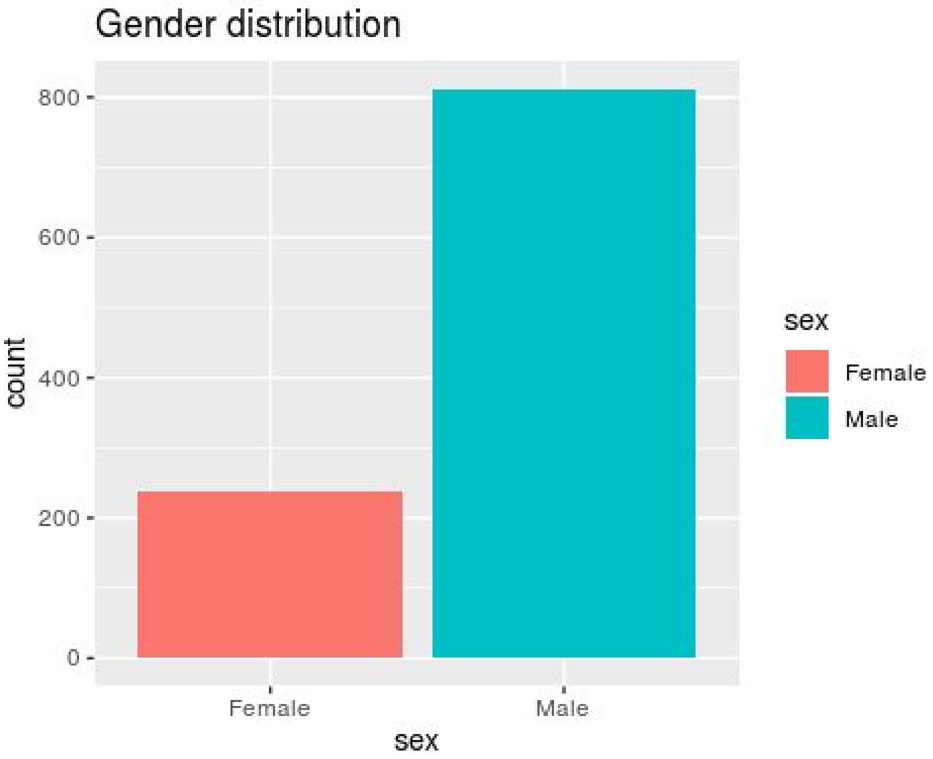
Data Visualisation

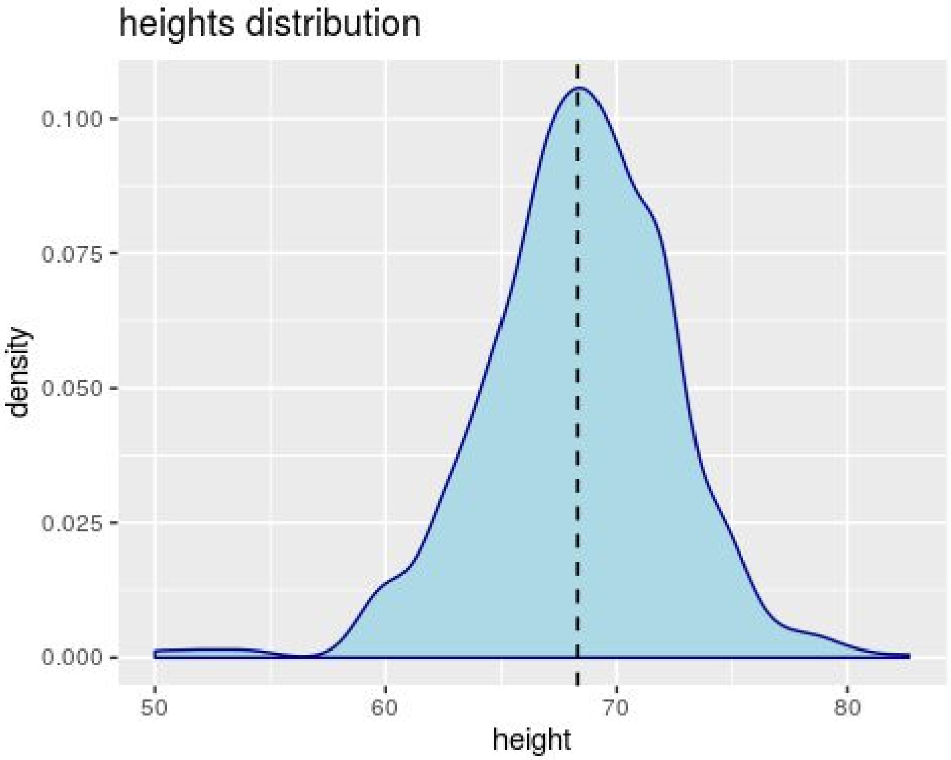
Date: 24/01/20 Lab-:1

I. Which graph you will use to plot data for gender distribution and height distribution? Plot and Justify. Do we need any plot to understand gender distribution?

```r
library(dslabs)
library(ggplot2)
heightsData <- data.frame(data=heights,stringsAsFactors=FALSE)
colnames(heightsData) <- c("sex", "height")
ggplot(data=heightsData)+geom_bar(mapping=aes(x=sex,fill=sex))+ggtitle("Gender distribution")
```



```
ggplot(data=heightsData,mapping=aes(x=height))+ggtitle("heights
