BIG DATA PROJECT

Study of Change in Annual Rainfall in Different Region

Group Members (Batch 1)

**1) Chaitali Kotasthane**

**2) Kunal Mandalya**

**3) Ruchit Furia**

**4) Mihir Makwana**

**5) Kaustubh Gaikwad**

Various Graph Representations

# 1) Bar Plot

The basic syntax to create a bar-chart in R is:

barplot(H, xlab , ylab ,main, names.arg ,col)

Following is the description of the parameters used:

• H is a vector or matrix containing numeric values used in bar chart.

• xlab is the label for x axis.

• ylab is the label for y axis.

• main is the title of the bar chart.

• names.arg is a vector of names appearing under each bar.

• col is used to give colors to the bars in the graph.

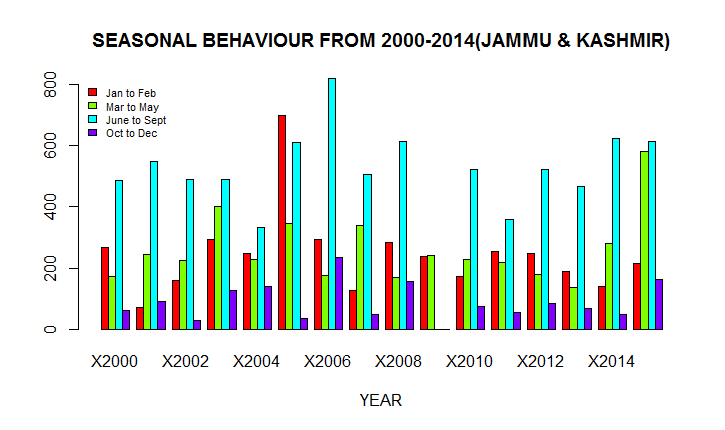


Fig. 1

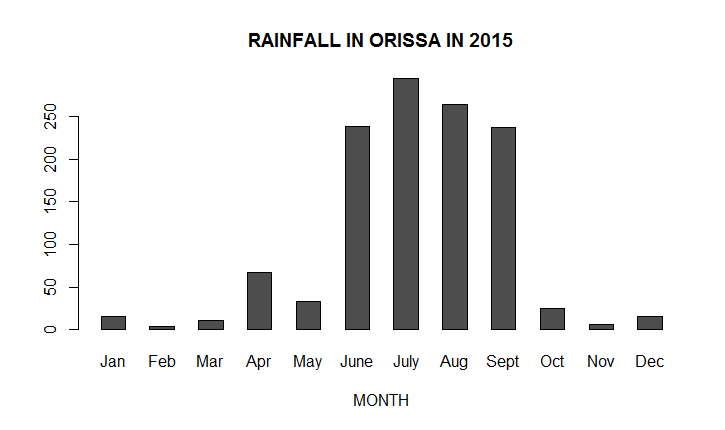


Fig.2

Fig.1: It is the graph representing annual rainfall in Jammu & Kashmir season-wise. X-axis represents year and Y-axis represents annual rainfall. The 4 different color shown in the graph represents the season.

Fig.2: It is the graph representing annual rainfall in Orissa month-wise. X-axis represents months and Y-axis represents annual rainfall.

## 2) Box Plot

The basic syntax to create a boxplot in R is:

ggplot(dataset10,aes(x=SubDiv,y=Annual))+geom\_boxplot()+scale\_y\_continuous(limits=c(0,4000),breaks=seq(0,4000,200))

Following is the description of the parameters used:

* dataset10 🡪 Dataset used for plotting the boxplot
* aes(x,y) 🡪 The x-axis and y-axis of the boxplot
* scale\_y\_continuous() 🡪 Sets the limits of the y-axis
* breaks() 🡪 Sets the scale of they-axis

