Biostatistics Report

Matej Rojec, Patrycja Ozgowicz

May 28, 2023

Contents

List of Figures

List of Tables

Abstract

In the project we analysed the impact of different covariets to understand the impact of them to glycosolated hemoglobin and to diabetes. We conducted a thorough exploratory data analysis, we fitted a gamma regression model, interpreted the model. We concluded that the covariates that have the biggest impact on glycosolated hemoglobin and therefore diabetes are total cholesterol, stabilized glucose, age and waist length in inches.

1 Introduction

The main goal of the project is to investigate the relationship between the glycosylated hemoglobin and the associated set of the observed risk factors using a gamma regression model. Glycosylated hemoglobin is a form of hemoglobin that is chemically linked to a sugar. The formation of the sugar-hemoglobin linkage indicates the presence of excessive sugar in the bloodstream. The level of glycosylated hemoglobin higher than 7% is usually taken as a positive diagnosis of diabetes. In this project, we develop the model, which relate glycosylated hemoglobin with other factors in the most effective way. Our data are related to African Americans blood test results and body measurements. In this work, firstly, we introduce the diabetes concept. Then we perform a thorough exploratory data analysis of the data and describe a mathematical framework. We use a gamma regression model, which fitting and selection is described in detail. At the end, we provide interpretation of the associated results with respect to the study objectives.
The benefit of the project may be identifying relevant features which significantly impact level of glycosylated hemoglobin and as a result cause diabetes. Such analysis may be useful for people who would like to avoid diabetes. In addition, it might also be valuable for doctors who will be able to better understand the basis of the disease.

1.1 Diabetes

Diabetes is a chronic (long-lasting) health condition that affects how the body turns food into energy. The body breaks down most of the food into sugar (glucose) and releases it into the bloodstream. When blood sugar goes up, it signals the pancreas to release insulin. Insulin acts like a key to let the blood sugar into body's cells for use as energy. With diabetes, the body doesn't make enough insulin or can't use it as well as it should. When there isn't enough insulin or cells stop responding to insulin, too much blood sugar stays in the bloodstream. There are three main types of diabetes [10]:

1. **Type 1 diabetes** is thought to be caused by an autoimmune reaction (the body attacks itself by mistake). This reaction stops the body from making insulin. Approximately $5 - 10\%$ of the people who have diabetes

have type 1. Symptoms of type 1 diabetes often develop quickly. It's usually diagnosed in children, teens, and young adults.

2. **Type 2 diabetes** is when the body doesn't use insulin well and can't keep blood sugar at normal levels. About $90-95\%$ of people with diabetes have type 2. It develops over many years and is usually diagnosed in adults (but more and more in children, teens, and young adults).

3. **Gestational Diabetes** develops in pregnant women who have never had diabetes. Gestational diabetes usually goes away after a baby is born. However, it increases the risk for type 2 diabetes later in life.

2 Exploratory data analysis

2.1 Data characteristics

We consider data, which consist of subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. The data set contains a data frame with 390 observations on the following 14 variables that are a mixture of categorical and numerical data types. Description of each of the variables is in the table 1.

Table 1: Description of attributes

Variable name	Type	Description
ID	integer	Subject Identification
CHOL	integer	Total Cholesterol
SGLU	integer	Stabilized Glucose
GHB	numeric	Glycosolated Hemoglobin
LOCATION	factor	A factor with levels: Buckingham and Louisa
AGE	integer	Age (years)
GENDER	factor	Gender: male or female
HHT	integer	Height (inches)
WHT	integer	Weight (pounds)
FRAME	factor	A factor with levels: small, medium and large
SBP	integer	First Systolic Blood Pressure
DSP	integer	First Diastolic Blood Pressure
W	integer	Waist (inches)
H	integer	Hip (inches)

Basic information about data is presented in the table 2.

Table 2: Basic information about dataset.

rows	390
columns	14
discrete columns	3
continuous columns	11
all missing columns	0
total missing values	32
complete rows	367
total observations	5460

2.2 Quick overview

Here's a brief description of each covariate in our dataset:

1. CHOL: CHOL refers to cholesterol levels, which are a measure of the amount of cholesterol present in the blood. High cholesterol levels can be associated with an increased risk of cardiovascular diseases.

2. SGLU: SGLU stands for blood sugar levels, specifically measuring the concentration of glucose in the blood. Abnormal blood sugar levels can be indicative of conditions such as diabetes or impaired glucose metabolism.

3. LOCATION: LOCATION represents the location or region of the individual. This variable is categorical, representing different geographic areas, representing a specific location metric.

4. AGE: AGE represents the age of the individual. Age is a significant covariate in many health-related models, as certain health conditions and risks can be associated with specific age ranges.

5. GENDER: GENDER represents the gender of the individual, categorized as male or female. Gender can influence various health factors and risks, making it an important covariate in many models.

6. HHT: HHT represents height, which is the vertical measurement of an individual's stature. Height is an important anthropometric measurement used in various health assessments.

7. WHT: WHT stands for weight, which represents the individual's body weight. Body weight can be a relevant factor in predicting health outcomes and risks.

8. FRAME: FRAME represents the body frame or body type of the individual. It categorize individuals as having a small, medium, or large frame, which can influence factors like body composition and metabolism.

9. SBP: SBP represents systolic blood pressure, which is the pressure in the arteries when the heart beats. Elevated systolic blood pressure can be an indicator of hypertension or other cardiovascular issues.

10. DSP: DSP refers to diastolic blood pressure, which is the pressure in the arteries when the heart is at rest between beats. High diastolic blood pressure can also be a sign of hypertension or other cardiovascular conditions.

11. W: W represents waist circumference, which is a measure of the waist size. Waist circumference is often used as an indicator of central obesity and can be associated with various health risks.

12. H: H represents hip circumference, which is a measure of the hip size. It is, similarly as waist circumference, often used as an indicator of obesity and can be associated with various health risks.

The distribution of the numerical variables can be seen in figure 1, but we will take a closer look at each covariate will be described in the next subchapters.

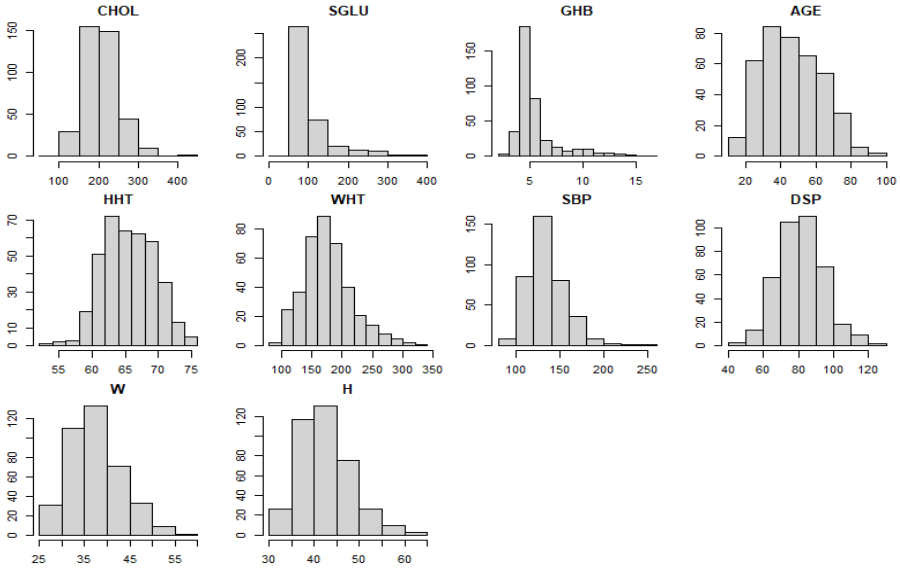

Figure 1: Distributions of the variables.

2.3 Outliers

An important part of any data analysis is seeing if out data includes outliers. We can spot these using the help of box plots. These can be seen in figure 2. We can see from the figure that the variables SGLU, GHB & CHOL have the highest dispersion around the mean. Hence they have the most outliers.

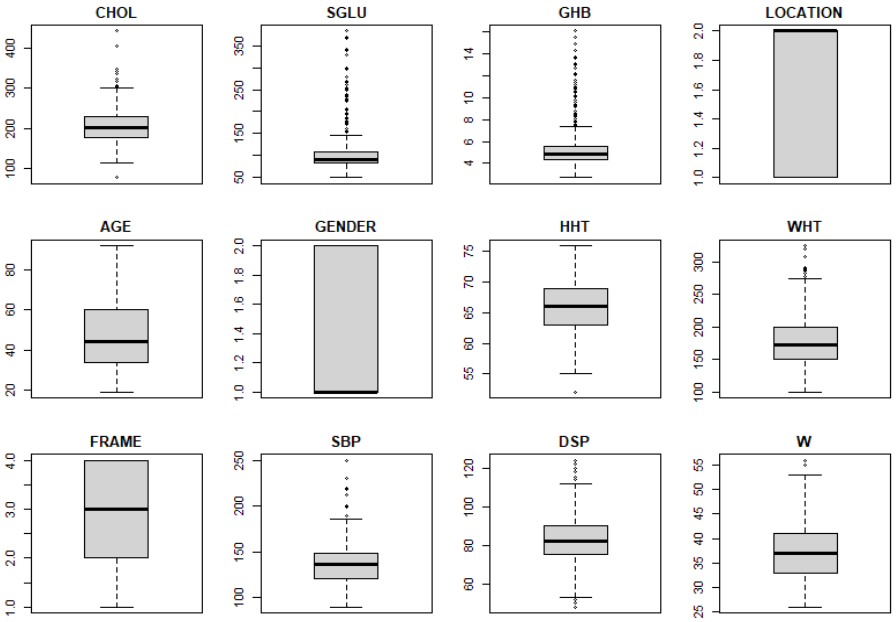

Figure 2: Outliers in the data set.

2.4 Missing values

Initially, missing values were marked in data as empty space in a cell. We changed it and represented by the symbol NA (not available). The occurrence of these values in the columns is shown in plot 3. From the graph, we can read that the biggest part of missing values is in column "FRAME". All rows without measurements in column "W" also had no data in column "H". The same with columns "SBP" and "DSP". In total, there are only 23 rows with missing values, which is $5,9\%$ of all rows. Since the relative amount of is low, we will remove these rows when fitting the models.

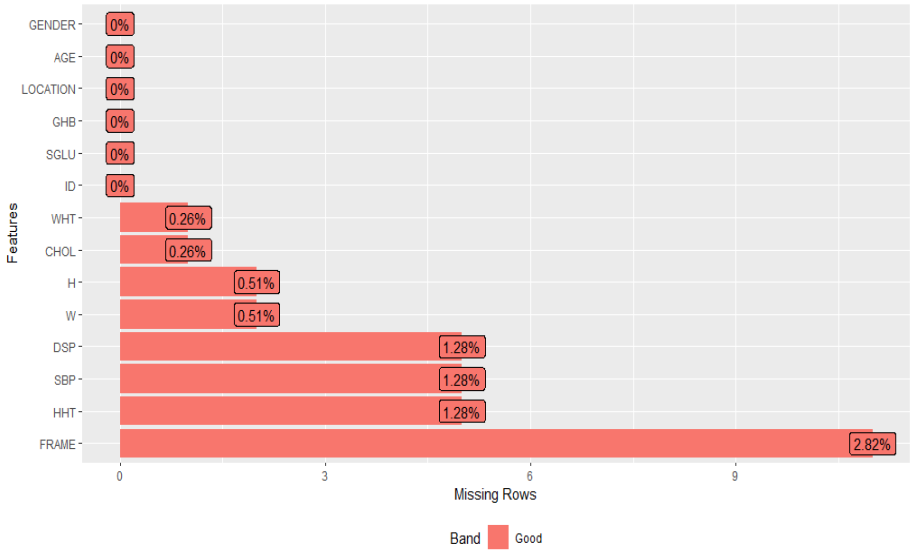

Figure 3: Occurrence of missing values in individual columns.

2.5 Analysis of individual variables

To get familiar with data, we have decided to plot all features. The first one is "ID". It is just an identifier and has no information about data, so we have decided to remove it. Next attributes are described in sections below.

