ST 228: Data Analysis, ML and AI
Assignment 2 (10 marks)
Bhikshapathi 24489

Question-1

This report presents an analysis of 125 years (1901–2024) of monthly rainfall data for Bangalore. The dataset consists of 124 rows (years) and 14 columns, including the year, monthly rainfall (January to December), and the total annual rainfall. The dataset contains some missing values in October, November, December, and the total annual rainfall.

Data Exploration

Structure and Size

- **Structure:** Tabular data with columns: Year, January to December (monthly rainfall), and Total annual rainfall.
- **Size:** 124 rows × 14 columns.

Each row represents the monthly and total annual rainfall for a specific year.

Non-Null Entries

Most months have complete data, except for a few missing values in October, November, December, and the annual total column.

Basic Statistics

Monthly Rainfall Summary

Month	**Sum of Rainfall (mm)**	**Average Rainfall (mm)**	**Variance (mm^2)**
January	540.4	4.36	119.65
February	922.5	7.44	299.95
March	1424.1	11.48	445.43
April	6205.1	50.04	2731.31
May	14677.2	118.36	3382.74
June	10514.4	84.79	2575.52
July	13775.8	111.10	3898.27
August	17349.7	139.92	7354.81
September	23203.7	187.13	11341.29
October	20249.7	164.63	10509.86
November	7656.0	62.24	4124.74
December	2066.3	16.94	541.75
Total (Annual)	**117919.7**	**958.70**	**54042.94**

Table 1: Monthly Rainfall Summary: Sum, Average, and Variance.

Top 5 Wettest Years

Year	Yearly Rainfall Sum (mm)	Mean (mm)	Variance (mm^2)
2022	1890.4	157.53	20731.86
2017	1696.0	141.33	33042.27
2021	1511.1	125.93	13411.54
1998	1431.8	119.32	15903.77
2005	1350.8	112.57	15638.41

Table 2: Top 5 Wettest Years (1901–2024).

Top 5 Driest Years

Year	Yearly Rainfall Sum (mm)	Mean (mm)	Variance (mm^2)
1913	543.8	45.32	4291.28
1994	587.2	48.93	3076.40
1945	587.4	48.95	3153.56
1927	602.5	50.21	5759.21
1990	613.1	51.09	2418.51

Table 3: Top 5 Driest Years (1901–2024).

