ACKNOWLEDGEMENT

I, Adrija Karmakar have taken efforts in this project. However, it would not have been possible without the kind help and support of many individuals and organizations. I would like to extend my sincere thanks to all of them. I am highly indebted to my supervisor ***Mr. JOYDEEP BASU*** and ***Ms. SOUMITA MODAK*** for his expert guidance and constant supervision as well as for providing necessary information regarding the project. I would like to express my gratitude towards the professors of the Statistics Department, Basanti Devi College for their kind co-operation and encouragement which helped me in completion of this project. I would also take this opportunity to thank *Minitab*® and *R-Software* for developing open-source software which helped me immensely in my dissertation. My thanks and appreciations also go to my friends for their constant support and other people who have willingly helped me out with their abilities. Last but not the least; I would also like to thank my parents and my fellow mates for being by my side in the time of doing the project.

CONTENT

ABSTRACT:

Rainfall is a prime input for various engineering design such as hydraulic structures, bridges and culverts, canals, storm water sewer and road drainage system. The detailed statistical analysis of each region is essential to estimate the relevant input value for design and analysis of engineering structures and also for crop planning. I choose Sub Himalayan West Bengal and Sikkim area for my project. I take the monthly rainfall data for a period of 50 years (1965-2015).Then I fit an appropriate model, plot graphs and forecast on my following data. This analysis will provide useful information for water resources planner, farmers and urban engineers to assess the availability of water and create the storage accordingly.

INTRODUCTION:

Water is vital for any life process and there can be no substitute for it. Water is also used for transportation, is a source of power and serves many other useful purposes for domestic consumption, agriculture and industry. The main important source of water in any area is rain and it has a dramatic effect on agriculture. Plants get their water supply from natural sources and through irrigation.

Picture 1: Rainfall in India

The Sub-Himalayan West Bengal and Sikkim is a small yet one of the rainiest pockets during the Monsoon season. Sub-Himalayan West Bengal comprises of mountains and foothills. The normal rainfall in Sub-Himalayan West Bengal and Sikkim is more than the other rainiest pockets like Assam and Meghalaya.

At present, a Trough from Central Pakistan and adjoining Punjab is extending up to Nagaland across North Haryana, North Uttar Pradesh, Centre of Low-Pressure Area (northeast and adjoining Bihar), North Bihar, Sub-Himalayan West Bengal, and Sikkim and Assam.

Active to Vigorous Monsoon conditions are prevailing over Northeast India since the last many days. Assam, Meghalaya and Arunachal Pradesh, in particular, have been receiving heavy to very heavy rains.

The average rainfall during the Monsoon season for Sub-Himalayan West Bengal and Sikkim is around 2006 mm. It is more than the seasonal rainfall of Assam, Meghalaya and Arunachal Pradesh, wherein the Monsoon average rainfall for these pockets is less than 1800 mm.
During Monsoon, it rains the heaviest in Coastal Karnataka which is followed by Konkan and Goa, Kerala, Sub-Himalayan West Bengal, and Sikkim, Assam and Meghalaya, Arunachal Pradesh and then Nagaland, Manipur, Mizoram, and Tripura.

At the time of Monsoon, this pocket has a different pattern that is during Active Monsoon, it rains the least over here. However, during Break Monsoon, this pocket gets drenched like anything and this is what we are witnessing at present.

Hasimara in West Bengal records hefty rains during this period as well. The monthly average of this place exceeds 1000 mm for the months of July and August. For the next four days, torrential rains would continue over the region and these rains will only subside when any system would build up in the Bay of Bengal.

Picture 2: Location map of the study area

TIME SERIES:

A time series is a sequence of data points that occur in successive order over some period of time. This can be contrasted with cross sectional data, which captures a point-in-time and the time is ordered.

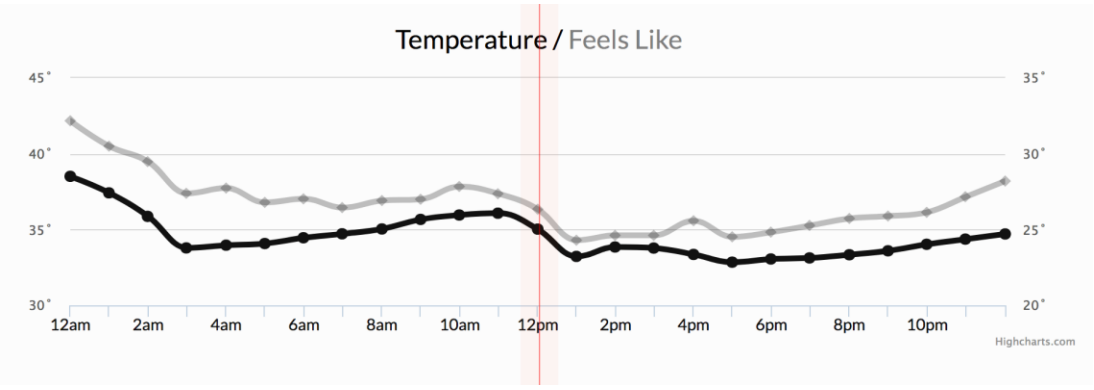
In investing, a time series tracks the movement of the chosen data points, such as a security's price, over a specified period of time with data points recorded at regular intervals. There is no minimum or maximum amount of time that must be included, allowing the data to be gathered in a way that provides the information being sought by the investor or analyst examining the activity.

➢ Goals of time series analysis:

1. Descriptive: Identify patterns in correlated data trends and seasonal variation.
2. Explanation: Understanding and modelling the data.
3. Forecasting: Prediction of short-term trends from previous patterns.
4. Intervention analysis: How does a single event change the time series.
5. Quality control: Deviations of a specified size indicate a problem.

Weather records, economic indicators and patient health evolution metrics — all are time series data. Time series data could also be server metrics, application performance monitoring, network data, sensor data, events, clicks and many other types of analytics data. Some real life examples are given below:

Notice how time — depicted at the bottom of the below chart — is the axis. Maximum temp of a place of a day.



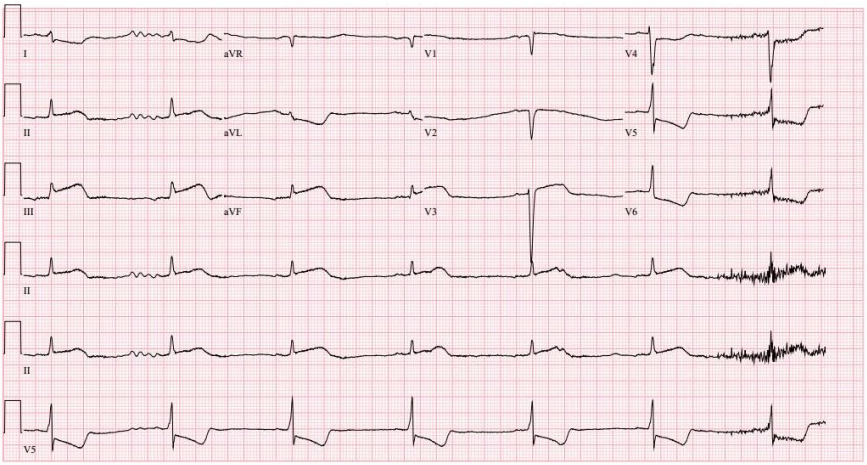
Example 1: Weather conditions

In the next chart below, note time as the axis over which stock price changes are measured. In investing, a time series tracks the movement of data points, such as a security's price over a

specified period of time with data points recorded at regular intervals. This can be tracked over the short term (such as a security's price on the hour over the course of a business day) or the long term (such as a security's price at close on the last day of every month over the course of five years).

Example 2: Stock Price Rate

Another familiar example of time series data is patient health monitoring, such as in an electrocardiogram (ECG), which monitors the heart's activity to show whether it is working normally.



Example 3: Heart Rate

➢ Uses of Time Series:

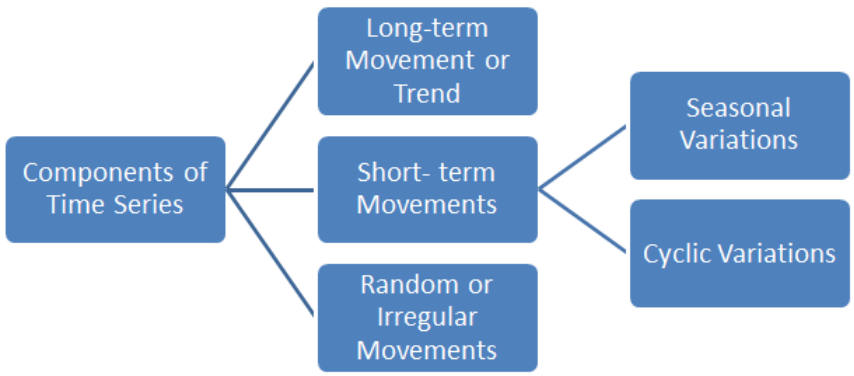
- The most important use of studying time series is that it helps us to predict the future behavior of the variable based on past experience
- It is helpful for business planning as it helps in comparing the actual current performance with the expected one.
- From time series, we get to study the past behavior of the phenomenon or the variable under consideration.
- We can compare the changes in the values of different variables at different times or places, etc.

➢ Components for Time Series Analysis:

The various reasons or the forces which affect the values of an observation in a time series are the components of a time series. The four categories of the components of time series are:

- Trend
- Seasonal Variations
- Cyclic Variations
- Random or Irregular movements

Seasonal and Cyclic Variations are the periodic changes or short-term fluctuations.



- **Trend:**

The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency. It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.

It is observable that the tendencies may increase, decrease or are stable in different sections of time. But the overall trend must be upward, downward or stable. The population, agricultural production, items manufactured, number of births and deaths, number of industry or any factory, number of schools or colleges are some of its examples showing some kind of tendencies of movement.

- **Periodic Fluctuations:**

There are some components in a time series which tend to repeat themselves over a certain period of time. They act in a regular spasmodic manner.

- **Seasonal Variations:**

These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

These variations come into play either because of the natural forces or man-made conventions. The various seasons or climatic conditions play an important role in seasonal variations. Such as production of crops depends on seasons, the sale of umbrella and raincoats in the rainy season, and the sale of electric fans and A.C. shoot up in summer seasons.

The effect of man-made conventions such as some festivals, customs, habits, fashions, and some occasions like marriage is easily noticeable. They recur themselves year after year. An upswing in a season should not be taken as an indicator of better business conditions.

- **Cyclic Variations:**

The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the 'Business Cycle'.

It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular are not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them.

- **Random or Irregular Movements:**

There is another factor which causes the variation in the variable under study. They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

➢ Mathematical Model for Time Series Analysis:

Mathematically, a time series is given as,

$y_t = f(t)$

Here, y_t is the value of the variable under study at time t. If the population is the variable under study at the various time period $t_1, t_2, t_3, \dots, t_n$. Then the time series is,

t: $t_1, t_2, t_3, \dots, t_n$

y_t: $y_{t1}, y_{t2}, y_{t3}, \dots, y_{tn}$

or, t: $t_1, t_2, t_3, \dots, t_n$

y_t: $y_1, y_2, y_3, \dots, y_n$

- **Additive Model for Time Series Analysis:**

If y_t is the time series value at time t. T_t, S_t, C_t, and R_t are the trend value, seasonal, cyclic and random fluctuations at time t respectively. According to the Additive Model, a time series can be expressed as

$y_t = T_t + S_t + C_t + R_t.$

This model assumes that all four components of the time series act independently of each other.

- **Multiplicative Model for Time Series Analysis:**

The multiplicative model assumes that the various components in a time series operate proportionately to each other. According to this model

$y_t = T_t \times S_t \times C_t \times R_t$

- **Mixed models:**

Different assumptions lead to different combinations of additive and multiplicative models as

$y_t = T_t + S_t + C_t R_t.$

The time series analysis can also be done using the model, $y_t = T_t + S_t \times C_t \times R_t$ or $y_t = T_t \times C_t + S_t \times R_t$ etc.

METHEDOLOGY:

The methodology adopted in this study is "Rainfall in Sub Himalayan West Bengal and Sikkim". From the study and analysis, variation in results among the plotting position methods is found to be insignificant.

In the present work various simple statistical procedures have been applied in order to reveal the monthly rainfall variability of 50 years of Sub Himalayan West Bengal and Sikkim. The process of fitting model and forecasting are mention below.

The following steps are given below:

1. Import Data
2. Converting to Time Series Object
3. Exploratory Time Series Data Analysis
4. Decompose
5. Extract the Random Part
6. Check stationary or not
7. Seasonal Plot
8. ACF
9. Use ARIMA Model on random part of the ts
10. Checking residuals for the model
11. Forecasting

DATA ANALYSIS:

A forecast is calculation or estimation of future events, especially for financial trends or coming weather. Until this year, forecasting was very helpful as a foundation to create any action or policy before facing any events. As an example, in the tropics region which several countries only had two seasons in a year (dry season and rainy season), many countries especially country which relies so much on agricultural commodities will need to forecast rainfall in term to decide the best time to start planting their products and maximizing their harvest. Another example is forecast can be used for a company to predict raw material prices movements and arranges the best strategy to maximize profit from it. We have 1965–2015 historical rainfall data and will try to fit a model with "R" Language.

Now, we discuss how we import our data set, how to convert data in time series data, how to fit a model and, forecasting on it; that shown below:

1. Import Data:

At first, we import our data set in R, through Notepad in **".txt"** format. Then we check the data type through **"class ()"** comment.

1.1 Code and Output:

```
> rain=scan ("C:/Users/Arun Karmakar/Downloads/Rainfall.txt")
Read 612 items
> class(rain)
[1] "numeric"
```

2. Converting to Time Series Object:

R has extensive facilities for analysing time series data. This section describes the **creation of a time series, seasonal decomposition**, **modelling with exponential and ARIMA models, and forecasting with the forecast package**.

The ***ts ()*** function will convert a numeric vector into an R time series object. The format is ***ts (vector, start=, end=, frequency=)*** where start and end are the times of the first and last observation and frequency is the number of observations per unit time **(1=annual, 4=quartly, 12=monthly, etc.).**

Now, we convert our data into time series object.

2.1 Code and Output:

```
>rain_ts <- ts (rain, frequency = 12, start = c (1965,1))
>rain_ts
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1965	0.2	22.0	48.2	69.5	232.4	444.6	697.6	830.2	345.7	39.5	32.7	0.2
1966	29.5	7.6	3.0	51.5	231.7	287.5	741.5	682.8	357.3	100.9	18.3	3.8
1967	2.3	0.1	98.8	85.3	213.7	727.5	901.6	274.1	409.4	128.7	7.9	11.1
1968	26.3	4.7	34.3	90.8	237.3	646.7	827.6	594.2	514.8	314.3	87.3	0.5
1969	11.5	2.8	45.1	77.5	273.3	504.8	550.5	448.2	431.2	45.8	34.1	0.0
1970	18.0	13.4	7.6	162.5	162.7	556.6	731.2	425.4	560.8	60.7	3.3	0.0
1971	28.4	21.3	63.0	239.8	244.5	757.6	599.5	542.0	375.4	328.7	24.8	1.9
1972	9.4	41.6	52.1	134.3	287.5	457.2	590.3	335.1	407.1	86.8	12.3	0.9
1973	13.8	35.9	40.8	82.9	324.3	679.8	426.5	424.7	474.1	281.3	16.6	3.9
1974	19.1	2.5	94.4	210.8	339.0	508.1	826.8	673.7	480.8	175.0	3.0	8.3
1975	20.4	29.9	15.6	83.7	251.1	493.2	808.2	263.8	515.7	174.6	1.0	6.6
1976	8.8	51.0	27.3	112.8	303.5	649.2	531.9	671.7	288.0	113.1	35.5	1.2
1977	6.6	16.0	67.6	245.6	308.1	466.0	502.9	627.8	316.0	263.7	60.8	25.3
1978	14.7	17.9	49.3	136.7	313.3	515.5	545.2	274.0	376.5	62.9	61.1	10.8
1979	8.1	26.8	15.1	100.6	142.0	261.7	741.7	432.4	427.0	286.5	31.7	38.1
1980	11.6	44.8	77.3	130.4	300.1	485.9	774.4	692.5	477.5	151.2	0.9	2.0
1981	48.8	23.0	77.8	185.6	273.7	468.6	833.1	540.9	453.2	38.1	9.2	17.3
1982	0.2	9.9	47.0	125.6	174.5	499.0	749.5	305.5	341.6	54.1	20.0	10.6
1983	26.7	56.9	35.3	87.7	335.3	557.9	758.4	295.6	535.8	125.1	3.2	20.8
1984	29.3	11.3	42.6	127.3	360.1	474.6	737.2	364.8	405.5	179.1	5.1	11.5
1985	2.2	30.5	43.2	69.5	358.6	532.7	907.4	372.7	464.5	211.9	17.4	27.3
1986	1.3	5.8	7.9	122.4	152.1	436.4	545.4	363.2	493.9	215.1	15.5	6.4
1987	2.8	35.3	79.8	152.7	148.0	451.0	791.9	836.6	530.3	150.3	5.6	1.2
1988	6.8	24.6	71.2	96.5	268.8	304.9	831.7	990.5	446.6	76.0	18.4	5.5
1989	21.0	41.6	45.8	58.1	358.4	551.3	629.8	407.8	593.6	98.9	30.7	17.6
1990	3.8	109.9	80.1	158.3	358.2	554.7	814.1	916.6	440.6	115.0	0.4	13.7
1991	41.5	24.2	45.8	102.7	283.7	691.4	592.6	500.9	689.5	67.8	1.7	30.4
1992	10.0	46.7	26.0	84.0	253.4	328.1	736.3	458.0	276.4	125.5	5.0	9.5
1993	37.6	35.2	58.8	136.9	320.5	449.6	723.9	521.6	352.2	228.8	28.4	2.4
1994	49.6	55.0	114.1	126.6	243.3	388.0	357.3	396.1	302.1	84.3	8.5	2.2
1995	17.9	41.1	50.6	84.2	349.6	859.0	775.3	539.2	638.5	99.7	88.6	14.0
1996	30.9	34.2	59.5	85.0	401.7	334.6	863.6	545.1	337.9	107.3	0.4	0.0
1997	18.9	31.3	76.4	119.5	166.9	613.3	488.1	479.0	443.9	41.3	16.3	56.5
1998	13.6	37.0	125.1	154.1	254.2	653.0	854.6	860.4	478.9	209.1	13.1	2.1
1999	8.7	6.2	17.9	156.0	335.4	550.5	831.7	751.6	404.1	246.0	10.7	4.5
2000	10.7	36.2	55.1	185.1	326.5	649.6	574.3	498.1	465.8	82.9	22.6	1.2
2001	4.0	20.5	50.2	134.8	347.2	472.2	399.0	424.5	434.9	282.8	36.6	5.9
2002	30.1	10.0	95.6	237.1	181.7	407.2	743.3	406.6	340.1	75.7	8.7	10.8
2003	17.0	79.2	90.1	171.4	211.7	516.6	703.0	354.0	321.4	223.0	16.4	19.2
2004	19.9	11.5	40.5	163.5	287.5	489.5	739.5	356.4	394.7	198.4	9.0	4.4
2005	19.5	21.0	132.1	155.5	254.0	446.5	704.3	564.5	229.9	271.1	8.8	0.5
2006	0.7	17.0	39.8	119.3	270.5	446.2	470.2	295.0	456.8	118.4	20.3	7.9
2007	3.1	81.1	46.6	160.1	215.1	485.1	686.3	412.4	458.6	79.5	8.8	2.0
2008	33.2	22.1	79.8	141.8	171.4	526.5	657.7	636.3	296.7	97.9	16.9	9.5
2009	6.0	40.3	55.0	115.2	308.7	350.6	456.1	568.4	174.6	235.1	9.1	5.6
2010	5.6	19.6	77.6	176.6	335.9	558.1	593.4	461.3	308.1	66.2	7.9	2.2
2011	8.5	19.9	71.2	135.0	247.8	419.8	612.3	470.3	356.3	46.7	26.7	4.3
2012	15.3	13.9	45.5	159.8	202.4	604.2	684.5	332.7	434.7	119.4	12.5	7.4
2013	3.0	23.6	32.1	114.7	296.5	404.9	588.4	416.3	308.0	199.8	16.1	2.7
2014	0.2	26.6	37.7	47.9	308.6	543.2	384.6	563.3	371.5	31.2	5.3	2.4
2015	15.7	15.0	64.8	149.0	304.6	508.2	393.3	626.6	354.9	53.6	23.8	9.0

3. Exploratory Time Series Data Analysis:

We have to visualize our rainfall data through time series plot(Line chart, values against time).

3.1. Plot the time series data:

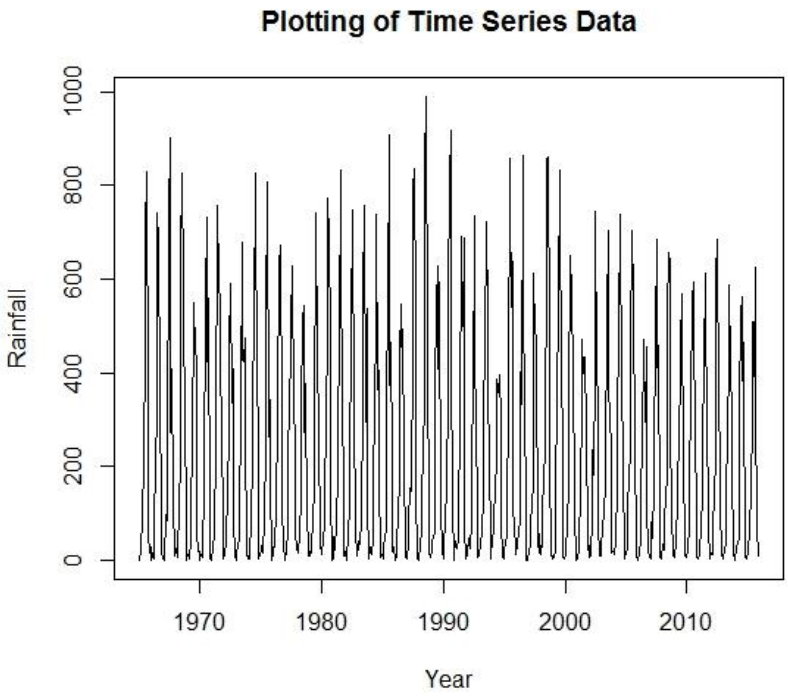
Now we are plotting Time Series Rainfall Data.

3.1.1. Code:

```
>plot(rain_ts)

>plot(rain_ts,
+  main = "Plotting of Time Series Data",
+ xlab = "Year",
+ ylab = "Rainfall")
```

3.1.2. Output(Diagram-1):



3.1.3. Interpretation(Diagram-1):

In the above diagram, x-axis shows **"year"** and y-axis shows **"Rainfall(mm)"**. This time series plot**(Rainfall in Sub Himalayan West Bengal and Sikim:1965-2015)** visualizes that rainfall has **seasonality pattern without any trends** occurred.

4. Decompose:

This is a useful abstraction. **Decomposition** is primarily used for time series analysis, and as an analysis tool it can be used to inform forecasting models on our problem.

It provides a structured way of thinking about a time series forecasting problem, both generally in terms of modelling complexity and specifically in terms of how to best capture each of these components in a model.

Each of these components are something we may need to think about and address during data preparation, model selection, and model tuning. We may address it explicitly in terms of modelling the trend and subtracting it from our data, or implicitly by providing enough history for an algorithm to model a trend if it may exist.

We may or may not be able to cleanly or perfectly break down our specific time series as an additive or multiplicative model.

Real-world problems are messy and noisy. There may be additive and multiplicative components. There may be an increasing trend followed by a decreasing trend. There may be non-repeating cycles mixed in with the repeating seasonality components.

Nevertheless, these abstract models provide a simple framework that we can use to analyse our data and explore ways to think about and forecast our problem.

Now, **we decompose the original data into trend, seasonal and random part.**

4.1. Code:

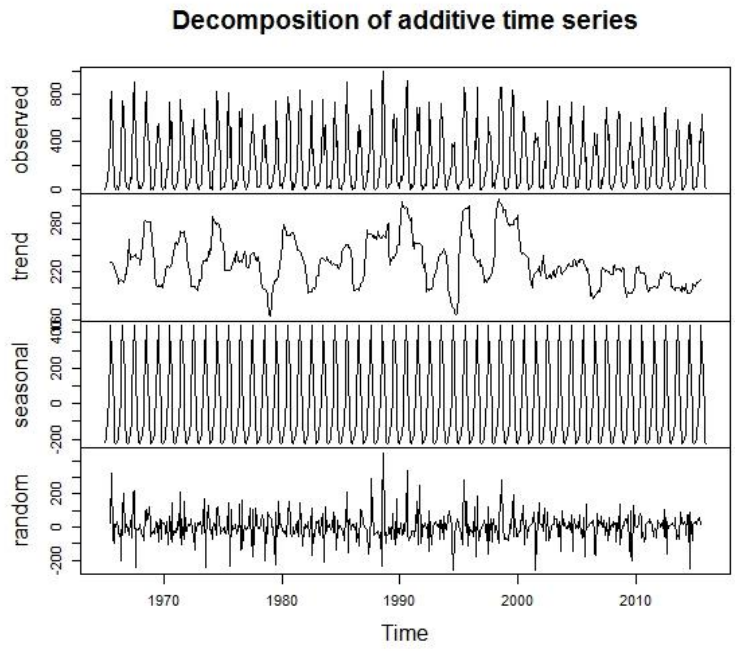
```
Yt=decompose(rain_ts)
```

4.2. Plot of Decomposition:

4.2.1. Code:

```
>plot(Yt)
```

4.2.2. Output(Diagram-2):



4.2.3. Interpretation(Diagram-2):

We have decomposed our time series data into more details based on **Trend, Seasonality, and Random component**. We have to gain more precise insight into rainfall behavior during the period 1965-2015.

5. Extract the Random Part:

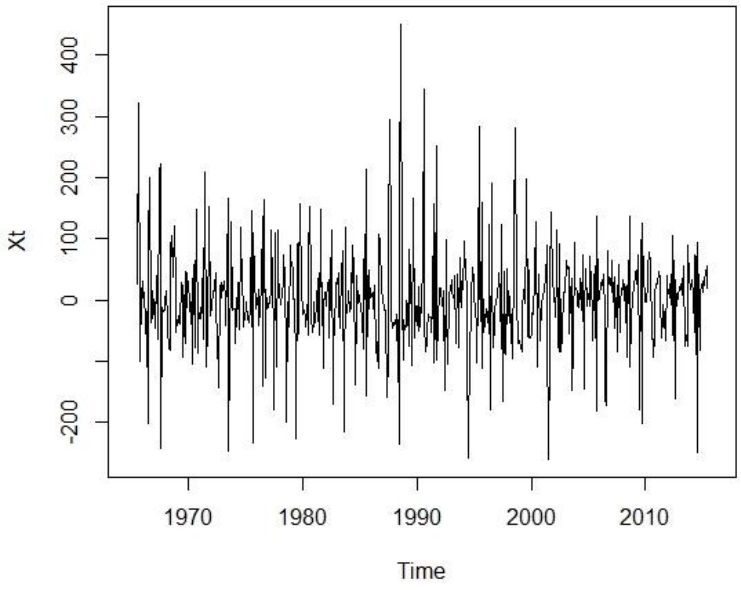
Let's extract the random part now.

5.1.1Code:

```
>Xt=Yt$random

>plot(Xt)
```

5.1.2. Output(Diagram-3):



5.1.3. Interpretation:(Diagram-3)

The above diagram, looks like white noise. So, it is stationary and we may confidently fit Stochastic models on it.

Lag Plot:

A lag plot checks whether a data set or time series is random or not. Random data should not exhibit any identifiable structure in the lag plot. Non-random structure in the lag plot indicates that the underlying data are not random. Several common patterns for lag plots are shown in the example below:

A lag is a fixed time displacement. For example, given a data set Y_1, Y_2, …, Y_n; $Y_2$and Y_7 have lag 5 since 7-2=5. Lag plots can be generated for any arbitrary lag, although the most commonly used lag is 1.

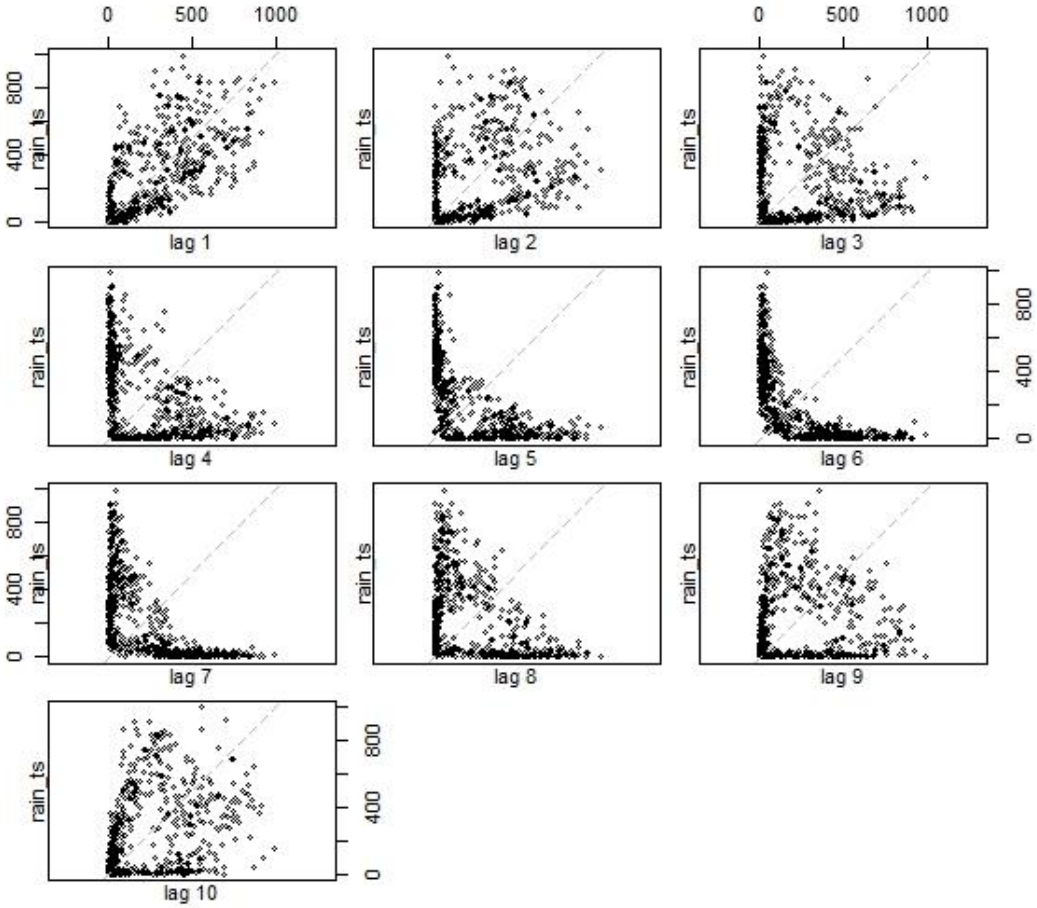
A plot of lag 1 is a plot of the values of Y_i versus Y_{i-1}.

- Vertical axis: Y_i for all i.
- Horizontal axis: $Y_{i-1}$for all i.

5.2.1. Code:

```
>lag.plot(rain_ts,lag=10)
```

5.2.2. Output(Diagram-4):



5.2.3.Interpretation:(Diagram-4):

In the above diagram we plot lag 1 to lag 10. By comparing all the lag plots we can conclude that, as the lag is increases the auto correlation decreases.

6. Check stationary or not:

Stationarity:

Stationarity is an important concept in the field of time series analysis with tremendous influence on how the data is perceived and predicted. When forecasting or predicting the future, most time series models assume that each point is independent of one another. The best indication of this is when the dataset of past instances is stationary. For data to be stationary, the statistical properties of a system do not change over time. This does not mean that the values for each data point have to be the same, but the overall behaviour of the data should remain constant. From a purely visual assessment, time plots that do not show trends or seasonality can be considered stationary. More numerical factors in support of stationarity include a constant mean and a constant variance.

Trend = When there is a long-term increase or decrease in the data.

Seasonality = Reoccurring pattern at a fixed and known frequency based on a time of the year, week, or day.

Now we check the random part is stationary or not by Augmented Dickey Fuller Test.

Augmented Dickey-Fuller Test (ADF):

In statistics and econometrics, an **Augmented Dickey–Fuller test** (**ADF**) tests the null hypothesis that a unit root is present in a time series sample. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models.

The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.

Testing Procedure:

The testing procedure for the ADF test is the same as for the Dickey–Fuller test but it is applied to the model,

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \ldots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t ,$$

Where α is a constant, β the coefficient on a time trend and *p*the lag order of the autoregressiveprocess.Imposing the constraints α=0 and β=0 correspondsto modelling a random walk and using the constraint β=0 corresponds to modelling a random walkwith a drift.

By including lags of the order *p* the ADF formulation allows for higher-order autoregressive processes. This means that the lag length *p* has to be determined when applying the test. One possible approach is to test down from high orders and examine the *t*-values on coefficients. An alternative approach is to examine information criteriasuch as the Akaikeinformation criterion, Bayesian information criterion or the Hannan–Quinn information criterion.

The unit root test is then carried out under the null hypothesis γ=0 against the alternative hypothesis of γ<0. Once a value for the test statistic,

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$$

is computed it can be compared to the relevant critical value for the Dickey–Fuller test. As this test is asymmetrical, we are only concerned with negative values of our test statistic DF_τ. If the calculated test statistic is less (more negative) than the critical value, then the null hypothesis of γ=0 is rejected and no unit root is present.

6.1. Code and Output:

```
>install. Packages("tseries")
>library(tseries)
Registered S3 method overwritten by 'quantmod':
  method            from
as.zoo.data.frame zoo

    'tseries' version: 0.10-48

    'tseries' is a package for time series analysis and computational
    finance.

    See 'library (help="tseries")' for details.

Warning message:
package 'tseries' was built under R version 3.6.3
```

```
>adf.test(Xt[7:606])

        Augmented Dickey-Fuller Test

data:  Xt[7:606]
Dickey-Fuller = -15.345, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(Xt[7:606]) : p-value smaller than printed p-value
```

6.2. Interpretation:

We know that the series is stationary if p-value is less than 0.05.
Here, our p-value is 0.01 that is less than 0.05. Therefore, the series is stationary.

We know that our data has a seasonality pattern. So, to explore more about our rainfall data seasonality; seasonal plot, seasonal-subseries plot, and seasonal box plot will provide a much more insightful explanation about our data.

7. Seasonal Plot:

Seasonal plots are a graphical tool to visualize and detect seasonality in a time series. Seasonal subseries plots involve the extraction of the seasons from a time series into asubseries. Based on a selected periodicity, it is an alternative plot that emphasizes the seasonal patterns are where the data for each season are collected together in separate mini time plots.

Seasonal plots enable the underlying seasonal pattern to be seen clearly, and also show the changes in seasonality over time. Especially, it allows to detect changes between different seasons, changes within a particular season over time.

However, this plot is only useful if the period of the seasonality is already known. In many cases, this will in fact be known. For example, monthly data typically has a period of 12. If the period is not known, an autocorrelation plot or spectral plot can be used to determine it. If there is a large number of an observation, then a **box plot** may be preferable.

Seasonal sub-series plots are formed by

- Vertical axis: response variable.
- Horizontal axis: time of year.

The horizontal line displays the mean value for each month over the time series.

The analyst must specify the length of the seasonal pattern before generating this plot. In most cases, the analyst will know this from the context of the problem and data collection

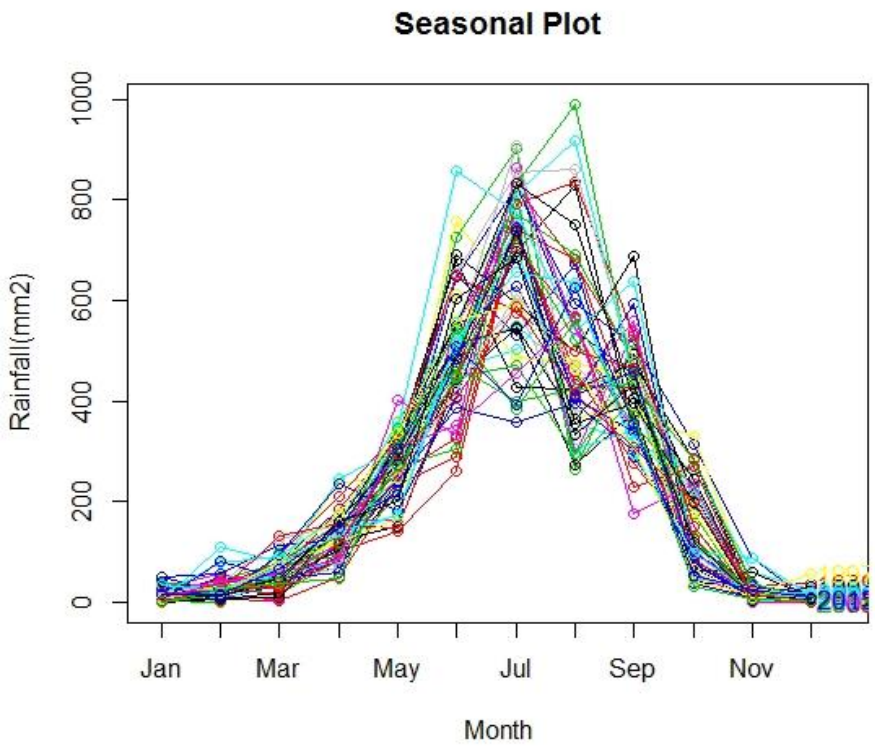
7.1.1.Code:

```
>install. Packages("forecast")

>library(forecast)

>seasonplot(rain_ts,year.labels = TRUE,col=1:13,
+ main="Seasonal Plot",ylab="Rainfall(mm2)")
```

7.1.2 Output(Diagram-5):



7.1.3.Interpretation(Diagram-5):

The seasonal plot (the above diagram) indeed shows a seasonal pattern that occurred

each year (1965-2015).

7.2. Seasonal Box plot:

Using seasonal box plot, we can more clearly see the data pattern.

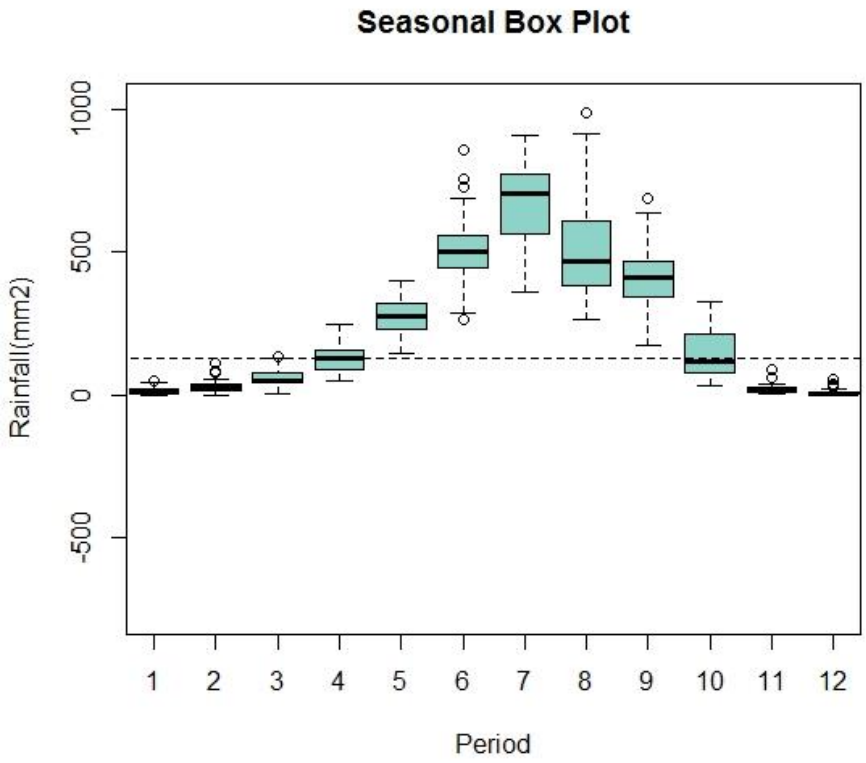
7.2.1. Code:

```
>install. Packages("tsutils")


>library(tsutils)

