**Rain Forecasting**

**Introduction**

Accurate rainfall forecasts are essential for Mumbai's water resource management. This case study examines the development of a machine learning model to predict monthly rainfall in Mumbai, leveraging historical data to improve water allocation, storage, and distribution strategies. Effective rainfall prediction can lead to optimized reservoir management, reduced operational costs, and a more consistent water supply throughout the year.

**Data Description**

The dataset used in this study consists of monthly rainfall data for Mumbai from 1901 to 2021. The data includes:

* Monthly rainfall amounts (in mm or a similar unit) for each month (January to December).
* The corresponding year.
* The annual total rainfall.

**Problem Statement**

The challenge is to develop a predictive model that can accurately forecast monthly rainfall in Mumbai.

This model will assist water authorities in making informed decisions regarding:

* Reservoir levels and releases.
* Water distribution planning.
* Drought mitigation strategies.
* Flood control measures.

**Dataset Information:**

The total number of columns is 14.

Month columns:- Jan, Feb, March, April, May, June, July, August, Sept, Oct, Nov, Dec

Other columns:- Years, Total

Total column is sum of month value total

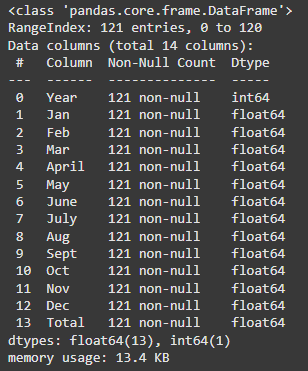
In the **Year** column data for years showing 1901 from 2021 records are available

Other columns have holding float datatype

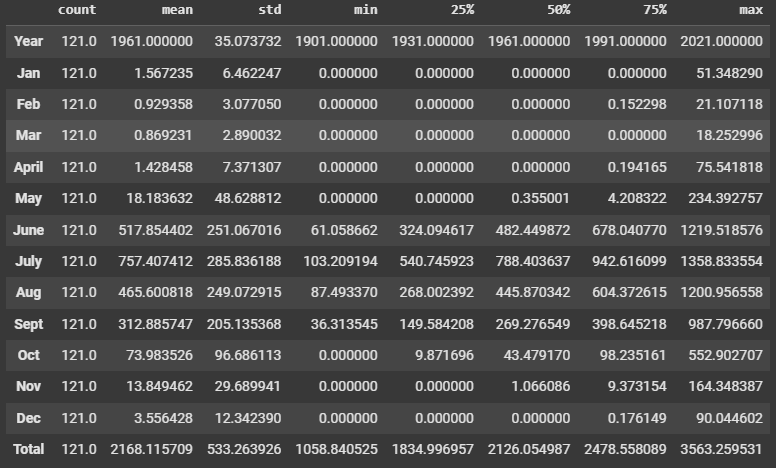
There are 121 entries 0 to 120

Data types: float64(13), int64(1)

memory usage: 13.4 KB



No, Duplicate values are present in the dataset. Also not have any null value present

After the data got checked duplication and null value evaluation **df.describe()** check description of the dataset

This is a very small set of datasets before creating any graphical institution, what we learned from this description.

This dataset total is based on a location in Mumbai and we see here these datasets are from 1901 to 2021 total is 121 years from 1901, with monthly and yearly precipitation patterns from 1901 to 2021 utilized for forecasting upcoming rain sessions of 2025 we lack some years

Precipitation data was analyzed by categorizing it into four seasons:

(1) **Non-monsoon** (January, February, March)

(2) **Pre-monsoon** (April, May)

(3) **Southwest monsoon** (June, July, August, September)

(4) **Northeast monsoon** (October, November, and December)

then basic statistical analysis finds out the maximum and minimum rainfall, mean, standard deviation, skewness, and Kurtosis.

**ASK**

**Why is there so much rain in Mumbai?**

The heavy rainfall in Mumbai is primarily due to the Southwest Monsoon and orographic influences from the nearby Western Ghats. Western Ghats also known as the **Sahyadri.**  These factors result in a prolonged wet season with intense rainfall.

**Why is Mumbai suffering from a water crisis?**

The first reason, Mumbai has a massive population, [As of the year 2024, the population of Mumbai, India was over 21.6 million inhabitants. This was a 1.77 per cent growth from last year. The historical trends indicate that the population of Mumbai has been steadily increasing since 1960.](https://www.statista.com/statistics/911012/india-population-in-mumbai/#statisticContainer) Leading to high water demand that often outstrips the available supply.

**Dependence on monsoon rains:**

The city largely relies on monsoon rains to fill its reservoirs, making it vulnerable to erratic rainfall patterns due to climate change

**Inefficient water distribution:**

Outdated pipelines and uneven distribution systems lead to significant water losses and inadequate access to clean water in certain areas, particularly slum

**Water extraction from distant sources:**

Most of Mumbai's water is sourced from dams located far away, which can cause issues with transportation and distribution

**Lack of water conservation practices:**

Inadequate awareness and implementation of water-saving measures within the city contribute to the problem

**How much water is needed in Mumbai in a year?**

According to recent reports, Mumbai currently requires around **4,463 million litres** of water per day (MLD) meaning the city needs approximately **1,626,000 million litres** of water per year based on current demand.

**How much water is supplied in Mumbai in the year?**

According to available data, Mumbai receives approximately 3,850 million litres of water per day, which translates to roughly 1.4 billion litres per year, supplied by the [BMC (Brihanmumbai Municipal Corporation)](https://www.mcgm.gov.in/irj/go/km/docs/documents/MCGM%20Department%20List/City%20Engineer/Deputy%20City%20Engineer%20(Planning%20and%20Design)/City%20Development%20Plan/Urban%20Basic%20Services.pdf)

**1. Water Demand:**

* Mumbai's daily water demand: 4,463 million litres (MLD)
* Mumbai's annual water demand: 4,463 MLD 365 days/year = 1,628,995 million litres (approximately 1.63 trillion litres)

**2. Water Supply:**

* Mumbai's daily water supply: 3,850 million litres (MLD)
* Mumbai's annual water supply: 3,850 MLD 365 days/year = 1,405,250 million litres (approximately 1.41 trillion litres)

**3. Water Deficit:**

* Annual water deficit: 1,628,995 million litres (demand) - 1,405,250 million litres (supply) = 223,745 million litres (approximately 223.75 billion litres)

**What questions should we ask BMC and stakeholders to learn more about the dataset, its limitations, requirements, and results?**

What are the BMC's needs and requirements for improved rainfall forecasting models? (e.g., lead time, accuracy, spatial resolution). How could the dataset we have be used to develop or improve these models?

How are rainfall forecasts incorporated into water management planning and operations?

What methods are currently used for rainfall forecasting in Mumbai?

How accurate are the current rainfall forecasts, and what are their limitations?

**Rainwater Harvesting:**

What is the potential for rainwater harvesting to supplement Mumbai's water supply, given the historical rainfall patterns?

What other data would be needed to complement the rainfall data and make it more useful for planning purposes (e.g., land use data, population data, evaporation rates)?

How could the historical rainfall dataset be used to inform urban planning decisions related to water management?

Is the BMC open to sharing its own rainfall data or other relevant water resources data with researchers or the public?

Does the BMC have any data or estimates on the potential contribution of rainwater harvesting to the city's water supply?

How does the BMC promote and incentivize rainwater harvesting in Mumbai?

What are the known limitations in the accuracy of current rainfall forecasts?

How frequently are the models updated or retrained?

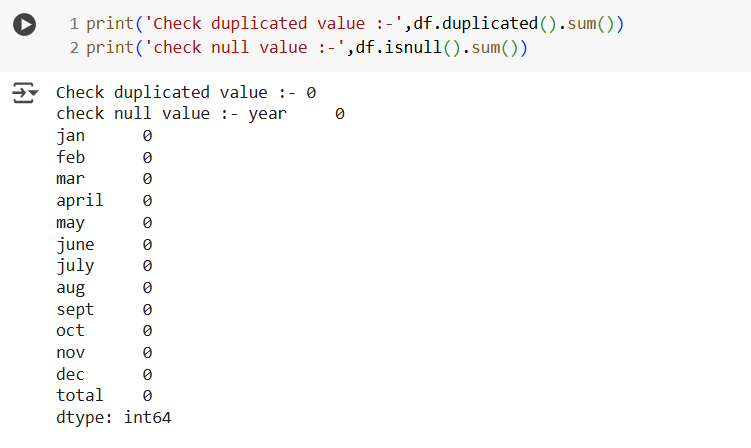
Are forecasts generated for specific locations within Mumbai, or are they more general for the entire metropolitan area?

What is the typical lead time for rainfall forecasts?

Are different lead times used for different purposes?

**Data Preprocessing**

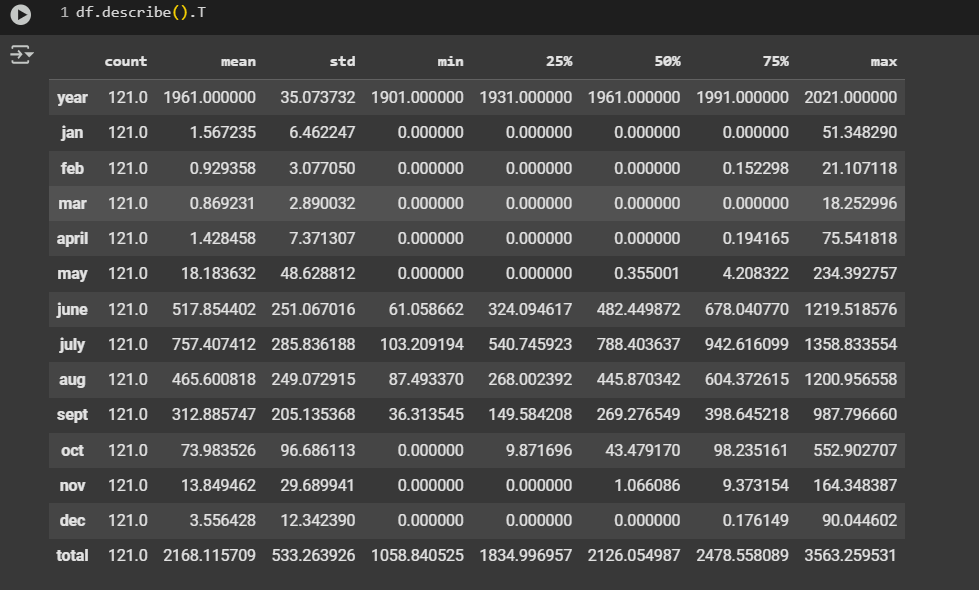
Checking Duplicated values and Missing values

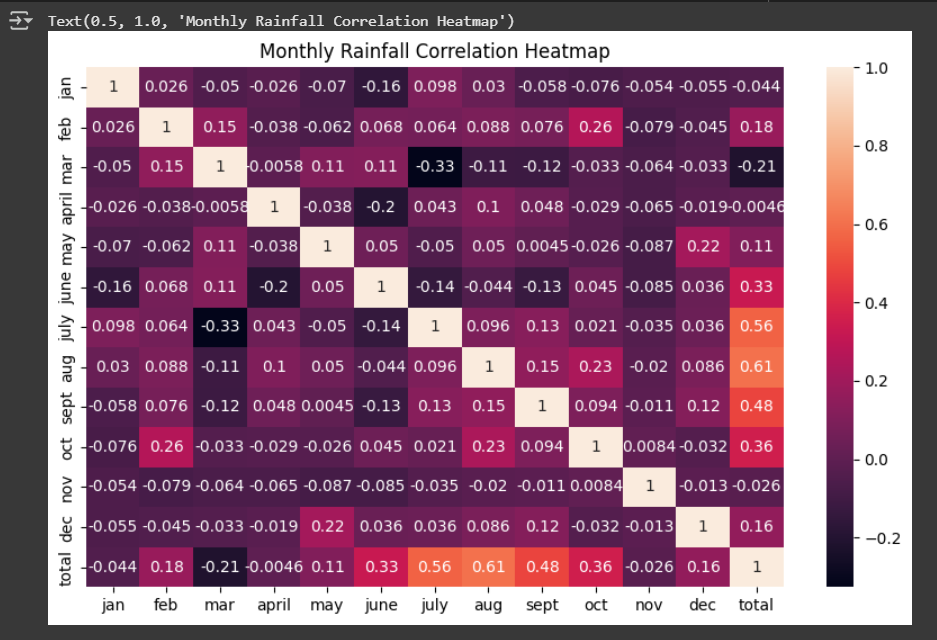
Upon examining the dataset, it was found that there are no duplicate values and no missing values present.

In year column shows 121 unique values from 1901 to 2021

Unique years:- [1901, 1902, 1903, 1904, 1905, 1906, 1907, 1908, 1909, 1910, 1911, 1912, 1913, 1914, 1915, 1916, 1917, 1918, 1919, 1920, 1921, 1922, 1923, 1924, 1925, 1926, 1927, 1928, 1929, 1930, 1931, 1932, 1933, 1934, 1935, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021]

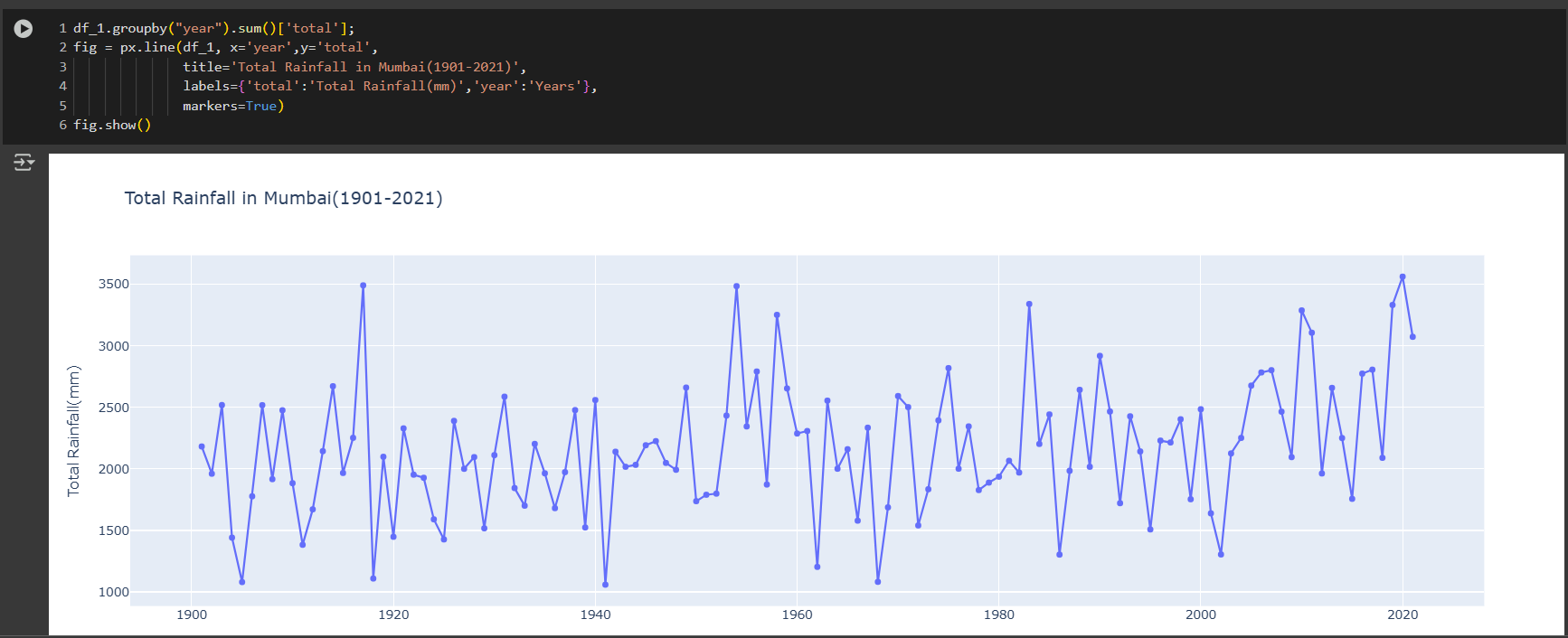
Review the dataset attributes and convert them to lowercase.

