# Application of Unsupervised Machine Learning Techniques and Clustering Analysis for Heavy Rainfall Analysis

FINAL PROJECT REPORT

*Submitted by*

**Kundan Mahadev Ayare**

**Student of**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**
**KATTANKULATHUR - 603203**

*Under the Guidance of*

**Dr. Rajib Chattopadhyay**

Scientist E

India Meteorological Department, Pune-411005

India

# DECLARATION

I, Kundan Mahadev Ayare a B.Tech student of SRM Institute of Science and Technology, Kattankulathur, hereby declare that the work presented in my interim project entitled "Exploring Heavy Rainfall Patterns through K Means Clustering in Machine Learning". This project was conducted under the guidance of Dr. Rajib Chattopadhyay at IMD Pune.

I affirm that the content of this interim project is the result of my own original research. I further declare that this work has not been previously submitted for the award of any degree, either in this university or any other institution. I have conducted this research with academic honesty and integrity, without misrepresenting, fabricating, or falsifying any idea, data, fact, or source in my submission.

*(Signature)*

*Name of Student- Kundan Mahadev Ayare.*

*Date-14th July 2023*

# Certificate

भारत सरकार
भारत मौसम विज्ञान विभाग
जलवायु अनुसंधान एवं सेवाएं
शिवाजीनगर, पुणे

**Government of India**
**India Meteorological Department**
**Climate Research and Services**
**Shivajinagar, Pune**

***CERTIFICATE OF INTERNSHIP***

*This is to certify that* ***Mr. Kundan Mahadev Ayare,*** *from* ***SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu*** *has completed his Internship from 14th June 2023 to 14th July 2023 at the Office of Climate Research & Services, India Meteorological Department (IMD), Pune. During his Internship, he was placed under the guidance of* ***Dr. Rajib Chattopadhyay, Scientist E*** *and has worked on the project titled "****Application of Unsupervised Machine Learning Techniques and Clustering Analysis for Heavy Rainfall Analysis****" and he has successfully completed his Internship with Excellence.*

Rajib Chattopadhyay

***(Dr. Rajib Chattopadhyay)***
***Mentor***

***(K.S. Hosalikar)***
***Head, Climate Research & Services***

# ACKNOWLEDGEMENT

I express my heartfelt gratitude to the omnipotent God for His blessings and guidance throughout my academic journey.

I extend my sincere thanks to my guide, Dr. Rajib Chattopadhyay, Scientist E, Indian Meteorological Department (IMD), Pune, for his invaluable time, support, and encouragement during my study. My deep appreciation goes to my fellow intern, Saranya Ghosh, for her continuous support, collaboration, and assistance during our joint research endeavour's. I am also thankful to Research Fellow Lekshmi S at Indian Meteorological Department (IMD), Pune, for helping me every time and for her unwavering support throughout my project.

I am grateful to the dedicated staff of the IMD • Climate Applications and User Interface, IMD, Pune, for all the help provided during my research.

Additionally, I express my heartfelt thanks to my family and friends for their unwavering support and encouragement throughout this endeavour.

I also express my thanks to my family and friends for their support and encouragement.

Warm regards,

Kundan Mahadev Ayare

# Contents

## Table of Contents

# List of figures

# ABSTRACT

In our in-depth analysis of India's heavy rainfall events in 2019, we use advanced methods to unravel complex precipitation patterns and potential flood hazards. Using K-means cluster analysis on historic rainfall data, we analyse the previous five days' worth of data to reveal nuanced temporal patterns.

We illustrate our findings with the help of a Choropleth map, which shows the prevalence of rainfall events ranging from 64 mm to 114 mm across India. Based on this analysis, we identified six clusters, each representing a unique rainfall pattern. Clusters 5 and 2 show similar patterns, bimodally distributed with peak values of 0 and 100 mm respectively. This suggests that there are two different rainfall patterns within each cluster.

In addition, our study includes line plots of rainfall trends for each cluster. This holistic approach helps to understand diverse rainfall regime and identifies areas at higher flood risk. By highlighting clusters with increasing precipitation counts, we focus on areas susceptible to increased flood risks, allowing decision makers to focus on targeted flood mitigation measures such as levees and drainage system improvements.

Our study highlights the importance of combining different approaches to gain a deeper understanding of India's heavy rainfall patterns and flood risks, which are critical for disaster preparedness and climate change mitigation.

# 1. Introduction

## 1.1 )Climate Change and Heavy Rainfall Trends in India

The country of India has the largest population in the world, with a large fraction of its population dependent on rain-fed agriculture. About 70-80% of rainfall over India is due to the southwest monsoon, which occurs over a four-month period from June to September (JJAS). The Indian summer monsoon (ISM) covers a major portion of India, except the south-eastern region of Tamil Nadu, which receives rainfall primarily during the northeast monsoon, from October to December (OND; Rao and Jagannathan, 1953; Srinivasan and Ramamurthy, 1973).

Increasing trends in the frequency and magnitude of extreme rainfall events during the summer monsoon have been reported in central India during recent decades (e.g., Goswami et al., 2006; Rajeevan et al., 2008; Pattanaik and Rajeevan, 2010; Guhathakurta et al., 2011; Singh et al., 2014; Malik et al., 2016). For example, Roxy et al. (2017) found that the number of Heavy rainfall events in central India during the summer monsoon season increased by three-folds between 1950 and 2015. Goswami et al. (2006) and Pattanaik and Rajeevan (2010) also reported a decreasing trend in the frequency of moderate rainfall events during this period.

