Domestic Violence Analysis

Sercan Tomaz

2024-07-14

```
df <- read.csv("domestic.violence.csv", sep = ",")
```

```
dim(df)
```

```
## [1] 347   7
```

```
head(df)
```

```
##   SL..No Age Education Employment Income Marital.status Violence
## 1      1  30 secondary unemployed      0        married      yes
## 2      2  47  tertiary unemployed      0        married       no
## 3      3  24  tertiary unemployed      0       unmarred       no
## 4      4  22  tertiary unemployed      0       unmarred       no
## 5      5  50   primary unemployed      0        married      yes
## 6      6  21  tertiary unemployed      0       unmarred      yes
```

```
str(df)
```

```
## 'data.frame':    347 obs. of  7 variables:
##  $ SL..No        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age           : int  30 47 24 22 50 21 30 27 20 18 ...
##  $ Education     : chr  "secondary" "tertiary" "tertiary" "tertiary" ...
##  $ Employment    : chr  "unemployed" "unemployed" "unemployed" "unemployed" ...
##  $ Income        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Marital.status: chr  "married" "married" "unmarred" "unmarred" ...
##  $ Violence      : chr  "yes" "no" "no" "no" ...
```

```
library(dplyr)
df <- select(df, -SL..No)
```

```
library(fastDummies)
df <- dummy_cols(df,
                 select_columns = c("Violence", "Marital.status"),
           remove_first_dummy = TRUE)
df$Marital.status_unmarred <- ifelse(df$Marital.status_unmarred == 1, 0, 1)
df <- select(df, -5 & -6)
names(df) [5] <- "Violence_Status"
names(df) [6] <- "Marital_Status"
df <- select(df, c(1, 2, 3, 4, 6, 5))
```

- When examining the Employment column, we observe an internationally employed column that appears to be categorized differently despite having the same label. Upon further inspection using the unique command, we notice that one of the employed columns contains a space. We will now correct this issue.
- Additionally, we will merge the Semi-Employed column into the main employed column to mitigate the imbalance in the classification problem. Subsequently, we will apply one-hot encoding to this category as well.

```
table(df$Employment)
```

```
##
##      employed     employed  semi employed    unemployed
##            23            3            47           274
```

```
unique(df$Employment)
```

```
## [1] "unemployed"    "semi employed" "employed"      "employed "
```

```
df$Employment <- ifelse(df$Employment == "employed ", "employed", df$Employment)
df$Employment <- ifelse(df$Employment == "semi employed", "employed", df$Employment)
```

- Now, upon re-examining the employment column, we observe that the data has been successfully modified.

```
table(df$Employment)
```

```
##
##   employed unemployed
##         73        274
```

- We assigned a dummy variable to the Employment column and subsequently restructured it so that employed entries are coded as 1 and unemployed entries as 0. Thereafter, we removed the original column and reindexed the columns accordingly.

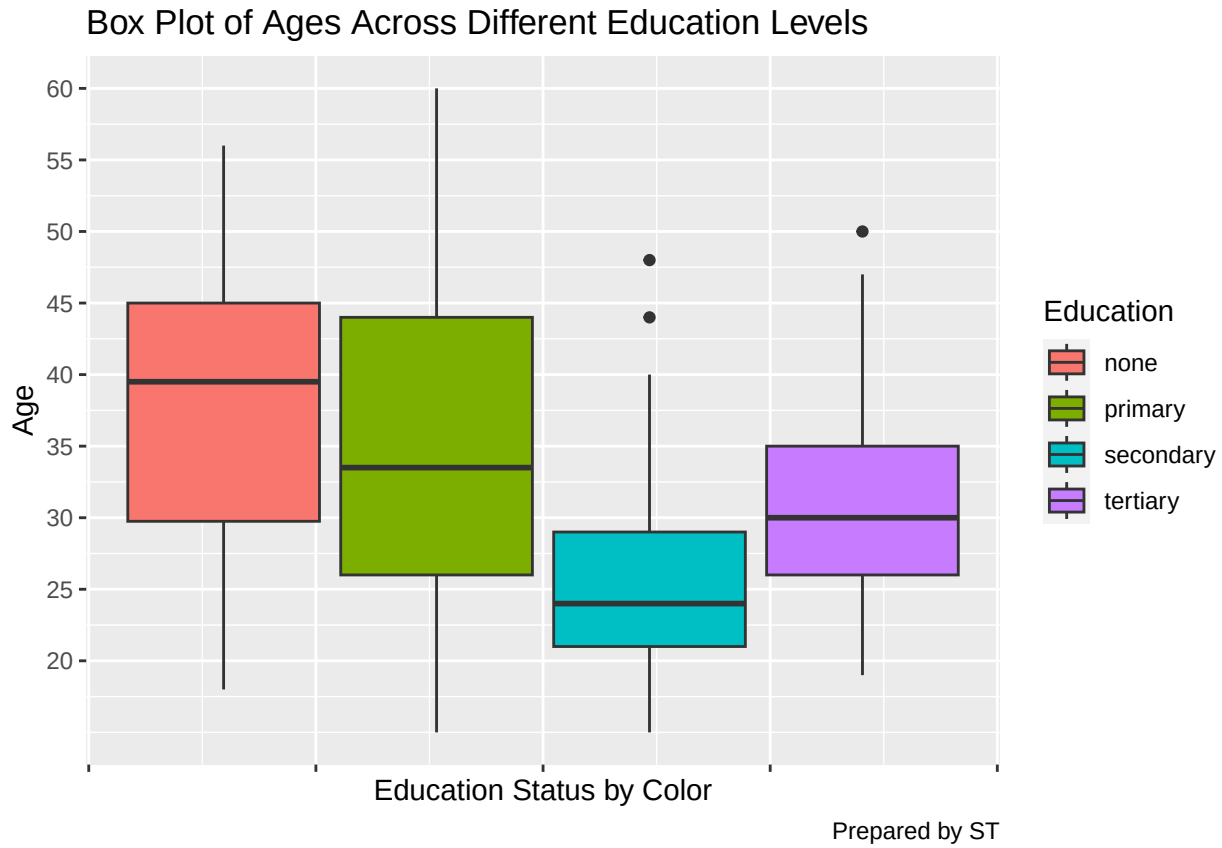
```
df <- dummy_cols(df,
                 select_columns = "Employment",
           remove_first_dummy = TRUE)
df$Employment_unemployed <- ifelse(df$Employment_unemployed == 1, 0, 1)
df <- select(df, -3)
df <- select(df, c(1, 2, 3, 4, 6, 5))
names(df) [5] <- "Employment_Status"
```

```
df <- dummy_cols(df,
                 select_columns = "Education",
           remove_first_dummy = TRUE)
df <- select(df, c(1, 2, 7, 8, 9, 3, 4, 5, 6))
```

- Upon examining the box plot, which displays the average ages across different education levels, the following observations can be made. The highest average age is observed among those without any formal education, whereas the lowest average age is found in the secondary education level.

- Additionally, outliers are present in both the secondary and tertiary education levels. Analyzing the average ages individually, we observe that individuals with no formal education have the highest average age at 37.5 years. This value is 34.6 years for primary education, 30.9 years for tertiary education, and the lowest average age of 25.1 years is seen in the secondary education level. On the other hand, the overall mean is 31.4

```
library(ggplot2)
df %>%
  ggplot(aes(y = Age)) +
  geom_boxplot(aes(fill = Education)) +
  theme(axis.text.x = element_blank()) +
  ggtitle("Box Plot of Ages Across Different Education Levels") +
  labs(x = "Education Status by Color",
       caption = "Prepared by ST") +
  scale_y_continuous(breaks = seq(20, 60, by = 5))
```

