CORELATION ANALYSIS:

```
* Corelation is a statistical measuare that extends to which two variable are linearly in relation
* It refers to the process of establishing the relationship between two variables

Types of Corelations:

                  Values = -1 , 0 , -1
    * -1 indicates strong negative corelation which means x increases, y decreases
    * 0  no association between two variables
    * +1 indicates strong positive corelation which means x increases, y increases



Methods of Corelations:
      * Pearson (measures the linear dependancy of two variables)
      * Spearman (compute the corelation between the rank of x and y variables)
      * Kendal (measures the correspondence between the x & y varibales)
```

Correlation Coefficient Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Pearson Method:

In [1]:

```
x = c(4,8,12,16)
y = c(5,10,15,20)
```

```
if method is not specified by default pearson method is used
```

In [2]:

```
cor.test(x,y, method="pearson")
```

```
	Pearson's product-moment correlation

data:  x and y
t = Inf, df = 2, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 1 1
sample estimates:
cor 
  1 
```

In [6]:

```
a = cor.test(x,y)

if (a$estimate ==1)
    {
    print("Strong Positive")
} else if(a$estimate==-1){
    print("Strong Negative")
} else{
    print("No Relation")
}
```

```
[1] "Strong Positive"
```

In [7]:

```
a = c(4,8,12,26)
b = c(5,10,15,30)
```

In [8]:

```
v = cor.test(a,b)

if (v$estimate ==1)
    {
    print("Strong Positive")
} else if(v$estimate==-1){
    print("Strong Negative")
} else{
    print("No Relation")
}
```

```
[1] "No Relation"
```

In [12]:

```
r = c(4,8,12,0)
t = c(5,0,5,30)
```

In [16]:

```
o = cor.test(r,t)

if (o$estimate ==1)
    {
    print("Strong Positive")
} else if(o$estimate<1){
    print("Strong Negative")
} else{
    print("No Relation")
}
```

```
[1] "Strong Negative"
```

In [ ]:

Statistical Summary:

```
            Exploring Data
                 -> Summarization
                 -> Visualisation
                 -> Normalization



       Data Summarization:
             * Mean   -> Sum of X / number of values
             * Median -> Center number
             * Mode   -> The number which is repeated multiple times



        Replacing missing values:
                 If there is a null value in the dataset then go for data summarisation to fill the null value.
```

```
Measure the Central Dispertion:
        -> Standard Deviation
        -> Variance
        -> Range
```

```
Measure of Central Dispertion:
        The measure of central dispertion express the scattering of the datapoints in terms of distance such as
        range or in terms of deviation from the central value such as variance and standard deviation.
```

Measure of Central Tendency:

In [41]:

```r
a =c(5,52,53,64,65,86,86,92)
mean = mean(a)
mode <- function(a) {
   uniqv <- unique(a)
   uniqv[which.max(tabulate(match(a, uniqv)))]
}
median = median(a)
result <- mode(a)


print(paste("Mean: ",mean))
print(paste("Mode: ",result))
print(paste("Median: ",median))
```

```
[1] "Mean:  62.875"
[1] "Mode:  86"
[1] "Median:  64.5"
```

In [43]:

```r
# creating a data frame:
df = data.frame(mean, result, median)
df
```

mean	result	median
62.875	86	64.5

Measure of Dispersion or Variability:

In [35]:

```r
A =c(5,52,53,64,65,86,86,92)
a = sort(A)
# Range:
range= max(a) - min(a)
# Variance:
variance= var(a)
# Standard Deviation:
std = sd(a)
# Dispersion:
dispersion = mean(a) - sd(a)

print(paste("Range: ", range))
print(paste("Variance: ", variance))
print(paste("Standard Deviation: ", std))
print(paste("Dispersion: ", dispersion))
```

```
[1] "Range:  87"
[1] "Variance:  784.125"
[1] "Standard Deviation:  28.0022320538917"
[1] "Dispersion:  34.8727679461083"
```