>seasplot(rain_ts,outplot=2,trend=FALSE,
+ main="Seasonal Box Plot",ylab="Rainfall(mm2)")
Results of statistical testing
Presence of trend not tested.
Evidence of seasonality: TRUE  (pval: 0)
```

7.2.2. Output(Diagram-6):



7.2.3. Interpretation (Diagram-6):

From the above diagram, the horizontal line indicates the rainfall value, means grouped by month; with using this information we have got the insight that, the rainfall (mm) is started in the month of May and it's going to increase. It takes the peak value on July and then it started to decrease in the month of October.

8. ACF:

ACF plot:

A **time series** is a sequence of measurements of the same variable(s) made over time. Usually, the measurements are made at evenly spaced times — for example, monthly or yearly. The coefficient of correlation between two values in a time series is called the **autocorrelation function** (**ACF**). In other words,

>Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.
>Autocorrelation measures the relationship between a variable's current value and its Past values.
>An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of negative 1 represents a perfect negative correlation.

ACF is useful because:

Help us uncover hidden patterns in our data and help us select the correct forecasting methods.

1. Help identify seasonality in our time series data.

2. Analysing the autocorrelation function (ACF) and partial autocorrelation function (PACF) in conjunction is necessary for selecting the appropriate ARIMA model for any time series prediction.

3. Assumption made by ACF: Weak stationary — meaning no systematic

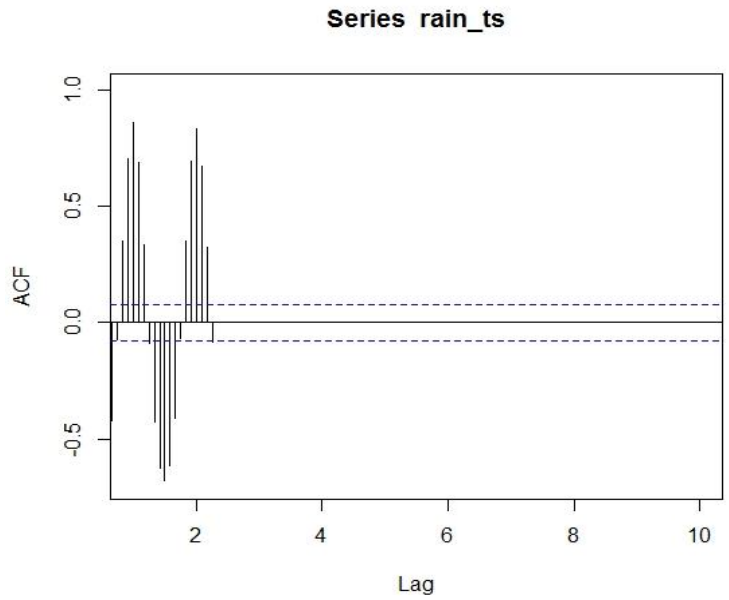
change in the mean, variance, and no systematic fluctuation. So, when performing ACF it is advisable to remove any trend present in the data and to make sure the data is stationary.

The first step in building the ARIMA model is to create an autocorrelation plot on stationary random part of the time series data.

8.1. Code:

```
>acf(rain_ts,na.action=na.pass,xlim=c(1,10))
```

8.2. Output(Diagram-7):



8.3.Interpretation (Diagram-7):

From the above ACF plot we get MA parameter q. There is a significant spike at lag 3. This ACF plot suggests that the appropriate model might be ARIMA (0,0,3). The dashed blue lines indicate the 95% confidence interval for the correlations are significantly different from zero.

9. Use ARIMA Model on random part of the ts:

Moving Average Method:

The moving average method is an improvement over the semi-average method and short-term fluctuations are eliminated by it. A moving average is defined as an average of fixed number of items in the time series which move through the series by dropping the top items of the previous averaged group and adding the next in each successive average.

Let (t_1, y_1), (t_2, y_2), …, (t_n, y_n) denote given time series $y_1, y_2, \ldots, y_n$ are the values of the variable y; corresponding to time periods $t_1, t_2, \ldots, t_n$, respectively.

The moving averages of order m are defined as,

$(y_1+y_2+\ldots+y_m)/m$; $(y_1+y_2+\ldots+y_{m+1})/m$;

Here $y_1+y_2+\ldots+y_m, y_2+y_3+\ldots+y_{m+1}, \ldots$ are called moving totals of m.

In using moving averages in estimating the trend, we shall have to decide as what should be the order of the moving averages. The order of the moving average should be equal to the length of the cycles in the time series. In case the order of the moving averages is given in the problem itself, then we shall use that order for computing the moving average. The order of the moving averages may either be odd or even.

The moving averages of order 3 are,

$(y_1+y_2+y_3)/3$; $(y_2+y_3+y_4)/3$;…; $(y_{n-2}+y_{n-1}+y_n)/3$

These moving averages are called the trend values. They are considered to correspond 2nd, 3rd… $(n-1)$ th years, respectively. Calculation of trend values by using moving averages of even order is slightly complicated. The following steps are involved in the method:

Step 1: In the first step, a group of beginning years (periods), which constitute cycle, is chosen for calculating the average. This average is placed in front of the mid-year of the group.

Step 2: Now delete the first-year value from the group and add a succeeding year value in the group. Find the average of the reconstituted group and place it in front of this group.

Step 3: If the number of years in a group is odd, middle year is located without any problem. But if the number of years in the group is even, the average of the averages in pairs is calculated and placed against the mid-year of the two.

Step 4: Repeat the Step 2 till all years of the data are exhausted.

Step 5: The moving averages calculated are considered as an artificially constructed time series.

Step 6: Plot the moving averages on a graph paper taking years along *x*-axis and moving averages along *y*-axis by choosing a proper scale.

Step 7: Join the plotted point in the sequence of time periods. The resulting graph provides the trend.

AIC:

The AkaikeInformation Criterion (AIC) is an estimator of out of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

In plain words, AIC is a single number score that can be used to determine which of multiple models is most likely to be the best model for a given dataset. It estimates models *relatively*, meaning that AIC scores are only useful in comparison with other AIC scores for the same dataset. A lower AIC score is better.

AIC is most frequently used in situations where one is not able to easily test the model's performance on a test set in standard machine learning practice (small data, or time series). AIC is particularly valuable for time series, because time series analysis' most valuable data is often the most recent, which is stuck in the validation and test sets. As a result, training on all the data and using AIC can

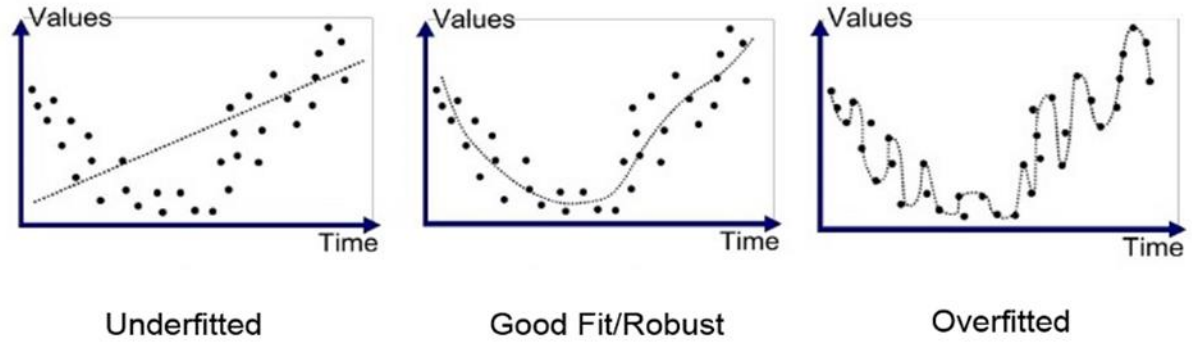
result in improved model selection over traditional train/validation/test model selection methods.

AIC works by evaluating the model's fit on the training data, and adding a penalty term for the complexity of the model (similar fundamentals to regularization). The desired result is to find the lowest possible AIC, which indicates the best balance of model fit with generalizability. This serves the eventual goal of maximizing fit on out-of-sample data.

AIC = -2ln(L)+2k,

AIC equation, where L = likelihood and k = number of parameters

AIC uses a model's maximum likelihood estimation (log-likelihood) as a measure of fit. Log-likelihood is a measure of how likely one is to see their observed data, given a model. The model with the maximum likelihood is the one that "fits" the data the best. The natural log of the likelihood is used as a computational convenience.
AIC is low for models with high log-likelihoods (the model fits the data better, which is what we want), but adds a penalty term for models with higher parameter complexity, since more parameters means a model is more likely to overfit to the training data.



The over fit model maximizes log-likelihood, since all data points fall exactly on the model's prediction. Penalizing parameter complexity counterbalances this, and leads to better fit.

Now we fit an appropriate MA model for my dataset.

9.1.Code and Output:

Fitting a Model:

```
>model_MA=arima(Xt, order=c(0,0,3))
>model_MA

Call:
arima(x = Xt, order = c(0, 0, 3))

Coefficients:
          ma1      ma2      ma3  intercept
      -0.3616  -0.4030  -0.2354     0.0082
s.e.   0.0386   0.0451   0.0375     0.0315

sigma^2 estimated as 5129:  log likelihood = -3416.96,  aic = 6843.91
```

9.2. Interpretation:

Here, order = c (p, d, q). If q=1, then the model is MA1; q=2, then the model is MA2; q=3, then the model is MA3.

We will set other model based on our suggestion with modifying MA.
To choose the best fit among all of the ARIMA models for our data, we will compare AIC value between those models. In our model, AIC(MA1)=6971.34, AIC(MA2)=6877.83, AIC(MA3)=6843.91.

The model with minimum AIC often is the best model for forecasting. Therefore, for our data MA3 has least AIC value, so MA3 is the bestmodel for forecasting for my dataset.

10. Checking residuals for the model:

Fitted values:

Each observation in a time series can be forecast using all previous observations. We call these **fitted values** and they are denoted by $\hat{y}_{t|t-1}$, meaning the forecast of y_t based on observations $y_1, \ldots, y_{t-1}$. We use these so often, we sometimes drop part of the subscript and just write $\hat{y}_t$ instead of $\hat{y}_{t|t-1}$. Fitted values always involve one-step forecasts.
Actually, fitted values are often not true forecasts because any parameters involved in the forecasting method are estimated using all available observations in the time series, including future observations. For example, if we use the average method, the fitted values are given by,

$$\hat{y}_t = \hat{c}$$

Where, $\hat{c}$ is the average computed over all available observations, including those at times *after* t. Similarly, for the drift method, the drift parameter is estimated using all available observations. In this case, the fitted values are given by,

$$\hat{y}_t = \hat{y}_{t-1} + \hat{c}$$

Where, $\hat{\mathbf{c}} = (\mathbf{y_T} - \mathbf{y_1})/(\mathbf{T} - \mathbf{1})$**.**In both cases, there is a parameter to be estimated from the data. The "hat" above the cc reminds us that this is an estimate. When the estimate of cc involves observations after time *t*, the fitted values are not true forecasts. On the other hand, forecasts do not involve any parameters, and so fitted values are true forecasts in such cases.

Residuals:

The "residuals" in a time series model are what is left over after fitting a model. For many (but not all) time series models, the residuals are equal to the difference between the observations and the corresponding fitted values:

$$e_t = y_t - \hat{y}_t$$

Residuals are useful in checking whether a model has adequately captured the information in the data. A good forecasting method will yield residuals with the following properties:

1. The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts.
2. The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.
3. Any forecasting method that does not satisfy these properties can be improved. However, that does not mean that forecasting methods that satisfy these properties cannot be improved. It is possible to have several different forecasting methods for the same data set, all of which satisfy these properties. Checking these properties is important in order to see whether a method is using all of the available information, but it is not a good way to select a forecasting method.

If either of these properties is not satisfied, then the forecasting method can be modified to give better forecasts. Adjusting for bias is easy: if the residuals have mean mm, then simply add mm to all forecasts and the bias problem is solved.

In addition to these essential properties, it is useful (but not necessary) for the

residuals to also have the following two properties.

3. The residuals have constant variance.
4. The residuals are normally distributed.

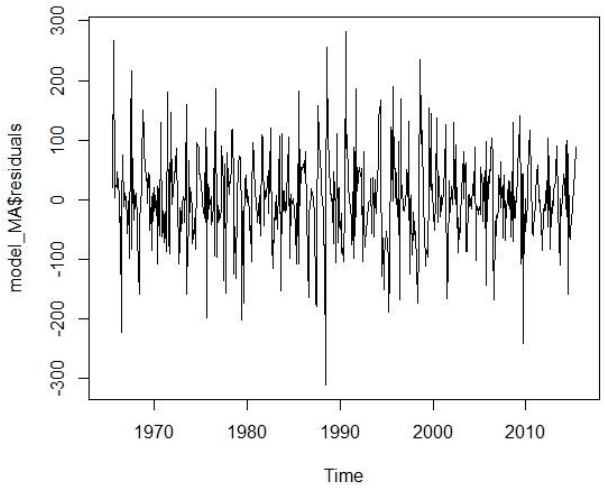
These two properties make the calculation of prediction intervals easier. However, a forecasting method that does not satisfy these properties cannot necessary be improved. Sometimes applying a Box-Cox transformation may assist with these properties, but otherwise there is usually little that you can do to ensure that your residuals have constant variance and a normal distribution. Instead, an alternative approach to obtaining prediction intervals is necessary. Again, we will not address how to do this until later in the book.

We also need to have residuals checked for this model to make sure this model will be appropriate for our time series forecasting.

10.1.1.Code:

```
>plot(model_MA$residuals)
```

10.1.2. Output(Diagram-8):



10.1.3. Interpretation (Diagram-8):

From the above diagram, we cannot interpret anything clearly. It looks like white noise.

ACF plot of Residuals:

With time series data, it is highly likely that the value of a variable observed in the current time period will be similar to its value in the previous period, or even the period before that, and so on. Therefore, when fitting a regression model to time series data, it is common to find autocorrelation in the residuals. In this case, the estimated model violates the assumption of no autocorrelation in the errors, and our forecasts may be inefficient — there is some information left over which should be accounted for in the model in order to obtain better forecasts. The forecasts from a model with auto correlated errors are still unbiased, and so are not "wrong," but they will usually have larger prediction intervals than they need to. Therefore, we should always look at an ACF plot of the residuals.

Another useful test of autocorrelation in the residuals designed to take account for the regression model is the **Breusch-Godfrey** test, also referred to as the LM (Lagrange Multiplier) test for serial correlation. It is used to test the joint hypothesis that there is no autocorrelation in the residuals up to a certain specified order.

A small p-value indicates there is significant autocorrelation remaining in the residuals.

The Breusch-Godfrey test is similar to the Ljung-Box test, but it is specifically designed for use with regression models.

Ljung-Box Test:

The Box-Ljungtest (1978) is a diagnostic tool used to test the lack of fit of a time series model.
The test is applied to the residuals of a time series after fitting an ARMA *(p,q)* model to the data. The test examines m autocorrelations of the residuals .If the autocorrelations are very small; we conclude that the model does not exhibit significant lack of fit.
In general, the Box-Ljung test is defined as:

H_0 : The model does not exhibit lack of fit

H_α :The model exhibits lack of fit.

Test: Given a time series Y of length n.
Statistic: The test Statistic is defined as

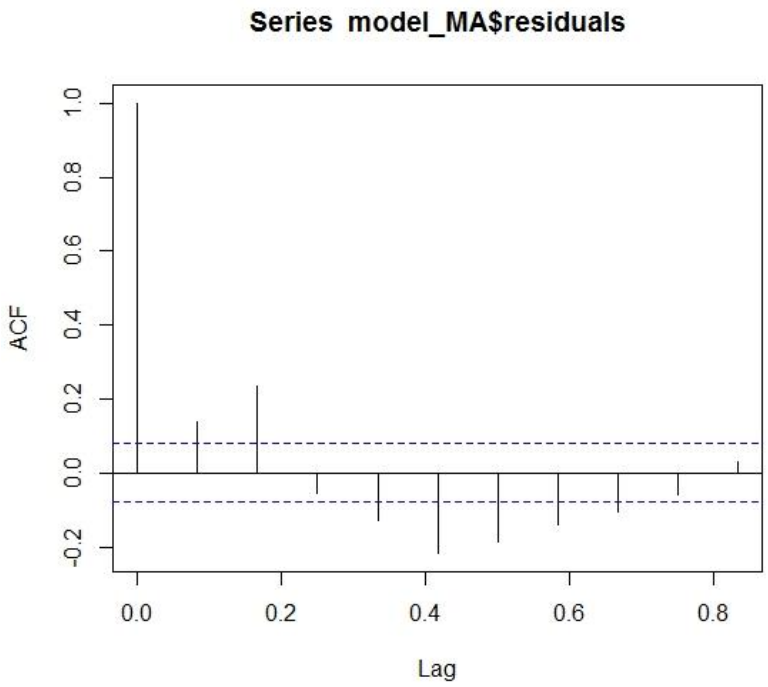
$$\boldsymbol{Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{r}^2{}_k}{n-k}}$$

Where,$\hat{r}_k$is the estimated autocorrelation of the series at lag *k* , and *m* is the number of lags being tested.
Significance Level: α
Critical Region : The Box- Ljung test rejects the null hypothesis(indicating that the model has significant lack of fit) if
$\boldsymbol{Q > \chi^2_{1-\alpha,h}}$
Where,$\chi^2_{1-\alpha,h}$ is the chi-square distribution table value with h degrees of freedom and significance level α.
Because the test is applied to residuals, the degrees of freedom must account for the estimated model parameters so that $h = m – p – q$, where *p* and *q* indicate the number of parameters from the ARMA (*p, q)* model fit to the data.

10.2.1. Code:

```
>acf(model_MA$residuals,na.action = na.pass,lag.max = 10)
```

10.2.2. Output(Diagram-9):



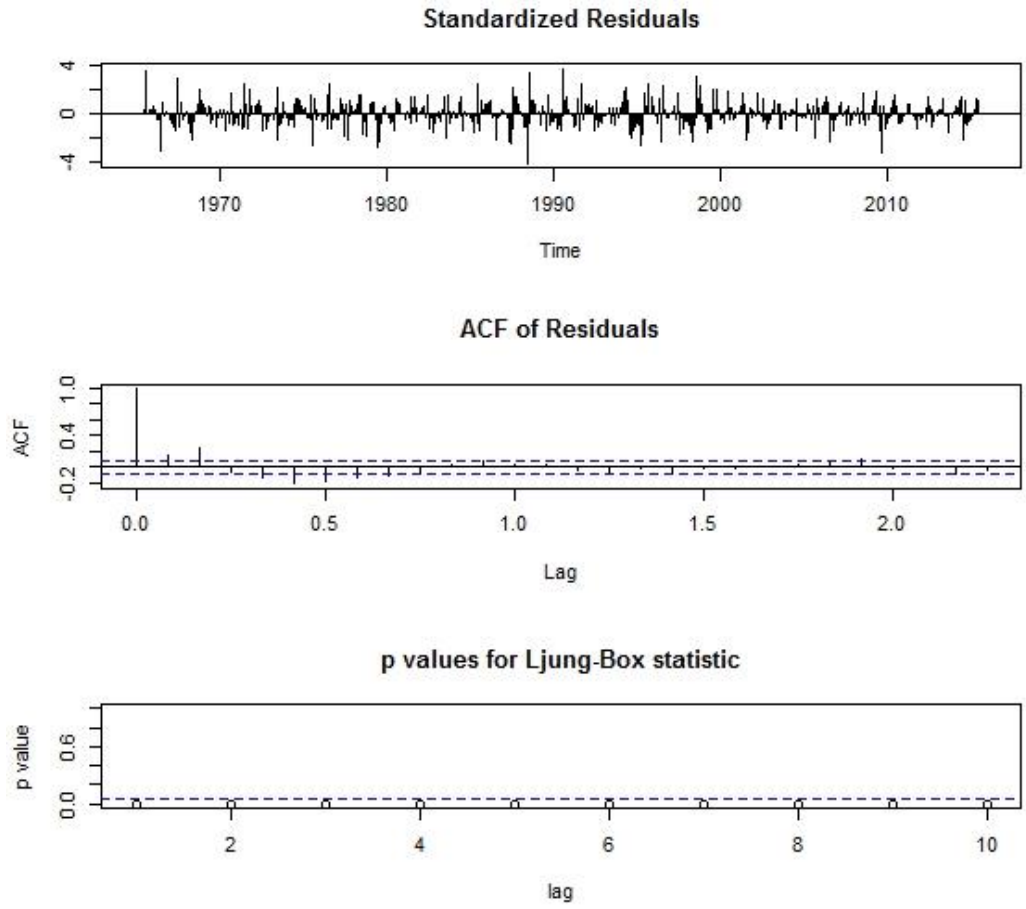
10.2.3. Interpretation:(Diagram-9)

From the ACF plot we can conclude that, the auto correlations are not significant with 95% confident.

10.3.1. Code:

```
>tsdiag(model_MA)
```

10.3.2. Output(Diagram-10):



10.3.3. Interpretation (Diagram-10):

Based on the Ljung-Box diagram we can say that the errors are random, and from the ACF plot of model residuals, we can conclude that this model is appropriate for forecasting.Since, its residuals show white noise behavior and uncorrelated against each other.

11.Forecasting:

Time series forecasting is a technique for the prediction of events through a sequence of time. The technique is used across many fields of study, from the geology to behavior to economics. The techniques predict future events by analyzing the trends of the past, on the assumption that future trends will hold similar to historical trends.

Making predictions about the future is called extrapolation in the classical statistical handling of time series data.

More modern fields focus on the topic and refer to it as time series forecasting.

Forecasting involves taking models fit on historical data and using them to predict future observations.

Descriptive models can borrow for the future (i.e., to smooth or remove noise), they only seek to best describe the data.

An important distinction in forecasting is that the future is completely unavailable and must only be estimated from what has already happened.

The purpose of time series analysis is generally twofold: to understand or model the stochastic mechanism that gives rise to an observed series and to predict or forecast the future values of a series based on the history of that series.

The skill of a time series forecasting model is determined by its performance at predicting the future. This is often at the expense of being able to explain why a specific prediction was made, confidence intervals and even better understanding the underlying causes behind the problem.

11.1.1. Code and Output:

```
>predict(model_MA,n.ahead = 12)
$pred
             Jan          Feb          Mar          Apr          May          Jun
2016 0.008171958 0.008171958 0.008171958 0.008171958 0.008171958 0.008171958
             Jul          Aug          Sep          Oct          Nov          Dec
2016 0.008171958 0.008171958 0.008171958 0.008171958 0.008171958 0.008171958
```

```
$se
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
2016 83.16851 83.16851 83.16851 83.16851 83.16851 83.16851 83.16851 83.16851
          Sep      Oct      Nov      Dec
2016 83.16851 83.16851 83.16851 83.16851
```

11.1.2. Interpretation:

We need to fix some further complex model for the data to get some better forecasting. But that is beyond our scope of study. So, I fail to predict a better forecasting result. I would further want to proceed this project in future.

CONCLUSION:

From my project "Time Series Data Analysis on Rainfall in Sub Himalayan West Bengal and Sikkim" we can know about the rainfall habit of Sub Himalayan West Bengal and Sikkim. If we have a look on all above graphs, we clearly conclude that, the rainfall (mm) in that region is mainly started in the month of May and it's going to increase. It takes the peak value on July and then it started to decrease in the month of October. We decompose our data and extract the random part, where we find some white noise. Then we check stationarity of our data set, and we come to a conclusion that our data is stationary and we can fit stochastic model on it. By using ARIMA model we fit MA 3 model. Then we check the residuals and conclude that, the errors are random and auto correlations are not significant with 95% confident.

We can't get a proper forecast from this model, as we work on only stochastic part, due to our study and syllabus limitations. We can fit more appropriate and suitable model for it and we can get better forecasting from that. I would further want to proceed this project in future.

REFERENCE:

1. Fundamental of Statistics (Volume-2) by A.M. Gun, M.K. Gupta & B. Dasgupta (Published by D. Chakraborty for The World Press Private Limited, Kolkata)

2. Fundamental of Applied Statistics by S.C. Gupta & V.K. Kapoor (Publisher: Sultan Chand & Sons Educational Publishers, New Delhi)

3. Time Series Analysis: With Applications in R (Book by Jonathan Cryer and Kung-sik Chan)

4. https://www.sciencedirect.com/topics/engineering/moving-average

5. https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced

6. https://www.toppr.com/guides/business-mathematics-and-statistics/time-series-analysis/components-of-time-series/

7. https://www.influxdata.com/what-is-time-series-data/#:~:text=Time%20series%20examples,other%20types%20of%20analytics%20data

APPENDIX:

Rainfall data of Sub Himalayan West Bengal and Sikkim

DATA SOURCE:

https://data.gov.in/catalog/rainfall-india?filters%5Bfield_catalog_reference%5D=1090541&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc