Homework 1

Jordan Garrett

```r
library(faraway)
library(MASS)
```

1)

```r
fun_functions <- function(x){
  std <-  sd(x)
  mean.AD <- (sum(abs(x-mean(x))))*(length(x)^-1)
  med.AD <- median(abs(x-median(x)))
  interq <- IQR(x)
  return(c("SD"=std,
           "Mean Absolute Deviastion" = mean.AD,
           "Median Absolute Deviation" = med.AD,
           "IQR"=interq))
}
```

2)

Apply above function to variables *pregnancy*, *diastolic*, *bmi*, *age*

```r
explore.pregnancy <- fun_functions(pima$pregnant)
explore.diastolic <- fun_functions(pima$diastolic)
explore.bmi <- fun_functions(pima$bmi)
explore.age <- fun_functions(pima$age)

explore.pregnancy
```

```
##                        SD  Mean Absolute Deviastion
##                  3.369578                  2.771620
## Median Absolute Deviation                       IQR
##                  2.000000                  5.000000
```

```r
explore.diastolic
```

```
##                        SD  Mean Absolute Deviastion
##                  19.35581                  12.63942
## Median Absolute Deviation                       IQR
##                   8.00000                  18.00000
```

```r
explore.bmi
```

```
##                         SD  Mean Absolute Deviastion
##                    7.88416                   5.84227
## Median Absolute Deviation                       IQR
##                    4.60000                   9.30000
```

```
explore.age
```

```
##                         SD  Mean Absolute Deviastion
##                  11.760232                  9.586405
## Median Absolute Deviation                       IQR
##                   7.000000                 17.000000
```

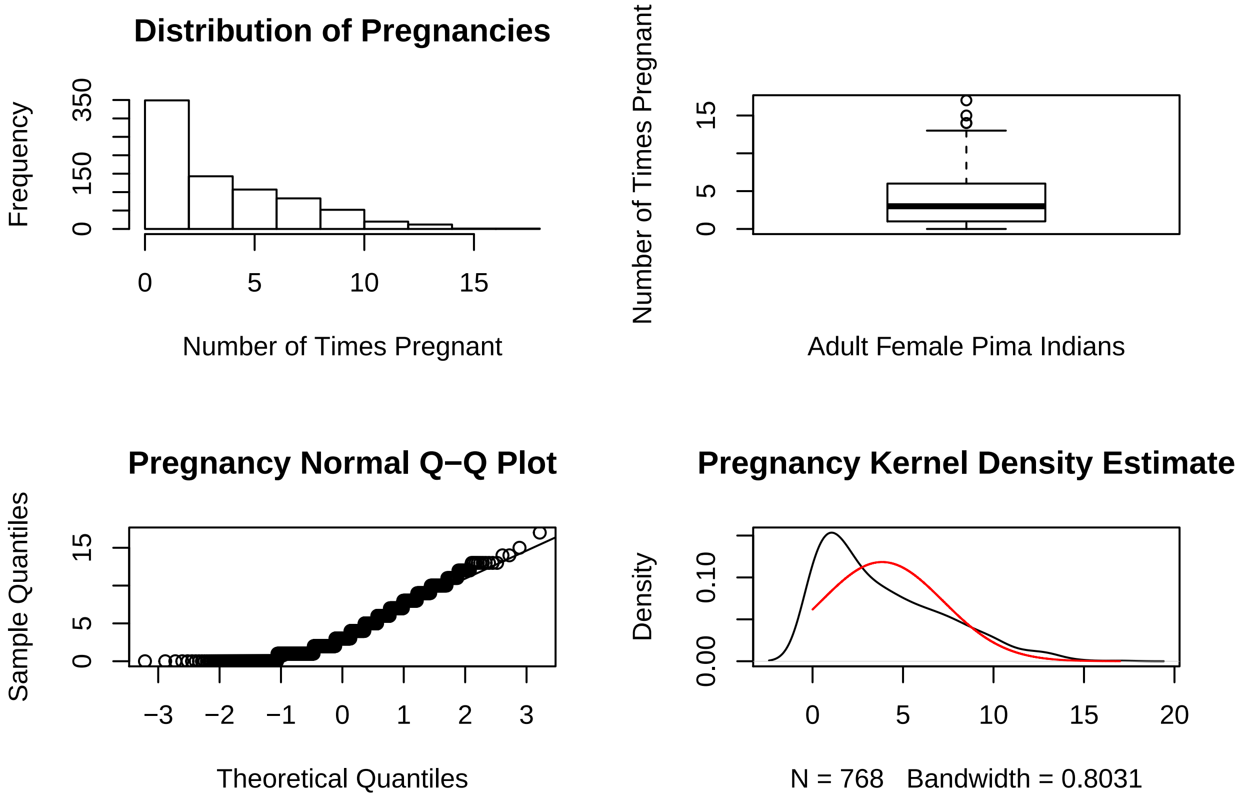
Create eda plot function

```
plot_distributions <- function(x,hist_xlab,hist_title,box_xlab,box_ylab,qq_title,kd_title){
  par(mfrow=c(2,2))
    hist(x, xlab=hist_xlab, main=hist_title)
    boxplot(x, xlab=box_xlab, ylab=box_ylab)
    qqnorm(x, main = paste(qq_title,'Normal Q-Q Plot'))
    qqline(x)
    plot(density(x), main=paste(kd_title,'Kernel Density Estimate'))
    y <- seq(min(x),max(x),0.001)
    lines(y,dnorm(y,mean=mean(x),sd=sd(x)),col="red") # compare with normal distribution
}
```

