**PROJECT 2 REPORT**

**Team No: 19**

**(k=9, seeds=1,2, years=2008,2009,2010,month=June(6))**

**Submitted By:**

**1)Sonika Sarangi Annaiah – 1001733497**

**2)Vaishakh Deepak Ghati - 1001738753**

**PROJECT OVERVIEW**

* The weather dataset for the state of Texas during the years 2006-2010 was provided to perform the analysis using k-means clustering.
* The task was to perform weather data analysis for the month of June in the years 2008,2009 and 2010.
* The datafiles contained weather data recordings for every hour of the day.
* The datafile for each year was pre processed to exactly obtain only the recordings for the month of June to perform the analysis.
* The dataframe of the month was queried. On the month data set we performed a GroupBy operation on STN number and aggregated the mean to get the monthly average values for each station which was given as input to perform the cluster analysis.
* SSE and jaccard values for the clusters were computed for Euclidean and Correlation measures for various years.
* For the comparison between various years: new datasets for each year wrt another year such that both the years had the same station numbers,- (sorted for cluster correspondence ) was obtained and jaccard values between various years was calculated.
* For cluster correspondence, the means of the clusters in both the years were compared by Euclidean and cosine similarity to check cluster of year1 corresponds to which cluster of year2. Based on this, the jaccard between various years were calculated.
* A graph plot of all the 4 attributes (Temp,STP,WDSP,DewP) values vs cluster values for various years were plotted to understand the change in weather Pattern across Station-clusters.
* Pre-processing was done mainly using Excel and R(lubridate and sqldf libraries)

**ANALYSIS**

**1)** For the year 2008 with k-value = 9 and seed=1

Euclidean Distance within Cluster SSE was found to be:

6.188944e+02 1.523877e+03 2.655364e+02 2.113011e+07 3.415473e+00 1.780552e+03 2.473835e+02 6.215257e+01 5.510252e+04

**SSE = 21189714.33**

K-means clustering with 9 clusters of sizes 32, 2, 14, 26, 35, 12, 3, 35, 2 and cluster means was obtained.

Pearson correlation within SSE was found to be :

1.481185e-15 3.333596e-06 1.121736e-10 1.319413e-13 2.160342e-11

2.634920e-16 0.000000e+00 3.296490e-15 0.000000e+00

**SSE = 3.334211098 x 10^(-6)**

K-means clustering with 9 clusters of sizes 36, 10, 45, 15, 2, 39, 1, 12, 1 and cluster means was obtained.

Jaccard value for Euclidean v/s Pearson is **0.5868403**.

**2)** For the year 2008 with k-value = 9 and seed=2

Euclidean Distance within Cluster SSE was found to be:

2.166056e+03 1.139174e+01 4.095683e+00 2.473835e+02 9.134023e+01

0.000000e+00 1.523877e+03 6.191138e+06 3.654880e+01

**SSE=6195218.691**

K-means clustering with 9 clusters of sizes 46, 15, 58, 3, 15, 1, 2, 10, 11 and cluster means was obtained.

Pearson correlation within cluster SSE was found to be :

1.481185e-15 1.319413e-13 2.352555e-10 7.851287e-10 3.296490e-15 2.634920e-16 3.333596e-06 2.810451e-12 0.000000e+00

**SSE=3.336 x 10^(-6)**

K-means clustering with 9 clusters of sizes 36, 15, 10, 35, 12, 39, 10, 3, 1 and cluster means was obtained.

Jaccard value for Euclidean v/s Pearson is **0.4991404**

**3)** For the year 2009 with k-value = 9 and seed=1

Euclidean Distance within Cluster SSE was found to be:

4.979357e+01 9.894912e+05 4.912541e+01 4.032057e-01 2.281432e+02 3.770685e+04 2.291909e+02 7.173365e+01 8.429391e+01

**SSE = 1027909.527**

K-means clustering with 9 clusters of sizes 7, 3, 20, 7, 49, 3, 17, 25, 40 and cluster means was obtained.

Pearson correlation within cluster SSE was found to be :

6.298816e-12 4.650913e-13 1.871452e-06 2.871722e-12 9.657769e-10 7.777784e-14 2.352571e-13 1.350220e-13 1.217448e-13

**SSE= 1.87 x 10^(-6)**

K-means clustering with 9 clusters of sizes 18, 28, 10, 32, 3, 26, 15, 32, 7 40 and cluster means was obtained.

Jaccard value for Euclidean v/s Pearson is **0.3452863**

**4)**For the year 2009 with k-value=9 and seed =2

Euclidean Distance within Cluster SSE was found to be:

7.173365e+01 9.297329e+01 2.144586e+06 4.300706e+02 8.429391e+01 2.389402e+02 1.798574e+01 4.979357e+01 3.770685e+04

**SSE = 2183278.627**

K-means clustering with 9 clusters of sizes 25, 17, 10, 7, 40, 48, 14, 7, 3 and cluster means was obtained.

