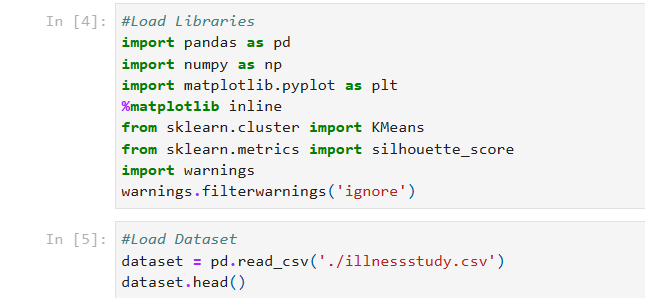


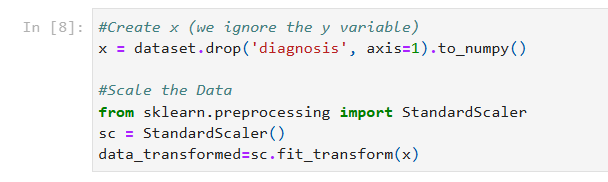
|  |  |
| --- | --- |
| Name | Kavinraj Muthuvel |
| Student ID | 100995545 |
| Subject | Introduction to Data Analytics |
| Date | November 25, 2024 |



**Libraries Imported:**

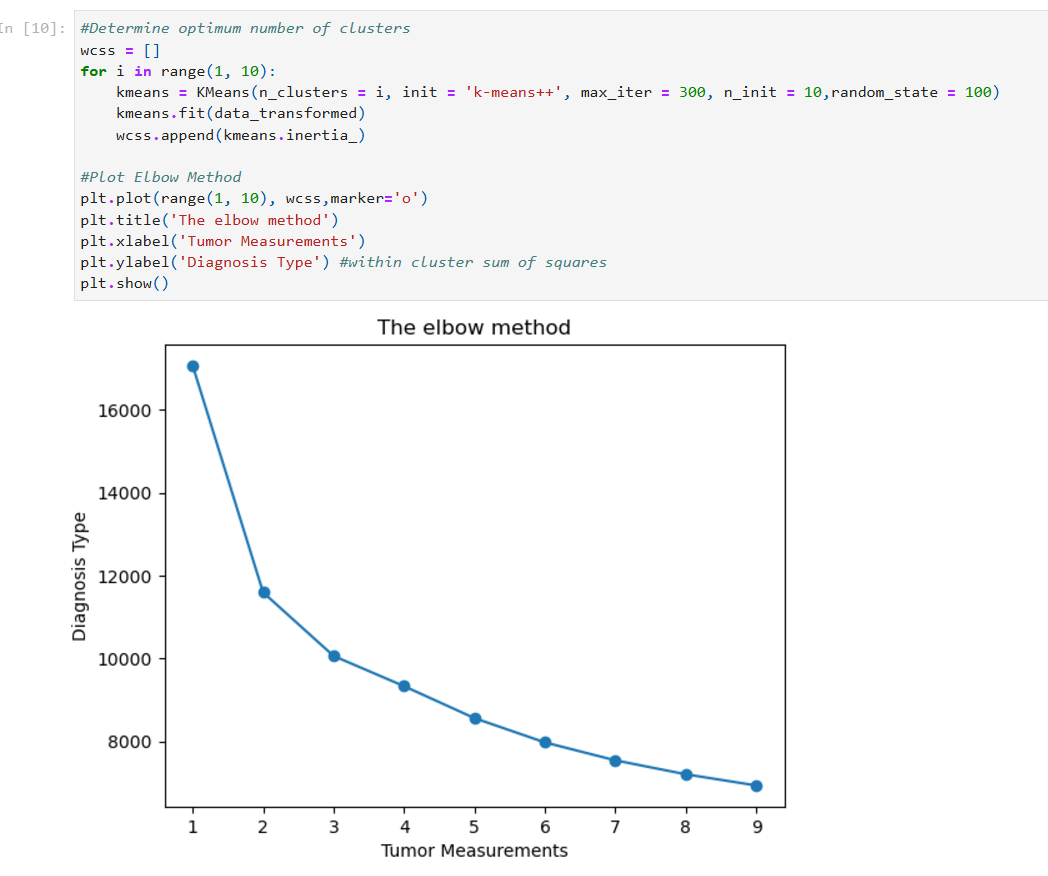
* pandas and numpy are used for data handling and numerical computations.
* matplotlib.pyplot helps visualize data.
* KMeans (from sklearn) performs clustering.
* silhouette\_score evaluates the quality of clusters.