Plot edas

- Pregnancy

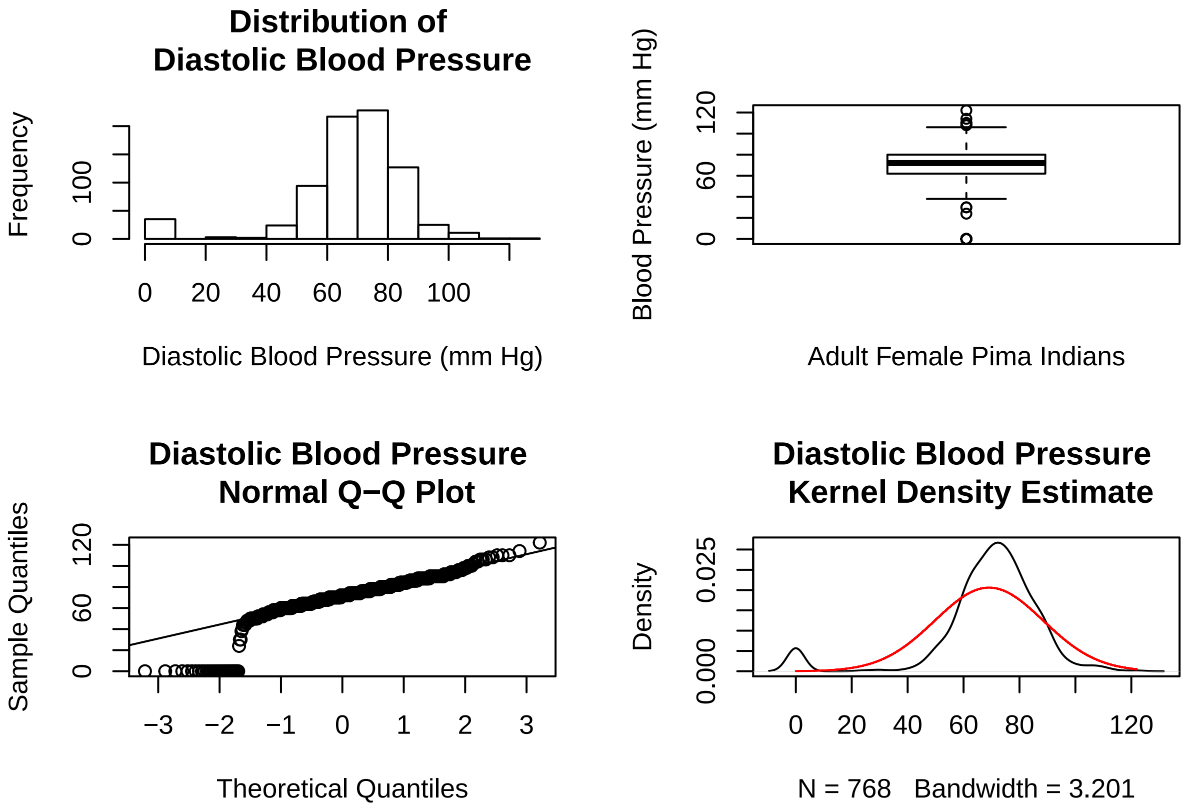
```
plot_distributions(pima$pregnant,'Number of Times Pregnant','Distribution of Pregnancies',
'Adult Female Pima Indians','Number of Times Pregnant','Pregnancy','Pregnancy')
```



Each of the plots indicate that the distribution of pregnancies is skewed right or positively skewed. Both the boxplot and slightly U-shaped Q-Q plot reveal that there are a few outliers within the 75th IQR.

- Diastolic

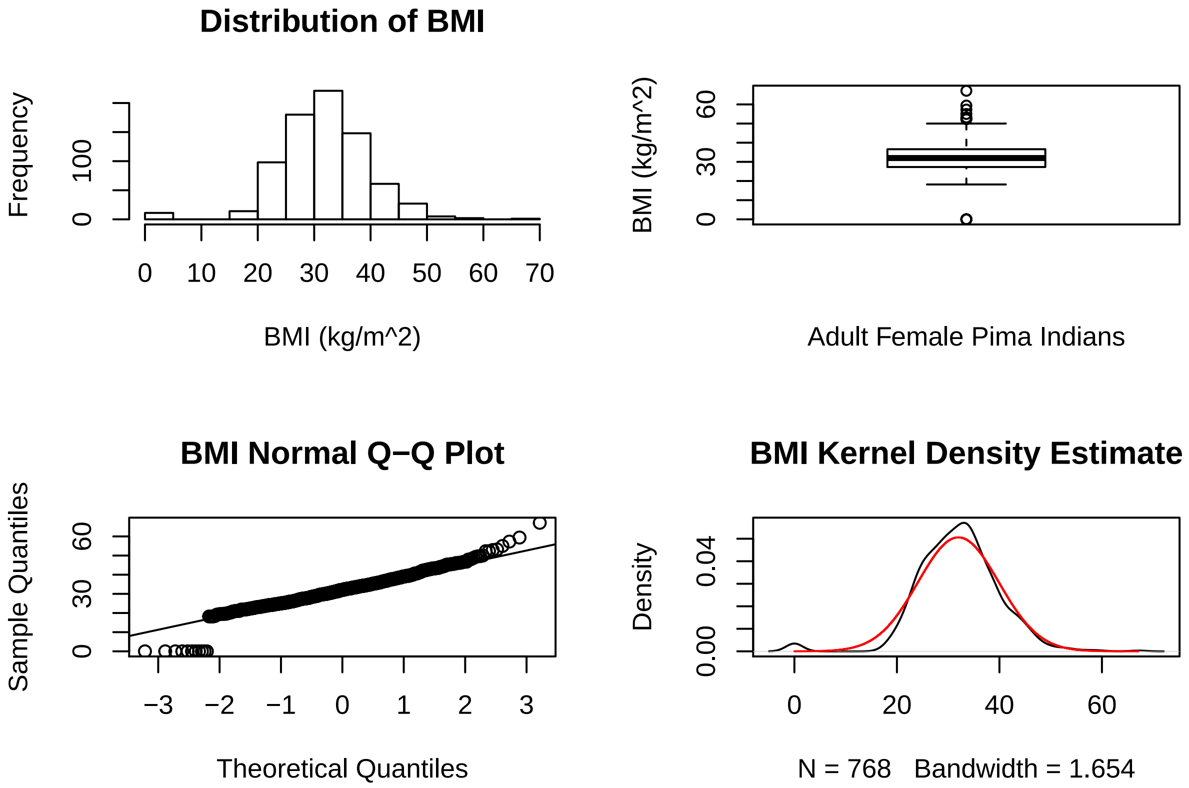
```
plot_distributions(pima$diastolic,'Diastolic Blood Pressure (mm Hg)',
                   'Distribution of \nDiastolic Blood Pressure',
                   'Adult Female Pima Indians',
                   'Blood Pressure (mm Hg)',
                   'Diastolic Blood Pressure \n',
                   'Diastolic Blood Pressure \n')
```



