Using Body Signals to Predict Smoking Status

Project Description

The provided dataset is a collection of basic health biological signal data aimed at determining the presence or absence of smoking based on these bio-signals. The target variable in this dataset is coded as follows:

- NO smoking: 0
- YES smoking: 1

The dataset includes various features that capture different aspects of an individual's health and well-being. These features include demographic information such as gender and age, anthropometric measurements like height, weight, and waist circumference, as well as various physiological indicators like blood pressure (systolic and diastolic), cholesterol levels, hemoglobin levels, urine protein levels, and more.

By analysing these bio-signals and their relationship with the smoking variable, we can gain insights into how these signals may be indicative of smoking behaviour. This information can be valuable for understanding the impact of smoking on health and developing strategies for smoking cessation or prevention.

Exploring the dataset and analysing the relationships between these bio-signals and smoking status can provide valuable insights into the association between smoking behaviour and health indicators.

Problem Definition

The problem is to investigate the relationship between smoking status and various body signals in a given dataset. The dataset contains information on multiple body signals.

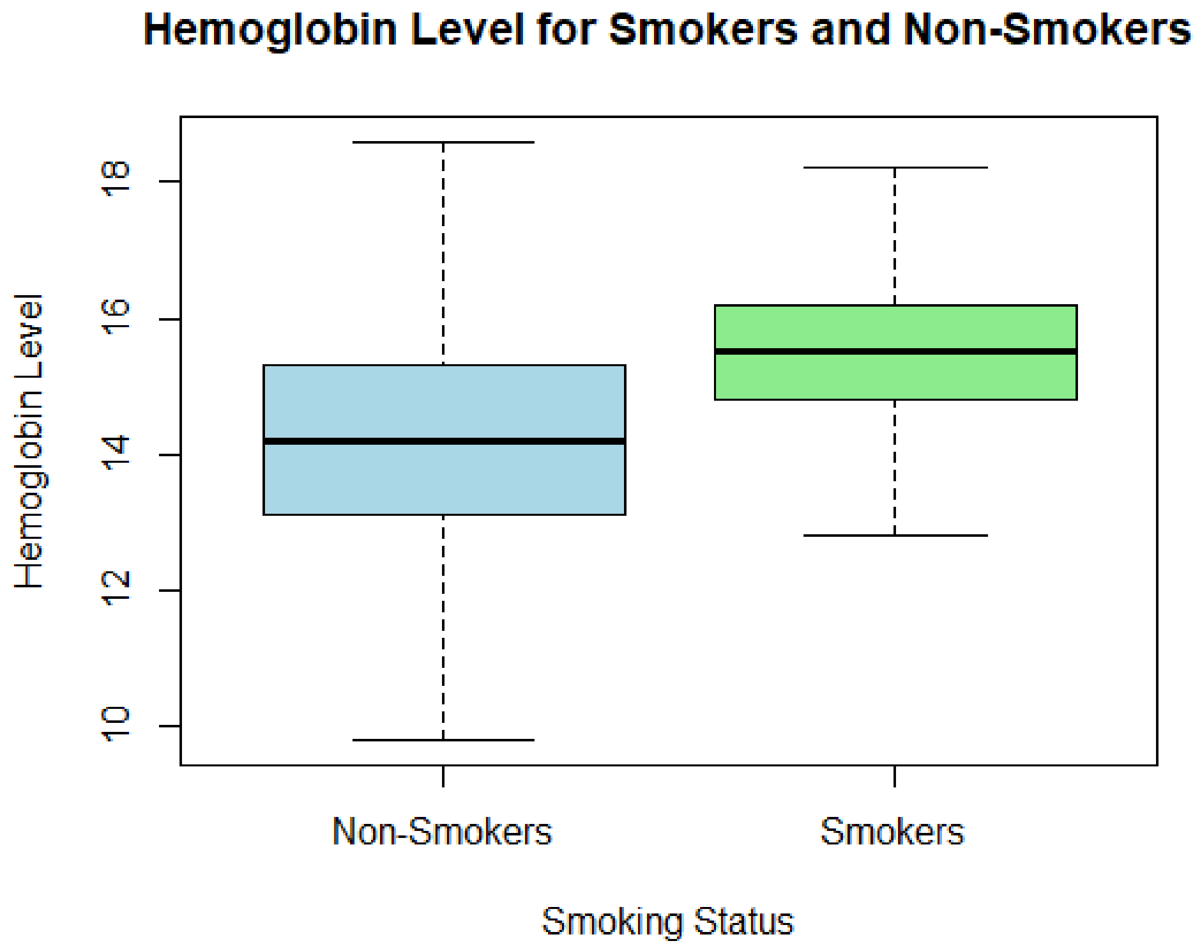
The goal is to understand if there are any significant differences or associations between smoking status and these body signals. This analysis aims to explore whether smoking has an impact on multiple body signals

Objective

The objective of this study is to examine the association between smoking status and various body signals in a given dataset.