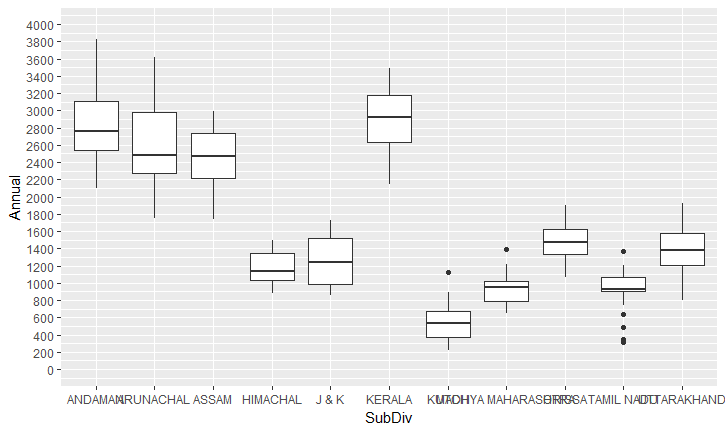
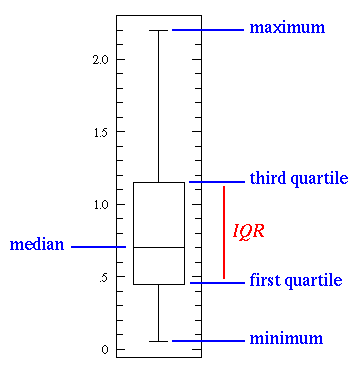


Fig. 3

Fig.3 Shows the Boxplot of the Annual Rainfall OF various SubDivisions in India in a year

Fig.3: The box plot is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum. In the simplest box plot the central rectangle spans the first quartile to the third quartile (the *interquartile range* or *IQR*). A segment inside the rectangle shows the median and "whiskers" above and below the box show the locations of the minimum and maximum.



This simplest possible box plot displays the full range of variation (from min to max), the likely range of variation (the *IQR*), and a typical value (the median). Not uncommonly real datasets will display

## 3) Histogram

A histogram is a plot that lets you discover, and show, the underlying frequency distribution of a set of continuous data. The shape of a histogram is its most obvious and informative characteristic: It helps in data mining to easily identify where a relatively large amount of data is situated. Generally, the X-axis represents the continuous variable which has been split into intervals while the values on the y-axis show how frequently these values on the x-axis occur in the data.

The basic syntax for creating a histogram using R using ggplot2 package is:

ggplot(dataset,aes(x=v)) +geom\_histogram(binwidth=a,col,fill) + labs(x,y) + facet\_wrap(column,scale=”free\_y”)

Following is the description of the parameters used:

• dataset is the dataset you want to plot your histogram about

• v=Column on the X-axis

• binwidth represents width of the bars while col is used to set color of the bars and fill represents outline color.

• xlab and ylab is used to give description of x-axis and y-axis

• scale=”free\_y” just shows the y-axis on each of the histogram plotted.

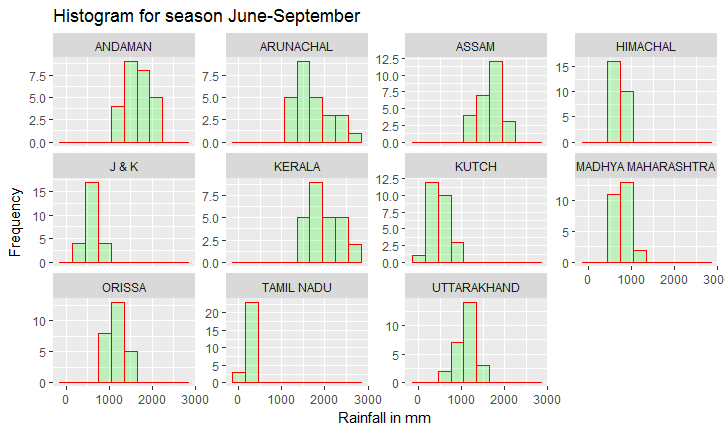
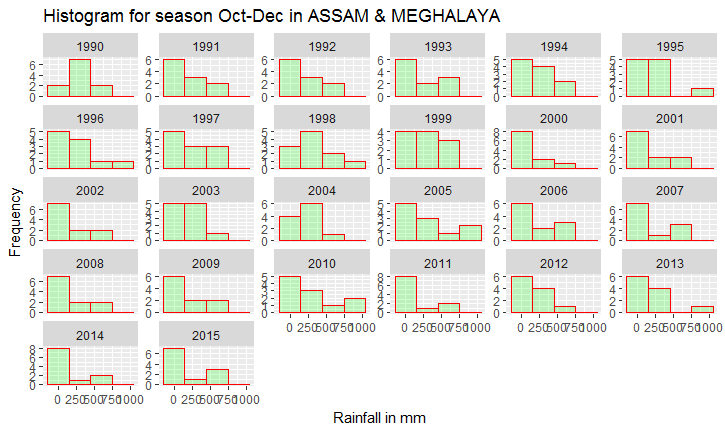


