Understanding different visualization methods using Trees, Rubber and oddbooks dataset

Trees

Let us check the structure of trees dataset

```
library(ggplot2)

summary(trees)
```

```
##      Girth           Height       Volume
##   Min.   : 8.30   Min.   :63   Min.   :10.20
##   1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
##   Median :12.90   Median :76   Median :24.20
##   Mean   :13.25   Mean   :76   Mean   :30.17
##   3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
##   Max.   :20.60   Max.   :87   Max.   :77.00
```

```
str(trees)
```

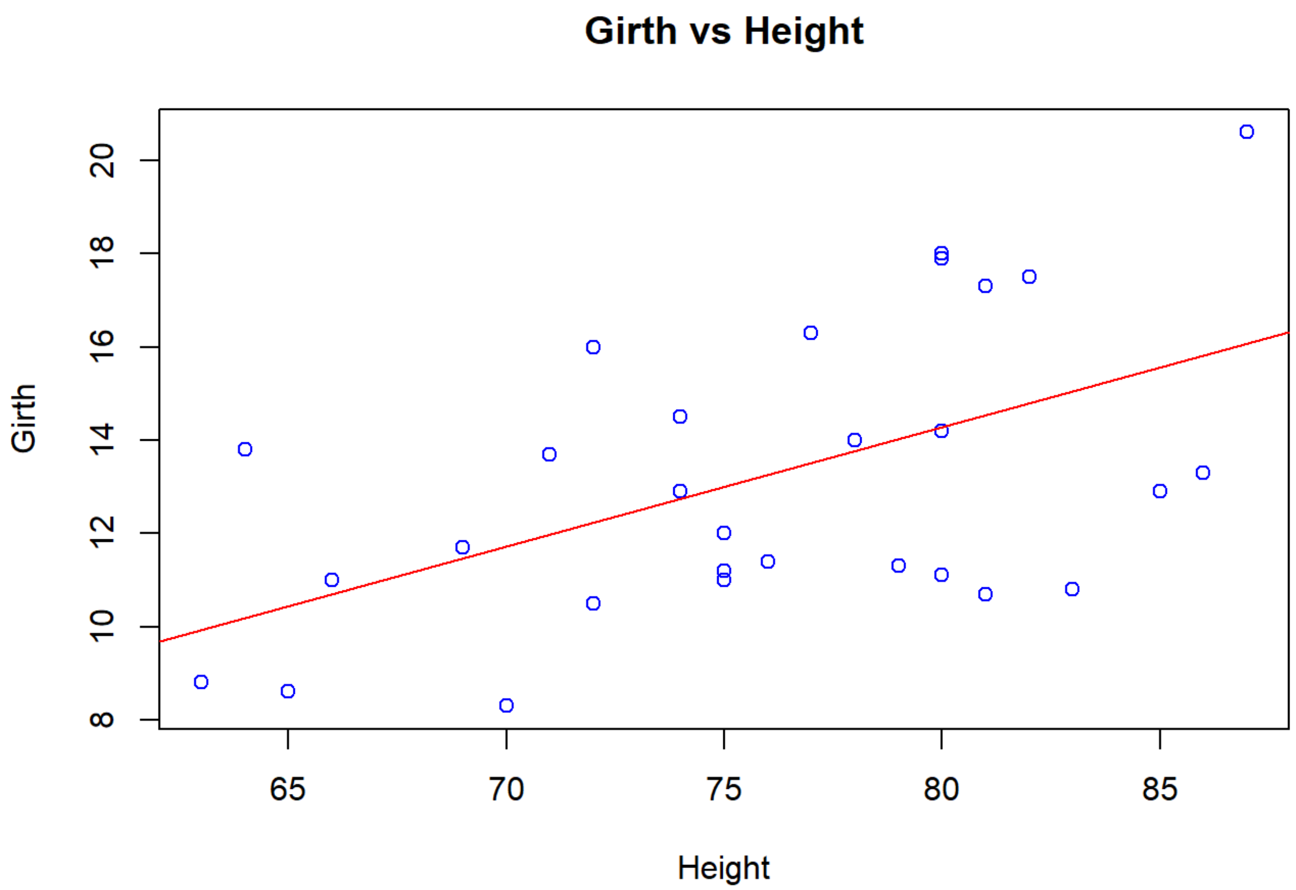
```
## 'data.frame':    31 obs. of  3 variables:
##  $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
##  $ Height: num  70 65 63 72 81 83 66 75 80 75 ...
##  $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
```

```
sapply(trees, is.factor)
```

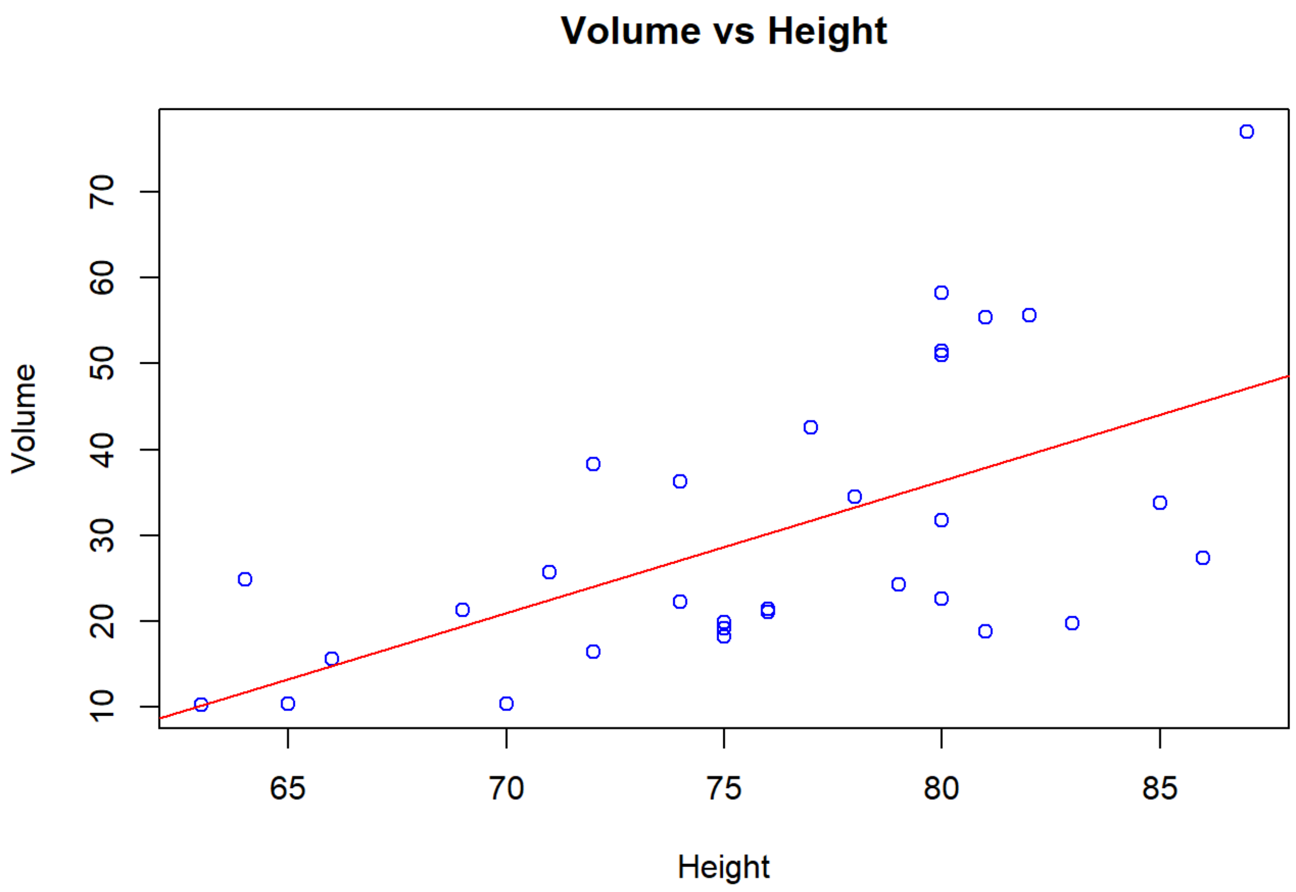
```
##  Girth Height Volume
##  FALSE  FALSE  FALSE
```

We make use of linear model and then plot it

```
model <- lm(Girth ~ Height, data = trees)
model1 <- lm(Volume ~ Height, data = trees)

plot(trees$Height, trees$Girth, main = 'Girth vs Height', xlab = 'Height', ylab = '
Girth', col = 'blue')

abline(model, col = 'red')
```

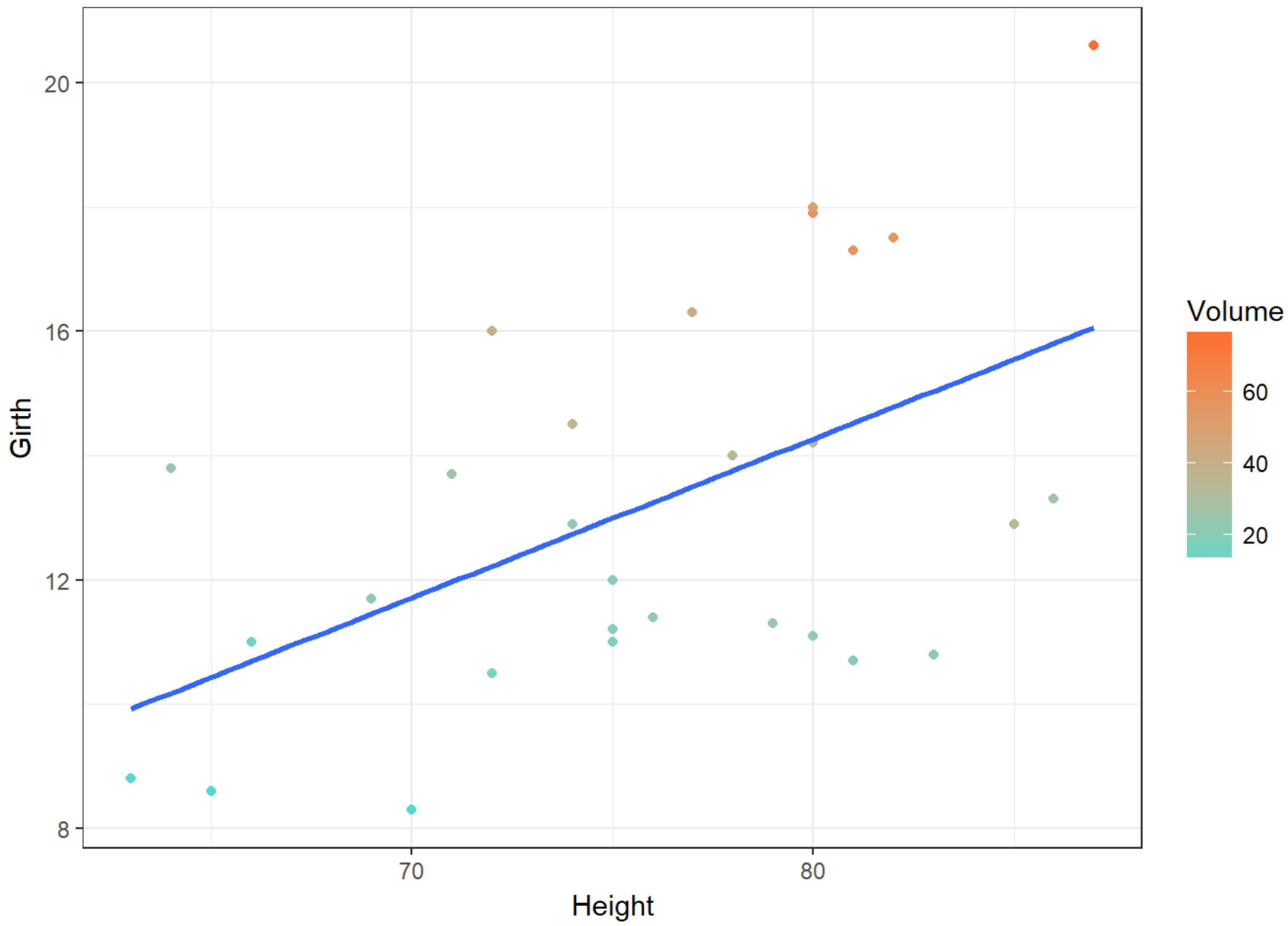



```
plot(trees$Height, trees$Volume, main = 'Volume vs Height', xlab = 'Height', ylab =
'Volume', col = 'blue')

abline(model1, col = 'red')
```

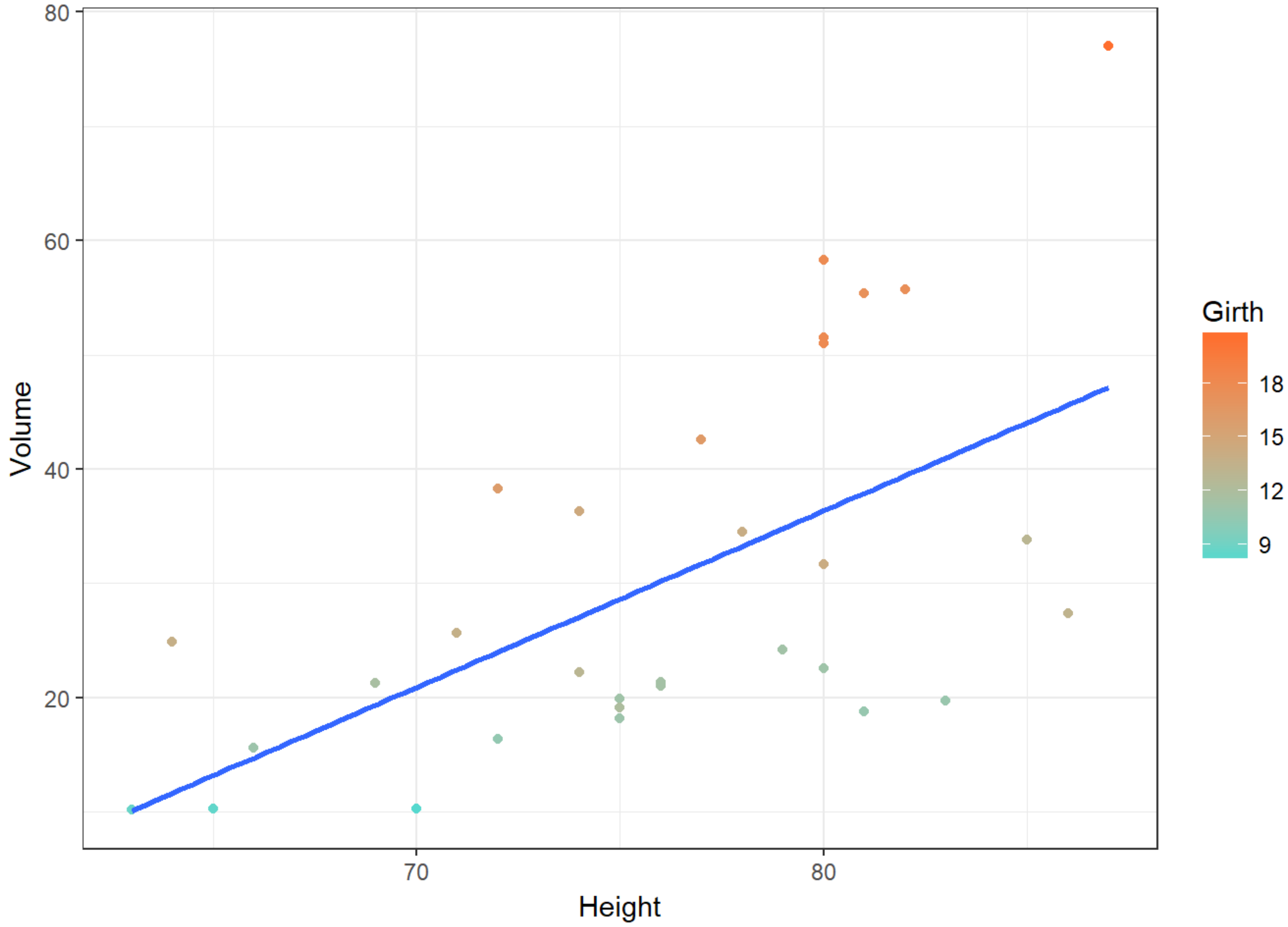


The same can be performed using ggplot

```
ggplot(trees, aes(Height,Girth), alpha = 0.4) + geom_point(aes(color = Volume)) + t
heme_bw() +geom_smooth(method = 'lm', formula = y ~x, se = F) + scale_color_continu
ous(high = "#FF6E2E", low = "#55D8CE")
```

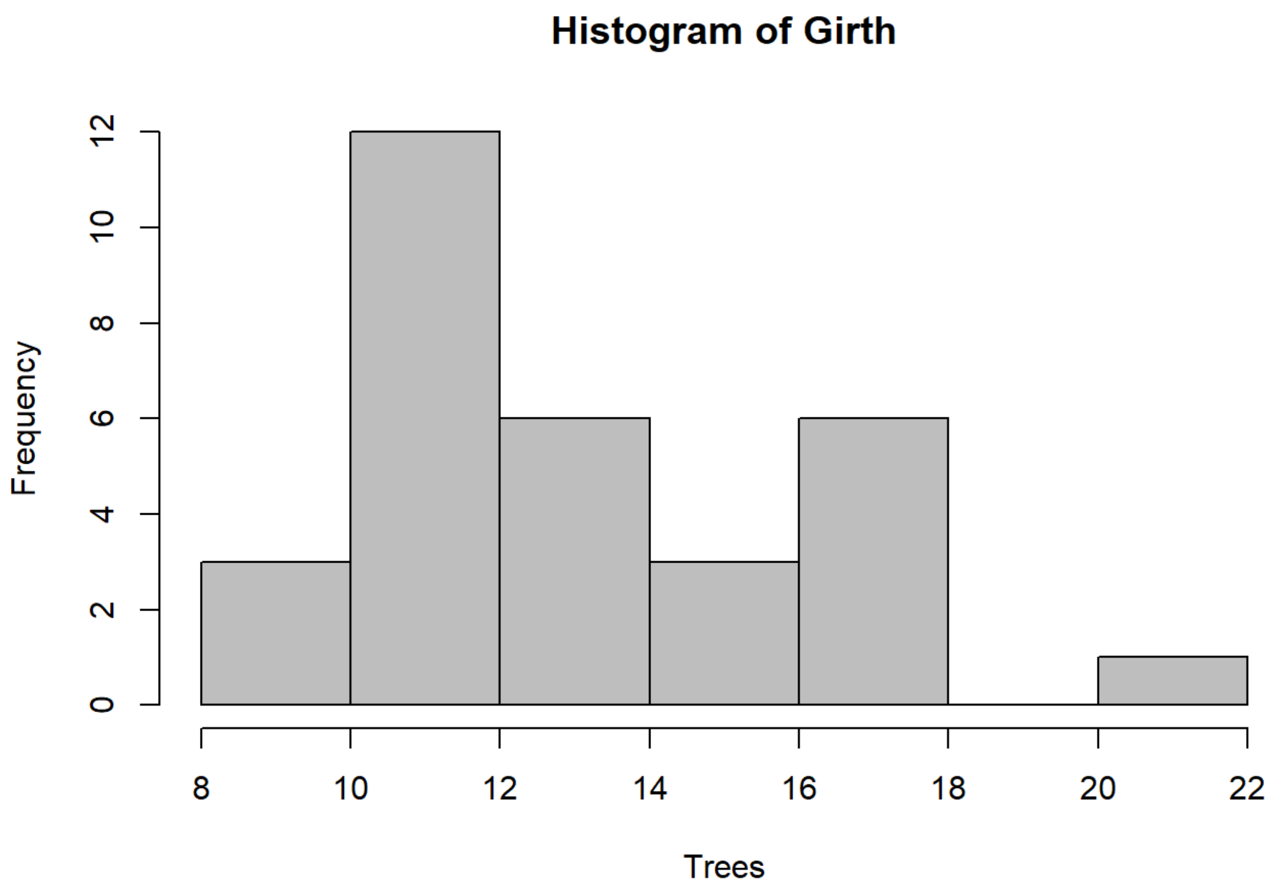



```
ggplot(trees, aes(Height,Volume), alpha = 0.4) + geom_point(aes(color = Girth)) + t
heme_bw() +geom_smooth(method = 'lm', formula = y ~x, se = F) + scale_color_continu
ous(high = "#FF6E2E", low = "#55D8CE")
```

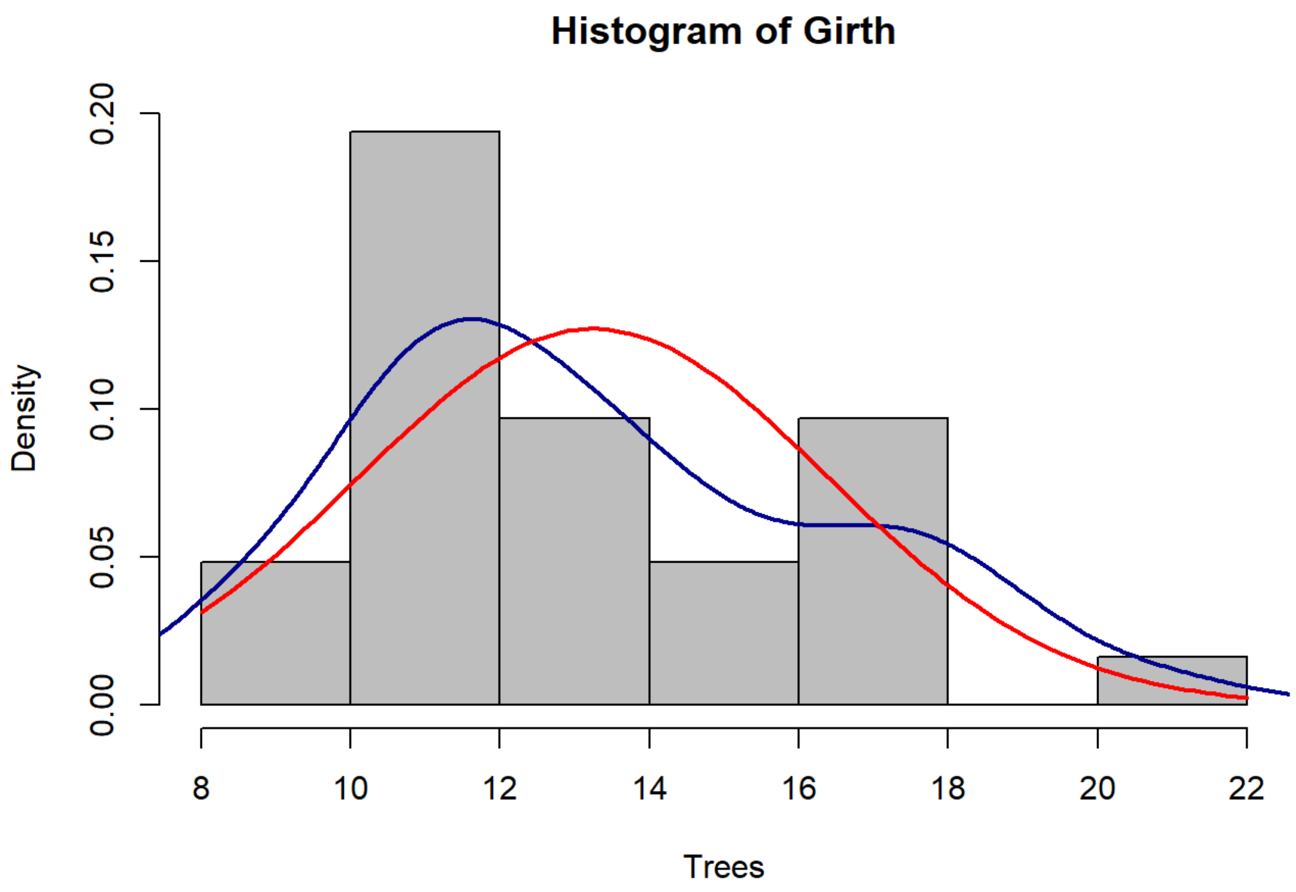


Histogram and density

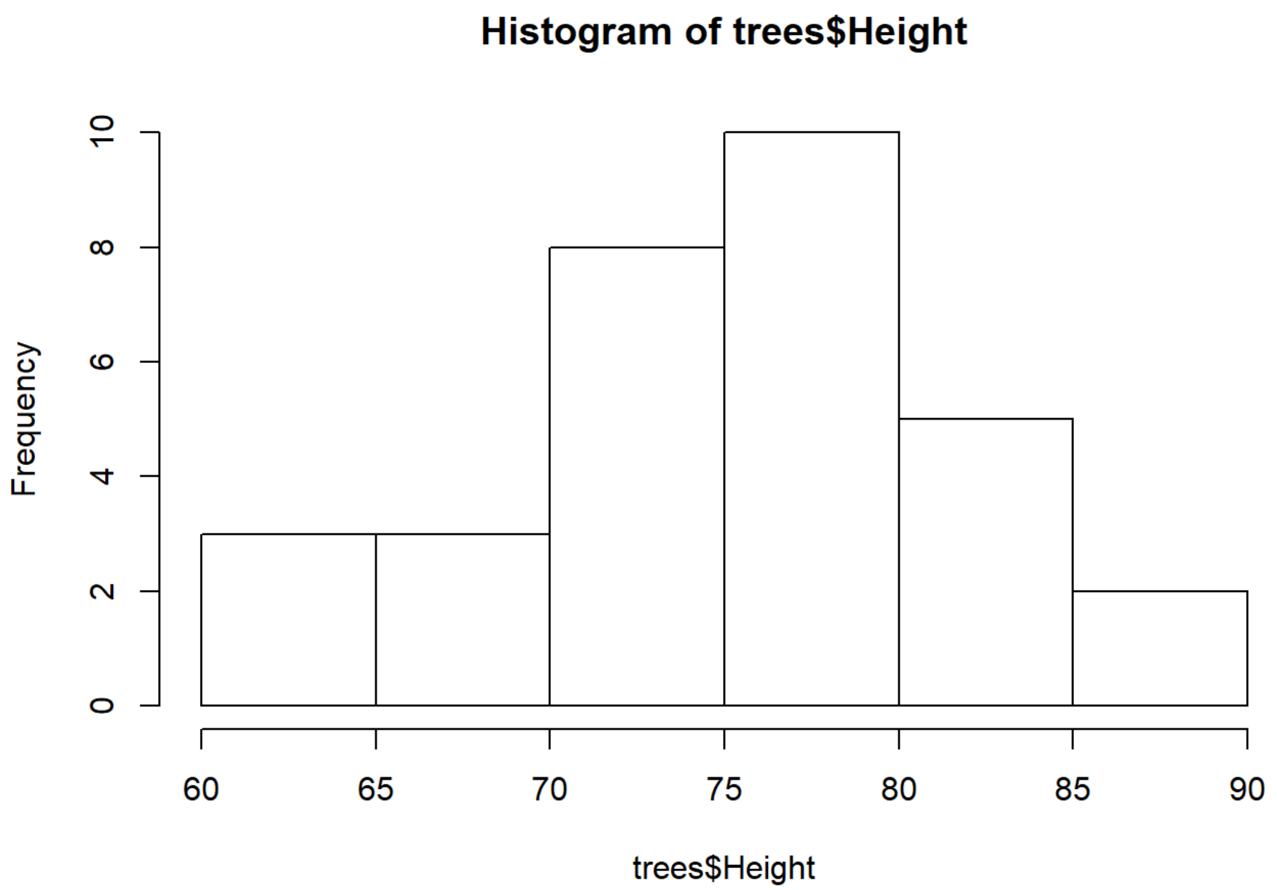
```
hist(trees$Girth, col = 'grey', xlab = "Trees", main = "Histogram of Girth", break
s = 5 )
```



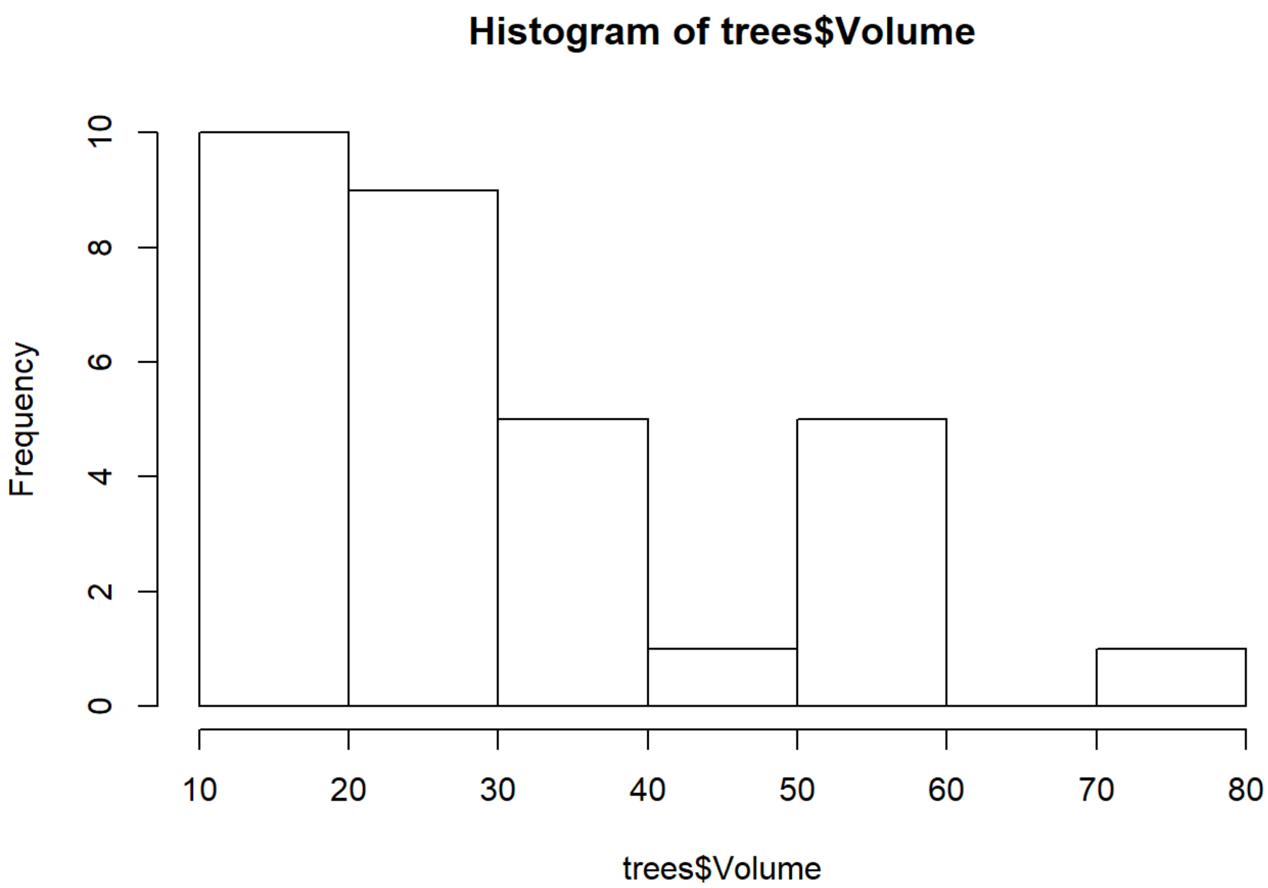

```
hist(trees$Girth, col = 'grey', xlab = "Trees", main = "Histogram of Girth", freq
= F)
d <- density(trees$Girth)
lines(d, lwd = 2, col = 'darkblue')

curve(dnorm(x, mean=mean(trees$Girth), sd=sd(trees$Girth)), add=TRUE, col='red', lw
d=2)
```




```
hist(trees$Height)
```

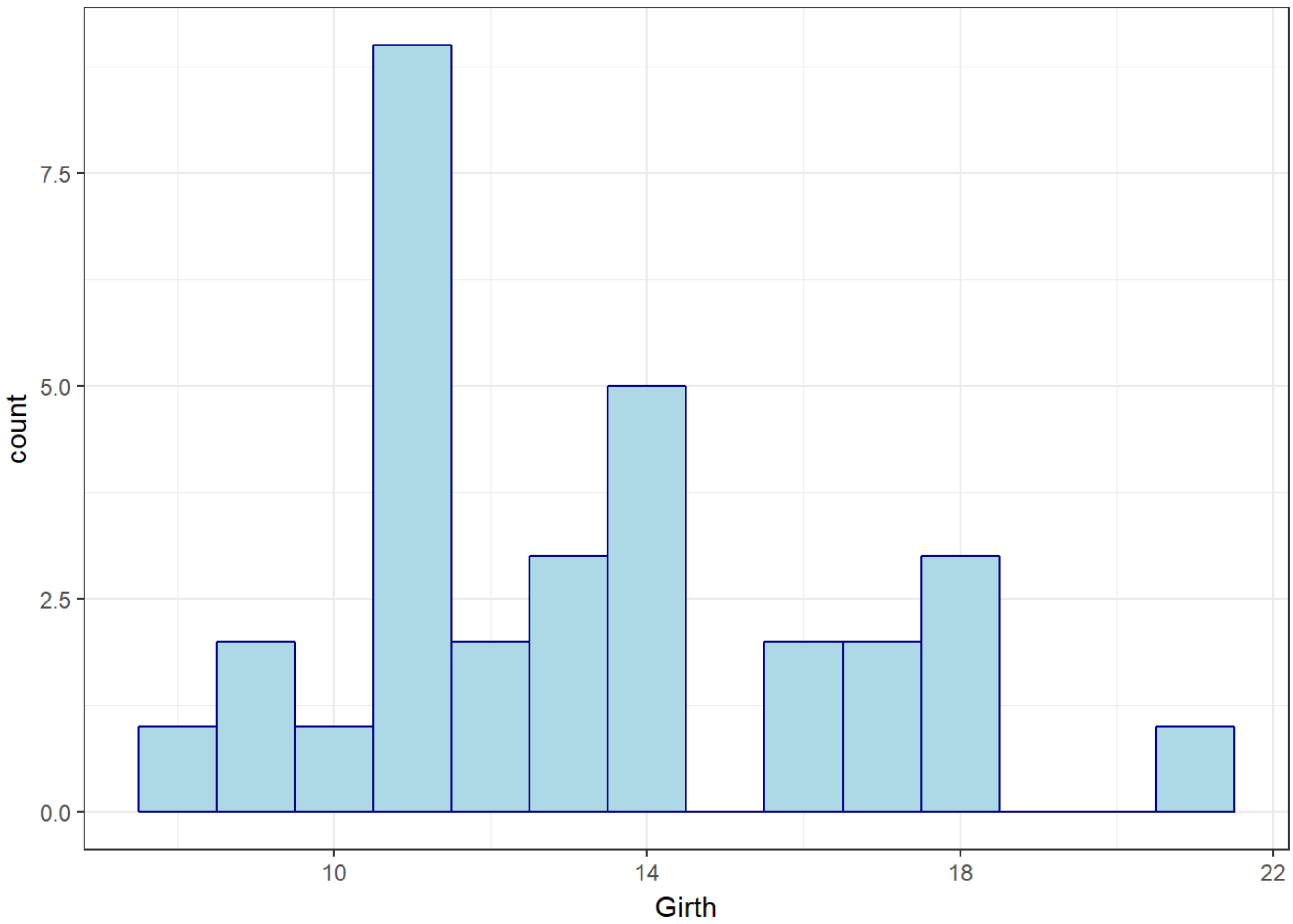


```
hist(trees$Volume)
```

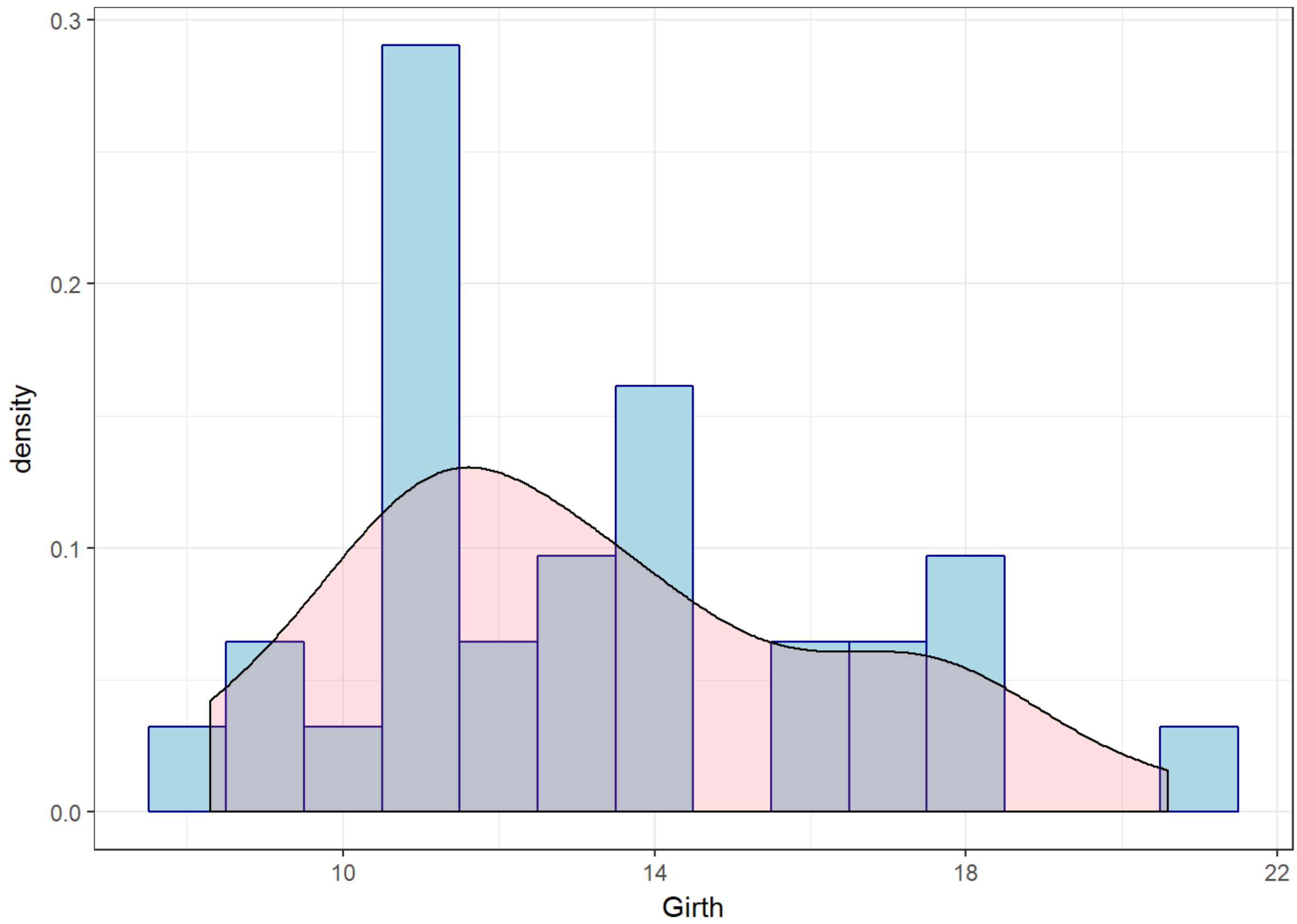


The same can be performed using ggplot

```
ggplot(trees, aes(Girth)) + geom_histogram(fill = 'lightblue', binwidth = 1, color
= 'darkblue') + theme_bw()
```

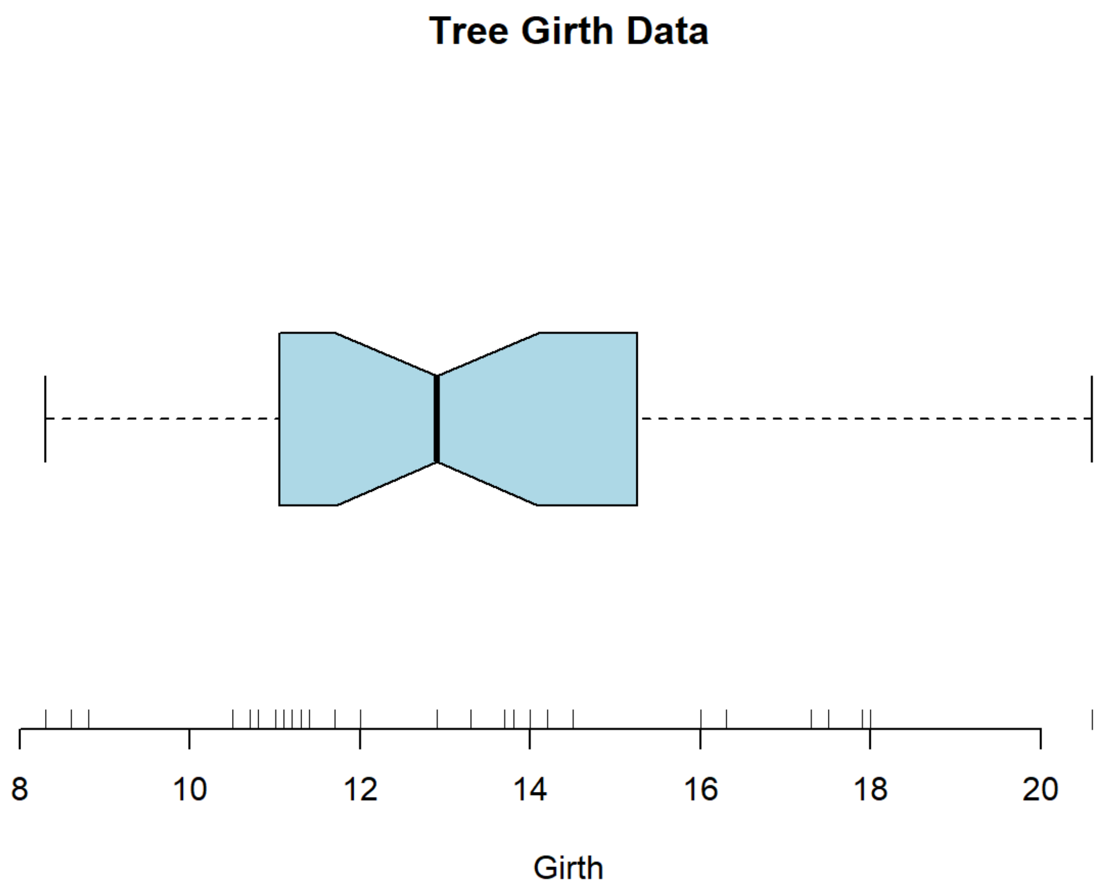



```
ggplot(trees, aes(Girth)) + geom_histogram(aes(y=..density..) ,binwidth = 1, color=
"darkblue", fill="lightblue" ) + theme_bw() + geom_density(alpha=.2, fill="#FF6666"
)
```

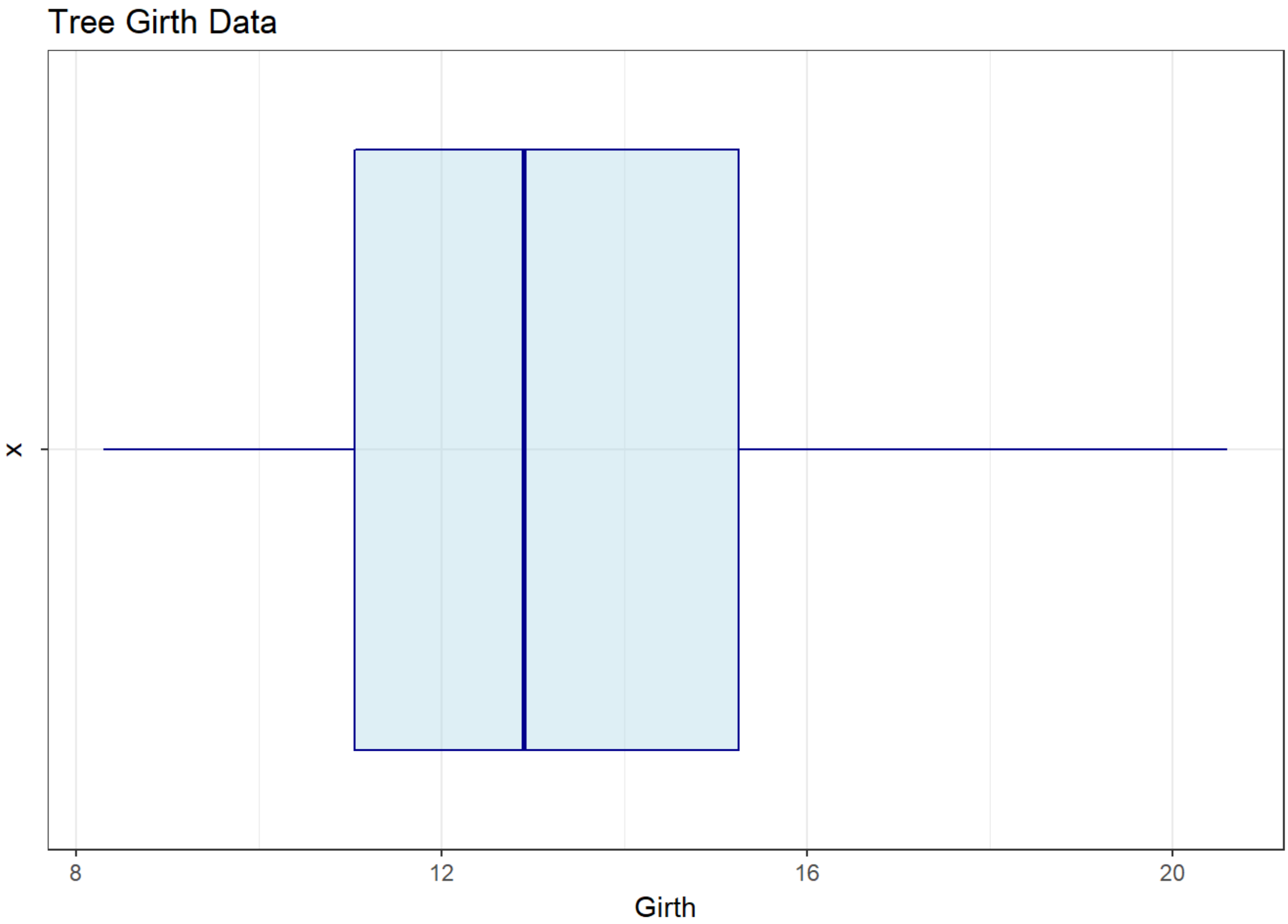


Boxplot and rug

```
boxplot(trees$Girth, col = 'lightblue', horizontal = T, xlab = 'Girth', main = 'Tr
ee Girth Data', frame.plot = F, boxwex = 0.6, notch = T)
rug(trees$Girth, side = 1)
```

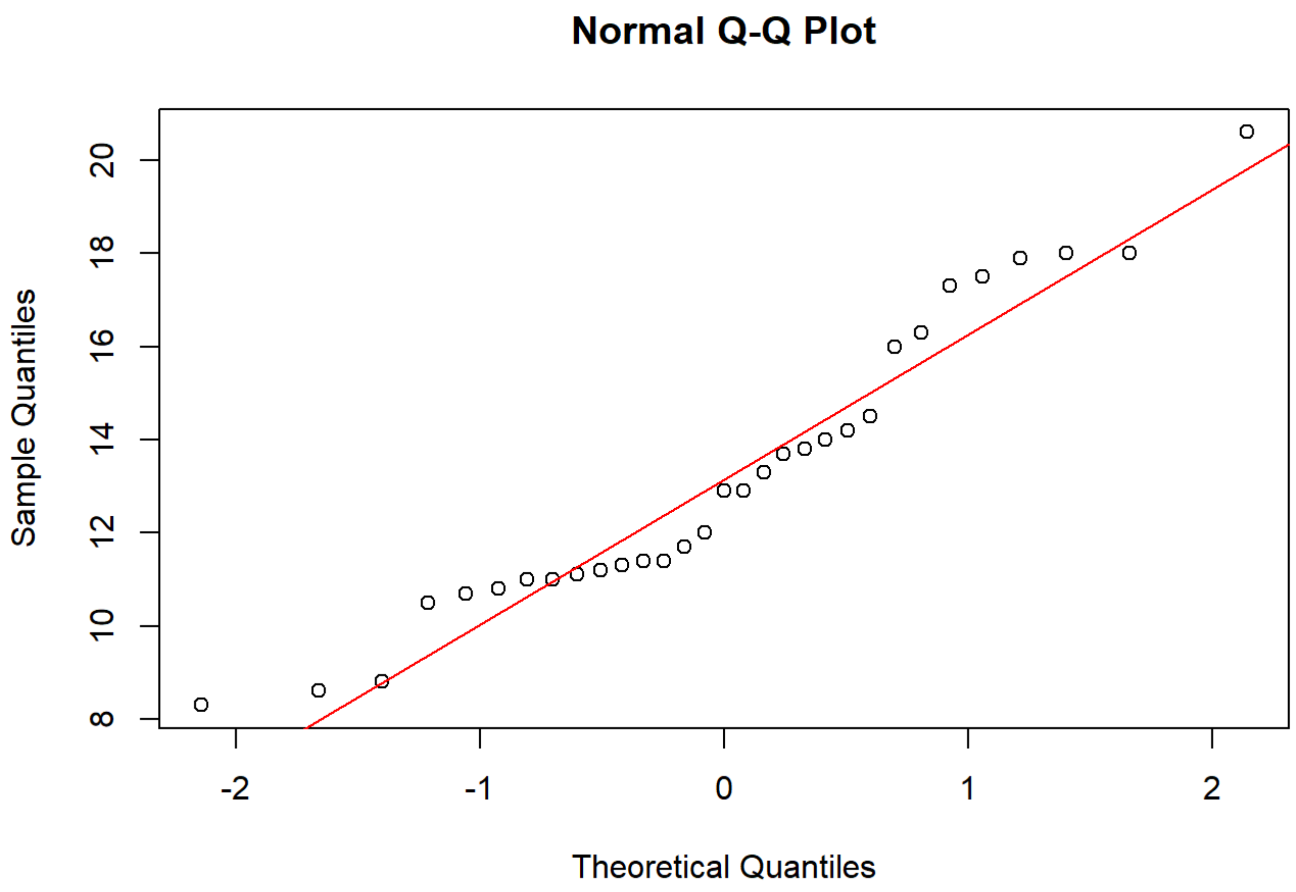



```
ggplot(trees, aes(x='',y=Girth)) + geom_boxplot(color = 'darkblue', fill = 'lightbl
ue', alpha = 0.4)+ theme_bw() + coord_flip() + ggtitle("Tree Girth Data")
```

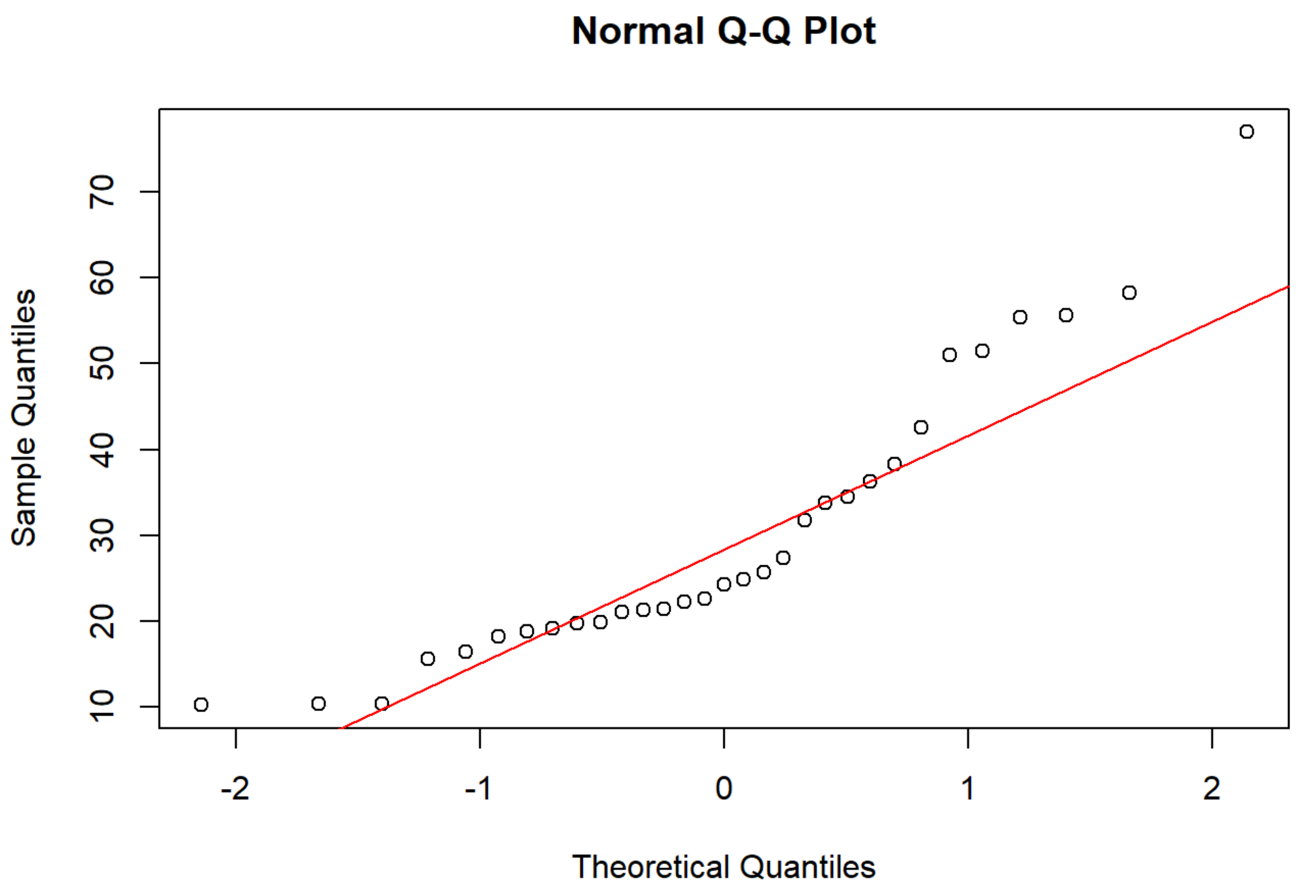


rnorm and qqnorm

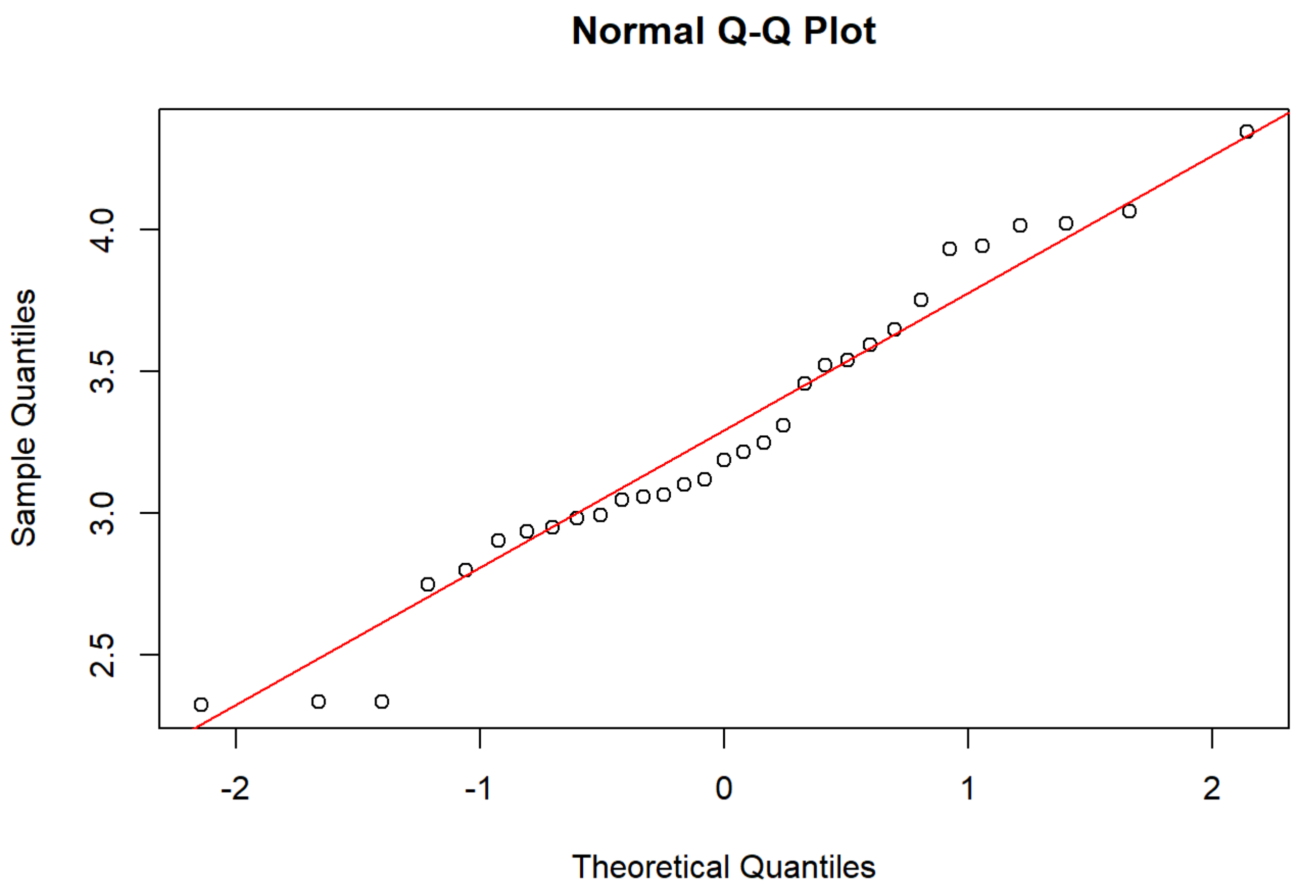
```
qqnorm(trees$Girth)
qqline(trees$Girth, col = 'red')
```



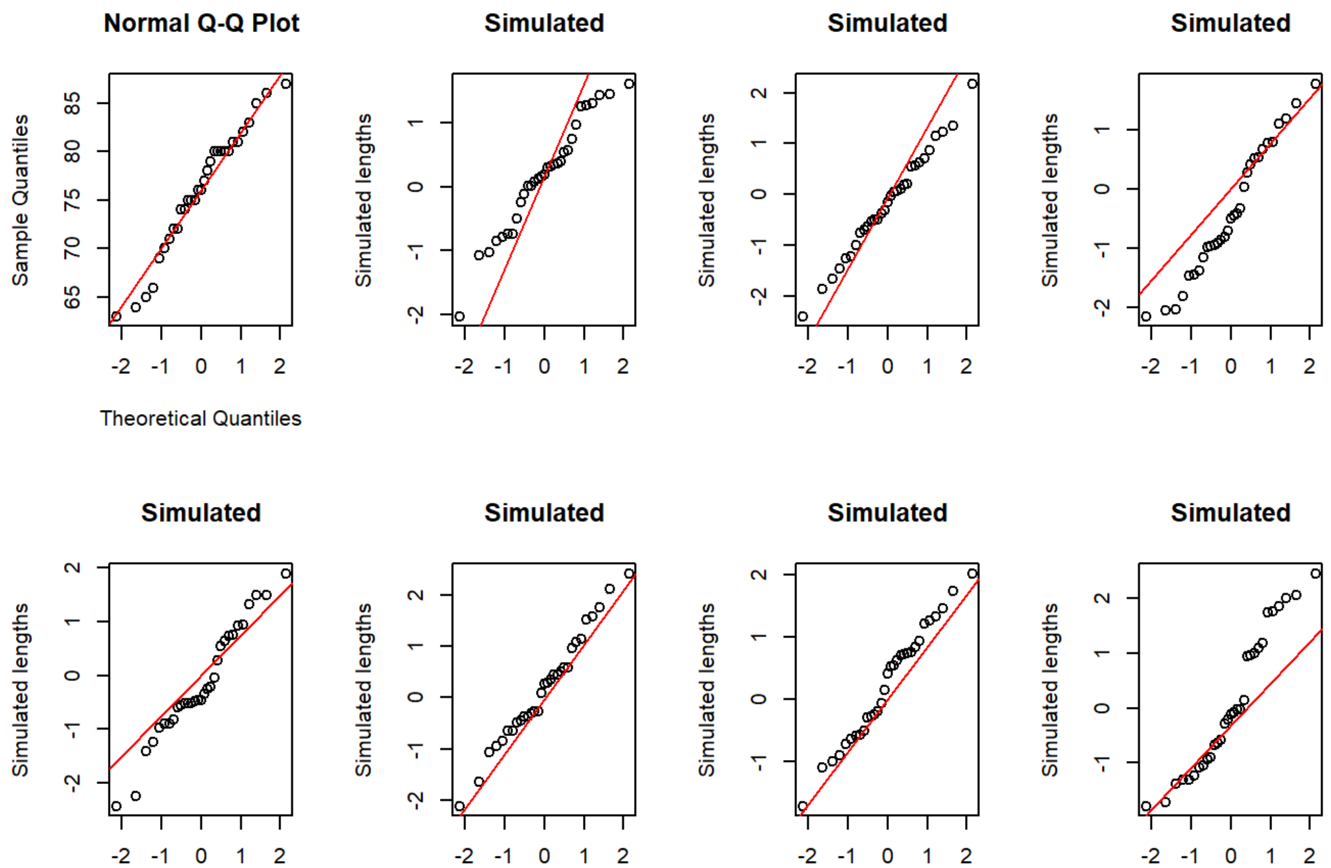

```
qqnorm(trees$Volume)
qqline(trees$Volume, col = 'red')
```




```
log.Volume <- log(trees$Volume)
qqnorm(log.Volume)
qqline(log.Volume, col = 'red')
```

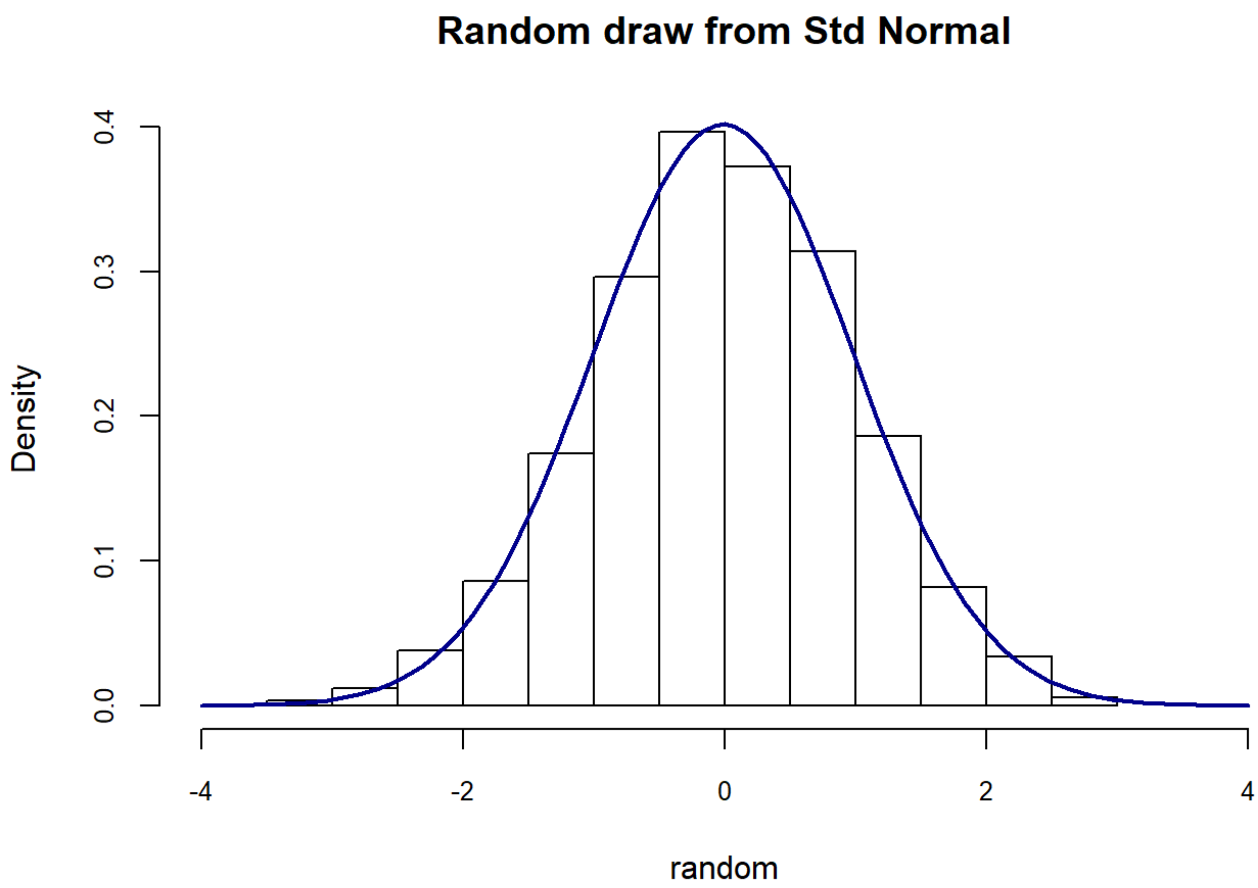



```
par(mfrow = c(2,4))
qqnorm(trees$Height)
qqline(trees$Height, col = 'red')
for (i in 1:7){ qqnorm(rnorm(31),xlab="", ylab="Simulated lengths",
                main="Simulated")
qqline(rnorm(31), col = 'red')
}
```



histogram for rnorm and dnorm

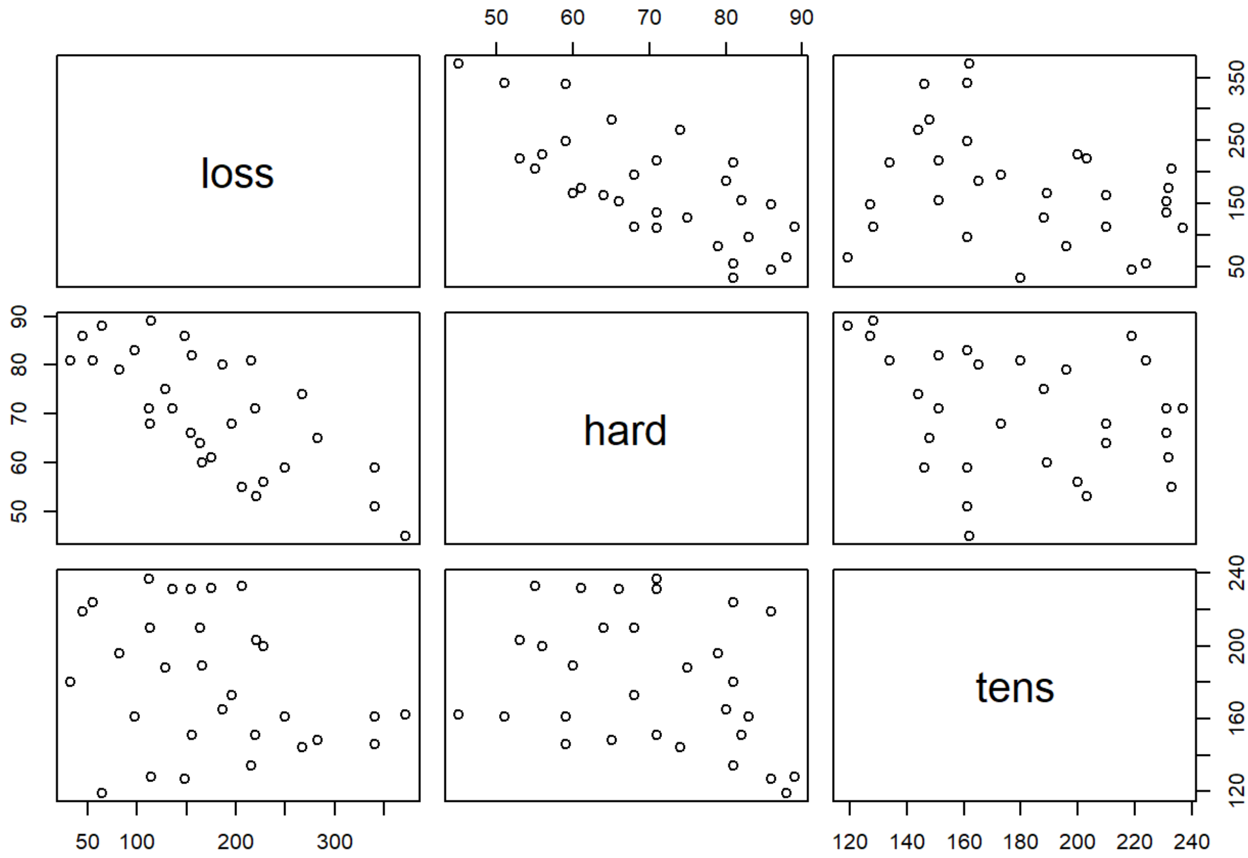
```
par(mfrow=c(1,1))
random <-  rnorm(1000,0,1)

hist(random, main="Random draw from Std Normal", cex.axis=.8, freq = F, xlim = c(-4
,4))
curve(dnorm(x, mean(random), sd(random)), add=TRUE, col="darkblue", lwd=2)
```



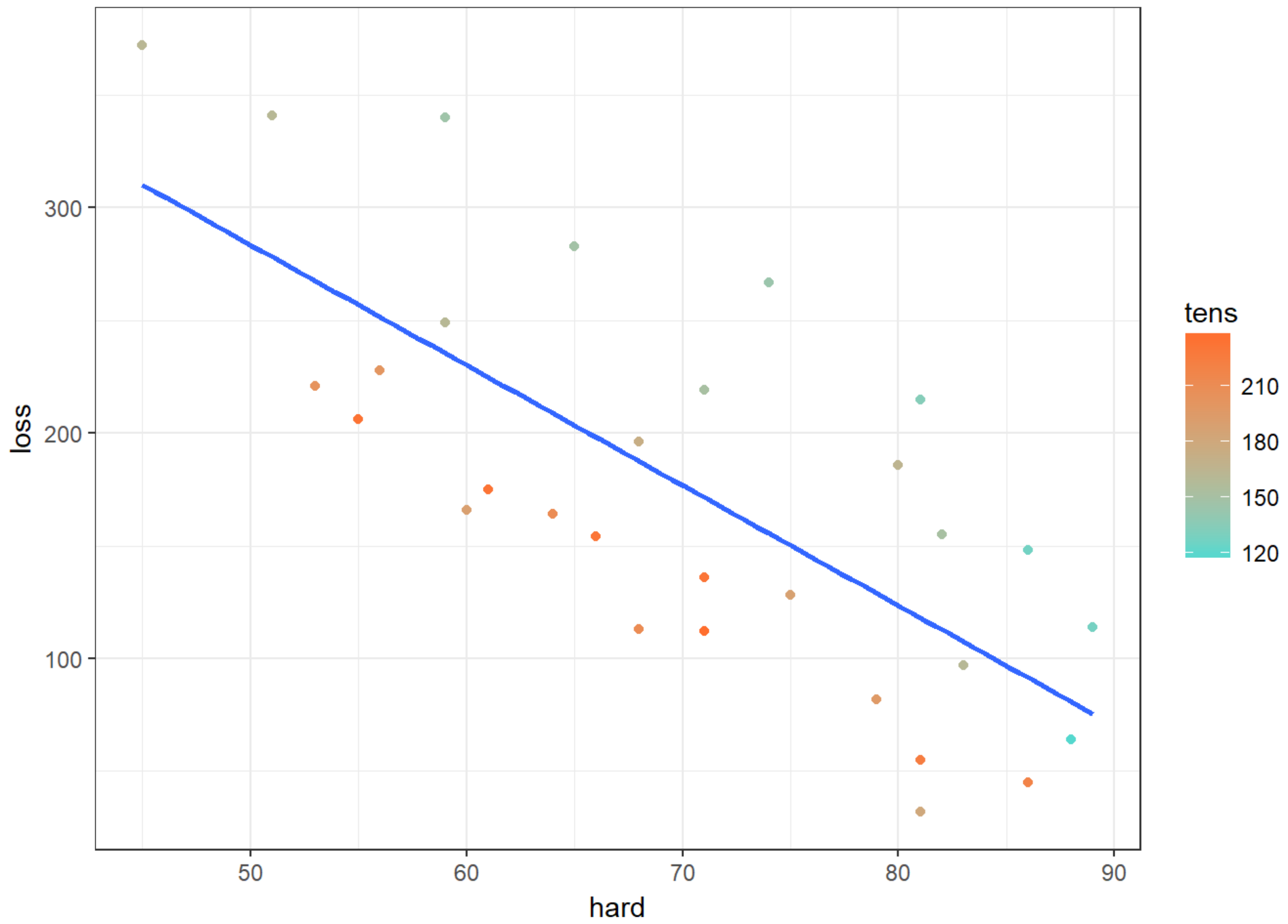
Rubber

Let us view and plot rubber dataset

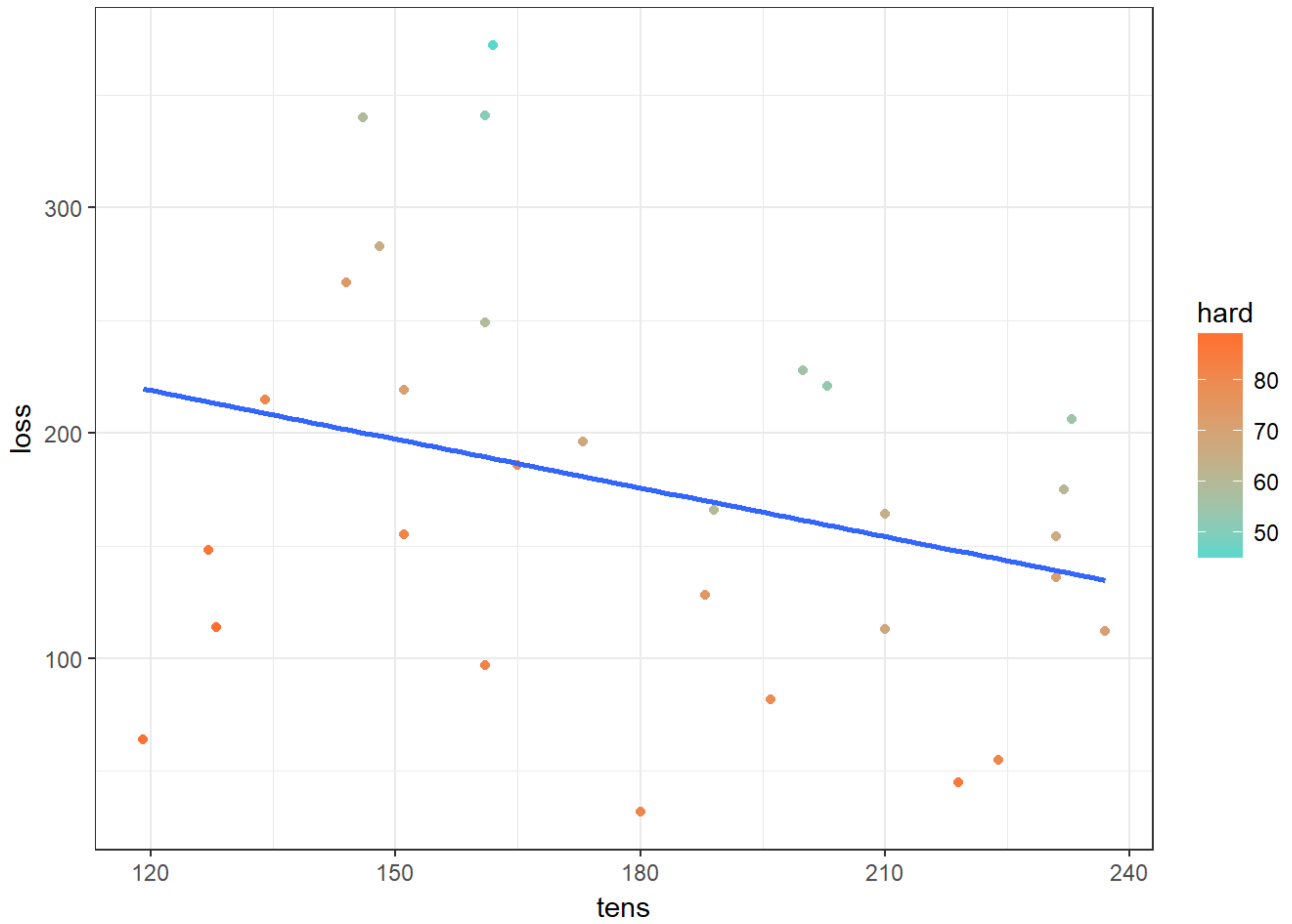
```
library(MASS)

plot(Rubber)
```




```
ggplot(Rubber, aes(hard, loss)) + geom_point(aes(color = tens)) + theme_bw() +geom_
smooth(method = 'lm', formula = y ~x, se = F) + scale_color_continuous(high = "#FF6
E2E", low = "#55D8CE")
```




```r
ggplot(Rubber, aes(tens, loss)) + geom_point(aes(color = hard)) + theme_bw() +geom_smooth(method = 'lm', formula = y ~x, se = F) + scale_color_continuous(high = "#FF6E2E", low = "#55D8CE")
```



We can see a negative correlation between loss and hard/tens

```
Rubber.lm <- lm(loss~hard+tens, data=Rubber)

summary(Rubber.lm)
```

```
##
## Call:
## lm(formula = loss ~ hard + tens, data = Rubber)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.385 -14.608   3.816  19.755  65.981
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 885.1611    61.7516  14.334 3.84e-14 ***
## hard         -6.5708     0.5832 -11.267 1.03e-11 ***
## tens         -1.3743     0.1943  -7.073 1.32e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.49 on 27 degrees of freedom
## Multiple R-squared:  0.8402, Adjusted R-squared:  0.8284
## F-statistic:    71 on 2 and 27 DF,  p-value: 1.767e-11
```

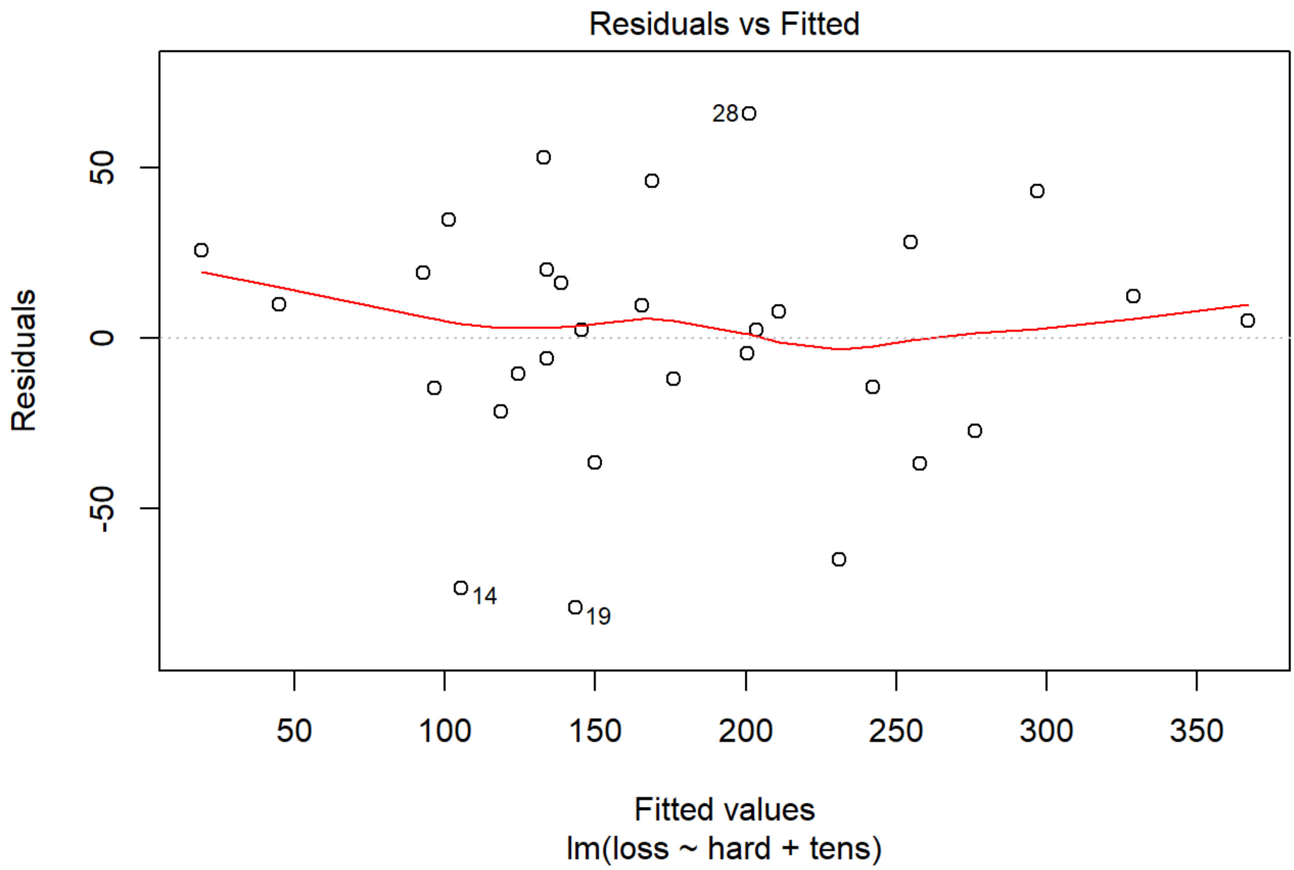
The intercept 885.16 indicates the loss at hardness = 0 and tensile strength = 0 The increase in hardness by 1 will lead to decrease in loss by 6.57 The increase in tensile strength by 1 will lead to decrease in loss by

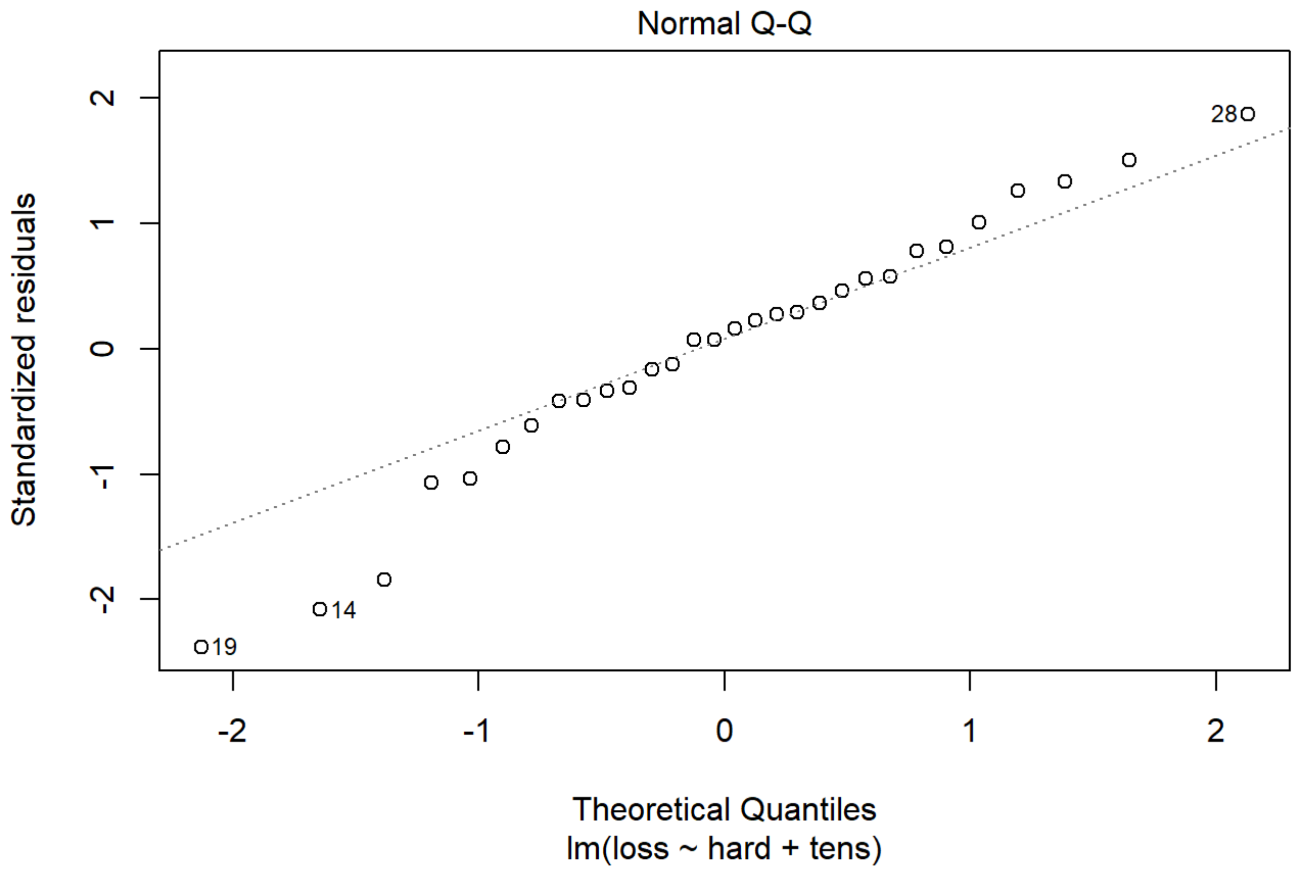
1.37 The p value is 0.001 which indicates the confidence of 99.999

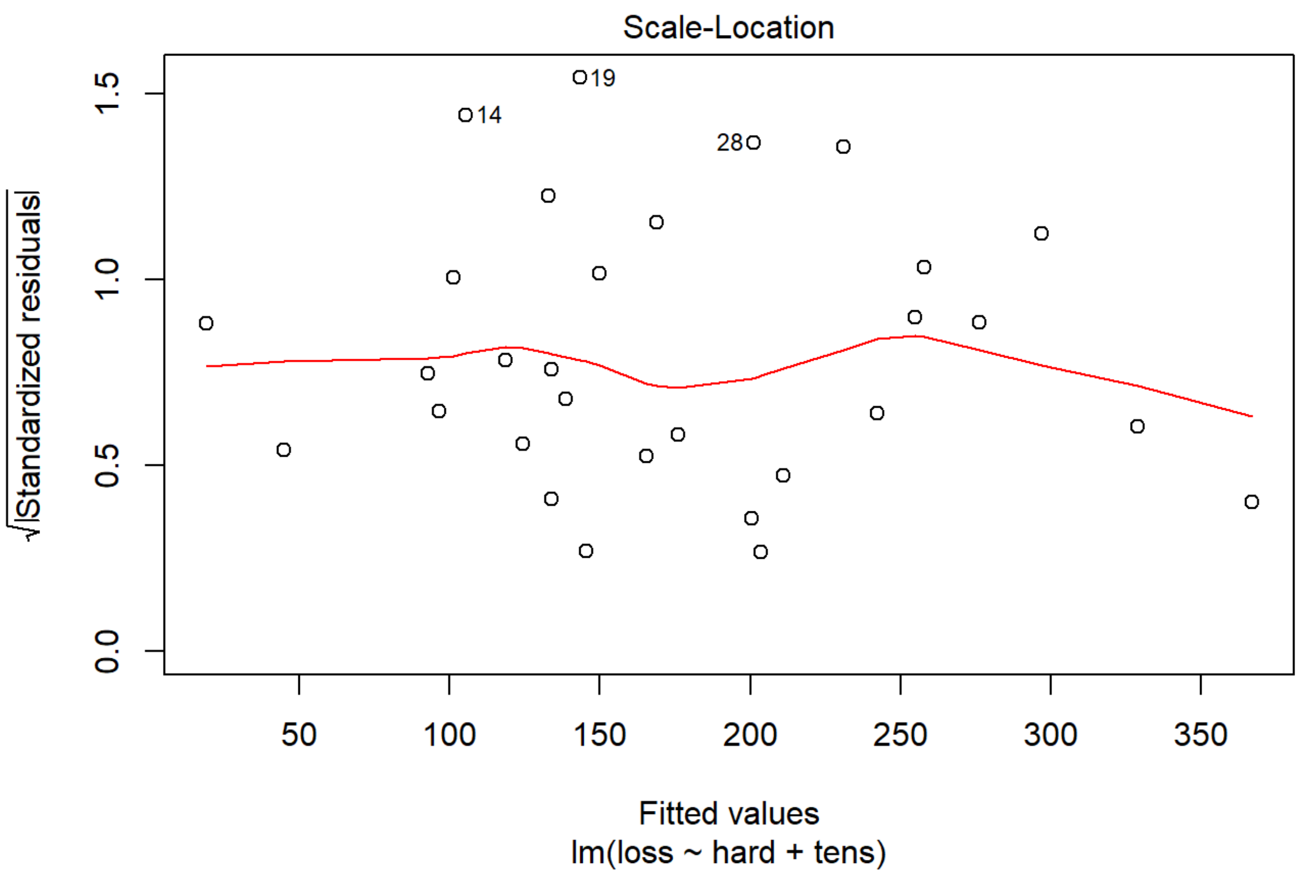
There is a negative corelation between loss and hardness/tensile strength

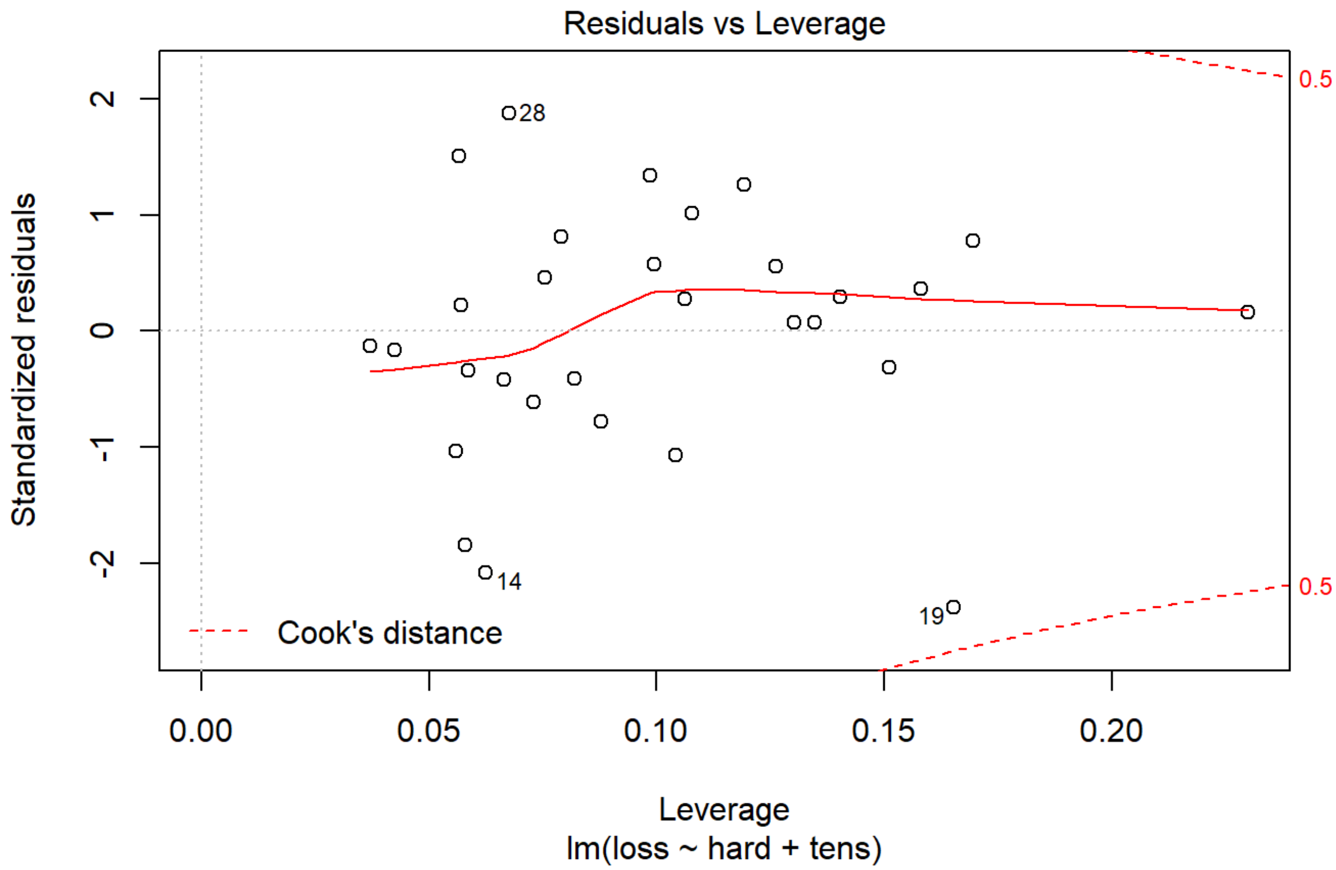
Let us plot the linear model using plot and termplot

```
plot(Rubber.lm)
```

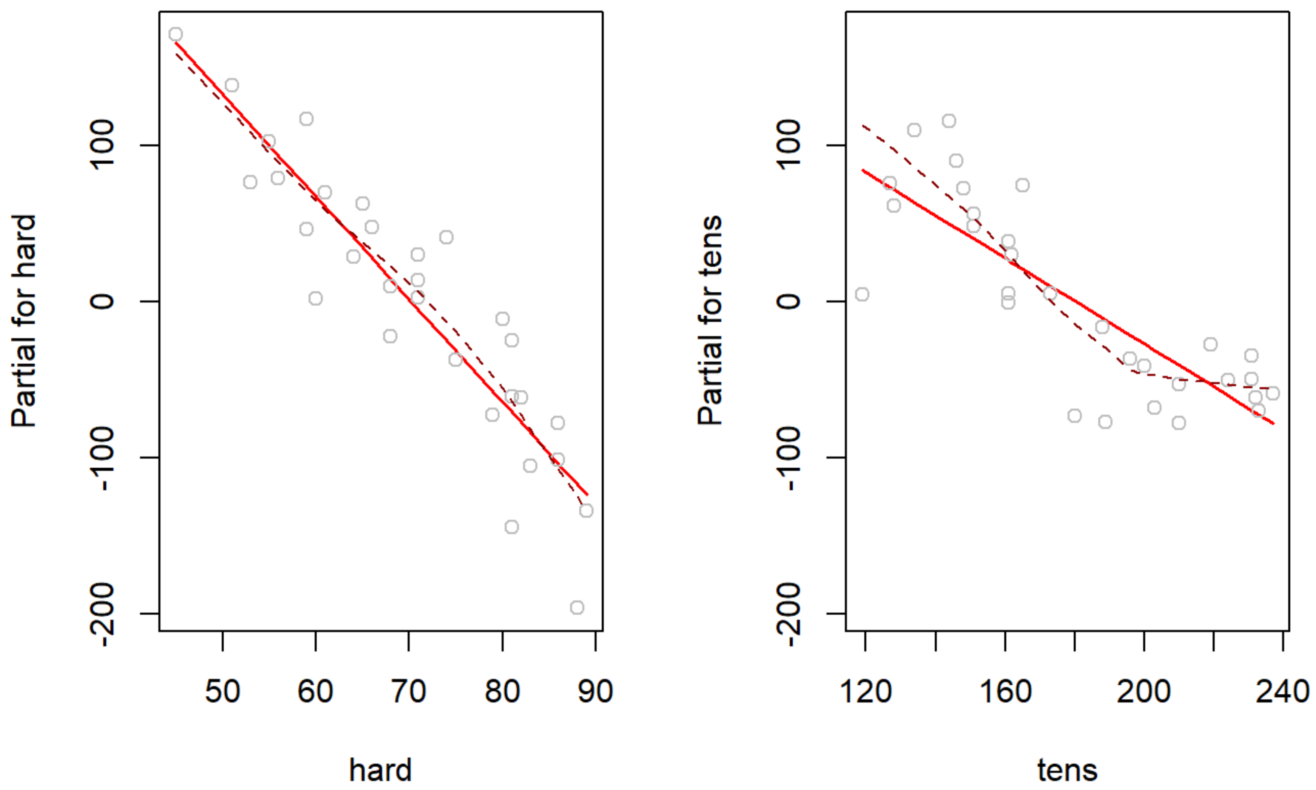



Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(loss ~ hard + tens)
28
14
19
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
lm(loss ~ hard + tens)
19
14
28
0.0
0.5
1.0
1.5
50
100
150
200
250
300
350




```
par(mfrow=c(1,2))
termplot(Rubber.lm, partial=TRUE, smooth=panel.smooth)
```




```
par(mfrow=c(1,1))
```

Oddbooks

Let us install the package DAAG which consists of Oddbooks dataset

```
library(DAAG)
```

```
## Warning: package 'DAAG' was built under R version 3.4.4
```

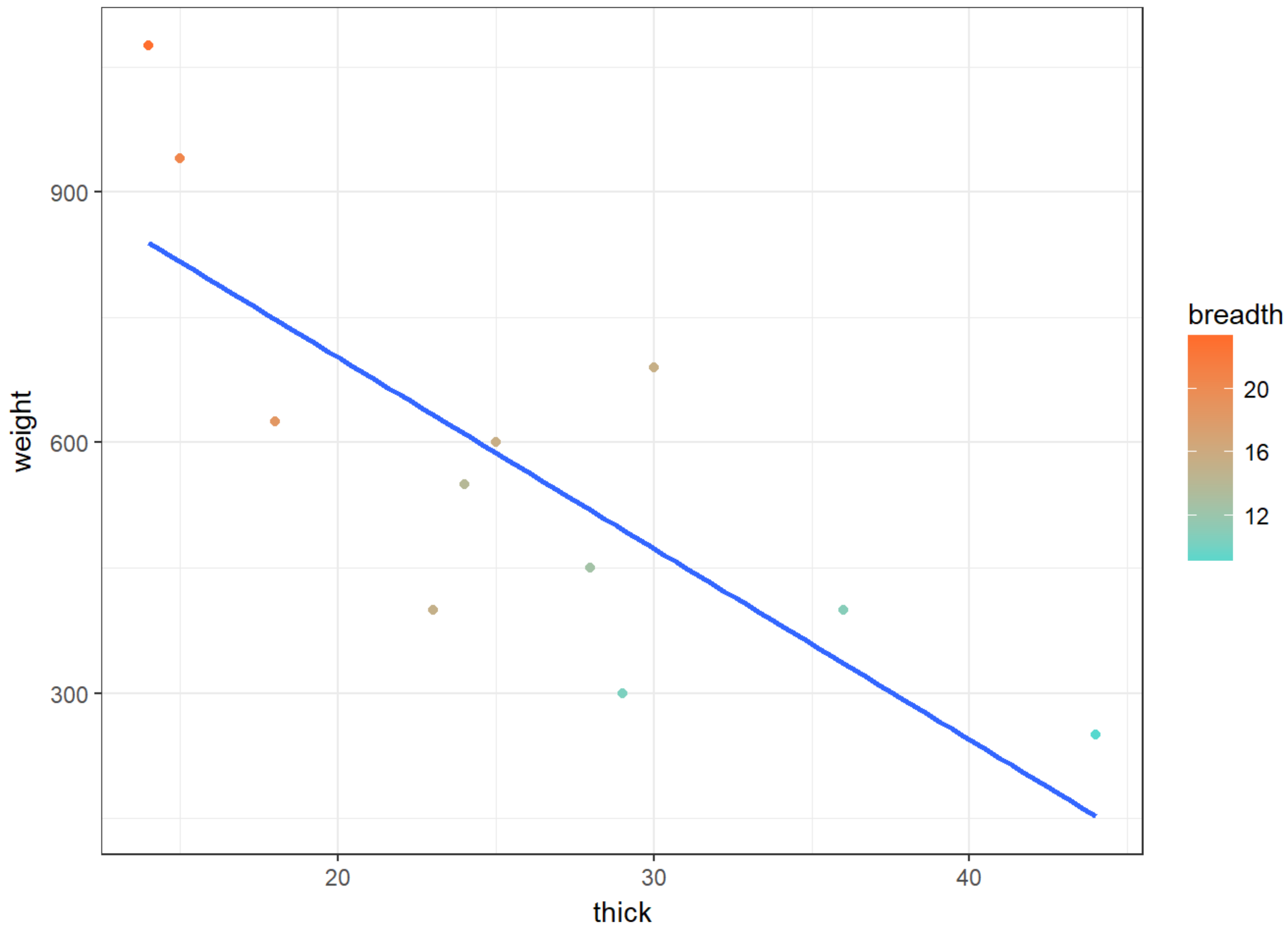
```
## Loading required package: lattice
```

```
##
## Attaching package: 'DAAG'
```

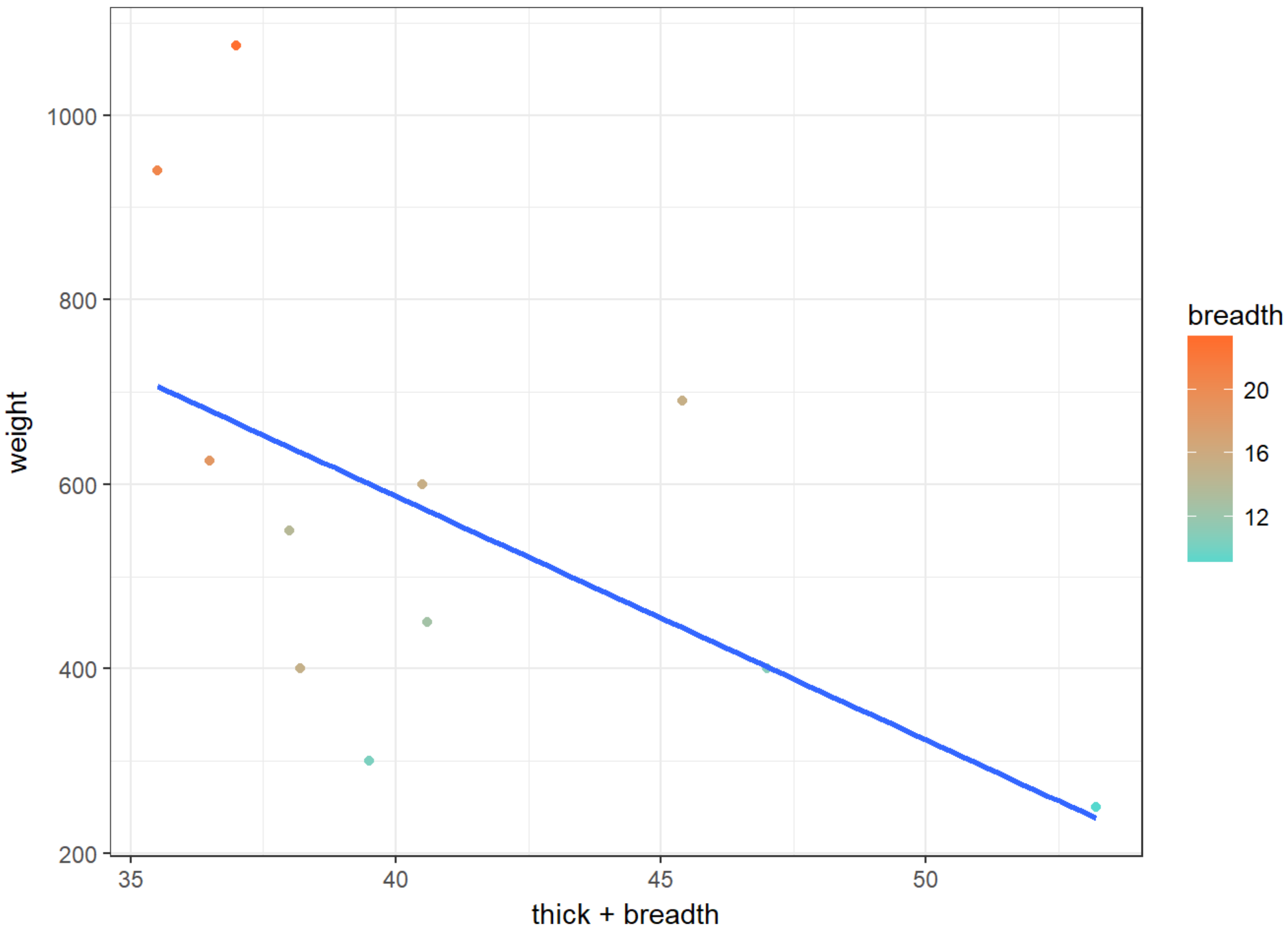
```
## The following object is masked from 'package:MASS':
##
##     hills
```

Let us plot using ggplot

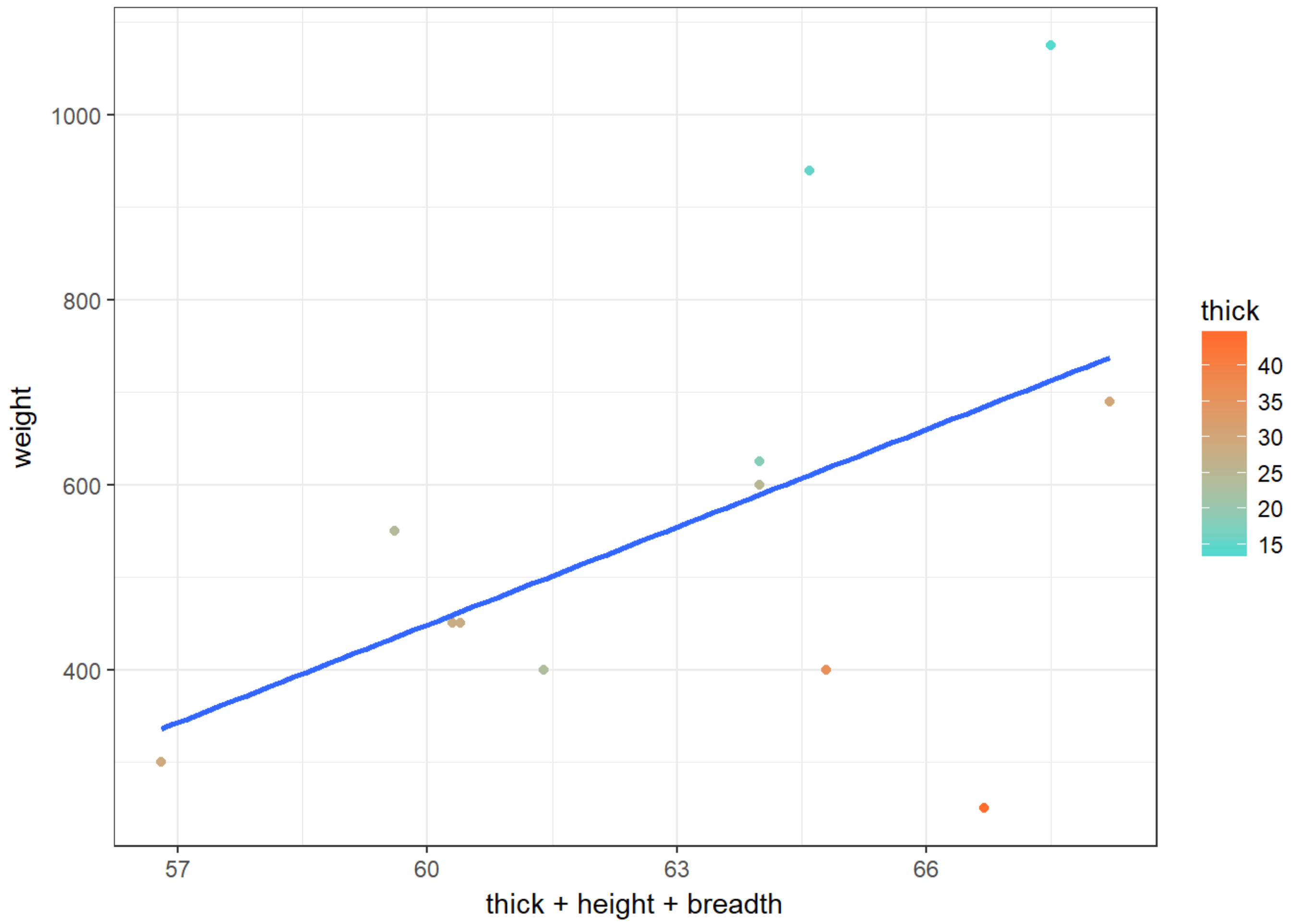
```
ggplot(oddbooks, aes(thick, weight)) + geom_point(aes(color = breadth)) + theme_bw(
) +geom_smooth(method = 'lm', formula = y ~x, se = F) + scale_color_continuous(high
= "#FF6E2E", low = "#55D8CE")
```




```
ggplot(oddbooks, aes(thick+ breadth, weight)) + geom_point(aes(color = breadth)) +
theme_bw() +geom_smooth(method = 'lm', formula = y ~x, se = F) + scale_color_contin
uous(high = "#FF6E2E", low = "#55D8CE")
```




```
ggplot(oddbooks, aes(thick+height+breadth, weight)) + geom_point(aes(color = thick)
) + theme_bw() +geom_smooth(method = 'lm', formula = y ~x, se = F) + scale_color_co
ntinuous(high = "#FF6E2E", low = "#55D8CE")
```



Using linear model on oddbooks

```
logbooks <- log(oddbooks)

logbooks.lm1 <- lm(weight~thick,data=logbooks)
summary(logbooks.lm1)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   9.691988  0.7076002 13.696983 8.345461e-08
## thick        -1.072579  0.2190487 -4.896534 6.263390e-04
```

There is a negative correlation between weight and thickness At 0 thickness the weight is 9.69 and with the increase in Weight by 1 the thickness will decrease by 1.07 This gives us a very weird result as generally weight should increase with increase in thickness

```
logbooks.lm2<-lm(weight~thick+height,data=logbooks)
 summary(logbooks.lm2)$coef
```

```
##               Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -1.2631920  3.5520303 -0.3556253 0.73031392
## thick        0.3129265  0.4723981  0.6624212 0.52429995
## height       2.1143070  0.6782222  3.1174254 0.01236986
```

There is a negative correlation between weight and thickness/height At 0 thickness/height the weight is -1.26 and with the increase in Weight by 1 the thickness will increase by 0.3 and height by 2.11 However we can see a low confidence interval for thickness and Weight

```
logbooks.lm3<-lm(weight~thick+height+breadth,data=logbooks)
  summary(logbooks.lm3)$coef
```

```
##                Estimate Std. Error    t value  Pr(>|t|)
## (Intercept) -0.7191177   3.216233 -0.2235900 0.8286803
## thick        0.4647506   0.434447  1.0697521 0.3159421
## height       0.1536690   1.273404  0.1206758 0.9069237
## breadth      1.8771865   1.069562  1.7550980 0.1173191
```

There is a negative correlation between weight and thickness/height/breadth At 0 thickness/height the weight is -0.71 and with the increase in Weight by 1 the thickness will increase by 0.46 and height by 0.15 and breadth by 1.87 However we can see a low confidence intervals

We can see very different results from oddboks dataset as books may be chosen in such a way to produce odd results