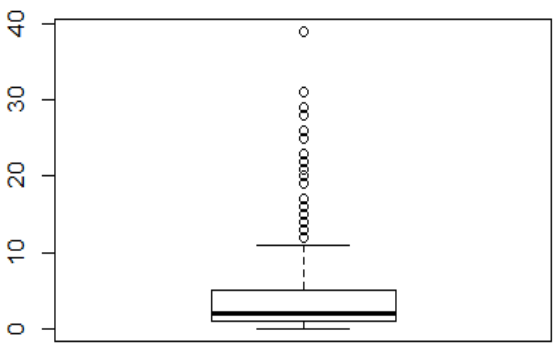
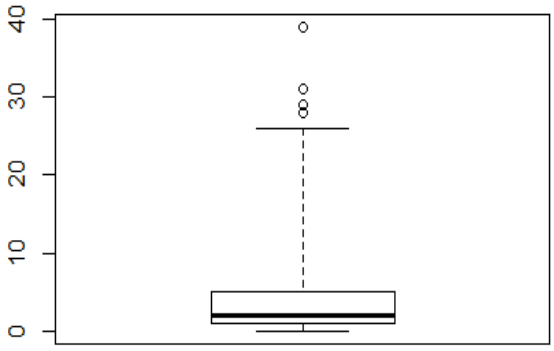
```r
#=============================================================================
#                    Using R: Data Understanding (Shark Attacks)
#=============================================================================

#-----------------------------------------------------------------------------
#(a) The data contains historical information ranging from early ...
#There are some issues that are highly timeliness and also important. Having high
quality data is the key to having good model leads to decision. Here, there are many
problems barricade us of having high quality data. There are missing data, not
complete data, they are not consistent means they do not follow the same pattern. Miss
spelling and abbreviation is also cause some problem. There are outliers that maybe
are by chance or intentional.

#-----------------------------------------------------------------------------
#(b) The data is not "clean", but at least the columns are at least ...
GSAF <- read.csv(file="Shark.csv", head=TRUE, sep=",")
GSAFdata <- GSAF[(GSAF$Year >= 2000),]

#-----------------------------------------------------------------------------
#(c) The Date field is currently stored as character field and ...
NewDate <- as.Date(GSAFdata$Date, "%d-%b-%y")
for (i in 1:length(NewDate)){
  if (!(is.na(NewDate[i])) & ((NewDate[i]) > as.Date("2016-01-01"))){
    NewDate[i] <- as.Date(GSAFdata$Date[i], "%d-%b-%Y")
  }
  if (is.na(NewDate[i])){
    NewDate[i] <- as.Date(GSAFdata$Date[i], "Reported %d-%b-%Y")
  }
}

#-----------------------------------------------------------------------------
#(d) What percentage of the new date field is missing? (Note: should be ...
mean(is.na(NewDate))
## [1] 0.02736466
#Because the Dates are not in a consistent format, some of the data will be missed.
With a short survey
#in the Dates, it can be understood most of the cases that the R could not assign a
date is because of lack
#of complete information.

#-----------------------------------------------------------------------------
#(e) Delete all of the records in GSAFdata that have missing ...
#It will delete all NA dates
GSAFdata$Date <- NewDate
GSAFdata <- GSAFdata[is.na(GSAFdata$Date)==F,]

#(Outliers)It will delete all dates which is greater than both the previous and the
next dates
countor=0
jj=0
end=length(GSAFdata$Date)-2
for(j in 2:end){
  if((GSAFdata$Date[j]>GSAFdata$Date[j-1]) & (GSAFdata$Date[j]>GSAFdata$Date[j+1])){
    jj <- cbind(jj,j)
    countor=countor+1
  }
}
jj <- jj[-1]
GSAFdata <- GSAFdata[-(jj),]
```

