# DATA MINNING - Project 2

**Team 16** –

Akshaya Sharma - 1001707344

Rohit Manhas - 1001657127

**Seed – 20**

**K – 10**

**Years – 2008,2009,2010**

***Overall Status:***

We read the concept of clustering approach to mining unsupervised learningusing the k-means approach.

We went through the dpylr, amap, ggmap package information given to us by Sir in the link as well as R tutorial. Understood the methods of tidyverse, factoextra library, stringr library and the arguments given to the method. Then completed the tasks required for the Project.

**Status:** Completed data cleaning, preprocessed the data using different approaches to get the required data in the proper format. The given weather data set in separate files for the state of Texas for the years 2008 to 2010 (3 years) for analysis.

Then we created three clusters of kmeans, Euclidean, Pearson correlation and Jaccard for the given years 2008, 2009 and 2010.

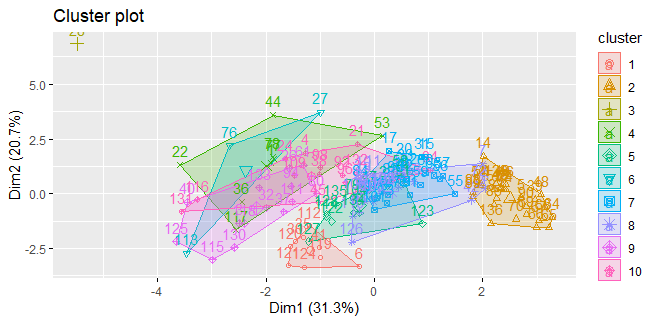
**Steps Followed: -**

1. Started with preprocessing of data by separating our month’s data from all the file.
2. Took average of all of all the days and then means of the whole month for every unique station.
3. Then we plot the clusters based on the data that we got after preprocessing.
4. After Kmeans we calculated the Pearson and Euclidean and also plotted the same.
5. Using Pearson and Euclidean calculations we analyzed the Jaccard Coefficient.

