HW6_110078503

110078503

3/25/2022

Thanks to 109065707, 110078506

Question 1)

The Verizon dataset this week is provided as a “wide” data frame. Let’s practice reshaping it to a “long” data frame. You may use either shape (wide or long) for your analyses in later questions.

```
verizon <- read.csv("verizon_wide.csv",header=TRUE)
#install.packages("reshape2")
library(reshape2)
```

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).

> I choose “reshape2” According to 2 articles (https://www.r-bloggers.com/2016/06/how-to-reshape-data-in-r-tidyr-vs-reshape2/ (https://www.r-bloggers.com/2016/06/how-to-reshape-data-in-r-tidyr-vs-reshape2/) & http://sanaitics.com/UploadedFiles/html_files/3862reshape2_vs_tidyr.html (http://sanaitics.com/UploadedFiles/html_files/3862reshape2_vs_tidyr.html)), tidyr’s aim is data tidying while reshape2 has the wider purpose of data reshaping and aggregating. And, functions in “tidyr” is more limited.

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

```
#reshape the versizon_wide.csv data to verizon_long
verizon_long <- melt(verizon,na.rm= TRUE)
```

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

```
names(verizon_long) <- c("company","response_time")
```

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

```
#show head of data
head(verizon_long)
```

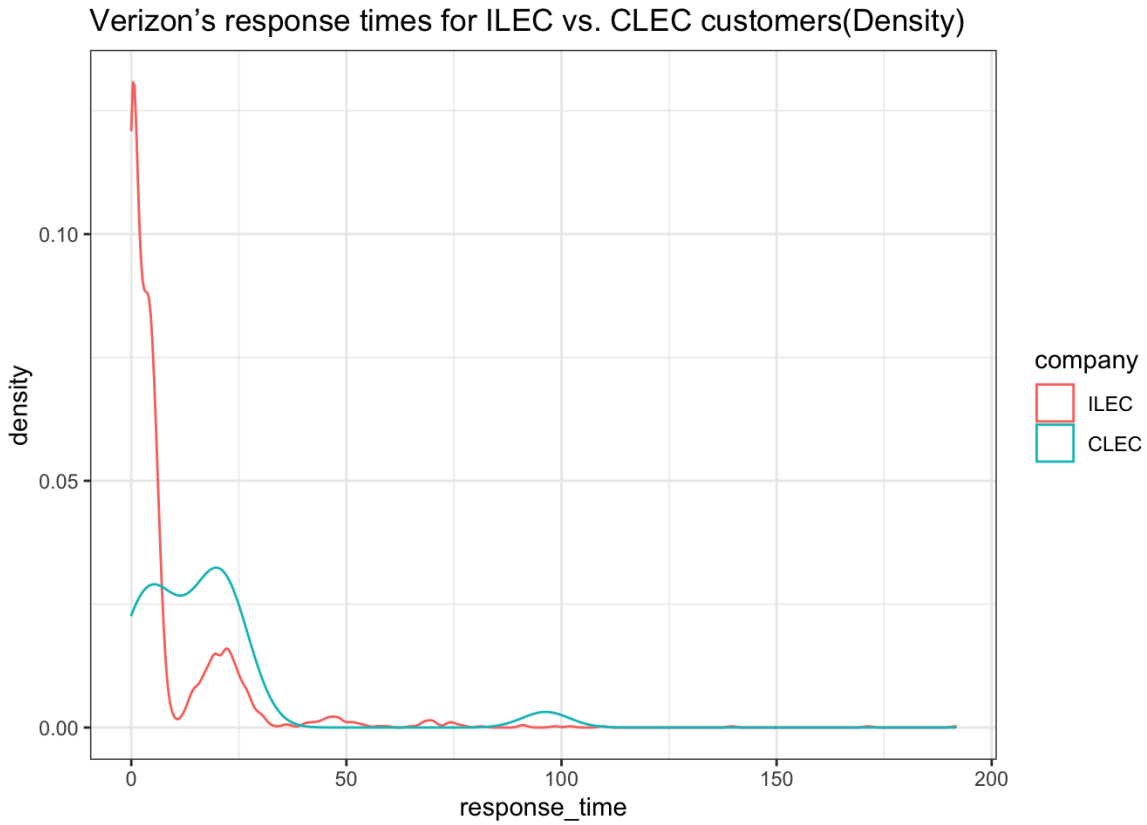
```
##   company response_time
## 1    ILEC         17.50
## 2    ILEC          2.40
## 3    ILEC          0.00
## 4    ILEC          0.65
## 5    ILEC         22.23
## 6    ILEC          1.20
```

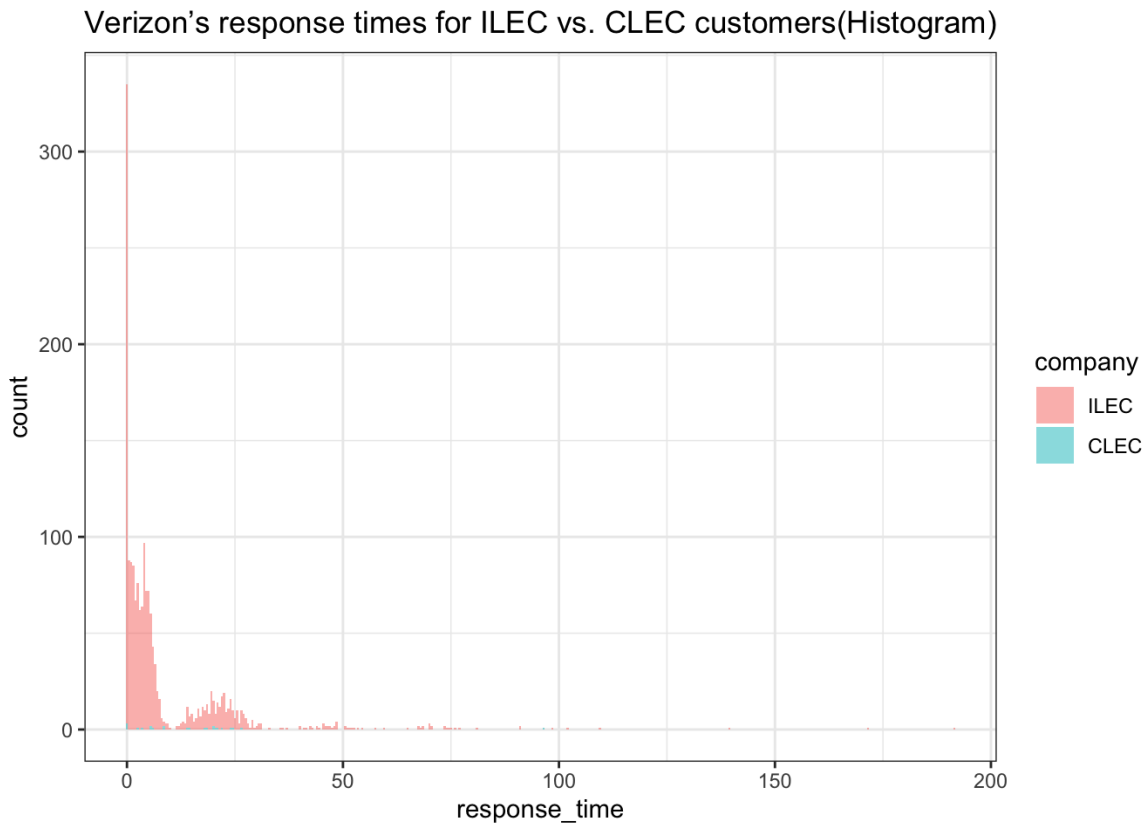
```
#show tail of data
tail(verizon_long)
```

```
##      company response_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

```
#install.packages("ggplot2")
library(ggplot2)

plot1_1 <- ggplot(verizon_long, aes(x = response_time)) +
  geom_density(aes(color = company)) +
  theme_bw() +
  labs(title = "Verizon's response times for ILEC vs. CLEC customers(Density)")
plot1_2 <- ggplot(verizon_long, aes(x=response_time, fill=company)) +
  geom_histogram(binwidth=.5, alpha=1/2) + theme_bw() + labs(title = "Verizon's response times for
 ILEC vs. CLEC customers(Histogram)")

plot1_1;plot1_2
```





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)

Null Hypothesis: The mean of response times for CLEC customers is **less than/equal** for ILEC customers (μ of RT for CLEC - μ of RT for ILEC <= 0)

Alternative Hypothesis: The mean of response times for CLEC customers is **greater** than for ILEC customers (μ of RT for CLEC - μ of RT for ILEC > 0)

b. Use the appropriate form of the t.test() function to test the difference between the mean of ILEC versus CLEC response times at 1% significance. For each of the following tests, show us the results and tell us whether you would reject the null hypothesis.

```
#install.packages("data.table")
require(data.table)
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:reshape2':
##
##     dcast, melt
```

```
verizon_long <- data.table(verizon_long)
#verizon_long[,shapiro.test(response_time), company] #data doesn't follow normal distribution, so u
se Ansari-Bradley test to test difference of two samples' variances
#ansari.test(response_time ~ company, verizon_long) #p<0.05, two samples' variances are different
```

```
#t test

#i. variances are equal
t.test(response_time~company, data = verizon_long, var.equal = TRUE, conf.level = 0.99)
```

```
##
##  Two Sample t-test
##
## data:  response_time by company
## t = -2.6125, df = 1685, p-value = 0.009068
## alternative hypothesis: true difference in means between group ILEC and group CLEC is not equal
to 0
## 99 percent confidence interval:
##  -16.0903564  -0.1046833
## sample estimates:
## mean in group ILEC mean in group CLEC
##           8.411611          16.509130
```

since p-value = 0.009068 is <0.01, **I can reject the null hypothesis**.

```
#ii. variances are not equal
t.test(response_time~company, data = verizon_long, var.equal = FALSE, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  response_time by company
## t = -1.9834, df = 22.346, p-value = 0.05975
## alternative hypothesis: true difference in means between group ILEC and group CLEC is not equal
to 0
## 99 percent confidence interval:
##  -19.588967   3.393927
## sample estimates:
## mean in group ILEC mean in group CLEC
##           8.411611          16.509130
```

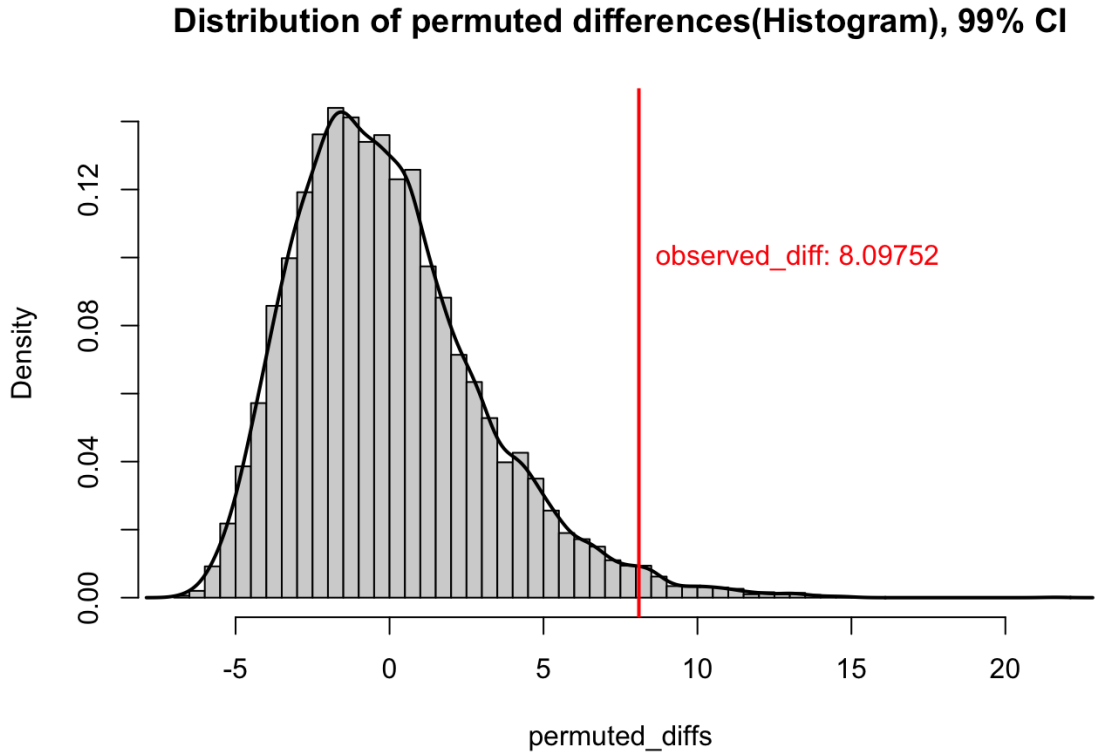
since p-value = 0.05975 is >0.01, **I cannot reject the null hypothesis**.

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

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

```
#round(runif(1) * 10^9) #random set.seed
set.seed(939581302)

#permuted difference(grouping)
n_ILEC <- nrow(verizon_long[company == "ILEC"])
n_CLEC <- nrow(verizon_long[company == "CLEC"])
groups <- c(rep("ILEC",n_ILEC),rep("CLEC",n_CLEC))

#permuted difference
permute_diff<-function(values, groups) {
  permuted <-sample(values,replace = FALSE)
  grouped <-split(permuted, groups)
  permuted_diff <- mean(grouped$CLEC,na.rm=TRUE)- mean(grouped$ILEC,na.rm=TRUE)}

nperms <- 10000
permuted_diffs <- replicate(nperms,permute_diff(verizon_long$response_time,verizon_long$company))

#observed difference
observed_diff <- mean(verizon$CLEC,na.rm=TRUE)- mean(verizon$ILEC,na.rm=TRUE)
```

```
#Visualize
#ci_99 <- quantile(permuted_diffs,probs = c(0.01,0.99))
hist(permuted_diffs, breaks = "fd", probability = TRUE, main = "Distribution of permuted difference
s(Histogram), 99% CI")
lines(density(permuted_diffs),lwd=2)
abline(v=observed_diff, col = "red", lwd = 2)
text(x=observed_diff, y= 0.1, "observed_diff: 8.09752", pos = 4, col = "red")
```




```
#abline(v=ci_99, lty= "dashed")
```

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

```
#one-tailed p-value
p_1tailed <- sum(permuted_diffs > observed_diff) /nperms
#two-tailed p-value
p_2tailed <- sum(abs(permuted_diffs) > observed_diff) /nperms

cat(" one-tailed p-value: ", p_1tailed, "\n"  , "two-tailed p-value: ", p_2tailed)
```

```
##  one-tailed p-value:  0.0185
##  two-tailed p-value:  0.0185
```

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

> No, since two-tailed p-value(0.0185) > 0.01, **I cannot reject the null hypothesis**.

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. You may use either the permutation approach (with either for-loops or the vectorized form) or the rank sum approach.

```
#Wilcoxon test(permutation approach)
companys <-split(x=verizon_long$response_time,f=verizon_long$company)
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.

```
n1 <- length(companys$CLEC)
n2 <- length(companys$ILEC)

wilcox_p_1tail<-1-pwilcox(W, n1, n2); wilcox_p_1tail
```

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

c. Run the Wilcoxon Test again using the wilcox.test() function in R – make sure you get the same W as part [a]. Show the results.

```
wilcox.test(companys$CLEC,companys$ILEC, alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  companys$CLEC and companys$ILEC
## W = 26820, p-value = 0.0004565
## alternative hypothesis: true location shift is greater than 0
```

```
#W = 26820 is the same as part [a]
```

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?

> p-value(0.0004565) < 0.01, **I can reject the null hypothesis**, which represents that the values of CLEC and ILEC are different.

Question 4)