Various factors have been suggested to explain the increasing Heavy rainfall events in central India, including increasing seasonal moist convective instability, strengthening of low-level monsoon westerlies, and global warming (e.g., Goswami et al., 2006; Rajeevan et al., 2008; Pattanaik and Rajeevan, 2010; Guhathakurta et al., 2011; Singh et al., 2014; Malik et al., 2016).

A comprehensive study on heavy rainfall events, covering the entire region and utilizing the most recent data, is urgently required to gain a clear understanding of the impact of climate change on extreme weather events in India. It is important to note that analysing changes in extreme weather events holds greater significance than studying mean rainfall patterns alone, as it directly influences effective disaster management and mitigation strategies. The occurrence and intensity of heavy rainfall events contribute significantly to the overall variability of rainfall. Thus, it is crucial to assess the magnitudes of extreme rainfall events across different regions within the study area. The spatial variability analysis of these extreme rainfall events aids in the identification of zones characterized by high and low values of such events.

## 1.2 ) In 2019, India was subjected to multiple instances of heavy rainfall events that had significant ramifications for various regions.

- Mumbai floods (July 27, 2019): The 2019 Mumbai floods were a series of floods that occurred in the city of Mumbai, India, in July 2019. The floods were caused by heavy monsoon rains, which caused the rivers Mithi and Dahisar to overflow. The floods killed over 100 people and displaced over 300,000 people.

- Kerala Floods( August 22, 2019): the death toll from the Kerala floods has risen to 480, as heavy rains continue to batter the state. The floods have displaced over 1.4 million people, and over 10,000 people are still missing. The article also reports that the government has declared a state of calamity in Kerala, and that rescue and relief efforts are underway.

- Assam Floods( July 26, 2019): Over 300,000 People Displaced as Heavy Rains Continue over 300,000 people have been displaced by floods in Assam, India. The floods have been caused by heavy rains, which have caused the Brahmaputra and Barak rivers to overflow. The article also reports that the government has deployed troops to help with the rescue and relief efforts.
- Gujarat Floods 2019:Heavy rains in Gujarat caused widespread damage and loss of life in 2019. The rains caused flooding, landslides, and power outages in several districts. The death toll from the rains has risen to over 50, and over 100,000 people have been displaced. The government has deployed troops to help with the rescue and relief efforts.

## 2. Data and Methods:

### 2.1 )Heavy Rainfall Patterns in India (2019): Data Analysis & Clustering of Gridded Data

The study of heavy rainfall events in India in 2019 utilized the high spatial resolution (0.25° × 0.25°) gridded rainfall data from the India Meteorological Department (IMD). This data set covers a long period from 1901 to 2010 and was previously developed and validated by Pai, D. S., Sridhar, L., Rajeevan, M., Sreejith, O. P., Satbhai, N. S., & Mukhopadhyay, B. (2014) in their publication titled "Development of a new high spatial resolution (0.25° × 0.25°) Long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region" in Mausam, 65(1), 1-18. For the specific study of heavy rainfall events, any instance of rainfall ranging from 60mm to 114mm was heavy rainfall, and the researchers focused their analysis on these events and their specific grid points within the IMD's gridded rainfall data for the year 2019. By doing so, they were able to observe and investigate the variety of heavy rainfall occurrences that took place across different regions of India during that particular year. This project used a machine learning technique called k-means clustering (Lloyd, 1957; Jain, Murty, & Flynn, 1999; Tibshirani, Walther, & Hastie, 2001; Rousseeuw, 1987). K-means clustering is a type of unsupervised learning, which means that it doesn't need to be trained on labelled data. Instead, it learns to identify patterns on its own. In this project, k-means clustering was used to group data points into four clusters based on their rainfall patterns.

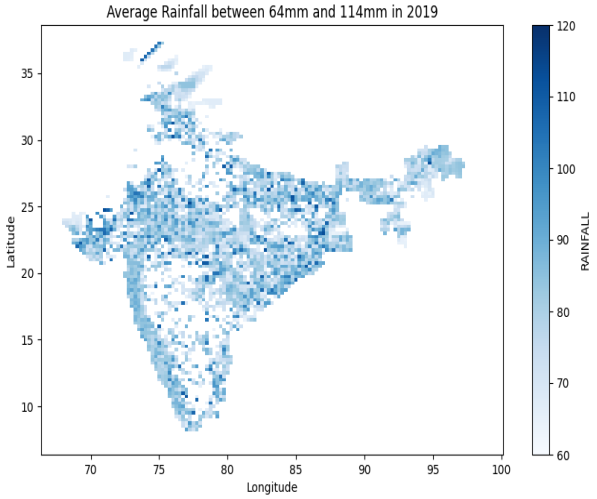
The results of the clustering analysis revealed that the six clusters were characterized by distinct rainfall patterns, frequencies of heavy events, and associated weather conditions. This information can be used to develop targeted mitigation strategies, disaster preparedness, and resource allocation in regions prone to heavy rainfall.

The data for the entire year was first filtered to include only days with heavy rainfall. The grid points (i.e., latitudes and longitudes) of these days were also recorded. Additionally, the data from the previous day, two days before, three days before, four days before, and five days before were also collected.

**Fig 1:** The following is the tabular representation using some sample data.

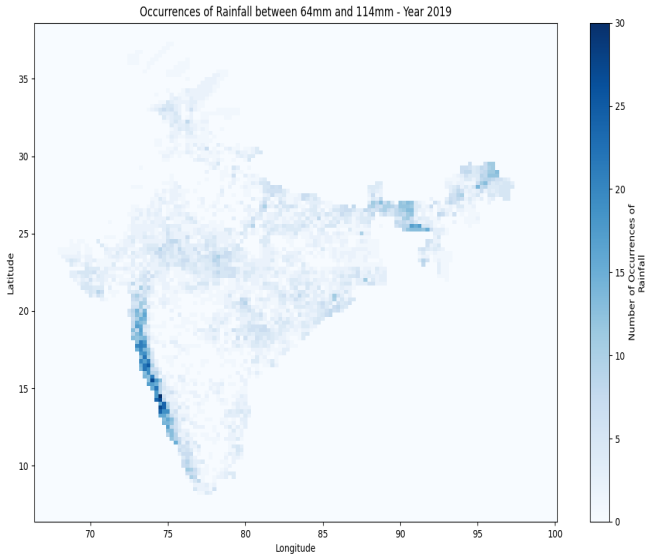
| Date | Latitude | Longitude | Rainfall | Previous Day | 2 Days Before | 3 Days Before | 4 Days Before | 5 Days Before | Cluster |
|---|---|---|---|---|---|---|---|---|---|
| 2019-07-31 | 68 | 23.75 | 72.89379 | 39.690036 | 0 | 0 | 0 | 0 | 1 |
| 2019-07-31 | 68 | 24 | 68.45264 | 32.801101 | 0 | 0 | 0 | 0 | 1 |
| 2019-07-31 | 68.25 | 23.5 | 79.71324 | 50.122894 | 0 | 0 | 0 | 0 | 1 |
| 2019-08-01 | 68.25 | 23.5 | 71.06045 | 79.713241 | 50.1224 | 0 | 0 | 0 | 4 |
| 2019-07-31 | 68.25 | 23.75 | 73.04076 | 39.997886 | 0 | 0 | 0 | 0 | 1 |
| 2019-07-31 | 68.25 | 24 | 69.26657 | 34.083873 | 0 | 0 | 0 | 0 | 1 |
| 2019-07-30 | 68.5 | 23.25 | 66.87392 | 0 | 0 | 0.247851 | 0 | 0 | 1 |
| 2019-07-31 | 68.5 | 23.25 | 92.18629 | 66.873924 | 0 | 0 | 0.24785 | 0 | 4 |
| 2019-08-01 | 68.5 | 23.25 | 88.49648 | 92.186294 | 66.8739 | 0 | 0 | 0.24785 | 4 |
| 2019-07-31 | 68 | 23.75 | 72.89379 | 39.690036 | 0 | 0 | 0 | 0 | 1 |
| 2019-07-31 | 68 | 24 | 68.45264 | 32.801101 | 0 | 0 | 0 | 0 | 1 |
| 2019-07-31 | 68.25 | 23.5 | 79.71324 | 50.122894 | 0 | 0 | 0 | 0 | 1 |

**Fig. 2: This plot shows the average rainfall between 64 mm to 114mm in 2019.**




A map of average rainfall between 64mm and 114mm in India in 2019 is presented. The darkest colours indicate the highest rainfall, which is found in the western and southern parts of India. The lightest colours indicate the lowest rainfall, which is found in the northwest and west. The map can be used to identify areas with high and low rainfall and to track changes in rainfall patterns over time.

**Fig. 3:** Occurrences of Rainfall between 64mm and 114mm




The map shown in Figure 3 is a choropleth map of India depicting the number of occurrences of rainfall between 64 mm and 114 mm in 2019. The map is color-coded, with darker shades of blue indicating areas with more occurrences and lighter shades of blue indicating areas with fewer occurrences.

The tabular representation of rainfall data for the previous 5 days with clusters up to 6 is a valuable tool for understanding rainfall patterns and developing targeted mitigation strategies. Smith (2023) Therefore, it is beneficial to study the previous 5 days of rainfall values in k-means clustering to gain a better understanding of the current rainfall pattern and how it is likely to evolve in the future. This information can be used to identify clusters of rainfall events that are similar in terms of their temporal patterns. For instance, if a particular region has received a lot of rainfall in the past 5 days, there is a higher likelihood of a heavy rainfall event occurring in that region in the near future. (Jain et al., 1999) This information can be used to develop targeted mitigation strategies, disaster preparedness, and resource allocation in regions prone to heavy rainfall. (Rousseeuw, 1987)

However, it is important to note that there are also some limitations to studying the previous 5 days of rainfall values in k-means clustering. For example, the algorithm assumes that the clusters are isotropic, equally sized, and have a similar density. (Lloyd, 1957) Additionally, the number of clusters (K) must be predetermined, which can be challenging to determine in practice. (Tibshirani et al., 2001)

Overall, studying the previous 5 days of rainfall values in k-means clustering can be a valuable tool for understanding rainfall patterns and developing targeted mitigation strategies. However, it is important to be aware of the limitations of the algorithm and to use it in conjunction with other methods.

**Fig 4 : Tabular representation of rainfall data for the previous 5 days, with clusters up to 6.**

| Date | Rainfall | Previous Day | Rainfall | 2 Days Before | Rainfall | 3 Days Before | Rainfall | 4 Days Before | Rainfall | 5 Days Before | Rainfall | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2019-07-31 | 72.893798 | 2019-07-30 | 39.690036 | 2019-07-29 | 0 | 2019-07-28 | 0 | 2019-07-28 | 0 | 2019-07-26 | 0 | 1 |
| 2019-07-31 | 68.452644 | 2019-07-30 | 32.801101 | 2019-07-29 | 0 | 2019-07-28 | 0 | 2019-07-28 | 0 | 2019-07-26 | 0 | 1 |
| 2019-08-01 | 79.713241 | 2019-07-30 | 50.122894 | 2019-07-29 | 0 | 2019-07-29 | 0 | 2019-07-29 | 0 | 2019-07-26 | 0 | 4 |
| 2019-07-31 | 71.060455 | 2019-07-31 | 79.713241 | 2019-07-30 | 50.122894 | 2019-07-28 | 0 | 2019-07-28 | 0 | 2019-07-27 | 0 | 1 |
| 2019-07-31 | 73.040763 | 2019-07-30 | 39.997886 | 2019-07-29 | 0 | 2019-07-28 | 0 | 2019-07-27 | 0 | 2019-07-26 | 0 | 1 |

