Solution

Libraries

```
library(dslabs)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(gghighlight)
library(UsingR)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units

##
## Attaching package: 'UsingR'
## The following object is masked from 'package:survival':
##
##     cancer
```

Question 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?

Ans: (i) For Gender Distribution,There is no need of Plot we can tell in percentage of gender distribution.but if we want to plot graph for gender distribution then bar plot is appropriate.

This two-category frequency table(sex & height) is the simplest form of a distribution. We don't really need to visualize it since one number describes everything we need to know: 23% are females and the rest are males. When there are more categories, then a simple barplot describes the distribution.

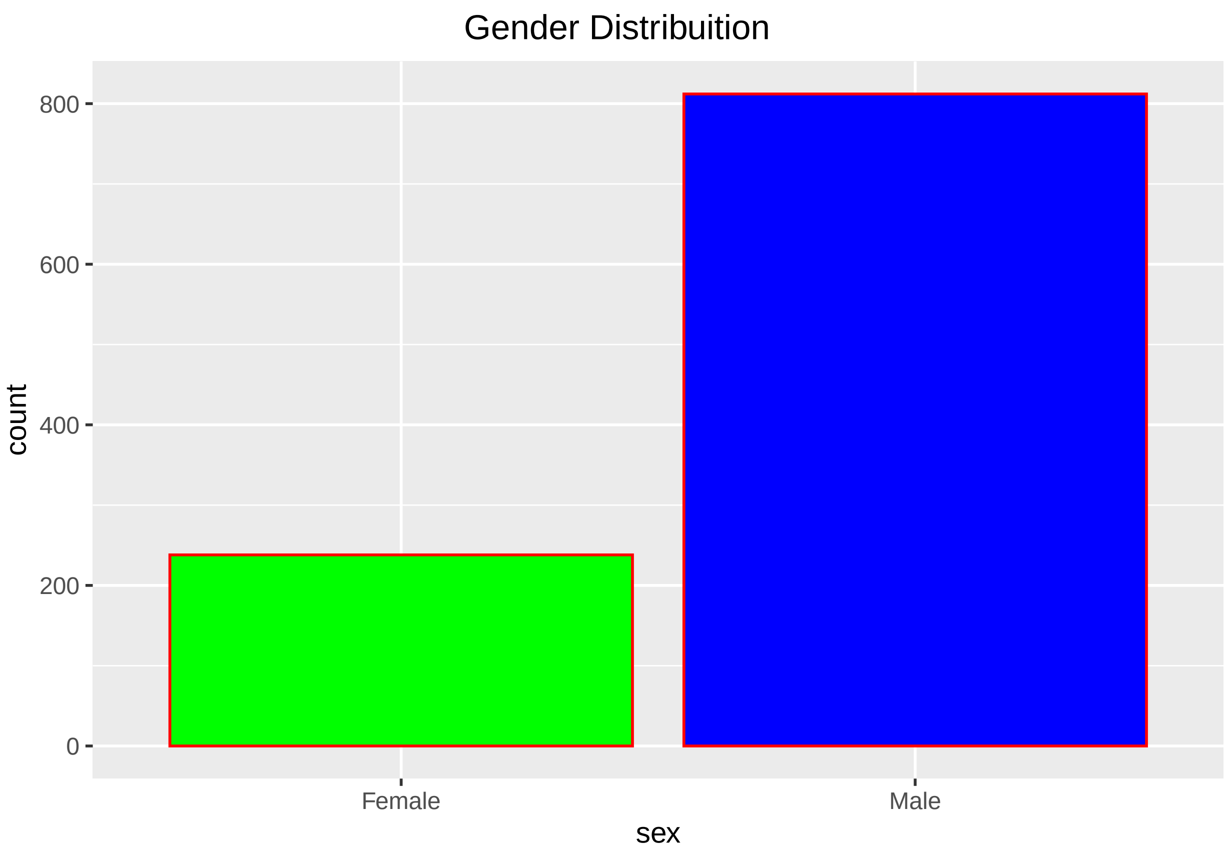
(ii) For Height Distribution, Density plot is appropraite.

An advantage Density Plots have over Histograms is that they're better at determining the distribution shape because they're not affected by the number of bins used (each bar used in a typical histogram)

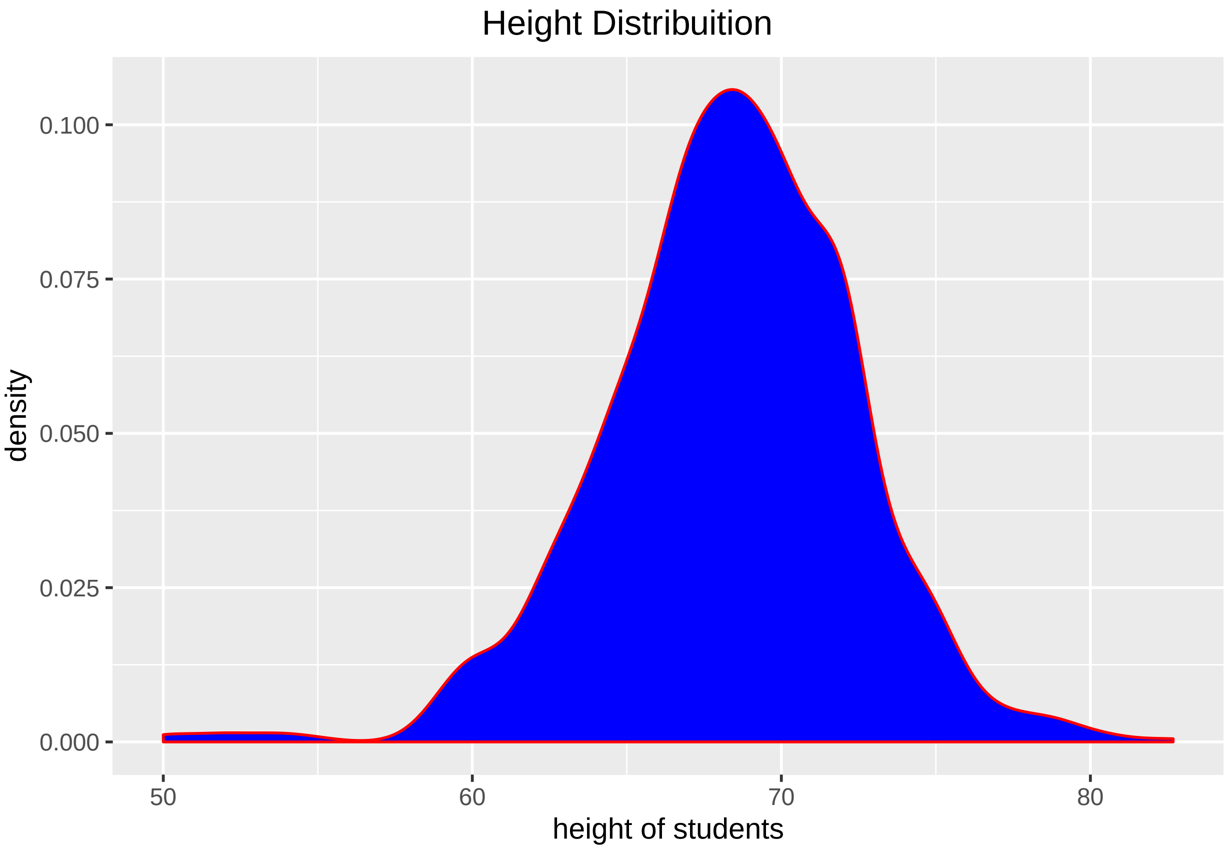
```
## proportion of male and female
prop.table(table(heights$sex))
```

```
##
##    Female      Male
## 0.2266667 0.7733333
```

```
##Gender Distribuition Plot
ggplot(data=heights)+geom_bar(mapping=aes(sex),color="red",fill=c("green","blue"))+
  ggtitle("Gender Distribuition")+
  theme(plot.title = element_text(hjust = 0.45))
```



```
## Height Distribution plot
ggplot(data=heights)+geom_density(mapping=aes(height),fill="blue",color="red")+
   ggtitle("Height Distribuition")+
  theme(plot.title = element_text(hjust = 0.45))+labs(x="height of students",y="density")
```