**Dataset Loaded:** The file illnessstudy.csv is read into a pandas DataFrame. The first five rows (head) of the dataset are displayed.



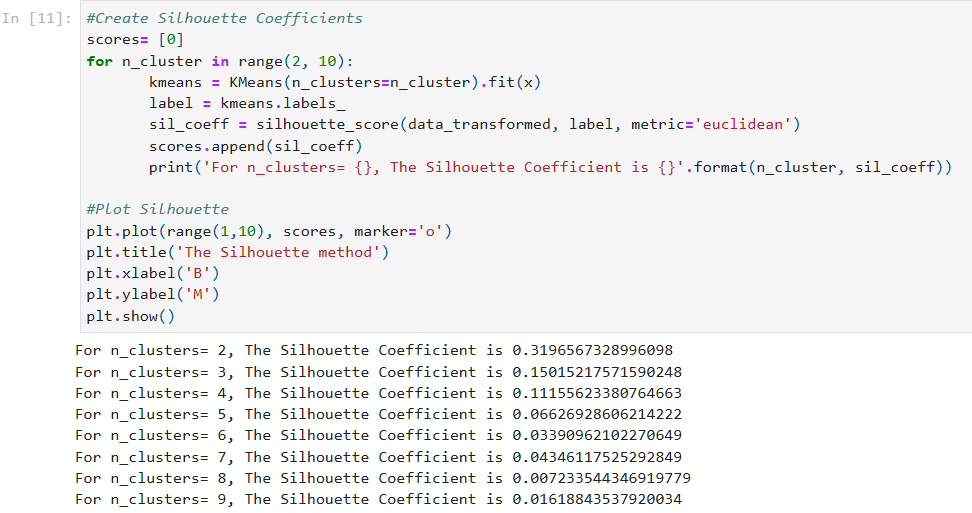
**Drop Diagnosis Column:** The diagnosis column (target variable) is dropped to focus on clustering based on feature values. The data is converted into a NumPy array for easier processing.

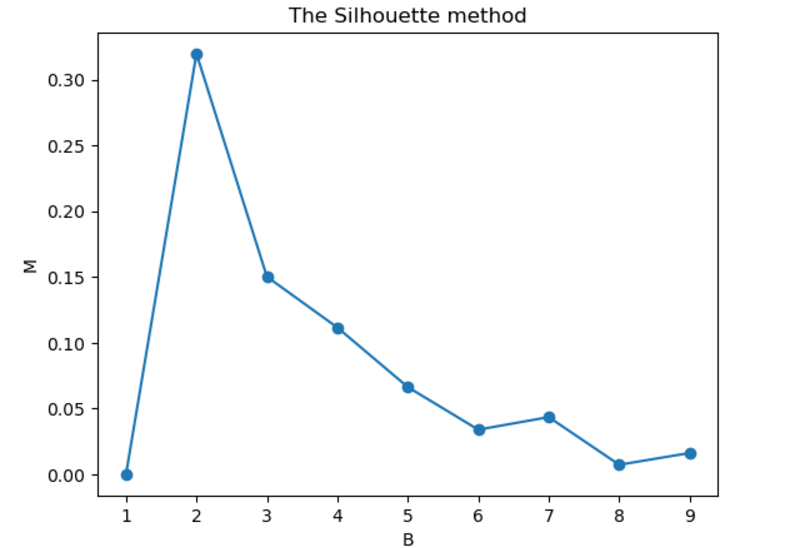
**Scaling Features:** StandardScaler standardizes the data by scaling each feature to have a mean of 0 and a standard deviation of 1. This ensures all features contribute equally during clustering.



