Cardiovascular disease detection

Tathagata Bardhan

Importing the dataset

```
rm(list=ls())
`cardio_train(project)` <- read.csv("C:/Users/user/OneDrive/Desktop/project-baan bapat/cardio_train(project).csv"
)
data=`cardio_train(project)`
head(data)
```

```
##   id   age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active
## 1  0 18393      2    168     62   110    80           1    1     0    0      1
## 2  1 20228      1    156     85   140    90           3    1     0    0      1
## 3  2 18857      1    165     64   130    70           3    1     0    0      0
## 4  3 17623      2    169     82   150   100           1    1     0    0      1
## 5  4 17474      1    156     56   100    60           1    1     0    0      0
## 6  8 21914      1    151     67   120    80           2    2     0    0      0
##   cardio
## 1      0
## 2      1
## 3      1
## 4      1
## 5      0
## 6      0
```

Data cleaning

Null value detection

```
sapply(data,function(x) sum(is.na(x)))
```

```
##          id         age      gender      height      weight       ap_hi
##           0           0           0           0           0           0
##       ap_lo cholesterol        gluc       smoke        alco      active
##           0           0           0           0           0           0
##      cardio
##           0
```

So there are no null values in the dataset

Identifying duplicate rows

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df=data %>% distinct(age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, active, cardio, .
keep_all=TRUE)
dim(df)
```

```
## [1] 69976    13
```

Although the id is unique for all the patients, there are 24 duplicate rows based on all other columns

Overview and summary statistics of the data

```
str(df)
```

```
## 'data.frame':    69976 obs. of  13 variables:
##  $ id         : int  0 1 2 3 4 8 9 12 13 14 ...
##  $ age        : int  18393 20228 18857 17623 17474 21914 22113 22584 17668 19834 ...
##  $ gender     : int  2 1 1 2 1 1 1 2 1 1 ...
##  $ height     : int  168 156 165 169 156 151 157 178 158 164 ...
##  $ weight     : num  62 85 64 82 56 67 93 95 71 68 ...
##  $ ap_hi      : int  110 140 130 150 100 120 130 130 110 110 ...
##  $ ap_lo      : int  80 90 70 100 60 80 80 90 70 60 ...
##  $ cholesterol: int  1 3 3 1 1 2 3 3 1 1 ...
##  $ gluc       : int  1 1 1 1 1 2 1 3 1 1 ...
##  $ smoke      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ alco       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ active     : int  1 1 0 1 0 0 1 1 1 0 ...
##  $ cardio     : int  0 1 1 1 0 0 0 1 0 0 ...
```

We require gender,cholesterol,gluc,smoke,alco,active,cardio variables to be Factors but these are continuous here. So we need explicit conversion

```
df[,c(3,8:13)]=lapply(df[,c(3,8:13)],function(x) as.factor(x))
str(df)
```

```
## 'data.frame':    69976 obs. of  13 variables:
##  $ id         : int  0 1 2 3 4 8 9 12 13 14 ...
##  $ age        : int  18393 20228 18857 17623 17474 21914 22113 22584 17668 19834 ...
##  $ gender     : Factor w/ 2 levels "1","2": 2 1 1 2 1 1 1 2 1 1 ...
##  $ height     : int  168 156 165 169 156 151 157 178 158 164 ...
##  $ weight     : num  62 85 64 82 56 67 93 95 71 68 ...
##  $ ap_hi      : int  110 140 130 150 100 120 130 130 110 110 ...
##  $ ap_lo      : int  80 90 70 100 60 80 80 90 70 60 ...
##  $ cholesterol: Factor w/ 3 levels "1","2","3": 1 3 3 1 1 2 3 3 1 1 ...
##  $ gluc       : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 1 3 1 1 ...
##  $ smoke      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ alco       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ active     : Factor w/ 2 levels "0","1": 2 2 1 2 1 1 2 2 2 1 ...
##  $ cardio     : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 1 1 ...
```