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

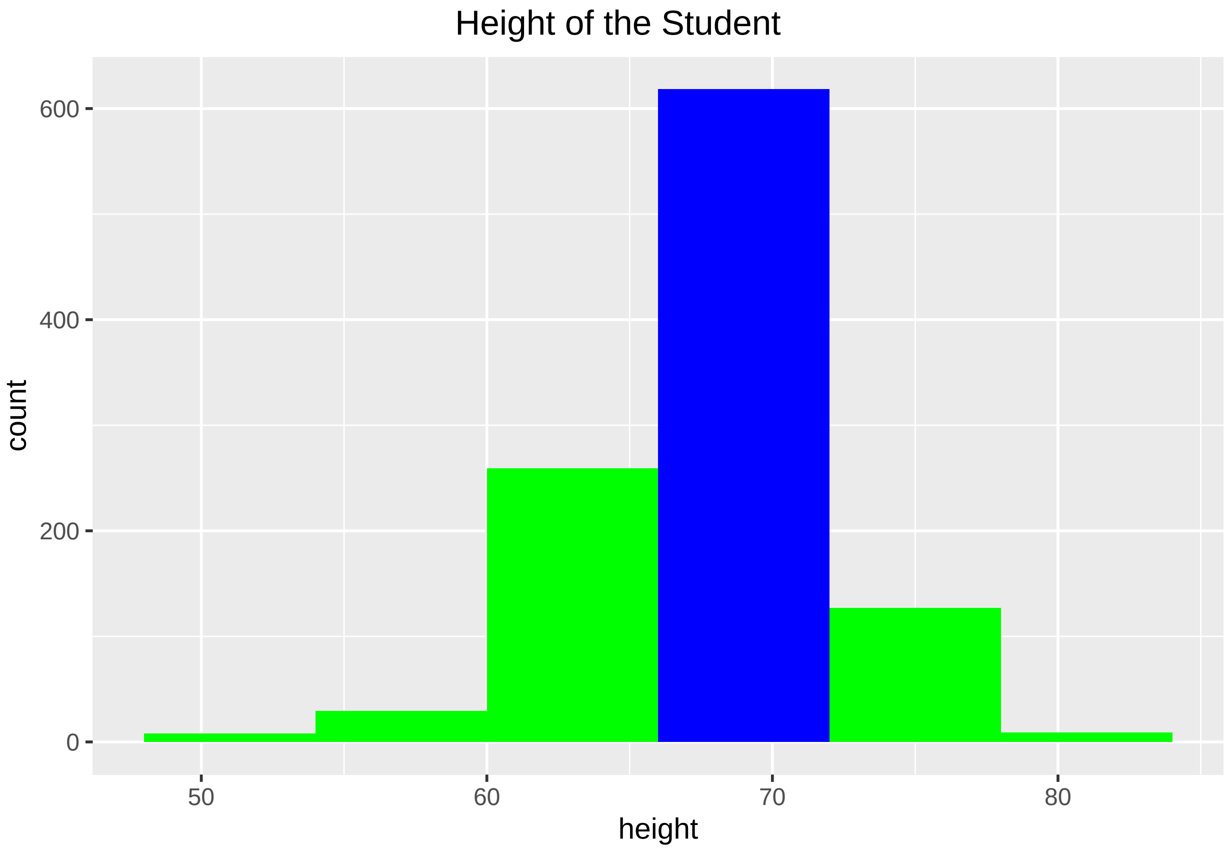
Ans: 65.04% of students have heights between 66 and 72 inches. In histogram Blue color shows height of the students between 66 inches and 72 inches. out of 1050, 683 student lies between 66 inches and 72 inches

```
## No of students lies between 66 inches and 72 inches
x <- heights$height[heights$height>=66 & heights$height<=72]
students=length(x)/nrow(heights)
students*100
```

```
## [1] 65.04762
```

```
## Histogram plot of students

ggplot(data=heights)+
  geom_histogram(mapping=aes(height),
      breaks=seq(48,84,by=6),fill=c("green","green","green","blue","green","green"))+
  ggtitle("Height of the Student")+
  theme(plot.title = element_text(hjust = 0.45))
```



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

Ans: The range of the data is from 50 to 84 with the majority (more than 95%) between 60 and 75 inches

Histograms are much preferred because they greatly facilitate answering such questions(What ranges contain 95% of the values?).

