اونيۏرسيتي مليسيا ڤهڠ السلطان عبد الله
UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

PROJECT
STATISTICAL MODELLING & SIMULATION (BSD3443)

NAME		**STUDENT ID**
1	TAN CHEK CHENG	SD21031
2	WONG ZI MING	SD21037
3	DAVID LAU KING LUEN	SD21044
4	TEAN JIN HE	SD21063
LECTURER	DR. NORYANTI MUHAMMAD	
DATE RELEASE	17 OCTOBER 2023 (TUESDAY)	
DURATION	17 OCTOBER 2023 – 22 JANUARY 2024	
SESSION/SEMESTER	SESSION 2023/2024 SEMESTER I	
METHOD OF SUBMISSION	Upload **R Code** (rfile) and **Output** as **report** (pdf) in KALAM	

Marks Distribution			
Course Outcomes	**Program Outcomes**	**Domain**	**Marks**
CLO2 Formulate statistical models for various problems in science, engineering, and industry. (C5, PLO2)	Critical Thinking & Scientific Approach	CTPS	**/ 10 (2%)**
CLO3 Manipulate statistical modelling theory and methodology in solving various applications using appropriate statistical software. (P4, PLO3)	Technical Skills	Psychomotor	**/ 15 (3%)**
CLO4: Demonstrate good interest and initiative for exploring issues in statistical modelling analysis for a given task (A3, PLO7)	Lifelong Learning	Affective	**/15 (2%)**
CLO5 Plan a business strategy by generating new ideas and innovation in the application of statistical modelling and simulation. (A3, PLO8)	Entrepreneur Skills	Affective	**/ 10 (3%)**
		Total	**/ 50**

BACKGROUND

Consider you are a Data Analyst at a certain department of organisation or company which objective to find and create a new or best solution to your organisation. Your organisation might consist of government sector, private company, industry, agency etc. Based on the data given, there are three main objectives you should focus on, which are:

1. To understand the data and variables by using any knowledge, method, and tools that you have learned.
2. To analyse the data and develop the best model which give the best solution to your organisation. Item you should do for the analysis;
    a. Investigate the basic model obtained from the data set.
    b. Develop the best model as your solution based on this course.

   c. Briefly explain your final model which consider as the best solution.

3. To plan the simple business of this task or assignment by considering developing an appropriate mobile apps/interactive dashboard/software or other appropriate device for the solution you obtained to be considered as your product.

Important Notes:

1. Please use the rubrics attached to guide your final report in order to get maximum marks.

APPENDIX: RUBRIC GUIDELINE FOR CLO2 (PLO2)

CRITERIA	LEVEL OF ACHIEVEMENT					WEIGHTAGE	SCORE (%)
	Inadequate **1**	**Emerging** **2**	**Developing** **3**	**Good** **4**	**Excellent** **5**		
Identifying main problem and stating the subsidiary aspect of the problem	Unable to identifies the main problem and fail to state the subsidiary aspect of the problem	identifies little the main problem and fail to state the subsidiary aspect of the problem	Identifies the main problem without stating the subsidiary aspect of the problem	Identifies the main problem and considering little stating the subsidiary aspect of the problem	Identifies and clearly states both the main and subsidiary aspect of the problem	0.1	
Collecting and analysing information	Collects inadequate viable information unable to perform analysis	Collects little viable information unable to perform analysis	Collects adequate information and performs analysis based on two sources	Collects adequate information and performs analysis based on more sources	Collects abundance of information and performs analysis based on multiple resources and justify.	0.1	
Identifying potential solutions	Identifies a single solution yet fail to present reasoning	Identifies two solutions yet fail to present reasoning	Identifies a few simple solutions and simple reasoning for the suggested solutions	Identifies more simple solutions and simple reasoning for the suggested solutions	Identifies and explains - accurately and thoroughly- multiple solutions and perspectives	0.1	
Selecting the most appropriate solution based on appropriate data or theory and provide alternative approach	Do not select any solutions that does not meet the required specifications.	Selects a solution that does not meet the required specifications.	Selects a reasonable solution but does not justify the solution. No discussion of alternate approaches included	Selects a reasonable solution and justify the solution. No discussion of alternate approaches included	Selects and articulates a solution based on appropriate data and discuss alternative approaches.	0.1	
						TOTAL	**/2**

APPENDIX: RUBRIC GUIDELINE FOR CLO3 (PLO3)

CRITERIA	LEVEL OF ACHIEVEMENT					WEIGHTAGE	SCORE (%)
	Inadequate **1**	**Emerging** **2**	**Developing** **3**	**Good** **4**	**Excellent** **5**		
Theory/ Knowledge	Very little knowledge provided, or information is incorrect	Some knowledge or information provided but missing all major points	Some knowledge or information provided but still missing some major points	Good knowledge observed, missing some minor points	Excellent knowledge observed; provides all necessary background principles	0.2	
Measurement/ Techniques/ Data Validation	Inappropriate measurement techniques are demonstrated	Partly correct measurement techniques are demonstrated, with partly valid data	Correct measurement techniques are demonstrated, with partly valid data	Good measurement techniques are demonstrated, with valid but not accurate data	Competent measurement techniques are demonstrated, with valid and accurate data	0.2	
Results	Lack of results / zero readability of the result. Poor originality, taking credits of others work	Partly complete result	Result presented but at low readability / some result presented. Reader has to guess some of the missing information. Less originality, copy paste here and then	Clear, neat presentation. All required results are presented. Readability. Complete with labels, title, axes, etc.	Very Clear, neat presentation. All required results are presented. High readability. Complete with labels, title, axes, etc.	0.2	
						TOTAL	**/3**

APPENDIX: RUBRIC GUIDELINE FOR CLO4 (PLO7)

Criteria	LEVEL OF ACHIEVEMENT					WEIGHTAGE	SCORE (%)
	Very Weak 1	Weak 2	Fair 3	Good 4	Very Good 5		
Interest	No interest in exploring issues for a given task	Demonstrate limited interest in exploring issues for a given task	Demonstrate sufficient interest in exploring issues for a given task	Demonstrate good interest for exploring issues for a given task	Demonstrate excellent interest in exploring issues for a given task	0.2	
Optimisation	Not able to retrieve information	Able to retrieve Information from minimal references	Able to retrieve information from sufficient reference	Able to retrieve information from many references	Able to retrieve Information from Maximum references	0.2	
						TOTAL	/2

APPENDIX: RUBRIC GUIDELINE FOR CLO5 (PLO8)

Criteria	LEVEL OF ACHIEVEMENT					WEIGHTAGE	SCORE (%)
	Very Weak 1	Weak 2	Fair 3	Good 4	Very Good 5		
Vision	No vision to solve problem.	Minimal vision to solve problem.	Satisfactory vision to solve problem.	Good vision to solve problem.	Excellent vision to solve problem.	0.2	
Passionate	Dislike to organise an entrepreneurial activity.	Minimal liking to organise an entrepreneurial activity.	Satisfactory liking and enjoys organising an entrepreneurial activity.	Likes and enjoys organising an entrepreneurial activity.	Passionate to organise an entrepreneurial activity.	0.2	
Entrepreneurial Opportunity	No entrepreneurial idea for value adding/solving customer needs.	Has unclear entrepreneurial idea for value adding/ solving customer needs and is not relevant to customer needs.	Business idea is clear but does not fulfil the realistic customer needs.	Business idea is clear and fulfils the customer needs.	Able to mobilise the idea to become opportunity according to the business strategy and fulfil the customer needs.	0.2	
						TOTAL	/3

TABLE OF CONTENT

1.0 INTRODUCTION

1.1 Case Study

Smoking is the process of burning a material—typically tobacco, cannabis, or opium—and then inhaling the smoke to become high. Depending on the material and the amount smoked, smoking can have either energizing or calming effects on the body. In addition to being extremely addictive, smoking is bad for your health since it increases your risk of developing heart disease, stroke, lung cancer, and chronic obstructive pulmonary disease (COPD). Many negative effects of smoking on health have been demonstrated. It has been discovered that smoking shortens the life expectancy of smokers overall and damages almost all of the body's organs. As of 2018, smoking is still a major contributor to avoidable morbidity and mortality worldwide, posing a persistent threat to global health. The World Health Organization analysis estimates that by 2030, smoking-related fatalities will account for 10 million deaths.

Evidence-based treatments to assist smokers in quitting have been proposed and promoted. Less than one-third of the participants were able to maintain sobriety, though. Many doctors felt that smoking cessation therapy was time-consuming and useless, thus they did not routinely offer it in their daily practices. The degree of nicotine dependence, the amount of carbon monoxide (CO) exhaled, the number of cigarettes smoked daily, the age at which smoking began, the history of failed quit attempts, marital status, emotional distress, temperament and impulsivity scores, and the desire to give up the habit are some of the variables that have been proposed to address this issue and determine which smokers would be more likely to quit successfully. However, when these characteristics are used individually for prediction, the results can be inconsistent and challenging for physicians and patients to comprehend and apply. Offering a prediction model could be a useful approach to comprehending each smoker's likelihood of stopping. In recent years, machine learning techniques have been used to construct health outcome prediction models.

1.2 Objectives

1. To develop a full model and do the comparison with a reduced model in order to choose the best prediction model used.
2. To predict the smoking status by considering various health-related variables
3. To visualize the prediction result.

2.0 DATA DESCRIPTION

Dataset: https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction/data

The data has been provided with the smoker status prediction and contains the following fields:

Attribute	Data Type	Description
age	Integer	Age of the individual in years
height.cm.	Integer	Height of the individual in centimeters
weight.kg.	Integer	Weight of the individual in kilograms
waist.cm.	Numeric	Waist circumference length in centimeters
eyesight.left.	Numeric	Eyesight in the left eye
eyesight.right.	Numeric	Eyesight in the right eye
hearing.left.	Integer	Hearing status in the left ear (1: normal, 2: impaired)
hearing.right.	Integer	Hearing status in the right ear (1: normal, 2: impaired)
systolic	Integer	Systolic blood pressure
relaxation	Integer	Diastolic blood pressure
fasting.blood.sugar	Integer	Fasting blood sugar level
Cholesterol	Integer	Total cholesterol level
triglyceride	Integer	Triglyceride level
HDL	Integer	High-density lipoprotein (HDL) cholesterol level
LDL	Integer	Low-density lipoprotein (LDL) cholesterol level
hemoglobin	Numeric	Hemoglobin level
Urine.protein	Integer	Urine protein status (1: present, 2: absent)
serum.creatinine	Numeric	Serum creatinine level
AST	Integer	Aspartate transaminase (AST) level

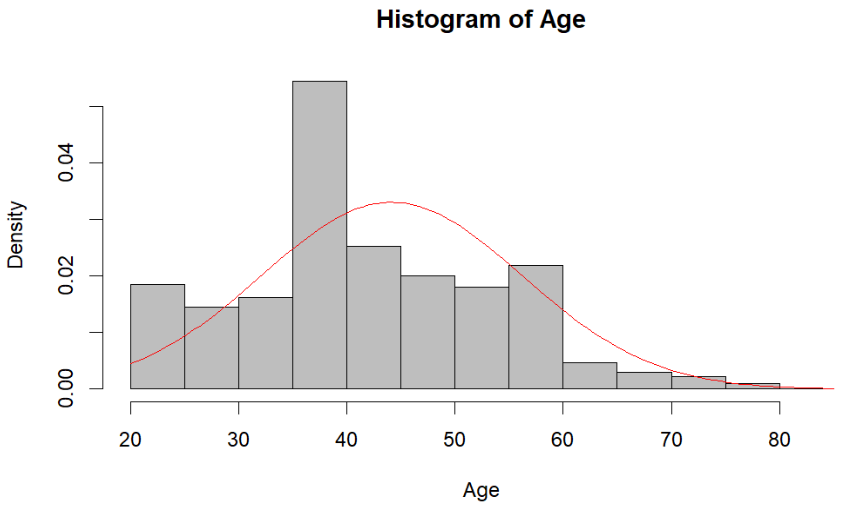
ALT	Integer	Alanine transaminase (ALT) level
Gtp	Integer	Gamma-glutamyl transferase (GTP) level
dental.caries	Integer	Dental caries status (1: present, 0: absent)
smoking	Integer	Smoking status (1: smoker, 0: non-smoker)

3.0 DATA ANALYSIS

3.1 Attributes Distributions

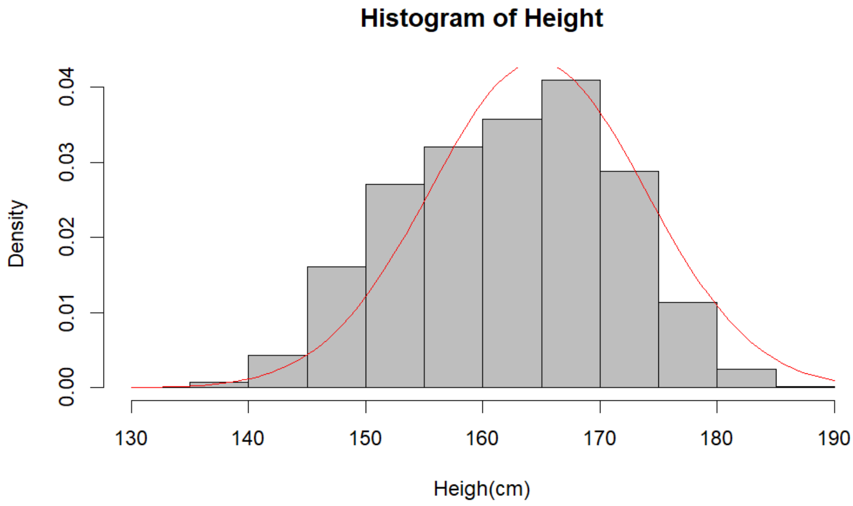
The attributes distributions shown below are followed by the significant attributes.

3.1.1 Age



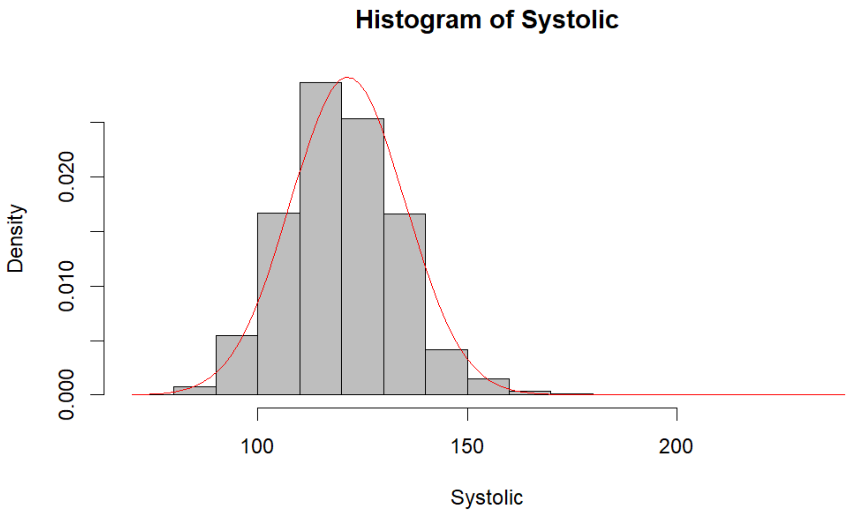
Based on the histogram above, the 'Age' distribution is approximately normally distributed since it is symmetric bell-shaped with the highest point at the mean. The most common age is around 40, as indicated by the highest point on the histogram and density curve.

3.1.2 Height



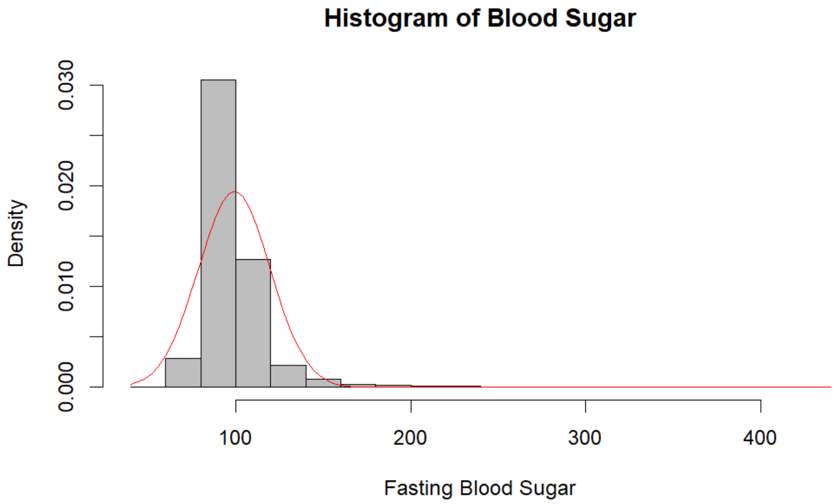
Based on the histogram above, the 'Height' distribution is approximately normally distributed since it is symmetric bell-shaped with the highest point at the mean. The most common height range is around 160-170 cm, as indicated by the highest point on the histogram and density curve. This means that most individuals in this dataset have a height within this range.

3.1.3 Systolic



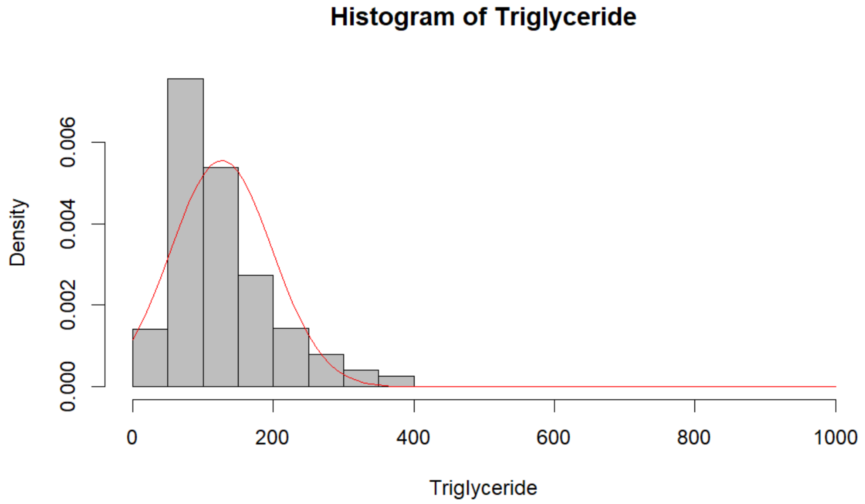
Based on the histogram above, the 'Systolic' distribution is approximately normally distributed since it is symmetric bell-shaped with the highest point at the mean. The most common systolic blood pressure is around 120, as indicated by the highest point on the histogram and density curve.

3.1.4 Blood Sugar



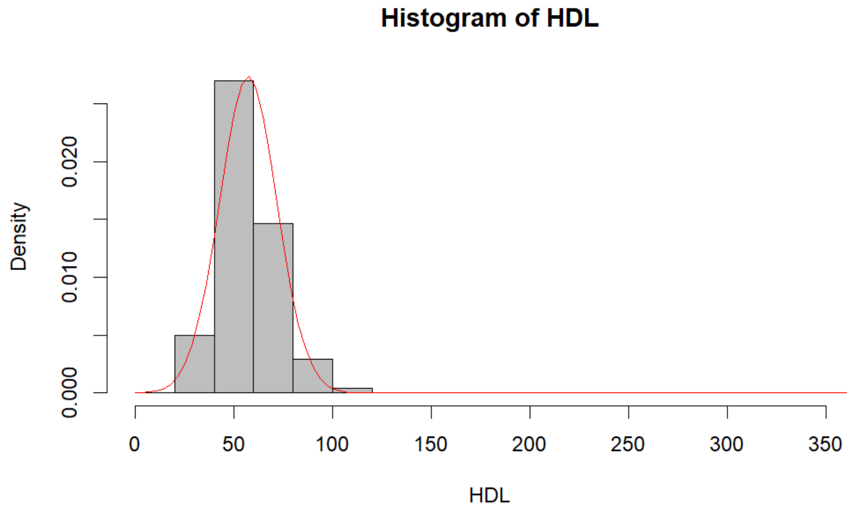
Based on the histogram above, the 'Blood Sugar' distribution is approximately skewed to the right with a peak at 100. This means that most individuals in this dataset have a fasting blood sugar level of around 100.

3.1.5 Triglyceride



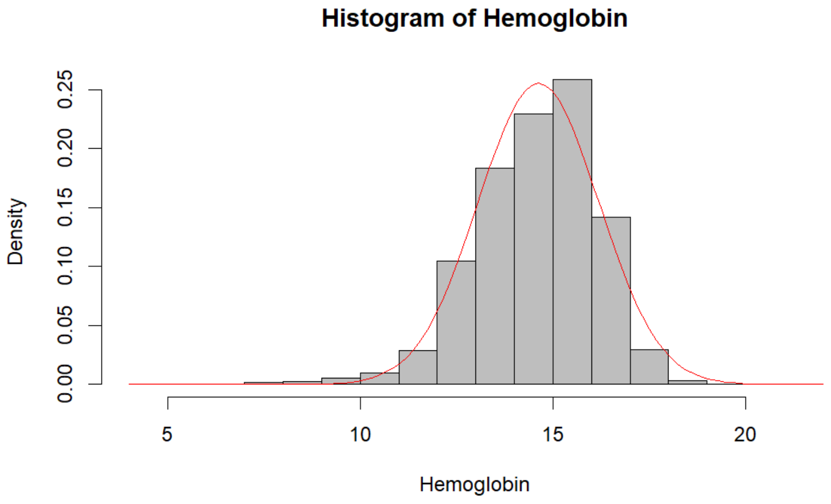
From the histogram and density curve, we can see that the distribution of triglyceride levels is skewed to the right, with a peak around 0-200. This means that most individuals in this dataset have low triglyceride levels, but there are also a significant number of individuals with higher levels.

3.1.6 HDL



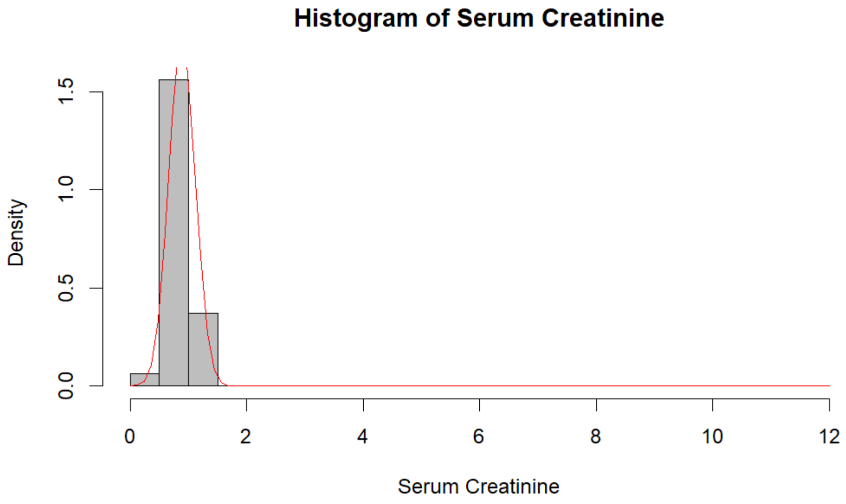
From the histogram and density curve, we can see that the distribution of HDL levels is skewed to the right, with a peak around 50-100. This means that most individuals in this dataset have their HDL levels within this range, but there are also a significant number of individuals with higher levels.

3.1.7 Hemoglobin



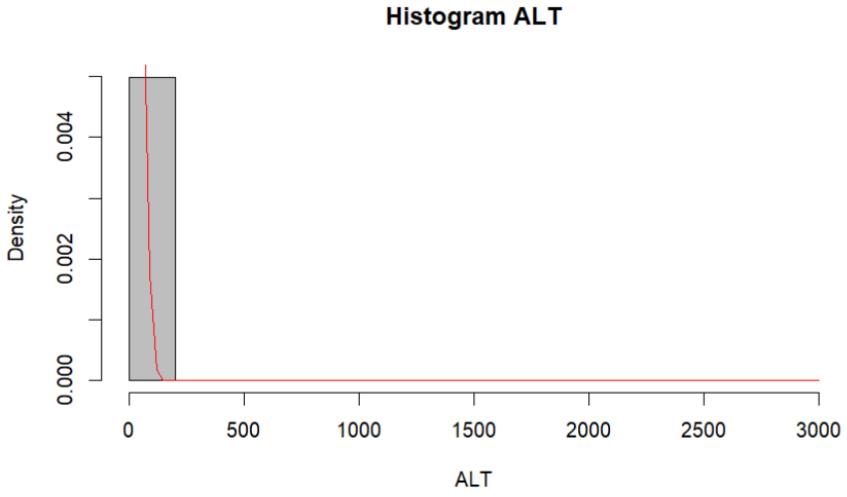
From the histogram and density curve, we can see that the distribution of hemoglobin levels is skewed to the left, with a peak around 15. This means that most individuals in this dataset have their hemoglobin levels within this range.

