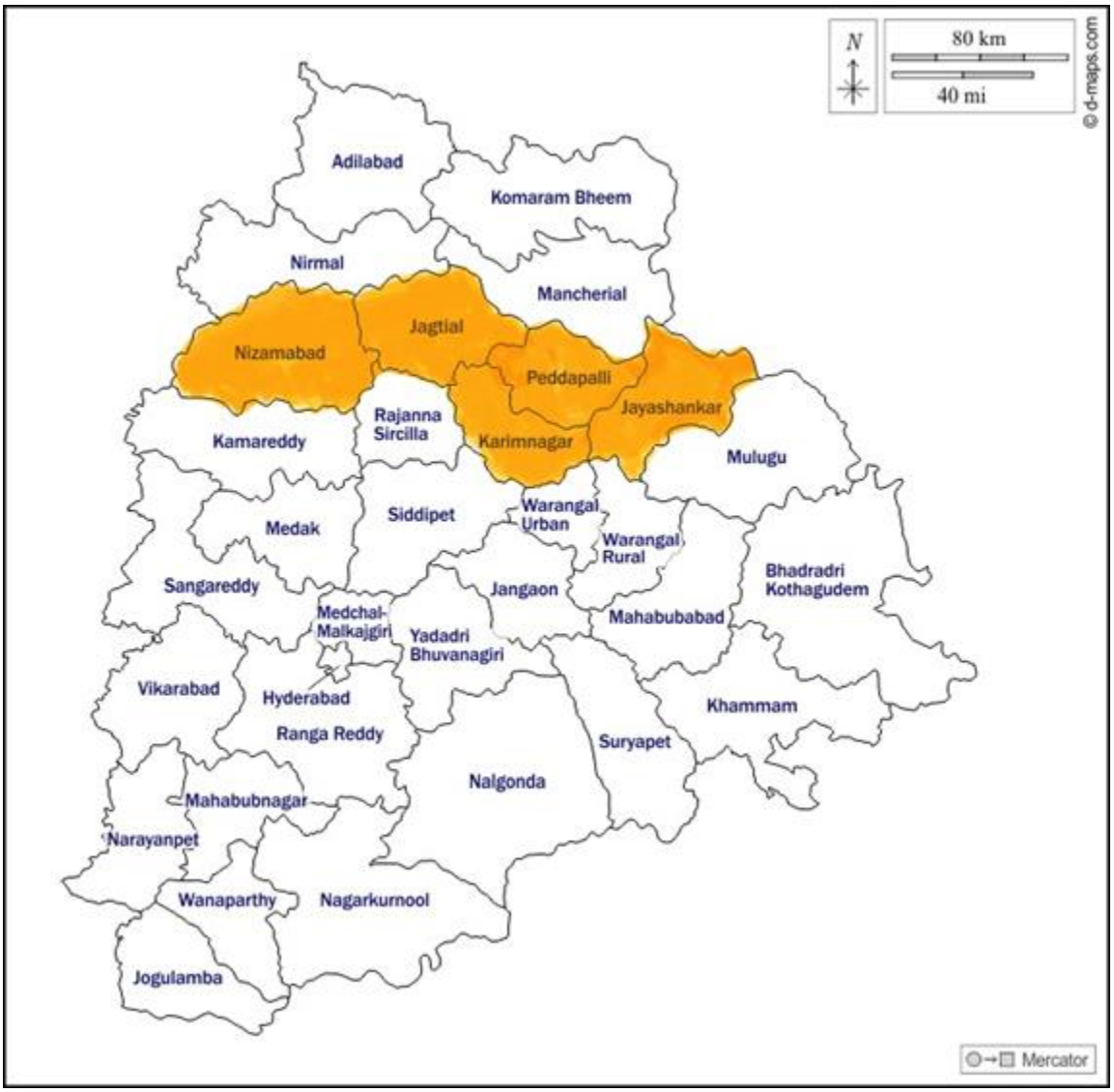
Statistical Analysis of IMD Quarter-Degree Rainfall Data in 5 Districts of Telangana State

Course: Applied Computational Laboratory Course code: CE60501

Name: Prashant Joshi, Swagatam Bora, Shrutee Jain

1. Introduction

The present study focuses on conducting a comprehensive statistical analysis of the IMD quarter-degree rainfall data spanning 71 years (1950-2021) for five districts in Telangana, namely Nizamabad, Jagtial, Pedapalli, Jayashankar Bhupalpally, and Karimnagar. The dataset comprises rainfall information across 22 grid points within the study area.



- Purpose

The primary purpose of this study is to understand the temporal and spatial patterns of rainfall in the specified Telangana districts over the past seven decades. Through rigorous statistical analysis, we aim to identify trends, and variations in the rainfall across the designated grid points.

- Scope

The scope of the study encompasses:

1. Temporal Analysis: Investigating the long-term trends and variations in rainfall over the 71-year period.

2. Spatial Analysis: Examining the spatial distribution of rainfall across the 22 grid points within the five districts.

3. District-Level Comparisons: Comparing and contrasting rainfall patterns among the districts to discern any distinctive features or trends.

4. Statistical Testing: Conducting various statistical tests to explore relationships, differences, and significant patterns in the rainfall data.

2. Data Description

The dataset under examination comprises IMD quarter-degree rainfall data for a span of 71 years, from 1950 to 2021. The study area encompasses five districts in Telangana: Nizamabad, Jagtial, Pedapalli, Jayashankar Bhupalpally, and Karimnagar.

2.1 Variables

Temporal Variable: Represented by the calendar years from 1950 to 2021, spanning the entire duration of the dataset.

Spatial Variable: Represented by the specific geographical location within each district where rainfall measurements were recorded. There are a total of 22 grid points.

Quantitative Variable: The main variable of interest, represents the amount of rainfall (in millimeters) recorded at each grid point for a given year.

2.2 Units

Year: The temporal variable is measured in calendar years (e.g., 1950, 1951, ..., 2021).

GridPoint: Spatial variable indicating the specific location within a district where rainfall data was recorded. Each grid point is at quarter degree representing an area almost equal to 27.75 X 27.75 km^2.

Rainfall: The quantitative variable is measured in millimeters, representing the amount of rainfall recorded at each grid point each year.

3. Objectives:

1. To calculate and analyze key measures of central tendency (mean, median) and dispersion (quantiles, interquartile range, standard deviation, variance) for the rainfall data across the specified grid points.

2. To compute and examine the yearly, monthly, seasonal, and daily variations in rainfall data for each district, providing insights into the temporal patterns and trends over the 71 years

3. To create time series visualizations for each district, illustrating the fluctuation of rainfall over the study period.

4. To investigate and establish the relationship between the district with the maximum rainfall and the district with the second maximum rainfall

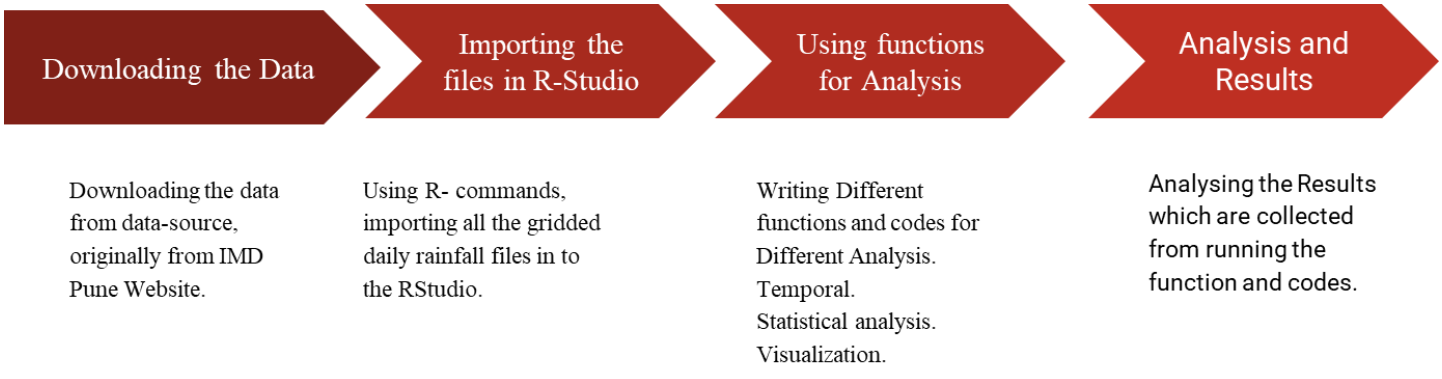
5. To generate autocorrelation plots for each district, revealing the degree of correlation between rainfall measurements at different time lags.

6. PCA: Principal component analysis.

7. Goodness of fit

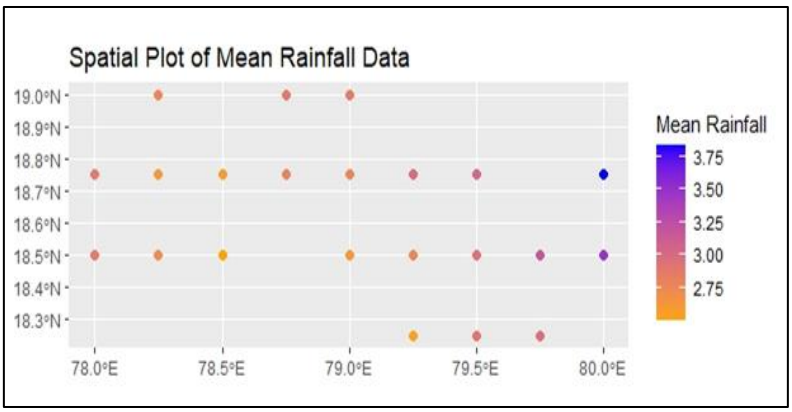
8. Hypothesis Testing (T-Test)

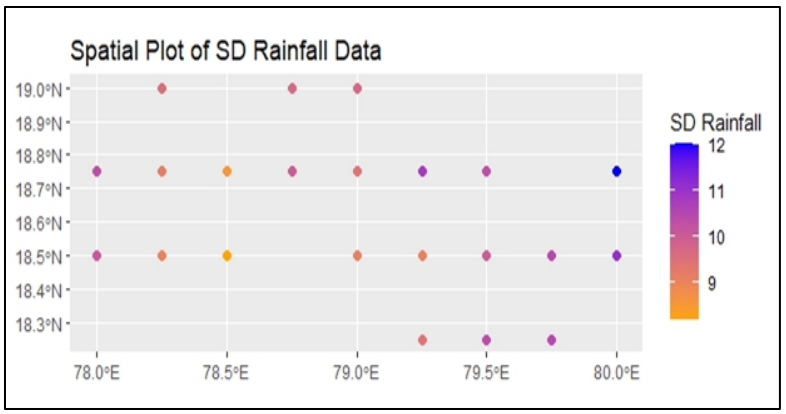
4. Data Preparation

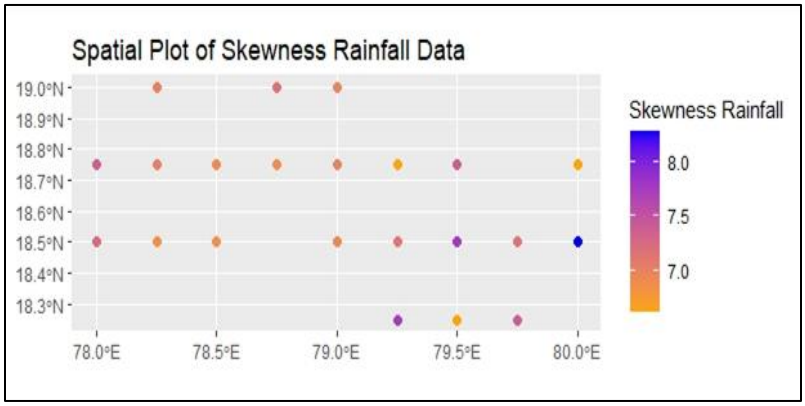


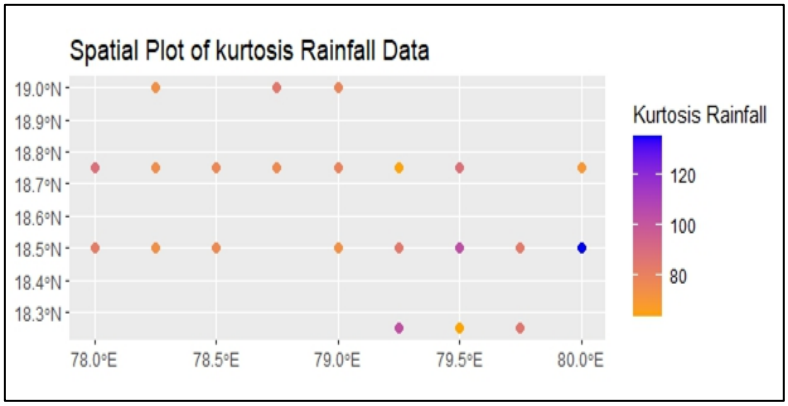
5. Exploratory Data Analysis (EDA)

A. Central Tendency and spread measures for all 5 districts:









- Mean Rainfall Plot:

The plot images shows the average rainfall across the grid points. The grid points depicted by blue dots receive higher average rainfall, whereas those with orange and yellow dots receive lower average rainfall. The statistical analysis reveals a significant correlation between geographical locations oriented in the north-east direction and higher levels of precipitation, suggesting that areas positioned in this cardinal direction are more likely to experience increased rainfall.

- Standard Deviation Plot:

The SD plot illustrates the variability or dispersion of rainfall data around the mean. Points with higher SD values suggest more variability in rainfall, indicating that some grid points experience a greater inconsistency in rainfall amounts. From the image, it appears that there are a few grid points with notably higher variability, as indicated by the blue points.