**Dataset Describe**



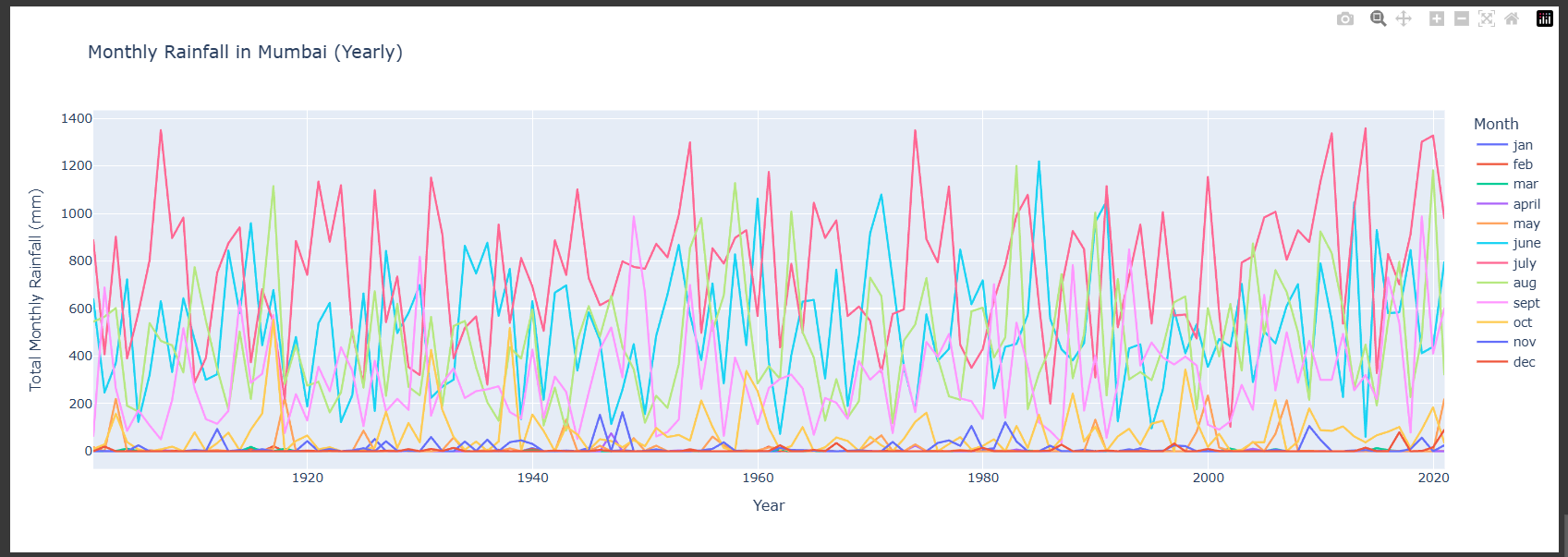
**Monthly Rainfall Correlation Heatmap**

* The correlation matrix of monthly rainfall shows that the monsoon months **(July, August, September)** are strongly correlated with each other and with total annual rainfall, indicating their significant impact on overall rainfall patterns.
* Non-monsoon months have weaker or negative correlations, highlighting their limited contribution to total rainfall and lack of dependence on each other.

**Total Rainfall In Mumbai(1901-2021)**

* The most prominent feature is the high degree of fluctuation in rainfall. Mumbai's rainfall is not consistent year after year. This suggests that relying on average rainfall figures might not be sufficient for planning and water management.
* short-term ups and downs trend, there doesn't appear to be a clear, long-term upward or downward trend in rainfall. This implies that there's no strong evidence from this graph alone to suggest a significant increase or decrease in overall rainfall in Mumbai over the past 121 years.
* A high degree of fluctuation in rainfall leads to **cyclical patterns**. Cycles are usually longer than seasonal and difficult to predict because they are irregular. Cyclicity can be caused by economic, political, or social factors that influence the data.

**Monthly Rainfall In Mumbai (Yearly)**

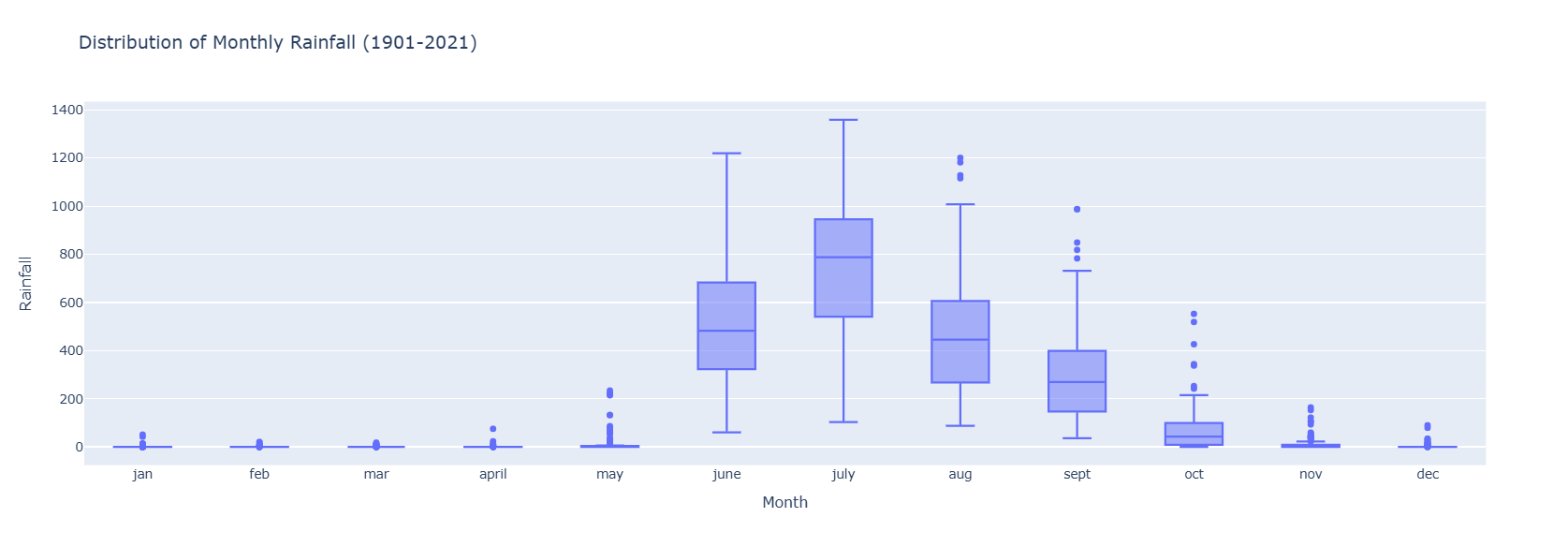


Visually, it's hard to discern a clear **upward or downward long-term trend** in the overall rainfall over the 100-year period (1920-2020). The rainfall seems to oscillate and vary naturally year by year without a definitive increasing or decreasing pattern across the decades.

* **Rainfall Intensity:** The chart shows that in peak monsoon months, the rainfall can exceed **1200 mm or even 1400 mm in a single month** in certain years. This highlights the potential for very intense rainfall events in Mumbai.
* **Low Rainfall Months:** Conversely, months like **January, February, March, April, May, October, November, and December** consistently show much lower rainfall, with their lines staying closer to the bottom of the chart. These represent the drier months in Mumbai.
* **High Rainfall Months:** By observing the peaks and the legend (even though colors might be a bit faded), it's highly likely that the months of **June, July, August, and September** display the highest rainfall. These months' lines will show prominent spikes throughout the years, indicating substantial rainfall during this period.

**Outliers Check & Treatment**

**Distribution Of Monthly Rainfall(1901 -2021)**

****

These box plots provide a clear visual comparison of rainfall patterns in Mumbai Monthly:

**January Rainfall:**

* **Very Low Median and IQR:** Both the median and the IQR are extremely low, showing that January is consistently dry in Mumbai.
* **Few Outliers:** There are very few outliers, and they are not as extreme as in December, indicating that significant rainfall in January is rare.

**February Rainfall:**

* **Low Median, Wider IQR than January, and Moderate Outliers:** The median is low, but the IQR is wider than in January, suggesting slightly more variability in February rainfall. There are also a few outliers, but they are generally less extreme than the highest outliers in December.
* **Slightly Less Skewed Than December:** The distribution is still right-skewed, but less so than December.

**March Rainfall:**

* **Very Low Median and IQR**: The box is extremely short and located very close to zero, showing that most March rainfall amounts are negligible.
* **Numerous Outliers**: Several outliers are present, indicating occasional years with slightly higher (but still relatively low) rainfall in March.
* **Overall**: March is consistently dry, with very little rainfall expected.

**April Rainfall:**

* **Low Median and IQR**: Similar to March, the box is short and near zero, indicating that April is also typically dry.
* **Fewer Outliers**: Compared to March, there are fewer outliers, but the presence of one high outlier suggests that while rare, April can experience significant rainfall events.
* **Overall:** April is predominantly dry, but slightly less so than March, with a very small chance of significant rainfall.

**May Rainfall:**

* **Slightly Higher Median and Wider IQR**: The box is somewhat taller and positioned slightly higher than the previous two months, indicating a potential increase in both the typical amount and variability of rainfall in May.
* **More Frequent Higher Rainfall**: The presence of more data points (including a few outliers) above the box suggests that May experiences more instances of higher rainfall compared to March and April.
* **Overall**: May marks a transition towards the wetter months, with pre-monsoon showers becoming more frequent and sometimes substantial.

**June Rainfall:**

* **Moderate Variability:** June rainfall shows a moderate degree of variability. While there's a cluster of points in the lower range, there are several instances of significantly higher rainfall.
* **No Clear Trend:** There's no obvious upward or downward trend in June rainfall over the years.
* **Typical Monsoon Onset:** June marks the typical onset of the monsoon in Mumbai, so this variability is expected.

**July Rainfall:**

* **Generally High and Consistent:** July rainfall is generally high and relatively consistent. Most data points are clustered in the upper range, indicating substantial rainfall in most Julys.
* **Less Variability Than June:** Compared to June, July rainfall shows less variability. There are fewer instances of exceptionally high or low rainfall.
* **Peak Monsoon Month:** July is typically the peak monsoon month in Mumbai, which is reflected in the consistently high rainfall amounts.

**August Rainfall:**

* **High and Relatively Consistent:** August, like July, shows high and relatively consistent rainfall. Most data points are clustered in the upper range.
* **Similar Pattern to July:** The rainfall pattern in August is quite similar to July, with consistently high amounts and less variability than June.

**September Rainfall:**

* **Decreasing Rainfall:** September shows a wider spread of data points compared to July and August, with a noticeable decrease in rainfall amounts in some years.
* **More Variability Than July/August:** September rainfall is more variable than July or August, indicating the start of the monsoon's withdrawal.
* **End of Monsoon:** September marks the end of the main monsoon season in Mumbai, which is reflected in the decreasing rainfall trend.

**October Rainfall:**

* High variability, moderately high typical rainfall, and frequent exceptionally high rainfall events (outliers). The rainfall distribution is skewed towards higher amounts. This confirms the high variability we saw in the line plots.
* The plot shows several outliers on the higher end, reinforcing the observation that October can have exceptionally high rainfall events. These outliers are far above the upper whisker, highlighting their unusualness.

**November Rainfall:**

* Low variability, low typical rainfall, with occasional (but rare) higher rainfall events. The rainfall distribution is also skewed towards higher amounts, but less so than October. This confirms the consistency and generally low rainfall in November.

**December Rainfall:**

* **Low Median, Wide IQR, and High Outliers:** The box is short and close to zero, indicating that the median and interquartile range (IQR) of December rainfall are low. However, there are several significant outliers, some reaching very high values. This confirms that while most Decembers are dry, there's a chance of substantial rainfall.
* **Skewness:** The distribution is heavily right-skewed, meaning there's a long tail of higher rainfall amounts.

Before processing with the next steps, treat the any outliers.

**1.Removing Outliers Using IQR (Interquartile Range) Method:**

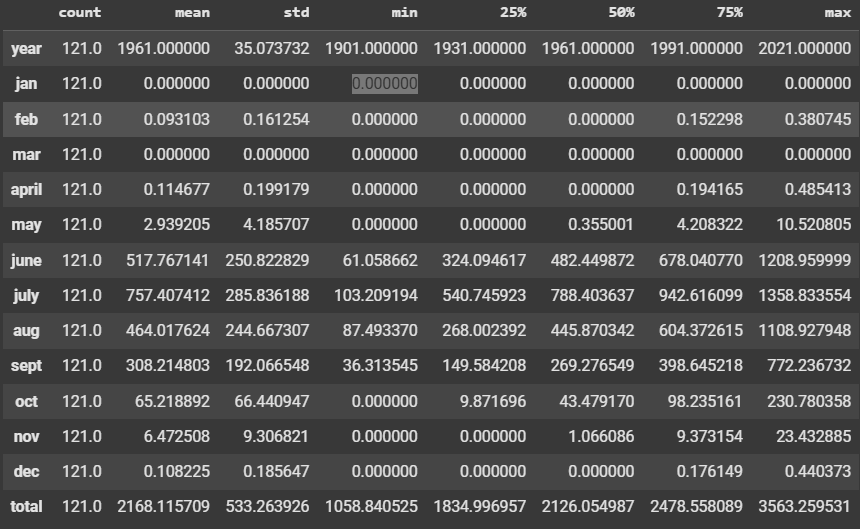
* Outliers are typically defined as values falling outside 1.5 times the IQR above Q3 or below Q1.
* In this case, for each month, rainfall data points that fall below the lower bound or above the upper bound are considered outliers and removed.
* The new boxplot will have no outliers, and the range of whiskers (the lines extending from the box) will be shorter, reflecting the more typical rainfall patterns.
* The data will look more condensed, and extremely high or low rainfall values will no longer be visible.
* Interpretation is Removing outliers and simplifies the data by focusing on the core range of rainfall for each month, making it easier to spot patterns in typical monthly rainfall.

**2. Capping Outliers:-**

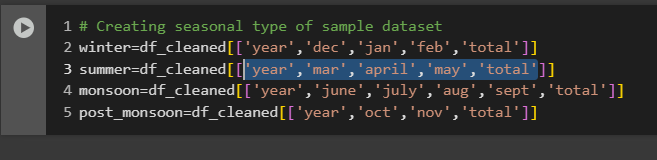
* Instead of removing outliers, this method caps them. Values beyond the upper or lower bounds (as determined by the IQR) are set to the respective bounds. This preserves all data points but restricts extreme values.
* The whiskers will extend only up to the capped values, giving a smoother representation of the data.
* Interpretation is Capping reduces the impact of extreme values while keeping the data intact. The boxplot will still display variation, but the risk of highly skewed results due to outliers is reduced.

**Summary of Handling Outliers :-**

* **Removing outliers using IQR**: This method strictly eliminates outliers, focusing on typical rainfall distributions, but at the cost of potentially discarding important extreme values.
* **Capping outliers**: This retains all data but limits the effect of extreme values, ensuring that outliers don’t skew the results too much while preserving the dataset's completeness.

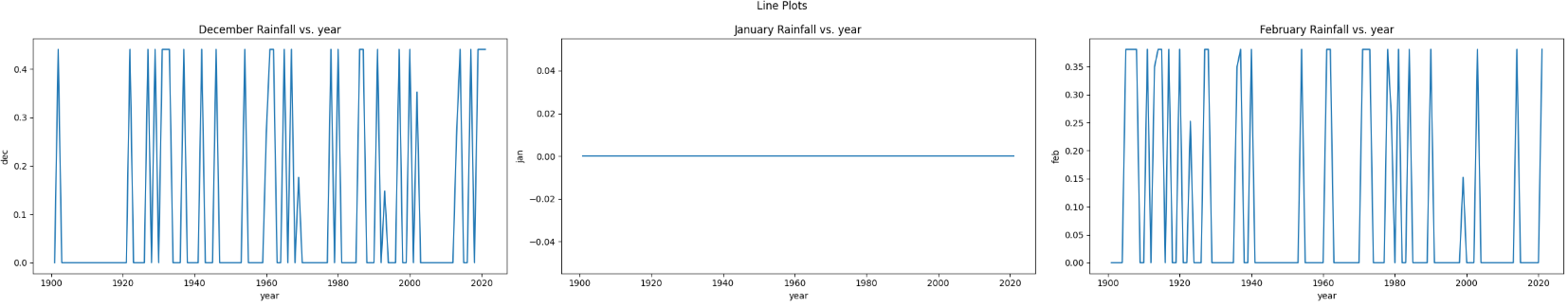


After this process create a split dataset into seasonal parts.

* A DataFrame called **post\_monsoon** is created for the post-monsoon season (Oct, Nov.
* A DataFrame called **monsoon** is created for the monsoon season (June, July, Aug, and Sept).
* A DataFrame called **summer** is created for the monsoon season (Mar, April, May).
* A DataFrame called **winter** is created for the monsoon season (Dec, Jan, Feb).

Now, let's examine the insights based on the seasons moving forward.

**WINTER:**

****In the winter session, the months are December, January and, February. We gain insight from the 1901 to 2021 data we have, analyzing the amount of rainfall during the winter session over the last 121 years.

