



**Department of Computer Science  
American International University-Bangladesh  
Mid Term Project -Report**

**Course Name: INTRODUCTION TO DATA SCIENCE**

**“A report on Data Pre-Processing”**

**Supervised By:**

Tohedul Islam

Associate Professor, Computer Science

**Submitted By:**

Md Shahadat Hossen

ID: 20-43083-1

Section: E

## **Project Title: Pre-processing on a Dataset.**

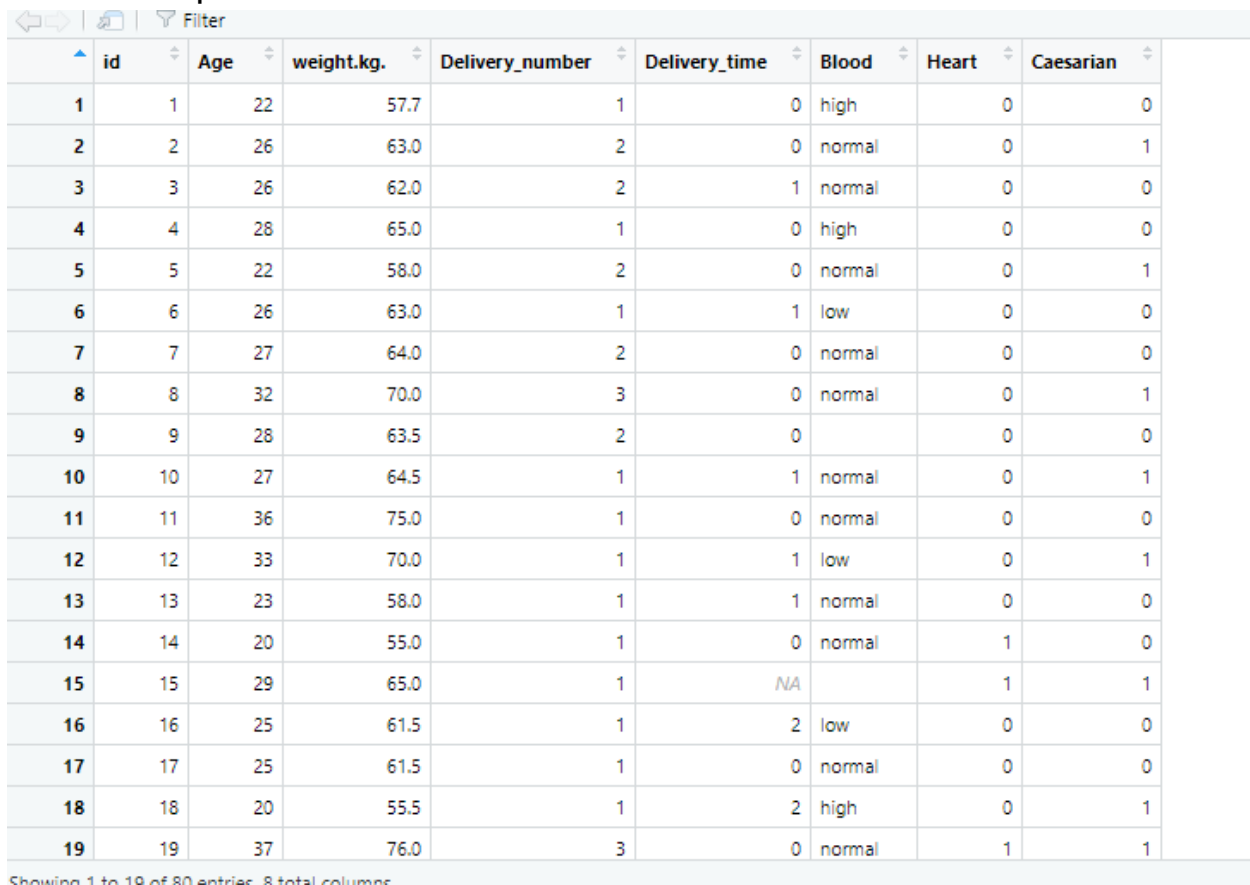
### **Project overview:**

Text, images, and videos are examples of messy, real-world data. In addition to having the potential to be inaccurate and inconsistent, they are frequently unfinished and lack a regular and uniform design. Machines prefer to read data as 1s and 0s and process it in a well-organized manner. Because of this, calculating structured data like integers and percentages is simple. Nevertheless, before analysis, unstructured data in the form of text and images must be prepared by cleaning and formatting. Data preparation describes the procedures that change or encrypt data so that a computer can quickly understand it. The algorithm must be rapid at interpreting the characteristics of the data in order for the model to produce accurate and precise predictions. the 5 major steps of data preprocessing: Data quality assessment, Data cleaning. Data transformation, Data reduction. To fill in missing values, smooth noisy data, resolve inconsistencies, and eliminate outliers, data cleaning is a phase in the data preprocessing process. Data integration is the phase of data preparation where data from various sources is combined into one massive data storage. data repository. Data transformation is the process of converting the value, structure, or format of high-quality data into new formats utilizing methods like scaling, normalization, and others.

