**Hierarchical Clustering**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name:Navya.P**

**Batch Id:WDMCO 050121**

**Topic: Hierarchical Clustering**

**1. Business Problem**

* 1. **Objective**
  2. **Constraints (if any)**

**2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**Using R and Python codes perform:**

**3. Data Pre-processing**

**2.1 Data Cleaning, Feature Engineering, etc.**

**4. Exploratory Data Analysis (EDA):**

**4.1. Summary**

**4.2. Univariate analysis**

**4.3. Bivariate analysis**

**5. Model Building**

**5.1 Build the model on the scaled data (try multiple options)**

**5.2 Perform the hierarchical clustering, visualize the clusters using dendrogram**

**5.3 Validate the clusters (try with different no. of clusters) – label the clusters and derive insights (compare the results from multiple approaches)**

**6. Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

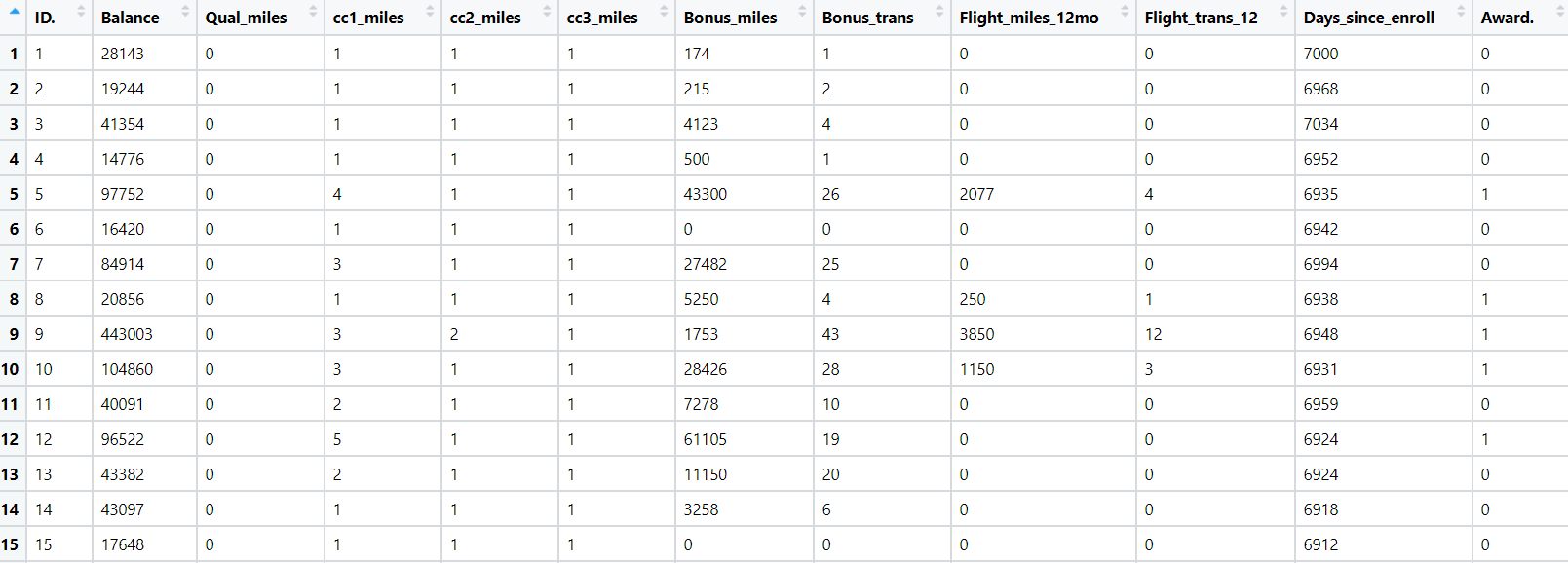
**Note:**

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the modules (elaborating on steps mentioned above)

**Problem Statement:**

1. Perform clustering for the airlines data to obtain optimum number of clusters. Draw the inferences from the clusters obtained. Refer to EastWestAirlines.xlsx dataset.

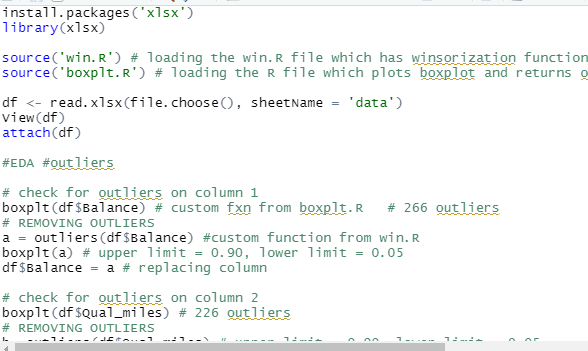


The Data pre-processing is done on the data set.

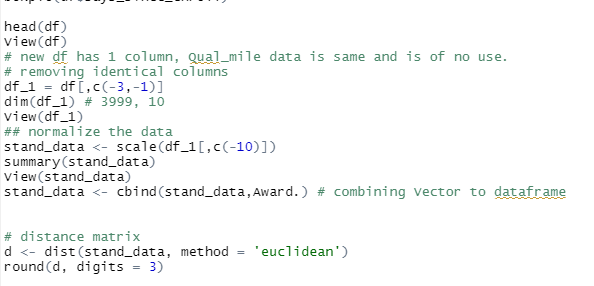
The outliers are adjusted using winsorization technique. Box plot visualization is done, to check the before and after outliers.

