HW06

107070008

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.

```
data <- read.csv("./verizon_wide.csv")
```

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 tidyr, because tidy output only var and value. However, reshape2 also output id
element.  In this question, we don't need id, so the formation od tidyr is more tidy to
me link: https://jtr13.github.io/spring19/hx2259_qz2351.html
```

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

```
library(tidyr)
long <- gather(data, na.rm = TRUE, key = "host", value = "load_time")
```

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

```
head(long)
```

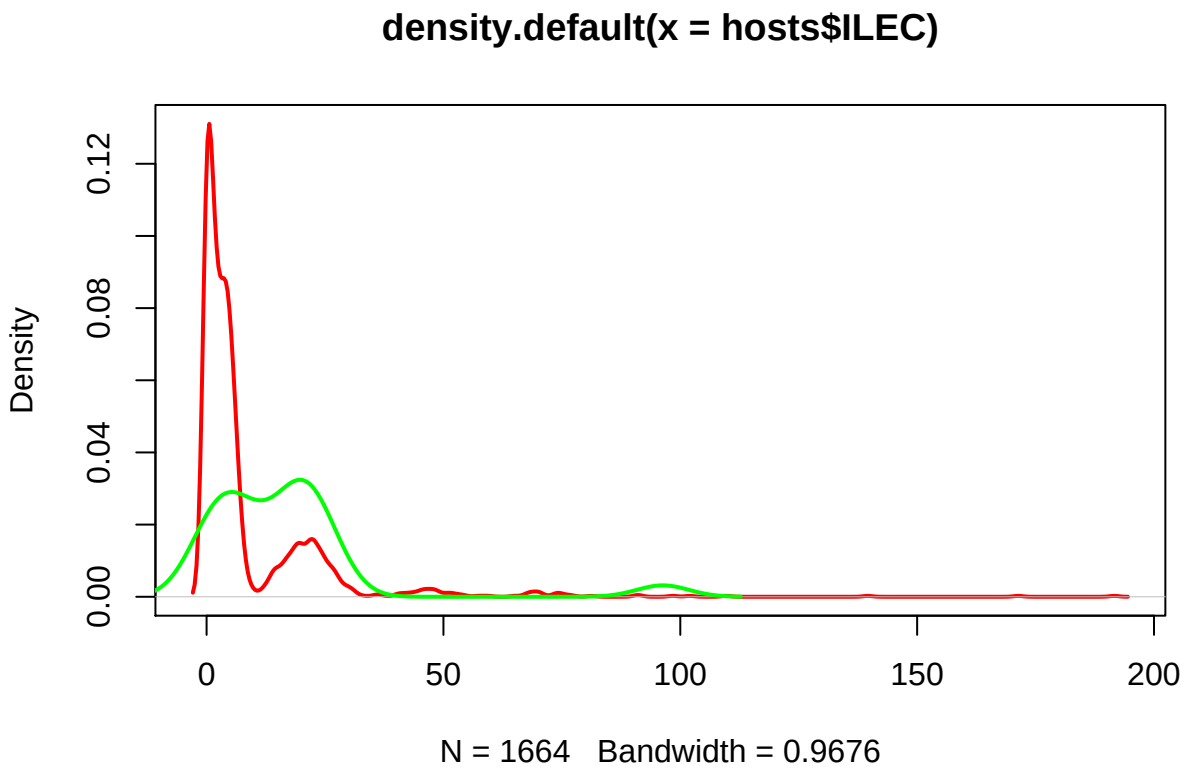
```
##   host load_time
## 1 ILEC     17.50
## 2 ILEC      2.40
## 3 ILEC      0.00
## 4 ILEC      0.65
## 5 ILEC     22.23
## 6 ILEC      1.20
```

```
tail(long)
```

```
##      host load_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

```
hosts <- split(x = long$load_time, f = long$host)
plot(density(hosts$ILEC), col = "red", lwd = 2)
lines(density(hosts$CLEC), col = "green", lwd = 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) `H0:ILEC = CLEC H1:ILEC > CLEC`

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.

c. Conduct the test assuming variances of the two populations are equal

```
t.test(hosts$ILEC, hosts$CLEC, conf.level = 0.99, alt = "two.sided", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  hosts$ILEC and hosts$CLEC
## t = -2.6125, df = 1685, p-value = 0.009068
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -16.0903564  -0.1046833
## sample estimates:
```

```
## mean of x mean of y
##  8.411611 16.509130
```

Since 0.09068(p-value) < 0.01(significant level), reject H0.

ii. Conduct the test assuming variances of the two populations are not equal

```
t.test(hosts$ILEC, hosts$CLEC, conf.level = 0.99, alt = "two.sided", var.equal = FALSE)
```

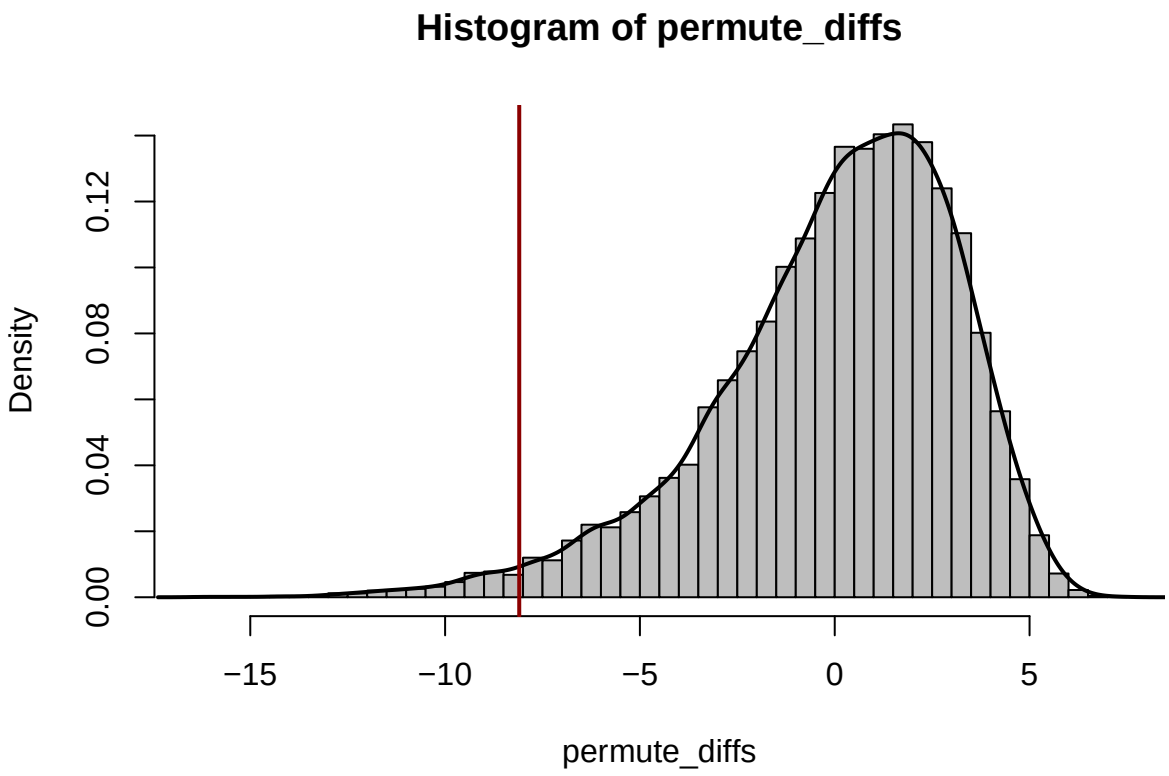
```
##
##  Welch Two Sample t-test
##
## data:  hosts$ILEC and hosts$CLEC
## t = -1.9834, df = 22.346, p-value = 0.05975
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -19.588967   3.393927
## sample estimates:
## mean of x mean of y
##  8.411611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

Since 0.05975(p-value) > 0.01(significant level), do not reject H0.

c. Use a permutation test to compare the means of ILEC vs. CLEC response times
d. Visualize the distribution of permuted differences, and indicate the observed difference as well.

ii. What are the one-tailed and two-tailed p-values of the permutation test?
iii. Would you reject the null hypothesis at 1% significance in a one-tailed test?

```
#i
ob_diff <- mean(hosts$ILEC) - mean(hosts$CLEC)
permute_diff <- function(values, groups){
  permuted <- sample(values, replace = FALSE)
  grouped <- split(permuted, groups)
  permute_diff <- mean(grouped$ILEC) - mean(grouped$CLEC)
}
nperms <- 10000
permute_diffs <- replicate(nperms, permute_diff(long$load_time, long$host))
hist(permute_diffs, col="grey", breaks = "fd", probability = TRUE)
lines(density(permute_diffs), lwd = 2)
abline(v=ob_diff, col="darkred", lwd = 2)
```




```
#ii
p_1tailed <- sum(permute_diffs > ob_diff) / nperms
p_1tailed
```

```
## [1] 0.9805
```

```
p_2tailed <- sum(abs(permute_diffs) > ob_diff) / nperms
p_2tailed
```

