Rainfall EDA and Time Series Data Analysis of Northern (Himachal Pradesh) and Southern (Tamilnadu) states of India

Final Semester Project

Name : Arnab Das

Registration Number : 012-1111-0609-20

Roll Number : 203012-21-0040

Semester : VI

Paper : DSE-B2

Rainfall EDA and Time Series Data Analysis of Northern (Himachal Pradesh) and Southern (Tamil nadu) states ofIndia.

Department Of Statistics, Asutosh College, Kolkata

Contents

Summary Of The Project

Rainfall is a natural process in which atmospheric vapour changes into water. The water so formed then travels from atmosphere to earth. The term precipitation is also used for rainfall. Rainfall forecasting is very important because heavy and irregular rainfall can have many impacts like destruction of crops and farms, damage of property so a better forecasting model is essential for an early warning that can minimize risks to life and property and also managing the agricultural farms in better way. This prediction mainly helps farmers and also water resources can be utilized efficiently. Rainfall prediction is a challenging task and the results should be accurate.

The has been collected from www.kaggle.com. We have taken the dataset of two different states namely HimachaL Pradesh and Tamil nadu from January 1980 to December 2017.The objective of the project is to perform time series analysis of the acquired data. Here, the AR(1) model is being used to do the same to forecast the rainfall coverage till the year 2024 for HimachaL Pradesh and Tamil nadu and successively conclude which sate receives the most rainfall over these years.

Introduction

Definition of One Centimetre Rainfall: Suppose the water precipitated on a certain plain area in the form of a rainfall is not lost in any manner and if there is no runoff and evaporation whatsoever then all the water will go on accumulating on the surface of the area in the form of a layer. When the layer of this deposited water is one centimetre thick, it is said that one centimetre rainfall has occurred.

Characteristics of Rainfall: Rain falling over a region is neither uniformly distributed nor is it constant over time. It is really difficult to predict when and how much of rain would fall. However it is possible to measure the amount of rain falling at any point and measurements from different point gives an idea of the rainfall pattern within an area.

In India, the rainfall is predominantly dictated by the monsoon climate. The monsoon in India arises from the reversal of the prevailing wind direction from Southwest to Northeast and results in three distinct seasons during the course of the year. The Southwest monsoon brings heavy rains over most of the country between June and October, and is referred to commonly as the ‘wet’ season.

Moisture laden winds sweep in from the Indian Ocean as low-pressure areas develop over the subcontinent and release their moisture in the form of heavy rainfall. Most of the annual rainfall in India comes at this time with the exception of in Tamil nadu, which receives over half of its rain during the North East monsoon from October to November.

The retreating monsoon brings relatively cool and dry weather to most of India as drier air from the Asian interior flows over the subcontinent. From November until February, temperatures remain cool and precipitation low.

In northern India it can become quite cold, with snow occurring in the Himalayas as weak cyclonic storms from the west settle over the mountains. Between March and June, the temperature and humidity begin to rise steadily in anticipation of the South West monsoon.

This pre-monsoonal period is often seen as a third distinct season although the post-monsoon in October also presents unique characteristics in the form of slightly cooler temperatures and occasional light drizzling rain. These transitional periods are also associated with the arrival of cyclonic tropical storms that batter the coastal areas of India with high winds, intense rain and wave activity.

Rainfall and temperature vary greatly depending on season and geographic location. Further, the timing and intensity of the monsoon is highly unpredictable. This results in a vastly unequal and unpredictable distribution over time and space. In general, the Northern half of the subcontinent sees greater extremes in temperature and rainfall with the former decreasing towards the north and the latter towards the west.

Measurement of Rainfall: One can measure the rain falling at a place by placing a measuring cylinder graduated in a length scale, commonly in mm. In this way, we are not measuring the volume of water that is stored in the cylinder, but the 'depth' of rainfall. The cylinder can be of any diameter, and we would expect the same 'depth' even for large diameter cylinders provided the rain that is falling is uniformly distributed in space.

Now think of a cylinder with a diameter as large as a town, or a district or a catchment of a river. Naturally, the rain falling on the entire area at any time would not be the same and what one would get would be an 'average depth'. Hence, to record the spatial variation of rain falling over an area, it is better to record the rain at a point using a standard sized measuring cylinder. Modern technology has helped to develop Radars, which measures rainfall over an entire region. However, this method is rather costly compared to the conventional recording and non-recording rain gauges which can be monitored easily with cheap labour.

Factors Affecting Rainfall: Factors which affect the rainfall are the following:

(i) Nearness to Sea: Coastal area receives more rainfall.
(ii) Presence of Mountains: Mountainous region receives more rainfall than plain area.
(iii) Direction of Wind: Movement of clouds depends upon the direction of wind. The area over which wind brings clouds will get rainfall.
(iv) Development of Forest: The area with thick forest gets more rainfall.
(v) Height of a Place above Sea-Level: At high altitudes temperature is low and hence when clouds reach that area they get cooled and precipitation takes place.

Methodology

The values of a variable recorded for different points or intervals of time for an individual or a population are called time series data. Yields of paddy, population or literacy rates of India in the last fifteen years, for example, are time series data. Again, values of a variable for different individuals in a group for the same point or interval of time are called spatial series data. Tea production in different tea-producing countries in a year or budget estimates of receipts of a country from different heads in a year are spatial series data. In both time series and spatial series, we may have cross-sectional data when each member of a group is classified on two or more characters.

For example, the population of a country in different years may be given, classified by sex and residential status. Similarly, the literacy rates in several states in a year may be given separately for males and females, besides for the whole population.

Exploratory Data Analysis: In statistics, exploratory data analysis is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. Primarily Exploratory Data Analysis is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data. Tukey defined data analysis in 1961 as: “Procedures for analysing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analysing data. “ In Exploratory Data Analysis we first check the measure of location. A measure of location is a measure of the center of a batch of a numbers. The most commonly used measure of location is Arithmetic Mean (or simply, Mean). The mean is defined as,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_{i.} \text{ i.e. } \overline{x} = \frac{\textit{Sum of the obsevations}}{\textit{Total No. of observations}}$$

The next common measure of location is, Median. Median is defined to be the middle value of an ordered data set. Quartiles are also used as measure of location. A p^{th} quartile divides the whole dataset into four equal parts; where p = 1(1)4. The 1st, 2nd, 3rd & 4th quartiles are denoted by Q1, Q2, Q3 & Q4 respectively. Clearly, Median is the 2nd Quartile. The difference between the 3rd and 1st quartile is known as Interquartile Range (IQR), which is a measure of dispersion. A measure of dispersion gives a numerical indication of the “scatteredness “ of a batch of numbers. The most common measure of dispersion is Variance, defined as,

Boxplot: A Boxplot is a graphical display invented by Tukey that shows a measure of location (the Median), a measure of dispersion (the Interquartile Range) and the presence of possible outliers and also gives an indication of the symmetry or skewness of the data. Outlier is a data point that differs significantly form other data points or observations. The outliers are detected by following method: If a data point is greater than Q3 + 1.5 × IQR or if the data point is less than Q1 - 1.5 × IQR then that data point is an outlier.

Construction of a Boxplot:

1. Horizontal lines are drawn at the median and at the upper and lower quartiles and are joined by vertical lines to produce the box.

2. A vertical line is drawn up from the upper quartile to the most extreme data point that is within a distance of 1.5 (IQR) of the upper quartile. A similarly defined vertical line is drawn down from the lower quartile. Short horizontal lines are added to mark the ends of these vertical lines.

3. Each data point beyond the ends of the vertical lines is marked with an asterisk or dot (* or ·).

Line Diagram: This diagram is meant for representing chronological data. In fact, it exhibits the relationship of the variable (e.g. sales of coffee of a company, productions of a crop) may be specified for individual points of time or for different period of time. In constructing a line diagram, two axis of co-ordinates are taken, the horizontal one for time and the vertical one for variable. The scale for each axis is then selected and the data are plotted as different points on the plane, the plotting of variable values being done against points of time or mid-points of the time interval (for time period). The successive points are now joined by straight line segments and the chart so obtained is called a line diagram for the given data. Two or mutually related time series data having same unit of measurement can be represented using the same axis of co-ordinates, by drawing a number of line diagrams, one for each series. These different line diagrams are mutually distinguished by using distinct pattern of lines such as broken lines, dotted lines or multiple-coloured lines. The resulting diagram is known as a Multiple Line Diagram. It is used for comparing two mutually related time series data e.g. if we want to compare the literacy rates for a number countries last 15 years, say, we may draw multiple line diagram.

