Student ID: 112077423

```
library(ggplot2)
library(compstatslib)
library(tidyr)
```

```
data <- read.csv('verizon_wide.csv', header=TRUE)
head(data)
```

```
##    ILEC  CLEC
## 1 17.50 26.62
## 2  2.40  8.60
## 3  0.00  0.00
## 4  0.65 21.15
## 5 22.23  8.33
## 6  1.20 20.28
```

Question 1

(a)

I'll use tidyr since reshape2 was deprecated for tidyr. A deprecated package is one where the maintainer has encouraged people to use other packages instead.

(b)

```
df_long <- gather(data, na.rm = TRUE, key = 'customer', value = 'time')
```

(c)

```
head(df_long)
```

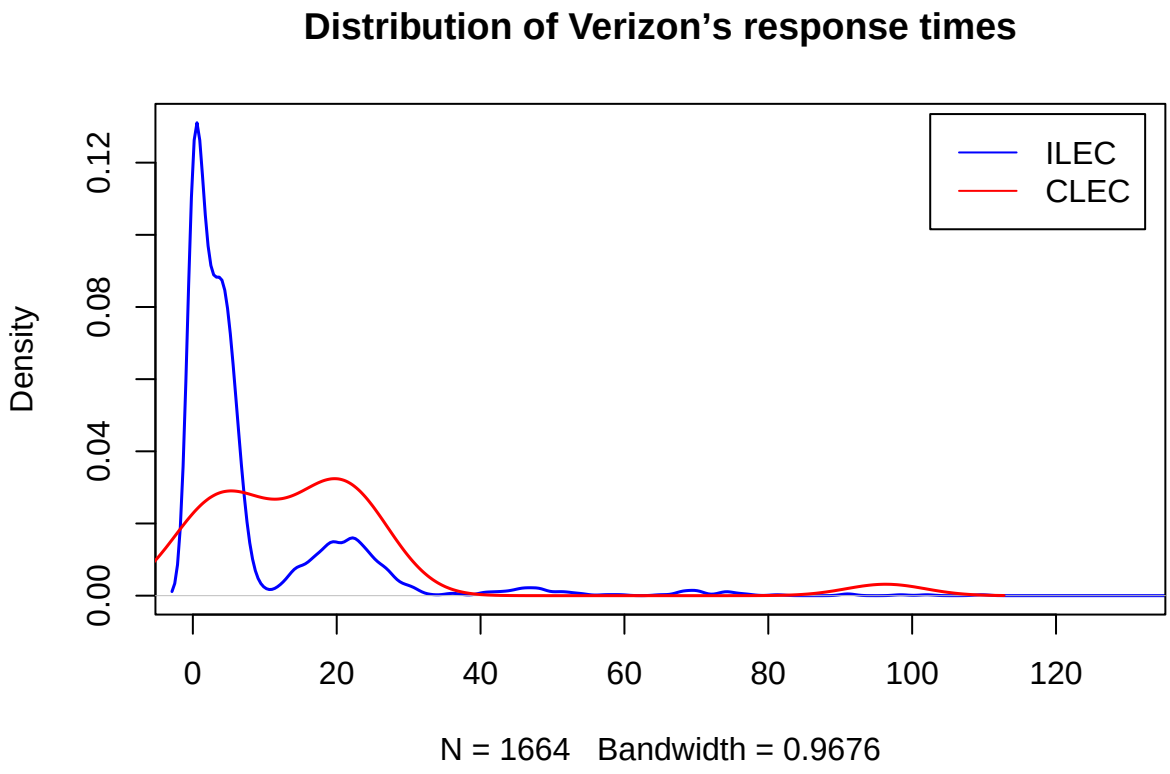
```
##   customer  time
## 1     ILEC 17.50
## 2     ILEC  2.40
## 3     ILEC  0.00
## 4     ILEC  0.65
## 5     ILEC 22.23
## 6     ILEC  1.20
```

```
tail(df_long)
```

```
##      customer  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)

```r
customers <- split(x=df_long$time, f=df_long$customer)
plot(density(customers$ILEC), col='blue', lwd=1.5, xlim=c(0, 130), main='Distribution of Verizon's respo
lines(density(customers$CLEC), col='red', lwd=1.5)
legend('topright', lty=1, legend=c("ILEC", "CLEC"), col=c("blue", "red"), inset=.02)
```



Question 2

(a)

H0: u(CLEC) - u(ILEC) <= 0

Ha: u(CLEC) - u(ILEC) > 0

(b)

```r
alpha <- 0.01
result_same <- t.test(customers$CLEC, customers$ILEC, alt="greater", var.equal=TRUE)
result_diff <- t.test(customers$CLEC, customers$ILEC, alt="greater", var.equal=FALSE)

print(result_same)
```

```
##
##  Two Sample t-test
##
## data:  customers$CLEC and customers$ILEC
```

```
## t = 2.6125, df = 1685, p-value = 0.004534
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.996491      Inf
## sample estimates:
## mean of x mean of y
## 16.509130  8.411611
```

```
print(result_diff)
```

```
##
##  Welch Two Sample t-test
##
## data:  customers$CLEC and customers$ILEC
## t = 1.9834, df = 22.346, p-value = 0.02987
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.091721      Inf
## sample estimates:
## mean of x mean of y
## 16.509130  8.411611
```

```
if (result_same$p.value < alpha) {
  print('Considering that population standard deviations are equal: reject null hypothesis')
} else {
  print('Considering that population standard deviations are equal: fail to reject null hypothesis')
}
```

```
## [1] "Considering that population standard deviations are equal: reject null hypothesis"
```

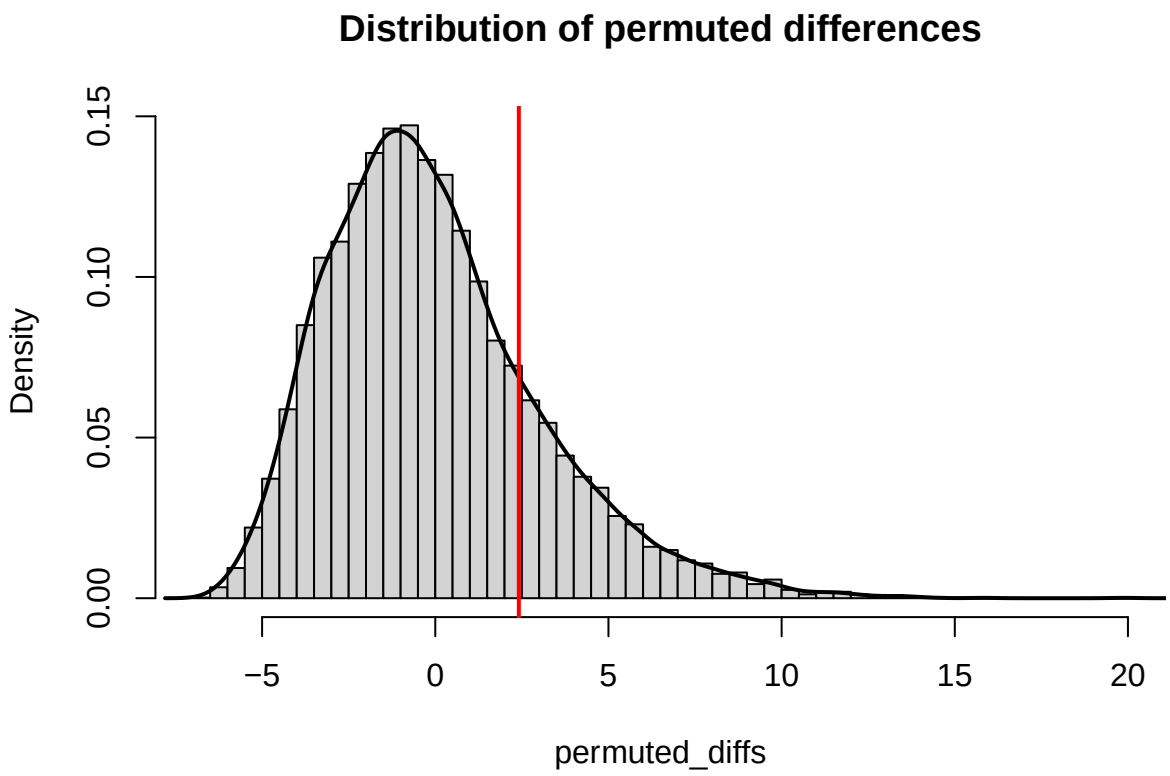
```
if (result_diff$p.value < alpha) {
  print('Considering that population standard deviations are not equal: reject null hypothesis')
} else {
  print('Considering that population standard deviations are not equal: fail to reject null hypothesis')
}
```

```
## [1] "Considering that population standard deviations are not equal: fail to reject null hypothesis"
```

(c)

```
observed_diff <- mean(customers$CLEC) - mean(customers$ILEC)

permute_diff <- function(values, groups) {
  permuted <- sample(values, replace = FALSE)
  grouped <- split(permuted, groups)
  permuted_diff <- mean(grouped$CLEC) - mean(grouped$ILEC)
  return(permuted_diff)
}

nperms <- 10000
permuted_diffs <- replicate(nperms, permute_diff(df_long$time, df_long$customer))
```

```r
hist(permuted_diffs, breaks = "fd", probability = TRUE, main='Distribution of permuted differences')
lines(density(permuted_diffs), lwd=2)
abline(v=mean(abs(permuted_diffs)), col='red', lw=2)
```




```r
p_1tailed <- sum(permuted_diffs > observed_diff) / nperms
p_2tailed <- sum(abs(permuted_diffs) > observed_diff) / nperms

out1 <- paste('one-tailed p-value:', p_1tailed)
out2 <- paste('two-tailed p-value:', p_2tailed)
cat(out1, out2, sep='\n')
```

```
## one-tailed p-value: 0.0175
## two-tailed p-value: 0.0175
```

```r
alpha = 0.01

if (p_1tailed < alpha) {
  print('According to one-tailed test: reject null hypothesis')
} else {
  print('According to one-tailed test: fail to reject null hypothesis')
}
```

```
## [1] "According to one-tailed test: fail to reject null hypothesis"
```

As we can see from the graph above, the mean difference is around 2.5. Also, we can notice some negative values indicating that in some cases CLEC mean time was smaller than ILEC mean time.

Question 3

(a-b)

```
gt_eq <- function(a, b) {
  ifelse(a > b, 1, 0) + ifelse(a == b, 0.5, 0)
}

W <- sum(outer(customers$CLEC, customers$ILEC, FUN = gt_eq))

n1 <- length(customers$CLEC)
n2 <- length(customers$ILEC)

wilcox_p_1tail <- 1 - pwilcox(W, n1, n2)
out1 <- paste('W statistic:', W)
out2 <- paste('one-tailed p-value for W:', wilcox_p_1tail)
cat(out1, out2, sep='\n')
```

```
## W statistic: 26820
## one-tailed p-value for W: 0.000368834147256192
```

(c)

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

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

(d)

Since the p-value is much smaller than the significance level (0.01) therefore we should reject the null hypothesis.

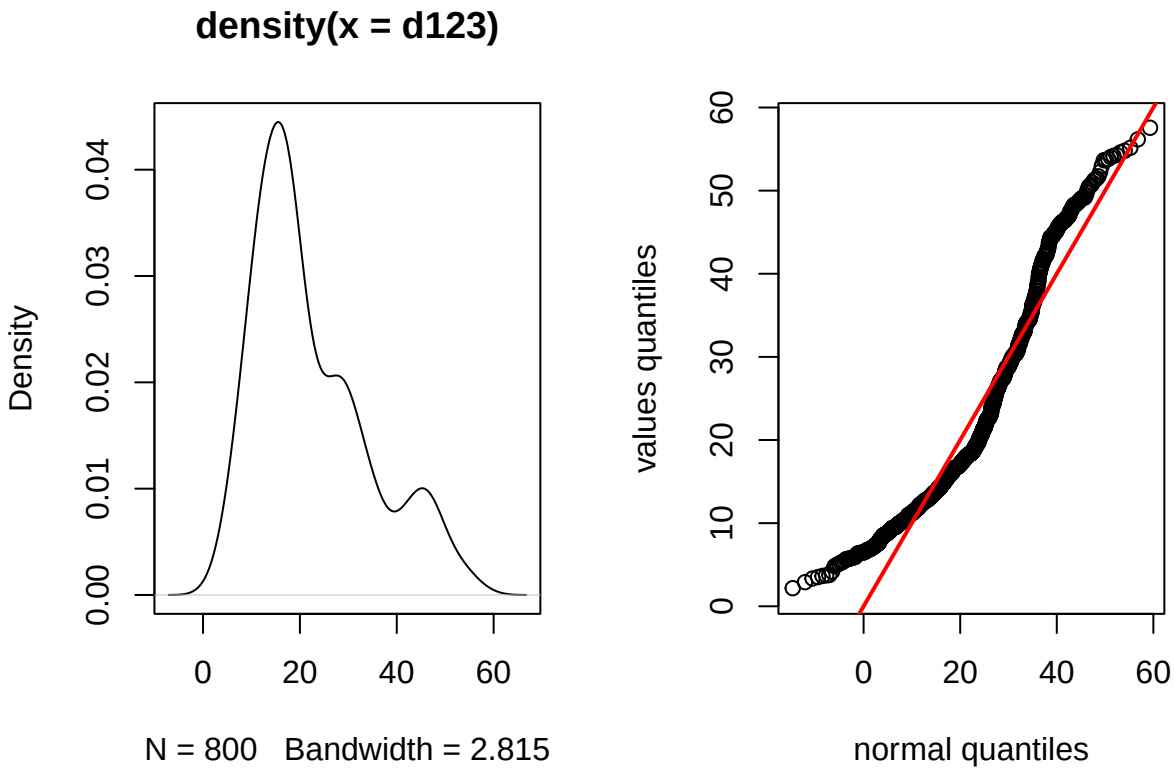
Question 4

(a)

