# Exploratory Data Analysis (EDA) Report: Indian Rainfall Data

**Submitted By:**

**Name:** Aryan Paratakke

**PRN:** 22070521070

**Section:** C

**Semester:** 7

**Batch:** 2022-2026

**Under the Guidance of:** Dr. Bhupesh Kumar Dewangan

**Course:** Data Science (CA1 Submission)

**Institute:** Symbiosis Institute of Technology (SIT), Nagpur

## 1. Introduction

The technical report is an in-depth Exploratory Data Analysis (EDA) of an Indian climate/rainfall data that belongs to the real world. The main goal of the given project is to elaborate on the nature of rainfall patterns of different districts and states in India, to draw key trends, distributions, and abnormalities. The data set combines the rainfall measurements on various time scales, such as daily, weekly, cumulative (seasonal/annual accumulated), and monthly.

The analysis is organised in such a way that at first it gives a national picture then brings a miniaturisation analysis of unique rainfall phenomena of Maharashtra, and ends with a comparative study in terms of comparing Maharashtra to its neighbours. This protocol is expected to reveal the insights regarding the variability of rainfall, how it may deviate from normal, and what factors may have an effect on it. The results of this EDA would be of essence in making informed decision in agricultural planning, water resource management and disaster preparedness especially in a climate-sensitive nation like India.

## 2. Dataset Description

The essence of this Exploratory Data Analysis project is anauthentic and full Indian Climate and Rainfall Dataset. The dataset will provide a diagnostic and prognostic data with all the detailed features of rainfall events comprising the associated cloud features and rainfall patterns based on satellites and surface incorporatingRadio Detection And Ranging (RADAR). It is this information that is especially provided to duty forecasters to have operational meteorological functions.

**Source of Data:** The dataset has been sourced from the**National Data & Analytics Platform (NDAP) under NITI Aayog, Government of India.**

- **Dataset Link:** https://ndap.niti.gov.in/dataset/7319?tab=profile

This is a rich data that was initially given as the original rainfall data.csv that provides a multi angle observation of precipitation patterns in India.

**GitHub Link:** https://github.com/Aryan152005/DS_Lab/tree/main/Rainfall_EDA_CA1

**Key characteristics and granularities of the dataset include:**

- Geographical Granularity:
  - srcDistrictName: Specifies the district where the reading was recorded.
  - srcStateName: Indicates the corresponding state or Union Territory.
- Temporal Granularity:
  Rainfall readings are available across various time scales:
  - Daily: Individual daily precipitation measurements.
  - Weekly: Aggregated rainfall data summarized over weekly periods.
  - Cumulative: Rainfall totals accumulated over specific longer durations (e.g., season-to-date or cumulative from a particular start date).
  - Monthly: Comprehensive monthly total rainfall figures.
- Rainfall Metrics:
  - Actual Values (Daily Actual, Monthly Actual, etc.): The recorded amount of rainfall in millimeters (mm) for the respective period.
  - Normal Values (Daily Normal, Monthly Normal, etc.): The historical average or expected rainfall for the given period.
  - Percentage Departure (Percentage of Daily Departure, Percentage of Monthly Departure, etc.): Quantifies the percentage deviation of actual rainfall from the its normal value, clearly indicating surplus (excess) or deficit conditions.
  - Category Labels (Daily Category, Monthly Category, etc.): Qualitative classifications (e.g., 'Normal', 'Excess', 'Deficient', 'Large Excess', 'Large Deficient', 'No Rain', 'Scanty') that categorize the rainfall performance based on its percentage departure from normal.
- Timeframe: The data cover various years and this offers a time-series of the data necessary in examining climate trends and long-term variability of the rains.

This rich lineage and native composition provide the data with great utility towards the application of meteorological analysis in a holistic manner helping in providing insights towards regional rainfall fluctuation, ferreting anomalies, and making informed choices in regards to water resource management.

The basic prerequisite of any kind of data analysis is proper loading of raw data and initial examination to learn about existing structure in a data set and realize the data quality problems in real-time.

## 2.1 Dataset Loading

Data of rainfalls was obtained in `original_rainfall_data.csv`. The loading was made using the Pandas library that is a foundation in the data manipulation in Python and that it is ensured that the data is properly imported into DataFrame. The first few rows give us the first clue in looking into the structure of the dataset:

**Table: First 5 Rows of the Dataset (Initial Load)**

| OBJI_ID | srcDistrictName | srcStateName | srcYear | srcCalendarDay | Daily Actual | Daily Normal | Percentage of Daily Departure | Daily Category | Weekly Date | Weekly Actual | Weekly Normal | Percentage of Weekly Departure | Weekly Category | Cumulative Date | Cumulative Actual | Cumulative Normal | Percentage of Cumulative Departure | Cumulative Category | Monthly Date | Monthly Acutual |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 573 | NICOBAR | ANDAMAN & NICOBAR | 2025 | 2025-07-23 | 2.3 | 11 | -79 | LD | 17-07-2025 To 23-07-2025 | 97.1 | 68.4 | 42 | E | 2025-06-01 | 402.2 | 474 | -15 | N | 01-07-2025 To 23-07-2025,,,, | 132.3 |
| 571 | NORTH & MIDDLE ANDAMAN | ANDAMAN & NICOBAR | 2025 | 2025-07-23 | 17.5 | 14.3 | 22 | E | 17-07-2025 To 23-07-2025 | 100.2 | 88.1 | 14 | N | 2025-06-01 | 1092.2 | 817.4 | 34 | E | 01-07-2025 To 23-07-2025,,,, | 189.4 |
| 572 | SOUTH ANDAMAN | ANDAMAN & NICOBAR | 2025 | 2025-07-23 | 6.1 | 12.8 | -52 | D | 17-07-2025 To 23-07-2025 | 140.8 | 83.3 | 69 | LE | 2025-06-01 | 844.6 | 712.6 | 19 | N | 01-07-2025 To 23-07-2025,,,, | 252.9 |
| 736 | ANAKAPALLI | ANDHRA PRADESH | 2025 | 2025-07-23 | 12.6 | 5.6 | 125 | LE | 23-07-2025 | 12.6 | 5.6 | 125 | LE | 2025-06-01 | 273.5 | 244.5 | 12 | N | 01-07-2025 To 23-07-2025,,,, | 139.3 |
| 94 | ANANTAPURAMU | ANDHRA PRADESH | 2025 | 2025-07-23 | 3.5 | 2.3 | 51 | E | 23-07-2025 | 3.5 | 2.3 | 51 | E | 2025-06-01 | 86.7 | 108.6 | -20 | D | 01-07-2025 To 23-07-2025,,,, | 28.2 |

## 2.3 Initial Data Information and Column Assessment

The df.info() method provided a concise summary of the DataFrame:

- **Total Entries:** The dataset contains 1,229,401 records.
- **Columns:** There are`24`columns in total.
- **Initial Data Types:**
    - Numerical columns like `OBJ_ID`, `srcYear`, `Daily Actual`, `Weekly Actual`, `Cumulative Actual`, `Monthly Actual` were correctly identified as`int64`or`float64`.
    - Categorical columns such as `srcDistrictName`, `srcStateName`, and `Daily Category` were `object` (string) type, which is appropriate.
- **Key Observations for Cleaning:** Importantly some of the columns that the data had to be obviously numerical (e.g. `Daily Normal`, `Percentage of Daily Departure`, `Weekly Normal`, `Monthly Normal` etc.) were of type `object`, meaning that they contained characters or irregular formatting. Also, all columns that had dates ( `srcCalendarDay`, `Daily Date`, `Weekly Date`, `Cumulative Date`, `Monthly Date`) were of type `object` and thus should be parsed as such, into the `datetime` object.
- **Missing Values (Preliminary):** According to the information provided by the `df.info()` output, several columns, which contained information on `Weekly` and `Cumulative` data, as well as `OBJ_ID` columns, consisted of a drastically lower non-null rate (about 416,000 out of 1.2 million) which indicates high numbers (about 66%) of missing data.

## 3. Data Cleaning and Feature Engineering

It is a phase in which the raw data was systematically converted into a clean, homogenous and analysis ready form. The various steps have dealt with particular aspects of data quality that make it possible to be confident in further statistical operations and visualization.

### 3.1 Standardizing Column Names

Irreconcilable column names may create fault and make the program hard to read. This was done by standardizing all the column names.

- **Process:** Leading and trailing whitespaces were removed from all column names. Specific typos identified during the initial inspection were corrected, such as renaming` Percentage of Cumulative Departure` to `Percentage of Cumulative Departure` and `Monthly Acutual` to `Monthly Actual`.
- **Impact:** This standardization ensures that columns can be accessed consistently and without ambiguity, improving the overall integrity and maintainability of the data processing pipeline.

### 3.2 Handling Missing Values

Missing data, if not appropriately managed, can lead to biased analyses and incorrect conclusions. A thorough assessment of missing values was conducted, and a specific imputation strategy was applied.

- **Missingness Profile:** The initial missing values analysis revealed the following profile:

**Table: Missing Values Information (Before Imputation)**

| Unit | Missing Count | Missing Percentage (%) |
|---|---|---|
| Percentage of Weekly Departure | 819493 | 66.6579 |
| Percentage of Cumulative Departure | 817724 | 66.514 |
| Weekly Actual | 817644 | 66.5075 |
| Weekly Date | 816900 | 66.447 |
| Cumulative Actual | 816213 | 66.3911 |
| Cumulative Date | 814461 | 66.2486 |
| Weekly Normal | 813177 | 66.1442 |
| Cumulative Normal | 813176 | 66.1441 |
| OBJ_ID | 813175 | 66.144 |
| Cumulative Category | 813175 | 66.144 |
| Weekly Category | 813175 | 66.144 |
| Percentage of Daily Departure | 17849 | 1.45185 |
| Daily Actual | 15606 | 1.2694 |
| Percentage of Monthly Departure | 6951 | 0.565397 |
| Monthly Actual | 5645 | 0.459167 |
| Monthly Normal | 56 | 0.00455506 |
| Daily Normal | 2 | 0.000162681 |

- Imputation Strategy:
  - For numerical columns representing rainfall measurements (Daily Actual, Weekly Normal, Monthly Actual, and their corresponding departure percentages), missing values (NaN) were imputed with 0.0. This approach is justified in the context of rainfall data, where a missing recording often implies an absence of rainfall for that period.
  - For categorical Category columns (Daily Category, Weekly Category, etc.), missing values were filled with the string 'Unknown'. This explicitly marks data points where a category could not be assigned.

- The high percentage of missing values (approx. 66%) in OBJ_ID and all Weekly and Cumulative columns was particularly noted. This suggests that data at these granularities might not be available for every daily record, rather than being true missing values in an erroneous sense.

- Outcome: Following imputation, all targeted numerical and categorical columns were rendered complete, significantly enhancing the robustness of subsequent analyses.

### 3.3 Correcting Data Types and Parsing Dates

Accurate data types are fundamental for correct statistical computations and time-series analysis. This step focused on converting columns to their appropriate data types.

- **Single Date Columns:** `srcCalendarDay` and `Cumulative Date` columns, which contained date strings in 'DD-MM-YYYY' format, were converted into `datetime` objects using `pd.to_datetime` with `dayfirst=True`. The `errors='coerce' `parameter was used to handle any unparseable entries gracefully by converting them to `NaT` (Not a Time).
- **Date Range Columns:** The `Weekly Date` and `Monthly Date` columns, which held date ranges (e.g., 'DD-MM-YYYY To DD-MM-YYYY'), were parsed into separate`_Start Date`and`_End Date``datetime`columns.
- **Numerical Columns Verification:** All 'Normal' and 'Percentage Departure' columns were confirmed to be`float64`, ensuring that are readiness for numerical operations.
- **Identifier Column:** The `OBJ_ID` column was converted to `Int64` (a nullable integer type in Pandas), which is suitable for identifier columns which that may contain missing entries.
- **Outcome:** The DataFrame is now populated with numerical and date columns typed correctly so today you can do right math and you can get time based aggregate.