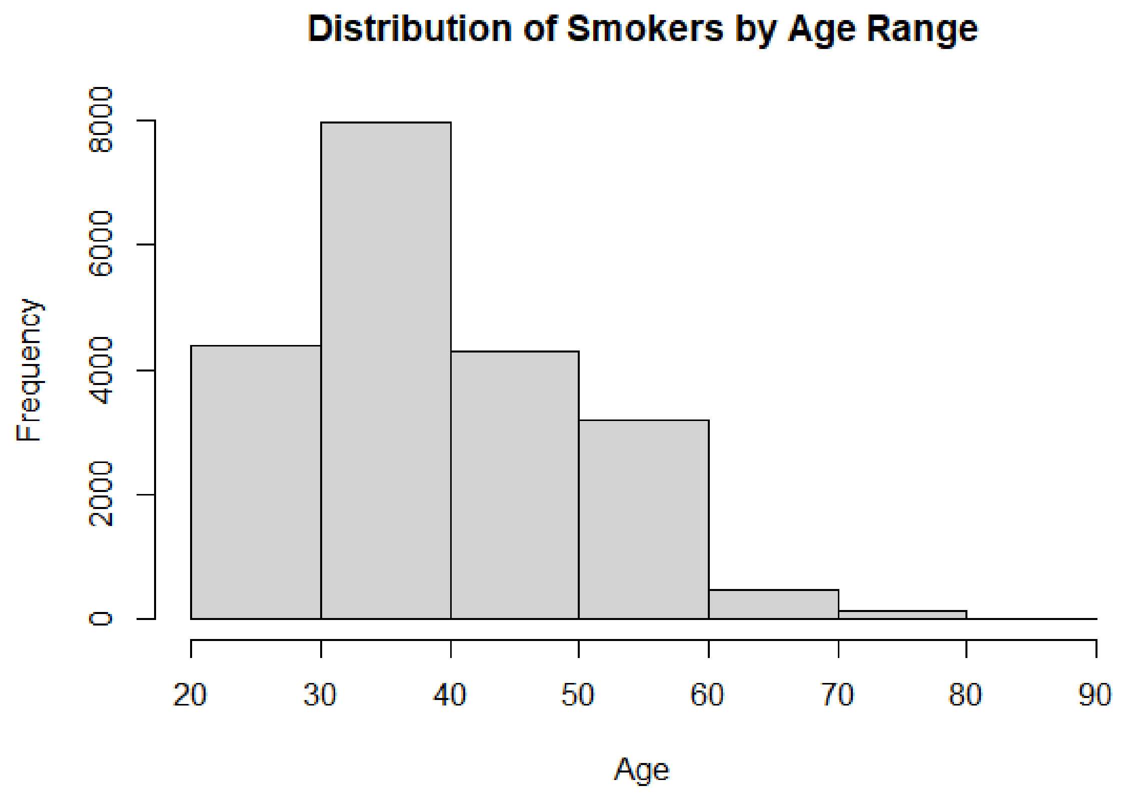
Dataset & Variables Description

- "ID" : record in index
- "gender" : Male and Females displayed as (M,F)
- "age" : Age of population
- "height.cm." : Height in CM
- "weight.kg." : Weight in KG
- "systolic" : First read in blood pressure ex. 120
- "diastolic" : Second read in blood pressure ex. 80
- "Fasting.blood.sugar”: measures sugar (glucose) in your blood
- "Cholesterol" : Cholesterol read
- "triglyceride" : Type of fat (lipid) found in your blood
- "hemoglobin" : Hemoglobin read
- "Urine.protein" : Read of excess of bloodborne proteins in urine
- "serum.creatinine" : a waste product in your blood that comes from your muscles.
- "dental.caries" : Tooth decay or dental cavities
- "smoking" : Value of smoker and non-smokers appears (1:Smokers , 0:Non-smokers)



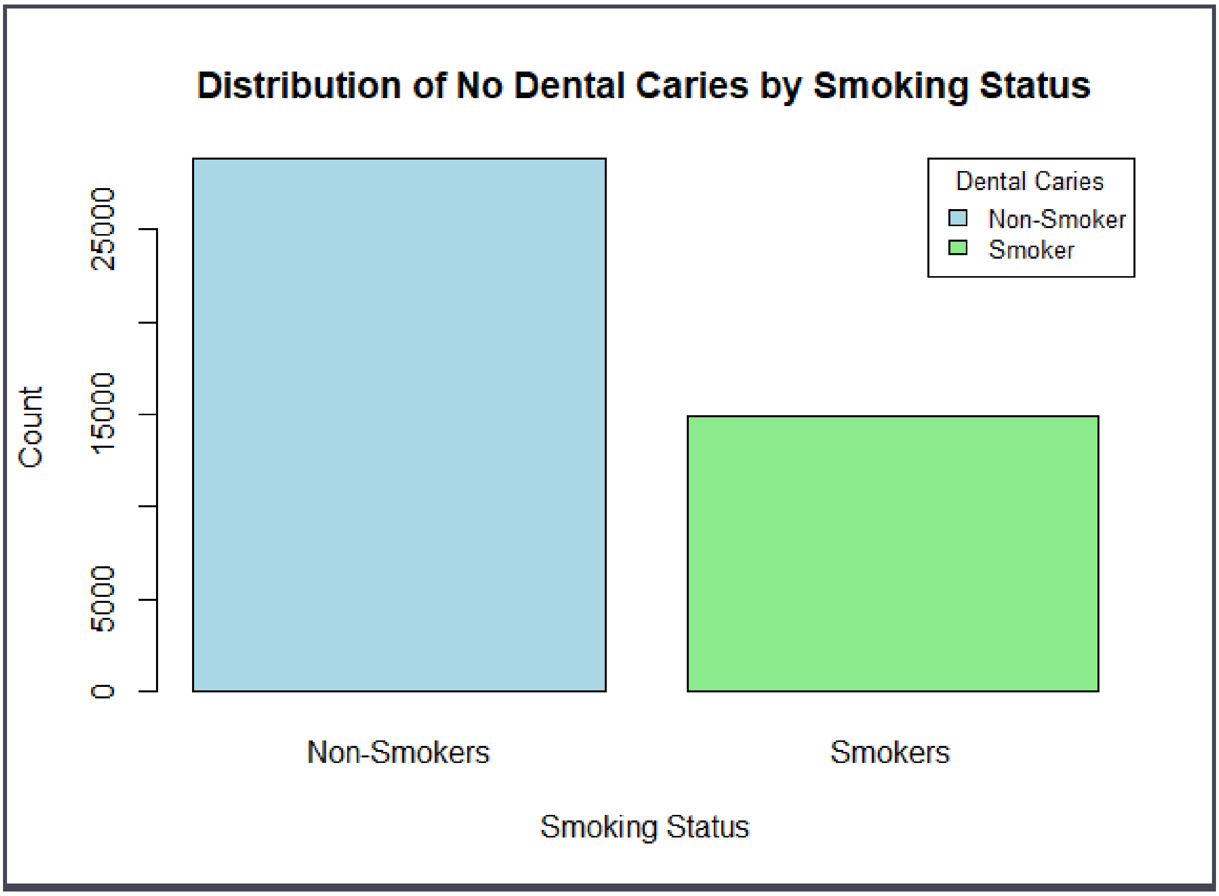
Observation: The median hemoglobin level for smokers is noticeably higher than that of non-smokers, Which suggests that there’s a potential association between smoking and high hemoglobin levels.

```
boxplot(hemoglobin ~ smoking, data = data,
        xlab = "Smoking Status", ylab = "Hemoglobin Level",
        main = "Hemoglobin Level for Smokers and Non-Smokers",
        col = c("lightblue", "lightgreen"),
        names = c("Non-Smokers", "Smokers"))
```



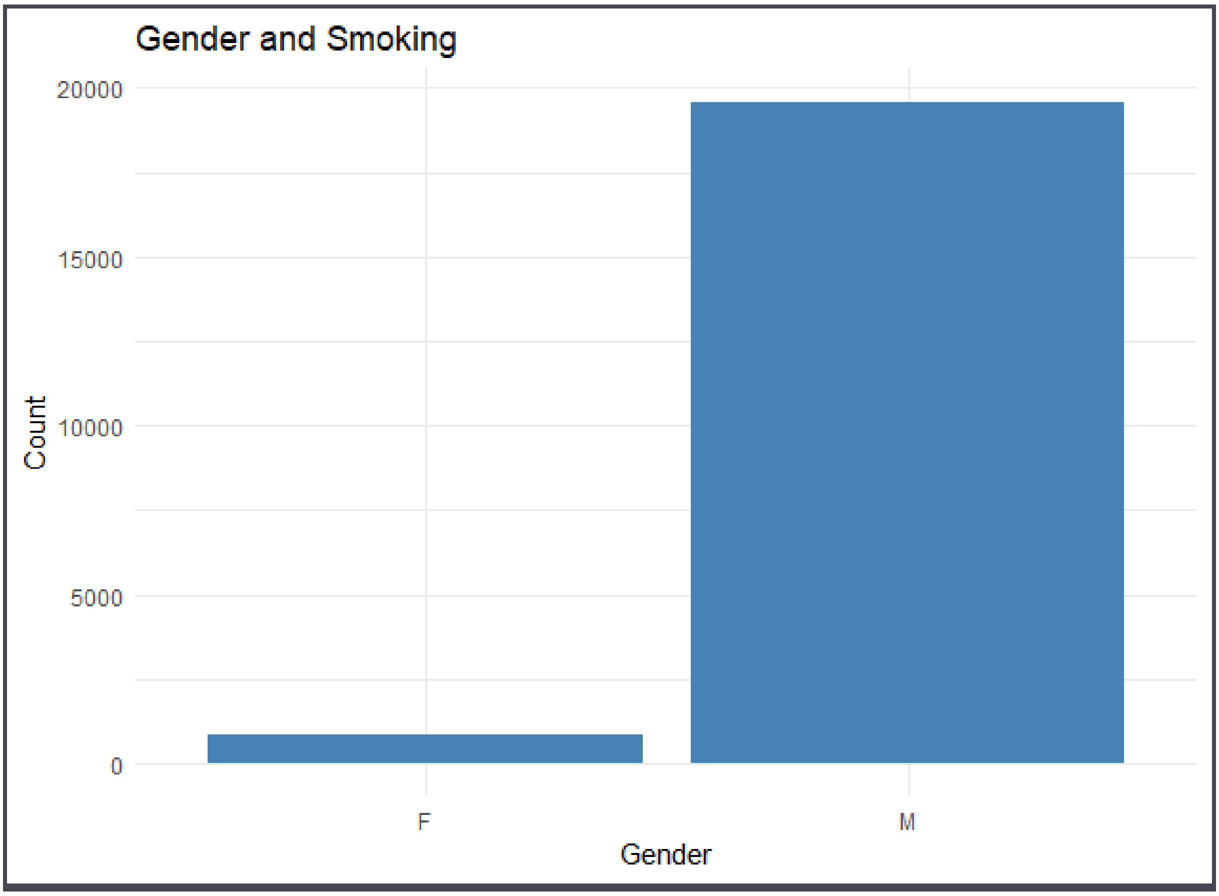
Observation: We observed that the histogram is right Skewed which indicates that the majority of smokers are younger people and the peak is from 30 to 40 years which are adults which are close to 8000 thousand people

```
hist(data$age[data$smoking == 1], breaks = 5,
        xlab = "Age", ylab = "Frequency",
        main = "Distribution of Smokers by Age Range")
```



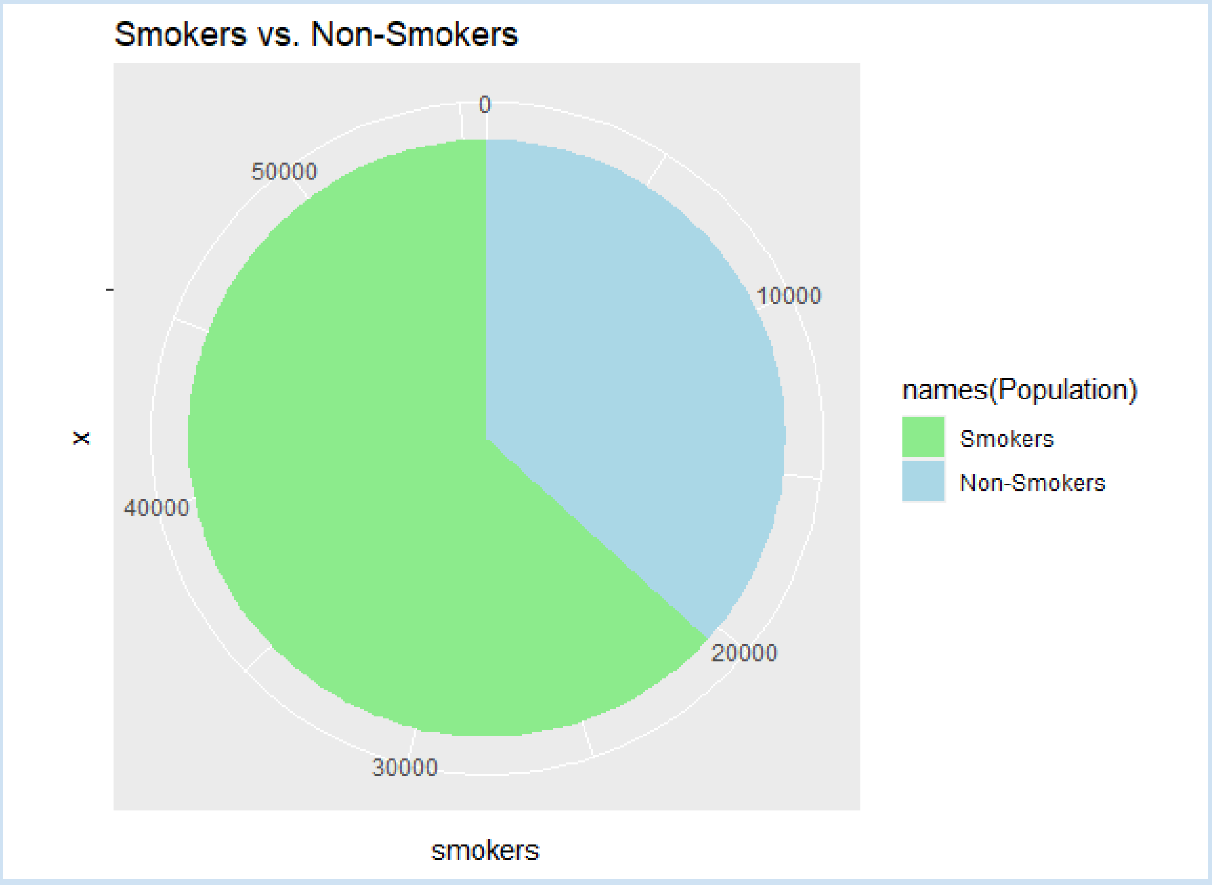
Observation: Upon analysing the bar plot comparing dental caries between smokers and non-smokers, we made a significant observation. It became evident that non-smokers had a notably lower prevalence of dental caries compared to smokers.

```
contingency_table <- table(data$dental.caries, data$smoking)
no_caries_table <- contingency_table[1, ]
barplot(no_caries_table, beside = TRUE, legend = TRUE,
        main = "Distribution of No Dental Caries by Smoking Status",
        xlab = "Smoking Status", ylab = "Count",
        col = c("lightblue", "lightgreen"),
        names = c("Non-Smokers", "Smokers"),
        args.legend = list(title = "Dental Caries",
                x = "topright",
                cex = 0.8,
                legend = c("Non-Smoker", "Smoker")))
```



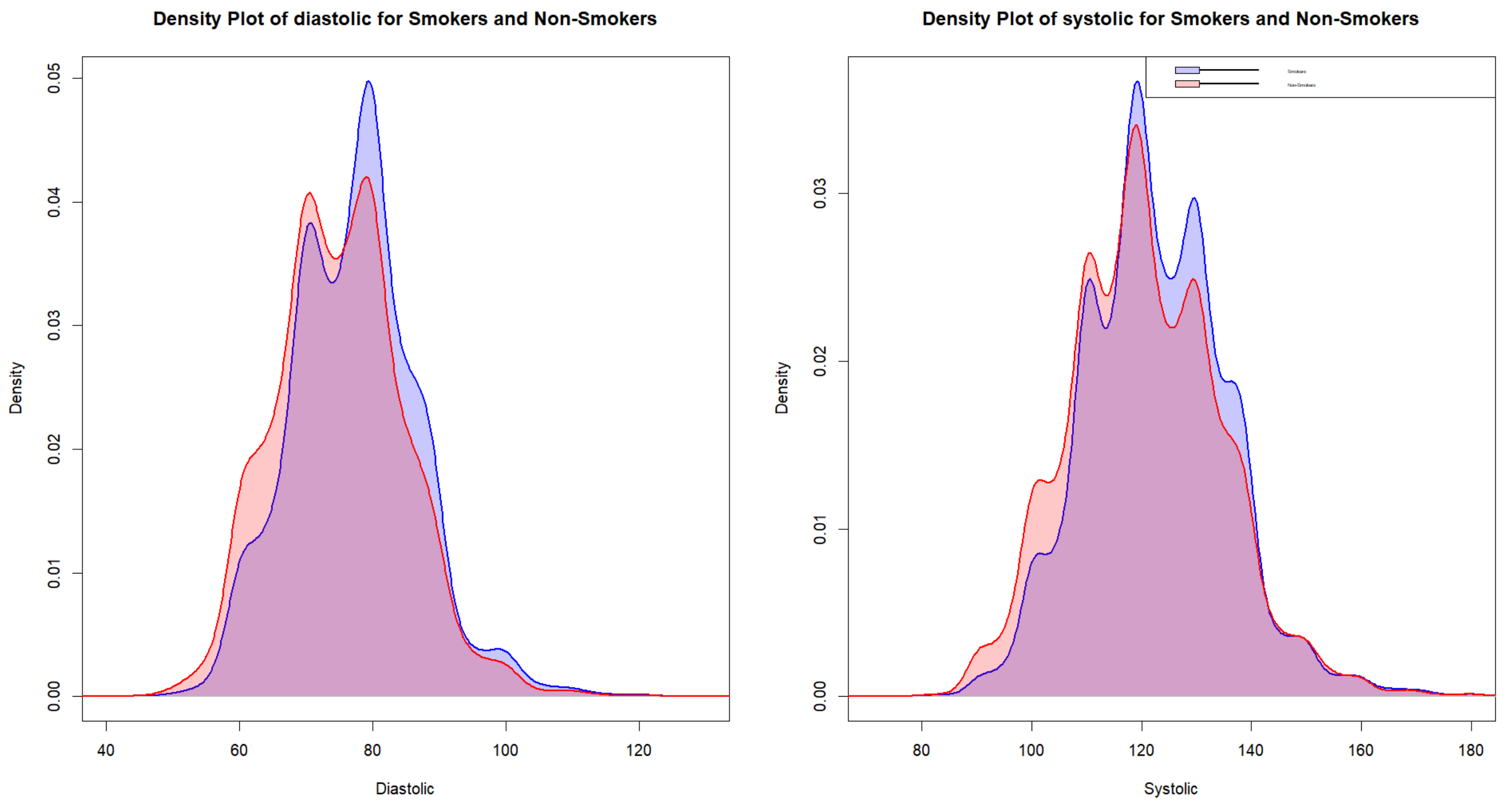
Observation: After analysing the data, we observed a statistically significant difference in the proportion of smokers between males and females. The proportion of smokers among males was found to be significantly greater than the proportion of smokers among females.

```
subset_data <- subset(data, smoking == 1)
bar_plot <- ggplot(subset_data, aes(x = gender)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Gender and Smoking",
           x = "Gender",
           y = "Count") +
  theme_minimal()
print(bar_plot)
```



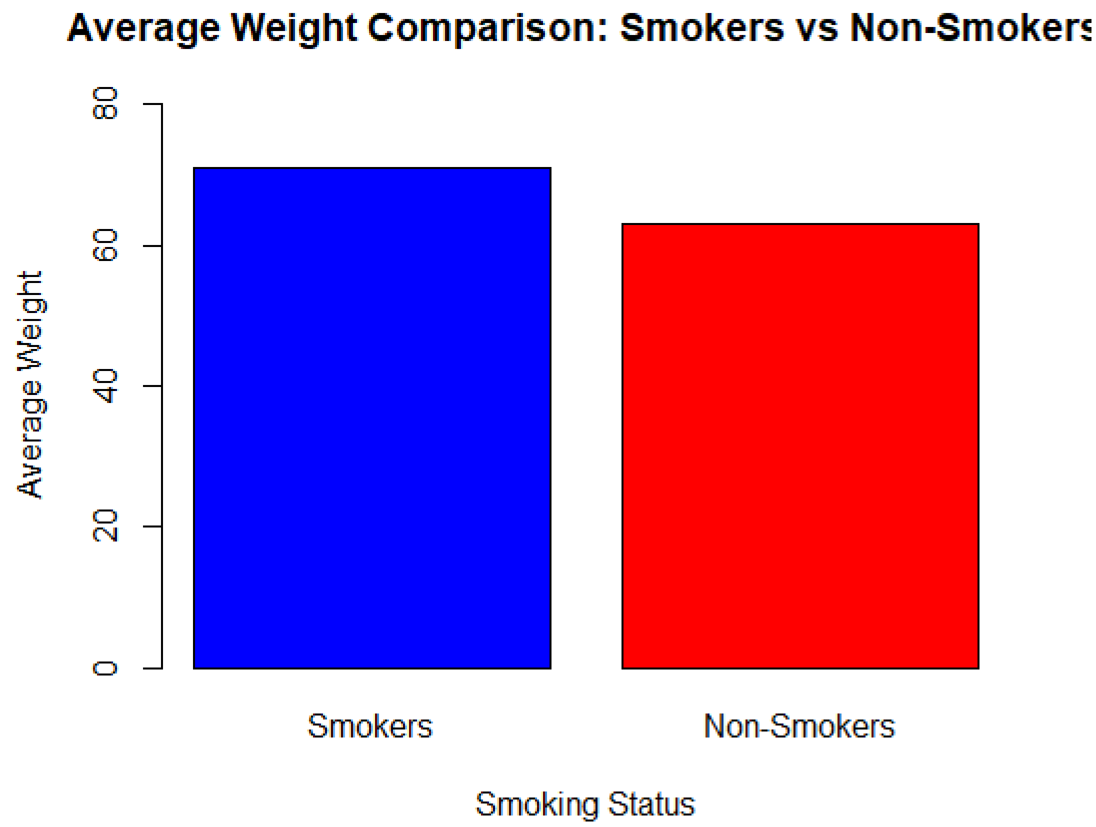
Observation: Upon analysing the data, we found that smokers constituted approximately 62% of the total population, while non-smokers accounted for approximately 38%. This significant difference in percentages suggests that smoking is prevalent among a substantial portion of our dataset.

```
Population <- table(data$smoking)
pie_chart <- ggplot(data.frame(Population), aes(x = "", y = smokers, fill =
names(Population))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  scale_fill_manual(values = c("1" = "lightblue", "0" = "lightgreen"),
                    labels = c("Smokers", "Non-Smokers")) +
  ggtitle("Smokers vs. Non-Smokers")
print(pie_chart)
```



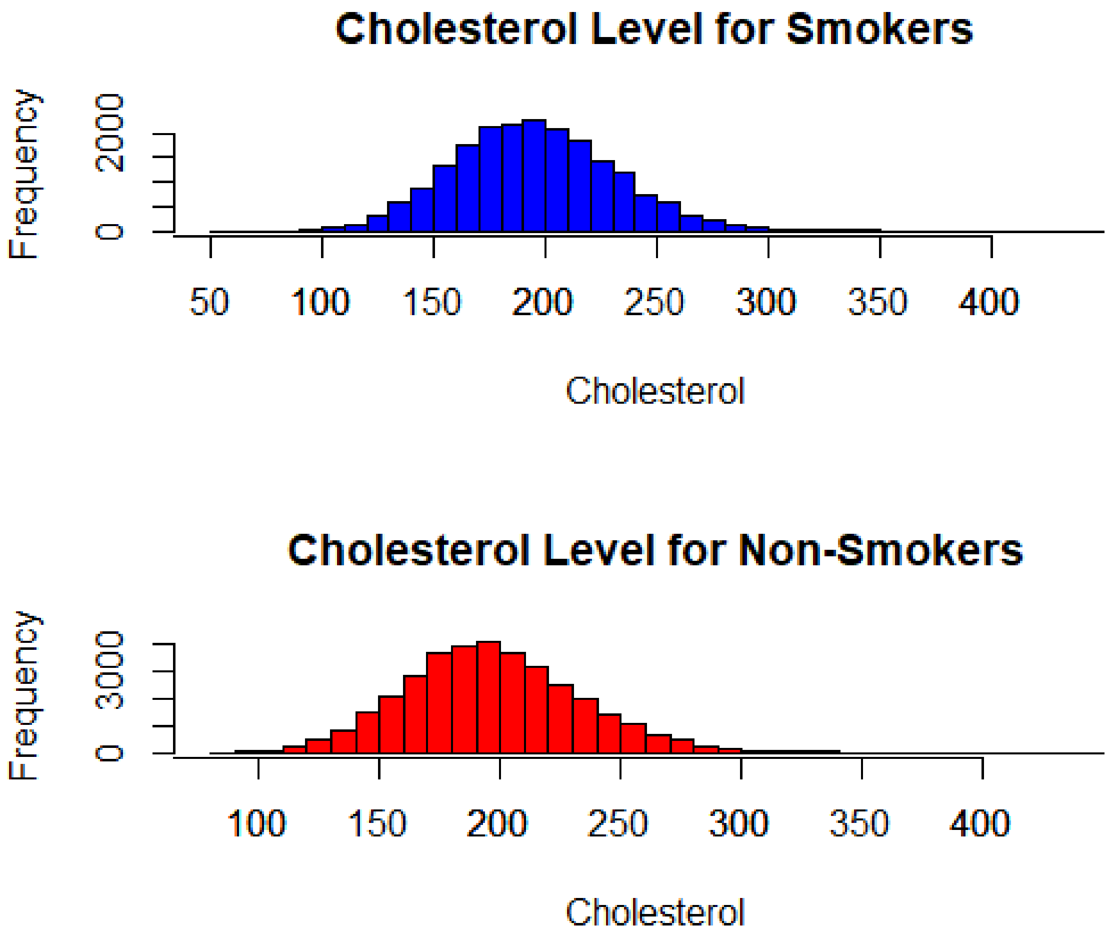
Smokers
Non-Smokers

Observation: The density plots of Diastolic and Systolic Blood Pressure for smokers and non-smokers show that the distributions are quite similar, And the overlapping areas indicate that there is a considerable overlap in the values of these variables between smokers and non-smokers. This suggests that there might not be a significant difference in their levels between the two groups.



Observation: The grouped bar plot reveals that the average weight of smokers is slightly higher compared to non-smokers, However, the difference in average weight between the two groups is not significant, as the ranges of the bars overlap substantially. This suggests that smoking may not have a substantial impact on weight in this dataset.

```
avg_weight_smokers <- mean(data$weight[data$smoking == 1])
avg_weight_non_smokers <- mean(data$weight[data$smoking == 0])
bar_data <- data.frame(Category = c("Smokers", "Non-Smokers"),
                       AverageWeight = c(avg_weight_smokers, avg_weight_non_smokers))
barplot(bar_data$AverageWeight, names.arg = bar_data$Category,
        xlab = "Smoking Status", ylab = "Average Weight",
        main = "Average Weight Comparison: Smokers vs Non-Smokers",
        col = c("blue", "red"), ylim = c(0, 80),
        beside = TRUE)
```



Observation: This indicates that the cholesterol levels for both smokers and non-smokers are relatively normally distributed. The absence of strong skewness or significant deviations suggests that Cholesterol levels in both groups are relatively consistent.

```
# Subset the data for smokers and non-smokers
smokers <- data[data$smoking == 1, ]
non_smokers <- data[data$smoking == 0, ]
par(mfrow = c(2, 1))
hist(smokers$Cholesterol, col = "blue", main = "Cholesterol Level for Smokers",
         xlab = "Cholesterol", ylab = "Frequency", breaks = 50)
axis(side = 1, at = seq(0, max(smokers$Cholesterol), by = 50))
hist(non_smokers$Cholesterol, col = "red", main = "Cholesterol Level for Non-Smokers",
         xlab = "Cholesterol", ylab = "Frequency", breaks = 50)
axis(side = 1, at = seq(0, max(non_smokers$Cholesterol), by = 50))
```

Data cleaning & Transformation

Checking for missing values: The code loops through each column and checks if there are any missing values in the dataset. It was found that there were no missing values present.

By using function : [anyNA(data)]

Checking for duplicates: By checking and removing any duplicate rows no duplicates was found

By using function : [anyDuplicated(data)]

Renaming columns : Original names changed to more descriptive names (relaxation) to (diastolic)

Removing of unnecessary data variables: height(cm) , waist(cm) , eyesight(left) , eyesight(right) , hearing(left) , hearing(right) , HDL , LDL , AST , ALT , GTP , GTP , Oral , tartar.

Hypothesis Testing:

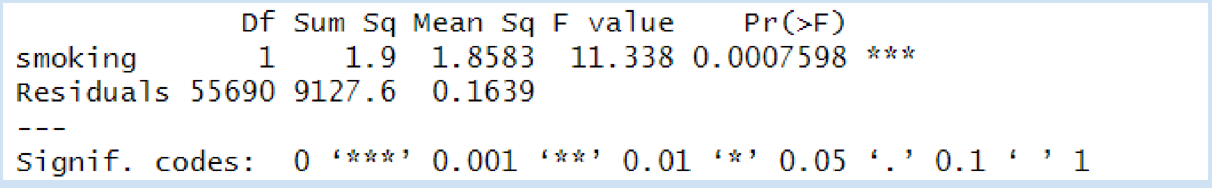
Test: Anova

Hypothesis: There is a significant association between smoking status (smokers vs. non-smokers) and the presence of urinary protein, and we'll test that now.

Null hypothesis(H0): there is no significant association between smoking status and the presence of urinary protein.

Alternative hypothesis(H1): there is a significant association between smoking status and the presence of urinary protein.

Conclusion: The p-value for the smoking variable in the ANOVA test is 0.0007598, which is less than 0.05, So we reject the null hypothesis and conclude that there is a significant association between smoking status and the presence of urinary protein.

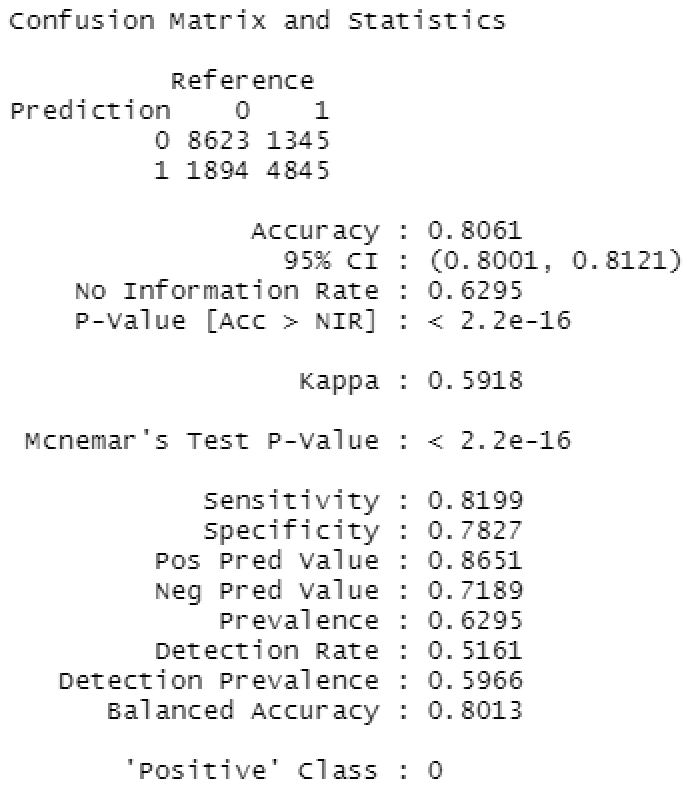


```
             Df Sum Sq Mean Sq F value    Pr(>F)    
smoking       1    1.9  1.8583  11.338 0.0007598 ***
Residuals 55690 9127.6  0.1639                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

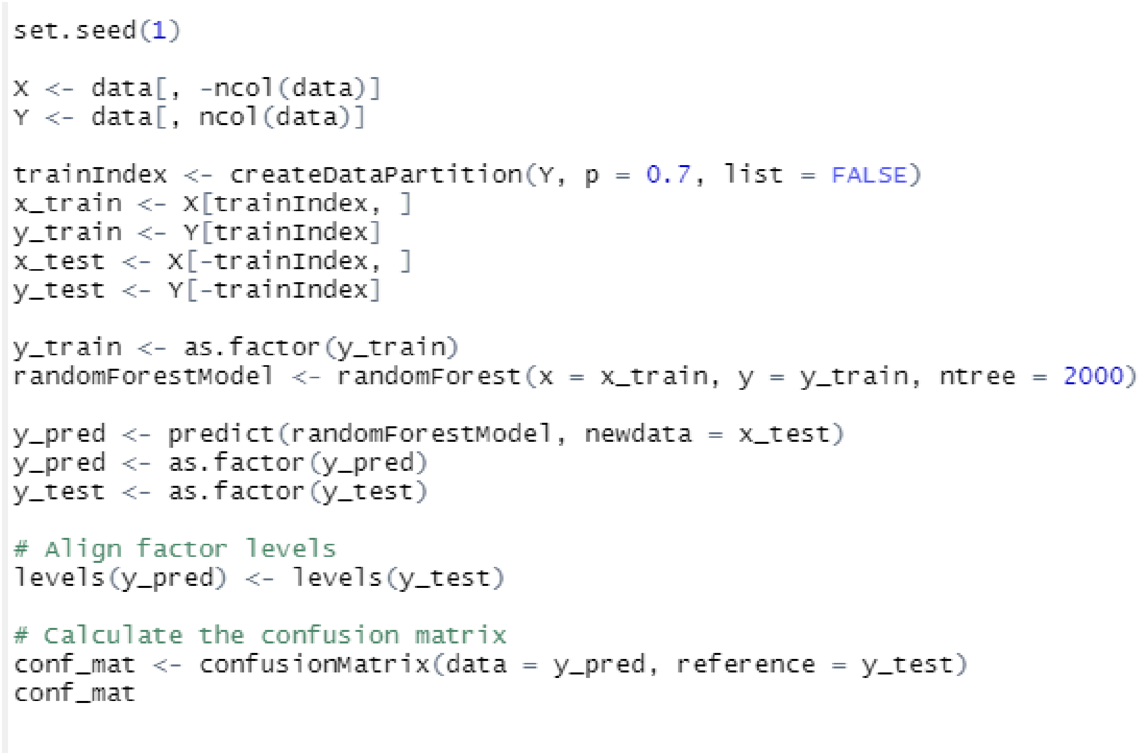
Data Analytics Technique Used:

Random Forest: Random Forest is an ensemble method that combines multiple decision trees to make predictions. Each decision tree in the random forest is trained on a random subset of the training data and the final prediction is made by aggregating the predictions of all the individual trees we have used Random Forest for its ability to handle large amounts of data, high-dimensional feature spaces, and feature interactions which we observed in our dataset.

The model had results which we can consider to be good enough and the confusion matrix is displayed below.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8623 1345
         1 1894 4845

               Accuracy : 0.8061
                 95% CI : (0.8001, 0.8121)
    No Information Rate : 0.6295
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5918

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8199
            Specificity : 0.7827
         Pos Pred Value : 0.8651
         Neg Pred Value : 0.7189
             Prevalence : 0.6295
         Detection Rate : 0.5161
   Detection Prevalence : 0.5966
      Balanced Accuracy : 0.8013

       'Positive' Class : 0
```



```r
set.seed(1)

X <- data[, -ncol(data)]
Y <- data[, ncol(data)]

trainIndex <- createDataPartition(Y, p = 0.7, list = FALSE)
x_train <- X[trainIndex, ]
y_train <- Y[trainIndex]
x_test <- X[-trainIndex, ]
y_test <- Y[-trainIndex]

y_train <- as.factor(y_train)
randomForestModel <- randomForest(x = x_train, y = y_train, ntree = 2000)

y_pred <- predict(randomForestModel, newdata = x_test)
y_pred <- as.factor(y_pred)
y_test <- as.factor(y_test)

# Align factor levels
levels(y_pred) <- levels(y_test)

# Calculate the confusion matrix
conf_mat <- confusionMatrix(data = y_pred, reference = y_test)
conf_mat
```