Clustering for Weather Pattern and Forecast Similarity on Bangalore Meteorological Data

Subha Chakraborty Roll No: MT2024156
Jyoti Singh Roll No: MT2024072

13th December 2024

Abstract

This report investigates the application of clustering techniques to analyze Bangalore's meteorological data from 2010 to 2024, aiming to identify recurring weather patterns and enhance the accuracy of weather forecasting.

The report explores two popular clustering methods, K-means and fuzzy C-means, in conjunction with Principal Component Analysis (PCA) for dimensionality reduction. A representative dataset from 2022 to 2024, was used for clustering. The evaluation of cluster validity is conducted through metrics such as "Within Cluster Sum of Squares" and "Silhouette Width," which assess the compactness and separation of clusters. The results show that fuzzy C-means, with a membership parameter of 2.9, provide superior performance in identifying meaningful clusters, demonstrating their potential over traditional K-means.

Furthermore, we were able to categorize Bangalore's weather into five distinct patterns, each corresponding to specific climatic conditions, which can significantly improve forecasting techniques by identifying similar weather scenarios. These findings highlight the importance of how clustering can be used for weather prediction and forecasting tailored to Bangalore's unique climate.

Keywords: Clustering, K-means, fuzzy C-means, Principal Component Analysis (PCA), Weather Forecasting, Bangalore.

Contents

1 Introduction **4**
1.1 Problem Statement 4
1.2 Motivation and Approach 4
1.3 The Data 4

2 Methodology **6**
2.1 Exploratory Data Analysis before preprocessing 6
2.2 Data Preprocessing 10
2.2.1 Principal Component Analysis 10
2.3 Clustering Algorithms 13
2.3.1 K-means 13
2.3.2 fuzzy C-means 15
2.3.3 WCSS Elbow Method fails for different values of membership for Fuzzy C-means 17

3 Results and Conclusion **22**

1 Introduction

1.1 Problem Statement

The problem statement involves using Bangalore's meteorological data, using clustering techniques, mainly K-Means and fuzzy C-means to analyze and identify recurring weather patterns observed in the city.

1.2 Motivation and Approach

Weather data analysis,in the context of weather forecasting is a fascinating domain with numerous approaches. Ours is an effective approach dealing with analyzing the data using clustering techniques and segmenting the weather data into distinct patterns.By doing this we can identify recurring weather conditions that share similar characteristics. This method allows for forecasting the future behavior of weather features, such as temperature or humidity, based on the patterns observed in historical data. Instead of predicting specific outcomes for each instance of weather data, clustering provides a more generalized approach, making it possible to forecast unknown weather characteristics by indentifying the weather pattern.

1.3 The Data

The meteorological data of Bangalore, with which we are dealing, range from 2010 to 2024, and have relevant characteristics such as temperature, humidity, cloud cover, dew point, wind speed, wind direction, rainfall,snow and so on. Figure 1 shows a plot of the relevant features.

