American International University-Bangladesh (AIUB)

Department of Computer Science

Faculty of Science & Technology (FST)

Group: 8

PROJECT TITLE

INTRODUCTION TO DATA SCIENCE MIDTERM PROJECT

Semester: Spring_24_25		**Section: F**
SN	Student Name	Student ID
01	AHBAB SAKALAN	22-48108-2
02	MOHAMMED SAIFUL ISLAM	22-48091-2
03	ABDULLAH AL MARUF	22-47997-2
04	NIDAN ALAM	22-47046-1

Dataset Overview:

This database called Diabetes Prediction provides expanded medical characteristics with demographic factors from individual patient records. The digital records contain seven features including age together with gender and BMI and hypertension and heart disease and smoking history and HbA1c level and blood glucose level. The diagnosis of diabetes constitutes a positive target variable outcome while a lack of diagnosis scores as negative. The Diabetes Prediction dataset serves as an appropriate foundation to train predictive models which estimate diabetes potential through patient characteristics.

The identified dataset enables healthcare professionals to detect diabetes risk patterns which enables them to establish preventive measures for susceptible groups. The system enables individualized treatment preparation by showing important risk indicators. Academic researchers benefit from this dataset because it reveals how different health and lifestyle elements affect diabetes formation. The information base enables predictive models along with providing detailed analysis about how demographic information and medical conditions relate to diabetes diagnosis.

Data Exploration

```
AllData <- read.csv("/Users/islam/Downloads/Data_Science/Data science LAb/Dataset(Updated)_MIdterm_sectoin(F).csv", header = TRUE, sep = ",")
AllData


head(AllData)
tail(AllData)
summary(AllData)
sapply(AllData, function(x) sum(is.na(x)))
```

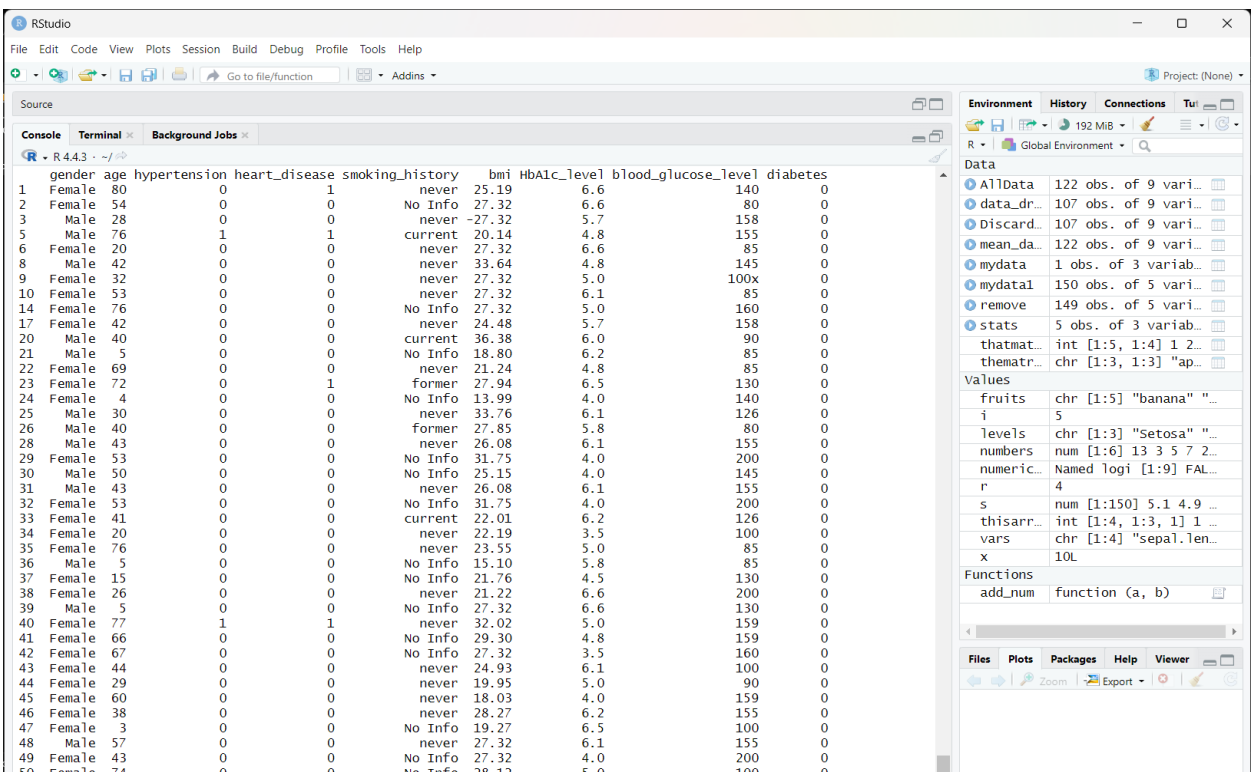
```
> AllData
    gender age hypertension heart_disease smoking_history    bmi HbA1c_level blood_glucose_level diabetes
1   Female  80            0             1           never  25.19         6.6                 140        0
2   Female  54            0             0         No Info  27.32         6.6                  80        0
3     Male  28            0             0           never -27.32         5.7                 158        0
4   Female  NA            0             0         current  23.45         5.0                 155        0
5     Male  76            1             1         current  20.14         4.8                 155        0
6   Female  20            0             0           never  27.32         6.6                  85        0
7     <NA>  79            0             0         No Info  23.86         5.7                  85        0
8     Male  42            0             0           never  33.64         4.8                 145        0
9   Female  32            0             0           never  27.32         5.0                 100        0
10  Female  53            0             0           never  27.32         6.1                  85        0
11  Female  54            0             0          former     NA         6.0                 100        0
12  Female  78           NA             0          former  36.05         5.0                 130        0
13  Female  67            0             0            <NA>  25.69         5.8                 200        0
14  Female  76            0             0         No Info  27.32         5.0                 160        0
15    <NA>  78            0             0         No Info  27.32         6.6                 126        0
16    Male  15            0             0           never     NA         6.1                 200        0
17  Female  42            0             0           never  24.48         5.7                 158        0
18  Female  42            0             0            <NA>  27.32         5.7                  80        0
19    Male  NA            0             0            ever  25.72         3.5                 159        0
20    Male  40            0             0         current  36.38         6.0                  90        0
21    Male   5            0             0         No Info  18.80         6.2                  85        0
22  Female  69            0             0           never  21.24         4.8                  85        0
23  Female  72            0             1          former  27.94         6.5                 130        0
24  Female   4            0             0         No Info  13.99         4.0                 140        0
25    Male  30            0             0           never  33.76         6.1                 126        0
26    Male  40            0             0          former  27.85         5.8                  80        0
```

```
> summary(AllData)
    gender              age          hypertension     heart_disease     smoking_history        bmi          HbA1c_level    blood_glucose_level    diabetes
 Length:122         Min.   :-33.00   Min.   :0.00000   Min.   :0.00000   Length:122         Min.   :-27.32   Min.   :3.500   Min.   : 80.0       Min.   :0.0000
 Class :character   1st Qu.: 38.50   1st Qu.:0.00000   1st Qu.:0.00000   Class :character   1st Qu.: 24.73   1st Qu.:5.700   1st Qu.:130.0       1st Qu.:0.0000
 Mode  :character   Median : 52.00   Median :0.00000   Median :0.00000   Mode  :character   Median : 27.32   Median :6.150   Median :155.0       Median :0.0000
                    Mean   : 53.19   Mean   :0.08333   Mean   :0.06557                      Mean   : 27.43   Mean   :6.255   Mean   :157.1       Mean   :0.4262
                    3rd Qu.: 67.75   3rd Qu.:0.00000   3rd Qu.:0.00000                      3rd Qu.: 29.26   3rd Qu.:6.600   3rd Qu.:160.0       3rd Qu.:1.0000
                    Max.   :290.00   Max.   :1.00000   Max.   :1.00000                      Max.   : 63.48   Max.   :9.000   Max.   :300.0       Max.   :1.0000
                    NA's   :4        NA's   :2                                              NA's   :2
```

Through summary function can show minimum value, mean, median, maximum value of a column. Through this operation can take a overview of data set.

