Submitted by Team 2

Aparna Krishna Bhat | Surbhi Zambad

Data Mining: Project 2

Clustering

**PROJECT OVERVIEW**

* The weather datasets for the state of Texas during the years 2006-2010 was provided to perform the analysis using k-means clustering and the datafiles containing weather data recordings for every hour of the day (.g files). Also, stations.csv containing Lat, Lon of the weather stations was provided.
* The task was to perform weather data analysis for the month of February for the years 2007, 2008 and 2009.
* The datafile for each year was preprocessed to exactly obtain only the recordings for the month of February to perform the analysis.
* Pre-processing was done using R (sqldf and dplyr libraries).
* The dataframe of the month was queried on the yearmoda\_hr column by performing sqldf function and using !duplicated() and the sorted data for a particular year was obtained.
* Distance matrix was plotted using the functions **get\_dist** and **fviz\_dist** from the **factoextra**R package to illustrate which stations have large dissimilarities (red) versus those that appear to be similar (teal) (Refer fig Similarity Matrix for the month February 2008 down in the report).
* We computed k-means using **Kmeans** function grouping the data into different clusters (k=2 to k=8) with different **nstart(4,10)** that attempts multiple initial configurations (random centroids) and used the best one (centroid).
* To determine optimal clusters size i.e. to choose the value for k we used Elbow method and kmeans function where the location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters (Size of k).
* SSE and Jaccard values for the clusters were computed for Euclidean and Correlation measures for the given years.
* For the comparison between given years: new datasets for each year w.r.t another year such that both the years had the same station numbers (sorted for cluster analysis using sqldf function ) was obtained and Jaccard values between various years are calculated.
* SSE and Jaccard values for the clusters for Euclidean and correlation measures between three years (Feb 2007, 2008, 2009) was computed taking unique station number into consideration.
* A graph plot of all the 4 attributes (DewP, WDSP , Temp, STP) values vs cluster values for given years were plotted using clustplot to understand the change in weather pattern across station-clusters.
* A graph plot using ggplot2 for the stations belonging to the same cluster was plotted on the Texas map.

**Packages Used**

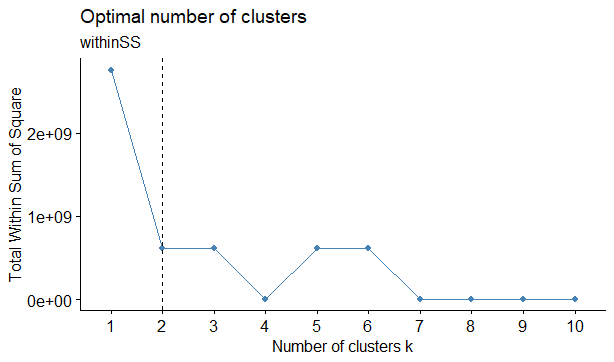
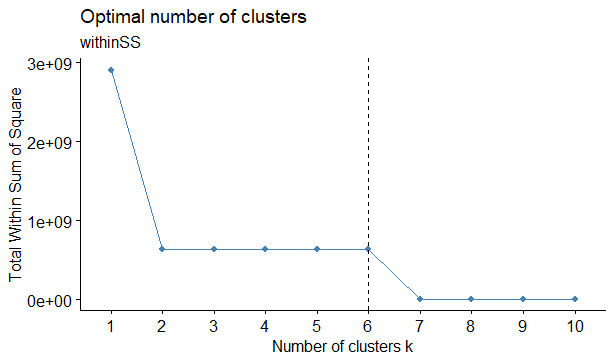
1. dplyr
2. sqldf
3. ggplot2
4. amap #For K-means
5. reprex
6. tidyverse # data manipulation
7. cluster # clustering algorithms
8. factoextra # clustering algorithms & visualization
9. fpc
10. NbClust
11. clv
12. clusteval
13. FactoMineR
14. lubridate

**DATA PRE-PROCESSING**

1. converted the .g files into .csv files by renaming them.
2. Using slice function deleted unused columns and renamed the column heading.
3. Using sqldf function extracted the distinct dataset by querying yearmoda\_hr column and used !duplicated() to remove the duplicate datasets.
4. Used write.csv() to create new sorted dataset csv file for further analysis.
5. Dropped the yearmoda\_hr, SNDP and PRCP columns.

