**PROJECT 2 REPORT**

**(years= (2007,2008,2009), month=June (6))**

**Submitted by:**

Team No. 6

Amrutha Kenchanagowdra -1001861876

Ajay Shivarudrappa – 1001720944

**PROJECT OVERVIEW**

* The weather dataset for the state of Texas for the years 2007-2009 has been provided to perform the analysis using k-means clustering.
* We have been asked to perform weather data analysis for the month of June for the years 2007,2008 and 2009.
* The datafiles contained weather data recordings for every hour of the day.
* The subset of the dataset was taken to get the specified columns as per the problem statement such as: **STN,yearModa\_hr,Temp,DewP,STP,WDSP and the columns were given a proper name such as: "station","date","temp","dewpoint","stationpressure","windspeed"**
* The datafile for each year was pre-processed to exactly obtain only the records for the month of June to perform the analysis.
* Group by was done on data set to get the station wise groups of for the **temp","dewpoint","stationpressure","windspeed”** attributes
* Stations present in station.csv were read and compared with the data and only common stations were taken as subset.
* Monthly average for each station was calculated with respect to all the 4 attributes specified so that each row in the pre-processed dataset represented monthly average for each station w.r.t each attribute.
* This pre-processed dataset was given as input to perform the cluster analysis.
* For each year ( June month), clustering analysis was done using **kmeans (package used: amap**) method for metrics such as Euclidean and Pearson and the plots were made to visualize the clusters formed for different k values from 2 to 8.
* Best k value was found based on elbow method and the similarity between the clusters were found using **Jaccard** (package used: **clusteval** ) and the comparison between the years for the same was done.
* SSE was computed for both pearson and Euclidean metric and the comparision was done.
* 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.
* For reading the data, we have converted the .g file into .csv file and have used it as such.
* For calculating the monthly averages for each station we have replaced 999 values with 0 and calculated the mean of each parameter.

**Cluster analysis for year 2007 data**

After preprocessing the clean dataset obtained contained the below number of stations and each of them included number of records as show below:

Before preprocessing: After Preprocessing:

A picture containing table

Description automatically generatedTable

Description automatically generated

**Case1:** Kmeans clustering was done for different k values between 2 to 8 and the clusters were plot. Randomly generate centroids generation was taken care by kmeans method with respect to the given k value at each step.

By keeping the same value for **seed: 30** we have tried plotting kmeans clusters.

**Euclidean metric**

1. For **k =2** below results were obtained:

K-means clustering with 2 clusters of sizes 45, 111 were obtained with the cluster means as show below.

Text

Description automatically generatedChart, scatter chart

Description automatically generated

1. For **k =3** below results were obtained:

K-means clustering with 3 clusters of sizes 8, 111, 37 were obtained with the cluster means as show below.

Text

Description automatically generatedChart, scatter chart

Description automatically generated

1. For **k = 7** below results were obtained:

K-means clustering with 7 clusters of sizes 22, 30, 32, 19, 45, 3, 5 were obtained with the cluster means as show below.

Cluster means:

A picture containing table

Description automatically generatedChart, scatter chart

Description automatically generated

We have tried kmeans clustering for k = 4, 5 as well and based on the analysis of the plots above, we can see that the clusters for Euclidean method are associated properly for k=3 and are overlapping in case of k=7 and association is not well in other cases and thus we can consider that the **best k for Euclidean method is k = 3**

**Pearson metric**

Kmeans clustering was done for pearson metric for different values of k between 2 to 8 and thus below results are obtained.

1. For **k =2** below results were obtained:

K-means clustering with 2 clusters of sizes 47, 109 were obtained

A picture containing text

Description automatically generatedChart, scatter chart

Description automatically generated

1. For **k =3 and k=7** respective below results were obtained:

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

We have tried kmeans clustering for k = 4, 5 as well and based on the analysis of the plots above, we can see that the clusters for Pearson method are associated properly for k=2 and are overlapping in case of k=7 and association is not well in other cases example in case of k=3 the within cluster spacing is more and thus we can consider that the **best k for Pearson method is k = 2**

**Case2:**

For the best k chosen, Random seed value (120) was taken and the kmeans clustering for both Euclidean and Pearson methods were computed, and the below results were obtained

**Euclidean metric:** K=3, seed = 120 **Pearson metric:** K=2, seed = 120

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Association remained same for both the best k values for Euclidean and person with the change in random seed value

**Case3 and case4:** **SSE comparison for the best k chosen**

The basic idea behind the cluster partitioning methods such as kmeans clustering is to define clusters such that the total intra cluster variation know as SSE (which is determined by the cluster parameter withinss: total within cluster sum of squares.) is minimized.

**Euclidean and Pearson metrics:**

SSE values for each of the k values between 2 to 8 were computed and the elbow curve was plot for the same. Elbow curve also helps us to verify the best k value which we have chosen in the previous step.

Chart, line chart

Description automatically generated Table

Description automatically generated

**Case5: Cluster similarity comparison w.r.t Jaccard coefficient**

We can see from the below table that the Pearson vs Euclidean(k=2) for the both the seed values have slightly higher Jaccard/almost simliar than the other cases, which shows that the weather remained constant w.r.t both metrics for the June month.

But when k =7 the similarity is drastically varying which is not true as a matter of fact as we are measuring the jaccard for same month/year, which means that the clustering with respect to k=7 is not good.

