Similarity Clustering

Joshua Durana

2022-10-09

Source: https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset (https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset) This data set contains data collected by IOT devices to detect smoke

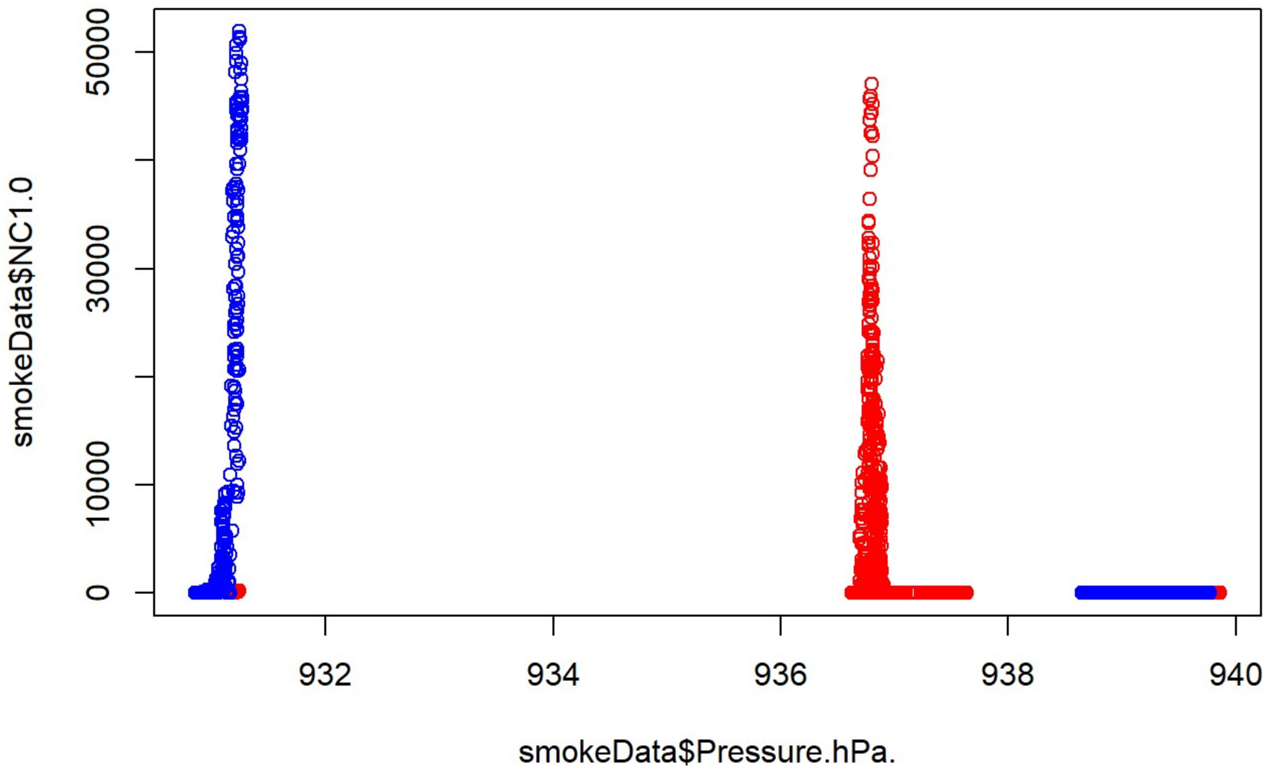
Load and Clean Data

```
smokeData <- read.csv("Data/smokeData.csv", header = TRUE)

#Remove unnecessary columns
smokeData <- subset(smokeData, select = -c(X,UTC,CNT))

#Convert fire alarm into factors
smokeData$Fire.Alarm <- as.factor(smokeData$Fire.Alarm)
```

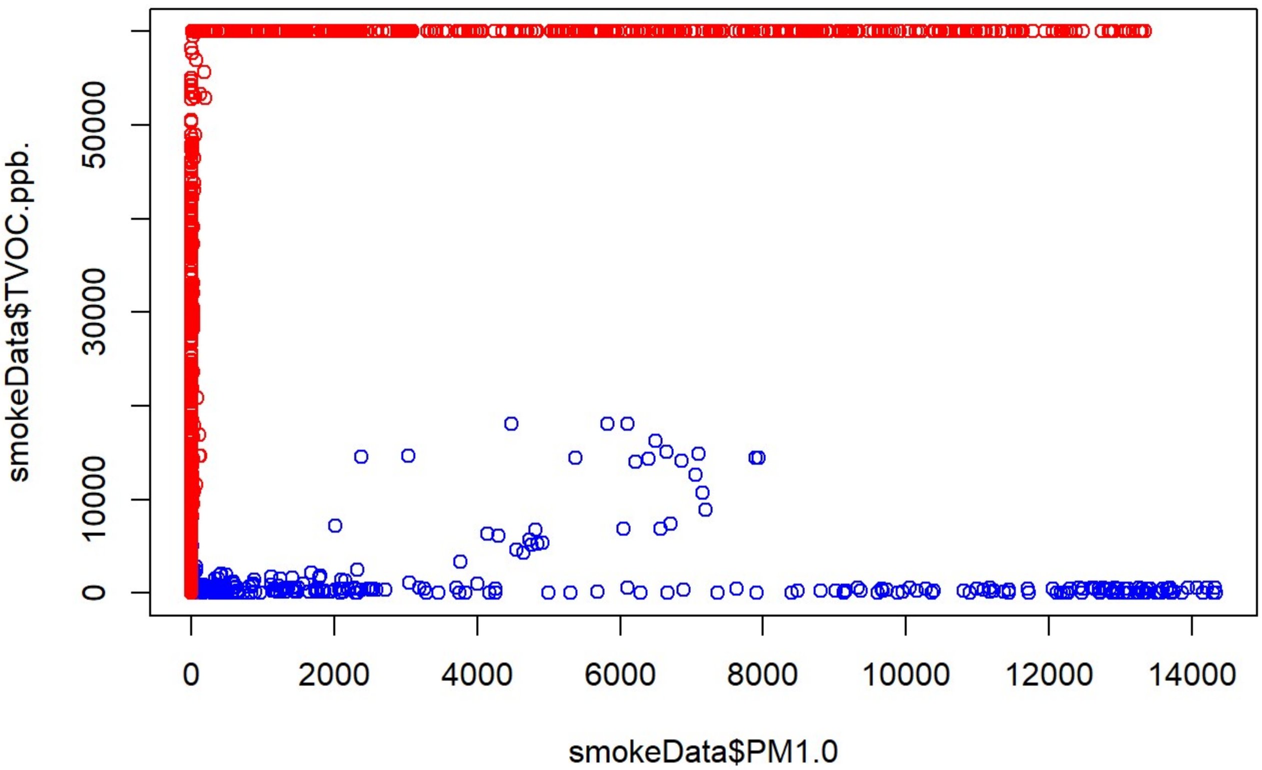
Data Exploration

We're trying to see whether we can use clustering to find clusters for Fire.Alarm

```
plot(smokeData$Pressure.hPa., smokeData$NC1.0, col = c("red", "blue")[unclass(smokeData$Fire.
Alarm)])
```




```
plot(smokeData$PM1.0, smokeData$TVOC.ppb., col = c("red", "blue")[unclass(smokeData$Fire.Alar
m)])
```



Pairs with pressure has good separation between the different factors, but usually have more than 3 clusters. The other plot has good separation with 2 clusters.

```
summary(smokeData)
```

```
##  Temperature.C.    Humidity...      TVOC.ppb.       eCO2.ppm.
##  Min.   :-22.01   Min.   :10.74   Min.   :    0   Min.   :  400
##  1st Qu.: 10.99   1st Qu.:47.53   1st Qu.:  130   1st Qu.:  400
##  Median : 20.13   Median :50.15   Median :  981   Median :  400
##  Mean   : 15.97   Mean   :48.54   Mean   : 1942   Mean   :  670
##  3rd Qu.: 25.41   3rd Qu.:53.24   3rd Qu.: 1189   3rd Qu.:  438
##  Max.   : 59.93   Max.   :75.20   Max.   :60000   Max.   :60000
##      Raw.H2       Raw.Ethanol    Pressure.hPa.       PM1.0
##  Min.   :10668   Min.   :15317   Min.   :930.9   Min.   :    0.00
##  1st Qu.:12830   1st Qu.:19435   1st Qu.:938.7   1st Qu.:    1.28
##  Median :12924   Median :19501   Median :938.8   Median :    1.81
##  Mean   :12942   Mean   :19754   Mean   :938.6   Mean   :  100.59
##  3rd Qu.:13109   3rd Qu.:20078   3rd Qu.:939.4   3rd Qu.:    2.09
##  Max.   :13803   Max.   :21410   Max.   :939.9   Max.   :14333.69
##      PM2.5              NC0.5              NC1.0              NC2.5
##  Min.   :    0.00   Min.   :    0.00   Min.   :    0.00   Min.   :    0.000
##  1st Qu.:    1.34   1st Qu.:    8.82   1st Qu.:    1.38   1st Qu.:    0.033
##  Median :    1.88   Median :   12.45   Median :    1.94   Median :    0.044
##  Mean   :  184.47   Mean   :  491.46   Mean   :  203.59   Mean   :   80.049
##  3rd Qu.:    2.18   3rd Qu.:   14.42   3rd Qu.:    2.25   3rd Qu.:    0.051
##  Max.   :45432.26   Max.   :61482.03   Max.   :51914.68   Max.   :30026.438
##  Fire.Alarm
##  0:17873
##  1:44757
##
##
##
##
```