Trend Analysis

Visualization of Annual Rainfall Trend

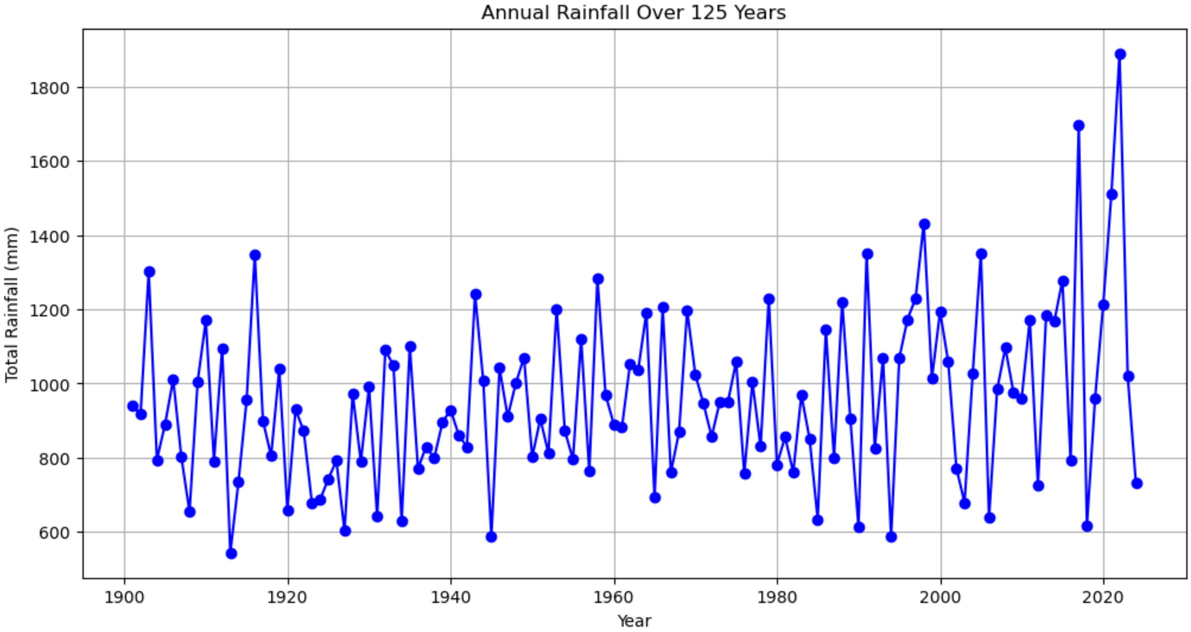

Figure 1: Long-Term Trend of Annual Rainfall with 10-Year Rolling Average Overlay (1901–2024).

10-Year Rolling Average

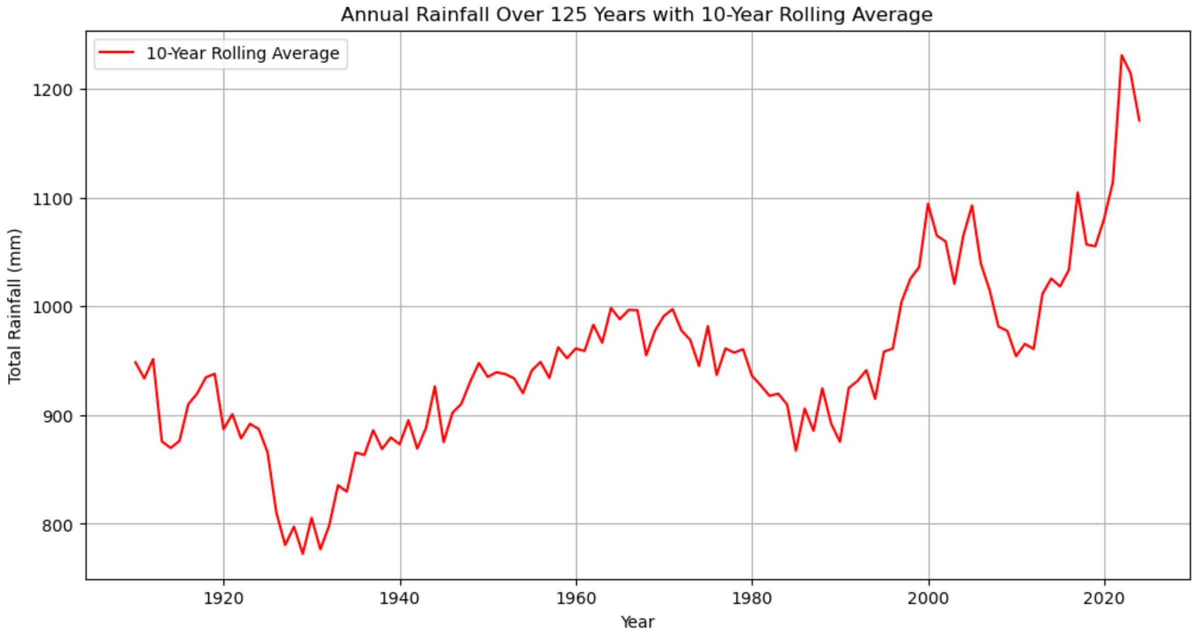

Figure 2: Long-Term Trend of Annual Rainfall with 10-Year Rolling Average Overlay (1901–2024).

