TD1

HUON Sopanha

5. How People Get Their News The Brunswick Research Organization surveyed 50 randomly selected individuals and asked them the primary way they received the daily news. Their choices were via newspaper (N), television (T), radio (R), or Internet (I).Construct a categorical frequency distribution for the data and interpret the results.

```
news<- c("N", "N", "T", "T", "T", "I", "R", "R", "I", "T",
           "I", "N", "R", "R", "I", "N", "N", "I", "T", "N",
           "I", "R", "T", "T", "T", "T", "N", "R", "R", "I",
           "R", "R", "I", "N", "T", "R", "T", "I", "I", "T",
           "T", "I", "N", "T", "T", "I", "R", "N", "R", "T")
```

```
library(tidyverse)
library(ggplot2)
table(news)%>%
  as.tibble()%>%
  mutate(freq = n,perc = (n/sum(n)*100))%>%
  select(-n)%>%
  bind_rows(tibble(news="Total",freq=sum(.$freq),perc=sum(.$perc)))%>%
  knitr::kable(format = 'markdown')
```

news	freq	perc
I	12	24
N	10	20
R	12	24
T	16	32

news	freq	perc
Total	50	100

7. Ages of the Vice Presidents at the Time of Their Death The ages of the Vice Presidents of the United States at the time of their death are listed below.

90 83 80 73 70 51 68 79 70 71 72 74 67 54 81 66 62 63 68 57 66 96 78 55 60 66 57 71 60 85 76 98 77 88 78 81 64 66 77 93 70

(a) Use the data to construct a frequency distribution with 6 classes.

(b) Find the relative frequency.

(c) Construct a histogram, frequency polygon, and ogive.

Anwser

a-Use the data to construct a frequency distribution with 6 classes.

```
library(tidyverse)
Age <- c(90,83,80,73,70,51,68,79,70,71,72,74,67,
         54,81,66,62,63,68,57,66,96,78,55,60,66,57,
         71,60,85,76,98,77,88,78,81,64,66,77,93,70)
number =round((max(Age)-min(Age))/6)
breakss<-seq(min(Age),max(Age)+number,by=number)
Age%>%
  cut(breaks=breakss,right=FALSE)->Classes
table(Classes)%>%
  as.tibble()%>%
  mutate(Freq=n)%>%
  select(-n)->datasets
```

```
datasets%>%
  knitr::kable(format = 'markdown')
```

Classes	Freq
[51,59)	5
[59,67)	9
[67,75)	11
[75,83)	9
[83,91)	4
[91,99)	3

b-Find the relative frequency.

```
datasets%>%
  mutate(Relative_freq=round(Freq*100/41,2))%>%
  bind_rows(tibble(Classes="Total",
                   Freq=sum(.$Freq),
                   Relative_freq=as.integer(sum(.$Relative_freq)))
            )%>%
knitr::kable(format = 'markdown')
```

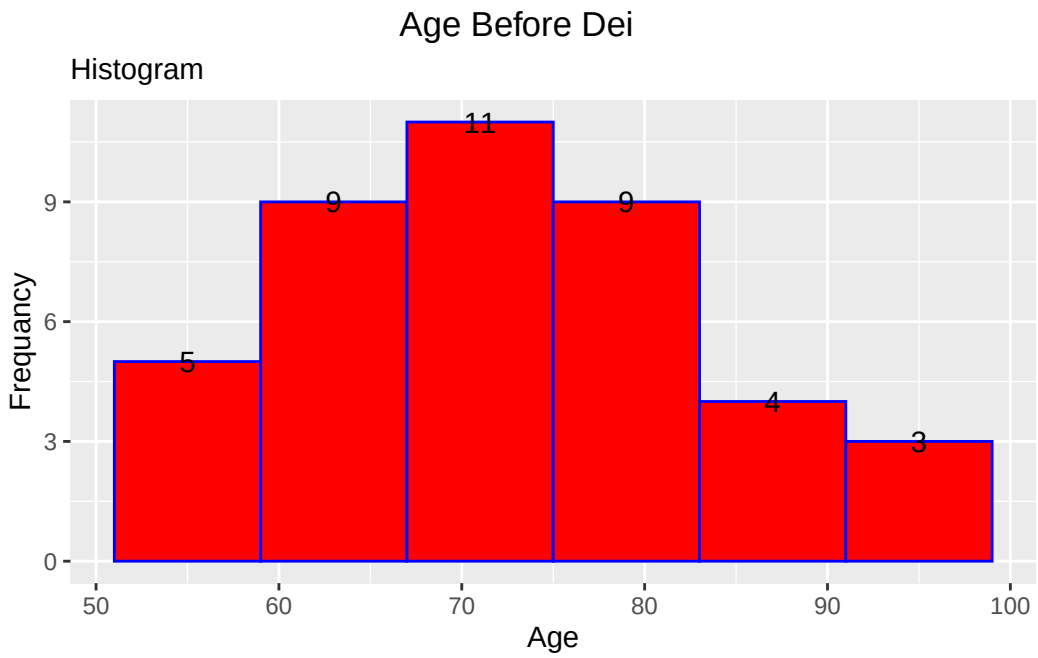
Classes	Freq	Relative_freq
[51,59)	5	12.20
[59,67)	9	21.95
[67,75)	11	26.83
[75,83)	9	21.95
[83,91)	4	9.76
[91,99)	3	7.32
Total	41	100.00

(c) Construct a histogram, frequency polygon, and ogive.

```
min=c()
for (n in 1:6) {
```

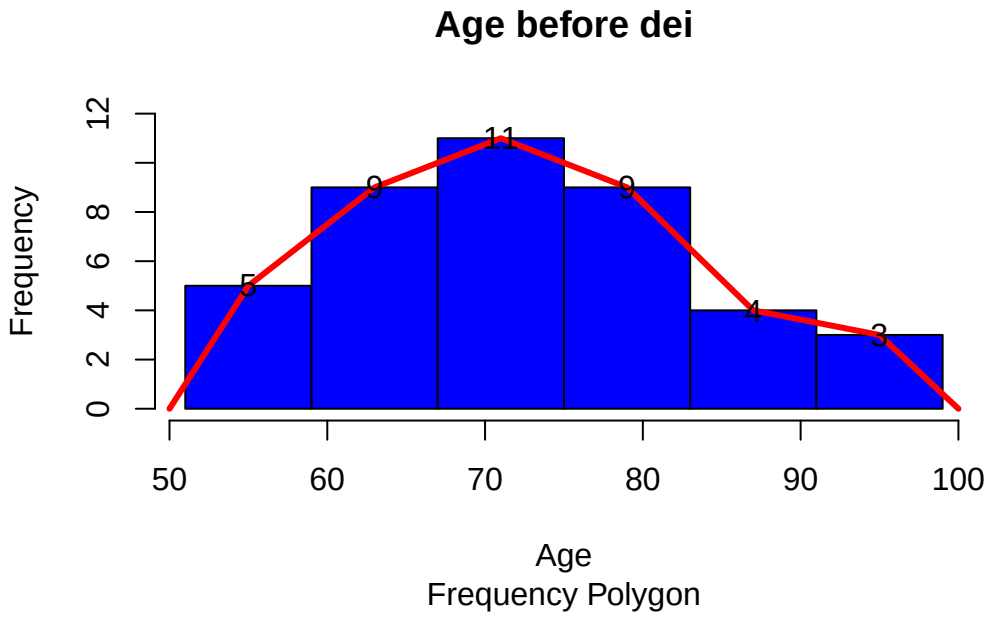
```
  r<-(breakss[n+1]+breakss[n])/2
  min=c(min,r)
}
Age%>%
  as.tibble()%>%
ggplot(aes(value)) +
  geom_histogram(breaks=breakss,
                 right=FALSE,
                 col='blue',
                 fill='red',)+
  ggtitle('Age Before Dei')+
  theme(plot.title = element_text(hjust=0.45))+
  labs(x='Age',y='Frequancy', subtitle = 'Histogram')+
  annotate("text", x =min, y=datasets$Freq, label =datasets$Freq)
```

```
Warning: The `right` argument of `stat_bin()` is deprecated as of ggplot2 2.1.0.
i Please use the `closed` argument instead.
```

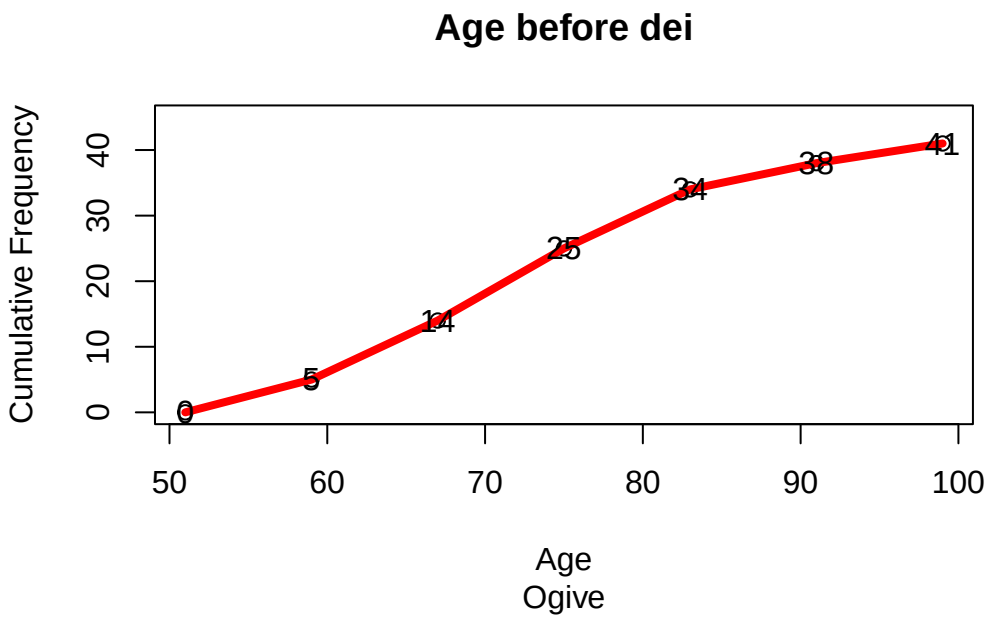



```
hist(Age,
     main='Age before dei',
     sub='Frequency Polygon',
```

```
     col="blue",
     xlab='Age',
     ylab='Frequency',
     ylim = c(0,12),
     breaks=breakss,
     right = FALSE)
lines(c(50,min,100),c(0,datasets$Freq,0),lwd=3,col='red')
text(min,datasets$Freq,datasets$Freq)%>%
knitr::kable(format = 'markdown')
```




```
c_freq = c(0,cumsum(table(Classes)))
plot(breakss, c_freq,
     xlab="Age",
     ylab="Cumulative Frequency",
     ylim=c(0,45))
title(main='Age before dei',sub='Ogive')
lines(breakss, c_freq,lwd=4,col='red')
text(breakss,c_freq,c_freq)
```

