# Topic: Hierarchical Clustering

## 1.) Perform clustering (Both hierarchical and K means clustering) for the airlines data to obtain optimum number of clusters.

Draw the inferences from the clusters obtained.

**Ans:**

**Importing the data set directly as there is no output variable.**

**Data Preprocessing:**

* Columns are rearranged to have the easy access of the variables.
* In R , dummy variable are created automatically when object(x) is created as model. Matrix command has inbuild feature of converting dummy variables or we can factorize the variables whereas, in python we need to write label encoder () code for creation of dummy variables.
* Removing unnecessary columns and checking for the NA value

**Normalizing the data:**

* Using the normalization function, the data is normalized and the data will come under same level.

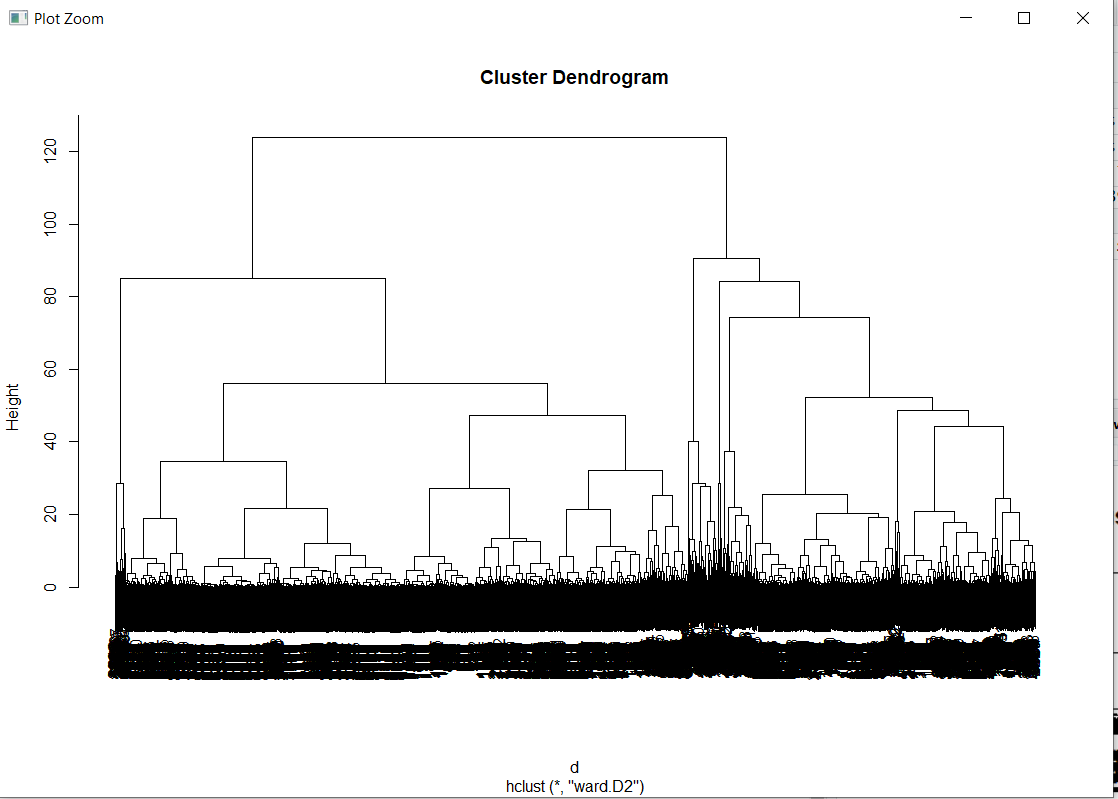
**def norm\_func(i):**

**x=(i-i.min()/i.std())**

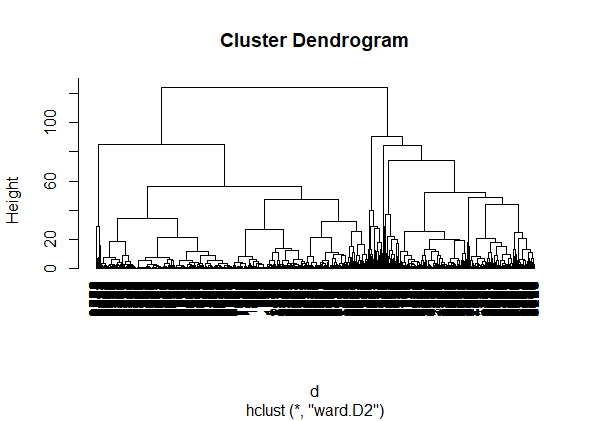
**return (x)**

**Dendogram:**

* IF Standardization/Normalization of data is not done then below issues will influence the model.
* Also Distance measure will be calculated incorrectly, if all the variables are not brought to equal weight.
* Largest scale will dominating the measure.
* The below dendogram generated with the given data set with standardization is as below.
* Using **ward.D2** method the dendogram is shown as below.



* Now the dendogram is divided into clusters removing the edge of the clusters on x-axis and the aggregate of each cluster is compared with the other clusters.

