HW6

110078509

20220326

Question 1

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

(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 above. However, there is a slightly difference.

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)

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.”

Even 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. However, I use dplyr frequently do deal with dataframe. So, both package are okay for me. Therefore, I will demo two way for the following question.

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

```
# 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
```

```
head(loads_long.re,2)
```

```
##    set Time
## 1 ILEC 17.5
## 2 ILEC  2.4
```

```
# By "tidyr"
loads_long.ti <- gather( raw.wide,na.rm = TRUE,
                     key = "set",
                     value = "Time")
head(loads_long.ti,2)
```

```
##    set Time
## 1 ILEC 17.5
## 2 ILEC  2.4
```

(c). Show us the “head” and “tail” of the data to show that the reshaping worked

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)
```

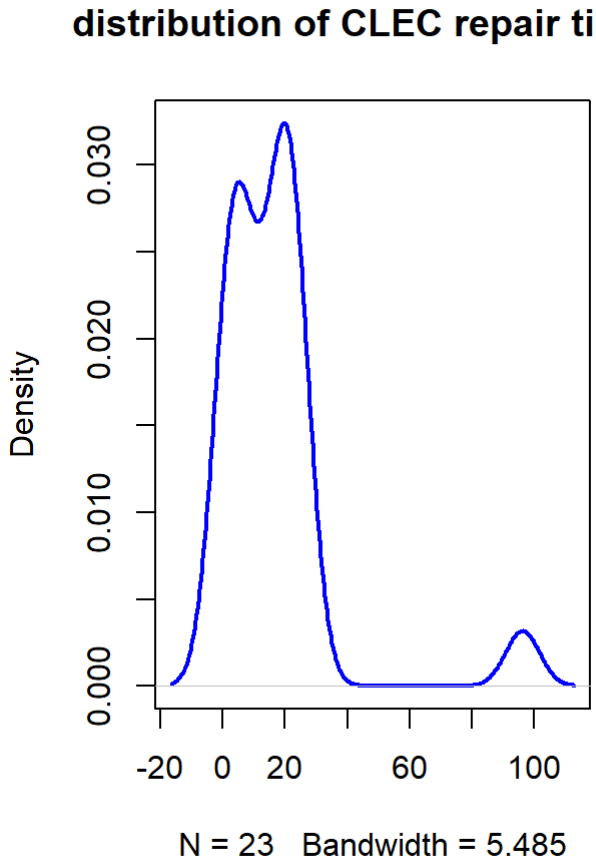
```
##       set  Time
## 1682 CLEC 24.20
## 1683 CLEC 22.13
## 1684 CLEC 18.57
## 1685 CLEC 20.00
## 1686 CLEC 14.13
## 1687 CLEC  5.80
```

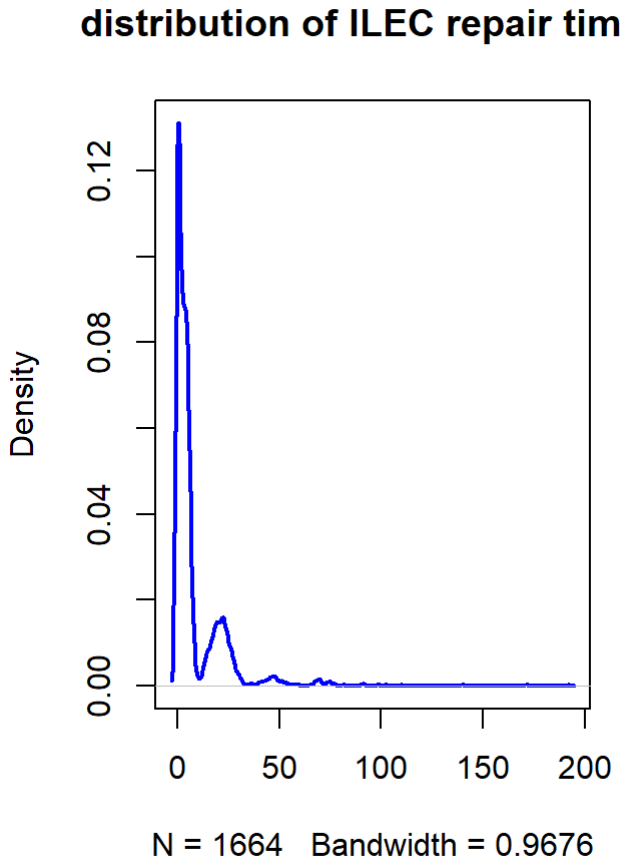
d. Visualize Verizon’s response times for ILEC vs. CLEC customers

