DATA 606 Data Project Proposal

Eric Lehmphul

10/31/2021

Data Preparation

Retrieved dataset from Kaggle: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset. It is a dataset relating to cardiovascular disease and relative variables of interest.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

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

```
# load data

cardio.data <- read.csv("https://raw.githubusercontent.com/SaneSky109/DATA606/main/Data_Project/Data/ca
```

```
# remove unecessary column: id
cardio.data <- cardio.data[,-1]
```

```
# create factors
cardio.data$cardio <- factor(cardio.data$cardio)
cardio.data$gender <- factor(cardio.data$gender)
cardio.data$cholesterol <- factor(cardio.data$cholesterol)
cardio.data$gluc <- factor(cardio.data$gluc)
cardio.data$smoke <- factor(cardio.data$smoke)
cardio.data$alco <- factor(cardio.data$alco)
cardio.data$active <- factor(cardio.data$active)
```

```
# rename factor levels

levels(cardio.data$cardio) <- c("No", "Yes")
levels(cardio.data$gender) <- c("Female", "Male")
levels(cardio.data$cholesterol) <- c("Normal", "Above_Normal", "Well_Above_Normal")
levels(cardio.data$gluc) <- c("Normal", "Above_Normal", "Well_Above_Normal")
levels(cardio.data$smoke) <- c("No", "Yes")
levels(cardio.data$alco) <- c("No", "Yes")
levels(cardio.data$active) <- c("No", "Yes")
```

```
# transform age since it is in days

cardio.data$age <- cardio.data$age/365

# remove outliers of ap_hi

# I am assuming the that these measures are errors and
# I am just dropping them due to problems it will cause with modeling
# Highest pressure recorded in an individual was 370/360.(https://pubmed.ncbi.nlm.nih.gov/7741618/)

summary(cardio.data$ap_hi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -150.0   120.0   120.0   128.8   140.0 16020.0
```

```
cardio.data <- cardio.data[cardio.data$ap_hi <= 370,]
cardio.data <- cardio.data[cardio.data$ap_hi > 0,]

summary(cardio.data$ap_hi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   120.0   120.0   126.7   140.0   309.0
```

```
# remove outliers of ap_lo


summary(cardio.data$ap_lo)
```

```
##     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   -70.00     80.00     80.00     96.65     90.00 11000.00
```

```
cardio.data <- cardio.data[cardio.data$ap_lo <= 360,]
cardio.data <- cardio.data[cardio.data$ap_lo > 0,]

summary(cardio.data$ap_lo)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   80.00   80.00   81.35   90.00  190.00
```

```
glimpse(cardio.data)
```

```
## Rows: 68,985
## Columns: 12
## $ age         <dbl> 50.39178, 55.41918, 51.66301, 48.28219, 47.87397, 60.03836~
## $ gender      <fct> Male, Female, Female, Male, Female, Female, Female, Male, ~
## $ height      <int> 168, 156, 165, 169, 156, 151, 157, 178, 158, 164, 169, 173~
## $ weight      <dbl> 62, 85, 64, 82, 56, 67, 93, 95, 71, 68, 80, 60, 60, 78, 95~
## $ ap_hi       <int> 110, 140, 130, 150, 100, 120, 130, 130, 110, 110, 120, 120~
## $ ap_lo       <int> 80, 90, 70, 100, 60, 80, 80, 90, 70, 60, 80, 80, 80, 70, 9~
## $ cholesterol <fct> Normal, Well_Above_Normal, Well_Above_Normal, Normal, Norm~
## $ gluc        <fct> Normal, Normal, Normal, Normal, Normal, Above_Normal, Norm~
## $ smoke       <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, No, Ye~
## $ alco        <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, No, Ye~
## $ active      <fct> Yes, Yes, No, Yes, No, No, Yes, Yes, Yes, No, Yes, Yes, No~
## $ cardio      <fct> No, Yes, Yes, Yes, No, No, No, Yes, No, No, No, No, No, No~
```

Research question

My research question is: **Do gender, age, body weight, body height, blood pressure, cholesterol, glucose levels, smoking, drinking alcohol and activity level of an individual significantly influence the likelihood of contracting cardiovascular disease?**

I aim to determine what variables are the most important determining factors to cardiovascular disease given the data presented in the dataset.

Cases

The cases are the number of people who participate in the medical examination. There were a total of 70,000 cases in the original data file. After data pre-processing, the number of cases is 68,985. This change is due to the removal of rows that seemed to be errors such as extremely high and low blood pressure (-1,000 or 15,000).

```
nrow(cardio.data)
```

```
## [1] 68985
```

Data collection

The data was collected from medical information given by patient and examination results. "All of the dataset values were collected at the moment of medical examination." (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset)

The data was downloaded from Kaggle (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset) and then I uploaded it to Github to be used to import the data into R.

Type of study

This is an observational study since the analysis is on events that have already occurred.

Data Source

The link to where I retrieved the data is: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Dependent Variable

The response variable is `cardio`. This is a qualitative variable since it is a categorical binary variable. `cardio` is an indicator variable that indicates whether or not someone has cardiovascular disease.

Independent Variable

There are multiple variables that I am considering for analysis. The list contains a group of both quantitative and qualitative variables:

- `age` (quantitative): Age of patient in years
- `gender` (qualitative): Gender of patient
- `height` (quantitative): Height of patient in cm
- `weight` (quantitative): Weight of patient in kg
- `ap_hi` (quantitative): Systolic blood pressure
- `ap_lo` (quantitative): Diastolic blood pressure
- `cholesterol` (qualitative): Cholesterol level of patient
- `smoke` (qualitative): Binary variable to determine if a patient smokes
- `alco` (qualitative): Binary variable to determine if a patient drinks alcohol
- `gluc` (qualitative): Glucose level of patient
- `active` (qualitative): Yes/No if patient is physically active