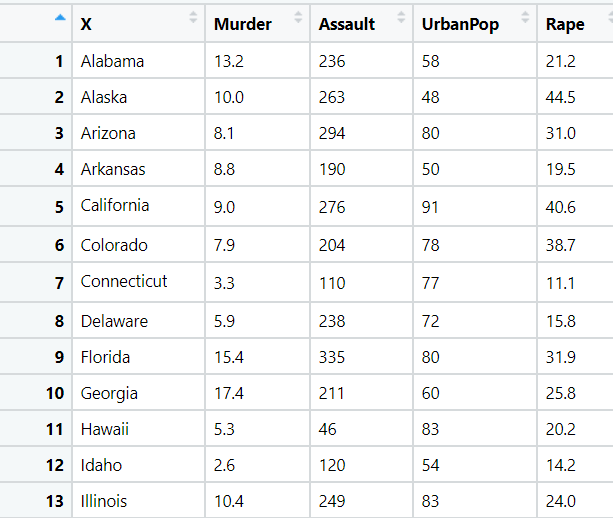


* Finally, the output is stored in the csv file into working directory **(getwd).**

**Inferences:**

* Cluster 5 has high Number of miles eligible for award and bonus\_trans in the past 12 months
* Cluster 4 has Number of miles earned with Small Business credit card in the past 12 months and Number of miles earned from bonus transn in the past 12 months
* Cluster 3 has Number of miles counted as qualifying for Topflight status and Number of miles earned with freq. flyer credit card in the past 12 months:
* Cluster 2 has Number of miles earned with Rewards credit card in the past 12 months
* Cluster 1 has least performance compared to other groups

2.) Perform Clustering for the crime data and identify the number of clusters formed and draw inferences.



**Ans:**

**Importing the data set directly as there is no output variable.**

**Data Preprocessing:**

* Columns are rearranged to have the easy access of the variables.
* In R , dummy variable are created automatically when object(x) is created as model. Matrix command has inbuild feature of converting dummy variables or we can factorize the variables whereas, in python we need to write label encoder () code for creation of dummy variables.
* Removing unnecessary columns and checking for the NA value

**Normalizing the data:**

* Using the normalization function, the data is normalized and the data will come under same level.

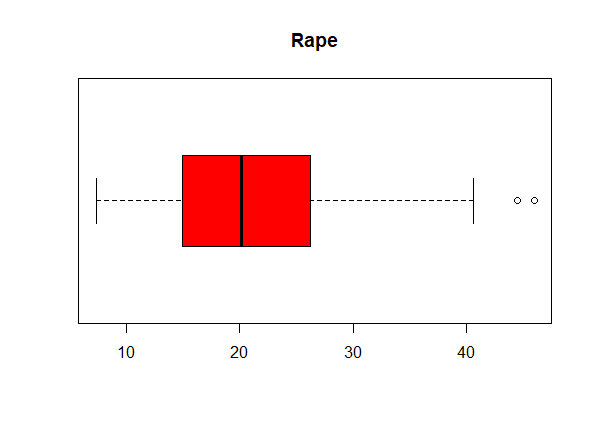
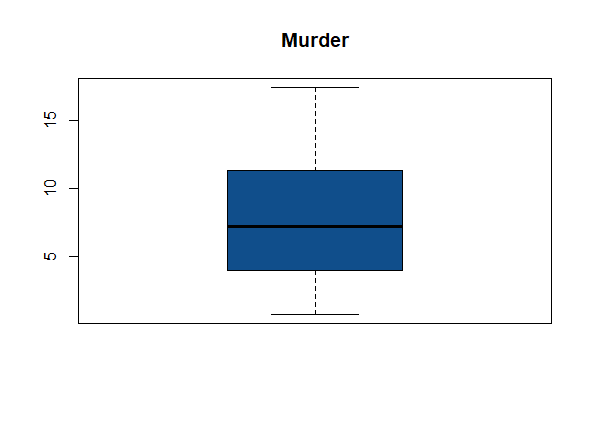
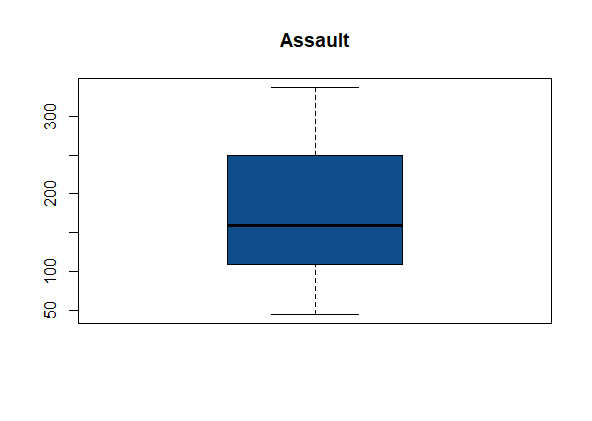
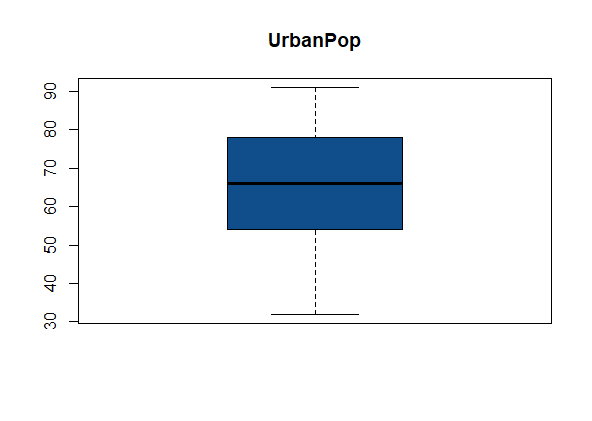
**def norm\_func(i):**