**OPTIMAL K VALUE**

For obtaining optimal K value we used the function kmeans, map\_dbl and plotted optimal number of clusters w.r.t to cluster k and total within sum of square. We used fviz\_nbclust() to plot the k values and obtained the elbow curve to find optimal k value.

**Fig**:- **Elbow curve for Feb 2007** **Fig:- Elbow curve for Feb 2009**

**ANALYSIS**

1. For the month **February 2007** with **seed=5, nstart=4, and k=3**

**Euclidean Distance** within Cluster SSE was found to be:

34.7159 827.7353 436432.9143

**SSE = 437295.365**

K-means clustering with 3 clusters of sizes 20, 124, 9 and cluster means were obtained.

**Pearson correlation** within SSE was found to be:

3.164241e-16 1.168329e-08 1.550169e-14

K-means clustering with 3 clusters of sizes 6, 144, 3 and cluster means were obtained.

**SSE = 1.16833058181141\*10^(-08)**

Jaccard value for Euclidean v/s Pearson is **0.75822686**

1. For the month **February 2007** with **seed = 30, nstart =10, and k=2**

**Euclidean Distance** within Cluster SSE was found to be:

4.796944e+07 8.277353e+02

**SSE =** **47970267.7353**

K-means clustering with 2 clusters of sizes 29, 124 and cluster means were obtained.

**Pearson correlation** within cluster SSE was found to be:

1.506864e-06 1.168329e-08

K-means clustering with 2 clusters of sizes 9, 144 and cluster means were obtained.

**SSE=** **1.51854729\*10^(-06)**

Jaccard value for Euclidean v/s Pearson is **0.746955**

1. For the month **February 2008** with **seed= 5, nstart=4, and k=4**

**Euclidean Distance** within Cluster SSE was found to be:

3512.7013 101.8226 938791.5167 2824.4259

**SSE = 945230.4665**

K-means clustering with 4 clusters of sizes 116, 18, 8, 19 and cluster means were obtained.

**Pearson correlation** within cluster SSE was found to be:

2.554972e-09 3.867764e-07 1.347809e-08 2.068301e-10

K-means clustering with 4 clusters of sizes 8, 5, 124, 24 and cluster means were obtained.

**SSE= 4.030162921\*10^(-07)**

Jaccard value for Euclidean v/s Pearson is **0.670425**

1. For the month **February 2008** with **seed = 30, nstart=10, and k=3**

**Euclidean Distance** within Cluster SSE was found to be:

101.8226 2463.6257 938791.5167

**SSE = 941356.965**

K-means clustering with 3 clusters of sizes 18, 135, 8 and cluster means were obtained.

**Pearson correlation** within cluster SSE was found to be:

1.746641e-09 3.867764e-07 2.554972e-09

K-means clustering with 3 clusters of sizes 148, 5, 8 and cluster means were obtained.

**SSE = 3.91078013\*10^(-07)**

Jaccard value for Euclidean v/s Pearson is **0.741483**

1. For the month **February 2009** with **seed=5, nstart=4, K=6**

**Euclidean Distance** within Cluster SSE was found to be:

2269.5424 261.4165 245.5641 596391.2711 0.0000 1822.0882

**SSE = 600989.8823**

K-means clustering with 6 clusters of sizes 22, 57, 25, 9, 1, 52 and cluster means were obtained.

**Pearson correlation** within cluster SSE was found to be:

8.939285e-09 3.254868e-06 1.858646e-09 1.830305e-11 9.535004e-09 7.083919e-09

K-means clustering with 6 clusters of sizes 32, 9, 25, 35, 21, 44 and cluster means were obtained.

**SSE = 3.28230315705\*10^(-06)**

Jaccard value for Euclidean v/s Pearson is **0.384089**

1. For the month **February 2009** with **seed =30 nstart=10 k = 6**

**Euclidean Distance** within Cluster SSE was found to be:

180.9329 1405.0142 0.0000 272.3448 2269.5424 596391.2711

**SSE = 600519.1054**

K-means clustering with 6 clusters of sizes 79, 32, 1, 23, 22, 9 and cluster means were obtained.

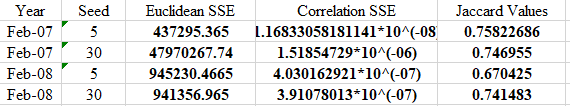
**Pearson correlation** within cluster SSE was found to be:

2.475625e-11 9.535004e-09 8.144502e-09 1.858646e-09 7.083919e-09 3.254868e-06

K-means clustering with 6 clusters of sizes 31, 44, 25, 36, 21, 9 and cluster means were obtained.