```
summary(df[,-c(1)])
```

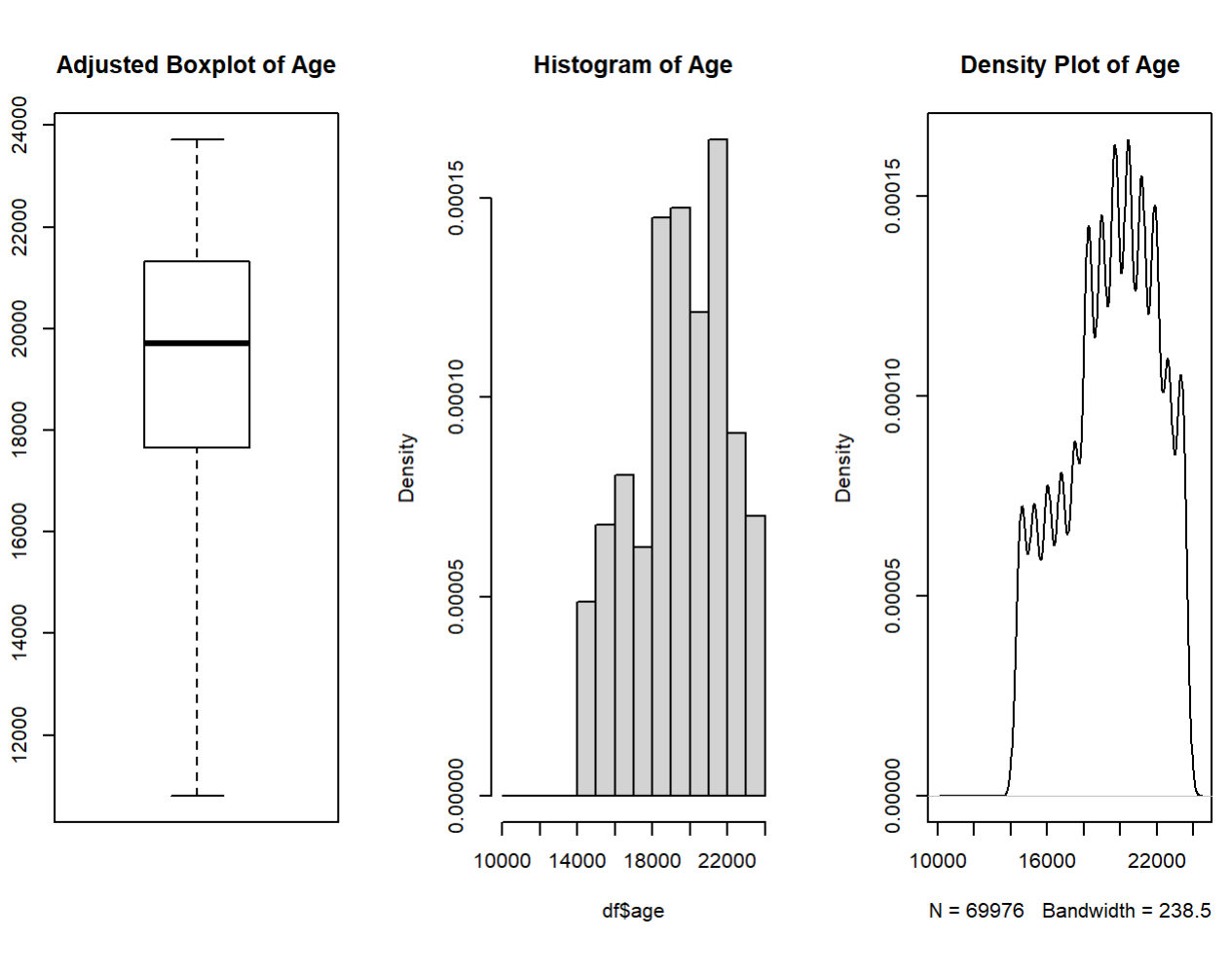
```
##       age        gender        height          weight           ap_hi
##  Min.   :10798   1:45509   Min.   : 55.0   Min.   : 10.00   Min.   : -150.0
##  1st Qu.:17664   2:24467   1st Qu.:159.0   1st Qu.: 65.00   1st Qu.:  120.0
##  Median :19703             Median :165.0   Median : 72.00   Median :  120.0
##  Mean   :19469             Mean   :164.4   Mean   : 74.21   Mean   :  128.8
##  3rd Qu.:21327             3rd Qu.:170.0   3rd Qu.: 82.00   3rd Qu.:  140.0
##  Max.   :23713             Max.   :250.0   Max.   :200.00   Max.   :16020.0
##      ap_lo          cholesterol gluc      smoke     alco      active
##  Min.   :  -70.00   1:52361     1:59455   0:63807   0:66212   0:13735
##  1st Qu.:   80.00   2: 9549     2: 5190   1: 6169   1: 3764   1:56241
##  Median :   80.00   3: 8066     3: 5331
##  Mean   :   96.64
##  3rd Qu.:   90.00
##  Max.   :11000.00
##  cardio
##  0:35004
##  1:34972
##
##
##
##
```

Thus the values for each variable are of much different scales from each other and so we may need to standardize these later

Univariate Analysis

For variable 'Age'