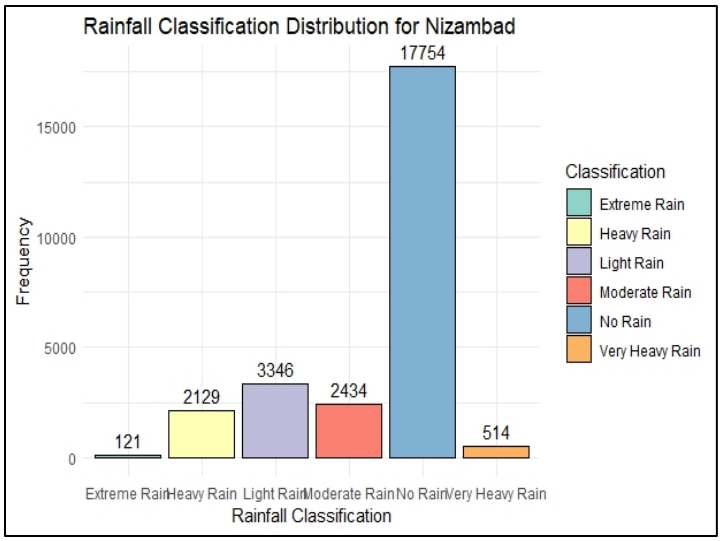
- Skewness Plot:

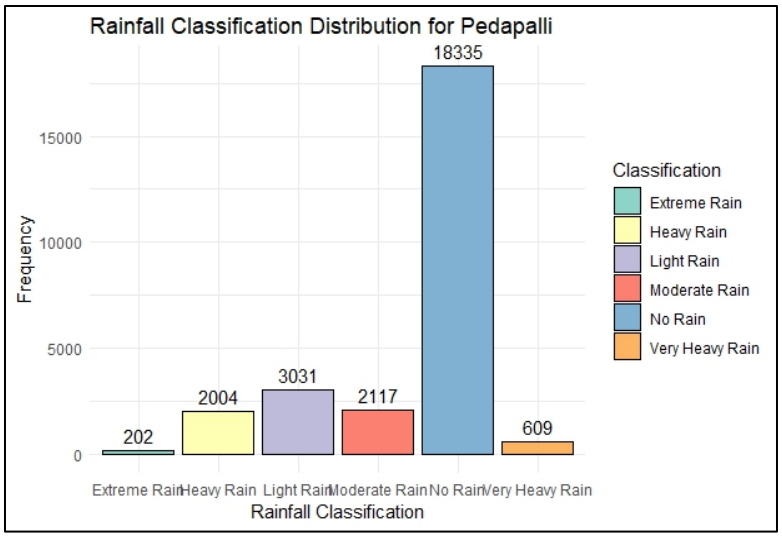
Skewness indicates the asymmetry of the rainfall distribution. A positive skewness value, which seems to be the case across all grid points, suggests that the rainfall distribution is right-skewed, meaning there are a few days with exceptionally high rainfall compared to the rest.

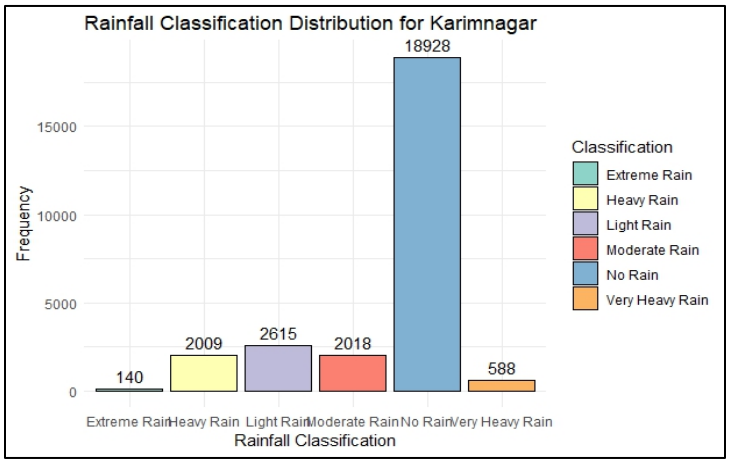
- Kurtosis Plot:

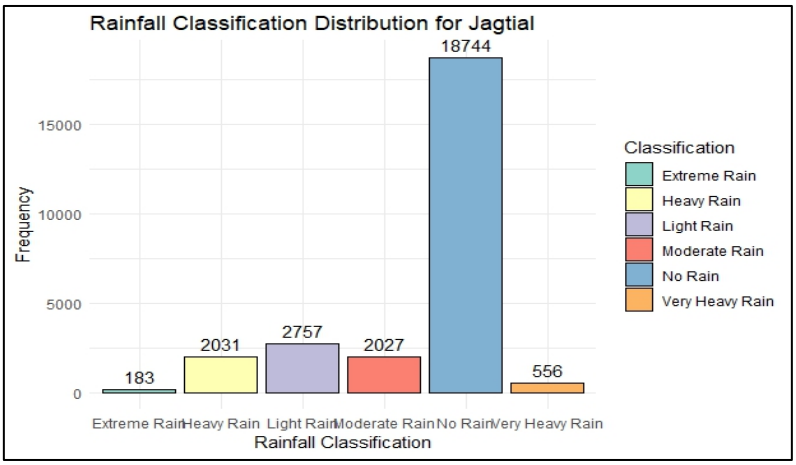
Kurtosis measures the 'tailedness' of the rainfall distribution. High kurtosis suggests that there are more frequent extreme rainfall events than would be expected in a normal distribution. The blue points indicate grid points with very high kurtosis, suggesting the presence of extreme values in the rainfall data, such as heavy rainfall days.

B. Rainfall Classification Distribution :



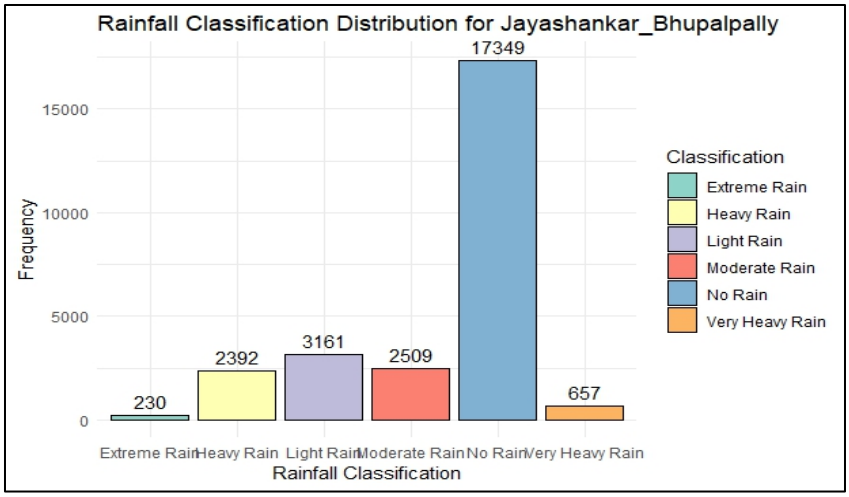






The bar charts provided represent the frequency of different rainfall classifications across five districts: Nizambad, Pedapalli, Jayashankar_Bhupalpally, Karimnagar, and Jagtial.

Each bar chart classifies rainfall into six categories: No Rain, Light Rain, Moderate Rain, Heavy Rain, Very Heavy Rain, and Extreme Rain.



Across all districts, the most common rainfall classification is 'No Rain', indicating a significant number of dry days.

'Extreme Rain' events are the least common in all districts, suggesting that such events are rare but may have significant implications when they do occur.

There is a noticeable frequency of 'Very Heavy Rain' across the districts, which is important for water resource management and flood risk assessment.

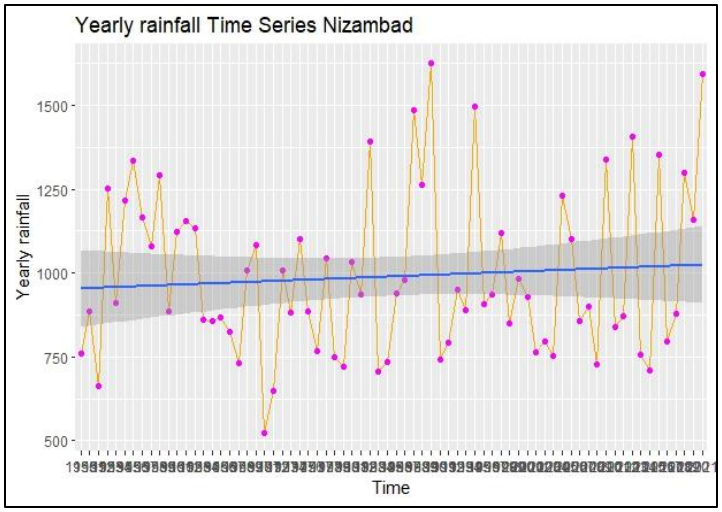
The frequency of 'Light Rain' and 'Moderate Rain' suggests that while there are many dry days, the regions also experience a fair amount of rainfall that may be beneficial for agriculture and replenishing water resources.

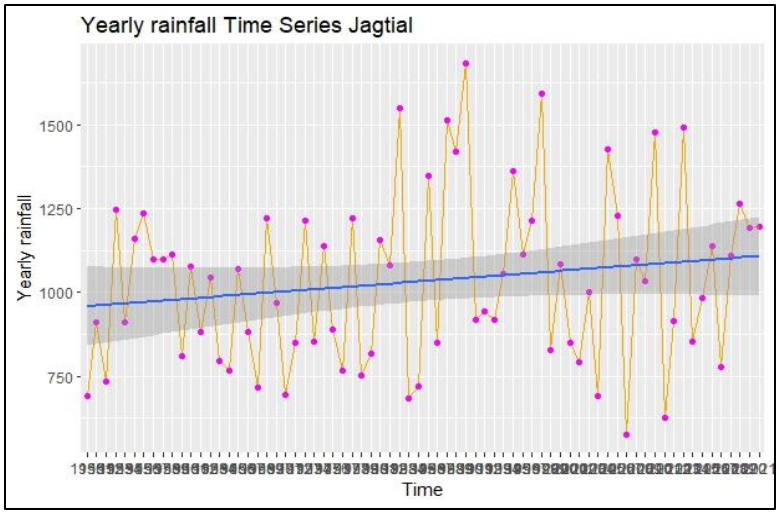
C. Data Preparation

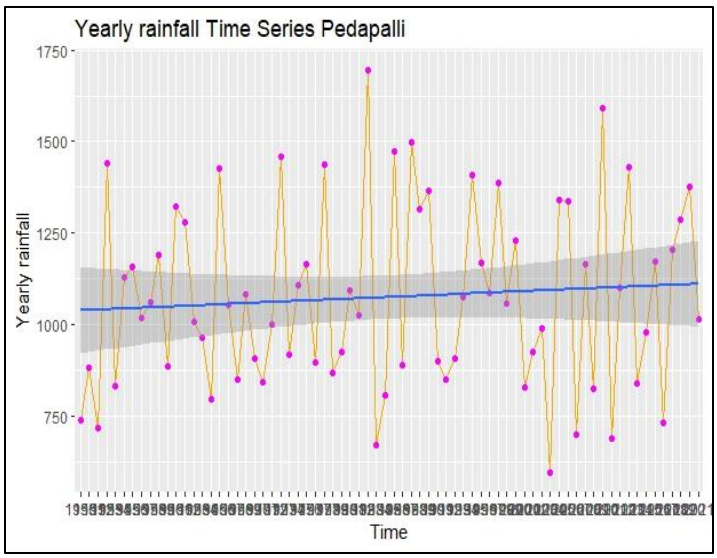
The provided script is an R function named GridSum designed to compute the average rainfall across multiple points within a district. This function takes a list of CSV files as input each representing different points within a district and aggregates their data into a single data frame, effectively consolidating the rainfall data of the district.

