An Assignment on Multivariate Analysis

Course Name: Multivariate Analysis

Course Code: PM-ASDS06

Submitted to:

Md. Habibur Rahman

Associate Professor

Department of Statistics, JU

Submitted by:

Mohammad Saiduzzaman Sayed

ID: 20215063

Batch: 5th

Sec: A

Professional Masters in

Applied Statistics and Data Science (PM-ASDS)

JAHANGIRNAGAR UNIVERSITY

Assignment on Multivariate Analysis

Problem 01:

In our "Mymensingh.csv" multivariate dataset, there are 672 rows which are observations, and 12 Columns, which are variables. After selecting mentioned variables in the Assignment which are Temperature (TEM), Dew point temperature (DPT), wind speed (WIS), Humidity (HUM), Sea level pressure (SLP), and Total Rainfall (T_RAN), and removing the missing values, now the dataset has 654 observations and 6 variables.

(i) Chi-square Q-Q plot:

A formal method for judging the joint normality of a data set is based on the squared generalized distances.

$$d_i^2 = (x_i - \bar{x})' \, S^{-1} \, (x_i - \bar{x}) \; for \; i = 1,2,3, \dots, n$$

To construct the chi-square plot

(i) Order the squared distances $d_i^2 = (x_i - \bar{x})' \, S^{-1} \, (x_i - \bar{x}) \; for \; i = 1,2,3, \dots, n$ from smallest to largest as $d_{(1)}^2 \leq d_{(2)}^2 \leq \cdots \leq d_{(n)}^2$

(ii) Calculate the quantiles $\chi_p^2 \left[\frac{n-i+0.5}{n}\right] \; for \; i = 1,2,3, \dots, n$

(iii) Plot the pairs of observations $\left(\chi_p^2 \left[\frac{n-i+0.5}{n}\right], d_{(i)}^2\right) \; for \; i = 1,2,3, \dots, n$ and examine the "straightness" of the outcome.

The Chi-square Q-Q plot for the distribution of all 6 variables:

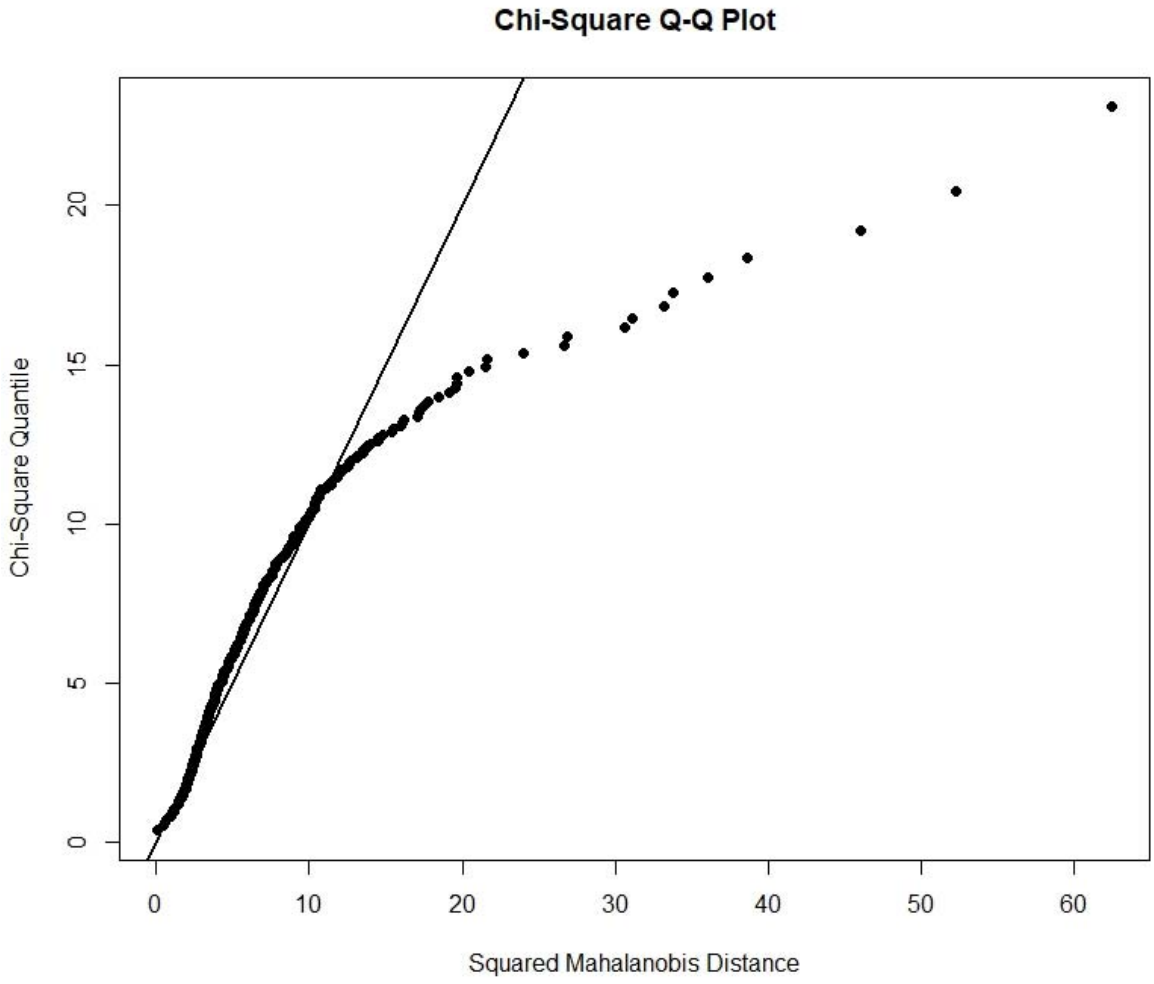

Fig: Chi-square Q-Q plot for Mymensingh.csv

Comment:

In the Multivariate Normality test using Chi-square Q-Q plot, If the line is 45-degree straight, then we can say our dataset is normally distributed.

In this case, the straight line is far from the 45-degree, which means our dataset is not normally distributed. And, we can say that from the Chi-square Q-Q plot, our dataset contains many outliers.

(ii) The new dataset extracting without outliers:

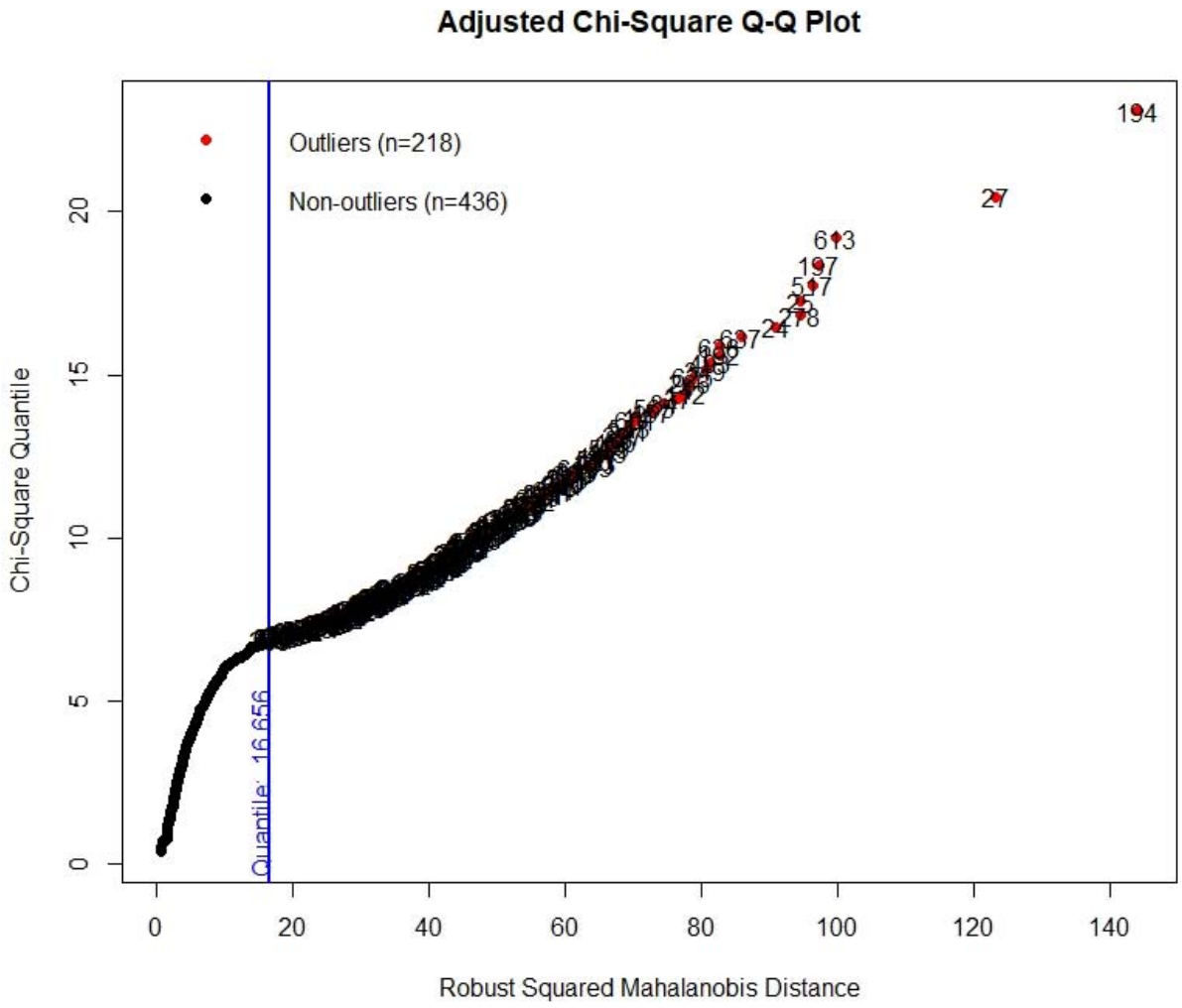

Fig: Adjusted Chi-square Q-Q plot

With the help of the Adjusted Chi-square Q-Q plot, we can see our dataset has 218 outliers and 436 non-outliers. After extracting our new dataset without outliers, the new dataset contains 436 observations and 6 variables.