2.5.1 Glycosylated Hemoglobin - GHB

Hemoglobin is the substance inside red blood cells that carries oxygen to the cells of the body. Glucose is a type of sugar in blood that comes from the food. Glucose molecules in the blood normally become stuck to hemoglobin molecules — this means the hemoglobin has become glycosylated (also referred to as hemoglobin A1c or HbA1c). As a person's blood sugar becomes higher, more of the person's hemoglobin becomes glycosylated. The glucose remains attached to the hemoglobin for the life of the red blood cell, or about 2 to 3 months. A blood test can measure the amount of glycosylated hemoglobin in the blood. [2]

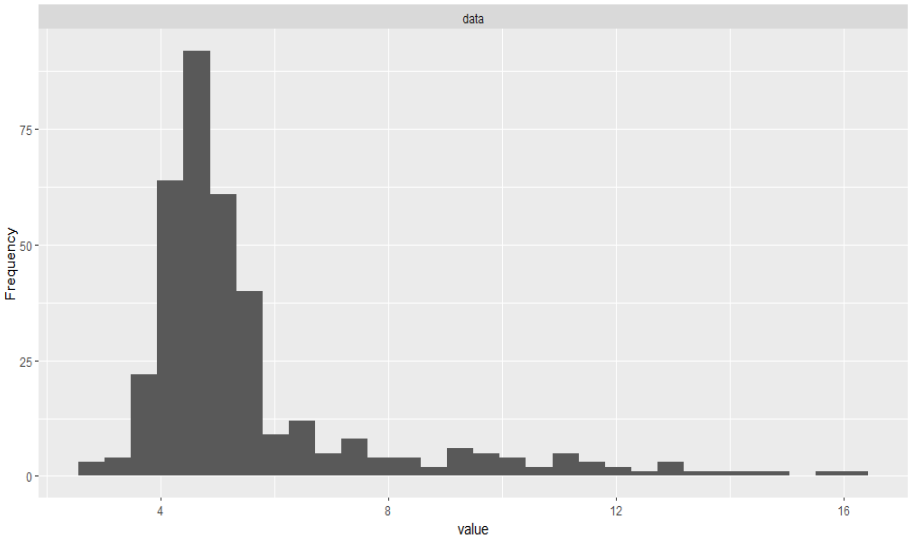

Figure 4: Histogram of feature "Glycosylated Hemoglobin".

Table 3: Short summary statistics for "Glycosylated Hemoglobin" feature.

	GHB
min	2.68
Q1	4.39
median	4.86
mean	5.60
Q3	5.63
max	16.11
var	4.97
sd	2.23
IQR	1.24

The glycosylated hemoglobin test shows what a person's average blood glucose level was for the 2 to 3 months before the test. The normal level for glycosylated hemoglobin is less than 7%. If the result is higher, we usually take it as a positive diagnosis of diabetes. On the plot 4 we observe that the distribution of the data is right-skewed. Also, Table 3 confirms that, saying that the mean is greater than the median. Both the mean and the median are below 7%, which indicate that most people are healthy. Indeed, there are 311 subject with glycosylated hemoglobin less than 7% and 56 above this level. It means 15.26% of all people have diabetes. Few of them have extremely high level of glycosylated hemoglobin, since we can see results above 10% and maxi-

mum value equal 16.11%. Here, we are dealing with unequally distributed data and imbalance class. Only for visualization purposes, we create a new variable named "diabetes" with levels yes and no, which distinguish healthy subjects from those with diabetes.

2.5.2 Total Cholesterol — CHOL

Feature "CHOL" is a measure of the total amount of cholesterol in subject's blood. Results are given in milligrams per deciliter (mg/dL).

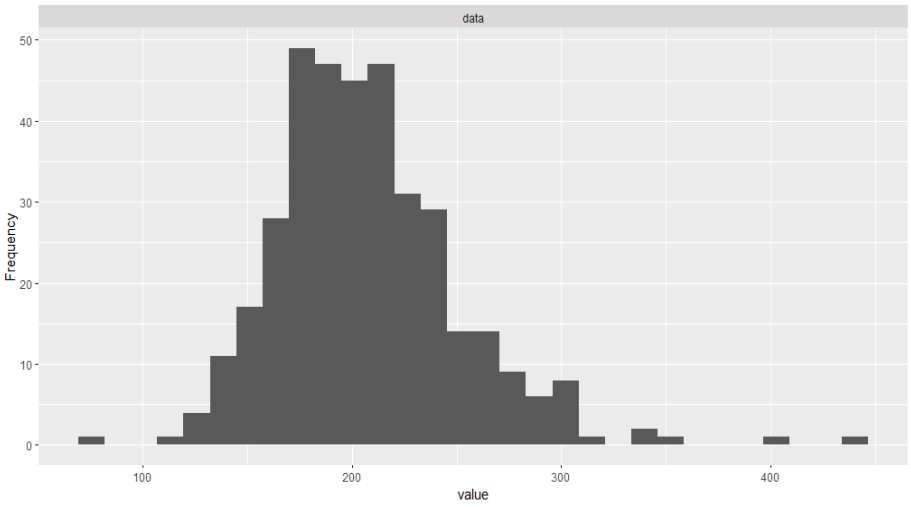

Figure 5: Histogram of feature "Total Cholesterol".

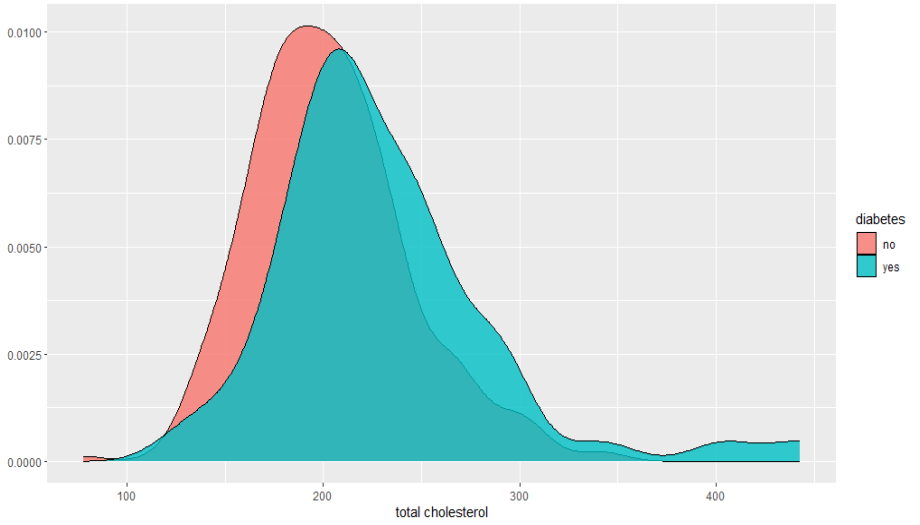

Figure 6: "Total Cholesterol" feature distribution by information whether a subject has diabetes or not.

Table 4: Short summary statistics for "Total Cholesterol" feature.

	CHOL
min	78.00
Q1	179.00
median	204.00
mean	207.54
Q3	229.00
max	443.00
var	1935.51
sd	43.99
IQR	50.00

From the histogram (figure 5) and Table 4, we see that the values are in the range between 78 and 443 mg/dL. The healthy range for total cholesterol in adults is considered to be from 125 to 200 mg/dL. Borderline high is from 200 to 239 mg/dL and high is at or above 240 mg/dL. Most of the time, very low cholesterol doesn't cause a problem, while high cholesterol is very dangerous. It can enter artery wall, damage its integrity and lead to the formation of atherosclerotic plaque (hardened deposits). It can lead to serious problems, e.g. block blood flow to heart, legs and arms or brain. Ideal total cholesterol level should be around 150. [1] From the table 4 we can observe that mean and median are slightly above the normal range. 46% of the study participants are in healthy range. Only two people are below 125 mg/dL and the rest are above

200 mg/dL. On the plot with distribution 6, we observe that most people who have diabetes have higher total cholesterol than those who are healthy.

2.5.3 Stabilized Glucose — SGLU

Glucose is a sugar that mainly comes from carbohydrates in the food and drinks. It's the body's main source of energy. Blood carries glucose to all the body's cells to use for energy. Several bodily processes help keep the blood glucose in a healthy range. Insulin, a hormone the pancreas makes, is the most significant contributor to maintaining healthy blood sugar. A healthy fasting blood glucose level for someone without diabetes is 70 to 99 mg/dL. A blood sugar result of 70 mg/dL or lower is usually considered low. If the fasting blood glucose level is 100 to 125 mg/dL, it usually means prediabetes. People with prediabetes have up to a 50% chance of developing Type 2 diabetes over the next five to ten years. But they can take steps to prevent Type 2 diabetes from developing. If the fasting blood glucose level is 126 mg/dl or higher, it is high and usually means the patient has diabetes. [9]

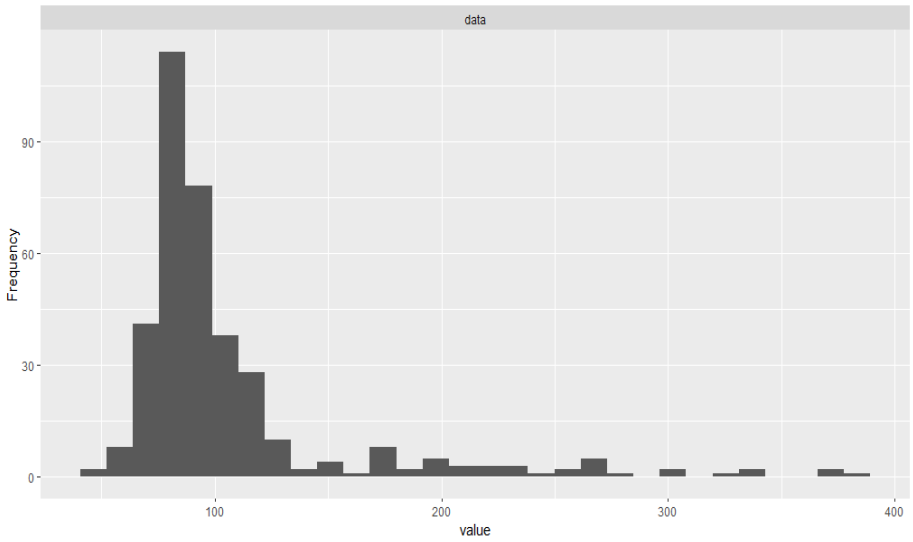

Figure 7: Histogram of feature "Stabilized Glucose".

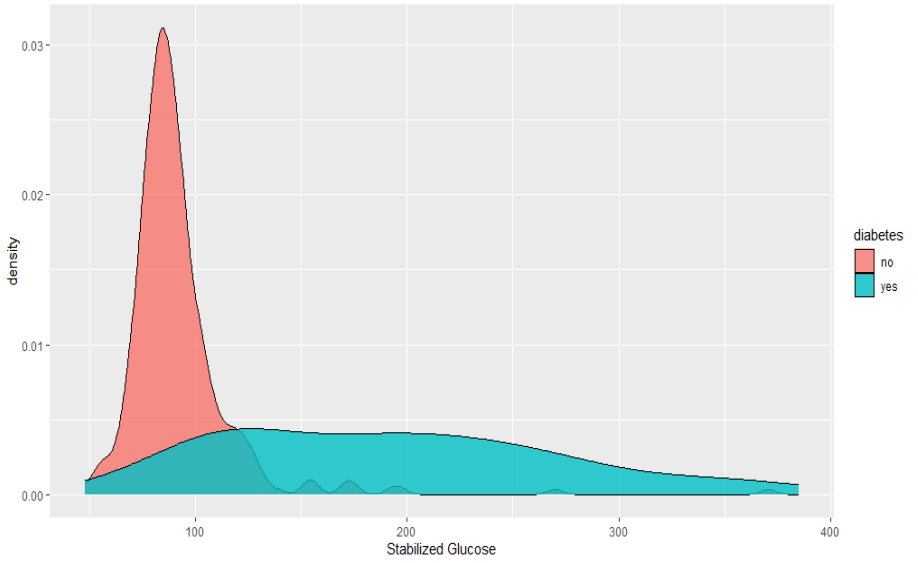

Figure 8: "Stabilized Glucose" feature distribution by information whether a subject has diabetes or not.

Table 5: Short summary statistics for "Stabilized Glucose" feature.

	SGLU
min	48.00
Q1	81.00
median	90.00
mean	107.33
Q3	108.00
max	385.00
var	2921.92
sd	54.05
IQR	27.00

Plot 7 presents the histogram of the "Stabilized Glucose" feature. We can observe that its distribution is really similar to the Glycosylated Hemoglobin one on figure 4. It's also right-skewed and have many outliers, which are much bigger than the mean. From the Table 5 we can read that 75% of subjects have glucose level below 108, which means they are not in the high range. However, maximum value equal to 385, which is extremely high. On the density plot 8 we can see that people can have diabetes, even if their glucose level is low. However, these are only few cases, and generally we can observe that most of the subjects with glucose level below 99 are healthy. We can see that almost

all people with high glucose level have diabetes. It indicates strong correlation between Stabilized Glucose and Glycosylated Hemoglobin.

2.5.4 First Blood Pressure — SBP

Blood pressure is the pressure of blood pushing against the walls of arteries. Arteries carry blood from the heart to other parts of the body. Blood pressure normally rises and falls throughout the day. Systolic blood pressure, measures the pressure in the arteries when the heart beats. It measures the force the heart exerts on the walls of the arteries each time it beats. Results are given in millimetres of mercury (mm Hg). Systolic blood pressure categories according to *The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure* [4]:

- Low: below 90;
- Normal: 90 - 119;
- At Risk: 120–139;
- High: above or at 140.

