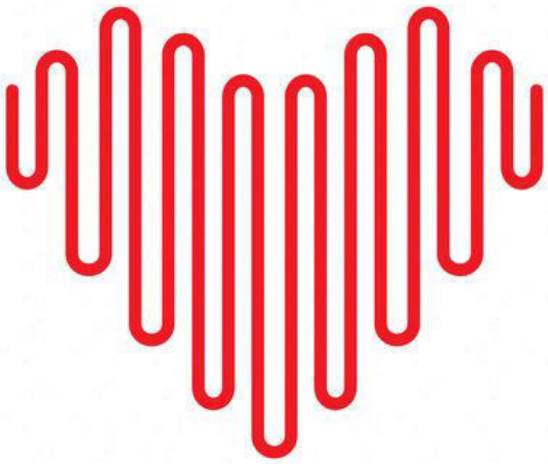
Analysis of Cardiovascular Diseases Dataset



Srijan Mallick

Dipan Banik

Oishik Dasgupta

Primary Objective:

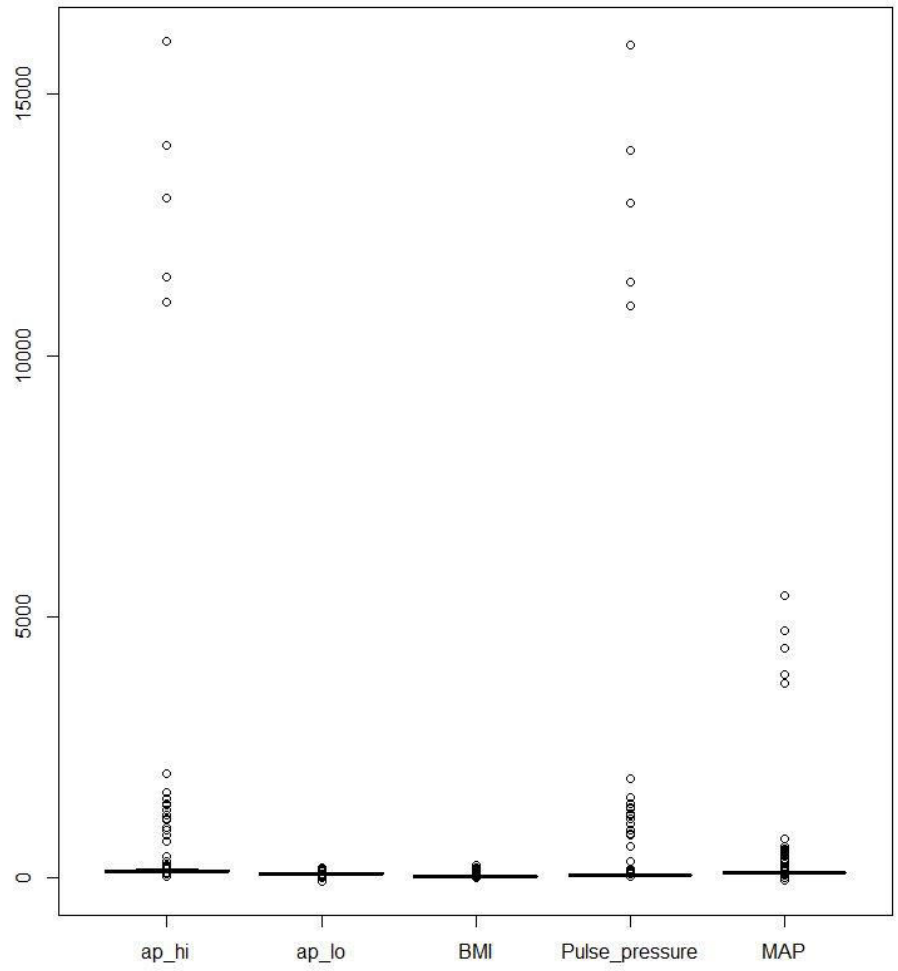
To determine the factors which contribute to the presence of cardiovasculardiseases.

In order to reach to a definitive answer to our objective, we need to answer some sub-questions first, namely,

- Are there any serious outliers in the dataset that might negatively influence our conclusions ?
- What is the measure of linear association between the variables in the dataset ?
- Do habits like smoking/drinking have any effect on the presence of cardiovascular diseases ?
- Does being physically active reduce the risk of heart diseases ?
- Does a person's height/weight make him/her more prone to heart diseases ?
- How is a person's systolic/diastolic blood pressure associated with the presence of cardiovascular diseases ?
- Are old people at a greater risk than the younger ones ?
- How is a person's cholesterol/glucose level associated with the presence of cardiovascular diseases ?
- Does obesity increase the risk of heart disease ?

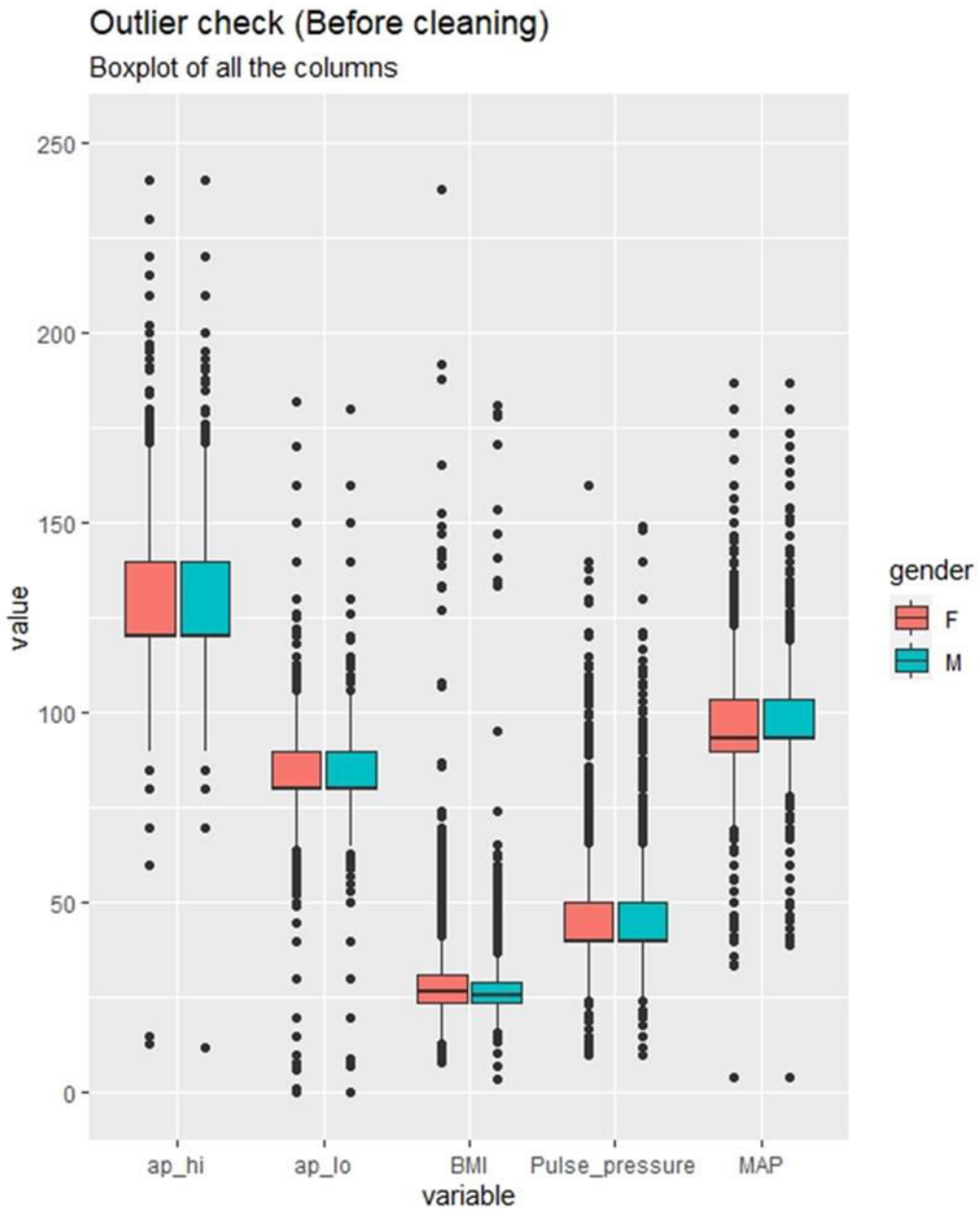
- **Are there any serious outliers in the dataset that might negatively influence our conclusion ?**

We begin our analysis by plotting a box-plot of some variables in our dataset which we felt might need some cleaning.



