PRACTICAL 8

Q1. Perform the following tasks:

Load the bank data set in Weka. Explain the characteristics of this dataset.

No. of attributes: 9

No. of instances: 300

Sum of weights : 300

Q2. Write down the following details regarding the attributes:

a. names

b. types

c. values

|  |  |  |  |
| --- | --- | --- | --- |
| Sr.no. | Attribute Name | Type | Values |
| 1 | Age | Numeric | Min:18 Max:67 Mean:42.57 StdDev: 14.22 |
| 2 | Sex | Nominal | Male,Female |
| 3 | Region | Nominal | Inner\_city, rural, town, suburban |
| 4 | Income | Numeric | Min: 5014.21 Max: 63130.1 |
| 5 | Married | Nominal | Yes, No |
| 6 | Children | Nominal | Yes, No |
| 7 | Car | Nominal | Yes, No |
| 8 | Mortgage | Nominal | Yes, No |
| 9 | Pep | Nominal | Yes, No |

Q3. Run the Simple K-Means clustering algorithm on the dataset. Consider no. of clusters as 6.

a. What do you mean by seed value?

The **seed** number (any integer) is the randomization for your initial **K** points. **K** represents the number of clusters.

b. What are the number of instances and percentage figures in each cluster?

Clustered Instances

0 74 ( 25%)

1 56 ( 19%)

2 31 ( 10%)

3 47 ( 16%)

4 43 ( 14%)

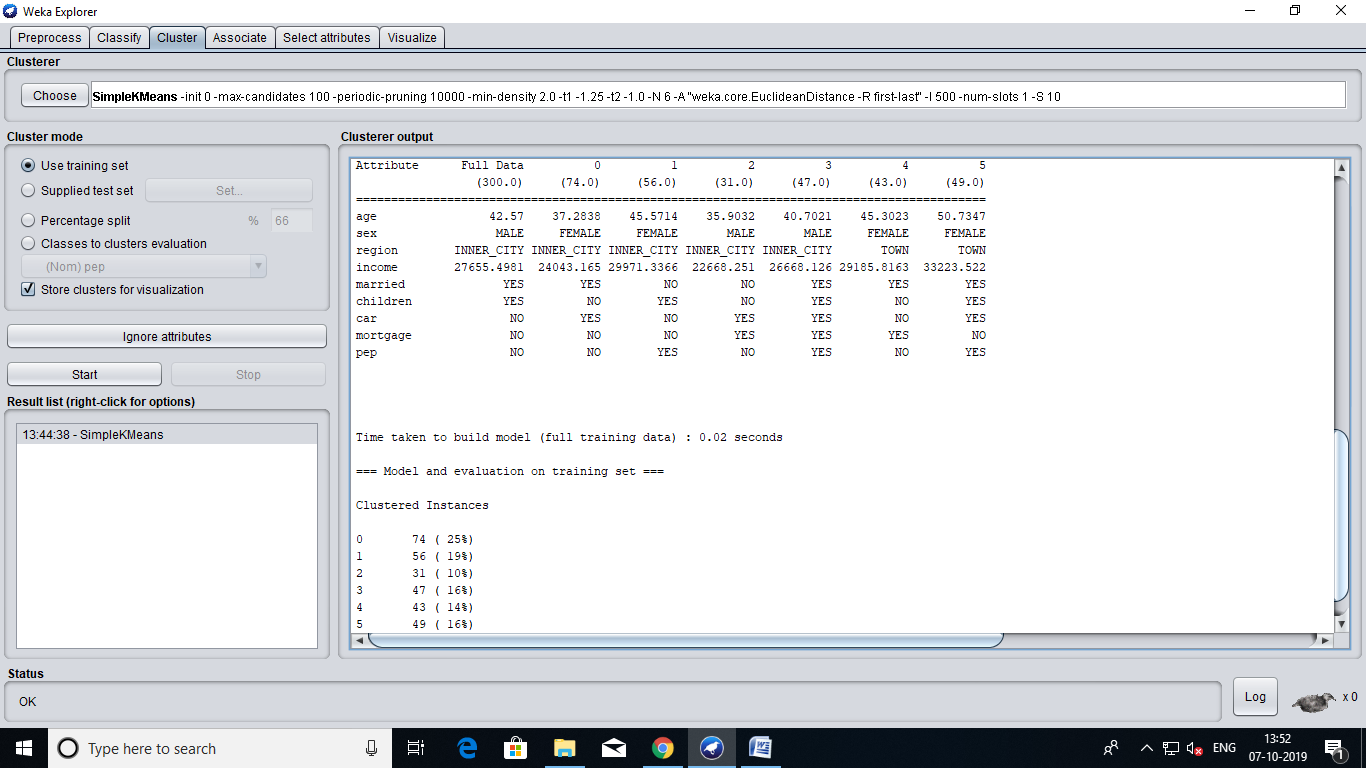
5 49 ( 16%)