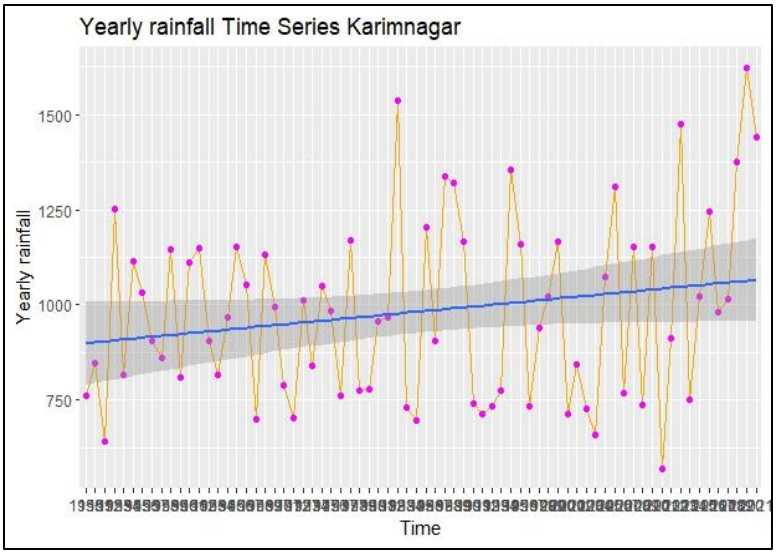
```r
#to find average of all points of a district
GridSum<- function(csv_files){
  totdata<- data.frame(DateTime= character(),rainfall= numeric() )

  for (csv_file in csv_files) {
    data <- read.csv(csv_file)
    totdata<- merge(totdata, data, by = "DateTime", all = TRUE)
  }

  totdata$rainfall<- rowSums(totdata[3:ncol(totdata)])
  totdata$rainfall<- totdata$rainfall/length(csv_files)
  return(totdata[,1:2])
}
```

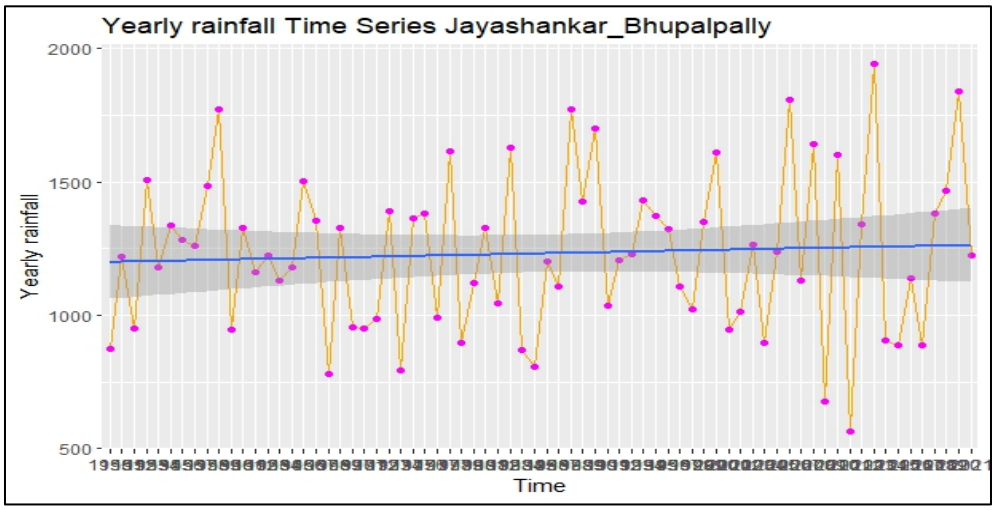
D. Yearly Time Series for Each District:











Each graph shows considerable year-to-year variability in rainfall, as indicated by the individual data points. There are years with particularly high rainfall, shown by peaks in the graph, and years with lower rainfall, shown by troughs.

The trend line (blue line), likely representing a moving average, provides a visual representation of the long-term trend in the data. In these graphs, the trend line seems to indicate a slight upward trend in annual rainfall over the period for most districts.

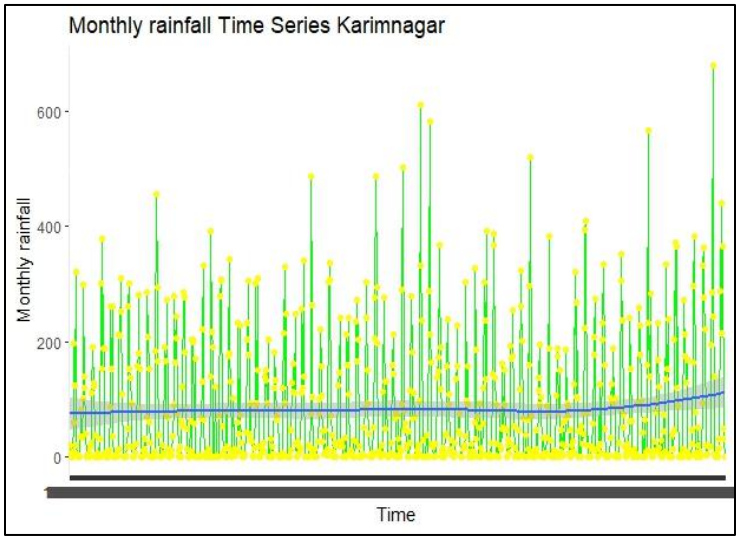
High Rainfall Years:
The points that significantly deviate above the trend line correspond to years with unusually high rainfall. These could be associated with specific climatic events or patterns, such as strong monsoon seasons or cyclones.

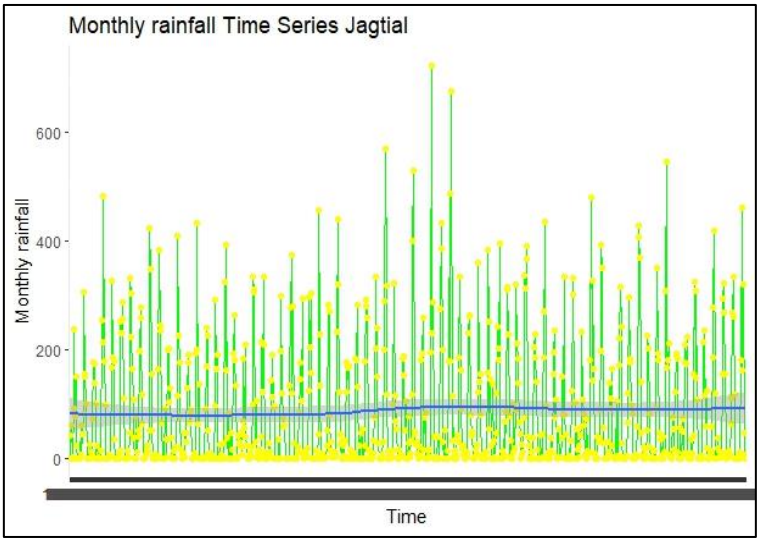
Low Rainfall Years:
Conversely, points below the trend line highlight years with less rainfall than average, which could have implications for water scarcity and agricultural output in those years.

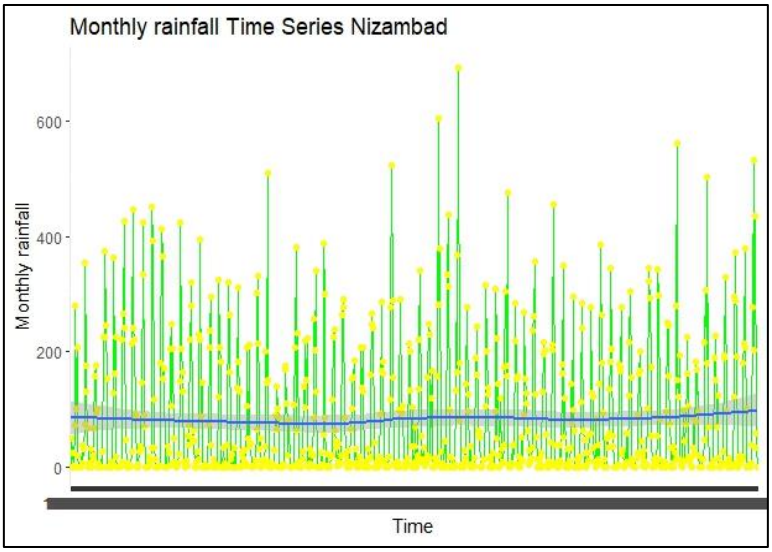
Consistency Across Districts:
While each district has its unique rainfall profile, there may be similarities in certain periods, suggesting regional climatic influences affecting rainfall patterns across multiple districts.

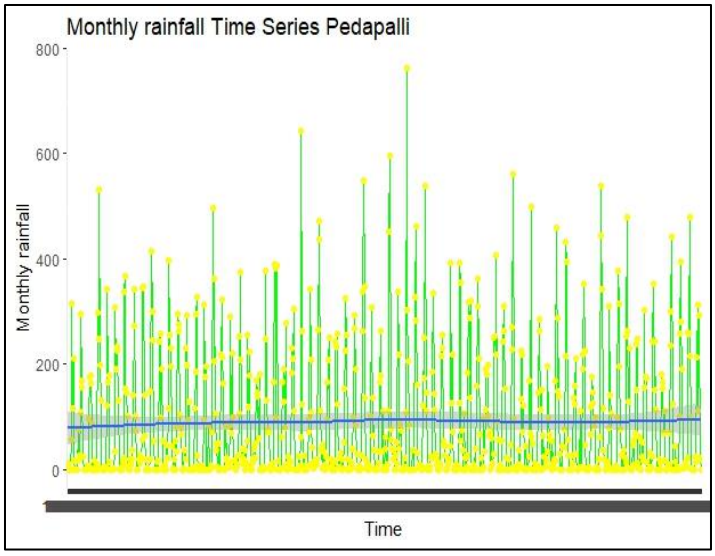
The shaded area around the trend line represent the confidence interval or variance, providing a visual guide to the reliability of the trend.

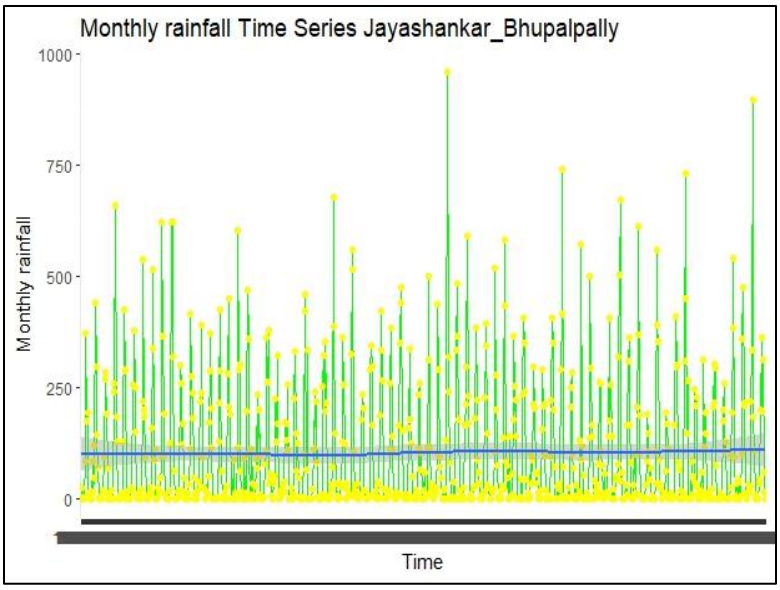
E. Monthly time series data for each district:



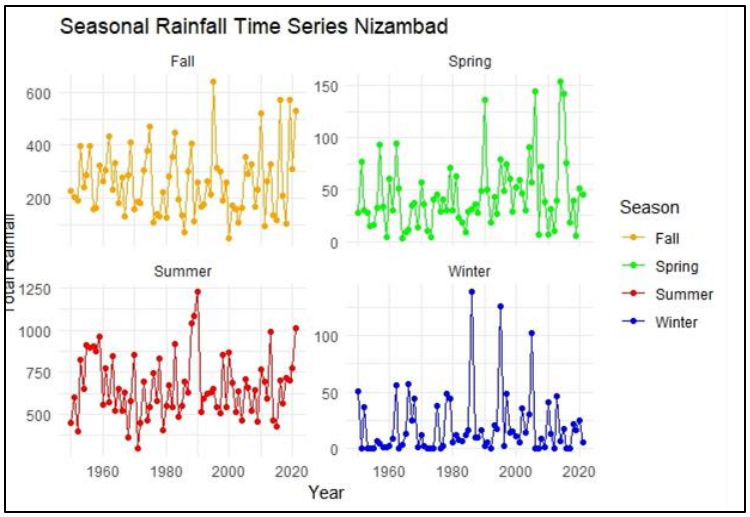


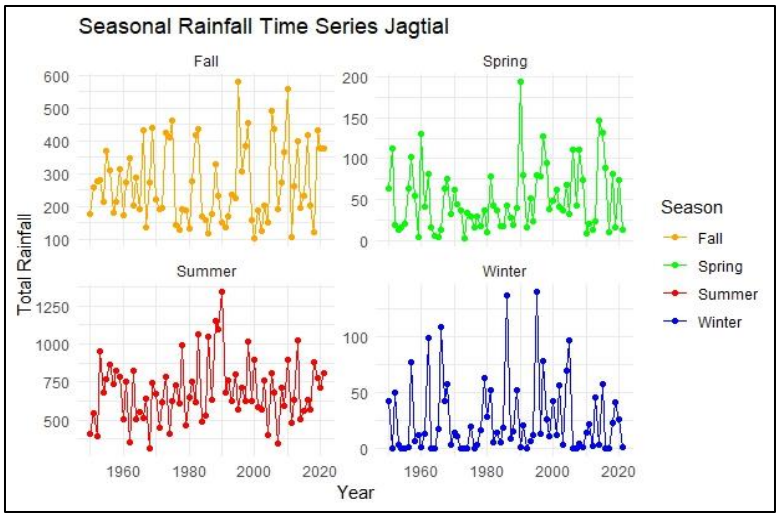

Monthly rainfall Time Series Nizambad
Monthly rainfall
600
400
200
0
Time

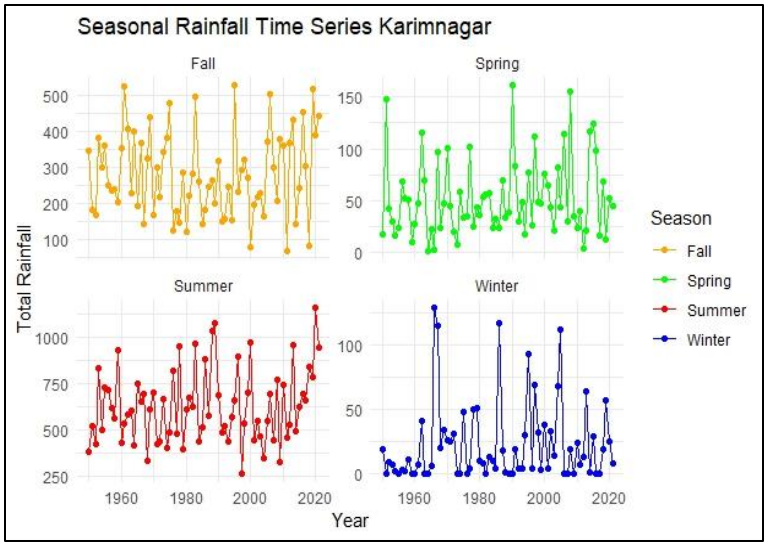

Monthly rainfall Time Series Pedapalli
800
600
400
200
0
Monthly rainfall
Time

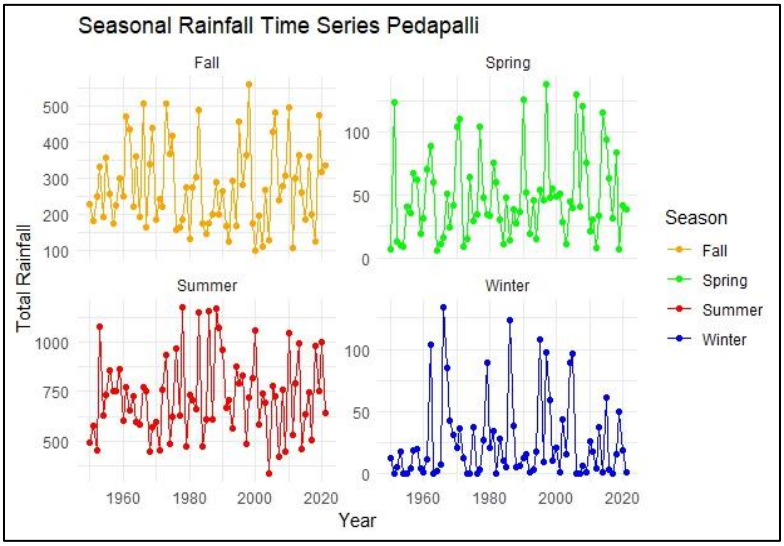

Monthly rainfall Time Series Jayashankar_Bhupalpally
1000
750
500
250
0
Monthly rainfall
Time

