Old Faithful

```
## Set this up for your own directory
dataDirectory <- "C:/Users/Jasdeep/Desktop/SEM 2 Waterloo/STAT
847/Assignments/dataDirectory"
path_concat <- function(path1, path2, sep="/") paste(path1, path2, sep = sep)
```

Karan Vijay Singh 20745105 STAT 847

In the Yellowstone National Park, Wyoming, USA there is a famous geyser called “Old Faithful” which erupts with some regularity. The physical model is thought to be something like the following illustration (from Rinehart (1969), p. 572, via):

describe what’s happening in the stages as follows:

``We do not discuss geological reasons for the fact that sometimes the cascading effect works down to the bottom of the tube while at other times it stops earlier. We simply note the phenomenon and discuss its consequences. Stages 3a and 3b are associated with short and long waiting times for the next eruption. In stage 3a, the system starts a new cycle partially filled with hot water so that the following heating time is shorter; at the new eruption the entire tube will be emptied, since part of the water had already been heated in the previous cycle."

For each eruption, the waiting time w between its beginning and the beginning of the previous eruption is recorded to the nearest minute and the duration d of the eruption is recorded to fractions of a minute.

Collected from August 1st until August 15th, 1985 the data record the 299 successive eruptions which occurred during this time. Though R. A. Hutchinson, the park geologist, collected similar data sets, it is not clear from the source whether or not this data set is one of them. Measurements had to be taken through the night and duration times for these eruptions were recorded only as being one of short, medium, or long (encoded here as 2, 3, or 4 minutes, respectively).

a. **(3 marks) Describe the target population/process $\pop{P}_{Target}$ you think scientific investigators have in mind for the above problem. Carefully define both what constitutes an individual unit of $\pop{P}_{Target}$ and how the collective of units is defined.**

Target Population = the eruptions which will occur in the future to which the researchers wish to find a pattern and draw a conclusion.

An Individual Unit = includes an eruption i.e. duration and waiting time of a single eruption.

Collective of units is defined as a set of eruptions (i.e. each eruption being a unit consisting of duration and waiting time) that will happen in the future.

b. **(4 marks) Describe a study population/process $\pop{P}_{Study}$ as it might have been available for the scientific investigators. Again, carefully define both what constitutes an individual unit of $\pop{P}_{Study}$ and how the collective of units is defined. Why might there be study error?**

Study population = the data sets collected by R. A. Hutchinson, the park geologist for the eruptions.

An Individual Unit = an eruption i.e. duration and waiting time of a single eruption from the collected dataset.

Collective of units is defined as a set of eruptions (each eruption being a unit consisting of duration and waiting time) that are in the study population.

There might be study error but the size of the error may not be knowable as the target population contains eruptions(unit) form the future.

c. **(4 marks) Describe the $\samp{S}$. Again, carefully define both what constitutes an individual unit of $\samp{S}$ and how the collective of units is defined. Why might there be sample error?**

Sample = the 299 succesive eruptions from August 1st until August 15th, 1985 on which the study is being performed.

An Individual Unit = an eruption i.e. duration and waiting time of a single eruption from 299 observations.

Collective of units is defined as a set of eruptions (each eruption being a unit consisting of duration and waiting time) that are in the sample.

There might be sample error because the sample I might choose, may contain eruptions recorded of only day time.Hence resulting in sample error.

d. **(2 marks) Imagine the process for selecting a sample. How might this process produce sampling bias?**

The different data sets collected by R. A. Hutchinson, the park geologist for the eruptions are the different samples available to us.There might be some datasets that contain only day time eruptions data or night time eruptions data which doesnot represent the population of interest fully. Hence it will result in producing sampling bias.

e. **(4 marks) For eruption i, d_i denotes its duration and w_i the time between its beginning and the beginning of the previous eruption. Given the above description of a physical model for how the geyser might work, explain why the independence of the variates in each of the following pairs might be of interest:**

i. w_i **and** d_i

If w_i and d_i are independent i.e. the ith wait time is independent of ith duration i.e it won't matter for how long the current wait time is, the duration of eruption will remain same.

ii. d_i **and** w_{i+1}

It means that the next wait time is independent of the current duration i.e. whether the current duration happens for long time or short time or in the other sense the geyser tube gets completely or partly empty, the next waiting time won't change.

iii. d_{i-1} **and** d_i

If d_{i-1} and d_i are independent i.e. the previous duration time is independent of current duration time, that means duration of the current eruption won't change whether the duration time of last eruption is short or long.

iv. w_{i-1} **and** w_i

If w_{i-1} and w_i are independent i.e. the previous wait time is independent of current wait time that means whether for previous eruption, the waiting time was long or short, the current wait time won't change depending uopn the previous wait time.

f. **(2 marks) Describe one other variate of potential interest which is implicitly defined in this data set? How would you determine its value?**

The time of the day i.e. whether day or night during which the eruptions were recorded. Its value can be determined from the duration values of eruptions as in the night time the duration vaues are rounded to 2,3,4 minutes.

g. **(3 marks) Imagine the measuring process. What problem(s) do you think might be associated with the measuring process? How might it manifest itself in terms of measuring bias and/or variabiliry?**

One main problem in measuring is the night time measurements which have been recorded as short, medium or long (2,3,4 minutes) and not accurate

Also the instrument or stop watch used to measure the durations may not be accurate.