c. What is the number of iterations that were required?

Number of iterations required: 4

d. What is the sum of squared errors? What does it represent?

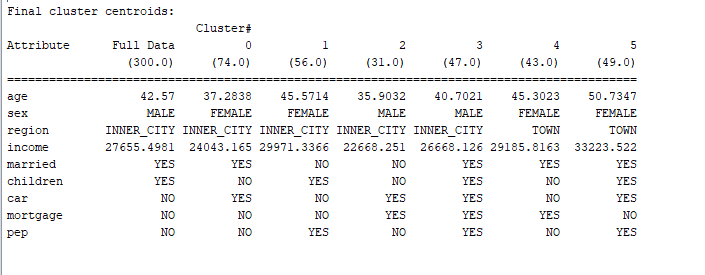
Within cluster sum of squared errors: 540.7387788014682



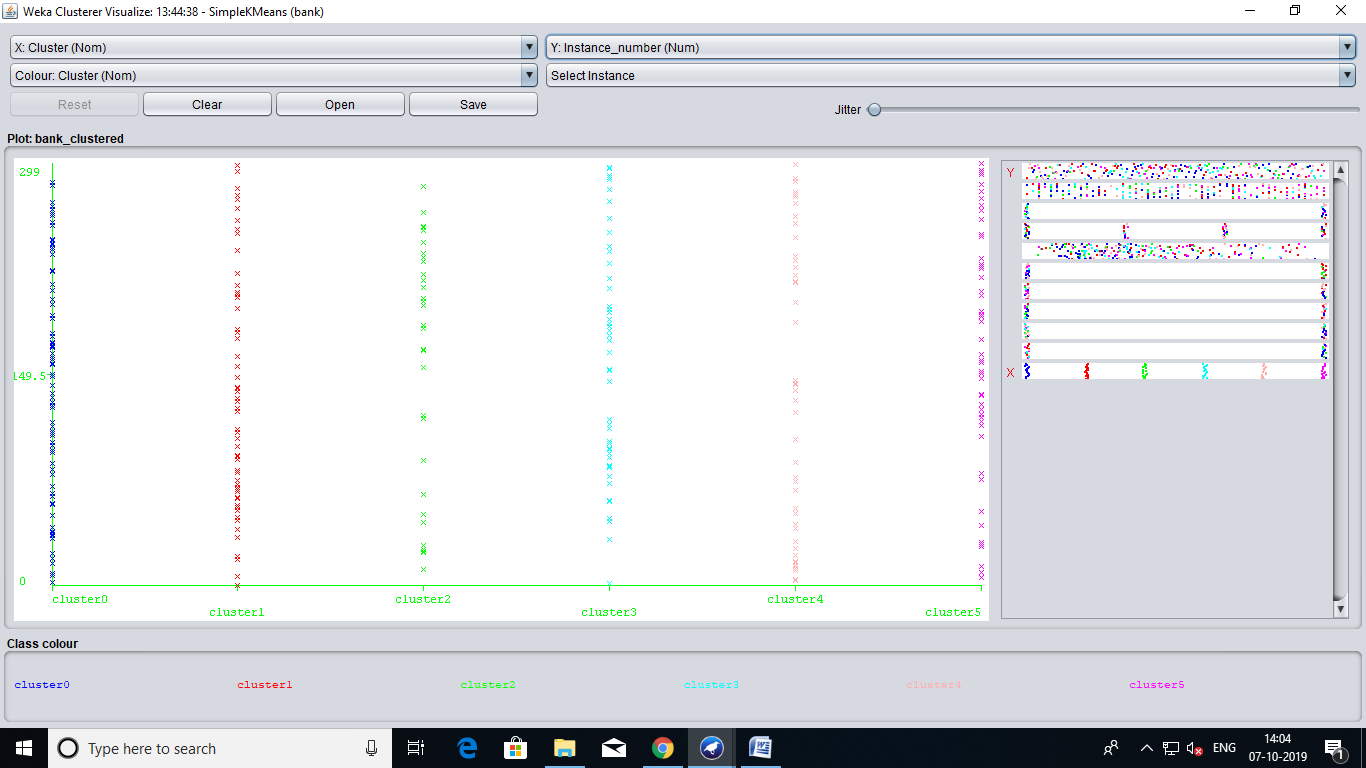
For each point, the error is the distance to the nearest cluster – To get SSE, we square these errors and sum them.

