**Name:** Shuvrima Alam, **Student ID:** 1001085726, **Team No.3**, **Date:** 10/24/2019

## Basic Information

The value of **k** used :5 | **Years** analyzed: 2006, 2007, 2008 | **Month**: February | **Method** used: KMeans clustering | Distance **metrics**/Similarity metrics used: Euclidean, Pearson correlation coefficient, Jaccard coefficient | **Seed** Values: 10, 150.

## Overall Status:

The first approach was to pre-process the data and get the records for month February. Subseting along with partial string matching was used to get values with '02' in YEARMODA variable for each dataframes. The time column was divided into year, month, day, hour for ease of average calculations using lubridate library. For each data frame the missing values were replaces with NA values in order to ignore the values during average calculations. Starting with year 2006, the aggregate function was used to calculate daily averages of temp, stp, dewp, wdsp with respect to day and station numbers variables and the values were saved in a vector.

Using the daily averages, a new vector with monthly averages were calculated in same manner with respect to station numbers. All the monthly average values for each weather variables were then saved in a new table with the corresponding station numbers using merge function. This data frame has be used for clustering using Kmeans function from amap library. First the seed value was set to 10 and clustering algorithm was used first for Euclidean metric and then for Pearson coefficient metric. The corresponding clusters were saved in two variables. Then the clustering process was repeated for each metric with a different seed value of 150. The whole process was repeated in same manner for year 2007 and 2008. To compare Euclidean and Pearson metrics for each seed value Jaccard coefficient was used using clusteval library in R and function cluster_similarity. Once the clusters were made for each year the few of the graphs were plotted for visualization purposes on Texas map using ggplot library and Google geocoding API.

In attempt to do year to year analysis of the weather, each cluster from one year has to be compared to every cluster in another year and the most similar ones are considered and compared against each other. These similarities are measured using Jaccard coefficient. Set based jaccard has to be used in order to pass the clusters sets where station numbers represent ids. Jaccard function was written and a loop was run which compared the clusters using that function. Using R functions to calculate intersection and union of sets, the Jaccard coefficients of each cluster in a particular year with respect to another cluster in other year are calculated.

Wherever intersections of data frames were required setdiff was used to sort the data and consider the stations present in all datasets for each year. Overall, project successfully maps stations to appropriate clusters based on weather similarity.

## File Descriptions:

| File Name | Function |
|---|---|
| proj2.R | Contains all the code written for this project and analysis. |
| hourly_2006 | Contains the data set for year 2006 imported and manipulated. The first line with variable names were changed by removing the commas in between for ease of data import. |

| | |
|---|---|
| hourly_2007 | Contains the data set for 2007 imported and manipulated. The first line with variable names were changed by removing the commas in between for ease of data import. |
| hourly_2008 | Contains the data set for 2008 imported and manipulated. The first line with variable names were changed by removing the commas in between for ease of data import. |

## List of Algorithms used:

*Only KMeans clustering algorithm from amap.*

## Division of Labor:

| **Task** | **Time Taken(hours)** |
|---|---|
| Reading the theory behind K-means clustering and related knowledge | 4 |
| Reading about the libraries or functions required for this project | 10 |
| Pre-processing data | 1 |
| Writing code for clustering and calculating similarities | 4 |
| Visualization of the clusters | 2 |
| Debugging | 3 |
| Writing Report and Analysis | 5 |
| *This project was completed **individually*** | **Overall Time Taken :29** |

## Problems Encountered and Handled:

| **Problem Type** | **Method of Handling** |
|---|---|
| Finding appropriate functions to use. | Due to many options available in R it can be confusing to choose the best one. Eventually, it did not matter if best function was chosen as long as it generated correct results as suggested in class. |
| Finding correct way to interpret. | The program generated results but analyzing those results required knowledge and correct understanding of the mining concepts. In order to do appropriate interpretation personal research was done and methods suggested in class was adapted. |
| Confusion regarding SSE and Jaccard coefficient. | The confusion as to if withinss produced by amap KMeans was really the SSE that was asked for lead me to spent some time but was cleared in class. How to figure out and analyze Jaccard for year to year was not clear but upon consulting was solved. |

