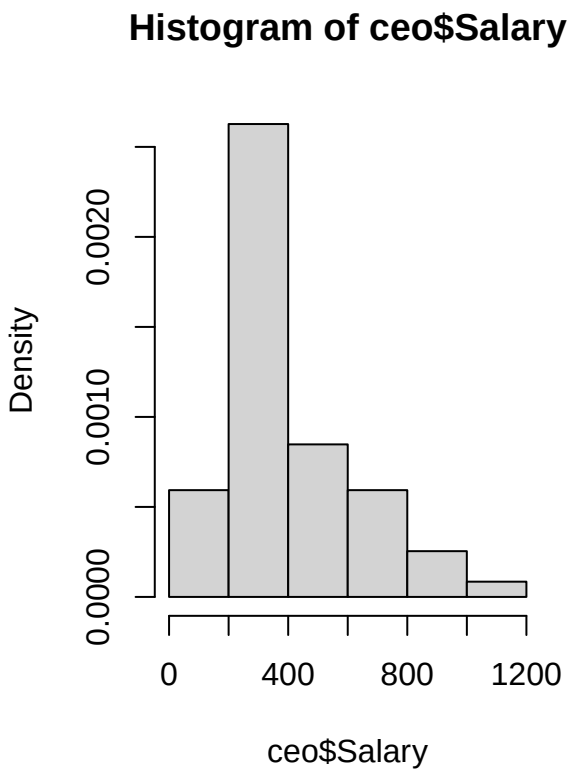
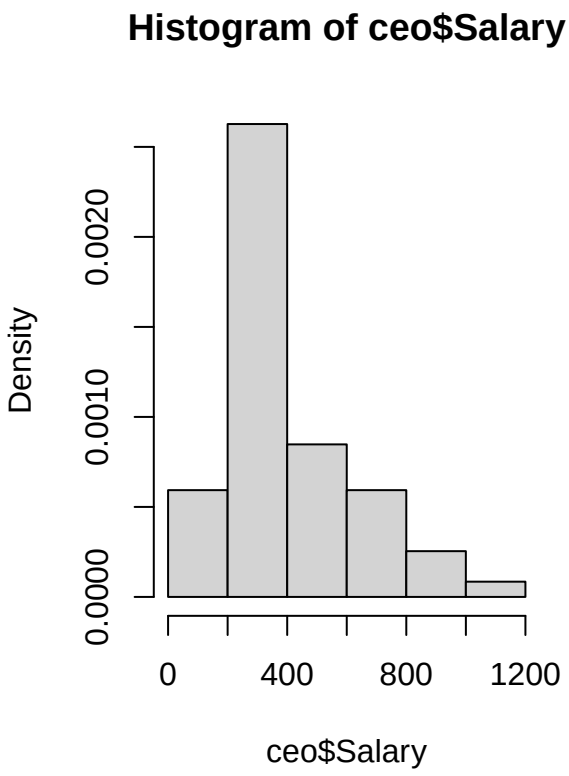
A4Q4

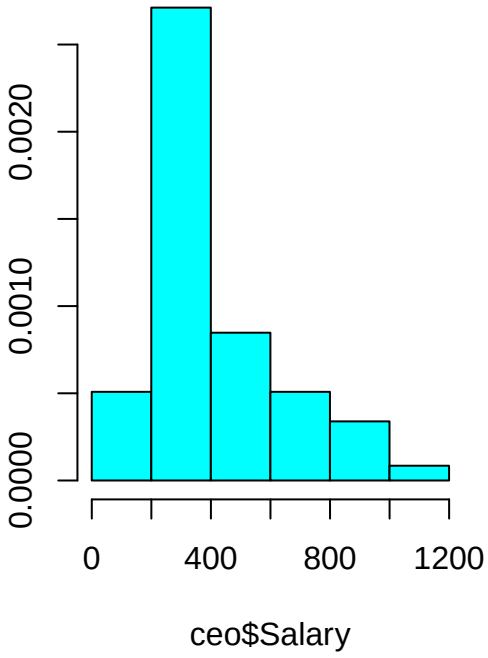
Undergraduate Student

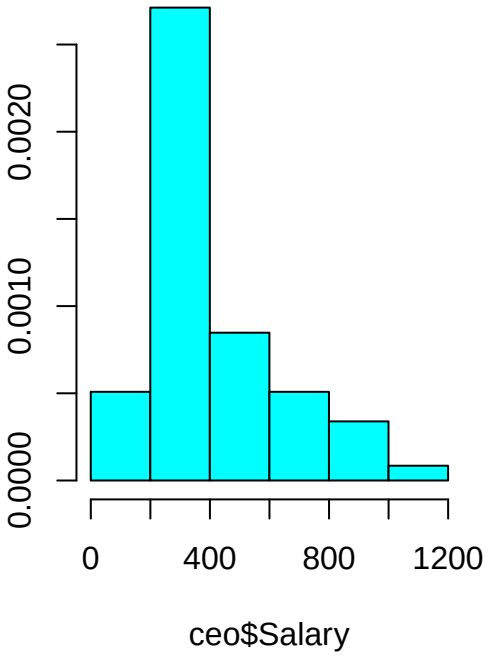
```
ceo <- read.csv("ceo.csv", header=T)[,-c(1)]
ceo <- na.omit(ceo)

cv.hist.fun <- function(x){
     ### histogram cross validation function
     n <- length(x)
     a <- min(x)
     b <- max(x)
     k <- 100
     nbins <- seq(1,n,length=k)  ###number of bins
     nbins <- round(nbins)
     h <- (b-a)/nbins            ###width of bins
     risk <- rep(0,k)
     for(i in 1:k){
          ###get counts N_j
          br <- seq(a,b,length=nbins[i]+1)
          N <- hist(x,breaks=br,plot=F)$counts
          risk[i] <- sum(N^2)/(n^2*h[i])  - (2/(h[i]*n*(n-1)))*sum(N*(N-1))
          }
     hbest <- h[risk==min(risk)]
     hbest <- hbest[1]  ###in case of tie take first (smallest) one
     mbest <- (b-a)/hbest   ###optimal number of bins
     list(risk=risk,nbins=nbins,h=h,mbest=mbest)
}

n<-length(ceo$Salary)
opt.h.hist<-cv.hist.fun(ceo$Salary)$mbest

## histograms
par(mfrow=c(1,2))
hist(ceo$Salary,probability=TRUE)
hist(ceo$Salary,probability=TRUE,breaks=opt.h.hist)
```





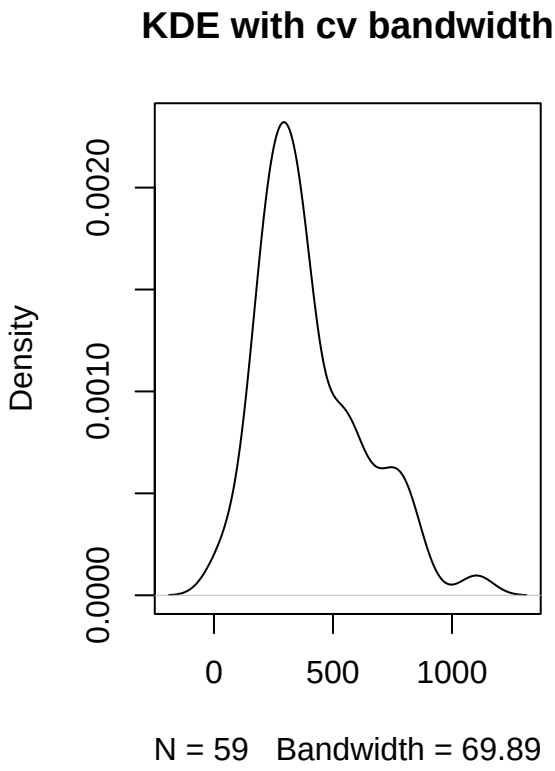

```
library(MASS)
truehist(ceo$Salary)
truehist(ceo$Salary,nbins=opt.h.hist)
```

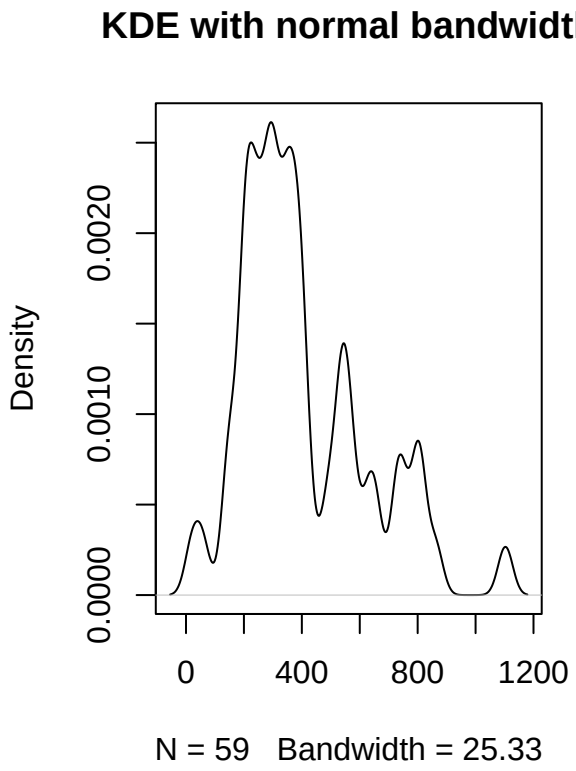




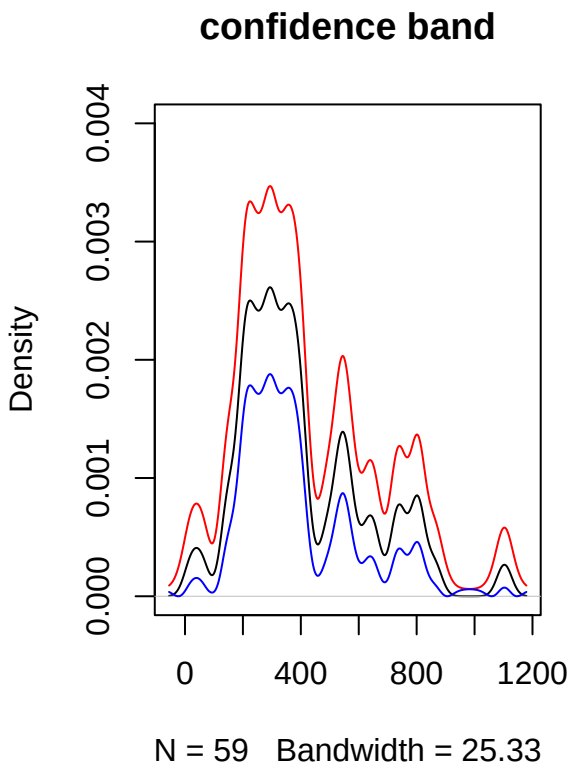

```
par(mfrow=c(1,2))

### bandwidth selection by cv
```

```r
h.cv<-ucv(ceo$Salary)
f.cv<-density(ceo$Salary,width=h.cv)
plot(f.cv ,main="KDE with cv bandwidth")

# normal reference
sigma.hat<-min(sd(ceo$Salary),IQR(ceo$Salary)/1.34)
h.normal<-1.06*sigma.hat/n^(0.2)
f.normal<-density(ceo$Salary,width=h.normal)
plot(f.normal,main="KDE with normal bandwidth")
```






```r
#95% confidence band
c <- (qnorm(0.05)/(2*1200/25.33))/2*sqrt(1200/25.33/n)
ln <- ifelse(sqrt(f.normal$y)-c>0,(sqrt(f.normal$y)-c)^2,0)
un <- (sqrt(f.normal$y)+c)^2
plot(f.normal,main="confidence band", ylim=c(0,0.004))
lines(f.normal$x, ln, col="red")
lines(f.normal$x, un, col="blue")
```

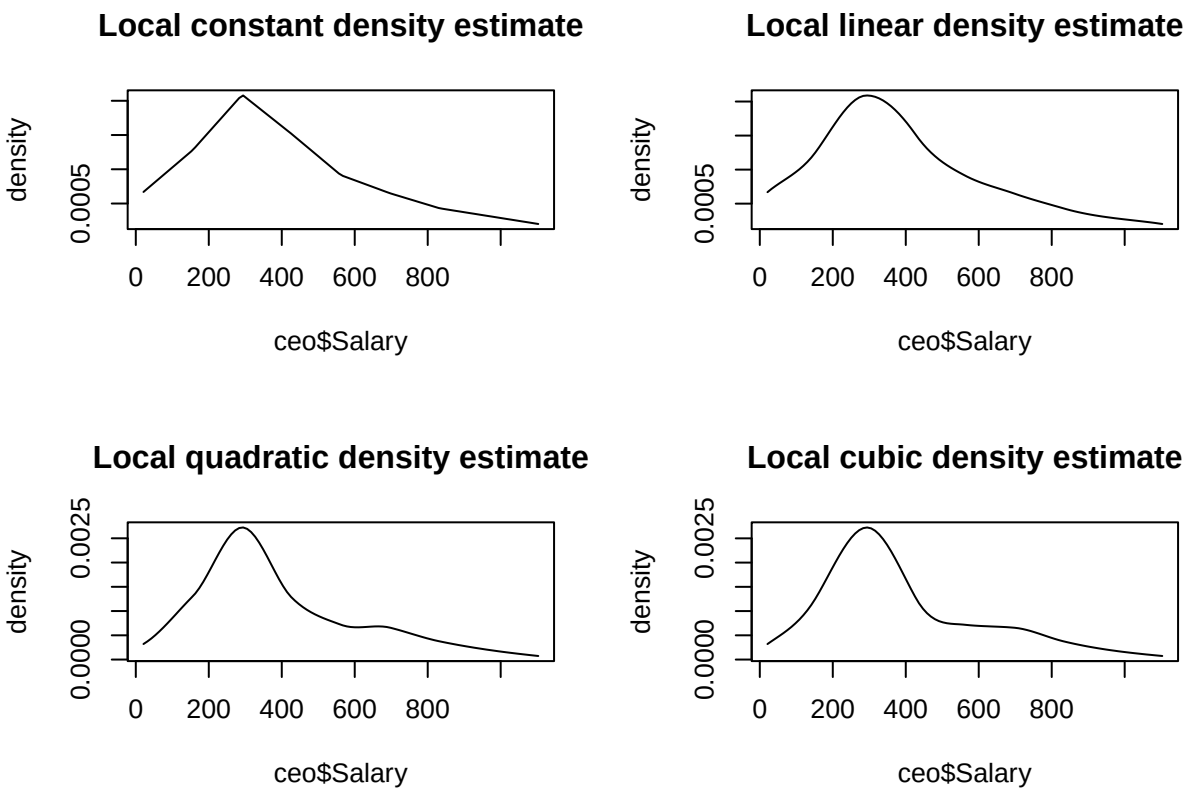


Comments: From the 95% confidence band, we see that all bumps fall in the band. This means that with 25.33 bandwidth, the bumps exist.

```
# kernel shapes
library(locfit)
```

```
## locfit 1.5-9.5    2022-03-01
```

```
par(mfrow=c(2,2))
f.loc0<-locfit(~ceo$Salary,deg=0,link="ident")
plot(f.loc0,main="Local constant density estimate")
f.loc1<-locfit(~ceo$Salary,deg=1,link="ident")
plot(f.loc1,main="Local linear density estimate")
f.loc2<-locfit(~ceo$Salary,deg=2,link="ident")
plot(f.loc2,main="Local quadratic density estimate")
f.loc3<-locfit(~ceo$Salary,deg=3,link="ident")
plot(f.loc3,main="Local cubic density estimate")
```



Comment: With constant density estimate, the curve is not smooth. When we use linear and above, the curve becomes smooth. However, when we go up to quadratic density estimate, we can see a clear small bump at right tail.