A5 (Q1)- Ram Yoogesh Gopu (20867060)

```
knitr::opts_chunk$set(echo = TRUE,
                      warning = FALSE,
                      message = FALSE,
                      fig.align = "center",
                      fig.width = 13,
                      fig.height = 18,
                      out.height = "40%")
set.seed(12314159)
library(loon.data)
library(loon)
```

```
## Loading required package: tcltk
```

```
library(gridExtra)

imageDirectory <- "./img"
dataDirectory <- "./data"
path_concat <- function(path1, path2, sep="/") paste(path1, path2, sep = sep)
```

Loading the dataset

```
library(MASS)
data(geyser)
```

(A)

The target population constitutes of all the eruptions from the Yellowstone National park. Individual unit is eruption. This eruption is constituted by by both waiting time and the duration.

(B)

Study population considers all the eruptions that has taken place till date. Individual unit is eruption. There might be study error because of various external factors such as the temperature during the study.

(C)

The dataset which was recorded from August 1st to August 15th, 1985 has 299 observations (Waiting time and the duration) which constitutes as the sample S. Individual Unit is eruption. There might be sample error because of the various factors which will affect the eruptions.

(D)

There are various ways in which it might produce sampling bias, (i.e) time, geological factors, temperature could affect the process and would fail to be accurate.

(E)

(i)

From the explanation we could infer that both the waiting time and the duration should be dependent on each other, i.e shorther the waiting time, longer the duration. So this pair would be of interest.

(ii)

These variates are dependent on each other, especially in 3a and 3b. There must be some external factors which could make these two independent, but from the above explanation they will be considered dependent unless explicitly mentioned so.

(iii)

So from the scenario we can say that the sucessive duration will be high if the precending duration is on lower range. on the contrary the sucessive duration cannot be predicted if the preceding duration is high. So it should be interesting to understand the relationship between these two variates.

(iv)

Similarly, the waiting time would be high for the eruption if the previous waiting time is less, which means they are dependent. On the contrary they are not dependent if the time is vice versa. So it should be interesting to understand the relationship between these two variates.

(F)

From the dataset, the other variate which would be of potential interest is the time required to finish the steaming process.

(G)

```r
digits <- function(number) {
  sampar <- array(rep(0, 10))
  for (i in number)
  {
    sampar[i + 1] <- sampar[i + 1] + 1
  }
  return(sampar)
}

h1 <- ((geyser$duration %% 1 ) * 10)
h2 <- (geyser$waiting %% 10)

# Performing the tests

chisq.test(digits(h1))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  digits(h1)
## X-squared = 152.61, df = 9, p-value < 2.2e-16
```

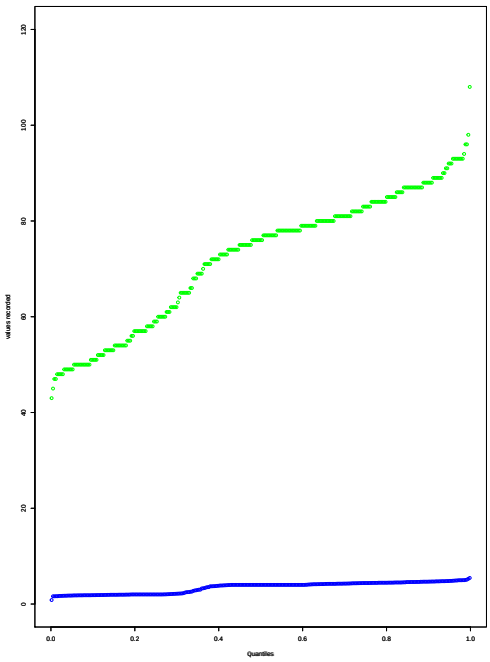
```r
chisq.test(digits(h2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  digits(h2)
## X-squared = 9.194, df = 9, p-value = 0.4196
```