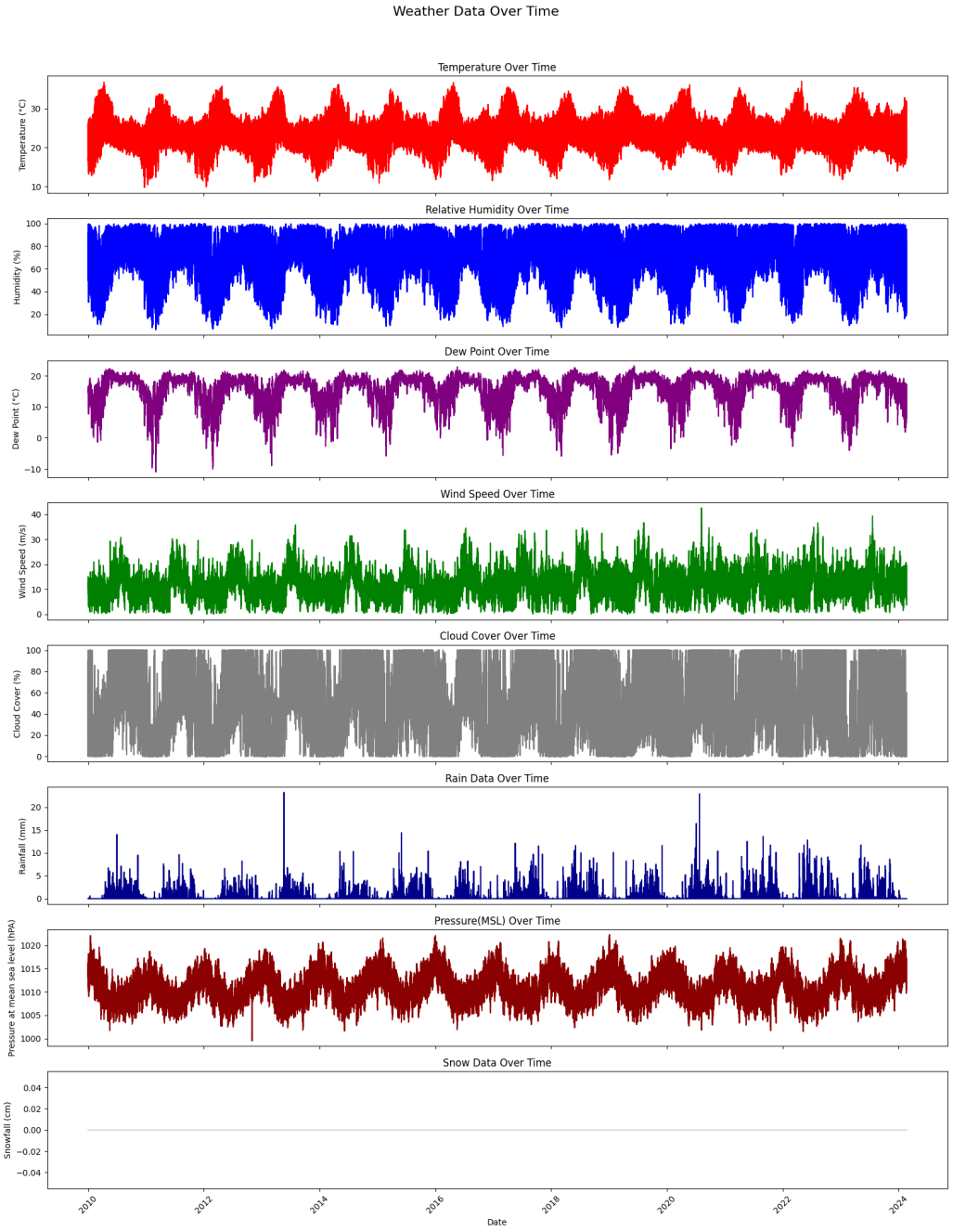

Figure 1: Plot of the relevant meteorological features of Bangalore from 2010 to 2024.

2 Methodology

This section describes the Exploratory Data Analysis(EDA) before the pre-processing, data preprocessing steps, clustering algorithms used (K-Means,fuzzy C-means), and evaluation metrics applied to assess the quality of clusters. The methodology ensures that the derived clusters are meaningful and actionable for weather forecasting.

2.1 Exploratory Data Analysis before preprocessing

This section deals with explaining the EDA steps performed to derive insights from the data and then proceed further with said insights.The first thing we did was make a observation that the data ,more or less,repeated after some cycles in the year range. We decided to work then only a snippet of the data from 2022 to 2024, which captured the repeating pattern.The Figure 2 on the next page captures the plot of said data.

The features snow and snowdepth where dropped from data beforehand as from the plot of the original data in Figure 1, it was obvious that snow and related features where irrelevant to bangalore.

A tabular representation of basic information of the respective feature can be found below in Table 1:

No	Column	Non-Null Count	Null Count	Dtype	Min	Max
1	temperature_2m	18744	0	float64	11.796	36.996
2	relative_humidity_2m	18744	0	float64	9.9298725	100.0
3	dew_point_2m	18744	0	float64	-4.054	21.846
4	apparent_temperature	18744	0	float64	10.869806	37.499794
5	precipitation	18744	0	float64	0.0	12.8
6	rain	18744	0	float64	0.0	12.8
7	pressure_msl	18744	0	float64	1001.5	1021.5
8	surface_pressure	18744	0	float64	903.1194	919.80994
9	cloud_cover	18744	0	float64	0.0	100.0
10	cloud_cover_low	18744	0	int64	0	100
11	cloud_cover_mid	18744	0	int64	0	100
12	cloud_cover_high	18744	0	int64	0	100
13	wind_speed	18744	0	float64	0.5091169	39.345543

Table 1: Summary of Meteorological Data Columns

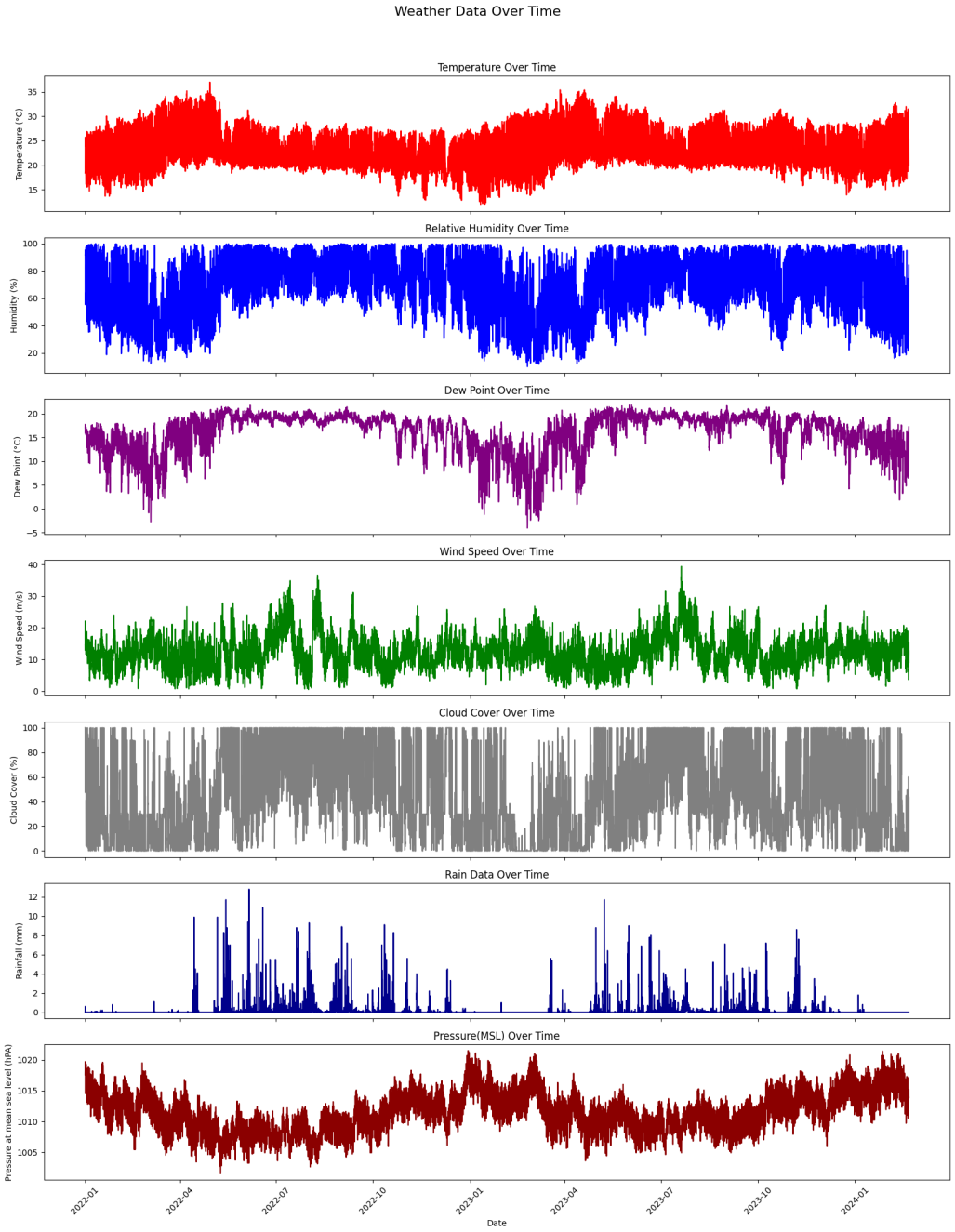

Figure 2: Plot of the relevant meteorological features of Bangalore from 2022 to 2024.