Also the person taking the measurements may stop the stop watch early while the eruption has not completely ended or delay it by few seconds after the eruption has ended.

h. **(10 marks) To assess the measuring systems, we might consider looking at the least significant parts of each measurement. For this the modulus arithmetic binary operator `%%` in `R` can be handy to find the least significant part of a measurement. For example `x %% 10` will return the rightmost digits in a non-negative integer `x` and `x %% 1` will return the fractional part of a non-negative real number `x`.**

Using the `%%` modulus operator to construct the appropriate datali, perform a Pearson chi-square goodness of fit (in each case use 10 non-overlapping equal size bins) to test each of the following hypotheses:

i. H_d **the fractional part of the duration follows a** $U[0,1]$ **distribution, ii.** H_w **the rightmost digit of the waiting time equiprobably any one of the digits** $0,1,2,\ldots,9$**.**

Summarize your findings (including showing your code). What do you conclude about the two measuring systems?

```
library(MASS)
duration <- geyser$duration



#no of bins
B <- 10
#probabilty of falling in each bin
probabilty <- rep(1/B,B)

#getting the fractional part of duration using modulus
fract_duration <- duration%%1

values <- hist(fract_duration, breaks = B,plot = FALSE)

observedFraction <- values$counts
#performing goodness of fit test for fractional part of duration
chisq.test(observedFraction, p = probabilty)

##
##  Chi-squared test for given probabilities
##
## data:  observedFraction
## X-squared = 153.61, df = 9, p-value < 2.2e-16

#very small value, hence evidence against the null hypothesis
```

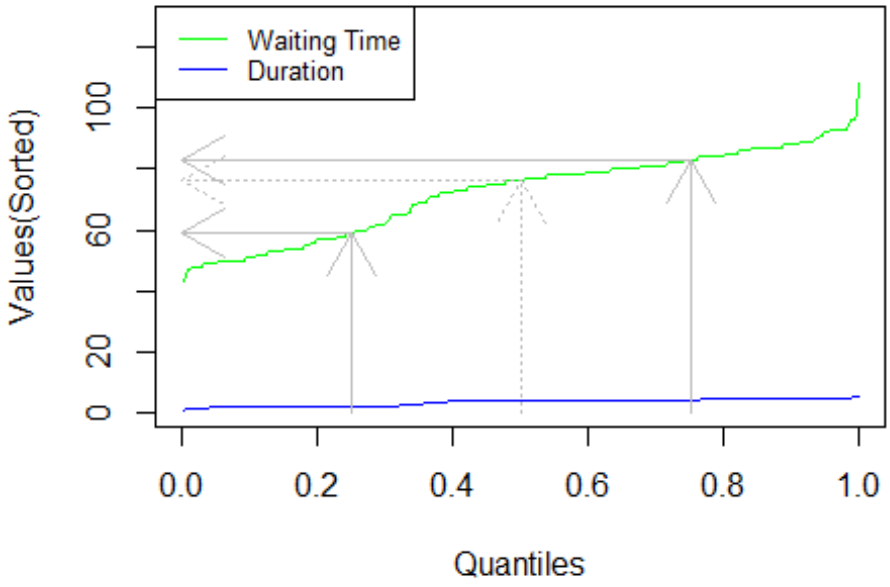
The p-value is very small. Hence, we have an evidence against the null hypothesis.Hence , we can say that the measuring system used to measure the waiting time is not accurate.

```
#part b
waitingTime <- geyser$waiting

#getting rightmost part of waiting time
rightmost_waitingTime <- waitingTime%%10

numCount <- table(rightmost_waitingTime)

#performing goodness of fit test for fractional part of duration
chisq.test(numCount, p = probabilty)

##
##  Chi-squared test for given probabilities
##
## data:  numCount
## X-squared = 9.194, df = 9, p-value = 0.4196
```

The p value is 0.4196 which is not small. Hence we cannot say that we have an evidence against the null hypothesis. Hence, we can say the measuring system to measure the duration is quite accurate.