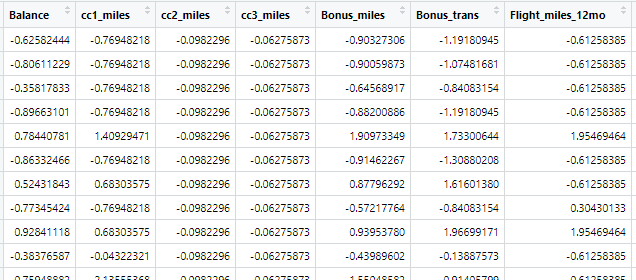
In R:

I have formed two R scripts separately for boxplot and winsorization.



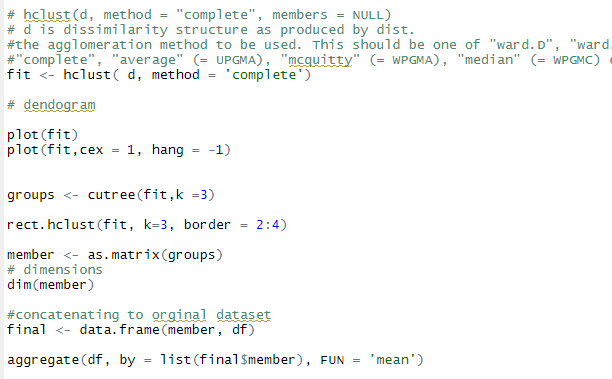
The outliers are checked using box plot and removed and then the columns are updated as per the values obtained. The obtained data frame is as follows. It has 1 identical column Qual\_miles. We have to remove the column for further analysis.

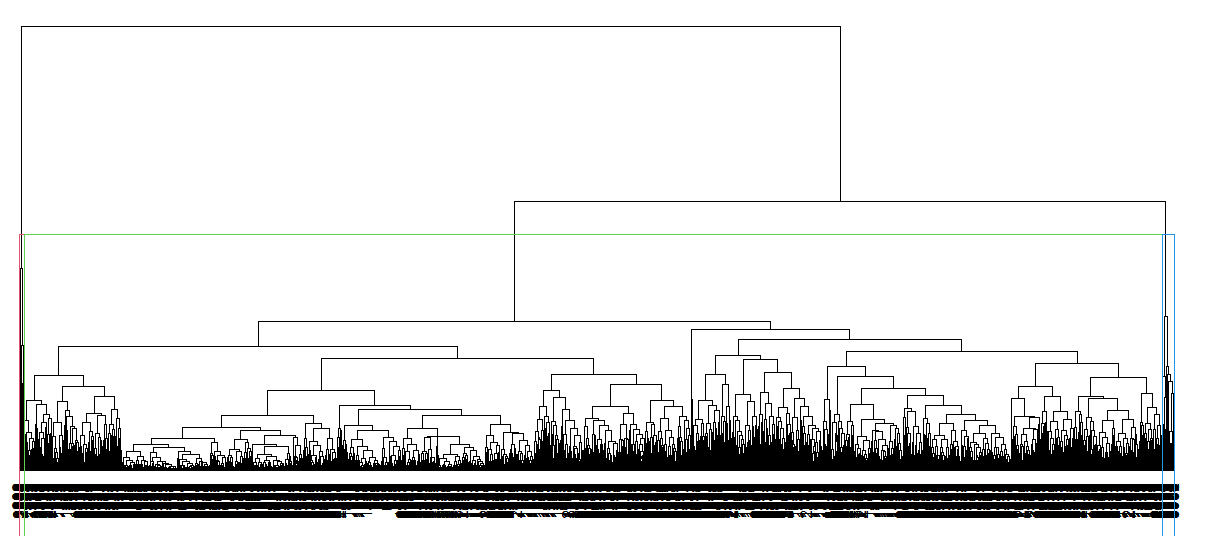


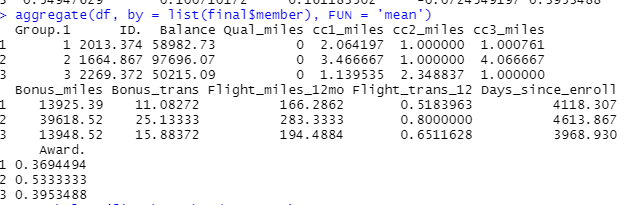


Now, we calculate the distance matrix i.e., using Euclidean distance method. . Hierarchical clustering is performed on this distance matrix. And the dendogram is viewed to see the no. of clusters and group appropriately.

Dendogram is viewed and sectionalized based on the no. of big clusters to analyse the data. Here k = 3 is appropriate. The matrix is then attached to original df and grouped with respect to the clusters mentioned to analyse the data.

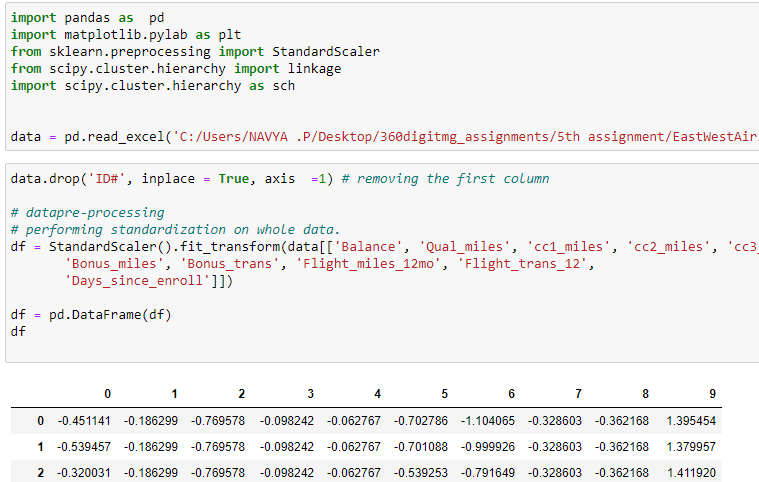


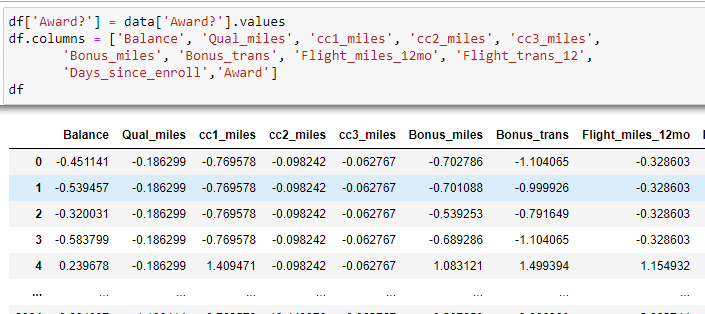


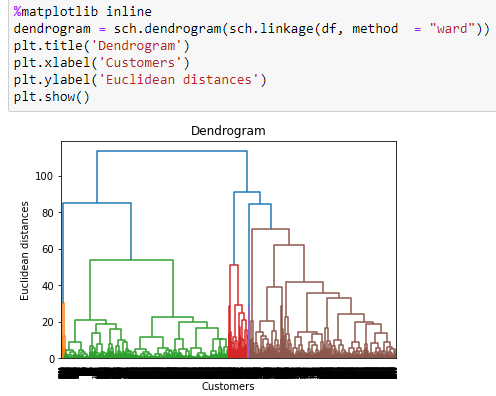


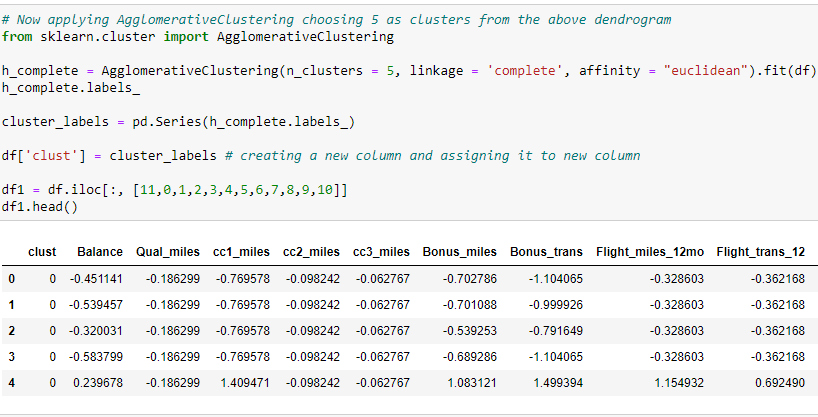
**IN PYTHON**

In python, the code is as follows.



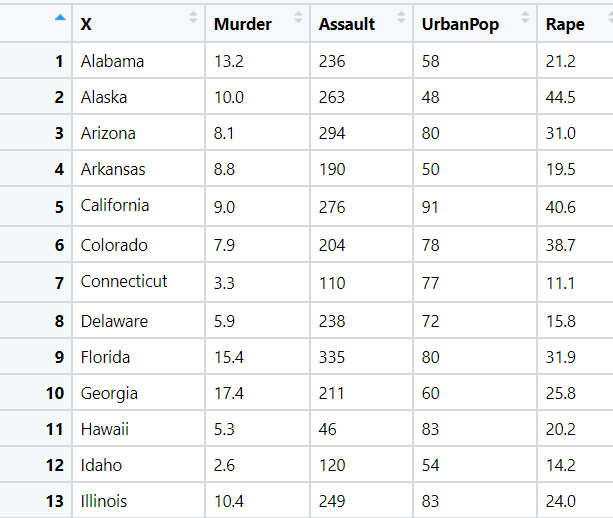






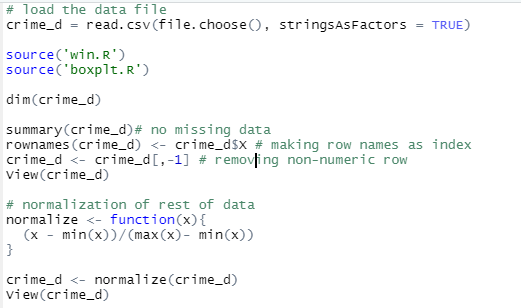
The data is clustered into 5 groups as per the dendogram shown.

1. Perform clustering for the crime data and identify the number of clusters formed and draw inferences. Refer to crime\_data.csv dataset.

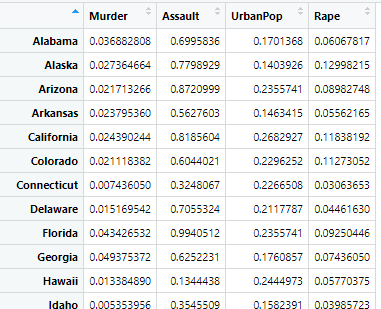


Load the data file. The dataset has 4 numeric columns and one categorical value of city. The city column is converted to index for better analysis

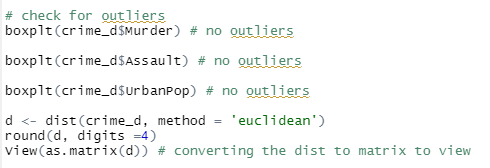
The data is normalized and then checked for outliers. The data has no missing values.



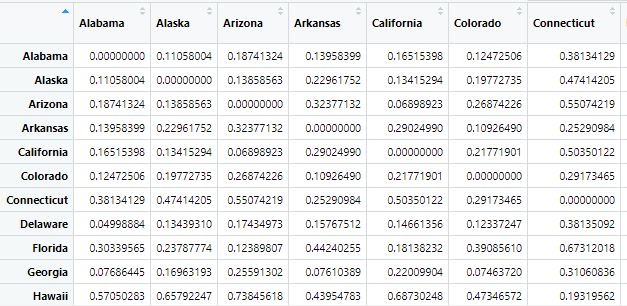
The normalized data



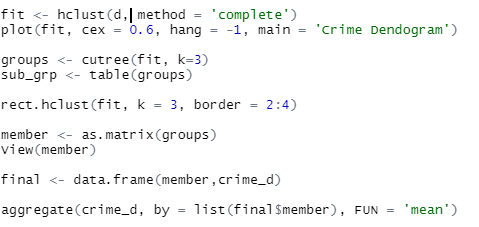
There are no outliers in the data. So, getting to next step



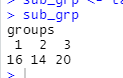
The dissimilarity matrix with Euclidean distance.



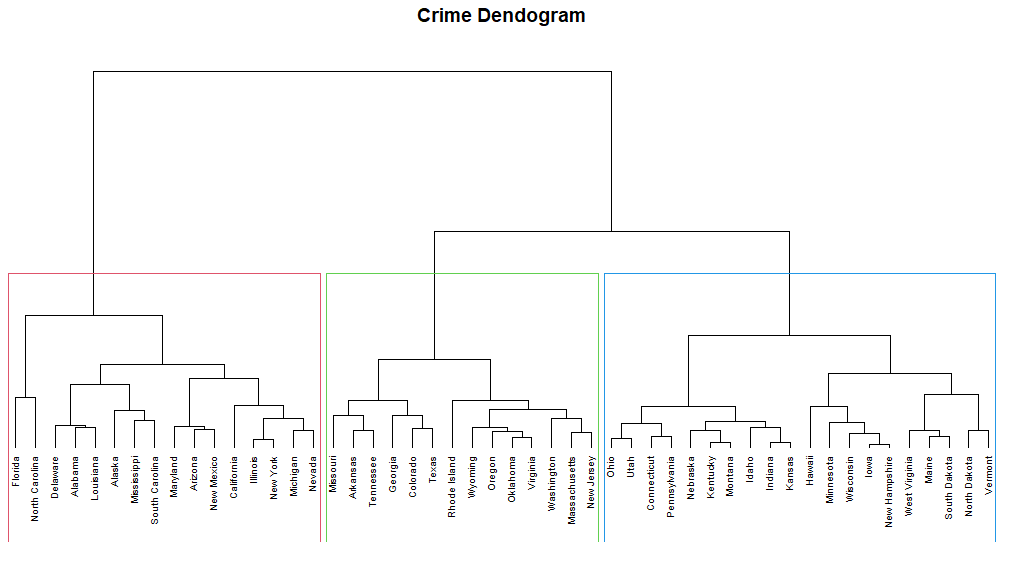
Proceeding to hclust to perform hierarchical clustering. After performing hierarchical clustering the distances are grouped into 3 clusters based on denogram.



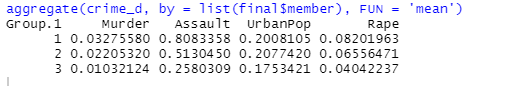
The groups are shown as follows. There are 16 cities in cluster 1, 14 cities in cluster 2, 20 in cluster 3.

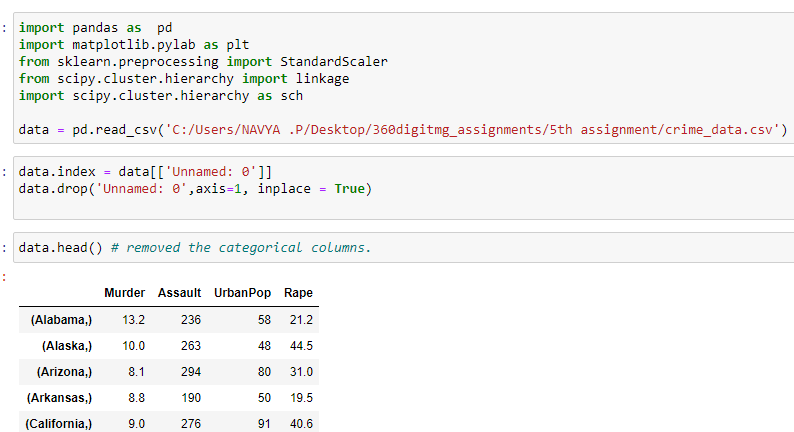


The dendogram is as follows.

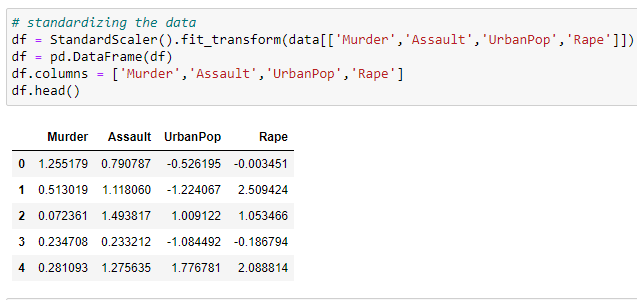


The details of grouped dataset is as follows.

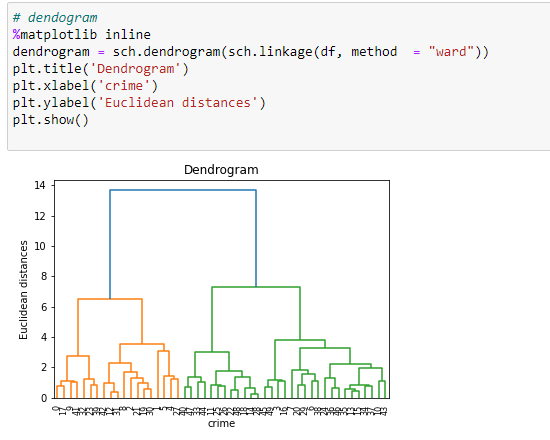


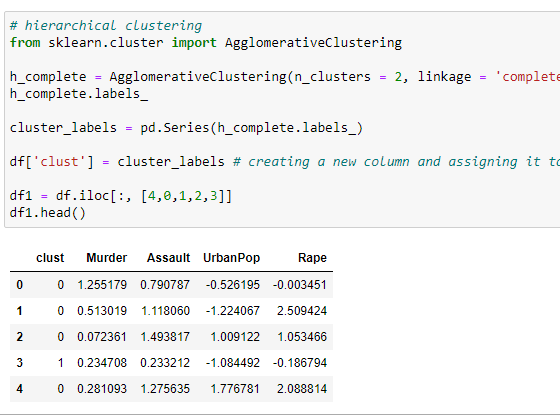
IN PYTHON :

Data pre-processing:



Dendogram:



Hierarchical agglomerative clustering

1. Perform clustering analysis on the telecom data set. The data is a mixture of both categorical and numerical data. It consists the number of customers who churn. Derive insights and get possible information on factors that may affect the churn decision. Refer to Telco\_customer\_churn.xlsx dataset.

Hint:

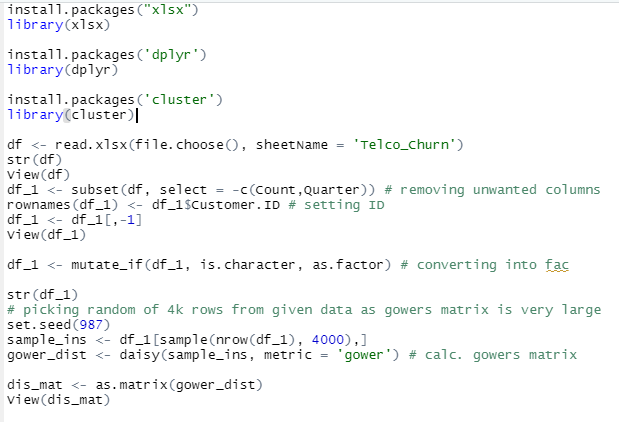
* Perform EDA and remove unwanted columns.
* Use Gower dissimilarity matrix, In R use daisy() function.



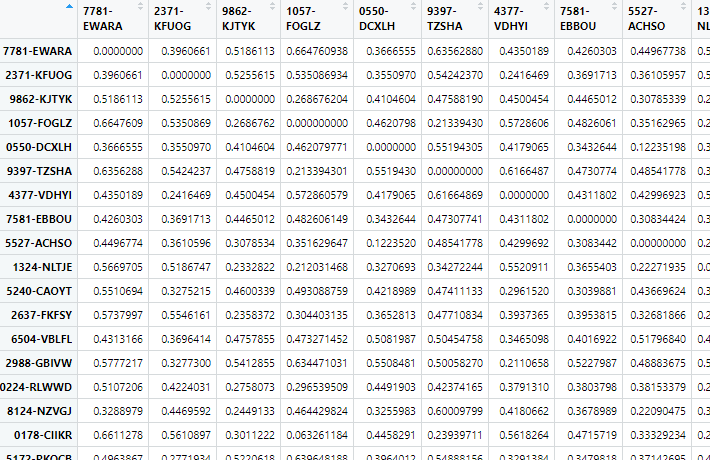
The data set has two unanted columns. They are removed and the whole dataset is passed on to daisy() function from cluster package.

The gowars metric is used to perform the row wise distance measure.

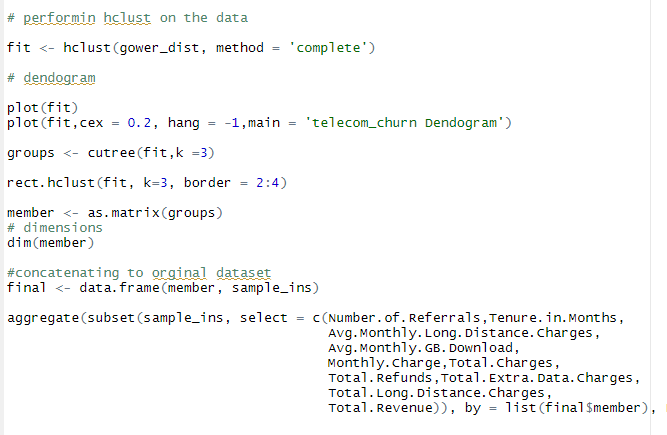
The R code is as follows:



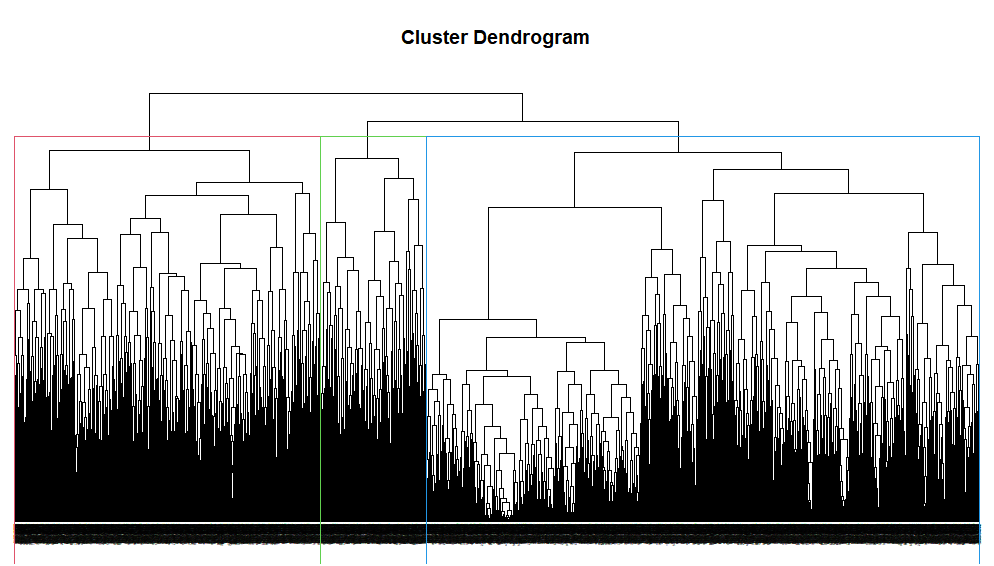
The distance matrix is as follows.