K-Means Clustering

smokeKCluster

```
#Scale Pressure and NC1.0
scaledCol <- sapply(smokeData[c(7, 11)], function(x) c(scale(x)))

#Kmeans
smokeKCluster <- kmeans(scaledCol, 2, nstart = 50)
smokeKCluster$withinss
```

```
## [1] 51302.29 23365.84
```

```
table(smokeData$Fire.Alarm, smokeKCluster$cluster)
```

```
##
##         1     2
##   0   493 17380
##   1  1121 43636
```

The within sum of squares seems pretty large and the table seems to show that the clusters and the fire alarm factor doesn't really correlate. This is most likely due to this pair's plot having 3 clusters of fire alarm values. It's shown when I use 3 clusters for the same pair

```
#Scale Pressure and NC1.0
scaledCol <- sapply(smokeData[c(7, 11)], function(x) c(scale(x)))

#Kmeans
smokeKCluster <- kmeans(scaledCol, 3, nstart = 50)
smokeKCluster$withinss
```

```
## [1]  9388.537  4505.644 26434.129
```

```
table(smokeData$Fire.Alarm, smokeKCluster$cluster)
```

```
##
##         1     2     3
##   0   270  6356 11247
##   1   101 43632  1024
```

```
#Scale TVOC and PM1.0
scaledCol <- sapply(smokeData[c(3, 8)], function(x) c(scale(x)))

#Kmeans
smokeKCluster <- kmeans(scaledCol, 2, nstart = 50)
smokeKCluster$withinss
```

```
## [1]  8809.696 30657.927
```

```
table(smokeData$Fire.Alarm, smokeKCluster$cluster)
```

```
##
##         1     2
##   0 16885   988
##   1 44641   116
```

The within sum of squares are big, but compared with the other cluster is much more smaller. The table also shows an improvement between correlation the clusters and fire alarm. This improvement is likely due to this pair's plot being separated by 2 main groups.