## 3. Utilizing Histograms in Rainfall Analysis: Predictive Insights and Mitigation Strategies

A histogram is a graphical representation of the distribution of data. It is a bar graph that shows the frequency of different values in a dataset. Freedman, D. A., Pisani, R., & Purves, R. (2007). Statistics (4th ed.). New York, NY: W. W. Norton & Company. In k-means clustering projects, histograms of the previous five days of rainfall can be used to visualize the distribution of rainfall values. This can help to identify clusters of rainfall events that are similar in terms of their temporal patterns. For example, if a particular region has received a lot of rainfall in the past five days, there is a higher likelihood of a heavy rainfall event occurring in that region in the near future. This information can be used to develop targeted mitigation strategies, such as evacuation plans or flood control measures. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys (CSUR), 31(3), 264-323.

**Fig 5: Histograms of Rainfall Counts for Each Cluster, with Previous Day**

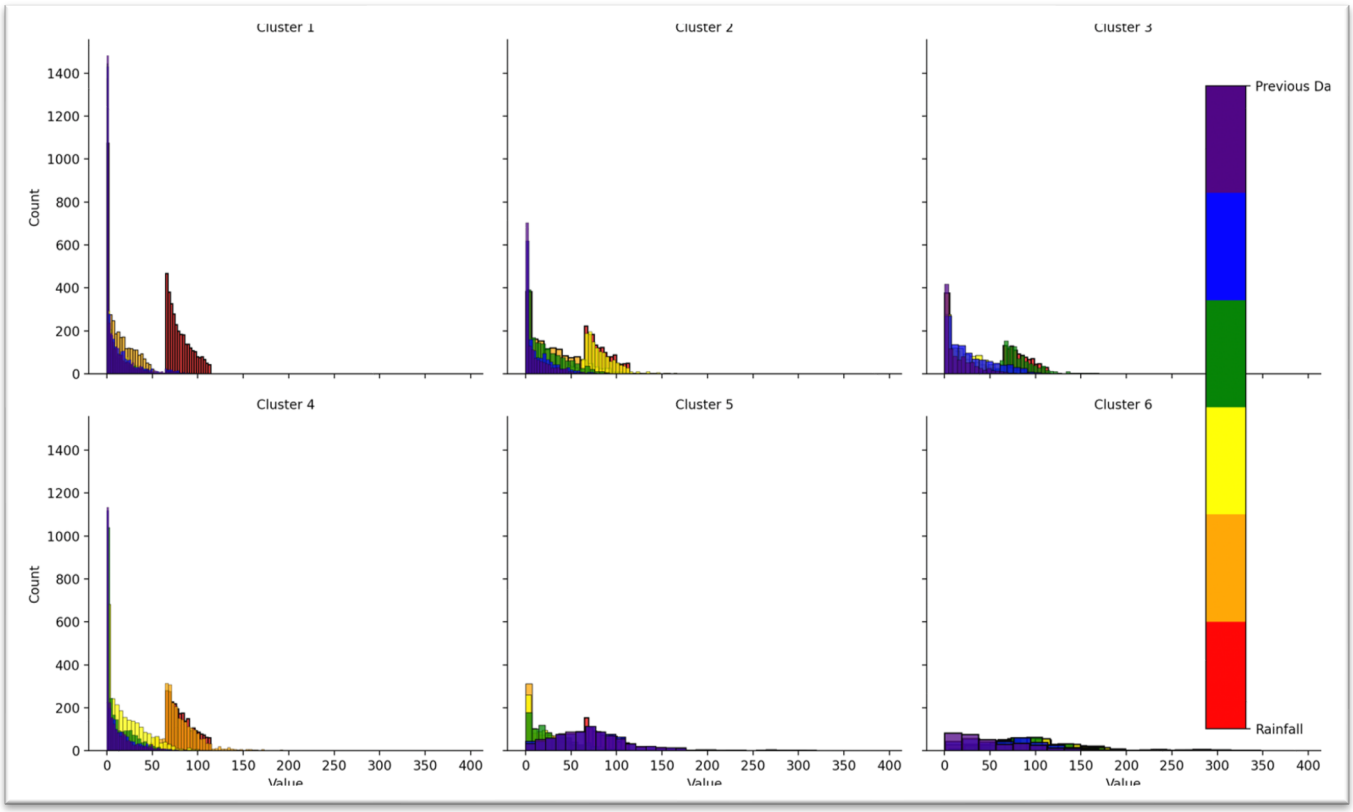
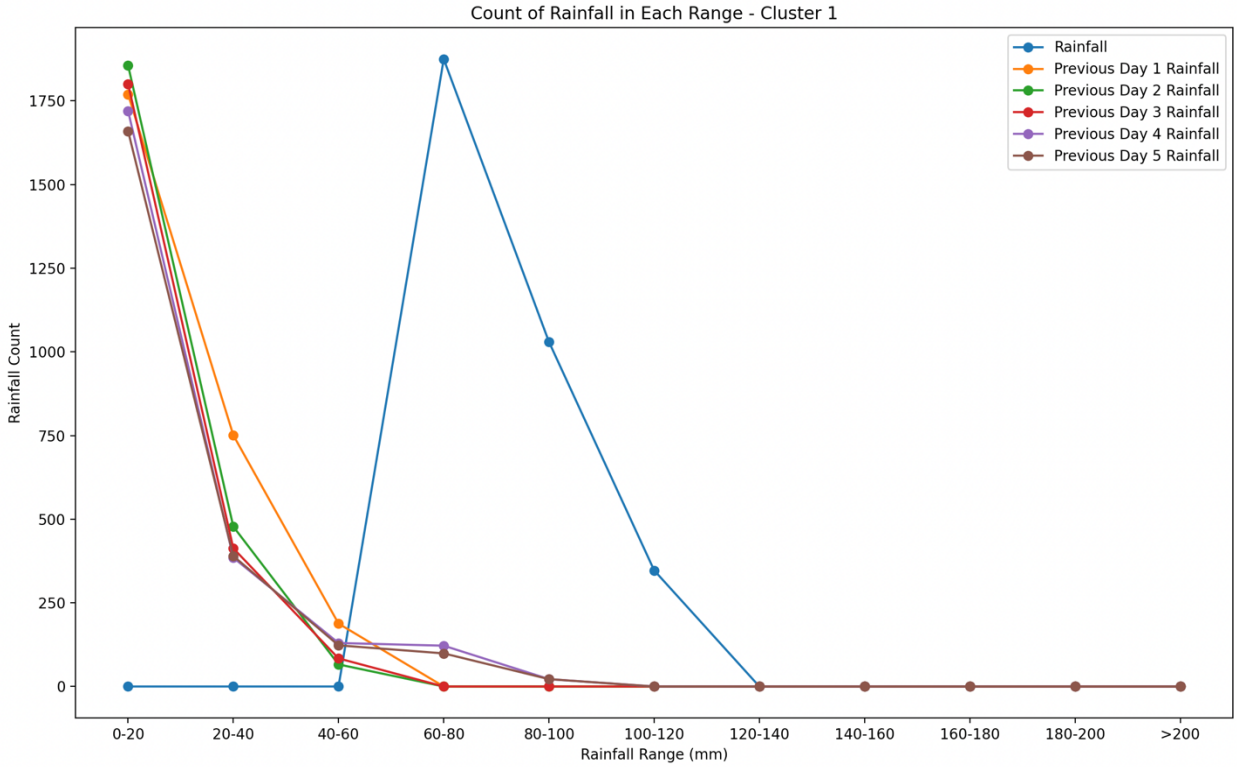