## Comparison analysis of SSE for Euclidian metric

The $**withinss**: is the within cluster sum of squares. So, it results in a vector with a number for each cluster. This value is expected to be as low as possible for each cluster, since we would like to have homogeneity within the clusters. Changing the seed value yields to some change in SSE in this case and bigger seed value yields to better results. The SSE for each cluster shows improvement and the better results are shaded in green.

| Seed Value 10 | Seed Value 150 |
|---|---|
| **SSE of clusters:** 9288604,59893,1527,1434834,533066 | **SSE of clusters:** 3597161, 5095665, 5692, 1527, 1434834 |
| **Total SSE:** 11317924 | **Total SSE:** 10134879 **(better)** |
| **Overall:** ***Seed Value 150 yields better results.*** | |

## Year 2007

| Seed Value 10 | Seed Value 150 |
|---|---|
| **SSE of clusters:** 8638722, 57468,1374,1434827,532805 | **SSE of clusters:** 3578424, 5291501, 5637, 1374 ,1434827 |
| **Total SSE:** 10665196 | **Total SSE:** 10311763 **(better)** |
| **Overall:** ***Seed Value 150 yields better results.*** | |

## Year 2008

| Seed Value 10 | Seed Value 150 |
|---|---|
| **SSE of clusters: 8407631, 60959, 1384, 1230469, 532873** | **SSE of clusters: 60959, 1384, 532873, 1230469, 8407631** |
| **Total SSE:** 10233316 | **Total SSE:** 10233316 |
| **Overall:** ***For no difference was observed.*** | |

To summarize change of seed value to a **higher number** yielded to slightly **better** results as far as SSE goes for year 2006 and 2007 but not for 2008. Most of the clusters for seed 150 either had same SSE values as ones for seed 10 or less. **WithinSS – The Within sum of squares measures the sum of the squared difference from the cluster center**. A smaller WithinSS (or SSE) means there is less variance in that cluster's data. For this analysis, better results were primarily judged on basis of total withinss for each clustering.

## Comparison analysis of SSE for Pearson metric

## Year 2006

| Seed Value 10 | Seed Value 150 |
|---|---|
| **SSE of clusters:** 1.843633e-21, 1.417531e-20, 7.784945e-19, 8.948697e-19 ,7.726941e-23 | **SSE of clusters:** 7.784945e-19 ,1.843633e-21, 8.948697e-19, 1.417531e-20, 7.726941e-23 |
| **Total SSE: 9.887** x **10^-8** | **Total SSE: 9.887** x **10^-8** |
| **Overall: No difference.** | |

## Year 2007

| Seed Value 10 | Seed Value 150 |
|---|---|
| **SSE of clusters:** 8.000151e-21, 3.311393e-20, 3.609413e-21 ,1.004568e-19, 3.023840e-21 | **SSE of clusters:** 1.004568e-19, 3.345319e-21 2.785424e-21, 3.311393e-20 ,8.000151e-21 |
| **Total SSE: 2.355 x 10^**-8 | **Total SSE: 2.317 x 10^**-8 |
| **Overall:** ***Seed Value 150 yielded slightly better results.*** | |

## Year 2008

| Seed Value 10 | Seed Value 150 |
|---|---|

| | |
|---|---|
| **SSE of clusters:** 1.586251e-20, 4.547562e-21, 7.446383e-22, 1.066789e-19, 6.280131e-19 | **SSE of clusters:** 5.050382e-21, 1.586251e-20, 1.066789e-19, 6.280131e-19, 7.361434e-22 |
| **Total SSE: 4.996 x 10^**-8 | **Total SSE: 5.030 x 10^**-8 |
| **Overall:** ***Seed Value 10 yielded better results.*** | |

Overall, for pearson there were not much differences between SSEs for different seed values. For year 2006, no difference was observed. For 2007, seed value 150 yielded better results and for year 2008 seed value 10 was better as lower withinss values are desired.

