DEPARTMENT OF STATISTICS
LOYOLA COLLEGE (AUTONOMOUS), NUNGAMBAKKAM
CHENNAI – 600034

FORECASTING DIABETES USING LOGISTIC REGRESSION

UST 6708

SUBMITTED TO
DEPARTMENT OF STATISTICS
IN PARTIAL FULFILLMENT OF THE DEGREE OF BACHELOR
OF SCIENCE IN STATISTCS

SUBMITTED BY
JOHN VARSHAN J (20-UST-033)

UNDER THE GUIDANCE OF

PROF. DR. SELVA ARUL PANDIYAN

DEPARTMENT OF STATISTICS,

LOYOLA COLLEGE, CHENNAI.

CERTIFICATE

THIS IS TO CERTIFY THAT THE BONAFIDE RECORD WORK ENTITLED **"FORECASTING DIABETES USING LOGISTIC REGRESSION"** SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF THE **DEGREE OF BACHELOR OF SCIENCE IN STATISTICS**, IS A BONAFIDE WORK DONE UNDER THE GUIDANCE OF **PROF. DR. SELVA ARUL PANDIYAN** SUBMITTED BY **JOHN VARSHAN J (20-UST-033)** IN THE **DEPARTMENT OF STATISTICS, LOYOLA COLLEGE (AUTONOMOUS), CHENNAI – 600034** DURING THE ACADEMIC YEAR **2020-2023.**

Internal Examiner **JOHN VARSHAN J**

DECLARATION

I HEREBY TO DECLARE THAT THE TITLED **"FORECASTING DIABETES USING LOGISTIC REGRESSION"** IS BASED ON THE SECONDARY DATA FROM THE ONLINE SOURCE 'KAGGLE' UNDER THE GUIDANCE OF **PROF. DR. SELVA ARUL PANDIYAN**, SUBMITTED IN PARTIAL FULFILMENT OF THE DEGREE OF BACHELOR SCIENCE IN STATISTICS. I FURTHER DECLARE, THAT THIS PROJECT WORK OR ANY OTHER PART OF IT'S NOT BEEN SUBMITTED BY ME ANYWHERE FOR THE AWARD OF ANY DEGREE, DIPLOMA OR OTHER SIMILAR TITLE BEFORE.

JOHN VARSHAN J

(20-UST-033)

ACKNOWLEDGEMENT

I express my sincere and deep filling thanks to our beloved principal who had been a great moral support to us and for providing an excellent environment at college. I am greatly indebted to record my respectful and sincere thanks to Head of the Department **DR. T. EDWIN PRABAKARAN,** for his valuable suggestion and encouragement during the period of this work. The meticulous guidance of my guide **PROF. Dr. SELVA ARUL PANDIYAN**, for his constant encouragement, inspiration and support at each and every stage of my project work which made the work get completed in time and for providing necessary guidance and encouragement for doing this project.

INDEX

FORECASTING DIABETES USING LOGISTIC REGRESSION

Objective

The goal is obvious and simple, predicting whether or not a patient has diabetes. It's a typical classification problem. But there is a thing should be considered. It need to put weight on prediction diabetics have diabetes. It's not a big problem if you predict normal person has diabetes. However it's a huge problem if you predict diabetic doesn't have diabetes. So, It'd better to prepare different criteria to estimate the model.

INTRODUCTION

Diabetes

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy. Sometimes your body doesn’t make enough—or any—insulin or doesn’t use insulin well. Glucose then stays in your blood and doesn’t reach your cells.

Over time, having too much glucose in your blood can cause health problems. Although diabetes has no cure, you can take steps to manage your diabetes and stay healthy. Sometimes people call diabetes “a touch of sugar” or “borderline diabetes.” These terms suggest that someone doesn’t really have diabetes or has a less serious case, but every case of diabetes is serious.

The most common types of diabetes are type 1, type 2, and gestational diabetes.

Type 1 diabetes

If you have type 1 diabetes, your body does not make insulin. Your immune system attacks and destroys the cells in your pancreas that make insulin. Type 1 diabetes is usually diagnosed in children and young adults, although it can appear at any age. People with type 1 diabetes need to take insulin every day to stay alive.

Type 2 diabetes

If you have type 2 diabetes, your body does not make or use insulin well. You can develop type 2 diabetes at any age, even during childhood.

However, this type of diabetes occurs most often in middle-aged and older people. Type 2 is the most common type of diabetes.

Gestational diabetes

Gestational diabetes develops in some women when they are pregnant. Most of the time, this type of diabetes goes away after the baby is born. However, if you've had gestational diabetes, you have a greater chance of developing type 2 diabetes later in life. Sometimes diabetes diagnosed during pregnancy is actually type 2 diabetes.

Prevention

- Losing weight and keeping it off. Weight control is an important part of diabetes prevention. You may be able to prevent or delay diabetes by losing 5 to 10% of your current weight. For example, if you weigh 200 pounds, your goal would be to lose between 10 to 20 pounds. And once you lose the weight, it is important that you don't gain it back.
- Following a healthy eating plan. It is important to reduce the amount of calories you eat and drink each day, so you can lose weight and keep it off. To do that, your diet should include smaller portions and less fat and sugar. You should also eat a variety of foods from each food group, including plenty of whole grains, fruits, and vegetables. It's also a good idea to limit red meat, and avoid processed meats.
- Get regular exercise. Exercise has many health benefits, including helping you to lose weight and lower your blood sugar levels. These both lower your risk of type 2 diabetes. Try to get at least 30 minutes of physical activity 5 days a week. If you have not been active, talk with your health care professional to figure out which types of exercise are best for you. You can start slowly and work up to your goal.
- Don't smoke. Smoking can contribute to insulin resistance, which can lead to type 2 diabetes. If you already smoke, try to quit.
- Talk to your health care provider to see whether there is anything else you can do to delay or to prevent type 2 diabetes. If you are at high risk, your provider may suggest that you take one of a few types of diabetes medicines.

About the dataset

This dataset is to predict whether the patients have diabetes or not by using the variables in the dataset. Which variable more significant to the outcome(diabetes) is analyzed by using the data.

The following are the variables used to find the diabetes in this dataset,

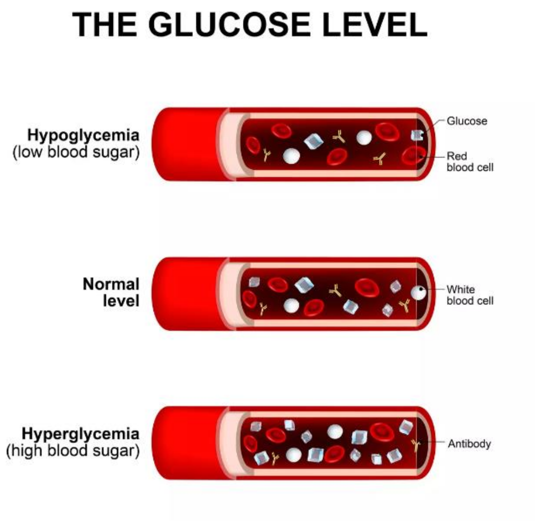
- cholesterol
- glucose
- hdl_chol
- chol_hdl_ratio
- age
- gender
- height
- weight
- bmi
- systolic_bp
- diastolic_bp
- hip
- waist_hip_ratio
- outcome

Cholesterol

Cholesterol comprises carbon, hydrogen, and oxygen. It is a waxy, fatty substance that is solid and white or light yellow. Its chemical formula is $C_{27}H_{46}O$. This means it has 27 atoms of carbon, 46 of hydrogen, and one of oxygen. Cholesterol’s structure consists trusted source of a central sterol nucleus of four hydrocarbon rings, which are hydrogen and carbon atoms with a circular arrangement a hydrocarbon tail, a chain of hydrogen and carbon atoms at the end of a molecule a hydroxyl group, which is one hydrogen atom bonded to one oxygen atom

The four hydrocarbon rings join together in the middle of the compound. The hydrocarbon tail attaches to one end, and the hydroxyl group attaches to the other.