(iii) Chi-square Q-Q plot for new dataset:

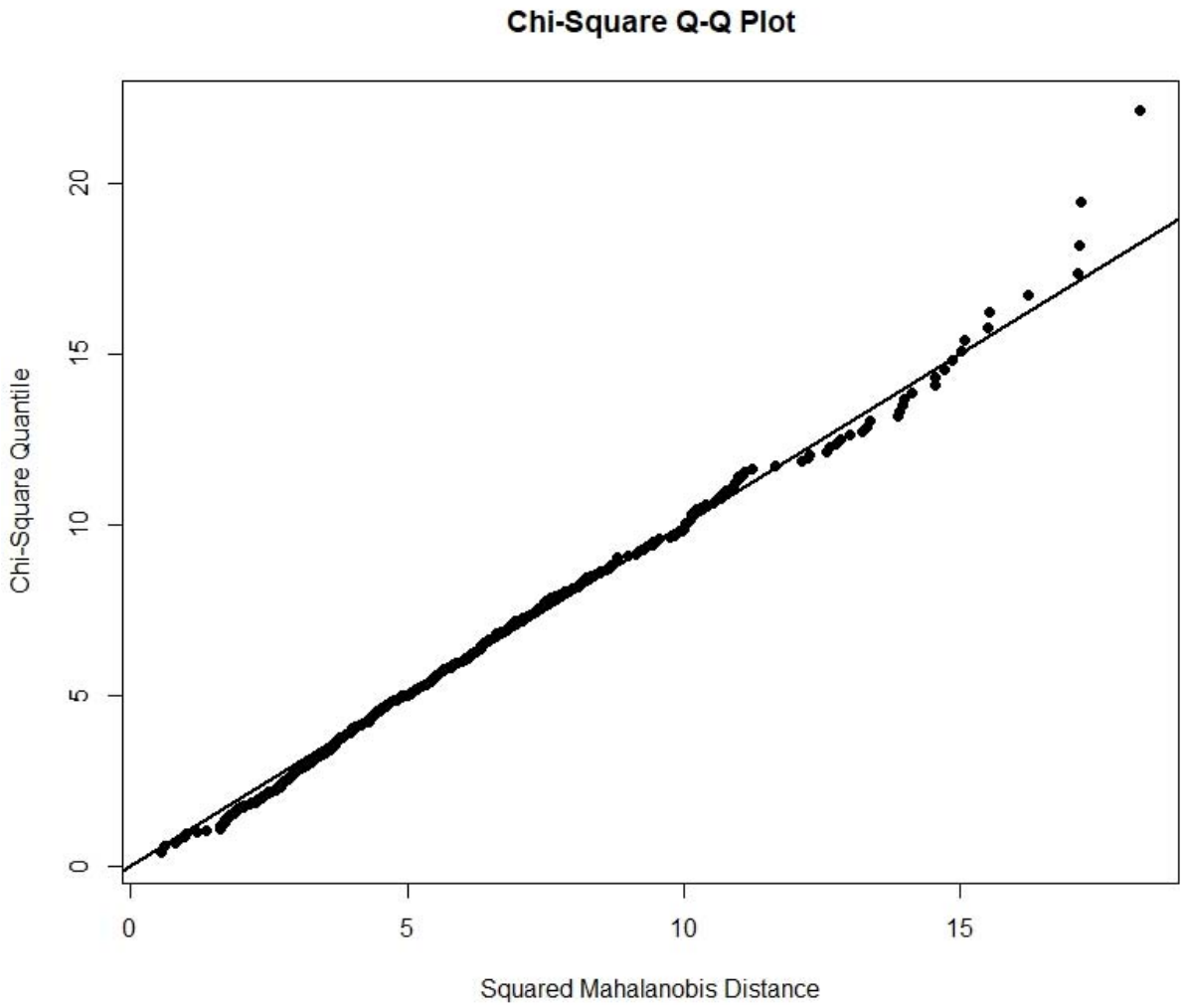

Fig: Chi-square Q-Q plot for new data

Comment:

For the new dataset, the straight line is close to the 45-degree, which means our new dataset (after removing the outliers) is approximately normally distributed.

(iv) Scatter plot for selected new data (except outliers):

The scatter plot helps us to identify what type of relationship between quantitative variables.

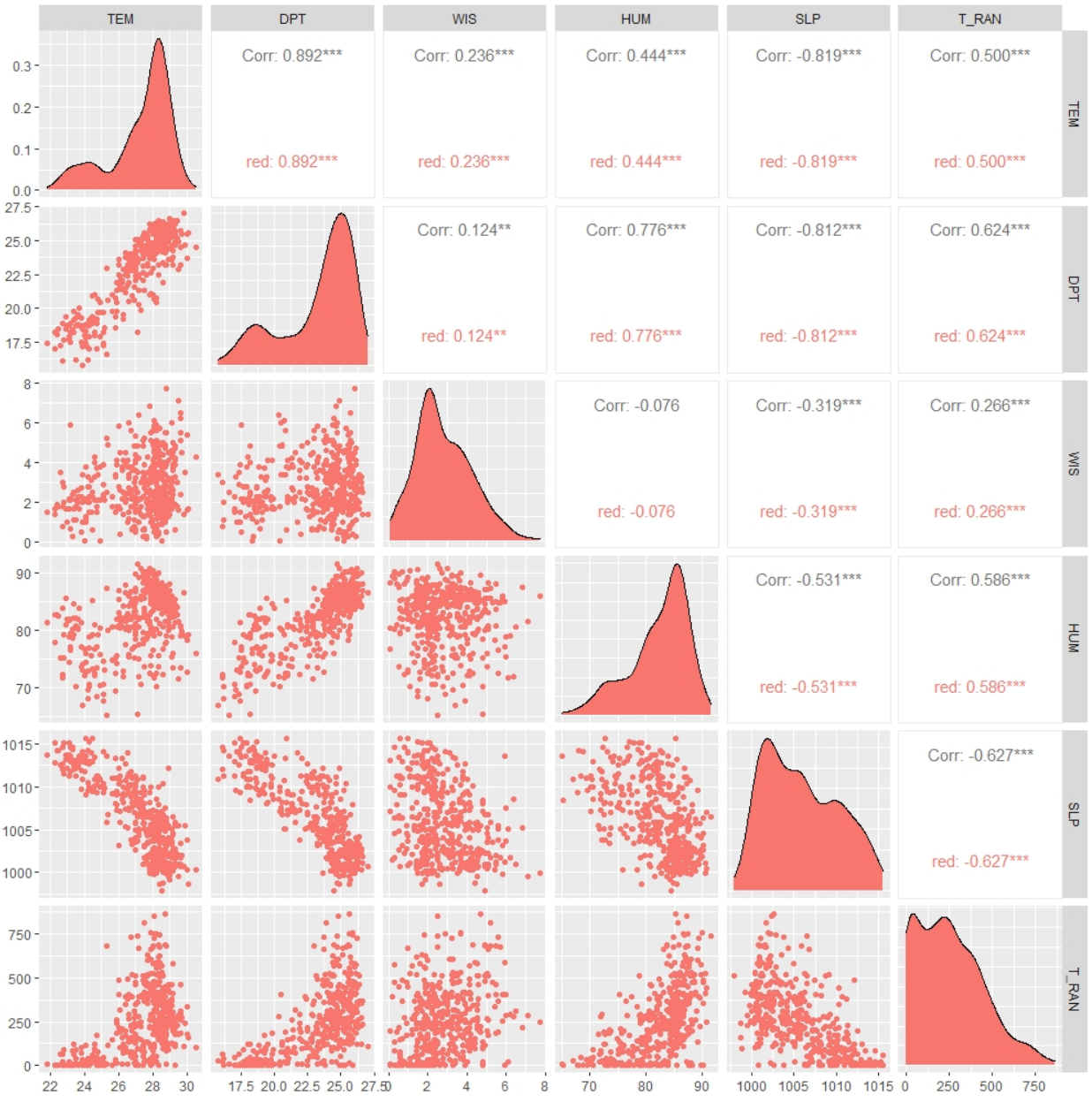

Fig: Scatter Plot for new data

(v) Mean Vector for new data:

The population mean:

$$\mu_p = \frac{1}{N}\sum_{i=1}^{N} x_{ip}$$

Then the mean vector:

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

Using the above formula, we get our mean vector.

TEM	DPT	WIS	HUM	SLP	T_RAN
27.275688	23.278211	2.798624	82.312844	1005.920642	249.325688

variance–covariance matrix for new data:

The population variance is,

$$\sigma_{pp} = \frac{1}{N}\sum_{i=1}^{n}(X_{ip} - \mu_p)^2$$

The population covariance is,

$$\sigma_{pq} = \frac{1}{N}\sum_{i=1}^{n}(X_{ip} - \mu_p)(X_{iq} - \mu_q)$$

Using the above formula, we get our variance-covariance matrix.

	TEM	DPT	WIS	HUM	SLP	T_RAN
TEM	3.3470398	4.4427564	0.6034607	4.2440900	-6.503865	177.50656
DPT	4.4427564	7.4140299	0.4725906	11.0396483	-9.595779	329.74389
WIS	0.6034607	0.4725906	1.9609636	-0.5593777	-1.940615	72.42551
HUM	4.2440900	11.0396483	-0.5593777	27.2774604	-12.030723	593.74316
SLP	-6.5038648	-9.5957791	-1.9406152	-12.0307232	18.827251	-528.00996
T_RAN	177.5065570	329.7438943	72.4255067	593.7431636	-528.009957	37684.41782

(vi) Scaling the dataset:

Standardization is a feature scaling technique. It is the process of rescaling data so that the data have a mean of '0' and standard deviation of '1'.

We can scale our dataset using below formula:

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

Here, $\bar{x}$ = mean

σ = standard deviation

The principal components for new data:

The random vector $\boldsymbol{X}' = [X_1, X_2, \dots, X_p]$ have the covariance matrix $\boldsymbol{\Sigma}$ with eigen values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, then the ith principal component is,

$$Y_i = \boldsymbol{e}_i'\boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \quad i = 1,2, \dots, p$$

Using above formula, we get,

	PC1	PC2	PC3	PC4	PC5	PC6
TEM	0.452		0.549	0.127	0.428	0.538
DPT	0.495	-0.156	0.188	-0.198	0.279	-0.760
WIS	0.145	0.867	-0.220	-0.415		
HUM	0.388	-0.442	-0.396	-0.596		0.365
SLP	-0.468	-0.141	-0.217	-0.177	0.826	
T_RAN	0.401		-0.641	0.621	0.200	

(vii) The scree plot for the different principal components:

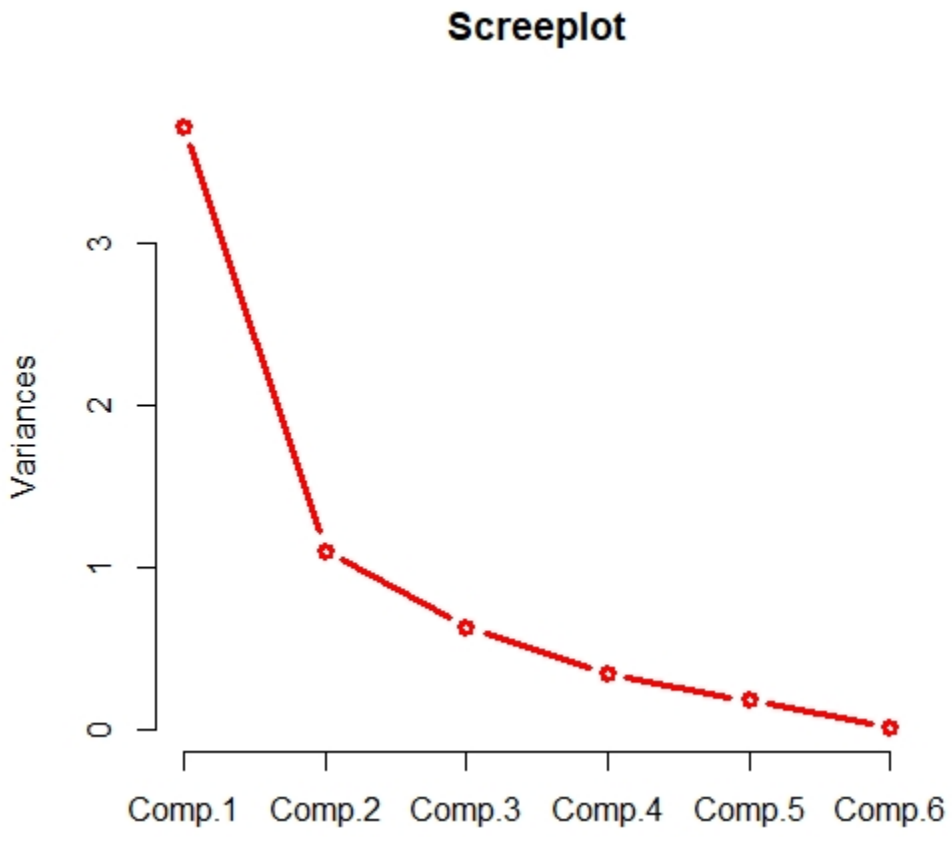

Fig: Scree Plot

Comment on scree plot:

To find out the important Principal components, we use the scree plot. We can plot each principal component against their variances. The "knee down point" of the graph where the eigenvalues seem to level off is found and components to the left of this point should be retained as important. In this case scree plot, Principal Component 1, 2 and 3 are enough to describe the data. And we ignore other principal components. So, there are three important principal components.

Biplot:

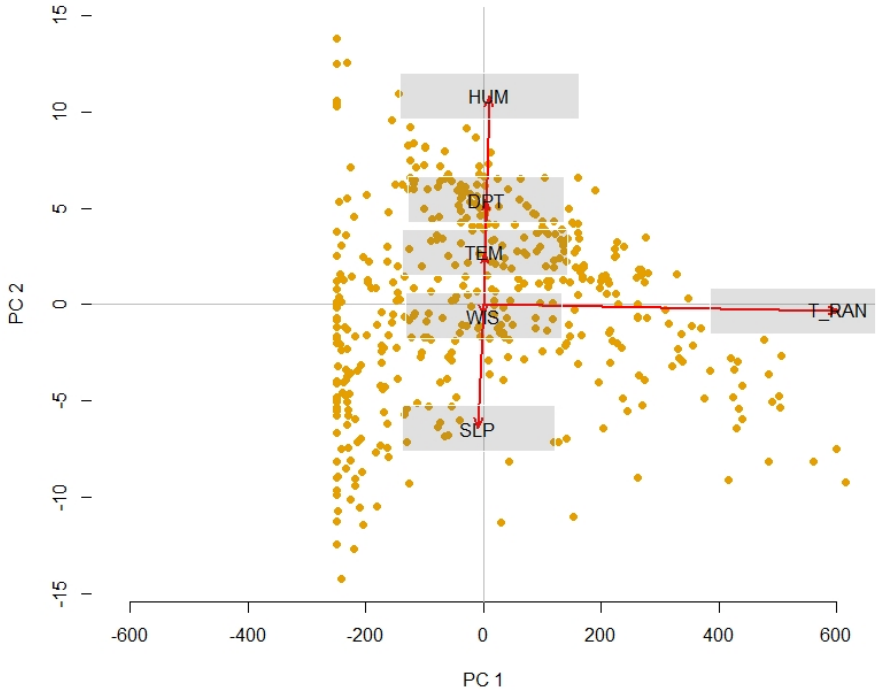

Fig: 2D Biplot

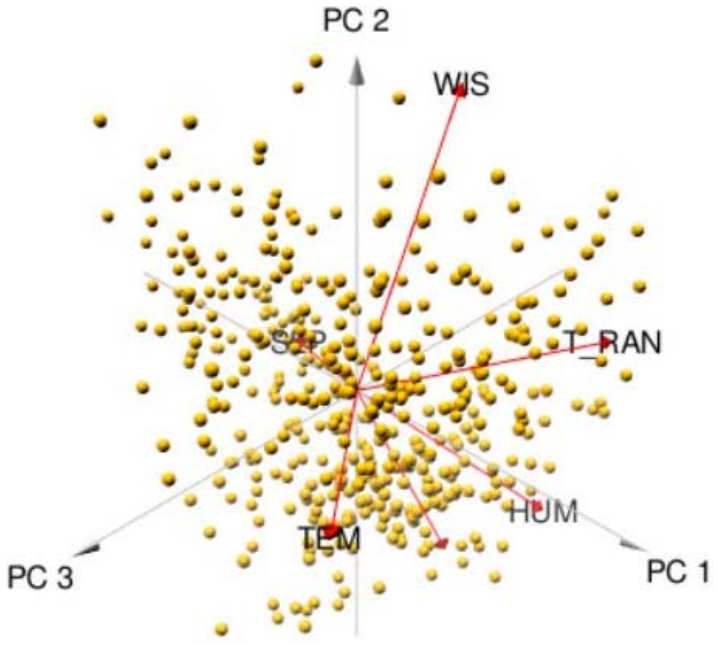

Fig: 3D Biplot

Comment on Biplot:

With the help of Biplot, we can plot information on the observations and the variables in a multidimensional dataset. And it represents the distances between observations and the relationships between variables. From our Biplot, relationship between variables:

TEM and DPT ~ Strong Positive Correlation (89%)

TEM and WIS ~ Positive Correlation (23%)

TEM and HUM ~ Positive Correlation (44%)

TEM and SLP ~ Strong Negative Correlation (- 81%)

TEM and T_RAN ~ Positive Correlation (49%)

DPT and WIS ~ Positive Correlation (12%)

DPT and HUM ~ Moderate Positive Correlation (77%)

DPT and SLP ~ Strong Negative Correlation (- 81%)

DPT and T_RAN ~ Positive Correlation (62%)

WIS and HUM ~ Negative Correlation (- 7%)

WIS and SLP ~ Negative Correlation (- 31%)

WIS and T_RAN ~ Positive Correlation (26%)

HUM and SLP ~ Negative Correlation (- 53%)

HUM and T_RAN ~ Positive Correlation (58%)

SLP and T_RAN ~ Negative Correlation (- 62%)

The percentage values gain from the correlation matrix.

(viii) Multiple Linear Regression:

We fit a regression line to predict the total rainfall (T_RAN) based on Temperature (TEM), Dew point temperature (DPT), wind speed (WIS), Humidity (HUM), Sea level pressure (SLP).

ANOVA Table: Analysis of Variance | Table Response: T_RAN

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
TEM	1	4095031	4095031	220.203	< 2.2e-16	***
DPT	1	2540833	2540833	136.629	< 2.2e-16	***
WIS	1	902911	902911	48.553	1.214e-11	***
HUM	1	245907	245907	13.223	0.00031	***
SLP	1	611507	611507	32.883	1.847e-08	***
Residuals	430	7996533	18597			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

Q-Q plots for errors and other important results:

Residuals vs Fitted plot:

A Residuals vs Fitted plot displays the spread of residuals over fitted values.

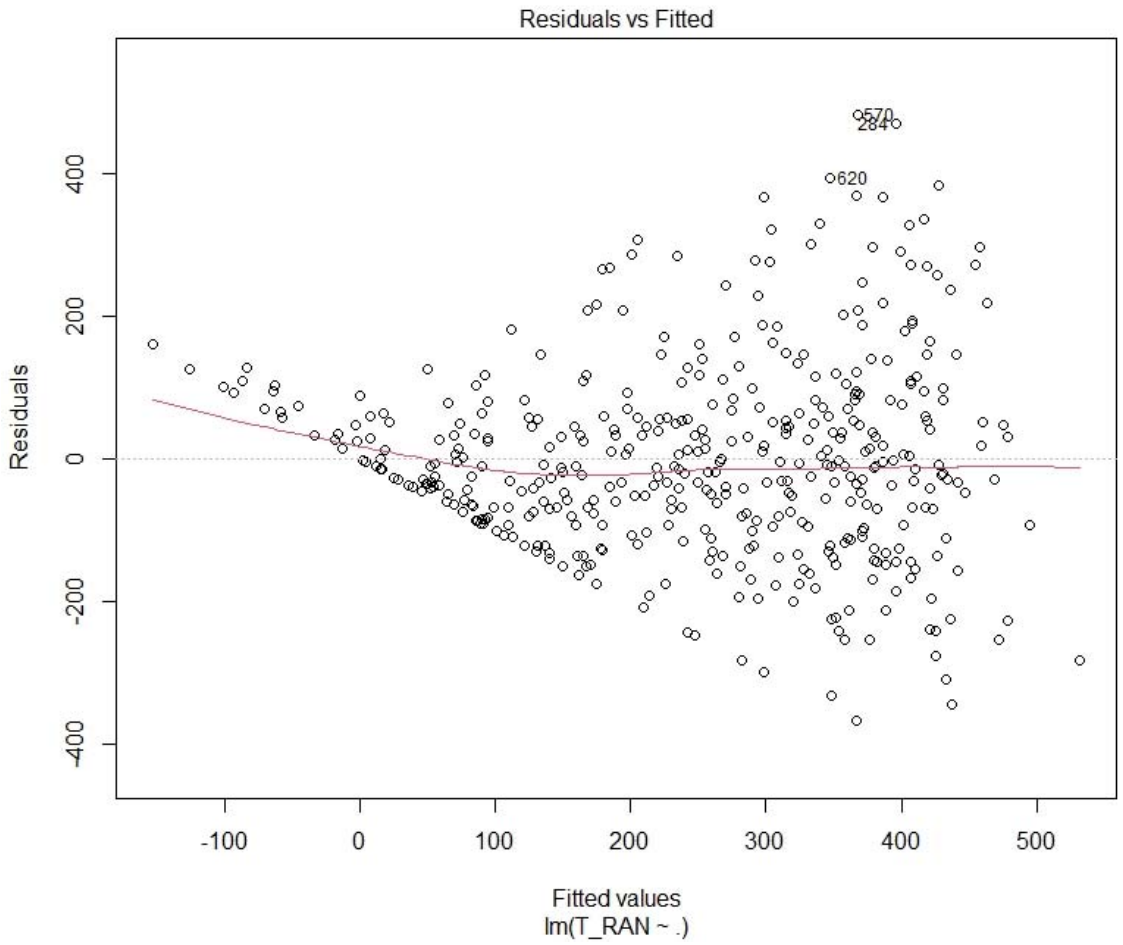

Fig: Residuals vs Fitted Plot

In the above Residual vs Fitted plot, we can see that the spread of the residuals tends to be higher for higher fitted values, and it does look like we would need to make any changes to the model.