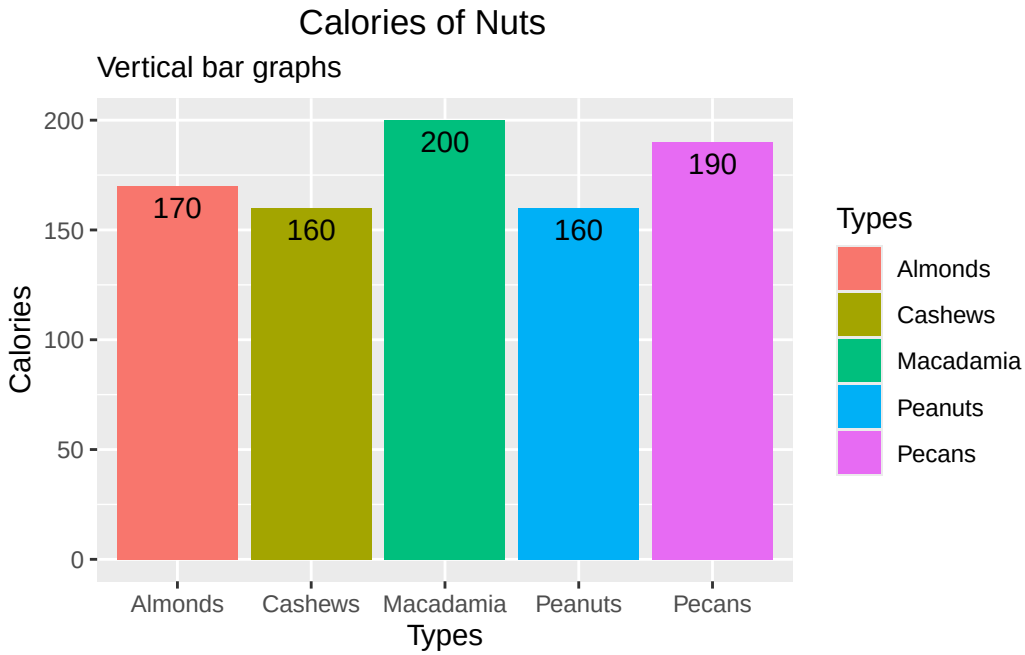


9. Calories of Nuts The data show the number of calories per ounce in selected types of nuts. Construct vertical and horizontal bar graphs for the data.

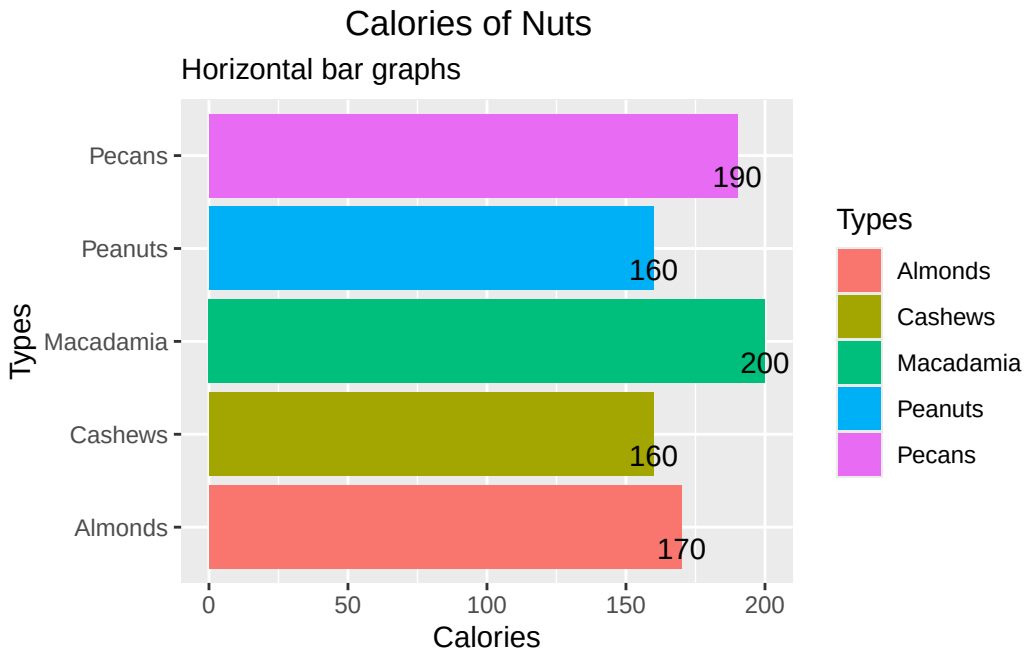
```
df<-tibble(
  Types=c('Peanuts','Almonds','Macadamia', 'Pecans','Cashews'),
   Calories=c(160,170,200,190,160)
)
knitr::kable(df,format = 'markdown')
```

Types	Calories
Peanuts	160
Almonds	170
Macadamia	200
Pecans	190
Cashews	160

```
df%>%
  ggplot(aes(Types,Calories,fill=Types))+
  geom_bar(stat = "identity")+
  ggtitle('Calories of Nuts',subtitle = "Vertical bar graphs")+
  theme(plot.title = element_text(hjust=0.45))+
  geom_text(aes(label=Calories),vjust=1.5)
```




```
df%>%
  ggplot(aes(Calories,Types,fill=Types))+
  geom_bar(stat = "identity")+
  ggtitle('Calories of Nuts',subtitle = "Horizontal bar graphs")+
  theme(plot.title = element_text(hjust=0.45))+
  geom_text(aes(label=Calories),vjust=1.5)
```



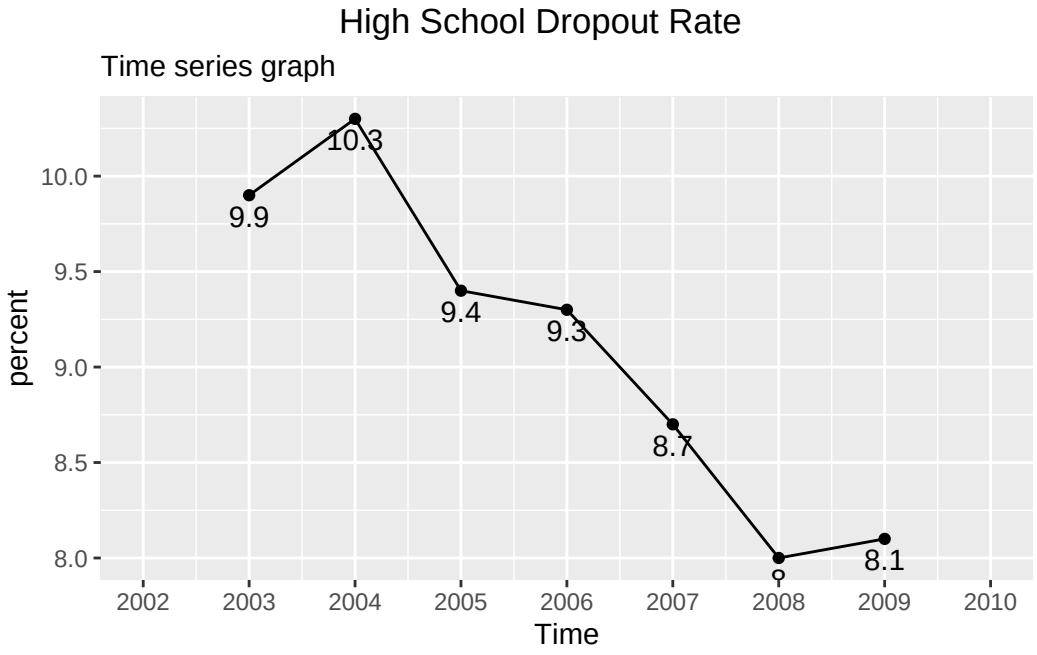
11. High School Dropout Rate The data show the high school dropout rate for students for

the years 2003 to 2009. Construct a time series graph and analyze the graph.

```
Rate<-tibble(
  Time=c(2003 ,2004 ,2005 ,2006 ,2007 ,2008 ,2009),
  percent = c(9.9 ,10.3 ,9.4 ,9.3, 8.7 ,8.0 ,8.1)
)
knitr::kable(Rate,format = 'markdown')
```

Time	percent
2003	9.9
2004	10.3
2005	9.4
2006	9.3
2007	8.7
2008	8.0
2009	8.1

```
ggplot(Rate,aes(Time,percent))+
  geom_line()+
  xlim(2002,2010)+
  geom_point()+
  geom_text(aes(label=percent),vjust=1.5)+
  scale_x_continuous(breaks = seq(2002,2010,by=1),limits = c(2002,2010))+
  ggtitle('High School Dropout Rate',subtitle = "Time series graph")+
  theme(plot.title = element_text(hjust=0.45))
```

