PROJECT REPORT – SP24-IN-INFO-B518-20814
Prof. Zeyana Hamid

Diabetes Prediction Dataset
A Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data

Adarsh Viswanath, Amol Prakash, Keerthika Sunchu, Meghana Darla, Samantha Sanjeev
[viswana, amolprak, ksunchu, mdarla, fnsama] @iu.edu
Indiana University, Indianapolis, USA

Abstract. The study explores the impact of age, BMI, HbA1c levels, and blood glucose concentrations on the likelihood of developing diabetes within a diverse patient population, using a dataset of 100,000 individuals. Employing non-parametric Kruskal-Wallis and Dunn's post-hoc tests due to the non-normal distribution of the data, significant differences were found. These results confirm the critical role these variables play in the risk and management of diabetes and highlight their potential in enhancing predictive models for early detection and targeted treatment strategies. This study demonstrates the efficacy of integrating advanced statistical techniques and machine learning to improve diabetes prediction and inform public health strategies.

***Keywords**:* Diabetes, prediction model, statistical testing, logistic regression

1. Project Scope

1.1 Introduction

Diabetes mellitus is a global health concern characterized by insufficient insulin production or utilization, leading to metabolic imbalances and beta cell deterioration. Type 1 diabetes, commonly diagnosed in children and young adults, necessitates insulin therapy. Type 2 diabetes, often linked to insulin resistance, requires lifestyle modifications and medication (Ganie et al., 2022). Diabetes mellitus is influenced by a combination of factors including age, BMI, HbA1C levels, and blood glucose concentrations, with each playing a pivotal role in the development and management of this metabolic disorder (Fletcher et al., 2022). Recent research highlights a surge in utilizing machine learning for predicting type 2 diabetes mellitus, reflecting a growing interest in leveraging technology to tackle this urgent challenge (Ganie et al., 2022).

1.2 Aim

The primary aim of this project is to evaluate the predictive power of various demographic and medical factors—specifically age, BMI, HbA1c levels, and blood glucose concentrations—on the likelihood of developing diabetes.

1.3 Purpose

The purpose of our research is to rigorously analyze and interpret the relationship between critical risk factors and the prevalence of diabetes in a diverse population. Through this analysis, we intend to provide evidence-based insights that can inform both clinical and public health strategies, potentially leading to more effective diabetes management and prevention programs.

1.4 Research Question

What is the impact of age, BMI, HbA1C levels, and blood glucose concentrations, on the likelihood of developing diabetes in a diverse patient population?

1.5 Hypothesis

Null Hypothesis: There is no significant impact of age, BMI, HbA1C levels, and blood glucose concentrations, on the likelihood of developing diabetes in a diverse patient population.

Alternate Hypothesis: There is a significant impact of age, BMI, HbA1C levels, and blood glucose concentrations, on the likelihood of developing diabetes in a diverse patient population.

2. Methodology

2.1 Diabetes Dataset Description

Variable	Category	Description
gender	Categorical	"Male", "Female" or "Other"
age	Numerical	Age of the patient (0-80 years)
heart disease	Categorical	0 indicates the patient doesn't have hypertension, 1 indicates if the patient has hypertension.
smoking history	Categorical	“Not current”, “Former”, “No Information”, “Current”, “Never”
BMI	Numerical	“Underweight" or” Normal”, ”Overweight”, “Obese”
HbA1c_level	Numerical	Average blood sugar level over the past 2-3 months
blood_glucose_level	Numerical	Amount of glucose in the bloodstream at a given time
diabetes	Numerical	1 indicates the presence of diabetes and 0 indicates the absence of diabetes.

2.2 Data Collection and Extraction

One of the most crucial steps in creating a machine-learning model is data collection. After filtering the dataset from Kaggle, we resulted in the finding of Diabetes Prediction Dataset.
The first step involved downloading the dataset and importing it into R Studio. The working directory was set using the set.wd() function and the data was loaded into R using the read.csv() function. This dataset comprises 100,000 rows and 9 columns. The dependent variable is “Diabetes” which is being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes. The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level.
Data source: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/code

Categorical Variables	Numerical Variables
Gender	Age
Smoking History	BMI
Hypertension	HbA1c Level
Heart Disease	Blood Glucose Level
Diabetes	

2.3 Data Cleaning and Outlier Treatment

The absence of null values was verified, and boxplots were utilized to identify potential outliers. It was confirmed that there were no null values for all the variables, and outliers were detected for the variable BMI, HbA1c_level, and blood_glucose_level. To address outliers detected in the boxplot, the Interquartile Range (IQR) method was applied. This technique

involves calculating the range between the first quartile (Q1) and the third quartile (Q3) of the data distribution, known as the IQR.

3. Data Analysis

3.1 Descriptive Statistics

We performed descriptive statistics using a dataset of 90,387 cleaned records to elucidate factors influencing diabetes. Categorically, the variables include gender and smoking history, both of which are qualitative in nature, and the binary variable, diabetes, where a value of '1' denotes the presence and '0' the absence of diabetes, affecting nearly 4.973% of the participants. This indicates a relatively low prevalence of diagnosed diabetes within the dataset.

Numerical variables are extensively analyzed to understand demographic and health-related dynamics. The age of participants ranges widely from 0.08 to 80 years, with the mean and median ages being 41.41 and 42 years, respectively, suggesting a predominantly middle-aged demographic. This is crucial for assessing age-related risk factors associated with diabetes.

Moreover, binary indicators for hypertension and heart disease are present in 6.565% and 3.562% of the dataset, respectively. These conditions are significant as they often correlate with higher diabetes risks. The BMI of study participants varies from 14.71 to 38.50, with a mean value of 26.32, indicating a distribution skewed towards higher BMI levels, a known risk factor for diabetes. Additionally, the HbA1c levels, ranging from 3.5 to 8.2 with a mean of 5.456, and blood glucose levels spanning from 80 to 240 mg/dL with a mean of 134.6 mg/dL, provide insights into the metabolic status of the cohort. These glycemic indicators are critical as they reflect long-term glucose control and immediate blood sugar levels, respectively, both of which are instrumental in diagnosing diabetes. The analysis of these descriptive statistics serves as a foundation for further statistical testing and predictive modeling.

3.2. Logistic Regression

Following the assumption that the outcome variable is binary and a large sample size, we performed logistic regression models that explored the impact of individual predictors (age, BMI, HbA1c levels, and blood glucose levels) on the likelihood of developing diabetes. We additionally also performed a multivariate model combining all predictors to assess the collective impact of these variables.

Predictor	Intercept Estimate	Coefficient Estimate	Odds Ratio	P-Value	Increase in Odds	AIC	Interpretation
Age	-5.6094	0.0511	1.0524	< 2e-16	5.2% per year	31471	Each additional year of age increases the odds of having diabetes by approximately 5.2%. The model indicates a strong relationship between age and diabetes, suggesting aging as a significant risk factor.
BMI	-6.814	0.139	1.149	< 2e-16	14.9% per unit	-	Each unit increase in BMI is associated with a 14.9% increase in the odds of diabetes. This reflects the critical role of BMI as a predictor, highlighting its

							importance in diabetes risk management.
HbA1c Level	-18.094	2.481	11.958	< 2e-16	1195.8% per unit	-	A one-unit increase in HbA1c level results in an approximately 1196% increase in the odds of diabetes. This dramatic rise underscores the strong predictive power of HbA1c levels in diabetes diagnosis.
Blood Glucose	-7.343	0.029	1.030	< 2e-16	3% per unit	-	Each unit increase in blood glucose level leads to a 3% increase in the odds of developing diabetes. Although significant, the impact is smaller than that of HbA1c levels.
Multivariate	-	-	-	-	-	19847.2	The initial AIC value of 19847.2 is the baseline for the full model. We could interpret that all predictor variables are important for predicting diabetes, as removing any of them increases the AIC, which indicates a worse model fit. The variables with the largest impact on AIC when removed are HbA1c level and blood glucose level, suggesting they are the most important predictors of diabetes in the model.

3.3. Statistical Analysis Using Non-Parametric Tests:

To examine the relationship between various predictors and the presence of diabetes, we employed non-parametric statistical tests given the non-normal distribution of our data. The initial analysis involved the Kruskal-Wallis test, a robust method for determining statistical differences across groups defined by a categorical variable—in this case, diabetes status (presence or absence).

Upon obtaining significant results from the Kruskal-Wallis tests, we proceeded with Dunn's post-hoc tests to conduct pairwise comparisons between the groups. This was necessary to pinpoint the specific differences between diabetic and non-diabetic individuals across age, BMI, HbA1c levels, and blood glucose levels. To control the inflation of Type I error due to multiple testing, we applied the Holm method, stringent adjustment procedure ensuring the reliability of our findings.

Statistical Test	Variables	Interpretation
Kruskal-Wallis Test	Age	Significant differences in age distributions between diabetic and non-diabetic individuals ($p < 2.2e-16$). Older age groups exhibit higher diabetes risk, suggesting age as a crucial factor in diabetes development

Kruskal-Wallis Test	BMI	Significant differences in BMI between individuals with and without diabetes ($p < 2.2e-16$). Higher BMI levels are commonly associated with increased diabetes risk, highlighting the importance of weight management in diabetes prevention and management.
Kruskal-Wallis Test	HbA1c Level	Significant variations in HbA1c levels between diabetic and non-diabetic groups ($p < 2.2e-16$). Elevated HbA1c levels indicate poor long-term glucose control, crucial for diagnosing diabetes and monitoring glucose management.
Kruskal-Wallis Test	Blood Glucose Level	Significant differences in blood glucose levels between groups ($p < 2.2e-16$). Directly correlated with diabetes, elevated blood glucose levels underscore their critical role in diabetes diagnosis.
Dunn's Post-hoc Test	Age	Confirms significant age differences between diabetic and non-diabetic individuals. Highlights the need for age-specific diabetes management strategies.
Dunn's Post-hoc Test	BMI	Validates significant BMI differences, emphasizing obesity as a key modifiable risk factor for diabetes.
Dunn's Post-hoc Test	HbA1c Level	Reinforces significant differences in HbA1c levels, underscoring the need for regular monitoring of this parameter in populations at risk of diabetes.
Dunn's Post-hoc Test	Blood Glucose Level	Further confirms significant differences in blood glucose levels, supporting the use of glucose monitoring as a primary tool in diabetes management and prevention.

3.4. Interpretation

The Kruskal-Wallis tests demonstrated significant statistical differences across all tested variables between groups with and without diabetes, confirming the substantial impact of each variable on the likelihood of developing the disease:

- Age: Chi-squared = 3963.6, df = 1, p-value < 2.2e-16. This result indicates that older individuals have significantly higher instances of diabetes, confirming age as a critical risk factor. The extremely low p-value suggests that the observed association is highly unlikely to occur by chance.
- BMI: Chi-squared = 1639.3, df = 1, p-value < 2.2e-16. A higher BMI is strongly associated with an increased prevalence of diabetes, underscoring obesity as a major indicator of diabetes risk. The p-value reinforces the statistical significance of BMI as a predictor.
- HbA1c Levels: Chi-squared = 4941, df = 1, p-value < 2.2e-16. Elevated HbA1c levels, indicative of poor long-term glucose control, are significantly more common among individuals with diabetes. The p-value confirms the robustness of HbA1c levels as a predictor of diabetes.
- Blood Glucose Levels: Chi-squared = 2415.4, df = 1, p-value < 2.2e-16. Similarly, higher immediate blood glucose levels are significantly associated with diabetes, with the p-value affirming this variable's reliability as a diagnostic marker.

Dunn's post-hoc tests further substantiated these findings, confirming significant differences between the diabetic and non-diabetic groups for each variable, consistently yielding p-values < 2e-16. The consistency of these extremely low p-values across all variables provides

convincing evidence against the null hypothesis, suggesting a statistically significant impact of age, BMI, HbA1c levels, and blood glucose levels on diabetes occurrence. These results validate the alternative hypothesis and highlight the importance of these factors in diabetes screening and risk assessment strategies.

3.5. Data Visualization

Visualizations were created for a better understanding of the distribution of individual variables. We used "ggplot2" for visualizing the distribution and relationships among the cleaned variables.

4. Results

Our analysis revealed significant differences between diabetic and non-diabetic individuals across multiple variables based on our statistical testing:

Age: The statistical tests confirmed significant differences in age distributions, indicating that older individuals have a higher likelihood of developing diabetes ($p < 2.2e\text{-}16$).

BMI: Significant disparities were found in BMI levels between the two groups, with higher BMI strongly associated with an increased risk of diabetes ($p < 2.2e\text{-}16$).

HbA1c Levels: Our findings highlighted that elevated HbA1c levels, indicative of poor glucose control, were more prevalent among diabetic patients ($p < 2.2e\text{-}16$).

Blood Glucose Levels: Similarly, blood glucose levels were significantly higher in diabetic individuals compared to non-diabetic individuals ($p < 2.2e\text{-}16$).

5. Conclusion

The analysis conducted on the diabetes prediction dataset successfully demonstrated significant statistical associations between diabetes and several key factors, namely age, BMI, HbA1c levels, and blood glucose levels. The evidence strongly supports the alternative hypothesis that these factors significantly impact the likelihood of developing diabetes in a diverse patient population. This underscores the importance of targeted interventions that address these modifiable risk factors to reduce the prevalence and impact of diabetes. Moreover, the use of advanced statistical techniques and machine learning in analyzing medical and demographic data offers a promising avenue for enhancing predictive accuracy and developing personalized diabetes management strategies. By focusing on early identification and management of risk factors identified in this study, it is possible to mitigate the progression of diabetes and improve patient outcomes.

6. References

Fletcher, B., Gulanick, M., & Lamendola, C. (2002). Risk factors for type 2 diabetes mellitus. *The Journal of Cardiovascular Nursing, 16*(2), 17–23. https://doi.org/10.1097/00005082-200201000-00003

Ganie, S. M., Malik, M. B., & Arif, T. (2022). Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches. *Journal of Diabetes and Metabolic Disorders, 21*(1), 339–352. https://doi.org/10.1007/s40200-022-00981-w

Appendix

- Codes for loading and exploring the dataset.

```{r}
library(readr)
diabetes_prediction_dataset <- read_csv("C:/Users/Meghana/Downloads/diabetes_prediction_dataset.csv")
View(diabetes_prediction_dataset)
```

```
Rows: 100000 Columns: 9── Column specification
```

```{r}
#checking the number of rows and columns
dim(diabetes_prediction_dataset)
```

```{r}
#Viewing the first few rows of the dataset
head(diabetes_prediction_dataset)
```

```{r}
column_names = names(diabetes_prediction_dataset)
print(column_names)
```

```{r}
#checking for datastructures of the dataset
str(diabetes_prediction_dataset)
```

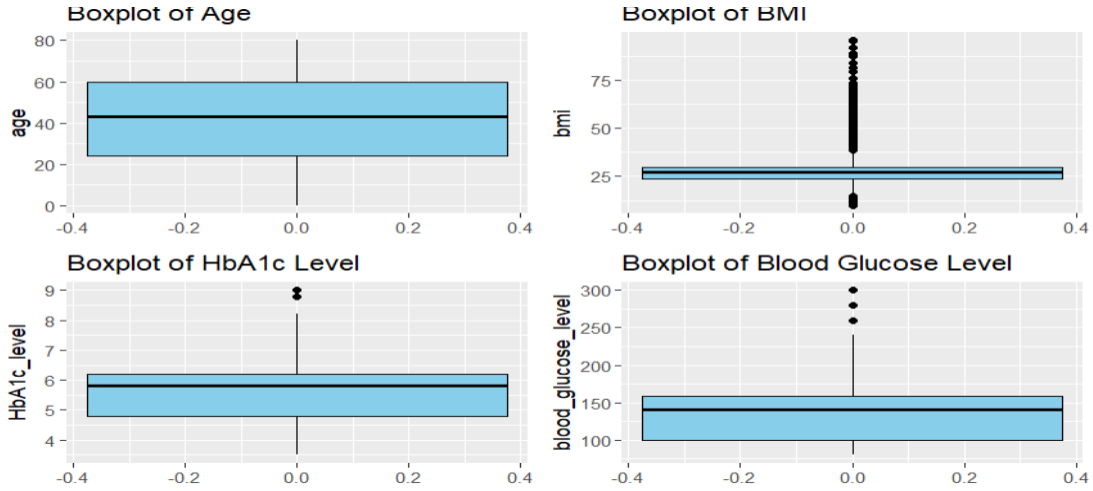
- Code for calculating the summary statistics of the dataset.

```{r}
# Summary of all the data attributes
summary(diabetes_prediction_dataset)
```

- Code for counting null values for each variable.

```
#Counting null values for each variable
```{r}
null_values_in_age <- sum(is.na(diabetes_prediction_dataset$age))
print(null_values_in_age)
null_values_in_bmi <- sum(is.na(diabetes_prediction_dataset$bmi))
print(null_values_in_bmi)
null_values_in_gender <- sum(is.na(diabetes_prediction_dataset$gender))
print(null_values_in_gender)
null_values_in_hypertension <- sum(is.na(diabetes_prediction_dataset$hypertension))
print(null_values_in_hypertension)
null_values_in_heart_disease <- sum(is.na(diabetes_prediction_dataset$heart_disease))
print(null_values_in_heart_disease)
null_values_in_smoking_history <- sum(is.na(diabetes_prediction_dataset$smoking_history))
print(null_values_in_smoking_history)
null_values_in_HbA1c_level <- sum(is.na(diabetes_prediction_dataset$HbA1c_level))
print(null_values_in_HbA1c_level)
null_values_in_blood_glucose_level <- sum(is.na(diabetes_prediction_dataset$blood_glucose_level))
print(null_values_in_blood_glucose_level)
null_values_in_diabetes <- sum(is.na(diabetes_prediction_dataset$diabetes))
print(null_values_in_diabetes)
```
```

- Code for checking outliers.

```
#Checking for outliers
```{r}
# Install patchwork package if not already installed
install.packages("patchwork")

# Load necessary libraries
library(ggplot2)
library(patchwork)

# Create boxplots for continuous variables
boxplot_age <- ggplot(diabetes_prediction_dataset, aes(y = age)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Boxplot of Age")

boxplot_bmi <- ggplot(diabetes_prediction_dataset, aes(y = bmi)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Boxplot of BMI")

boxplot_HbA1c <- ggplot(diabetes_prediction_dataset, aes(y = HbA1c_level)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Boxplot of HbA1c Level")

boxplot_blood_glucose <- ggplot(diabetes_prediction_dataset, aes(y = blood_glucose_level)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Boxplot of Blood Glucose Level")

# Arrange boxplots using patchwork
arranged_plots <- boxplot_age + boxplot_bmi + boxplot_HbA1c + boxplot_blood_glucose
arranged_plots <- arranged_plots + plot_layout(ncol = 2)

# Print the arranged plots
arranged_plots
```

- Code for displaying boxplots with outliers.

```
# Calculate and print outliers for each variable
outliers_age <- boxplot.stats(diabetes_prediction_dataset$age)$out
outliers_bmi <- boxplot.stats(diabetes_prediction_dataset$bmi)$out
outliers_HbA1c <- boxplot.stats(diabetes_prediction_dataset$HbA1c_level)$out
outliers_blood_glucose <- boxplot.stats(diabetes_prediction_dataset$blood_glucose_level)$out

print("Outliers in Age:")
print(outliers_age)

print("Outliers in BMI:")
print(outliers_bmi)

print("Outliers in HbA1c Level:")
print(outliers_HbA1c)

print("Outliers in Blood Glucose Level:")
print(outliers_blood_glucose)
```
```



- Codes for removing the outliers.

```
#Removing the outliers
```{r}
# Load the dplyr package
library(dplyr)

# Calculate the IQR for HbA1c level
Q1_HbA1c <- quantile(diabetes_prediction_dataset$HbA1c_level, 0.25, na.rm = TRUE)  # First Quartile (25th
percentile)
Q3_HbA1c <- quantile(diabetes_prediction_dataset$HbA1c_level, 0.75, na.rm = TRUE)  # Third Quartile (75th
percentile)
IQR_HbA1c <- Q3_HbA1c - Q1_HbA1c  # Interquartile Range

# Define outlier thresholds
lower_bound_HbA1c <- Q1_HbA1c - 1.5 * IQR_HbA1c
upper_bound_HbA1c <- Q3_HbA1c + 1.5 * IQR_HbA1c

# Filter out outliers for HbA1c level
cleaned_dataset <- diabetes_prediction_dataset %>%
  filter(HbA1c_level >= lower_bound_HbA1c & HbA1c_level <= upper_bound_HbA1c)

# Calculate the IQR for blood glucose level
Q1_blood_glucose <- quantile(diabetes_prediction_dataset$blood_glucose_level, 0.25, na.rm = TRUE)  # First
Quartile (25th percentile)
Q3_blood_glucose <- quantile(diabetes_prediction_dataset$blood_glucose_level, 0.75, na.rm = TRUE)  # Third
Quartile (75th percentile)
IQR_blood_glucose <- Q3_blood_glucose - Q1_blood_glucose  # Interquartile Range

# Define outlier thresholds
lower_bound_blood_glucose <- Q1_blood_glucose - 1.5 * IQR_blood_glucose
upper_bound_blood_glucose <- Q3_blood_glucose + 1.5 * IQR_blood_glucose

# Filter out outliers for blood glucose level
cleaned_dataset <- cleaned_dataset %>%
  filter(blood_glucose_level >= lower_bound_blood_glucose & blood_glucose_level <= upper_bound_blood_glucose)
```

```
# Calculate the IQR for BMI
Q1_bmi <- quantile(diabetes_prediction_dataset$bmi, 0.25, na.rm = TRUE)  # First Quartile (25th percentile)
Q3_bmi <- quantile(diabetes_prediction_dataset$bmi, 0.75, na.rm = TRUE)  # Third Quartile (75th percentile)
IQR_bmi <- Q3_bmi - Q1_bmi  # Interquartile Range

# Define outlier thresholds
lower_bound_bmi <- Q1_bmi - 1.5 * IQR_bmi
upper_bound_bmi <- Q3_bmi + 1.5 * IQR_bmi

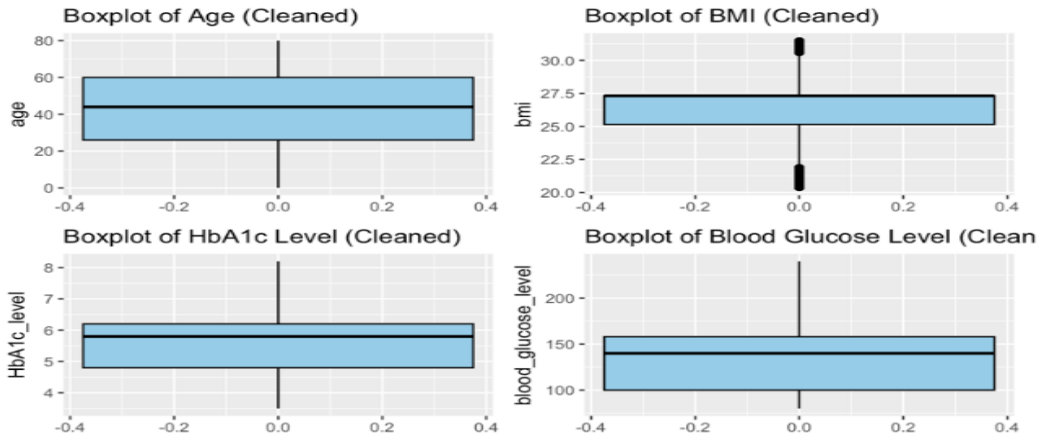
# Filter out outliers for BMI
cleaned_dataset <- cleaned_dataset %>%
  filter(bmi >= lower_bound_bmi & bmi <= upper_bound_bmi)

# View the dimensions of the cleaned dataset to confirm rows are dropped
dim(cleaned_dataset)

```
```

- Code for boxplots after removing outliers.

```r
# Load necessary libraries
library(ggplot2)
library(patchwork)
# Re-create the boxplots for the cleaned continuous variables
boxplot_age_clean <- ggplot(cleaned_dataset, aes(y = age)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Boxplot of Age (Cleaned)")

boxplot_bmi_clean <- ggplot(cleaned_dataset, aes(y = bmi)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Boxplot of BMI (Cleaned)")

boxplot_HbA1c_clean <- ggplot(cleaned_dataset, aes(y = HbA1c_level)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Boxplot of HbA1c Level (Cleaned)")

boxplot_blood_glucose_clean <- ggplot(cleaned_dataset, aes(y = blood_glucose_level)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Boxplot of Blood Glucose Level (Cleaned)")

# Arrange the cleaned boxplots using patchwork
arranged_plots_clean <- boxplot_age_clean + boxplot_bmi_clean + boxplot_HbA1c_clean + boxplot_blood_glucose_clean
arranged_plots_clean <- arranged_plots_clean + plot_layout(ncol = 2)

# Print the arranged plots
arranged_plots_clean

```



- Codes for calculating summary statistics for cleaned dataset

#Calculating summary statistics for the cleaned dataset

```r
#checking the number of rows and columns
dim(cleaned_dataset)
```

```r
head(cleaned_dataset)
```

```r
# Summary of all the data attributes
summary(cleaned_dataset)
```

- Codes for calculation frequency and percentage of each variable, along with central tendency and measures of dispersion

#Gender

```r
# Count or frequency of each gender
gender_frequency <- table(cleaned_dataset$gender)

# Displaying frequency of each gender
print("Frequency of each gender:")
print(gender_frequency)

# Percentage or proportion of each gender
gender_percentage <- prop.table(gender_frequency) * 100

# Displaying percentage or proportion of each gender
print("Percentage of each gender:")
print(gender_percentage)
```

#Hypertension

```r
# Count or frequency of hypertension
hypertension_frequency <- table(cleaned_dataset$hypertension)
# Percentage or proportion of hypertension
hypertension_percentage <- prop.table(hypertension_frequency) * 100

# Displaying frequency and percentage of hypertension
print("Frequency of Hypertension:")
print(hypertension_frequency)
print("Percentage of Hypertension:")
print(hypertension_percentage)
```

#Age
```{r}
# Measures of Central Tendency
mean_age <- mean(cleaned_dataset$age)
median_age <- median(cleaned_dataset$age)

# Displaying Measures of Central Tendency
print("Measures of Central Tendency:")
print(paste("Mean Age:", mean_age))
print(paste("Median Age:", median_age))

# Measures of Dispersion/Spread
range_age <- range(cleaned_dataset$age)
variance_age <- var(cleaned_dataset$age)
sd_age <- sd(cleaned_dataset$age)
iqr_age <- IQR(cleaned_dataset$age)

# Displaying Measures of Dispersion/Spread
print("Measures of Dispersion/Spread:")
print(paste("Range of Age:", range_age))
print(paste("Variance of Age:", variance_age))
print(paste("Standard Deviation of Age:", sd_age))
print(paste("Interquartile Range of Age:", iqr_age))

# Minimum and Maximum values
min_age <- min(cleaned_dataset$age)
max_age <- max(cleaned_dataset$age)

# Displaying Minimum and Maximum values
print("Minimum and Maximum Age:")
print(paste("Minimum Age:", min_age))
print(paste("Maximum Age:", max_age))

# Percentiles
percentiles_age <- quantile(cleaned_dataset$age, c(0.25, 0.5, 0.75))

print("Percentiles of Age:")
```

#Heart Disease
```{r}
# Count or frequency of heart disease
heart_disease_frequency <- table(cleaned_dataset$heart_disease)
# Percentage or proportion of heart disease
heart_disease_percentage <- prop.table(heart_disease_frequency) * 100

# Displaying frequency and percentage of heart disease
print("Frequency of Heart Disease:")
print(heart_disease_frequency)
print("Percentage of Heart Disease:")
print(heart_disease_percentage)
```

#Smoking History
```{r}
# Count or frequency of smoking history
smoking_history_frequency <- table(cleaned_dataset$smoking_history)
# Percentage or proportion of smoking history
smoking_history_percentage <- prop.table(smoking_history_frequency) * 100

# Displaying frequency and percentage of smoking history
print("Frequency of Smoking History:")
print(smoking_history_frequency)
print("Percentage of Smoking History:")
print(smoking_history_percentage)
```

```r
# Measures of Central Tendency for BMI
mean_bmi <- mean(cleaned_dataset$bmi)
median_bmi <- median(cleaned_dataset$bmi)

# Displaying Measures of Central Tendency for BMI
print("Measures of Central Tendency for BMI:")
print(paste("Mean BMI:", mean_bmi))
print(paste("Median BMI:", median_bmi))

# Measures of Dispersion/Spread for BMI
range_bmi <- range(cleaned_dataset$bmi)
variance_bmi <- var(cleaned_dataset$bmi)
sd_bmi <- sd(cleaned_dataset$bmi)
iqr_bmi <- IQR(cleaned_dataset$bmi)

# Displaying Measures of Dispersion/Spread for BMI
print("Measures of Dispersion/Spread for BMI:")
print(paste("Range of BMI:", range_bmi))
print(paste("Variance of BMI:", variance_bmi))
print(paste("Standard Deviation of BMI:", sd_bmi))
print(paste("Interquartile Range of BMI:", iqr_bmi))

# Minimum and Maximum values for BMI
min_bmi <- min(cleaned_dataset$bmi)
max_bmi <- max(cleaned_dataset$bmi)

# Displaying Minimum and Maximum values for BMI
print("Minimum and Maximum BMI:")
print(paste("Minimum BMI:", min_bmi))
print(paste("Maximum BMI:", max_bmi))

# Percentiles for BMI
percentiles_bmi <- quantile(cleaned_dataset$bmi, c(0.25, 0.5, 0.75))

# Displaying Percentiles for BMI
print("Percentiles of BMI:")
print(percentiles_bmi)
```

```r
# Measures of Central Tendency for HbA1c Level
mean_hba1c <- mean(cleaned_dataset$HbA1c_level)
median_hba1c <- median(cleaned_dataset$HbA1c_level)

# Displaying Measures of Central Tendency for HbA1c Level
print("Measures of Central Tendency for HbA1c Level:")
print(paste("Mean HbA1c Level:", mean_hba1c))
print(paste("Median HbA1c Level:", median_hba1c))

# Measures of Dispersion/Spread for HbA1c Level
range_hba1c <- range(cleaned_dataset$HbA1c_level)
variance_hba1c <- var(cleaned_dataset$HbA1c_level)
sd_hba1c <- sd(cleaned_dataset$HbA1c_level)
iqr_hba1c <- IQR(cleaned_dataset$HbA1c_level)

# Displaying Measures of Dispersion/Spread for HbA1c Level
print("Measures of Dispersion/Spread for HbA1c Level:")
print(paste("Range of HbA1c Level:", range_hba1c))
print(paste("Variance of HbA1c Level:", variance_hba1c))
print(paste("Standard Deviation of HbA1c Level:", sd_hba1c))
print(paste("Interquartile Range of HbA1c Level:", iqr_hba1c))

# Minimum and Maximum values for HbA1c Level
min_hba1c <- min(cleaned_dataset$HbA1c_level)
max_hba1c <- max(cleaned_dataset$HbA1c_level)

# Displaying Minimum and Maximum values for HbA1c Level
print("Minimum and Maximum HbA1c Level:")
print(paste("Minimum HbA1c Level:", min_hba1c))
print(paste("Maximum HbA1c Level:", max_hba1c))

# Percentiles for HbA1c Level
percentiles_hba1c <- quantile(cleaned_dataset$HbA1c_level, c(0.25, 0.5, 0.75))

# Displaying Percentiles for HbA1c Level
print("Percentiles of HbA1c Level:")
print(percentiles_hba1c)
```

```r
# Measures of Central Tendency for Blood Glucose Level
mean_blood_glucose <- mean(cleaned_dataset$blood_glucose_level)
median_blood_glucose <- median(cleaned_dataset$blood_glucose_level)

# Displaying Measures of Central Tendency for Blood Glucose Level
print("Measures of Central Tendency for Blood Glucose Level:")
print(paste("Mean Blood Glucose Level:", mean_blood_glucose))
print(paste("Median Blood Glucose Level:", median_blood_glucose))

# Measures of Dispersion/Spread for Blood Glucose Level
range_blood_glucose <- range(cleaned_dataset$blood_glucose_level)
variance_blood_glucose <- var(cleaned_dataset$blood_glucose_level)
sd_blood_glucose <- sd(cleaned_dataset$blood_glucose_level)
iqr_blood_glucose <- IQR(cleaned_dataset$blood_glucose_level)

# Displaying Measures of Dispersion/Spread for Blood Glucose Level
print("Measures of Dispersion/Spread for Blood Glucose Level:")
print(paste("Range of Blood Glucose Level:", range_blood_glucose))
print(paste("Variance of Blood Glucose Level:", variance_blood_glucose))
print(paste("Standard Deviation of Blood Glucose Level:", sd_blood_glucose))
print(paste("Interquartile Range of Blood Glucose Level:", iqr_blood_glucose))

# Minimum and Maximum values for Blood Glucose Level
min_blood_glucose <- min(cleaned_dataset$blood_glucose_level)
max_blood_glucose <- max(cleaned_dataset$blood_glucose_level)

# Displaying Minimum and Maximum values for Blood Glucose Level
print("Minimum and Maximum Blood Glucose Level:")
print(paste("Minimum Blood Glucose Level:", min_blood_glucose))
print(paste("Maximum Blood Glucose Level:", max_blood_glucose))

# Percentiles for Blood Glucose Level
percentiles_blood_glucose <- quantile(cleaned_dataset$blood_glucose_level, c(0.25, 0.5, 0.75))

# Displaying Percentiles for Blood Glucose Level
print("Percentiles of Blood Glucose Level:")
print(percentiles_blood_glucose)
```

#Diabetes

```r
# Count or frequency of diabetes
diabetes_frequency <- table(cleaned_dataset$diabetes)
print("Frequency of Diabetes:")
print(diabetes_frequency)

# Percentage or proportion of diabetes
diabetes_percentage <- prop.table(diabetes_frequency) * 100
print("Percentage of Diabetes:")
print(diabetes_percentage)
```

- Codes for visualization

#Visualization for the cleaned dataset

```r
library(ggplot2)
# Gender (Categorical Variable)
ggplot(cleaned_dataset, aes(x = gender)) +
  geom_bar(fill = "#99CCFF", color = "black") +  # Custom color for bars
  labs(title = "Distribution of Gender")
```

```r
# Age (Continuous Variable)
ggplot(cleaned_dataset, aes(x = age)) +
  geom_histogram(fill = "#FF9999", color = "black", bins = 30) +  # Custom color for bars
  labs(title = "Distribution of Age")
```

```r
# Hypertension (Categorical Variable)
ggplot(cleaned_dataset, aes(x = factor(hypertension))) +
  geom_bar(fill = "#66CC99", color = "black") +  # Custom color for bars
  labs(title = "Distribution of Hypertension")

# Heart Disease (Categorical Variable)
ggplot(cleaned_dataset, aes(x = factor(heart_disease))) +
  geom_bar(fill = "#FFCC66", color = "black") +  # Custom color for bars
  labs(title = "Distribution of Heart Disease")

# Smoking History (Categorical Variable)
ggplot(cleaned_dataset, aes(x = smoking_history)) +
  geom_bar(fill = "#FF99CC", color = "black") +  # Custom color for bars
  labs(title = "Distribution of Smoking History")
```

```r
# Create a pie chart for gender
pie(table(diabetes_prediction_dataset$gender),
    main = "Distribution of gender",
    col = c("cyan", "pink"),
    labels = paste(names(table(diabetes_prediction_dataset$gender)), "\n",
table(diabetes_prediction_dataset$gender), sep = ""))
```

```r
# Create a pie chart for Smoking history
pie(table(cleaned_dataset$smoking_history),
    main = "Distribution of smoking history",
    col = c("lightblue", "pink", "yellow", "black", "cyan", "green"),
    labels = paste(names(table(cleaned_dataset$smoking_history)), "\n", table(cleaned_dataset$smoking_history),
sep = ""))
```

```r
# BMI (Continuous Variable)
ggplot(cleaned_dataset, aes(x = bmi)) +
  geom_histogram(fill = "#99CCFF", color = "black", bins = 30) +  # Custom color for bars
  labs(title = "Distribution of BMI")

# HbA1c_level (Continuous Variable)
ggplot(cleaned_dataset, aes(x = HbA1c_level)) +
  geom_histogram(fill = "#FF9999", color = "black", bins = 30) +  # Custom color for bars
  labs(title = "Distribution of HbA1c Level")

# Blood Glucose Level (Continuous Variable)
ggplot(cleaned_dataset, aes(x = blood_glucose_level)) +
  geom_histogram(fill = "#66CC99", color = "black", bins = 30) +  # Custom color for bars
  labs(title = "Distribution of Blood Glucose Level")
```

- Codes for testing normality of the data

```r
#Normality Testing
```{r}
# Perform Shapiro-Wilk test on various columns
shapiro_age <- shapiro.test(cleaned_dataset$age)
shapiro_bmi <- shapiro.test(cleaned_dataset$bmi)
shapiro_hba1c <- shapiro.test(cleaned_dataset$HbA1c_level)
shapiro_glucose <- shapiro.test(cleaned_dataset$blood_glucose_level)

# Print the results
print("Shapiro-Wilk Test for Age:")
print(shapiro_age)

print("Shapiro-Wilk Test for BMI:")
print(shapiro_bmi)

print("Shapiro-Wilk Test for HbA1c Level:")
print(shapiro_hba1c)

print("Shapiro-Wilk Test for Blood Glucose Level:")
print(shapiro_glucose)
```
```

- Codes for Q-Q plots and histograms

```r
# Histogram for Age with Normal Curve
hist(cleaned_dataset$age, freq = FALSE, main = "Age Distribution")
curve(dnorm(x, mean = mean(cleaned_dataset$age), sd = sd(cleaned_dataset$age)), add = TRUE, col = "red")

# Histogram for BMI with Normal Curve
hist(cleaned_dataset$bmi, freq = FALSE, main = "BMI Distribution")
curve(dnorm(x, mean = mean(cleaned_dataset$bmi), sd = sd(cleaned_dataset$bmi)), add = TRUE, col = "red")

# Histogram for HbA1c Level with Normal Curve
hist(cleaned_dataset$HbA1c_level, freq = FALSE, main = "HbA1c Level Distribution")
curve(dnorm(x, mean = mean(cleaned_dataset$HbA1c_level), sd = sd(cleaned_dataset$HbA1c_level)), add = TRUE, col =
"red")

# Histogram for Blood Glucose Level with Normal Curve
hist(cleaned_dataset$blood_glucose_level, freq = FALSE, main = "Blood Glucose Level Distribution")
curve(dnorm(x, mean = mean(cleaned_dataset$blood_glucose_level), sd = sd(cleaned_dataset$blood_glucose_level)),
add = TRUE, col = "red")
```

```r
#Creating Q-Q plots and histograms with normal curve for assessing normality.
# Q-Q Plot for Age
# Function to create Q-Q plots
create_qq_plot <- function(data, column) {
  qqnorm(data[[column]], main = paste("Q-Q Plot of", column))
  qqline(data[[column]], col = "red", lwd = 2)
}

# Create Q-Q plots for each continuous variable
create_qq_plot(cleaned_dataset, "age")
create_qq_plot(cleaned_dataset, "bmi")
create_qq_plot(cleaned_dataset, "HbA1c_level")
create_qq_plot(cleaned_dataset, "blood_glucose_level")
```

- Codes for logistic regression model

```r
# Logistic regression with diabetes as the binary outcome and age as the predictor
logistic_model1 <- glm(diabetes ~ age, data = cleaned_dataset, family = binomial())

# Summary of the logistic model
summary(logistic_model1)

# Calculate the odds ratios from the model's coefficients
odds_ratios <- exp(coef(logistic_model1))

# Print the odds ratios
print(odds_ratios)
```

```r
# Logistic regression with diabetes as the binary outcome and bmi as the predictor
logistic_model6 <- glm(diabetes ~ bmi, data = cleaned_dataset, family = binomial())

# Summary of the logistic model
summary(logistic_model6)

# Calculate the odds ratios from the model's coefficients
odds_ratios <- exp(coef(logistic_model6))

# Print the odds ratios
print(odds_ratios)
```

```r
# Logistic regression with diabetes as the binary outcome and HbA1c as the predictor
logistic_model7 <- glm(diabetes ~ HbA1c_level, data = cleaned_dataset, family = binomial())

# Summary of the logistic model
summary(logistic_model7)

# Calculate the odds ratios from the model's coefficients
odds_ratios <- exp(coef(logistic_model7))

# Print the odds ratios
print(odds_ratios)
```

```r
# Logistic regression with diabetes as the binary outcome and blood glucose as the predictor
logistic_model8 <- glm(diabetes ~ blood_glucose_level, data = cleaned_dataset, family = binomial())

# Summary of the logistic model
summary(logistic_model8)

# Calculate the odds ratios from the model's coefficients
odds_ratios <- exp(coef(logistic_model8))

# Print the odds ratios
print(odds_ratios)
```

- Multivariate logistic regression code

```r
# Multivariate Logistic Regression Model
full_model <- glm(diabetes ~ age + gender + hypertension + heart_disease + smoking_history + bmi + HbA1c_level +
blood_glucose_level,
                  family = binomial(), data = cleaned_dataset)


stepwise_model <- step(full_model, direction="both")

# Print the summary of the refined model
summary(stepwise_model)
```

- Code for non-parametric testing (Kruskal Wallis)

```r
# Load necessary library
library(stats)

# Perform Kruskal-Wallis test for age across different levels of diabetes
kruskal_result_age <- kruskal.test(age ~ diabetes, data = cleaned_dataset)
print(kruskal_result_age)

# Perform Kruskal-Wallis test for BMI across different levels of diabetes
kruskal_result_bmi <- kruskal.test(bmi ~ diabetes, data = cleaned_dataset)
print(kruskal_result_bmi)

# Perform Kruskal-Wallis test for HbA1c level across different levels of diabetes
kruskal_result_HbA1c <- kruskal.test(HbA1c_level ~ diabetes, data = cleaned_dataset)
print(kruskal_result_HbA1c)

# Perform Kruskal-Wallis test for blood glucose level across different levels of diabetes
kruskal_result_blood_glucose <- kruskal.test(blood_glucose_level ~ diabetes, data = cleaned_dataset)
print(kruskal_result_blood_glucose)

kruskal_result_blood_glucose <- kruskal.test(smoking_history ~ diabetes, data = cleaned_dataset)
print(kruskal_result_blood_glucose)
```

- Code for Dunn's post-hoc test

```{r}
# Install and load necessary library
install.packages("PMCMRplus")
library(PMCMRplus)

# Perform Kruskal-Wallis test for age across different levels of diabetes
kruskal_result_age <- kruskal.test(age ~ diabetes, data = cleaned_dataset)

# Check the result
print(kruskal_result_age)

# Perform post-hoc analysis for age using pairwise Wilcoxon rank sum tests with Bonferroni adjustment
posthoc_age <- PMCMRplus::kwAllPairsDunnTest(cleaned_dataset$age, cleaned_dataset$diabetes, method =
"bonferroni")

# Print the post-hoc analysis result
print(posthoc_age)

# Similarly, perform post-hoc analysis for other variables like BMI, HbA1c level, and blood glucose level
posthoc_bmi <- PMCMRplus::kwAllPairsDunnTest(cleaned_dataset$bmi, cleaned_dataset$diabetes, method =
"bonferroni")
posthoc_HbA1c <- PMCMRplus::kwAllPairsDunnTest(cleaned_dataset$HbA1c_level, cleaned_dataset$diabetes, method =
"bonferroni")
posthoc_blood_glucose <- PMCMRplus::kwAllPairsDunnTest(cleaned_dataset$blood_glucose_level,
cleaned_dataset$diabetes, method = "bonferroni")

# Print the post-hoc analysis results for BMI, HbA1c level, and blood glucose level
print(posthoc_bmi)
print(posthoc_HbA1c)
print(posthoc_blood_glucose)
```