Project II: Clustering

CSE 5334: Data Mining
Name: Hamza Reza Pavel
Student ID: 1001741797
Group: 2

Introduction

In this project, we are going to use the K-means clustering algorithm on provided weather dataset to understand and analyze the unsupervised approach to mining. The K-means clustering algorithm organizes *n* observations into *k* clusters based on distance-based similarity measures. The most common distance measure used in the K-means algorithm is the Euclidean distance measure. But other distance measures such as Manhattan, Cosine, Correlation, and Hamming distance measures.

The provided dataset is the weather data of the state of Texas from 2006 to 2010 divided into files by year, collected from around 200 weather stations. For this project, our group is going to use a subset of the entire dataset consisting of weather station data of the months of February and August for the years 2006, 2008, and 2009.

Analysis of the Dataset

The dataset consists of hourly weather data collected from the stations over a span of 5 years. The dataset has a total of 19 columns and 5,839,608 rows. After the preprocessing steps, we split our dataset into 6 subsets, two for each assigned year representing the months. Hence, we get 6 subsets of a dataset representing the data collected from February 2006, August 2006, February 2008, August 2008, February 2009, and August 2009. Also, the K-means algorithm is greatly affected by the dimensionality of the dataset. Our main dataset had 19 different columns. We remove the unnecessary columns from our dataset and only keep the following 5 columns:

Table 1: Important fields of the dataset.

STN	Station number for the location. Unique identifier for each of the stations.
YearMoDa_H	The timestamp when a particular entry was collected.
Temp	Mean temperature of that hour in degrees Fahrenheit to tenths.
DewP	Mean dew point in that hour in degrees Fahrenheit to tenths.
STP	Mean station pressure for that hour in millibars to tenths.
WDSP	Mean wind speed for the hour in knots to tenths.

Once the dataset is split into the expected 6 subsets based on the YearMoDa_H time stamp, we get rid of the timestamps column too. We only use Temp, DewP. STP, and WDSP in our K-Means clustering algorithm to measure the distance between two points and ignore the STN field as this is an identifier for the stations.

Data Preprocessing

This project required a significant amount of time in data preprocessing steps as the K-means clustering algorithm is greatly affected by noise, outliers, and dimensionality of the data. For data preprocessing, the following steps were taken:

i. First, we load the files containing the weather data for the year of 2006, 2008, and 2009 and bind them into a single data-frame.
ii. From this data-frame, we get rid of the unnecessary columns. Our initial dataframe had a total of 19 columns, after this step, we are left with 6 columns.
iii. In this step, we parse the *YearMoDa_H* column and create a DateTime object from the entries of this column. We replace the strings of this column with the obtained DateTime objects.
iv. In this step, we create 6 subsets of the data using the DateTime column. Our subset of data is for the months of February and August for the years 2006, 2008, and 2009. Each data subset represents either of the two months from one of the three years.
v. For each of the data subset, we process them even further using the following steps:
    a. There are multiple entries for each of the weather stations. To cluster the weather station, each station should have one entry. We do this by taking the mean of Temp, DewP, STP, and WDSP of each month. While taking the mean, we ignore the missing values which are represented by 999.9 or 9999.9.
    b. The data of each field has a different unit and range. A field with higher values might have more impact when calculating the distance measures. To circumvent this problem, we scale the data-frame using the default scaling function of R.
vi. We write the preprocessed data-frames in files for further using them in the analysis.

The following table shows a summary statistics of our 6 data subsets before scaling:

Table 2: Summary statistics of the data subsets.