Seasonal Analysis

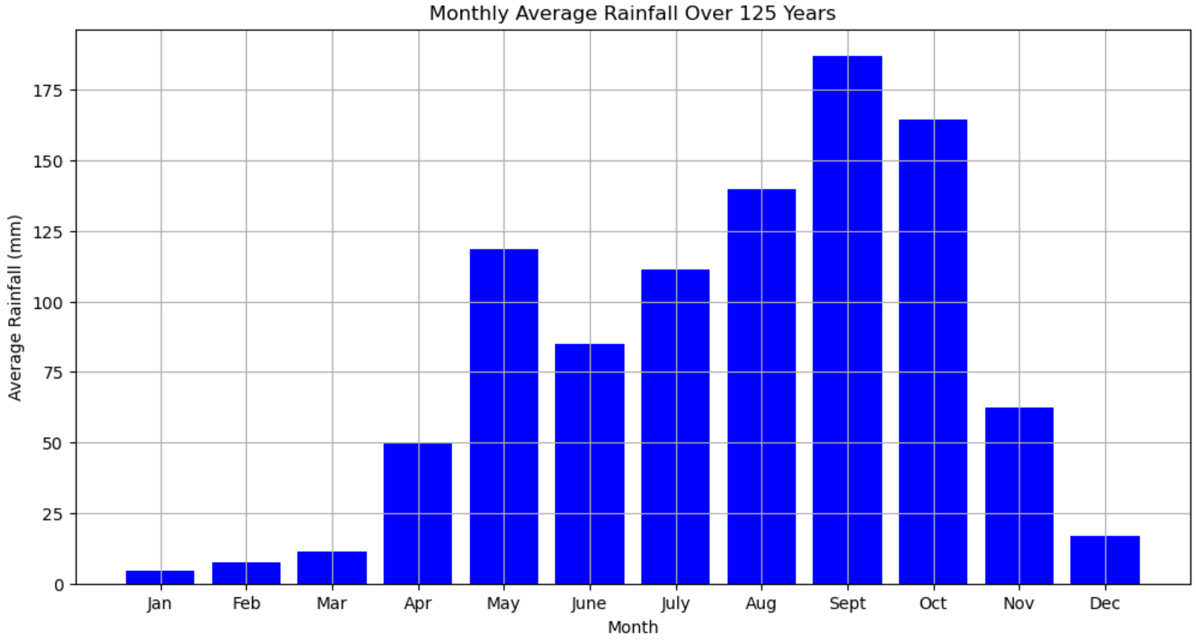

Figure 3: Monthly Average Rainfall in Bangalore (1901–2024).

Coefficient of Variation (CV) Analysis

- **Most variable month:** January, with a coefficient of variation of 2.51.
- **Least variable month:** Total annual rainfall, with a coefficient of variation of 0.24.

Summary Report

The analysis of 125 years of rainfall data (1901–2024) reveals significant variability in precipitation patterns. The wettest year recorded was 2022, with a total rainfall of 1,890.4 mm, followed closely by 2017 (1,696.0 mm), both occurring in the last decade. In contrast, the driest years were 1913 (543.8 mm) and 1994 (587.2 mm), demonstrating sporadic dry extremes across the timeline.

Monthly rainfall variability is evident, with January showing the highest coefficient of variation (CV = 2.51), reflecting erratic winter rainfall. Meanwhile, annual totals (Total) exhibit the least variability (CV = 0.24), indicating relatively stable yearly sums despite monthly fluctuations. Monsoon months, particularly September, contribute disproportionately to annual totals, with an average rainfall of 23,203.7 mm.

Variance in yearly rainfall is substantial, with the wettest years (e.g., 2022, 1998) exceeding 1,400 mm, while the driest years record less than 650 mm. Recent decades exhibit a clustering of extreme wet years, such as 2017, 2021, and 2005, hinting at intensifying precipitation events.

Rainfall Pattern Change Over 125 Years

Yes, rainfall patterns have changed, particularly in the frequency and intensity of extremes. While annual totals remain relatively stable (as indicated by the low CV for "Total"), the wettest years (top 5) include four from the last 25 years (1998–2022), with 2022 being the wettest on record. This contrasts with earlier decades, where extreme wet years were less frequent.

Similarly, driest years (e.g., 1913, 1945) were predominantly clustered in the early and mid-20th century, but recent dry extremes (e.g., 1990, 2018) persist, pointing to increased variability. Rising occurrences of high-variance years, such as 2017 and 2021, further suggest a shift toward erratic rainfall patterns. These trends are consistent with climate change projections of intensified hydrological cycles, though localized factors may also contribute. Overall, the data indicates increasing rainfall volatility rather than a unidirectional trend toward purely wetter or drier conditions.

