HW (Week 6)

109006206

QUESTION 1

PART A

Refference Link : https://stackoverflow.com/questions/57410033/tidyr-vs-dplyr-reshape2

reshape2 focuses on reshaping data, it provides more flexibility for transforming wide data to long format especially when dealing with multiple variables that need to be reshaped simultaneously.

PART B

```
setwd("/Users/olivia/Documents/Documents/Study/Semester 6/BACS/HW5")
verizon<- read.csv("verizon_wide.csv")
library(reshape2)
loads_long<-melt(verizon, na.rm= TRUE,
                 variable.name= "PhoneService",
                 value.name= "ResponseTime")
```

PART C

```
head(loads_long)
```

```
##   PhoneService ResponseTime
## 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)
```

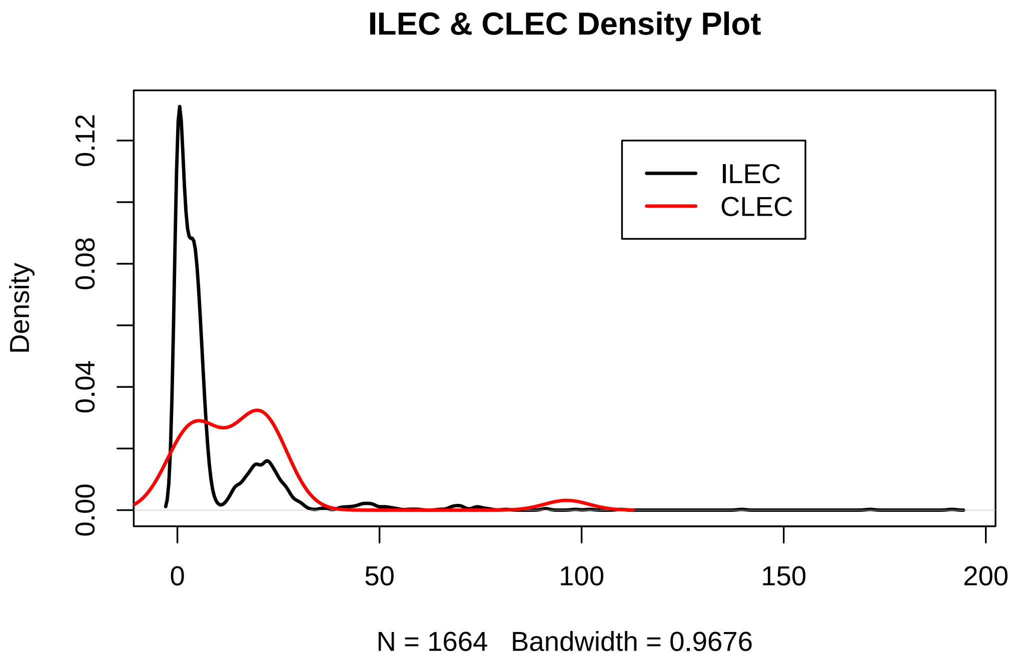
```
##      PhoneService ResponseTime
## 1682         CLEC        24.20
## 1683         CLEC        22.13
## 1684         CLEC        18.57
## 1685          CLEC        20.00
## 1686          CLEC        14.13
## 1687          CLEC         5.80
```

PART D

```
ILEC<-subset(loads_long, loads_long$PhoneService == "ILEC")
CLEC<-subset(loads_long, loads_long$PhoneService == "CLEC")

plot(density(ILEC$ResponseTime),lwd = 2, main = "ILEC & CLEC Density Plot")
lines(density(CLEC$ResponseTime),lwd = 2, col = "red")

legend(110,
       0.12,
       c("ILEC","CLEC"),
       lwd = c(2,2),
       lty = c("solid","solid"),
       col = c("black","red"))
```



QUESTION 2

PART A

Null: The Mean of Response Time for CLEC equal or less than ILEC.

Alternative: The Mean of Response Time for CLEC is greater than ILEC.