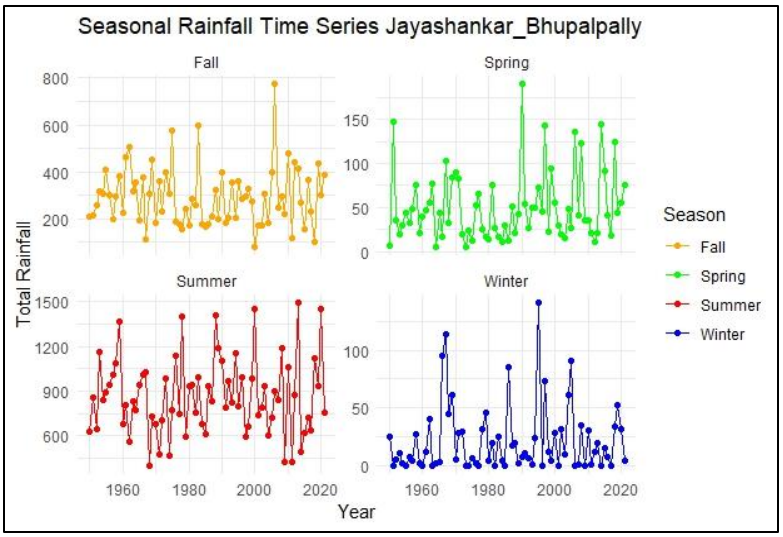
F. Seasonal Time series Data:





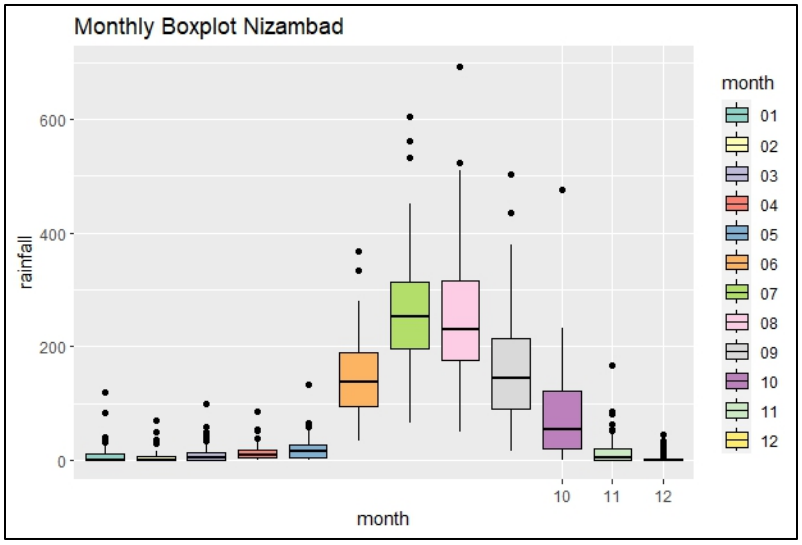


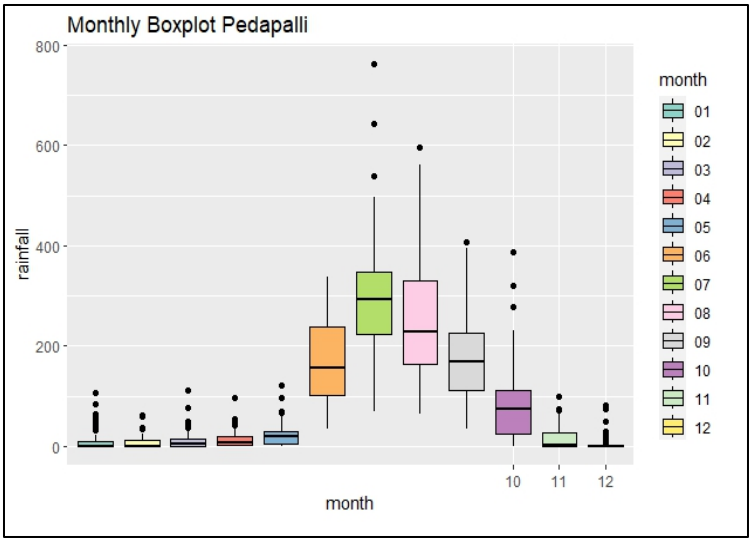


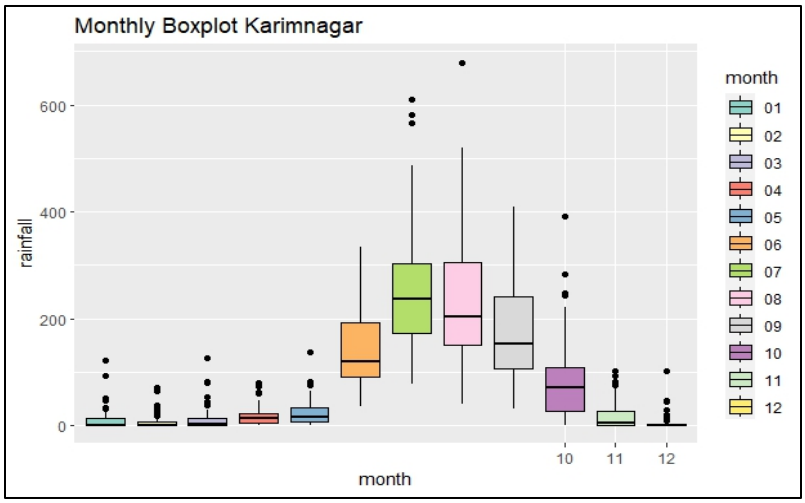


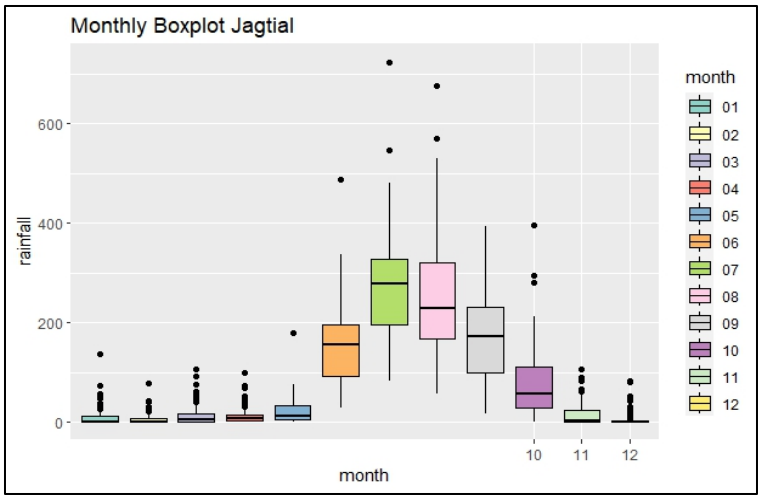
There is substantial variability in rainfall from year to year within each season, indicating that some years have more intense seasonal rainfall than others. The variability is most pronounced in the summer and fall, which is consistent with the variability of the monsoon. The charts show some years where the rainfall during the monsoon season (summer) is notably low, which could have implications for drought conditions and water scarcity. While all districts show a similar seasonal pattern, there are differences in the intensity and variability of rainfall across the districts.

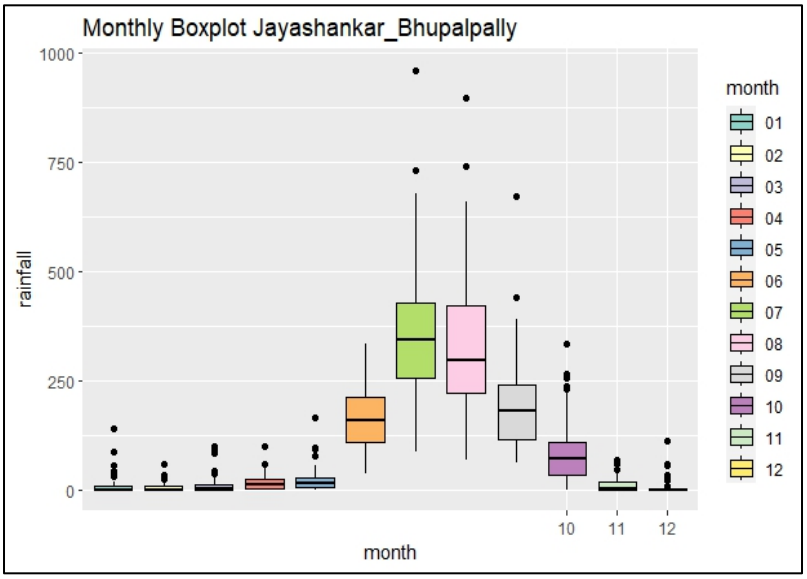
G. Monthly Boxplots:









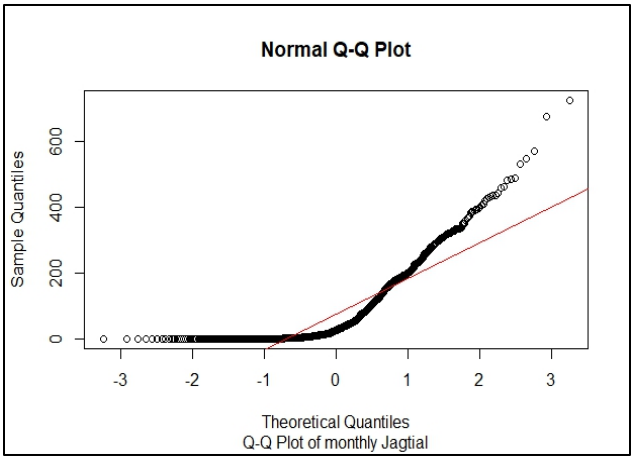


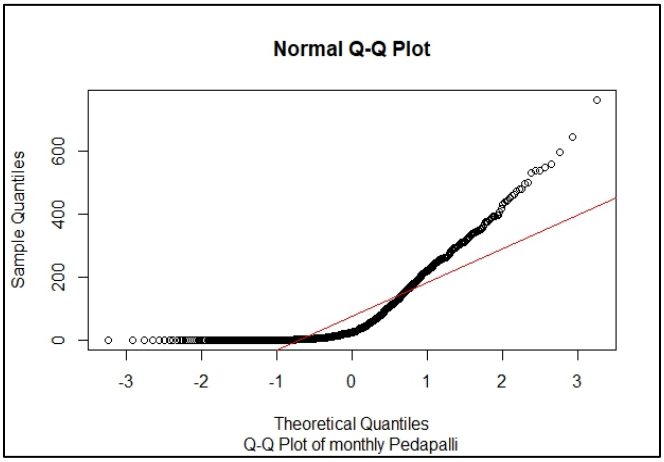
The boxplots provided represent the distribution of monthly rainfall data averaged over the period from 1950 to 2021 for five districts: Nizamabad, Jayashankar_Bhupalpally, Pedapalli, Karimnagar, and Jagtial. For all districts, the monsoon months (typically June to September) show higher medians and larger IQRs, indicating not only more rainfall but also greater variability during the monsoon season.

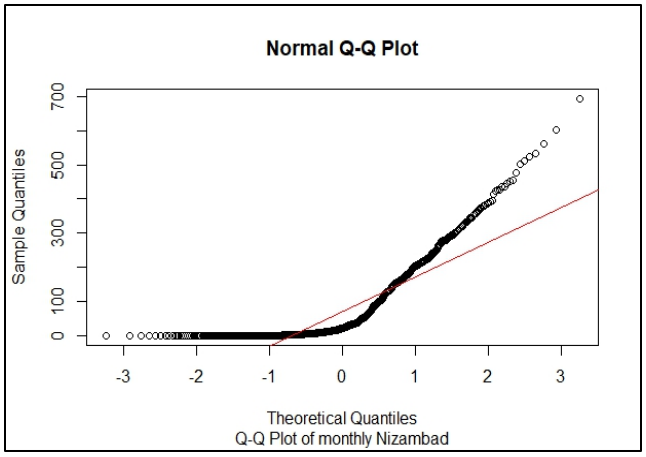
The non-monsoon months show lower medians and smaller IQRs, reflecting less rainfall and less variability. The points above or below the whiskers represent outliers, which indicate months with unusually high or low rainfall for the season. The presence of outliers is particularly noticeable during the monsoon months, suggesting that there are occasional extreme rainfall events.

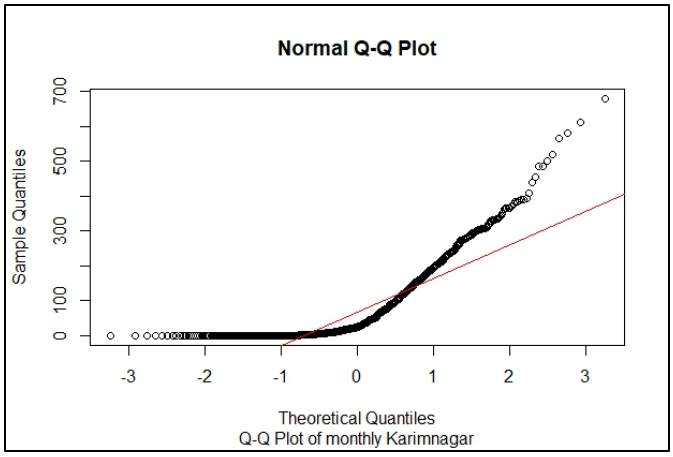
Comparing the boxplots across the districts, there may be slight variations in the timing and intensity of the monsoon, as suggested by differences in the heights of the boxes and positions of the medians. Some districts may experience an earlier or later onset of the monsoon, which could be reflected in the boxplots. Months with longer whiskers have a wider range of rainfall amounts, showing more inconsistency in the amount of rainfall received.

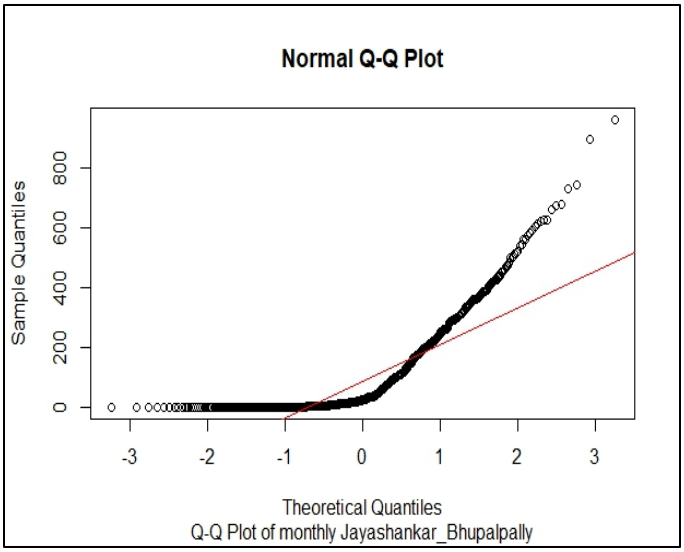
H. Monthly Quantile Plots:





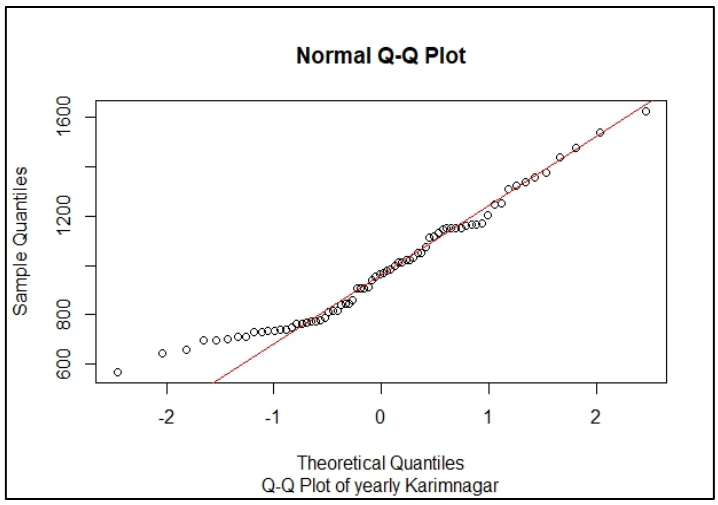


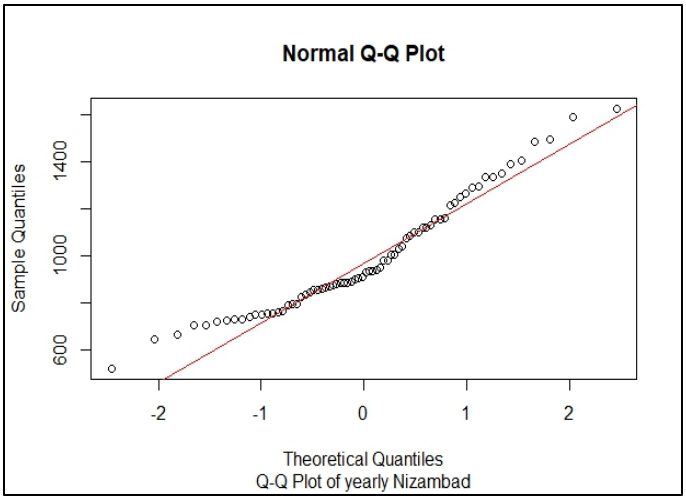


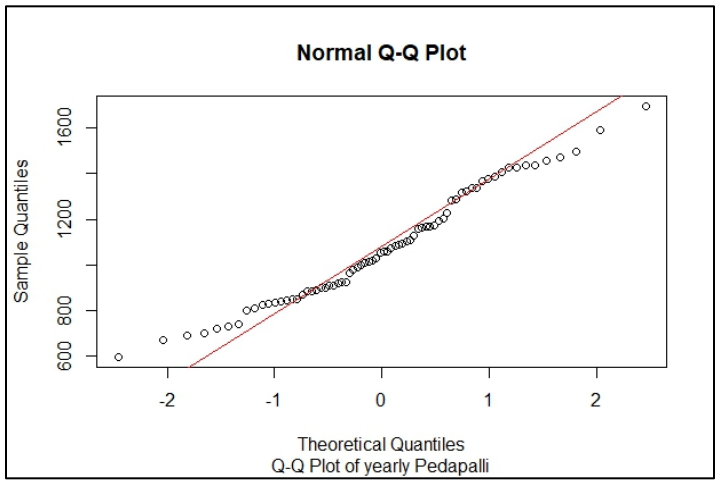


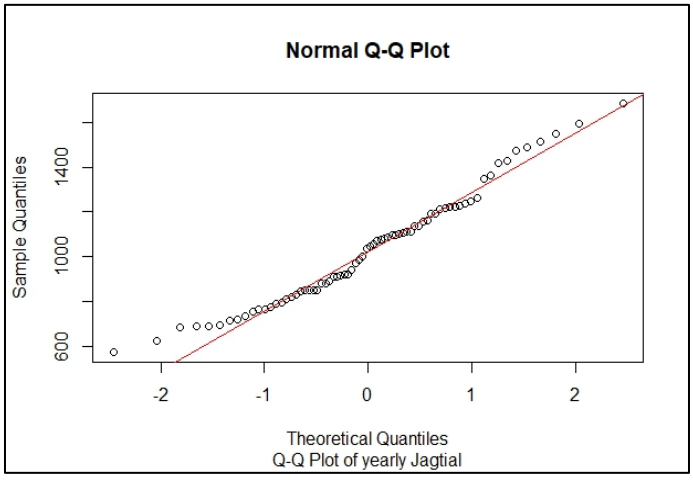
The upward curve at the right end of the plots suggests that the rainfall data have a heavy right tail. This means there are more extreme values, or higher rainfall months, than what would be expected in a normal distribution. The low-lying points at the left end of the plots indicate that there are also more months with very low rainfall than would be expected if the data were normally distributed. In the central part of the plots, where the data points are closer to the line, the rainfall data are more in line with what would be expected from a normal distribution. In these plots, outliers are evident at both ends, suggesting the presence of both unusually wet and unusually dry months.

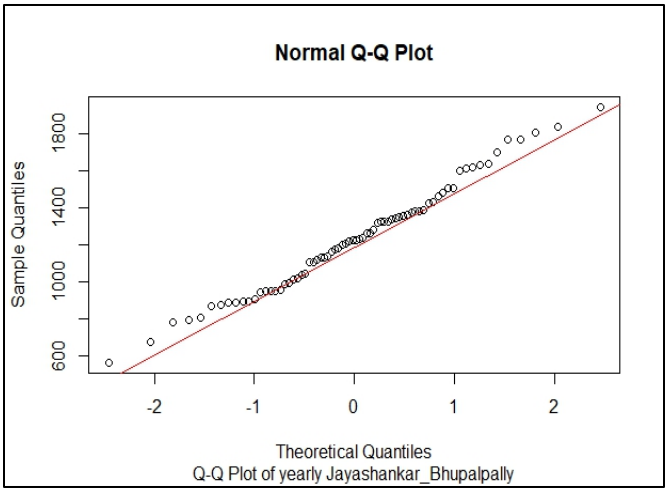
I. Yearly Quantile Plots:





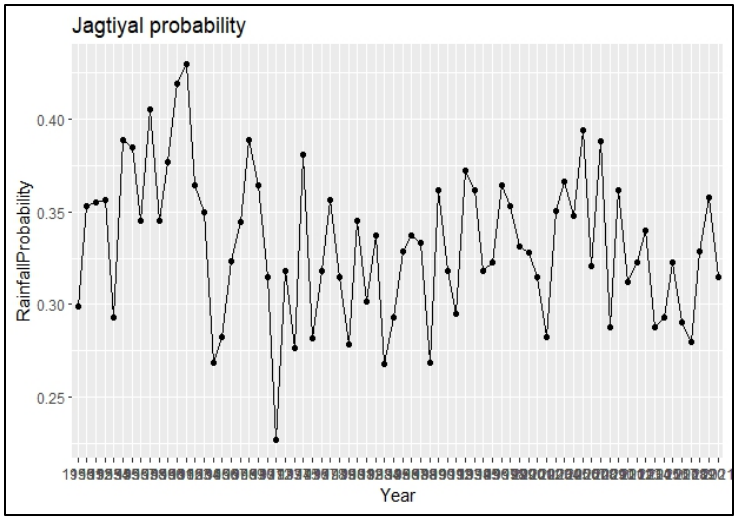


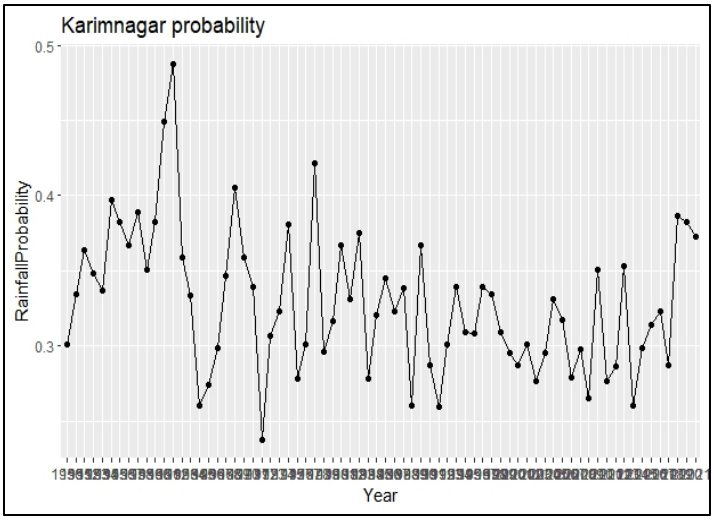


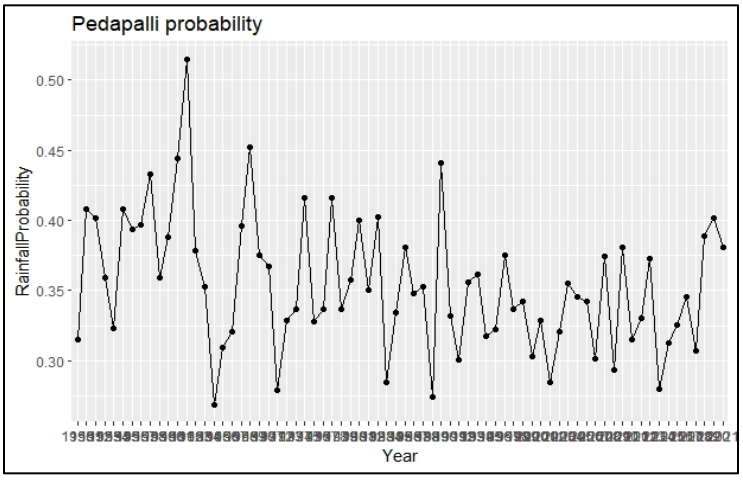


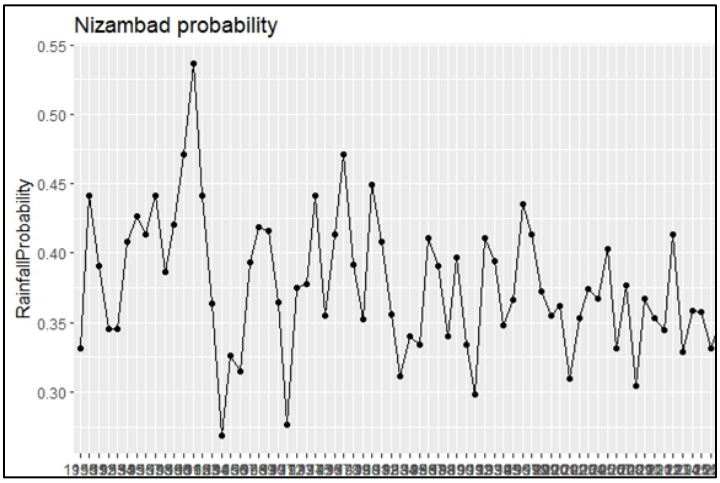
The plots for the five districts show that the points lie close to the line in the middle portion of the data, indicating that the central part of the distribution is close to normal. In the provided plots, there is some deviation at the ends, but it is not as pronounced as in the monthly Q-Q plots, indicating that the yearly data are closer to a normal distribution than the monthly data. Compared to the monthly Q-Q plots, the yearly Q-Q plots show a closer adherence to normality.