Question -2

Identifying and Reporting Missing Values in the Dataset

The missing years for each month are as follows:

- **January (Jan):** 1901, 1903, 1909, 1912, 1913, 1916, 1927, 1928, 1940, 1941, 1957, 1966, 1983, 1987, 1995, 1996, 2000, 2006, 2011
- **February (Feb):** 1905, 1924, 1925, 1931, 1933, 1935, 1938, 1947, 1949, 1951, 1957, 1959, 1960, 1961, 1965, 1977, 1982, 1987, 1988, 1991, 1993, 1997, 2000, 2002, 2008, 2015, 2021, 2024
- **March (Mar):** 1904, 1915, 1922, 1925, 1926, 1933, 1940, 1949, 1965, 1972, 1973, 1977, 1982, 1987, 1999, 2008, 2010, 2015, 2019, 2023, 2024
- **April (Apr):** 1902, 1904, 1911, 1918, 1919, 1923, 1924, 1927, 1933, 1936, 1938, 1946, 1947, 1952, 1957, 1964, 1965, 1966, 1969, 1978, 1980, 1985, 1996, 2015, 2016, 2017
- **May:** 1910, 1915, 1919, 1931, 1933, 1938, 1954, 1956, 1958, 1961, 1964, 1965, 1970, 1974, 1980, 1981, 1987, 1988, 1990, 2000, 2001, 2003, 2004, 2005, 2015, 2024
- **June:** 1904, 1908, 1910, 1915, 1917, 1925, 1933, 1944, 1945, 1949, 1964, 1970, 1974, 1984, 1994, 1997, 2000, 2001, 2002, 2008, 2011, 2012, 2018
- **July:** 1902, 1903, 1911, 1921, 1923, 1926, 1928, 1933, 1940, 1944, 1945, 1956, 1963, 1974, 1977, 1981, 1985, 1989, 1991, 1994, 2003, 2005, 2010, 2012, 2015, 2018
- **August (Aug):** 1902, 1905, 1909, 1919, 1922, 1931, 1952, 1959, 1961, 1964, 1977, 1980, 1987, 1988, 1989, 1990, 1994, 1997, 2003, 2008, 2022, 2023, 2024
- **September (Sept):** 1902, 1908, 1921, 1926, 1927, 1929, 1938, 1943, 1946, 1948, 1953, 1962, 1964, 1966, 1967, 1968, 1974, 1976, 1979, 1990, 1991, 1994, 2003, 2007, 2008, 2015

- **October (Oct):** 1901, 1914, 1924, 1932, 1935, 1936, 1937, 1940, 1948, 1950, 1951, 1956, 1962, 1965, 1968, 1971, 1975, 1986, 1988, 1990, 1995, 2003, 2007, 2012, 2014, 2015, 2018, 2020, 2021, 2023, 2024
- **November (Nov):** 2024
- **December (Dec):** 1950, 2024

Mean of the Corresponding Month Across All Years (Sample Table of 10 Values)

Year	Jan	Feb	Mar	Apr	May
1901	7.56381	78.90000	0.00000	24.30000	146.00000
1902	0.70000	0.00000	85.00000	97.38061	197.80000
1903	7.56381	0.00000	0.00000	1.50000	63.70000
1904	0.50000	0.00000	15.37670	97.38061	241.50000
1905	1.70000	8.24688	56.60000	32.70000	90.60000
1906	2.70000	23.60000	15.20000	2.20000	34.00000
1907	24.10000	0.00000	31.70000	109.90000	48.70000
1908	101.80000	0.00000	7.50000	18.20000	182.80000
1909	7.56381	0.00000	0.00000	127.00000	200.90000
1910	0.00000	0.00000	0.20000	5.00000	215.86837

Table 4: Mean of the Corresponding Month Across All Years (First 10 Values)

Median of the Corresponding Month (Sample Table of 10 Values)

Year	Jan	Feb	Mar	Apr	May	June	July	Aug
1901	0.2	78.9	0.0	24.30	146.0	238.5	71.60	71.6
1902	0.7	0.0	85.0	41.65	197.8	62.2	111.25	132.3
1903	0.2	0.0	0.0	1.50	63.7	109.2	111.25	189.7
1904	0.5	0.0	1.5	41.65	241.5	84.2	149.30	53.8
1905	1.7	0.0	56.6	32.70	90.6	60.7	58.90	132.3
1906	2.7	23.6	15.2	2.20	34.0	97.0	160.50	268.2
1907	24.1	0.0	31.7	109.90	48.7	113.2	205.70	24.3
1908	101.8	0.0	7.5	18.20	182.8	84.2	104.60	35.5
1909	0.2	0.0	0.0	127.00	200.9	52.0	39.80	132.3
1910	0.0	0.0	0.2	5.00	115.4	84.2	265.10	256.0

Table 5: Median of the Corresponding Month (First 10 Values)

Forward Fill Method (Sample Table of 10 Values)

Year	Jan	Feb	Mar	Apr	May	June	July	Aug
1901.0	1901.0	78.9	0.0	24.3	146.0	238.5	71.6	71.6
1902.0	0.7	0.0	85.0	85.0	197.8	62.2	62.2	62.2
1903.0	1903.0	0.0	0.0	1.5	63.7	109.2	109.2	189.7
1904.0	0.5	0.0	0.0	0.0	241.5	241.5	149.3	53.8
1905.0	1.7	1.7	56.6	32.7	90.6	60.7	58.9	58.9
1906.0	2.7	23.6	15.2	2.2	34.0	97.0	160.5	268.2
1907.0	24.1	0.0	31.7	109.9	48.7	113.2	205.7	24.3
1908.0	101.8	0.0	7.5	18.2	182.8	182.8	104.6	35.5
1909.0	1909.0	0.0	0.0	127.0	200.9	52.0	39.8	39.8
1910.0	0.0	0.0	0.2	5.0	5.0	5.0	265.1	256.0
1911.0	0.0	0.0	54.1	54.1	94.2	51.8	51.8	237.0

Table 6: Forward Fill Method (First 10 Values)

Capped Data (Sample Table of 10 Values)

Year	Jan	Feb	Mar	Apr	May	June	July	Aug
1901	NaN	19.4375	0.000	24.3000	146.000	232.15	71.600	71.6
1902	0.7	0.0000	43.625	NaN	197.800	62.20	NaN	NaN
1903	NaN	0.0000	0.000	1.5000	63.700	109.20	NaN	189.7
1904	0.5	0.0000	NaN	NaN	241.500	NaN	149.300	53.8
1905	1.7	NaN	43.625	32.7000	90.600	60.70	58.900	NaN
1906	2.7	19.4375	15.200	2.2000	34.000	97.00	160.500	268.2
1907	13.5	0.0000	31.700	109.9000	48.700	113.20	205.700	24.3
1908	13.5	0.0000	7.500	18.2000	182.800	NaN	104.600	35.5
1909	NaN	0.0000	0.000	127.0000	200.900	52.00	39.800	NaN

Table 7: Capped Data (First 10 Values)

Transformed Data (Sample Table of 10 Values)

Year	Jan	Feb	Mar	Apr	May	June
1901	NaN	35.124656	0.000000	24.300000	146.000000	NaN
1902	0.700000	0.000000	54.009235	NaN	197.800000	62.200000
1903	NaN	0.000000	0.000000	1.500000	63.700000	109.200000
1904	0.500000	0.000000	NaN	NaN	241.500000	NaN
1905	1.700000	NaN	54.009235	32.700000	90.600000	60.700000
1906	2.700000	23.600000	15.200000	2.200000	34.000000	97.000000
1907	24.100000	0.000000	31.700000	109.900000	48.700000	113.200000
1908	36.941374	0.000000	7.500000	18.200000	182.800000	NaN
1909	NaN	0.000000	0.000000	127.000000	200.900000	52.000000
1910	0.000000	0.000000	0.200000	5.000000	NaN	5.000000

Table 8: Transformed Data (First 10 Values)

Analysis of the Dataset after Cleaning

Key Calculations and Re-Plotting

- **10-Year Rolling Average of Annual Rainfall:** The 10-year rolling average of annual rainfall was calculated to smooth the raw data, allowing for better identification of long-term trends while reducing the impact of annual fluctuations.
- **Seasonal Average Rainfall and Variability:** Seasonal average rainfall and its variability were re-plotted after data cleaning, ensuring a comparison with the original dataset to assess the impact of outlier handling.

