**Hierarchical Clustering**

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Ritendu Bhattacharyya**

**Batch ID:** DSWDMOD 190522H

**Topic: Hierarchical Clustering**

**Problem Statements:**

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.



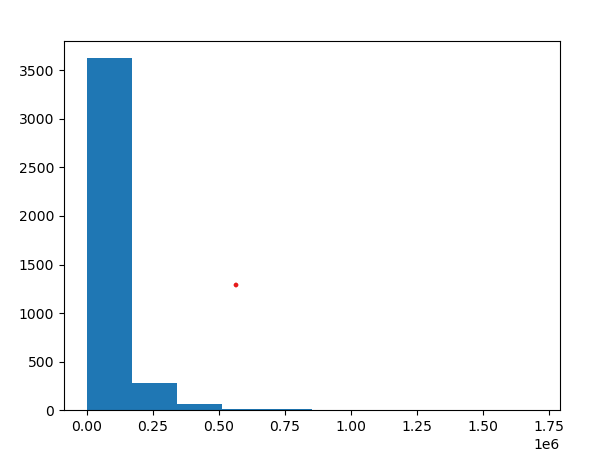
**Ans –**

After importing the data first dropping useless column first.

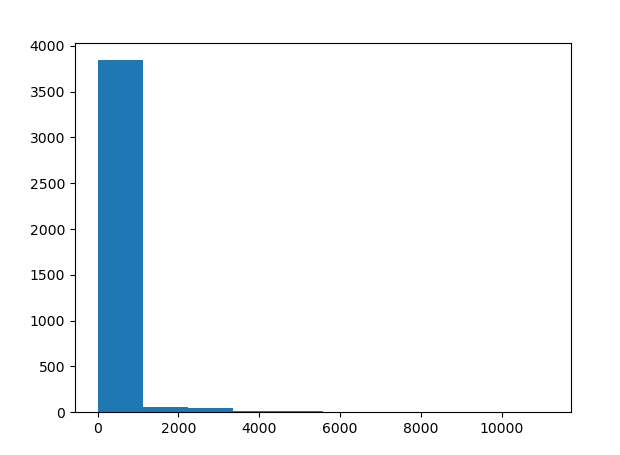
df.drop(['ID#'], axis=1, inplace=True)

then analyzing the histogram of the data.