Both the sterol nucleus and hydrocarbon tail do not mix with water, so this structure cannot travel through the bloodstream alone. For this reason, cholesterol combines with proteins to create lipoproteins, which can travel through the blood to reach cells that need them.



Glucose

When someone eats food, their digestive system breaks the food down into its most basic molecular parts. Starch is broken down into the most basic unit of sugar, which is called glucose. Blood glucose is a measurement of the amount of sugar present in the blood. Every tissue and cell in the body, including the brain, requires glucose to function, repair, and grow. This is because glucose is used to make fuel for every cell in the body. The mitochondria in cells break glucose down into even smaller molecules of cellular energy called adenosine triphosphate (ATP). ATP is the basic unit of energy for living organisms, whether that organism is a human, a mouse, or a flea.

Measuring blood glucose levels is important because some illnesses or diseases can cause blood glucose to become too high or too low. If blood glucose levels are at an extreme, catastrophic illness or even death can occur unless there is medical intervention. Central maintaining blood glucose homeostasis are two hormones, insulin and glucagon, both produced by the pancreas and released into the bloodstream in response to changes in blood glucose.

Hdl_Cholesterol

High-density lipoprotein (HDL) cholesterol functions to help clear fats from your bloodstream. As a result, it is known as the "good" cholesterol, in comparison to low-density lipoprotein cholesterol (LDL), which is known as the "bad" kind. Find out what these two types of cholesterol do for you, what your test results mean, and what you can do to improve your cholesterol levels.

Systolic blood pressure

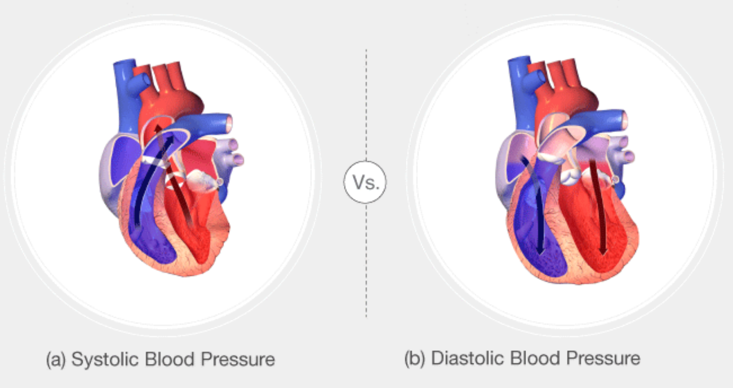
During a heartbeat, the heart pushes blood out into the arteries. Systolic pressure is the measure of this pressure within the arteries while the heart beats. This phase, known as systole, is the point at which blood pressure is the highest.

Systolic blood pressure is considered normal when the reading is below 120 mmHg (millimeters of mercury) while a person is sitting quietly at rest.

Systolic pressure below 90 mmHg is considered low and may require intervention and management from your healthcare provider. If you get multiple systolic pressure readings above 180 mmHg, it is considered dangerously high.

Diastolic blood pressure

The heart rests between beats so it can refill with blood. The pause between beats is called diastole. Your diastolic blood pressure is the measurement during this pause before the next heartbeat.



Normal diastolic blood pressure during quiet rest is below 80 mmHg.1 If you have high blood pressure, the diastolic number is often higher even during quiet rest.

Chapter - II

METHODOLOGY

Logistic regression

This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds.

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. `1For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate the how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit**.**

There are three types of logistic regression models, which are defined based on categorical response.

Binary logistic regression: In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes (e.g. 0 or 1). Some popular examples of its use include predicting if an e-mail is spam or not spam or if a tumor is malignant or not malignant. Within logistic regression, this is the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.

Multinomial logistic regression: In this type of logistic regression model, the dependent variable has three or more possible outcomes; however, these values have no specified order. For example, movie studios want to predict what genre of film a moviegoer is likely to see to market films more effectively. A multinomial logistic regression model can help the studio to determine the strength of influence a person's age, gender, and dating status may have on the type of film that they prefer. The studio can then orient an advertising campaign of a specific movie toward a group of people likely to go see it.

Ordinal logistic regression: This type of logistic regression model is leveraged when the response variable has three or more possible outcome, but in this case, these values do have a defined order. Examples of ordinal responses include grading scales from A to F or rating scales from 1 to 5

****

This data is extracted from Kaggle, which is the website for dataset for many projects. There are some tools and methods I used to analyze this data which is I mentioned below,

- Python
- R language

These are the main tools used in this project. There are many methodology used in these tools

Python

- **Python's** statistics is a built-in Python library for descriptive statistics. You can use it if your datasets are not too large or if you can't rely on importing other libraries.
- NumPy is a third-party library for numerical computing, optimized for working with single- and multi-dimensional arrays. Its primary

type is the array type called ndarray. This library contains many routines for statistical analysis.

- **SciPy** is a third-party library for scientific computing based on NumPy. It offers additional functionality compared to NumPy, including scipy.stats for statistical analysis.
- **pandas** is a third-party library for numerical computing based on NumPy. It excels in handling labeled one-dimensional (1D) data with Series objects and two-dimensional (2D) data with DataFrame objects.
- **Matplotlib** is a third-party library for data visualization. It works well in combination with NumPy, SciPy, and pandas

R language

R is a language that is designed for statistical computing, graphical data analysis, and scientific research. It is usually preferred for data visualization as it offers flexibility and minimum required coding through its packages.

- Univariate
- Bivariate
- Outliers removal
- Logistic regression
- Stepwise regression
- ROC curve

these are some the methodology that I used in r language for analyzing this data.

Univariate

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data**.**

Bivariate

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them.

Bivariate analysis can be helpful in testing simple hypotheses of association. Bivariate analysis can help determine to what extent it becomes easier to know and predict a value for one variable.

If the dependent variable, the one whose value is determined to some extent by the other, independent variable is a categorical variable, such as the preferred brand of cereal, then probit or logit regression (or multinomial probit or multinomial logit) can be used. If both variables are ordinal, meaning they are ranked in a sequence as first, second, etc., then a rank correlation coefficient can be computed. If just the dependent variable is ordinal, ordered probit or ordered logit can be used. If the dependent variable is continuous-either interval level or ratio level, such as a temperature scale or an income scale then simple regression can be used.

Outlier removal

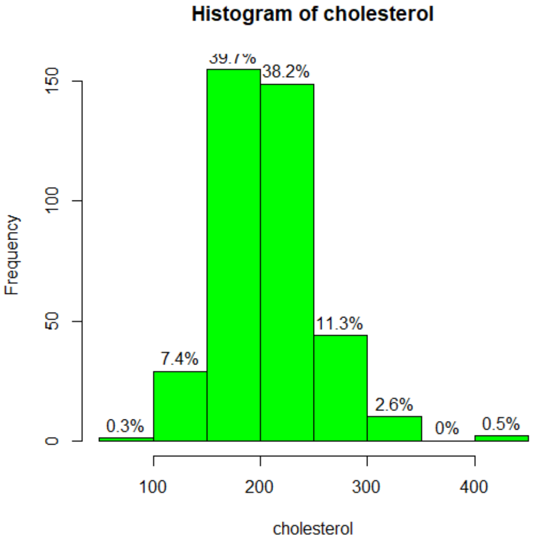
An Overview of outliers and why it’s important for a data scientist to identify and remove them from data. Understand different techniques for outlier treatment: trimming, capping, treating as a missing value, and discretization. Understanding different plots and libraries for visualizing and treating outliers in a dataset.