Histograms sacrifice just a bit of information to produce plots that are much easier to interpret.

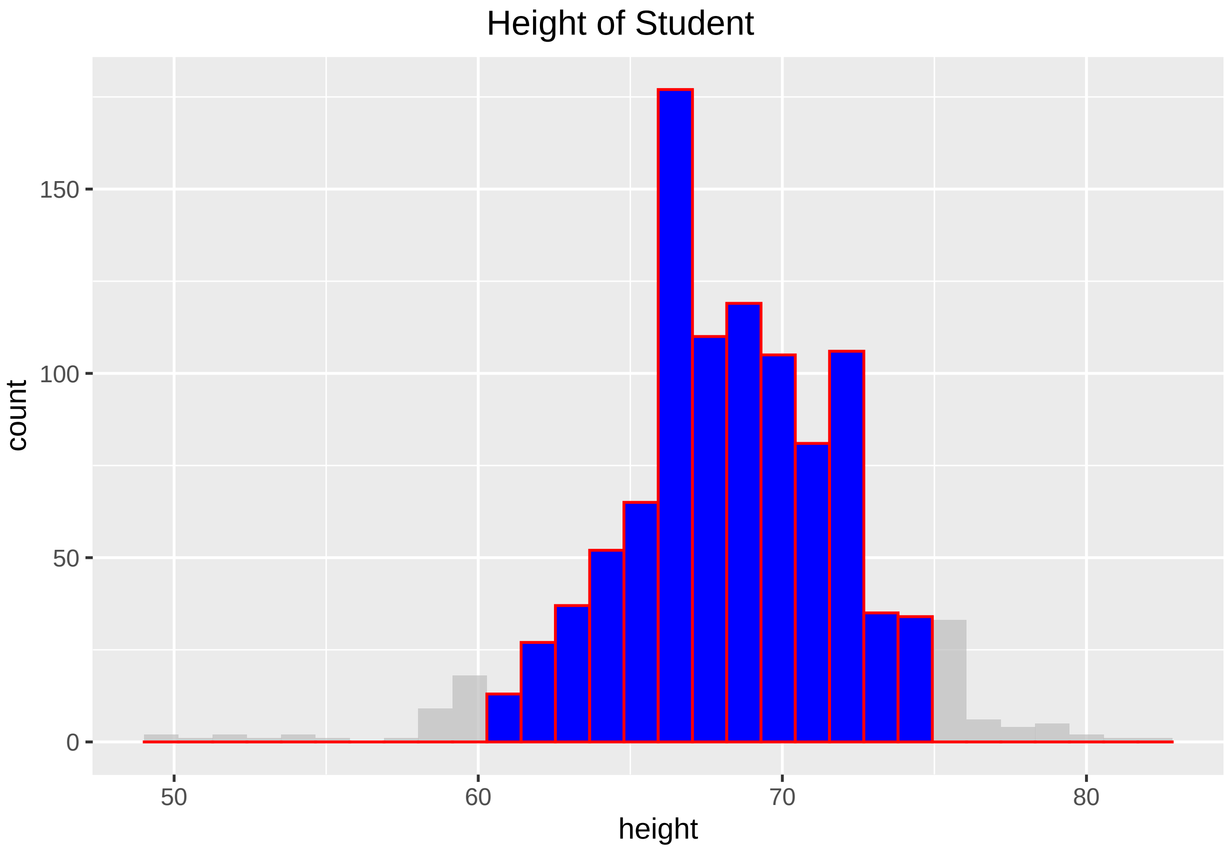
The heights are close to symmetric around 69 Therefore, the histogram above is not only easy to interpret, but also provides almost all the information contained in the raw list

In histogram plot “blue” colors shows majority of the students (95.23%, almost 1000 students out of 1050 students) lies in this range.

```
x <- heights$height[heights$height>=60 & heights$height<=75]
students=length(x)/nrow(heights)
students
```

```
## [1] 0.952381
```

```
ggplot(data=heights)+geom_histogram(mapping=aes(height),bins=30,color="red",fill="blue")+
  ggtitle(" Height of Student")+
  theme(plot.title = element_text(hjust = 0.45))+
  gghighlight(height>60 & height<75)
```



