**DATA MINING Project II: CLUSTERING**

**Team No: 11**

**Members: Vamsee Krishna B, Nikhileshwar I**

**What are given?**

We are given with the weather data set in separate files for the state of Texas for the years 2006 to 2010 (5 years) for analysis using k-means clustering.

**Pre-Processing:**

We have taken the given file and imported into the R Studio using the read.csv function and we have started the Preprocessing.

Initially we are working with the data set corresponding to the year 2006.

Initially we have separated the data column wise and then we have added the headers

We have given column names to all the columns in our data set

In our dataset, we have a column called **yearModa\_hr** using this column we have separated the year, month, day and hour Using the Lubridate R studio library after that we have started the actual preprocessing of the data given to us.

The seed value given to our team is 25 and a month specific to our team is October so we're working on the month of October

So from the main data set we have made a subset containing just the data of October month.

So from the complete subset we're taking the sample of 4000 records because the sample size given to our team is 4000

And now we are dividing the data set into training and testing data sets as the training and testing data split for our team is 80-20 we are converting 80% of our data into training set an 20% of our data into testing set

In the data set we currently have we have some unknown values in the form of 9999.9 we are substituting these values using the mean of the entire column, So that the data won't be affected.

In our given data set we are given the data of every hour. For every state there will be unique station numbers defined and as we are collecting the data for every hour we will have a station number for each and every row in our data set there is a high possibility that there are there will be duplicate records that is station numbers In our data. so we want to prepare data set with only unique station numbers for achieving this we are taking the monthly mean group by the station number so that we will get any unique data set containing station members with no duplicate values we will be aligning with the respective rows in our data set.

**Clustering and plotting using K-Means:**

For the following column names we are running the K-Means:

* Temp
* DewP
* STP
* WDSP

Initially we will be setting the seed value as 25 and do the following .

For the K means implementation we are given the value of K's 5 so we will be getting 5 clusters And so we will be getting 5 centers and we will be plotting the point accordingly.

we will be implementing the K means for 2 distances that is Euclidean and Pearson

for each column starting with temperature we will be applying the K means algorithm giving the euclidean as a distance measure and we will be plotting it on the graph as per the centers of each cluster.

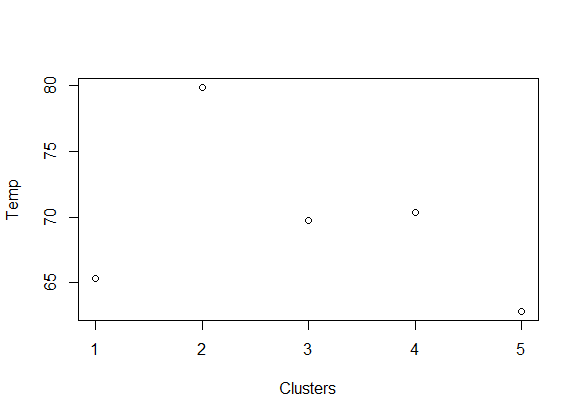
At once we will be getting 5 points on the graph.

We will be repeating the same for the remaining 3 columns also.

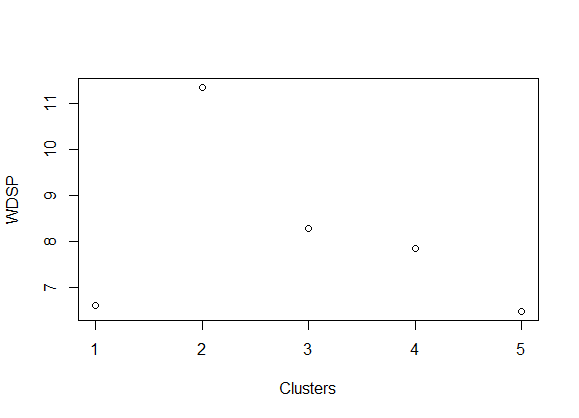
\*Now we will repeat the same what we did above now modeling but by changing the seed value to 20 .

\*Below are some plots.

Year 2006, seed 25, Euclidean and for column Temp.



Year 2006, seed 20, Euclidean and for column WDSP.



Year 2007, seed 25, Euclidean and for column DewP.



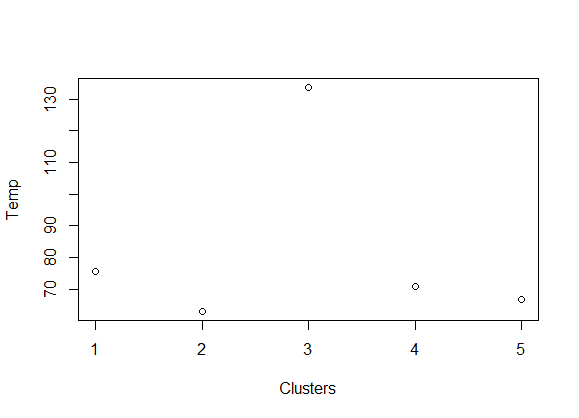
Year 2008, seed 25, Euclidean and for column STP.



