R Project

Ntsika Mahle Mdingi

2025-03-27

Introduction

In this project I analysed the “Pima Indians Diabetes” dataset from Kaggle using R. The goal of this project is to clean the dataset, explore statistical relationships, visualize key trends, and extract meaningful insights about diabetes risk factors. The dataset contains various medical variables along with an outcome variable indicating whether an individual has diabetes (1 = diabetic, 0 = non-diabetic).

Data Cleaning And Viewing

Loading data

```
Diabetes <- read.csv("diabetes.csv")
```

Checking the data for missing values and removing invalid values (0)

```
sum(is.na(Diabetes))
```

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

```
Diabetes_Cleaned <- Diabetes[Diabetes$Glucose != 0 &
                             Diabetes$BloodPressure != 0 &
                             Diabetes$SkinThickness != 0 &
                             Diabetes$Insulin != 0 &
                             Diabetes$BMI != 0 &
                             Diabetes$Age != 0, ]
```

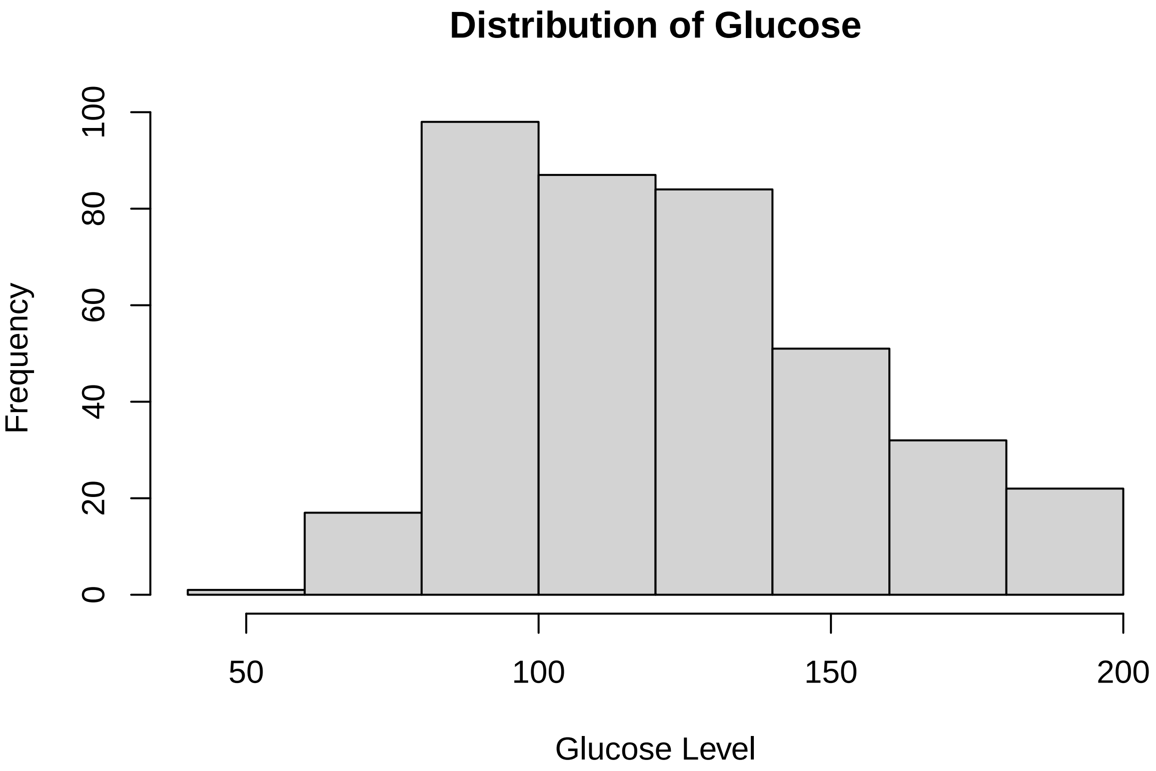
Statistical view of cleaned data

```
print(summary(Diabetes_Cleaned))
```

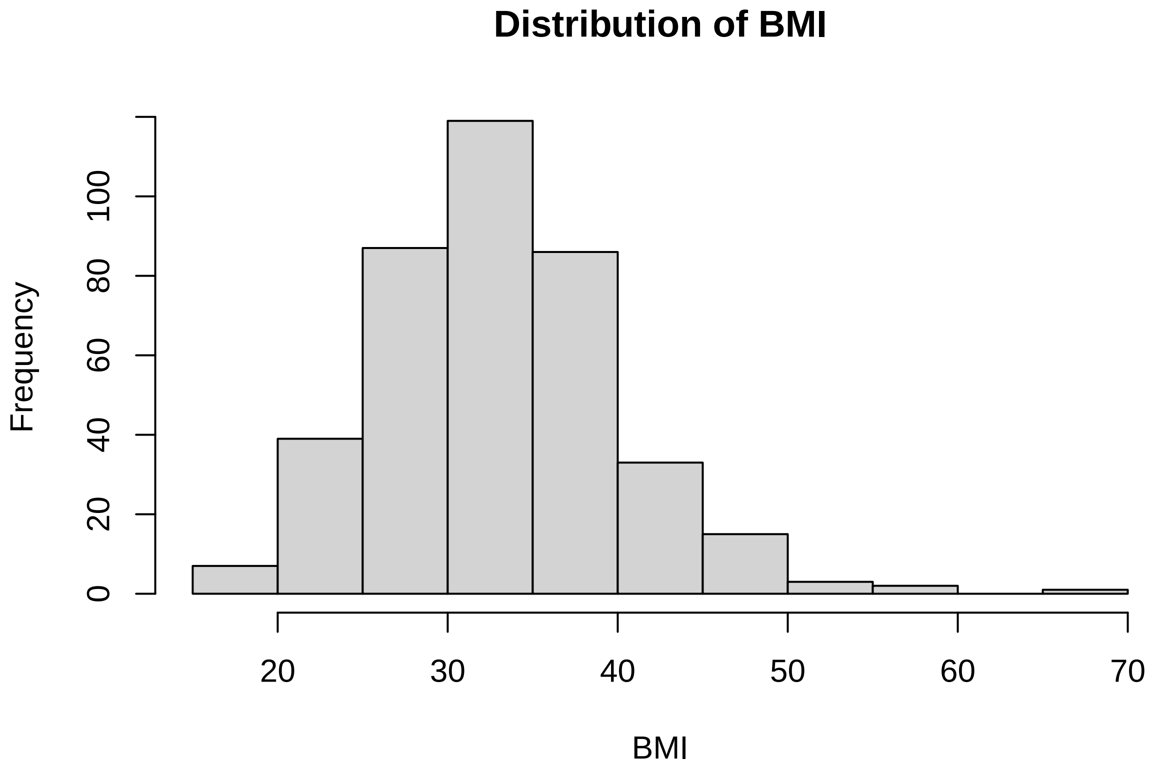
```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   : 56.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:21.00
##  Median : 2.000   Median :119.0   Median : 70.00   Median :29.00
##  Mean   : 3.301   Mean   :122.6   Mean   : 70.66   Mean   :29.15
##  3rd Qu.: 5.000   3rd Qu.:143.0   3rd Qu.: 78.00   3rd Qu.:37.00
##  Max.   :17.000   Max.   :198.0   Max.   :110.00   Max.   :63.00
##     Insulin            BMI        DiabetesPedigreeFunction      Age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0850           Min.   :21.00
##  1st Qu.: 76.75   1st Qu.:28.40   1st Qu.:0.2697           1st Qu.:23.00
##  Median :125.50   Median :33.20   Median :0.4495           Median :27.00
##  Mean   :156.06   Mean   :33.09   Mean   :0.5230           Mean   :30.86
##  3rd Qu.:190.00   3rd Qu.:37.10   3rd Qu.:0.6870           3rd Qu.:36.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##     Outcome
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3316
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

Data Visualization

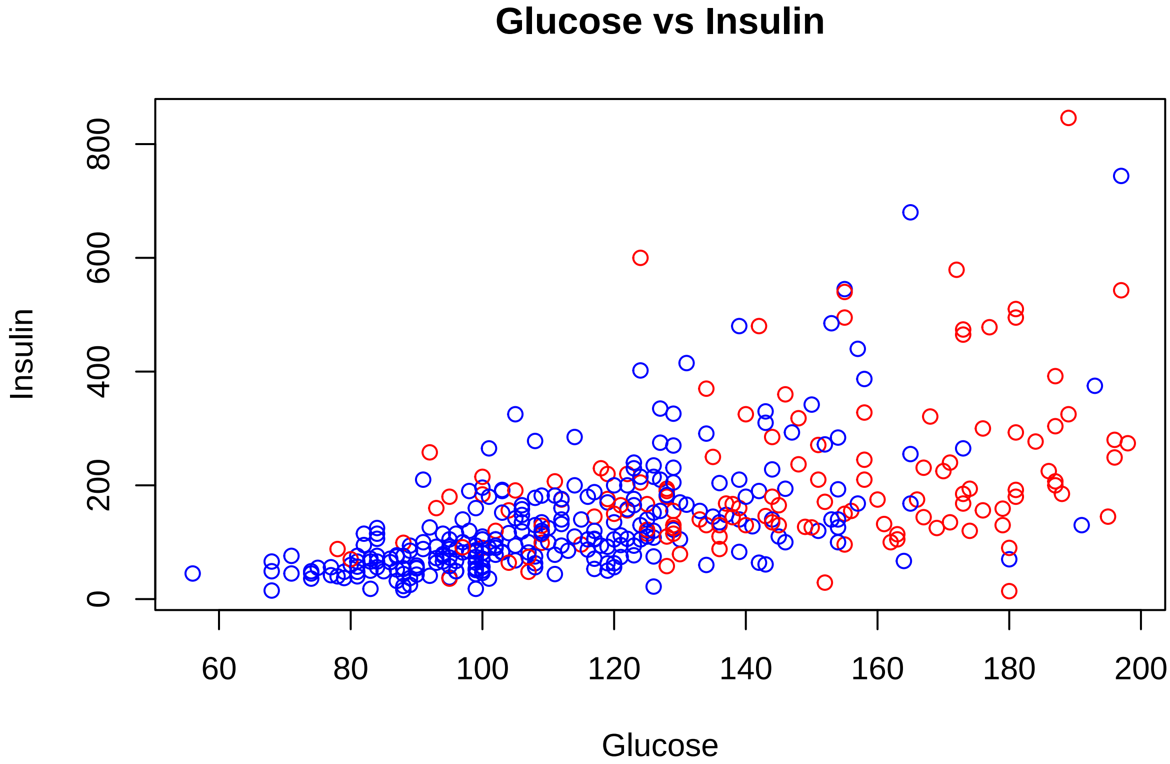
```r
hist(Diabetes_Cleaned$Glucose, main = "Distribution of Glucose", xlab = "Glucose Level")
```



```r
hist(Diabetes_Cleaned$BMI, main = "Distribution of BMI", xlab = "BMI")
```



```
plot(Diabetes_Cleaned$Glucose, Diabetes_Cleaned$Insulin,
     main = "Glucose vs Insulin",
     xlab = "Glucose", ylab = "Insulin",
     col = ifelse(Diabetes_Cleaned$Outcome == 1, "red", "blue"))
```



Statistical Correlation

```r
print(cor(Diabetes_Cleaned[, c("Glucose", "BMI", "Age", "Insulin", "BloodPressure", "DiabetesPedigreeFu
```

```
##                            Glucose       BMI        Age   Insulin BloodPressure
## Glucose                  1.0000000 0.2095159 0.34364150 0.5812230     0.2100266
## BMI                      0.2095159 1.0000000 0.06981380 0.2263965     0.3044034
## Age                      0.3436415 0.0698138 1.00000000 0.2170820     0.3000389
## Insulin                  0.5812230 0.2263965 0.21708199 1.0000000     0.0985115
## BloodPressure            0.2100266 0.3044034 0.30003895 0.0985115     1.0000000
## DiabetesPedigreeFunction 0.1401802 0.1587710 0.08502911 0.1359058    -0.0159711
## Outcome                  0.5157027 0.2701184 0.35080380 0.3014292     0.1926733
##                          DiabetesPedigreeFunction   Outcome
## Glucose                                0.14018018 0.5157027
## BMI                                    0.15877104 0.2701184
## Age                                    0.08502911 0.3508038
## Insulin                                0.13590578 0.3014292
## BloodPressure                         -0.01597110 0.1926733
## DiabetesPedigreeFunction               1.00000000 0.2093295
## Outcome                                0.20932951 1.0000000
```

The results indicate a moderate positive correlation (0.58) between glucose levels and insulin levels. Additionally, glucose shows a positive correlation with diabetes outcome, meaning individuals with higher glucose levels are more likely to be diabetic.

Key Takeaways

Glucose and Diabetes:

There is a clear positive correlation between glucose levels and diabetes outcome. Individuals with higher glucose readings are more likely to be diabetic.

Insulin and Glucose:

The scatterplot and correlation matrix suggest that higher glucose levels are often paired with lower insulin levels in diabetic individuals, which aligns with what is seen in diabetes, particularly Type 1 diabetes where insulin production is impaired.

BMI and Diabetes:

While glucose remains the strongest predictor, BMI also shows a relationship with diabetes, indicating that obesity or higher body fat percentages contribute to an increased risk of developing the condition.

Age Factor:

Individuals tend to have a higher likelihood of diabetes, suggesting that age is a risk factor that should be considered alongside other variables.

Conclusion

The findings suggest that high glucose levels are a key indicator of diabetes, often accompanied by lower insulin levels in diabetic individuals. BMI and age also contribute to the risk, reinforcing the importance of maintaining a healthy lifestyle and early screening for diabetes prevention.

Future analysis could explore predictive modeling to assess how well these variables can classify diabetes cases or introduce additional data sources to enhance understanding of diabetes risk factors.