**SSE = 3.28151482725\*10^(-06)**

Jaccard value for Euclidean v/s Pearson is **0.351395**



In the above comparisons, we can analyse from the SSE for the Euclidean and SSE for the Pearson’s correlation that SSE for the Pearson correlation is a better clustering technique as it is having lesser SSE value and, lower SSE value suggests better results. One reason, for this behaviour is that correlation uses covariance. For each cluster using different seed values and nstart values for given years we observed that Pearson correlation performed better.

For using different Jaccard values considering different seed values, the value of Jaccard between Pearson correlation and Euclidean varied from 0.1-0.3-0.7 (seed values = 5, 30). This indicates that cluster vectors are nearly same with same datasets and changing only the distance measure.

1. **Year to Year Comparisons – (2007,2008), (2007,2009), (2008,2009)**

1. (2007, 2008) with seed=5: - **Euclidean Jaccard value** = 0.22630696,

**Correlation Jaccard value** = 0.22741734

2. (2007, 2009) with seed=5: - **Euclidean Jaccard value** = 0.227494,

**Correlation Jaccard value** = 0.1716489

3. (2008,2009) with seed=5: - **Euclidean Jaccard value** = 0.179685

**Correlation Jaccard value** = 0.166916

For the year (2007,2008) we got higher Jaccard value than the year (2008,2009). This implies that more weather change occurred between the year (2008,2009) as compared to the year (2007,2008). It can be observed that the Jaccard values for the years (2008,2009) was the least compared to the other years. Overall Jaccard value for Correlation is less than Euclidean Jaccard value.

1. **SSE and Jaccard comparison for three years (2007\_2008\_2009)** with **seed = 5, nstart=4, and k = 5**

**Euclidean Distance** within Cluster SSE was found to be:

74486.6774 1263.8807 2532.8629 0.0000 558.0277

**SSE = 78841.4487**

K-means clustering with 5 clusters of sizes 11, 97, 27, 1, 34 and cluster means were obtained.

**Pearson correlation** within cluster SSE was found to be:

0.000000e+00 5.710025e-11 1.352582e-12 2.158753e-11 4.355796e-08

K-means clustering with 5 clusters of sizes 1, 59, 37, 62, 11 and cluster means were obtained.

**SSE = 4.3638000362\*10^(-08)**

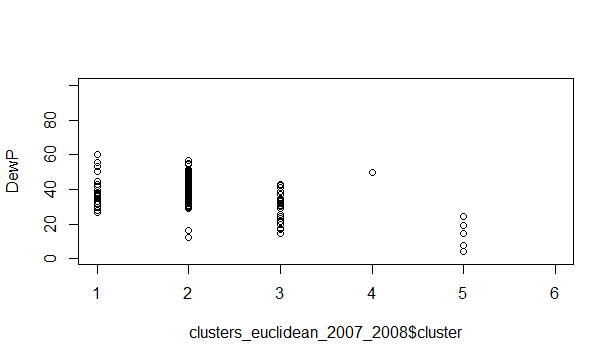
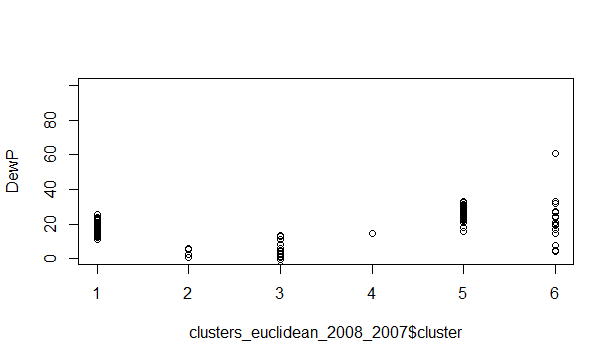
Jaccard value for Euclidean v/s Pearson is **0.4437509**

Here we can see that the SSE value for Euclidean is higher than the SSE value for the Pearson correlation. This implies that Pearson correlation is better technique for cluster analysis while taking three years into consideration as lesser the SSE value better the performance.