In [24]:

```
# quartile:
b= quantile(a)
b
```

0%
5
25%
52.75
50%
64.5
75%
86
100%
92

In [37]:

```
df1 = data.frame(range, variance, dispersion, Q1=b[2], Q2=b[3], Q3=b[4], row.names="Data Set 1")
df1
```

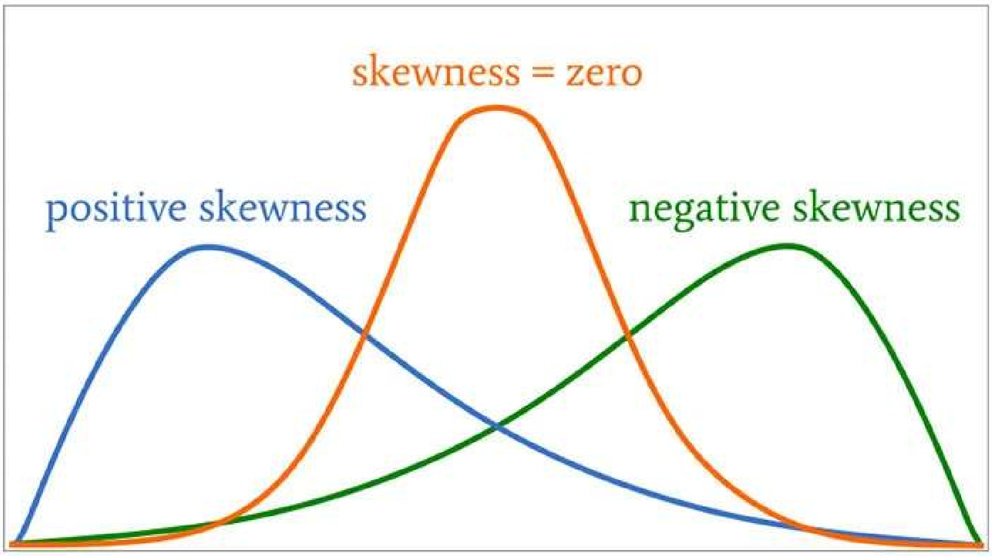
	range	variance	dispersion	Q1	Q2	Q3
Data Set 1	87	784.125	34.87277	52.75	64.5	86

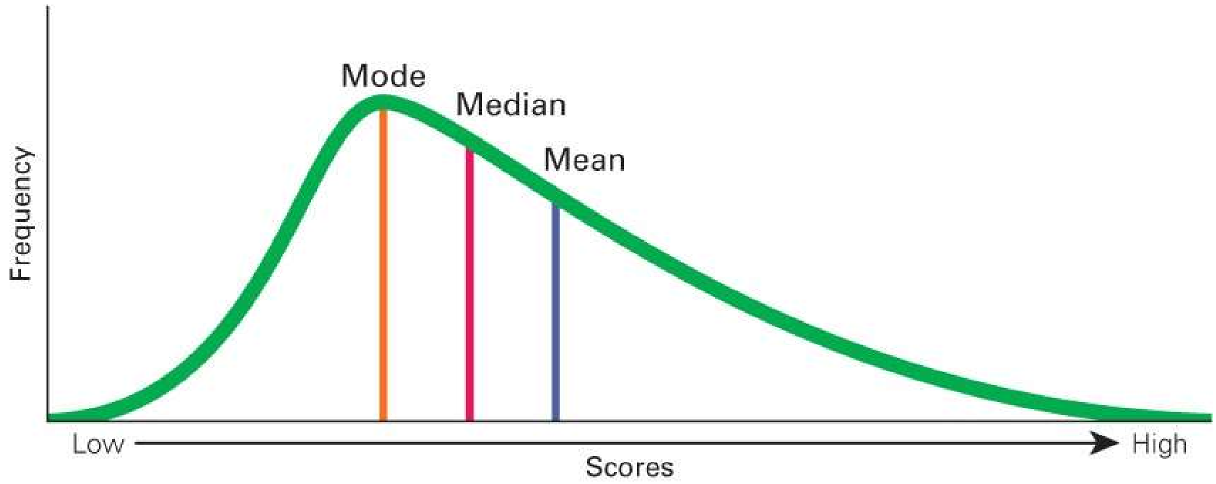
Symmetrical Distribution:

```
* A frequency distribution is said to be symmetrical if the frequencies are equally distributed on both
  the sides of central values.
* A symmetrical distibution may be either bell-shaped or U-shaped.
* Mean = Median = Mode
```

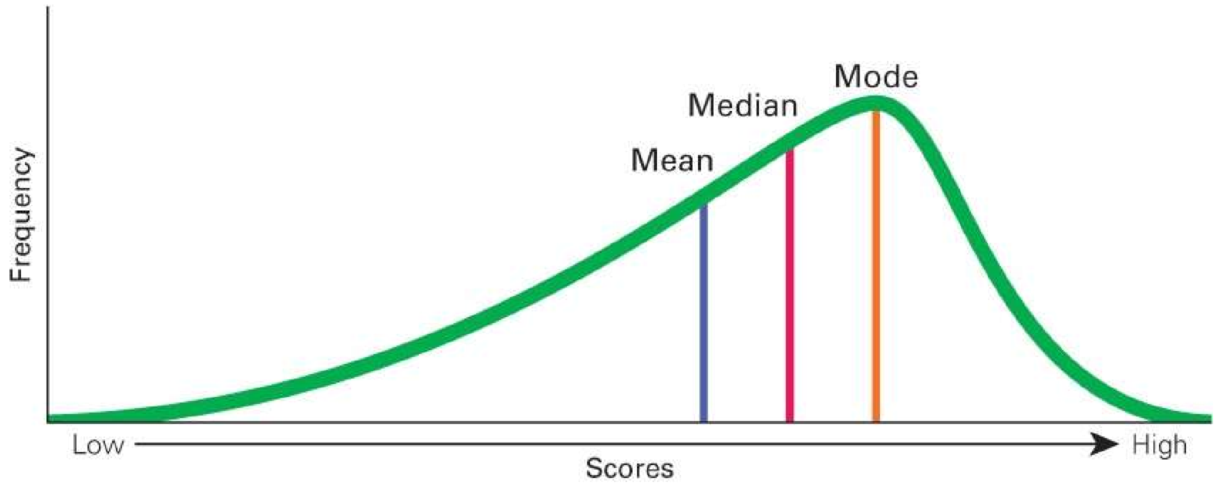
Skewed Distribution:

```
Types of Skewness:
        * Positive Skewness (Mean exceeds Mode and Median)
        * Negative Skewness (Mode exceeds Mean and Median)
        * No Skewness       (Mean Median Mode are equal)
```





(a) Right-skewed distribution



(b) Left-skewed distribution

Pearson's Coefficient of Skewness:

```
* This method is most frequently used for measuring skewness.


Formula:

                                  Mean  - Mode
  Pearsons Coefficient   =    --------------------
                               Standard Deviation


Formula: (When Mode cannot be determined we can use this formula)
                                 Mean - (3*Median - 2*Mean)
   Pearsons Coefficient    =    --------------------------
                                    Standard Deviation
```

In [46]:

```
install.packages("e1071")
```

```
  There is a binary version available but the source version is later:
      binary source needs_compilation
e1071  1.7-6 1.7-12              TRUE

  Binaries will be installed

Warning message:
"package 'e1071' is in use and will not be installed"
```

In [47]:

```
library("e1071")
```

In [48]:

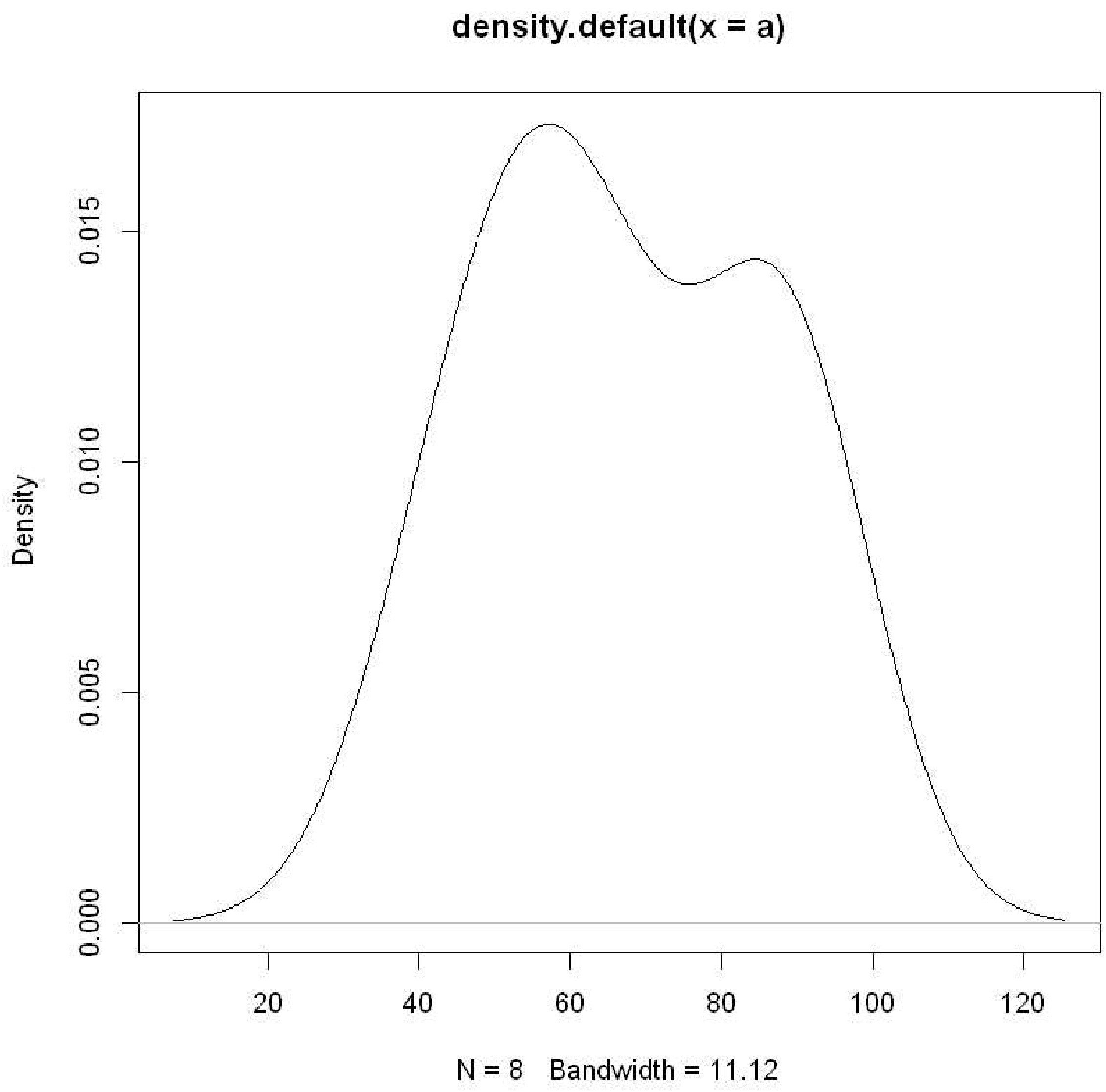
```
a = c(41, 52, 53, 64, 65, 86, 86, 92)
skewness(a)
```

0.0542062901835262

In [49]:

```
skewness(a)
plot(density(a))
```

0.0542062901835262



In [54]:

```
data()
```

In [60]:

```
data(mtcars)
```

In [61]:

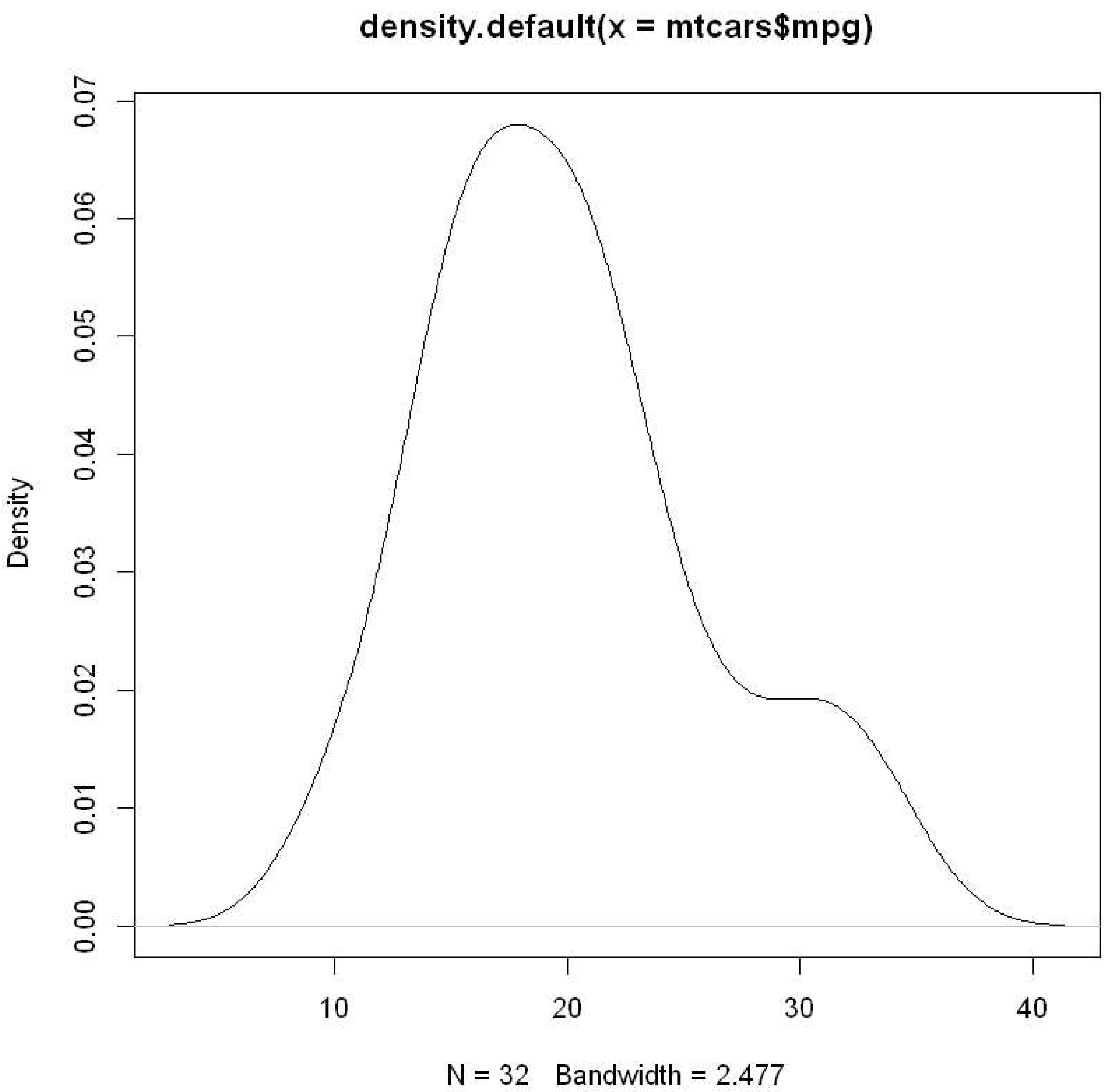
```
mtcars$mpg
```

21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26 30.4 15.8 19.7 15 21.4

In [65]:

```
skewness(mtcars$mpg)
plot(density(mtcars$mpg))
```

0.610655017573288



	Measure of Tendency				Measure of Dispersion or Variability						
	Mean	Median	Mode	Range	Standard Deviation	Variable	Q1	Q2	Q3	Dispression
Dataset 1	67.4	64.5	86	51	18.7	350.8	52.5	64.5	86	48.7
Dataset 1	62.8	64.5	86	87	28	784.1	52.5	64.5	86	34.8

In [ ]: