MATH1324 Assignment 1

Modeling Body Measurements

Student Details

Student Name (s3795594, Vignesh Gopalakrishnan)

Problem Statement

Write a clear and concise problem statement that guides your investigation. Explain which variable you choose and outline the approaches taken for normal distribution fitting.

The aim of the study is to study the body measurements database, tidy the data, take a variable and examine the summary statistics and perform different empirical and visual tests for checking the normality of the variable for males and females seperately. After analysing different variables, I chose to concentrate on Chest Diameter (che.di) because the males and females the spread differ for males and females. I have used Shapiro-Wilk's test and Jarque-Bera's test for testing normality empirically and Density plot, Q-Q plot and histogram overlapped with the normal curve for testing the normality visually.

Load Packages

```
library("readr")
library("magrittr")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("ggplot2")
library("ggpubr")
library("tseries")
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

Data

Import the body measurements data and prepare it for analysis. Show your code.

```
body_dim <- read_csv("G:/Intro to Stati/bdims.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
body_dim
```

bia.di	bii.di	bit.di	che.de	che.di	elb.di	wri.di	kne.di	ank.di	sho.gi
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
42.9	26.0	31.5	17.7	28.0	13.1	10.4	18.8	14.1	106.2
43.7	28.5	33.5	16.9	30.8	14.0	11.8	20.6	15.1	110.5
40.1	28.2	33.3	20.9	31.7	13.9	10.9	19.7	14.1	115.1
44.3	29.9	34.0	18.4	28.2	13.9	11.2	20.9	15.0	104.5
42.5	29.9	34.0	21.5	29.4	15.2	11.6	20.7	14.9	107.5

bia.di <dbl>	bii.di <dbl>	bit.di <dbl>	che.de <dbl>	che.di <dbl>	elb.di <dbl>	wri.di <dbl>	kne.di <dbl>	ank.di <dbl>	sho.gi <dbl>
43.3	27.0	31.5	19.6	31.3	14.0	11.5	18.8	13.9	119.8
43.5	30.0	34.0	21.9	31.7	16.1	12.5	20.8	15.6	123.5
44.4	29.8	33.2	21.8	28.8	15.1	11.9	21.0	14.6	120.4
43.5	26.5	32.1	15.5	27.5	14.1	11.2	18.9	13.2	111.0
42.0	28.0	34.0	22.5	28.0	15.6	12.0	21.1	15.0	119.5

1-10 of 507 rows | 1-10 of 25 columns Previous **1** 2 3 4 5 6 ... 51 Next

```
body_dim$sex <- factor(body_dim$sex,levels = c(0,1), labels= c("Female", "Male")) #define the variable sex as a factor and d
efining labels for it
body_dim_male <- body_dim %>% filter(sex == "Male") #filtering out the male observations
body_dim_female <- body_dim %>% filter(sex == "Female") #filtering out the male observations
body_dim_male
```

bia.di <dbl>	bii.di <dbl>	bit.di <dbl>	che.de <dbl>	che.di <dbl>	elb.di <dbl>	wri.di <dbl>	kne.di <dbl>	ank.di <dbl>	sho.gi <dbl>
42.9	26.0	31.5	17.7	28.0	13.1	10.4	18.8	14.1	106.2
43.7	28.5	33.5	16.9	30.8	14.0	11.8	20.6	15.1	110.5
40.1	28.2	33.3	20.9	31.7	13.9	10.9	19.7	14.1	115.1
44.3	29.9	34.0	18.4	28.2	13.9	11.2	20.9	15.0	104.5
42.5	29.9	34.0	21.5	29.4	15.2	11.6	20.7	14.9	107.5
43.3	27.0	31.5	19.6	31.3	14.0	11.5	18.8	13.9	119.8
43.5	30.0	34.0	21.9	31.7	16.1	12.5	20.8	15.6	123.5
44.4	29.8	33.2	21.8	28.8	15.1	11.9	21.0	14.6	120.4
43.5	26.5	32.1	15.5	27.5	14.1	11.2	18.9	13.2	111.0
42.0	28.0	34.0	22.5	28.0	15.6	12.0	21.1	15.0	119.5

1-10 of 247 rows | 1-10 of 25 columns Previous **1** 2 3 4 5 6 ... 25 Next