Finding and Handling Missing Values

```r
AllData$gender[AllData$gender == ""] <- NA
AllData$smoking_history[AllData$smoking_history == ""] <- NA
AllData$gender
AllData$smoking_history


data_drop <- na.omit(AllData)
write.csv(data_drop, "/Users/islam/Downloads/Data_Science/Data science LAb/Discard_instatnse.csv", row.names = FALSE)
data_drop
```



First replace empty rows by NA in categorical columns such as gender, smoking history. then where NA found drop the row (**discard instances**)

Handling invalid value:

```
AllData$blood_glucose_level <- parse_number(as.character(AllData$blood_glucose_level))
AllData
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Source

Console Terminal Background Jobs

R 4.4.3 · ~/

```
> AllData$blood_glucose_level <- parse_number(as.character(AllData$blood_glucose_level))
> AllData
    gender age hypertension heart_disease smoking_history    bmi HbA1c_level blood_glucose_level diabetes
1   Female  80            0             1          never  25.19         6.6                 140        0
2   Female  54            0             0         No Info  27.32         6.6                  80        0
3     Male  28            0             0          never -27.32         5.7                 158        0
4   Female  NA            0             0        current  23.45         5.0                 155        0
5     Male  76            1             1        current  20.14         4.8                 155        0
6   Female  20            0             0          never  27.32         6.6                  85        0
7     <NA>  79            0             0         No Info  23.86         5.7                  85        0
8     Male  42            0             0          never  33.64         4.8                 145        0
9   Female  32            0             0          never  27.32         5.0                 100        0
10  Female  53            0             0          never  27.32         6.1                  85        0
11  Female  54            0             0         former     NA         6.0                 100        0
12  Female  78           NA             0         former  36.05         5.0                 130        0
13  Female  67            0             0           <NA>  25.69         5.8                 200        0
14  Female  76            0             0         No Info  27.32         5.0                 160        0
15    <NA>  78            0             0         No Info  27.32         6.6                 126        0
16    Male  15            0             0          never     NA         6.1                 200        0
17  Female  42            0             0          never  24.48         5.7                 158        0
18  Female  42            0             0           <NA>  27.32         5.7                  80        0
19    Male  NA            0             0           ever  25.72         3.5                 159        0
20    Male  40            0             0        current  36.38         6.0                  90        0
21    Male   5            0             0         No Info  18.80         6.2                  85        0
22  Female  69            0             0          never  21.24         4.8                  85        0
23  Female  72            0             1         former  27.94         6.5                 130        0
24  Female   4            0             0         No Info  13.99         4.0                 140        0
25    Male  30            0             0          never  33.76         6.1                 126        0
26    Male  40            0             0         former  27.85         5.8                  80        0
27    Male -33           NA             0          never  26.47         4.0                 158        0
28    Male  43            0             0          never  26.08         6.1                 155        0
29  Female  53            0             0         No Info  31.75         4.0                 200        0
30    Male  50            0             0         No Info  25.15         4.0                 145        0
```

Fix invalid value like 100x. convert string value into integer using parse number function

```r
AllData$blood_glucose_level <- parse_number(as.character(AllData$blood_glucose_level))
AllData
unique_values <- unique(AllData$gender)
unique_values
AllData$gender[AllData$gender %in% c("Malee")] <- "Male"
AllData$gender[AllData$gender %in% c("Femalee")] <- "Female"
AllData
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Source

Console Terminal Background Jobs

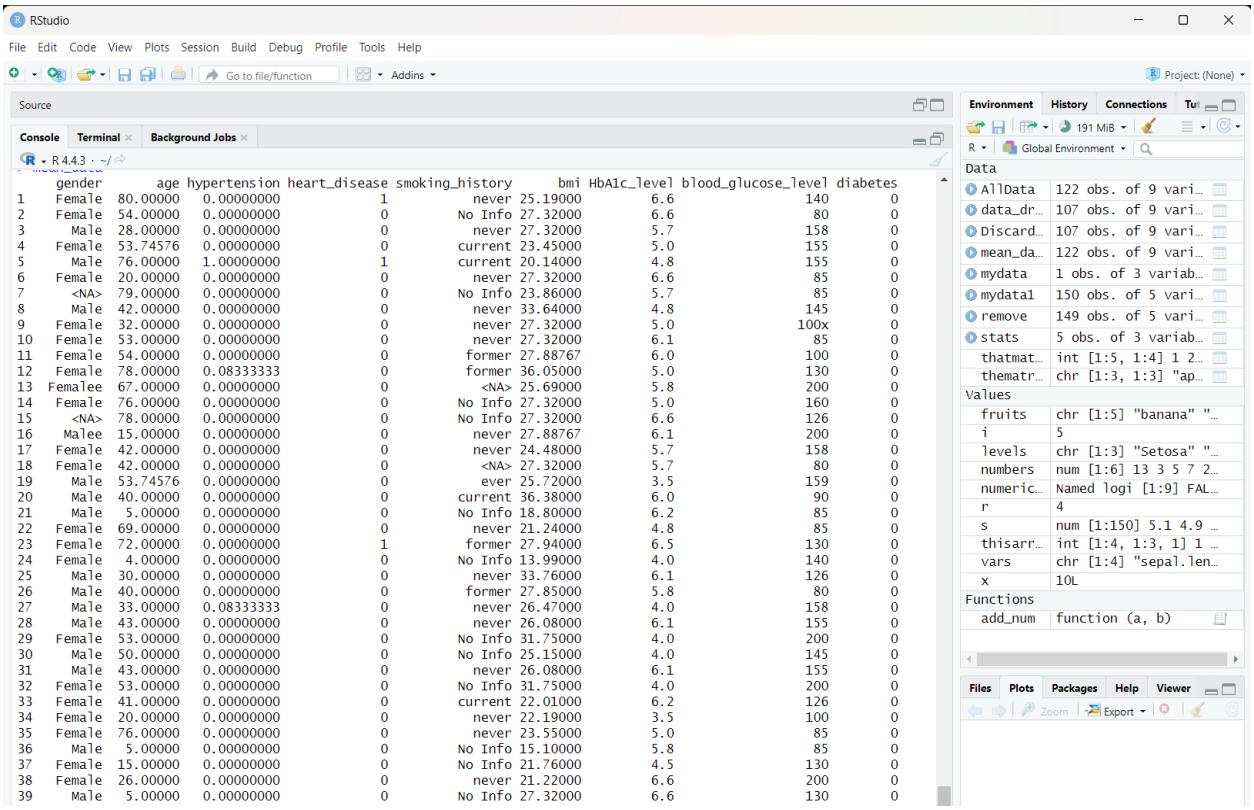
R 4.4.3 · ~/

```
> unique_values <- unique(AllData$gender)
> unique_values
[1] "Female" "Male"   NA
> AllData$gender[AllData$gender %in% c("Malee")] <- "Male"
> AllData$gender[AllData$gender %in% c("Femalee")] <- "Female"
> AllData
    gender age hypertension heart_disease smoking_history    bmi HbA1c_level blood_glucose_level diabetes
1   Female  80            0             1           never  25.19         6.6                 140        0
2   Female  54            0             0         No Info  27.32         6.6                  80        0
3     Male  28            0             0           never -27.32         5.7                 158        0
4   Female  NA            0             0         current  23.45         5.0                 155        0
5     Male  76            1             1         current  20.14         4.8                 155        0
6   Female  20            0             0           never  27.32         6.6                  85        0
7     <NA>  79            0             0         No Info  23.86         5.7                  85        0
8     Male  42            0             0           never  33.64         4.8                 145        0
9   Female  32            0             0           never  27.32         5.0                 100        0
10  Female  53            0             0           never  27.32         6.1                  85        0
11  Female  54            0             0          former     NA         6.0                 100        0
12  Female  78           NA             0          former  36.05         5.0                 130        0
13  Female  67            0             0            <NA>  25.69         5.8                 200        0
14  Female  76            0             0         No Info  27.32         5.0                 160        0
15    <NA>  78            0             0         No Info  27.32         6.6                 126        0
16    Male  15            0             0           never     NA         6.1                 200        0
17  Female  42            0             0           never  24.48         5.7                 158        0
18  Female  42            0             0            <NA>  27.32         5.7                  80        0
19    Male  NA            0             0            ever  25.72         3.5                 159        0
20    Male  40            0             0         current  36.38         6.0                  90        0
21    Male   5            0             0         No Info  18.80         6.2                  85        0
22  Female  69            0             0           never  21.24         4.8                  85        0
23  Female  72            0             1          former  27.94         6.5                 130        0
24  Female   4            0             0         No Info  13.99         4.0                 140        0
25    Male  30            0             0           never  33.76         6.1                 126        0
26    Male  40            0             0          former  27.85         5.8                  80        0
27    Male -33           NA             0           never  26.47         4.0                 158        0
28    Male  43            0             0           never  26.08         6.1                 155        0
29  Female  53            0             0         No Info  31.75         4.0                 200        0
30    Male  50            0             0         No Info  25.15         4.0                 145        0
31    Male  43            0             0           never  26.08         6.1                 155        0
32  Female  53            0             0         No Info  31.75         4.0                 200        0
33  Female  41            0             0         current  22.01         6.2                 126        0
```

Detect unique value then fix invalid value such as femalee to female and malee to male.

Replace by Average Value:

```
numeric_columns <- sapply(AllData,is.numeric)

mean_data <- AllData
mean_data[numeric_columns] <- lapply(AllData[numeric_columns], abs)
mean_data[numeric_columns] <-lapply(mean_data[numeric_columns], function(x) ifelse(is.na(x),mean(x,na.rm=TRUE),x))
mean_data
write.csv(mean_data,"/Users/islam/Downloads/Data_Science/Data science LAb/replace_by_mean.csv",row.names = FALSE)
```

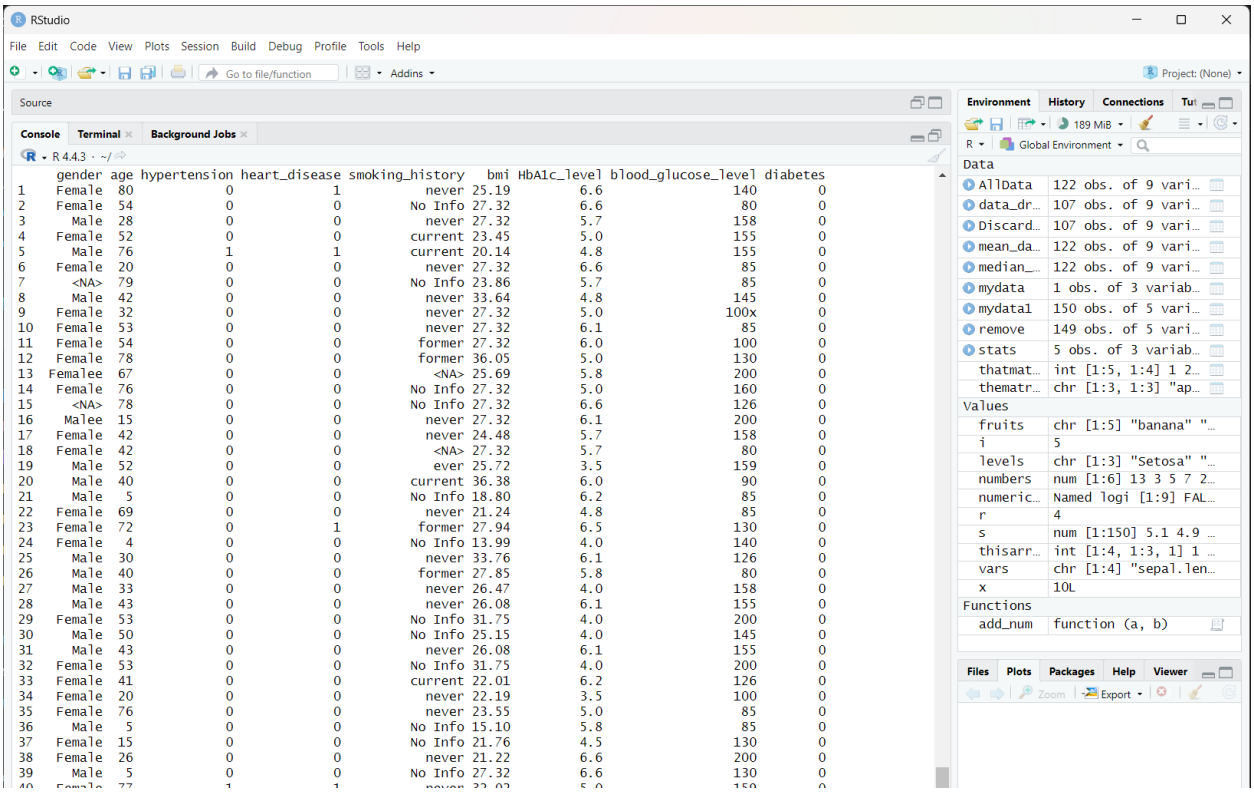


	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
1	Female	80.00000	0.00000000	1	never	25.19000	6.6	140	0
2	Female	54.00000	0.00000000	0	No Info	27.32000	6.6	80	0
3	Male	28.00000	0.00000000	0	never	27.32000	5.7	158	0
4	Female	53.74576	0.00000000	0	current	23.45000	5.0	155	0
5	Male	76.00000	1.00000000	1	current	20.14000	4.8	155	0
6	Female	20.00000	0.00000000	0	never	27.32000	6.6	85	0
7	<NA>	79.00000	0.00000000	0	No Info	23.86000	5.7	85	0
8	Male	42.00000	0.00000000	0	never	33.64000	4.8	145	0
9	Female	32.00000	0.00000000	0	never	27.32000	5.0	100x	0
10	Female	53.00000	0.00000000	0	never	27.32000	6.1	85	0
11	Female	54.00000	0.00000000	0	former	27.88767	6.0	100	0
12	Female	78.00000	0.08333333	0	former	36.05000	5.0	130	0
13	Femalee	67.00000	0.00000000	0	<NA>	25.69000	5.8	200	0
14	Female	76.00000	0.00000000	0	No Info	27.32000	5.0	160	0
15	<NA>	78.00000	0.00000000	0	No Info	27.32000	6.6	126	0
16	Malee	15.00000	0.00000000	0	never	27.88767	6.1	200	0
17	Female	42.00000	0.00000000	0	never	24.48000	5.7	158	0
18	Female	42.00000	0.00000000	0	<NA>	27.32000	5.7	80	0
19	Male	53.74576	0.00000000	0	ever	25.72000	3.5	159	0
20	Male	40.00000	0.00000000	0	current	36.38000	6.0	90	0
21	Male	5.00000	0.00000000	0	No Info	18.80000	6.2	85	0
22	Female	69.00000	0.00000000	0	never	21.24000	4.8	85	0
23	Female	72.00000	0.00000000	1	former	27.94000	6.5	130	0
24	Female	4.00000	0.00000000	0	No Info	13.99000	4.0	140	0
25	Male	30.00000	0.00000000	0	never	33.76000	6.1	126	0
26	Male	40.00000	0.00000000	0	former	27.85000	5.8	80	0
27	Male	33.00000	0.08333333	0	never	26.47000	4.0	158	0
28	Male	43.00000	0.00000000	0	never	26.08000	6.1	155	0
29	Female	53.00000	0.00000000	0	No Info	31.75000	4.0	200	0
30	Male	50.00000	0.00000000	0	No Info	25.15000	4.0	145	0
31	Male	43.00000	0.00000000	0	never	26.08000	6.1	155	0
32	Female	53.00000	0.00000000	0	No Info	31.75000	4.0	200	0
33	Female	41.00000	0.00000000	0	current	22.01000	6.2	126	0
34	Female	20.00000	0.00000000	0	never	22.19000	3.5	100	0
35	Female	76.00000	0.00000000	0	never	23.55000	5.0	85	0
36	Male	5.00000	0.00000000	0	No Info	15.10000	5.8	85	0
37	Female	15.00000	0.00000000	0	No Info	21.76000	4.5	130	0
38	Female	26.00000	0.00000000	0	never	21.22000	6.6	200	0
39	Male	5.00000	0.00000000	0	No Info	27.32000	6.6	130	0

Find numerical columns then negative values convert to positive by abs() function. Then where NA is found replace by the mean value. Mean calculated by mean function.(**Average Value)**

Replace by median:

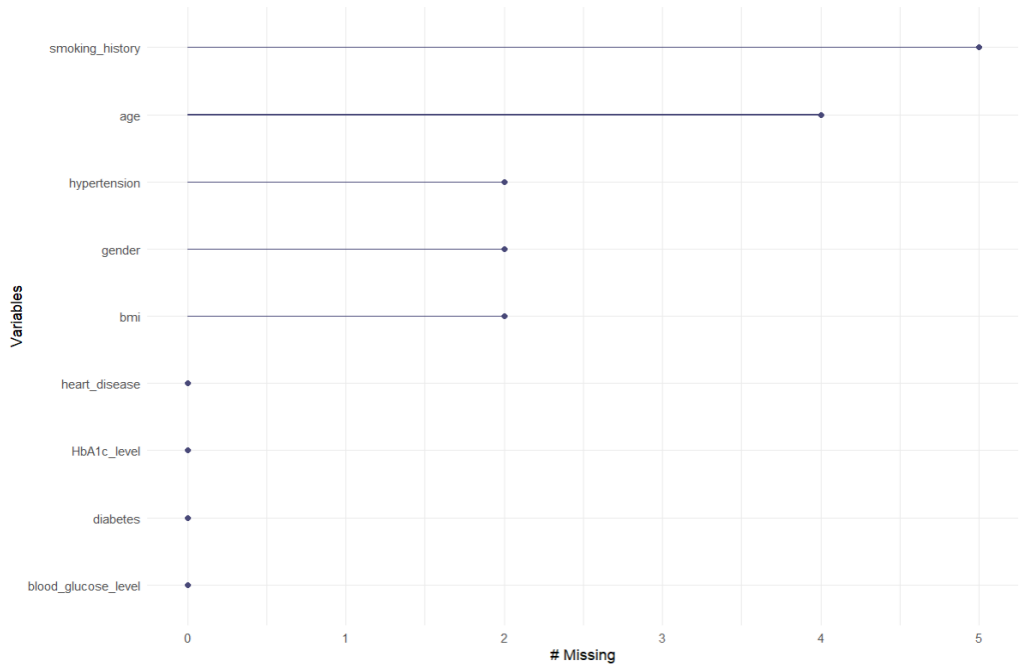
```
median_data <- AllData
median_data[numeric_columns] <- lapply(AllData[numeric_columns], abs)
median_data[numeric_columns] <-lapply(median_data[numeric_columns], function(x) ifelse(is.na(x),median(x,na.rm=TRUE),x))
median_data
write.csv(median_data,"/Users/islam/Downloads/Data_Science/Data science LAb/replace_by_median.csv",row.names = FALSE)
```



In numerical columns negative values convert to positive by abs function. Then where NA is found replace by the median value. median calculated by median function.

missing values graph

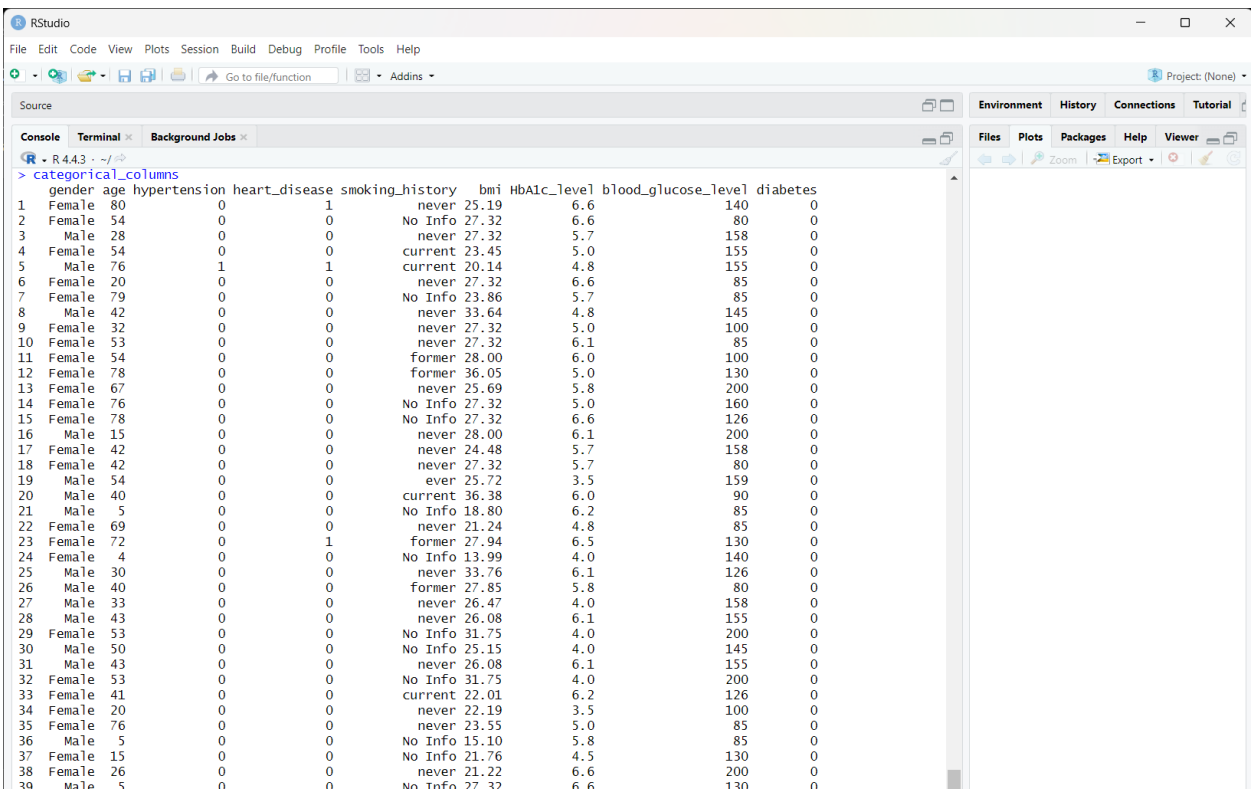
```
gg_miss_var(AllData)
```



missing values on a graph

Replace by Most Frequent/Average Value

```
mean_data_age <- AllData
mean_data_age$age <- abs(mean_data_age$age)
age_mean <- round(mean(mean_data_age$age, na.rm = TRUE))
mean_data_age$age[is.na(mean_data_age$age)] <- age_mean
mean_data_age



mean_data_bmi <- mean_data_age
mean_data_bmi$bmi <- abs(mean_data_bmi$bmi)
bmi_mean <- round(mean(mean_data_bmi$bmi, na.rm = TRUE))
mean_data_bmi$bmi[is.na(mean_data_bmi$bmi)] <- bmi_mean
mean_data_bmi


categorical_columns <- mean_data_bmi
gender_frequency <- table(mean_data_bmi$gender)
most_frequent_gender <- names(gender_frequency)[which.max(gender_frequency)]
categorical_columns$gender[is.na(mean_data_bmi$gender)] <- most_frequent_gender
categorical_columns


smoking_history_frequency <- table(mean_data_bmi$smoking_history)
most_frequent_smoking_history <- names(smoking_history_frequency)[which.max(smoking_history_frequency)]
categorical_columns$smoking_history[is.na(mean_data_bmi$smoking_history)] <- most_frequent_smoking_history
categorical_columns


hypertension_frequency <- table(mean_data_bmi$hypertension)
most_frequent_hypertension <- names(hypertension_frequency)[which.max(hypertension_frequency)]
categorical_columns$hypertension[is.na(mean_data_bmi$hypertension)] <- most_frequent_hypertension
categorical_columns
write.csv(median_data,"/Users/islam/Downloads/Data_Science/Data science LAb/replace_for_categorical.csv",row.names = FALSE)
```



```
> categorical_columns
    gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
1   Female  80            0            1           never 25.19         6.6                 140        0
2   Female  54            0            0         No Info 27.32         6.6                  80        0
3     Male  28            0            0           never 27.32         5.7                 158        0
4   Female  54            0            0         current 23.45         5.0                 155        0
5     Male  76            1            1         current 20.14         4.8                 155        0
6   Female  20            0            0           never 27.32         6.6                  85        0
7   Female  79            0            0         No Info 23.86         5.7                  85        0
8     Male  42            0            0           never 33.64         4.8                 145        0
9   Female  32            0            0           never 27.32         5.0                 100        0
10  Female  53            0            0           never 27.32         6.1                  85        0
11  Female  54            0            0          former 28.00         6.0                 100        0
12  Female  78            0            0          former 36.05         5.0                 130        0
13  Female  67            0            0           never 25.69         5.8                 200        0
14  Female  76            0            0         No Info 27.32         5.0                 160        0
15  Female  78            0            0         No Info 27.32         6.6                 126        0
16    Male  15            0            0           never 28.00         6.1                 200        0
17  Female  42            0            0           never 24.48         5.7                 158        0
18  Female  42            0            0           never 27.32         5.7                  80        0
19    Male  54            0            0            ever 25.72         3.5                 159        0
20    Male  40            0            0         current 36.38         6.0                  90        0
21    Male   5            0            0         No Info 18.80         6.2                  85        0
22  Female  69            0            0           never 21.24         4.8                  85        0
23  Female  72            0            1          former 27.94         6.5                 130        0
24  Female   4            0            0         No Info 13.99         4.0                 140        0
25    Male  30            0            0           never 33.76         6.1                 126        0
26    Male  40            0            0          former 27.85         5.8                  80        0
27    Male  33            0            0           never 26.47         4.0                 158        0
28    Male  43            0            0           never 26.08         6.1                 155        0
29  Female  53            0            0         No Info 31.75         4.0                 200        0
30    Male  50            0            0         No Info 25.15         4.0                 145        0
31    Male  43            0            0           never 26.08         6.1                 155        0
32  Female  53            0            0         No Info 31.75         4.0                 200        0
33  Female  41            0            0         current 22.01         6.2                 126        0
34  Female  20            0            0           never 22.19         3.5                 100        0
35  Female  76            0            0           never 23.55         5.0                  85        0
36    Male   5            0            0         No Info 15.10         5.8                  85        0
37  Female  15            0            0         No Info 21.76         4.5                 130        0
38  Female  26            0            0           never 21.22         6.6                 200        0
39    Male   5            0            0         No Info 27.32         6.6                 130        0
```

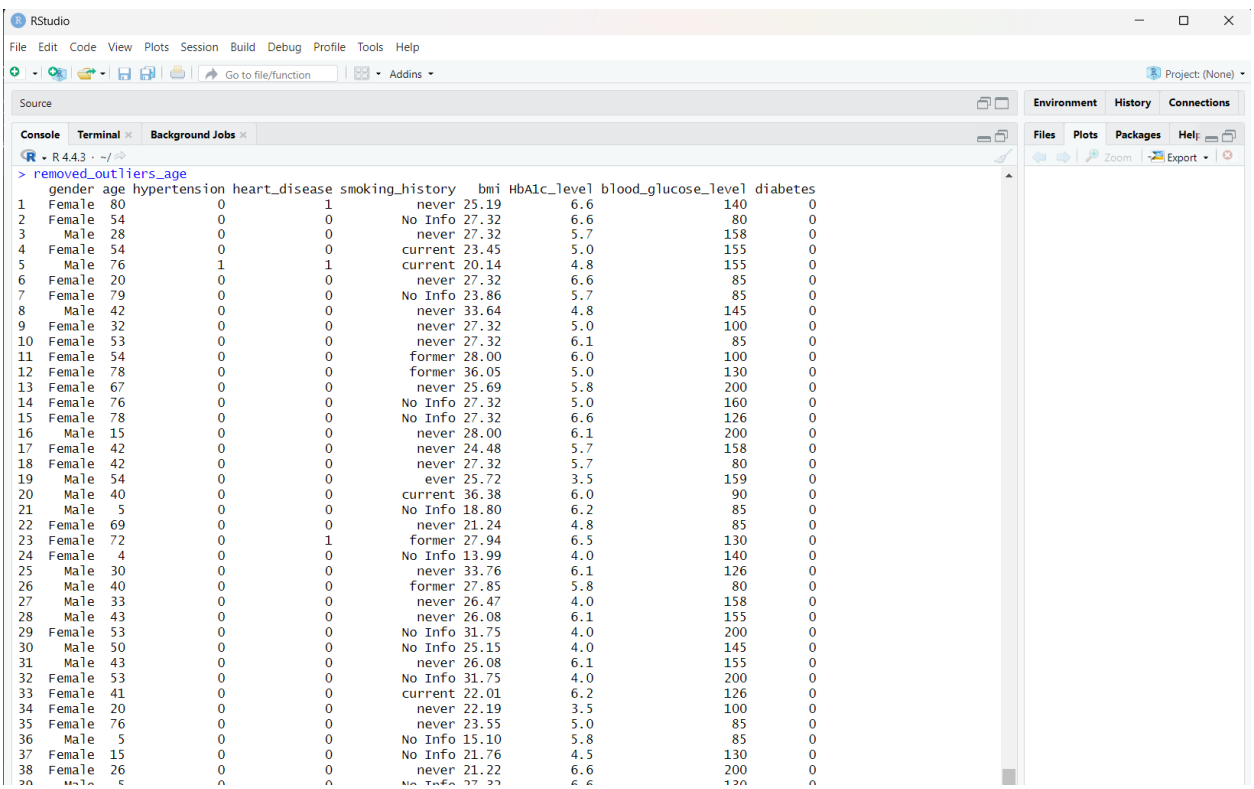
Numerical value replaces by average value and categorical value replace by most frequent value

Detect outliers and handle those values:

```
outlier <- read.csv("/Users/islam/Downloads/Data_Science/Data science LAb/replace_for_categorical.csv", header = TRUE, sep = ",")
Q1_age <- quantile(outlier$age, 0.25)
Q3_age <- quantile(outlier$age, 0.75)
IQR_value_age <- Q3_age - Q1_age
Q1_age
Q3_age

lower_bound <- Q1_age - 1.5 * IQR_value_age
upper_bound <- Q3_age + 1.5 * IQR_value_age

outliers_age<- outlier[outlier$age < lower_bound | outlier$age > upper_bound, ]
outliers_age
removed_outliers_age <- outlier[!(outlier$age %in% outliers_age$age), ]
removed_outliers_age
write.csv(removed_outliers_age,"/Users/islam/Downloads/Data_Science/Data science LAb/after_remove_age_outliers.csv",row.names = FALSE)
```

```
114 Female  64            0              0             ever 49.27          8.2                 140        1
> outliers_age
    gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
54  Female 290            0              0     not current 30.22         5.7                 100        0
121 Female 280            0              0         No Info 27.32         8.8                 159        1
>
```

outlier detected for age column using IQR method then removed outlier



Data after removing outlier for age column

```r
outlier_2 <- read.csv("/Users/islam/Downloads/Data_Science/Data science LAb/after_remove_age_outliers.csv", header = TRUE, sep = ",")
Q1_bmi <- quantile(outlier_2$bmi, 0.25)
Q3_bmi <- quantile(outlier_2$bmi, 0.75)
IQR_value_bmi <- Q3_bmi - Q1_bmi

lower_bound_bmi <- Q1_bmi - 1.5 * IQR_value_bmi
upper_bound_bmi <- Q3_bmi + 1.5 * IQR_value_bmi

outliers_bmi <- outlier_2[outlier_2$bmi < lower_bound_bmi | outlier_2$bmi > upper_bound_bmi, ]
outliers_bmi
removed_outliers_bmi <- outlier_2[outlier_2$bmi >= lower_bound_bmi & outlier_2$bmi <= upper_bound_bmi, ]
removed_outliers_bmi
write.csv(removed_outliers_bmi,"/Users/islam/Downloads/Data_Science/Data science LAb/after_remove_bmi_outliers.csv",row.names = FALSE)
```

```
> outliers_bmi <- outlier_2[outlier_2$bmi < lower_bound_bmi | outlier_2$bmi > upper_bound_bmi, ]
> outliers_bmi
    gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
12  Female  78            0             0          former 36.05         5.0                 130        0
20    Male  40            0             0         current 36.38         6.0                  90        0
24  Female   4            0             0         No Info 13.99         4.0                 140        0
36    Male   5            0             0         No Info 15.10         5.8                  85        0
45  Female  60            0             0           never 18.03         4.0                 159        0
59    Male   7            0             0         No Info 15.94         5.8                 158        0
60    Male   3            0             0         No Info 15.80         6.2                  90        0
65  Female  11            0             0         No Info 17.98         6.5                 159        0
75    Male  50            0             0          former 37.16         9.0                 159        1
76  Female  67            0             0           never 63.48         8.8                 155        1
84  Female  47            0             0           never 36.49         7.5                 155        1
86  Female  61            0             0     not current 39.36         9.0                 140        1
96  Female  80            0             0          former 36.18         6.5                 200        1
97  Female  52            1             0           never 50.30         6.6                 155        1
104 Female  68            0             0         No Info 40.31         7.5                 260        1
107   Male  48            1             0         current 36.12         6.8                 140        1
109   Male  37            0             0           never 37.24         7.0                 126        1
111 Female  59            0             0          former 43.41         6.2                 160        1
113 Female  64            0             0            ever 49.27         8.2                 140        1
114   Male  43            0             0           never 39.00         8.8                 220        1
>
```

Detected outlier for bmi column, using IQR method

```
> removed_outliers_bmi <- outlier_2[outlier_2$bmi >= lower_bound_bmi & outlier_2$bmi <= upper_bound_bmi, ]
> removed_outliers_bmi
    gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
1   Female  80            0             1           never 25.19         6.6                 140        0
2   Female  54            0             0         No Info 27.32         6.6                  80        0
3     Male  28            0             0           never 27.32         5.7                 158        0
4   Female  54            0             0         current 23.45         5.0                 155        0
5     Male  76            1             1         current 20.14         4.8                 155        0
6   Female  20            0             0           never 27.32         6.6                  85        0
7   Female  79            0             0         No Info 23.86         5.7                  85        0
8     Male  42            0             0           never 33.64         4.8                 145        0
9   Female  32            0             0           never 27.32         5.0                 100        0
10  Female  53            0             0           never 27.32         6.1                  85        0
11  Female  54            0             0          former 28.00         6.0                 100        0
13  Female  67            0             0           never 25.69         5.8                 200        0
14  Female  76            0             0         No Info 27.32         5.0                 160        0
15  Female  78            0             0         No Info 27.32         6.6                 126        0
16    Male  15            0             0           never 28.00         6.1                 200        0
17  Female  42            0             0           never 24.48         5.7                 158        0
18  Female  42            0             0           never 27.32         5.7                  80        0
19    Male  54            0             0            ever 25.72         3.5                 159        0
21    Male   5            0             0         No Info 18.80         6.2                  85        0
22  Female  69            0             0           never 21.24         4.8                  85        0
23  Female  72            0             1          former 27.94         6.5                 130        0
25    Male  30            0             0           never 33.76         6.1                 126        0
26    Male  40            0             0          former 27.85         5.8                  80        0
27    Male  33            0             0           never 26.47         4.0                 158        0
28    Male  43            0             0           never 26.08         6.1                 155        0
29  Female  53            0             0         No Info 31.75         4.0                 200        0
30    Male  50            0             0         No Info 25.15         4.0                 145        0
31    Male  43            0             0           never 26.08         6.1                 155        0
32  Female  53            0             0         No Info 31.75         4.0                 200        0
33  Female  41            0             0         current 22.01         6.2                 126        0
34  Female  20            0             0           never 22.19         3.5                 100        0
35  Female  76            0             0           never 23.55         5.0                  85        0
37  Female  15            0             0         No Info 21.76         4.5                 130        0
38  Female  26            0             0           never 21.22         6.6                 200        0
39    Male   5            0             0         No Info 27.32         6.6                 130        0
40  Female  77            1             1           never 32.02         5.0                 159        0
41  Female  66            0             0         No Info 29.30         4.8                 159        0
```

After detecting the outlier remove the outlier for bmi column

```r
outlier_3 <- read.csv("/Users/islam/Downloads/Data_Science/Data science LAb/after_remove_bmi_outliers.csv", header = TRUE, sep = ",")

Q1_blood_glucose_level <- quantile(outlier_3$blood_glucose_level, 0.25)
Q3_blood_glucose_level <- quantile(outlier_3$blood_glucose_level, 0.75)
IQR_value_blood_glucose_level <- Q3_blood_glucose_level - Q1_blood_glucose_level

lower_bound_blood_glucose_level <- Q1_blood_glucose_level - 1.5 * IQR_value_blood_glucose_level
upper_bound_blood_glucose_level <- Q3_blood_glucose_level + 1.5 * IQR_value_blood_glucose_level

outliers_blood_glucose_level <- outlier_3[outlier_3$blood_glucose_level < lower_bound_blood_glucose_level |
                                          outlier_3$blood_glucose_level > upper_bound_blood_glucose_level, ]
outliers_blood_glucose_level

removed_outliers_blood_glucose_level <- outlier_3[outlier_3$blood_glucose_level >= lower_bound_blood_glucose_level &
                                                  outlier_3$blood_glucose_level <= upper_bound_blood_glucose_level, ]
removed_outliers_blood_glucose_level
write.csv(removed_outliers_blood_glucose_level, "/Users/islam/Downloads/Data_Science/Data science LAb/after_remove_blood_glucose_level_outliers.csv", row.names = FALSE)
```

```
> outliers_blood_glucose_level
   gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
64   Male  50            1             0         current 27.32         5.7                 260        1
69 Female  60            0             0           never 27.32         7.5                 300        1
71 Female  80            1             0           never 27.32         6.8                 280        1
72 Female  54            0             0           never 31.70         6.5                 280        1
74   Male  53            0             0         current 30.80         6.6                 280        1
76 Female  43            0             0           never 26.71         6.5                 300        1
87   Male  80            0             1          former 24.36         7.5                 280        1
93 Female  73            0             0           never 35.56         5.8                 260        1
96   Male  62            0             0     not current 32.19         5.8                 300        1
```

Detected outlier for blood_glucose_level column, using IQR method

```
> removed_outliers_blood_glucose_level
    gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
1   Female  80            0             1           never 25.19         6.6                 140        0
2   Female  54            0             0         No Info 27.32         6.6                  80        0
3     Male  28            0             0           never 27.32         5.7                 158        0
4   Female  54            0             0         current 23.45         5.0                 155        0
5     Male  76            1             1         current 20.14         4.8                 155        0
6   Female  20            0             0           never 27.32         6.6                  85        0
7   Female  79            0             0         No Info 23.86         5.7                  85        0
8     Male  42            0             0           never 33.64         4.8                 145        0
9   Female  32            0             0           never 27.32         5.0                 100        0
10  Female  53            0             0           never 27.32         6.1                  85        0
11  Female  54            0             0          former 28.00         6.0                 100        0
12  Female  67            0             0           never 25.69         5.8                 200        0
13  Female  76            0             0         No Info 27.32         5.0                 160        0
14  Female  78            0             0         No Info 27.32         6.6                 126        0
15    Male  15            0             0           never 28.00         6.1                 200        0
16  Female  42            0             0           never 24.48         5.7                 158        0
17  Female  42            0             0           never 27.32         5.7                  80        0
18    Male  54            0             0            ever 25.72         3.5                 159        0
19    Male   5            0             0         No Info 18.80         6.2                  85        0
20  Female  69            0             0           never 21.24         4.8                  85        0
21  Female  72            0             1          former 27.94         6.5                 130        0
22    Male  30            0             0           never 33.76         6.1                 126        0
23    Male  40            0             0          former 27.85         5.8                  80        0
24    Male  33            0             0           never 26.47         4.0                 158        0
25    Male  43            0             0           never 26.08         6.1                 155        0
26  Female  53            0             0         No Info 31.75         4.0                 200        0
27    Male  50            0             0         No Info 25.15         4.0                 145        0
28    Male  43            0             0           never 26.08         6.1                 155        0
29  Female  53            0             0         No Info 31.75         4.0                 200        0
30  Female  41            0             0         current 22.01         6.2                 126        0
31  Female  20            0             0           never 22.19         3.5                 100        0
32  Female  76            0             0           never 23.55         5.0                  85        0
33  Female  15            0             0         No Info 21.76         4.5                 130        0
34  Female  26            0             0           never 21.22         6.6                 200        0
35    Male   5            0             0         No Info 27.32         6.6                 130        0
36  Female  77            1             1           never 32.02         5.0                 159        0
37  Female  66            0             0         No Info 29.30         4.8                 159        0
38  Female  67            0             0         No Info 27.32         3.5                 160        0
```

After detecting the outlier remove the outlier for blood_glucose_level column

convert attributes from numeric to categorical or categorical to numeric

```
convert_atribute <- categorical_columns

convert_atribute$gender<- factor(categorical_columns$gender, levels = c("Male", "Female"), labels = c(1, 0))
convert_atribute$smoking_history <- factor(categorical_columns$smoking_history, levels = c("never", "current", "former", "not current", "ever", "No Info"), labels = c(1, 2,
convert_atribute$hypertension <- factor(categorical_columns$hypertension, levels = c(0, 1), labels = c("NO" , "YES"))
convert_atribute$diabetes <- factor(categorical_columns$diabetes, levels = c(0, 1), labels = c( "positive", "negative"))
convert_atribute
write.csv(convert_atribute, "/Users/islam/Downloads/Data_Science/Data science LAb/categorical_to_numerical_numerical_to_categorical.csv", row.names = FALSE)
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Source | Environment

Console | Terminal | Background Jobs | Files

R 4.4.3 · ~/

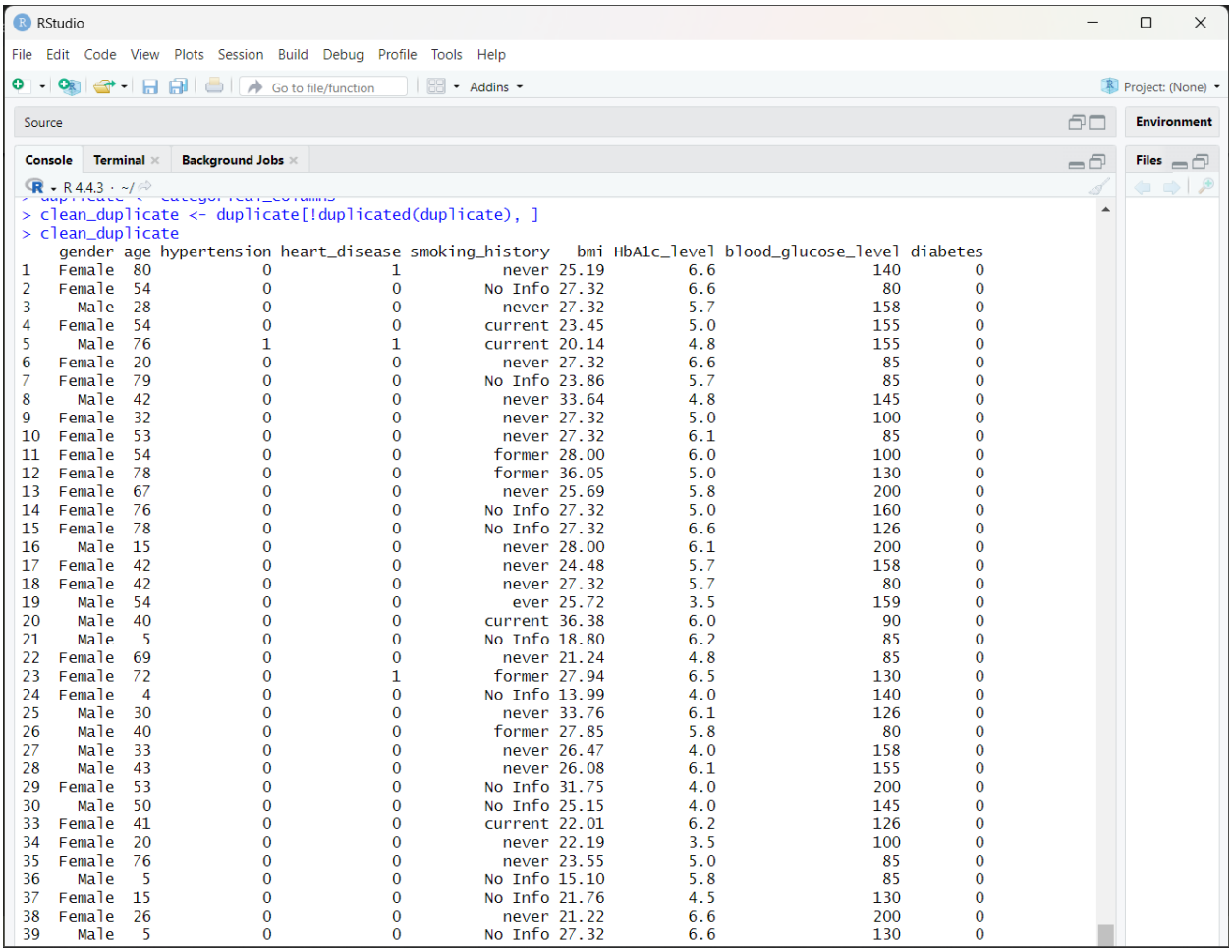
```
   gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
1       0  80           NO            1               1 25.19         6.6                 140 positive
2       0  54           NO            0               6 27.32         6.6                  80 positive
3       1  28           NO            0               1 27.32         5.7                 158 positive
4       0  54           NO            0               2 23.45         5.0                 155 positive
5       1  76          YES            1               2 20.14         4.8                 155 positive
6       0  20           NO            0               1 27.32         6.6                  85 positive
7       0  79           NO            0               6 23.86         5.7                  85 positive
8       1  42           NO            0               1 33.64         4.8                 145 positive
9       0  32           NO            0               1 27.32         5.0                 100 positive
10      0  53           NO            0               1 27.32         6.1                  85 positive
11      0  54           NO            0               3 28.00         6.0                 100 positive
12      0  78           NO            0               3 36.05         5.0                 130 positive
13      0  67           NO            0               1 25.69         5.8                 200 positive
14      0  76           NO            0               6 27.32         5.0                 160 positive
15      0  78           NO            0               6 27.32         6.6                 126 positive
16      1  15           NO            0               1 28.00         6.1                 200 positive
17      0  42           NO            0               1 24.48         5.7                 158 positive
18      0  42           NO            0               1 27.32         5.7                  80 positive
19      1  54           NO            0               5 25.72         3.5                 159 positive
20      1  40           NO            0               2 36.38         6.0                  90 positive
21      1   5           NO            0               6 18.80         6.2                  85 positive
22      0  69           NO            0               1 21.24         4.8                  85 positive
23      0  72           NO            1               3 27.94         6.5                 130 positive
24      0   4           NO            0               6 13.99         4.0                 140 positive
25      1  30           NO            0               1 33.76         6.1                 126 positive
26      1  40           NO            0               3 27.85         5.8                  80 positive
27      1  33           NO            0               1 26.47         4.0                 158 positive
28      1  43           NO            0               1 26.08         6.1                 155 positive
29      0  53           NO            0               6 31.75         4.0                 200 positive
30      1  50           NO            0               6 25.15         4.0                 145 positive
31      1  43           NO            0               1 26.08         6.1                 155 positive
32      0  53           NO            0               6 31.75         4.0                 200 positive
33      0  41           NO            0               2 22.01         6.2                 126 positive
34      0  20           NO            0               1 22.19         3.5                 100 positive
35      0  76           NO            0               1 23.55         5.0                  85 positive
36      1   5           NO            0               6 15.10         5.8                  85 positive
37      0  15           NO            0               6 21.76         4.5                 130 positive
38      0  26           NO            0               1 21.22         6.6                 200 positive
39      1   5           NO            0               6 27.32         6.6                 130 positive
```

convert attributes from numeric to categorical or categorical to numeric. Gender column and smoking_history column convert categorical to numeric. For gender Where male =1 and female=0.

Hypertension and diabetes column convert attributes from numeric. Where 0 replace by NO and 1 replace by YES, for diabetes column 0 replace by positive and 1 replace by negative.

remove duplicate rows

```
duplicate <- categorical_columns

clean_duplicate <- duplicate[!duplicated(duplicate), ]
clean_duplicate
write.csv(clean_duplicate, "/Users/islam/Downloads/Data_Science/Data science LAb/remove_duplicate.csv", row.names = FALSE)
```



	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
1	Female	80	0	1	never	25.19	6.6	140	0
2	Female	54	0	0	No Info	27.32	6.6	80	0
3	Male	28	0	0	never	27.32	5.7	158	0
4	Female	54	0	0	current	23.45	5.0	155	0
5	Male	76	1	1	current	20.14	4.8	155	0
6	Female	20	0	0	never	27.32	6.6	85	0
7	Female	79	0	0	No Info	23.86	5.7	85	0
8	Male	42	0	0	never	33.64	4.8	145	0
9	Female	32	0	0	never	27.32	5.0	100	0
10	Female	53	0	0	never	27.32	6.1	85	0
11	Female	54	0	0	former	28.00	6.0	100	0
12	Female	78	0	0	former	36.05	5.0	130	0
13	Female	67	0	0	never	25.69	5.8	200	0
14	Female	76	0	0	No Info	27.32	5.0	160	0
15	Female	78	0	0	No Info	27.32	6.6	126	0
16	Male	15	0	0	never	28.00	6.1	200	0
17	Female	42	0	0	never	24.48	5.7	158	0
18	Female	42	0	0	never	27.32	5.7	80	0
19	Male	54	0	0	ever	25.72	3.5	159	0
20	Male	40	0	0	current	36.38	6.0	90	0
21	Male	5	0	0	No Info	18.80	6.2	85	0
22	Female	69	0	0	never	21.24	4.8	85	0
23	Female	72	0	1	former	27.94	6.5	130	0
24	Female	4	0	0	No Info	13.99	4.0	140	0
25	Male	30	0	0	never	33.76	6.1	126	0
26	Male	40	0	0	former	27.85	5.8	80	0
27	Male	33	0	0	never	26.47	4.0	158	0
28	Male	43	0	0	never	26.08	6.1	155	0
29	Female	53	0	0	No Info	31.75	4.0	200	0
30	Male	50	0	0	No Info	25.15	4.0	145	0
33	Female	41	0	0	current	22.01	6.2	126	0
34	Female	20	0	0	never	22.19	3.5	100	0
35	Female	76	0	0	never	23.55	5.0	85	0
36	Male	5	0	0	No Info	15.10	5.8	85	0
37	Female	15	0	0	No Info	21.76	4.5	130	0
38	Female	26	0	0	never	21.22	6.6	200	0
39	Male	5	0	0	No Info	27.32	6.6	130	0

Remove duplicate by finding duplicate values through duplicated () function

Normalization:

```
min_max_norm <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}
normalization_data <- categorical_columns

normalization_data$bmi <- min_max_norm(normalization_data$bmi)
normalization_data$bmi
normalization_data$blood_glucose_level  <- min_max_norm(normalization_data$blood_glucose_level )
normalization_data$blood_glucose_level
normalization_data
write.csv(new_data, "/Users/islam/Downloads/Data_Science/Data science LAb/normalized_data.csv", row.names = FALSE)
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Source | Environment

Console | Terminal | Background Jobs | Files

R 4.4.3 · ~/

```
> normalization_data
    gender age hypertension heart_disease smoking_history        bmi HbA1c_level blood_glucose_level diabetes
1   Female  80            0             1           never 0.22630835         6.6          0.27272727        0
2   Female  54            0             0         No Info 0.26934734         6.6          0.00000000        0
3     Male  28            0             0           never 0.26934734         5.7          0.35454545        0
4   Female  54            0             0         current 0.19114973         5.0          0.34090909        0
5     Male  76            1             1         current 0.12426753         4.8          0.34090909        0
6   Female  20            0             0           never 0.26934734         6.6          0.02272727        0
7   Female  79            0             0         No Info 0.19943423         5.7          0.02272727        0
8     Male  42            0             0           never 0.39704991         4.8          0.29545455        0
9   Female  32            0             0           never 0.26934734         5.0          0.09090909        0
10  Female  53            0             0           never 0.26934734         6.1          0.02272727        0
11  Female  54            0             0          former 0.28308749         6.0          0.09090909        0
12  Female  78            0             0          former 0.44574662         5.0          0.22727273        0
13  Female  67            0             0           never 0.23641140         5.8          0.54545455        0
14  Female  76            0             0         No Info 0.26934734         5.0          0.36363636        0
15  Female  78            0             0         No Info 0.26934734         6.6          0.20909091        0
16    Male  15            0             0           never 0.28308749         6.1          0.54545455        0
17  Female  42            0             0           never 0.21196201         5.7          0.35454545        0
18  Female  42            0             0           never 0.26934734         5.7          0.00000000        0
19    Male  54            0             0            ever 0.23701758         3.5          0.35909091        0
20    Male  40            0             0         current 0.45241463         6.0          0.04545455        0
21    Male   5            0             0         No Info 0.09719135         6.2          0.02272727        0
22  Female  69            0             0           never 0.14649424         4.8          0.02272727        0
23  Female  72            0             1          former 0.28187513         6.5          0.22727273        0
24  Female   4            0             0         No Info 0.00000000         4.0          0.27272727        0
25    Male  30            0             0           never 0.39947464         6.1          0.20909091        0
26    Male  40            0             0          former 0.28005658         5.8          0.00000000        0
27    Male  33            0             0           never 0.25217216         4.0          0.35454545        0
28    Male  43            0             0           never 0.24429178         6.1          0.34090909        0
29  Female  53            0             0         No Info 0.35886038         4.0          0.54545455        0
30    Male  50            0             0         No Info 0.22550010         4.0          0.29545455        0
31    Male  43            0             0           never 0.24429178         6.1          0.34090909        0
32  Female  53            0             0         No Info 0.35886038         4.0          0.54545455        0
33  Female  41            0             0         current 0.16205294         6.2          0.20909091        0
34  Female  20            0             0           never 0.16569004         3.5          0.09090909        0
35  Female  76            0             0           never 0.19317034         5.0          0.02272727        0
36    Male   5            0             0         No Info 0.02242877         5.8          0.02272727        0
37  Female  15            0             0         No Info 0.15700141         4.5          0.22727273        0
38  Female  26            0             0           never 0.14609012         6.6          0.54545455        0
```

Min-Max normalization to the bmi column and blood_glucose_level in a dataset called categorical_columns. First create a function which contain the formula of min-max. then invoke the function by column value.

filtering methods to filter the data

```
filtered_data2 <- filter(categorical_columns, categorical_columns$gender == "Male" & (categorical_columns$age>=25 & categorical_columns$age<50))
filtered_data2
```

```
> filtered_data2
   gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
1    Male  28            0             0           never 27.32         5.7                 158        0
2    Male  42            0             0           never 33.64         4.8                 145        0
3    Male  40            0             0         current 36.38         6.0                  90        0
4    Male  30            0             0           never 33.76         6.1                 126        0
5    Male  40            0             0          former 27.85         5.8                  80        0
6    Male  33            0             0           never 26.47         4.0                 158        0
7    Male  43            0             0           never 26.08         6.1                 155        0
8    Male  43            0             0           never 26.08         6.1                 155        0
9    Male  43            0             0         No Info 23.04         5.7                 160        0
10   Male  43            0             0           never 27.32         3.5                 126        0
11   Male  34            0             0           never 31.16         5.8                  90        0
12   Male  29            0             0         current 25.41         6.1                 130        1
13   Male  48            1             0         current 36.12         6.8                 140        1
14   Male  37            0             0           never 37.24         7.0                 126        1
15   Male  43            0             0           never 39.00         8.8                 220        1
16   Male  43            0             0           never 22.43         7.0                 160        1
17   Male  33            1             0            ever 25.94         9.0                 140        1
18   Male  43            0             0            ever 19.46         9.0                 130        1
```

Filtering gender column where gender Male and age is between 25 and 50.

```
filtered_data2 <- filter(categorical_columns, categorical_columns$gender == "Female" & smoking_history == "ever")
filtered_data2
write.csv(filtered_data2, "/Users/islam/Downloads/Data_Science/Data science LAb/filter2.csv", row.names = FALSE)
```

```
> filtered_data2 <- filter(categorical_columns, categorical_columns$gender == "Female" & smoking_history == "ever")
> filtered_data2
  gender age hypertension heart_disease smoking_history   bmi HbA1c_level blood_glucose_level diabetes
1 Female  59            0             1            ever 23.11         6.5                 200        0
2 Female  64            0             0            ever 49.27         8.2                 140        1
```

Filtering female those who had ever smoking history

imbalanced data set into the balanced data:

Sampling:

```
table(AllData$diabetes)
df_over <- ovun.sample(diabetes ~ ., data = AllData, method = "over",
                       N = max(table(AllData$diabetes)) * 2)$data
table(df_over$diabetes)
write.csv(df_over, "/Users/abdullahalmaruf/Desktop/DATA_SCIENCE_PROJECT/balanced_over.csv",
          row.names = FALSE)
```

```
> table(AllData$diabetes)

 0  1
68 51
```

Before oversampling on diabetes features in dataset.

```
> table(df_over$diabetes)

 0  1
68 68
```

Output after oversampling on diabetes features in dataset.

```
df_under <- ovun.sample(diabetes ~ ., data = AllData, method = "under",
                        N = min(table(AllData$diabetes)) * 2)$data
table(df_under$diabetes)
write.csv(df_under, "/Users/abdullahalmaruf/Desktop/DATA_SCIENCE_PROJECT/balanced_under.csv",
          row.names = FALSE)
```

```
> table(df_under$diabetes)

 0  1
51 51
```

Undersampling on diabetes features in dataset.

Split Training and Testing data:

```
set.seed(135)
split <- sample.split(AllData$diabetes, SplitRatio = 0.8)
train_data <- subset(AllData, split == TRUE)
test_data  <- subset(AllData, split == FALSE)
nrow(train_data)
nrow(test_data)

write.csv(train_data, "/Users/abdullahalmaruf/Desktop/DATA_SCIENCE_PROJECT/train_data.csv",
          row.names = FALSE)
write.csv(test_data, "/Users/abdullahalmaruf/Desktop/DATA_SCIENCE_PROJECT/test_data.csv",
          row.names = FALSE)
```

```
> nrow(train_data)
[1] 95
> nrow(test_data)
[1] 24
>
```

This code splits the dataset `AllData` into training and testing sets based on a 80-20 ratio using the `sample.split()` function from the `caTools` package. The training data is stored in `train_data`, and the testing data is stored in `test_data`. The number of rows in each dataset is printed using `nrow()`.

Compare the central tendencies (mean, median, mode) of Age across different groups of Gender and interpret the results :

```
data$gender <- as.factor(data$gender)


gender_age_stats <- data %>%
  group_by(gender) %>%
  summarise(
    Mean_Age = mean(age, na.rm = TRUE),
    Median_Age = median(age, na.rm = TRUE),
    Mode_Age = as.numeric(names(sort(table(age), decreasing = TRUE)[1]))
  )

print("Central Tendencies of Age by Gender:")
print(gender_age_stats)
```

```
> print(gender_age_stats)
# A tibble: 2 × 4
  gender Mean_Age Median_Age Mode_Age
  <fct>     <dbl>      <dbl>    <dbl>
1 Female     57.8         54       43
2 Male       47.2         49       43
>
```

Calculating mean, median, and mode of age grouped by gender. Females have a higher mean and median age compared to males. The most common age (mode) is 43 for both genders.

Compare the central tendencies (mean, median, mode) of Age across hypertension and interpret the results:

```r
data$hypertension <- as.factor(data$hypertension)

hypertension_age_stats <- data %>%
  group_by(hypertension) %>%
  summarise(
    Mean_Age = mean(age, na.rm = TRUE),
    Median_Age = median(age, na.rm = TRUE),
    Mode_Age = as.numeric(names(sort(table(age), decreasing = TRUE)[1]))
  )


print("Central Tendencies of Age by Hypertension:")
print(hypertension_age_stats)
```

```
> print(hypertension_age_stats)
# A tibble: 2 × 4
  hypertension Mean_Age Median_Age Mode_Age
  <fct>           <dbl>      <dbl>    <dbl>
1 0                53.0         52       43
2 1                61.5         60       33
>
```

Calculate mean, median, and mode of age grouped by hypertension (0 = No, 1 = Yes). People with hypertension have higher mean and median ages compared to those without. Mode shifts from 43 (no hypertension) to 33 (hypertension).

Compare the Spread (Range, IQR, Variance, Standard Deviation) of Age across different groups of Gender and interpret the results:

```r
data$gender <- as.factor(data$gender)


gender_age_spread <- data %>%
  group_by(gender) %>%
  summarise(
    Min_Age = min(age, na.rm = TRUE),
    Max_Age = max(age, na.rm = TRUE),
    Range_Age = max(age, na.rm = TRUE) - min(age, na.rm = TRUE),
    IQR_Age = IQR(age, na.rm = TRUE),
    Variance_Age = var(age, na.rm = TRUE),
    SD_Age = sd(age, na.rm = TRUE)
  )


print("Spread of Age by Gender:")
print(gender_age_spread)
```

```
> print(gender_age_spread)
# A tibble: 2 × 7
  gender Min_Age Max_Age Range_Age IQR_Age Variance_Age SD_Age
  <fct>    <int>   <int>     <int>   <dbl>        <dbl>  <dbl>
1 Female       3     290       287    28          1878.   43.3
2 Male         3      80        77    26.5         470.   21.7
> z
```

Calculates range, IQR, variance, and standard deviation of age grouped by gender. Females show much wider age spread, higher variance, and higher standard deviation than males.

Conclusion:

We performed essential data preprocessing steps, including handling missing values, outliers, duplicates, and invalid data. Attributes were converted as needed, and continuous variables were normalized. We balanced the dataset, applied filtering, and split the data for training and testing. Finally, by comparing the central tendencies and spread of age across gender and hypertension groups, we gained key insights into the dataset's structure.