Heart Disease Analysis

1. Importing the Data

```
# install.packages("ggplot2")
# install.packages("dplyr")
library(ggplot2)
```

```
## Warning: package ’ggplot2’ was built under R version 4.1.3
```

```
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.1.3
```

```
##
## 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
```

```
hd <- read.csv("heart_2020_cleaned.csv")
# Usually, I check for duplicates. However there is no unique id variable in
# place, checking for duplicates will only result in false positives.

head(hd$MentalHealth, 40)
```

```
##  [1] 30  0 30  0  0  0  0  0  0  0  0  0  0  0 30  0  2 30  0  0  0  5 15 30  0
## [26] 30  8  0  0  0  0  4  0  5  0  0  0  3  0  2
```

```
n <- c(18:24)
hd2 <- hd[-which(hd$SleepTime %in% n), ]
summary(hd2)
```

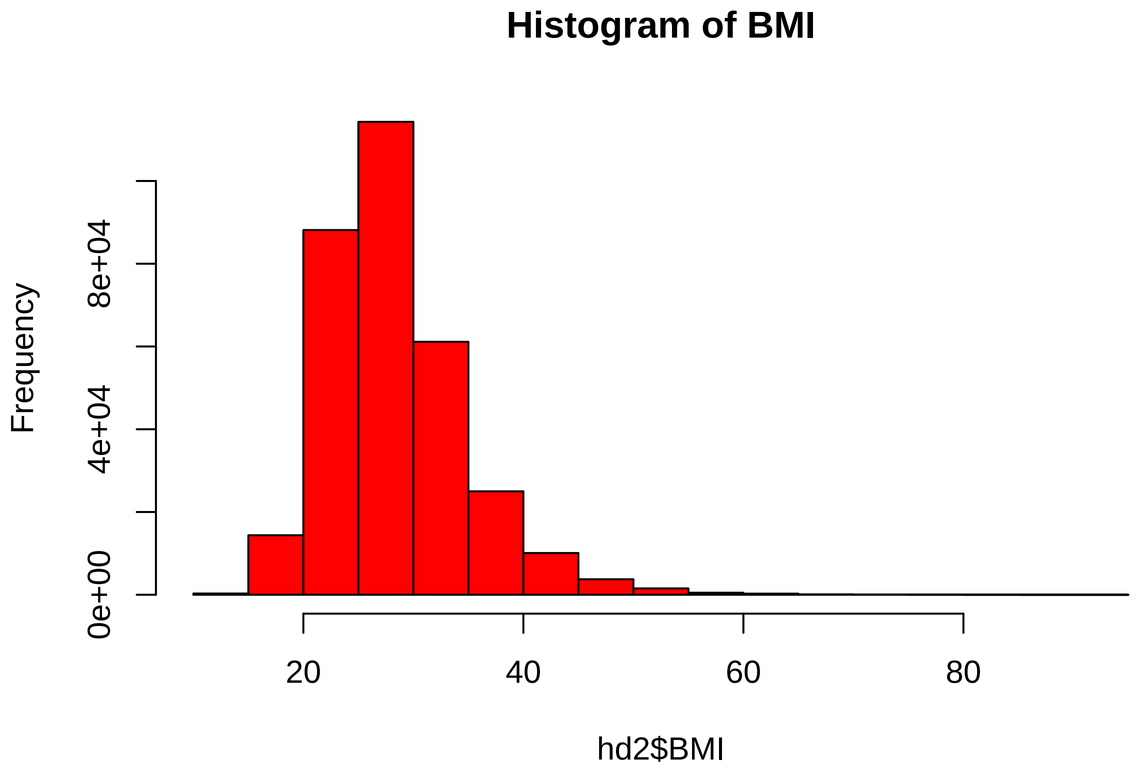
```
##  HeartDisease            BMI          Smoking          AlcoholDrinking
##  Length:319582      Min.   :12.02   Length:319582      Length:319582
##  Class :character   1st Qu.:24.03   Class :character   Class :character
##  Mode  :character   Median :27.32   Mode  :character   Mode  :character
##                     Mean   :28.32
##                     3rd Qu.:31.42
##                     Max.   :94.85
##      Stroke          PhysicalHealth    MentalHealth    DiffWalking
##  Length:319582      Min.   : 0.000   Min.   : 0.000   Length:319582
##  Class :character   1st Qu.: 0.000   1st Qu.: 0.000   Class :character
##  Mode  :character   Median : 0.000   Median : 0.000   Mode  :character
##                     Mean   : 3.366   Mean   : 3.894
##                     3rd Qu.: 2.000   3rd Qu.: 3.000
##                     Max.   :30.000   Max.   :30.000
##       Sex            AgeCategory           Race             Diabetic
##  Length:319582      Length:319582      Length:319582      Length:319582
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  PhysicalActivity    GenHealth            SleepTime         Asthma
##  Length:319582      Length:319582      Min.   : 1.000   Length:319582
##  Class :character   Class :character   1st Qu.: 6.000   Class :character
##  Mode  :character   Mode  :character   Median : 7.000   Mode  :character
##                                        Mean   : 7.089
##                                        3rd Qu.: 8.000
##                                        Max.   :17.000
##  KidneyDisease       SkinCancer
##  Length:319582      Length:319582
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
dim(hd2)
```

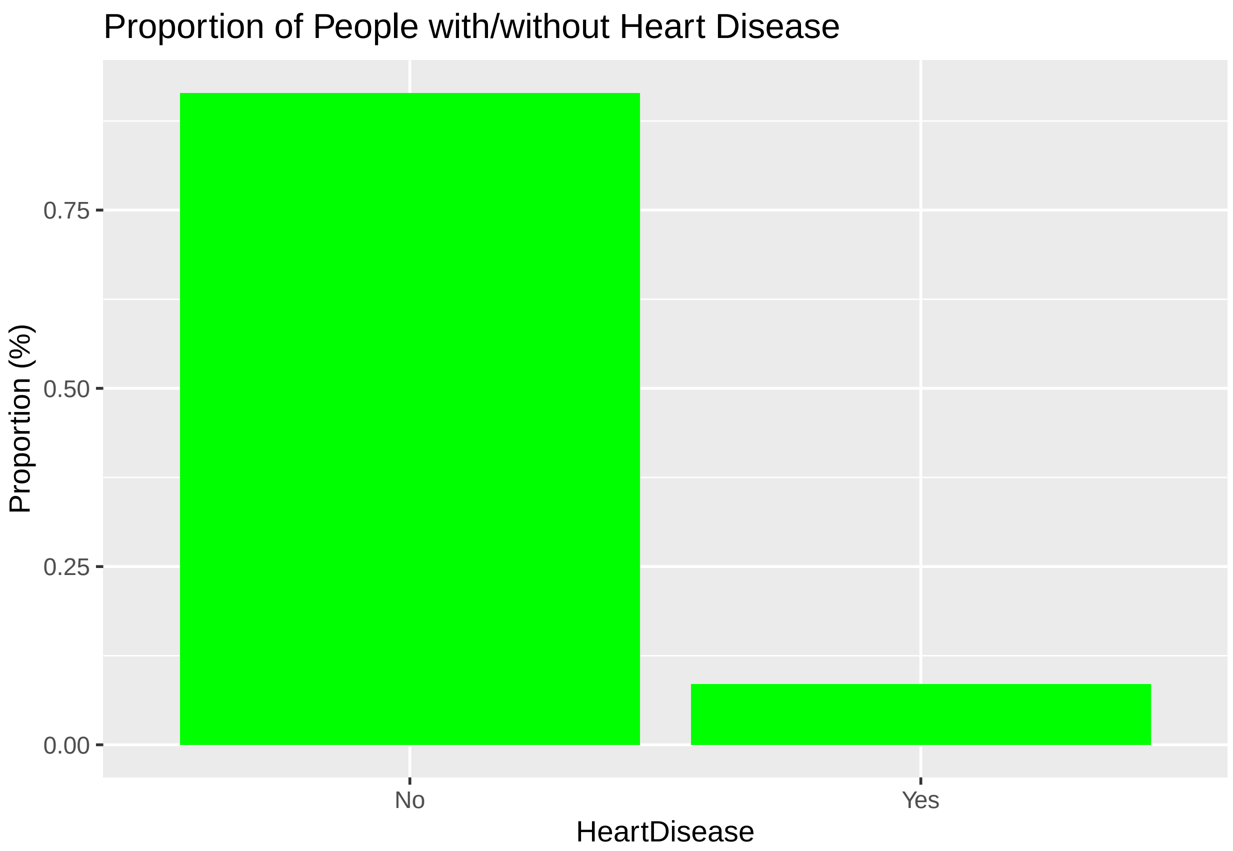
```
## [1] 319582     18
```

2. Exploring Data Through Visualization

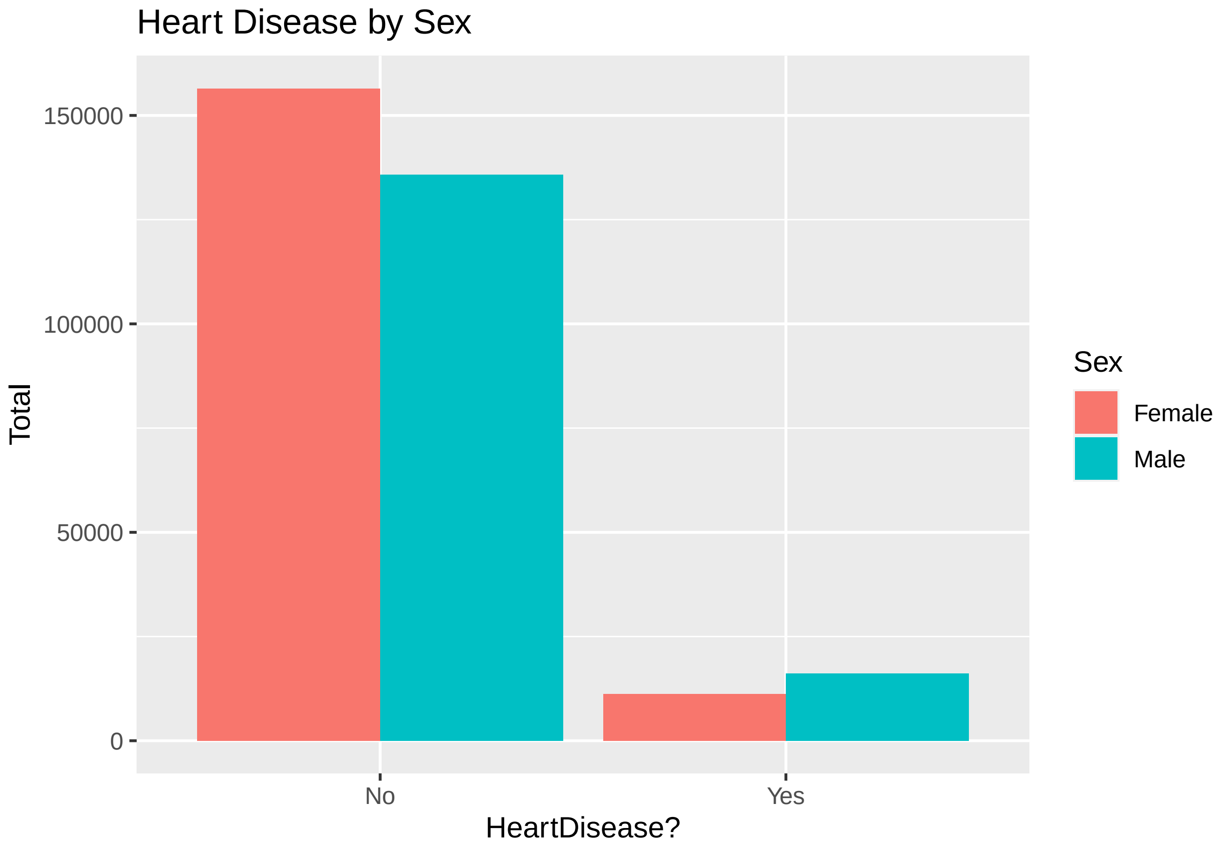
```
hist(hd2$BMI, col = "red", main = "Histogram of BMI")
```



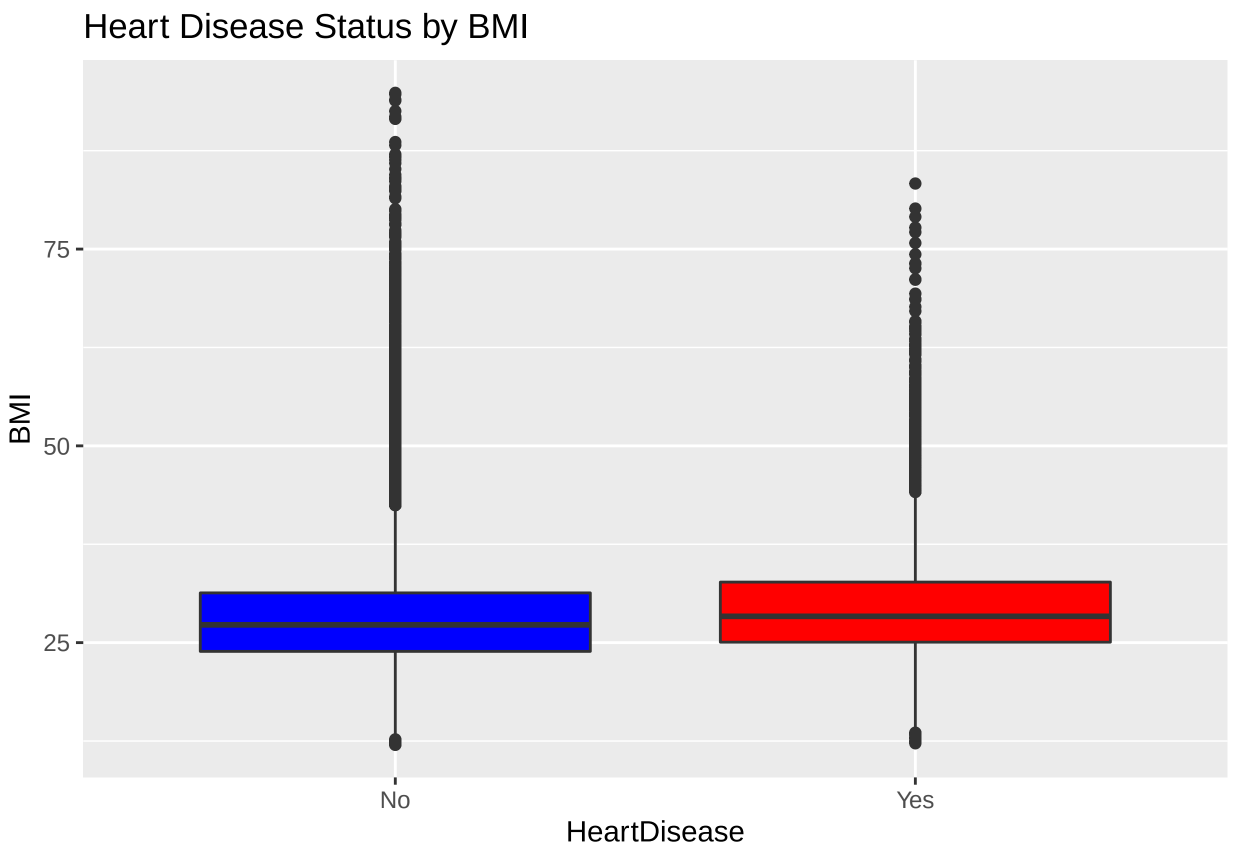
```
ggplot(hd2, aes(x = HeartDisease)) +
  geom_bar(aes(y = (..count..)/sum(..count..)), fill = "green") +
  ylab("Proportion (%)") +
  ggtitle("Proportion of People with/without Heart Disease")
```



```
hd_sx <- hd2 %>%
  select(Sex, HeartDisease) %>%
  group_by(Sex) %>%
  count(HeartDisease)
ggplot(hd_sx, aes(x = HeartDisease, y = n, fill = Sex)) +
  geom_bar(stat="identity", position = "dodge") +
  xlab("HeartDisease?") +
  ylab("Total") +
  ggtitle("Heart Disease by Sex")
```



```
ggplot(hd2, aes(x = HeartDisease, y = BMI, fill = HeartDisease)) +
  geom_boxplot(fill = c("blue", "red")) +
  ggtitle("Heart Disease Status by BMI")
```



3. Creating our Prediction Model.

```r
# install.packages("randomforest")
# install.packages("caret")

# First we gotta change the HeartDisease variable into a binary one
# consisting of 1' and 0's so we can use them in modeling.
hd_alt <- hd2 %>%
  mutate(HeartDisease = ifelse(HeartDisease == 'Yes', 1, 0))

class(hd_alt$HeartDisease)
```

```
## [1] "numeric"
```

```r
hd_alt$HeartDisease <- as.factor(hd_alt$HeartDisease) # Changes variable into factor

# Now we finally run a random forest model on the data (as per usual) to see
# if we can somehow obtain a model that can predict heart disease in people.
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.1.3
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(caret)
```

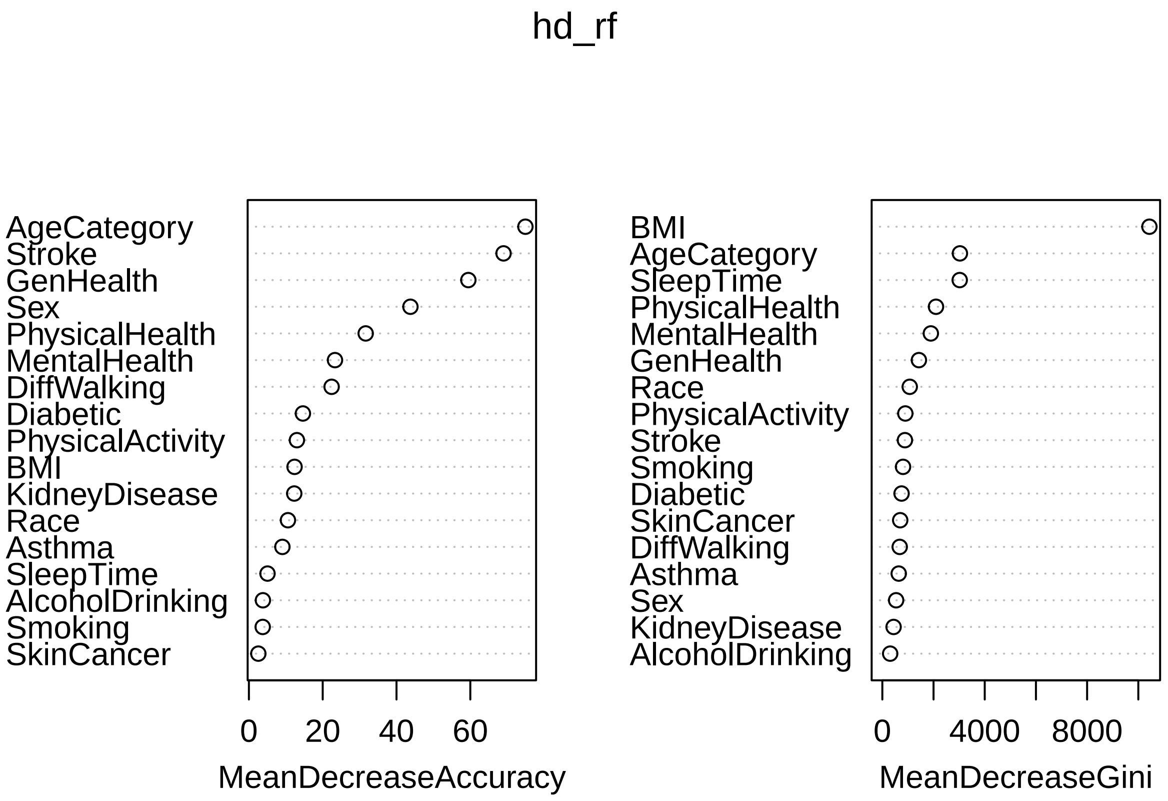
```
## Warning: package 'caret' was built under R version 4.1.3
```

```
## Loading required package: lattice
```

```
set.seed(5)
s <- createDataPartition(hd_alt$HeartDisease, p = 0.6, list = F)
train <- hd_alt[s, ]
test <- hd_alt[-s, ]

hd_rf <- randomForest(HeartDisease ~ ., train, mtry = 17,
                       importance = T, na.action = na.omit, ntree = 150)

hd_pred <- predict(hd_rf, test)
varImpPlot(hd_rf)
```



```
table_hd <- table("original" = test$HeartDisease, "prediction" = hd_pred)
table_hd
```

```
##           prediction
## original      0      1
##         0 114517   2385
##         1   9531   1399
```

```
prop.table(table_hd)
```

```
##           prediction
## original           0          1
##         0 0.89583985 0.01865730
##         1 0.07455880 0.01094405
```

```
acc <- sum(diag(table_hd)) / sum(table_hd)
acc # Calculation of the prediction accuracy.
```

```
## [1] 0.9067839
```