Hierarchical Clustering

```
library(flexclust)
```

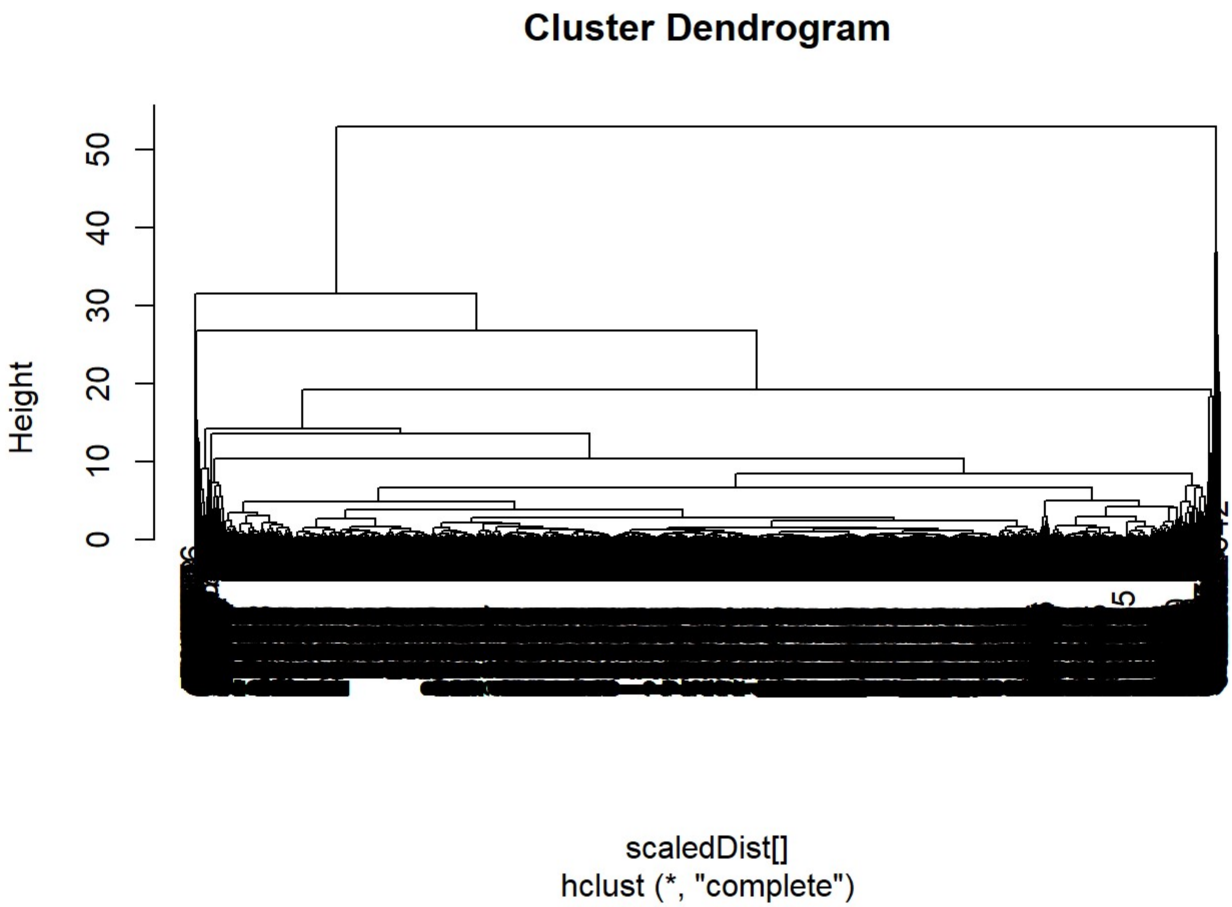
```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
#Subset Data
set.seed(10622)

i <- sample(1:nrow(smokeData), .2*nrow(smokeData), replace = FALSE)
smokeSubset <- smokeData[i,]

#Scale Data
scaledDist <- dist(scale(smokeSubset[1:12]))

#Hierarchical Clustering
smokeHClust <- hclust(scaledDist[])

plot(smokeHClust)
```




```
for (cut in 2:40)
{
  smokeCut <- cutree(smokeHClust, cut)
  smokeTable <- table(smokeCut, smokeSubset$Fire.Alarm)
  smokeRI <- randIndex(smokeTable)
  print(paste(cut,"RI: ", smokeRI))
}
```

```
## [1] "2 RI:  0.00891496092765294"
## [1] "3 RI:  0.00892401632882631"
## [1] "4 RI:  0.00863646054648794"
## [1] "5 RI:  0.01954936794346"
## [1] "6 RI:  0.0195611879592891"
## [1] "7 RI:  0.0195533154045274"
## [1] "8 RI:  0.0195551172316176"
## [1] "9 RI:  0.0279465511033111"
## [1] "10 RI:  0.0279423000772641"
## [1] "11 RI:  0.0279977850383266"
## [1] "12 RI:  0.027991348035457"
## [1] "13 RI:  0.0279907407464338"
## [1] "14 RI:  0.0594533140291848"
## [1] "15 RI:  0.0594531625532121"
## [1] "16 RI:  0.059033827397995"
## [1] "17 RI:  0.0590266317630248"
## [1] "18 RI:  0.0590262279700544"
## [1] "19 RI:  0.0590286284622342"
## [1] "20 RI:  0.0590284770398663"
## [1] "21 RI:  0.0590250700391533"
## [1] "22 RI:  0.05902360629221"
## [1] "23 RI:  0.0590232277371135"
## [1] "24 RI:  0.0590094988466643"
## [1] "25 RI:  0.0590083379515025"
## [1] "26 RI:  0.0445550411203805"
## [1] "27 RI:  0.044552054033946"
## [1] "28 RI:  0.044551755325511"
## [1] "29 RI:  0.044550361353317"
## [1] "30 RI:  0.0444411860255761"
## [1] "31 RI:  0.0444410615695896"
## [1] "32 RI:  0.0444393191864699"
## [1] "33 RI:  0.0444390951658768"
## [1] "34 RI:  0.122289343396101"
## [1] "35 RI:  0.122289244306837"
## [1] "36 RI:  0.122289194762207"
## [1] "37 RI:  0.122289145217578"
## [1] "38 RI:  0.12228909567295"
## [1] "39 RI:  0.122273043267543"
## [1] "40 RI:  0.122272746001792"
```

