In this document we will try to find best configuration for k-means clustering and compare its performance with EM (Expectation Maximization) clustering using Weka.

Choice of K:

Let's try different values of K.

K=2:

```
=== Run information ===

Scheme:       weka.clusterers.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 firs
Relation:     occupancy_detection_training_dataset
Instances:    8143
Attributes:   6
              Temperature
              Humidity
              Light
              CO2
              HumidityRatio
Ignored:
              Occupancy
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 16
Within cluster sum of squared errors: 1188.274502028947

Initial starting points (random):

Cluster 0: 22.03,26.34,469,1031,0.004312
Cluster 1: 19.5,27.1,0,456,0.003795

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute       Full Data         0         1
                 (8143.0)  (1517.0)  (6626.0)
=============================================
Temperature       20.6195   21.6826   20.3761
Humidity          25.7321   31.3724   24.4408
Light            119.5194  329.1936   71.5152
CO2              606.5462 1123.0857  488.2863
HumidityRatio      0.0039     0.005    0.0036




Time taken to build model (full training data) : 0.14 seconds

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

Clustered Instances

0       1517 ( 19%)
1       6626 ( 81%)
```

K=3:

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1
Relation:     occupancy_detection_training_dataset
Instances:    8143
Attributes:   6
              Temperature
              Humidity
              Light
              CO2
              HumidityRatio
Ignored:
              Occupancy
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 17
Within cluster sum of squared errors: 865.6600062432577

Initial starting points (random):

Cluster 0: 22.03,26.34,469,1031,0.004312
Cluster 1: 19.5,27.1,0,456,0.003795
Cluster 2: 20.6,31.45,438,1050.5,0.00472

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute       Full Data          0          1          2
                 (8143.0)   (2118.0)   (5197.0)    (828.0)
=========================================================
Temperature       20.6195    21.8615    19.9963    21.3542
Humidity          25.7321     22.348    25.6772    34.7332
Light            119.5194    300.235    14.3161   317.5694
CO2              606.5462   714.3305   450.0043 1313.3835
HumidityRatio      0.0039     0.0036     0.0037     0.0055



Time taken to build model (full training data) : 0.06 seconds

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

Clustered Instances

0      2118 ( 26%)
1      5197 ( 64%)
2       828 ( 10%)
```

K=4:

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1
Relation:     occupancy_detection_training_dataset
Instances:    8143
Attributes:   6
              Temperature
              Humidity
              Light
              CO2
              HumidityRatio
Ignored:
              Occupancy
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 18
Within cluster sum of squared errors: 489.2325189340183

Initial starting points (random):

Cluster 0: 22.03,26.34,469,1031,0.004312
Cluster 1: 19.5,27.1,0,456,0.003795
Cluster 2: 20.6,31.45,438,1050.5,0.00472
Cluster 3: 21.1,24.92,0,442.25,0.003852

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute       Full Data         0         1         2         3
                 (8143.0)  (1936.0)  (3042.0)   (659.0)  (2506.0)
============================================================================
Temperature       20.6195   21.9581   19.7311   21.4412   20.4477
Humidity          25.7321   23.4223   29.6181   35.3866   20.2605
Light            119.5194  310.5329   25.5352  338.8126   28.3714
CO2              606.5462  736.6508  465.4083 1437.2617  458.9076
HumidityRatio      0.0039    0.0038    0.0042    0.0056     0.003


Time taken to build model (full training data) : 0.06 seconds

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

Clustered Instances

0      1936 ( 24%)
1      3042 ( 37%)
2       659 (  8%)
3      2506 ( 31%)
```

K=5:

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1
Relation:     occupancy_detection_training_dataset
Instances:    8143
Attributes:   6
              Temperature
              Humidity
              Light
              CO2
              HumidityRatio
Ignored:
              Occupancy
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 405.7899303636782

Initial starting points (random):

Cluster 0: 22.03,26.34,469,1031,0.004312
Cluster 1: 19.5,27.1,0,456,0.003795
Cluster 2: 20.6,31.45,438,1050.5,0.00472
Cluster 3: 21.1,24.92,0,442.25,0.003852
Cluster 4: 22.92,16.89,193.5,441.5,0.002913