Relevant Summary Statistics

Summary Statistics

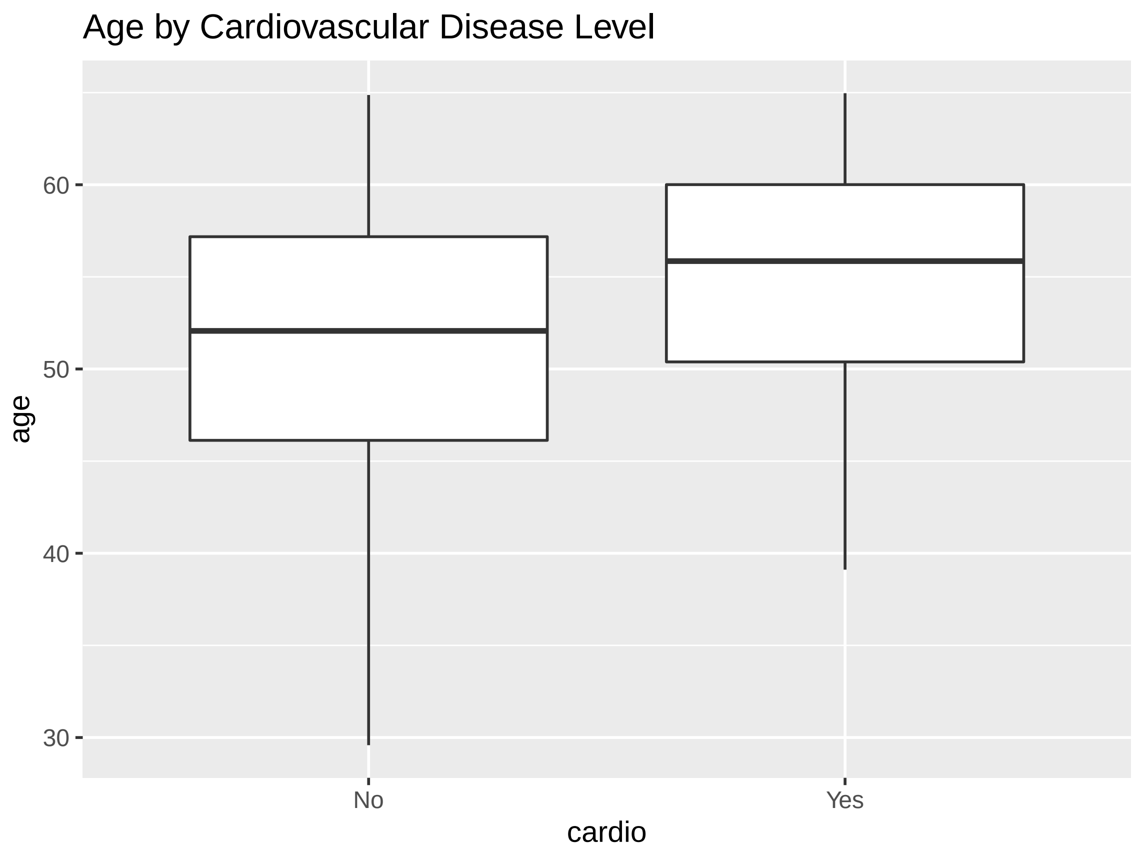
```
summary(cardio.data)
```

```
##       age           gender          height          weight      
##  Min.   :29.58   Female:44932   Min.   : 55.0   Min.   : 11.00  
##  1st Qu.:48.37   Male  :24053   1st Qu.:159.0   1st Qu.: 65.00  
##  Median :53.98                  Median :165.0   Median : 72.00  
##  Mean   :53.33                  Mean   :164.4   Mean   : 74.12  
##  3rd Qu.:58.42                  3rd Qu.:170.0   3rd Qu.: 82.00  
##  Max.   :64.97                  Max.   :250.0   Max.   :200.00  
##      ap_hi           ap_lo                   cholesterol   
##  Min.   :  7.0   Min.   :  1.00   Normal           :51747  
##  1st Qu.:120.0   1st Qu.: 80.00   Above_Normal     : 9339  
##  Median :120.0   Median : 80.00   Well_Above_Normal: 7899  
##  Mean   :126.3   Mean   : 81.35                            
##  3rd Qu.:140.0   3rd Qu.: 90.00                            
##  Max.   :240.0   Max.   :190.00                            
##                  gluc       smoke        alco       active      cardio    
##  Normal           :58650   No :62924   No :65288   No :13571   No :34844  
##  Above_Normal     : 5088   Yes: 6061   Yes: 3697   Yes:55414   Yes:34141  
##  Well_Above_Normal: 5247                                                  
##                                                                           
##                                                                           
## 
```

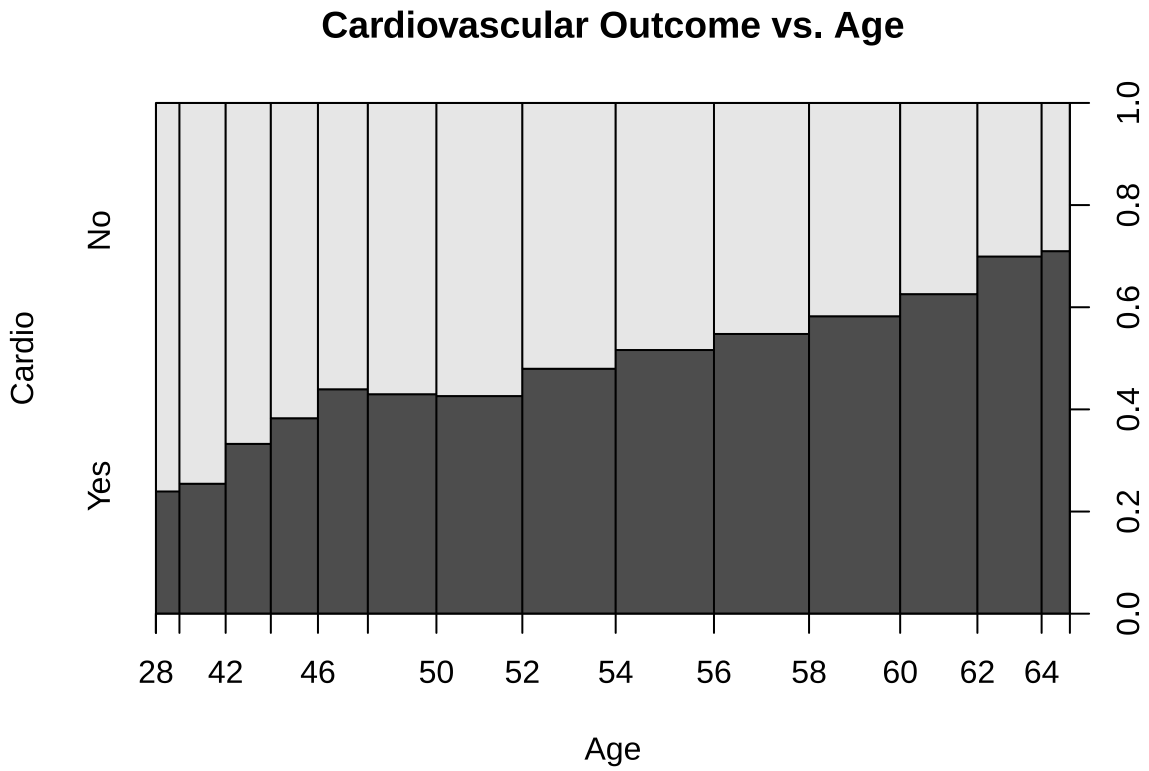
Visualizations

Cardiovascular Disease Outcome by Age

```
ggplot(cardio.data, aes(x=cardio, y=age)) +
  geom_boxplot() +
  ggtitle("Age by Cardiovascular Disease Level")
```