Normal Q-Q plot for errors:

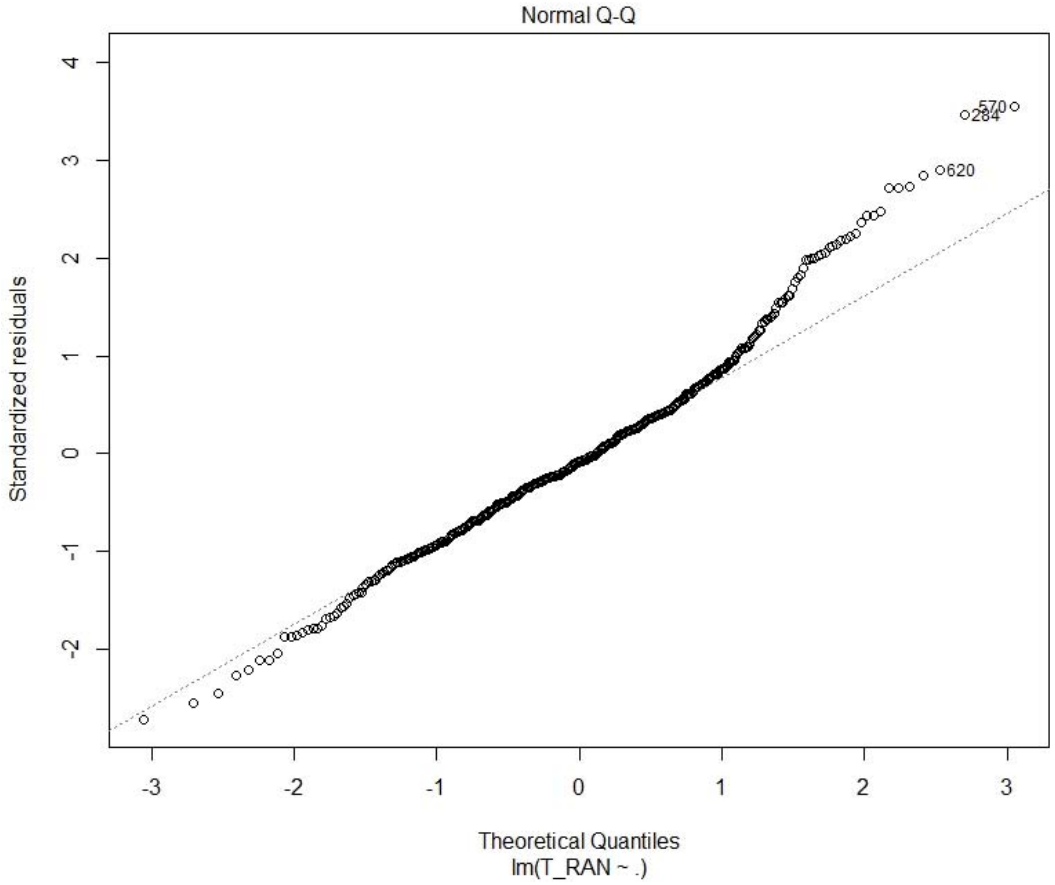

Fig: Normal Q-Q Plot for error

In the above Normal Q-Q plot, residuals are approximately normally distributed because the straight line is close to 45-degree.

Scale Location plot:

A scale-location plot displays the fitted values of a regression model. By this plot, we can check whether the red line is approximately horizontal or not. And we can check whether the spread around the red line with the fitted values does vary or not.

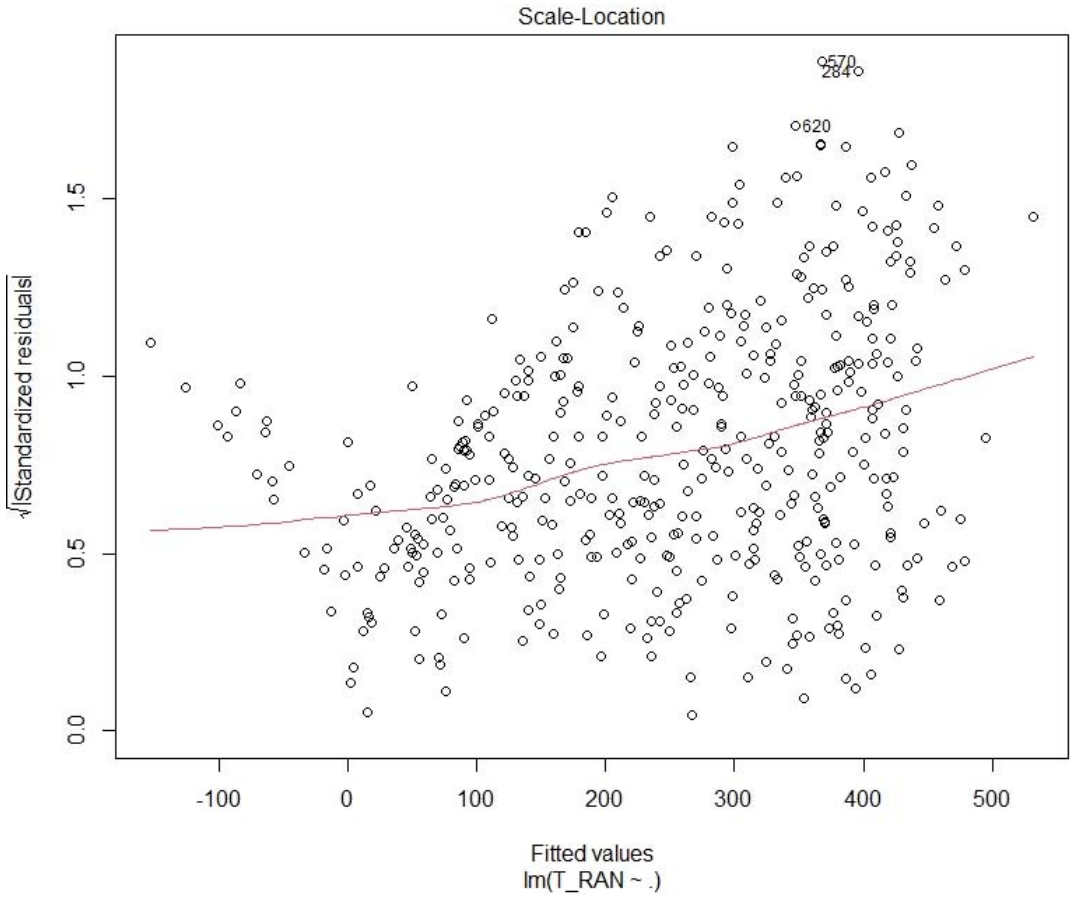

Fig: Scale-Location Plot

In the above plot, the red line is approximately horizontal. And the fitted values spread around the red line.

Residuals vs Leverage plot:

A Residuals vs Leverage plot represents to identify influential observations in a regression model. If any point in the plot falls outside of Cook's distance (the red dashed lines) then it is considered to be an influential observation.

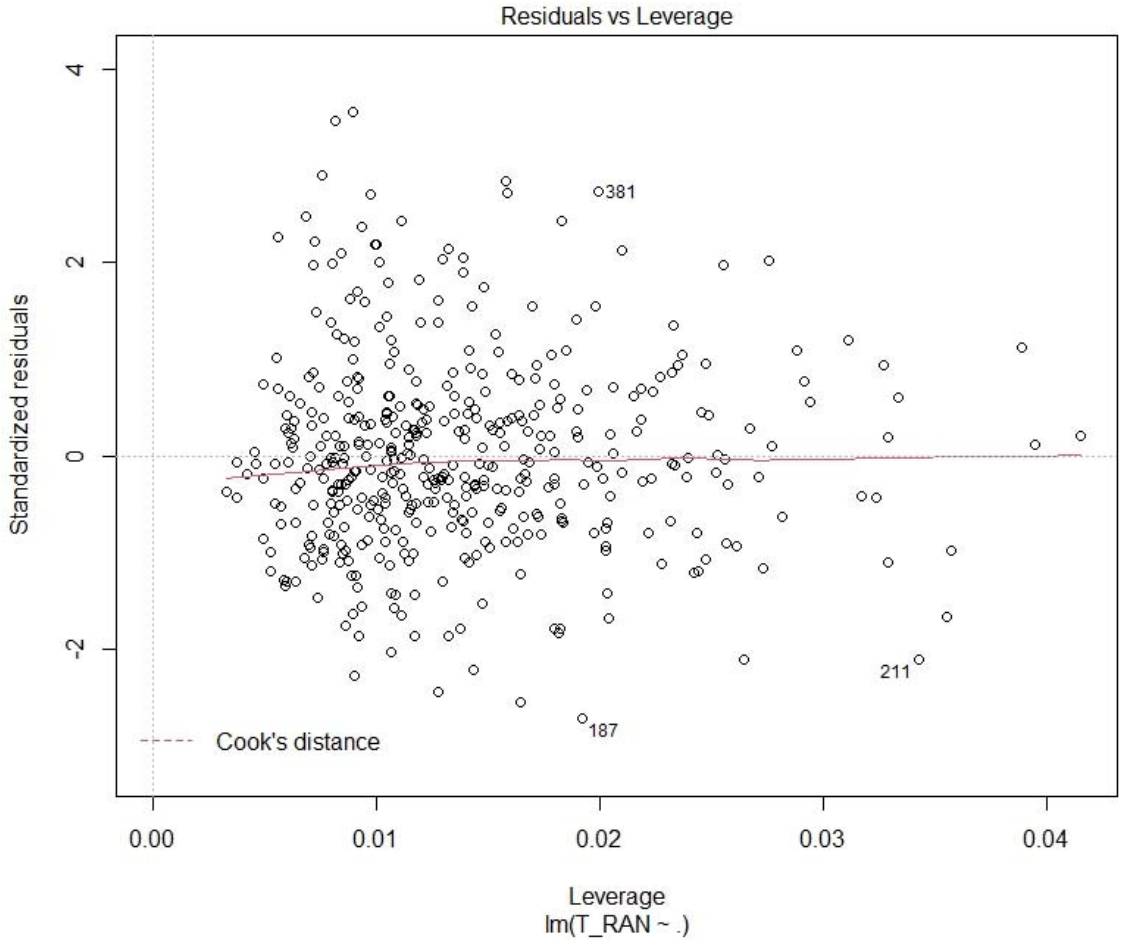

Fig: Residuals vs Leverage

In the above plot, we can see that observation #187 lies closest to the border of Cook's distance, but it doesn't fall outside of the dashed line. This means there are not any influential points in our regression model.

