

Data Analytics & visualization CS634(LAB_2)

Rathod Dharmraj 202061008 Mtech CSE

1/31/2021

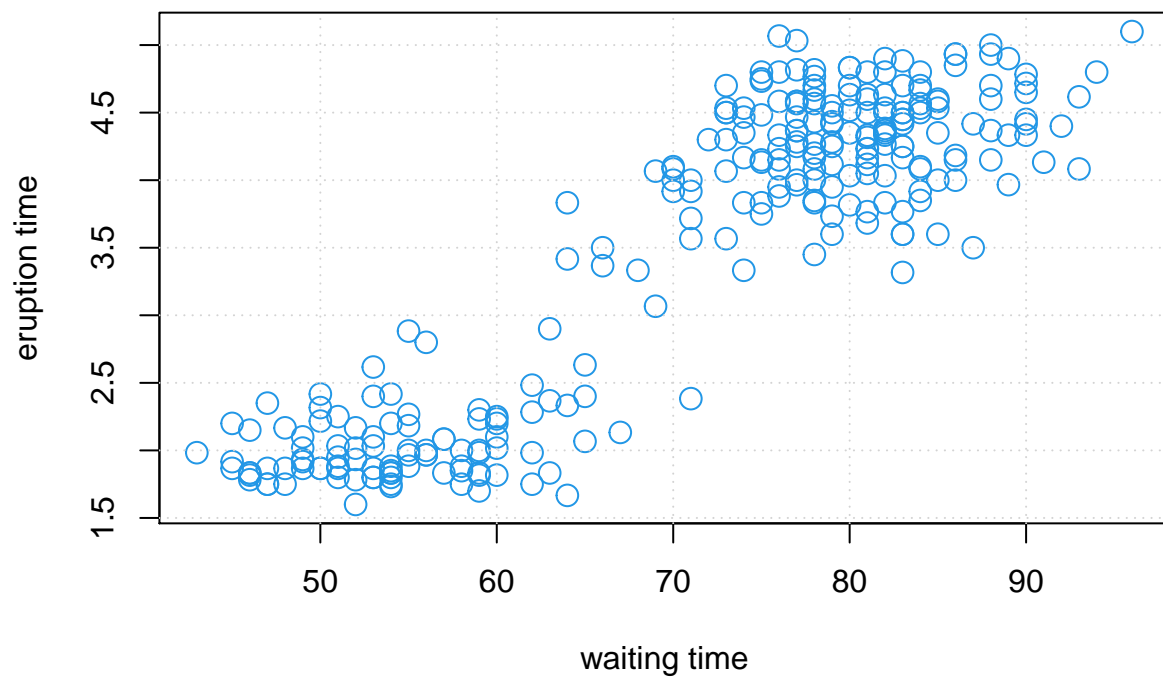
Question (1)

Use various dataset available in R to Construct following plots:

- (1) Scatter Plot
- (2) Violin Plot
- (3) Q-Q Plot

Dataset = faithful

```
#faithful  
plot(faithful$waiting, faithful$eruptions, pch = 21, col = 4, cex = 1.5, xlab = "waiting time", ylab = "eruption time",  
grid())
```

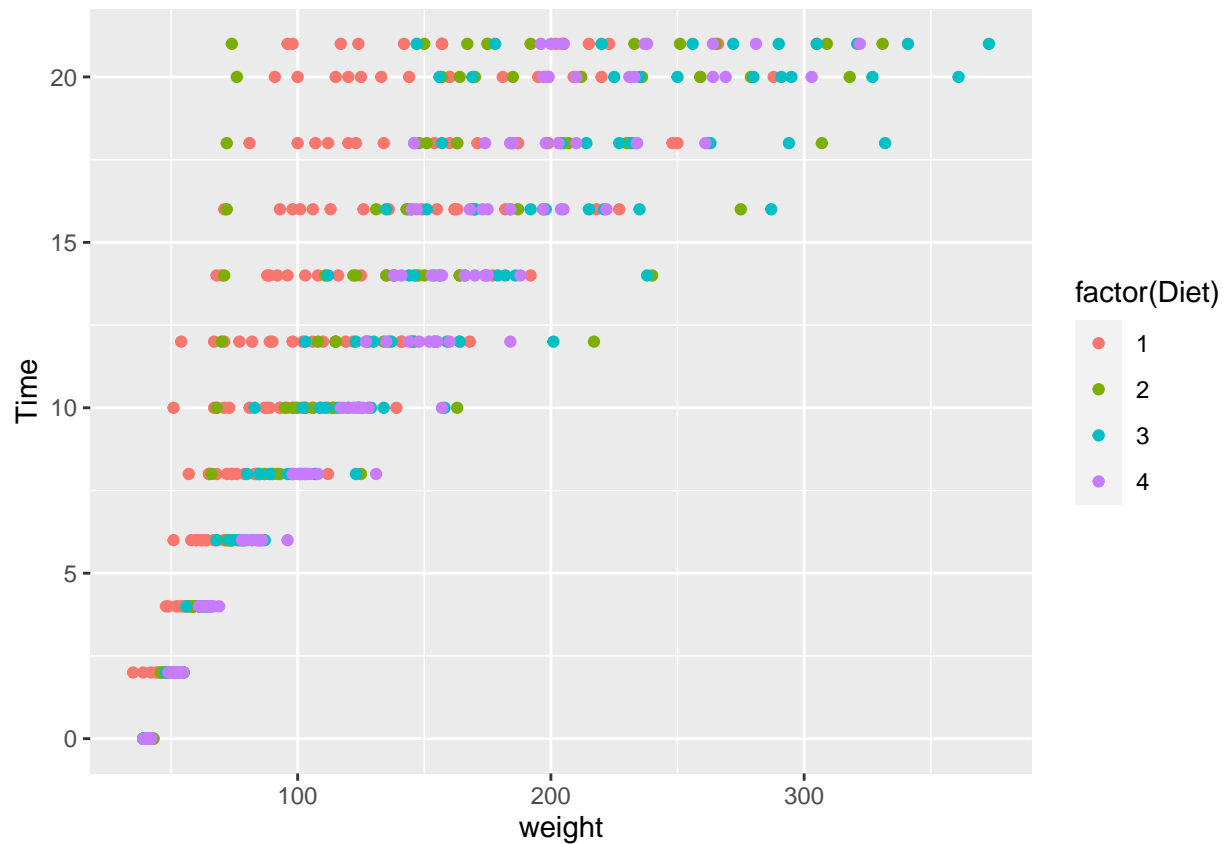


we can see there are two clear clusters. The first cluster is of shorter wait time so, shorter eruption time. The second cluster is of longer wait time and as a result, a longer eruption time. “if the geyser waits a longer time

before erupting, more pressure builds up and when it is finally released there will be longer eruption”.

#ChickWeight

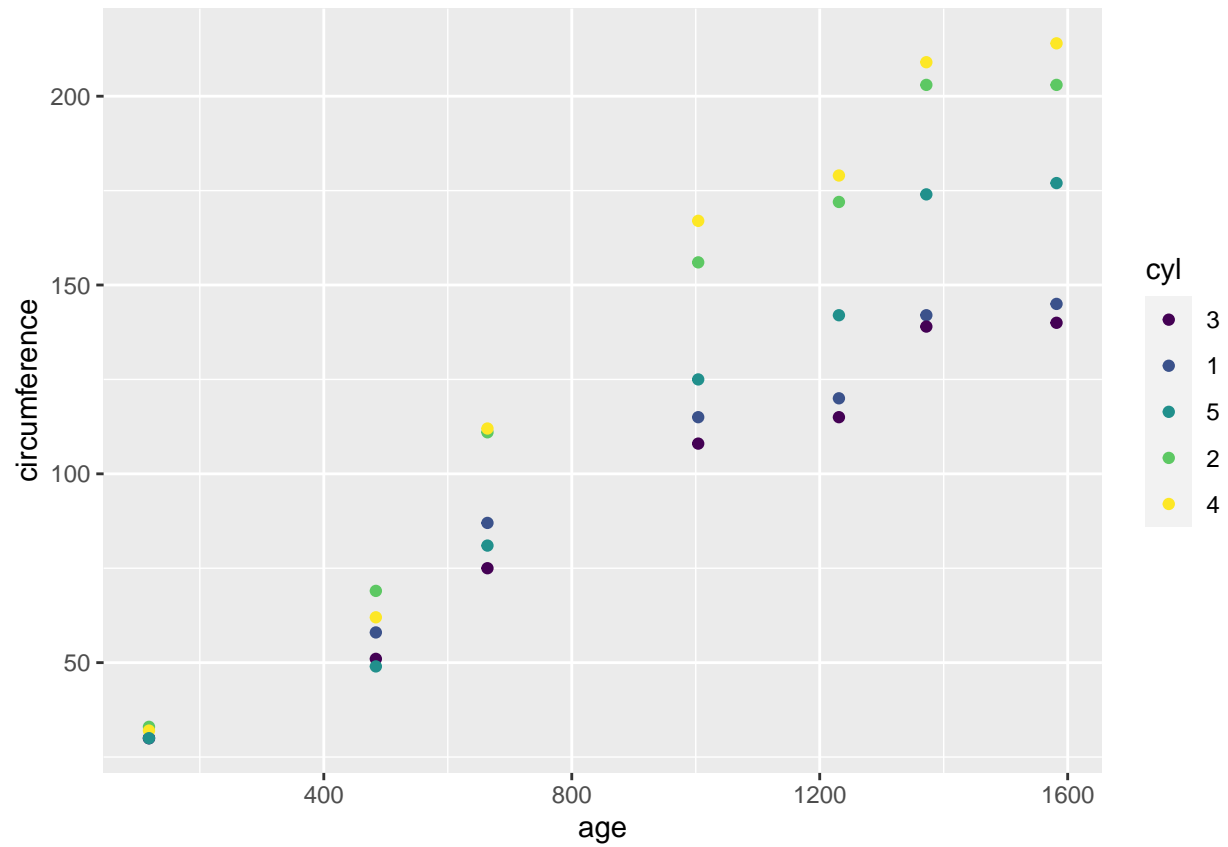
```
library(ggplot2)
ggplot(ChickWeight, aes(x = weight, y = Time, color = cyl)) +
  geom_point(aes(color = factor(Diet)))
```



plot is showing that how the different diet differ with time.