distribution")+
        geom_density(color="darkblue", fill="lightblue") +
        geom_vline(aes(xintercept = mean(height)),
             linetype = "dashed", size = 0.6)
```



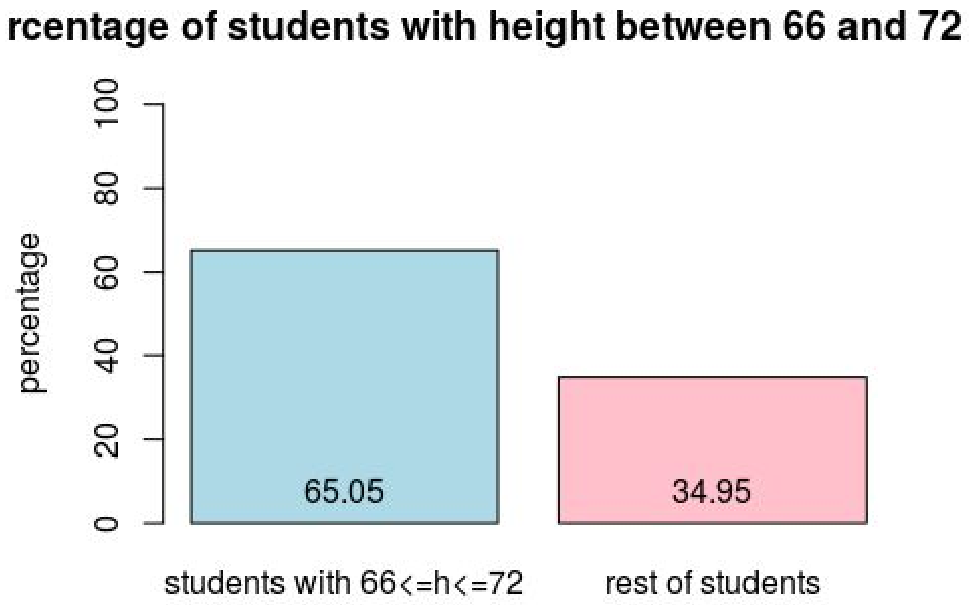
1)As gender variable is a discrete variable barplot will be a best way to show the distribution.

2)As height is a continuous variable density curve will show us best distribution.

3)I think we do not need any plot to describe binary variables like sex.Because knowing one category distribution will tell us the other category distribution(noOfFemales = total-noOfMales in our case).