Q4. What do you mean by centroid? What does it represent? Tabulate the characteristics of the centroid of each cluster.

The centroid is (typically) the mean of the points in the cluster.



Q5. Visualize the results of this clustering (let the X-axis represent the cluster name, and the Y axis represent the instance number)



a. Is there a significant variation in age between clusters?

No

b. Which clusters are predominated by males and which clusters are predominated by females?

Clusters predominated by males: Cluster 2, Cluster 3

Clusters predominated by females: Cluster 0, Cluster 1, Cluster 4, Cluster 5

b. What can be said about the variation of income between clusters?

c. Which clusters are dominated by married people and which clusters are dominated by unmarried people?

Clusters predominated by married: Cluster 0, Cluster 3, Cluster 4, Cluster 5

Clusters predominated by unmarried: Cluster 1, Cluster 2

d. How do the clusters differ with respect to the number of children?

Clusters predominated by children: Cluster 1, Cluster 3, Cluster 5

Clusters predominated by “no children” : Cluster 0, Cluster 4

Cluster 2 has almost same distribution in both yes and no classes.

e. Which cluster has the highest number of people with cars?

Cluster 0

f. What can be said about the variation of mortgage holdings between clusters?

Clusters 0,1 and 5 have least mortgage holdings.

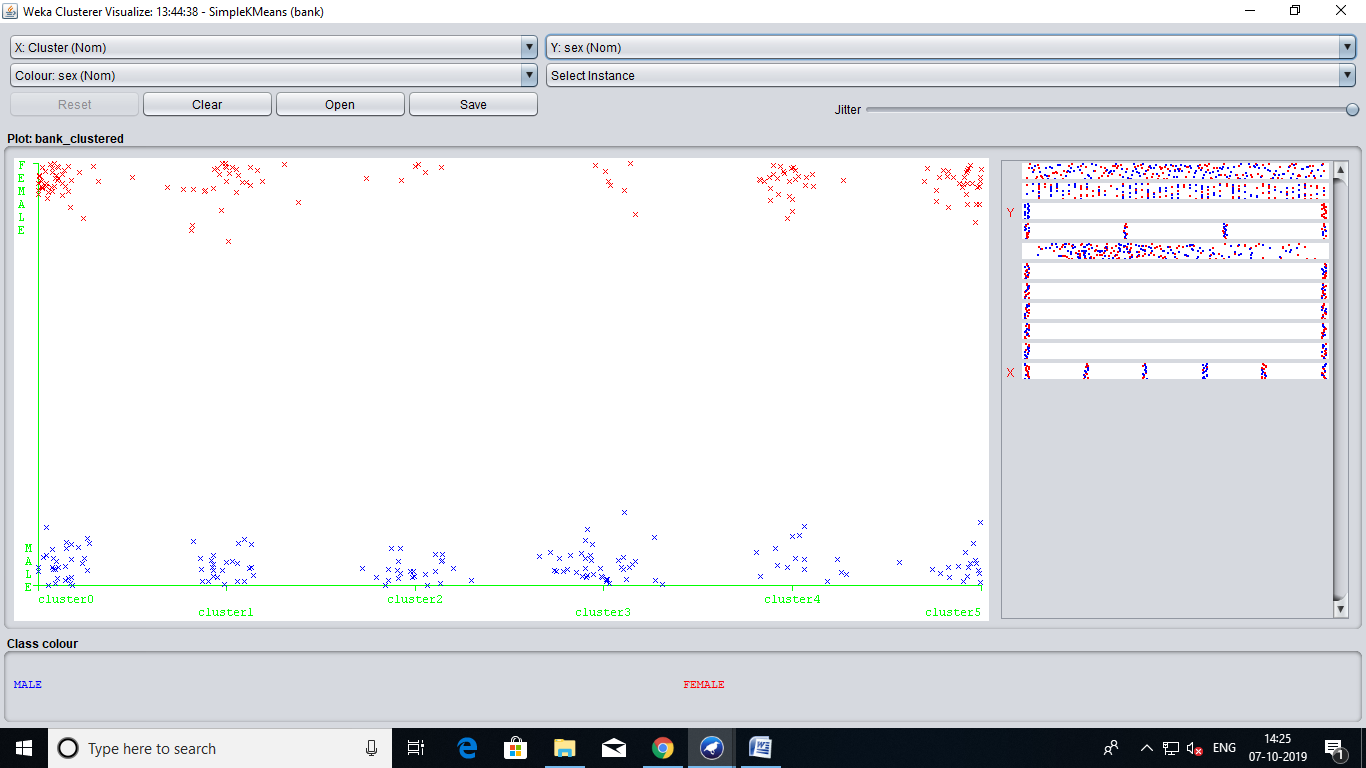
g. Which clusters comprise mostly of people who buy the PEP product and which ones are