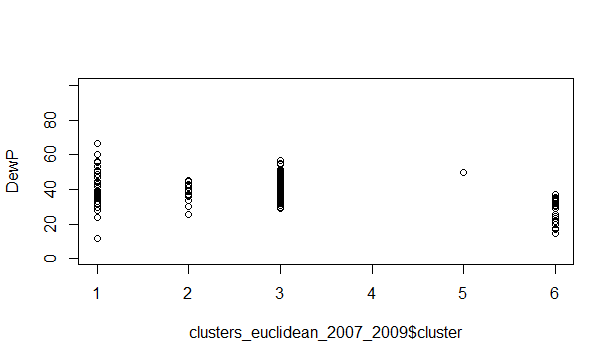
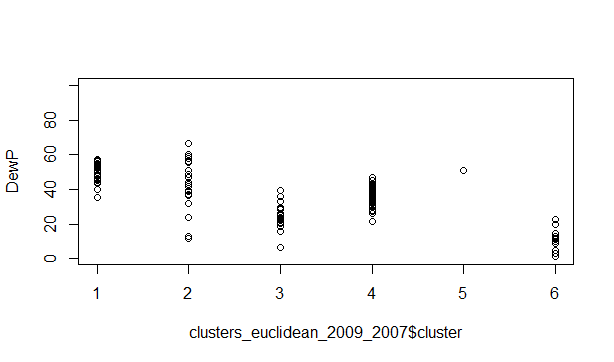
**VISUALIZATION**

(More visualization results are present in the visualization folder)

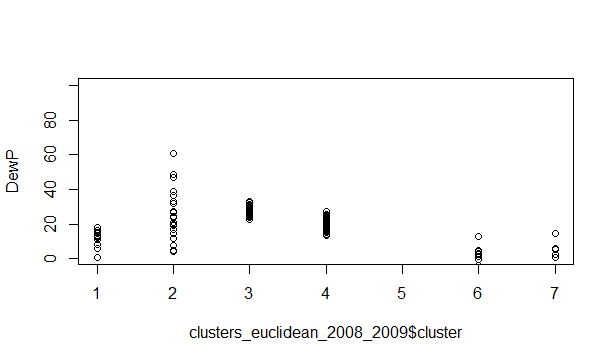
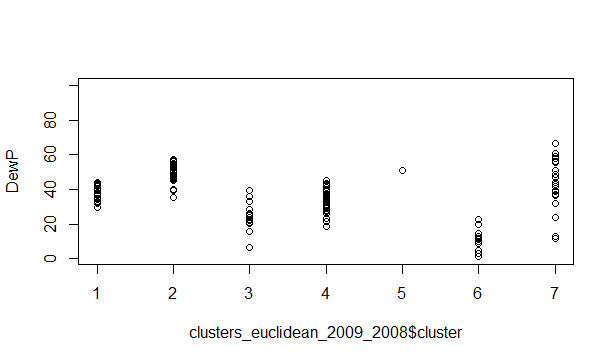
**DewP:** DewP analysis for Euclidean across various clusters

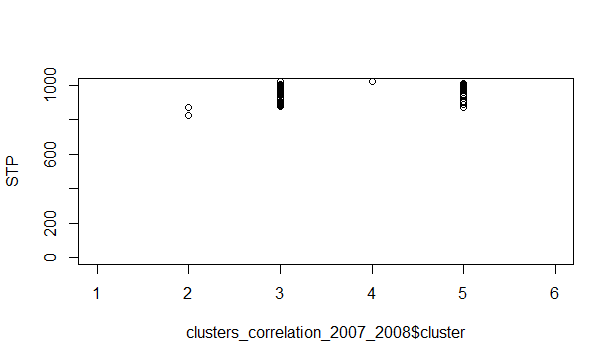
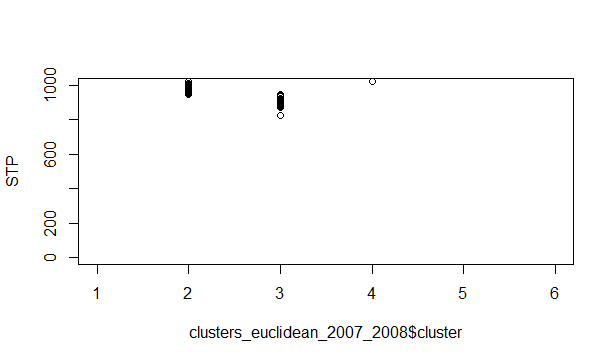
**Fig:- Euclidean 2007\_2008 for DewP Fig:- Euclidean 2008\_2007 DewP**

