EDA A5 Q1

The data on `duration` and `waiting` times are contained in the `geyser` data set found in the `MASS` package. Load this as

```
source(file = "/Users/rudranibhadra/Downloads/graphicalTests.R", echo = FALSE)
source(file = "/Users/rudranibhadra/Downloads/generateData.R", echo = FALSE)
source(file = "/Users/rudranibhadra/Downloads/numericalTests.R", echo = FALSE)
library(MASS)
data(geyser)
```

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

The target population consists of the all the eruptions of the geyser Old Faithful (even in the near future). The units are the individual eruptions for which the waiting time w and duration d are the set of units.

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

The study population consists of eruptions of the geyser Old Faithful which have occurred till date. The units are the individual eruptions recorded for which the waiting time w and duration d are the set of units. Study error can occur since it may not fully represent the target population which consists of all the geyser eruptions of Old Faithful.

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

The sample consists of 299 successive eruptions which occurred during August 1st until August 15th, 1985. The units are the individual eruptions recorded for which the waiting time w and duration d are the set of units. Sample error can occur since it may not fully represent the study population which consists of all the geyser eruptions of Old Faithful which have occurred till date.

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

The sample of eruptions were chosen from a period of only 15 days and the measurements were taken at night. It does not include measurements taken during the day. Sampling bias can occur because of weather, temperature, humidity etc or from the fact that the units with long/short waiting times or with short/medium/long duaration times may be sampled with lower probability and so any of these can be underrepresented in the sample.

e. *(4 marks)*
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
   ii. d_i and w_{i+1}
   iii. d_{i-1} and d_i
   iv. w_{i-1} and w_i

f. It see if the waiting times and the duration of an eruption affect each other. If there is more water in the tube, the waiting time is longer, it means more water to heat so the duration time will be longer.

ii. To see if the duration of an eruption is affected by the waiting time of the next eruption. If the duration is long, more water evaporates which leaves more space in the tube. So more water needs to fill in the tube so the next waiting time will be longer.

iii. To see if the duration of two consecutive durations are related or not since longer duration indicates more water to evaporate which leaves more space in the tube. So more cold water will fill the geyser and the water at the bottom will be under higher pressure. It will evaporate at higher temperatures and so the duration time may be longer in the next eruption.

iv. To see if the waiting time of two consecutive durations are correlated or not since longer the waiting time, longer the duration. So more water needs to be refilled in the geyser and the waiting time for the next eruption might be longer.

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 actual heating (waiting) time of certain eruption can be an implicit variate which can be defined as $w_i - d_{i-1}$.

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 variability?

There might be errors in calculating the duration times of the eruptions since the measurements were taken only at night as the duration times were only recorded as 2,3 or 4 mins. The fractional part isnt calculated. There also might be errors in recording the values via different geologists and different instruments. This might give the measuring bias as the average of all possible measurement errors. If the values differ too much from the average, it might give rise to variability.

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 data set, 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?

```
hd<-geyser['duration']%%1
hw<-geyser['waiting']%%10

hd<-as.vector(hd[['duration']])
hw<-as.vector(hw[['waiting']])

#hd1<-table(cut(hd,10))
#hw1<-table(cut(hw,10))
#chisq.test(hd1)
#chisq.test(hw1)
m <-10
minv <- min(hw)
maxv <- max(hw)
breaks <- c(minv, minv + cumsum(rep.int((maxv - minv) / m, m-1)), maxv)

f1<-hist(hd,breaks=10,plot = FALSE)

f2<-hist(hw,breaks=breaks,plot = FALSE)

# i
chisq.test(f1$counts)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  f1$counts
## X-squared = 153.61, df = 9, p-value < 2.2e-16
```

```
# ii
chisq.test(f2$counts)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  f2$counts
## X-squared = 9.194, df = 9, p-value = 0.4196
```

i. Since the p value is much lesser than 0.05, it shows strong evidence against the null hypothesis. So the H_d most likely does not follow a uniform distribution.

ii. Since the p value is greater than 0.05, it shows weak evidence against the null hypothesis. So the H_w most likely equiprobably any one of the digits $0, 1, 2, \ldots, 9$.