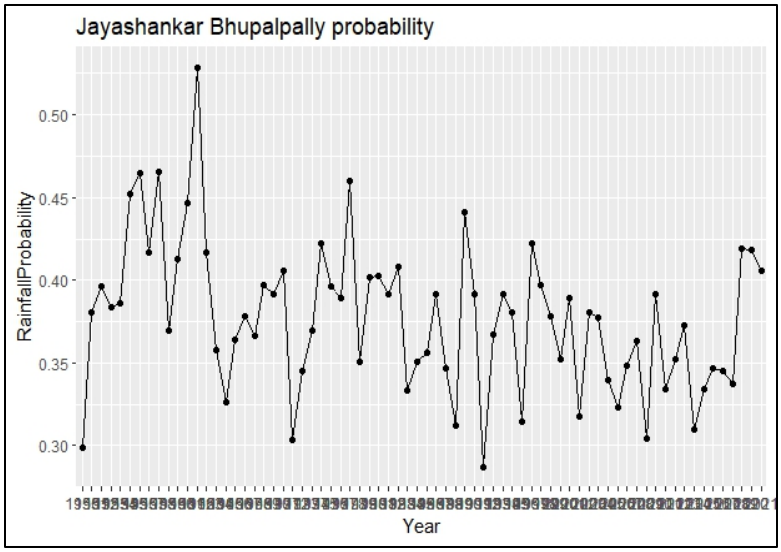
J. Probability of Rainfall:







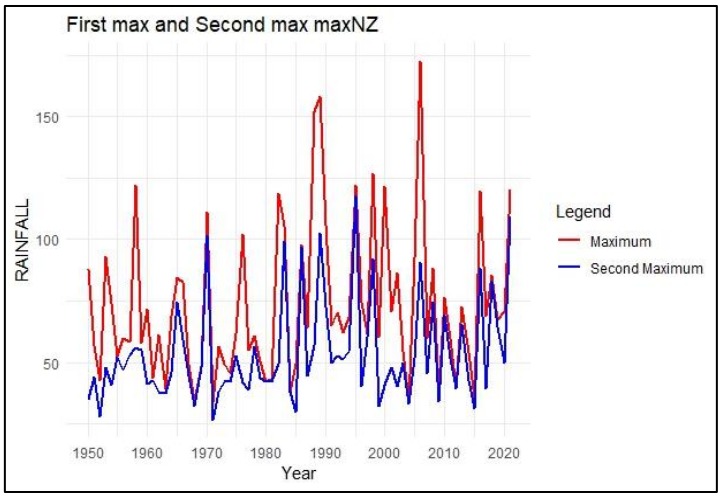




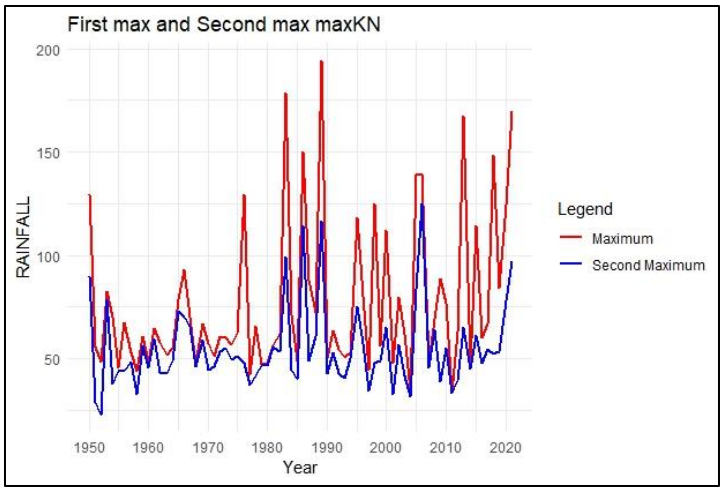
The provided images are plots showing the probability of rainfall for five districts over a series of years. The probability values fluctuate over the years, indicating variability in the likelihood of rainfall.

There doesn't appear to be a clear upward or downward trend, suggesting that the probability of rainfall has not consistently increased or decreased over the period displayed. Sharp peaks and troughs indicate years with significantly higher or lower probabilities of rainfall than surrounding years.

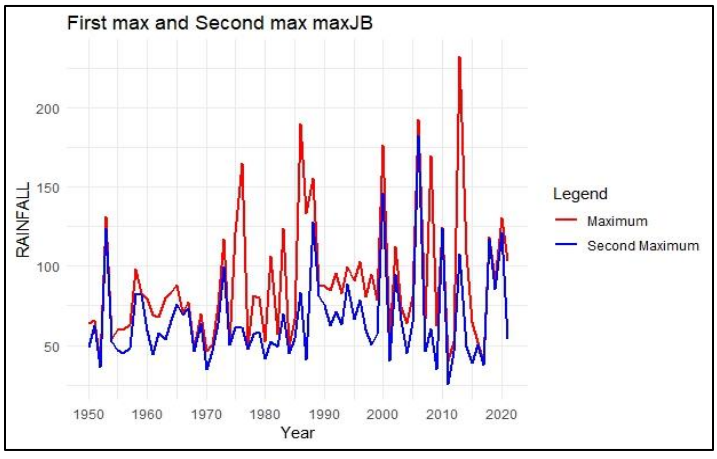
K. Relation between Annual Maximum and Second Maximum:



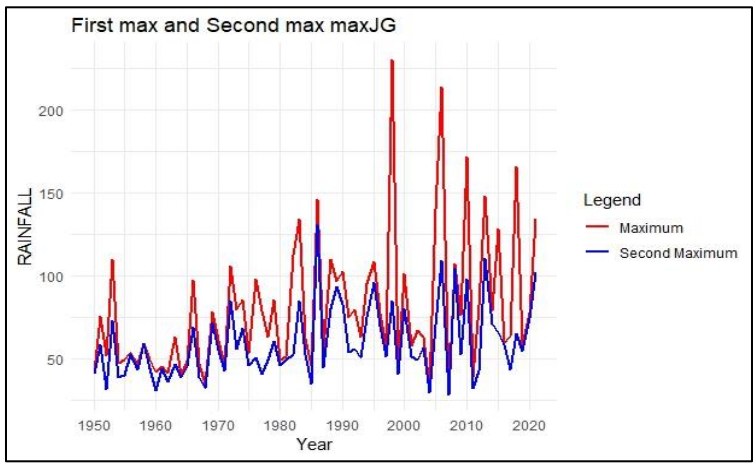
Correlation:0.7118



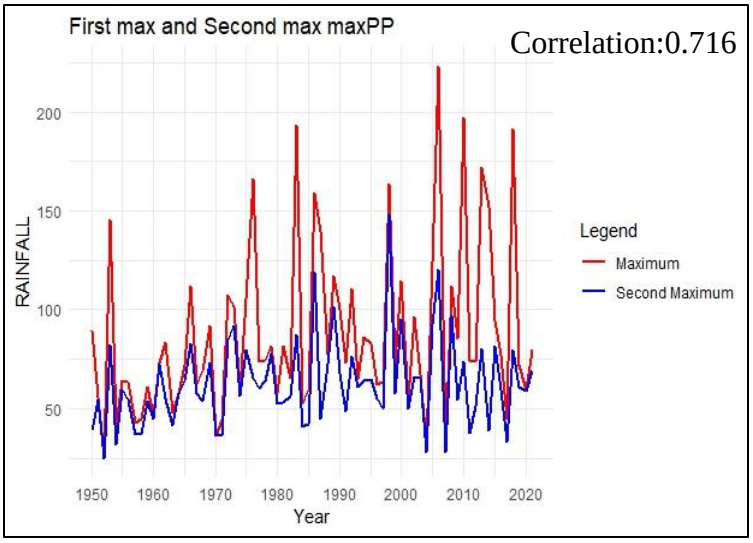
Correlation:0.7940



Correlation:0.7290

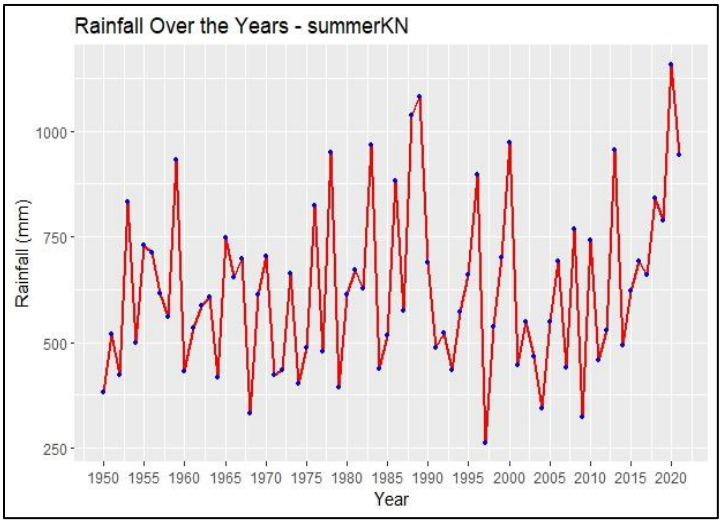


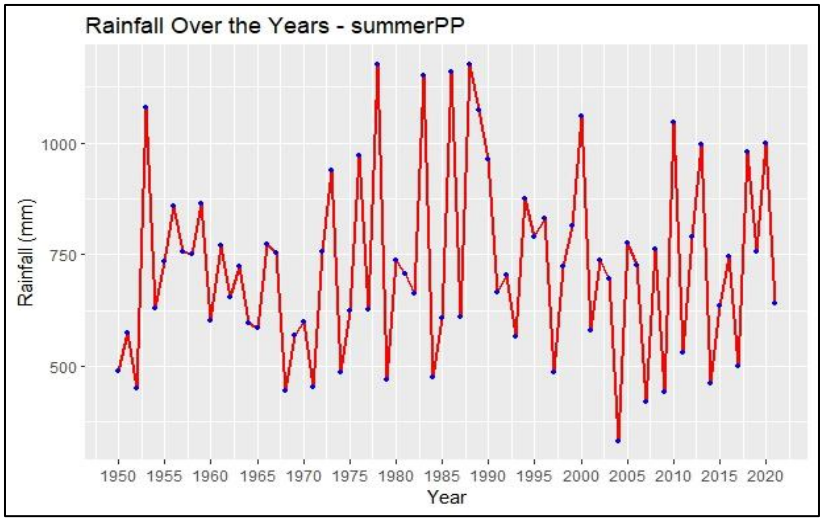
Correlation:0.801



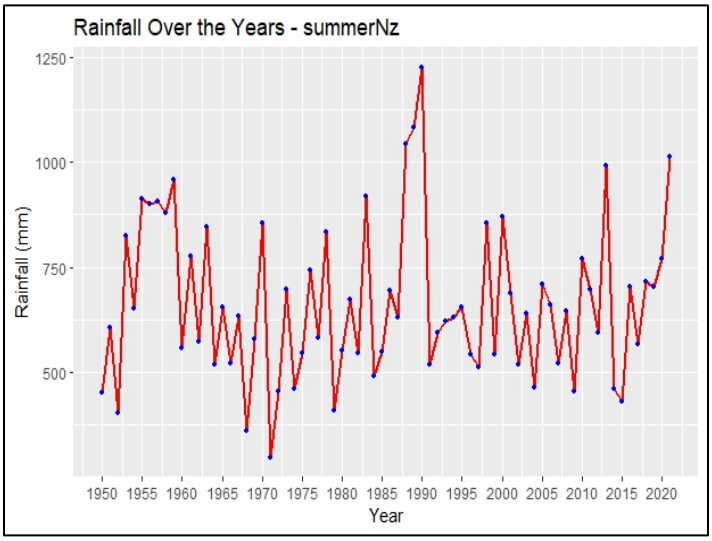
Annual Maximum Rainfall: The red lines represent the highest annual rainfall recorded in each year for the respective districts. Second Maximum Rainfall: The blue lines indicate the second-highest annual rainfall amounts, providing a sense of the variability and frequency of heavy rainfall events. The distance between the maximum and second maximum lines within each year could indicate the consistency of rainfall events; closer lines might suggest a more even distribution of heavy rainfall events throughout the year. The height of the peaks in both maximum and second maximum rainfall indicates the extremity of rainfall events. A consistent pattern between the first and second maximum across years could suggest that the district experiences regular extreme rainfall events. Wide gaps between the first and second maximum might indicate that the highest rainfall event of the year is an outlier, significantly surpassing other heavy rainfall events.

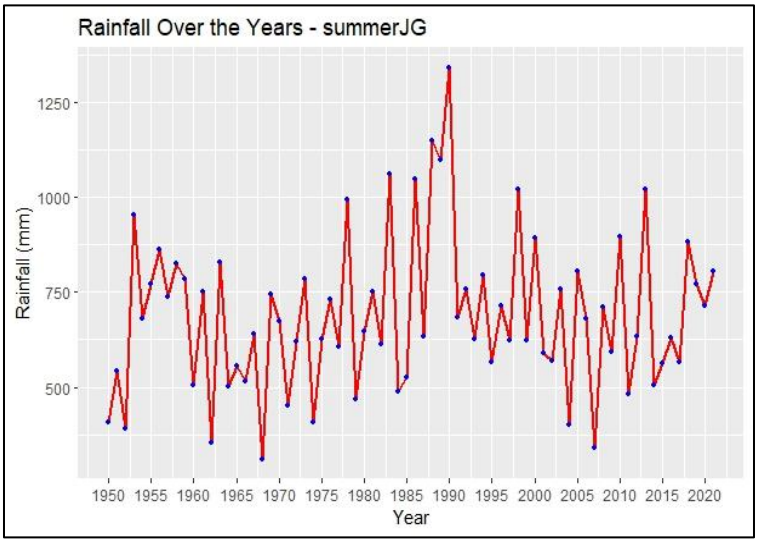
L. Summer/Monsoon Analysis – Climatology

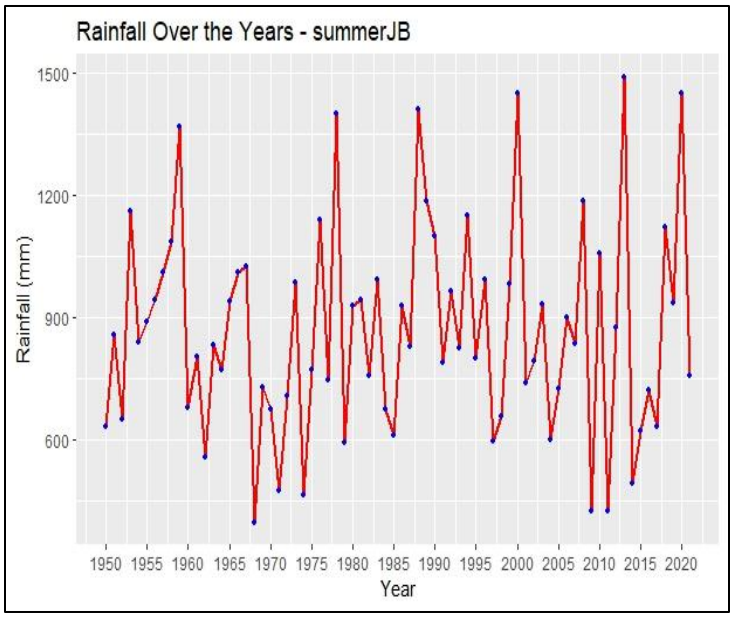




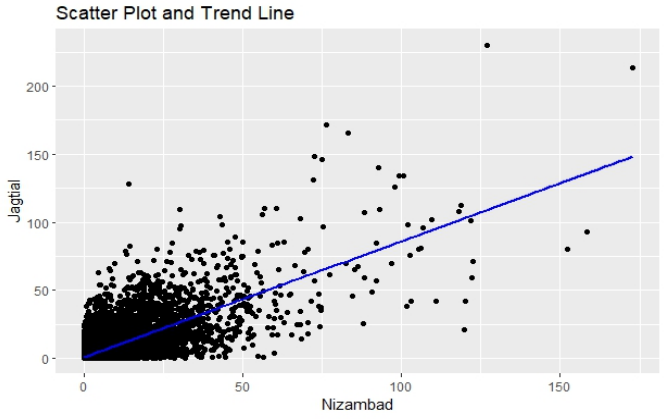
	NAME	MEAN	MEDIAN	MODE	STANDARD_DEVIATION	SKEWNESS	KURTOSIS	IQR	MAD
1	summerNz	664.9108	636.9756	297.0467	186.3525	0.6515817	3.067776	235.9392	117.2293
2	summerJG	688.8124	660.4882	310.0550	204.3679	0.6430108	3.445192	229.2535	121.5663
3	summerJB	866.6029	834.0426	395.9115	257.8416	0.5028804	2.905421	318.0862	157.7124
4	summerKN	627.4847	609.9003	261.3746	199.5527	0.5895038	2.685698	257.9586	131.4596
5	summerPP	724.3203	723.3946	331.5494	204.6294	0.4865015	2.555993	235.9202	131.1058

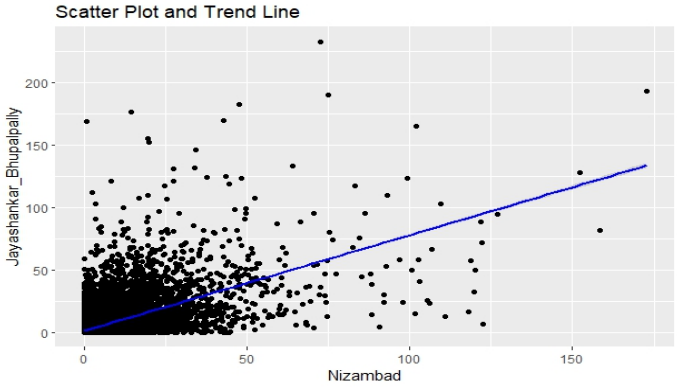


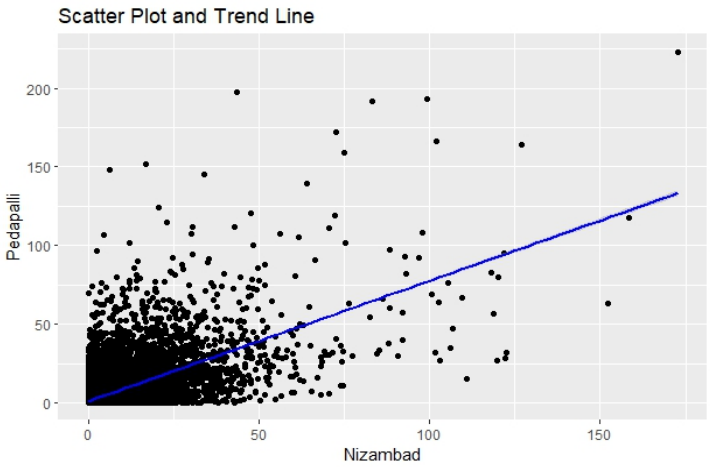


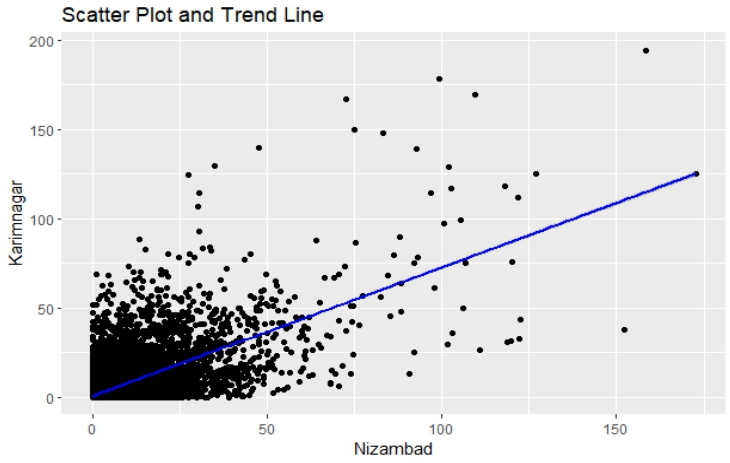


M. Scatter Plot with trend line:



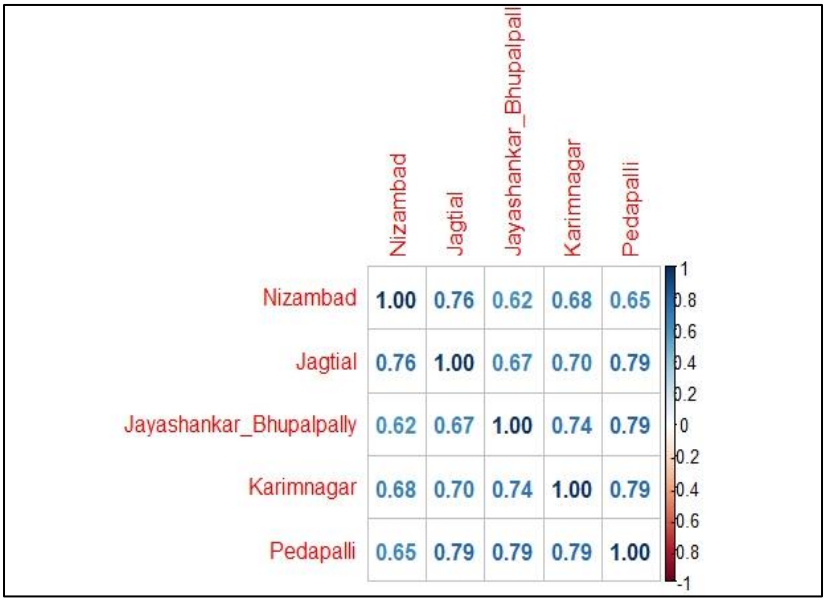




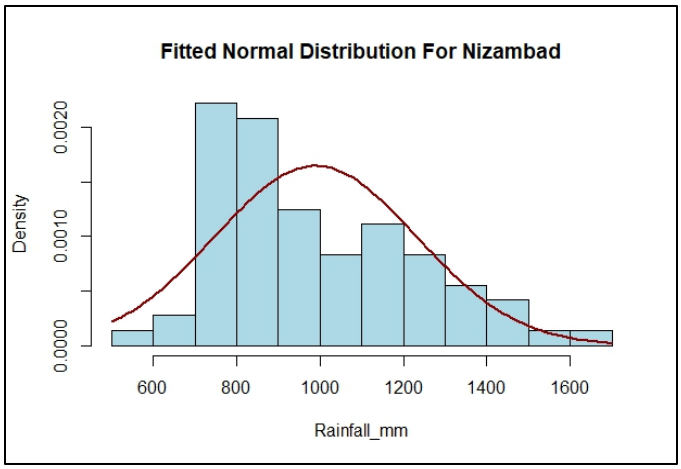


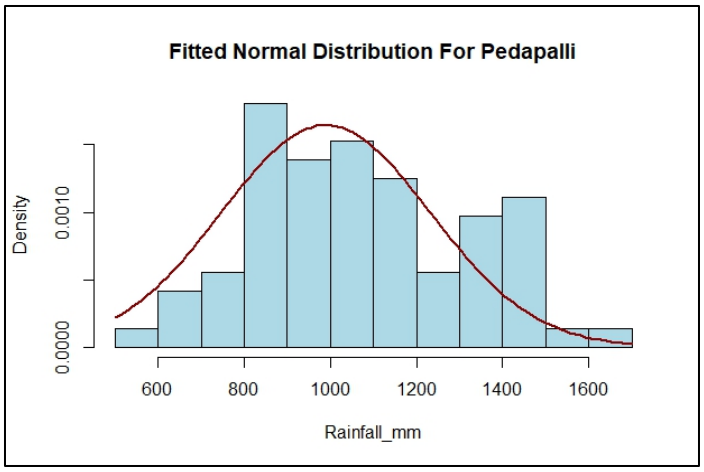
N. Correlation between the districts:

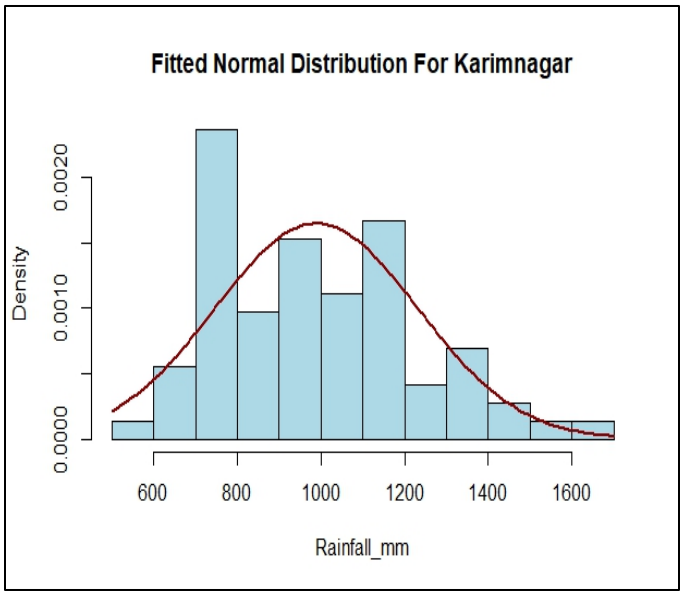
The correlation matrix shows the pairwise correlation coefficients between the rainfall data of different districts. These coefficients range from -1 to 1, where: 1 indicates a perfect positive correlation, meaning rainfall in one district increases or decreases in lockstep with rainfall in another. 0 indicates no correlation, meaning there is no linear relationship between rainfall in the districts. -1 indicates a perfect negative correlation, meaning rainfall in one district increases as rainfall in another decreases, and vice versa. Nizamabad has moderately strong positive correlations with all other districts, the strongest being with Jagtial (0.76) and the weakest with Pedapalli (0.65). Jagtial shows a similar pattern, with strong correlations with all districts, especially Pedapalli (0.79). Jayashankar_Bhupalpally has somewhat weaker correlations with the other districts, suggesting its rainfall pattern is less synchronized with the others, potentially due to localized climatic influences. Karimnagar and Pedapalli have high correlations with each other (0.79) and with other districts, suggesting a strong regional similarity in rainfall patterns. Overall, the matrix indicates a general positive correlation in rainfall patterns across the districts, with no negative correlations, which means that all districts tend to experience increases or decreases in rainfall around the same times.