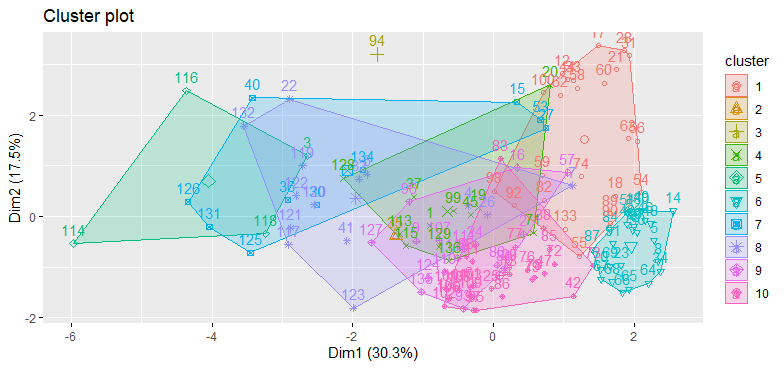
Seed = 20

Kmeans : –

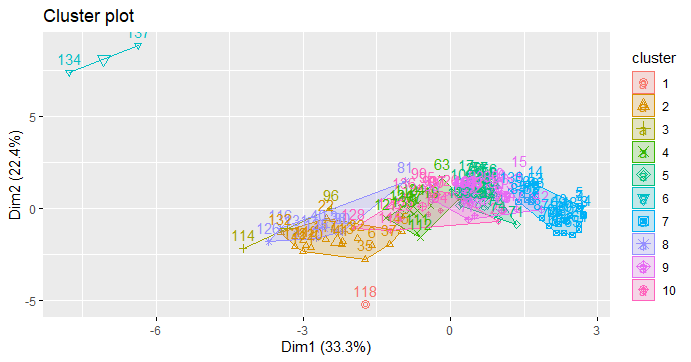
2008



2009



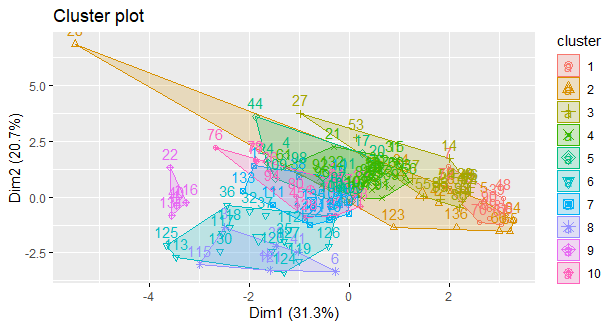
2010



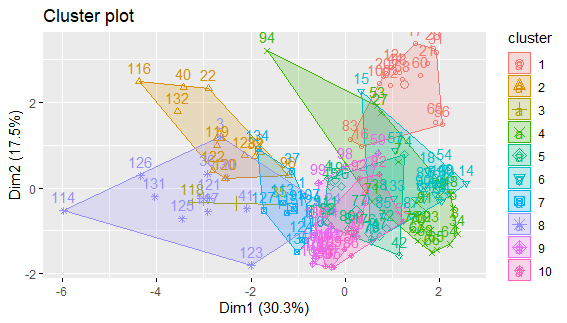
**Analysis** :- Due to change of center in Kmeans and the year change we can analyze that the clusters which were near to each other in 2008, by the time 2009 and 2010 they changed their respective positions.

**Pearson**

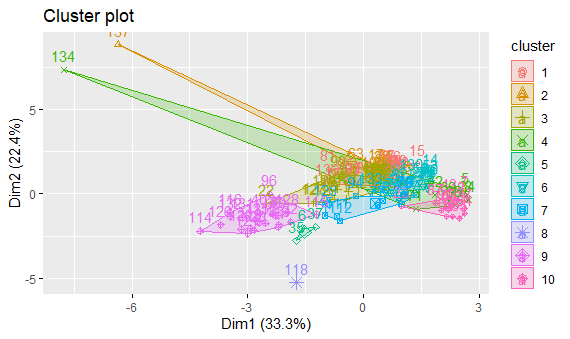
2008



2009



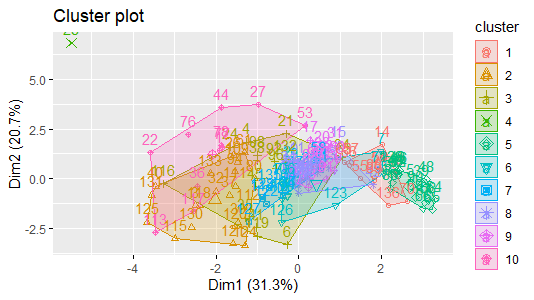
2010



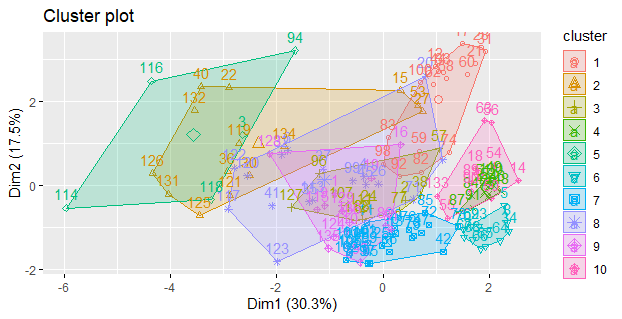
**Analysis** :- Due to change of center and the year change we can analyze that the clusters which were near to each other in 2008, by the time 2009 and 2010 they changed their respective positions.

Euclidean :-

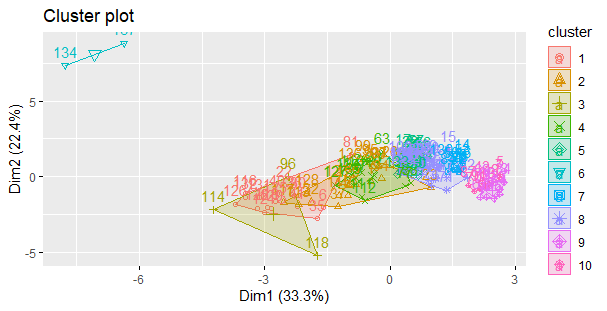
2008



2009



2010

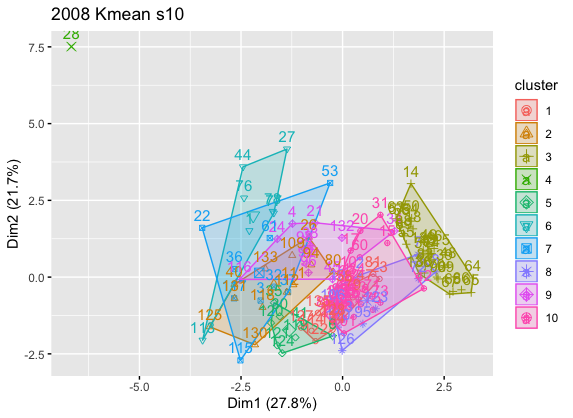


**Analysis** :- Due to change of center and the year change we can analyze that the clusters which were near to each other in 2008, by the time 2009 and 2010 they changed their respective positions