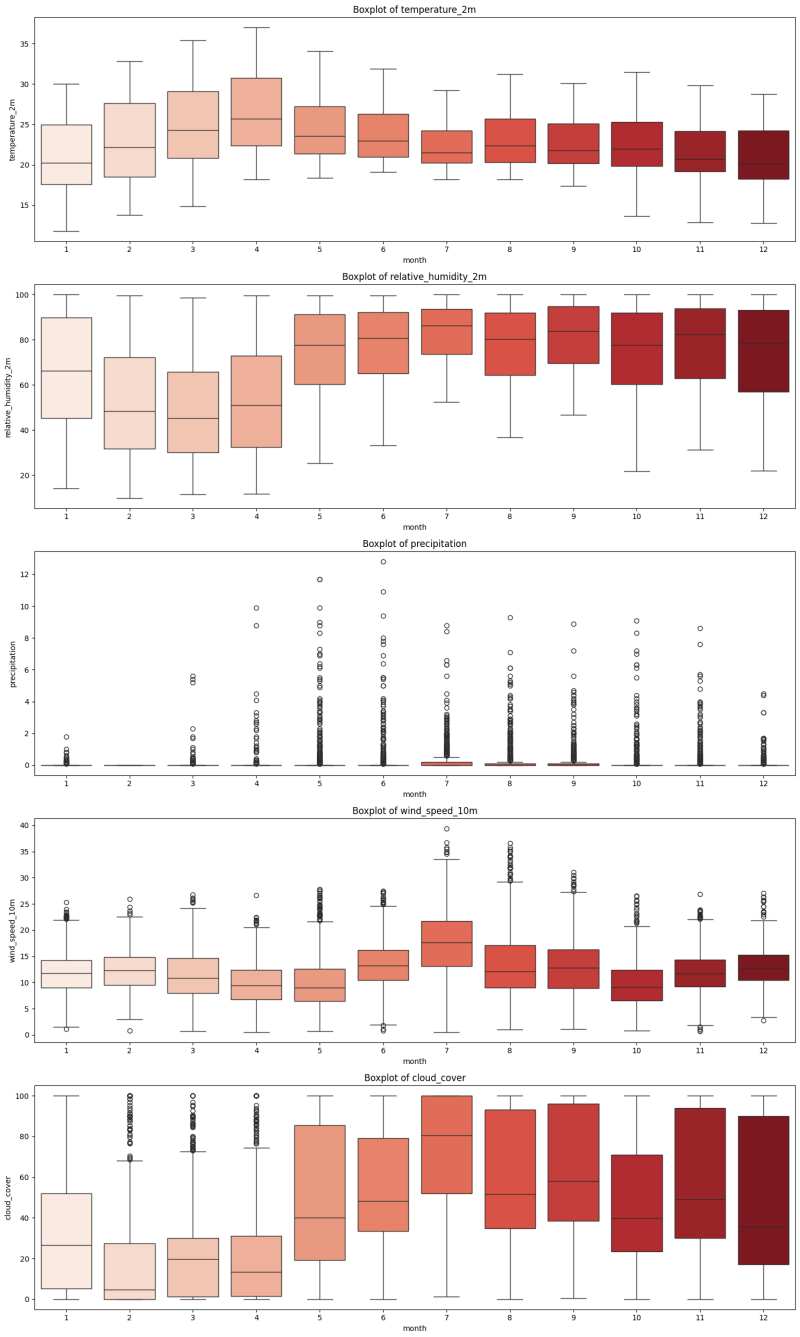

Figure 3: BoxPlot of some of the features of the weather data

To analyze the data distribution, we did boxplots(as shown in Figure 3) of some

of the features.Note that precipitation column gives same information as the rain column,as observed by the correlation matrix in Figure 4. We observed that temperature and wind speed exhibit fairly symmetric distributions around the monthly medians, where as humidity and cloud cover shows slight skewness(both left and right) across various months.The precipitation data shows a more right-skewed pattern, with most months having low to moderate levels but a few months experiencing much higher spikes. So we made the observation that precipitation levels are more variable and unpredictable compared to the other climate metrics in the dataset.

The Figure 4 represents a correlation matrix of the features of weather we made to make further analysis.