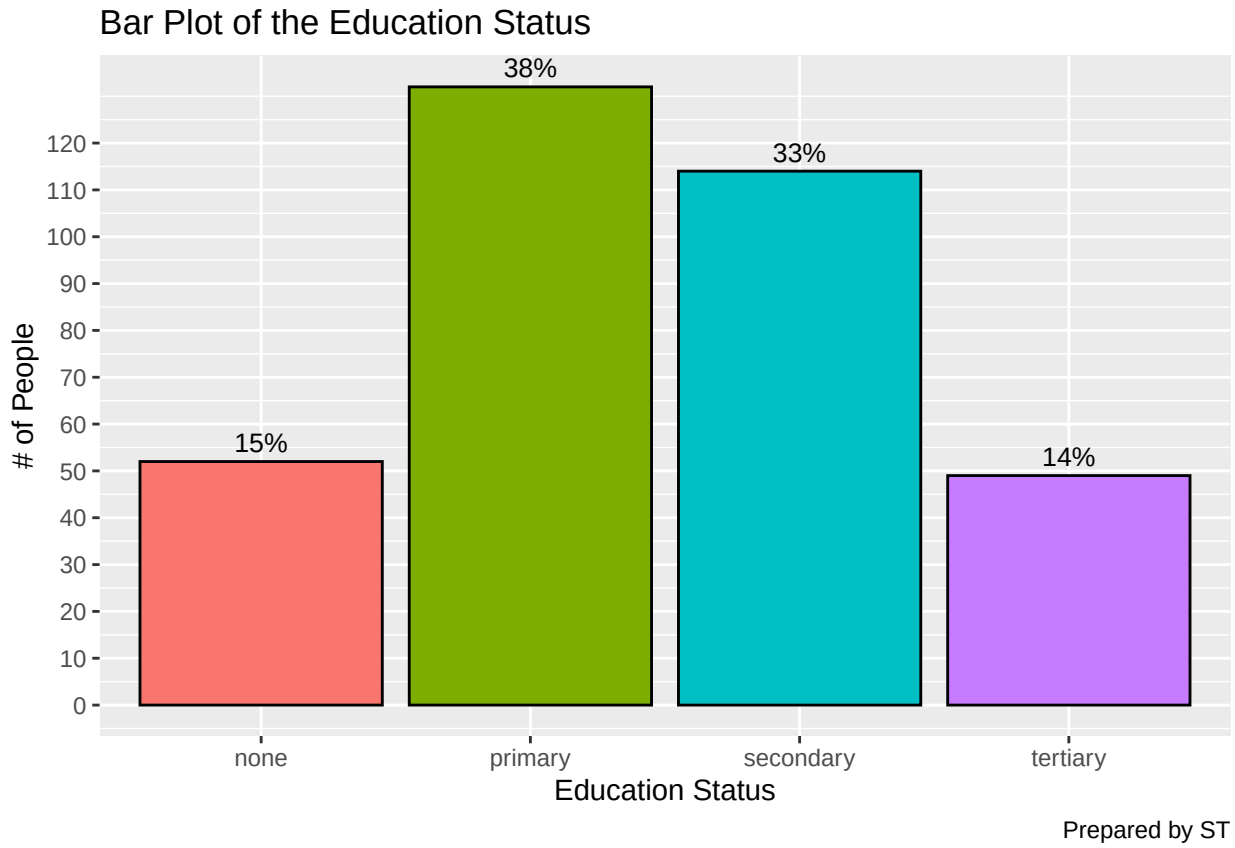



```
mean.ages.all <-df %>%
  group_by(Education) %>%
  summarise(mean.age.by.education = round(mean(Age), 1)) %>%
  rbind(mean(df$Age))
mean.ages.all [5, 1] <- "Overall Mean"
mean.ages.all$mean.age.by.education <- round(mean.ages.all$mean.age.by.education, 1)
```

- Upon examining our bar plot, we observe that individuals with primary education status are in the majority, while those with tertiary education status are in the minority. Additionally, the number of

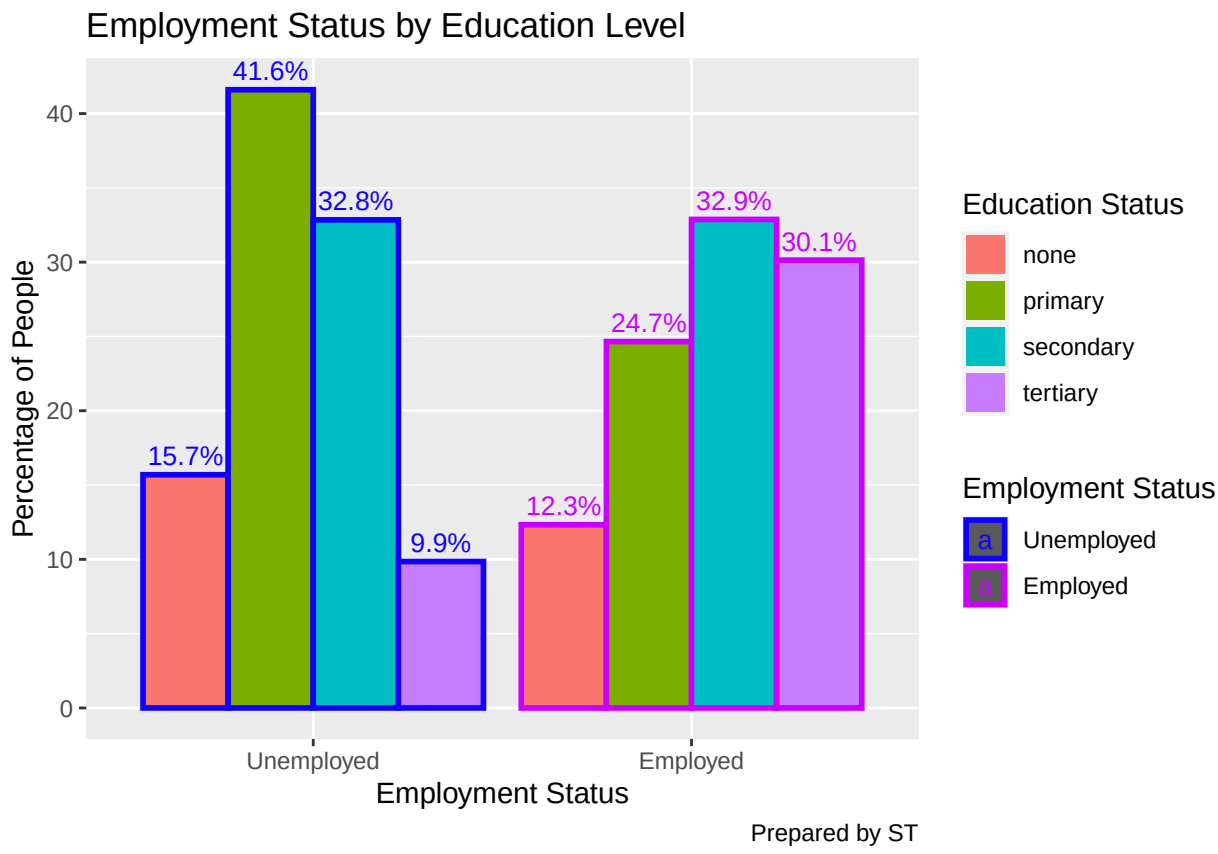
individuals with secondary education status is 114, whereas the number of individuals without any formal education is 52.

```
ggplot(df, aes(x = Education, fill = Education)) +
  geom_bar(color = "black", show.legend = FALSE) +
  scale_y_continuous(breaks = seq(0, 120, by = 10)) +
  ggtitle("Bar Plot of the Education Status") +
  labs(x = "Education Status",
       y = "# of People",
       caption = "Prepared by ST") +
  geom_text(stat = "count", aes(label = paste0(round(..count../sum(..count..)*100), "%")),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5)
```



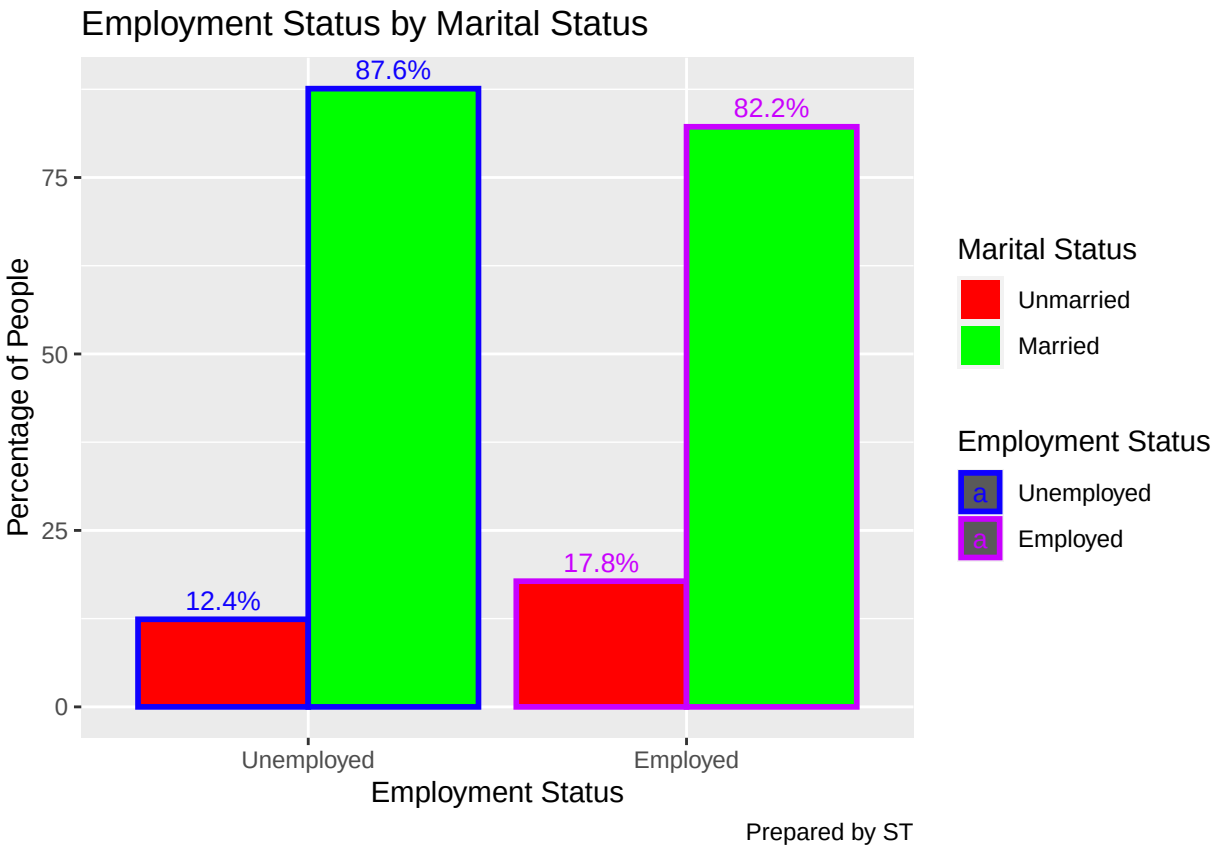
- The data indicates a clear trend where individuals with higher education levels (secondary and tertiary) are more likely to be employed.
- A significant number of individuals with only primary education or no education are unemployed, highlighting a potential area for policy intervention to improve educational opportunities and employment rates.
- While higher education generally correlates with better employment status, it does not guarantee employment, as evidenced by the 27 unemployed individuals with tertiary education.
- This comparison suggests that primary education might be more common among the unemployed, whereas tertiary education shows a trend towards employment.

```r
df_count_3 <- df %>%
  group_by(Employment_Status, Education) %>%
  summarise(count = n()) %>%
  group_by(Employment_Status) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(df_count_3, aes(x = factor(Employment_Status), y = percentage, fill = Education, color = factor(
  geom_bar(position = "dodge", stat = "identity", size = 1) +
  labs(title = "Employment Status by Education Level",
       x = "Employment Status",
       y = "Percentage of People",
       fill = "Education Status",
       caption = "Prepared by ST") +
  scale_color_manual(values = c("0" = "#0F00FF",
                                "1" = "#C900FF"),
                     name = "Employment Status",
                     labels = c("0" = "Unemployed",
                                "1" = "Employed")) +
  scale_x_discrete(labels = c("0" = "Unemployed",
                              "1" = "Employed")) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5)
```