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute       Full Data          0          1          2          3          4
                 (8143.0)    (927.0)   (3014.0)    (648.0)   (2560.0)    (994.0)
=============================================================================
Temperature       20.6195     22.072    19.7277    21.4552    20.4455    21.8723
Humidity          25.7321    26.4234    29.6536    35.4341     20.731    19.7519
Light            119.5194   296.5377     24.591   338.4653    12.1946   375.9506
CO2              606.5462    838.968   464.3329  1446.1076   454.0272   666.4949
HumidityRatio      0.0039     0.0043     0.0042     0.0056     0.0031     0.0032



Time taken to build model (full training data) : 0.03 seconds

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

Clustered Instances

0       927 ( 11%)
1      3014 ( 37%)
2       648 (  8%)
3      2560 ( 31%)
4       994 ( 12%)
```

K=6:

```
Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1
Relation:     occupancy_detection_training_dataset
Instances:    8143
Attributes:   6
              Temperature
              Humidity
              Light
              CO2
              HumidityRatio
Ignored:
              Occupancy
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 326.3284756965421

Initial starting points (random):

Cluster 0: 22.03,26.34,469,1031,0.004312
Cluster 1: 19.5,27.1,0,456,0.003795
Cluster 2: 20.6,31.45,438,1050.5,0.00472
Cluster 3: 21.1,24.92,0,442.25,0.003852
Cluster 4: 22.92,16.89,193.5,441.5,0.002913
Cluster 5: 19.45,27,0,470,0.003767

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute       Full Data          0          1          2          3          4          5
                 (8143.0)   (641.0)  (2873.0)   (653.0)  (1066.0)  (1001.0)  (1909.0)
=============================================================================================
Temperature       20.6195   22.3157   19.6911   21.4487   21.0182    21.878    20.281
Humidity          25.7321   26.5385   29.8187   35.4023   24.9441   19.7334   19.5888
Light            119.5194   401.451   21.7813   339.255   30.7621  373.1897   13.3315
CO2              606.5462  956.1158  468.1223 1442.8043  476.3223  668.9981  451.4104
HumidityRatio      0.0039    0.0044    0.0042    0.0056    0.0038    0.0032    0.0029


Time taken to build model (full training data) : 0.03 seconds

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

Clustered Instances

0       641 (  8%)
1      2873 ( 35%)
2       653 (  8%)
3      1066 ( 13%)
4      1001 ( 12%)
5      1909 ( 23%)
```

According to Total within sum of squares (TWSS) rule or elbow rule. It looks like optimal value of K will be either 3 or 4. Because, at these points the TWSS stops dropping fast. But, according to domain knowledge and nature of the problem we are trying to solve there should be only two natural clusters one for Yes and one for No. To get accurate measure Complete Clustering needs to be performed i.e. every data point must belong to some cluster and every cluster should be exclusive i.e. each data point should belong to one cluster only as person can either be present or not present. Due to these reasons I will choose value of K = 2 that is against the elbow rule but, according to domain knowledge and knowing the fact that Weka ignores additional clusters (exceeding number of class labels), I will choose K= 2.

Clusters description:

Following are the values of initial centroids and final centroids for K = 2.

```
Initial starting points (random):

Cluster 0: 22.03,26.34,469,1031,0.004312
Cluster 1: 19.5,27.1,0,456,0.003795

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute       Full Data          0          1
                 (8143.0)   (1517.0)   (6626.0)
=============================================
Temperature       20.6195    21.6826    20.3761
Humidity          25.7321    31.3724    24.4408
Light            119.5194   329.1936    71.5152
CO2              606.5462 1123.0857   488.2863
HumidityRatio      0.0039     0.005     0.0036
```

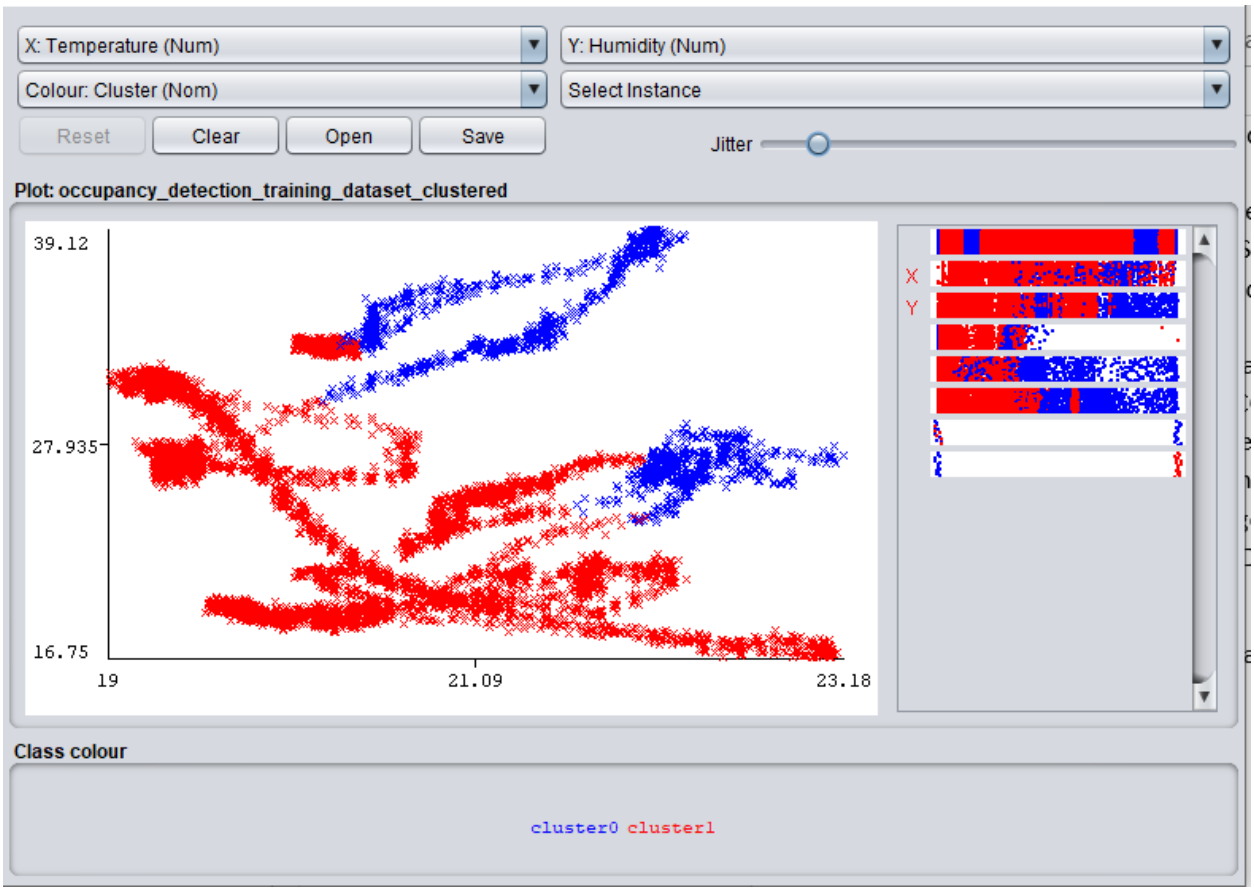
Cluster 0 is for occupancy = Yes and Cluster 1 is for occupancy = No.

The final values of centroids are sensible according to general knowledge of Science. Mean of Temperature for cluster 0 is higher than cluster 1, this indicates if the person is in the room then temperature of room increases, this corresponds to science rules as human body emits heat resulting in increased temperatures. Similar trend for the humidity as well, as activities like cooking, washing etc. and people's breath provides the primary source of moisture that cause humidity indoors. If the lights are switched on or if they are brighter then there is higher probability that someone is inside the room and that is why average value of Light for cluster 0 is higher. Human body exhales CO2 resulting in increased percentage of CO2 in the room where human is present, so Cluster 0 has higher average of CO2. Humidity ratio is the ratio of weight of moisture to the weight of dry air in the air. As discussed before, presence of the person increases the temperature and warm air contains more moisture resulting in low percentage of dry air and presence of person also increases the humidity which cause

the humidity ratio to increase. Due to these reasons Humidity ratio average of Cluster 0 is higher than Cluster 1.

So, average values of the attributes of the chosen cluster centers are justified.

Cluster Visualization:



For cluster visualization I have chosen two attributes (Temperature and Humidity). These two attributes are extremely correlated and they display clear distinction between two clusters. As it can be viewed from above that there is clear separation between cluster 0 (Yes occupancy) and cluster 1 (No occupancy) where Humidity increases with the increase in Temperature, which makes this observation interesting. For those cases where Temperature is high and Humidity is close to average then the clusters are not clearly separated from each other. For cases where Temperatures is high and Humidity is low then cluster 1 gets easily separated from cluster 0. This brings to the conclusion that Humidity plays a very important role in deciding occupancy where high values indicates the presence of some person and low values indicates that no one is not present.

Clusters evaluation with the help of class labels:

Let’s compare the accuracy of clusters with different values of K. Value of K will be 2, 3, 4. 2 was my choice and 3, 4 was the choice suggested by elbow rule.

K=2:

```
Within cluster sum of squared errors: 1188.274502028947

Initial starting points (random):

Cluster 0: 22.03,26.34,469,1031,0.004312
Cluster 1: 19.5,27.1,0,456,0.003795

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute       Full Data         0         1
                 (8143.0)  (1517.0)  (6626.0)
=============================================
Temperature       20.6195   21.6826   20.3761
Humidity          25.7321   31.3724   24.4408
Light            119.5194  329.1936   71.5152
CO2              606.5462 1123.0857  488.2863
HumidityRatio      0.0039     0.005    0.0036



Time taken to build model (full training data) : 0.03 seconds

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

Clustered Instances

0      1517 ( 19%)
1      6626 ( 81%)


Class attribute: Occupancy
Classes to Clusters:

    0    1  <-- assigned to cluster
 1019  710 | Y
  498 5916 | N

Cluster 0 <-- Y
Cluster 1 <-- N

Incorrectly clustered instances :       1208.0   14.8348 %
```

K=3:

```
Initial starting points (random):

Cluster 0: 22.03,26.34,469,1031,0.004312
Cluster 1: 19.5,27.1,0,456,0.003795
Cluster 2: 20.6,31.45,438,1050.5,0.00472

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute       Full Data          0          1          2
                 (8143.0)   (2118.0)   (5197.0)    (828.0)
=========================================================
Temperature       20.6195    21.8615    19.9963    21.3542
Humidity          25.7321     22.348    25.6772    34.7332
Light            119.5194    300.235    14.3161   317.5694
CO2              606.5462   714.3305   450.0043 1313.3835
HumidityRatio      0.0039     0.0036     0.0037     0.0055



Time taken to build model (full training data) : 0.03 seconds

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

Clustered Instances

0      2118 ( 26%)
1      5197 ( 64%)
2       828 ( 10%)


Class attribute: Occupancy
Classes to Clusters:

    0    1    2  <-- assigned to cluster
 1086   65  578 | Y
 1032 5132  250 | N

Cluster 0 <-- Y
Cluster 1 <-- N
Cluster 2 <-- No class

Incorrectly clustered instances :       1925.0   23.6399 %
```

K=4:

```
Initial starting points (random):

Cluster 0: 22.03,26.34,469,1031,0.004312
Cluster 1: 19.5,27.1,0,456,0.003795
Cluster 2: 20.6,31.45,438,1050.5,0.00472
Cluster 3: 21.1,24.92,0,442.25,0.003852

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute       Full Data          0          1          2          3
                 (8143.0)   (1936.0)   (3042.0)    (659.0)   (2506.0)
=============================================================================
Temperature       20.6195    21.9581    19.7311    21.4412    20.4477
Humidity          25.7321    23.4223    29.6181    35.3866    20.2605
Light            119.5194   310.5329    25.5352   338.8126    28.3714
CO2              606.5462   736.6508   465.4083  1437.2617   458.9076
HumidityRatio      0.0039     0.0038     0.0042     0.0056     0.003



Time taken to build model (full training data) : 0.03 seconds

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

Clustered Instances

0      1936 ( 24%)
1      3042 ( 37%)
2       659 (  8%)
3      2506 ( 31%)


Class attribute: Occupancy
Classes to Clusters:

    0    1    2    3  <-- assigned to cluster
 1045   95  493   96 | Y
  891 2947  166 2410 | N

