# American International University-Bangladesh



# **Mid Project**

# **Submitted to:**

NAME: TOHEDUL ISLAM

Subject: INTRODUCTION TO DATA SCIENCE

**SECTION: A** 

# **Submitted by:**

Name	ID
Sayeem Bin Ezaz	20-43770-2
Ridita Zaman Adikta	20-43679-2
Maliha Mehjabin	20-42508-1

Date of Submission: 11.11.2023

#### **Short Summary of Dataset:**

The dataset contains information related to factors contributing to heart attacks and consists of 1319 samples, each comprising nine fields. The fields include age, gender, heart rate, systolic and diastolic blood pressure, blood sugar, CK-MB, and Test-Troponin. Gender is represented as 0 for Female and 1 for Male, while CK-MB and Test-Troponin may indicate enzyme or troponin measurements, respectively. The dataset's output field, labeled "Class," categorizes the presence of heart attacks into "Negative" for absence and "Positive" for presence. It's worth noting that there are missing values, particularly in the gender and CK-MB fields, and some unusual data values that would require further data cleaning and preprocessing for meaningful analysis.

#### **Data Import and View:**

**Code:** MidData<- read.csv("V:/MidProject.csv",header=TRUE ,sep=",")

#### MidData

>	> MidData									
				pressurehight	pressurelow		class			
1	64	male	66	160	83		negative			
2	21	male	94	98	46		positive			
3	55	male	64	-160	77		negative			
4	64	male	70	120	55		positive			
5	55	male	64	112	65	300	negative			
6		female	61	112	58	87	negative			
7		female	40	179	68		negative			
8	63	male	60	214	82	87	positive			
9		female	60	NA	81		negative			
10	67		61	160	95	100	negative			
11		female	60	166	90	102	negative			
1.2	63	female	60	150	10		negative			
13	64	male	60	199	5	92	positive			
14		female	94	122	67	97	negative			
1.5	47	male	76	120	70	319	negative			
16		male	81	NA	66	134	positive			
1.7	86	female	73	114	68	87	positive			
18		female	70	100	68	96	negative			
19	37	female	72	107	86	274	negative			
20	45	male	60	109	65	89	positive			
21	60	male	92	151	78	301	negative			
22	48	male	135	98	60	100	positive			
23	52	male	76	109	85	227	positive			
24	30	male	63	110	68	107	positive			
25	NA.	male	63	320	63	269	positive			
26	72	male	64	106	68	111	positive			
27	42	male	65	150	68	101	negative			
28	72	female	64	325	60	95	negative			
29	47	female	66	134	57	279	positive			
30	63	male	66	135	5.5		negative			
31	. 54	male	125	131	82	95	positive			
32	35	male	62	137	61	321	negative			
33	68	male	61	121	49	98	positive			
34	56	female	60	145	62	105	negative			
35	5 50	male	61	136	70	136	positive			
36	64	male	58	156	76	82	positive			
37	NA.	male	60	166	82	117	negative			
38	64	male	65	155	75	107	negative			

Description: Data is read into a variable called MidData from a CSV file.

# **Unwanted Sign or Invalid value corrected from Pressurehight:**

**Code:** MidData\$pressurehight<-gsub("-","",MidData\$pressurehight)

#### MidData

- > MidData\$pressurehight<-gsub("-","",MidData\$pressurehight)</pre>
- > MidData

> MIGDALA										
	age	gender	impluse	pressurehight	pressurelow	glucose	class			
1	64	male	66	160	83		negative			
2	21	male	94	98	46	296	positive			
3	55	male	64	160	77	270	negative			
4	64	male	70	120	55	270	positive			
5	55	male	64	112	65	300	negative			
6	58	female	61	112	58	87	negative			
7	32	female	40	179	68	102	negative			
8	63	male	60	214	82	87	positive			
9	44	female	60	<na></na>	81	135	negative			
10	67		61	160	95	100	negative			
11	NA	female	60	166	90	102	negative			
12	63	female	60	150	10	198	negative			
13	64	male	60	199	5	92	positive			
14	54	female	94	122	67	97	negative			
15	47	male	76	120	70	319	negative			
16	61	male	81	<na></na>	66	134	positive			
17	86	female	73	114	68	87	positive			
18	45	female	70	100	68	96	negative			
19	37	female	72	107	86	274	negative			
20	45	male	60	109	65	89	positive			
21	60	male	92	151	78	301	negative			
22	48	male	135	98	60	100	positive			
23	52	male	76	109	85	227	positive			
24	30	male	63	110	68	107	positive			
25	NA	male	63	320	63	269	positive			
26	72	male	64	106	68	111	positive			
27	42	male	65	150	68	101	negative			
28	72	female	64	325	60	95	negative			
29	47	female	66	134	57		positive			
30	63	male	66	135	55	166	negative			
31	54	male	125	131	82		positive			
32	35	male	62	137	61		negative			
33	68	male	61	121	49		positive			
34	56	female	60	145	62		negative			
		-								

**Description:** In this dataset gsub() function is used to corrected or remove unwanted signs. There was a "-" sign at preassureheight column. Which may create problem to analysis the data or during work with data.