Pearson correlation within cluster SSE was found to be :

6.497342e-12 2.027273e-12 1.217448e-13 9.657769e-10 9.768096e-14 2.341712e-11 0.000000e+00 1.871452e-06 3.045416e-11

**SSE = 1.871 x 10^(-6)**

K-means clustering with 9 clusters of sizes 48, 17, 7, 3, 25, 2, 1, 10, 58 and cluster means was obtained.

Jaccard value for Euclidean v/s Pearson is **0.4322326**

**5)**For the year 2010 with k-value = 9 and seed=1

K-means clustering with 9 clusters of sizes 3, 33, 19, 18, 1, 1, 10, 44, 54 and cluster means was obtained.

Euclidean Distance within Cluster SSE was found to be:

 1.679548e+03 6.537328e+01 3.418608e+01 1.941475e+02 0.000000e+00 0.000000e+00 2.272087e+06 5.793618e+01 1.657394e+02

**SSE = 2274283.922**

Pearson correlation within cluster SSE was found to be :

1.341661e-13 3.382373e-06 9.026845e-11 4.543377e-13 1.602775e-13

3.978156e-12 9.097021e-12 1.202942e-14 1.871473e-12

**SSE = =3.3824 x 10^(-6)**

K-means clustering with 9 clusters of sizes 7, 11, 3, 11, 24, 28, 37, 14, 48 and cluster means was obtained.

Jaccard value for Euclidean v/s Pearson is **0.4023097**

**6)**For the year 2010 with k-value = 9 and seed = 2

K-means clustering with 9 clusters of sizes 30, 7, 17, 13, 35, 33, 1, 37, 10 and cluster means was obtained.

Euclidean Distance within Cluster SSE was found to be:

5.375172e+01 4.611342e+02 8.085016e+01 7.275536e+00 8.824591e+01 2.202722e+01 0.000000e+00 6.879241e+02 2.272087e+06

**SSE = 2273488**

Pearson correlation within cluster SSE was found to be :

4.012201e-12 5.315925e-11 2.453005e-12 1.180525e-12 9.026845e-11 6.048497e-12 1.511341e-11 5.285988e-13 3.382373e-06

**SSE =3.38 x 10^(-6)**

K-means clustering with 9 clusters of sizes 34, 3, 37, 32, 3, 14, 34, 15, 11 and cluster means was obtained.

Jaccard value for Euclidean v/s Pearson is **0.3558183**

**INTERPRETATION**

In all the previous comparisons, we see the intra cluster SSE and sum of SSE is lesser for Pearsons correlation. So Pearson is a better clustering technique using SSE metric(lower SSE means better). One reason, for this behaviour is that correlation uses covariance between the 2 datasets and std deviation. Therefore, data scaling is handled, and fact that say [3,10000] and [100,20000] in this distance is completely predominated by parameter2 and undermines parameter1 in clustering. But Correlation, handles these cases very well.

For each cluster and in total for all years and all seed values, we see that Correlation performs better.

Using Jaccard values: For different seed values, the value of Jaccard between Pearsons and Euclidean varies from 0.5-0.6-0.7. (observed for seed values =25,26,30,40)

(This indicates that because it’s the same dataset and only distance measures are changing , cluster vectors are nearly similar. As cluster points are assigned to the same cluster in both the metrics )

**7)Year to Year Comparisons –(2008,2009) ,(2009,2010), (2008,10)- for common stations.**

1. (2008,2009) – seed=1: - **Euclidean Jaccard value**=0.2566762,

**Pearsons** **Jaccard value**= 0.1838278

2. (2009,2010) – seed=1: - **Euclidean Jaccard value**=0.2960862,

**Pearsons** **Jaccard value**= 0.717094

2. (2008,2010) – seed=1: - **Euclidean Jaccard value**=0.2286184

**Pearsons** **Jaccard value**= 0.1851531

**INTERPRETATION**

For this and many seeds we see that for the year 2009,2010 has higher jaccard value than 2008-2009 years. This means stations did not shift clusters to the extent they did in 2008-09. Implying that more weather change happened more between 8-09 as compared to 9-10 for the month of June.

Also for many seeds, it was observed that Jaccard value between 2008-10 was the lowest of the 3. Since we have 2009-10 as the highest , definitely between 2008-10, many fluctuations have happened since the time difference is 2 years. Also we see that Pearsons Jaccard is lower than Euclidean Jaccard.