Significant Variables:

From our model summary, to predict our T_RAN variable, WIS and SLP are the most significant, and HUM is less significant than WIS and SLP. TEM and DPT are not significant for this model.

Summary:

Here, R-squared is 0.5122, which means the TEM, DPT, WIS, HUM, SLP variables can explain the T-RAN variable with 51.22% accuracy. So, our model is not well fitted in this situation.

Problem 2

Table A: Maximum Temperature

Table A represents the maximum temperature of different locations in Bangladesh for some period of time. Here, 34 stations with some period and their monthly basis maximum temperature. Based on our problem, our variables are 12 months and observations are row-wise stations.

Hierarchical Clustering:

Distance Matrix:

To do Hierarchical clustering, we need to calculate the distance matrix. The length of a straight line drawn from one cluster to another is used to calculate the distance between them and here we are using Euclidean distance to find the distance matrix. The straight-line distance between two arbitrary points P and Q with coordinates $P = (x_1, x_2, \dots, x_p)$ and $Q = (y_1, y_2, \dots, y_p)$ is given by

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

Single Linkage:

Distances or similarities between two objects can be used as inputs to a single linkage algorithm. Groups are formed from the individuals by merging nearest neighbors, where the term nearest neighbor refers to the minimum distance. We can use a dendrogram to visualize the results of single linkage clustering.

Comment on dendrogram and Cluster allocated:

In the dendrogram below, we can see the similarities between some stations such as Barisal and Patuakhali, Bhola and Khepupara etc. Also, the height of the dendrogram indicates the order in which the clusters were joined.

Observations are allocated to clusters by drawing a horizontal line through the dendrogram. Observations that are joined together below the line are in clusters. In the example below, we have two clusters. One cluster combines Rangpur, Dinajpur and Sayedpur and the second cluster combining other 31 stations.

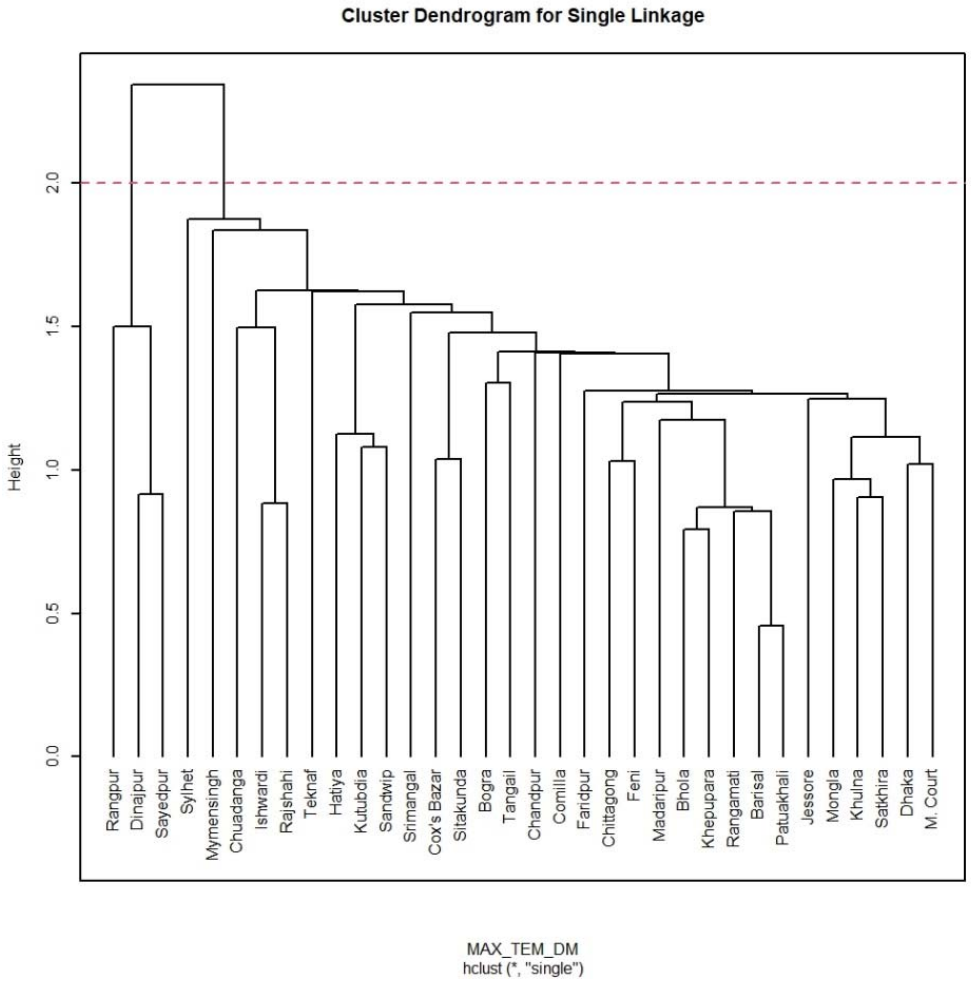

Fig: Single Linkage Cluster Dendrogram for Maximum Temperature

Complete Linkage:

In Complete linkage clustering, the distance between clusters is determined by the distance between the two objects, one from each cluster, that are most distant. Thus, complete linkage ensures that all items in a cluster are within some maximum distance of each other. Also, for complete linkage clustering, We can use a dendrogram to visualize the results.

Comment on dendrogram and Cluster allocated:

In the example below, we merged six subgroups into two clusters. One cluster combines Mongla, Khulna, Satkhira, Ishwardi, Rajshahi, Chuadanga, Jessore and the second cluster combining other 27 stations.

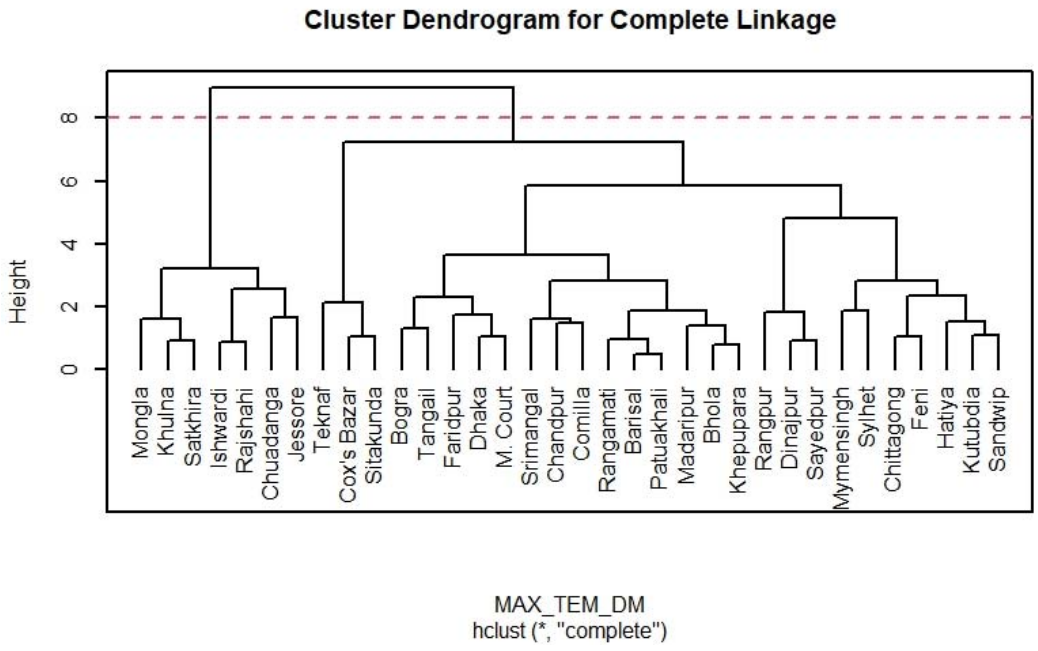

Fig: Complete Linkage Cluster Dendrogram for Maximum Temperature

Average linkage:

Average linkage considers the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster.

Comment on dendrogram and Cluster allocated:

In the example below, we merged seven subgroups into two clusters. One cluster combines Mongla, Khulna, Satkhira, Ishwardi, Rajshahi, Chuadanga, Jessore and the second cluster combining other 27 stations.

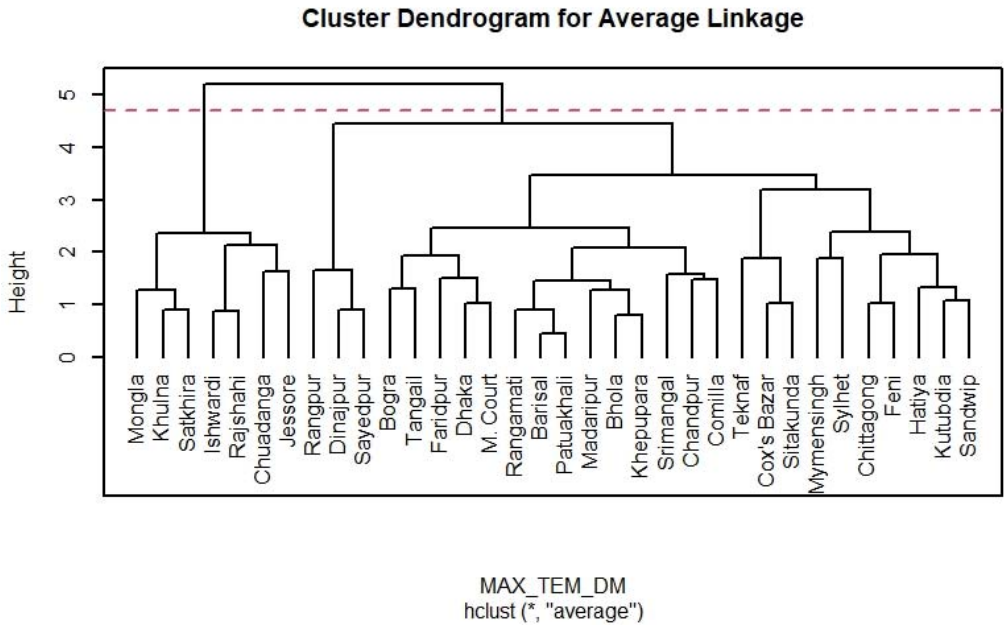

Fig: Average Linkage Cluster Dendrogram for Maximum Temperature

Ward's linkage:

Ward considered hierarchical clustering algorithms that were based on minimizing information loss when connecting two groups.