# Replace blank values with null values:

# **Code:**

MidData[MidData == ""] <- NA

## MidData

>	. MidD	ata[Mi	dD:	ata "	"] <- NΔ							
	> MidData[MidData == ""] <- NA > MidData											
1			er	impluse	pressurehight	pressurelow	glucose	class				
1	. 6	4 ma	1e	66	160	83	160	negative				
2	2	1 ma	٦e	94	98	46	296	positive				
3	5	5 ma	٦e	64	160	77	270	negative				
4	6	4 ma	٦e	70	120	55	270	positive				
5	5	5 ma	٦e	64	112	65	300	negative				
6	5	8 fema	٦e	61	112	58	87	negative				
7	3	2 fema	٦e	40	179	68	102	negative				
8				60	214	82	87	positive				
9	4	4 fema	٦e	60	<na></na>	81	135	negative				
1	.0 6			61	160	95	100	negative				
1	.1 N	A fema	٦e	60	166	90	102	negative				
1	.2 6	3 fema	٦e	60	150	10	198	negative				
1	.3 6	4 ma	٦e	60	199	5	92	positive				
		4 fema	٦e	94	122	67	97	negative				
1	.5 4			76	120	70	319	negative				
1	6 6			81	<na></na>	66	134	positive				
		6 fema		73	114	68	87	positive				
		5 fema		70	100	68	96	negative				
	.9 3			72	107	86	274	negative				
	0 4			60	109	65	89	positive				
	1 6			92	151	78		negative				
	2 4			135	98	60		positive				
2	3 5			76	109	85	227	positive				
	4 3			63	110	68		positive				
2	5 N			63	320	63	269	positive				
	6 7			64	106	68	111	positive				
	7 4			65	150	68		negative				
		2 fema		64	325	60		negative				
		7 fema		66	134	57		positive				
	0 6			66	135	55		negative				
	1 5			125	131	82		positive				
3	י כ	5 - 112	۱۸	67	1 27	61	271	nogativo				

**Description:** Removing blank spaces with null value is important. Cause sometime blank space takes place as not null value.

#### **Summary of Dataset:**

#### Code:

MidData\$pressurehight<-as.integer(MidData\$pressurehight)

#### summary(MidData)

```
> MidData$pressurehight<-as.integer(MidData$pressurehight)
```

```
> summary(MidData)
     age
                        gender
                                              impluse
                                                               pressurehight
                                                                                   pressurelow
                                                                                                        glucose
                                                                                                                            class
       : 19.00
                                          Min. : 40.00
1st Qu.: 62.00
                                                                                                     Min. : 66.00
1st Qu.: 97.25
Min.
                    Length:150
                                                               Min. : 85.0
1st Qu.:110.8
                                                                                  Min. : 5.00
1st Qu.:60.25
                                                                                                                        Length:150
1st Qu.: 46.00
Median : 56.00
Mean : 56.14
                    Class :character
                                                                                                                        Class :character
                                                               Median :122.5
Mean :129.2
                                                      74.00
                                                                                                     Median :116.00
                    Mode :character
                                          Median :
                                                                                  Median :69.00
                                                                                                                         Mode :character
                                           Mean
                                                  : 81.98
                                                                                  Mean
                                                                                         :68.95
                                                                                                     Mean
                                                                                                            :148.65
 3rd Qu.: 64.00
                                           3rd Qu.: 83.00
                                                               3rd Qu.:140.0
                                                                                  3rd Qu.:80.00
                                                                                                     3rd Qu.:179.25
Max. :155.00
NA's :5
                                           Max. :1111.00
                                                               Max. :325.0
                                                                                  Max.
                                                                                         :95.00
                                                               NA's
                                                                        :2
```

**Description:** This uses the summary() function to create a summary of the MidData data. For removing unwanted sign data type of preassurehight column changed. So we use as integer function again to change the data type & get the proper summary of data.

#### **Display Dataset:**

#### Code: str(MidData)

```
> str(MidData)
'data.frame':
                150 obs. of 7 variables:
                : int 64 21 55 64 55 58 32 63 44 67 ...
 $ age
                       "male" "male" "male" ...
 $ gender
                : chr
                : int
 $ impluse
                       66 94 64 70 64 61 40 60 60 61 ...
                       160 98 160 120 112 112 179 214 NA 160 ...
 $ pressurehight: int
 $ pressurelow
               : int
                       83 46 77 55 65 58 68 82 81 95 ...
 $ glucose
                       160 296 270 270 300 87 102 87 135 100 ...
                 int
                       "negative" "positive" "negative" "positive" ...
 $ class
                : chr
```

**Description:** R object structures are shown with "str." The data's contents are displayed using the string "str". An alternate function to show the output summary is str(MidData), particularly in cases where the data set is large.