**December Rainfall:**

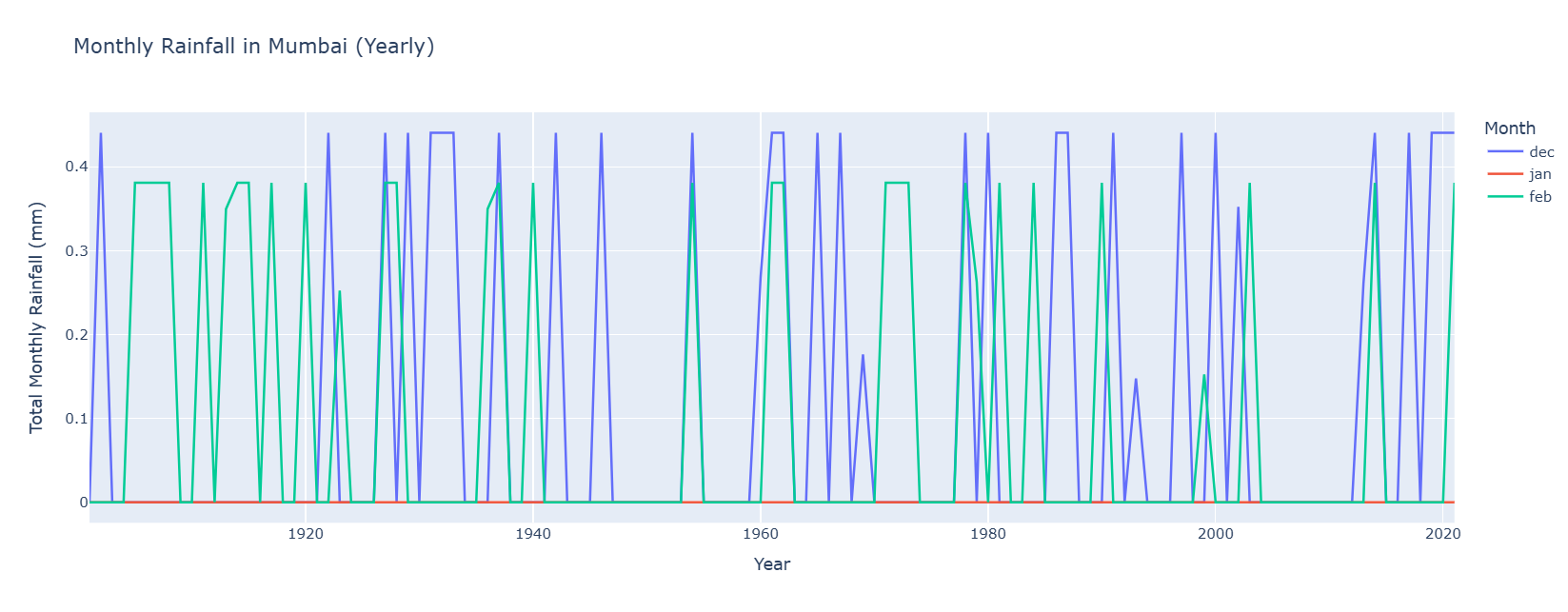
* **Generally Low with Intermittent Spikes:** December rainfall is typically low, with the line hovering near zero for many years. However, there are intermittent spikes, some quite significant, indicating occasional years with unusually high rainfall for December.
* **Potential for Unseasonal Showers:** These spikes suggest that while generally dry, December can experience unseasonal rainfall events, possibly related to passing weather systems or cyclones.

**January Rainfall:**

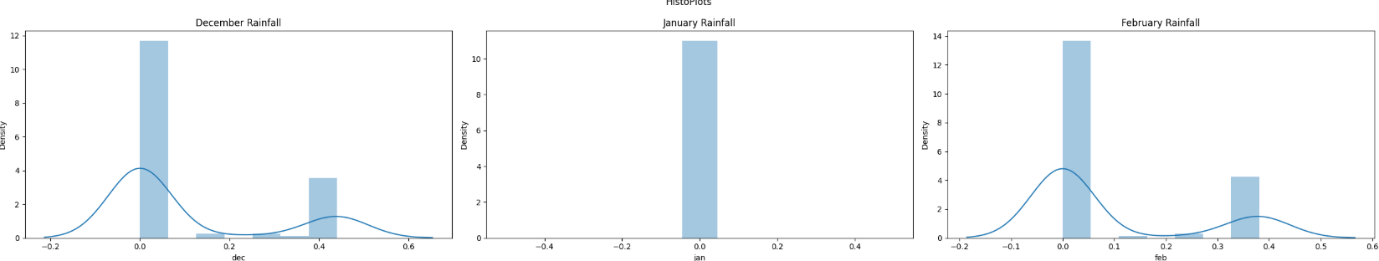
* **Low and Relatively Consistent:** January rainfall is generally low and more consistent compared to December. The line mostly stays close to zero, with fewer and less dramatic spikes.
* **Drier Than December:** Overall, January appears to be a drier month than December in Mumbai, with less variability and fewer instances of significant rainfall.

**February Rainfall:**

* **Low with Occasional Higher Rainfall:** February also generally experiences low rainfall, but with occasional years of more substantial amounts.
* **More Variability Than January, Less Than December:** February shows slightly more variability in rainfall compared to January but less than December. There are some noticeable peaks, though not as frequent or as high as in December.

**Monthly Rainfall In Mumbai(Yearly)**

These line plots reveal the following about the winter rainfall patterns in Mumbai:

* **December:** Generally dry but prone to occasional, sometimes significant, unseasonal rainfall events.
* **January:** The driest of the three months, with consistently low rainfall and few instances of higher amounts.
* **February:** Also mostly dry, but with slightly more variability and occasional higher rainfall compared to January, though less so than December.

These histograms confirm the trends observed in the line plots:

* **December:** Exhibits high variability with a bimodal distribution, suggesting two distinct patterns: a large number of dry Decembers and a significant number of moderately wet Decembers, along with occasional very wet Decembers.

The distribution is positively skewed (right-skewed), with a tail extending towards higher rainfall amounts. This indicates that while less frequent, Mumbai does experience years with very high December rainfall.

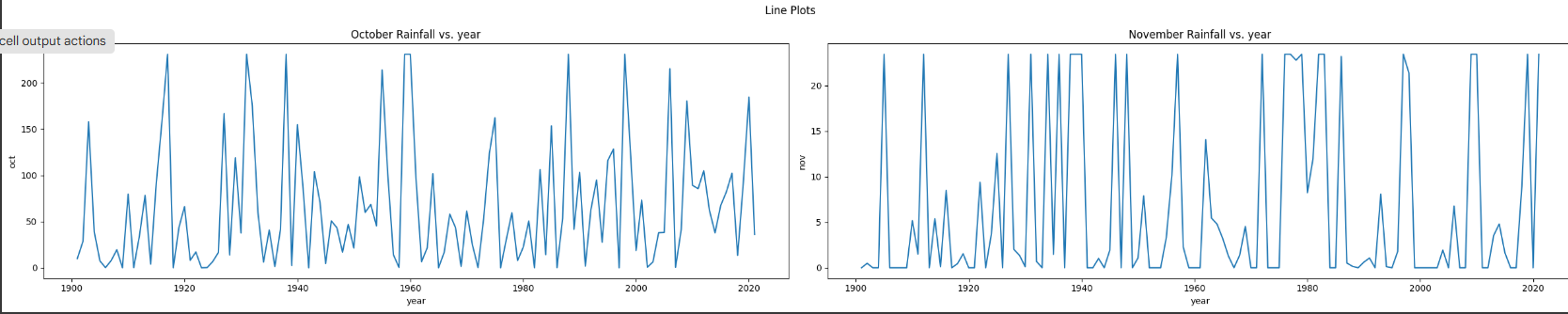
* **January:** Is predominantly dry, with most years having negligible rainfall.
* **February:**

Shows a pattern between January and December, with more instances of light rain compared to January but less intense or frequent rainfall compared to December.

The distribution is positively skewed, indicating that occasional higher rainfall events are possible, though less frequent and less intense than in December.

**POST MONSOON:**

In the post monsoon session, the months are October, November. We gain insight from the 1901 to 2021 data we have, analyzing the amount of rainfall during the winter session over the last 121 years.

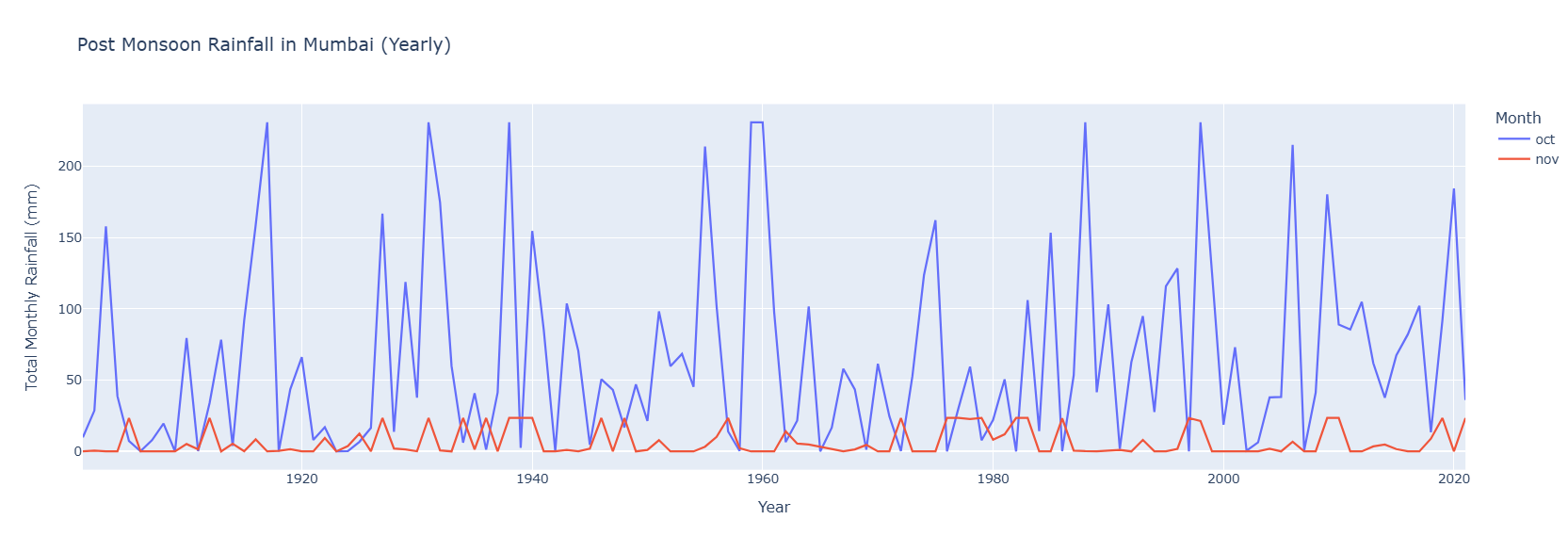
****

**October Rainfall Vs Years:**

* **Highly Variable:** October rainfall in Mumbai is highly variable, showing large fluctuations from year to year.
* **Significant Rainfall Events:** The plot reveals several instances of very high rainfall in October, indicating that Mumbai can experience significant rainfall events well into the post-monsoon season.
* **No Clear Trend:** While there are fluctuations, there's no discernible upward or downward trend in October rainfall over the 120 years.

**November Rainfall Vs Years:**

* **Generally Low:** November rainfall in Mumbai is generally low, with the line mostly hovering near zero.
* **Occasional Spikes:** There are a few instances of higher rainfall in November, but these are less frequent and less intense than the rainfall events observed in October.
* **No Clear Trend:** Similar to October, there's no clear long-term trend in November rainfall.

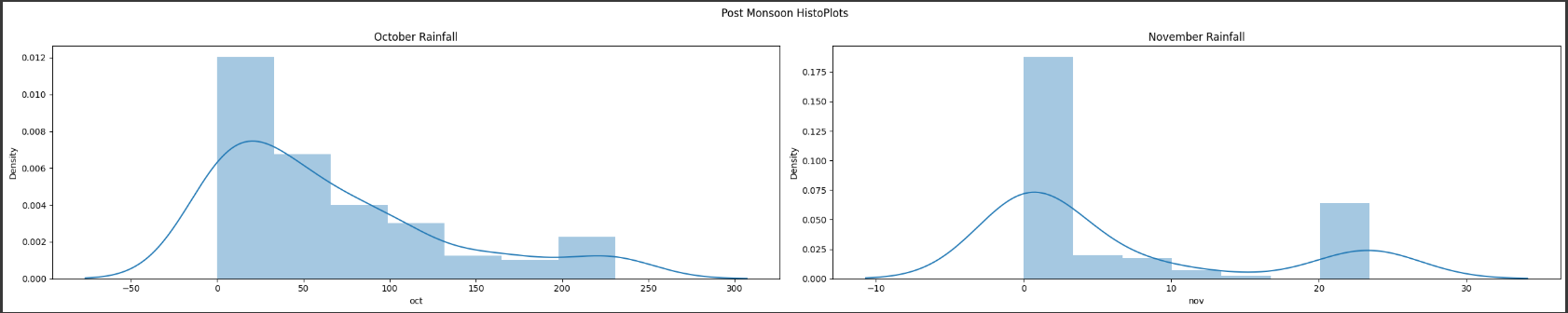


This chart displays the post-monsoon rainfall in Mumbai from 1900 to 2020, specifically for the months of October and November

* **November Rainfall**: November rainfall remains consistently low, generally below 25 mm, suggesting a drier period compared to October.

November generally experiences low rainfall, indicating a transition to a drier period. This information is crucial for agriculture and water resource planning, as it suggests a potential need for irrigation and water conservation.

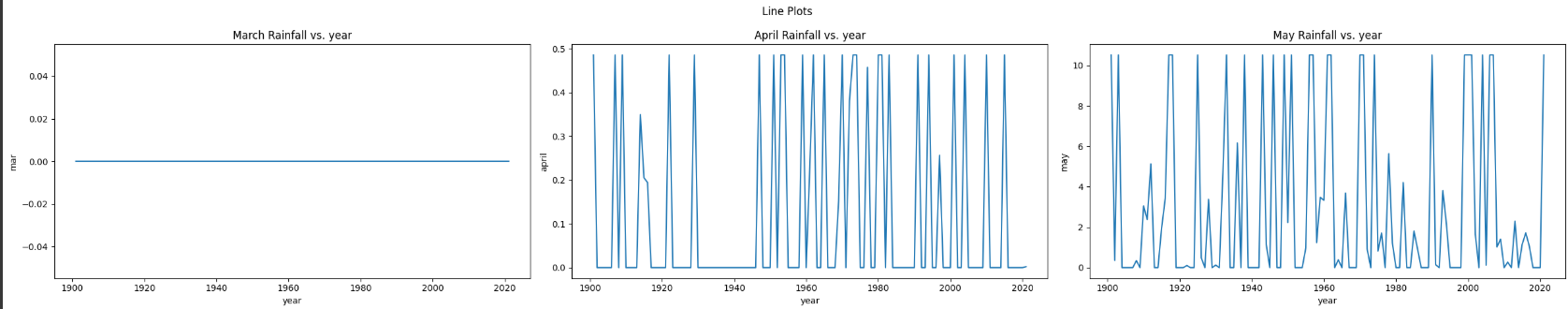
* **October Rainfall**: The blue line, representing October rainfall, shows significantly higher levels of precipitation compared to November (red line) throughout the entire period.



These histograms confirm the trends observed in the line plots:

* **October:** Exhibits high variability with a bimodal distribution, suggesting two distinct patterns: a large number of dry Octobers and a significant number of moderately wet Octobers, along with occasional very wet Octobers.
* **November:** Is predominantly dry, with most years having negligible rainfall. However, it also shows a strong positive skew, indicating that it's capable of experiencing occasional, unusually high rainfall events, although these are much less frequent and intense than in October.

**SUMMER:**

**** In the summer session, the months are March, April, May. We gain insight from the 1901 to 2021 data we have, analyzing the amount of rainfall during the winter session over the last 121 years.

**March Rainfall Vs Years:**

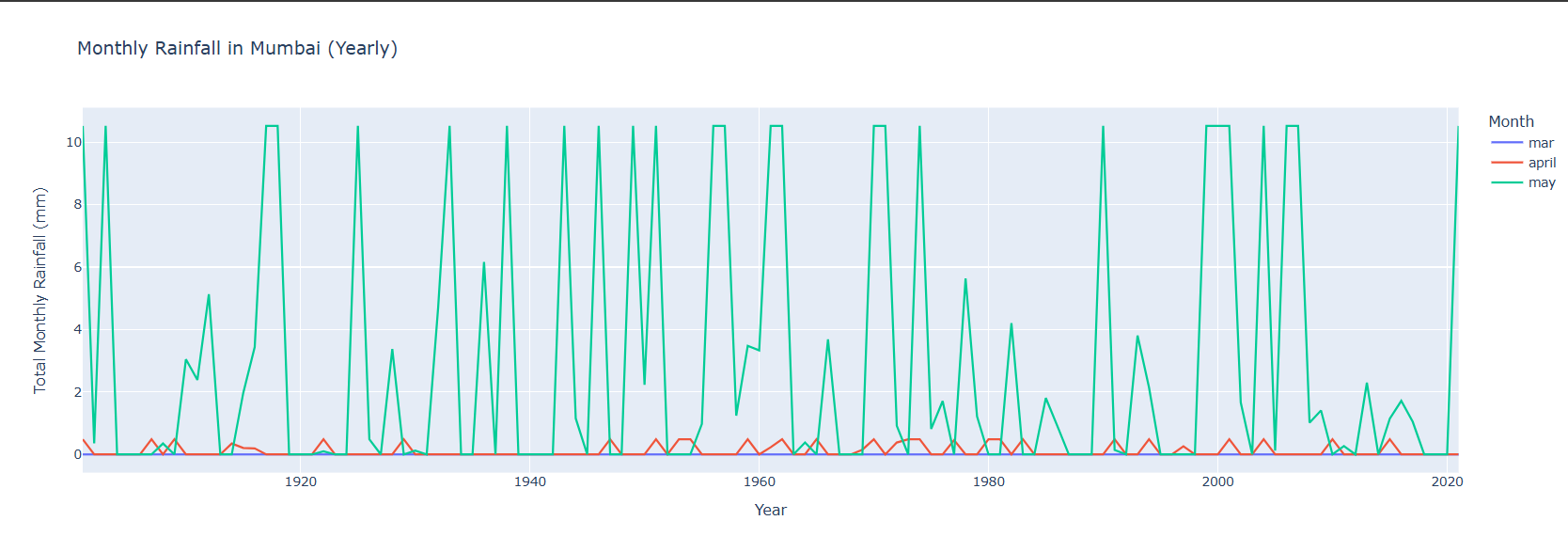
* **Consistently Low:** The line is almost entirely flat at zero, indicating very little to no rainfall in March across all years.
* **Extremely Dry:** This confirms that March is a consistently dry month in Mumbai.