### 3.4 Addressing Negative and Outlier Rainfall Values

Such physical measurements as rainfall cannot be negative. Examining the numerical summaries critically, there was a value of -4.00mm in the`Monthly Actual` which showed that there was a data error.

- **Correction:** All negative values observed in 'Actual' rainfall columns (specifically`Monthly Actual`) were identified and replaced with`0.0`.
- **Outcome:** This made all of the rainfall measurements physically sound. After correction, all columns of values in the column of Actual rainfall were found to have minimum value of`0.00 mm`, which confirms to have reliable data in which meteorological interpretations can be carried out.

### 3.5 Correcting Average Annual Rainfall Calculation

An initial aggregate sum for "Average Annual Rainfall across India" yielded an unrealistically high value, indicating an aggregation error.

- **Refinement:** This was modified to calculate the`Total Monthly Actual Rainfall per Year`as a summation of`Monthly Actual values within the group`selected by`srcYear`grouped over the full dataset. This is a record of the amount of rainfall in each year. Then an average of these annual sums was taken in order to come up with a more significant "Approximate Average Annual Rainfall inside India."
- **Observation:** A declining pattern against the years (2021) to 2025 was observed in the corrected annual totals that may indicate that the data in newer years is not fully complete compared to the hypothetical climatic trend.

- **Table: Total Monthly Actual Rainfall per Year (Aggregated Across All States/Districts)**

| srcYear | Total Annual Rainfall |
|---|---|
| 2021.00 | 55473980.80 |
| 2022.00 | 58087780.10 |
| 2023.00 | 52294945.40 |
| 2024.00 | 24250461.90 |
| 2025.00 | 6978086.50 |

- **Insight:** It has fixed this metric of `39417050.94 mm` which has more clarity at such high level when it comes to rainfall of India in entire period of the dataset and later it can be used to compare among the states.

### 3.6 Deriving Time-Based Features

Took existing date columns and created new features to make advanced time-series analysis possible and allow aggregations based on various temporal dimensions.

- **New Features:** `Month`, `DayOfWeek` (0=Monday, 6=Sunday), `Quarter`, and `DayOfYear` were extracted from the `srcCalendarDay` column.
- **Impact:** T hese native integer features make it simple to do seasonal analysis, trending and pattern discovery in different time scales.

---

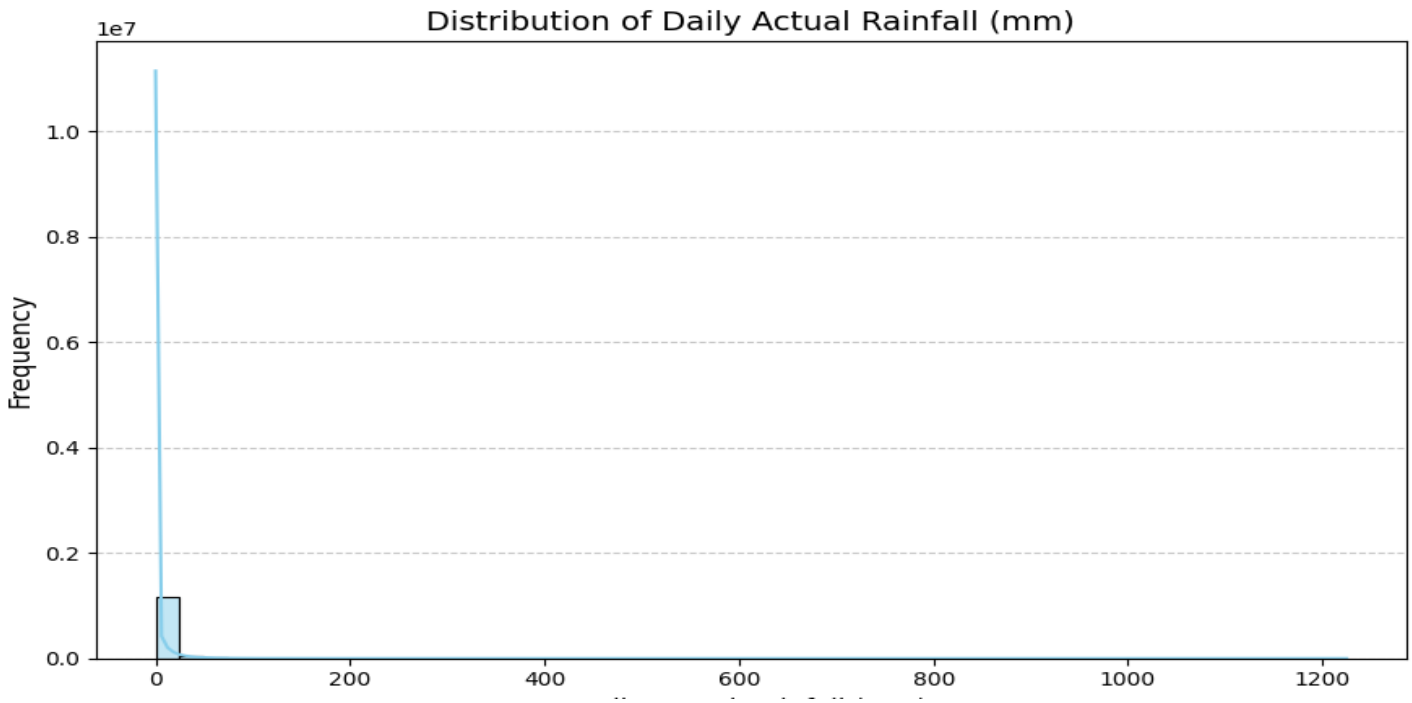
## 4. Univariate Analysis and India-Level Visualizations

This Here is the distribution of individual rainfall measures across India, and that gives the background of the national rainfall characteristics.

### 4.1 Distribution of Key Rainfall Metrics

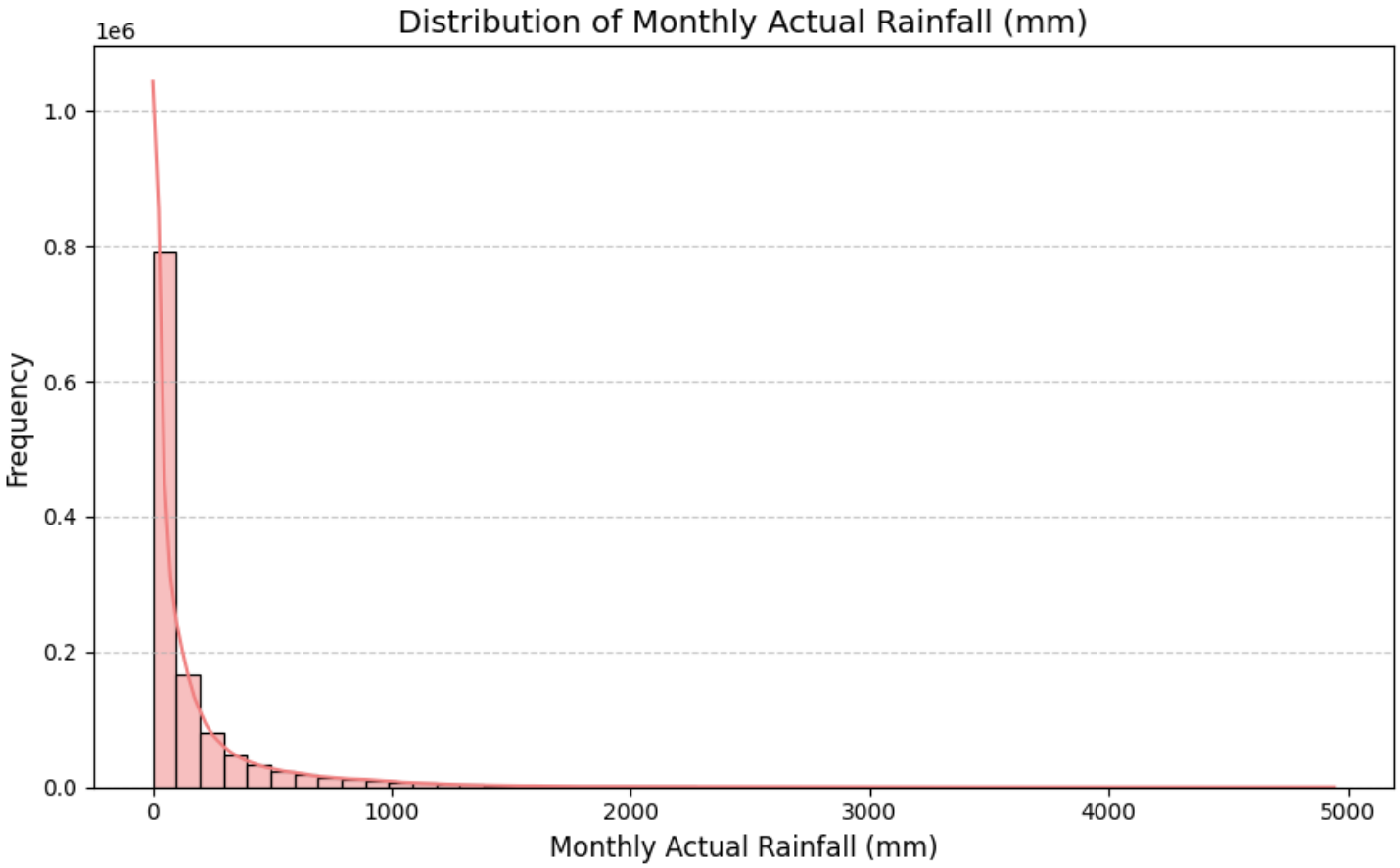
Histograms and Kernel Density Estimate (KDE) plots were employed to visualize the frequency and probability density of`Daily Actual`rainfall, `Monthly Actual` rainfall, and `Percentage of Daily Departure`.

**Figure: Distribution of Daily Actual Rainfall (mm)**



- **Observation:** In the histogram, the bar taking up an extremely tall place is at`0 mm`, which represents a huge majority of India in the daily observations reported no rain. The KDE curve also exhibits the point where a sharp peak shows up at`0 mm` and then quite a steep graph is shown further. The X-axis incorporates values as large as`1200 mm` and this indicates that, although this is a rare scenario, there are still very high rain fall events that happen daily hence there is a long right tail on the distribution.
- **Insight:** Such a distribution pattern is typical of meteorological data because the days of intensive and short-term precipitation come just after the days of no rain. It puts special emphasis on extreme variability of daily rainfall throughout the heterogeneous subcontinent of India.