```
body_dim_female
```

bia.di <dbl>	bii.di <dbl>	bit.di <dbl>	che.de <dbl>	che.di <dbl>	elb.di <dbl>	wri.di <dbl>	kne.di <dbl>	ank.di <dbl>	sho.gi <dbl>
37.6	25.0	31.3	16.2	24.9	11.2	9.2	17.0	12.3	95.0
36.7	26.4	31.0	16.8	24.5	12.1	9.9	19.3	12.8	99.5
34.8	25.9	30.2	16.4	24.2	11.3	8.9	17.0	12.2	88.0
36.6	27.9	31.8	19.3	24.9	12.3	9.5	18.6	13.0	97.0
35.5	28.2	31.0	18.2	26.2	11.5	9.1	17.2	12.4	103.3
37.0	28.0	32.0	15.1	25.7	12.5	10.0	17.2	13.2	93.5
35.5	26.5	29.2	15.4	24.5	12.3	9.4	17.2	12.0	93.3
37.4	30.2	33.2	18.8	26.6	13.3	10.7	19.8	13.8	94.5
37.8	29.0	32.6	18.6	25.0	12.1	9.8	17.8	12.7	98.6
38.6	28.8	33.2	19.7	29.4	13.4	11.5	20.9	13.2	115.5

1-10 of 260 rows | 1-10 of 25 columns Previous **1** 2 3 4 5 6 ... 26 Next

Summary Statistics

Calculate descriptive statistics (i.e., mean, median, standard deviation, first and third quartile, interquartile range, minimum and maximum values) of the selected measurement grouped by sex.

```
#Male Observations
body_dim_male$che.di %>% summary() #produce mean, median first,third quartile, minimum and maximum values
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.70   28.65   29.90   29.95   31.40   35.60
```

```
body_dim_male$che.di %>% sd() #standard deviation
```

```
## [1] 2.083108
```

```
body_dim_male$che.di %>% IQR() #interquartile range
```

```
## [1] 2.75
```

```
#Female Observations
body_dim_female$che.di %>% summary() #produce mean, median first,third quartile, minimum and maximum values
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    22.2    24.9    25.9    26.1    27.1    33.2
```

```
body_dim_female$che.di %>% sd() #standard deviation
```

```
## [1] 1.818808
```

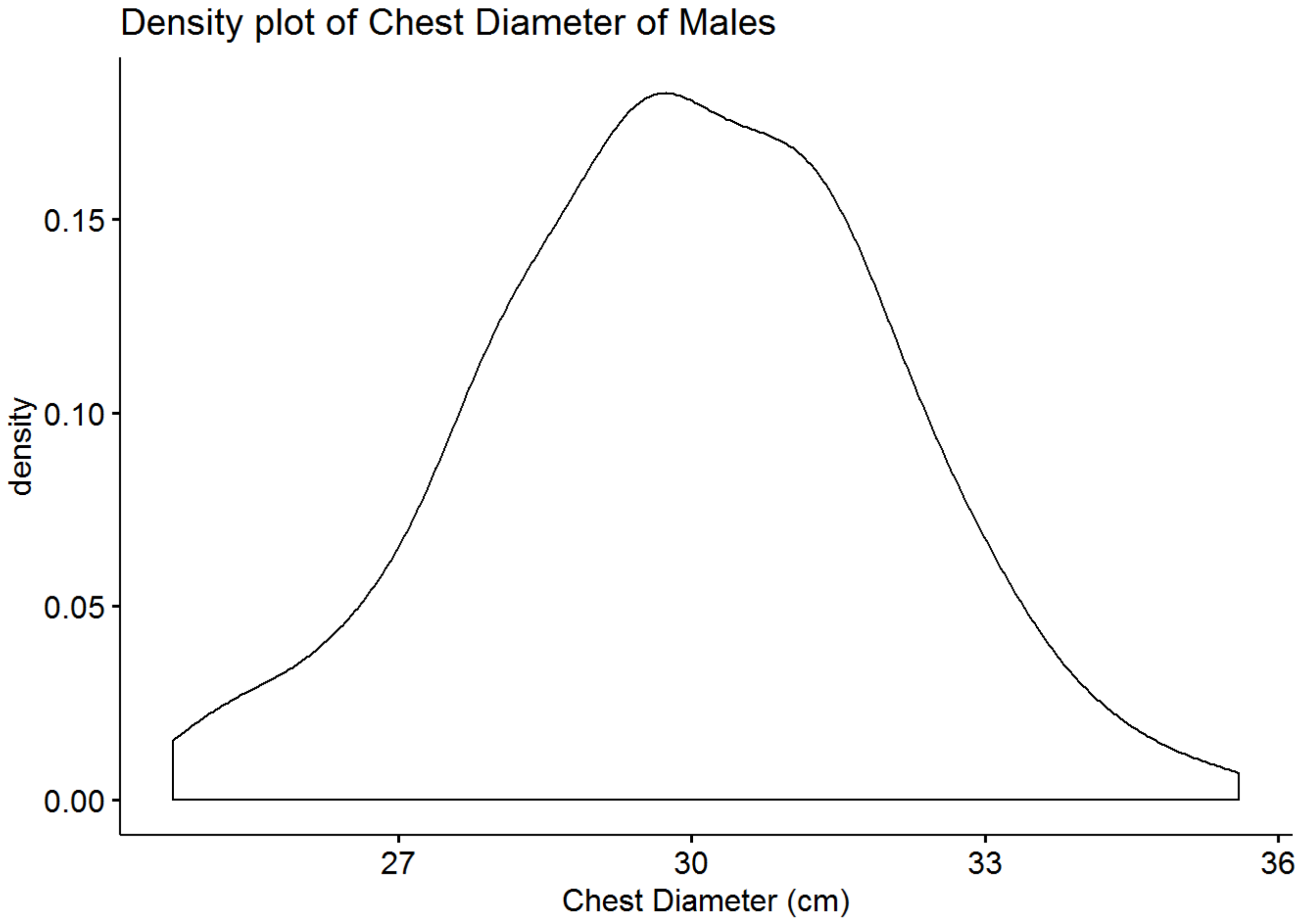
```
body_dim_female$che.di %>% IQR() #interquartile range
```

```
## [1] 2.2
```

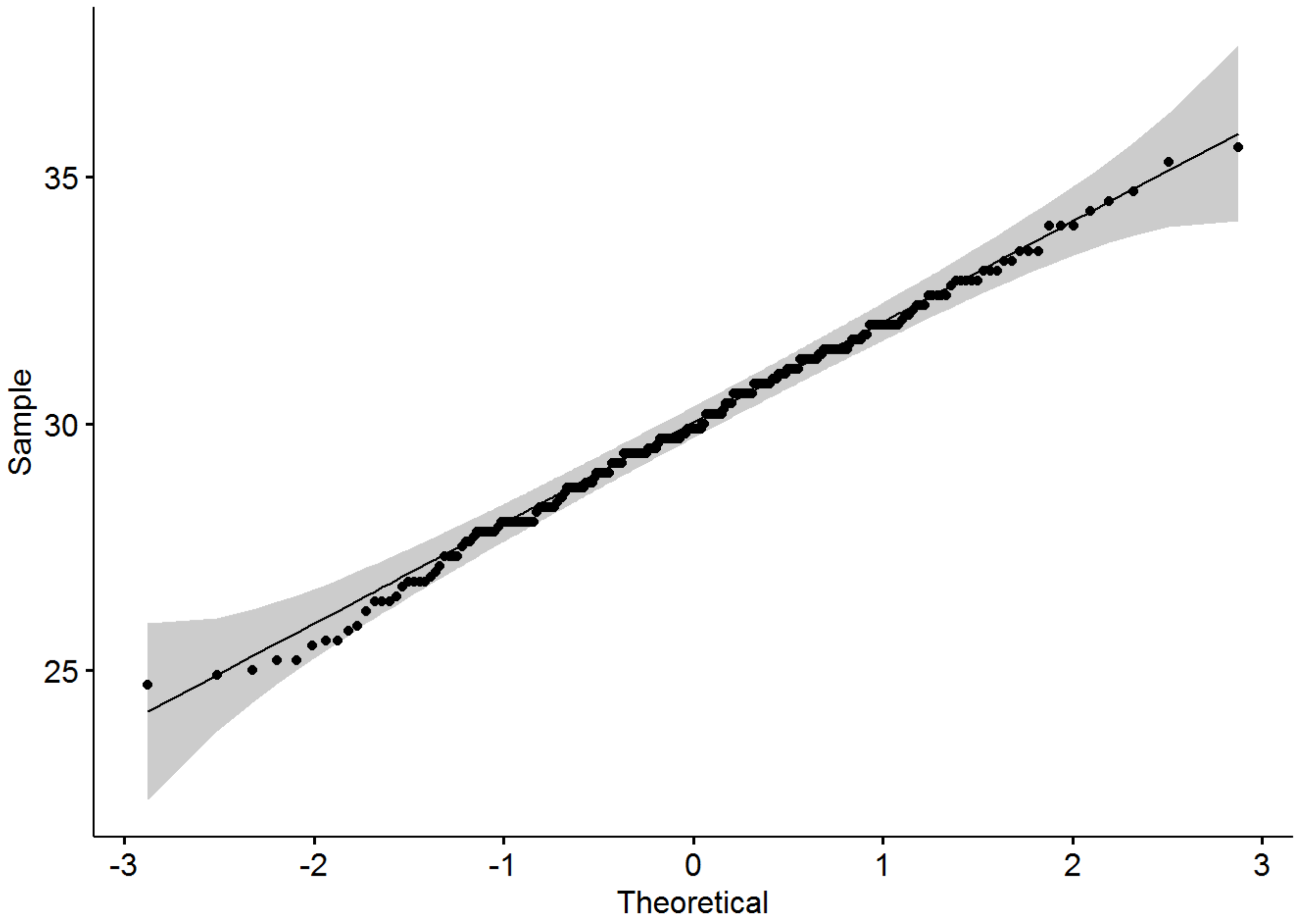
Distribution Fitting

Compare the empirical distribution of selected body measurement to a normal distribution separately in men and in women. You need to do this visually by plotting the histogram with normal distribution overlay. Show your code.

```
#Male Observations
x <- seq(min(body_dim_male$che.di),max(body_dim_male$che.di))
mu <- mean(body_dim_male$che.di)
sd <- sd(body_dim_male$che.di)
ggdensity(body_dim_male$che.di,
          main = "Density plot of Chest Diameter of Males",
          xlab = "Chest Diameter (cm)") #plotting the density plot for testing normality visually
```




```
ggqqplot(body_dim_male$che.di) #plotting Q-Q to draw a visual correlation between a given sample and the normal distribution
```



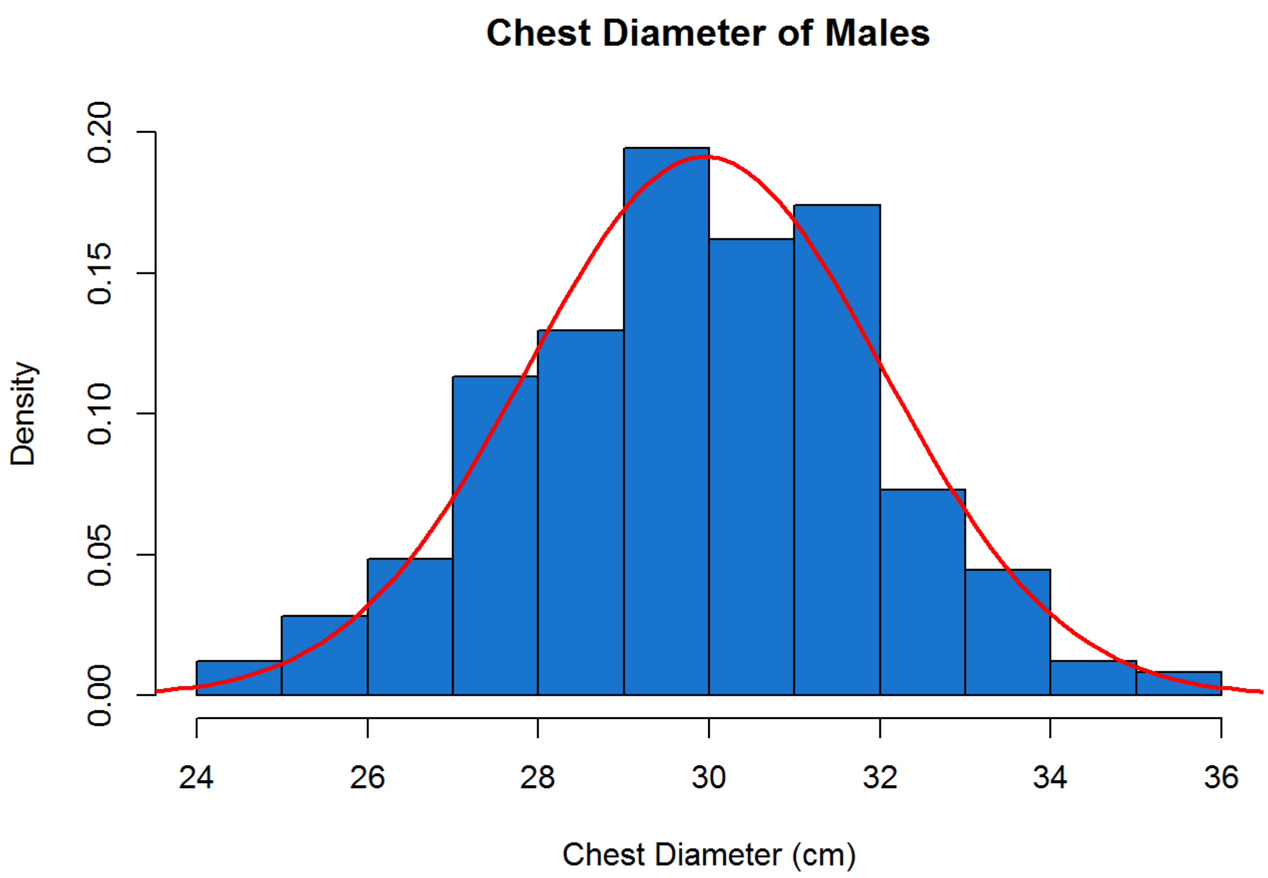

```
shapiro.test(body_dim_male$che.di) #performing the Shapiro-Wilk's test for testing normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  body_dim_male$che.di
## W = 0.99577, p-value = 0.739
```

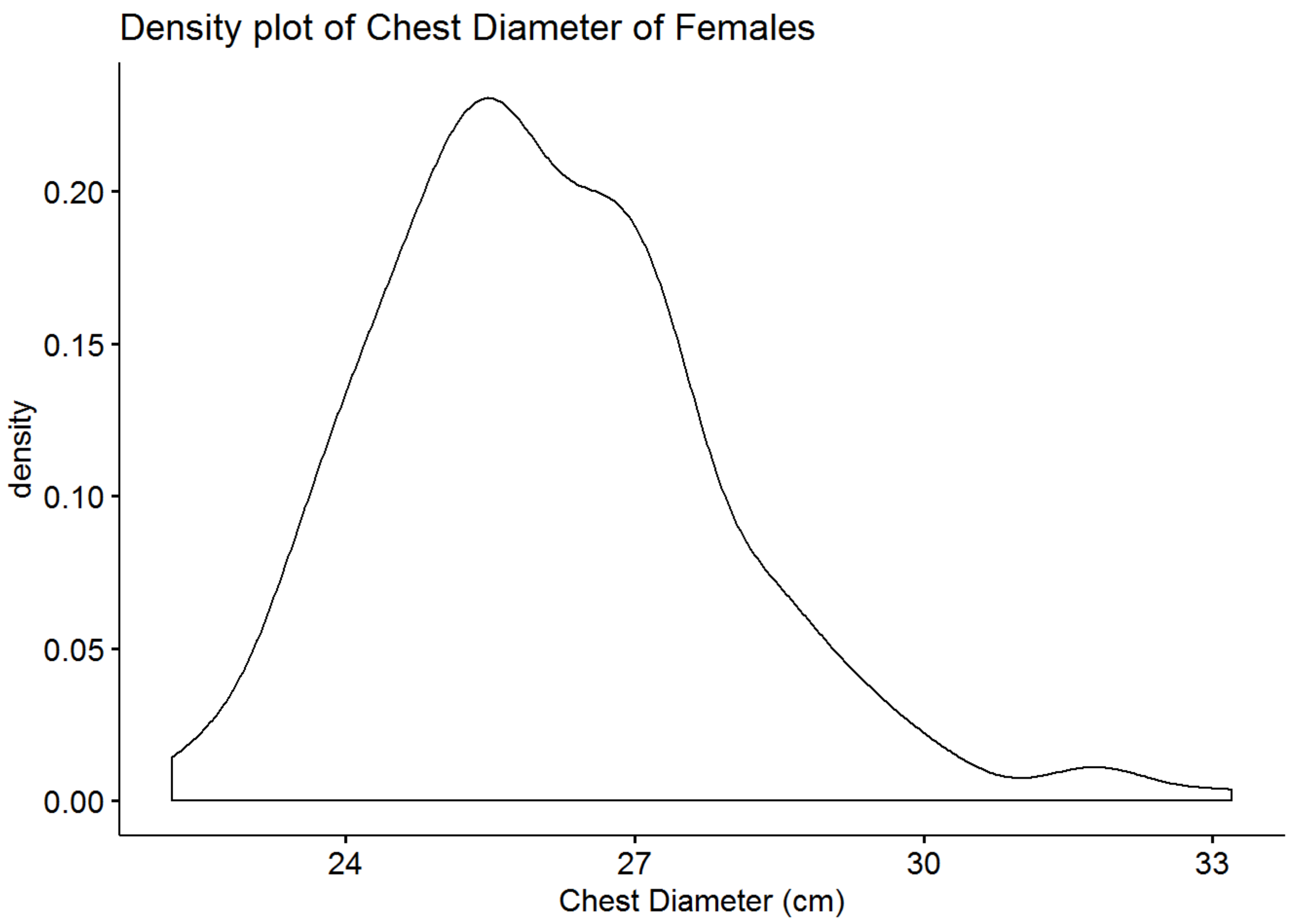
```
jarque.bera.test(body_dim_male$che.di) #performing the Jarque-Bera test for testing normality
```

```
##
##  Jarque Bera Test
##
## data:  body_dim_male$che.di
## X-squared = 0.39304, df = 2, p-value = 0.8216
```

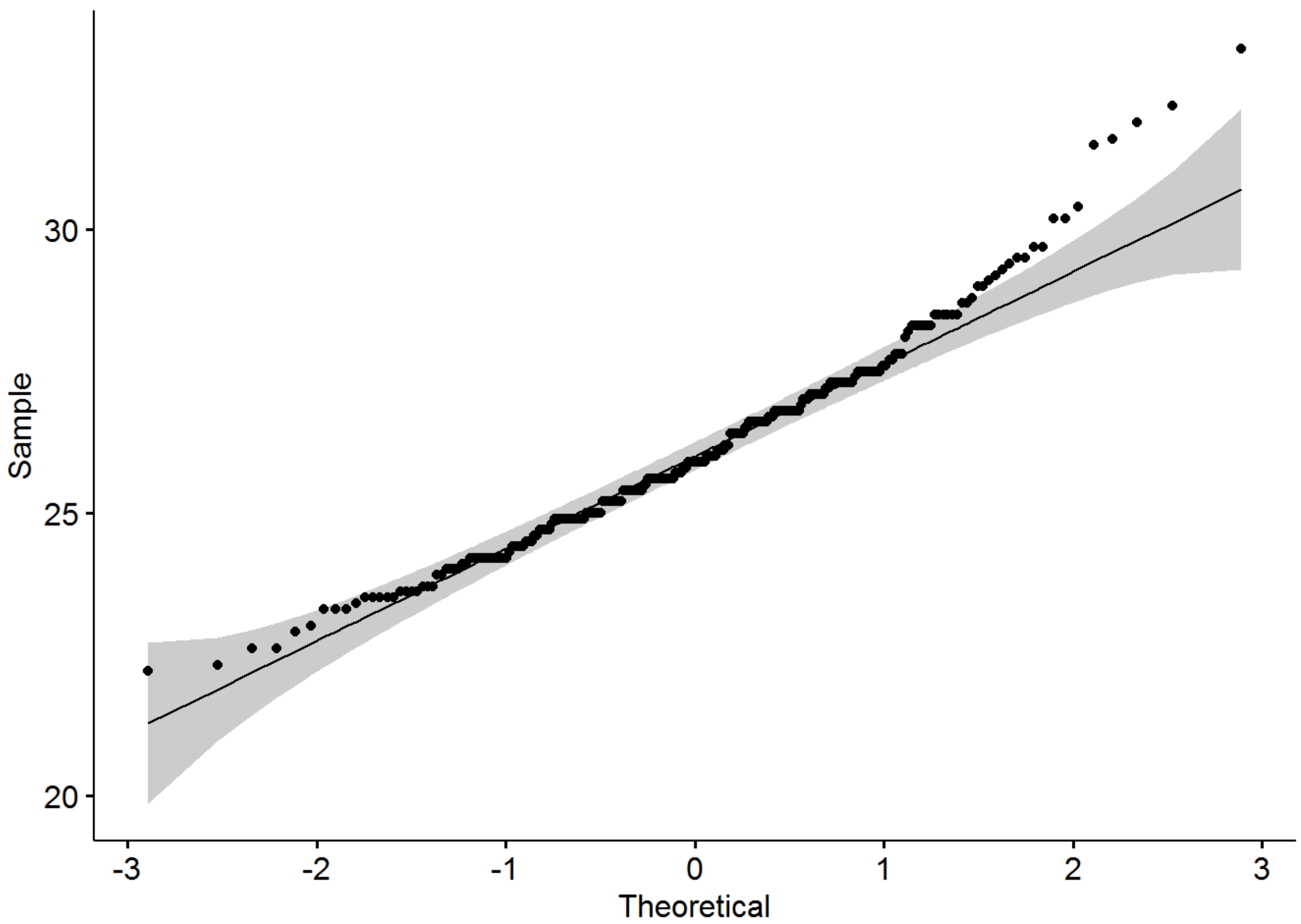
```
body_dim_male$che.di %>% hist(xlab="Chest Diameter (cm)", ylim = c(0, 0.2),
                              main="Chest Diameter of Males",
                              prob=TRUE,col = "dodgerblue3") #plotting the histogram of the observations
curve(dnorm(x,mu,sd),xlim = c(mu-sd*4, mu+sd*4),col="red", add=TRUE, lwd= 2) #overlapping the normal distribution curve over histogram for visual normality testing
```



```
#Female Observations
x <- seq(min(body_dim_female$che.di),max(body_dim_female$che.di))
mu <- mean(body_dim_female$che.di)
sd <- sd(body_dim_female$che.di)
ggdensity(body_dim_female$che.di,
          main = "Density plot of Chest Diameter of Females",
          xlab = "Chest Diameter (cm)")#plotting the density plot for testing normality visually
```




```
ggqqplot(body_dim_female$che.di) #plotting Q-Q to draw a visual correlation between a given sample and the normal distributi
on
```



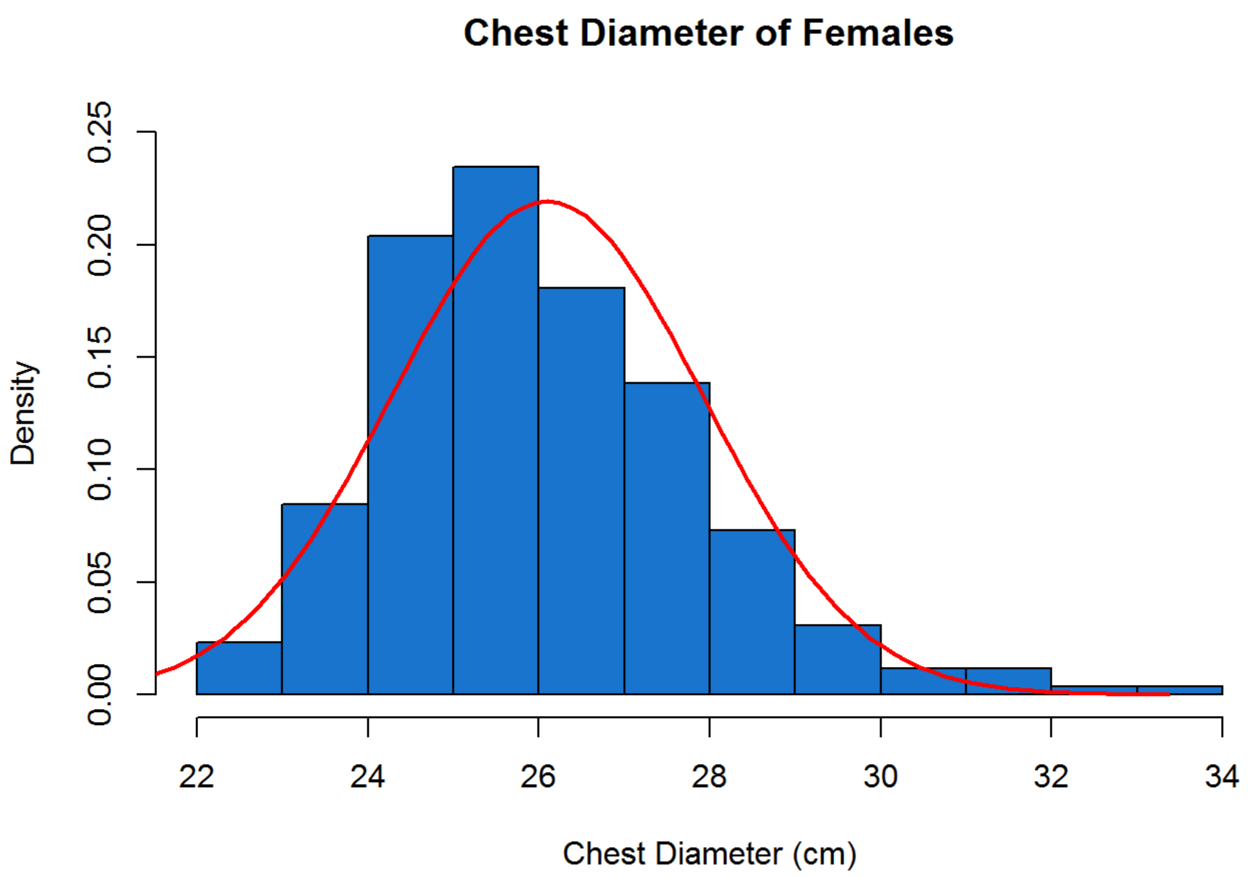

```
shapiro.test(body_dim_female$che.di) #performing the Shapiro-Wilk’s test for testing normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  body_dim_female$che.di
## W = 0.96777, p-value = 1.357e-05
```

```
jarque.bera.test(body_dim_female$che.di) #performing the Jarque-Bera test for testing normality
```

```
##
##  Jarque Bera Test
##
## data:  body_dim_female$che.di
## X-squared = 38.533, df = 2, p-value = 4.293e-09
```

```
body_dim_female$che.di %>% hist(xlab="Chest Diameter (cm)", ylim = c(0, 0.25),
                              main="Chest Diameter of Females",
                              prob=TRUE,col = "dodgerblue3")#plotting the histogram of the observations
curve(dnorm(x,mu,sd),xlim = c(mu-sd*4, mu+sd*4),col="red", add=TRUE, lwd= 2) #overlapping the normal distribution curve over histogram for visual normality testing
```



Interpretation

Going back to your problem statement, what insight has been gained from the investigation? Discuss the extent to how your theoretical normal distribution fits the empirical data.

After checking the summary statistics I found that the mean and median for chest diameter in males are almost similar but it is different for females. This means that the chest diameter for males are symmetric where for females it is right skewed. Futher, the density plot for males shows a symmetric normal curve, and the sample and theoretical values do not vary much in Q-Q plot hinting normal distribution. To further confirm this empirical tests namely Shapiro-Wilk’s test and Jarque-Bera’s test were performed on male chest diameter. Both the tests showed a p value >0.05 confirming normal distribution. Finally this was visually verified by plotting histogram with a normal distribution overlap. All these agree for a perfect normal distribution for male chest diameter. However, the density plot for females shows a right skewed normal curve, the sample and theoretical values vary in Q-Q plot hinting variation from a theoretic normal curve. To further explore this empirical tests namely Shapiro-Wilk’s test and Jarque-Bera’s test were performed on female chest diameter. Both the tests showed a p value < 0.05 confirming a non normal distribution. Finally this was visually verified by plotting histogram with a normal distribution overlap which shows right skewness. All these confirm that female chest diameter do not follow a normal distribution.