# **Missing Value Detection (Numeric Value):**

Code: is.na(MidData)

## > is.na(MidData)

> IS.Na(MIGDALA)									
			gender	impluse	pressurehight	pressurelow	glucose	class	
		FALSE	FALSE		FALSE	FALSE			
		FALSE	FALSE		FALSE	FALSE			
	[3,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[4,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[5,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[6,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[7,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[8,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[9,]	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	
		FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[11,]	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[12,]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	
	[13,]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	
	[14,]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	
	[15,]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	
	[16,]	FALSE	FALSE	FALSE	TRUE	FALSE		FALSE	
	[17,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[18,]	FALSE	FALSE	FALSE	FALSE	FALSE		FALSE	
	[19,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[20,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[21,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[22,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[23,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[24,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[25,]	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[26,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[27,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[28,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[29,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[30,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[31,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[32,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
		FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[34,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	[35,]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	
	۲36.٦	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	

**Description:** A data frame or vector's missing values can be found using the is.na() function. With each element being TRUE if the associated element in the data frame or vector is missing and FALSE otherwise, it yields a logical vector with the same length as the input data.

#### **Missing Value Count in Each Column**

**Code:** colSums(is.na(MidData))

```
> colSums(is.na(MidData))
    age    gender    impluse pressurehight pressurelow    glucose    class
    5     3     0     2     0     0     0
```

**Description:** The number of null elements in each column has been determined using the colSums(is.na()) function.

#### **Specific Missing Value row Number**

**Code:** which(is.na(MidData\$age))

```
> which(is.na(MidData$age))
[1] 11 25 37 74 122
> |
```

**Description:** The which() function is used for getting exact number of row is missing. We used this for age to see the missing values.

#### **Standard Deviation of all numeric values:**

```
Code: age<-MidData$age
sd(age)
impluse<-MidData$impluse
sd(impluse)
pressurehight<-MidData$pressurehight
sd(pressurehight)
pressurelow<-MidData$pressurelow
sd(pressurelow)
glucose<-MidData$glucose
sd(glucose)
```

```
> age<-MidData$age
> sd(age)
[1] 17.1392
> impluse<-MidData$impluse
> sd(impluse )
[1] 14.78093
> pressurehight<-MidData$pressurehight
> sd(pressurehight)
[1] 32.58258
> pressurelow<-MidData$pressurelow
> sd(pressurelow)
[1] 13.61182
> glucose<-MidData$glucose
> sd(glucose)
[1] 73.58597
```

**Description:** The sd() function computes the standard deviation of the column values. The sd() function in R is a built-in function that computes the standard deviation, which is a measure of the amount of variation or dispersion in a set of values. This step is done after all the data cleaning process.

# **Handling invalid data/outliers in the data set:**

# **Removing Missing Value:**

**Code:** RemovedMidData<-na.omit(MidData)

#### RemovedMidData

- > RemovedMidData<-na.omit(MidData)</pre>
- > RemovedMidData

	age gender impluse pressurehight pressurelow glucose class									
_	_	_		-	•		class			
1	64	male	66	160	83		negative			
2	21	male	94	98	46		positive			
3	55	male	64	160	77		negative			
4	64	male	70	120	55		positive			
5	55	male	64	112	65		negative			
6		female	61	112	58		negative			
7		female	40	179	68		negative			
8	63	male	60	214	82		positive			
12		female	60	150	10		negative			
13		male	60	199	5		positive			
14	54	female	94	122	67		negative			
15	47	male	76	120	70	319	negative			
17	86	female	73	114	68	87	positive			
18	45	female	70	100	68	96	negative			
19	37	female	72	107	86	274	negative			
20	45	male	60	109	65	89	positive			
21	. 60	male	92	151	78	301	negative			
22	48	male	135	98	60	100	positive			
23	52	male	76	109	85	227	positive			
24	30	male	63	110	68	107	positive			
26	72	male	64	106	68	111	positive			
27	42	male	65	150	68	101	negative			
28	72	female	64	325	60	95	negative			
29	47	female	66	134	57	279	positive			
30	63	male	66	135	55	166	negative			
31	. 54	male	125	131	82	95	positive			
32	35	male	62	137	61	321	negative			
33	68	male	61	121	49		positive			
34	56	female	60	145	62	105	negative			
35	50	male	61	136	70		positive			
36	64	male	58	156	76		positive			
38	64	male	65	155	75		negative			
•					7-	100				

**Description:** The na.omit() function is used to remove rows from the data MidData that have missing values. The data that has missing values removed is then stored in a new data frame called RemovedMidData.