**Figure: Distribution of Monthly Actual Rainfall (mm)**

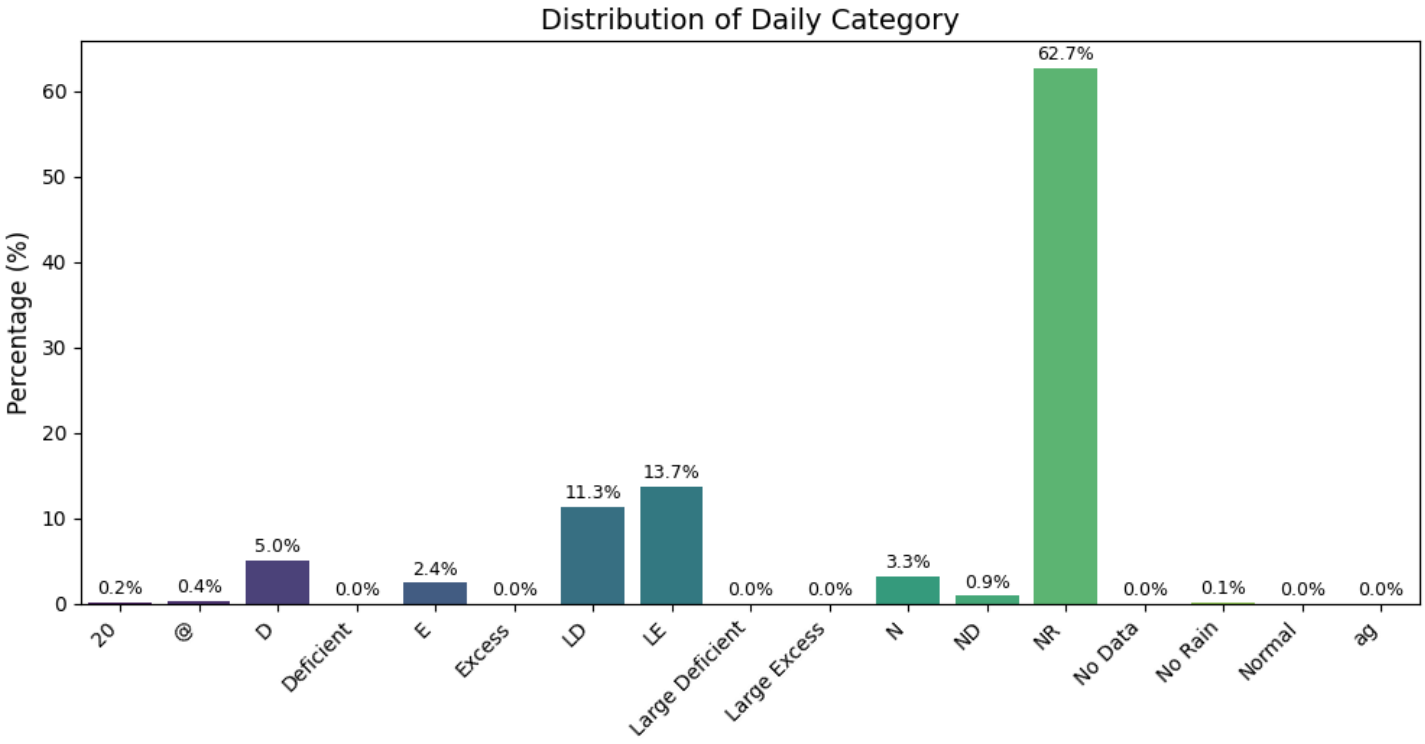



- **Observation:** As with the case of daily rain, the distribution of`Monthly Actual`rainfall is right- skewed in nature, meaning that it has a large mode at`0 mm`. But as compared to actuals of the day this peak is not very strong, and the distribution is much wider stretching to`5000 mm`.
- **Insight:** The summation of rainfall during one month is the reason that makes the range of total rainfall values to be more varied. Whereas there may be many months when one can only record dry conditions, when the rainfall is recorded during a monthly cycle the chance is higher that some rainfall will be recorded, or a significant quantity, than when the outcome only has to occur once in a given day.

## 4.2 Rainfall Category Distribution

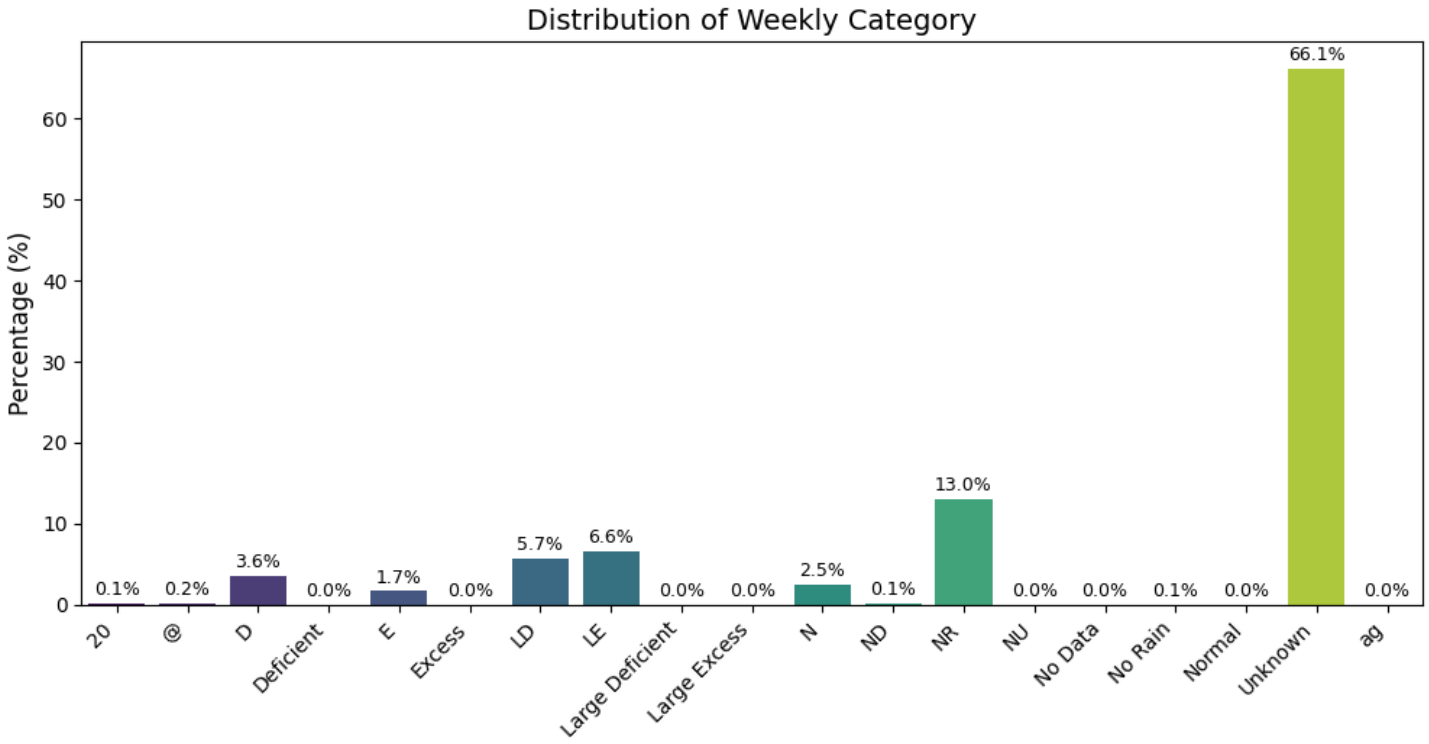
To depict the proportion of qualitative types of rainfall (e.g. Normal, Excess, Deficient) and visualise them at various levels of timescale bar charts were employed.

**Figure: Distribution of Daily Category**



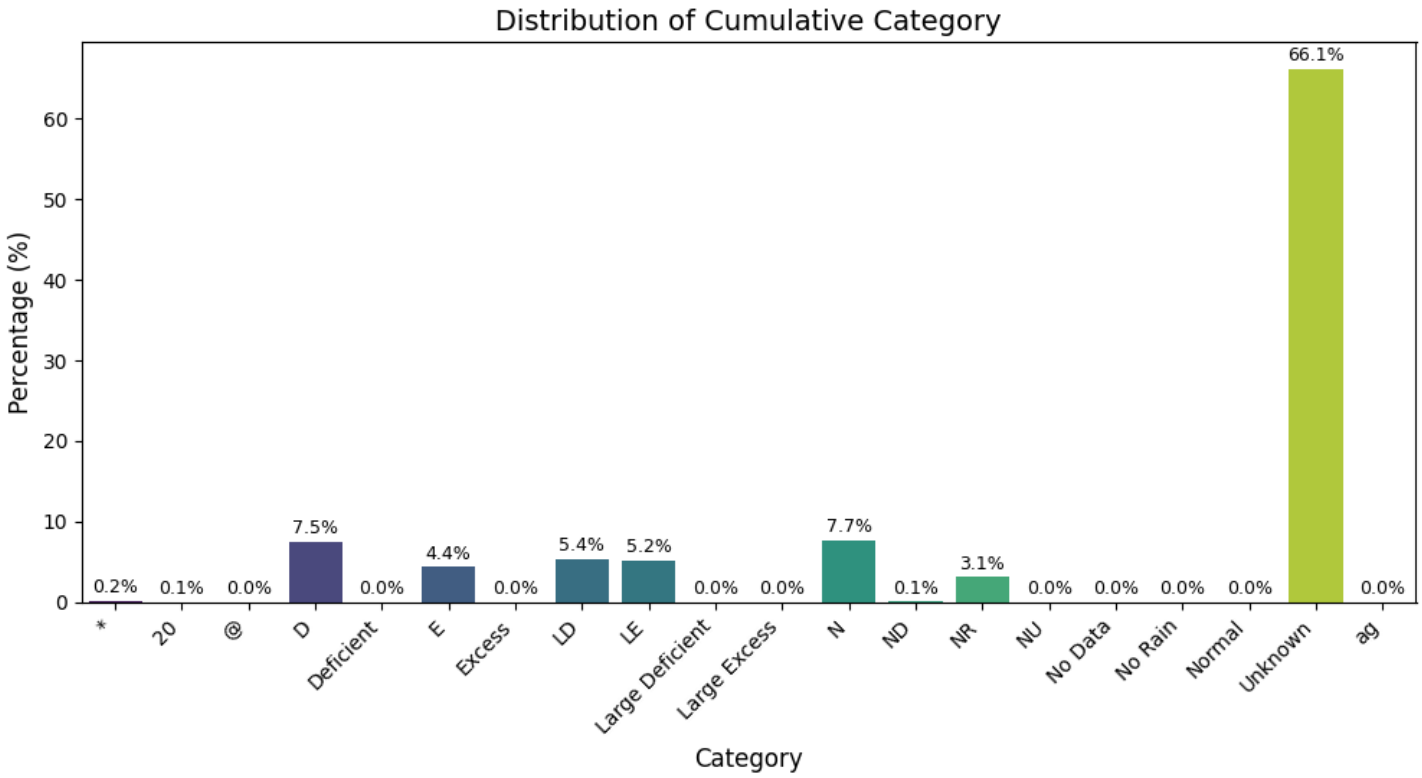

- **Observation:** The dominant type of no rain category of the NR('No Rain') is overwhelming with about `62.7%` share of observations on a daily basis. The next values after the above process are 'LE' (Large Excess) as`13.7%` and LD (Large Deficient) as`11.3%`.
- **Insight:** This quantitative spread supports the numerical fact that that a huge proportion of the days across India receive no rainfall at all. Moreover, even the rainfalls that do take place tend to fall into the extreme areas (that is extremely high or extremely low), thus demonstrating the unpredictability of the day-to-day weather conditions.