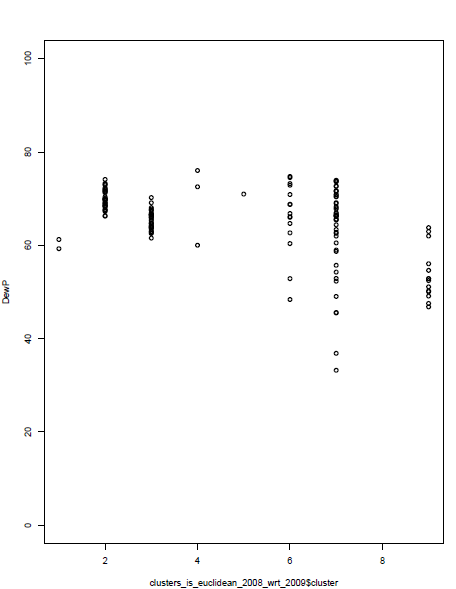
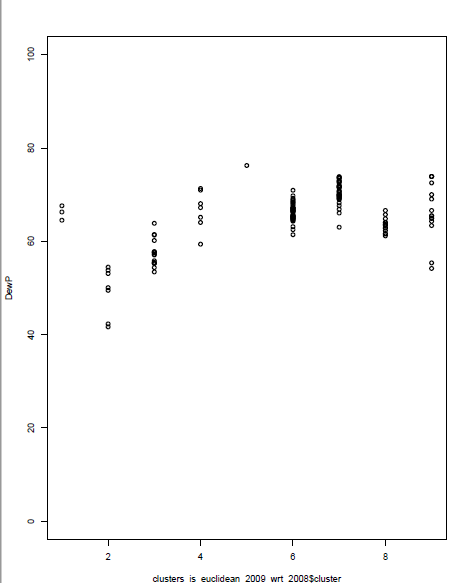
For the years: 2008-9 and 9-10 Jaccard values of Euclidean is observed to be higher than that of Correlation, indicating that for clusters tend to show larger(although not much) similarity when using Euclidean as compared to correlation , as Euclidean can be skewed for some dimensions that dominate by scale. Therefore, Euclidean might has a better tendency to say, no weather change has occurred than Pearson. Furthermore, if the clusters are converging – Euclidean can make it difficult for points to change clusters than correlation.(only if weather data points change drastically, the new data point will be picked up by a new cluster in Euclidean as compared to Pearsons.) However, owing to lower values of Jaccard, across years, we see that weather change is happening.

**VISUALISATION**

(Few more visualisation of clusters is present in the visualisation folder of the project)

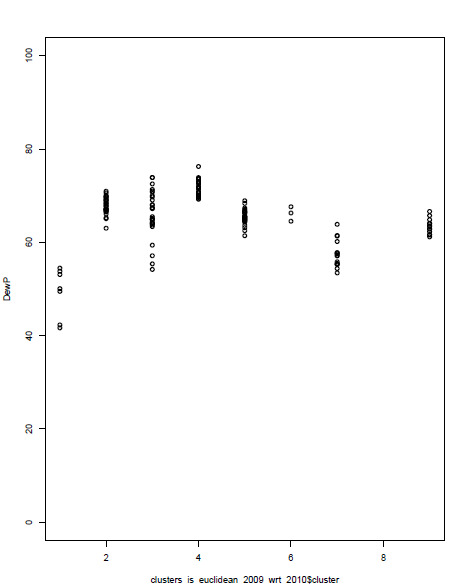
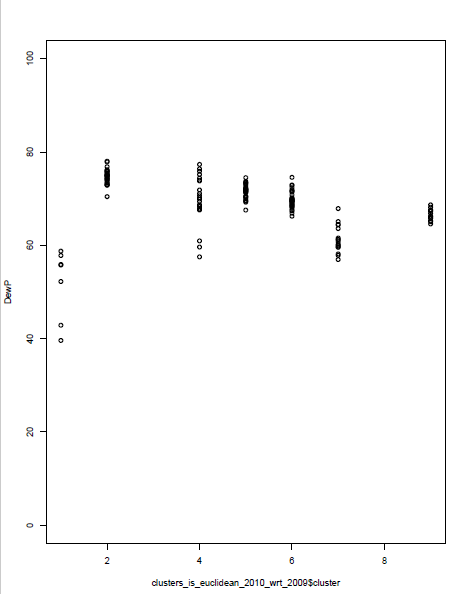
**Dewp:** 2008-09 DewP values across various clusters.