i. **(12 marks) Plot the sample quantiles of both the duration and the waiting times on the same plot (use a different colour for each variate). Show your plot and the code used to generate it. By referring to the relevant features of the sample quantiles, separately describe the distribution of each variate and compare the two distributions to one another. Now compare the two distributions by constructing an appropriate quantile-quantile plot and referring to its relevant features. Again show the plot and the code.**

```
library(MASS)

duration <- geyser$duration
waitingTime <- geyser$waiting

xValues <- ppoints(n = length(waitingTime))

waitingTimeRange <- range(waitingTime)
durationRange <- range(duration)
min_y <- min(durationRange[1],waitingTimeRange[1])
max_y <- max(durationRange[2],waitingTimeRange[2])+20

y_range <- c(min_y,max_y)
plot(x=xValues, y=sort(duration), col = "blue" , type ="l", ylim =y_range,
xlab = "Quantiles", ylab = "Values(Sorted)" )
lines(x = xValues , y = sort(waitingTime), col ="green")
legend("topleft", legend=c("Waiting Time", "Duration"),
       col=c("green", "blue"), lty=1:1, cex=0.8)
#1st Quartile
arrows(0.25,0,0.25,59, lty = 1,col ="grey")
arrows(0.25,59,0,59, lty=1,col ="grey")
#Median
arrows(0.5,0,0.5,76, lty=3,col ="grey")
arrows(0.5,76,0,76, lty=3,col ="grey")
#3rd Quartile
arrows(0.75,0,0.75,83, lty=1,col ="grey")
arrows(0.75,83,0,83, lty=1,col ="grey")
```



Distribution of Duration: Min. : 0.833

1st Qu.: 2.000

Median : 4.000

Mean : 3.460

3rd Qu.: 4.383

Max. :5.450

Mid Quartile : 3.19

As can be seen from the data i.e median and mid quartile values are not same.Therefore we can say that distribution of duration is skewed.

The grey lines above in the graph indicate the 1st Quartile, median and 3rd quartile for waiting time.

Distribution of Waiting Time: Min. : 43

1st Qu.: 59

Median : 76

Mean : 72.314

3rd Qu.: 83

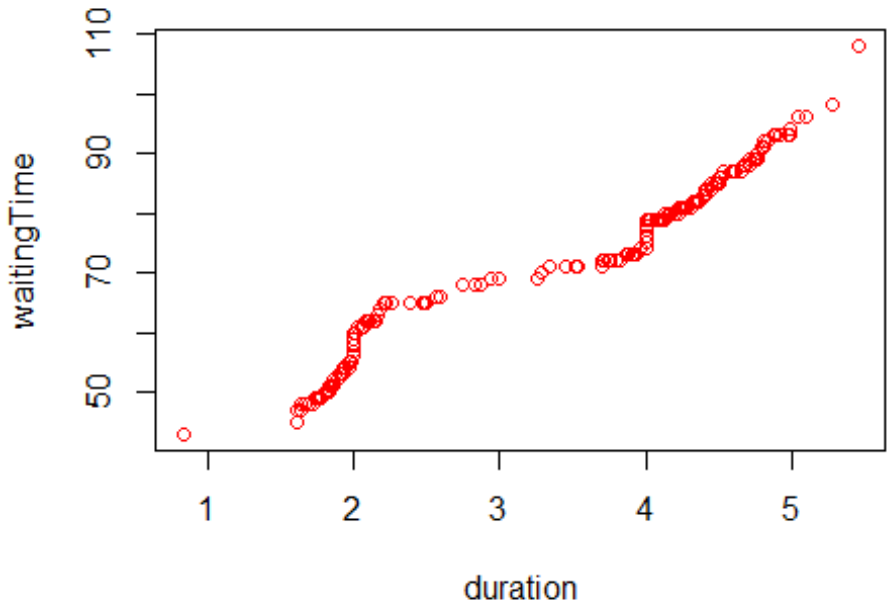
Max. :108

Mid Quartile :71

As can be seen from the data i.e median and mid quartile values are not same.Therefore we can say that distribution of waiting time is skewed.

The duration and waiting time do not come from the same distribution.

```
#plotting the quantile quantile plot for comparing the distributions
qqplot(duration,waitingTime, col="red")
```



As we can see from the plot that both waiting time and duration do not follow the same distribution as the graph is not a straight line.

j. **(10 marks) Consider whether the waiting times w_i. We might ask whether waiting times are independently distributed. One way to test this is to compare each waiting time w_i with that one that occurred exactly k eruptions previously, namely w_{i-k}, the so called ``lagged k" value. For $k \geq 1$, there will be $n - k$ pairs (w_{i-k}, w_i) which could be assessed for independence.**

Rather than consider the original waiting times, use the function `transform2uniform()` on the waiting time to give values $u_i = \hat{Q}_W(w_i)$. So we now consider the independence of u_i and its lag k value u_{i-k}.