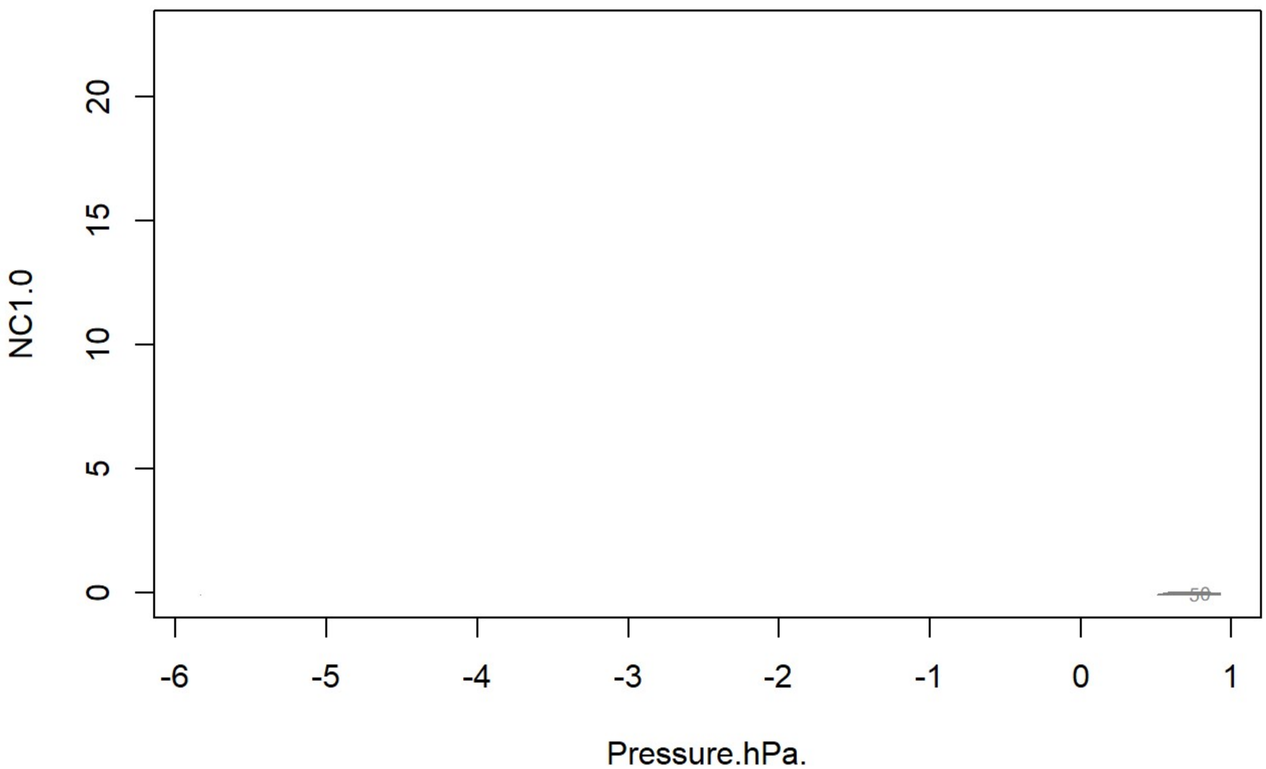
The best clustering results seems to be starting at cut 34. This is most likely due to the clustering over fitting at the lower heights. This clustering method seems much better with lower dimension data sets to reduce the risk of over fitting.

