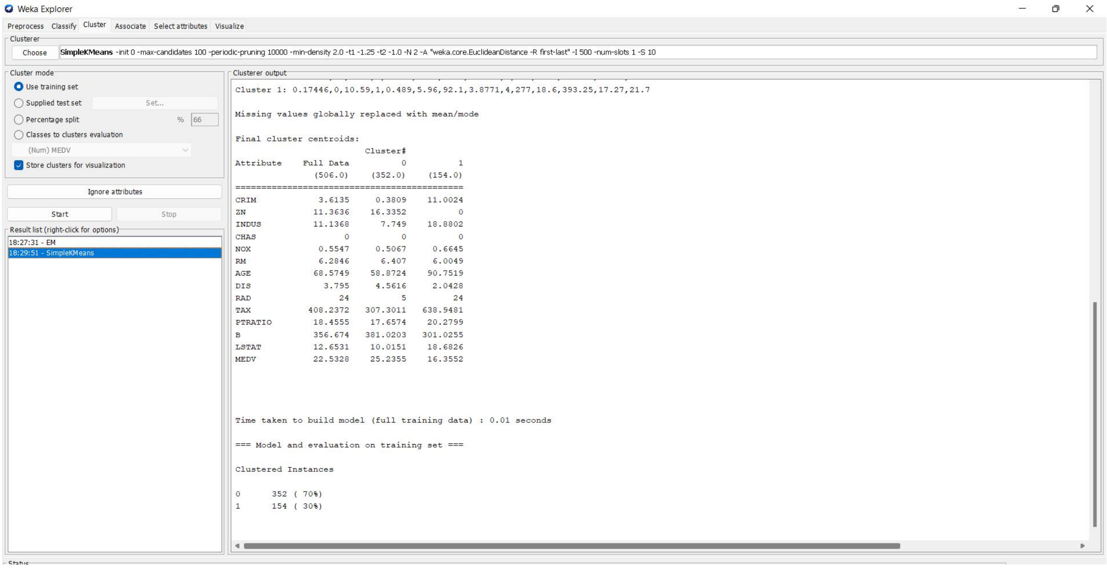
EXPERIMENT-6

DATA SEGMENTATION BY K- MEANS CLUSTER USING WEKA AND R-TOOL

Name: S.G.DEVSACHIN
Reg.No: 192111088
Subject: CSA1672, Data warehouse and data mining

OUTPUT:

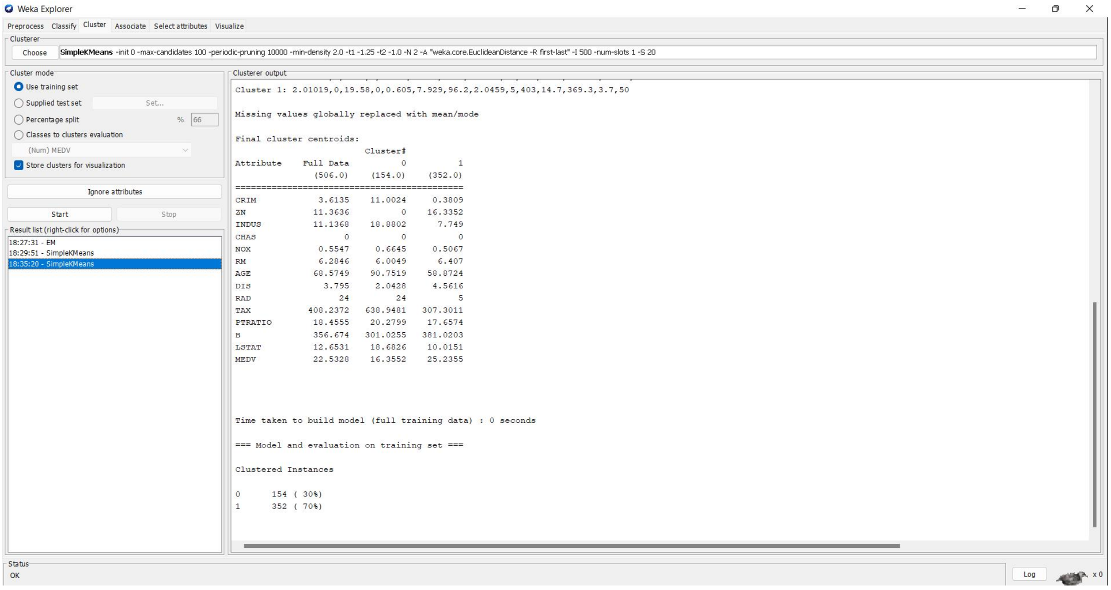
1)**Choose a set of attributes for clustering and give a motivation.**