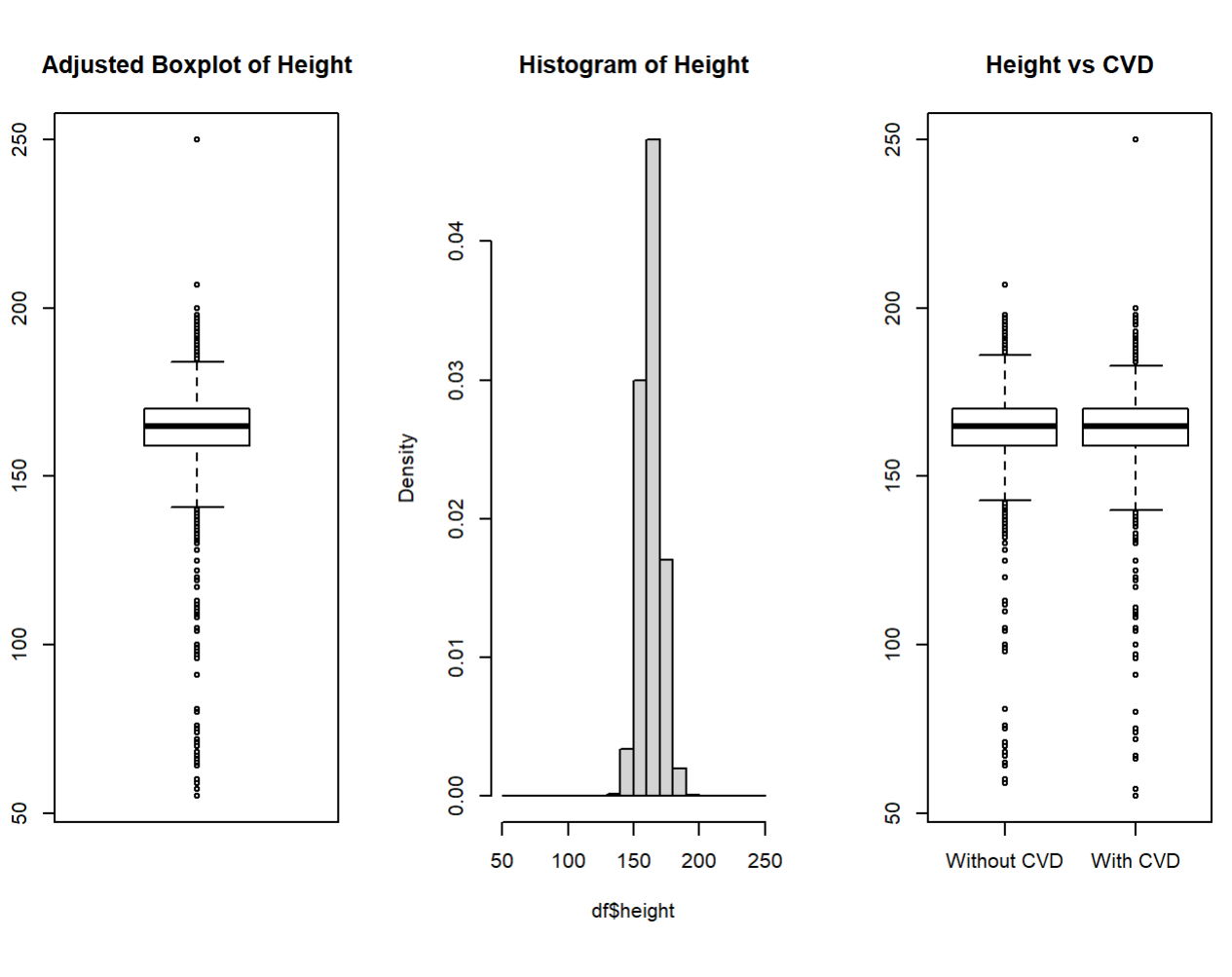
```
par(mfrow=c(1,3))
library(robustbase)
adjbox(df$age,main="Adjusted Boxplot of Age")
hist(df$age,main="Histogram of Age",prob=TRUE)
plot(density(df$age),main="Density Plot of Age")
```



- From the summary statistics of Age we see that here we have people aged from 29 years to 65 years approx and more precisely majority of the patients are between 49 years to 60 years
- The distribution of Age is slightly negatively skewed
- Adjusted boxplot is preferred here, which shows no presence of outliers

For variable 'Height'

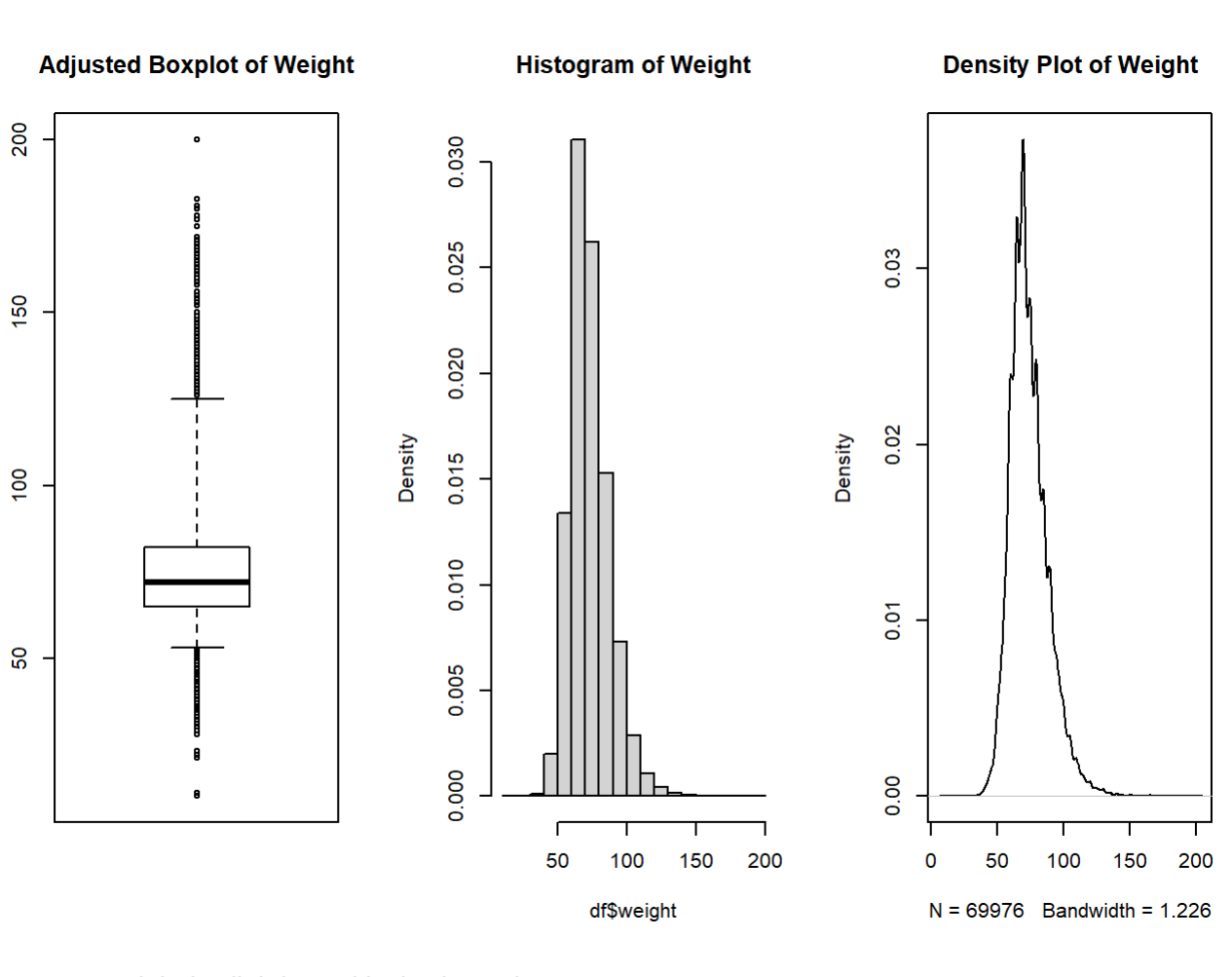
```
par(mfrow=c(1,3))
adjbox(df$height,main="Adjusted Boxplot of Height")
hist(df$height,main="Histogram of Height",prob=TRUE)
adjbox(df$height~df$cardio,names=c("Without CVD","With CVD"),main="Height vs CVD")
```



- Heights are more or less symmetrically distributed
- Most of the patients are of height 160 to 170 cm
- On an average, patients with or without CVD have more or less same height.
  - Thus height is not an important factor in detecting CVD
  - So, we are bothered about the outliers present in variable Height

For variable 'Weight'

```
par(mfrow=c(1,3))
adjbox(df$weight,main="Adjusted Boxplot of Weight")
hist(df$weight,main="Histogram of Weight",prob=TRUE)
plot(density(df$weight),main="Density Plot of Weight")
```

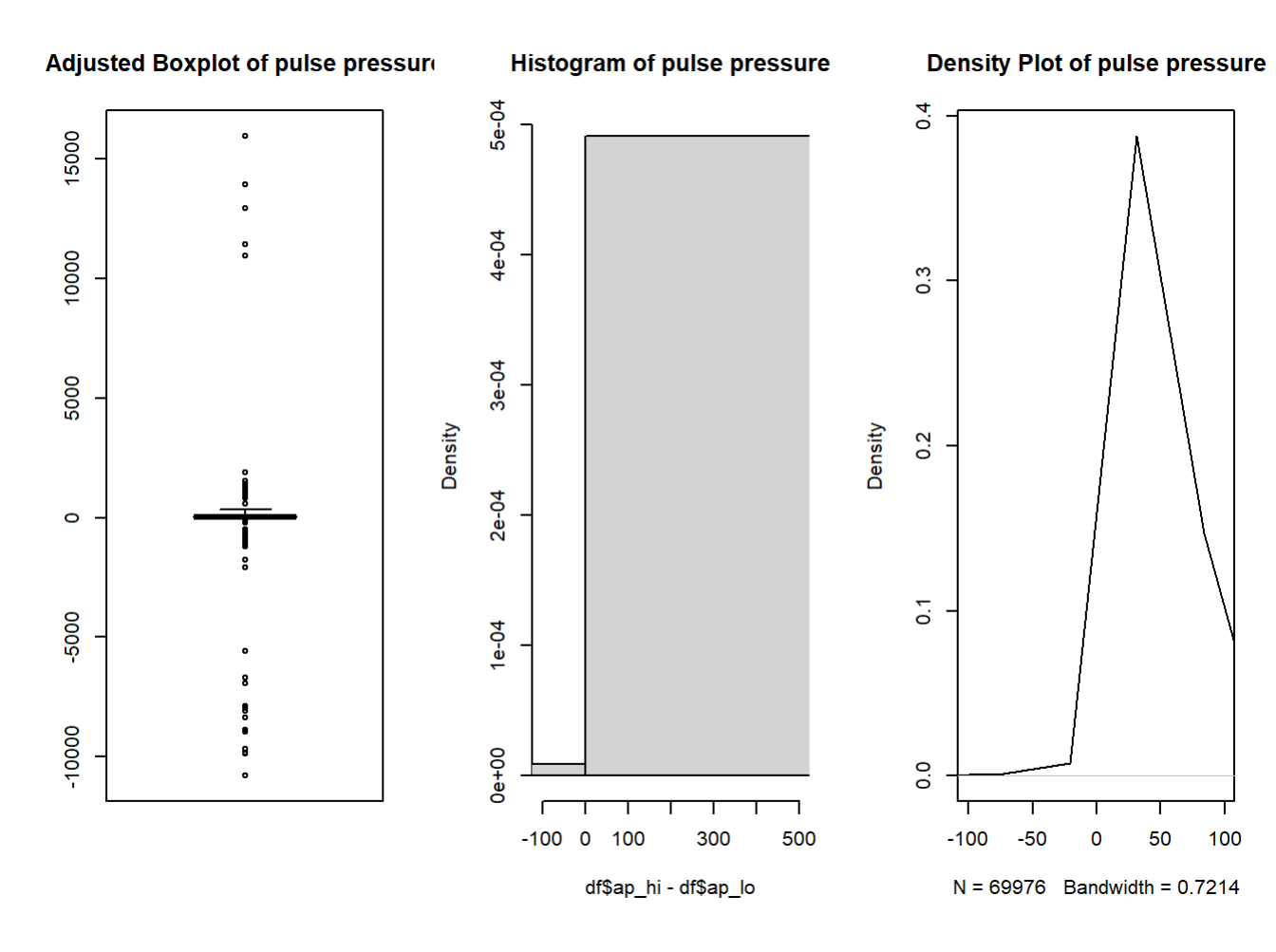


- Here Weight is slightly positively skewed
- Majority of people in our data have weight 60 to 80 kg
- From the summary statistics we can argue that Weight of 10 kgs is not possible for a person 29 years of Age and above.
  - So, we suspect that outliers beyond the lower whisker of the boxplot of Weight are only the meaningful ones that should be removed.