# **Removing outliers from Impulse Column:**

#### **Code:**

MidData\$impluse[MidData\$impluse > 190] <- NA

#### MidData

- > MidData\$impluse[MidData\$impluse > 190] <- NA
- > MidData

> MIGDALA									
	gender	impluse		pressurelow		class			
64	male	66	160	83	160	negative			
21	male	94	98	46	296	positive			
55		64	160	77	270	negative			
64		70	120	55	270	positive			
55	male	64	112	65	300	negative			
58	female	61	112	58	87	negative			
32	female	NA	179	68	102	negative			
63	male	60	214	82	87	positive			
44	female	60	NA	81	135	negative			
67	<na></na>	61	160	95	100	negative			
. NA	female	60	166	90	102	negative			
63	female	60	150	10	198	negative			
64	male	60	199	5	92	positive			
54	female	94	122	67	97	negative			
47	male	76	120	70	319	negative			
61	male	81	NA	66	134	positive			
86	female	73	114	68	87	positive			
3 45	female	70	100	68	96	negative			
37	female	72	107	86	274	negative			
45	male	60	109	65	89	positive			
. 60	male	92	151	78	301	negative			
48	male	NA	98	60	100	positive			
52	male	76	109	85	227	positive			
30	male	63	110	68	107	positive			
NA.	male	63	320	63	269	positive			
72	male	64	106	68	111	positive			
42	male	65	150	68	101	negative			
72	female	64	325	60	95	negative			
47	female	66	134	57	279	positive			
63	male	66	135	55	166	negative			
. 54	male	NA	131	82	95	positive			
35	male	62	137	61	321	negative			
68	male	61	121	49	98	positive			
56	female	60	145	62	105	negative			
	age 64 21 55 64 55 58 32 63 44 67 NA 63 64 47 63 64 47 65 61 48 45 45 45 45 45 45 47 48 47 48 48 47 48 48 48 49 49 49 49 49 49 49 49 49 49 49 49 49	age gender 64 male 21 male 55 male 64 male 55 male 64 male 58 female 32 female 63 male 44 female 63 female 64 male 64 male 65 female 65 female 66 male 67 shale 68 female 69 and female 69 and female 60 male 60 male 60 male 61 male 62 male 63 female 64 male 65 female 66 male 67 male 68 male	age gender impluse 64 male 66 21 male 94 55 male 64 64 male 70 55 male 64 58 female 61 32 female NA 63 male 60 44 female 60 67 <na> 61 NA female 60 63 female 60 64 male 60 65 female 76 66 male 81 78 female 76 66 male 81 78 female 70 78 female 70 78 male 60 79 male 60 70 male 92 70 male 63 71 male 63 72 male 64 74 male 63 72 male 64 74 male 65 75 female 64 76 male 65 76 male 66 77 male 66 78 male 66 79 male 66 70 male 66 70 male 66 71 male 66 72 male 66 73 male 66 74 male 66 75 male 66 76 male 66</na>	age gender impluse pressurehight 64 male 66 160 21 male 94 98 55 male 64 160 64 male 70 120 55 male 64 112 58 female 61 112 32 female NA 179 63 male 60 214 44 female 60 NA 67 <na> 61 160 NA female 60 166 63 female 60 150 64 male 60 150 65 female 60 150 66 male 81 NA 67 female 76 120 68 female 73 114 68 female 70 100 69 37 female 72 107 60 male 92 151 60 male 92 151 60 male 92 151 61 NA male 63 120 62 MA male 63 120 63 MA MA 98 64 MA 98 65 MA MA 98 65 MA MA 98 66 MA MA 98 67 MA MA 98 68 MA 98 68 MA MA 98 68 MA MA 98 69 MA MA 98 60 MA MA 98 60 MA MA 98 61 MA MA 98 62 MA MA 98 63 MA MA MA 98 64 MA 98 65 MA MA 98 66 MA MA 98 67 MA MA 98 68 MA MA 98 68 MA MA 98 68 MA MA 98 69 MA MA 98 69 MA MA 98 60 MA MA 98 60 MA MA 98 61 MA MA 98 62 MA MA 98 63 MA MA 98 64 MA 98 65 MA MA 98 66 MA MA 98 67 MA MA 98 68 MA MA 98 69 MA MA 98 69 MA MA 98 60 MA MA 98 61 MA MA 98 62 MA MA 98 63 MA MA MA 98 64 MA MA 98 65 MA MA MA 98 66 MA MA 131 67 MA MA 131 68 MA MA 131</na>	age gender impluse 64         male 66         160         83           21         male 94         98         46           55         male 64         160         77           64         male 70         120         55           55         male 64         112         65           58         female 61         112         58           32         female 81         NA         179         68           63         male 60         214         82           44         female 60         NA         81           67 <na>         61         160         95           8         64         male 60         166         90           8         64         male 60         150         10           8         64         male 60         150         10           8         64         male 60         199         5           8         64         male 76         120         70           9         45         female 72         107         86           9         37         female 72         107         86           9         45         mal</na>	age         gender         impluse         pressurehight         pressurelow         glucose           64         male         66         160         83         160           21         male         94         98         46         296           55         male         64         160         77         270           64         male         70         120         55         270           55         male         64         112         65         300           58         female         61         112         58         87           32         female         61         112         58         87           32         female         60         214         82         87           44         female         60         NA         81         135           67 <na>         61         160         95         100           8         64         male         60         150         10         198           8         64         male         60         150         10         198           8         64         male         60         150</na>			

## **Description:**

The maximum range of human Impulse is 190 beasts per minutes. From the MidData dataset the noisy impulses of human's are replaced with null value which are out of the range.

#### **Removing missing values of Gender:**

#### Code:

37

38

40

56.13793

64.00000

34.00000

44 00000

male

male

male

male

60

65

96

94

MidData <- MidData[!is.na(MidData\$gender),]

> MidData <- MidData[!is.na(MidData\$gender),]</pre> > MidData age gender impluse pressurehight pressurelow glucose 64.00000 1 male 66 160 83 160 negative 2 21,00000 male 94 98 46 296 positive 55.00000 64 160 77 3 male 270 negative 70 55 64.00000 120 270 positive 4 male 5 55.00000 male 64 112 65 300 negative 6 58.00000 female 61 58 87 negative 112 32.00000 female 102 negative 7 40 179 68 8 63.00000 male 60 82 87 positive 214 9 44.00000 female 60 135 135 negative 81 11 56.13793 female 60 90 102 negative 166 12 63.00000 female 60 150 10 198 negative 13 64.00000 male 60 199 5 92 positive 54.00000 female 94 67 14 122 97 negative 15 47.00000 male 76 120 70 319 negative 61.00000 16 male 81 135 66 134 positive 17 86.00000 female 73 114 68 87 positive 18 70 45.00000 female 100 68 96 negative 72 19 37.00000 female 107 86 274 negative 20 45.00000 60 male 109 65 89 positive 21 60.00000 male 92 151 78 301 negative 22 48.00000 135 male 98 60 100 positive 227 positive 23 52.00000 male 76 109 85 24 30.00000 male 63 110 68 107 positive 25 56.13793 male 63 320 63 269 positive 26 72.00000 male 64 106 68 111 positive 42.00000 27 male 65 150 68 101 negative 28 64 72.00000 female 325 95 negative 60 47.00000 female 57 279 positive 29 66 134 30 63.00000 55 male 66 135 166 negative 31 54.00000 male 125 131 82 95 positive 32 35.00000 321 negative male 62 137 61 68.00000 33 male 61 121 49 98 positive 34 56.00000 female 62 105 negative 60 145 35 50.00000 136 positive male 61 136 70 64.00000 76 82 positive 36 male 58 156

**Description:**!is.na() function is used for removing missing values of specific column. The missing gender column may impact the nature of our dataset, so we removed gender from our MidData.

166

155

105

91

82

75

75

52

117 negative

107 negative

136 positive

208 negative

# Replace Missing values With Mean, Median, Mode

#### Replace missing values with mean value:

**Code:** Mean\_age<-mean(MidData\$age,na.rm = TRUE)

Mean\_age

MidData\$age[is.na(MidData\$age)]<-Mean\_age

MidData

- > Mean\_age<-mean(MidData\$age,na.rm = TRUE)</pre>
- > Mean\_age

[1] 56.09313

- > MidData\$age[is.na(MidData\$age)]<-Mean\_age</pre>
- > MidData

	age	gender	impluse	pressurehight	pressurelow	glucose	class
1	64.00000	male	66	160	83	160	negative
2	21.00000	male	94	98	46	296	positive
3	55.00000	male	64	160	77	270	negative
4	64.00000	male	70	120	55	270	positive
5	55.00000	male	64	112	65	300	negative
6	58.00000	female	61	112	58	87	negative
7	32.00000	female	40	179	68	102	negative
8	63.00000	male	60	214	82	87	positive
9	44.00000	female	60	NA	81	135	negative
11	56.13793	female	60	166	90	102	negative
12	63.00000	female	60	150	10	198	negative
13	64.00000	male	60	199	5	92	positive
14	54.00000	female	94	122	67	97	negative
15	47.00000	male	76	120	70	319	negative
16	61.00000	male	81	NA	66	134	positive
17	86.00000	female	73	114	68	87	positive
18	45.00000	female	70	100	68	96	negative
19	37.00000	female	72	107	86	274	negative
20	45.00000	male	60	109	65	89	positive
21	60.00000	male	92	151	78	301	negative
22	48.00000	male	135	98	60	100	positive
23	52.00000	male	76	109	85	227	positive
24	30.00000	male	63	110	68	107	positive
25	56.13793	male	63	320	63	269	positive
26	72.00000	male	64	106	68	111	positive
27	42.00000	male	65	150	68	101	negative
28	72.00000	female	64	325	60	95	negative
29	47.00000	female	66	134	57	279	positive
30	63.00000	male	66	135	55	166	negative
31	54.00000	male	125	131	82	95	positive
32	35.00000	male	62	137	61	321	negative
33	68.00000	male	61	121	49	98	positive
34	56.00000	female	60	145	62	105	negative
ъ		1 . 1 .1		1 6.1 " "	1 6 , 1	.1 1	1.1

**Description:** Calculated the mean value of the "age" column first, and then replaced the missing values with the calculated mean value. We don't want to consider missing values so we used na.rm=TRUE function

#### Replace missing values of Impulse using median method

**Code:** Median\_impulse<-median(MidData\$impluse,na.rm = TRUE)

Median\_impulse

MidData\$impluse[is.na(MidData\$impluse)]<-Median\_impulse

MidData

- > Median\_impulse<-median(MidData\$impluse, na.rm = TRUE)</pre>
- > Median\_impulse