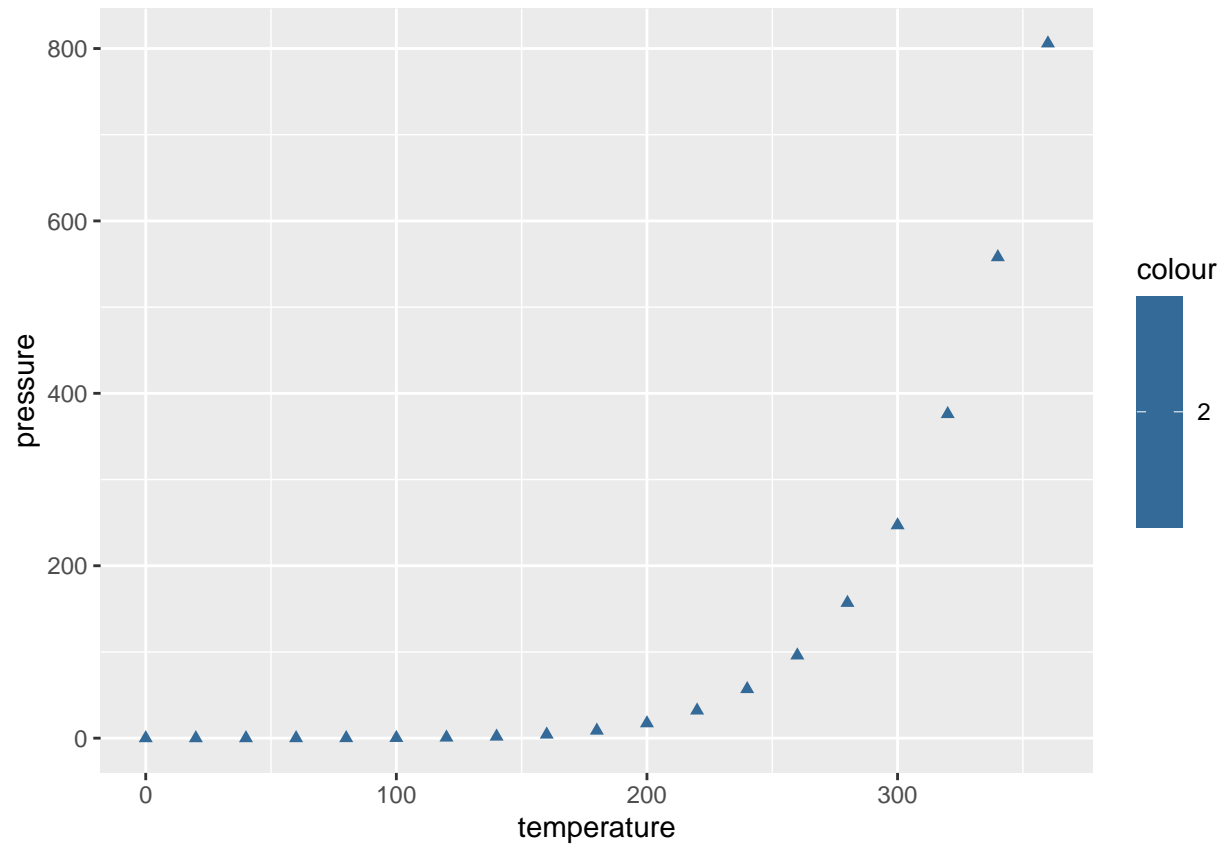
#Orange

```
library(ggplot2)
ggplot(Orange, aes(x = age, y = circumference, color = cyl)) +
  geom_point(aes(color = factor(Tree)))
```



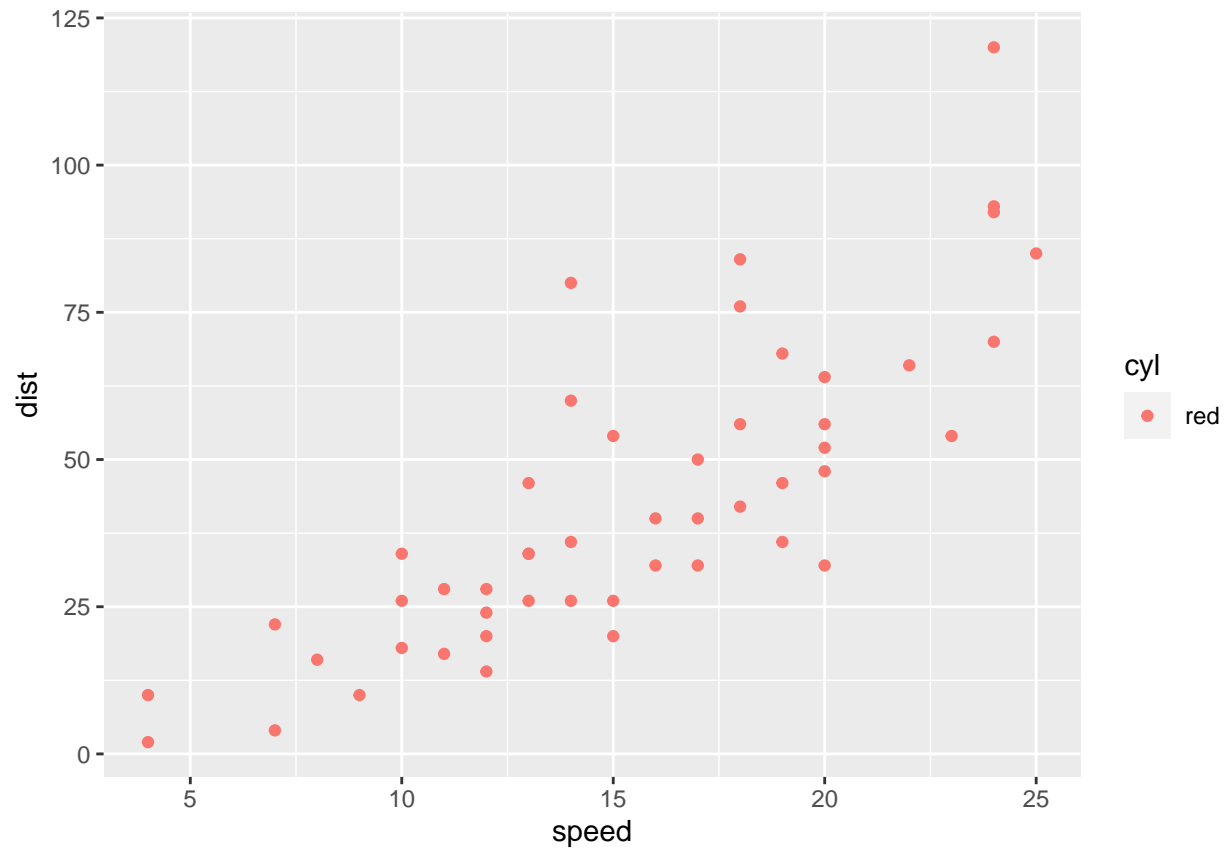
with this plot we can understand how circumference of the tree is increasing with its age.

```
#pressure
library(ggplot2)
ggplot(pressure, aes(x =temperature , y = pressure)) +
geom_point(aes(color = 2),pch =17)
```



with this plot we can understnad that after 200 temperature the pressure is increasing exponentially.

```
#cars
library(ggplot2)
ggplot(cars, aes(x =speed , y = dist,color=cyl)) +
geom_point(aes(color='red'),pch =19)
```



here we can see scatter plot distance vs speed. how much ft cars is going after pressing the break.

```
library(vioplot)
```

```
## Loading required package: sm
```

```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

```
## Loading required package: zoo
```

##

```
## Attaching package: 'zoo'
```

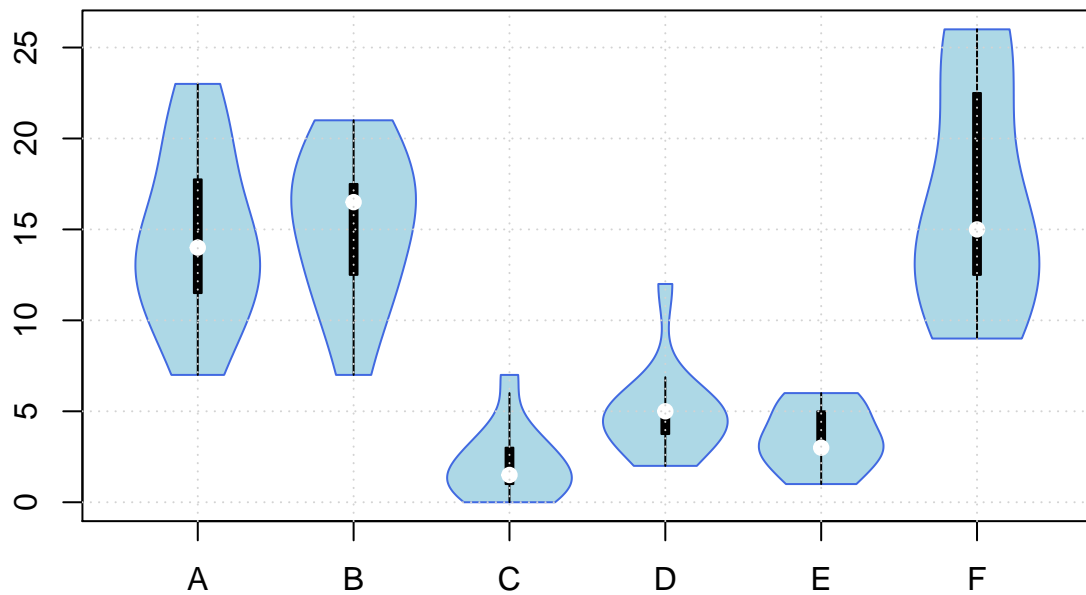
```
## The following objects are masked from 'package:base':
```

##

