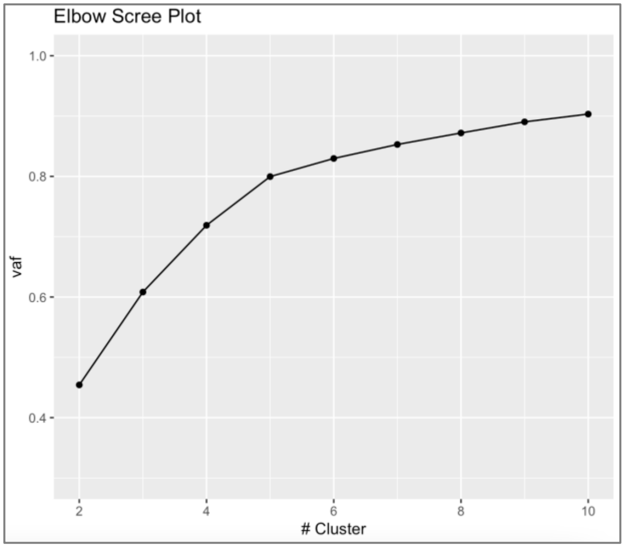
K-Means

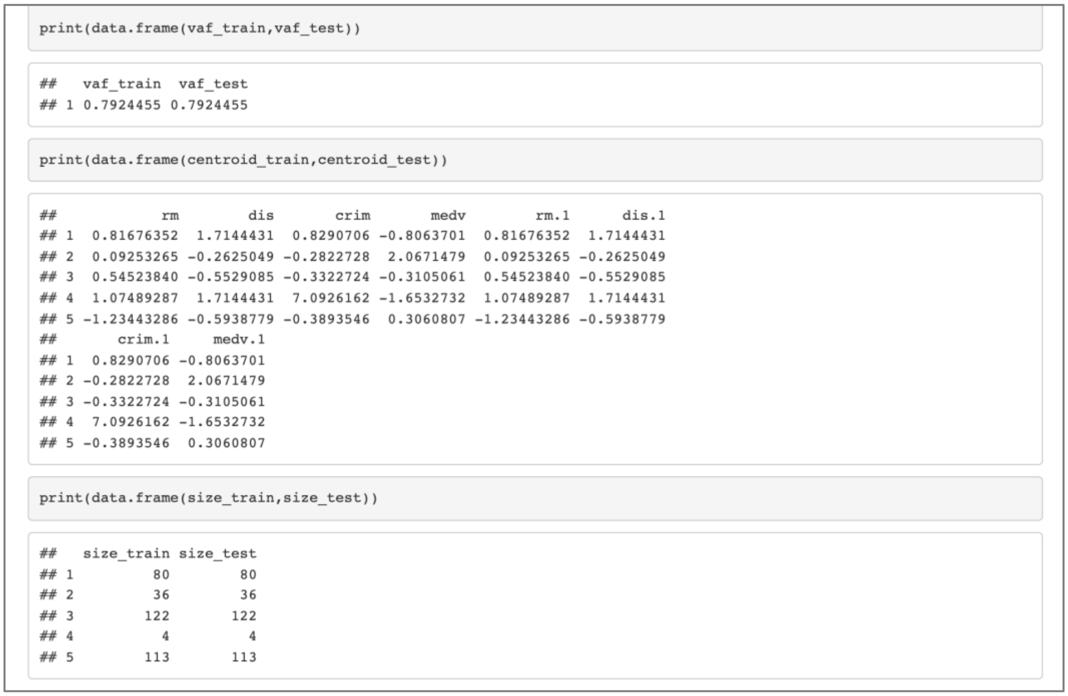
Azizha Zeinita

Perform a Scree test

As we can see from the scree plot (elbow method), the best k to perform k-means is 5



Choose 1 K-means solution (the best K from the last step) to retain from the many solutions that you have generated.



```
print(data.frame(vaf_train,vaf_test))
```

```
##   vaf_train  vaf_test
## 1 0.7924455 0.7924455
```

```
print(data.frame(centroid_train,centroid_test))
```