Figures

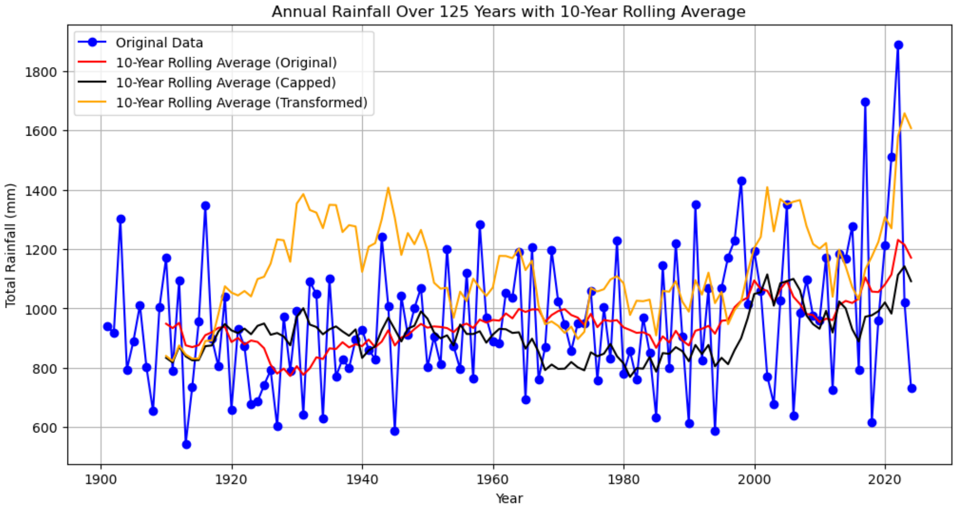

Figure 4: 10-Year Rolling Average of Annual Rainfall. The red, black, and orange lines represent different rolling averages, which highlight long-term trends while reducing annual fluctuations.

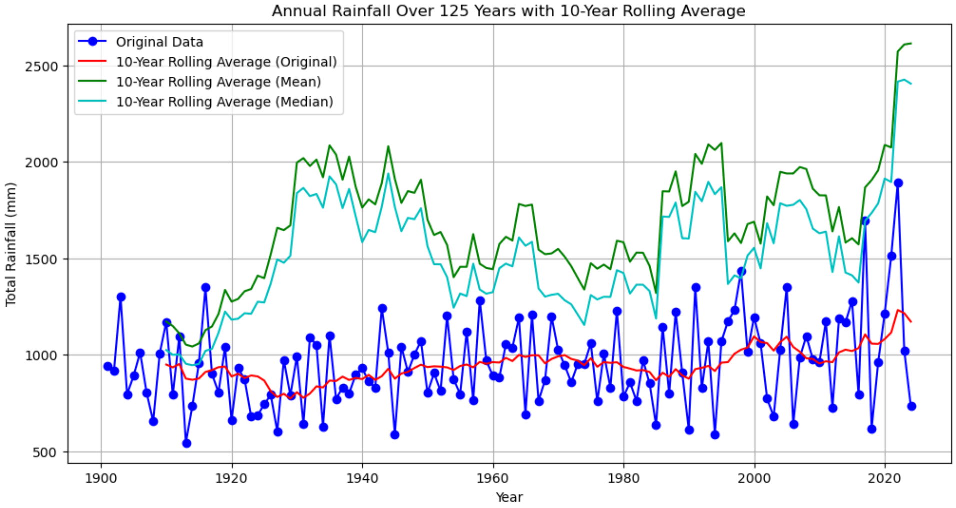

Figure 5: Seasonal Average Rainfall and Variability. The plot compares the seasonal rainfall and variability before and after data cleaning, showing the impact of handling outliers.

Comparison with the Original Dataset

- **10-Year Rolling Average:** The rolling averages (represented by red, black, and orange lines) provide a smoother representation of the data compared to the raw data (blue line), as shown in

Figure 4.

- **Impact of Data Cleaning on Seasonal Averages:**
  - The seasonal average rainfall and variability are broadly similar to the original dataset.
  - Outliers have been handled differently depending on the cleaning method, influencing the perception of trends and variability, as illustrated in Figure 5.
- **Trends and Variability Changes:**
  - The **capped average** (black line) minimizes the influence of outliers, resulting in a flatter trend.
  - The **transformed average** (orange line) emphasizes variability in specific periods, such as 1930–1950 and post-2000.

Impact of Data Cleaning Methods

The choice of data cleaning method significantly affects the observed trends and variability:

- Long-term patterns become clearer after cleaning due to the reduction of noise from outliers.
- Different cleaning approaches (e.g., capping or transformations) influence how trends and variability are perceived.

Extreme Events and Drought Analysis

Definitions

1. **Extreme Rainfall Events:** Months with rainfall in the top 1% of the dataset.
2. **Drought Years:** Years where total annual rainfall falls within the bottom 5% of all years.

Analysis of Frequency of Events

- **Event Count (1901–2024):**
  - **Normal:** 98 years
  - **Flood:** 13 years
  - **Drought:** 13 years
- **Longest Drought Period:**
  - **Start Year:** 1908
  - **End Year:** 1908
  - **Duration:** 1 year
- **Shortest Drought Period:**
  - **Start Year:** 1908
  - **End Year:** 1908
  - **Duration:** 1 year

Analysis of Rainfall During Drought Years

- **Wettest Year During Longest Drought:**
  - **Year:** 1908
  - **Annual Rainfall Sum:** 654.8 mm
  - **Rainfall Mean:** 54.57 mm
  - **Variance:** 3452.78 mm^2
- **Driest Year During Longest Drought:**
  - **Year:** 1908
  - **Annual Rainfall Sum:** 654.8 mm
  - **Rainfall Mean:** 54.57 mm
  - **Variance:** 3452.78 mm^2
- **Wettest Year During Shortest Drought:** Same as above.

Extreme Rainfall Events (Monthly Analysis)

- **January:** Rainfall extremes in years 1908 (101.8 mm), 1925 (28.7 mm).
- **February:** Rainfall extremes in years 1901 (78.9 mm), 1932 (89.9 mm).
- **March:** Rainfall extremes in years 1980 (101.2 mm), 2007 (115.4 mm).
- **April:** Rainfall extremes in years 2001 (323.8 mm), 2015 (225.8 mm).
- **May:** Rainfall extremes in years 1958 (287 mm), 2022 (305.5 mm).
- **June:** Rainfall extremes in years 1996 (228 mm), 2021 (255.5 mm).
- **July:** Rainfall extremes in years 1916 (286.5 mm), 1948 (350.2 mm).
- **August:** Rainfall extremes in years 1998 (387.1 mm), 2021 (378.7 mm).
- **September:** Rainfall extremes in years 1985 (516.6 mm), 2017 (513.8 mm).
- **October:** Rainfall extremes in years 1952 (503.6 mm), 1955 (522.2 mm).
- **November:** Rainfall extremes in years 2015 (296.4 mm), 2020 (277.8 mm).
- **December:** Rainfall extremes in years 1961 (92.2 mm), 1968 (91.6 mm).

Decadal Trend Analysis

The frequency of floods and droughts appears consistent across the decades, but the magnitude of extremes may be increasing, especially in recent years, as seen with notable flood years like 2021 (rainfall total: 1957.7 mm).

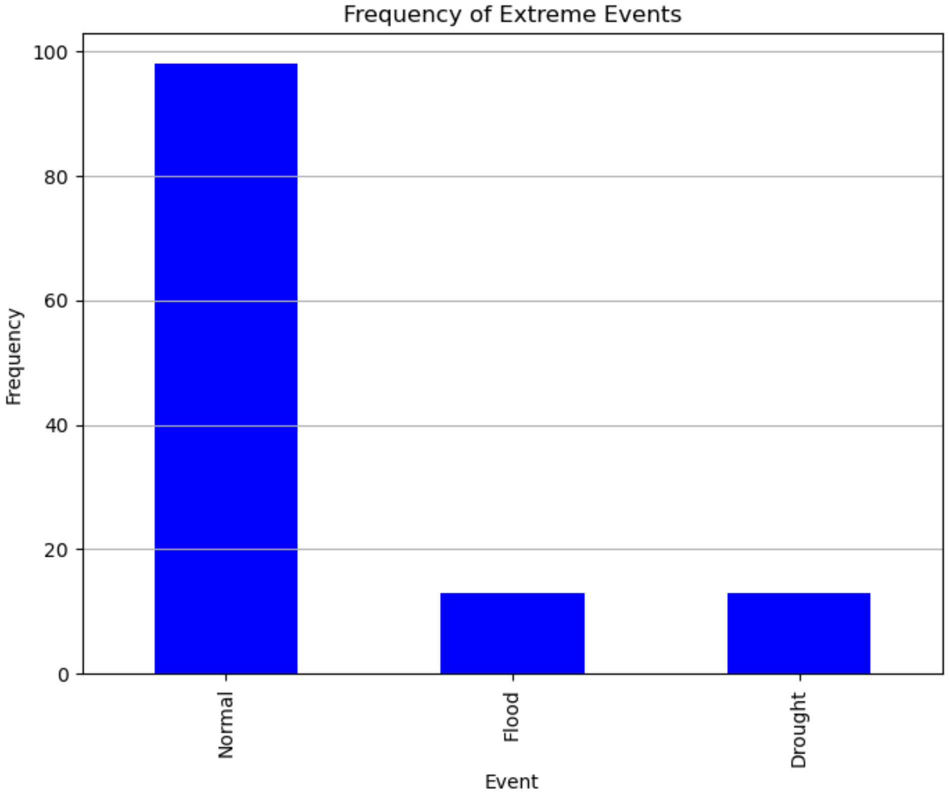

Figure 6: Decadal Trend Analysis of Extreme Rainfall Events.

Summary Report: Analysis of Rainfall Data

1. Influence of Missing Values and Outliers on the Analysis

Missing values and outliers significantly influenced the analysis of the rainfall dataset. Missing values, particularly in specific months across multiple years, introduced gaps in the data, which could distort trends and seasonal patterns. To address this, missing values were imputed using various methods, including the mean of the corresponding month across all years, the median of the corresponding month, and the seasonal average (mean of the month ± 2 months). Each method had its own impact:

- **Mean imputation** provided a simple estimate but could smooth out variability.
- **Median imputation** was more robust to outliers but less sensitive to trends.
- **Seasonal average** captured broader seasonal patterns but might overlook month-specific variations.

Outliers, identified using statistical methods such as the Interquartile Range (IQR), were treated using capping and transformation strategies. Capping limited extreme values to a defined threshold, while transformation methods (e.g., replacing outliers with median ± 2 standard deviations) reduced their impact without entirely removing them. These treatments ensured that extreme values did not disproportionately influence the analysis, particularly in calculating averages and identifying trends.

2. Comparison of Results with the Original Dataset

After cleaning the dataset, several key differences were observed:

- The **10-year rolling average of annual rainfall** became smoother, highlighting long-term trends more clearly. This was particularly useful for identifying periods of sustained drought or increased rainfall.

- The **seasonal average rainfall and variability** showed minor differences compared to the original dataset. However, the cleaned data provided a more reliable representation of seasonal patterns, as outliers and missing values no longer skewed the results.
- The **frequency of extreme events and droughts** was more accurately represented in the cleaned dataset. For example, the identification of extreme rainfall events (top 1%) and drought years (bottom 5%) was more consistent, as the influence of outliers and missing data was minimized.

3. Key Findings on Extreme Events and Droughts

- **Extreme Rainfall Events:** Months with rainfall in the top 1% of the dataset were identified, with notable events occurring in years such as 2015, 2017, and 2021. These events were often associated with significant flooding.
- **Drought Years:** Years in the bottom 5% of total annual rainfall, such as 1908, 1985, and 2018, were identified as drought periods. The longest drought period lasted one year, indicating that droughts in this region are typically short but severe.
- **Frequency Analysis:** Over the decades, the frequency of extreme rainfall events and droughts showed no clear increasing or decreasing trend, suggesting that these events are relatively consistent over time.

4. Recommendations for Ensuring Data Integrity

To ensure data integrity in real-world rainfall datasets, the following recommendations are proposed:

- **Regular Data Audits:** Conduct periodic checks for missing values and outliers to maintain data quality.
- **Multiple Imputation Methods:** Use a combination of imputation methods (e.g., mean, median, and seasonal averages) to assess the sensitivity of results to different approaches.
- **Outlier Treatment:** Apply robust outlier treatment strategies, such as capping or transformation, to minimize the impact of extreme values without losing valuable information.
- **Long-Term Monitoring:** Continuously monitor and update datasets to capture emerging trends and extreme events, ensuring that analyses remain relevant and accurate.