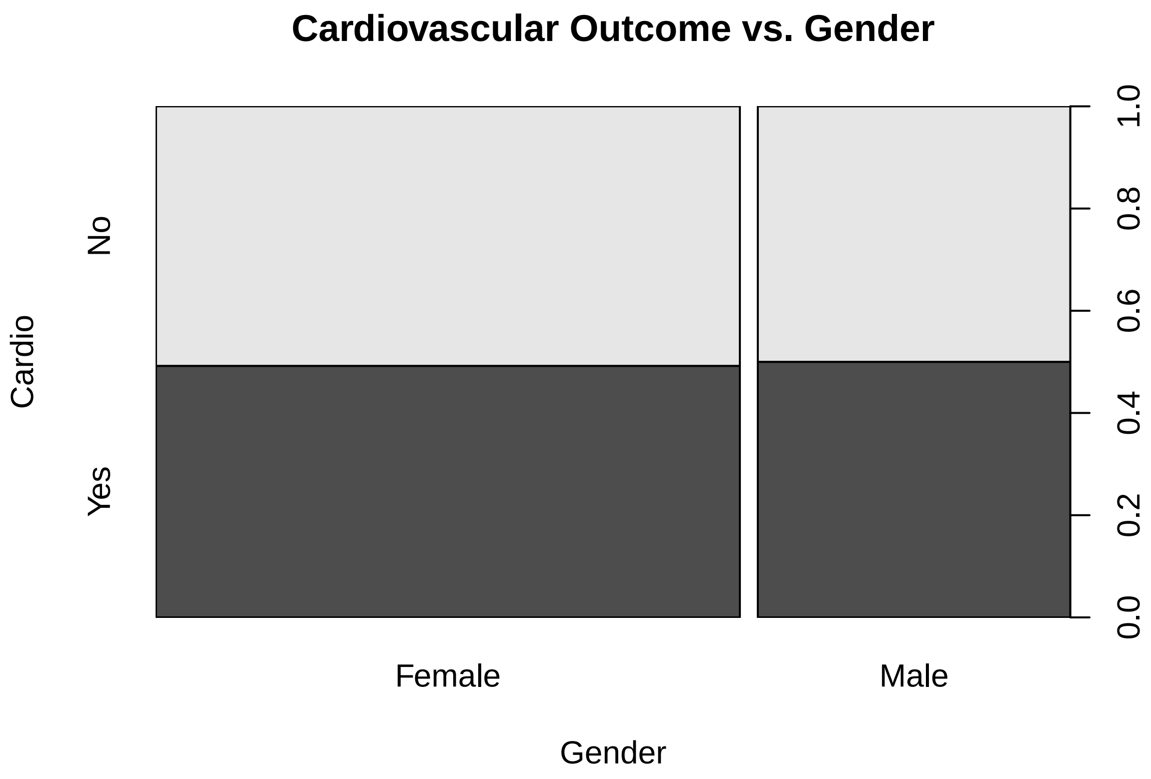


```
plot(cardio.data$cardio ~ cardio.data$age, xlab = "Age",ylab = "Cardio", main = "Cardiovascular Outcome
```


Cardiovascular Outcome vs. Age
Cardio
No
Yes
1.0
0.8
0.6
0.4
0.2
0.0
28
42
46
50
52
54
56
58
60
62
64
Age

Cardiovascular Disease Outcome by Gender

```
plot(cardio.data$cardio ~ cardio.data$gender, xlab = "Gender",ylab = "Cardio", main = "Cardiovascular Ou
```

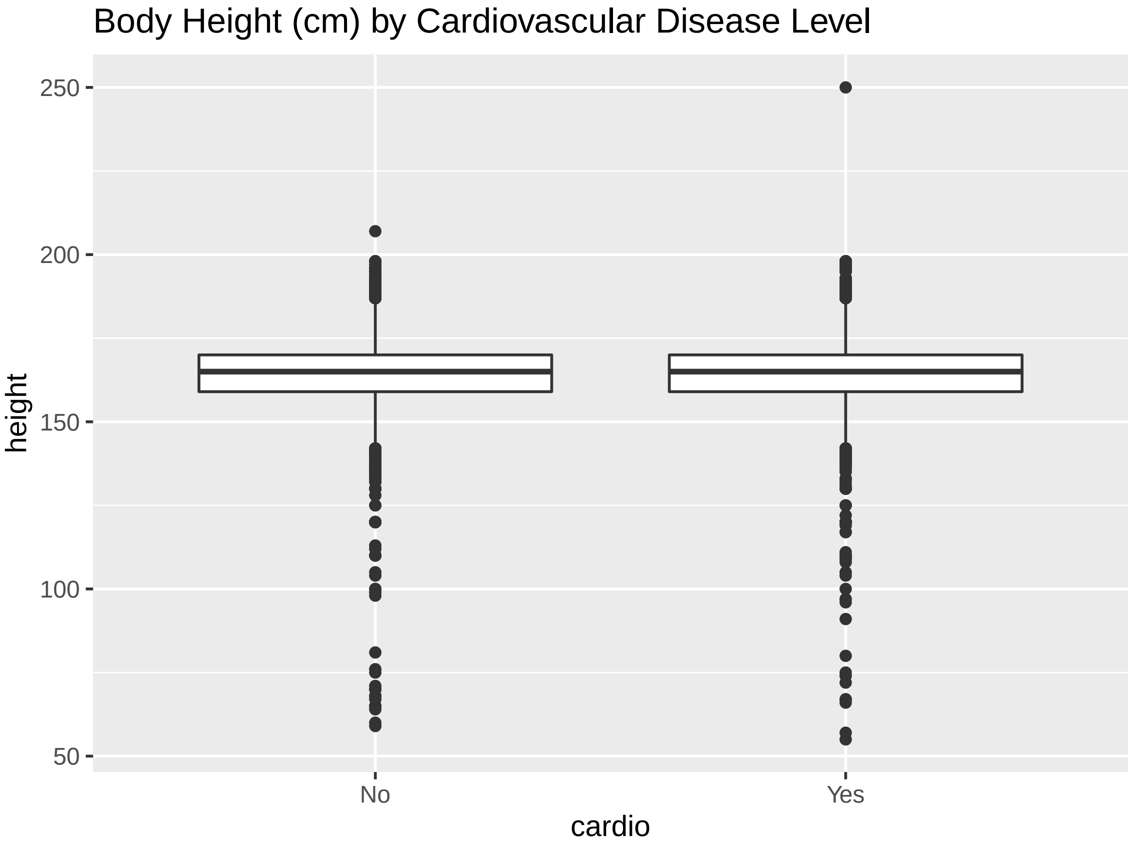


```
cardio.data %>%
  group_by(gender) %>%
  count(cardio)
```

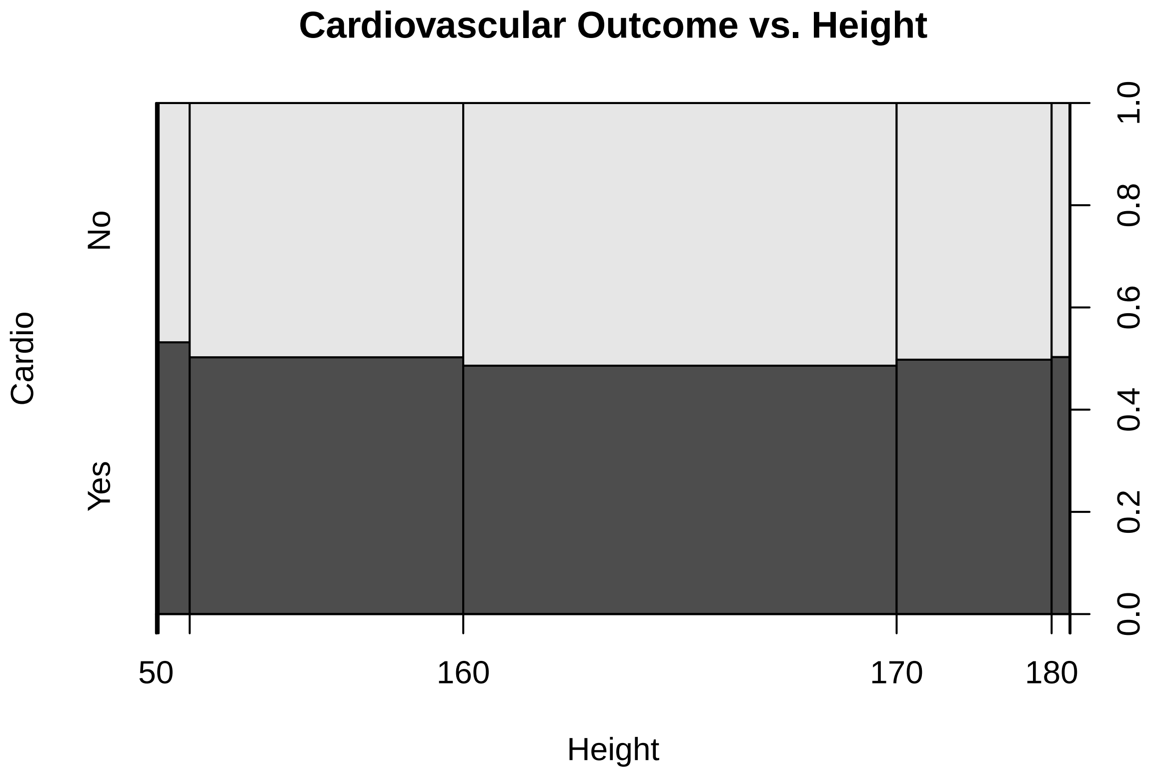
```
## # A tibble: 4 x 3
## # Groups:   gender [2]
##   gender cardio     n
##   <fct>  <fct>  <int>
## 1 Female No     22819
## 2 Female Yes    22113
## 3 Male   No     12025
## 4 Male   Yes    12028
```

