Heart Data Set Analysis

```r
# Exploratory Analysis of the Data Set:

library(ggplot2)
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
```

```r
library(tidyr)

setwd("/Users/faizshaikh/Downloads/archive")
```

```
Warning: The working directory was changed to /Users/faizshaikh/Downloads/archive inside a notebook chunk. The wo
rking directory will be reset when the chunk is finished running. Use the knitr root.dir option in the setup chun
k to change the working directory for notebook chunks.
```

```r
heart_data <- read.csv("heart.csv")

# Viewing the structure of the data set:
str(heart_data)
```

```
'data.frame':	303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
 $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
 $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
 $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
 $ exng    : int  0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slp     : int  0 0 2 2 2 1 1 2 2 2 ...
 $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ thall   : int  1 2 2 2 2 1 2 3 3 2 ...
 $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Summary statistics of the numeric variables:
summary(heart_data[c("age", "trtbps", "chol", "thalachh")])
```

```
      age            trtbps           chol          thalachh
 Min.   :29.00   Min.   : 94.0   Min.   :126.0   Min.   : 71.0
 1st Qu.:47.50   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:133.5
 Median :55.00   Median :130.0   Median :240.0   Median :153.0
 Mean   :54.37   Mean   :131.6   Mean   :246.3   Mean   :149.6
 3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:166.0
 Max.   :77.00   Max.   :200.0   Max.   :564.0   Max.   :202.0
```

```r
# Creating frequency table for the categorical variables:
table(heart_data$Sex)
```

```
< table of extent 0 >
```

```r
table(heart_data$exang)
```

```
< table of extent 0 >
```

```r
table(heart_data$ca)
```

```

  0   1   2   3   4 
175  65  38  20   5 
```

```r
table(heart_data$cp)
```

```

  0   1   2   3 
143  50  87  23 
```

```r
table(heart_data$fbs)
```

```

  0   1 
258  45 
```

```r
table(heart_data$rest_ecg)
```

```
< table of extent 0 >
```

```r
table(heart_data$target)
```

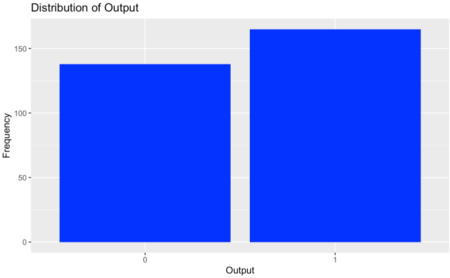
```
< table of extent 0 >
```

```r
# Plotting distributions of the numeric variables:
numeric_vars <- c("age", "trtbps", "chol", "thalachh")
numeric_plots <- lapply(numeric_vars, function(var) {
  ggplot(heart_data, aes_string(x = var)) +
    geom_histogram(binwidth = 5, fill = "blue", color = "black") +
    labs(title = paste("Distribution of", var), x = var, y = "Frequency")
})
```

```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
Please use tidy evaluation idioms with `aes()`.
See also `vignette("ggplot2-in-packages")` for more information.
```

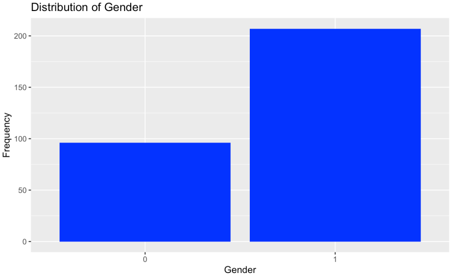
```r
output_print <- ggplot(heart_data, aes(x = as.factor(output))) +
             geom_bar(fill = "blue") +
             scale_y_continuous(labels = scales::comma) +
             labs(title = "Distribution of Output", x = "Output", y = "Frequency")
```

```
print(output_print)
```



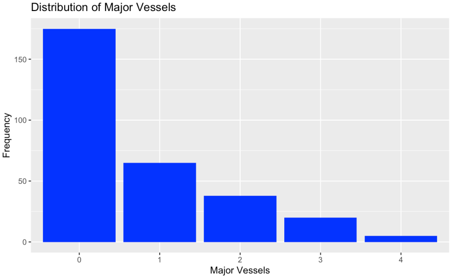
Hide

```
gender <- ggplot(heart_data, aes(x = as.factor(sex))) +
             geom_bar(fill = "blue") +
             scale_y_continuous(labels = scales::comma) +
             labs(title = "Distribution of Gender", x = "Gender", y = "Frequency")

print(gender)
```



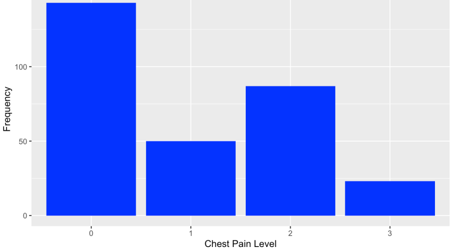
Hide

```
caa_print <- ggplot(heart_data, aes(x = as.factor(caa))) +
             geom_bar(fill = "blue") +
             scale_y_continuous(labels = scales::comma) +
             labs(title = "Distribution of Major Vessels", x = "Major Vessels", y = "Frequency")

print(caa_print)
```



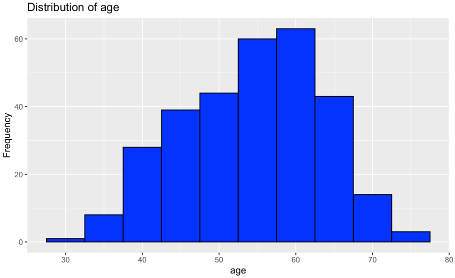
Hide

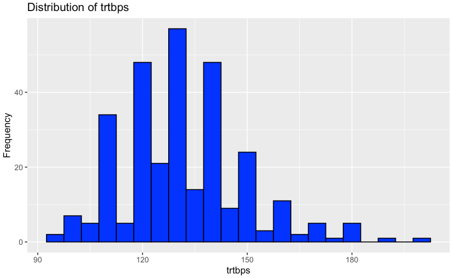
```
cp_print <- ggplot(heart_data, aes(x = as.factor(cp))) +
             geom_bar(fill = "blue") +
             scale_y_continuous(labels = scales::comma) +
             labs(title = "Distribution of Chest Pain", x = "Chest Pain Level", y = "Frequency")

print(cp_print)
```

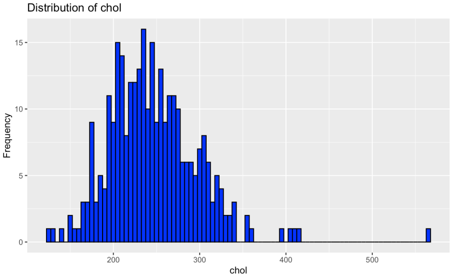


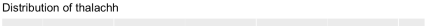
Hide

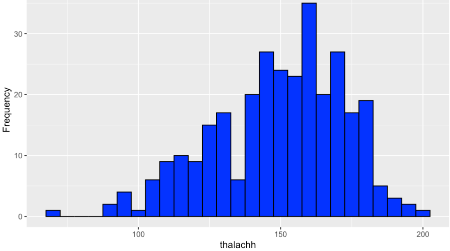
```
# Printing the plots
for (plot in numeric_plots) {
  print(plot)
}
```


Distribution of age


Distribution of trtbps


Distribution of chol


Distribution of thalachh



Hide

```
library(rpart)
library(caret)
library(e1071)

setwd("/Users/faizshaikh/Downloads/archive")
```

```
Warning: The working directory was changed to /Users/faizshaikh/Downloads/archive inside a notebook chunk. The wo
rking directory will be reset when the chunk is finished running. Use the knitr root.dir option in the setup chun
k to change the working directory for notebook chunks.
```

Hide

```
heart_data <- read.csv("heart.csv")

set.seed(100)

# Factorizing the categorical variables.
heart_data$sex <- as.factor(heart_data$sex)
heart_data$cp <- as.factor(heart_data$cp)
heart_data$fbs <- as.factor(heart_data$fbs)
heart_data$restecg <- as.factor(heart_data$restecg)
heart_data$exng <- as.factor(heart_data$exng)
heart_data$slp <- as.factor(heart_data$slp)
heart_data$caa <- as.factor(heart_data$caa)
heart_data$thall <- as.factor(heart_data$thall)

# Building the decision tree model using RPart()
model <- rpart(output ~ cp + fbs + restecg + exng + slp + caa + thall, data = heart_data, method = "class", cp =
0.000015, minsplit = 10, minbucket = 2)
predictions <- predict(model, newdata = heart_data, type = "class")

# Creating the confusion matrix for RPart:
print("Confusion Matrix for R-Part:")
```

```
[1] "Confusion Matrix for R-Part:"
```

Hide

```
cm <- table(heart_data$output, predictions)
print(cm)
```

```
   predictions
      0   1
  0 117  21
  1  14 151
```

Hide

```
# ==========================================================================================

# Building the Naïve-Bayes model:
nb_model <- naiveBayes(output ~ cp + fbs + restecg + exng + slp + caa + thall, data = heart_data)
nb_predictions <- predict(nb_model, newdata = heart_data)

# Creating the confusion matrix for Naïve Bayes:
nb_cm <- table(heart_data$output, nb_predictions)
print("Confusion Matrix for Naïve Bayes:")
```

```
[1] "Confusion Matrix for Naïve Bayes:"
```

Hide

```
print(nb_cm)
```

```
   nb_predictions
      0   1
  0 114  24
  1  16 149
```

Hide

```
# ==========================================================================================

# Calculating accuracy for Naïve-Bayes (based upon confusion matrix results):
nb_accuracy <- sum(diag(nb_cm)) / sum(nb_cm)
print(paste("Accuracy for Naïve Bayes:", nb_accuracy))
```

```
[1] "Accuracy for Naïve Bayes: 0.867986798679868"
```

Hide

```
# Calculating accuracy for RPart (based upon confusion matrix results):
rpart_accuracy <- sum(diag(cm)) / sum(cm)
print(paste("Accuracy for rpart:", rpart_accuracy))
```

```
[1] "Accuracy for rpart: 0.884488448844885"
```