Comment on dendrogram and Cluster allocated:

In the example below, we merged five subgroups into two clusters. One cluster combines Mongla, Khulna, Satkhira, Ishwardi, Rajshahi, Chuadanga, Jessore and the second cluster combining other 27 stations.

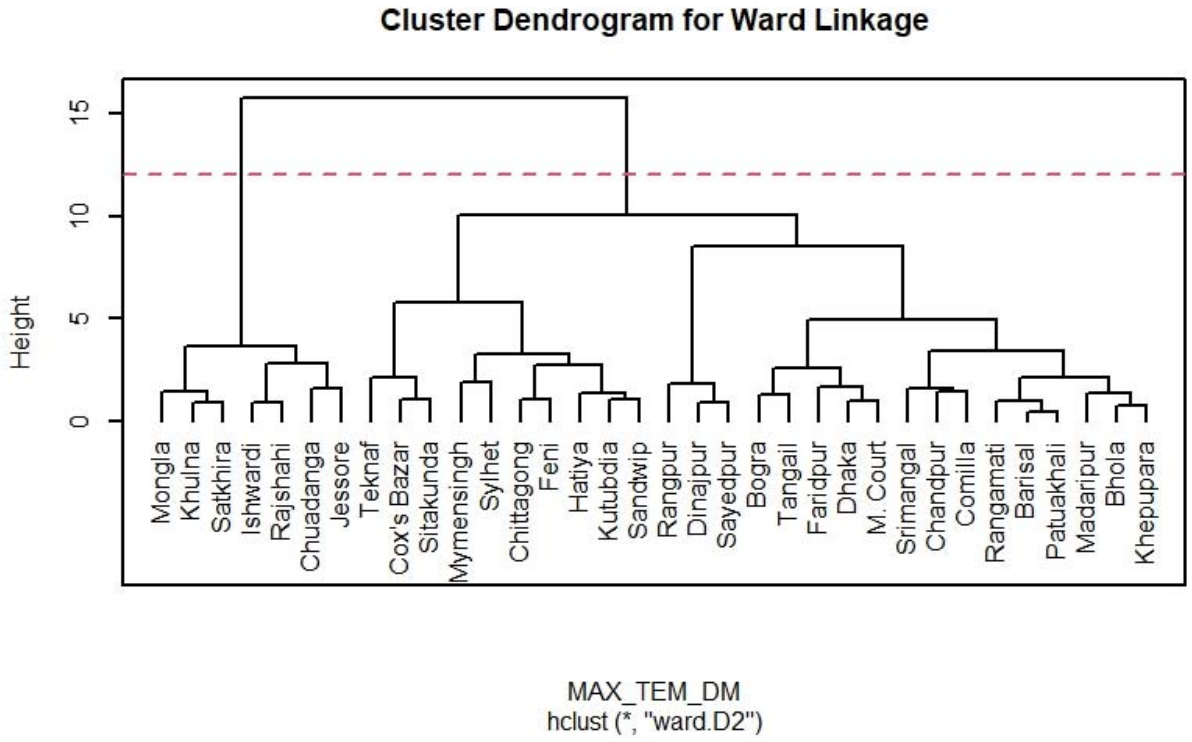

Fig: Ward's Linkage Cluster Dendrogram for Maximum Temperature

Non-hierarchical clustering:

K-means Clustering:

MacQueen suggests the term K-means for describing an algorithm of his that assigns each item to the cluster having the nearest centroid. The process is composed of three steps:

1. Partition the items into K initial clusters.

2. Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. (Distance is usually computed using Euclidean distance with either standardized or unstandardized observations.) Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

3. Repeat Step 2 until no more reassignments take place.

Optimal number of clusters:

To find the optimal number of clusters for k-means clustering we draw a plot where the Number of Clusters corresponds with Within Groups Sum of Squares. If we see the below plot, After the number 4 cluster, the line is getting narrow. So, we decided our optimal number of clusters is 4.

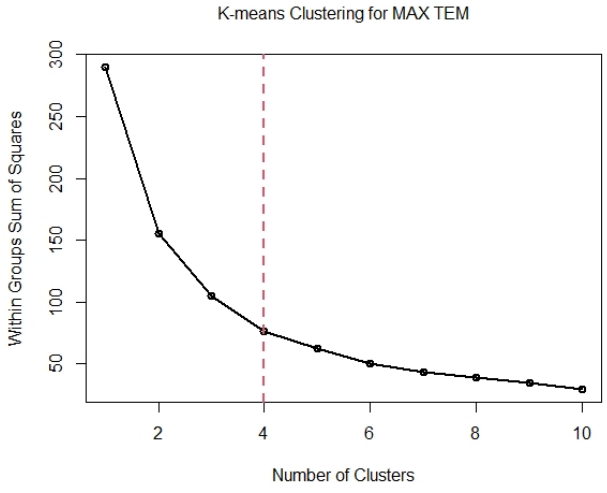

Fig: Finding Optimal Number of Clusters

After applying the k-means algorithm, our results are given below:

Clusters:

Cluster-1	Cluster-2	Cluster-3	Cluster-4
Chuadanga	Dinajpur	Chittagong	Barisal
Ishwardi	Mymensingh	Cox's Bazar	Bhola
Jessore	Rangpur	Feni	Bogra
Khulna	Sayedpur	Hatiya	Chadpur
M. Court		Khepupara	Comilla
Mongla		Kutubdia	Dhaka
Rajshahi		Sandwip	Faridpur
Satkhira		Sitakunda	Madaripur
		Sylhet	Patuakhali
		Teknaf	Rangamati
			Srimangal
			Tangail

Total Sum of Squares: 289.7985

Within-cluster Sum of Squares:

17.46875	6.79000	28.75200	23.50000

Total within-cluster sum of squares: 76.51075

Between-cluster Sum of Squares: 213.2878

Clustering Visualization for MAX TEM:

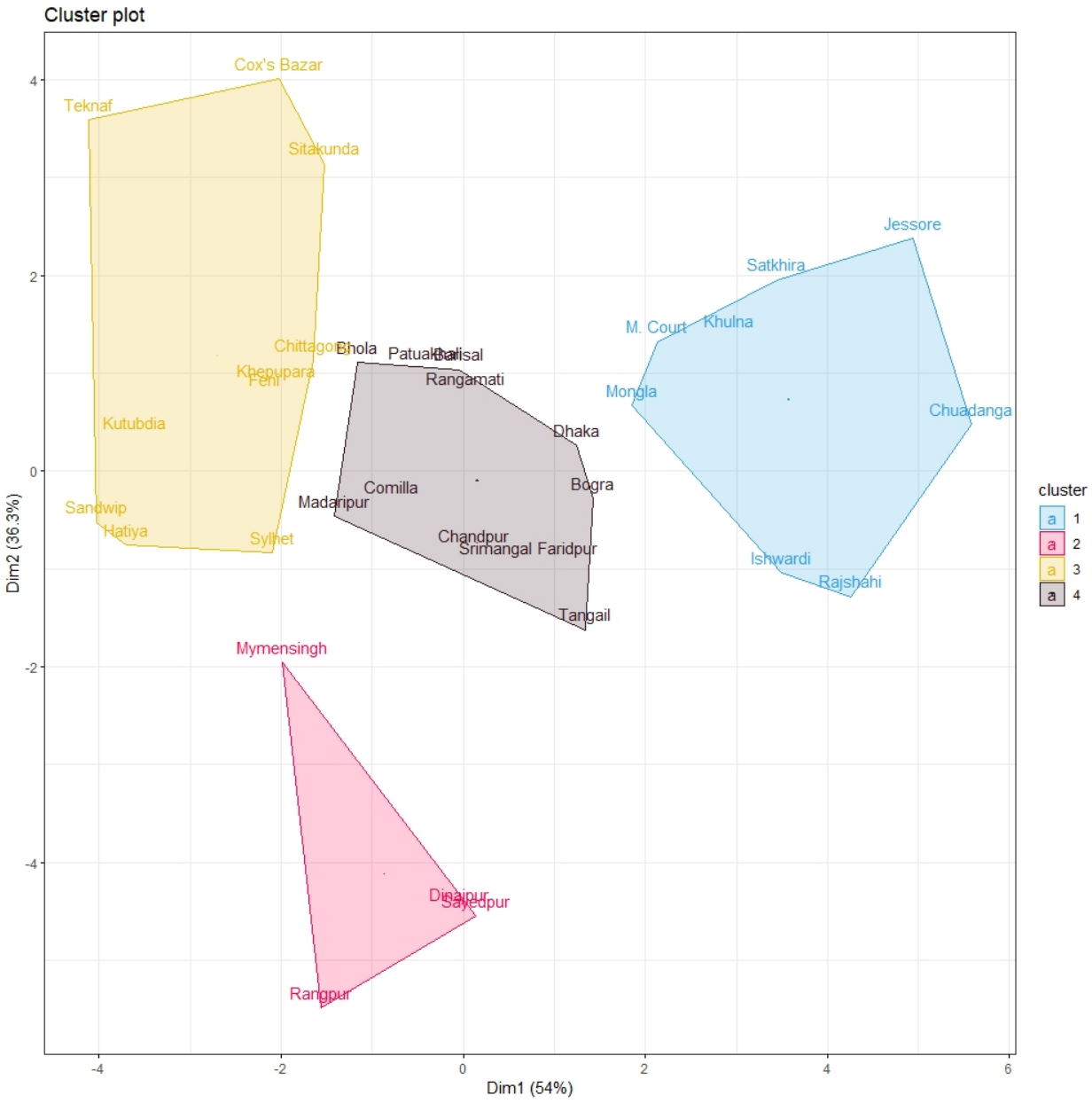

Fig: Clusters Visualization for Maximum Temperature

Table B: Minimum Temperature

Table B represents the minimum temperature of different locations in Bangladesh for some period. Here, 34 stations with some period and their monthly basis minimum temperature. Based on our problem, our variables are 12 months and observations are row-wise stations.

Hierarchical Clustering:

After calculating the distance matrix, now we are calculating the single linkage, complete linkage, average linkage, and ward linkage for Minimum Temperature.

Single Linkage:

Comment on dendrogram and Cluster allocated:

In the dendrogram below, we have four subgroups. We merged four subgroups into two clusters. Only Sylhet in One cluster and the second cluster combining other 33 stations.

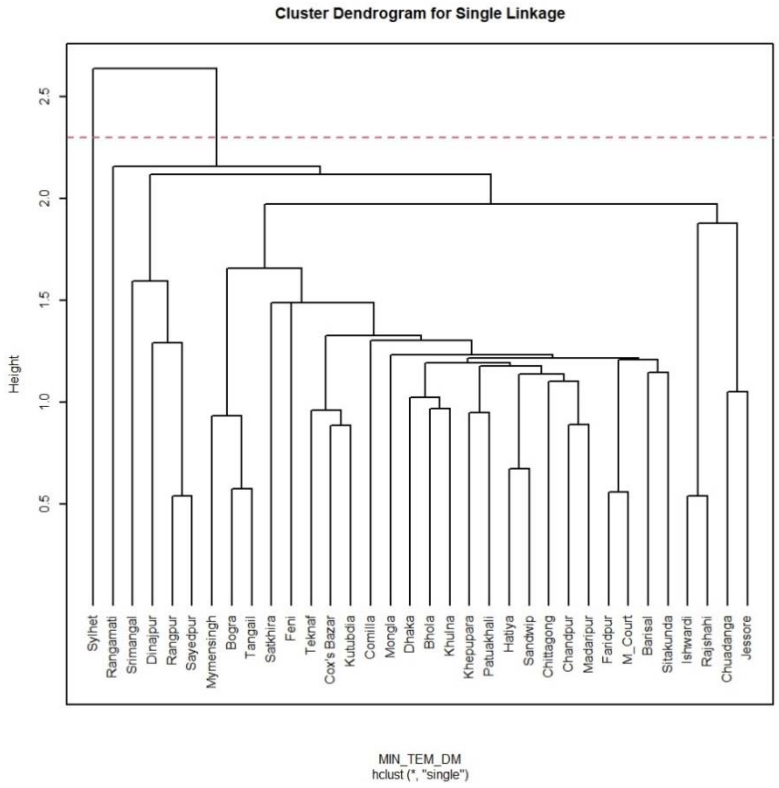

Fig: Single Linkage Cluster Dendrogram for Minimum Temperature

Complete Linkage:

Comment on dendrogram and Cluster allocated:

We have four subgroups from the below dendrogram. We merged four subgroups into two clusters. One cluster combines Srimangal, Dinajpur, Rangpur, Sayedpur, Mymensingh, Bogra, Tangail, Ishwardi, Rajshahi, Chuadanga, Jessore and the second cluster combining other 23 stations.

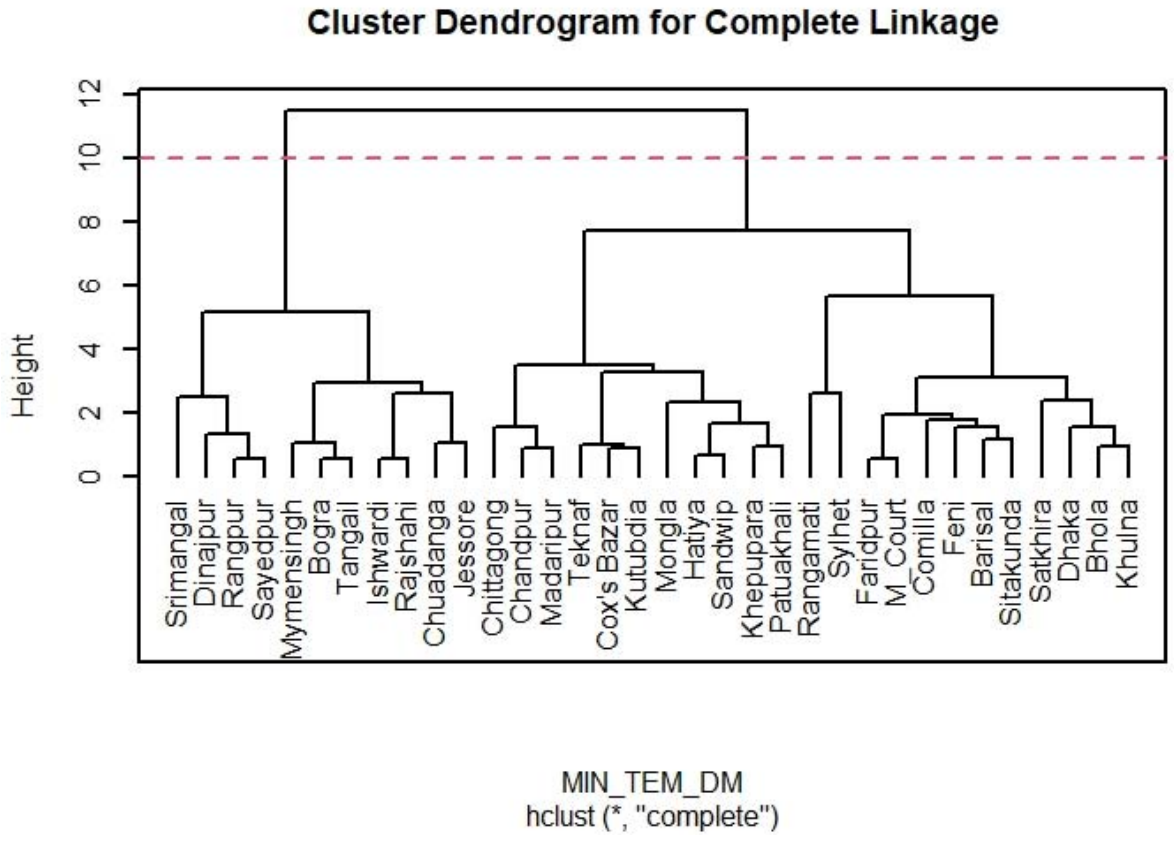

Fig: Complete Linkage Cluster Dendrogram for Minimum Temperature

Average linkage:

Comment on dendrogram and Cluster allocated:

We have three subgroups from the below dendrogram. we merged three subgroups into two clusters. One cluster combines Srimangal, Dinajpur, Rangpur, Sayedpur and the second cluster combining other 30 stations.

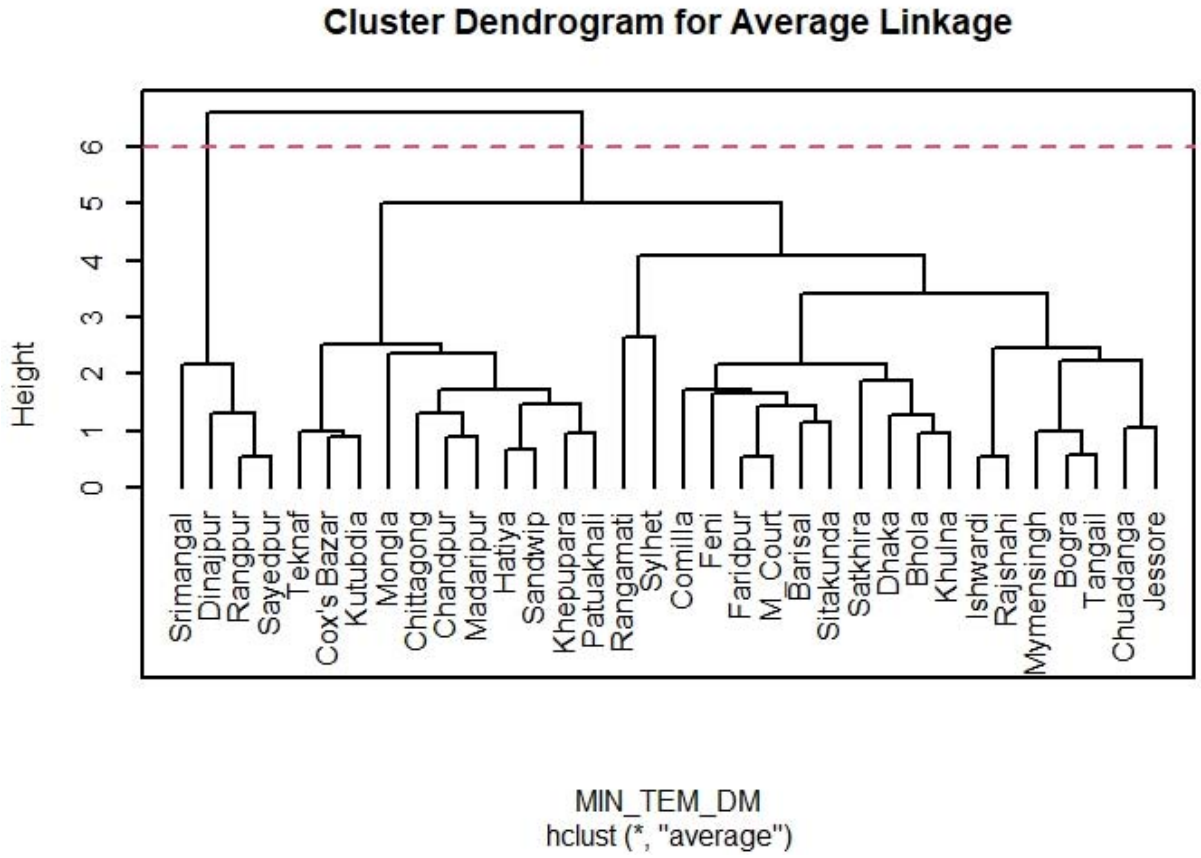

Fig: Average Linkage Cluster Dendrogram for Minimum Temperature

Ward's linkage:

Comment on dendrogram and Cluster allocated:

Here, we have four subgroups from the below dendrogram. We merged four subgroups into two clusters. One cluster combines Srimangal, Dinajpur, Rangpur, Sayedpur, Mymensingh, Bogra, Tangail, Ishwardi, Rajshahi, Chuadanga, Jessore and the second cluster combining other 23 stations.

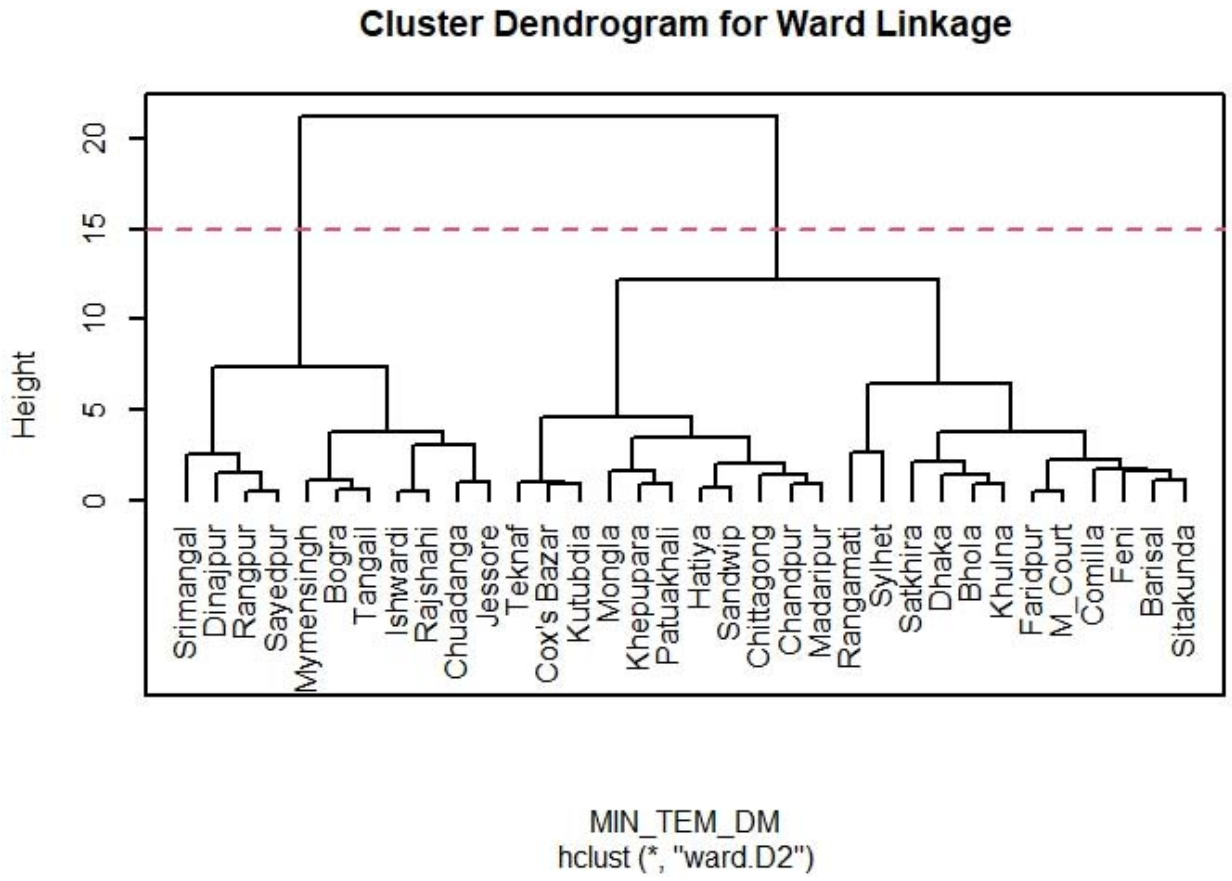

Fig: Ward's Linkage Cluster Dendrogram for Minimum Temperature

Non-hierarchical clustering:

K-means Clustering:

Optimal number of clusters:

If we see the below plot, After the number 5 cluster, the line is getting narrow. So, we decided our optimal number of clusters is 5.

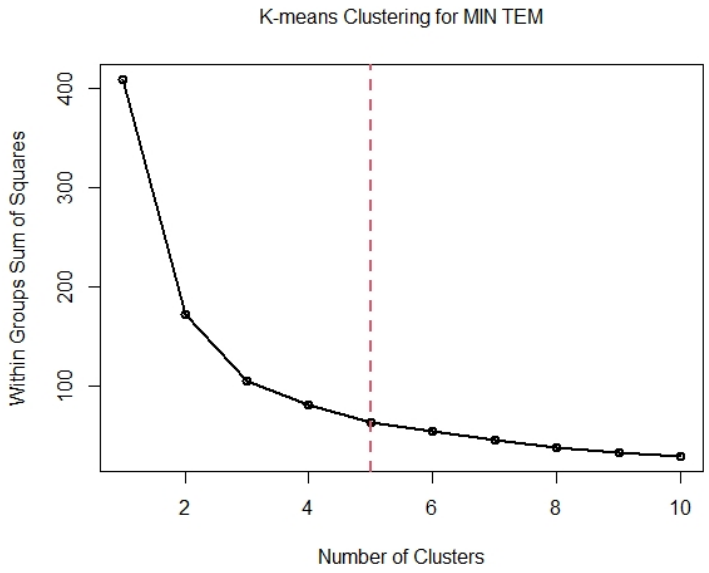

Fig: Finding Optimal Number of Clusters

After applying the k-means algorithm, our results are given below:

Cluster means:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	12.33	15.57	20.32	23.68	24.82	25.87	25.84	25.96	25.59	23.76	19.05	14.00
2	12.95	15.00	19.00	21.90	23.45	24.75	25.00	25.05	24.60	23.10	19.25	14.90
3	11.14	14.17	18.80	22.97	24.36	25.81	26.10	26.24	25.63	23.21	17.97	12.97
4	14.18	16.87	21.12	24.17	25.29	25.85	25.63	25.71	25.53	24.37	20.47	15.95
5	10.50	13.08	17.40	21.20	23.20	25.10	25.73	26.00	25.15	22.42	17.05	12.55

Clusters:

Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster-5
Barisal	Rangamati	Bogra	Chandpur	Dinajpur
Bhola	Sylhet	Chuadanga	Chittagong	Rangpur
Comilla		Ishwardi	Cox's Bazar	Srimangal
Dhaka		Jessore	Hatiya	Sayedpur
Faridpur		Mymensingh	Khepupara	
Feni		Rajshahi	Kutubdia	
Khulna		Tangail	Madaripur	
M. Court			Mongla	
Satkhira			Patuakhali	
Sitakunda			Sandwip	
			Teknaf	

Total Sum of Squares: 409.2535

Within-cluster Sum of Squares:

17.19700	3.48500	13.62000	23.47636	4.57250

Total within-cluster sum of squares: 62.35086

Between-cluster Sum of Squares: 346.9027

Clustering Visualization for MIN TEM:

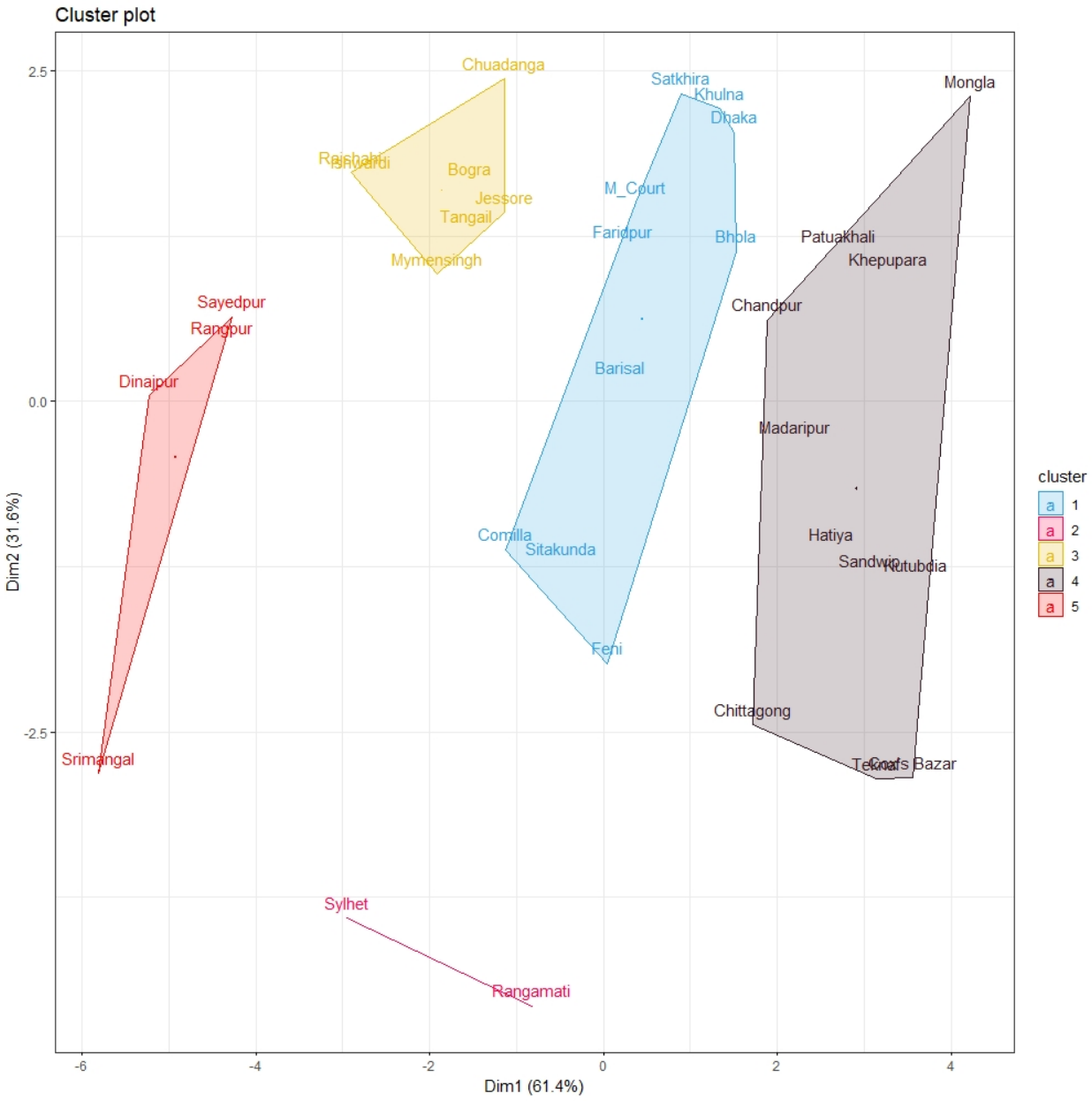

Fig: Clusters Visualization for Minimum Temperature

Comparison between Hierarchical and Non-hierarchical clustering results in Maximum temperature and Minimum temperatures:

Hierarchical clustering comparison between Maximum temperature and Minimum temperatures:

Maximum Temperature	Minimum Temperature
Single Linkage Comparison	
Some subgroups are distributed into two clusters.	Four subgroups are distributed into two clusters.
Complete Linkage Comparison	
Six subgroups are distributed into two clusters.	Four subgroups are distributed into two clusters.
Average Linkage Comparison	
Seven subgroups are distributed into two clusters.	Three subgroups are distributed into two clusters.
Ward's Linkage Comparison	
Five subgroups are distributed into two clusters.	Four subgroups are distributed into two clusters.

Non-hierarchical clustering comparison between Maximum temperature and Minimum temperatures:

K-means Clustering Comparison		
Basis for comparison	Maximum Temperature	Minimum Temperature
1. Optimal number of clusters, k	k = 4.	k = 5.
2. Number of stations in each cluster	Cluster-1: 8 Cluster-2: 4 Cluster-3: 10 Cluster-4: 12	Cluster-1: 10 Cluster-2: 2 Ckuster-3: 7 Cluster-4: 11 Cluster-5: 4
3. Total Sum of Squares	289.7985	409.2535
4. Within-cluster Sum of Squares	17.46875 6.79000 28.75200 23.50000	17.19700 3.48500 13.62000 23.47636 4.57250
5. Total within-cluster sum of squares	76.51075	62.35086
6. Between-cluster Sum of Squares	213.2878	346.9027

THE END