**Fig: - Euclidean 2007\_2009 for DewP Fig: - Euclidean 2009\_2007 for DewP**

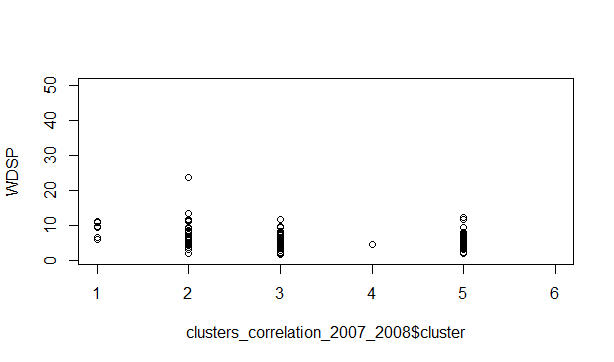
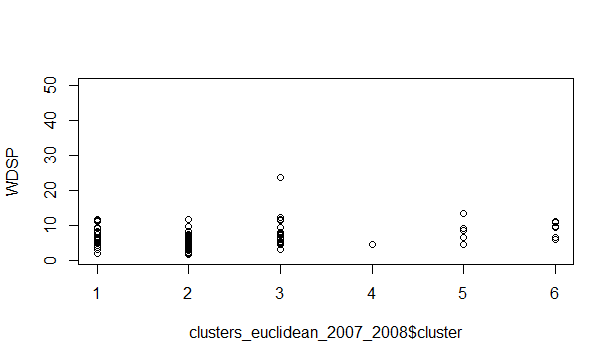
 

**Fig: - Euclidean 2008\_2009 for DewP Fig: - Euclidean 2009\_2008 for DewP**

**STP:** STP analysis for Euclidean and correlation for the year 2007\_2008 across various cluster 

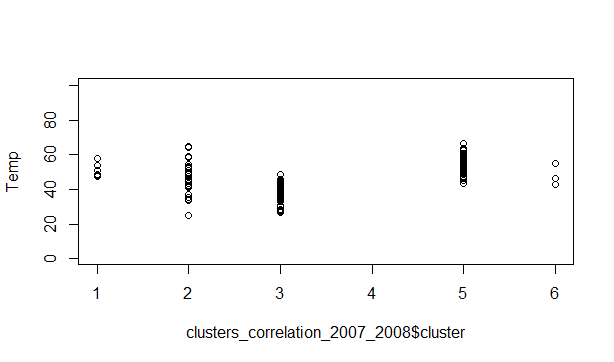
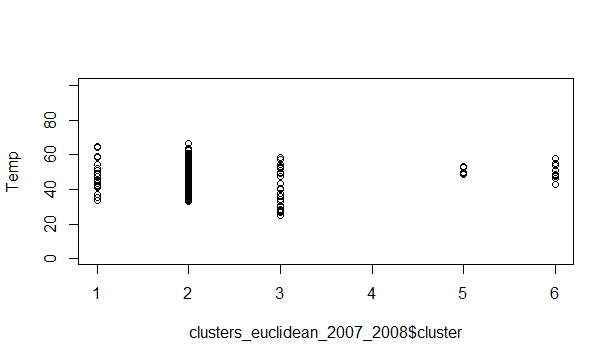
**Fig: - Correlation 2007\_2008 for STP Fig: - Euclidean 2007\_2008 for STP**