**Table

Description automatically generated**

**Cluster analysis for year 2008 data**

Similar procedure has been followed for year 2008 as that of 2007 and the preprocessing was done to get the June month data

**Case1:** Kmeans clustering was done for different k values between 2 to 8 and the clusters were plot. Randomly generate centroids generation was taken care by kmeans method with respect to the given k value at each step. After plotting the different k means clusters for k =2 to 8, we have analyzed the cluster association and found the **best k for Euclidean and Pearson measure => k= 3**

**Euclidean Pearson**

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

**Case2:**

For the best k chosen, Random seed value (120) was taken and the kmeans clustering for both Euclidean and Pearson methods were computed, and the below results were obtained

**Euclidean metric:** K=3, seed = 120 **Pearson metric:** K=3, seed = 120

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Association remained same for both the best k values for Euclidean and person with the change in random seed value

**Case3 and case4:** **SSE comparison for the best k chosen**

**Euclidean and Pearson metrics:**

SSE values for each of the k values between 2 to 8 were computed and the elbow curve was plot for the same. Elbow curve also helps us to verify the best k value which we have chosen in the previous step.

**Chart, line chart

Description automatically generated** **Table

Description automatically generated**

**Case5: Cluster similarity comparison w.r.t Jaccard coefficient**

We can see from the below table that the Pearson vs Euclidean(k=3) for the both the seed values are simliar, which shows that the weather remained constant w.r.t both metrics for the June month.

But when k =7 the similarity is drastically varying which is not true as a matter of fact as we are measuring the jaccard for same month/year, which means that the clustering with respect to k=7 is not good.

Table

Description automatically generated

**Cluster analysis for year 2009 data**

Similar procedure has been followed for year 2009 as that of 2007 and the preprocessing was done to get the June month data

**Case1:** Kmeans clustering was done for different k values between 2 to 8 and the clusters were plot. Randomly generate centroids generation was taken care by kmeans method with respect to the given k value at each step. After plotting the different k means clusters for k =2 to 8, we have analyzed the cluster association and found the **best k for Euclidean and Pearson measure => k= 2**

**Euclidean Pearson**

**Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated**

We can infer from the above graphs that the Euclidean has done better association than the Pearson.

**Case2:**

For the best k chosen, Random seed value (120) was taken and the kmeans clustering for both Euclidean and Pearson methods were computed, and the below results were obtained

**Euclidean metric:** K=2, seed = 120 **Pearson metric:** K=2, seed = 120

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Association remained almost same for both the best k values for Euclidean and person with the change in random seed value

**Case3 and case4:** **SSE comparison for the best k chosen**

**Euclidean and Pearson metrics:**

SSE values for each of the k values between 2 to 8 were computed and the elbow curve was plot for the same. Elbow curve also helps us to verify the best k value which we have chosen in the previous step. Table shows the SSE values computed for each of the bestk w.r.t different seed values.

Chart, line chart, histogram

Description automatically generatedTable

Description automatically generated

**Case5: Cluster similarity comparison w.r.t Jaccard coefficient**

We can see from the below table that the Pearson vs Euclidean(k=2) for the both the seed values are simliar, which shows that the weather remained constant w.r.t both metrics for the June month.

But when k =7 the similarity is drastically varying which is not true as a matter of fact as we are measuring the jaccard for same month/year, which means that the clustering with respect to k=7 is not good.

Table

Description automatically generated

**Case6:**

**Comparison between 2007 vs 2008**

The data for year 2007 and 2008 were compared for the common stations to be found and the subset was taken to select the data records with common stations.

Kmeans clusters were computed with the best k for each year and the cluster comparison was done using Jaccard coefficient and the below results were obtained

**A picture containing table

Description automatically generated**

We can see that the Euclidean metric did not capture the similarity between the weathers accurately whereas the Pearson metric has captured it well when the best k value for Pearson was computed for both the years

**Comparison between 2008 vs 2009**

Similarly, here too, the data which has common station numbers between both the years was subset and taken for comparisons between years 2008 and 2009 and the kmeans clusters for best k values were computed and below results were obtained.

Table

Description automatically generated

From the above table we can see that the similarity between the weathers is low which means that the weather has changed over time from year 2008 to year 2009

**Comparison between 2007 vs 2009**

The data which has common station numbers between both the years was subset and taken for comparisons between years 2007 and 2009 and the kmeans clusters for best k values were computed and below results were obtained.

Table

Description automatically generated

From the above table we can see that the similarity between the weathers is low which means that the weather has changed over time from year 2007 to year 2009

**Visualization: for year 2007 and 2008**

Map

Description automatically generated

**FILE NAMES**

**src directory:**

june2007.R: Preprocessing of June 2007 data and clustering is handled here.

june2008.R: Preprocessing of June 2008 data and clustering is handled here.

june2009.R: Preprocessing of June 2009 data and clustering is handled here.

Year\_comparison: comparision between years 2007 vs 2008 vs 2009 is handled here.

map.R: plotting weather stations for each year on Texas map.

**DATAFiles directory:**

.csv files: datafiles converted and used for project

**Vizualization:**

Various visualization plots are present here.

**CHALLENGES**

1.Interpreting the meaning of Jaccard values and its usability and fluctuation to different datasets.

2.kmeans parameters understanding and plotting of clusters and plotting the clusters on Texas map

3.SSE computation and elbow method interpretation

**Division of Labor**

Amrutha:

I have handled the Preprocessing of all the data files, handled clustering of year 2007 data.

Worked on the report preparation.

Ajay:

Handled clustering of year 2008 and 2009 data and worked on the map generation for comparison part.