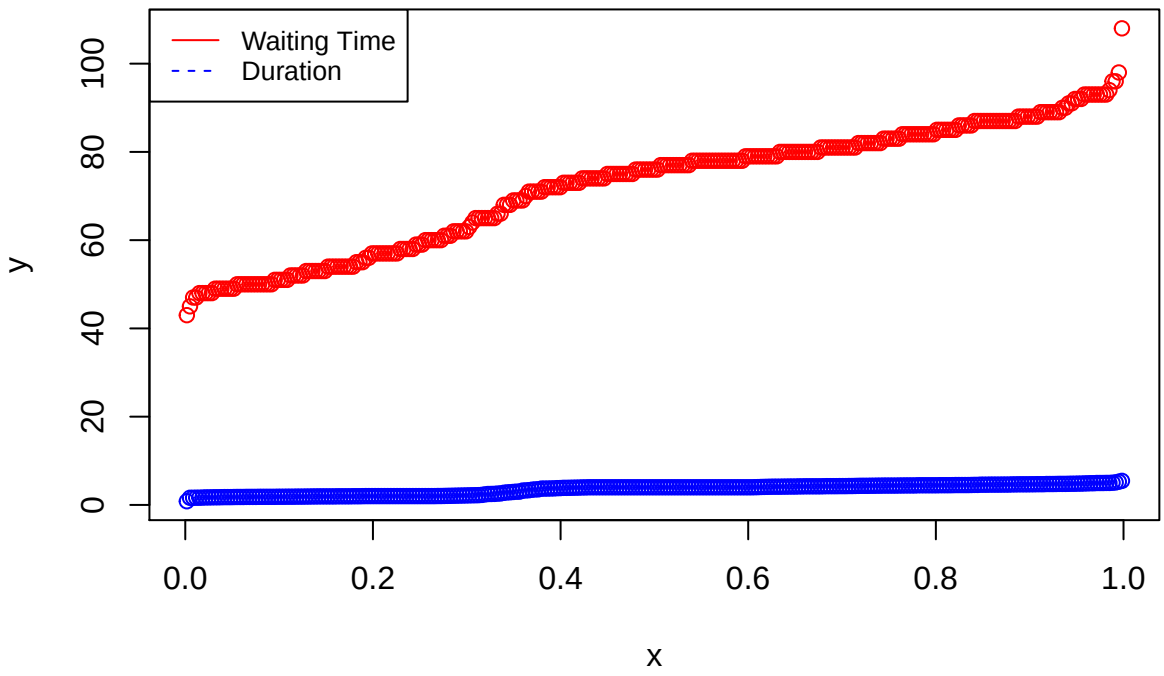
Since H_d most likely doesnt follow a uniform distribution, its not a good measuring system.

Since H_w most likely follows a equiprobable distribution, its a good measuring system.

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.

```
n<-nrow(geyser)
x<-ppoints(n)
y<-sort(geyser$duration)
y1<-sort(geyser$waiting)
plot(x,y,col='blue',ylim = range(c(y,y1)))
points(x,y1, col = "red")

legend("topleft", legend=c("Waiting Time", "Duration"), col=c("red", "blue"), lty=1:2, cex=0.8)
```

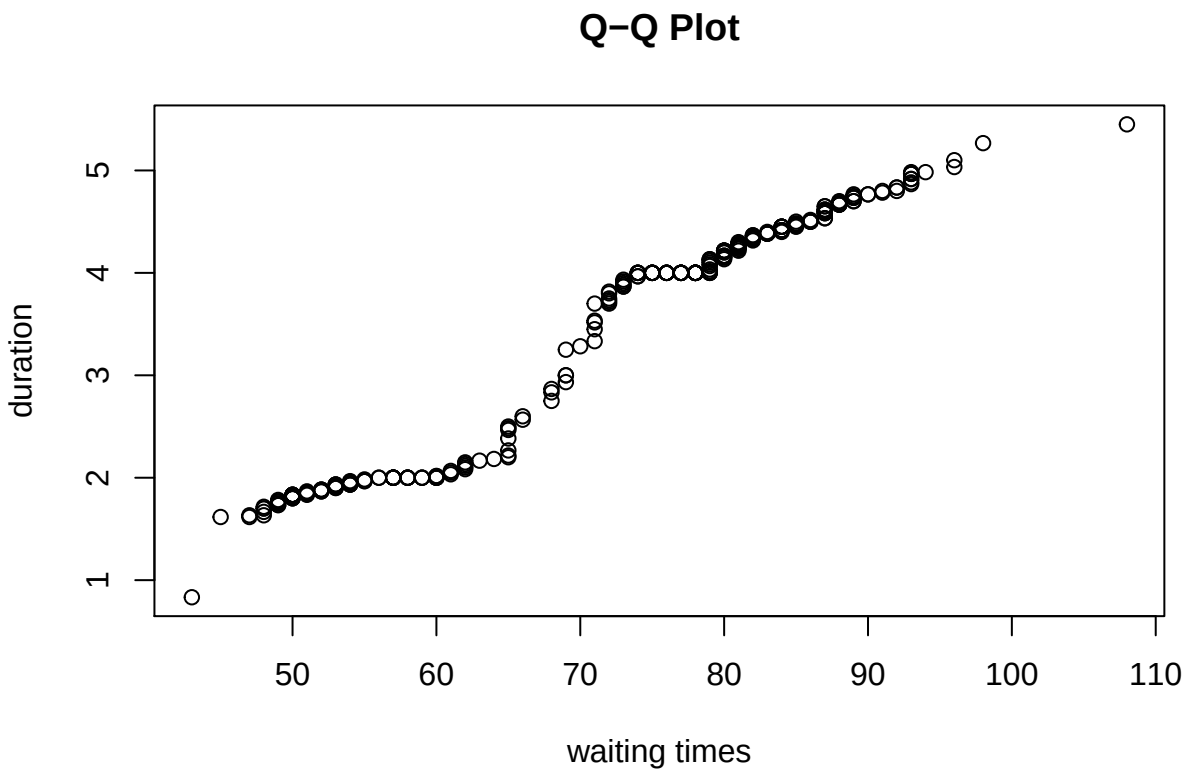


For the waiting times, the median is around 75. Its interquartile range seems to range from 60 to 80.

For the duration, the median is around 4. Its interquartile range seems to range from 2 to 5.

From the above sample quantile plot, we can see that the range of waiting times is higher than the range of duration time.The median of the waiting times is higher than the median of the duration times. The difference in slopes in the middle tells us that waiting times has a higher interquartile range than durations.

```
qqplot(geyser$waiting,geyser$duration,xlab='waiting times',ylab='duration',main='Q-Q Plot')
```



The above qqplot shows that these two data do not appear to have come from populations with the same distribution since it does not form a straight line.

j. *(10 marks)* Consider the waiting times w_i. We might ask whether waiting times could have been 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. A scatterplot of these pairs could be used to assess independendence.

Alternatively, we might first transform them to values which should be more nearly uniformly distributed. To that end, define

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

Now use the function `transform2uniform()` on the waiting times to give values $u_i = \widehat{Q}_W(w_i)$. You will now consider the independence of u_i and its lag k value u_{i-k}. If they are independent, the scatterplot of the $n - k$ pairs (u_{i-k}, u_i) should look like **uniform scatter** in the unit square.