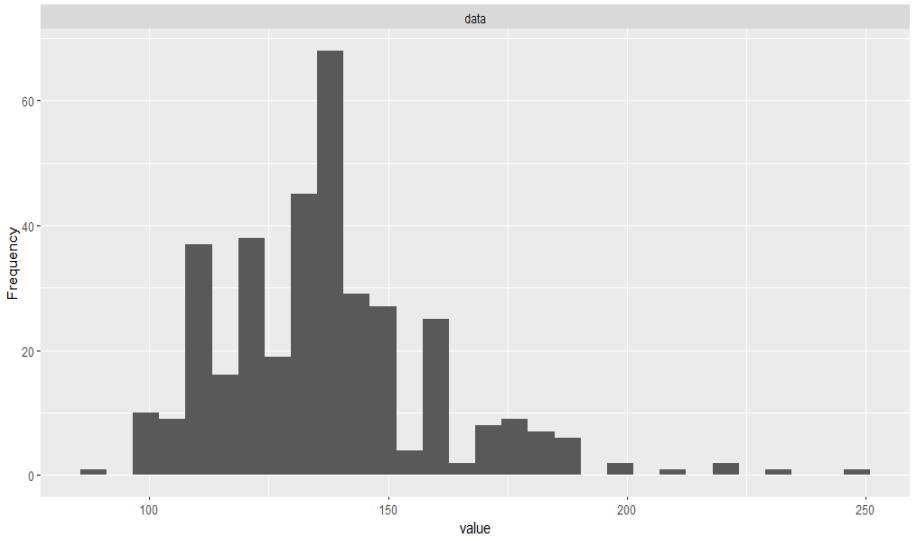

Figure 9: Histogram of feature "First Systolic Blood Pressure".

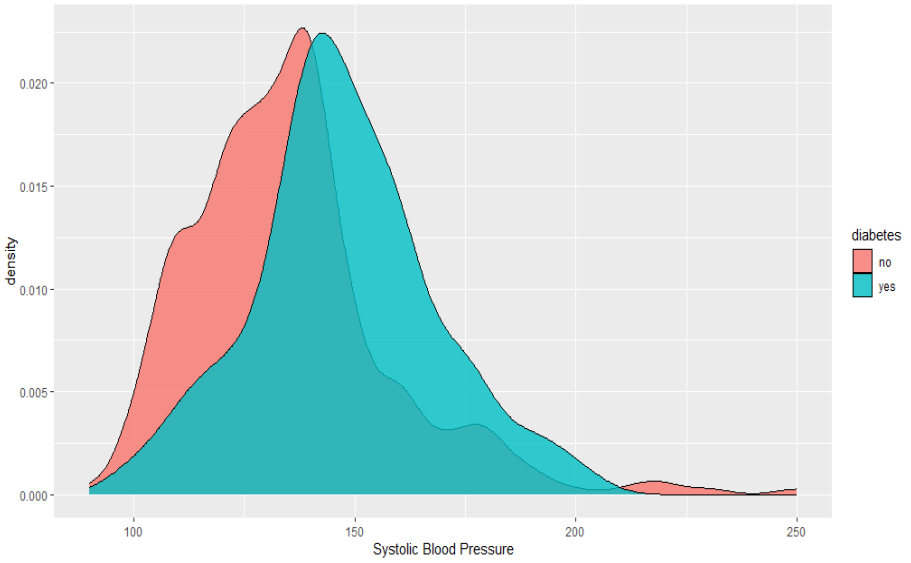

Figure 10: "First Systolic Blood Pressure" feature distribution by information whether a subject has diabetes or not.

Table 6: Short summary statistics for "First Systolic Blood Pressure" feature.

	SBP
min	90.00
Q1	121.50
median	136.00
mean	137.11
Q3	148.00
max	250.00
var	531.71
sd	23.06
IQR	26.50

On the plot 9 we see distribution of systolic blood pressure measurement results. We can observe few extremely high values above 180 mm Hg. Those are really risky results, which pose a threat to the health or life of the patient, especially the maximum value equal to 250 mmHg. The lowest value read from the table 6 is equal to 90, which means there is no subject with too low blood pressure. Only less than 25% of people qualify to the normal category. The rest are at risk or have high blood pressure. Figure 10 presents a density plot of "First Systolic Blood Pressure" feature divided by information whether a subject has diabetes or not. We can observe that the results of people with

diabetes are higher than those of healthy people. The vast majority of people with diabetes have systolic blood pressure above the normal range. However, people with systolic blood pressure higher than 250 do not have diabetes.

2.5.5 First Diastolic Blood Pressure — DSP

Diastolic blood pressure, measures the pressure in your arteries when the heart rests between beats. Results are also given in millimetres of mercury (mm Hg). Diastolic blood pressure categories according to *The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure* [4]:

- Low: below 60;
- Normal: 69-79;
- At Risk: 80–89;
- High: above or at 90.

Over the years, research has found that both systolic and diastolic blood pressures are equally important in monitoring heart health. However, most studies show a greater risk of stroke and heart disease related to higher systolic pressures compared with elevated diastolic pressures. [3]

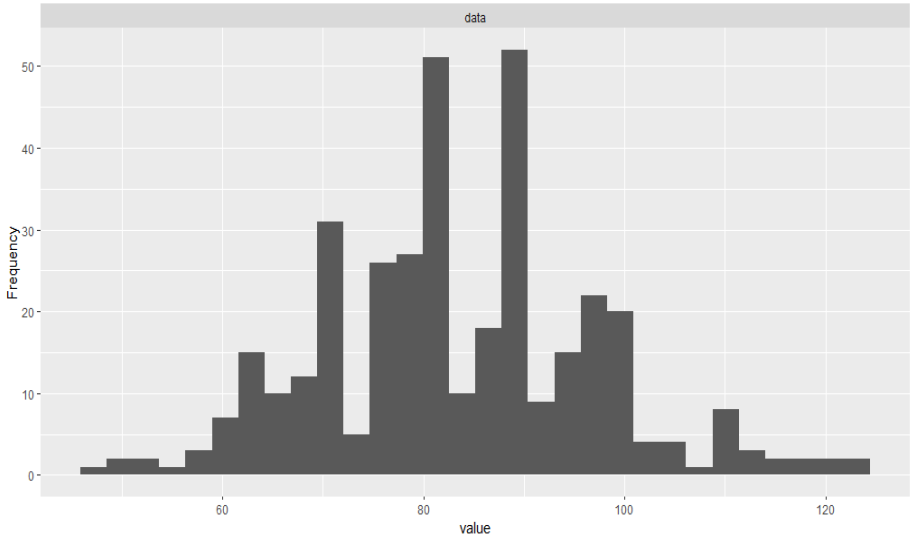

Figure 11: Histogram of feature "First Diastolic Blood Pressure".

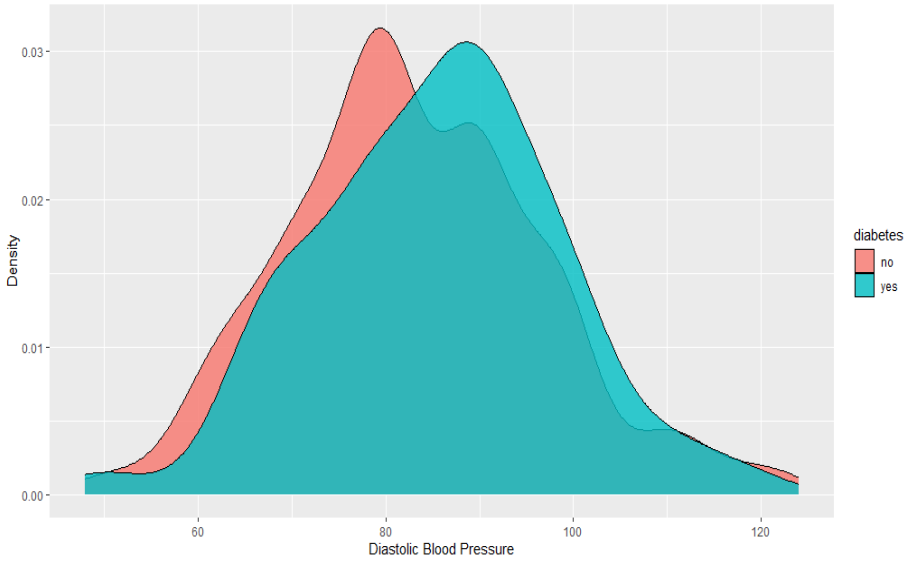

Figure 12: "First Diastolic Blood Pressure" feature distribution by information whether a subject has diabetes or not.

Table 7: Short summary statistics for "First Diastolic Blood Pressure" feature.

	DBP
min	48.00
Q1	75.00
median	82.00
mean	83.40
Q3	92.00
max	124.00
var	186.37
sd	13.65
IQR	17.00

From the plot 11 we can observe that most of the measurements are between 70 and 90, but there are some outliers in both directions. According to the table 7 minium value is 48 and maximum is 124. It means we have people with too low and too high blood pressure. However, there are many more patients with high diastolic blood pressure — 128. By looking at the density plot 12 we can't see any significant trend that would indicate that higher diastolic blood pressure leads to diabetes. The two densities have similar shape, the blue one, describing the distribution of people with diabetes, is slightly moved to the right comparing with the red one.

2.5.6 Age

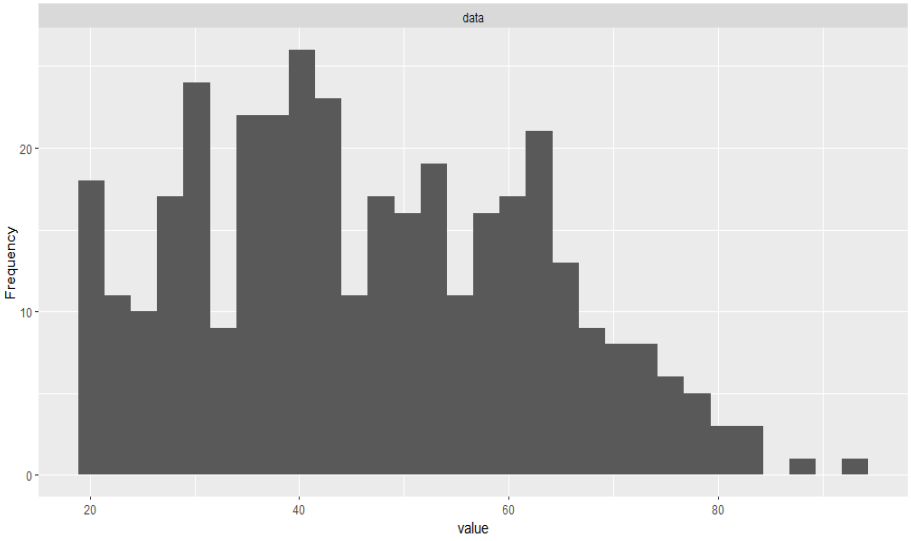

Figure 13: Histogram of feature "Age".

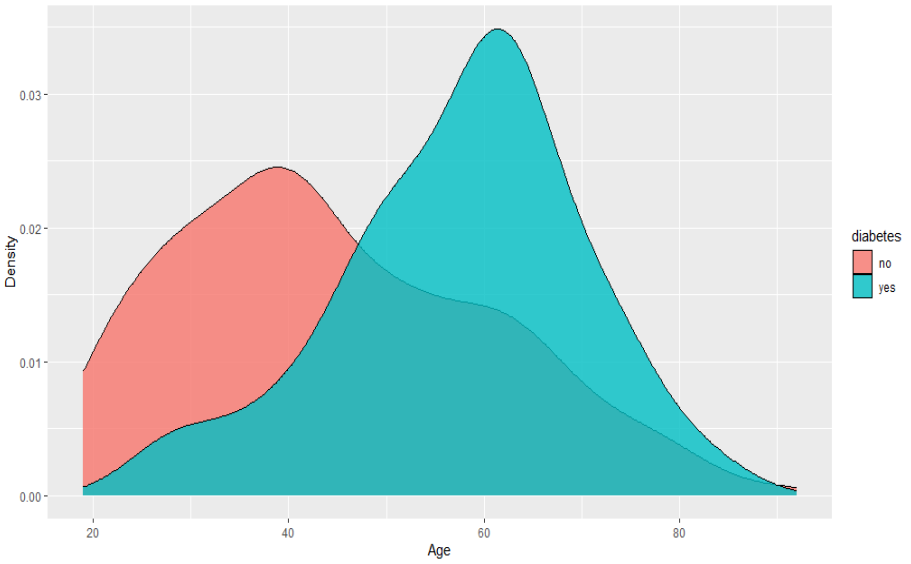

Figure 14: "Age" feature distribution by information whether a subject has diabetes or not.

Table 8: Short summary statistics for "Age" feature.