## Comparison and analysis the clusters from different metric using Jaccard

### Year 2006

The values in the tables below show Jaccard coefficients for Euclidean metric vs Pearson metric for each seed values using clusteval library. Jaccard is a **measure** of **similarity** for the two sets of data.

| **Euclidean Metric vs. Pearson Metric** | |
|---|---|
| **Jaccard Coefficient:** | 0.2275813 (seed 10)     0.2293993 (seed 150) |
| **Overall:** *According to these jaccard values euclidean and pearson metric yields to very different clusterings in cases of both seed values. Although higher seed value lead to slightly more similar clustering.* | |

### Year 2007

| **Euclidean Metric vs. Pearson Metric** | |
|---|---|
| **Jaccard Coefficient:** | 0.2928986 (seed 10)     0.29253 (seed 150) |
| **Overall:** *According to these jaccard values euclidean and pearson metric yields to very different clusterings in cases of both seed values. Although higher seed value lead to slightly more similar clustering.* | |

### Year 2008

| **Euclidean Metric vs. Pearson Metric** | |
|---|---|
| **Jaccard Coefficient:** | 0.2278448 (seed 10)     0.2299816 (seed 150) |
| **Overall:** *According to these jaccard values euclidean and pearson metric yields to very different clusterings in cases of both seed values. Although higher seed value lead to slightly more similar clustering.* | |

The above results make sense as for KMeans generally euclidean is used. Pearson and Euclidean are likely to yield different clusterings. K-means does *not* minimize distances. It minimizes the *sum of squared 1-dimensional deviations* (SSE) which is mathematically equivalent to squared Euclidean distance, so it does minimize *Euclidean* distances, as a mathematical side effect.

## Year-wise Analysis

Seed Value considered: 10 Method: **Euclidean**

Using Jaccard coefficient formula below tables were obtained.

### Comparison of Y1(2006),Y2(2007)

The cells colored same indicate most similar clusters.

| | Year 2006 c1 | Year 2006 c2 | Year 2006 c3 | Year 2006 c4 | Year 2006 c5 |
|---|---|---|---|---|---|
| Year 2007 c1 | 0 (approx) | 0.8 | 0 | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| Year 2007 c2 | 0(approx) | 0 | 0.9051724 | 0 | 0 |
| Year 2007 c3 | 0(approx) | 0 | 0.02586207 | 0 | 0 |
| Year 2007 c4 | 0.8846154 | 0 | 0 | 0.07692308 | 0.03846154 |
| Year 2007 c5 | 0(approx) | 0 | 0.06896552 | 0 | 0 |

Between 2006 and 2007 there has been slight weather changes. Cluster 1 from year 2006 for example is compared to Cluster 4 in Year 2007 due to highest Jaccard coefficient. However, there is no 100 % overlap (Jaccard coefficient of 1) which indicates difference in weather clusters and hence weather change and climate shift has happened. Same can be said for other other cells that have same color.

**Comparison of Y2(2007),Y3(2008)**

| | Year 2007 c1 | Year 2007 c2 | Year 2007 c3 | Year 2007 c4 | Year 2007 c5 |
|---|---|---|---|---|---|
| Year 2008 c1 | 0 | 0 | 0 | 0.8846154 | 0 |
| Year 2008 c2 | 0.8 | 0 | 0 | 0 | 0 |
| Year 2008 c3 | 0 | 0.9051724 | 0.02586207 | 0 | 0.06896552 |
| Year 2008 c4 | 0 | 0 | 0 | 0.07692308 | 0 |
| Year 2008 c5 | 0 | 0 | 0 | 0.03846154 | 0 |

Least weather change has been for stations located in cluster 2 in 2007 and stations in cluster 3 of year 2007.

**Comparison of Y1(2006),Y2(2007),Y3(2008)**

**Y1 vs. Y3**

| | Year 2006 c1 | Year 2006 c2 | Year 2006 c3 | Year 2006 c4 | Year 2006 c5 |
|---|---|---|---|---|---|
| Year 2008 c1 | 1 | 0 | 0 | 0 | 0 |
| Year 2008 c2 | 0 | 1 | 0 | 0 | 0 |
| Year 2008 c3 | 0 | 0 | 1 | 0 | 0 |
| Year 2008 c4 | 0 | 0 | 0 | 1 | 0 |
| Year 2008 c5 | 0 | 0 | 0 | 0 | 1 |

Between year 2006 and 2008 the weather changed in 2007 somewhat from 2006 but then in 2008 it became similar to 2006 in the month of February as indicated by Jaccard coefficient according to our three year analysis.

