OBJECTIVE 1

```
# Step 1: Data Preparation
# Load necessary libraries
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
```

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-7
```

The above linraries are used for this prohect. The

```
# Load the data
data <- read.csv("diabetes.csv")
```

In this project, I utilized the Pima Indian Diabetes dataset to fulfill the objectives of the course.

#Data Exploration

The summary function to provides a comprehensive overview of the central tendencies, dispersions, and distributions of each variable in the dataset, facilitating initial data understanding and identifying potential data quality issues.

```
# Data exploration
summary(data)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  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        DiabetesPedigreeFunction      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
```

tThe insights I considered are:

Many key variables like Glucose and BMI show zeros, likely representing missing data needing cleanup. Variables such as Age and Glucose vary widely, reflecting diverse health profiles among participants. Non-diabetic cases outnumber diabetic cases, indicating an imbalance that could affect model outcomes.

The str function quickly summarizes the structure of a dataset, including the types and formats of each variable, the number of observations, and the organization of the data frame, which helps in understanding the composition and readiness of the data for further analysis.

```
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 ...
```

```
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
```

#Dara Cleaning

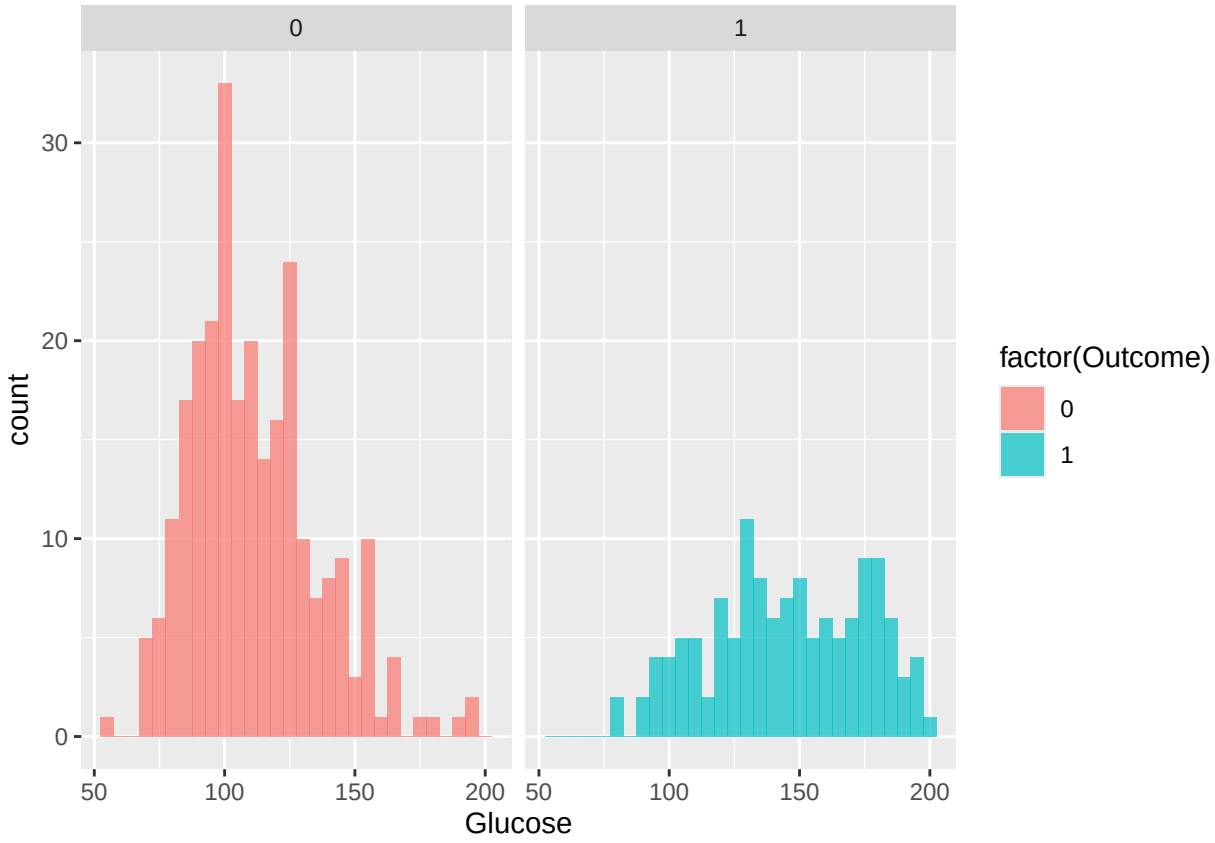
```
# Handling missing values and outliers
# Assuming zero in some columns (like Glucose, BloodPressure, SkinThickness, Insulin, BMI) represents m
zero_fields <- c("Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI")
data[zero_fields] <- lapply(data[zero_fields], function(x) replace(x, x == 0, NA))
data <- na.omit(data)
```

In the above step, I replaced zero values in key variables like Glucose and BMI with NA, considering them as missing data. I then removed all rows with any NA values, cleaning the dataset of incomplete records. This process improves data accuracy and reliability for subsequent statistical analysis and modeling.

#Exploratory Data Analysis (EDA)

To effectively visualize and analyze the distribution and relationships between key variables such as glucose, age, and diabetes outcome, in my opinion, histograms and scatter plots provide a clear, graphical representation of the data, which helps in detecting underlying patterns, trends, and outliers. This visual approach is crucial for understanding how these variables interact and influence the likelihood of diabetes, aiding in more informed model development and hypothesis testing.

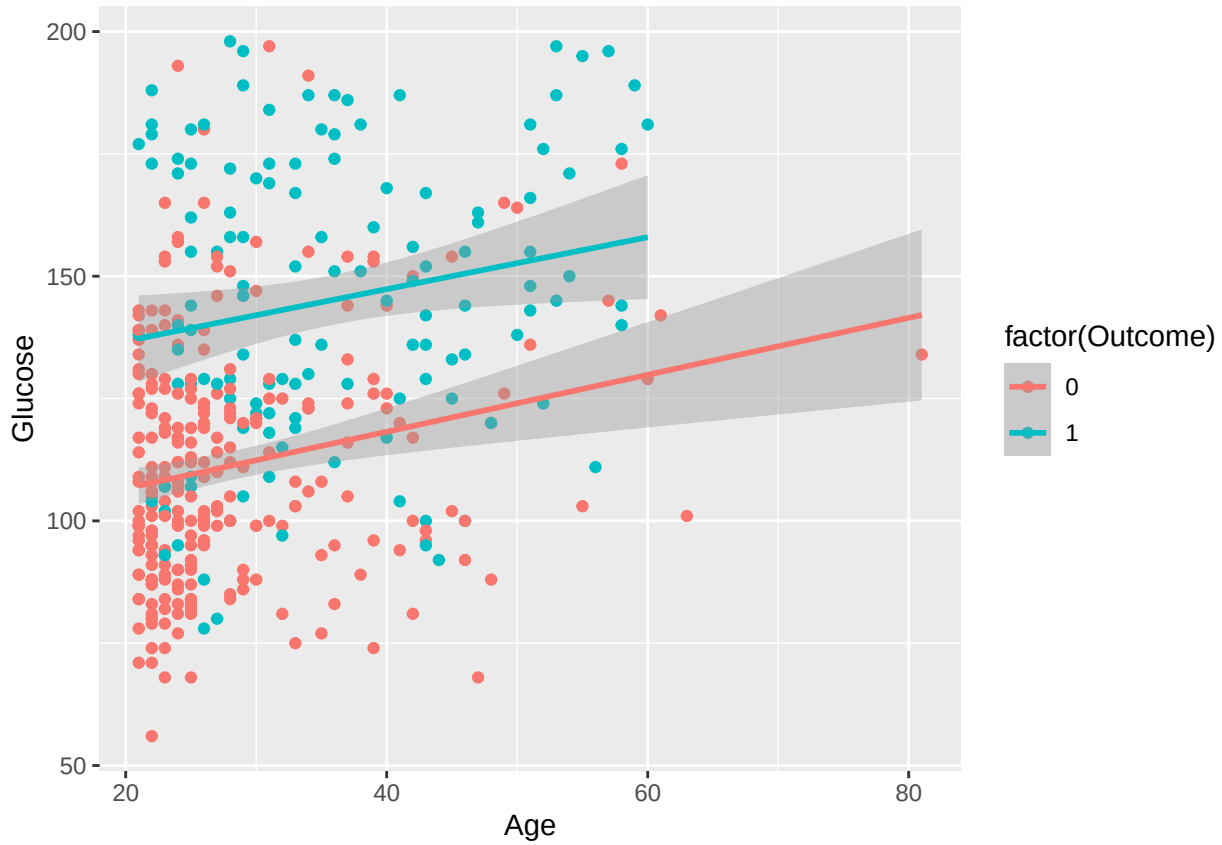
```
# Step 2: Exploratory Data Analysis (EDA)
# Visualizing distributions
ggplot(data, aes(x = Glucose, fill = factor(Outcome))) + geom_histogram(binwidth = 5, alpha = 0.7) + fa
```



The facetted histogram that compares the distribution of glucose levels for individuals with and without diabetes, represented by the facets labeled "0" (no diabetes) and "1" (diabetes), respectively. The red histogram (facet 0) shows the frequency of glucose levels for non-diabetic individuals, while the blue histogram (facet 1) corresponds to diabetic individuals. Notably, the distribution for the diabetic group tends to have higher glucose levels, with the peak shifted towards the right compared to the non-diabetic group. This visual contrast highlights the association between higher glucose levels and the presence of diabetes, a critical insight for understanding risk factors within the dataset.

```
ggplot(data, aes(x = Age, y = Glucose, color = factor(Outcome))) + geom_point() + geom_smooth(method =
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This scatter plot shows individual data points representing the relationship between age and glucose levels, differentiated by diabetes outcome, where red dots indicate non-diabetic individuals (Outcome 0) and blue dots indicate diabetic individuals (Outcome 1). Trend lines—red for non-diabetic and blue for diabetic—suggest an upward trend, indicating that glucose levels may increase with age for both groups, but with a steeper incline for diabetic individuals. The shaded areas around the lines represent confidence intervals, providing a visual sense of the variability and reliability of the estimated relationship. Creating this plot is essential because it visually explores the potential interaction between age and glucose levels as they relate to the presence of diabetes, which can inform subsequent analytical decisions and model feature selection.

```
# Calculating summary statistics
summary(data)
```

```
##   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
```

#Model Fitting

```
# Step 3: Model Fitting
# Splitting the dataset
set.seed(123)
trainIndex <- createDataPartition(data$Outcome, p = 0.7, list = FALSE)
trainData <- data[trainIndex, ]
testData <- data[-trainIndex, ]
```

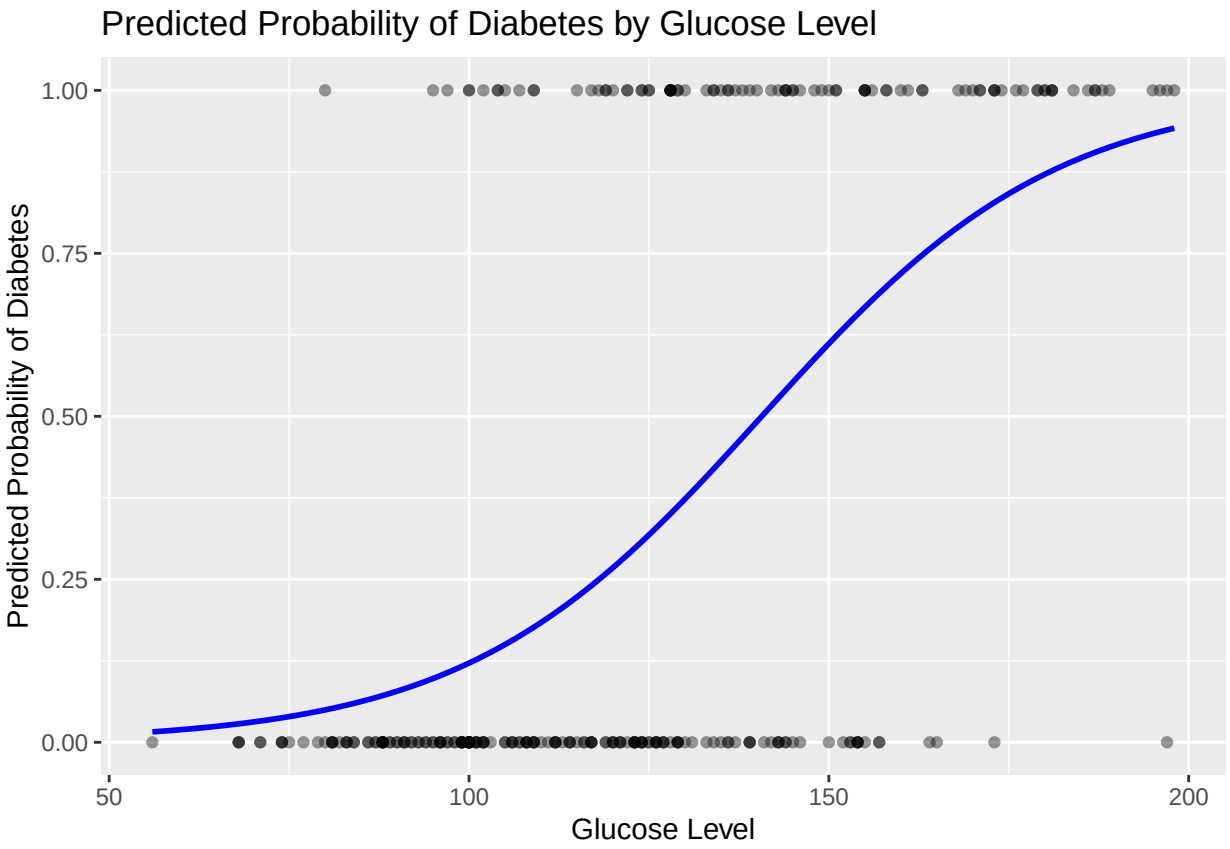
In this step, the dataset is being split into two parts: a training set and a test set. The set.seed(123) function ensures that the random selection of data is reproducible. The createDataPartition function from the caret package is used to divide the dataset, allocating 70% of the data to the training set (trainData) and the remaining 30% to the test set (testData). This partitioning is stratified on the Outcome variable to maintain the proportion of cases with and without diabetes in both sets. This split is crucial for training the model on one subset of data and then evaluating its performance on a separate, unseen subset to assess its predictive ability.

```
# Fitting logistic regression model
fit_logistic <- glm(Outcome ~ ., family = binomial, data = trainData)
```

```
ggplot(trainData, aes_string(x = "Glucose", y = "Outcome")) +
  geom_point(alpha = 0.4) +
  stat_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE, color = "blue") +
  labs(title = "Predicted Probability of Diabetes by Glucose Level",
       x = "Glucose Level",
       y = "Predicted Probability of Diabetes")
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


Predicted Probability of Diabetes by Glucose Level
Predicted Probability of Diabetes
1.00
0.75
0.50
0.25
0.00
50
100
150
200
Glucose Level