Conduct a line up test for independence of u_{i-k} and u_i for each of

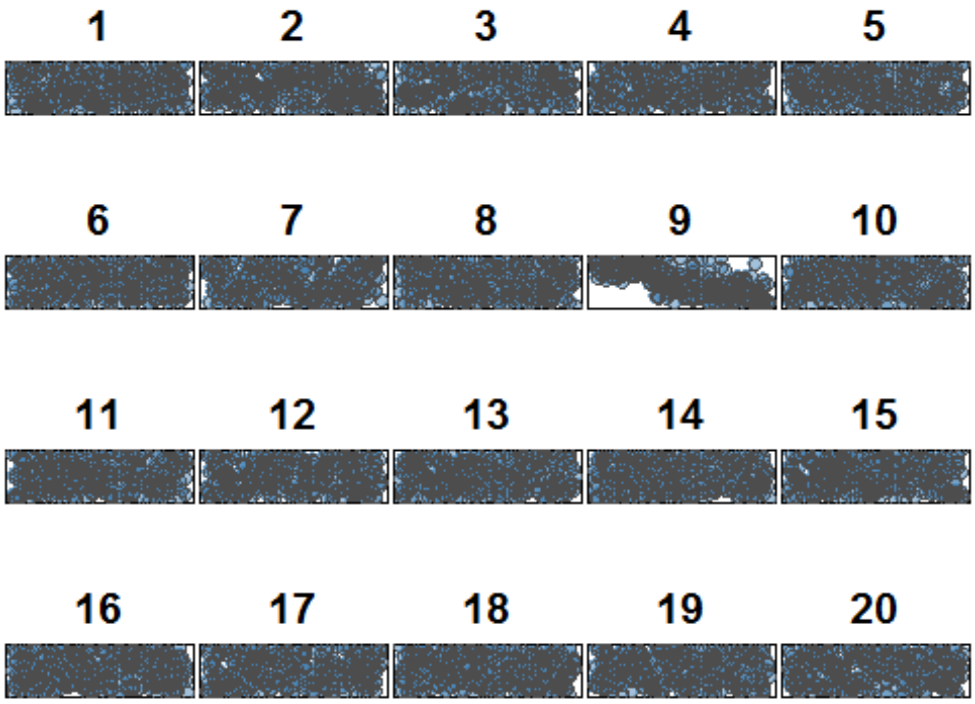
i. $k = 1$, the immediately preceding eruption, and ii. $k = 22$, the eruption occurring roughly the day before.

Show your code for constructing the necessary data and the lineup plots. What do you conclude about the dependence between waiting times?

```
library(MASS)
duration <- geyser$duration
waitingTime <- geyser$waiting
n <- length(waitingTime)

#transform2uniform function
transform2uniform <- function(x, a = if(length(x) <= 10) 3/8 else 1/2, ...)
{(rank(x, ...) - a) / length(x)}


#function to get mix waiting times to check for their independence
mixCoordinates <- function(oldFaith){
  n <- length(oldFaith$x)
  stopifnot(n == length(oldFaith$y))
  oldFaith_samples <- sample(oldFaith$x, n , replace = FALSE)
  data.frame(x = oldFaith_samples, y= oldFaith$y)
}

#function to get n-k pairs of waiting times
getWaitingTimeComb <- function(oldFaith, k){
  waitIK <- c()
  waitI <- c()
  wTime <- oldFaith$waiting
  n <- length(wTime)
  for(i in k:n){
    if(i > k){
      waitIK <- c(waitIK, wTime[i-k])
      waitI <- c(waitI, wTime[i])
    }
  }
  return(data.frame(x = waitIK, y = waitI))
}

uniform_wTime <- transform2uniform(waitingTime)
uniform_wDur <- transform2uniform(duration)
uniform_oldFaith <- data.frame(waiting = uniform_wTime, duration =
uniform_wDur)

#my actual uniform data
pairs_22 <- getWaitingTimeComb(uniform_oldFaith,22)
pairs_1 <- getWaitingTimeComb(uniform_oldFaith,1)

#scatter plot
```

