lab_3_4.R

Mohsin Chunawala

2024-11-06

```
# Load necessary libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

# Load the data
# Replace the path with the correct path to your CSV file if needed
data <- read.csv("D:/MCA/Assignments/Rprogramming/lab 3 and 4/Dataset1_Mortality.csv")

# Part 3: Descriptive Statistics

# Part A: Number of variables and observations
num_variables <- ncol(data)
num_observations <- nrow(data)
cat("Variables:", num_variables, "Observations:", num_observations, "\n")
```

```
## Variables: 7 Observations: 200
```

```
# Part B: Extract variable 2 and 3 as vectors
vector2 <- data[[2]]
vector3 <- data[[3]]
cat("Vector for Variable 2:\n", vector2, "\n")
```

```
## Vector for Variable 2:
##  176 167 170 173 170 165 174 171 180 166 176 172 165 164 167 168 171 181 172 178 175 187 170 181 179 170 170 178 175 185
173 179 173 179 176 179 174 172 182 156 168 173 173 177 179 175 168 170 169 177 172 189 178 170 166 170 170 179 168 175 191
182 174 176 172 167 163 171 175 170 170 172 178 170 175 173 171 160 173 176 167 185 177 167 178 180 181 175 166 174 177 170
164 164 170 171 176 168 168 175 162 179 161 170 165 175 172 166 174 170 183 186 173 173 164 169 165 170 171 167 164 174 174
162 176 159 160 167 172 182 166 180 165 170 171 172 168 161 191 163 168 183 186 163 171 171 179 175 172 178 170 180 169 175
176 169 187 167 183 170 169 175 165 175 178 185 186 173 175 170 156 168 170 173 177 165 166 162 173 177 172 188 172 168 156
168 179 161 176 173 160 179 166 159 175 174 172 170 165 178
```

```
cat("Vector for Variable 3:\n", vector3, "\n")
```

```
## Vector for Variable 3:
##  77 56 80 89 71 62 75 68 100 74 63 53 69 82 76 86 86 74 70 84 79 89 72 86 70 78 82 72 81 80 83 75 75 79 77 79 74 71 83 77
53 77 83 90 80 61 86 78 96 81 72 73 70 74 69 77 74 79 74 97 92 101 82 76 93 65 76 84 68 64 79 74 67 69 96 86 69 59 71 66 63
110 97 75 94 69 70 80 72 74 76 65 65 62 80 73 77 75 66 87 70 84 63 58 67 84 75 75 69 81 73 96 77 81 67 80 94 64 64 89 68 70
76 72 99 57 63 72 77 81 65 85 76 68 72 81 58 75 99 64 73 92 92 71 71 73 68 80 68 110 60 88 71 66 65 73 87 68 91 81 82 86 77
67 83 79 90 85 70 65 70 64 73 87 81 76 63 63 73 77 86 83 77 66 61 68 74 61 73 67 74 78 68 82 80 57 91 60 74 81
```

```
# Part C: Count the different blood groups in the dataset
num_blood_groups <- length(unique(data$BLOOD))
cat("Different Blood Groups:", num_blood_groups, "\n")
```

```
## Different Blood Groups: 4
```

```
# Part D: List unique SMOKE categories
unique_smoke <- unique(data$SMOKE)
cat("Unique SMOKE Categories:\n")
```

```
## Unique SMOKE Categories:
```

```
print(unique_smoke)
```

```
## [1] "nonsmo" "sigare" "pipe"
```

```
# Part E: Count individuals with CHOL levels above 300
# Convert CHOL to numeric in case of any non-numeric values
data$CHOL <- as.numeric(as.character(data$CHOL))

# Check for any NA values after conversion
if (any(is.na(data$CHOL))) {
  cat("Warning: Some CHOL values could not be converted to numeric and are set to NA.\n")
}

high_CHOL <- sum(data$CHOL > 300, na.rm = TRUE)
cat("Individuals with CHOL > 300:", high_CHOL, "\n")
```

```
## Individuals with CHOL > 300: 12
```

```
# Part F: Mean HEIGHT for individuals where mortality is 'alive'
mean_HEIGHT_alive <- mean(data$HEIGHT[data$MORT == "alive"], na.rm = TRUE)
cat("Mean HEIGHT for Mortality Alive:", mean_HEIGHT_alive, "\n")
```

```
## Mean HEIGHT for Mortality Alive: 172.517
```

```
# Part G: AGE of the tallest person with O Blood Group
tallest_o <- data %>% filter(BLOOD == "O") %>% arrange(desc(HEIGHT)) %>% slice(1) %>% select(AGE)
cat("AGE of Tallest O Blood Group Person:\n")
```

```
## AGE of Tallest O Blood Group Person:
```

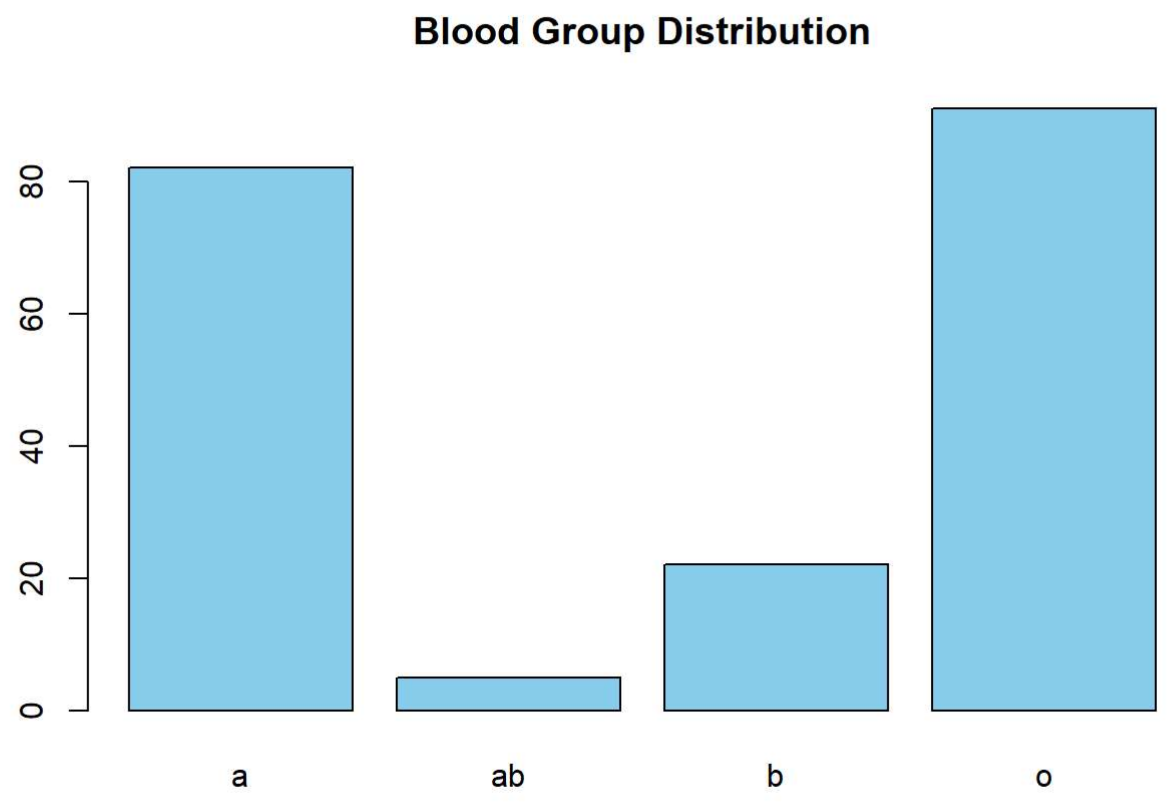
```
print(tallest_o)
```

```
## [1] AGE
## <0 rows> (or 0-length row.names)
```

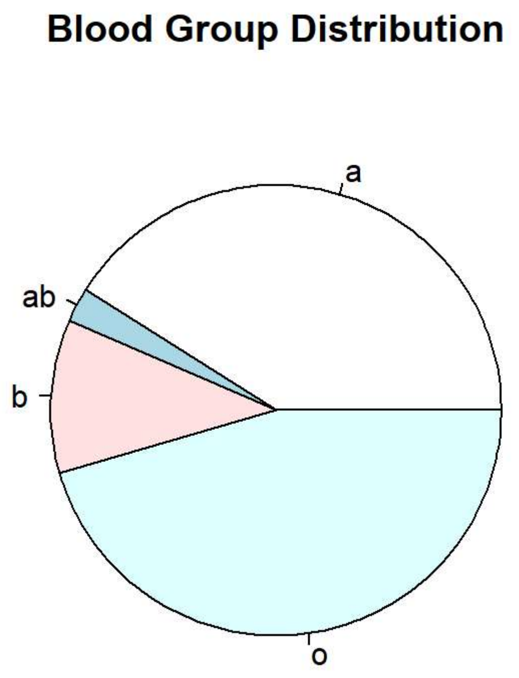
```
# Part H: Count nonsmokers who are alive and below 40 years old
nonsmokers_alive_below_40 <- sum(data$SMOKE == "no" & data$MORT == "alive" & data$AGE < 40, na.rm = TRUE)
cat("Nonsmokers Alive Below 40:", nonsmokers_alive_below_40, "\n")
```

```
## Nonsmokers Alive Below 40: 0
```

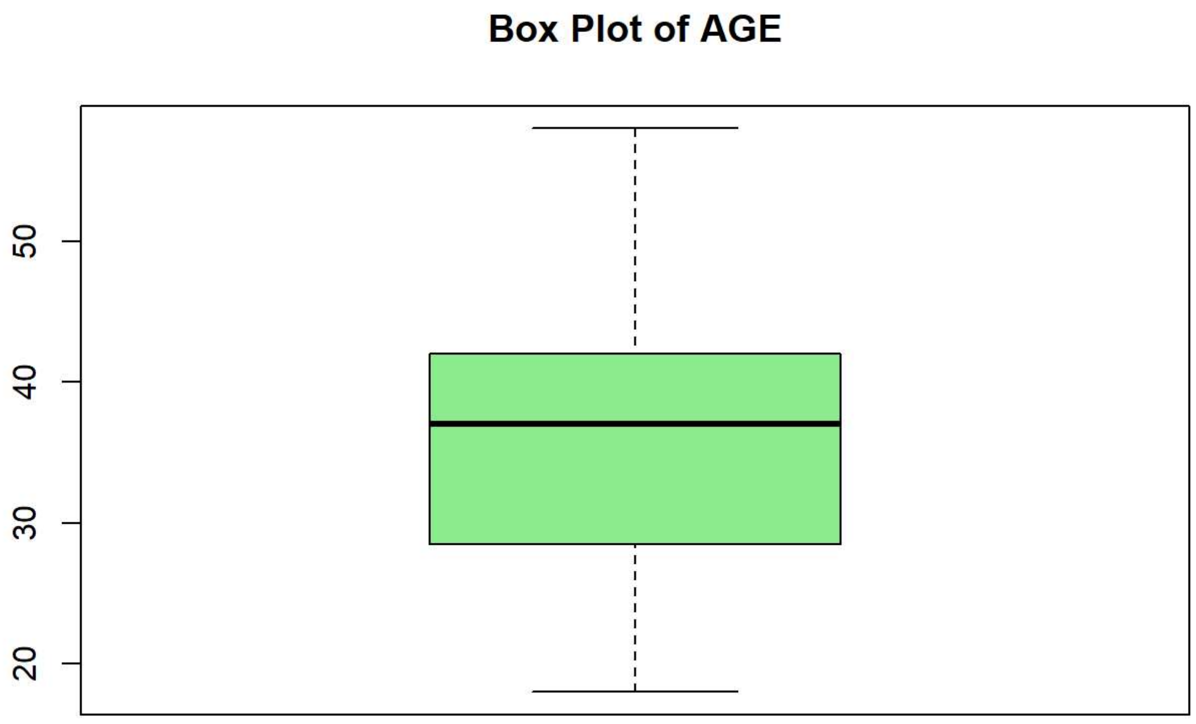
```
# Part 4: Data Visualization

# Part A: Single-variable plots
# Bar plot of BloodGroup
barplot(table(data$BLOOD), main="Blood Group Distribution", col="skyblue")
```



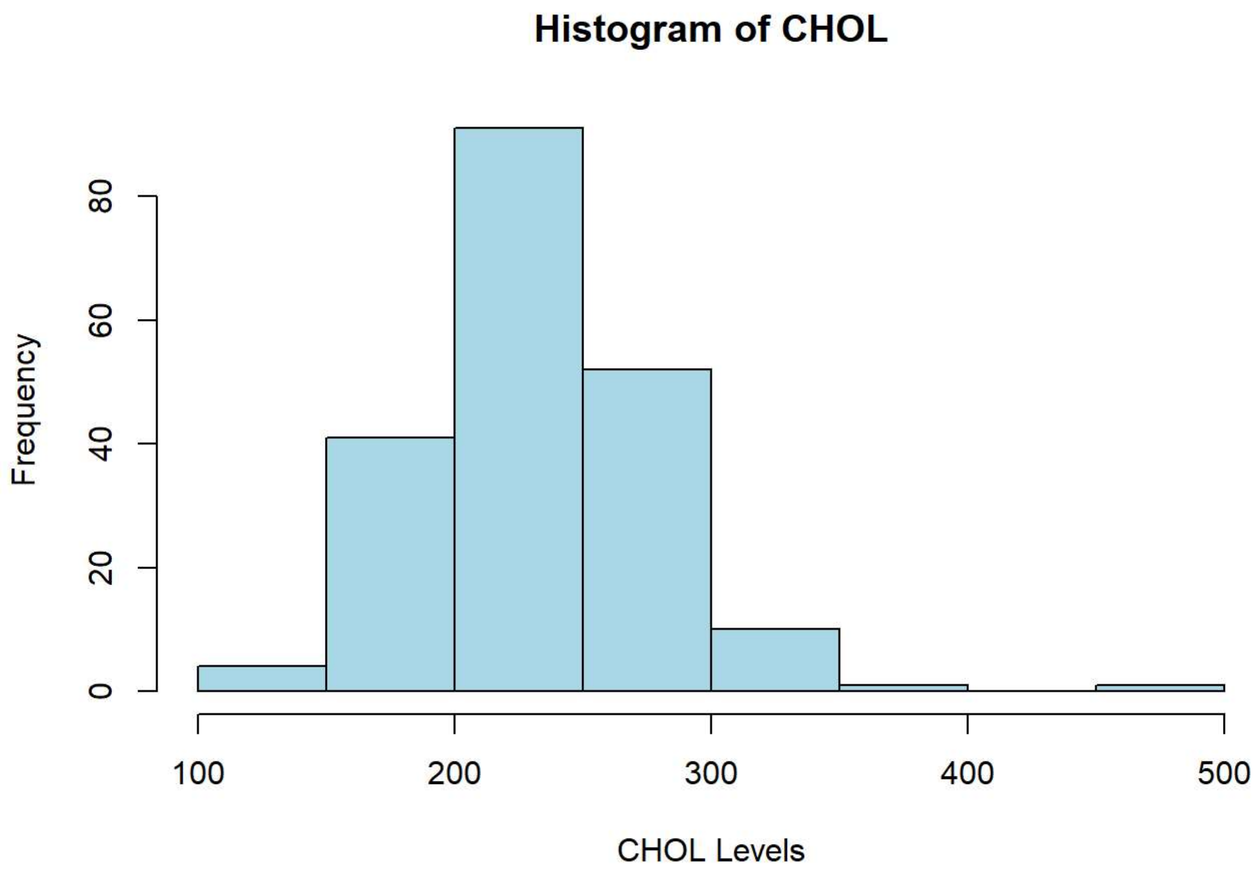
```
# Pie chart of BloodGroup
pie(table(data$BLOOD), main="Blood Group Distribution")
```



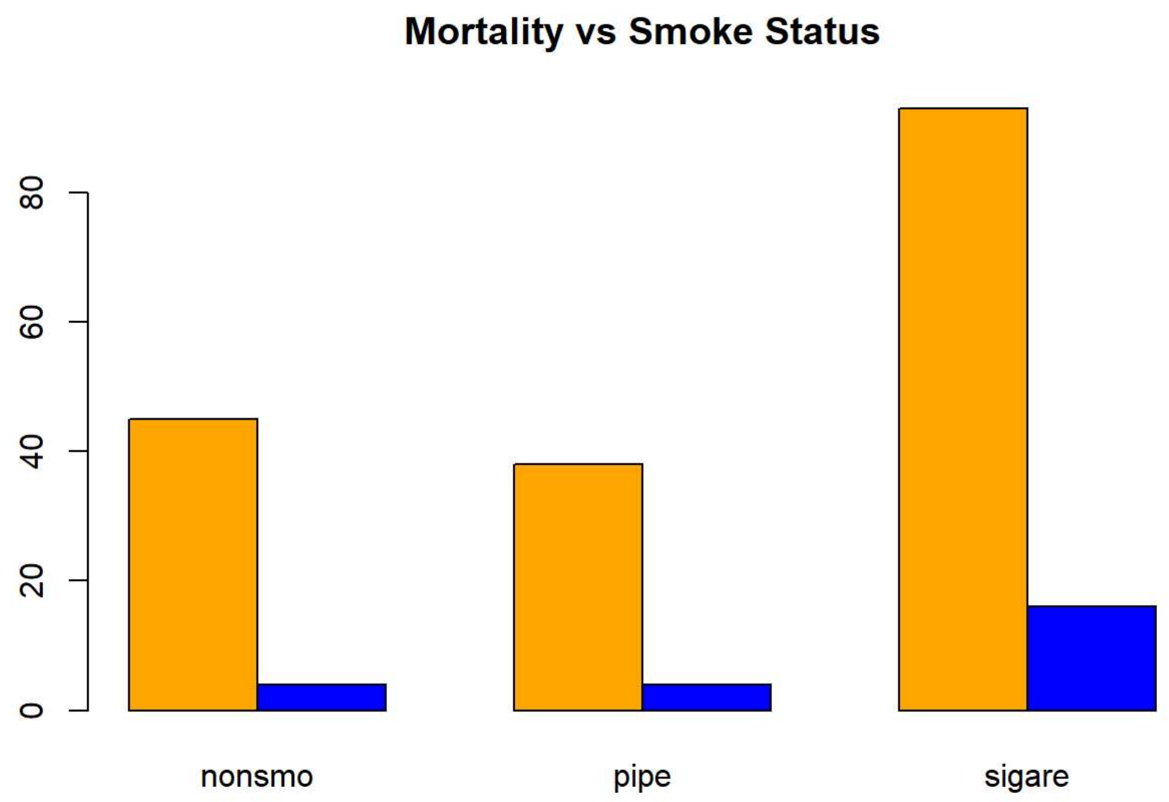
```
# Box plot of AGE
boxplot(data$AGE, main="Box Plot of AGE", col="lightgreen")
```



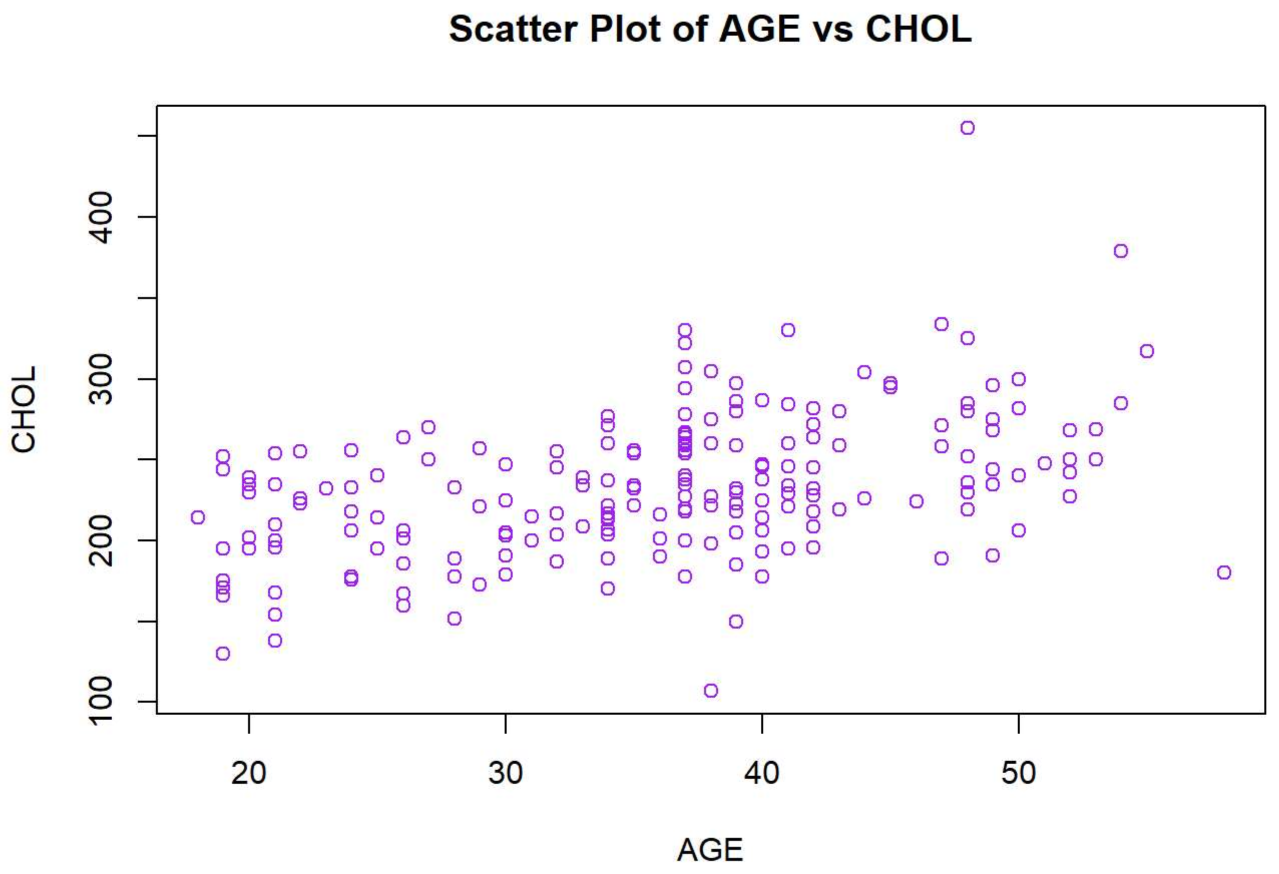
```
# Histogram of CHOL
hist(data$CHOL, main="Histogram of CHOL", col="lightblue", xlab="CHOL Levels")
```



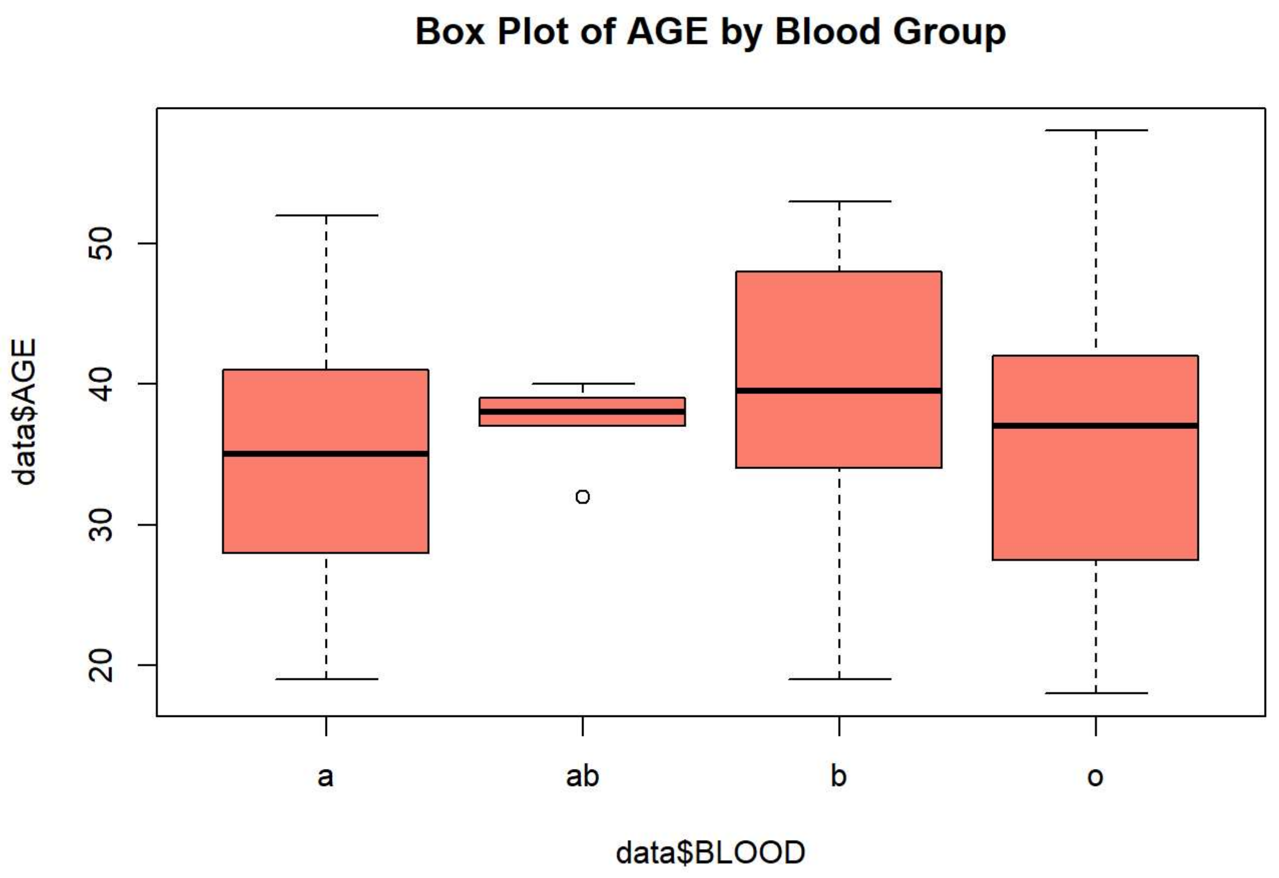
```
# Part B: Two-variable plots
# Bar plot for Mortality and SMOKE
with(data, barplot(table(MORT, SMOKE), beside=TRUE, col=c("orange", "blue"), main="Mortality vs Smoke Status"))
```



```
# Scatter plot for AGE and CHOL
plot(data$AGE, data$CHOL, main="Scatter Plot of AGE vs CHOL", xlab="AGE", ylab="CHOL", col="purple")
```



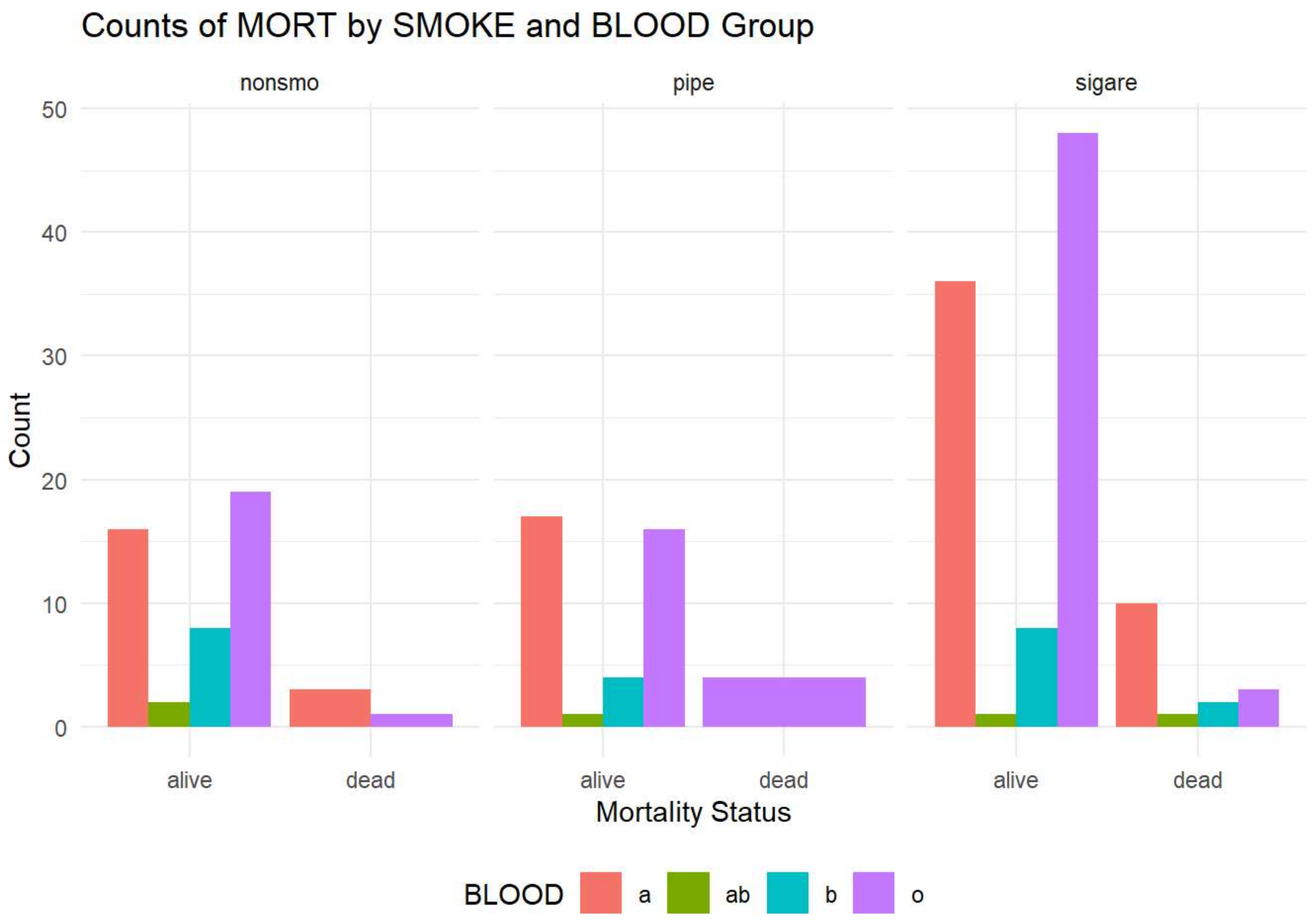
```
# Box plot for AGE by BloodGroup
boxplot(data$AGE ~ data$BLOOD, main="Box Plot of AGE by Blood Group", col="salmon")
```



```
# Part C: Multivariable 2D Bar Plot with ggplot2

# Summarize data for the 2D bar plot
mort_smoke_blood_summary <- data %>%
  group_by(MORT, SMOKE, BLOOD) %>%
  summarise(Count = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'MORT', 'SMOKE'. You can override using the
## `.groups` argument.
```

```
# Create a 2D bar plot
ggplot(mort_smoke_blood_summary, aes(x = MORT, y = Count, fill = BLOOD)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ SMOKE) +
  labs(title = "Counts of MORT by SMOKE and BLOOD Group",
       x = "Mortality Status",
       y = "Count") +
  theme_minimal() +
  theme(legend.position = "bottom")
```


Counts of MORT by SMOKE and BLOOD Group
nonsmo
pipe
sigare
Count
alive
dead
alive
dead
alive
dead
Mortality Status
BLOOD
a
ab
b
o