II. Show using plot, what percentage of students have heights between 66 inches and 72 inches?

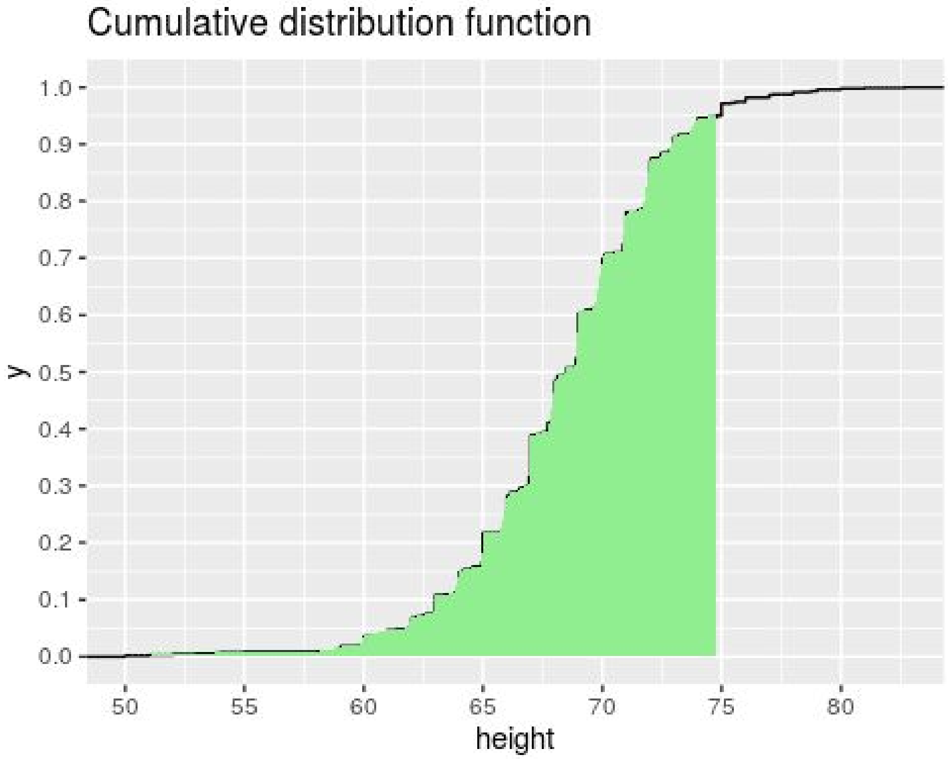
```
heights6672 <- heightsData[heightsData[,"height"]>=66 &
heightsData[,"height"]<=72,]
sixtySixTo72 <- nrow(heights6672)
rest <- nrow(heightsData)-nrow(heights6672)
percent_Val <-
c(sixtySixTo72*100/nrow(heightsData),rest*100/nrow(heightsData))
mybar <- barplot(percent_Val,names.arg = c('students with 66<=h<=72','rest of
students'),ylim=c(0,100),ylab='percentage',col=c('lightblue','pink'),main='pe
rcentage of students with height between 66 and 72 inches')
text(mybar, 0, round(percent_Val, 2),cex=1, pos=3)
```



III. What range of heights contains 95% of the data? Which graph is effective to show such analysis? Plot and give justification.

```r
cdf <- ggplot(heightsData, aes(x=height))
+stat_ecdf(geom="step")+ggtitle("Cumulative distribution function")+
  theme(panel.grid.minor = element_line(color = 'white')
        ,panel.grid.major = element_line(color = 'white'))+
  scale_x_continuous(breaks =
seq(min(heightsData$height),max(heightsData$height),by=5), limits =
c(min(heightsData$height),max(heightsData$height)))+
scale_y_continuous(breaks = seq(0,1,by=0.1), limits = c(0,1))
cdf.data <- (ggplot_build(cdf))$data[[1]][1:2]
cdf_95percent <- data.frame(dplyr::filter(cdf.data, cdf.data$x>=0 &
cdf.data$y<=0.95))
cdf+geom_area(data=data.frame(x=cdf_95percent[2],y=cdf_95percent[1]),
            aes(x=x, y=y), fill="lightgreen",main="cumulative distribution
function")
```

```
## Warning: Ignoring unknown parameters: main
```



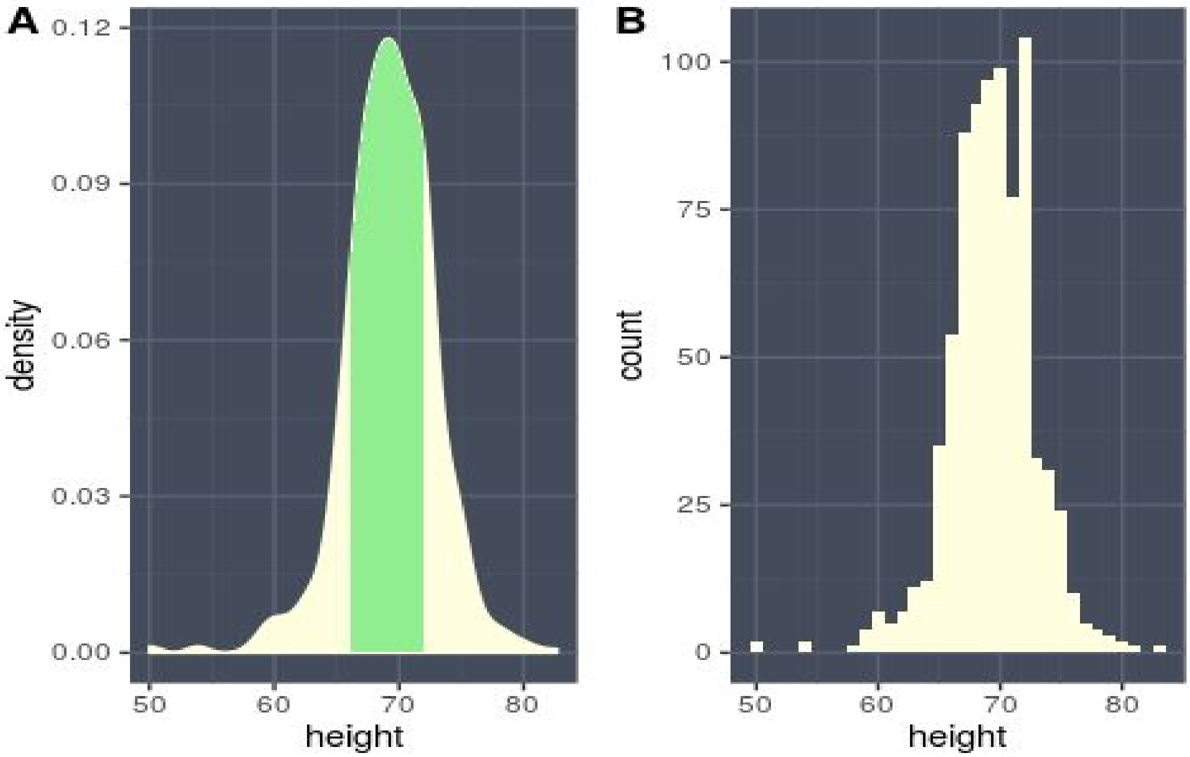
1)Cumulative distribution function will give us the area or percentage of students till some value x.

2) As we know highest value y can take is 1 when all the height data is completed.So at y = 0.95 we get 95% of data get covered.The corresponding x value tells us the maximum height at which 95% is covered.

3)As we can observe from the cumulative distribution function 95% of the data lies in the range 50 to approximately 75.

Q2)Use smoothed density curve to plot the height of Male students, highlighting the students with height between 66 and 72 inches. For the same data, plot and justify how is smoothed density graph different from histogram.

```r
library(tidyverse)

out<-data.frame(data.matrix(heightsData))
maleStudents <- out[out[, "sex"] == 2,]
str(maleStudents)

## 'data.frame':    812 obs. of  2 variables:
##  $ sex   : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ height: num  75 70 68 74 61 67 72 72 69 68 ...

density<-ggplot(maleStudents, aes(x=height)) +
  geom_density(fill = 'lightyellow', color = 'lightyellow')+
  theme(panel.background = element_rect(fill = '#444B5A')
        ,panel.grid.minor = element_line(color = '#4d5566')
        ,panel.grid.major = element_line(color = '#586174'))


density.data <- ggplot_build(density)
x1 <- min(which(density.data$data[[1]]$x >=66))
x2 <- max(which(density.data$data[[1]]$x <=72))
ggplot1 <-
density+geom_area(data=data.frame(x=density.data$data[[1]]$x[x1:x2],
                       y=density.data$data[[1]]$y[x1:x2]),
            aes(x=x, y=y), fill="lightgreen")

ggplot2<-ggplot(data=maleStudents)+
  geom_histogram(mapping=aes(x=height),binwidth = 1,fill="lightyellow")+
  theme(panel.background = element_rect(fill = '#444B5A')
        ,panel.grid.minor = element_line(color = '#4d5566')
        ,panel.grid.major = element_line(color = '#586174'))


cowplot::plot_grid(ggplot1, ggplot2, labels = "AUTO")
```

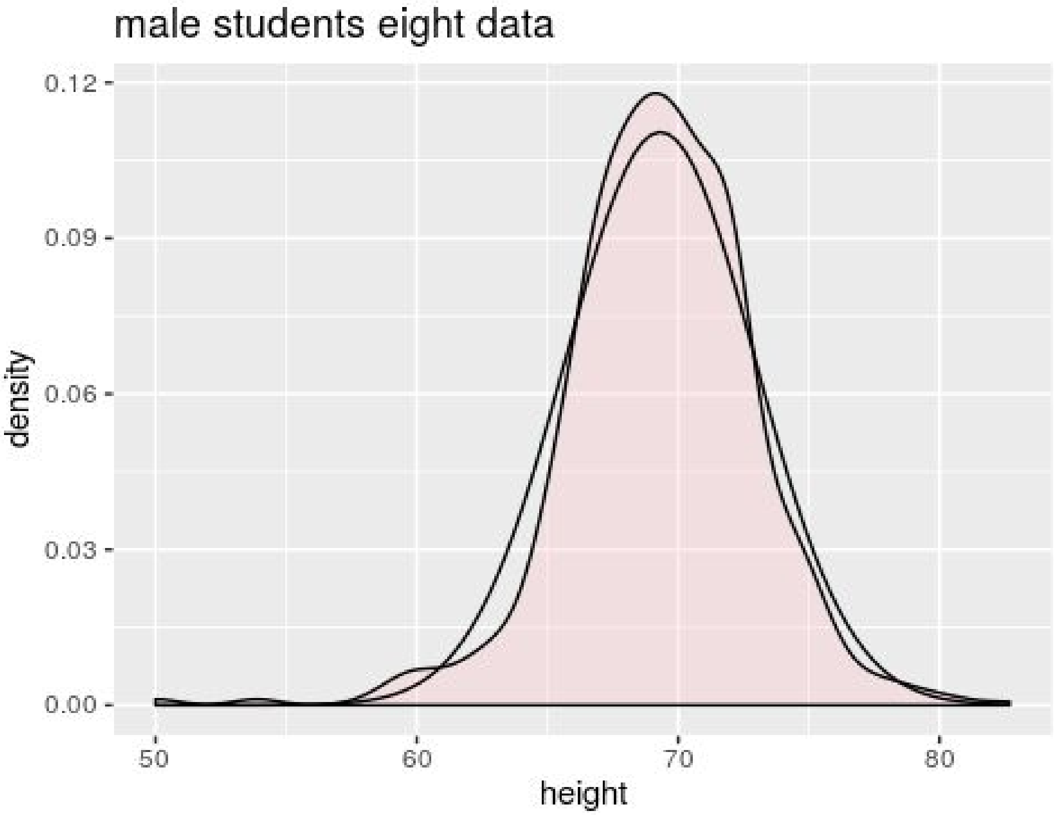


1) In the smooth density, we do not have sharp edges at the interval boundaries where as histogram has.

2) In smooth density a continuous curve is drawn at every individual data point and all of these curves are then added together to make a single smooth density estimation, where as in histogram bins are plotted at some range of discrete values(eg.66.01-66.99)

Q3.Does male height follow normal distribution? Justify your answer with suitable smoothed density plots. Also, answer what percentage of values lies within 1.5 standard deviation of the mean.

```r
library("tigerstats")

ggplot(maleStudents, aes(x = height)) +
  geom_density(fill = 'pink', color = 'black',alpha=0.3)+ggtitle("male
students eight data")+
  theme(panel.grid.major = element_line(size = 0.5, linetype = 'solid',
                                colour = "white"),
  panel.grid.minor = element_line(size = 0.25, linetype = 'solid',
                                colour = "white"))+

stat_function(fun=dnorm,args=list(mean(maleStudents$height),sd(maleStudents$h
eight)))
```



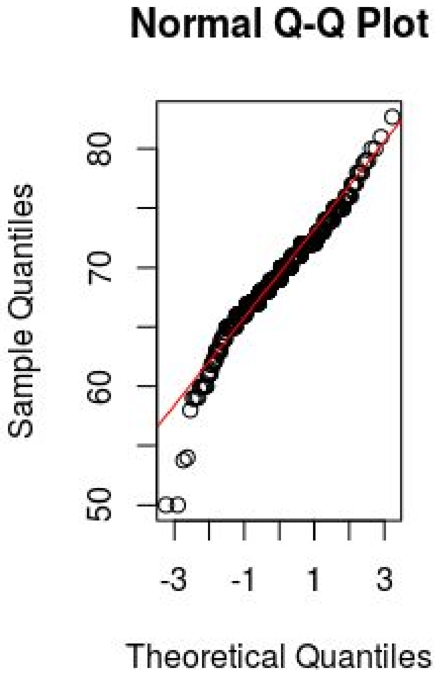
```
lower <- mean(maleStudents$height)-1.5
upper <- mean(maleStudents$height)+1.5
sdStudents <- filter(maleStudents,maleStudents$height>=lower &
maleStudents$height <= upper)
cat("\npercentage of students within 1.5 standard devaition from mean
is:",(nrow(sdStudents)*100)/nrow(maleStudents))

##

1)percentage of students within 1.5 standard devaition from mean is: 33.99015
```

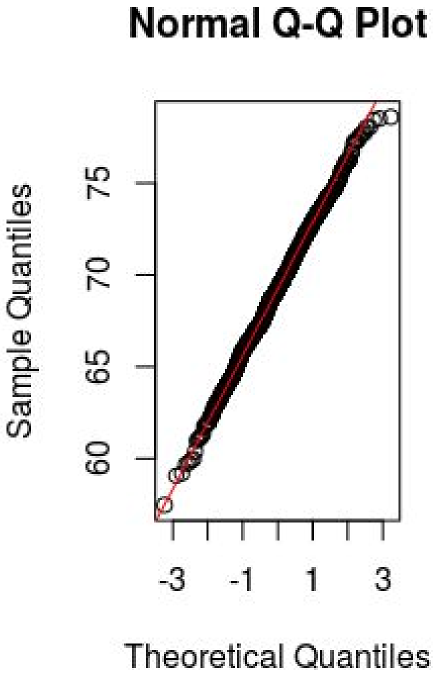
2)Yes male students height data almost follows normal distribution.

3)As you can see from the plot that pink curve is actual data which is approximately matching with the normal distribution with mean=mean(maleStudendsHeight) and stdev = stdev(maleStudentsHeight)

Q4.Plot a quantile-quantile plot (QQ Plot) to check whether the Male height distribution is well approximated by the normal distribution.

```
library(dslabs)
library(ggplot2)
par(mfrow=c(1,2))
qqnorm(maleStudents$height);
qqline(maleStudents$height, col = 2)
me = mean(maleStudents$height)
sd = sqrt(var(maleStudents$height))
data.norm <- rnorm(nrow(maleStudents),me,sd)
qqnorm(data.norm);
qqline(data.norm, col = 2)
```





We can see 2 qq plots here:

1.First one is the QQplot drawn using actual data

2.Second one is the QQplot drawn using by drawing some random samples of same size, same mean and variance as of our male students height from normal distribution

3.We can clearly see that in first plot our points are not fitting the straight line where as it is almost perfectly fitting in second plot

4.We can clearly say that male students height distribution is not well approximated using normal distribution