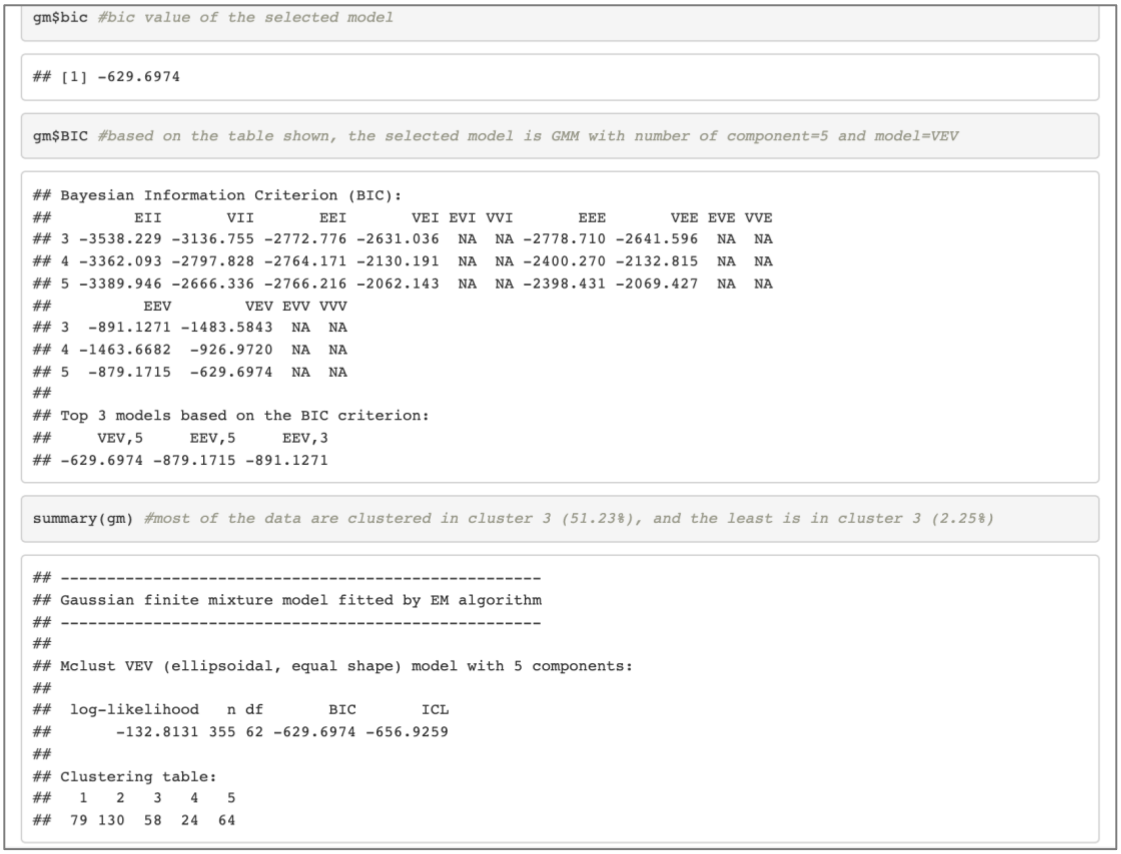
```
##            rm        dis       crim       medv         rm.1      dis.1
## 1  0.81676352  1.7144431  0.8290706 -0.8063701  0.81676352  1.7144431
## 2  0.09253265 -0.2625049 -0.2822728  2.0671479  0.09253265 -0.2625049
## 3  0.54523840 -0.5529085 -0.3322724 -0.3105061  0.54523840 -0.5529085
## 4  1.07489287  1.7144431  7.0926162 -1.6532732  1.07489287  1.7144431
## 5 -1.23443286 -0.5938779 -0.3893546  0.3060807 -1.23443286 -0.5938779
##       crim.1     medv.1
## 1  0.8290706 -0.8063701
## 2 -0.2822728  2.0671479
## 3 -0.3322724 -0.3105061
## 4  7.0926162 -1.6532732
## 5 -0.3893546  0.3060807
```

```
print(data.frame(size_train,size_test))
```

```
##   size_train size_test
## 1         80        80
## 2         36        36
## 3        122       122
## 4          4         4
## 5        113       113
```

VAF, centroid and cluster size comparison between train and test are presenting a decent level of stability. It means that scree test has good selection of K

Generate 3-5 Gaussian Mixtures (GM) | 8. Choose one solution & do interpretation



```
gm$bic #bic value of the selected model
```

```
## [1] -629.6974
```

```
gm$BIC #based on the table shown, the selected model is GMM with number of component=5 and model=VEV
```