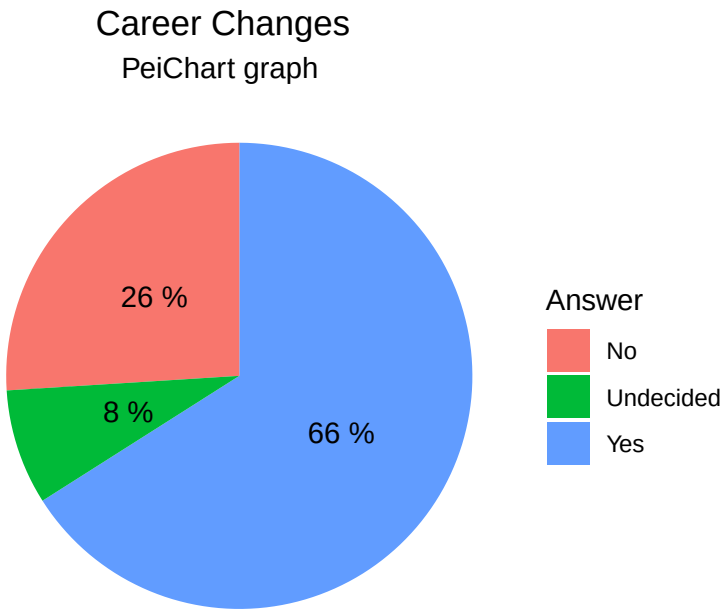


13. Career Changes A survey asked if people would like to spend the rest of their careers with their present employers. The results are shown. Construct a pie graph for the data and analyze the results.

```
Career<-tibble(
  Answer=c('Yes','No','Undecided'),
   people= c(660,260,80)
)
knitr::kable(Career,format = 'markdown')
```

Answer	people
Yes	660
No	260
Undecided	80

```
ggplot(Career,aes('',people,fill = Answer))+
  geom_bar(stat = "identity",width = 1)+
  coord_polar('y',start=0)+
  theme_void()+
  geom_text(aes(label=paste(people*100/sum(people), "%", sep=" ")),position=position_stack(v
    ggtitle('Career Changes',subtitle = "PeiChart graph")+
  theme(plot.title = element_text(hjust=0.45),plot.subtitle = element_text(hjust=0.45))
```



15. Songs on CDs The data show the number of songs on each of 40 CDs from the author's collection. Construct a dotplot for the data and comment on the graph.

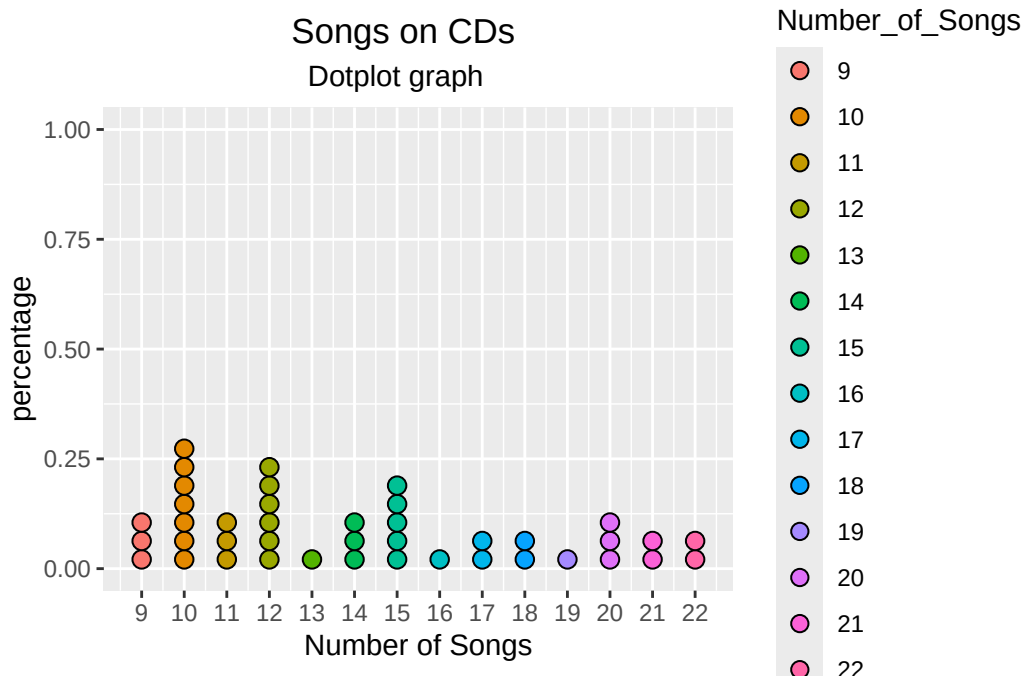
```
song<-c(10, 14, 18, 11, 11, 15, 16, 10, 10, 17, 10, 15, 22, 9, 14, 12, 18, 12, 12, 15, 21, 2
PM<-c('null','null','null','null',song)
M<-matrix(PM,nrow =9,ncol = 5 )
knitr::kable(M,format = 'markdown')
```

null	15	14	15	15
null	16	12	10	11
null	10	18	19	12
null	10	12	20	12
10	17	12	21	9
14	10	15	10	14
18	15	21	17	20
11	22	22	9	12
11	9	20	13	10

```
table(song)%>%
  as.tibble()%>%
  set_names('Number of song','freq')%>%
  bind_rows(tibble('Number of song'="Total",freq=sum(.$freq)))%>%
knitr::kable(format = 'markdown')
```

Number of song	freq
9	3
10	7
11	3
12	6
13	1
14	3
15	5
16	1
17	2
18	2
19	1
20	3
21	2
22	2
Total	41

```
df<-as.tibble(song)
Number_of_Songs<-factor(df$value)
ggplot(df,aes(x=value,fill =Number_of_Songs))+
  geom_dotplot()+
  xlab('Number of Songs')+
  ylab('percentage')+
  ggtitle(' Songs on CDs',subtitle = "Dotplot graph")+
  theme(plot.title = element_text(hjust=0.45),plot.subtitle = element_text(hjust=0.45))+
  xlim(8,23)+
  scale_x_continuous(breaks=seq(7,24,1))
```



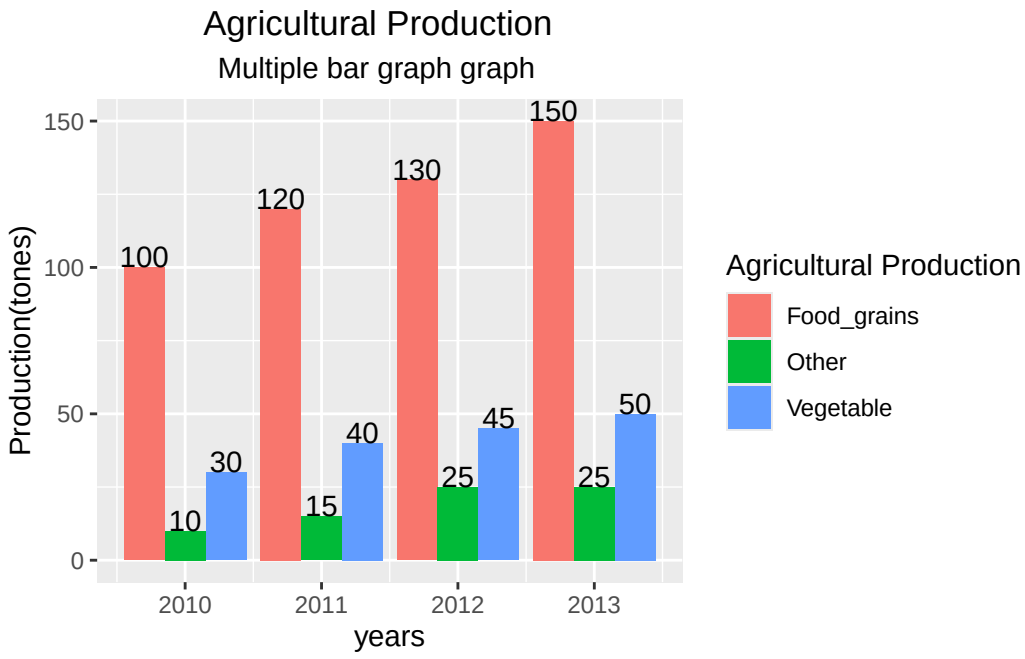
17. Draw a multiple bar graph for the following data which represented agricultural production for the priod from 2010-2013.

```
data<-tibble(
  years=c(2010,2011,2012,2013),
  Food_grains=c(100,120,130,150),
  Vegetable=c(30,40,45,50),
  Other=c(10,15,25,25)
```

```
)
knitr::kable(data,format = 'markdown')
```

years	Food_grains	Vegetable	Other
2010	100	30	10
2011	120	40	15
2012	130	45	25
2013	150	50	25

```
data%>%
  pivot_longer(cols=-years,
               names_to = 'Produce',
               values_to = 'Quantity')%>%
  ggplot(aes(x=years,y=Quantity,fill = Produce))+
  geom_col(position = 'dodge')+
  guides(fill=guide_legend(title = 'Agricultural Production'))+
  labs(y='Production(tones)')+
    geom_text(aes(label=Quantity),position=position_dodge(0.9),vjust=0)+
  ggtitle('Agricultural Production',subtitle = "Multiple bar graph graph")+
  theme(plot.title = element_text(hjust=0.45),plot.subtitle = element_text(hjust=0.45))
```



14.Peyton Manning's Colts Career Peyton Manning played for the Indianapolis Colts for 14 years. (He did no t play in 2011.) The data show the number of touch-downs he scored for the years 1998–2010. Construct a dotplot for the data and comment on the graph.

```
Peyton<-c(26 ,33 ,27, 49, 31, 27, 33, 26, 26, 29, 28, 31, 33)
PM<-c('null',Peyton)
M<-matrix(PM,nrow =2,ncol = 7 )
knitr::kable(M,format = 'markdown')
```