	AGE
min	19.00
Q1	34.00
median	45.00
mean	46.68
Q3	60.00
max	92.00
var	264.95
sd	16.28
IQR	26.00

On the plot 13 we can see the distribution of the age of people who took part in the study. We consider a group of adults with mean of age equal around 47 years according to the Table 8. Half of the group is in age between 34 and 60. The most frequent age is 40, the median is 45. Around 14% of the surveyed population are elderly people over 65 years of age. On the density plot 14 we can see that blue density has a peak above 60 years, which indicate that older people have diabetes more frequently than the younger in our population. It is not surprisingly, because research shows that older people are especially at risk of diabetes. This is, among others, related to increasing insulin resistance and impaired pancreatic islet function with ageing. [5]

2.5.7 Height — HHT

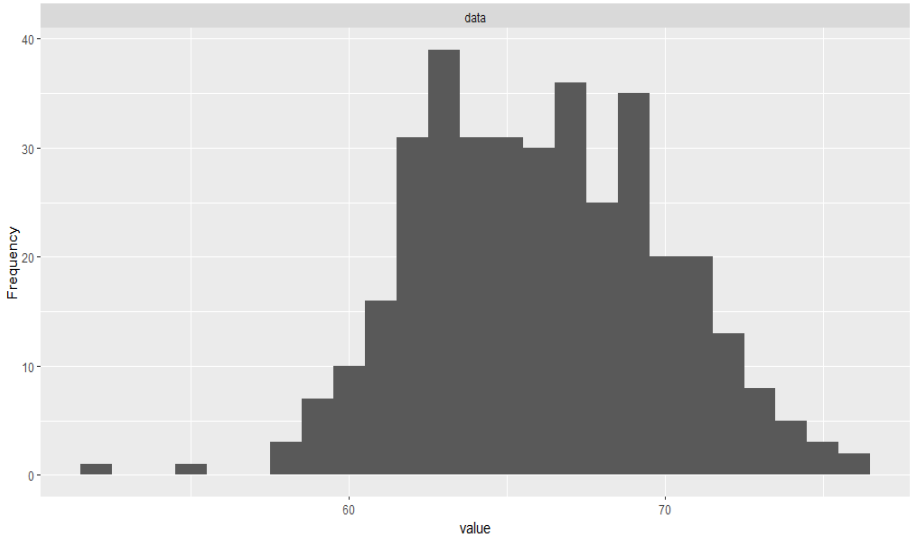

Figure 15: Histogram of feature "Height".

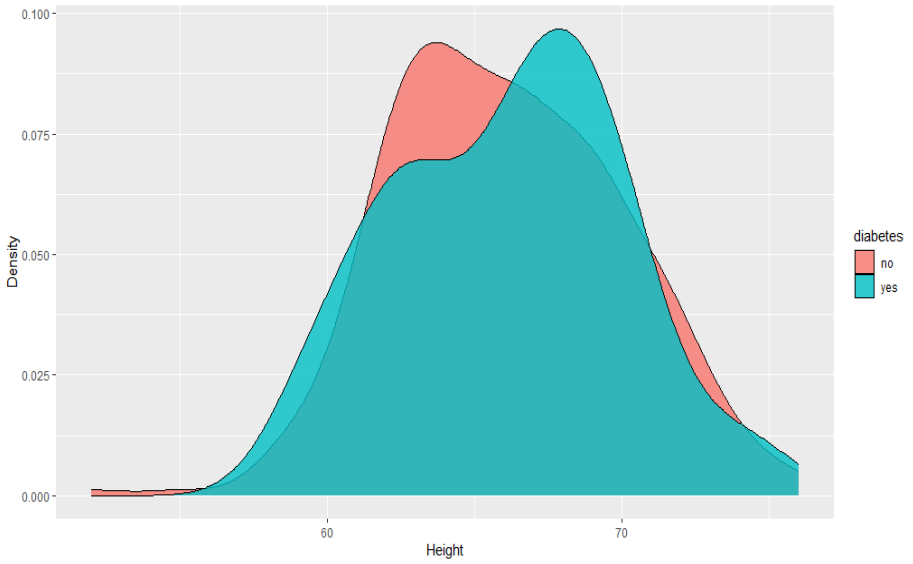

Figure 16: "Height" feature distribution by information whether a subject has diabetes or not.

Table 9: Short summary statistics for "Height" feature.

	HHT
min	52.00
Q1	63.00
median	66.00
mean	66.05
Q3	69.00
max	76.00
var	15.05
sd	3.88
IQR	6.00

"Height" attribute is a height measurement of subjects, given in inches. Plot 15 presents the histogram of the height. From the plot, we see that most of the result are between 60 and 70 inches. From the table 9 we read that mean and median are close to each other, which indicate distribution close to the symmetric one. There is only one outlier, minimum of the results — 52 inches. On the plot 16 we can't see any obvious dependencies between height and diabetes.

2.5.8 Weight — WHT

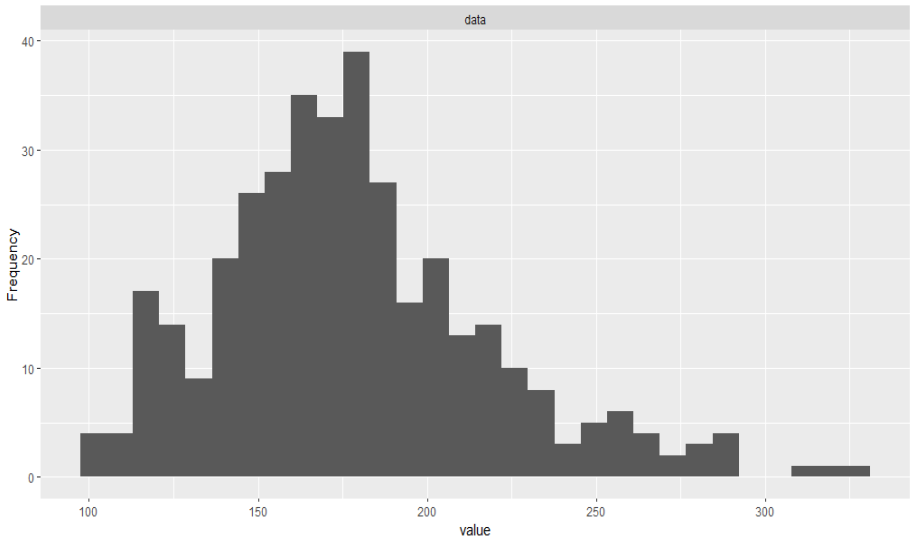

Figure 17: Histogram of feature "Weight".

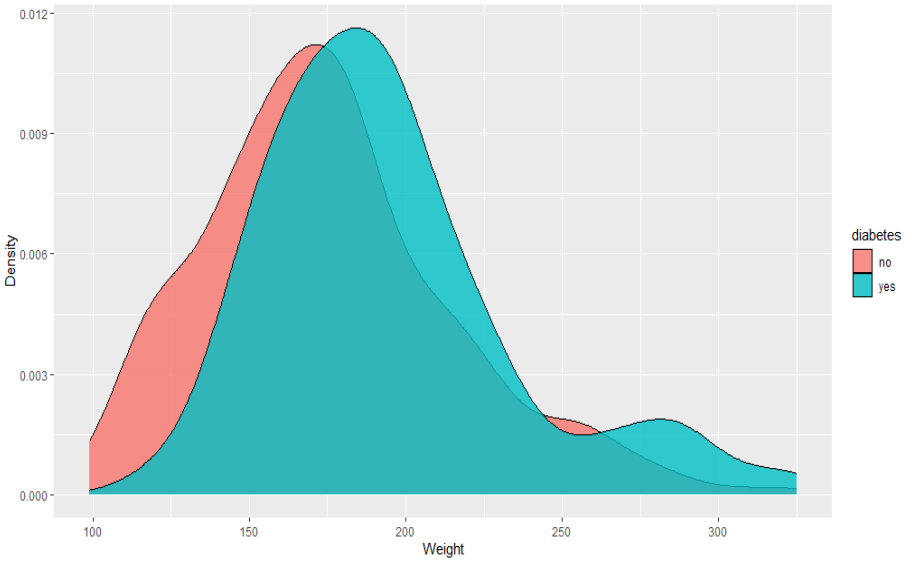

Figure 18: "Weight" feature distribution by information whether a subject has diabetes or not.

Table 10: Short summary statistics for "Weight" feature.

	WHT
min	99.00
Q1	151.00
median	174.00
mean	178.12
Q3	200.00
max	325.00
var	1650.45
sd	40.62
IQR	49.00

"Weight" attribute is a weight measurement of subjects, given in pounds. From the plot 18 and table 10, we can see that results vary from 99 to 325 pounds. It means we consider very different patients. Half of them weigh between 151 and 200 pounds. Standard deviation, which tells how widely the values are spread around the average weight is equal to 40.62. Hence, it can be said that dispersion of weight measurements is quite large. There are many outliers, which are higher than the mean. On the figure 16 we can observe that the blue density plot 18 is slightly moved to the right comparing with the red one. It means people with diabetes are usually a bit heavier than the healthy subjects.

2.5.9 Waist — W

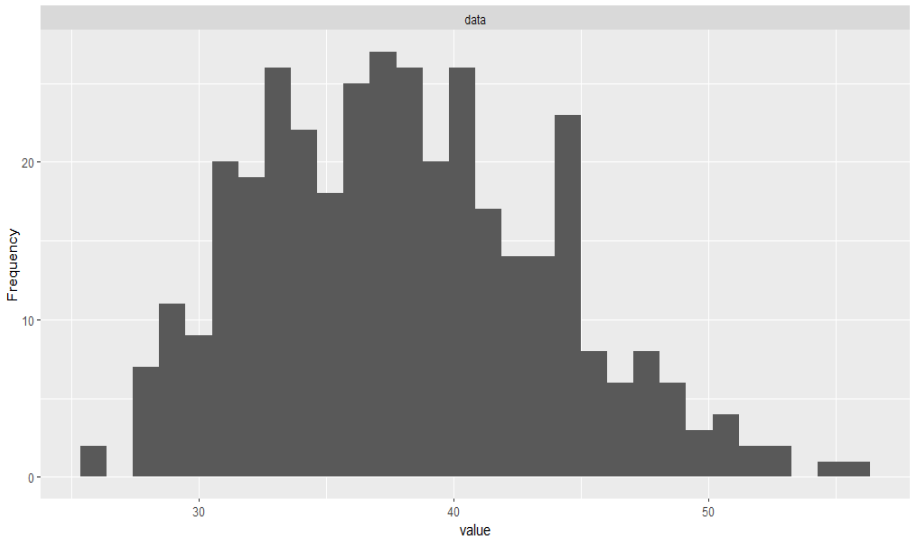

Figure 19: Histogram of feature "Waist".

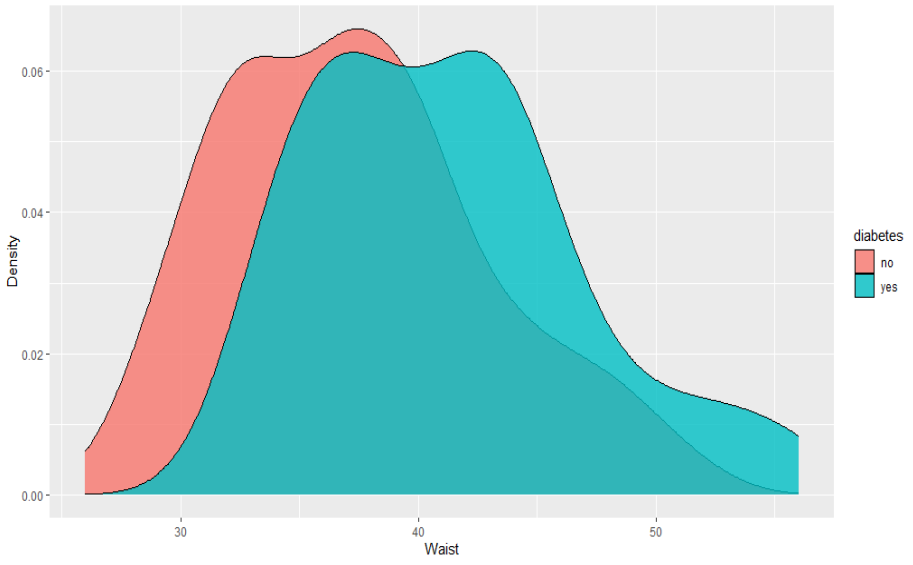

Figure 20: "Waist" feature distribution by information whether a subject has diabetes or not.

Table 11: Short summary statistics for "Waist" feature.

	W
min	26.00
Q1	33.00
median	37.00
mean	37.93
Q3	41.50
max	56.00
var	33.65
sd	5.80
IQR	8.50

Plot 19 and Table 11 describe waist measurement of subjects, given in inches. Its distribution is close to symmetric, cumulated between 33 and 41.5 inches. Density plots on the figure 20 look similar, the blue one is shifted to the right in relation to the red one. It means that people with diabetes tend to have higher waist circumference.