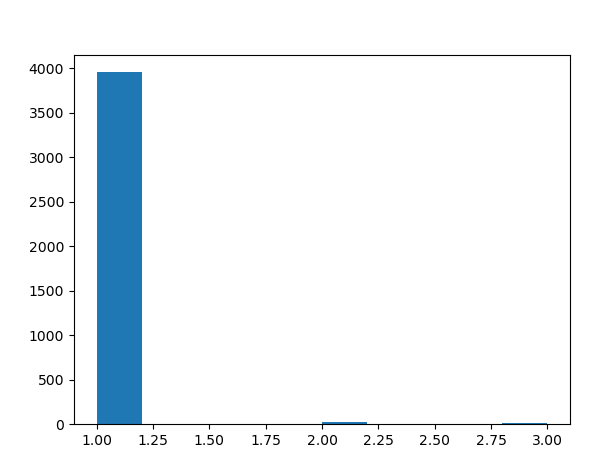
df.Balance



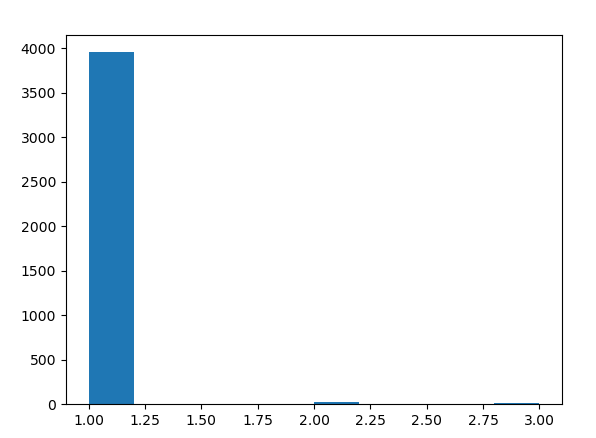
df.Qual\_miles



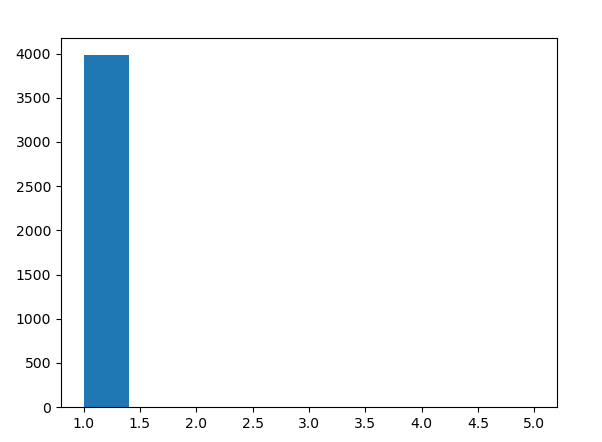
df.cc1\_miles



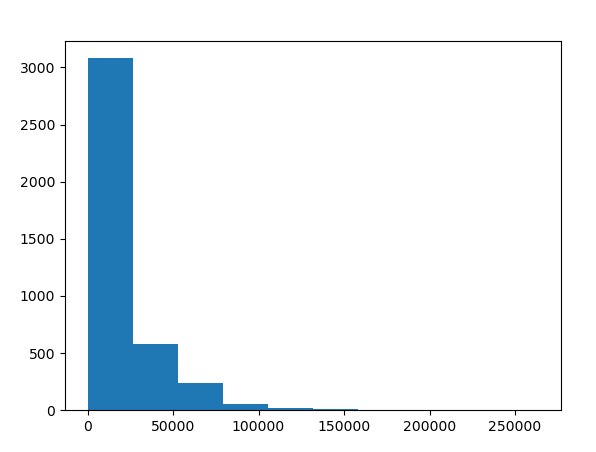
df.cc2\_miles



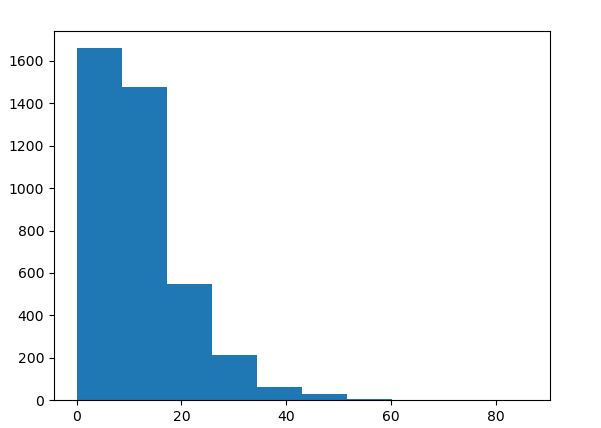
df.cc3\_miles



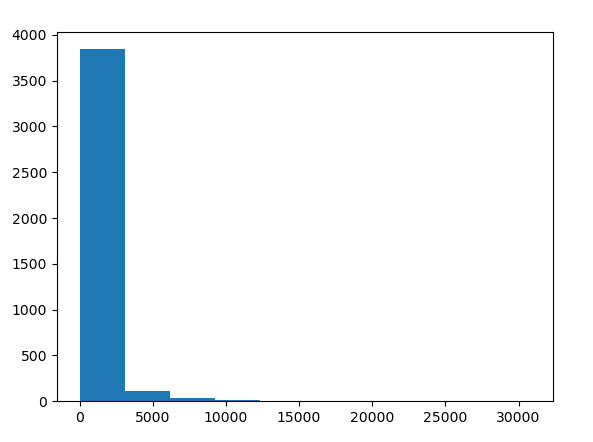
df.Bonus\_miles



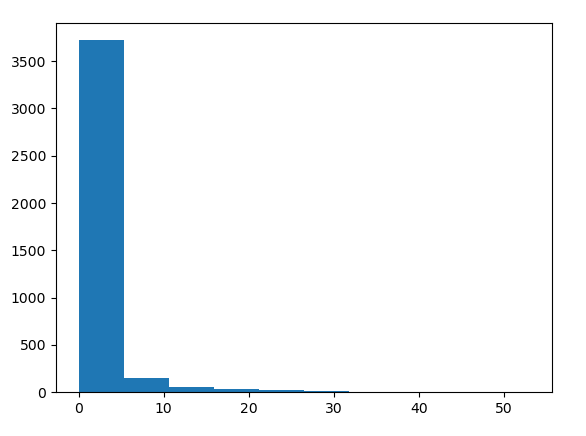
df.Bonus\_trans



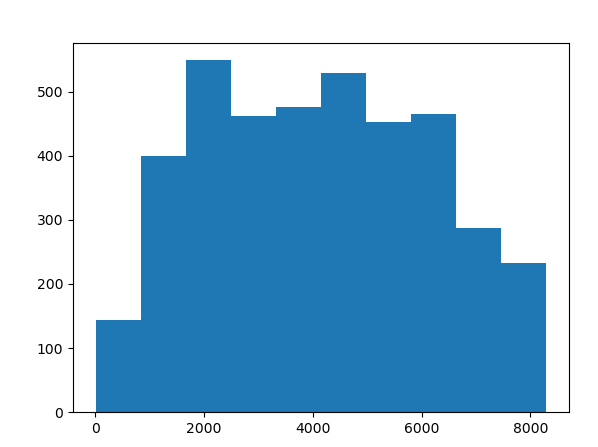
df.Flight\_miles\_12mo



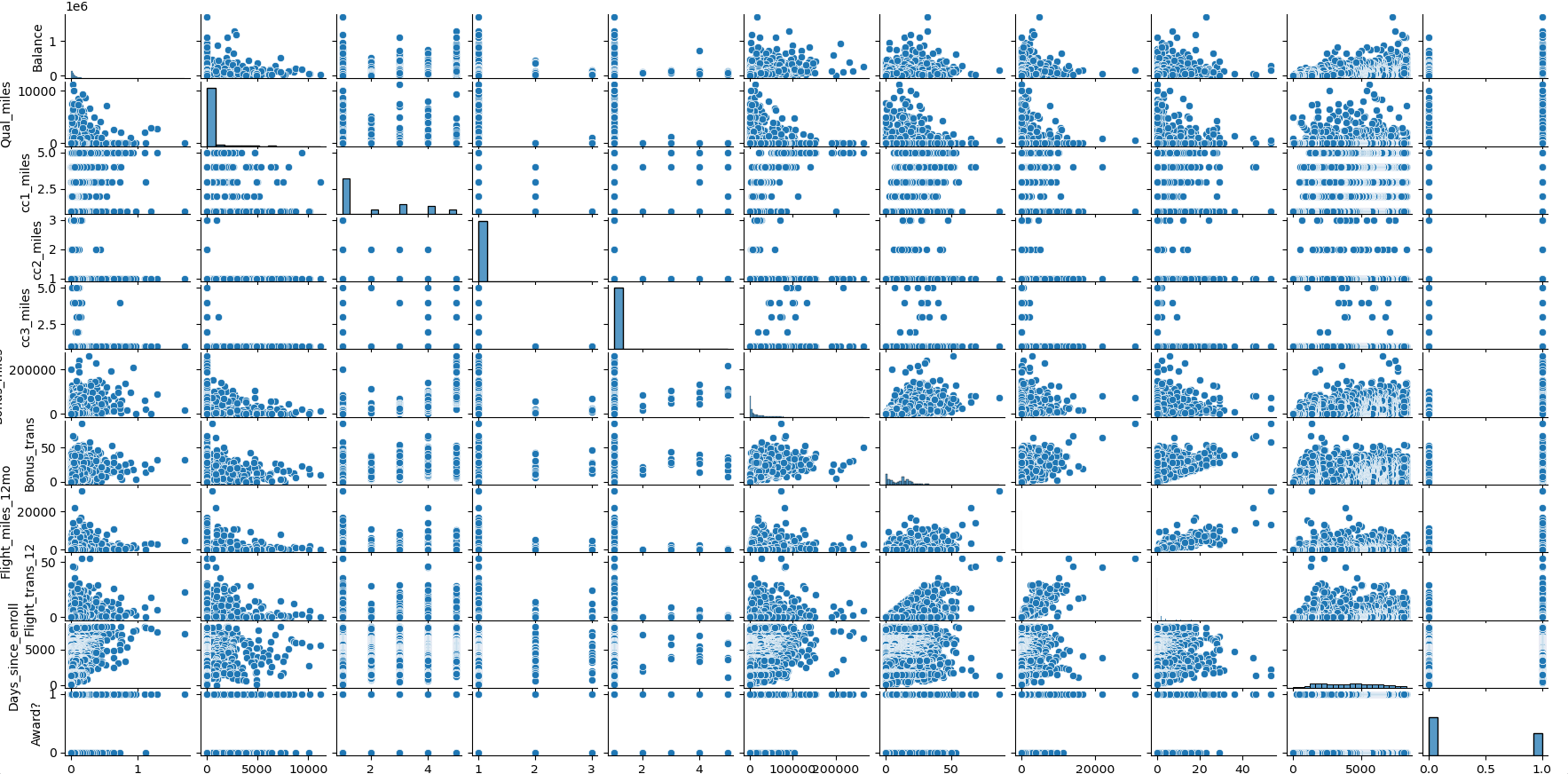
df.Flight\_trans\_12



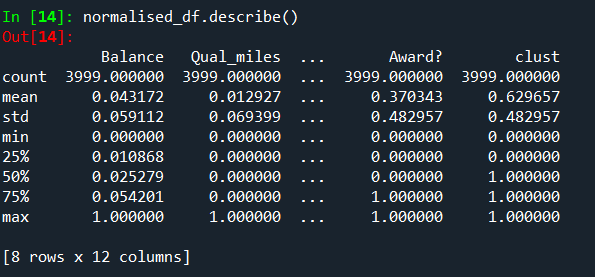
df.Days\_since\_enroll



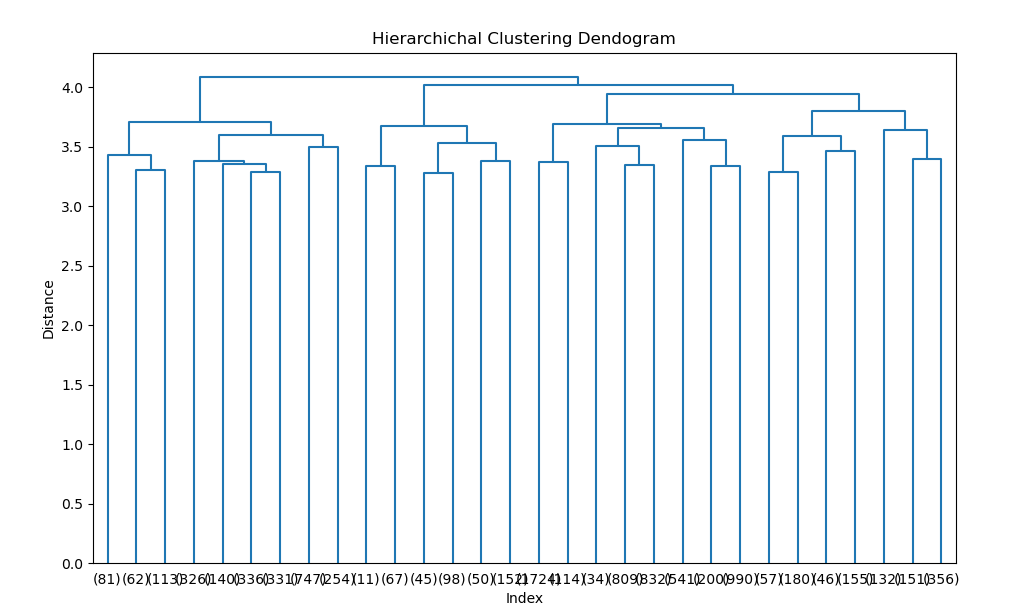
Doing the pair plot for the dataset



Normalized the data



After that applied dedrogram with linkage matrix



Clustered the entire data in 2 clusters.

1st Cluster : Frequent flyers

2nd Cluster: Less frequent flyers

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

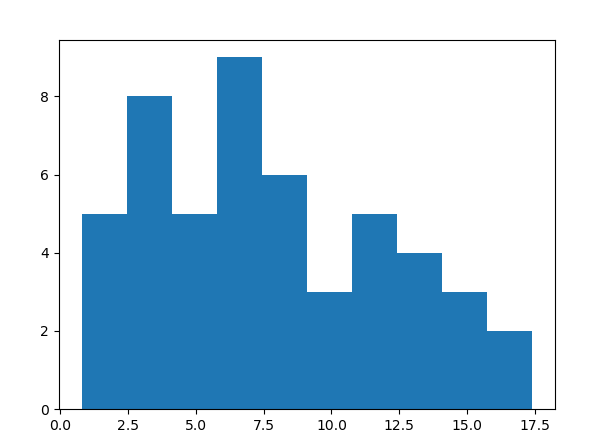


After importing the dataframe, dropped the unwanted column.

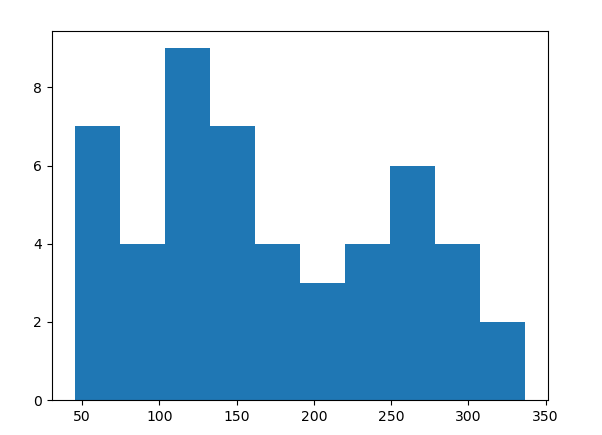
crime\_data.drop(['Unnamed: 0'], axis=1, inplace=True)

started analyzing the histogram

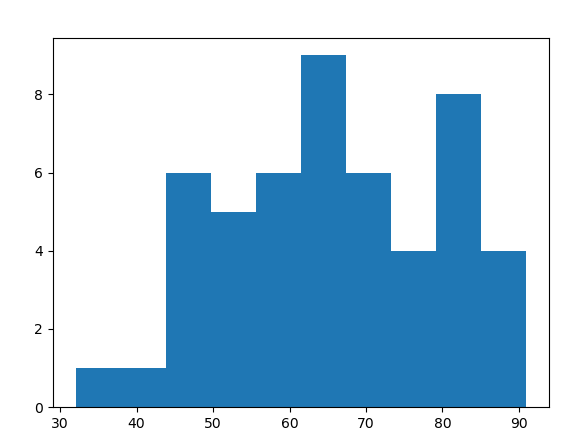
crime\_data.Murder



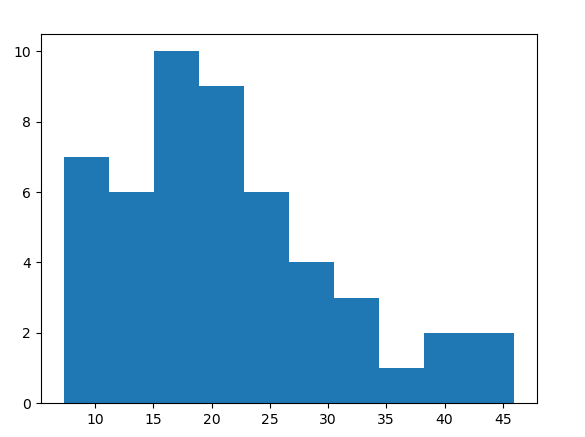