```
df$weight[df$weight<adjboxStats(df$weight)$fence[1]]=adjboxStats(df$weight)$fence[1]
```

For pulse pressure

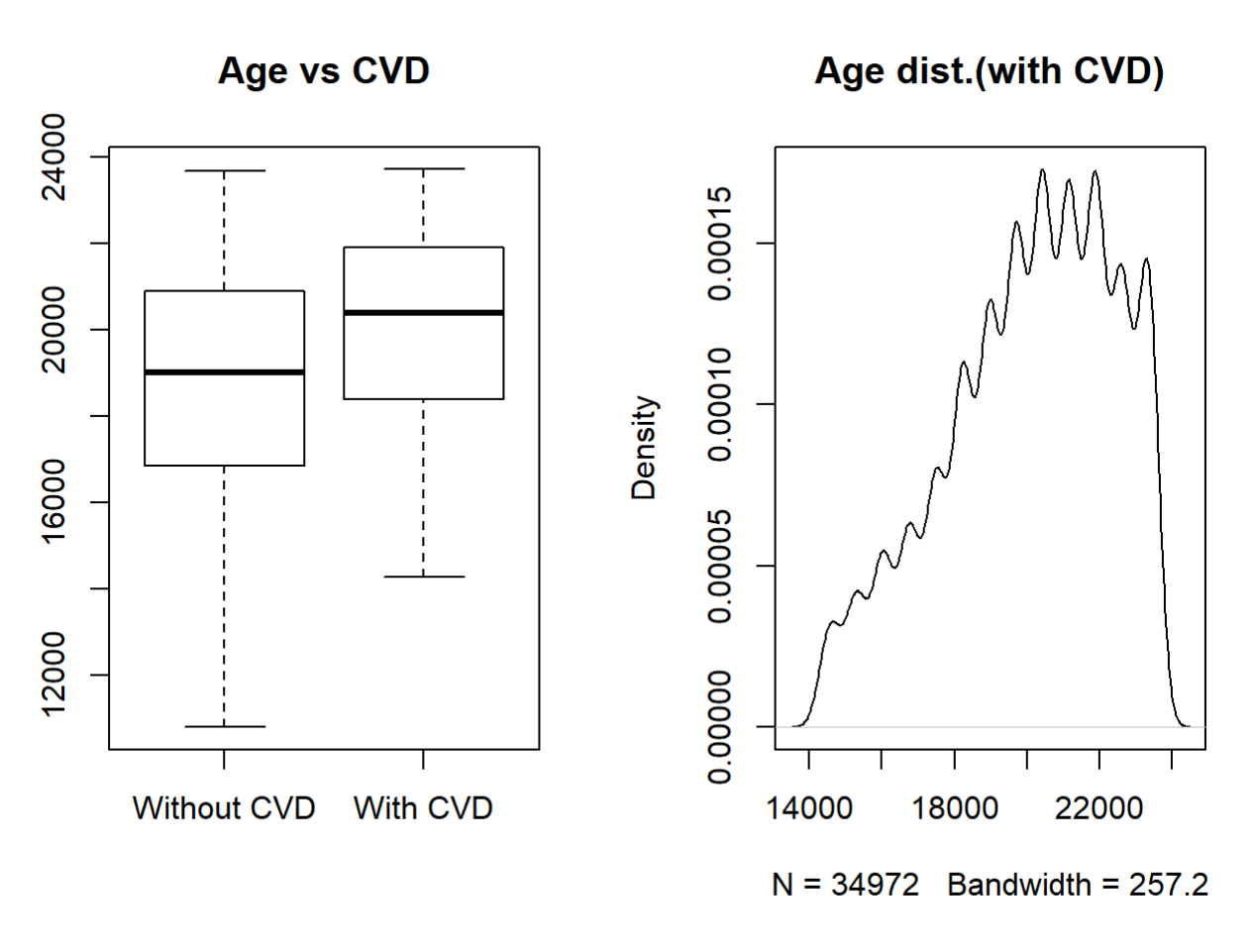
```
par(mfrow=c(1,3))
adjbox(df$ap_hi-df$ap_lo,main="Adjusted Boxplot of pulse pressure")
hist(df$ap_hi-df$ap_lo,xlim=c(-100,500),main="Histogram of pulse pressure",prob=TRUE)
plot(density(df$ap_hi-df$ap_lo),xlim=c(-100,100),main="Density Plot of pulse pressure")
```



Bivariate Analysis

Comparison of 'Age' with CVD

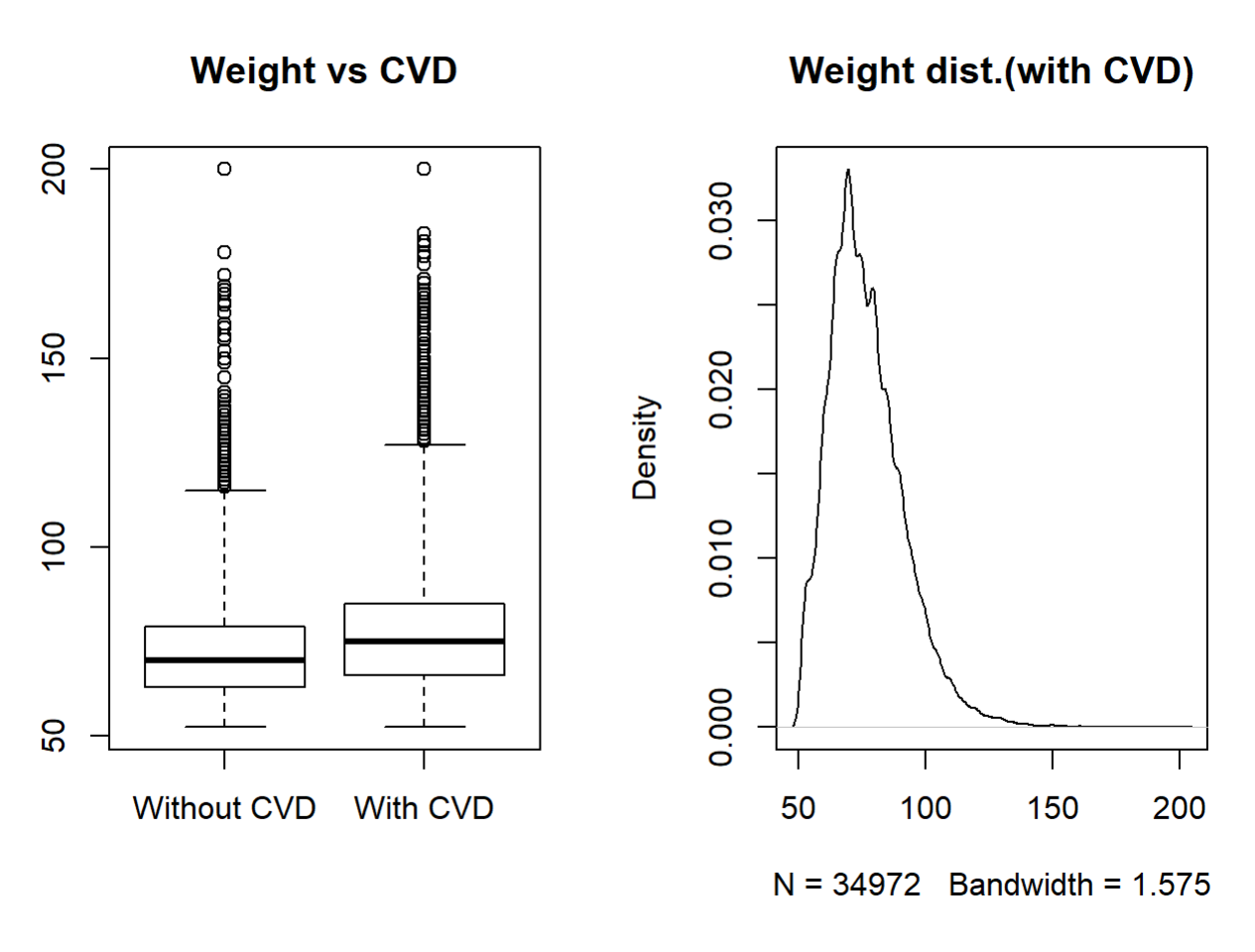
```
par(mfrow=c(1,2))
adjbox(df$age~df$cardio,names=c("Without CVD","With CVD"),main="Age vs CVD")
plot(density(df$age[df$cardio==1]),main="Age dist.(with CVD)")
```



- Patients who are suffering from CVD are, on an average, older than patients who donot suffer from CVD
- More precisely patients having age between 20,000-22,000 days i.e. between 55 to 60 years approximately suffer mostly from CVD

Comparison of 'Weight' with CVD

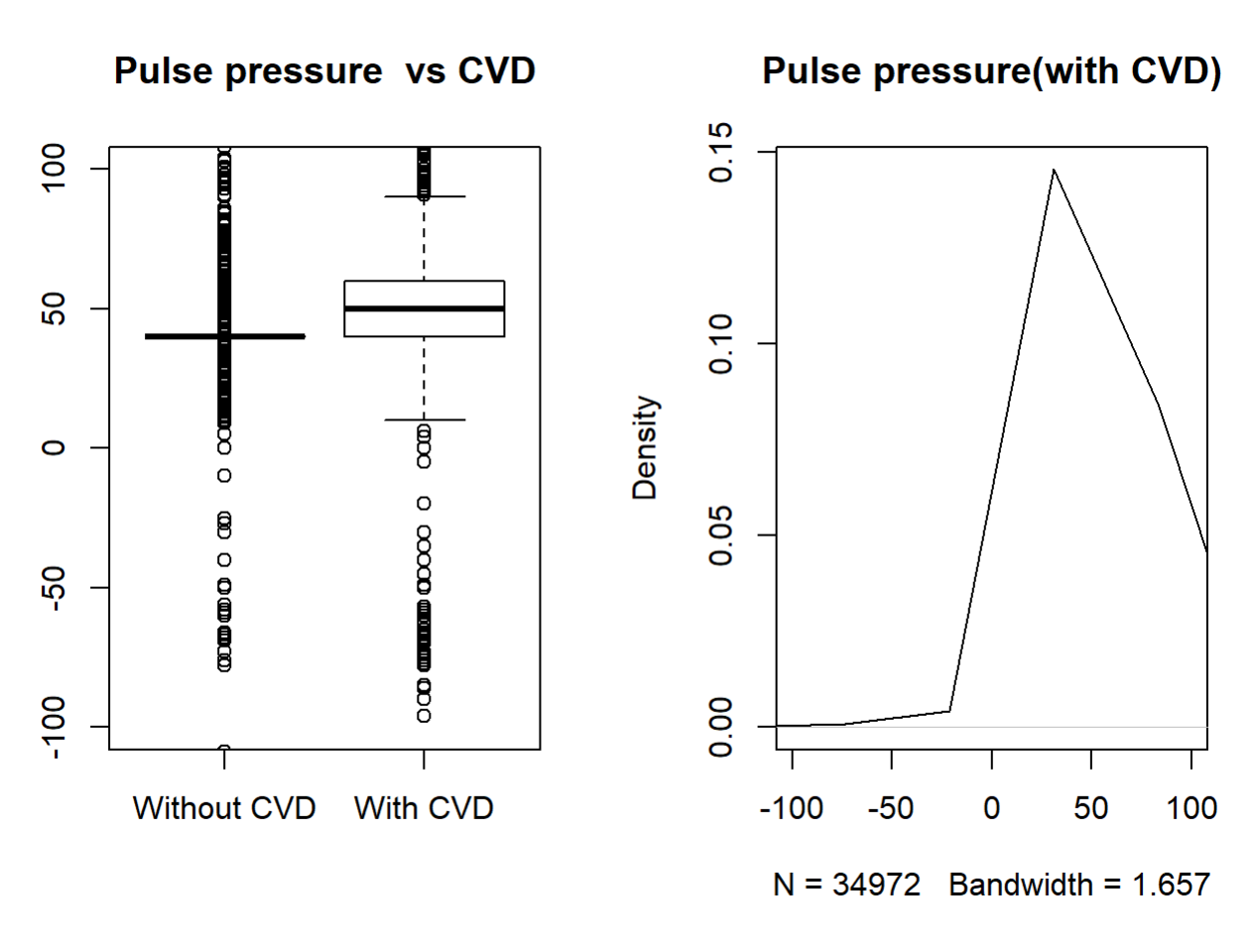
```
par(mfrow=c(1,2))
adjbox(df$weight~df$cardio,names=c("Without CVD","With CVD"),main="Weight vs CVD")
plot(density(df$weight[df$cardio==1]),main="Weight dist.(with CVD)")
```



- Thus, on an average, people who suffer from CVD have more weight than those who don't
- Infact heart diseases are mostly occuring for patients whose weights are within 65-80 kgs

Comparison of 'Pulse pressure' with CVD

```
par(mfrow=c(1,2))
adjbox((df$ap_hi-df$ap_lo)~df$cardio,ylim=c(-100,100),names=c("Without CVD","With CVD"),main="Pulse pressure  vs
 CVD")
plot(density(df$ap_hi[df$cardio==1]-df$ap_lo[df$cardio==1]),xlim=c(-100,100),main="Pulse pressure(with CVD)")
```



- On an average, people suffering from CVD have higher Pulse pressure
- Many people suffering from CVD have Pulse pressure higher than normal range i.e. greater than 60 mm Hg
- So, high pulse pressure can be a good indication of CVD

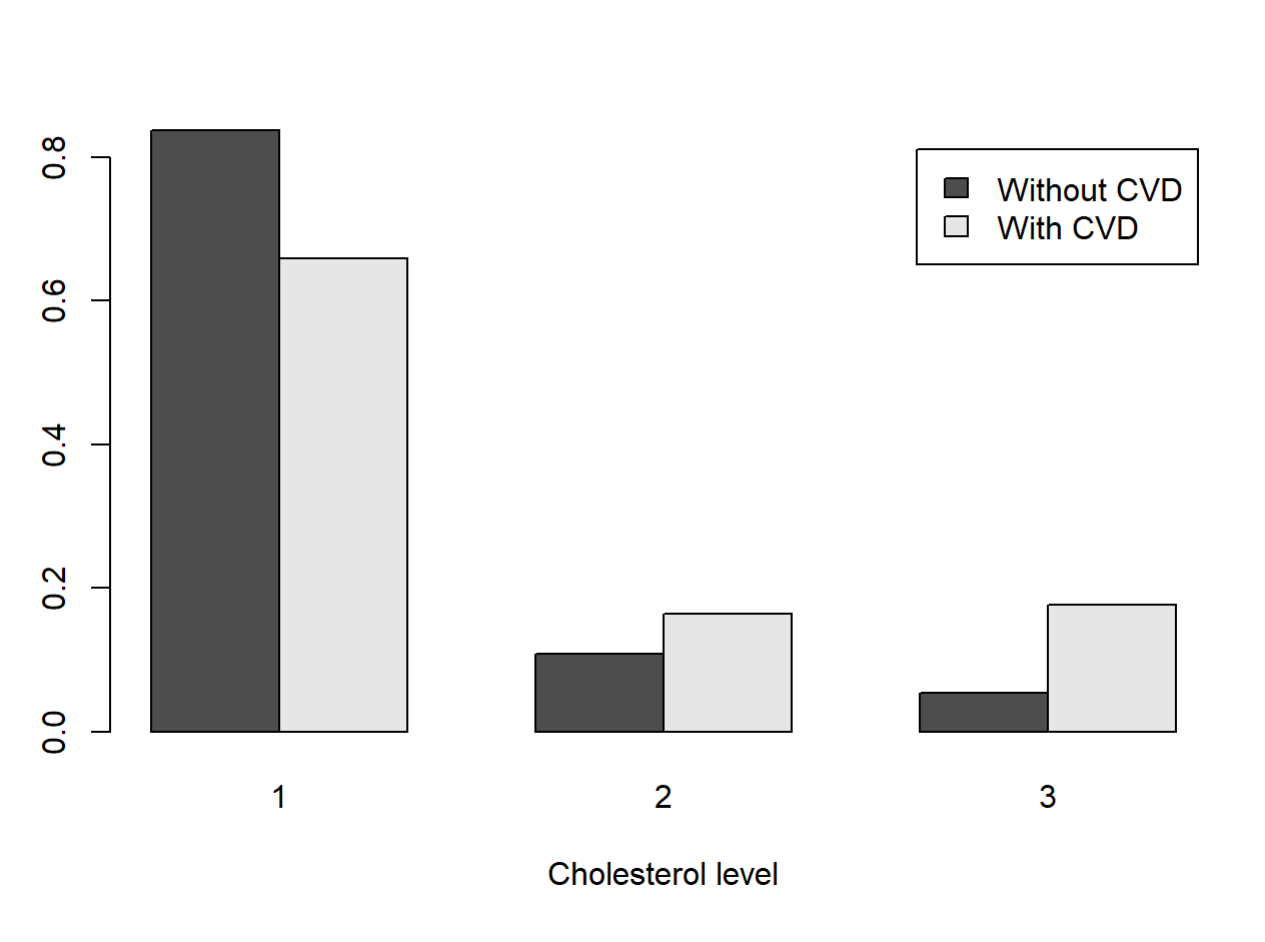
Comparison of CVD with 'Cholesterol level'

```
library(gmodels)
CrossTable(df$cholesterol,df$cardio,prop.c = F,prop.r = F,prop.t = F,prop.chisq = F,chisq = T)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  69976
##
##
##               | df$cardio
## df$cholesterol |         0 |         1 | Row Total |
## ---------------|-----------|-----------|-----------|
##              1 |     29313 |     23048 |     52361 |
## ---------------|-----------|-----------|-----------|
##              2 |      3799 |      5750 |      9549 |
## ---------------|-----------|-----------|-----------|
##              3 |      1892 |      6174 |      8066 |
## ---------------|-----------|-----------|-----------|
##   Column Total |     35004 |     34972 |     69976 |
## ---------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  3421.399     d.f. =  2     p =  0
##
##
##
```

Chi-sq test of Association shows dependence of presence of CVD on Cholesterol level

```
par(mfrow=c(1,1))
mytable=table(df$cardio,df$cholesterol)
mytable2=prop.table(mytable, 1)
barplot(mytable2,beside = TRUE,legend.text = c("Without CVD","With CVD"),xlab="Cholesterol level")
```



- Patients having Cholesterol level well above normal suffer very much from CVD
- Patients having Cholesterol in normal level are much less sufferers of CVD

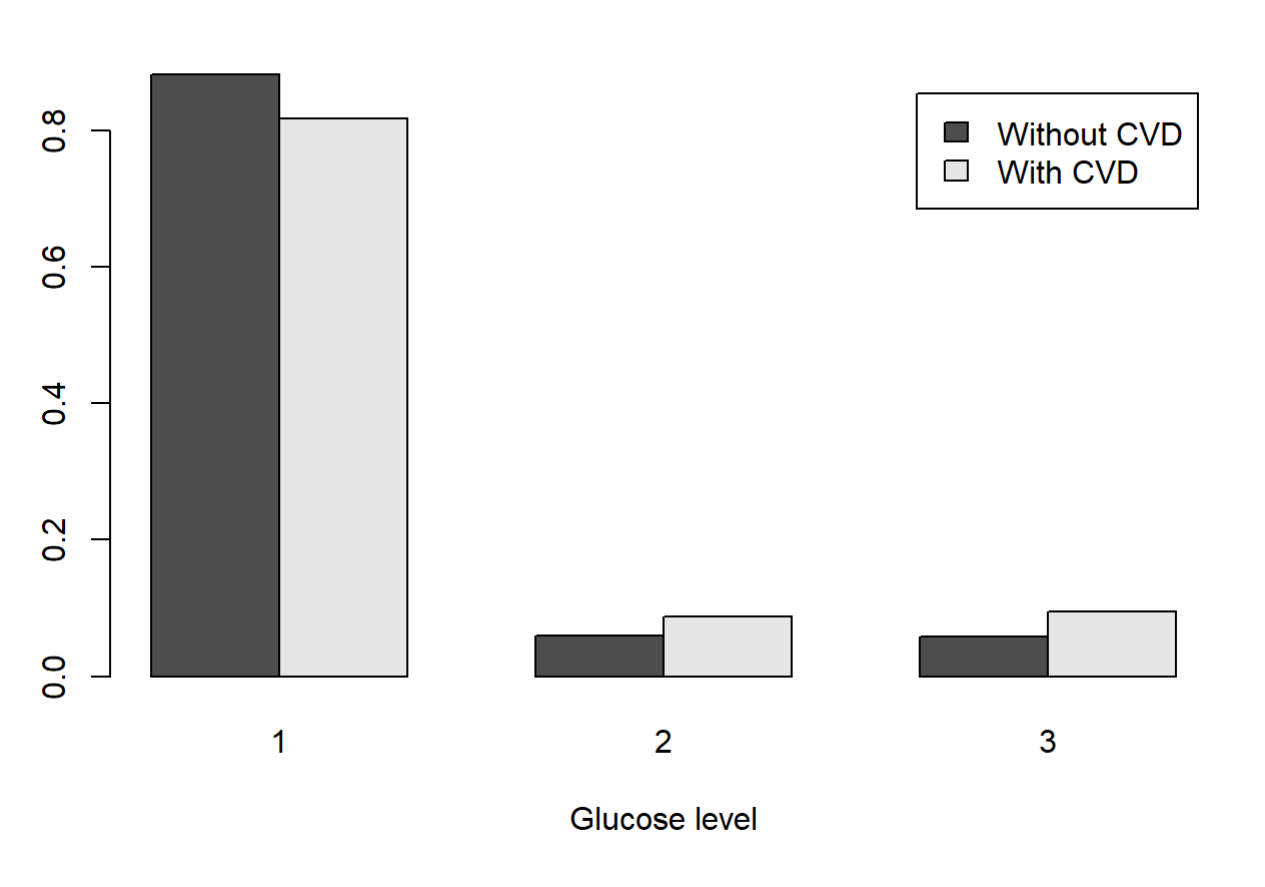
Comparison of CVD with 'Glucose level'

```
CrossTable(df$gluc,df$cardio,prop.c = F,prop.r = F,prop.t = F,prop.chisq = F,chisq = T)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  69976
##
##
##              | df$cardio
##      df$gluc |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |     30877 |     28578 |     59455 |
## -------------|-----------|-----------|-----------|
##            2 |      2112 |      3078 |      5190 |
## -------------|-----------|-----------|-----------|
##            3 |      2015 |      3316 |      5331 |
## -------------|-----------|-----------|-----------|
## Column Total |     35004 |     34972 |     69976 |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  586.1834     d.f. =  2     p =  5.150929e-128
##
##
##
```

Chi-sq test of Association tells that presence of CVD depends upon Glucose level

```
par(mfrow=c(1,1))
mytable=table(df$cardio,df$gluc)
mytable3=prop.table(mytable, 1)
barplot(mytable3,beside = TRUE,legend.text = c("Without CVD","With CVD"),xlab="Glucose level")
```



- Patients having Glucose level well above normal have more tendency to develop CVD
- Patients having normal Glucose level suffers comparatively less from CVD

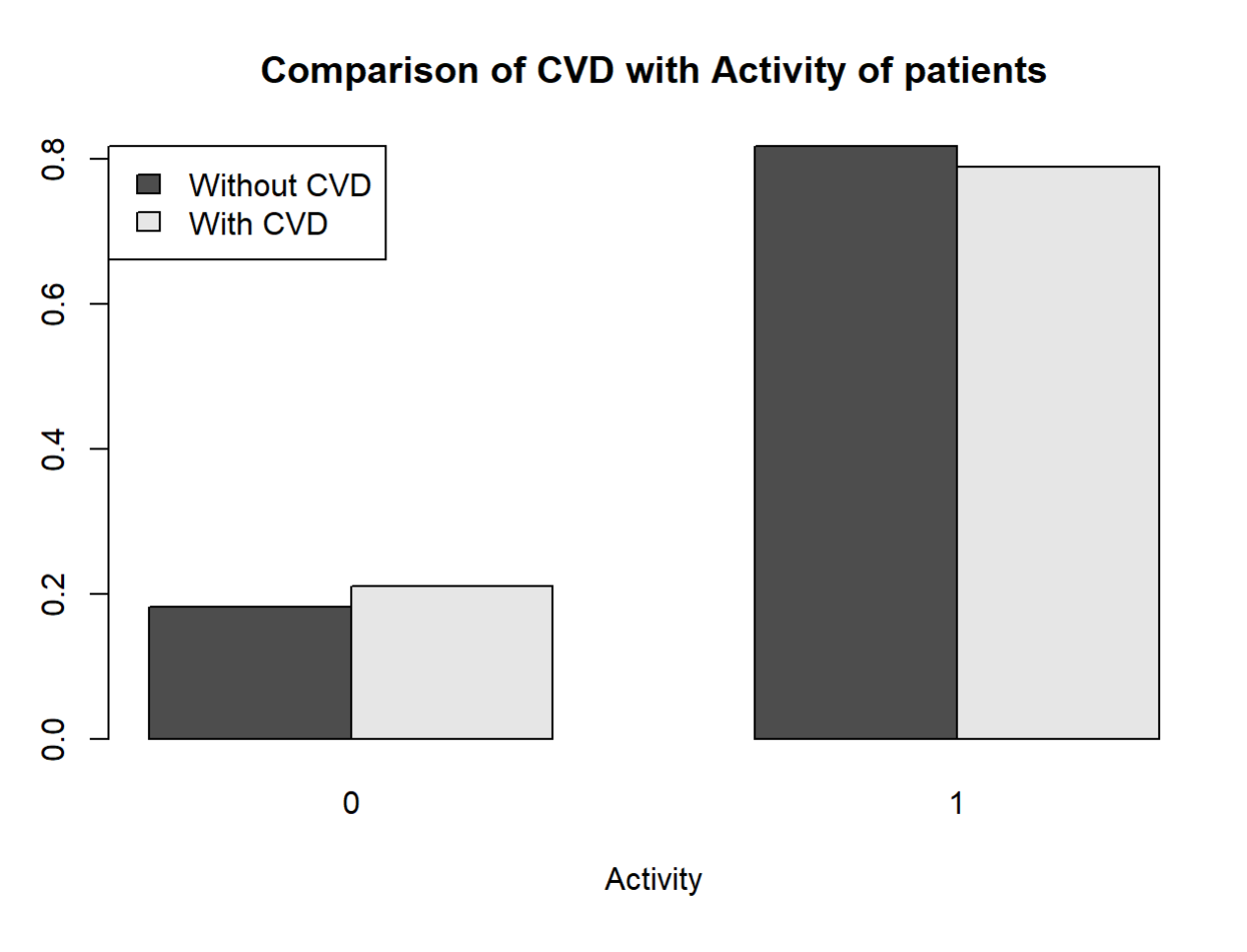
Comparison of CVD with 'Activity'

```
CrossTable(df$active,df$cardio,prop.c = F,prop.r = F,prop.t = F,prop.chisq = F,chisq = T)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  69976
##
##
##              | df$cardio
##    df$active |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |      6375 |      7360 |     13735 |
## -------------|-----------|-----------|-----------|
##            1 |     28629 |     27612 |     56241 |
## -------------|-----------|-----------|-----------|
## Column Total |     35004 |     34972 |     69976 |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  89.01457     d.f. =  1     p =  3.919161e-21
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  88.83506     d.f. =  1     p =  4.29142e-21
##
##
```

From the results of Chi-sq test of Association we see that presence of CVD depends on physical activity of patients

```
par(mfrow=c(1,1))
mytable=table(df$cardio,df$active)
mytable4=prop.table(mytable, 1)
barplot(mytable4,beside = TRUE,legend.text = c("Without CVD","With CVD"),args.legend = list(x = "topleft"),xlab=
"Activity",main="Comparison of CVD with Activity of patients")
```



- There are more CVD patients among those who are not active
- There is a bit lesser tendency to develop CVD among the people who are active

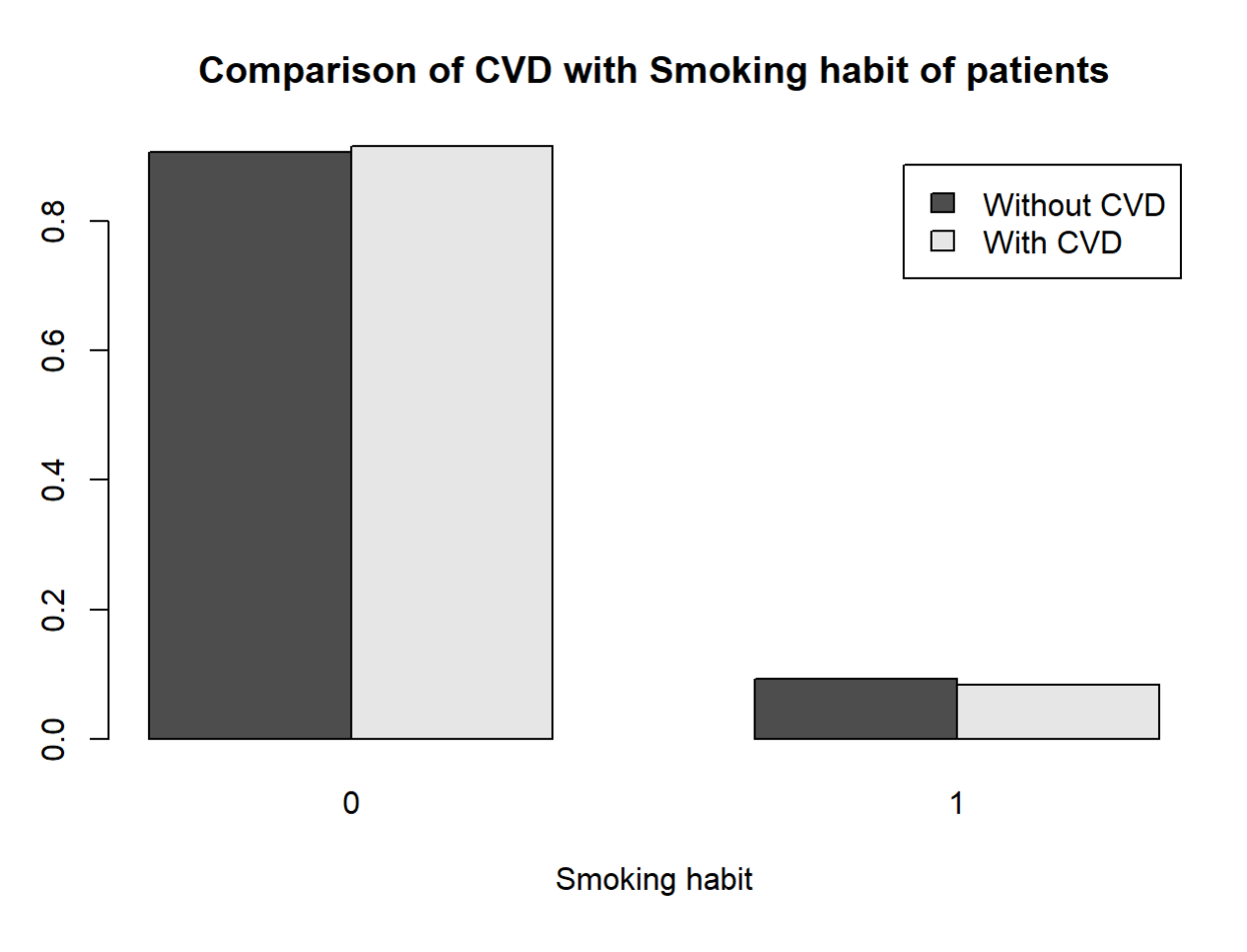
Comparison of CVD with 'Smoking'

```
CrossTable(df$smoke,df$cardio,prop.c = F,prop.r = F,prop.t = F,prop.chisq = F,chisq = T)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  69976
##
##
##              | df$cardio
##     df$smoke |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |     31764 |     32043 |     63807 |
## -------------|-----------|-----------|-----------|
##            1 |      3240 |      2929 |      6169 |
## -------------|-----------|-----------|-----------|
## Column Total |     35004 |     34972 |     69976 |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  16.88387     d.f. =  1     p =  3.973787e-05
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  16.77447     d.f. =  1     p =  4.209577e-05
##
##
```

Chi-sq test of Association indicates dependence of CVD on smoking habits

```
par(mfrow=c(1,1))
mytable=table(df$cardio,df$smoke)
mytable5=prop.table(mytable, 1)
barplot(mytable5,beside = TRUE,legend.text = c("Without CVD","With CVD"),xlab="Smoking habit",main="Comparison of
CVD with Smoking habit of patients")
```



Pairwise correlation coefficients

```
round((cor(df[,-c(1,3,8:13)])),4)
```

```
##            age  height weight  ap_hi  ap_lo
## age     1.0000 -0.0815 0.0533 0.0208 0.0176
## height -0.0815  1.0000 0.2852 0.0055 0.0062
## weight  0.0533  0.2852 1.0000 0.0306 0.0439
## ap_hi   0.0208  0.0055 0.0306 1.0000 0.0161
## ap_lo   0.0176  0.0062 0.0439 0.0161 1.0000
```

```
cor.test(df$ap_hi,df$ap_lo)$p.value
```

```
## [1] 2.092353e-05
```

There is significant correlation between variables 'ap_hi' and 'ap_lo' which may cause the problem of multicollinearity later