```
#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")
```





(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: The difference in means is less than or equal to 0 H1: The difference in means is larger than 0.

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

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

- Ans:

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.

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]])
}
```

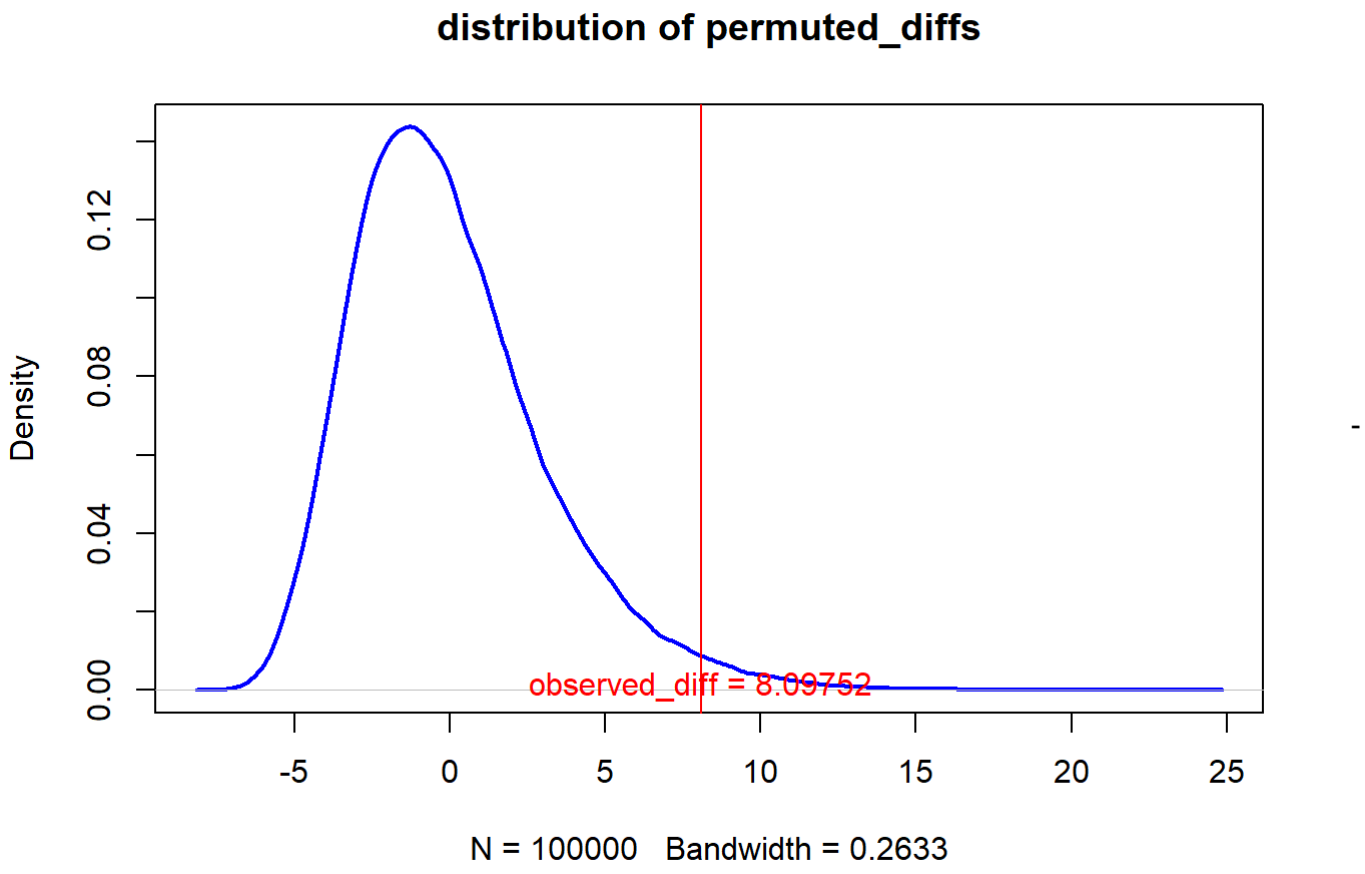
i. Visualize the distribution of permuted differences, and indicate the observed difference as well.