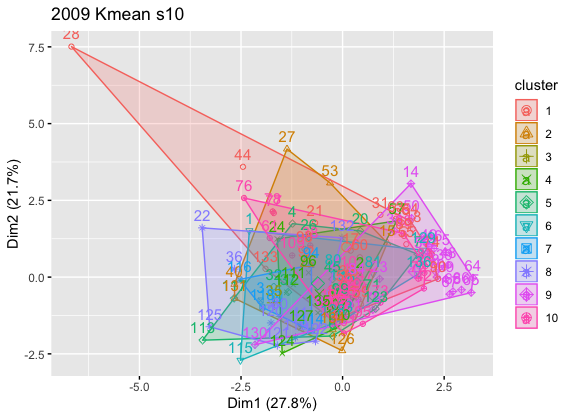
Seed = 10

Kmeans : –

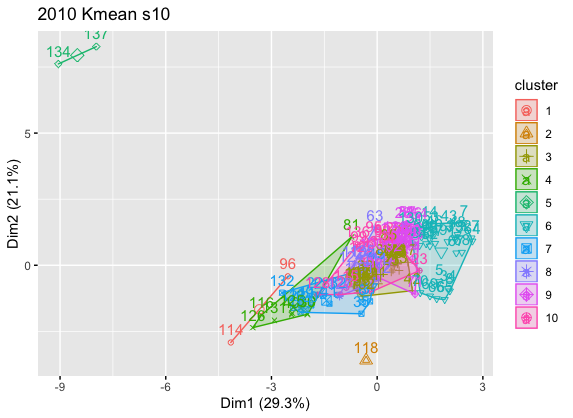
2008



2009



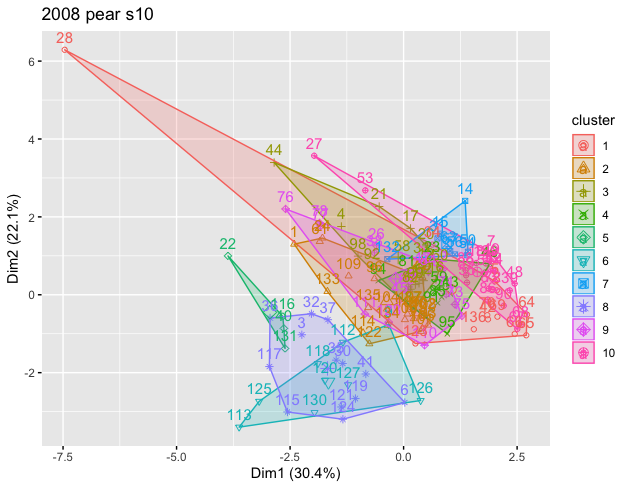
2010



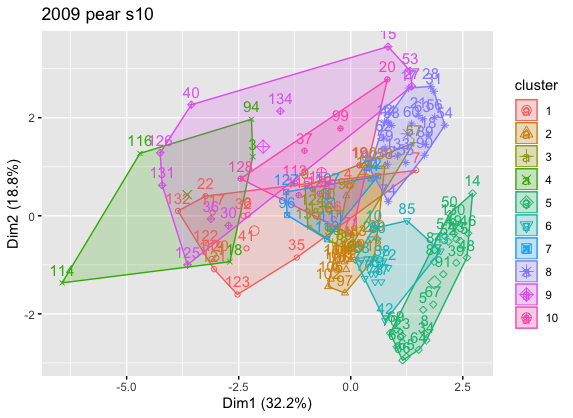
**Analysis** :- Due to change of seed we can analyze that the clusters changed their respective positions