3.1.8 Serum Creatinine



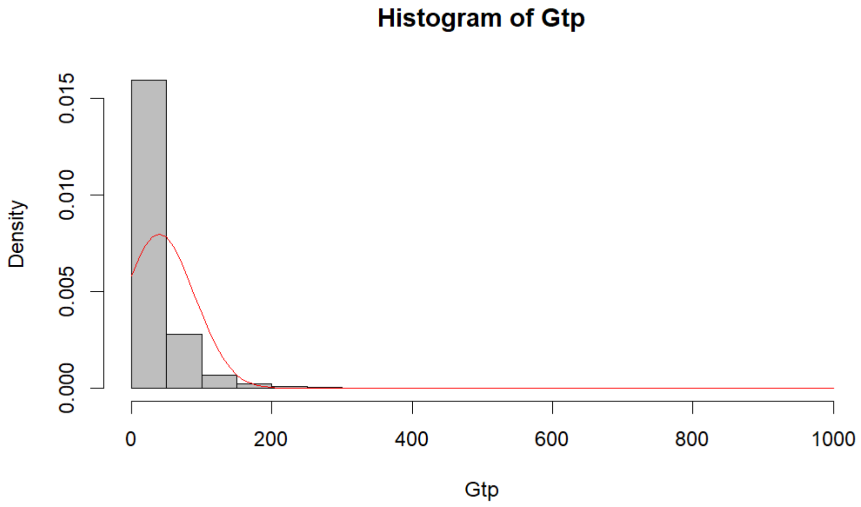
From the histogram and density curve, we can see that the distribution of serum creatinine levels is skewed to the right, with a peak around 1-2. This means that most individuals in this dataset have their serum creatinine levels within this range, but there are also a significant number of individuals with higher levels.

3.1.9 ALT



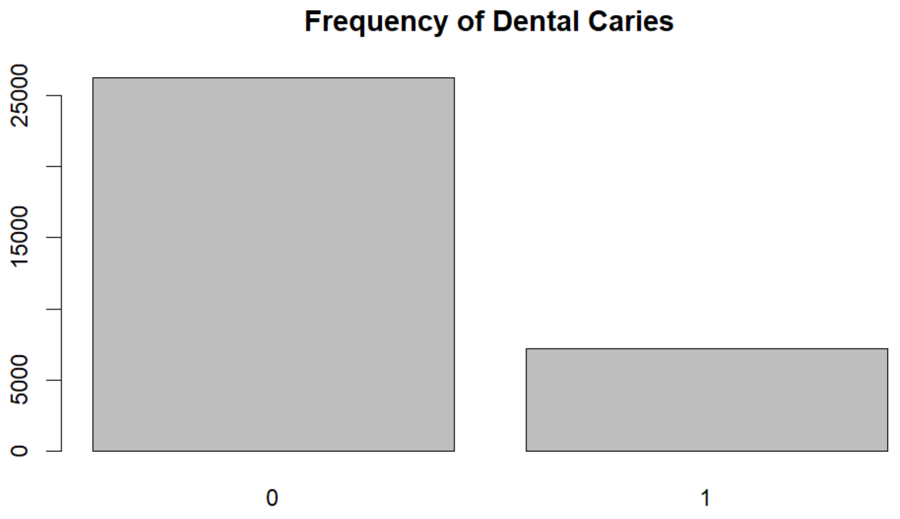
From the histogram and density curve, we can see that the distribution of ALT levels is skewed to the right, with a peak around 0-500. This means that most individuals in this dataset have their ALT levels within this range, but there are also a significant number of individuals with higher levels.

3.1.10 Gtp



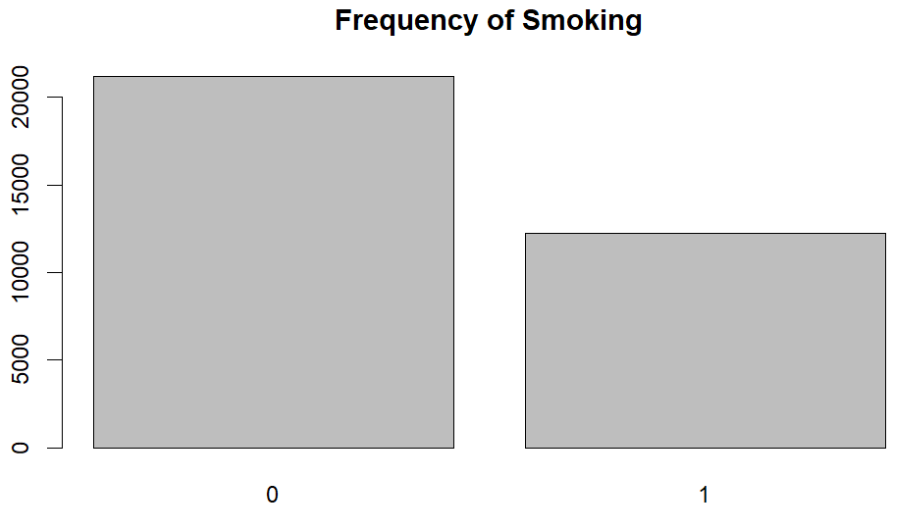
From the histogram and density curve, we can see that the distribution of Gtp levels is skewed to the right, with a peak around 0-200. This means that most individuals in this dataset have their Gtp levels within this range, but there are also a significant number of individuals with higher levels.

3.1.11 Dental Caries



Category "0" represents no dental caries while category "1" represents dental caries. From the bar graph, we can see that the category "0" has a significantly higher frequency of dental caries than the category "1" with around 25000 counts while category "1" has a frequency of around 7500 counts. This means that most of the people do not have dental caries.

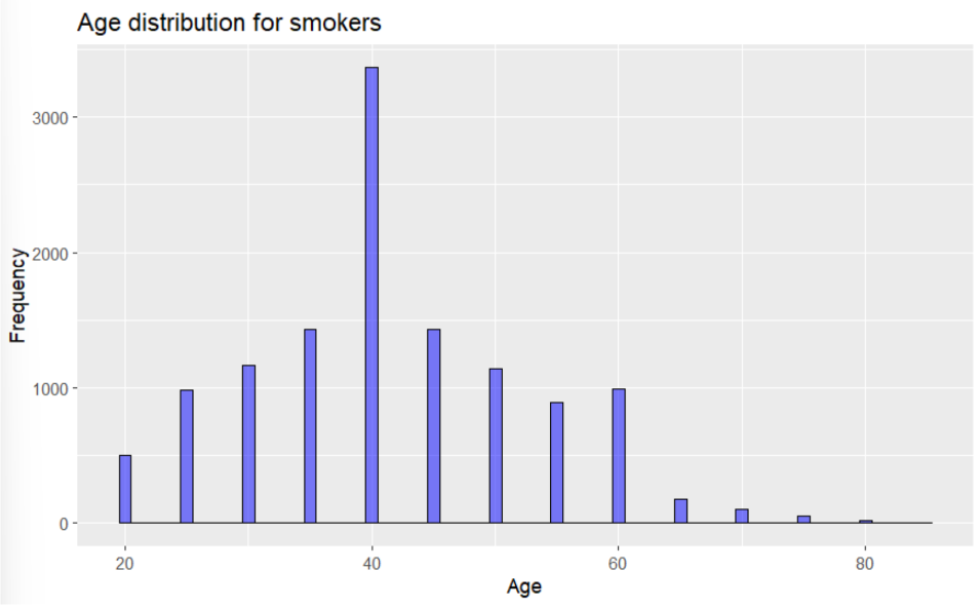
3.1.12 Smoking

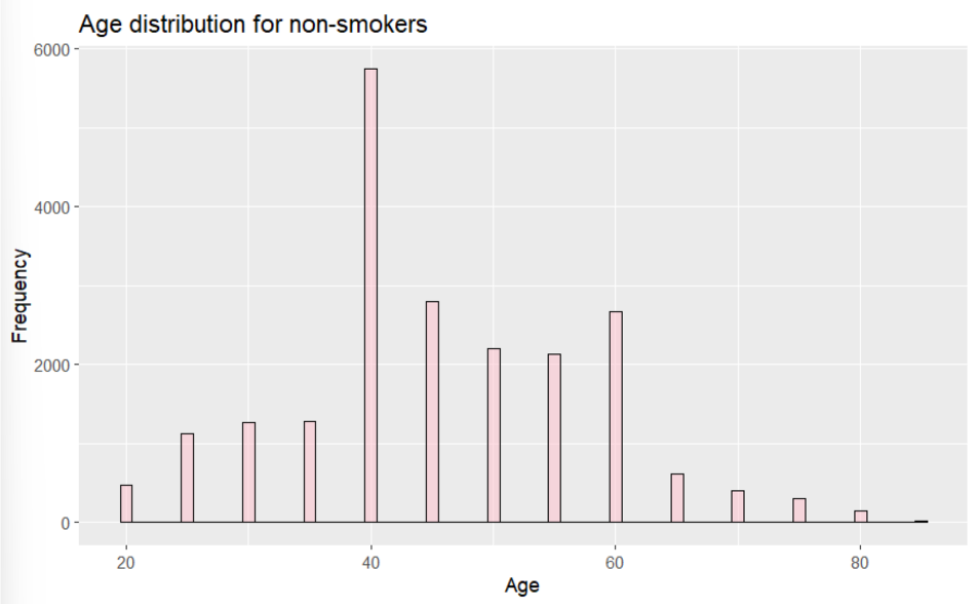


Category "0" represents no smoking while category "1" represents smoking. From the bar graph, we can see that the category "0" has a significantly higher frequency of smoking than the category "1" with around 20000 counts while category "1" has a frequency of around 12500 counts. This means that most of the people are non-smokers.

3.2 Exploratory Analysis and Visualization

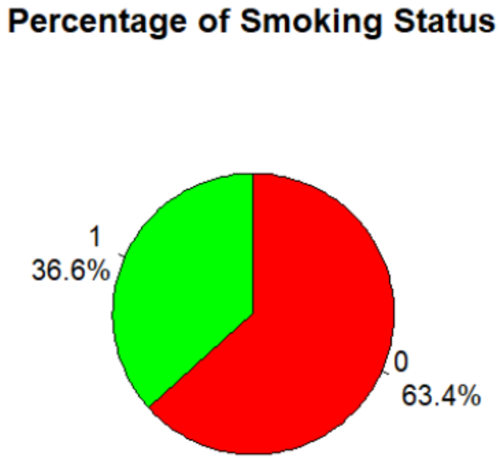
3.2.1 Age Distribution for Smokers and Non-smokers





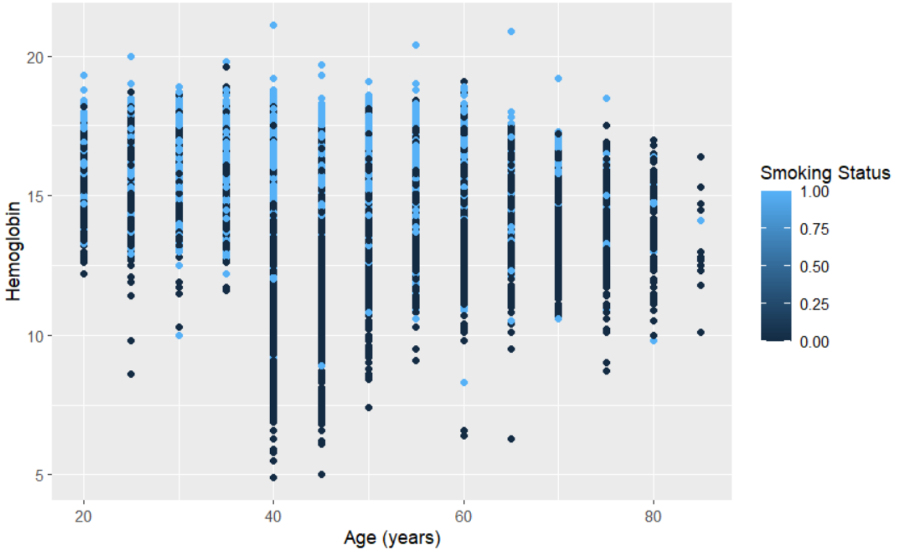
Based on both the histograms, the frequency of smokers and non-smokers is the highest at age 40 while the lowest is at age 80 respectively.

3.2.2 Percentage of Smoking Status



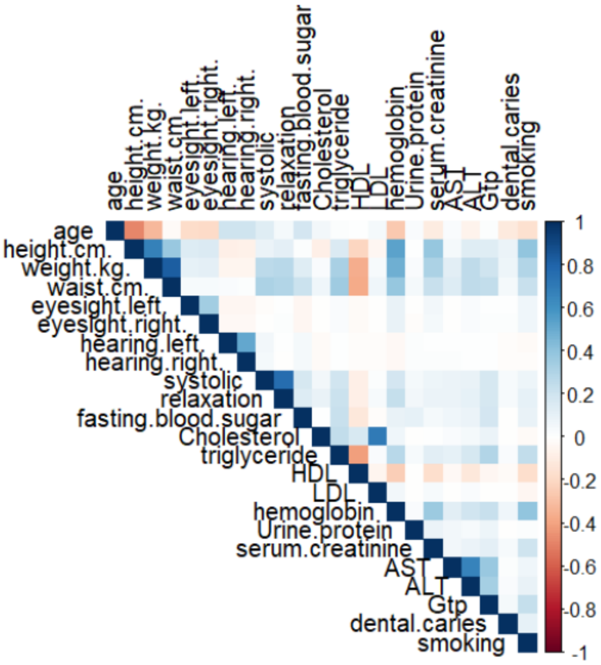
Based on the pie chart, non-smokers have a higher percentage of smoking status of 63.4% while another 36.6% are smokers.

3.2.3 Hemoglobin vs Age



Based on this scatter plot, we can see that the light blue colour represents the person who is a smoker and the darker blue colour represents the person who is not a smoker. Based on the scatter plot, people who are older are likely not a smoker. Moreover, people who smoke have an overly high level of hemoglobin. This is because the carbon monoxide in cigarette smoke blocks oxygen attachment to the red cells' empty hemoglobin slots, causing the body to increase red blood cell production.

3.2.4 Correlation Matrix among the Attributes

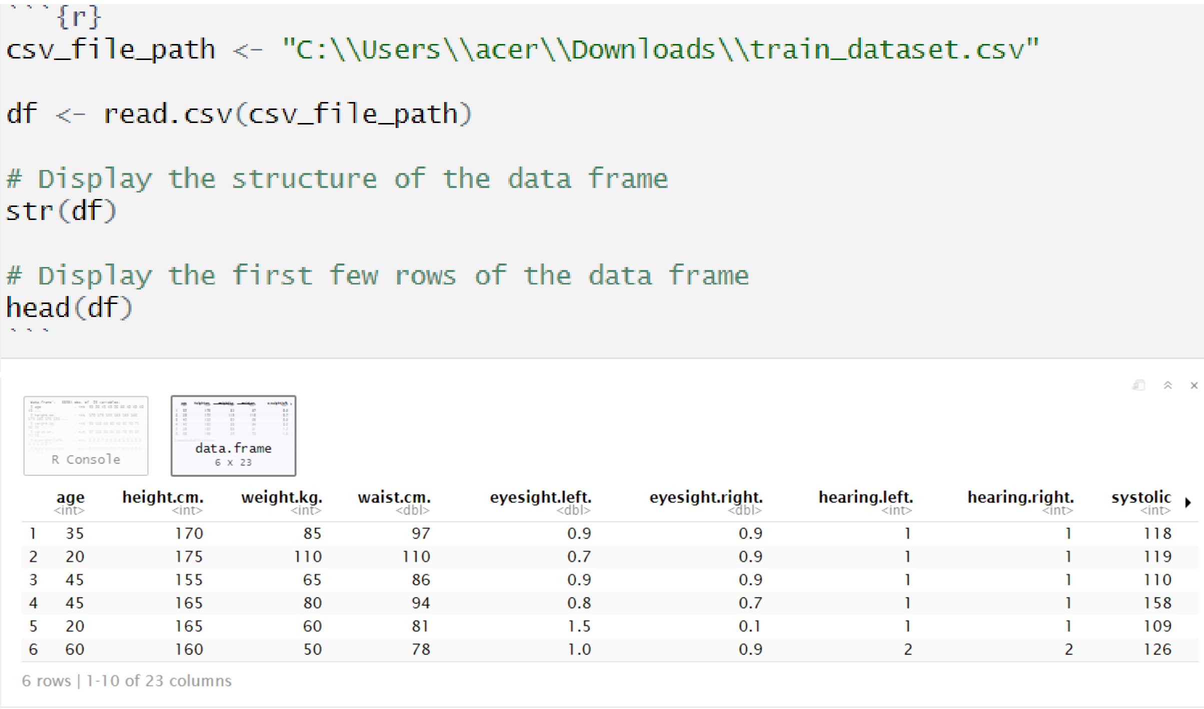


From the heatmap, the darker blue colour represents the strong correlation between the variables. There appears to be a number that shows a negative number, which means that when the number is negative, the data has a relationship that is rather than reversed. If the number is approaching zero, the variables do not correlate at all.

Based on the heatmap, when investigating the correlation between the response variable smoking and the explanatory variables, the explanatory variables that have a weak positive correlation with the highest to the response variable smoking among the attributes in this dataset are height.cm., weight.kg., waist,cm., triglyceride, hemoglobin, serum.creatinine and Gtp.

4.0 DATA MODELLING

Extract and load the dataset into R studio



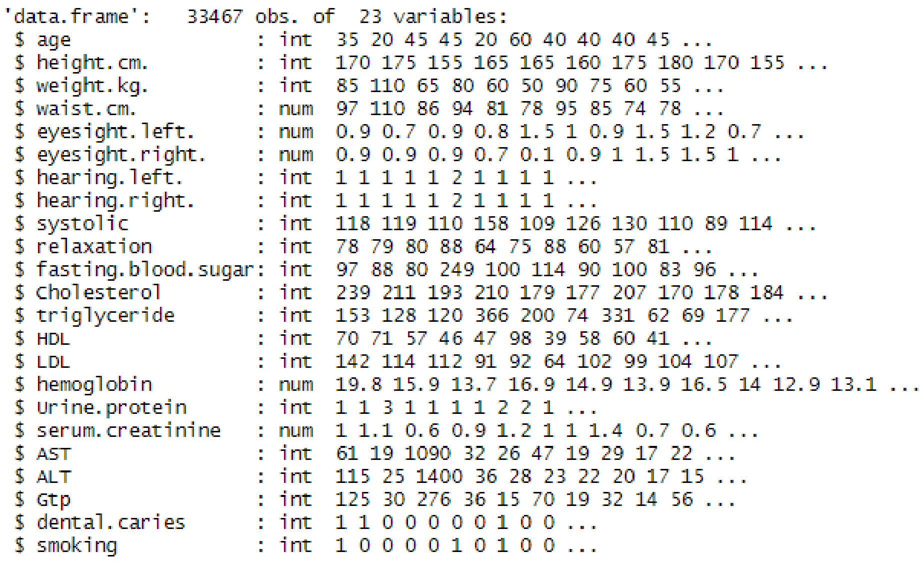
~~~
```{r}
csv_file_path <- "C:\\Users\\acer\\Downloads\\train_dataset.csv"

df <- read.csv(csv_file_path)

# Display the structure of the data frame
str(df)

# Display the first few rows of the data frame
head(df)
```
~~~

	age	height.cm.	weight.kg.	waist.cm.	eyesight.left.	eyesight.right.	hearing.left.	hearing.right.	systolic
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<int>	<int>	<int>
1	35	170	85	97	0.9	0.9	1	1	118
2	20	175	110	110	0.7	0.9	1	1	119
3	45	155	65	86	0.9	0.9	1	1	110
4	45	165	80	94	0.8	0.7	1	1	158
5	20	165	60	81	1.5	0.1	1	1	109
6	60	160	50	78	1.0	0.9	2	2	126

6 rows | 1-10 of 23 columns

Figure 4.1 Loading Data

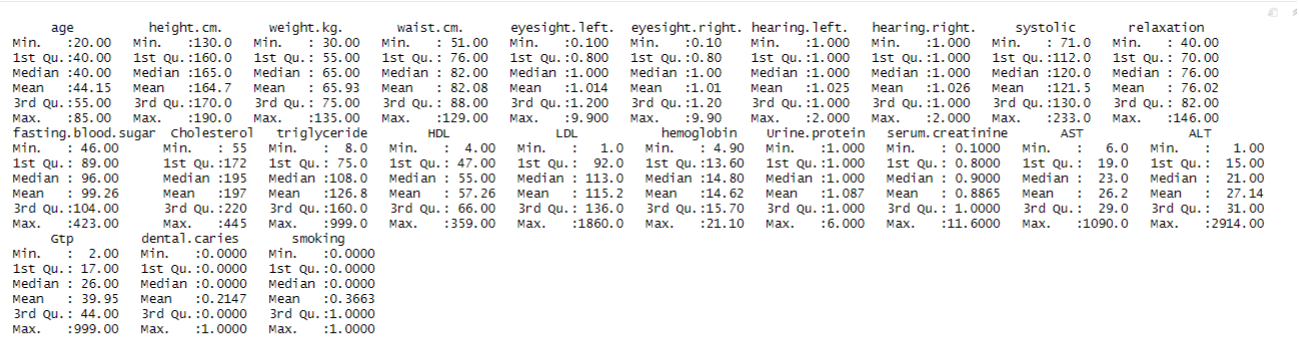
The dataset is loaded into R studio. The figure above shows the head of the data frame imported



```
'data.frame':	33467 obs. of  23 variables:
 $ age                : int  35 20 45 45 20 60 40 40 40 45 ...
 $ height.cm.         : int  170 175 155 165 165 160 175 180 170 155 ...
 $ weight.kg.         : int  85 110 65 80 60 50 90 75 60 55 ...
 $ waist.cm.          : num  97 110 86 94 81 78 95 85 74 78 ...
 $ eyesight.left.     : num  0.9 0.7 0.9 0.8 1.5 1 0.9 1.5 1.2 0.7 ...
 $ eyesight.right.    : num  0.9 0.9 0.9 0.7 0.1 0.9 1 1.5 1.5 1 ...
 $ hearing.left.      : int  1 1 1 1 1 2 1 1 1 1 ...
 $ hearing.right.     : int  1 1 1 1 1 2 1 1 1 1 ...
 $ systolic           : int  118 119 110 158 109 126 130 110 89 114 ...
 $ relaxation         : int  78 79 80 88 64 75 88 60 57 81 ...
 $ fasting.blood.sugar: int  97 88 80 249 100 114 90 100 83 96 ...
 $ Cholesterol        : int  239 211 193 210 179 177 207 170 178 184 ...
 $ triglyceride       : int  153 128 120 366 200 74 331 62 69 177 ...
 $ HDL                : int  70 71 57 46 47 98 39 58 60 41 ...
 $ LDL                : int  142 114 112 91 92 64 102 99 104 107 ...
 $ hemoglobin         : num  19.8 15.9 13.7 16.9 14.9 13.9 16.5 14 12.9 13.1 ...
 $ Urine.protein      : int  1 1 3 1 1 1 1 2 2 1 ...
 $ serum.creatinine   : num  1 1.1 0.6 0.9 1.2 1 1 1.4 0.7 0.6 ...
 $ AST                : int  61 19 1090 32 26 47 19 29 17 22 ...
 $ ALT                : int  115 25 1400 36 28 23 22 20 17 15 ...
 $ Gtp                : int  125 30 276 36 15 70 19 32 14 56 ...
 $ dental.caries      : int  1 1 0 0 0 0 0 1 0 0 ...
 $ smoking            : int  1 0 0 0 0 1 0 1 0 0 ...
```

Figure 4.2 Data structure

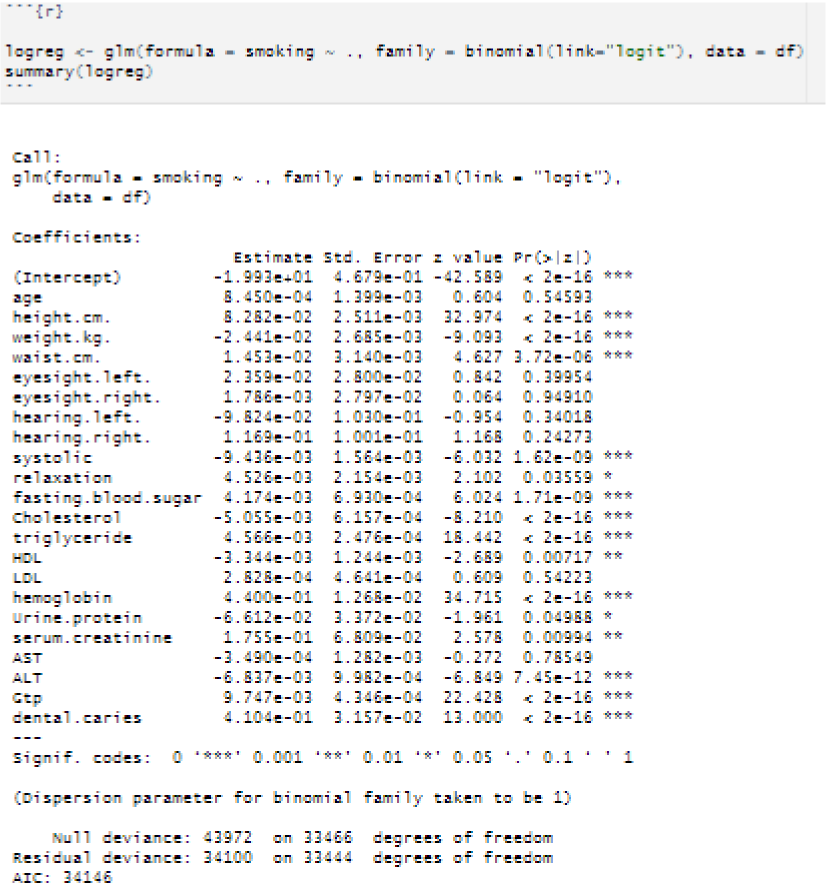
The figure shows all the attribute names within the dataframe and their data type respectively.



```
      age          height.cm.      weight.kg.        waist.cm.      eyesight.left.  eyesight.right.  hearing.left.   hearing.right.     systolic        relaxation   
 Min.   :20.00   Min.   :130.0   Min.   : 30.00   Min.   : 51.00   Min.   :0.100   Min.   :0.10    Min.   :1.000   Min.   :1.000   Min.   : 71.0   Min.   : 40.00  
 1st Qu.:40.00   1st Qu.:160.0   1st Qu.: 55.00   1st Qu.: 76.00   1st Qu.:0.800   1st Qu.:0.80    1st Qu.:1.000   1st Qu.:1.000   1st Qu.:112.0   1st Qu.: 70.00  
 Median :40.00   Median :165.0   Median : 65.00   Median : 82.00   Median :1.000   Median :1.00    Median :1.000   Median :1.000   Median :120.0   Median : 76.00  
 Mean   :44.15   Mean   :164.7   Mean   : 65.93   Mean   : 82.08   Mean   :1.014   Mean   :1.01    Mean   :1.025   Mean   :1.026   Mean   :121.5   Mean   : 76.02  
 3rd Qu.:55.00   3rd Qu.:170.0   3rd Qu.: 75.00   3rd Qu.: 88.00   3rd Qu.:1.200   3rd Qu.:1.20    3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:130.0   3rd Qu.: 82.00  
 Max.   :85.00   Max.   :190.0   Max.   :135.00   Max.   :129.00   Max.   :9.900   Max.   :9.90    Max.   :2.000   Max.   :2.000   Max.   :233.0   Max.   :146.00  
 fasting.blood.sugar  Cholesterol   triglyceride        HDL               LDL            hemoglobin    Urine.protein   serum.creatinine      AST               ALT         
 Min.   : 46.00      Min.   : 55   Min.   :  8.0   Min.   :  4.00   Min.   :   1.0   Min.   : 4.90   Min.   :1.000   Min.   : 0.1000   Min.   :   6.0   Min.   :   1.00  
 1st Qu.: 89.00      1st Qu.:172   1st Qu.: 75.0   1st Qu.: 47.00   1st Qu.:  92.0   1st Qu.:13.60   1st Qu.:1.000   1st Qu.: 0.8000   1st Qu.:  19.0   1st Qu.:  15.00  
 Median : 96.00      Median :195   Median :108.0   Median : 55.00   Median : 113.0   Median :14.80   Median :1.000   Median : 0.9000   Median :  23.0   Median :  21.00  
 Mean   : 99.26      Mean   :197   Mean   :126.8   Mean   : 57.26   Mean   : 115.2   Mean   :14.62   Mean   :1.087   Mean   : 0.8865   Mean   :  26.2   Mean   :  27.14  
 3rd Qu.:104.00      3rd Qu.:220   3rd Qu.:160.0   3rd Qu.: 66.00   3rd Qu.: 136.0   3rd Qu.:15.70   3rd Qu.:1.000   3rd Qu.: 1.0000   3rd Qu.:  29.0   3rd Qu.:  31.00  
 Max.   :423.00      Max.   :445   Max.   :999.0   Max.   :359.00   Max.   :1860.0   Max.   :21.10   Max.   :6.000   Max.   :11.6000   Max.   :1090.0   Max.   :2914.00  
      Gtp         dental.caries       smoking      
 Min.   :  2.00   Min.   :0.0000   Min.   :0.0000  
 1st Qu.: 17.00   1st Qu.:0.0000   1st Qu.:0.0000  
 Median : 26.00   Median :0.0000   Median :0.0000  
 Mean   : 39.95   Mean   :0.2147   Mean   :0.3663  
 3rd Qu.: 44.00   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :999.00   Max.   :1.0000   Max.   :1.0000  
```

Figure 4.3 Summary of dataframe

The provided summary depicts various statistical measures for each attribute in the dataset. It includes information such as the minimum and maximum values, quartiles, and the mean for all the attributes involved.

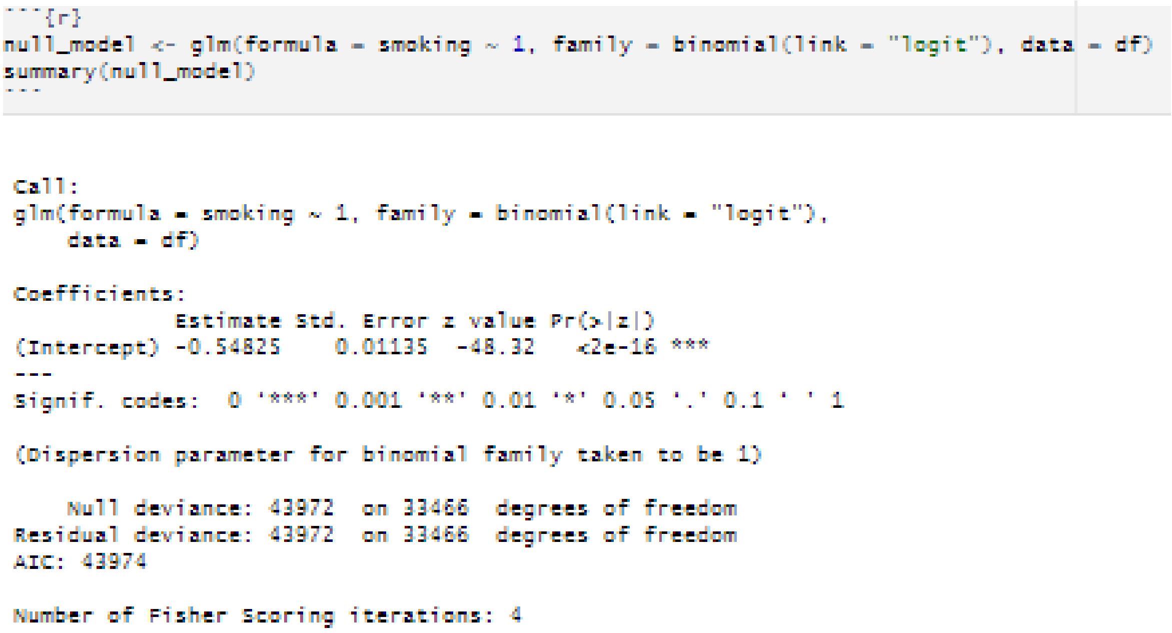


~~~
```{r}

logreg <- glm(formula = smoking ~ ., family = binomial(link="logit"), data = df)
summary(logreg)
```
~~~

```
Call:
glm(formula = smoking ~ ., family = binomial(link = "logit"), 
    data = df)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -1.993e+01  4.679e-01 -42.589  < 2e-16 ***
age                  8.450e-04  1.399e-03   0.604  0.54593    
height.cm.           8.282e-02  2.511e-03  32.974  < 2e-16 ***
weight.kg.          -2.441e-02  2.685e-03  -9.093  < 2e-16 ***
waist.cm.            1.453e-02  3.140e-03   4.627 3.72e-06 ***
eyesight.left.       2.359e-02  2.800e-02   0.842  0.39954    
eyesight.right.      1.786e-03  2.797e-02   0.064  0.94910    
hearing.left.       -9.824e-02  1.030e-01  -0.954  0.34018    
hearing.right.       1.169e-01  1.001e-01   1.168  0.24273    
systolic            -9.436e-03  1.564e-03  -6.032 1.62e-09 ***
relaxation           4.526e-03  2.154e-03   2.102  0.03559 *  
fasting.blood.sugar  4.174e-03  6.930e-04   6.024 1.71e-09 ***
Cholesterol         -5.055e-03  6.157e-04  -8.210  < 2e-16 ***
triglyceride         4.566e-03  2.476e-04  18.442  < 2e-16 ***
HDL                 -3.344e-03  1.244e-03  -2.689  0.00717 ** 
LDL                  2.828e-04  4.641e-04   0.609  0.54223    
hemoglobin           4.400e-01  1.268e-02  34.715  < 2e-16 ***
Urine.protein       -6.612e-02  3.372e-02  -1.961  0.04988 *  
serum.creatinine     1.755e-01  6.809e-02   2.578  0.00994 ** 
AST                 -3.490e-04  1.282e-03  -0.272  0.78549    
ALT                 -6.837e-03  9.982e-04  -6.849 7.45e-12 ***
Gtp                  9.747e-03  4.346e-04  22.428  < 2e-16 ***
dental.caries        4.104e-01  3.157e-02  13.000  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43972  on 33466  degrees of freedom
Residual deviance: 34100  on 33444  degrees of freedom
AIC: 34146
```

Figure 4.4 Full Logistic Model

All the attributes are used to fit into the basic logistic model with binomial distribution. The target variable is “smoking”. The figure above shows the null deviance and residual deviance. The AIC value is 34146



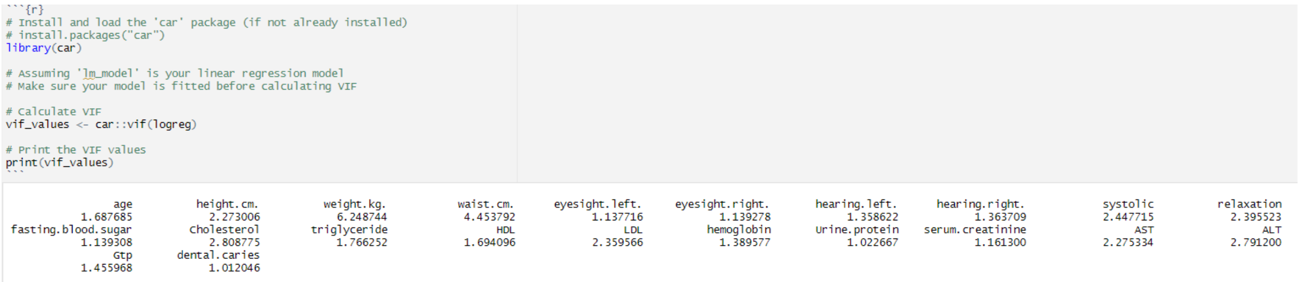
~~~
```{r}
null_model <- glm(formula = smoking ~ 1, family = binomial(link = "logit"), data = df)
summary(null_model)
```
~~~

```
Call:
glm(formula = smoking ~ 1, family = binomial(link = "logit"), 
    data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.54825    0.01135  -48.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43972  on 33466  degrees of freedom
Residual deviance: 43972  on 33466  degrees of freedom
AIC: 43974

Number of Fisher Scoring iterations: 4
```

Figure 4.5 Null Model

The figure above shows the null model with AIC value of 43974. The AIC (Akaike Information Criterion) is a measure that balances the goodness of fit of a model with its complexity, penalizing models that have more parameters. The goal is to find a model that fits the data well but is not overly complex. In the context of model selection criteria like AIC, lower values are preferred.

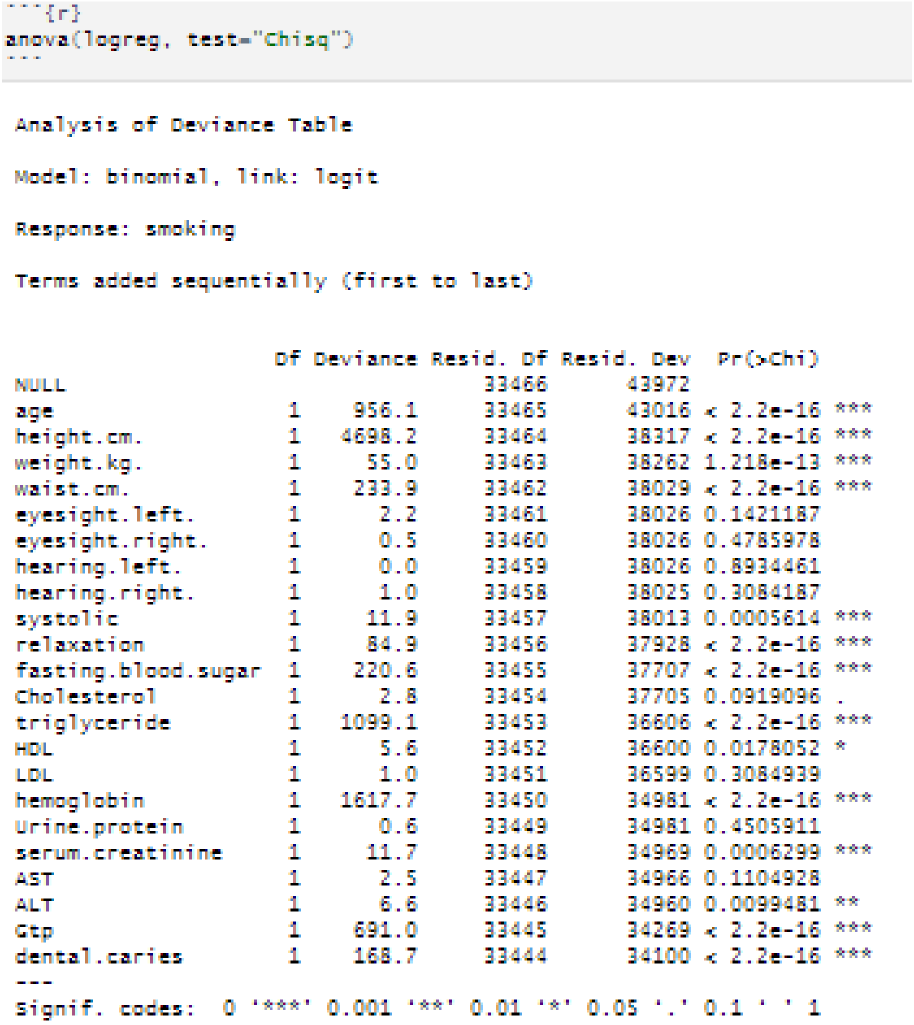
By comparing the two models, we can conclude that the null model does not offer a better fit as it has a higher AIC value than the full model.



~~~
```{r}
# Install and load the 'car' package (if not already installed)
# install.packages("car")
library(car)

# Assuming 'lm_model' is your linear regression model
# Make sure your model is fitted before calculating VIF

# Calculate VIF
vif_values <- car::vif(logreg)

# Print the VIF values
print(vif_values)
```
~~~

```
              age        height.cm.        weight.kg.         waist.cm.    eyesight.left.   eyesight.right.     hearing.left.    hearing.right.          systolic        relaxation 
         1.687685          2.273006          6.248744          4.453792          1.137716          1.139278          1.358622          1.363709          2.447715          2.395523 
fasting.blood.sugar       Cholesterol      triglyceride               HDL               LDL        hemoglobin     Urine.protein  serum.creatinine               AST               ALT 
         1.139308          2.808775          1.766252          1.694096          2.359566          1.389577          1.022667          1.161300          2.275334          2.791200 
              Gtp     dental.caries 
         1.455968          1.012046 
```

Figure 4.6 Multicollinearity Checking



~~~
```{r}
anova(logreg, test="Chisq")
```
~~~

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: smoking

Terms added sequentially (first to last)


                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                33466      43972              
age                  1    956.1     33465      43016 < 2.2e-16 ***
height.cm.           1   4698.2     33464      38317 < 2.2e-16 ***
weight.kg.           1     55.0     33463      38262 1.218e-13 ***
waist.cm.            1    233.9     33462      38029 < 2.2e-16 ***
eyesight.left.       1      2.2     33461      38026 0.1421187    
eyesight.right.      1      0.5     33460      38026 0.4785978    
hearing.left.        1      0.0     33459      38026 0.8934461    
hearing.right.       1      1.0     33458      38025 0.3084187    
systolic             1     11.9     33457      38013 0.0005614 ***
relaxation           1     84.9     33456      37928 < 2.2e-16 ***
fasting.blood.sugar  1    220.6     33455      37707 < 2.2e-16 ***
Cholesterol          1      2.8     33454      37705 0.0919096 .  
triglyceride         1   1099.1     33453      36606 < 2.2e-16 ***
HDL                  1      5.6     33452      36600 0.0178052 *  
LDL                  1      1.0     33451      36599 0.3084939    
hemoglobin           1   1617.7     33450      34981 < 2.2e-16 ***
Urine.protein        1      0.6     33449      34981 0.4505911    
serum.creatinine     1     11.7     33448      34969 0.0006299 ***
AST                  1      2.5     33447      34966 0.1104928    
ALT                  1      6.6     33446      34960 0.0099481 ** 
Gtp                  1    691.0     33445      34269 < 2.2e-16 ***
dental.caries        1    168.7     33444      34100 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.7 ANOVA test

############################ Results & Interpretation###########################

Significance Codes:

'***': Very highly significant (p-value < 0.001).

'**': Highly significant (p-value < 0.01).

'*': Significant at a 5% level (p-value < 0.05).

' ': Not significant (p-value > 0.1).

Results from the full logistic regression model show that there are some insignificant variables such as"age ", "eyesight.left.", "eyesight.right.", "hearing.left. ", "hearing.right. ", "Cholesterol ", "LDL ","AST " " and "relaxation" based on its respective p-values.

Variance Inflation Factors (VIF).

The Variance Inflation Factor (VIF) measures the inflation in the coefficient of the independent variable due to the collinearities among the other independent variables.

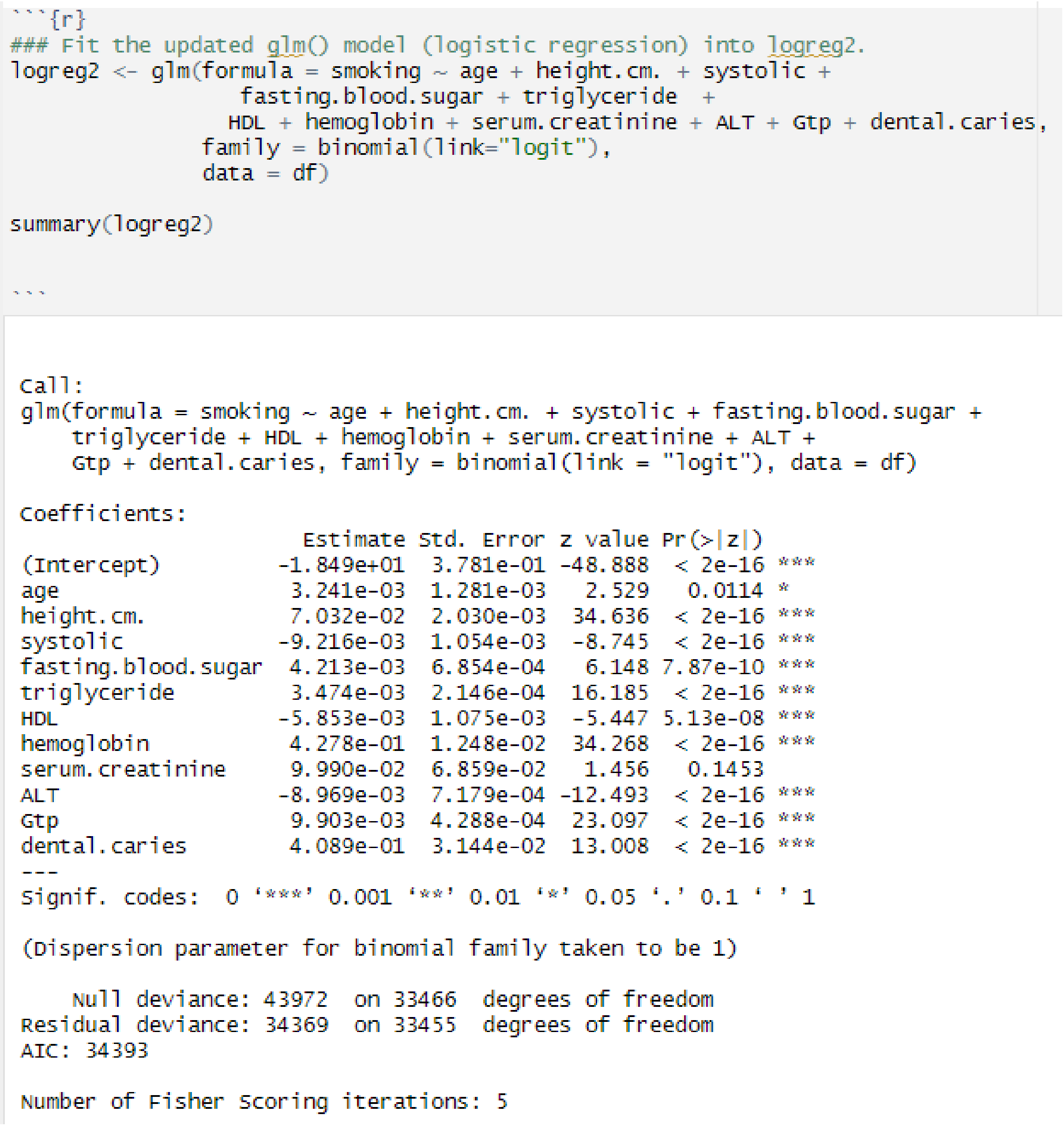
A VIF of 1 means that the regression coefficient is not inflated by the presence of the other predictors, and hence multicollinearity does not exist.

Ideally, the Variance Inflation Factors are below 5.

Results from multicollinearity with VIF test shows that weight.kg. (VIF: 6.23): This variable has a relatively high VIF, indicating that its variance is inflated due to its correlation with other predictors and waist.cm. (VIF: 4.43): While this VIF is above 2, it is not extremely high. However, it suggests some correlation with other predictors.

Results from the ANOVA test show that "eyesight.left.", "eyesight.right.", "hearing.left.   ", "hearing.right. ", "Cholesterol ", "LDL", "Urine.protein", and "AST " are insignificant to the fitted model.

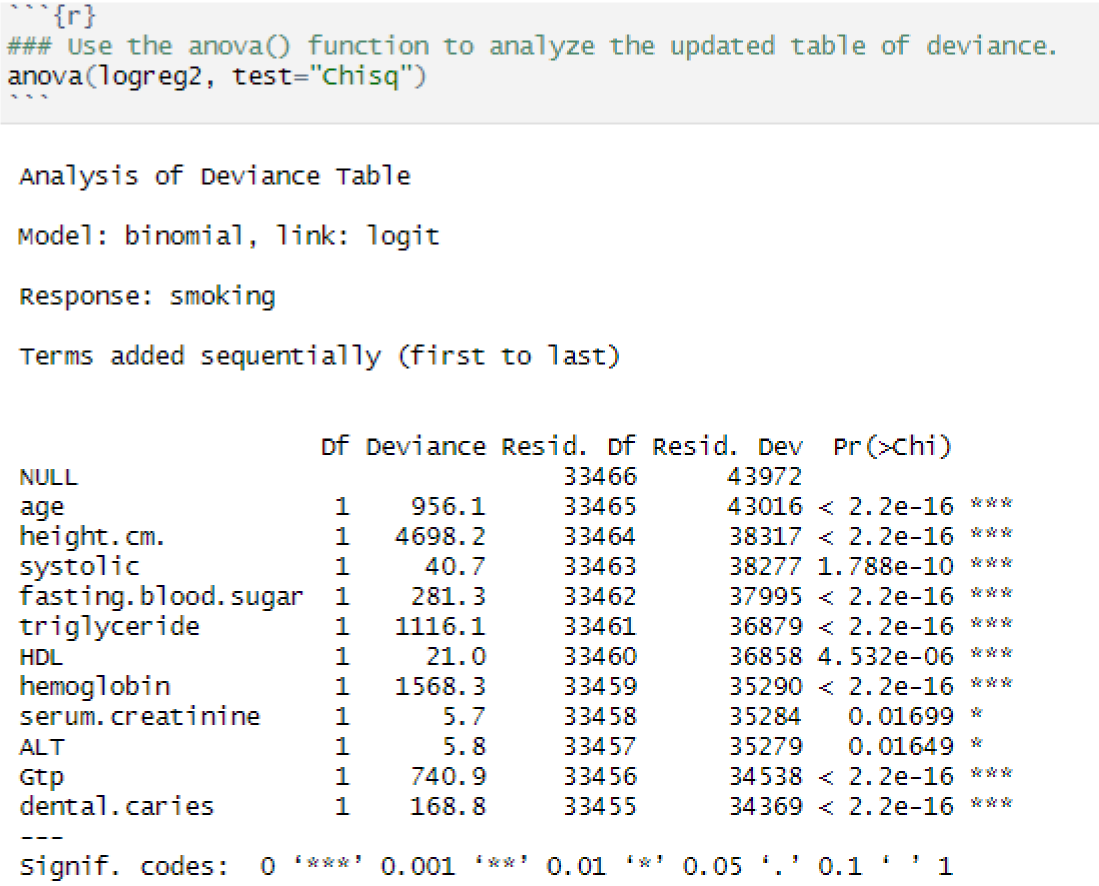
Hence, fit the updated glm() model (logistic regression) as a reduced model without the insignificant variables and non-collinear variables into logreg2 in order to develop the best model in our case study.



~~~
```{r}
### Fit the updated glm() model (logistic regression) into logreg2.
logreg2 <- glm(formula = smoking ~ age + height.cm. + systolic +
                  fasting.blood.sugar + triglyceride  +
                 HDL + hemoglobin + serum.creatinine + ALT + Gtp + dental.caries,
               family = binomial(link="logit"),
               data = df)

summary(logreg2)

```
~~~

```
Call:
glm(formula = smoking ~ age + height.cm. + systolic + fasting.blood.sugar +
    triglyceride + HDL + hemoglobin + serum.creatinine + ALT +
    Gtp + dental.caries, family = binomial(link = "logit"), data = df)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.849e+01  3.781e-01 -48.888  < 2e-16 ***
age                  3.241e-03  1.281e-03   2.529   0.0114 *
height.cm.           7.032e-02  2.030e-03  34.636  < 2e-16 ***
systolic            -9.216e-03  1.054e-03  -8.745  < 2e-16 ***
fasting.blood.sugar  4.213e-03  6.854e-04   6.148 7.87e-10 ***
triglyceride         3.474e-03  2.146e-04  16.185  < 2e-16 ***
HDL                 -5.853e-03  1.075e-03  -5.447 5.13e-08 ***
hemoglobin           4.278e-01  1.248e-02  34.268  < 2e-16 ***
serum.creatinine     9.990e-02  6.859e-02   1.456   0.1453
ALT                 -8.969e-03  7.179e-04 -12.493  < 2e-16 ***
Gtp                  9.903e-03  4.288e-04  23.097  < 2e-16 ***
dental.caries        4.089e-01  3.144e-02  13.008  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43972  on 33466  degrees of freedom
Residual deviance: 34369  on 33455  degrees of freedom
AIC: 34393

Number of Fisher Scoring iterations: 5
```

Figure 4.8 Reduced Model



~~~
```{r}
### Use the anova() function to analyze the updated table of deviance.
anova(logreg2, test="Chisq")
```
~~~

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: smoking

Terms added sequentially (first to last)


                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                33466      43972              
age                  1    956.1     33465      43016 < 2.2e-16 ***
height.cm.           1   4698.2     33464      38317 < 2.2e-16 ***
systolic             1     40.7     33463      38277 1.788e-10 ***
fasting.blood.sugar  1    281.3     33462      37995 < 2.2e-16 ***
triglyceride         1   1116.1     33461      36879 < 2.2e-16 ***
HDL                  1     21.0     33460      36858 4.532e-06 ***
hemoglobin           1   1568.3     33459      35290 < 2.2e-16 ***
serum.creatinine     1      5.7     33458      35284   0.01699 *  
ALT                  1      5.8     33457      35279   0.01649 *  
Gtp                  1    740.9     33456      34538 < 2.2e-16 ***
dental.caries        1    168.8     33455      34369 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4.9 Reduced Model ANOVA test

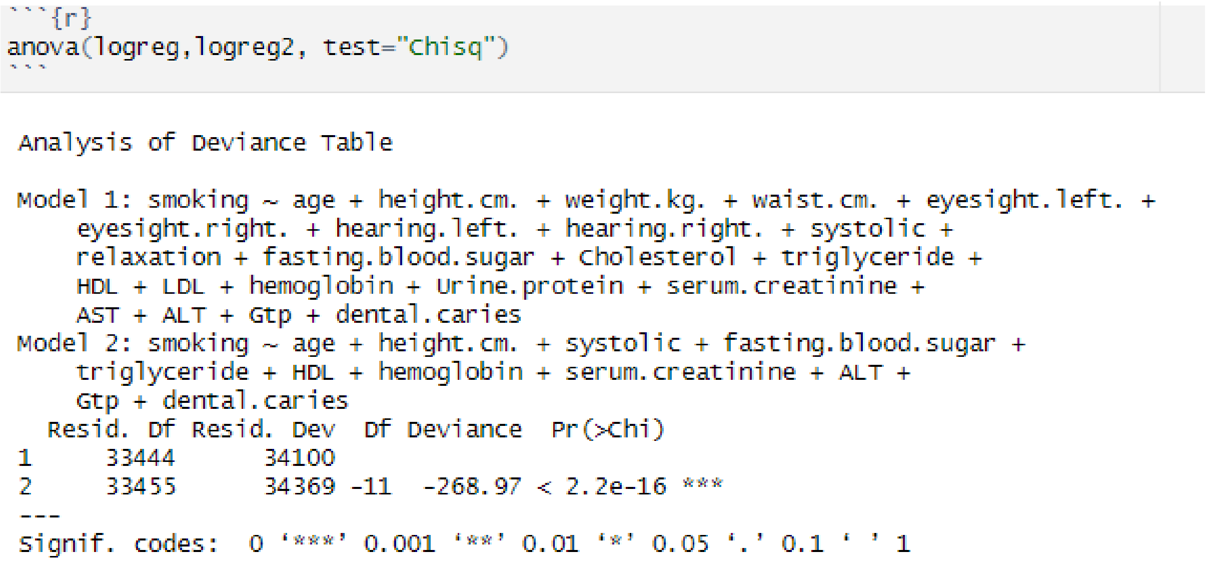
############################ Results & Interpretation #############################

Results from the updated fitted logistic regression model show that only "serum.creatinine" variable is insignificant based on its p-value.

Additionally, results from the ANOVA test show all variables, including "serum.creatinine ", are significant to the fitted model.

This is judged by the low deviance residuals as well as the Pr(>Chi) of > .05, respectively.

Hence, the "serum.creatinine " variable is kept in the fitted model.



~~~
```{r}
anova(logreg,logreg2, test="Chisq")
```
~~~

```
Analysis of Deviance Table

Model 1: smoking ~ age + height.cm. + weight.kg. + waist.cm. + eyesight.left. +
    eyesight.right. + hearing.left. + hearing.right. + systolic +
    relaxation + fasting.blood.sugar + Cholesterol + triglyceride +
    HDL + LDL + hemoglobin + Urine.protein + serum.creatinine +
    AST + ALT + Gtp + dental.caries
Model 2: smoking ~ age + height.cm. + systolic + fasting.blood.sugar +
    triglyceride + HDL + hemoglobin + serum.creatinine + ALT +
    Gtp + dental.caries
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1     33444      34100
2     33455      34369 -11  -268.97 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4.10 Comparison of full model and reduced model

H o: The full model offers better fit

H 1: The full model does not offer better fit

The p-value = 0..0000 and is less than α=0.05, *H*o is rejected.

Therefore, we can conclude that 0 the reduced model offers better fit than the full model

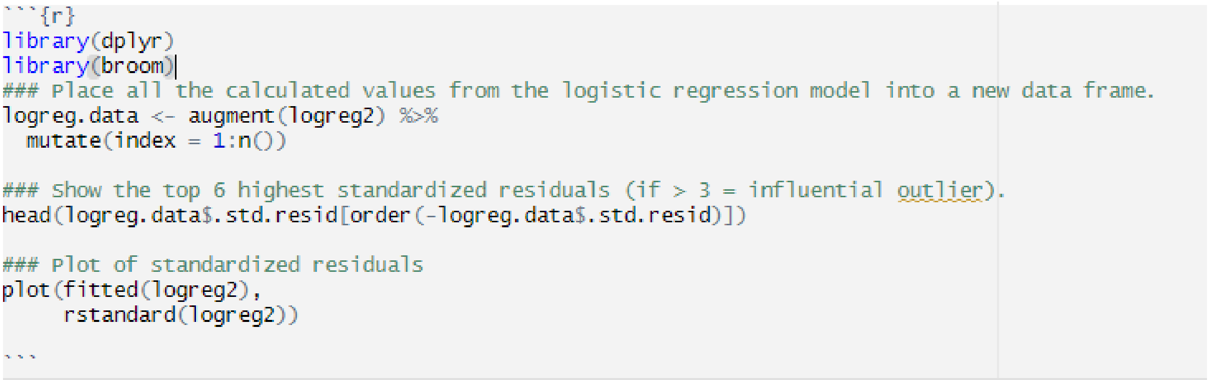
Test for the Absence of Strongly Influential Outliers

############################### Notes ###

Test using standardized residuals and Cook's Distance.

Standardized residual values > 3 = influential outlier.

Cook's D value > Cook's D Threshold (4/N) = influential outlier.



~~~
```{r}
library(dplyr)
library(broom)
### Place all the calculated values from the logistic regression model into a new data frame.
logreg.data <- augment(logreg2) %>%
  mutate(index = 1:n())

### Show the top 6 highest standardized residuals (if > 3 = influential outlier).
head(logreg.data$.std.resid[order(-logreg.data$.std.resid)])

### Plot of standardized residuals
plot(fitted(logreg2),
     rstandard(logreg2))

```
~~~

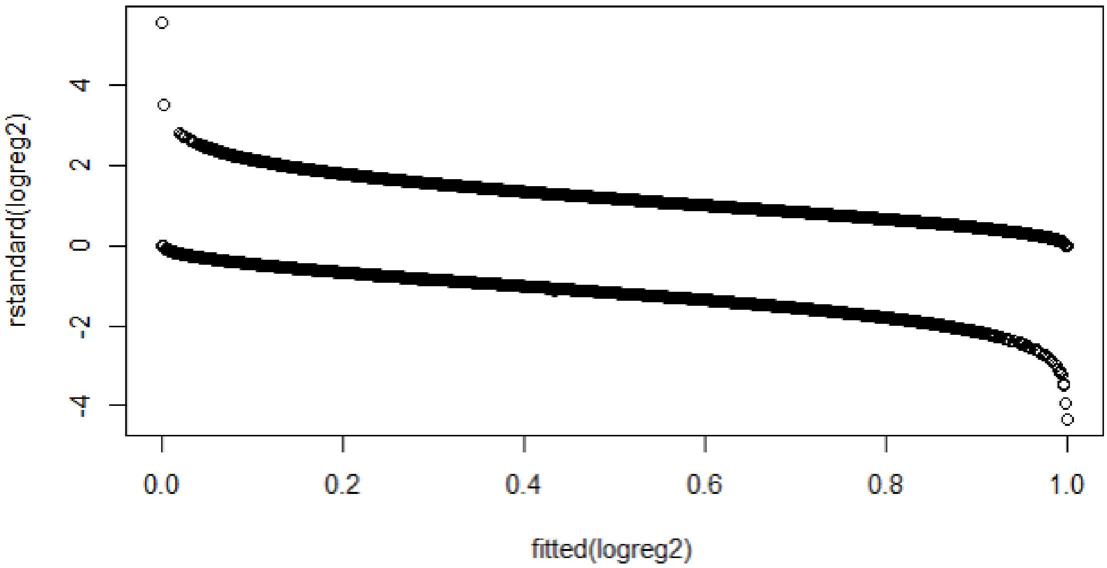
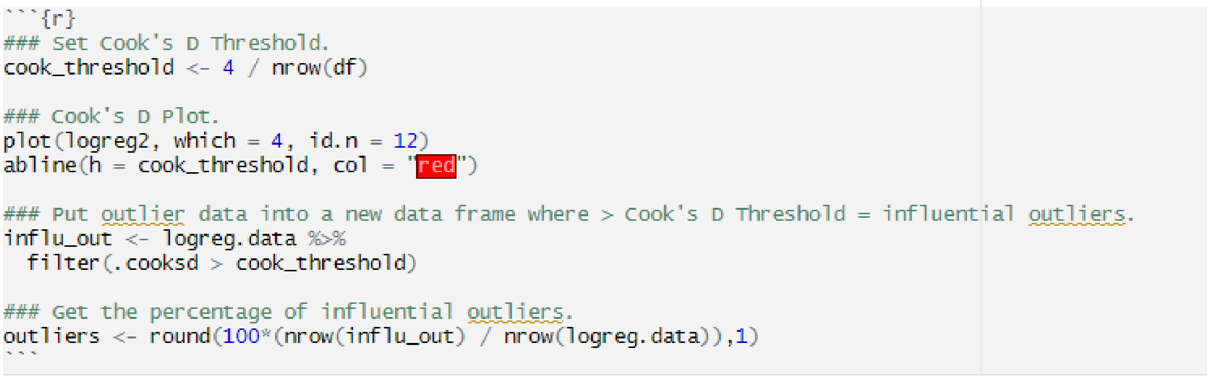


Figure 4.11 Standardized Residual Plot



~~~
```{r}
### Set Cook's D Threshold.
cook_threshold <- 4 / nrow(df)

### Cook's D Plot.
plot(logreg2, which = 4, id.n = 12)
abline(h = cook_threshold, col = "red")

### Put outlier data into a new data frame where > Cook's D Threshold = influential outliers.
influ_out <- logreg.data %>%
  filter(.cooksd > cook_threshold)

### Get the percentage of influential outliers.
outliers <- round(100*(nrow(influ_out) / nrow(logreg.data)),1)
```
~~~

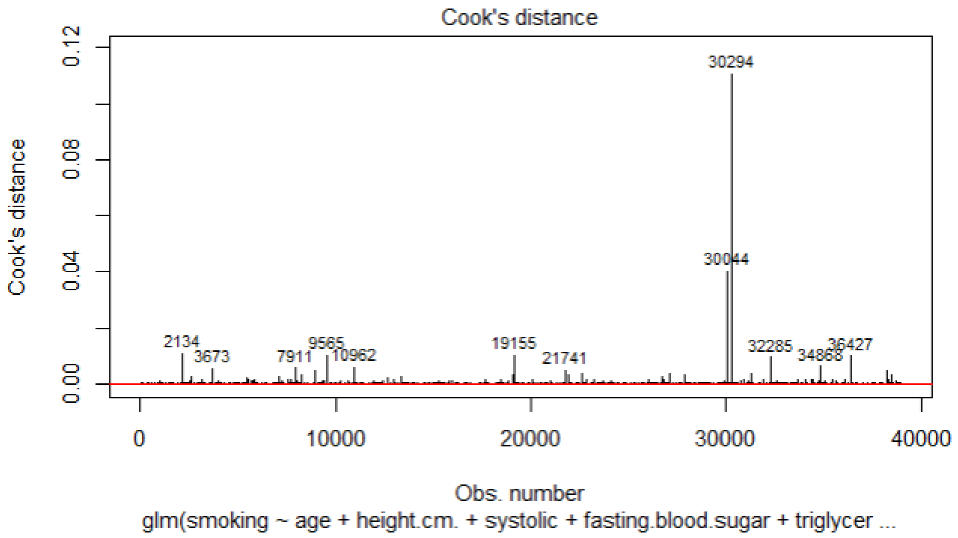
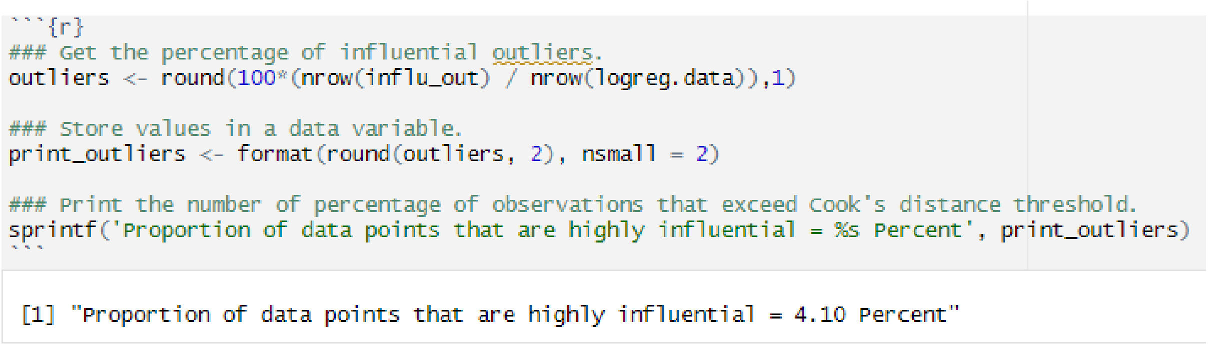


Figure 4.12 Cook Distance Plot



~~~
```{r}
### Get the percentage of influential outliers.
outliers <- round(100*(nrow(influ_out) / nrow(logreg.data)),1)

### Store values in a data variable.
print_outliers <- format(round(outliers, 2), nsmall = 2)

### Print the number of percentage of observations that exceed Cook's distance threshold.
sprintf('Proportion of data points that are highly influential = %s Percent', print_outliers)
```
~~~

```
[1] "Proportion of data points that are highly influential = 4.10 Percent"
```

Figure 4.13 Percentage of Influential Outliers

######################### Results & Interpretation #########################

Standardized Residuals.

Results show that none of the data points of the fitted model consists of any outliers.

Cook's Distance.

In addition, based on the pre-defined threshold (4/N), only 4.1% of the data points are in the outlier zone, which is small as well.

This highlighted section outlines the justification for choosing the final model, which is regarded as the most favourable solution.

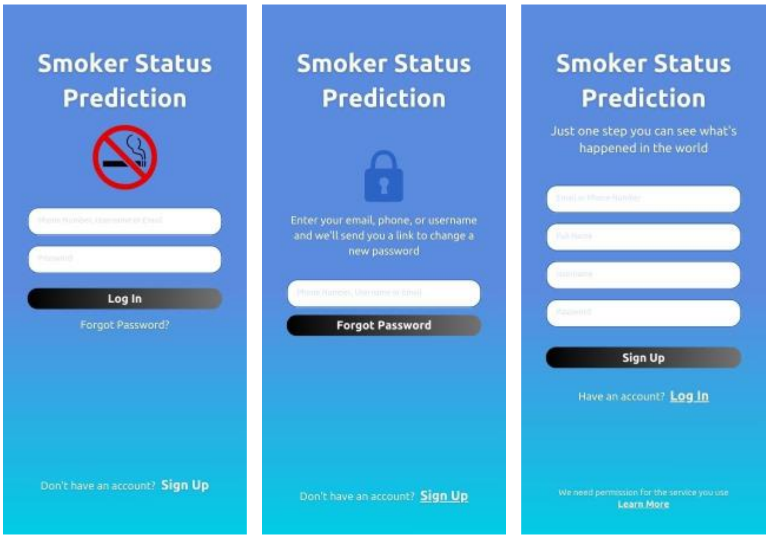
The final logistic regression model was constructed to predict smoking status by considering various health-related variables. Through a stepwise refinement process, insignificant predictors and potential multicollinearity issues were addressed. The resulting model includes significant predictors such as age, height, systolic blood pressure, fasting blood sugar, triglyceride levels, HDL cholesterol, hemoglobin levels, ALT (Alanine Aminotransferase), GTP (Gamma-Glutamyl Transferase), and the presence of dental caries. Notably, serum creatinine levels were found to be statistically insignificant in the full model. Then the reduced model is developed with better fit performance after comparison with full model. Diagnostic tests, including VIF for multicollinearity, standardized residuals and Cook's Distance for influential outliers, and deviance tests, were conducted to ensure the model's robustness. The findings indicate that the model provides a significant fit without strong evidence of multicollinearity or influential outliers. Interpretation of the coefficients suggests how each predictor contributes to the likelihood of smoking based on individual health characteristics, providing a valuable tool for understanding and predicting smoking behavior.

5.0 END PRODUCT

5.1 Introduction

A mobile application for smoker status prediction is a tool that allows people to quickly and efficiently identify whether he or she is a smoker or not. The software would utilize a person's health information such as age, height, weight, blood pressure, hemoglobin, urine protein, and so on to predict a person's smoker status. Furthermore. This tool was developed as a mobile application with a user-friendly interface that allows users to input their preferred data for predictions and review the expected results at the end of pages.

5.2 Smoker Status Prediction Interface



Smoker Status Prediction

Phone Number, Username or Email
Password
Log In
Forgot Password?
Don't have an account? Sign Up

Smoker Status Prediction

Enter your email, phone, or username and we'll send you a link to change a new password

Phone Number, Username or Email
Forgot Password
Don't have an account? Sign Up

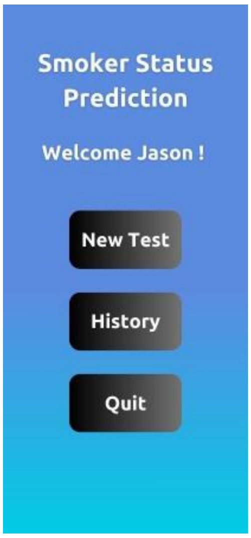
Smoker Status Prediction

Just one step you can see what's happened in the world

We need permission for the service you use

Step 1:

The user can log in by filling up their phone number, username or email and password. If a user forgets their password, they can click the ‘Forget Password?’ button and it will redirect the user to a page where they can reset their password via a link after entering their phone number, username or email. If a user is new to this application and has not yet signed up, they can do so by clicking the ‘Sign Up’ button below and creating a new account.

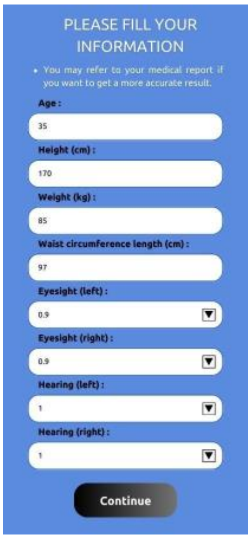


Smoker Status Prediction

Welcome Jason !

Step 2:

Once logged in to an account, it will jump to the next page which is the main menu of the application. It shows that there is the title of the application 'Smoker Status Prediction' on the top and a greeting to the user below the title. After that, three buttons are given to the user to select, the user can choose the 'New Test' button to start to predict the smoker status. If the user has tested the prediction, they can click the 'History' button to review the preview results. Also, the user can choose the 'Quit' button to exit the interface.



PLEASE FILL YOUR INFORMATION

- You may refer to your medical report if you want to get a more accurate result.

Age : 35

Height (cm) : 170

Weight (kg) : 85

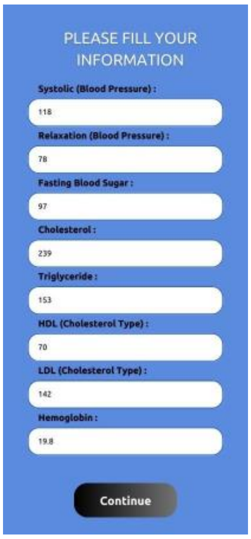
Waist circumference length (cm) : 97

Eyesight (left) : 0.9 ▼

Eyesight (right) : 0.9 ▼

Hearing (left) : 1 ▼

Hearing (right) : 1 ▼



PLEASE FILL YOUR INFORMATION

Systolic (Blood Pressure) : 118

Relaxation (Blood Pressure) : 78

Fasting Blood Sugar : 97

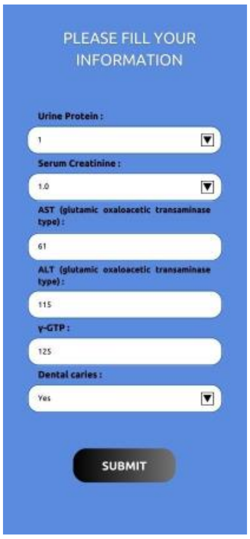
Cholesterol : 239

Triglyceride : 153

HDL (Cholesterol Type) : 70

LDL (Cholesterol Type) : 142

Hemoglobin : 19.8



PLEASE FILL YOUR INFORMATION

Urine Protein : 1 ▼

Serum Creatinine : 1.0 ▼

AST (glutamic oxaloacetic transaminase type) : 61

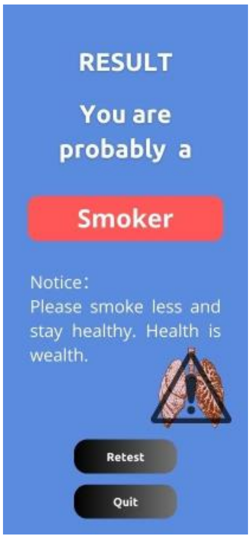
ALT (glutamic oxaloacetic transaminase type) : 115

γ-GTP : 125

Dental caries : Yes ▼

Step 3:

On these pages, the user can fill up their medical information to predict the result. If the user forgets or is not sure about their information, they may refer to their medical check-up report to enter the data in order to get more accurate results at the end. Moreover, the figures show the sample information of a person to test whether he/she is a smoker or non-smoker. After filling up the data, the user can click the 'SUBMIT' button to start the prediction.



RESULT

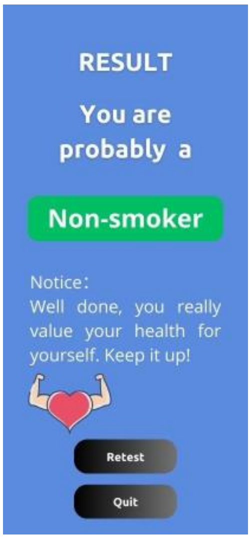
You are probably a

Smoker

Notice:
Please smoke less and stay healthy. Health is wealth.

Step 4 a:

After submitting the data, the result will be shown in the figure above. For the sample data, the result shows the user is probably a smoker based on his / her medical information. In addition, there is also a notice to motivate the user to smoke less and stay healthy. Then, there are two buttons shown below the page which are the 'Retest' and 'Quit' buttons. The user can click the 'Retest' button to re-do the prediction again if they are not satisfied with the result or they can click the 'Quit' button to exit the interface.



RESULT

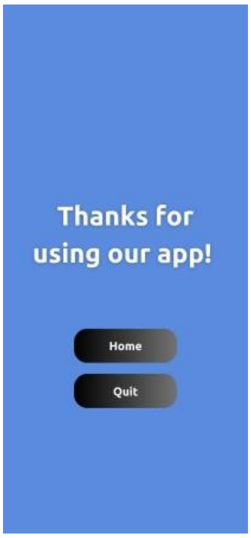
You are probably a

Non-smoker

Notice:
Well done, you really value your health for yourself. Keep it up!

Step 4 b:

This page shows a good result which the user is probably not a smoker based on the information given. Moreover, there is also a notice to the user and two buttons are shown below the page which are the ‘Retest’ and ‘Quit’ buttons. The user can click the ‘Retest’ button to re-do the prediction again or they can click the ‘Quit’ button to exit the interface.



Thanks for using our app!

Step 5:

Lastly, if a user wants to log back into the home page, they can click the ‘Home’ button, which will take them to the login page. Otherwise, they can click the ‘Quit’ button to exit the interface.

6.0 CONCLUSION

In a nutshell, the conclusion of the project on statistical modelling and simulation for predicting smoking status using health-related variables is that a logistic regression model was developed to predict smoking status based on various health characteristics. The model was refined through a stepwise process, addressing predictors and potential multicollinearity issues. The resulting model includes significant predictors such as age, height, blood pressure, cholesterol levels, and other health-related variables. Diagnostic tests were conducted to ensure the model's robustness, indicating a significant fit without strong evidence of multicollinearity or influential outliers. The model provides a valuable tool for understanding and predicting smoking behaviour. Additionally, a mobile application was developed as the end product, providing a user-friendly interface for inputting medical information and predicting smoking status. This project demonstrates the potential of machine learning techniques in constructing health outcome prediction models and offers a practical tool for individuals and healthcare professionals to comprehend and apply smoking cessation strategies.

Moreover, this project is an excellent representation of the critical role that entrepreneurship skills play in a variety of fields: it requires proficiency in statistical analysis, application development, data analysis, predictive modelling, and business planning. This project highlights the entrepreneurial skills needed to interpret data, create user-friendly applications, and innovate in business strategy. It involves everything from identifying predictors for smoking status to developing a mobile app for health prediction and using machine learning techniques for outcome modelling. It emphasises how important it is to use data-driven insights to make strategic decisions and promote industry growth, and how important entrepreneurial savvy is to generating value and developing creative solutions.

7.0 REFERENCE

Centers for Disease Control and Prevention. “CDC - Fact Sheet - Health Effects of Cigarette Smoking - Smoking & Tobacco Use.” *Smoking and Tobacco Use*, U.S. Department of Health & Human Services, 29 Oct. 2021, www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm.

How Does Smoking Affects Red Blood Cell Count? [Fact Checked!]. 8 Sept. 2022, scienceoxygen.com/how-does-smoking-affects-red-blood-cell-count/. Accessed 7 Jan. 2024.

“Tobacco: Health Benefits of Smoking Cessation.” *Www.who.int*, 25 Feb. 2020, www.who.int/news-room/questions-and-answers/item/tobacco-health-benefits-of-smoking-cessation.

“Why Is Hemoglobin Higher in Smokers? – Heimduo.” *Heimduo.org*, heimduo.org/why-is-hemoglobin-higher-in-smokers/. Accessed 7 Jan. 2024.

World Health Organization. “Tobacco.” *World Health Organization*, 31 July 2023, www.who.int/news-room/fact-sheets/detail/tobacco.

SMS Project

Code ▾

Group 3

Hide

```
csv_file_path <- "C:/UMP SEM 5/BSD3443 STATISTICAL MODELLING AND SIMULATION/new title Group P
roject/train_dataset.csv"

df <- read.csv(csv_file_path)

# Display the first few rows of the data frame
head(df)
```

	a.. <int>	height.cm. <int>	weight.kg. <int>	waist.cm. <dbl>	eyesight.left. <dbl>	eyesight.right. <dbl>	hearing.left. <int>
1	35	170	85	97	0.9	0.9	1
2	20	175	110	110	0.7	0.9	1
3	45	155	65	86	0.9	0.9	1
4	45	165	80	94	0.8	0.7	1
5	20	165	60	81	1.5	0.1	1
6	60	160	50	78	1.0	0.9	2

6 rows | 1-9 of 23 columns

Hide

```
#get column names
names(df)
```

```
 [1] "age"                 "height.cm."          "weight.kg."          "waist.cm."
"eyesight.left."      "eyesight.right."
 [7] "hearing.left."       "hearing.right."      "systolic"            "relaxation"
"fasting.blood.sugar" "Cholesterol"
[13] "triglyceride"        "HDL"                 "LDL"                 "hemoglobin"
"Urine.protein"       "serum.creatinine"
[19] "AST"                 "ALT"                 "Gtp"                 "dental.caries"
"smoking"
```

Hide

```
#check if any null value
sum(is.na(df))
```

```
[1] 0
```

Hide

```
#Check if any duplicated data
sum(duplicated(df))
```

```
[1] 5517
```

Hide

```
duplicated_rows <- df[duplicated(df),]
duplicated_rows
```

	a..	height.cm.	weight.kg.	waist.cm.	eyesight.left.	eyesight.right.	hearing.le
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<i
702	20	165	60	78.0	0.9	0.2	
1132	50	175	70	78.5	1.5	1.0	
1144	50	160	65	82.0	1.0	1.0	
1233	45	155	45	68.0	0.2	0.2	
1346	25	165	85	92.3	1.2	1.2	
1463	40	165	65	80.0	1.0	1.2	
1485	45	155	60	75.0	1.0	1.0	
1709	50	170	70	86.0	0.5	0.5	
1750	75	150	50	75.0	0.4	0.7	
1785	40	165	60	75.0	1.2	0.9	

1-10 of 5,517 rows | 1-8 of 23 columns Previous **1** 2 3 4 5 6 ... 100 Next

Hide

```
df<-unique(df)
df
```

	a..	height.cm.	weight.kg.	waist.cm.	eyesight.left.	eyesight.right.	hearing.lef
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<int
1	35	170	85	97.0	0.9	0.9	
2	20	175	110	110.0	0.7	0.9	
3	45	155	65	86.0	0.9	0.9	
4	45	165	80	94.0	0.8	0.7	
5	20	165	60	81.0	1.5	0.1	
6	60	160	50	78.0	1.0	0.9	
7	40	175	90	95.0	0.9	1.0	

	a..	height.cm.	weight.kg.	waist.cm.	eyesight.left.	eyesight.right.	hearing.lef
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<int
8	40	180	75	85.0	1.5	1.5	
9	40	170	60	74.0	1.2	1.5	
10	45	155	55	78.0	0.7	1.0	

Hide

```
#Display the list structure
str(df)
```

```
'data.frame':	33467 obs. of  23 variables:
 $ age                : int  35 20 45 45 20 60 40 40 40 45 ...
 $ height.cm.         : int  170 175 155 165 165 160 175 180 170 155 ...
 $ weight.kg.         : int  85 110 65 80 60 50 90 75 60 55 ...
 $ waist.cm.          : num  97 110 86 94 81 78 95 85 74 78 ...
 $ eyesight.left.     : num  0.9 0.7 0.9 0.8 1.5 1 0.9 1.5 1.2 0.7 ...
 $ eyesight.right.    : num  0.9 0.9 0.9 0.7 0.1 0.9 1 1.5 1.5 1 ...
 $ hearing.left.      : int  1 1 1 1 1 2 1 1 1 1 ...
 $ hearing.right.     : int  1 1 1 1 1 2 1 1 1 1 ...
 $ systolic           : int  118 119 110 158 109 126 130 110 89 114 ...
 $ relaxation         : int  78 79 80 88 64 75 88 60 57 81 ...
 $ fasting.blood.sugar: int  97 88 80 249 100 114 90 100 83 96 ...
 $ Cholesterol        : int  239 211 193 210 179 177 207 170 178 184 ...
 $ triglyceride       : int  153 128 120 366 200 74 331 62 69 177 ...
 $ HDL                : int  70 71 57 46 47 98 39 58 60 41 ...
 $ LDL                : int  142 114 112 91 92 64 102 99 104 107 ...
 $ hemoglobin         : num  19.8 15.9 13.7 16.9 14.9 13.9 16.5 14 12.9 13.1 ...
 $ Urine.protein      : int  1 1 3 1 1 1 1 2 2 1 ...
 $ serum.creatinine   : num  1 1.1 0.6 0.9 1.2 1 1 1.4 0.7 0.6 ...
 $ AST                : int  61 19 1090 32 26 47 19 29 17 22 ...
 $ ALT                : int  115 25 1400 36 28 23 22 20 17 15 ...
 $ Gtp                : int  125 30 276 36 15 70 19 32 14 56 ...
 $ dental.caries      : int  1 1 0 0 0 0 0 1 0 0 ...
 $ smoking            : int  1 0 0 0 0 1 0 1 0 0 ...
```

Hide

```
# get number of rows and columns of data frame
dim(df)
```

```
[1] 33467    23
```

Hide

```
# check null values in dataset
colSums(is.na(df))
```

```
                age          height.cm.          weight.kg.           waist.cm.      eyesigh
t.left.     eyesight.right.
                  0                   0                   0                   0
0                   0
      hearing.left.      hearing.right.            systolic          relaxation fasting.bloo
d.sugar         Cholesterol
                  0                   0                   0                   0
0                   0
       triglyceride                 HDL                 LDL          hemoglobin       Urine.p
rotein    serum.creatinine
                  0                   0                   0                   0
0                   0
                AST                 ALT                 Gtp       dental.caries             s
moking
                  0                   0                   0                   0
0
```

Hide

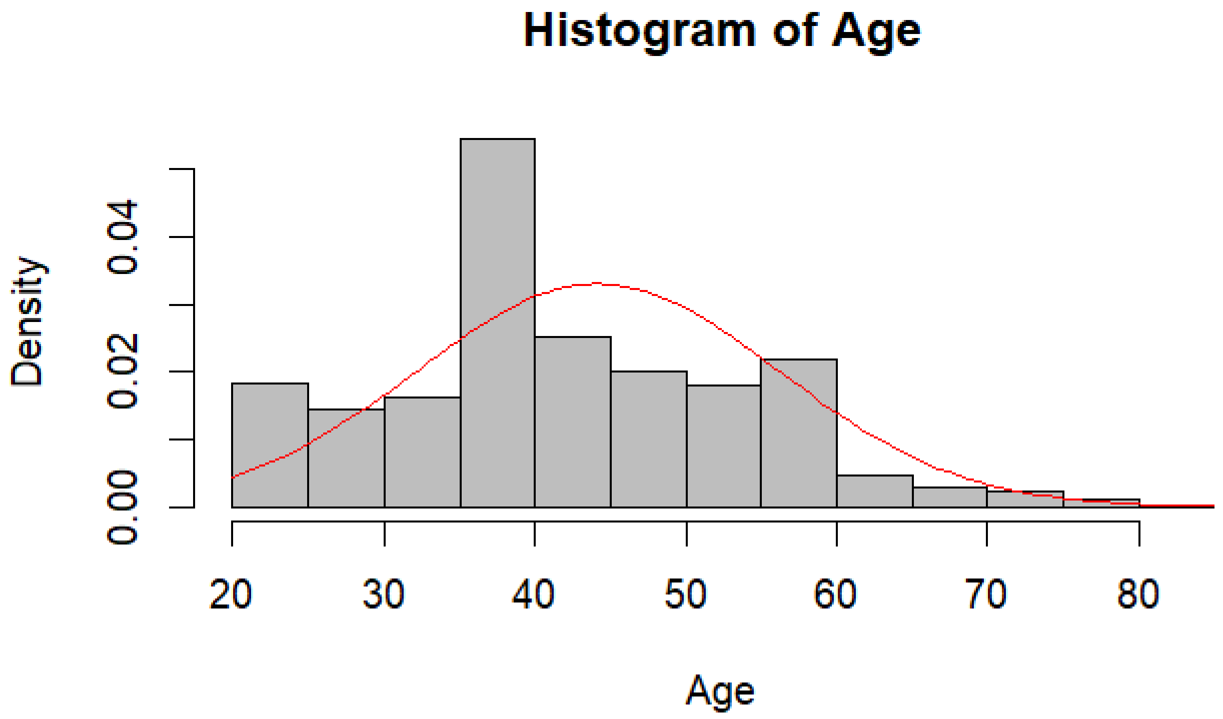
```
summary(df)
```

```
      age          height.cm.      weight.kg.       waist.cm.      eyesight.left.  eyesight.r
ight. hearing.left.   hearing.right.
 Min.   :20.00   Min.   :130.0   Min.   : 30.00   Min.   : 51.00   Min.   :0.100   Min.   :0.
10    Min.   :1.000   Min.   :1.000
 1st Qu.:40.00   1st Qu.:160.0   1st Qu.: 55.00   1st Qu.: 76.00   1st Qu.:0.800   1st Qu.:0.
80    1st Qu.:1.000   1st Qu.:1.000
 Median :40.00   Median :165.0   Median : 65.00   Median : 82.00   Median :1.000   Median :1.
00    Median :1.000   Median :1.000
 Mean   :44.15   Mean   :164.7   Mean   : 65.93   Mean   : 82.08   Mean   :1.014   Mean   :1.
01    Mean   :1.025   Mean   :1.026
 3rd Qu.:55.00   3rd Qu.:170.0   3rd Qu.: 75.00   3rd Qu.: 88.00   3rd Qu.:1.200   3rd Qu.:1.
20    3rd Qu.:1.000   3rd Qu.:1.000
 Max.   :85.00   Max.   :190.0   Max.   :135.00   Max.   :129.00   Max.   :9.900   Max.   :9.
90    Max.   :2.000   Max.   :2.000
    systolic       relaxation     fasting.blood.sugar  Cholesterol   triglyceride        HDL
LDL           hemoglobin
 Min.   : 71.0   Min.   : 40.00   Min.   : 46.00      Min.   : 55   Min.   :  8.0   Min.   :
4.00   Min.   :   1.0   Min.   : 4.90
 1st Qu.:112.0   1st Qu.: 70.00   1st Qu.: 89.00      1st Qu.:172   1st Qu.: 75.0   1st Qu.:
47.00   1st Qu.:  92.0   1st Qu.:13.60
 Median :120.0   Median : 76.00   Median : 96.00      Median :195   Median :108.0   Median :
55.00   Median : 113.0   Median :14.80
 Mean   :121.5   Mean   : 76.02   Mean   : 99.26      Mean   :197   Mean   :126.8   Mean   :
57.26   Mean   : 115.2   Mean   :14.62
 3rd Qu.:130.0   3rd Qu.: 82.00   3rd Qu.:104.00      3rd Qu.:220   3rd Qu.:160.0   3rd Qu.:
66.00   3rd Qu.: 136.0   3rd Qu.:15.70
 Max.   :233.0   Max.   :146.00   Max.   :423.00      Max.   :445   Max.   :999.0   Max.   :3
59.00   Max.   :1860.0   Max.   :21.10
 Urine.protein   serum.creatinine       AST              ALT               Gtp         denta
l.caries       smoking
 Min.   :1.000   Min.   : 0.1000   Min.   :   6.0   Min.   :   1.00   Min.   :  2.00   Min.   
:0.0000   Min.   :0.0000
 1st Qu.:1.000   1st Qu.: 0.8000   1st Qu.:  19.0   1st Qu.:  15.00   1st Qu.: 17.00   1st Q
u.:0.0000   1st Qu.:0.0000
 Median :1.000   Median : 0.9000   Median :  23.0   Median :  21.00   Median : 26.00   Median
:0.0000   Median :0.0000
 Mean   :1.087   Mean   : 0.8865   Mean   :  26.2   Mean   :  27.14   Mean   : 39.95   Mean   
:0.2147   Mean   :0.3663
 3rd Qu.:1.000   3rd Qu.: 1.0000   3rd Qu.:  29.0   3rd Qu.:  31.00   3rd Qu.: 44.00   3rd Q
u.:0.0000   3rd Qu.:1.0000
 Max.   :6.000   Max.   :11.6000   Max.   :1090.0   Max.   :2914.00   Max.   :999.00   Max.   
:1.0000   Max.   :1.0000
```

#Exploratory Data Analysis (EDA)

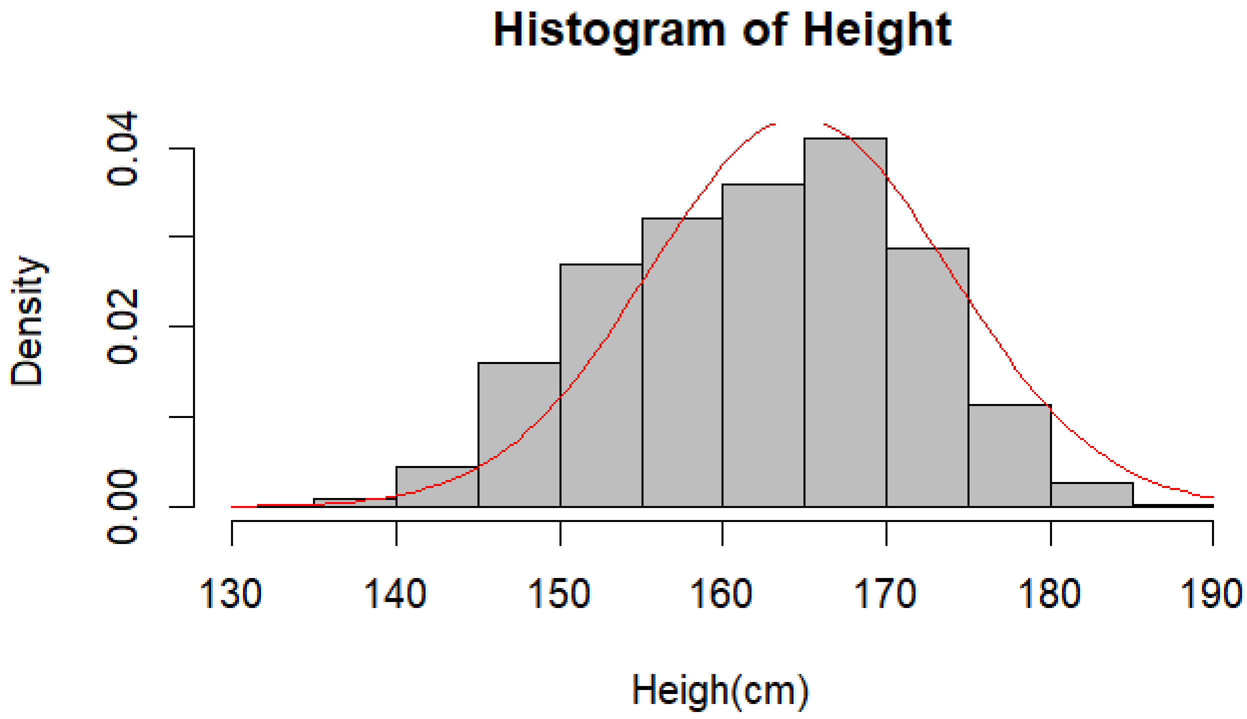
Hide

```
hist(df$age, freq=FALSE, col="gray", xlab="Age", main = "Histogram of Age")
curve(dnorm(x, mean=mean(df$age), sd=sd(df$age)), add=TRUE, col="red") #line
```



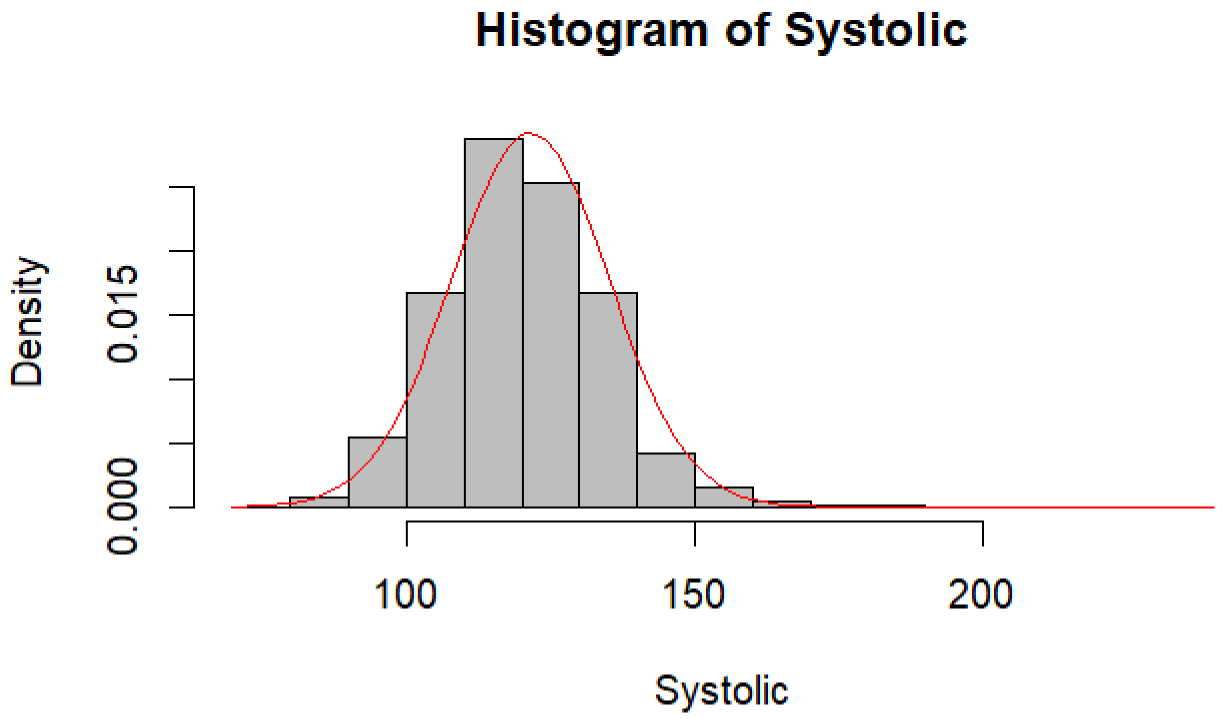
Hide

```
hist(df$height.cm., freq=FALSE, col="gray", xlab="Heigh(cm)", main = "Histogram of Height")
curve(dnorm(x, mean=mean(df$height.cm.), sd=sd(df$height.cm.)), add=TRUE, col="red") #line
```



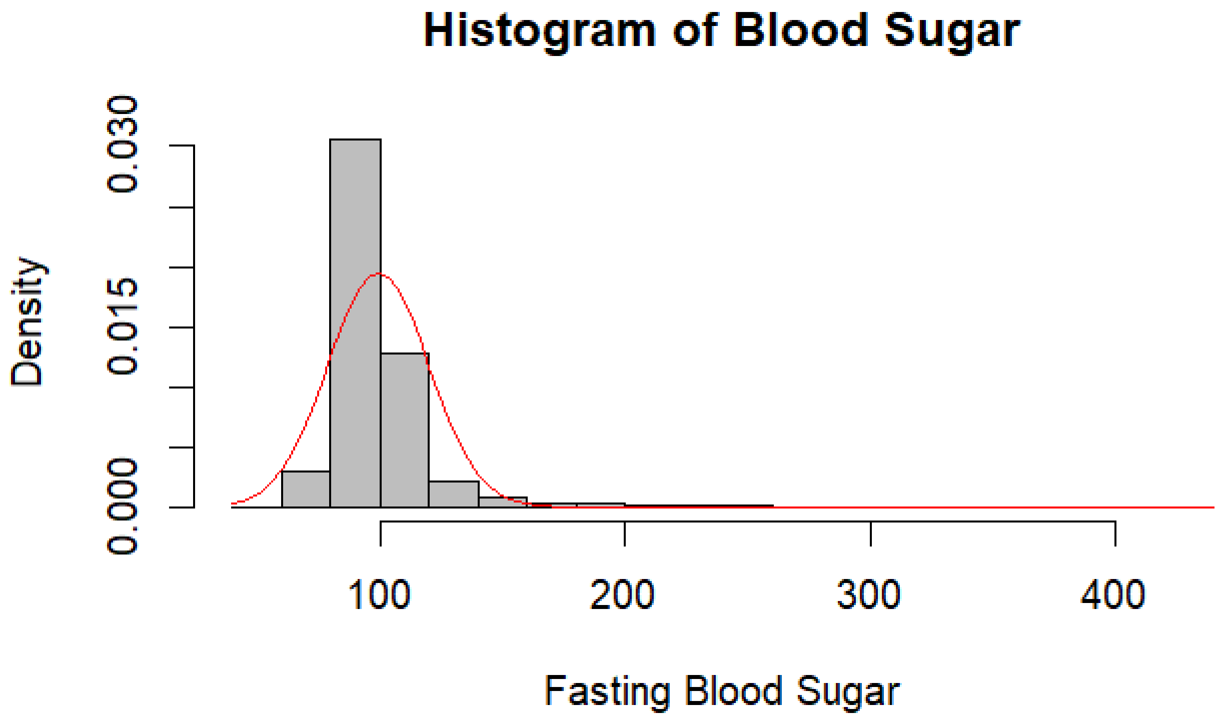
Hide

```
hist(df$systolic, freq=FALSE, col="gray", xlab="Systolic", main = "Histogram of Systolic")
curve(dnorm(x, mean=mean(df$systolic), sd=sd(df$systolic)), add=TRUE, col="red") #line
```



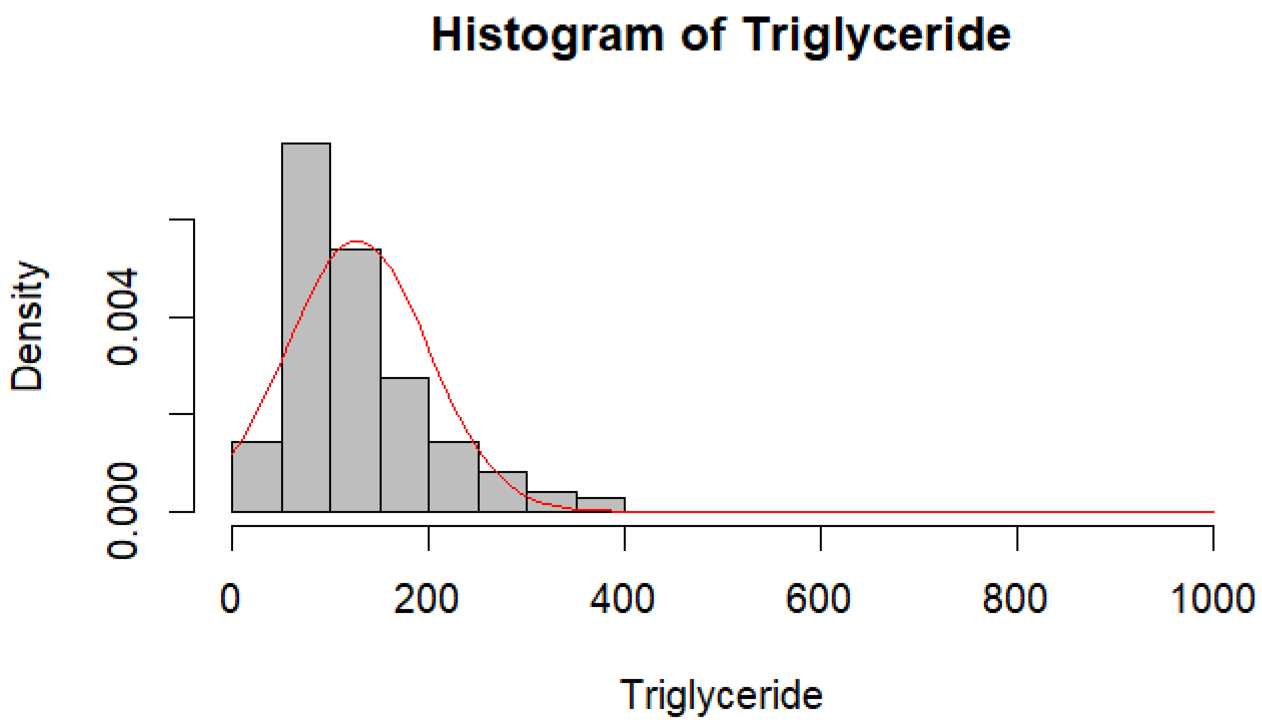
Hide

```
hist(df$fasting.blood.sugar, freq=FALSE, col="gray", xlab="Fasting Blood Sugar", main = "Histogram of Blood Sugar")
curve(dnorm(x, mean=mean(df$fasting.blood.sugar), sd=sd(df$fasting.blood.sugar)), add=TRUE, col="red") #line
```



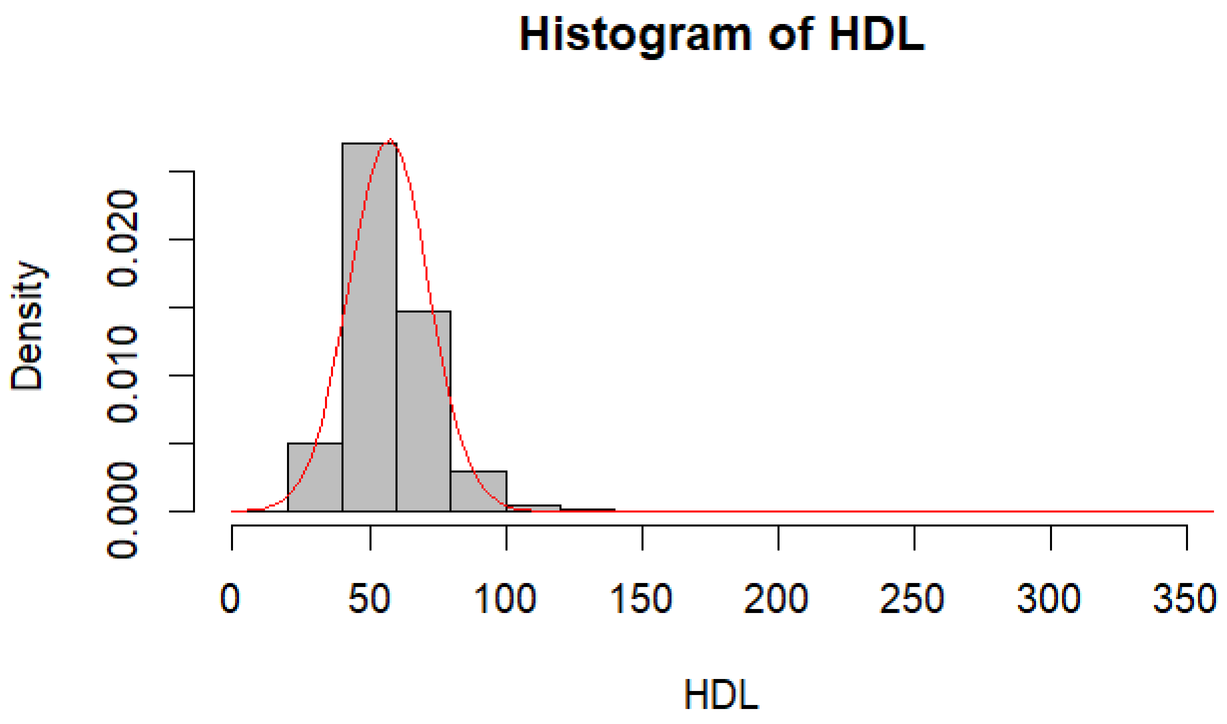
Hide

```
hist(df$triglyceride, freq=FALSE, col="gray", xlab="Triglyceride", main = "Histogram of Triglyceride")
curve(dnorm(x, mean=mean(df$triglyceride), sd=sd(df$triglyceride)), add=TRUE, col="red") #line
```



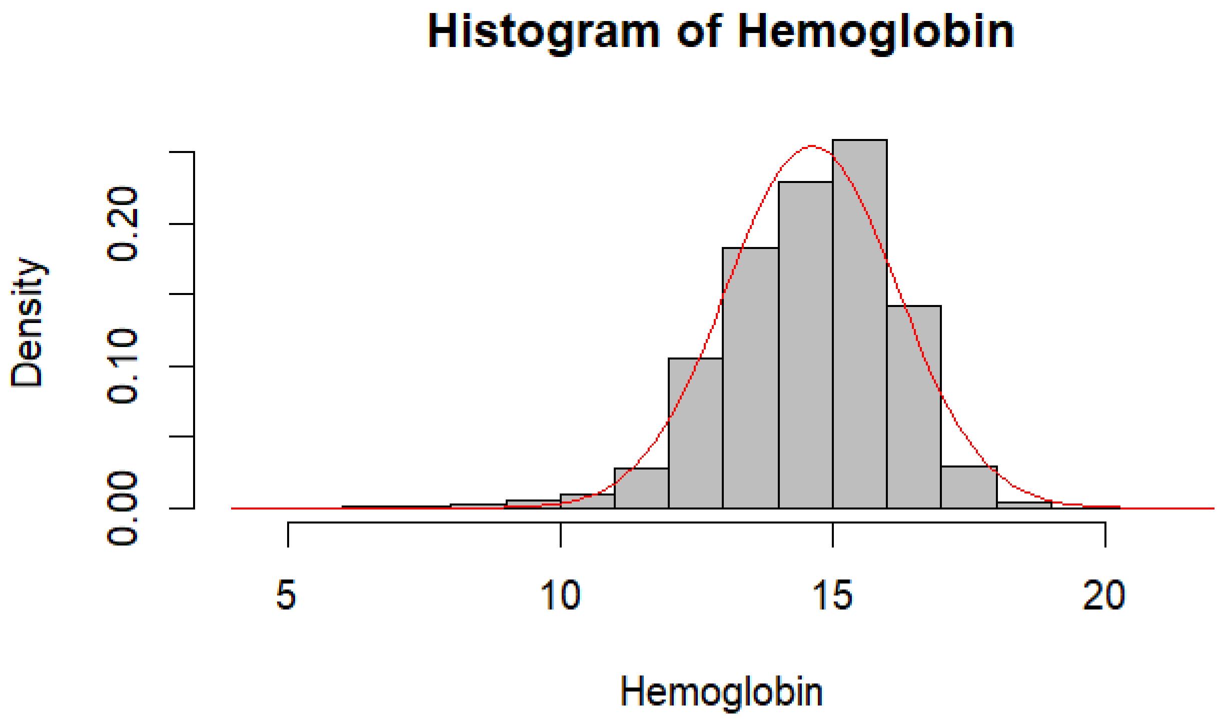
Hide

```
hist(df$HDL, freq=FALSE, col="gray", xlab="HDL", main = "Histogram of HDL")
curve(dnorm(x, mean=mean(df$HDL), sd=sd(df$HDL)), add=TRUE, col="red") #line
```



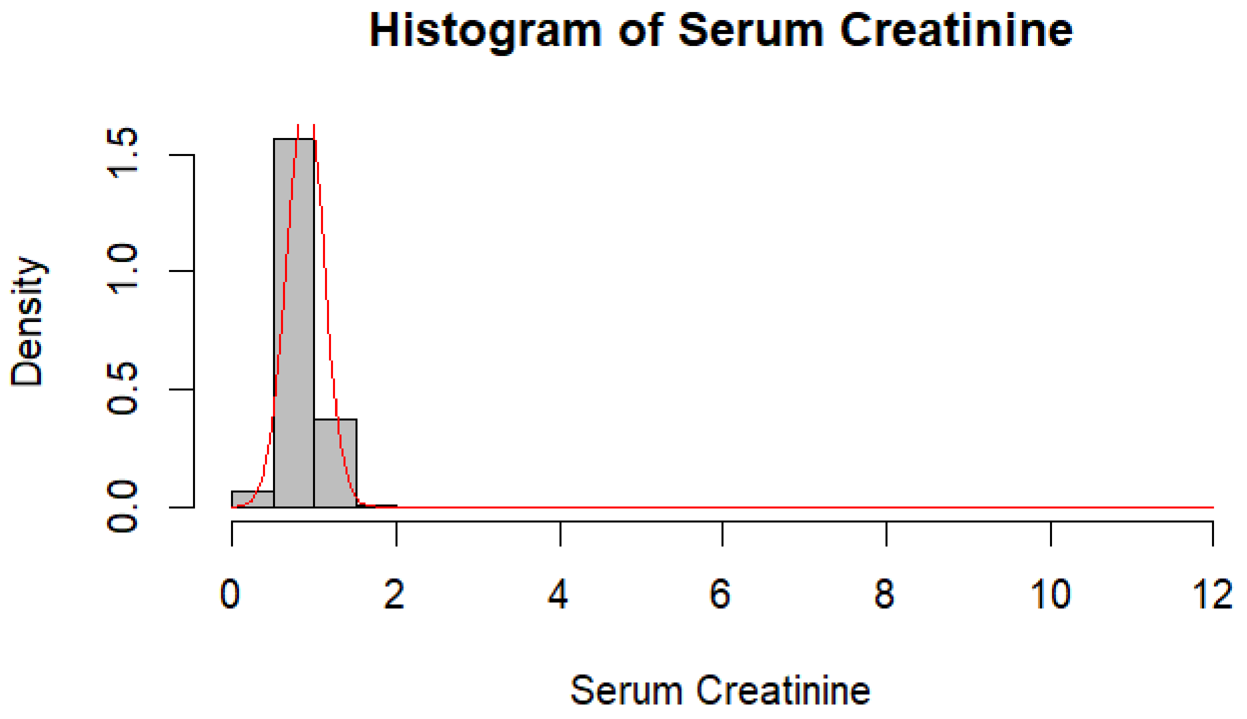
Hide

```
hist(df$hemoglobin, freq=FALSE, col="gray", xlab="Hemoglobin", main = "Histogram of Hemoglobin")
curve(dnorm(x, mean=mean(df$hemoglobin), sd=sd(df$hemoglobin)), add=TRUE, col="red") #line
```



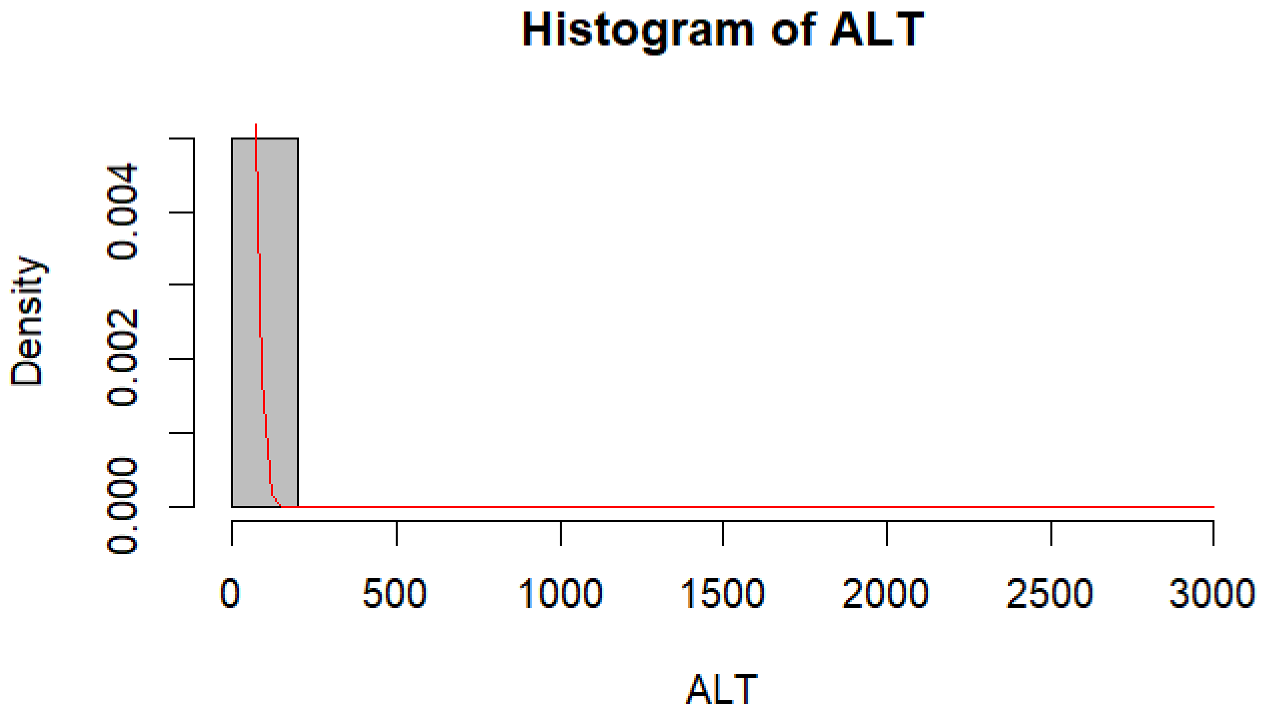
Hide

```
hist(df$serum.creatinine, freq=FALSE, col="gray", xlab="Serum Creatinine", main = "Histogram of Serum Creatinine")
curve(dnorm(x, mean=mean(df$serum.creatinine), sd=sd(df$serum.creatinine)), add=TRUE, col="red") #line
```



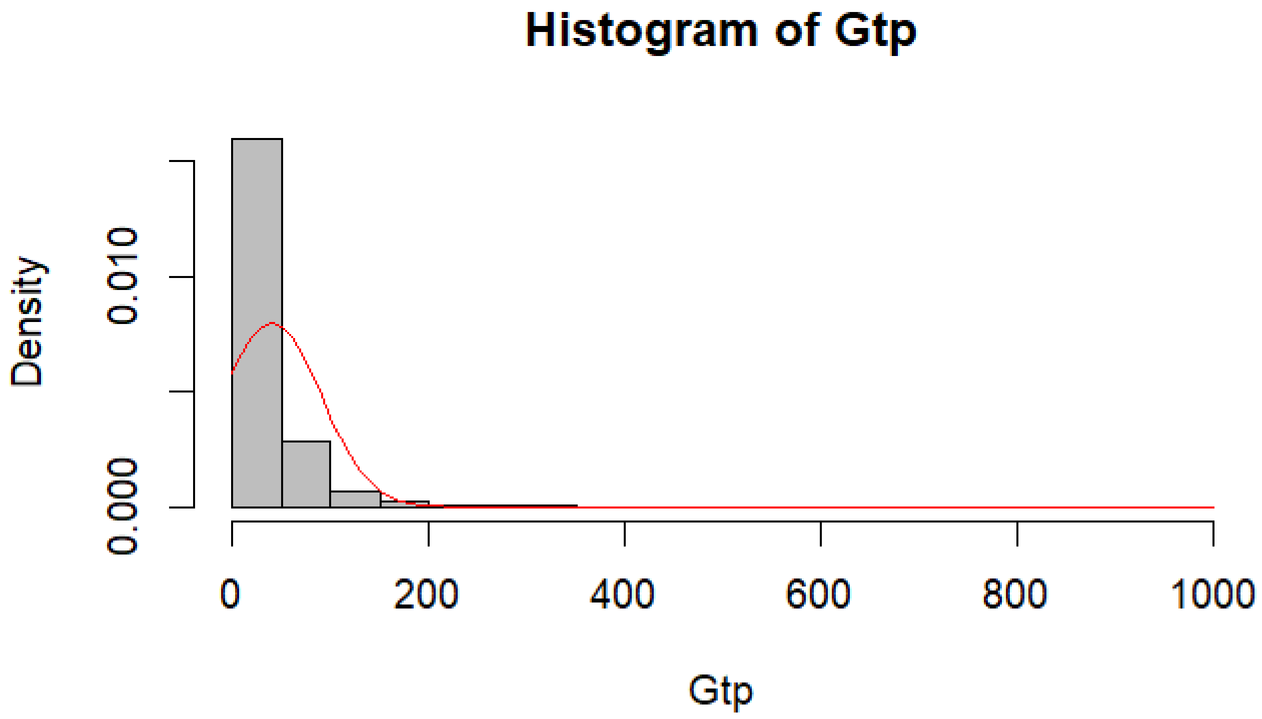
Hide

```
hist(df$ALT, freq=FALSE, col="gray", xlab="ALT", main = "Histogram of ALT")
curve(dnorm(x, mean=mean(df$ALT), sd=sd(df$ALT)), add=TRUE, col="red") #line
```



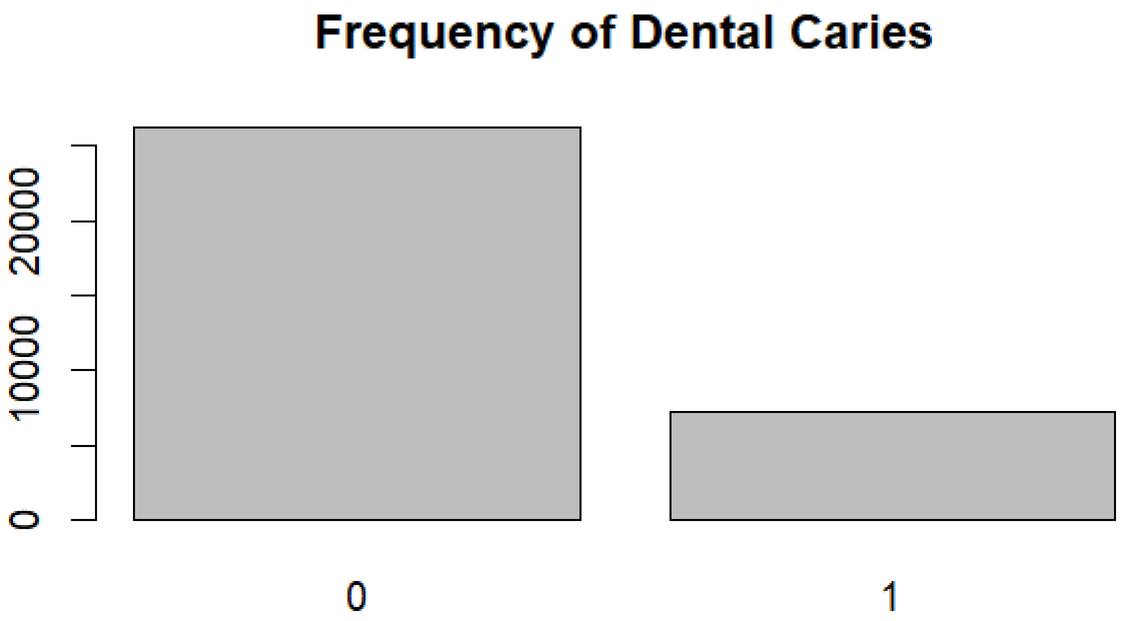
Hide

```
hist(df$Gtp, freq=FALSE, col="gray", xlab="Gtp", main = "Histogram of Gtp")
curve(dnorm(x, mean=mean(df$Gtp), sd=sd(df$Gtp)), add=TRUE, col="red") #line
```



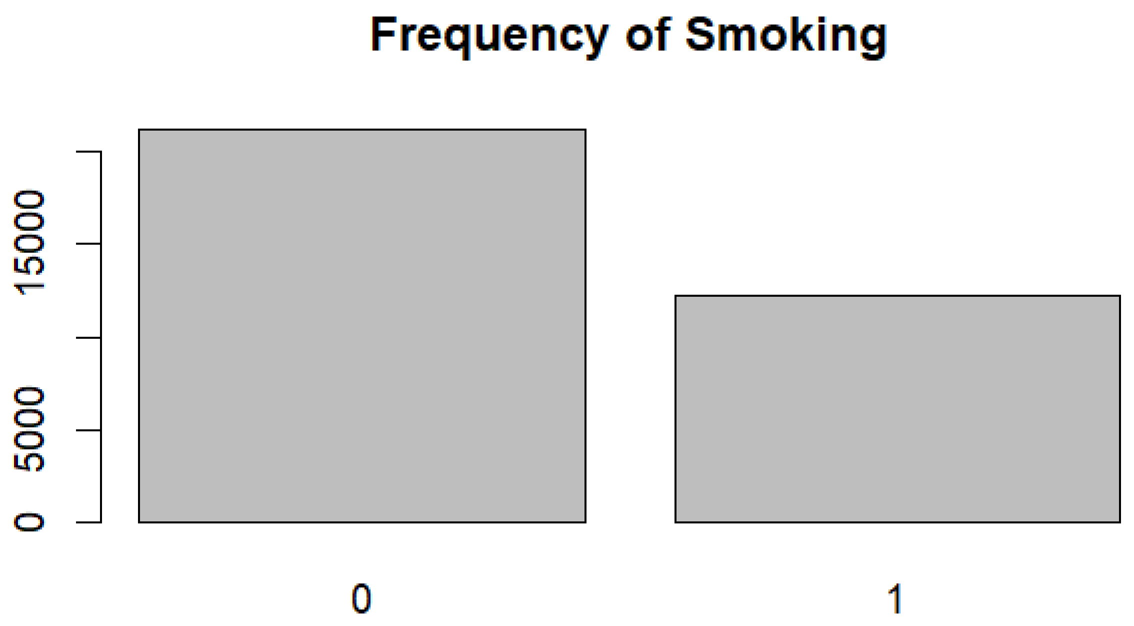
Hide

```
barplot(table(df$dental.caries), main = "Frequency of Dental Caries")
```



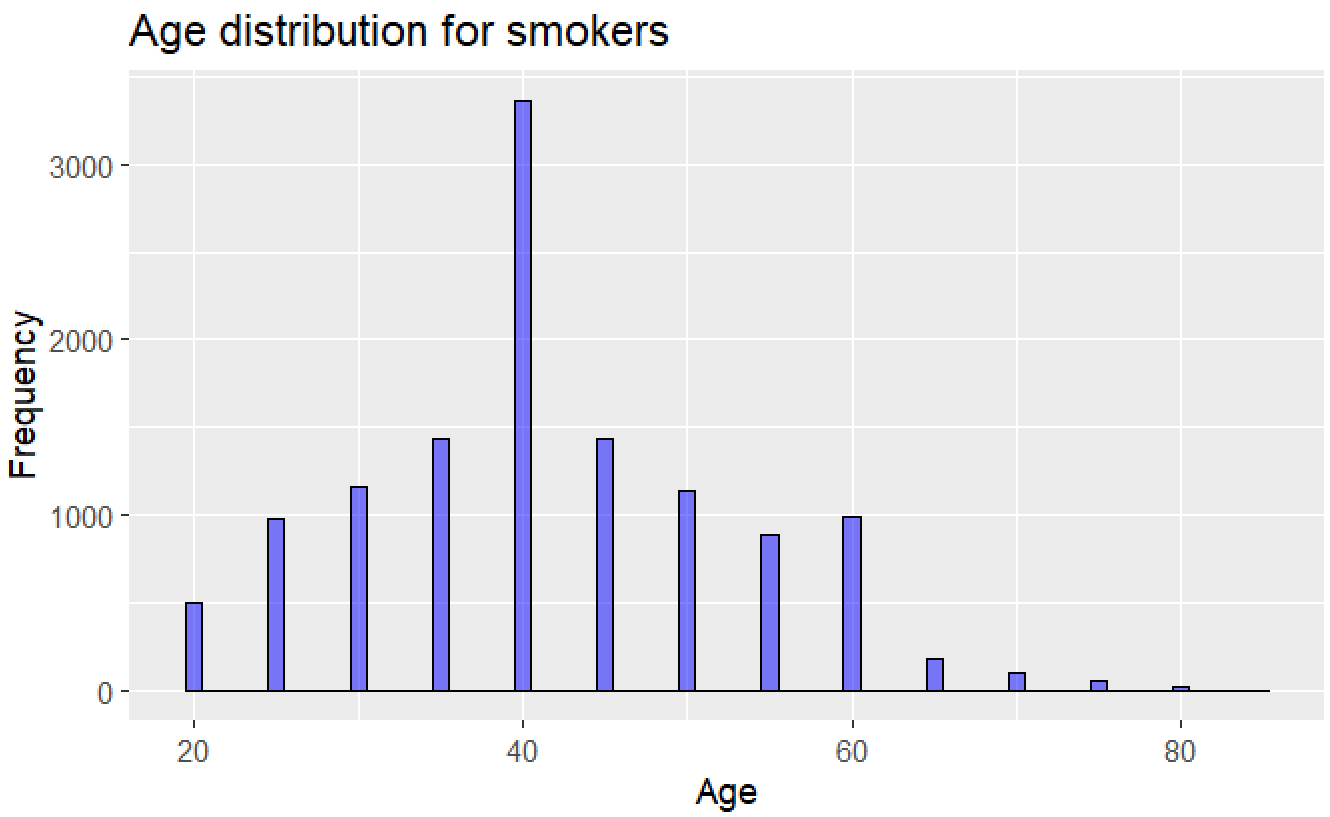
Hide

```
barplot(table(df$smoking), main = "Frequency of Smoking")
```



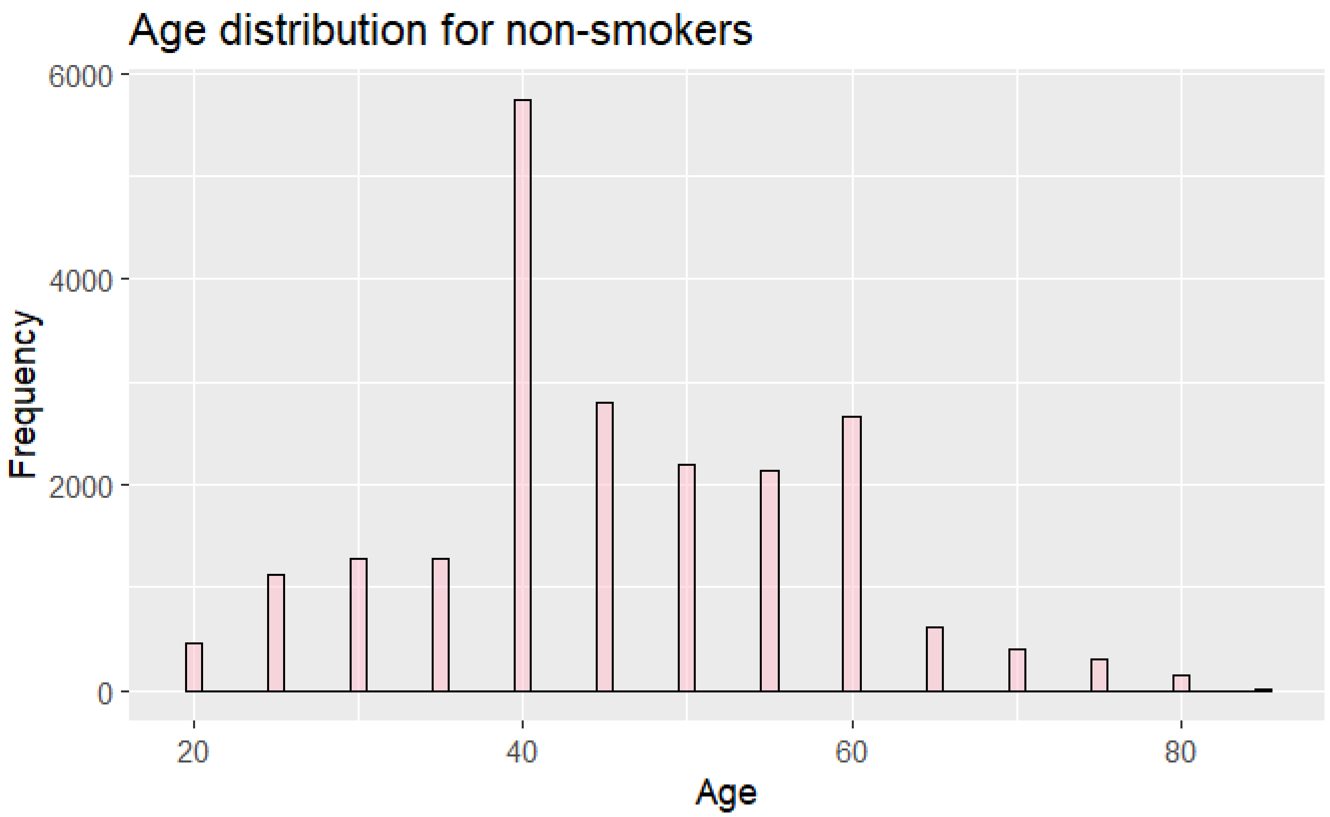
Hide

```
# Assuming you have the 'ggplot2' library installed. If not, install it using install.packages('ggplot2')
library(ggplot2)

# Histogram for smokers
ggplot(subset(df, smoking == 1), aes(x = age)) +
  geom_histogram(binwidth = 1, color = "black", fill = "blue", alpha = 0.5) +
  labs(title = "Age distribution for smokers") +
  xlab("Age") +
  ylab("Frequency")
```


Age distribution for smokers

Hide

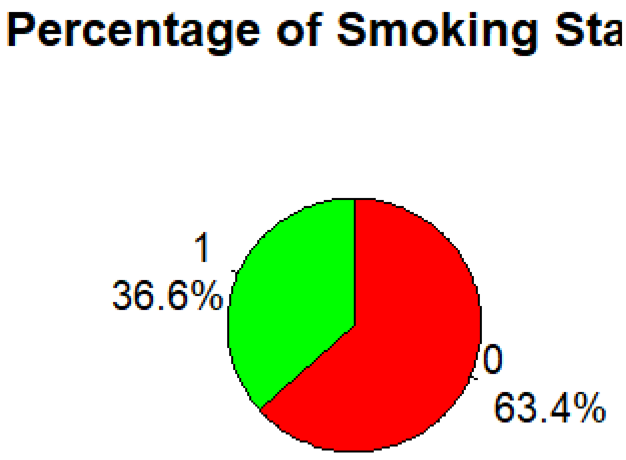
```
# Histogram for non-smokers
ggplot(subset(df, smoking == 0), aes(x = age)) +
  geom_histogram(binwidth = 1, color = "black", fill = "pink", alpha = 0.5) +
  labs(title = "Age distribution for non-smokers") +
  xlab("Age") +
  ylab("Frequency")
```


Age distribution for non-smokers

Based on both the histograms, the frequency of smokers and non-smokers is the highest at age 40 while the lowest is at age 80 respectively.

Hide

```
library(ggplot2)
library(plotrix)
library(scales)
# Pie chart for 'smoking'
smoking_counts <- table(df$smoking)

# Calculate percentages and format labels
percentages <- smoking_counts / sum(smoking_counts) * 100
labels_with_percentages <- paste(names(percentages), "\n", sprintf("%.1f%%", percentages))

# Plotting side by side with percentages
par(mfrow = c(1, 2), mar = c(5, 5, 2, 2))

pie(smoking_counts, labels = labels_with_percentages, col = c("red", "green"),
    main = "Percentage of Smoking Status", clockwise = TRUE,
    shadow = TRUE, explode = c(0, 0.1), percent = TRUE)
```



Hide

```
NA
NA
```

Based on the pie chart, non-smokers have a higher percentage of smoking status of 63.4% while another 36.6% are smokers.

Hide

```
round(cor(df),4)
```

```
                        age height.cm. weight.kg. waist.cm. eyesight.left. eyesight.right. he
aring.left. hearing.right. systolic relaxation
age                  1.0000    -0.4809    -0.3255   -0.0271        -0.1882         -0.1918
0.2018          0.2066   0.1327     0.0484
height.cm.          -0.4809     1.0000     0.6744    0.3781         0.1489          0.1556
-0.0808         -0.0777   0.0820     0.1149
weight.kg.          -0.3255     0.6744     1.0000    0.8247         0.1068          0.1136
-0.0519         -0.0526   0.2683     0.2736
waist.cm.           -0.0271     0.3781     0.8247    1.0000         0.0307          0.0396
0.0215          0.0197   0.3193     0.2930
eyesight.left.      -0.1882     0.1489     0.1068    0.0307         1.0000          0.3444
-0.0465         -0.0484  -0.0139     0.0072
eyesight.right.     -0.1918     0.1556     0.1136    0.0396         0.3444          1.0000
-0.0374         -0.0352  -0.0083     0.0146
hearing.left.        0.2018    -0.0808    -0.0519    0.0215        -0.0465         -0.0374
1.0000          0.5152   0.0570     0.0093
hearing.right.       0.2066    -0.0777    -0.0526    0.0197        -0.0484         -0.0352
0.5152          1.0000   0.0482    -0.0012
systolic             0.1327     0.0820     0.2683    0.3193        -0.0139         -0.0083
0.0570          0.0482   1.0000     0.7606
relaxation           0.0484     0.1149     0.2736    0.2930         0.0072          0.0146
0.0093         -0.0012   0.7606     1.0000
fasting.blood.sugar  0.1833     0.0155     0.1362    0.2138        -0.0417         -0.0424
0.0421          0.0450   0.1724     0.1496
Cholesterol          0.0580    -0.0809     0.0254    0.0635        -0.0077         -0.0089
-0.0270         -0.0242   0.0596     0.0947
triglyceride         0.0147     0.1576     0.3228    0.3595         0.0225          0.0257
0.0047          0.0015   0.1971     0.2153
HDL                  0.0085    -0.2143    -0.3607   -0.3762        -0.0209         -0.0239
-0.0196         -0.0144  -0.0842    -0.0884
LDL                  0.0422    -0.0465     0.0368    0.0672        -0.0093         -0.0091
-0.0193         -0.0179   0.0136     0.0358
hemoglobin          -0.2652     0.5367     0.4888    0.3804         0.0956          0.0963
-0.0319         -0.0318   0.1851     0.2347
Urine.protein        0.0273     0.0060     0.0284    0.0423        -0.0088         -0.0118
0.0168          0.0179   0.0480     0.0521
serum.creatinine    -0.1035     0.3801     0.3175    0.2272         0.0578          0.0464
0.0022          0.0105   0.0732     0.0879
AST                  0.0248     0.0498     0.1289    0.1463        -0.0057         -0.0021
0.0079          0.0100   0.0863     0.0865
ALT                 -0.0681     0.1315     0.2513    0.2504         0.0216          0.0276
0.0004          0.0014   0.0935     0.1039
Gtp                  0.0113     0.1383     0.2087    0.2403         0.0029          0.0171
0.0084          0.0050   0.1714     0.1786
dental.caries       -0.1145     0.0829     0.0767    0.0476         0.0077          0.0169
-0.0186         -0.0165   0.0344     0.0363
smoking             -0.1671     0.3934     0.2981    0.2228         0.0601          0.0620
-0.0233         -0.0181   0.0704     0.1045
                    fasting.blood.sugar Cholesterol triglyceride     HDL     LDL hemoglobin U
rine.protein serum.creatinine     AST     ALT
age                              0.1833      0.0580       0.0147  0.0085  0.0422    -0.2652
0.0273           -0.1035  0.0248 -0.0681
height.cm.                       0.0155     -0.0809       0.1576 -0.2143 -0.0465     0.5367
0.0060            0.3801  0.0498  0.1315
weight.kg.                       0.1362      0.0254       0.3228 -0.3607  0.0368     0.4888
```

```
0.0284           0.3175  0.1289  0.2513
waist.cm.                        0.2138      0.0635       0.3595 -0.3762  0.0672     0.3804
0.0423           0.2272  0.1463  0.2504
eyesight.left.                  -0.0417     -0.0077       0.0225 -0.0209 -0.0093     0.0956
-0.0088           0.0578 -0.0057  0.0216
eyesight.right.                 -0.0424     -0.0089       0.0257 -0.0239 -0.0091     0.0963
-0.0118           0.0464 -0.0021  0.0276
hearing.left.                    0.0421     -0.0270       0.0047 -0.0196 -0.0193    -0.0319
0.0168           0.0022  0.0079  0.0004
hearing.right.                   0.0450     -0.0242       0.0015 -0.0144 -0.0179    -0.0318
0.0179           0.0105  0.0100  0.0014
systolic                         0.1724      0.0596       0.1971 -0.0842  0.0136     0.1851
0.0480           0.0732  0.0863  0.0935
relaxation                       0.1496      0.0947       0.2153 -0.0884  0.0358     0.2347
0.0521           0.0879  0.0865  0.1039
fasting.blood.sugar              1.0000      0.0091       0.2264 -0.1222 -0.0185     0.0985
0.1055           0.0433  0.0693  0.0888
Cholesterol                      0.0091      1.0000       0.2436  0.1684  0.7066     0.0670
-0.0012           0.0063  0.0254  0.0473
triglyceride                     0.2264      0.2436       1.0000 -0.4153  0.0221     0.2724
0.0330           0.1234  0.1037  0.1744
HDL                             -0.1222      0.1684      -0.4153  1.0000 -0.0560    -0.2409
-0.0085          -0.1719 -0.0362 -0.1298
LDL                             -0.0185      0.7066       0.0221 -0.0560  1.0000     0.0548
-0.0086           0.0319 -0.0017  0.0277
hemoglobin                       0.0985      0.0670       0.2724 -0.2409  0.0548     1.0000
0.0238           0.3661  0.1239  0.1992
Urine.protein                    0.1055     -0.0012       0.0330 -0.0085 -0.0086     0.0238
1.0000           0.0862  0.0539  0.0387
serum.creatinine                 0.0433      0.0063       0.1234 -0.1719  0.0319     0.3661
0.0862           1.0000  0.0516  0.0861
AST                              0.0693      0.0254       0.1037 -0.0362 -0.0017     0.1239
0.0539           0.0516  1.0000  0.6671
ALT                              0.0888      0.0473       0.1744 -0.1298  0.0277     0.1992
0.0387           0.0861  0.6671  1.0000
Gtp                              0.1801      0.0835       0.2988 -0.0480 -0.0088     0.2228
0.0706           0.1112  0.3790  0.3307
dental.caries                   -0.0022     -0.0004       0.0296 -0.0273 -0.0002     0.0697
-0.0010           0.0316  0.0131  0.0259
smoking                          0.0978     -0.0296       0.2495 -0.1774 -0.0420     0.3983
0.0134           0.2090  0.0635  0.0973
                        Gtp dental.caries smoking
age                  0.0113       -0.1145 -0.1671
height.cm.           0.1383        0.0829  0.3934
weight.kg.           0.2087        0.0767  0.2981
waist.cm.            0.2403        0.0476  0.2228
eyesight.left.       0.0029        0.0077  0.0601
eyesight.right.      0.0171        0.0169  0.0620
hearing.left.        0.0084       -0.0186 -0.0233
hearing.right.       0.0050       -0.0165 -0.0181
systolic             0.1714        0.0344  0.0704
relaxation           0.1786        0.0363  0.1045
fasting.blood.sugar  0.1801       -0.0022  0.0978
Cholesterol          0.0835       -0.0004 -0.0296
triglyceride         0.2988        0.0296  0.2495
HDL                 -0.0480       -0.0273 -0.1774
```

```
LDL                  -0.0088       -0.0002 -0.0420
hemoglobin            0.2228        0.0697  0.3983
Urine.protein         0.0706       -0.0010  0.0134
serum.creatinine      0.1112        0.0316  0.2090
AST                   0.3790        0.0131  0.0635
ALT                   0.3307        0.0259  0.0973
Gtp                   1.0000        0.0454  0.2380
dental.caries         0.0454        1.0000  0.1064
smoking               0.2380        0.1064  1.0000
```

Based on the correlation shown above, the variables that have positive correlation with response variables smoking are height.cm., weight.kg., waist.cm., triglyceride, hemoglobin, serum.creatinine and Gtp.

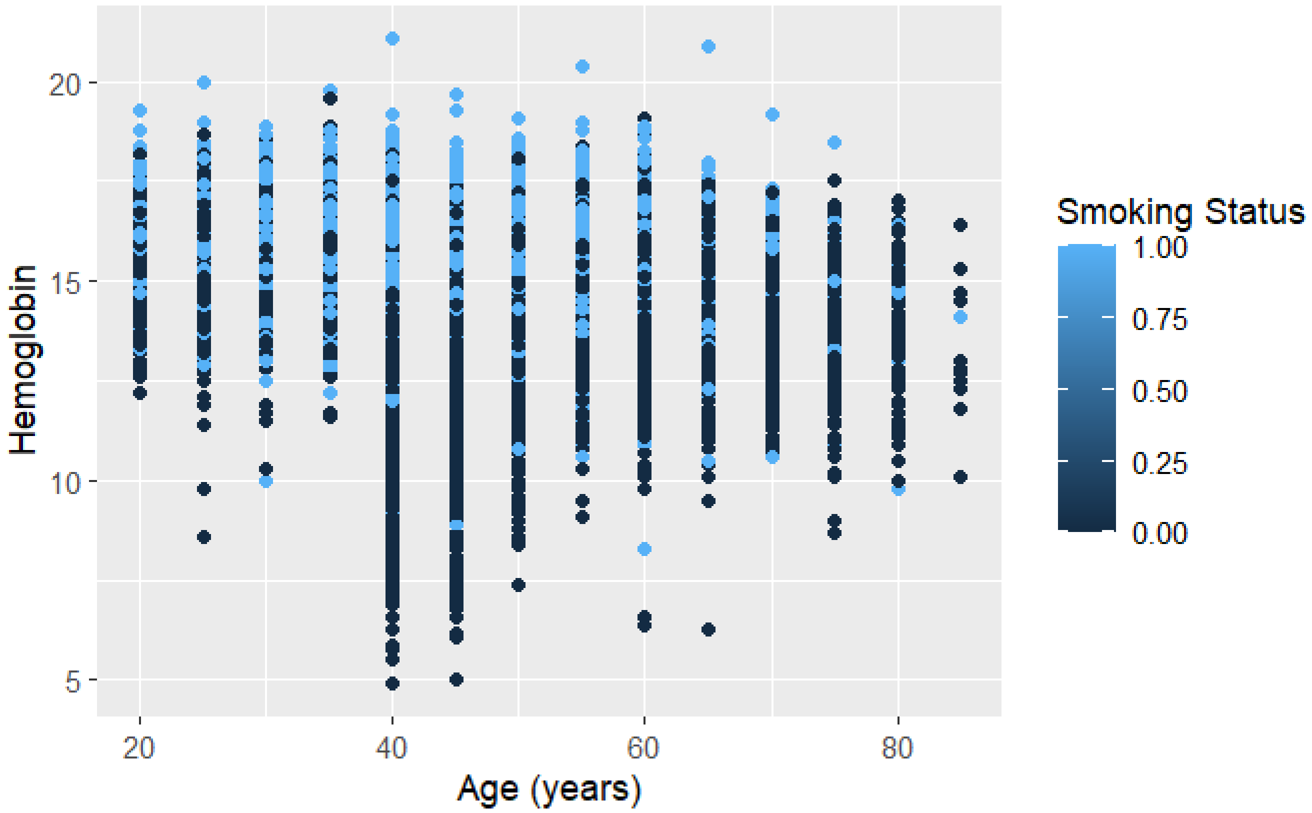
Hide