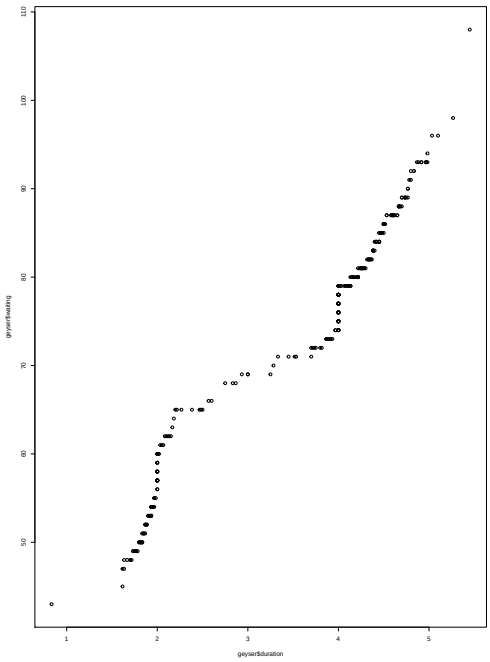
So from the results we can conclude that the p-value is really low (0.05), and the null hypothesis is rejected for H_d (Which means that it doesnt follow a U[0,1] distribution). Furthermore the p-value is > 0.05 and the null hypotheis is considered for H_w (the rightmost digit of the waiting time equiprobably any of the digits 0, 1, 2)

(I)

```r
dt <- ppoints(geyser$duration)
wt <- ppoints(geyser$waiting)

plot(dt, sort(geyser$duration), col = "blue", ylab = "values recorded", xlab = "quantiles",  ylim = c(0

par(new = TRUE)

plot(wt, sort(geyser$waiting), col = "green", ylab = "values recorded", xlab = "Quantiles", ylim = c(0,
```




```r
qqplot(geyser$duration, geyser$waiting)
```

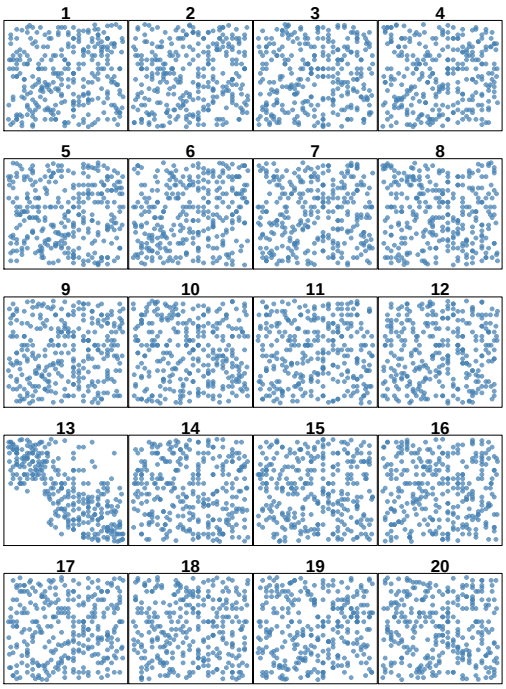


(J)

```r
transform2uniform <- function(x,
```

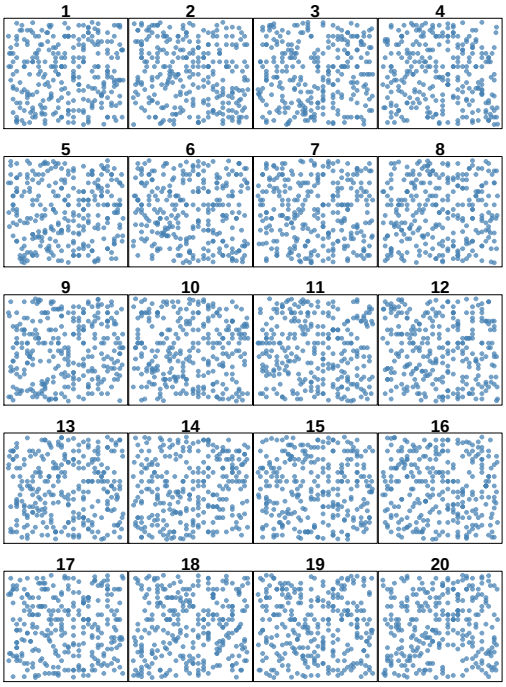
```
                                a = if(length(x) <= 10) 3/8 else 1/2,
                                ...) {
  (rank(x, ...) - a) / length(x)
}
```

```
# getting transformed values
source("graphicalTests.R")
source("numericalTests.R")
source("generateData.R")

geyser$t <- transform2uniform(geyser$waiting)

ck <- function(value, k = 1){
  return(list(x = value[1: (length(value) - k)], y = value[(k + 1) : (length(value))]))
}

genrator <- function (value) {
  val_y <- sample(value$y, length(value$x), replace = FALSE)
  return(list(x = value$x, y = val_y))
}

k1 <- ck(geyser$t)
k2 <- ck(geyser$t, 22)

lineup(k1, generateSubject = mixCoords, showSubject = showScatter)
```




```
## $trueLoc
## [1] "log(2.2397447421778e+102, base=16) - 72"
```

```
lineup(k2, generateSubject = mixCoords, showSubject = showScatter)
```