The clusters who are not shaded are not part of my analysis as not much details can be derived from their values.

Seed Value considered: 150 Method: **Pearson**

Using Jaccard coefficient formula below tables were obtained.

**Comparison of Y1(2006),Y2(2007)**

| | Year 2006 c1 | Year 2006 c2 | Year 2006 c3 | Year 2006 c4 | Year 2006 c5 |
|---|---|---|---|---|---|
| Year 2007 c1 | 0 (approx) | 1 | 0 | 0 | 0 |
| Year 2007 c2 | 0(approx) | 0 | 0.9051724 | 0 | 0 |
| Year 2007 c3 | 0(approx) | 0 | 0.02586207 | 1 | 0 |
| Year 2007 c4 | 0.884 | 0 | 0 | 0 | 0.03846154 |
| Year 2007 c5 | 0(approx) | 0 | 0.06896552 | 0 | 0 |

For intersection of the stations Pearson and Euclidean metrics' Jaccard coefficients do not show any difference and the results and conclusion about weather remains same as Euclidean metric analysis/clustering.

## Comparison of Y2(2007),Y3(2008)

| | Year 2007 c1 | Year 2007 c2 | Year 2007 c3 | Year 2007 c4 | Year 2007 c5 |
|---|---|---|---|---|---|
| Year 2008 c1 | 0 | 0 | 0 | 0.8846154 | 0 |
| Year 2008 c2 | 0.9 | 0 | 0 | 0 | 0 |
| Year 2008 c3 | 0 | 0.9051724 | 0.02586207 | 0 | 0.06896552 |
| Year 2008 c4 | 0 | 0 | 0 | 0.07692308 | 0 |
| Year 2008 c5 | 0 | 0 | 0 | 0.03846154 | 0 |

Again, not a significant change from Euclidean outcome. The weather analysis would be same.

## Comparison of Y1(2006), Y2(2007),Y3(2008)

### Y1 vs. Y3

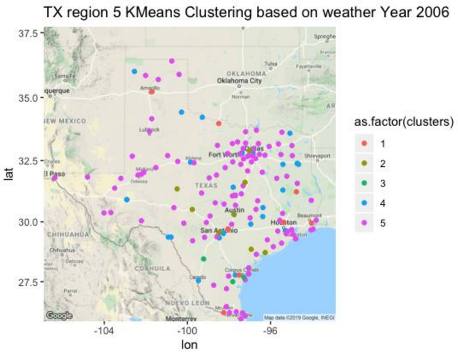
| | Year 2006 c1 | Year 2006 c2 | Year 2006 c3 | Year 2006 c4 | Year 2006 c5 |
|---|---|---|---|---|---|
| Year 2008 c1 | 1 | 0 | 0 | 0 | 0 |
| Year 2008 c2 | 0 | 0.9 | 0 | 0 | 0 |
| Year 2008 c3 | 0 | 0 | 1 | 0 | 0 |
| Year 2008 c4 | 0 | 0 | 0 | 1 | 0 |
| Year 2008 c5 | 0 | 0 | 0 | 0 | 1 |

Same conclusion as Euclidean for the weather. Only slight change in Jaccard of c2. Color code can be followed for detailed analysis picture. Same colors have been compared for weather change analysis.

## Visualization Plots

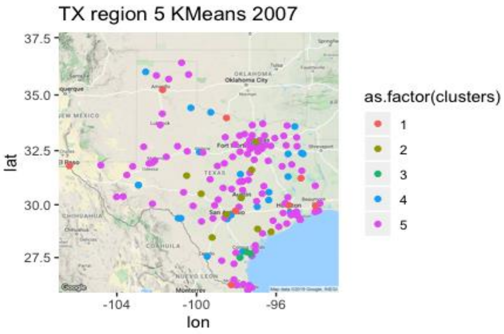
**Spatial Coverage:** TX, **Temporal Coverages:** 2006, 2007, 2008

