Report on Analysis of Pima Indians Diabetes Dataset

2024-12-12

#1**Introduction**

The Pima Indians Diabetes dataset is useful in exploring the health risk factors attached to diabetes within Pima Indian women. Diabetes is a present day global problem that has affected millions of people, with some communities like that of the Pima Indians being most affected. Because of the various genetic, environmental, and lifestyle factors, which are involved in the development of the disease, it remains to be a major area interest to most public health systems (Forouhi & Wareham, 2010). The identification of the frequent predictors and risk factors with the specific high-risk groups can result in effective prevention and reduction of the international burden of diabetes. It is within this context that this study aims to establish the significant determinants of diabetes, comprising of glucose parameters, BMIs, insulin levels, and other various measures in a bid to enhance the knowledge concerning the disorders that characterize diabetes. This paper considers these relationships by using a dataset including various factors like pregnancies, blood pressure, and diabetes pedigree that lead to diabetes development (Katsarou *et al.*, 2017).

As a result, this analysis using exploratory data analysis and summary statistics and predictive modeling seeks to identify the best recommendations to be offered to the relevant authorities concerning the policies and interventions recommended in the prevention and management of diabetes. Therefore, the results of this study can significantly extend existing knowledge of diabetes risk profiles, specifically, facilitate designing early identification, adequate treatment, and behavioral change approaches. However, this research should be used to indicate the importance of data in solving one of the biggest health challenges in the modern world.

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## Loading required package: lattice
##
##
## Attaching package: 'caret'
##
##
## The following object is masked from 'package:purrr':
##
##     lift
##
##
##
## Attaching package: 'reshape2'
##
```

```
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
## 'data.frame':    768 obs. of  9 variables:
##  $ Pregnancies      : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose          : int  148 85 183 89 137 116 78 115 197 125 ...
##  $ Blood.Pressure   : int  72 66 64 66 40 74 50 0 70 96 ...
##  $ Skin.Thickness   : int  35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin          : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI              : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ Diabetes.Pedigree: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age              : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ Outcome          : int  1 0 1 0 1 0 1 0 1 1 ...
```

#2**Data Description and Source**

The dataset contains 768 observations, each comprising nine health-related attributes: pregnancies, plasma glucose, blood pressure, SI, diabetes pedigree, age, the binary value is equal to 1 if a woman developed diabetes and 0 otherwise (DeFronzo *et al.*, 2015). It is possible to outline the outcome variable as a crucial tool helping to differentiate between diabetic and non-diabetic people and perform more targeted analysis.

The data was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases accessible in the kaggle website in order to investigate the rate of diabetes in a population that has a high risk outcome of developing the disease (Pima Indians diabetes database, 2000). This is a strong point of this dataset as it complies with all the FAIR principles: the data set is findable, accessible, interoperable, and reusable. These principles improve the usefulness of the current dataset as a reference for international research. The ethical factors were also enabled to observe how the provided dataset will be of benefit to everyone and particularly when addressing issues to do with health to the vulnerable population. The dataset's rich diversity, particularly in glucose and BMI values, provides a robust basis for statistical and predictive analysis, allowing for the exploration of different relationships between health indicators and diabetes.

#3**Summary Statistics**

```
## [1] 0
```

```
##       Pregnancies           Glucose    Blood.Pressure    Skin.Thickness
##                 0                 0                 0                 0
##           Insulin               BMI Diabetes.Pedigree               Age
##                 0                 0                 0                 0
##           Outcome
##                 0
```

```
##   Pregnancies        Glucose      Blood.Pressure   Skin.Thickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        Diabetes.Pedigree      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780    Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437    1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725    Median :29.00
```

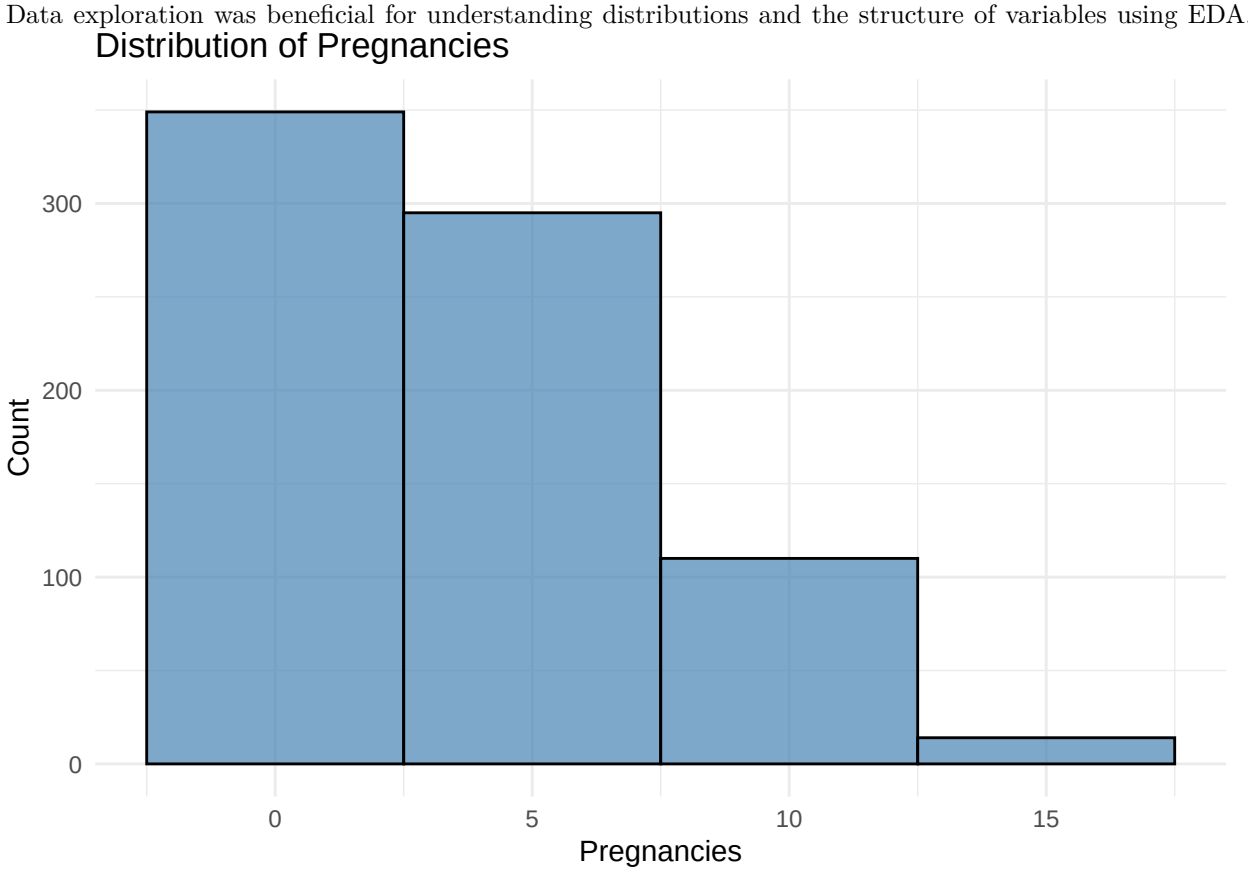
```
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719    Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262    3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200    Max.   :81.00
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

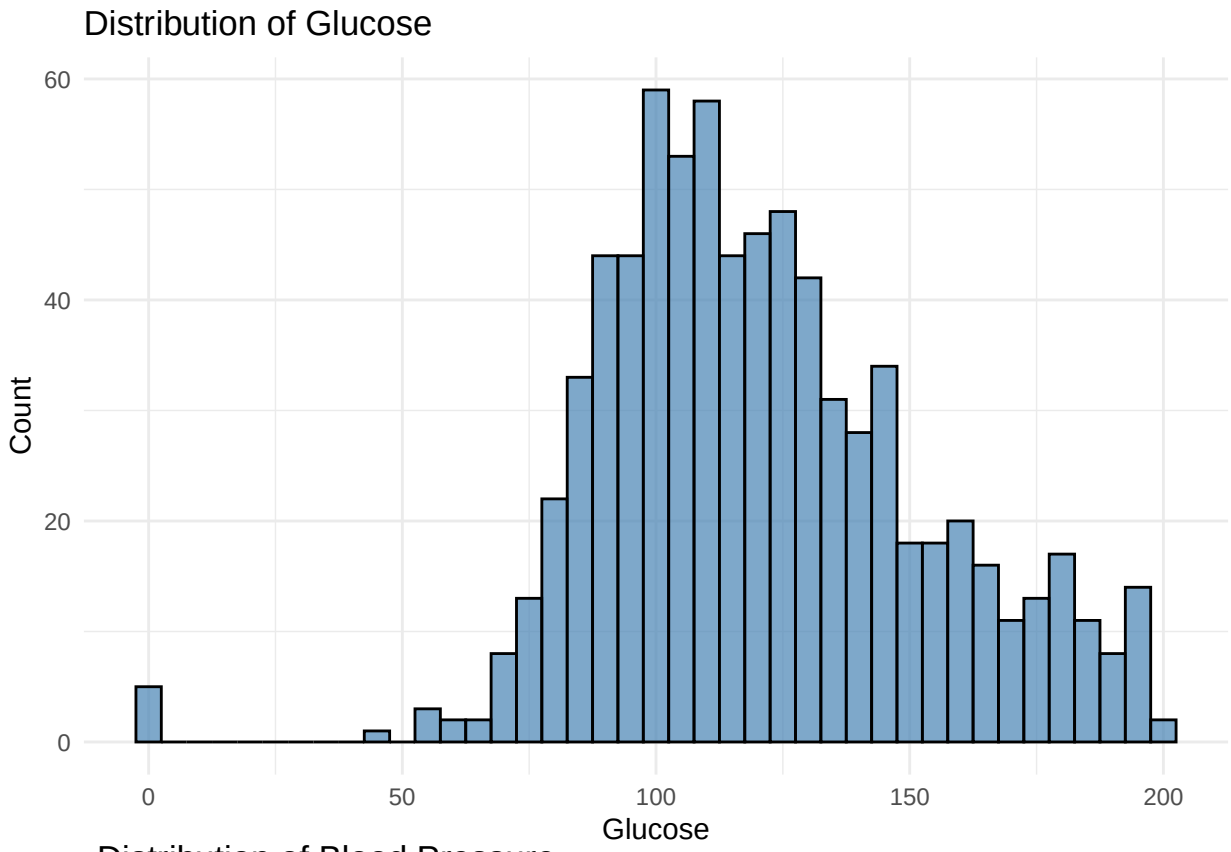
The dataset revealed high volatility in the selected dataset that is pertinent for estimating the likelihood of developing diabetes. For example, glucose level with a mean of roughly 120 and standard deviation of 31 implying that glycemic control fluctuates quite greatly among the participants. Participants represented a group of varied nutritional status with average BMI of approximately 32, from underweight to morbid obesity. Similarly, the number of pregnancies varies significantly from zero to seventeen, meaning that their reproductive career has been diverse.

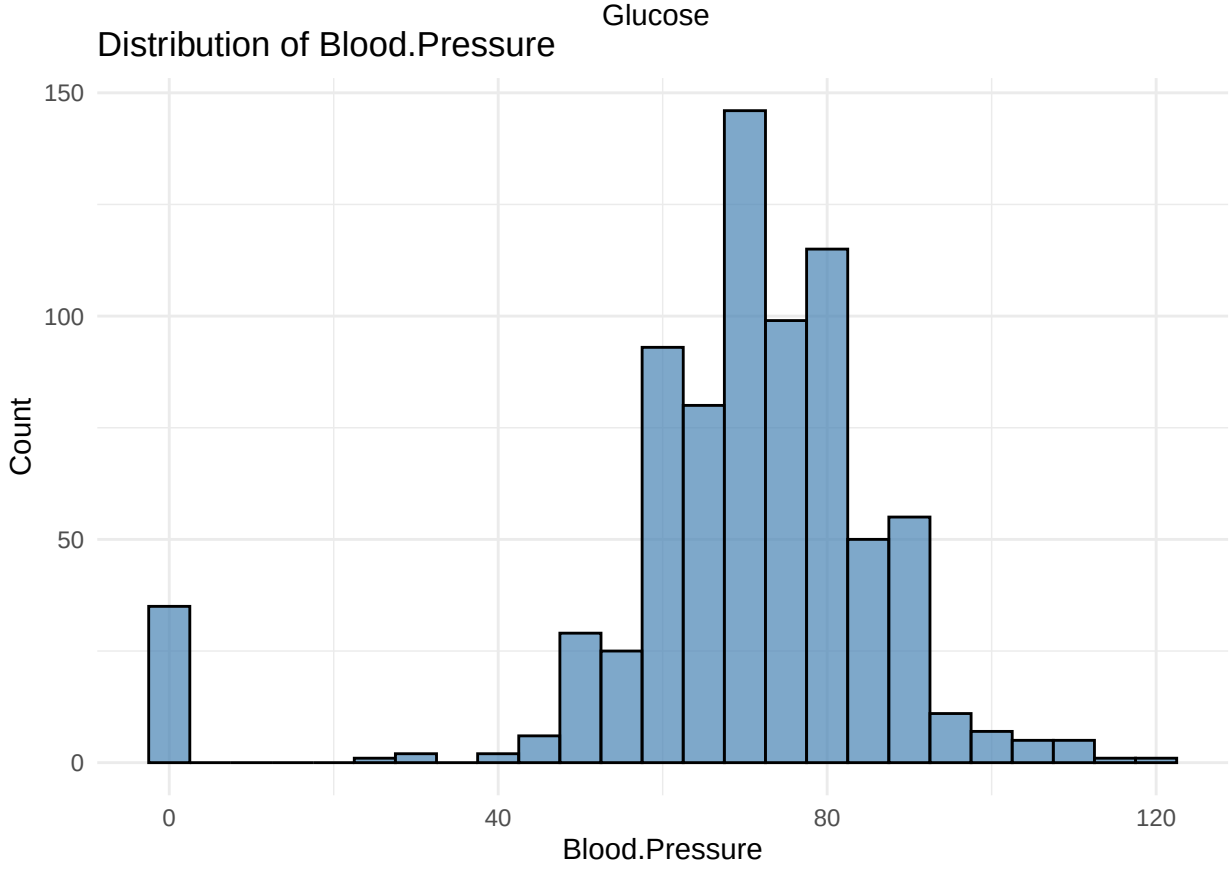
Other traits with such a large variance include blood pressure, skin measurement, and insulin levels, indicating that the determination of the risk of diabetes is not easy. The distribution of the binary outcome variable is almost equal, as about 35 percent of participants in the dataset were diagnosed with diabetes. This balance also improves the robustness of the predictive model because this approach avoids the risk of overemphasising between diabetic and non-diabetic categories. The dispersion of data also differs, which means that statistical models are formed with greater reliability, indicating the dependence between indicators and diabetes.

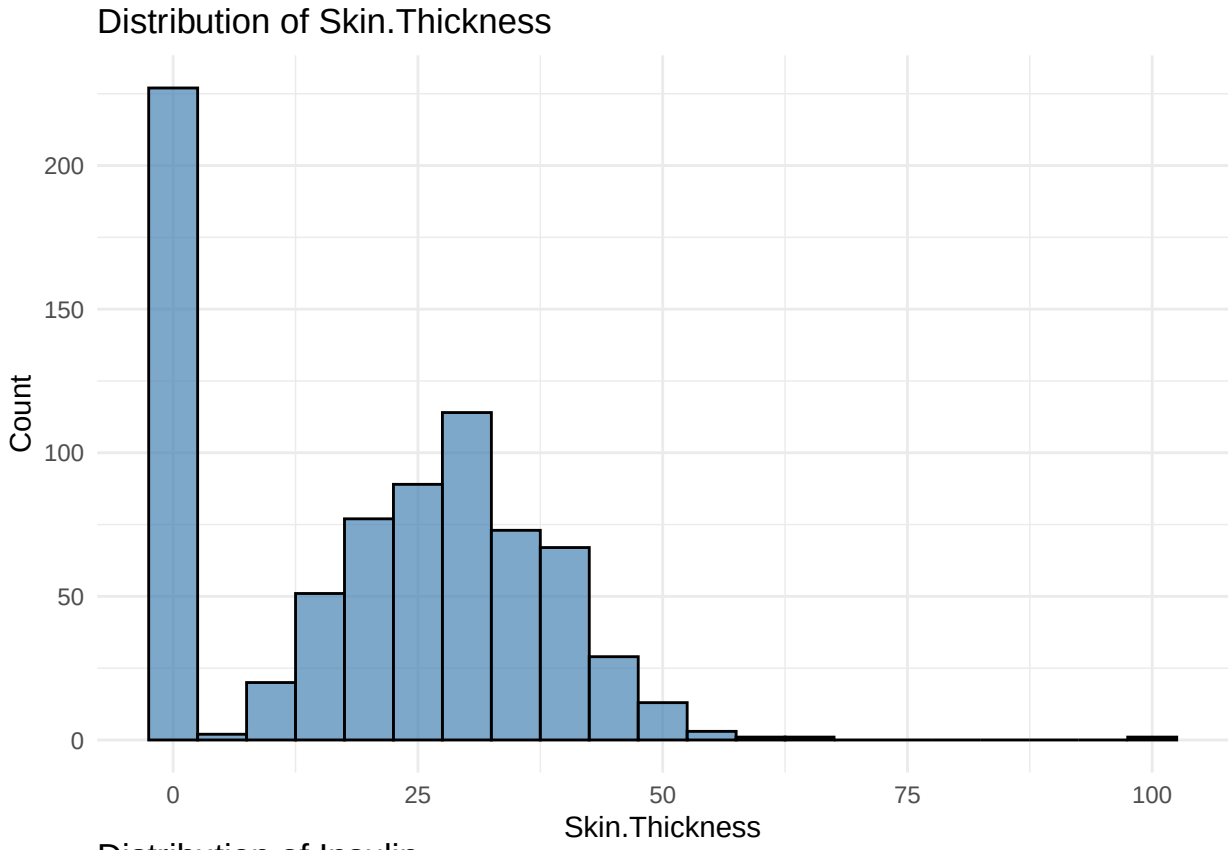
#4**Exploratory Data Analysis**

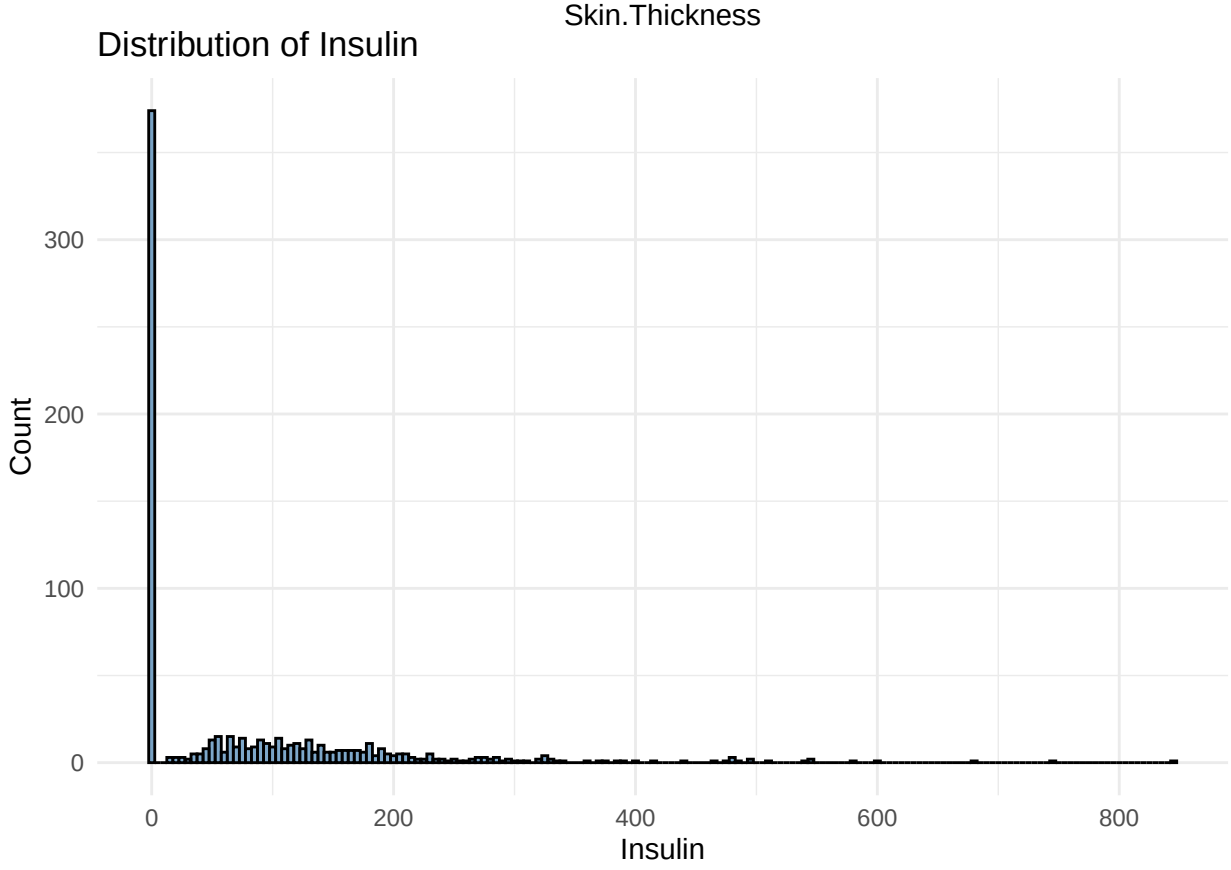
Data exploration was beneficial for understanding distributions and the structure of variables using EDA.

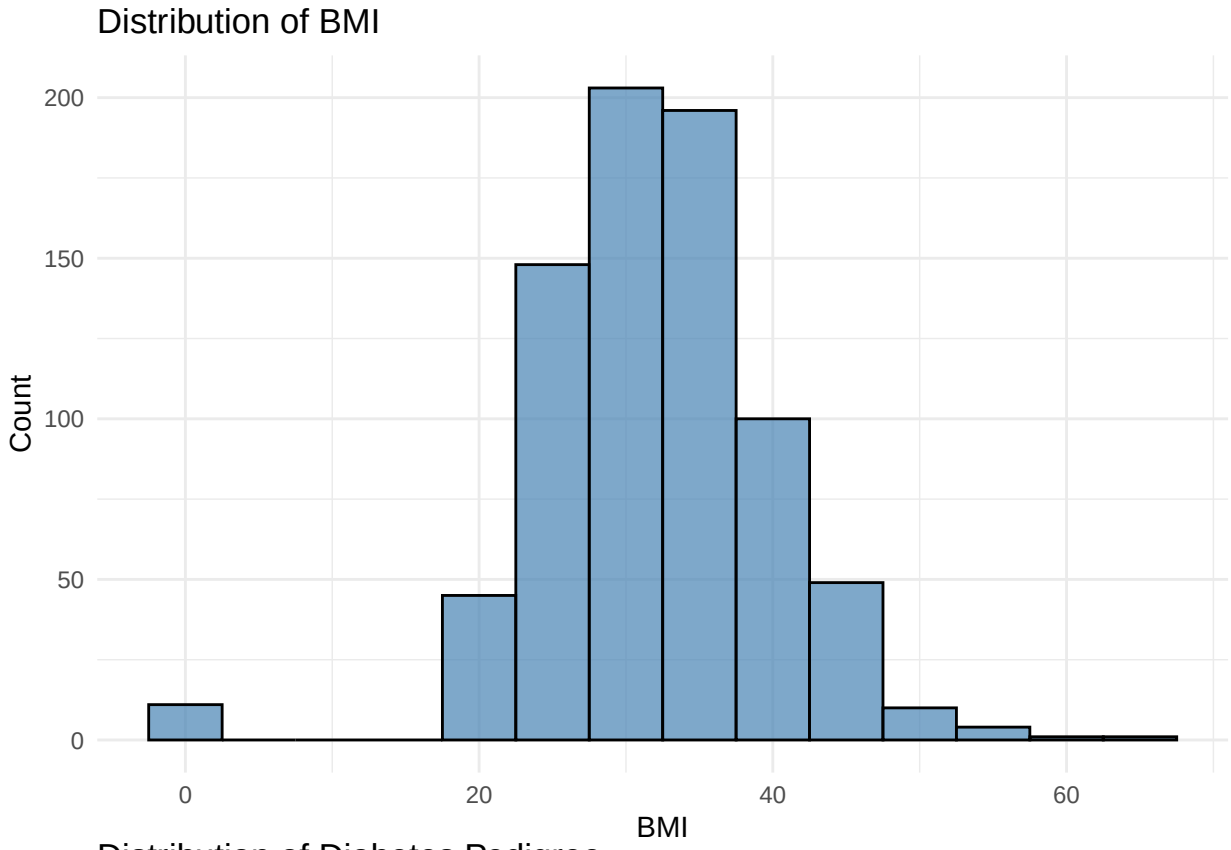


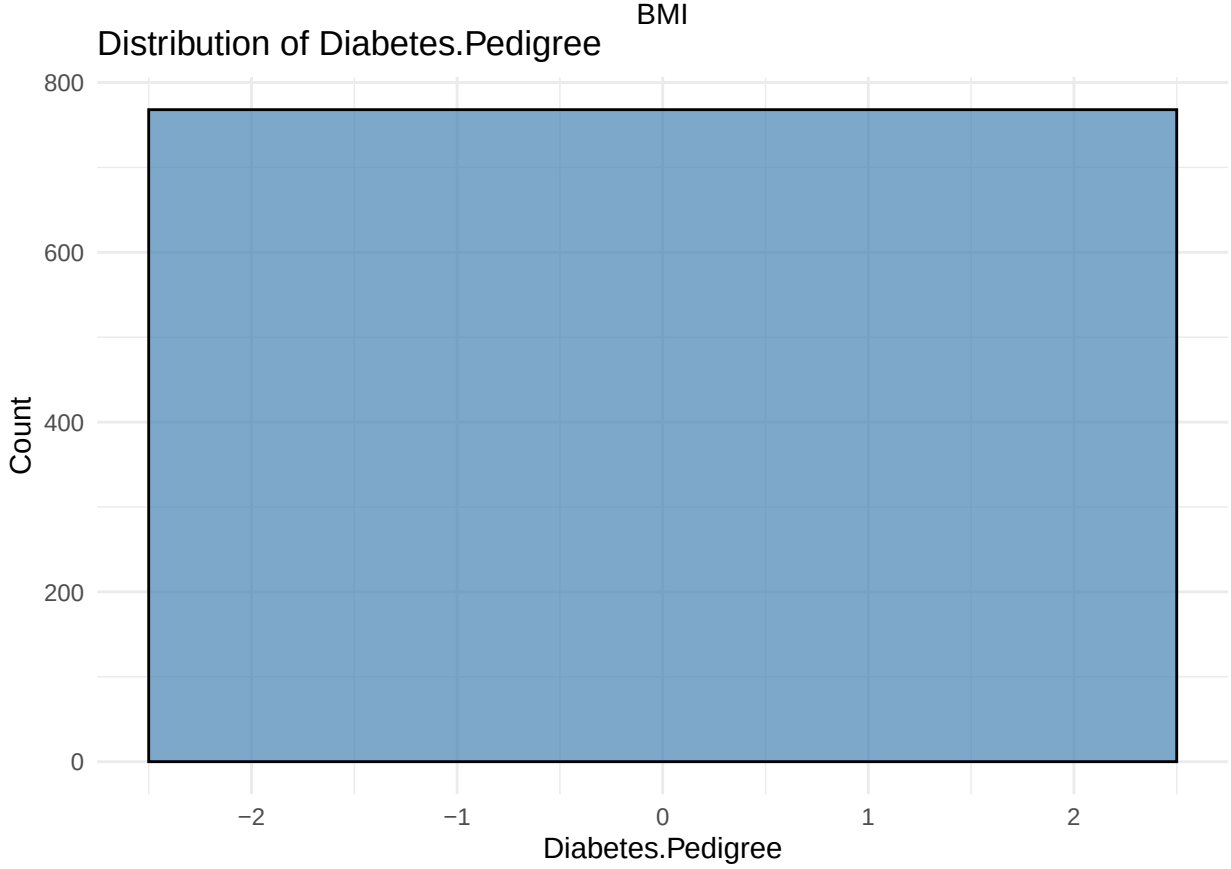

Distribution of Glucose
Count
60
40
20
0
0
50
100
150
200
Glucose

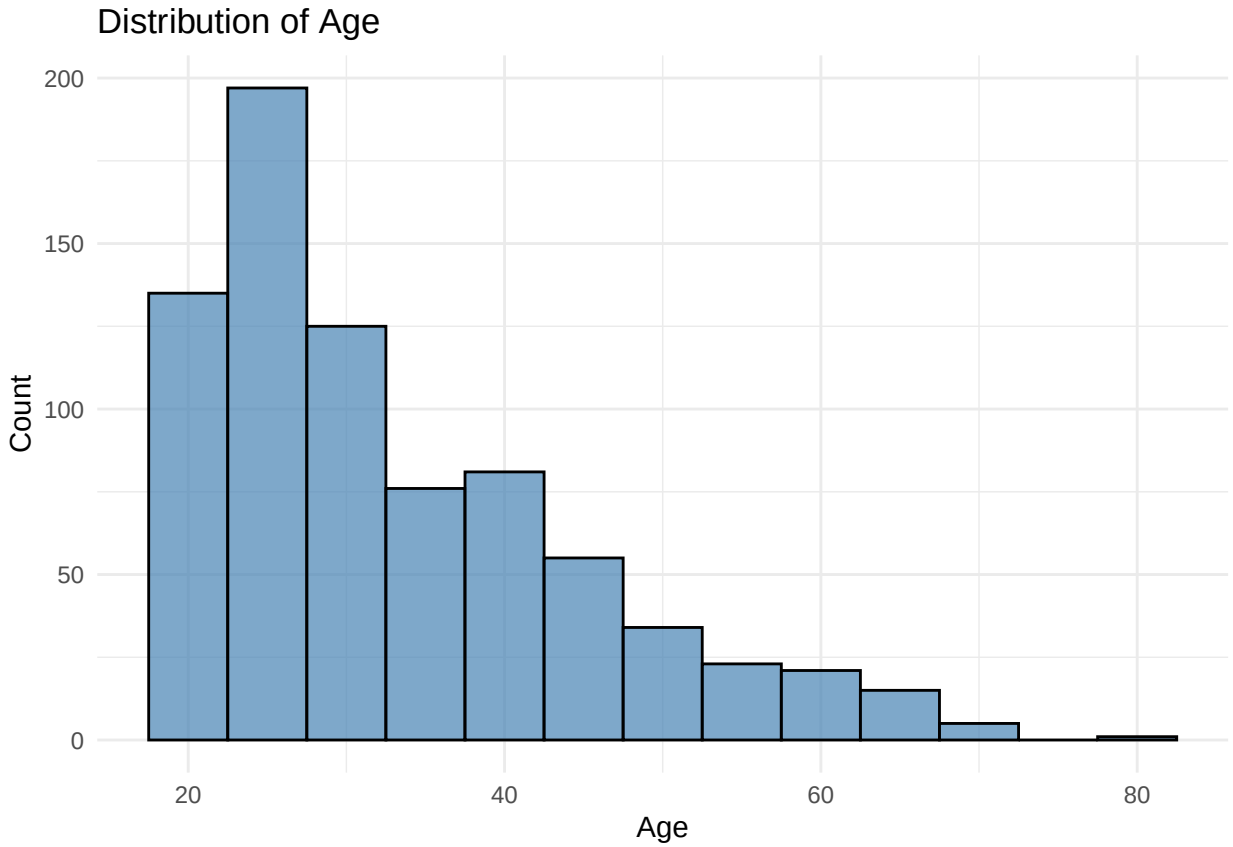

Distribution of Blood.Pressure
Count
150
100
50
0
0
40
80
120
Blood.Pressure


Distribution of Skin.Thickness
Count
200
150
100
50
0
0
25
50
75
100
Skin.Thickness

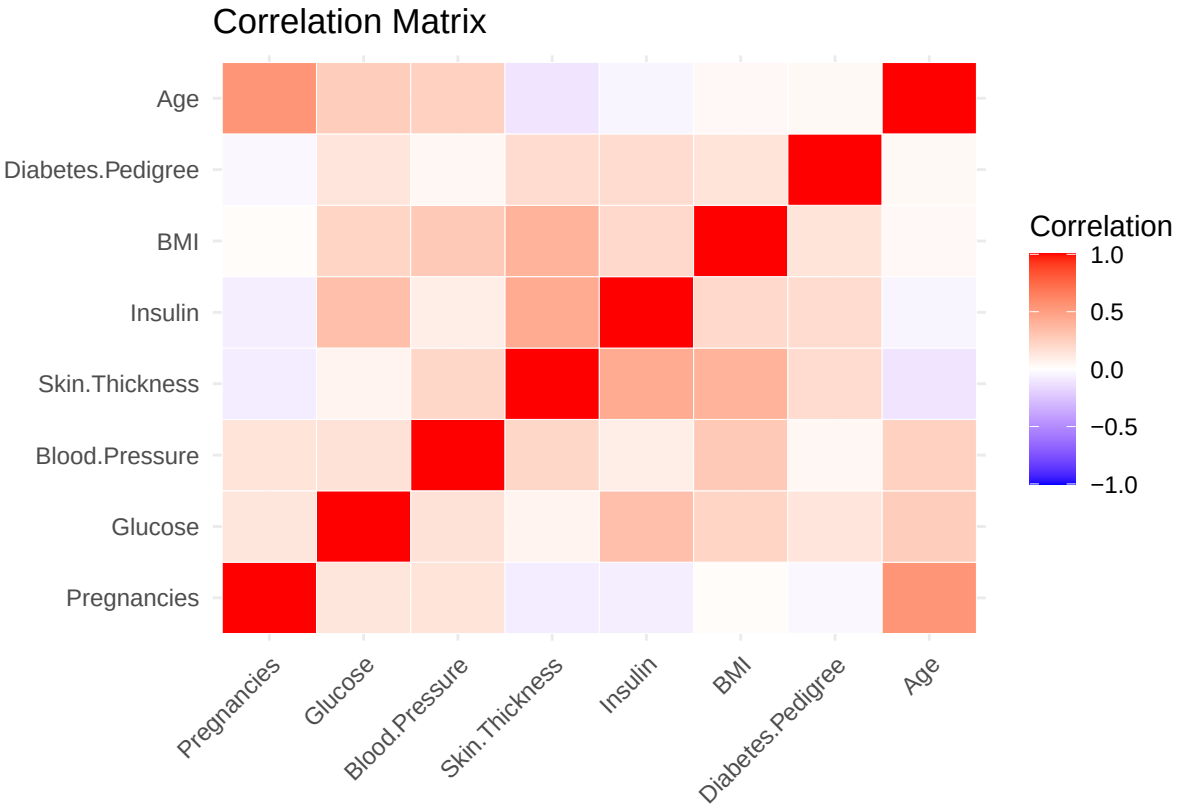

Distribution of Insulin
Count
300
200
100
0
0
200
400
600
800
Insulin


Distribution of BMI
Count
200
150
100
50
0
0
20
40
60
BMI


Distribution of Diabetes.Pedigree
Count
800
600
400
200
0
−2
−1
0
1
2
Diabetes.Pedigree



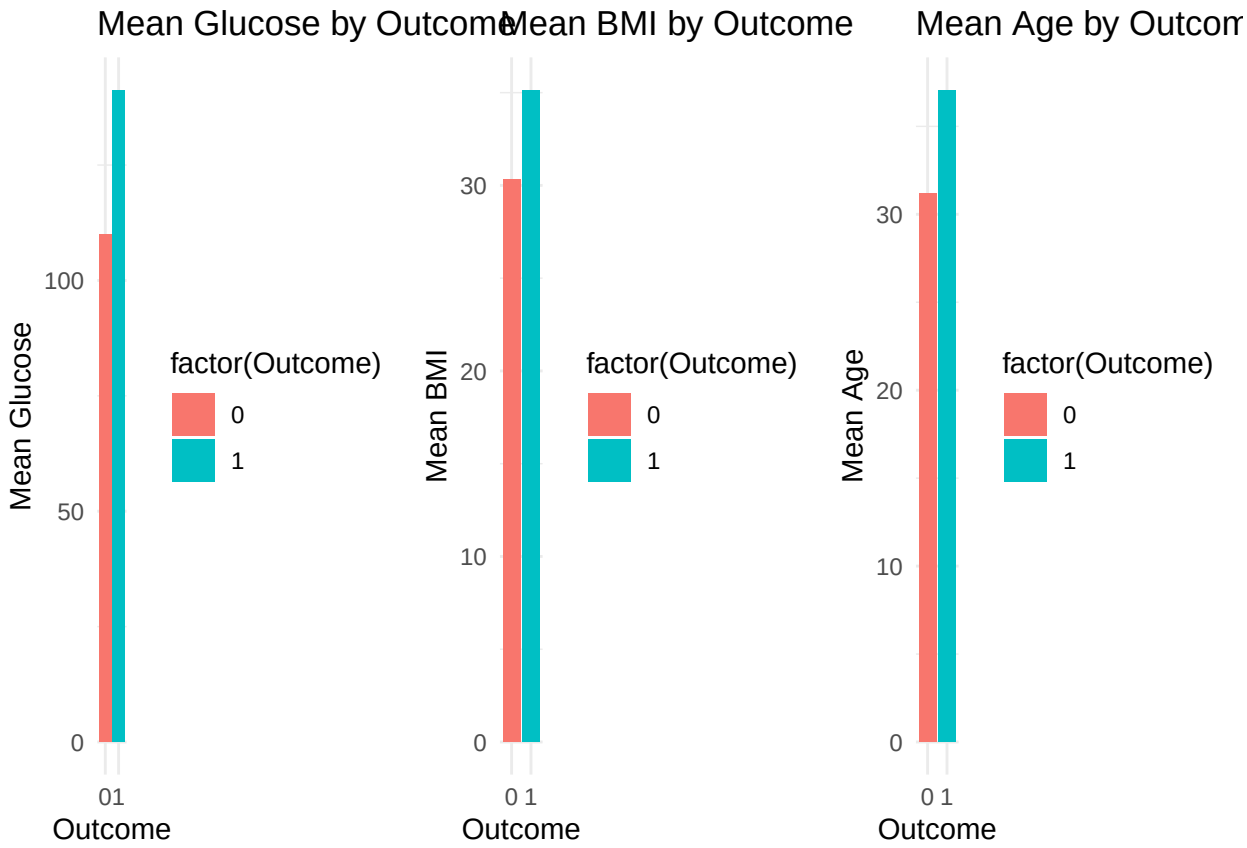
Some broad patterns were illustrated with histograms of continuous variables. In terms of distribution, glucose for instance, the distribution was positively skewed since majority of participants were close to normal glucose levels while few participants are at high abnormal glucose level known as hyperglycemia. BMI was close to normally distributed, leaned slightly to the right, indicating that the percent of people who could be classified as overweight or obese was apparent. Age was rather evenly distributed, with the largest number of participants between 21 and 40 years of age, which is the most common age for developing type 2 diabetes.



Additional analysis through correlation analysis gave further understanding of the nature of the different variables. A heatmap was used to confirm the findings that glucose levels were positively related to diabetes outcomes with the coefficient of +0.49. Thus, there were valid associations between BMI and diabetes pedigree scores and the overall diabetes outcomes, and the increased value of genetic risk factors presented as the effect of lifestyle behaviours. These insights enabled the subsequent predictive modeling by first determining what variables are most useful for diabetes classification.

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```



The three bar charts give a comparative view about the mean glucose level, mean BMI and mean age of the participants according to the diabetes outcome. The first chart shows that the relevant mean glucose level in diabetic people is about 141 while the same value for people without diabetes is equal to 109 thus making glucose a key marker of diabetes. The second chart reveals that the mean BMI is significantly higher also in the diabetic group, reinforcing obesity as a significant risk factor. On third chart, it can be seen that the participants of the diabetic group are somewhat older than participants in the non-diabetic group, which is also demonstrated in numerous studies indicating that chances of getting diabetes increase with age.

#5**Predictive Modeling**

```
##
## Call:
## glm(formula = Outcome ~ ., family = binomial, data = train_data)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -8.2216062  0.7781491 -10.566  < 2e-16 ***
## Pregnancies             0.1185211  0.0357929   3.311 0.000929 ***
## Glucose                 0.0352886  0.0041848   8.433  < 2e-16 ***
## Blood.Pressure         -0.0130815  0.0057276  -2.284 0.022374 *
## Skin.Thickness         -0.0009780  0.0075288  -0.130 0.896648
## Insulin                -0.0009111  0.0009841  -0.926 0.354533
## BMI                     0.0861702  0.0166338   5.180 2.21e-07 ***
## Diabetes.Pedigree       0.7824888  0.3212008   2.436 0.014845 *
## Age                     0.0152434  0.0102944   1.481 0.138676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 797.28  on 614  degrees of freedom
## Residual deviance: 583.64  on 606  degrees of freedom
## AIC: 601.64
##
## Number of Fisher Scoring iterations: 5
```

For this study, the medical records data was used to develop a logistic regression model for predicting diabetes outcomes. The dataset was split between the training set, which was 80%, and the test set which was 20%. Thus, glucose levels, BMI, and diabetes pedigree were taken as key predictors, as they were predictive of diabetes outcomes with a statistically significant measure. The most relevant predictors were glucose levels which have the p-value less than 0.001 proving its importance for diabetes diagnosis.

```
##                   Metric     Value
## Accuracy       Accuracy 0.7908497
## Sensitivity Sensitivity 0.9009901
## Specificity Specificity 0.5769231
```

Analyzing the results of the accuracy of the model with the test dataset it reached about 79 % accuracy. It further revealed that specificity, which shows ability of a screening tool to correctly identify non-diabetic people, was at 98%, while the sensitivity that reveals a tool's ability to accurately identify diabetic patients was at 90%. Specificity was, however, relatively moderate at 58% which indicates that the model struggled to differentiate patients who do not have diabetes.

#6**Discussion**

The findings of this study emphasizes that diabetes mellitus is a polygenic disorder resulting from a processes involving physiological, genetic, and ecological factors. Glucose levels, BMI, and diabetes pedigree emerged

as the most significant predictors that were consistent with medical wisdom and supported their place in diabetes pathogenesis (Kaul *et al.*, 2013). High BMI strengthens the presumption of obesity as the worsening enabler for advancing metabolic stress and insulin resistance. The addition of diabetes pedigree as a predictor emphasizes on genetic susceptibility which confirms the heralded role of genes in the development of this disease. The high sensitivity that a logistic regression model offers for PH interventions would mean that most people with diabetes would be correctly diagnosed. The sensitivity of such disturbances allows early detection, thus providing chance for timely intervention and, subsequently, reduction of severe consequences and enhancement of the long-term prognosis. However, due to the moderate specificity of the model, a high risk of false positive identification is observed, which will lead to the unnecessary diagnosis and treatment of the diseases for people without diabetes. To overcome this limitation, a more extensive database with more variables should be included in the analysis these are; diet, exercise and economical factors all of which are associated with diabetes.

#7**Recommendations**

The following managerial implications can be derived from the research work: Glucose and BMI surveillance should therefore be key governmental and community health initiatives. Positive changes in diet and the amount of exercise could greatly decrease the onset of diabetes in high-risk databases such as the Pima Indians. Efforts should also be made to include data collection on other factors defining diabetes risk, including the lifestyle and environment. The use of advanced analytical techniques could further enhance predictive modeling. Machine learning algorithms, for instance, can capture complex, nonlinear relationships between variables, potentially improving specificity without compromising sensitivity. Moreover, improving the awareness of clinic and community about glycemic control and weight management could promote the population-based preventive actions leading to lower incidence of diabetes.

#8**Conclusion**

This paper focuses on glucose level, Body Mass Index, and diabetes pedigree as the most important components that can be used to diagnose diabetes amongst the Pima Indian women. Such findings are useful in designing interventions that will help to reduce stated diabetes risk in high risk populations. Despite the high predictive analytics of the logistic regression model, more complex models with additional predictors should be formulated in the future. In filling these gaps, there is a huge potential for researchers and policymakers to make changes that will reduce the impact of diabetes and enhance health in those populations.

#9**References**

DeFronzo, R. A., Ferrannini, E., Groop, L., Henry, R. R., Herman, W. H., Holst, J. J., ... & Weiss, R. (2015). Type 2 diabetes mellitus. Nature reviews Disease primers, 1(1), 1-22.
Forouhi, N. G., & Wareham, N. J. (2010). Epidemiology of diabetes. Medicine, 38(11), 602-606. Katsarou, A., Gudbjörnsdottir, S., Rawshani, A., Dabelea, D., Bonifacio, E., Anderson, B. J., ... & Lernmark, Å. (2017). Type 1 diabetes mellitus. Nature reviews Disease primers, 3(1), 1-17.
Kaul, K., Tarr, J. M., Ahmad, S. I., Kohner, E. M., & Chibber, R. (2013). Introduction to diabetes mellitus. Diabetes, 1-11
Pima Indians diabetes database. (2000). Retrieved from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

#10**Code Appendix**

```
##
##
## processing file: Analysis Pima.Rmd

##   |                                                                      |

## output file: appendix-code.R

## [1] "appendix-code.R"
```

```
## setwd("~/")
## # Load necessary libraries
## library(tidyverse)
## library(caret)
## library(ggplot2)
## library(reshape2)
##
## # Load the dataset
## data <- read.csv("diabetes.csv")
##
## # View the structure of the data
## str(data)
##
## # Check if any missing values exist in the entire dataset
## sum(is.na(data))
##
## # Check if there are missing values for each column
## colSums(is.na(data))
##
## # Summary statistics
## summary_stats <- summary(data)
## print(summary_stats)
##
## # Visualize the data distribution
## for (col in colnames(data)[-ncol(data)]) {
##   p <- ggplot(data, aes_string(x = col)) +
##     geom_histogram(binwidth = 5, fill = "steelblue", color = "black", alpha = 0.7) +
##     labs(title = paste("Distribution of", col), x = col, y = "Count") +
##     theme_minimal()
##   print(p)
## }
##
## # Generate a correlation heatmap
## correlation_matrix <- data %>%
##   select(-Outcome) %>%
##   cor()
## heatmap <- ggplot(melt(correlation_matrix), aes(Var1, Var2, fill = value)) +
##   geom_tile(color = "white") +
##   scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1), spa
##   theme_minimal() +
##   labs(title = "Correlation Matrix", x = "", y = "") +
##   theme(axis.text.x = element_text(angle = 45, hjust = 1))
## print(heatmap)
##
## # Calculate the mean of Glucose, BMI, and Age for each Outcome group
## mean_values <- data %>%
##   group_by(Outcome) %>%
##   summarize(
##     Mean_Glucose = mean(Glucose, na.rm = TRUE),
##     Mean_BMI = mean(BMI, na.rm = TRUE),
##     Mean_Age = mean(Age, na.rm = TRUE)
##   )
##
## # Load necessary libraries
```

```
## library(tidyverse)
## library(gridExtra)
##
## # Generate a bar plot for the Mean of Glucose by Outcome
## plot_glucose <- ggplot(mean_values, aes(x = factor(Outcome), y = Mean_Glucose, fill = factor(Outcome)
##   geom_bar(stat = "identity", position = "dodge") +
##   labs(title = "Mean Glucose by Outcome", x = "Outcome", y = "Mean Glucose") +
##   theme_minimal()
##
## # Generate a bar plot for the Mean of BMI by Outcome
## plot_bmi <- ggplot(mean_values, aes(x = factor(Outcome), y = Mean_BMI, fill = factor(Outcome))) +
##   geom_bar(stat = "identity", position = "dodge") +
##   labs(title = "Mean BMI by Outcome", x = "Outcome", y = "Mean BMI") +
##   theme_minimal()
##
## # Generate a bar plot for the Mean of Age by Outcome
## plot_age <- ggplot(mean_values, aes(x = factor(Outcome), y = Mean_Age, fill = factor(Outcome))) +
##   geom_bar(stat = "identity", position = "dodge") +
##   labs(title = "Mean Age by Outcome", x = "Outcome", y = "Mean Age") +
##   theme_minimal()
##
## # Arrange the plots in a single row
## grid.arrange(plot_glucose, plot_bmi, plot_age, ncol = 3)
##
## # Train-Test Split
## set.seed(123)
## train_index <- createDataPartition(data$Outcome, p = 0.8, list = FALSE)
## train_data <- data[train_index, ]
## test_data <- data[-train_index, ]
##
## # Logistic Regression Model
## model <- glm(Outcome ~ ., data = train_data, family = binomial)
## summary(model)
##
## # Model Evaluation
## predictions <- predict(model, test_data, type = "response")
## threshold <- 0.5
## test_data$Predicted <- ifelse(predictions > threshold, 1, 0)
## conf_matrix <- confusionMatrix(as.factor(test_data$Predicted), as.factor(test_data$Outcome))
##
## # Model Performance Metrics
## performance <- data.frame(
##   Metric = c("Accuracy", "Sensitivity", "Specificity"),
##   Value = c(conf_matrix$overall["Accuracy"],
##             conf_matrix$byClass["Sensitivity"],
##             conf_matrix$byClass["Specificity"])
## )
## print(performance)
##
## # Visualize Performance Metrics
## p <- ggplot(performance, aes(x = Metric, y = Value, fill = Metric)) +
##   geom_bar(stat = "identity", color = "black", alpha = 0.7) +
##   labs(title = "Model Performance Metrics", x = "Metric", y = "Value") +
##   theme_minimal()
```

```
## print(p)
##
## knitr::purl("Analysis Pima.Rmd", output = "appendix-code.R", documentation = 0)
## cat(readLines("appendix-code.R"), sep = "\n")
```