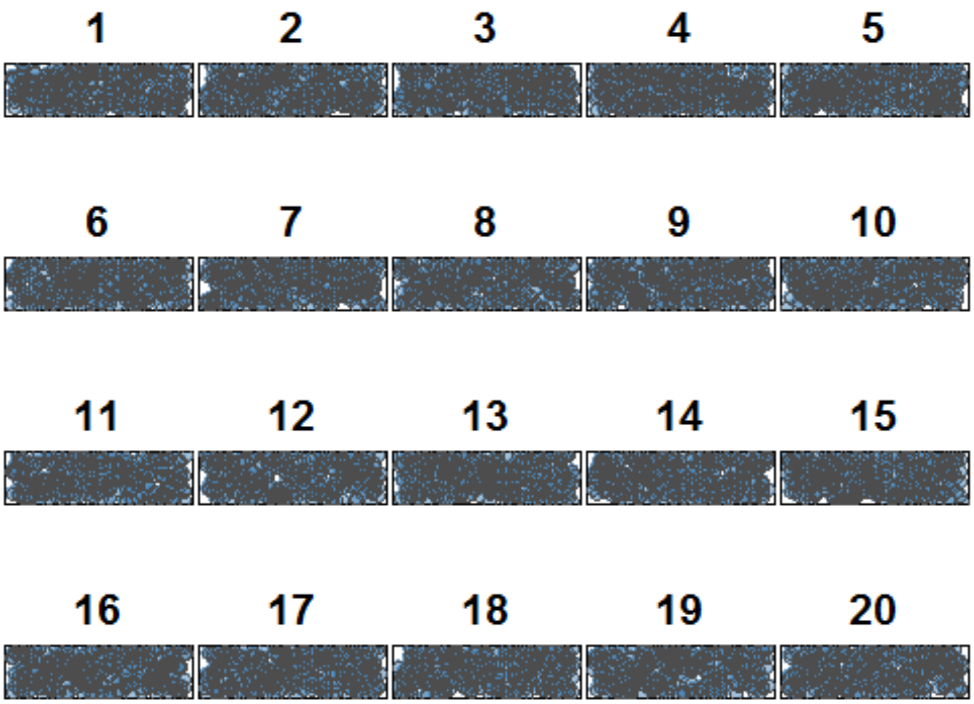
```r
scatterPlot <- function(data, subjectNo) {
  plot(data$x, data$y, # Assume data has these named components
       main=paste(subjectNo), cex.main = 2,# display subject number
       ylab="", xlab="", xaxt="n", yaxt="n", # remove axes
       pch=19, col=adjustcolor("steelblue", 0.5), cex=1.75 # the points
  )
  points(data$x, data$y, pch=1, col="grey30", cex=1.5) # adds point outlines
}

#hide location function
hideLocation <- function(trueLoc, nSubjects){
  possibleBaseVals <- 3:min(2*nSubjects, 50)
  # remove easy base values
  possibleBaseVals <- possibleBaseVals[possibleBaseVals != 10 &
possibleBaseVals != 5]
  base <- sample(possibleBaseVals, 1)
  offset <- sample(5:min(5*nSubjects, 125),1)
  # return location information (trueLoc hidden)
  list(trueLoc = paste0("log(",base^(trueLoc + offset), ", base=",base,") -
", offset))
}

#reveal Location
revealLocation <- function(hiddenLocation){
  eval(parse(text=hiddenLocation$trueLoc))}

#Lineup Function to generate
lineup <- function(data, showSubject=NULL, generateSubject=NULL,
                   trueLoc=NULL, layout =c(5,4)
) {
  # Get the total number of subjects
  nSubjects <- layout[1] * layout[2]
  # Get the location to be used for the real data
  if (is.null(trueLoc)) {trueLoc <- sample(1:nSubjects, 1)}
  # Error checking
  if (is.null(showSubject)) {stop("need a plot function for the subject")}
  if (is.null(generateSubject)) {stop("need a function to generate subject")}
  # Local function to decide which subject to present
  presentSubject <- function(subjectNo) {
    if(subjectNo != trueLoc) {data <- generateSubject(data)}
    showSubject(data, subjectNo) }
  # This does the plotting
  savePar <- par(mfrow=layout, mar=c(2.5, 0.1, 3, 0.1), oma=rep(0,4))
  sapply(1:nSubjects, FUN = presentSubject)
  par(savePar)
  #print(trueLoc)
  # hide the true location but return information to reconstruct it.
  hideLocation(trueLoc, nSubjects)
}
```

```
#set.seed(35521)
answer <- lineup(pairs_1,
                 generateSubject = mixCoordinates,
                 showSubject = scatterPlot,
                 layout=c(4,5))
```



As it can be clearly seen from the lineup plots that one of them is visually different from all others(probabilty of observing something atleast as strange as the observed data =1/20) and that corresponds to the data, which clearly indicates strong evidence against the null hypothesis that ith waiting time is independent of i-1 th waiting time.

```
answer <- lineup(pairs_22,
                 generateSubject = mixCoordinates,
                 showSubject = scatterPlot,
                 layout=c(4,5))
```