Some outliers represent true values from natural variation in the population. Other outliers may result from incorrect data entry, equipment malfunctions, or other measurement errors. An outlier isn’t always a form of dirty or incorrect data, so you have to be careful with them in data cleansing. What you should do with an outlier depends on its
most likely cause

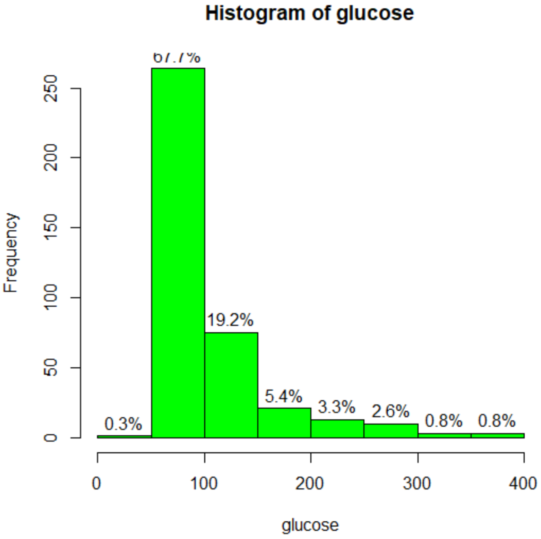
Chapter - III

VISUALIZATION AND ANALIZATION

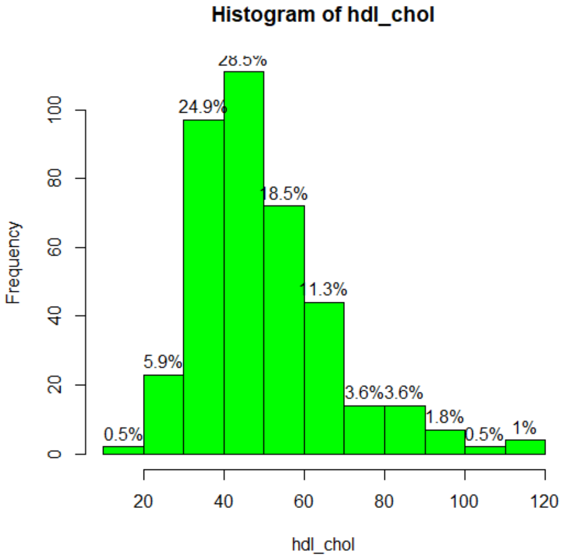
Univariate



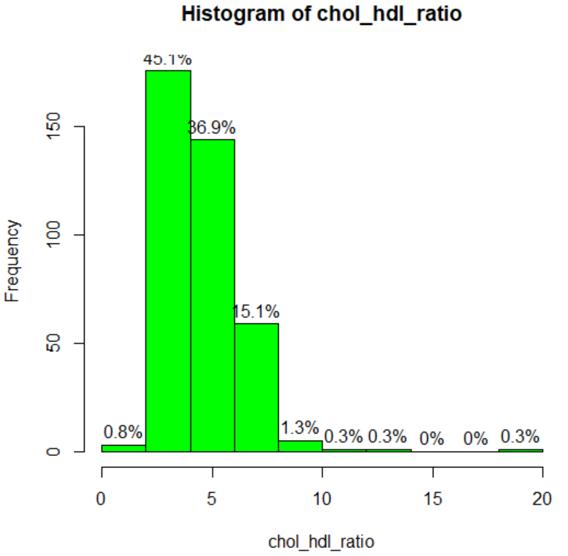
The visual representation of the variable Cholesterol is visualized here, the maximum of 39.7% patients have high Cholesterol and minimum of 0.3% of the patients have low Cholesterol.



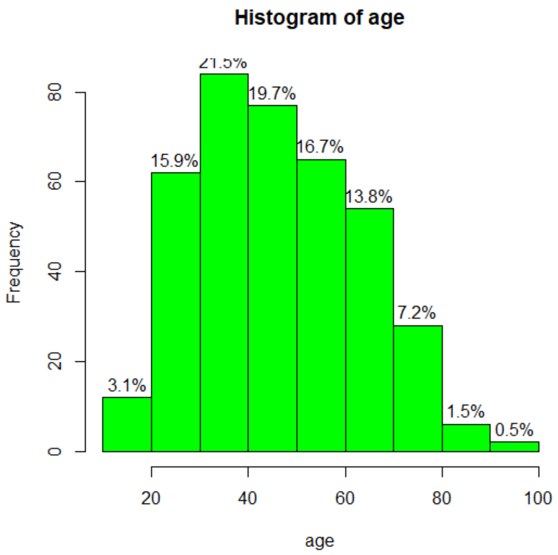
This is the visual representation of the variable glucose, here 67% of the patients has hig glucose level and 0.3% of the patients has low glucose level.



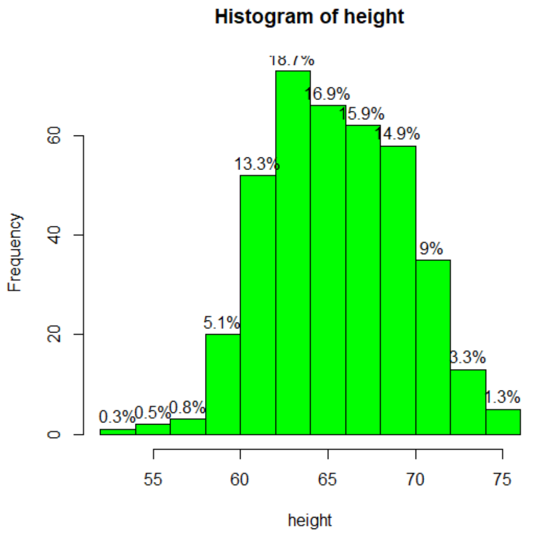
This is the visual representation of the variable hdl_chol which is the **high-density lipoprotein** cholesterol, here 28.5% of the patients have high level and 0.5% of the patients have low level.



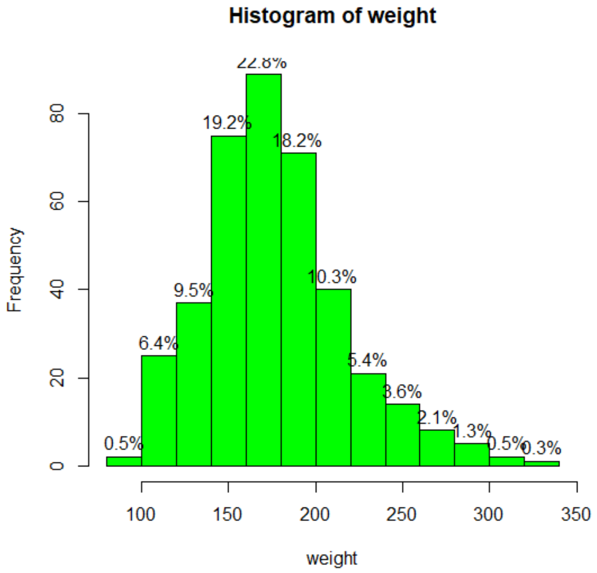
Visual representation of the variable chol_hdl_ratio is visualized. Here 45.1% of the patients has high **hdl cholesterl ratio** and 0.3% of the patints has low ratio.