	Temp	DewP	STP	WDSP
February, 2006 Shape: 119 X 5	Min. : 40.14 1st Qu.: 51.09 Median: 55.92 Mean: 58.39 3rd Qu.: 63.96 Max. :108.02	Min. :11.45 1st Qu.:29.93 Median:34.94 Mean:34.29 3rd Qu.:40.70 Max. :51.69	Min. : 834.6 1st Qu.: 970.6 Median: 995.9 Mean: 981.4 3rd Qu.:1008.4 Max. :1019.8	Min. : 4.098 1st Qu.: 7.047 Median : .004 Mean 8.045 3rd Qu.: 8.823 Max. :18.518
August, 206 Shape: 117 X 5	Min. :73.04 1st Qu.:84.03 Median :86.13 Mean :85.39 3rd Qu.:87.63 Max. :90.65	Min. :58.73 1st Qu.:64.66 Median :66.51 Mean :67.33 3rd Qu.:70.64 Max. :76.51	Min. : 836.7 1st Qu.: 958.2 Median : 990.1 Mean : 977.0 3rd Qu.:1003.7 Max. :1014.8	Min. : 1.524 1st Qu.: 5.184 Median : 6.687 Mean : 6.553 3rd Qu.: 7.787 Max. :12.165
February, 2008 Shape: 136 X 5	Min. : 45.47 1st Qu.: 54.08 Median : 59.47 Mean : 61.44 3rd Qu.: 65.94 Max. :103.59	Min. :14.16 1st Qu.:32.41 Median :38.26 Mean :37.94 3rd Qu.:46.39 Max. :58.25	Min. : 832.0 1st Qu.: 966.2 Median : 992.9 Mean : 978.6 3rd Qu.:1005.4 Max. :1018.7	Min. : 3.957 1st Qu.: 7.282 Median : 8.420 Mean : 8.400 3rd Qu.: 9.315 Max. :21.866
August, 2008 Shape:135 X 5	Min. :71.26 1st Qu.:81.49 Median :83.02 Mean :82.58 3rd Qu.:84.3 Max. :92.22	Min. :53.54 1st Qu.:65.10 Median :68.59 Mean :67.96 3rd Qu.:71.68 Max. :78.24	Min. : 835.0 1st Qu.: 964.9 Median : 989.1 Mean : 975.7 3rd Qu.:1001.0 Max. :1012.9	Min. : 1.418 1st Qu.: 4.682 Median : 5.668 Mean : 5.599 3rd Qu.: 6.434 Max. :11.852
February, 2009	Min. :43.07	Min. : 9.11	Min. : 834.6	Min. : 5.172

Shape:136 X 5	1st Qu.:56.20 Median :58.80 Mean :59.95 3rd Qu.:63.01 Max. :94.91	1st Qu.:32.71 Median :38.42 Mean :37.53 3rd Qu.:43.54 Max. :58.60	1st Qu.: 968.6 Median : 995.8 Mean : 981.3 3rd Qu.:1008.0 Max. :1021.3	1st Qu.: 7.681 Median : 8.888 Mean : 8.873 3rd Qu.:10.009 Max. :18.390
August, 2009 Shape:137 X 5	Min. :73.96 1st Qu.:83.20 Median :84.97 Mean :84.72 3rd Qu.:86.76 Max. :92.71	Min. :50.25 1st Qu.:62.44 Median :66.58 Mean :66.14 3rd Qu.:70.50 Max. :77.64	Min. : 837.2 1st Qu.: 966.6 Median : 991.9 Mean : 978.3 3rd Qu.:1003.4 Max. :1015.5	Min. : 2.232 1st Qu.: 5.297 Median : 6.848 Mean : 6.633 3rd Qu.: 7.961 Max. :11.348

From a quick glance over the table, we notice that minimum temperature is lowest during February among the two months. This is expected as February is the end of winter. What is interesting is that maximum temperature for each year also lies within February, not August. The Dew point is pretty consistent across the years. There is no discernible change in STP over the years or by month. The wind speed is also consistent over the years with higher min, max, and mean in February of the same year than August. There is not much deviation in windspeed across years.

Data Collection and Analysis Tasks

I. & II. Determining K value from Elbow Curve

The SSE or Sum of Squared Error is the sum of squared distance between the centroid and each point in the cluster. Our goal for clustering is to cluster with a value of K that provides sufficiently low value of total SSE. If we plot SSE against the K value, we will notice that after a certain K value, the change in SSE is not that significant. Due to the elbow-like shape of this graph, it's called the Elbow graph and is a good measure for determining a good K value. As we have two data subset for each year, we only use one to the determine K value for that year.