[1] 74

- > MidData\$impluse[is.na(MidData\$impluse)]<-Median\_impulse</pre>
- > MidData

ubata						
age	_	impluse	-	pressurelow	glucose	class
64.00000		66		83		negative
				46	296	positive
55.00000	male	64	160	77	270	negative
64.00000		70	120	55	270	positive
55.00000	male	64	112	65	300	negative
58.00000	female	61	112	58	87	negative
32.00000	female	40	179	68	102	negative
63.00000	male	60	214	82	87	positive
44.00000	female	60	NA	81	135	negative
67.00000	<na></na>	61	160	95	100	negative
56.13793	female	60	166	90	102	negative
63.00000		60	150	10	198	negative
64.00000	male	60	199	5	92	positive
54.00000	female	94	122	67	97	negative
47.00000	male	76	120	70	319	negative
61.00000	male	81	NA	66	134	positive
86.00000	female	73	114	68	87	positive
45.00000	female	70	100	68	96	negative
37.00000	female	72	107	86	274	negative
45.00000	male	60	109	65	89	positive
60.00000	male	92	151	78	301	negative
48.00000	male	135	98	60	100	positive
52.00000	male	76	109	85	227	positive
30.00000	male	63	110	68	107	positive
56.13793	male	63	320	63	269	positive
72.00000	male	64	106	68	111	positive
42.00000	male	65	150	68	101	negative
72.00000	female	64	325	60	95	negative
47.00000	female	66	134	57	279	positive
63.00000	male	66	135	55	166	negative
54.00000	male	125	131	82	95	positive
35.00000	male	62	137	61	321	negative
68 00000	mala	61	1 71	40	ΩQ	nocitivo
	age 64.00000 21.00000 55.00000 64.00000 55.00000 32.00000 63.00000 44.00000 67.00000 56.13793 63.00000 47.00000 61.00000 45.00000 45.00000 45.00000 45.00000 45.00000 45.00000 45.00000 47.00000 48.00000 56.13793 72.00000 47.00000 47.00000 56.13793 72.00000 47.00000 56.13793 72.00000 47.00000 56.13793 72.00000 47.00000 56.13793	age gender 64.00000 male 21.00000 male 55.00000 male 64.00000 male 55.00000 male 55.00000 male 63.00000 female 63.00000 female 67.00000 solution 64.00000 male 64.00000 male 64.00000 male 64.00000 male 64.00000 male 65.00000 female 45.00000 female 45.00000 male 37.00000 male 60.00000 male	age gender impluse 64.00000 male 66 21.00000 male 94 55.00000 male 64 64.00000 male 70 55.00000 male 64 58.00000 female 61 32.00000 female 60 44.00000 female 60 67.00000 NA> 61 56.13793 female 60 64.00000 female 60 64.00000 male 60 54.00000 female 76 61.00000 male 76 61.00000 male 76 61.00000 female 73 45.00000 female 70 37.00000 female 72 45.00000 male 60 60.00000 male 60 61.3793 male 63 52.00000 male 63 56.13793 male 63 72.00000 male 64 42.00000 male 65 72.00000 female 64 47.00000 female 66 63.00000 male 66 63.00000 male 66 63.00000 male 66 63.00000 male 66	age gender impluse pressurehight 64.00000 male 66 160 21.00000 male 94 98 55.00000 male 64 160 64.00000 male 70 120 55.00000 male 64 112 58.00000 female 61 112 32.00000 female 60 214 44.00000 female 60 NA 67.00000 <na> 61 160 63.00000 female 60 166 63.00000 female 60 150 64.00000 female 60 150 64.00000 male 60 199 54.00000 male 60 199 54.00000 female 76 120 61.00000 male 76 120 61.00000 male 76 120 61.00000 female 70 100 37.00000 female 72 107 45.00000 female 72 107 45.00000 male 76 109 60.00000 male 60 109 60.00000 male 63 110 56.13793 male 63 320 72.00000 male 64 106 42.00000 male 64 106 42.00000 male 65 150 72.00000 female 66 134 63.00000 male 66 135 54.00000 male 66 135</na>	age gender         impluse pressurehight         pressurelow           64.00000         male         66         160         83           21.00000         male         94         98         46           55.00000         male         64         160         77           64.00000         male         70         120         55           55.00000         male         64         112         65           58.00000         female         61         112         58           32.00000         female         60         179         68           63.00000         female         60         214         82           44.00000         female         60         NA         81           67.00000         female         60         NA         81           67.00000         female         60         166         90           63.00000         female         60         150         10           64.00000         female         60         199         5           54.00000         female         76         120         70           61.00000         female         70         100         68	age gender         impluse         pressurehight         pressurelow         glucose           64.00000         male         66         160         83         160           21.00000         male         94         98         46         296           55.00000         male         64         160         77         270           64.0000         male         64         112         65         300           58.0000         female         61         112         58         87           32.00000         female         60         214         82         87           44.0000         female         60         NA         81         135           67.00000         female         60         NA         81         135           67.00000         female         60         NA         81         135           67.00000         female         60         166         90         102           63.00000         female         60         150         10         198           64.00000         male         60         199         5         92           54.00000         male         81         NA

**Description:** Calculated the median value of the "impulse" column first, and then replaced the missing values with the calculated median value. We don't want to consider missing values so we used na.rm=TRUE function