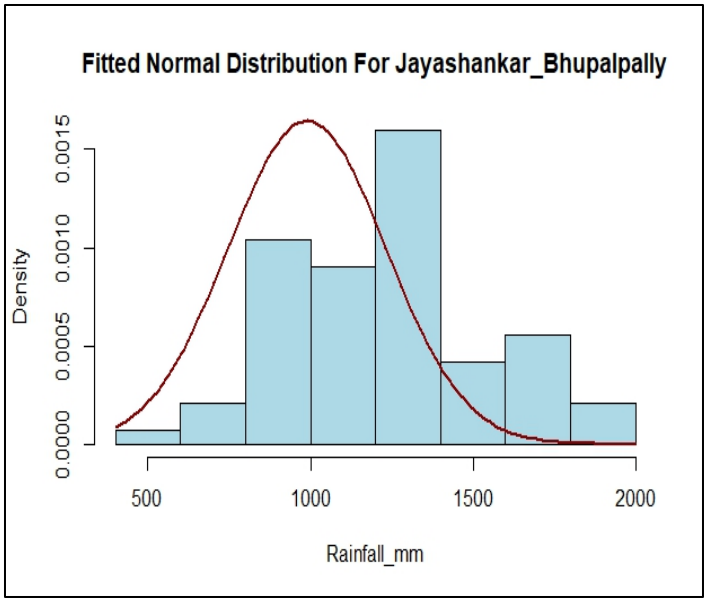


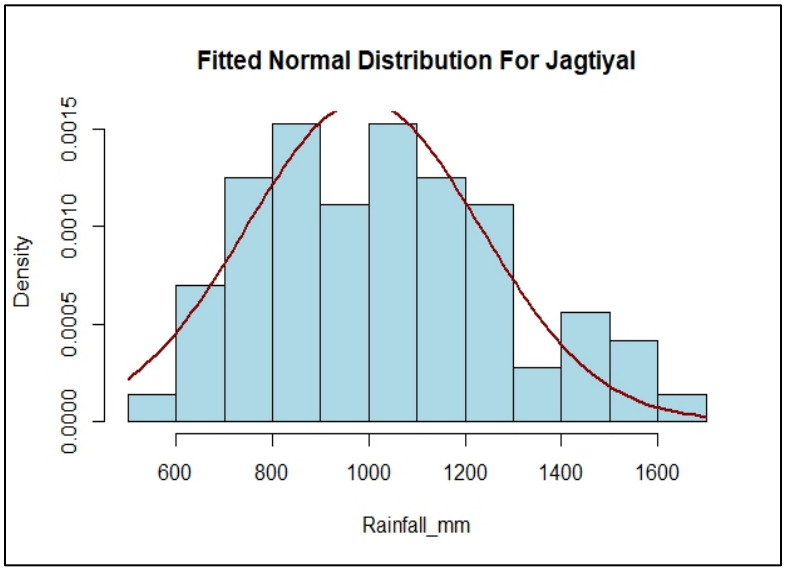
O. Fitted Normal Distribution over yearly data:



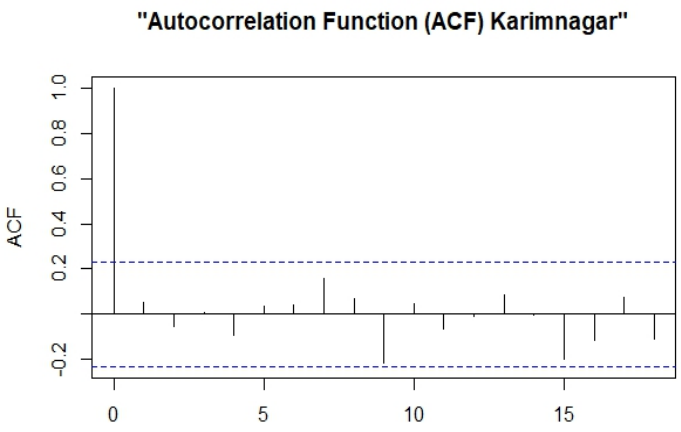


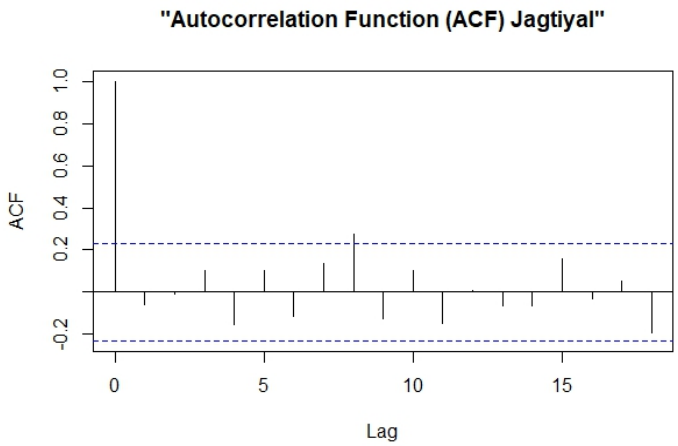


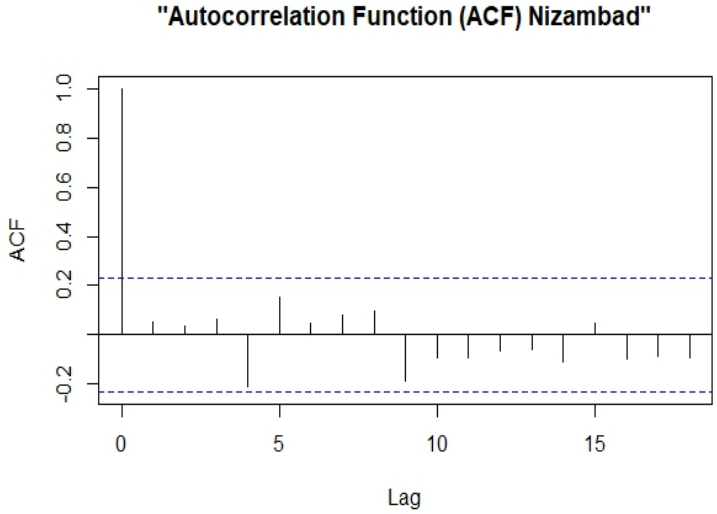


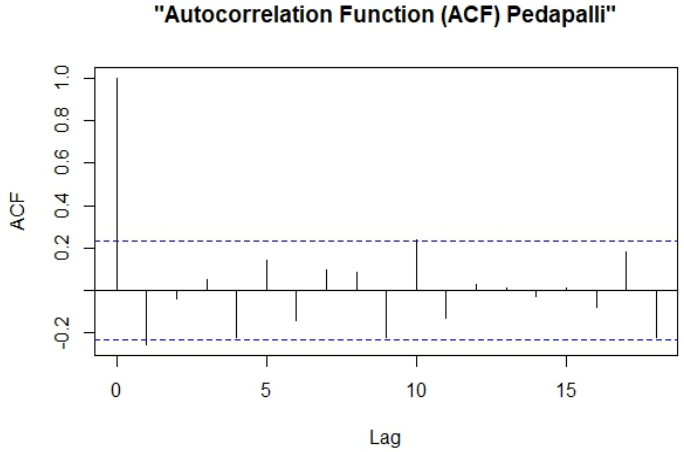


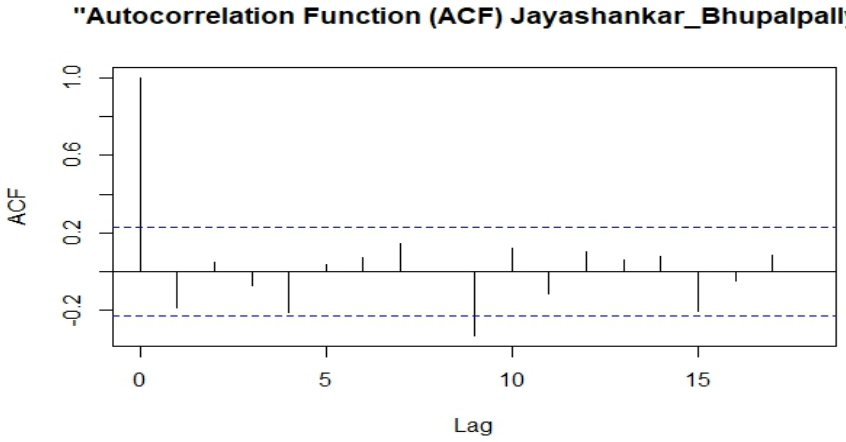
P. Autocorrelation Plots:



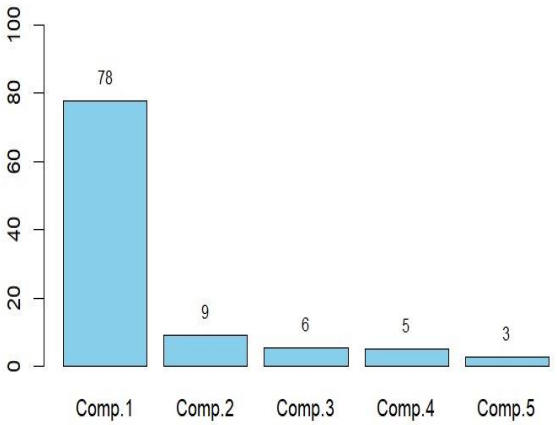




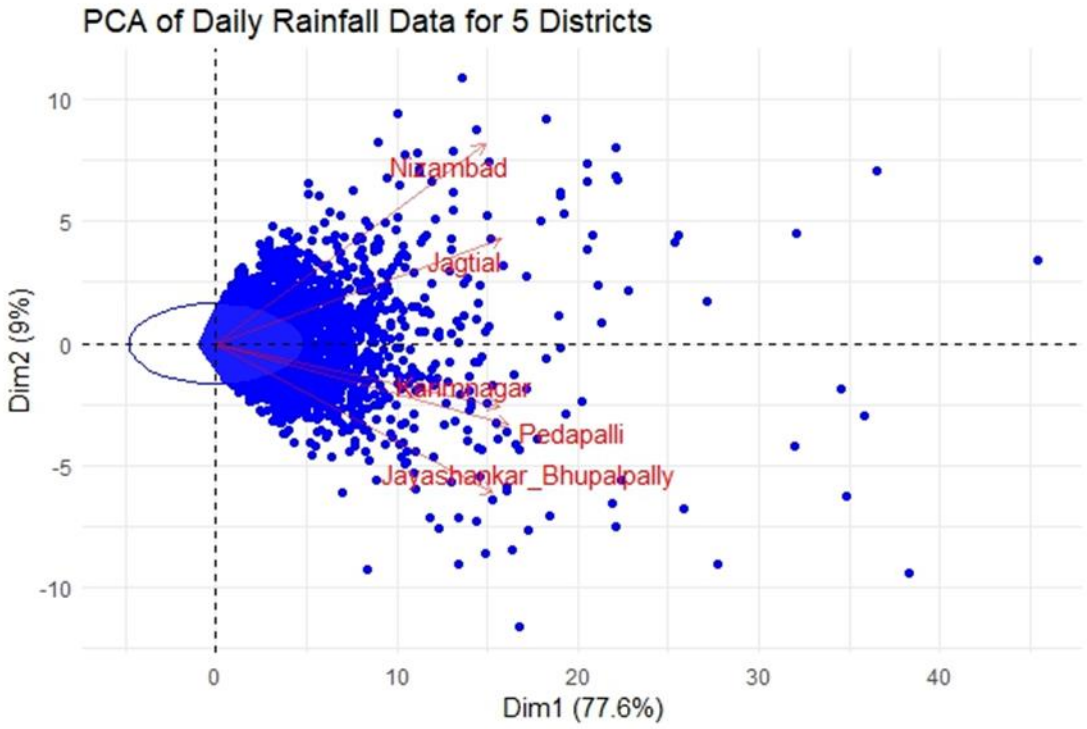




Q. Principal Component Analysis (PCA):




```
#PCA_____________________________________________________________________________
combine_riyal<- as.data.frame(cbind(Date=Nizambad$DateTime, combined_data))
SD1<- scale(combine_riyal[,2:6], center = TRUE, scale=TRUE)
y<- cov.wt(SD1)
R <- princomp(combine_riyal[,2:6], scores = TRUE, covmat = y)
v<- ((R$sdev^2)/5)*100
barplot(v, ylim = c(0, 100), col='skyblue')
text(x = barplot(v, plot = FALSE), y = v + 1, label =round(v), pos = 3, cex = 0.8, col = "black")
```



- T- test (hypothesis testing)

```
t_test_result <- t.test(yearlyNZ$rainfall, yearlyJB$rainfall)
print(t_test_result)

# Welch Two Sample t-test
#
# data:  yearlyNZ$rainfall and yearlyJB$rainfall
# t = -5.3392, df = 136.97, p-value = 3.779e-07
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   -331.2103 -152.1798
# sample estimates:
#   mean of x mean of y
# 989.2699 1230.9650
```

Null Hypothesis= there is no significance difference between the mean
Alternative Hypothesis= there is difference in mean.
Mean of Nz is significantly smaller than mean of JB, since p almost zero, we can reject the null hypothesis.

```
t_test_result <- t.test(yearlyNZ$rainfall, yearlyJG$rainfall)
print(t_test_result)

# Welch Two Sample t-test
#
# data:  yearlyNZ$rainfall and yearlyJG$rainfall
# t = -1.0623, df = 141.76, p-value = 0.2899
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   -126.34299   38.01817
# sample estimates:
#   mean of x mean of y
# 989.2699 1033.4323
```

Null Hypothesis= there is no significance difference between the mean
Alternative Hypothesis= there is difference in mean.
Mean of Nz is smaller than mean of JG but p is > 0.05, so there is not enough evidence to reject the null hypothesis.

6. Conclusion

- The mean of the rainfall data was found to be more towards the NE direction.
- The most amount of rainfall is received by district Jayashankar Bhupalpally.
- There is high amount of rainfall in the month of June, July, August and September.
- There is a decreasing trend of rainfall probability, representing less frequency.
- Rainfall follows normal distribution in all the districts.

References

Statistical Methods in Water Resources, Dennis R. Helsel,Robert M. Hirsch.
www.stackoverflow.com
Google Images.
Rpubs.com
Github.com
Special Thanks to Dr. Satish, Mohd. Azharuddin Class on PCA and Prof. Balaji's class on Hypothesis Testing.