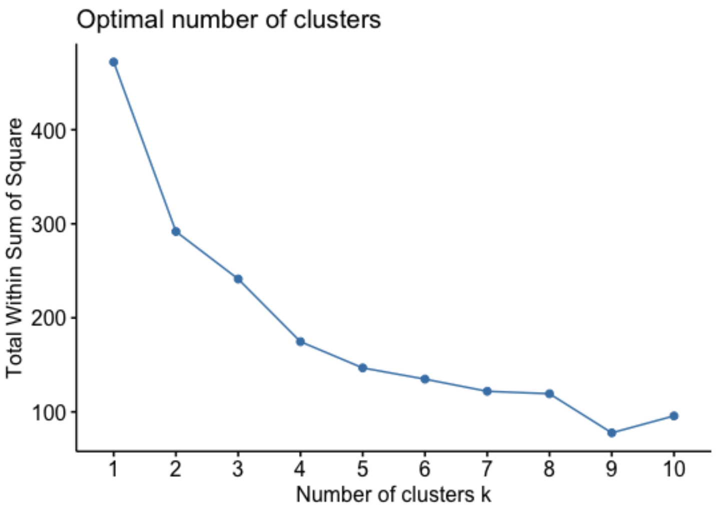


Fig. 1: Elbow Graph for Y 2006.

Above plot in Fig. 1 is drawn using the subset of data consisting of February 2006. As we can observe, there is not much change the in SSE after K value of 6. So we can say, 6 is a good K value for the data of Year 2006.

For Year 2008, we can see from Fig. 2 that a good K value could be 6 as there is no significant change in SSE after K value of 6.

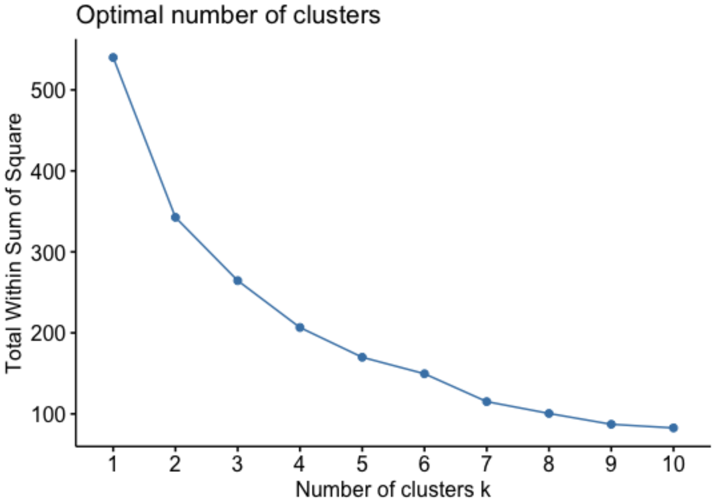


Fig. 2: Elbow Graph for Y 2008.

Similarly, for the year 2009, if we draw the elbow graph, we can observe that a good K value is 7.

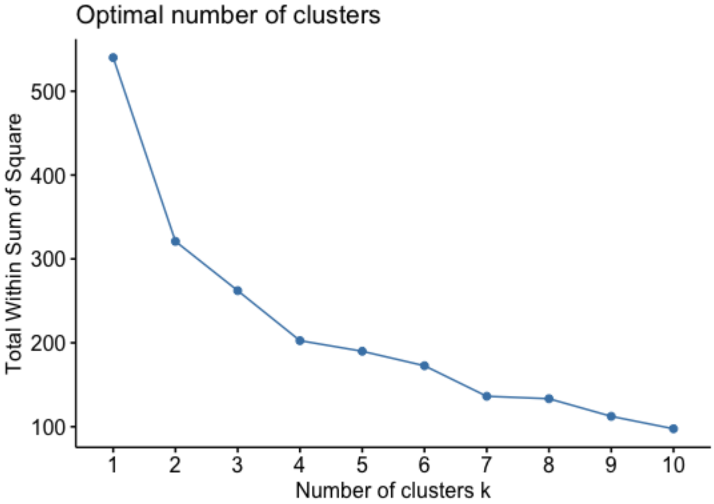


Fig. 3: Elbow graph for Y 2009.

The following table summarizes the K value for clustK-valueyear:

Year	K value
2006	6
2008	6
2009	7

III. & IV. Analysis of Cluster For different Seed Value

Different initial partitions can result in different centroid and clusters in K-means clustering. To analyze the impact of different starting points, we are going to run K-means clustering on each years data with two different seed values and compare the results.

For Y2006: First, we run K-the means on data subset of February 2006 using the seed value 06271993. We repeat the process again for the same dataset using the seed value 39917260. The collected relevant information is in Table 4.

