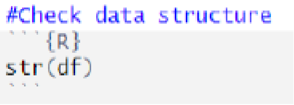
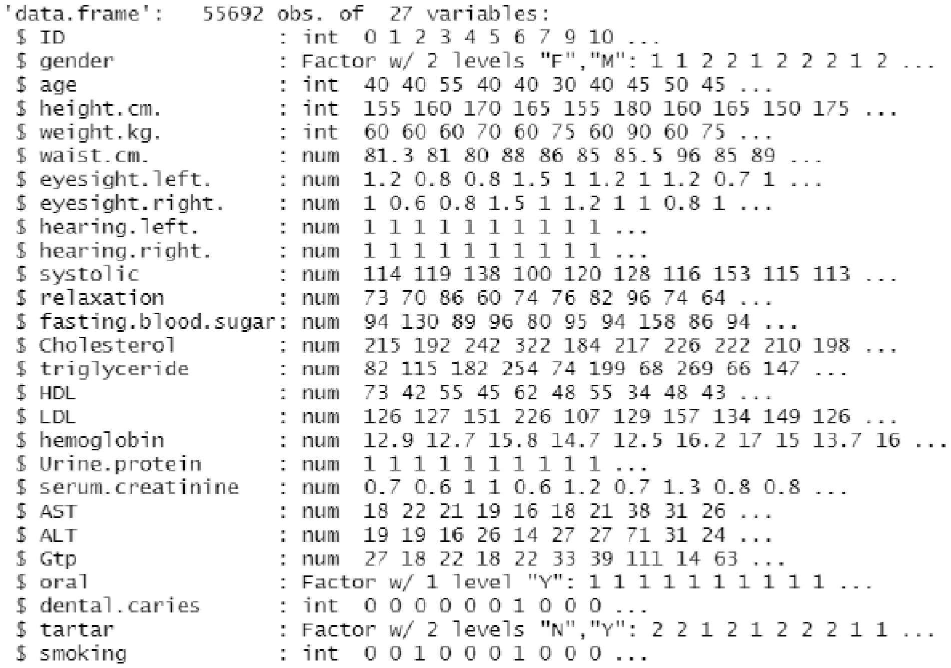
Exploratory Data Analysis

1. Data Structure and data summary

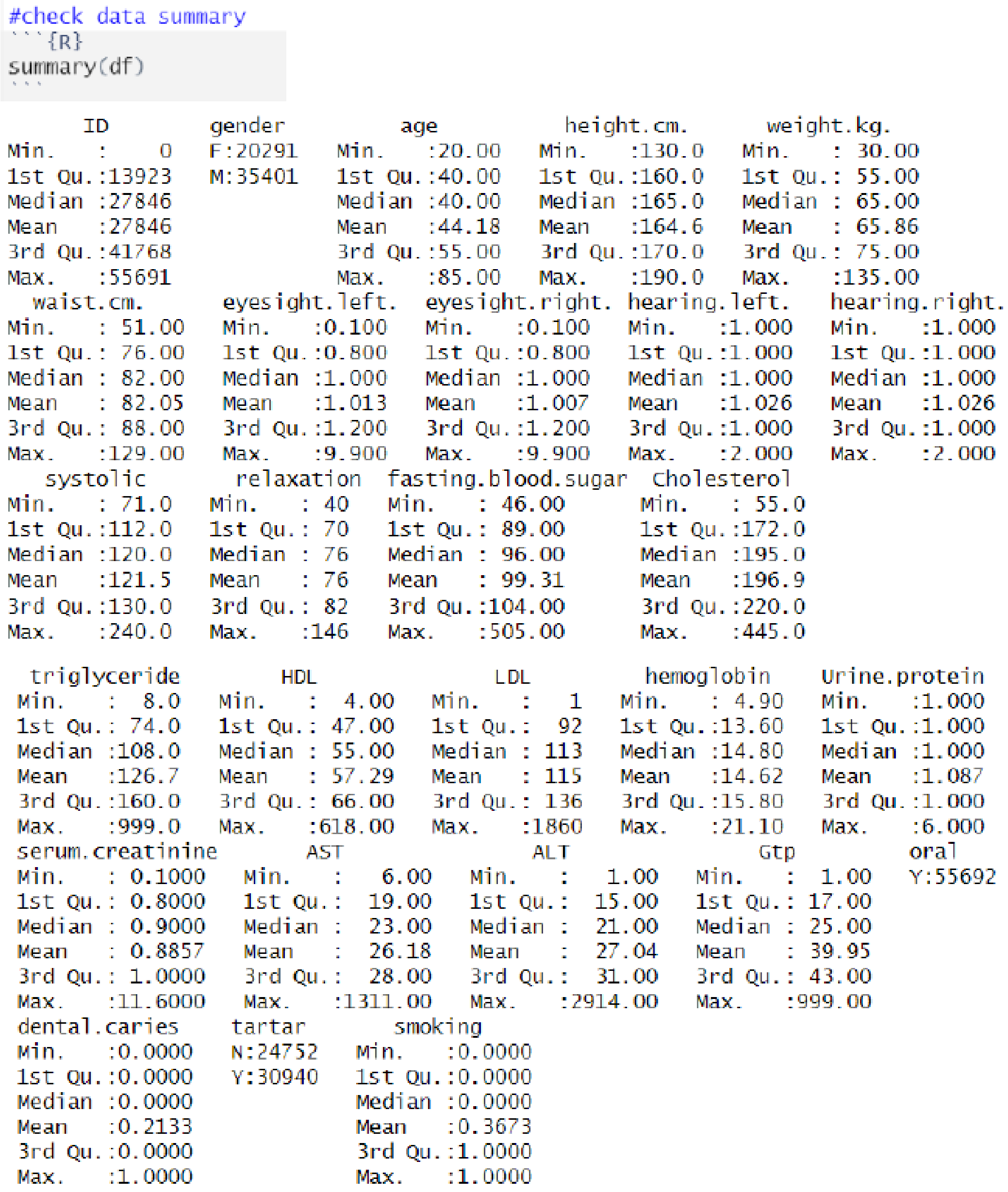
Firstly, the data structure and data summary are checked.



~~~
#Check data structure
```{R}
str(df)
```
~~~



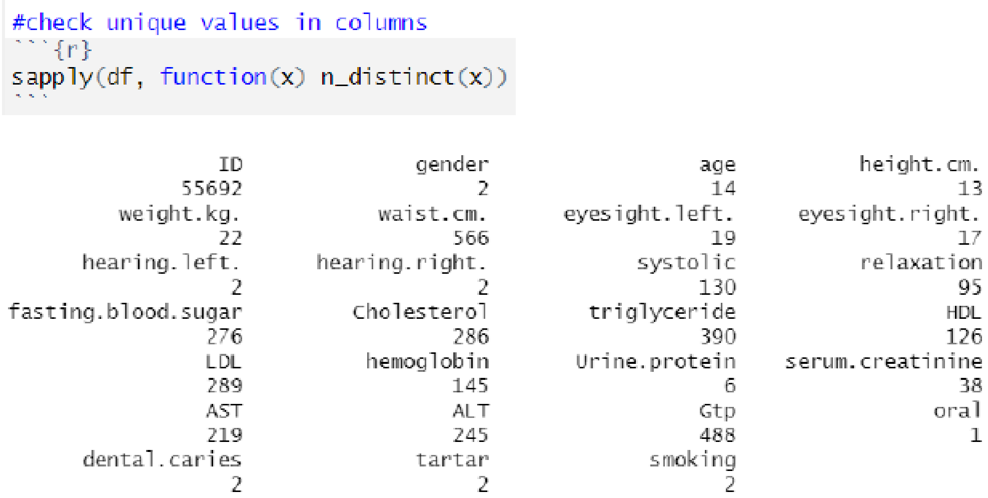
```
'data.frame':	55692 obs. of  27 variables:
 $ ID                 : int  0 1 2 3 4 5 6 7 9 10 ...
 $ gender             : Factor w/ 2 levels "F","M": 1 1 2 2 1 2 2 2 1 2 ...
 $ age                : int  40 40 55 40 40 30 40 45 50 45 ...
 $ height.cm.         : int  155 160 170 165 155 180 160 165 150 175 ...
 $ weight.kg.         : int  60 60 60 70 60 75 60 90 60 75 ...
 $ waist.cm.          : num  81.3 81 80 88 86 85 85.5 96 85 89 ...
 $ eyesight.left.     : num  1.2 0.8 0.8 1.5 1 1.2 1 1.2 0.7 1 ...
 $ eyesight.right.    : num  1 0.6 0.8 1.5 1 1.2 1 1 0.8 1 ...
 $ hearing.left.      : num  1 1 1 1 1 1 1 1 1 1 ...
 $ hearing.right.     : num  1 1 1 1 1 1 1 1 1 1 ...
 $ systolic           : num  114 119 138 100 120 128 116 153 115 113 ...
 $ relaxation         : num  73 70 86 60 74 76 82 96 74 64 ...
 $ fasting.blood.sugar: num  94 130 89 96 80 95 94 158 86 94 ...
 $ Cholesterol        : num  215 192 242 322 184 217 226 222 210 198 ...
 $ triglyceride       : num  82 115 182 254 74 199 68 269 66 147 ...
 $ HDL                : num  73 42 55 45 62 48 55 34 48 43 ...
 $ LDL                : num  126 127 151 226 107 129 157 134 149 126 ...
 $ hemoglobin         : num  12.9 12.7 15.8 14.7 12.5 16.2 17 15 13.7 16 ...
 $ Urine.protein      : num  1 1 1 1 1 1 1 1 1 1 ...
 $ serum.creatinine   : num  0.7 0.6 1 1 0.6 1.2 0.7 1.3 0.8 0.8 ...
 $ AST                : num  18 22 21 19 16 18 21 38 31 26 ...
 $ ALT                : num  19 19 16 26 14 27 27 71 31 24 ...
 $ Gtp                : num  27 18 22 18 22 33 39 111 14 63 ...
 $ oral               : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
 $ dental.caries      : int  0 0 0 0 0 0 1 0 0 0 ...
 $ tartar             : Factor w/ 2 levels "N","Y": 2 2 1 2 1 2 2 2 1 1 ...
 $ smoking            : int  0 0 1 0 0 0 1 0 0 0 ...
```



~~~~
#check data summary
```{R}
summary(df)
```
~~~~

```
       ID           gender         age           height.cm.       weight.kg.    
 Min.   :    0   F:20291   Min.   :20.00   Min.   :130.0   Min.   : 30.00  
 1st Qu.:13923   M:35401   1st Qu.:40.00   1st Qu.:160.0   1st Qu.: 55.00  
 Median :27846             Median :40.00   Median :165.0   Median : 65.00  
 Mean   :27846             Mean   :44.18   Mean   :164.6   Mean   : 65.86  
 3rd Qu.:41768             3rd Qu.:55.00   3rd Qu.:170.0   3rd Qu.: 75.00  
 Max.   :55691             Max.   :85.00   Max.   :190.0   Max.   :135.00  
    waist.cm.       eyesight.left.  eyesight.right.  hearing.left.    hearing.right.
 Min.   : 51.00   Min.   :0.100   Min.   :0.100   Min.   :1.000   Min.   :1.000  
 1st Qu.: 76.00   1st Qu.:0.800   1st Qu.:0.800   1st Qu.:1.000   1st Qu.:1.000  
 Median : 82.00   Median :1.000   Median :1.000   Median :1.000   Median :1.000  
 Mean   : 82.05   Mean   :1.013   Mean   :1.007   Mean   :1.026   Mean   :1.026  
 3rd Qu.: 88.00   3rd Qu.:1.200   3rd Qu.:1.200   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :129.00   Max.   :9.900   Max.   :9.900   Max.   :2.000   Max.   :2.000  
    systolic        relaxation   fasting.blood.sugar  Cholesterol   
 Min.   : 71.0   Min.   : 40   Min.   : 46.00       Min.   : 55.0  
 1st Qu.:112.0   1st Qu.: 70   1st Qu.: 89.00       1st Qu.:172.0  
 Median :120.0   Median : 76   Median : 96.00       Median :195.0  
 Mean   :121.5   Mean   : 76   Mean   : 99.31       Mean   :196.9  
 3rd Qu.:130.0   3rd Qu.: 82   3rd Qu.:104.00       3rd Qu.:220.0  
 Max.   :240.0   Max.   :146   Max.   :505.00       Max.   :445.0  

  triglyceride        HDL              LDL           hemoglobin     Urine.protein 
 Min.   :  8.0   Min.   :  4.00   Min.   :   1   Min.   : 4.90   Min.   :1.000  
 1st Qu.: 74.0   1st Qu.: 47.00   1st Qu.:  92   1st Qu.:13.60   1st Qu.:1.000  
 Median :108.0   Median : 55.00   Median : 113   Median :14.80   Median :1.000  
 Mean   :126.7   Mean   : 57.29   Mean   : 115   Mean   :14.62   Mean   :1.087  
 3rd Qu.:160.0   3rd Qu.: 66.00   3rd Qu.: 136   3rd Qu.:15.80   3rd Qu.:1.000  
 Max.   :999.0   Max.   :618.00   Max.   :1860   Max.   :21.10   Max.   :6.000  
 serum.creatinine       AST               ALT               Gtp           oral    
 Min.   : 0.1000   Min.   :   6.00   Min.   :   1.00   Min.   :  1.00   Y:55692  
 1st Qu.: 0.8000   1st Qu.:  19.00   1st Qu.:  15.00   1st Qu.: 17.00            
 Median : 0.9000   Median :  23.00   Median :  21.00   Median : 25.00            
 Mean   : 0.8857   Mean   :  26.18   Mean   :  27.04   Mean   : 39.95            
 3rd Qu.: 1.0000   3rd Qu.:  28.00   3rd Qu.:  31.00   3rd Qu.: 43.00            
 Max.   :11.6000   Max.   :1311.00   Max.   :2914.00   Max.   :999.00            
 dental.caries      tartar        smoking      
 Min.   :0.0000   N:24752   Min.   :0.0000  
 1st Qu.:0.0000   Y:30940   1st Qu.:0.0000  
 Median :0.0000             Median :0.0000  
 Mean   :0.2133             Mean   :0.3673  
 3rd Qu.:0.0000             3rd Qu.:1.0000  
 Max.   :1.0000             Max.   :1.0000  
```



~~~~
#check unique values in columns
```{r}
sapply(df, function(x) n_distinct(x))
```
~~~~

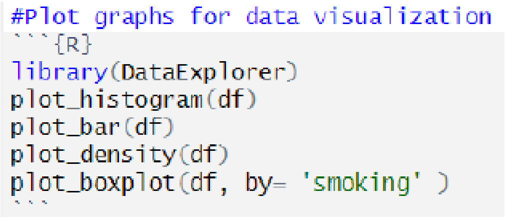
```
                 ID              gender                 age          height.cm. 
              55692                   2                  14                  13 
          weight.kg.           waist.cm.      eyesight.left.     eyesight.right. 
                 22                 566                  19                  17 
       hearing.left.      hearing.right.            systolic          relaxation 
                  2                   2                 130                  95 
fasting.blood.sugar         Cholesterol        triglyceride                 HDL 
                276                 286                 390                 126 
                LDL          hemoglobin       Urine.protein    serum.creatinine 
                289                 145                   6                  38 
                AST                 ALT                 Gtp                oral 
                219                 245                 488                   1 
      dental.caries              tartar             smoking 
                  2                   2                   2 
```

Originally, the dataset does not contain missing values. However, 1% of missing data is introduced to the dataset in the next step to experiment on imputing missing values using missForest package.
The data structure and summary revealed multiple observations:

- All variables are presented in numeric or integer form except gender, oral and tartar which are presented in factor form.
- No missing values across the dataset.
- The values in age, height and weight variables are multiples of 5, the observations have been grouped accordingly.
- The range of value input for hearing.left, hearing.right, urine.protein, dental.cares and smoking variables is relatively small.
- The 'oral' variable has the same input of value 1 throughout the entire dataset.
- Spelling mistake on dental.caries column name.

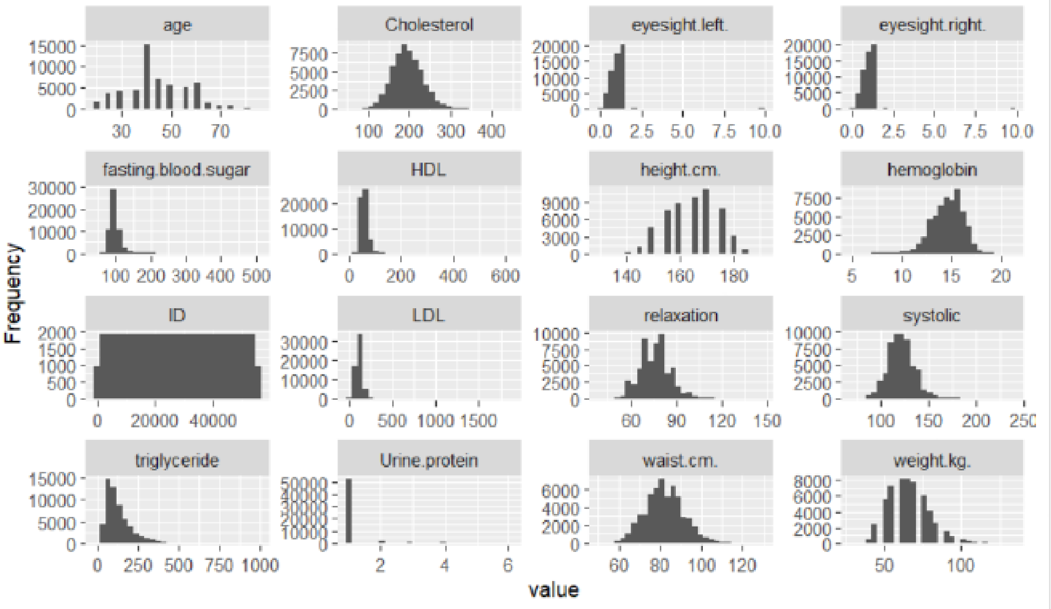
2. Data visualization

In order to understand the data distribution better, the data is further explored with data visualization using various plots including histogram, density plot, bar graph and box plot.



~~~
#Plot graphs for data visualization
```{R}
library(DataExplorer)
plot_histogram(df)
plot_bar(df)
plot_density(df)
plot_boxplot(df, by= 'smoking' )
```
~~~

Histogram



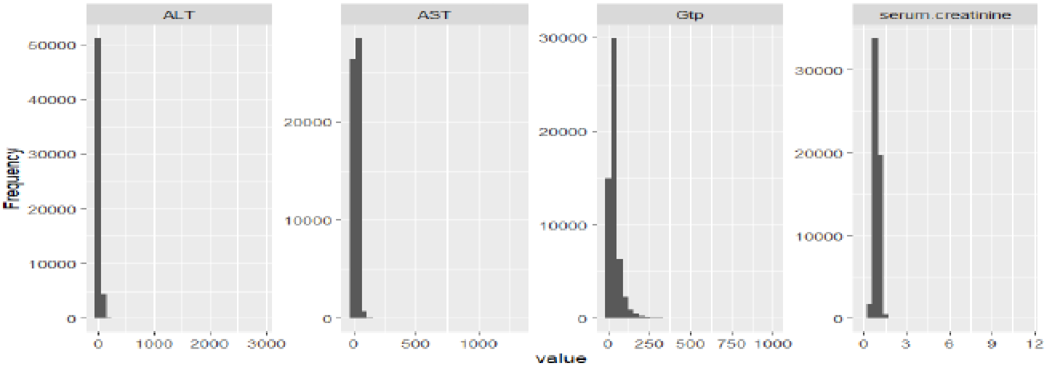


Figure 1 Histograms of Continuous Variable

Histograms show the data distribution of non-factor data type variables. From the histogram, it is noticed that there are gaps in between for age, height, and weight variables. Cholesterol, fasting blood sugar, HDL, LDL, relaxation, systolic, waist show a Gaussian distribution. On the other hand, hemoglobin histogram is slightly left-skewed whereas triglyceride is right-skewed. ALT, AST Gtp, serum.creatinine have a small range of values, where most of the observations fall under the minimum value.

<u>Density plot</u>

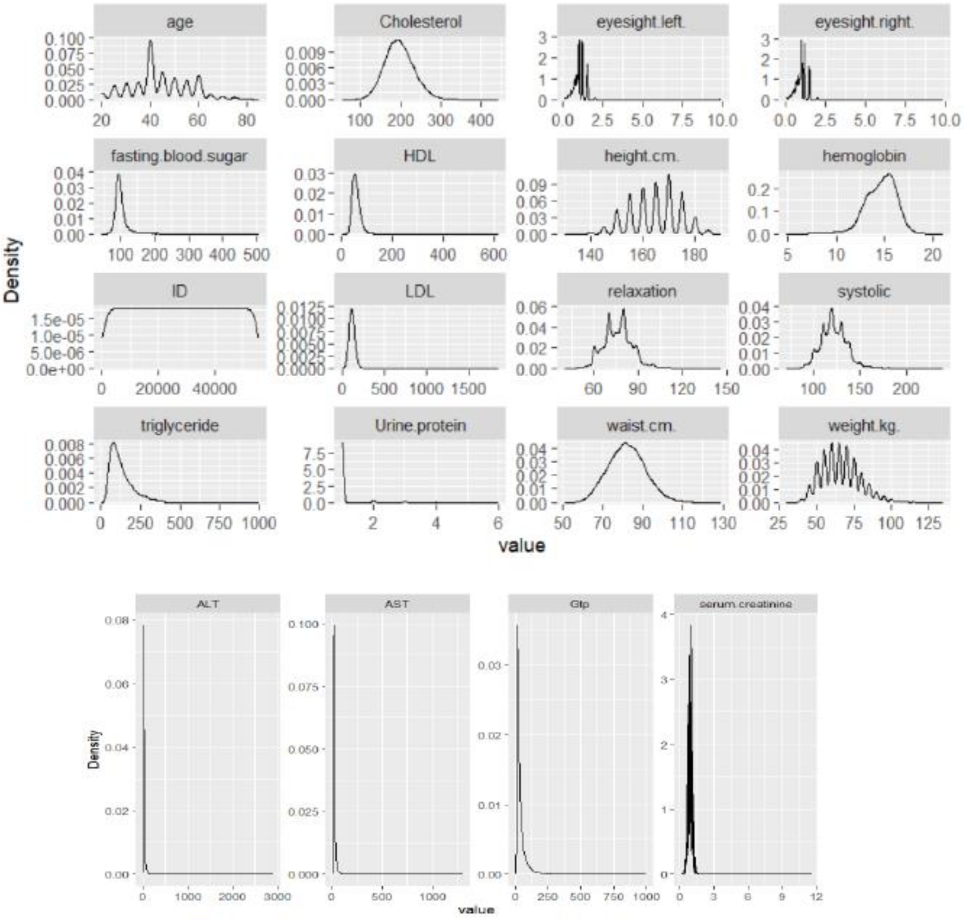


Figure 2 Density Plot of Continuous Variable

From the density plots, similar observations were noticed. The spiky curve in age, heigh weight further confirm on there is a gap in between. The density plots for Cholesterol, fasting blood sugar, HDL, LDL, relaxation, systolic, waist show a bell-curve, whereas the density plot of hemoglobin and triglycerides is slightly skewed. Due to the narrow range of values in ALT, AST, Gtp and serum creatinine, these variables show a spiky appearance density plot.

Bar Plot

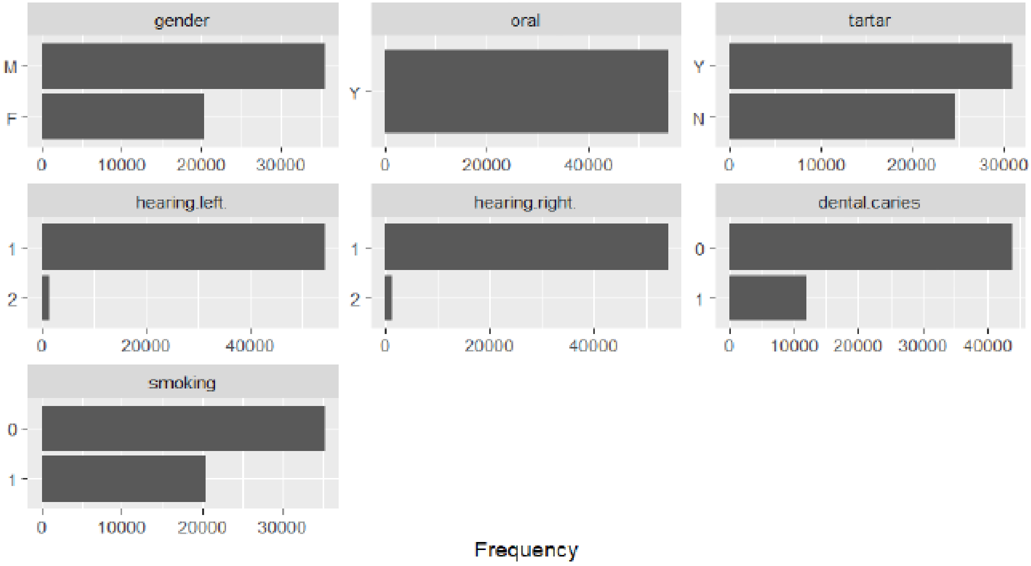


Figure 3 Bar Plots of Categorical Variable

The bar plots show that oral has only one input value throughout the entire dataset. Thus, this variable should be dropped. Furthermore, it is confirmed that hearing.left, hearing.right, dental.cares and smoking have only 2 unique values. Therefore, these columns should be converted to factor data type.

Box Plot

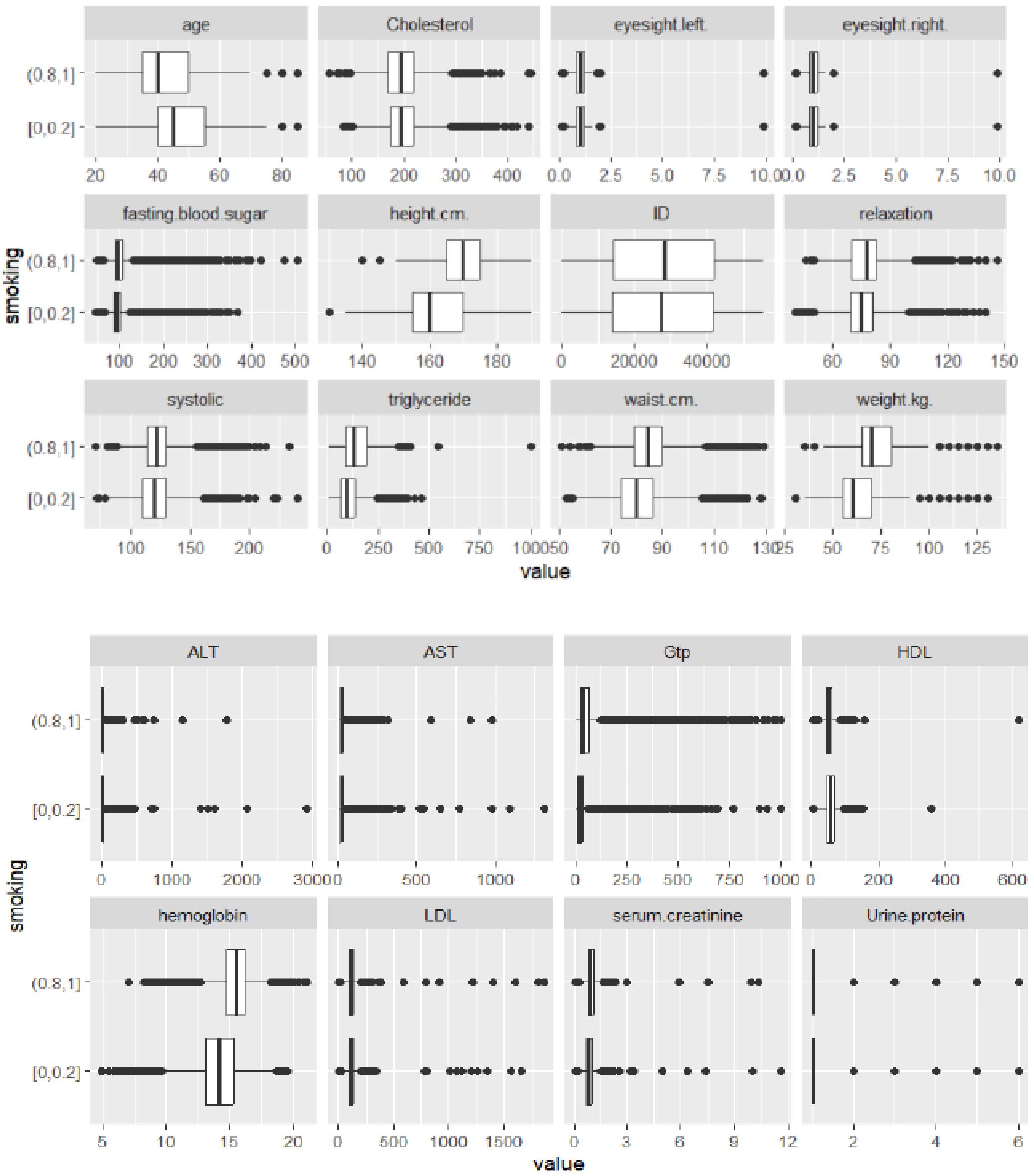
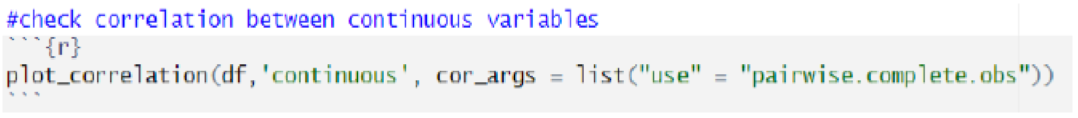


Figure 4 Boxplot of Variables

Outliers are noticed in the boxplots. However, noticed that most of the variables have high number of outliers. Removing all outliers may cause information loss as the high readings of each biological profile could have a certain level of relationship with smoking status. Therefore, only extreme outliers will be removed from the dataset. Extreme outliers are a single point which was noticed at the very far end from the other outliers. Extreme value in cholesterol, systolic, fasting blood sugar, triglyceride, ALT, AST and HDL will be removed. Specifically, observations with cholesterol value larger than 400, systolic value less than 50 and more than 250, fasting blood sugar value more than 450, triglyceride value more than 500, ALT and AST with value more than 500, and HDL with value more than 300 will be removed.