**Figure: Distribution of Weekly Category**



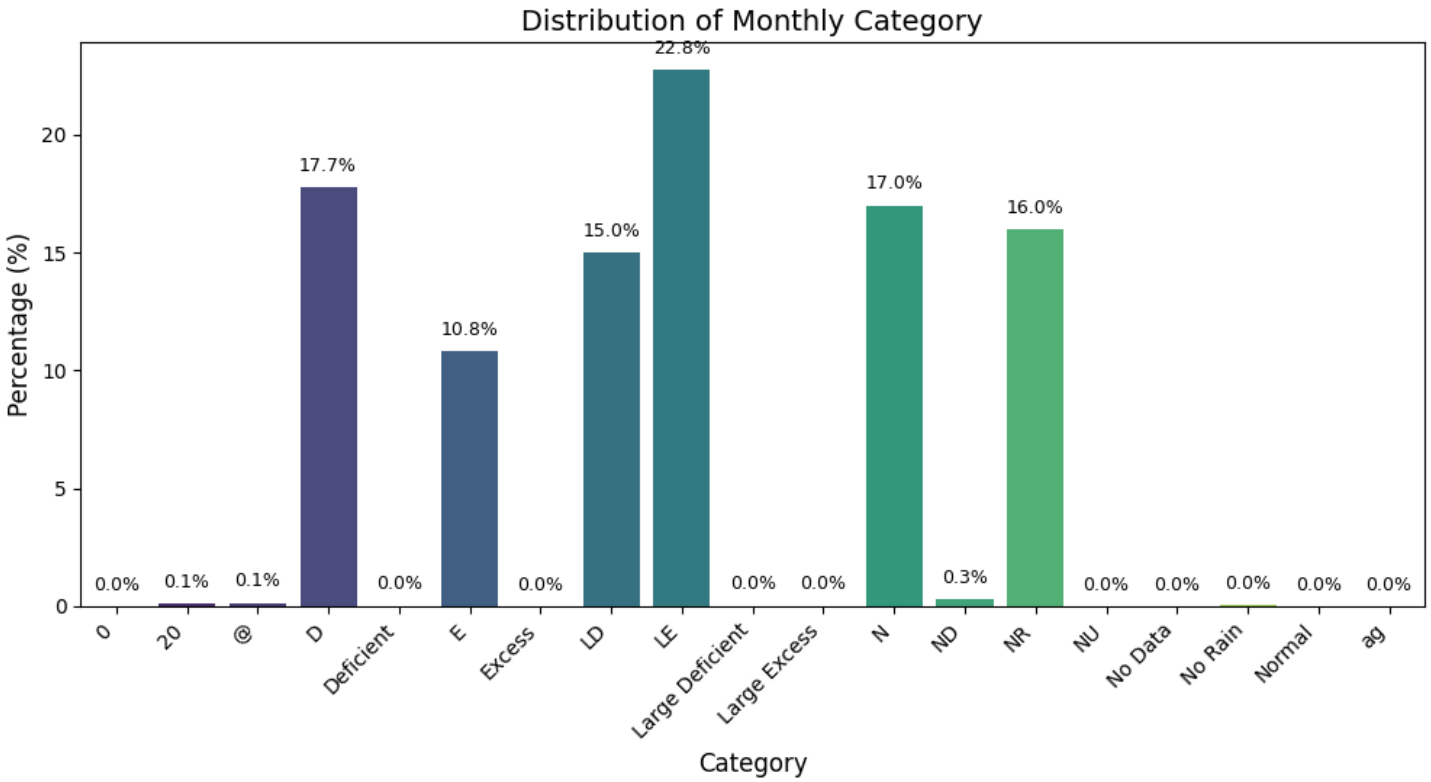

- **Observation:** This is because the category of the unknown takes up the significant of the weekly records `66.1%` and directly linked to the percentage of missing data on weekly records which was filled up. Of the known categories, the one with the highest frequency is the `13.0%` of the records falling under category NR (No Rain).
- **Insight:** The prevalence of Unknown highlights a serious issue with regard to unavailability of data at the level of weekly data. The No Rain condition still shows a significant figure in case of weekly data availability with any indication of a delayed dry spell or extended spell or period when there was no interesting gain in any week.

**Figure: Distribution of Cumulative Category**




- **Observation:** As in weekly data the Unknown category prevails also with the cumulative records comprising`66.1%`of the total. The available categories among them include the following: 'N' (Normal) with a percentage of`7.7%` and the close trailing second ranked category is the 'D' (Deficient) with a percentage of`7.5%`.
- **Insight:** The very high number of Unknown points to the nature of data incompleteness over cumulative periods. However, available records show an almost equal split between the categories of Normal and Deficient cumulative rainfall thus there is a balance between normal and less than normal seasonal totals.

**Figure: Distribution of Monthly Category**

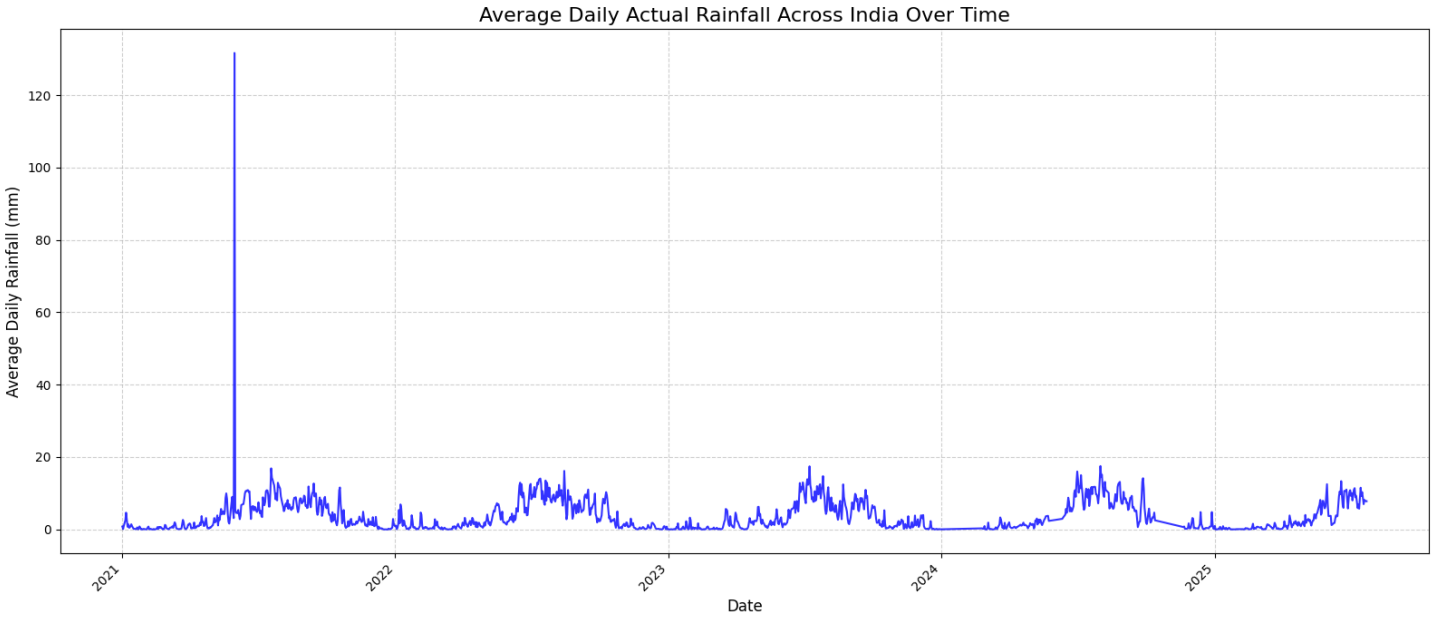



- **Observation:** The most common of the monthly category comes in form of `LE` (Large Excess) at `22.8%`. This is then followed by D (Deficient) at`17.7%` and N (Normal) at`17.0%`, NR (No Rain) at`16.0%` and LD (Large Deficient) at`15.0%`. An observation was also the fact that names of the categories such as D, Deficient, etc., denote the same idea, which is an aspect of future similarities in data standardization.
- **Insight:** India has a high percentage of months at a monthly level with levels of rains which are highly more than normal and this could have indicated high levels of monsoons. Yet, there is a significant size of months to be faced with the status of Defense or No Rain as well, which highlights the fluctuation across months and the co-occurrence of excess and deficit situations.

## 4.3 Overall Time Series

A line plot illustrating the average`Daily Actual`rainfall across India over time provides a macro-level view of national rainfall trends and seasonality.

**Figure: Average Daily Actual Rainfall Across India Over Time**

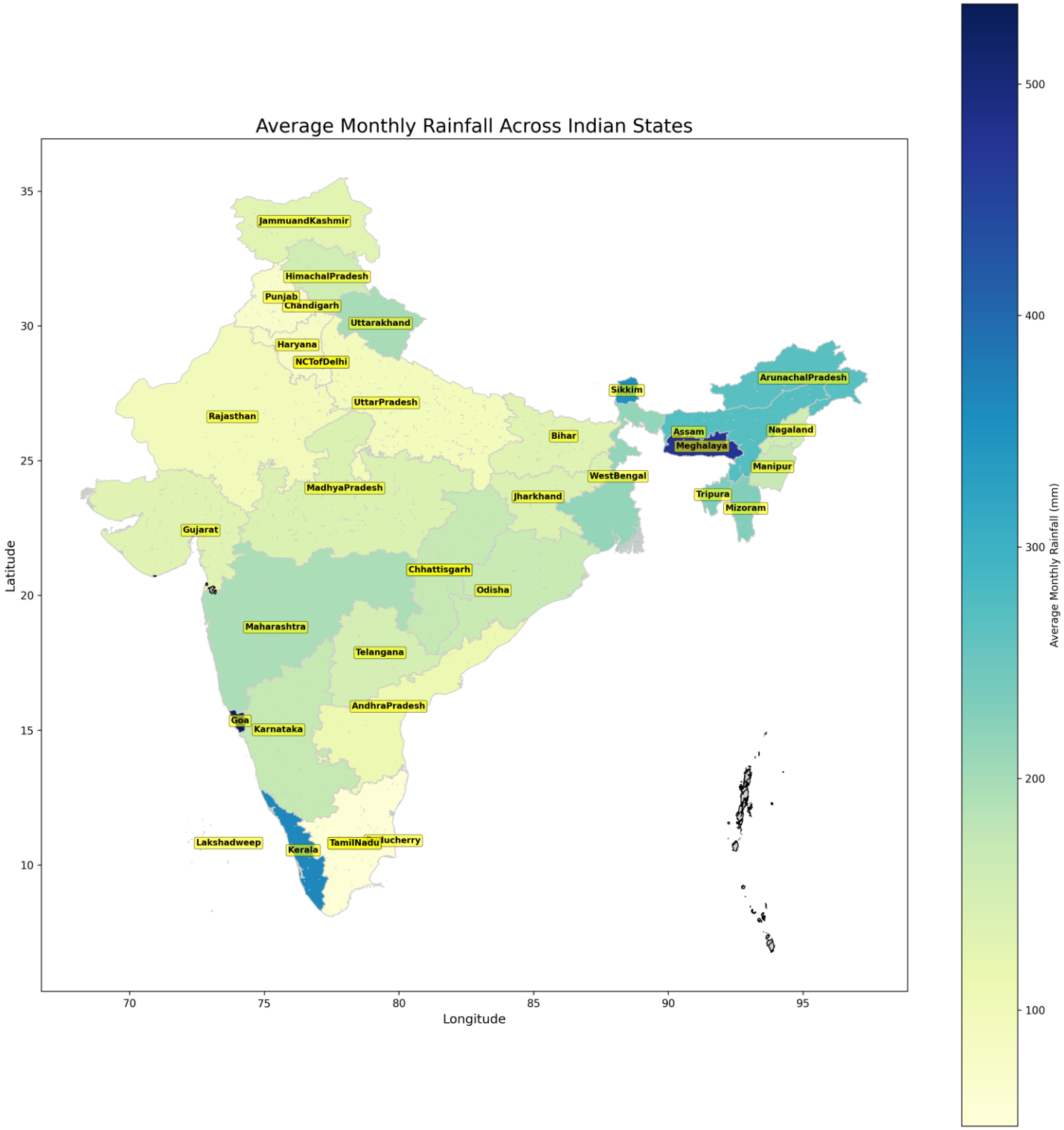



- **Observation:** In the time series plot, it may be observed that there are some clear annual cycles. High values of average daily rainfall are repeatedly noticed during the months of the usual Indian monsoon season, which spans over a period of about six months; half a year; of June to September, every year. Months of the year other than these tend to record an average rainfall that is much lower. Where the data density or peak height appears to decline significantly by the later years (e.g., 2024, 2025) this may be indicative that data is not fully collected during those recent years.
- **Insight:** It can be seen that in time series plot, there are evident annual cycles. The average daily rainfall in the months of the regular Indian monsoon season which stretches across the time duration of approximately half a year; six months; June to September, every year, is often seen with high values. The average rainfall in other months of the year is far much lower. In case the data density or the peak heights seem to fall dramatically in the later years (e.g. 2024, 2025) this can suggest that in those recent years, data is not fully retrieved..