1. **Within-Cluster Sum of Squares (WCSS):** For each cluster size (i from 1 to 9), WCSS is calculated, representing the compactness of the clusters.
2. **Elbow Method:** A line plot is generated to identify the "elbow point," where the WCSS sharply decreases and then flattens, indicating the optimal number of clusters.

**Output:**  
An elbow plot with WCSS on the y-axis and cluster numbers on the x-axis. The elbow is at **2 clusters**, suggesting two meaningful groups.

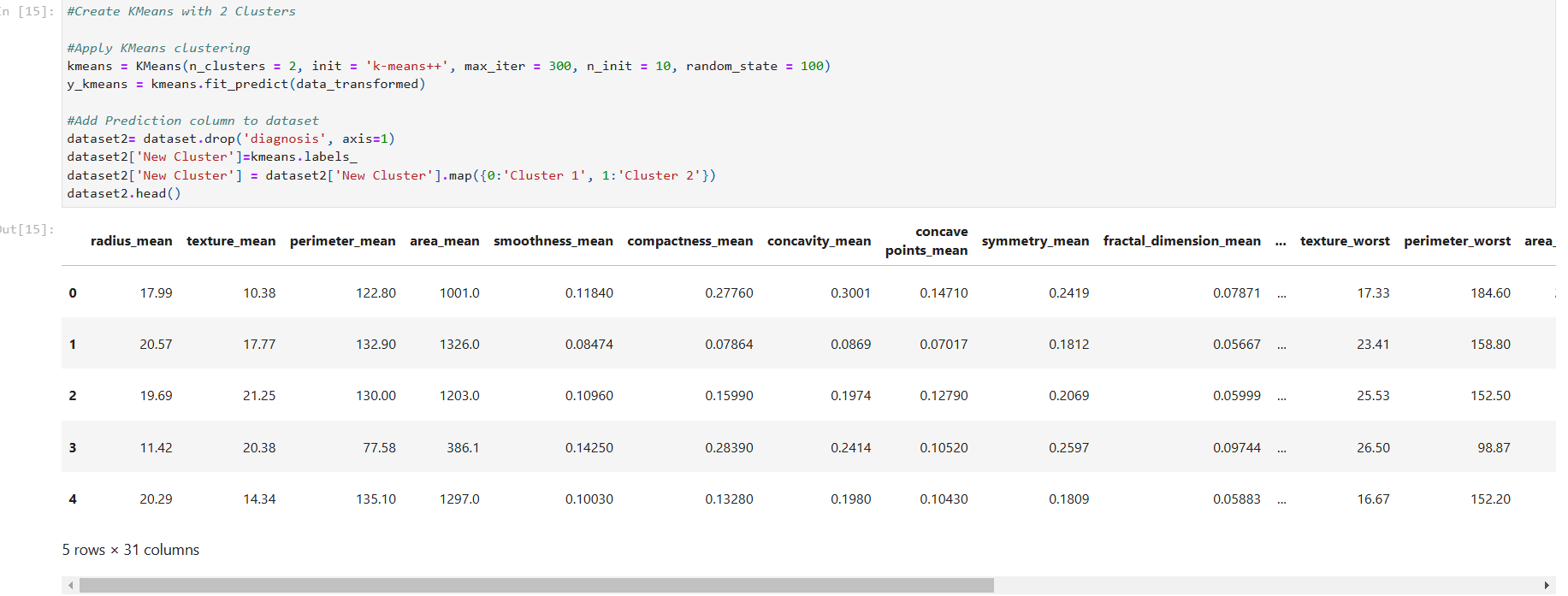




1. **Silhouette Coefficient:** Measures how well-separated the clusters are. A higher value indicates better-defined clusters.
2. **Loop Through Cluster Numbers:** The silhouette coefficient is computed for cluster sizes from 2 to 9. The results are plotted to visualize the best number of clusters.

