Mid Term Project
Introduction to Data Science
Zerin Tasnim
ID: 20-43032-1
Section: D

Import the data set as csv and print the data set:

```
1  mydata<-read.csv("E:/10th semester/Data Science/Project/Dataset_midterm.csv",header=TRUE,sep=",")
2  mydata
```

Here is the code of import the dataset as csv file. In this code also has the location of the csv dataset file.

Output:

```
> mydata<-read.csv("E:/10th semester/Data Science/Project/Dataset_midterm.csv",header=TRUE,sep=",")
> mydata
   id Age weight.kg. Delivery_number
1   1  22       57.7               1
2   2  26       63.0               2
3   3  26       62.0               2
4   4  28       65.0               1
5   5  22       58.0               2
6   6  26       63.0               1
7   7  27       64.0               2
8   8  32       70.0               3
9   9  28       63.5               2
10 10  27       64.5               1
11 11  36       75.0               1
12 12  33       70.0               1
13 13  23       58.0               1
14 14  20       55.0               1
15 15  29       65.0               1
```

16	16	25	61.5	1
17	17	25	61.5	1
18	18	20	55.5	1
19	19	37	76.0	3
20	20	24	56.6	1
21	21	26	62.0	1
22	22	33	75.0	2
23	23	25	62.0	1
24	24	27	65.0	NA
25	25	20	55.0	1
26	26	18	49.0	NA
27	27	18	50.0	1
28	28	30	68.0	1
29	29	32	73.0	1
30	30	26	62.5	2
31	31	25	58.0	1
32	32	40	82.0	1
33	33	32	68.0	2
34	34	27	63.0	2
35	35	26	59.0	2
36	36	28	66.0	3
37	37	33	75.0	1
38	38	31	69.0	2
39	39	31	63.0	1
40	40	26	59.0	1
41	41	27	63.0	1
42	42	19	51.0	1
43	43	36	73.0	1
44	44	22	57.0	1
45	45	36	72.5	4
46	46	28	62.5	3
47	47	26	NA	1
48	48	32	67.5	2
49	49	26	62.5	2
50	50	NA	NA	2
51	51	33	68.5	3
52	52	21	53.0	2
53	53	30	68.0	3
54	54	35	74.0	1
55	55	29	63.5	2
56	56	25	59.0	2
57	57	32	67.5	3
58	58	95	110.0	1
59	59	26	61.5	1
60	60	30	67.5	2
61	61	22	58.5	1
62	62	NA	NA	1
63	63	32	67.0	2
64	64	32	67.0	2
65	65	31	66.0	1
66	66	35	72.0	2
67	67	28	62.5	3
68	68	29	64.5	2
69	69	25	62.0	1
70	70	27	61.0	2
71	71	90	105.0	1

```
72 72  29       65.0              1
73 73  28       64.0              2
74 74  32       69.0              3
75 75  38       75.0              3
76 76  27       62.5              2
77 77  33       66.0              4
78 78  NA       63.0              2
79 79  25       58.0              1
80 80  24       57.0              2
   Delivery_time  Blood Heart Caesarian
1              0   high     0         0
2              0 normal     0         1
3              1 normal     0         0
4              0   high     0         0
5              0 normal     0         1
6              1    low     0         0
7              0 normal     0         0
8              0 normal     0         1
9              0            0         0
```

```
10             1 normal     0         1
11             0 normal     0         0
12             1    low     0         1
13             1 normal     0         0
14             0 normal     1         0
15            NA            1         1
16             2    low     0         0
17             0 normal     0         0
18             2   high     0         1
19             0 normal     1         1
20             2    low     1         1
21             1 normal     0         0
22             0    low     1         1
23             1   high     0         0
24            NA    low     1         1
25             0   high     1         1
26             0 normal     0         0
27            NA   high     1         1
28             0 normal     0         0
```

```
29             0   high     1         1
30             1 normal     1         0
31             0    low     0         0
32             0 normal     1         1
33             0   high     1         1
34             0 normal     1         1
35             2 normal     0         1
36             0   high     0         1
37             1 normal     0         0
38             2 normal     0         0
39             0 normal     0         0
40             2    low     1         1
41             0   high     1         1
42             0 normal     0         1
43             1   high     0         1
44             0 normal     0         1
45             0   high     1         1
46             0 normal     1         1
47             0 normal     0         0
```

```
48              0   high      1         1
49              2 normal      0         0
50              0    low      1         1
51              2 normal      1         0
52              1    low      1         1
53              2   high      0         0
54              1    low      0         0
55              0 normal      1         1
56              0 normal      0         0
57              1    low      1         1
58              0    low      0         1
59              0   high      0         1
60              1   high      1        NA
61              2   high      0         0
62              0 normal      0         1
63              0    low      0         1
64              0 normal      1         1
65              2   high      1         0
66              0 normal      0         1
67              0 normal      0         1
68              0 normal      1         0
69              0    low      0         1
70              2    low      0         0
71              0    low      0         1
72              2             1         1
73              0 normal      0         0
74              0 normal      1         0
75              2   high      1         1
76              1 normal      0         0
77              0 normal      0        NA
78              1   high      0         1
79              2    low      0         1
80              2 normal      0         0
> |
```

It is the output of the dataset which is import in Rstudio.

To see the column name of the data set:

```
3  names(mydata)
```

This code is to see the column name of the dataset. Here with this code can see the attributes names.

Output:

```
> names(mydata)
[1] "id"              "Age"
[3] "weight.kg."      "Delivery_number"
[5] "Delivery_time"   "Blood"
[7] "Heart"           "Caesarian"
> |
```

The output of the name() function where we can see the attributes of the dataset.

Summary of the structure of data set:

```
4  str(mydata)
```

Here is the code to see the summary of the structure of dataset.

Output:

```
> str(mydata)
'data.frame':   80 obs. of  8 variables:
 $ id             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Age            : int  22 26 26 28 22 26 27 32 28 27 ...
 $ weight.kg.     : num  57.7 63 62 65 58 63 64 70 63.5 64.5 ...
 $ Delivery_number: int  1 2 2 1 2 1 2 3 2 1 ...
 $ Delivery_time  : int  0 0 1 0 0 1 0 0 0 1 ...
 $ Blood          : chr  "high" "normal" "normal" "high" ...
 $ Heart          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Caesarian      : int  0 1 0 0 1 0 0 1 0 1 ...
>
```

In the output we can see the summary of the structure of the dataset. The dataset has 80 observations (rows) and 8 variables (columns). Also here is showed the data types of the dataset.

Descriptive Statistics Using summary() Function:

```
5 summary(mydata)
```

Here is the code to see the descriptive Statistics. To see descriptive statistic we use the summary() function.

Output:

```
> summary(mydata)
       id              Age
 Min.   : 1.00   Min.   :18.00
 1st Qu.:20.75   1st Qu.:25.00
 Median :40.50   Median :28.00
 Mean   :40.50   Mean   :29.68
 3rd Qu.:60.25   3rd Qu.:32.00
 Max.   :80.00   Max.   :95.00
                 NA's   :3
   weight.kg.     Delivery_number
 Min.   : 49.00   Min.   :1.000
 1st Qu.: 61.00   1st Qu.:1.000
 Median : 63.50   Median :1.500
 Mean   : 65.13   Mean   :1.679
 3rd Qu.: 68.00   3rd Qu.:2.000
 Max.   :110.00   Max.   :4.000
 NA's   :3        NA's   :2
 Delivery_time       Blood

 Min.   :0.0000   Length:80
 1st Qu.:0.0000   Class :character
 Median :0.0000   Mode  :character
 Mean   :0.6234
 3rd Qu.:1.0000
 Max.   :2.0000
 NA's   :3
     Heart          Caesarian
 Min.   :0.000   Min.   :0.0000
 1st Qu.:0.000   1st Qu.:0.0000
 Median :0.000   Median :1.0000
 Mean   :0.375   Mean   :0.5641
 3rd Qu.:1.000   3rd Qu.:1.0000
 Max.   :1.000   Max.   :1.0000
                 NA's   :2
>
```

In the output here min, max, median, and mean are shown.

Counting number of Missing values in each column:

```
8 colSums(is.na(mydata))
```

Here the code for counting number of missing values in each column.

Output:

```
> colSums(is.na(mydata))
             id             Age
              0               3
     weight.kg. Delivery_number
              3               2
  Delivery_time           Blood
              3               0
          Heart       Caesarian
              0               2
>
```

In the output we can see in age and weight has 3 missing values, Delivary_number and Caesarian has 2 missing values and id and heart has no missing value.