Pearson:-

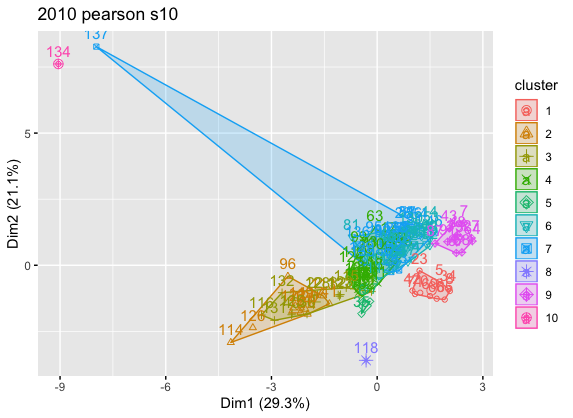
2008



2009



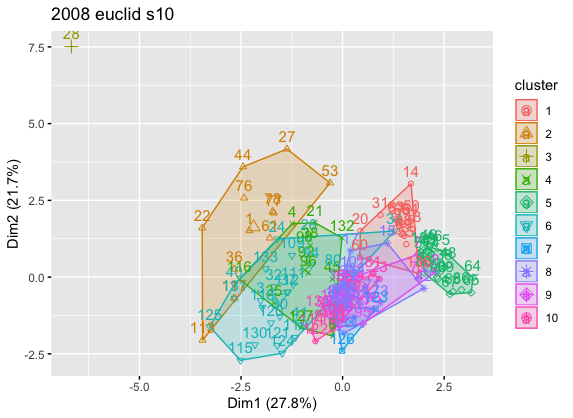
2010



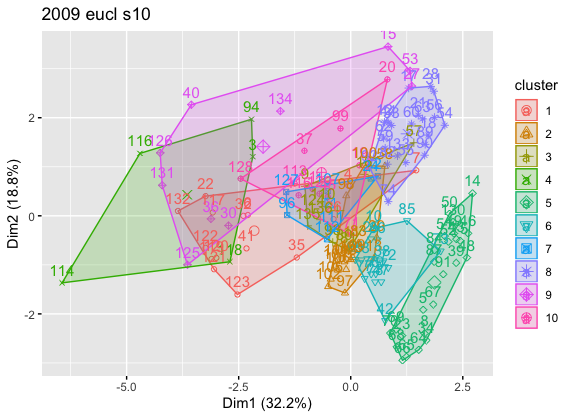
**Analysis** :- Due to change of seed we can analyze that the clusters changed their respective positions