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

```
#iii
t.test(hosts$ILEC, hosts$CLEC, conf.level = 0.99, alt = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  hosts$ILEC and hosts$CLEC
## t = -2.6125, df = 1685, p-value = 0.9955
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  -15.3149      Inf
## sample estimates:
## mean of x mean of y
##  8.411611 16.509130
```

```
t.test(hosts$ILEC, hosts$CLEC, conf.level = 0.99, alt = "greater", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  hosts$ILEC and hosts$CLEC
## t = -1.9834, df = 22.346, p-value = 0.9701
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  -18.3259      Inf
## sample estimates:
## mean of x mean of y
##  8.411611 16.509130
```

Both p-values 0.9701 and 0.9955 are bigger than significance level, so we should not reject H0.

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.

```
time_ranks <- rank(long$load_time)
ranked_groups <- split(time_ranks, long$host)
U1 <- sum(ranked_groups$ILEC)
n1 <- length(hosts$ILEC)
n2 <- length(hosts$CLEC)
W <- U1 - (n1*(n1+1))/2
W
```

```
## [1] 11452
```

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

```
wilcox_p1tail <- 1 - pwilcox(W, n1, n2)
wilcox_p1tail
```

```
## [1] 0.9996305
```

```
wilcox_p2tail <- 2 * wilcox_p1tail
wilcox_p2tail
```

```
## [1] 1.999261
```

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(hosts$ILEC, hosts$CLEC, alt = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hosts$ILEC and hosts$CLEC
## W = 11452, p-value = 0.9995
## alternative hypothesis: true location shift is greater than 0
```

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? `Not reject, because p-value is greater than 0.01.`

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.

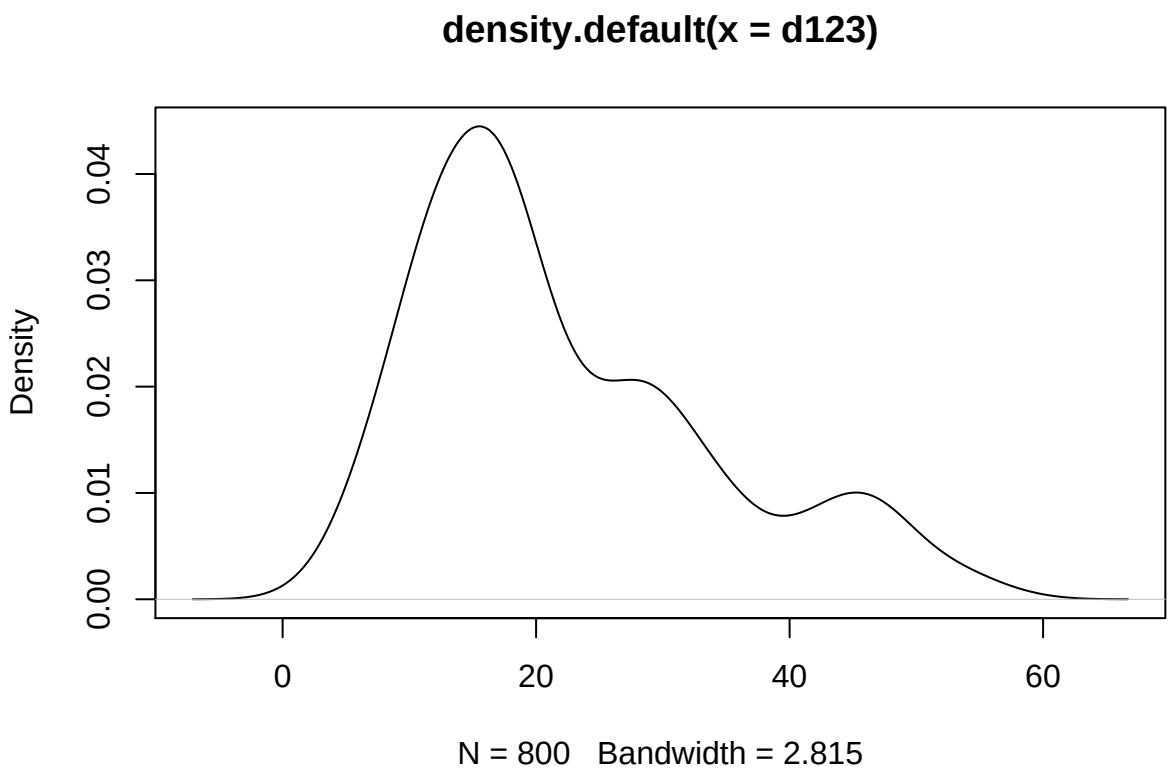
a. Follow the following steps to create a function to see how a distribution of values compares to a perfectly normal distribution. The ellipses (. . .) in the steps below indicate where you should write your own code.

```
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="red", lwd=2)
}
```

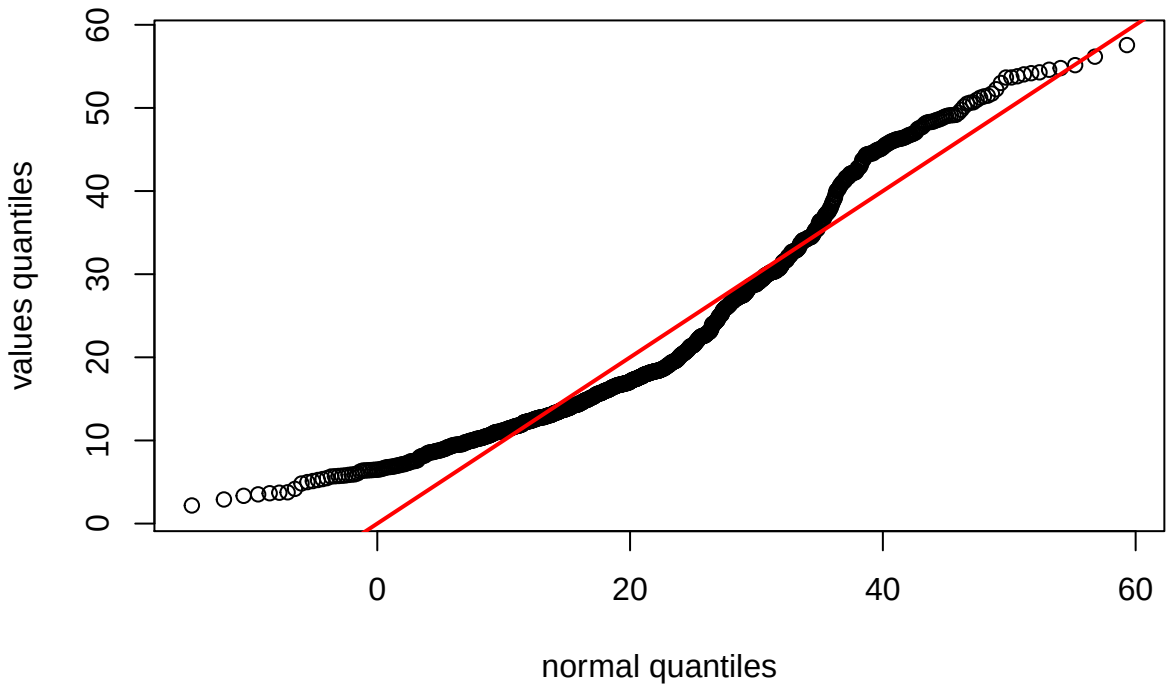
b. Confirm that your function works by running it against the values of our d123 distribution from week 3 and checking that it looks like the plot on the right:

```
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)

plot(density(d123))
```

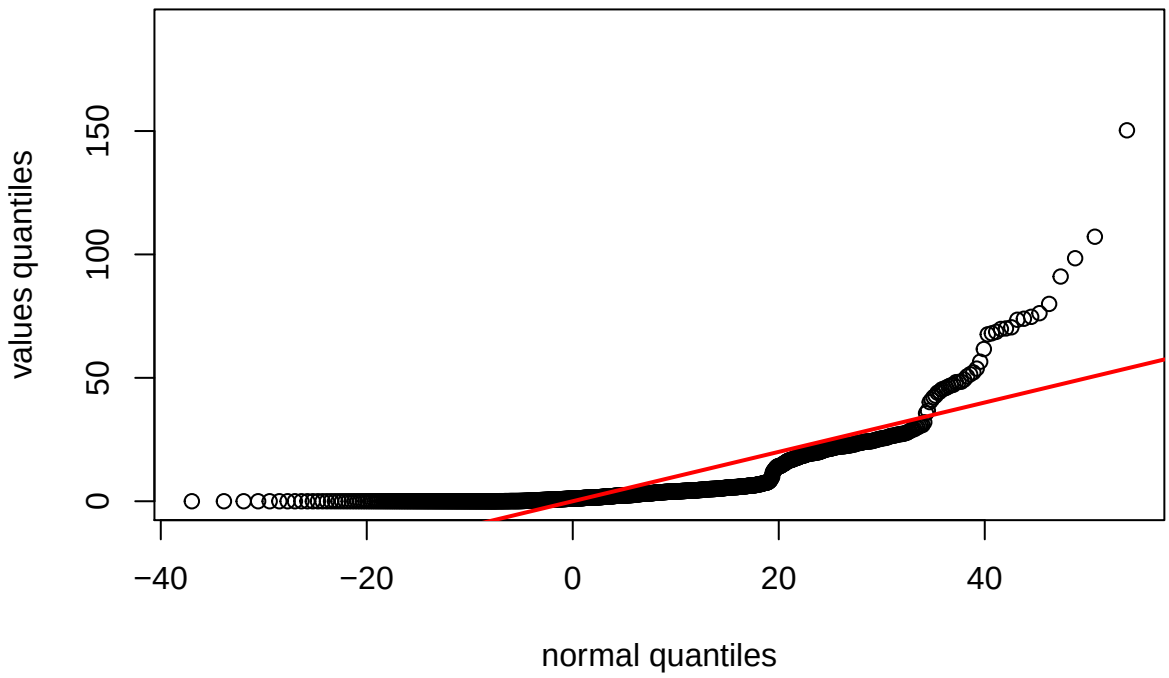



```
norm_qq_plot(d123)
```

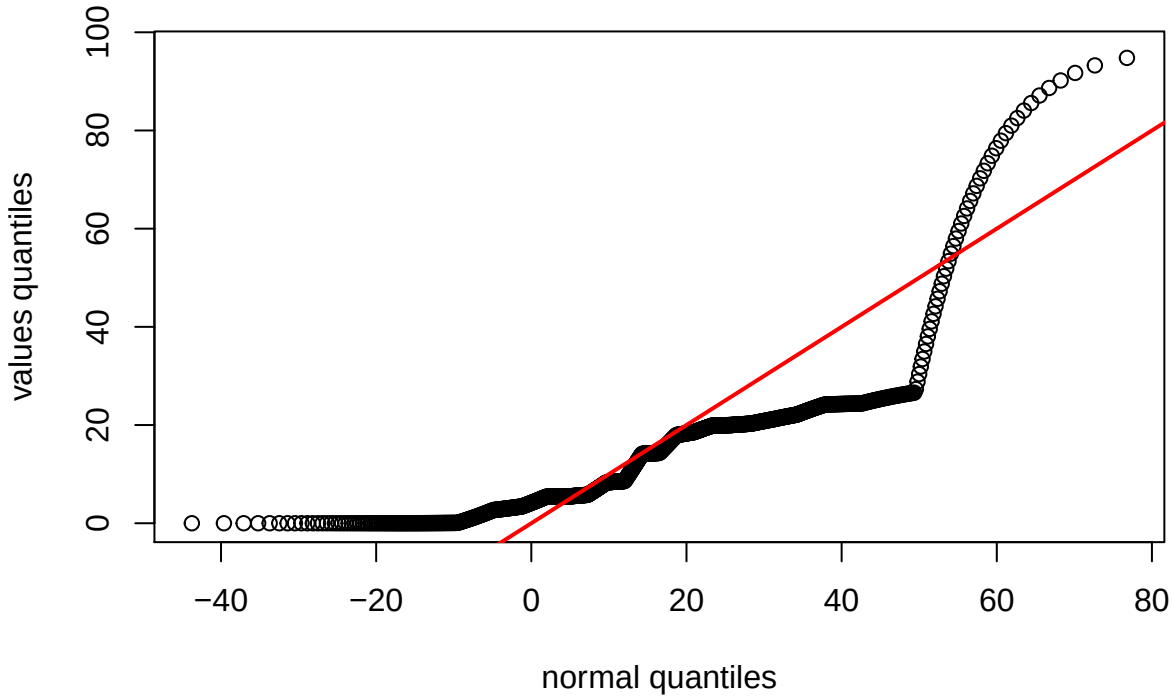


c. Use your normal Q-Q plot function to check if the values from each of the CLEC and ILEC samples we compared in question 2 could be normally distributed. What's your conclusion?

```
norm_qq_plot(hosts$ILEC)
```




```
norm_qq_plot(hosts$CLEC)
```



In ILEC's plot, the upper tail is a little bit far away from red line, so I think it is a "Skwed Right" histogram.On the other hand, CLEC's plot is normally distributed, because of its symmetry.