Question 1(a)

R Code:

```
set.seed(1716)
x1 <- rnorm(200, 10, 2)

x2 <- rnorm(200, 10, 2)
```

```
y1 <- summary(x1)
y2 <- summary(x2)
```

```
#plot x1 and x2 in same graph with different color
plot(x1, col = ("red"), pch = 16, main = "Overlaying graph of two
samples", xlab = "Observations", ylab = "Values")
points(x2, col = ("blue"), pch = 15)
legend("bottomleft", legend = c("x1", "x2"), col = c("red", "blue"),
pch = c(16, 15))
```

Outputs:

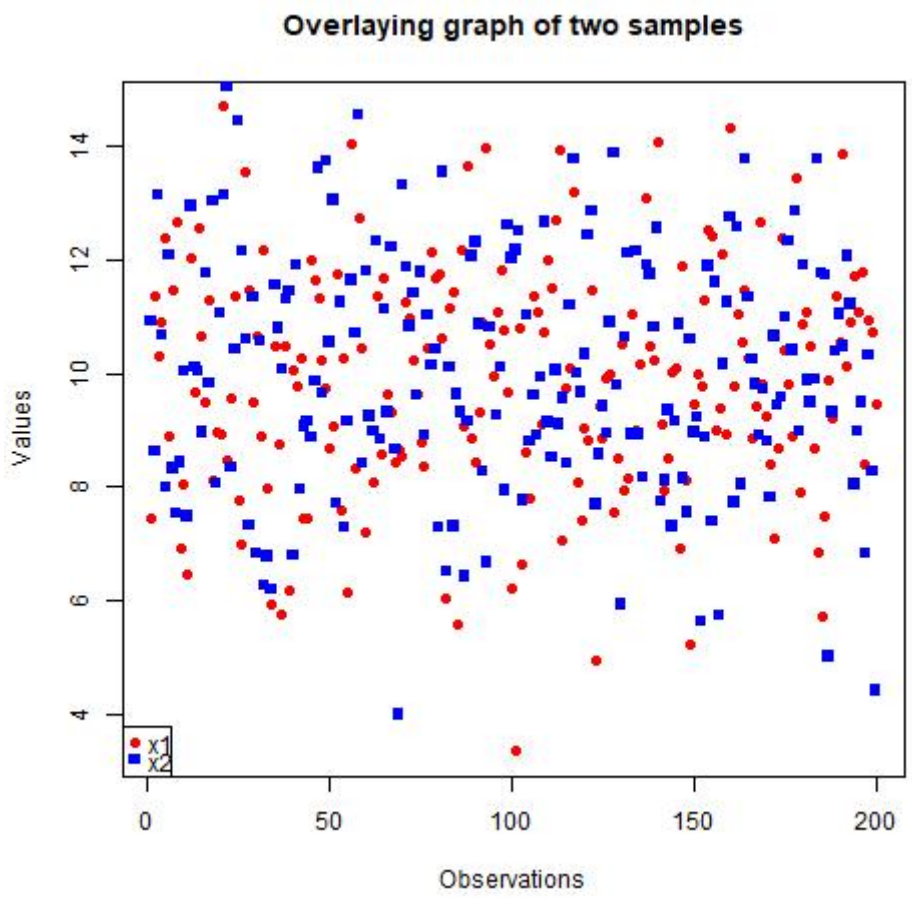
```
r$> y1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.348   8.568   9.902   9.857  11.289  14.709

r$> y2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.992   8.768  10.021  10.039  11.623  15.445
```

Comparison:

From the outputs it is visible that the summary statistics function generated the minimum, maximum, mean, median etc which differs for two samples. The reason of not being exactly the same of the two samples are the seed values though for both cases N(10,2) and the observation size is 200. For x1, the seed value is set to 1716 (ID = 21301716), thus the sample generates the random number with this exact same value. However, for x2 there is no seed value which made x2 a pseudo random sample set where it is arbitrary and can not be the same value set as sample set x1. The graph of the two sample set illustrates the difference of two samples.

Graph:



Question 1(b)

R Code:

```
set.seed(1716)
x1 <- rnorm(200, 5, 7)
set.seed(1716)
x2 <- rnorm(200, 5, 2)

y1 <- summary(x1)
y2 <- summary(x2)
```

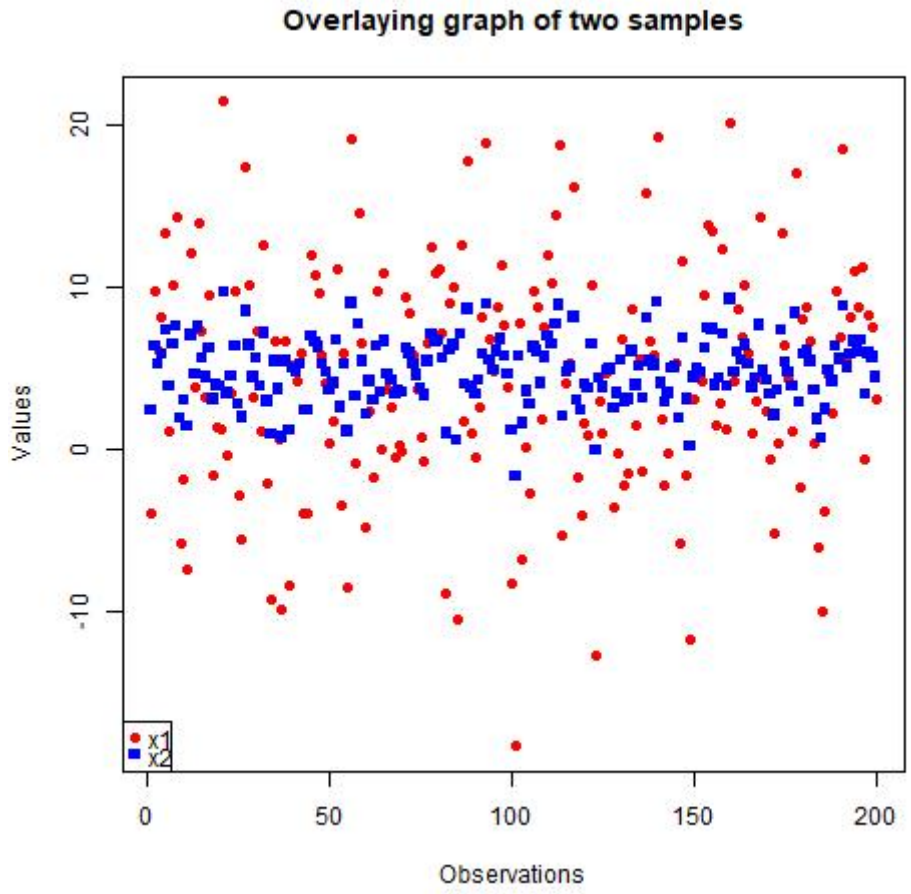
```
#plot x1 and x2 in same graph with different color
plot(x1, col = ("red"), pch = 16, main = "Overlaying graph of two
samples", xlab = "Observations", ylab = "Values")
points(x2, col = ("blue"), pch = 15)
legend("bottomleft", legend = c("x1", "x2"), col = c("red", "blue"),
pch = c(16, 15))
```

Outputs:

```
r$> y1
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-18.28274  -0.01115   4.65555   4.49775   9.51276  21.48075

r$> y2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.652   3.568   4.902   4.857   6.289   9.709
```

Graph:



Comparison:

The output and the graph illustrates the difference between two samples very clearly though they have the same seed value for this trial. Using the seed value of 1716 both samples generates the exactly same random numbers with 200 observations and the same mean that is 5. However, the result differs because of the variance. For first sample the variance is 7 and the other one is 2. The different value of the variance creates the difference in the maximum achievable value and also creates difference in the minimum achievable value for both samples. Thus, it has not achieved the same result.

Question 2(a)

R Code:

```
setwd("D:\\BRACU\\14th Semester\\STA301\\Project 1")
load("D:\\BRACU\\14th Semester\\STA301\\Project 1\\algae.RData")


#Create a new variable by combining the small and medium size rivers in
one category.
dataset_temp <- algae
dataset_temp <- dataset_temp[!is.na(dataset_temp$mnO2),]
```

```
dataset_temp$combined <- ifelse(dataset_temp$size == "large", "large",
"small/medium")
View(dataset_temp)
```

Question 2(b)

R Code:

```
#Analyze chem2 value for rivers of small/medium and large sizes,
separately, using summary statistics and different plots.

x1 <- dataset_temp$mnO2[dataset_temp$combined == "small/medium"]
x2 <- dataset_temp$mnO2[dataset_temp$combined == "large"]
```

```
y1 <- summary(x1)
y2 <- summary(x2)
```

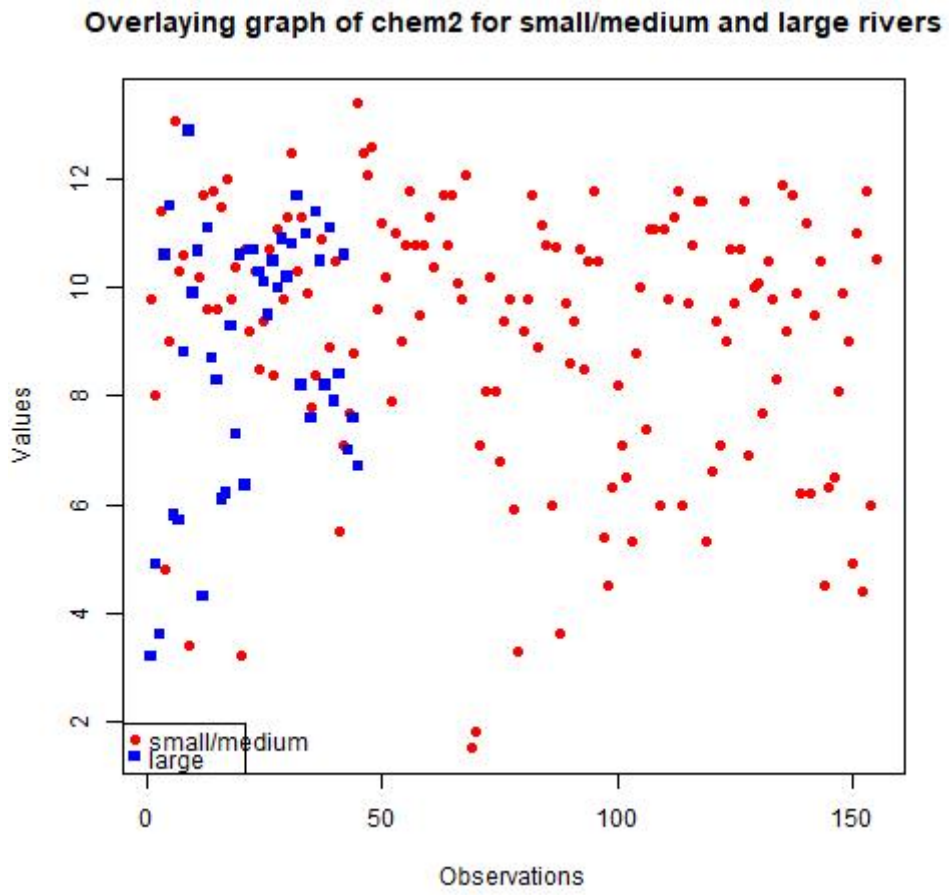
```
#plot x1 and x2 in same graph with different color
plot(x1, col = ("red"), pch = 16, main = "Overlaying graph of chem2 for
small/medium and large size rivers", xlab = "Observations", ylab =
"Values")
points(x2, col = ("blue"), pch = 15)
legend("bottomleft", legend = c("small/medium", "large"), col = c("red",
"blue"), pch = c(16, 15))
```

Outputs:

```
r$> y1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.500   8.000   9.800   9.202  10.800  13.400

r$> y2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.200   7.300   9.500   8.832  10.680  12.900
```

Graph:


Overlaying graph of chem2 for small/medium and large rivers
Values
Observations
small/medium
large
0
50
100
150
2
4
6
8
10
12

Question 2(c)

R Code:

```
#Is there a significant difference in mean chem2 value for rivers of
small/medium and large sizes? Carry out an appropriate test. Include
the appropriate hypotheses, test statistic value, p-value, and
conclusion.

test <- t.test(y1, y2, paired = TRUE)
```

Outputs:

```
r$> source("d:\\BRACU\\14th Semester\\STA301\\Project 1\\project1.R", encoding = "UTF-8")

r$> test

        Paired t-test

data:  y1 and y2
t = 0.13498, df = 5, p-value = 0.8979
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.8733770  0.9701787
sample estimates:
mean difference
     0.04840087
```