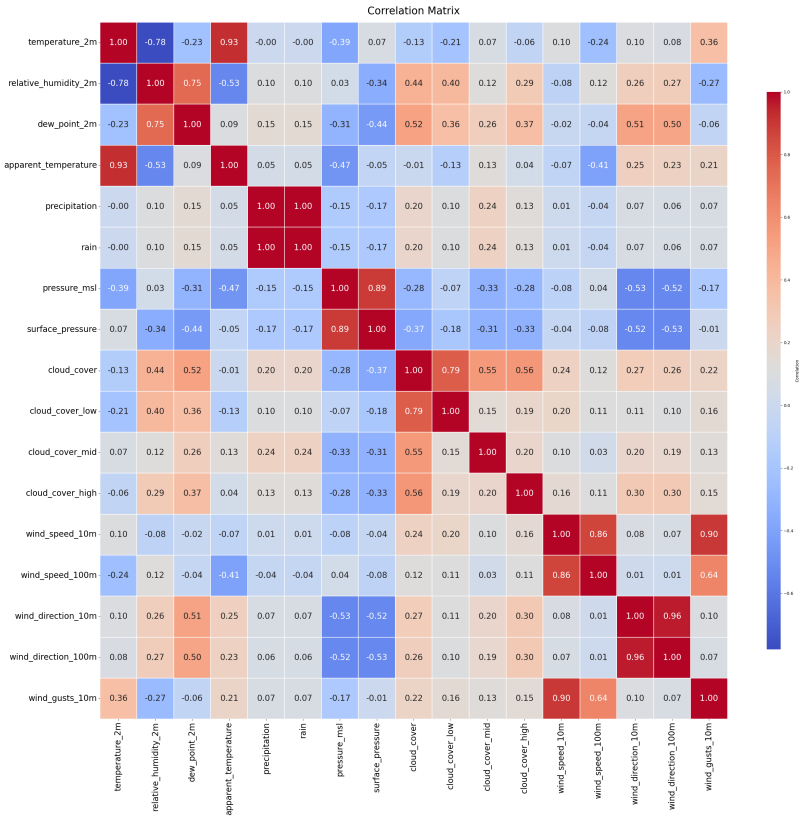

Figure 4: Correlation Matrix of weather data

2.2 Data Preprocessing

The first step we did was standardize the data, by implementing the below formula in python code:

$$x' = \frac{x - \mu}{\sigma}$$

Where:

- x' is the standardized value of x
- μ is the mean of the column x
- σ is the standard deviation of the column x

Then we decided, that the best approach forward would be to perform Principal Component Analysis(PCA) and project the data into only the required number of components. Again as a result of PCA, the components of data are also mostly linearly independent(although this is a side-effect of PCA not its primary purpose).

2.2.1 Principal Component Analysis

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

$$\mathbf{v}_k =_{\mathbf{v}} \left(\mathbf{v}^T\mathbf{C}\mathbf{v}\right)$$

$$\mathbf{X}_{PCA} = \mathbf{X}\mathbf{V}_k$$

Where:

- X is the original mean-centered data
- $\mathbf{X}_{PCA}$is the component projected data
- $\mathbf{C}$ is the covariance matrix of the data.
- $\mathbf{v}_k$ is the eigenvector corresponding to the k^{th} principal component.
- $\mathbf{v}$ is a candidate eigenvector.

To perform PCA, it is important to analyze how many components is actually needed. In other words how many components are enough to capture and explain the information(variance) of the data. To approach this we did two plots: Scree Plot and Cumulative Sum Plot.

Both these plots analyzed how much of the variance ratio was explained with respective to the principal components.

$$\text{variance\_ratio} = \frac{\lambda_k}{\sum_{i=1}^{n} \lambda_i}$$

Where:

- λ_k is the eigenvalue corresponding to the k^{th} principal component.
- λ_i is the eigenvalue for the i^{th} principal component.
- n is the total number of principal components.

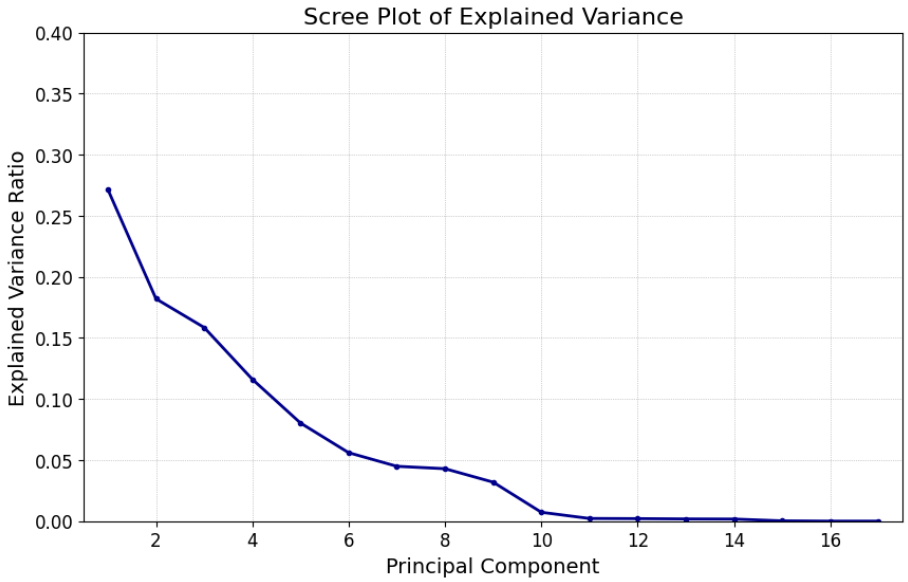

Figure 5: Screeplot

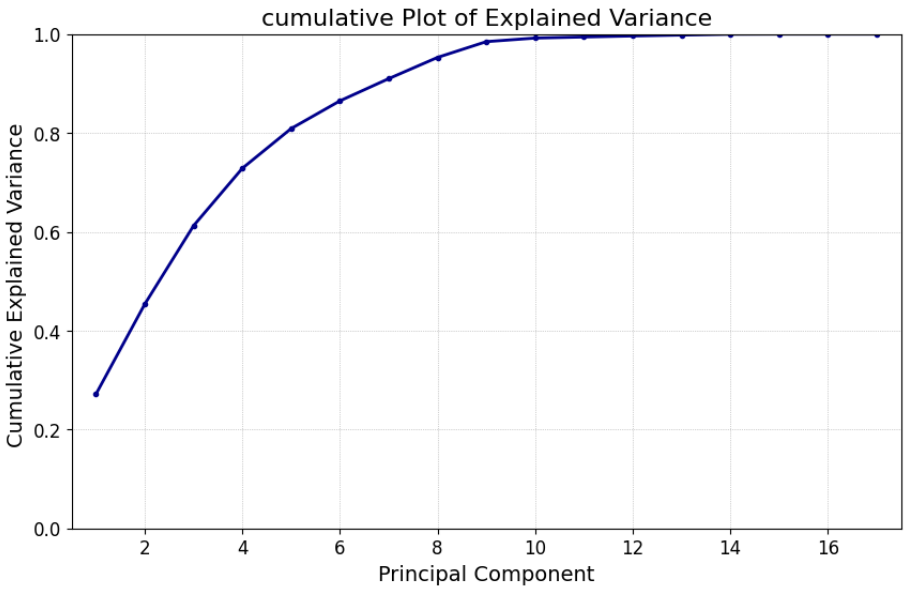

Figure 6: Cumulativesum Plot for variance ratio vs components

As observed from both the plots in Figure:5 and Figure:6, the number of components which correctly captured and explained the data(via variance) was 10. Hence, we proceeded to project data into 10 components.Also from Figure:7 we can conclude that the components are linearly independent from correlation matrix.

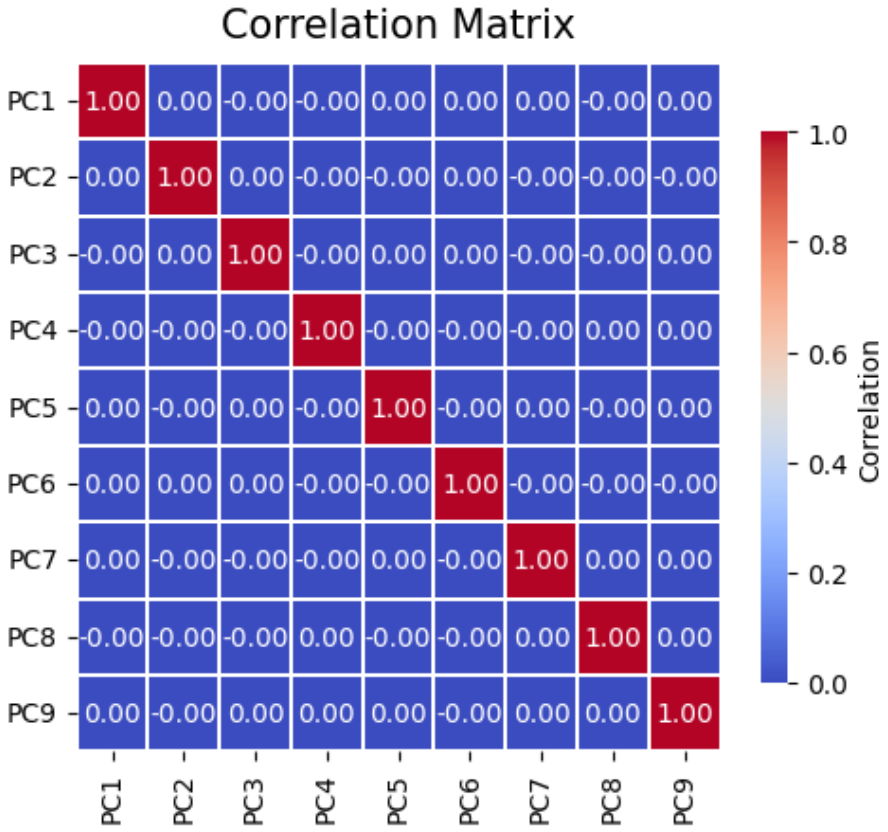

Figure 7: Correlation matrix of the 10 principal components

2.3 Clustering Algorithms

Our First Clustering Algorithm we wanted to experiment with ,was K-means,

2.3.1 K-means

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}_{\{c_i=k\}} \|\mathbf{x}_i - \mathbf{v}_k\|^2$$

Where:

- N is the number of data points in the dataset.
- K is the number of clusters.
- $\mathbf{x}_i$ is the i^{th} data point.

- $\mathbf{v}_k$ is the k^{th} centroid.
- c_i is the index of the centroid closest to the i^{th} point.
- $\mathbf{1}_{\{c_i=k\}}$ is an indicator function that is 1 if data point $\mathbf{x}_i$ belongs to cluster k, otherwise 0.

As we all know K-means is a hard clustering algorithm, which will give me K clusters. Now the problem statement kind of demands us to find what the optimal K is, as we don't know how many possible broad patterns are there in Bangalore's weather. We can determine this by measuring how good is the quality of clusters being made with different values of k. To measure the quality one of the metrics we used was 'Within Cluster Sum of Squares'.

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{v}_k\|^2$$

Where:

- K is the number of clusters.
- C_k represents the points belonging to the k^{th} cluster.
- $\mathbf{x}_i$ is a data point in the cluster C_k.
- $\mathbf{v}_k$ is the centroid of the k^{th} cluster.
- $\|\mathbf{x}_i - \mathbf{v}_k\|$ is the Euclidean distance between the data point $\mathbf{x}_i$ and the centroid $\mathbf{v}_k$.

We plotted a graph for WCSS vs K clusters and our goal was to find at which cluster does the WCSS stop improving, which means that said number of clusters was optimal.Basically the Elbow Method.