```
##      as.Date, as.Date.numeric
```

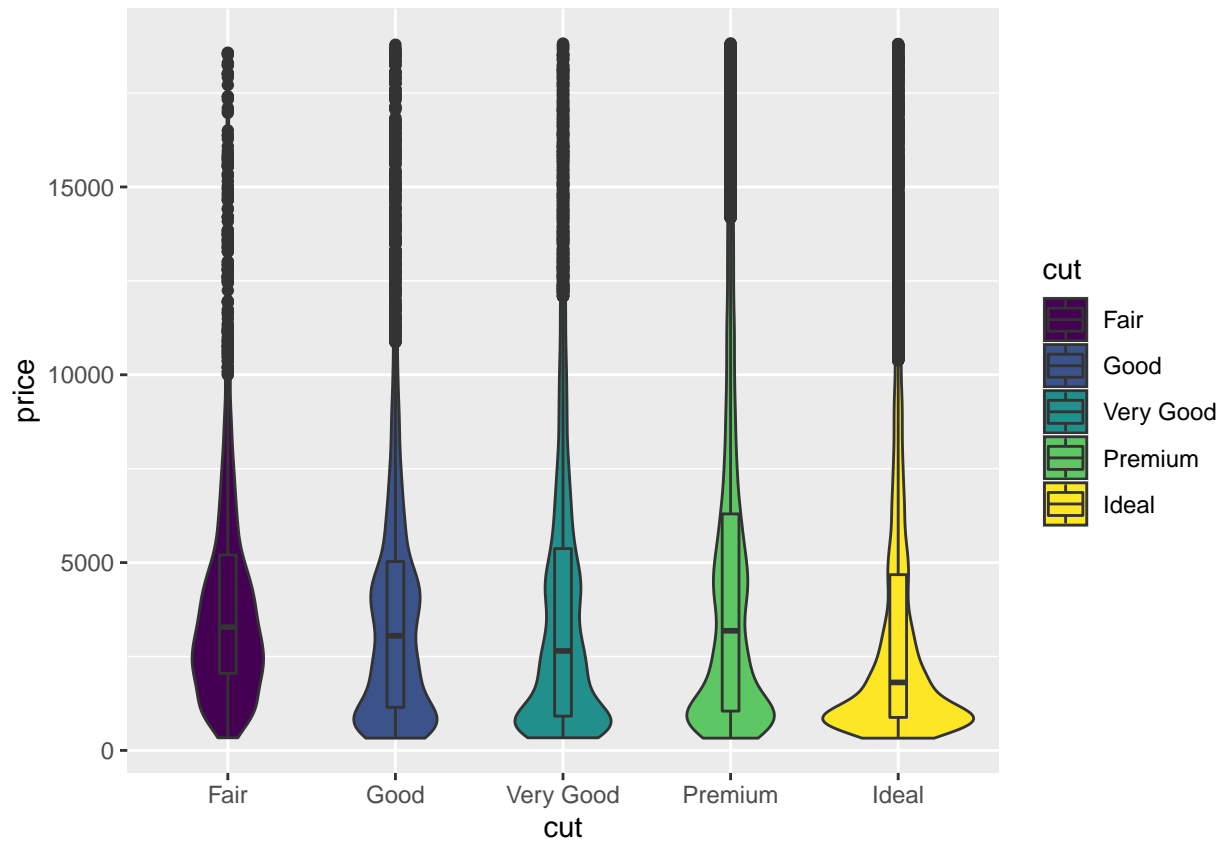
#InsectSprays

```
vioplot(InsectSprays$count[InsectSprays$spray=="A"], InsectSprays$count[InsectSprays$spray=="B"], InsectSprays$count[InsectSprays$spray=="C"], InsectSprays$count[InsectSprays$spray=="D"])
```

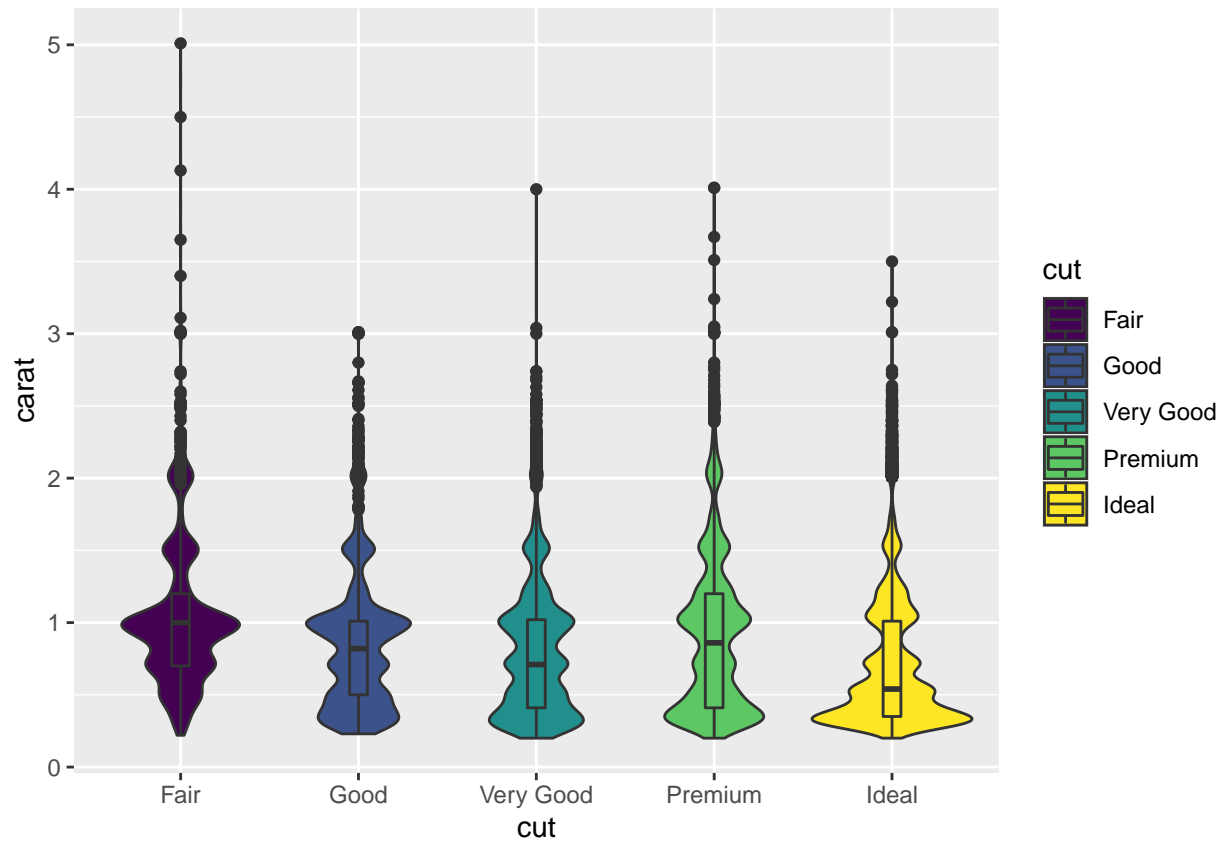


by this violin plot we can say that how the values of one insecticides changes with another what is the distribution of that insecticides and we can compare with others.

```
#Puromycin
library(ggplot2)
ggplot(diamonds,aes(x= cut,y=price,fill = cut))+geom_violin()+geom_boxplot(width = 0.1)
```

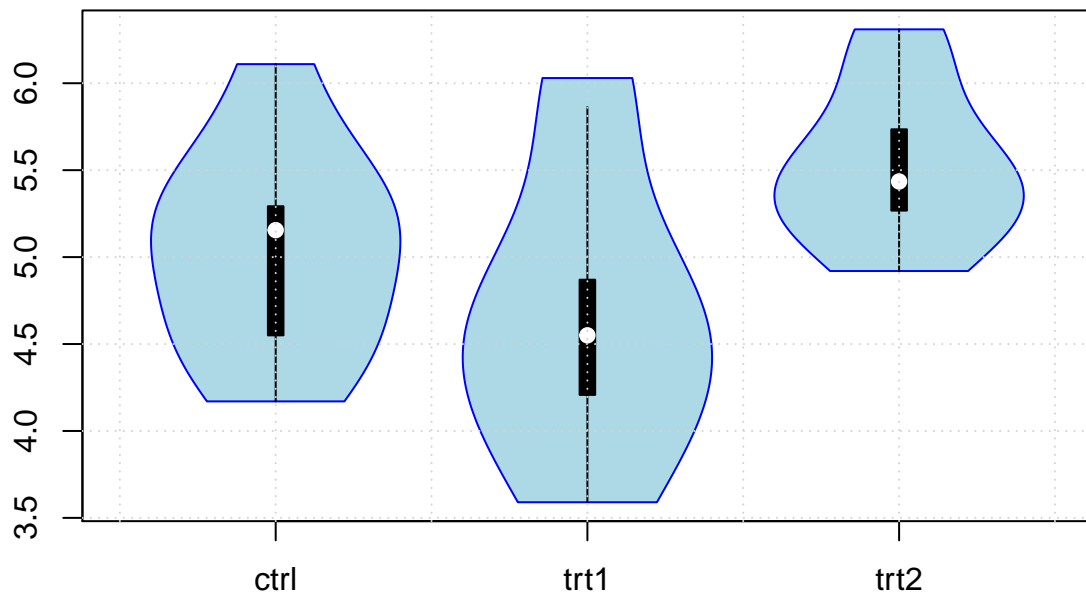


