Report

Fidan Nabiyeva

November 7, 2023

1. **Abstract**

This project report outlines the comprehensive process of conducting Exploratory Data Analysis (EDA), preprocessing, and the implementation of various clustering algorithms on a customer dataset. The primary goal is to gain insights into customer behavior and preferences through clustering techniques, specifically focusing on K-means, K-prototype, and Agglomerative clustering.

1. **Introduction**

In the contemporary landscape of data-driven decision-making, businesses wield the power of analytics to navigate the complexities of customer behavior. For example, finding a relationship between people age and their income, how they related to their education, where they live or kind of information. Our focus in this investigation lies within the realm of customer data, a rich tapestry woven with variables such as age, income, marital status, occupation, settlement size, education, and sex.

By leveraging unsupervised learning techniques, specifically clustering, we aim to uncover natural groupings and relationships within the multifaceted landscape of customer data. The goal is to discern meaningful segments, each characterized by unique combinations of age, income, marital status, occupation, settlement size, education, and sex.

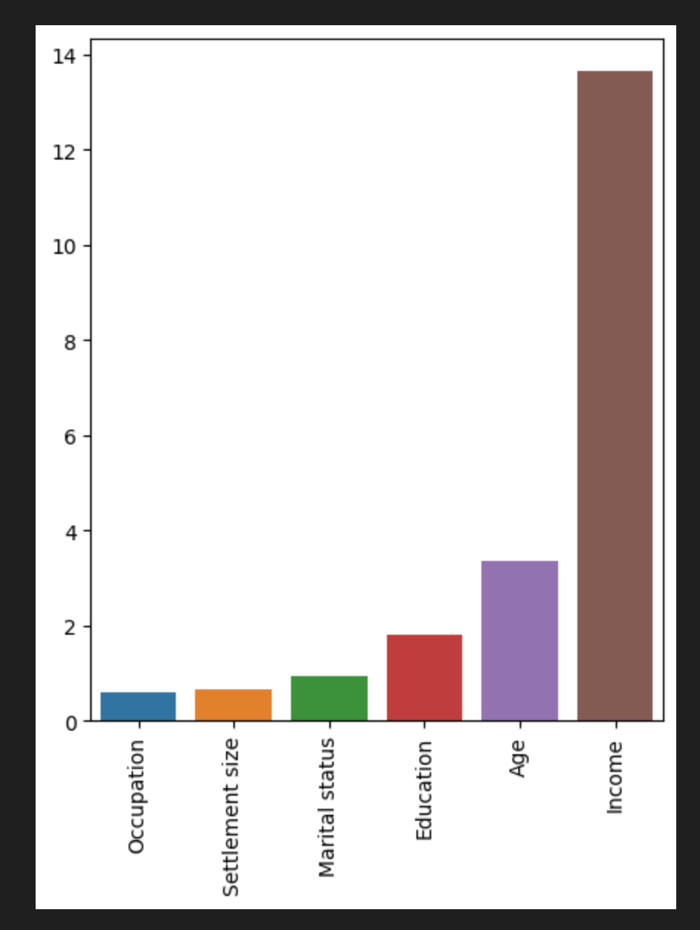
In the next steps we will talk about how we prepare our data to modelling, how we apply these algorithms, what kind of relationships we can see.

1. **EDA**

Our data set consist of 9 features and 2000 rows. The ‘Unnamed: 0’ and ‘ID’ columns are same as index. For that reason, we will not use them, so we can drop them. There will be 7 features: 5 of them are consist of categorical data and 2 of them numerical.



There are null values in ‘Marital status’, ‘Age’, ‘Education’, ‘Income’, ‘Occupation’, ‘Settlement size’. Maximum null values are in ‘Income’ that 273. The minimum are in ‘Occupation’ that 12.



The percentage of null values

While looking at data we can clearly see that there are some misspelling words in ‘Marital status’. So we are going to cheeck uniquie values of ‘Marital status’ and other categorical columns to see if we have any other misspelling word or same meaning words.