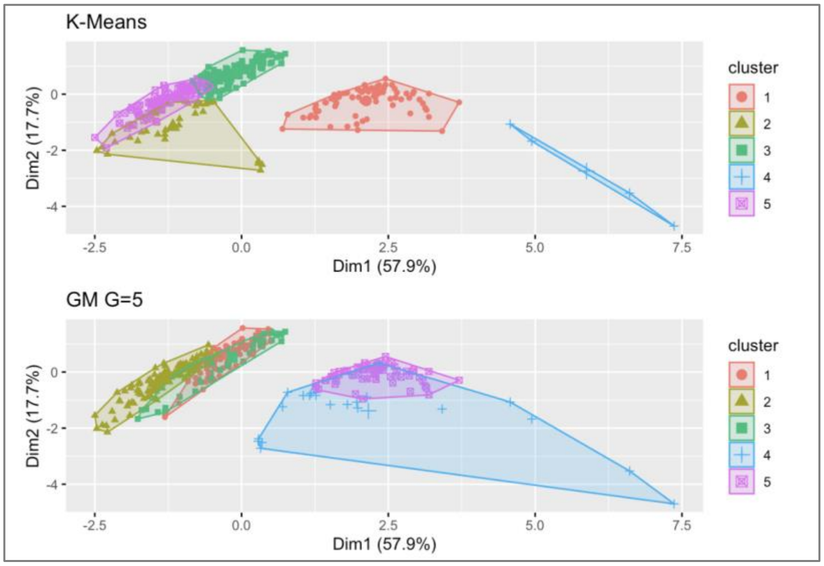
```
## Bayesian Information Criterion (BIC):
##         EII       VII       EEI       VEI EVI VVI       EEE       VEE EVE VVE
## 3 -3538.229 -3136.755 -2772.776 -2631.036  NA  NA -2778.710 -2641.596  NA  NA
## 4 -3362.093 -2797.828 -2764.171 -2130.191  NA  NA -2400.270 -2132.815  NA  NA
## 5 -3389.946 -2666.336 -2766.216 -2062.143  NA  NA -2398.431 -2069.427  NA  NA
##          EEV        VEV EVV VVV
## 3  -891.1271 -1483.5843  NA  NA
## 4 -1463.6682  -926.9720  NA  NA
## 5  -879.1715  -629.6974  NA  NA
##
## Top 3 models based on the BIC criterion:
##     VEV,5     EEV,5     EEV,3
## -629.6974 -879.1715 -891.1271
```

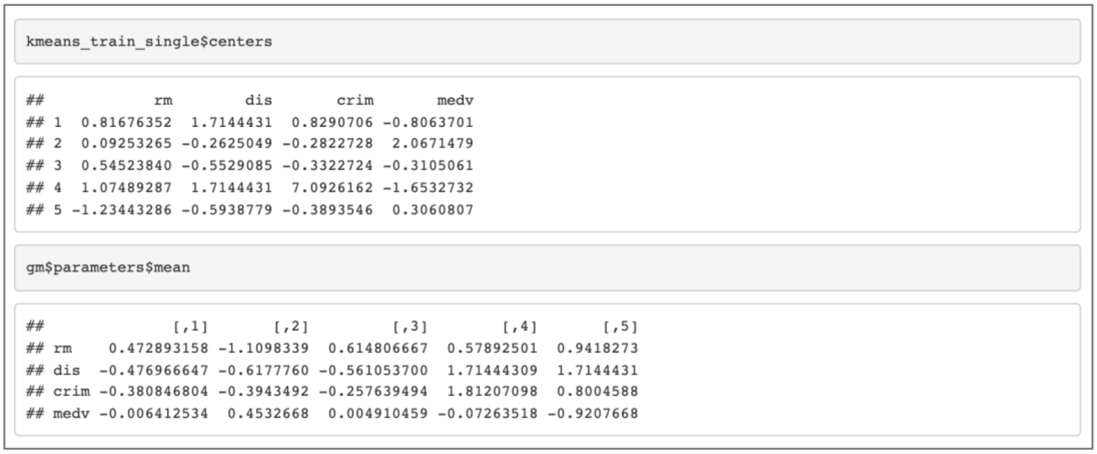
```
summary(gm) #most of the data are clustered in cluster 3 (51.23%), and the least is in cluster 3 (2.25%)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEV (ellipsoidal, equal shape) model with 5 components:
##
##  log-likelihood   n df       BIC       ICL
##       -132.8131 355 62 -629.6974 -656.9259
##
## Clustering table:
##   1   2   3   4   5
##  79 130  58  24  64
```

gm$bic   : bic value of the selected model
gm$BIC   : based on the table shown, the selected model is GMM with number of component=5 and model=VEV
summary(gm) : most of the data are clustered in cluster 2 (36.62%), and the least is in cluster 4 (26.76%)

Build a GM model with the best components on train data and compare it with the train KMeans solution from an interpretability perspective.





```
kmeans_train_single$centers
```

```
##             rm        dis       crim       medv
## 1  0.81676352  1.7144431  0.8290706 -0.8063701
## 2  0.09253265 -0.2625049 -0.2822728  2.0671479
## 3  0.54523840 -0.5529085 -0.3322724 -0.3105061
## 4  1.07489287  1.7144431  7.0926162 -1.6532732
## 5 -1.23443286 -0.5938779 -0.3893546  0.3060807
```

```
gm$parameters$mean
```