```
ggplot(diamonds,aes(x= cut,y=carat,fill = cut))+geom_violin()+geom_boxplot(width = 0.1)
```



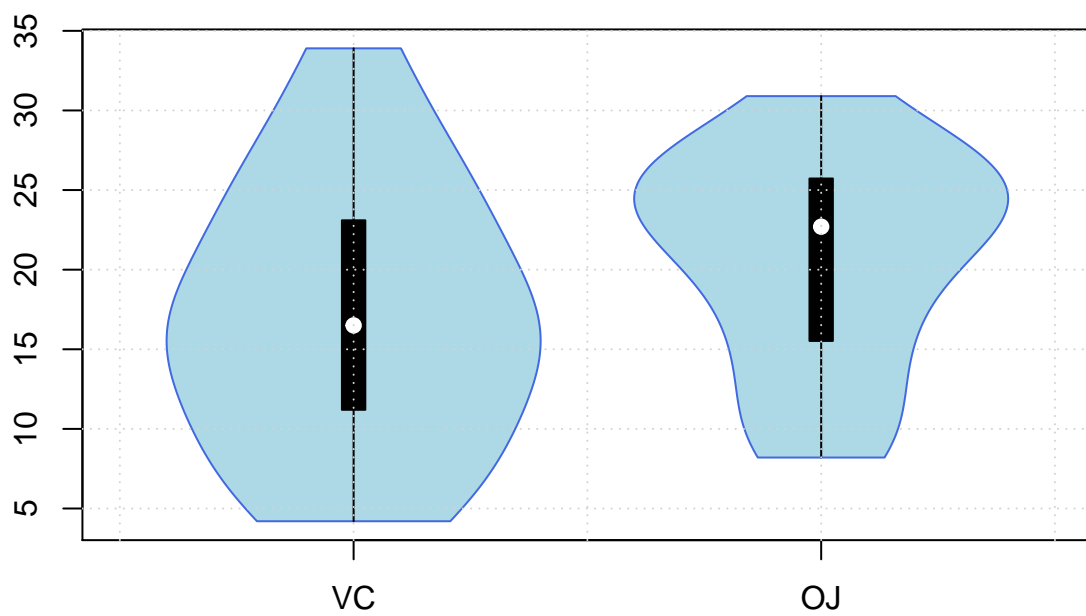
Here in figure 2 we can see outliers and distribution of the cut over carat and in figure one we can see distribution of the cut over price.

```
vioplot(PlantGrowth$weight[PlantGrowth$group=="ctrl"],PlantGrowth$weight[PlantGrowth$group=="trt1"],PlantGrowth$group,
        grid())
```

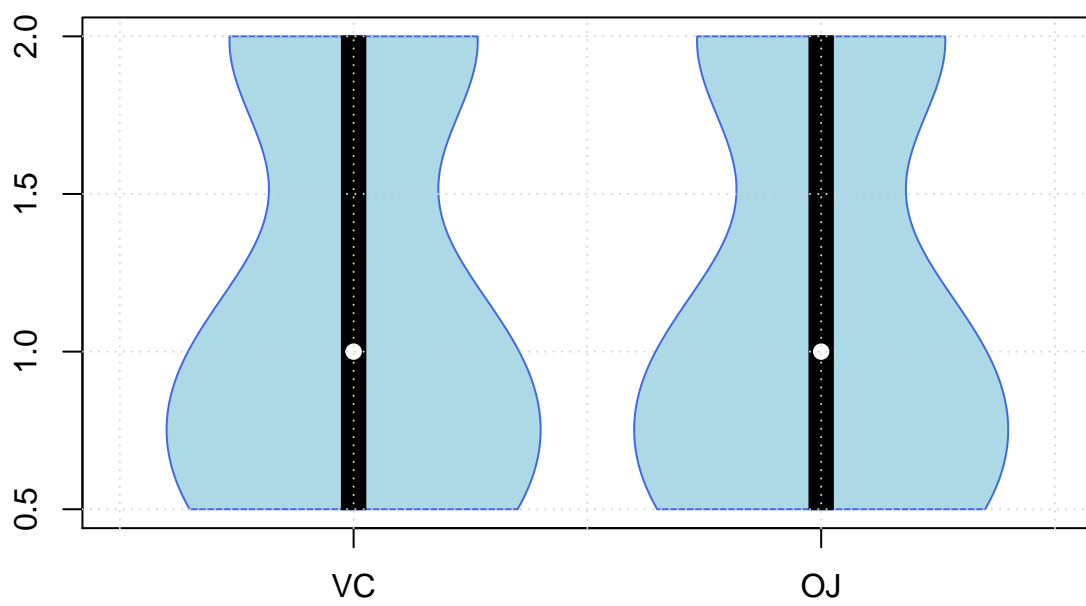



```
#ToothGrowth
```

```
vioplot(ToothGrowth$len[ToothGrowth$supp=="VC"],ToothGrowth$len[ToothGrowth$supp=="OJ"],names=c("VC","OJ"),
grid())
```



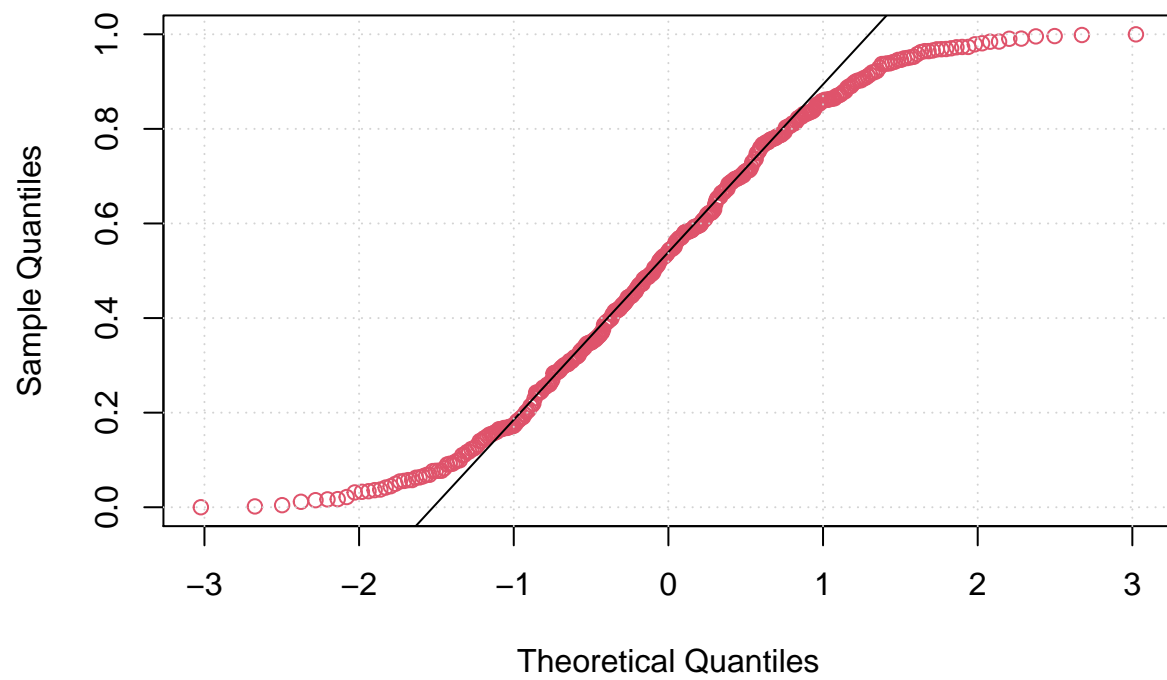
```
vioplot(ToothGrowth$dose[ToothGrowth$supp=="VC"],ToothGrowth$dose[ToothGrowth$supp=="OJ"],names=c("VC",
grid()
```