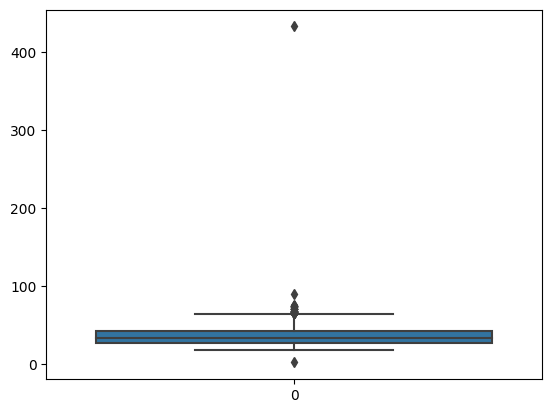
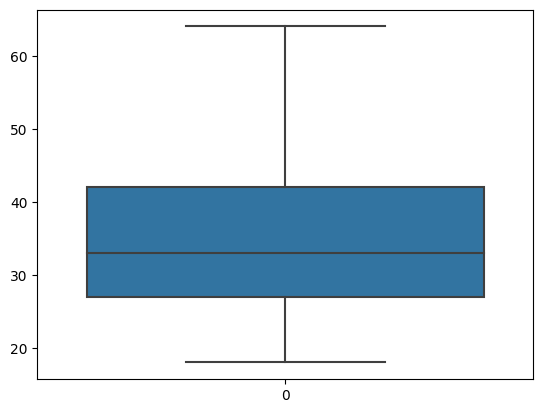
Also with the help of scatterplot we clearly see that there are plenty outliers in our ‘Age’ and ‘Income’ columns.

1. **Preprocessing**
   1. ***Fixing Invalid values and null values***

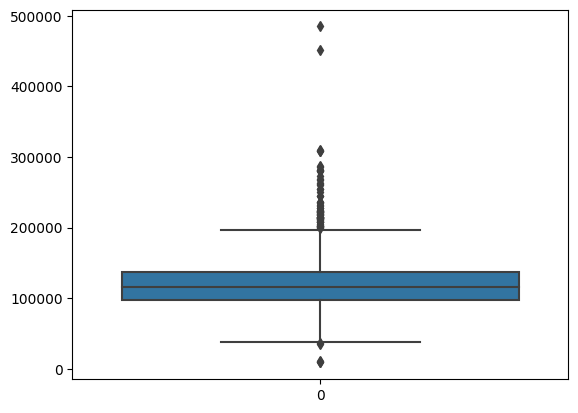
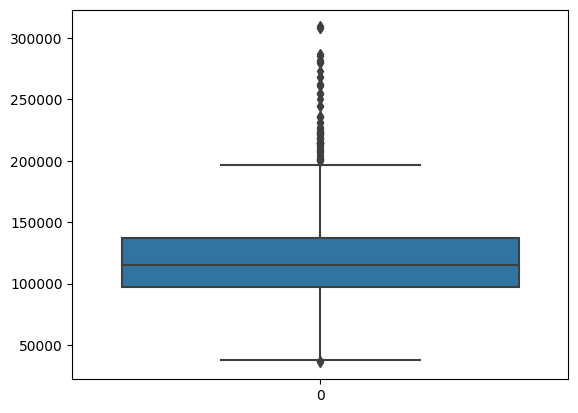
In first step misspelled words or invalid values is fixed. We have done changes in ‘Marital status’ and ‘Settlement size’. After that we do Null value and Outlier cleaning. While Null values of categorical columns is filled with most frequent values in special conditions. And null values of numerical data is filled with mean values.

* 1. ***Outlier cleaning***

‘Age’ column has 3 big outliers which age is 90, 433 and 2 which are dropped. The rest of values which greater than upper limit are considered as retire age and are replaced with upper limit of the boxplot that calculated with IQR(Inter Quantile Range).

In ‘Income’ we dropped big values and small values that are far away from rest of outliers. But rest of outliers was stayed because people can have variety amount of income.

* 1. ***Normalization and Encoding***

For doing normalization data is copied. Standard Scaler is used for normalization.

Label and One-hot encoding applied to categorical columns based on that they are ordered or not. Created 3 copies of data which are: normalized, encoded, encoded and normalized.

Data is ready for modelling.