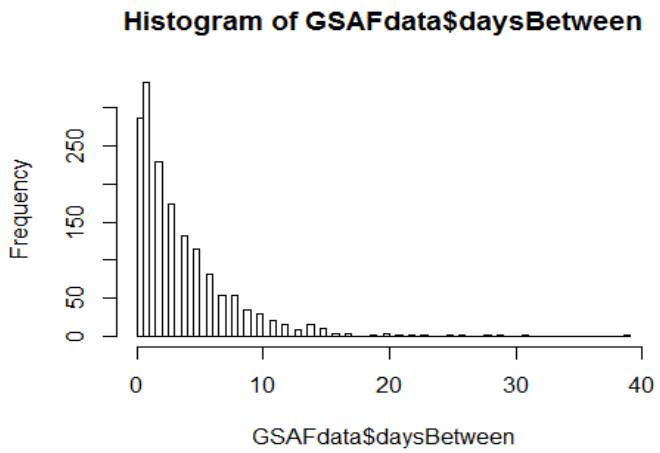
```r
#(Outliers)It will delete all dates which is smaller than both the previous and the
next dates
jj=0
countor=0
end=length(GSAFdata$Date)-2
for(k in 2:end){
  if((GSAFdata$Date[k]<GSAFdata$Date[k-1]) & (GSAFdata$Date[k]<GSAFdata$Date[k+1])){
    jj <- cbind(jj,k)
    countor=countor+1
  }
}
jj <- jj[-1]
GSAFdata <- GSAFdata[-(jj),]

#-------------------------------------------------------------------------------
#(f) It has been said that shark attacks occur as Poisson process, ...

#i. Use the diff command to help you create a vector daysBetween ...
daysBetween <- c(NA, diff(GSAFdata$Date))
GSAFdata <- data.frame(daysBetween,GSAFdata)

#ii. Run and comment on the results from boxplot and adjbox for GSAFdata$daysBetween.
boxplot(GSAFdata$daysBetween)

library(robustbase)
adjbox(GSAFdata$daysBetween)
```



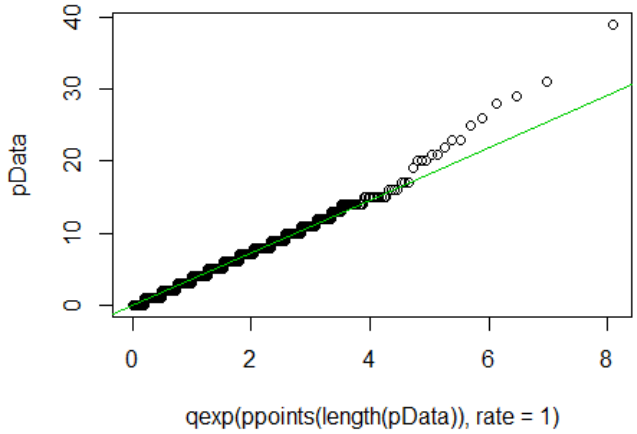



```r
#As can be seen in boxplot we can find many outliers While when we change to adjbox
the number of outliers decrease to 4 dates (because boxplot use normal quantiles). By
a survey on the data it is logical that for about more than a month there would not be
any report of shark attacks. It can be for many reason, for instance, in some month
people do not spend their time near ocean.

#iii. Is the Grubb's test, the Generalized ESD test, both, ...

# helper function for Generalized ESD (adapted from code available tackoverflow.com)
################################################################################
# Compute the critical value for ESD Test
esd.critical <- function(alpha, n, i) {
  p = 1 - alpha/(2*(n-i+1))
  t = qt(p,(n-i-1))
  return(t*(n-i) / sqrt((n-i-1+t**2)*(n-i+1)))}
#main function
removeoutliers = function(y,k=5,alpha=0.05) {
  o <- 0
  if (k<1 || k >= length(y))
    stop ("the number of suspected outliers, k, must be in [1,n-1]")
  ## Define values and vectors.
  y2 = y
```

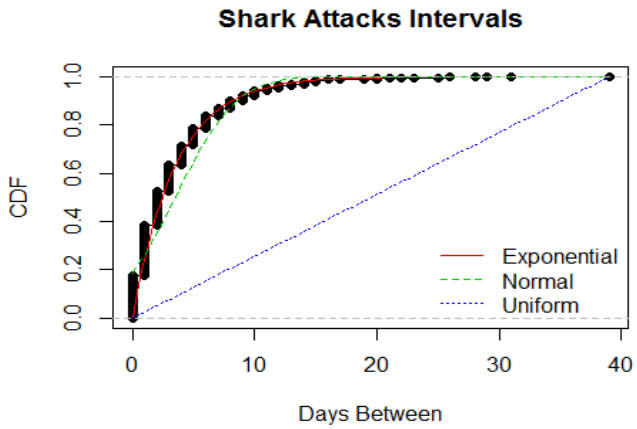
```
  n = length(y)
  toremove = 0
  tval<-NULL
  ris<-NULL
  ## Compute test statistic until r values have been removed from the sample.
  for (i in 1:k){
    if(sd(y2)==0) break
    ares = abs(y2 - mean(y2))/sd(y2)
    Ri = max(ares)
    y2 = y2[ares!=Ri]
    tval<-c(tval,esd.critical(alpha,n,i))
    ris<-c(ris,Ri)
    ## Compute critical value.
    if(Ri>esd.critical(alpha,n,i))
      toremove = i}
  # Values to keep
  if(toremove>0){
    outlierLevel = sort(abs(y-mean(y)),decreasing=TRUE)[toremove]
    o = y[abs(y-mean(y)) >= outlierLevel]
    y = y[abs(y-mean(y)) < outlierLevel]}
  RVAL <- list(numOutliers=toremove,outliers=o,cleandata=y,critical=tval,teststat=ris)
  return (RVAL)}
################################################################################

library(outliers)
grubbs.test(GSAFdata$daysBetween)
removeoutliers(GSAFdata$daysBetween[2:length(GSAFdata$daysBetween)],20,0.05)
hist(GSAFdata$daysBetween, breaks=100)
```

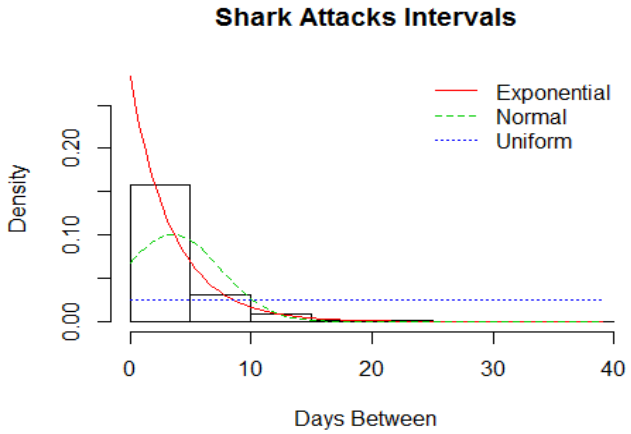


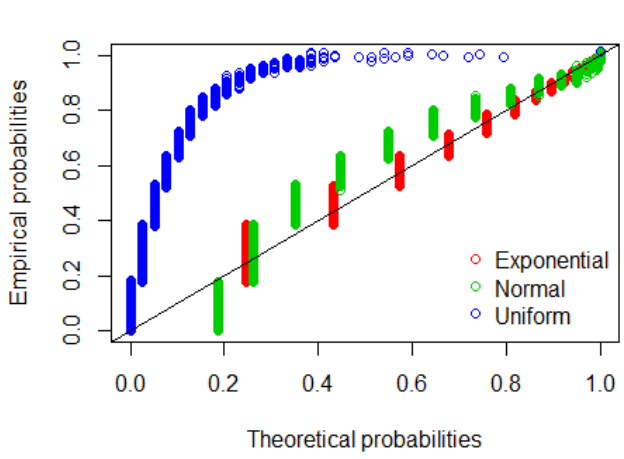

```
#These 2 tests are for normal distribution and as can be seen they are not. Also,
#Using the Grubbs test which is for only one variable it could not be resulted that
there is an outlier because the p-value is so small(comparing to alpha)
#Using the Generalized ESD test it can be resulted that there is 12 outliers
#However, as I described before having this numbers in our data set is logical and
does not need to remove any data as outlier
#It should be noted that some wrong data is already deleted from the data in section
(e) and the gap could have been the result of deleted records too.

#-------------------------------------------------------------------------------
#(g) Use the qqplot and qqline commands to help you visually ...
pData <- GSAFdata$daysBetween[is.na(GSAFdata$daysBetween)==F]
qqplot(qexp(ppoints(length(pData)), rate=1), pData); qqline(pData, distribution=qexp,
col=3)
```