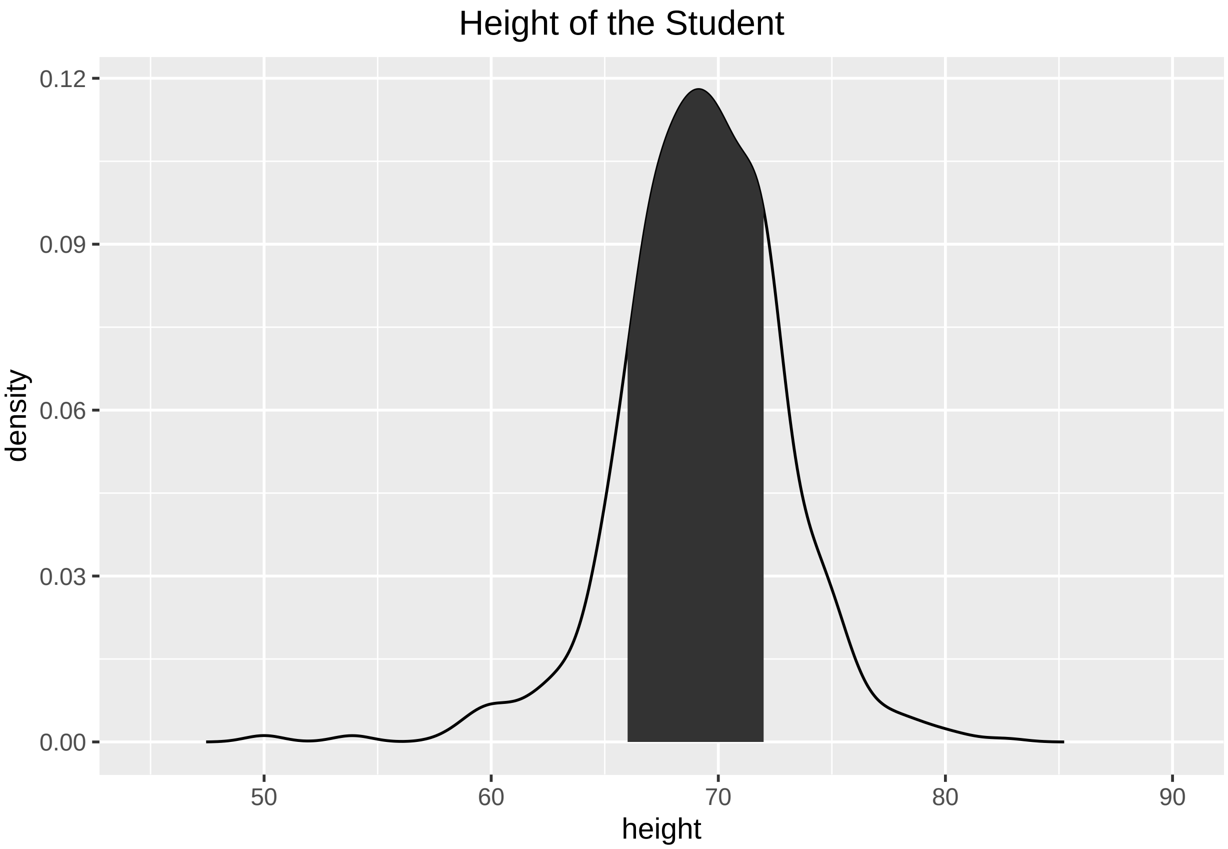
##Question 2 Use smoothed density curve to plot the height of Male students, highlighting the students with height between 66 and 72 inches. For the same data, plot and justify how is smoothed density graph different from histogram.

Ans: A Density Plot visualises the distribution of data over a continuous interval or time period. This chart is a variation of a Histogram that uses kernel smoothing to plot values, allowing for smoother distributions by smoothing out the noise. The peaks of a Density Plot help display where values are concentrated over the interval.