PART B

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

```
t.test(verizon$CLEC,verizon$ILEC, alt ="greater",var.equal=TRUE)
```

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

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

```
t.test(verizon$CLEC,verizon$ILEC, alt = "greater",var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  verizon$CLEC and verizon$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
```

PART C

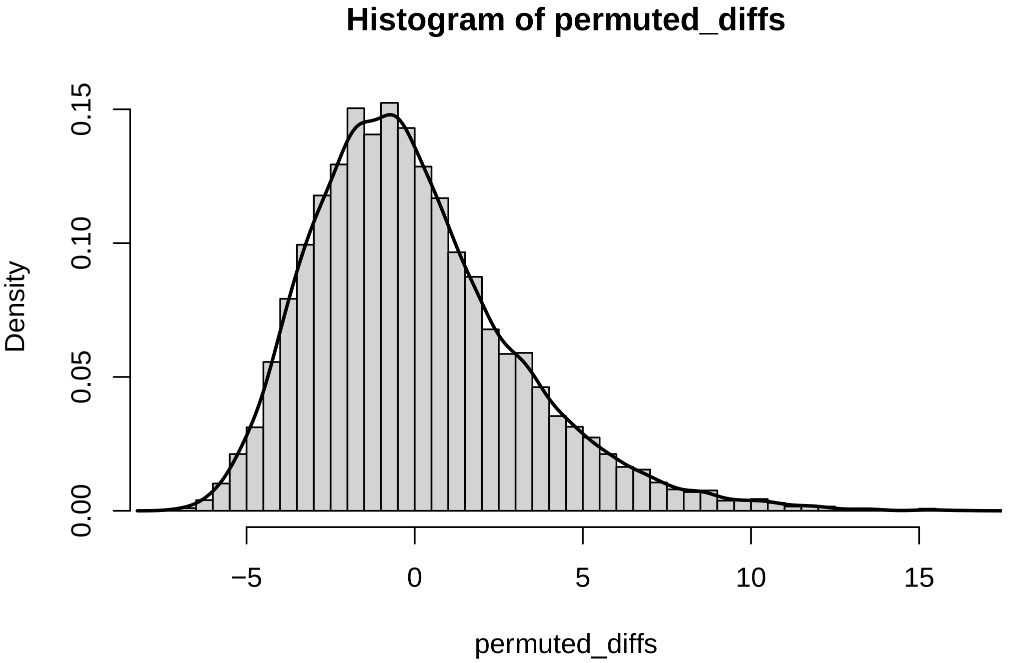
i) Visualize the distribution of permuted differences, and indicate the observed difference as well.

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

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

nperms <- 10000

permuted_diffs<-replicate(nperms, permute_diff(loads_long$ResponseTime, loads_long$PhoneService))
hist(permuted_diffs, breaks = "fd", probability = TRUE)
lines(density(permuted_diffs), lwd=2)
```



ii) What are the one-tailed and two-tailed p-values of the permutation

one-tailed p-values: 0.0182

two-tailed p-values: 0.0182

iii) Would you reject the null hypothesis at 1% significance in a one-tailed test? No , I would not reject the null hypothesis at 1% significance in one-tailed test since the p-value shows that p>0.01

QUESTION 3

PART A

a) Compute the W statistic comparing the values. You may use either the permutation approach (try the functional form) or the rank sum approach.

```r
time_ranks <-rank(loads_long$ResponseTime)
ranked_groups<-split(time_ranks, loads_long$PhoneService)
U1 <-sum(ranked_groups$ILEC)
n1 <-length(verizon$ILEC)
W<-U1 -(n1 * (n1 + 1))/2
```

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

