677 final project

Boyu Chen

2022-05-12

4.25

For this question, I reference from stackoverflow:

https://stackoverflow.com/questions/24211595/order-statistics-in-r?msclkid=fd6683dac56711ecbfcea9bd8a172395

```
# Get pdf function
f<-function(x, mu=0, sigma=1) dunif(x, mu,sigma)
# Get cdf function
F<-function(x, mu=0, sigma=1) punif(x,mu,sigma, lower.tail=FALSE)
#Find the distribution of the order statistics
integrand <- function(x,r,n){
  x*(1-F(x))^(r-1)*F(x)^(n-r)*f(x)
}
#Get expectation function
E <- function(r,n) {
(1/beta(r,n-r+1)) * integrate(integrand,-Inf,Inf, r, n)$value
}
# Get the approximate function
medianprrox<-function(k,n){
  m<-(k-1/3)/(n+1/3)
  return(m)
}
```

So we can get

```
E(2.5,5)
```

```
## [1] 0.4166667
```

```
medianprrox(2.5,5)
```

```
## [1] 0.40625
```

```
E(5,10)
```

```
## [1] 0.4545455
```

```
medianprrox(5,10)
```

```
## [1] 0.4516129
```

By compare these two result, we can see they are similar.

4.27

Import data

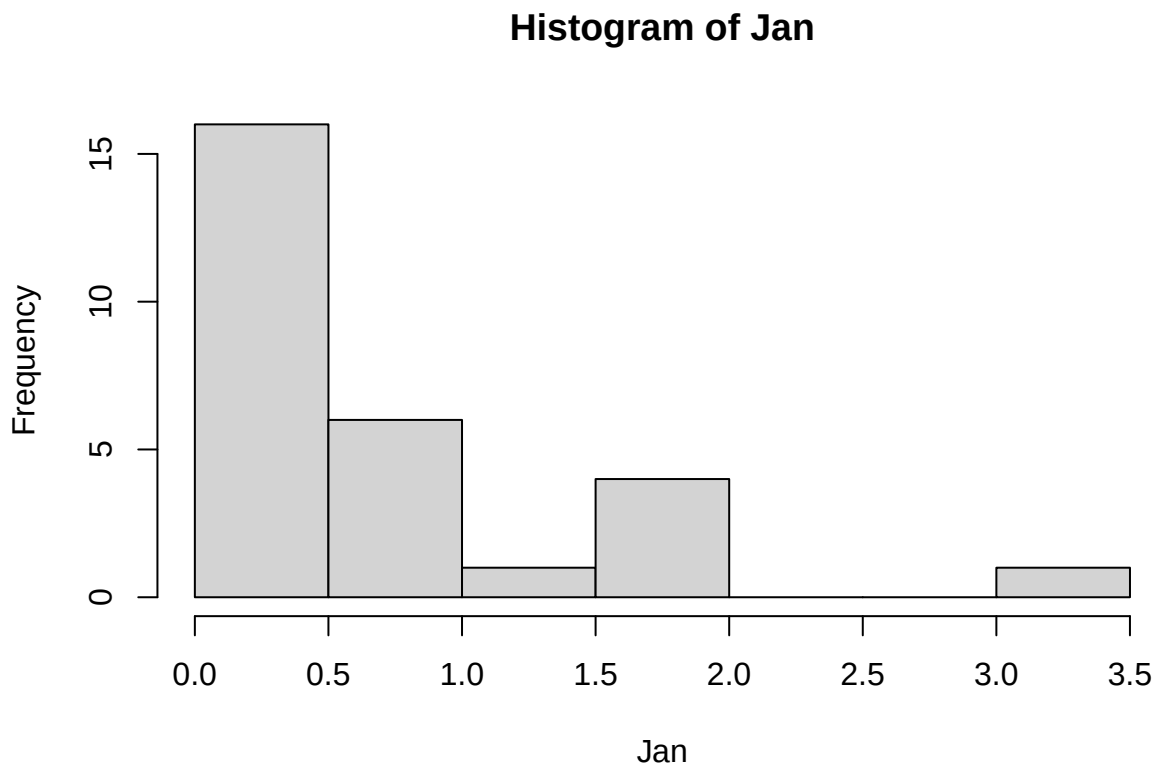
```
Jan<-c(0.15,0.25,0.10,0.20,1.85,1.97,0.80,0.20,0.10,0.50,0.82,0.40,1.80,0.20,1.12,      1.83,0.45,3.17
Jul<-c(0.30,0.22,0.10,0.12,0.20,0.10,0.10,0.10,0.10,0.10,0.10,0.17,0.20,2.80,0.85,     0.10,0.10,1.23,
0.1,0.2,0.1)
```

part a

```
summary(Jan)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1875  0.4250  0.7196  0.9000  3.1700
```

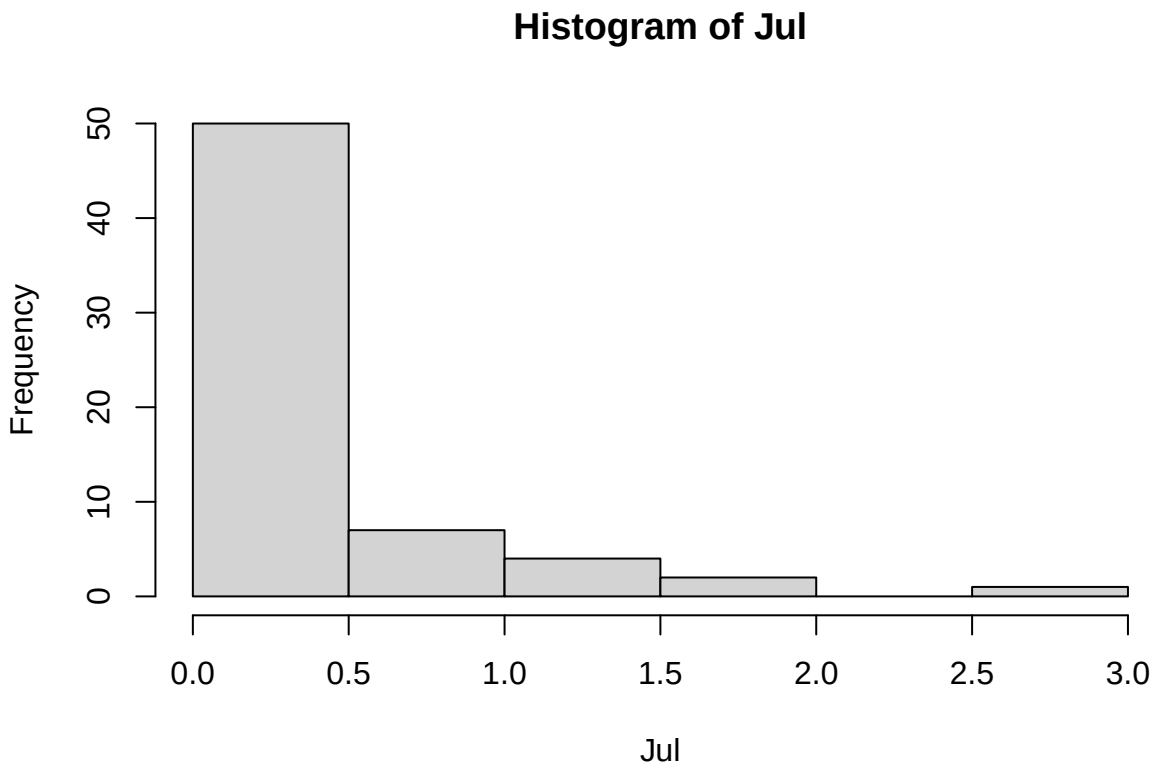
```
hist(Jan)
```



```
summary(Jul)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.1000  0.2000  0.3931  0.4275  2.8000
```

```
hist(Jul)
```

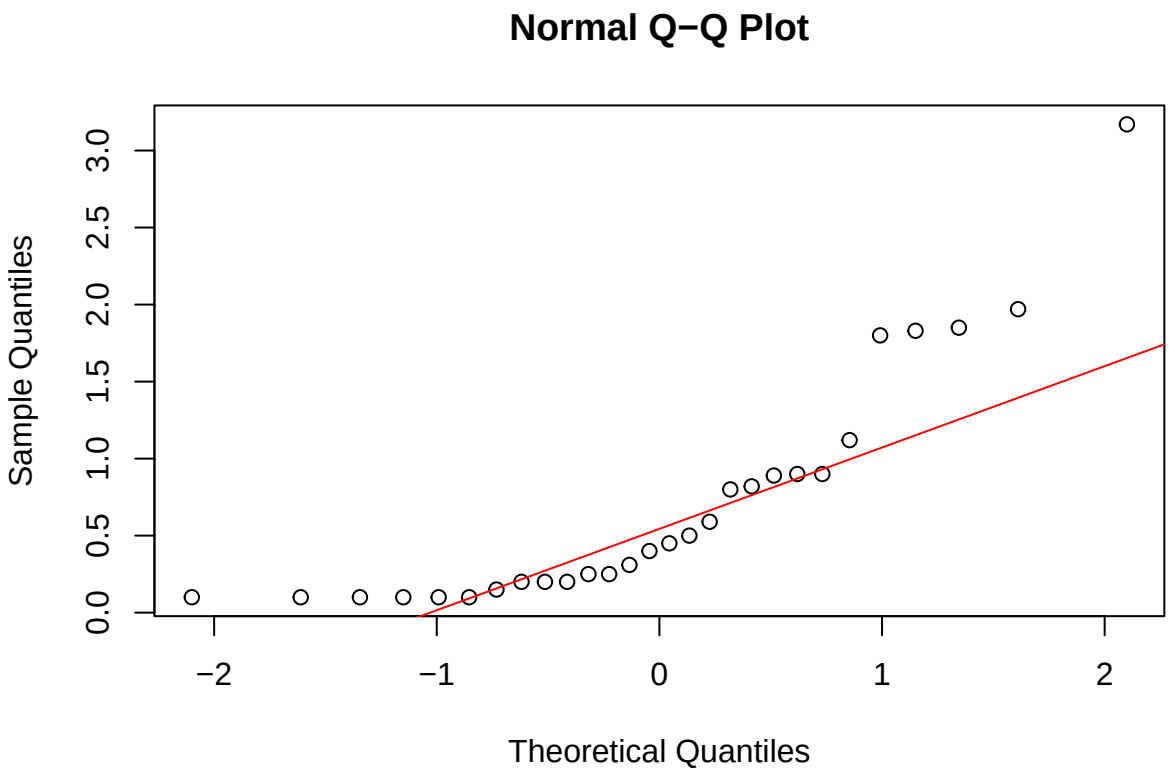


By comparing these two months, we can find that the value of Min.,1st Qu., Median,Mean 3rd Qu.,Max. for Jul is smaller than Jan's.

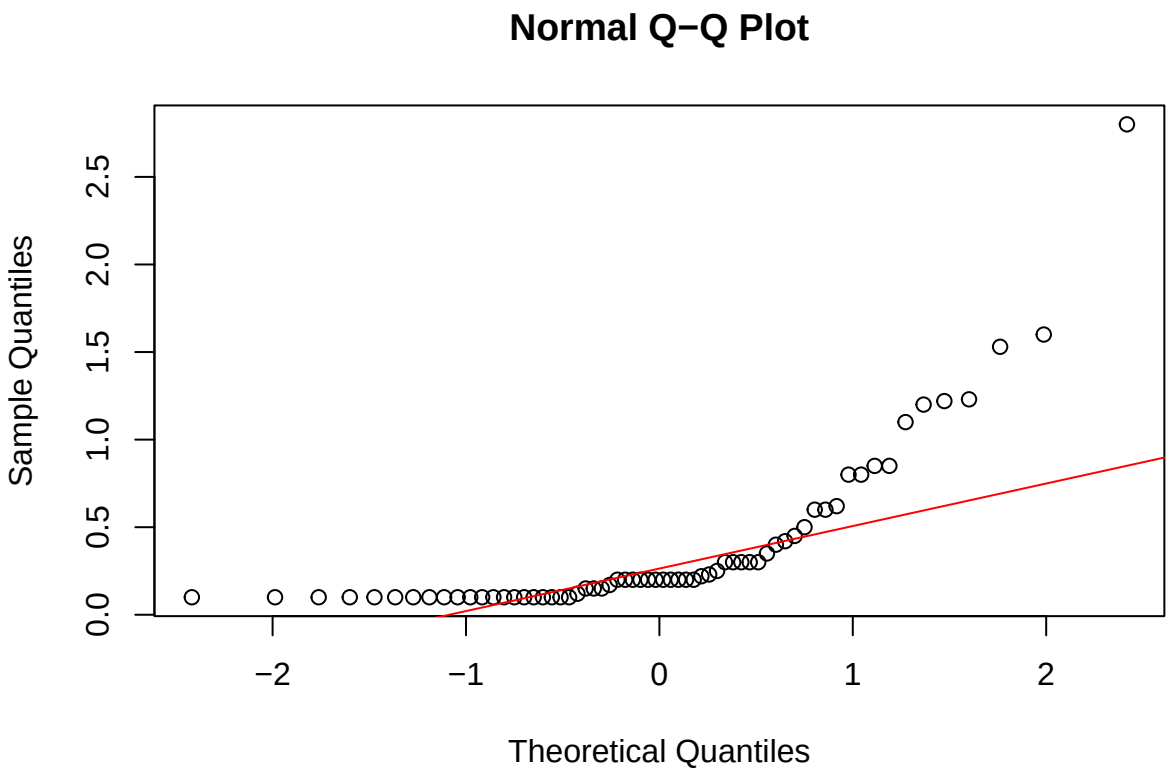
part b

Create qqplot

```
qqnorm(Jan, pch = 1)
qqline(Jan, col = "red", lwd = 1)
```

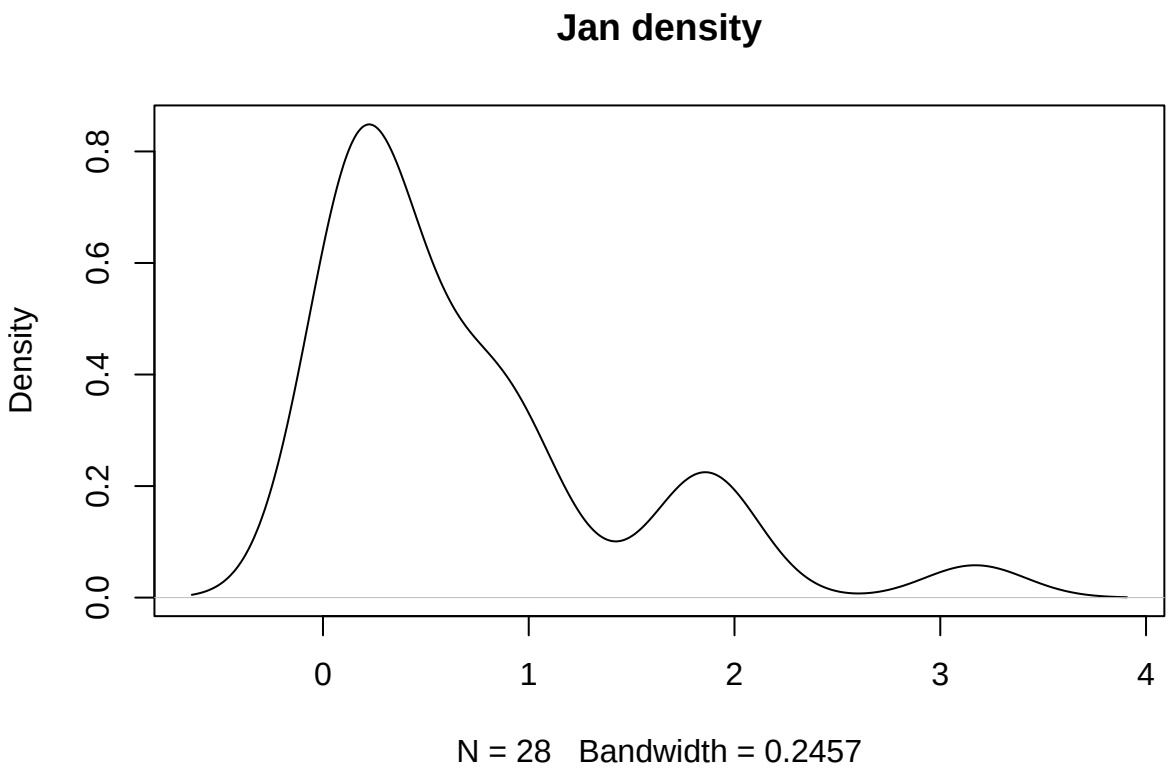



```
qqnorm(Jul, pch = 1)
qqline(Jul, col = "red", lwd = 1)
```

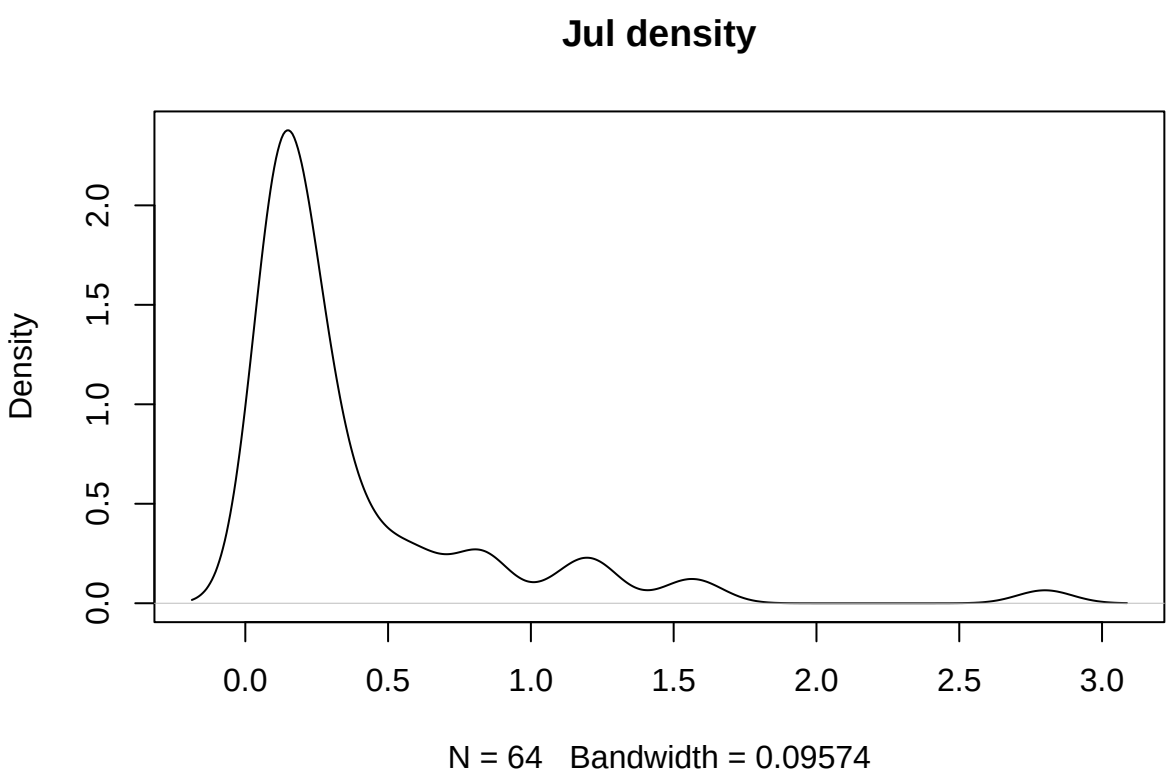


To decide which kind of model is the most appropriate, I decide to use density plot.

```
plot(density(Jan),main='Jan density')
```




```
plot(density(Jul),main='Jul density')
```



Based on the qqplots, we can find the sample data doesn't look like the normal distribution. So according to the density plot, the data looks like following the gamma distribution. So I think gamma distribution is reasonable.

part c

To fit a gamma model, I reference from this source:

https://www.statology.org/fit-gamma-distribution-to-dataset-in-r/

```
fit.Jan <- fitdist(Jan, distr = "gamma", method = "mle")
summary(fit.Jan)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.056222  0.2497495
## rate  1.467650  0.4396202
## Loglikelihood:  -18.7616   AIC:  41.5232   BIC:  44.18761
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.7893943
## rate  0.7893943 1.0000000
```

Standard error is 0.2497495 0.4396202

```
fit.Jul <- fitdist(Jul, distr = "gamma", method = "mle")
summary(fit.Jul)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape 1.196419  0.1891196
## rate  3.043403  0.5936302
## Loglikelihood:  -3.634886   AIC:  11.26977   BIC:  15.58754
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.8103948
## rate  0.8103948 1.0000000
```

Standard error is 0.1891196 0.5936302

Find MLE

```
# MLE for Jan
exp(fit.Jan$loglik)
```

```
## [1] 7.11117e-09
```

```
# MLE for Jul
exp(fit.Jul$loglik)
```

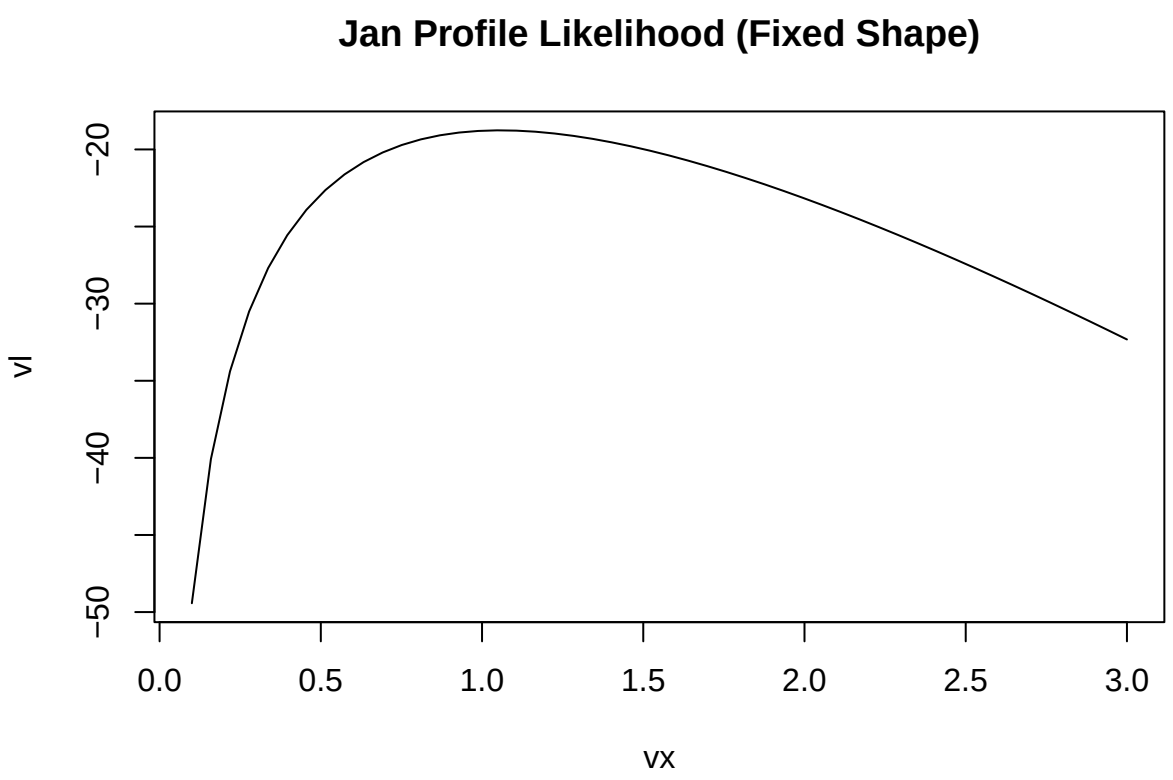
```
## [1] 0.02638693
```

By comparing MLE, we can find Jan's MLE is smaller than Jul' MlE. So we can say the model for Jul is better than Jan's. Parameter comparison: Jul's alpha and beta is larger than Jan's alpha and beta.

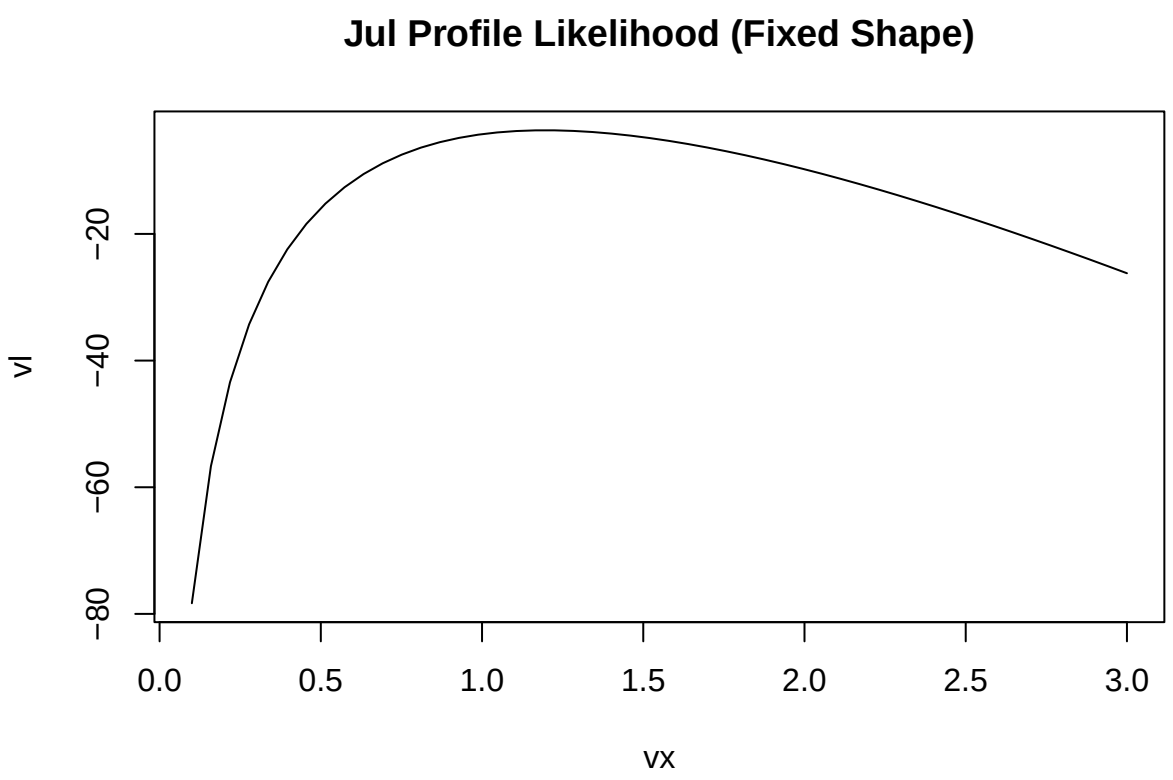
Creating Likelihood profile I reference this resource:

https://www.r-bloggers.com/2015/11/profile-likelihood/

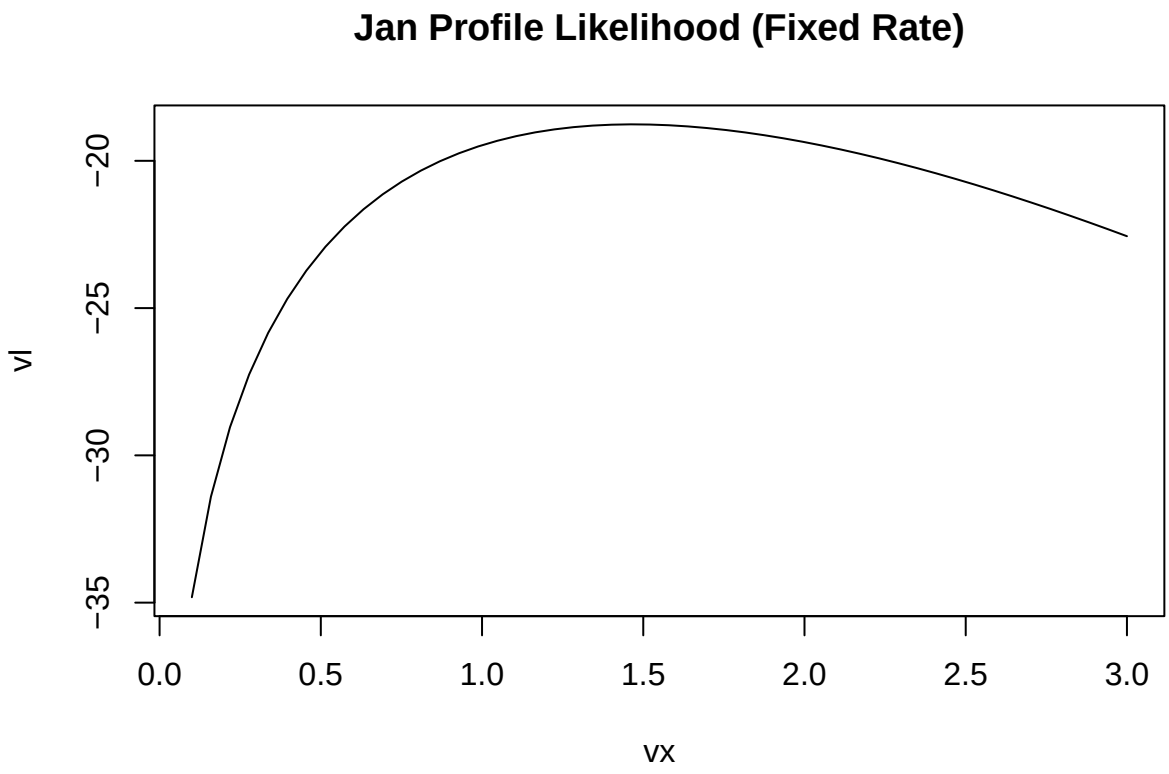
```
prof_log_lik=function(a){
b=(optim(1,function(z) -sum(log(dgamma(Jan,a,z)))))$par
return(-sum(log(dgamma(Jan,a,b))))
}
vx=seq(.1,3,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main='Jan Profile Likelihood (Fixed Shape)')
```



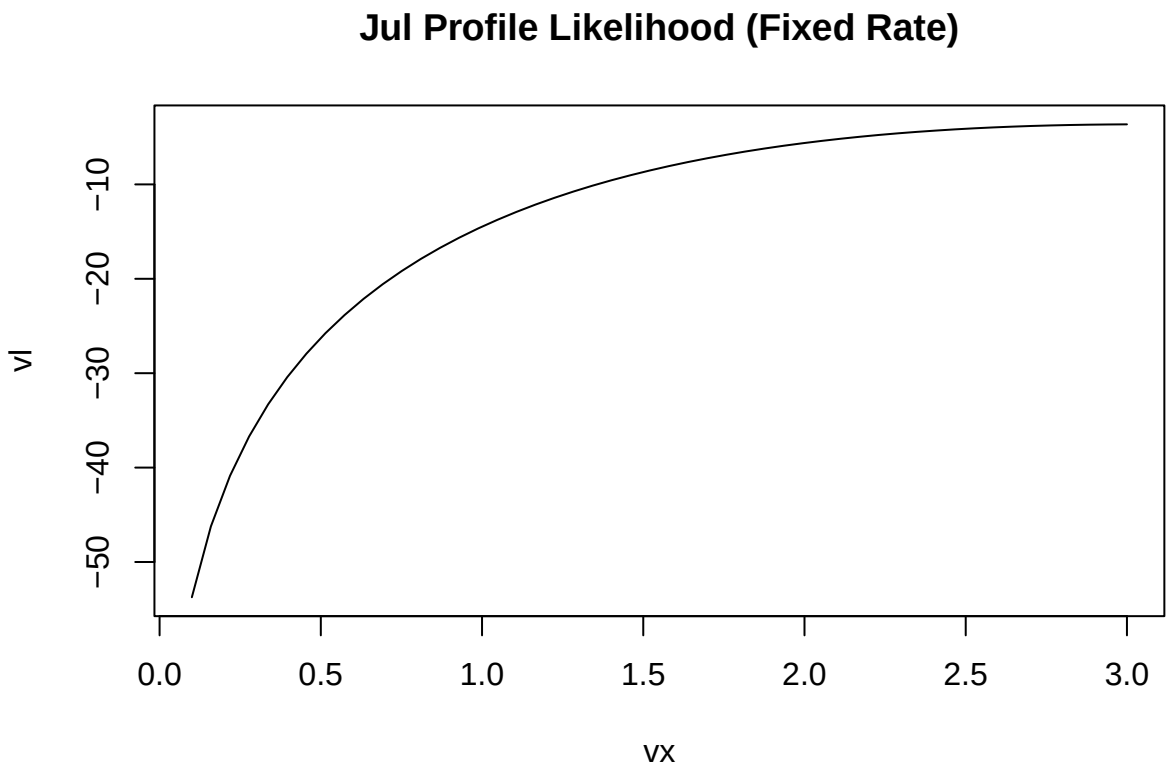

```
prof_log_lik=function(a){
b=(optim(1,function(z) -sum(log(dgamma(Jul,a,z)))))$par
return(-sum(log(dgamma(Jul,a,b))))
}
vx=seq(.1,3,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main='Jul Profile Likelihood (Fixed Shape)')
```




```
prof_log_lik=function(z){
a=(optim(1,function(a) -sum(log(dgamma(Jan,a,z)))))$par
return(-sum(log(dgamma(Jan,a,z))))
}
vx=seq(.1,3,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main='Jan Profile Likelihood (Fixed Rate)')
```




```
prof_log_lik=function(z){
a=(optim(1,function(a) -sum(log(dgamma(Jul,a,z)))))$par
return(-sum(log(dgamma(Jul,a,z))))
}
vx=seq(.1,3,length=50)
vl=-Vectorize(prof_log_lik)(vx)
plot(vx,vl,type="l",main='Jul Profile Likelihood (Fixed Rate)')
```



part d

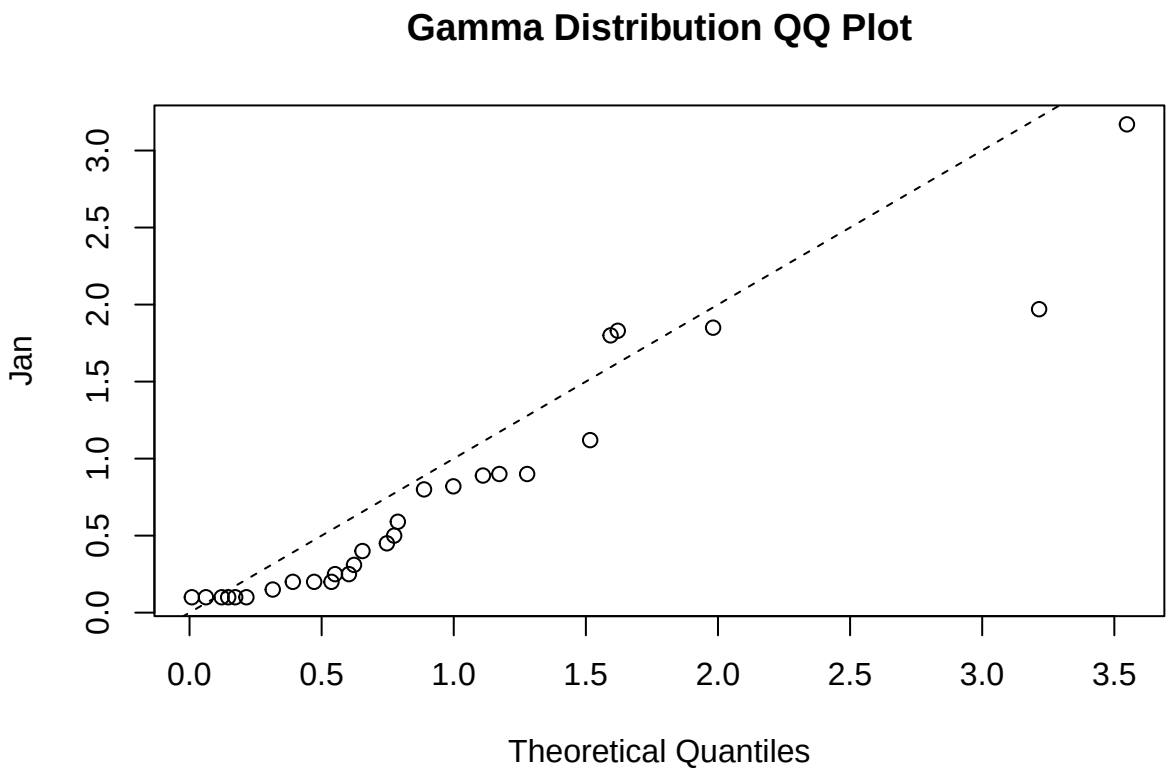
For this part, I referenced from this source:

https://github.com/qPharmetra/qpToolkit/blob/master/R/qqGamma.r

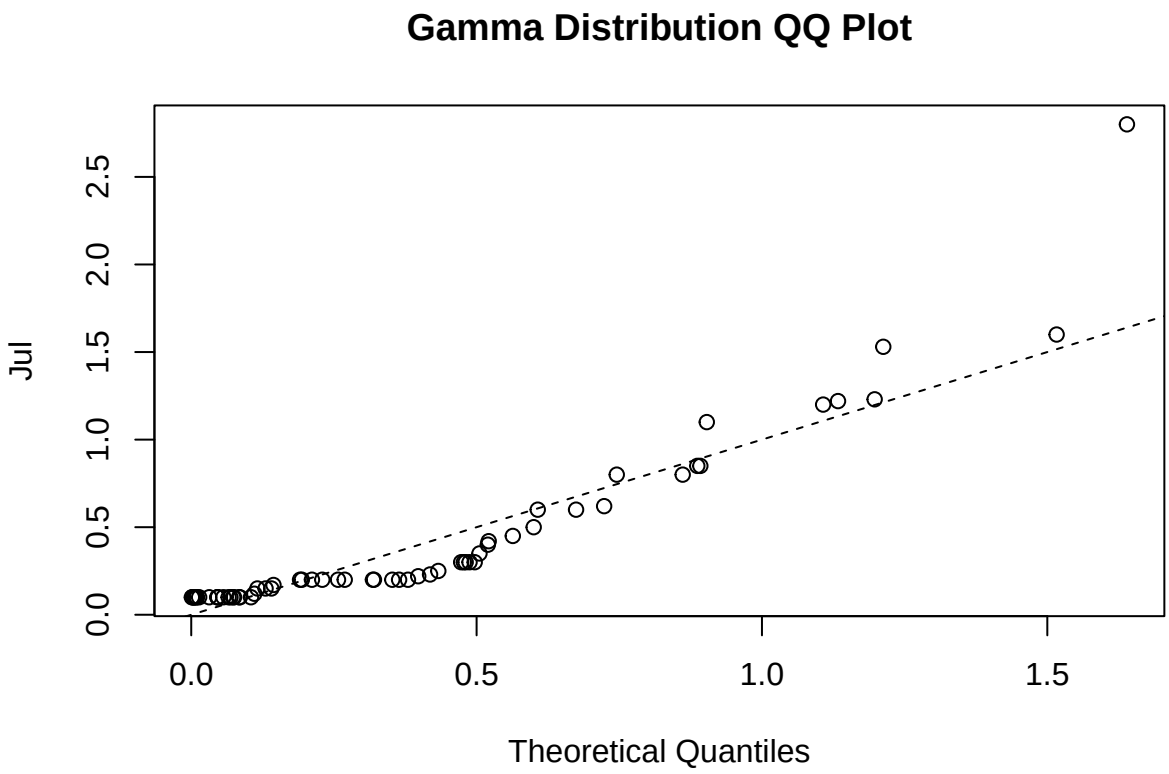
```
qqGamma <- function(x
                    , ylab = deparse(substitute(x))
                    , xlab = "Theoretical Quantiles"
                    , main = "Gamma Distribution QQ Plot",...)
{
    # Plot qq-plot for gamma distributed variable

    xx = x[!is.na(x)]
    aa = (mean(xx))^2 / var(xx)
    ss = var(xx) / mean(xx)
    test = rgamma(length(xx), shape = aa, scale = ss)

    qqplot(test, xx, xlab = xlab, ylab = ylab, main = main,...)
    abline(0,1, lty = 2)
}
qqGamma(Jan)
```


Gamma Distribution QQ Plot
Jan
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Theoretical Quantiles

```
qqGamma(Jul)
```

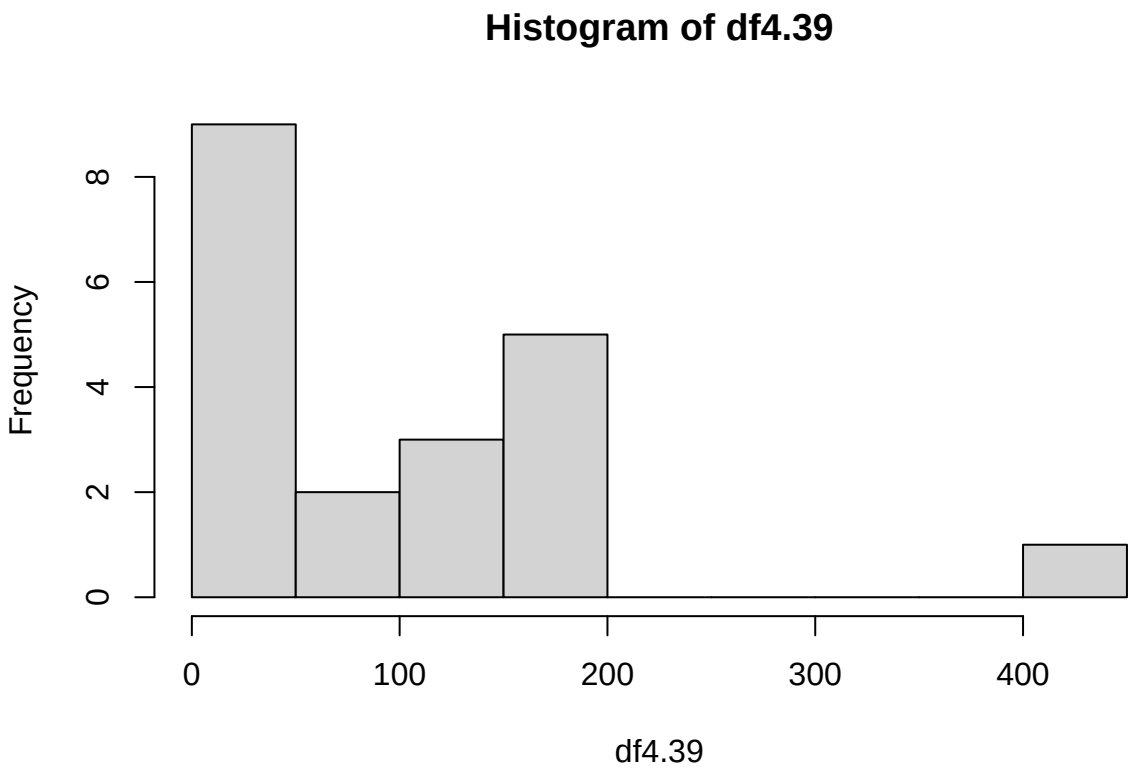


From the plot, it is obvious that Jul model fit better.

4.39

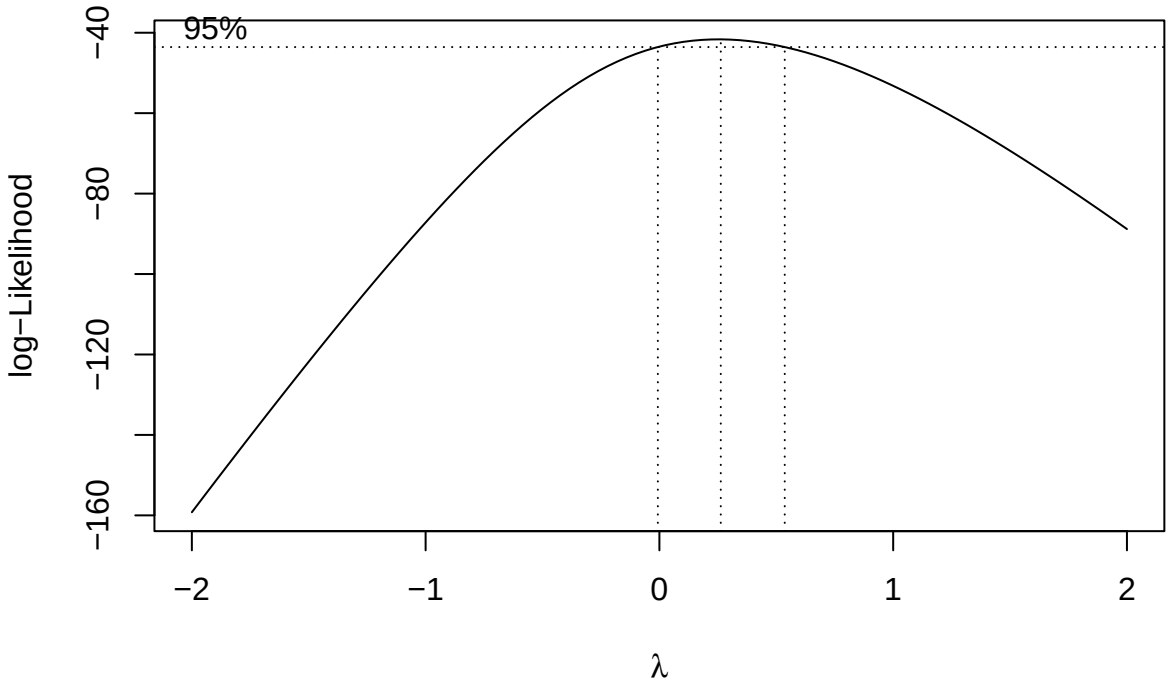
First import data

```
df4.39<-c(0.4,1.0,1.9,3.0,5.5,8.1,12.1,25.6,50.0,56.0,70.0,115.0,115.0,119.5,154.5,157.0,175.0,179.0,18
hist(df4.39)
```



Conduct Box-Cox transformation

```
b<-boxcox(lm(df4.39~1))
```

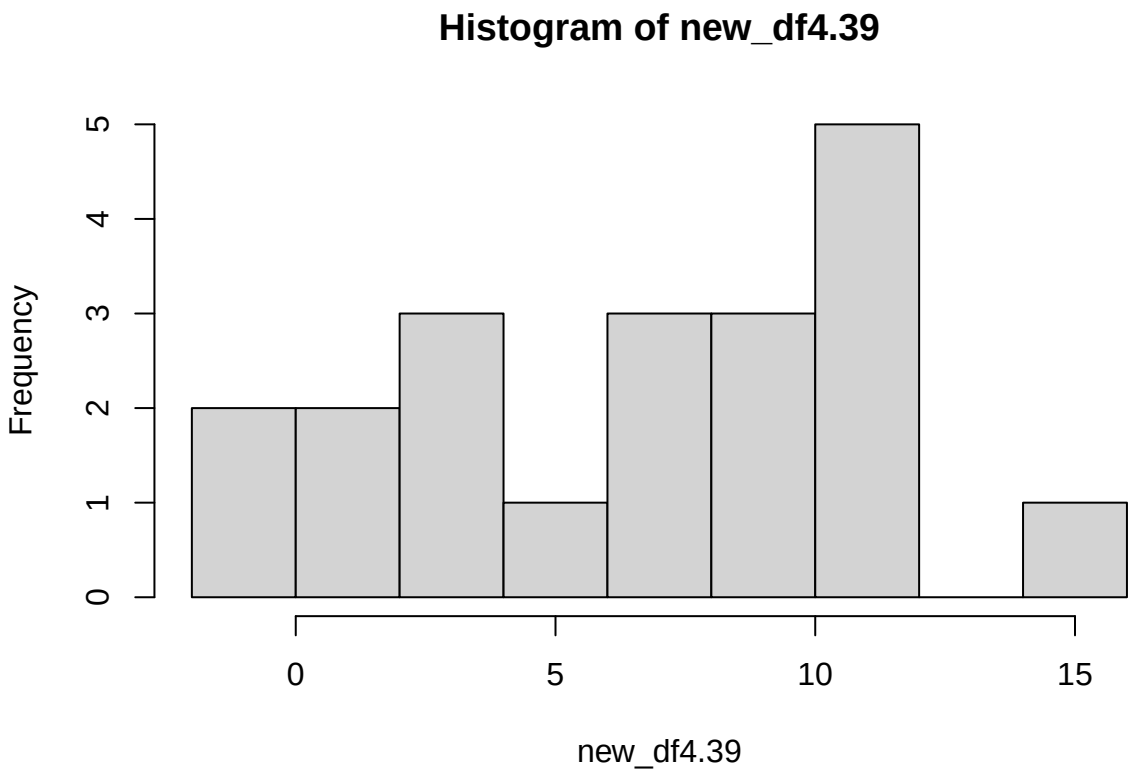


Find exact lambda value

```
lambda <- b$x[which.max(b$y)]
lambda
```

```
## [1] 0.2626263
```

```
# We can find the exact lambda value is 0.2626263
new_df4.39 <- (df4.39 ^ lambda - 1) / lambda
hist(new_df4.39)
```



What I learned

Through this project, I learned how to create gamma distribution model in R and construct likelihood profile.