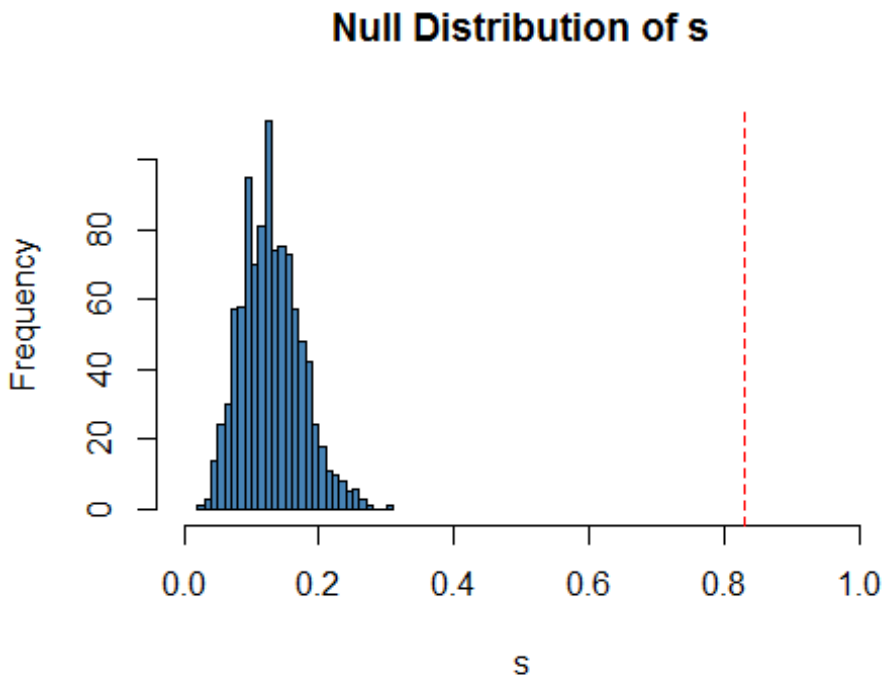


As it can be clearly seen from the lineup plots that it is difficult to visually differentiate or identify the plot that corresponds to the data or that is different from other plots, Hence we do not have evidence against the null hypothesis that the ith waiting time is independent of i-22 th waiting time

k. **(10 marks) Consider now the relationship between w_i and d_i. By fitting and appropriately summarising a `loess()` smooth of d_i on w_i carry out a significance test of the hypothesis that duration is independent of waiting time. Show your code. What do you conclude from this test.**

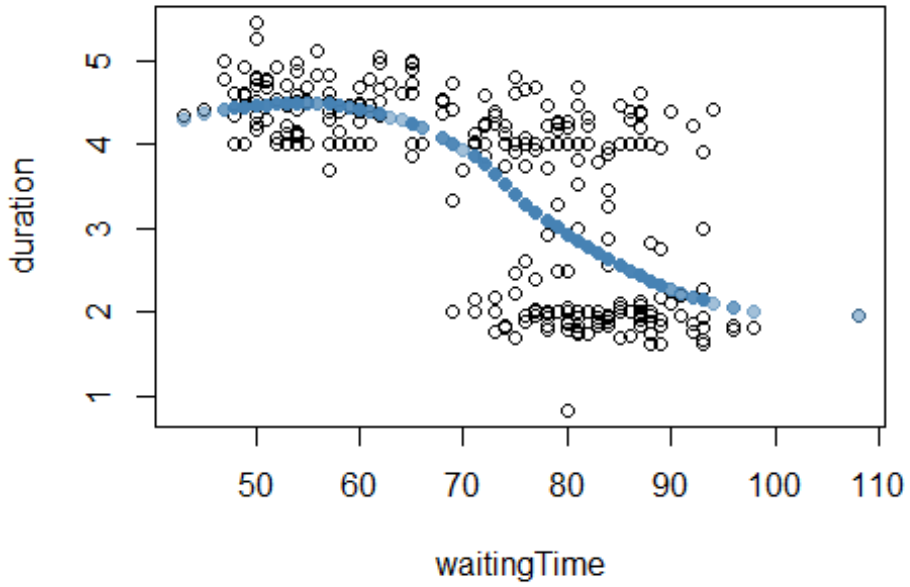
```
library(MASS)

duration <- geyser$duration #dependent
waitingTime <- geyser$waiting #

n <- length(waitingTime)

#function to get mix waiting times to check for their independence
mixCoordinates <- function(oldFaith){
  n <- length(oldFaith$waiting)
  stopifnot(n == length(oldFaith$duration))
  oldFaith_samples <- sample(oldFaith$waiting, n , replace = FALSE)
  data.frame(x = oldFaith_samples, y= oldFaith$duration)
}

#fitting on sample data
stdDev_Samples <-c()
```

```
for (i in 1:1000){
  mixCoord <- mixCoordinates(geyser)
  smoothFit <- loess(mixCoord$y ~ mixCoord$x, family = "symmetric")
  stdDev_Samples <-c(stdDev_Samples, sd(smoothFit$fitted))
}
limit_x <-c(0,1)
hist(stdDev_Samples,breaks = 30,col = "steelblue",xlim = limit_x,main = "Null
Distribution of s", xlab = "s")

#fitting in actual data
smoothFitRobust <- loess(duration ~ waitingTime , data = geyser, family =
"symmetric")
s <- sd(smoothFitRobust$fitted)
abline(v=s, col="red",lty=2)
```