Model Clustering

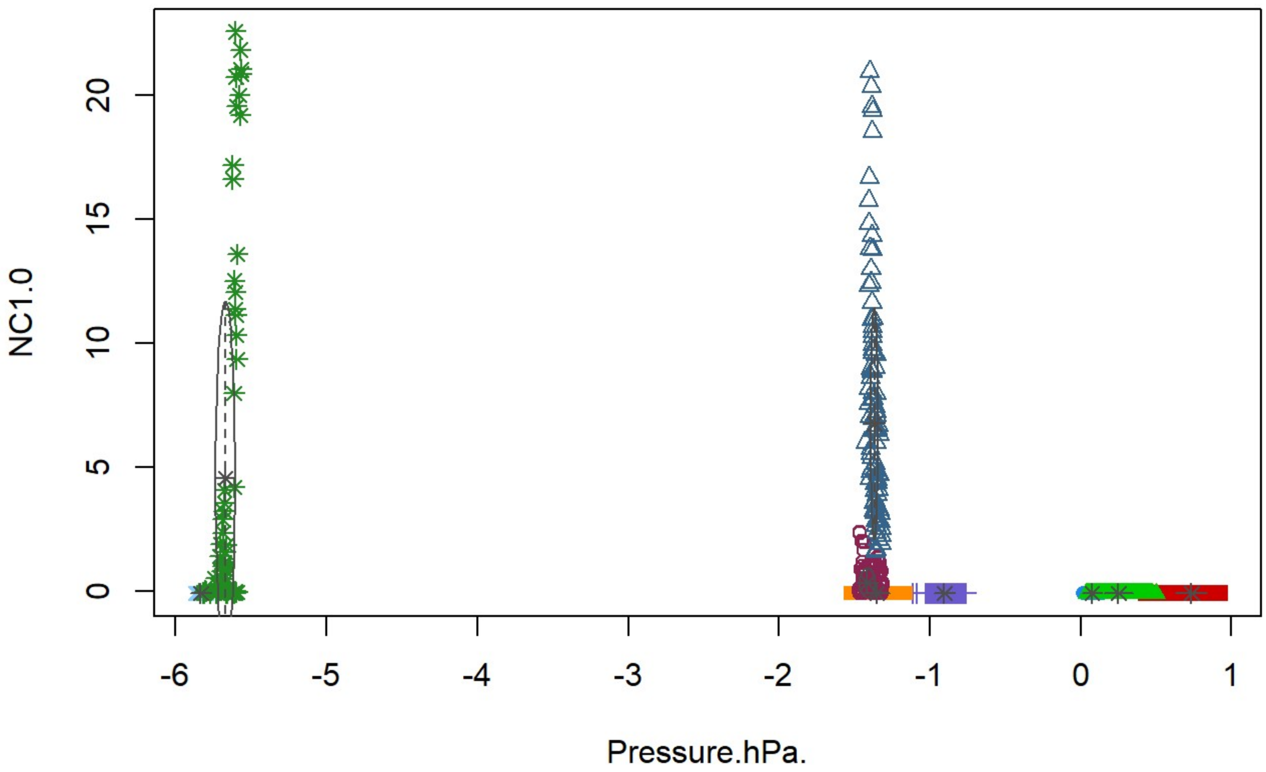
```
library(mclust)
```

```
## Package 'mclust' version 5.4.10
## Type 'citation("mclust")' for citing this R package in publications.
```

```
scaledCol <- sapply(smokeSubset[c(7,11)], function(x) c(scale(x)))

smokeDens <- densityMclust(scaledCol)
plot(smokeDens, what="density")
```




```
smokeModel <- Mclust(scaledCol)
plot(smokeModel, what = "classification")
```




```
summary(smokeModel)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVI (diagonal, varying volume and shape) model with 9 components:
##
##  log-likelihood     n df      BIC    ICL
##         84371.2 12526 44 168327.2 167397
##
## Clustering table:
##    1    2    3    4    5    6    7    8    9
## 3651 3728 2651 1524  513  156  132   71  100
```

This model didn't really cluster the points by whether it triggered a fire alarm. The model seems to be very sensitive, because even close data points are separated in different clusters.

The best clustering model at predicting seems to be K-Means due to being able to specify how many clusters you want. The other 2 models overfit the data. The hierarchy clustering was correct on having 2 clusters, while the model clustering had 9 clusters. But, clustering isn't really used for prediction, it's mainly used to gather insights about the data.