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

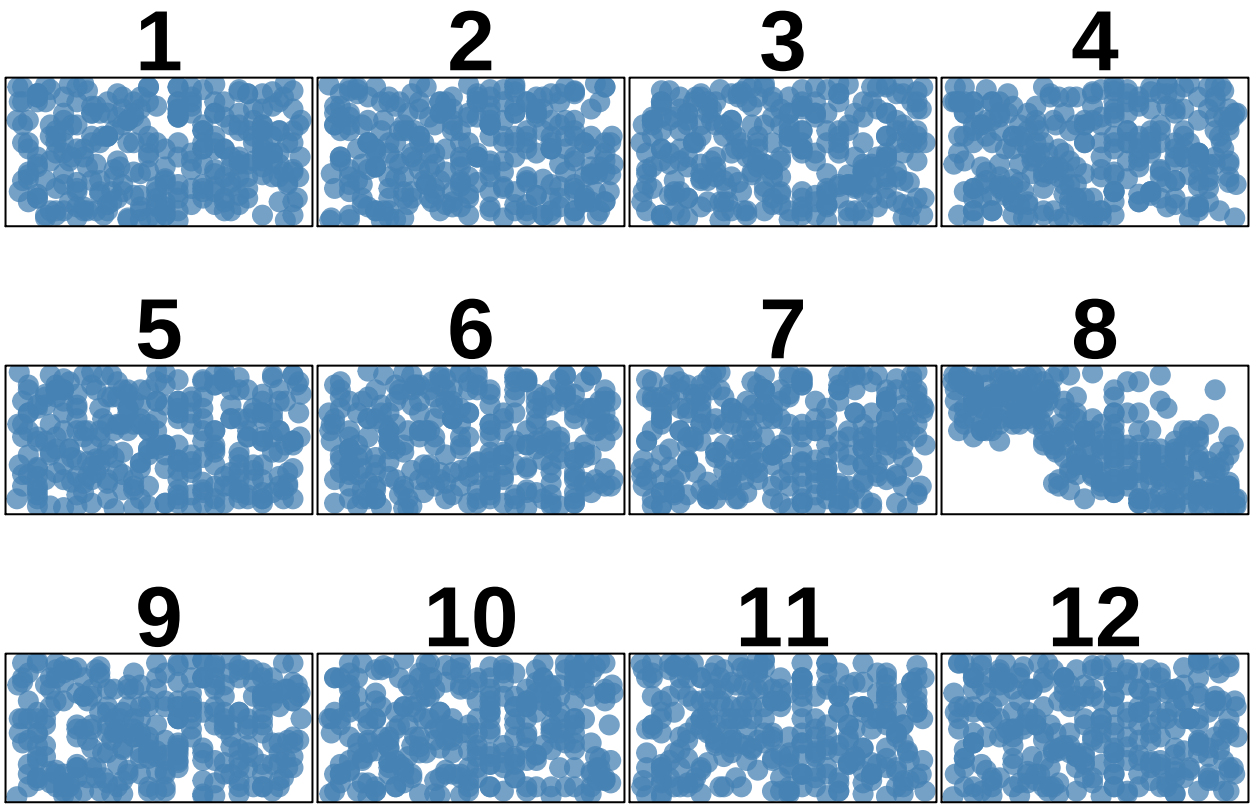
i. $k = 1$, the immediately preceding eruption, and
ii. $k = 22$, the eruption occurring roughly the day before.

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

```
t1<-transform2uniform(geyser$waiting)

#k=1
j=0
t11<-c()
t22<-c()
r<-length(geyser$waiting)
for(i in 1:(r-1)){
  t11[j]<-t1[i+1]
  t22[j]<-t1[i]
  j=j+1
}

data1 <- list(y = t11, x = t22)
lineup(data1,
generateSubject = mixCoords,
showSubject = showScatter,layout =c(3,4))
```

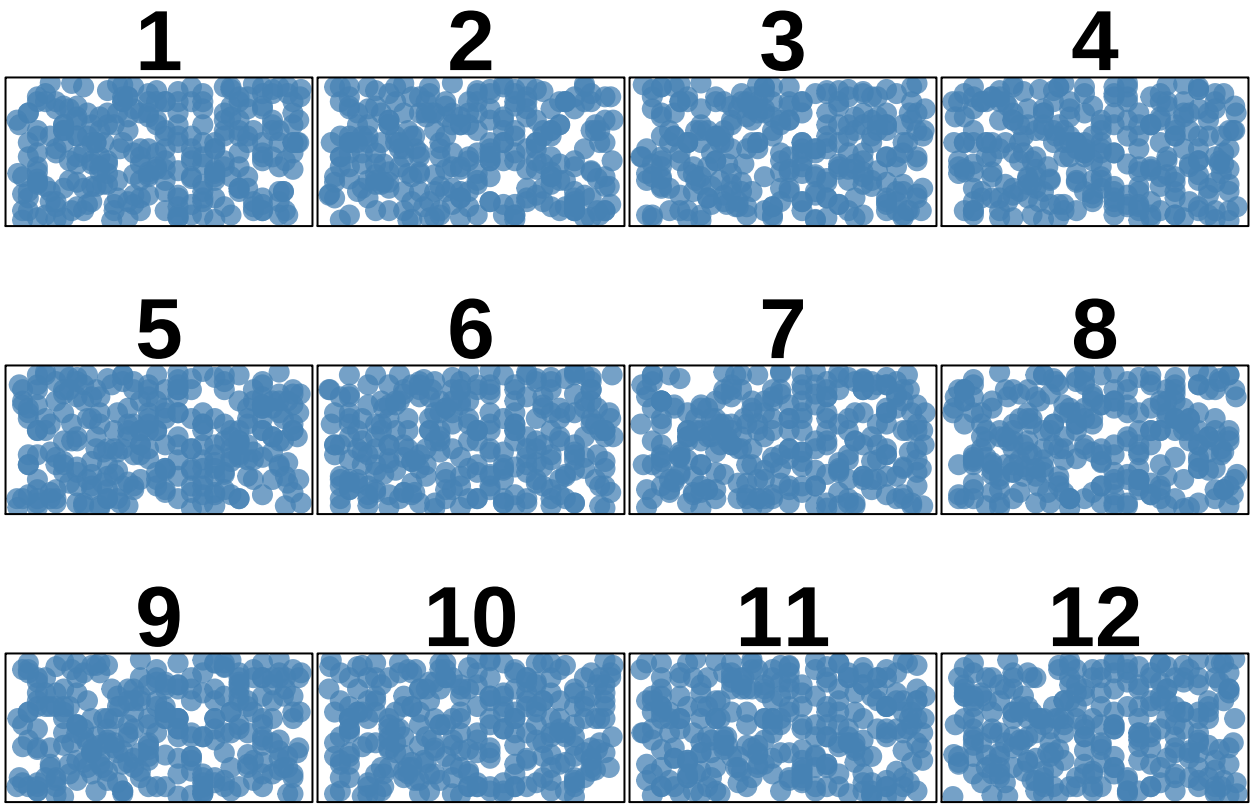



```
## $trueLoc
## [1] "log(2.57110087081438e+61, base=16) - 43"
```

```
#k=22
j=0
t33<-c()
t44<-c()
for(i in 1:(r-22)){
```

```
  #print(i)
  t33[j]<-t1[i+22]
  t44[j]<-t1[i]
  j=j+1
}

data2 <- list(y = t33, x = t44)
lineup(data2,
generateSubject = mixCoords,
showSubject = showScatter,layout =c(3,4))
```




```
## $trueLoc
## [1] "log(6.62473726694924e+24, base=12) - 15"
```

When k=1, the plot which seems to rejects the null hypothesis is the one which shows a negative correlation between current and previous waiting times. Which means they are most likely dependent on each other. When k=22, all the plots show a uniform distribution. The waiting times in this case appear to be independent of each other.

k. *(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

l. $k = 1$, the immediately preceding eruption, and

ii. $k = 22$, the eruption occurring roughly the day before.