Cluster 0 <-- Y
Cluster 1 <-- N
Cluster 2 <-- No class
Cluster 3 <-- No class


Incorrectly clustered instances :       4151.0   50.9763 %
```

As we increase the value of K the accuracy goes down. Which justifies my choice of K, accuracy is best at K = 2. Poorer accuracy for large values of K could be due to the fact that WEKA ignores additional clusters (exceeding number of class labels). We can conclude that there are only two natural clusters in our dataset one for No occupancy and One for Yes occupancy. Choice of clusters more than 2 does not make sense especially according to the implementation of algorithm in WEKA.

EM Clustering:

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     occupancy_detection_training_dataset-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0
Instances:    8143
Attributes:   7
              Temperature
              Humidity
              Light
              CO2
              HumidityRatio
Ignored:
              Occupancy
              cluster
Test mode:    Classes to clusters evaluation on training data

=== Clustering model (full training set) ===


EM
==

Number of clusters selected by cross validation: 3
Number of iterations performed: 12


                Cluster
Attribute              0        1        2
                  (0.54)   (0.08)   (0.38)
============================================
Temperature
  mean            20.0563   21.541  21.2359
  std. dev.        0.6036   1.0829   0.9761

Humidity
  mean            25.2251  18.4867  28.0072
  std. dev.        4.8792   1.0042   5.4728

Light
  mean                  0 163.1983 282.6882
  std. dev.        0.0003 141.8593   220.91
```

```
CO2
  mean           448.4962 459.4068 865.9671
  std. dev.       15.2802  26.3819 392.3509

HumidityRatio
  mean             0.0037   0.0029   0.0044
  std. dev.        0.0007   0.0001   0.0009


Time taken to build model (full training data) : 17.51 seconds

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

Clustered Instances

0      4579 ( 56%)
1       570 (  7%)
2      2994 ( 37%)


Log likelihood: -3.49075


Class attribute: Occupancy
Classes to Clusters:

    0    1    2  <-- assigned to cluster
    0   27 1702 | Y
 4579  543 1292 | N

Cluster 0 <-- N
Cluster 1 <-- No class
Cluster 2 <-- Y

Incorrectly clustered instances :       1862.0   22.8663 %
```

Cluster 0 is for No Occupancy and Cluster 2 is for Yes Occupancy while Cluster 1 is No Class.

Value of K:

EM clustering algorithm chose the value of K = 3 which is different from my previous choice of K = 2. So, it disagrees with my choice of K. But, in terms of accuracy measure choice of K = 2 proves to be right as the percentage of Incorrectly Clustered Instances is greater in case of EM Clustering with K = 3.

Comparison with K-means:

EM clustering performs better to identify groups that are overlapping or if the attributes are independent. In our problem and dataset there is no scenario of overlapping groups. At any given time either someone will be in the room/building or no one will be there. That is the reason why EM clustering prediction accuracy is lower as compared to K-means (K=2). K-means (K=2) makes choices based on distance measures while EM Clustering makes decision according to probabilities and likelihood that's why K-means (K=2) performs better in our case (non-overlapping and dependent attributes). Also, EM Clustering chose three clusters and there are only two natural groups in the dataset, WEKA ignores the additional clusters ultimately bringing down the accuracy. Overall EM Clustering gives the general overview of the clusters but, in our case we need concrete and well defined

explanation of the clusters which is given by clusters formed by K-Means(K=2). These reason makes K-means (K=2) more suitable for our case.

Clusters description:

Clusters can be described intuitively in terms of standard deviation calculated by EM Clustering. Standard deviation of the attributes for (No Occupancy) Cluster are lower as compared to Standard deviation of attributes for (Yes Occupancy). There is obvious reason for this behavior. For the cases when there is no occupancy i.e. no one is present in the room or building the environment variables stays more or less same and that is why the cluster formed for No occupancy is much more tight and close. While on the other hand the environment variables changes significantly according to number of persons in the room/building. Like increase in Temperature of the room/building is directly proportional to number of people in it. Similar behavior is for other environment variables like CO2, humidity etc. Due to this reason the cluster for Yes occupancy is much more widely spread or sparse as compared to that of Cluster for No occupancy.

Explanation for both K-means (K=2) and EM Clustering are justifiable and sensible according to laws of science. But explanation of clusters, using mean in case of K-means (K=2) was much more intuitive and simpler and presence of no class cluster (cluster 1) formed by EM Cluster prohibits us to give final verdict. So, explanation of clusters formed by K-means (K=2) is better.