- It appears that a significant majority of both employed and unemployed individuals are married. However, the proportion of married individuals is higher among the unemployed compared to the employed. Unmarried individuals constitute a smaller proportion in both employment statuses.
- A slightly higher proportion of employed individuals are unmarried compared to their unemployed counterparts.

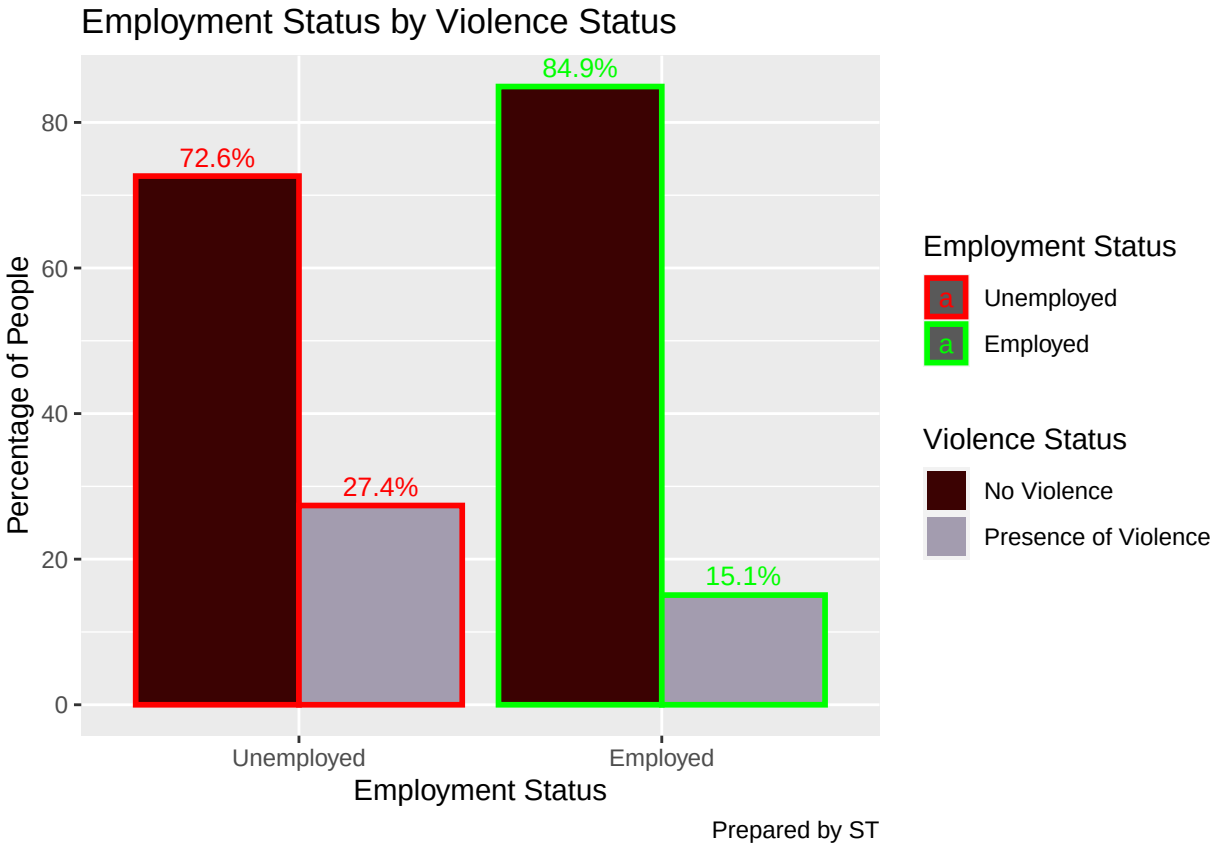
```
df_count_2 <- df %>%
  group_by(Employment_Status, Marital_Status) %>%
  summarise(count = n()) %>%
  mutate(percentage_2 = count / sum(count) * 100)

ggplot(df_count_2, aes(x = factor(Employment_Status), y = percentage_2, fill = factor(Marital_Status),
  geom_bar(position = "dodge", stat = "identity", size = 1) +
  labs(title = "Employment Status by Marital Status",
       x = "Employment Status",
       y = "Percentage of People",
       fill = "Marital Status",
       caption = "Prepared by ST") +
  scale_color_manual(values = c("0" = "#0F00FF",
                                "1" = "#C900FF"),
                     name = "Employment Status",
                     labels = c("0" = "Unemployed",
                                "1" = "Employed")) +
  scale_x_discrete(labels = c("0" = "Unemployed",
                              "1" = "Employed")) +
  scale_fill_manual(values = c("0" = "red",
                               "1" = "green"),
                    name = "Marital Status",
                    labels = c("0" = "Unmarried",
                               "1" = "Married")) +
  geom_text(aes(label = paste0(round(percentage_2, 1), "%")),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5)
```



- Unemployment appears to correlate with a higher likelihood of experiencing violence, while employment, though offering some protection, does not eliminate the risk entirely. Addressing these disparities requires comprehensive strategies that address societal, economic, and safety factors affecting individuals in both employment sectors.

```
df_count <- df %>%
  group_by(Employment_Status, Violence_Status) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(df_count, aes(x = factor(Employment_Status),
                     y = percentage,
                     fill = factor(Violence_Status),
                     color = factor(Employment_Status))) +
  geom_bar(stat = "identity", position = "dodge", size = 1) +
  labs(title = "Employment Status by Violence Status",
       x = "Employment Status",
       y = "Percentage of People",
       fill = "Violence Status",
       caption = "Prepared by ST") +
  scale_color_manual(values = c("0" = "red",
                                "1" = "green"),
                     name = "Employment Status",
                     labels = c("0" = "Unemployed",
                                "1" = "Employed")) +
```

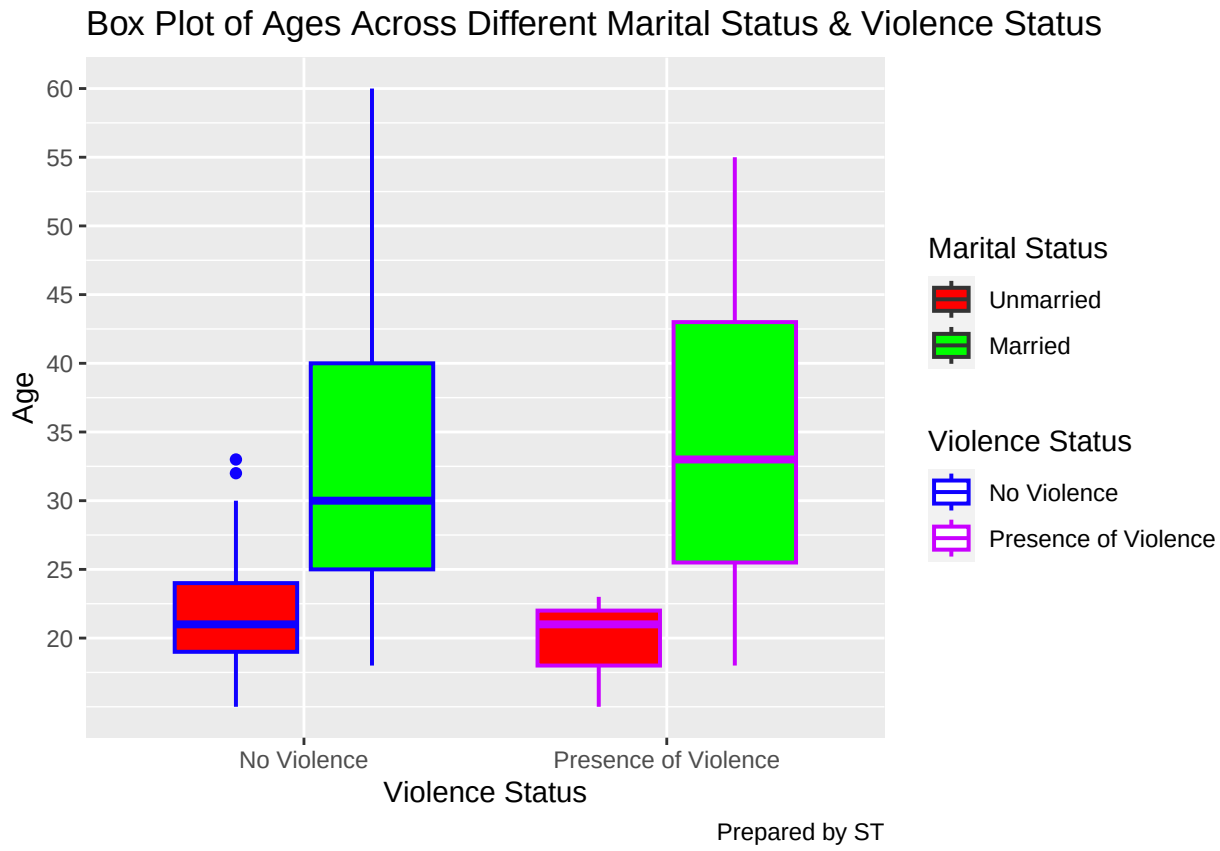
```
  scale_x_discrete(labels = c("0" = "Unemployed",
                              "1" = "Employed")) +
  scale_fill_manual(values = c("0" = "#3B0202",
                               "1" = "#A39CAF"),
                    name = "Violence Status",
                    labels = c("0" = "No Violence",
                               "1" = "Presence of Violence")) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5)
```



- The box plot presents a clear contrast in age distribution between unmarried and married individuals. For the unmarried group (represented by the red box plot), the median age is approximately 22 years, with the interquartile range (IQR) predominantly spanning from around 20 to 25 years. A notable feature in this group is the presence of outliers above 30 years, indicating that while the majority of unmarried individuals are young, there are a few significantly older individuals in this category.
- In contrast, the married group (represented by the green box plot) has a median age of approximately 32 years. The IQR for this group extends from around 25 to 40 years, showcasing a wider age range compared to the unmarried group. This broader range suggests more variability in the ages of married individuals. Additionally, there are no apparent outliers in the married group, indicating a more consistent age distribution within this category.
- Comparing the two groups reveals that married individuals tend to be older than unmarried individuals, as evidenced by the higher median age and broader age range. The significant difference in median ages suggests that marriage is more prevalent among older individuals in this dataset. Furthermore,

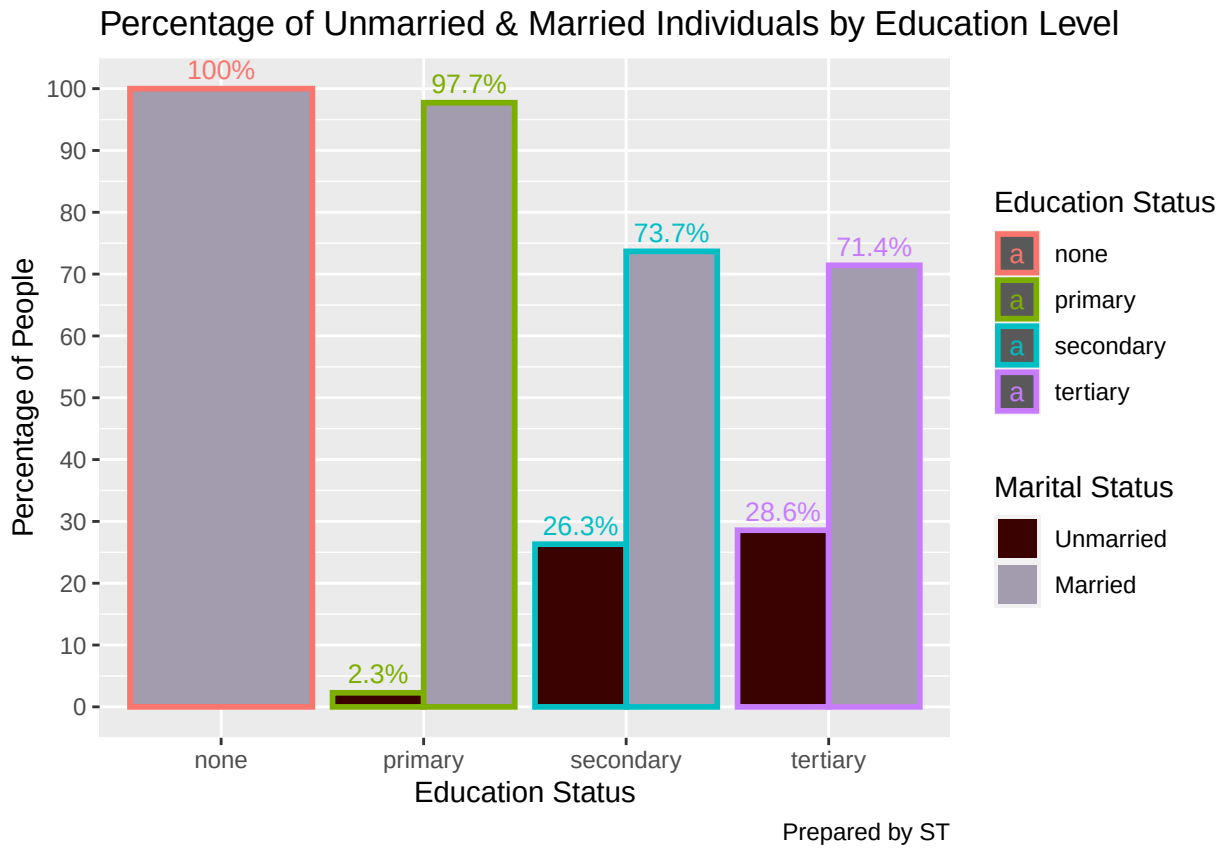
the greater variability in the married group indicates a wider span of ages, whereas the unmarried group shows a more concentrated age range around the early twenties. These differences highlight the potential life stage differences between the two groups, with marriage being more common among older individuals.

```
ggplot(df, aes(x = factor(Violence_Status), y = Age)) +
  geom_boxplot(aes(fill = factor(Marital_Status),
                   color = factor(Violence_Status)),
               size = 0.7) +
  scale_y_continuous(breaks = seq(20, 60, by = 5)) +
  scale_fill_manual(values = c("0" = "red",
                               "1" = "green"),
                    name = "Marital Status",
                    labels = c("0" = "Unmarried",
                               "1" = "Married")) +
  scale_x_discrete(labels = c("0" = "No Violence",
                              "1" = "Presence of Violence")) +
  labs(x = "Violence Status",
       caption = "Prepared by ST") +
  ggtitle("Box Plot of Ages Across Different Marital Status & Violence Status") +
  scale_color_manual(values = c("0" = "#0F00FF",
                                "1" = "#C900FF"),
                     name = "Violence Status",
                     labels = c("0" = "No Violence",
                                "1" = "Presence of Violence"))
```



- All individuals with no education are married. This indicates a strong correlation between having no formal education and being married.
- The vast majority of individuals with primary education are married. A small minority, 2.3%, are unmarried.
- A significant portion of individuals with secondary education are unmarried (26.3%). However, the majority (73.7%) are married, indicating that secondary education does not heavily favor one marital status over the other as strongly as primary or none education statuses do.
- Similar to secondary education, a noticeable percentage of individuals with tertiary education are unmarried (28.6%), while the majority are married (71.4%). This suggests a somewhat balanced distribution in this education level, with a slight inclination towards being married.
- The bar plot indicates that individuals with higher levels of education (secondary and tertiary) have a more balanced distribution between unmarried and married statuses. In contrast, individuals with lower levels of education (none and primary) are predominantly married.

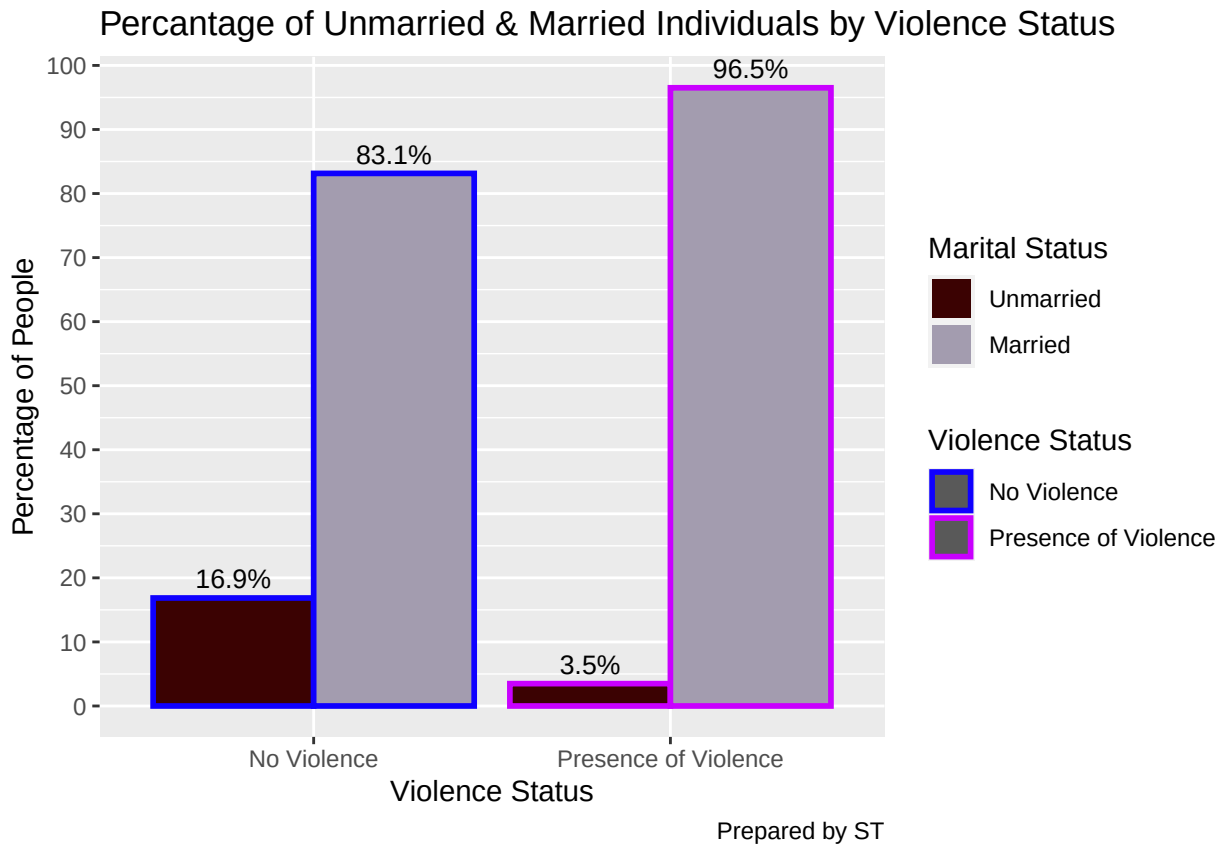
```
df_count_4 <- df %>%
  group_by(Education, Marital_Status) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(df_count_4, aes(x = Education, y = percentage, fill = factor(Marital_Status), color = Education))
  geom_bar(size = 1, position = "dodge", stat = "identity") +
  scale_y_continuous(breaks = seq(0, 120, by = 10)) +
  scale_fill_manual(values = c("0" = "#3B0202",
                               "1" = "#A39CAF"),
                    name = "Marital Status",
                    labels = c("0" = "Unmarried",
                               "1" = "Married")) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5) +
  labs(x = "Education Status",
       y = "Percentage of People",
       caption = "Prepared by ST",
       color = "Education Status") +
  ggtitle("Percentage of Unmarried & Married Individuals by Education Level")
```



- The data presented in the chart highlights a significant disparity in the experience of violence between married and unmarried individuals. Married individuals are much more likely to report experiencing violence compared to their unmarried counterparts. Conversely, the majority of individuals who do not experience violence are predominantly married.
- These findings suggest that marital status plays a crucial role in the likelihood of experiencing violence. The markedly higher incidence of violence among married individuals warrants further investigation to understand the underlying causes. It is imperative to explore why married individuals are more susceptible to violence and to develop targeted interventions and policies to address and mitigate this issue. The low incidence of violence among unmarried individuals also provides a point of comparison that may help in understanding the dynamics at play in different marital statuses.
- Further research is necessary to delve deeper into these correlations and to formulate effective strategies for prevention and support for those affected by violence, particularly within the context of marriage.

```
df_count_5 <- df %>%
  group_by(Violence_Status, Marital_Status) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(df_count_5, aes(x = factor(Violence_Status),
                       y = percentage)) +
  geom_bar(position = "dodge",
           stat = "identity",
           aes(fill = factor(Marital_Status),
               color = factor(Violence_Status)),
           size = 1) +
```

```
  scale_y_continuous(breaks = seq(0, 100, by = 10)) +
  scale_fill_manual(values = c("0" = "#3B0202",
                               "1" = "#A39CAF"),
                    name = "Marital Status",
                    labels = c("0" = "Unmarried",
                               "1" = "Married")) +
  scale_color_manual(values = c("0" = "#0F00FF",
                                "1" = "#C900FF"),
                     name = "Violence Status",
                     labels = c("0" = "No Violence",
                                "1" = "Presence of Violence")) +
  scale_x_discrete(labels = c("0" = "No Violence",
                              "1" = "Presence of Violence")) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_dodge2(width = 0.9),
            vjust = -0.5,
            size = 3.5,
            padding = 0.1) +
  labs(x = "Violence Status",
       y = "Percentage of People",
       caption = "Prepared by ST") +
  ggtitle("Percantage of Unmarried & Married Individuals by Violence Status")
```

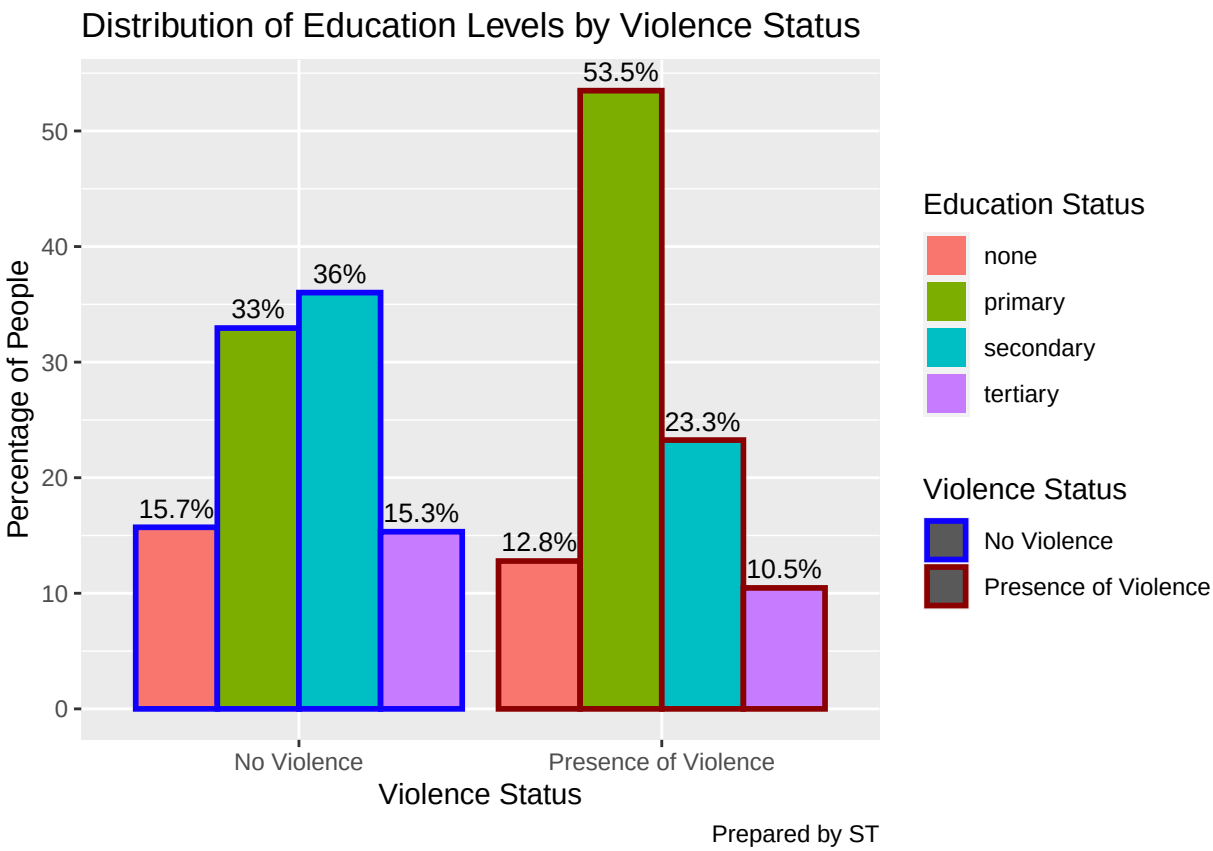


- The data indicates a significant disparity in the experience of violence based on education levels. Among those who did not experience violence, secondary education is the most prevalent, followed by

primary education. In contrast, the majority of individuals who experienced violence had only primary education, highlighting a potential vulnerability within this group.

- The chart underscores the potential link between education level and the experience of violence. The high incidence of violence among individuals with only primary education points to the need for targeted interventions and educational programs to address this vulnerability. Additionally, the lower prevalence of violence among individuals with higher education levels suggests that enhancing educational opportunities could be a strategic approach to mitigate the risk of violence.
- Further research should explore the socio-economic factors that contribute to these patterns and develop comprehensive strategies to support and protect individuals with lower educational attainment, particularly those with primary education. The data provides a foundation for policymakers to consider education as a critical factor in violence prevention and to promote educational initiatives as part of broader violence mitigation efforts.

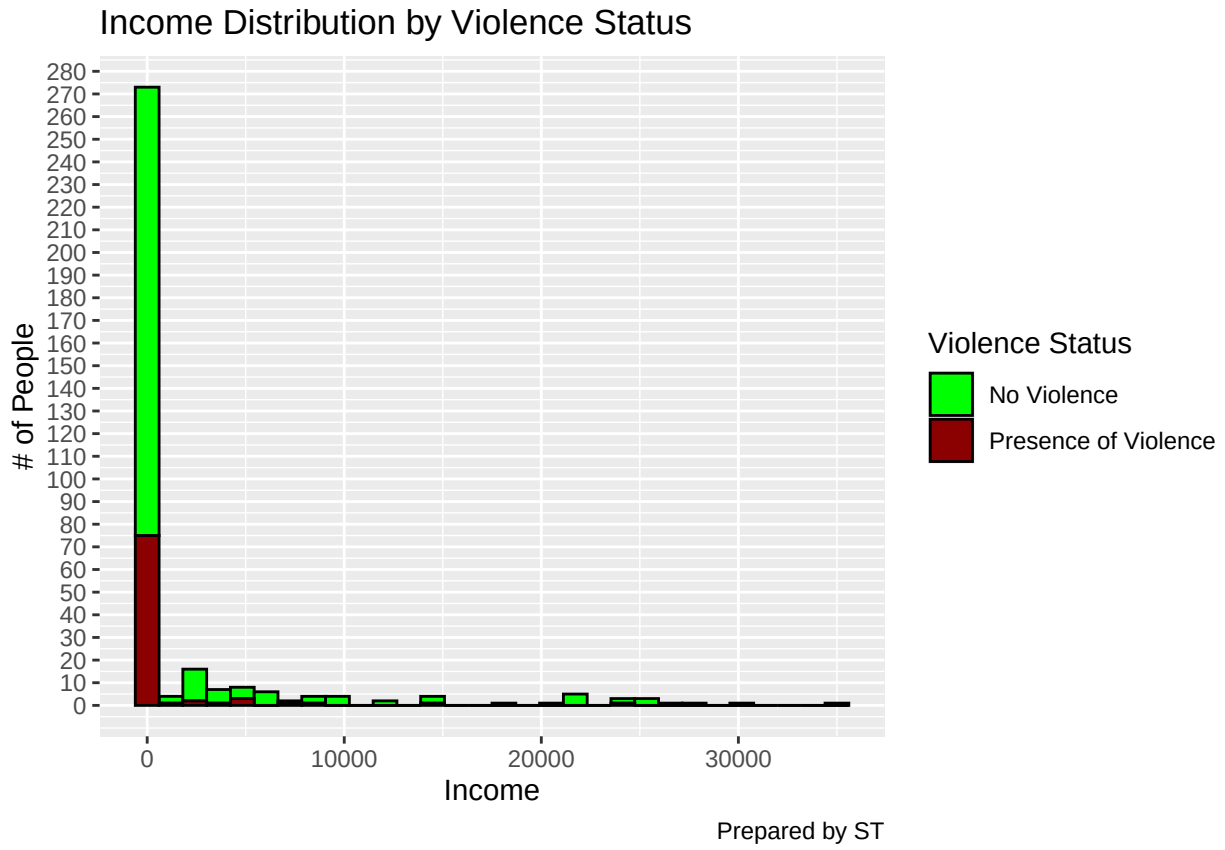
```r
df_count_6 <- df %>%
  group_by(Violence_Status, Education) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(df_count_6, aes(x = factor(Violence_Status),
                       y = percentage)) +
  geom_bar(position = "dodge",
           stat = "identity",
           aes(fill = factor(Education),
               color = factor(Violence_Status)),
           size = 1) +
  scale_y_continuous(breaks = seq(0, 50, by = 10)) +
  labs(x = "Violence Status",
       y = "Percentage of People",
       caption = "Prepared by ST",
       fill = "Education Status") +
  ggtitle("Distribution of Education Levels by Violence Status") +
  scale_x_discrete(labels = c("0" = "No Violence",
                              "1" = "Presence of Violence")) +
  scale_color_manual(values = c("0" = "#0F00FF",
                                "1" = "darkred"),
                     name = "Violence Status",
                     labels = c("0" = "No Violence",
                                "1" = "Presence of Violence")) +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_dodge2(width = 0.9),
            vjust = -0.5,
            size = 3.5,
            padding = 0.1)
```



- The chart illustrates the distribution of income among individuals, segmented by their experience of violence. The x-axis represents the income levels, while the y-axis denotes the number of people. The green bars indicate individuals who have not experienced violence, and the maroon bars represent those who have experienced violence.
- The dominant observation is the large number of individuals with zero income, which includes both those who have and have not experienced violence. However, the proportion of individuals not experiencing violence is substantially higher within this group.
- The chart indicates that the incidence of violence is more prominent among individuals with zero income compared to those with any positive income. As income levels rise, the occurrence of violence diminishes significantly.
- Higher income levels have a very sparse representation in the dataset, with almost no individuals reporting violence.

```
ggplot(df, aes(x = Income)) +
  geom_histogram(aes(fill = factor(Violence_Status)),
                 color = "black",
                 bins = 30) +
  scale_y_continuous(breaks = seq(0, 300, by = 10)) +
  scale_fill_manual(values = c("0" = "green",
                               "1" = "darkred"),
                    name = "Violence Status",
                    labels = c("0" = "No Violence",
                               "1" = "Presence of Violence")) +
  labs(x = "Income",
       y = "# of People",
```

```
      caption = "Prepared by ST") +
  ggtitle("Income Distribution by Violence Status")
```

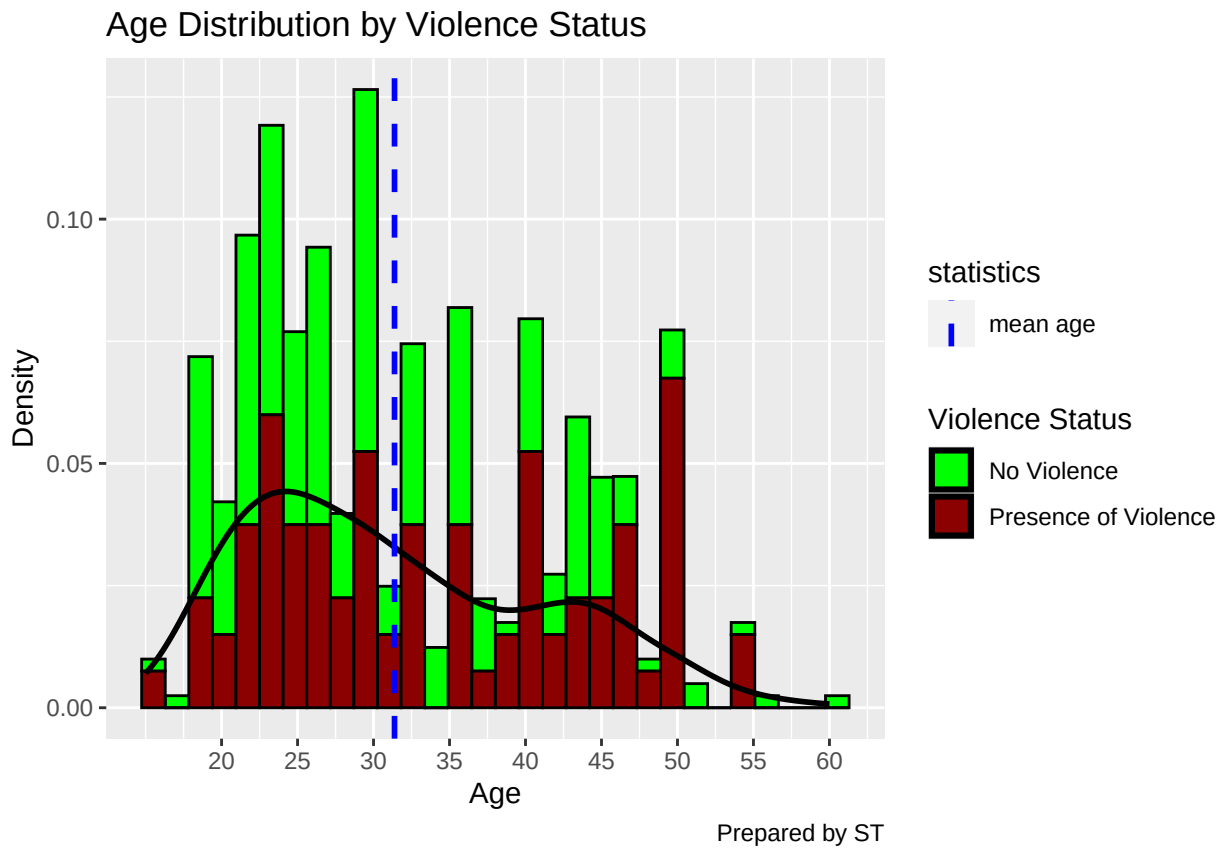


- The graph shows the age distribution of individuals with a history of victimization by violence.
- The data includes individuals between the ages of 20 and 60.
- The data is represented by density bars, which show the likelihood of victimization by violence for each age group.
- According to the graph, the risk of victimization by violence peaks between the ages of 30 and 40.
- The risk of victimization by violence in this age group is higher than in the 20 and 50 age groups.
- The risk of victimization by violence is lower for individuals aged 60 and older.

```
df %>%
  ggplot(aes(x = Age)) +
  geom_histogram(aes(y = ..density.., fill = factor(Violence_Status)),
                 color = "black") +
  geom_density(size = 1) +
  ggtitle("Age Distribution by Violence Status") +
  labs(y = "Density",
       caption = "Prepared by ST") +
  geom_vline(aes(xintercept = mean(Age),
                 color = "mean age"),
             linetype = "dashed",
             size = 1) +
  scale_x_continuous(breaks = seq(20, 60, by = 5)) +
  scale_color_manual(name = "statistics",
```

```
                      values = c("mean age" = "blue")) +
  scale_fill_manual(values = c("0" = "green",
                               "1" = "darkred"),
                    name = "Violence Status",
                    labels = c("0" = "No Violence",
                               "1" = "Presence of Violence"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- The provided chart visually represents the correlations and relationship intensities among a set of variables, utilizing Pearson correlation coefficients and a color scale to denote the direction (positive in blue, negative in red) and magnitude of these correlations. Additionally, the small pie charts within each cell offer a visual emphasis on the strength of these correlations.
- Key observations from the chart are as follows:

Income:

- There is a high positive correlation with Education_tertiary (0.71), indicating that higher education levels are associated with higher income.
- A moderate positive correlation with Age (0.40) suggests that income tends to increase with age. Education_primary:
- Exhibits a negative correlation with both Education_secondary (-0.32) and Education_tertiary (-0.55). This implies an inverse relationship between primary education levels and higher education

levels, indicating that individuals with only primary education are less likely to pursue secondary or tertiary education.

Age:

- Shows a negative correlation with Education_primary (-0.46) and Education_secondary (-0.28). This may indicate that as age increases, the prevalence of only primary or secondary education decreases, possibly due to older individuals having had fewer opportunities for higher education.

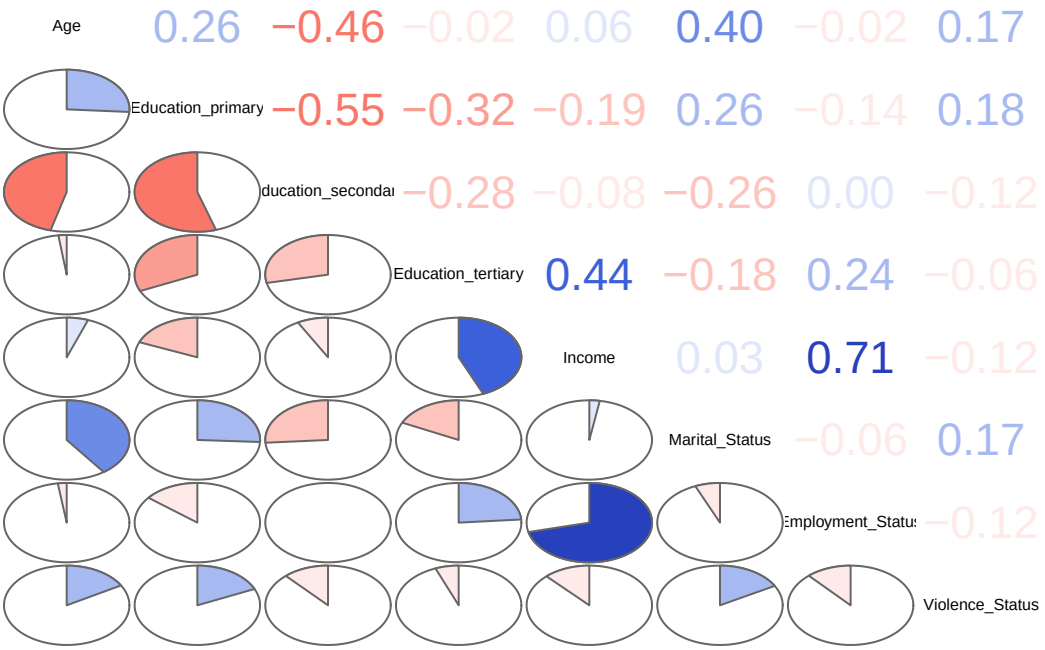
Employment_Status:

- Displays a positive correlation with Income (0.24), suggesting that employment status positively influences income, with employed individuals likely to earn more.

Marital_Status and Violence_Status:

- The correlations with other variables are generally low or insignificant, indicating weak relationships between these variables and the others examined in this analysis.

```
library(corrgram)
corrgram(df, lower.panel = panel.pie,
         upper.panel = panel.cor)
```



Machine Learning Models

- Now, we will build models using different machine learning algorithms and evaluate the performance of these models using various metrics.
- For applying machine learning models, we designate our response variable and other necessary variables as a factor.
- First, let's split our dataset into a training set and a test set using a ratio of 0.7.

Classification Tree

```
library(caTools)
df$Violence_Status <- as.factor(df$Violence_Status)
df$Education_primary <- as.factor(df$Education_primary)
df$Education_secondary <- as.factor(df$Education_secondary)
df$Education_tertiary <- as.factor(df$Education_tertiary)
df$Marital_Status <- as.factor(df$Marital_Status)
df$Employment_Status <- as.factor(df$Employment_Status)
set.seed(101)
split <- sample.split(df$Violence_Status, SplitRatio = 0.7)
test  <- subset(df, split == FALSE)
train <- subset(df, split == TRUE)
```

```
library(caret)
library(rpart)
library(rpart.plot)
param.grid.ct <- expand.grid(cp = seq(0.01, 1, by = 0.01))
```

```
set.seed(101)
ctrl.ct <- trainControl(method = "cv",
                        number = 5)
```

```
set.seed(101)
parameter.search.ct <- train(Violence_Status ~.,
                             data = train,
                             method = "rpart",
                             trControl = ctrl.ct,
                             tuneGrid = param.grid.ct)
```

```
parameter.search.ct$bestTune$cp
```

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

```
ct.model <- rpart(Violence_Status ~.,
                  data = train,
                  cp = 0.09)
```

```
ct.preds <- predict(ct.model, test, type = "class")
```

```r
confusionMatrix(ct.preds, test$Violence_Status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 78 26
##          1  0  0
##
##                Accuracy : 0.75
##                  95% CI : (0.6555, 0.8297)
##     No Information Rate : 0.75
##     P-Value [Acc > NIR] : 0.5524
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 9.443e-07
##
##             Sensitivity : 1.00
##             Specificity : 0.00
##          Pos Pred Value : 0.75
##          Neg Pred Value :  NaN
##              Prevalence : 0.75
##          Detection Rate : 0.75
##    Detection Prevalence : 1.00
##       Balanced Accuracy : 0.50
##
##        'Positive' Class : 0
##
```

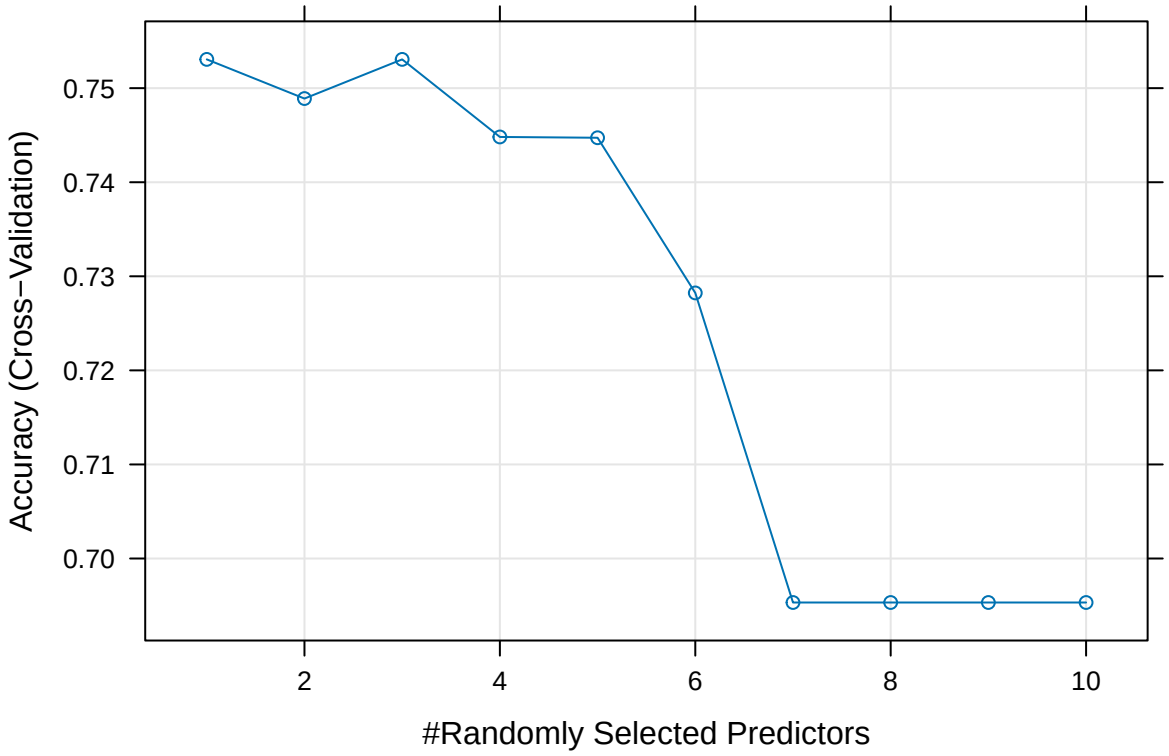
Random Forest

```r
param.grid.rf <- expand.grid(mtry = 1:10)
```

```r
set.seed(101)
ctrl.rf <- trainControl(method = "cv",
                        number = 5)
```

```r
parameter.search.rf <- train(Violence_Status ~.,
                             data = train,
                             method = "rf",
                             trControl = ctrl.rf,
                             tuneGrid = param.grid.rf)
```

```r
plot(parameter.search.rf)
```




```
library(randomForest)
set.seed(101)
rf.model <- randomForest(Violence_Status~.,
                         train,
                         mtry = parameter.search.rf$bestTune$mtry,
                         ntree = 10)
```

```
rf.preds <- predict(rf.model, test)
```

```
confusionMatrix(rf.preds, test$Violence_Status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 73 24
##          1  5  2
##
##                Accuracy : 0.7212
##                  95% CI : (0.6247, 0.8046)
##     No Information Rate : 0.75
##     P-Value [Acc > NIR] : 0.7880308
##
##                   Kappa : 0.0169
##
```

```
##  Mcnemar's Test P-Value : 0.0008302
##
##             Sensitivity : 0.93590
##             Specificity : 0.07692
##          Pos Pred Value : 0.75258
##          Neg Pred Value : 0.28571
##              Prevalence : 0.75000
##          Detection Rate : 0.70192
##    Detection Prevalence : 0.93269
##       Balanced Accuracy : 0.50641
##
##        'Positive' Class : 0
##
```

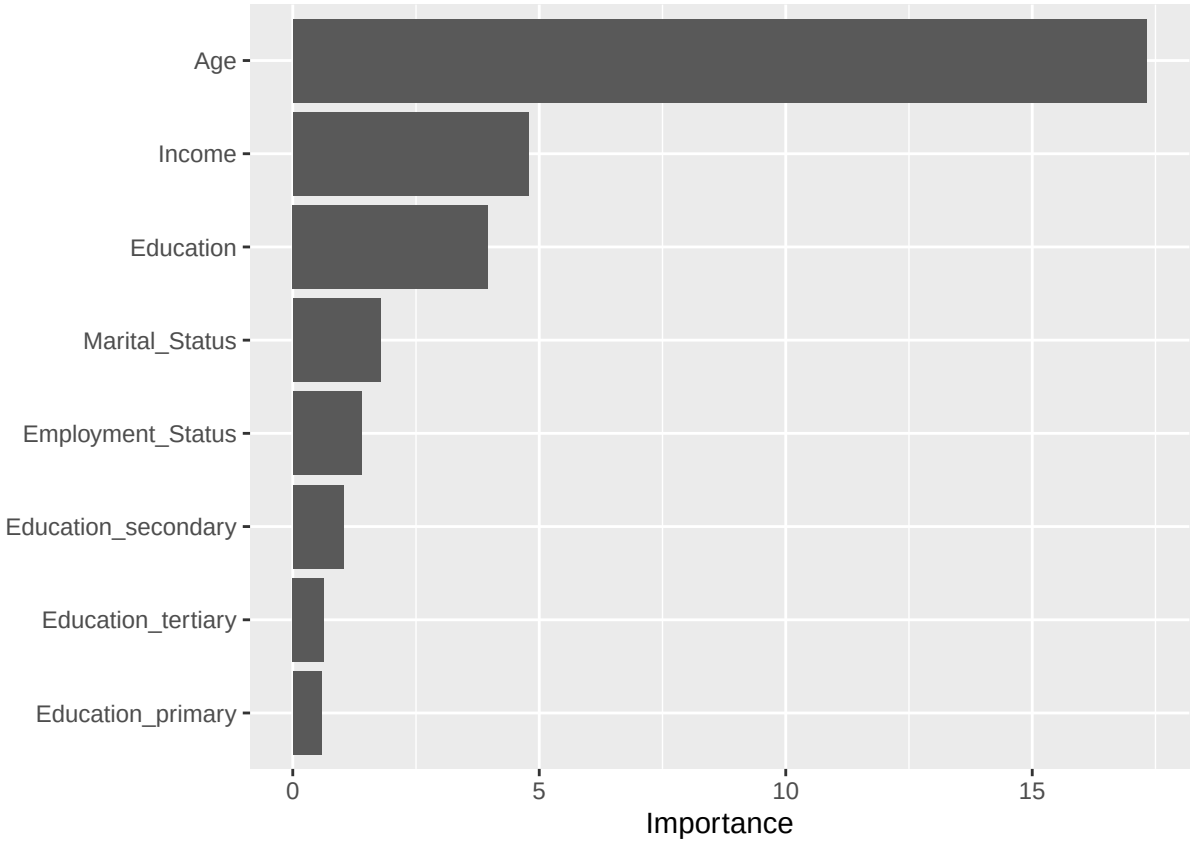
```
library(vip)
```

```
## Warning: package 'vip' was built under R version 4.3.3
```

```
##
## Attaching package: 'vip'
```

```
## The following object is masked from 'package:utils':
##
##     vi
```

```
vip(rf.model)
```



Gradient Boosting Model

```
library(gbm)
param.grid.gbm <- expand.grid(n.trees = c(5, 20, 50, 100, 300),
                              shrinkage = c(0.01, 0.1, 0.3),
                              interaction.depth = c(1, 2, 3, 4, 5),
                              n.minobsinnode = c(5, 10, 15, 20))
```

```
ctrl.gbm <- trainControl(method = "cv",
                         number = 5)
```

```
gbm.preds <- predict(gbm.model, test)
```

```
confusionMatrix(gbm.preds, test$Violence_Status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 74 24
##          1  4  2
##
##                Accuracy : 0.7308
##                  95% CI : (0.6349, 0.8131)
##     No Information Rate : 0.75
##     P-Value [Acc > NIR] : 0.7186083
##
##                   Kappa : 0.0345
##
##  Mcnemar's Test P-Value : 0.0003298
##
##             Sensitivity : 0.94872
##             Specificity : 0.07692
##          Pos Pred Value : 0.75510
##          Neg Pred Value : 0.33333
##              Prevalence : 0.75000
##          Detection Rate : 0.71154
##    Detection Prevalence : 0.94231
##       Balanced Accuracy : 0.51282
##
##        'Positive' Class : 0
##
```

Comparison and Recommendations:

Accuracy:

- Classification Tree (CT): 0.75
- Random Forest (RF): 0.7212
- Gradient Boosting (GB): 0.7308

- The Classification Tree has the highest accuracy at 75%, but accuracy alone does not provide a complete picture.

Sensitivity and Specificity:

- CT: Sensitivity = 1.00, Specificity = 0.00
- RF: Sensitivity = 0.93590, Specificity = 0.07692
- GB: Sensitivity = 0.94872, Specificity = 0.07692
- The Classification Tree has perfect sensitivity but zero specificity, meaning it classifies all positive cases correctly but fails to correctly classify any negative cases. Both Random Forest and Gradient Boosting have high sensitivity but very low specificity.

Kappa:

- CT: 0
- RF: 0.0169
- GB: 0.0345
- The Kappa statistic, which accounts for chance agreement, is highest for Gradient Boosting, though still low across all models.

McNemar's Test:

- CT: 9.443e-07
- RF: 0.0008302
- GB: 0.0003298
- All models show statistically significant results from McNemar's test, indicating differences in the paired proportions of outcomes.

Balanced Accuracy:

- CT: 0.50
- RF: 0.50641
- GB: 0.51282
- Balanced Accuracy, which considers both sensitivity and specificity, is slightly higher for Gradient Boosting.

Conclusion:

- While the Classification Tree has the highest accuracy, its zero specificity makes it unsuitable for applications where correctly identifying negative cases is important. Both Random Forest and Gradient Boosting offer a better balance between sensitivity and specificity, although they still have low specificity.
- Gradient Boosting shows slightly better overall performance with the highest Balanced Accuracy and Kappa, making it the preferred choice among the three. However, given the low specificity across all models, further model tuning or exploration of alternative models might be necessary to achieve better classification performance.