**x=(i-i.min()/i.std())**

**return (x)**

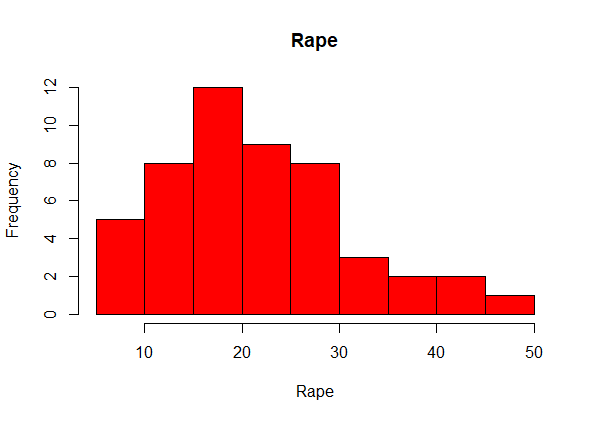
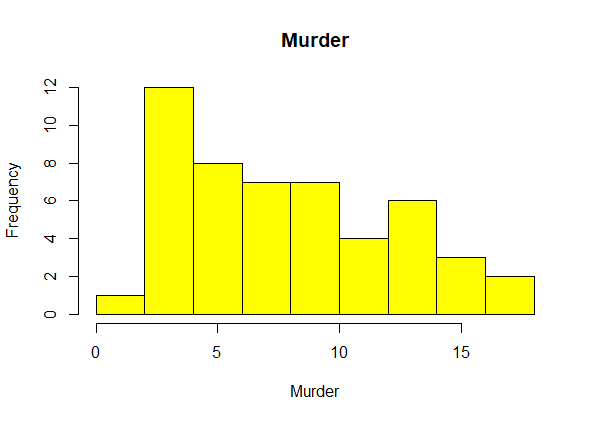
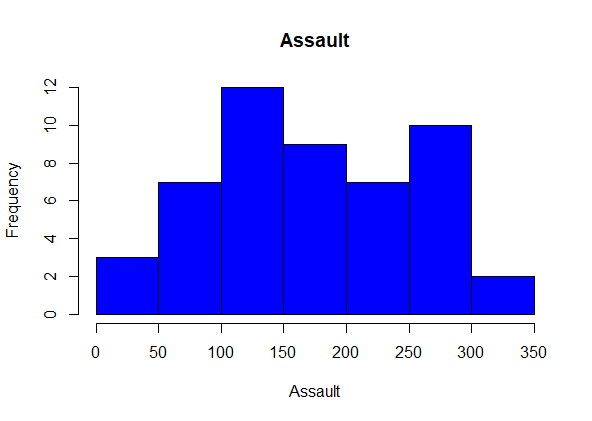
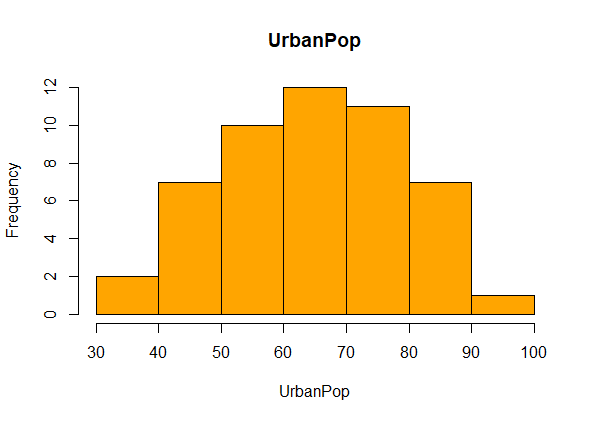
**Exploratory Data Analysis:**

**Box Plot Representation:**



* From the above graphical Representation, its clearly shows there exists no outliers