**WDSP :** WDSP analysis for Euclidean and correlation for the year 2007\_2008 across various cluster

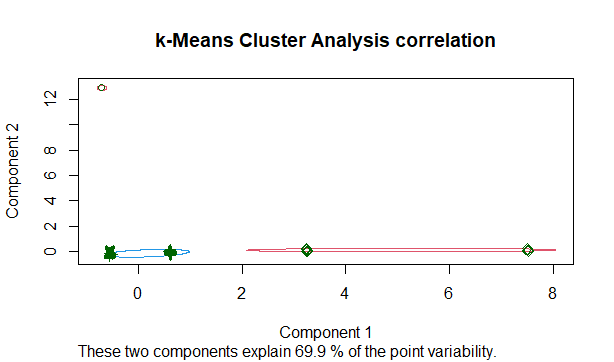
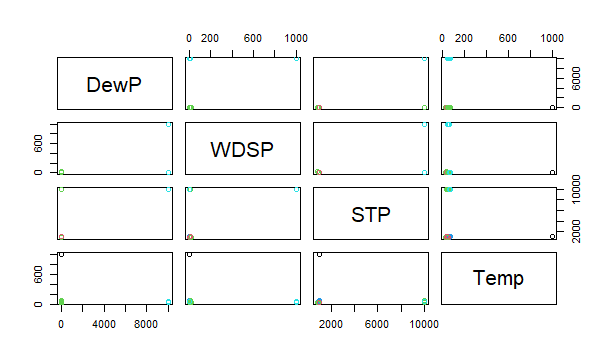
 

**Fig: - Correlation 2007\_2008 for WDSP Fig: - Euclidean 2007\_2008 for WDSP**

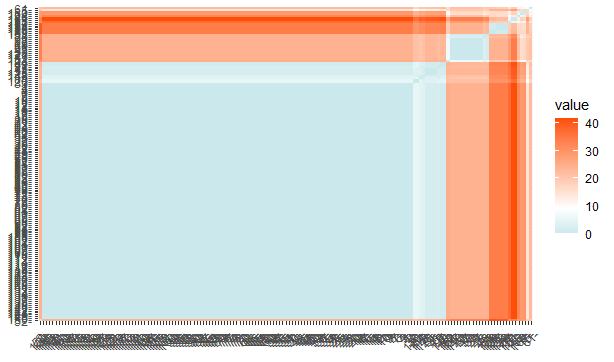
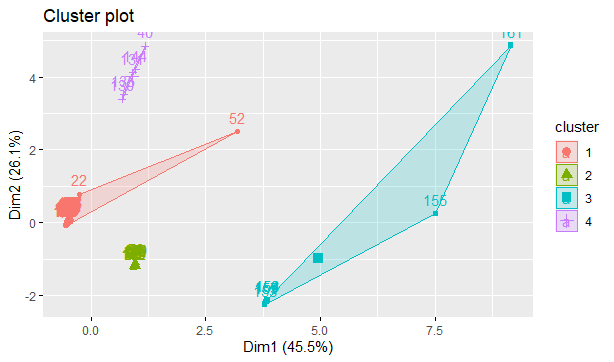
**Temp:** Temp analysis for Euclidean and correlation for the year 2007\_2008 across various cluster

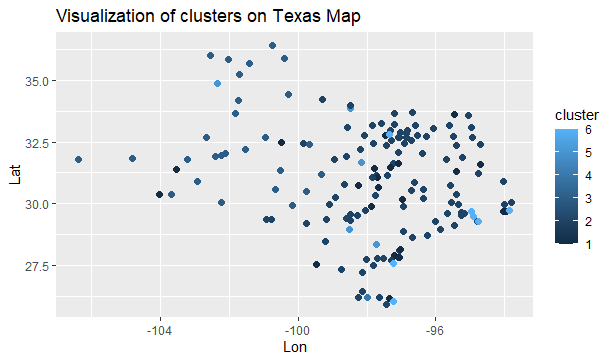
**Fig: - Correlation 2007\_2008 for Temp Fig: - Euclidean 2007\_2008 for Temp**



**Fig:- Attribute wise distribution for the years 07, 08, 09 Fig:- k-means cluster analysis for correlation**

**Fig:- Similarity Matrix for the month February 2008 Fig:- Cluster plot for the month Feb 2008 (k=4)**



**Fig:- Data visualization of clusters on Texas map for February (2007,2008)**

We plotted the graph using the Lat, Lon provided in the stations.csv for the month Feb(2007,2008) using station numbers and clusters.

**CHALLENGES**

1. Interpreting the meaning of Jaccard values and its usability and fluctuation to different datasets. Also interpreting how Euclidean and Pearson can perform for attributes with different range.
2. Handling a Huge dataset and many folds of pre-processing the dataset.
3. Data visualization of cluster on the Texas map.

**FILE NAMES**

1. R\_files\_team2 folder contains all the R source file used in the project.
2. visualization\_team2 folder has all the graph plots and analysis. (Many years wise visualization excluded from the report due to page limitation are present in this folder).
3. dataset\_team2 folder has various datasets related to pre-processing and final input dataset.