Plot - 1

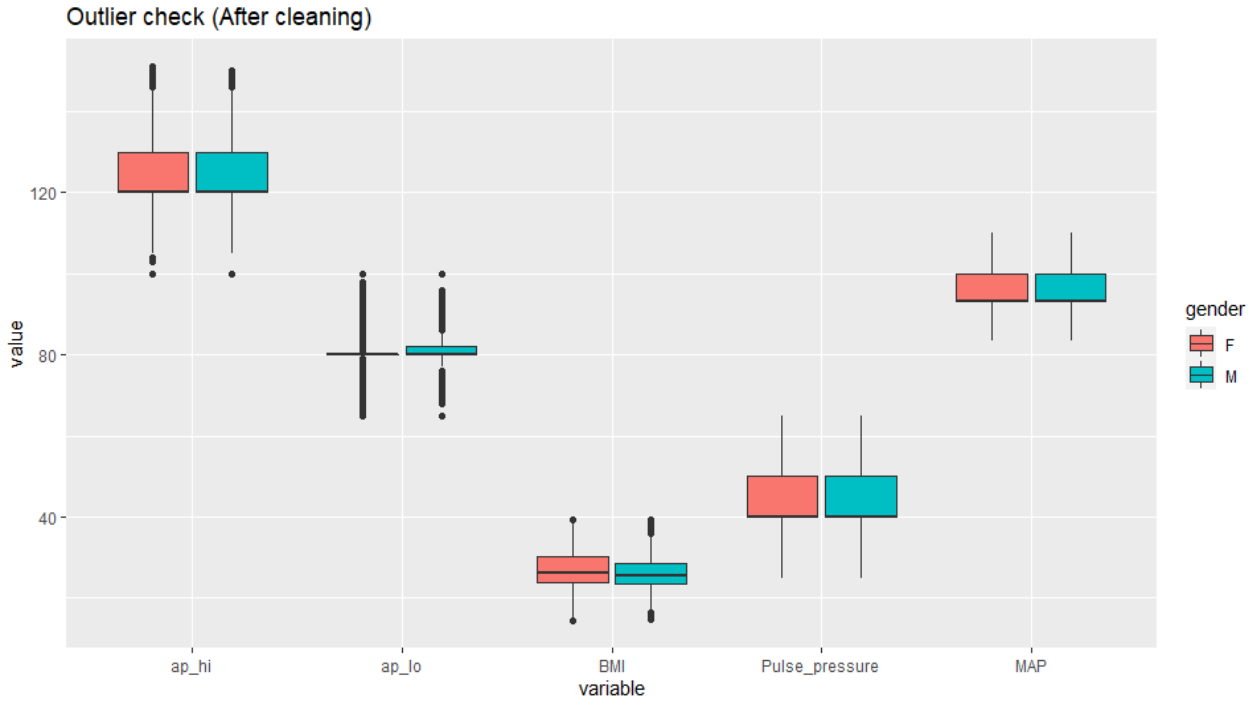
Before removing outliers, the boxplot of certain key factors that might dictate the presence of cardiovascular diseases looks inconclusive at the first glance due to the presence of some extreme and absurd values.



Plot-2

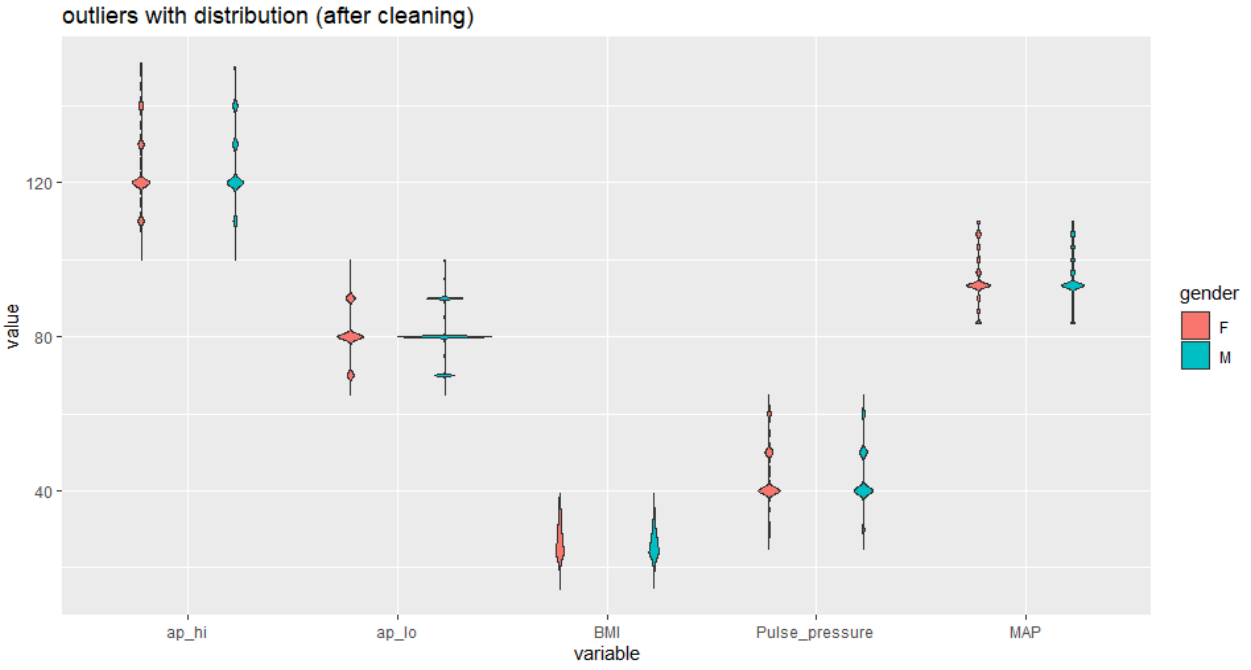
We manually remove some absurd values, based on observation, and further group the variables by the categorical variable ‘gender’ in order to get a clearer picture of the distribution of outliers in our data.

Now we formally remove the outliers by the quantile formula.



Plot - 3

We also plot the violin plot to get a better sense of the distributions of the variables we have plotted previously.