#### Replace missing values of Pressurehight using mode method

```
Code:
custom_mode <- function(x) {</pre>
 ux <- unique(x)
 ux[which.max(tabulate(match(x, ux)))]
Mode pressurehight <-custom mode(MidData$pressurehight)
Mode pressurehight
Midata$pressurehight[is.na(MidData$pressurehight)]<-Mode_pressurehight
MidData
> Mode_pressurehight <-custom_mode(MidData$pressurehight)</pre>
> Mode_pressurehight
> MidData$pressurehight[is.na(MidData$pressurehight)]<-Mode_pressurehight</p>
> MidData
           age gender impluse pressurehight pressurelow glucose
1
      64.00000
                  male
                             66
                                           160
                                                          83
                                                                 160 negative
2
                             94
      21.00000
                  male
                                            98
                                                          46
                                                                 296 positive
                             64
                                                          77
3
      55.00000
                  male
                                           160
                                                                 270 negative
                             70
                                                          55
4
      64.00000
                  male
                                           120
                                                                 270 positive
5
                             64
                                                          65
      55.00000
                  male
                                           112
                                                                 300 negative
6
      58.00000 female
                             61
                                           112
                                                          58
                                                                  87 negative
      32.00000 female
                             40
                                           179
                                                          68
                                                                 102 negative
8
      63.00000
                  male
                             60
                                           214
                                                          82
                                                                  87 positive
      44.00000 female
                             60
                                           135
                                                          81
                                                                 135 negative
10
      67.00000
                  <NA>
                             61
                                           160
                                                          95
                                                                 100 negative
                                                          90
11
      56.13793 female
                             60
                                           166
                                                                 102 negative
12
     63.00000 female
                             60
                                           150
                                                          10
                                                                 198 negative
13
     64.00000
                             60
                                           199
                                                          5
                                                                  92 positive
                  male
                             94
14
      54.00000 female
                                           122
                                                          67
                                                                  97 negative
15
      47.00000
                  male
                             76
                                           120
                                                          70
                                                                 319 negative
16
     61.00000
                  male
                             81
                                           135
                                                          66
                                                                 134 positive
                             73
17
      86.00000 female
                                           114
                                                          68
                                                                  87 positive
18
      45.00000 female
                             70
                                           100
                                                          68
                                                                  96 negative
      37.00000 female
19
                             72
                                                                 274 negative
                                           107
                                                          86
20
     45.00000
                  male
                             60
                                           109
                                                          65
                                                                  89 positive
21
     60.00000
                  male
                             92
                                                          78
                                           151
                                                                 301 negative
22
     48.00000
                  male
                            135
                                            98
                                                          60
                                                                 100 positive
23
      52.00000
                  male
                             76
                                           109
                                                          85
                                                                 227 positive
24
      30.00000
                  male
                             63
                                           110
                                                          68
                                                                 107 positive
25
                             63
      56.13793
                  male
                                           320
                                                          63
                                                                 269 positive
26
     72.00000
                  male
                             64
                                           106
                                                          68
                                                                 111 positive
27
     42.00000
                  male
                             65
                                           150
                                                          68
                                                                 101 negative
28
      72.00000 female
                             64
                                           325
                                                          60
                                                                  95 negative
                                                          57
29
      47.00000 female
                             66
                                           134
                                                                 279 positive
30
     63.00000
                                                          55
                  male
                             66
                                           135
                                                                 166 negative
```

**Description:** Calculated the mode value of the "pressurehight " column first, and then replaced the missing values with the calculated mode value. We don't get mode function by default so we used Custom\_mode function to calculate mode & replace it.

131

82

95 positive

31

54.00000

male

125

# **Data Annotation:**

#### **Annotating Gender Column:**

#### Code:

MidData\$gender < -factor(MidData\$gender, levels = c("male", "female"), labels = c(1,2))

#### MidData

```
> MidData$gender<-factor(MidData$gender,levels = c("male","female"),labels = c(1,2))
> MidData
           age gender impluse pressurehight pressurelow glucose
                                                                         class
1
     64.00000
                                           160
                     1
                                                                 160 negative
2
     21.00000
                     1
                            94
                                            98
                                                         46
                                                                 296 positive
3
     55.00000
                     1
                            64
                                           160
                                                         77
                                                                 270 negative
                     1
                            70
4
     64.00000
                                           120
                                                         55
                                                                 270 positive
5
                     1
                                                         65
     55.00000
                            64
                                           112
                                                                 300 negative
                     2
6
     58.00000
                            61
                                                         58
                                           112
                                                                  87 negative
                     2
7
     32.00000
                            40
                                           179
                                                         68
                                                                 102 negative
     63.00000
                     1
                            60
8
                                           214
                                                         82
                                                                  87 positive
9
     44.00000
                     2
                            60
                                           135
                                                         81
                                                                 135 negative
                     2
11
     56.13793
                            60
                                           166
                                                         90
                                                                 102 negative
12
     63.00000
                     2
                            60
                                           150
                                                         10
                                                                 198 negative
13
                     1
                                           199
     64.00000
                             60
                                                          5
                                                                  92 positive
                     2
14
                            94
                                           122
                                                         67
                                                                  97 negative
     54.00000
15
     47.00000
                     1
                            76
                                           120
                                                         70
                                                                 319 negative
16
     61.00000
                     1
                             81
                                           135
                                                         66
                                                                 134 positive
                     2
17
     86.00000
                             73
                                           114
                                                         68
                                                                  87 positive
                     2
18
                             70
     45.00000
                                           100
                                                         68
                                                                  96 negative
                     2
19
                            72
     37.00000
                                           107
                                                         86
                                                                 274 negative
                     1
20
     45.00000
                            60
                                           109
                                                         65
                                                                  89 positive
21
     60.00000
                     1
                            92
                                           151
                                                         78
                                                                 301 negative
22
     48.00000
                     1
                           135
                                                         60
                                            98
                                                                 100 positive
                            76
23
     52.00000
                     1
                                           109
                                                         85
                                                                 227 positive
24
     30.00000
                     1
                            63
                                           110
                                                         68
                                                                 107 positive
25
     56.13793
                     1
                            63
                                           320
                                                         63
                                                                 269 positive
26
     72.00000
                     1
                            64
                                           106
                                                         68
                                                                 111 positive
27
     42.00000
                     1
                            65
                                           150
                                                         68
                                                                 101 negative
                     2
28
     72.00000
                            64
                                           325
                                                         60
                                                                  95 negative
                     2
29
     47.00000
                            66
                                           134
                                                         57
                                                                 279 positive
                     1
30
     63.00000
                            66
                                           135
                                                         55
                                                                 166 negative
31
     54.00000
                     1
                           125
                                           131
                                                         82
                                                                  95 positive
32
     35.00000
                     1
                            62
                                           137
                                                         61
                                                                 321 negative
33
     68.00000
                     1
                            61
                                                         49
                                           121
                                                                  98 positive
                     2
34
     56.00000
                            60
                                                         62
                                           145
                                                                 105 negative
35
     50.00000
                     1
                            61
                                           136
                                                         70
                                                                 136 positive
```

