Limitations of Log Transformation

Consider the following data set X. We wish to see the assumption of normality is valid for subsequent analyses.

```
X <- c(25, 10090, 132, 5941, 161, 2236, 38, 1011, 177, 431, 0,
       355, 7024, 19, 6771, 2750, 1324, 2705, 96, 215)

shapiro.test(X)

# shapiro.test(X)
#
#        Shapiro-Wilk normality test
#
# data:  X
# W = 0.72826, p-value = 8.8e-05
```

The p-value is 8.8e-05 i.e. very very low. We reject the null hypothesis that this sample is drawn for a normally distributed population of values.

We now investigate if a log-transformation would be useful. However, there is a value of 0 in the data set, for which a logarithm can not be computed.

```
log(X)
#  [1] 3.218876 9.219300 4.882802 8.689633 5.081404 7.712444 3.637586 6.918695
#  [9] 5.176150 6.066108     -Inf 5.872118 8.857088 2.944439 8.820404 7.919356
# [17] 7.188413 7.902857 4.564348 5.370638
```

As we can see below, the Shapiro-Wilk test fails to give us a valid answer. (`p-value = NA`).

```
shapiro.test(log(X))

#
#        Shapiro-Wilk normality test
#
# data:  log(X)
# W = NaN, p-value = NA
```

We may try out a different transformation instead. (see Tukey's Ladder).