**April Rainfall Vs Years:**

* **Mostly Low with Occasional Spikes:** April rainfall is also generally low, but there are a few instances of higher rainfall, represented by the spikes in the line.
* **Unseasonal Showers:** These spikes suggest that April can experience occasional unseasonal showers or storms, although the amounts are still relatively modest.

**May Rainfall Vs Years:**

* **More Variable and Higher Amounts:** May shows the most significant variability in rainfall among the three months. There are more frequent and higher rainfall amounts compared to March and April.
* **Transition Month:** This indicates that May is a transition month, experiencing pre-monsoon showers and thunderstorms that contribute to more substantial rainfall.



This chart displays the post-monsoon rainfall in Mumbai from 1900 to 2020, specifically for October and November

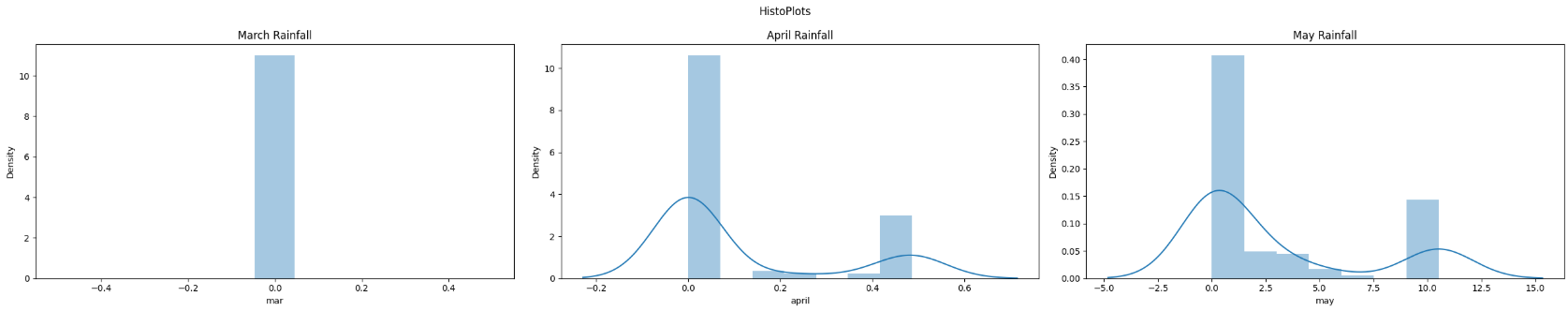
**March :**

* The purple line for March is almost entirely flat at zero for the entire period. This indicates that March is an extremely dry month in Mumbai with virtually no rainfall across the years.

**April :**

* The red line for April mostly hovers near zero, indicating low rainfall in most years. However, there are a few instances where the line shows spikes, indicating occasional years with slightly higher rainfall in April.
* These spikes suggest that April can experience occasional unseasonal showers or storms, although the amounts are still relatively modest.

**May :**

* The green line for May shows the most significant variability in rainfall among the three months. There are more frequent and higher rainfall amounts compared to March and April.
* This indicates that May is a transition month, experiencing pre-monsoon showers and thunderstorms that contribute to more substantial rainfall.
* While not definitive, there might be a slight increasing trend in May rainfall in recent years, with more frequent and slightly higher peaks. However, further statistical analysis would be needed to confirm this.

**March Rainfall:**

* **Almost Exclusively Zero:** The vast majority of years have no rainfall in March, represented by the tall bar at the 0 mm bin.
* **Rare Trace Amounts:** There might be a few years with trace amounts of rainfall, but they are insignificant compared to the dominant zero rainfall.

**April Rainfall:**

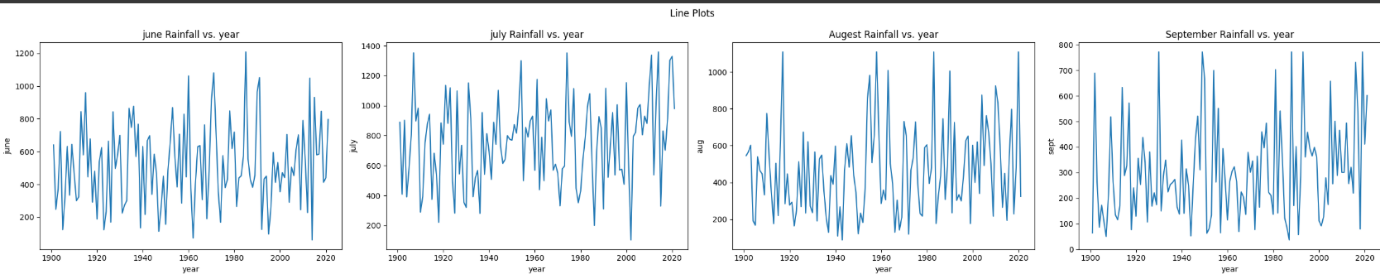
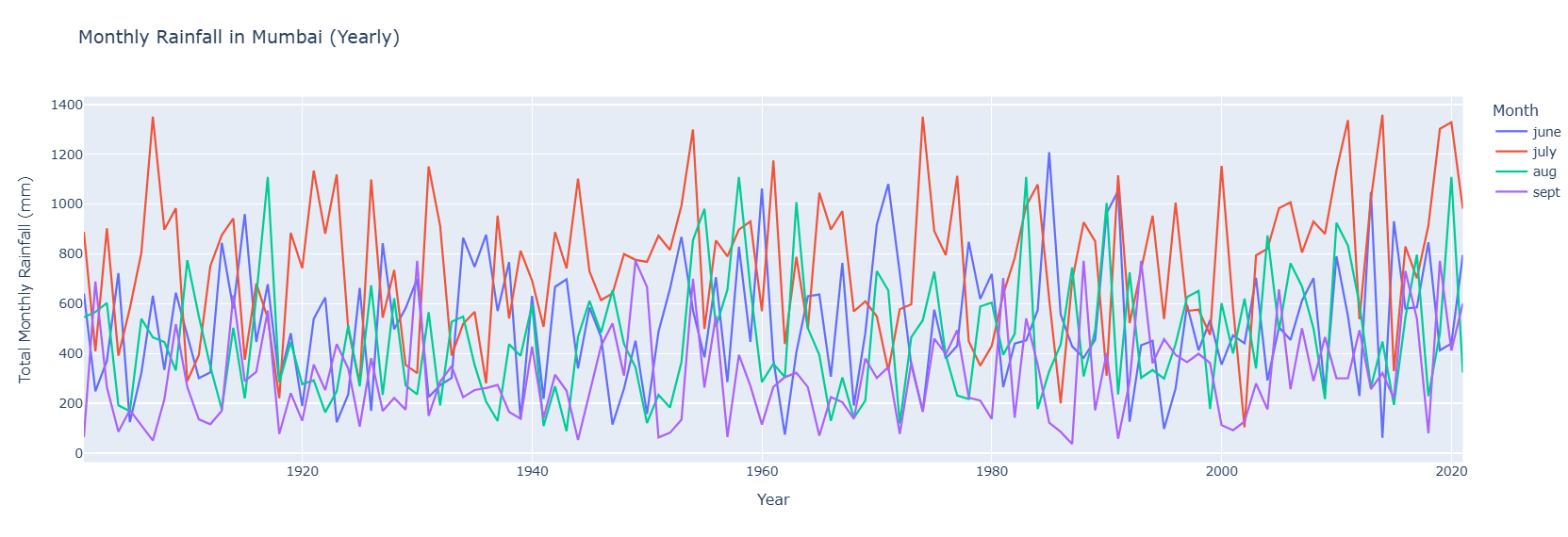
* **Mostly Zero with a Few Occasional Light Showers:** Most years have no rainfall, but there's a small bar indicating a few years with light showers (around 5-10 mm).
* **Possibly One Event with Slightly Higher Rainfall:** One year might have received slightly higher rainfall (around 20 mm), but it's still a relatively small amount.

**May Rainfall:**

* **More Variable and Higher Amounts:** May shows more variability, with a wider spread of rainfall amounts. There's still a significant portion of years with low rainfall, but there are more instances of moderate rainfall (around 10-20 mm) and even a few years with higher amounts (around 30-40 mm).
* **Possible Bimodal Distribution:** The histogram hints at a possible bimodal distribution, with one peak near zero and another in the moderate rainfall range. However, more data might be needed to confirm this.

**MONSOON:**

In the summer session, the months are March, April, May. We gain insight from the 1901 to 2021 data we have, analyzing the amount of rainfall during the winter session over the last 121 years.

****

**June Rainfall Vs Years :**

* June rainfall shows a moderate degree of variability. There are fluctuations throughout the period, with some years experiencing significantly higher rainfall than others.
* The plot reveals instances of June rainfall exceeding 1000 mm, indicating the possibility of heavy rainfall events or extended periods of rain.
* There's no clear upward or downward trend in June rainfall over the 120 years.

**July Rainfall Vs Years:**

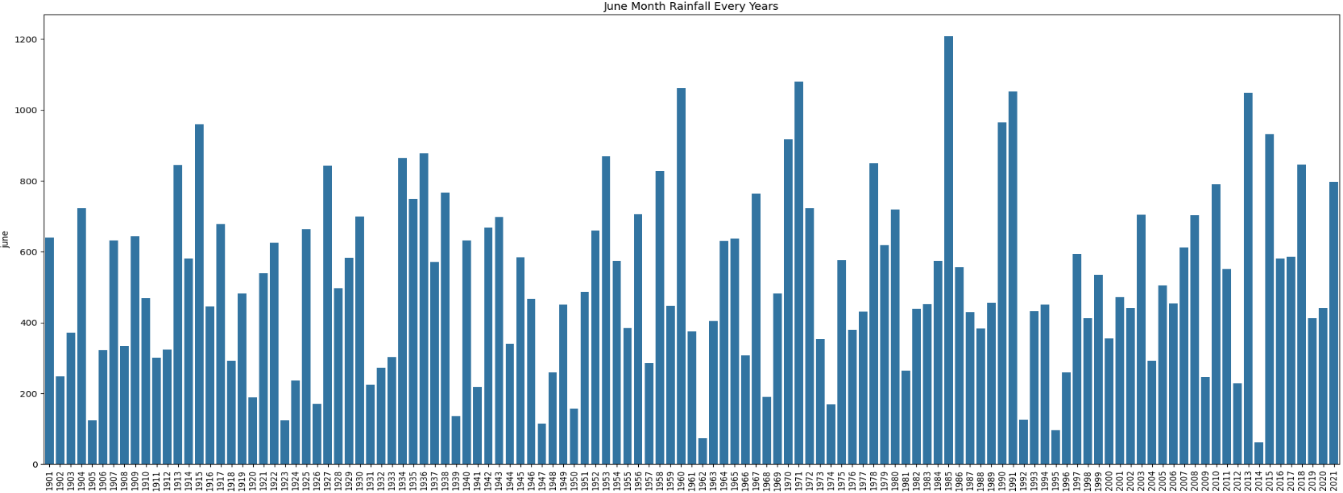
* July rainfall is generally high and relatively consistent, mostly staying above 600 mm. This indicates that July is a core monsoon month, delivering substantial rainfall in most years.
* Compared to June, July rainfall shows less fluctuation, suggesting a more stable pattern of heavy rainfall.
* There might be a slight upward trend in recent years, but it's not definitive and would require further statistical testing.

**August Rainfall Vs Years:**

* August, like July, shows high and relatively consistent rainfall, mostly staying above 600 mm.
* The rainfall pattern in August is quite similar to July, with consistently high amounts and less variability than in June.

**September Rainfall Vs Years:**

* September rainfall shows a wider spread of values compared to July and August, with a noticeable decrease in rainfall amounts in some years. This indicates the start of the monsoon's withdrawal.
* September rainfall is more variable than July or August, suggesting a less predictable pattern as the monsoon weakens.

**June Month Rainfall Every Year**

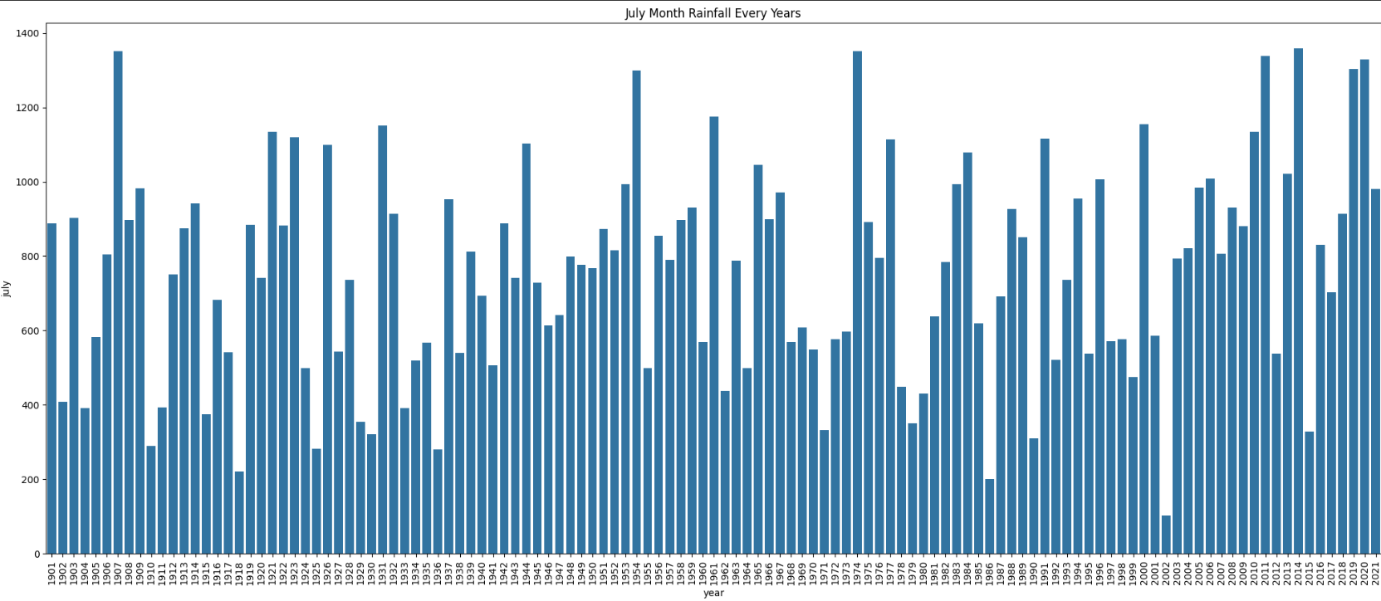
**High Variability** – Rainfall in June varies significantly across the years, with some years having extreme highs and others having very low rainfall.

**Extreme Rainfall Events** – Some years (e.g., around 1915, 1961, 1988, and 2013) show exceptionally high rainfall, possibly indicating monsoon surges or extreme weather conditions.

**Dry Years** – There are years with notably low rainfall, which could suggest drought-like conditions.

**Recent Trends** – The recent years (post-2000) seem to have a mix of moderate and high rainfall, indicating possible climate fluctuations or changing monsoon patterns.

**Long-Term Patterns** – The pattern suggests cycles of high and low rainfall, which might be influenced by larger climatic phenomena like El Niño or La Niña.

 **July Month Rainfall Every Year**

**High Variability:**

* The rainfall levels show significant fluctuations between years, indicating inconsistent monsoon patterns.

**Extreme Rainfall Events:**

* Some years have **exceptionally high rainfall** (e.g., around **1907, 1975, 2005, and 2019**).
* These peaks might correspond to **strong monsoons or extreme weather events**.