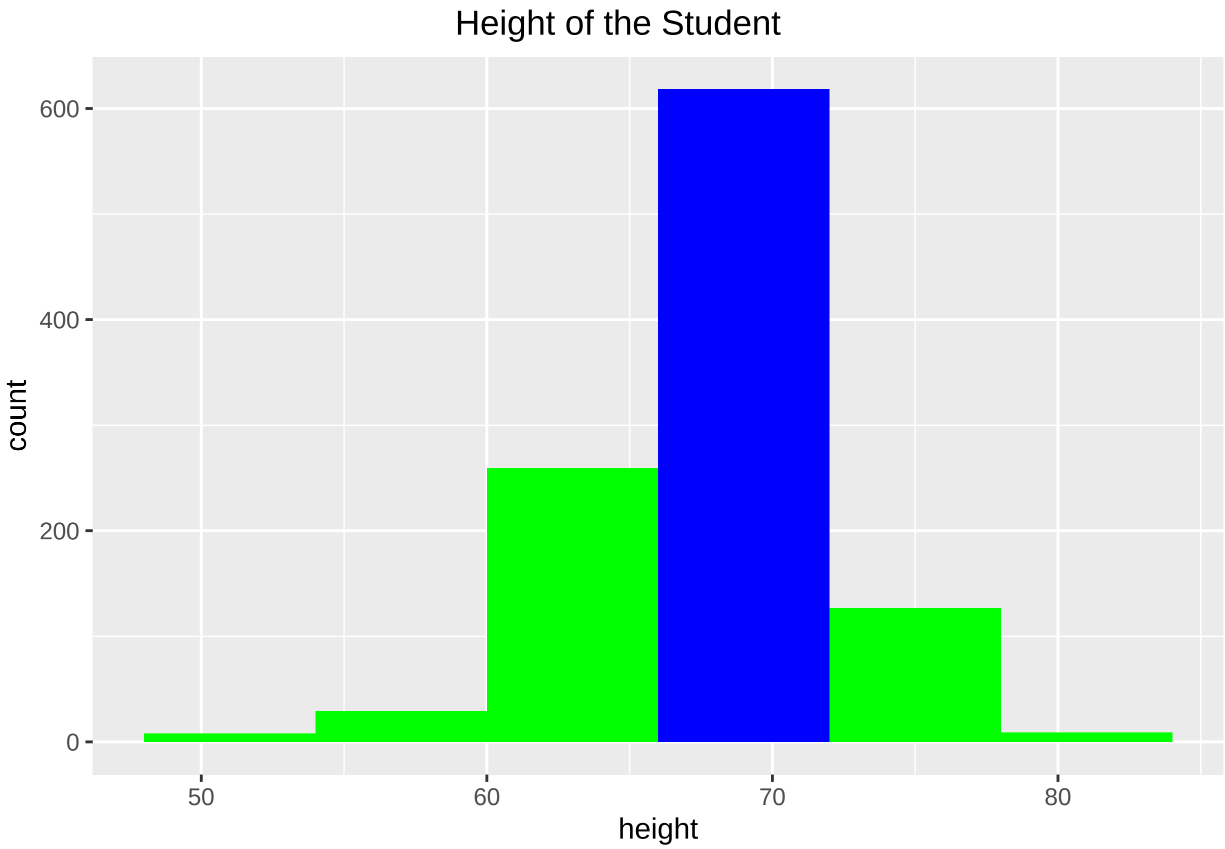
An advantage Density Plots have over Histograms is that they're better at determining the distribution shape because they're not affected by the number of bins used (each bar used in a typical histogram).

```r
## Density Plot
male<-subset(heights,sex=="Male")
data<-with(density(male$height),data.frame(x,y))
ggplot(data,aes(x=x,y=y))+
  geom_line()+
  geom_area(aes(x=ifelse(x>66 & x<72,x,0)))+xlim(45,90)+
  ggtitle("Height of the Student")+
  theme(plot.title = element_text(hjust = 0.45))+labs(x="height",y="density")
```

```
## Warning: Removed 430 rows containing missing values (position_stack).
```



```
##Histogram Plot
ggplot(data=heights)+
  geom_histogram(mapping=aes(height),
      breaks=seq(48,84,by=6),fill=c("green","green","green","blue","green","green"))+
  ggtitle("Height of the Student")+
  theme(plot.title = element_text(hjust = 0.45))
```



##Question 3 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.

Ans: Yes, Male height follow normal distribution. Height is sexually dimorphic and statistically it is more or less normally distributed, but with heavy tails.

In our dataset 88.79% of values lies within 1.5 standard deviation of the mean.

