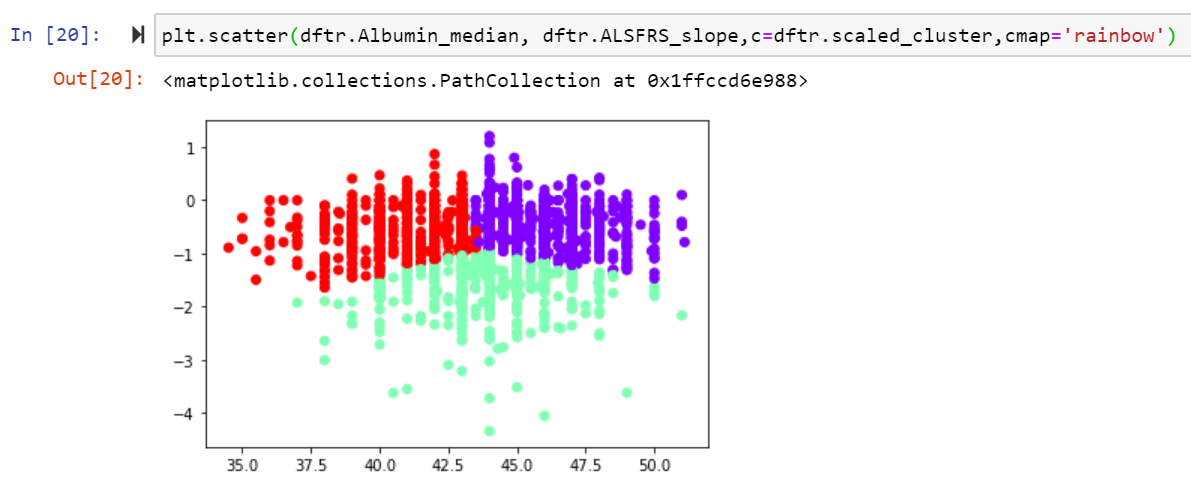
* Now, I have followed the same steps as I did in step2 and created the new cluster classification based on the standardised data, i.e., (scaled\_cluster).



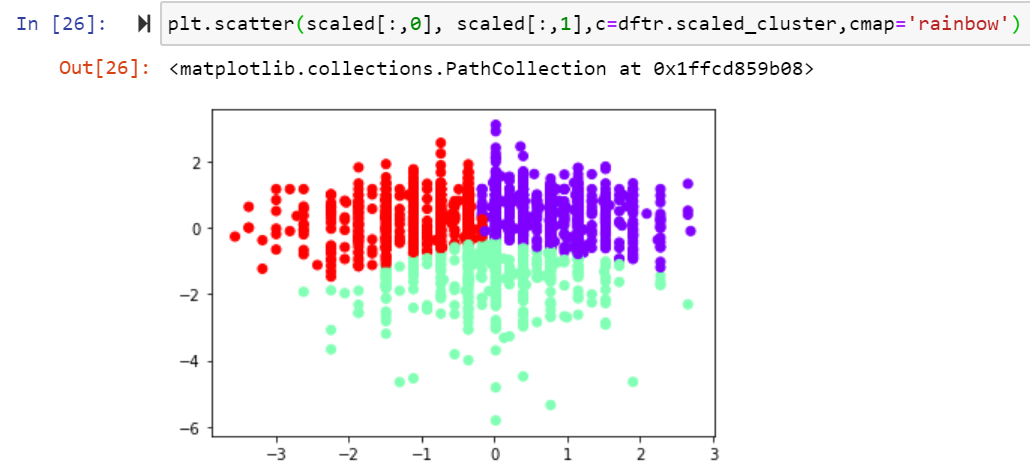
* Then, I have added this newly created variable to the DFTR dataframe.

**Step7: Create new clusters from standardised data using scatter plot:**

* This is the scatter plot of original data (or original scale) but using the new clusters based on standardised data.

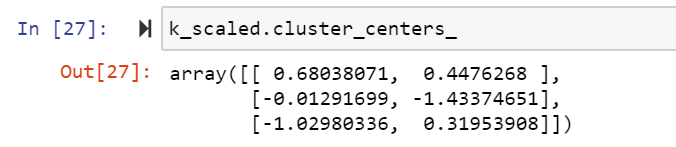


* This is the scatter plot of standardised data using the new clusters based on standardised data.
* You can identify the scale difference to that of the above plot.



**Evaluating the model performance by report the center of clusters:**

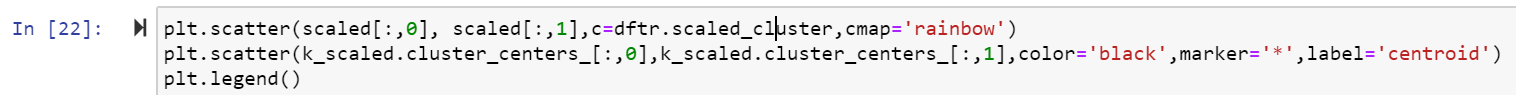
* The center of clusters can be identified by using “cluster\_centers\_” function from sklearn.