## 4.4 India-Level Monthly Rainfall Distribution Map

In order to give the national picture of the rainfall distribution and consequently focus on the state of Maharashtra, a static choropleth map was created which showed the average of the monthly rainfall in all states of India. It uses aggregated rainfall data on a state-by-state level and state geographical information..

**Figure: Average Monthly Rainfall Across Indian States**




- **Observation:** The map brings out clearly the diffused rain zones in India. States that are mostly affected by Western Ghats areas, i.e., Goa, coastal Karnataka, Kerala, and sections of Maharashtra as well as the North-

Eastern states including Arunachal Pradesh, Assam, Meghalaya, Nagaland, Manipur, Mizoram, and Tripura can be seen in darker shades of blue. This means that they receive a much greater average monthly rainfall on account of orographic influences and direct exposure to currents of monsoons. States in the arid and semi-arid areas (e.g., Rajasthan, Gujarat, and interior Maharashtra) are depicted in paler tones of yellow/green, the uniform signal that the average annual rainfall of the mentioned states is lower. Central plains and the south states have moderate color saturations, and this could be attributed to the presence of different intensity of monsoons. More importantly, some territories (e.g., Andaman & Nicobar Islands, Dadra & Nagar Haveli and Daman & Diu, Jammu & Kashmir, Ladakh) are presented in the light grey with the hatch and the purpose of these displays is to indicate that there are no data available. This is mainly caused by the constant naming inconsistencies or confusing administrative boundaries representation in the GeoJSON that could not be completed by the dynamic fuzzy matching.

- **Insight:** This national scale map is an eloquent graphical synthesis of macro-climatic patterns of India. It highlights the enormous influence of the topography (consisting of high mountain ranges and the coast) as well as the fact that the branches of the monsoon did not penetrate the subcontinent equally. The high versus low rainfall areas are polar opposites and this is an instant clue on regional water resources availability. Although the improved mapping reduced incompatibilities, the other regions of absence of information indicate the current implementation difficulties of joining the data in the real world with the geospatial boundaries stipulating the importance of ensuring that in such studies, databases are carefully harmonized. This map is also very useful in establishing the broader hydrological context into which the more detailed, state-by-state, explorations would go.

---

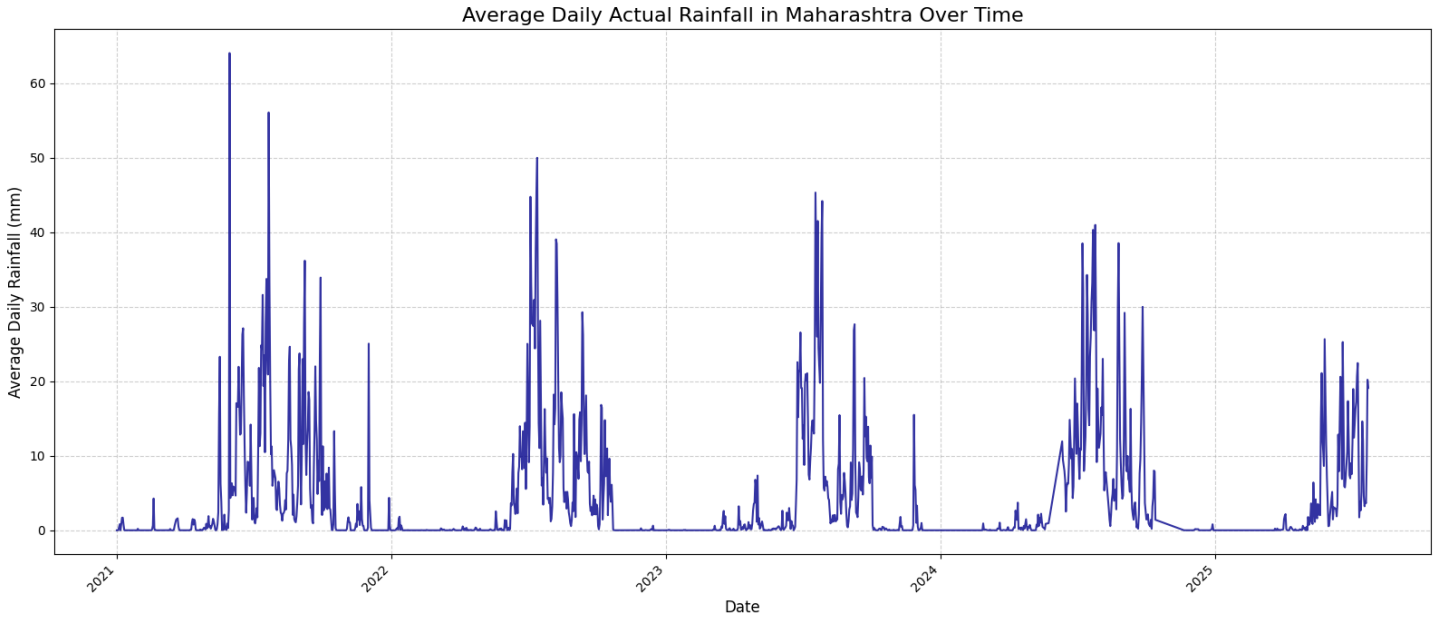
## 5. Zooming into Maharashtra - State-Level Analysis

This section provides a focused and detailed examination of rainfall patterns within the state of Maharashtra, a region of significant agricultural and economic importance.

### 5.1 Maharashtra State-Level Rainfall Trends

Analyzing the average daily actual rainfall over time specifically for Maharashtra reveals its state-specific seasonal patterns and any temporal anomalies.

**Figure: Average Daily Actual Rainfall in Maharashtra Over Time**

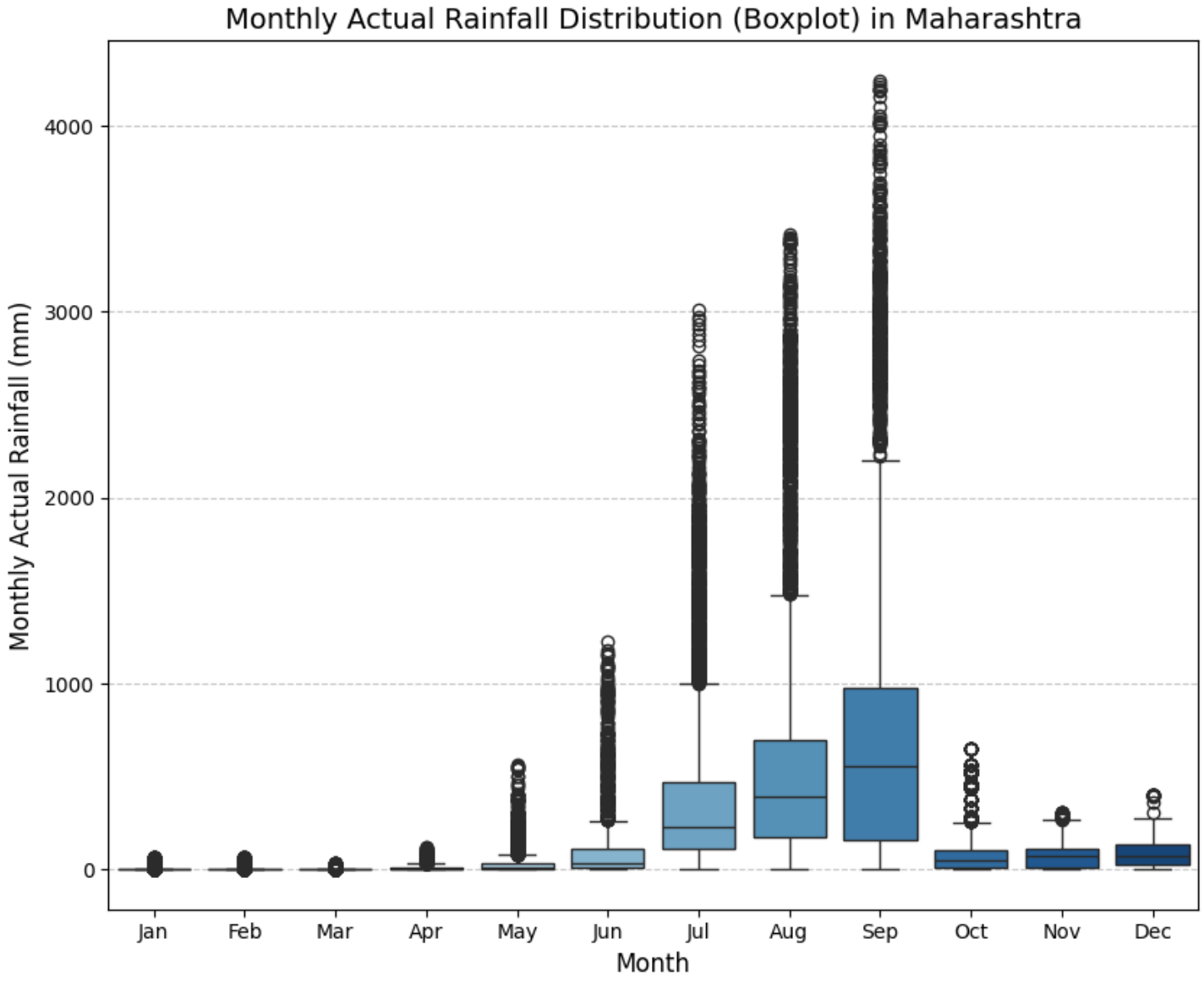


- **Observation:** The trend observed in such time series plot with filtering done only on Maharashtra shows clear evidence of a strong annual rainfall spike during the months of monsoon (usually between June and September). These peaks are overall rainy days in the state with an average rainfall, and the rainfall in these days is just high as compared to the days of other months since the rainfall is minimal in other months. The trend of changing time series in general in Maharashtra follows the national trend but shows the peculiarities of the monsoons in the state.
- **Insight:** The high seasonality can be very important in the planning of agriculture as the state largely relies on the rain-fed crops. It is extremely important to monitor the occurrence and strength of such peaks of monsoon as well as its reduction to allow managing water resources and estimating the threats of droughts or floods on the state scale.