2.5.10 Hip — H

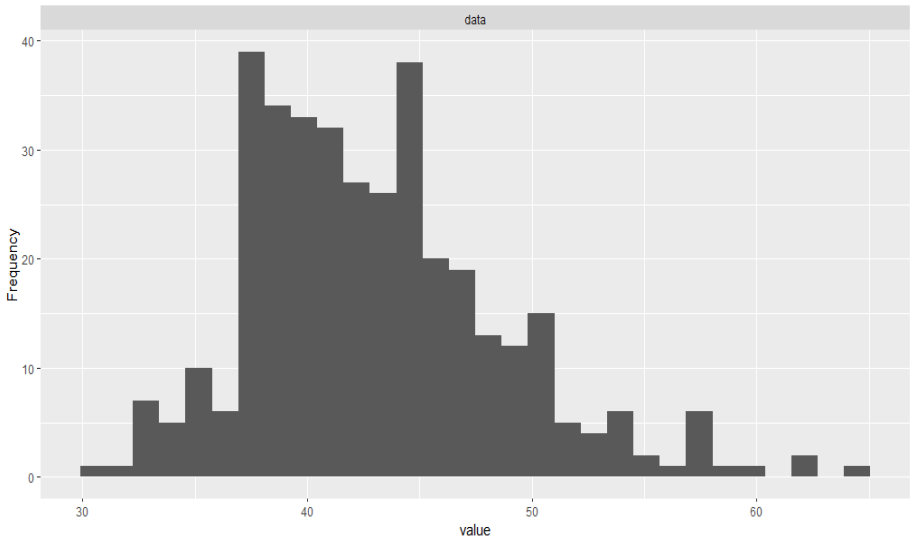

Figure 21: Histogram of feature "Hip".

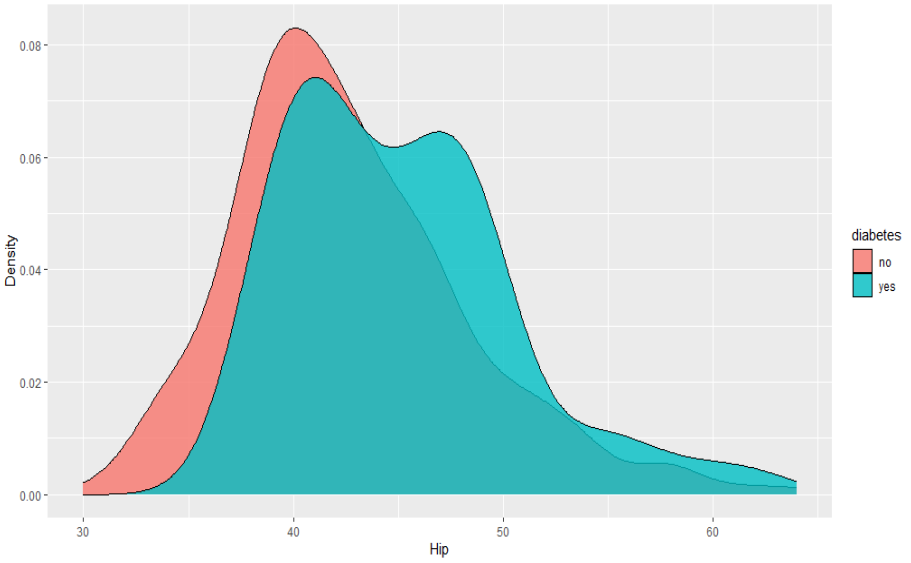

Figure 22: "Hip" feature distribution by information whether a subject has diabetes or not.

Table 12: Short summary statistics for "Hip" feature.

	H
min	30.00
Q1	39.00
median	42.00
mean	43.04
Q3	46.00
max	64.00
var	31.75
sd	5.63
IQR	7.00

Hip measurement of subjects, given in inches, is the last numeric attribute. Its histogram presented on the plot 19 suggest there are few people with a bit of flesh. From Table 11 we can read that maximum hip measurement is 64 inches. We observe outliers, but it is hard to presuppose subjects' obesity. We don't know their gender, height or body frame. From the density plots on the figure 20, we see that patients with the smallest hip measurements are healthy. Then, from value 38 inches, diabetes patient show up. It may indicate, that the higher hip measurement the biggest probability of having diabetes.

2.5.11 Location

The data contains information about two communities, defined by the geographic boundaries of two counties in United States: Buckingham County and Louisa County. These counties are demographically similar. They are not contiguous with each other, being separated by over 20 miles, and they are in different health districts. Buckingham County had a total population of 11.926 in 1990, of which 4.656 (39%) were African American. Louisa County had a total population in 1990 of 20.325, of which 5.233 (26%) were African American. The two counties were of similar size (about 500 square miles each), had similar African-American populations. [6]

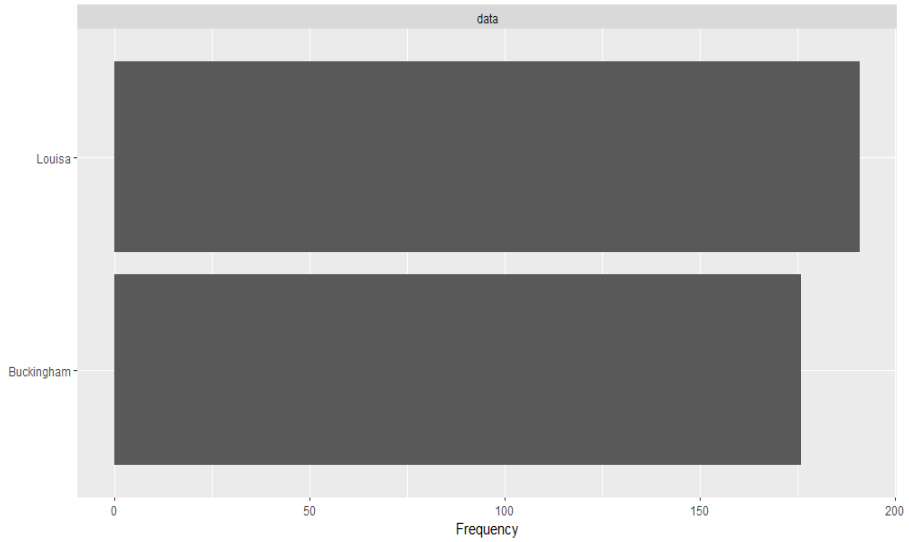

Figure 23: Bar plot of feature "Location".

Table 13: Contingency table for "Location" and information about diabetes.

diabetes	no	yes	Total
LOCATION			
Buckingham	147 (83.5%)	29 (16.5%)	176 (100.0%)
Louisa	164 (85.9%)	27 (14.1%)	191 (100.0%)
Total	311 (84.7%)	56 (15.3%)	367 (100.0%)

In the chart, we see that people are fairly evenly divided by place of residence. Slightly more people are from Louisa County. In order to see what is the distribution of people with diabetes along people from the particular county, we have prepared contingency table 13. We can read that the percentage of

diabetes cases in both counties is similar, 16.5% in Buckingham and 14.1% in Louisa. This may indicate that place of residence may not have a significant impact on the diabetes morbidity.

2.5.12 Gender

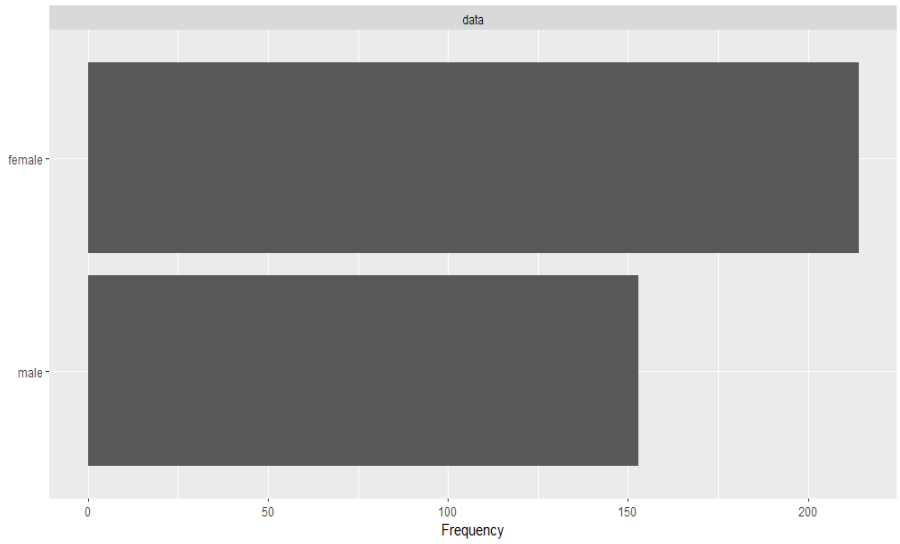

Figure 24: Bar plot of feature "Gender".

Table 14: Contingency table for "Gender" and information about diabetes.

diabetes	no	yes	Total
GENDER			
female	182 (85.0%)	32 (15.0%)	214 (100.0%)
male	129 (84.3%)	24 (15.7%)	153 (100.0%)
Total	311 (84.7%)	56 (15.3%)	367 (100.0%)

Our data contains two sex: Male and Female. As we can observe on the plot 24 there is more woman in our data, around 58% of all rows. Distribution over the diabetes morbidity is presented in the table 14. 15% of women 15.7% of men have diabetes, which indicate that there is no strong correlation between gender and the diabetes morbidity.

2.5.13 Frame

Feature "Frame" indicate body frame size. Bone structure varies in size from person to person. In order to determine a person's optimal weight, researchers

added frame size as a factor. Body frame or bone structure can impact ideal height to weight ratio and total body mass index. Human body frame sizes are categorized into three categories: small frame, medium frame and large frame. There are three basic methods for determining body frame size:

1. The wrist method
2. The elbow breadth method
3. Finger measurement method

The first one is the most standard one. It is to measure the circumference of the wrist using a tape measure. Then, depending on the gender and height of a person, there are different range for each frame. [7]

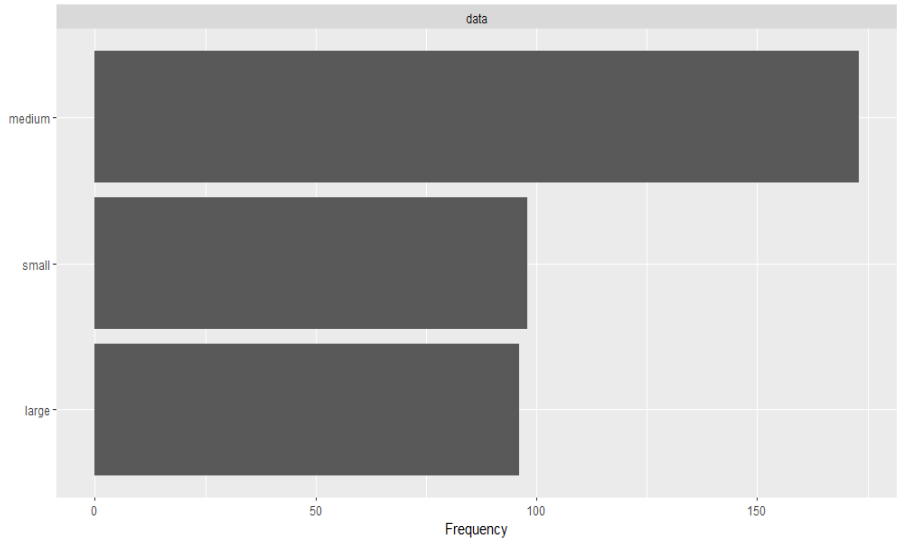

Figure 25: Bar plot of feature "Frame".

Table 15: Contingency table for "Frame" and information about diabetes.

diabetes	no	yes	Total
FRAME			
large	74 (77.1%)	22 (22.9%)	96 (100.0%)
medium	148 (85.5%)	25 (14.5%)	173 (100.0%)
small	89 (90.8%)	9 (9.2%)	98 (100.0%)
Total	311 (84.7%)	56 (15.3%)	367 (100.0%)

From the bar plot of "Frame" feature, we can observe that the most numerous frame is "medium", which is almost twice as big as the other two. Levels

"small" and "large" have similar count, 98 and 96 respectively. When looking at the table 15 we can clearly see that the percentage of diabetes among a given group increases with increasing body size. Almost 23% of all people with "large" body size have diabetes, while it is only 9.2% among "small" frame.

2.6 Correlation

The correlations between GHB and the other variables can be see from figure 26. We can see a strong connection between GHB and CHOL, SGLU, AGE and weight, the higher value these variables have the higher the GHB is going to be. The biggest dependency can be seen between GHB and SGLU. It's quite intuitive, since we know that as a person's blood sugar becomes higher (bigger SGLU), more of the person's hemoglobin becomes glycosylated (bigger GHB). The other variables don't seem to have such a strong correlation to GHB.