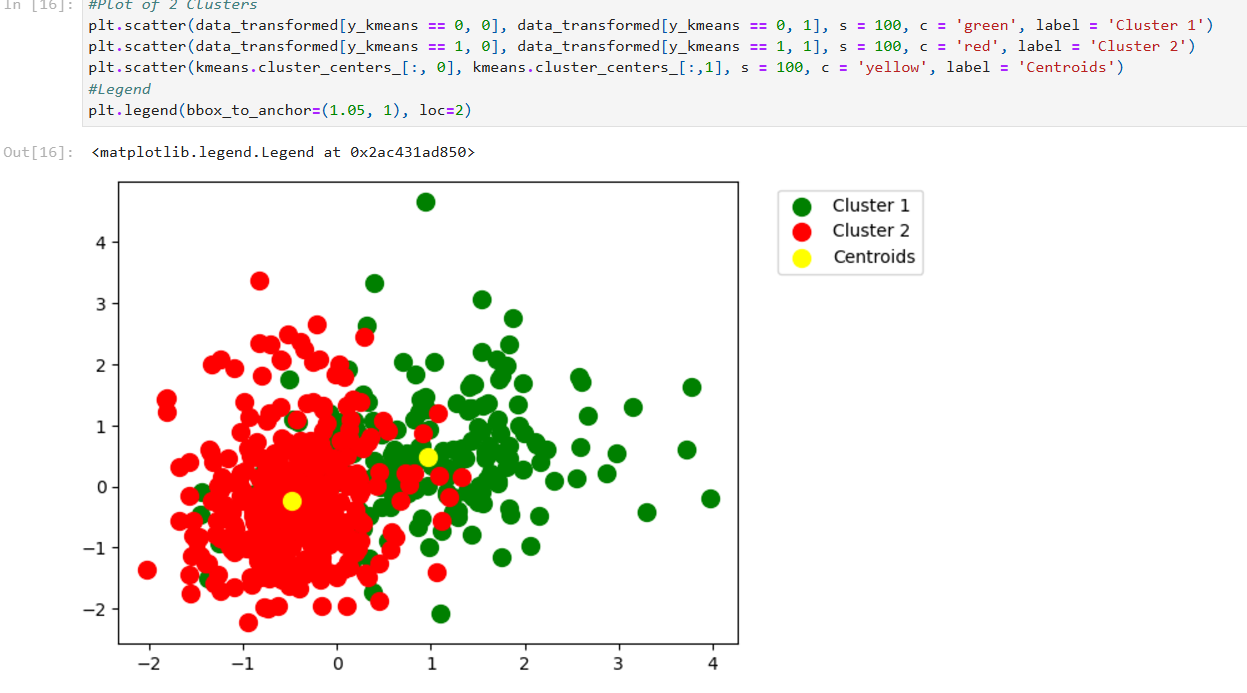
**Output:**

* The silhouette coefficient is highest for **2 clusters (0.319)**, confirming the choice from the elbow method.
* The coefficient decreases with more clusters, showing diminished cluster clarity.



1. **Fit KMeans:** KMeans is applied with 2 clusters (optimal number identified earlier). Cluster labels are assigned to all data points.
2. **Add Cluster Labels:** A new column, New Cluster, is added to the dataset to store the cluster assignments. These clusters likely correspond to malignant and benign tumors.

**Output:**  
The first five rows of the dataset with the new column New Cluster, indicating whether each tumor belongs to cluster1 or cluster 2.



1. **Scatter Plot:** Each cluster is represented by a color (green for Cluster 1 and red for Cluster 2).
2. **Centroids:** Cluster centroids are shown in yellow, representing the average position of each cluster.

**Output:**  
The scatter plot visually shows two well-separated clusters with their centroids, confirming the effectiveness of the clustering.