## Import csv file:

```
middataset <- read.csv("D:/midtermdataset.csv",header=TRUE,sep=",")  
middataset
```

Here we import the csv file to R studio



	id	Age	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
1	1	22	57.7	1	0	high	0	0
2	2	26	63.0	2	0	normal	0	1
3	3	26	62.0	2	1	normal	0	0
4	4	28	65.0	1	0	high	0	0
5	5	22	58.0	2	0	normal	0	1
6	6	26	63.0	1	1	low	0	0
7	7	27	64.0	2	0	normal	0	0
8	8	32	70.0	3	0	normal	0	1
9	9	28	63.5	2	0		0	0
10	10	27	64.5	1	1	normal	0	1
11	11	36	75.0	1	0	normal	0	0
12	12	33	70.0	1	1	low	0	1
13	13	23	58.0	1	1	normal	0	0
14	14	20	55.0	1	0	normal	1	0
15	15	29	65.0	1	NA		1	1
16	16	25	61.5	1	2	low	0	0
17	17	25	61.5	1	0	normal	0	0
18	18	20	55.5	1	2	high	0	1
19	19	37	76.0	3	0	normal	1	1

Showing 1 to 19 of 80 entries, 8 total columns

## Shape of the dataset:

Code-

```
summary(middataset)
```

```
> summary(middataset)
      id      Age      weight.kg.  Delivery_number Delivery_time
Min.   : 1.00   Min.   :18.00   Min.   : 49.00   Min.   :1.000   Min.   :0.0000
1st Qu.:20.75   1st Qu.:25.00   1st Qu.: 61.00   1st Qu.:1.000   1st Qu.:0.0000
Median :40.50   Median :28.00   Median : 63.50   Median :1.500   Median :0.0000
Mean   :40.50   Mean   :29.68   Mean   : 65.13   Mean   :1.679   Mean   :0.6234
3rd Qu.:60.25   3rd Qu.:32.00   3rd Qu.: 68.00   3rd Qu.:2.000   3rd Qu.:1.0000
Max.   :80.00   Max.   :95.00   Max.   :110.00   Max.   :4.000   Max.   :2.0000
      NA's      :3      NA's      :3      NA's      :2      NA's      :3

      Blood      Heart      Caesarian
Length:80      Min.   :0.000   Min.   :0.0000
Class :character 1st Qu.:0.000   1st Qu.:0.0000
Mode  :character 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
```

---

**Missing value handling :**  
**counting null value (NA) for each attribute:**

Code-  
 colsums(is.na(middataset))

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

**Detect NA value and delete that all row:**

Code-  
 deleteNa <- na.omit(middataset)

In this case we detect the the NA values and delete it for data clearing

	id	Age	weight.kg.	Delivery_number	Delivery_time	Blood	Heart	Caesarian
7	7	27	64.0	2	0	normal	0	0
8	8	32	70.0	3	0	normal	0	1
9	9	28	63.5	2	0		0	0
10	10	27	64.5	1	1	normal	0	1
11	11	36	75.0	1	0	normal	0	0
12	12	33	70.0	1	1	low	0	1
13	13	23	58.0	1	1	normal	0	0
14	14	20	55.0	1	0	normal	1	0
16	16	25	61.5	1	2	low	0	0
17	17	25	61.5	1	0	normal	0	0
18	18	20	55.5	1	2	high	0	1
19	19	37	76.0	3	0	normal	1	1
20	20	24	56.6	1	2	low	1	1
21	21	26	62.0	1	1	normal	0	0
22	22	33	75.0	2	0	low	1	1
23	23	25	62.0	1	1	high	0	0
25	25	20	55.0	1	0	high	1	1
28	28	30	68.0	1	0	normal	0	0
29	29	32	73.0	1	0	high	1	1

Showing 6 to 25 of 70 entries. 8 total columns

Here, we can see id no – 24,26 rows are deleted because of missing value.

### another way of missing value handling:

For every row and column we have many missing values . so finding the missing values we use mean, median and mode for removing NA values.