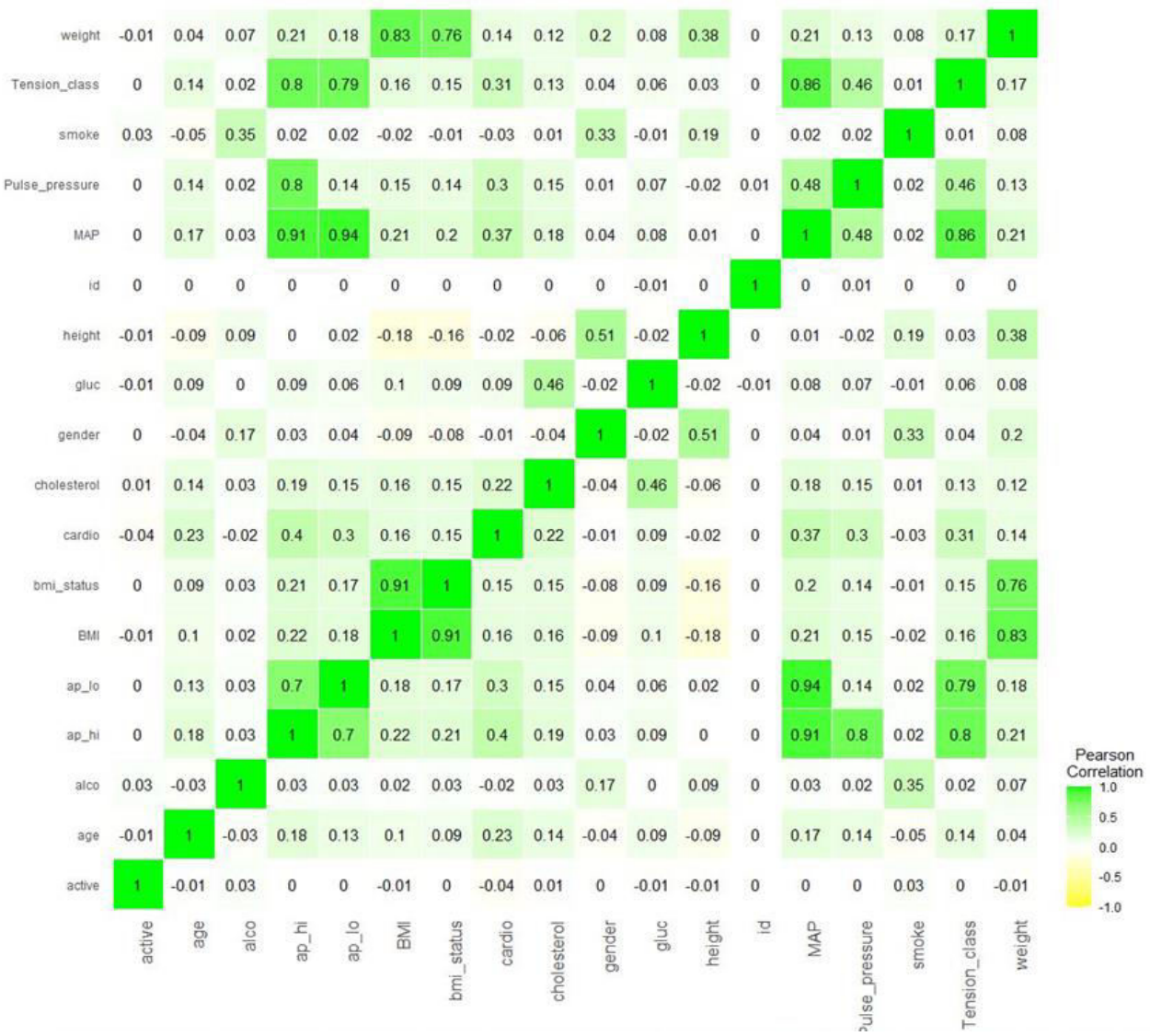


Plot - 4

We have created some additional variables from the existing ones, in order to reach our primary objective, namely,

i. Pulse Pressure = (ap_hi – ap_lo)
ii. Mean Arterial Pressure = ap_lo+(Pulsepressure/3)
iii. BMI = weight/(height)^2 (weight in kg, height in m)
iv. Tension class
v. BMI status

- **What is the measure of linear association between the variables in the dataset ?**

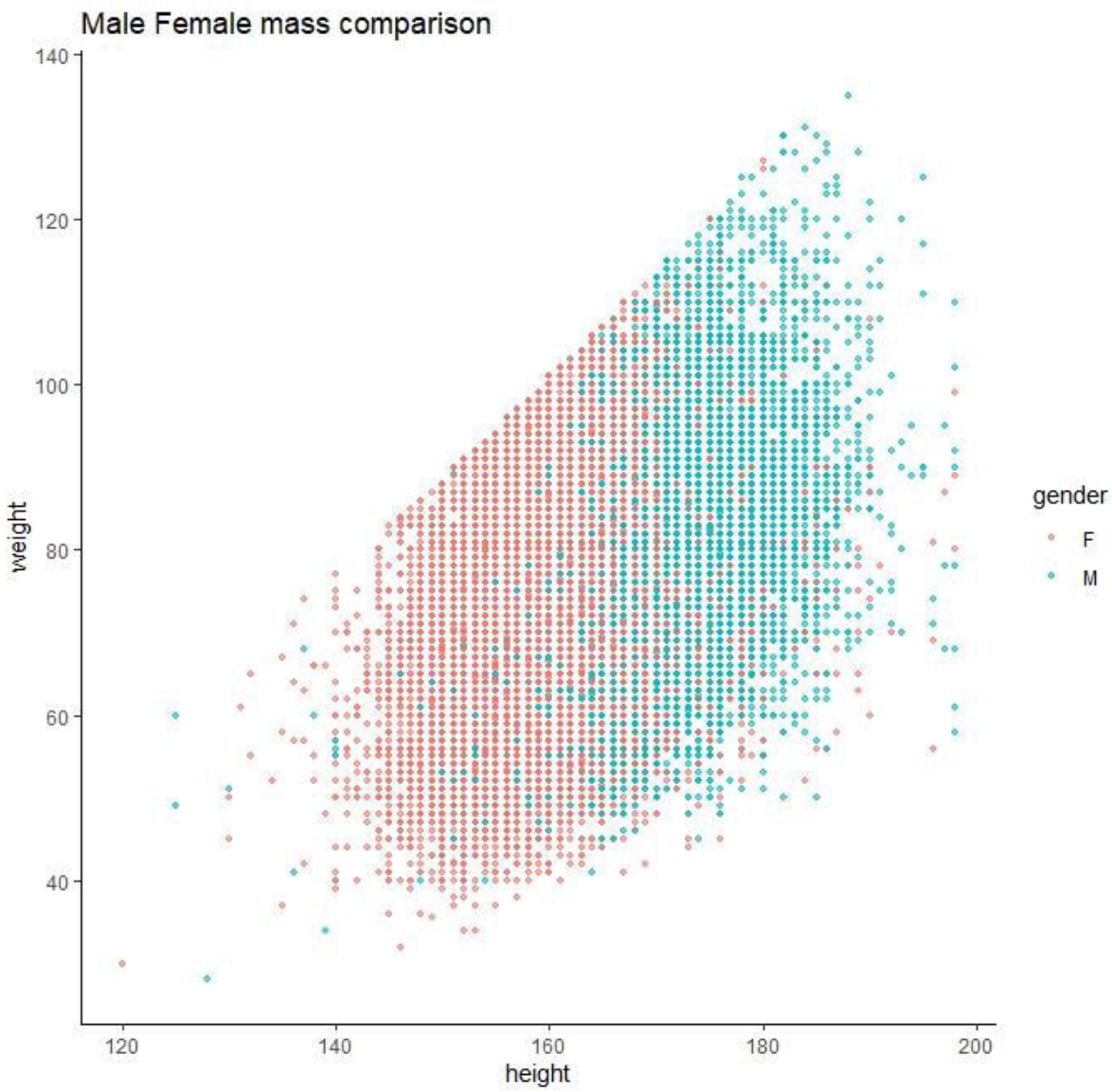


Plot - 5

Looking at the row corresponding to the 'cardio' variable, we observe that there is a moderate to strong linear association between the presence of cardiovascular diseases (cardio) and :

i. Age
ii. Systolic blood pressure (ap_hi)
iii. Diastolic blood pressure (ap_lo)
iv. Cholesterol

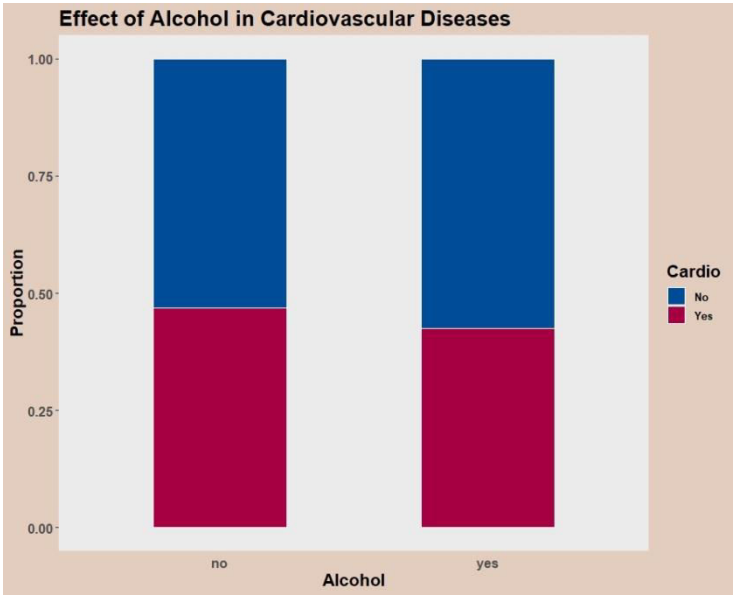
Let us check if the overall mass diversity of the whole data follows the natural trend or not,



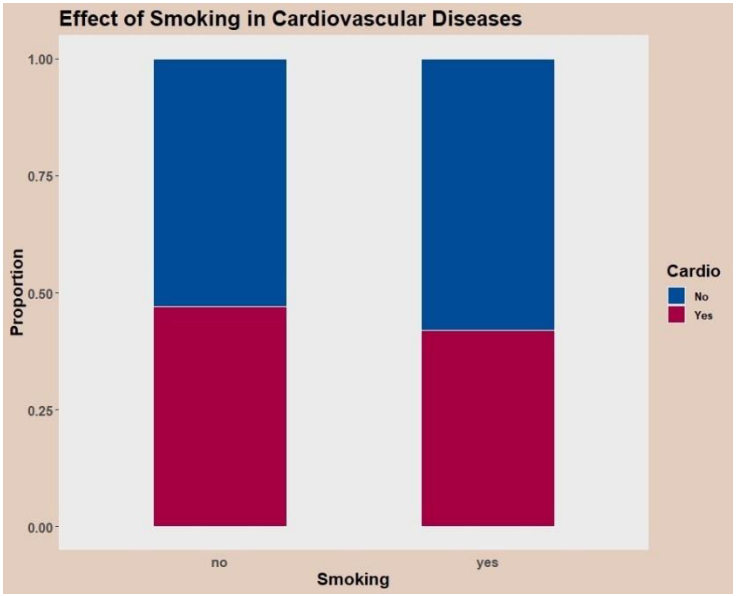
Plot - 6

It is clear for the scatterplot that the distribution of height and weight of the observed sample set follows the natural trend, and is positively correlated, as is also evident from the heatmap.