Euclidean: 2008 Euclidean: 2009

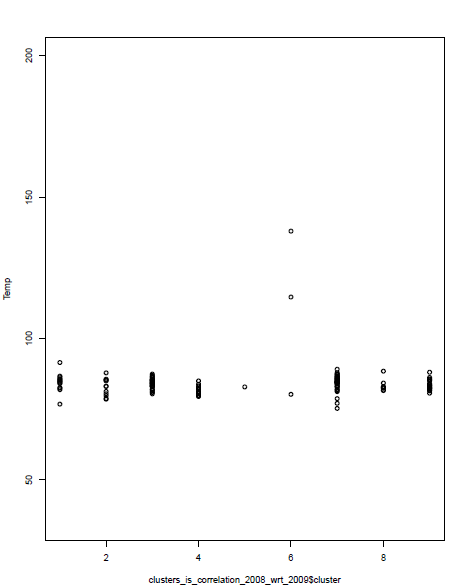
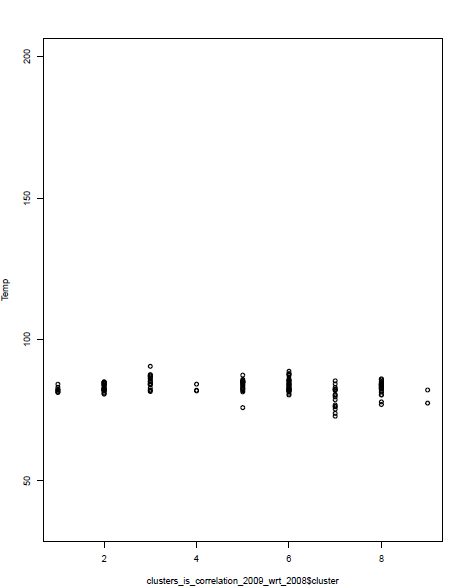
**Dewp:** 2009-10 DewP values across various clusters.

Euclidean: 2009 Euclidean: 2010

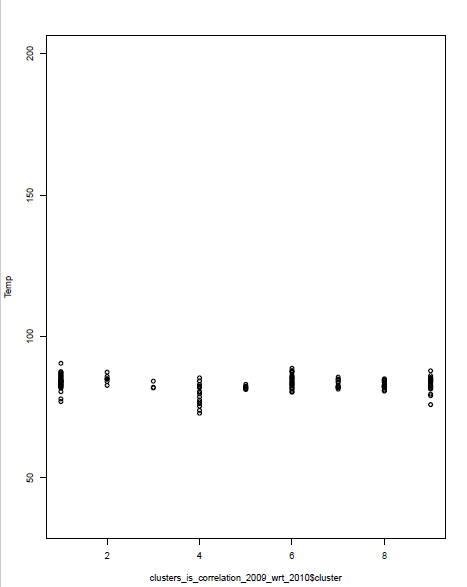
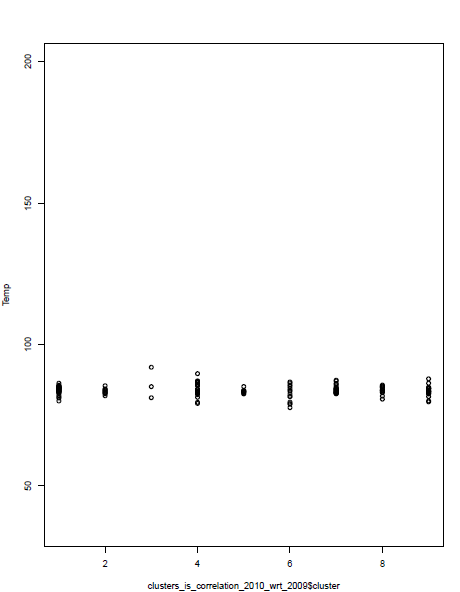
 

**Temp:**

Pearsons correlation 2008 Pearsons correlation 2009

Pearsons correlation 2009 Pearsons correlation 2010

Similar analysis for WDSP AND STP was performed. (Present in the visualisation folder of the project main folder).

There are about 120-130 common stations, and even few stations that go to another cluster will see a reduction of jaccard value that may not be evident from these clusterplots.

**CHALLENGES**

1. How to Map cluster to cluster across years. (Handled by sorting and coding an algorithm to compute cluster-means similarity)
2. Interpreting the meaning of Jaccard values and its usability and fluctuation to different datasets. Also interpreting how Euclidean and Pearson can perform for attributes with different scale/range. (Enhancing domain knowledge)
3. Handling a Huge dataset and many fold pre-processing. (File compressions, Excel operations and reduce the creation of irrelevant dataframes by logically structuring the code)

**FILE NAMES**

1. Rfiles folder contains all the R source file used in the project.
2. Visualisation folder has all the graph plots and analysis. (Many year wise visualisation excluded from the report are present in this folder)
3. Dataset folder has various datasets related to pre-processing and final input dataset. ( .g files are excluded)