Fig. 4

Fig.4: This histogram represents rainfall in June-September in all the states of a particular year. The X-axis represents the continuous variable rainfall in mm and X-axis denotes its frequency.

Fig. 5

One can visually conclude that Tamil Nadu had the lowest rainfall while Arunachal and Kerala witnessed maximum rainfall.

Fig.5: This histogram represents Annual Rainfall in the season Oct-Dec in the state Assam & Meghalaya in the year 1990-2015.X-axis is the annual rainfall while Y-axis denotes its frequency. For eg. -In 2013, the state witnessed rainfall of 2500-5000mm for approx. 4 times.

One can easily identify that in 2010, the state received maximum rainfall of 7000mm almost 2 times.

## 4) Line Graph

This graph can also be constructed with the help of ggplot2 using the geom\_smooth() function as shown. Line chart (in the below mentioned graph) basically gives us the summary of all the states and their variation in annual rainfall over the entire range of years 1990-2015.One can understand this plotting as-

1) First, points are plotted against each state’s annual rainfall in the particular year.

2) These points of a particular state help determine the approximate curve to be formed thus giving us variation of annual rainfall over the range of years.

One can also apply this graph in data mining as it shows how a particular variable varies with respect to some quantity.

The basic syntax to create a line chart in R using ggplot2 is:

ggplot(dataset,aes(x,y,color)+geom\_smooth(se=FALSE)

Following is the description of the parameters used:

• dataset is the dataset you want to plot your line graph on

• x,y represent X-axis and Y-axis variables respectively(column names) while color represents that the line graph be divided into n colors where n stands for no of different variations in the column passed to it.(See Fig.6)

• geom\_smooth() function actually plots the curve depending on the graph formed by ggplot().

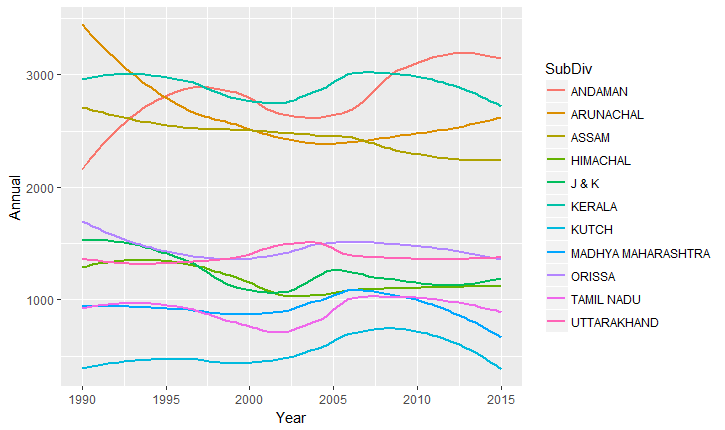


Fig. 6

Fig.6: Here,we use color=”SubDiv” which suggests that 11 different colors be used to represent the 11 states in the SubDiv column.

One can conclude that the annual rainfall of Andaman has been approximately increasing from the year 1990.