Table 4: Cluster specific information for Feb Y2006 data		
Seed Value	06271993	39917260
Cluster wise SSE	17.27244,34.29947,20.31742,27.65861, 18.59899,0.00000	27.65861,0.00000,34.29947,19.44582 18.59899,18.15260
Total SSE	118.14	118.15
Clusters of Size	34, 19, 32, 11, 22, 1	11, 1, 19, 38, 22, 28

As we observe in the collected data in Table 4, the total SSE and cluster wise SSE is almost same. The order of the cluster is different. The size of the clusters are slightly different where the biggest cluster size is 34 for one seed value and 38 for the other seed value.

The data for August 2006 can be observed in Table 5. There is no difference in total SSE and cluster-wise SSE for this subset of data. Though the order of the clusters are different, their sizes are same.

Table 5: Cluster specific information for Aug Y2006 data		
Seed Value	06271993	39917260
Cluster wise SSE	14.613703,20.367042,15.192135 ,9.673499, 8.905610,19.173264	15.192135,14.613703,19.173264, 8.905610,20.367042,9.673499
Total SSE	87.92	87.92
Clusters of Size	15, 22, 25, 9, 28, 18	15, 28, 18, 9, 22, 25

The Table 6 contains the K-means output data for subset of data for February, 2008. Same as previous subset, this subset has same total SSE for both seed values. The cluster size and cluster wise SSE seems to be similar too.

Table 6: Cluster specific information for Feb Y2008 data		
Seed Value	06271993	39917260
Cluster wise SSE	0.00000,34.91532,30.99770,14.18479, 23.17364,20.97171	21.15609,34.91532,32.04344,12.94113 0.00000,23.17364
Total SSE	124.22	124.22
Clusters of Size	46, 35, 21, 20, 1, 13	46, 1, 20, 13, 35, 21

The Table 7 contains output of K-means for August 2008, the Table 8 contains output of K-means for February 2009, and finally the Table 9 contains output for August 2009.

Table 7: Cluster specific information for Aug Y2008 data		
Seed Value	06271993	39917260
Cluster wise SSE	9.085464,29.326249,25.569372,16.114342 ,18.183341,11.463821	18.183341,29.326249,11.463821,25.569372 ,9.085464,16.114342
Total SSE	109.74	109.74
Clusters of Size	39, 22, 13, 37, 6, 18	6, 22, 18, 39, 37, 13

There are no significant difference in total SSE and cluster-wise SSE for these subsets of the dataset. For all 3 subset, the cluster size for different seed value seems to be similar too.

Table 8: Cluster specific information for Feb Y2009 data		
Seed Value	06271993	39917260

Cluster wise SSE	29.26451,12.49987,18.79686 0.00000,23.38123,22.65392,16.03742	22.65392,23.38123,16.03742 ,29.26451,0.00000,12.49987, 18.79686
Total SSE	122.63	122.63
Clusters of Size	6, 26, 31, 26, 1, 20, 26	26, 1, 31, 26, 26, 6, 20

Even though K-means algorithm is greatly affected by initial points of the clusters, the output for our subset of datasets seems to be consistent except for the February2006 subset. The reason could be we used a high random sample value of 50 when running the K-means algorithm. This means the K-means function took 50 different set of initial starting points and choose the best one.

Table 9: Cluster specific information for Aug Y2009 data

Seed Value	06271993	39917260
Cluster wise SSE	16.283990,18.304966,10.857419 ,9.002647,21.946292,22.399694, 10.085684	18.304966,10.085684,16.283990 22.399694,21.946292,9.002647, 10.857419
Total SSE	108.88	108.88
Clusters of Size	20, 20, 15, 31, 18, 23, 10	20, 10, 15, 20, 23, 18, 31

V. Analysis of Change in Cluster by year based on Jaccard Coefficient

For analysis of similarity of the clusters, we use the Jaccard Coefficient. The Jaccard coefficient or Jaccard index measures the similarity between two sets and it is the ratio of intersection of the set divided by the union of the sets. In our project, we are clustering the weather stations based on the weather data. We use the *clusteval* library of R to calculate the similarity between two clusters. We create a 6X6 matrix where each cell of the matrix is a Jaccard index between clusters of two subset of data. We then compare year-wise similarity using this matrix.