**Low Rainfall Years:**

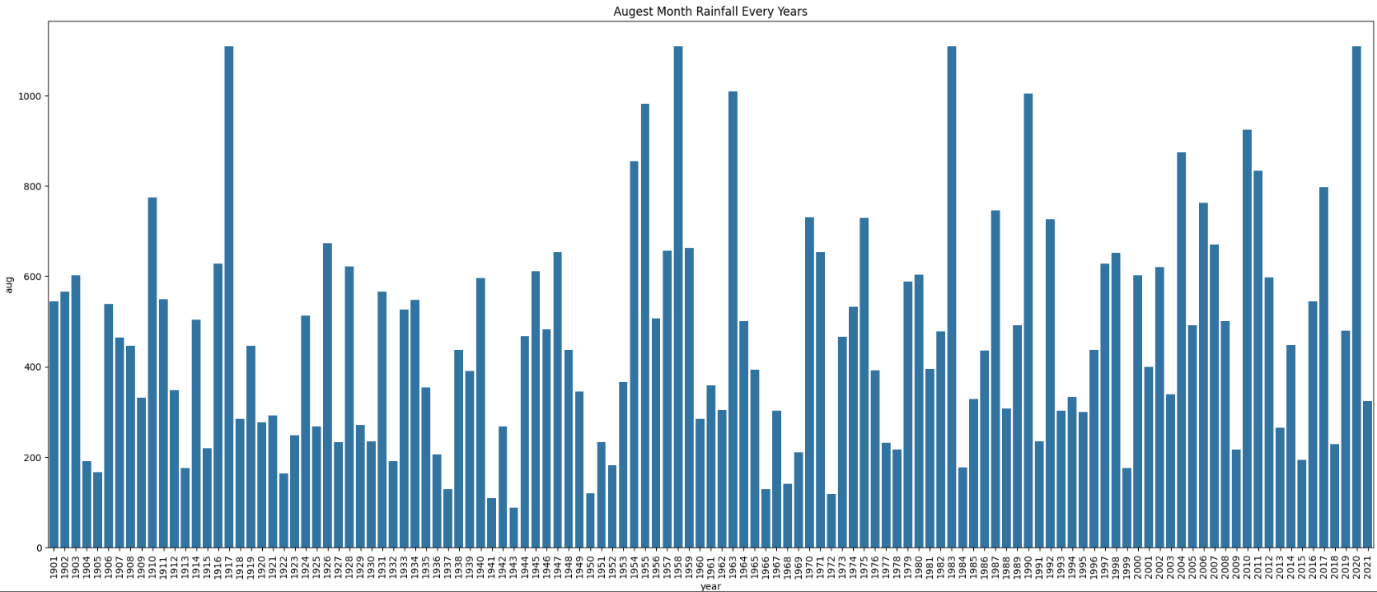
* There are years with **notably lower rainfall** (e.g., **1918, 1972, 2002, 2015**), which could indicate **drought or weak monsoon seasons**.

**Recent Trends:**

* Rainfall in the **2000s and 2010s** has had both **high and low variations**, indicating that climate variability continues.
* Some of the highest rainfall years are seen **after 2000**, possibly linked to **changing climate patterns**.

**Long-Term Patterns:**

* The data suggests possible **cyclical variations** with **periodic high and low rainfall** events. This could be influenced by larger climatic factors such as **El Niño, La Niña, or global warming effects**.

**August Month Rainfall Every Year**

**Higher Rainfall Compared to September:**

* The **August rainfall appears to be more intense** compared to September
* Many years have **rainfall exceeding 600 mm**, with some extreme years crossing **1000 mm**.

**Extreme Peaks in Certain Years:**

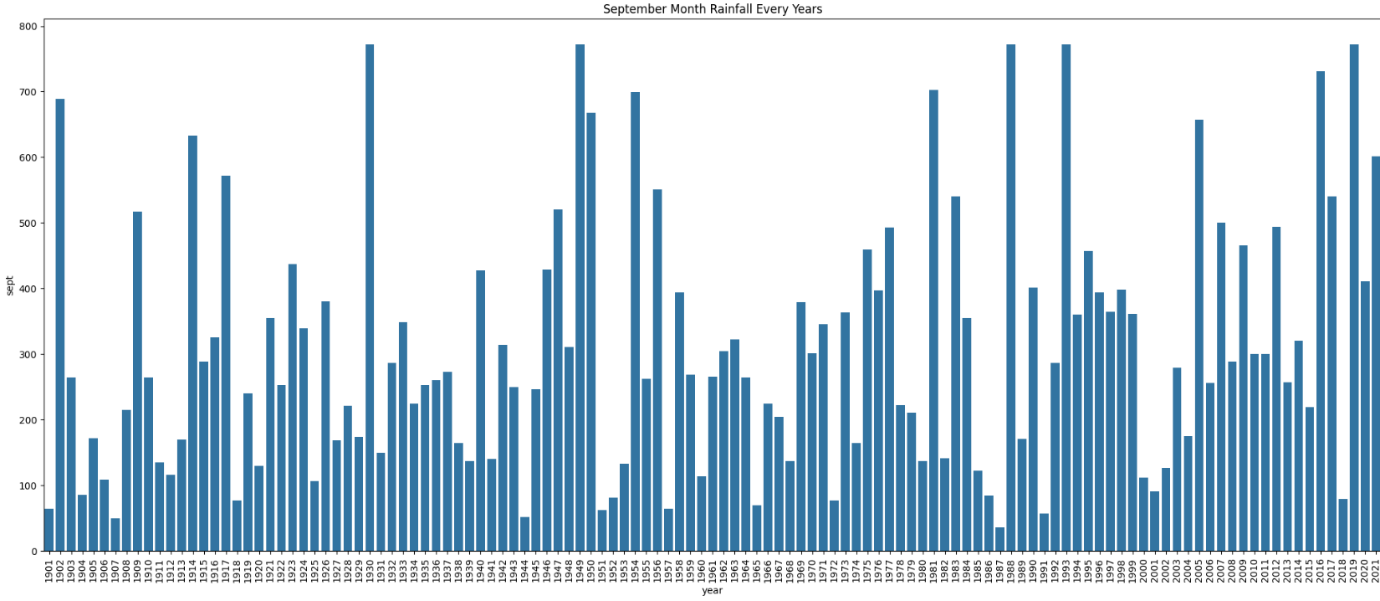
* **1917, 1958, 1976, 1988, and 2018** show significantly **higher rainfall**, suggesting stronger monsoons or unusual weather patterns.
* These peaks may indicate **flood-prone years** due to excessive rainfall

**Frequent Heavy Rainfall Years:**

* Unlike September, where rainfall fluctuates more unpredictably, **August consistently shows years of high rainfall**.
* This suggests that **August is a crucial monsoon month**, contributing heavily to annual rainfall.

**Recent Trends:**

* Post-2000, **many years show consistently high rainfall**, indicating a possible **increase in August monsoon intensity** due to climate change.
* The number of **moderate to high rainfall occurrences has increased**.

**September Month Rainfall Every Year**

**Extreme Peaks in Certain Years:**

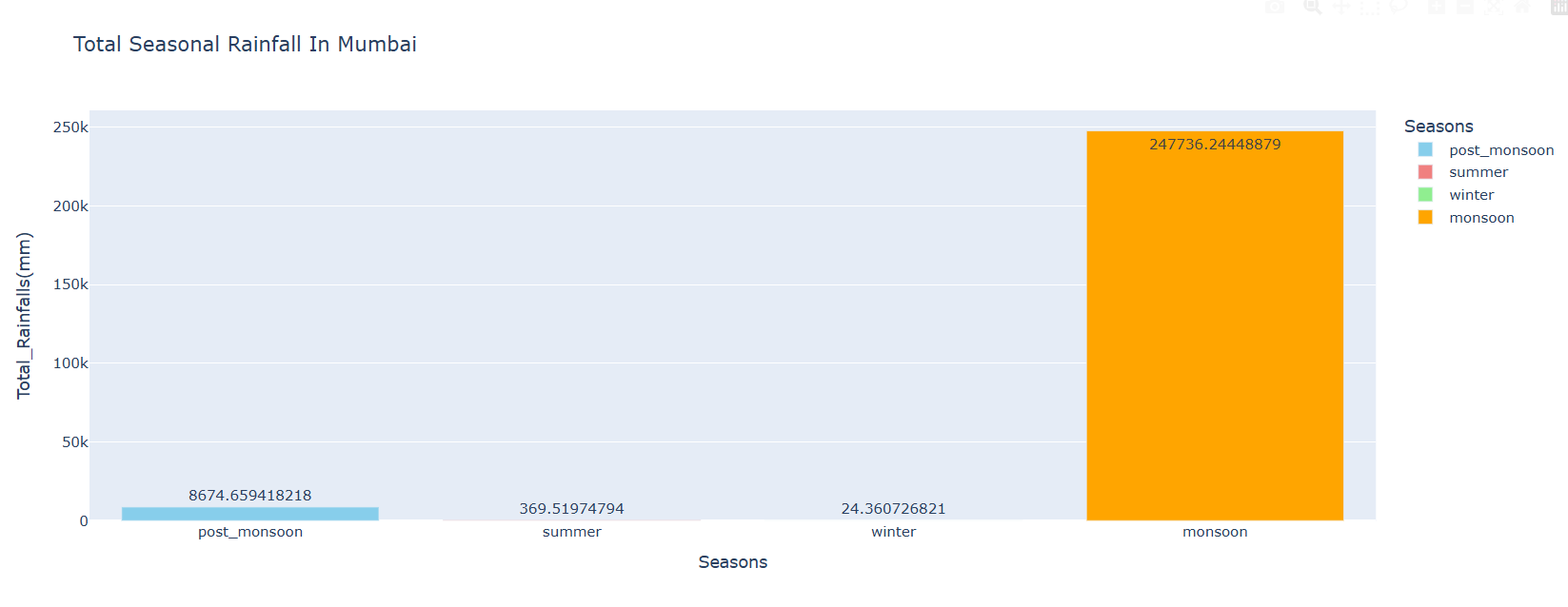
* Some years, such as **1923, 1954, 1988, and 2019**, show **significantly higher rainfall** than others. These extreme values suggest that in certain years, monsoons extended longer or were more intense.

**High Variability:**

* The rainfall in September varies significantly across different years, with some years experiencing extreme peaks while others have lower values.

**Long-Term Patterns:**

* The rainfall distribution does not show a clear **increasing or decreasing trend**, suggesting that **September rainfall remains highly unpredictable**.
* However, there are periods where rainfall appears to be consistently higher (e.g., **1970s and 1980s**) compared to others.
* In the most recent years, rainfall seems to be more **frequent and intense**, possibly due to climate change effects.
* Some years, such as **post-2000**, indicate more frequent moderate to high rainfall occurrences.

**Total Seasonal Rainfall In Mumbai**

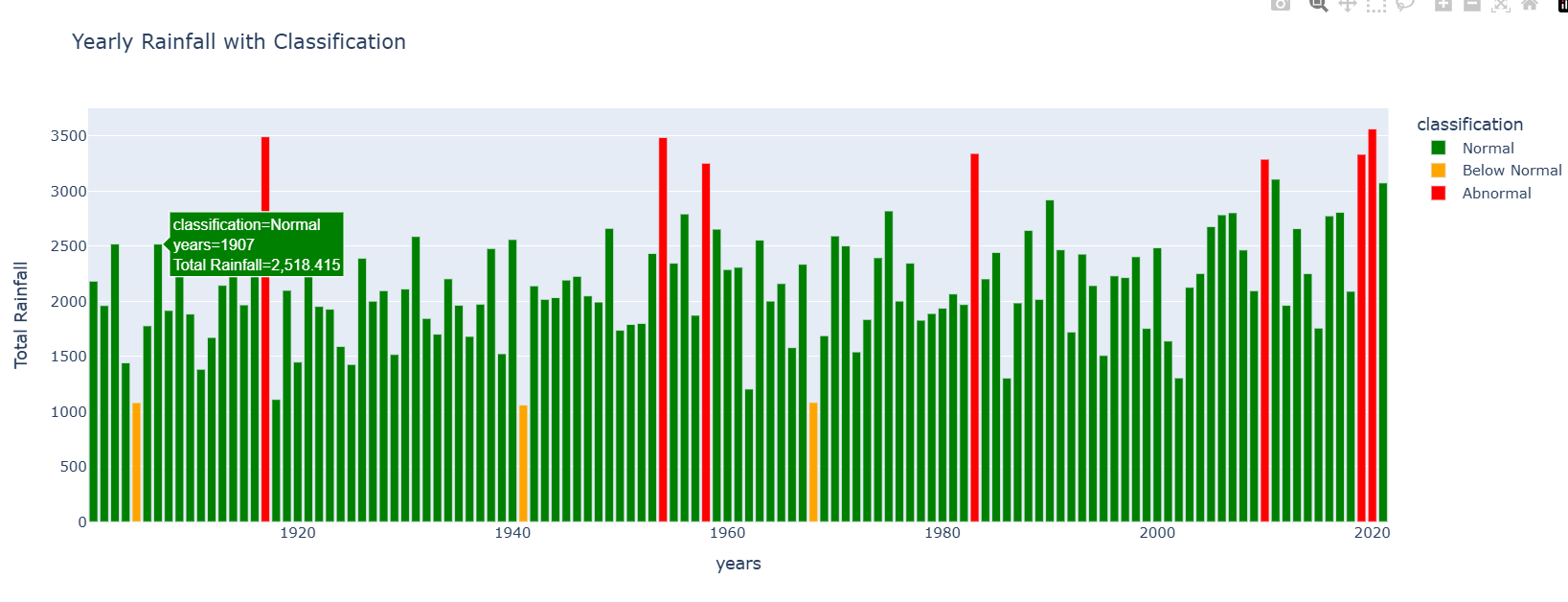
 **Monsoon Dominance:** The monsoon season (orange bar) shows significantly higher rainfall (247,736 mm) compared to all other seasons. This is expected, as the monsoon season is the primary rainy season in Mumbai.

 **Post-Monsoon Moderate Rainfall:** The post-monsoon season (blue bar) exhibits a moderate amount of rainfall (8,674 mm), considerably less than the monsoon but still substantial compared to summer and winter.

 **Summer and Winter Scarcity:** Both the summer (red bar) and winter (green bar) seasons show very low rainfall amounts (369 mm and 24 mm, respectively), indicating dry periods.

**Yearly Rainfall With Classification**

Let's turn raw yearly rainfall data into insightful categories! By analyzing the totals, we can classify them into engaging categories like "**Normal**," "**Abnormal**," and “**Below Normal**”. This way, we can better understand the trends and patterns in our rainfall data, making it easier to discuss and interpret!

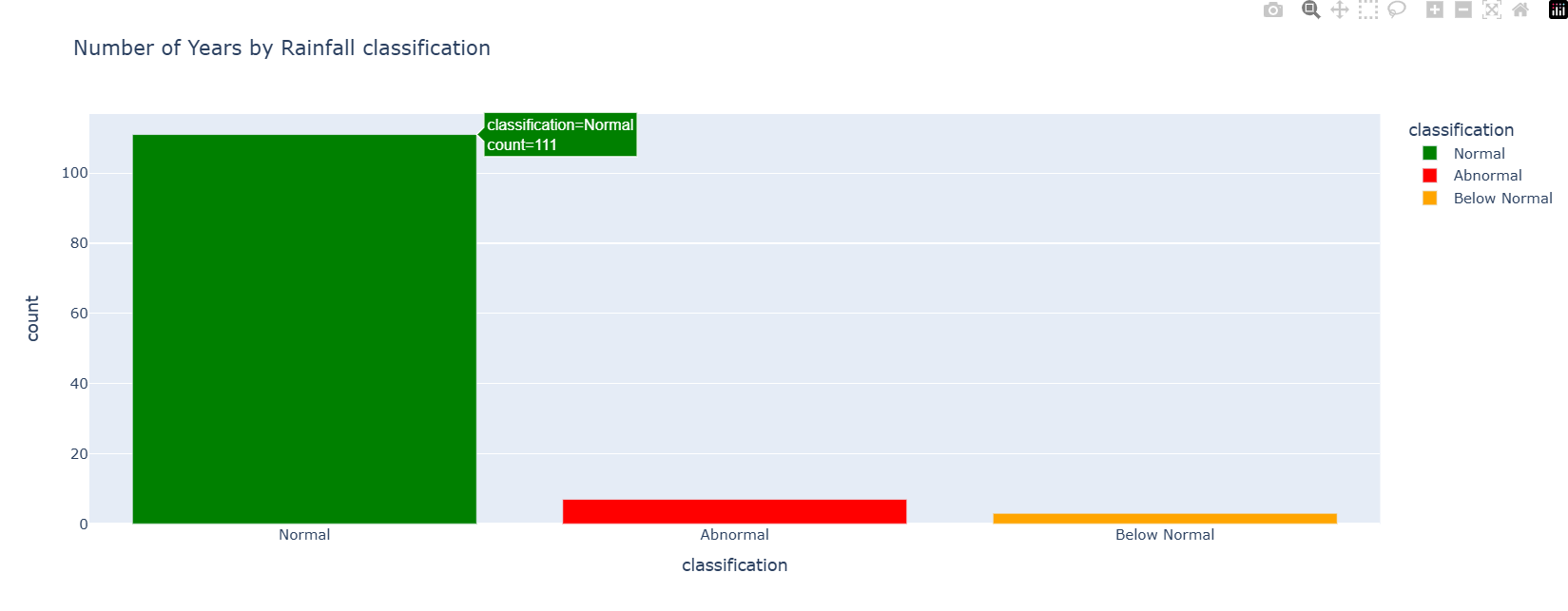


 **Visual Overview of Rainfall Variability:** The chart provides a quick visual assessment of how rainfall has varied across the years. You can see periods with consistently high rainfall, periods with fluctuations, and periods with generally lower rainfall.

