

R Project Report

SL	NAME	ID
1	Raufull Islam Rauf	21-45779-3
2	Nazmul Hasan Emon	21-45829-3

Description of Dataset: There was 80 instances in the dataset of pregnant patient where attributes are Patient_id, Age, Gender, weight.kg., Delivery_number, Delivery_time, Blood, Heart, Caesarian. Here, **Caesarian** is the **Target/Output** class. This is a **supervised** dataset. There is some duplicate value, Noisy value, missing value, outliers etc. Some value given into categorical value which should be converted into numerical value or vice versa.

Code:

library(dplyr) - Used for data manipulation.

library(Amelia) - Used for handling missing data.

library(ggplot2) - Used for data visualization

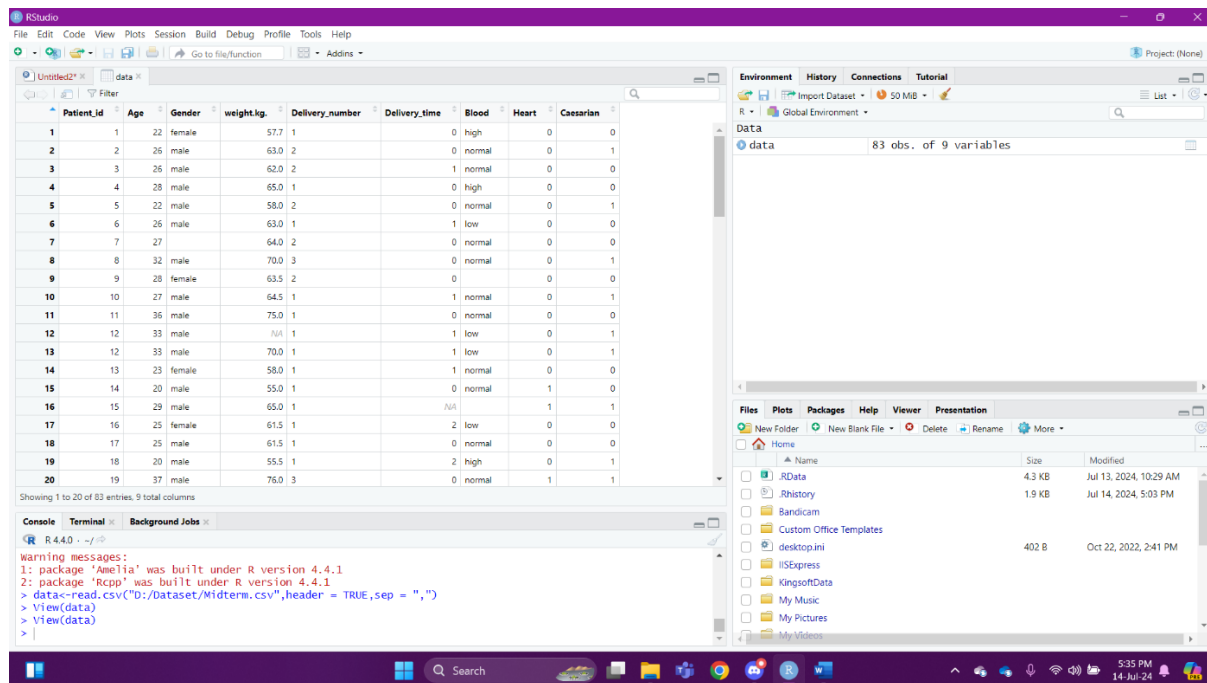
library(caret) - Contains functions to streamline the process of creating predictive models.

library(modeest) - Provides functions to compute statistical modes.

library(smotefamily) - Contains functions for oversampling and undersampling techniques.

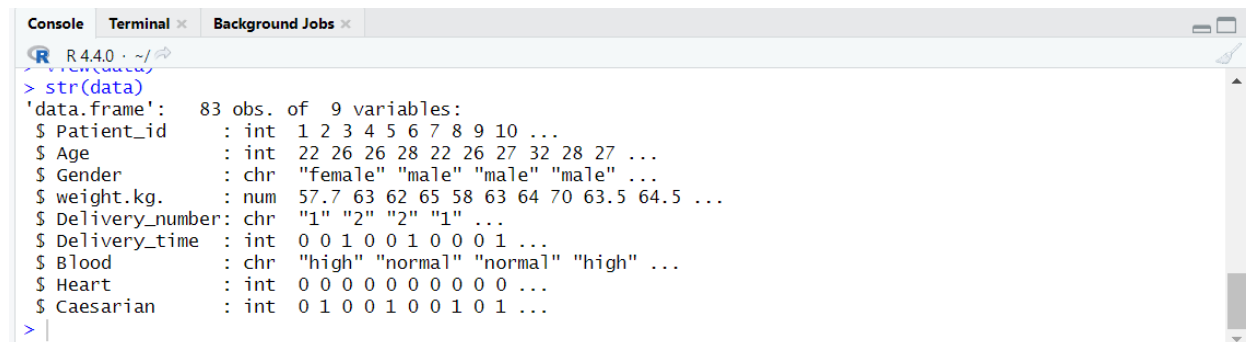
```
data<-read.csv("D:/Dataset/Midterm.csv",header = TRUE,sep = ",")
```

- Reads the CSV file into a data frame.



str(data)

- To see Data Type of variables in Data frame



colSums(is.na(data))

- Checks missing value in each column

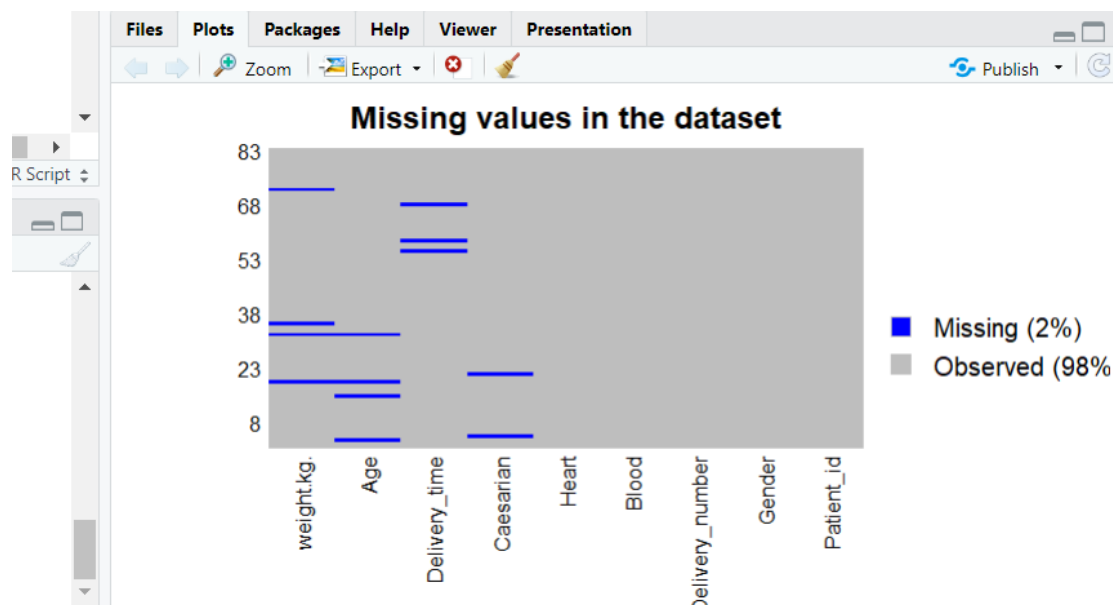
```

R 4.4.0 ~ /
$ weight.kg. : num  57.7 63 62 65 58 63 64 70 63.5 64.5 ...
$ Delivery_number: chr  "1" "2" "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 ...
> colSums(is.na(data))
  Patient_id      Age      Gender weight.kg. Delivery_number Delivery_time
         0         4         0         4         0         3
  Blood      Heart Caesarian
         0         0         2
>

```

```
missmap(data, main = "Missing values in the dataset", col = c("blue", "grey"), legend = TRUE)
```

- Visualization of missing value in a graph.



```
data$Blood <- factor(data$Blood, levels = c('low', 'normal', 'high'), labels = c(0, 1, 2))
```

- Converting Blood pressure into Categorical to Numerical value to find NA.

Patient_id	Age	Gender	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
1	1	22	female	57.7	1	0 2	0	0
2	2	26	male	63.0	2	0 1	0	1
3	3	26	male	62.0	2	1 1	0	0
4	4	28	male	65.0	1	0 2	0	0
5	5	22	male	58.0	2	0 1	0	1
6	6	26	male	63.0	1	1 0	0	0
7	7	27		64.0	2	0 1	0	0
8	8	32	male	70.0	3	0 1	0	1
9	9	28	female	63.5	2	0 NA	0	0
10	10	27	male	64.5	1	1 1	0	1
11	11	36	male	75.0	1	0 1	0	0
12	12	33	male	NA	1	1 0	0	1
13	12	33	male	70.0	1	1 0	0	1
14	13	23	female	58.0	1	1 1	0	0
15	14	20	male	55.0	1	0 1	1	0
16	15	29	male	65.0	1	NA NA	1	1

Showing 1 to 17 of 83 entries, 9 total columns

```
data$Delivery_number <- factor(data$Delivery_number, levels = c('1', '2', '3', '4'), labels = c(1, 2, 3, 4))
```

- Converting No. of Delivery into Categorical to Numerical value to find NA.

Patient_id	Age	Gender	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
1	1	22	female	57.7	1	0 2	0	0
2	2	26	male	63.0	2	0 1	0	1
3	3	26	male	62.0	2	1 1	0	0
4	4	28	male	65.0	1	0 2	0	0
5	5	22	male	58.0	2	0 1	0	1
6	6	26	male	63.0	1	1 0	0	0
7	7	27		64.0	2	0 1	0	0
8	8	32	male	70.0	3	0 1	0	1
9	9	28	female	63.5	2	0 NA	0	0
10	10	27	male	64.5	1	1 1	0	1
11	11	36	male	75.0	1	0 1	0	0
12	12	33	male	NA	1	1 0	0	1
13	12	33	male	70.0	1	1 0	0	1
14	13	23	female	58.0	1	1 1	0	0
15	14	20	male	55.0	1	0 1	1	0
16	15	29	male	65.0	1	NA NA	1	1

Showing 1 to 17 of 83 entries, 9 total columns

```
data$Gender <- 'female'
```

- Handling Noisy Value in Dataset

	Patient_id	Age	Gender	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
1	1	22	female	57.7	1	0	2	0	0
2	2	26	female	63.0	2	0	1	0	1
3	3	26	female	62.0	2	1	1	0	0
4	4	28	female	65.0	1	0	2	0	0
5	5	22	female	58.0	2	0	1	0	1
6	6	26	female	63.0	1	1	0	0	0
7	7	27	female	64.0	2	0	1	0	0
8	8	32	female	70.0	3	0	1	0	1
9	9	28	female	63.5	2	0	NA	0	0
10	10	27	female	64.5	1	1	1	0	1
11	11	36	female	75.0	1	0	1	0	0
12	12	33	female	NA	1	1	0	0	1
13	12	33	female	70.0	1	1	0	0	1
14	13	23	female	58.0	1	1	1	0	0
15	14	20	female	55.0	1	0	1	1	0
16	15	29	female	65.0	1	NA	NA	1	1

```
data <- na.omit(data)
```

- Removing rows with at least one missing value

	Patient_id	Age	Gender	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
1	1	22	female	57.7	1	0	2	0	0
2	2	26	female	63.0	2	0	1	0	1
3	3	26	female	62.0	2	1	1	0	0
4	4	28	female	65.0	1	0	2	0	0
5	5	22	female	58.0	2	0	1	0	1
6	6	26	female	63.0	1	1	0	0	0
7	7	27	female	64.0	2	0	1	0	0
8	8	32	female	70.0	3	0	1	0	1
10	10	27	female	64.5	1	1	1	0	1
11	11	36	female	75.0	1	0	1	0	0
13	12	33	female	70.0	1	1	0	0	1
14	13	23	female	58.0	1	1	1	0	0
15	14	20	female	55.0	1	0	1	1	0
17	16	25	female	61.5	1	2	0	0	0
18	17	25	female	61.5	1	0	1	0	0
19	18	20	female	55.5	1	2	2	0	1

Showing 1 to 17 of 68 entries, 9 total columns

```
colSums(is.na(data))
```

- After handling all the missing values, the NA count of All Column is 0.

```
> data <- na.omit(data)
> colSums(is.na(data))
Patient_id      Age      Gender      weight.kg. Delivery_number Delivery_time
0              0              0              0              0              0
Blood          Heart      Caesarian
0              0              0
```

```
data <- data <- distinct(data)
```

- Handling duplicate values in dataset.

Patient_id	Age	Gender	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
1	1	22	female	57.7	1	0	2	0
2	2	26	female	63.0	2	0	1	0
3	3	26	female	62.0	2	1	1	0
4	4	28	female	65.0	1	0	2	0
5	5	22	female	58.0	2	0	1	0
6	6	26	female	63.0	1	1	0	0
7	7	27	female	64.0	2	0	1	0
8	8	32	female	64.0	3	0	1	0
9	10	27	female	64.5	1	1	1	0
10	11	36	female	75.0	1	0	1	0
11	12	33	female	70.0	1	1	0	0
12	13	23	female	58.0	1	1	1	0
13	14	20	female	55.0	1	0	1	1
14	16	25	female	61.5	1	2	0	0
15	17	25	female	61.5	1	0	1	0
16	18	20	female	55.5	1	2	2	0

```
data <- subset(data, Age >= 18 & Age <= 45 & weight.kg. >= 50 & weight.kg. <= 90)
```

- Filtering the outliers which are irrelevant

59	70	27	female	62.0	0	2	2	0	0	0
60	71	90	female	130.0	1	0	0	0	0	1
61	73	135	female	64.0	2	0	1	0	0	0
62	74	32	female	69.0	3	0	1	1	1	0
63	75	38	female	75.0	3	2	2	1	1	1
64	76	27	female	62.5	2	1	1	0	0	1
65	79	25	female	140.0	1	2	0	0	0	1
66	80	120	female	57.0	2	2	1	0	0	0

Then,

55	65	31	female	66.0	1	2	2	1	1
56	67	28	female	62.5	3	0	1	0	1
57	68	29	female	64.5	2	0	1	1	0
58	69	25	female	62.0	1	0	0	0	1
59	70	27	female	61.0	2	2	0	0	0
62	74	32	female	69.0	3	0	1	1	0
63	75	38	female	75.0	3	2	2	1	1
64	76	27	female	62.5	2	1	1	0	1

Showing 45 to 61 of 61 entries, 9 total columns

```
data$Caesarian <- factor(data$Caesarian, levels = c(0, 1), labels = c("No", "Yes"))
```

```
data$Heart <- factor(data$Heart, levels = c(0, 1), labels = c("apt", "inept"))
```

```
data$Delivery_time <- factor(data$Delivery_time, levels = c(0, 1, 2), labels = c("timely", "premature", "latecomer"))
```

```
data$Blood <- factor(data$Blood, levels = c(0, 1, 2), labels = c("low", "normal", "high"))
```

- Converting Caesarian, Heart Condition, Delivery time and Blood pressure into Numerical to Catagorical values as per Data frame.

Mid_Project.R* × data ×

Filter

	Patient_id	Age	Gender	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
1	1	22	female	57.7	1	timely	high	apt	No
2	2	26	female	63.0	2	timely	normal	apt	Yes
3	3	26	female	62.0	2	premature	normal	apt	No
4	4	28	female	65.0	1	timely	high	apt	No
5	5	22	female	58.0	2	timely	normal	apt	Yes
6	6	26	female	63.0	1	premature	low	apt	No
7	7	27	female	64.0	2	timely	normal	apt	No
8	8	32	female	70.0	3	timely	normal	apt	Yes
9	10	27	female	64.5	1	premature	normal	apt	Yes
10	11	36	female	75.0	1	timely	normal	apt	No
11	12	33	female	70.0	1	premature	low	apt	Yes
12	13	23	female	58.0	1	premature	normal	apt	No
13	14	20	female	55.0	1	timely	normal	inept	No
14	16	25	female	61.5	1	latecomer	low	apt	No
15	17	25	female	61.5	1	timely	normal	apt	No
16	18	20	female	55.5	1	latecomer	high	apt	Yes

Showing 1 to 17 of 61 entries, 9 total columns

```
yes_count <- sum(data$Caesarian == "Yes")
```

```
print(yes_count)
```

```
No_count <- sum(data$Caesarian == "No")
```

```
print(No_count)
```

- Counting number of Caesarian and Non-Caesarian Patient in Dataset

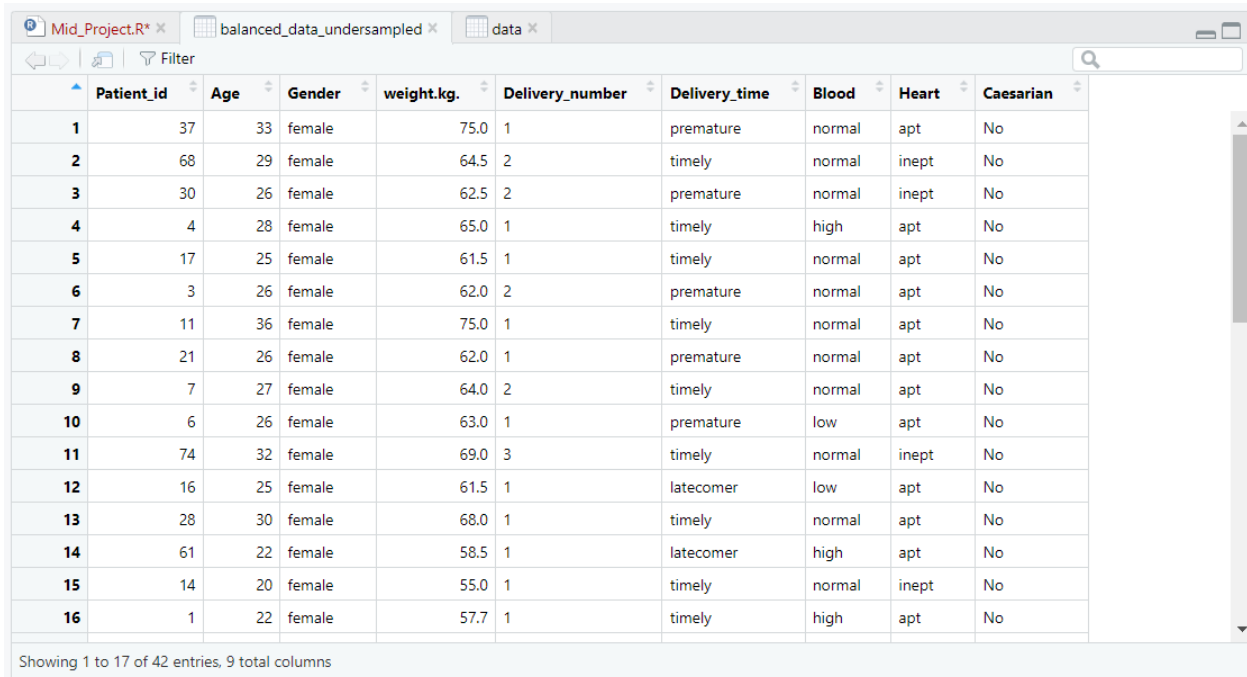
```
> yes_count <- sum(data$Caesarian == "Yes")
> print(yes_count)
[1] 40
> No_count <- sum(data$Caesarian == "No")
> print(No_count)
[1] 21
>
```

```
set.seed(123)
```

```
balanced_data_undersampled <- downSample(x = data[, -which(names(data) == "Caesarian")],
```

```
      y = data$Caesarian, yname = "Caesarian")
```

- Reproducing the data and balancing the dataset using Under sampling method where we decrease the majority class to be equal to minority class.



	Patient_id	Age	Gender	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
1	37	33	female	75.0	1	premature	normal	apt	No
2	68	29	female	64.5	2	timely	normal	inept	No
3	30	26	female	62.5	2	premature	normal	inept	No
4	4	28	female	65.0	1	timely	high	apt	No
5	17	25	female	61.5	1	timely	normal	apt	No
6	3	26	female	62.0	2	premature	normal	apt	No
7	11	36	female	75.0	1	timely	normal	apt	No
8	21	26	female	62.0	1	premature	normal	apt	No
9	7	27	female	64.0	2	timely	normal	apt	No
10	6	26	female	63.0	1	premature	low	apt	No
11	74	32	female	69.0	3	timely	normal	inept	No
12	16	25	female	61.5	1	latecomer	low	apt	No
13	28	30	female	68.0	1	timely	normal	apt	No
14	61	22	female	58.5	1	latecomer	high	apt	No
15	14	20	female	55.0	1	timely	normal	inept	No
16	1	22	female	57.7	1	timely	high	apt	No

Showing 1 to 17 of 42 entries, 9 total columns

New, yes and no counter

```
yes_count <- sum(balanced_data_undersampled$Caesarian == "Yes")
```

```
print(yes_count)
```

```
No_count <- sum(balanced_data_undersampled$Caesarian == "No")
```

```
print(No_count)
```

```
> yes_count <- sum(balanced_data_undersampled$Caesarian == "Yes")
> print(yes_count)
[1] 21
> No_count <- sum(balanced_data_undersampled$Caesarian == "No")
> print(No_count)
[1] 21
> |
```

```
balanced_data_stats <- data.frame(
```

```
  Variable = c("Age", "Weight"),
```

```
  Mean = c(mean(balanced_data_undersampled$Age),
mean(balanced_data_undersampled$weight.kg.)),
```

```
  Median = c(median(balanced_data_undersampled$Age),
median(balanced_data_undersampled$weight.kg.)),
```

```
  Mode = c(as.numeric(names(sort(table(balanced_data_undersampled$Age), decreasing=TRUE)[1])),
as.numeric(names(sort(table(balanced_data_undersampled$weight.kg.), decreasing=TRUE)[1]))),
```

```
  SD = c(sd(balanced_data_undersampled$Age), sd(balanced_data_undersampled$weight.kg.)),
```

```
  Variance = c(var(balanced_data_undersampled$Age), var(balanced_data_undersampled$weight.kg.)),
```

```
  Range = l(list(range(balanced_data_undersampled$Age),
range(balanced_data_undersampled$weight.kg.)))
```

```
)
```

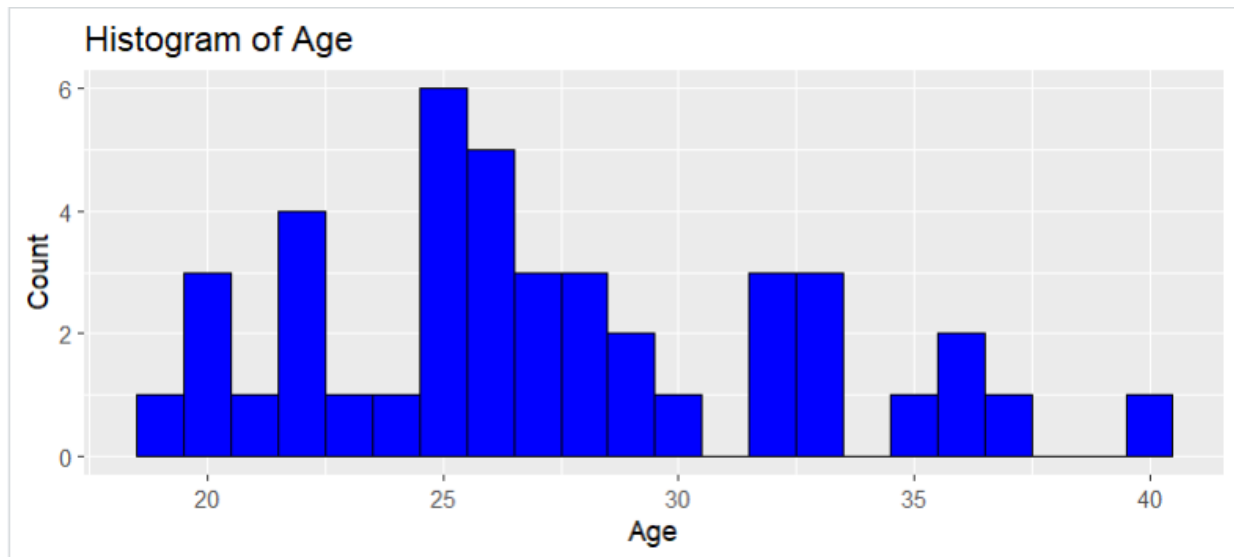
```
print(balanced_data_stats)
```

- Finding Measure of Central Tendency (Mean, Mode, Median) and Measure of Spread (Range, Standard Deviation and Variance)

```
>
> print(balanced_data_stats)
  Variable    Mean Median Mode      SD Variance  Range
1    Age 27.30952  26.00   25 5.167856 26.70674 19, 40
2  weight 63.51905  62.25   62 6.964167 48.49963 51, 82
> |
```

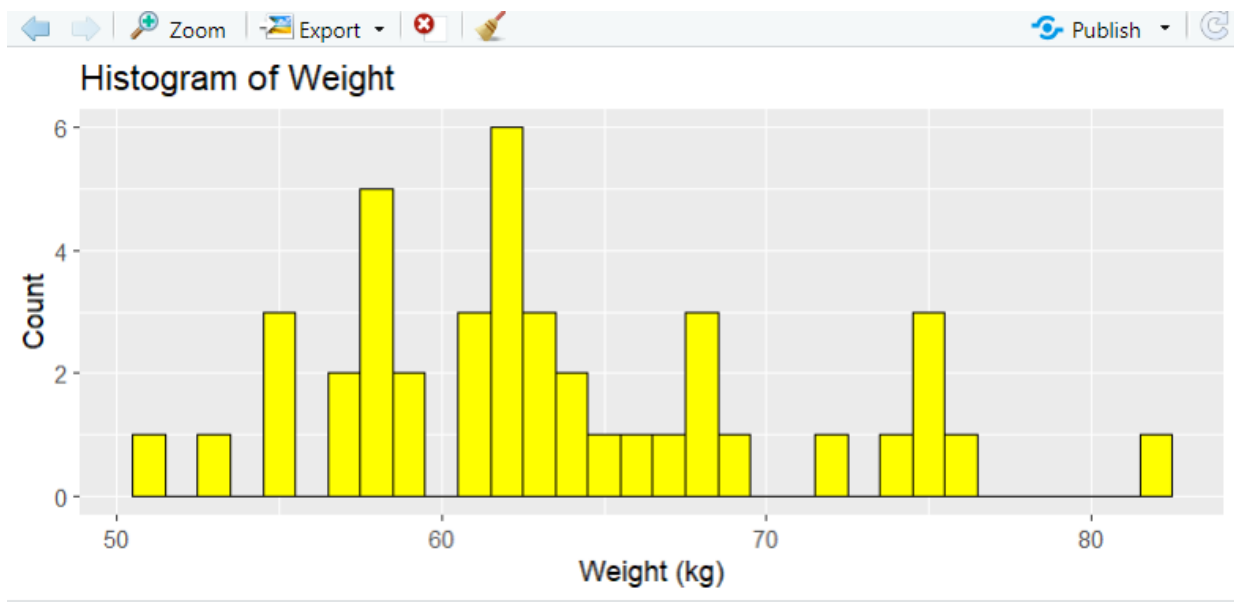
```
ggplot(balanced_data_undersampled, aes(x = Age)) + geom_histogram(binwidth = 1, fill = "skyblue",  
color = "black") + labs (title = "Histogram of Age", x = "Age", y = "Count")
```

- Histogram of Age which is continuous attribute



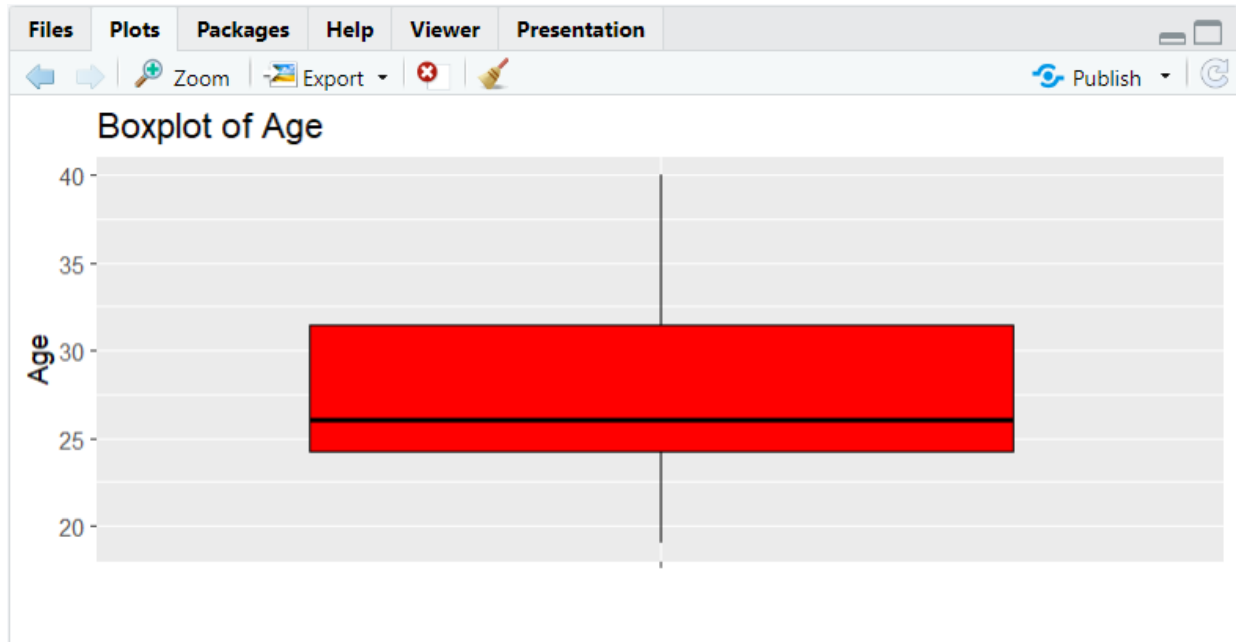
```
ggplot(balanced_data_undersampled, aes(x = weight.kg.)) + geom_histogram(binwidth = 1, fill =  
"yellow", color = "black") + labs(title = "Histogram of Weight", x = "Weight (kg)", y = "Count")
```

- Histogram of Weight which is continuous attribute



```
ggplot(balanced_data_undersampled, aes(x = "", y = Age)) + geom_boxplot(fill = "red", color = "black") +  
labs (title = "Boxplot of Age", x = "", y = "Age")
```

- Box Plot of Age which is continuous attribute



```
ggplot(balanced_data_undersampled, aes(x = "", y = weight.kg.)) + geom_boxplot(fill = "green", color =  
"black") + labs(title = "Boxplot of Weight", x = "", y = "Weight (kg)")
```

- Box Plot of Wright which is continuous attribute

