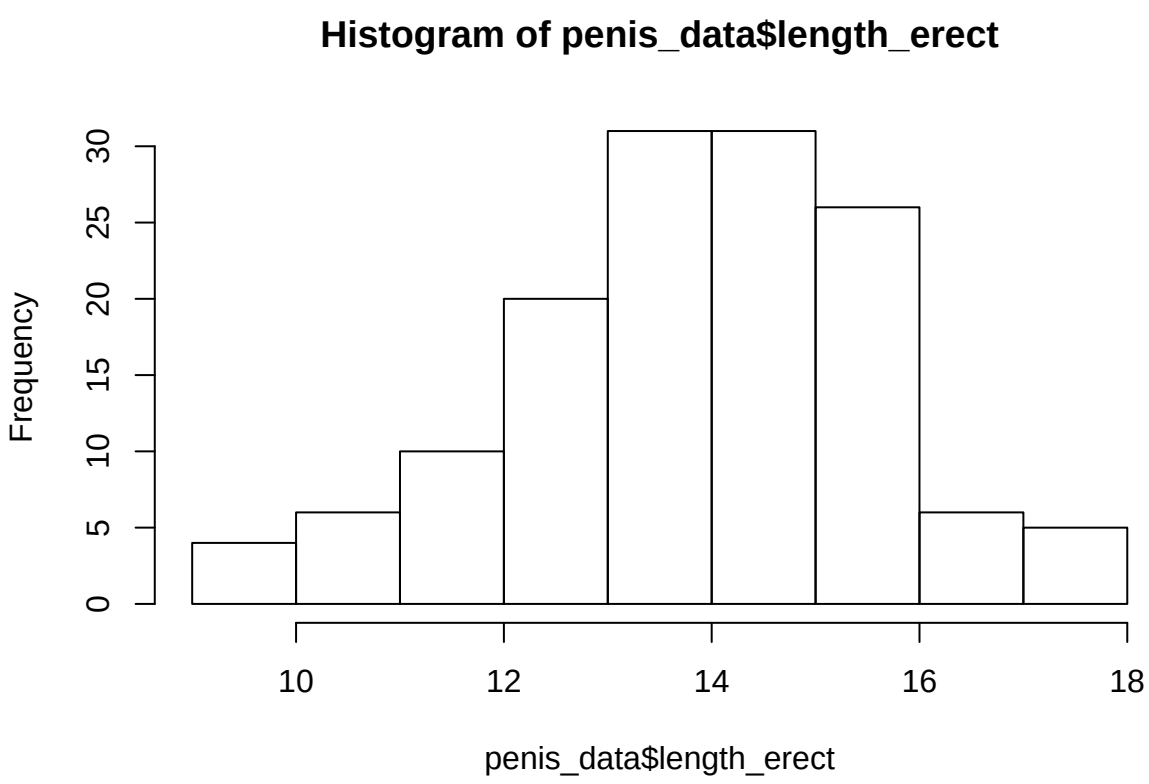
Penis Measurements Across the World

```
library(ggplot2)
library(dplyr)
```

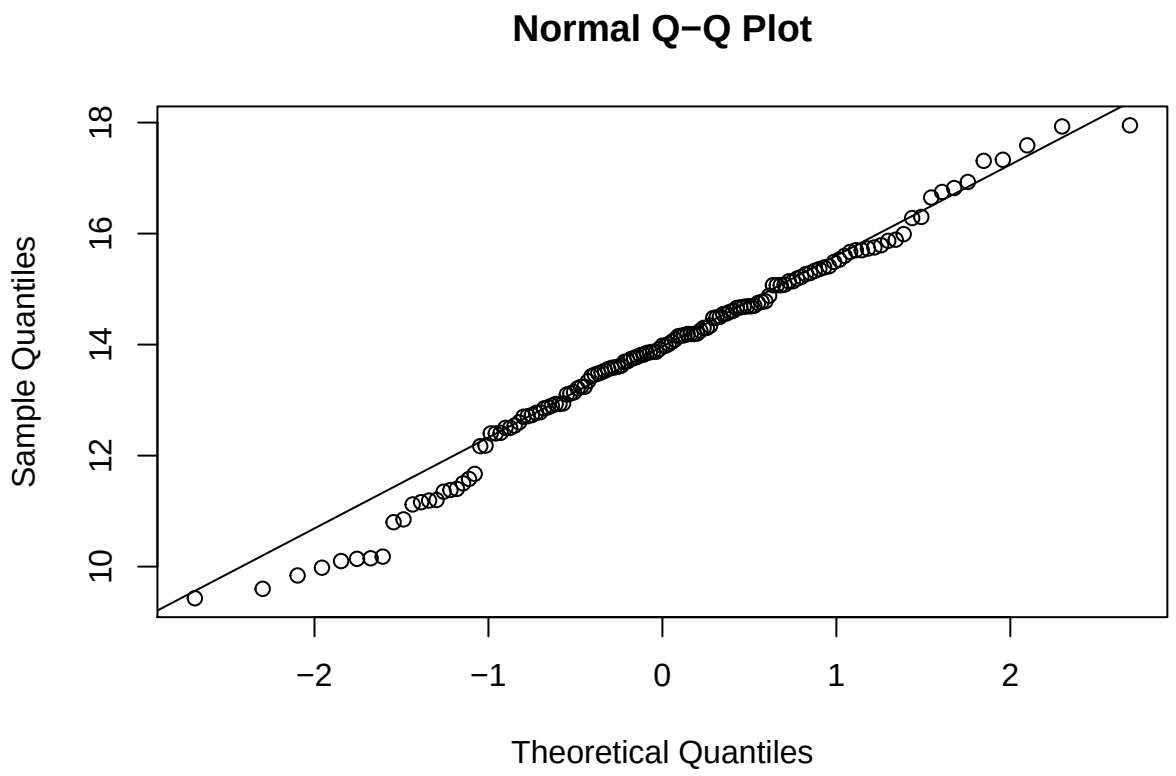
```
##
## Attaching package: 'dplyr'

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

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

```
# read in dataset
penis_data <- read.csv("/Users/Dohyun/Desktop/projects/Penis-Project/world_penis_dataset/penis.csv")
```

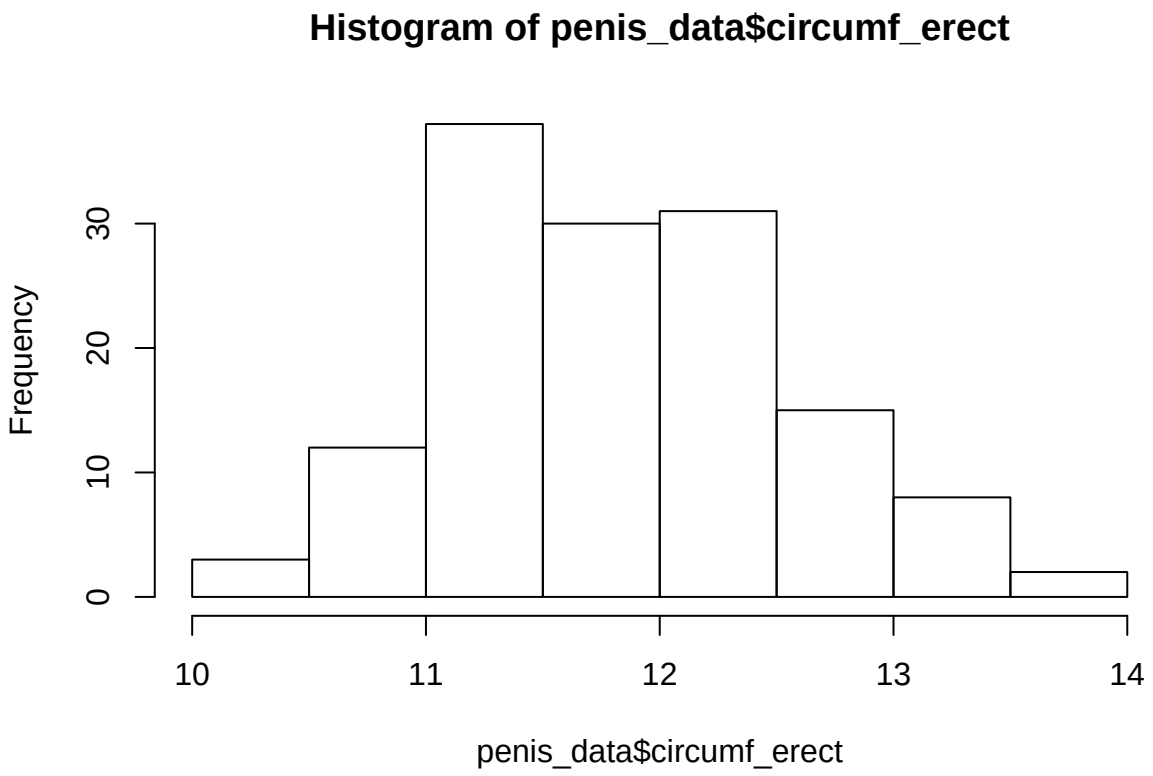
```
#check normality of erect length means

#using a histogram
hist(penis_data$length_erect)
```



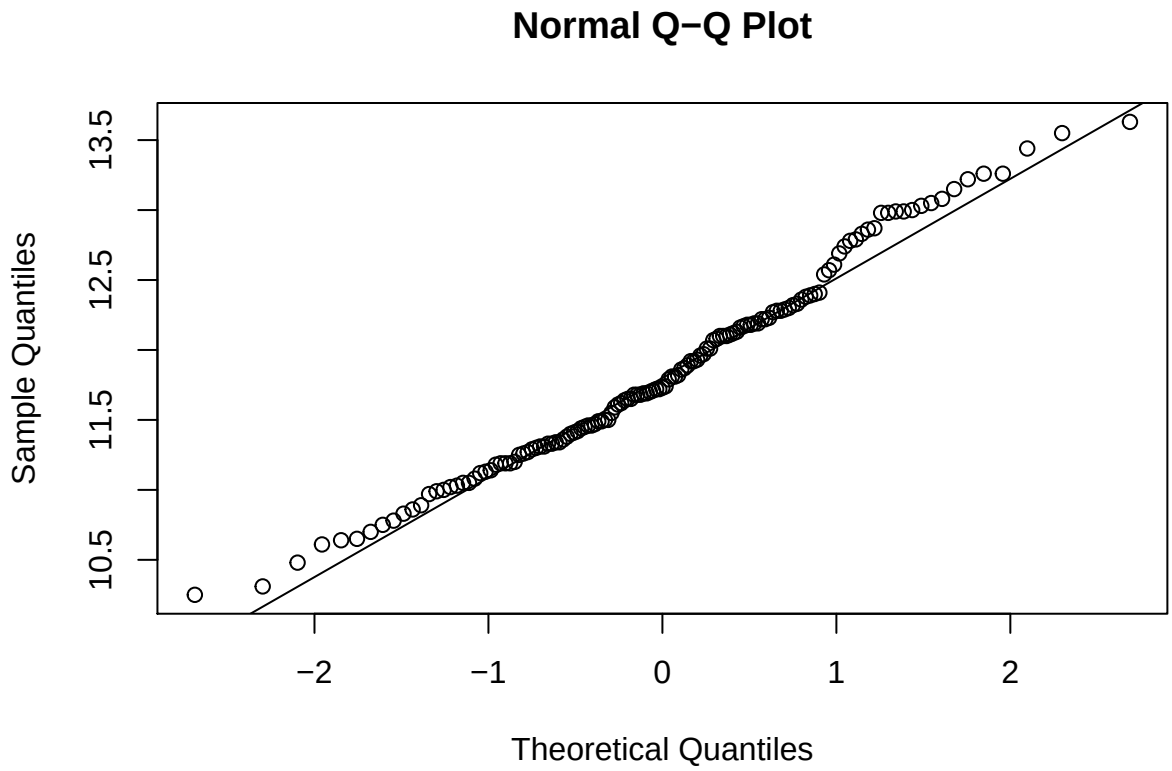
```r
#NPP plot
qqnorm(penis_data$length_erect)
qqline(penis_data$length_erect)
```




```r
#check normality of erect girth means

#using a histogram
hist(penis_data$circumf_erect)
```




```r
#NPP plot
qqnorm(penis_data$circumf_erect)
qqline(penis_data$circumf_erect)
```



```r
t.test(penis_data$length_erect)
```

```
##
##  One Sample t-test
##
## data:  penis_data$length_erect
## t = 91.633, df = 138, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  13.55726 14.15526
## sample estimates:
## mean of x
##  13.85626
```

```r
t.test(penis_data$circumf_erect)
```

```
##
##  One Sample t-test
##
## data:  penis_data$circumf_erect
## t = 192.43, df = 138, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  11.71941 11.96275
## sample estimates:
## mean of x
##  11.84108
```

Confidence interval for mean erect length is 13.56-14.16 cm. Confidence interval for mean erect girth is 11.72-11.96 cm.

```r
#check for overlaps between both methods
self_reported_data <- filter(penis_data, Method == "Self reported")
measured_data <- filter(penis_data, Method == "Measured")



t.test(self_reported_data$length_erect)
```

```
##
##  One Sample t-test
##
## data:  self_reported_data$length_erect
## t = 79.429, df = 50, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  14.33515 15.07897
## sample estimates:
## mean of x
##  14.70706
```

```r
#CI for mean self-reported length: 14.33-15.08
t.test(measured_data$length_erect)
```

```
##
##  One Sample t-test
##
## data:  measured_data$length_erect
## t = 68.323, df = 87, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  12.97443 13.75193
## sample estimates:
## mean of x
##  13.36318
```

```r
#CI for mean measured length: 12.97-13.75

t.test(self_reported_data$circumf_erect)
```

```
##
##  One Sample t-test
##
## data:  self_reported_data$circumf_erect
## t = 139.45, df = 50, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  11.90118 12.24902
## sample estimates:
## mean of x
##   12.0751
```

```r
#CI for mean self-reported girth: 11.55-11.86
t.test(measured_data$circumf_erect)
```

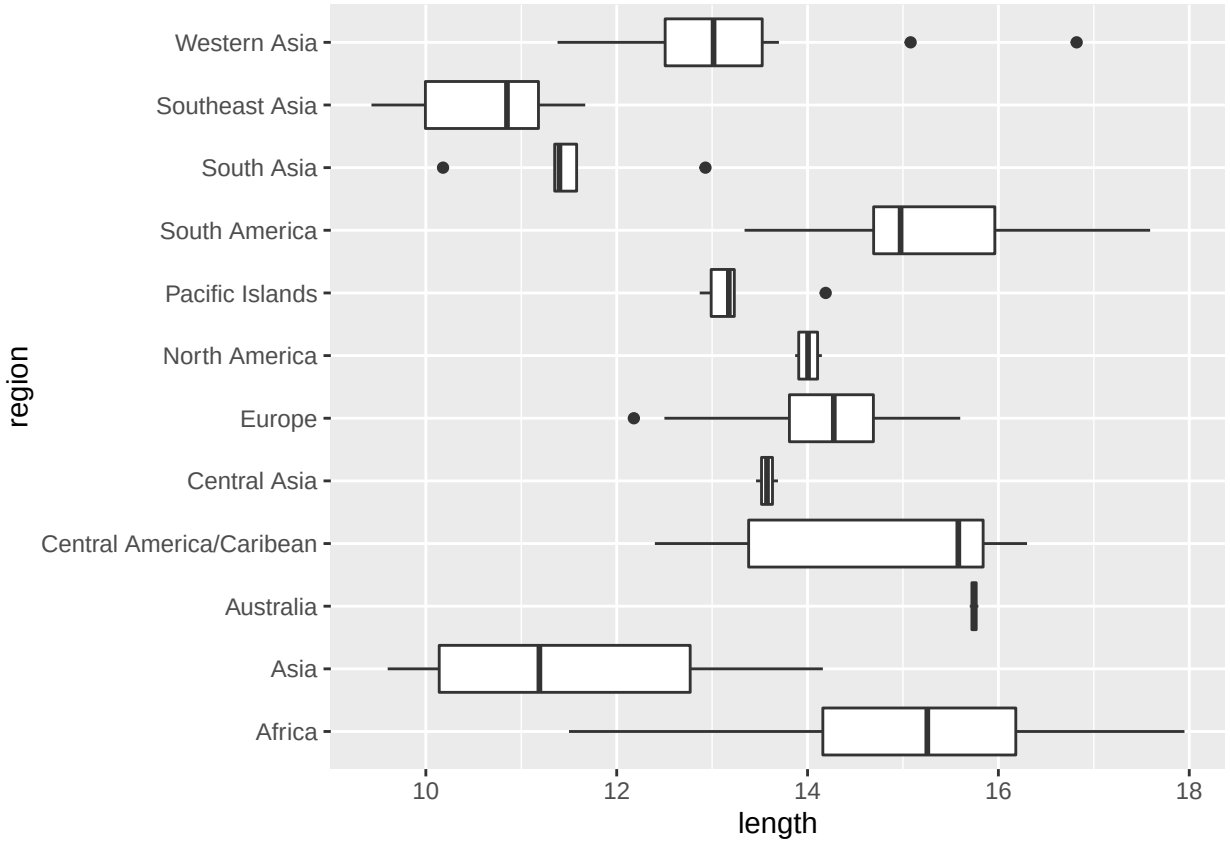
```
##
##  One Sample t-test
##
## data:  measured_data$circumf_erect
## t = 146.16, df = 87, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  11.54628 11.86463
## sample estimates:
## mean of x
##  11.70545
```

```r
##CI for mean measure length: 11.90-12.25
```

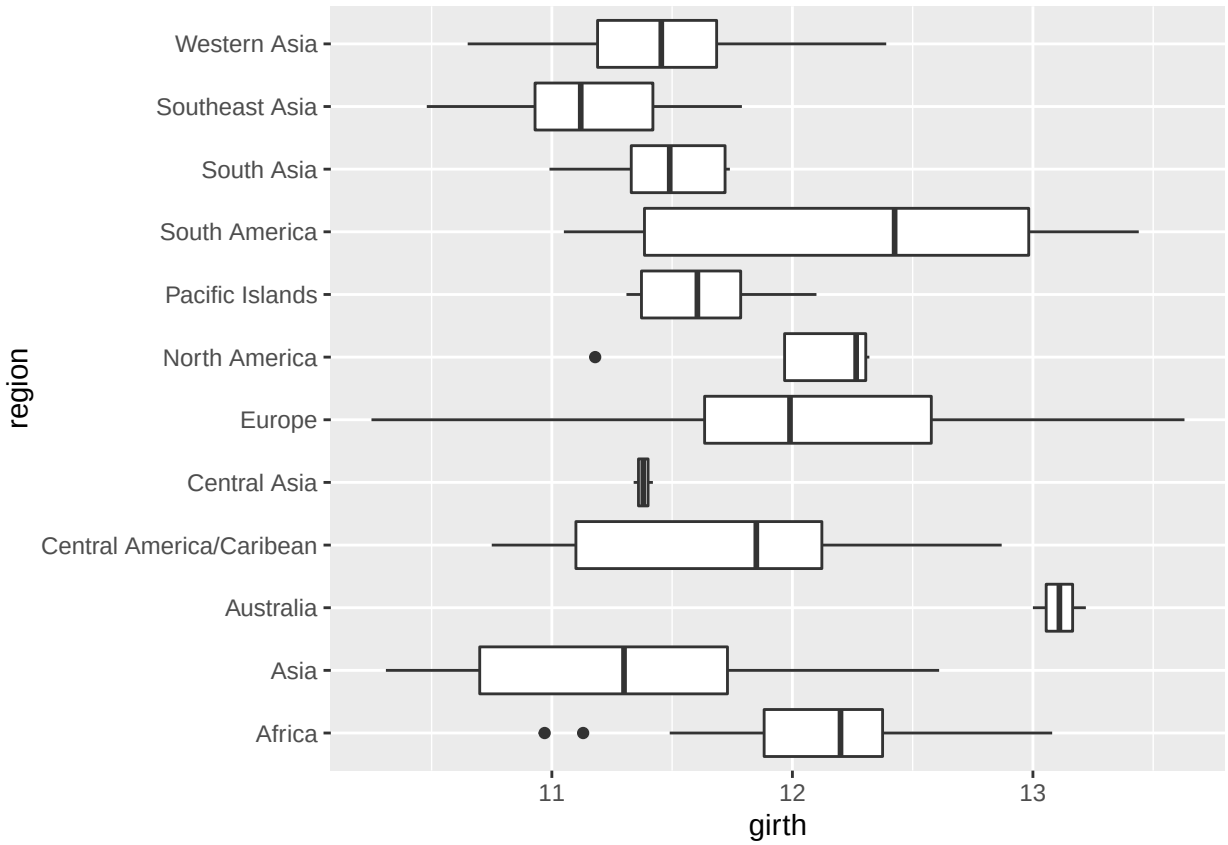
Note that we we only care about the erect length and girth it provides a better standard of measurement. Flaccid measurements will always vary depending on body and outside temperature and different conditions like health.

The confidence intervals for the self reported and measured don't overlap, and on average the self-reported data shows a higher range of measurements, which could skew the data by a couple centimeters. It is not certain whether the self-reported measurements are honest, but it is not out of the question whether they are.

```
size_length <- length(penis_data$length_erect)
size_girth <- length(penis_data$circumf_erect)
region <- penis_data[,"Region"]
length <- penis_data[,"length_erect"]
girth <- penis_data[,"circumf_erect"]

#boxplot of the regions
bp <- ggplot(penis_data, aes(x = region, y = length)) +
  geom_boxplot()
bp + coord_flip()
```




```
#boxplot of the regions vs girth
bp2 <- ggplot(penis_data, aes(x = region, y = girth)) +
  geom_boxplot()
bp2 + coord_flip()
```



What is the relationship between length and girth? Is length a good indicator for girth and vice versa?

```
#create a linear regression model bt length & girth
fit1 <- lm(girth ~ length)
summary(fit1)
```

```
##
## Call:
## lm(formula = girth ~ length)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32413 -0.43201 -0.04332  0.41533  1.63634
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.55684    0.39479  21.675  < 2e-16 ***
## length       0.23702    0.02826   8.387 5.42e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5919 on 137 degrees of freedom
## Multiple R-squared:  0.3393, Adjusted R-squared:  0.3344
## F-statistic: 70.34 on 1 and 137 DF,  p-value: 5.416e-14
```

From the p-value of our slope (which is 5.42e-14), we are able to tell that our slope isn't zero, which simply just tells us that there is a relationship between length and girth. However, our R-squared value of 0.34

tells us that there is a fairly weak correlation between the two variables, which could lead us to believe that length is not a good indicator of girth and vice versa.