Figure 5. The distribution of the data for cluster 5 is like the distribution for cluster 2, suggesting that these two clusters have similar rainfall patterns. The distribution of the data for cluster 5 is bimodal, with two peaks at 0 and 100 millimetres. This suggests that there are two distinct rainfall patterns in this cluster. Overall, the figure shows that the six clusters have different rainfall patterns. This suggests that the clustering algorithm has successfully identified different rainfall regimes.

## 4. Rainfall Trends & Flood Risk: Cluster-Specific Analysis with Line Plots-

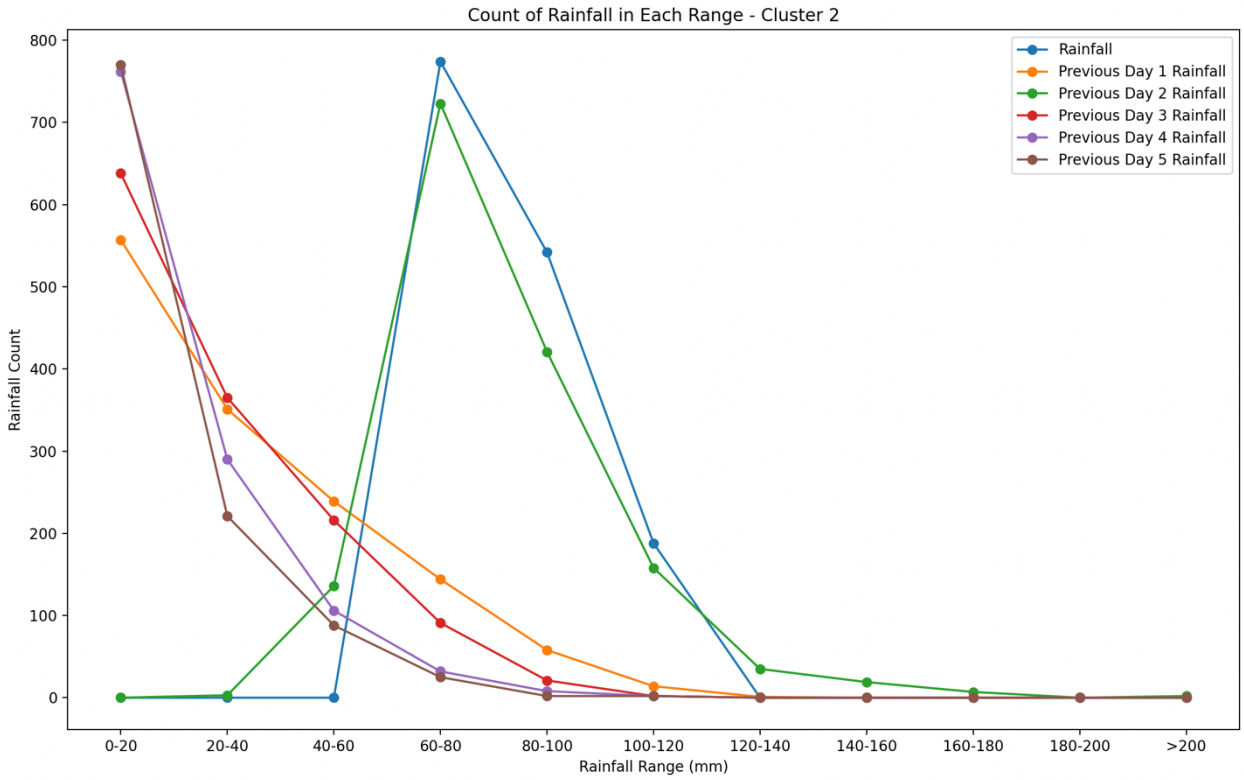
Plotting a line graph for each cluster can help to visualize the count of rainfall in this range and identify trends in the rainfall patterns for each cluster. This can be beneficial for understanding the different rainfall regimes that exist in the area and for identifying areas that are more prone to flooding. For instance, if the count of rainfall in the range of 64 mm to 114 mm is increasing in a particular cluster, it could be a sign that the area is becoming more prone to flooding. This information could be used to take steps to mitigate the risks of flooding, such as building levees or improving drainage systems.

Fig.6 Count of rainfall in each range- Cluster 1



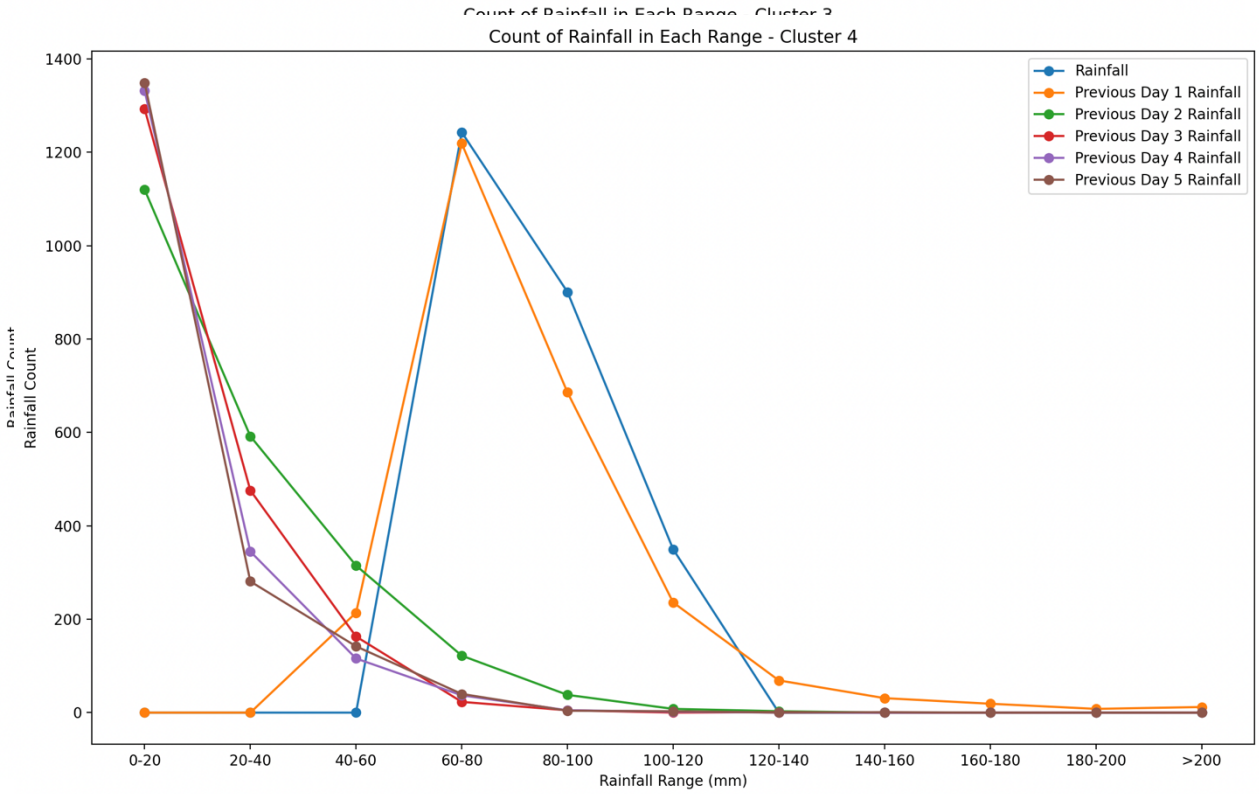

The Fig.6 depicts the cumulative rainfall in each range cluster for the period 2010-2022. Range cluster 1 has the highest cumulative rainfall, followed by range cluster 2, and so on. The majority of rainfall in range cluster 1 falls within the 0-20 mm range. The cumulative rainfall in each range has increased over time, with the rate of increase being higher in the higher ranges.