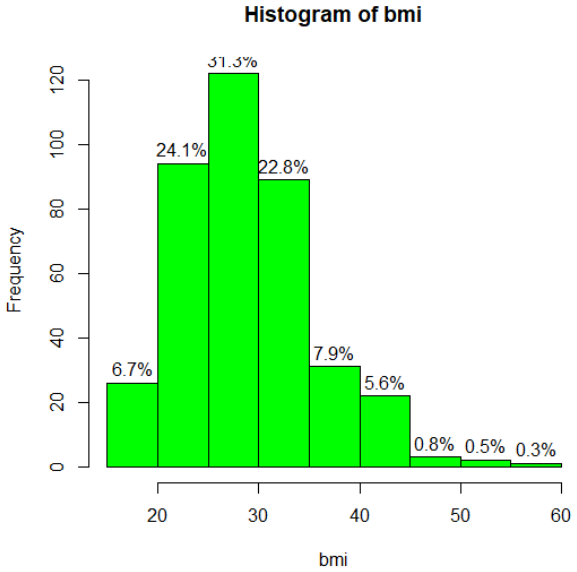
Here this histogram shows that about 21.5% of the patients were the age between 30 and 40, and 0.5% of the patients were 90 to 100.



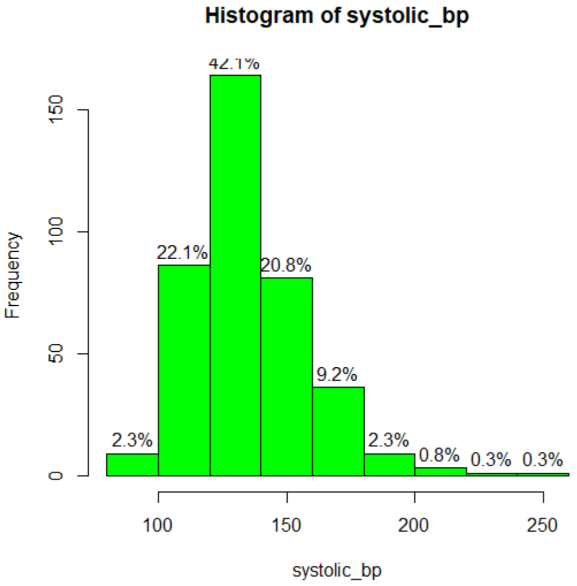
The visual representation of height shows that 18.7% of the patients were taller and 0.3% of the patients were short.



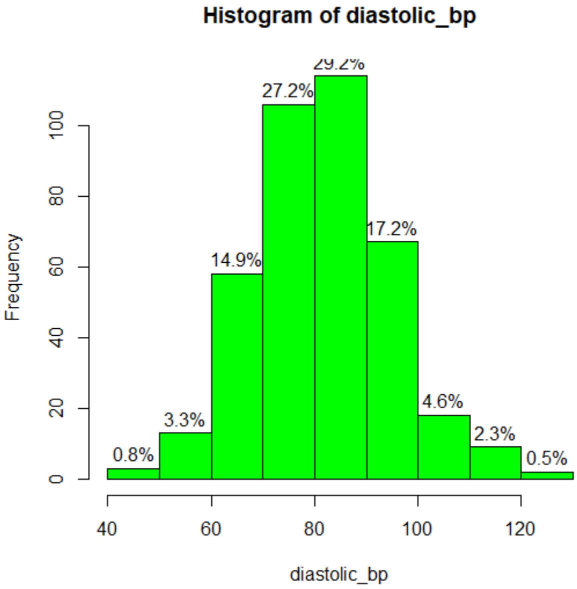
According to the visual representation of the data, 22.8% of the patients weights between 150 to 200 lbs, and 0.3% patients were over weight.



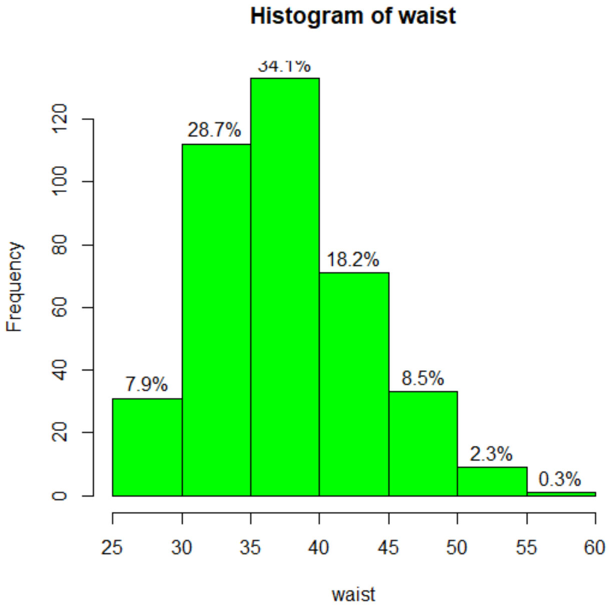
This histogram shows that 31.3% of the patients has the bmi between 25 to 30, and 0.3% of the patients has the bmi between 55 to 60.



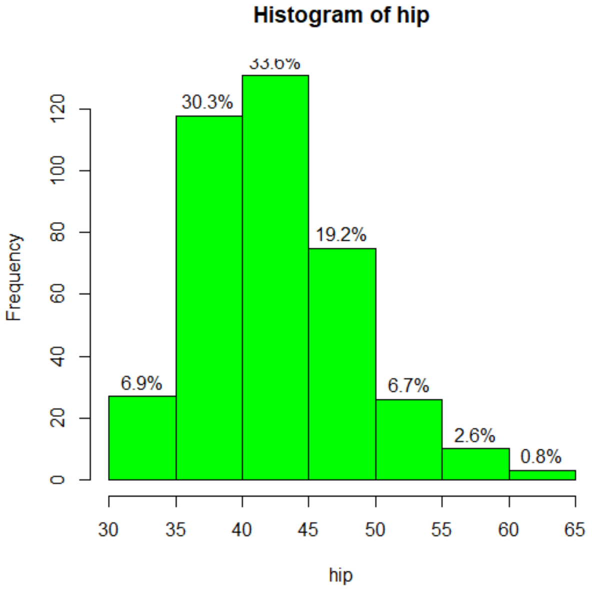
Here 42.1% of the patients has the high systolic bp and 0.3% of the patients has low systolic bp.



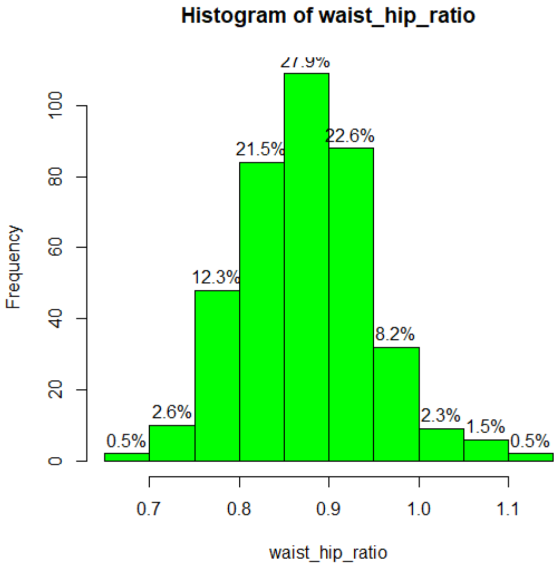
Here 29.2% of the patients has the high diastolic bp and 0.5% of the patients has low systolic bp.



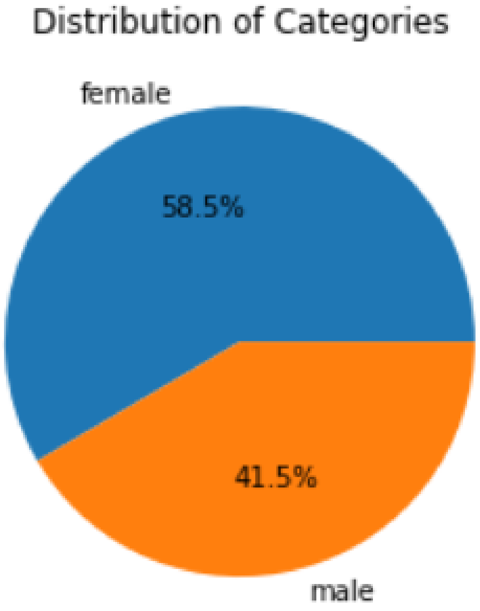
This histogram shows 34.1% of the patients has waist between 35 to 40, 0.3% of the patients has between 55 to 60.



This histogram shows 33.6% of the patients has hip between 40 to 45, 0.8% of the patients has between 60 to 65.



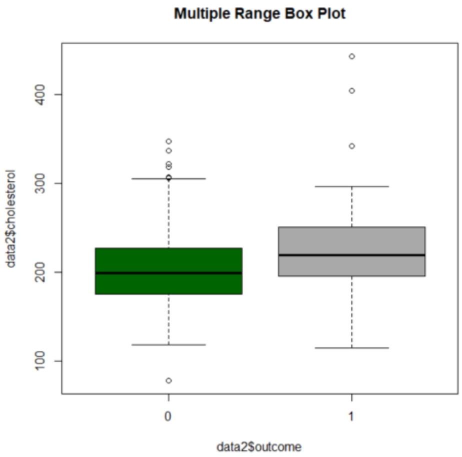
This histogram shows 27.9% of the patients have the ratio between 0.85 to 0.9, and 0.5% of the patients has less than 0.7 and greater than 1.1



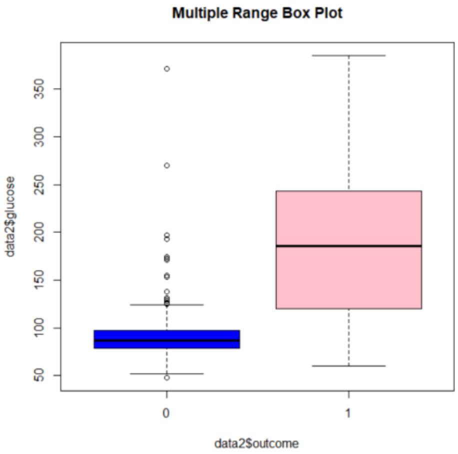
This is categorical variable so pie chat is used for data visualization. Female patients were higher compared to male patients.

Bivariate

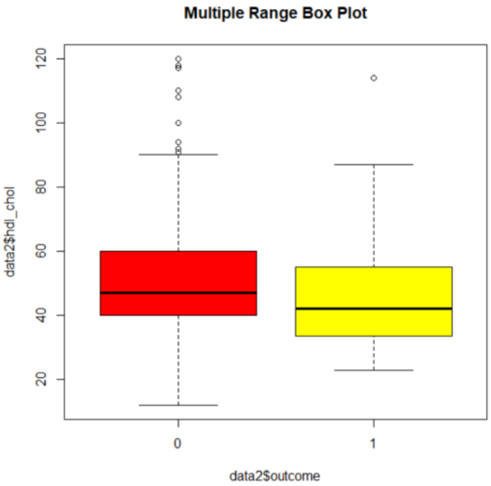
Cholesterol vs Outcome



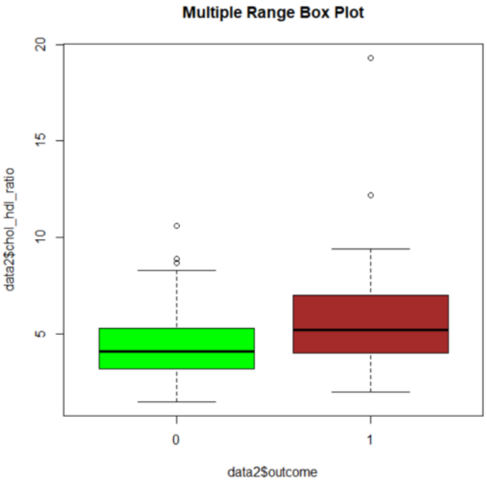
Glucose vs Outcome



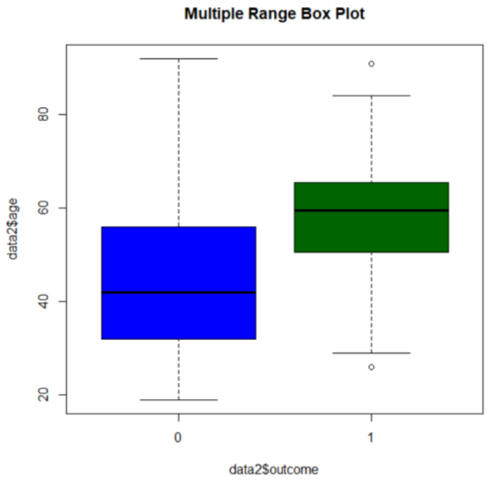
Hdl_chol vs Outcome



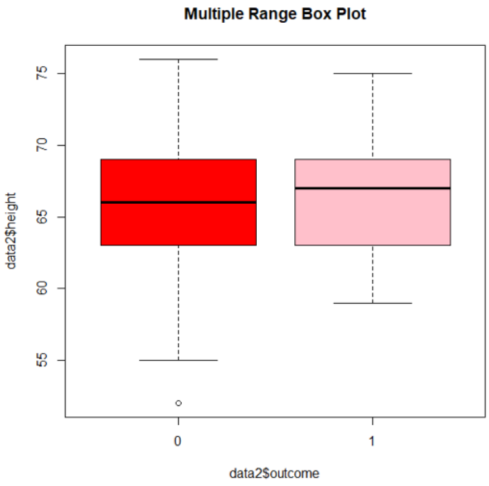
chol_hdl_ratio vs Outcome



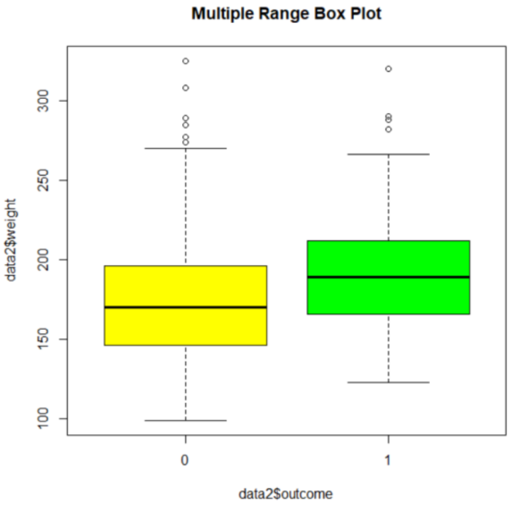
Age vs Outcome



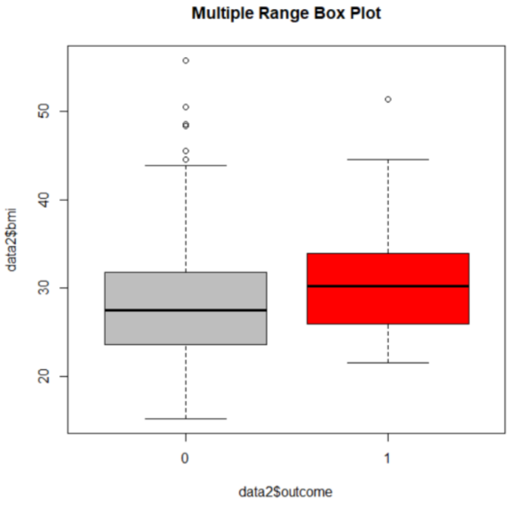
Height vs Outcome



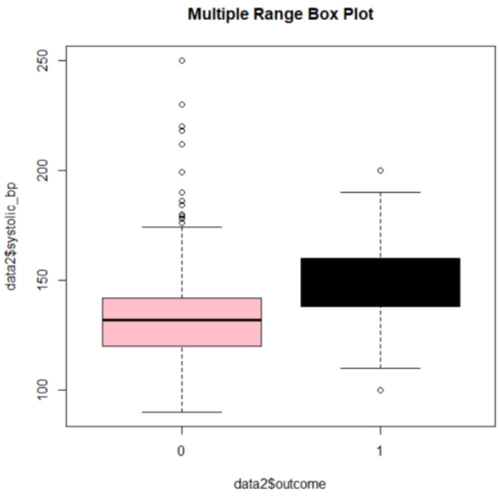
Weight vs Outcome



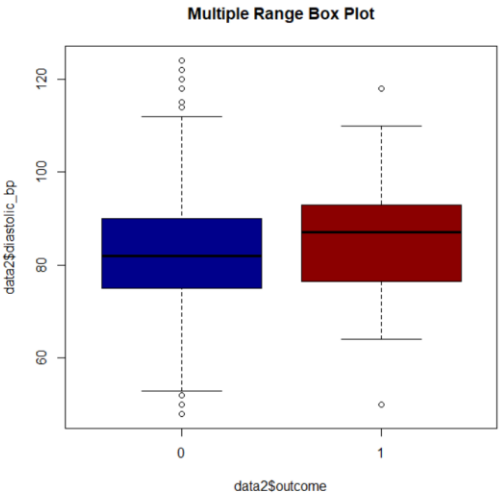
Bmi vs Outcome



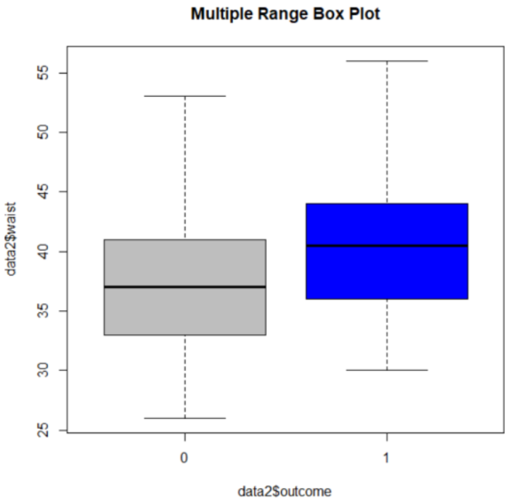
Systolic_bp vs Outcome



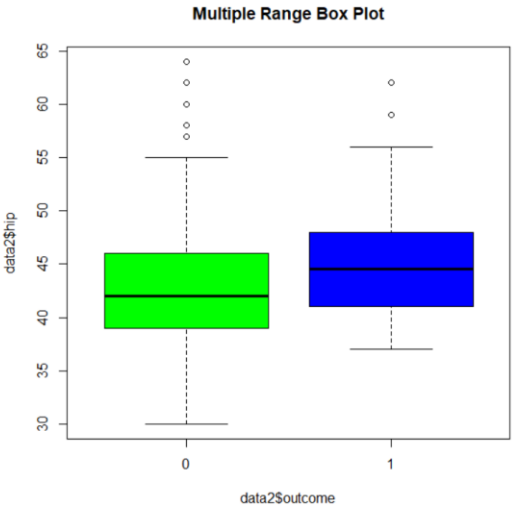
Diastolic_bp vs Outcome



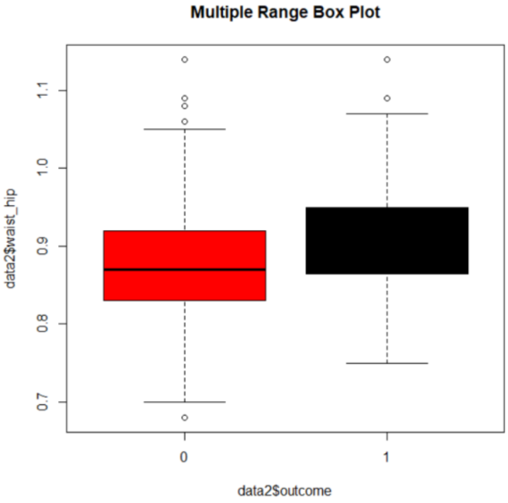
Waist vs Outcome



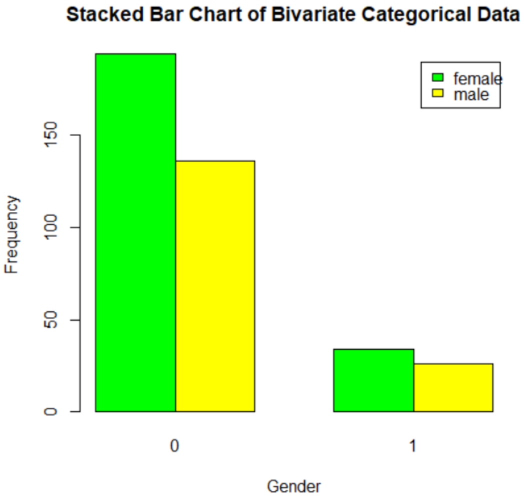
Hip vs Outcome



Waist_hip_ratio vs Outcome



Gender vs Outcome



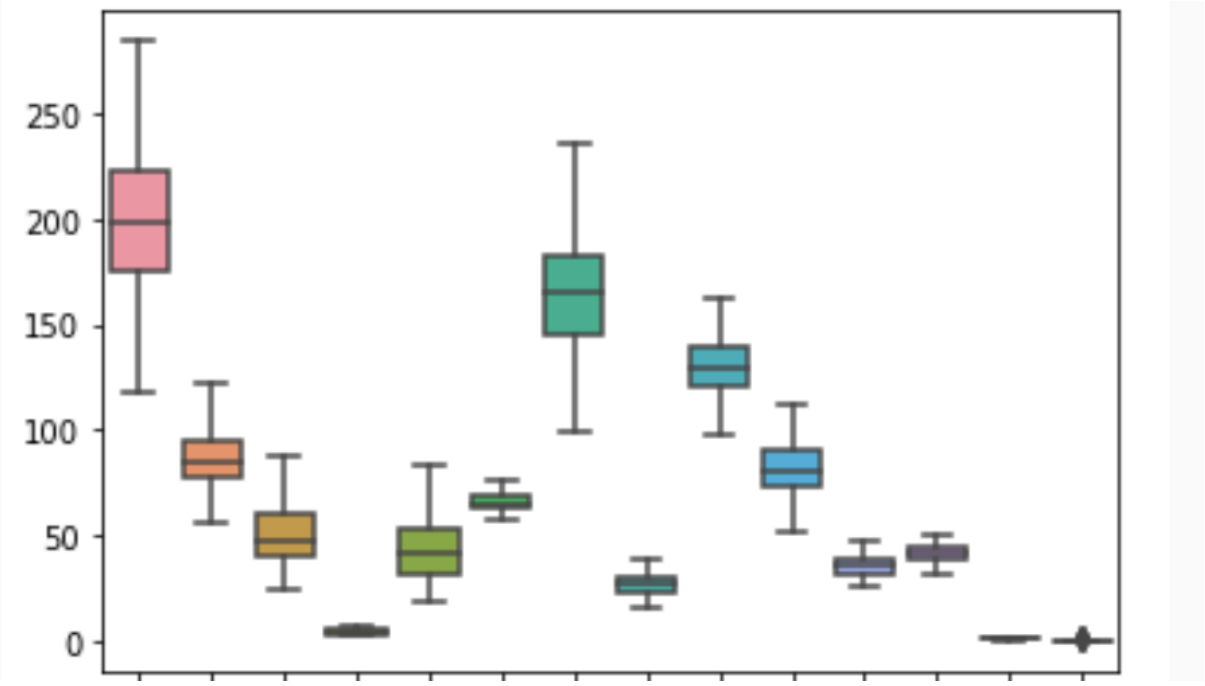
Here is the box plot which shows the comparison of each variable to the outcome, but only one independent variable is categorical so I used histogram for bivariate.