crime\_data.Assault



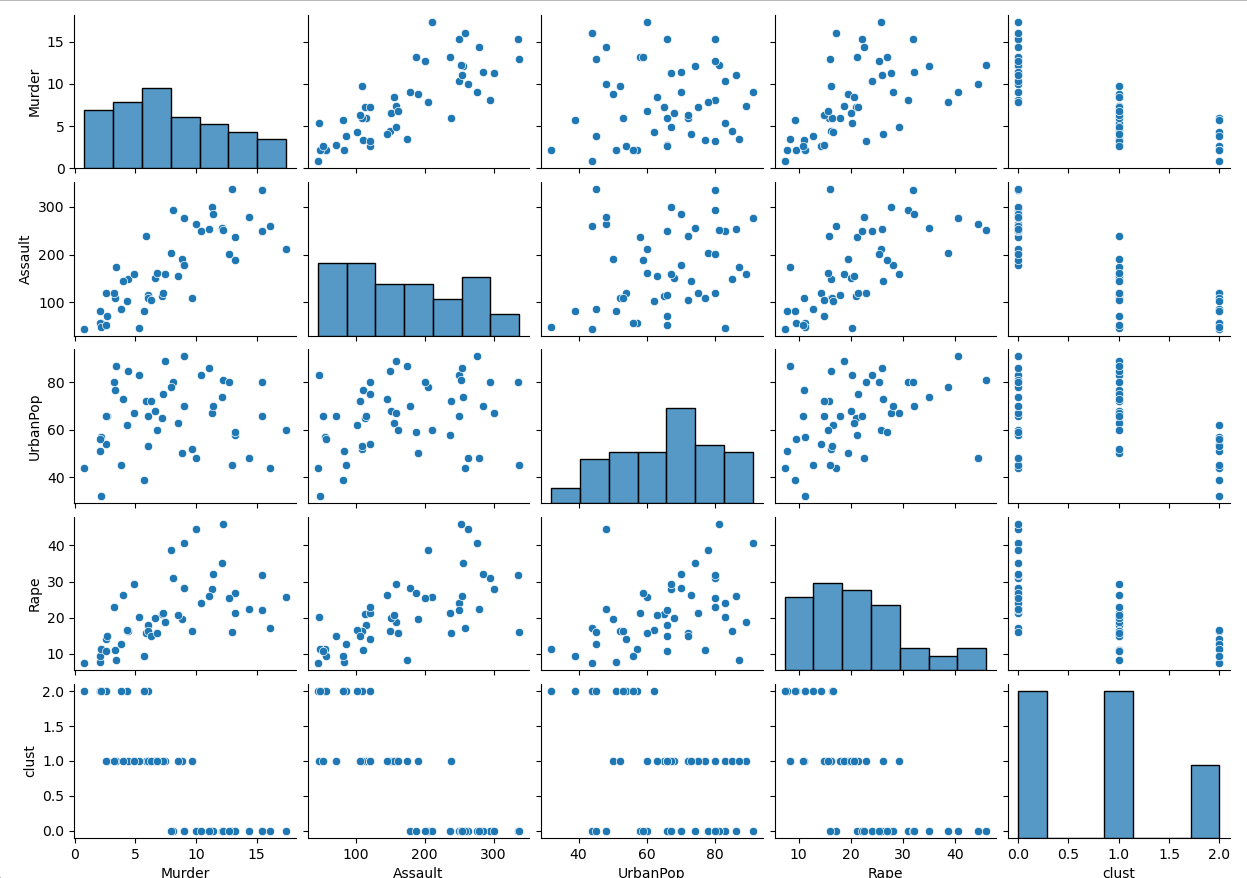
crime\_data.UrbanPop



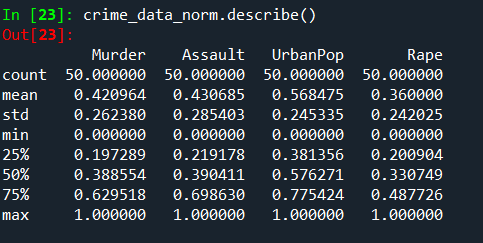
crime\_data.Rape



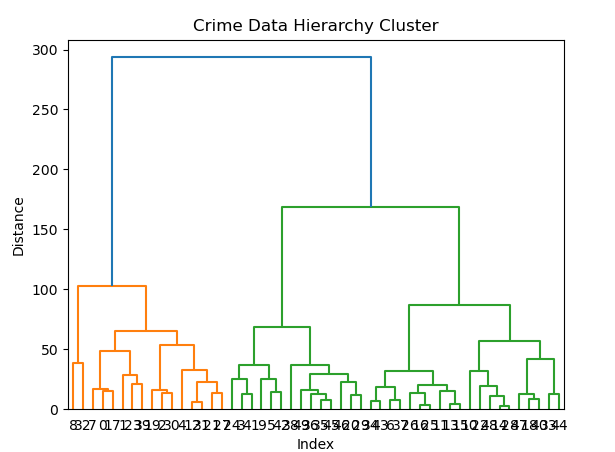
Now analyzing the pairplot of the dataset



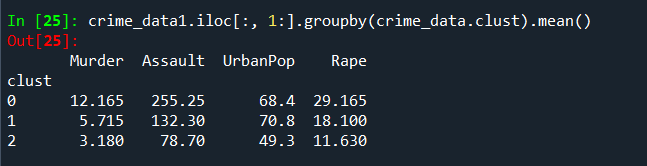
Normalized the data



Plotting the dedrogram



Divided into 3 clusters

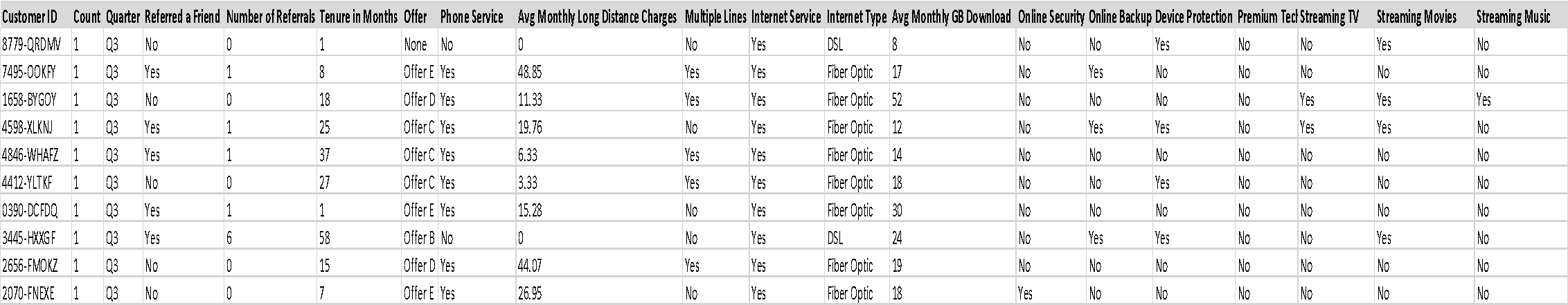


1st Cluster – High crime areas

2nd Cluster – Medium crime areas

3rd Cluster – Low crime areas

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



After getting the data, dropped the zero var data

columns\_with\_var = tele\_customer\_churn.var() != 0

columns = tele\_customer\_churn.columns

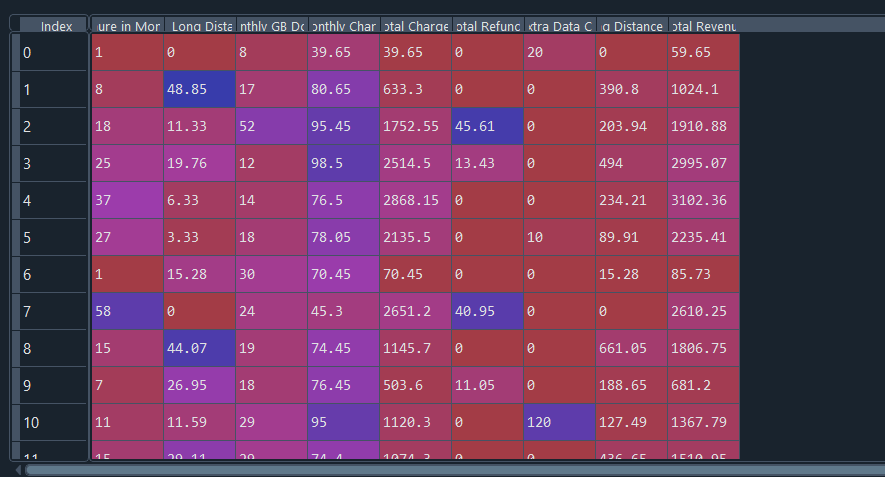
for i in columns:

if(i in columns\_with\_var.index):

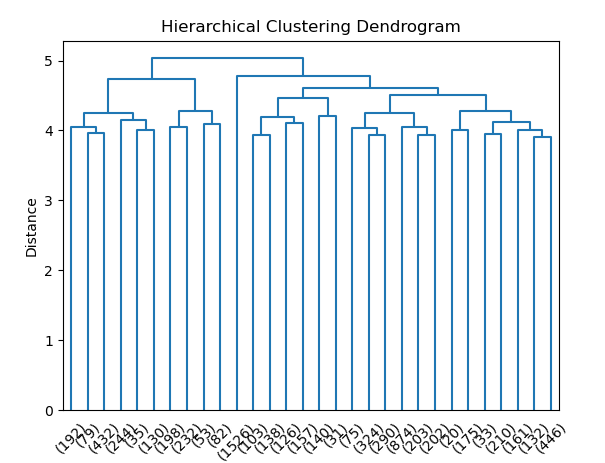
if(columns\_with\_var[i] == 0):

tele\_customer\_churn.drop(i, axis=1, inplace=True)

Divided the object data into a data set



After that did onehot encoding in the categorical data and standardization in numerical data. Then merged those two data and plotted the dendrogram



Divided the data into 3 clusters.