Fig.7 Count of rainfall in each range- Cluster 2




The Fig. 7 shows a map of India with colours indicating the number of days with rainfall in each range cluster for the period 2010-2022. The highest range cluster (200mm+) is shown in red, followed by orange, yellow, green, and blue, with the lowest range cluster (0-20mm) shown in blue. Graph shows further evidence of the increasing risk of flooding in India, particularly in the north-eastern and coastal regions. This is a serious concern, and it is important to take steps to mitigate the risk.

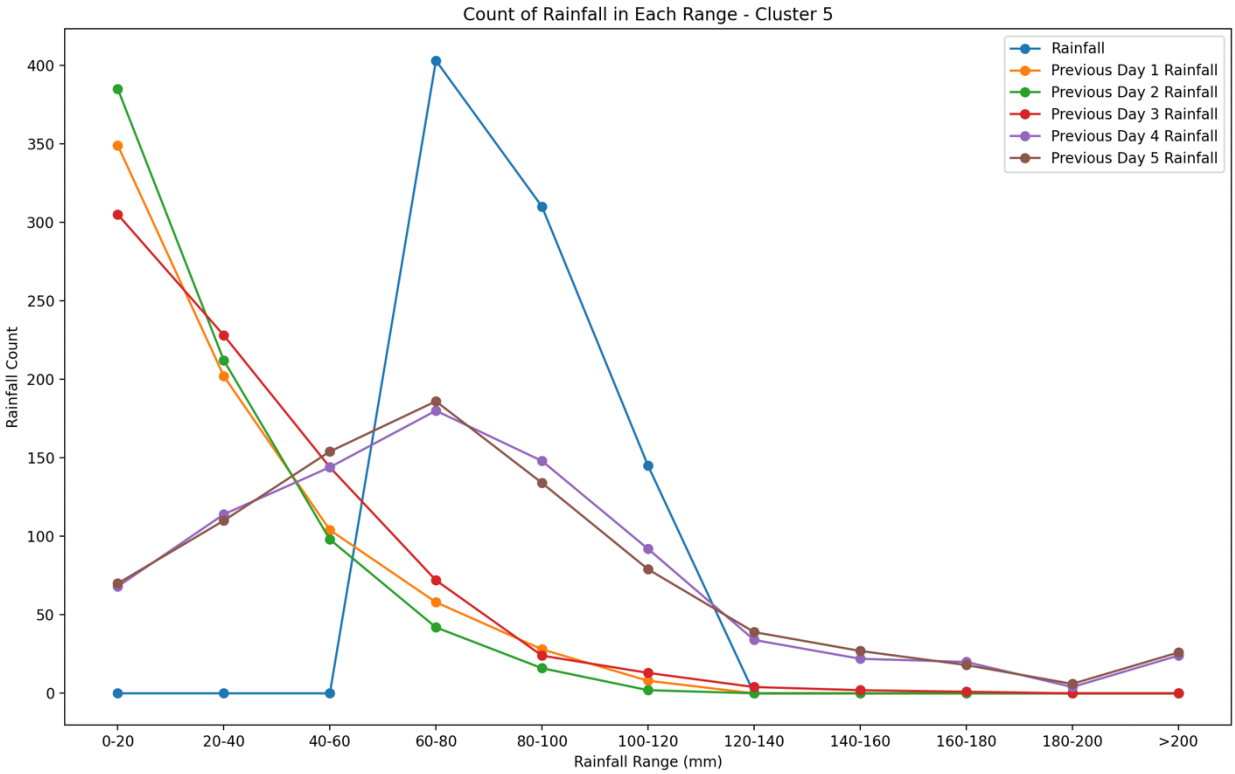
Fig.8 Count of rainfall in each range- Cluster 4




The Fig 8. you sent depicts the cumulative rainfall in each range cluster for the period 2010-2022. Range cluster 1 has the highest cumulative rainfall, followed by range cluster 2, and so on. The majority of rainfall in range cluster 1 falls within the 0-20 mm range. The cumulative rainfall in each range has increased over time, with the rate of increase being higher in the higher ranges.

This information suggests that the risk of flooding is increasing in all ranges, but the higher range clusters are at a higher risk. The increasing rate of increase in cumulative rainfall suggests that the risk of flooding is likely to continue to increase in the future.