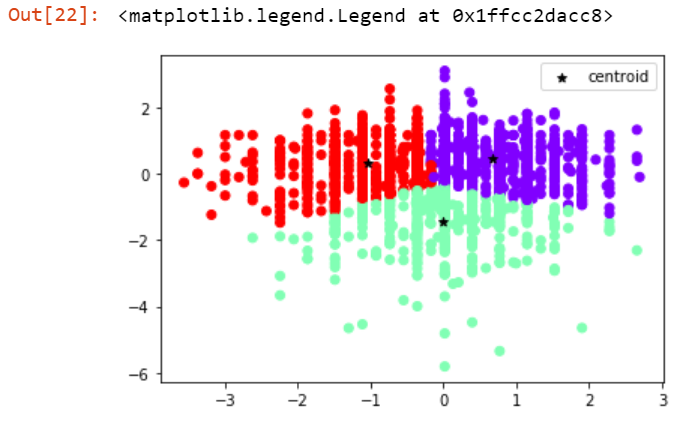


* Above are the centers for the 3 clusters.

**Visualize the final clustering result:**

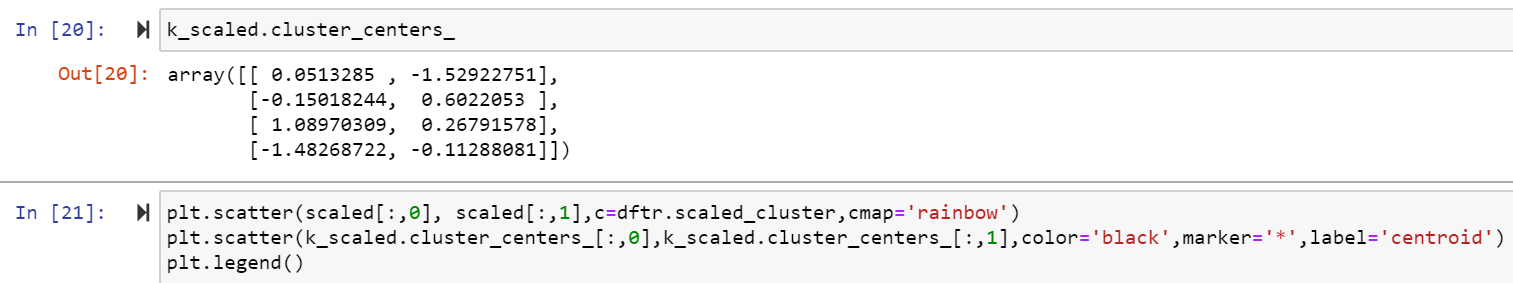
* Now, the final plot with the 3 clusters can be seen even with the center of clusters for the standardised data.

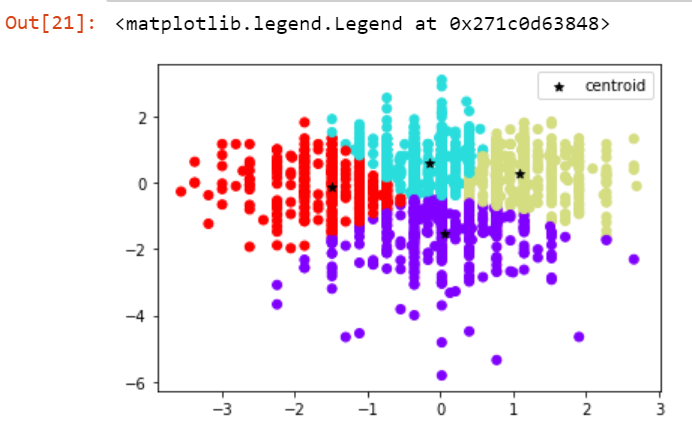




**K-Means (n=4) (HW3\_Sobha\_k=4.ipynb):**

* I have followed the entire process as I did for **K-Means (n=3),** but selected k=4 and below are the results.





**Conclusion:**

* By looking at the k=3 and k=4, I conclude that the k=3 is the optimal number for the Albumin\_median and ALSFRS\_slope clusters over k=4, as with the x and y axis scale the k=3 scatter plot clusters has certain definitions when compared to the k=4 scatter plot.
* K=3 scatter plot defines, when ALSFRS\_slope <=~-1 there is no much disease condition, however with the increase in the ALSFRS\_slope over -1 and under Albumin\_median <=0 categorises as one category and >0 as another category. We attribute some patterns based on this cluster when compared to the k=4.
* Elbow technique also suggests, =4 as optimal number and beyond that there is not much decrease in the wcss.