The value of standard deviation for the data is 0.82879.On doing it 1000 times, the observed value of s= 0.82879 looks unusual.The probability of observing something at least as strange as the observed data is 0.This corresponds to an observed significance level of 0 and provides very strong evidence against the hypothesis i.e. that duration is independent of waiting time.

```
plot(x = waitingTime, y= duration)
lines(x = waitingTime, y =smoothFitRobust$fitted,pch=19,
col=adjustcolor("steelblue", 0.5), type = "p")
```



As it can be seen, that the fitted line is not a straight line parallel to the x-axis.Hence, evidence against the null hypothesis i.e. fitted durations are not independent of waiting time.

l. **(12 marks) Consider the possible dependence of the ith duration d_i on that duration, d_{i-k}, lagged k behind. Using a two-dimensional kernel density estimate as a means to display the data (without the data points), conduct a lineup test of independence using joint density contours for each of**

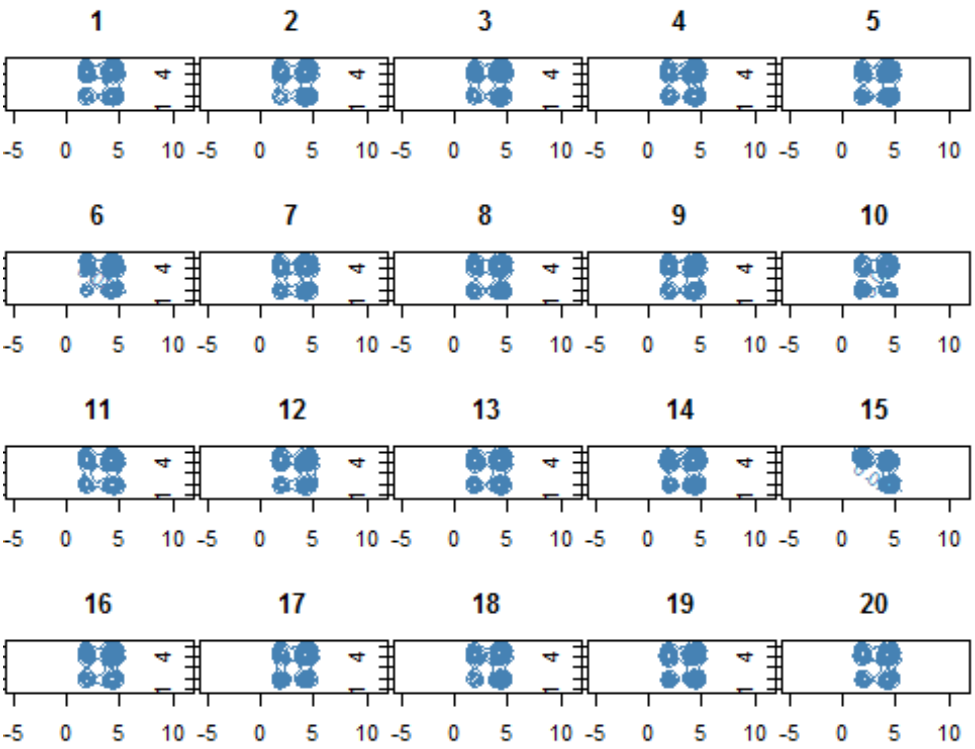
i. $k = 1$, the immediately preceding eruption, and ii. $k = 22$, the eruption occurring roughly the day before.

Show your code for constructing the necessary data and the lineup plots. What do you conclude about the dependence between durations lengths?

```
library(MASS)
duration <- geyser$duration
waitingTime <- geyser$waiting
n <- length(waitingTime)

#function to get n-k pairs of duration times
getDurationComb <- function(oldFaith, k){
  durationIK <- c()
  durationI <- c()
  durationTime <- oldFaith$duration
  n <- length(durationTime)
  for(i in k:n){
```

```
    if(i > k){
      durationIK <- c(durationIK, durationTime[i-k])
      durationI <- c(durationI, durationTime[i])
    }
  }
  return(data.frame(x = durationIK, y = durationI))
}

#contour plot
contourPlot <- function(data, subjectNo) {
  densities <- kde2d(data$x,data$y)
  contour(densities, asp=1,
          main = subjectNo,col="steelblue")

}

#hide location function
hideLocation <- function(trueLoc, nSubjects){
  possibleBaseVals <- 3:min(2*nSubjects, 50)
  # remove easy base values
  possibleBaseVals <- possibleBaseVals[possibleBaseVals != 10 &
possibleBaseVals != 5]
  base <- sample(possibleBaseVals, 1)
  offset <- sample(5:min(5*nSubjects, 125),1)
  # return location information (trueLoc hidden)
  list(trueLoc = paste0("log(",base^(trueLoc + offset), ", base=",base,") -
", offset))
}

#reveal Location
revealLocation <- function(hiddenLocation){
  eval(parse(text=hiddenLocation$trueLoc))}

#Lineup Function to generate
lineup <- function(data, showSubject=NULL, generateSubject=NULL,
                   trueLoc=NULL, layout =c(5,4)
) {
  # Get the total number of subjects
  nSubjects <- layout[1] * layout[2]
  # Get the location to be used for the real data
  if (is.null(trueLoc)) {trueLoc <- sample(1:nSubjects, 1)}
  # Error checking
  if (is.null(showSubject)) {stop("need a plot function for the subject")}
  if (is.null(generateSubject)) {stop("need a function to generate subject")}
  # Local function to decide which subject to present
  presentSubject <- function(subjectNo) {
    if(subjectNo != trueLoc) {data <- generateSubject(data)}
    showSubject(data, subjectNo) }
  # This does the plotting
```

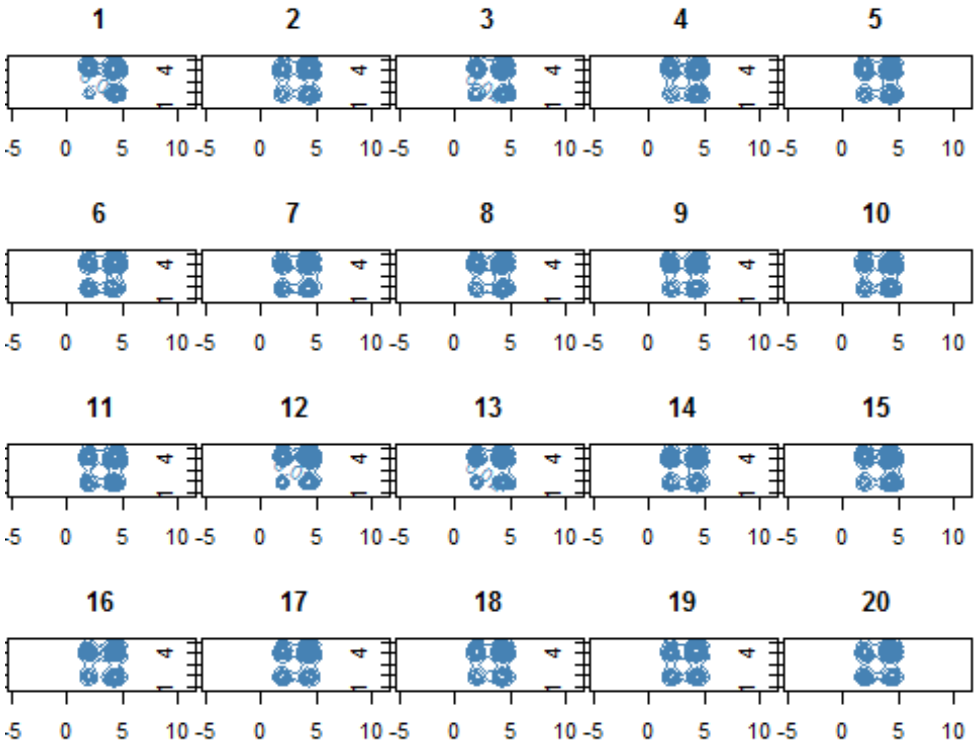
```
  savePar <- par(mfrow=layout, mar=c(2.5, 0.1, 3, 0.1), oma=rep(0,4))
  sapply(1:nSubjects, FUN = presentSubject)
  par(savePar)
  #print(trueLoc)
  # hide the true location but return information to reconstruct it.
  hideLocation(trueLoc, nSubjects)
}

#function to get mix waiting times to check for their independence
mixCoordinates <- function(oldFaith){
  n <- length(oldFaith$x)
  stopifnot(n == length(oldFaith$y))
  oldFaith_samples <- sample(oldFaith$x, n , replace = FALSE)
  data.frame(x = oldFaith_samples, y= oldFaith$y)
}

#my data
pairs_1 <- getDurationComb(geyser,1)
#set.seed(35521)
answer <- lineup(pairs_1,
                 generateSubject = mixCoordinates,
                 showSubject = contourPlot,
                 layout=c(4,5))
```



As it can be clearly seen from the lineup plots that one of them is visually different from all others(probabilty of observing something atleast as strange as the observed data =1/20)

and that corresponds to the data, which clearly indicates strong evidence against the null hypothesis that ith duration is independent of i-1 th duration.

```
pairs_22 <- getDurationComb(geyser,22)
answer <- lineup(pairs_22,
                 generateSubject = mixCoordinates,
                 showSubject = contourPlot,
                 layout=c(4,5))
```



As it can be clearly seen from the lineup plots that it is difficult to visually differentiate or identify the plot that corresponds to the data or that is different from other plots, Hence we do not have evidence against the null hypothesis that the ith duration is independent of i-22 th duration.