The kernel density estimate, histogram, and Q-Q plot all indicate that the distribution of diastolic blood pressure is bimodal. This is most likely due to missing data being entered in as a "0" rather than "N/A", since it is not possible to have a blood pressure of 0 if you are alive. The boxplot suggests that there are outliers in our distribution, although this may change once missing values are properly coded. If missing data was excluded, the distribution would be fairly normal.

- BMI

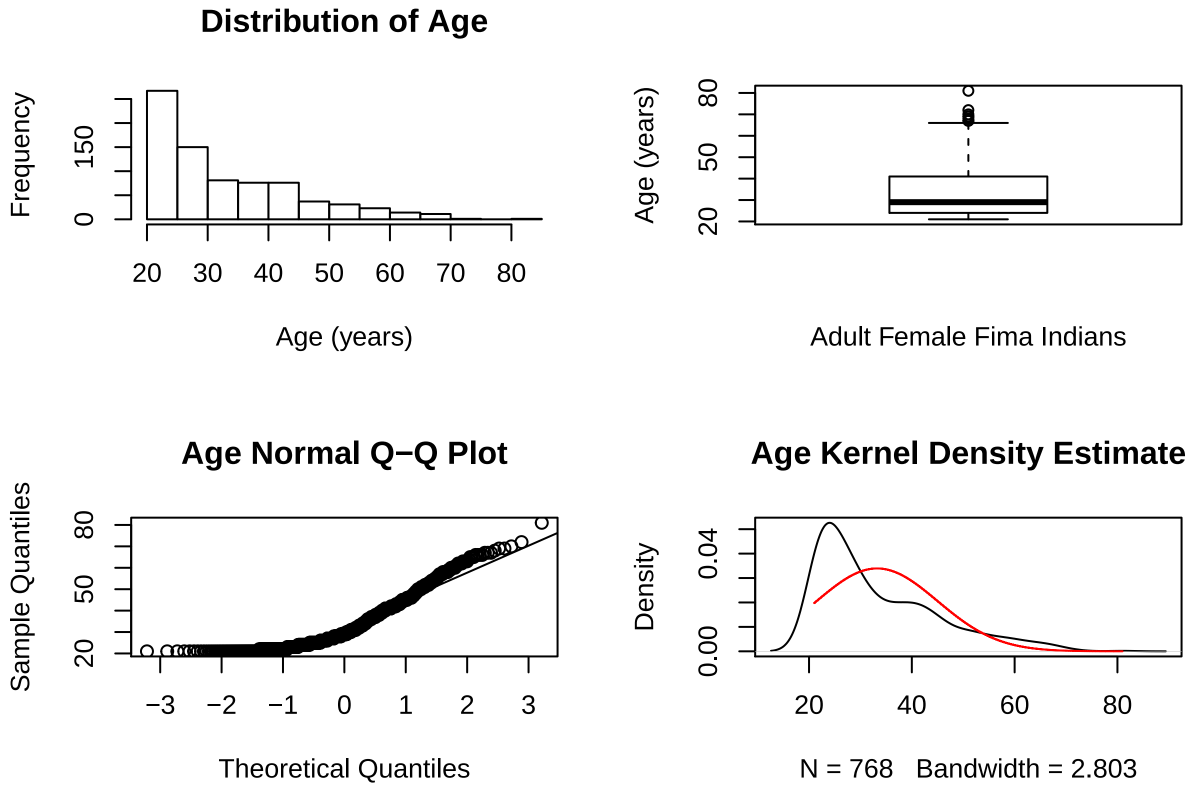
```
plot_distributions(pima$bmi,
                   'BMI (kg/m^2)',
                   'Distribution of BMI',
                   'Adult Female Pima Indians',
                   'BMI (kg/m^2)','BMI','BMI')
```



Once again, missing values in the data produce a bimodal distribution. It is apparent that there are incorretly coded missing values since it is not possible to have a bmi of 0. If missing values were excluded, then the distribution of bmi would have a slight positive skew. This is most apparent in the histogram. Boxplot and Q-Q plot results indicate that there are outliers within the 75th IQR of the data.

- Age

```r
plot_distributions(pima$age,'Age (years)',
                    'Distribution of Age',
                    'Adult Female Fima Indians',
                    'Age (years)','Age','Age')
```