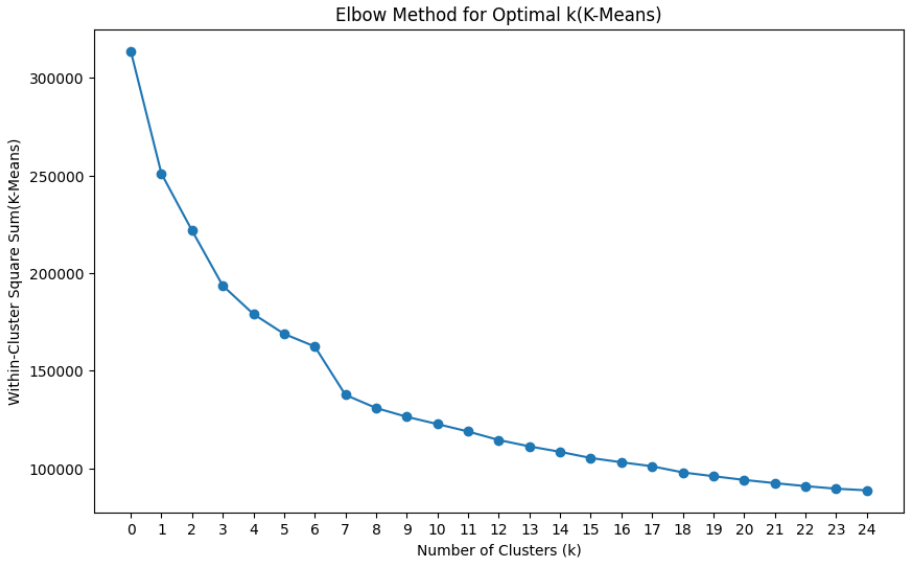

Figure 8: Elbow method for K-means

From the graph Figure:8 it is clear that 7 clusters can be formed. However, due to the hard clustering nature of K-means,this many weather pattern might not be accurate, as hard clustering misunderstands how real life weather instance data actually is. A better approach would be a soft clustering way. But this is just our assumption.We wanted to see mathematical evidence of that.

2.3.2 fuzzy C-means

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} (u_{ik}^{m}) \left\| \mathbf{x}_i - \mathbf{v}_k \right\|^2$$

Where:

- N is the number of data points.
- K is the number of clusters.
- $\mathbf{x}_i$ is the i^{th} data point.
- $\mathbf{v}_k$ is the k^{th} centroid.
- u_{ik} is the membership degree of the i^{th} data point in the k^{th} cluster.
- m is the fuzziness parameter.

$$\mathbf{v}_k = \frac{\sum_{i=1}^{N} u_{ik}^m \mathbf{x}_i}{\sum_{i=1}^{N} u_{ik}^m}$$

Where:

- $\mathbf{v}_k$ is the centroid of the k^{th} cluster.
- u_{ik} is the membership degree of the i^{th} data point in the k^{th} cluster.
- m is the fuzziness parameter.
- $\mathbf{x}_i$ is the i^{th} data point.

$$u_{ik} = \frac{1}{\sum_{j=1}^{K} \left(\frac{\|\mathbf{x}_i - \mathbf{v}_k\|}{\|\mathbf{x}_i - \mathbf{v}_j\|} \right)^{2/(m-1)}}$$

Where:

- u_{ik} is the membership degree of the i^{th} data point in the k^{th} cluster.
- m is the fuzziness parameter.
- $\mathbf{x}_i$ is the i^{th} data point.
- $\mathbf{v}_k$ and $\mathbf{v}_j$ are centroids of clusters k and j respectively.

We then proceeded with another algorithm similiar to K-means, but instead of hard clustering, it went with a soft clustering approach, considering the membership of the weather instances to clusters, rather than committing the weather instance to a particular cluster. So we further analyzed the C-means approach using a within cluster sum of squares approach with a membership of 1.5 and contrasted with K-means.

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ik}^m \left\| \mathbf{x}_i - \mathbf{v}_k \right\|^2$$

Where:

- $\mathbf{x}_i$ is the i^{th} data point.
- $\mathbf{v}_k$ is the centroid of the k^{th} cluster.
- u_{ik} is the membership degree of the i^{th} data point in the k^{th} cluster.
- m is the fuzziness parameter.

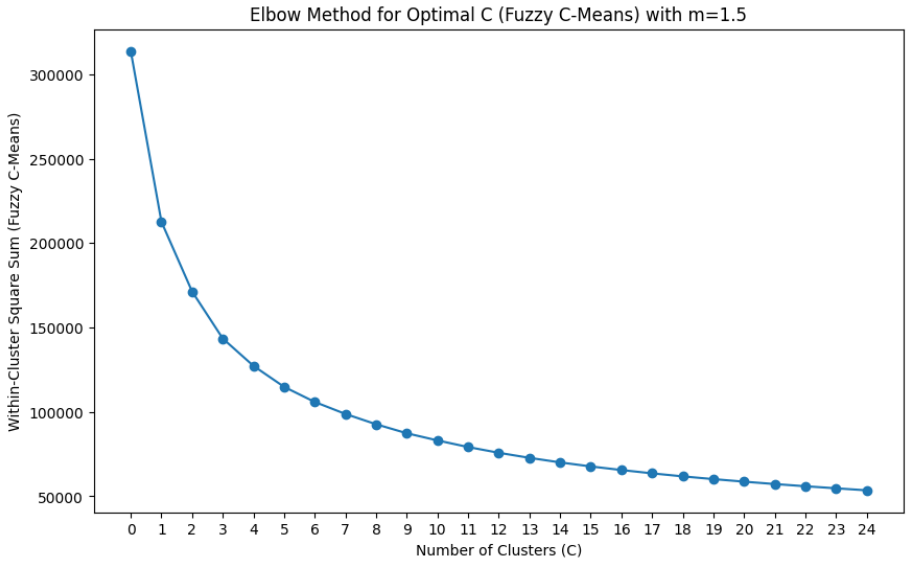

Figure 9: Elbow method for Fuzzy C-means

In Figure: 9, we can that even for smaller values of C(number of clusters), we get the same WCSS we get for optimal K in K-means. It means that cohesion of the clusters formed in K is not good as the square sums within the clusters is quite large. The clusters are not very good.

2.3.3 WCSS Elbow Method fails for different values of membership for Fuzzy C-means

Now although one might naively assume that with different values of membership, the one which gives the least WCSS, I will go with that, but that approach is wrong. Although WCSS helps us understand cohension, good clusters also need good separation, and incase of soft clustering we have to worry about any point affecting the cohension value of a cluster, not necessarily the one it is most likely to belong to hence. Hence, for that a better metrics is needed. Below figures show the elbow method for different membership values showing why it fails.