Time Series Analysis: If we observe or record numerical features of an individual or a population for different points or intervals of time, the set of observations forms a time series. The population of a country over a number of years, temperature of a place noted on different days of a week, the weight of an animal recorded on different months, and yearly production of wheat are some common examples of time series data. In fact, the familiar time series data, such as economic and demographic series, are usually available at equal intervals of time like days, years, etc. and hence throughout our discussion we shall consider only series of values given at equispaced intervals of time. If Yt be the value of the time series at time t, then mathematically it may be represented by a function of t. In case of period data, t is to be considered as the mid-point of the t-th period. It should be noted that past time series data are analysed to detect the nature of variations in the data and subsequently to enable one to plan the future judiciously. Such analysis is also important for business forecasting. Some preliminary adjustments to the time series data are sometimes necessary to make them amenable for statistical analysis. The raw data may not be comparable for several reasons. It is known that the number of days in different calendar months, as well as the number of working days in various months are not the same. It is, therefore, essential to convert monthly data into a standard period by necessary adjustment. For instance, to make the monthly data on industrial production comparable, the figure for each month is to be divided by the number of working days in the concerned month. Thus, the figures are obtained per working-day basis. The data related to populations or geographical regions should be transformed to per capita or per unit basis. Again, with changes in the price-level, the purchasing power of money changes. Hence monetary data are also not comparable over time. In order to bring the figures to a comparable basis, the effect of price-changes are eliminated dividing the current period figure by a suitable price index number (e.g., the wholesale price index number of a country for data on the national income of the country) of the current period with the selected base period. This technique is called deflation. When time series data are graphically exhibited, the variations can be readily observed. Apparently, the graph represents an overall picture of haphazard Movement, but in reality it is not so. It is observed that at least a part of the changes, known as the systematic part, can be accounted for and the remaining part is the irregular. The main factors constituting the systematic part are, Secular Trend (simply, Trend), Seasonal Variation and Cyclical Variation. It is noteworthy that some or all of these components may be present in a given time series and their isolation is useful for miscellaneous purposes. The value (Yt) of a time series at any time (t) is considered as the resultant Of the combined effect of secular trend (Tt), seasonal (St), cyclical (Ct) and Irregular (It) variations. In one approach, Yt is assumed to be the product of the above four components,

$$Y_t = T_t \text{ x } S_t \text{ x } C_t \text{ x } I_t \ ,$$

which is the multiplicative or product model. In the other approach, the following

additive or sum model is considered:

$$Y_t = T_t + S_t + C_t + I_t.$$

In product model, Tt has the same unit as that of Yt whereas St, Ct and It are unit free; but in sum model, all the components have same unit as that of Yt. Now we shall discuss the components one by one as follows: The term secular trend, or simply trend, means the smooth, regular, long-term movement of a series when observed over a long period of time. A series may show an upward or a downward trend, or may remain more or less at a constant level. For example, an upward trend may be observed in data on population, and a downward trend in money value; again, a series of barometric readings of a particular place may remain more or less at the same level. It should be noted that a series may change its course after some time, but sudden or frequent Changes are inconsistent with the idea of trend. By seasonal variations we meant a periodic movement (that is, a movement that repeats itself at regular intervals of time), where the period is not longer than one year. Monthly expenditures of a family, quarterly sales in a departmental Store and number of books issued from a library on different days of a week are some examples where seasonal variations are prominent. These variations in Economic time series may be attributed to two broad factors, namely, (i) climatic Changes of various seasons (e.g., the sale of ice-cream, demand for electric fans Goes up in summer), and (ii) habits, customs and conventions followed by the People at different times (for instance, sale of certain consumer goods increases during festival months). In some cases, the study of this component is of prime Importance. For example, a study of seasonal variations in the demand of different goods is very essential for efficient running of any departmental store. By cyclical fluctuations we mean the oscillatory movement in a time series with period of oscillation more than one year. One complete period is called a cycle. These variations, though more or less regular, are not necessarily periodic. Such fluctuations in a time series are usually attributed to a business cycle' comprising four successive phases, viz. prosperity for boom), recession, depression and recovery. The swing from boom to recovery and back again to boom is found to vary in time span. Most of the economic series, like series pertaining to prices Illustration are affected by business cycles. By irregular fluctuations we mean those variations which are either completely unaccountable or are caused by unpredictable events like floods, wars, earthquakes, strikes, etc. This category of movements includes all types of variations that are not covered by the other three components and, thus, may be regarded as residual variation. Method of moving averages: In this technique, a series of arithmetic means, each of m successive observations of the given data, is computed and these means are referred to as the moving averages of period m, where m is the average period of the cycles or a multiple of it. To begin with, we take the first m values; at the next stage, exclude the first and include the (m + 1)th value and so on. We repeat this process until we reach at the last set of m values. Each mean is placed against the mid-point of the time interval that it covers. If m is odd, the moving averages correspond to the tabulated times for which the time series is given. On the other hand, when m is even, each moving average falls midway between two tabulated time values. In this case, a subsequent two-item moving average is computed to make the resulting moving average values correspond to given times. These moving averages are the trend values for the corresponding times.

AUTO-REGRESSION SERIES: The value of a series at any time ‘t’ may depend upon its own value at times t-1, t-2,..., t-k, (say), the relationship being linear. A series represented by the recurrence relation:

$Y_t = f\ (y_{t-1}, y_{t-2}, ..., y_{t-k}) + \in_t = a_1\ y_{t-1} + a_2\ y_{t-2} + ... + a_k\ y_{t-k} + \in_t$

where f is a linear function and $\in_i$ is the 'disturbance' function such that $\in_i$ ‘s are identically and independently distributed (i.i.d.) random variables, $\in_i \sim N\ (0, \sigma^2)$, is known as 'auto regressive' series of order k. Under certain conditions to be satisfied by $a_1, a_2, ..., a_k$ the generating series can be shown to represent an oscillatory time series. In the following sequences, we shall assume (which we can do without loss of generality) that y_t's are measured from their means so that $E(y_t) = 0$.

AUTOREGRESSIVE MODEL: The notation AR(p’) refers to Autoregressive model of order p’. The AR(p’) is written as,

$$y_t = c + \sum_{i=1}^{p'} \varphi_i\ y_{t-i} + \in_t$$

Result

Himachal Pradesh:

Summary of the Data:

```
>view(himachal)
>#checking the  class of the data
>class(himachal)
>class(himachal)
[1]   "tbl_df  "          "tbl  "         "data.frame  "
```

Fig 1.2: Class of the obtained Data

```
>#making the class of the dataset from   "data frame  " to   "time series  "
>himachal$DATE=as.Date(himachal$DATE,format=   "%y-%m-%d  ")
>Rainfall<-ts(himachal$'RAINFALL IN MM',start=1980,end=c(2017,12),frequency=12)
>#checking the class of dataset again
>class(Rainfall)
[1]   "ts  "
>start(Rainfall)
[1]  1980    1
>end(Rainfall)
[1]  2017    12
```

Fig 1.3: converting the data from   "data frame " form to   "time series " form in R

Also, in fig 1.3 we can see that we have the data (Rainfall) from January 1980('1' for January) to December 2017 ('12' for December).

```
> summary(Rainfall)
Min.   1st Qu.  Median   Mean    3rd Qu.     Max.
0.00   36.20    75.60        99.63    136.75      477.10
```

Fig 1.4: Summary of the Data

We can see that the median rainfall in mm 75.60 form fig 1.4. Also, the mean rainfall in mm is 99.63. The maximum rainfall in mm is 477.10 and the minimum rainfall in mm is 0.00. We have, Q1 = 36.20 and Q3 = 136.75.

```
>plot(Rainfall)
```

