

# Diabetes:

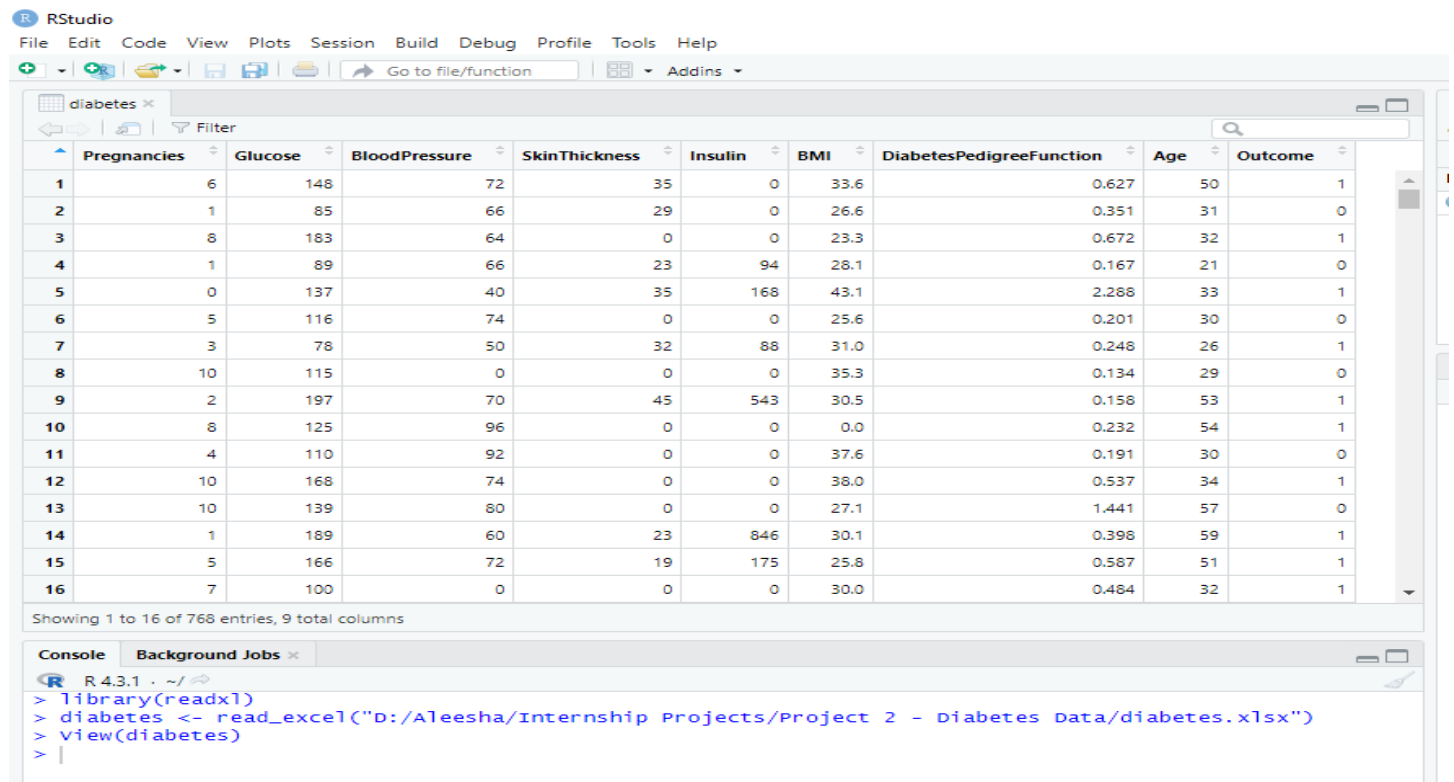
Diabetes is a serious health problem affecting many people worldwide. As a data analyst, I'm here to explore the data about diabetes. We'll use this information to better understand the disease, its causes, and how to improve care for those who have it. Let's dive into the numbers and see what insights we can uncover.

## About Dataset:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.<sup>2</sup> From the data set in the (.csv) File We can find several variables, some of them are independent (several medical predictor variables) and only one target dependent variable (Outcome).

## Import CSV File into R Studio:

Importing a diabetes CSV file into R Studio is a fundamental step in data analysis, enabling us to explore and visualize data, gain insights, and make informed decisions.



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

diabetes x

|    | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI  | DiabetesPedigreeFunction | Age | Outcome |
|----|-------------|---------|---------------|---------------|---------|------|--------------------------|-----|---------|
| 1  | 6           | 148     | 72            | 35            | 0       | 33.6 | 0.627                    | 50  | 1       |
| 2  | 1           | 85      | 66            | 29            | 0       | 26.6 | 0.351                    | 31  | 0       |
| 3  | 8           | 183     | 64            | 0             | 0       | 23.3 | 0.672                    | 32  | 1       |
| 4  | 1           | 89      | 66            | 23            | 94      | 28.1 | 0.167                    | 21  | 0       |
| 5  | 0           | 137     | 40            | 35            | 168     | 43.1 | 2.288                    | 33  | 1       |
| 6  | 5           | 116     | 74            | 0             | 0       | 25.6 | 0.201                    | 30  | 0       |
| 7  | 3           | 78      | 50            | 32            | 88      | 31.0 | 0.248                    | 26  | 1       |
| 8  | 10          | 115     | 0             | 0             | 0       | 35.3 | 0.134                    | 29  | 0       |
| 9  | 2           | 197     | 70            | 45            | 543     | 30.5 | 0.158                    | 53  | 1       |
| 10 | 8           | 125     | 96            | 0             | 0       | 0.0  | 0.232                    | 54  | 1       |
| 11 | 4           | 110     | 92            | 0             | 0       | 37.6 | 0.191                    | 30  | 0       |
| 12 | 10          | 168     | 74            | 0             | 0       | 38.0 | 0.537                    | 34  | 1       |
| 13 | 10          | 139     | 80            | 0             | 0       | 27.1 | 1.441                    | 57  | 0       |
| 14 | 1           | 189     | 60            | 23            | 846     | 30.1 | 0.398                    | 59  | 1       |
| 15 | 5           | 166     | 72            | 19            | 175     | 25.8 | 0.587                    | 51  | 1       |
| 16 | 7           | 100     | 0             | 0             | 0       | 30.0 | 0.484                    | 32  | 1       |

Showing 1 to 16 of 768 entries, 9 total columns

Console Background Jobs x

```
R 4.3.1 . ~/
> library(readxl)
> diabetes <- read_excel("D:/Aleesha/Internship Projects/Project 2 - Diabetes Data/diabetes.xlsx")
> view(diabetes)
> |
```

# Data Exploration and Validation:

Before diving into analysis, it's a good practice to check a few lines of the imported data to ensure it matches your expectations.

```
Console Background Jobs x
R 4.3.1 ~ /
> d
# A tibble: 768 x 9
  Pregnancies Glucose BloodPressure skinThickness Insulin BMI DiabetesPedigreeFunction Age
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1      6     148      72      35      0    33.6      0.627    50
2      1      85      66      29      0    26.6      0.351    31
3      8     183      64      0      0    23.3      0.672    32
4      1      89      66      23     94    28.1      0.167    21
5      0     137      40      35    168    43.1      2.29     33
6      5     116      74      0      0    25.6      0.201    30
7      3      78      50      32     88    31      0.248    26
8     10     115       0      0      0    35.3      0.134    29
9      2     197      70      45    543    30.5      0.158    53
10     8     125      96      0      0      0      0.232    54
# i 758 more rows
# i 1 more variable: Outcome <dbl>
# i use `print(n = ...)` to see more rows
```

To confirm whether the data was imported correctly and if any lines are missing, you can check the number of rows and columns in the table using the "dim()" function in R Studio.

```
> dim(d)
[1] 768 9
```

To understand the data's type and structure, you can use the "str()" function in R Studio, which provides a summary of the data frame's attributes and variable types.

```
> str(d)
tibble [768 x 9] (s3: tbl_df/tbl/data.frame)
 $ Pregnancies      : num [1:768] 6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose          : num [1:768] 148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure    : num [1:768] 72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness    : num [1:768] 35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin          : num [1:768] 0 0 0 94 168 0 88 0 543 0 ...
 $ BMI              : num [1:768] 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num [1:768] 0.627 0.351 0.672 0.167 2.288 ...
 $ Age              : num [1:768] 50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome          : num [1:768] 1 0 1 0 1 0 1 0 1 1 ...
```

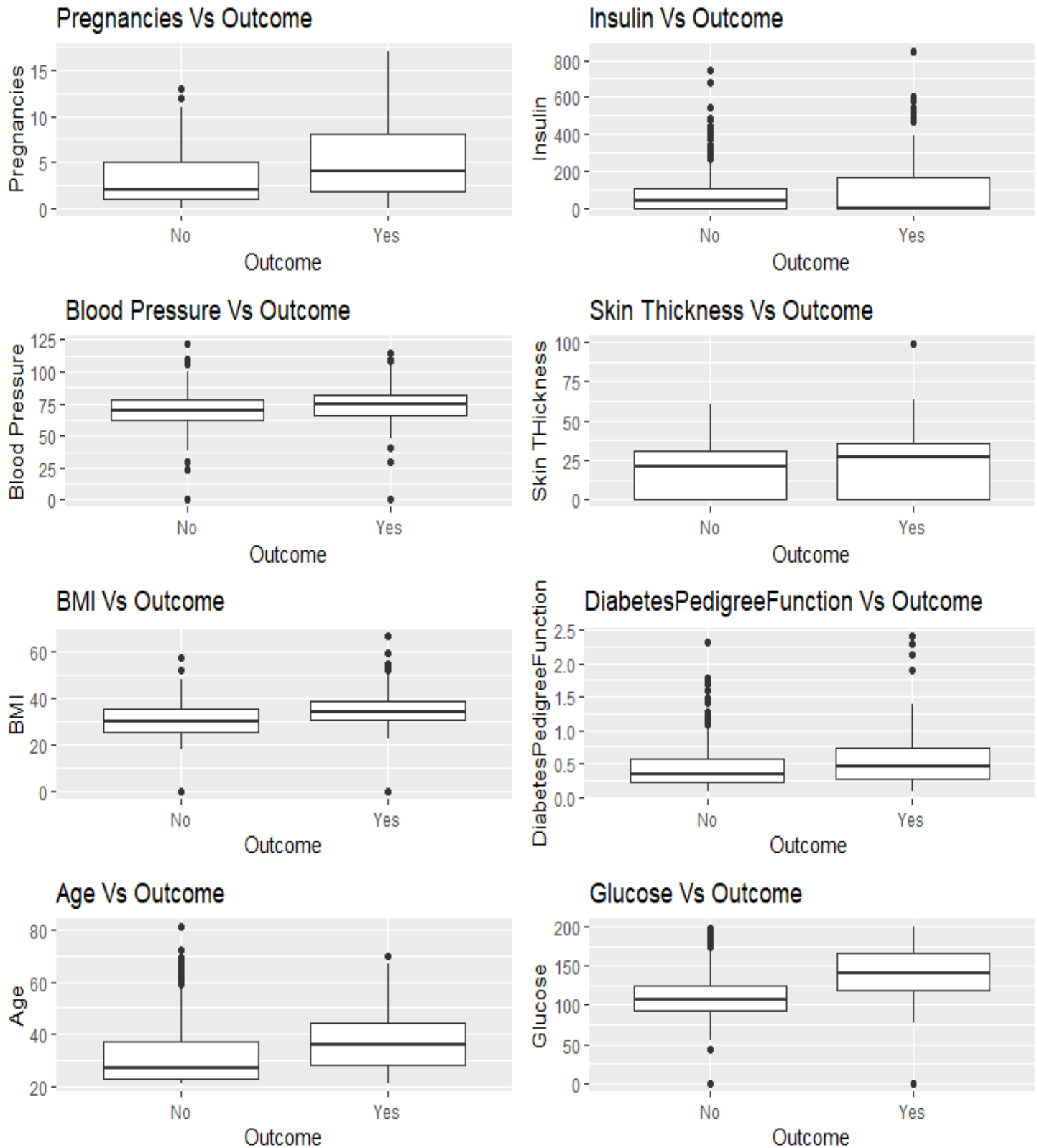
To maintain data integrity, it's crucial to check for the presence of null values in columns, ensuring all data points are complete and reliable.

```
> #check the null values of each column
> null_counts<-colSums(is.na(data))
> null_counts
      Pregnancies      Glucose      BloodPressure      skinThickness      Insulin
      0            0            0            0            0
      BMI DiabetesPedigreeFunction      Age      Outcome      age_cat
      0            0            0            0            0
```

With our data exploration and validation complete, it's time to proceed to the next step of data cleaning.

## Data Cleaning:

Now, let's proceed to check for outliers in the data, as outliers are data points that deviate significantly from the majority of the data and can impact the accuracy of our analysis.

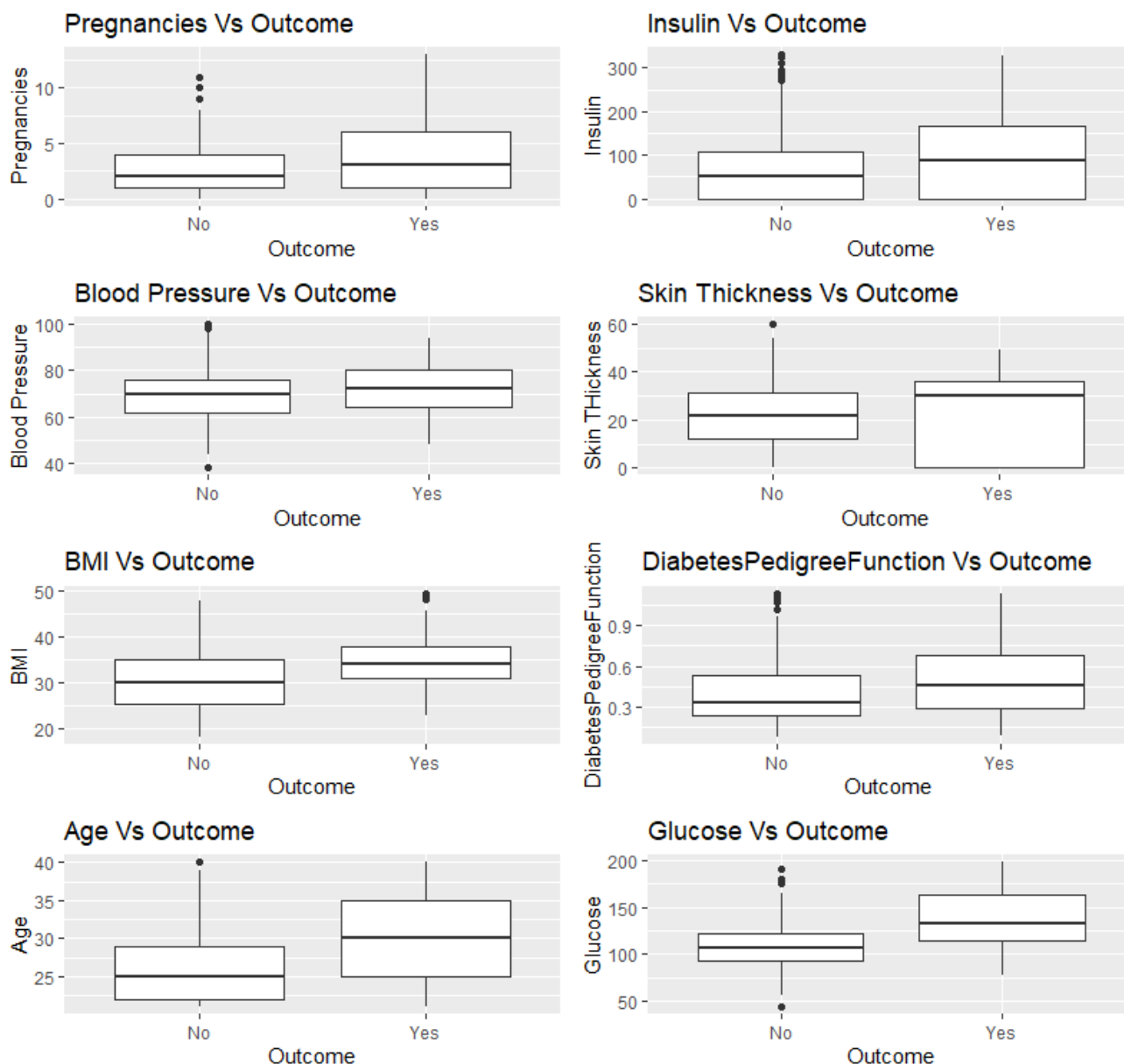


In the preceding figure, it's evident that there is a significant presence of outliers in variables such as Insulin, BMI, Blood Pressure, and DiabetesPedigreeFunction. To ensure the integrity of our analysis, it is essential that we proceed to remove these outliers from the dataset and store the cleaned data in another variable, which we will name **"data."**

```
# Define a function to remove outliers based on IQR for all columns
remove_outliers_all <- function(dataframe) {
  for (col in names(dataframe)) {
    Q1 <- quantile(dataframe[[col]], 0.25)
    Q3 <- quantile(dataframe[[col]], 0.75)
    IQR <- Q3 - Q1
    lower_bound <- Q1 - 1.5 * IQR
    upper_bound <- Q3 + 1.5 * IQR
    dataframe <- dataframe[!(dataframe[[col]] < lower_bound | dataframe[[col]] > upper_bound), ]
  }
  return(dataframe)
}

# Apply the function to remove outliers from all columns
data <- remove_outliers_all(d)
```

After removing outliers, it is imperative to recheck the data for any remaining outliers to confirm the effectiveness of the cleaning process.



Before removing outliers, the dataset contains **786 rows** and **9 columns**. After the outlier removal process, which we've stored in a new variable named **"data"**, we are left with **485 rows** and **9 columns**, highlighting the impact of this data cleaning step on our dataset's size and integrity.

```
> dim(data)
[1] 485 9
> data[1:5,]
# A tibble: 5 x 9
  Pregnancies Glucose BloodPressure SkinThickness Insulin BMI DiabetesPedigreeFunction Age Outcome
  <dbl>      <dbl>      <dbl>      <dbl>      <dbl> <dbl> <dbl>      <dbl> <dbl>
1         1         85         66         29         0  26.6  0.351      31     0
2         8        183         64         0         0  23.3  0.672      32     1
3         1         89         66         23        94  28.1  0.167      21     0
4         5        116         74         0         0  25.6  0.201      30     0
5         3         78         50        32        88  31     0.248      26     1
```

With the completion of our data cleaning phase, we are now prepared to transition to the data processing stage.

## Data Processing:

In our diabetes dataset analysis, we took a closer look at four important factors: pregnancies, glucose levels, age, and BMI. By grouping and organizing this information, we made it easier to uncover meaningful patterns and insights. This process helps us understand how these factors are related to diabetes and can guide us in making informed decisions about managing and preventing the condition. By categorizing and studying these columns, we hope to find valuable information that can improve our understanding of diabetes and lead to better healthcare outcomes for those affected by it.

### Making Categories of Age:

We have observed a range of ages among the patients in our dataset, with the youngest being **21** and the oldest **40**. To simplify our analysis and gain a more comprehensive understanding of the age-related trends, we have stratified these ages into two distinct categories: the first category, denoted as **(21-30)**, represents individuals between the ages of **21** and **30**, while the second category, denoted as **(31-40)**, encompasses those aged **31 to 40**. This categorization allows us to explore potential age-specific patterns and associations with greater precision and granularity.

```
> data$age_cat<-ifelse(data$Age >20 & data$Age <=30,1,ifelse(data$Age >30 & data$Age <=40,2,0))
```

### Making Categories of Pregnancies:

In light of the pregnancy data within our dataset, where the minimum recorded pregnancies are **0** and the maximum is **11**, we have logically divided this range into three distinct categories for more precise analysis. The first category, designated as **(0-5)**, encapsulates patients with 0 to 5 pregnancies. The second category, marked as **(6-10)**, encompasses patients with **6 to 10** pregnancies. Lastly, the third category, identified as **(11-15)**, represents patients with **11 to 15** pregnancies.

```
> data$preg_cat<- ifelse(data$Pregnancies>=0 & data$Pregnancies<=5,1,ifelse(data$Pregnancies>5 & data$Pregnancies<=10,2,ifelse(data$Pregnancies>11 & data$Pregnancies<=15,3,0))
```

### Making Categories of Glucose:

Considering the range of **glucose levels** in our patient dataset, with a minimum value of **44** and a maximum value of **198**, we have thoughtfully classified this spectrum into three distinct categories for more refined analysis. The first category, labelled as **(44-79)**, represents **lower** glucose levels, while the second category, denoted as **(80-99)**, signifies **normal** glucose levels. Lastly, the third category, defined as **(100-200)**, indicates **higher** glucose levels.

```
> data$glu_cat<- ifelse(data$Glucose>=44 & data$Glucose<70,1,ifelse(data$Glucose>=70 & data$Glucose<=99,2,ifelse(data$Glucose>99 & data$Glucose<=200,3,0)))
```

### Making Categories of BMI:

To enhance our analysis, we've categorized patient BMI levels into three distinct groups based on the observed range: **(0-18.4)** for **low**, **(18.5-24.8)** for **normal**, and **(24.9 and above)** for **high** BMI values, allowing for a more precise examination of BMI-related trends.

```
> data$BMI_cat<- ifelse(data$BMI>=0 & data$BMI<18.5,1,ifelse(data$BMI>=18.5 & data$BMI<24.9,2,ifelse(data$BMI>=24.9,3,0)))
```

With the completion of our data processing stage, we must now perform a thorough verification of the dataset to ensure the accuracy of the added categories column. This step is crucial in maintaining the integrity of our analysis and validating that the categorization has been correctly implemented.

```
> data
# A tibble: 485 x 13
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome age_cat preg_cat glu_cat BMI_cat
  <dbl>    <dbl>      <dbl>      <dbl>      <dbl> <dbl>      <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1         1      85        66         29        0  26.6      0.351    31      0      2      1      2      3
2         8     183        64          0        0  23.3      0.672    32      1      2      2      3      2
3         1      89        66         23       94  28.1      0.167    21      0      1      1      2      3
4         5     116        74          0        0  25.6      0.201    30      0      1      1      3      3
5         3      78        50         32       88  31       0.248    26      1      1      1      2      3
6         4     110        92          0        0  37.6      0.191    30      0      1      1      3      3
7        10     168        74          0        0  38       0.537    34      1      2      2      3      3
8         0     118        84         47      230  45.8      0.551    31      1      2      1      3      3
9         7     107        74          0        0  29.6      0.254    31      1      2      2      3      3
10        1     115        70         30       96  34.6      0.529    32      1      2      1      3      3
# i 475 more rows
# i Use `print(n = ...)` to see more rows
```

# Data Analysing and Visualizing:

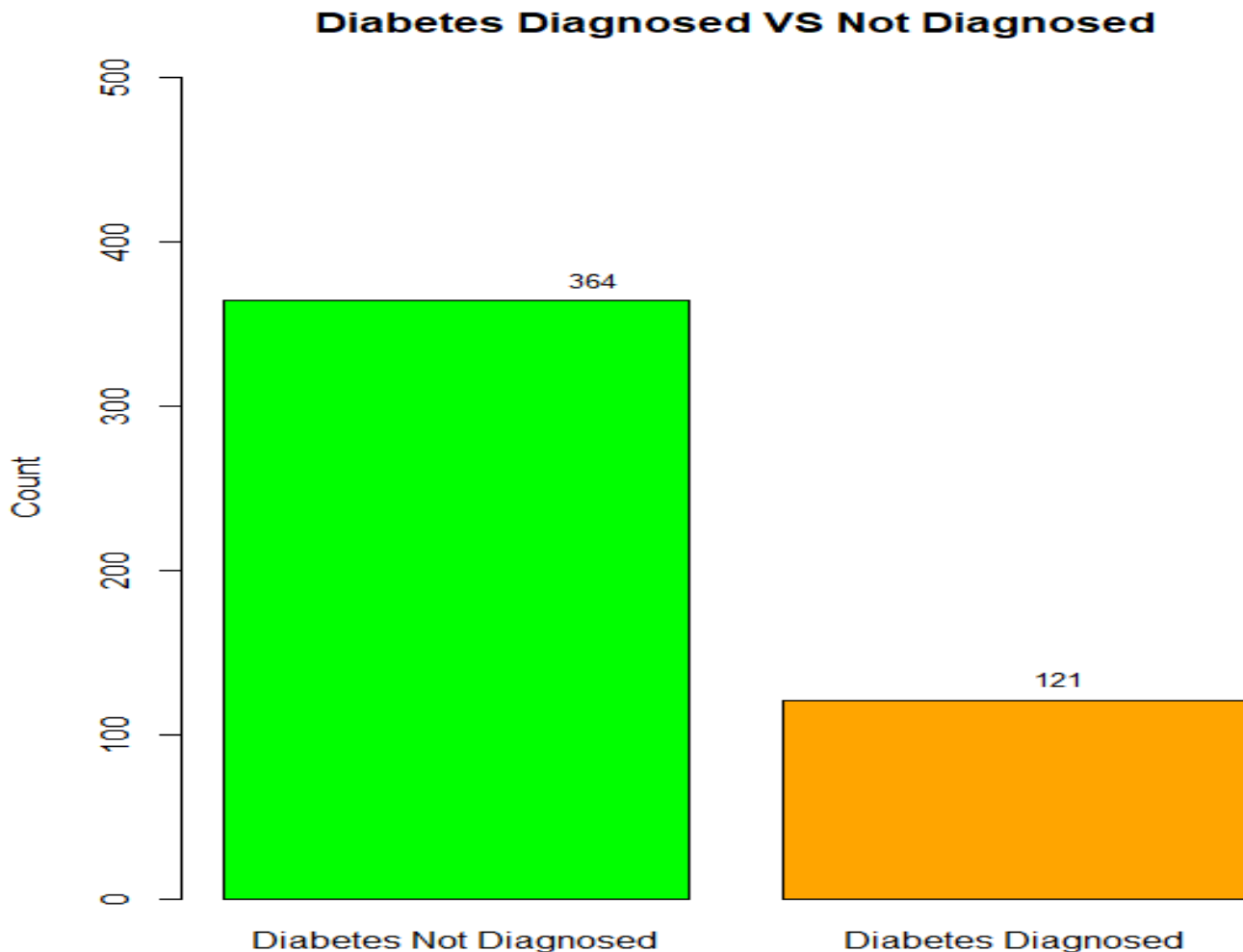
At this stage, we begin the process of analysing and visualizing our data, paving the way to draw meaningful conclusions and insights from the dataset.

## 1) DIABETES DIAGNOSED VS NOT DIGANOSED:

Based on the bar plot analysis, out of the total 485 records in our dataset after removing outliers, it is evident that 364 patients have not been diagnosed with diabetes, while 121 patients have received a diabetes diagnosis.

```
> diagnose<- table(data$outcome)
> diagnose

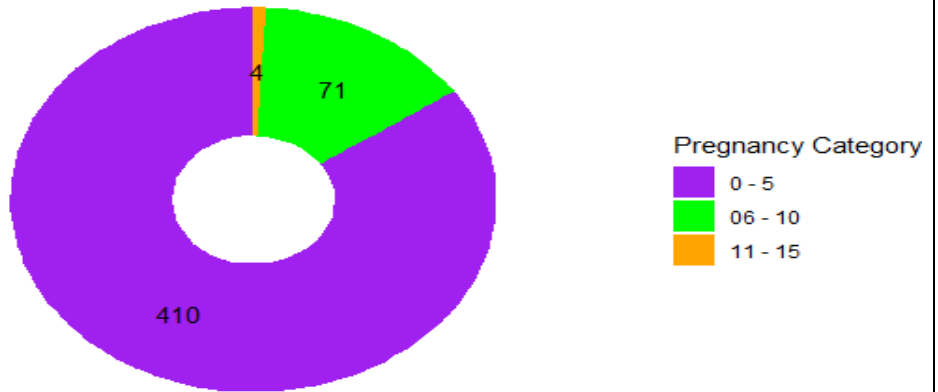
 0  1 
364 121
> 
> barplot(diagnose,ylim=c(0,500),col=c("green","orange"),names.arg=c("Diabetes Not Diagnosed","Diabetes Diagnosed"), main="Diabetes Diagnosed VS Not Diagnosed", ylab="Count")
> max_values<-max(diagnose)
> text(1:2 , diagnose, labels = diagnose, pos=3, offset = 0.5, cex=0.8, col = "black")
> 
, 1
```



## 2) NUMBER OF PATIENTS IN EACH CATEGORY:

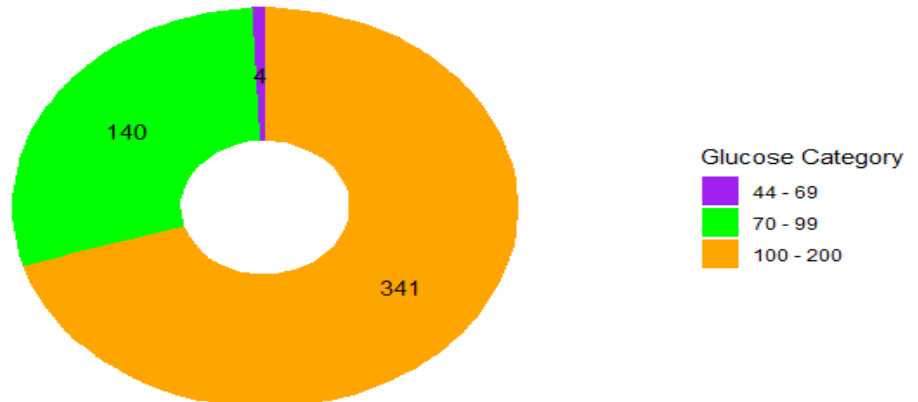
### PREGNANCIES CATEGORY:

In this chart, we can observe that **410** patients have pregnancies falling within the range of **0** to **5**, while **71** patients have pregnancies within the range of **6** to **10**. Interestingly, only **4** patients have pregnancies within the range of **11** to **15**.



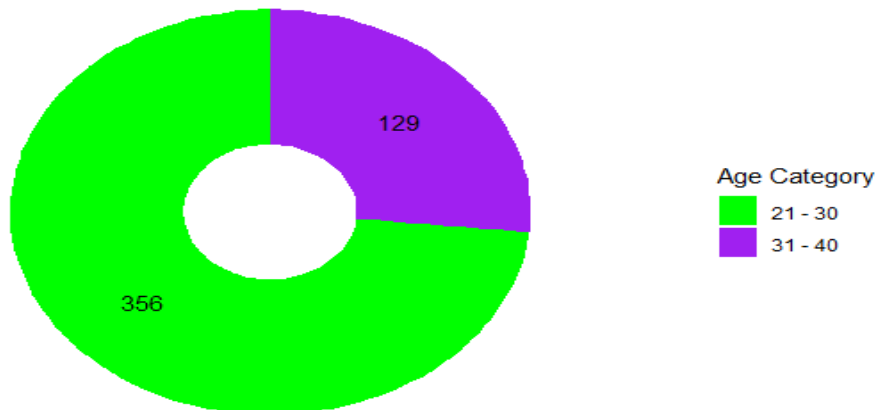
### GLUCOSE LEVEL CATEGORY:

In this chart, we analyse that **4** patients have glucose levels falling within the range of **44** to **69**, while **140** patients have glucose levels within the range of **70** to **99**. Remarkably, only **341** patients have glucose levels within the range of **100** to **200**.



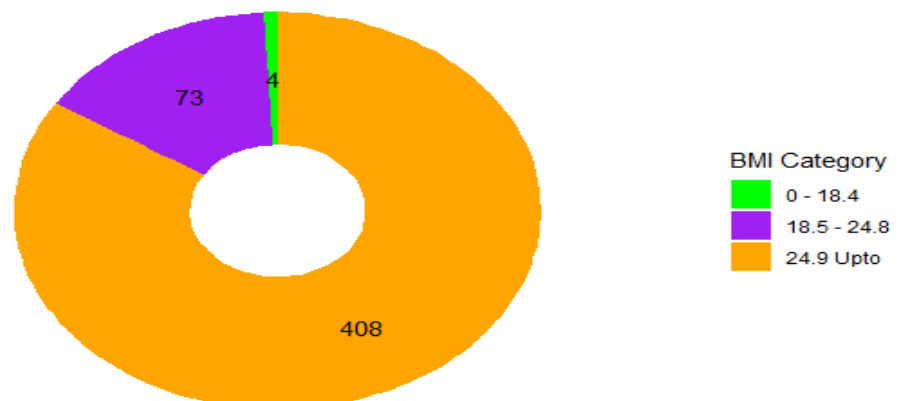
### AGE CATEGORY:

In this chart, we analyse that **356** patients have an age falling within the range of **21** to **30**, while **129** patients have an age within the range of **31** to **40**.



### BMI CATEGORY:

In this chart, we analyse that only **4** patients have a BMI falling within the range of **0** to **18.4**, while **73** patients have a BMI within the range of **18.5** to **24.8**. Remarkably, **408** patients have a BMI within the range of **24.9** and above.





### 3) CORRELATION MATRIX:

A correlation matrix is a table displaying the correlation coefficients between multiple variables, revealing the strength and direction of their relationships. We draw a correlation matrix of our data to assess and understand the relationships between various variables, aiding in data analysis and decision-making.

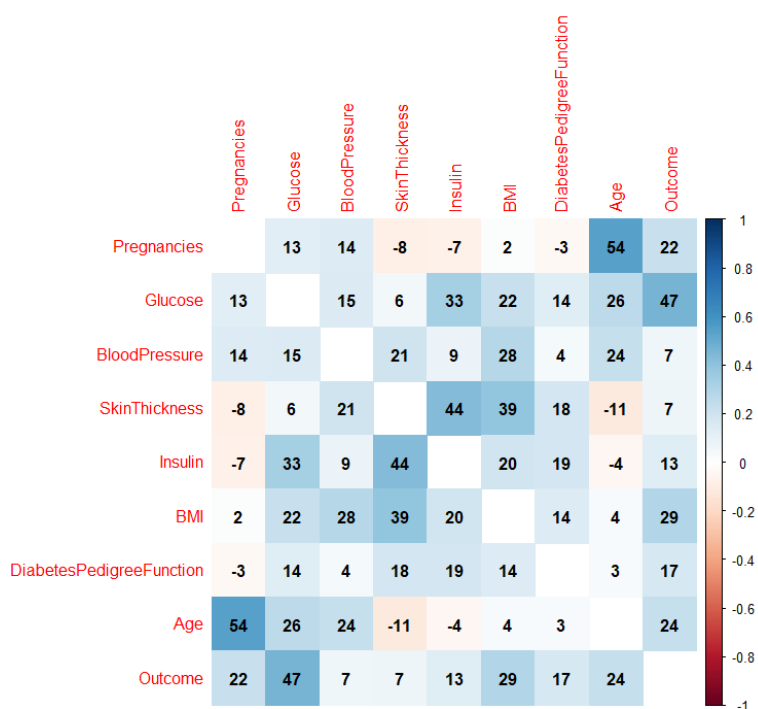
```
> correlation_matrix<-cor(diabetes)
> correlation_matrix
```

|                          | Pregnancies | Glucose    | BloodPressure | SkinThickness | Insulin     | BMI        | DiabetesPedigreeFunction | Age         | Outcome    |
|--------------------------|-------------|------------|---------------|---------------|-------------|------------|--------------------------|-------------|------------|
| Pregnancies              | 1.0000000   | 0.12945867 | 0.14128198    | -0.08167177   | -0.07353461 | 0.01768309 | -0.03352267              | 0.54434123  | 0.22189815 |
| Glucose                  | 0.12945867  | 1.00000000 | 0.15258959    | 0.05732789    | 0.33135711  | 0.22107107 | 0.13733730               | 0.26351432  | 0.46658140 |
| BloodPressure            | 0.14128198  | 0.15258959 | 1.00000000    | 0.20737054    | 0.08893338  | 0.28180529 | 0.04126495               | 0.23952795  | 0.06506836 |
| SkinThickness            | -0.08167177 | 0.05732789 | 0.20737054    | 1.00000000    | 0.43678257  | 0.39257320 | 0.18392757               | -0.11397026 | 0.07475223 |
| Insulin                  | -0.07353461 | 0.33135711 | 0.08893338    | 0.43678257    | 1.00000000  | 0.19785906 | 0.18507093               | -0.04216295 | 0.13054795 |
| BMI                      | 0.01768309  | 0.22107107 | 0.28180529    | 0.39257320    | 0.19785906  | 1.00000000 | 0.14064695               | 0.03624187  | 0.29269466 |
| DiabetesPedigreeFunction | -0.03352267 | 0.13733730 | 0.04126495    | 0.18392757    | 0.18507093  | 0.14064695 | 1.00000000               | 0.03356131  | 0.17384407 |
| Age                      | 0.54434123  | 0.26351432 | 0.23952795    | -0.11397026   | -0.04216295 | 0.03624187 | 0.03356131               | 1.00000000  | 0.23835598 |
| Outcome                  | 0.22189815  | 0.46658140 | 0.06506836    | 0.07475223    | 0.13054795  | 0.29269466 | 0.17384407               | 0.23835598  | 1.00000000 |

To gain a clearer understanding of these matrix values, we plot them into a heatmap, which provides a visual representation of variable relationships for enhanced comprehension and analysis.

```
> corplot(correlation_matrix,
method="color",
type="full",
tl.cex= 0.7,
tl.color= "black",
diag= FALSE,
addrect = 3,
addCoef.col = "black",
addCoef.cex=0.1,
addCoefasPercent = TRUE
```

In the provided heatmap, the correlation matrix values range from -1 to 1, where -1 signifies a highly negative correlation, 0 indicates no correlation, and 1 represents a strong positive correlation between the variables. The values within each block box denote the percentage of the relationship strength between the respective variables, aiding in understanding their interconnections.



#### 4) AGE CATEGORIES VS DIABETES OUTCOMES:

In our dataset, we initially narrowed our analysis to isolate the age category and outcome columns. By doing so, we aimed to directly assess the relationship between age categories and diabetes outcomes, setting aside the influence of other variables for this specific analysis.

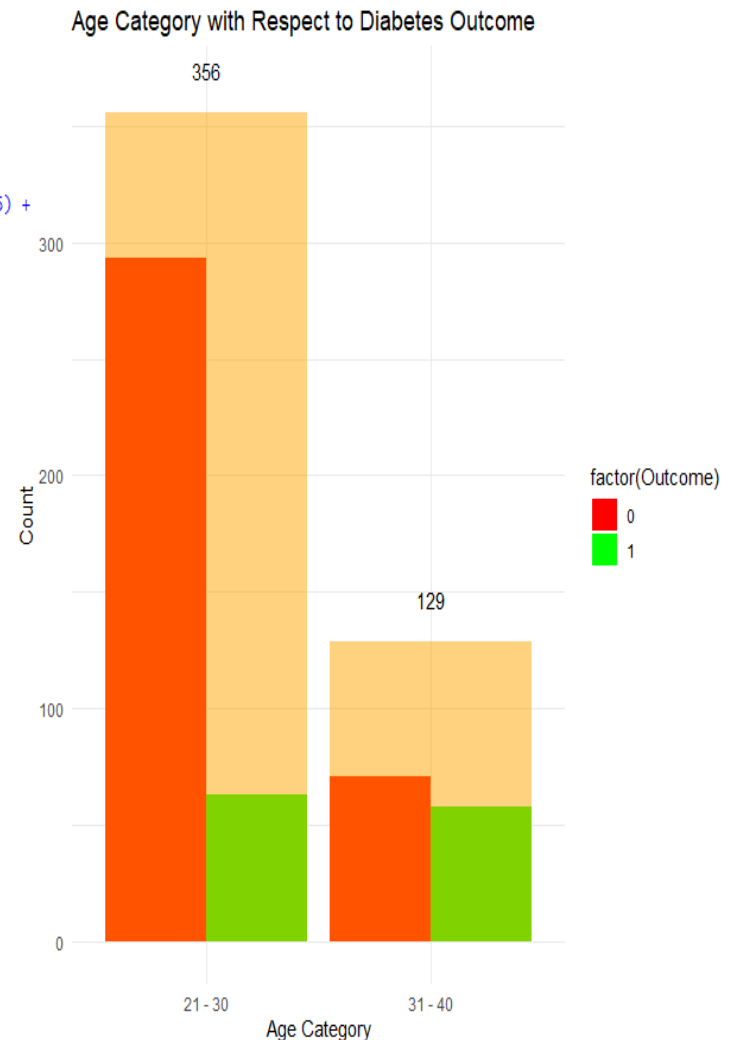
```
> data[1:3,]
# A tibble: 3 x 10
  Pregnancies Glucose BloodPressure SkinThickness Insulin BMI DiabetesPedigreeFunction Age Outcome age_cat
  <dbl>      <dbl>      <dbl>      <dbl>      <dbl> <dbl> <dbl>      <dbl> <dbl> <dbl>
1         1        85         66         29         0  26.6  0.351      31         0         2
2         8       183         64          0         0  23.3  0.672      32         1         2
3         1        89         66         23        94  28.1  0.167      21         0         1

> df<-data[,c(10,9)]
> df[1:3,]
# A tibble: 3 x 2
  age_cat Outcome
  <dbl>      <dbl>
1         2         0
2         2         1
3         1         0
```

After removing outliers, we have a total of **485 patients**. Among them, there are **356 patients** in the **(21-30)** age group and **129 patients** in the **(31-40)** age group. Using a bar plot, we examined the relationship between age categories **(21-30 and 31-40)** and diabetes outcomes. The plot reveals that a higher percentage of patients in the **31-40** age group were diagnosed with diabetes compared to the **21 - 30** age group.

```
custom_labels <- c("21 - 30", "31 - 40")
```

```
ggplot(df, aes(x = factor(age_cat))) +
  geom_bar(aes(fill = factor(Outcome)), position = "dodge") +
  geom_bar(data = df %>% group_by(age_cat) %>% summarise(total_count = n()),
    aes(x = factor(age_cat), y = total_count), fill = "orange",
    stat = "identity", position = position_dodge(width = 0.9), alpha = 0.5) +
  geom_text(data = df %>% group_by(age_cat) %>% summarise(total_count = n()),
    aes(x = factor(age_cat), y = total_count + 10, label = total_count),
    vjust = -0.5) +
  labs(title = "Age Category with Respect to Diabetes Outcome",
    x = "Age Category",
    y = "Count") +
  scale_fill_manual(values = c("0" = "red", "1" = "green")) +
  scale_x_discrete(labels = custom_labels) +
  theme_minimal()
```



## 5) PREGNANCIES CATEGORIES VS DIABETES OUTCOMES:

In our dataset, we initially focused our analysis on isolating the pregnancies category and outcome columns. By doing so, our goal was to directly examine the relationship between the number of pregnancies and diabetes outcomes, while temporarily disregarding the potential impact of other variables for this specific analyse

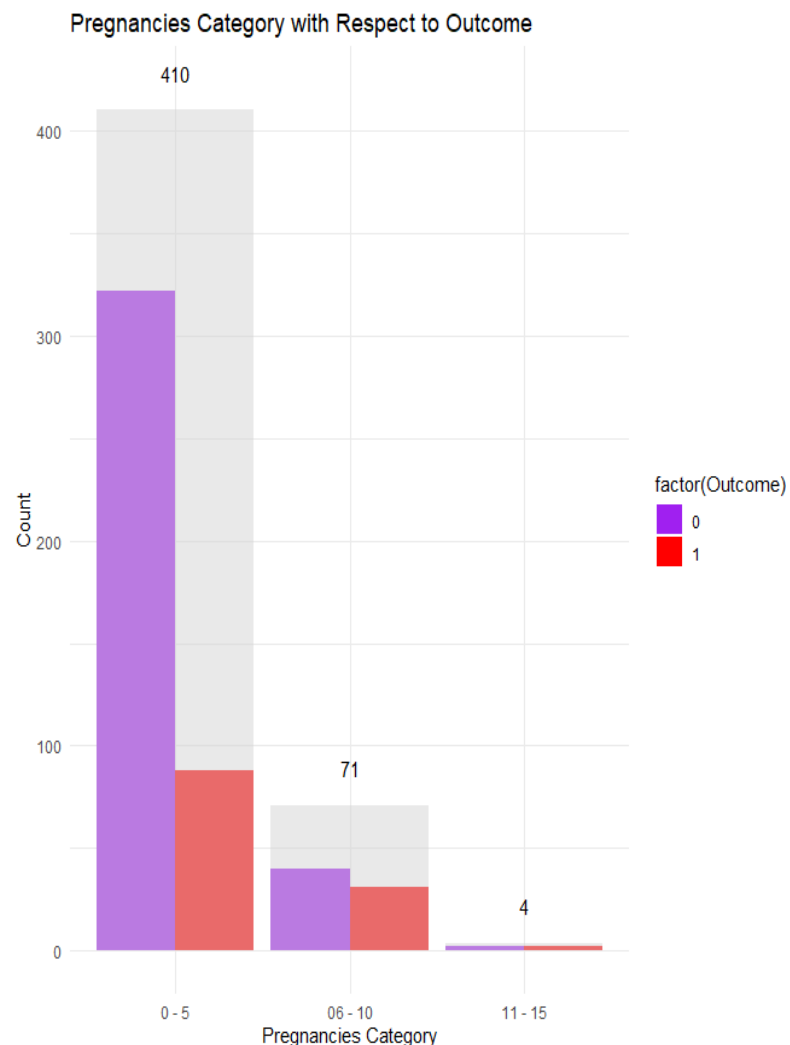
```
> data[1:3,]
# A tibble: 3 x 13
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome age_cat preg_cat glu_cat BMI_cat
  <dbl>      <dbl>      <dbl>      <dbl>      <dbl>  <dbl> <dbl>      <dbl> <dbl> <dbl> <dbl> <dbl>
1         1        85         66         29         0    26.6  0.351     31         0         2         1         2         3
2         8       183         64          0         0    23.3  0.672     32         1         2         2         3         2
3         1        89         66         23        94    28.1  0.167     21         0         1         1         2         3

> preg<-data[,c(11,9)]
> preg[1:3,]
# A tibble: 3 x 2
  preg_cat Outcome
  <dbl>      <dbl>
1         1         0
2         2         1
3         1         0
```

The total of **485** patients. Among them, there are **410** patients in the **(0-5)** pregnancies category, **71** patients in the **(6-10)** pregnancies category, and **4** patients in the **(11-15)** pregnancies category. Using a bar plot, we examined the relationship between pregnancy categories **(0-5, 6-10, and 11-15)** and diabetes outcomes. The plot reveals that a higher percentage of patients in the **(0-5)** pregnancies category was diagnosed with diabetes compared to the **(6 -10)** pregnancies category, while the **(11-15)** pregnancies category had very few patients

```
> custom_labels <- c("0 - 5", "06 - 10", "11 - 15")

ggplot(preg, aes(x = factor(preg_cat))) +
  geom_bar(aes(fill = factor(Outcome)), position =
"dodge") +
  geom_bar(data = preg %>% group_by(preg_cat) %>%
summarise(total_count = n()),
  aes(x = factor(preg_cat), y = total_count
), fill = "lightgray",
  stat = "identity", position =
position_dodge(width = 0.9), alpha = 0.5) +
  geom_text(data = preg %>% group_by(preg_cat) %>%
summarise(total_count = n()),
  aes(x = factor(preg_cat), y = total_count
+ 10, label = total_count),
  vjust = -0.5) +
  labs(title = "Pregnancies Category with Respect to
Outcome",
  x = "Pregnancies Category",
  y = "Count") +
  scale_fill_manual(values = c("0" = "purple", "1" =
"red")) +
  scale_x_discrete(labels = custom_labels) +
  theme_minimal()
```



## 6) GLUCOSE CATEGORIES VS DIABETES OUTCOMES:

In our dataset, we initially concentrated our analysis on isolating the glucose category and outcome columns. This approach allowed us to directly investigate the association between glucose levels and diabetes outcomes, while temporarily setting aside the potential influence of other variables for this specific analysis.

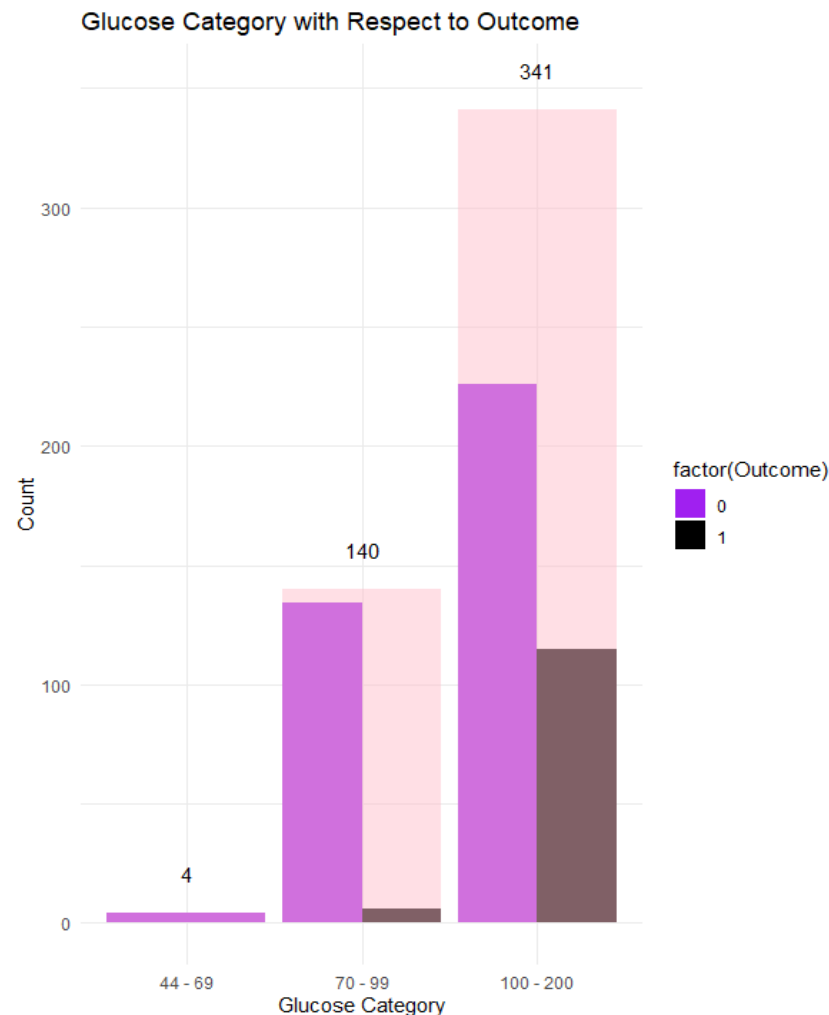
```
> data[1:3,]
# A tibble: 3 x 13
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome age_cat preg_cat glu_cat BMI_cat
  <dbl>      <dbl>      <dbl>      <dbl>      <dbl> <dbl> <dbl>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1         1         85         66         29         0  26.6  0.351      31     0     2     1     2     3
2         8        183         64         0         0  23.3  0.672      32     1     2     2     3     2
3         1         89         66         23        94  28.1  0.167      21     0     1     1     2     3

> glu<-data[,c(12,9)]
> glu[1:3,]
# A tibble: 3 x 2
  glu_cat Outcome
  <dbl>      <dbl>
1     2         0
2     3         1
3     2         0
```

We have a total of **485** patients. Among them, there are **4** patients in the **(44-69)** glucose category, **140** patients in the **(70-99)** glucose category, and **341** patients in the **(100-200)** glucose category. Using a bar plot, we examined the relationship between glucose categories **(44-69, 70-99, and 100-200)** and diabetes outcomes. The plot reveals that a higher percentage of patients in the **(100-200)** glucose category was diagnosed with diabetes compared to the **(44-69)** and **(70-99)** glucose categories.

```
> custom_labels <- c("44 - 69", "70 - 99", "100 - 200")
)

ggplot(glu, aes(x = factor(glu_cat))) +
  geom_bar(aes(fill = factor(Outcome)), position = "dodge") +
  geom_bar(data = glu %>% group_by(glu_cat) %>% summarise(total_count = n()),
    aes(x = factor(glu_cat), y = total_count), fill = "pink",
    stat = "identity", position = position_dodge(width = 0.9), alpha = 0.5) +
  geom_text(data = glu %>% group_by(glu_cat) %>% summarise(total_count = n()),
    aes(x = factor(glu_cat), y = total_count + 10, label = total_count),
    vjust = -0.5) +
  labs(title = "Glucose Category with Respect to Outcome",
    x = "Glucose Category",
    y = "count") +
  scale_fill_manual(values = c("0" = "purple", "1" = "black")) +
  scale_x_discrete(labels = custom_labels) +
  theme_minimal()
```



## 7) BMI CATEGORIES VS DIABETES OUTCOMES:

In our dataset, our initial focus was on isolating the BMI category and outcome columns. This approach enabled us to directly explore the relationship between BMI values and diabetes outcomes, while temporarily disregarding the potential effects of other variables for this specific analysis.

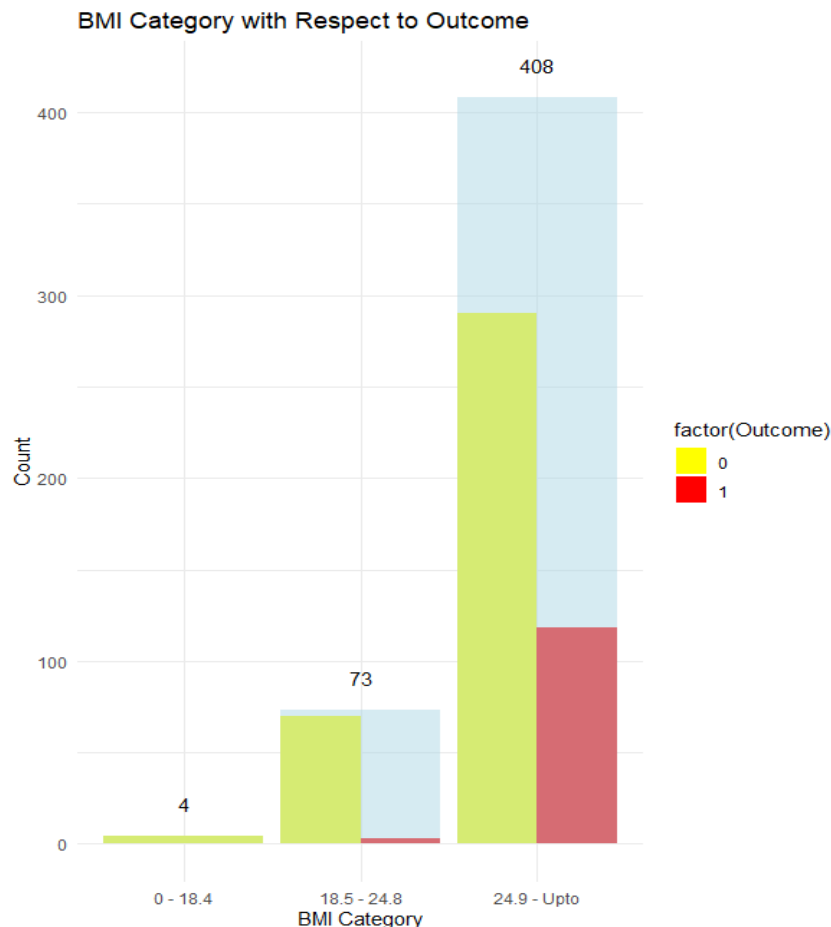
```
> data[1:3,]
# A tibble: 3 x 13
  Pregnancies Glucose BloodPressure SkinThickness Insulin BMI DiabetesPedigreeFunction Age Outcome age_cat preg_cat glu_cat BMI_cat
  <dbl>      <dbl>      <dbl>      <dbl>      <dbl> <dbl>      <dbl>      <dbl> <dbl> <dbl> <dbl> <dbl>
1         1        85         66         29         0  26.6      0.351     31         0         2         1         2         3
2         8       183         64          0         0  23.3      0.672     32         1         2         2         3         2
3         1        89         66         23        94  28.1      0.167     21         0         1         1         2         3
```

```
> BMic<-data[,c(13,9)]
> BMic[1:3,]
# A tibble: 3 x 2
  BMI_cat Outcome
  <dbl>      <dbl>
1         3         0
2         2         1
3         3         0
```

We have a total of **485** patients. Among them, there are **4** patients in the **(0-18.4)** BMI category, **73** patients in the **(18.5-24.8)** BMI category, and **408** patients in the **(24.9-UPTO)** BMI category. Using a bar plot, we examined the relationship between BMI categories (**0-18.4**, **18.5-24.8**, and **24.9-upto**) and diabetes outcomes. The plot reveals that a higher percentage of patients in the **(24.9- up to)** BMI category was diagnosed with diabetes compared to the **(0-18.4)** and **(18.5-24.8)** BMI categories.

```
> custom_labels <- c("0 - 18.4", "18.5 - 24.8", "24.9 - upto")

ggplot(BMic, aes(x = factor(BMI_cat))) +
  geom_bar(aes(fill = factor(Outcome)),
    position = "dodge") +
  geom_bar(data = BMic %>% group_by(BMI_cat) %
    >% summarise(total_count = n()),
    aes(x = factor(BMI_cat), y =
      total_count), fill = "lightblue",
    stat = "identity", position =
      position_dodge(width = 0.9), alpha = 0.5) +
  geom_text(data = BMic %>% group_by(BMI_cat) %
    >% summarise(total_count = n()),
    aes(x = factor(BMI_cat), y =
      total_count + 10, label = total_count),
    vjust = -0.5) +
  labs(title = "BMI Category with Respect to
    Outcome",
    x = "BMI Category",
    y = "Count") +
  scale_fill_manual(values = c("0" = "yellow",
    "1" = "red")) +
  scale_x_discrete(labels = custom_labels) +
  theme_minimal()
```



## 8) CORRELATION COEFFICIENTS WITH OUTCOME:

Correlation coefficients with the "Outcome" variable measure the strength and direction of linear relationships between each of the other variables (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age) and whether an individual has diabetes or not.

```
> correlation_vector<-cor(data)[,"Outcome"]
> correlation_vector
```

| Pregnancies | Glucose                  | BloodPressure | SkinThickness | Insulin    |
|-------------|--------------------------|---------------|---------------|------------|
| 0.20907176  | 0.48330114               | 0.10841405    | 0.07902087    | 0.12588378 |
| BMI         | DiabetesPedigreeFunction | Age           | Outcome       |            |
| 0.28588543  | 0.17878545               | 0.28120704    | 1.00000000    |            |

In the correlation coefficient analysis, we exclude the "Outcome" variable from the results since there's no need to measure its correlation with itself, allowing us to focus on the relationships between other variables and the diabetes outcome.

```
> #remove Outcome vector
> correlation_vector<-correlation_vector[-9]
> correlation_vector
```

| Pregnancies | Glucose                  | BloodPressure | SkinThickness | Insulin    |
|-------------|--------------------------|---------------|---------------|------------|
| 0.20907176  | 0.48330114               | 0.10841405    | 0.07902087    | 0.12588378 |
| BMI         | DiabetesPedigreeFunction | Age           |               |            |
| 0.28588543  | 0.17878545               | 0.28120704    |               |            |

After calculating the coefficient values, we can determine the names of the coefficients associated with each variable. This information is crucial for plotting a graph to visualize and better understand the relationships between variables and the diabetes outcome.

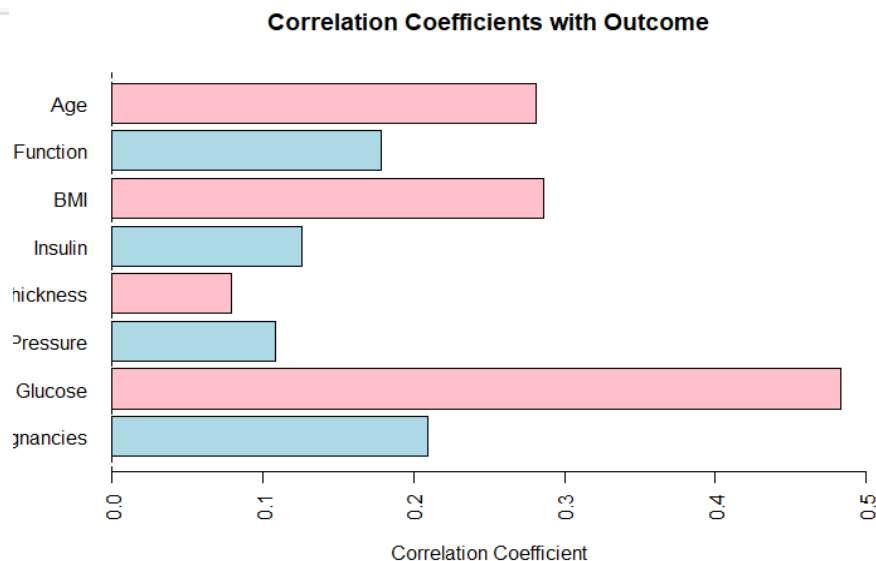
```
> #get the features names
> features_name<-names(data)[-9]
> features_name
```

| [1] | "Pregnancies" | "Glucose"                  | "BloodPressure" | "SkinThickness" | "Insulin" |
|-----|---------------|----------------------------|-----------------|-----------------|-----------|
| [6] | "BMI"         | "DiabetesPedigreeFunction" | "Age"           |                 |           |

After extracting the feature names and plotting a graph of coefficient values, our analysis indicates that the features Glucose, Age, and BMI are highly correlated with the diabetes outcome, suggesting that these variables have a significant impact on predicting diabetes.

```
> color_vector <- rep(c("lightblue", "pink"), length.out =
length(correlation_vector))

barplot(correlation_vector,
  main = "Correlation Coefficients with Outcome",
  xlab = "Correlation Coefficient",
  col = color_vector,
  xlim = c(0, 0.5),
  horiz = TRUE,
  names.arg = features_name,
  las = 2)
abline(v = 0, lty = 2, col = "black")
```



## 9) DECISION TREE ANALYSIS

Decision tree analysis is a machine learning technique used to model and predict outcomes by recursively splitting data based on the most significant variables. In this context, we are utilizing the variables Glucose, Age, and BMI due to their high impact on the diabetes outcome, as determined by coefficient analysis. These variables will be key factors in constructing the decision tree model to predict diabetes outcomes.

```
> library(rpart)
> library(rpart.plot)
> trees<-rpart(d$Outcome~d$Age+d$Glucose+d$BMI)
> rpart.plot(trees)
```

The following decision tree illustrates:

- **First Layer (Root):** If **glucose** is smaller than **155**, the patient is predicted not to have diabetes; if glucose is greater than or equal to **155**, the patient is predicted to have diabetes.
- **Second Layer:** If **glucose** is smaller than **102**, than patient is predicted not to have diabetes; otherwise, the patient is predicted to have diabetes.
- **Third Layer:** If **age** is smaller than **29** than patient is predicted not to have diabetes; otherwise, the patient is predicted to have diabetes.
- **Fourth Layer:** BMI is greater than or equal to **31**, the patient is predicted to have diabetes.