```r
n2 <-length(na.omit(verizon$CLEC))
wilcox_p_1tail <- 1 -pwilcox(W, n1, n2)
wilcox_p_1tail
```

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

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.

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

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  verizon$ILEC and verizon$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 similar?

We cannot reject the null hypothesis since there is not enough evidence and the p-value is above 0.01

Question 4

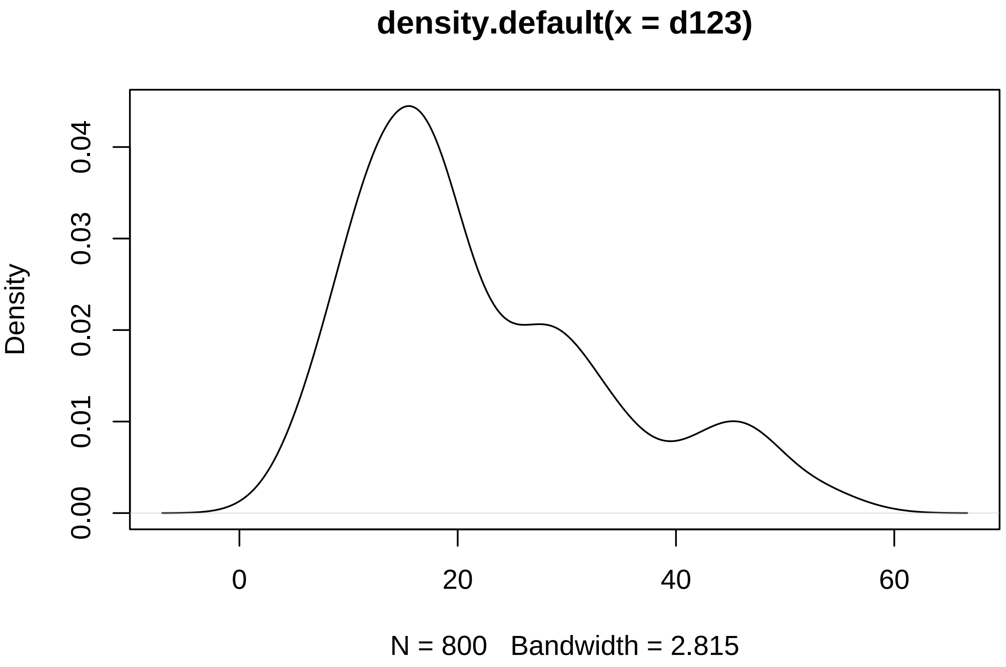
PART A

```
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)
}
```

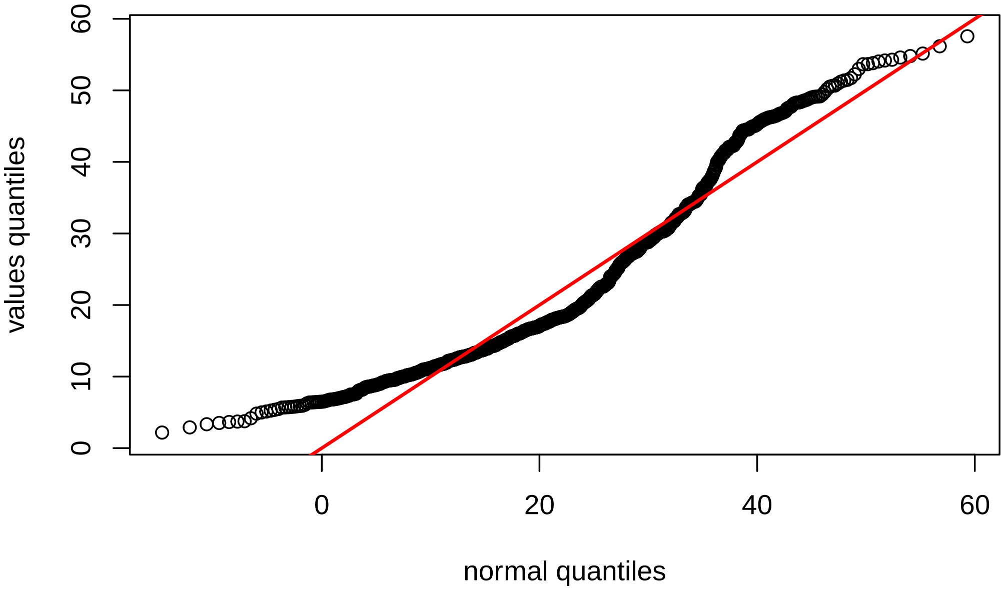
PART B

```
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))
```



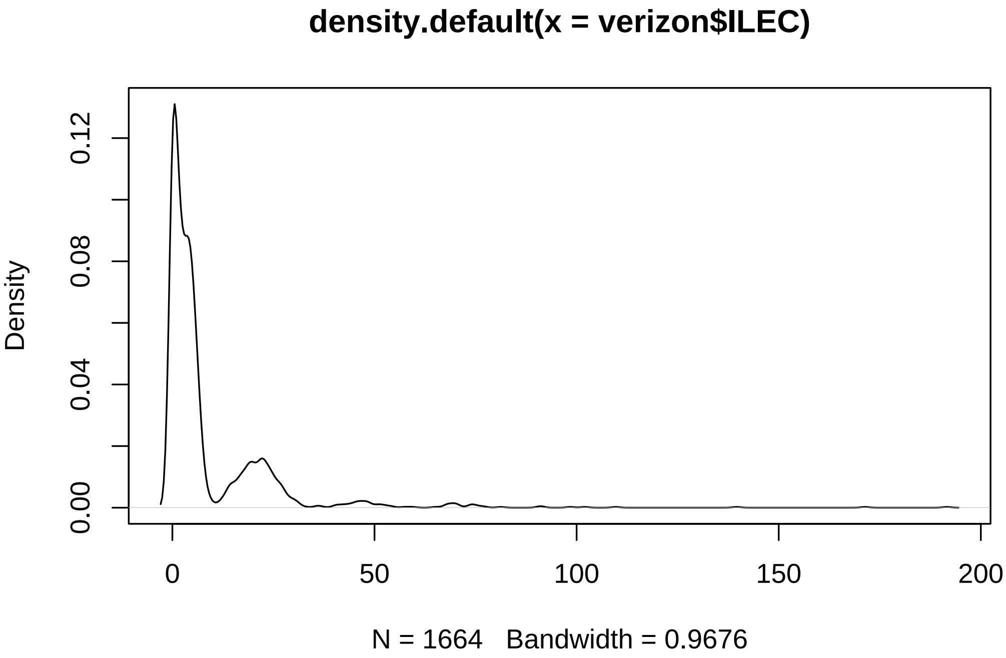
```
norm_qq_plot(d123)
```



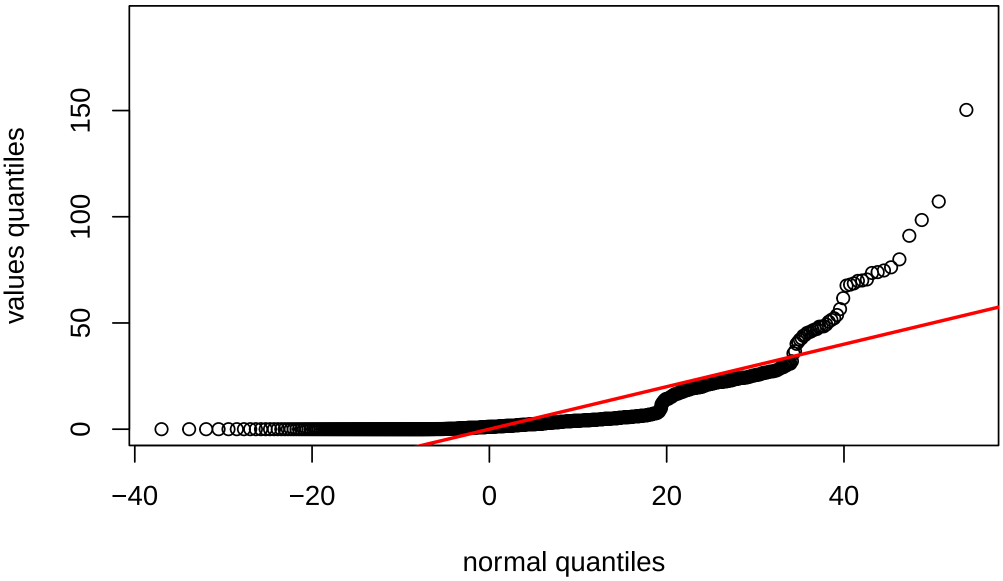
The plot is Rightly-Skewed. Most of the data is distributed on the left side with a long "tail" of data extending out to the right. Looking at the Q-Q plot we can see that the last two theoretical quantiles for this dataset should be around 60, when in fact those quantiles are greater than 50. The red line shows where