## 5) Pie Chart

The basic syntax for creating a pie-chart using the R is:

pie(x, labels, radius, main, col, clockwise)

Following is the description of the parameters used:

• x is a vector containing the numeric values used in the pie chart.

• labels is used to give description to the slices.

• radius indicates the radius of the circle of the pie chart.(value between -1 and +1).

• main indicates the title of the chart.

• col indicates the color palette.

• clockwise is a logical value indicating if the slices are drawn clockwise or anti clockwise.

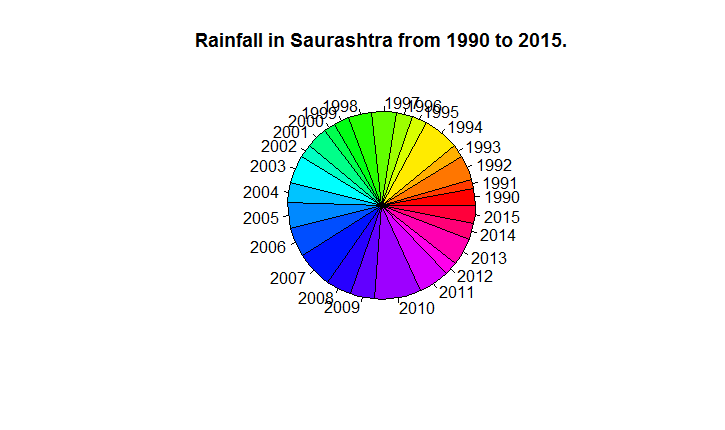


Fig. 9

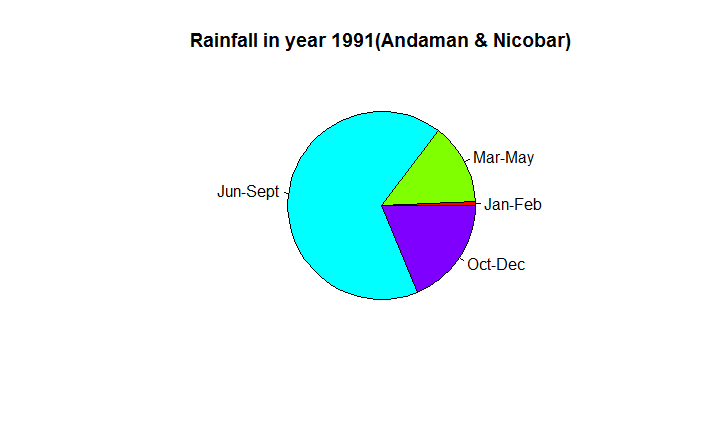


Fig. 10

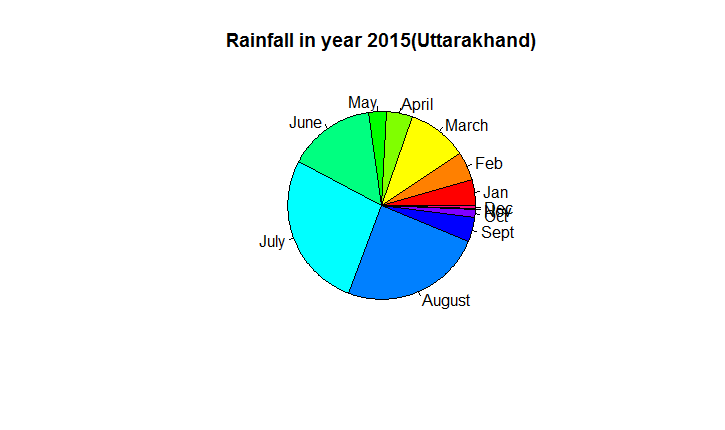


Fig. 11

Fig.9: This Pie Chart represents the annual rainfall in Saurashtra between the year 1990-2015.

Fig.10: This Pie Chart represents the annual rainfall in 1990 for Andaman &Nicobar, during different seasons

Fig.11: This Pie Chart represents the annual rainfall in 2015 for Uttarkhand, during different months.

## 6) Linear Regression

The general mathematical equation for a linear regression is:

y = ax+b

Following is the description of the parameters used:

• y is the response variable.

• x is the predictor variable.

• a and b are constants which are called the coefficients.

The basic syntax for lm() function in linear regression is:

lm(formula,data)

Following is the description of the parameters used:

• formula is a symbol presenting the relation between x and y.

• data is the vector on which the formula will be applied.

The basic syntax for predict() in linear regression is:

predict(object, newdata)

Following is the description of the parameters used:

• object is the formula which is already created using the lm() function.

• newdata is the vector containing the new value for predictor variable.

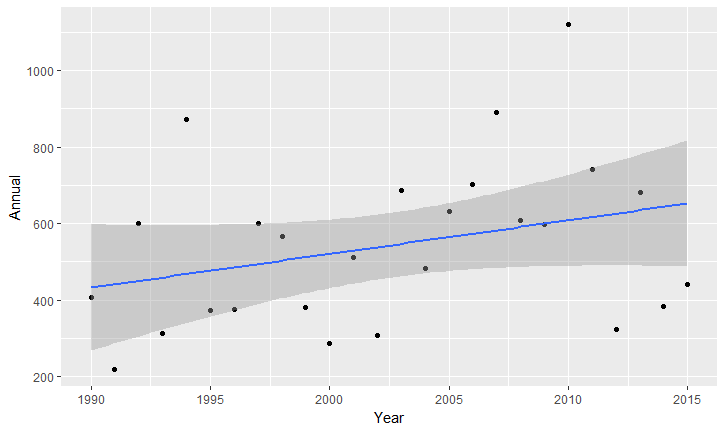


Fig. 12

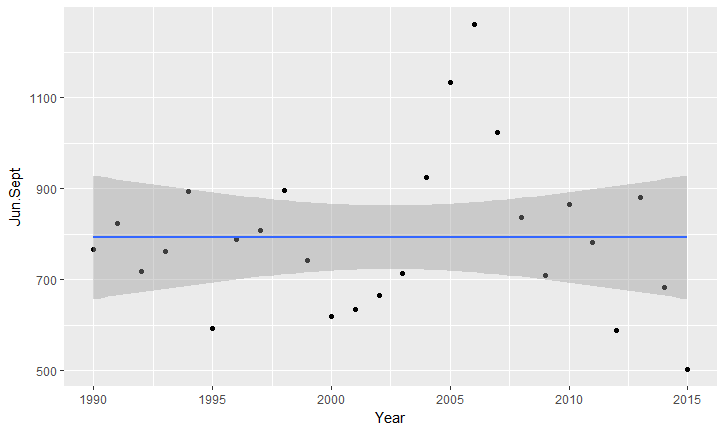


Fig. 13

Fig.12:

Extracted data contains amount of annual rainfall readings along with years for the region of Saurashtra, here Year is plotted along the x axis and Annual rainfall along Y axis. As there are variations seen in amount of rainfall every year, we have plotted the points accordingly, which is fed to compiler(R) which gives a relation and produces a line with positive slope which describes that there is linear increase in annual rainfall every year which can be further used to predict amount of rainfall in forthcoming years. For eg: If we were to predict annual rainfall at year 2020 in Saurashtra region the amount of rainfall is found to be about 696.165 mm

Fig.13:

Extracted data contains amount of annual rainfall readings along with years for the region of Madras during the months of JUNE-SEPTEMBER, here Year is plotted along the x axis and Annual rainfall along Y axis. As there are variations seen in amount of rainfall every year, we have plotted the points accordingly, which is fed to compiler(R) which gives a relation and produces a line with almost zero slope which describes that there is consistency in annual rainfall every year which can be further used to predict amount of rainfall in forthcoming years. For eg: If we were to predict amount of rainfall in year 2020 in Madras region during JUNE-SEPTEMBER the amount of rainfall is found to be about 792.9236 mm