```
#randu
```

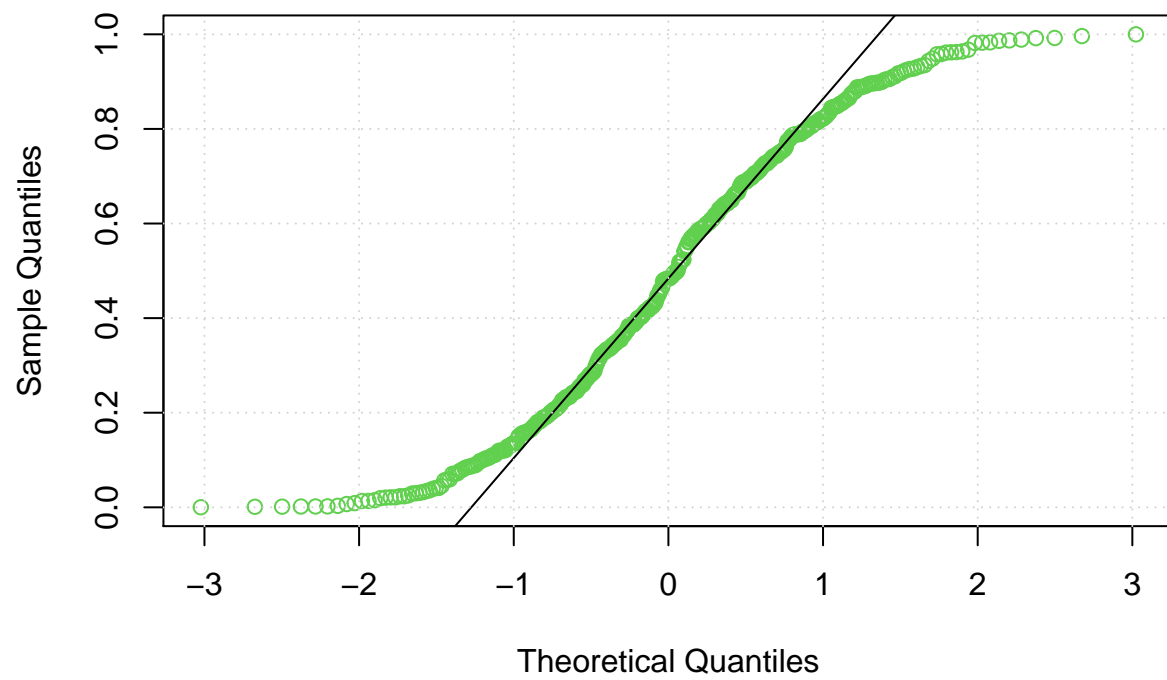
```
qqnorm(randu$x,col=2)  
qqline(randu$x)  
grid()
```

Normal Q-Q Plot



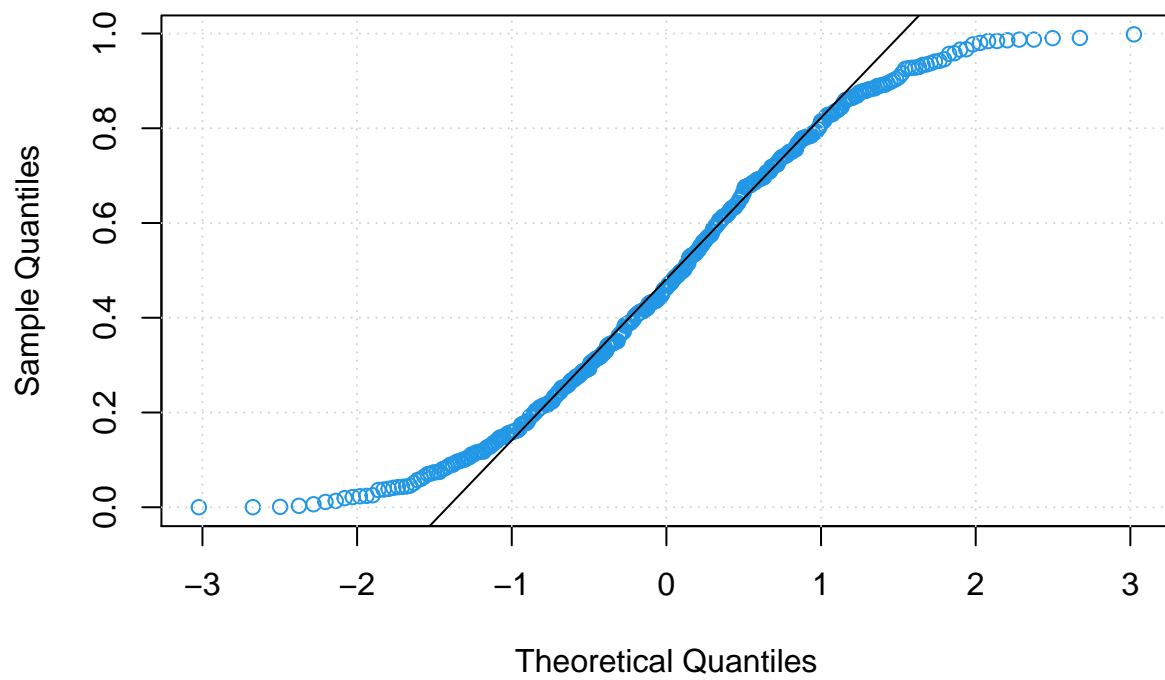
```
qqnorm(randu$y,col=3)  
qqline(randu$y)  
grid()
```

Normal Q-Q Plot



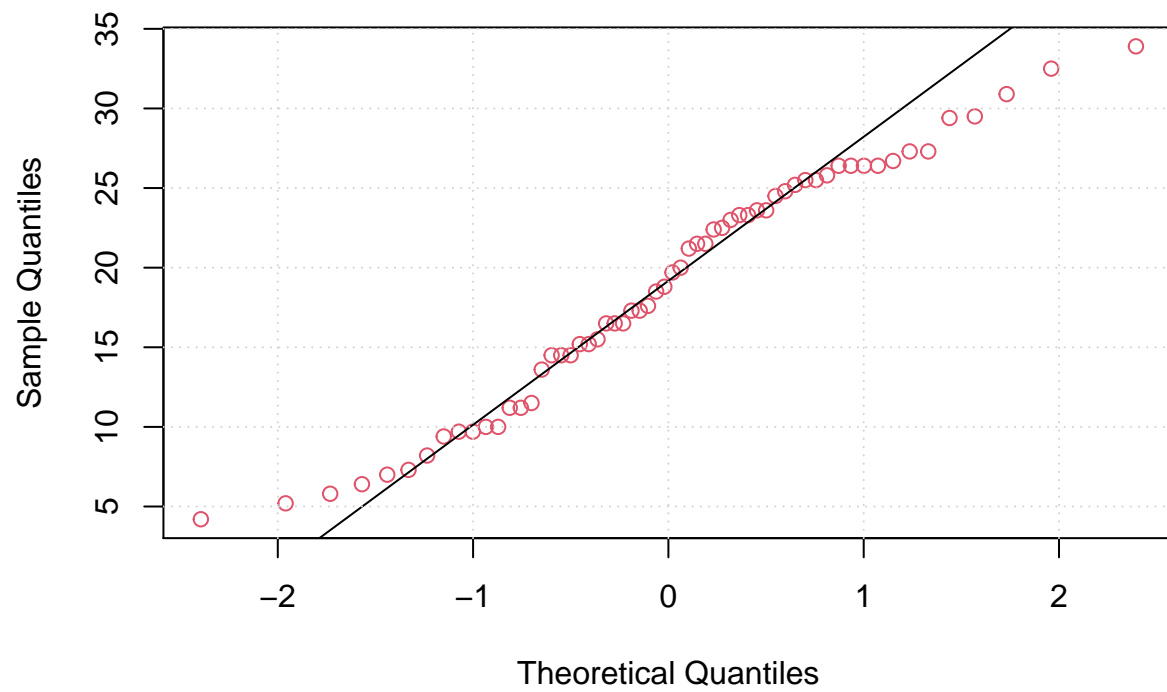
```
qqnorm(randu$z,col=4)  
qqline(randu$z)  
grid()
```

Normal Q-Q Plot



```
#ToothGrowth  
qqnorm(ToothGrowth$len, col=10)  
qqline(ToothGrowth$len)  
grid()
```

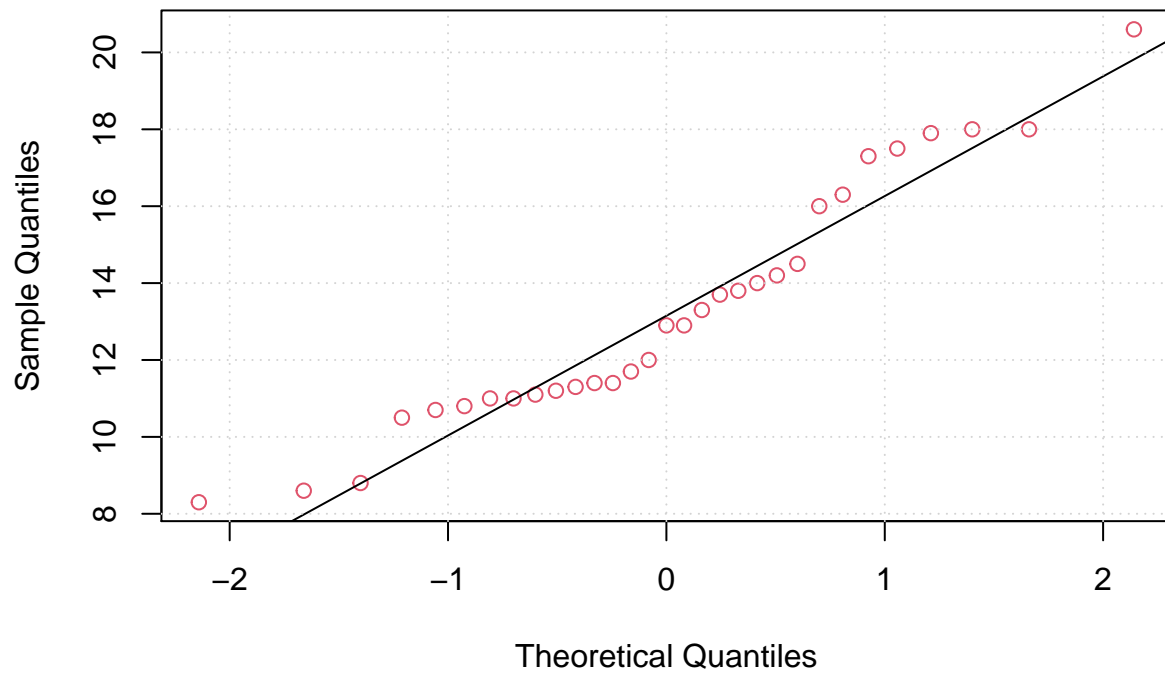
Normal Q-Q Plot



```
#trees
```

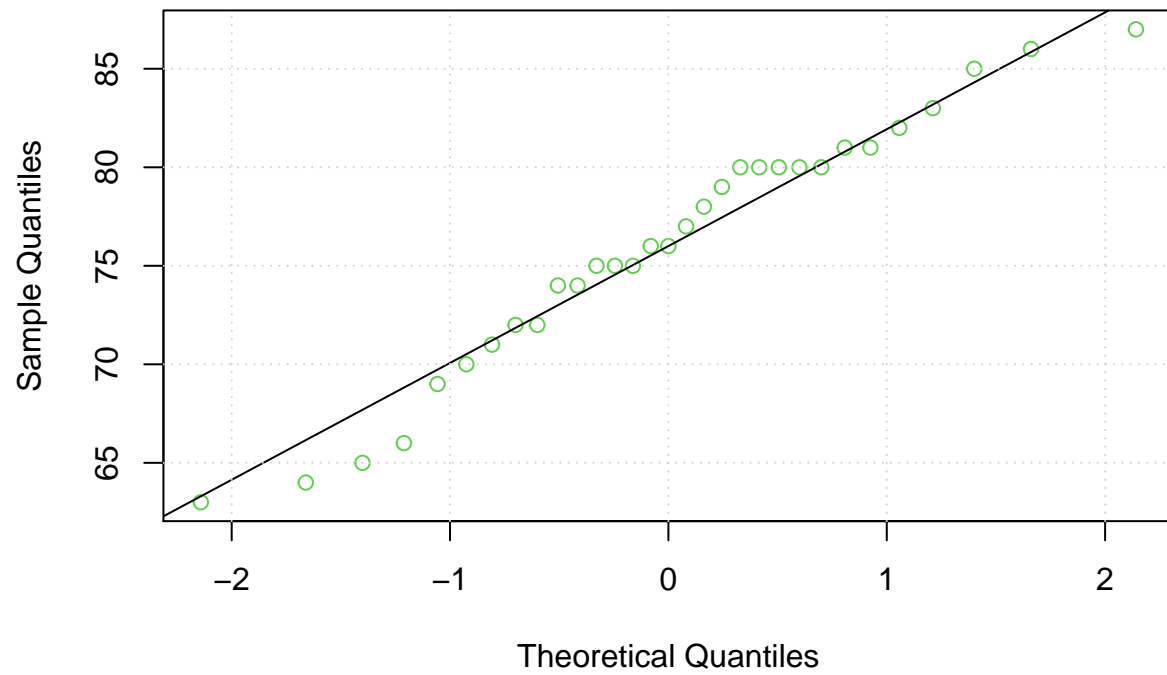
```
qqnorm(trees$Girth,col=2)  
qqline(trees$Girth)  
grid()
```

Normal Q-Q Plot



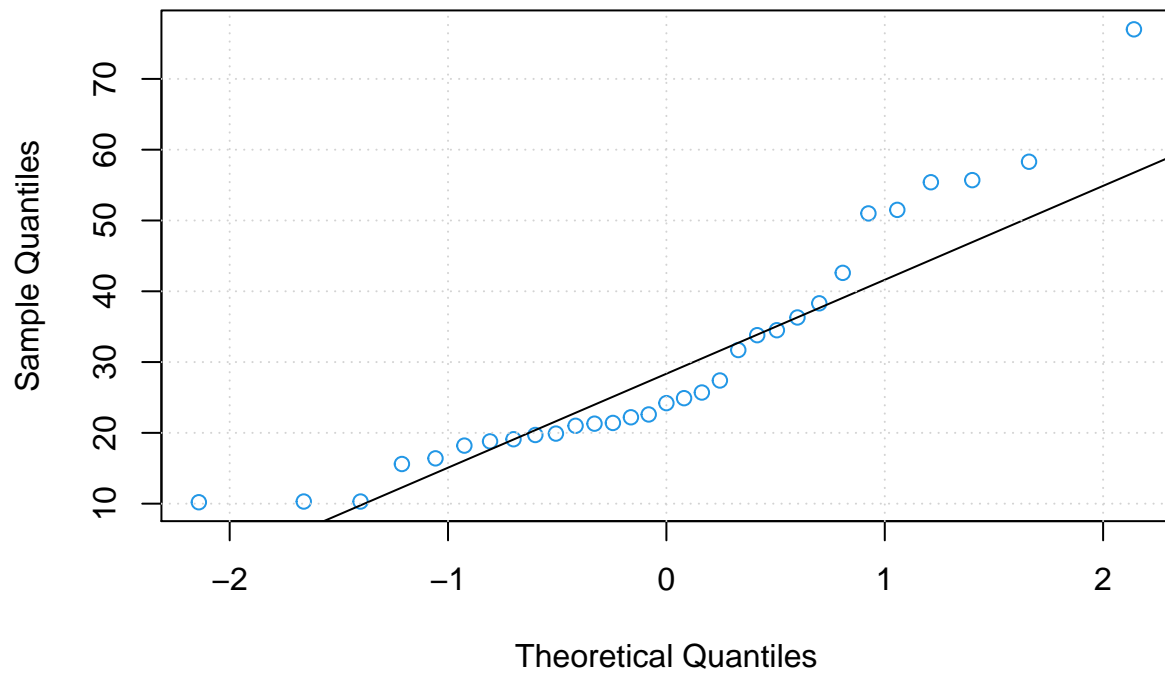
```
qqnorm(trees$Height,col=3)
qqline(trees$Height)
grid()
```


Normal Q-Q Plot



```
qqnorm(trees$Volume,col=4)
qqline(trees$Volume)
grid()
```

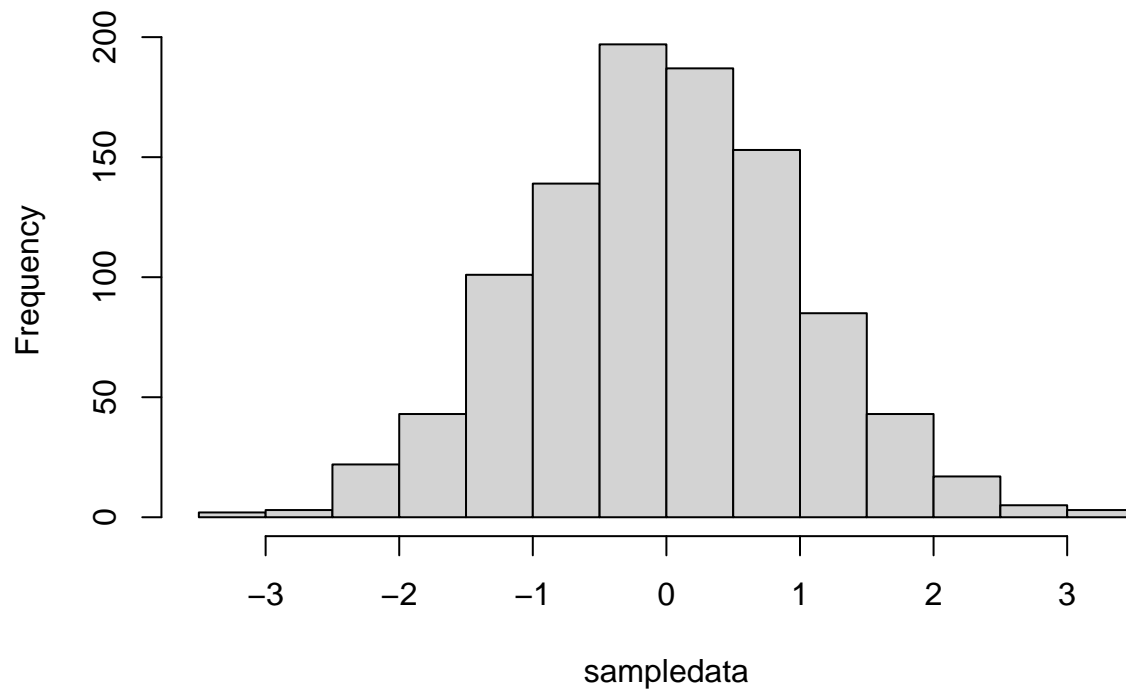
Normal Q-Q Plot



Question (2) Generate N number of random samples with different parameters and plot Q-Q plot with comparison to standard Normal Distribution.

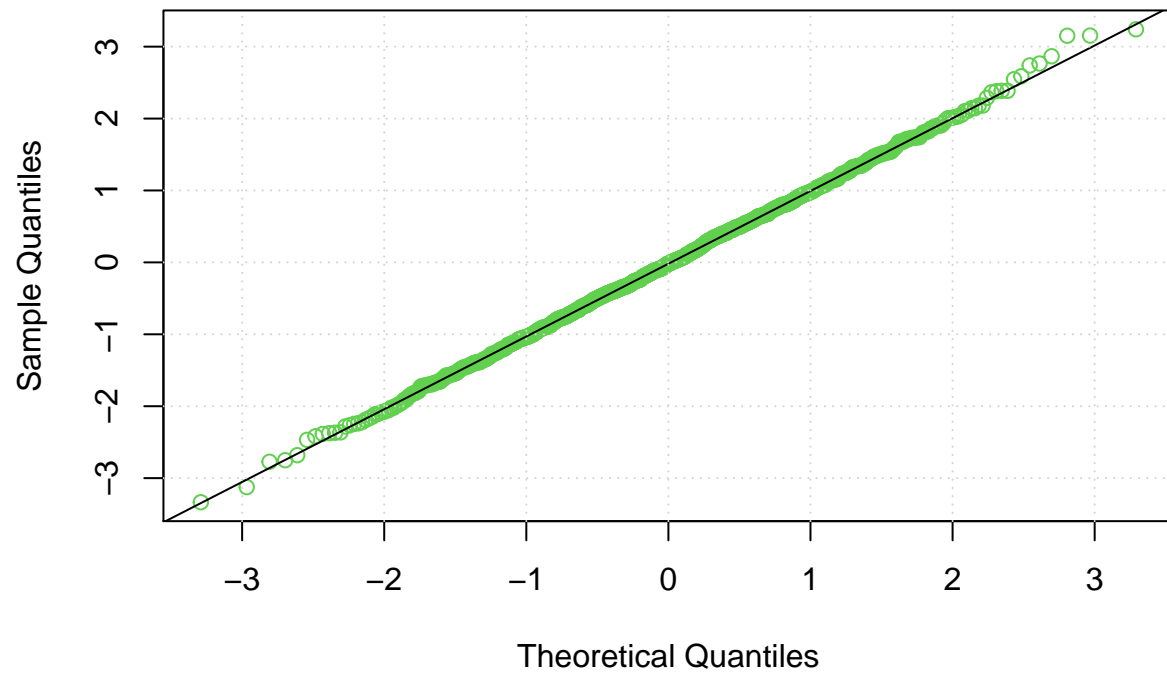
```
sampldata <-rnorm(1000,0,1)
sampldatan <-runif(1000,0,1)
hist(sampldata)
```

Histogram of sampledata



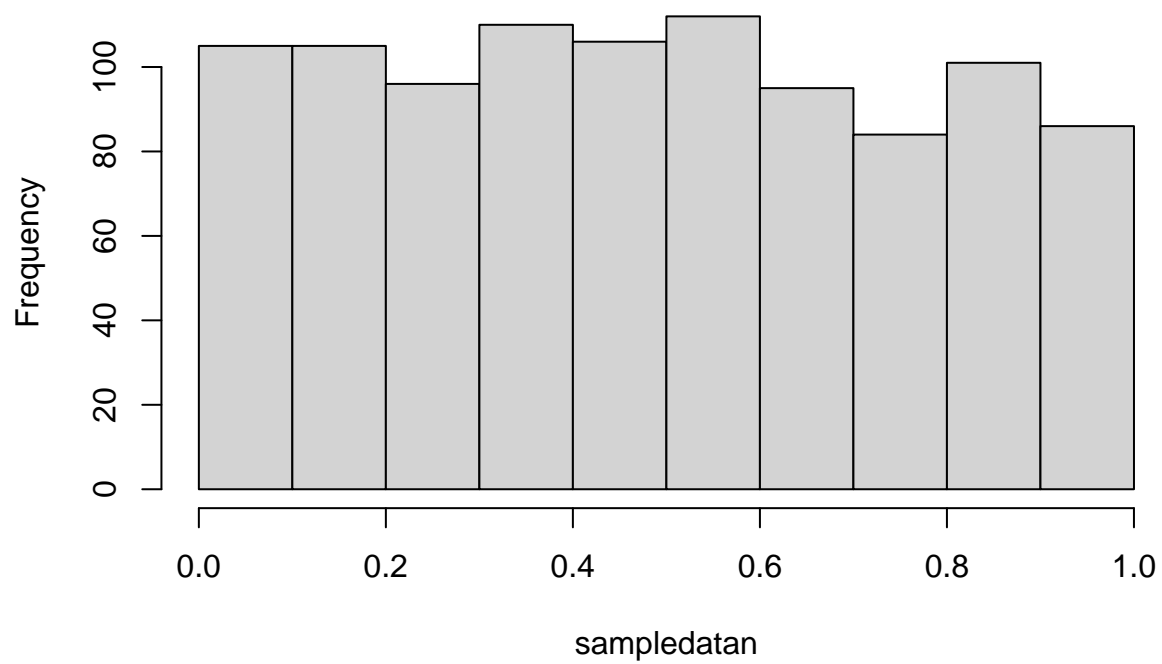
```
qqnorm(sampledata,col=3)
qqline(sampledata)
grid()
```

Normal Q-Q Plot



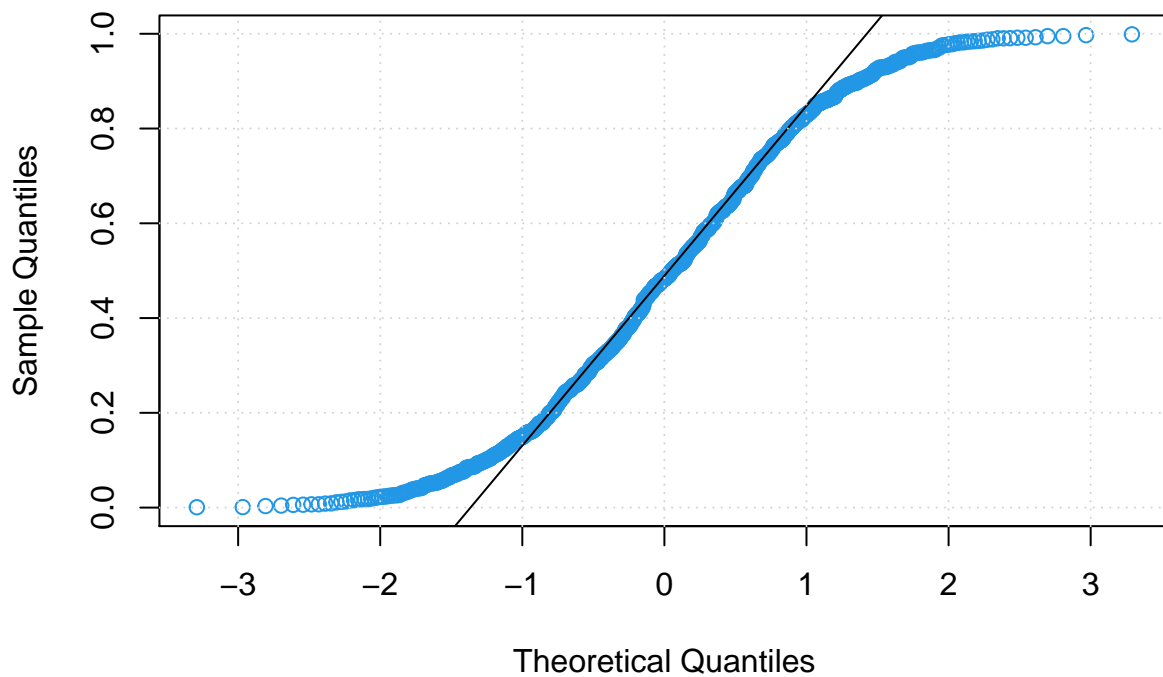
```
hist(sampledata)
```

Histogram of sampledatan



```
qqnorm(sampledatan,col=4)  
qqline(sampledatan)  
grid()
```

Normal Q-Q Plot



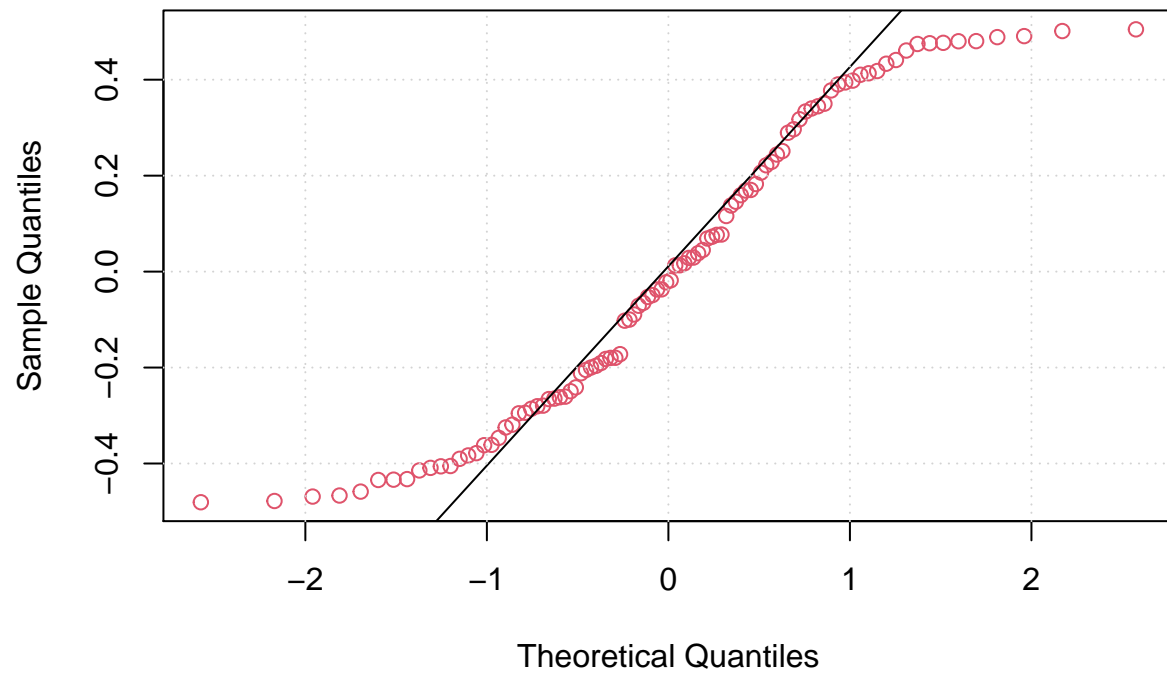
this is how we can identify the normal distribution using q-q plot.

```
library(tibble)
randomvariables<-matrix(data=rep(0,25000), nrow = 100, ncol = 5)
for (i in c(1:5)){
  if(i%%2==0){
    randomvariables[,i] <-rnorm(100, 0.5, 0.2)
  } else{
    randomvariables[,i]<-runif(100,0,1)
  }
  randomvariables[,i]<-randomvariables[,i]-mean(randomvariables[,i])
}

samplespacemixed<-data.frame(data=randomvariables)