Remove missing values from data set:

```
9  mydata_remove<-na.omit(mydata)
```

With this code we can remove the missing or null values from the dataset.

Output:

	id	Age	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
48	48	32	67.5	2	0	high	1	1
49	49	26	62.5	2	2	normal	0	0
51	51	33	68.5	3	2	normal	1	0
52	52	21	53.0	2	1	low	1	1
53	53	30	68.0	3	2	high	0	0
54	54	35	74.0	1	1	low	0	0
55	55	29	63.5	2	0	normal	1	1
56	56	25	59.0	2	0	normal	0	0
57	57	32	67.5	3	1	low	1	1
58	58	95	110.0	1	0	low	0	1
59	59	26	61.5	1	0	high	0	1
61	61	22	58.5	1	2	high	0	0

Here we can see the missing or null values are remove with the full instance.

Finding the Standard deviation of the Attributes:

```
10  s<-mydata_remove$Age
11  sd(s)
12  s<-mydata_remove$weight.kg.
13  sd(s)
14  s<-mydata_remove$Delivery_number
15  sd(s)
16  s<-mydata_remove$Delivery_time
17  sd(s)
18  s<-mydata_remove$Heart
19  sd(s)
20  s<-mydata_remove$Caesarian
21  sd(s)
```

This code is present the standard deviation of the attributes.

Output:

```
> s<-mydata_remove$Age
> sd(s)
[1] 11.76199
> s<-mydata_remove$weight.kg.
> sd(s)
[1] 9.514509
> s<-mydata_remove$Delivery_number
> sd(s)
[1] 0.7749975
> s<-mydata_remove$Delivery_time
> sd(s)
[1] 0.8320694
> s<-mydata_remove$Heart
> sd(s)
[1] 0.4826171
> s<-mydata_remove$Caesarian
> sd(s)
[1] 0.5017567
```

Here is the sd of age is 11.76199, weight is 9.514509, delivary_number is 0.7749975, delivary_time is 0.8320694, heart is 0.4826171 and caesarian is 0.5017567.

Descriptive Statistics Using summary() Function without Missing value:

```
23 summary(mydata_remove)
```

This is the descriptive statistics with summary() function without missing value.

Output:

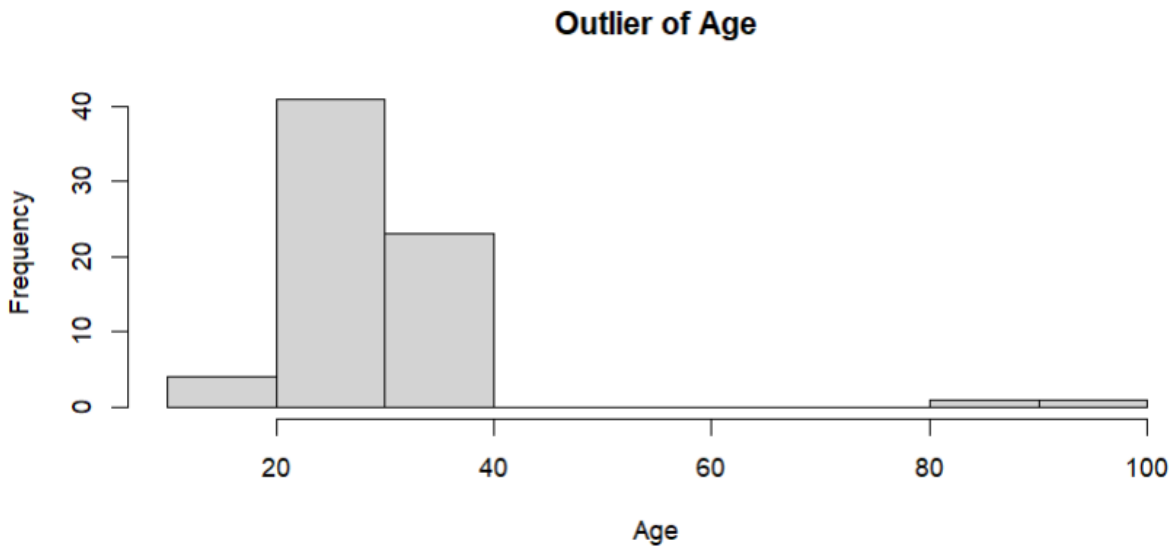
```
> summary(mydata_remove)
       id              Age          weight.kg.     Delivery_number Delivery_time
 Min.   : 1.00   Min.   :19.00   Min.   : 51.00   Min.   :1.000   Min.   :0.0000
 1st Qu.:19.25   1st Qu.:25.00   1st Qu.: 61.12   1st Qu.:1.000   1st Qu.:0.0000
 Median :39.50   Median :28.00   Median : 63.25   Median :1.500   Median :0.0000
 Mean   :39.63   Mean   :30.06   Mean   : 65.56   Mean   :1.671   Mean   :0.6571
 3rd Qu.:58.75   3rd Qu.:32.00   3rd Qu.: 68.38   3rd Qu.:2.000   3rd Qu.:1.0000
 Max.   :80.00   Max.   :95.00   Max.   :110.00   Max.   :4.000   Max.   :2.0000
    Blood               Heart          Caesarian
 Length:70          Min.   :0.0000   Min.   :0.0000
 Class :character   1st Qu.:0.0000   1st Qu.:0.0000
 Mode  :character   Median :0.0000   Median :1.0000
                    Mean   :0.3571   Mean   :0.5429
                    3rd Qu.:1.0000   3rd Qu.:1.0000
                    Max.   :1.0000   Max.   :1.0000
>
```