- **Do habits like smoking/drinking have any effect on the presence of cardiovascular diseases?**



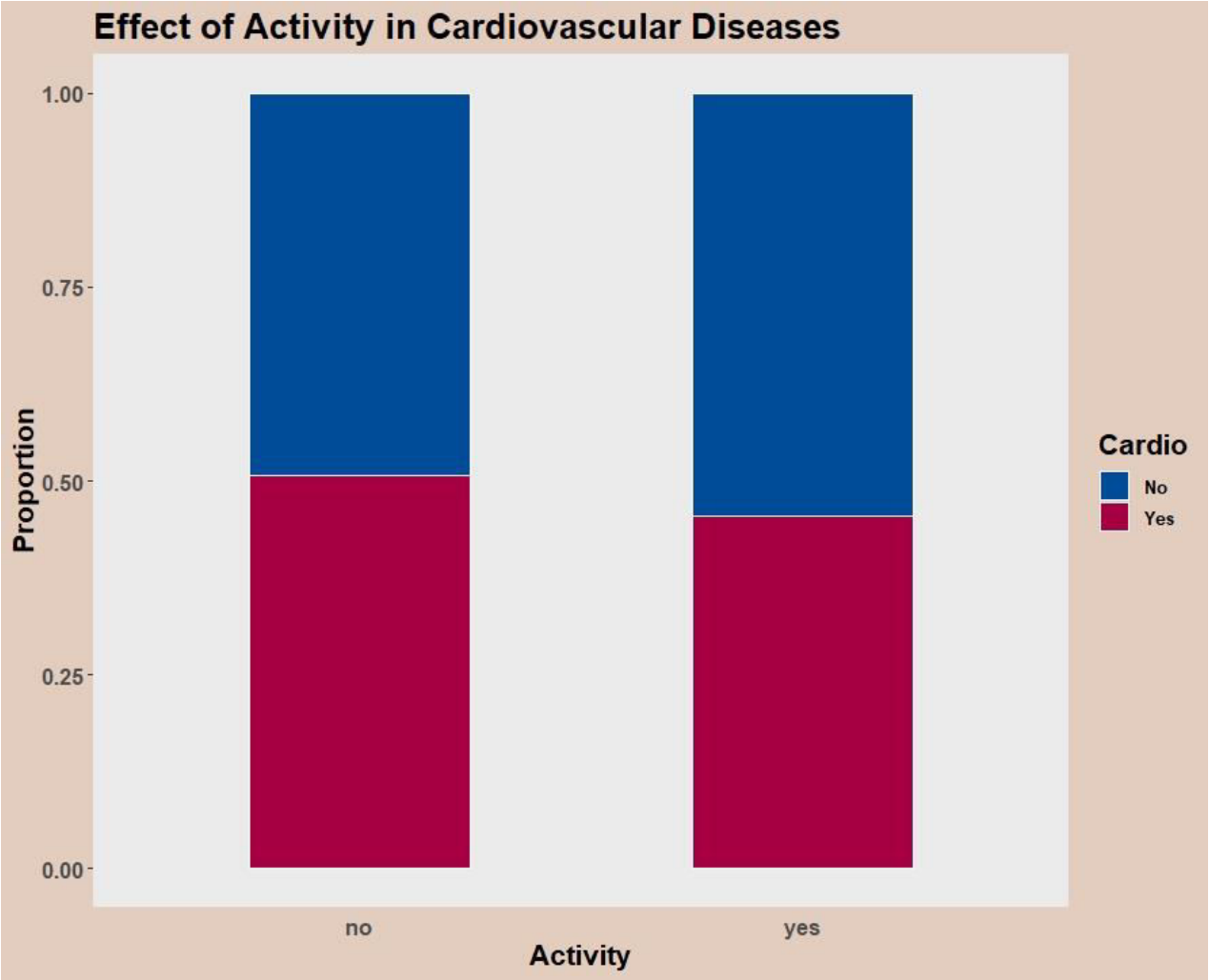
Plot - 7



Plot - 8

Our chosen dataset is highly imbalanced when it comes to smokers/non-smokers and alcohol consumers/non-consumers. Due to the dominance of one group over the other, it is not possible to derive any meaningful observation from the plot.

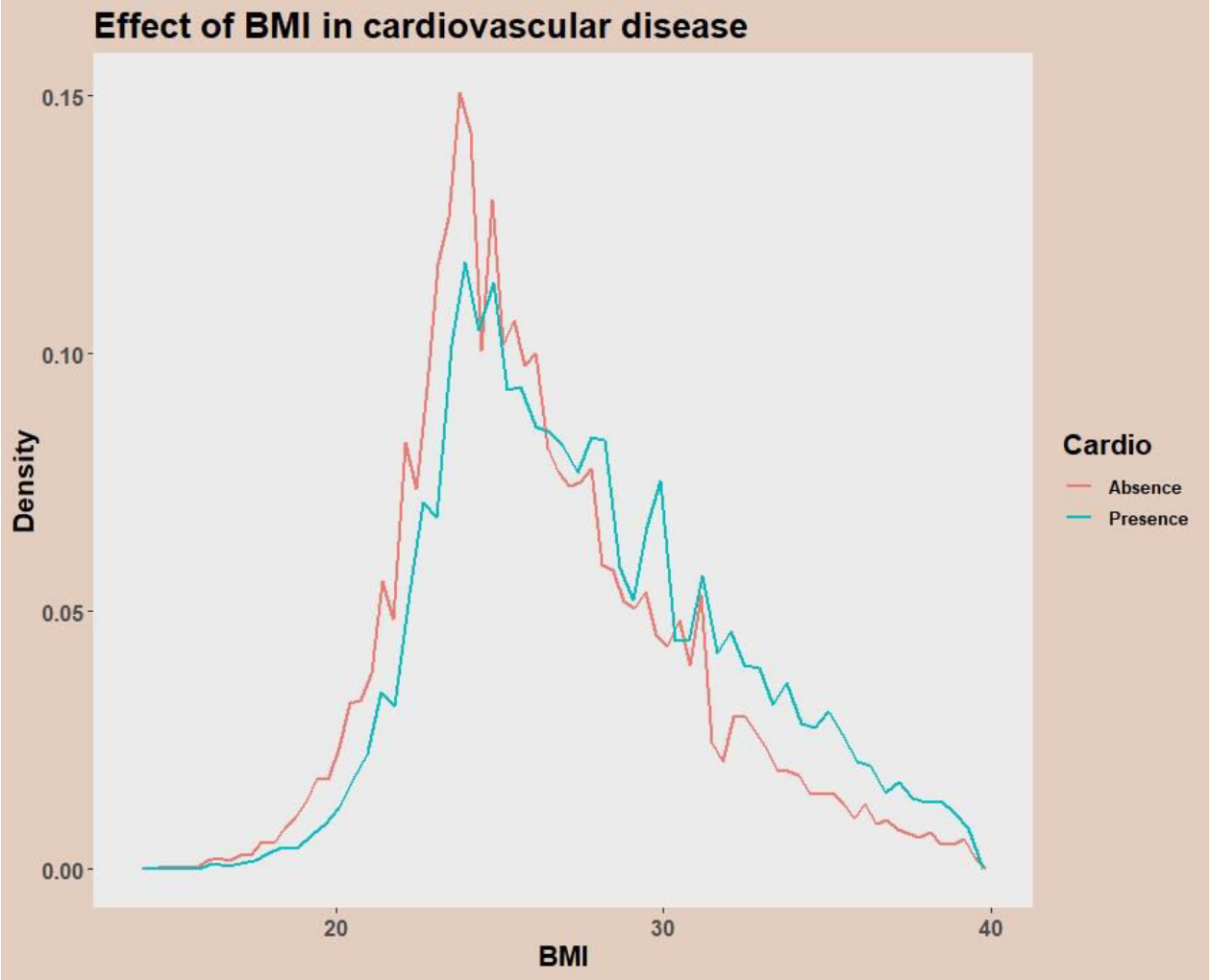
- **Does being physically active reduce the risk of heart diseases ?**



Plot - 9

Observation: People who are physically active have a lesser risk of cardiovascular diseases than those who are not.