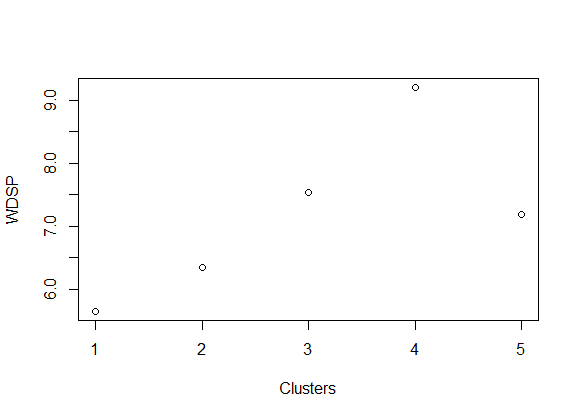
\*We will be repeating the same process giving the Pearson a distance measure.

\*Below are some plots.

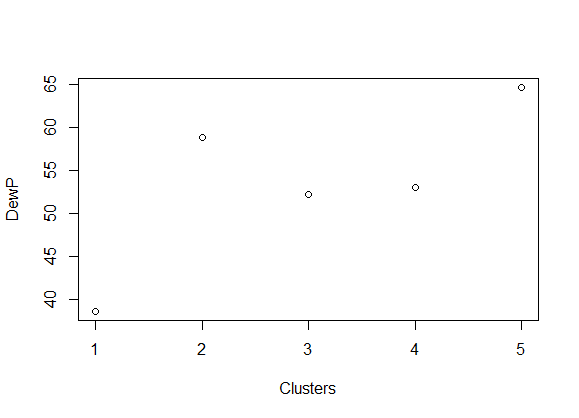
Year 2006, seed 25, Pearson and for column Temp.



Year 2006, seed 20, Pearson and for column WDSP.



Year 2007, seed 25, Pearson and for column DewP.



Year 2008, seed 25, Pearson and for column STP.

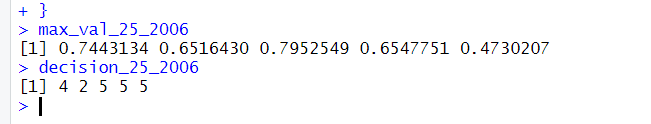


**Comparison and Analysis using Jaccard:**

For achieving this we are installing a library ClustEval and using the function cluster\_similarity in that we are declaring 2 arrays one for the max value where the maximum similarity is stored and other for the decision where the actual cluster value is stored.

For calculating this, we are iterating a for loop with “i” value as 5 as the K value is 5 for us and passing the labels defined above, seed as 25 for year 2006 for now and the similarity s Jaccard.

After Executing it we got the below output which indicated which clusters are most similar in the Euclidean and Pearson.



**Analysis:**

The above output says that the 1st cluster in Euclidean is similar to 4th cluster in Pearson with 0.7443134 similarity.

The 2nd cluster in Euclidean is similar to 2nd cluster in Pearson with 0.6516430 similarity

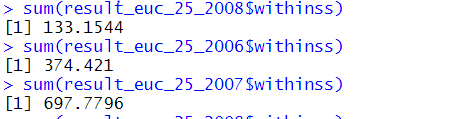
The 3rd cluster in Euclidean us similar to 5th cluster in Pearson with 0.7952549 similarity

The 4th cluster in Euclidean us similar to 5th cluster in Pearson with 0.6547751 similarity

The 5th cluster in Euclidean us similar to 5th cluster in Pearson with 0.4730207 similarity

**Comparing the two clustering’s using SSE for the Euclidean correlation coefficient metric:**

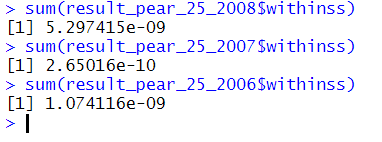
Below is the computed output for Euclidean correlation coefficient metric for all the 3 years



The Above output infers that in the year 2008 the clustering went well with less SSE.

**Comparing the two clustering’s using SSE for the Pearson correlation coefficient metric:**

Below is the computed output for Pearson correlation coefficient metric for all the 3 years



The Above output infers that in the year 2006 the clustering went well with less SSE.

**3 difficulties faced and resolved:**

* Separating the column **yearModa\_hr** into separate columns, we achieved this by usingLubridate library functions.
* Eliminating the missing values which were prepopulated to be 9999.99, we did this by looping through the entire column and replacing that value with the mean.
* Grouping the entire dataset to get the distinct station numbers, we did this by using the aggregate function using group by station number.

**References:**

<https://cran.r-project.org/web/packages/amap/index.html>

<https://www.geeksforgeeks.org/k-means-clustering-introduction/>

<https://www.datacamp.com/community/tutorials/k-means-clustering-r>

<https://stackoverflow.com/questions/46789010/error-in-bind-rows-x-id-column-cant-be-converted-from-factor-to-numeric>

<https://data.library.virginia.edu/working-with-dates-and-time-in-r-using-the-lubridate-package/>