Fig 1.5: R Code to plot the Data

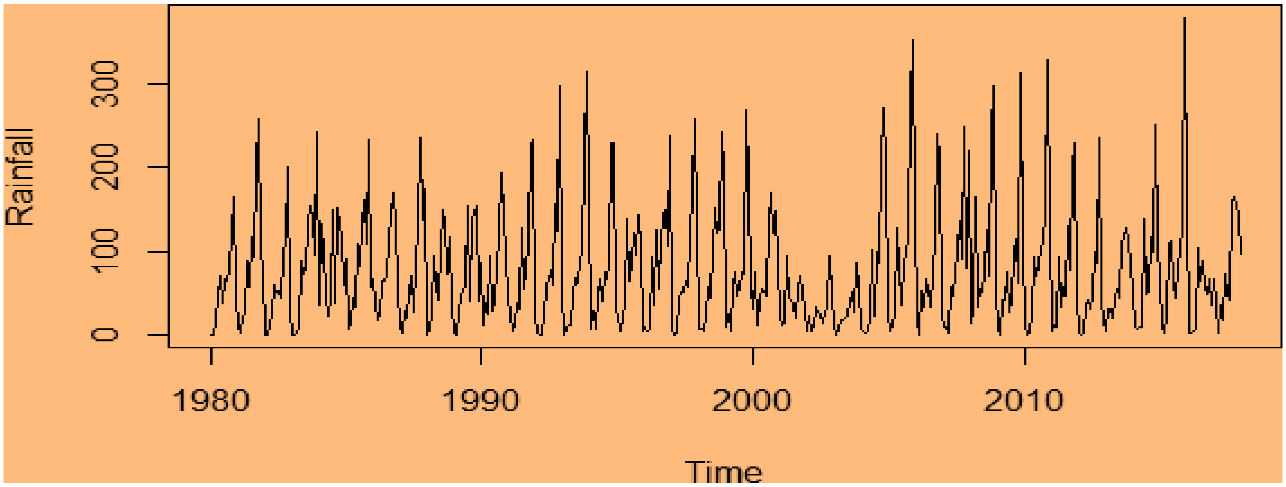


Fig 1.6: Plotted Data in RStudio

```
>#Getting a boxplot by Cycle
>boxplot(Rainfall~cycle(Rainfall))
```

Fig 1.7: Getting Boxplot of the Data Monthwise

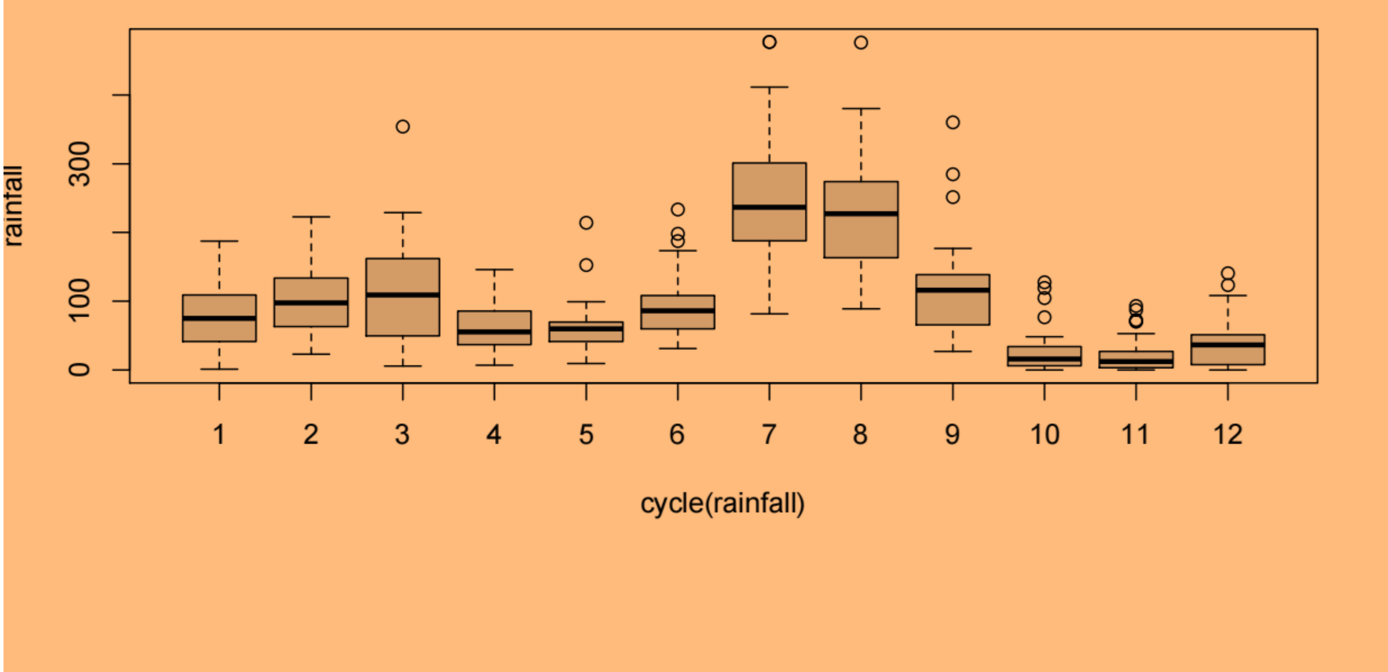


Fig 1.8: Boxplot of the Data Monthwise

From fig 1.8 we can see that the median rainfall in mm of each month is different. We can see that the month of October and November receives maximum rainfall throughout the year.

```
>data2<-decompose(Rainfall,   “additive   “)
```

```
>plot(data2)
```