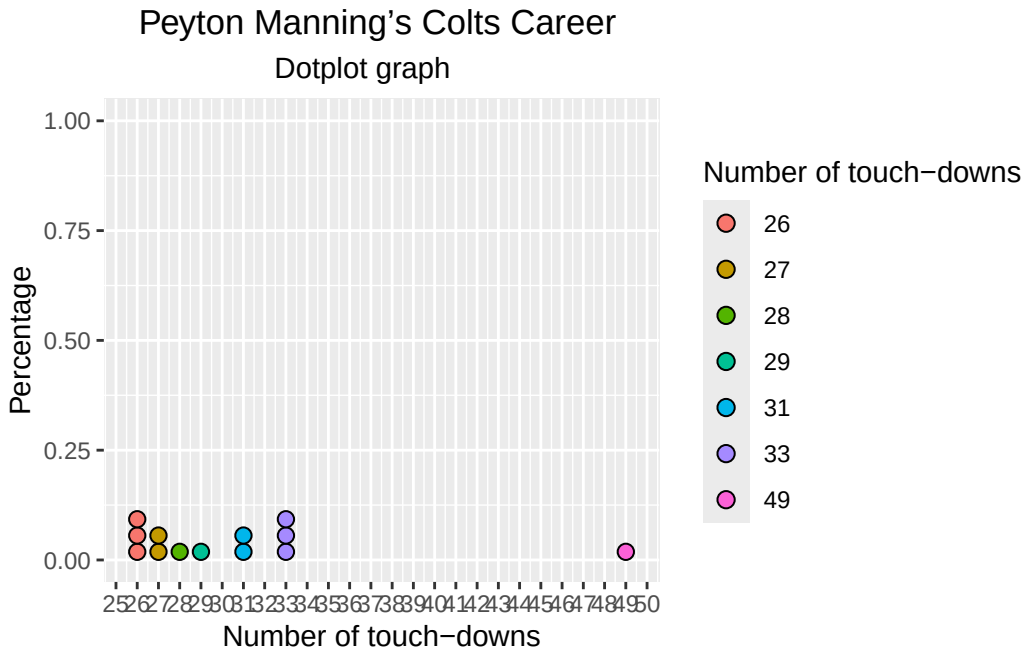
null	33	49	27	26	29	31
26	27	31	33	26	28	33

```
table(Peyton)%>%
  as.tibble()%>%
  set_names('number of touch-downs','freq')%>%
  bind_rows(tibble('number of touch-downs'="Total",freq=sum(.$freq)))%>%
knitr::kable(format = 'markdown')
```

number of touch-downs	freq
26	3
27	2
28	1
29	1
31	2
33	3
49	1
Total	13

```
df<-as.tibble(Peyton)
people<-factor(df$value)
ggplot(df,aes(x=value,fill =people))+
  geom_dotplot()+
  xlab('Number of touch-downs')+
  ylab('Percentage')+
  ggtitle('Peyton Manning's Colts Career',subtitle = "Dotplot graph")+
```

```
  theme(plot.title = element_text(hjust=0.45),plot.subtitle = element_text(hjust=0.45))+
  xlim(25,50)+
  scale_x_continuous(breaks=seq(25,50,1))+
    guides(fill=guide_legend(title = 'Number of touch-downs'))
```



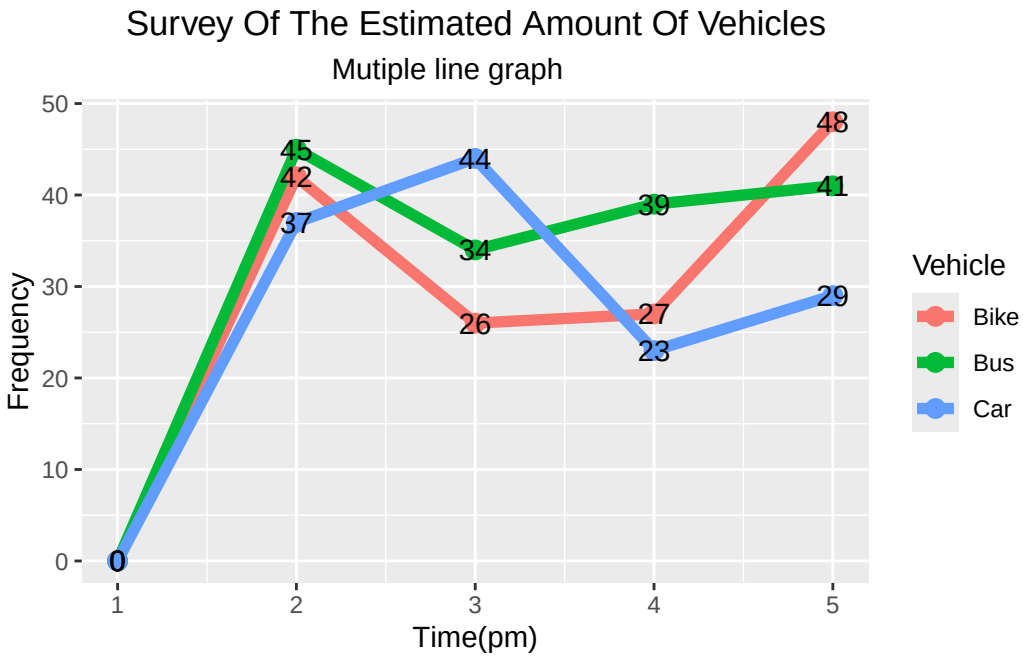
16. The traffic situation in X-City is getting worse, and it is high time a solution was offered.The company hired to work on the project took a survey of the estimated amount of vehicles that move on the road daily and for various intervals. The result of this survey is illustrated in the table below.Construct a multiple line graph to visualize the data. Hence, determine the vehicle with the highest frequency and that with the lowest frequency.

```
vehicle <-tibble(
  time = c('1-2pm','2-3pm','3-4pm','4-5pm'),
  Car = c(37,44,23,29),
  Bus = c(45,34,39,41),
  Bike =c(42,26,27,48)
```

```
)
knitr::kable(vehicle,format = 'markdown')
```

time	Car	Bus	Bike
1-2pm	37	45	42
2-3pm	44	34	26
3-4pm	23	39	27
4-5pm	29	41	48

```
vehicle$time<-c(2,3,4,5)
vehicle%>%
bind_rows(tibble(time=1,Car=0,Bus=0,Bike=0))->vehicle
vehicle%>%
    pivot_longer(cols=-time,
                 names_to = 'Vehicle',
                 values_to = 'Frequency')->df
df[order(df$Vehicle,decreasing = FALSE),]->df
df%>%
  ggplot(aes(x=time,y=Frequency,colour = factor(Vehicle)))+
  geom_line(size=2)+
  geom_point(size=3)+
  annotate("text", x =df$time, y=df$Frequency, label =df$Frequency)+
  guides(colour=guide_legend(title = 'Vehicle'))+
  xlab('Time(pm)')+
  ggtitle(' Survey Of The Estimated Amount Of Vehicles',subtitle = "Mutiple line graph")+
  theme(plot.title = element_text(hjust=0.45),plot.subtitle = element_text(hjust=0.45))
```


Survey Of The Estimated Amount Of Vehicles
Mutiple line graph
Frequency
Time(pm)
Vehicle
Bike
Bus
Car
0
45
42
37
44
34
26
39
27
23
48
41
29

At 5pm:

-Bike is the hightest frequency.

-Car is the lowest frequency.

At 4pm:

-Bus is the hightest frequency.

-Car is the lowest frequency.

At 3pm:

-Car is the hightest frequency.

-Bike is the lowest frequency.

At 2pm:

-Bus is the hightest frequency.

-Car is the lowest frequency.

19.Households of Four Television Networks Asurvey showed the number of viewers and number of households off our television networks. Find the average number of viewers,usingtheweightedmean.

```
Huoseholds<-c(1.4,0.8,0.3,1.6)
Viewers_in_millions<-c(1.6,0.8,0.4,1.8)
df<-tibble(Huoseholds,Viewers_in_millions)
knitr::kable(df,format = 'markdown')
```

Huoseholds	Viewers_in_millions
1.4	1.6
0.8	0.8
0.3	0.4
1.6	1.8

```
sum1<-0
sum2<-0
for (i in (1:4)){
  sum1=sum1+(Huoseholds[i]*Viewers_in_millions[i])
  sum2=sum2+(Huoseholds[i])
}
message("Average depends on weight is ",round(sum1/sum2,2) )
```

```
Average depends on weight is 1.43
```

26. Checkeachdatasetforoutliers.

(a) 14,18,27,26,19,13,5,25
(b) 112,157,192,116,153,129,131

```
check_int<-function(a){
  n=(a*10)%%10
  if(n==0){
    return(TRUE)
  }else{
    return(FALSE)
  }
}
```

```
Q<-function(a){
    a<-sort(a)
  n=length(a)
  c1=(n*25)/100
  c3=(n*75)/100
  if(check_int(c1)==TRUE){
    Q1=(a[c1]+a[c1+1])/2
  }else{
    Q1=a[round(c1,0)]
```

```
  }
  if(check_int(c3)==TRUE){
    Q3=(a[c3]+a[c3+1])/2
  }else{
    Q3=a[round(c3,0)]
  }
  return(c(Q1,Q3))
}
find_outier<-function(a){
  a<-sort(a)
  n=length(a)
  Q13=Q(a)
  Q1=Q13[1]
  Q3=Q13[2]
  IQR = Q3 - Q1
  Domain = seq(round(Q1-1.5*IQR,0),round(Q3+1.5*IQR,0))
  outier<-c()
  for (i in  1:n){
      if(a[i]%in%Domain==FALSE){
        outier<-c(outier,a[i])
    }
  }
  if(length(outier)==0){
    message('There is not outier in data')
  }else if(length(outier)==1){
    message('Outier in data is ', outier)
  }else{
    message('Outier in data are: ')
    for (i in 1:length(outier)) {
      print(outier[i])
    }
  }
}
```

(a)14,18,27,26,19,13,5,25

```
a<-c(14,18,27,26,19,13,5,25)
find_outier(a)
```

```
There is not outier in data
```

(b) 112,157,192,116,153,129,131

```
b<-c(112,157,192,116,153,129,131)
find_outier(b)
```

```
There is not outier in data
```

27. The following sample data are them id term examination test scores for 30 students:

55 60 91 85 60 70 89 99 59 67

72 82 60 68 57 74 64 70 68 91

89 90 83 40 79 85 71 80 76 81

a. Find the mean,mode,median,variance,standard deviation, 1,and 3 of the data.

b. Construct a frequency table with 5 classes.

c.Using the grouped data formula,find the mean,mode,median,variance,standard deviation, 1,and 3 for the table in part(b)and compare it to the results in part(a).

d. Construct a histogram and comment on the shape of the distribution.

e. Find the percentile values of 55,60,and 74.

```
mean<-function(a){
  return(sum(a)/length(a))
}
mode<-function(a){
  table(a)%>%
    as.tibble()->df
  r<-max(df[,'n'])
  return(df[df$n==r,'a'])
}
median<-function(a){
  sort(a)->a
  n<-length(a)
  if(n%%2==0){
    m=(a[n/2]+a[(n/2)+1])/2
  }else{
```

```
    m=a[(n+1)/2]
  }
  return(m)
}
varaint<-function(a){
  mean(a)->n
  s=c()
  for (i in 1:length(a)) {
    s=c(s,(abs(a[i]-n))^2)
  }
  return(sum(s)/(length(a)-1))
}
classify<-function(a,n){
  number =round((max(a)-min(a))/n)
  breakss<-seq(min(a),max(a)+number,by=number)
  a%>%
    cut(breaks=breakss,right=FALSE)->Classes
  table(Classes)%>%
    as.tibble()%>%
    mutate(Freq=n)%>%
    select(-n)%>%
    return()
}
```

a. Find the mean,mode,median,variance,standard deviation, 1,and 3 of the data.

```
data<-c(55,60,91,85,60,70,89,99,59,67,72,82,60,68,57,74,64,70,68,91,89,90,83,40,79,85,71,80,
data1<-data
message('Mean of data is ',round(mean(data),2))
```

```
Mean of data is 73.83
```

```
mode(data)%>%
  mutate(Mode=a)%>%
  select(-a)%>%
  knitr::kable(format = 'markdown')
```

Mode
60

```
message("Median of data is ",median(data))
```

```
Median of data is 73
```

```
message('Varaint of data is ',round(varaint(data),3))
```

```
Varaint of data is 182.557
```

```
message('Standaed deviation of data is ',round(sqrt(varaint(data)),2))
```

```
Standaed deviation of data is 13.51
```

```
Q13=Q(data)
Q1=Q13[1]
Q3=Q13[2]
message('Q1 of data is ',Q1)
```

```
Q1 of data is 64
```

```
message('Q3 of data is ',Q3)
```

```
Q3 of data is 83
```

b. Construct a frequency table with 5 classes.

```
knitr::kable(classify(data,5),format = 'markdown')
```

Classes	Freq
[40,52)	1
[52,64)	6

Classes	Freq
[64,76)	9
[76,88)	8
[88,100)	6

c.Using the grouped data formula,find the mean,mode,median,variance,standard deviation, 1,and 3 for the table in part(b)and compare it to the results in part(a).

```
midpoint<-function(b){
    mid<-c()
  for (i in 1:length(b)-1) {
    r<-(b[i]+b[i+1])/2
    mid<-c(mid,r)
  }
    return(mid)
}
mean_class<-function(a,b){
  mid=midpoint(b)
  a%>%
    mutate(Mid_range=mid)%>%
    mutate(XM=Freq*Mid_range)->a
a$XM%>%
  sum()->sum
  return(sum/sum(a$Freq))
}
varaint_class<-function(a,b){
    mid=midpoint(b)
    mean_class(a,b)->n
  a%>%
    mutate(Mid_range=mid)%>%
    mutate(V=(Mid_range-n)^2)->a
  sum(a$V)/(sum(a$Freq)-1)%>%
    return()
}
```

```
number =round((max(data)-min(data))/5)
b<-seq(min(data),max(data)+number,by=number)
data<-classify(data,5)
```

```r
message('Mean of data is ',round(mean_class(data,b),2))
```

```
Mean of data is 74.8
```

```r
message('Varaint of data is ',round(varaint_class(data,b),2))
```

```
Varaint of data is 53.63
```

```r
message('Standaed deviation of data is ',round(sqrt(varaint_class(data,b)),2))
```

```
Standaed deviation of data is 7.32
```

Median

$Q_i = l + \frac{(\frac{i \times n}{4} - CF)}{f} \times h$

n is number of all data

```r
data%>%
  mutate(upper=b[2:length(b)])->x
# find comulative
cf<-c(x$Freq[1])
  for (i in 1:length(x$Freq)-1) {
    r<-cf[i]+x$Freq[i+1]
    cf<-c(cf,r)
  }
x%>%
  mutate(CF=cf)->x
# find all data
x$Freq%>%
    sum()/2->t # find medium
# t=15 close to 16
Cf=7 # Commulative of medium
l=64 # lower of Medium
f=9 # frequency of medium
median=l+((t-Cf)/f)*number
message("Median of data is ", round(median,2))
```

```
Median of data is 74.67
```

Q1 and Q3

```
x$Freq%>%
    sum()/4->t # find position Q1
# t=7.5 in 64-76
Cf=1 # Commulative of lower Q1
l=52 # lower of Q1
f=6 # frequency of Q1
Q1=l+((t-Cf)/f)*number
message("Q1 of data is ", round(Q1,2))
```

```
Q1 of data is 65
```

```
x$Freq%>%
    sum()*3/4->t # find position
# t=22.5 in 76-88
Cf=16 # Commulative of Q3
l=76 # lower of Q3
f=8 # frequency of Q3
Q3=l+((t-Cf)/f)*number
message("Q3 of data is ", round(Q3,2))
```

```
Q3 of data is 85.75
```

#Mode

$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

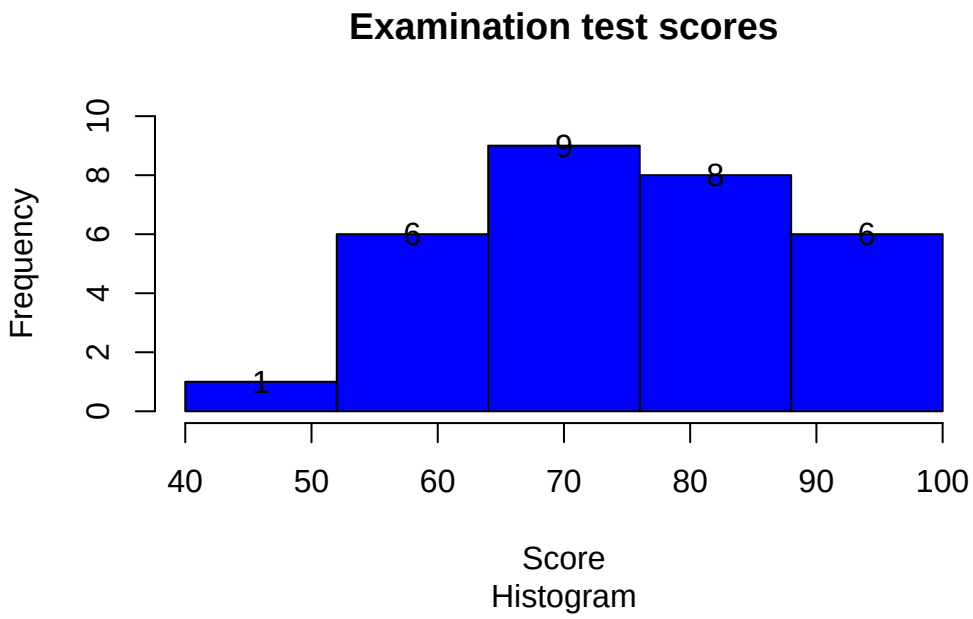
```
# highest frequency is 9 in 64-76
l=64
f1=9
f2=8# upper frequency column
f0=6# lower frequency column
Mode=l+number*(f1-f0)/(2*f1-f0-f2)
message("Mode of data is ", round(Mode,2))
```

```
Mode of data is 73
```

By part B and part A , We can said that the gruoped data have better statistic than the point data.It mean each data in part B are closer than each data in part A. so Part B is better than Part A

d. Construct a histogram and comment on the shape of the distribution.

```
hist(data1,
     main='Examination test scores',
     sub='Histogram',
     col="blue",
     xlab='Score',
     ylab='Frequency',
     ylim=c(0,10),
     breaks=b,
     right = FALSE)
text(midpoint(b),data$Freq,data$Freq)%>%
knitr::kable(format = 'markdown')
```




```
sk=(mean_class(data,b)-median)*3/sqrt(varaint_class(data,b))
message('Skewness = ',round(sk,2))
```

```
Skewness = 0.05
```

By the Skewness value we can said that graph is not symmetric.Therefor it is not approximate normal distribution.

e. Find the percentile values of 55,60,and 74.

```
percentile<-function(a,n){
  a<-sort(a)
  count=0
  for(i in 1:length(a)){
    if(a[i]<n){
      count=count+1
    }
  }
  return(round((count+0.5)*100/length(a),2))
}
```

```
message('Percentile of 55 is ',percentile(data1,55),'%')
```

```
Percentile of 55 is 5%
```

```
message('Percentile of 60 is ',percentile(data1,60),'%')
```

```
Percentile of 60 is 15%
```

```
message('Percentile of 74 is ',percentile(data1,74),'%')
```

```
Percentile of 74 is 51.67%
```

28. For the following data:

6.3 2.9 4.5 1.1 1.8 4.0 1.2 3.1 2.0 4.0

7.0 2.8 4.3 5.3 2.9 8.3 4.4 2.8 3.1 5.6

4.5 4.5 5.7 0.5 6.2 3.7 0.9 2.4 3.0 3.5

(a) Find the mean,mode,median,variance, standard deviation, 1, 3,and 90th percentile.

(b)Construct a frequency table with 5 classes.

(c)Using the grouped data formula,find the mean,mode,median,variance,standard

deviation, 1, 3,and 90th percentile for the frequency table constructed in part(b) and compare it to the

results in part(a). (d)Construct a histogram,and comment on the shape of the data.

(a) Find the mean,mode,median,variance, standard deviation, 1, 3,and 90th percentile.

```
position_per<-function(a,p){
  a<-sort(a)
  n=length(a)
  c=(n*p)/100
  if(check_int(c)==TRUE){
    Q=(a[c]+a[c+1])/2
  }else{
    Q=a[round(c,0)]
  }
  return(Q)
}
```

```
data<-c(6.3,2.9,4.5,1.1,1.8,4.0,1.2,3.1,2.0,4.0,7.0,2.8,4.3,2.9,8.3,4.4,2.8,3.1,5.6,4.5,4.5,5
data1<-data
message('Mean of data is ',round(mean(data),2))
```

```
Mean of data is 3.74
```

```
mode(data)%>%
  mutate(Mode=a)%>%
  select(-a)%>%
  knitr::kable(format = 'markdown')
```

Mode
4.5

```
message("Median of data is ",median(data))
```

```
Median of data is 3.6
```

```
message('Varaint of data is ',round(varaint(data),3))
```

```
Varaint of data is 3.502
```

```r
message('Standaed deviation of data is ',round(sqrt(varaint(data)),2))
```

```
Standaed deviation of data is 1.87
```

```r
Q13=Q(data)
Q1=Q13[1]
Q3=Q13[2]
message('Q1 of data is ',Q1)
```

```
Q1 of data is 2.8
```

```r
message('Q3 of data is ',Q3)
```

```
Q3 of data is 4.5
```

```r
message('90th percentile of data is ',position_per(data,90))
```

```
90th percentile of data is 6.25
```

(b)Construct a frequency table with 5 classes.

```r
classify_<-function(a,n){
  number =round((max(a)-min(a))/n,1)
  br<-seq(min(a),max(a)+number,by=number)
  a%>%
    cut(breaks=br,right=FALSE)%>%
    table()%>%
    as.tibble()%>%
    mutate(Freq=c(6,9,8,5,2))%>%
    select(-n)%>%
    return()
}
```

```r
knitr::kable(classify_(data,5),format = 'markdown')
```

.	Freq
[0.5,2.1)	6
[2.1,3.7)	9
[3.7,5.3)	8
[5.3,6.9)	5
[6.9,8.5)	2

(c)Using the grouped data formula,find the mean,mode,median,variance,standard deviation, 1, 3,and 90th percentile for the frequency table constructed in part(b) and compare it to the results in part(a).

```
number =round((max(data)-min(data))/5,1)
b<-seq(min(data),max(data)+number,by=number)
data<-classify_(data,5)
```

```
message('Mean of data is ',round(mean_class(data,b),2))
```

```
Mean of data is 3.86
```

```
message('Varaint of data is ',round(varaint_class(data,b),2))
```

```
Varaint of data is 0.95
```

```
message('Standaed deviation of data is ',round(sqrt(varaint_class(data,b)),2))
```

```
Standaed deviation of data is 0.98
```

Median

$Q_i = l + \frac{(\frac{i \times n}{4} - CF)}{f} \times h$ n is number of all data

```
data%>%
  mutate(upper=b[2:length(b)])->x
# find comulative
cf<-c(x$Freq[1])
  for (i in 1:length(x$Freq)-1) {
    r<-cf[i]+x$Freq[i+1]
    cf<-c(cf,r)
  }
x%>%
  mutate(CF=cf)->x
knitr::kable(x,format = 'markdown')
```

.	Freq	upper	CF
[0.5,2.1)	6	2.1	6
[2.1,3.7)	9	3.7	15
[3.7,5.3)	8	5.3	23
[5.3,6.9)	5	6.9	28
[6.9,8.5)	2	8.5	30

```
# find all data
x$Freq%>%
    sum()/2->t
# find medium
# t=15 close to 20
Cf=6 # Commulative of medium
l=2.1 # lower of Medium
f=9 # frequency of medium
median=l+((t-Cf)/f)*number
message("Median of data is ", round(median,2))
```

```
Median of data is 3.7
```

Q1 and Q3

```
x$Freq%>%
    sum()/4->t# find position Q1
# t=7.5 in [2.1,3.7)
```

```r
Cf=0 # Commulative of Q1
l=0.5 # lower of Q1
f=6 # frequency of Q1
Q1=l+((t-Cf)/f)*number
message("Q1 of data is ", round(Q1,2))
```

```
Q1 of data is 2.5
```

```r
x$Freq%>%
    sum()*3/4->t # find position
# t=22.5 in [3.7,5.3)
Cf=15 # Commulative of Q3
l=3.7 # lower of Q3
f=8 # frequency of Q3
Q3=l+((t-Cf)/f)*number
message("Q3 of data is ", round(Q3,2))
```

```
Q3 of data is 5.2
```

#Mode

$Mode = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

```r
# highest frequency is 9 in [2.1,3.7)
l=2.1
f1=9
f2=8# upper frequency column
f0=6# lower frequency column
Mode=l+number*(f1-f0)/(2*f1-f0-f2)
message("Mode of data is ", round(Mode,2))
```

```
Mode of data is 3.3
```

90th percentile

$P_i = l + \frac{(\frac{i \times n}{100} - CF_<)}{f} \times h$

$CF_<$ is cumulative frequency of the class previous to percentile class

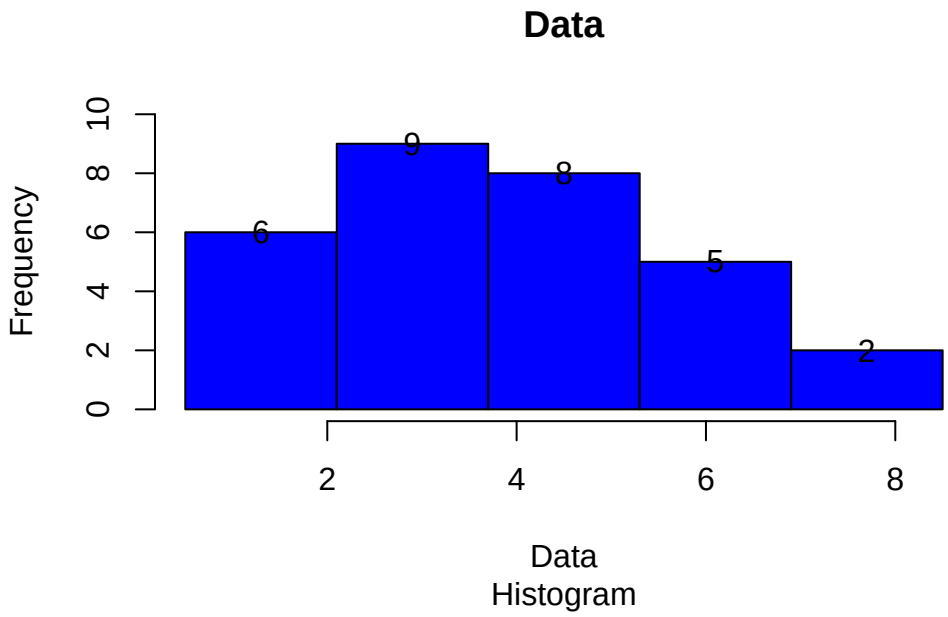
```
x$Freq%>%
    sum()*90/100->t # find position
# t=27 in [5.3,6.9)
Cf=23
l=5.3 # lower of Q3
f=5 # frequency of Q3
Q=l+((t-Cf)/f)*number
message("90th percentile of data is ", round(Q,2))
```

```
90th percentile of data is 6.58
```

By part B and part A , We can said that the gruoped data have better statistic than the point data.It mean each data in part B are closer than each data in part A. so Part B is better than Part A

d. Construct a histogram and comment on the shape of the distribution.

```
hist(data1,
     main='Data',
     sub='Histogram',
     col="blue",
     xlab='Data',
     ylab='Frequency',
     ylim=c(0,10),
     breaks=b,
     right = FALSE)
text(midpoint(b),data$Freq,data$Freq)%>%
knitr::kable(format = 'markdown')
```




```
sk=(mean_class(data,b)-median)*3/sqrt(varaint_class(data,b))
message('Skewness = ',round(sk,2))
```

```
Skewness = 0.49
```

By the Skewness value we can said that graph is not symmetric.Therefor it is not approximate normal distribution. It is positively Skewness.

29. In recent years,due to low interestrates , many home owners refinanced their home mort gages.Linda Lahey is a mortgage of ficer at Down River Federal Saving sand Loan.

Below is the amount refinanced for 20 loans she processed last week. The data are reported in thousands of dollar sand arranged from smallest to largest.

59.2 59.5 61.6 65.5 66.6 72.9 74.8 77.3 79.2 83.7+

85.6 85.8 86.6 87.0 87.1 90.2 93.3 98.6 100.2 100.7

a. Find the median,first quartile ,and third quartile.

b. Find the 26 th and 83 rd percentiles.

c.Draw a box plot of the data and comment on the shape of the distribution.

a. Find the median,first quartile ,and third quartile.

```
Q<-function(a){
    a<-sort(a)
  n=length(a)
  c1=(n*25)/100
  c3=(n*75)/100
  if(check_int(c1)==TRUE){
    Q1=(a[c1]+a[c1+1])/2
  }else{
    Q1=a[round(c1,0)]
  }
  if(check_int(c3)==TRUE){
    Q3=(a[c3]+a[c3+1])/2
  }else{
    Q3=a[round(c3,0)]
  }
  return(c(Q1,Q3))
}
```

```
data<-c(59.2,59.5,61.6,65.5,66.6,72.9,
        74.8,77.3,79.2,83.7,85.6,85.8,
        86.6,87.0,87.1,90.2,93.3,98.6,
        100.2,100.7)
message("Median of data is ",median(data))
```

```
Median of data is 84.65
```

```
Q13=Q(data)
Q1=Q13[1]
Q3=Q13[2]
message('Q1 of data is ',Q1)
```

```
Q1 of data is 69.75
```

```
message('Q3 of data is ',Q3)
```

```
Q3 of data is 88.65
```

b. Find the 26th and 83rd percentiles.

```
message('26th percentile of data is ',position_per(data,26))
```

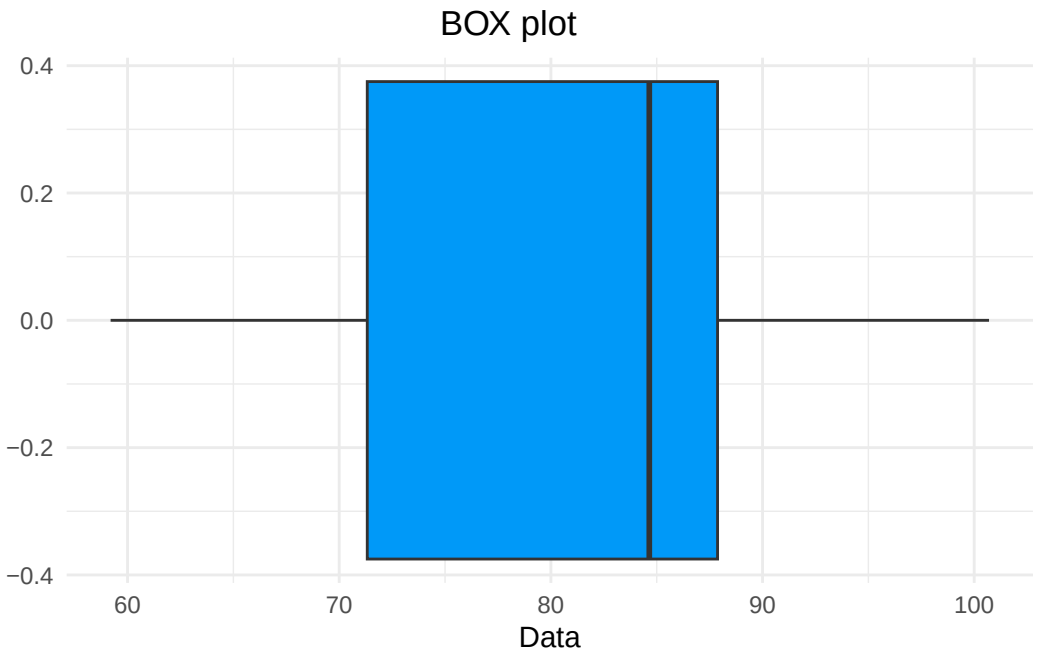
```
26th percentile of data is 66.6
```

```
message('83th percentile of data is ',position_per(data,83))
```

```
83th percentile of data is 93.3
```

c.Draw a boxplot of the data and comment on the shape of the distribution.

```
data%>%
  as.tibble()%>%
  ggplot(aes(x=value))+
  theme_minimal()+
  geom_boxplot(fill = "#0099f8")+
  xlab('Data')+
  ggtitle('BOX plot')+
  theme(plot.title = element_text(hjust=0.45))
```



By graph we can said that graph is not symmetric.Therefor it is not approximate normal distribution. It is negatively skewed.

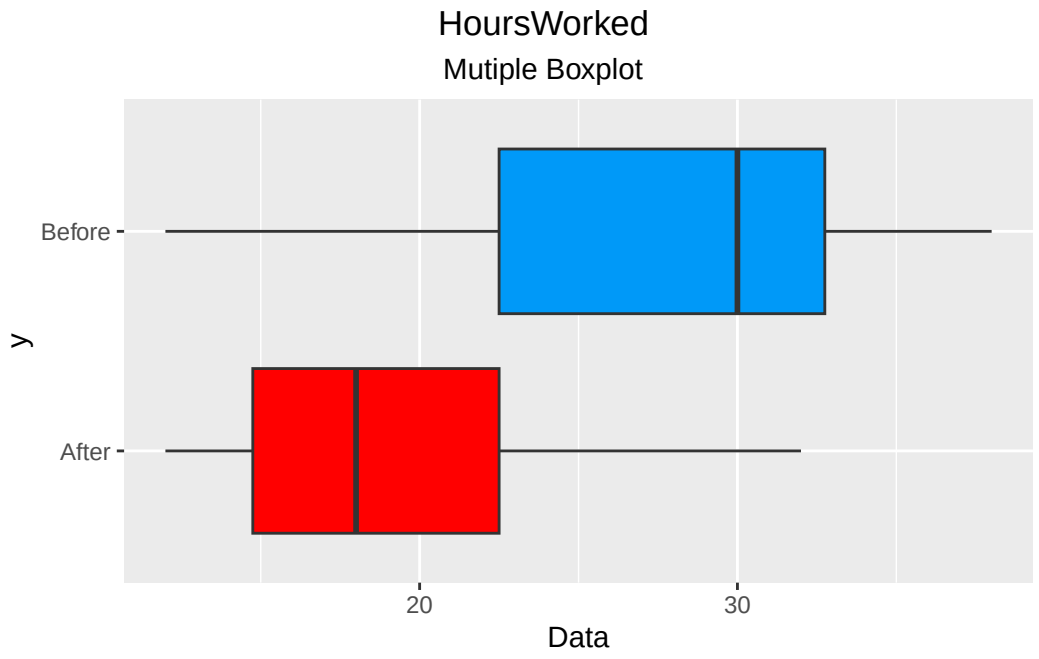
30.Hours Worked The data shown herere present the number of hours that 12 part-time employee statoy store worked during the weeks before and after Christmas .Construct two boxplots and compare the distributions.

```
Before<-c(38,16,18,24,12,30,35,32,31,30,24,35)
After<-c(26,15,12,18,24,32,14,18,16,18,22,12)
data<-tibble(Before,After)
knitr::kable(data,format = 'markdown')
```

Before	After
38	26
16	15
18	12
24	18
12	24

Before	After
30	32
35	14
32	18
31	16
30	18
24	22
35	12

```
data%>%
  ggplot()+
  geom_boxplot(aes(x=Before,y='Before'),fill = "#0099f8")+
  geom_boxplot(aes(x=After,y='After'),fill = "red")+
  xlab('Data')+
  ggtitle('HoursWorked',subtitle = "Mutiple Boxplot")+
  theme(plot.title = element_text(hjust=0.45),plot.subtitle = element_text(hjust=0.45))
```



By the graph After is more approximate normal distribution than Before. After is positively skewed . Before is negatively skewed.

31.Many times in statistics it is necessary to see if a set of data values is approximately normally distributed.There are special techniques that can beused.

One technique is to draw a histogram for the data and see if it is approximately bell-shaped. (Note: It does not have to be exactly symmetric to bebell-shaped.)The number sofbranches of the 50top libraries are shown.

67 84 80 77 97 59 62 37 33 42

36 54 18 12 19 33 49 24 25 22

24 29 9 21 21 24 31 17 15 21

13 19 19 22 22 30 41 22 18 20

26 33 14 14 16 22 26 10 16 24

1. Construct a frequency distribution for the data.
2. Construct a histogram for the data.

3.Describe the shape of the histogram.

4. Based on your answer to question 3,do you feel that the distribution is approximately normal?
5. Find the mean and standard deviation for the data.
6. What percent of the data values fall within 1 standard deviation of the mean?
7. What percent of the data values fall within 2 standard deviations of the mean?
8. What percent of the data values fall within 3 standard deviations of the mean?
9. Does your answer help support the conclusion you reached in question 4? Explain.

1. Construct a frequency distribution for the data.

```r
data<-c(67, 84, 80, 77, 97, 59, 62, 37, 33, 42, 36, 54, 18, 12,
        19, 33, 49, 24, 25, 22, 24, 29, 9, 21, 21, 24, 31, 17,
        15, 21, 13, 19, 19, 22, 22, 30, 41, 22, 18, 20, 26, 33,
        14, 14, 16, 22, 26, 10, 16, 24)
data%>%
  as.tibble()%>%
  table()%>%
  as.tibble()%>%
  mutate(Freq=n)%>%
  select(-n)->df
  knitr::kable(df,format = 'markdown')
```

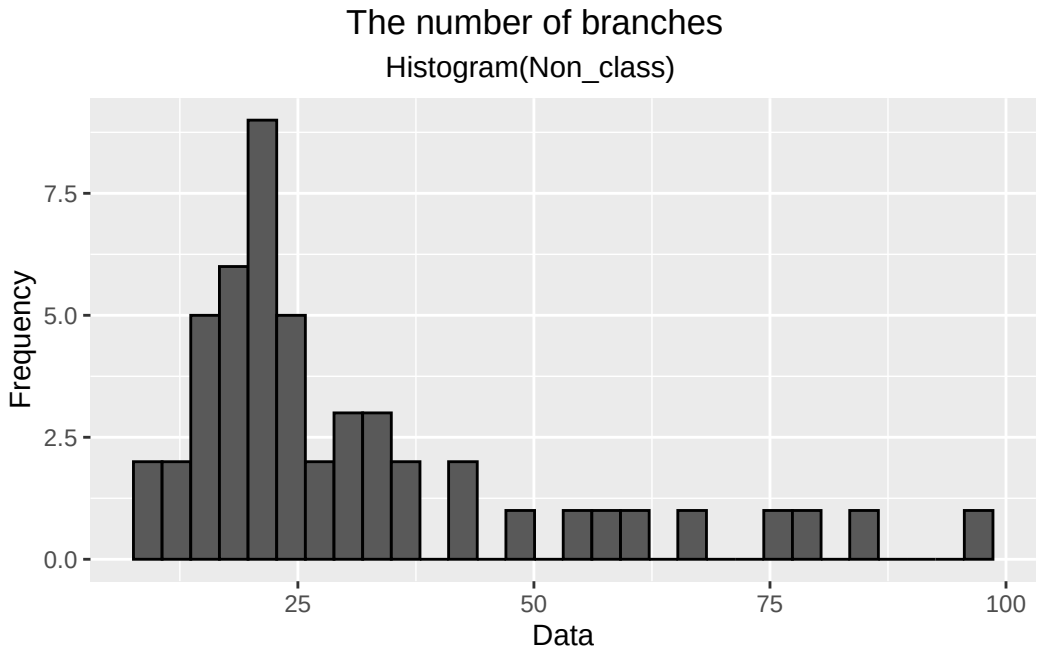
value	Freq
9	1
10	1
12	1
13	1
14	2
15	1
16	2
17	1
18	2
19	3
20	1
21	3
22	5
24	4
25	1
26	2
29	1
30	1
31	1
33	3
36	1
37	1
41	1
42	1
49	1
54	1
59	1
62	1

value	Freq
67	1
77	1
80	1
84	1
97	1

2. Construct a histogram for the data.

Non classify

```
data%>%
  as.tibble()%>%
  ggplot(aes(x=value))+
  geom_histogram(color='black')+
  labs(x='Data',y='Frequency')+
  ggtitle('The number of branches',subtitle = "Histogram(Non_class)")+
  theme(plot.title = element_text(hjust=0.45),plot.subtitle = element_text(hjust=0.45))
```

