**Task 1: Dataset Identification and Background on Analysis**

**1.0 Introduction**

Diabetes is one of the most common chronic diseases worldwide, affecting millions of people. Early detection is crucial for effective management and prevention of complications. In this study, we analyze a **diabetes prediction** dataset obtained from Kaggle, which includes various risk factors such as **age, BMI, hypertension, heart disease, HbA1c\_level, and smoking history**. These factors play a significant role in determining a person’s likelihood of developing diabetes.

By examining this dataset, its aim to identify patterns and trends that contribute to diabetes. Understanding these risk factors will help in developing better prevention strategies and improving early diagnosis. The dataset provides a comprehensive set of patient attributes, making it valuable for gaining insights into how lifestyle and health conditions influence diabetes. Through this study, we hope to provide meaningful recommendations that can support healthcare professionals and individuals in managing and reducing the risk of diabetes.

**2.0 Identification of dataset analysis background**

**2.1 Objective of the analysis**

The **objective of this analysis** is to identify **key risk factors** associated with diabetes and develop predictive insights that can contribute to **early detection and better disease management**. By analysing correlations between variables such as **age, BMI, and blood glucose levels,** this study aims to determine their influence on diabetes. Additionally, the research will assess how **lifestyle choices, such as smoking history, impact diabetes prevalence**. The insights gained from this analysis can help in designing **preventive healthcare strategies and awareness programs** to reduce diabetes-related complications.

**2.2 Scope of the analysis**

The **scope of this analysis** includes several key aspects. The study will focus on **identifying relationships between major health indicators and diabetes status**. Specifically, the analysis will explore **how age, BMI, blood glucose levels, and smoking history contribute to diabetes risk**. Additionally, it will examine **demographic trends by analysing the distribution of diabetes cases based on gender and age groups**. However, certain aspects are will be not used for example hypertension, heart diseases and HbA1c\_level.

**2.3 Target Audience**

The findings of this study will benefit **various stakeholders**. **Healthcare professionals** can use these insights to **identify high-risk patients** and implement **early intervention strategies**. **Researchers and data scientists** working in **medical AI and predictive modeling** can use this dataset for further model development, improving the accuracy of diabetes prediction algorithms. **Public health organizations** can leverage the results to design **awareness programs** and **policy recommendations** aimed at diabetes prevention. Lastly, the **general public** will benefit from understanding their **risk levels based on lifestyle and medical conditions**, enabling them to take **preventive actions** such as dietary changes and lifestyle modifications.

**Task 2**: **Gathering Relevant Data**

**3.0 Dataset**

**3.1 Reasons for data selection:**

For this analysis, selecting a suitable dataset is crucial to ensuring accurate and meaningful results. The **Diabetes Prediction Dataset**, obtained from **Kaggle**, is chosen due to its **comprehensive set of attributes related to diabetes risk factors**. This dataset includes **100,000 patient records**, making it statistically significant for generating reliable insights. The dataset provides a diverse range of variables, including **age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level**, all of which are relevant for understanding diabetes risk.

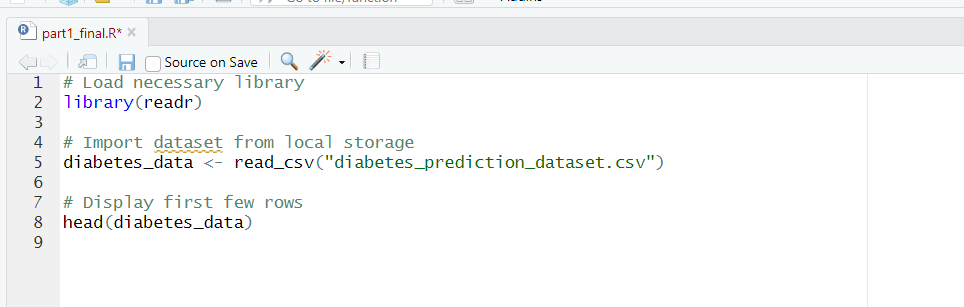
**3.2 Select suitable dataset to fulfil the objective:**

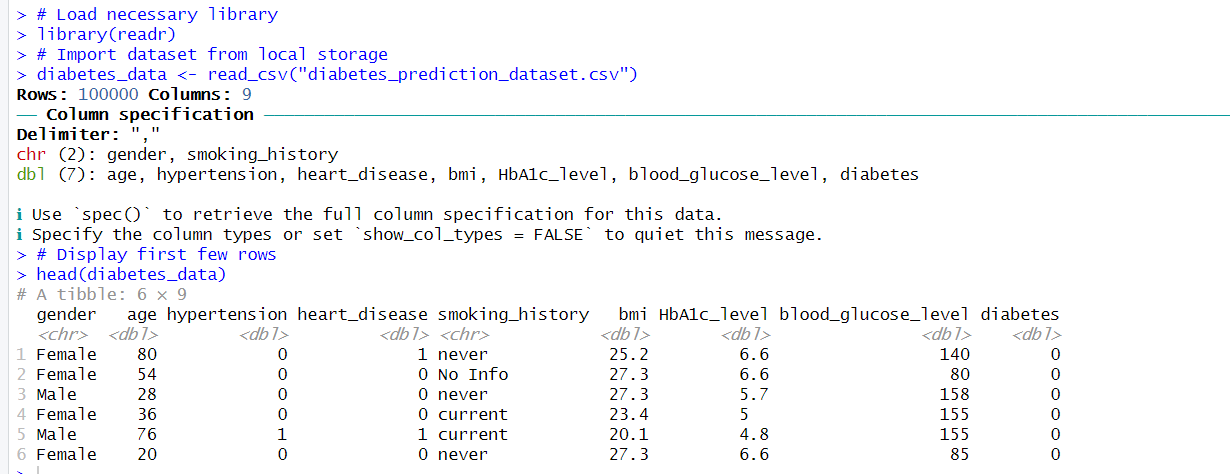
The **justification for selecting this dataset** is based on its **relevance to the study objective**. Since the goal is to identify key factors contributing to diabetes, the dataset provides **both medical and lifestyle-related indicators**, allowing for a **comprehensive analysis**. Unlike smaller datasets, this dataset offers a **large sample size**, which enhances **data reliability and reduces bias**. Additionally, it includes biometric **indicators** such as **HbA1c level and blood glucose level**,

**3.3 Dataset source and evidence:**

The dataset is obtained from **Kaggle**, a well-known online platform that provides high-quality datasets for research and data science projects. The dataset can be accessed at Kaggle: <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?resource=download>

Additionally, this dataset can be loaded into **RStudio** using CSV import functions. The dataset is available in **CSV format**, which allows easy loading and preprocessing using R. Below is the **code snippet** demonstrating how the dataset can be loaded into RStudio:





**3.4 Specify suitable variables:**

In this study, key variables have been selected to analyze factors associated with diabetes risk. The **independent variables** include **age, gender, BMI, smoking history, and blood glucose level**, as these factors play a significant role in influencing an individual's likelihood of developing diabetes. The **dependent variable** is **diabetes status** (1 = Yes, 0 = No), which indicates whether a person has been diagnosed with diabetes.

By examining the relationship between these independent variables and diabetes occurrence, this analysis aims to identify patterns that contribute to early detection and preventive healthcare strategies. The dataset is well-suited for this research due to its size, data quality, and relevance to diabetes risk assessment. By leveraging these variables, the study seeks to provide meaningful insights for medical professionals, researchers, and individuals at risk of diabetes.

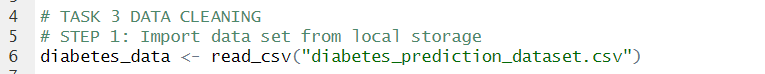
**Task 3: Data Cleaning - Handling Missing or Incomplete Data**

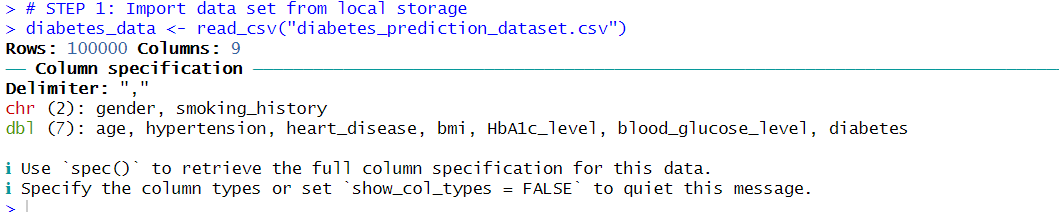
**4.0 Data cleaning**

Upon analyzing the dataset, it was observed that the smoking\_history column contained the value "No Info", which indicates missing information. Since this value does not contribute meaningful insights to the analysis, it was decided to treat it as missing data (NULL in database terms or NA in R).

To address this issue, the following steps were taken:

1. **Import data set from local storage:**

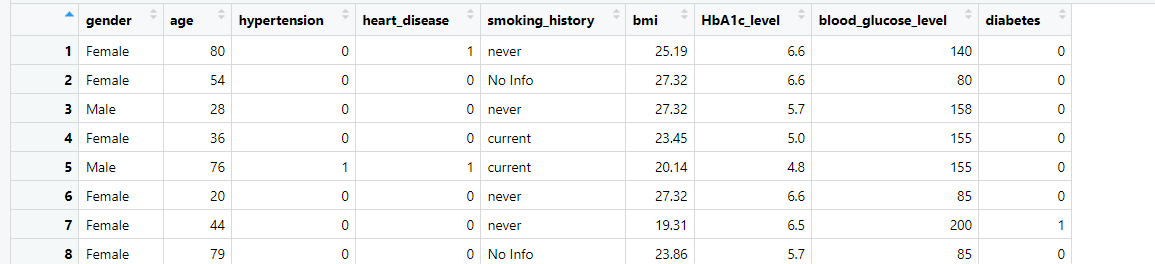
****

****

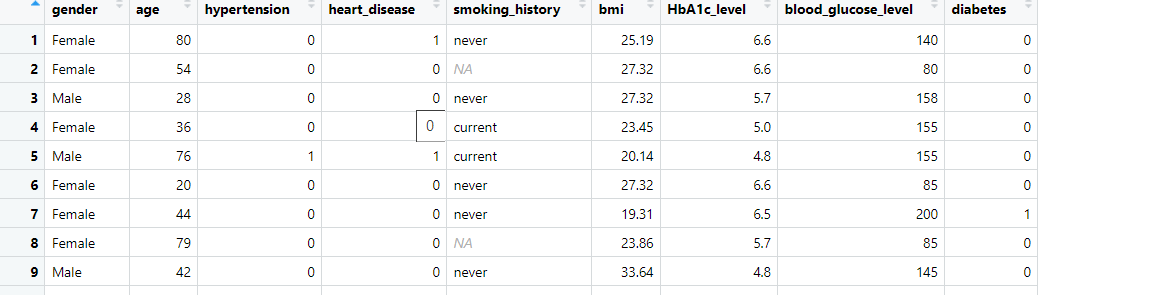
* **Replace "No Info" with NA in the 'smoking\_history' column:**

****

* This action should be performed by replacing "No info" with "NA," as "No info" indicates NULL or NA.
* **Before Replace “No info” to “NA”**

****

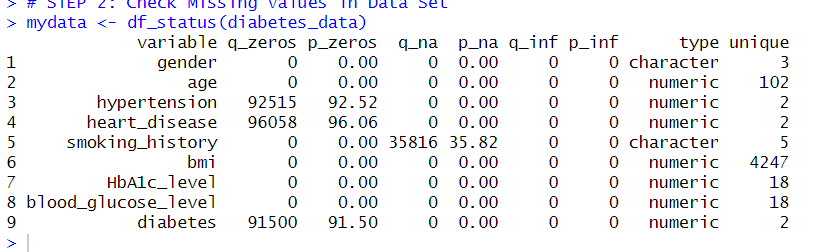
* “No info” in smoking\_history column
* **After replace “No info” to “NA”**

****

* After replace “No info” to “NA”

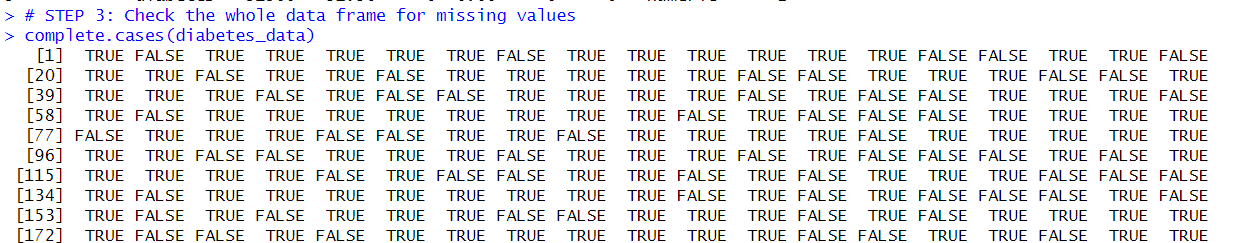
1. **Check missing values in data set:**

****

****

* The image displays the output of mydata <- df\_status(diabetes\_data), showing that hypertension, heart disease, and diabetes have a quantity of zero. However, in this dataset, 0 actually indicates "No," while 1 indicates "Yes." Therefore, the quantity of zeros should be ignored. Meanwhile, the smoking history contains "NA" values, which should be addressed appropriately.

1. **Check the whole data frame for missing values:**



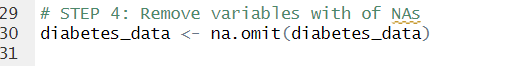
* The image shows the output of complete.cases(diabetes\_data), which checks for missing values in the dataset. TRUE indicates a valid (non-missing) value, while FALSE means a missing value in that column. Rows with only TRUE have no missing data, whereas those with FALSE contain missing values. This helps identify incomplete records for data cleaning.

1. **Remove variables with of Nas:**

* **Before**

****

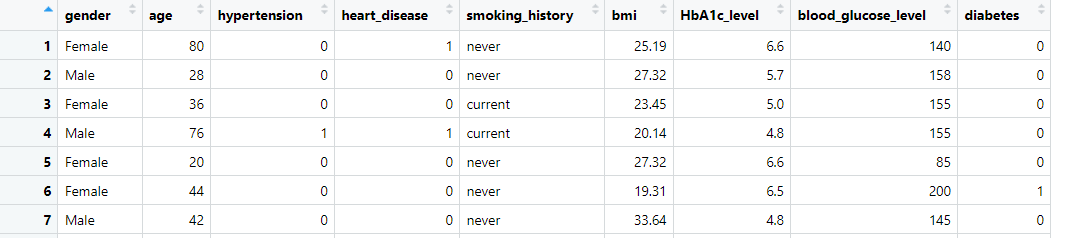
* Number of objects in data set before cleaning is 100000 and it have 9 variables.

****

* Remove “NA” values using command diabetes\_data <- na.omit(diabetes\_data)
* **After remove “NA” values**

****

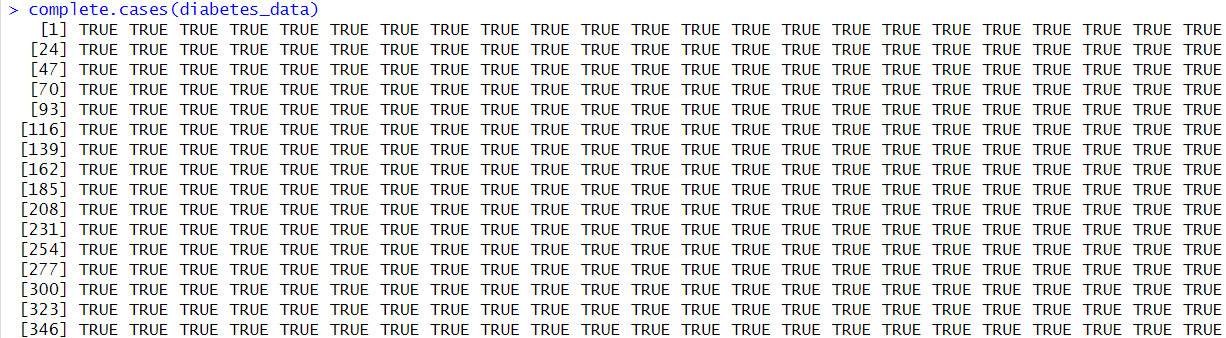
* After removing “NA” the objects are reducing to 64184 and the variables are still remained 9

****

* As shown above no more “NA” values available.

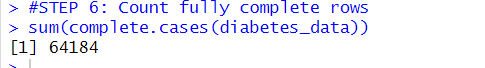
1. **Check for missing values:**

****

****

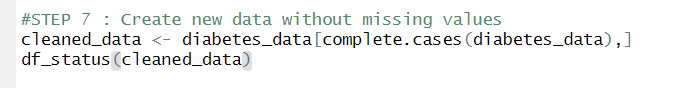
* This image shows the output of `complete.cases(diabetes\_data)`, where all values are “TRUE”, indicating that there are no missing values in the dataset. This means the data is fully complete and does not require further cleaning for missing values.

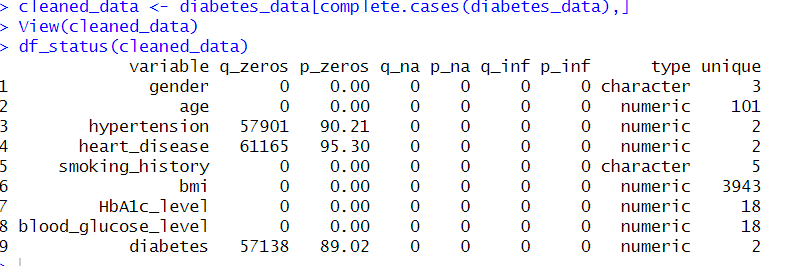
1. **Count fully complete rows:**

****

* The output [1] 64184 indicates that there are **64,184 rows** in diabetes\_data that do not contain any missing values.

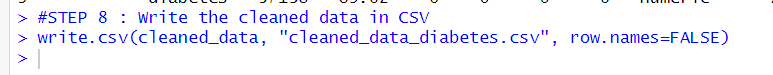
1. **Create new data without missing values:**

****

****

* After confirm that the data is clean, thus new data set can be created based on cleaned data set.

1. **Write the cleaned data in new csv file:**



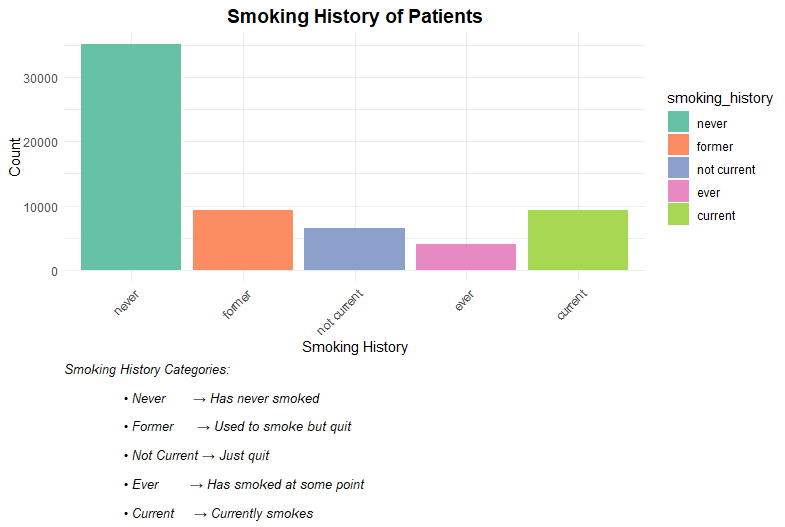
* Write it in csv file.

**Task 4: Analyze the data to extract meaningful insights and trends**

**5.0 Data Analysis (EDA) and Visualisation**

**5.1 Univariate data analysis:**

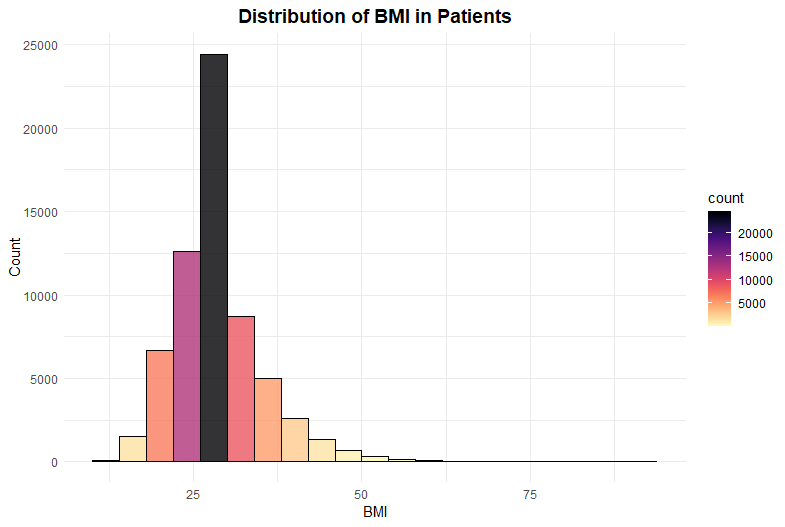
* **5.1.1 Univariate 1 Smoking History of Patient ( no diabetes status )**

****

The bar chart shows the smoking history of patients, categorized into five groups. The **"never"** category has the highest count, with over 30,000 patients who have never smoked. The **"former"** category follows, with around 10,000 patients who used to smoke but quit. The **"not current"** category, representing those who recently quit smoking, has a slightly lower count than the former category. The **"ever"** category, which includes individuals who have smoked at some point in their lives, has even fewer patients. Lastly, the **"current"** category represents those who still smoke, with a count higher than **"ever"** but much lower than **"never."**

In summary, the majority of **patients have never smoked**, while a significant number have quit smoking. The number of current smokers is relatively lower, suggesting that many people either never start smoking or eventually quit.

* **5.1.2 Univariate 2 Distribution of BMI in Patients ( no diabetes status )**

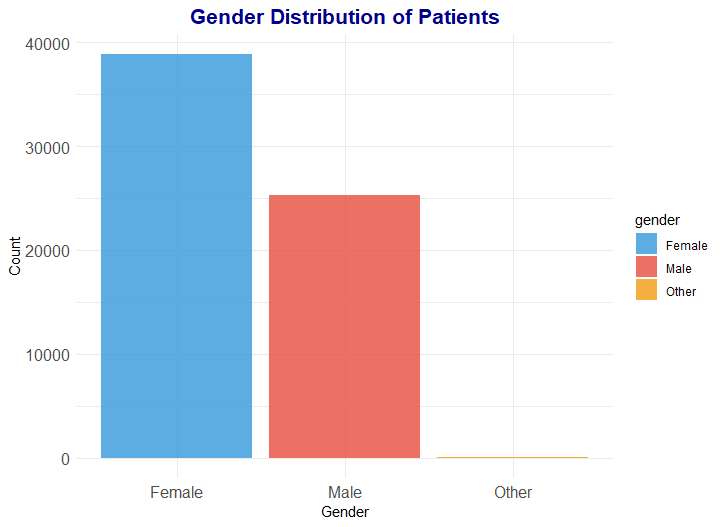
****

The histogram represents the **distribution of BMI in patients**, showing how BMI values are spread across the dataset. The majority of patients have a **BMI around 25**, as indicated by the tallest bar in black, which represents more than **20,000 patients**. The distribution is skewed towards the lower BMI values, with most patients falling within the **20 to 35 range**, where the frequency remains relatively high. As BMI increases beyond **35**, the count starts to decrease, with fewer patients in higher BMI ranges. The number of patients with a **BMI over 50** is significantly lower, as seen from the shorter bars in lighter colors.

The color gradient, ranging from **black to yellow**, visually represents the patient count, with **black indicating the highest frequency** and **lighter shades representing lower counts**. This helps to easily identify the most common BMI range in the dataset.

In summary, most patients have a **BMI close to 25**, which is typically considered within the normal to overweight range. The dataset shows a steady decline in patient count as BMI increases, suggesting that **extremely high BMI values are less common**. This distribution highlights that while some patients have a **higher BMI**, the majority are clustered around the **normal to moderately overweight range**

* **5.1.3 Univariate 3 Gender Distribution of Patient ( no diabetes status )**

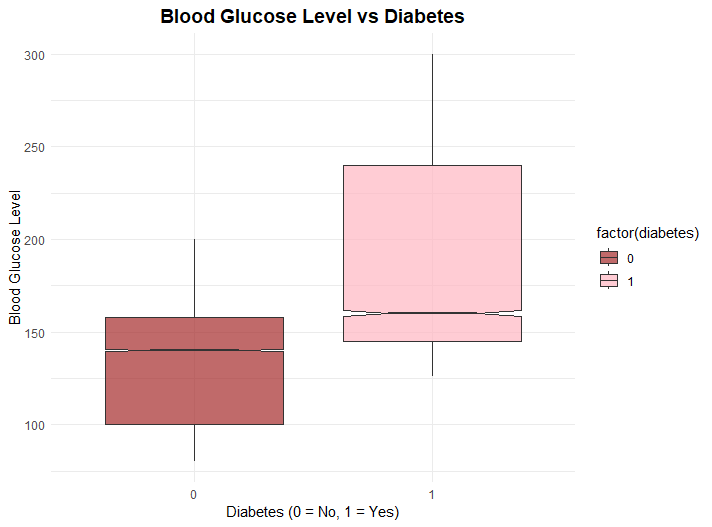
****

The bar chart represents the gender distribution of patients in the dataset. The x-axis displays three gender categories: Female, Male, and Other, while the y-axis represents the number of patients. The chart uses different colors to distinguish each category, with blue for females, red for males, and orange for the "Other" category. The data shows that the majority of patients are female, with around 40,000 individuals, followed by male patients, who account for approximately 25,000. The "Other" category has a significantly lower number of patients.

In conclusion this suggests that females make up the largest proportion of patients in this dataset, while the "Other" gender category is much less represented.

**5.2 Bivariate data analysis:**

* **5.2.1 Bivariate 1 Blood Glucose Level vs Diabetes**

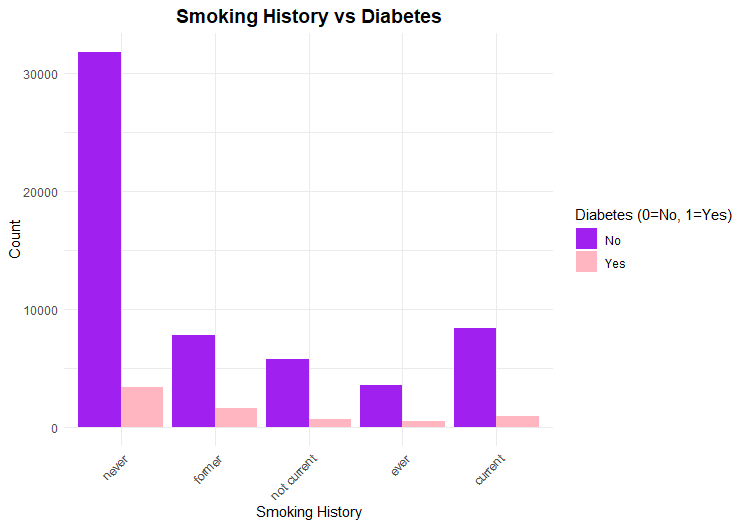
****

The boxplot compares **blood glucose levels between diabetic and non-diabetic patients.** The x-axis represents diabetes status, with **0 indicating non-diabetic individuals and 1 representing those with diabetes,** while the y-axis shows blood glucose levels. The non-diabetic group, represented by the dark red box, has a **median blood glucose level around 130 to 140.** The interquartile range (IQR), which includes the middle 50% of values, falls roughly **between 100 and 160,** indicating a more stable blood glucose level. Although there are some outliers, they are not extreme.

In contrast, the diabetic group, shown by the light pink box, has a **higher median blood glucose level, around 150 to 160**. The IQR is much wider, ranging from about 120 to 250, demonstrating greater variability in blood glucose levels among diabetic patients. This group also has more extreme outliers, with some values exceeding 300, suggesting that blood sugar levels fluctuate significantly in this population.

Overall, the data indicates that **diabetic patients generally have higher blood glucose levels than non-diabetic individuals**. The greater variation in glucose levels among diabetics may be due to differences in disease severity, treatment, or lifestyle factors. This analysis highlights the importance of regular monitoring and management of blood glucose levels in diabetic patients to prevent complications associated with extreme fluctuations.

* **5.2.2 Bivariate 2 Smoking History vs Diabetes**

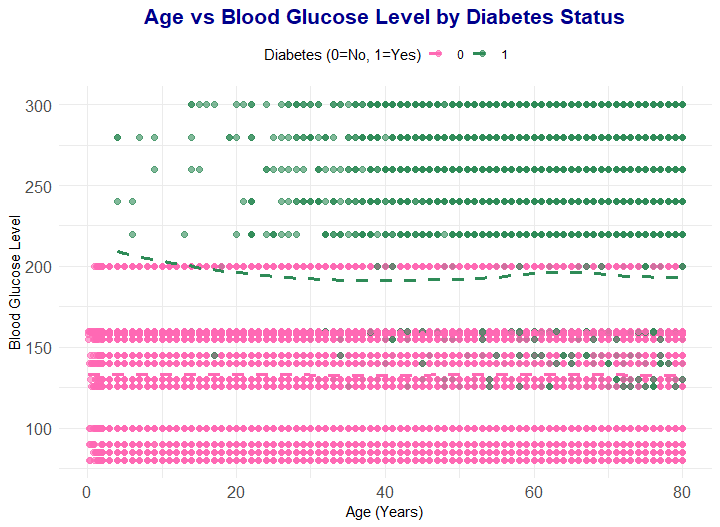
****

The bar chart illustrates the **relationship between smoking history and diabetes status.** The x-axis represents different categories of smoking history, **including "never," "former," "not current," "ever," and "current,**" while the y-axis indicates the count of individuals in each category. The data is further divided into two groups based on diabetes status, **non-diabetic individuals, represented by purple bars, and diabetic individuals, represented by pink bars**.

The chart shows that the majority of individuals, both diabetic and non-diabetic, have never smoked. This category has the highest count, with a significant number of non-diabetic individuals compared to diabetics. The second largest group consists of former smokers, followed by those who are classified as "not current" smokers, meaning they may have smoked in the past but do not currently smoke. The "ever" and "current" smoker categories have the lowest counts, though there are still noticeable differences between diabetic and non-diabetic individuals.

Across all smoking categories, the number of non-diabetic individuals is consistently higher than diabetic individuals. However, a notable trend is that in each smoking group, there are still diabetic individuals present, **suggesting that smoking history may be a relevant factor in diabetes risk**. The relatively higher proportion of diabetics among former and current smokers could indicate a potential link between smoking and an increased risk of developing diabetes. This visualization emphasizes the importance of lifestyle factors, such as smoking habits, in understanding diabetes prevalence and possible risk factors.

* **5.2.3 Bivariate 3 Age vs Blood Glucose Level by Diabetes Status**



The scatter plot shows the relationship between age and blood glucose levels, divided by diabetes status. The x-axis represents age in years, while the y-axis shows blood glucose levels. The pink dots represent people without diabetes (0), and the green dots represent people with diabetes (1).

From the graph, can see that people with diabetes usually have higher blood glucose levels than those without diabetes. Most non-diabetic people have blood glucose levels below 200, while diabetic people often have levels above 250, with some going over 300. The pink dots are mostly below 200, showing that non-diabetic people have stable blood glucose levels. On the other hand, the green dots are more spread out, meaning diabetic people have a wider range of blood sugar levels.

Looking at different ages, blood glucose levels for non-diabetic people stay steady across all ages. For diabetic people, high blood glucose levels appear at all ages, meaning age does not seem to change blood sugar levels much in this group.

In conclusion the graph shows that **having diabetes is linked to high blood glucose levels**. People with diabetes have much higher blood sugar compared to those without diabetes. Also, **blood sugar levels remain high in diabetic people at all ages**, which means **managing blood sugar is important for all diabetic individuals, no matter their age.**

Coding :

# Load necessary library

install.packages("funModeling")

library(funModeling)

install.packages("tidyverse")

library(tidyverse)

install.packages("Himsc")

library(Hmisc)

install.packages('ggplot2')

library(ggplot2)

library(readr)

library(dplyr)

# TASK 3 DATA CLEANING

# STEP 1: Import data set from local storage

diabetes\_data <- read\_csv("diabetes\_prediction\_dataset.csv")

#Replace "No Info" with NA in the 'smoking\_history' column

diabetes\_data$smoking\_history[diabetes\_data$smoking\_history == "No Info"] <- NA

# STEP 2: Check Missing Values in Data Set

mydata <- df\_status(diabetes\_data)

# STEP 3: Check the whole data frame for missing values

complete.cases(diabetes\_data)

# STEP 4: Remove variables with of NAs

diabetes\_data <- na.omit(diabetes\_data)

# STEP 5: Check for missing values

complete.cases(diabetes\_data)

#STEP 6: Count fully complete rows

sum(complete.cases(diabetes\_data))

#STEP 7 : Create new data without missing values

cleaned\_data <- diabetes\_data[complete.cases(diabetes\_data),]

df\_status(cleaned\_data)

#STEP 8 : Write the cleaned data in CSV

write.csv(cleaned\_data, "cleaned\_data\_diabetes.csv", row.names=FALSE)

library(dplyr)

diabetes\_data <- diabetes\_data %>%

mutate(

hypertension = factor(hypertension, levels = c(0,1), labels = c("No", "Yes")),

heart\_disease = factor(heart\_disease, levels = c(0,1), labels = c("No", "Yes")),

diabetes = factor(diabetes, levels = c(0,1), labels = c("No", "Yes"))

)

#TASK 4

library(ggplot2)

library(viridis)

# call cleaned data csv

cleaned\_diabetes\_data <- read\_csv("cleaned\_data\_diabetes.csv")

# Univariate - Numerical: Histogram of BMI

ggplot(diabetes\_data, aes(x=bmi, fill=..count..)) +

geom\_histogram(binwidth=4, color="black", alpha=0.8) +

scale\_fill\_viridis(option="magma", direction=-1) +

labs(title="Distribution of BMI in Patients", x="BMI", y="Count") +

theme\_minimal() +

theme(plot.title = element\_text(hjust = 0.5, face="bold", size=14))

# Univariate - Categorical: Bar Chart of Smoking History

# Reorder smoking\_history levels from least to most harmful

diabetes\_data$smoking\_history <- factor(diabetes\_data$smoking\_history,

levels = c("never", "former", "not current", "ever", "current"))

# Create the bar chart with reordered categories

ggplot(diabetes\_data, aes(x=smoking\_history, fill=smoking\_history)) +

geom\_bar() +

scale\_fill\_brewer(palette="Set2") +

labs(

title = "Smoking History of Patients",

x = "Smoking History",

y = "Count",

caption = "Smoking History Categories:\n

• Never → Has never smoked\n

• Former → Used to smoke but quit\n

• Not Current → Just quit\n

• Ever → Has smoked at some point\n

• Current → Currently smokes"

) +

theme\_minimal() +

theme(

axis.text.x = element\_text(angle=45, hjust=1),

plot.title = element\_text(hjust = 0.5, face="bold", size=14),

plot.caption = element\_text(hjust=0, size=10, face="italic") # Align left & italic caption

)

# Univarite 3

# Load required libraries

library(ggplot2)

# Load the dataset

diabetes\_data <- read.csv("cleaned\_data\_diabetes.csv")

# Create a bar chart for gender distribution

ggplot(diabetes\_data, aes(x=gender, fill=gender)) +

geom\_bar(alpha=0.8) + # Adjust transparency for aesthetics

scale\_fill\_manual(values=c("#3498db", "#e74c3c", "#f39c12")) + # Added a third color

labs(

title="Gender Distribution of Patients",

x="Gender",

y="Count"

) +

theme\_minimal() +

theme(

plot.title = element\_text(hjust=0.5, face="bold", size=16, color="darkblue"),

axis.text = element\_text(size=12),

legend.position = "right" # Keep legend for clarity

)

# Bivariate - Numerical vs Categorical: Boxplot of Blood Glucose vs Diabetes

ggplot(diabetes\_data, aes(x=factor(diabetes), y=blood\_glucose\_level, fill=factor(diabetes))) +

geom\_boxplot(outlier.color="red", notch=TRUE, alpha=0.7) +

scale\_fill\_manual(values=c("brown", "lightpink")) +

labs(title="Blood Glucose Level vs Diabetes", x="Diabetes (0 = No, 1 = Yes)", y="Blood Glucose Level") +

theme\_minimal() +

theme(plot.title = element\_text(hjust = 0.5, face="bold", size=14))

# Bivariate - Categorical vs Categorical: Side-by-Side (Dodge) Bar Chart (Smoking History vs Diabetes)

ggplot(diabetes\_data, aes(x=smoking\_history, fill=factor(diabetes))) +

geom\_bar(position="dodge") + # Side-by-side bars instead of stacking

scale\_fill\_manual(values=c("purple", "lightpink")) +

labs(title="Smoking History vs Diabetes", x="Smoking History", y="Count", fill="Diabetes (0=No, 1=Yes)") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle=45, hjust=1),

plot.title = element\_text(hjust = 0.5, face="bold", size=14))

# Bivariate 3

# Load required libraries

library(ggplot2)

# Load the dataset

diabetes\_data <- read.csv("cleaned\_data\_diabetes.csv")

# Create a scatter plot: Age vs Blood Glucose Level

ggplot(diabetes\_data, aes(x=age, y=blood\_glucose\_level, color=factor(diabetes))) +

geom\_point(alpha=0.6, size=2) + # Scatter points with transparency

geom\_smooth(method="loess", se=FALSE, linetype="dashed", size=1.2) + # Smooth trend line

scale\_color\_manual(values=c("#FF69B4", "#2E8B57")) + # Aesthetic color scheme

labs(

title="Age vs Blood Glucose Level by Diabetes Status",

x="Age (Years)",

y="Blood Glucose Level",

color="Diabetes (0=No, 1=Yes)"

) +

theme\_minimal() +

theme(

plot.title = element\_text(hjust=0.5, face="bold", size=16, color="darkblue"),

axis.text = element\_text(size=12),

legend.position = "top"

)

)