Fig 1.9: R Code to check the components

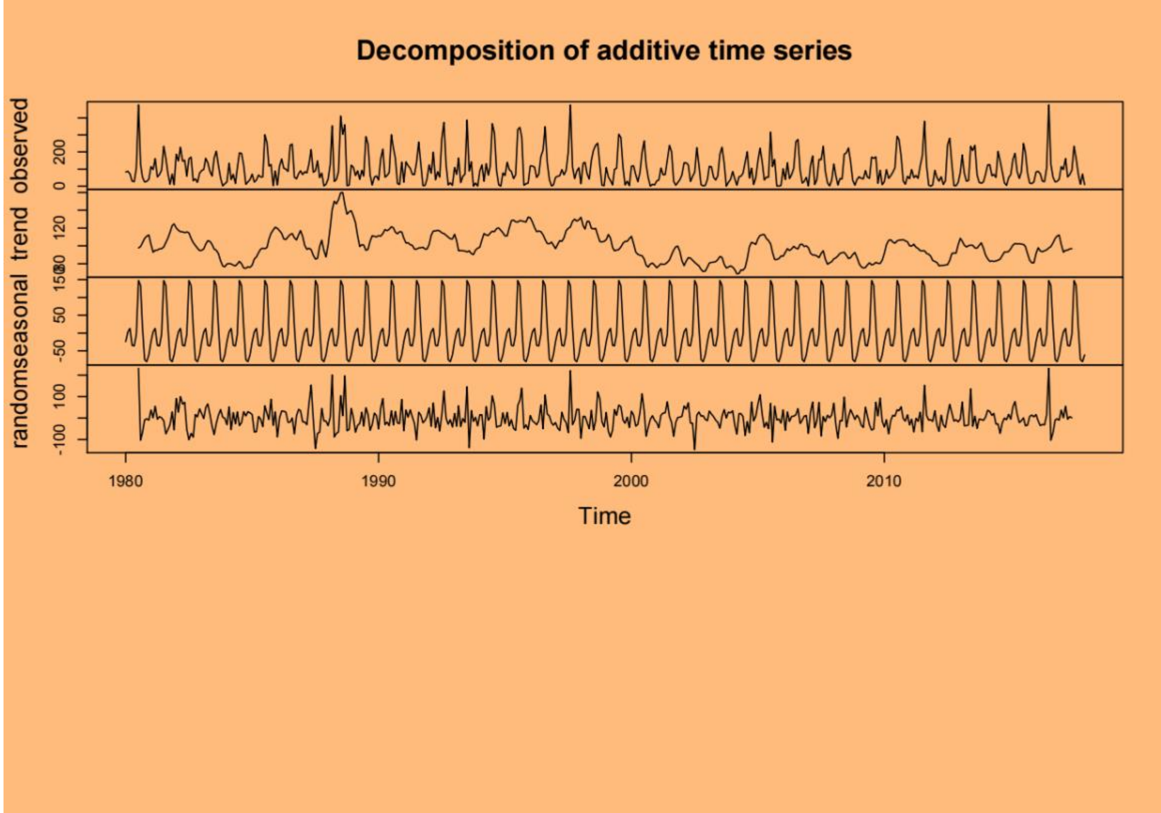


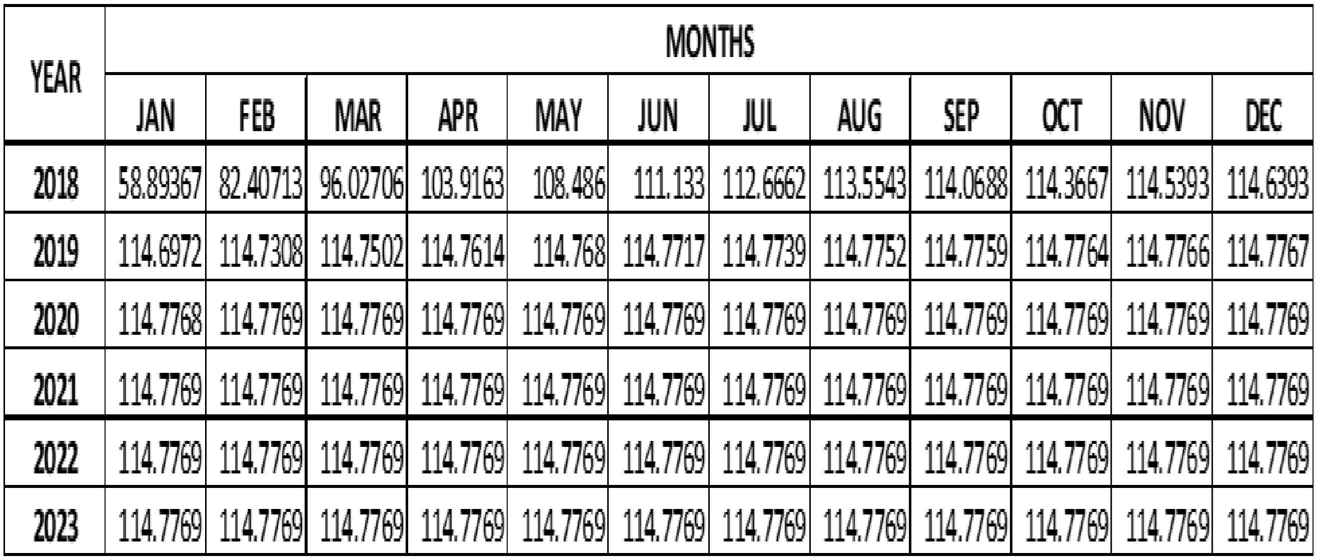
Fig 1.10: The Components present in the Data (Additive Model)

```
> ar1=arima(Rainfall,c(1,0,0))
> ar1
Call:
arima(x = Rainfall, order = c(1, 0, 0))
Coefficients:
         ar1        intercept
      0.3747      99.4829
s.e. 0.0434       6.1438
sigma^2 estimated as 6749: log likelihood = -2657.41, aic = 5320.82
```

Fig 1.11: Fitted AR (1) Model

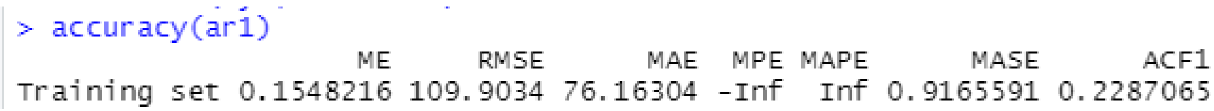
The fitted AR(1) model is, $y_t = 99.4829 + 0.3747\, y_{t-1} + \in_t$

Forecasted Rainfall (in mm) for the next 6 years using AR(1):



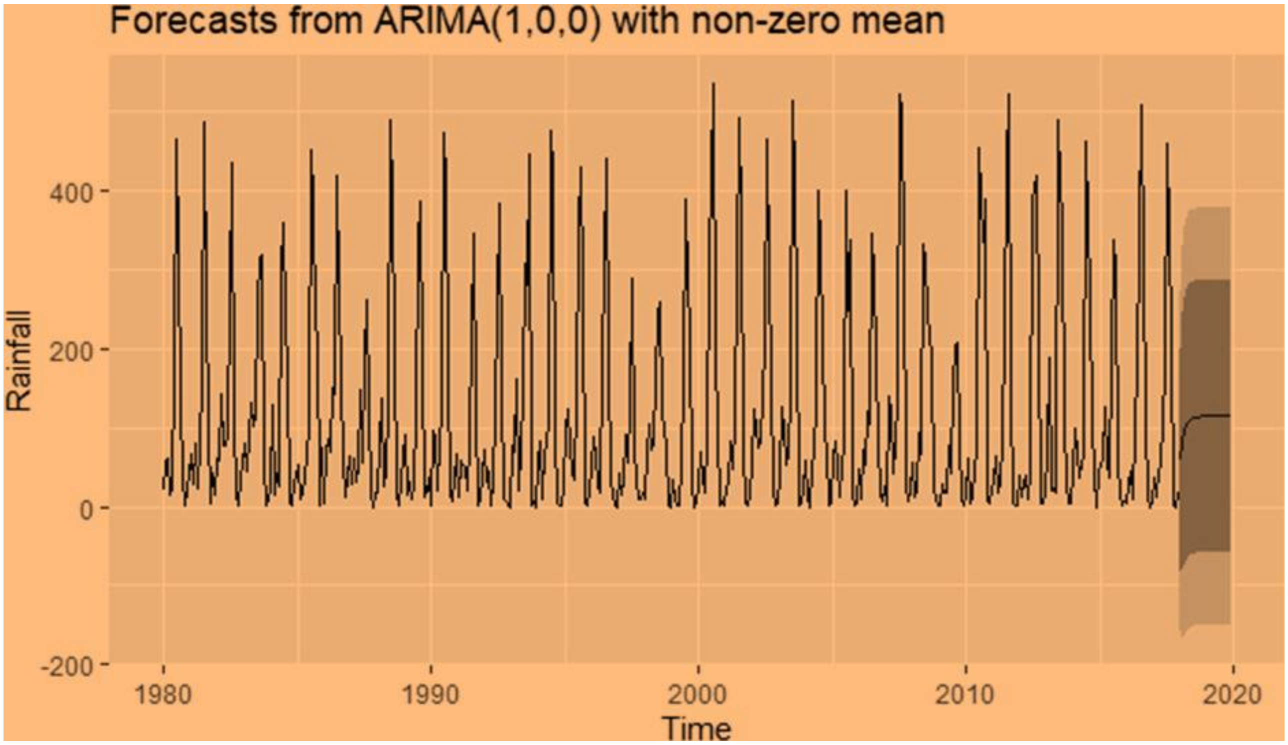
YEAR	MONTHS											
	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
2018	58.89367	82.40713	96.02706	103.9163	108.486	111.133	112.6662	113.5543	114.0688	114.3667	114.5393	114.6393
2019	114.6972	114.7308	114.7502	114.7614	114.768	114.7717	114.7739	114.7752	114.7759	114.7764	114.7766	114.7767
2020	114.7768	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769
2021	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769
2022	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769
2023	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769	114.7769

Fig 1.12: Plotted Forecasted Rainfall coverage using AR(1) model with the Observed rainfall (with 80% and 95% Confidence Intervals)



```
> accuracy(ar1)
                    ME     RMSE      MAE  MPE MAPE      MASE      ACF1
Training set 0.1548216 109.9034 76.16304 -Inf  Inf 0.9165591 0.2287065
```

Fig 1.13: Accuracy of the fitted model



MAPE is calculated as the average of(actual-predicted/abs(actual). This means that the function will return -Inf, Inf or NaN if the actual is zero. Due to the instability at or near zero, SMAPE or MAPE are often used as alternatives.

When we have a MASE=1, that means the model is exactly as good as just picking the last observation. An MASE=0.5, means that our model has doubled the prediction accuracy. The lower, the better. When MASE >1, that means the model needs a lot of improvement.

Result For Forecasted Rainfall Coverage in HIMACHAL PRADESH from 2018-2023

```
>View(himachal)
>#checking the class of the data
>class(himachal)
[1]   "tbl_df  "          "tbl  "         "data.frame  "
```

Fig 1.15: Class of the obtained Data

```
>#making the class of the dataset from   "data frame  " to   "time series  "
>himachal$DATE=as.Date(himachal$DATE,format=   "%y-%m-%d  ")
>Rainfall<-ts(himachal$'RAINFALL IN MM',start=2018,end=c(2023,12),frequency=12)
>#checking the class of dataset again
>class(Rainfall)
[1]   "ts  "
>start(Rainfall)
[1]  2018    1
>end(Rainfall)
[1]  2023    12
```

Fig 1.16: converting the data from   "data frame " form to   "time series " form in R

Also, in fig 1.16 we can see that we have the data (Rainfall) from January 2018 ('1' for January) to December 2023 ('12' for December).

```
> summary(Rainfall)
Min.   1st Qu.  Median   Mean    3rd Qu.     Max.
85.54  90.44     100.76     110.36     136.75      477.10
```

Fig 1.17: Summary of the Data

We can see that the median of forecasted rainfall in mm 85.54 form fig 1.17. Also, the mean rainfall in mm is 90.44. The maximum rainfall in mm is 477.10 and the minimum rainfall in mm is 85.54 We have, Q1 = 90.44 and Q3 = 136.75.

