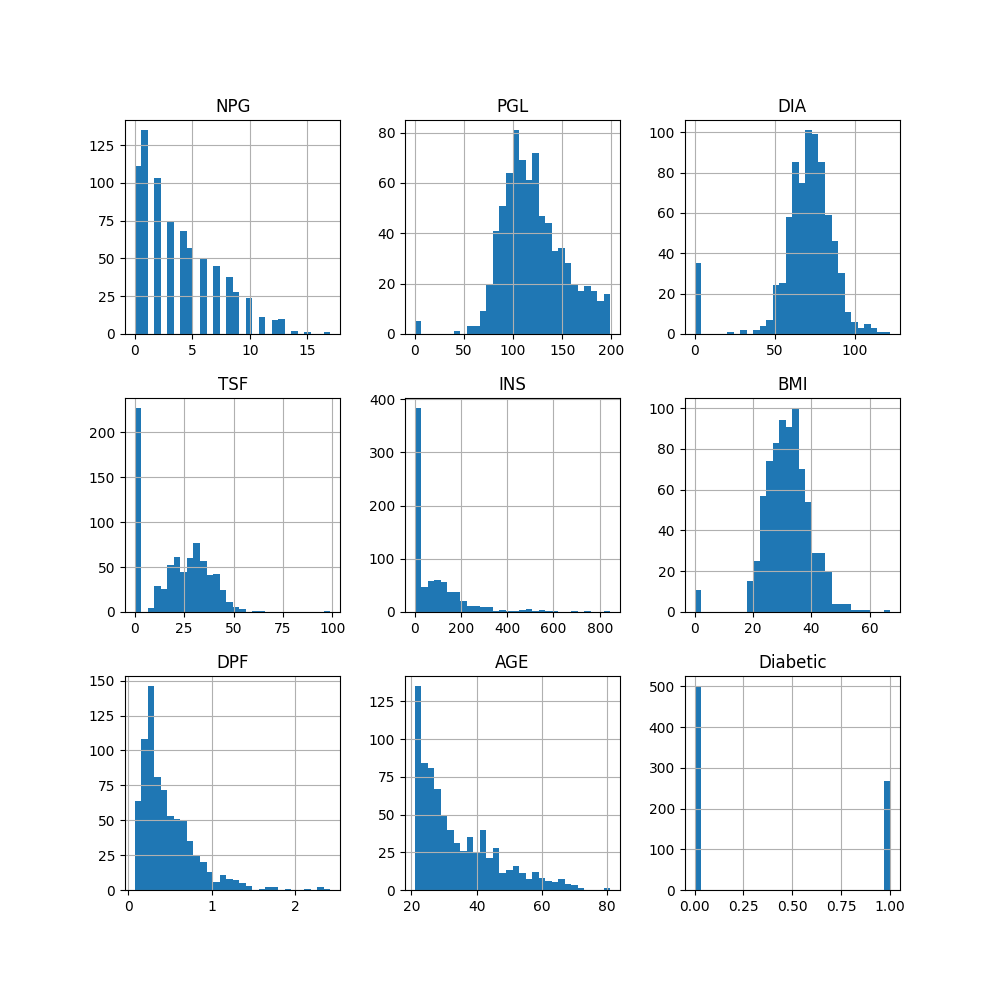
**Introduction**

In this project for COMP4388, we aim to build a machine learning model to predict diabetes. Using a specific dataset, we will analyze and identify key factors that contribute to diabetes. Our goal is to understand the relationship between these factors and the likelihood of diabetes, applying statistical and machine learning techniques. This project will help us gain insights into diabetes prediction and the effective use of data in healthcare.

**Data And Analysis**

Here are the summary statistics of all attributes in the diabetes dataset(before Cleaning ):

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute** | **Count** | **Mean** | **Std Dev** | **Min** | **25th Pct** | **Median** | **75th Pct** | **Max** |
| NPG | 768 | 3.85 | 3.37 | 0.00 | 1.00 | 3.00 | 6.00 | 17.00 |
| PGL | 768 | 120.89 | 31.97 | 0.00 | 99.00 | 117.00 | 140.25 | 199.00 |
| DIA | 768 | 69.11 | 19.36 | 0.00 | 62.00 | 72.00 | 80.00 | 122.00 |
| TSF | 768 | 20.54 | 15.95 | 0.00 | 0.00 | 23.00 | 32.00 | 99.00 |
| INS | 768 | 79.80 | 115.24 | 0.00 | 0.00 | 30.50 | 127.25 | 846.00 |
| BMI | 768 | 31.99 | 7.88 | 0.00 | 27.30 | 32.00 | 36.60 | 67.10 |
| DPF | 768 | 0.47 | 0.33 | 0.08 | 0.24 | 0.37 | 0.63 | 2.42 |
| AGE | 768 | 33.24 | 11.76 | 21.00 | 24.00 | 29.00 | 41.00 | 81.00 |
| Diabetic | 768 | 0.35 | 0.48 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |



**NPG (Number of Pregnancies):** Most women in the study have had few pregnancies. A small number have had many pregnancies.

**PGL (Plasma Glucose Level):** People's blood sugar levels are mostly around the middle range. Few people have very low or very high levels.

**DIA (Diastolic Blood Pressure):** Most people have a blood pressure that is not too low or too high, but right in a middle range.

**TSF (Triceps Skin Fold Thickness):** A lot of people have a thin skin fold on their arm, while very few have a thick one.

**INS (2-Hour Serum Insulin**): The insulin levels in most people's blood after two hours are quite low, with only a few having higher levels.

**BMI (Body Mass Index):** Most people have a BMI that suggests they are at a normal weight or a bit above. Some have a BMI that is much higher.

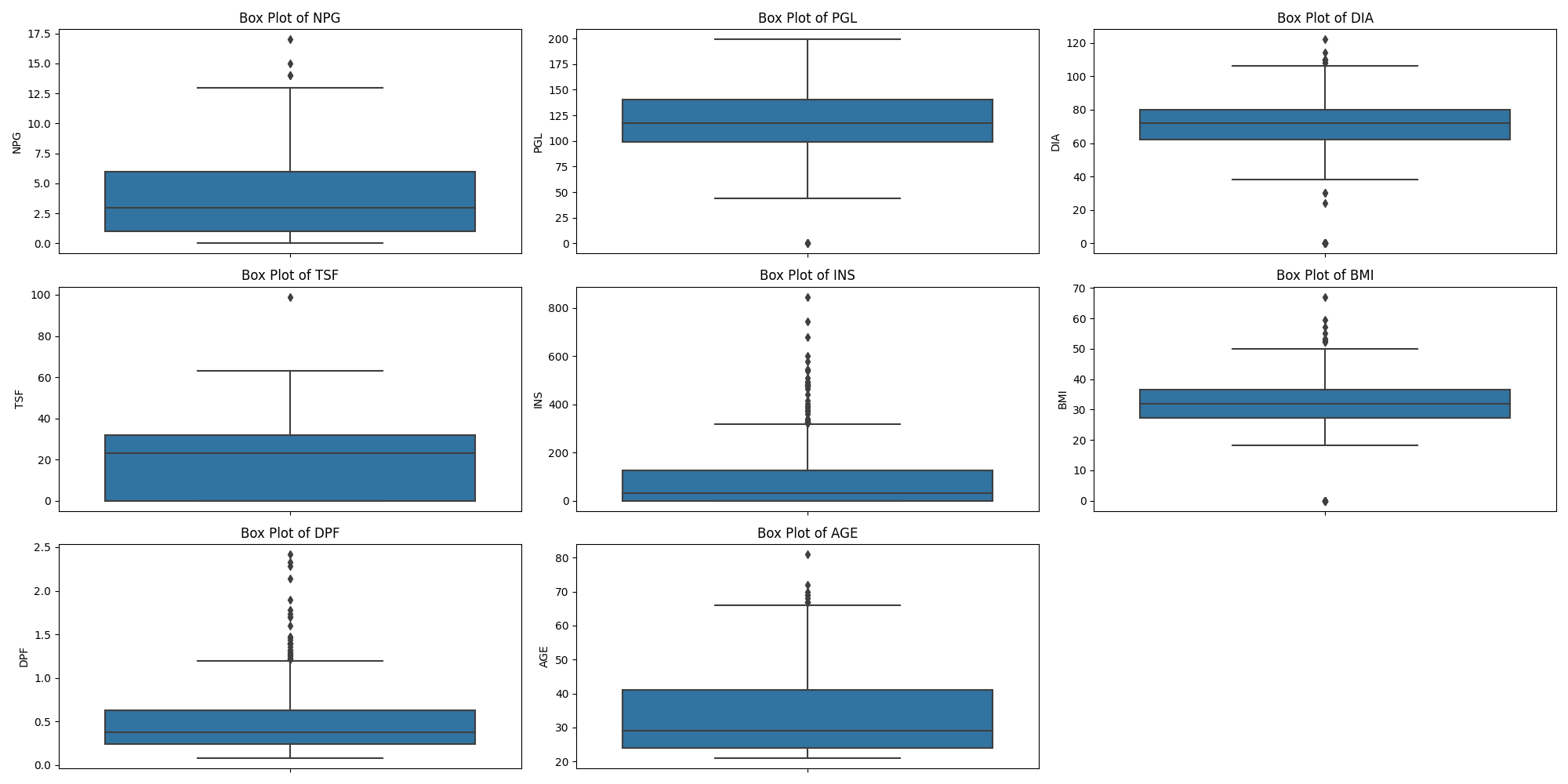
**DPF (Diabetes Pedigree Function):** Most people have a low score for diabetes based on their family history, with very few having a high score.

**AGE**: The people in the study are mostly younger, with fewer older individuals.

**Diabetic (Class Label):** There are more people without diabetes than with diabetes in this study.

For my diabetes dataset, we did not find any missing or duplicate records. However, there were biologically unrealistic zero values in key columns such as plasma glucose level, blood pressure, skin thickness, insulin, and BMI and there is outliers in my dataset . I am check it using a Box Plot approach to operate with outliers .

Therefore This data must be cleaned by replacing it with appropriate values to increase the strength of the models later.

The below images to illustrate the Outliers using Box Plot using Python :there fore , in the Next Page I am explained how to Handle the Noise in the dataset . (how to handle the zero’s that not realistic Only , and the outliers ) .

DataCleaning

Handling Zeros :

Some Rows contain zero values for variables where these values are biologically implausible, such as blood pressure and body mass index.

Plasma Glucose Level (PGL): 5 zeros

Diastolic Blood Pressure (DIA): 35 zeros

Triceps Skin Fold Thickness (TSF): 227 zeros

(INS): 374 zeros

Body Mass Index (BMI): 11 zeros

These zeros are likely placeholders for missing data. To address this, we will replace the zeros with appropriate values that best represent the underlying distribution of the data. One common method is to use the median or Mean of non-zero values within the same column.

provided that the number of zero entries is not so large that it does not significantly skew the distribution of the data

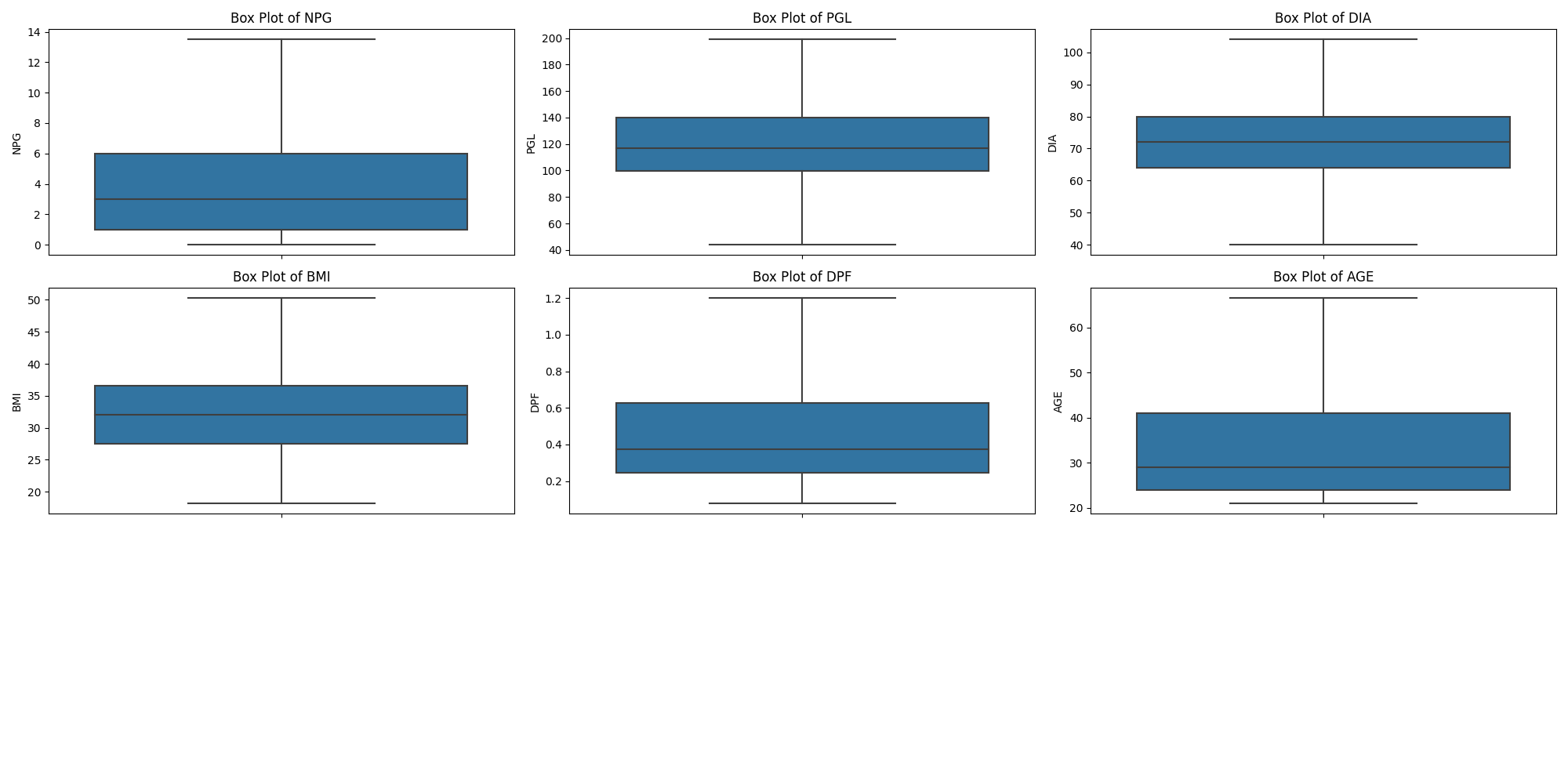
so I am **EXECLUDE** THE (**TNS** AND **INS**) COULMN’S

The rationale behind this decision is the excessive number of zero values present in these column . It is necessary to remove these variables to prevent skewed results and maintain the power of our predictive modelss.

For remining columns I am replace the Zeros by Median .

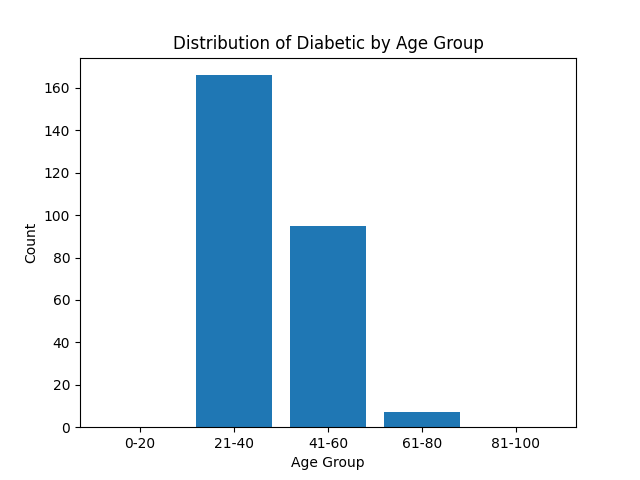
**Handling Outliers :**

cap\_outliers, designed to reduce outliers. It calculates the acceptable data range using the interquartile range (IQR) and sets extreme values accordingly. Values below Q1 - 1.5IQR or above Q3 + 1.5IQR are set to these limits. This method maintains the integrity of the data while minimizing the impact of outliers.

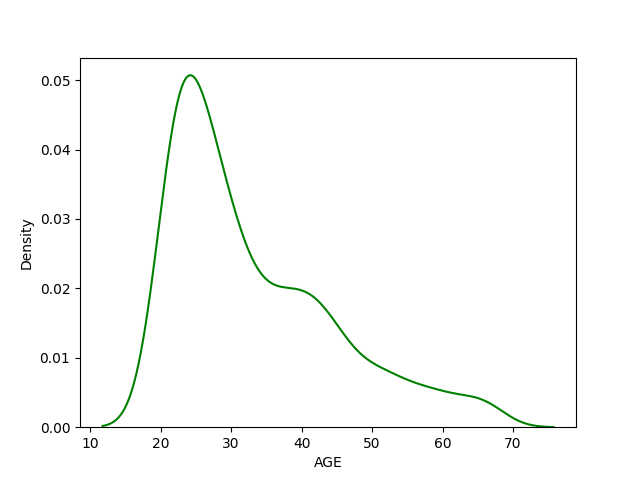


Histogram for Age Group :

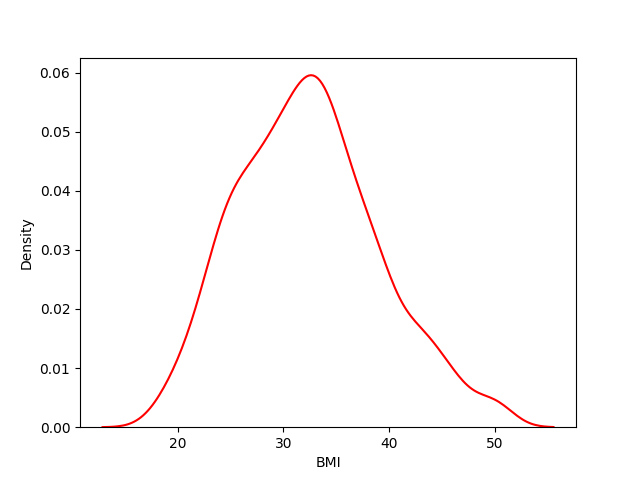
The chart shows the number of diabetic people in different age ranges. Most diabetics are between 21-40 years old, with fewer in older age groups.



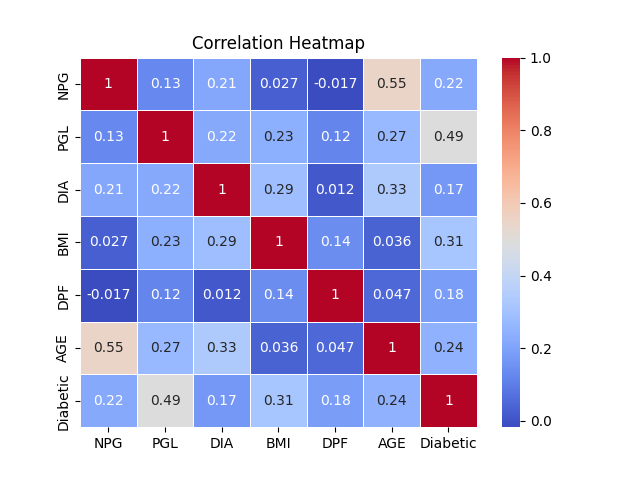
density plot for age, showing a peak in the younger age range and a gradual decline as age increases



density plot for BMI, with a peak around the middle range and tails off towards the lower and higher BMI values.



Visualize the correlation between all features



NPG (Number of Pregnancies) shows a notable positive correlation with age, which makes sense as the number of pregnancies can increase as women get older.

PGL (Plasma Glucose Level) has a moderately strong positive correlation with the diabetic outcome, indicating higher glucose levels could be associated with diabetes.

DIA (Diastolic Blood Pressure) and BMI (Body Mass Index) both show some positive correlation with the diabetic outcome, suggesting higher blood pressure and higher BMI could be associated with an increased risk of diabetes.

DPF (Diabetes Pedigree Function) and AGE show weaker correlations with other variables, but they do have some positive correlation with the diabetic outcome, indicating they might play a role in the risk of diabetes.

The correlation between AGE and NPG is the strongest in the dataset, which is expected as the likelihood of having more pregnancies increases with age.

Normalization of Data (Feature Scalling ) :

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

In this project I applied the Min-Max scale, which is to normalize the data. This method transforms features by scaling each one to a specific range, from 0 to 1. The formula used is (value - min) / (max - min).

Linear Regression Model’s

**Model 1 : Using All Independent Features :**

In the predictive model, we created Model 1 using all available independent features. We set aside 80% of our dataset for training purposes, while the remaining 20% was used to test the effectiveness of the model. Model parameters, represented by theta values and the intercept indicates the weight contributed by each The

**Parameters** : THETA’s [21.3827259 9.62138014 15.70274169 -8.54130898 0.90395886 1.30752398]

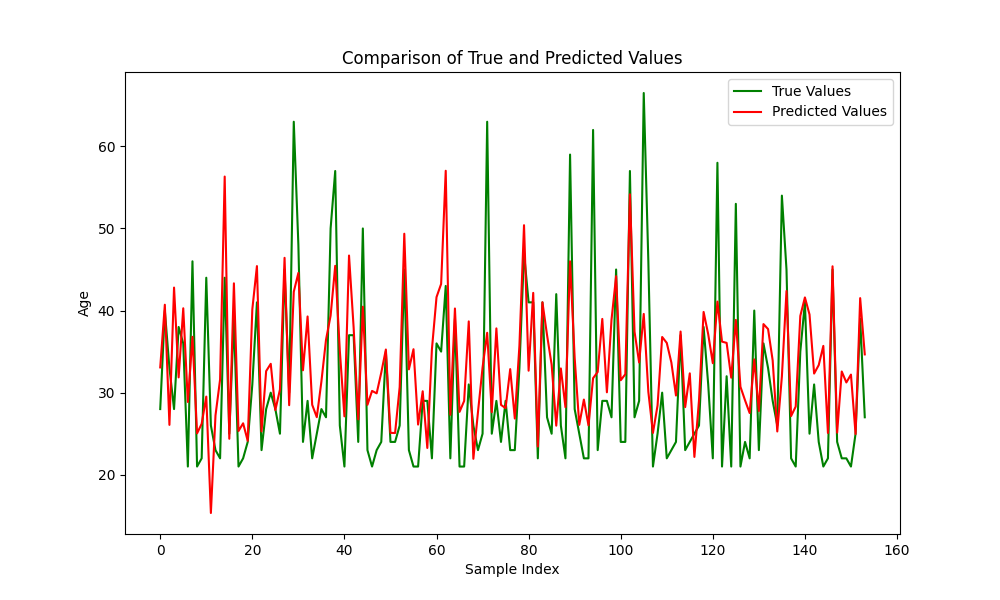
**Intercept** : 17.923492265140965

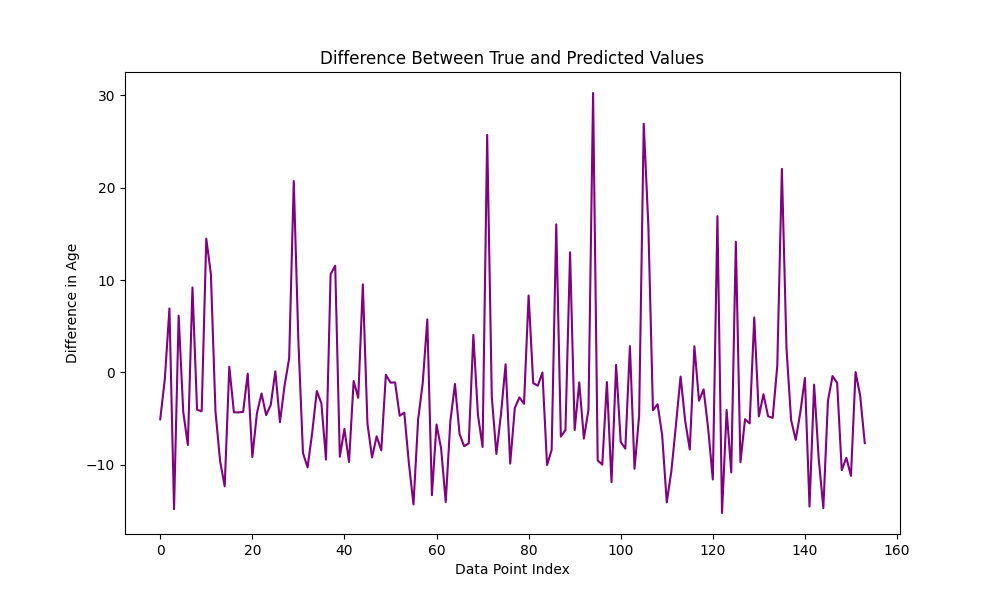
Resulting regression equation is:

H(x) = 21.3827259 X1 + 9.62138014 X2 + 15.70274169 X3 - 8.54130898 X4 + 0.90395886 X6 + 1.30752398 X7 + 17.923492265140965

**MSE : 75.73554269601364**

**RMSE : 8.70261700271899**

****

****

**Model2:**I used only the Number of Pregnancies (NPG) feature to predict age because it has the strongest association with age of 0.55, as shown in our data. This makes our model simpler and focuses on what matters most. By choosing the most relevant factor, our model is easier to understand and avoids using too much unnecessary information, which can make models too complex or less accurate.

**Results Model 2 :**

**Parameter :** 1.88129217

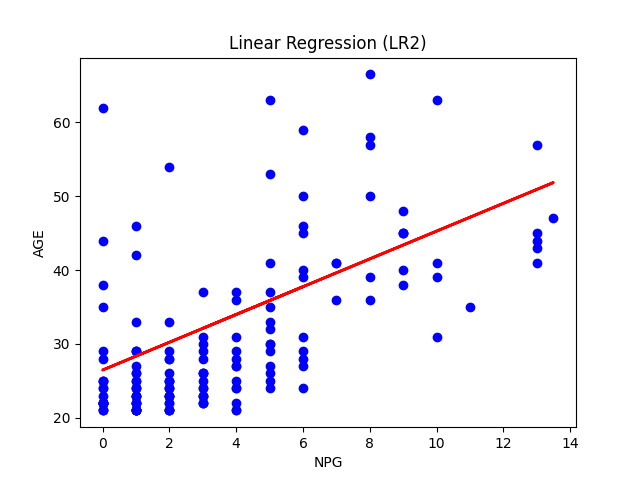
**Intercept** : 26.462249841681263

The Hypothesis Equation : f(x) = 1.88129217\*X1 + 26.462249841681263

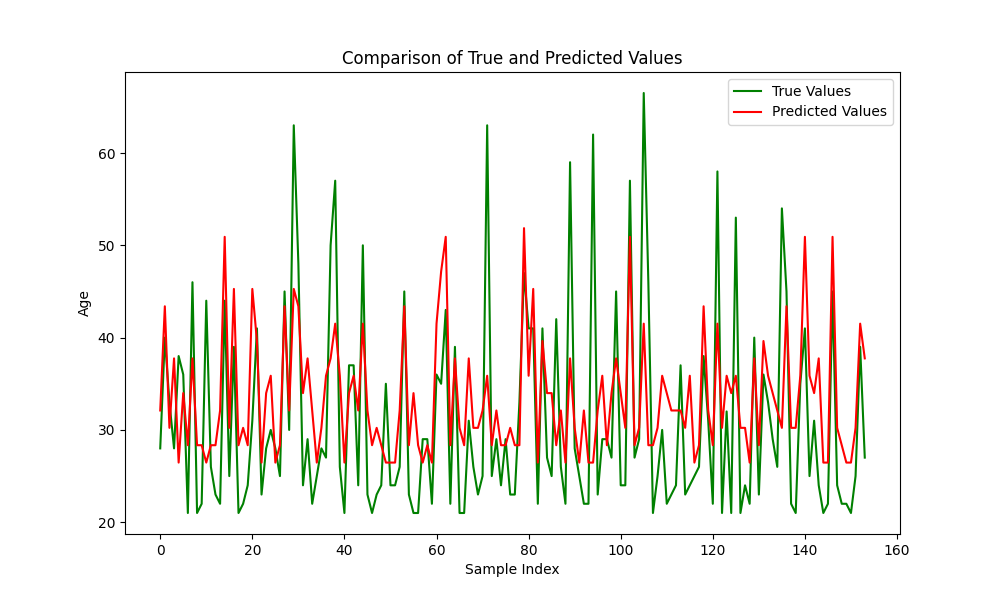
MSE : 77.17033765611646

RMSE : 8.784664914276267

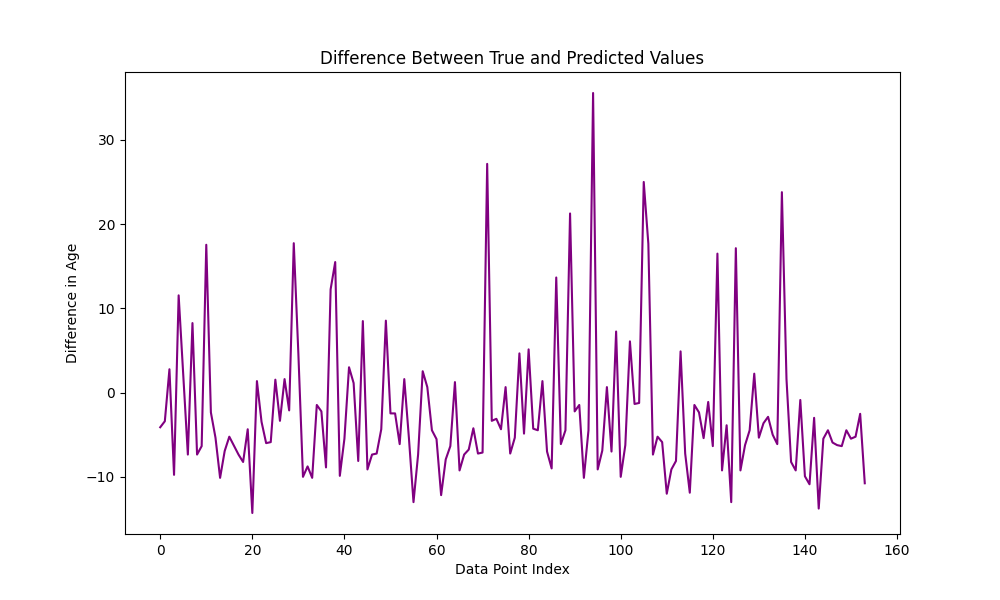
The red line indicates the best fitting line from the linear regression model (LR2), showing the expected relationship between NPG and age. The scatter plot indicates an increasing trend, meaning that as the number of pregnancies increases, age tends to increase as well, which is consistent with the positive association we observed previously.



Comparing the actual and predicted age values from the linear regression model (LR2). The green line represents the true age of individuals in the data set, and the red line represents the ages predicted by the model based on the number of pregnancies (NPG). This comparison helps in visually assessing the accuracy of the model's predictions against the actual data.



Explain the differences between actual and expected age values. This graph, often referred to as a residual plot, shows the discrepancies between model predictions and real-world data. Points above the zero line on the y-axis indicate where the model underestimated age, and points below the zero line indicate where the model overestimated. The closer these points are to the zero line, the more accurate the model's predictions are.



**Model 3 :**   
  
We used “number of pregnancies” (NPG), “plasma glucose level” (PGL), and “diastolic blood pressure” (DIA) as predictors because they show the strongest relationships with diabetes. These three factors give us a clearer picture of diabetes risk, making our model simple and effective.

**The Parameters** : [22.35762182 9.59770578 12.96864066]

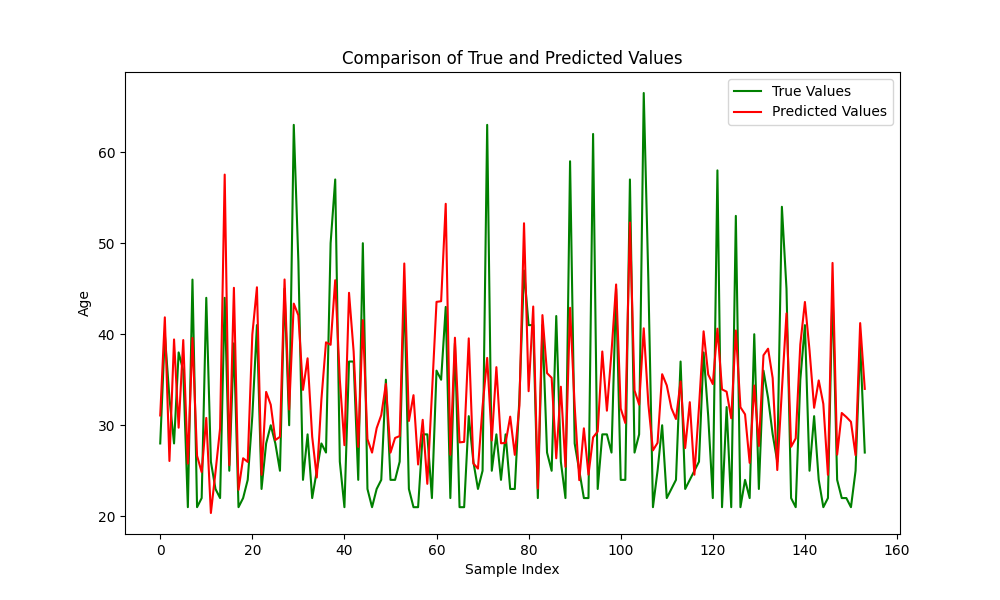
**Intercept** : 15.97927637823209

Hypothesis : 22.35762182 \*X1 + 9.59770578 \*X2 + 12.96864066 \*X3 + 15.97927637823209

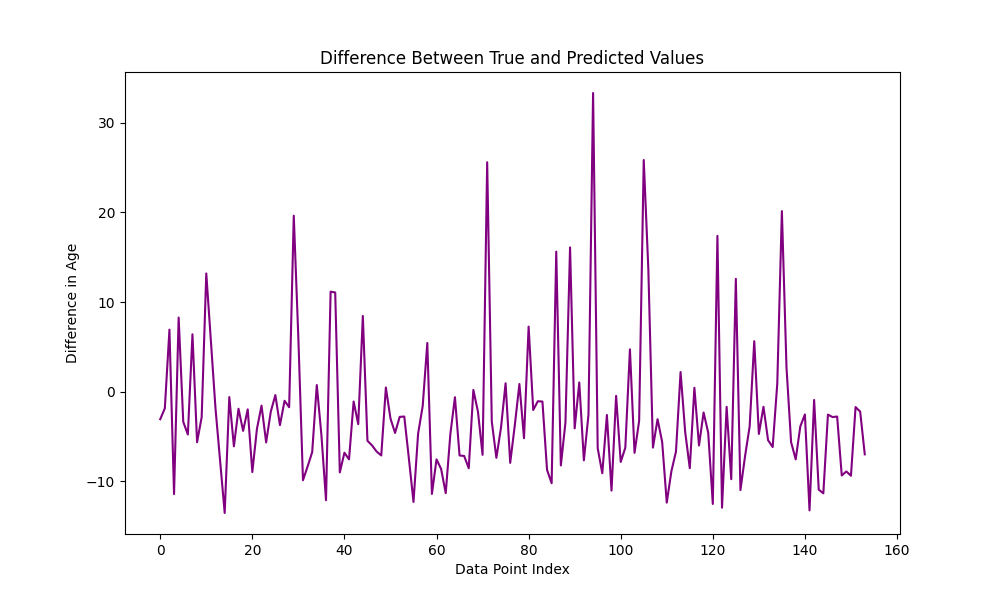
MSE : 68.95367309551023

RMSE : 8.303834842740445

A comparison between the actual ages and the ages predicted by our model, as both follow a similar pattern across the dataset.



Differences between actual and predicted ages, highlighting variation in model accuracy for each data point.



Compare the performance for Model’s

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **RMSE** |
| Model 1 | 75.735543 | 8.702617 |
| Model 2 | 77.170338 | 8.784665 |
| Model 3 | 68.953673 | 8.303835 |

I compared three different models to see which one predicted age best. Model 1 used all the information but   
It reached a Mean Squared Error (MSE) of 75.74 and a Root Mean Squared Error (RMSE) of 8.70.  
was probably confused by some of it.

Model 2 It got an MSE of 77.17 and an RMSE of 8.78, which are a bit higher than Model 1 This suggests that while 'NPG' is important, it's not enough on its own to predict someone's age accurately

She chose Model 3 It performs the best with an MSE of 68.95 and an RMSE of 8.30

The lower these values are, the closer the model's predictions are to the real ages