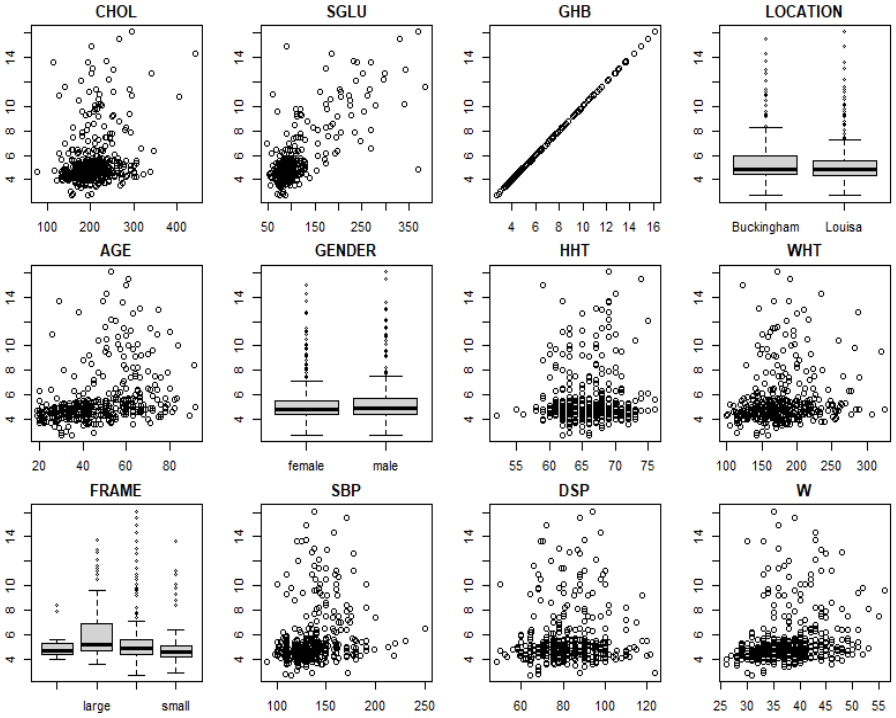

Figure 26: Bar plot of feature "Frame".

2.7 Inovation

To the data set we added two more variables, to see if they would influence the final model. These two variables are BMI (Body Mass Index) and Hip-to-Waist Ratio.

BMI is a measure of body fat based on an individual's weight and height. It is widely used as a screening tool to categorize individuals into different weight status categories. The formula to calculate BMI is:

$$\text{BMI} = \frac{\text{weight (in kilograms)}}{\text{height (in meters)}^2}.$$

The resulting BMI value can be interpreted using standard BMI categories, such as underweight, normal weight, overweight, and obese.

Hip-to-Waist Ratio is a measure of the distribution of fat in the body, particularly around the hips and waist. It is used as an indicator of abdominal obesity and is associated with an increased risk of certain health conditions, such as cardiovascular disease and type 2 diabetes. The formula to calculate is:

$$\text{Hip-to-Waist Ratio} = \frac{\text{hip circumference}}{\text{waist circumference}}.$$

Generally, a lower ratio indicates a healthier distribution of body fat, with less fat accumulating around the waist.

3 Methodology

3.1 Generalized linear models (GML)

Generalized Linear Models (GLM), introduced by Nelder and Wedderburn, synthesize the normal linear model that has a linear regression structure and have in common that the response variable belongs to the exponential distribution family. We say that the random variable X has distribution belonging to the (dispersion) exponential family if its p.d.f. or p.m.f. can be written in the form

$$f(x \mid \theta, \phi) = \exp\left\{\frac{x\theta - b(\theta)}{a(\phi)} + c(x, \phi)\right\}, \tag{1}$$

where θ and ϕ are scalar parameters, $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are known real functions. [8]

3.2 Gamma distribution

The gamma distribution is a continuous probability distribution that is often used to model positive continuous data with a skewed distribution, such as durations, waiting times, or financial data. It is characterized by two parameters: shape (a) and rate (b). The shape parameter determines the shape of the distribution, while the rate parameter determines the rate at which the distribution decays.

The probability density function (PDF) of the gamma distribution can be expressed using the shape parameter and the rate parameter as:

$$f(x; a, b) = (b^a/\Gamma(a)) \cdot x^{(a-1)} \cdot e^{bx}. \tag{2}$$

3.2.1 Gamma distribution belong to the (dispersion) exponential family

In this section we prove that gamma distribution belong to the (dispersion) exponential family, by representing its density function in the form of equation 1:

$$f(x;a,b) = \frac{b^a}{\Gamma(a)} \cdot x^{(a-1)} \cdot e^{-bx} =$$

$$= \exp\left\{-bx + \log(b^a) + \log(x^{a-1}) - \log(\Gamma(a))\right\} =$$

$$= \exp\left\{-bx + a\log(b) + (a-1)\log(x) - \log(\Gamma(a))\right\} =$$

$$= \exp\left\{\frac{-\frac{b}{a}x + \log(b)}{\frac{1}{a}} + (a-1)\log(x) - \log(\Gamma(a))\right\}.$$

Hence we have

$$\theta = -\frac{b}{a},$$

$$\phi = \frac{1}{a},$$

and

$$b = -\theta a = \frac{-\theta}{\phi}.$$

Then

$$\log(b) = \log(-\theta) - \log(\phi),$$

and the density function has a form

$$f(x;a,b) = \exp\left\{\frac{\theta x - (-\log(-\theta))}{\phi} - \frac{\log(\phi)}{\phi} + \left(\frac{1}{\phi} - 1\right)\log(x) - \log\left(\Gamma\left(\frac{1}{\phi}\right)\right)\right\}$$

It shows that Gamma distribution belongs the (dispersion) exponential family with

$$a(\phi) = \phi = \frac{1}{a},$$

$$b(\theta) = -\log(-\theta) = -\log(b/a),$$

$$c(x,\phi) = \frac{-\log(\phi)}{\phi} + \left(\frac{1}{\phi} - 1\right)\log(x) - \log\left(\Gamma\left(\frac{1}{\phi}\right)\right) = a\log(a) + (a-1)\log(x) - \log(\Gamma(a)).$$

3.3 Gamma generalized regression model

The gamma generalized regression model is a statistical model used to analyze data that follows a gamma distribution. It is an extension of the generalized linear model (GLM) framework, which allows for modeling of response variables that have non-normal error distributions. In the gamma regression model, the mean of the gamma distribution is related to the covariates through a link function. The most commonly used link function is the logarithmic link, which takes the form:

$$\log(E(Y)) = \mathbf{X}\beta, \tag{3}$$

where $E(Y)$ is the expected value of the response variable Y, X is the matrix of covariates, β is the vector of regression coefficients, and log denotes the natural logarithm. The model assumes that the response variable has a gamma distribution with a mean equal to $e^{X\beta}$.

The gamma generalized regression model assumes that the response variable follows a gamma distribution with a mean and a dispersion parameter. The dispersion parameter captures the variability of the response variable around the mean. The model assumes that the logarithm of the mean is a linear combination of the covariates.

The model can be estimated using maximum likelihood estimation (MLE), where the parameters are estimated by maximizing the likelihood function given the observed data. The MLE estimates provide information about the relationships between the covariates and the response variable.

It is important to note that the gamma generalized regression model assumes that the response variable is strictly positive, as the gamma distribution is only defined for positive values. If the response variable includes zeros or negative values, alternative models may be more appropriate.

From figure 4 we can see that GHB seems to have a gamma distribution, as such using a gamma regression model to fit the data makes sense.

3.4 Akaike information criterion (AIC)

The Akaike information criterion (AIC) is a criterion for selection of regression models. It is based on the log-likelihood function plus a correction factor as penalty of the model complexity, whose statistic is given by

$$AIC = -2\log L(\hat{\boldsymbol{\theta}}|\mathcal{D}) + 2p,$$

where $L(\boldsymbol{\theta}|\mathcal{D})$ is the likelihood function, $\mathcal{D}$ is the dataset (n observations) and θ is the maximum likelihood estimator of the parameter θ of dimension p, under the current statistical model. A low value for AIC indicates a better fit. [8]

4 Model fitting and model selection

We will use R to fit the final model using the gamma family. When implementing a gamma GLM in R, the log link function is often employed to relate the mean value of the response variable to the linear predictor as in equation (3).

Note: In R when we fit the model the rate parameter in equation (2)*, the b is fixed to be 1, and only a is being estimated and the other betas.*

Our selection process to get the final model is as follows. We start with a full model, meaning that we assume a gamma family glm model with a log link function, where GHB is a (linear) function of all the other variables described in the exploratory data analysis chapter, including the two new variables. For these model, we estimate the parameters.

We then used the step() function in R for stepwise model selection. It iteratively adds or removes covariates from a model based on specified criteria, in our case the Akaike Information Criterion (AIC), to find the best-fitting model, aiming to find the model that minimizes the AIC value. In our case, the step() function is applied to the full model, which is a gamma regression model fitted using the glm() function.

The stepwise selection process involves evaluating different models by adding or removing one covariate at a time and comparing the resulting models based on the AIC criterion. The step() function searches through the potential variable combinations to find the model that minimizes the selected criterion. The AIC of the full model was 1181.9. The reduction process of the model can be found in the appendix B.

The reduced model that we got using this is GHB as a function of CHOL, SGLU, AGE and W. The AIC of the reduced model was 1170.1. After getting the model, we need to evaluate it further to see if we need to add some variables. In the table 17 there is a quick overview of the models coefficients. The same table full the full model can be found in the appendix A.

Table 16: Overview of the reduced models coefficients.

Variable	Estimate	Std. Error	t value	p-value	2.5% CI	97.5% CI
(Intercept)	0.783	0.094	8.302	$1.92 \cdot 10^{-15}$	0.599	0.968
CHOL	0.001	0.000	2.703	0.00719	0.000	0.001
SGLU	0.004	0.000	15.696	$< 2 \cdot 10^{-16}$	0.003	0.004
AGE	0.003	0.001	4.106	$4.95 \cdot 10^{-5}$	0.002	0.005
W	0.005	0.002	2.296	0.02225	0.001	0.009

1. Intercept: The estimated intercept is 0.7833443. The p-value for the intercept is $1.92 \cdot 10^{-15}$.

2. CHOL: For a one-unit increase in CHOL, the expected value of GHB is expected to increase by a factor of $\exp(0.0007656) = 1.000766$. The p-value for CHOL is 0.00719, indicating that it is statistically significant at a significance level of 0.01.

3. SGLU: For a one-unit increase in SGLU, the expected value of GHB is expected to increase by a factor of $\exp(0.0037727) = 1.00378$. The p-value for SGLU is very small ($< 2 \cdot 10^{-16}$), indicating high statistical significance.

4. AGE: For a one-unit increase in AGE, the expected value of GHB is expected to increase by a factor of $\exp(0.0032112) = 1.003217$. The p-value for AGE is very small ($4.95 \cdot 10^{-5}$), indicating high statistical significance.

5. W: For a one-unit increase in W, the expected value of GHB is expected to increase by a factor of $\exp(0.0049923) = 1.004999$. The p-value for W is 0.02225, indicating that it is statistically significant at a significance level of 0.05.

To evaluate the model, it is good to do hypothesis testing. Firstly let's test the hypothesis $H_0 : \beta_0 = 0$ against $\beta_0 \neq 0$. Let's note that:

$$T_0 = \frac{\hat{\beta}_0 - 0}{\sqrt{\text{Var}(\hat{\beta}_0)}} \sim_{H_0} t_{(372)}.$$

The empirical test statistic is:

$$t_0 = \frac{0.7833443 - 0}{0.0943607} = 8.302.$$

From this we can calculate the p-value as:

$$P(|T_0| > |t_0| \mid H_0) = 2 \cdot (1 - F_{t(372)}(t_0)) \approx 1.9 \cdot 10^{-15}.$$

The conclusion is that we reject H_0 for the levels of significance $\alpha > 1.9 \cdot 10^{-15}$, respectively, so there is strong evidence for the rejection of rejecting the hypotheses, given the usual α of 1%, 5%.

From the table 17, we can see that all the p-values are below 0.05, so we can reject the hypothesis that $\beta_i = 0$ for $i = 0, 1, 2, 3, 4$ given the usual α of 5%.

We also took a look at the 95% confidence intervals for each of the coefficients. These also supports the fact that none of the betas should be 0.

We tried adding to the model FRAME covariates, as in the full model the p-value of the FRAMEmedium was less than 0.1. However, when adding this to the reduced model the AIC increased and the values of all of the FRAME covariates was more then 0.1, so we decided to remove them from the final model.

From this two points, we can conclude that the model cannot be reduced further down. Taking all this into account, we can now conclude that the reduced model is the final model.

5 Model Interpretation and model evaluation

5.1 Residuals vs fitted

Firstly we can analyse the residuals vs the fitted values. This can be seen in figure 27. We can see that most of the residuals, are randomly scattered around zero. This suggests that the model captures the linear relationship

between the covariates and the response variable effectively. It indicates that the model's predictions are, on average, unbiased, with no systematic patterns left unexplained. This is a positive indication of a good fit to the data.

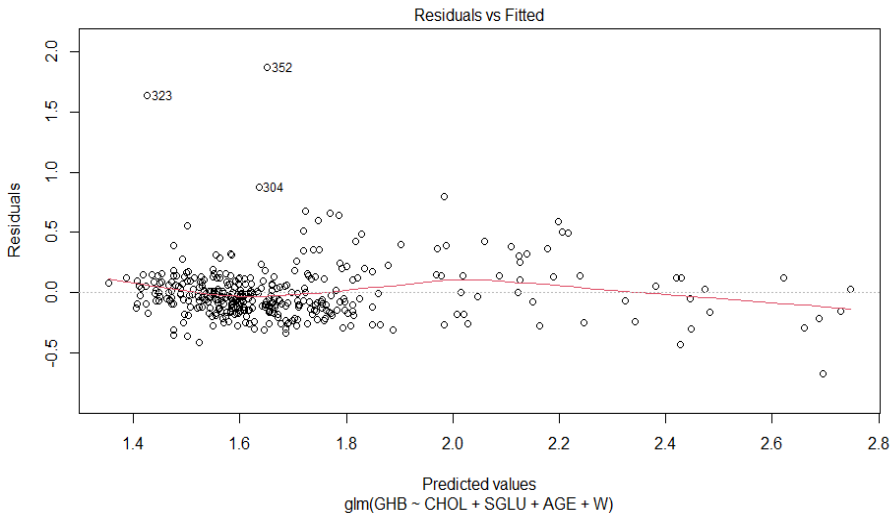

Figure 27: Residuals vs fitted values.

5.2 Normal Q-Q plot of residuals

Secondly, we can analyse the normal Q-Q (Quantile-Quantile) plot of residuals for the selected model. The plot is displaying the residuals of the model. The graph is displayed in figure 28. Since the points closely follow the diagonal line, it suggests that the residuals are normally distributed. This is a positive indication of a good fit to the data, since the model assumes the normal distribution of the residuals.

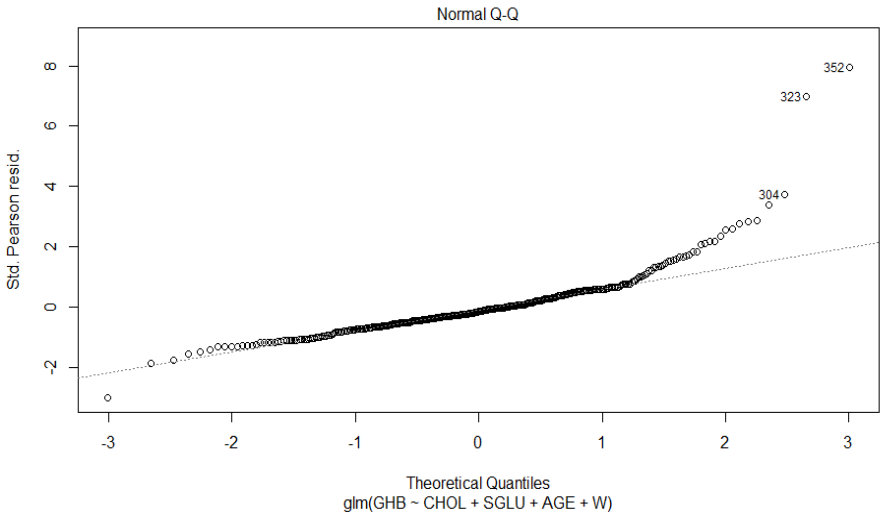

Figure 28: Normal Q-Q plot.

5.3 Scale location

Thirdly, let's look at the Scale-Location plot, that is seen in figure 29. This plot is a diagnostic tool used to assess the assumption of constant variance, or homoscedasticity, in a regression model. It displays the square root of the absolute standardized residuals on the y-axis and the fitted values or predicted values from the model on the x-axis. Each data point represents an observation from the dataset.

From the figure we can see that the points are randomly scattered around a horizontal line with no discernible pattern, it suggests that the spread of residuals is approximately constant across the range of fitted values. This indicates that the assumption of constant variance is met, which is desirable.

The implications of homoscedasticity is important, as it implies the validity of statistical tests and confidence intervals.

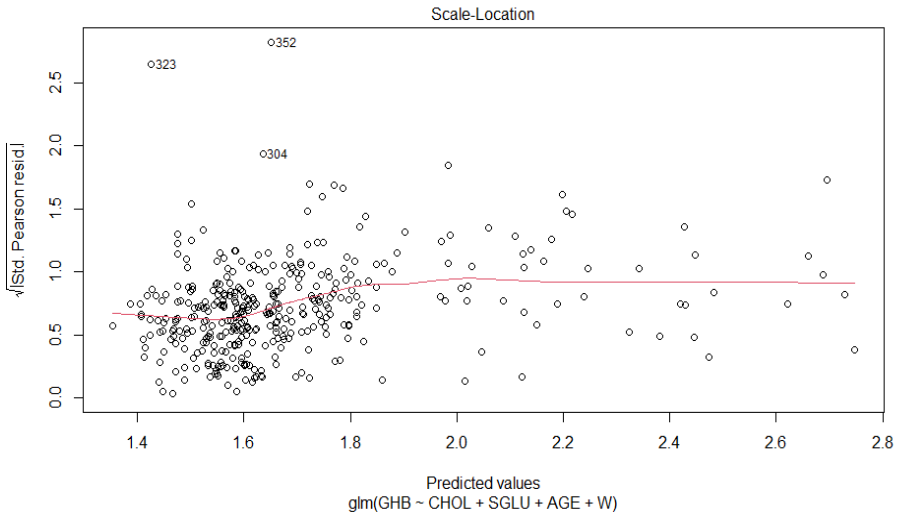

Figure 29: Scale location graph.

5.4 Cook's distance plot

Let's also take a look at the Cook's distance plot, generated in figure 30. The Cook's distance is a measure used to identify influential observations in a regression model. It quantifies the impact of each observation on the overall fit of the model by considering both the leverage (how extreme an observation's covariates are) and the influence (how much an observation affects the model's coefficients).

We took a look at the rows with high Cook's distance values (higher than 0.05). Assessing their impact involves considering their leverage, influence on the model coefficients, and potential effects on the overall fit of the model. After taking a closer look at this rows we did not notice any extreme exhibit extreme behavior, measurement errors, or other issues, so we decided to keep them in the final model.

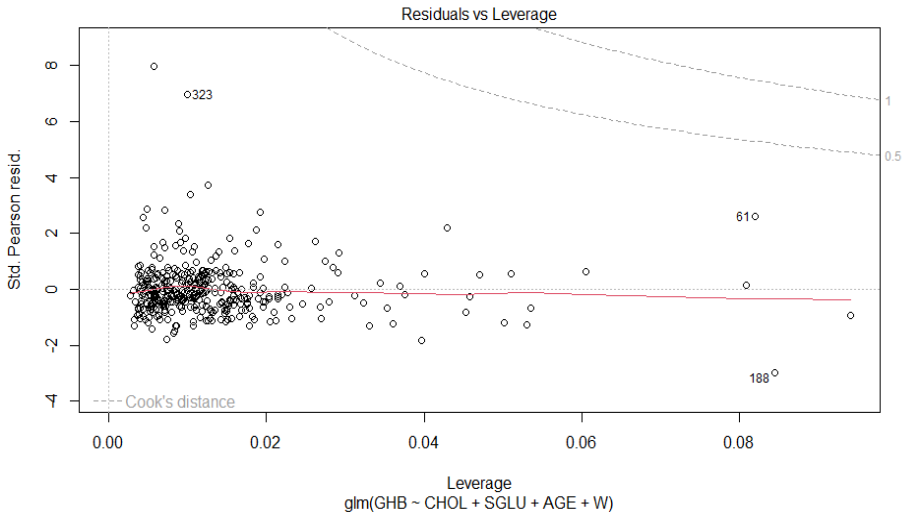

Figure 30: Cook's distance plot.

5.5 Interpretation

Our final model consists of four covariates, that is, total cholesterol, stabilized glucose, age and waist. From the coefficients we can infer that a higher value of any of this covariates means a higher value glycosolated hemoglobin. From the coefficients and the $p-$values we can see that all of these covariates play an important role to glycosolated hemoglobin. From this we can gather that the higher the CHOL, SGLU, AGE or W, the higher the risk of having diabetes. Meaning that if we would like to reduce the chance of diabetes lower levels of CHOL, SGLU and W would be needed.

For interpreting the model it is also important to see if the fitted model accurately describes the initial data. We can see from figure 31 that the distribution of the fitted model is similar to the initial data, indicating that the model is not a bad fit.

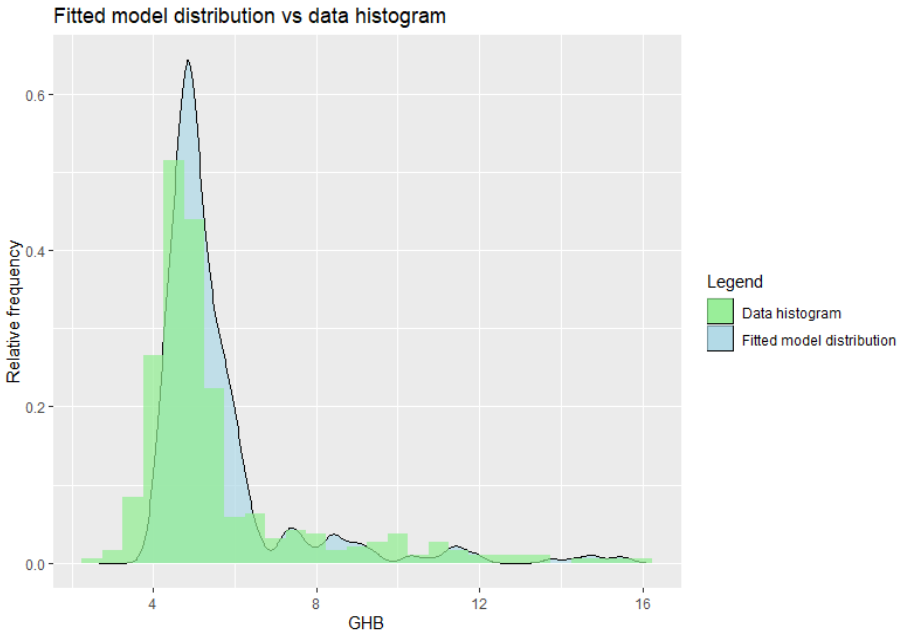

Figure 31: The distribution of the fitted model to the initial data.

6 Conclusion

The first part of the project was a general overview of the data and provide background information. Since our data have few features related to blood tests, we provide some information about reference ranges. We also prepared the data for further investigation, by removing unnecessary column "ID" and missing values.
In the second part of the project, we explained methodology, which we followed in this work. We introduced the concept of generalized linear model, gamma distribution and proved that gamma distribution can be used in GLM. Then explained gamma regression model and Akaike information criterion, which is used in model selection.
The fourth section was dedicated to fitting the model and choosing the best one in terms of AIC value. It turned out that the most significant features which influent "Glycosylated Hemoglobin" are "Total Cholesterol", "Stabilized Glucose", "Age" and "Waist". These attributes are significant with significance level $\alpha = 0.05$. It means that they influent level of glycosylated hemoglobin, which high value, cause diabetes.
In the fifth section, we evaluate a selected model using graphical tools: we plotted residuals vs fitted values, normal QQ plot, scale location and Cook's distance graph. We interpreted the result. It led us to the conclusion that the selected model reflects real data quite good.

A Full model table

Table 17: Overview of the full model's coefficients.