TAMILNADU:-

Summary of the Data:

```
>view(tamilnanu)
>#checking the  class of the data
>class(tamilnadu)
>class(tamilnadu)
[1]    “tbl_df   “              “tbl   “          “data.frame   “
```

Fig 1.2: Class of the obtained Data

```
>#making the class of the dataset from    “data frame   “ to    “time series   “
>tamilnadu$DATE=as.Date(tamilnadu$DATE,format=    “%y-%m-%d   “)
>Rainfall<-ts(tamilnadu$’RAINFALL IN MM’,start=1980,end=c(2017,12),frequency=12)
>#checking the class of dataset again
>class(Rainfall)
[1]    “ts   “
>start(Rainfall)
[1]  1980     1
>end(Rainfall)
[1]  2017    12
```

Fig 2.3: converting the data from   “data frame   “ form to   “time series   “ form in R

Also, in fig 2.3 we can see that we have the data (Rainfall) from January 1980 (‘1’ for

January) to December 2017('12' for December).

```
> summary(Rainfall)
Min.   1st Qu.  Median   Mean    3rd Qu.      Max.
0.00   23.70      55.90      75.54     110.88      379.80
```

Fig 2.4: Summary of the Data

We can see that the median rainfall in mm 55.90 form fig 2.5. Also, the mean rainfall in mm is 75.54. The maximum rainfall in mm is 379.80 and the minimum rainfall in mm is 0.00. We have, Q1 = 23.70 and Q3 = 110.88.

```
>plot(Rainfall)
>
```

Fig 2.5: R Code to plot the Data

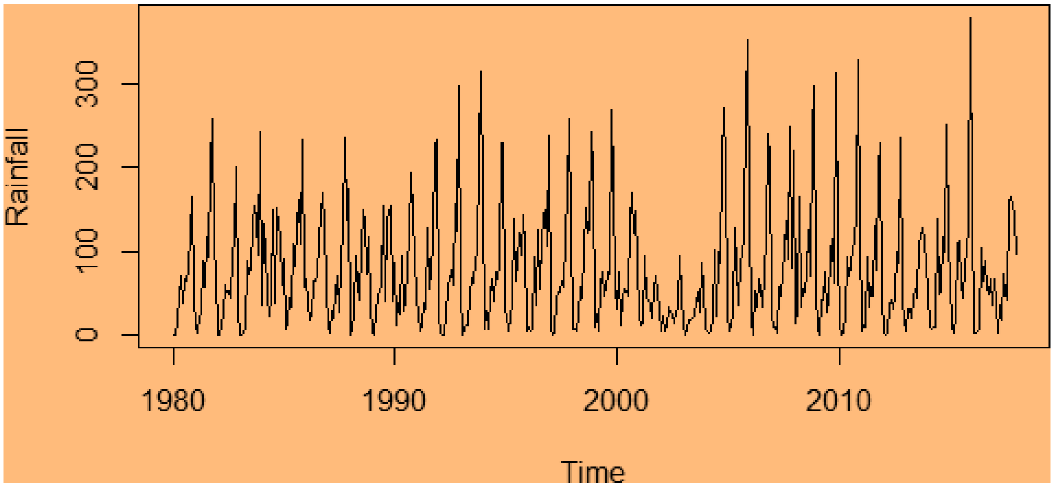


Fig 2.6: Plotted Data in RStudio

```
>#Getting a boxplot by cycle
>boxplot(Rainfall~cycle(Rainfall))
```

Fig 2.7: Getting Boxplot of the Data Monthwise

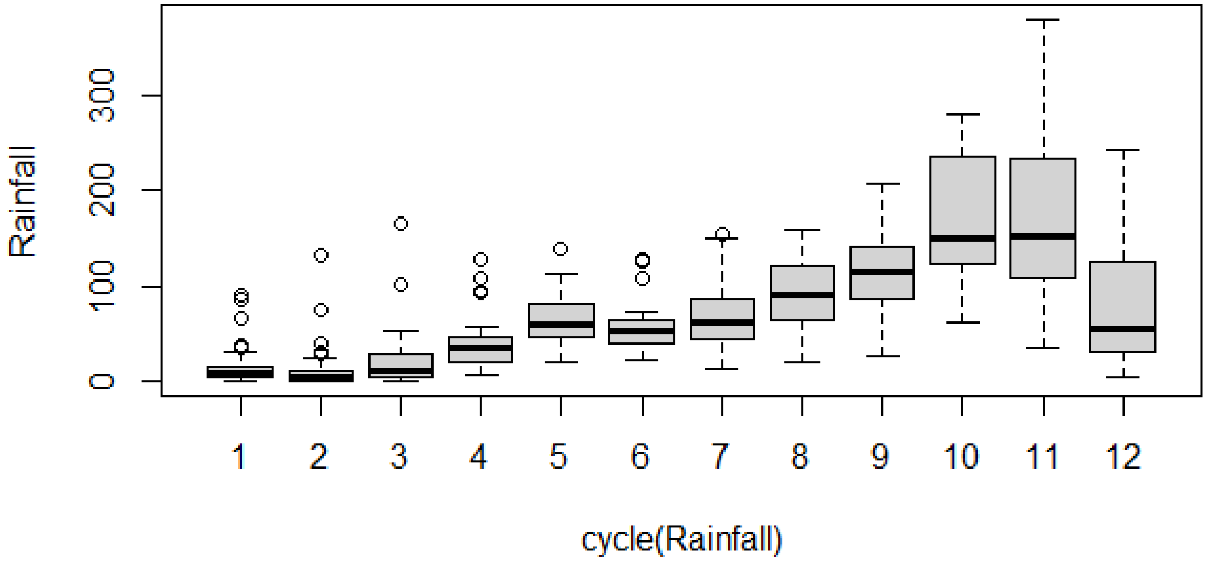


Fig 2.8: Boxplot of the Data Monthwise

From fig 2.8 we can see that the median rainfall in mm of each month is different. We can see that the month of July and August receives maximum rainfall throughout the year.

```
>data3<-decompose(Rainfall,    “additive “)
>plot(data3)
```

Fig 2.9: R Code to check the components

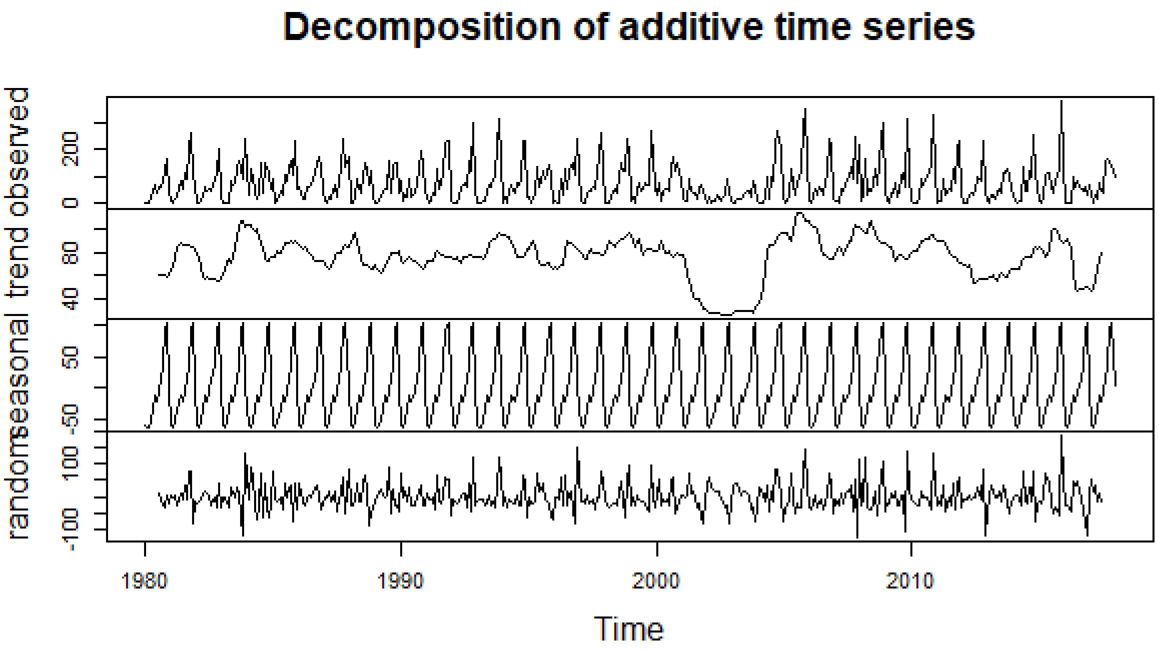
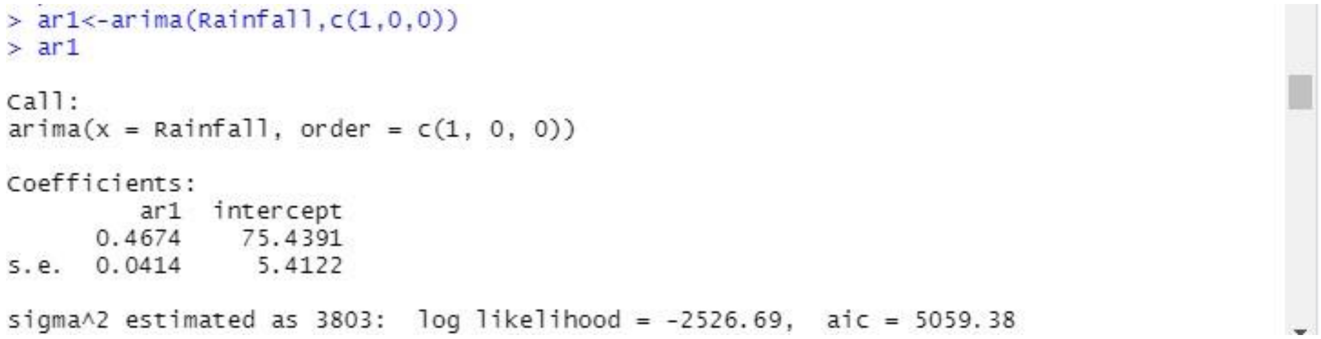


