VISUALIZATION

CHENNAI MATHEMATICAL INSTITUTE

MASTERS IN DATA SCIENCE

Diabetes for PIMA Indian female Population

Author:
Abhishek Chakraborty

Roll No:
MDS202202

Instructor:
Sourish Das

Email:
abhishekc@cmi.ac.in

December 9, 2022

Contents

1 Introduction **3**

1.1 Aims and Objectives . . . 3

2 Dataset Description **3**

2.1 Pregnancies . . . 3

2.2 Glucose . . . 3

2.3 BloodPressure . . . 4

2.4 SkinThickness . . . 4

2.5 Insulin . . . 4

2.6 Body Mass Index (BMI) . . . 4

2.7 DiabetesPedigreeFunction . . . 4

2.8 Age . . . 4

2.9 Outcome . . . 4

3 Graphical Presentation of Key Variables **5**

4 Summary of Analysis **10**

5 Conclusion **10**

List of Figures

1 boxplot . . . 5

2 Distribution of Variables . . . 6

3 Impact of Age over Outcome . . . 7

4 Glucose Level and Blood Sugar against Age . . . 7

5 Distribution of BMI . . . 8

6 Relationship between Glucose and Diabetes Pedigree Function vs Pregnancies . . . 8

7 Correlation Matrix of the attributes . . . 9

Abstract

This project will be used to show the relationship between the attributes of the PIMA Indians Diabetes dataset for females. The main focus of this paper will be to show if the indicator variable Outcome, i.e. whether a person has diabetes or not is related to the other attributes, using Visualization and Exploratory Data Analysis (EDA).

1 Introduction

PIMA Indians are native Americans who are based in Arizona Area. In this project, we have taken the PIMA Indians Diabetes dataset for females. We will try to visualise the attributes and do some Exploratory Data Analysis (EDA) and try to show how the attributes are related to each other through the use of R Programming.

1.1 Aims and Objectives

- Visualization of different attributes of the dataset.
- Exploratory Data Analysis of the dataset.
- Trying to visually represent the relationship between the causes of diabetes among PIMA Indian Females.
- Visually analysing whether a factor is actually significant in causing diabetes among PIMA Indian Females.

2 Dataset Description

This dataset contains of 8 predictor (independent) variables and one target (dependent) variable, namely Outcome. The predictor variables are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age.

In the next subsections of this section, we will try to provide a small description of each of the attributes of the dataset.

2.1 Pregnancies

This column of the dataset shows the number of times a PIMA Indian female has gotten pregnant in her entire life so far. We can show how number of pregnancies affect the chance of that person being diabetic. This is a quantitative numeric variable.

2.2 Glucose

This column of the dataset shows the Plasma Glucose concentration in 2 hours of an oral Glucose Tolerance Test. This is a continuous numeric variable. We can show either this correctly predicts whether a person has diabetes or not.

2.3 BloodPressure

This column of the dataset shows the Diastolic Blood Pressure in mm Hg of a person. In later part of this report we will show whether BloodPressure correctly predicts whether a person has diabetes or not. This is a continuous numeric variable.

2.4 SkinThickness

This column of the dataset shows the Triceps skin fold thickness in mm of PIMA Indian females who are under consideration. This column will be useful to show whether obesity is somewhat related to someone having diabetes or not. Although skin thickness is not related to obesity linearly, but we can assume they are somewhat correlated. This is a continuous numeric variable.

2.5 Insulin

This column of the dataset shows the 2-Hour serum insulin in mu U/ml. This column will be useful to show whether insulin levels are related to whether a person has diabetes or not. This is a continuous numeric variable.

2.6 Body Mass Index (BMI)

This column of the dataset shows the Body Mass Index (Weight in kg / (Height in m)^2) of PIMA Indian Female. This will actually show whether a person with maintained height-mass ratio or a person with obese body or malnutrition-ed body will have higher chance of having diabetes. This is a continuous numeric variable.

2.7 DiabetesPedigreeFunction

This column of the dataset depicts the Diabetes Pedigree Function for the given population. Diabetes Pedigree Function is defined as - **Diabetes Pedigree Function**: indicates the function which scores likelihood of diabetes based on family history. This is a continuous numeric variable.

2.8 Age

This column of the dataset depicts the age of the population. Later on in this report we will show whether the chance of having diabetes increases with age or not. This is a continuous numeric variable.