- **Does a person's height/weight make him/her more prone to heart diseases?**



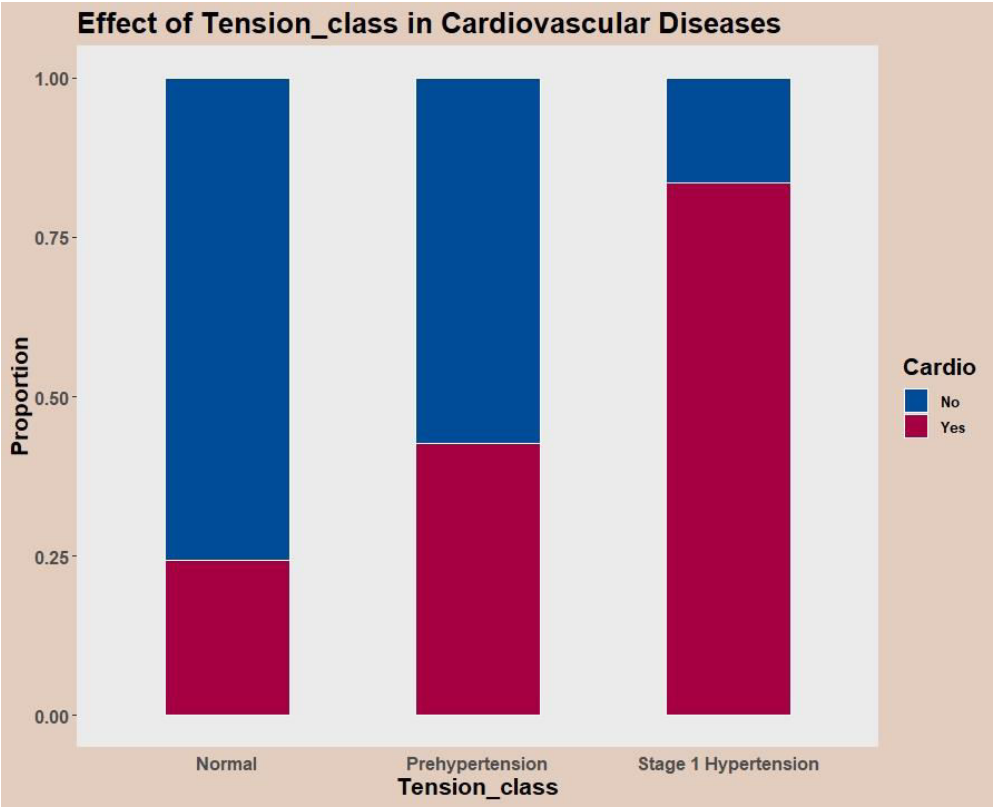
Plot - 10

BMI captures both height and weight so we use it to determine the latter's effect on cardiovascular diseases. From the frequency polygon above, we observe that at lower (<25) BMI values, absence of heart diseases predominates the presence of the same, i.e. over-weight and obese people are at a higher risk of having cardiovascular diseases.

- **How is a person's systolic/diastolic blood pressure associated with the presence of cardiovascular diseases ?**

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

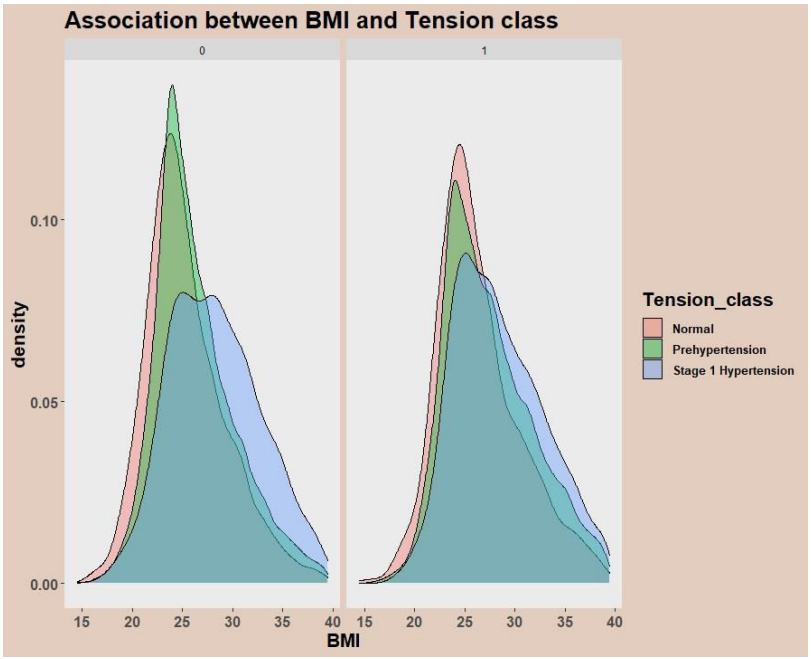
We shall now analyse the effect of tension class on heart diseases, for which we group systolic and diastolic pressure values in our dataset in groups according to the above table.



Plot - 11

Observation: The risk of having cardiovascular disease increases as we move from Normal to Pre-hypertension to Stage 1 Hypertension levels of Tension Class.

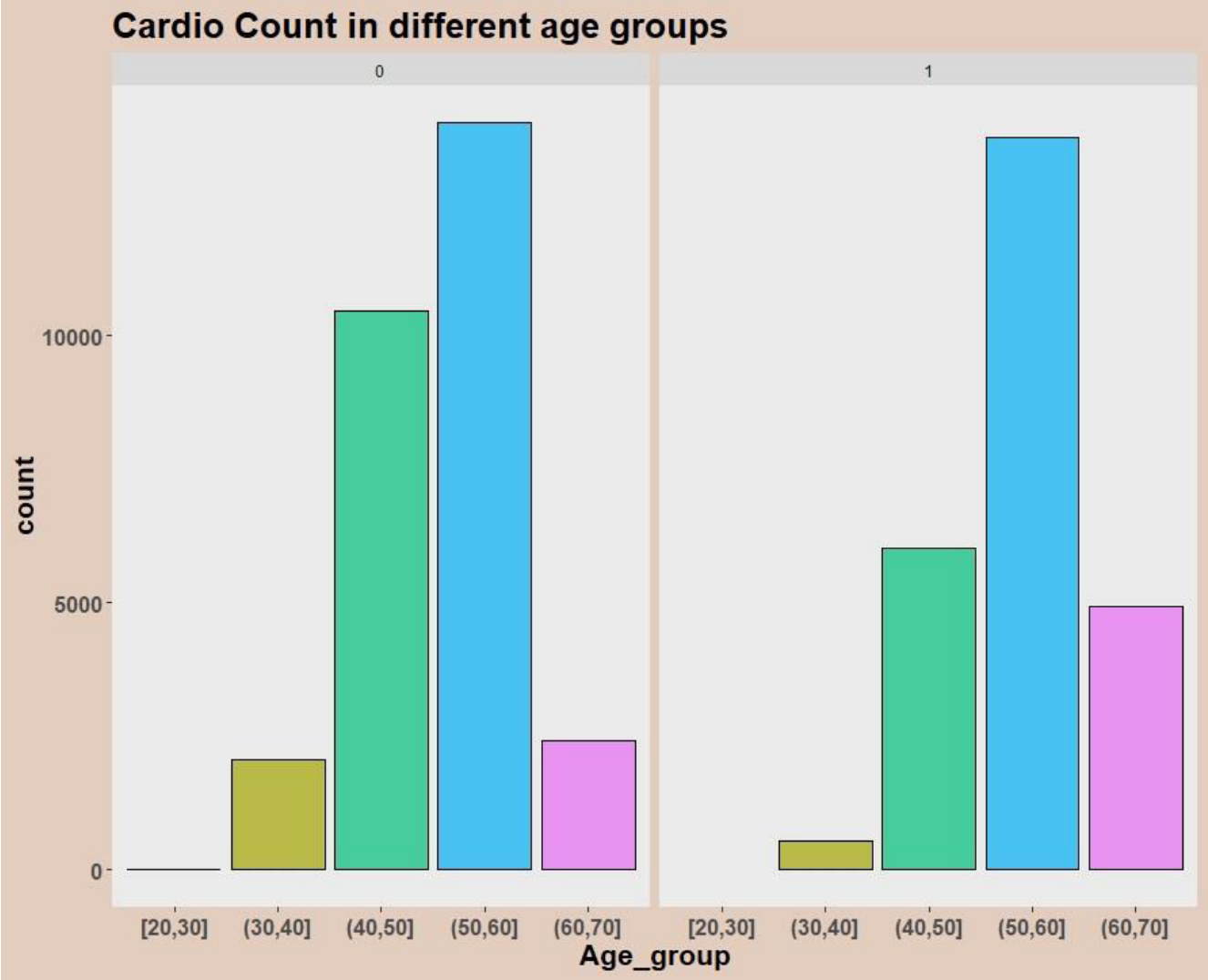
Plot - 12



Plot - 13

Observation: As we move higher up along the Tension class, we observe a steady increase in the mean BMI, which proves to be a strong indicator of heart diseases.

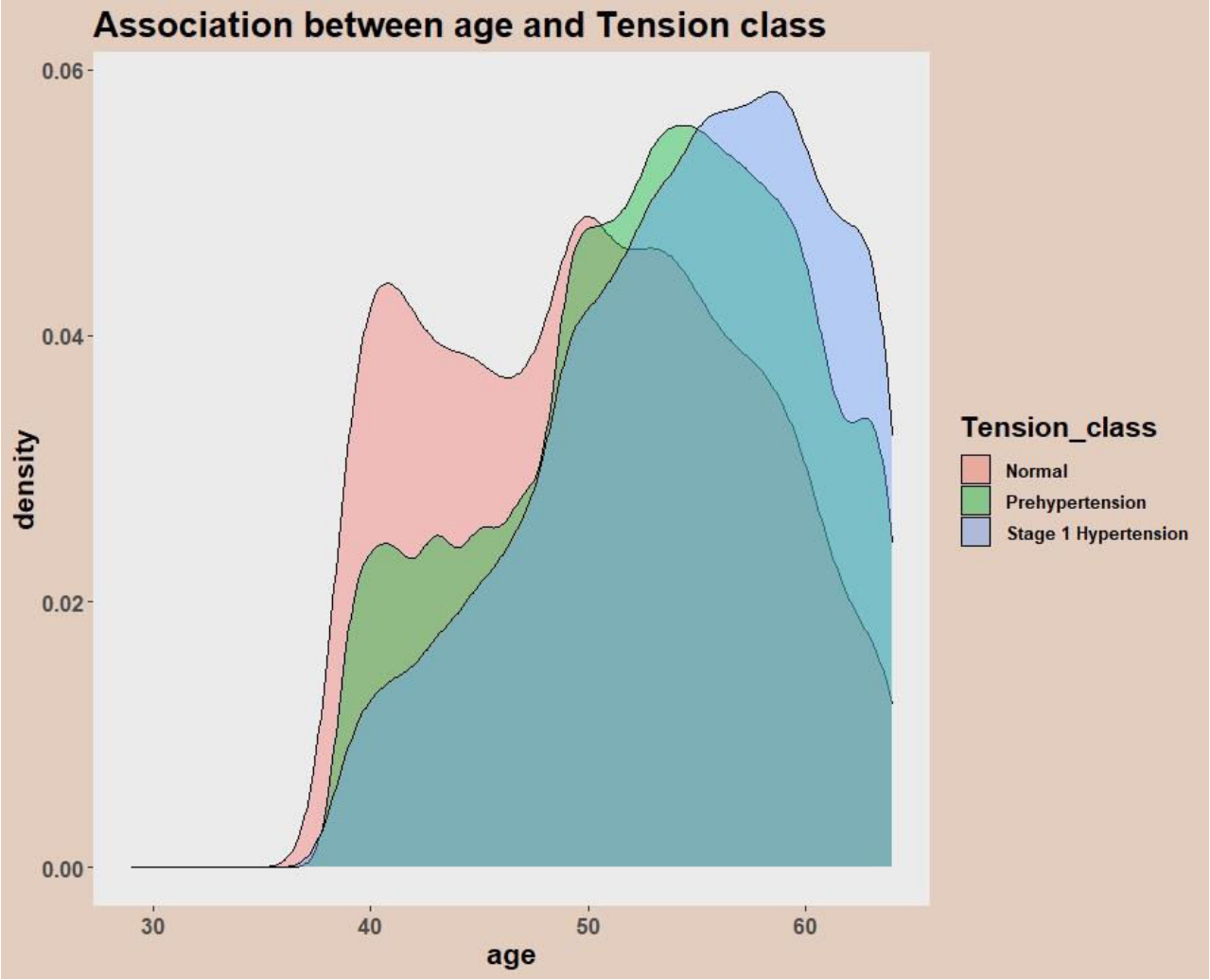
- **Are old people at a greater risk than the younger ones?**



Plot - 14

Observation:

In general, the risk of cardiovascular disease is greater in older people than in younger ones.

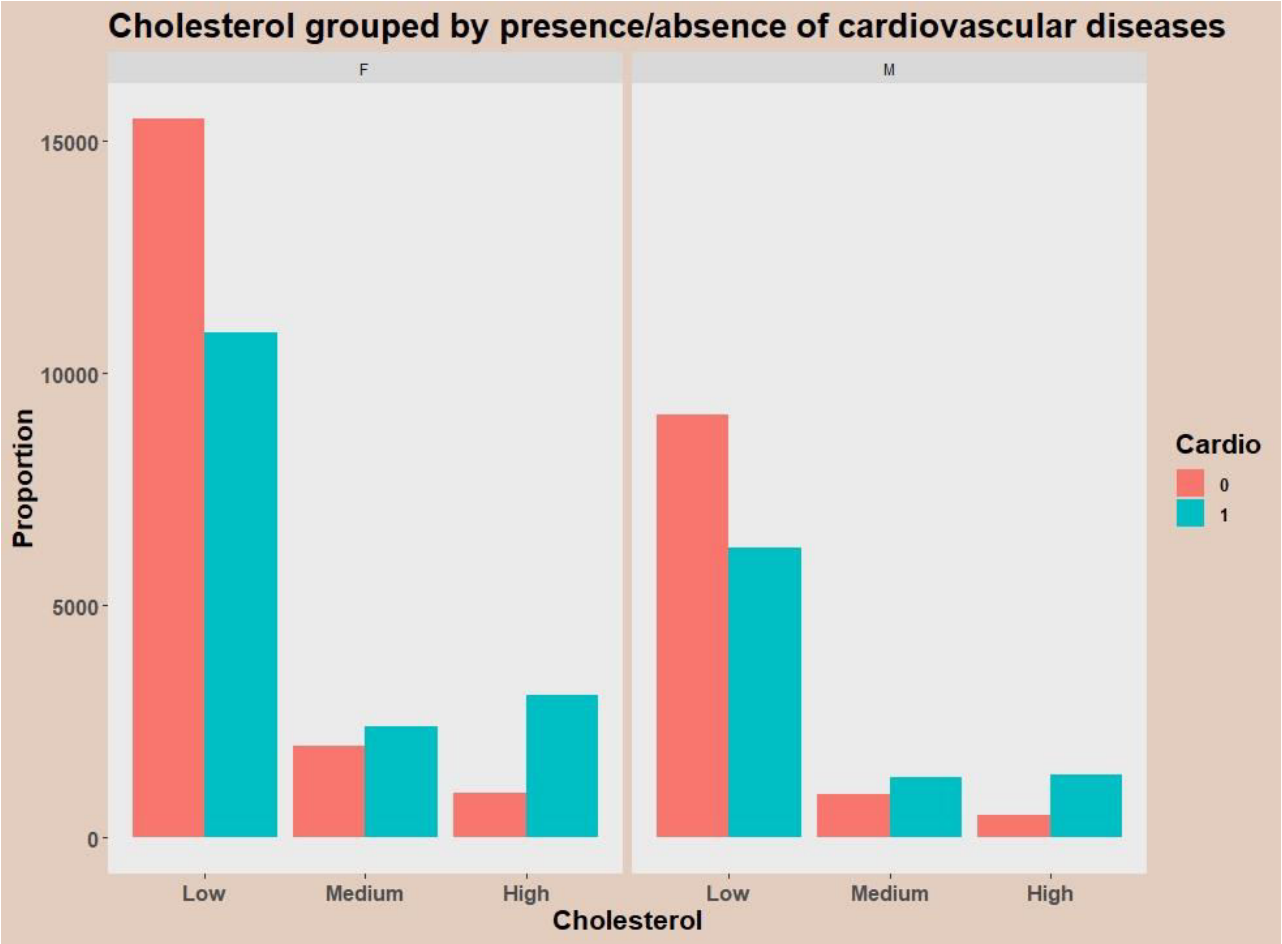


Plot - 15

Observation:

Higher age is associated with higher level in the Tension class, which again is a strong indicator of cardiovascular diseases. This helps solidify our conjecture that higher age is associated greater risk of heart diseases.