Fig 2.10: The Components present in the Data (Additive Model)

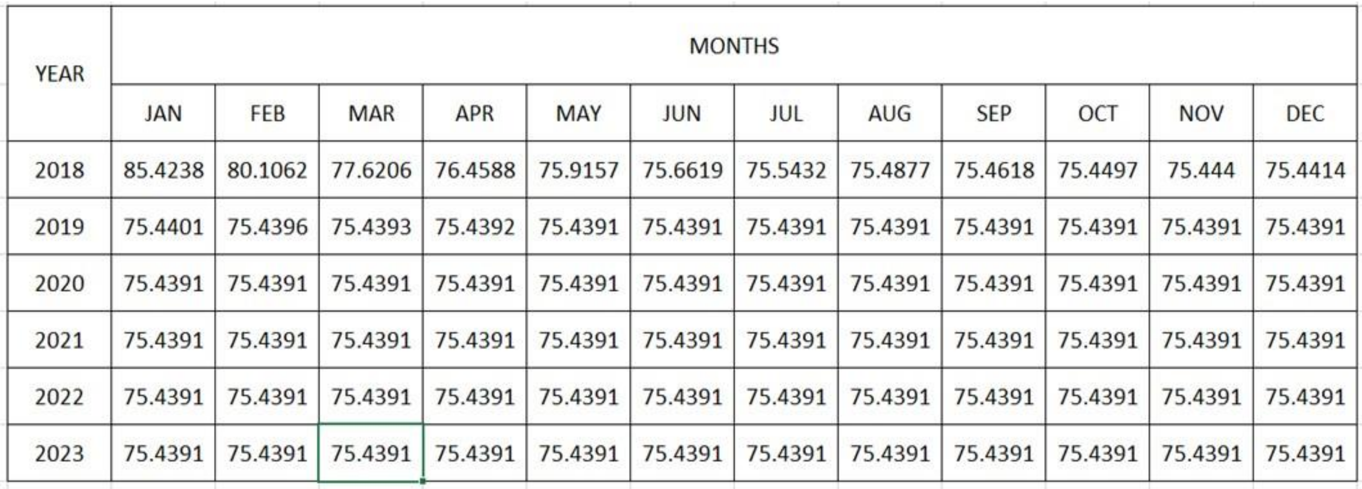


```
> ar1<-arima(Rainfall,c(1,0,0))
> ar1

Call:
arima(x = Rainfall, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.4674    75.4391
s.e.  0.0414     5.4122

sigma^2 estimated as 3803:  log likelihood = -2526.69,  aic = 5059.38
```

Fig 2.11: Fitted AR (1) Model

The fitted AR(1) model is, $y_t = 75.4391 + 5.4122\, y_{t-1} + \in_t$

Forecasted Rainfall (in mm) for the next 6 years using AR(1):



YEAR	MONTHS											
	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
2018	85.4238	80.1062	77.6206	76.4588	75.9157	75.6619	75.5432	75.4877	75.4618	75.4497	75.444	75.4414
2019	75.4401	75.4396	75.4393	75.4392	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391
2020	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391
2021	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391
2022	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391
2023	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391	75.4391

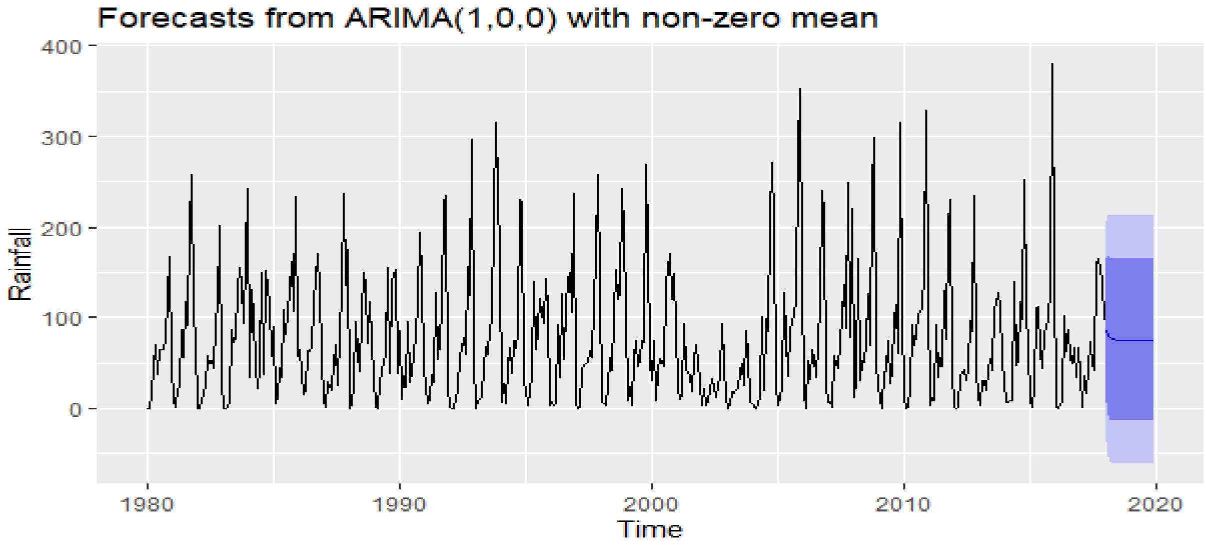
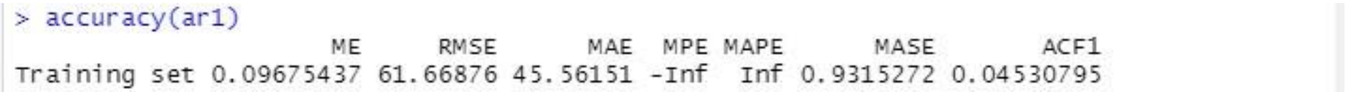


Fig 2.12: Plotted Forecasted Prices using AR(1) model with the Observed rainfall (with 80% and 95% Confidence Intervals)



```
> accuracy(ar1)
                     ME     RMSE      MAE  MPE MAPE      MASE       ACF1
Training set 0.09675437 61.66876 45.56151 -Inf  Inf 0.9315272 0.04530795
```

Fig 2.13: Accuracy of the fitted model

MAPE is calculated as the average of(actual-predicted/abs(actual). This means that the function will return -Inf, Inf or NaN if the actual is zero. Due to the instability at or near zero, SMAPE or MAPE are often used as alternatives.

When we have a MASE=1, that means the model is exactly as good as just picking the last observation. An MASE=0.5, means that our model has doubled the prediction accuracy. The lower, the better. When MASE >1, that means the model needs a lot of improvement.

Fig 2.14: Line Diagram of Observed Rainfall with the Predicted Rainfall

The Line Diagram of the two sets (observed prices and fitted Rainfall (in mm)) of data also show that our model is able to fit the dataset almost perfectly.

Result For Forecasted Rainfall Coverage in Tamil Nadu from 2018-2023

```
>View(tamilnadu)
>#checking the class of the data
>class(tamilnadu)
[1]   “tbl_df  “          “tbl  “         “data.frame  “
```

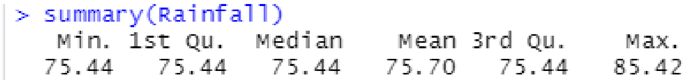
Fig 2.15: Class of the obtained Data

```
>#making the class of the dataset from   “data frame  “ to   “time series  “
>tamilnadu$DATE=as.Date(tamilnadu$DATE,format=   “%y-%m-%d  “)
>Rainfall<-ts(tamilnadu$’RAINFALL IN MM’,start=2018,end=c(2023,12),frequency=12)
>#checking the class of dataset again
>class(Rainfall)
[1]   “ts  “
>start(Rainfall)
[1]  2018   1
>end(Rainfall)
[1]  2023   12
```

Fig 2.16: converting the data from   “data frame “ form to   “time series “ form in R

Fig 2.16: converting the data from "data frame " form to "time series " form in R

Also, in fig 2.16 we can see that we have the data (Rainfall) from January 2018 ('1' for January) to December 2023 ('12' for December).



```
> summary(Rainfall)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  75.44   75.44   75.44   75.70   75.44   85.42 
```

Fig 2.17: Summary of the Data

We can see that the median of forecasted rainfall in mm 75.44 from fig 2.17. Also, the mean rainfall in mm is 75.70. The maximum rainfall in mm is 75.44 and the minimum rainfall in mm is 75.44. We have, Q1 = 75.44 and Q3 = 75.44

Conclusion

We have analysed the monthly data of Rainfall in Tamil nadu and HimachaL Pradesh fromthe year 1980-2017 in this project. From the basic Time Series Analysis, we have found that the datas follows a sideways trend over the years for Tamil nadu and HimachaL Pradesh. We have also tried to fit a model using Autoregressive model. We have fitted AR (1) model to the data and forecasted the monthly Rainfall coveragefor both the states for the next 6 years. From the forecasted/predicted Rainfall datawe can see that the trendline at first increases and then becomes constant. We cansee that the models quickly converge to flat lines (if there is no differencing), or tothe appropriate linear trend (if the order of differencing is 1). To be noted that theinitial forecasts are not flat it only looks flat as we are forecasting a long way out.

From the summary of the forecasted datas of Tamil nadu and HimachaL Pradesh from the year 2018-2023 we can see that HimachaL Pradesh is going to receive more rainfall coverage over the years than Tamil nadu.

References

Sources of Theory:

1. A.M. Gun, M.K. Gupta, B. Dasgupta (2017): Fundamentals of Statistics (Volume One & Two), World Press.

2. S.C. Gupta, V.K. Kapoor (2021): Fundamentals of Applied Statistics, Sultan Chand & Sons.

3. Avril Coghlan (2018): A Little Book of R for Time Series (Release 0.2).

4. Wayne A. Woodward, Henry L. Gray, Alan C. Elliott (2016): Applied Time Series Analysis with R (Second Edition), CRC Press, A Chapman & Hall Book.

5. Paul S.P. Cowpertwait, Andrew V. Metcalfe (2008): Introductory Time Series with R, Springer.

Software Used:

R version 4.0.5

RStudio

Microsoft Excel 2019

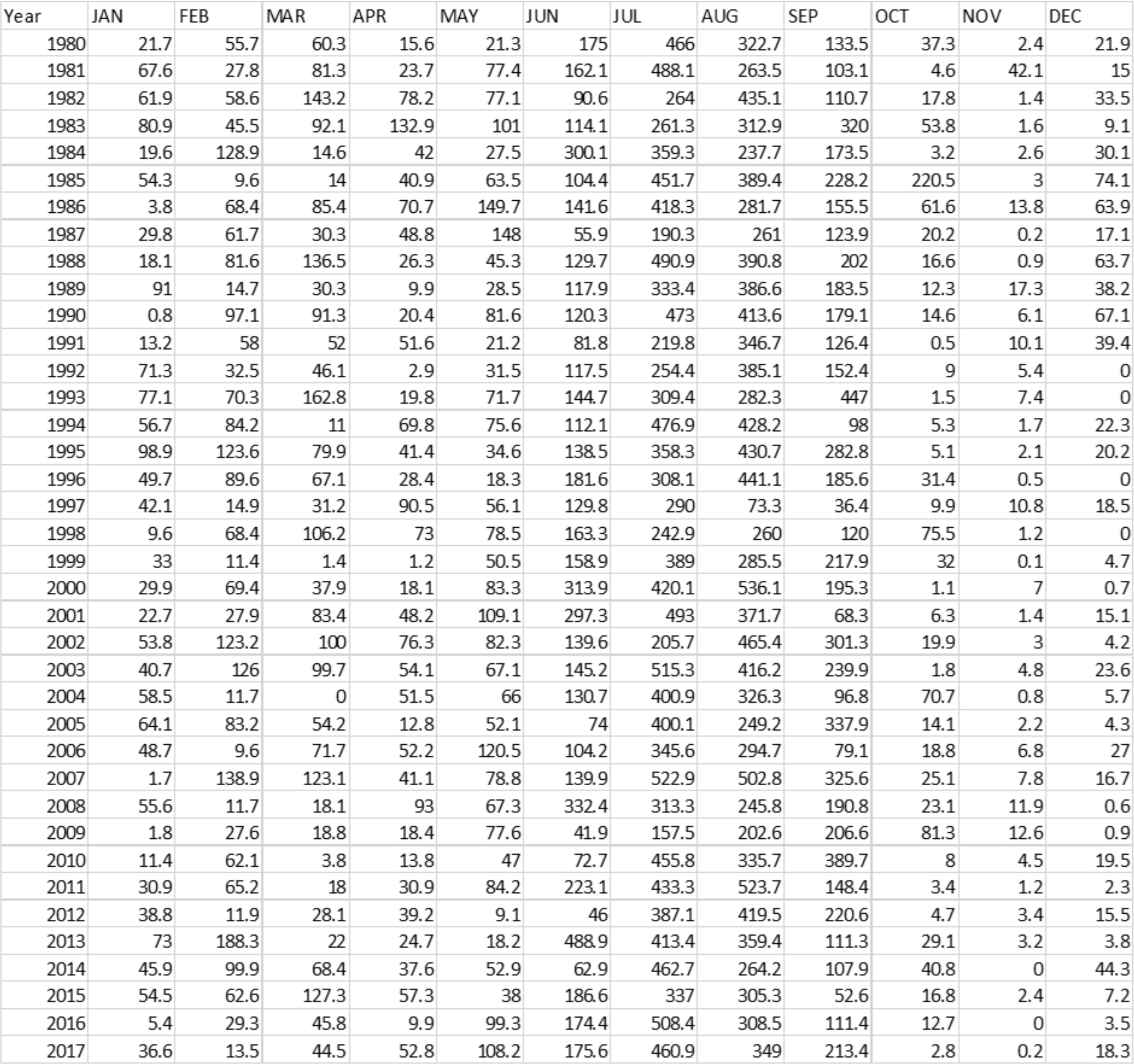
Microsoft Word 2019

Appendix

The Original Dataset:

Collected Monthly Data on Rainfall(in mm) from January 1980– December 2017:

Himachal Pradesh :-



Year	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1980	21.7	55.7	60.3	15.6	21.3	175	466	322.7	133.5	37.3	2.4	21.9
1981	67.6	27.8	81.3	23.7	77.4	162.1	488.1	263.5	103.1	4.6	42.1	15
1982	61.9	58.6	143.2	78.2	77.1	90.6	264	435.1	110.7	17.8	1.4	33.5
1983	80.9	45.5	92.1	132.9	101	114.1	261.3	312.9	320	53.8	1.6	9.1
1984	19.6	128.9	14.6	42	27.5	300.1	359.3	237.7	173.5	3.2	2.6	30.1
1985	54.3	9.6	14	40.9	63.5	104.4	451.7	389.4	228.2	220.5	3	74.1
1986	3.8	68.4	85.4	70.7	149.7	141.6	418.3	281.7	155.5	61.6	13.8	63.9
1987	29.8	61.7	30.3	48.8	148	55.9	190.3	261	123.9	20.2	0.2	17.1
1988	18.1	81.6	136.5	26.3	45.3	129.7	490.9	390.8	202	16.6	0.9	63.7
1989	91	14.7	30.3	9.9	28.5	117.9	333.4	386.6	183.5	12.3	17.3	38.2
1990	0.8	97.1	91.3	20.4	81.6	120.3	473	413.6	179.1	14.6	6.1	67.1
1991	13.2	58	52	51.6	21.2	81.8	219.8	346.7	126.4	0.5	10.1	39.4
1992	71.3	32.5	46.1	2.9	31.5	117.5	254.4	385.1	152.4	9	5.4	0
1993	77.1	70.3	162.8	19.8	71.7	144.7	309.4	282.3	447	1.5	7.4	0
1994	56.7	84.2	11	69.8	75.6	112.1	476.9	428.2	98	5.3	1.7	22.3
1995	98.9	123.6	79.9	41.4	34.6	138.5	358.3	430.7	282.8	5.1	2.1	20.2
1996	49.7	89.6	67.1	28.4	18.3	181.6	308.1	441.1	185.6	31.4	0.5	0
1997	42.1	14.9	31.2	90.5	56.1	129.8	290	73.3	36.4	9.9	10.8	18.5
1998	9.6	68.4	106.2	73	78.5	163.3	242.9	260	120	75.5	1.2	0
1999	33	11.4	1.4	1.2	50.5	158.9	389	285.5	217.9	32	0.1	4.7
2000	29.9	69.4	37.9	18.1	83.3	313.9	420.1	536.1	195.3	1.1	7	0.7
2001	22.7	27.9	83.4	48.2	109.1	297.3	493	371.7	68.3	6.3	1.4	15.1
2002	53.8	123.2	100	76.3	82.3	139.6	205.7	465.4	301.3	19.9	3	4.2
2003	40.7	126	99.7	54.1	67.1	145.2	515.3	416.2	239.9	1.8	4.8	23.6
2004	58.5	11.7	0	51.5	66	130.7	400.9	326.3	96.8	70.7	0.8	5.7
2005	64.1	83.2	54.2	12.8	52.1	74	400.1	249.2	337.9	14.1	2.2	4.3
2006	48.7	9.6	71.7	52.2	120.5	104.2	345.6	294.7	79.1	18.8	6.8	27
2007	1.7	138.9	123.1	41.1	78.8	139.9	522.9	502.8	325.6	25.1	7.8	16.7
2008	55.6	11.7	18.1	93	67.3	332.4	313.3	245.8	190.8	23.1	11.9	0.6
2009	1.8	27.6	18.8	18.4	77.6	41.9	157.5	202.6	206.6	81.3	12.6	0.9
2010	11.4	62.1	3.8	13.8	47	72.7	455.8	335.7	389.7	8	4.5	19.5
2011	30.9	65.2	18	30.9	84.2	223.1	433.3	523.7	148.4	3.4	1.2	2.3
2012	38.8	11.9	28.1	39.2	9.1	46	387.1	419.5	220.6	4.7	3.4	15.5
2013	73	188.3	22	24.7	18.2	488.9	413.4	359.4	111.3	29.1	3.2	3.8
2014	45.9	99.9	68.4	37.6	52.9	62.9	462.7	264.2	107.9	40.8	0	44.3
2015	54.5	62.6	127.3	57.3	38	186.6	337	305.3	52.6	16.8	2.4	7.2
2016	5.4	29.3	45.8	9.9	99.3	174.4	508.4	308.5	111.4	12.7	0	3.5
2017	36.6	13.5	44.5	52.8	108.2	175.6	460.9	349	213.4	2.8	0.2	18.3

Tamil nadu:

Year	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1980	0.1	0	15.4	45.9	69.2	37.7	64.9	64.2	76.3	122.3	166.2	47
1981	8.9	1.3	21.7	23.7	87.2	57	117	91.8	200.5	258	106.4	72.5
1982	0.2	0.1	8.6	25.2	58.8	49.8	52.5	44.5	93.4	114	200.8	30.9
1983	0.2	0.2	1.9	5.5	86.9	72.9	77.9	132	154.6	138.7	94.5	242.6
1984	34.5	131.3	101.7	45.4	21.6	43.6	149.5	38	151.3	134.8	113.5	58.5
1985	89.7	5.2	11.9	43.9	33.4	108.8	81.2	129.7	161.6	107.3	233.3	57.4
1986	65.5	39	16.1	20.9	62.7	63.4	66.9	115.2	142.3	170.5	133.4	57.1
1987	8	1.2	28.7	20	50.5	70.5	26.1	82.8	127	236.7	136.9	175
1988	0.2	3.8	27.4	94.6	59.1	41.1	102.6	149.8	136.1	71.8	117.6	33.1
1989	2.7	0	27.9	40.8	53.6	57.4	154.9	39.7	137.3	148.6	154	39.8
1990	84.8	10.2	36.2	24.3	94.9	28.9	40	80.1	119.6	194.1	144	46.1
1991	24.2	4.6	9.7	37.4	29.5	128.1	54.7	71.7	114	225.1	234.4	21.4
1992	3.1	0.2	0	22.5	58.3	66.4	77.9	59.9	157.3	123.9	297.2	52.1
1993	0.1	7	9.9	11.2	46.6	68.2	60.3	88.8	96.7	214.5	315.9	163.1
1994	7	27.6	5.5	48.3	66.6	38.6	75.3	66.1	84.2	230.6	229.1	27
1995	23.8	3.3	16.7	39.5	139	64.6	86.7	121.1	94	143.1	107.9	3.7
1996	7.6	3.6	5	91.5	34	125.3	55.4	112.4	141.2	149.7	106	237.6
1997	7.7	0.2	2.3	42	49.2	50.4	63.5	56.4	116	169.5	258.1	135.2
1998	6.9	5.7	4.1	23.6	55.8	43	101.9	152.9	121.8	124.7	242	197.4
1999	8.3	24.1	3.9	56.9	74.5	46.4	55.3	74.8	73.1	268.9	182.8	53.6
2000	30.3	74.5	9.4	41.6	54.7	52.2	45.8	136.1	169.8	120.3	148.6	88.9
2001	21.8	10.9	13	93.8	45	39.9	37.2	18.7	53.8	69.5	52.3	27.6
2002	2.9	22.1	3.9	9.2	32.6	23.7	11.8	26.5	32.2	93.7	47.2	12.2
2003	0.3	4.1	18	17.1	19.8	22.5	38.7	49.3	26.1	86	59.5	7.1
2004	2.8	0.7	3	16.2	101.1	21.2	98.6	85.1	208.3	271.1	204.5	25
2005	4.1	11.1	24.4	128	80.6	35.7	87.9	93.3	117.9	280.5	353.4	148.5
2006	15.3	0.2	52.4	32.6	65.2	57.2	33.6	73.4	116.3	240.4	215.1	26.1
2007	7.2	7.5	1.8	58	44.9	73.1	101	136.5	89.1	248.9	79.2	219.9
2008	11.7	29.1	164.7	31.5	53.7	51.1	73.1	126.5	70.7	242.7	298.5	50.1
2009	7.9	0	41	41.4	74.8	27	42.1	96.9	114.4	62.1	314.7	106.2
2010	11.8	0.2	1.9	22.9	91.9	70	81.4	102.9	111.1	148.1	328.6	124.5
2011	4.3	11.2	8	91.5	33.4	56	45.5	128.9	76	200.4	230.5	41
2012	3	0.1	2.5	35.5	41.9	30.1	46.5	98	84.9	235.2	44.5	14
2013	3.9	30.9	30	20.3	42	54.6	42.7	110.7	113.5	127.9	112.3	53.2
2014	7.4	6.1	8.1	8.3	139.1	47.8	50.6	117.7	98.9	252.2	110.8	66
2017	8.3	2.3	21.7	108.8	112.4	62.4	43.5	81.6	98.4	132.6	379.8	152.8
2016	2.5	0.8	3.1	6.3	102.5	63	86.8	55.1	49	65.8	33.8	66.5
2017	37.3	1.1	35.4	17.3	73.5	47.9	42	159.4	165.3	155.5	141.5	96.8

Source of Data: https://www.kaggle.com/datasets/saisaran2/rainfall-data-from-1901-to-2017-for-india?resource=download

Acknowledgement

I, Arnab Das, would like to thank my respected Supervisor Dr. Sakha Bhattacharya for his continuous, patient & valuable guidance and support for execution of the project titled “Rainfall EDA and Time Series Data Analysis of Northern (HimachaL Pradesh) and Southern (Tamil nadu) states of India “. I am grateful to him for his valuable suggestions and inputs enabling me to successfully completemy project with ease. Without his assistance the project could not be finished at all. In addition, I would also like to thank our Head of the Statistics Department, Asutosh College Dr. Dhiman Dutta and other faculty members Dr.Shirsendu Mukherjee, Dr. Parthasarathi Bera & Prof. Oindrila Bose for their constant support. Without their valuable knowledge and assistance it was not possible to successfully complete the project.