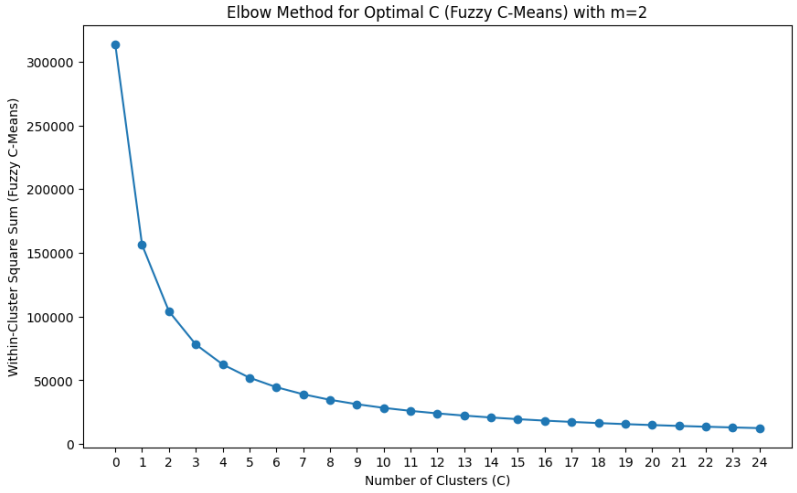

Figure 10: Elbow method for Fuzzy C-means for membership 2

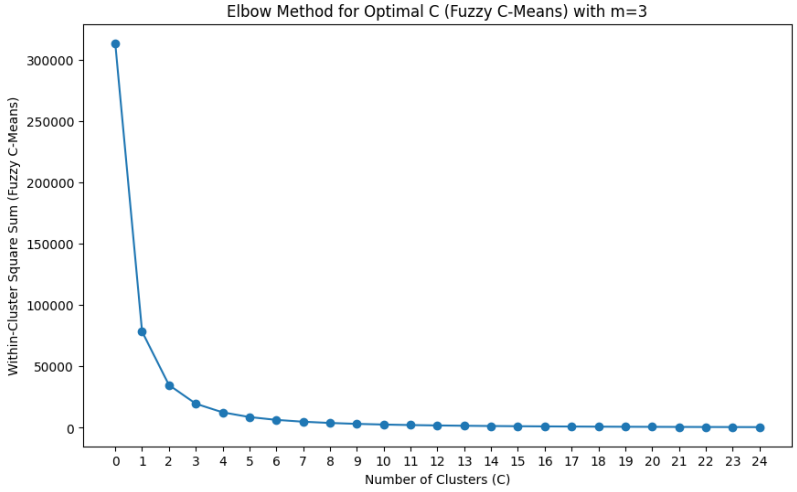

Figure 11: Elbow method for Fuzzy C-means for membership 3

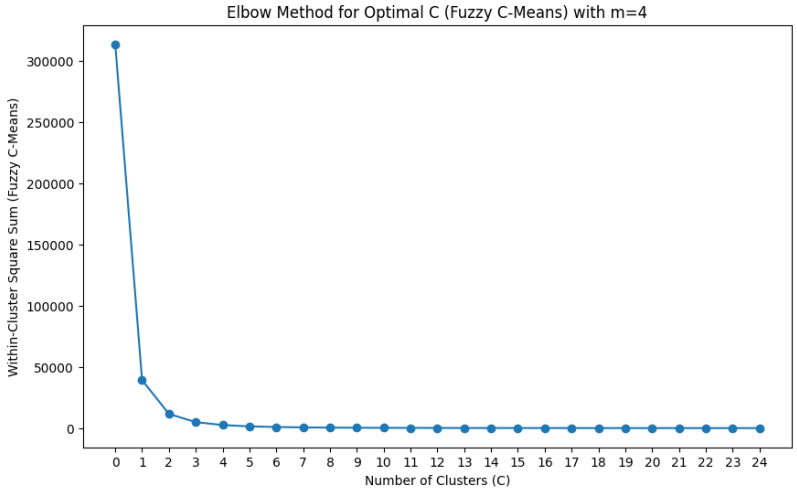

Figure 12: Elbow method for Fuzzy C-means for membership 4

So a better metric is Silhoutte Score, which considers both the separation and cohesion of clusters, where there is more cohesion withion clusters and clusters are well separated. Higher the score, better is the clustering.

$$s_i = \begin{cases} \frac{b_i - a_i}{\max(a_i, b_i)} & \text{if } \max(a_i, b_i) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Where:

- a_i is the mean intra-cluster distance, calculated using memberships and the centroid of the cluster to which $\mathbf{x}_i$ belongs.
- b_i is the mean nearest-cluster distance, considering memberships and centroids of other clusters.

$$a_i = \frac{\sum_{c=1}^{K} \left(u_{ic}^m\right) \|\mathbf{x}_i - \mathbf{v}_c\|}{\sum_{c=1}^{K} u_{ic}^m}$$

Where:

- u_{ic} is the membership degree of $\mathbf{x}_i$ in cluster c.
- $\mathbf{v}_c$ is the centroid of cluster c.
- m is the fuzziness parameter.

$$b_i = \min_{c' \neq c} \left(\frac{\sum_{j=1}^{N} u_{jc'}^{m} \left\| \mathbf{x}_j - \mathbf{v}_{c'} \right\|}{\sum_{j=1}^{N} u_{jc'}^{m}} \right)$$

Where:

- $u_{jc'}$ is the membership degree of data point $\mathbf{x}_j$ in cluster c'.
- $\mathbf{v}_{c'}$ is the centroid of cluster c'.
- m is the fuzziness parameter.

$$\text{Silhouette Score} = \frac{1}{N} \sum_{i=1}^{N} s_i$$

Where:

- N is the total number of data points.
- s_i is the silhouette score for the i^{th} data point.

Using this fuzzy Silhoutte score, we plotted a graph for the fuzzy Silhoutte score vs membership(fuzziness) paratmeter, with a good enough inference considered 5 clusters from the previous graphs. (An inference can be made from elbow methods of K-means and C-means that 5 is a good enough cluster number, although elbow method cannot give us membership).

