EDA on Diabetes Dataset

Hekma Magdy

Introduction

This report provides an Exploratory Data Analysis (EDA) of the Diabetes dataset .
The goal is to explore the patterns, distributions, and relationships between variables, and to understand the factors related to diabetes outcomes.

Dataset Description

The dataset used in this project was obtained from **Kaggle** (Diabetes Dataset). It contains 768 observations and 9 variables related to diagnostic measurements for diabetes prediction.

Running Code

Loading data

```
data = read.csv("C:/Users/Soliman - Store/Downloads/archive/diabetes.csv")
```

Summary Statistics

```
head(data)
```

```
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
1           6     148            72            35       0 33.6
2           1      85            66            29       0 26.6
3           8     183            64             0       0 23.3
4           1      89            66            23      94 28.1
5           0     137            40            35     168 43.1
6           5     116            74             0       0 25.6
```

```
  DiabetesPedigreeFunction Age Outcome
1                    0.627  50       1
2                    0.351  31       0
3                    0.672  32       1
4                    0.167  21       0
5                    2.288  33       1
6                    0.201  30       0
```

```r
str(data)
```

```
'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 ...
 $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness           : 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 ...
 $ DiabetesPedigreeFunction: 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 ...
```

```r
nrow(data)
```

```
[1] 768
```

```r
ncol(data)
```

```
[1] 9
```

```r
dim(data)
```

```
[1] 768   9
```

```r
library(dplyr)
library(tidyr)

numeric_cols <- c("Pregnancies", "Glucose", "BloodPressure", "SkinThickness",
                  "Insulin", "BMI", "DiabetesPedigreeFunction", "Age")
summary_table = data %>%
  select(all_of(numeric_cols)) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
  group_by(Variable) %>%
  summarise(
    Mean   = mean(Value, na.rm = TRUE),
    Median = median(Value, na.rm = TRUE),
    SD     = sd(Value, na.rm = TRUE),
    Min    = min(Value, na.rm = TRUE),
    Max    = max(Value, na.rm = TRUE)
  )
summary_table
```

```
# A tibble: 8 x 6
  Variable                    Mean  Median      SD    Min    Max
  <chr>                      <dbl>   <dbl>   <dbl>  <dbl>  <dbl>
1 Age                       33.2    29      11.8   21      81
2 BMI                       32.0    32       7.88   0      67.1
3 BloodPressure             69.1    72      19.4    0     122
4 DiabetesPedigreeFunction   0.472   0.372   0.331  0.078   2.42
5 Glucose                  121.    117      32.0    0     199
6 Insulin                   79.8    30.5   115.     0     846
7 Pregnancies                3.85    3       3.37   0      17
8 SkinThickness             20.5    23      16.0    0      99
```

Data Cleaning

During the initial exploration, several variables were found to contain zero values (e.g., **Glucose**, **BloodPressure**, **SkinThickness**, **Insulin**, and **BMI**). These values are not physiologically possible and should be treated as missing data. Therefore, zeros in these variables were replaced with `NA` for more accurate analysis.

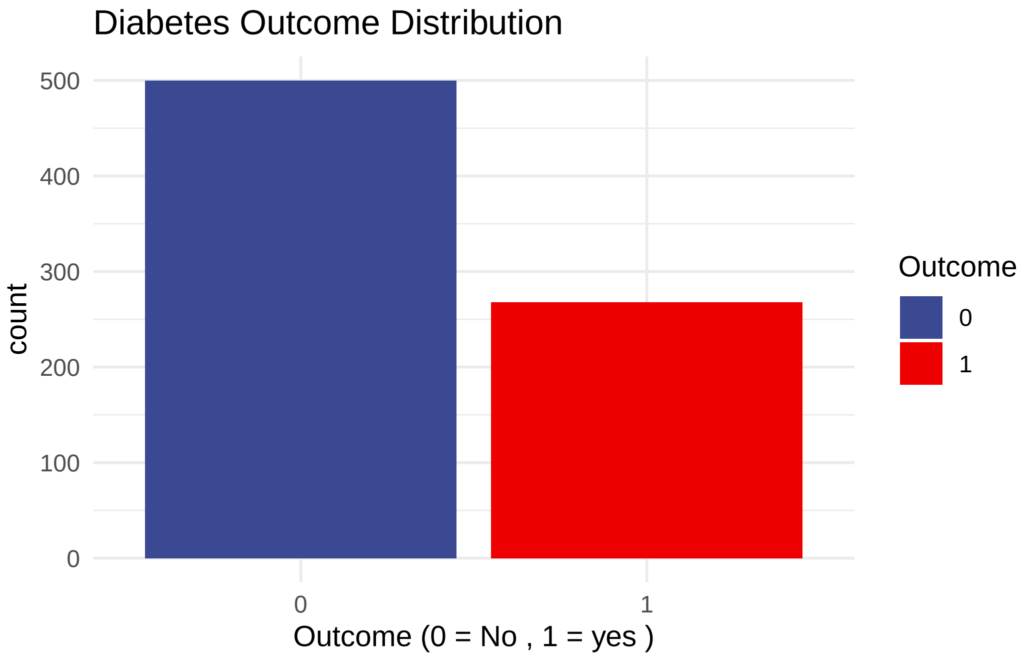
```
data$Glucose[data$Glucose == 0] <- NA
data$BloodPressure[data$BloodPressure == 0] <- NA
data$SkinThickness[data$SkinThickness == 0] <- NA
data$Insulin[data$Insulin == 0] <- NA
data$BMI[data$BMI == 0] <- NA
colSums(is.na(data))
```

```
              Pregnancies                  Glucose            BloodPressure
                        0                        5                       35
            SkinThickness                  Insulin                      BMI
```

```
                     227                      374                       11
DiabetesPedigreeFunction                      Age                  Outcome
                       0                        0                        0
```

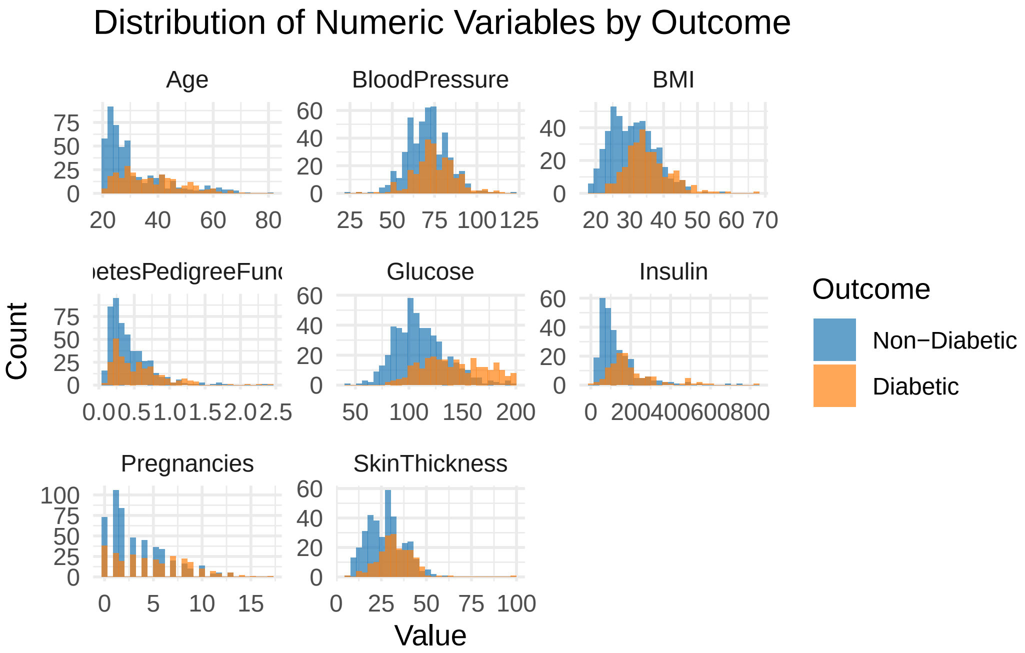
Outcome (Bar Plot)

```r
library(ggplot2)
library(ggsci)
data $Outcome= as.factor(data$Outcome)
ggplot(data = data , mapping = aes(x = Outcome , fill = Outcome))  +
  geom_bar() +
  scale_fill_manual(values = pal_aaas()(2)) +
  labs(title = "Diabetes Outcome Distribution" , x = "Outcome (0 = No , 1 = yes )" , y = "cou
  theme_minimal()
```



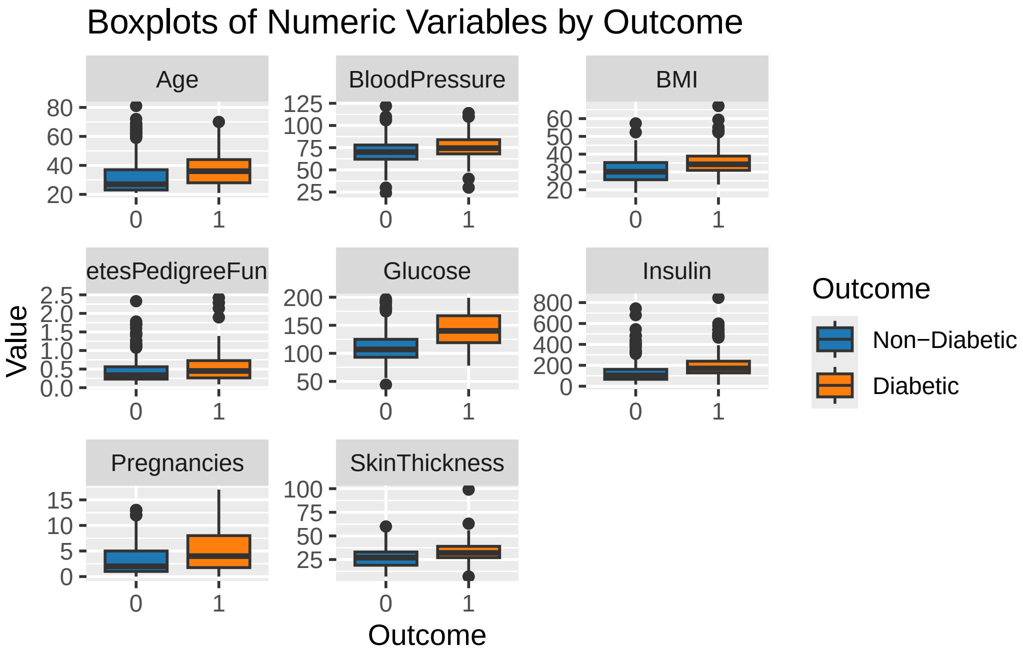
Histogram plot for each variable

```r
library(tidyverse)
data_long = data %>%
  gather(key = "variable" , value = "Value" , -Outcome)
ggplot(data_long, aes(x = Value, fill = Outcome)) +
  geom_histogram(bins = 30, alpha = 0.7, position = "identity") +
  facet_wrap(~variable, scales = "free") +
  theme_minimal() +
  labs(title = "Distribution of Numeric Variables by Outcome",
       x = "Value",
       y = "Count") +
  scale_fill_manual(values = c("0" = "#1f77b4", "1" = "#ff7f0e"),
                    name = "Outcome",
                    labels = c("Non-Diabetic", "Diabetic"))
```



Boxplots for numerical variables according to outcome

```
ggplot(data_long, aes(x = Outcome, y = Value, fill = Outcome)) +
  geom_boxplot() +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Boxplots of Numeric Variables by Outcome",
       x = "Outcome",
       y = "Value") +
  scale_fill_manual(values = c("0" = "#1f77b4", "1" = "#ff7f0e"),
                    name = "Outcome",
                    labels = c("Non-Diabetic", "Diabetic"))
```



Correlation Heatmap

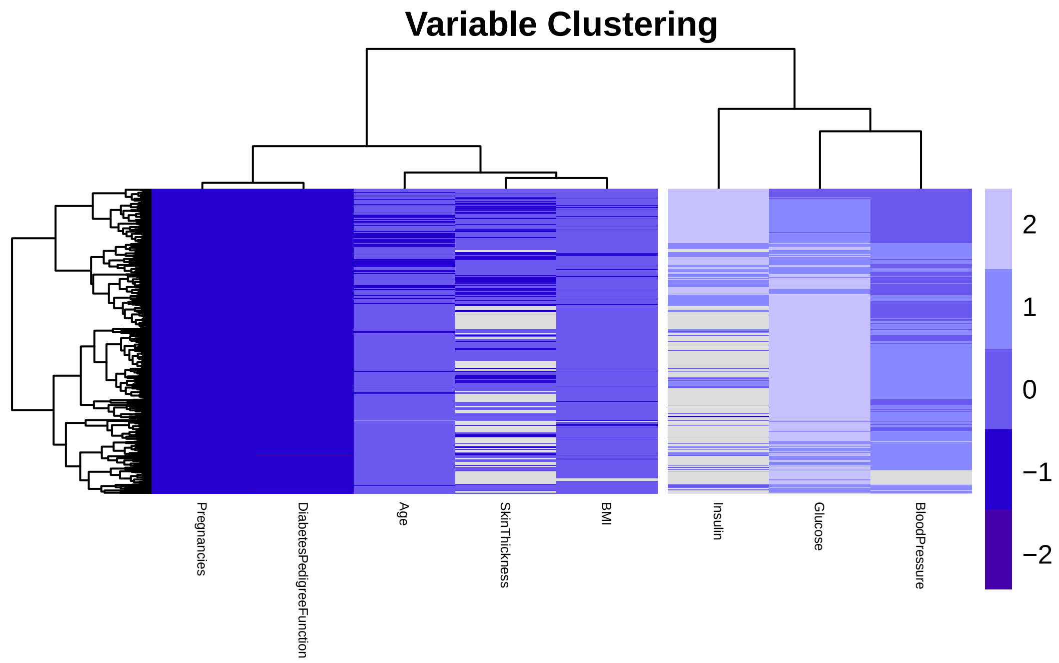
```r
library(ggsci)
data_without_outcome = data[-9]
str(data)
```

```
'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 ...
 $ BloodPressure           : int  72 66 64 66 40 74 50 NA 70 96 ...
 $ SkinThickness           : int  35 29 NA 23 35 NA 32 NA 45 NA ...
 $ Insulin                 : int  NA NA NA 94 168 NA 88 NA 543 NA ...
 $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 NA ...
 $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
 $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome                 : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 1 2 2 ...
```

```r
newdata_matrix = as.matrix(data_without_outcome)
dim(newdata_matrix)
```

```
[1] 768   8
```

```
library(pheatmap)
pheatmap(mat = newdata_matrix , scale = "row" , cluster_cols = T , cluster_rows = T ,
         cutree_cols = 2 , main = "Variable Clustering" , fontsize_row = 7 ,
         fontsize_col = 5, color = ggsci::pal_gsea()(5))
```



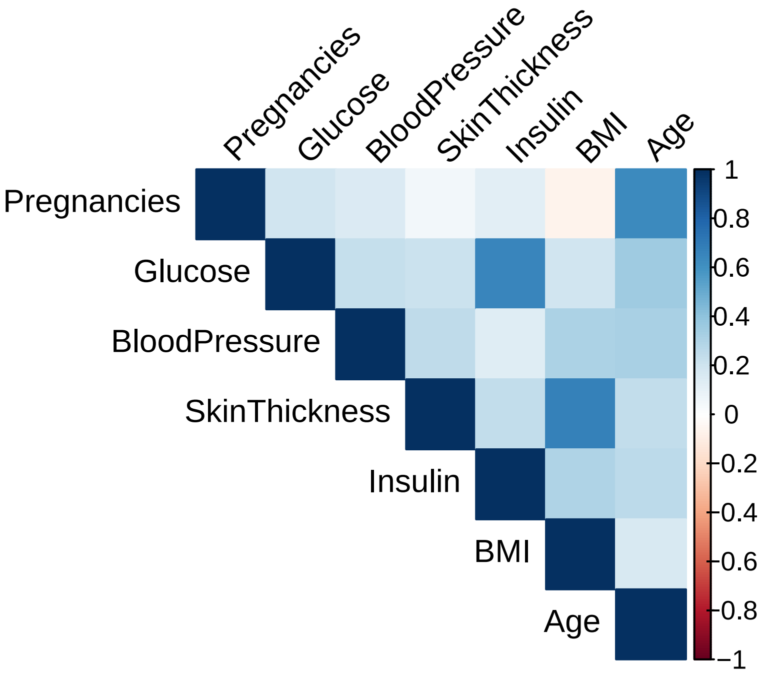
Spearman Rank Correlation Matrix

```
vars <- c("Pregnancies", "Glucose", "BloodPressure", "SkinThickness",
          "Insulin", "BMI", "Age")
cor_matrix  <- cor(data[, vars] ,method = "spearman" , use = "complete.obs")
print(cor_matrix)
```

```
              Pregnancies   Glucose BloodPressure SkinThickness   Insulin
Pregnancies    1.00000000 0.1904815     0.1524140    0.05475868 0.1231537
Glucose        0.19048148 1.0000000     0.2366093    0.21583824 0.6589582
BloodPressure  0.15241404 0.2366093     1.0000000    0.25010618 0.1316389
SkinThickness  0.05475868 0.2158382     0.2501062    1.00000000 0.2411450
Insulin        0.12315371 0.6589582     0.1316389    0.24114499 1.0000000
BMI           -0.06555144 0.1990712     0.3174275    0.67439293 0.3009061
Age            0.63365655 0.3503047     0.3292441    0.24150672 0.2607474
                      BMI       Age
Pregnancies   -0.06555144 0.6336566
```

```
Glucose        0.19907115 0.3503047
BloodPressure  0.31742747 0.3292441
SkinThickness  0.67439293 0.2415067
Insulin        0.30090608 0.2607474
BMI            1.00000000 0.1669629
Age            0.16696290 1.0000000
```

```
library(corrplot)
corrplot(cor_matrix , method = "color" , type = "upper", tl.col = "black", tl.srt = 45)
```



Conclusion

The exploratory data analysis (EDA) provided several important insights into the diabetes dataset. The outcome distribution showed a clear separation between diabetic and non-diabetic individuals, with more non-diabetic cases present. Visualizations such as histograms and boxplots revealed that key variables including **Glucose, BMI, and Insulin** exhibited noticeable differences between the two outcome groups, indicating their strong potential as predictors of diabetes.

The **Correlation Heatmap** further enhanced the analysis by highlighting the relationships between variables. For example, **Age and Pregnancies** showed a moderate positive correlation, which aligns with common demographic patterns. Similarly, **Insulin and SkinThickness** were positively correlated, suggesting a biological link. On the other hand, variables like **BloodPressure** and **Insulin** showed weaker correlations with most others, indicating more independent contributions. Importantly, none of the variables showed correlations high enough to suggest severe overlap between variables, meaning they can provide unique inf

The **clustered heatmap** also revealed patterns of similarity among variables and individuals, providing an additional perspective on how features group together. These findings confirm that variables such as **Glucose, BMI, and Insulin** are central in differentiating between diabetic and non-diabetic cases, while demographic measures like **Age and Pregnancies** contribute additional context.

Overall, this EDA not only highlighted important predictors but also provided a foundation for future **statistical modeling and machine learning approaches** to improve diabetes prediction and risk assessment.