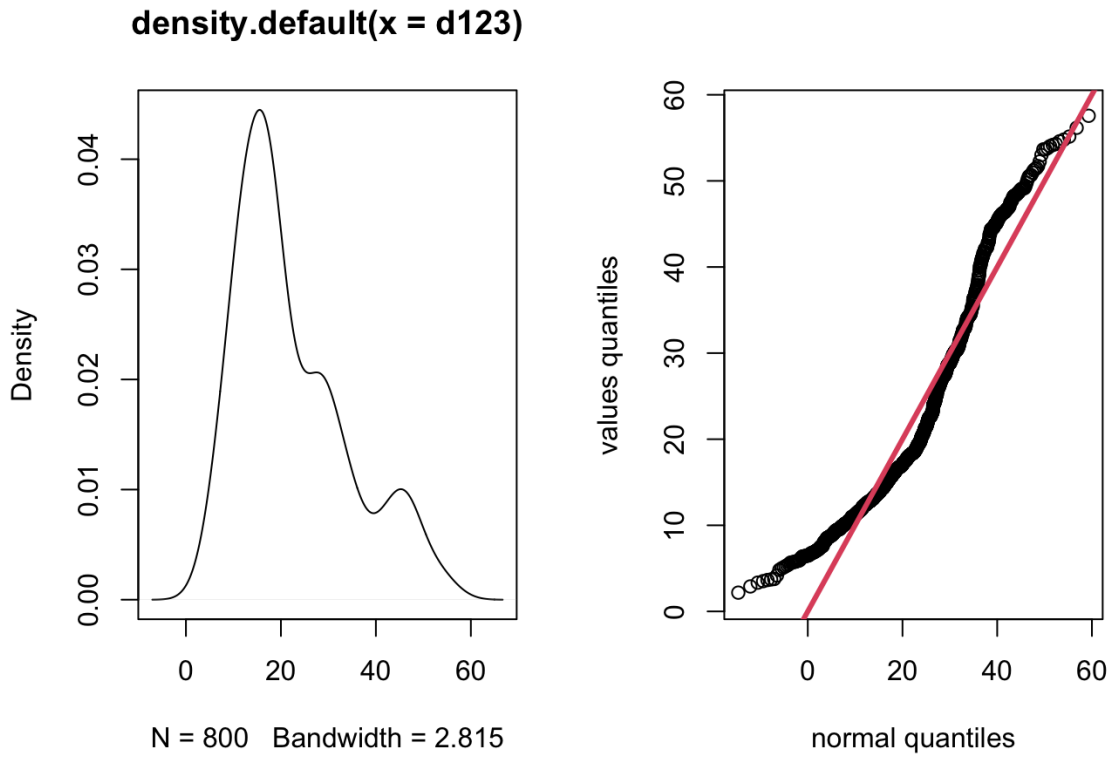
One of the assumptions of some classical statistical tests is that our population data should be roughly normal. Let's explore one way of visualizing whether a sample of data is normally distributed.

a. Follow the following steps to create a function to see how a distribution of values compares to a perfectly normal distribution. The ellipses (…) in the steps below indicate where you should write your own code.

```
#normal quantile-quantile plot #normal Q-Q plot
norm_qq_plot <- function(values) {
   probs1000 <- seq(0, 1, 0.001)
   q_vals <- quantile(values, probs = probs1000)
   q_norm <- qnorm(probs1000 ,mean = mean(values), sd = sd(values))
   plot(q_norm, q_vals, xlab="normal quantiles", ylab="values quantiles")
   abline(a=0,b=1,col=2,lwd=3)
}
```

b. Confirm that your function works by running it against the values of our d123 distribution from week 3 and checking that it looks like the plot on the right:

```
#test, use d123
set.seed(978234)
d1 <-rnorm(n=500,mean=15,sd=5)
d2 <-rnorm(n=200,mean=30,sd=5)
d3 <-rnorm(n=100,mean=45,sd=5)
d123 <-c(d1, d2, d3)

par(mfrow=c(1,2))
plot(density(d123))
norm_qq_plot(d123)
```

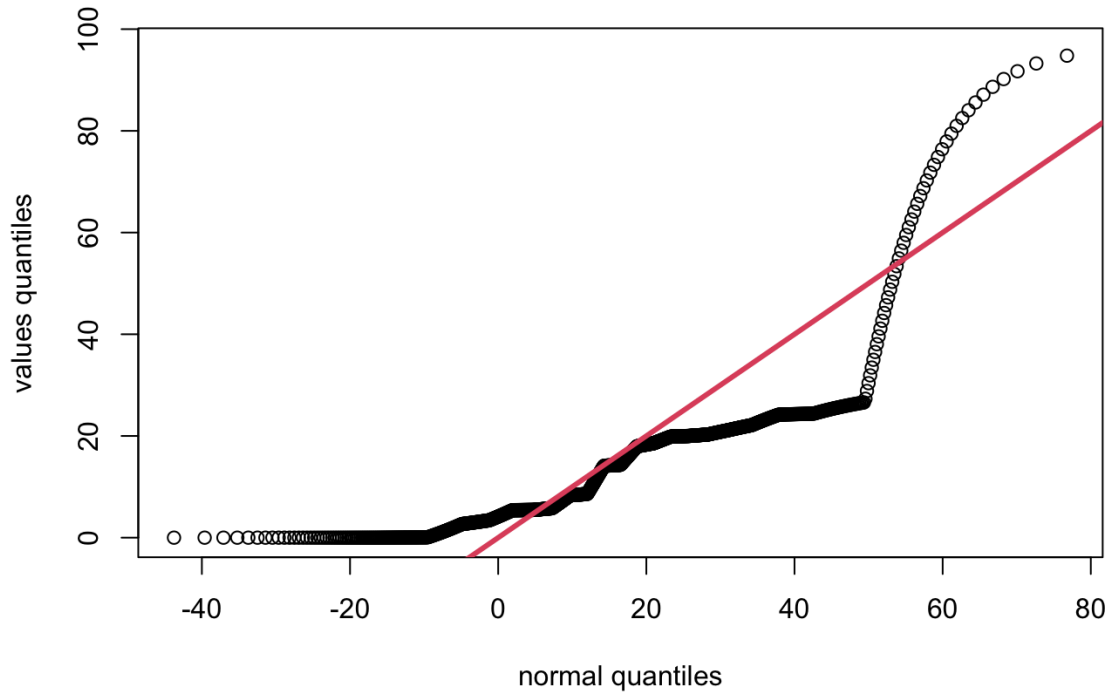


Interpret the plot you produced (see this article on how to interpret normal Q-Q plots) and tell us if it suggests whether d123 is normally distributed or not.

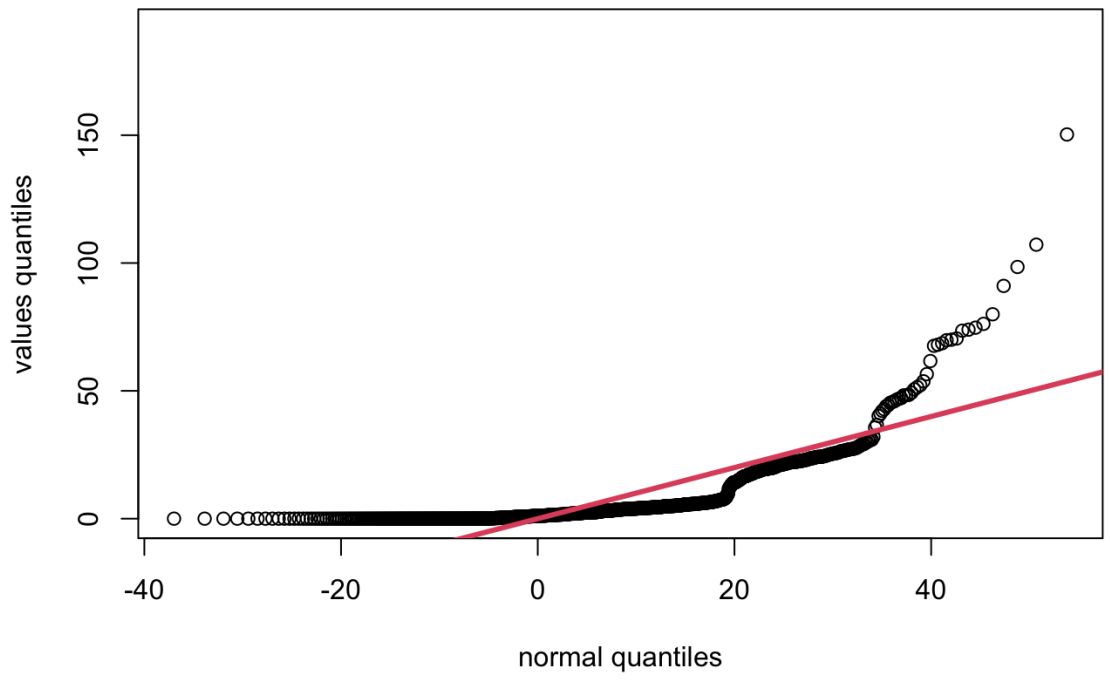
> It's unlikely normally distributed because the points stray from linearity, which the data is deviating from the theoretical distribution(normal quantiles).

c. Use your normal Q-Q plot function to check if the values from each of the CLEC and ILEC samples we compared in question 2 could be normally distributed. What's your conclusion?

```
norm_qq_plot(companys$CLEC)
```




```
norm_qq_plot(companys$ILEC)
```



Both two samples are unlikely normally distributed. Therefore, we cannot use methods that are not for normal distributed data(i.e., t test).