DataMining - Gym

Diogo Cruz André Eiras

Dataset

```
#Importing the dataset
df <- read.csv("gym.csv", stringsAsFactors = TRUE)
df_numeric <- df %>% select(where(is.numeric))
```

O dataset escolhido para este trabalho chama-se "*Gym Memebers Exercise Dataset*" tendo sido recolhido no *Kaggle*. O dataset tem informação detalhada sobre a rotina de exercícios, atributos físicos e dados demográficos de membros de um ginásio, sendo composto por 15 variáveis cada uma com 973 observações.

```
cat("Número de observações:",nrow(df), "\n")
```

```
Número de observações: 973
```

```
cat("Número de variávei:",ncol(df), "\n")
```

```
Número de variávei: 15
```

```
sapply(df, class)
```

```
                    Age                  Gender 
              "integer"                "factor" 
           Weight..kg.              Height..m. 
              "numeric"                "numeric" 
             Max_BPM                  Avg_BPM 
              "integer"                "integer" 
           Resting_BPM Session_Duration..hours. 
```

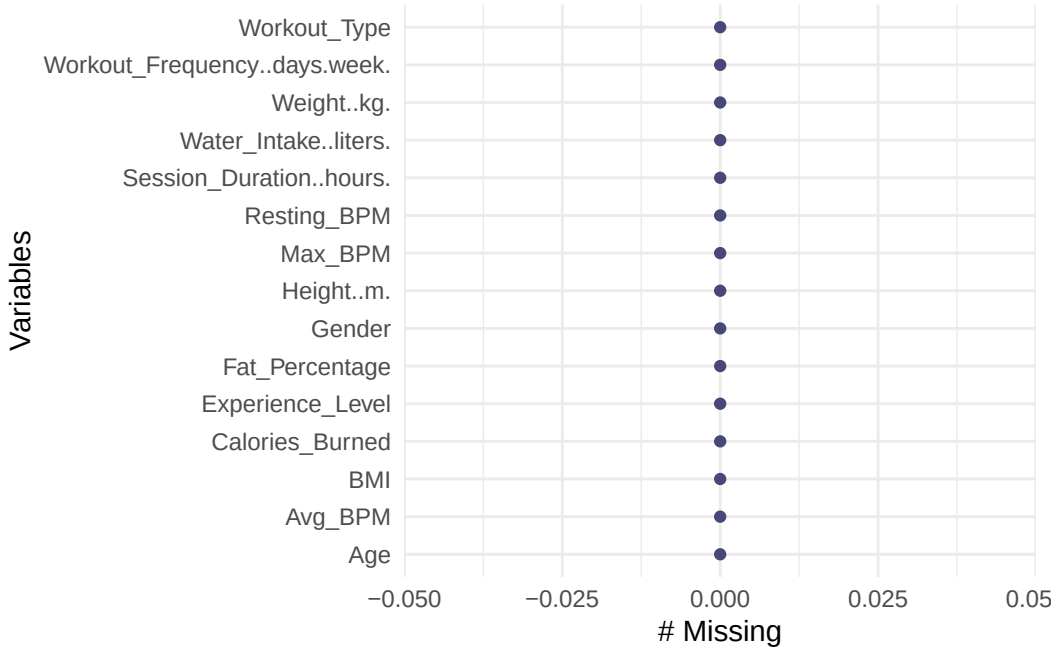
```
                    "integer"                     "numeric"
               Calories_Burned                  Workout_Type
                    "numeric"                      "factor"
                Fat_Percentage          Water_Intake..liters.
                    "numeric"                     "numeric"
Workout_Frequency..days.week.               Experience_Level
                    "integer"                     "integer"
                          BMI
                    "numeric"
```

A lista seguinte indica o significado e tipo de cada uma das variáveis presentes.

- `Age` - Varíavel Quantitativa Contínua que indica a idade do membro.
- `Gender` - Varíavel Categórica que indica o género do membro.
- `Weigh` - Variável Quantitativa Contínua com a massa do membro em kg.
- `Height` - Variável Quantitativa Contínua com a altura do membro em metros.
- `Max BPM` - Varíavel Quantitativa Contínua que indica a frequência cardíaca máxima atingida pelo membro durante o seu treino em batimentos por minuto.
- `Avg BPM` - Varíavel Quantitativa Contínua que indica a frequência cardíaca média do membro durante o seu treino em batimentos por minuto.
- `Resting BPM` - Varíavel Quantitativa Contínua que indica a frequência cardíaca do membro em descanso antes do treino.
- `Session Duration` - Varíavel Quantitativa Contínua que indica a duração do treino em horas.
- `Calories Burned` - Variável Quantitativa Contínua que indica a quantidade de calorias gastas durante o treino em kCal.
- `Workout Type` - Varíavel Categórica que inidica o tipo de treino realizado.
- `Fat Percentage` - Variável Quantitativa Contínua que indica a percentagem de massa gorda do membro.
- `Water Intake` - Variável Quantitativa Contínua que indica o número de litros de água ingeridos diáriamente pelo membro.
- `Workout Frequency` - Variável Quantitativa Discreta que indica o número de dias por semana em que o membro treinou.
- `Experience Level` - Variável Quantitativa Discreta que indica o nível de experiência (1 a 3) do membro.
- `BMI` - Variável Quantitativa Contínua que indica o Índice de Massa Corporal do membro.

Como é possível ver no gráfico seguinte, o dataset não apresenta dados omissos.

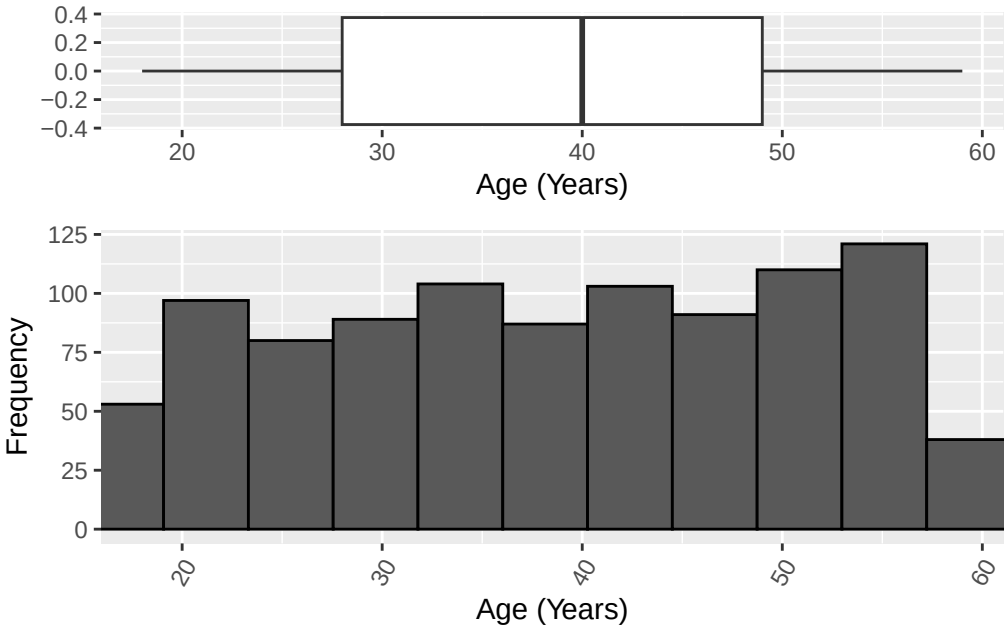
```
gg_miss_var(df)
```



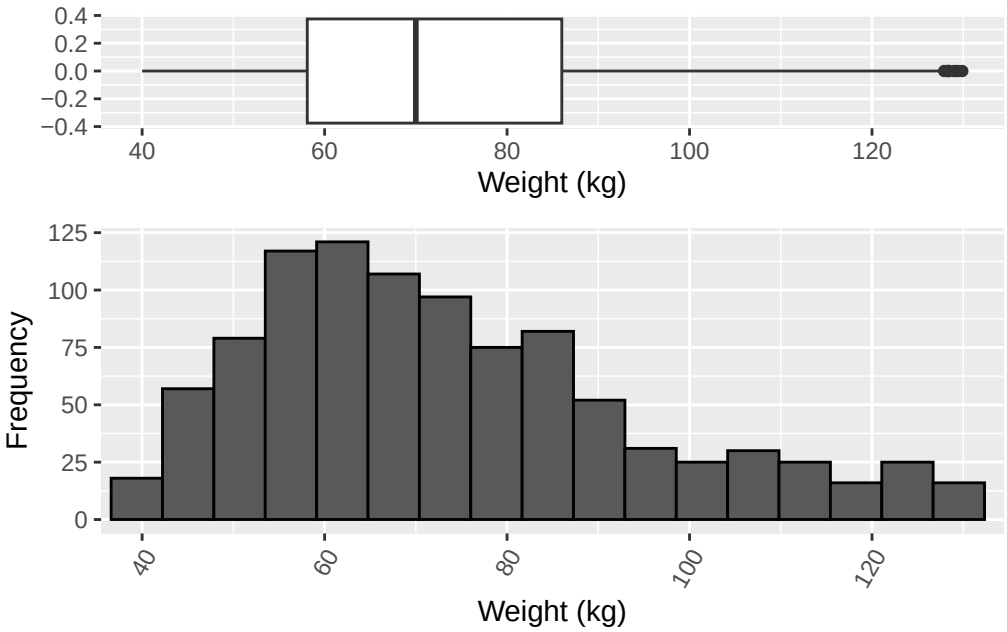
Análise Exploratória do Dataset

De modo a analisar a distribuição das nossas variáveis fizemos boxplot com histograma para cada variavel numérica. Podemos ver que a idade e as três variáveis referentes à frequência cardíaca dos mesmbros apresentam uma distribuição práticamente uniforme. Já as restantes variáveis apresentam distribuições com uma forma aproximadamente normal mas com assimetrias.

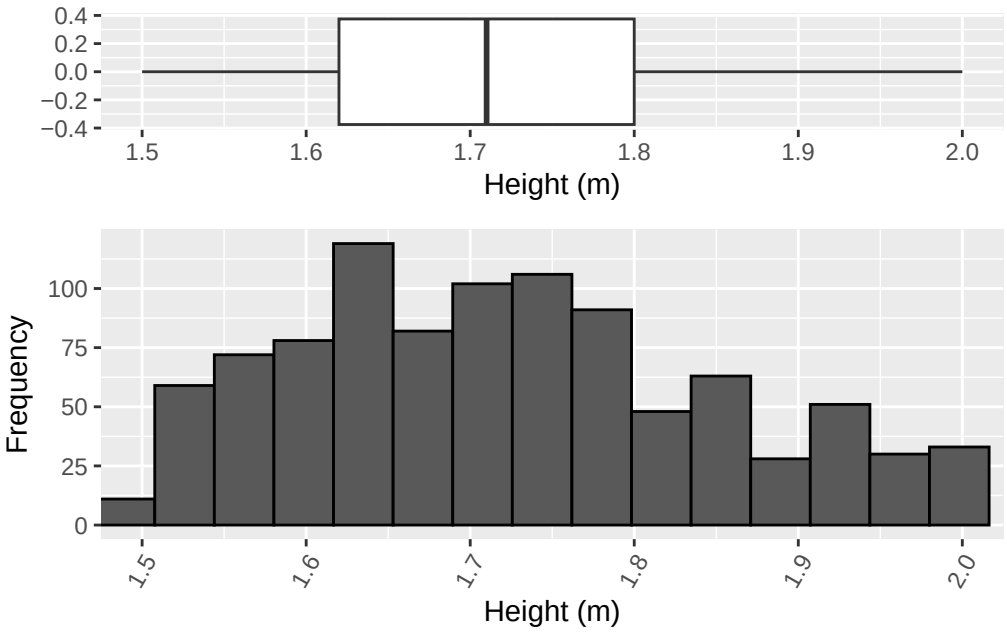
```
hist_and_box(df, df$Age, "Age (Years)")
```



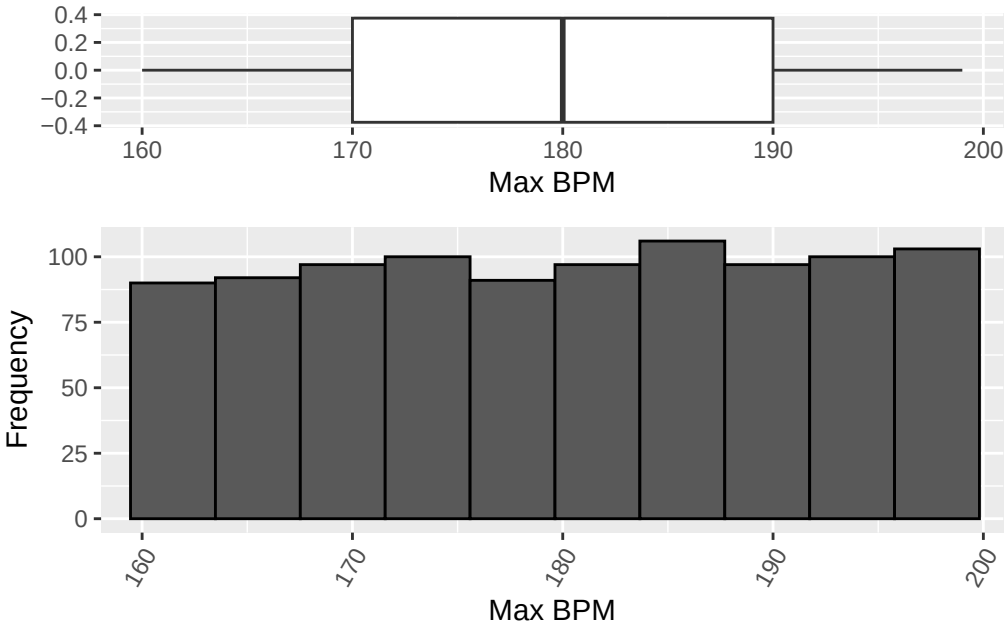

```
hist_and_box(df, df$Weight..kg., "Weight (kg)")
```



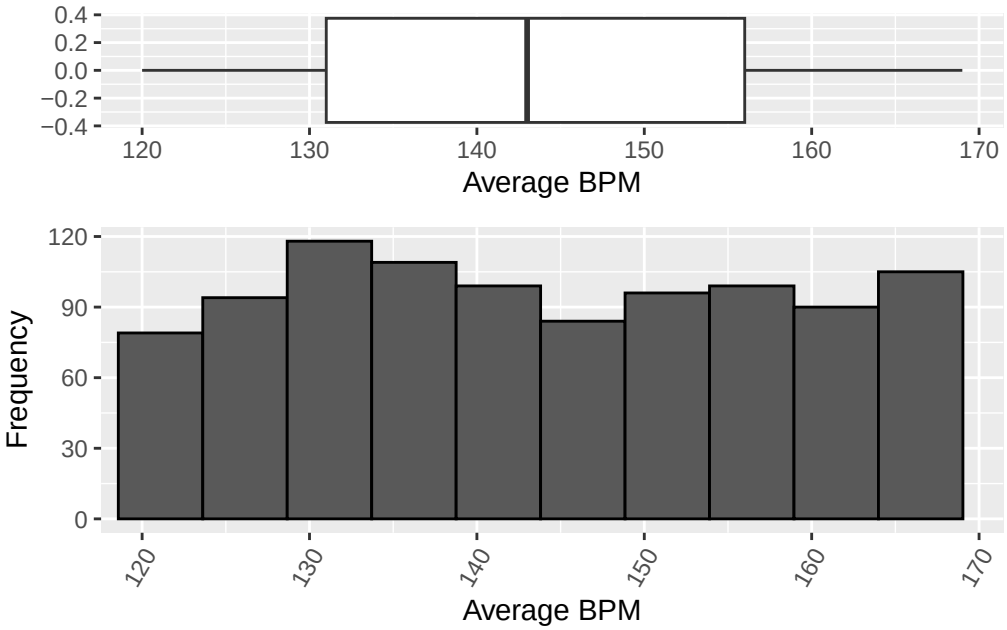
```
hist_and_box(df, df$Height..m., "Height (m)")
```



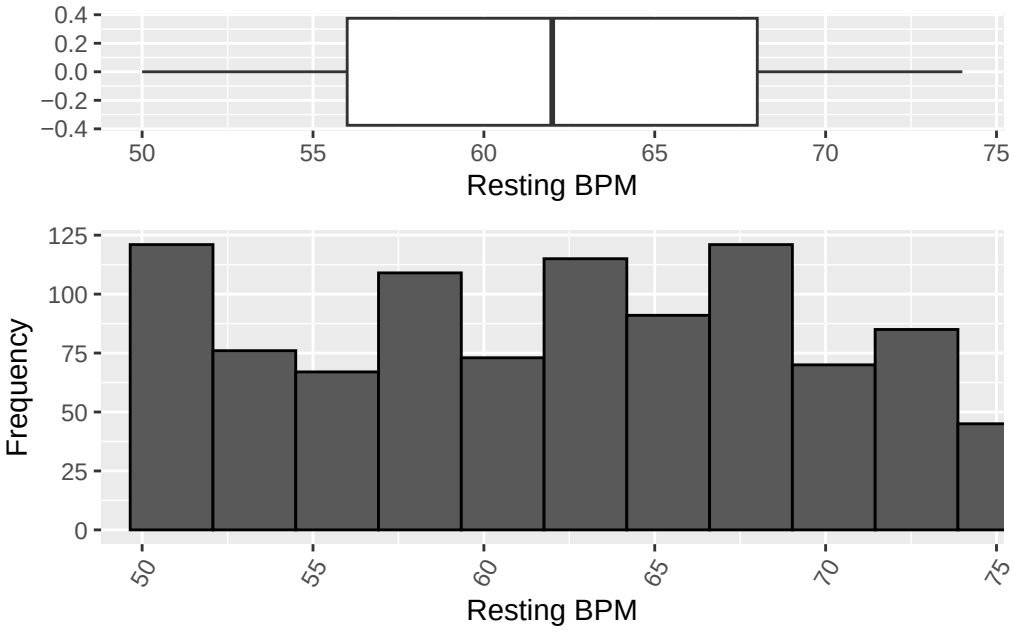

```
hist_and_box(df, df$Max_BPM, "Max BPM")
```



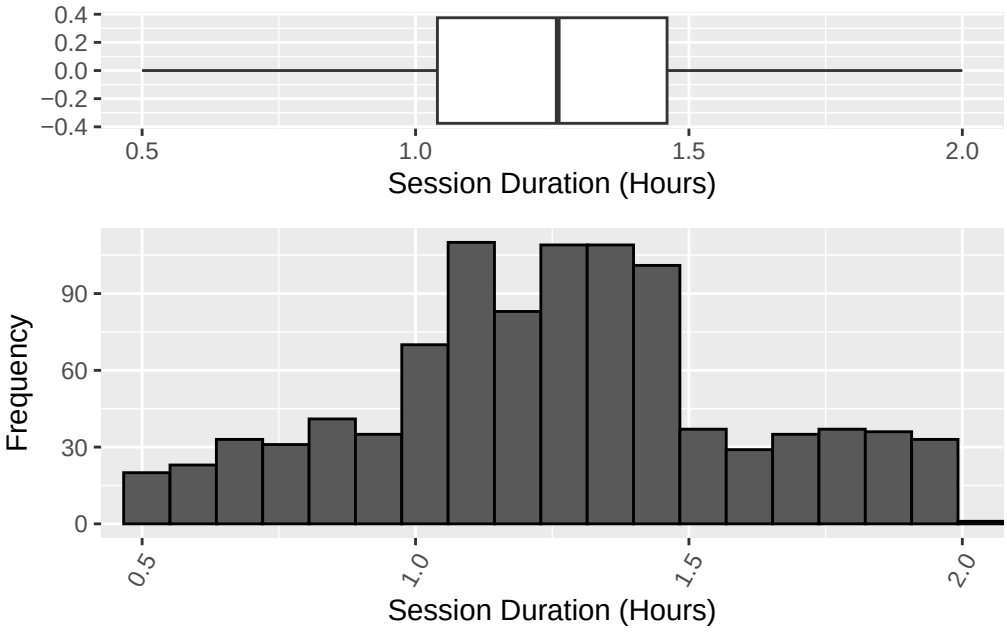
```
hist_and_box(df, df$Avg_BPM, "Average BPM")
```



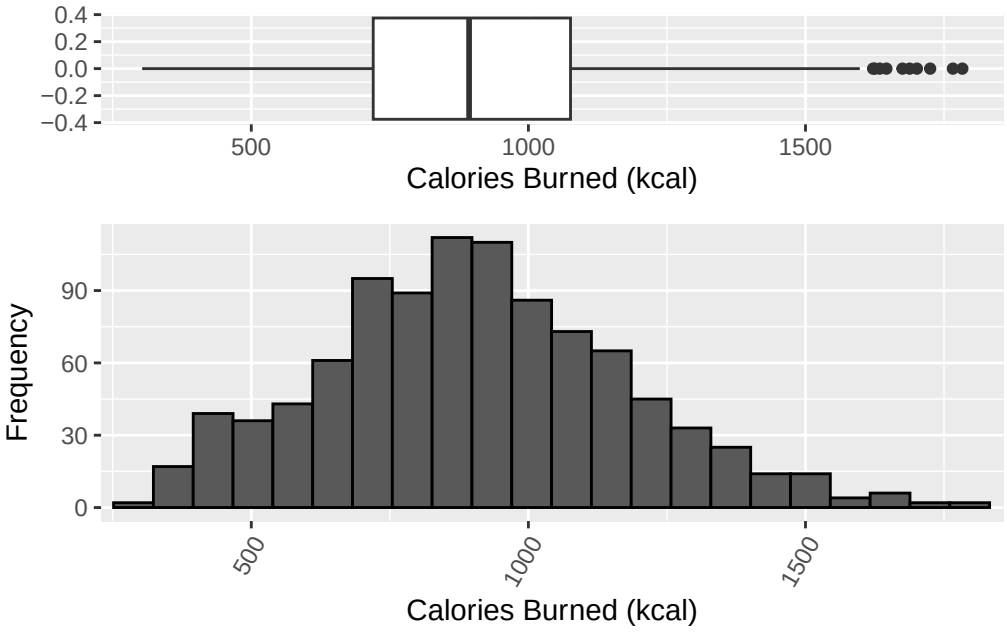

```
hist_and_box(df, df$Resting_BPM, "Resting BPM")
```



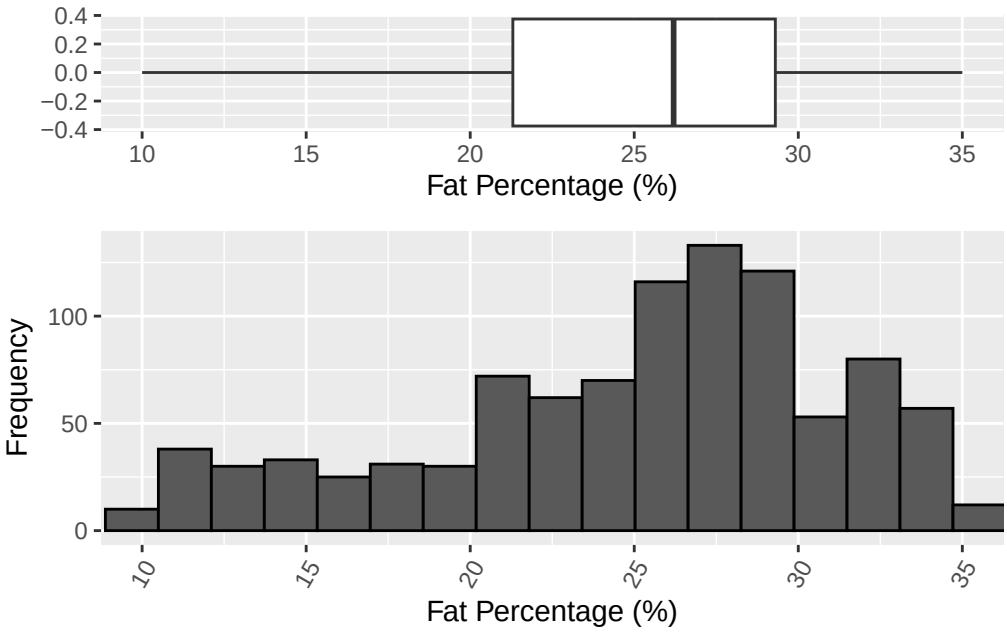
```r
hist_and_box(df, df$Session_Duration..hours., "Session Duration (Hours)")
```



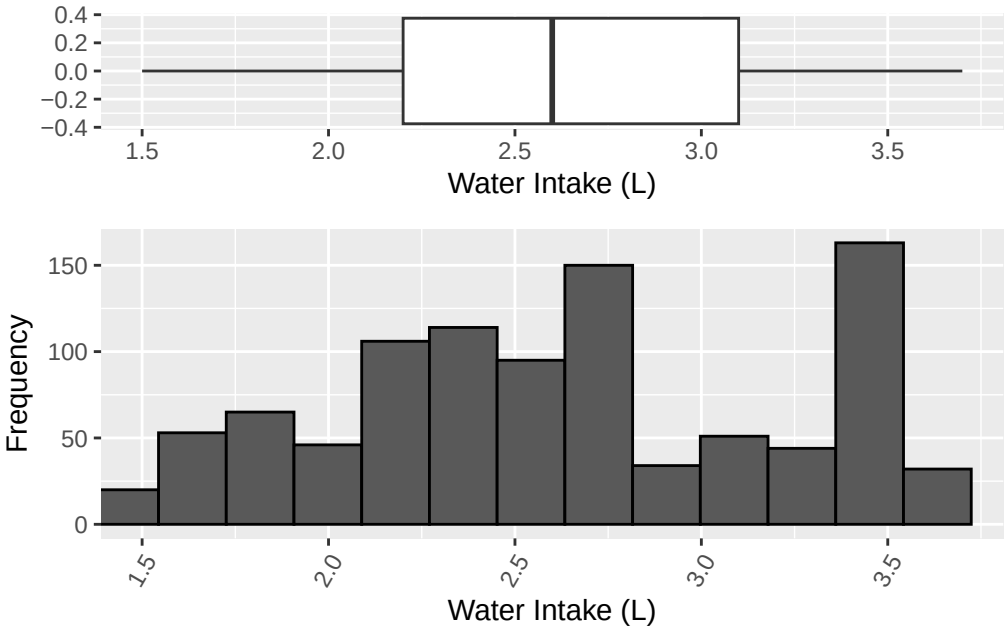

```r
hist_and_box(df, df$Calories_Burned, "Calories Burned (kcal)")
```



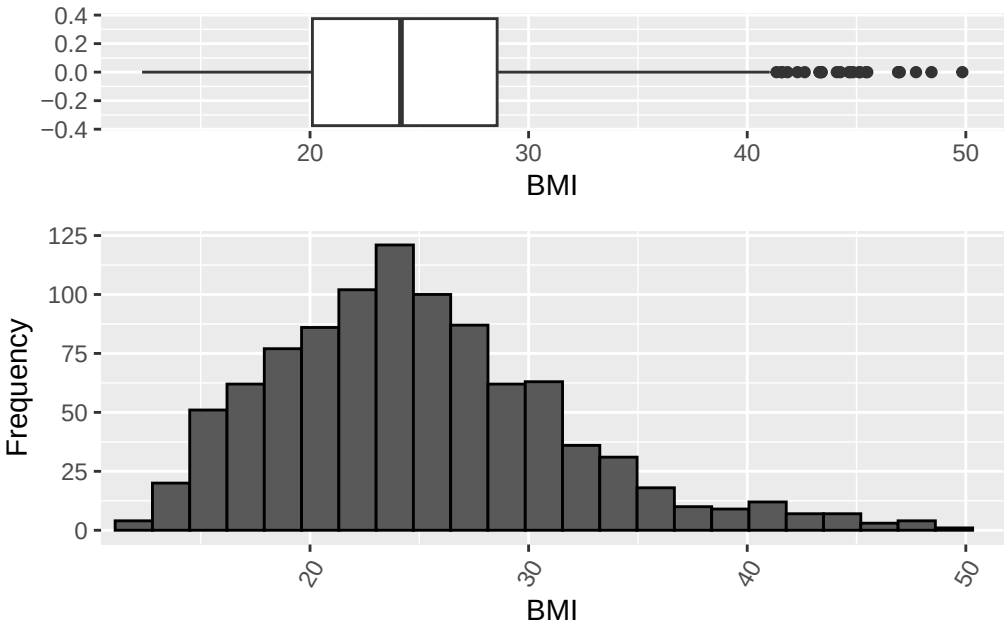
```
hist_and_box(df, df$Fat_Percentage, "Fat Percentage (%)")
```




```
hist_and_box(df, df$Water_Intake..liters., "Water Intake (L)")
```



```r
hist_and_box(df, df$BMI, "BMI")
```



A tabela seguinte mostra as estatísticas descritivas das variáveis numéricas do dataset.

```r
summary_stats <- data.frame(
  Variable = c("Age (Years)", "Weight (kg)", "Height (m)", "Max BPM", "Average BPM",
               "Resting BPM", "Session Duration (Hours)", "Calories Burned (kcal)",
               "Fat Percentage (%)", "Water Intake (l)", "BMI"),
  Mean = round(c(mean(df$Age, na.rm = TRUE), mean(df$Weight..kg., na.rm = TRUE),
                 mean(df$Height..m., na.rm = TRUE), mean(df$Max_BPM, na.rm = TRUE),
                 mean(df$Avg_BPM, na.rm = TRUE), mean(df$Resting_BPM, na.rm = TRUE),
                 mean(df$Session_Duration..hours., na.rm = TRUE), mean(df$Calories_Burned, na
                 mean(df$Fat_Percentage, na.rm = TRUE), mean(df$Water_Intake..liters., na.rm
                 mean(df$BMI, na.rm = TRUE)), 2),
  SD = round(c(sd(df$Age, na.rm = TRUE), sd(df$Weight..kg., na.rm = TRUE),
               sd(df$Height..m., na.rm = TRUE), sd(df$Max_BPM, na.rm = TRUE),
               sd(df$Avg_BPM, na.rm = TRUE), sd(df$Resting_BPM, na.rm = TRUE),
               sd(df$Session_Duration..hours., na.rm = TRUE), sd(df$Calories_Burned, na.rm =
               sd(df$Fat_Percentage, na.rm = TRUE), sd(df$Water_Intake..liters., na.rm = TRU
               sd(df$BMI, na.rm = TRUE)), 2),
  Median = round(c(median(df$Age, na.rm = TRUE), median(df$Weight..kg., na.rm = TRUE),
                   median(df$Height..m., na.rm = TRUE), median(df$Max_BPM, na.rm = TRUE),
                   median(df$Avg_BPM, na.rm = TRUE), median(df$Resting_BPM, na.rm = TRUE),
```

```
                  median(df$Session_Duration..hours., na.rm = TRUE), median(df$Calories_Burn
                  median(df$Fat_Percentage, na.rm = TRUE), median(df$Water_Intake..liters.,
                  median(df$BMI, na.rm = TRUE)), 2),
  IQR = round(c(IQR(df$Age, na.rm = TRUE), IQR(df$Weight..kg., na.rm = TRUE),
                IQR(df$Height..m., na.rm = TRUE), IQR(df$Max_BPM, na.rm = TRUE),
                IQR(df$Avg_BPM, na.rm = TRUE), IQR(df$Resting_BPM, na.rm = TRUE),
                IQR(df$Session_Duration..hours., na.rm = TRUE), IQR(df$Calories_Burned, na.rm
                IQR(df$Fat_Percentage, na.rm = TRUE), IQR(df$Water_Intake..liters., na.rm = T
                IQR(df$BMI, na.rm = TRUE)), 2),
  Min = round(c(min(df$Age, na.rm = TRUE), min(df$Weight..kg., na.rm = TRUE),
                min(df$Height..m., na.rm = TRUE), min(df$Max_BPM, na.rm = TRUE),
                min(df$Avg_BPM, na.rm = TRUE), min(df$Resting_BPM, na.rm = TRUE),
                min(df$Session_Duration..hours., na.rm = TRUE), min(df$Calories_Burned, na.rm
                min(df$Fat_Percentage, na.rm = TRUE), min(df$Water_Intake..liters., na.rm = T
                min(df$BMI, na.rm = TRUE)), 2),
  Max = round(c(max(df$Age, na.rm = TRUE), max(df$Weight..kg., na.rm = TRUE),
                max(df$Height..m., na.rm = TRUE), max(df$Max_BPM, na.rm = TRUE),
                max(df$Avg_BPM, na.rm = TRUE), max(df$Resting_BPM, na.rm = TRUE),
                max(df$Session_Duration..hours., na.rm = TRUE), max(df$Calories_Burned, na.rm
                max(df$Fat_Percentage, na.rm = TRUE), max(df$Water_Intake..liters., na.rm = T
                max(df$BMI, na.rm = TRUE)), 2)
)

# Exibir a tabela com kable (na ordem: Mean, SD, Median, IQR, Min, Max)
kable(summary_stats, format = "pipe", digits = 2, caption = "Estatísticas Descritivas das Var
```

Table 1: Estatísticas Descritivas das Variáveis

Variable	Mean	SD	Median	IQR	Min	Max
Age (Years)	38.68	12.18	40.00	21.00	18.00	59.00
Weight (kg)	73.85	21.21	70.00	27.90	40.00	129.90
Height (m)	1.72	0.13	1.71	0.18	1.50	2.00
Max BPM	179.88	11.53	180.00	20.00	160.00	199.00
Average BPM	143.77	14.35	143.00	25.00	120.00	169.00
Resting BPM	62.22	7.33	62.00	12.00	50.00	74.00
Session Duration (Hours)	1.26	0.34	1.26	0.42	0.50	2.00
Calories Burned (kcal)	905.42	272.64	893.00	356.00	303.00	1783.00
Fat Percentage (%)	24.98	6.26	26.20	8.00	10.00	35.00
Water Intake (l)	2.63	0.60	2.60	0.90	1.50	3.70
BMI	24.91	6.66	24.16	8.45	12.32	49.84