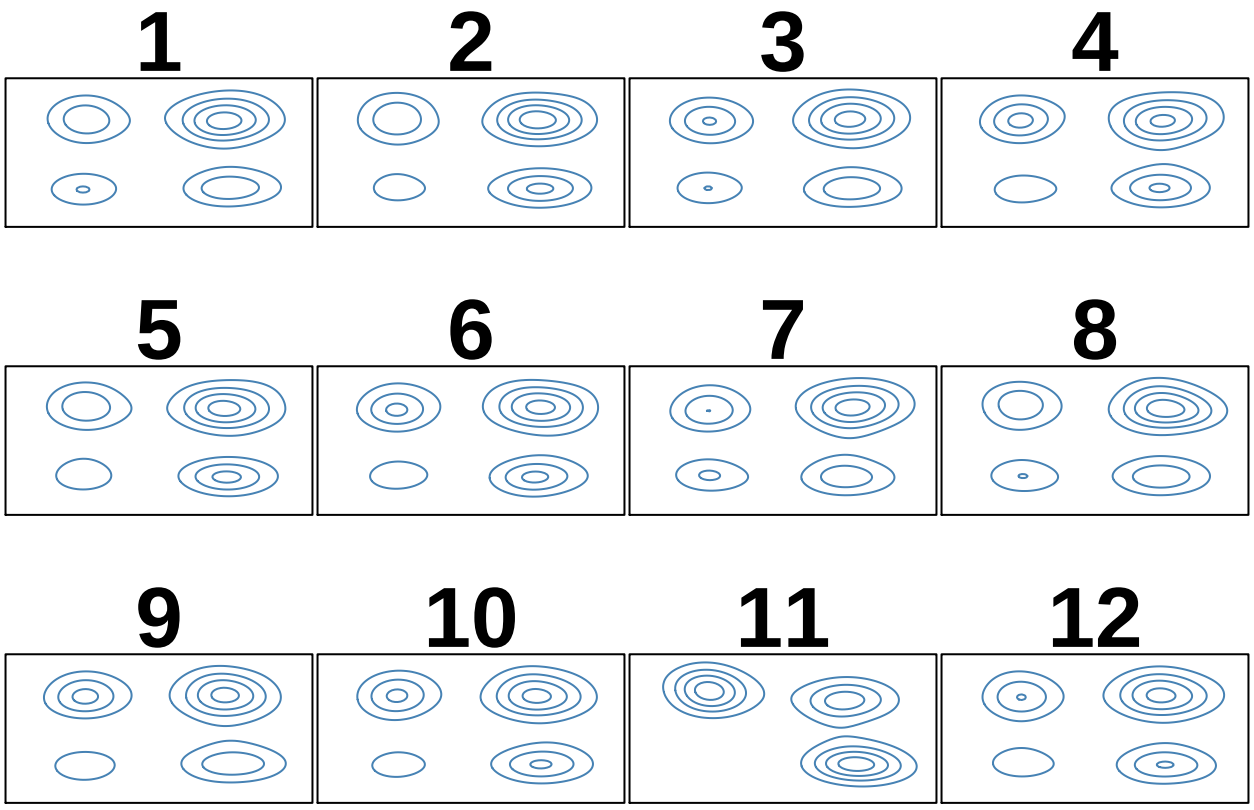
Show your code for constructing the necessary data and the lineup plots. What do you conclude about the

dependence between durations lengths?

```
l<-as.vector(geyser[['duration']])

#k=1
j=0
z11<-c()
z22<-c()
r<-length(geyser$duration)
for(i in 1:(r-1)){
  z11[j]<-l[i+1]
  z22[j]<-l[i]
  j=j+1
}


data3 <- list(y = z11, x = z22)
lineup(data3,
generateSubject = mixCoords,
showSubject = showDensityContours,
layout=c(3, 4))
```