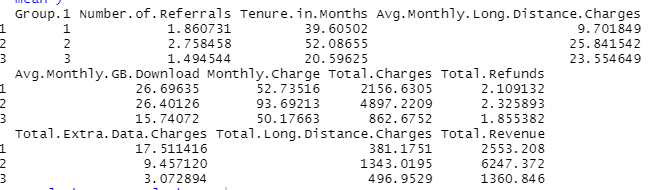
The hierarchical clustering is done based on the above matrix.



The dendogram is as follows.



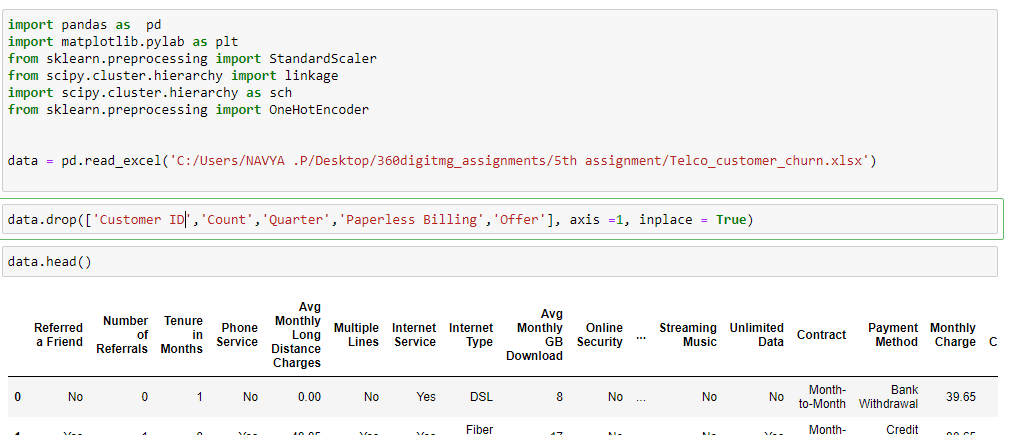
The result of hclust is as follows:

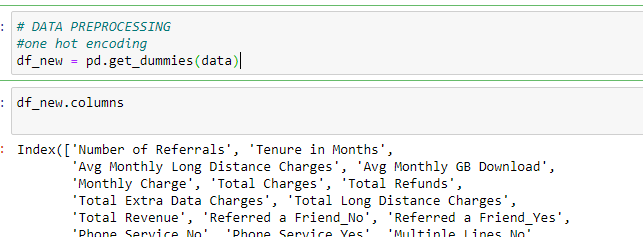


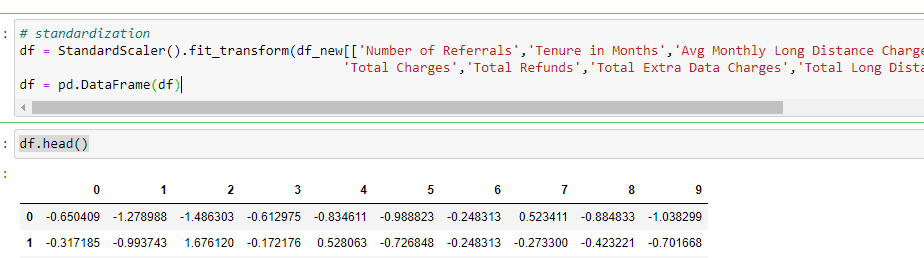
IN PYTHON:

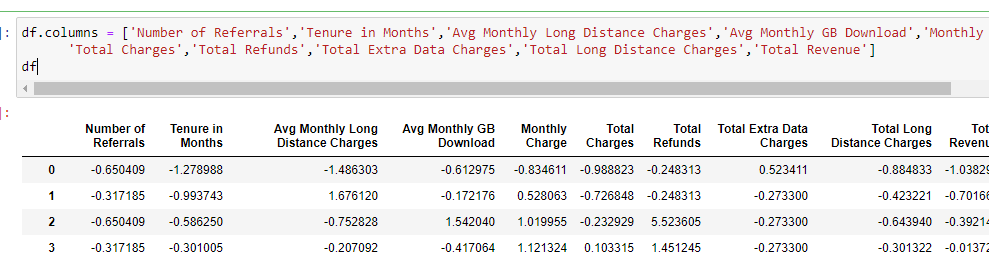
The data pre-processing is done, by removing the unwanted columns.

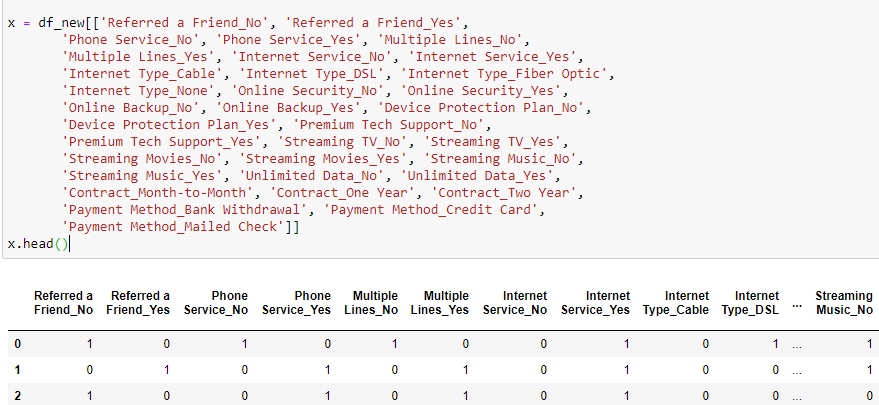
The categorical data is converted into one hot encoding. The numerical data is standardized.

