HW6

110078509 And credit to 110078503

20220326

Question 1

(a). Pick a reshaping package (we discussed two in class) – research them online and tell us why you picked it over others (provide any helpful links that supported your decision).

Ans:

First, both of the tidyr (gather) and reshape(melt) can complete the task as proof below. However, there is a slightly difference.

According to the link:

tortrial: https://jtr13.github.io/spring19/hx2259_qz2351.html (https://jtr13.github.io/spring19/hx2259_qz2351.html)

And the explanation: https://localcoder.org/tidyrgather-vs-reshape2melt-on-matrices (https://localcoder.org/tidyrgather-vs-reshape2melt-on-matrices)

mentioned that “Note that tidyr is designed for use in conjunction with dplyr, so you should always load both.” Though we don’t have to install dplyr, we still can implement tidyr(gather) in our current edition, however, it’s suggested to install these 2 packages at the same time for better package compatibility. Personally, I use dplyr frequently to deal with dataframe. So, both package are acceptable for me. Therefore, I will demo two way for the following question.

Last but not least, the tidyr is more limited for the reshape work. Hence, I would recommend “reshape(melt)” to the people who like to use a “easier” package setting.

(b). Show the code to reshape the versizon_wide.csv data

```
#We got plenty NA data in the col 'CLEC'. Set na.rm = TRUE
raw.wide <- read.csv("verizon_wide.csv", header = T)

# By "reshape2"
loads_long.re <- melt( data = raw.wide,na.rm = TRUE,
                   variable.name = "set",
                   value.name = "Time")
```

```
## No id variables; using all as measure variables
```

```
# By "tidyr"
loads_long.ti <- gather( raw.wide,na.rm = TRUE,
                     key = "set",
                     value = "Time")
```

p.s: In scenario of question (b) , melt func from reshape2 required us to set the var.id and show the warming “No id variables; using all as measure variables” It would not affect the result because if its blank, it will use all non-measured variables to replace it.(‘Warming’ would not affect the result, it still can compile successfully)

(c). head & tail

```
head(loads_long.ti)
```

```
##    set  Time
## 1 ILEC 17.50
## 2 ILEC  2.40
## 3 ILEC  0.00
## 4 ILEC  0.65
## 5 ILEC 22.23
## 6 ILEC  1.20
```

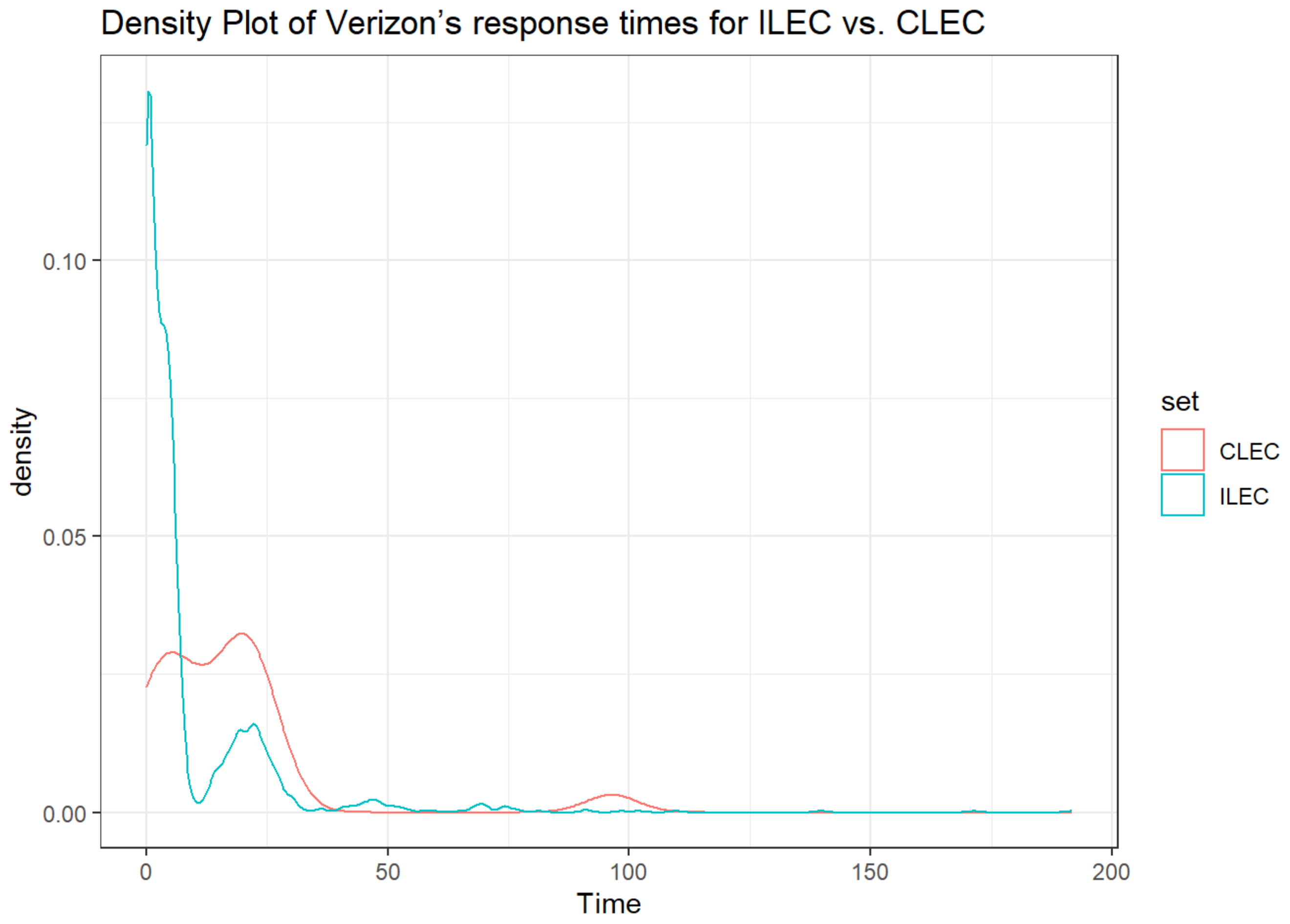
```
tail(loads_long.ti)
```

```
##       set  Time
## 1682 CLEC 24.20
## 1683 CLEC 22.13
## 1684 CLEC 18.57
## 1685 CLEC 20.00
## 1686 CLEC 14.13
## 1687 CLEC  5.80
```

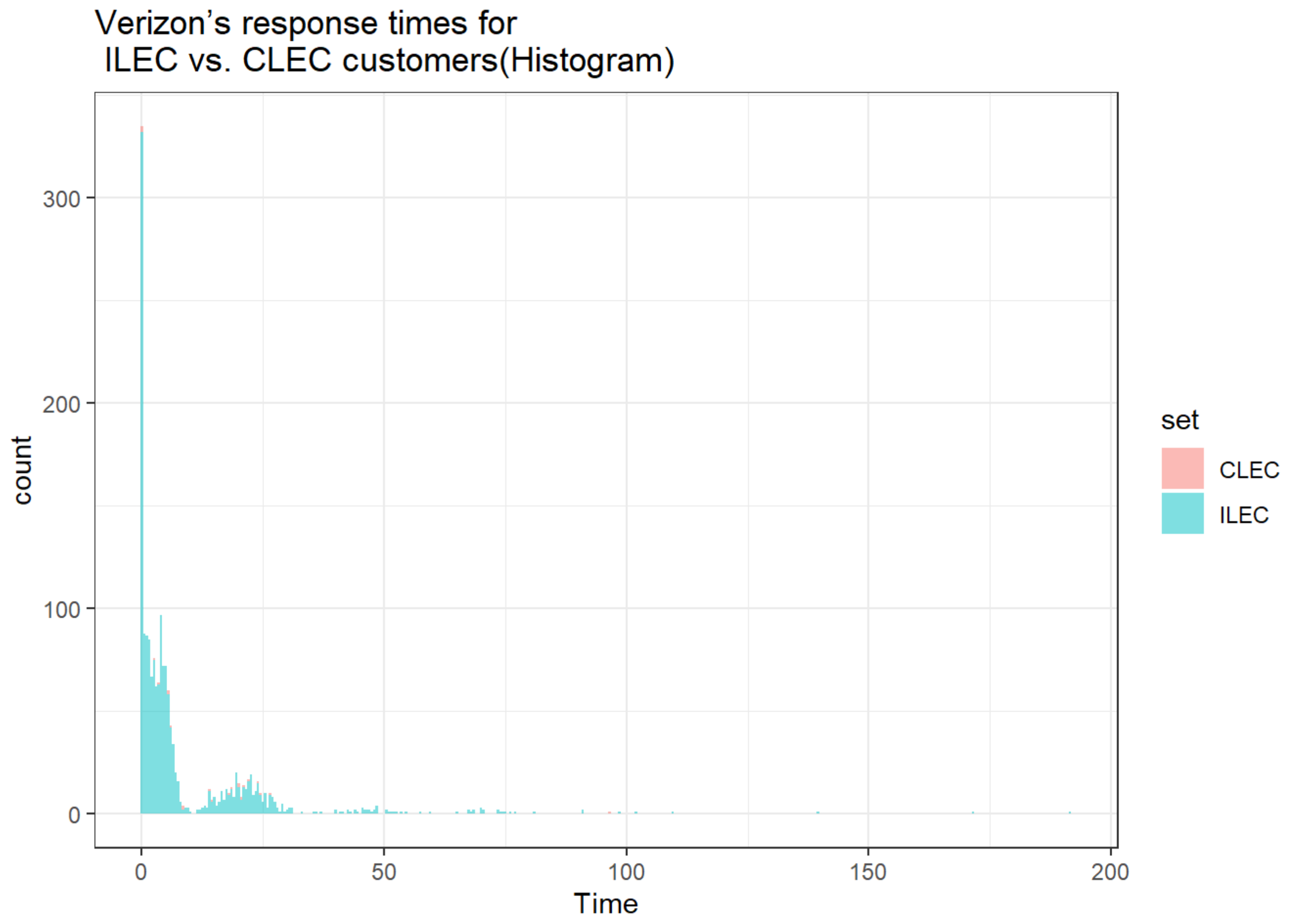
(d). Visualize Verizon’s response times for ILEC vs. CLEC customers

- Plot ILEC , CLEC together

```
ggplot(loads_long.ti, aes(x = Time)) +
    geom_density(aes(color = set)) +
    theme_bw() +
    labs(title = "Density Plot of Verizon’s response times for ILEC vs. CLEC ")
```



```
ggplot(loads_long.ti,aes(x=Time, fill=set))+ geom_histogram(binwidth=.5, alpha=1/2)+theme_bw
() +labs(title = "Verizon’s response times for
 ILEC vs. CLEC customers(Histogram)")
```

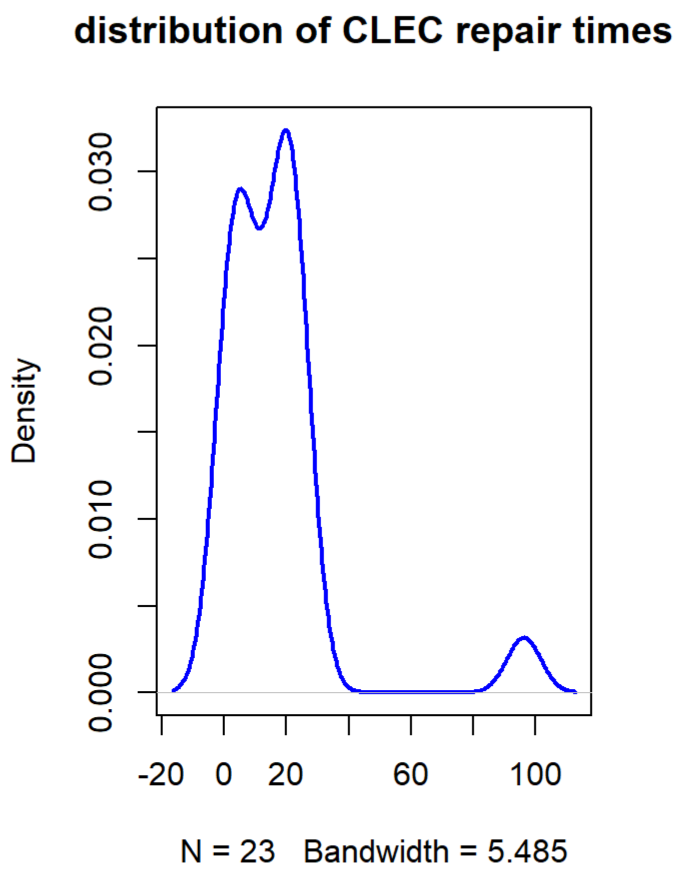


- Plot ILEC , CLEC separately with build-in function

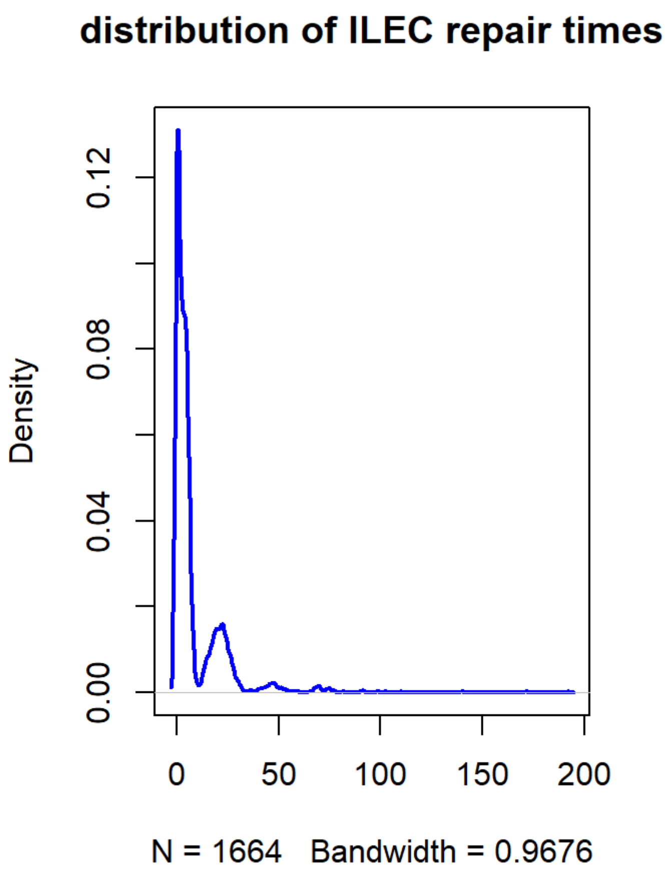
```
#Set -  CLEC
clec <- loads_long.ti %>% filter( loads_long.ti$set == "CLEC")
clec.t<-clec$Time

#Set -  ILEC
ilec<-loads_long.ti %>% filter( loads_long.ti$set == "ILEC")
ilec.t<-ilec$Time


#Set -  CLEC & ILEC Graph
par(mfrow=c(1,2))
plot(density(clec.t), lwd=2, col="blue", main="distribution of CLEC repair times")
plot(density(ilec.t), lwd=2, col="blue", main="distribution of ILEC repair times")
```



distribution of CLEC repair times



distribution of ILEC repair times

Question 2

Let’s test if the mean of response times for CLEC customers is greater than for ILEC customers

---

a.State the appropriate null and alternative hypotheses (one-tailed)

- Ans:

Let

the difference in means = the mean (CLEC customers) - the mean (ILEC customers)

H0: μ of Resonse Time for CLEC - μ of Resonse Time for ILEC <= 0 H0: μ of Resonse Time for CLEC - μ of Resonse Time for ILEC > 0

---

b.Test the difference between the mean of ILEC versus CLEC response times at 1% significance.

(b-i). Conduct the test assuming variances of the two populations are equal

Preface:

Because we have two set of data(clec.t,ilec.t), and the service time of the each set of user supposed to be independent in the current assumption.And the variance of each are set as equal so we can skip the variance test (var.test).

We could conduct Two-Sample independent Student t-Test as below:

```
# Student's Two-Sample t-Test
# var.equal=TRUE
t.test(raw.wide$CLEC, raw.wide$ILEC,
       alternative = "greater",
       conf.level = 0.99,
       mu = 0,
       var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  raw.wide$CLEC and raw.wide$ILEC
## t = 2.6125, df = 1685, p-value = 0.004534
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  0.8801387       Inf
## sample estimates:
## mean of x mean of y
## 16.509130  8.411611
```

Ans:

The p-value is 0.004534 < 0.01. Hence, we reject H0. Then we accept alternative hypothesis that true difference in means is larger than 0.

(b-ii). Conduct the test assuming variances of the two populations are not equal

```
# Welch's Two-Sample t-Test
# var.equal= FALSE

t.test(raw.wide$CLEC, raw.wide$ILEC,
       alternative = "greater",
       conf.level = 0.99,
       mu = 0,
       var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  raw.wide$CLEC and raw.wide$ILEC
## t = 1.9834, df = 22.346, p-value = 0.02987
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  -2.130858       Inf
## sample estimates:
## mean of x mean of y
## 16.509130  8.411611
```

Ans:

The p-value is 0.02987 > 0.01. Hence, we cannot reject H0. Therefore, the true difference in means is less than or equal to 0

c. Use a permutation test to compare the means of ILEC vs. CLEC response times

```
# Permutation: Method definition

permute_diff <- function(values, groups) {
    permuted <- sample(values, replace = FALSE)
    grouped <- split(permuted, groups)
    permuted_diff <- mean(grouped[[1]]) - mean(grouped[[2]])
}
```

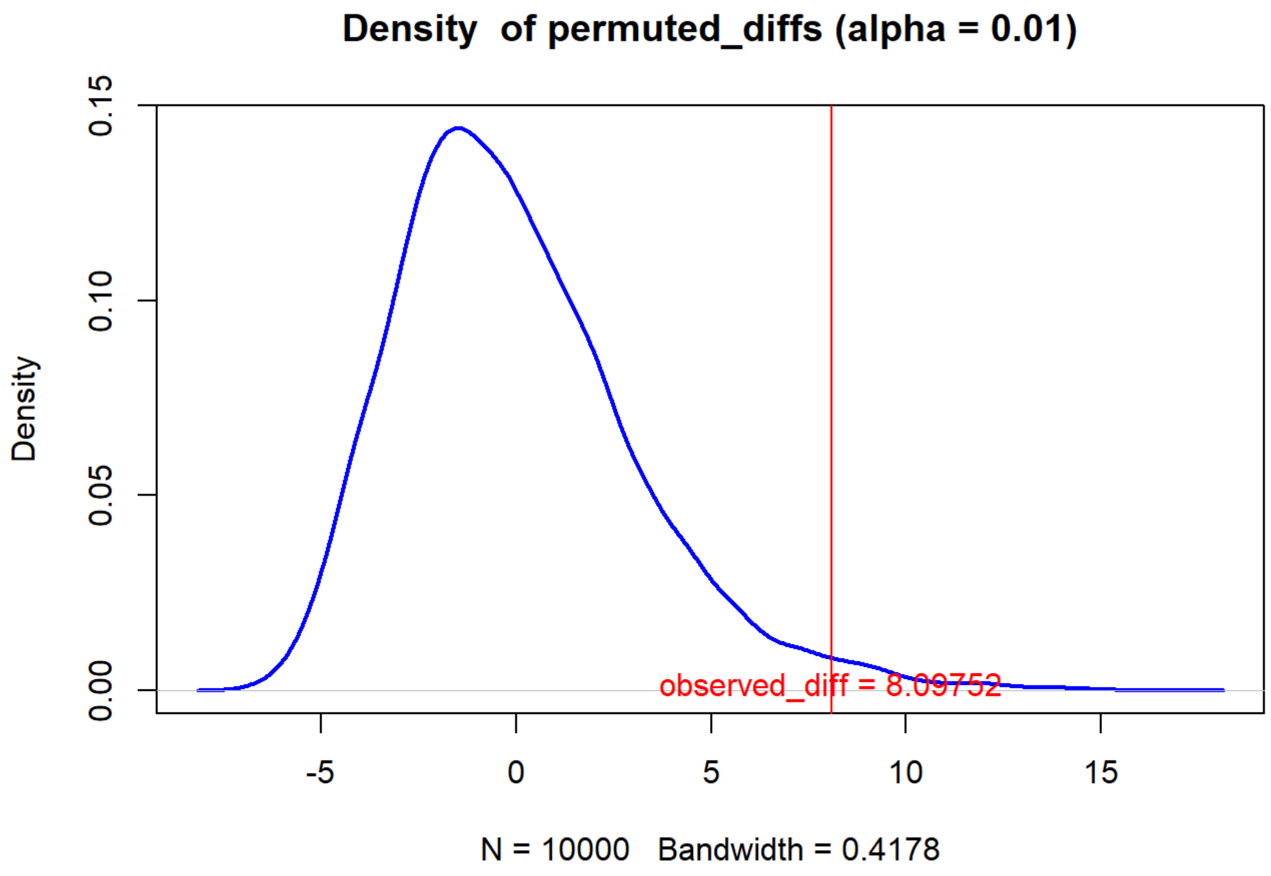
(c-i). Visualize the distribution of permuted differences, and indicate the observed difference as well.

```
# Observed Difference
observed_diff <- mean(clec.t) - mean(ilec.t);observed_diff
```

```
## [1] 8.09752
```

```
#
nperms <- 10000
permuted_diffs <- replicate(nperms, permute_diff(loads_long.ti$Time, loads_long.ti$set))

# Plot the distribution
plot(density(permuted_diffs), lwd=2, col="blue", main="Density  of permuted_diffs (alpha = 0.
01)")
abline(v=observed_diff, col = "red") #8.09752
text(x =observed_diff, y =0.001 , 'observed_diff = 8.09752', col = "red", cex = 1)
```



Ans:

The observed difference = 8.09752 as shown above.

(c-ii). What are the one-tailed and two-tailed p-values of the permutation test?

```
p_1tailed <- sum(permuted_diffs > observed_diff) / nperms

p_2tailed <- sum(abs(permuted_diffs) > observed_diff) / nperms

sprintf("one-tailed p-values: %f  &  two-tailed p-values:  %f ",p_1tailed, p_2tailed )
```

```
## [1] "one-tailed p-values: 0.018800  &  two-tailed p-values:  0.018800 "
```

Ans:

p-values: 0.019200 in 10000 permutations in both.

(c-iii). Would you reject the null hypothesis at 1% significance in a one-tailed test?

Ans:

No, I would not reject the null hypothesis at 1% significance in a one-tailed test because

0.019200 > 1% as the number of permutations = 10000.

Question 3

Let's use the Wilcoxon test to see if the response times for CLEC are different than ILEC.

a. Compute the W statistic comparing the values.

the permutation approach / or the rank sum approach.

Wilcoxon test

```
companys <-split(x=loads_long.ti$Time,f=loads_long.ti$set)
gt_eq<-function(a, b) {ifelse(a > b, 1, 0) +ifelse(a == b, 0.5, 0)}
W <- sum(outer(companys$CLEC,companys$ILEC,FUN =gt_eq));W
```

```
## [1] 26820
```

b. Compute the one-tailed p-value for W.

```
length_clec.t <- length(clec.t)
length_ilec.t <- length(ilec.t)
wilcox_p_1tail<-1- pwilcox(W, length_clec.t, length_ilec.t); wilcox_p_1tail
```

```
## [1] 0.0003688341
```

- Ans: one-tailed p-value for W: 0.0003688341

c. Run Wilcoxon Test again using the wilcox.test() function

Make sure you get the same W as part [a]. Show the results.

```
wilcox.test(clec.t,ilec.t, alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  clec.t and ilec.t
## W = 26820, p-value = 0.0004565
## alternative hypothesis: true location shift is greater than 0
```

Ans

W = 26820 is the same as the previous answer.

(d). At 1% significance and one-tailed, would you reject the null

hypothesis that the values of CLEC and ILEC are different from one another?

Ans

p-value =0.0004565 < 0.01, reject H0, and accept H1. It represents that there is significant different between the CLEC and ILEC.

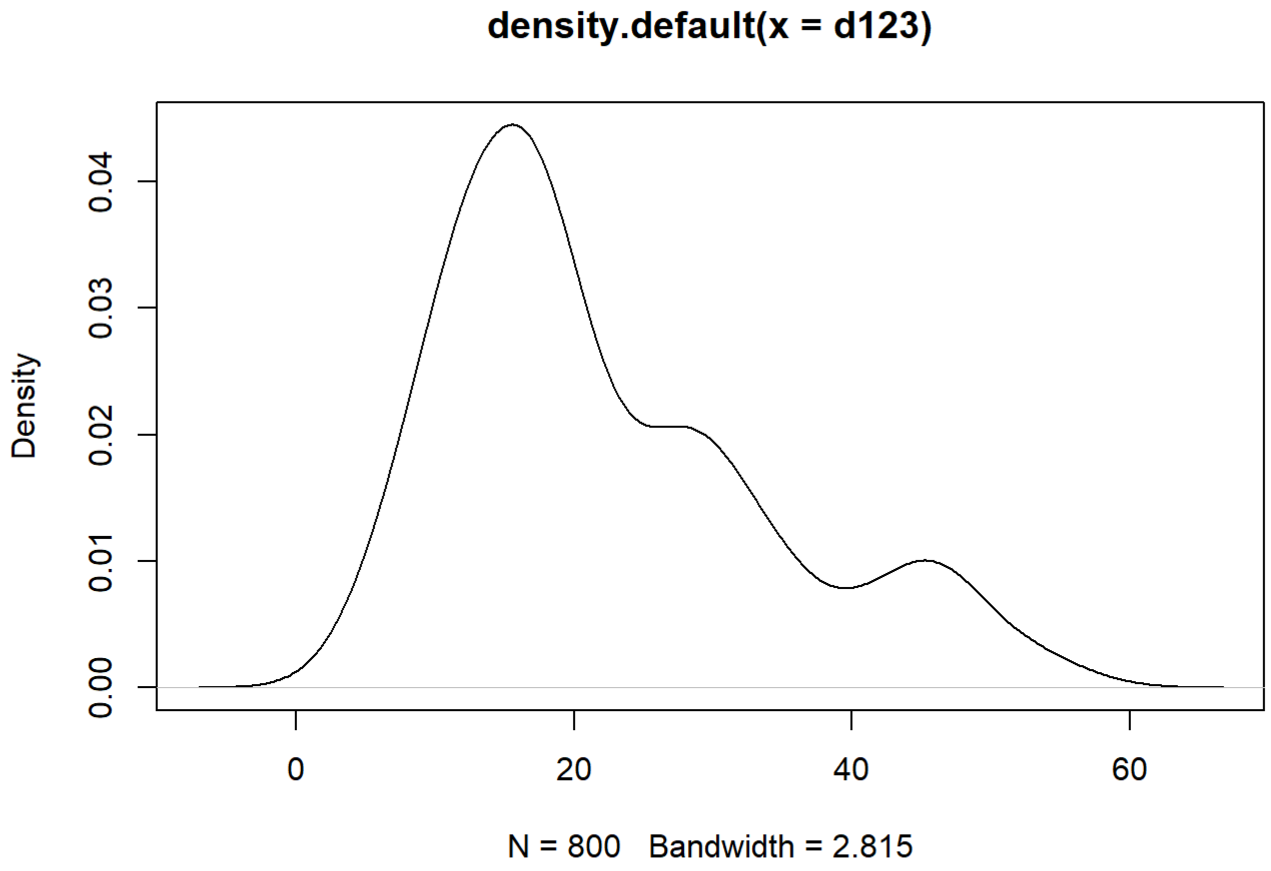
Question 4

a. make the function norm_qq_plot

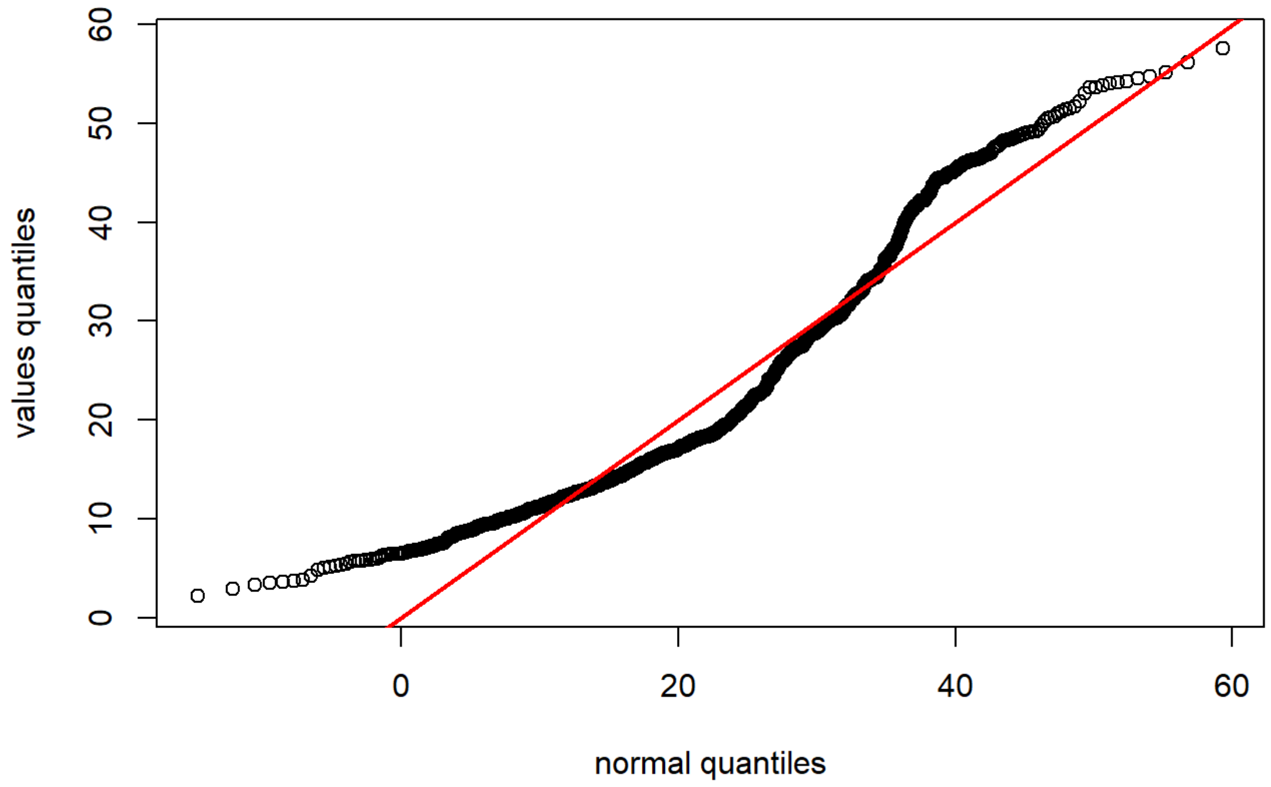
```
norm_qq_plot <- function(values){
    probs1000 <- seq(0, 1, 0.001);
    q_vals <- quantile(na.omit(values), probs1000);
    q_norm<-qnorm(probs1000,mean(values),sd(values));
    plot(q_norm, q_vals, xlab="normal quantiles", ylab="values quantiles");
    abline(0, 1, col="red", lwd=2)
}
```

b. Confirm that your function norm_qq_plot via d123 provided

```
plot(density(d123))
```



```
norm_qq_plot(d123)
```

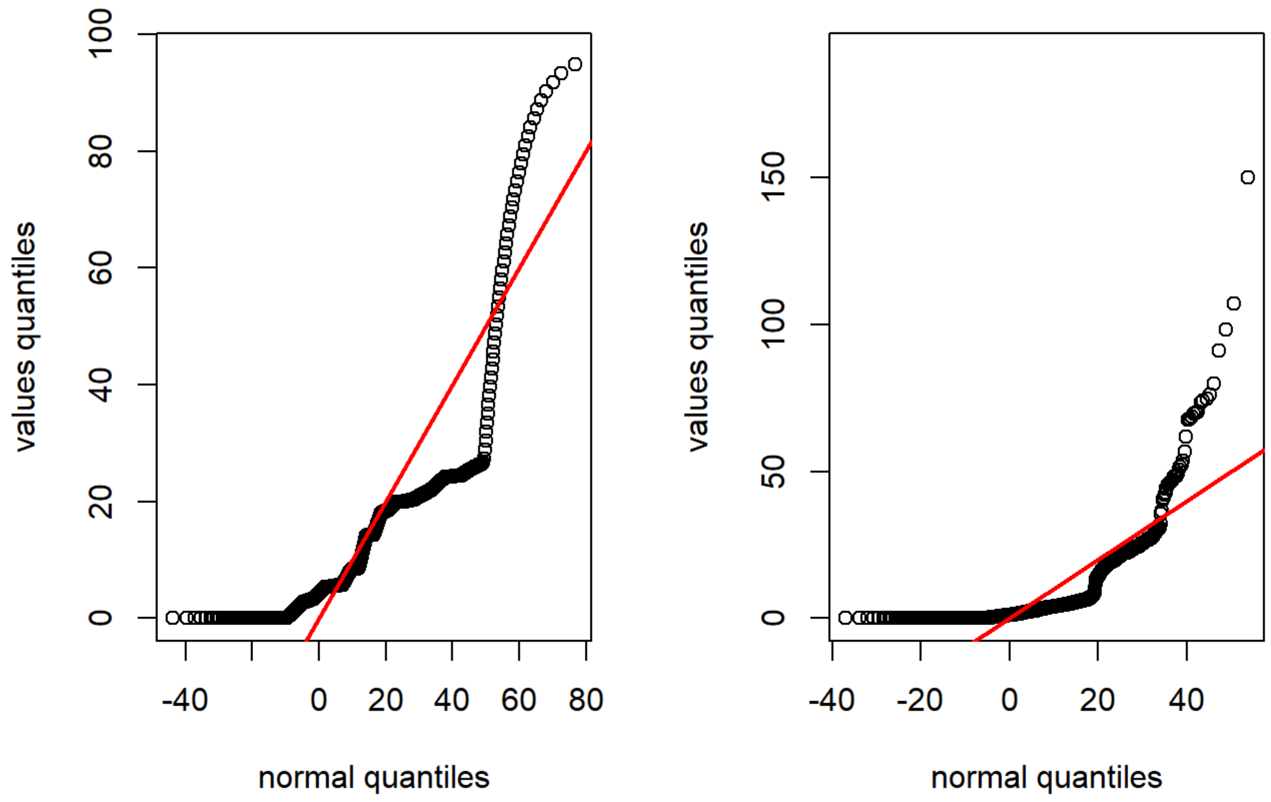


Explain:

It show that dataset d123 has “fat tails”. It means that there’re less data located in the central part of the distribution. For the terms of quantitles, the Q1 is much less than the theoretical normal distribution quantile and the last quantile is much bigger than the last theoretical quantile.So d123 is not normal distribution.

c. Apply the CLEC and ILEC samples to Q-Q plot function. What’s your conclusion?

```
par(mfrow=c(1,2))
norm_qq_plot(clec.t)
norm_qq_plot(ilec.t)
```



Ans:

In the Q-Q Plot, the scatter points that are away from the straight line and tend to be horizontal are the locations of the peaks of the data distribution. According to the plot above, these 2 dataset are right-tail(Positive Skew). Therefore, they are not normal distribution.