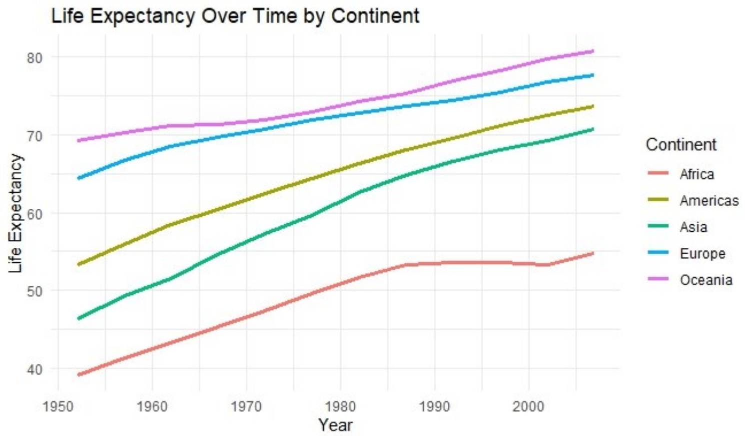
- **Problem 1: How has life expectancy changed over time in different continents?**

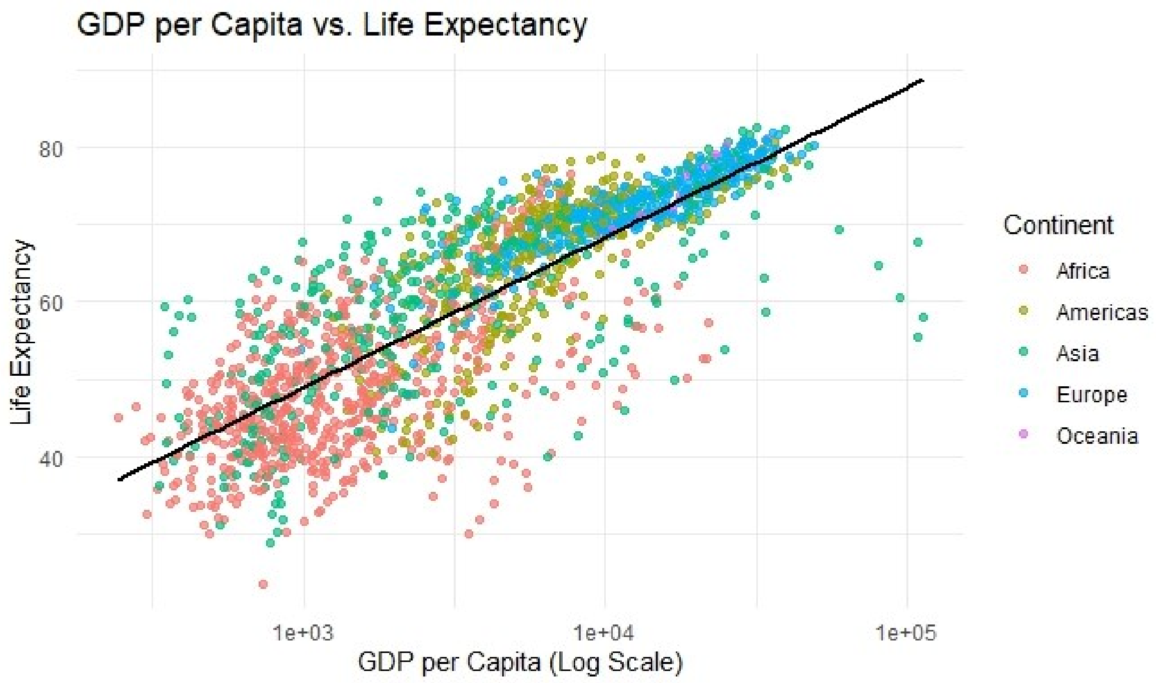
```
install.packages("gapminder")

library(gapminder)

data<-gapminder

# Load necessary libraries

library(ggplot2)

library(dplyr)

# Plot life expectancy over time by continent

ggplot(gapminder, aes(x = year, y = lifeExp, color = continent, group = continent)) +

  geom_line(stat = "summary", fun = "mean", size = 1.2) +

  labs(title = "Life Expectancy Over Time by Continent",

       x = "Year",

       y = "Life Expectancy",

       color = "Continent") + theme_minimal()
```



- This plot will show how life expectancy has increased across different continents. Typically, we might observe more rapid increases in Asia and Africa in recent decades. But slow increase shows for Europe and Americas continent.

- **Problem-2: Is there a relationship between GDP per capita and life expectancy?**

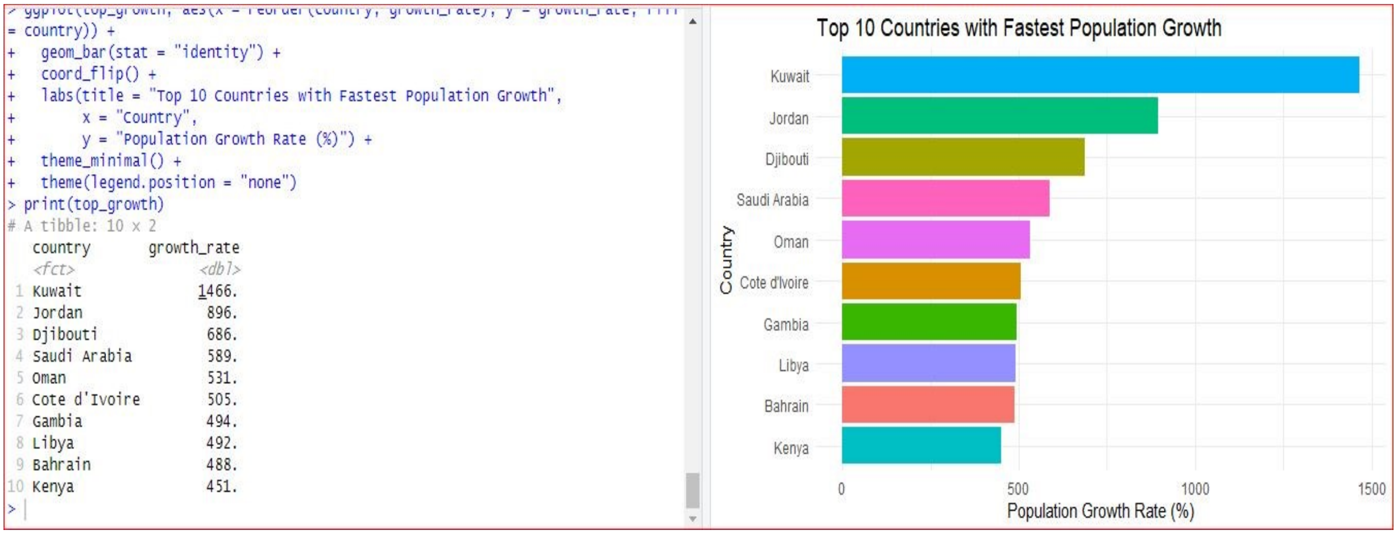
```
ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, color = continent)) +
  geom_point(alpha = 0.7) +
  scale_x_log10() +  # Log scale for better visualization
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "GDP per Capita vs. Life Expectancy",
       x = "GDP per Capita (Log Scale)",
       y = "Life Expectancy",
       color = "Continent") + theme_minimal()
```



This scatter plot shows the positive correlation between GDP per capita and Life expectancy trend.

Problem-03: Which countries have experienced the fastest population growth?

```
pop_growth <- gapminder %>%

                    group_by(country) %>%

                    arrange(year) %>%

                     summarise(growth_rate = (last(pop) - first(pop)) / first(pop) * 100) %>%

                    arrange(desc(growth_rate))

# Top 10 countries with fastest population growth

top_growth <- head(pop_growth, 10)

print(top_growth)

# Bar plot of the top 10 countries with the fastest population growth

ggplot(top_growth, aes(x = reorder(country, growth_rate), y = growth_rate, fill = country)) +

                 geom_bar(stat = "identity") + coord_flip() + labs(title = "Top 10 Countries with
                 Fastest Population Growth",x = "Country",

                  y = "Population Growth Rate (%)") +theme_minimal() +theme(legend.position =
                 "none")
```



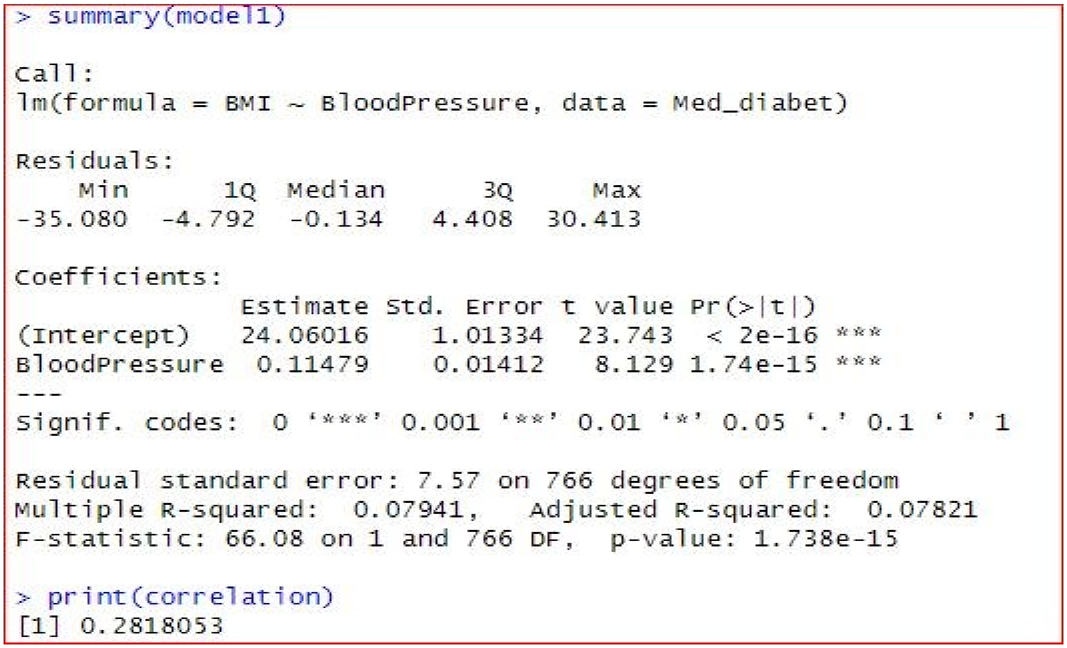
```
> ggplot(top_growth, aes(x = reorder(country, growth_rate), y = growth_rate, fill
= country)) +
+   geom_bar(stat = "identity") +
+   coord_flip() +
+   labs(title = "Top 10 Countries with Fastest Population Growth",
+        x = "Country",
+        y = "Population Growth Rate (%)") +
+   theme_minimal() +
+   theme(legend.position = "none")
> print(top_growth)
# A tibble: 10 x 2
   country       growth_rate
   <fct>               <dbl>
 1 Kuwait              1466.
 2 Jordan               896.
 3 Djibouti             686.
 4 Saudi Arabia         589.
 5 Oman                 531.
 6 Cote d'Ivoire        505.
 7 Gambia               494.
 8 Libya                492.
 9 Bahrain              488.
10 Kenya                451.
> 
```

Kuwait is the fastest population growth country all over the continent

Assignment-03

Problem 1: Examine How Blood Pressure Influences BMI

```
# Calculate correlation

correlation <- cor(Med_diabet$BloodPressure, Med_diabet$BMI, use="complete.obs")

#Visualize the relationship using a scatter plot.

ggplot(Med_diabet, aes(x=BloodPressure, y=BMI)) +

  geom_point(color='blue') +

  geom_smooth(method='lm', color='red') +

  labs(title="Blood Pressure vs BMI", x="Blood Pressure (mm Hg)", y="BMI")

#Examine the influence of Blood Pressure on BMI using a linear regression model.

model1 <- lm(BMI ~ BloodPressure, data=Med_diabet)

summary(model1)

print(correlation)
```



```
> summary(model1)

Call:
lm(formula = BMI ~ BloodPressure, data = Med_diabet)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.080  -4.792  -0.134   4.408  30.413 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   24.06016    1.01334  23.743  < 2e-16 ***
BloodPressure  0.11479    0.01412   8.129 1.74e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.57 on 766 degrees of freedom
Multiple R-squared:  0.07941,	Adjusted R-squared:  0.07821 
F-statistic: 66.08 on 1 and 766 DF,  p-value: 1.738e-15

> print(correlation)
[1] 0.2818053
```

figure-01

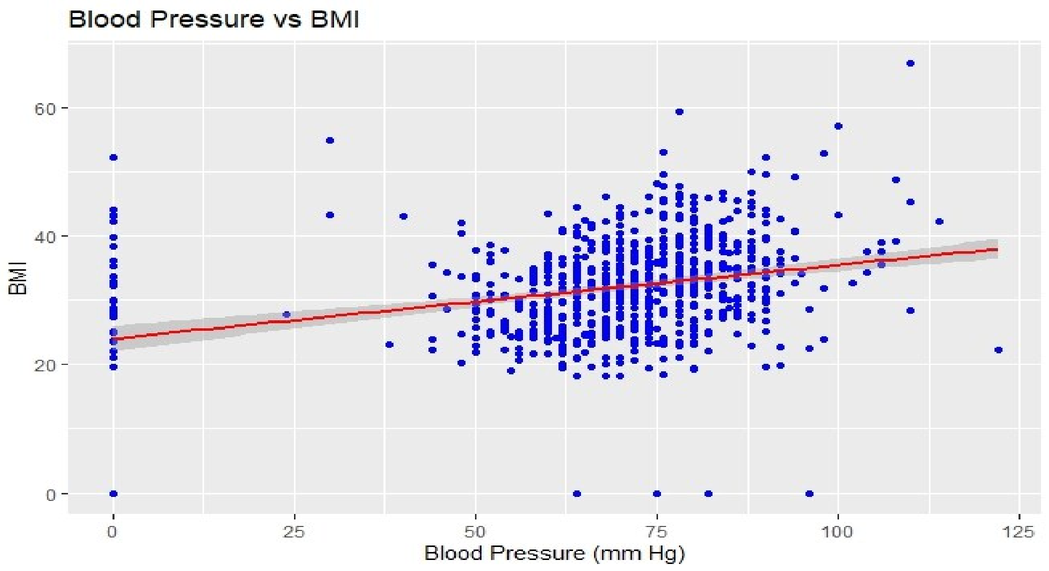


figure-02

My Interpretation:

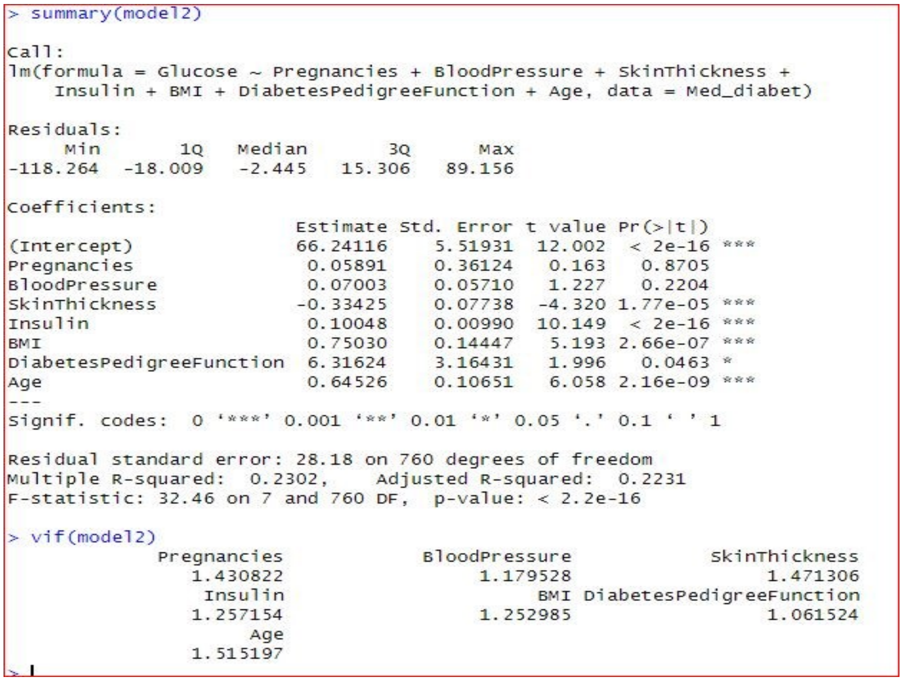
From figure-1, The multiple R Square indicates that **7.941% of the variability** in BMI is explained by BloodPressure. the low R-squared and weak correlation suggest that other factors influence BMI more than BloodPressure. **F-statistic:** 66.08 is the overall significance of the model. And **p-value:** 1.738e-15 Indicates the model is statistically significant overall, i.e., BloodPressure is a significant predictor of BMI.

The correlation between BloodPressure and BMI is 0.2818, which indicates a weak positive relationship.the relationship is positive means BloodPressure increases, BMI tends to increase slightly.

Problem 2: Investigate How Other Variables Affect Glucose Levels

```
library (corrplot)

#Correlation Matrix

variables <- Med_diabet[, c("Glucose", "Pregnancies", "BloodPressure", "SkinThickness",

                "Insulin", "BMI", "DiabetesPedigreeFunction", "Age")]

cor_matrix <- cor(variables, use="complete.obs")
```

```
corrplot(cor_matrix, method="circle")

#Build a multiple linear regression model to analyze how other variables affect Glucose levels.

library(car)

model2 <- lm(Glucose ~ Pregnancies + BloodPressure + SkinThickness +

             Insulin + BMI + DiabetesPedigreeFunction + Age, data= Med_diabet)

summary(model2)

# Check multicollinearity using VIF (Variance Inflation Factor)

vif(model2)
```



```
> summary(model2)

Call:
lm(formula = Glucose ~ Pregnancies + BloodPressure + SkinThickness + 
    Insulin + BMI + DiabetesPedigreeFunction + Age, data = Med_diabet)

Residuals:
     Min       1Q   Median       3Q      Max 
-118.264  -18.009   -2.445   15.306   89.156 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              66.24116    5.51931  12.002  < 2e-16 ***
Pregnancies               0.05891    0.36124   0.163   0.8705    
BloodPressure             0.07003    0.05710   1.227   0.2204    
SkinThickness            -0.33425    0.07738  -4.320 1.77e-05 ***
Insulin                   0.10048    0.00990  10.149  < 2e-16 ***
BMI                       0.75030    0.14447   5.193 2.66e-07 ***
DiabetesPedigreeFunction  6.31624    3.16431   1.996   0.0463 *  
Age                       0.64526    0.10651   6.058 2.16e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.18 on 760 degrees of freedom
Multiple R-squared:  0.2302,    Adjusted R-squared:  0.2231 
F-statistic: 32.46 on 7 and 760 DF,  p-value: < 2.2e-16

> vif(model2)
             Pregnancies            BloodPressure            SkinThickness 
                1.430822                 1.179528                 1.471306 
                 Insulin                      BMI DiabetesPedigreeFunction 
                1.257154                 1.252985                 1.061524 
                     Age 
                1.515197 
> 
```

figure-03

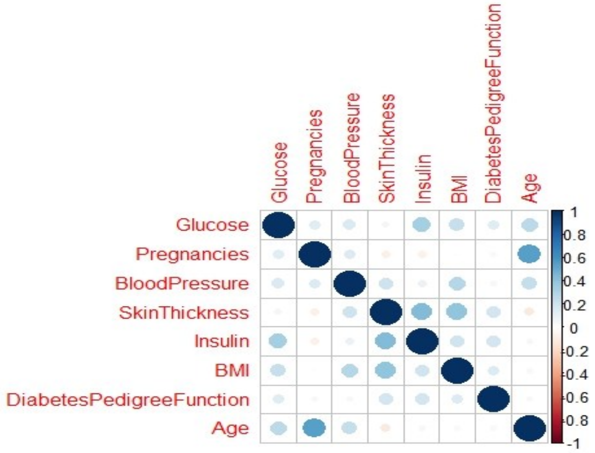


Figure-04

My Interpretation:

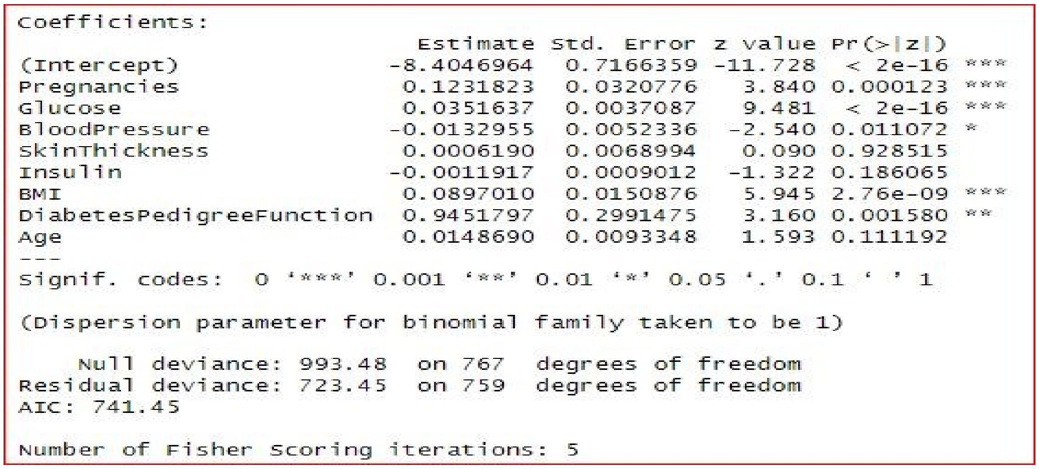
From the above Figure-03, The VIF(Variance Inflation Factor) values are all below 2, indicating no significant multicollinearity. VIF values higher than 5-10 would indicate that some predictors are highly correlated. there is no multicollinearity relationship, so all variables are save.

from this p values, we can see that there is no significant effect on Glucose of Pregnancies and Bloodpressure.and adjusted R square is 22.31% which is not good.To increase the R square value ,need to develop feature engineering .

Problem 3: Identify the Most Important Variable Influencing Diabetes Risk

```
#Logistic Regression (Identify the significant predictors for Diabetes risk.)

library(randomForest)

model3 <- glm(Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +

          Insulin + BMI + DiabetesPedigreeFunction + Age,

         data=Med_diabet, family=binomial)

summary(model3)

# Feature Importance using Random Forest (To determine the most important variable, use a Random Forest model.)

set.seed(123)

# Build Random Forest model
```

```
rf_model <- randomForest(Outcome ~ Pregnancies + Glucose + BloodPressure +
                         SkinThickness + Insulin + BMI +
                         DiabetesPedigreeFunction + Age,
                        data=Med_diabet, importance=TRUE)
#importance=TRUE-understand which features contribute the most to predictions
varImpPlot(rf_model)
importance(rf_model)
```



```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -8.4046964  0.7166359 -11.728  < 2e-16 ***
Pregnancies              0.1231823  0.0320776   3.840 0.000123 ***
Glucose                  0.0351637  0.0037087   9.481  < 2e-16 ***
BloodPressure           -0.0132955  0.0052336  -2.540 0.011072 *  
SkinThickness            0.0006190  0.0068994   0.090 0.928515    
Insulin                 -0.0011917  0.0009012  -1.322 0.186065    
BMI                      0.0897010  0.0150876   5.945 2.76e-09 ***
DiabetesPedigreeFunction 0.9451797  0.2991475   3.160 0.001580 ** 
Age                      0.0148690  0.0093348   1.593 0.111192    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 759  degrees of freedom
AIC: 741.45

Number of Fisher Scoring iterations: 5
```

figure-5

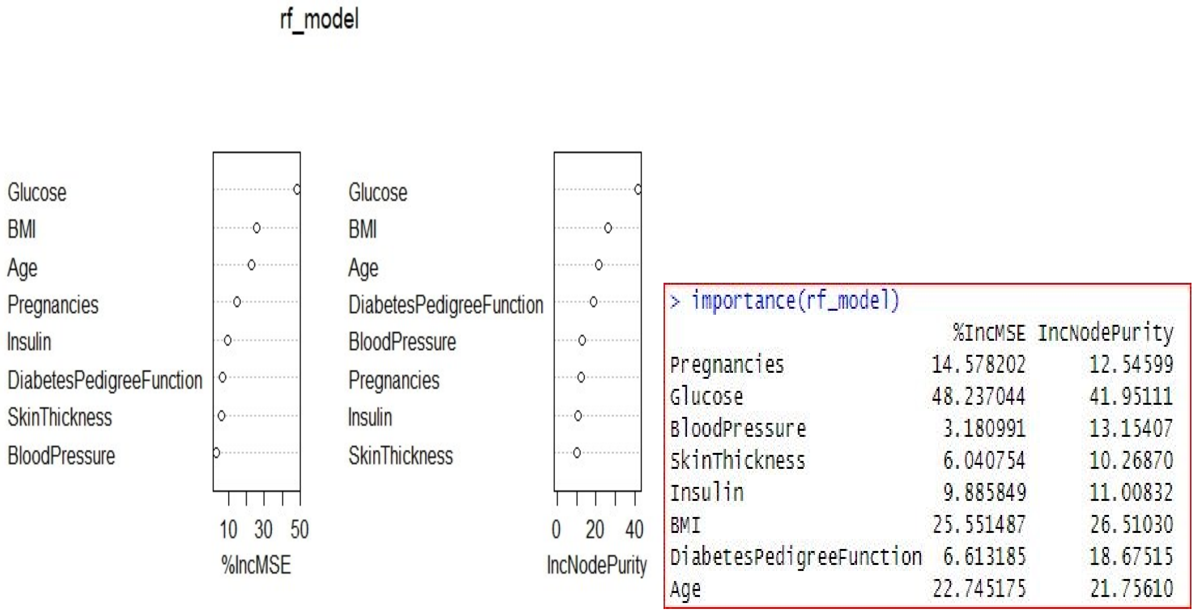
My Interpretation:

From the figure-5,

The logistic regression model suggests that variables like **Pregnancies**, **Glucose**, **BMI**, **DiabetesPedigreeFunction**, and **BloodPressure** has statistically significant effects on the outcome variable because these has the P-value less than 0.05.

on the Otherhand, The Variables like **SkinThickness**, **Insulin**, and **Age** are **not statistically significant** in this model because this is greater than 0.05.

The model perform well fitted because the **residual deviance** is significantly lower than the **null deviance**, it means your model with predictors is doing a much better job of fitting. And Lower AIC(Better fit of model) depends on the Lower **residual deviance.**



```
> importance(rf_model)
                            %IncMSE IncNodePurity
Pregnancies               14.578202      12.54599
Glucose                   48.237044      41.95111
BloodPressure              3.180991      13.15407
SkinThickness              6.040754      10.26870
Insulin                    9.885849      11.00832
BMI                       25.551487      26.51030
DiabetesPedigreeFunction   6.613185      18.67515
Age                       22.745175      21.75610
```

figure-06

From the figure-06:

Glucose and BMI are the most important features because they cause the largest increase in MSE when permuted. And, BloodPressure and SkinThickness show the least importance since permuting them minimally affects the model's accuracy.

On the other side, Glucose, BMI, and Age have the highest IncNodePurity, confirming their significant role in classifying the Outcome Variable. And, SkinThickness and Insulin are less influential, contributing less to node purity.