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

```
Cluster 1: 0.17446,0,10.59,1,0.489,5.96,92.1,3.8771,4,277,18.6,393.25,17.27,21.7

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
               (506.0)    (352.0)    (154.0)
============================================
CRIM            3.6135     0.3809    11.0024
ZN             11.3636    16.3352          0
INDUS          11.1368      7.749    18.8802
CHAS                 0          0          0
NOX             0.5547     0.5067     0.6645
RM              6.2846      6.407     6.0049
AGE            68.5749    58.8724    90.7519
DIS              3.795     4.5616     2.0428
RAD                 24          5         24
TAX           408.2372   307.3011   638.9481
PTRATIO        18.4555    17.6574    20.2799
B              356.674   381.0203   301.0255
LSTAT          12.6531    10.0151    18.6826
MEDV           22.5328    25.2355    16.3552




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      352 ( 70%)
1      154 ( 30%)
```

2)Experiment with atleast 2 different number of clusters but with same seed values:



Weka Explorer

Clusterer: SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 20

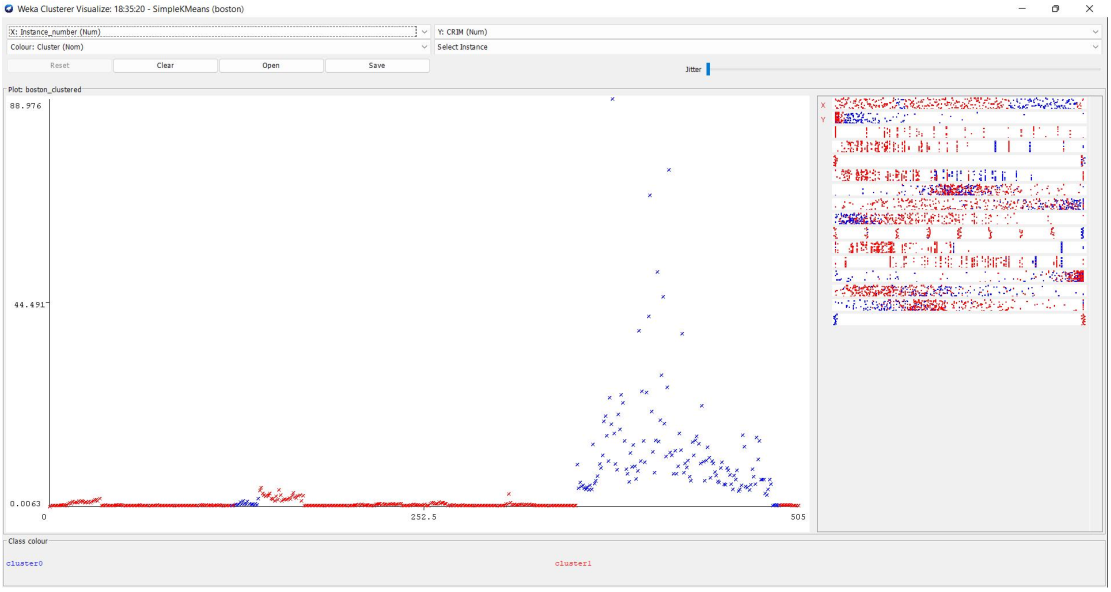
Cluster mode:
- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation (Num) MEDV
- ☑ Store clusters for visualization

Result list (right-click for options)
- 18:27:31 - EM
- 18:29:51 - SimpleKMeans
- 18:35:20 - SimpleKMeans