Table 10: Jaccard Similarity matrix for the datasets

	Y2006_Feb	Y2006_Aug	Y2008_Feb	Y2008_Aug	Y2009_Feb	Y2009_Aug
Y2006_Feb	1	0.143	0.13	0.12	0.1147	0.09
Y2006_Aug	0.14	1	0.134	0.11	0.108	0.107
Y2008_Feb	0.1327247	0.1347741	1	0.3139148	0.2740943	0.1906239
Y2008_Aug	0.12	0.11	0.3139148	1	0.2085048	0.1952575
Y2009_Feb	0.1147	0.108	0.2740943	0.2085048	1	0.2624529
Y2009_Aug	0.09	0.107	0.1906239	0.1952575	0.2624529	1

a. (Y1, Y2): In our case, we can consider 2006 as Y1 and 2008 as Y2. As we can see from the Jaccard matrix, the score for Jaccard Coefficient for month of February for Y1 and Y2 is 0.13 and similarity for the month of August for these two years is 0.11. This entails the weather between February of these two years are more similar than August. Interesting thing is, for Y1 month of February had more similarity with August of Y1 than similarity between February of Y1 and Y2. Similarly, August of Y1 is more similar to February of Y2 rather than August of Y2.

b. (Y2, Y3): In our case 2008 is Y2 and 2009 is Y3. Similarity between month of February and August of these two years are 0.27 and 0.19. The similarity between same months increased when compared to same months between Y1 and Y2. One reason could be, as years progressed more data was collected giving better mean Temp, Dewp, etc to be used in K-means. One interesting between these two years data is the month of August. August of Y2 is more similar to February of Y2 than August of Y3.

c. (Y1, Y3): In our case 2006 is Y1 and 2009 is Y3. The similarity between the cluster of weather stations for the month of February and August between these two years are 0.117 and 0.107. Similarity between month of February and August for Y1 is 0.14 and Similarity between month of February and August for Y3 is 0.26. It seems that same month across years has less similarity than different month across a year. This should not be the case. The cause of discrepancy could be presence of outlier in Y1 data. We can support this argument using data from Table 4, where there is a cluster of size one in the subset of data in month of February, 2006.

VI. Visualization of Similar Stations

For visualizing the clusters in Texas map, we plot the Latitude and Longitude of the stations into the Texas map using *ggmap*. We use color to group the stations belonging to same cluster The Figure 4 shows the clusters of weather station for the months of February and August of the years 2006, 2008, and 2009. Two interesting things to note from the figure is, the weather in the DFW area remains almost same across years and months, meaning the weather stations in that region belongs to the same cluster. This same thing could be observed in the coastal area of Texas which have similar weather across the year for same months.

Overall Status

Finished the project in entirety by completing the mentioned tasks. Started the project by getting familiar with K-means, kmeans function in R, and cluster evaluation library in R. Then downloaded and preprocessed the dataset. After that, played around with a smaller subset of the dataset consisting of 20 rows. Applied kmeans on the sample dataset for different numbers of K. Finally, performed the analysis mentioned in the project document on the assigned subset of data for our group.

File Descriptions

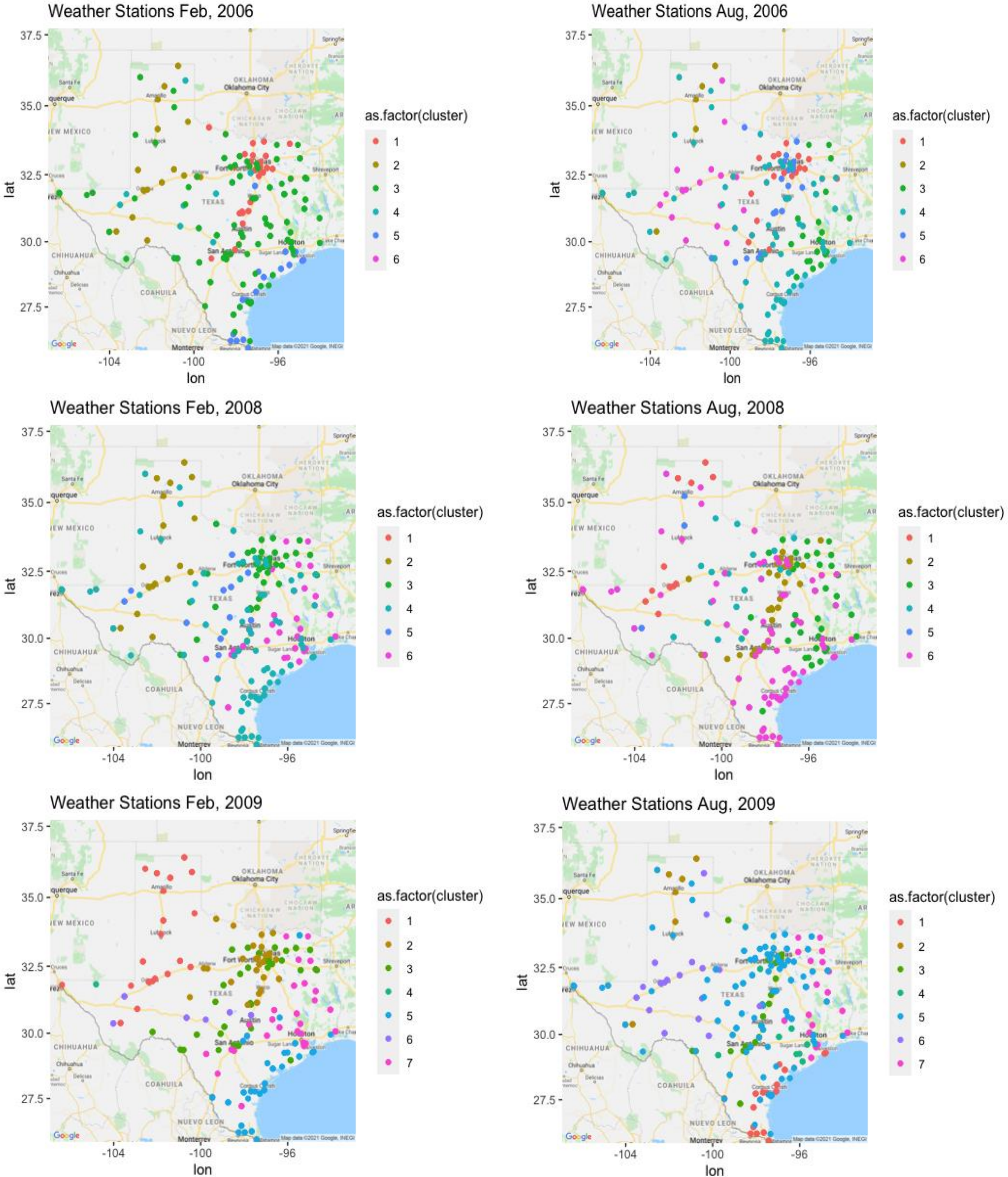
Apart from the project report file (this file), there are two additional directory. The directory named *source_code* contains the R scripts used for the project. The *source_code* folder contains three R script files:

a. **data_preprocessing.r**: This file contains code related to preprocessing the provided dataset. Once the dataset is preprocessed, this script writes the output in 6 different CSV files.

b. **kmeans_analysis.r**: This script contains the analysis related code of the project. The output CSV files from *data_preprocessing.R* scripts are required to run this script.

c. **plotting_map.R**: This file contains the visualization related code used to plot the cluster in Texas map. This file also requires the output CSV files generated by *data_preprocessing.R* script.

Another included directory is the *preprocessed_dataset* directory. This directory contains the preprocessed subset of the entire dataset assigned to our group in the naming format yYEAR_Month.csv where YEAR is either 2006,2008, or 2009 and month is either February or August.



Weather Stations Feb, 2006

Weather Stations Aug, 2006

Weather Stations Feb, 2008

Weather Stations Aug, 2008

Weather Stations Feb, 2009

Weather Stations Aug, 2009

Fig. 4: Weather station clusters plotted in Texas map.

Division of Labor

The entire project was done alone.

Challenges Encountered

1. Preprocessing the data set was more challenging than project 1. There were a lot of missing values. For some stations, there were no valid entries for some years. As we could not remove the rows with missing values, we had to ignore them when calculating the means.
2. Calculating the Jaccard coefficient of two clusters. There were cases where a cluster C1 of Y1's data did not match C1 of Y2, but matched with C3 of Y2. To circumvent the edge cases and complexity, I used the *clusteval* library of R, which provides comprehensives tools to compare and analyze the performance of clusters.
3. Plotting the visualization of the clusters in Texas map. This steps required filtering between multiple data-frame and identifying the relevant stations for the clusters.