**Description:** This method is more helpful to represent data in more meaningful ways. Here for Gender column has 1 & 2 values, which considered to represent "male" & "female" respectably. We replaced "male" value with 1 & "female" value with 2.

#### **Annotating Class Column:**

#### Code:

 $MidData\class < -factor(MidData\class, levels = c("positive", "negative"), labels = c(1,2))$ 

#### MidData

> MidData\$class<-factor(MidData\$class,levels = c("positive", "negative"), labels = c(1,2)) > MidData age gender impluse pressurehight pressurelow glucose class 64.00000 21.00000 55.00000 64.00000 55.00000 58.00000 32.00000 63.00000 44.00000 56.13793 63.00000 64.00000 54.00000 47.00000 61.00000 86.00000 45.00000 37.00000 45.00000 60.00000 48.00000 52.00000 30.00000 56.13793 72.00000 42.00000 72.00000 47.00000 63.00000 54.00000 35.00000 68.00000 56.00000 50.00000 

**Description:** This method is more helpful to represent data in more meaningful ways. Here for Class column has 1 & 2 values, which considered to represent "positive" & "negative" respectably. We replaced "positive" value with 1 & "negative" value with 2.

# **Histogram & BoxPlot:**

# Historgeam for all numeric columns:

## **Code:**

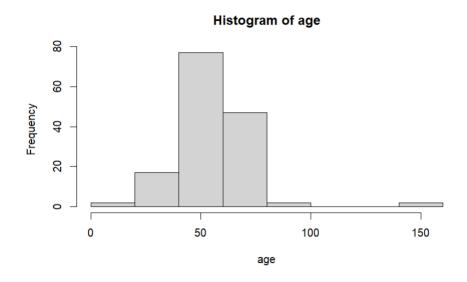
hist(age)

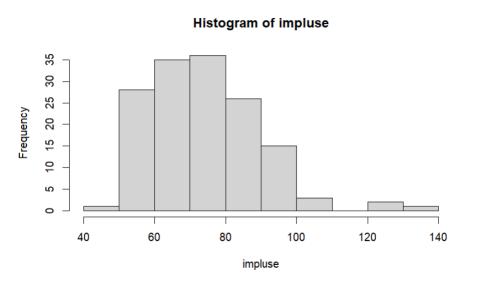
hist(impluse)

hist(pressurehight)

hist(pressurelow)

hist(gluecose)





#### Histogram of pressurehight Frequency pressurehight Histogram of pressurelow pressurelow Histogram of glucose Frequency

**Description:** The code hist() generates a histogram of the values in the columns, allowing us to see how the data is distributed.

glucose

#### **Boxplot of MidData:**

#### **Code:**

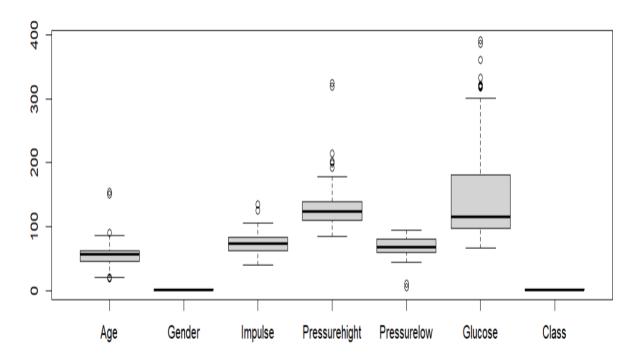
boxplot(MidData\$age, MidData\$gender, MidData\$impluse, MidData\$pressurehight,

MidData\$pressurelow,MidData\$glucose, MidData\$class,

main="Box Plots for MidData",

names=c("Age", "Gender", "Impulse", "Pressurehight", "Pressurelow", "Glucose", "Class"))

## **Box Plots for MidData**



**Description:** This technic is used for identify data points that are fall outside of the range of the most of the values. **boxplot** function is used for plotting boxes, **main** is used for providing name, **names**= is used for providing name each boxes.