#### Age:

We can adjust missing value of age by mean, median, mode values.

#### Mean:

Code-

```
middataset$Age[is.na(middataset$Age)] = mean(middataset$Age, na.rm = TRUE)
```

47	47	26.00000	NA	1	0	normal	0	0
48	48	32.00000	67.5	2	0	high	1	1
49	49	26.00000	62.5	2	2	normal	0	0
50	50	29.67532	NA	2	0	low	1	1
51	51	33.00000	68.5	3	2	normal	1	0
52	52	21.00000	53.0	2	1	low	1	1

Here, mean value = 29.67532  
Which we can notice into id - 50

### Median:

Code-

```
middataset$Age[is.na(middataset$Age)] = median(middataset$Age, na.rm = TRUE)
```

48	48	32	67.5	2	0	high	1	1
49	49	26	62.5	2	2	normal	0	0
50	50	26	NA	2	0	low	1	1
51	51	33	68.5	3	2	normal	1	0
52	52	21	53.0	2	1	low	1	1

Here, median value =26 (id- 50)

### Mode:

Code-

```
library(DescTools)
```

```
middataset$Age[is.na(middataset$Age)] <- Mode(middataset$Age, na.rm = TRUE)
```

48	48	32	67.5	2	0	high	1	1
49	49	26	62.5	2	2	normal	0	0
50	50	26	NA	2	0	low	1	1
51	51	33	68.5	3	2	normal	1	0
52	52	21	53.0	2	1	low	1	1

Here, mode value = 26(id -50)

So, for age mode and median value are same.so we can use one.

### Weight:

#### Mean:

Code-

```
middataset$weight.kg.[is.na(middataset$weight.kg.)] <-  
mean(middataset$weight.kg., na.rm = TRUE)
```

46	46	28	62.50000	3	0	normal	1	1
47	47	26	65.12727	1	0	normal	0	0
48	48	32	67.50000	2	0	high	1	1
49	49	26	62.50000	2	2	normal	0	0
50	50	NA	65.12727	2	0	low	1	1
51	51	33	68.50000	3	2	normal	1	0

Here, mean value = 65.12727 (id – 47,50)

#### Median:

Code-

```
middataset$weight.kg.[is.na(middataset$weight.kg.)] <-  
median(middataset$weight.kg., na.rm = TRUE)
```

47	47	26	63.5	1	0	normal	0	0
48	48	32	67.5	2	0	high	1	1
49	49	26	62.5	2	2	normal	0	0
50	50	NA	63.5	2	0	low	1	1
51	51	33	68.5	3	2	normal	1	0

Here, median value =63.5 (id- 47,50)

#### Mode:

Code-

```
library(DescTools)
```

```
middataset$weight.kg.[is.na(middataset$weight.kg.)] <-  
Mode(middataset$weight.kg., na.rm = TRUE)
```

46	46	28	62.5	3	0	normal	1	1
47	47	26	63.0	1	0	normal	0	0
48	48	32	67.5	2	0	high	1	1
49	49	26	62.5	2	2	normal	0	0
50	50	NA	63.0	2	0	low	1	1
51	51	33	68.5	3	2	normal	1	0
52	52	21	53.0	2	1	low	1	1

Here, mode value = 63 (id -47,50)

### Delivery Number:

We can't do mean,median of delivery number because meaning of dataset {1,2,3,4}

### Mode:

Code-

```
library(DescTools)
```

```
middataset$Delivery_number[is.na(middataset$Delivery_number)] <-
```

```
Mode(middataset$Delivery_number, na.rm = TRUE)
```

23	23	25	62.0	1	1	high	0	0
24	24	27	65.0	1	NA	low	1	1
25	25	20	55.0	1	0	high	1	1
26	26	18	49.0	1	0	normal	0	0
27	27	18	50.0	1	NA	high	1	1

Here, mode value = 1 (id -24,26)

### Delivery Time:

We can't do mean,median of delivery time because meaning of dataset {0,1,2}

0 = timely

1 = premature

2 = latecomer



**Mode:**

Code-

library(DescTools)

middataset\$Delivery\_time[is.na(middataset\$Delivery\_time)] &lt;-

Mode(middataset\$Delivery\_time, na.rm = TRUE)

24	24	27	65.0	NA	0	low	1	1
25	25	20	55.0	1	0	high	1	1
26	26	18	49.0	NA	0	normal	0	0
27	27	18	50.0	1	0	high	1	1
28	28	30	68.0	1	0	normal	0	0
29	29	32	73.0	1	0	high	1	1

Here, mode value = 0 (id -24,27)

**Heart problem:**

No missing value of heart problem.

**Caesarian:**

We can't do mean, median of delivery time because meaning of dataset {0,1}

0 means "No"

1 means "YES"

Code-

library(DescTools)

middataset\$Caesarian[is.na(middataset\$Caesarian)] &lt;-

Mode(middataset\$Caesarian, na.rm = TRUE)

59	59	26	61.5	1	0	high	0	1
60	60	30	67.5	2	1	high	1	1
61	61	22	58.5	1	2	high	0	0
62	62	NA	NA	1	0	normal	0	1
63	63	32	67.0	2	0	low	0	1
64	64	32	67.0	2	0	normal	1	1
65	65	31	66.0	1	2	high	1	0
66	66	35	72.0	2	0	normal	0	1
67	67	28	62.5	3	0	normal	0	1
68	68	29	64.5	2	0	normal	1	0
69	69	25	62.0	1	0	low	0	1
70	70	27	61.0	2	2	low	0	0
71	71	90	105.0	1	0	low	0	1
72	72	29	65.0	1	2		1	1
73	73	28	64.0	2	0	normal	0	0
74	74	32	69.0	3	0	normal	1	0
75	75	38	75.0	3	2	high	1	1
76	76	27	62.5	2	1	normal	0	0
77	77	33	66.0	4	0	normal	0	1

Showing 59 to 77 of 80 entries 8 total columns

Here, mode value = 1 (id -60,77)

## Data visualization:

### Noisy data-

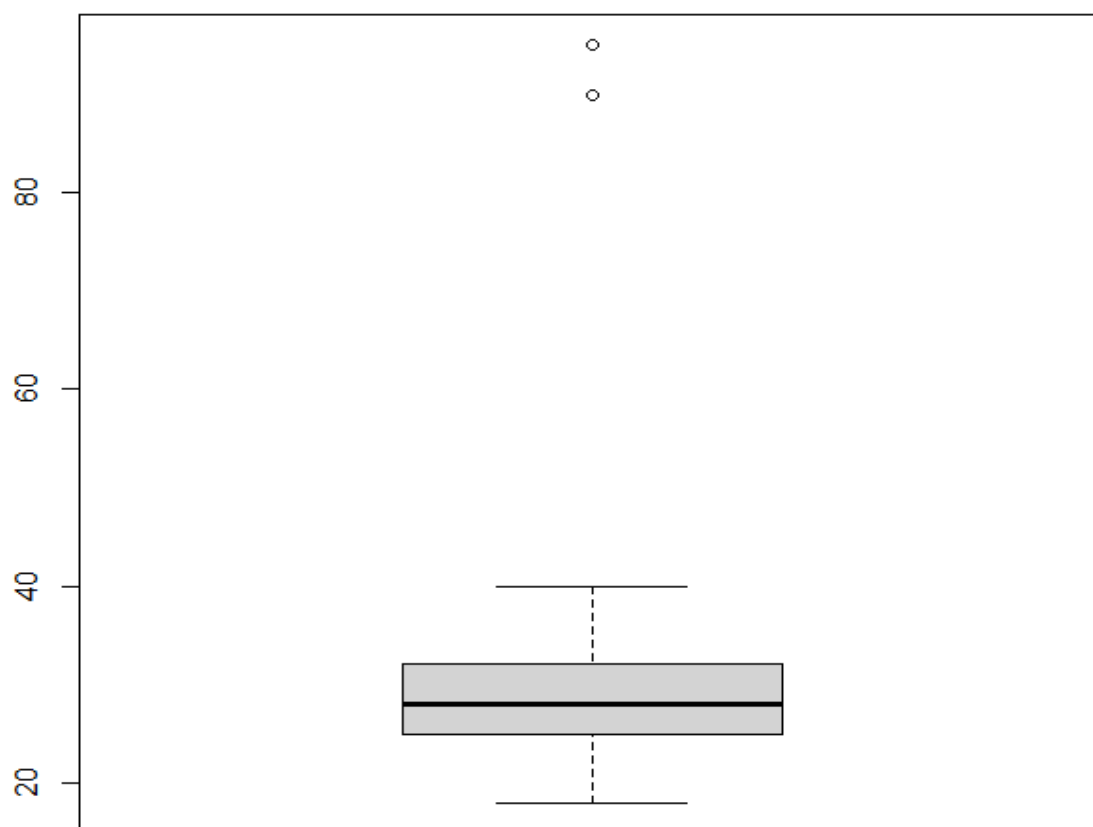
#### Age:

Code-

```
ageboxplot<-boxplot(middataset$Age)
```

```
Outlier<-ageboxplot$out
```

Outlier



Here, two 80 up age are noisy data

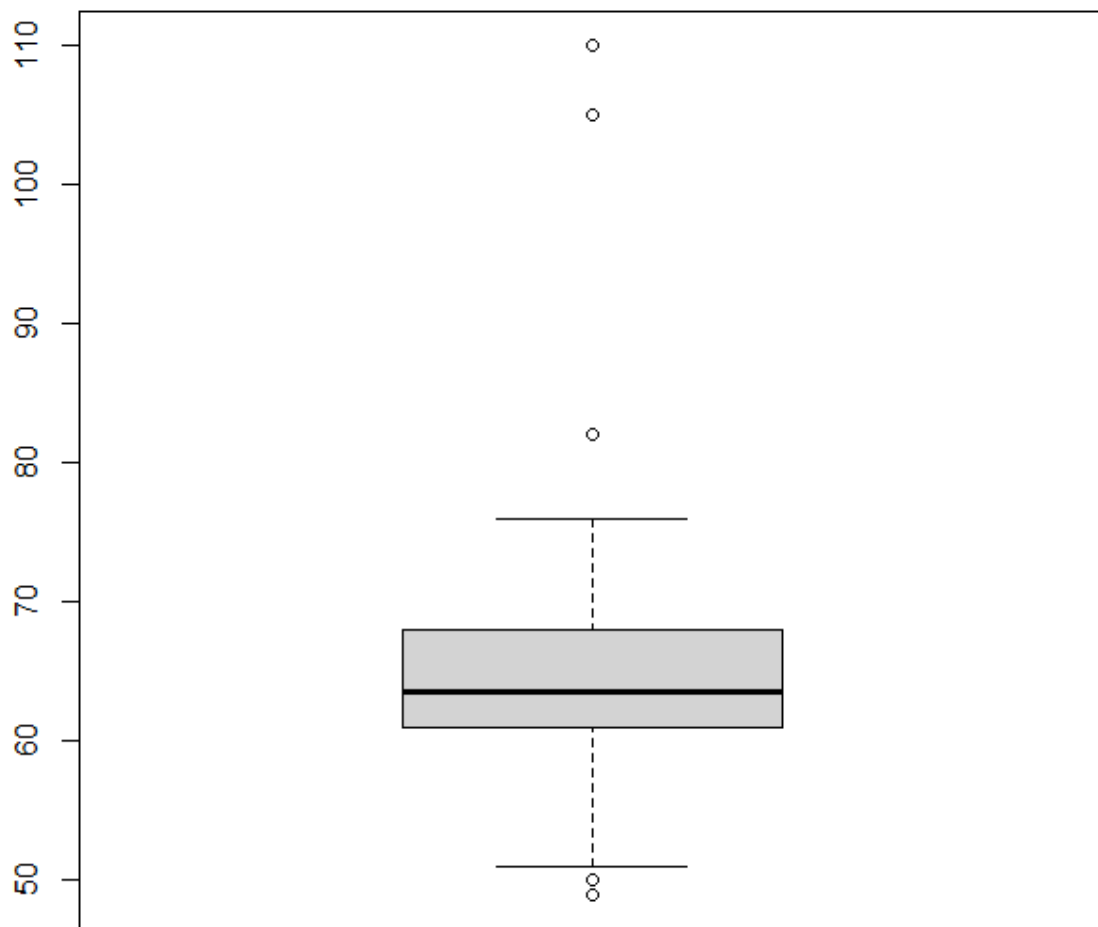
**weight:**

Code-

```
weightboxplot<-boxplot(middataset$weight.kg.)
```

```
Outliers<-weightboxplot$out
```

Outliers



Here, one 80 up & two 100 up kg weight are noisy data

## **Conclusion Of The project:**

We will gradually improve the data and process it using R language constructs and methods. The dataset was better and cleaner after all data preparation techniques were successfully used. But, I wasn't required to employ the entire technique for this assignment. I gained knowledge of the industry's most recent data and data preprocessing. Increase the number of tools in your toolkit. Preprocessing the data will increase the correctness of your dataset. Values that are incorrect or missing as a result of human error or other issues are eliminated. increased reliability. Perhaps more crucially, I had numerous issues when working with the data. In a single instance, I created an inaccurate column, added it to the data frame, and obtained the incorrect results. With the help of the R language, frameworks, and techniques, we will improve the data and process it. After all data cleaning procedures were successfully applied, the dataset was nicer and cleaner.