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

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

```
#
nperms <- 100000
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="distribution of permuted_diffs")
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.

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: 0.018810  &  two-tailed p-values:  0.018810 "
```

- *Ans:*

1.7920% in 100,000 permutations in both.

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

- *Ans:*

Yes, I would not reject the null hypothesis at 1% significance in a one-tailed test because

0.017920 > 0.01 as the number of permutations = 100,000.

Question 4

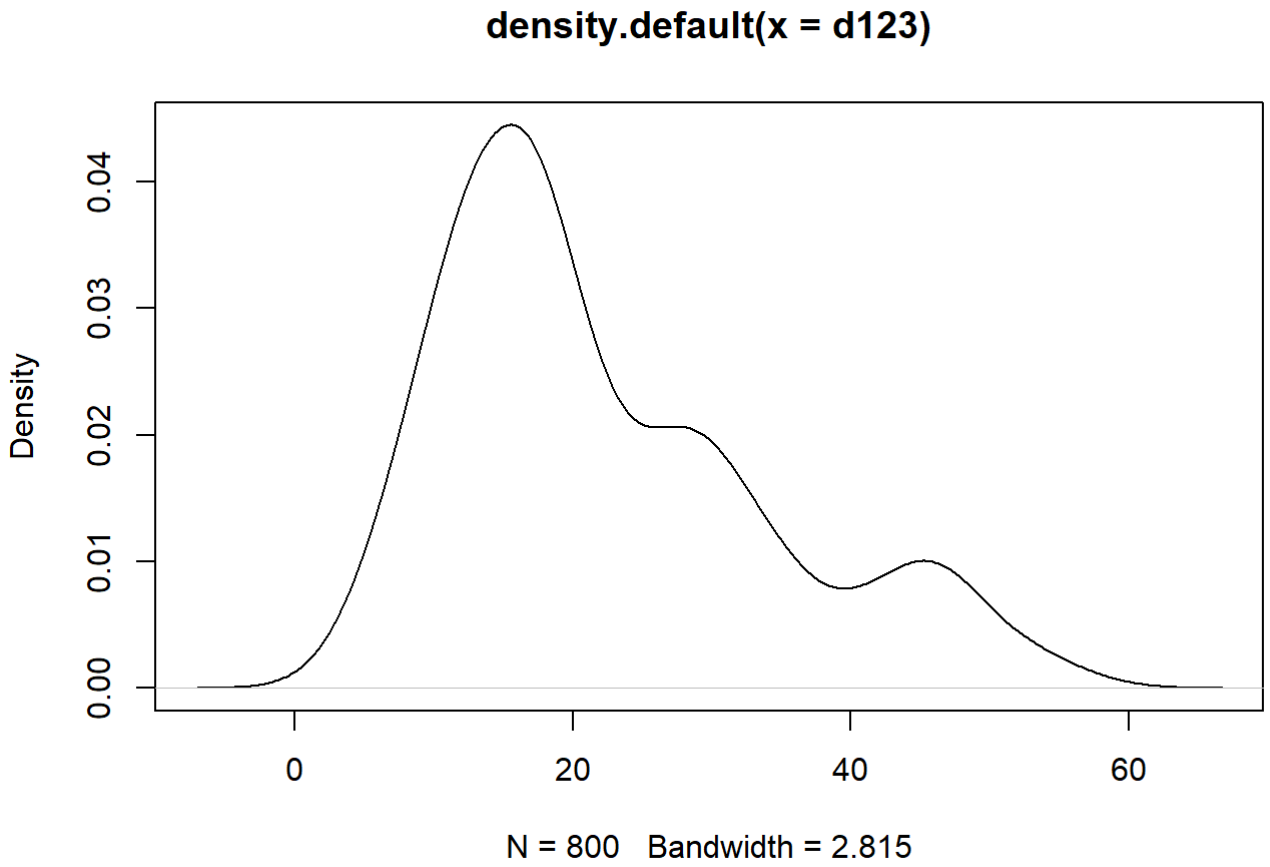
(Q4 - a) make the function norm_qq_plot

```
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)
}
```

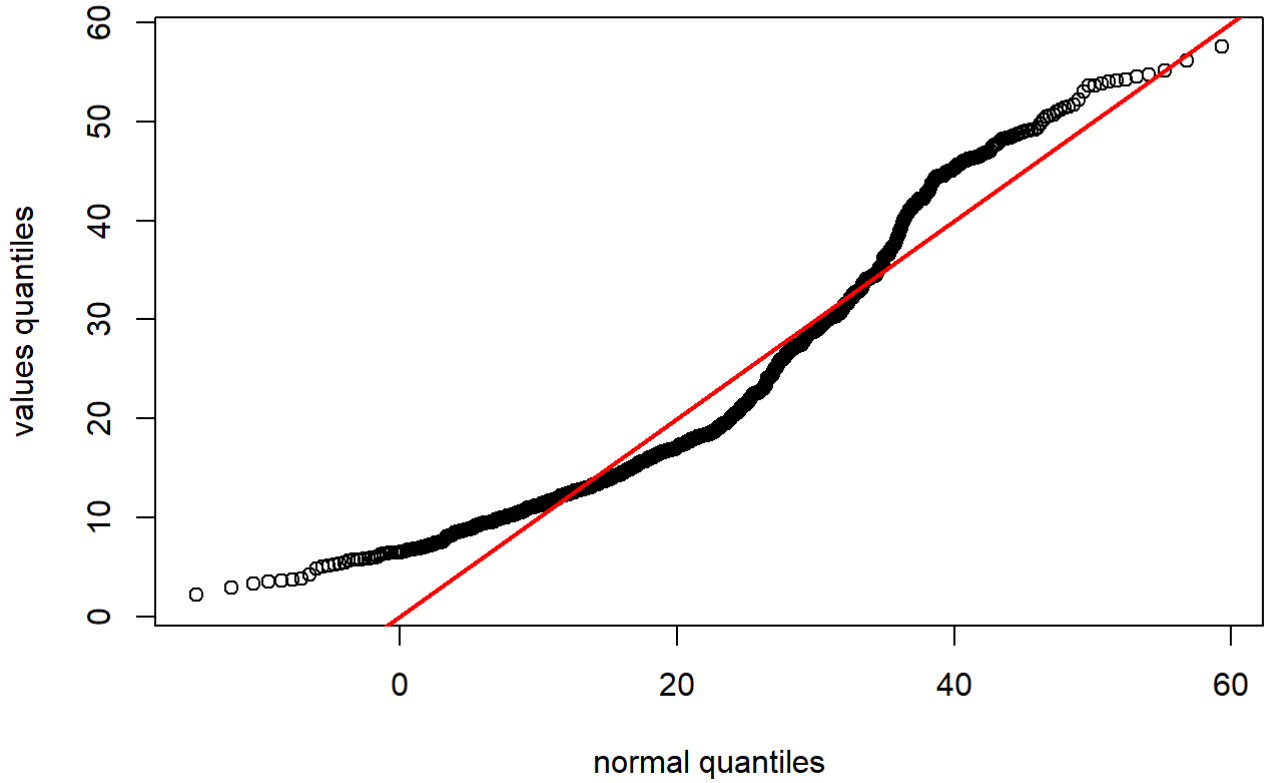
(Q4 - b) Confirm that your function norm_qq_plot via d123 provided

For better code quality, the constructing code of the d123 is hided

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




```
norm_qq_plot(d123)
```

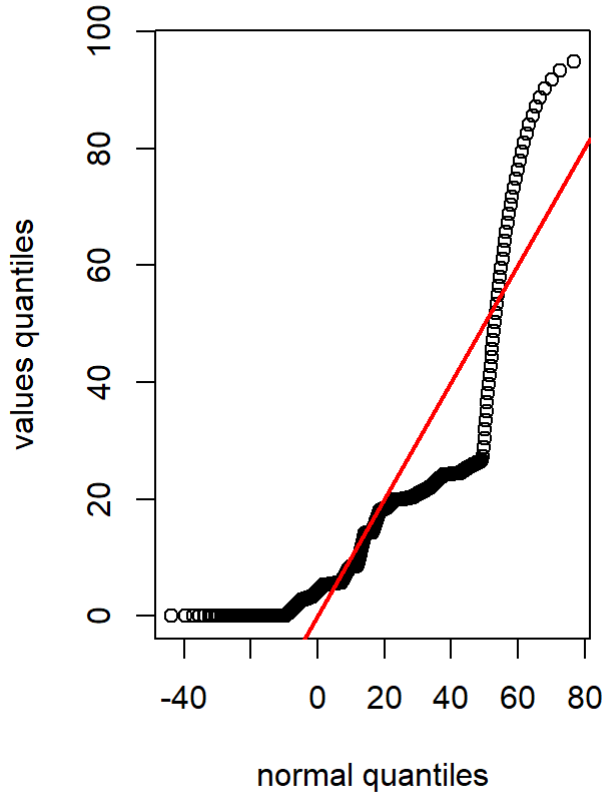


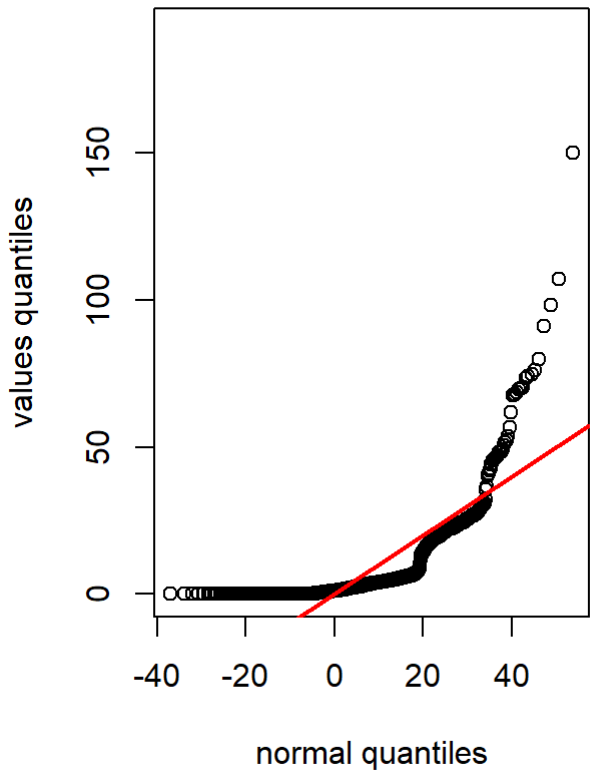
>It show that dataset d123 has “fat tails”. It means that there’re less data located in the central part of the distribution. For the terms of quantitles, the Q1 is much less than the theoretical normal distribution quantile and the last quantile is much bigger than the last theoretical quantile.So d123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

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.