For an approximately normal data set,86.64% of values lies within 1.5 standard deviation of the mean

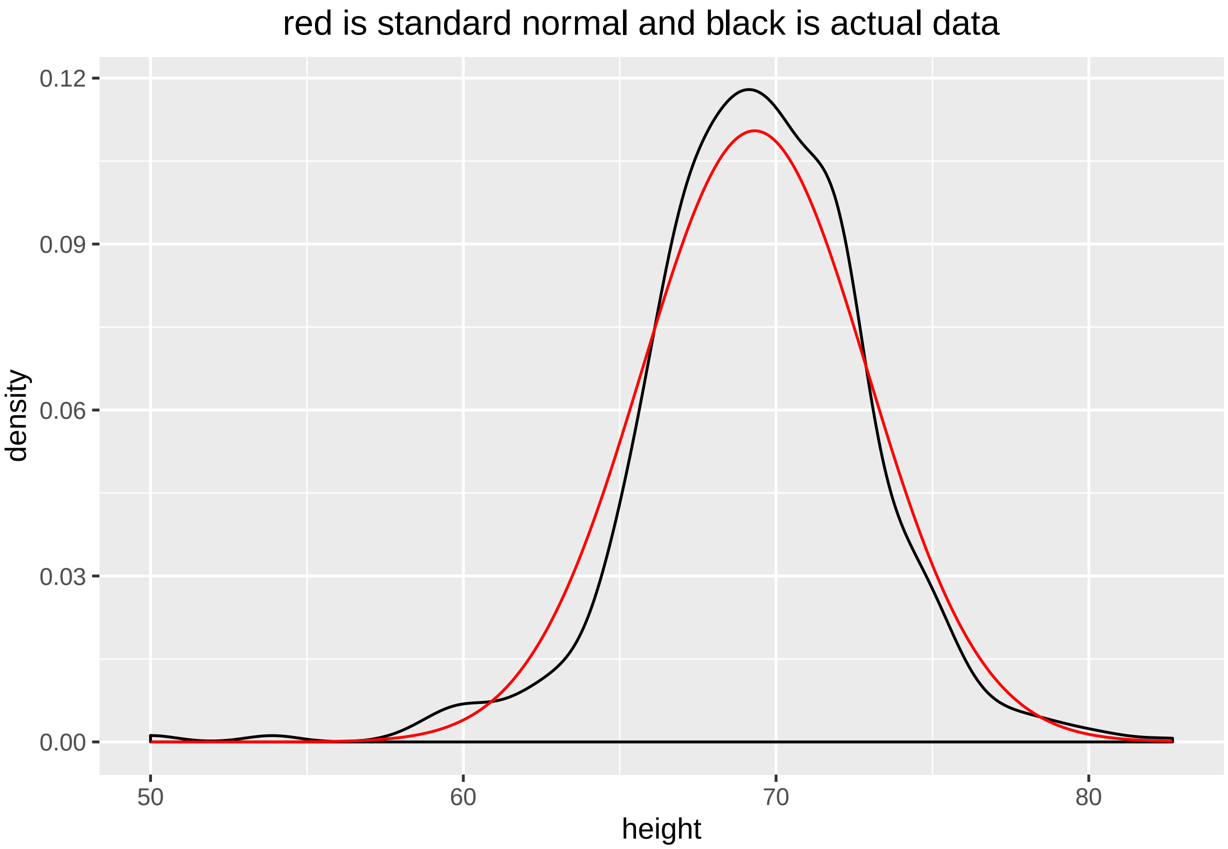
```
male_height=filter(heights,sex=="Male")

actual_plot<- ggplot(male_height, aes(x=height)) +
  geom_density(color="black")

x <- seq(min(male_height$height),
         max(male_height$height), length.out=800)

df <- with(actual_plot,
           data.frame(x = x, y = dnorm(x, mean(male_height$height), sd(male_height$height))))

actual_plot + geom_line(data = df, aes(x = x, y = y), color = "red")+labs(x ="height")+
ggtitle("red is standard normal and black is actual data")+
  theme(plot.title = element_text(hjust = 0.45))
```



```
values_lies_within= filter(male_height,abs(height-mean(male_height$height)) <= 1.5*sd(male_height$heigh

xy<-count(values_lies_within)*100/count(male_height)

 xy
```

```
##         n
## 1 88.7931
```