 **Frequency of Classifications:** By observing the colors, you can get a sense of how frequently each classification (Normal, Below Normal, Abnormal) has occurred over the years.

 **Identification of Extreme Years:** The red bars immediately draw attention to years with abnormally high or low rainfall.

 **Specific Year Information:** The annotation box shows how specific year information can be highlighted.



After conducting a thorough classification of the data, we reviewed the counts for each category: Normal, Abnormal, and Below Normal.

1. **Normal**: This category is visually represented by a green bar, indicating a strong status. It boasts a count of 111, signifying a substantial majority of the total data assessed.

2. **Abnormal**: The Abnormal category is denoted by a striking red bar, indicating potential concerns or deviations from the norm. It has a count of 7, suggesting that there are a small number of entries flagged for further investigation.

3. **Below Normal**: Lastly, entries that fall into the Below Normal category are represented by an orange bar. This category has a count of 3, reflecting a limited number of instances that require attention or intervention.

This classification highlights the distribution of data across these categories and underscores where particular focus may be needed.

The Indian Meteorological Department (IMD) uses a colour-coded warning system to indicate the severity of weather conditions. Here are the details of the scale with corresponding numbers for rainfall:

**Green Alert**: No warning. Normal conditions.

**Yellow Alert:** Be aware. Heavy rainfall ranging from 64.5 mm to 115.5 mm in 24 hours.

**Orange Alert:** Be prepared. Very heavy rainfall ranging from 115.6 mm to 204.4 mm in 24 hours.

**Red Alert:** Take action. Extremely heavy rainfall exceeding 204.5 mm in 24 hours.

**Feature Engineering**

This code maps categorical text values in two different columns—**"classification"** and **"category"**—to numerical values using Python's **map()** function in Pandas.

**1. Classification Column**

* **Mapping:**
  + 'Below Normal' → **0**, 'Normal' → **1**, Abnormal' → **2**
* The mapping is applied to the "**classification**" column in **df\_cleaned**.

**2. Category Column**

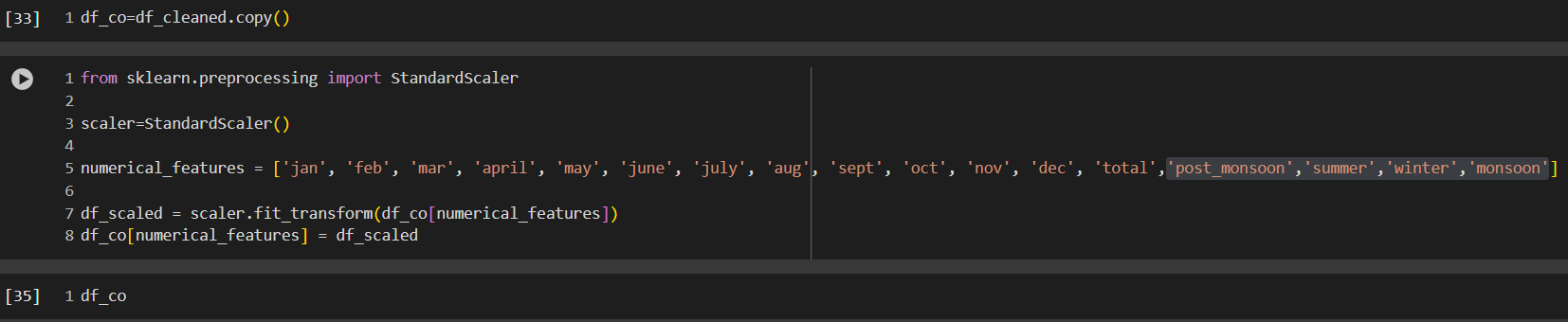
* **Mapping:**
  + 'Medium' → **0**, 'High' → **1**, 'Very High' → **2**, 'Extreme' → **3**
* The mapping is applied to the "**category**" column in **df\_cleaned**.

**Purpose & Usage**

* Converts categorical text values into **numerical labels** for **better analysis, visualization, and machine learning compatibility**.
* The **.map()** function ensures that each categorical value is correctly replaced with its respective numerical value in the dataset.

**Standard Scaler**

Here we use the Standard Scaling technique (Z-score normalization) to numerical features in the dataset

Standardization ensures that all numerical features have a **consistent scale**, making it easier for **machine learning models to perform optimally**.

It helps with **gradient descent convergence** and prevents **features with large values from dominating** others in algorithms like **KNN, Logistic Regression, and Neural Networks**.

**Machine Learning Models Building**

**Data Splitting & Feature Scaling**

This section prepares the dataset for **machine learning** by splitting it into training and testing sets.

**Step-by-Step Breakdown**

**1. Import train\_test\_split**

* train\_test\_split is used to divide the dataset into training and testing subsets.

**2. Define Features (X) and Target (y)**

* **X (Features)**: Includes monthly rainfall and seasonal data.
* **y (Target Variable)**: The total column, representing the total rainfall.

**3. Split Data into Training and Testing Sets**

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

* **test\_size=0.2** → 20% of the data is used for testing, while 80% is for training.
* **random\_state=42** → Ensures reproducibility, so the same split occurs every time.

**Why This is Important?**

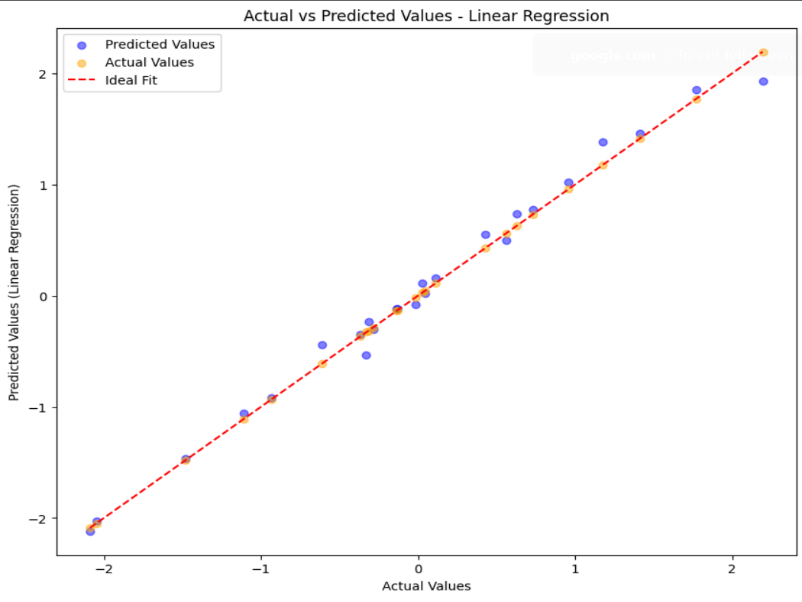
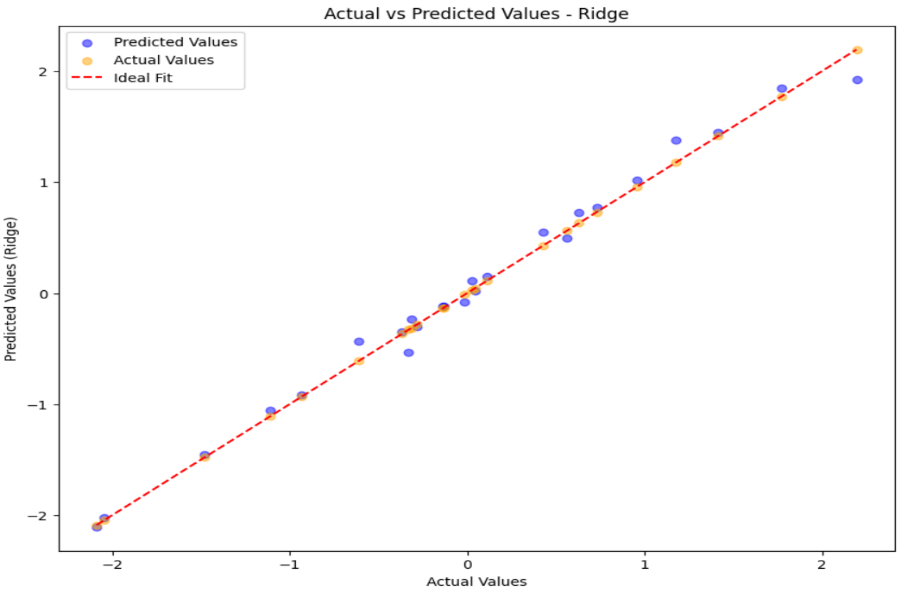
**Prevents Overfitting**: Ensures the model generalizes well by evaluating performance on unseen data. **Ensures Balanced Training & Testing**: Keeps a fair ratio between training and testing data.

**Machine Learning Models Used in the Analysis**

1. **Linear Regression** – A simple model that finds the best linear relationship between input features and the target variable.
2. **Ridge Regression** – A regularized version of Linear Regression that prevents overfitting using L2 regularization.
3. **Lasso Regression** – Similar to Ridge but uses L1 regularization, which can shrink some coefficients to zero, effectively selecting features.
4. **Decision Tree Regressor** – A tree-based model that splits data into smaller subsets but can overfit without pruning.
5. **Random Forest Regressor** – An ensemble of Decision Trees that improves generalization by averaging multiple tree predictions.
6. **K-Neighbors Regressor** – A non-parametric model that predicts based on the average of the k-nearest training points.
7. **AdaBoost Regressor** – A boosting model that combines weak learners (usually Decision Trees) to improve performance iteratively.
8. **Gradient Boosting Regressor** – Another boosting model that builds trees sequentially, correcting previous errors.
9. **XGBRegressor** – A highly optimized version of Gradient Boosting from the XGBoost library, known for its speed and accuracy.

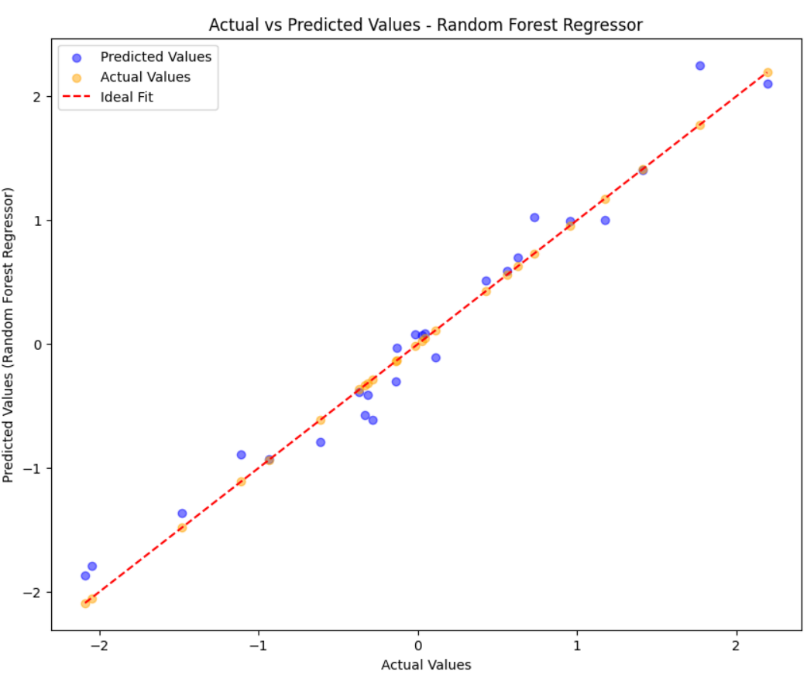
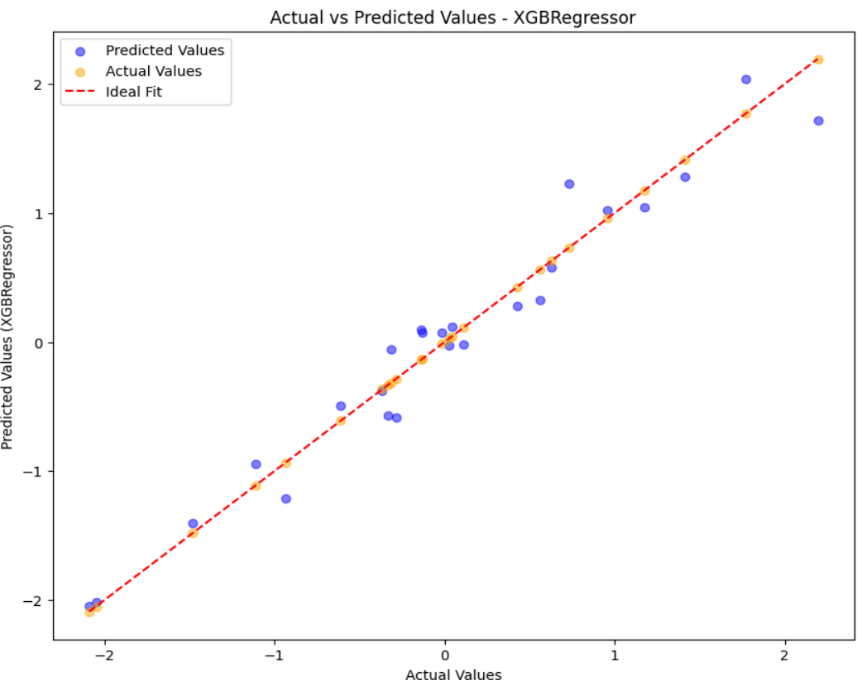
**Best Performing Models (Balanced Train & Test Performance)**

**Linear Regression**  
**Ridge Regression**  
**Random Forest Regressor**  
**XGBRegressor**  
**AdaBoost Regressor**

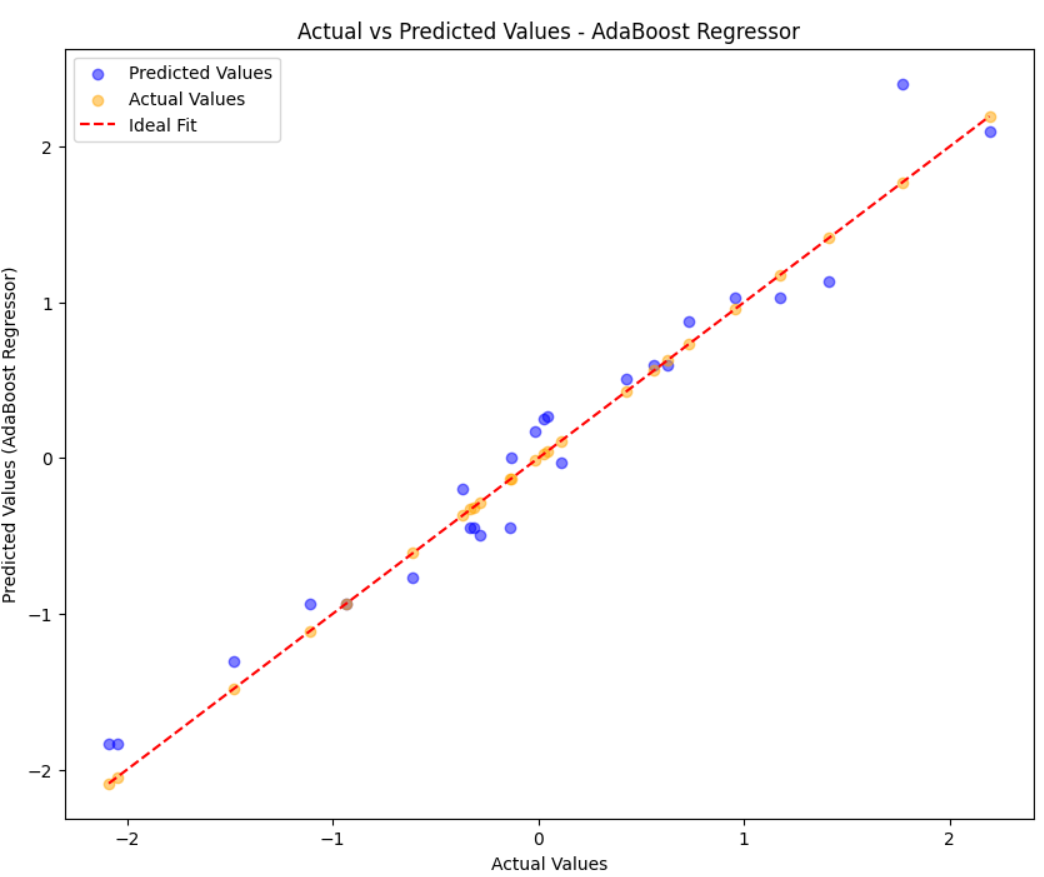
* **Linear Regression & Ridge Regression**
  + Very similar performance, with an **R² ~ 0.99** on both train and test sets.
  + **Root Mean Squared Error (RMSE)** is very low, meaning good generalization.
  + **Ridge Regression** slightly outperforms Linear Regression on the test set.
* **Random Forest & XGBRegressor**
  + **Slight overfitting** (Train R² = 1.00 but Test R² ≈ 0.96).
  + **Random Forest** has slightly better generalization than XGB.
  + **Both are robust models** but may need **hyperparameter tuning** to reduce overfitting.