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 of details of customers related to their auto insurance. Refer to Autoinsurance.csv dataset.



After importing the data, dropped the useless columns.

auto\_insurance.drop(['Customer', 'State', 'Education', 'Gender', 'Location Code', 'Marital Status', 'Sales Channel', 'Vehicle Class', 'Vehicle Size'], axis=1, inplace=True)

extracted object data into one dataset and numerical data into another dataset

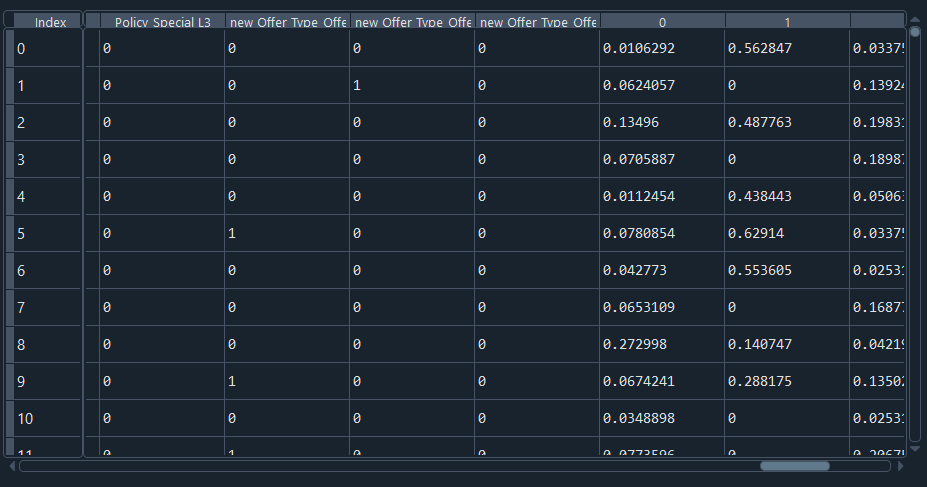
auto\_insurance\_object = auto\_insurance.select\_dtypes('object')

auto\_insurance\_numeric = auto\_insurance.select\_dtypes(['int64', 'float64'])

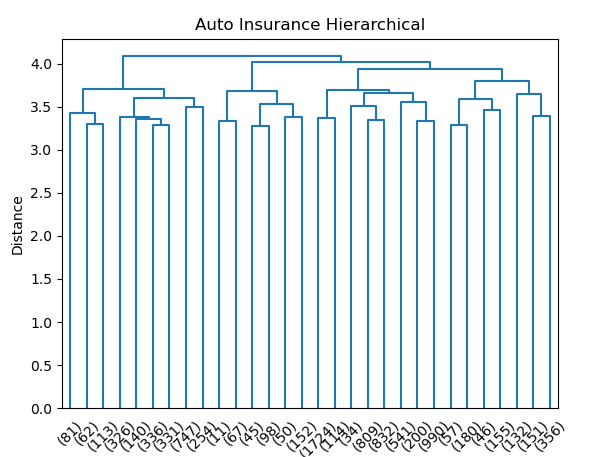
did one hot encoding in the object data

did normalization in the numerical data

after that concat those and got one dataset which will consist of all numerical data



Plotted the dendrogram over this dataset



Divided the dataset into 3 clusters

Did head of those clusters and analyzed the data

auto\_insurance[auto\_insurance['cluster']==0].head()

auto\_insurance[auto\_insurance['cluster']==1].head()

auto\_insurance[auto\_insurance['cluster']==2].head()

**Hints:**

**1. Business Problem**

* 1. **What is the business objective?**
  2. **Are there any constraints?**

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



**3. Data Pre-processing**

**3.1 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 and visualize the clusters using dendrogram.**

**5.3 Validate the clusters (try with different number of clusters) – label the clusters and derive insights (compare the results from multiple approaches).**

**6. Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**