The primary use of the box plot is to identify whether the given data has outlier or not, in this data we have outlier we need to remove the outlier in order to get the perfect model it should be eliminated.

As we see the visual representation of bivariate we can see small-small circles in the plot those are the outlies in that respective variables.

Now by using python code I'm removing outliers from the data and extract the .csv data for further analyzation.

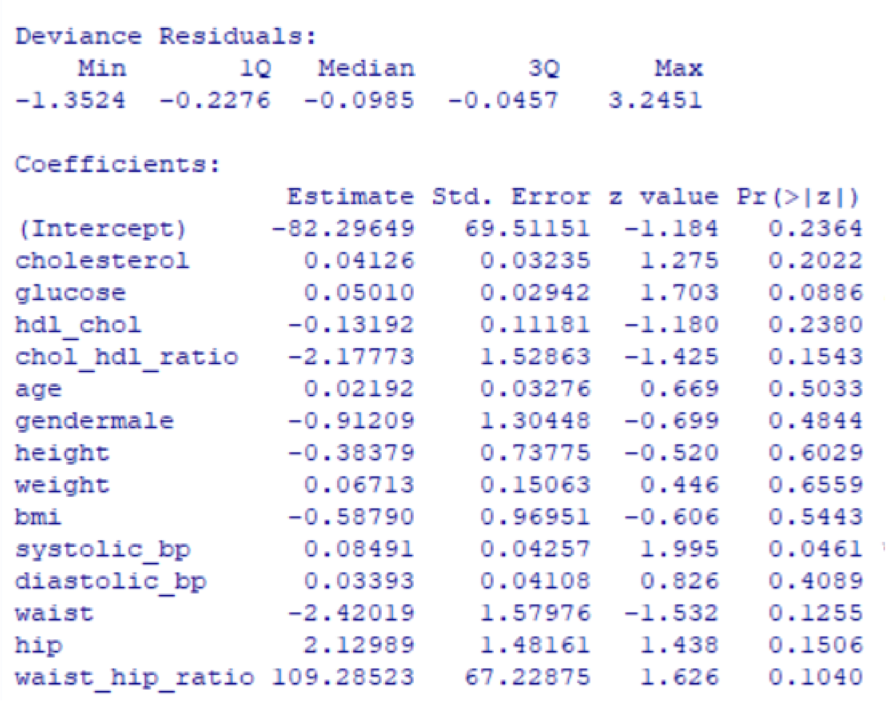
Visual representation of variables after removing Outliers



As the above box plot shows there is no outliers in the data, the data was extracted after removing the outliers. There are no small circles outside the box plot so the outliers were removed and the new data was extracted.

Logistic regression

Building Logistic Regression Model: Initially we built the model with all the variables and found that there are many variables are insignificant (have high p-value). We need to reduce the number of variables.

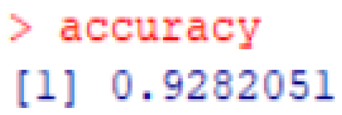


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3524  -0.2276  -0.0985  -0.0457   3.2451

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -82.29649   69.51151  -1.184   0.2364
cholesterol       0.04126    0.03235   1.275   0.2022
glucose           0.05010    0.02942   1.703   0.0886
hdl_chol         -0.13192    0.11181  -1.180   0.2380
chol_hdl_ratio   -2.17773    1.52863  -1.425   0.1543
age               0.02192    0.03276   0.669   0.5033
gendermale       -0.91209    1.30448  -0.699   0.4844
height           -0.38379    0.73775  -0.520   0.6029
weight            0.06713    0.15063   0.446   0.6559
bmi              -0.58790    0.96951  -0.606   0.5443
systolic_bp       0.08491    0.04257   1.995   0.0461
diastolic_bp      0.03393    0.04108   0.826   0.4089
waist            -2.42019    1.57976  -1.532   0.1255
hip               2.12989    1.48161   1.438   0.1506
waist_hip_ratio 109.28523   67.22875   1.626   0.1040
```

As we see here the p value most variables were not significant, so by using stepwise regression I'm removing the non-significant variables and keeping the most significant variable.

By using that this model will be the most efficient model, and by using this logistic regression found the accuracy of the model

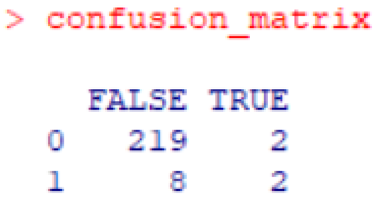


```
> accuracy
[1] 0.9282051
```

The accuracy of the overall model is 0.928 which is a good model, now after removing the non-significant variables the accuracy might varies.

Confusion Matrix

A Confusion matrix is an **N x N matrix** used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.



```
> confusion_matrix

   FALSE TRUE
0    219    2
1      8    2
```

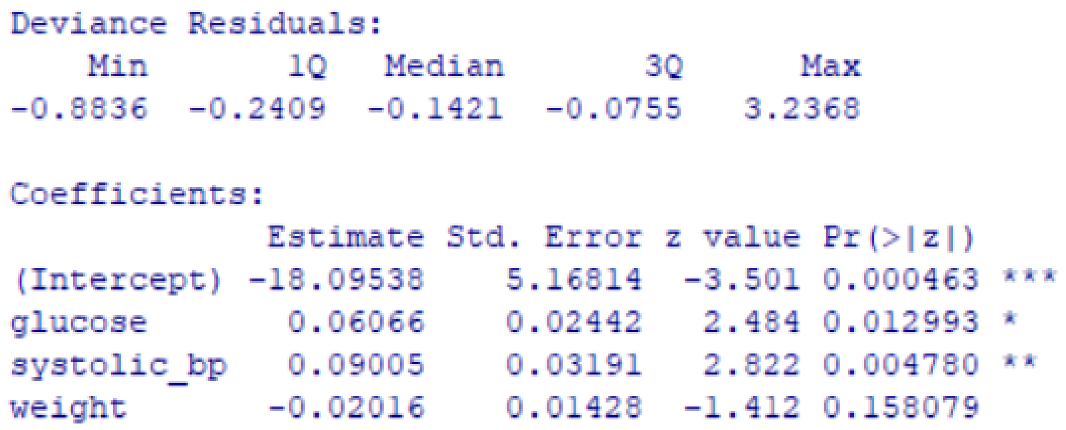
Stepwise regression

Stepwise regression is a step-by-step process of constructing a model by introducing or eliminating predictor variables. First, the variables undergo T-tests and F-tests. Then, predictor variables are individually tested to fit a linear regression model.

There are two main types of stepwise regression:

Forward Selection – In forward selection, the algorithm starts with an empty model and iteratively adds variables to the model until no further improvement is made.

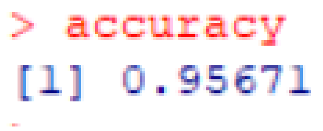
Backward Elimination – In backward elimination, the algorithm starts with a model that includes all variables and iteratively removes variables until no further improvement is made.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8836  -0.2409  -0.1421  -0.0755   3.2368

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.09538    5.16814  -3.501 0.000463 ***
glucose       0.06066    0.02442   2.484 0.012993 *
systolic_bp   0.09005    0.03191   2.822 0.004780 **
weight       -0.02016    0.01428  -1.412 0.158079
```

From the above output, the variables glucose and systolic blood pressure are more significant to the outcome. By using the stepwise regression most of the non-significant variables were eliminated. Compared to other variables Weight is close to the significant, but it was not the sufficient one.



```
> accuracy
[1] 0.95671
```

As mentioned above the accuracy of the model varies from above. The most significant variables have the accuracy of 0.956 which is better than the previous model.

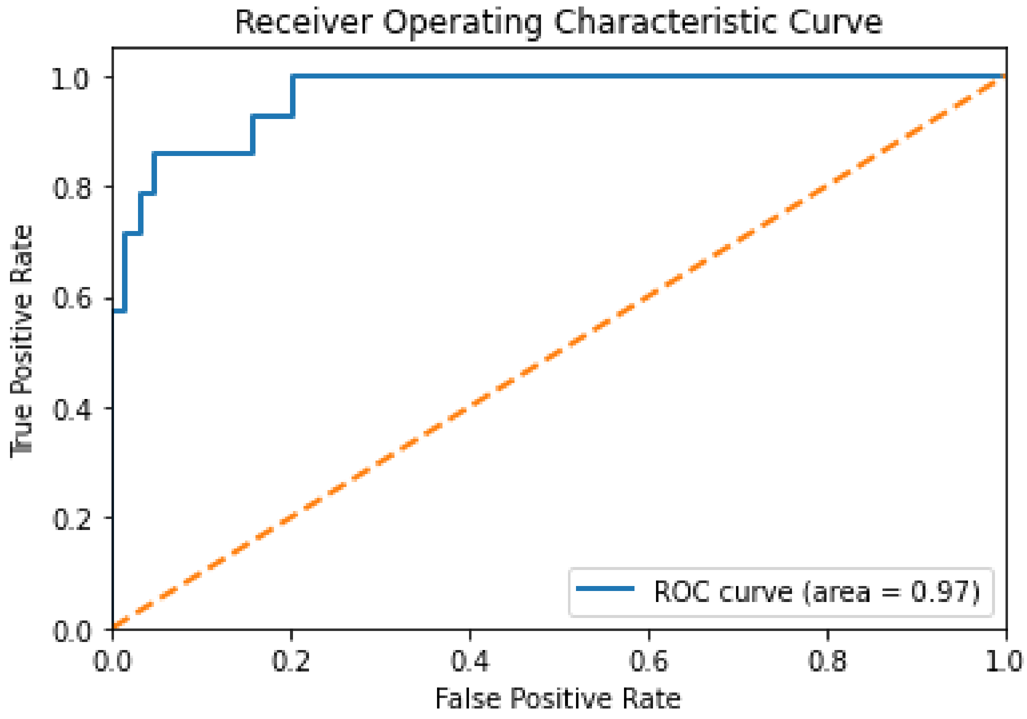
ROC curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters,

- True Positive Rate
- False Positive Rate

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

To compute the points in an ROC curve, we could evaluate a logistic regression model many times with different classification thresholds, but this would be inefficient. Fortunately, there's an efficient, sorting-based algorithm that can provide this information for us, called AUC.



The Person is diabetic or non-diabetic the model then make the prediction. The Area under the curve is 0.97 which is good enough to make a prediction if a patient is diabetic or not.

Chapter – IV

INTERPRETATION

Univariate

Histogram is used for univariate analysis, how the data is distributed and visualizing the data using histogram. It is also useful to check the pattern of the data

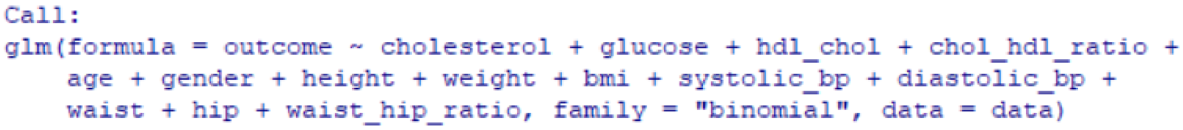
Bivariate

Bivariate, box plot is used to compare each variable to the depend variable. There are outliers in the data and the outliers were removed and the whole data is visualized in a single plot.

Logistic regression

Logistic regression is used in the overall data and there is many variables which are not significant to the dependent variable, here is the model for the overall data,

$$\begin{aligned}\text{Outcome} = {} & \beta_1\,(\text{cholesterol}) + \beta_2\,(\text{glucose}) + \beta_3\,(\text{hdl\_chol}) + \beta_4\,(\text{chol\_hdl\_ratio}) + \\ & \beta_5\,(\text{age}) + \beta_6\,(\text{gender}) + \beta_7\,(\text{height}) + \beta_8(\text{weight}) + \beta_9\,(\text{bmi}) + \\ & \beta_{10}\,(\text{systolic\_bp}) + \beta_{11}\,(\text{diastolic\_bp}) + \beta_{12}\,(\text{waist}) + \beta_{13}\,(\text{hip}) + \\ & \beta_{14}\,(\text{waist\_hip\_ratio})\end{aligned}$$



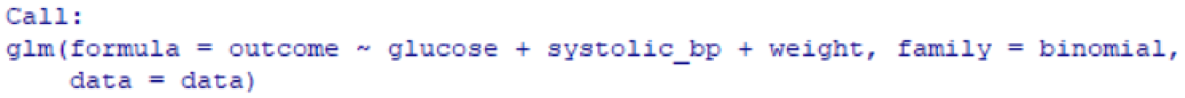
```
Call:
glm(formula = outcome ~ cholesterol + glucose + hdl_chol + chol_hdl_ratio +
    age + gender + height + weight + bmi + systolic_bp + diastolic_bp +
    waist + hip + waist_hip_ratio, family = "binomial", data = data)
```

This is the model for the overall data, were all the variables were fitted in the model building.

Stepwise regression

Most significant variables were separated from the data using the stepwise regression

Outcome = β_1 (glucose) + β_2 (hdl_chol) + β_3 (systolic_bp) + β_4 (weight)



```
Call:
glm(formula = outcome ~ glucose + systolic_bp + weight, family = binomial,
    data = data)
```

Glucose and Systolic blood pressure are the two variables which are more significant to the depend variable.

The p value of glucose is 0.012993 and p value of systolic_bp is 0.00478 both are less than 0.05

Confusion matrix

219 false and 2 true values were predicted correctly and 8 false and 2 true values were predicted wrongly. Here most of the values were predicted correctly.

ROC curve

ROC curve shows the area which the data is distributed. The logistic regression model had an accuracy of 92% and an AUC in the ROC curve of 0.97. The AUC in the ROC curve of 0.97. The most significant model accuracy of 95% and an AUC in the ROC of 0.97.