**XGBRegressor**

* **Train R² = 1.00, but Test R² = 0.96**
* performs well but may need regularization

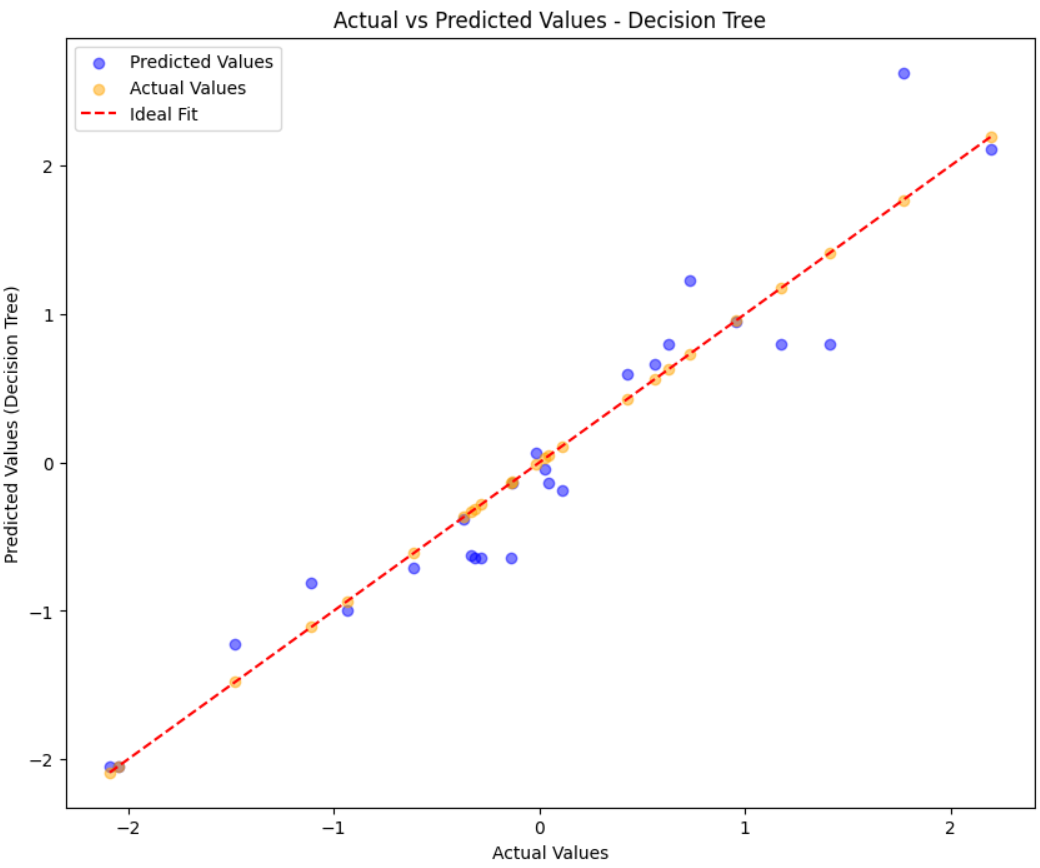


* **AdaBoost Regressor**
  + Good balance with **Train R² = 0.98 and Test R² = 0.94**.
  + Not as strong as XGBoost but still a solid choice.



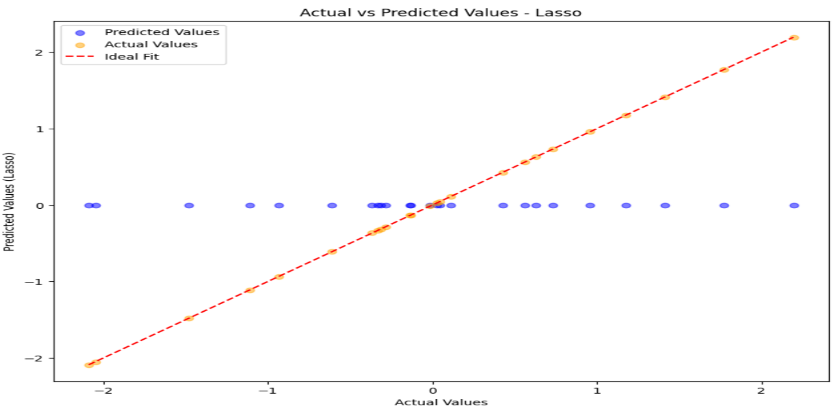
**Overfitting Models**

**Decision Tree Regressor**

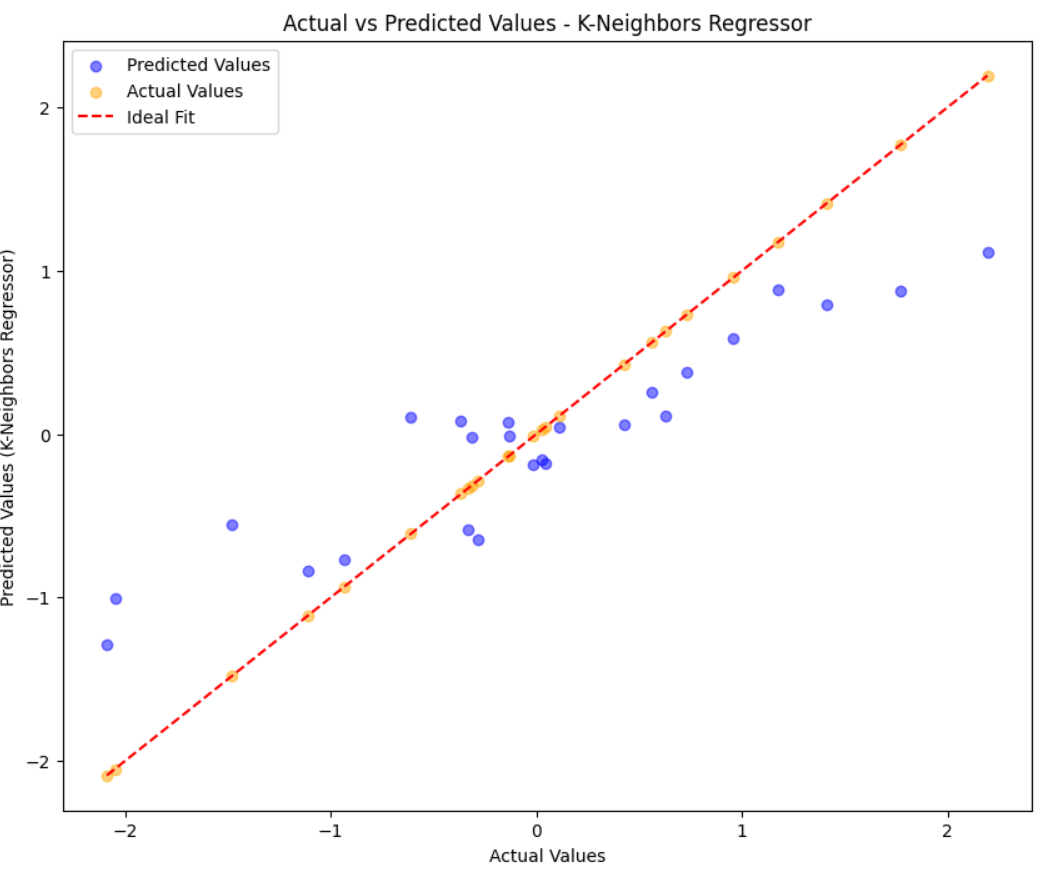
* **Train R² = 1.00, but Test R² = 0.91** → **Clearly overfitting**
* **RMSE for Test Set = 0.3070**, which is **higher than other models**

**Underperforming Models**

**Lasso Regression (Performs the Worst)**

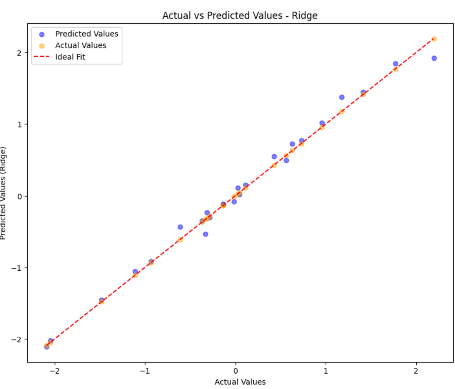
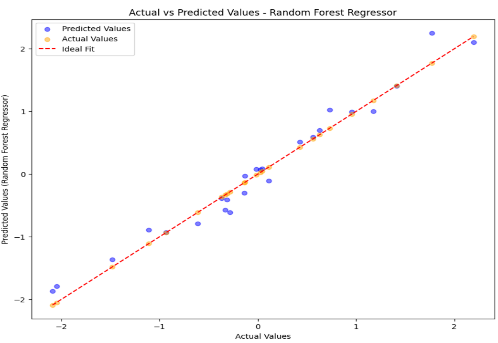
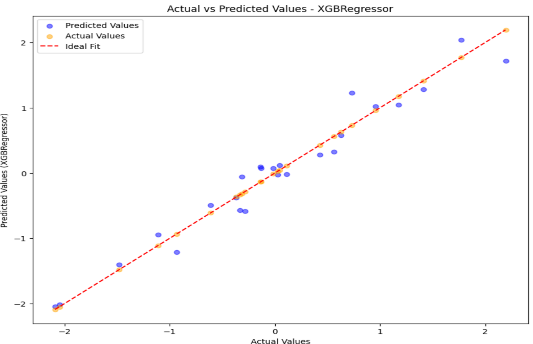
* **R² = 0.00 (Train), -0.0001 (Test)** → Basically **predicting the mean**
* **RMSE is too high (≈1.0)**
* **Why?**
  + Lasso adds **L1 regularization**, which **shrinks coefficients to zero**.
  + This suggests **the dataset may not have many redundant features** Lasso is too aggressive in removing important features.

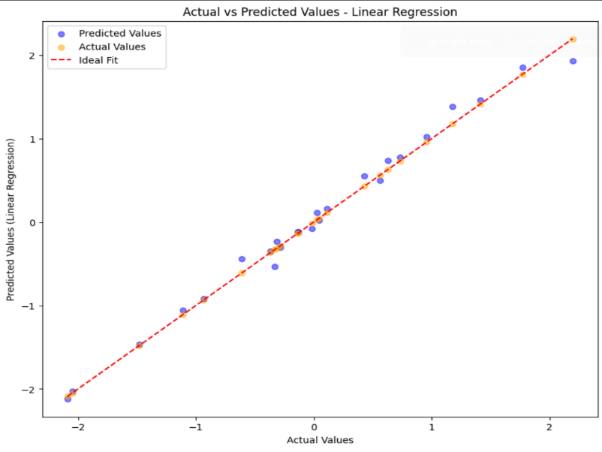
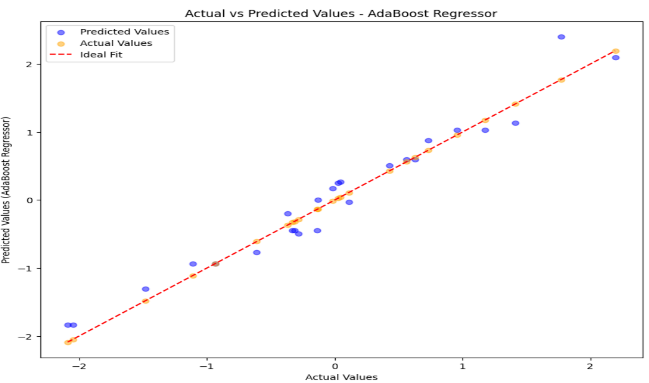
**K-Neighbors Regressor**

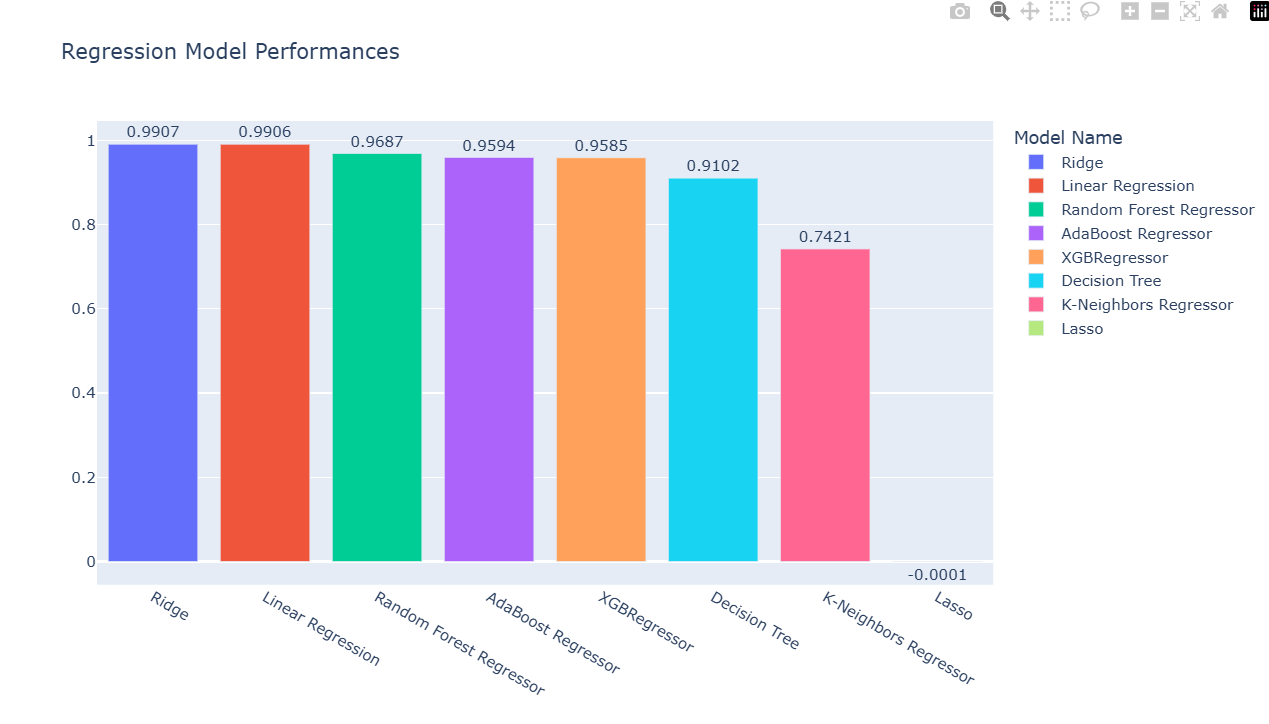
* **Train R² = 0.76, Test R² = 0.74** → **Poor performance compared to others**
* **High RMSE & MAE** suggest that **KNN is not suitable for this dataset**
*  Trying tuning **K (neighbors) parameter** or use tree-based methods.

**Final Recommendation**

The best models for deployment, considering minimal overfitting and strong test R² scores, are:

1. **Ridge Regression (R² = 0.9907)** – Best performance with regularization to prevent overfitting.
2. **Random Forest Regressor (R² = 0.9666)** – Strong performance using ensemble learning.
3. **XGBoost Regressor (R² = 0.9585)** – High accuracy with boosting techniques.
4. **AdaBoost Regressor (R² = 0.9491)** – Reliable model with boosting improvements.