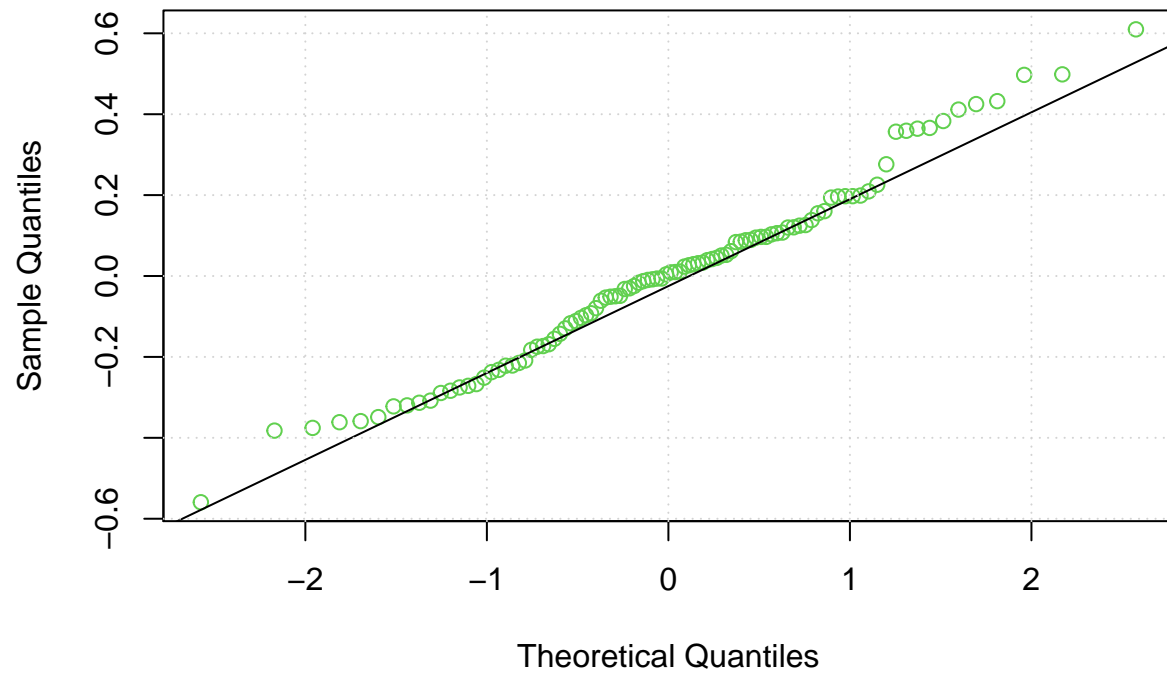
qqnorm(samplespacemixed$data.1,col =2)
qqline(samplespacemixed$data.1)
grid()
```

Normal Q-Q Plot



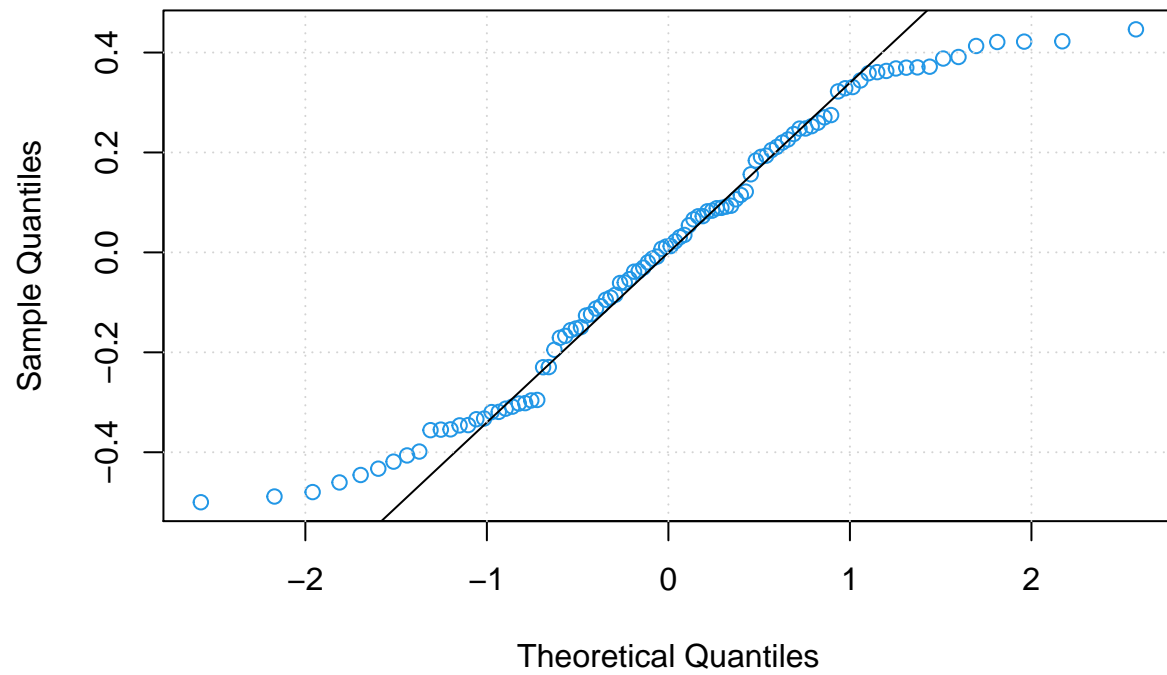
```
qqnorm(samplespacemixed$data.2,col =3)  
qqline(samplespacemixed$data.2)  
grid()
```

Normal Q-Q Plot



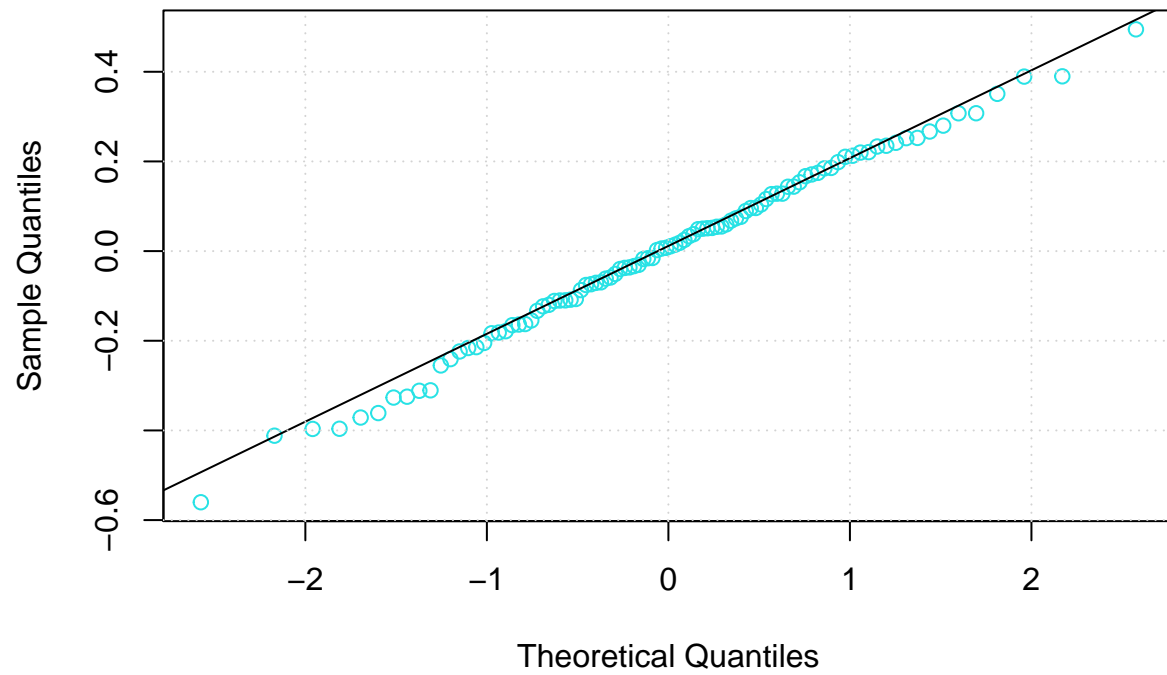
```
qqnorm(samplespacemixed$data.3,col =4)  
qqline(samplespacemixed$data.3)  
grid()
```


Normal Q-Q Plot

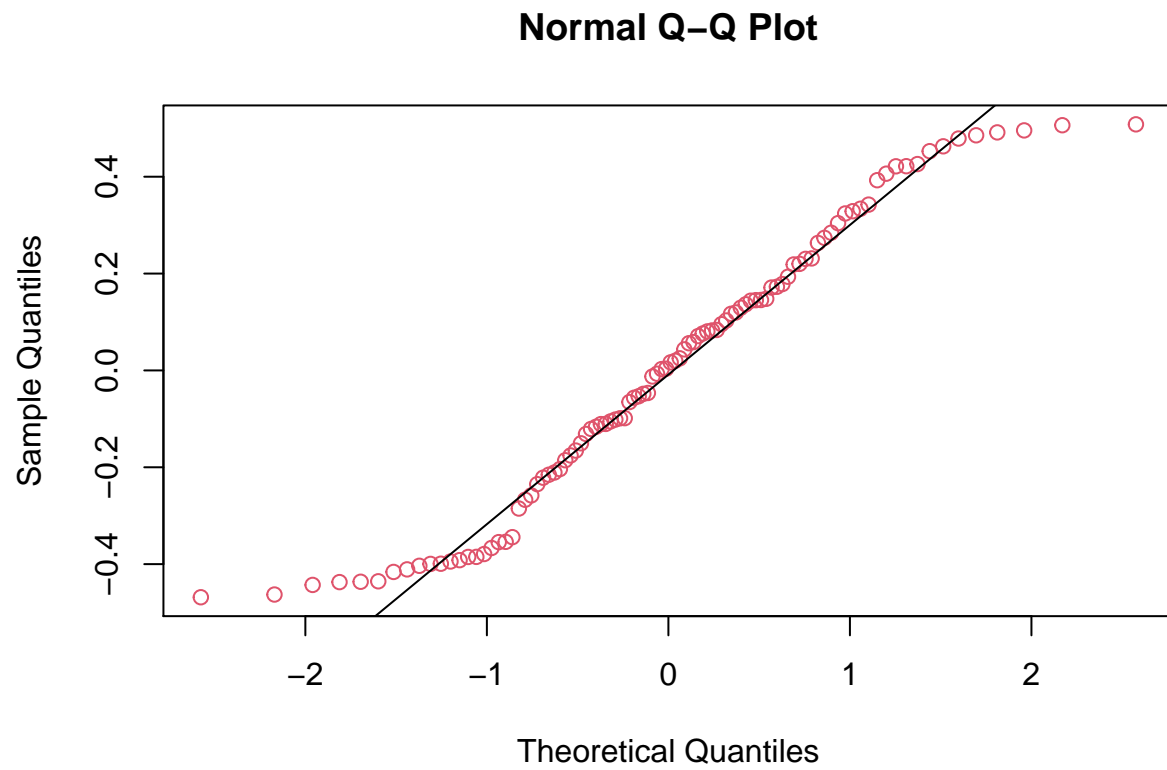


```
qqnorm(samplespacemixed$data.4,col =5)  
qqline(samplespacemixed$data.4)  
grid()
```

Normal Q-Q Plot



```
qqnorm(samplespacemixed$data.5,col =10)  
qqline(samplespacemixed$data.5)
```



this are the five different plots by which we can understand that is it follow the normal distribution or not. if it is no qqline then it is normal distribution otherwise not. we can say that 2 and 4th dataset slightly follows the normal distribution and others are not following.