TABLE OF CONTENTS

QUESTION NO 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.

SUMMARY DATA:

Balance	Qual_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award.
Min. : 0	Min. : 0.0	Min. : 0	Min. : 0.0	Min. : 0.0	Min. : 0.000	Min. : 2	Min. :0.0000
1st Qu.: 18528	1st Qu.: 0.0	1st Qu.: 1250	1st Qu.: 3.0	1st Qu.: 0.0	1st Qu.: 0.000	1st Qu.:2330	1st Qu.:0.0000
Median : 43097	Median : 0.0	Median : 7171	Median :12.0	Median : 0.0	Median : 0.000	Median :4096	Median :0.0000
Mean : 73601	Mean : 144.1	Mean : 17145	Mean :11.6	Mean : 460.1	Mean : 1.374	Mean :4119	Mean :0.3703
3rd Qu.: 92404	3rd Qu.: 0.0	3rd Qu.: 23801	3rd Qu.:17.0	3rd Qu.: 311.0	3rd Qu.: 1.000	3rd Qu.:5790	3rd Qu.:1.0000
Max. :1704838	Max. :11148.0	Max. :263685	Max. :86.0	Max. :30817.0	Max. :53.000	Max. :8296	Max. :1.0000

Data contains 12 columns, out of which 3 column contains factor values.

HIERARCHICAL CLUSTERING:

Before going to hierarchical clustering, I build a function for normalization as well as converting dummy variable for the factor columns. I created another function to check whether which of the 8-linkage method is applicable for my hierarchical clustering.

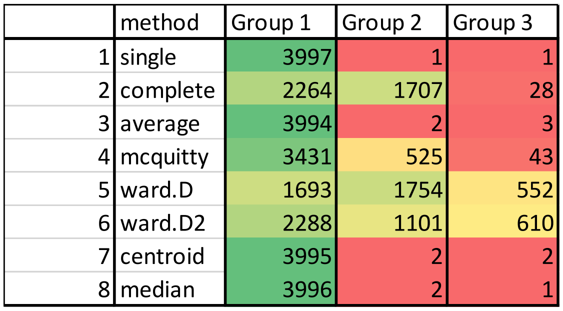
- normalize_dummy (x) # Here x should be in data frame

  No matter whether your data contains factor column or not. In this function all numeric columns will be transformed to normalize and the factor columns get their dummies respectively.
- all_hclust (dist,k) # Here dist is distance matrix, and k is no of clusters (cuttree) you want.

  This function is for clustering purpose, where we can see the results of all the clusters generated by the 8 linkages "single", "complete", "average", "mcquitty", "ward.D", "ward.D2", "centroid","median". So, now I don't have to check for each and every clusters. I can choose as my desire.

STEPS PERFORMED:

1. Converted the type of the 3 numeric columns i.e. (cc1_miles, cc2_miles, cc3_miles) as factor using as.factor.
2. Using these two functions I normalize my data using normalize_dummy.
3. Calculate the distance matrix using dist function.
4. Calculated all the 8 possible hierarchical-clusters using function all_hclust.
5. Viewed the no of points in each cluster.



	method	Group 1	Group 2	Group 3
1	single	3997	1	1
2	complete	2264	1707	28
3	average	3994	2	3
4	mcquitty	3431	525	43
5	ward.D	1693	1754	552
6	ward.D2	2288	1101	610
7	centroid	3995	2	2
8	median	3996	2	1

