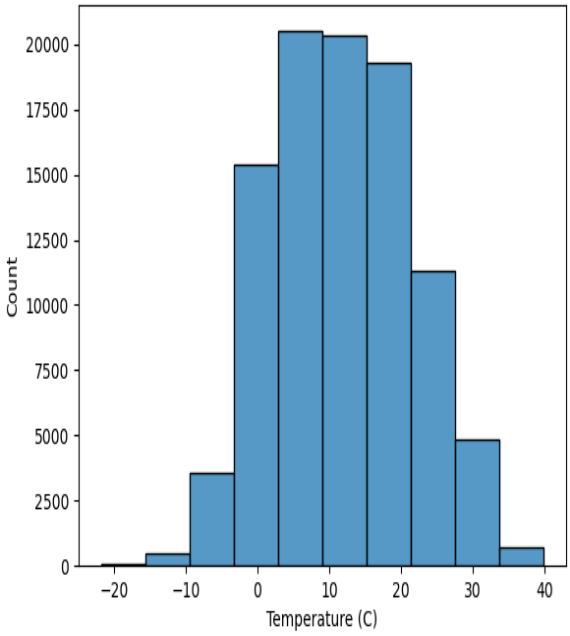
Clustering and fitting (Weather dataset)

Name : Sai Krishna V

Student_id: 23022047

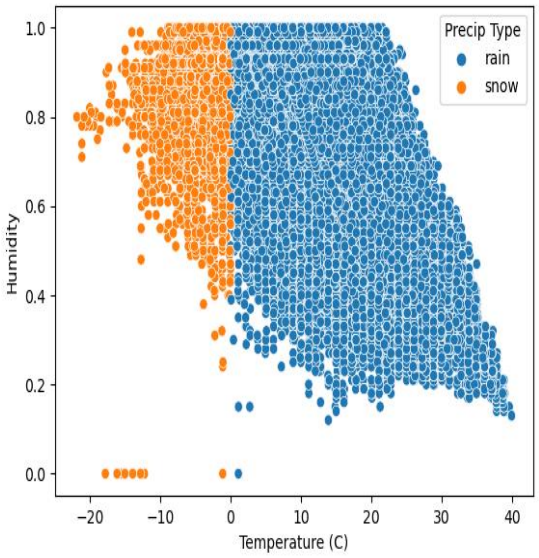
Abstract: This dataset has summary information that includes an abstract weather summary along with temperature, pressure, humidity, wind direction, and speed columns.

Histogram: The image is a histogram showing the distribution of temperature values in degrees Celsius. The x-axis is labeled “Temperature in C” and ranges from -20 to 40 degrees Celsius. The y-axis is labeled “Count” and shows the frequency of occurrences, ranging from 0 to 20,000. The histogram has several bars representing different temperature intervals. The tallest bars are centered around 0 to 20 degrees Celsius, indicating these temperatures have the highest frequency in the dataset.



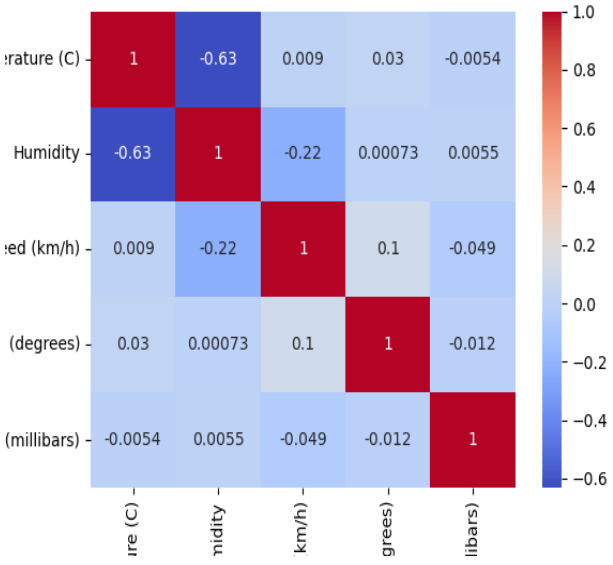
Scatter: The image is a scatter plot graph that illustrates the relationship between temperature, humidity, and precipitation type (rain or snow). The horizontal axis represents temperature in degrees Celsius, ranging from below -20 to above 30 degrees. The vertical axis represents humidity as a proportion from 0 to 1. Two types of precipitation are depicted by different colored dots: orange for rain and blue for snow.

The data points indicate that rain is more likely to occur at higher temperatures and a wide range of humidity levels, while snow occurs at lower temperatures across a similarly broad range of humidity. The transition from snow to rain appears to occur around the 0-degree Celsius mark, which is consistent with the freezing point of water. The plot suggests that as the temperature rises above freezing, precipitation is more likely to fall as rain rather than snow.

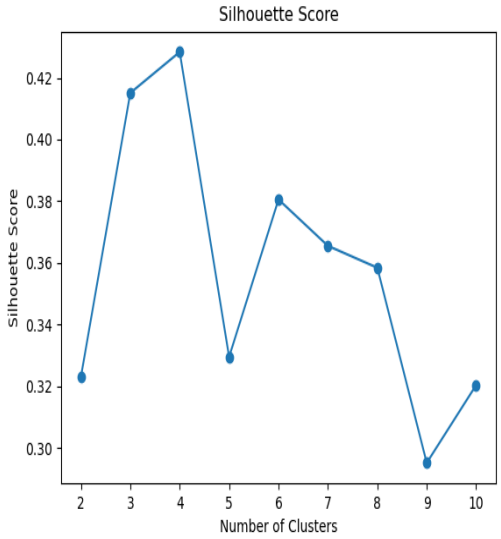


Heatmap:

The image displays a heat map correlating different weather-related variables: temperature in Celsius, humidity, wind speed in kilometers per hour, wind direction in degrees, and atmospheric pressure in millibars. Each variable correlates with itself perfectly, as indicated by the value of 1 on the diagonal. There is a strong negative correlation between temperature and humidity, with a value of -0.63, suggesting that as temperature increases, humidity tends to decrease. Other correlations between variables are weaker, as indicated by values closer to zero. The color gradient on the right side of the image indicates the correlation strength, with red representing a stronger positive correlation, blue a stronger negative correlation, and white indicating no correlation.



Sillhoute:

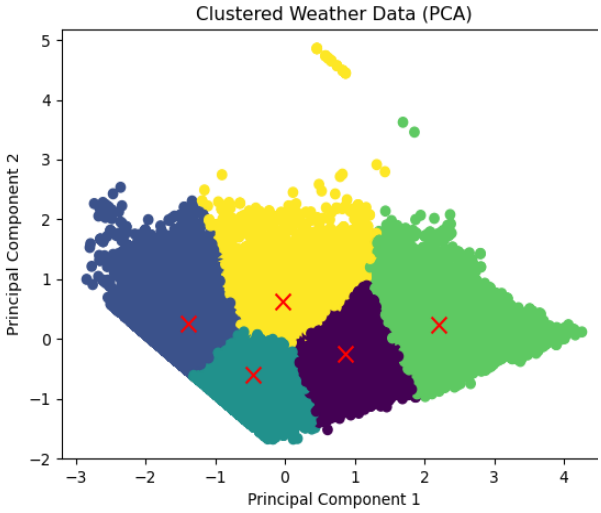


The image presents a line graph titled “Silhouette Score,” plotting the silhouette score values against the number of clusters. The x-axis is labeled “Number of Clusters” and ranges from 2 to 10. The y-axis represents the silhouette score, but the exact range is not visible. The line graph shows that the silhouette score peaks at 3 clusters, dips at 4, rises again at 5, and then generally trends downward as the number of clusters increases from 5 to 10, with some fluctuations. The graph suggests that 3 clusters might be the optimal number for whatever clustering algorithm was used, as it has the highest silhouette score, indicating better-defined and separated clusters.

Clutering:The image shows a scatter plot titled “Clustered Weather Data (PCA),” indicating that the data has been processed using Principal Component Analysis (PCA), a statistical technique used for dimensionality reduction. The plot has two axes: “Principal Component 1” on the x-axis and “Principal Component 2” on the y-axis. Points are

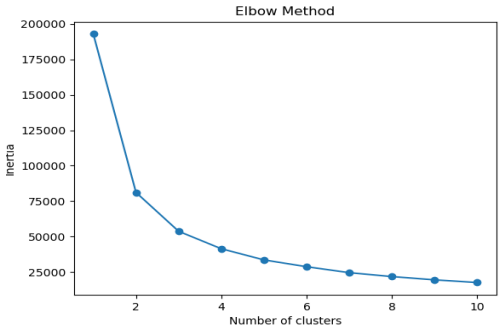
colored differently, suggesting the presence of distinct clusters within the data. Red “X”

marks appear to indicate the central points or centroids of the clusters. The plot illustrates how PCA can be used to simplify complex data into clusters for easier analysis and interpretation.

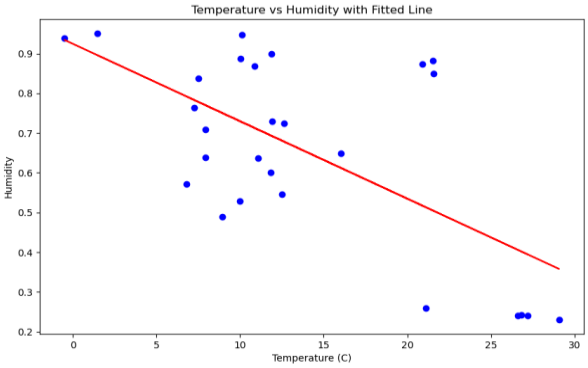


Elbow :The image is a graphical representation of the Elbow Method used in cluster analysis. The graph plots “Inertia” on the y-axis against the “Number of clusters” on the x-axis. The inertia appears to decrease sharply as the number of clusters increases from 2 to around 4 and then gradually levels off as the number of clusters approaches 10. This pattern indicates that the optimal number of clusters for the given dataset could be found at the “elbow” point, where the rate of decrease in inertia significantly slows down, which in this case appears to be around the 4-cluster mark. The Elbow Method is commonly used to determine the appropriate number of clusters to use in k-means clustering by identifying the point at which the addition of more clusters does not significantly

improve the fit of the model.



Fitting: The image displays a scatter plot graph comparing temperature on the x-axis to humidity . A fitted line is included to show the trend, which indicates a negative correlation between temperature and humidity; as the temperature increases, humidity tends to decrease. The data points are spread out, suggesting some variability in the relationship between the two variables.



Github link:
https://github.com/SaiKrishna200120/Clustering-and-fitting.git