1. **Clustering**

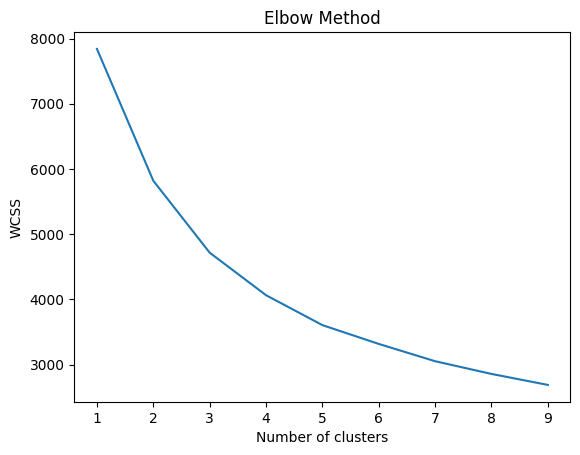
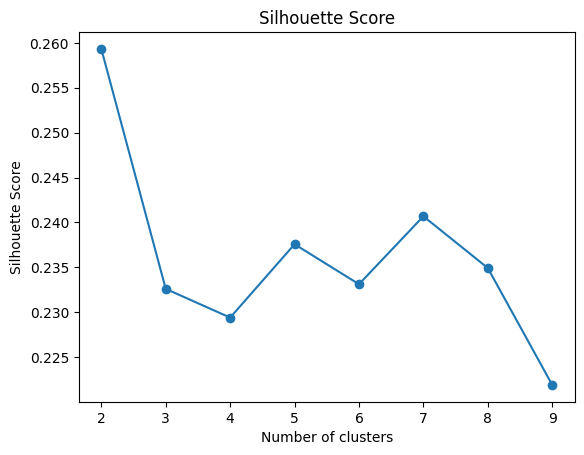
Clustering is an unsupervised Machine Learning method which is used to find similarities between dataset’s objects. For this dataset 3 different algorithms, namely, K-Means, K-prototypes and Agglomerative clustering have been used. And tried both normalized data and normal data for comparing.

* 1. ***K-means***

K-means is one of the most popular unsupervised machine learning algorithm because of its simplicity and effectiveness. K-means is an iterative algorithm that tries to split the dataset into K pre-defined distinct non-overlapping clusters where each data point belongs to only one group. But it is useful using it with numerical data because it doesn’t know how to handle with categorical data.

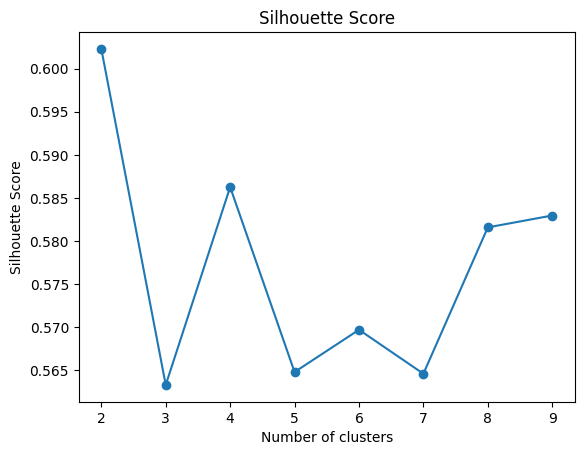
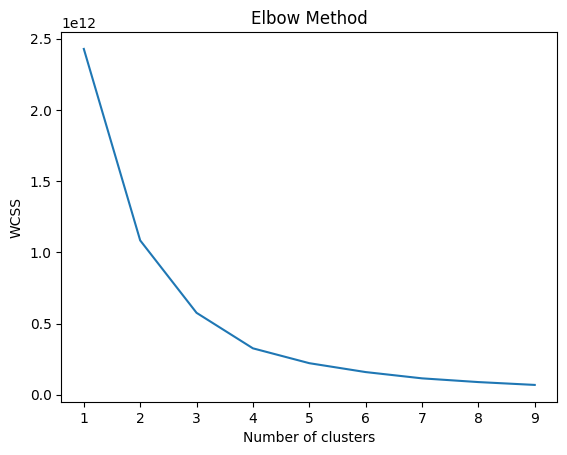
We can still use K-means algorithm with encoded data points. The main task is to determine the number of clusters needed. For this purpose, Elbow method and silhouette scores have been used.

For normalized data:

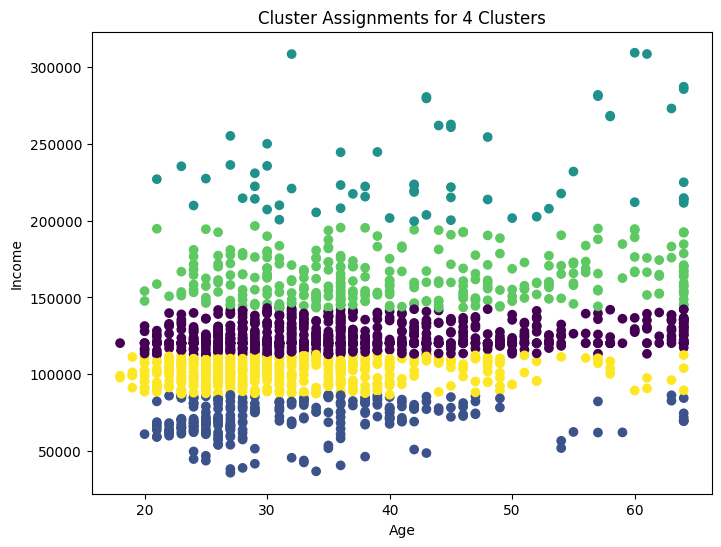
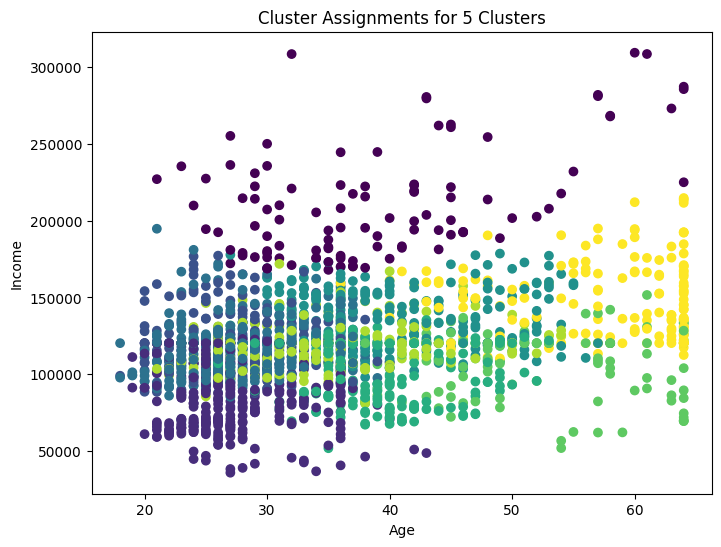
Silhouette Score: 0.23874554988173513

For normal data:



Silhouette Score: 0.5648159945357671

While there are not so much difference on elbow method but huge difference on Silhouette scores. In normalized data 5 is best number of clusters but in normal data 4 is best.

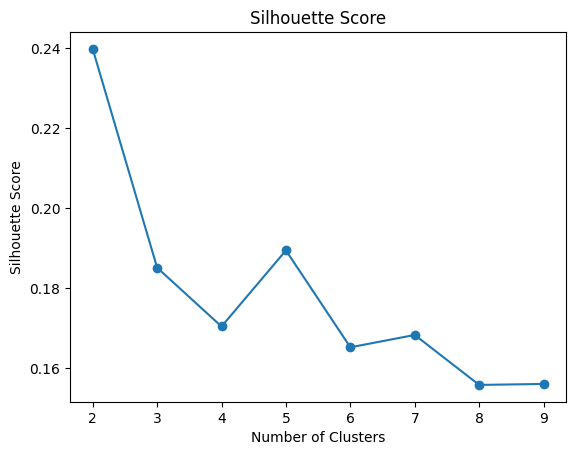
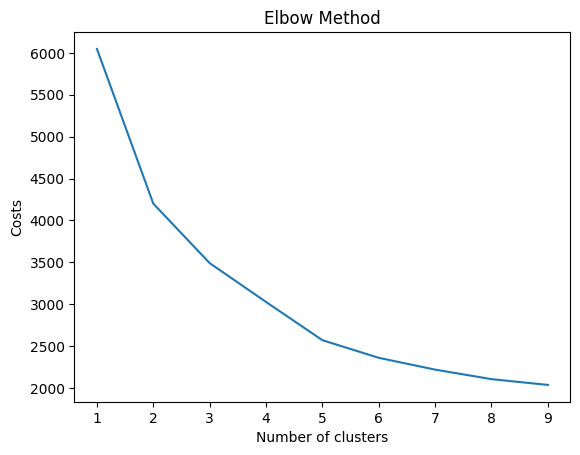


The result is obvious that normalized data divide clusters more effectively bur normal data just divide it based on income even if it is Silhouette score is more.

* 1. ***K-*** ***prototype***

This is a clustering model which includes a combination of k-means and k-modes models to achieve clustering of data points around certain prototypes (similar concept as centroids). It calculates Euclidean distances for numerical variables and Similarities for categorical variables and uses an assumed gamma as weightage (which decides the preference towards categorical or numerical variables), to assign points to a cluster.

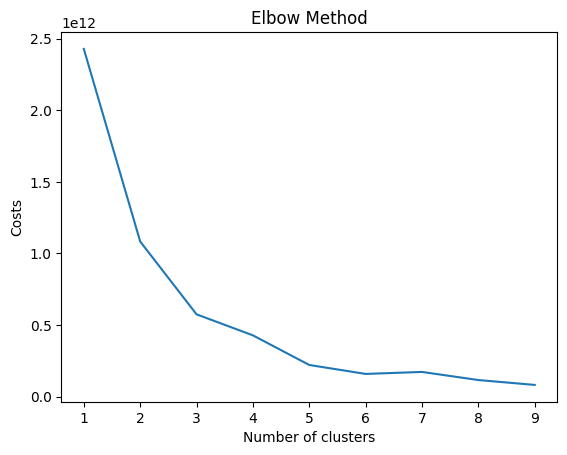
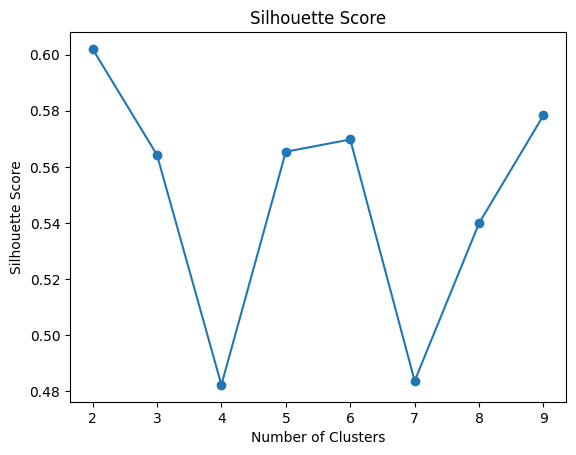
No encoded data is required to use this algorithm. So for normalized data graphs are:



Silhouette Score: 0.19005805539652504

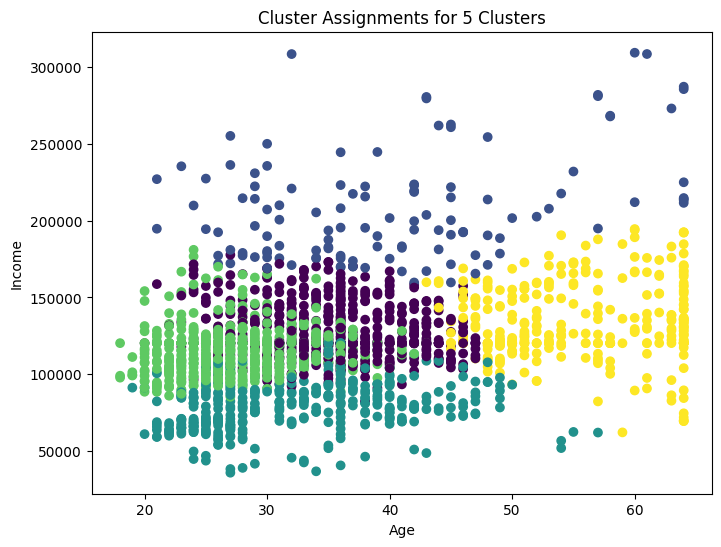
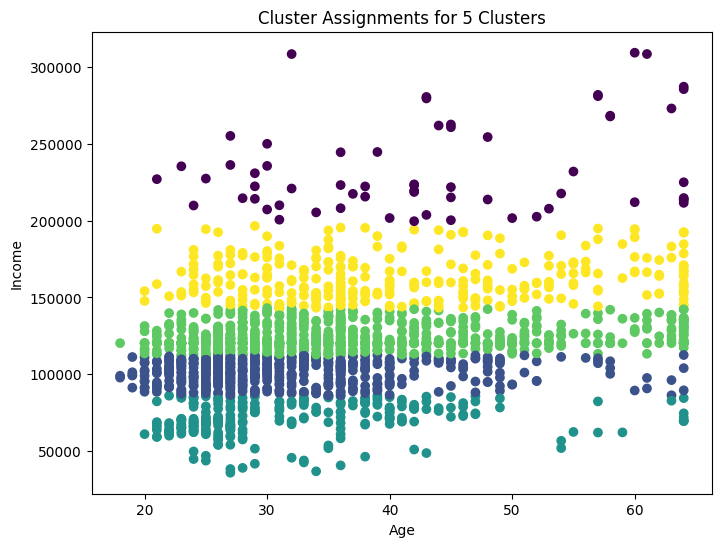
Considering these 2 graphs we can choose best cluster value as 5 again.

For normal data:

Silhouette Score: 0.5653179904886042

The number of clusters is 5 In these graphs too. But Silhouette scores still have huge gap. The normalized data divide clusters better again.