Euclidean:-

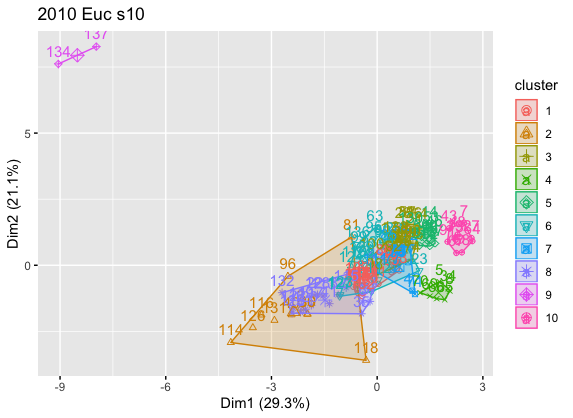
2008



2009



2010



**Analysis** :- Due to change of seed we can analyze that the clusters changed their respective positions

Metric using Jaccard coefficient – Euclidean 2008 – Pearson 2008

[1] 0.3043478

[1] 0.2758621

[1] 0.1153846

[1] 0.1

[1] 0.3214286

[1] 0.1052632

[1] 0.2380952

[1] 0.2105263

[1] 0.3125

[1] 0.125

[1] 0.2108408

Metric using Jaccard coefficient – Euclidean 2009 – Pearson 2009

[1] 0.5714286

[1] 0.3

[1] 0.2173913

[1] 0.3913043

[1] 0.1666667

[1] 0.55

[1] 0.3928571

[1] 0.1666667

[1] 0.3684211

[1] 0.3043478

[1] 0.3429084

Metric using Jaccard coefficient – Euclidean 2010 – Pearson 2010

[1] 0.56

[1] 0.1470588

[1] 0.3333333

[1] 0.2894737

[1] 0.3157895

[1] 0.1111111

[1] 0.625

[1] 0.2941176

[1] 0.3333333

[1] 0.4666667

[1] 0.3475884

Metric using Jaccard coefficient – Euclidean 2008 – Euclidean 2009

[1] 0.5714286

[1] 0.1666667

[1] 0.1851852

[1] 0.05882353

[1] 0.3043478

[1] 0.1363636

[1] 0.1304348

[1] 0.3333333

[1] 0.3333333

[1] 0.1904762

[1] 0.2410393

Metric using Jaccard coefficient – Euclidean 2008 – Pearson 2010

[1] 0.2727273

[1] 0.21875

[1] 0.2222222

[1] 0.0625

[1] 0.2608696

[1] 0.1666667

[1] 0.0952381

[1] 0.325

[1] 0.24

[1] 0.1111111

[1] 0.1975085

Metric using Jaccard coefficient – Euclidean 2009 – Euclidean 2010

[1] 0.2692308

[1] 0.4545455

[1] 0.4736842

[1] 0.4666667

[1] 0.3333333

[1] 0.6153846

[1] 0.5526316

[1] 0.2857143

[1] 0.2692308

[1] 0.3684211

[1] 0.4088843

Metric using Jaccard coefficient – Pearson 2008 – Pearson 2009

[1] 0.3043478

[1] 0.25

[1] 0.3571429

[1] 0.3030303

[1] 0.2173913

[1] 0.2173913

[1] 0.1818182

[1] 0.2857143

[1] 0.25

[1] 0.2592593

[1] 0.2626095

Metric using Jaccard coefficient – Pearson 2009 – Pearson 2010

[1] 0.3333333

[1] 0.3846154

[1] 0.36

[1] 0.2702703

[1] 0.2777778

[1] 0.2758621

[1] 0.15

[1] 0.2272727

[1] 0.1363636

[1] 0.3043478

[1] 0.2719843

|  |
| --- |
| Euclidean 2008 SSE witness  [1] 53.199481 204.517180 287.395975 0.000000 35.129228  [6] 149.796310 13.248531 3.812838 7.008736 1024.523351  Euclidean 2009 SSE witness  [1] 8.936552 319.489045 95.159251 16.650690 1255.456776  [6] 21.982456 5.281652 167.314855 24.993190 13.665331  Euclidean 2010 SSE witness  [1] 61.749193 42.615884 4857.300726 112.961357 5.719119  [6] 3048.729138 6.776695 9.734819 8.620088 9.107915  Euclidean 2008 SSE witness Mean  [1] 1778.632  Euclidean 2009 SSE witness Mean  [1] 1928.93  Euclidean 2010 SSE witness Mean  [1] 8163.315  Pearson 2008 SSE witness  [1] 1.135317e-11 2.062593e-12 1.208757e-11 1.937462e-11 3.237853e-12  [6] 2.773728e-11 5.711527e-12 1.703175e-11 1.545369e-11 7.575820e-12  Pearson 2009 SSE witness  [1] 7.165153e-13 2.519154e-11 3.288088e-11 3.034773e-15 5.324671e-12  [6] 1.221687e-11 3.595266e-12 1.187684e-10 3.321346e-14 3.003445e-11  Pearson 2010 SSE witness  [1] 3.028794e-12 5.822985e-10 2.725911e-11 1.528593e-08 1.645013e-09  [6] 1.783904e-12 6.071133e-11 0.000000e+00 1.202711e-10 8.858464e-13  Pearson 2008 SSE witness Mean  [1] 1.216259e-10  Pearson 2009 SSE witness Mean  [1] 2.287649e-10  Pearson 2010 SSE witness Mean  [1] 1.772718e-08 |
|  |
|  |

***Division of Labor:***

Did the project together. Hold meeting 3 hour every day for the week and finished the project. I did the preprocessing part and Akshaya did clustering and we collectively did project report.

***Problems encountered and how you handled them :-***

1. Removing all impurities from data and after that we preprocessed the large dataset for the very first time, so we took much more time than required in preprocessing.
2. Encountered error –no more for aggregation and then we came to know that we need to remove the missing values.
3. While calculating Jaccard we were getting less values than required, so after calculating manually we came to know we are missing few values and eventually used loop to capture all the values.

**Visualization :-**