Cardiovascular Disease Outcome by Height

```
ggplot(cardio.data, aes(x=cardio, y=height)) +
  geom_boxplot() +
  ggtitle("Body Height (cm) by Cardiovascular Disease Level")
```

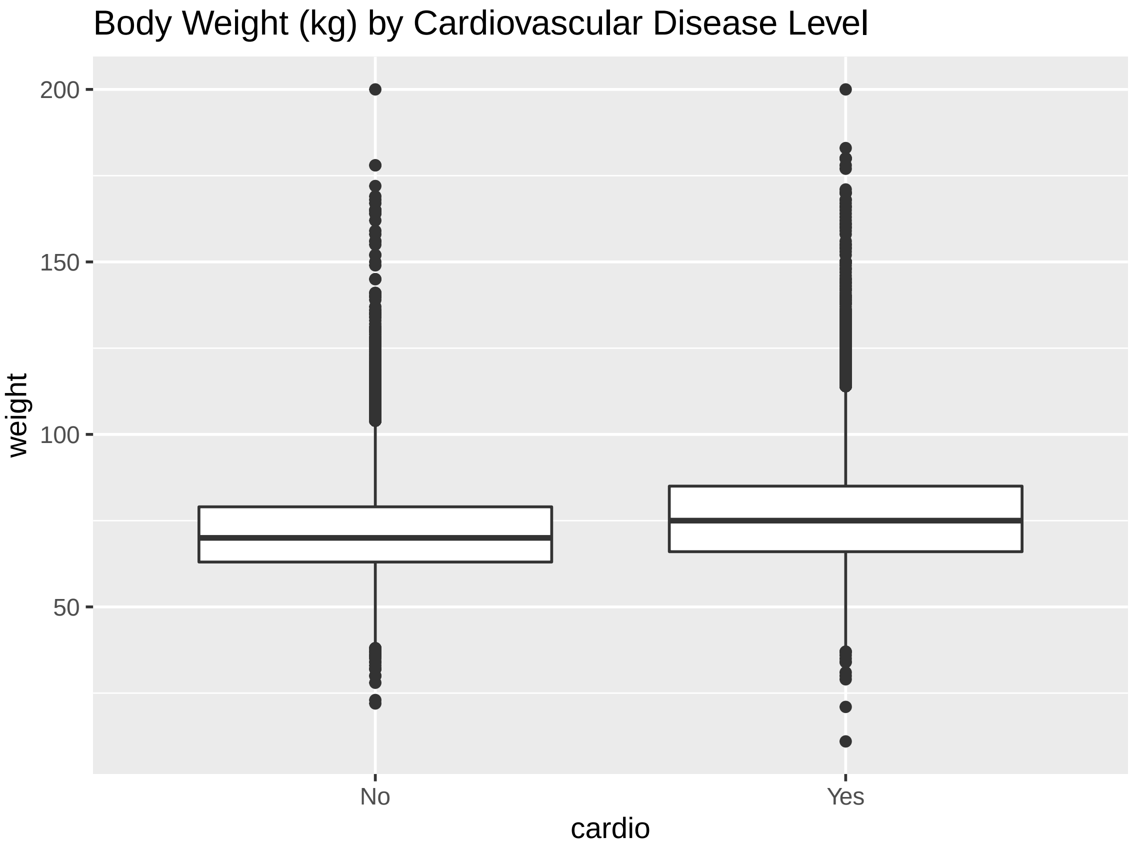


```
plot(cardio.data$cardio ~ cardio.data$height, xlab = "Height",ylab = "Cardio", main = "Cardiovascular Ou
```

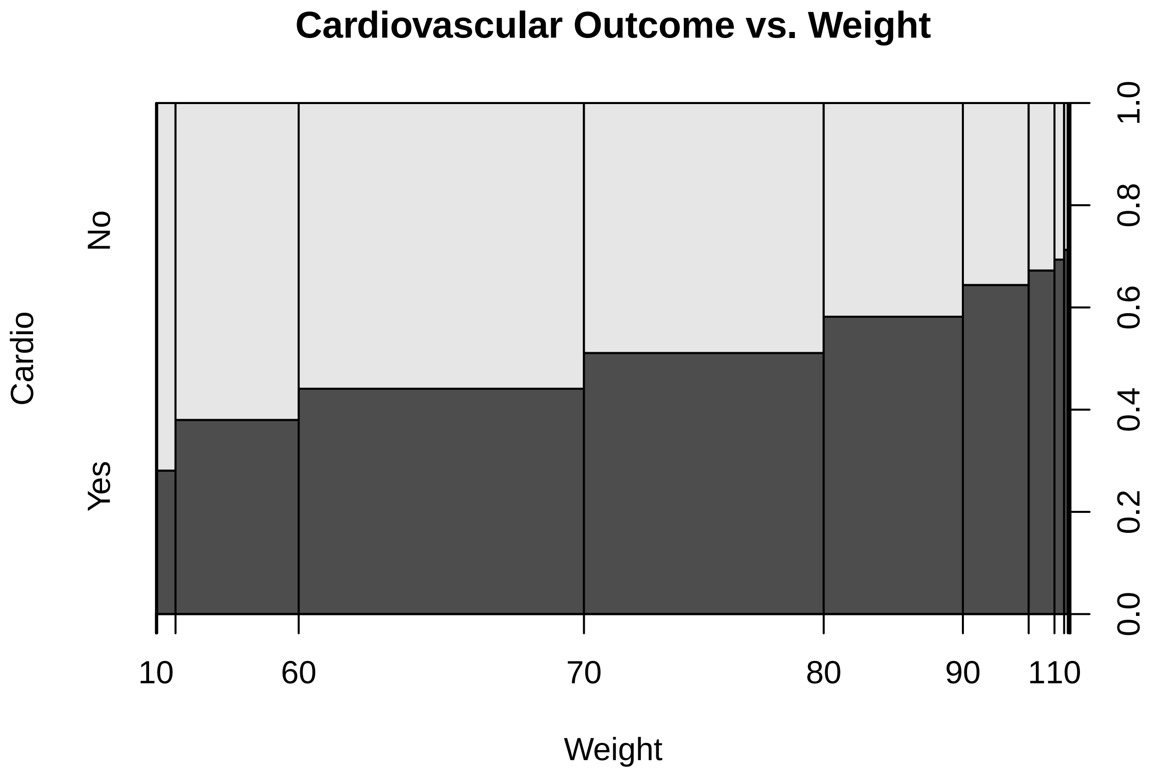

Cardiovascular Outcome vs. Height
Cardio
No
Yes
Height
50
160
170
180
0.0
0.2
0.4
0.6
0.8
1.0

Cardiovascular Disease Outcome by Weight

```
ggplot(cardio.data, aes(x=cardio, y=weight)) +
  geom_boxplot() +
  ggtitle("Body Weight (kg) by Cardiovascular Disease Level")
```

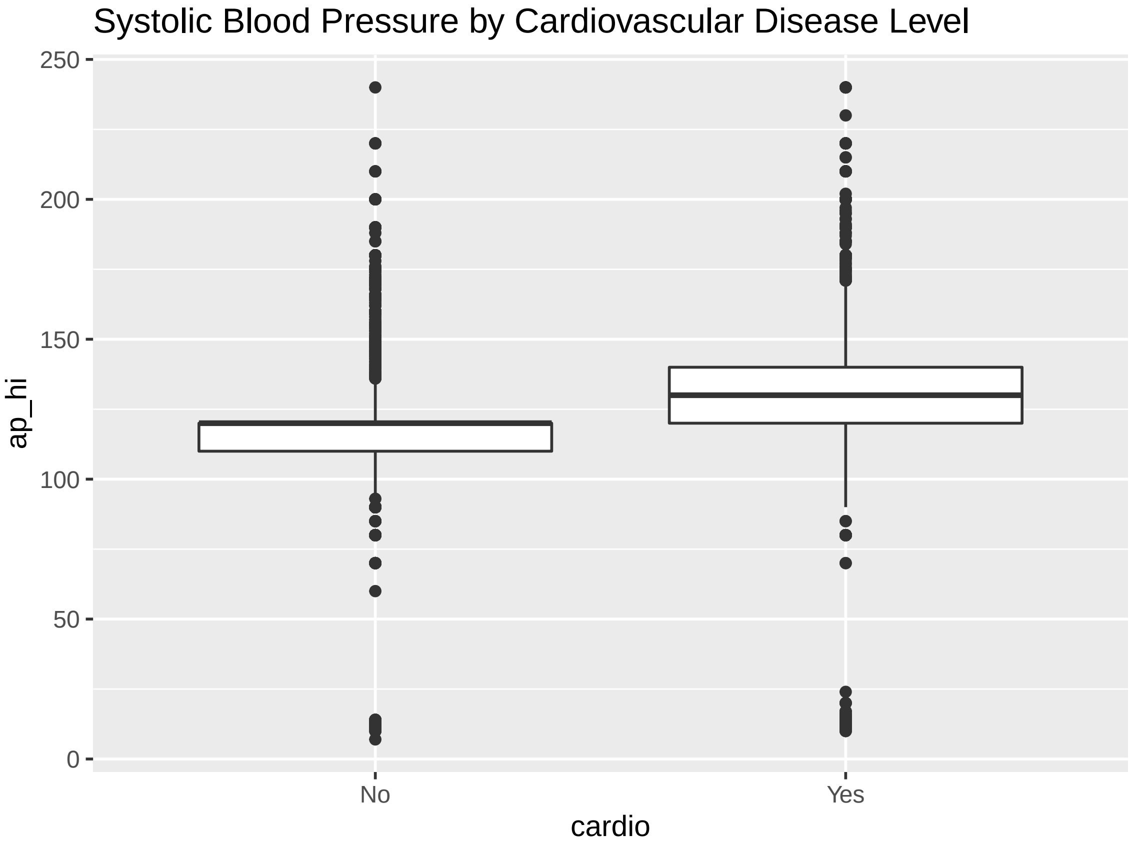


```
plot(cardio.data$cardio ~ cardio.data$weight, xlab = "Weight",ylab = "Cardio", main = "Cardiovascular Ou
```

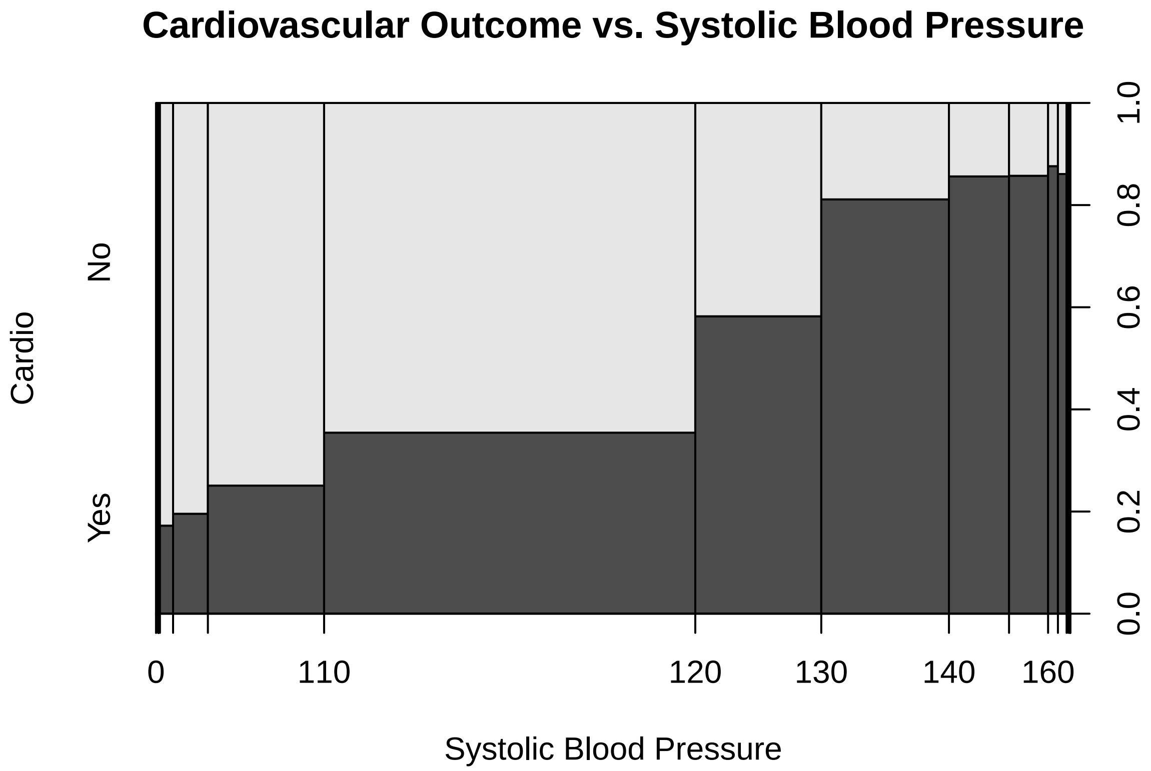

Cardiovascular Outcome vs. Weight
Cardio
No
Yes
1.0
0.8
0.6
0.4
0.2
0.0
10
60
70
80
90
110
Weight

Cardiovascular Disease Outcome by Systolic blood pressure

```
ggplot(cardio.data, aes(x=cardio, y=ap_hi)) +
  geom_boxplot() +
  ggtitle("Systolic Blood Pressure by Cardiovascular Disease Level")
```

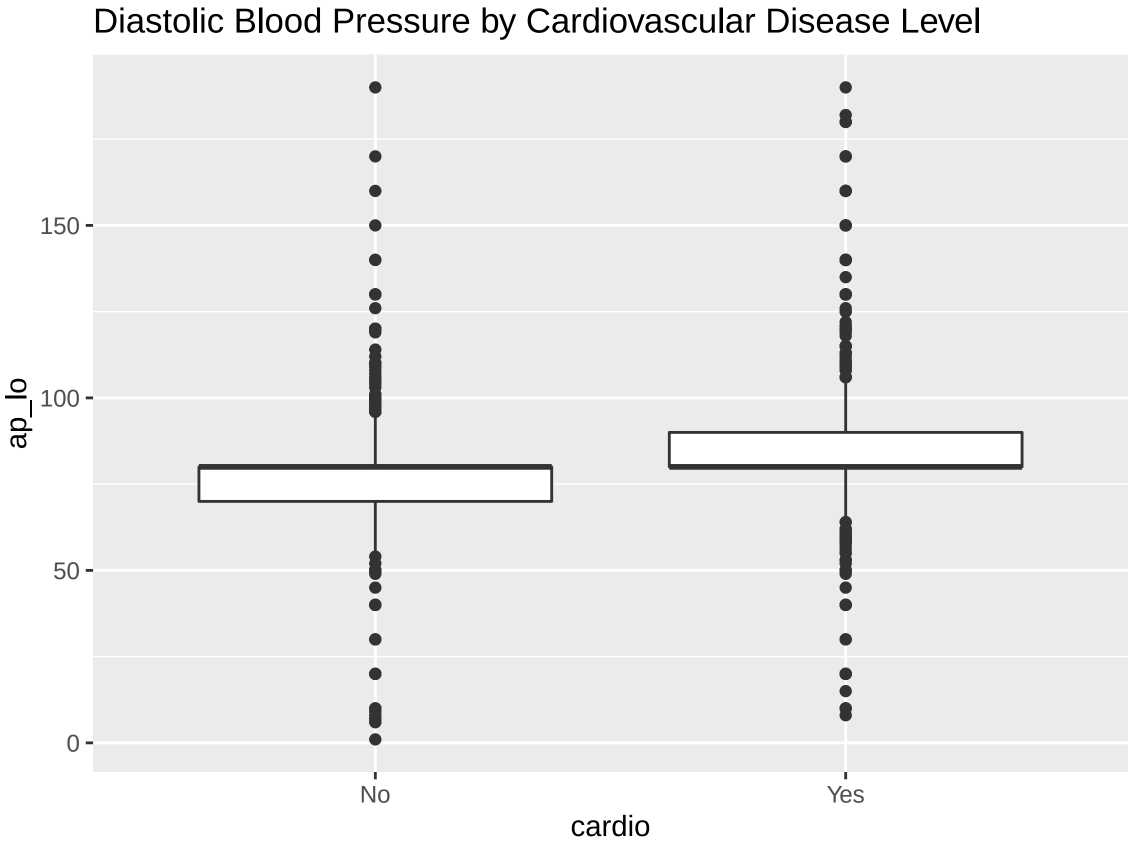


```
plot(cardio.data$cardio ~ cardio.data$ap_hi, xlab = "Systolic Blood Pressure",ylab = "Cardio", main = "
```

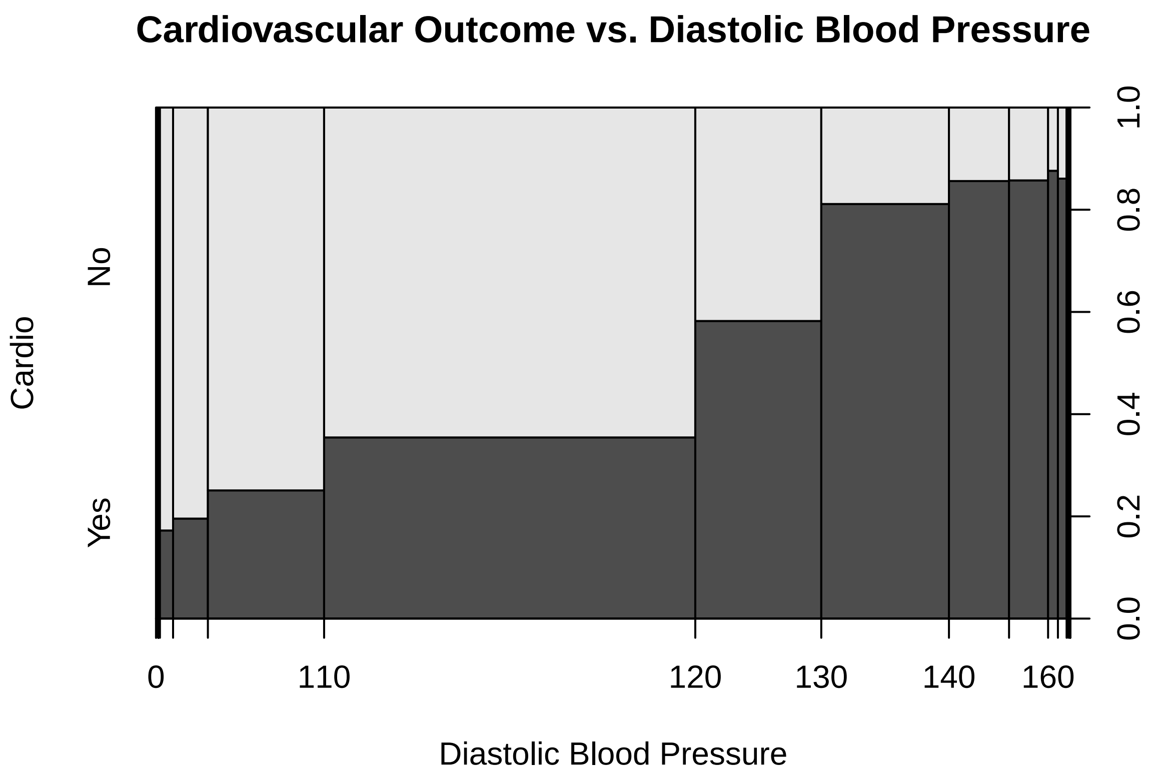

Cardiovascular Outcome vs. Systolic Blood Pressure
Cardio
No
Yes
1.0
0.8
0.6
0.4
0.2
0.0
0
110
120
130
140
160
Systolic Blood Pressure

Cardiovascular Disease Outcome by Diastolic blood pressure

```
ggplot(cardio.data, aes(x=cardio, y=ap_lo)) +
  geom_boxplot() +
  ggtitle("Diastolic Blood Pressure by Cardiovascular Disease Level")
```

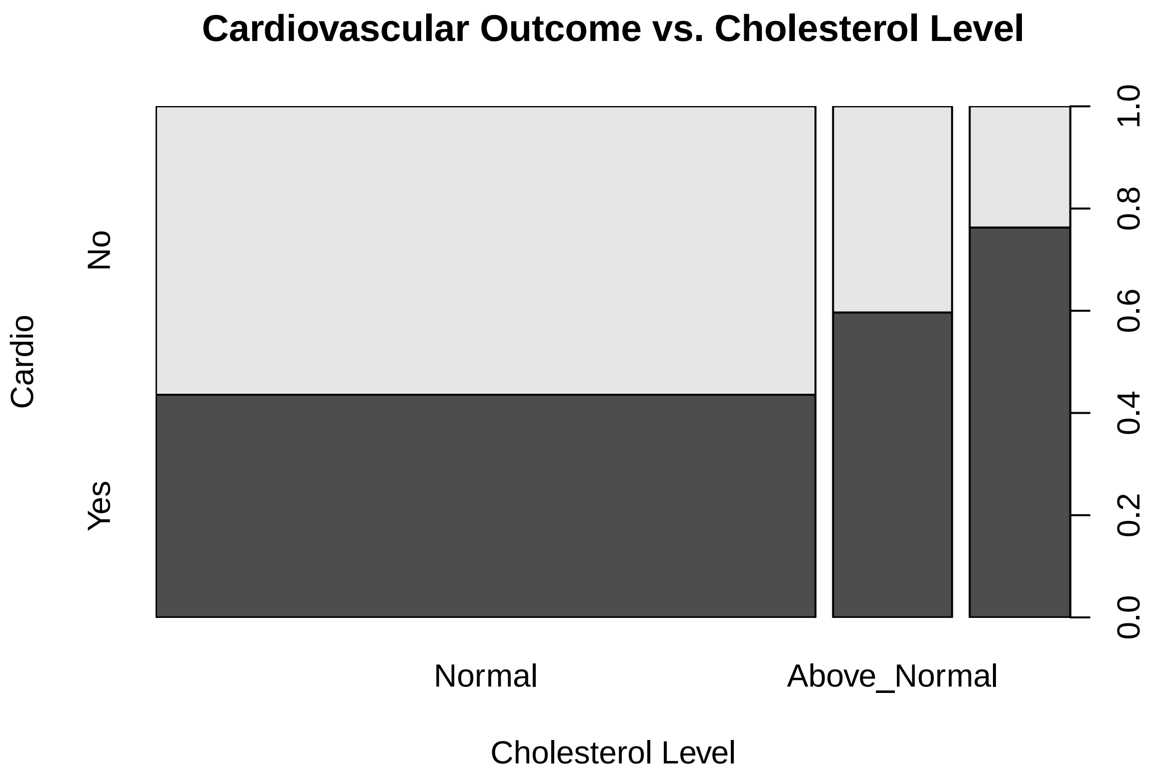


```
plot(cardio.data$cardio ~ cardio.data$ap_hi, xlab = "Diastolic Blood Pressure",ylab = "Cardio", main =
```


Cardiovascular Outcome vs. Diastolic Blood Pressure
Cardio
No
Yes
1.0
0.8
0.6
0.4
0.2
0.0
0
110
120
130
140
160
Diastolic Blood Pressure

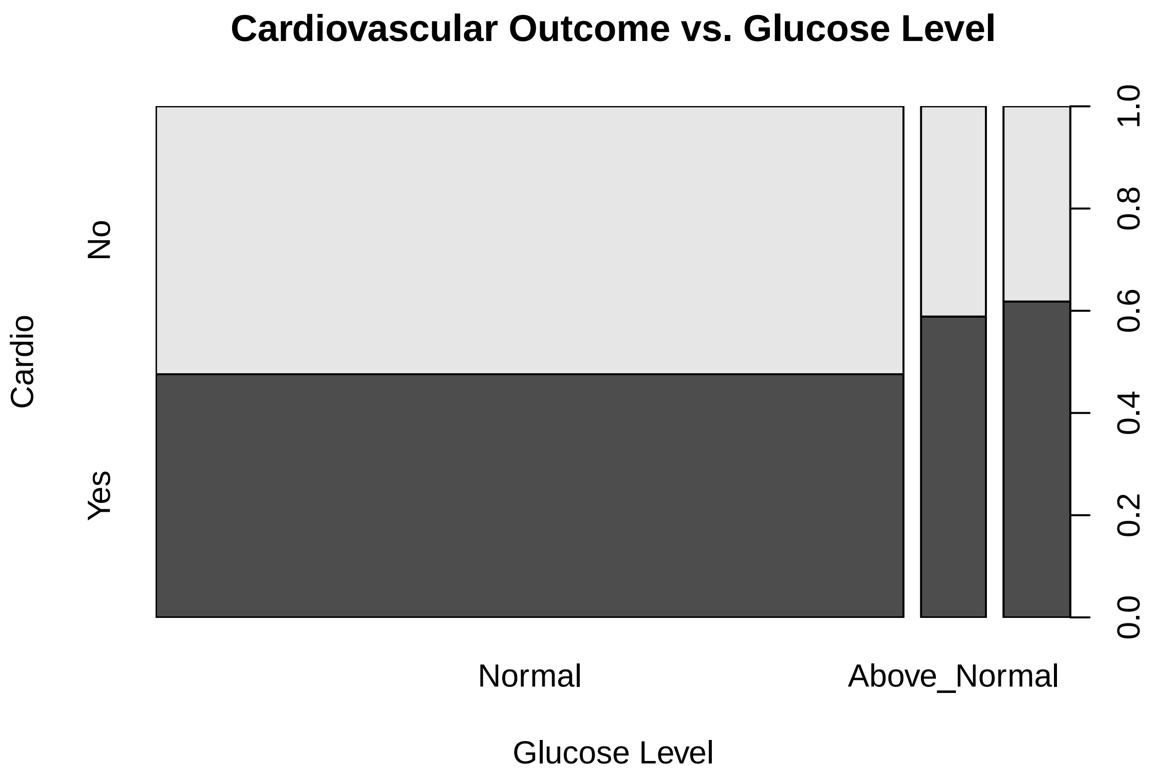
Cardiovascular Disease Outcome by Cholesterol Level

```
plot(cardio.data$cardio ~ cardio.data$cholesterol, xlab = "Cholesterol Level",ylab = "Cardio", main = "
```



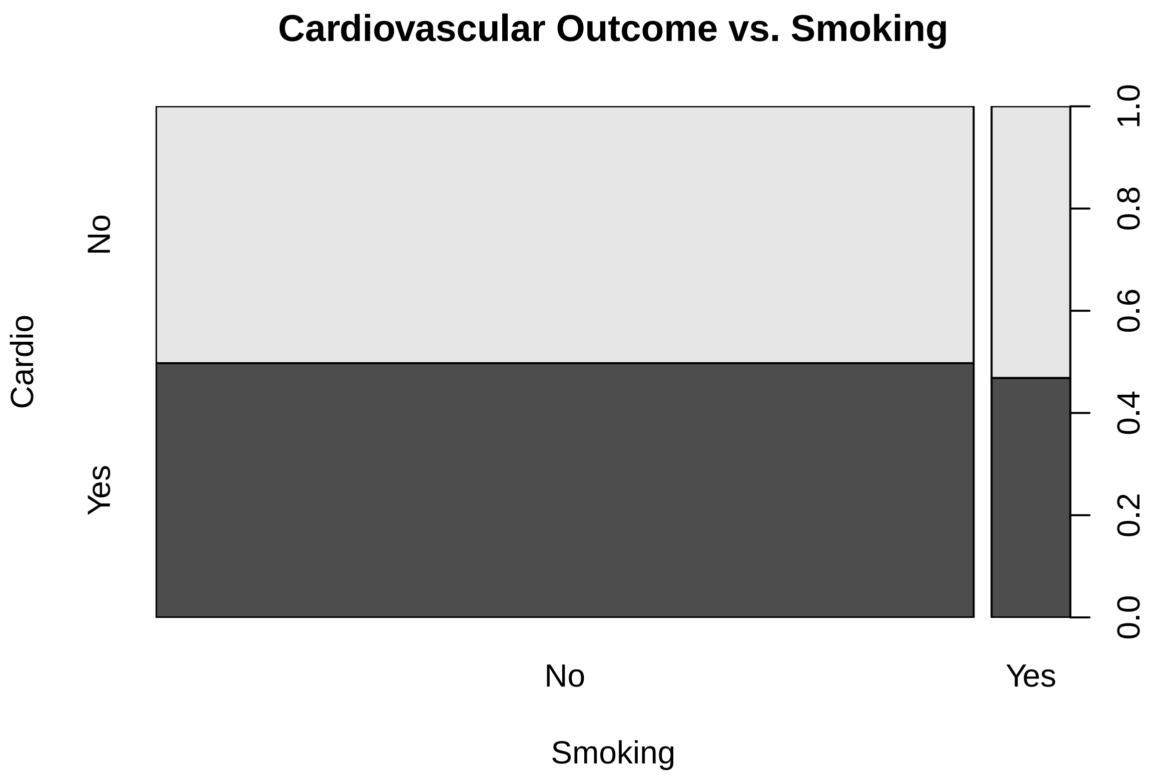
Cardiovascular Disease Outcome by Glucose Level

```r
plot(cardio.data$cardio ~ cardio.data$gluc, xlab = "Glucose Level",ylab = "Cardio", main = "Cardiovascul
```



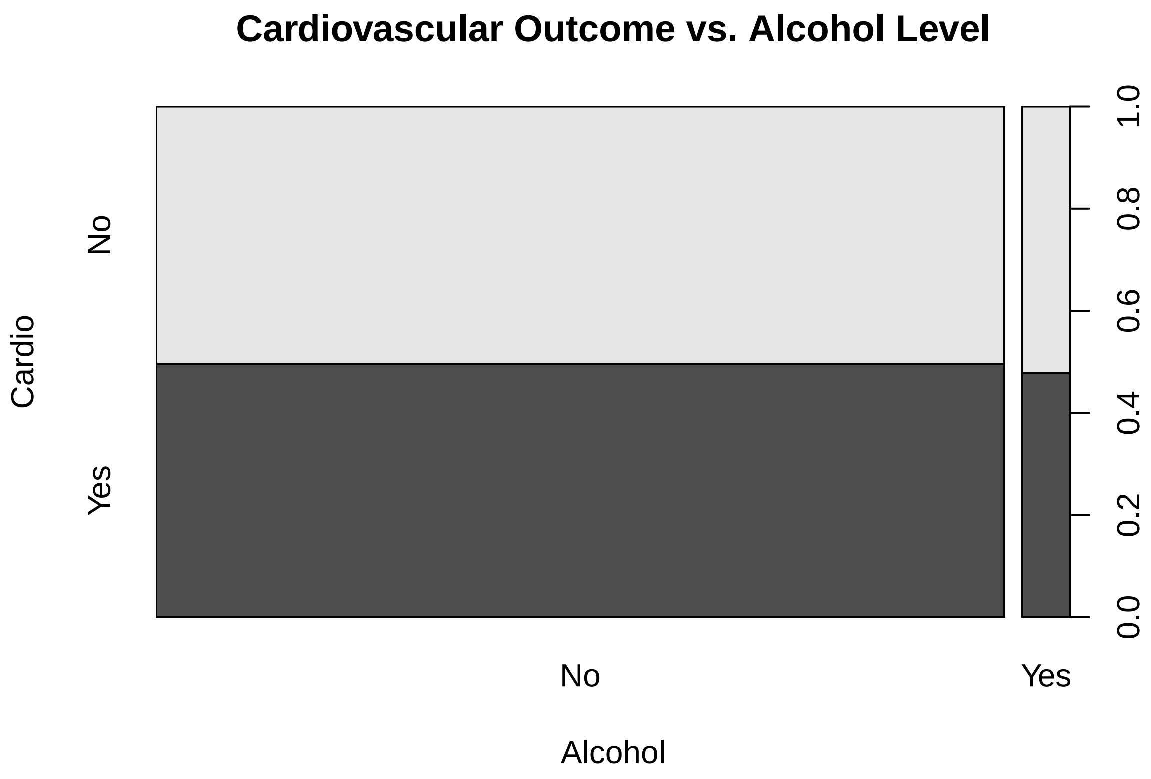
Cardiovascular Disease Outcome by Smoking Level

```
plot(cardio.data$cardio ~ cardio.data$smoke, xlab = "Smoking",ylab = "Cardio", main = "Cardiovascular Ou
```



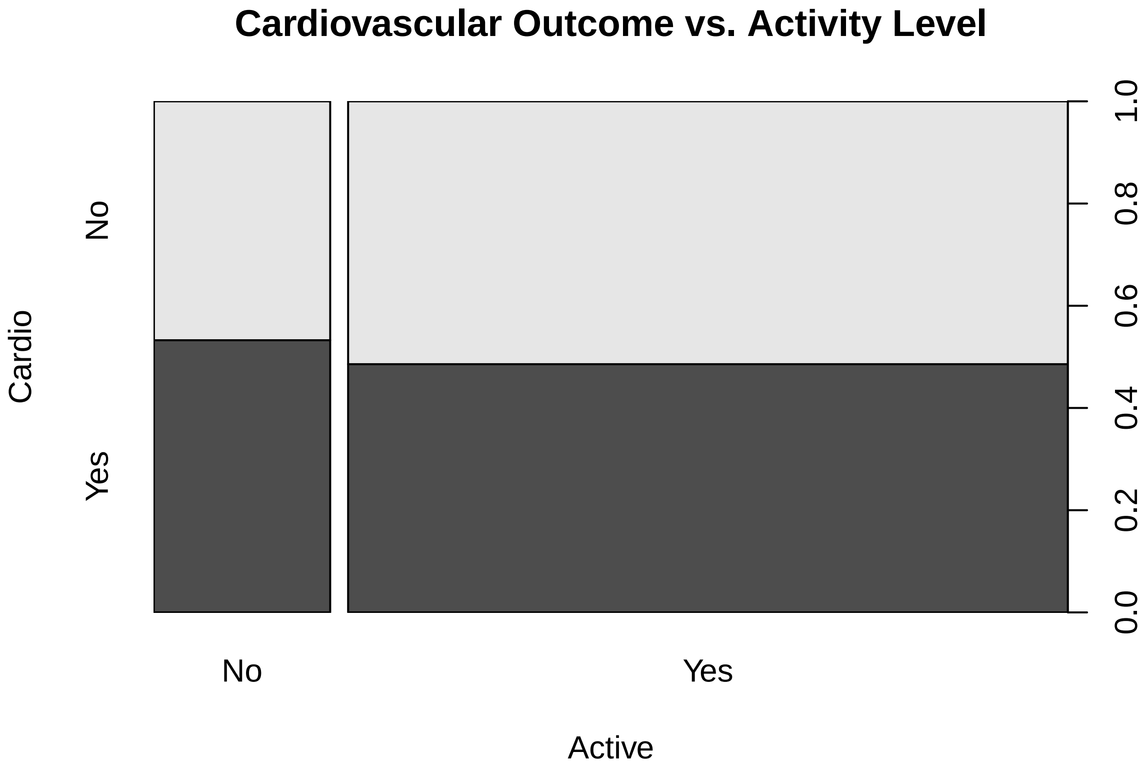
Cardiovascular Disease Outcome by Alcohol Level

```
plot(cardio.data$cardio ~ cardio.data$alco, xlab = "Alcohol",ylab = "Cardio", main = "Cardiovascular Ou
```



Cardiovascular Disease Outcome by Activity Level

```r
plot(cardio.data$cardio ~ cardio.data$active, xlab = "Active",ylab = "Cardio", main = "Cardiovascular Ou
```


Cardiovascular Outcome vs. Activity Level
Cardio
No
Yes
No
Yes
Active