Without missing value here are the min, max, mean and median. In R language there is no syntax to find mode.

Outliers of Attributes:

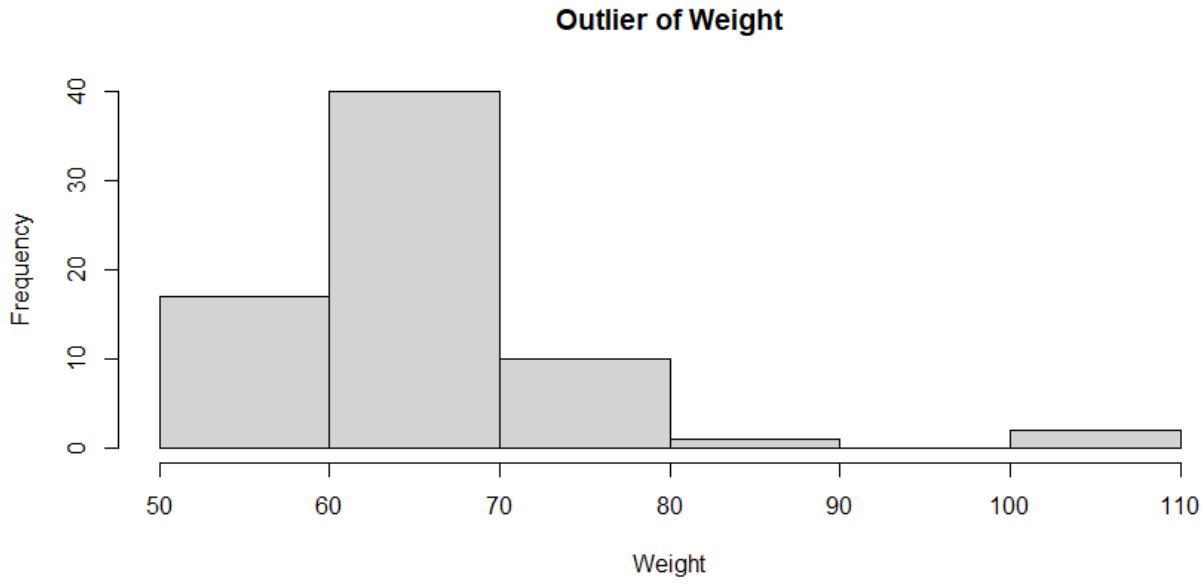
```
49 hist(mydata_remove$Age,xlab = "Age",main = "Outlier of Age", breaks = sqrt(nrow(mydata_remove)))
```

Output:



This histogram is present the outlier of the Age. Here 80-100 is the outlier of the age attribute.

```
51  hist(mydata_remove$weight.kg.,xlab = "Weight",main = "Outlier of Weight", breaks = sqrt(nrow(mydata_remove)))
```



Here 100-110 is the outlier of the weight attribute.

Annotate attributes:

```
37 mydata_remove["Blood"][mydata_remove["Blood"] == "normal"]<-1
38 mydata_remove["Blood"][mydata_remove["Blood"] == "high"]<-2
39 mydata_remove["Blood"][mydata_remove["Blood"] == "low"]<-3
```

Here annotate normal as 1, high as 2, low as 3 from Blood.

Output:

```
> mydata_remove
   id Age weight.kg. Delivery_number Delivery_time Blood Heart Caesarian
1   1  22       57.7               1             0     2     0          0
2   2  26       63.0               2             0     1     0          1
3   3  26       62.0               2             1     1     0          0
4   4  28       65.0               1             0     2     0          0
5   5  22       58.0               2             0     1     0          1
6   6  26       63.0               1             1     3     0          0
7   7  27       64.0               2             0     1     0          0
8   8  32       70.0               3             0     1     0          1
9   9  28       63.5               2             0           0          0
10 10  27       64.5               1             1     1     0          1
11 11  36       75.0               1             0     1     0          0
12 12  33       70.0               1             1     3     0          1
13 13  23       58.0               1             1     1     0          0
14 14  20       55.0               1             0     1     1          0
16 16  25       61.5               1             2     3     0          0
17 17  25       61.5               1             0     1     0          0
18 18  20       55.5               1             2     2     0          1
```

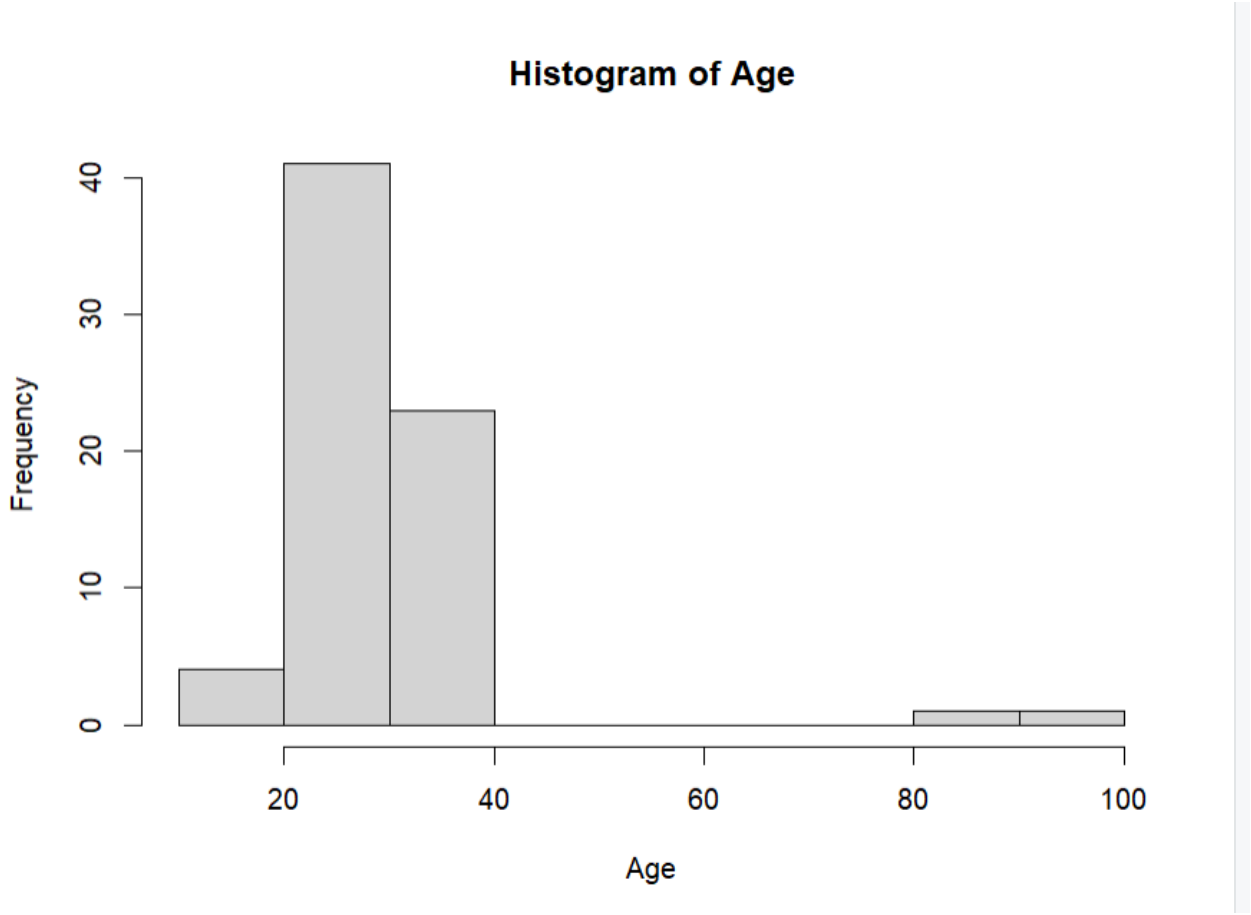
Here the output represents the Blood attribute with numeric type.

Histogram:

Age:

```
27 Age<-mydata_remove$Age
28 hist(Age)
```

Output:

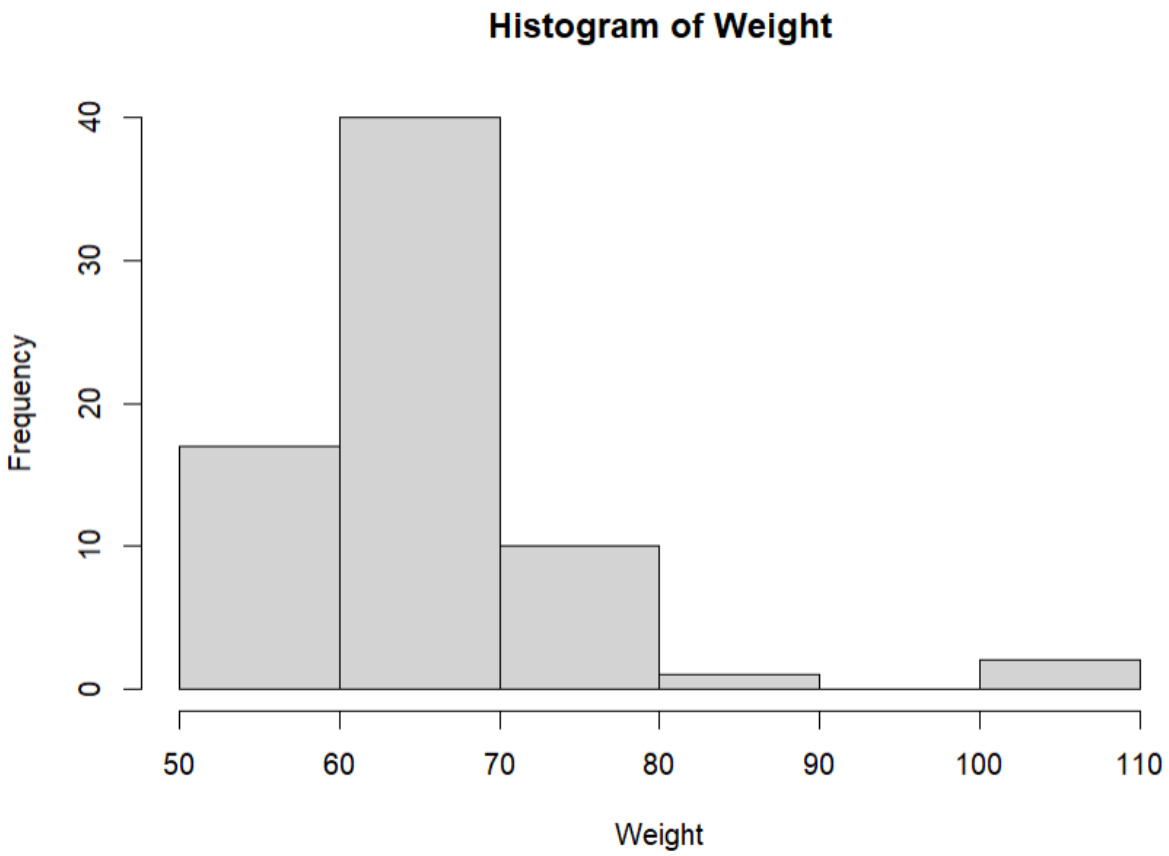


20–30-year-old people are more in the age histogram. Then people of 30-40 years. Then 10-20 then the lowest 80-100.

Weight:

```
29 Weight<-mydata_remove$weight.kg.
30 hist(Weight)
```

Output:



According to the weighted histogram, people who 60-70 kg have the highest frequency. Then 50-60kg, 70-80kg, 100-110 kg and the least are 80-90kg.