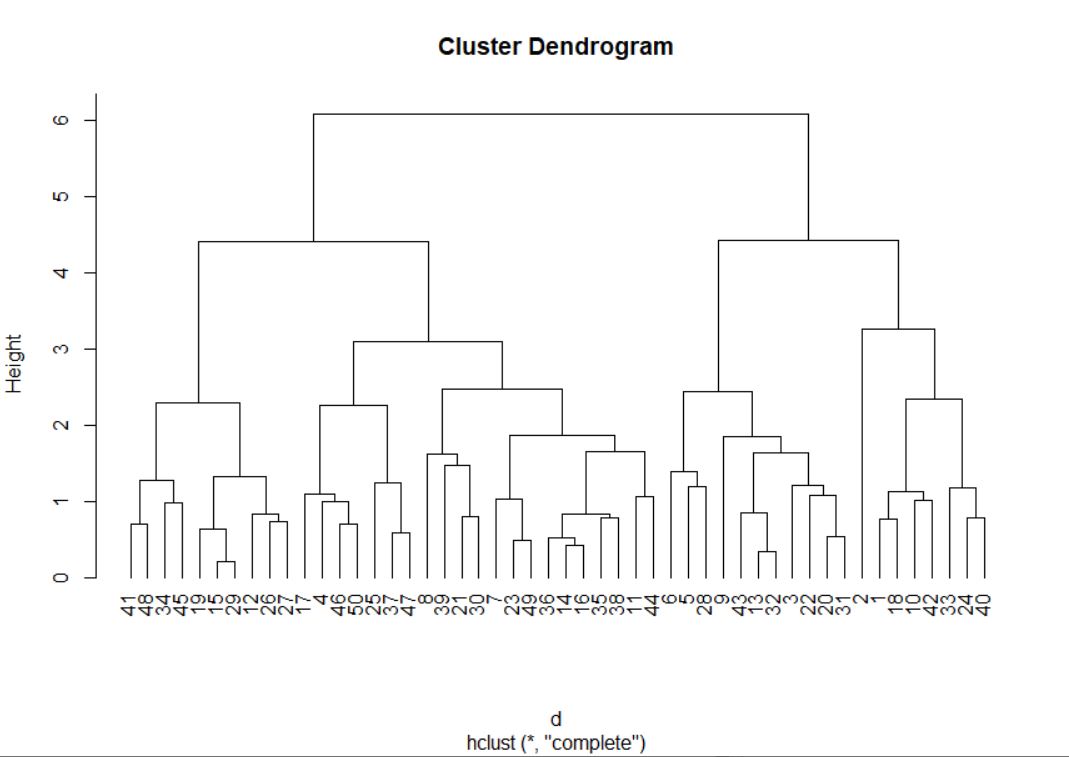
**Histogram Representation:**



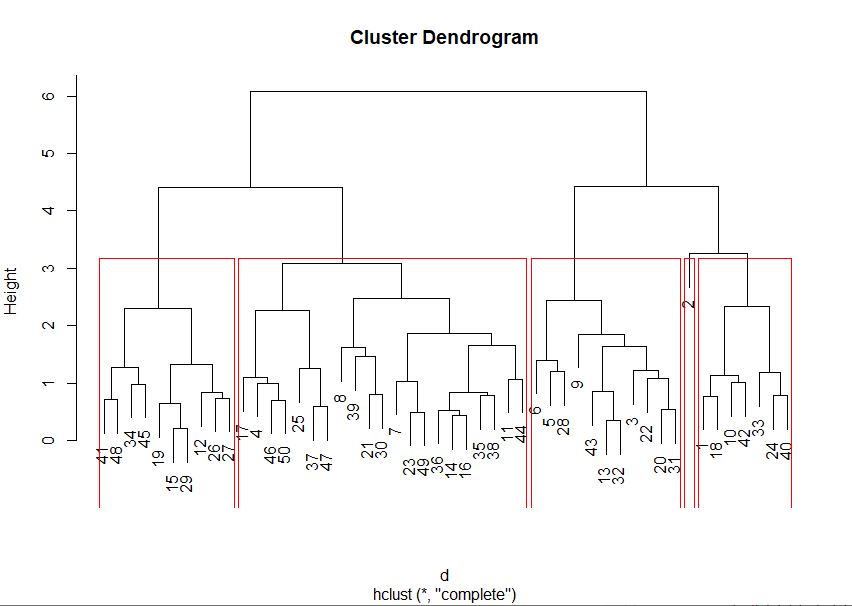
* From the above histogram, its clearly shows the data is normally distributed across all variables.

**Dendrogram:**

* If Standardization/Normalization of data is not done then below issues will influence the model.
* Also Distance measure will be calculated incorrectly, if all the variables are not brought to equal weight.
* Largest scale dominating the measure
* The below dendogram generated with the given data set with standardization is as below.
* Using “**Complete”** method the dendogram is shows as below.



* Now the dendogram is divided into clusters separately with red bordered line as shown below and the aggregate of each cluster is compared with the other clusters.

****

**Inferences:**

* Cluster 5 has least number of crimes compared to all the other groups but urban pops are high.
* Cluster 4 has high Murder, Assault and rape rate.
* Cluster 2 also has high Murder, Assault and rape rate more than Cluster 4.
* Cluster 3 has high Assault rate.

# Hints:

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

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

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy Ex: Feature Engineering, Hyper Parameter tuning, etc.
3. All the codes (executable programs) are running without errors
4. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here