Correlation Plot



~~~
#check correlation between continuous variables
```{r}
plot_correlation(df,'continuous', cor_args = list("use" = "pairwise.complete.obs"))
```
~~~

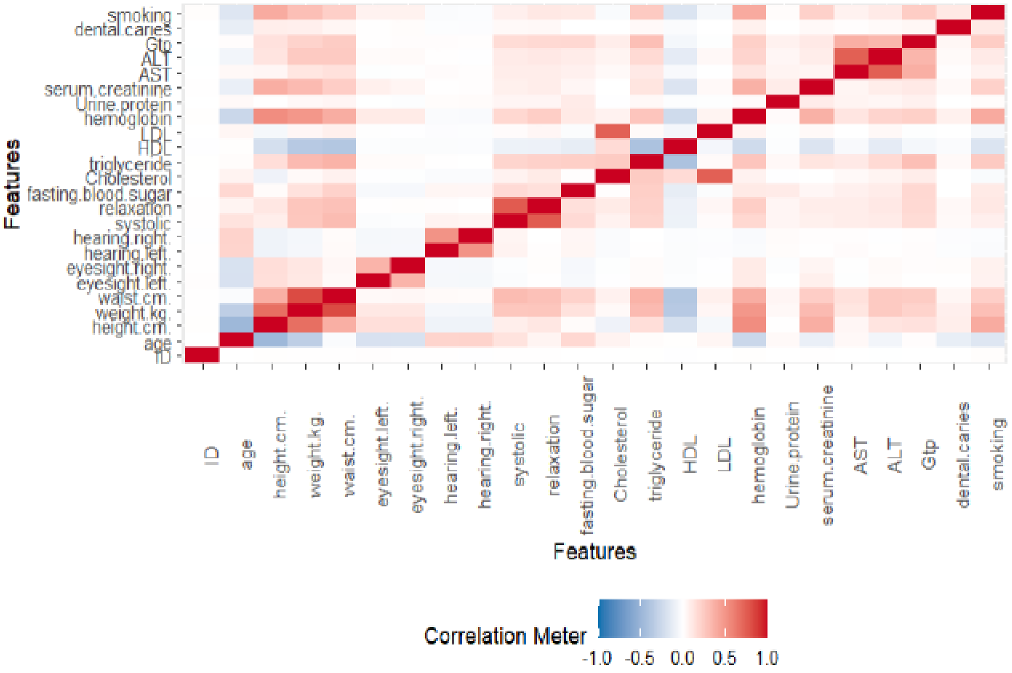


Figure 5 Correlation Plot of Continuous Variables

Few observations are observed from the correlation heatmap:

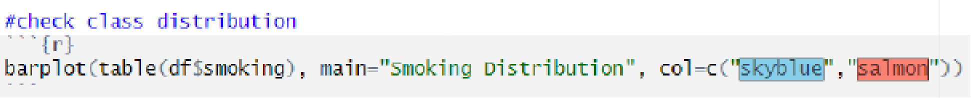
- Most of the variables are positively correlated with each other whereas only age and HDL are negatively correlated with most other variables.

- HDL is negatively correlated with most of the variables except cholesterol. Specifically, there is a strong negative correlation between HDL and triglyceride, waist and weight.

- Age is also negatively correlated with almost half of the variables. Specifically, age has strong negative correlation with height, weight and hemoglobin.

- Eyesight (left and right), hearing (left and right), urine.protein and dental cares have weak to no correlation with other variables. These variables can be excluded from further analysis.

Class Distribution



~~~
#check class distribution
```{r}
barplot(table(df$smoking), main="Smoking Distribution", col=c("skyblue","salmon"))
```
~~~

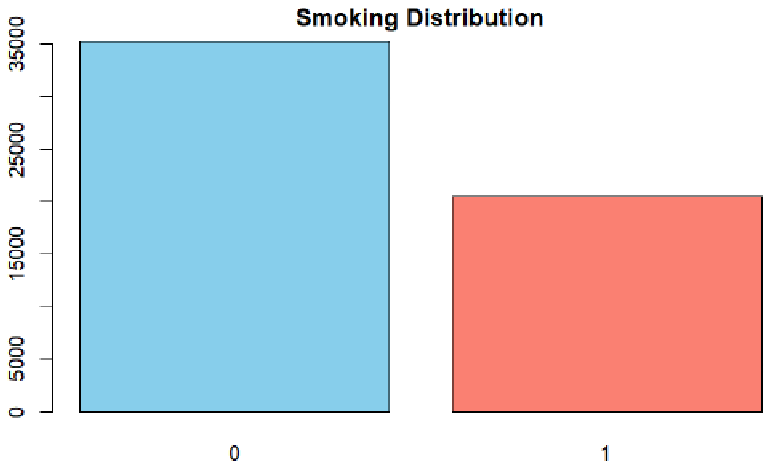
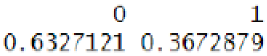


Figure 10 Class Distribution



```
        0         1 
0.6327121 0.3672879 
```

The barplot shows that the smoking class 0 and class 1 has a ratio of 0.63 to 0.37, which is still in an acceptable ratio.