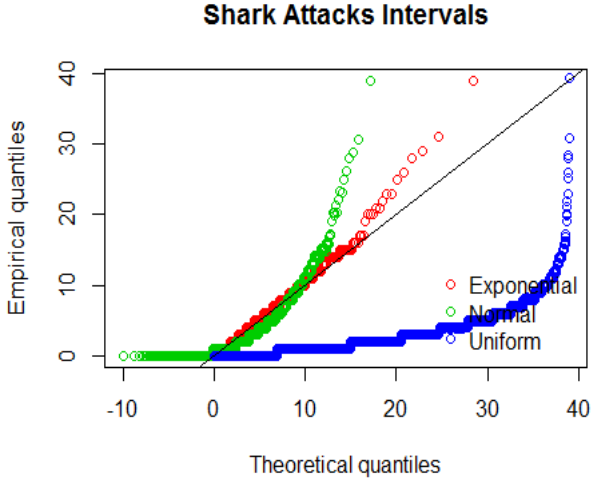



```
#As can be seen it can be concluded that the data is exponentially distributed
#for the first part of the graph which most of the data are there is ok but with
#some parts which contains very small number of data it is not matched.
#Overall the exponential distribution can be observed.

#-------------------------------------------------------------------------------
#(h) Use the package fitdistrplus and the commands such as cdfcomp, ...
library(fitdistrplus)
fitexp <- fitdist(pData, "exp", method="mme")
fitn <- fitdist(pData, "norm")
fituni <- fitdist(pData, "unif")
cdfcomp(list(fitexp, fitn, fituni), legendtext=c("Exponential", "Normal", "Uniform"),
        main="Shark Attacks Intervals", xlab="Days Between", lines01=T)
denscomp(list(fitexp, fitn, fituni), legendtext=c("Exponential", "Normal", "Uniform"),
         main="Shark Attacks Intervals", xlab="Days Between")
ppcomp(list(fitexp, fitn, fituni), legendtext=c("Exponential", "Normal", "Uniform"),
       main="Shark Attacks Intervals")
qqcomp(list(fitexp, fitn, fituni), legendtext=c("Exponential", "Normal", "Uniform"),
       main="Shark Attacks Intervals")
```

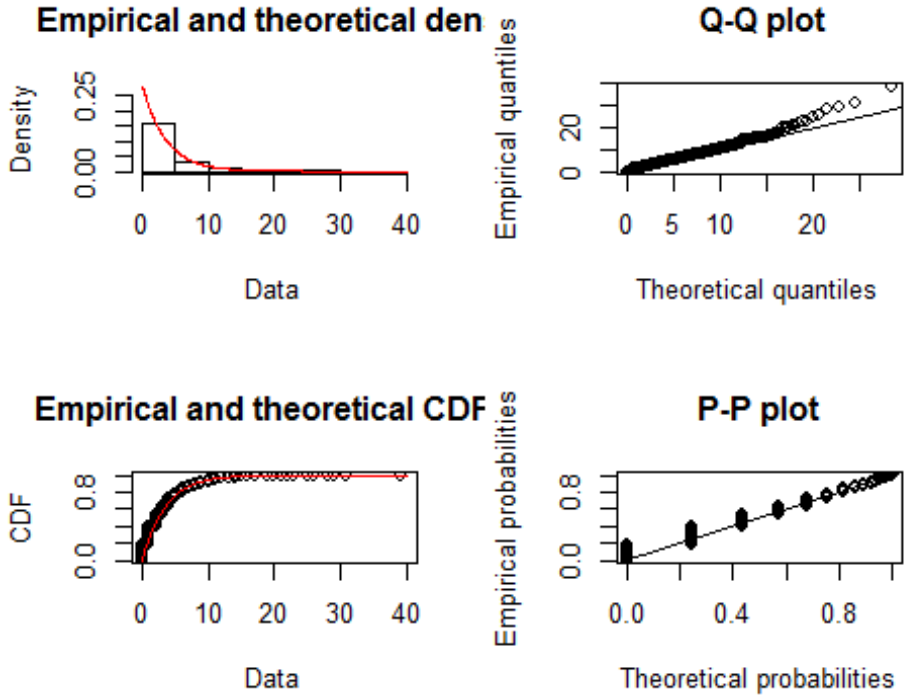








```
fitresult <- gofstat(list(fitexp, fitn, fituni),fitnames=c("Exponential", "Normal",
"Uniform"))
plot(fitexp)
```




```
#Considering all the distribution tested above, exponential is the best match for
daysBetween.

#-------------------------------------------------------------------------------
#(i) How do you respond to the claim that shark attacks ...

#Considering all, there is no obvious and conclusive answer and one can judge by all
the evidences and experience. In my opinion exponential is the best fit among all the
assessed distribution for time between attacks so it is reasonable that one conclude
the shark attacks occurs as Poisson process.
```