Assignment 4

Maharaj Teertha Deb, 40227747

2024-02-27

Section 3.1

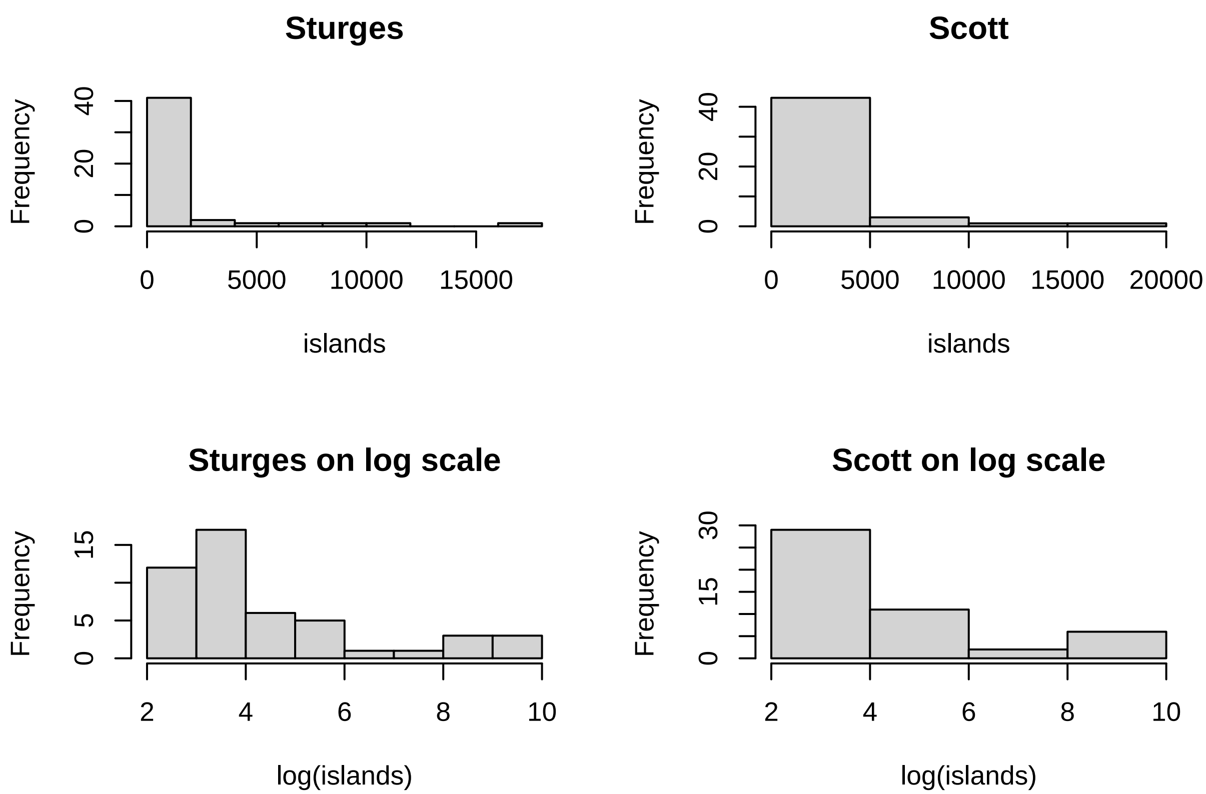
Problem 1 :

Consider the islands vector discussed in this section.

(a) Compare the histograms that result when using breaks based on Sturges' and Scott's rules. Make this comparison on the log scale and on the original scale.

(a) Answer:

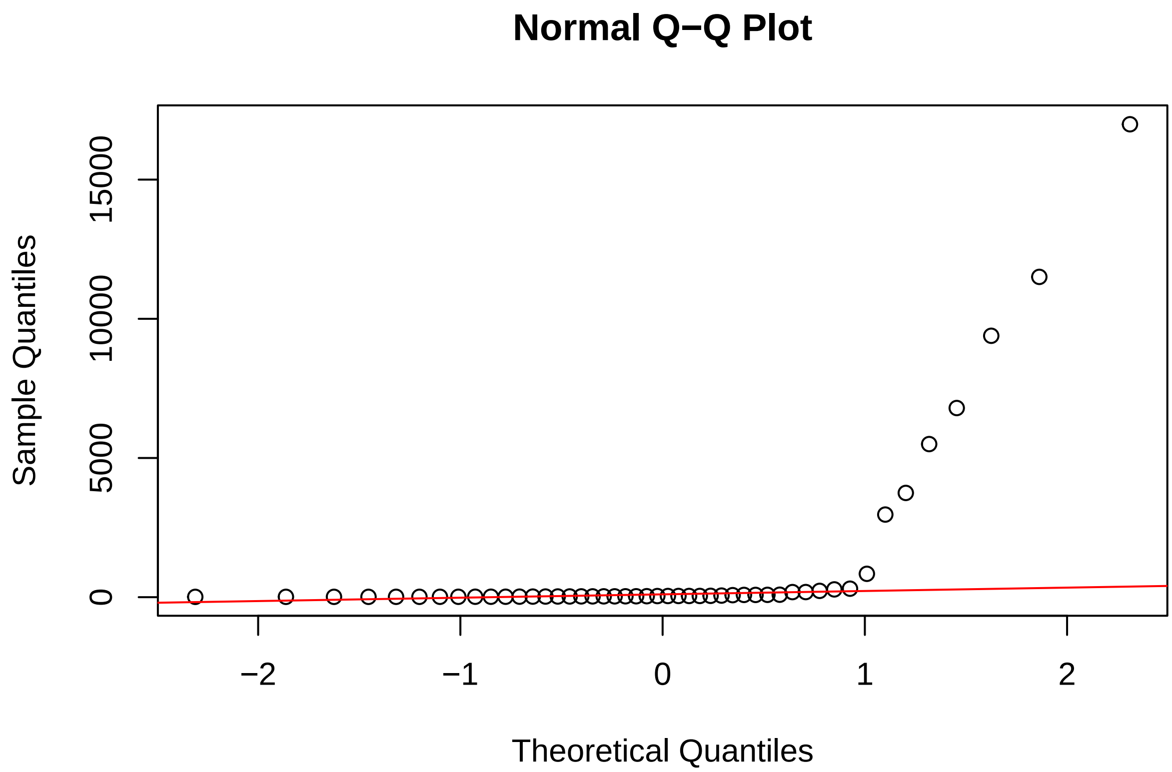
```
data(islands)

par(mfrow=c(2,2))
hist(islands, breaks = "Sturges", main = "Sturges")
hist(islands, breaks = "Scott", main = "Scott")
hist(log(islands), breaks = "Sturges", main = "Sturges on log scale")
hist(log(islands), breaks = "Scott", main = "Scott on log scale")
```



(b) Construct a normal QQ plot, and compare the result with the plots in Figure 3.13; which one is most similar, and what does this tell you about this data set?

(b) Answer:

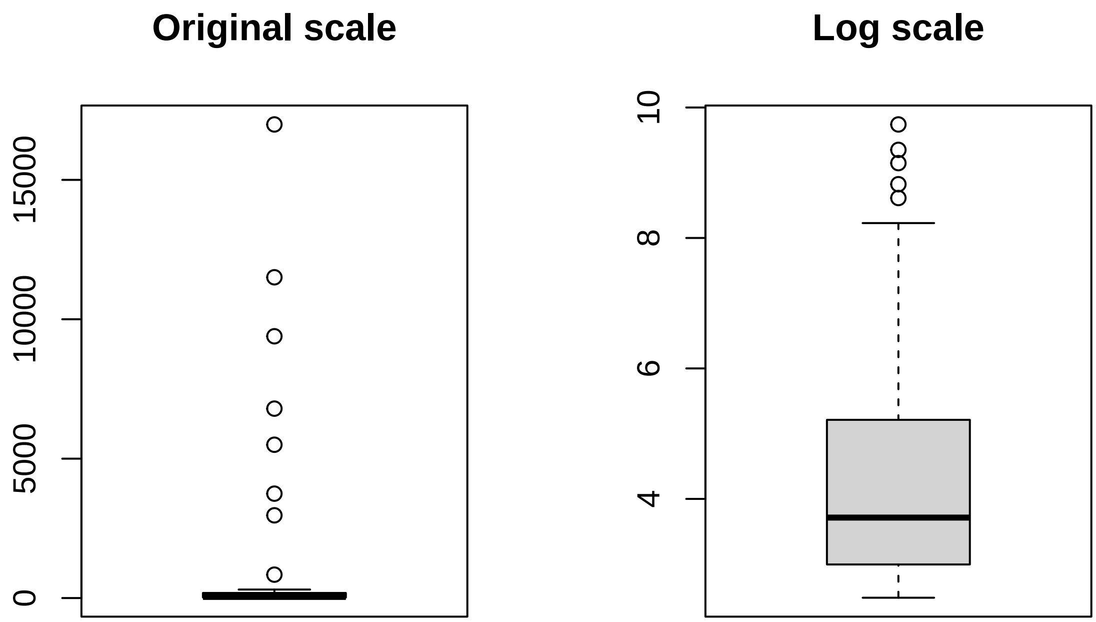
```r
qqnorm(islands)
qqline(islands, col = "red")
```



(c) Construct a boxplot for these data on the log scale as well as the original scale.

(c) Answer:

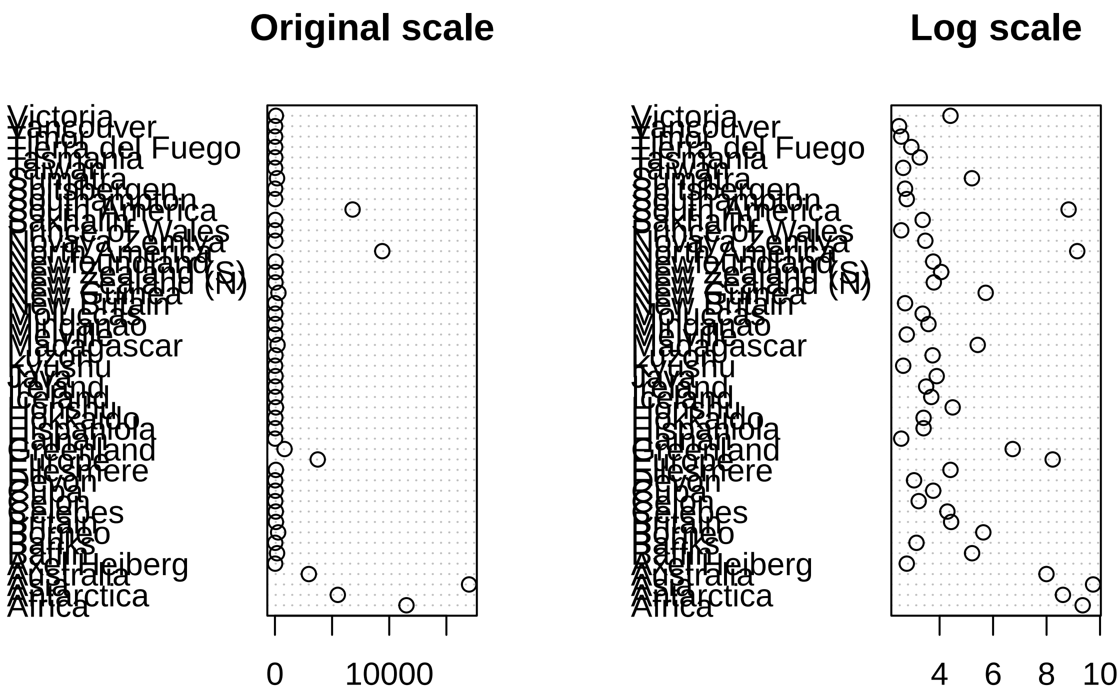
```
par(mfrow=c(1,2))
boxplot(islands, main = "Original scale")
boxplot(log(islands), main = "Log scale")
```



(d) Construct a dot chart of the areas. Is a log transformation needed here?

(d) Answer:

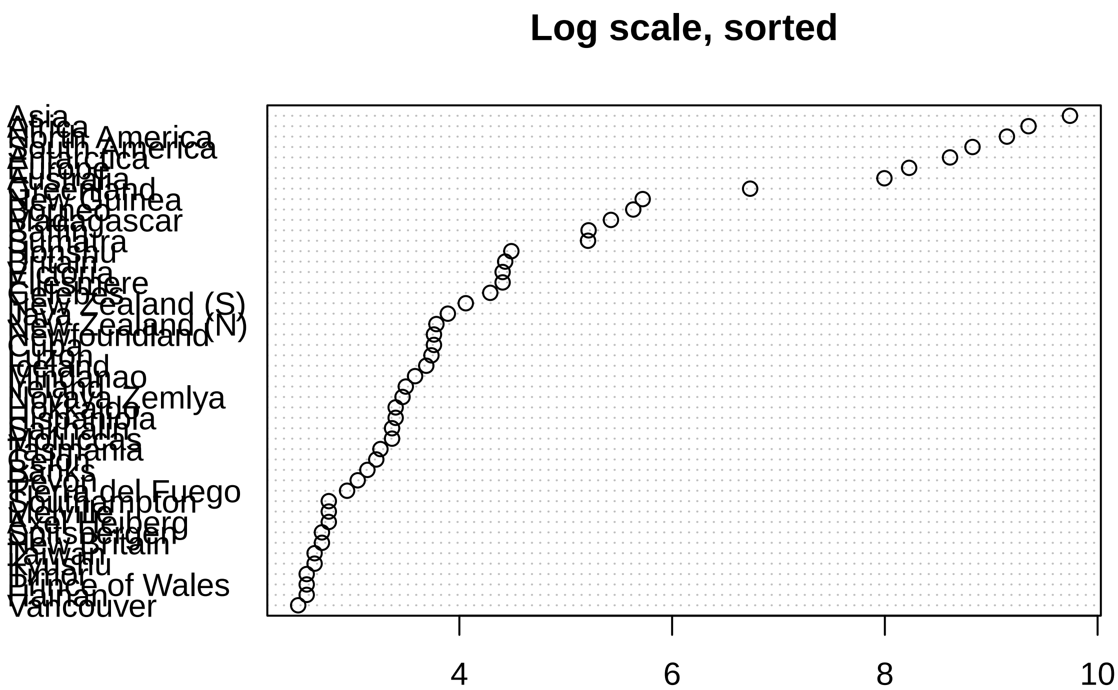
```
par(mfrow=c(1,2))
dotchart(islands,main="Original scale")
dotchart(log(islands),main="Log scale")
```



(e) Which form of graphic do you think is most appropriate for displaying these data?

(e) Answer:

```
dotchart(sort(log(islands)), main = "Log scale, sorted")
```



Problem 3:

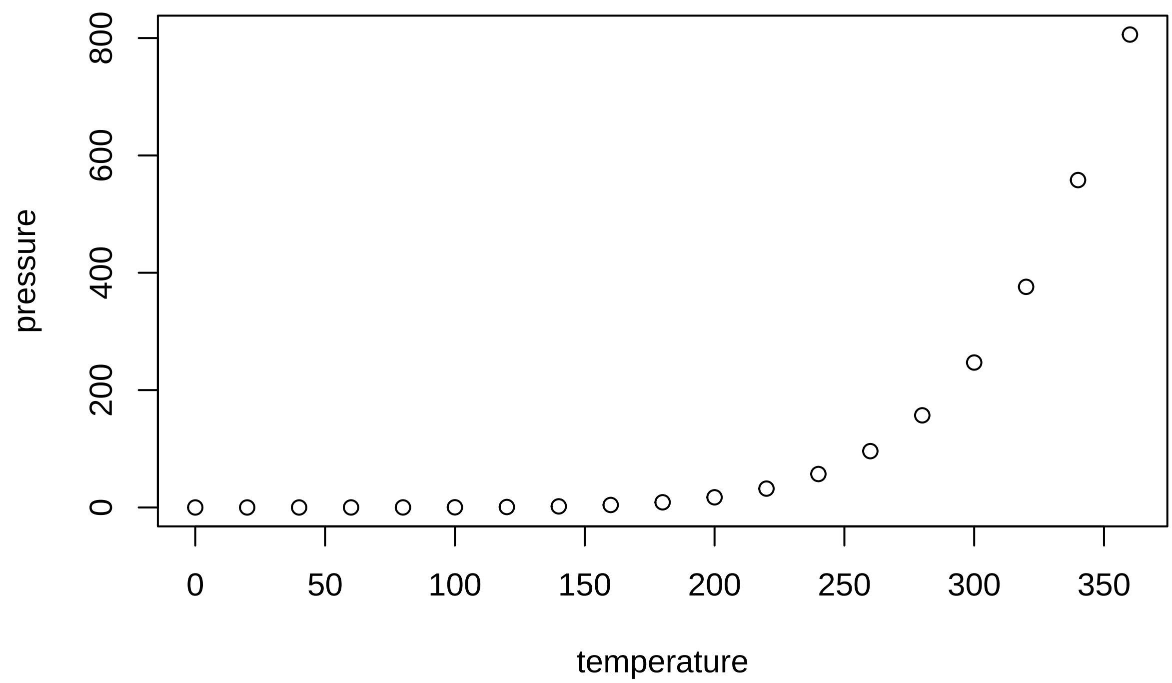
Consider the pressure data frame. There are two columns temperature and pressure

(a)

Construct a scatterplot with pressure on the vertical axis and temperature on the horizontal axis. Are the variables related linearly or nonlinearly?

(a) Answer:

```r
plot(pressure ~ temperature, data = pressure)# Calculate correlation coefficient
```




```r
correlation <- cor(pressure$temperature, pressure$pressure)

# Check for linearity
if(abs(correlation) < 0.005) {
  cat("The variables are related linearly (correlation =", correlation, ").\n")
} else {
  cat("The variables are related nonlinearly (correlation =", correlation, ").\n")
}
```

```
## The variables are related nonlinearly (correlation = 0.7577923 ).
```

(b)

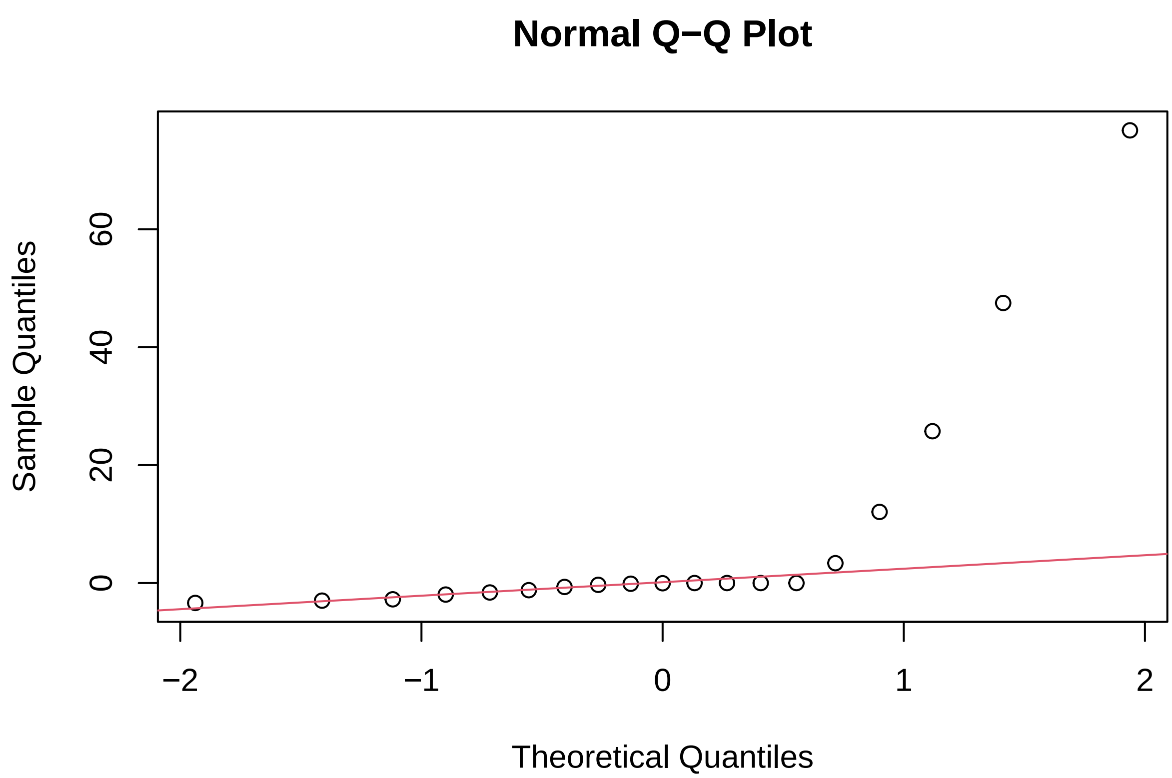
The graph of the following function passes through the plotted points reasonably well: $y = (0.168 + 0.007x)^{\frac{20}{3}}$. The differences between the pressure values predicted by the curve and the observed pressure values are called residuals. Here is a way to calculate them:

$$\text{residuals} < -with(\text{pressure }, \text{pressure} - (0.168 + 0.007 \times \text{temperature})^{\frac{20}{3}}$$

Construct a normal QQ plot of these residuals and decide whether they are normally distributed or whether they follow a skewed distribution.

(b) Answer:

```
# Calculate residuals
predicted_pressure <- (0.168 + 0.007 * pressure$temperature)^(20/3)
residuals <- pressure$pressure - predicted_pressure

# Construct a normal QQ plot
qqnorm(residuals)
qqline(residuals, col = 2)
```




```
# Assess normality
shapiro.test(residuals)
```

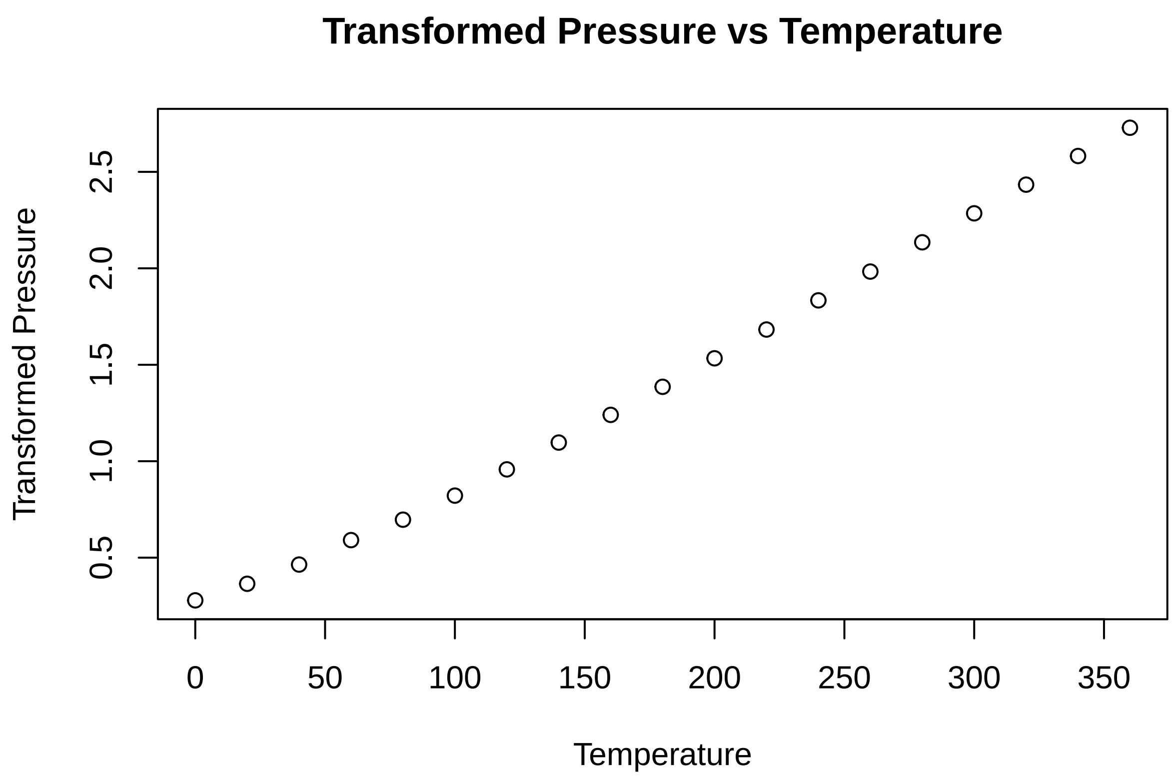
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.55893, p-value = 1.751e-06
```

(c)

Now, apply the power transformation $y^{\frac{3}{20}}$ to the pressure data values. Plot these transformed values against temperature. Is a linear or nonlinear relationship evident now?

(c) Answer:

```r
# Apply the power transformation y^(3/20) to pressure data values
transformed_pressure <- pressure$pressure^(3/20)

# Plot transformed values against temperature
plot(pressure$temperature, transformed_pressure,
     xlab = "Temperature", ylab = "Transformed Pressure",
     main = "Transformed Pressure vs Temperature")
```




```r
# Calculate correlation coefficient between transformed pressure and temperature
correlation_transformed <- cor(pressure$temperature, transformed_pressure)

# Check for linearity
if(abs(correlation_transformed) < 0.005) {
  cat("The relationship between transformed pressure and temperature appears to
      be linear (correlation =", correlation_transformed, ").\n")
} else {
  cat("The relationship between transformed pressure and temperature appears to
      be nonlinear (correlation =", correlation_transformed, ").\n")
}
```

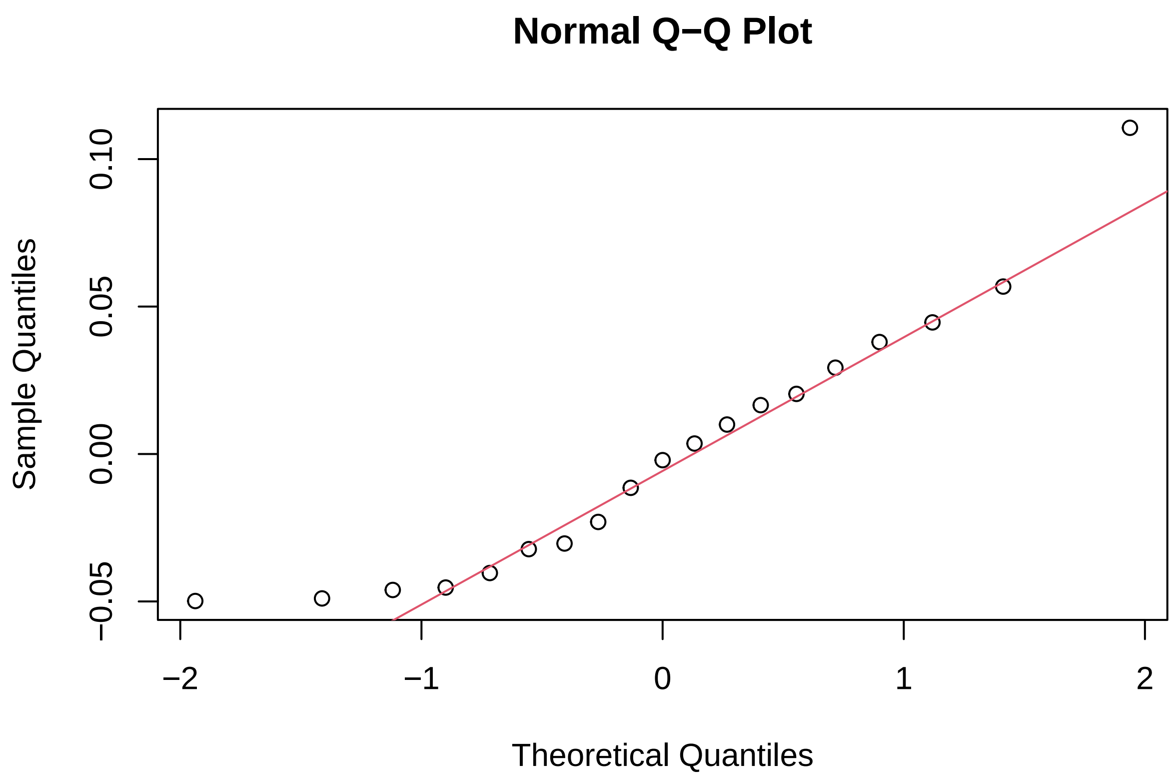
```
## The relationship between transformed pressure and temperature appears to
##       be nonlinear (correlation = 0.9984827 ).
```

(d)

Calculate residuals for the difference between transformed pressure values and those predicted by the straight line. Obtain a normal QQ plot, and decide whether the residuals follow a normal distribution or not.

(d) Answer:

```r
# Fit a linear regression model to transformed pressure values and temperature
lm_model_transformed <- lm(transformed_pressure ~ pressure$temperature)

# Predict transformed pressure values using the linear model
predicted_transformed_pressure <- predict(lm_model_transformed)

# Calculate residuals
residuals_transformed <- transformed_pressure - predicted_transformed_pressure

# Construct a normal QQ plot of residuals
qqnorm(residuals_transformed)
qqline(residuals_transformed, col = 2)
```




```r
# Assess normality
shapiro.test(residuals_transformed)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  residuals_transformed
## W = 0.92153, p-value = 0.1208
```

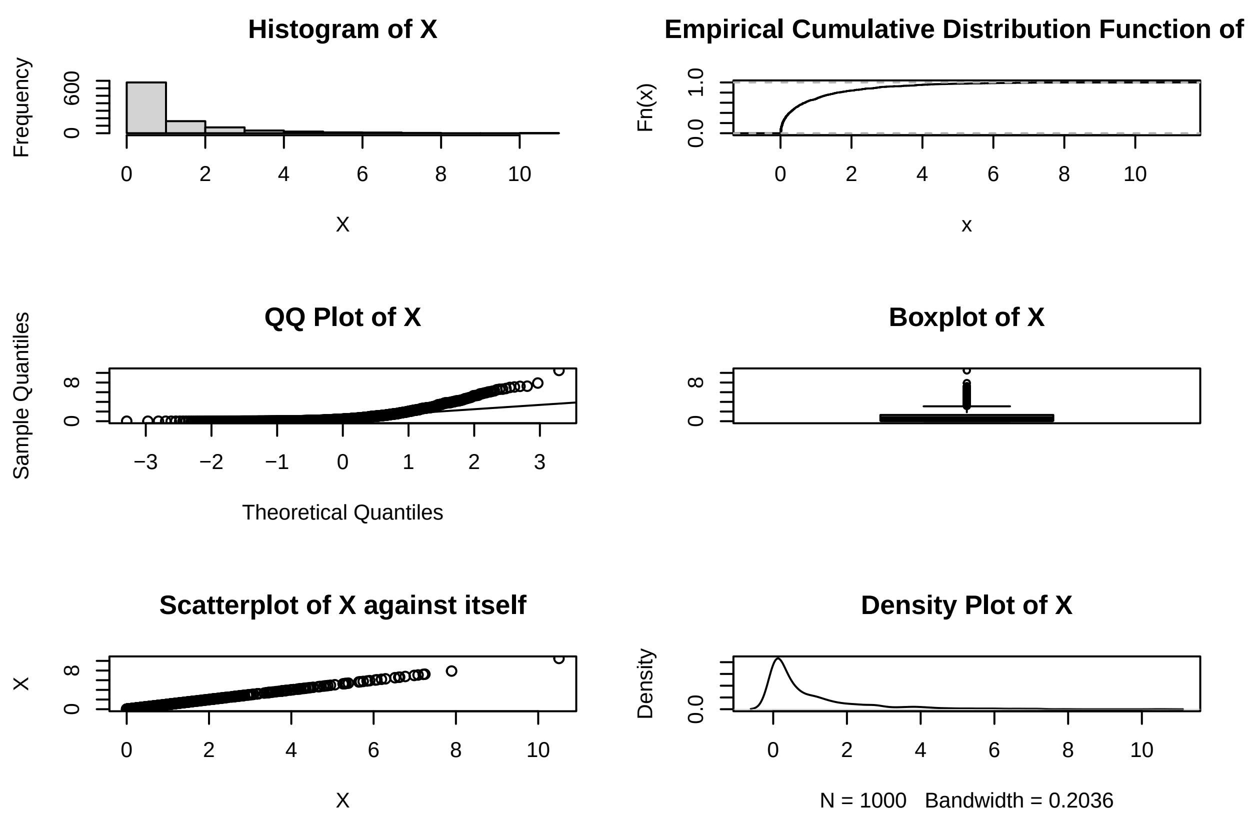
Section 3.3

Problem 4

Re-do the plots of Figure 3.18, but this time apply the six graphics functions to a variable X which is defined by $X = Z^2$, where Z is a simulated standard normal random variable as in the example. Comment on the changes that you see in the plots. (The variable X is an example of a chi-squared random variable on one degree of freedom.) Refer to the previous question.

Answer:

```r
# Load necessary library
library(MASS)  # For mvrnorm function

# Generate simulated standard normal random variable Z
set.seed(123)  # for reproducibility
Z <- rnorm(1000)

# Define variable X = Z^2
X <- Z^2

# Plotting Figure 3.18 with variable X
par(mfrow = c(3, 2))  # Set up a 3x2 layout for plots

# Plot histogram
hist(X, main = "Histogram of X")

# Plot ECDF
plot(ecdf(X), main = "Empirical Cumulative Distribution Function of X")

# Plot QQ plot
qqnorm(X, main = "QQ Plot of X")
qqline(X)

# Plot boxplot
boxplot(X, main = "Boxplot of X")

# Plot scatterplot of X against itself
plot(X, X, xlab = "X", ylab = "X", main = "Scatterplot of X against itself")

# Plot density plot
plot(density(X), main = "Density Plot of X")
```



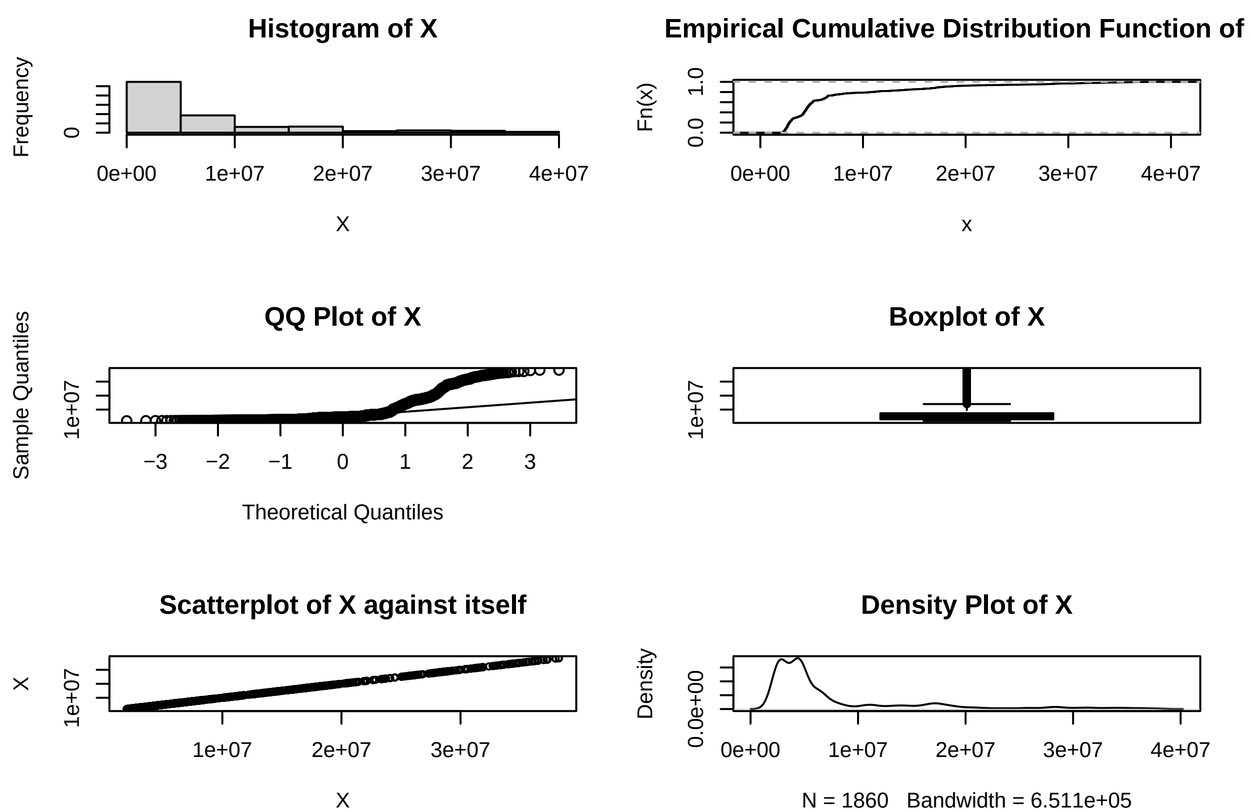
Problem 5:

Construct another set of six plots as in Figure 3.18, but this time applied to the data in EuStockMarkets. (i.e., apply the code of Figure 3.18 to Z, where $Z =$ EuStockMarkets). Comment on the results, and use the summary()} function to gain further insight. Which of the six plots are useful descriptors of this dataset, and which may be limited in their usefulness?

Answer:

```
# Now, apply the same set of plots to EuStockMarkets data
data("EuStockMarkets")

# Plotting Figure 3.18 with EuStockMarkets data
Z <- EuStockMarkets[, 1]  # Using the first column of EuStockMarkets data
X <- Z^2

par(mfrow = c(3, 2))  # Set up a 3x2 layout for plots

# Plot histogram
hist(X, main = "Histogram of X")

# Plot ECDF
plot(ecdf(X), main = "Empirical Cumulative Distribution Function of X")
```

```r
# Plot QQ plot
qqnorm(X, main = "QQ Plot of X")
qqline(X)

# Plot boxplot
boxplot(X, main = "Boxplot of X")

# Plot scatterplot matrix
plot(X, X, xlab = "X", ylab = "X", main = "Scatterplot of X against itself")

# Plot density plot
plot(density(X), main = "Density Plot of X")
```




```r
# Comment on the results and use summary() function to gain further insight
summary(X)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  1966557  3041894  4582019  7580367  7411286 38267709
```