Overview

The main purpose of our project is to analyse the clusters produced by applying k-means clustering on weather data. In this project we use four variables namely temperature, dew point, wind speed, station pressure in order to group stations which have similar weather. This is done using the following steps:

* Firstly, we went through the provided problem statement and the requirements to complete this project.
* We analysed the data set provided and removed the attributes which will not be used to cluster the stations.
* In order to calculate the daily and monthly average, we separated the year attribute which contained year, date, day, hours in a single string into different attributes.
* We converted the provided values which are considered as “missing values” into NA.
* We calculated the average of each attributes ignoring the NA values-if present- without ignoring the records, grouping each station.
* If any Nan values are present in the average data frame, it is replaced with the column mean containing Nan.
* We performed the k means for the dataset given after scaling the attributes.
* We performed the above step using Euclidean and Pearson as distance metrics with different seed for each of the metric.
* We plotted the results obtained by K-means – x-axis has clusters and y-axis has each of the four attributes.
* We compared each metric with different seed values using Jaccard similarity coefficient.
* We went through the above process again to compare the data in different years and compare the clusters obtained in from the above process for each year using Jaccard similarity.

File Description

**clustering\_2007.r**: This file contains the code which pre-processes the data set and performs k-means on the 2007 data set.

**clustering\_2008.r**: This file contains the code which pre-processes the data set and performs k-means on the 2008 data set.

**clustering\_2009.r**: This file contains the code which pre-processes the data set and performs k-means on the 2009 data set.

**YeartoYear.r**: This file contains the code to compare jaccard coefficient between clusters obtained from three different years.

Division of Work

Saikrishna : Performed pre-processing for 2007 data set and clustering in 2007 and 2008 data sets.

Deepika : Performed pre-processing for 2008 and 2009 data sets and clustering in 2009 data set.

Problem Faced

* We faced problems while reading data file, faced problem with multiple delimiters.
* While plotting the graph, we encountered problem with scaled data.
* We got Nan values in the attribute average data frame, which caused error in performing K-means.

Choosing Best K Values:

We have plotted the graph between number of clusters and Total SSE and identified the best K value.

Euclidean with seed=10(2007 Data)

Chart, line chart, scatter chart

Description automatically generated

For k=2 to 8, we didn’t get an elbow curve behaviour. Hence, we have chosen K with minimum SSE i.e, K=2

Pearson with seed=10(2007 Data)

Chart, line chart

Description automatically generated

For k=2 to 8, we didn’t get an elbow curve behaviour. Hence, we have chosen K with minimum SSE i.e, K=7

Euclidean with seed=10(2008 Data)

Chart, line chart, scatter chart

Description automatically generated

For k=2 to 8, we didn’t get an elbow curve behaviour. Hence, we have chosen K with minimum SSE i.e, K=2

Pearson with seed=10(2008 Data)

Chart, line chart, scatter chart

Description automatically generated

For k=2 to 8, we didn’t get an elbow curve behaviour. Hence, we have chosen K with minimum SSE i.e, K=2.

Euclidean with seed=10(2009 Data)

Chart, line chart, scatter chart

Description automatically generated

For k=2 to 8, we didn’t get an elbow curve behaviour. Hence, we have chosen K with minimum SSE i.e, K=2.

Pearson with seed=10(2009 Data)

Chart, line chart, scatter chart

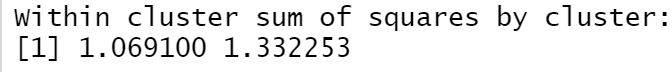
Description automatically generated

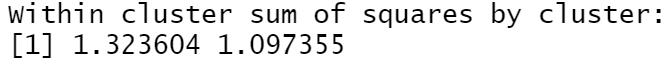
For k=2 to 8, we didn’t get an elbow curve behaviour. But from k=9 it’s linear, so, the elbow point for the above curve is k=8.

Analysis

**Comparing SSE for Euclidean (2007 data)**

SSE for Euclidean with seed (50):- 2.420959 SSE for Euclidean with seed (10):- 2.401353

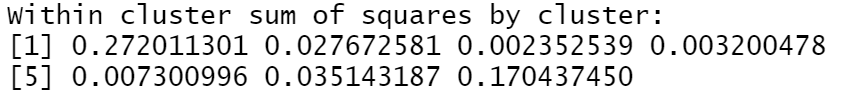


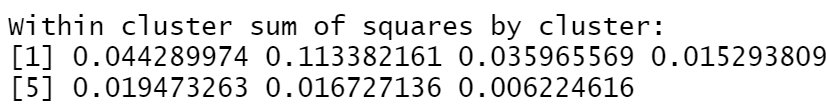


From the above we can see that Euclidean with seed 10 has performed better as the initial random point selection in seed 10 has produced better clusters.