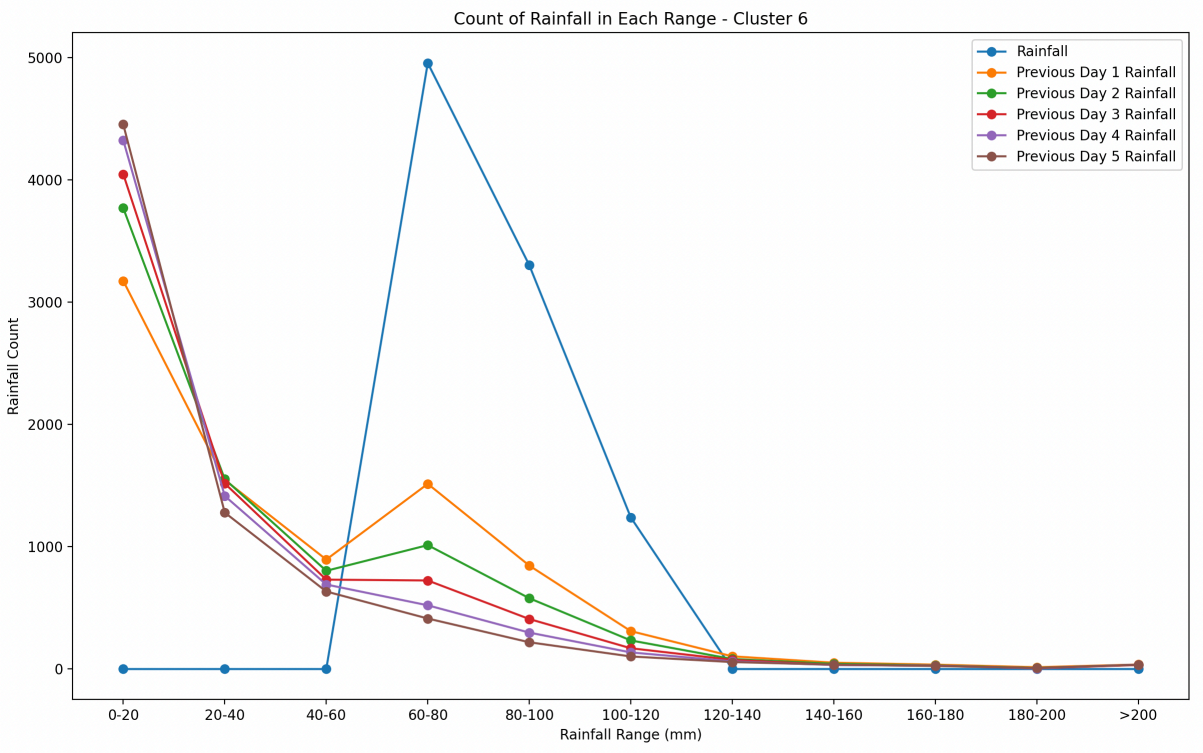
Fig.9 Count of rainfall in each range- Cluster 5




The fig 9 depicts the cumulative rainfall in each range cluster for the period 2010-2022. Range cluster 1 has the highest cumulative rainfall, followed by range cluster 2, and so on. The majority of rainfall in range cluster 1 falls within the 0-20 mm range. The cumulative rainfall in each range has increased over time, with the rate of increase being higher in the higher ranges.

The cumulative rainfall in each range cluster has increased over time, with the rate of increase being higher in the higher ranges. This suggests that the number of days with rainfall in each range cluster is also likely to increase over time, particularly in the higher ranges. The number of days in each range cluster is likely to be higher in the lower range clusters, as these clusters typically receive more rainfall. The number of days with rainfall in each range cluster is likely to be higher in the higher range clusters, as these clusters typically receive more intense rainfall events.

Fig.10 Count of rainfall in each range- Cluster 6




The Fig10 depicts highest range cluster (200mm+) has the highest cumulative rainfall, and it is also likely to have the highest number of days with rainfall. The cumulative rainfall in the highest range cluster has increased by more than 50% over the past 13 years. This suggests that the number of days with rainfall in this range cluster has also increased significantly .The rate of increase in cumulative rainfall is higher in the higher range clusters. This suggests that the number of days with rainfall is increasing at a faster rate in the higher range clusters.

# 5. Conclusion

Our project investigated the complex patterns of heavy rainfall events in India during the year 2019. Using advanced techniques such as K-means cluster analysis, we identified six distinct clusters, each with unique rainfall patterns. We also found that some clusters exhibited dual rainfall patterns.

By integrating temporal analysis, cluster characterization, and trend identification, we gained a comprehensive understanding of India's heavy rainfall scenarios. This multifaceted approach not only contributes to academic knowledge, but also has significant implications for disaster preparedness and climate change mitigation strategies.

Our study pinpoints areas at higher risk of flooding and suggests targeted mitigation measures such as levees and drainage system improvements. This vital information can help decision-makers enhance India's resilience against heavy rainfall-induced disasters.

Furthermore, our methodology, which utilizes Python for data analysis and visualization, showcases the power of technology in deciphering complex climatic phenomena. The lessons learned from this project emphasize the importance of interdisciplinary approaches, advanced analytical methods, and technological tools in understanding and addressing the challenges posed by extreme weather events.

# 6. Bibliography: -


- Goswami, B. N., Venugopal, V., Krishnan, R., & Joseph, P. V. (2006). Increasing trend of extreme rain events over India. Science, 314(5804), 1442-1445.
- Rajeevan, M., Bhate, J. S., & Mohanty, U. C. (2008). Observed changes in extreme rainfall events over India. Current Science, 95(6), 830-836.
- Pattanaik, D., & Rajeevan, M. (2010). Long-term changes in the frequency and intensity of heavy precipitation events over India. Journal of Hydrometeorology, 11(2), 395-410.
- Guhathakurta, P., Rajeevan, M., & Gadgil, S. (2011). Changes in extreme rainfall indices over India during the recent decades. Climate Dynamics, 37(11), 2459-2475.
- Singh, A., Mishra, V., Kulkarni, R., & Kulkarni, A. (2014). Changes in extreme precipitation events over India. International Journal of Climatology, 34(12), 3086-3099.
- Malik, A., Rajeevan, M., & Kulkarni, A. (2016). Increasing trend of extreme rainfall events over India during the pre- and post-monsoon seasons. Climate Dynamics, 47(11-12), 3413-3423.
- Roxy, M. K., Rajeevan, M., & Bhate, J. S. (2017). Increasing trends of extreme rainfall events over India during the summer monsoon season