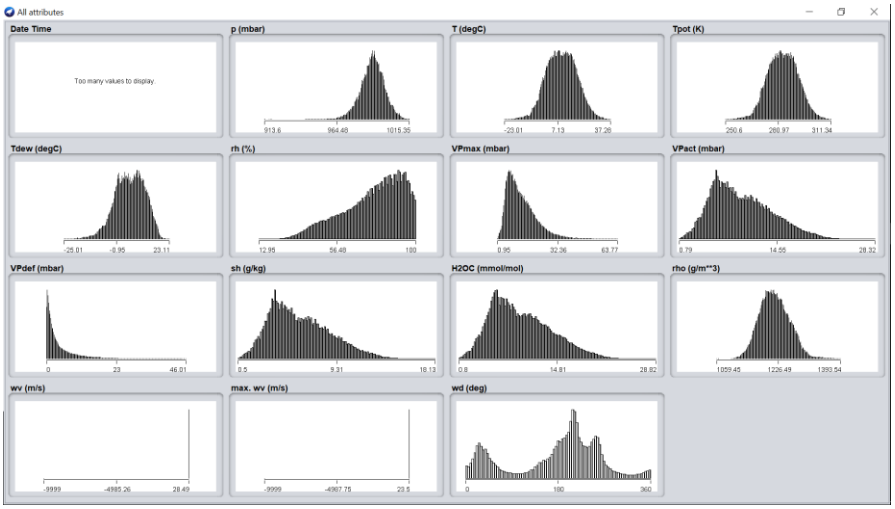
Jeeno Doria

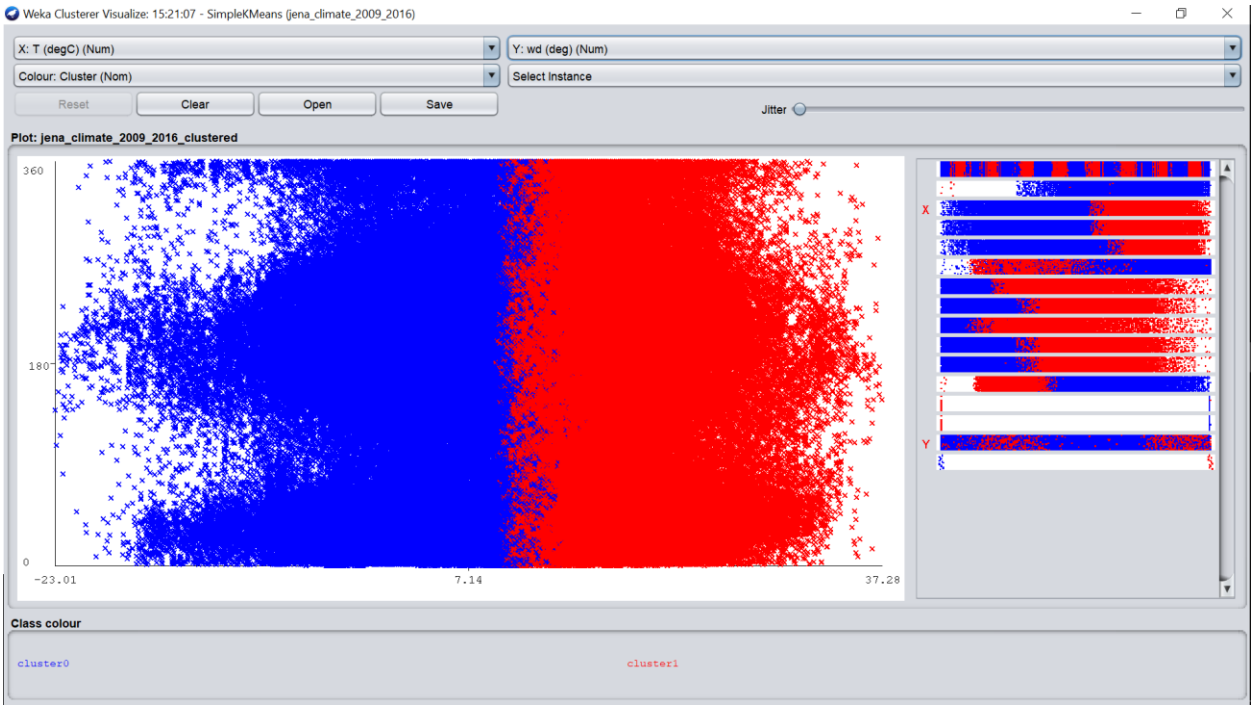
CS488

March 10, 2018

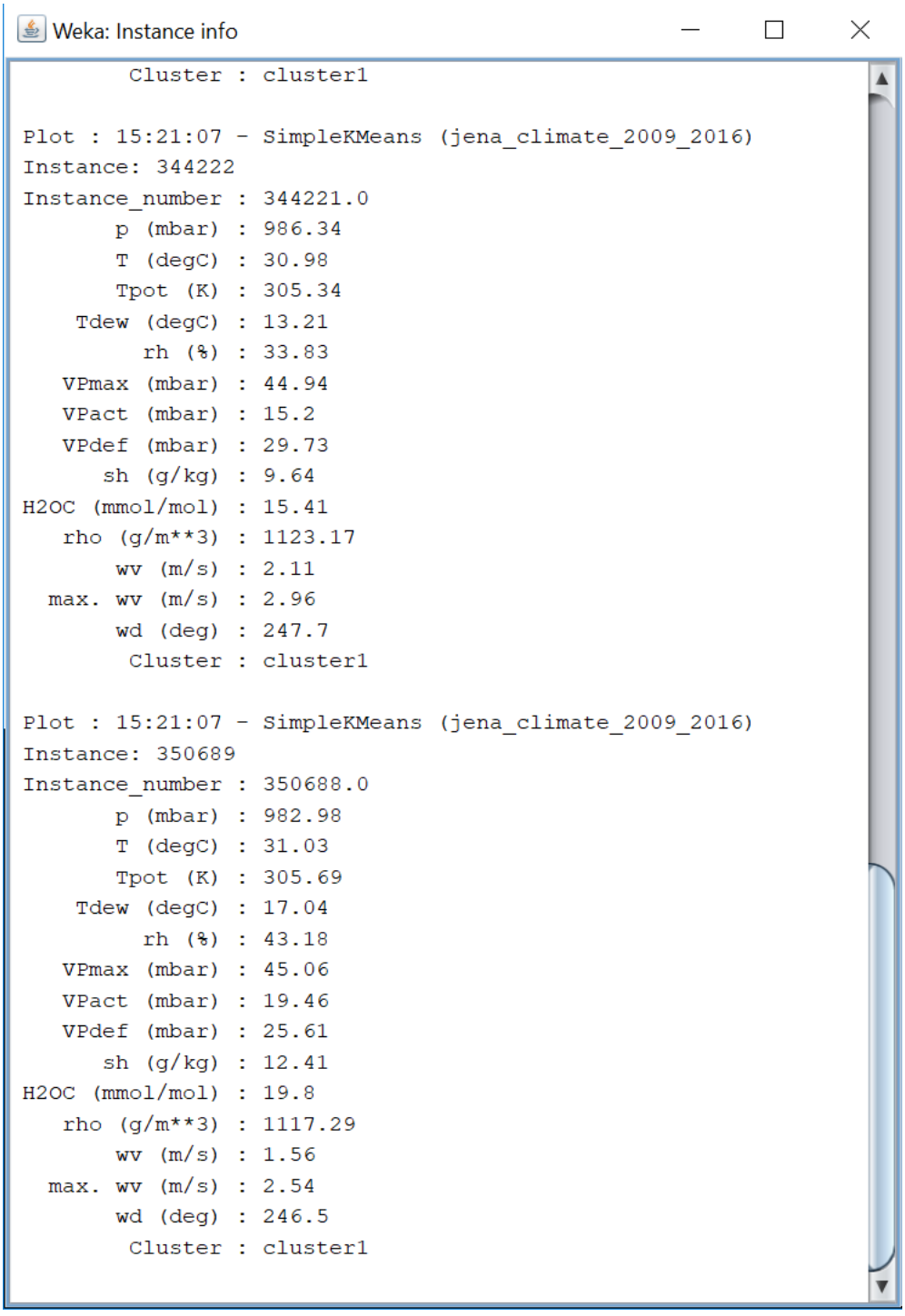
The dataset used in this project was a recording of the weather climate in Jena, Germany. There are 14 sensor readings that were recorded every 10 minutes in a span of 7 years.

Running one experiment using the k-mean clustering data mining algorithm I retrieved the following data analysis:





```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25
Relation:     jena_climate_2009_2016
Instances:    420551
Attributes:   15
              Date Time
              p (mbar)
              T (degC)
              Tpot (K)
              Tdew (degC)
              rh (%)
              VPmax (mbar)
              VPact (mbar)
              VPdef (mbar)
              sh (g/kg)
              H2OC (mmol/mol)
              rho (g/m**3)
              wv (m/s)
              max. wv (m/s)
              wd (deg)
Test mode:    user supplied test set: 420551 instances


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


kMeans
======

Number of iterations: 23
Within cluster sum of squared errors: 489105.8632929072

Initial starting points (random):

Cluster 0: '12.09.2015 21:10:00',988.08,19.43,293.6,12.03,62.23,22.61,14.07,8.54,8.9,14.24,1170.09,1.32,2.08,207.1
Cluster 1: '01.08.2013 13:00:00',991.11,27.9,301.83,15.7,47.47,37.64,17.87,19.77,11.29,18.03,1139.01,1.48,3.56,211.6

Missing values globally replaced with mean/mode


Final cluster centroids:
                                                Cluster#
Attribute                         Full Data             0                   1
                                 (420551.0)    (233051.0)          (187500.0)
=============================================================================
Date Time              01.07.2010 00:10:00 01.07.2010 00:10:00 01.07.2010 00:20:00
p (mbar)                          989.2128            989.4138            988.9629
T (degC)                            9.4501              3.3694             17.0082
Tpot (K)                          283.4927            277.3782            291.0928
Tdew (degC)                         4.9559              0.3045             10.7372
rh (%)                             76.0083             81.576              69.0879
VPmax (mbar)                       13.5763              8.2261             20.2261
VPact (mbar)                        9.5338              6.5647             13.2241
VPdef (mbar)                        4.0424              1.6614              7.0019
sh (g/kg)                           6.0224              4.1392              8.3631
H2OC (mmol/mol)                     9.6402              6.6364             13.3738
rho (g/m**3)                     1216.0627             1243.78            1181.6119
wv (m/s)                            1.7022              2.1174              1.1862
max. wv (m/s)                       3.0566              3.4681               2.545
wd (deg)                          174.7437            173.8208            175.8909



Time taken to build model (full training data) : 6.34 seconds



 === Evaluation on test set ===
 Clustered Instances

 0      233051 ( 55%)
 1      187500 ( 45%)
```


Weka: Instance info
Cluster : cluster1
Plot : 15:21:07 - SimpleKMeans (jena_climate_2009_2016)
Instance: 344222
Instance_number : 344221.0
p (mbar) : 986.34
T (degC) : 30.98
Tpot (K) : 305.34
Tdew (degC) : 13.21
rh (%) : 33.83
VPmax (mbar) : 44.94
VPact (mbar) : 15.2
VPdef (mbar) : 29.73
sh (g/kg) : 9.64
H2OC (mmol/mol) : 15.41
rho (g/m**3) : 1123.17
wv (m/s) : 2.11
max. wv (m/s) : 2.96
wd (deg) : 247.7
Cluster : cluster1
Plot : 15:21:07 - SimpleKMeans (jena_climate_2009_2016)
Instance: 350689
Instance_number : 350688.0
p (mbar) : 982.98
T (degC) : 31.03
Tpot (K) : 305.69
Tdew (degC) : 17.04
rh (%) : 43.18
VPmax (mbar) : 45.06
VPact (mbar) : 19.46
VPdef (mbar) : 25.61
sh (g/kg) : 12.41
H2OC (mmol/mol) : 19.8
rho (g/m**3) : 1117.29
wv (m/s) : 1.56
max. wv (m/s) : 2.54
wd (deg) : 246.5
Cluster : cluster1

After running it through Weka, I applied the data to a python program that implements k-mean clustering.

The following graph is my experimental result