**Comparing SSE for Pearson (2007 data)**

SSE for Pearson with seed (50):- 0.5181185 SSE for Pearson with seed (10):- 0.2513565

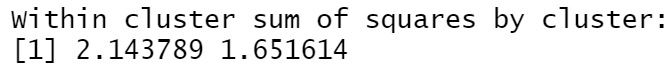


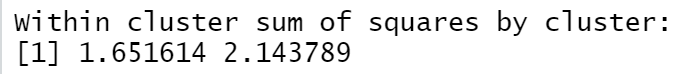


From the above we can see that Pearson with seed 10 has performed better as the initial random point selection in seed 10 has produced better clusters. From the images we can also see that there is larger individual sse for each clusters compared to seed 10.

**Comparing SSE for Euclidean (2008 data)**

SSE for Euclidean with seed (50):- 3.795403 SSE for Euclidean with seed (10):- 3.795403

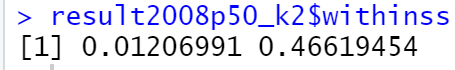


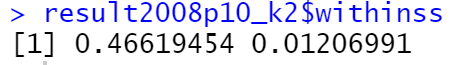


From the above we can see that Pearson with seed 10 and 50 have performed same. From the images we can also see that individual sse for each clusters are same.

**Comparing SSE for Pearson (2008 data)**

SSE for Pearson with seed (50):- 0.4782644 SSE for Pearson with seed (10):- 0.4782644





From the above we can see that Pearson with seed 10 and 50 have performed same. From the images we can also see that individual sse for each clusters are same.

**Comparing SSE for Euclidean (2009 data)**

SSE for Euclideanwith seed (50):- 2.182732 SSE for Euclideanwith seed (10):- 2.182732

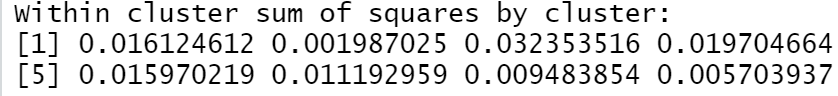


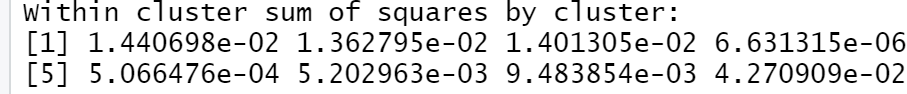


From the above we can see that Pearson with seed 10 and 50 have performed same. From the images we can also see that individual sse for each clusters are same.

**Comparing SSE for Pearson (2009 data)**

SSE for Pearson with seed (50):- 0.1125208 SSE for Pearson with seed (10):- 0.09995716





From the above we can see that Pearson with seed 10 has performed better as the initial random point selection in seed 10 has produced better clusters.

**Comparing Clusters Euclidean and Pearson using Jaccard coefficient (2007 data)**

Jaccard similarity between e10 and p10: 0.9132005 between c1 and c1

Jaccard similarity between e10 and p50: 0.866499 between c1 and c1

Jaccard similarity between e50 and p50: 0

Jaccard similarity between e50 and p10: 0.8806957 between c1 and c1

From the above we can see that Euclidean with seed 10 and pearson with seed 10 has maximum similarity between cluster 1 in Euclidean and cluster 1 in pearson, pearson with seed 10 and Euclidean with seed 50 has maximum similarity between clusters 1 in seed 10 and 1 in seed 50 with value of 0.88. Euclidean 50 and pearson 50 does not have any similarity.

**Comparing Clusters Euclidean and Pearson using Jaccard coefficient (2008 data)**

Jaccard similarity between e10 and p10: 0. 540913 between c1 and c1

Jaccard similarity between e10 and p50: 0. 540913 between c1 and c1

Jaccard similarity between e50 and p50: 0. 540913 between c1 and c1

Jaccard similarity between e50 and p10: 0.540913 between c1 and c1

From the above we can see that all combinations have same maximum similarity between clusters 1 in both of them.

**Comparing Clusters Euclidean and Pearson using Jaccard coefficient (2009 data)**

Jaccard similarity between e10 and p10: 0.6335677 between c1 and c8

Jaccard similarity between e10 and p50: 0.612069 between c1 and c3

Jaccard similarity between e50 and p50: 0

Jaccard similarity between e50 and p10: 0.6335677 between c1 and c8

From the above we can see that Euclidean with seed 10 and pearson with seed 10, Euclidean with seed 50 and pearson with seed 10 have maximum similarity between clusters 1 in Euclidean and cluster 8 in pearson, Euclidean with seed 10 and pearson with seed 50 has the similarity of : 0.612069 as between cluster 1 in Euclidean and cluster 3 in pearson. Euclidean 50 and pearson 50 does not have any similarity.

**Comparing different metrics in different years using Jaccard coefficient**

Jaccard similarity between Euclidean 2007 and 2008:-

Seed value 10 in both:- 0.8954327 between 1 and 1 seed value 50 in both:- 0.8954327 between 1 and 1

Jaccard similarity between Pearson 2007 and 2008:-

Seed value 10 in both:- 0.5949183 between 1 and 1 seed value 50 in both:- 0.5397709 between 2 and 1

Jaccard similarity between Euclidean 2008 and 2009:-

Seed value 10 in both:- 0.9455415 between 1 and 1 seed value 50 in both 0.9455415 between 1 and 1

Jaccard similarity between Pearson 2008 and 2009:-

Seed value 10 in both:- 0.5479545 between 1 and 3 seed value 50 in both:- 0.5214838 between 1 and 1

Jaccard similarity between Euclidean 2007 and 2009:-

Seed value 10 in both 0.9442421 between 1 and 1 seed value 50 in both 0.9442421 between 1 and 1

Jaccard similarity between Pearson 2007 and 2009:-

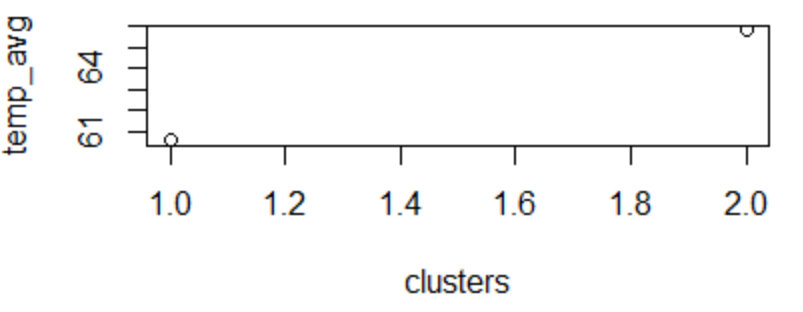
Seed value 10 in both:- 0.8962641 between 3 and 4 seed value 50 in both:- 0.8508751 between 4 and 4

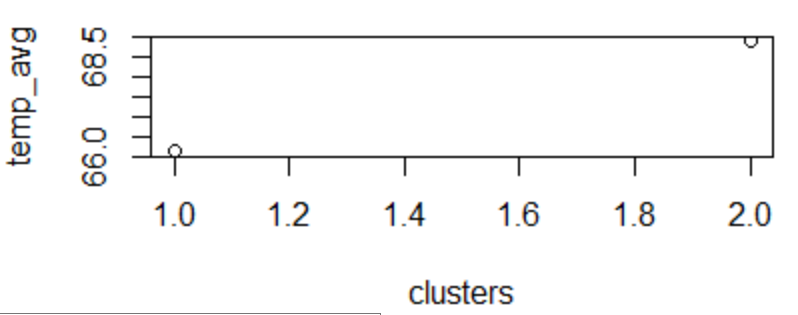
From the above values we can see that for Euclidean, weathers are almost similar in all years. Now Coming to Pearson weathers of 2007 and 2009 are more similar and the weathers of 2007&2008, 2008&2009 are dissimilar.

**Visualizing clusters(Seed=10)**

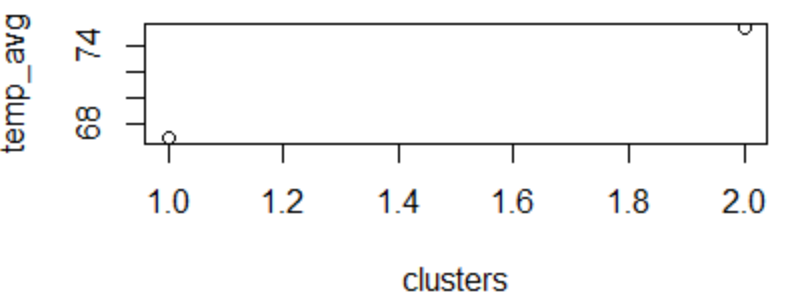
Temperature vs Clusters:-

2007:-

2008:-

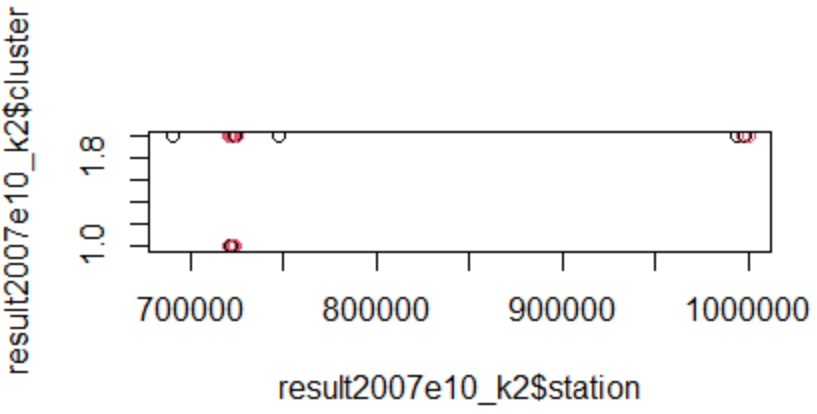


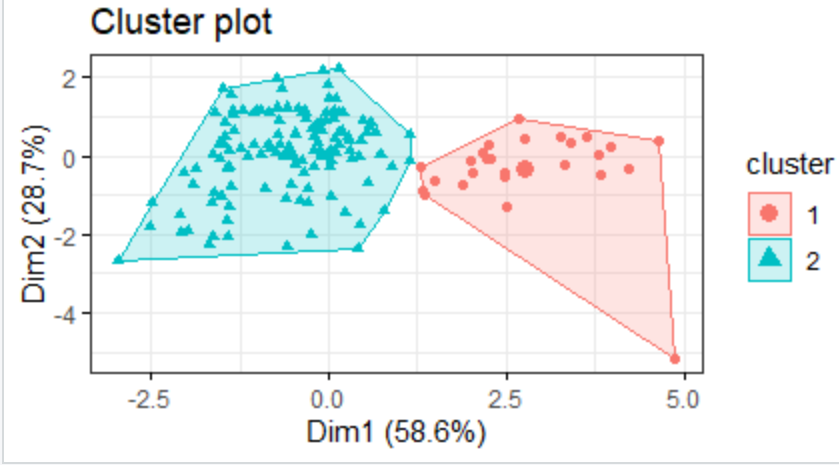
2009:-



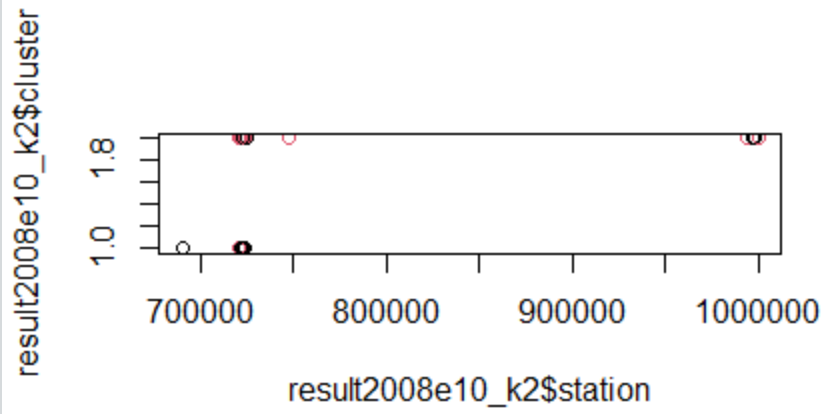
Clusters vs Stations:-

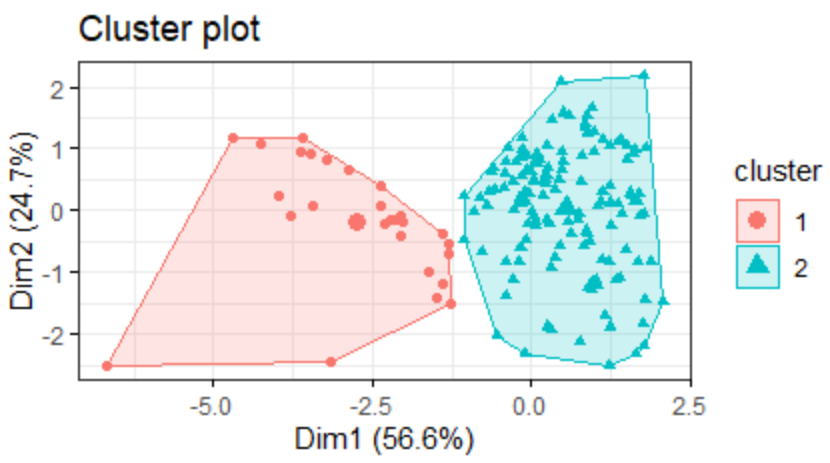
2007:





2008:-





2009:-