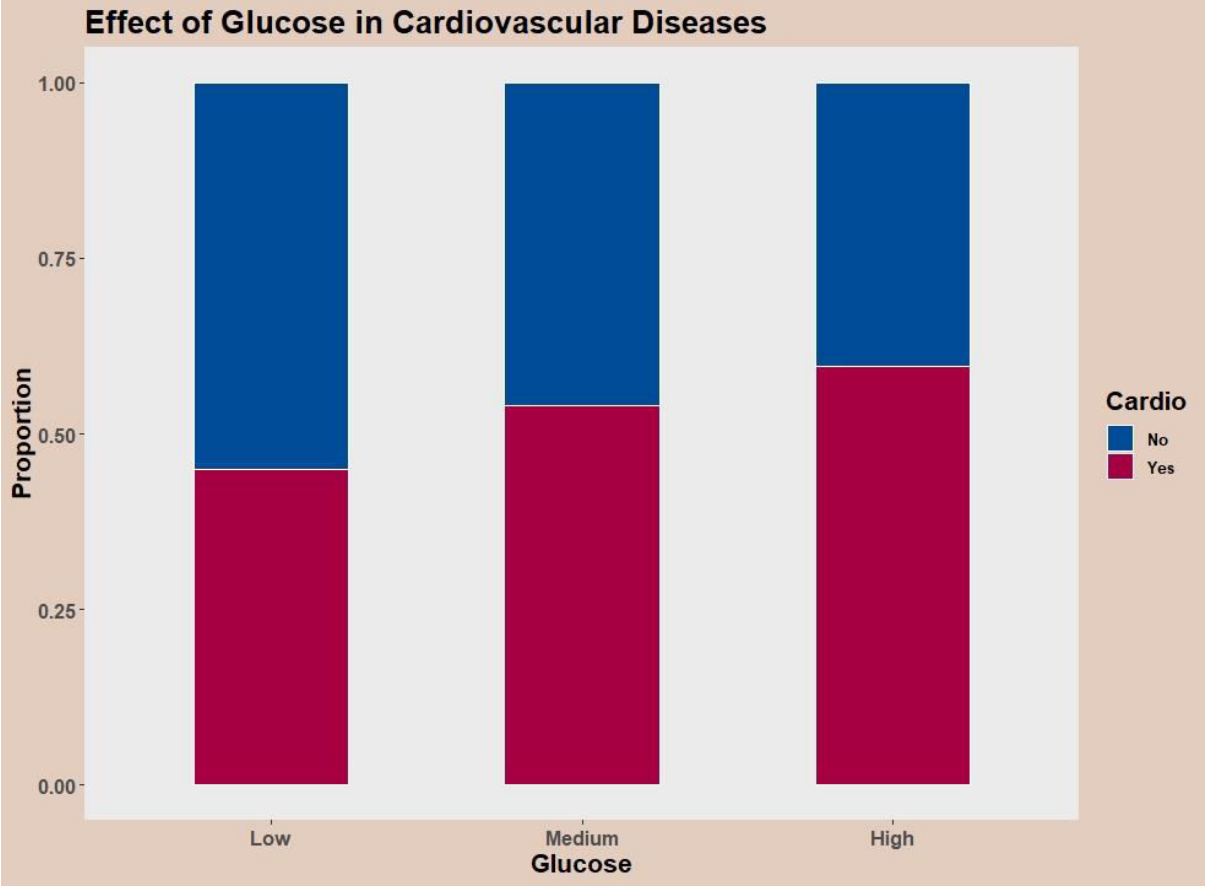
- **How is a person's cholesterol/glucose level associated with the presence of cardiovascular diseases?**



Plot - 16

Observation:

People with normal cholesterol levels are less likely to have heart diseases, while those who have their cholesterol levels above and well above normal are more prone to the same.



Plot - 17

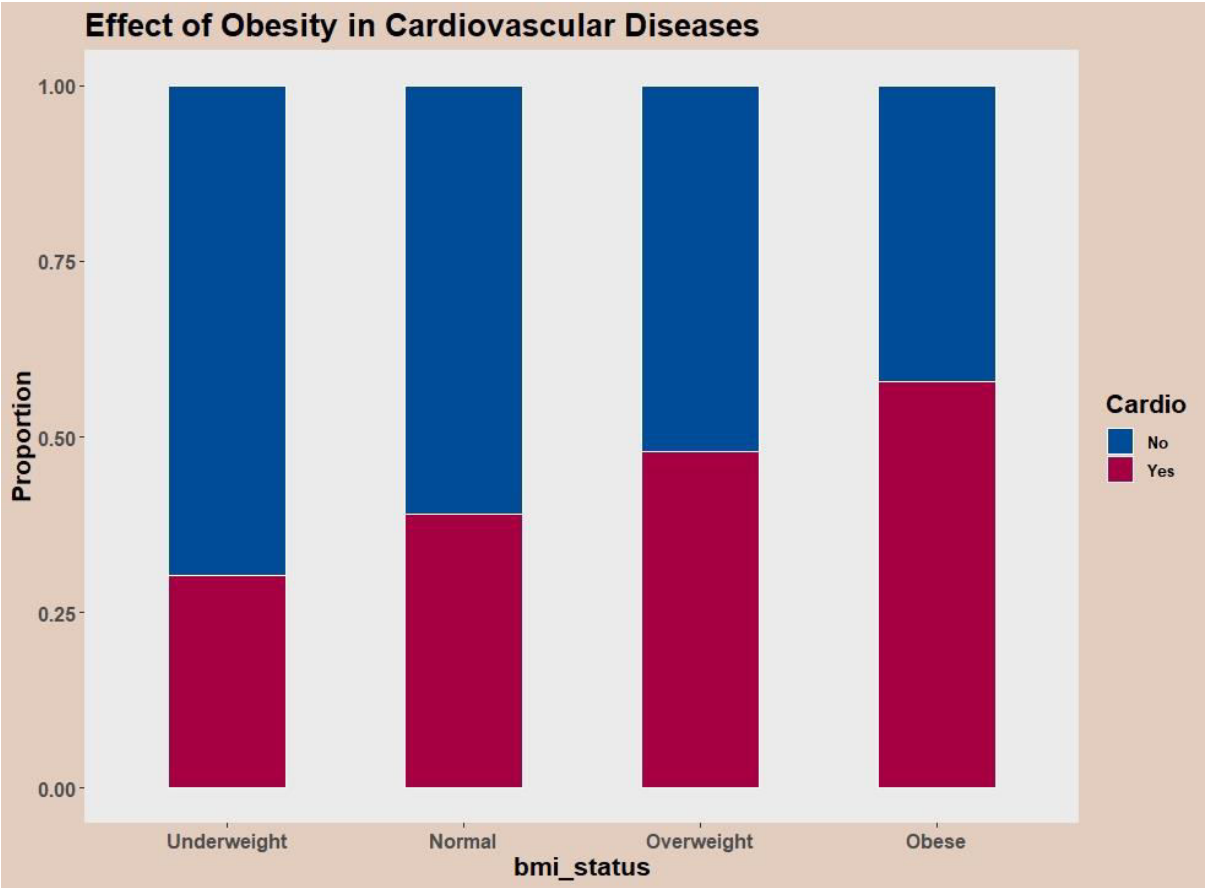
Observation:

People with normal glucose levels are less likely to have heart diseases, while those who have their cholesterol levels above and well above normal are more prone to the same.