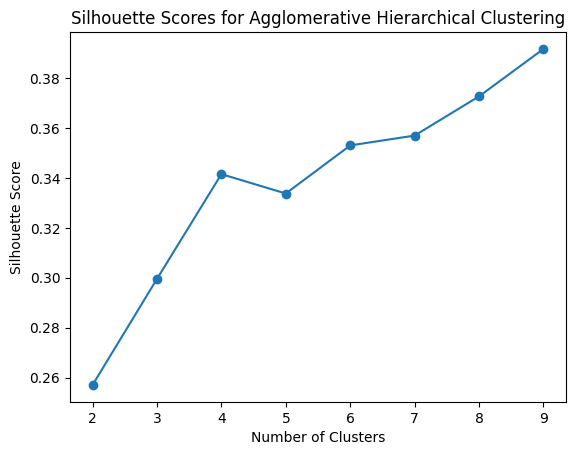
 

* 1. ***Agglomerative hierarchical clustering with gower distance***

Agglomerative clustering, unlike K-Means, is hierarchical clustering technique. It starts treating each data point as a cluster. Then it merges all ”clusters” into big cluster until the number of clusters is one. The result is tree-like object called dendrogram.

For agglomerative clustering we can’t use Elbow method since there’s no cluster centers in this algorithm. But still we can use Silhouette score.

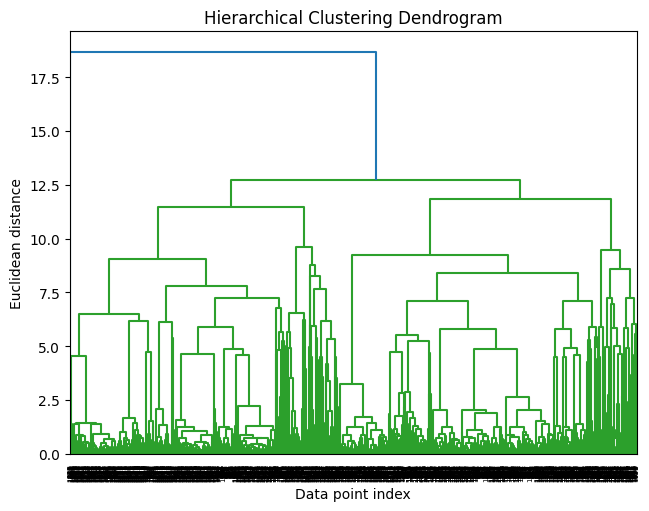
In Agglomerative clustering normalized and normal data both give approximately same result so only one of them will be given.



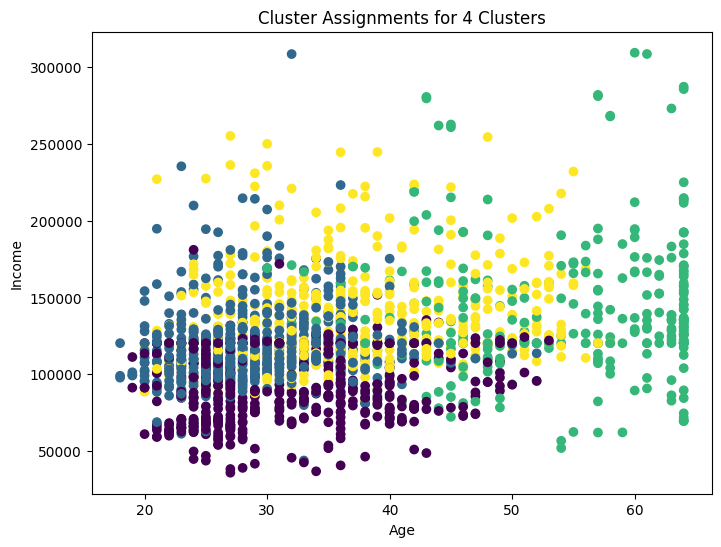
The Silhouette Score: 0.34147194

This is not best score for normal data but still best score for normalized data.

The dendrogram graph:



The output of division of 4 cluster with agglomerative clustering:



1. **Conclusion**

In conclusion, this project demonstrated the efficacy of clustering techniques in uncovering patterns and relationships within customer data. Normalized data consistently outperformed normal data in cluster division, emphasizing the importance of preprocessing steps. The choice of clustering algorithm should align with the nature of the dataset, considering both numerical and categorical variables. The findings from this project can guide businesses in tailoring their strategies based on distinct customer segments, ultimately enhancing decision-making processes.