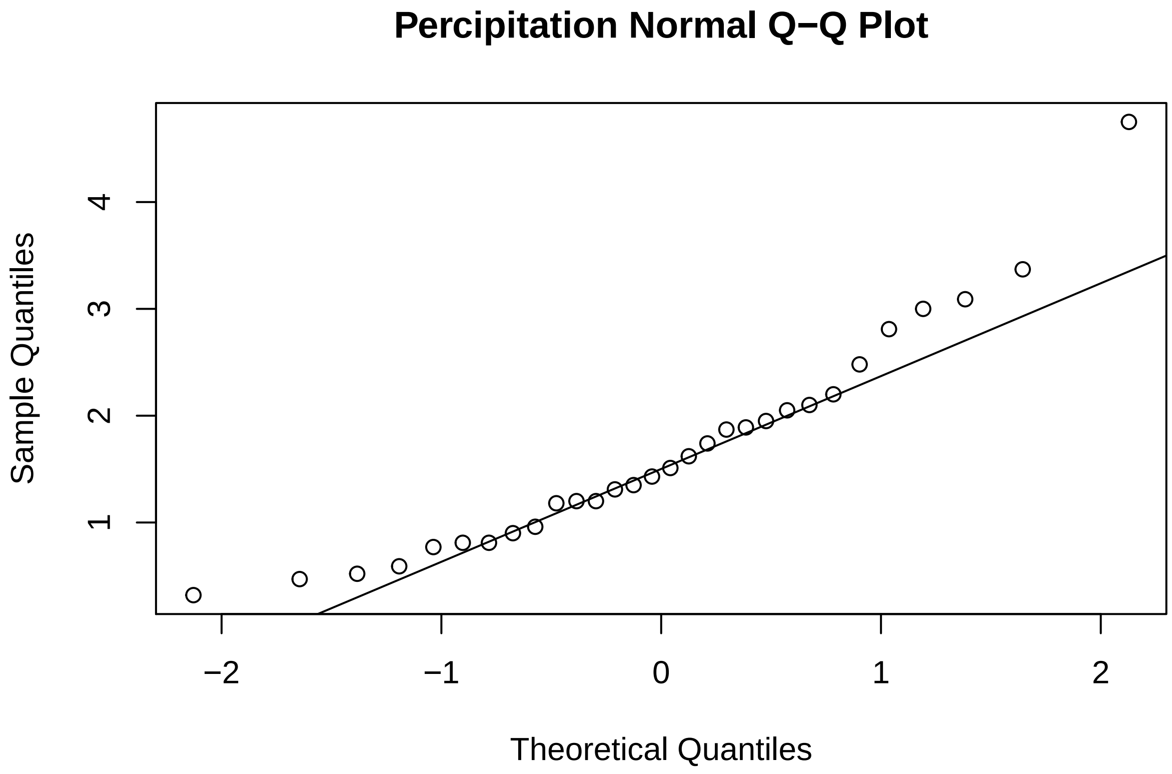
The distribution of age is positively skewed (or skewed right). The age of the sample was between 20-80 years old, with a few outliers above the age of 70. The s/u-shaped Q-Q plot indicates that the data is slightly under dispersed. The kernel density estimate suggests that the distribution is also slighty multimodal.

3)

```
precipt <- c(0.77, 1.20, 3.00, 1.62, 2.81, 2.48, 1.74, 0.47, 3.09, 1.31,
            1.87, 0.96, 0.81, 1.43, 1.51, 0.32, 1.18, 1.89, 1.20, 3.37,
            2.10, 0.59, 1.35, 0.90, 1.95, 2.20, 0.52, 0.81, 4.75, 2.05)
```

a)

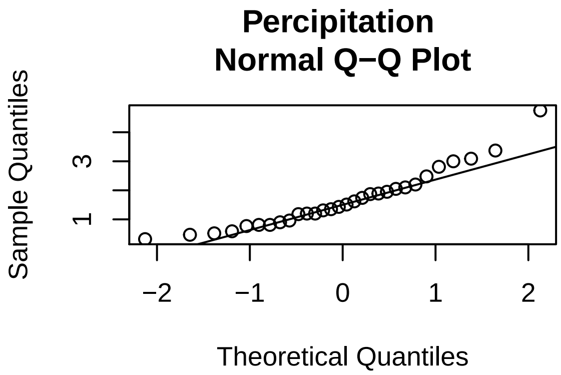
```
qqnorm(precipt, main = 'Percipitation Normal Q-Q Plot')
qqline(precipt)
```

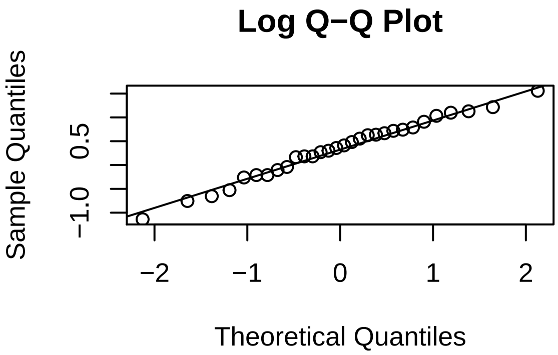


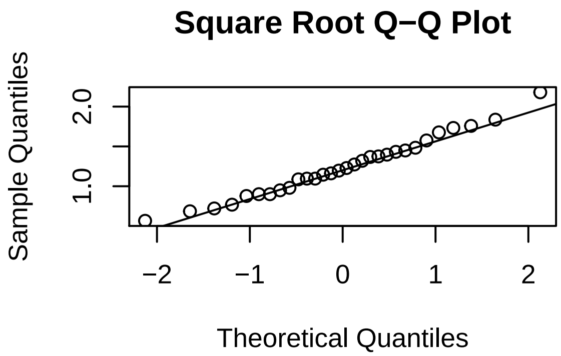
The percipitation data is not normally distributed, but is positively skewed as indicated by the slight U-shape of the plot.

b)

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

qqnorm(precipt, main = 'Percipitation \nNormal Q-Q Plot')
qqline(precipt)

qqnorm(log(precipt), main = 'Log Q-Q Plot')
qqline(log(precipt))

qqnorm(sqrt(precipt), main='Square Root Q-Q Plot')
qqline(sqrt(precipt))
```





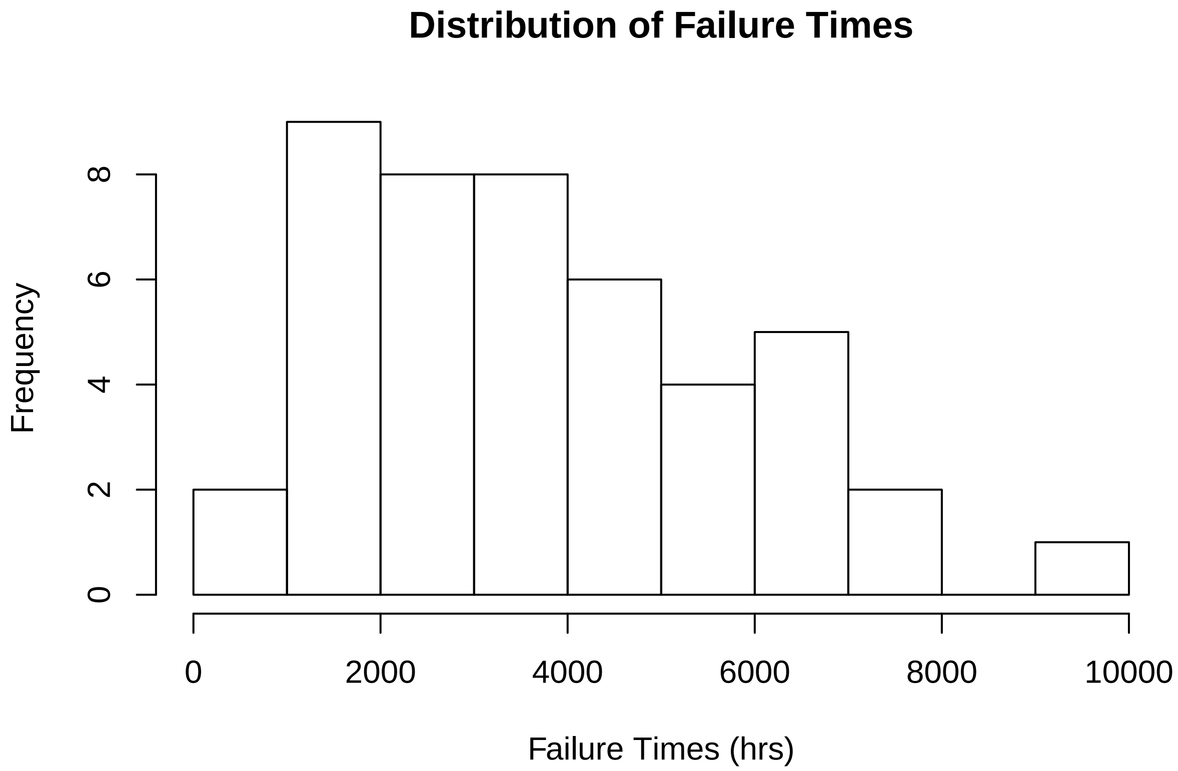


Log and square root transforms of the precipitation data are the closest to satisfying the normality assumption.

4)

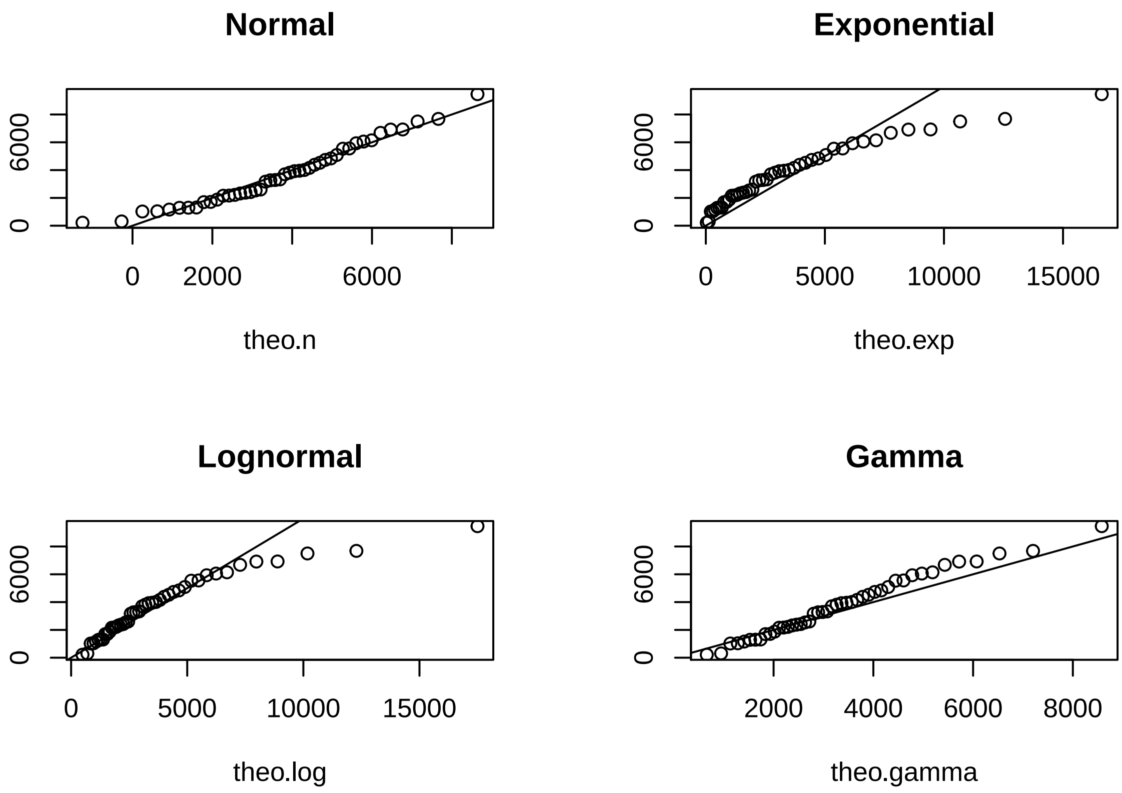
```
failure.trans <- c(4381,3953,2603,2320,1161,3286,6914,4007,3168,
              2376,7498,3923,9460,4525,2168,1288,5085,2217,
              6922,218,1309,1875,1023,1697,1038,3699,6142,
              4732,3330,4159,2537,3814,2157,7683,5539,4839,
              6052,2420,5556,309,1295,3266,6679,1711,5931)
```

```
hist(failure.trans, main = "Distribution of Failure Times",xlab='Failure Times (hrs)')
```



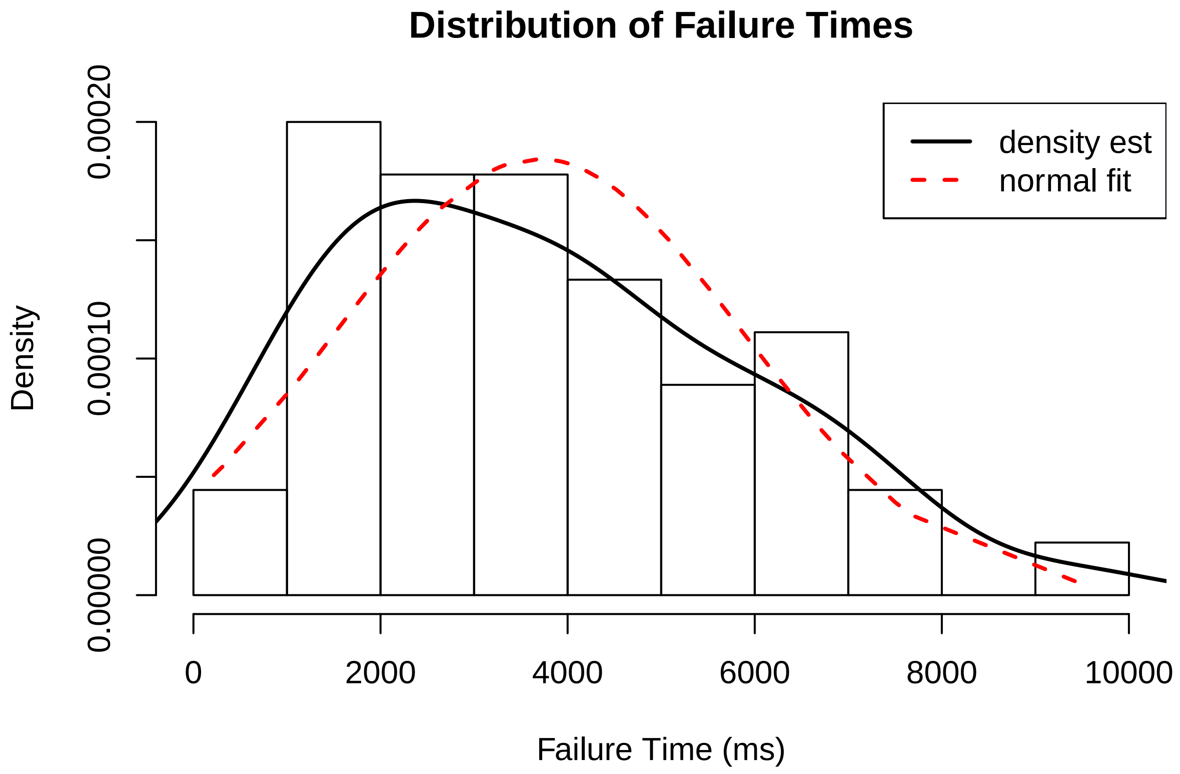
```r
par(mfrow=c(2,2))
fit.n <- fitdistr(failure.trans,"normal",lower=0.0011) # fit normal distribution
theo.n <- qnorm(ppoints(failure.trans), mean=fit.n$estimate[1],
                sd=fit.n$estimate[2])

fit.exp <- fitdistr(failure.trans,"exponential",lower=0.0011) # fit exponential distribution
theo.exp <- qexp(ppoints(failure.trans), rate=fit.exp$estimate)

fit.log <- fitdistr(failure.trans,"lognormal",lower=0.0011)
theo.log <- qlnorm(ppoints(failure.trans), meanlog=fit.log$estimate[1],
                   sdlog=fit.log$estimate[2])

fit.gamma <- fitdistr(failure.trans,"gamma",lower=0.0011)
theo.gamma <- qgamma(ppoints(failure.trans), fit.gamma$estimate[1],
                     fit.gamma$estimate[2])

par(mfrow=c(2,2))

qqplot(theo.n,failure.trans,ylab="", main="Normal")
abline(0,1)

qqplot(theo.exp,failure.trans,ylab="", main="Exponential")
abline(0,1)

qqplot(theo.log,failure.trans,ylab="", main="Lognormal")
abline(0,1)
```

```r
qqplot(theo.gamma,failure.trans,ylab="", main="Gamma")
abline(0,1)
```



Q-Q plots indicate that the distribution of failure time closely resembles a normal distribution.

```r
hist(failure.trans,freq=F, main = 'Distribution of Failure Times',xlab='Failure Time (ms)')
#density
lines(density(failure.trans),xlim=c(0,2),lwd=2)

#normal
lines(sort(failure.trans),dnorm(sort(failure.trans),fit.n$est[1],
                                fit.n$est[2]),col='red', lwd=2, lty=2)

legend("topright",
       c("density est","normal fit"),
       col=c("black","red"),lwd=2,lty=c(1,2))
```

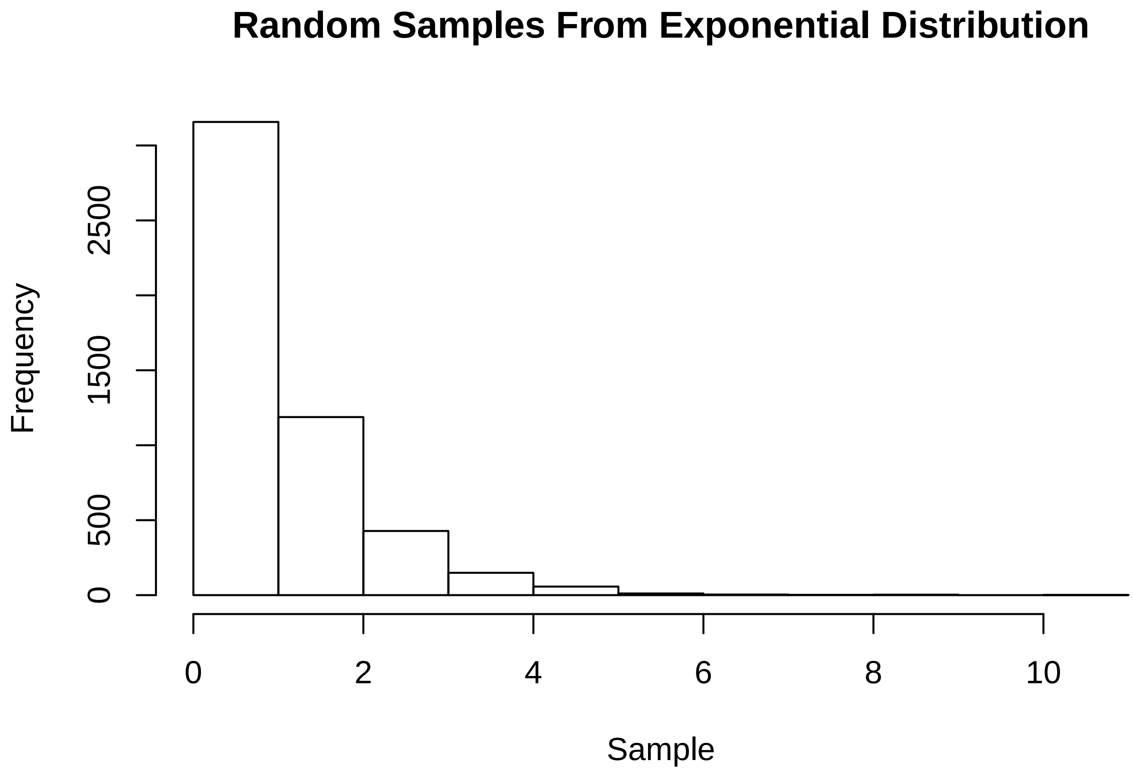


Subsequent plotting of the normal and gamma distribution confirm the results of the Q-Q plot. The normal distribution encompasses the kurtosis of the observed data, although it fails to capture the observed distribution's positive skew.

5)

```
fake.data_1 <- rexp(5000,1)
```

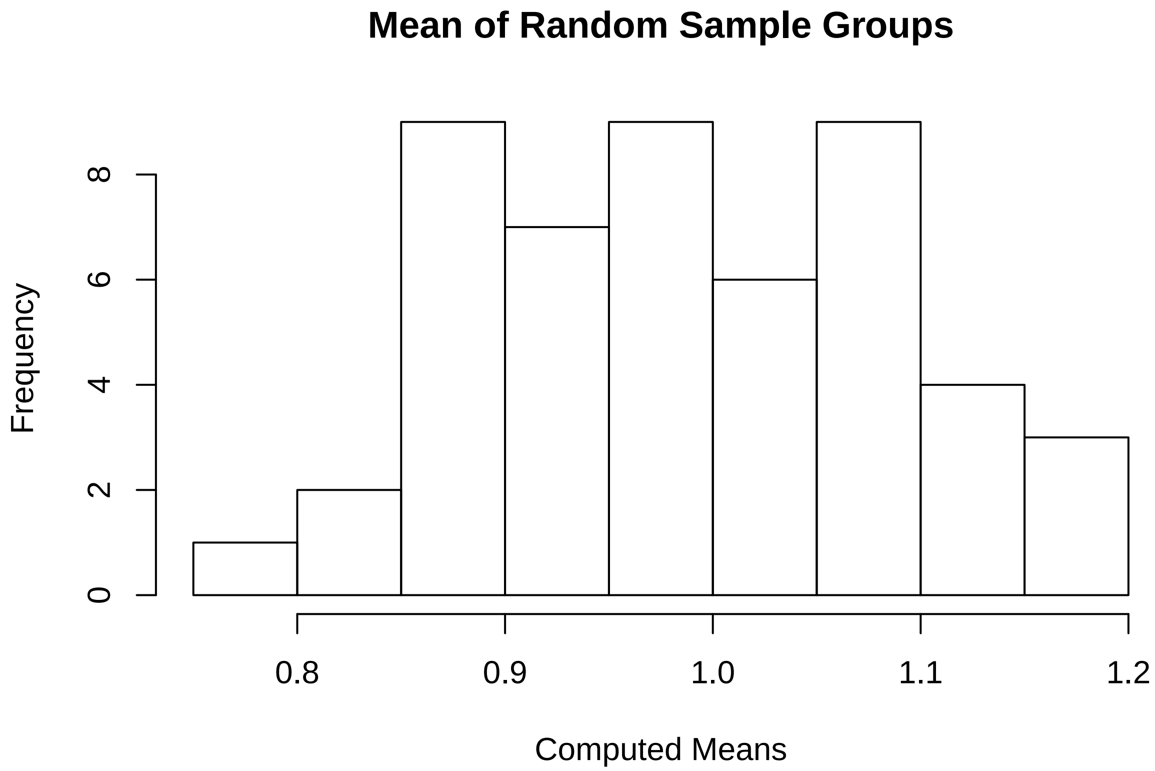
```
hist(fake.data_1, main='Random Samples From Exponential Distribution',
     xlab='Sample')
```



The distribution is positively skewed.

```
n <- 50
nn <- 100
fake.data.groupd_1 <- split(fake.data_1,rep(1:n, each=1))
```