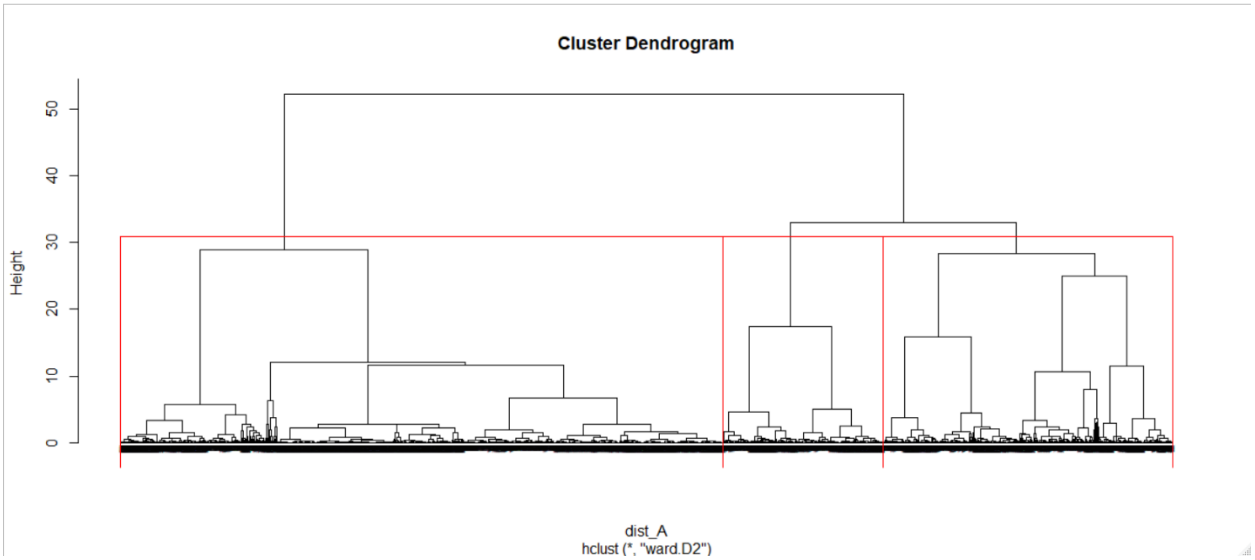
6. From the above plot we come up with the best method will be ward.D2.
7. Then we find out for the mean for numeric data and mode for categorical data as the representative of my clusters.

Group.1	Balance	Qual_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award.	cc1_miles
1	50475.78	143.98	3913.68	6.68	462.87	1.39	3754.43	0.25	1.00
2	116321.43	162.23	43138.42	19.10	536.66	1.62	4763.46	0.57	4.00
3	83234.79	111.94	19856.30	16.55	311.24	0.87	4320.37	0.46	3.00

8. Here in this above table Green labeled are Higher Values, Red values are Least values, and Yellow are moderate values.
9. Then I take a subsample from the whole record and perform the same clustering Again to see the Expected values in the Grouped data.

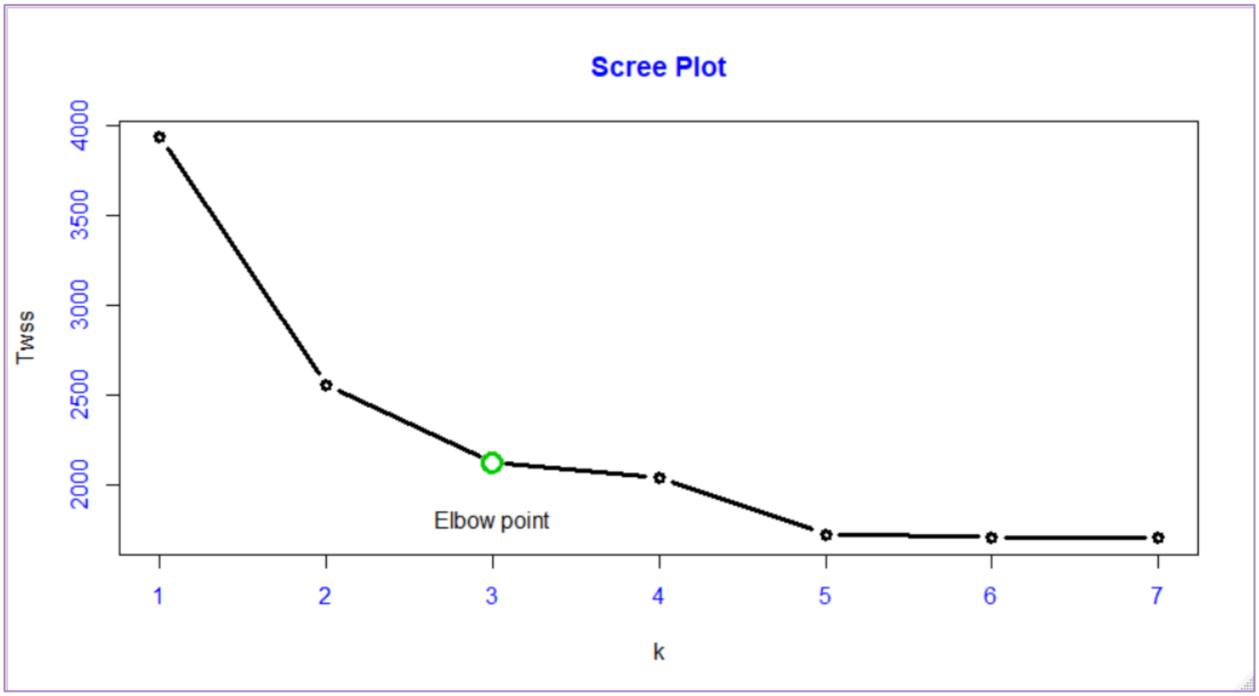
Group.1	Balance	Qual_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award.	cc1_miles
1	49055.92	152.08	3740.67	6.59	449.81	1.35	3734.89	0.24	1.00
2	85906.15	90.32	19830.39	16.38	315.73	0.87	4403.89	0.48	4.00
3	118905.84	143.07	43995.69	19.25	549.26	1.68	4733.21	0.58	3.00

10. Looking at the 2 patterns in the above two tables here we can see cluster 2 in sampled data is similar as in cluster 3 in sampled data.
    we can see that the pattern is differ only for column Qual_miles and cc1_miles.
    As overall allocation is kind similar in case of both Sampled and Whole data, so we can say that we may rely in this grouping.
11. Then I plotted the Dendrogram. And saved it as a pdf file, as we can’t visualize it properly inside our RStudio.



K-MEANS CLUSTERING

For Deciding my K value, i.e. deciding my number of clusters to take, I go for my Scree plot with different values of k from 1 to 10.



Here I got my Elbow point as 5 so I decide my k as 3.

And I perform my Clustering for k as 3

Group no →	1	2	3
Number of Records	568	1710	1721

Group.1	Balance	Bonus_ miles	Bonus_ trans	Flight_miles _12mo	Flight_trans _12	Days_since _enroll	Award.
1	87144.18	7986.81	9.49	1137.38	3.48	4330.85	1.00
2	104895.91	34789.08	18.21	457.59	1.36	4606.83	0.53
3	38037.08	2635.91	5.73	238.96	0.70	3563.35	0.00

From Above We can say that the 3rd cluster is always having the least values for each and every column.

CONCLUSION:

Group 1 represents the 14% of the travelers, which are most frequent travelers according to our data set, also all of the travelers are awarded with offers as we can see award = 1. Their Flight miles in 12 months are very much as compare to other groups. These Customers are associated with the East West Airlines since a long time.

Group 2 represents 42% of the travelers (customers) who are with once or twice travelled in the whole year as flight transition is 1.36, out of which about 53% are awarded with offers. So, we may say that this group consists of Middle-class Travelers.

Group 3 represents 43% of the travelers, who are not at all awarded as they are our non-frequent and new customers. They have flight transitions over the 12 months is nearly zero. So, we may consider them as our non-frequent travelers.

From the above we can say that, we should focus on the major part of our customers i.e. 43%, which are in our group 3. If we do some concession or offers for them, maybe we come up with profit in our business. Those who are in group 1, those are our frequent travelers, so they don’t bother about any kind of concessions.

QUESTION NO 2

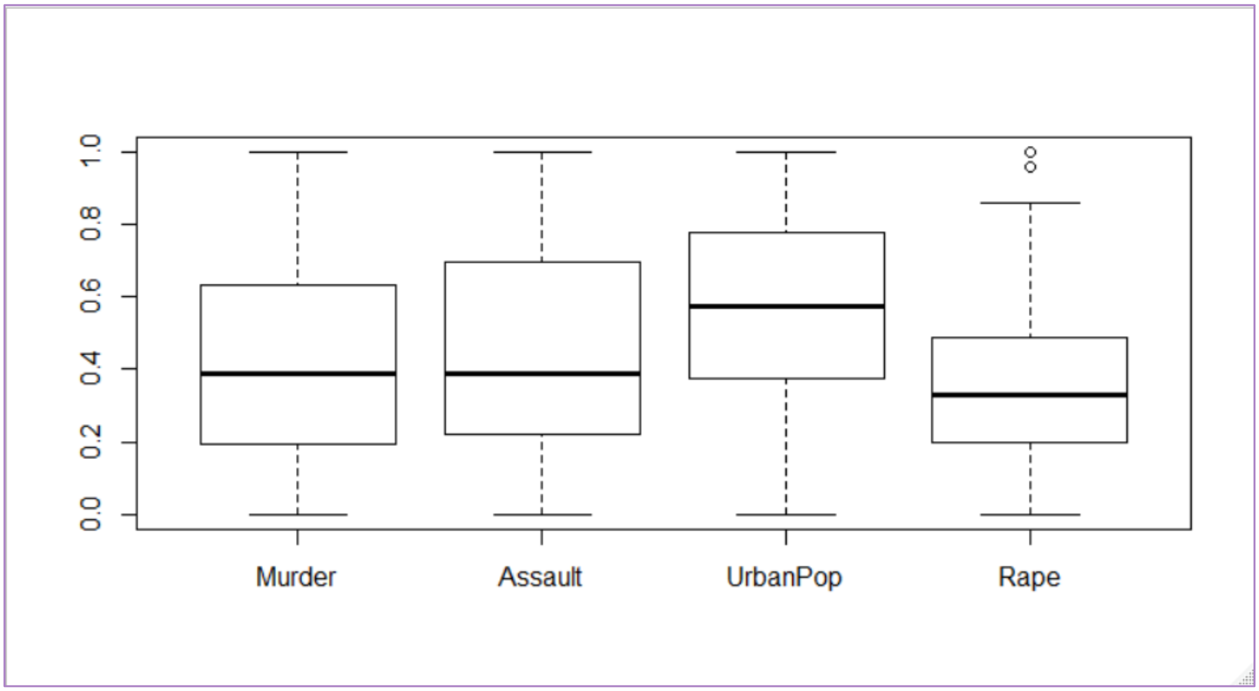
Perform Clustering for the crime data and identify the number of clusters formed and draw inferences.

SUMMARY DATA :

Murder	Assault	Urban Pop	Rape
Min. : 0.800	Min. : 45.0	Min. :32.00	Min. : 7.30
1st Qu.: 4.075	1st Qu.:109.0	1st Qu.:54.50	1st Qu.:15.07
Median : 7.250	Median :159.0	Median :66.00	Median :20.10
Mean : 7.788	Mean :170.8	Mean :65.54	Mean :21.23
3rd Qu.:11.250	3rd Qu.:249.0	3rd Qu.:77.75	3rd Qu.:26.18
Max. :17.400	Max. :337.0	Max. :91.00	Max. :46.00

Mean and median of every column seems like approximately same, so we can say there may be no outliers, if than in very less numbers.

BOXPLOT:



From the boxplot for normalized data, we can see only 2 outliers in the column “Rape”. So we may proceed with this Data for my clustering.

STEPS PERFORMED

1. Normalized the using the function I defined prior i.e. normalize_dummy,.
2. Calculated distance among each and every records in my data, using the Euclidian distance and the function dist.
3. Calculated all possible clusters for my crime data using my function all_hclust. With number of cluster as 3.
4. The Table of linkage methods and the number of records in each cluster can be seen by this table.

	METHOD	GROUP 1	GROUP 2	GROUP 3
1	**SINGLE**	**48**	**1**	**1**
2	**COMPLETE**	**20**	**20**	**10**
3	**AVERAGE**	**19**	**1**	**30**
4	**MCQUITTY**	**22**	**21**	**7**
5	**WARD.D**	**20**	**18**	**12**
6	**WARD.D2**	**20**	**18**	**12**
7	**CENTROID**	**48**	**1**	**1**
8	**MEDIAN**	**37**	**1**	**12**

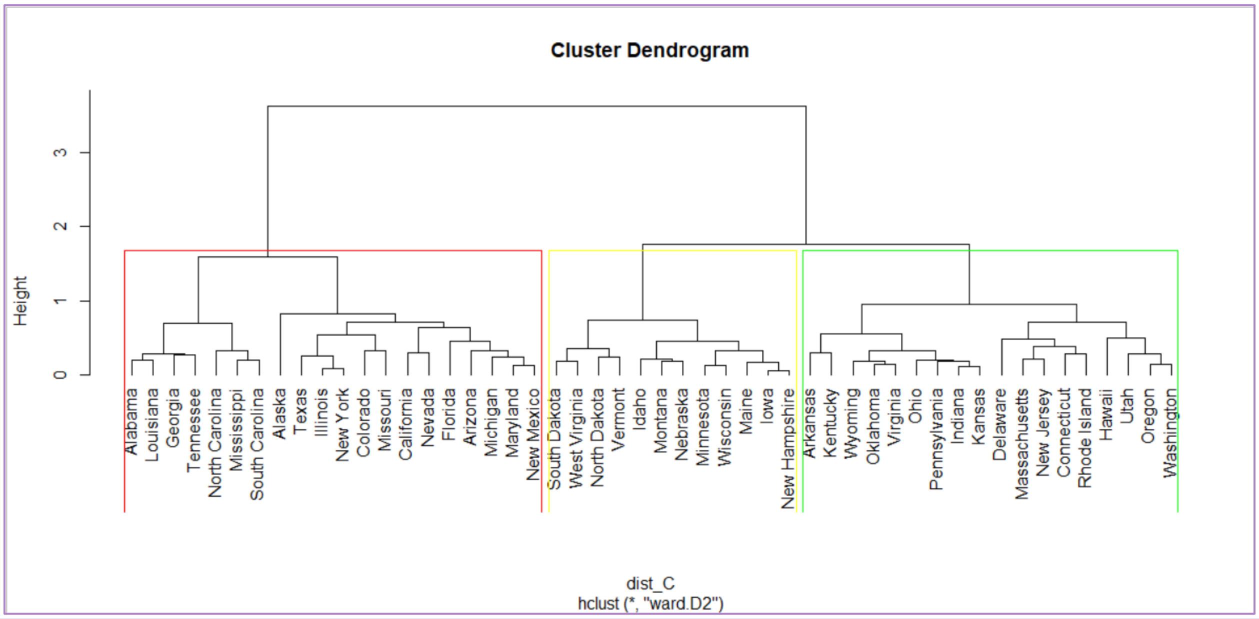
5. From this I may consider “complete”, or “ward.D2”, so I proceed with the ward.D2 clustering method calculated expected values for each column with respect to each groups.

Group.1	Murder	Assault	UrbanPop	Rape
1.00	12.17	255.25	68.40	29.17
2.00	6.06	140.06	71.33	18.68
3.00	3.09	76.00	52.08	11.83

6. Dendrogram:
7. From the Dendrogram we can classify the Place names and their Crime Rate.
8. Then we performed the same Experiment to check the stability of clustering with respect to a new sample from the Crime data.

Population	Murder	Assault	UrbanPop	Rape
1.00	12.17	255.25	68.40	29.17
2.00	6.06	140.06	71.33	18.68
3.00	3.09	76.00	52.08	11.83

Sample	Murder	Assault	UrbanPop	Rape
1	12.18182	260.6364	69.45455	29.85455
2	3.136364	74.63636	55.63636	12.87273
3	6.35	152.8333	70.94444	20.45



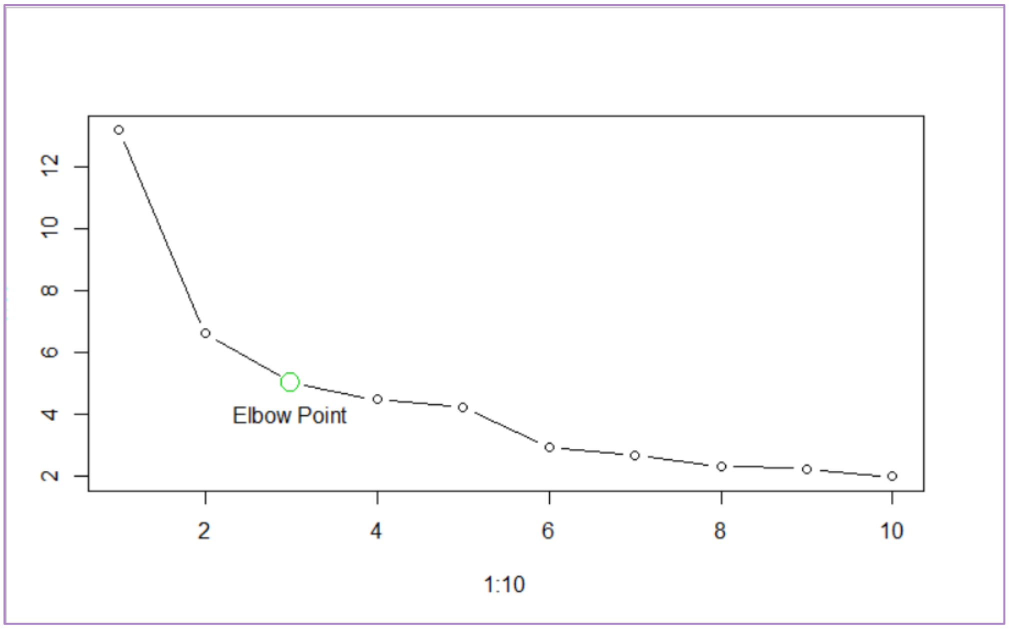
CONCLUSION:

If we look at the clusters here, group 1 from population is similar as group 1 in sample 1, mean difference is very little. Group 2 from population is similar as group 3 in sample.

So we may say that those places which belongs to group 1 are at high risk, and places which belongs to group 2 is also considered as a Risky place for visitors, but little bit less compare to group 1, among these 3 groups , the places belongs to group 3 may be considered as least risk for the tourists.

K-MEANS CLUSTERING

To choose my best k value for performing K-Means clustering, I do my Scree Plot



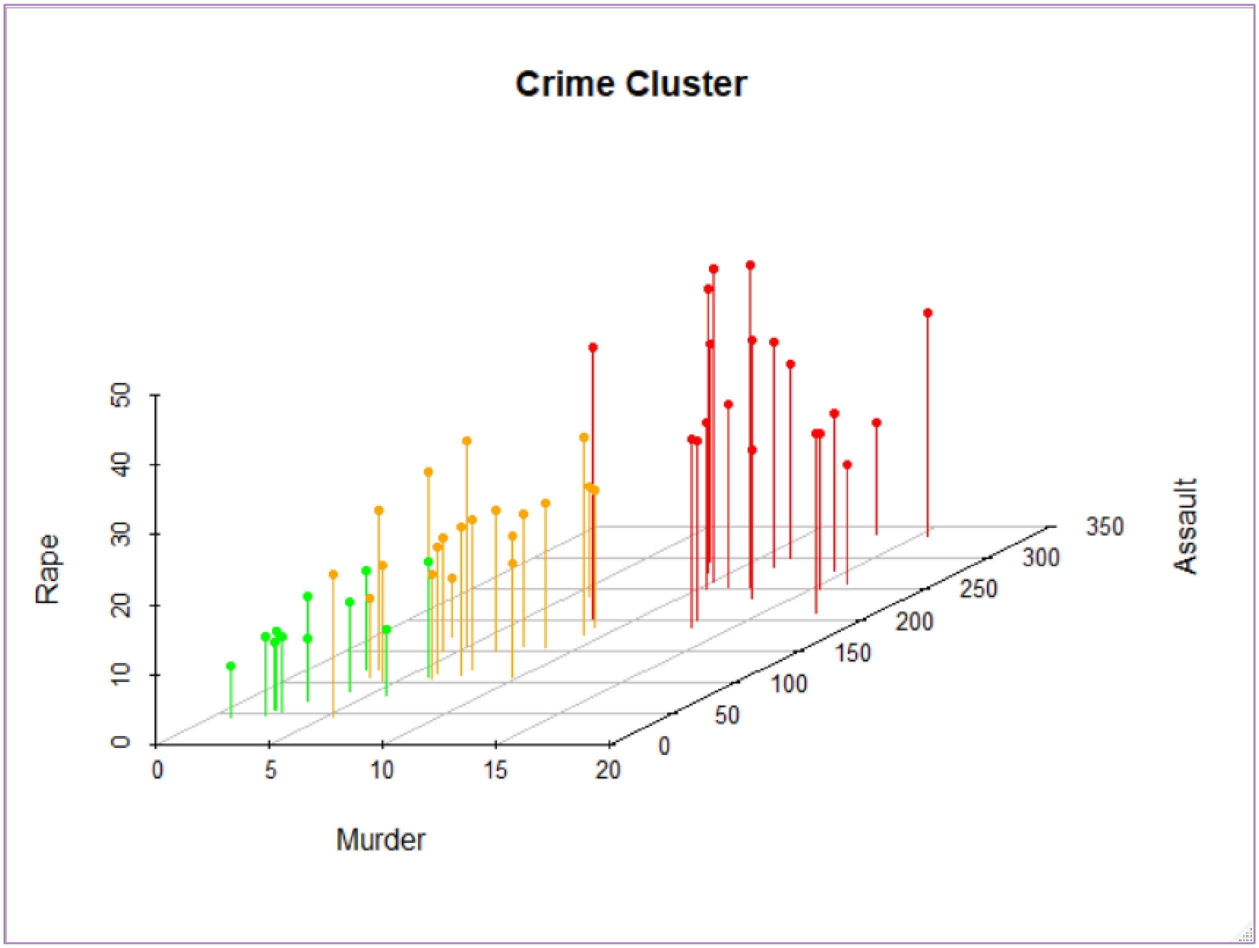
Here I got my elbow point for k as 3, so I perform my K-Means Clustering with k (number of cluster as 3).

I come up with the Result in the Tabular format.

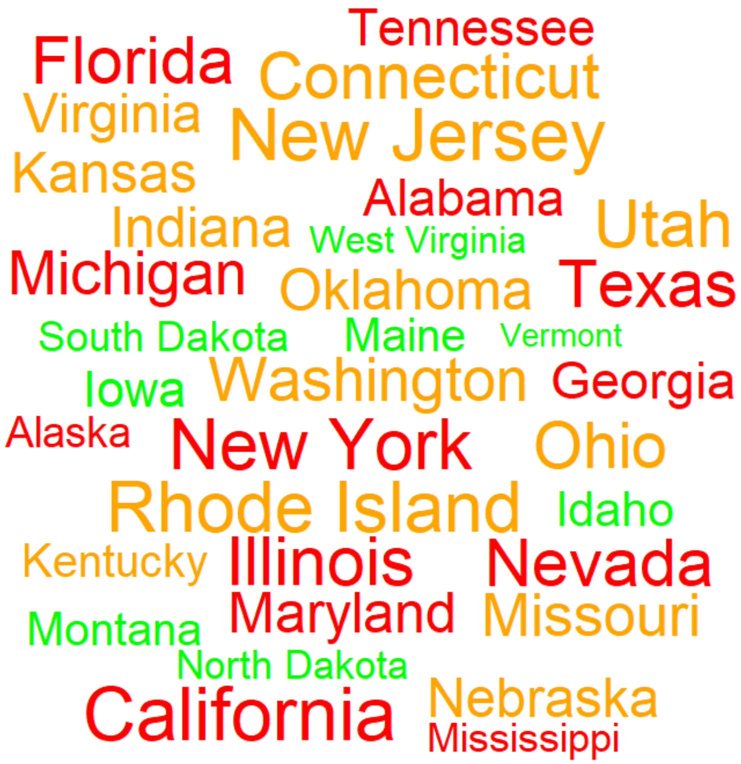
Group.1	Murder	Assault	UrbanPop	Rape
1	2.981818	73.63636	51.18182	11.40909
2	12.33158	259.3158	68.31579	29.21579
3	6.115	140.05	70.8	19.05

The Expected values in each of our ariables with respect to the groups of clusters matches to the output of the Hierarchical Clustering.

SCATTER PLOT:



WORD CLOUD:



This word cloud TEXT SIZE is based on the population of the locations and TEXT COLOUR based on the crime rate.

Red represents higher crime rate compare to all, yellow represents moderate and green represents least crime rate areas.

CONCLUSION:

Here each dot is representing each location, based on color green colored dots have low chances of crime in the location, whereas yellow for moderate crime and red for High rate in crime.

We may Focus on our group 2 i.e. red areas where crime rate is very high, we may deploy more Interceptors for those areas which locations are in group 2 with exotic series of vehicles with armed force for night duty.

Based on group 3 where we can find moderate crime, there we may create Awareness for the safety of people and make the Police department quick responsive introducing some sports series vehicles for them.

Sorry for the funny conclusion.