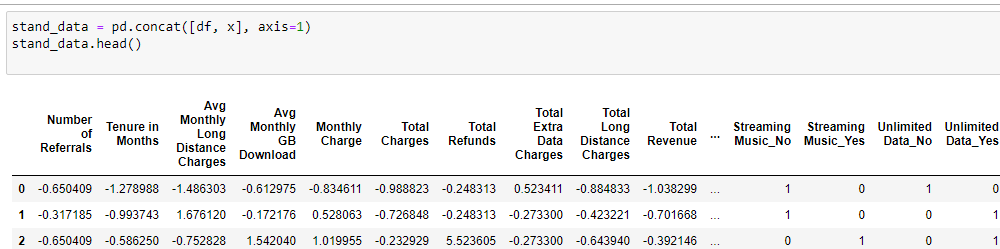


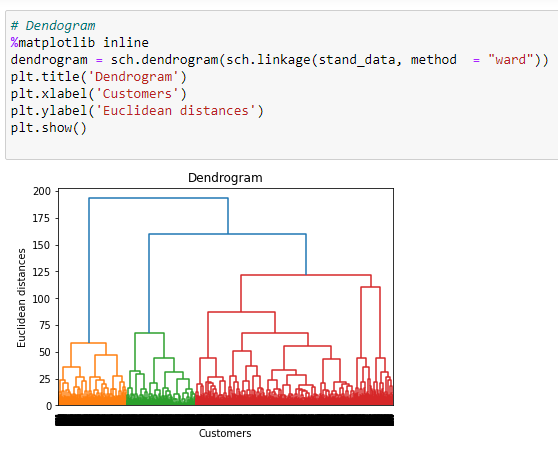


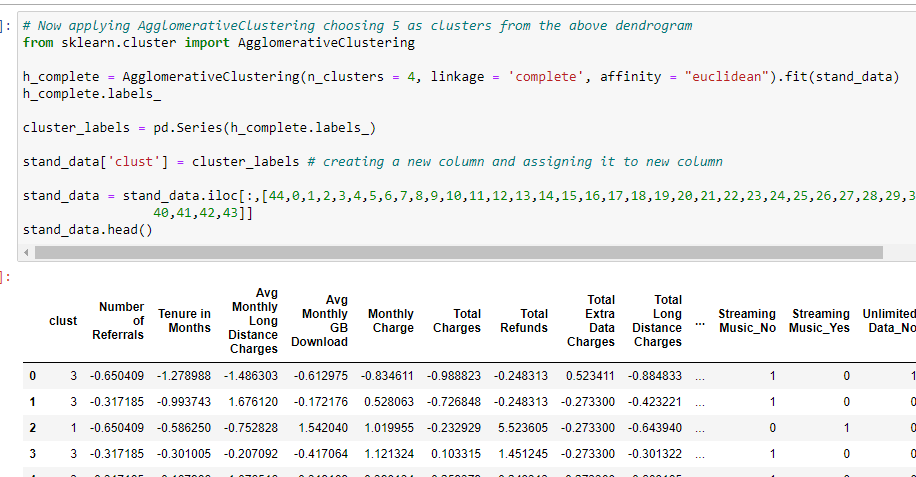












1. Perform clustering on mixed data convert the categorical variables to numeric by using dummies or Label Encoding and perform normalization techniques. The data set consists details of customers related to auto insurance. Refer to Autoinsurance.csv dataset.

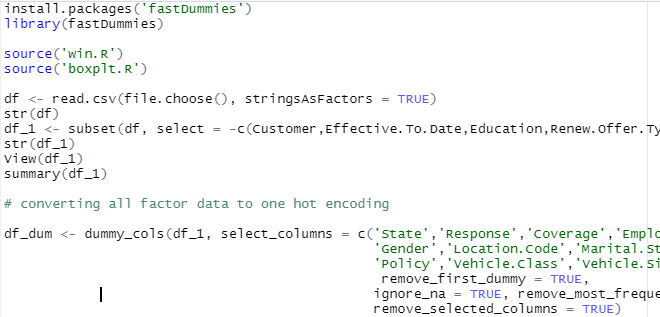


The data has unwanted data. Those columns are Customer, Effective.To.Date, Education, Renew.Offer.Type, Sales.Channel.

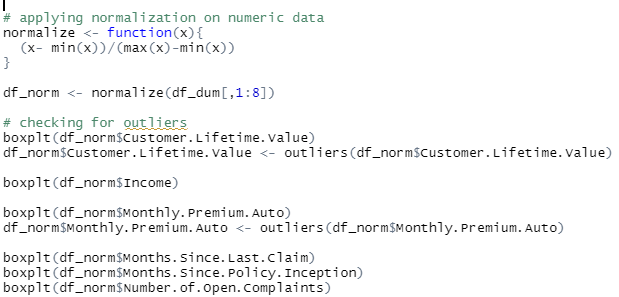
The above four columns are deleted. The remaining data has 8 numerical columns and 11 factor / categorical values.

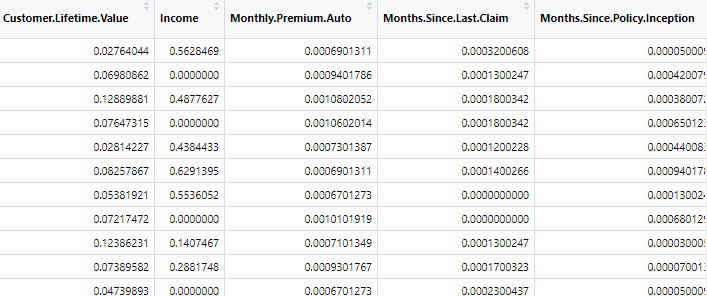
They are converted using one hot encoding.

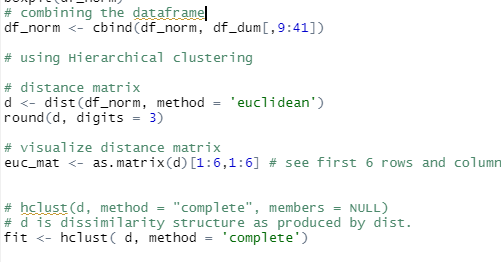
Later normalization is done for numeric data. This is the data pre-processing done on the data.

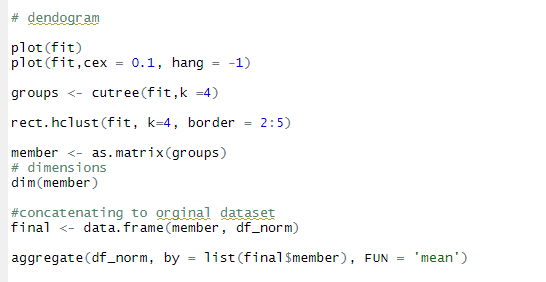


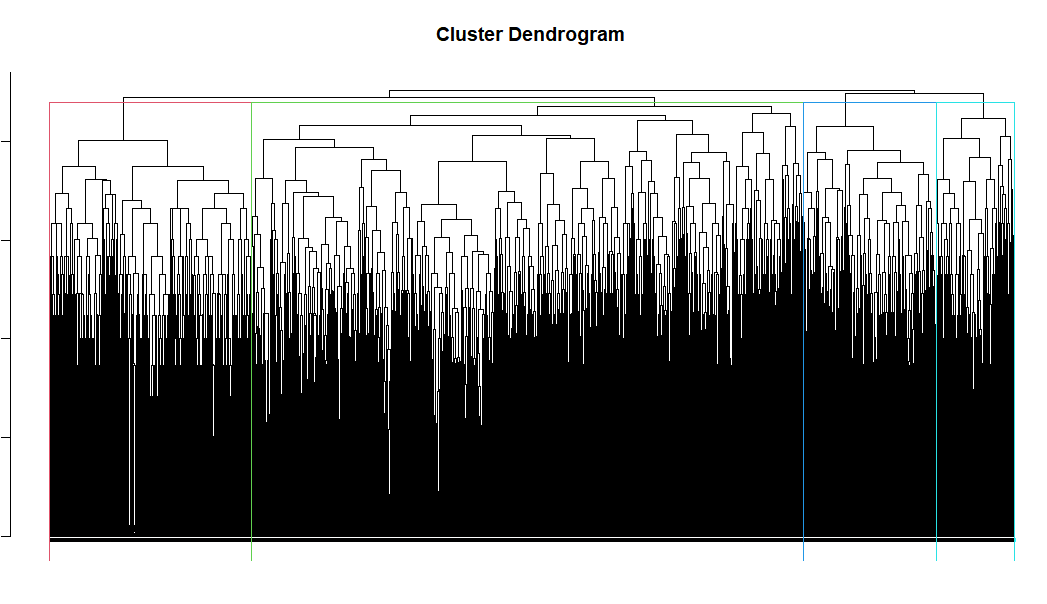
Applying normalization,

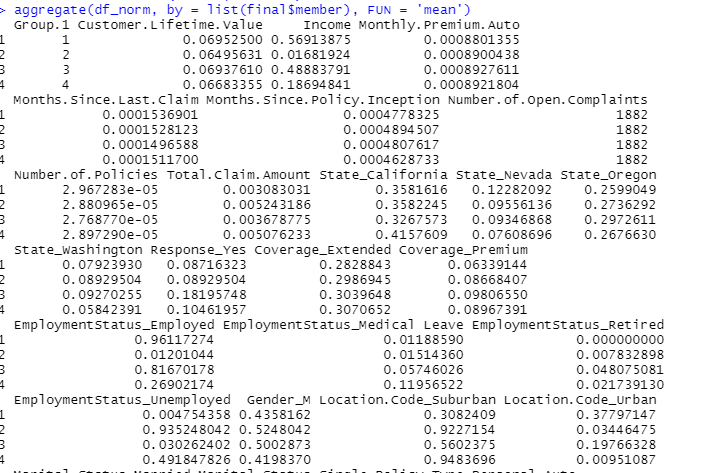


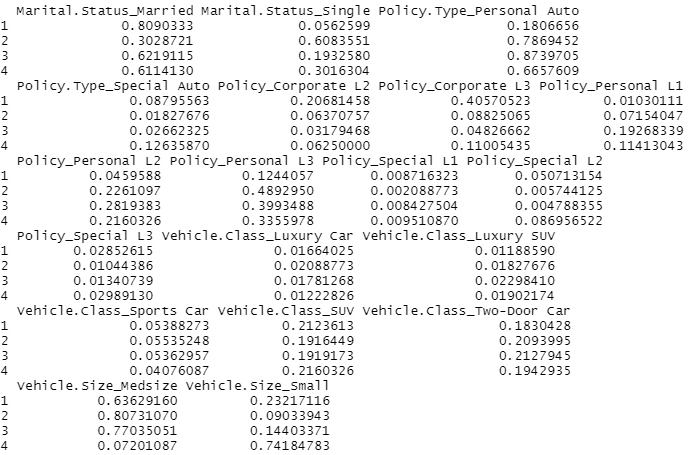












IN PYTHON :