## 5.2 Monthly Rainfall Distribution in Maharashtra

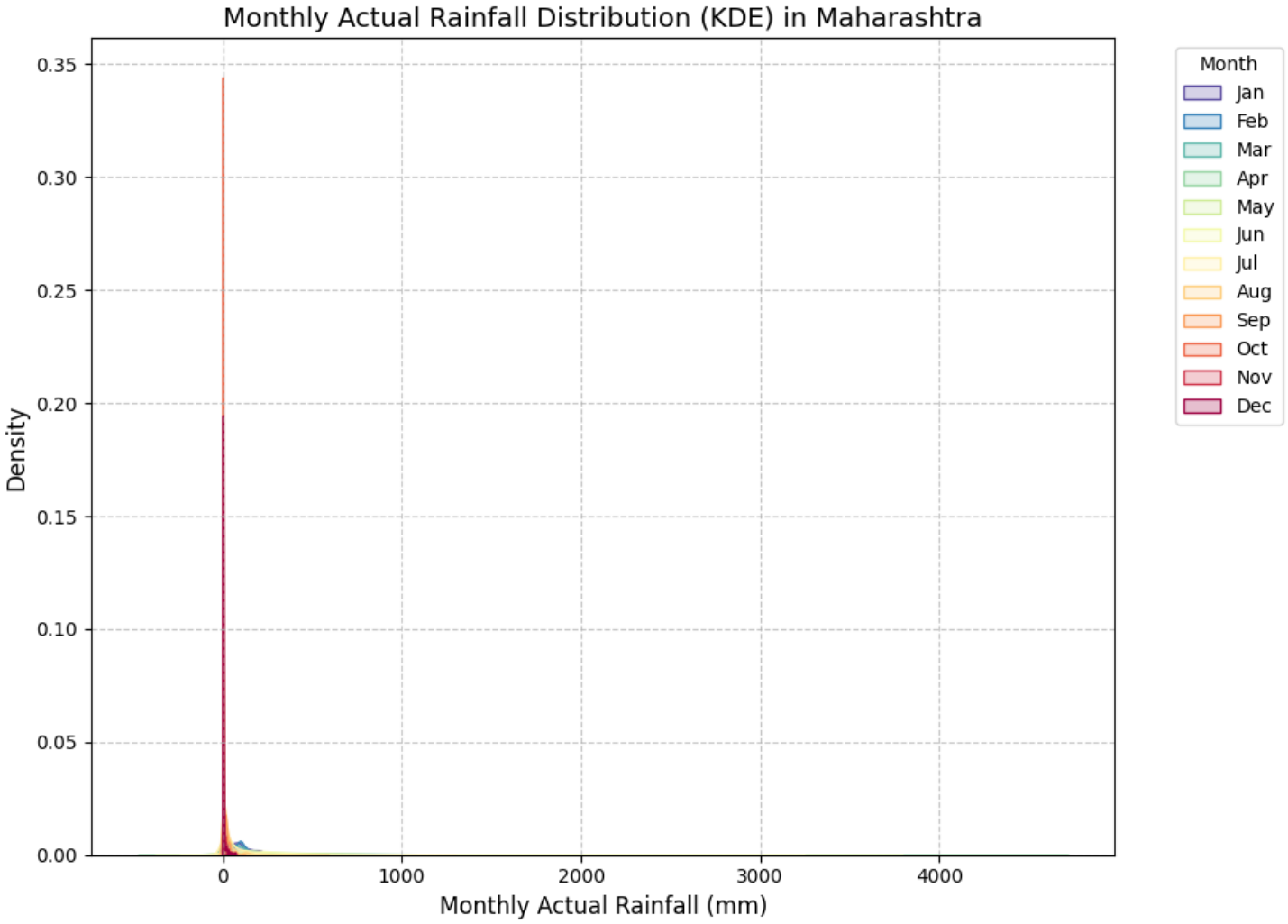
Boxplots and KDE plots were utilized to provide a granular view of the distribution and density of monthly actual rainfall within Maharashtra, analyzed across different months.

**Figure: Monthly Actual Rainfall Distribution (Boxplot) in Maharashtra**



- **Observation:** The monsoon months (June, July, August, September) and the drier months are well separated by the boxplot. These months record much higher median rain values during monsoon, and the interquartile ranges (IQRs) are much higher which means that there is greater variability and higher range of rainfall values. The months that do not fall within the monsoon window have either very low or no median rainfall with boxes close to the baseline. The other structural features that most often can be encountered during monsoon months are outliers which are individual points outside the whiskers and represent extreme occurrences of high rainfall.
- **Insight:** This visualization eloquently illustrates the very high concentration of the annual rainfall of Maharashtra, during the few months of monsoons. The high IQRs and outliers contribute to the high susceptibility of the state to episodes of heavy rainfall (causing local flooding) and the variability between months (affecting crop cycles).

**Figure: Monthly Actual Rainfall Distribution (KDE) in Maharashtra**




- **Observation:** The further analysis of the density of rainfall by a month is also given through the KDE plot. During monsoon months, a wider and flatter probability density curve can be visualized because the probability density curves are shown to lie on the higher end of rainfall. On the contrary, in case of dry months the curves are very much peaked around`0 mm` and tend to decrease drastically afterwards.
- **Insight:** This density expression strengthens the season Mulitplicity of rainfall, which gives a subtle interest to the probability of different rainfall quantities in each month. It points out how despite the fact that it is monsoon, there is a significant distribution of potential rainfall volumes, which means that we should have strong water management practices.

## 5.3 District-wise Breakdown of Average Rainfall in Maharashtra

This section provides a granular analysis of rainfall distribution at the district level within Maharashtra, combining quantitative summaries with a static geographical map.

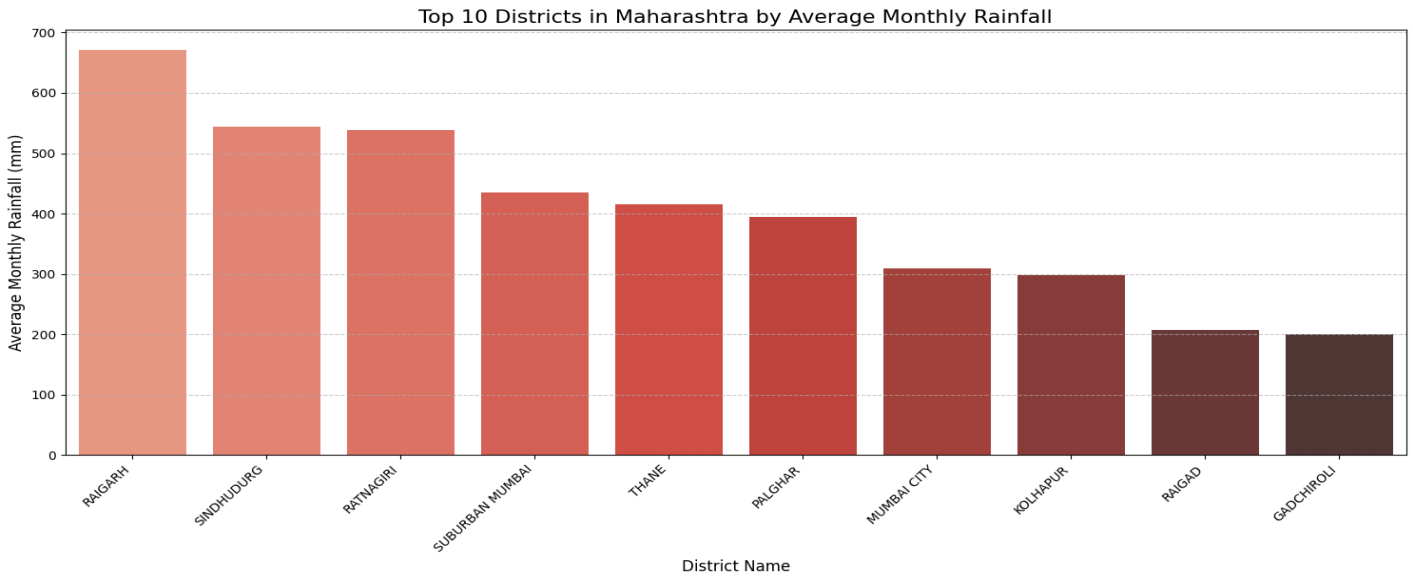
**Table: Top 10 Districts by Average Monthly Actual Rainfall in Maharashtra**

| srcDistrictName | Monthly Actual |
|---|---|
| RAIGARH | 670.79 |
| SINDHUDURG | 544.32 |
| RATNAGIRI | 538.31 |
| SUBURBAN MUMBAI | 435.33 |
| THANE | 415.42 |
| PALGHAR | 394.58 |
| MUMBAI CITY | 309.38 |
| KOLHAPUR | 297.66 |
| RAIGAD | 207.05 |
| GADCHIROLI | 200.34 |

**Table: Bottom 10 Districts by Average Monthly Actual Rainfall in Maharashtra**

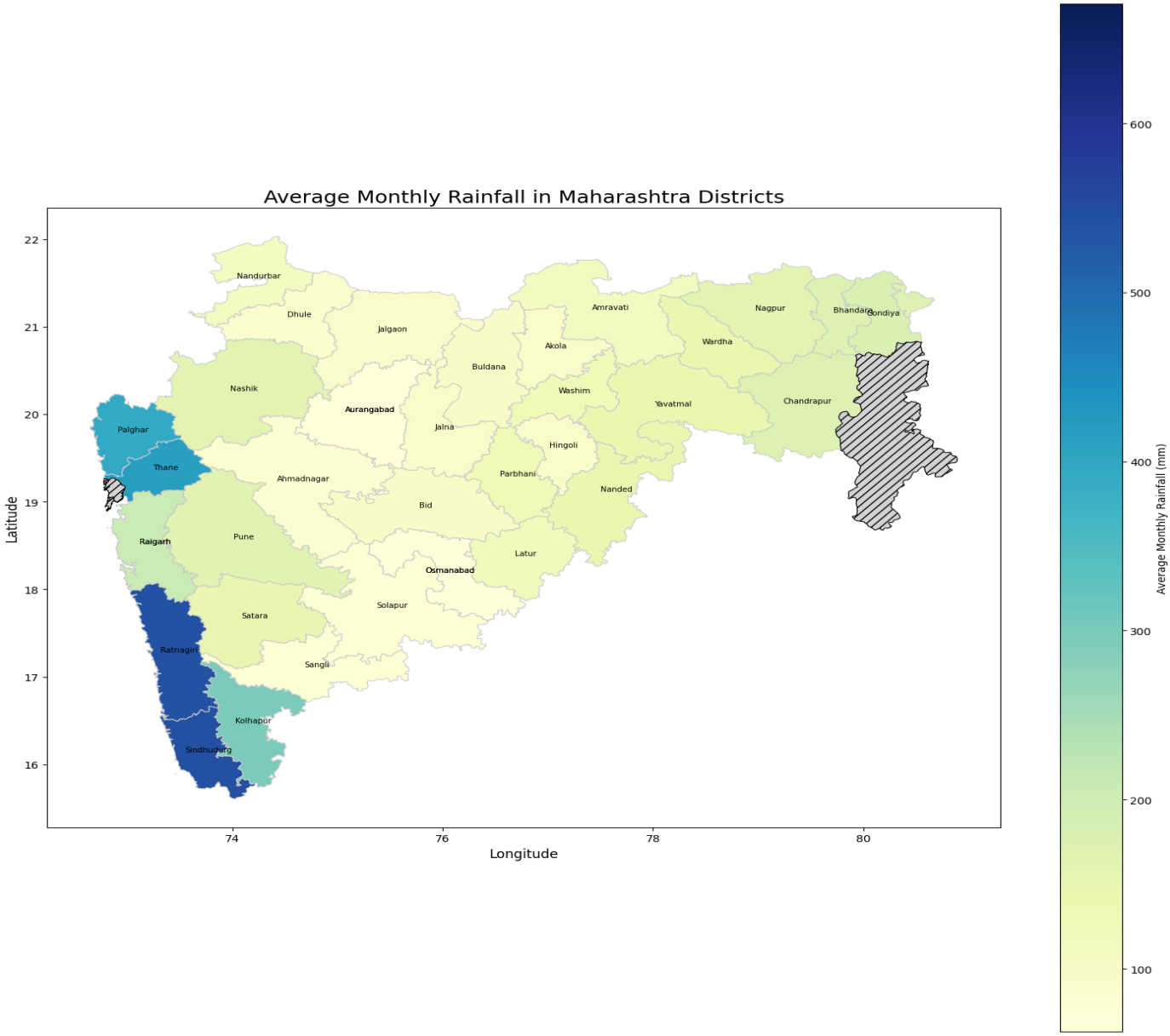
| srcDistrictName | Monthly Actual |
|---|---|
| JALNA | 94.85 |
| HINGOLI | 94.43 |
| AKOLA | 93.85 |
| JALGAON | 93.54 |
| DHULE | 88.47 |
| AHMADNAGAR | 81.19 |
| SANGLI | 76.93 |
| SOLAPUR | 76.77 |
| DHARASHIV | 67.51 |
| CHHATRAPATI SAMBHAJINAGAR | 62.78 |
| | |