**Next Steps:**

**Hyperparameter tuning**

**Hyperparameter tuning** are using with models are

 **XGBoost Regressor**

 **Gradient Boost Regressor**

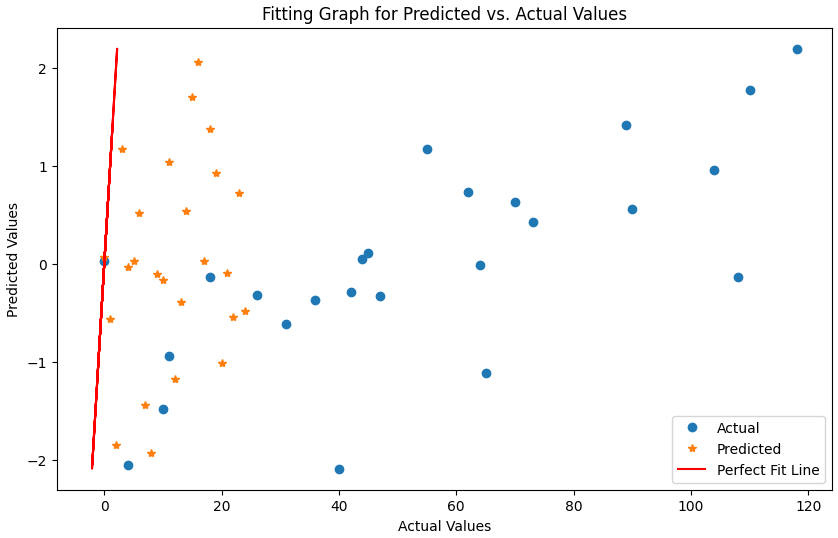
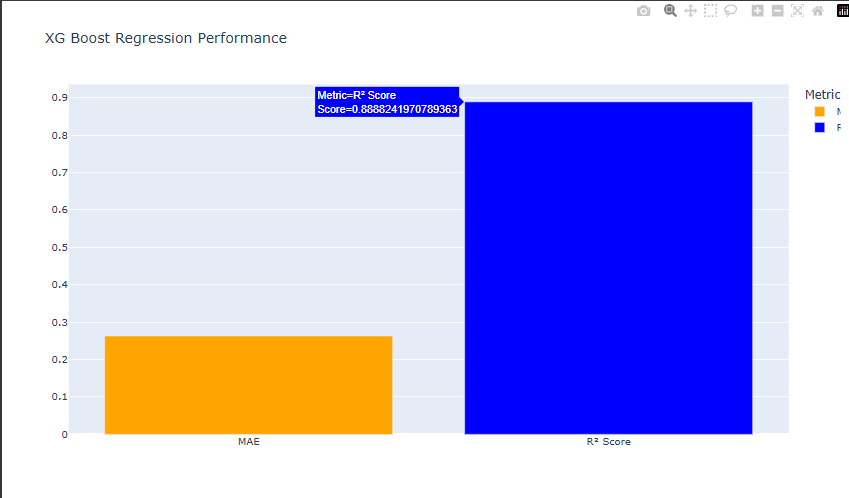
 **Random Forest Regressor**

 **Decision Tree Regressor**

**Comparison of Performance Of XGBoost Regressor Model**

| **Model** | **MAE (Lower is better)** | **R² Score (Higher is better)** |
| --- | --- | --- |
| XGBoost | **0.4317** | **0.8186** |
| XGBoost (Hyperparameter Tuned) | **0.2629** | **0.8888** |

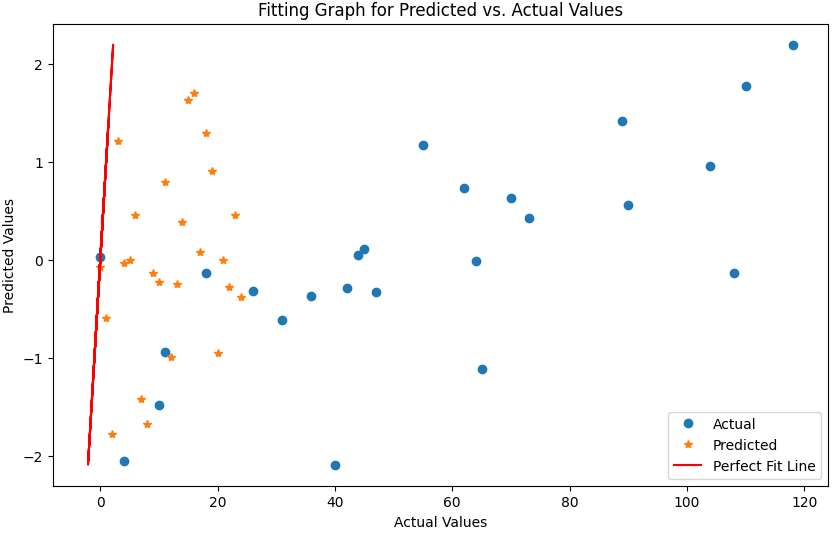
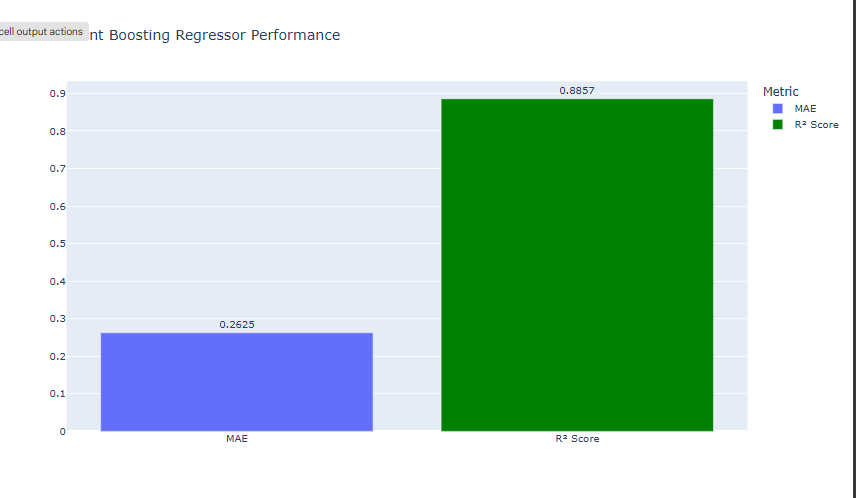
**Key Insights:**

* **MAE Reduced:** From **0.4317** to **0.2629**, showing that tuning improved prediction accuracy.
* **R² Score Increased:** From **0.8186** to **0.8888**, indicating an improvement of the model.
* **Overall Impact:** Hyperparameter tuning **has reduced the error and improved the model's ability to explain variance**, making predictions more reliable.

**Comparison of Performance Gradient Boost Regressor**

| **Model** | **MAE (Lower is better)** | **R² Score (Higher is better)** |
| --- | --- | --- |
| Gradient Boosting (Hyperparameter) | **0.2625** | **0.8857** |

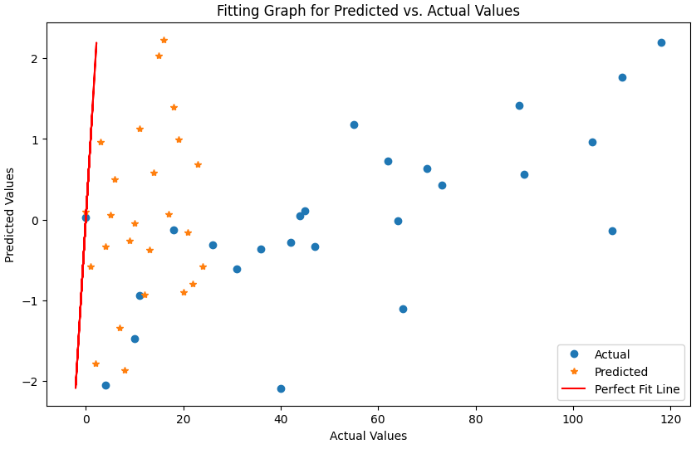
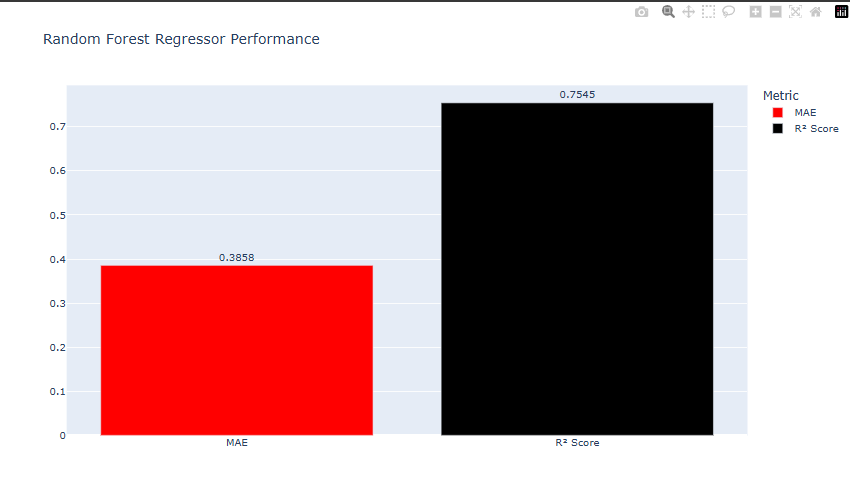
**Key Insights**

* **MAE Reduced:** **0.2625**, prediction accuracy.
* **R² Score Increased:** **0.8857**, meaning the model fits the data better.
* **Performance Boost:** Hyperparameter tuning led to an **increase in accuracy and better model generalisation**.

**Comparison of Performance Random Forest Regressor**

| **Model** | **MAE (Lower is better)** | **R² Score (Higher is better)** |
| --- | --- | --- |
| Random Forest (Default) | **0.4173** | **0.7531** |
| Random Forest (Hyperparameter) | **0.3857** | **0.7545** |

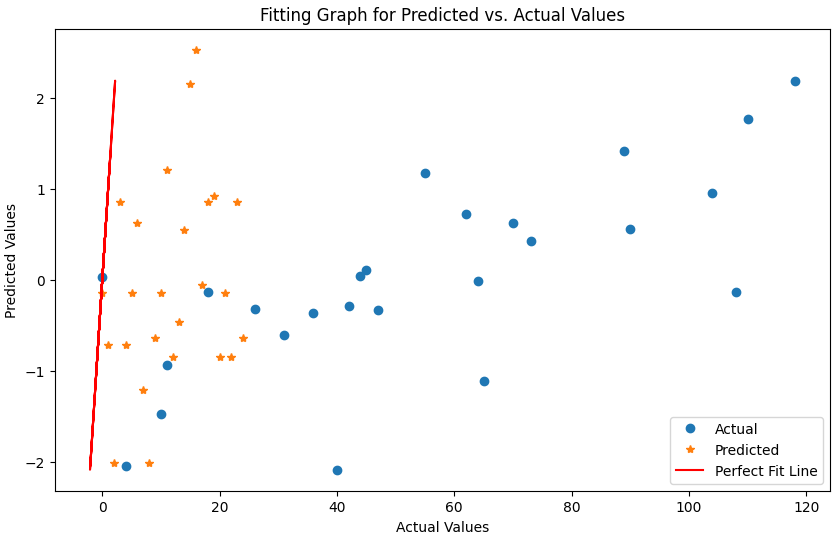
**Key Insights**

* **MAE Reduced:** From **0.4173** → **0.3857**, improving prediction accuracy.
* **R² Score Increased:** From **0.7531** → **0.7545**, meaning the model fits the data better but the default accuracy is more

**Comparison of Performance Decision Tree**

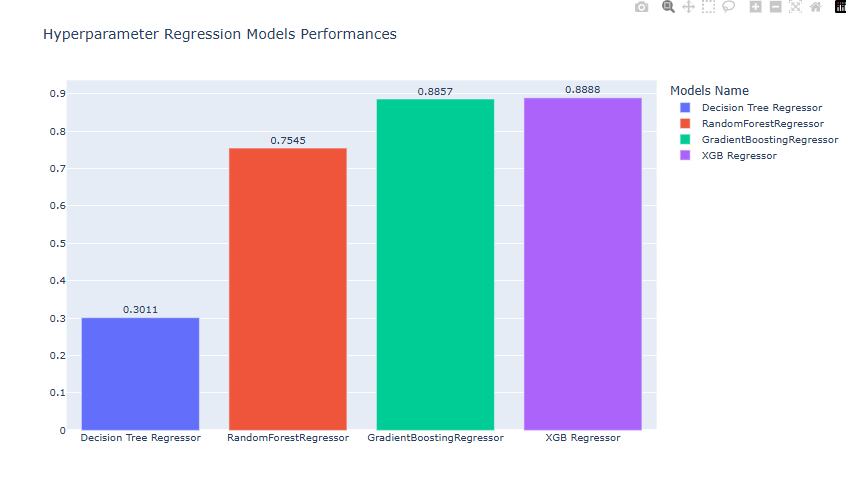
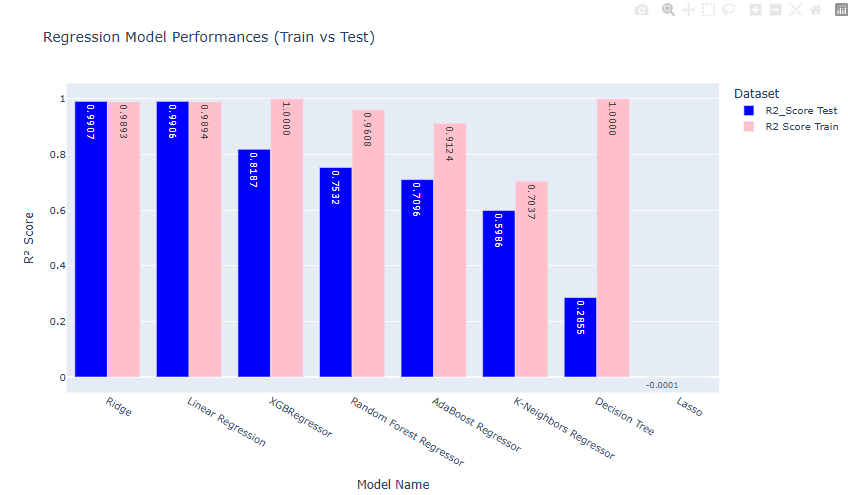
| **Model** | **MAE (Lower is better)** | **R² Score (Higher is better)** |
| --- | --- | --- |
| Decision Tree (Default) | **0.6412** | **0.2855** |
| Decision Tree (Hyperparameter) | **0.7322** | **0.3011** |

**Key Insights**

* **MAE Increased:** From **0.6412 → 0.7322,** indicating a decrease in prediction accuracy.
* **R² Score Slightly Increased: From 0.2855 → 0.3011, suggesting a marginal improvement in the model's ability to explain variance.**
* **Performance Boost:** Hyperparameter tuning led to a **better fit and improved predictions**.

**Visualise The All Models Comparisons**

**Final Conclusion**

1. **Hyperparameter Tuning is Crucial for Complex Models:** The right chart demonstrates the significant positive impact of hyperparameter tuning on complex models like Random Forest, Gradient Boosting, and XGBoost. This optimization allows them to achieve much better performance.
2. **Decision Tree Suffers from Tuning:** The decrease in Decision Tree's performance after tuning suggests that careful tuning is essential, and sometimes default parameters might be optimal.
3. **XGBoost performed best overall after hyperparameter tuning**, making it the most suitable model for this dataset.
4. **Ridge & Linear Regression:** Show consistent performance across train and test sets, indicating good generalization and minimal overfitting.
5. **XGBoost Emerges as the Top Performer (After Tuning):** After hyperparameter tuning, XGBoost achieves the highest R² score, making it the most effective model among those considered

**Best Model Selection**

➡ Based on the insights from both analyses:

Based on the observed issues with the **XGBoost** hyperparameter model trained and the strong performance of **Ridge Regression**, Ridge Regression is the best model for this dataset.

**SNAP Technique**

**Understanding Mumbai's Climate:**

* **Tropical Monsoon Climate:** Mumbai experiences a tropical monsoon climate with distinct wet and dry seasons.
* **Wet Season (June to September):** Heavy rainfall, high humidity.
* **Dry Season (October to May):** Relatively dry with moderate temperatures.
* **Warm Temperatures Year-Round:** Mumbai doesn't have a true **"winter”**

**Reinterpreting the SHAP Plot for Mumbai:**

**1. July and June Still Significant, but for Different Reasons:**

* **High Feature Values (Red):** Still indicate high values, likely representing high rainfall and humidity during the monsoon season.
* **Positive Impact:** In Mumbai, this would strongly correlate with the wet season, significantly impacting whatever the model is predicting (e.g., rainfall intensity, flood risk, water levels).

**2. January, March, and February Now Indicate Dry Season:**

* **Low Feature Values (Blue):** Now represent the dry season, with lower rainfall and humidity.
* **Negative Impact:** This means that lower values in these months (i.e., less rainfall) are associated with lower predicted values by the model.

**3. May, April, November, and December: Transition Periods:**

* **Moderate Positive Impact:** These months might represent the transition periods between the wet and dry seasons.
* **Feature Values:** Would reflect the gradual changes in rainfall and humidity.

**4. August, September, and October: Peak Monsoon Impact:**

* **Spread Around Zero:** While still part of the monsoon, the spread in SHAP values might indicate variability in rainfall intensity during these months.
* **Overall Positive Trend:** Still a positive trend, but potentially more nuanced than June and July.

**Contextual Considerations for Mumbai:**

* **Coastal Location:** Mumbai's coastal location influences its climate, moderating temperature fluctuations.
* **Urban Environment:** The urban heat island effect might slightly affect the temperature patterns.

**In summary, with the location Mumbai, the SHAP plot now highlights the strong influence of the monsoon season and the dry season on the model's output. The insights are consistent with Mumbai's tropical monsoon climate and the significant impact of rainfall and humidity on the predicted variable.**