the points would fall if the dataset were normally distributed.Thus, this plot is not normally distributed.

PART C

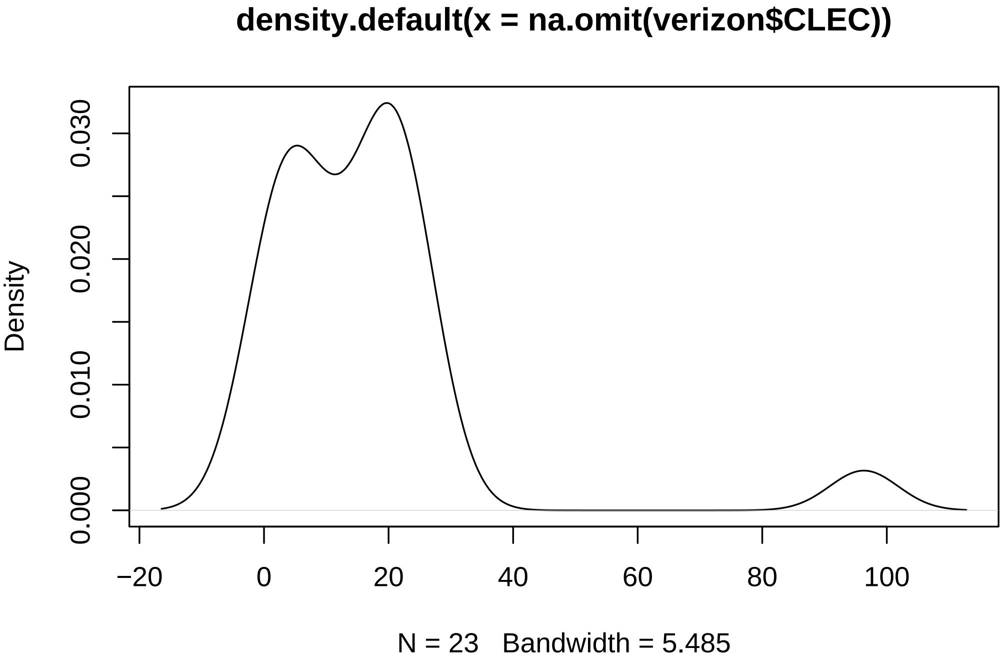
```
#ILEC
plot(density(verizon$ILEC))
```



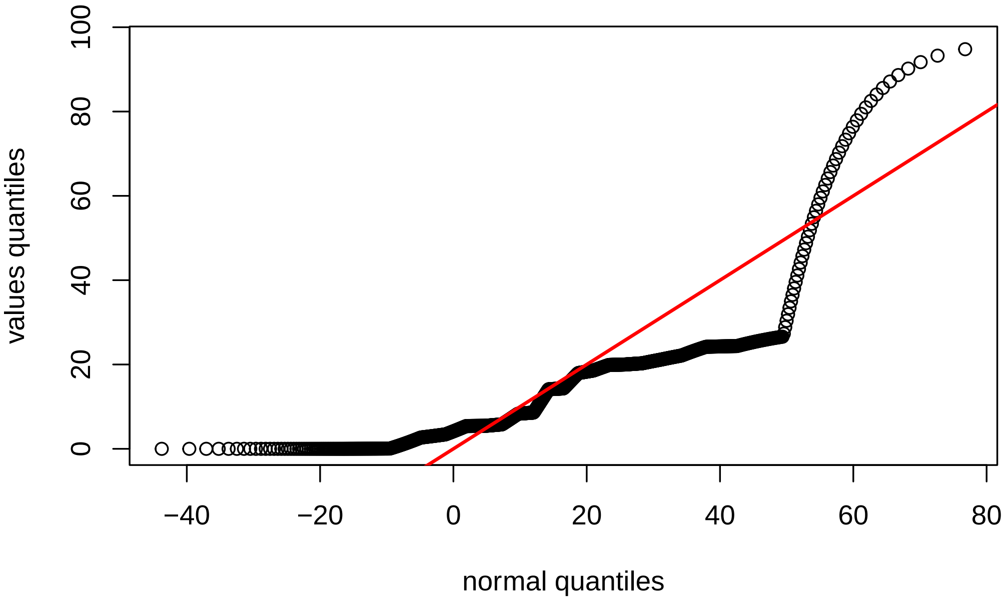
```
norm_qq_plot(verizon$ILEC)
```



```r
#CLEC
plot(density(na.omit(verizon$CLEC)))
```



```r
norm_qq_plot(na.omit(verizon$CLEC))
```



Both of ILEC and CLEC are not normally distributed since most of the data points doesn't sit around in the red lines.