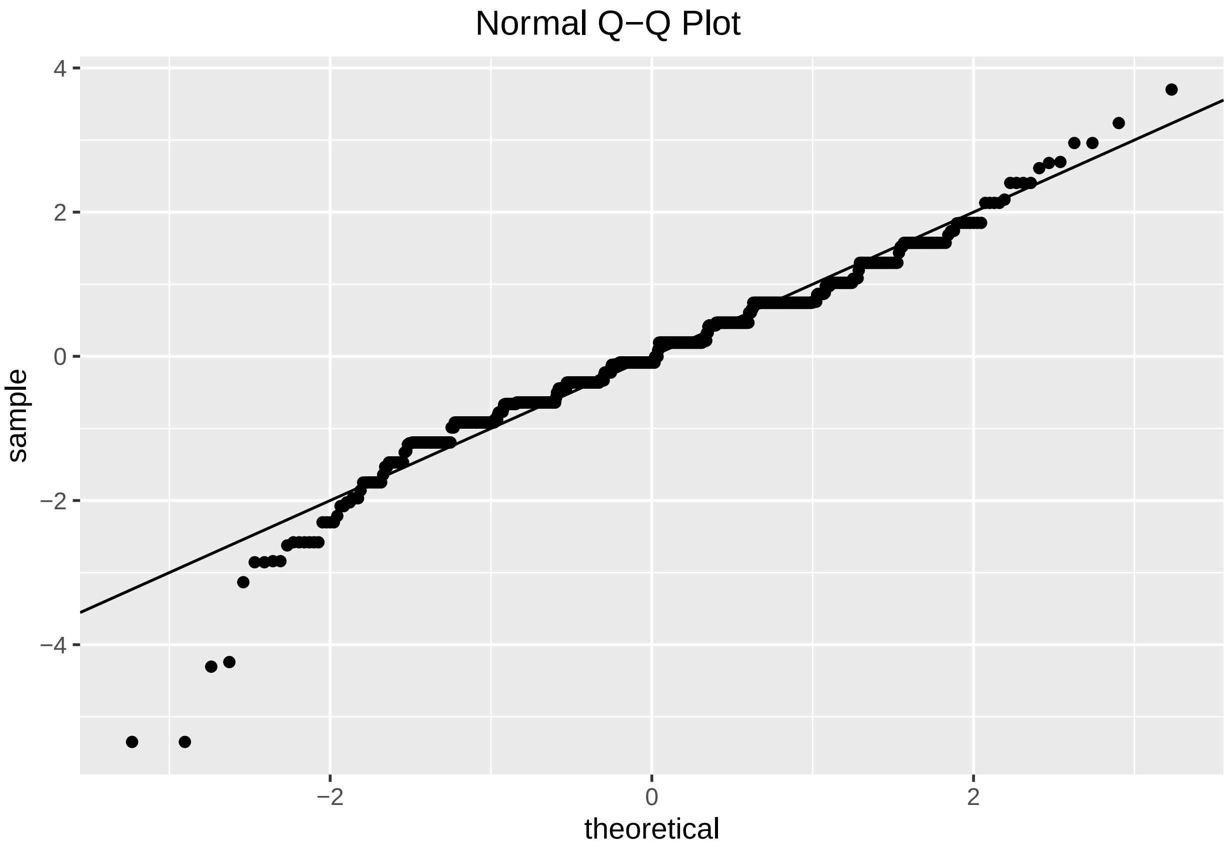
##Question 4 Plot a quantile-quantile plot (QQ Plot) to check whether the Male height distribution is well approximated by the normal distribution.

Ans:

Normal Q-Q Plot” provides a graphical way to determine the level of normality. The black line indicates the values your sample should adhere to if the distribution was normal. The dots are your actual data. If the dots fall exactly on the black line, then your data are normal

```
## QQ plot

heights %>%
  filter(sex=="Male") %>%
  ggplot(aes(sample = scale(height))) +
  geom_qq() +
  geom_abline()+
  ggtitle("Normal Q-Q Plot") +
  theme(plot.title = element_text(hjust = 0.45))
```


Normal Q−Q Plot
4
2
0
−2
−4
sample
−2
0
2
theoretical