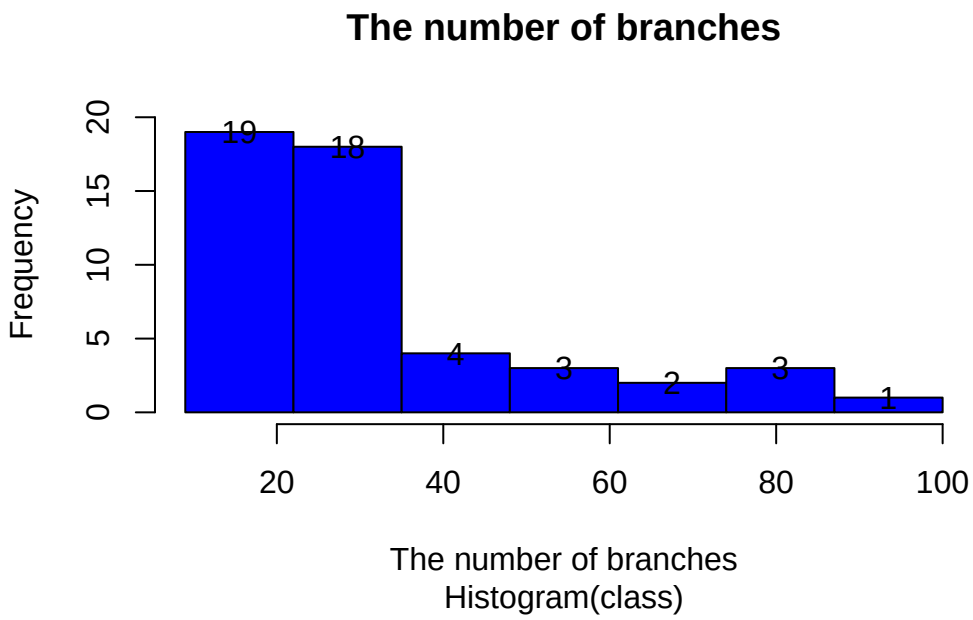


We should classify with $\sqrt{n}$ class

```
classify(data,round(sqrt(length(data)),0))->data1
number =round((max(data)-min(data))/round(sqrt(length(data)),0),0)
b<-seq(min(data),max(data)+number,by=number)
data1
```

```
# A tibble: 7 x 2
  Classes   Freq
  <chr>    <int>
1 [9,22)      19
2 [22,35)     18
3 [35,48)      4
4 [48,61)      3
5 [61,74)      2
6 [74,87)      3
7 [87,100)     1
```

```
hist(data,
     main='The number of branches',
     sub='Histogram(class)',
     col="blue",
     xlab='The number of branches',
     ylab='Frequency',
     ylim=c(0,20),
     breaks=b,
     right = FALSE)
text(midpoint(b),data1$Freq,data1$Freq)%>%
knitr::kable(format = 'markdown')
```



3.Describe the shape of the histogram.

By histogram we can said that the data is approximate positively skewed or Bimodal.

4. Based on your answer to question 3,do you feel that the distribution is approximately normal?

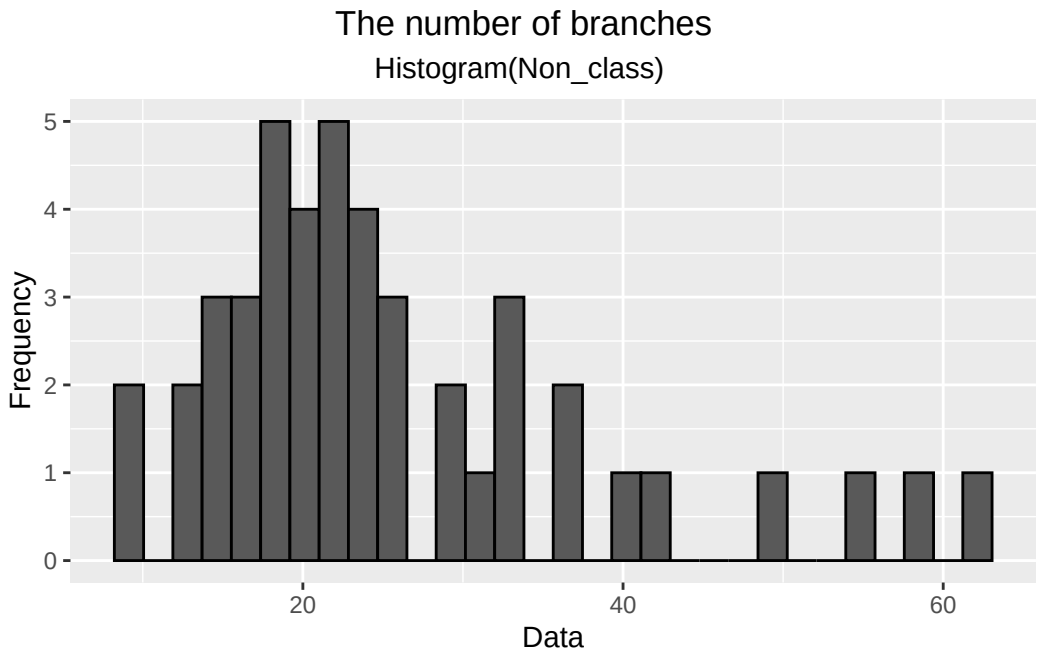
```
find_outier(data)
```

```
Outier in data are:

[1] 67
[1] 77
[1] 80
[1] 84
[1] 97
```

```
remove<-function(a,x){
  return(a[a!=x])
}
```

```
data2=data
for (i in c(67,77,80,84,97)) {
  data2<-remove(data2,i)
}
data2%>%
  as.tibble()%>%
  ggplot(aes(x=value))+
  geom_histogram(color='black')+
  labs(x='Data',y='Frequency')+
  ggtitle('The number of branches',subtitle = "Histogram(Non_class)")+
  theme(plot.title = element_text(hjust=0.45),plot.subtitle = element_text(hjust=0.45))
```

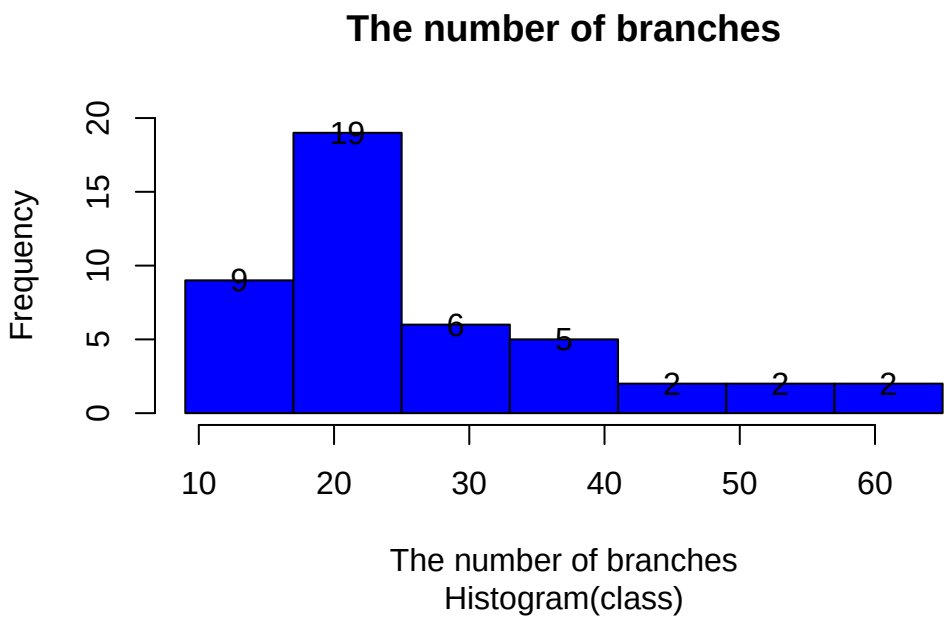



```
classify(data2,round(sqrt(length(data2)),0))->data3
number =round((max(data2)-min(data2))/round(sqrt(length(data2)),0),0)
b<-seq(min(data2),max(data2)+number,by=number)
data3
```

```
# A tibble: 7 x 2
```

```
  Classes  Freq
  <chr>   <int>
1 [9,17)      9
2 [17,25)    19
3 [25,33)     6
4 [33,41)     5
5 [41,49)     2
6 [49,57)     2
7 [57,65)     2
```

```
hist(data2,
     main='The number of branches',
     sub='Histogram(class)',
     col="blue",
     xlab='The number of branches',
     ylab='Frequency',
     ylim=c(0,20),
     breaks=b,
     right = FALSE)
text(midpoint(b),data3$Freq,data3$Freq)%>%
knitr::kable(format = 'markdown')
```



When I cut outlier I see the shape of graph is more approximate bell_shape. But I fell it still

is not bell shape.

5. Find the mean and standard deviation for the data..

```
mean(data)
```

```
[1] 31.38
```

```
message('Mean of data is ',round(mean(data),2))
```

```
Mean of data is 31.38
```

```
message('Varaint of data is ',round(varaint(data),2))
```

```
Varaint of data is 425.22
```

```
message('Standaed deviation of data is ',round(sqrt(varaint(data)),2))
```

```
Standaed deviation of data is 20.62
```

```
probability<-function(a,n,m){
  d=0
  for (i in 1:length(a)) {
    if(a[i]<n && a[i]>m){
      d=d+1
    }
  }
  return(d/length(a))
}
```

6. What percent of the data values fall within 1 standard deviation of the mean?

Let x is data

$P(|X - mean| < 1SD) = P(mean - SD < X < mean + SD)$

```
P=probability(data,mean(data)+sqrt(varaint(data)),mean(data)-sqrt(varaint(data)))
message('The percent of the data values fall within 1 standard deviations of the mean is ',r
```

```
The percent of the data values fall within 1 standard deviations of the mean is 80%
```

7. What percent of the data values fall within 2 standard deviations of the mean?

$P(|X - mean| < 2SD) = P(mean - 2SD < X < mean + 2SD)$

```
P=probability(data,mean(data)+2*sqrt(varaint(data)),mean(data)-2*sqrt(varaint(data)))
message('The percent of the data values fall within 2 standard deviations of the mean is ',r
```

```
The percent of the data values fall within 2 standard deviations of the mean is 92%
```

8. What percent of the data values fall within 3 standard deviations of the mean?

$P(|X - mean| < 3SD) = P(mean - 3SD < X < mean + 3SD)$

```
P=probability(data,mean(data)+3*sqrt(varaint(data)),mean(data)-3*sqrt(varaint(data)))
message('The percent of the data values fall within 3 standard deviations of the mean is ',r
```

```
The percent of the data values fall within 3 standard deviations of the mean is 98%
```

9. Does your answer help support the conclusion you reached in question 4? Explain.

Yes! it does. Base on question 2 and 3 the data is not bell shape.