**Using the R for Log File Analysis**

**What is a Log file? :**

Log files are used to track a session by using several parameters like UserId, SessionId, IP address, Path of the session, referral, browser, status code, byte size, and Response time.

**Uses of Analyzing Web Log files :**

Google analytics gives only mass collection of data which do not contain the data of specific user type or the behaviour of a particular user in a particular session is not given by the Google Analytics.  
  
By extracting the log files we can have the details of specific users and so we can generate algorithms to view their behavioural pattern.

**Analyzing the Log files :**

**R code :**

*library(ggplot2)*

#read Log file

*Log <- read.table("D:\\dc++\\workspace 2\\sample.log",header = F,sep = "," , comment.char ='#',stringsAsFactors = F)*

#give Coloumn names

*colnames(Log) <- c("Date","Time","Sample Case","Warning","Error Info")*

#combining date and time

*time <- paste(Log$Date,Log$Time)*

*Log <- cbind(Log,time)*

*rm(time)*

#Change charecter argument to time argument

*Log$time <- as.POSIXlt(Log$time)*

#seperating as days

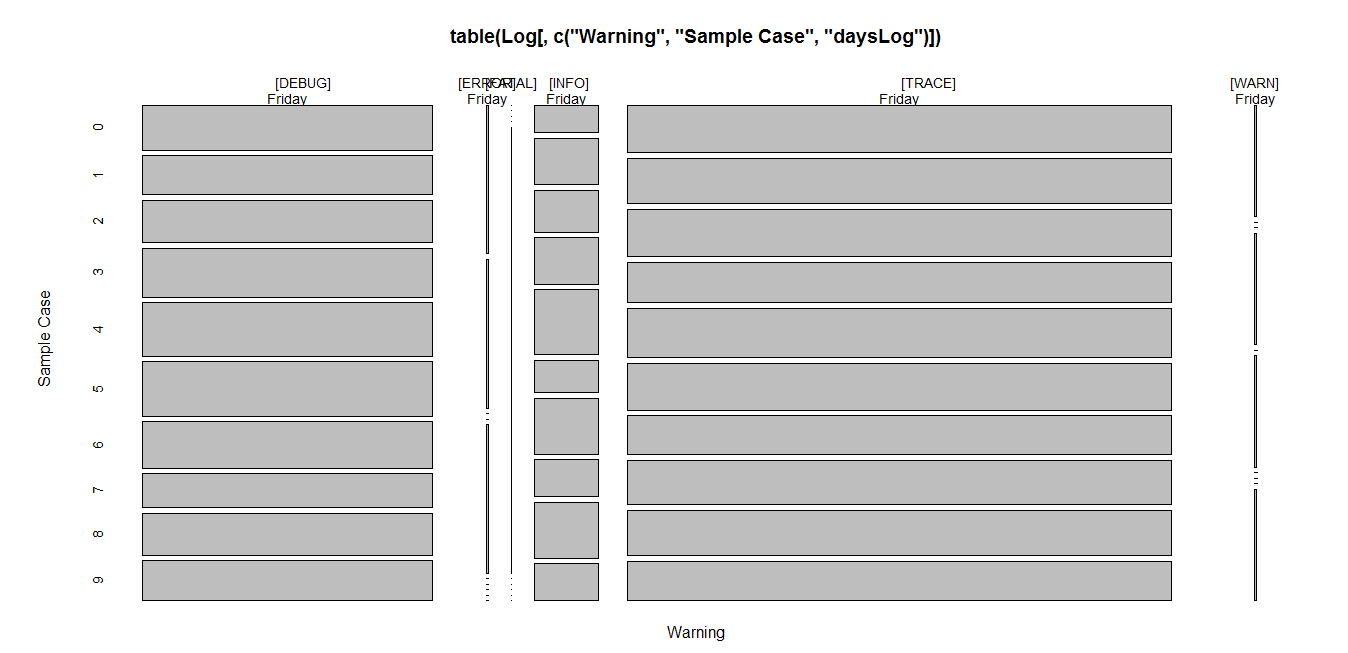
*daysLog <- format(Log$time,"%A")*

*daysLog <- factor(daysLog,levels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"))*

*Log <- cbind(Log,daysLog)*

*Log$`Sample Case` <- substr(Log$`Sample Case`,start = 12,stop = 12)*

*mosaicplot(table(Log[,c("Warning","Sample Case","daysLog")]),cex.axis =0.9)*



*######################################################################################*

*Log2 <- read.table("D:/dc++/workspace 2/access.log",header = F,sep = " ", na.strings = "-",stringsAsFactors = F)*

*Log2$V4 <- substr(Log2$V4,start = 2,stop=21)*

*Log2$V4 <- sub(":"," ",Log2$V4)*

*Log2$V4 <- as.POSIXlt(as.character(Log2$V4))*

*Log2$V4 <- as.POSIXlt(Log2$V4,format =" %d/%b/%Y %H:%M:%S")*

*Newv4 <- strptime(Log2$V4, format = "%Y-%m-%d %H:%M:%S")*

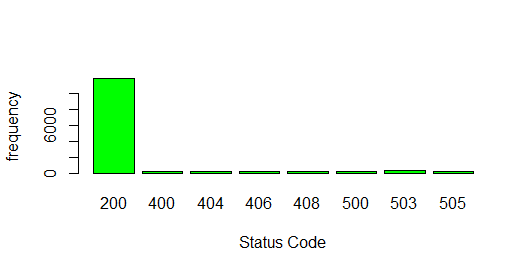
*weekdays <- format(Newv4,"%A")*

*weekdays <- factor(weekdays, levels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"))*

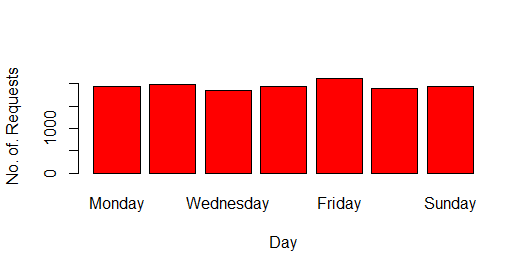
*hours <- format(Newv4 , "%H")*

*colnames(Log2) <- c("IP","UserID", "Session Id" , "Time" ,"Path" , "Status Code" , "Size in bytes" ,"Referral","Browser","Response Time")*

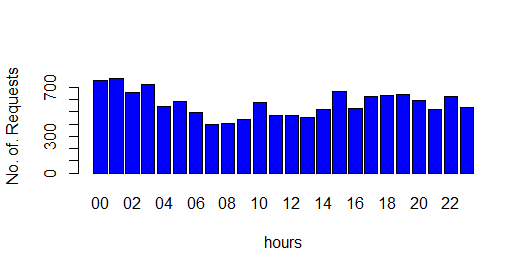
*barplot(table(Log2$`Status Code`),col= "green",xlab = "Status Code",ylab = "frequency")*



*barplot(table(weekdays),col = "red",xlab = "Day",ylab = "No. of. Requests")*



*barplot(table(hours),col = "blue",xlab = "hours",ylab = "No. of. Requests")*



*HACK <- subset(Log2, Path %in% c( "GET /?q=node/add HTTP/1.1", "GET*

*/?q=user/register HTTP/1.1", "GET /?q=node/add HTTP/1.0", "GET*

*/?q=user/register HTTP/1.0" ))*

#to find whether any person trying to hack the website but no one requested to hack as we got HACK as an empty set.