```
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)

```
par(mfrow=c(1,2))

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)
```

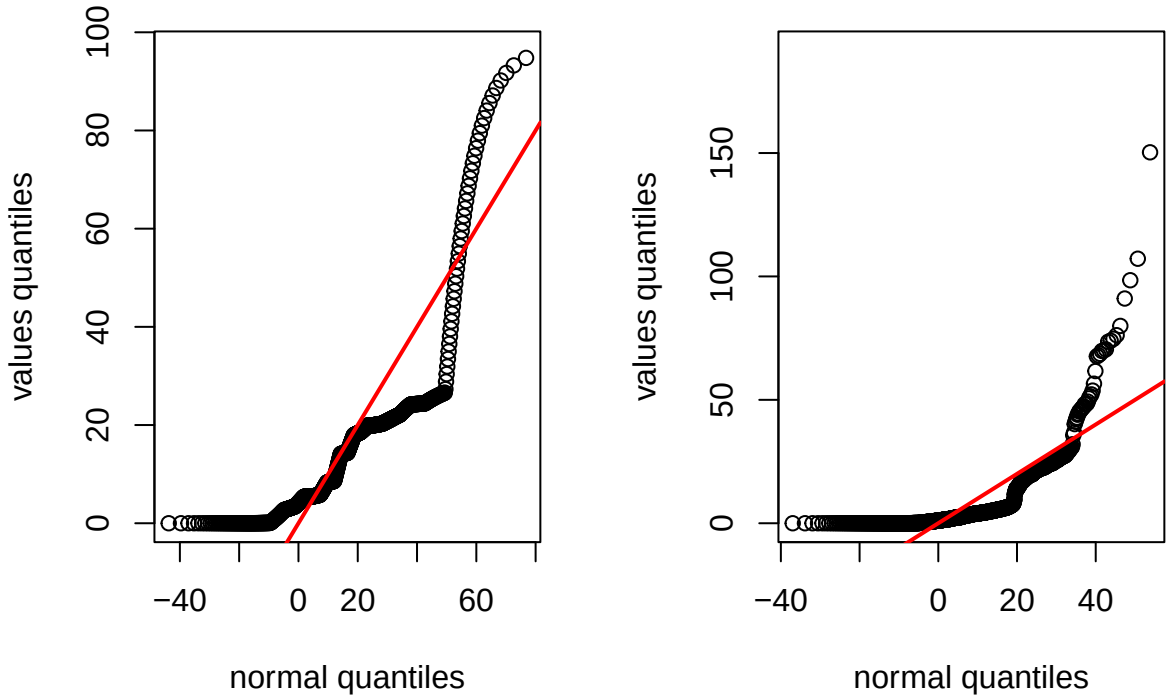


The red line shows where the points would fall if the dataset were normally distributed. Looking at the Q-Q plot for the second graph we can see that the theoretical quantile should be around 40, when in fact it is around 45. The point's trend upward shows that the actual quantiles are much greater than the theoretical quantiles.

If the data is normally distributed, the points will fall on the 45-degree reference line. If the data is not normally distributed, the points will deviate from the reference line. As we can see from the graph, points deviate a bit therefore we can't say that data is not really normally distributed.

(c)

```
par(mfrow=c(1,2))
norm_qq_plot(customers$CLEC)
norm_qq_plot(customers$ILEC)
```



As we can see, both CLEC and ILEC samples are not normally distributed.