2.9 Outcome

Outcome is a categorical variable which takes the value 1 if a person has diabetes and takes the value 0 if a person does not have diabetes.

3 Graphical Presentation of Key Variables

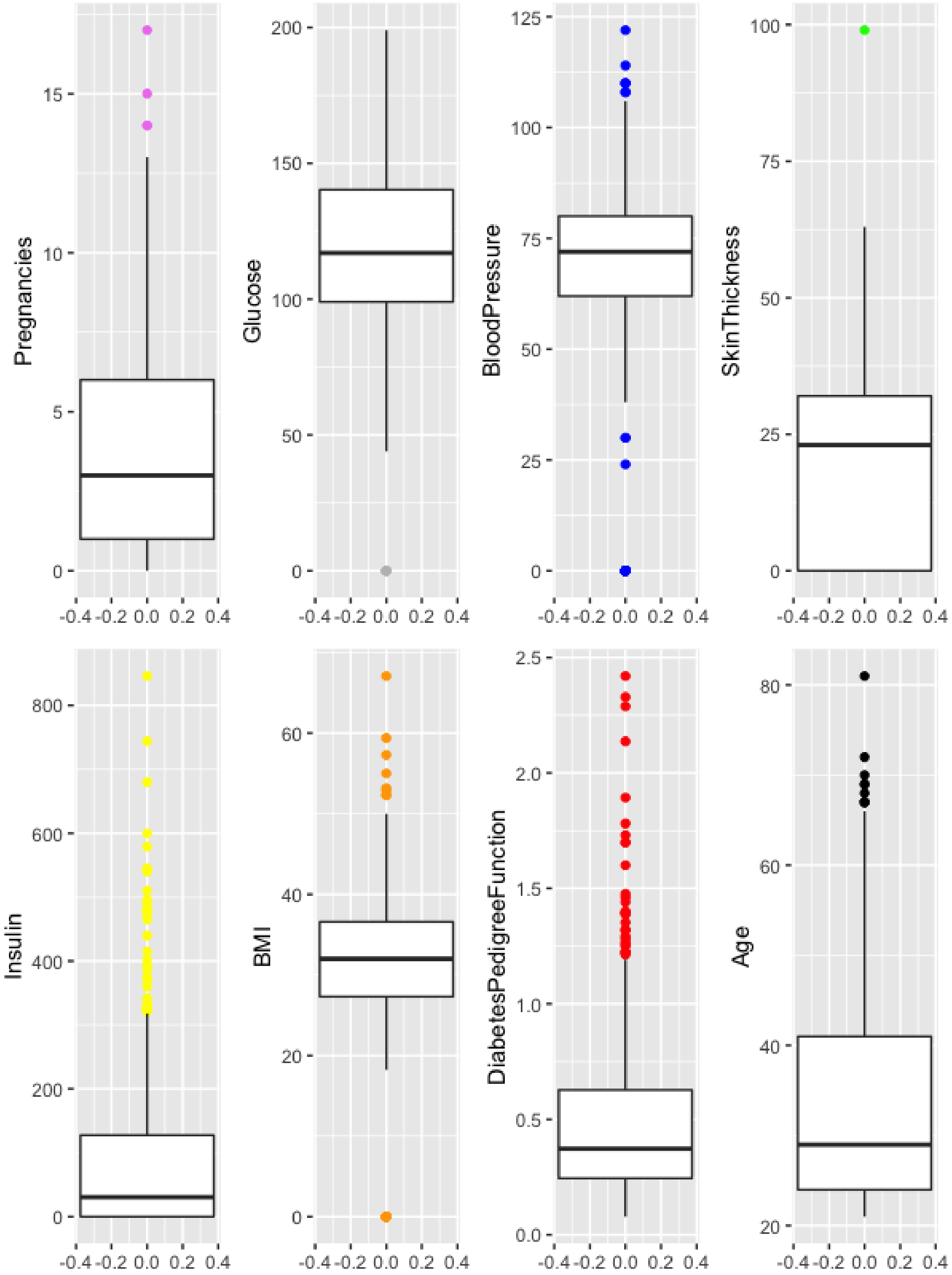

Figure 1: boxplot

The above figure shows the boxplots of the attributes of our dataset. A boxplot is often referred as Five Point Summary. So in each of the boxplots, the minimum value, maximum value, first, second and third quartile are represented for each of the variables.

