CSG1132 A2 Notes

Martin Ponce

Wednesday, September 24, 2014

Intro

Theme 1

Research question

Is gender related to size of a user's Facebook network?

Thesis statement

Gender is related to the size of a user's Facebook network.

Theme 2

Research question

Is gender related to intensity of Facebook use?

Thesis statement

Gender is related to the intensity of Facebook use.

Variables

Gender and network size

- Gender
- Facebook friends
- Close friends
- Sociability

Gender and intensity of Facebook use

- Gender
- Facebook logins
- Facebook hours

Working with the data

To import SPSS data:

```
require(foreign);

fbDataset.raw <- read.spss(
    "./src/Ass2-dataset.sav",
    use.value.labels=TRUE,
    to.data.frame=TRUE
);
```

Raw data set consists of 61 observations, 10 variables.

```
summary(fbDataset.raw);
```

```
##       Age           Gender       FB_Logins         Hours
##  Min.   :17.0   Min.   :0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:19.0   1st Qu.:1.00   1st Qu.: 4.75   1st Qu.: 3.75
##  Median :20.0   Median :1.00   Median : 9.00   Median : 7.00
##  Mean   :20.8   Mean   :0.95   Mean   :12.45   Mean   :12.78
##  3rd Qu.:22.0   3rd Qu.:1.00   3rd Qu.:16.00   3rd Qu.:16.00
##  Max.   :29.0   Max.   :1.00   Max.   :59.00   Max.   :76.00
##  NA's   :12     NA's   :13     NA's   :13      NA's   :13
##    FB_friends  Close_friends  Sociability    Extraversion   Self_esteem
##  Min.   :  0   Min.   : 5    Min.   :2.00   Min.   : 5.0   Min.   : 9
##  1st Qu.:133   1st Qu.:13    1st Qu.:3.00   1st Qu.:10.0   1st Qu.:14
##  Median :258   Median :18    Median :4.00   Median :12.0   Median :18
##  Mean   :273   Mean   :23    Mean   :3.62   Mean   :13.1   Mean   :17
##  3rd Qu.:371   3rd Qu.:33    3rd Qu.:4.00   3rd Qu.:15.0   3rd Qu.:20
##  Max.   :798   Max.   :83    Max.   :5.00   Max.   :23.0   Max.   :23
##  NA's   :15    NA's   :12    NA's   :12     NA's   :12     NA's   :12
##  Social_anxiety
##  Min.   : 0.0
##  1st Qu.: 5.0
##  Median :10.0
##  Mean   :24.2
##  3rd Qu.:55.0
##  Max.   :80.0
##  NA's   :12
```

Data cleaning

Removing NA values

Created new object called fbDataset.rmNA with all observations with NA values removed:

```
fbDataset.rmNA <- na.omit(fbDataset.raw);
```

56 observations remain.

Removing Nil values

Removed observations that have 0 (Nil) as FB_Login:

```
fbDataset.rmNil <- subset(fbDataset.rmNA, FB_Logins > 0);
```

53 observations remain.

Removing Outliers

FB Logins Removing FB Login outliers > 50

```
fbDataset.rmOut <- subset(fbDataset.rmNil, FB_Logins < 50);
```

51 observations remain.

FB Hours Removing FB Hours outliers > 50

```
fbDataset.rmOut <- subset(fbDataset.rmOut, Hours < 50);
```

50 observations remain.

Close Friends Removing Close Friends outliers > 70

```
fbDataset.rmOut <- subset(fbDataset.rmOut, Close_friends < 70);
```

48 observations remain

Attach final dataset

```
# make new set called final, assign clean data to it
fbDataset.final <- fbDataset.rmOut

# attach final dataset so I can refer straight to the variable
attach(fbDataset.final)
```

Summary post data clense

```
summary(fbDataset.final);
```

```
##       Age           Gender        FB_Logins         Hours
##  Min.   :17.0   Min.   :0.000   Min.   : 1.00   Min.   : 2.00
##  1st Qu.:18.0   1st Qu.:1.000   1st Qu.: 5.75   1st Qu.: 3.75
##  Median :19.0   Median :1.000   Median : 9.00   Median : 7.00
##  Mean   :20.6   Mean   :0.938   Mean   :11.23   Mean   :10.52
##  3rd Qu.:22.0   3rd Qu.:1.000   3rd Qu.:16.00   3rd Qu.:15.25
```

```
##  Max.   :29.0   Max.   :1.000   Max.   :34.00   Max.   :31.00
##    FB_friends  Close_friends   Sociability    Extraversion   Self_esteem
##  Min.   : 33   Min.   : 6.0   Min.   :2.00   Min.   : 5.0   Min.   : 9.0
##  1st Qu.:163   1st Qu.:12.8   1st Qu.:3.00   1st Qu.:11.0   1st Qu.:14.0
##  Median :275   Median :19.0   Median :4.00   Median :13.0   Median :18.0
##  Mean   :291   Mean   :21.7   Mean   :3.67   Mean   :13.4   Mean   :17.1
##  3rd Qu.:376   3rd Qu.:26.5   3rd Qu.:4.00   3rd Qu.:16.2   3rd Qu.:20.2
##  Max.   :798   Max.   :53.0   Max.   :5.00   Max.   :23.0   Max.   :23.0
##  Social_anxiety
##  Min.   : 0.0
##  1st Qu.: 5.0
##  Median :11.5
##  Mean   :23.8
##  3rd Qu.:55.0
##  Max.   :70.0
```

Central tendency

Gender

```
mean(Gender);
```

```
## [1] 0.9375
```

```
median(Gender);
```

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

mvf() = mode

```
mfv(Gender);
```

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

```
sd(Gender);
```

```
## [1] 0.2446
```

```
skewness(Gender);
```

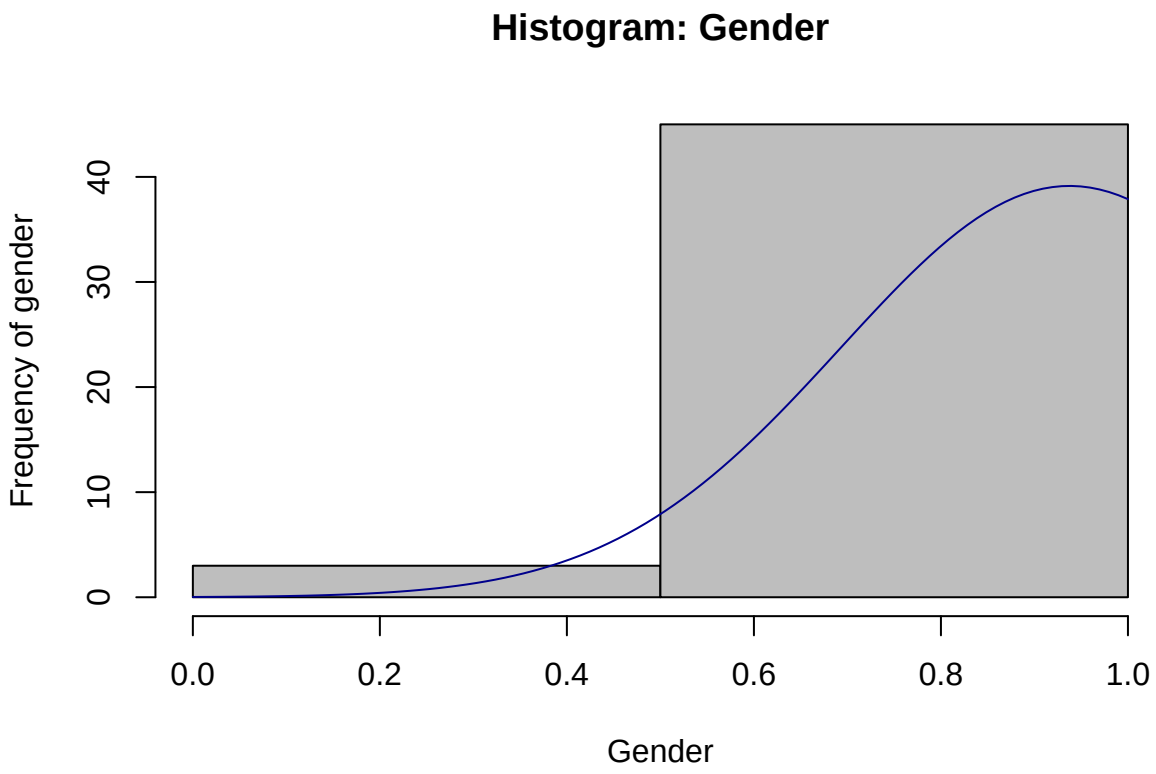
```
## [1] -3.502
## attr(,"method")
## [1] "moment"
```

```
kurtosis(Gender);
```

```
## [1] 10.49
```

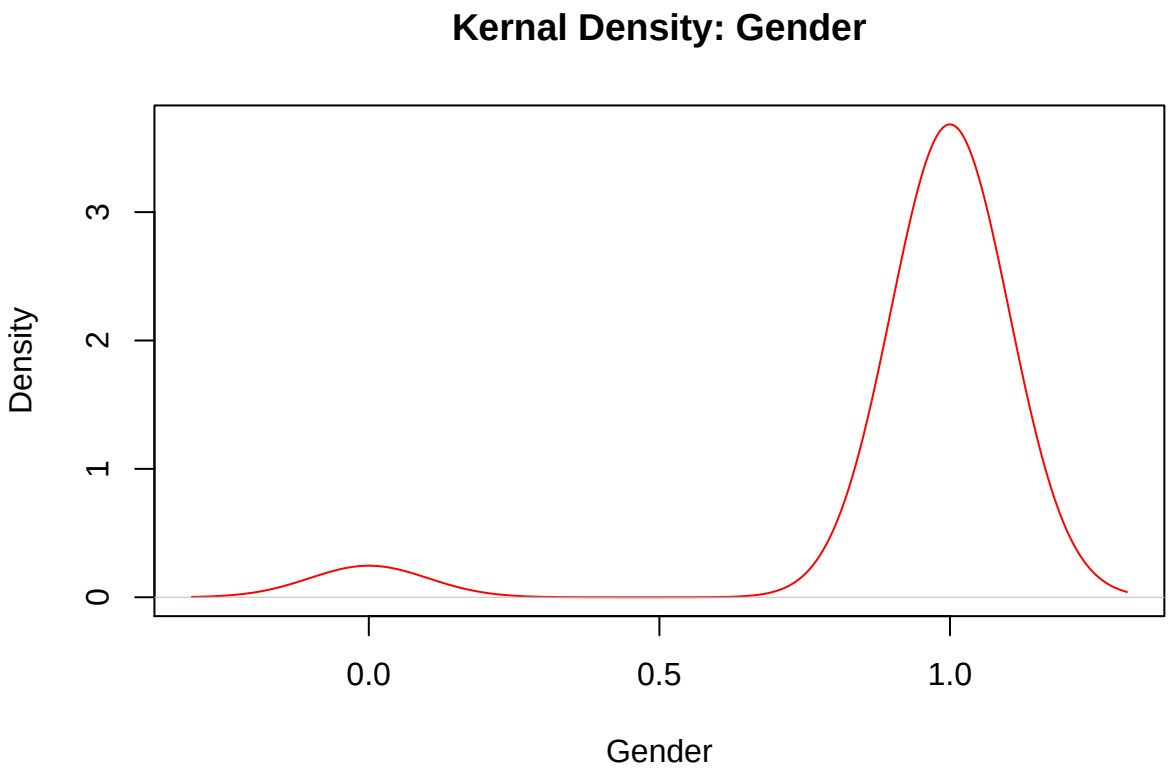
Histogram

```
h <- hist(
    Gender,
    breaks = 2,
    main = "Histogram: Gender",
    ylab = "Frequency of gender",
    xlab = "Gender",
    col = "grey"
);

# this draws the normal distribution curve over the histogram
xfit <- seq(
    min(Gender),
    max(Gender),
    length = 100
);

yfit <- dnorm(
    xfit,
    mean = mean(Gender), ,
    sd = sd(Gender)
);

yfit <- yfit * diff(h$mids[1:2]) * length(Gender);

lines(
    xfit,
    yfit,
    col = "darkblue"
);
```



Kernal density

```
density <- density(Gender);

plot(density,
    main = "Kernal Density: Gender",
    xlab = "Gender",
    ylab = "Density",
    col = "red"
);
```



FB Logins

```
mean(FB_Logins);
```

```
## [1] 11.23
```

```
median(FB_Logins);
```

```
## [1] 9
```

mvf() = mode

```
mfv(FB_Logins);
```

```
## [1] 6
```

```
sd(FB_Logins);
```

```
## [1] 8.933
```

```r
skewness(FB_Logins);
```

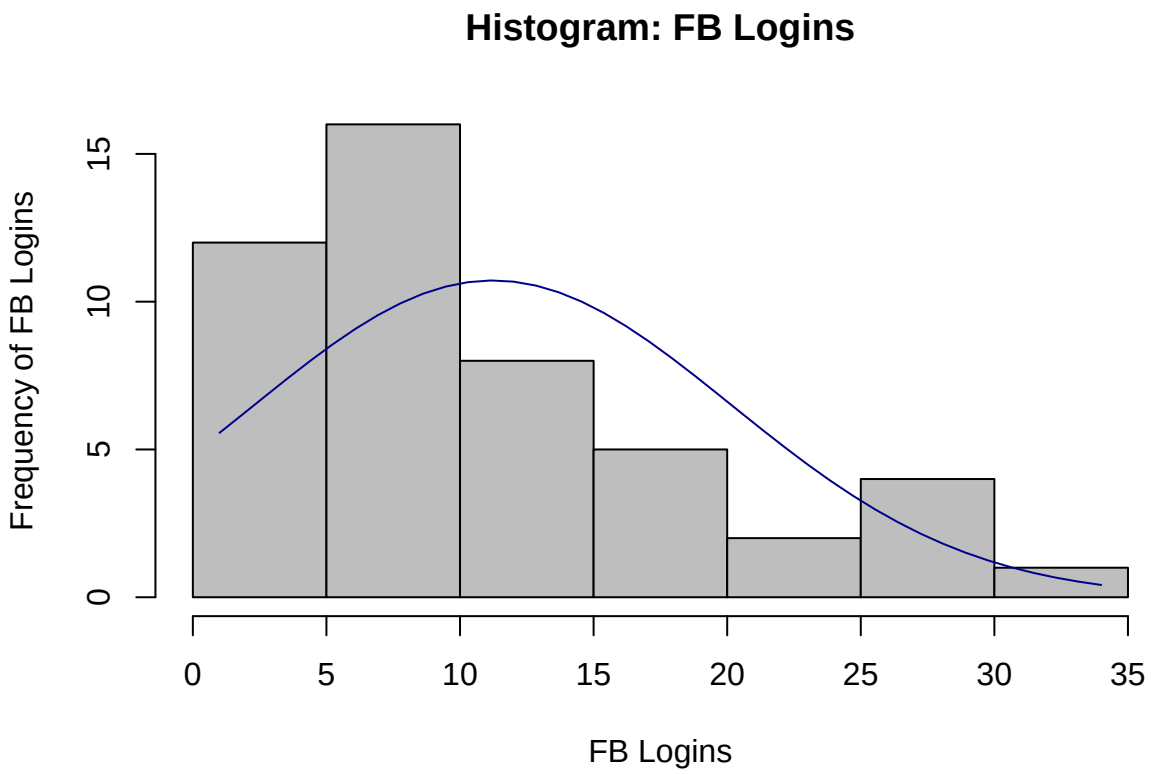
```
## [1] 0.8424
## attr(,"method")
## [1] "moment"
```

```r
kurtosis(FB_Logins);
```

```
## [1] -0.2736
```

Histogram

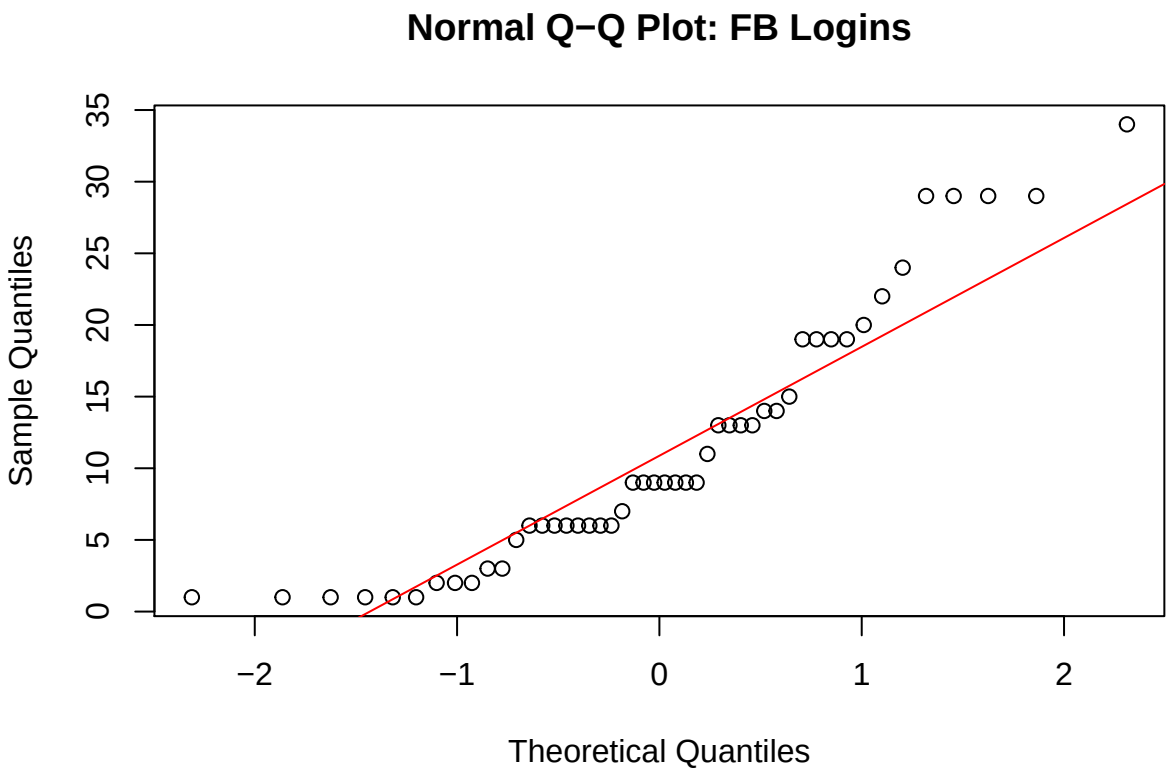
```r
h <- hist(
    FB_Logins,
    main = "Histogram: FB Logins",
    ylab = "Frequency of FB Logins",
    xlab = "FB Logins",
    col = "grey"
);

xfit <- seq(
    min(FB_Logins),
    max(FB_Logins),
    length=40
);

yfit <- dnorm(
    xfit,
    mean=mean(FB_Logins), ,
    sd=sd(FB_Logins)
);

yfit <- yfit * diff(h$mids[1:2]) * length(FB_Logins);

lines(
    xfit,
    yfit,
    col = "darkblue"
);
```



Q-Q plot

Identifies normality of distribution.

```
qqnorm(FB_Logins, main = "Normal Q-Q Plot: FB Logins");
qqline(FB_Logins, col = "red");
```



The majority of points do not fall on the expected normal distribution line. The distribution of **FB Logins** is not normal. Utilise non-parametric methods.

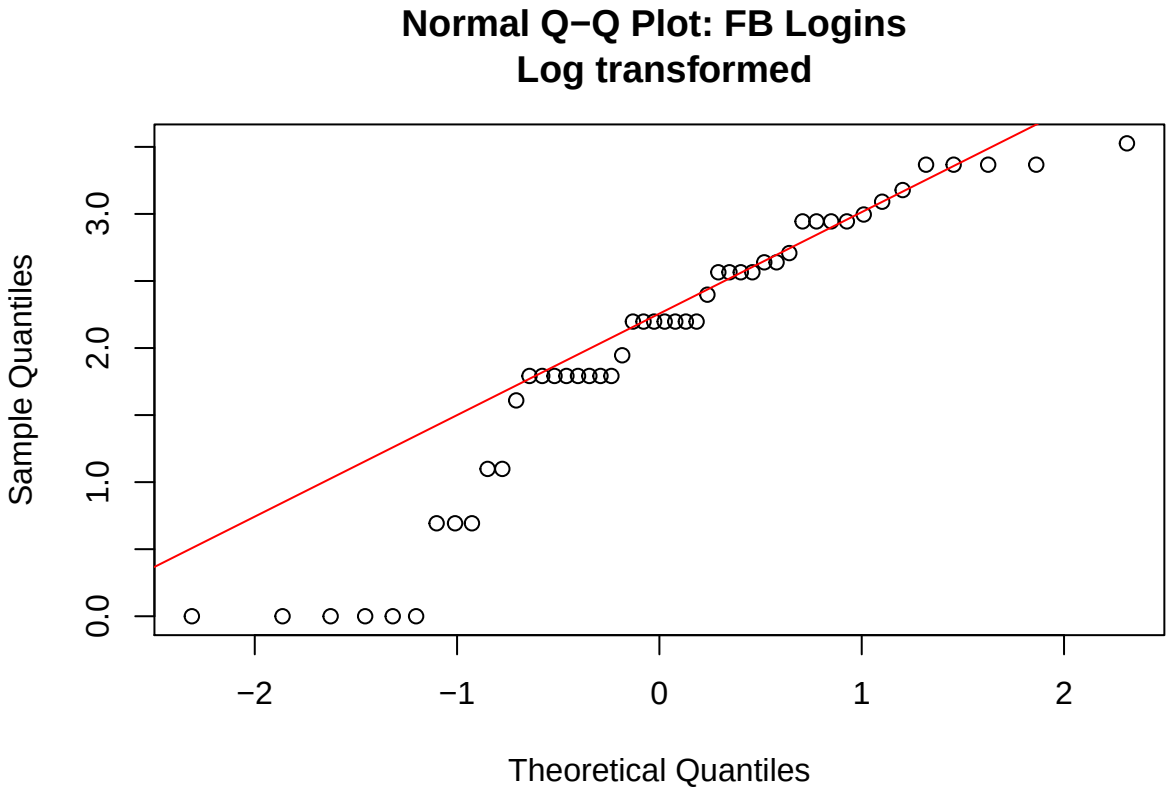
Log transformation

Log transformation may provide a normal distribution. Will test again for normality after transformation.

```
log.FB_Logins <- log(FB_Logins)
```

Testing for normality.

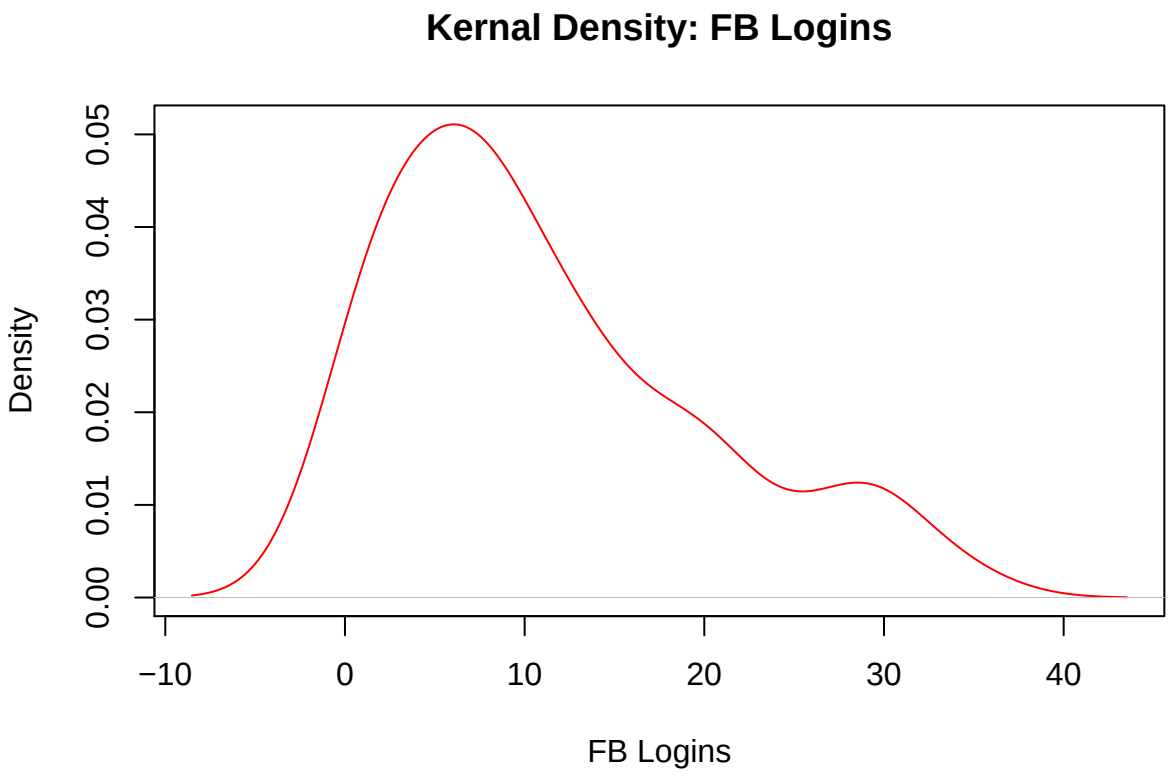
```
qqnorm(log.FB_Logins, main = "Normal Q-Q Plot: FB Logins \n Log transformed");
qqline(log.FB_Logins, col = "red");
```



Again, the majority of data points do not fall on the expected normal distribution line. Non-parametric methods still apply.

Kernal density

```
density <- density(FB_Logins);

plot(
    density,
    main = "Kernal Density: FB Logins",
    xlab = "FB Logins",
    ylab = "Density",
    col = "red"
);
```



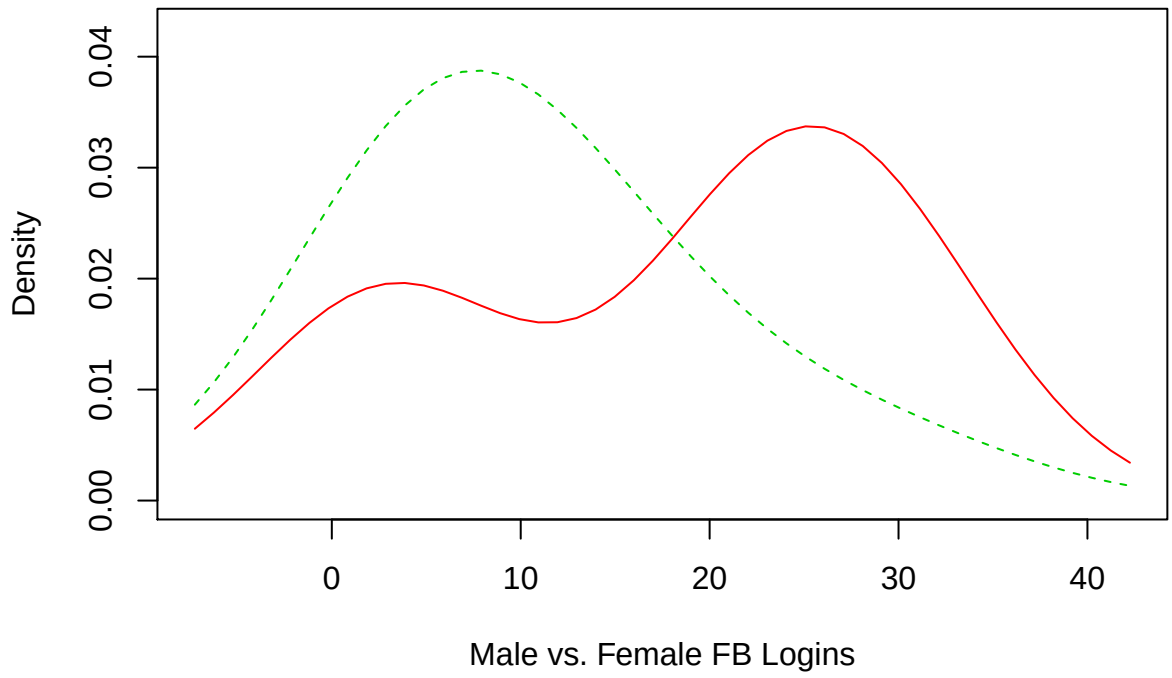
Note: Can't get legend to appear.

```
library(sm);
```

```
## Package 'sm', version 2.2-5.4: type help(sm) for summary information
```

```
#plot.new();

male_logins <- subset(fbDataset.final$FB_Logins, fbDataset.final$Gender == 0);
female_logins <- subset (fbDataset.final$FB_Logins, fbDataset.final$Gender == 1);

logins.f <- factor(
    fbDataset.final,
    levels = c(0, 1),
    labels = c("Female", "Male")
);

# the comparison plot

sm.density.compare(
    fbDataset.final$FB_Logins,
    fbDataset.final$Gender,
    xlab = "Male vs. Female FB Logins",
    main = "FB Login distribution by gender"
);
```

```
# legend

colfill <- c(
    2:(2 + length (levels(logins.f)))
);

legend(
    235,
    0.014,
    legend = c(
        "Female",
        "Male",
        fill = colfill
    )
);
```



FB Hours

```
mean(Hours);
```

```
## [1] 10.52
```

```r
median(Hours);
```

```
## [1] 7
```

mvf() = mode

```r
mfv(Hours);
```

```
## [1] 8
```

```r
sd(Hours);
```

```
## [1] 8.786
```

```r
skewness(Hours);
```

```
## [1] 1.013
## attr(,"method")
## [1] "moment"
```

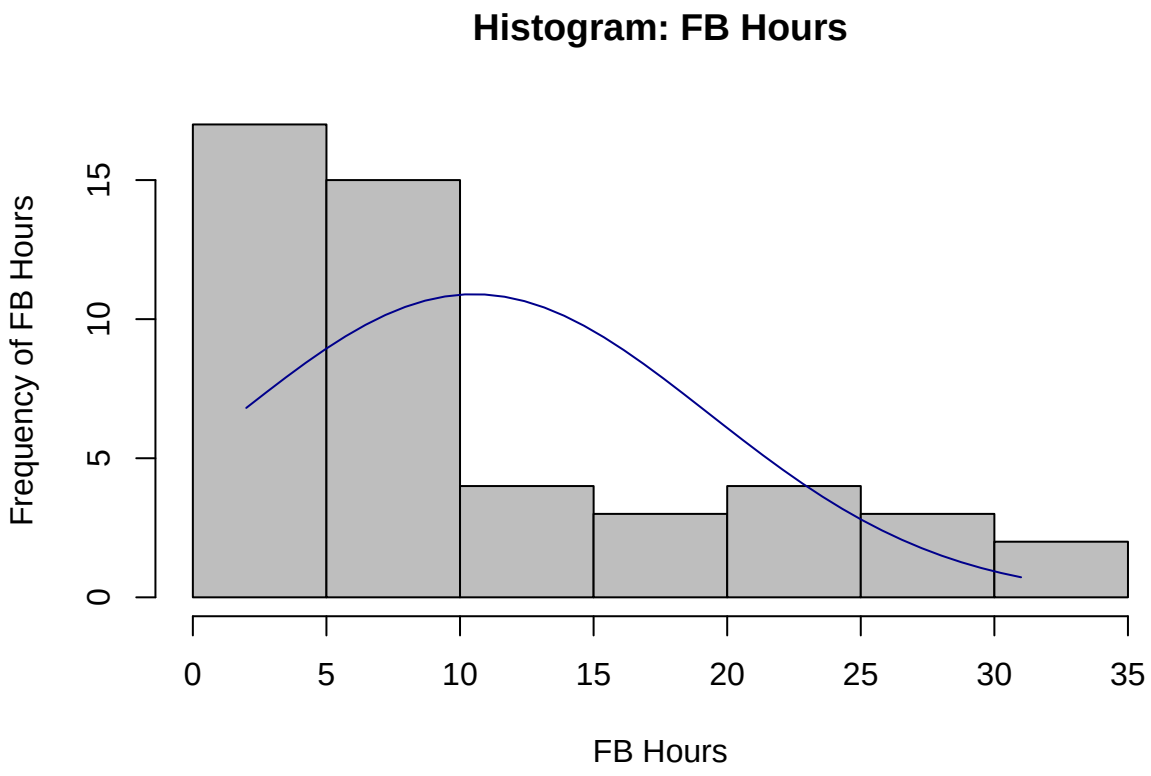
```r
kurtosis(Hours);
```

```
## [1] -0.316
```

Histogram

```r
h <- hist(
    Hours,
    main = "Histogram: FB Hours",
    ylab = "Frequency of FB Hours",
    xlab = "FB Hours",
    col = "grey"
);

xfit <- seq(
    min(Hours),
    max(Hours),
    length=40
);

yfit <- dnorm(
    xfit,
    mean = mean(Hours), ,
    sd = sd(Hours)
);

yfit <- yfit * diff(h$mids[1:2]) * length(Hours);
```

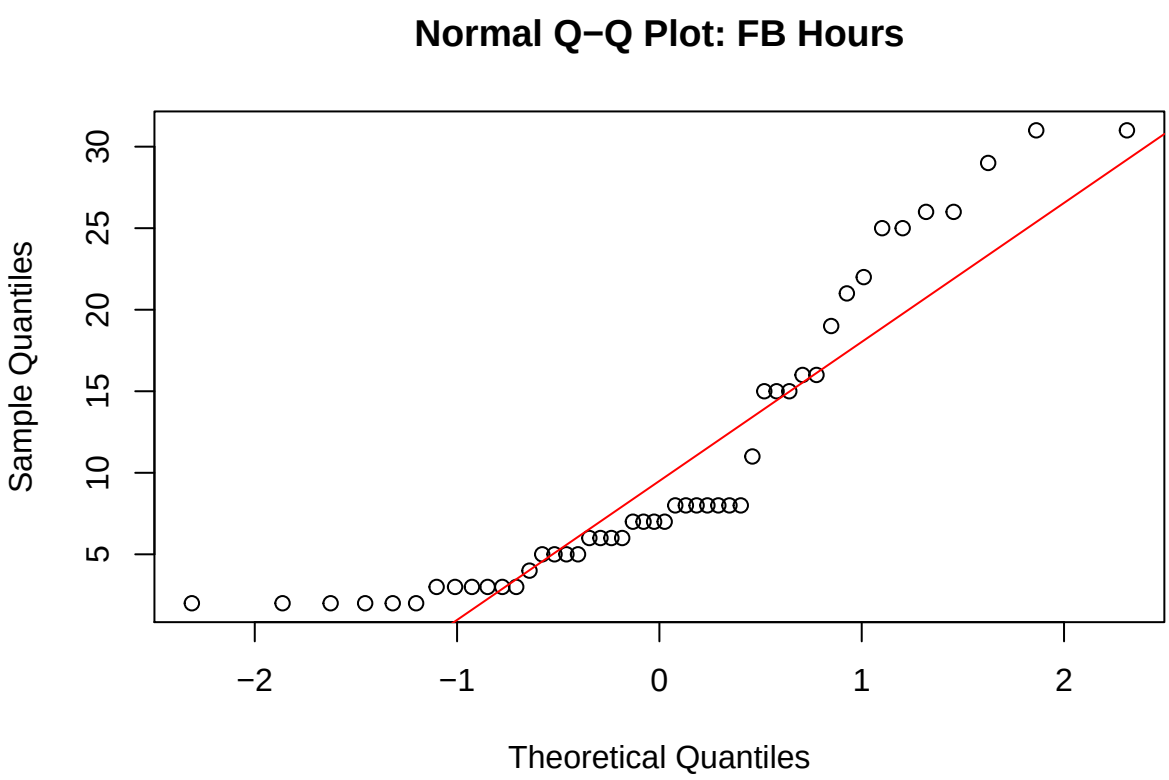
```
lines(
    xfit,
    yfit,
    col = "darkblue"
);
```



Q-Q plot

Identifies normality of distribution.

```
qqnorm(Hours, main = "Normal Q-Q Plot: FB Hours");
qqline(Hours, col = "red");
```



The majority of points do not fall on the expected normal distribution line. The distribution of **Hours** is not normal. Utilise non-parametric methods.

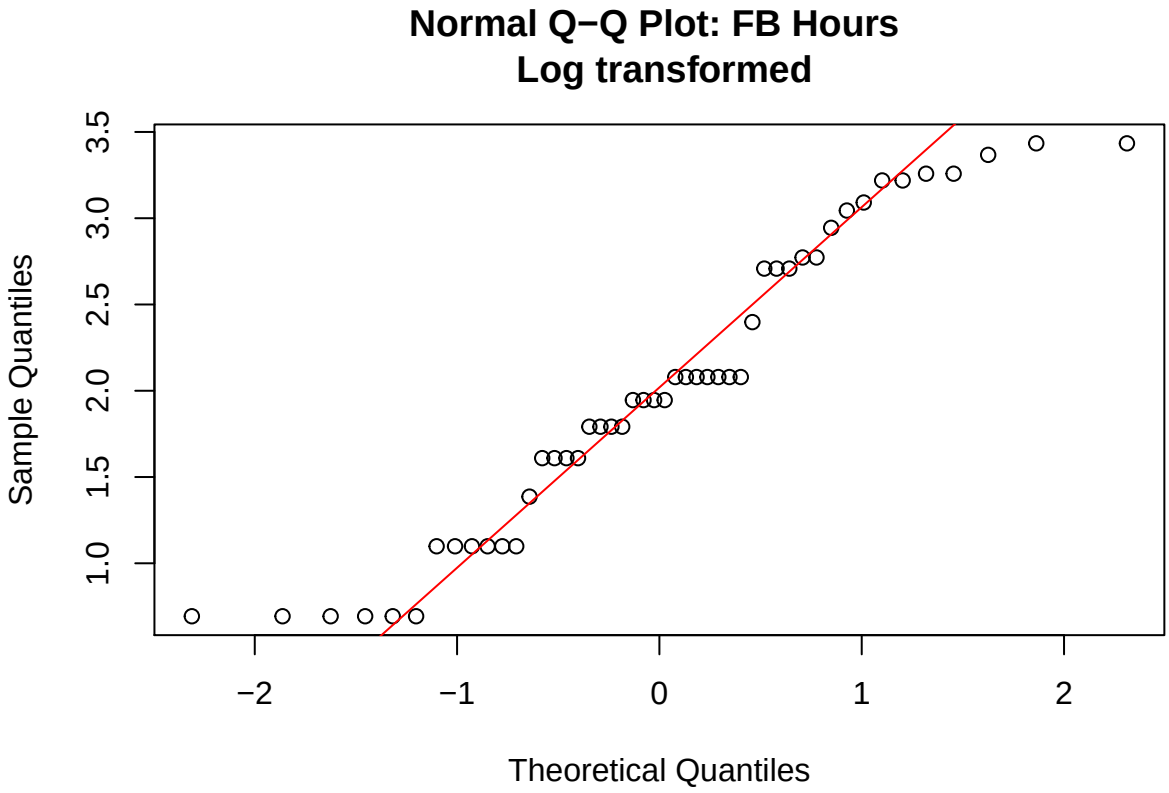
Log transformation

Log transformation may provide a normal distribution. Will test again for normality after transformation.

```
log.Hours <- log(Hours)
```

Testing for normality.

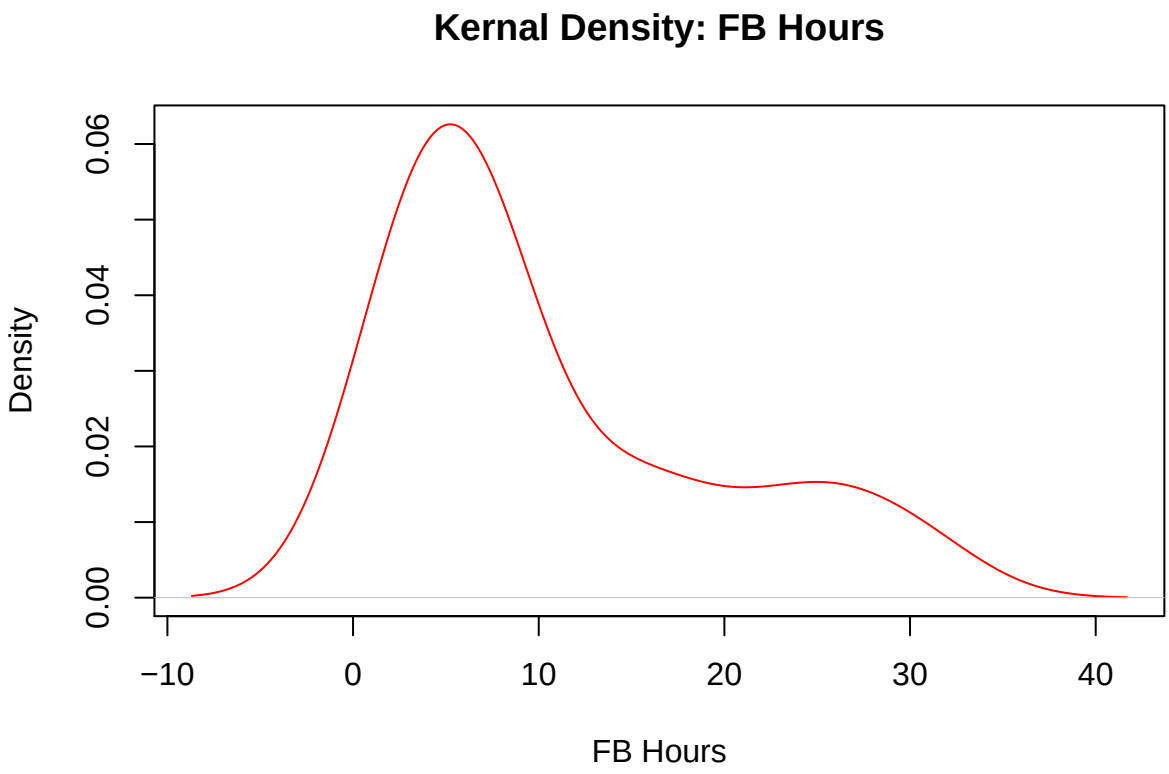
```
qqnorm(log.Hours, main = "Normal Q-Q Plot: FB Hours \n Log transformed");
qqline(log.Hours, col = "red");
```



Again, the majority of data points do not fall on the expected normal distribution line. Non-parametric methods still apply.

Kernal density

```
density <- density(Hours);

plot(density,
    main = "Kernal Density: FB Hours",
    xlab = "FB Hours",
    ylab = "Density",
    col = "red"
);
```



Facebook friends

```
mean(FB_friends);
```

```
## [1] 290.7
```

```
median(FB_friends);
```

```
## [1] 275
```

mvf() = mode

```
mfv(FB_friends);
```

```
## [1] 186 298
```

```
sd(FB_friends);
```

```
## [1] 176
```

```r
skewness(FB_friends);
```

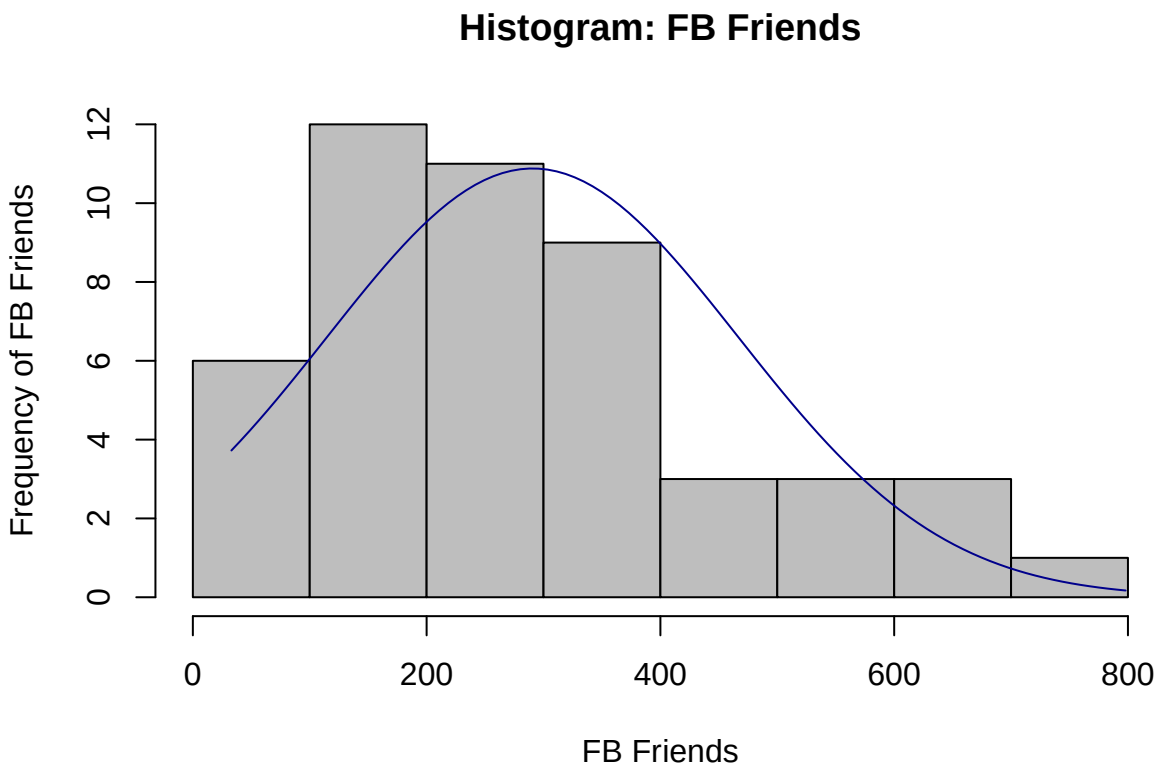
```
## [1] 0.7959
## attr(,"method")
## [1] "moment"
```

```r
kurtosis(FB_friends);
```

```
## [1] 0.04759
```

Histogram

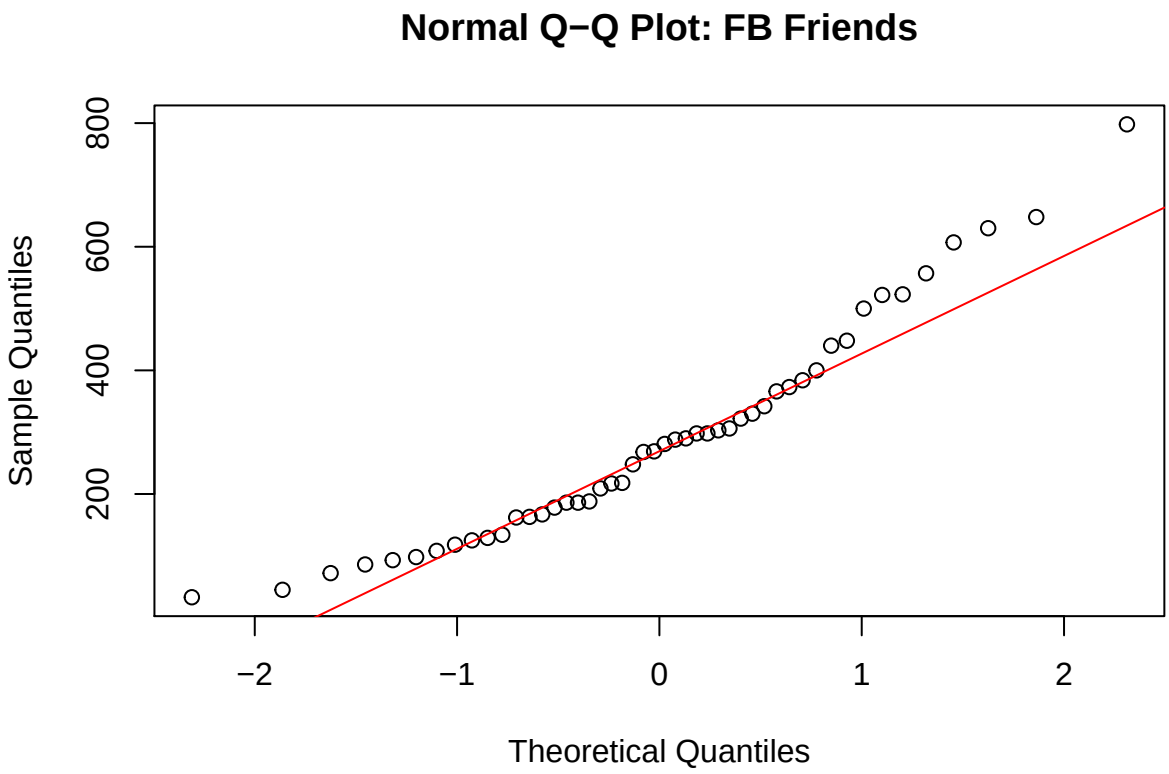
```r
h <- hist(
    FB_friends,
    main = "Histogram: FB Friends",
    ylab = "Frequency of FB Friends",
    xlab = "FB Friends",
    col = "grey"
);

xfit <- seq(
    min(FB_friends),
    max(FB_friends),
    length = 100
);

yfit <- dnorm(
    xfit,
    mean = mean(FB_friends), ,
    sd = sd(FB_friends)
);

yfit <- yfit * diff(h$mids[1:2]) * length(FB_friends);

lines(
    xfit,
    yfit,
    col = "darkblue"
);
```



Q-Q plot

Identifies normality of distribution.

```
qqnorm(FB_friends, main = "Normal Q-Q Plot: FB Friends");
qqline(FB_friends, col = "red");
```



The majority of points do not fall on the expected normal distribution line. The distribution of **FB Friends** is not normal. Utilise non-parametric methods.

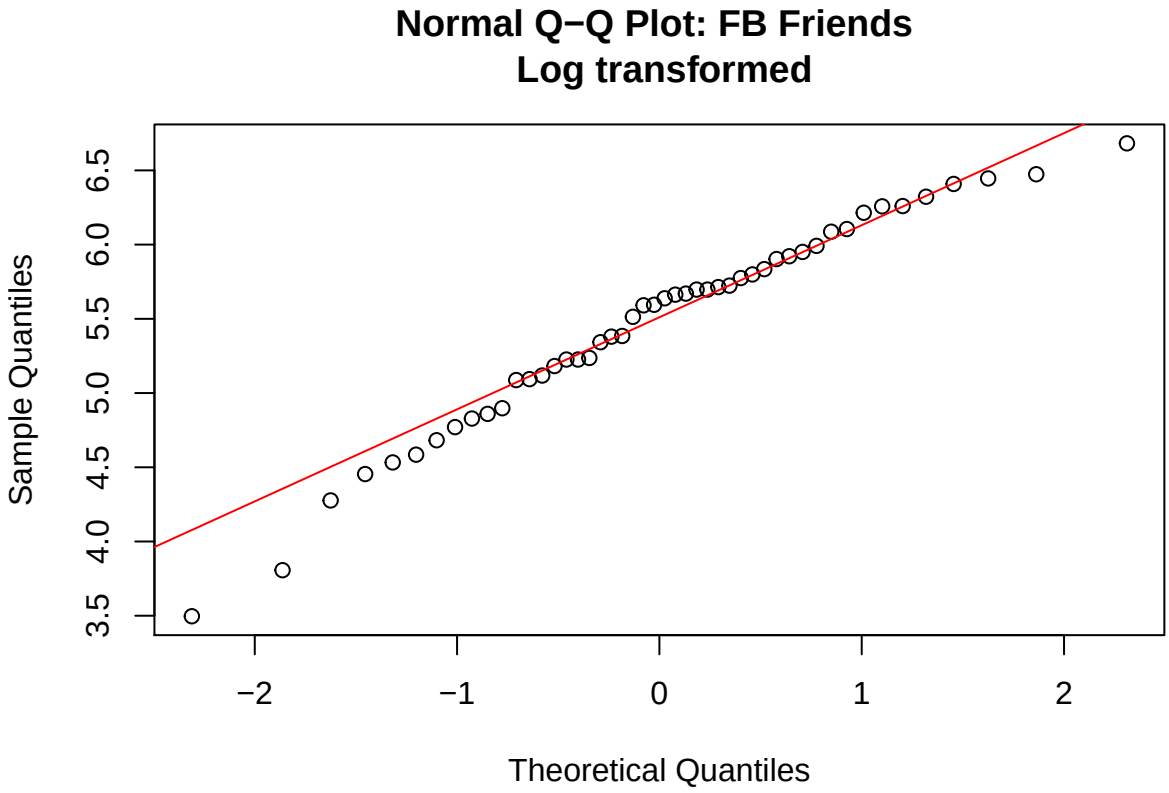
Log transformation

Log transformation may provide a normal distribution. Will test again for normality after transformation.

```
log.FB_friends <- log(FB_friends)
```

Testing for normality.

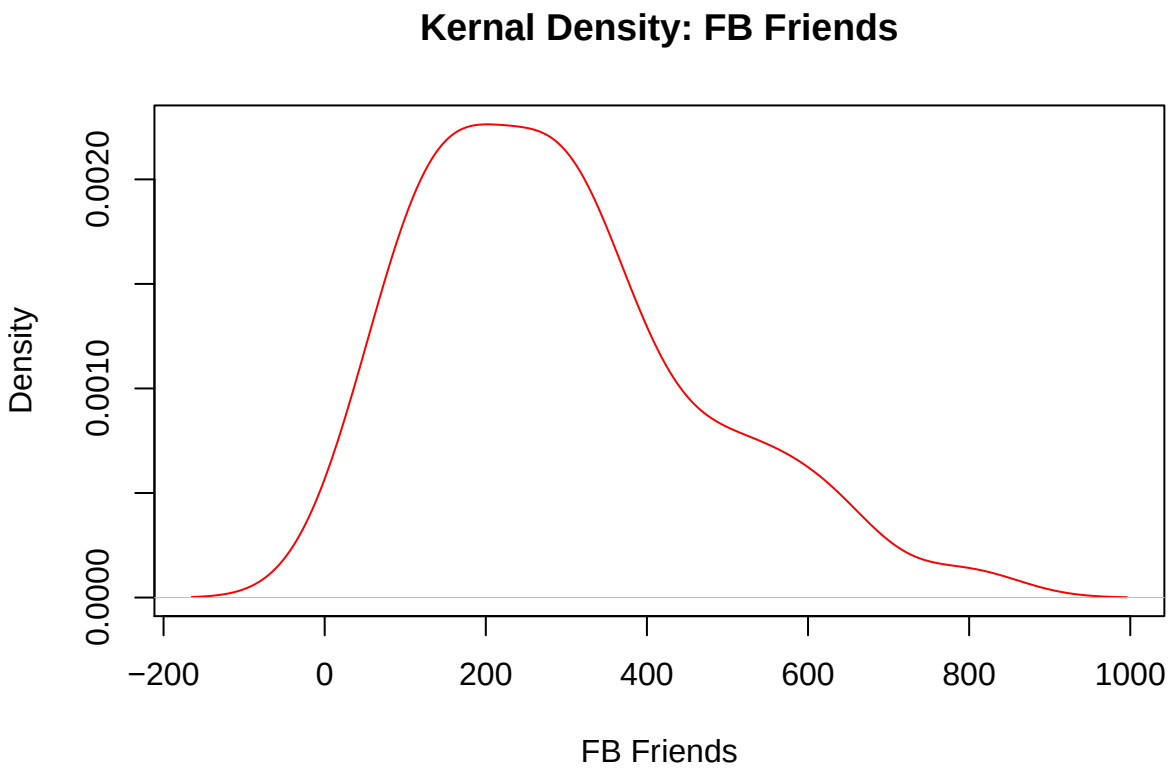
```
qqnorm(log.FB_friends, main = "Normal Q-Q Plot: FB Friends \n Log transformed");
qqline(log.FB_friends, col = "red");
```



While approaching normality, the majority of data points do not fall on the expected normal distribution line. Non-parametric methods still apply.

Kernal density

```
density <- density(FB_friends);

plot(density,
    main = "Kernal Density: FB Friends",
    xlab = "FB Friends",
    ylab = "Density",
    col = "red"
);
```



Close friends

```
mean(Close_friends)
```

```
## [1] 21.73
```

```
median(Close_friends)
```

```
## [1] 19
```

mvf() = mode

```
mfv(Close_friends)
```

```
## [1] 23
```

```
sd(Close_friends)
```

```
## [1] 12.29
```

```r
skewness(Close_friends)
```

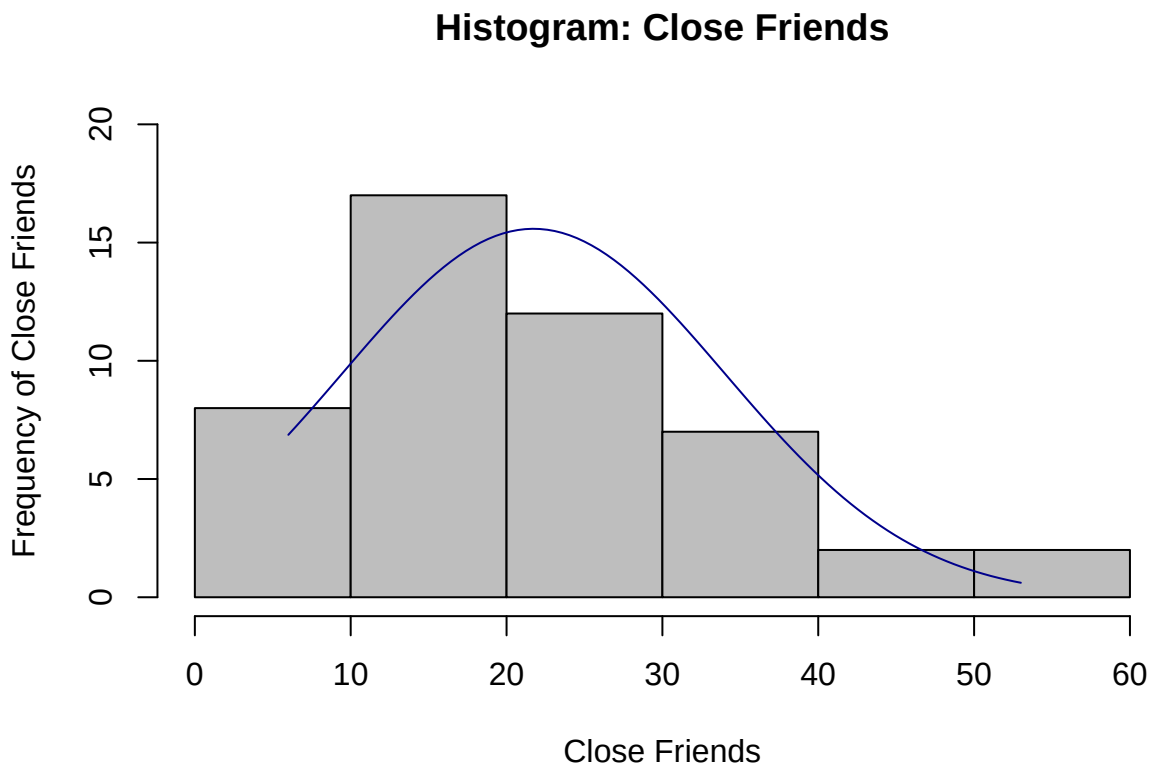
```
## [1] 0.9668
## attr(,"method")
## [1] "moment"
```

```r
kurtosis(Close_friends)
```

```
## [1] 0.1483
```

Histogram

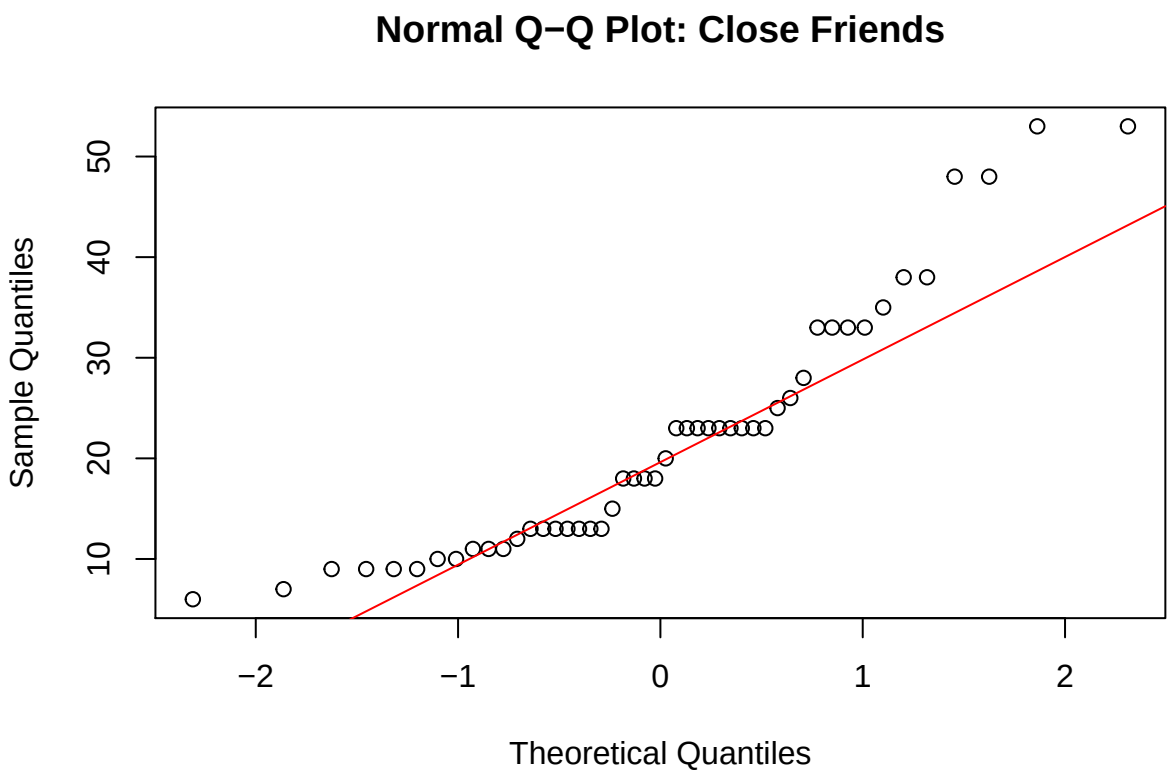
```r
h <- hist(
    Close_friends,
    main = "Histogram: Close Friends",
    ylab = "Frequency of Close Friends",
    xlab = "Close Friends",
    col = "grey",
    breaks = 5,

    # change the limit for x-axis, ylim for y-axis
    xlim = c(0, 60),
    ylim = c(0, 20),
);

xfit <- seq(
    min(Close_friends),
    max(Close_friends),
    length = 100
);

yfit <- dnorm(
    xfit,
    mean = mean(Close_friends), ,
    sd = sd(Close_friends)
);

yfit <- yfit * diff(h$mids[1:2]) * length(Close_friends);

lines(
    xfit,
    yfit,
    col = "darkblue"
);
```



Q-Q plot

Identifies normality of distribution.

```
qqnorm(Close_friends, main = "Normal Q-Q Plot: Close Friends");
qqline(Close_friends, col = "red");
```



The majority of points do not fall on the expected normal distribution line. The distribution of **Close friends** is not normal. Utilise non-parametric methods.

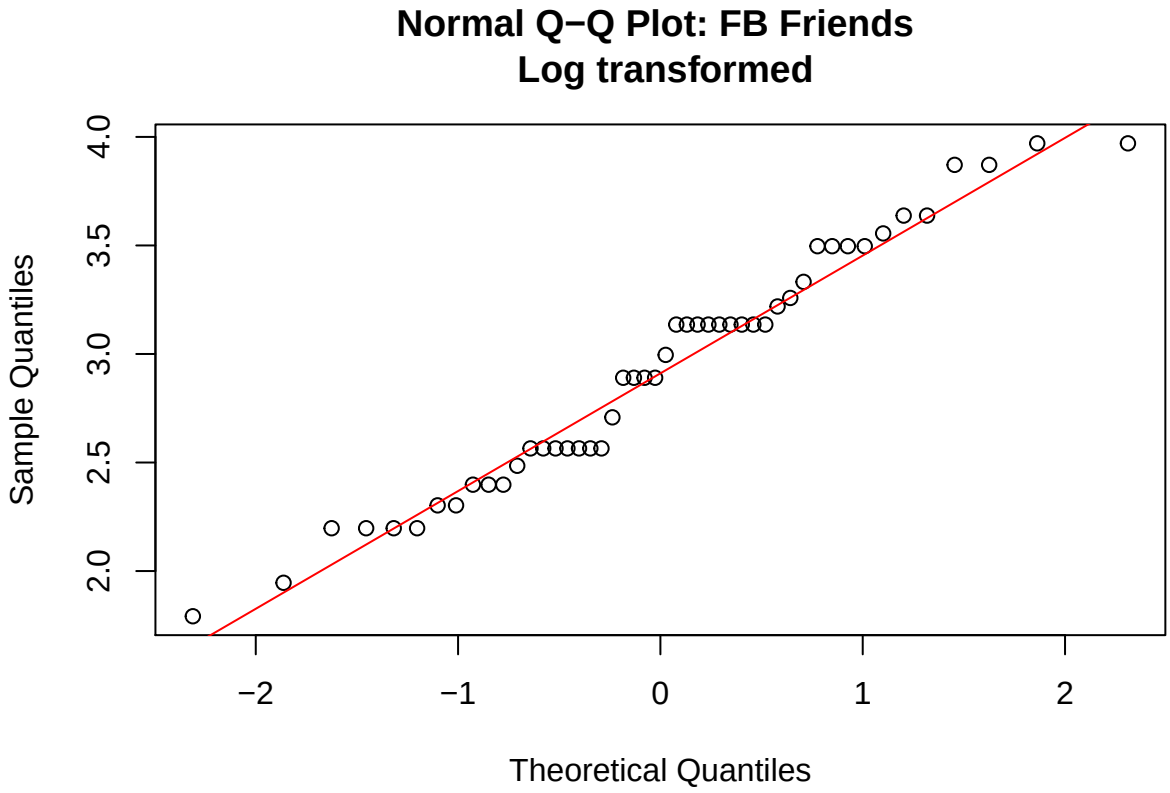
Log transformation

Log transformation may provide a normal distribution. Will test again for normality after transformation.

```
log.Close_friends <- log(Close_friends)
```

Testing for normality.

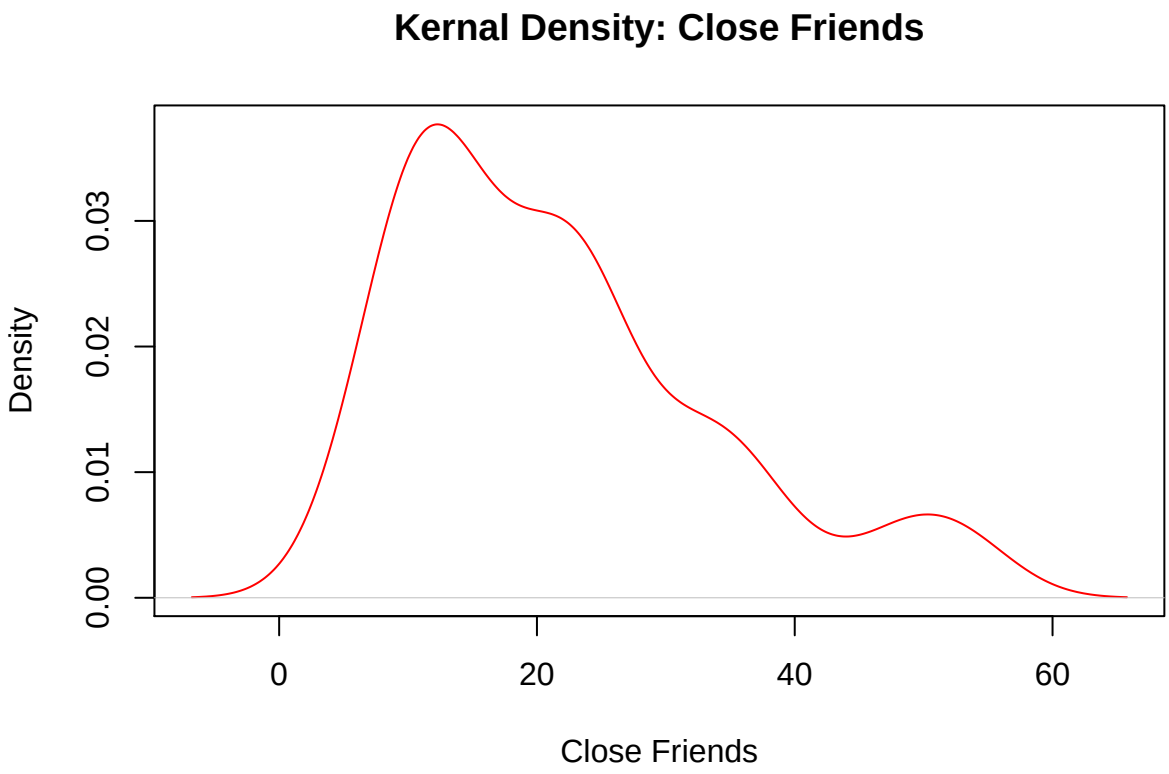
```
qqnorm(log.Close_friends, , main = "Normal Q-Q Plot: FB Friends \n Log transformed");
qqline(log.Close_friends, col = "red");
```



Again, the majority of data points do not fall on the expected normal distribution line. Non-parametric methods still apply.

Kernal density

```
density <- density(Close_friends);

plot(
    density,
    main = "Kernal Density: Close Friends",
    xlab = "Close Friends",
    ylab = "Density",
    col = "red"
);
```



Sociability

```
mean(Sociability)
```

```
## [1] 3.667
```

```
median(Sociability)
```

```
## [1] 4
```

mvf() = mode

```
mfv(Sociability)
```

```
## [1] 4
```

```
sd(Sociability)
```

```
## [1] 0.7532
```

```r
skewness(Sociability)
```

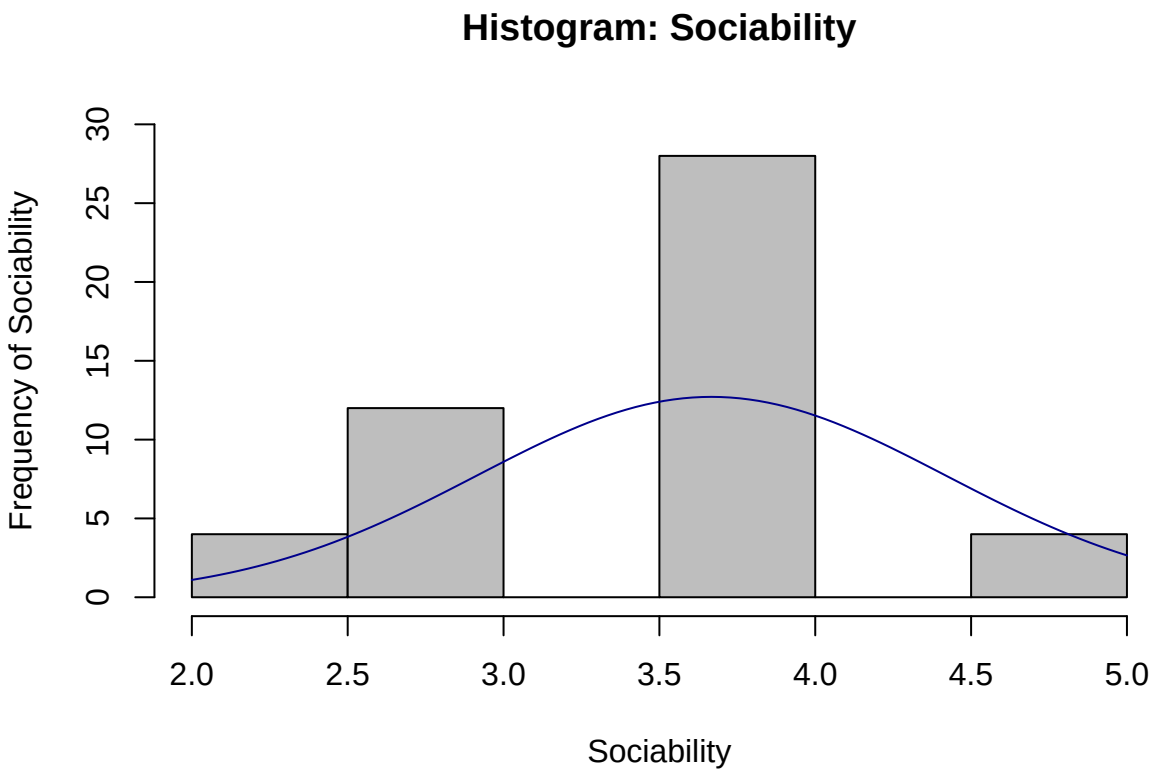
```
## [1] -0.5633
## attr(,"method")
## [1] "moment"
```

```r
kurtosis(Sociability)
```

```
## [1] -0.008646
```

Histogram

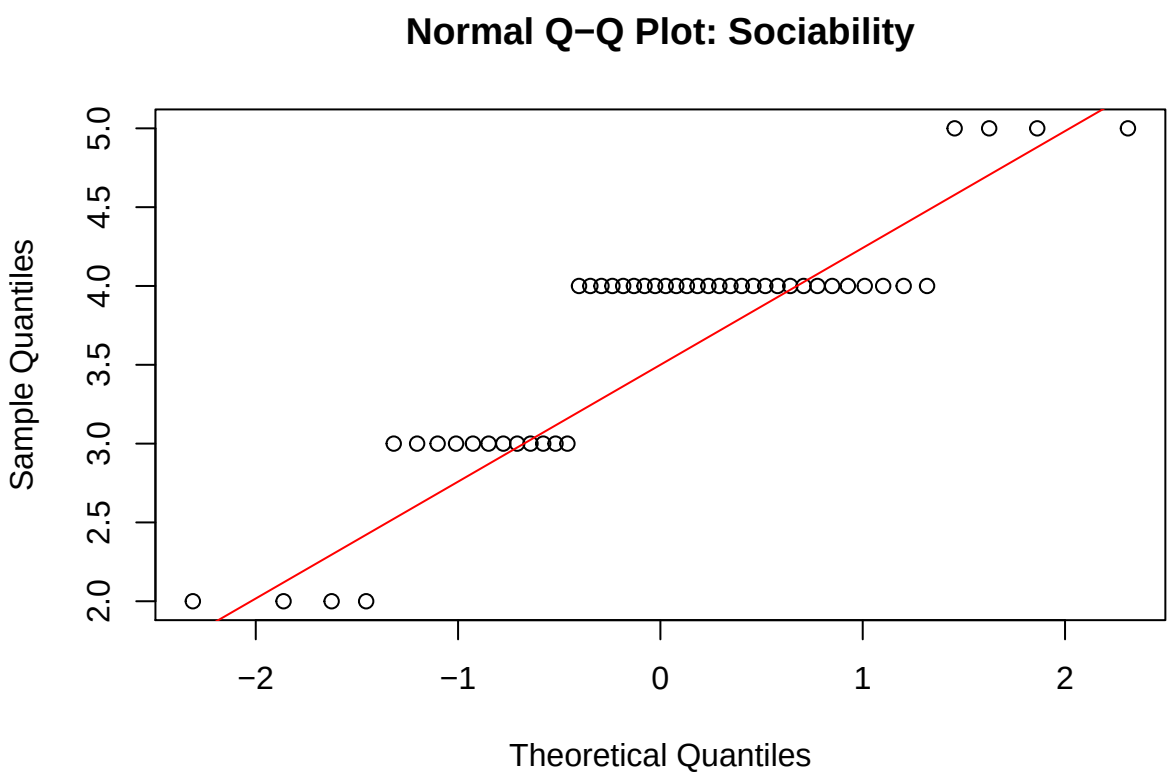
```r
h <- hist(
    Sociability,
    main = "Histogram: Sociability",
    ylab = "Frequency of Sociability",
    xlab = "Sociability",
    col = "grey",
    ylim = c(0, 30)
);

xfit <- seq(
    min(Sociability),
    max(Sociability),
    length = 100
);

yfit <- dnorm(
    xfit,
    mean = mean(Sociability), ,
    sd = sd(Sociability)
);

yfit <- yfit * diff(h$mids[1:2]) * length(Sociability);

lines(
    xfit,
    yfit,
    col = "darkblue"
);
```



Q-Q plot

Identifies normality of distribution.

```
qqnorm(Sociability, main = "Normal Q-Q Plot: Sociability");
qqline(Sociability, col = "red");
```



The majority of points do not fall on the expected normal distribution line. The distribution of **Sociability** is not normal. Utilise non-parametric methods.

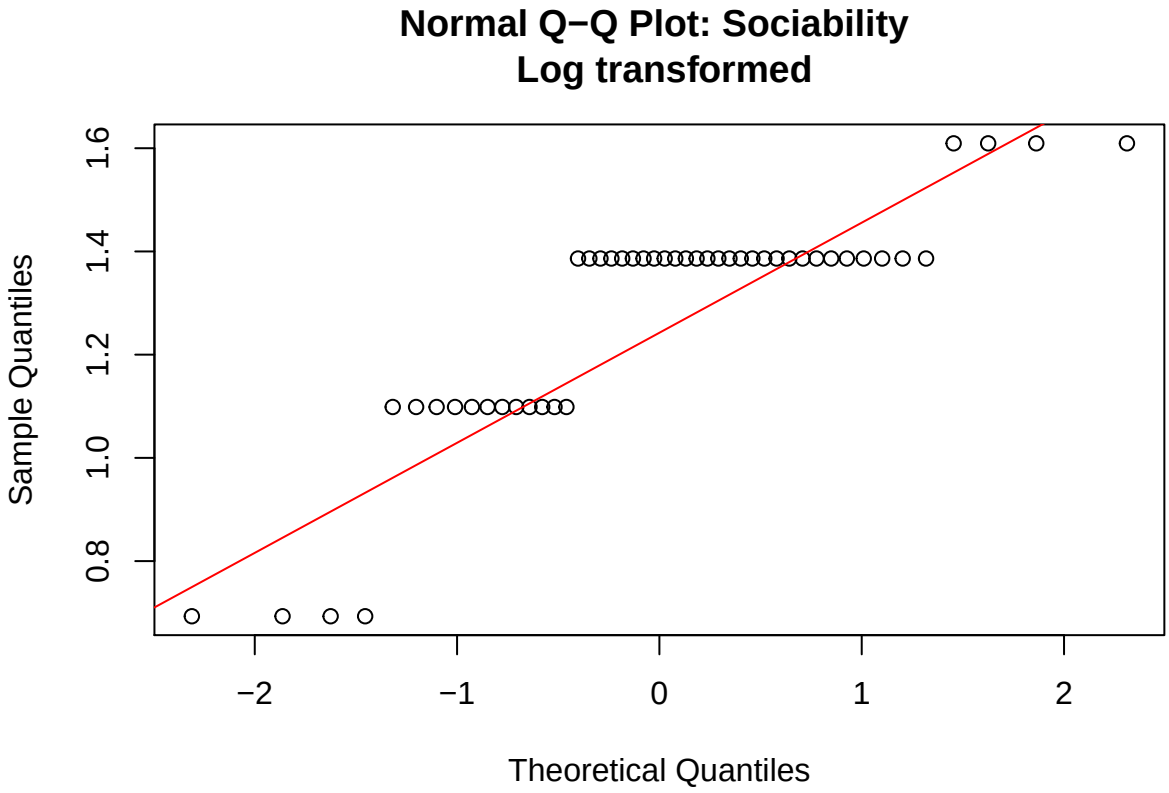
Log transformation

Log transformation may provide a normal distribution. Will test again for normality after transformation.

```
log.Sociability <- log(Sociability)
```

Testing for normality.

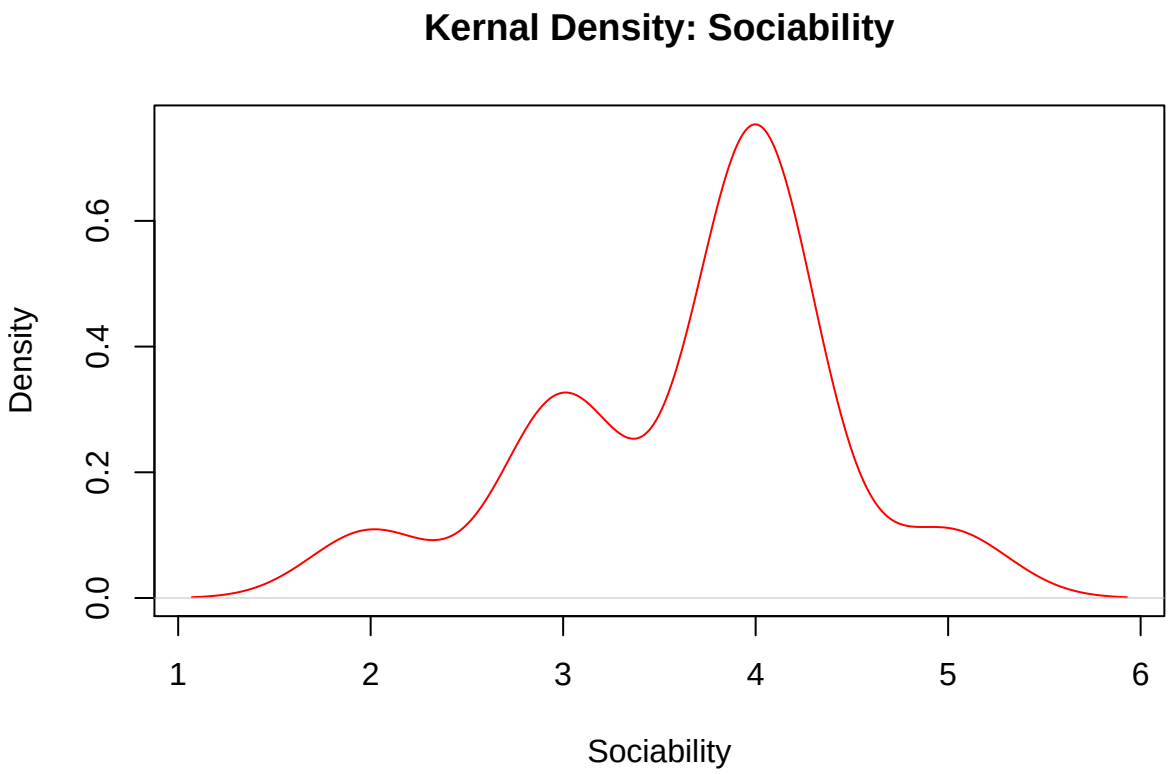
```
qqnorm(log.Sociability, , main = "Normal Q-Q Plot: Sociability \n Log transformed");
qqline(log.Sociability, col = "red");
```



Again, the majority of data points do not fall on the expected normal distribution line. Non-parametric methods still apply.

Kernal density

```
density <- density(Sociability);

plot(
    density,
    main = "Kernal Density: Sociability",
    xlab = "Sociability",
    ylab = "Density",
    col = "red"
);
```



Testing

Non-parametric tests chosen because the variables are non-normally distributed.

Tests chosen

- Mann-Whitney-Wilcoxon
  - See r-tutor.com
- Spearman's rho
  - See r-bloggers.com

Theme 1: Gender is related to network size

Facebook friends and gender

```
wilcox.test(FB_friends ~ Gender);
```

Mann-Whitney-Wilcoxon test

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  FB_friends by Gender
## W = 106, p-value = 0.1055
## alternative hypothesis: true location shift is not equal to 0
```

```
cor.test(Gender, FB_friends, method = "spearman");
```

Spearman's rho test

```
##
##  Spearman's rank correlation rho
##
## data:  Gender and FB_friends
## S = 22831, p-value = 0.1015
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##     rho
## -0.2392
```

Close friends and gender

```
wilcox.test(Close_friends ~ Gender);
```

Mann-Whitney-Wilcoxon test

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Close_friends by Gender
## W = 80.5, p-value = 0.5923
## alternative hypothesis: true location shift is not equal to 0
```

```
cor.test(Gender, Close_friends, method = "spearman");
```

Spearman's rho test

```
##
##  Spearman's rank correlation rho
##
## data:  Gender and Close_friends
## S = 19921, p-value = 0.5831
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## -0.08124
```

Sociability and gender

```
wilcox.test(Sociability ~ Gender);
```

Mann-Whitney-Wilcoxon test

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Sociability by Gender
## W = 73.5, p-value = 0.7915
## alternative hypothesis: true location shift is not equal to 0
```

```
cor.test(Sociability, Gender, method = "spearman");
```

Spearman's rho test

```
##
##  Spearman's rank correlation rho
##
## data:  Sociability and Gender
## S = 19199, p-value = 0.7765
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## -0.04207
```