comprised of people who do not buy the PEP product?

Clusters which comprise mostly of people who buy PEP product : 1,3,5

Clusters which comprise mostly of people who do not buy PEP product : 0,2,4

Q6. Select sex attribute as the color dimension and visualize the result, the distribution of males and females in each cluster. (Note: In this case, by changing the color dimension to other attributes, we can see their distribution within each of the clusters.)



Q7. Finally, save the resulting data set which included each instance along with its assigned cluster. To do so, we click the Save button in the visualization window and save the result as the file “bank -kmeans.arff”.

Q8. Run the SimpleKMeans algorithm for values of K (no. of clusters) ranging from 2 to 12.

Tabulate the sum of squared errors for each run. What do you observe about the trend of the

sum of squared errors?

|  |  |
| --- | --- |
| K | Within cluster sum of squared errors: |
| 2 | 775.1756576878267 |
| 3 | 681.9387170881896 |
| 4 | 625.4296762510829 |
| 5 | 563.7995027565171 |
| 6 | 540.7387788014682 |
| 7 | 524.3484219577417 |
| 8 | 488.1650647851437 |
| 9 | 470.6564571657609 |
| 10 | 466.86014843503983 |
| 11 | 443.0877285538886 |
| 12 | 420.93289947544827 |

As the value of k increases, the cluster sum of squared error decreases.

Q9. For the run with K=12, answer the following questions:

a. Is there a significant variation in age between clusters?

b. Which clusters are predominated by males and which clusters are predominated by females?

Clusters pre-dominated by males: 2,3,7,8,9,10,11

Clusters pre-dominated by females: 0,1,4,5,6

c. How do the clusters differ with respect to the number of children?

Clusters pre-dominated by children: 1,3,5,7,9,11

Clusters pre-dominated by “no-children”: 0,2,4,6,8,10

d. Which clusters comprise of people who buy the PEP product and which ones are comprised of people who do not buy the PEP product?

Clusters which comprise mostly of people who buy PEP product : 0,1,3,5,9

Clusters which comprise mostly of people who do not buy PEP product : 2,4,6,7,8,10,11

e. Do you see any differences in your ability to evaluate the characteristics of clusters generated for K=6 versus K=12? Why does this difference arise?

Within cluster sum of squared errors for k=6 : 540.7387788014682

Within cluster sum of squared errors for k=12 : 420.93289947544827

**Part 2**

Use DBSCAN clustering algorithm on the same dataset and observe the results. What canyou

conclude about DBSCAN and simple k means algorithm?

For DBSCAN:

Within cluster sum of squared errors: 775.1756576878267