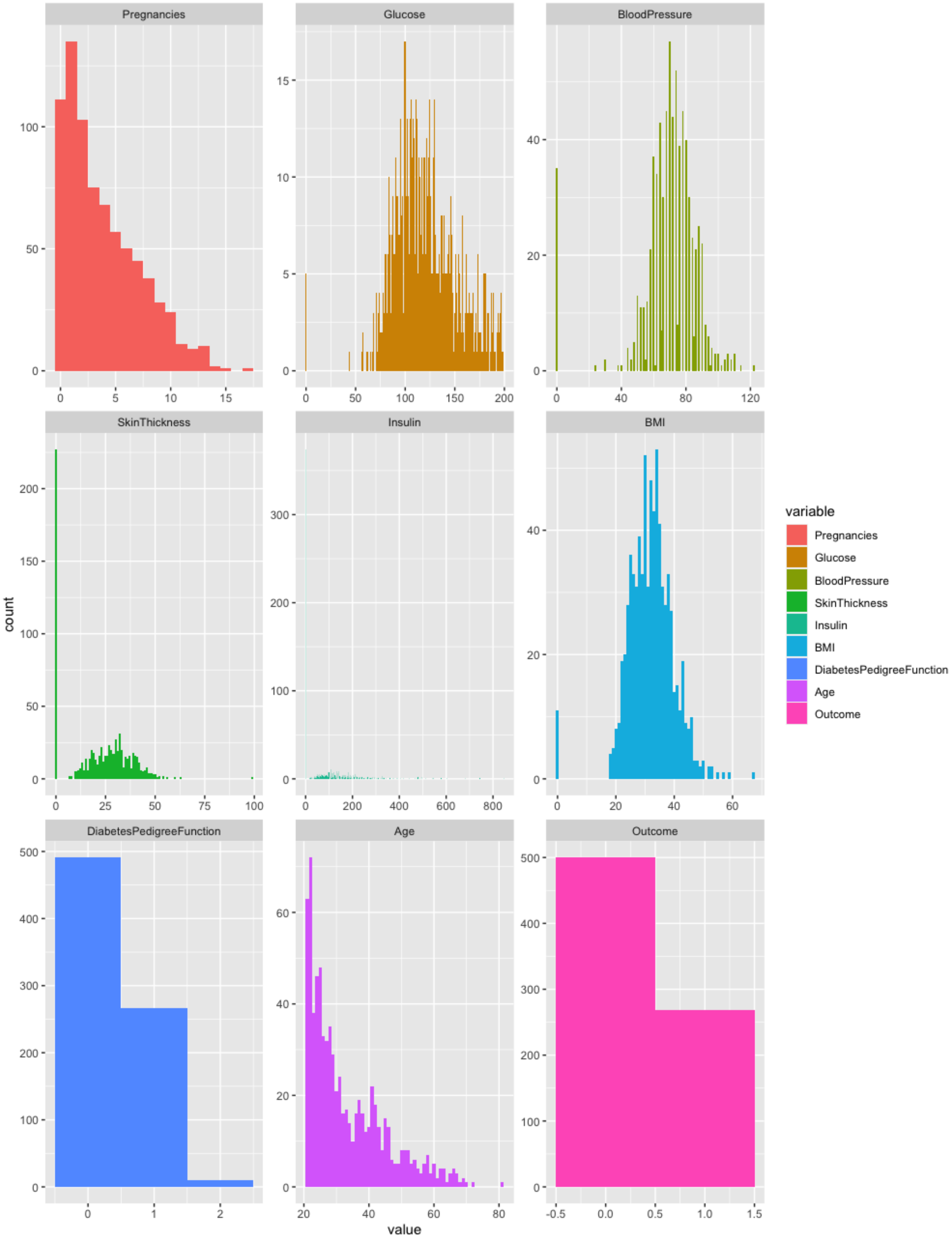

Figure 2: Distribution of Variables

This shows the distributions of the variables of the dataset. This figure (Figure 1) actually depicts the histograms of the variables and shows whether the variables are positively skewed or negatively skewed or symmetric.

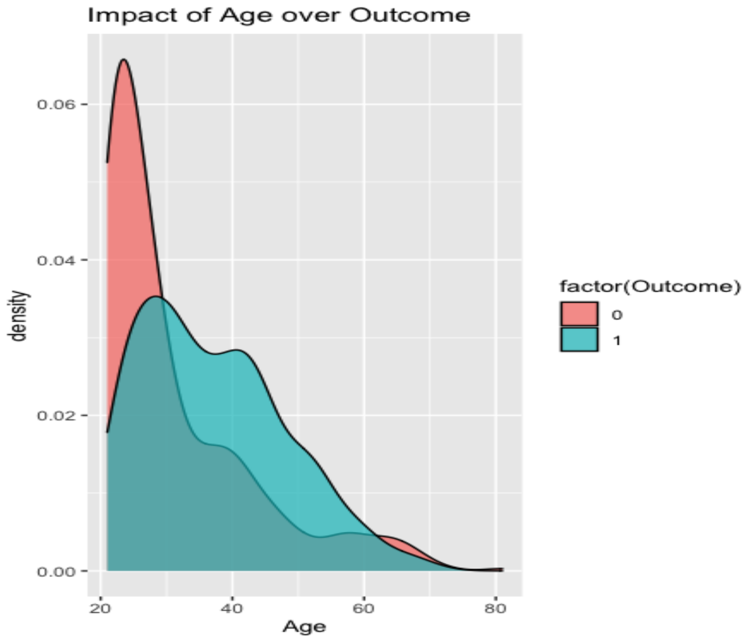

Figure 3: Impact of Age over Outcome

This figure shows how the distribution of Outcome is dependent on the attribute Age. As we can see from the figure that the chance of someone having glucose increases as they get older. Like we can see that the chance that a PIMA Indian female has diabetes is very less if their age is between 20 - 30. And the chance of having diabetes is comparatively more if a PIMA Indian female is of the age group 60+.

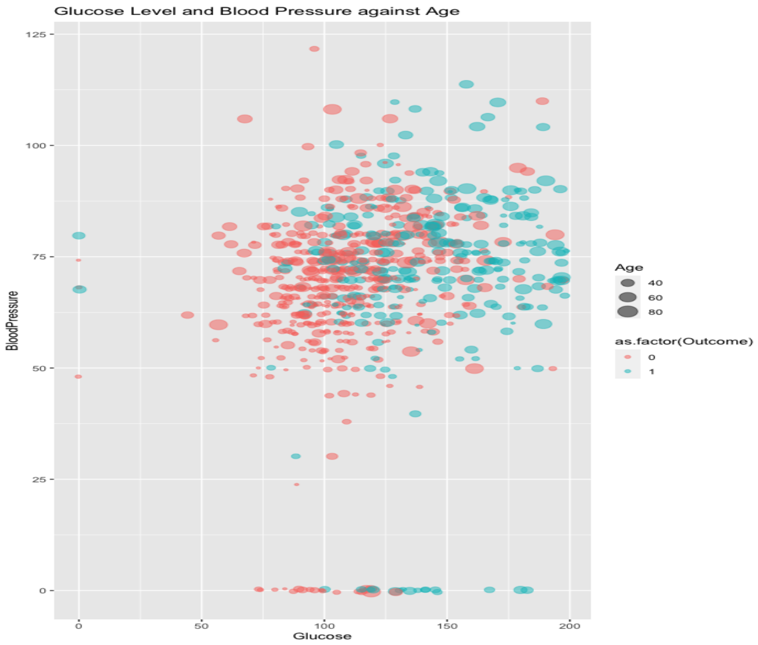

Figure 4: Glucose Level and Blood Sugar against Age

The above figure shows whether with age, the increasing (decreasing) blood pressure affects the glucose levels of the population under consideration. As we can see that as age increases, the increasing blood pressure is somewhat related to increasing glucose levels and vice versa.

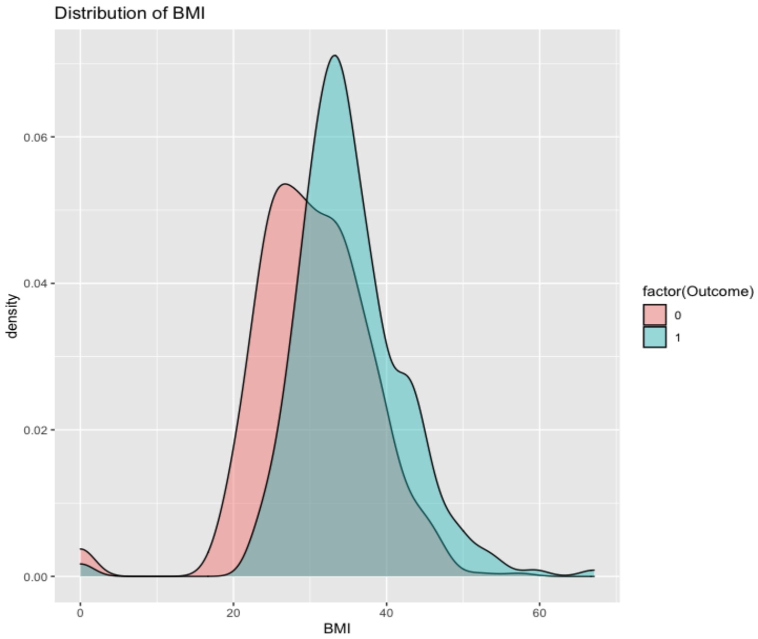

Figure 5: Distribution of BMI

The above figure shows how Body Mass Index (BMI), which is defined as Weight in kg / (Height in m)^2 is distributed and how it affects whether a person has diabetes or not. Now we know that the perfect range of BMI for women is 20 - 24 and any value of BMI above that implies that the person in concern is considered overweight. As we can see that as a person has a tendency to be overweight, the chance that the person has diabetes increases.

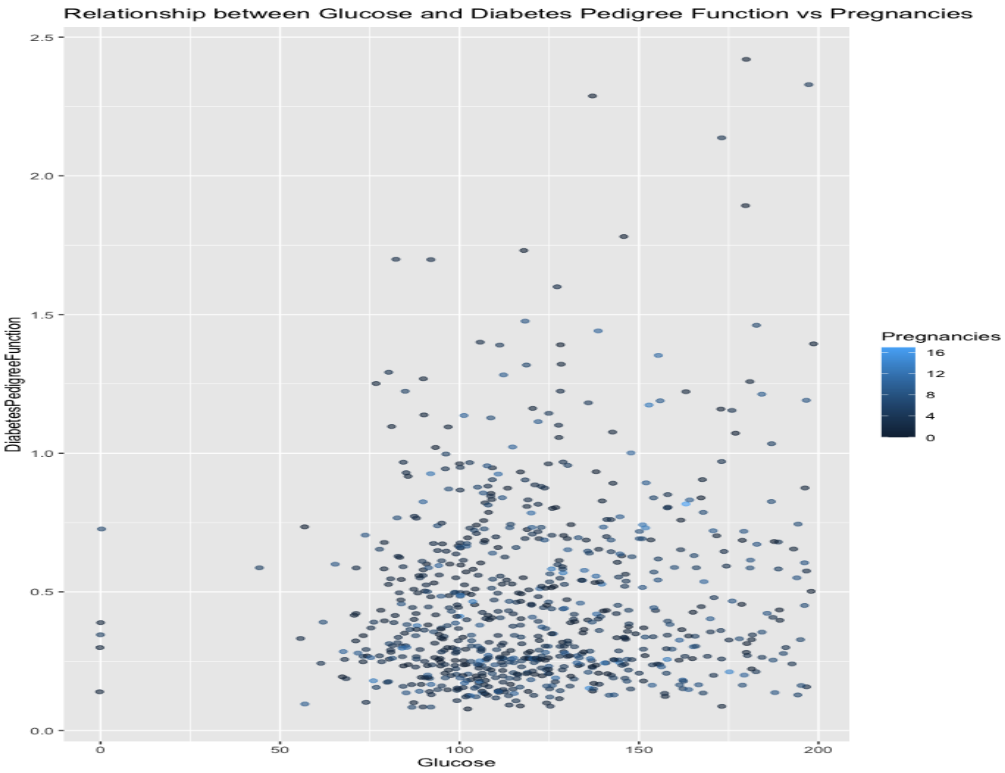

Figure 6: Relationship between Glucose and Diabetes Pedigree Function vs Pregnancies

Figure 6 shows how Glucose level of a person is related to Diabetes Pedigree Function based on the number of times she got pregnant. From the graph it is not quite clear whether these two attributes are linearly related to each other or not, but there is a chance that if a person has hereditary tendency of diabetes, then her glucose level tends to be more than normal. Also, there is a non uniform trend that the more times a person got pregnant, the more their chances of having higher glucose levels in their body.

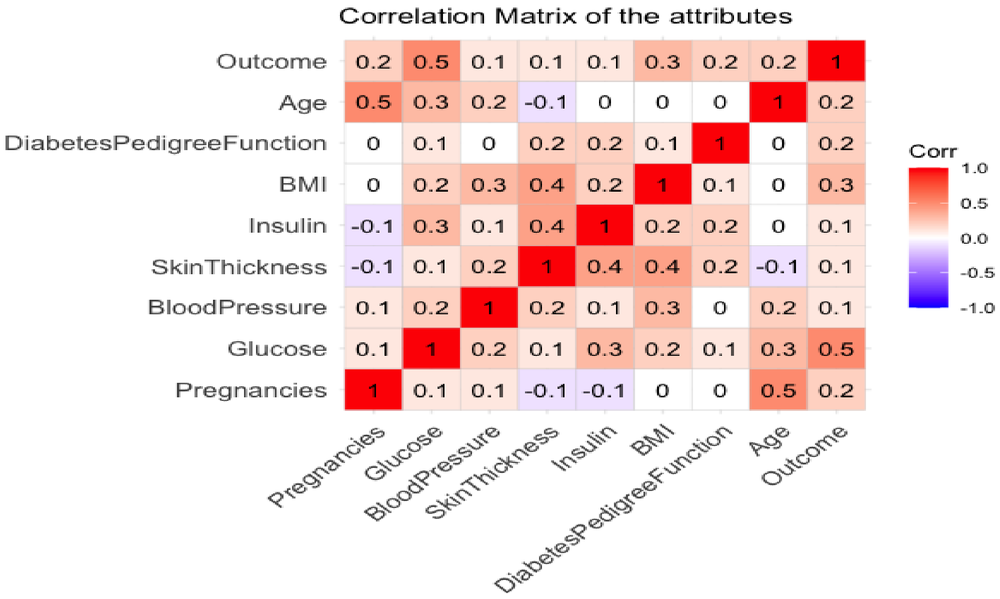

Figure 7: Correlation Matrix of the attributes

Figure 7 shows the correlation matrix of the attributes of the dataset. The correlation matrix depicts the correlation coefficient between pairs of variables of the dataset. We can see some positive linear associations and some negative associations between the variables. Clearly a positive entry in the correlation matrix means that the corresponding variables are related such that if one increases (decreases) the other one also increases (decreases). Similarly a negative entry in the correlation matrix implies that the corresponding variables are related in such a way that if one increases (decreases) the other one decreases (increases).

4 Summary of Analysis

The main purpose of this project was to visually depict how each of the predictor variable affects the chance of a person having diabetes or not. With the help of ggplot in R, I have tried to show how the independent variable Outcome is related, if it is, to the predictor variables. I will try to note down my findings below.

- As age increases, the chance of a PIMA Indian female person having diabetes increases, but not uniformly, which means that only age is not responsible for a PIMA Indian female person having diabetes. There are other factors which weigh in more or less.
- We see when the glucose level and blood pressure of a person under consideration are somewhat correlated. As age increases, the chance that a person has higher levels of blood pressure or glucose level also somewhat increases.
- We see for a person having BMI in the range of perfect BMI, the chance of having diabetes decreases, whereas for a person having higher or lower BMI has a higher chance of having diabetes.

5 Conclusion

Throughout the course of this project, we have seen that there are various factors that dictate whether a person will have diabetes or not. None of the factors can be completely viable for the occurrence of diabetes in a person. But the combined effect of all the factors under consideration is what is responsible for diabetes. Although the list of factors, i.e. the predictor variables is not exhaustive, but we can say that the factors under condition are quite relevant. In later part of the project, through the use of RShiny dashboard, we will see more extensively how these factors affect the chances of a PIMA Indian female having diabetes.

Github Link of the repository : https://github.com/AbhishekDS1729/Visualization-Project-DS1

Dashboard Link : https://abhishekchakrabortyds.shinyapps.io/Diabetes/

Youtube Video of the dashboard : https://youtu.be/qwAuXONedGg