Clusterer output

```
Cluster 1: 2.01019,0,19.58,0,0.605,7.929,96.2,2.0459,5,403,14.7,369.3,3.7,50

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
               (506.0)    (154.0)    (352.0)
============================================
CRIM            3.6135    11.0024     0.3809
ZN             11.3636          0    16.3352
INDUS          11.1368    18.8802      7.749
CHAS                 0          0          0
NOX             0.5547     0.6645     0.5067
RM              6.2846     6.0049      6.407
AGE            68.5749    90.7519    58.8724
DIS              3.795     2.0428     4.5616
RAD                 24         24          5
TAX           408.2372   638.9481   307.3011
PTRATIO        18.4555    20.2799    17.6574
B              356.674   301.0255   381.0203
LSTAT          12.6531    18.6826    10.0151
MEDV           22.5328    16.3552    25.2355




Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      154 ( 30%)
1      352 ( 70%)
```

Status: OK



Weka Clusterer Visualize: 18:35:20 - SimpleKMeans (boston)

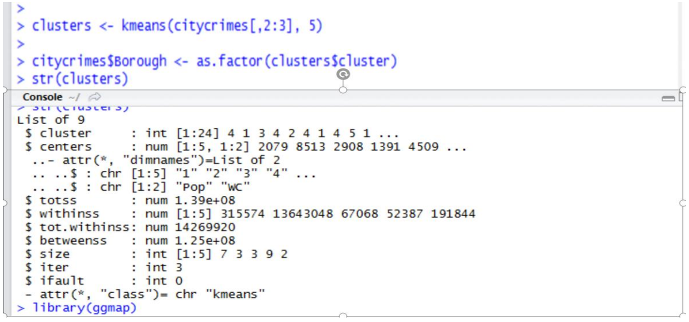
X: Instance_number (Num)

Y: CRIM (Num)

Colour: Cluster (Nom)

Plot: boston_clustered

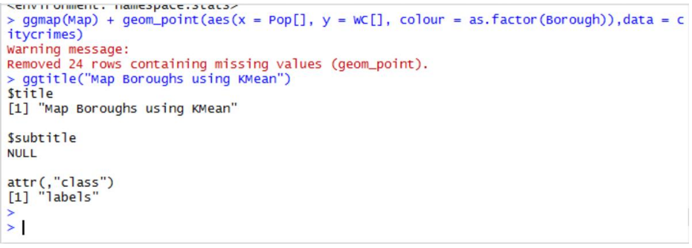
USING R-TOOL



```
>
> clusters <- kmeans(citycrimes[,2:3], 5)
>
> citycrimes$Borough <- as.factor(clusters$cluster)
> str(clusters)
```

Console ~/

```
List of 9
 $ cluster     : int [1:24] 4 1 3 4 2 4 1 4 5 1 ...
 $ centers     : num [1:5, 1:2] 2079 8513 2908 1391 4509 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:5] "1" "2" "3" "4" ...
  .. ..$ : chr [1:2] "Pop" "WC"
 $ totss       : num 1.39e+08
 $ withinss    : num [1:5] 315574 13643048 67068 52387 191844
 $ tot.withinss: num 14269920
 $ betweenss   : num 1.25e+08
 $ size        : int [1:5] 7 3 3 9 2
 $ iter        : int 3
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
> library(ggmap)
```



```
<environment: namespace:stats>
> ggmap(Map) + geom_point(aes(x = Pop[], y = WC[], colour = as.factor(Borough)),data = c
itycrimes)
Warning message:
Removed 24 rows containing missing values (geom_point).
> ggtitle("Map Boroughs using KMean")
$title
[1] "Map Boroughs using KMean"

$subtitle
NULL

attr(,"class")
[1] "labels"
>
> |
```

Map Boroughs using KMean
37.4
37.2
37.0
36.8
lat
-94.8
-94.6
-94.4
-94.2
lon
as.factor(Borough)
● 1
● 2
● 3
● 4
● 5
Mindenmines
Frontenac
Beulah
Kenoma
Pittsburg
Golden City
Nashville
Cherokee
Opolis
Jasper
Weir
Scammon
Asbury
Carytown
Waco
Alba
Avilla
Columbus
Crestline
Carl Junction
Oronogo
Carthage
Webb City
La Russell
Reeds
Joplin
Riverton
Galena
Fidelity
Sarcoxie
Lowell
Leawood
Baxter Springs
Diamond
Wentworth
Picher
Quapaw
Commerce
Miami
Peoria
Granby
Neosho
Stark City
Seneca
Wyandotte
Google
Map data ©2018 Google