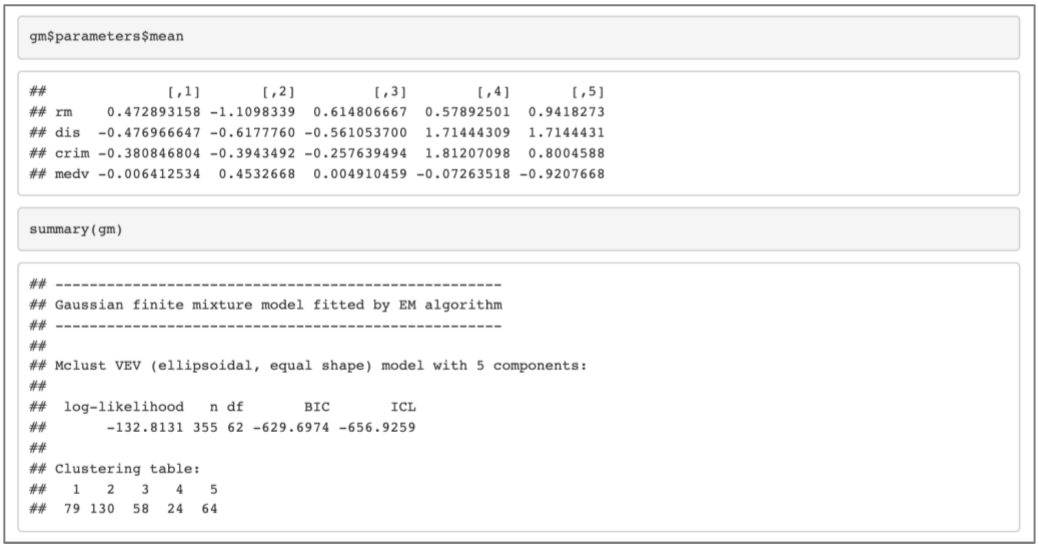
```
##              [,1]       [,2]         [,3]        [,4]       [,5]
## rm    0.472893158 -1.1098339  0.614806667  0.57892501  0.9418273
## dis  -0.476966647 -0.6177760 -0.561053700  1.71444309  1.7144431
## crim -0.380846804 -0.3943492 -0.257639494  1.81207098  0.8004588
## medv -0.006412534  0.4532668  0.004910459 -0.07263518 -0.9207668
```

Based on the visualization, data distribution is not circular, ideally GM clustering is better than K-Means clustering. As we can see, cluster 2,3 and 5 of K-Means clustering is tend to be forced to fit. In GM clustering, cluster 1,2, and 3 are also overlap as well as cluster 4 and 5, however GM is clustered buy normal distribution, not only by mean. The highest mean in K-Means is far above di average, besides, relation between variables in GM more make sense. In conclusion, GM model is chosen to be the best model.

K-Means

Azizha Zeinita

Summarize results and interpret the clusters/segments you choose as your final solution.



```
gm$parameters$mean
```

```
##                [,1]       [,2]        [,3]        [,4]       [,5]
## rm    0.472893158 -1.1098339  0.614806667  0.57892501  0.9418273
## dis  -0.476966647 -0.6177760 -0.561053700  1.71444309  1.7144431
## crim -0.380846804 -0.3943492 -0.257639494  1.81207098  0.8004588
## medv -0.006412534  0.4532668  0.004910459 -0.07263518 -0.9207668
```

```
summary(gm)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEV (ellipsoidal, equal shape) model with 5 components:
##
##  log-likelihood   n df       BIC       ICL
##       -132.8131 355 62 -629.6974 -656.9259
##
## Clustering table:
##   1   2   3   4   5
##  79 130  58  24  64
```

Cluster 1 : It has the third lowest median-price which the house area has 2nd lowest per capita crime rate, 3rd closest distance to employment center, but the also the 2nd lowest number of room.

Cluster 2 : Its the highest median-price with the lowest crime rate and closest distance to employment center, but it has the lowest number of room.

Cluster 3 : This cluster has the 2nd highest median-price and related to the 3rd lowest crime rate and 2nd farthest distance to employment center, but it has the 2nd highest number of room.

Cluster 4 : It has the 2nd lowest median-price that related to the highest criminal rate and the farthest distance to employee center, but it has the 3rd highest number of room.

Cluster 5 : The lowest median-price that related to the 2nd highest crime rate, the farest distance to emploment center (same as Cluster 4), and the highest number of room

Most of datapoints are clustered to cluster 2, which the increasement of median price is related the reduction of crime rate and distance to employment center, but having the lowest number of room compared to the other clusters. Characteristic of this cluster tends to consider crime rate and distance to employment center as the more essential considerations compared to the number of room in the house.