**Figure: Top 10 Districts in Maharashtra by Average Monthly Rainfall**



- **Observation:** The bar chart substantiates the existence of a considerable difference between the aggregate monthly rains in Maharashtra in a numerical method by utilization of locales of the state. The top ones are the coastal districts of`RAIGARH`(670.79 mm),`SINDHUDURG`(544.32 mm) and`RATNAGIRI`(538.31 mm) with proximity being a clear indication towards their location in the Western Ghats and direct access to the monsoon winds. The districts of Mumbai (`SUBURBAN MUMBAI`,`MUMBAI CITY`,`THANE`,`PALGHAR`) also come in the category of the high-rainfall areas. On the other hand, the rainfall is entirely different in some districts (such as - `CHHATRAPATI SAMBHAJINAGAR` - 62.78 mm, `DHARASHIV` - 67.51 mm and `SOLAPUR` - 76.77mm) giving an average less rainfall per month, which makes them into rain-shadow areas of the state.
- **Insight:** This this gives us a great idea of what the different geography means to the micro-climates of Maharashtra and mytho-geography. Western ghats make an effective barrier leading to heavy orographic precipitation of the windward (Konkan) side and the development of rain-shadow effects in the leeward region (parts of central Maharashtra).

**Figure: Average Monthly Rainfall in Maharashtra Districts (Static Map)**



- **Observation:** The static choropleth map is also a very eloquent display of the spatial distribution of the average monthly rainfall in Maharashtra. The darker blues with the higher rainfall are mostly towards the western coastal districts, which falls in line with the results of the bar chart. And the intensity of the color changes, becoming lighter and lighter as one moves east, across central Maharashtra, into the rain-shadow areas. Regions that are shown as light grey and hatched pattern (e.g., some parts of eastern Maharashtra and perhaps certain areas of Mumbai) are those districts whose rainfall value could not be correlated exactly with the geographical extent as shown by the GADM GeoJSON data.
- **Insight:** This map would be of great help when one visually identifies areas of high rainfall (i.e., the Konkan belt) and deficit areas in the state. It enables the rapid evaluation of the availability of water in the system and the regions likely to be affected by droughts and necessary policy responses. The No Data areas exemplify the difficulties in practical aspects of combining incompatible data and differing data standards of naming and definitions of administrative boundaries, which frequently necessitate time-consuming one by one mapping or complicated fuzzy matching strategies to provide full geospatial coverage.

---

## 6. Comparative Analysis - Maharashtra vs. Neighboring States

To provide a broader regional context, this section compares Maharashtra's rainfall characteristics with its border-sharing states: Gujarat, Madhya Pradesh (MP), Chhattisgarh, Telangana, Karnataka, and Goa.

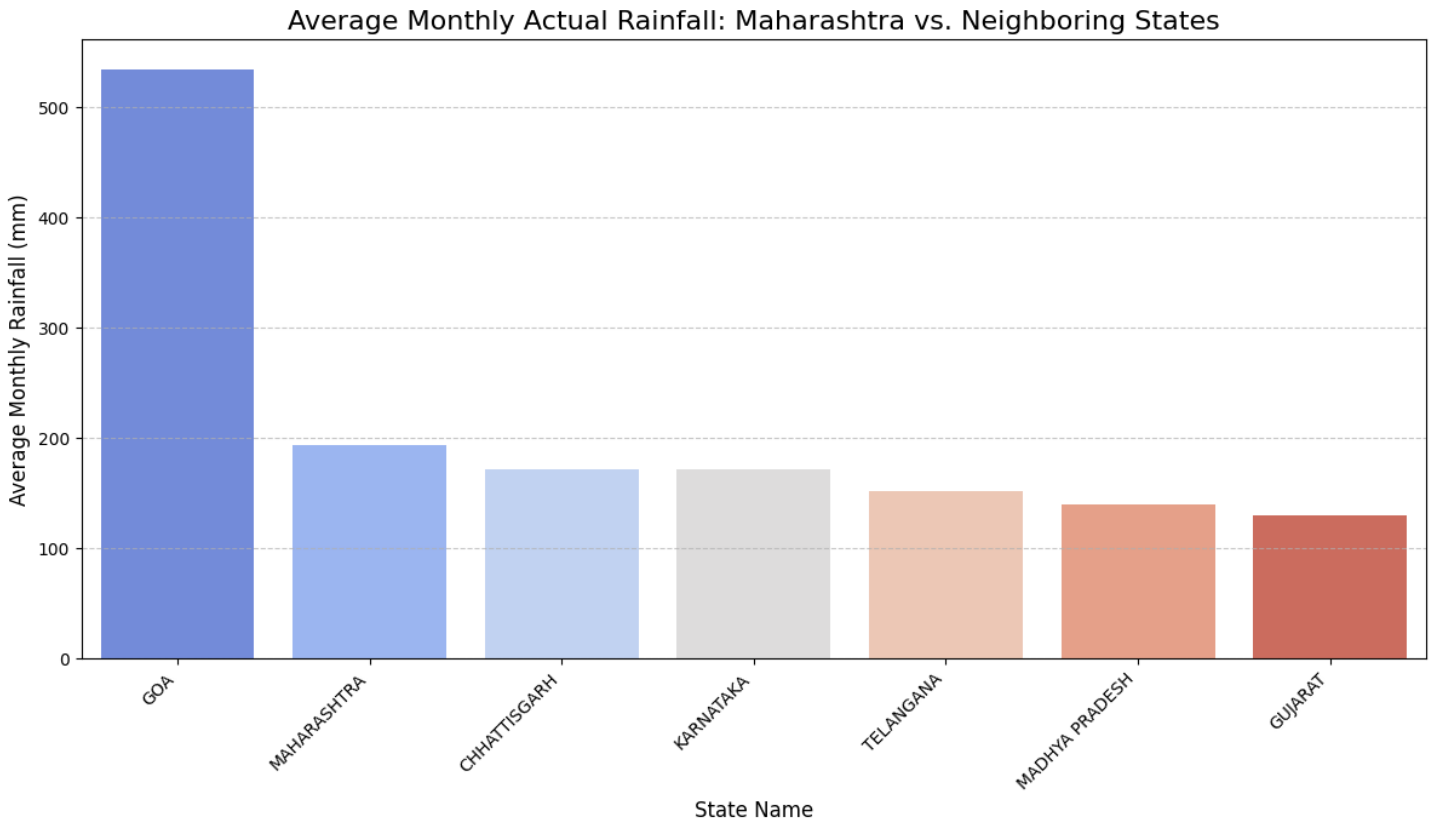
### 6.1 Average Monthly Rainfall Comparison

A bar chart was generated to directly compare the average monthly actual rainfall across Maharashtra and its neighboring states.

**Table: Average Monthly Rainfall (mm) for Maharashtra and Neighboring States**

| srcStateName | Monthly Actual |
|---|---|
| GOA | 534.58 |
| MAHARASHTRA | 193.90 |
| CHHATTISGARH | 172.03 |
| KARNATAKA | 171.41 |
| TELANGANA | 151.63 |
| MADHYA PRADESH | 139.34 |
| GUJARAT | 130.47 |
| | |

**Figure: Average Monthly Actual Rainfall: Maharashtra vs. Neighboring States**

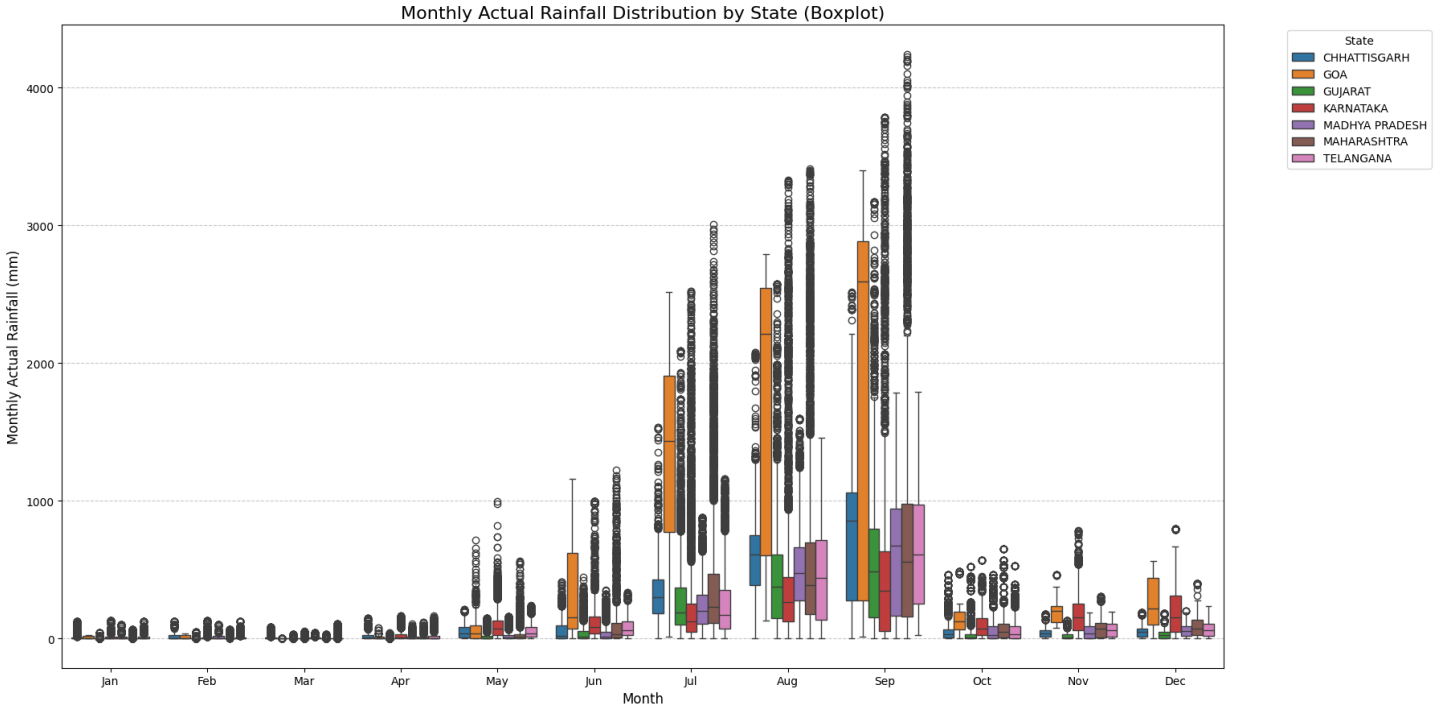



- **Observation:** Goa has the highest average monthly rainfall of`534.58 mm` and thus having the highest ranking in confirming it as a coastal region with intense rainfall. The other state on second position is Maharashtra with `193.90 mm`. Chhattisgarh (`172.03 mm`) and Karnataka (`171.41 mm`) are also close in range when compared to average rainfall. The average monthly rainfall is steeply decreasing in Telangana (`151.63 mm`), Madhya Pradesh (`139.34 mm`), and Gujarat (`130.47 mm`) respectively.
- **Insight:** This comparison brings about a clear view of how the climate is different in the west and central parts of India. The extremely high level of rainfall in Goa is a direct result of being on the windward side of the Western Ghats. The location of Maharashtra also makes it vulnerable to a definite overall monsoon impact and thus it fares better as far as the volume of rainfall is concerned as compared to its neighbors to the north (Gujarat and MP) and east (Telangana) but of course it has lower rainfall as compared to the concentrated band of coastal Goa.

## 6.2 Monthly Rainfall Distribution Comparison (Boxplots)

Boxplots were employed to compare the distribution of monthly actual rainfall across Maharashtra and its neighboring states, month by month, offering insights into variability and seasonality.

**Figure: Monthly Actual Rainfall Distribution by State (Boxplot)**




- **Observation:** All the boxplots are showing a unique pattern of seasonal rainfall peak at different seasons but mostly in the monsoon season between June to September. Goa records the largest medians of rainfall and largest interquartile ranges (IQRs), indicating very heavy rain which is highly variable, in these months of monsoons. Maharashtra boasts of good monsoon rains also with a lot of the rain coming in between June and September. Inner states such as Madhya Pradesh and Gujarat tend to show lower values in medians and narrow IQRs implying weaker or less consistent monsoons than in the western coastal states. The fact that there are a lot of outliers in all the states, especially the ones related to monsoon, points to regular occurrence of extreme precipitation events.
- **Insight:** The minute level monthly difference shows variations in features of the monsoon in the area. States that have larger medians and IQRs (such as Goa and coastal Maharashtra) are more vulnerable to flooding as well as enjoy more water security. On the other hand, those states that have lower medians and a smaller IQR might be more susceptible to drought, which highlights the varying problems of approaches to agriculture and control of water in the area.

## 6.3 Rainfall Departure Category Comparison (Heatmap)

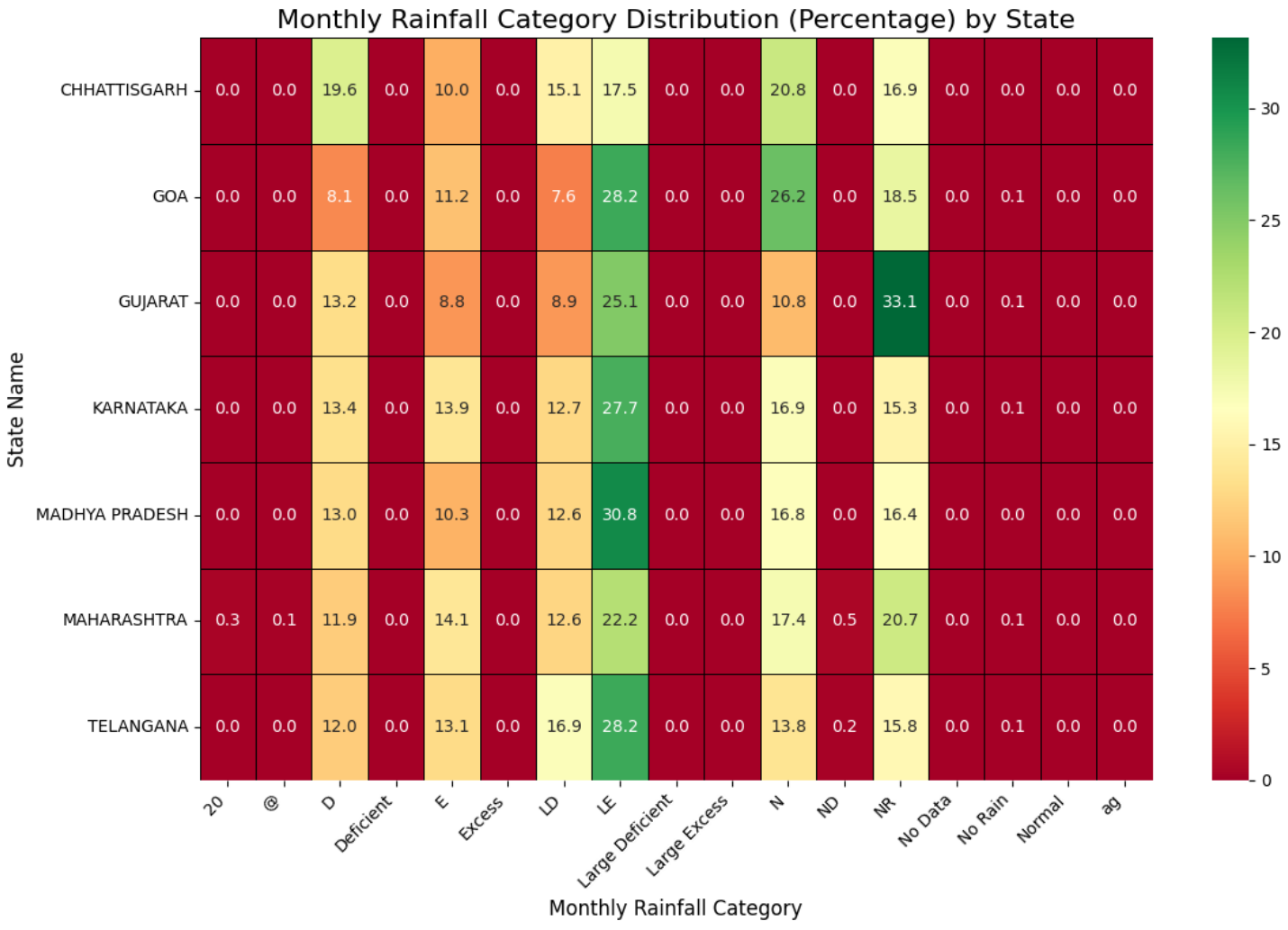
A heatmap visualizes the percentage distribution of dominant rainfall categories for each state, providing a qualitative comparison of rainfall performance relative to historical norms.

**Table: Monthly Rainfall Category Percentages per State**

| Monthly Category | CHHATTISGARH | GOA | GUJARAT | KARNATAKA | MADHYA PRADESH | MAHARASHTRA | TELANGANA |
|---|---|---|---|---|---|---|---|
| D | 19.6 | 8.1 | 13.2 | 13.4 | 13.0 | 11.9 | 12.0 |
| E | 10.0 | 11.2 | 8.8 | 13.9 | 10.3 | 14.1 | 13.1 |
| LD | 15.1 | 7.6 | 8.9 | 12.7 | 12.6 | 12.6 | 16.9 |
| LE | 17.5 | 28.2 | 25.1 | 27.7 | 30.8 | 22.2 | 28.2 |
| N | 20.8 | 26.2 | 10.8 | 16.9 | 16.8 | 17.4 | 13.8 |
| NR | 16.9 | 18.5 | 33.1 | 15.3 | 16.4 | 20.7 | 15.8 |
| Unknown | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 |

*(Note: For every category, those with redundant names such as 'Deficient', 'Excess', 'Normal', 'No Rain', 'Large Deficient', 'Large Excess', having values of 0.0 percent are omitted since there are no values, in this table, after standardization or are blank are left out as far as brevity is concerned. This is done because they do occur but do not have values in this table after standardization or are blank.)*

**Figure: Monthly Rainfall Category Distribution (Percentage) by State**



- **Observation:** Heatmap is a visual representation of the comparison of percentage distribution of various monthly rain fall categories across the states where `LE` (Large Excess) is a prominent category in most of the states especially in the cases of Madhya Pradesh (`30.8%`), Karnataka (`27.7%`), Telangana (`28.2%`), and Goa (`28.2%`). Maharashtra too possesses a good percentage of `LE`(`22.2%`). On the other hand, `NR`(No Rain) is very high at Gujarat (`33.1%`) and Goa (`18.5%`) and `D`(Deficient) is comparatively high at Chhattisgarh (`19.6%`) and Maharashtra (`11.9%`).
- **Insight:** This contrast does show that even though many states regularly had months in which rainfall was far-above-normal, the states are not equivalent in the way they are susceptible to droughts and-little-rain situations. Gujarat recorded high percentage of no rain which implies longer dry periods. The distribution in Maharashtra is more balanced and implies that it is a climate not only of surplus but also deficit rainfall which agrees with different agro-climatic regions of the state. During preprocessing, it was remarked that duplication of column names existed in the raw data (e.g., a column called 'D' and another column called, Deficient) which could easily be simplified to protrude a consolidated appearance.

---

## 7. Conclusion

The whole Exploratory Data Analysis performed on the dataset about Indian rainfall has brought to light some important conclusions about the intricate spatio-temporal nature of rain in the country, and perhaps more importantly, the state of Maharashtra and the regional perspective. The process of data cleaning and feature engineering were strict, and their quality was high and reliable for further analysis.

This EDA can confirm such main conclusions:

- **Extreme Skewness of Rainfall:** Skewness of rainfall distribution is high at both daily and monthly scales, which means that most of the times go with no or minimal rainfall. But these are interrupted by very severe episodes of rainfall which are however not frequent. This is the `on-off` nature of rainfall which is a characteristic.
- **Predominance of Deficits and Extremes:** The major part of the observed daily data is 100 % deficit (no rain). On monthly levels, `Large Excess`rainfalls are common but `Deficient` situations are also largely high, which indicates the vulnerability to both floods and droughts.
- **Strong Monsoon Seasonality:** The Indian monsoon towers over all other causes of rain and has clear and identified seasonal peaks in June-September nationwide (as well as within Maharashtra). Agricultural productivity and water security are determined by the monsoon performance level.
- **Major Regional Inequality in Maharashtra:** There are substantial regional disparites in Maharashtra because of the effects of geographical influence, especially the western ghats, on rainfall distribution. There is an unusually high rainfall on the east coast districts (e.g., Raigarh:`670.79 mm` average monthly) as compared to the comparatively drier rain-shadow areas in central Maharashtra (e.g., Chhatrapati Sambhajinagar:`62.78 mm` average monthly).
- **Different Regional Climates:** Comparative assessment along the fringes with other states shows the existence of varied rain patterns. Goa is the most outstanding with an extremely high average rainfall (`534.58 mm`) even as Maharashtra lies in a middle position with more rainfall as compared to the central and the northern states but not as much compared to the very concentrated areas on the Southwest Coast. There is also much variation by state in the balance of 'Excess vs. Deficient months.
- **Data Integrity Challenges:** EDA process identified typical data issues that occur in practice - inconsistencies in naming of districts in the rain data and geographic files creating issues with mapping

(done manually), redundant categories that were labeled. These were important to overcome to succeed in correct visualization and interpretation.

This EDA is an excellent source of analytical principles, and therefore, it can be used to make very important observations about the rainfall patterns of India that are important on a scientific and a policy-making basis as well as on climate adaptation and response strategies.