```
hist(sapply(fake.data.groupd_1, mean),
     main = 'Mean of Random Sample Groups', xlab='Computed Means')
```



The two histograms do not have the same shape. Jack knifing the data creates a second distribution that is more normal due to the central limit thereom (continuosly taking the mean).

6)

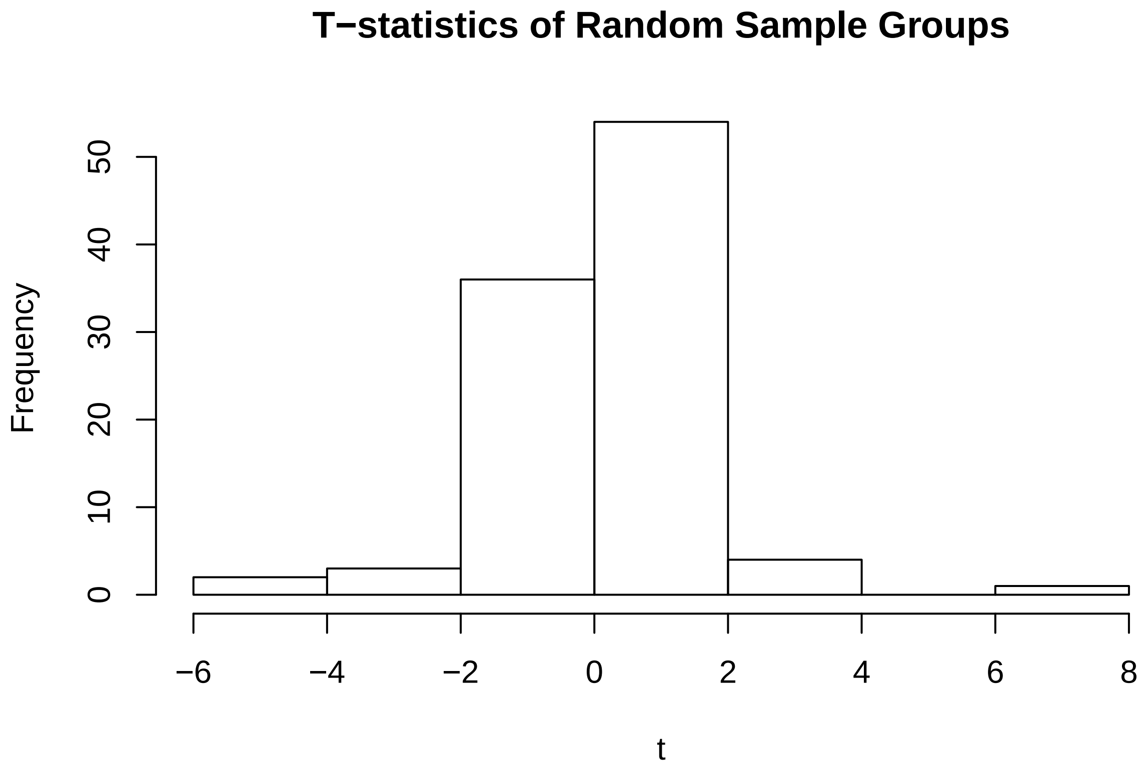
```
fake.data_2 <- rnorm(600,mean=10,sd=5)
```

```
n <- 100
nn <- 6
fake.data.groupd_2 <- split(fake.data_2,rep(1:n,each=1))
```

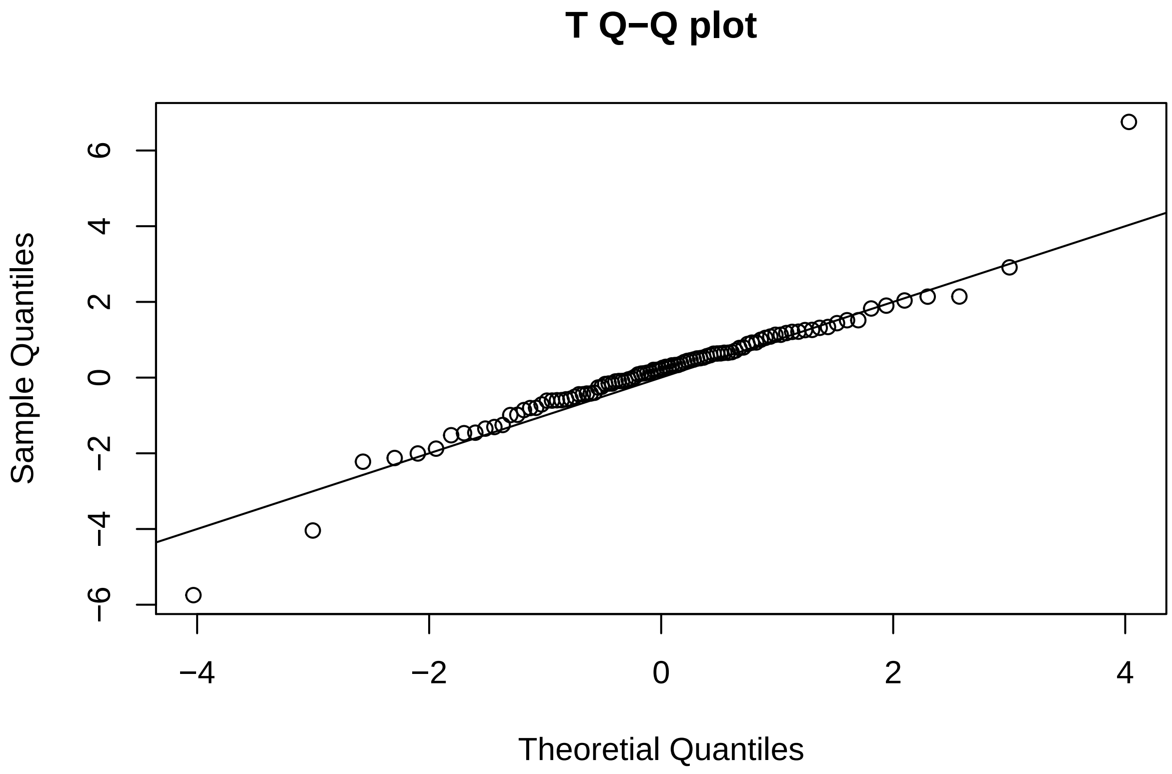
```
custom_function <- function(x){
  output <- (mean(x)-10)/sqrt(var(x)/6)
  return(output)
}
```

I expect the statistic to follow a Normal t-distribution since we are computing the t-statistic for each sample.

```
hist(sapply(fake.data.groupd_2,custom_function),
     main='T-statistics of Random Sample Groups',xlab='t')
```



```r
plot(qt(ppoints(sapply(fake.data.groupd_2,custom_function)),5),
     sort(sapply(fake.data.groupd_2,custom_function)),
     main="T Q-Q plot", xlab="Theoretial Quantiles",
     ylab="Sample Quantiles")
abline(0,1)
```



The generated distribution closely approximates a Normal t-distribution.