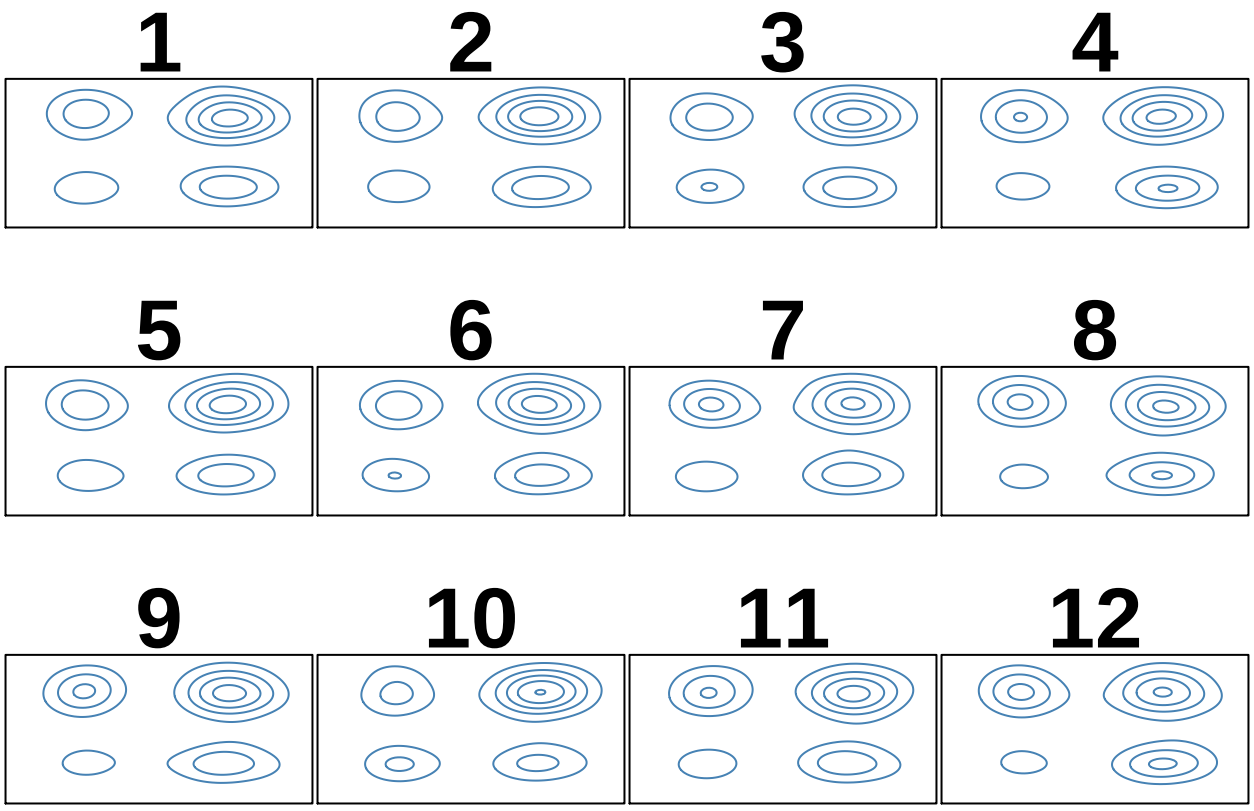



```
## $trueLoc
## [1] "log(8.14539297859635e+89, base=21) - 57"
```

```
#k=22
j=0
```

```
z33<-c()
z44<-c()
for(i in 1:(r-22)){
  #print(i)
  z33[j]<-l[i+22]
  z44[j]<-l[i]
  j=j+1
}


data4 <- list(y = z33, x = z44)
lineup(data4,
generateSubject = mixCoords,
showSubject = showDensityContours,
layout=c(3, 4))
```




```
## $trueLoc
## [1] "log(288230376151711744, base=4) - 24"
```

When k=1, the plot which seems to rejects the null hypothesis is the one which shows three density contour plots for duration lengths. Which means they most likely have some dependency on each other. When k=22, all the plots appear to be fairly similar. The duration times in this case appear to be independent of each other.