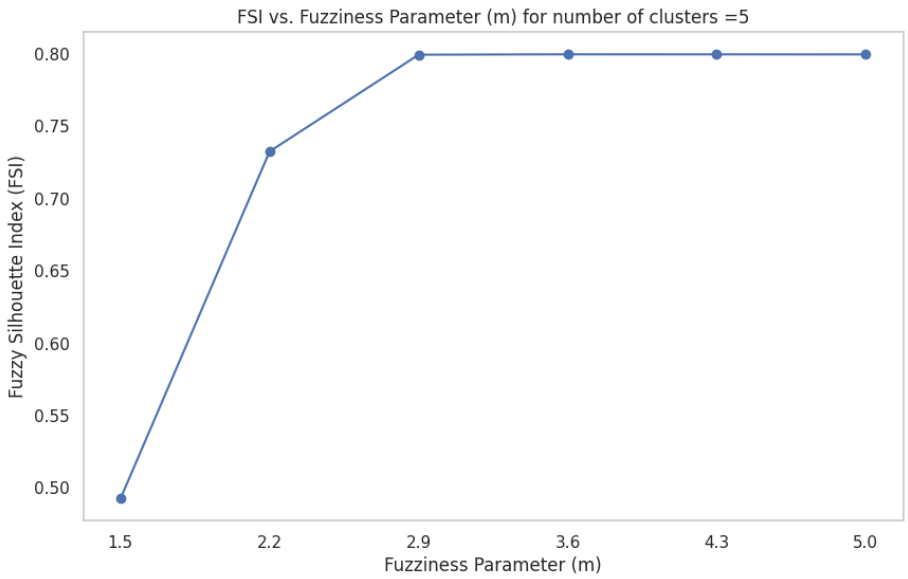

Figure 13: Fuzzy Silhoutte Index vs fuzzyness(membership)

From Figure 13, we can conclude that a membership of 2.9 is good enough for proceeding with the clustering.

3 Results and Conclusion

Finally we can cluster Bangalore's weather into 5 distinct clusters a.k.a weather patterns. Now, it's time to analyze what the mean. Observing the various feature values of the clusters, following generalizations can be made about the weather patterns.

Cluster	**Temperature (°C)**	**Humidity (%)**	**Wind Speed (m/s)**	**Rainfall (mm)**
0	11.80–37.00 (23.96)	9.93–99.69 (53.36)	0.72–26.76 (11.49)	0.0–1.7 (0.007)
1	14.80–35.20 (22.90)	17.56–100.00 (79.70)	0.51–39.35 (12.82)	0.0–12.8 (0.209)
2	24.30–26.65 (25.50)	48.41–66.26 (57.74)	11.62–21.99 (16.47)	0.0–0.3 (0.071)
3	12.05–26.40 (18.50)	43.22–100.00 (89.05)	1.53–26.83 (12.28)	0.0–3.9 (0.022)
4	22.80–35.40 (29.39)	16.66–77.45 (42.27)	1.84–25.96 (11.10)	0.0–2.8 (0.043)

Table 2: Cluster characteristics including ranges and averages (in parentheses).

Based on the generalizations, we decided our cluster types to be as follows:

Cluster	**Type**
0	Mild, Dry, with Sparse Clouds
1	Mild, Humid, A Little Cloudy with Frequent Rain
2	Hot, Cloudy with No Significant Rain
3	Colder, Humid, A Little Cloudy with Occasional Rain
4	Very Hot, with Sparse Clouds and Less Frequent Rain

Table 3: Cluster types based on weather characteristics.

We also wanted to know which weather pattern can be viewed in which month and what are the chances of it occuring. Hence, we plotted a graph to observe it:

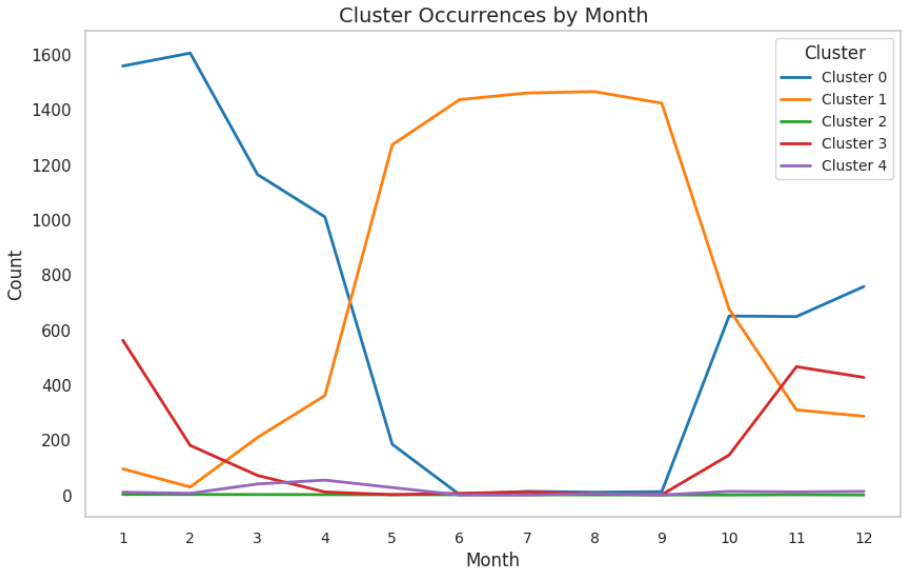

Figure 14: weather patterns(clusters)counts vs month

Hence, you can see from the graph which weather type predominantly occurs in which month, for example, from the months of June to September, expect a mild humid climate with frequent rain. Therefore, this report highlights how this project can be used to identify climate of Bangalore across various months, and if certain meteorological data for a weather instance are known, we can correlate it to one of the weather types and make an educated guess about other meteorological data for that instance.That is how forecast via similarity. Hence, we believe whatever goals we had set for ourselves with our project we were able to achieve them.