### Euclidean Metric KMeans 2006 (Seed :10)



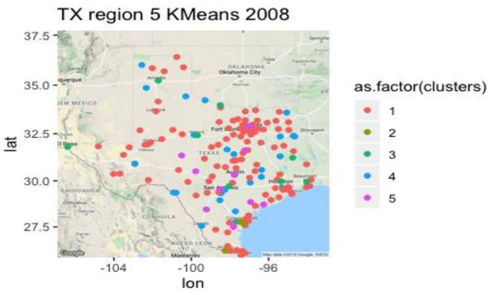
The color represents each cluster and these clusters are based on 4 weather dimensions using K-Means. Clusters with same color have similar weather and they are located on different Texas regions as shown on map. Most of the stations on Texas have similar weather as majority points are pink-colored. Very few stations have drastically different weather as represented by the orange colored dots.

### Euclidean Metric KMeans 2007 (Seed: 10)



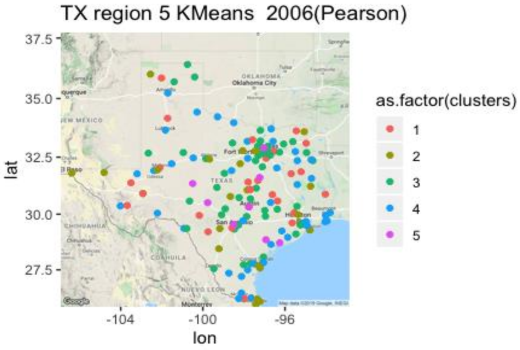
Most of stations still belong to pink colored cluster and have similar weather in Texas. However, there were some changes for instance the change in weather of station in El Paso which now belongs to orange cluster indicating a different weather type and that station previously belonged to majority cluster in year 2006.

## **Euclidean Metric KMeans 2008(Seed :10)**

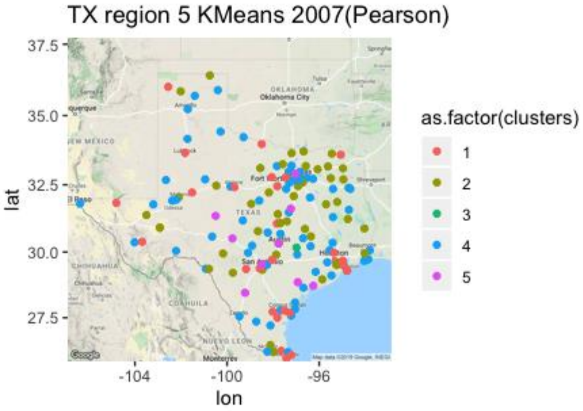


The colors are randomly assigned and hence now the majority points cluster are orange. Not a drastic weather change has been observed after two years in case of most stations. Most stations in the regions where weather was similar in past years still are clustered together around same region.

## **Pearson Metric KMeans 2006 (Seed: 150)**

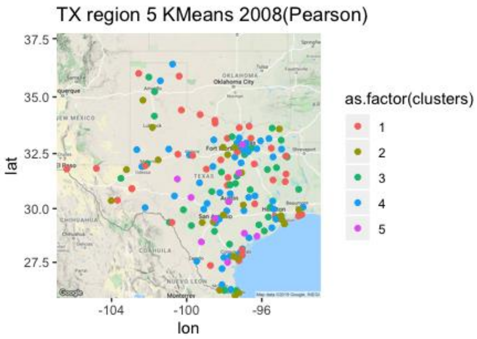


## Pearson Metric KMeans 2007(Seed: 150)



This graph for 2007 compared to graph for 2006 using Pearson metric shows more weather change in stations over one year compared to the case when Euclidean was used.

## Pearson Metric KMeans 2008(Seed: 150)



Again, a quite a few number of stations had change in weather and now belong to other weather clusters both from 2007 and 2006. But overall the changes are not drastic. Pearson with higher seed value definitely showed more diverse weather groups compared to Euclidean Graphs with lower seed value.

To summarize these visualization plots of stations with respect to weather on Texas map show that most stations in Texas have same type of weather. Few slight changes can also be observed for few stations over the year. Changes in weather over 3 years have been ongoing although the results are not drastic for the month of February.