```
#scatter plot
# Load packages
library(tidyverse)

# Load data
data(df)
```

```
Warning: data set ‘df’ not found
```

Hide

```
# Create a scatter plot of mpg, disp, and cyl
# Create a scatter plot of age, height, and smoking
ggplot(df, aes(x = age, y = hemoglobin , color = smoking)) +
  geom_point() +
  labs(x = "Age (years)", y = "Hemoglobin", color = "Smoking Status")
```



Based on this scatter plot, we can see that the light blue colour represents the person who is a smoker and the darker blue colour represents the person who is not a smoker. Based on the scatter plot, people who are older are likely not a smoker. Moreover, people who smoke have an overly high level of hemoglobin. This is because the carbon monoxide in cigarette smoke blocks oxygen attachment to the red cells' empty hemoglobin slots, causing the body to increase red blood cell production.

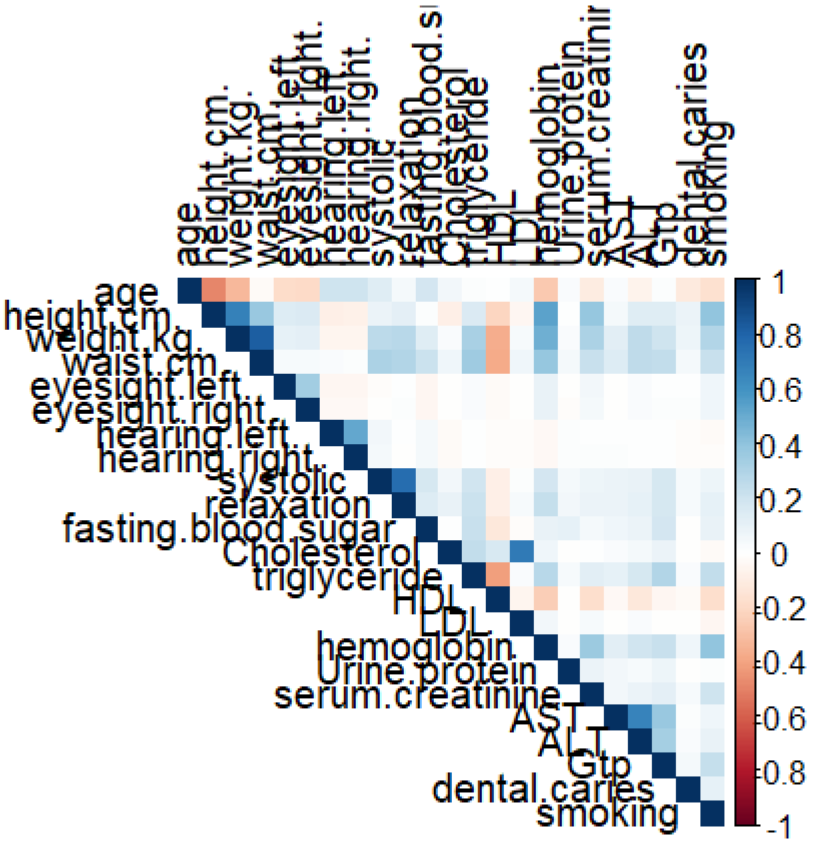
Hide

```
# Load packages
library(tidyverse)
library(corrplot)

# Load data (Replace 'df' with your actual dataframe)
data(df)
```

```
Warning: data set ‘df’ not found
```

Hide

```
# Calculate correlation matrix
cor_mat <- cor(df)

# Plot correlation matrix with larger size
corrplot(cor_mat, method = "color", type = "upper", tl.col = "black", tl.srt = 90)
```



Hide

```
NA
NA
```

From the heatmap, the darker blue colour represents the strong correlation between the variables. There appears to be a number that shows a negative number, which means that when the number is negative, the data has a relationship that is rather than reversed.

If the number is approaching zero, the variables do not correlate at all.

Based on the heatmap, when investigating the correlation between the response variable smoking and the explanatory variables, the explanatory variables that have a weak positive correlation with the highest to the response variable smoking among the attributes in this dataset are height.cm., weight.kg., waist,cm., triglyceride, hemoglobin, serum.creatinine and Gtp.

a. Investigate the basic model obtained from the data set.

Hide

```
logreg <- glm(formula = smoking ~ ., family = binomial(link="logit"), data = df)
summary(logreg)
```

```

Call:
glm(formula = smoking ~ ., family = binomial(link = "logit"), 
    data = df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.3034  -0.8208  -0.3947   0.9515   4.7542  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -1.993e+01  4.679e-01 -42.589  < 2e-16 ***
age                  8.450e-04  1.399e-03   0.604  0.54593    
height.cm.           8.282e-02  2.511e-03  32.974  < 2e-16 ***
weight.kg.          -2.441e-02  2.685e-03  -9.093  < 2e-16 ***
waist.cm.            1.453e-02  3.140e-03   4.627 3.72e-06 ***
eyesight.left.       2.359e-02  2.800e-02   0.842  0.39954    
eyesight.right.      1.786e-03  2.797e-02   0.064  0.94910    
hearing.left.       -9.824e-02  1.030e-01  -0.954  0.34018    
hearing.right.       1.169e-01  1.001e-01   1.168  0.24273    
systolic            -9.436e-03  1.564e-03  -6.032 1.62e-09 ***
relaxation           4.526e-03  2.154e-03   2.102  0.03559 *  
fasting.blood.sugar  4.174e-03  6.930e-04   6.024 1.71e-09 ***
Cholesterol         -5.055e-03  6.157e-04  -8.210  < 2e-16 ***
triglyceride         4.566e-03  2.476e-04  18.442  < 2e-16 ***
HDL                 -3.344e-03  1.244e-03  -2.689  0.00717 ** 
LDL                  2.828e-04  4.641e-04   0.609  0.54223    
hemoglobin           4.400e-01  1.268e-02  34.715  < 2e-16 ***
Urine.protein       -6.612e-02  3.372e-02  -1.961  0.04988 *  
serum.creatinine     1.755e-01  6.809e-02   2.578  0.00994 ** 
AST                 -3.490e-04  1.282e-03  -0.272  0.78549    
ALT                 -6.837e-03  9.982e-04  -6.849 7.45e-12 ***
Gtp                  9.747e-03  4.346e-04  22.428  < 2e-16 ***
dental.caries        4.104e-01  3.157e-02  13.000  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43972  on 33466  degrees of freedom
Residual deviance: 34100  on 33444  degrees of freedom
AIC: 34146

Number of Fisher Scoring iterations: 5
```

Hide

```
null_model <- glm(formula = smoking ~ 1, family = binomial(link="logit"), data = df)
summary(null_model)
```

```

Call:
glm(formula = smoking ~ 1, family = binomial(link = "logit"), 
    data = df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.9551  -0.9551  -0.9551   1.4173   1.4173  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.54825    0.01135  -48.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43972  on 33466  degrees of freedom
Residual deviance: 43972  on 33466  degrees of freedom
AIC: 43974

Number of Fisher Scoring iterations: 4
```

Hide

```
 #Install and load the 'car' package (if not already installed)
 #install.packages("car")
library(car)

# Assuming 'lm_model' is your linear regression model
# Make sure your model is fitted before calculating VIF

# Calculate VIF
vif_values <- car::vif(logreg)

# Print the VIF values
print(vif_values)
```

```
                age           height.cm.           weight.kg.            waist.cm.      eyesigh
t.left.     eyesight.right.       hearing.left. 
           1.687685             2.273006             6.248744             4.453792             1.
137716             1.139278             1.358622 
     hearing.right.             systolic           relaxation fasting.blood.sugar           Chole
sterol         triglyceride                  HDL 
           1.363709             2.447715             2.395523             1.139308             2.
808775             1.766252             1.694096 
                LDL           hemoglobin        Urine.protein    serum.creatinine              
AST                  ALT                  Gtp 
           2.359566             1.389577             1.022667             1.161300             2.
275334             2.791200             1.455968 
      dental.caries 
           1.012046 
```

Hide

```
anova(logreg, test="Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: smoking

Terms added sequentially (first to last)


                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                33466      43972              
age                  1    956.1     33465      43016 < 2.2e-16 ***
height.cm.           1   4698.2     33464      38317 < 2.2e-16 ***
weight.kg.           1     55.0     33463      38262 1.218e-13 ***
waist.cm.            1    233.9     33462      38029 < 2.2e-16 ***
eyesight.left.       1      2.2     33461      38026 0.1421187    
eyesight.right.      1      0.5     33460      38026 0.4785978    
hearing.left.        1      0.0     33459      38026 0.8934461    
hearing.right.       1      1.0     33458      38025 0.3084187    
systolic             1     11.9     33457      38013 0.0005614 ***
relaxation           1     84.9     33456      37928 < 2.2e-16 ***
fasting.blood.sugar  1    220.6     33455      37707 < 2.2e-16 ***
Cholesterol          1      2.8     33454      37705 0.0919096 .  
triglyceride         1   1099.1     33453      36606 < 2.2e-16 ***
HDL                  1      5.6     33452      36600 0.0178052 *  
LDL                  1      1.0     33451      36599 0.3084939    
hemoglobin           1   1617.7     33450      34981 < 2.2e-16 ***
Urine.protein        1      0.6     33449      34981 0.4505911    
serum.creatinine     1     11.7     33448      34969 0.0006299 ***
AST                  1      2.5     33447      34966 0.1104928    
ALT                  1      6.6     33446      34960 0.0099481 ** 
Gtp                  1    691.0     33445      34269 < 2.2e-16 ***
dental.caries        1    168.7     33444      34100 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results & Interpretation

Significance Codes: ‘***’: Very highly significant (p-value < 0.001).*** ’’: Highly significant (p-value < 0.01). ”*: Significant at a 5% level (p-value < 0.05). ’ ’: Not significant (p-value > 0.1).

Results from the full logistic regression model show that there are some insignificant variables such as”age “,”eyesight.left.”, “eyesight.right.”, “hearing.left.”, “hearing.right.”, “Cholesterol”, “LDL”,“AST” ” and “relaxation” based on its respective p-values.

Variance Inflation Factors (VIF).

The Variance Inflation Factor (VIF) measures the inflation in the coefficient of the independent variable due to the collinearities among the other independent variables.

A VIF of 1 means that the regression coefficient is not inflated by the presence of the other predictors, and hence multicollinearity does not exist.

Ideally, the Variance Inflation Factors are below 5.

Results from multicollinearity with VIF test shows that weight.kg. (VIF: 6.23): This variable has a relatively high VIF, indicating that its variance is inflated due to its correlation with other predictors and waist.cm. (VIF: 4.43): While this VIF is above 2, it is not extremely high. However, it suggests some correlation with other predictors.

Results from the ANOVA test show that “eyesight.left.”, “eyesight.right.”, “hearing.left.”, “hearing.right.”, “Cholesterol”, “LDL”, “Urine.protein”, and “AST” are insignificant to the fitted model.

Hence, fit the updated glm() model (logistic regression) as a reduced model without the insignificant variables and non-collinear variables into logreg2 in order to develop the best model in our case study.

b. Develop the best model as your solution based on this course.

Hide

```
### Fit the updated glm() model (logistic regression) into logreg2.
logreg2 <- glm(formula = smoking ~ age + height.cm. + systolic +
                  fasting.blood.sugar + triglyceride  +
                 HDL + hemoglobin + serum.creatinine + ALT + Gtp + dental.caries,
               family = binomial(link="logit"),
               data = df)

summary(logreg2)
```

```

Call:
glm(formula = smoking ~ age + height.cm. + systolic + fasting.blood.sugar + 
    triglyceride + HDL + hemoglobin + serum.creatinine + ALT + 
    Gtp + dental.caries, family = binomial(link = "logit"), data = df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.3039  -0.8216  -0.4068   0.9589   5.4841  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -1.849e+01  3.781e-01 -48.888  < 2e-16 ***
age                  3.241e-03  1.281e-03   2.529   0.0114 *  
height.cm.           7.032e-02  2.030e-03  34.636  < 2e-16 ***
systolic            -9.216e-03  1.054e-03  -8.745  < 2e-16 ***
fasting.blood.sugar  4.213e-03  6.854e-04   6.148 7.87e-10 ***
triglyceride         3.474e-03  2.146e-04  16.185  < 2e-16 ***
HDL                 -5.853e-03  1.075e-03  -5.447 5.13e-08 ***
hemoglobin           4.278e-01  1.248e-02  34.268  < 2e-16 ***
serum.creatinine     9.990e-02  6.859e-02   1.456   0.1453    
ALT                 -8.969e-03  7.179e-04 -12.493  < 2e-16 ***
Gtp                  9.903e-03  4.288e-04  23.097  < 2e-16 ***
dental.caries        4.089e-01  3.144e-02  13.008  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43972  on 33466  degrees of freedom
Residual deviance: 34369  on 33455  degrees of freedom
AIC: 34393

Number of Fisher Scoring iterations: 5
```

Hide

```
### Use the anova() function to analyze the updated table of deviance.
anova(logreg2, test="Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: smoking

Terms added sequentially (first to last)

                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                33466      43972
age                  1    956.1     33465      43016 < 2.2e-16
height.cm.           1   4698.2     33464      38317 < 2.2e-16
systolic             1     40.7     33463      38277 1.788e-10
fasting.blood.sugar  1    281.3     33462      37995 < 2.2e-16
triglyceride         1   1116.1     33461      36879 < 2.2e-16
HDL                  1     21.0     33460      36858 4.532e-06
hemoglobin           1   1568.3     33459      35290 < 2.2e-16
serum.creatinine     1      5.7     33458      35284   0.01699
ALT                  1      5.8     33457      35279   0.01649
Gtp                  1    740.9     33456      34538 < 2.2e-16
dental.caries        1    168.8     33455      34369 < 2.2e-16

NULL
age                 ***
height.cm.          ***
systolic            ***
fasting.blood.sugar ***
triglyceride        ***
HDL                 ***
hemoglobin          ***
serum.creatinine    *
ALT                 *
Gtp                 ***
dental.caries       ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results & Interpretation

Results from the updated fitted logistic regression model shows that only “serum.creatinine” variable is insignificant based on its p-value.

Additionally, results from the ANOVA test shows all variables, including “serum.creatinine”, are significant to the fitted model.

This is judged by the low deviance residuals as well as the Pr(>Chi) of > 0.05, respectively.

Hence, the “serum.creatinine” variable is kept in the fitted model.

Hide

```
car::vif(logreg2)
```

```
                age          height.cm.            systolic 
           1.421810            1.509450            1.118942 
fasting.blood.sugar        triglyceride                 HDL 
           1.116290            1.345118            1.269627 
         hemoglobin    serum.creatinine                 ALT 
           1.362957            1.157226            1.413102 
                Gtp       dental.caries 
           1.443219            1.011522 
```

Results show that all the variables have a VIF value of < 5 .

Test for the Absence of Strongly Influential Outliers

Notes

Test using standardized residuals and Cook’s Distance.

Standardized residuals values > 3 = influential outlier.

Cook’s D value > Cook’s D Threshold (4/N) = influential outlier.

Hide

```
### Place all the calculated values from the logistic regression model into a new data frame.
library(dplyr)
library(magrittr)
```

```


Attaching package: ‘magrittr’

The following object is masked from ‘package:purrr’:

    set_names

The following object is masked from ‘package:tidyr’:

    extract
```

Hide

```
library(tidyverse)
library(broom)
```

```
Warning: package ‘broom’ was built under R version 4.2.3
```

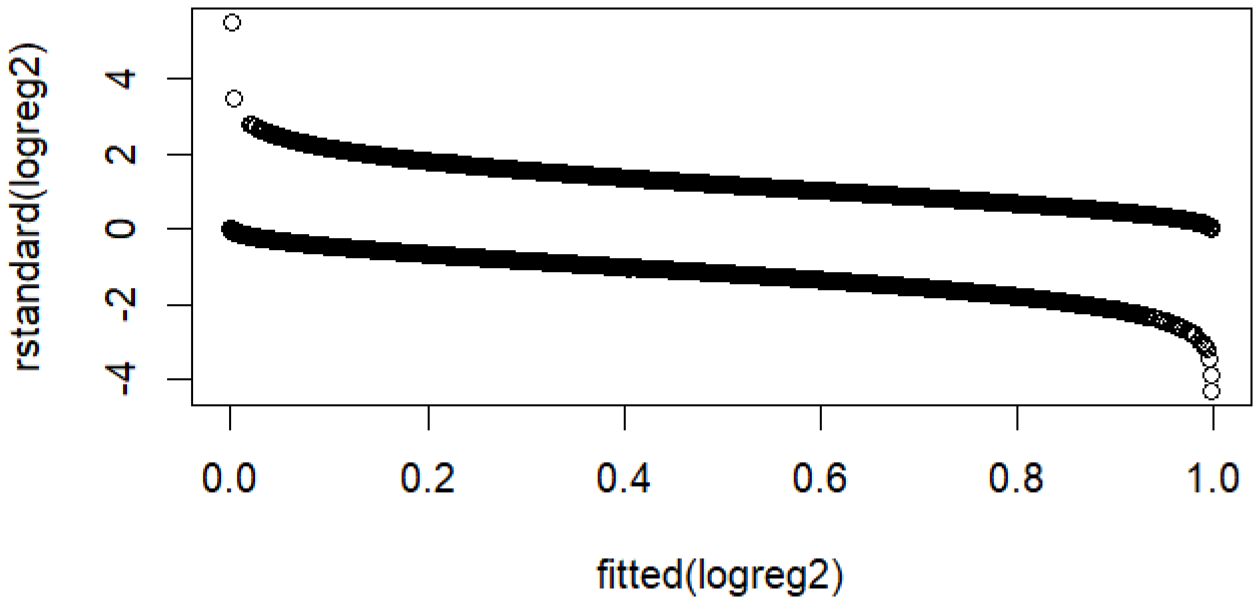
Hide

```
logreg.data <- augment(logreg2) %>%
  mutate(index = 1:n())

### Show the top 6 highest standardized residuals (if > 3 = influential outlier).
head(logreg.data$.std.resid[order(-logreg.data$.std.resid)])
```

```
[1] 5.484120 3.451080 2.800086 2.791547 2.771910 2.737196
```

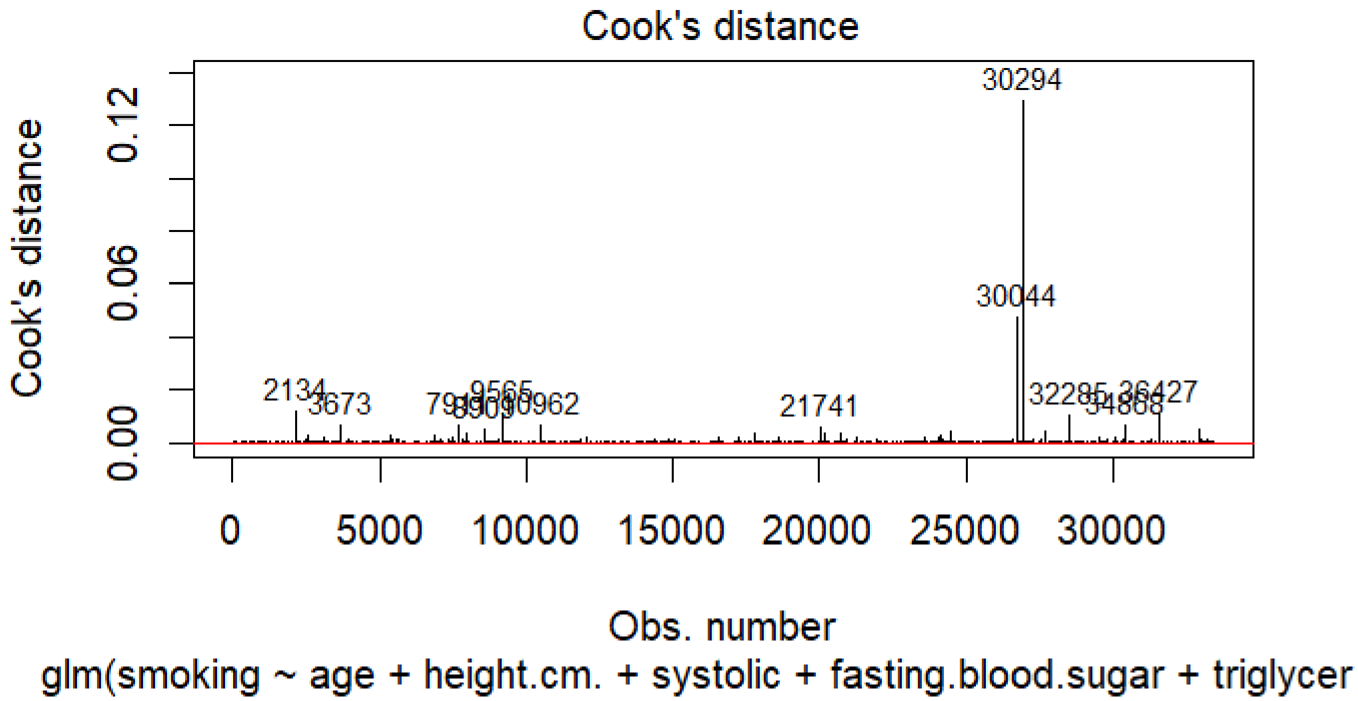
Hide

```
### Plot of standardized residuals
plot(fitted(logreg2),
     rstandard(logreg2))
```



Hide

```
### Set Cook's D Threshold.
cook_threshold <- 4 / nrow(df)

### Cook's D Plot.
plot(logreg2, which = 4, id.n = 12)
abline(h = cook_threshold, col = "red")
```



Hide

```
### Put outlier data into a new data frame where > Cook's D Threshold = influential outliers.
influ_out <- logreg.data %>%
  filter(.cooksd > cook_threshold)

### Get the percentage of influential outliers.
outliers <- round(100*(nrow(influ_out) / nrow(logreg.data)),1)
```

Hide

```
### Get the percentage of influential outliers.
outliers <- round(100*(nrow(influ_out) / nrow(logreg.data)),1)

### Store values in a data variable.
print_outliers <- format(round(outliers, 2), nsmall = 2)

### Print the number of percentage of observations that exceed Cook's distance threshold.
sprintf('Proportion of data points that are highly influential = %s Percent', print_outliers)
```

```
[1] "Proportion of data points that are highly influential = 4.10 Percent"
```

Results & Interpretation

Standardized Residuals.

Results show that none of the data points of the fitted model consist of any outliers.

Cook's Distance.

In addition, based on the pre-defined threshold (4/N), only 4.1% of the data points are in the outlier zone, which is small as well.

c. Briefly explain your final model which consider as the best solution.

The final logistic regression model was constructed to predict smoking status by considering various health-related variables. Through a stepwise refinement process, insignificant predictors and potential multicollinearity issues were addressed. The resulting model includes significant predictors such as age, height, systolic blood pressure, fasting blood sugar, triglyceride levels, HDL cholesterol, hemoglobin levels, ALT (Alanine Aminotransferase), GTP (Gamma-Glutamyl Transferase), and the presence of dental caries. Notably, serum creatinine levels were found to be statistically insignificant in the updated model. Diagnostic tests, including VIF for multicollinearity, standardized residuals and Cook’s Distance for influential outliers, and deviance tests, were conducted to ensure the model’s robustness. The findings indicate that the model provides a significant fit without strong evidence of multicollinearity or influential outliers. Interpretation of the coefficients suggests how each predictor contributes to the likelihood of smoking based on individual health characteristics, providing a valuable tool for understanding and predicting smoking behavior.