Variable	Estimate	Std. Error	t value	p-value	2.5% CI	97.5% CI
(Intercept)	0.369	0.936	0.394	0.693	-1.461	2.188
CHOL	0.001	0.000	2.683	0.008	0.000	0.001
SGLU	0.004	0.000	15.688	$< 2 \cdot 10^{-16}$	0.003	0.004
LOCATIONLouisa	-0.040	0.025	-1.591	0.112	-0.090	0.009
AGE	0.003	0.001	2.962	0.003	0.001	0.005
GENDERmale	0.014	0.040	0.359	0.720	-0.064	0.093
HHT	0.003	0.014	0.208	0.836	-0.025	0.031
WHT	-0.001	0.003	-0.464	0.643	-0.006	0.004
FRAMElarge	0.102	0.079	1.289	0.198	-0.056	0.254
FRAMEmedium	0.144	0.078	1.855	0.064	-0.011	0.292
FRAMEsmall	0.125	0.081	1.551	0.122	-0.036	0.281
SBP	0.000	0.001	0.607	0.544	-0.001	0.002
DSP	-0.001	0.001	-0.703	0.483	-0.003	0.002
W	0.008	0.005	1.570	0.117	-0.002	0.018
H	0.004	0.006	0.743	0.458	-0.007	0.016
BMI	0.002	0.016	0.100	0.920	-0.029	0.033

B R model code output

The output of the code used to generate the final model (file 'model.R') can be found below:

```
> source("~/model.R")
Start:  AIC=1181.93
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + HHT + WHT + FRAME +
    SBP + DSP + W + H + BMI + H

           Df Deviance    AIC
- BMI       1   16.452 1179.9
- HHT       1   16.454 1180.0
- GENDER    1   16.459 1180.1
- WHT       1   16.464 1180.2
- FRAME     3   16.687 1180.2
- SBP       1   16.472 1180.3
- DSP       1   16.479 1180.4
- H         1   16.482 1180.5
<none>          16.452 1181.9
- W         1   16.585 1182.4
- LOCATION  1   16.590 1182.5
```

```
- CHOL      1   16.852 1187.3
- AGE       1   16.928 1188.7
- SGLU      1   31.238 1451.0

Step:  AIC=1179.94
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + HHT + WHT + FRAME +
    SBP + DSP + W + H

           Df Deviance    AIC
- HHT       1   16.459 1178.1
- GENDER    1   16.460 1178.1
- FRAME     3   16.689 1178.3
- SBP       1   16.472 1178.3
- DSP       1   16.479 1178.4
- H         1   16.488 1178.6
- WHT       1   16.523 1179.2
<none>          16.452 1179.9
- W         1   16.585 1180.4
- LOCATION  1   16.592 1180.5
- CHOL      1   16.852 1185.3
- AGE       1   16.933 1186.8
- SGLU      1   31.296 1451.0

Step:  AIC=1178.09
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + WHT + FRAME + SBP +
    DSP + W + H

           Df Deviance    AIC
- SBP       1   16.479 1176.5
- GENDER    1   16.485 1176.6
- DSP       1   16.486 1176.6
- H         1   16.490 1176.7
- FRAME     3   16.708 1176.7
- WHT       1   16.523 1177.3
<none>          16.459 1178.1
- W         1   16.587 1178.5
- LOCATION  1   16.613 1178.9
- CHOL      1   16.855 1183.4
- AGE       1   16.935 1184.9
- SGLU      1   31.461 1452.9

Step:  AIC=1176.56
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + WHT + FRAME + DSP +
    W + H

           Df Deviance    AIC
```

```
- DSP       1   16.488 1174.7
- GENDER    1   16.507 1175.1
- FRAME     3   16.724 1175.1
- H         1   16.516 1175.2
- WHT       1   16.552 1175.9
<none>          16.479 1176.6
- W         1   16.610 1177.0
- LOCATION  1   16.637 1177.5
- CHOL      1   16.875 1181.9
- AGE       1   17.161 1187.2
- SGLU      1   31.572 1453.7

Step:  AIC=1174.77
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + WHT + FRAME + W +
    H

           Df Deviance    AIC
- FRAME     3   16.729 1173.2
- GENDER    1   16.514 1173.2
- H         1   16.525 1173.5
- WHT       1   16.564 1174.2
<none>          16.488 1174.8
- W         1   16.618 1175.2
- LOCATION  1   16.649 1175.8
- CHOL      1   16.875 1179.9
- AGE       1   17.165 1185.3
- SGLU      1   31.681 1454.1

Step:  AIC=1174.28
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + WHT + W + H

           Df Deviance    AIC
- GENDER    1   16.740 1172.5
- H         1   16.755 1172.7
- WHT       1   16.793 1173.4
- W         1   16.838 1174.3
<none>          16.729 1174.3
- LOCATION  1   16.843 1174.3
- CHOL      1   17.176 1180.4
- AGE       1   17.342 1183.4
- SGLU      1   31.859 1447.0

Step:  AIC=1172.51
GHB ~ CHOL + SGLU + LOCATION + AGE + WHT + W + H

           Df Deviance    AIC
```

```
- H         1   16.755 1170.8
- WHT       1   16.794 1171.5
- LOCATION  1   16.848 1172.5
- W         1   16.848 1172.5
<none>          16.740 1172.5
- CHOL      1   17.180 1178.5
- AGE       1   17.399 1182.5
- SGLU      1   31.873 1445.2

Step:  AIC=1170.86
GHB ~ CHOL + SGLU + LOCATION + AGE + WHT + W

           Df Deviance    AIC
- WHT       1   16.796 1169.6
- LOCATION  1   16.853 1170.6
<none>          16.755 1170.9
- W         1   16.943 1172.3
- CHOL      1   17.195 1176.8
- AGE       1   17.402 1180.6
- SGLU      1   31.936 1444.4

Step:  AIC=1169.79
GHB ~ CHOL + SGLU + LOCATION + AGE + W

           Df Deviance    AIC
- LOCATION  1   16.898 1169.6
<none>          16.796 1169.8
- W         1   17.060 1172.5
- CHOL      1   17.237 1175.7
- AGE       1   17.716 1184.4
- SGLU      1   31.994 1441.8

Step:  AIC=1170.09
GHB ~ CHOL + SGLU + AGE + W

Waiting for profiling to be done...
Waiting for profiling to be done...
    CHOL SGLU  GHB   LOCATION AGE GENDER HHT WHT  FRAME SBP DSP  W  H      BMI HIP_WAIST_RAT
64   223   75 4.25 Buckingham  22 female  62 137 medium 120  70 28 35 25.05489        1.2500
199  207   77 4.82 Buckingham  68   male  55 130  small 199 115 29 33 30.21157        1.1379
> source("~/FMF mag/2.semester/Biostatistics/Project/model.R")
Start:  AIC=1183.51
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + HHT + WHT + FRAME +
    SBP + DSP + W + H + BMI + HIP_WAIST_RATIO

                Df Deviance    AIC
```

```
- BMI              1   16.434 1181.5
- HHT              1   16.434 1181.5
- W                1   16.437 1181.6
- GENDER           1   16.439 1181.6
- WHT              1   16.441 1181.6
- FRAME            3   16.669 1181.8
- HIP_WAIST_RATIO  1   16.452 1181.8
- SBP              1   16.456 1181.9
- DSP              1   16.462 1182.0
- H                1   16.465 1182.1
<none>                 16.434 1183.5
- LOCATION         1   16.579 1184.2
- CHOL             1   16.837 1188.9
- AGE              1   16.896 1190.0
- SGLU             1   31.225 1452.8

Step:  AIC=1181.51
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + HHT + WHT + FRAME +
    SBP + DSP + W + H + HIP_WAIST_RATIO

                  Df Deviance    AIC
- W                1   16.437 1179.6
- GENDER           1   16.439 1179.6
- HHT              1   16.439 1179.6
- FRAME            3   16.669 1179.8
- HIP_WAIST_RATIO  1   16.452 1179.8
- SBP              1   16.456 1179.9
- DSP              1   16.462 1180.0
- H                1   16.467 1180.1
- WHT              1   16.501 1180.7
<none>                 16.434 1181.5
- LOCATION         1   16.581 1182.2
- CHOL             1   16.837 1186.9
- AGE              1   16.898 1188.0
- SGLU             1   31.291 1452.8

Step:  AIC=1179.58
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + HHT + WHT + FRAME +
    SBP + DSP + H + HIP_WAIST_RATIO

                  Df Deviance    AIC
- HHT              1   16.443 1177.7
- GENDER           1   16.443 1177.7
- FRAME            3   16.673 1177.9
- SBP              1   16.458 1178.0
- DSP              1   16.465 1178.1
```

```
- WHT              1   16.510 1178.9
<none>                 16.437 1179.6
- LOCATION         1   16.581 1180.2
- HIP_WAIST_RATIO  1   16.585 1180.3
- H                1   16.642 1181.4
- CHOL             1   16.839 1185.0
- AGE              1   16.902 1186.2
- SGLU             1   31.293 1451.4

Step:  AIC=1177.72
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + WHT + FRAME + SBP +
    DSP + H + HIP_WAIST_RATIO

                  Df Deviance    AIC
- SBP              1   16.465 1176.1
- GENDER           1   16.466 1176.2
- DSP              1   16.471 1176.2
- FRAME            3   16.692 1176.3
- WHT              1   16.510 1177.0
<none>                 16.443 1177.7
- HIP_WAIST_RATIO  1   16.587 1178.4
- LOCATION         1   16.601 1178.7
- H                1   16.643 1179.4
- CHOL             1   16.841 1183.1
- AGE              1   16.904 1184.2
- SGLU             1   31.455 1453.3

Step:  AIC=1176.22
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + WHT + FRAME + DSP +
    H + HIP_WAIST_RATIO

                  Df Deviance    AIC
- DSP              1   16.474 1174.4
- GENDER           1   16.489 1174.7
- FRAME            3   16.709 1174.8
- WHT              1   16.540 1175.6
<none>                 16.465 1176.2
- HIP_WAIST_RATIO  1   16.610 1176.9
- LOCATION         1   16.627 1177.2
- H                1   16.683 1178.3
- CHOL             1   16.863 1181.6
- AGE              1   17.131 1186.6
- SGLU             1   31.570 1454.2

Step:  AIC=1174.44
GHB ~ CHOL + SGLU + LOCATION + AGE + GENDER + WHT + FRAME + H +
```

```
    HIP_WAIST_RATIO

                  Df Deviance    AIC
- GENDER           1   16.497 1172.9
- FRAME            3   16.715 1172.9
- WHT              1   16.552 1173.9
<none>                 16.474 1174.4
- HIP_WAIST_RATIO  1   16.618 1175.1
- LOCATION         1   16.640 1175.5
- H                1   16.690 1176.5
- CHOL             1   16.863 1179.7
- AGE              1   17.136 1184.7
- SGLU             1   31.679 1454.6

Step:  AIC=1172.96
GHB ~ CHOL + SGLU + LOCATION + AGE + WHT + FRAME + H + HIP_WAIST_RATIO

                  Df Deviance    AIC
- FRAME            3   16.723 1171.2
- WHT              1   16.552 1172.0
<none>                 16.497 1173.0
- HIP_WAIST_RATIO  1   16.639 1173.6
- LOCATION         1   16.651 1173.8
- H                1   16.714 1175.0
- CHOL             1   16.877 1178.0
- AGE              1   17.198 1183.9
- SGLU             1   31.699 1452.2

Step:  AIC=1172.14
GHB ~ CHOL + SGLU + LOCATION + AGE + WHT + H + HIP_WAIST_RATIO

                  Df Deviance    AIC
- WHT              1   16.783 1171.2
<none>                 16.723 1172.1
- LOCATION         1   16.835 1172.2
- HIP_WAIST_RATIO  1   16.848 1172.4
- H                1   16.920 1173.7
- CHOL             1   17.165 1178.2
- AGE              1   17.362 1181.8
- SGLU             1   31.871 1445.7

Step:  AIC=1171.5
GHB ~ CHOL + SGLU + LOCATION + AGE + H + HIP_WAIST_RATIO

                  Df Deviance    AIC
- HIP_WAIST_RATIO  1   16.852 1170.7
```

```
- LOCATION         1   16.889 1171.4
<none>                 16.783 1171.5
- H                1   17.004 1173.5
- CHOL             1   17.224 1177.5
- AGE              1   17.656 1185.2
- SGLU             1   31.974 1443.0

Step:  AIC=1171.05
GHB ~ CHOL + SGLU + LOCATION + AGE + H

           Df Deviance    AIC
<none>          16.852 1171.0
- LOCATION  1   16.990 1171.5
- H         1   17.060 1172.8
- CHOL      1   17.307 1177.3
- AGE       1   17.918 1188.3
- SGLU      1   32.645 1453.9
    CHOL SGLU  GHB   LOCATION AGE GENDER HHT WHT  FRAME SBP DSP  W  H      BMI
    HIP_WAIST_RATIO
64   223   75 4.25 Buckingham  22 female  62 137 medium 120  70 28 35 25.05489
1.250000
199  207   77 4.82 Buckingham  68   male  55 130  small 199 115 29 33 30.21157
1.137931
```

References

[1] *Cholesterol Numbers and What They Mean* [viewed 28. 05. 2023], can be found at https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean

[2] *Glycosylated hemoglobin* [viewed 28. 05. 2023], can be found at https://www.humanitas.net/treatments/glycosylated-hemoglobin/

[3] *Which blood pressure number is important?* [viewed 28. 05. 2023], can be found at https://www.health.harvard.edu/staying-healthy/which-blood-pressure-number-is-important

[4] The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (2003 Guideline)

[5] M. Sue Kirkman, Vanessa Jones Briscoe, Nathaniel Clark, Hermes Florez, Linda B. Haas, Jeffrey B. Halter, Elbert S. Huang, Mary T. Korytkowski, Medha N. Munshi, Peggy Soule Odegard, Richard E. Pratley, Carrie S. Swift; Diabetes in Older Adults. Diabetes Care 1 December 2012; 35 (12): 2650–2664

[6] Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart HL (1997) A trial of church-based smoking cessation interventions for rural African Americans. Preventive Medicine 26:92-101.

[7] *Body Frame Size Measuring Tables* [viewed 28. 05. 2023], can be found at https://www.disabled-world.com/calculators-charts/body-frame.php

[8] Silva, G.L. (2023). Lecture Notes of Biostatistics. Lisbon, Instituto Superior Tecnico

[9] *Blood Glucose (Sugar) Test* [viewed 28. 05. 2023], can be found at https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test

[10] *What is Diabetes?* [viewed 28. 05. 2023], can be found at https://www.cdc.gov/diabetes/basics/diabetes.html