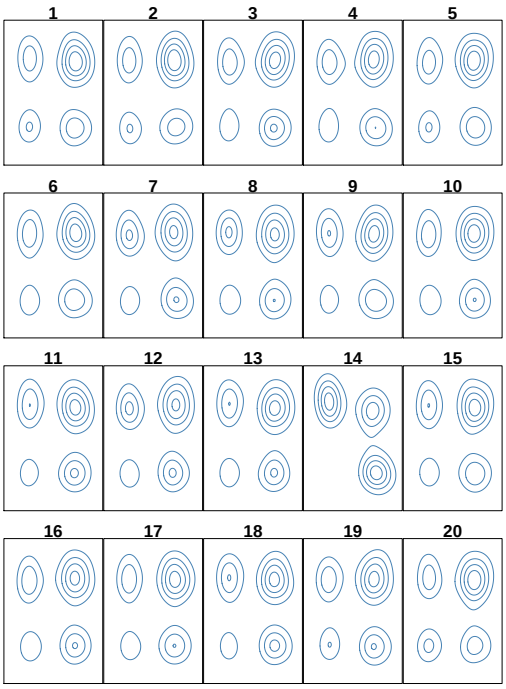

```
## $trueLoc
## [1] "log(1.02349036907747e+21, base=6) - 14"
```

for k = 1 from plot 1, there is a pattern which describes when x decreases, y increases (Box 13). This is against the hypothesis

for K = 22 from plot 2, there is a lack of pattern, which conveys that both are indpendent. i.e there is no evidence against the hypothesis.

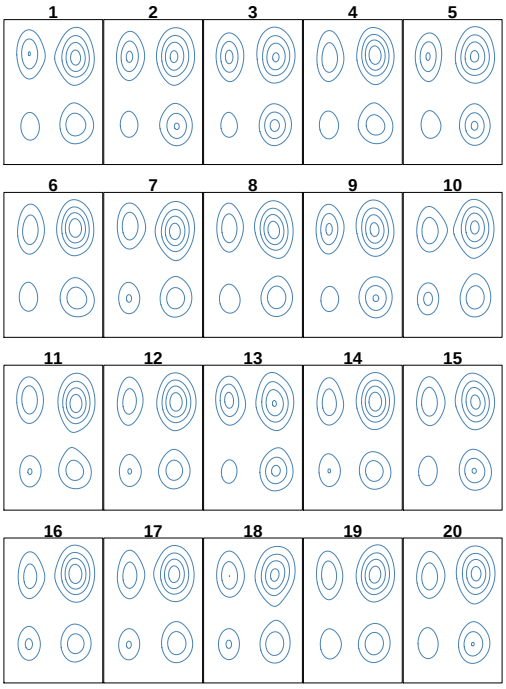
(K)

```
source("graphicalTests.R")

geyser$dt <- transform2uniform(geyser$duration)

denest <- function(value){
  val_y <- sample(value$y, length(value$x), replace = FALSE)
  return (list(x = value$x, y  = val_y))
}

k1 <- ck(geyser$duration)
k2 <- ck(geyser$duration, 22)

lineup(k1, generateSubject = denest, showSubject = showDensityContours, layout=c(4, 5))
```




```
## $trueLoc
## [1] "log(5.37533968658949e+91, base=12) - 71"
```

```
lineup(k2, generateSubject = denest, showSubject = showDensityContours, layout=c(4, 5))
```




```
## $trueLoc
## [1] "log(2.69599466671506e+67, base=16) - 37"
```

for K1 from the plot from box 1, it is evident that there remains a pattern between d_{i - 1} and d_i.This is an evidence which suggests that these two values are not likely to be independent.

for K = 22 from the second plot, there is a lack of pattern in any of the displayed boxes. So we can conclude that they are independent.