- **Does obesity increase the risk of heart disease?**

Weight Categories	BMI (kg/m^2)
Underweight	< 18.5
Healthy Weight	18.5-24.9
Overweight	25-29.9
Obese	30-34.9
Severely Obese	35-39.9
Morbidly Obese	≥40

We shall now analyse the effect of BMI status on heart diseases, for which we group BMI values in our dataset in groups according to the above table.



Plot - 18

Observation: Obese people are at a much greater risk of getting heart diseases than people with normal body weight.

CONCLUSION

From our analysis of the cardiovascular disease dataset, it is evident that the main contributing factors of heart diseases are :

i. Age
ii. Activity
iii. BMI / BMI status
iv. Systolic, diastolic blood pressure / Tension class
v. Cholesterol level
vi. Glucose level

There are also some inter-relations between the factors listed above, which we have established through our analysis.

APPENDIX

```
################################### loading the data ###################################
card <- read.csv('C:\\Users\\oishi\\Desktop\\PDS FINAL PROJECT\\cardio_train.csv',sep=';')

############################### loading necessary libraries ############################
library(dplyr)
library(ggplot2)
library(cowplot)
library(reshape)

################################### modifying the data ###################################
card1<-card
head(card1)
sapply(card1,function(x) sum(is.na(x)))                    # no NA values

#column - 'age'
card1 <- within(card1, age <- age%/%365); head(card1)
max(card1$age)
min(card1$age)

#column - 'height', 'weight', 'ap_hi', 'ap_lo'
card1 %>% select(height,weight,ap_hi,ap_lo) %>% apply(.,2,max)
```

```
card1 %>% select(height,weight,ap_hi,ap_lo) %>% apply(.,2,min)

card1 %>% select(ap_hi,ap_lo) %>% apply(.,2,function(x) sum(x<0))    # number of negative values seperately in both columns

sum(card1$ap_hi<0 | card1$ap_lo<0)                                   # number of rows with ap_hi or ap_lo being negative


#new columns - 'BMI', 'BP', 'Pulse_pressure', 'Age_group', 'Tension_class', 'MAP', 'BMI_status'

card1<- card1 %>% mutate(BMI = weight/(height*height)*100*100)

card1$BMI <- round(card1$BMI,digits = 2)

card1<-card1 %>% filter(BMI<250)

card1 <- card1 %>% mutate(Pulse_pressure = ap_hi-ap_lo)

card1<-card1 %>% filter(Pulse_pressure>=10)

card1 <- card1 %>% mutate(Age_group = cut(age,breaks=c(20,30,40,50,60,70),include.lowest = T))

card1<-card1 %>% mutate(MAP = ap_lo+(Pulse_pressure)/3)

f1 <- function(x,y){
  if(x<120 & y<80){
    return(1)          # 1 --> Normal
  }
  else if(x<139 | y<89){
    return(2)          # 2 --> Prehypertension
  }
  else if(x<159 | y<99){
    return(3)          # 3 --> Stage 1 Hypertension
  }
  else{
```

```
    return(4)           # 4 --> Stage 2 Hypertension
  }
}
card1<-card1 %>% mutate(Tension_class=mapply(f1,ap_hi,ap_lo))
######BMI      Weight Status
#Below 18.5     Underweight ---------->1
#18.5-24.9      Normal      ---------->2
#25.0-29.9      Overweight  ---------->3
#30.0 and Above      Obese   ---------->4
card1$bmi_status<-
ifelse(card1$BMI<18.5,1,ifelse((card1$BMI<25)&(card1$BMI>=18.
5),2,ifelse((card1$BMI<30)&(card1$BMI>=25),3,4)))
################################### data with positive
cardiovascular diseases ##############################
card_yes<-card1 %>% filter(cardio==1)
nrow(card_yes)


######################################### Cleaning the
data #######################################
# boxplot to check outliers
card1melt<-melt(card1[,c(3,6,7,14,15,17)],id.var="gender")
##### PLOT 1: Type G. 4 or more variables
boxplot(card1[,c(6,7,14,15,17)])


##### PLOT2: TYPE G. 4 or more variables
ggplot(card1melt, aes(y=value, x=variable)) +
  geom_boxplot(aes(fill=as.factor(gender))) +
  scale_fill_discrete(name="gender",labels=c("F","M")) +
```

```
  ylim(0,250) +
  labs(title = 'Outlier check (Before cleaning)')



outliers <- function(x) {
  Q1 <- quantile(x, probs=.25)
  Q3 <- quantile(x, probs=.75)
  IQR = Q3-Q1
  lower = Q1 - (IQR*1.5)
  upper = Q3 + (IQR*1.5)
  x > upper | x < lower
}

remove_outliers <- function(card3, cols = names(card3)) {
  for (i in cols) {
    card3 <- card3[!outliers(card3[[i]]),]
  }
  card3
}

card2 <- remove_outliers(card1,c('ap_hi','ap_lo','BMI','Pulse_pressure','MAP'))

# boxplot for double check
card2melt<-melt(card2[,c(3,6,7,14,15,17)],id.var="gender")

##### PLOT 3: Type G. 4 or more variables
```

```
ggplot(card2melt, aes(variable,value)) +
geom_boxplot(aes(fill=as.factor(gender))) +
 labs(title = 'Outlier check (After cleaning)') +
scale_fill_discrete(name="gender",labels=c("F","M"))

##### PLOT 4: Type G. 4 or more variables

ggplot(card2melt, aes(variable,value)) +
geom_violin(aes(fill=as.factor(gender))) +
scale_fill_discrete(name="gender",labels=c("F","M")) +
ggtitle('outliers with distribution (after cleaning)')


########################################## plotting for insights ##################################

# heatmap of the whole dataset

card3<-card2

cormat <- round(cor(subset(card3, select = -c(Age_group))),2)

melted_cormat <- melt(cormat,na.rm=TRUE)

head(melted_cormat)

##### PLOT 5: G. 4 or more variables

hmap_card<-ggplot(data = melted_cormat, aes(x=X2, y=X1, fill=value)) + geom_tile(color = "white")+
 geom_text(aes(X2,X1, label = value), color = "black", size = 4)+
 scale_fill_gradient2(low = "yellow", high = "green",
                      limit = c(-1,1), space = "Lab",
                      name="  Pearson\nCorrelation")+ #The function scale_fill_gradient2 is used with the argument limit = c(-1,1) as correlation coefficients range from -1 to 1.
 theme_minimal()+
 theme(axis.text.x = element_text(angle = 90, vjust = 1,
                                  size = 12, hjust = 1))+
```

```
  coord_fixed()  #coord_fixed() : this function ensures that one unit on the x-axis is the same length as one unit on the y-axis.

hmap_card+
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        axis.ticks = element_blank(),
        legend.justification = c(1, 0))


####### categorizing the entries of the categorical columns
card2$gender <- factor(card2$gender,levels = c('1','2'), labels= c('F','M'))
card2$smoke <- factor(card2$smoke,levels = c('0','1'), labels= c('no','yes'))
card2$alco <- factor(card2$alco,levels = c('0','1'), labels= c('no','yes'))
card2$active <- factor(card2$active,levels = c('0','1'), labels= c('no','yes'))
card2$Tension_class<-factor(card2$Tension_class,levels = c('1','2','3','4'), labels= c('Normal','Prehypertension','Stage 1 Hypertension','Stage 2 Hypertension'))
card2$bmi_status<-factor(card2$bmi_status,levels = c('1','2','3','4'), labels= c('Underweight','Normal','Overweight','Obese'))
head(card2)


#male female height weight comparison w.r.t. gender
```

```
##### PLOT 6:F. three variables (one categorical and two continuous)

ggplot(card2,aes(x=height,weight)) + geom_point(aes(colour=gender),alpha=0.6,size=1) + ggtitle('Male Female mass comparison') +theme_classic()


#alcohol vs cardiovascular diseases

##### PLOT 7:E. two variables (both categorical)


df1<-card2%>%group_by(alco,cardio)%>%summarise(Count=n())

df2<-df1%>%group_by(alco)%>%summarise(tot=sum(Count))

df1<-merge(df1,df2,by="alco")%>%mutate(Proportion=Count/tot)


#plotting the proportion


ggplot(df1,aes(x=factor(alco),y=Proportion)) +

  geom_col(aes(fill=factor(cardio)),color="white",width = 0.5) +

  scale_fill_manual(name="Cardio",labels=c("No","Yes"),values = c("#004C99","#A70042")) +


theme(title=element_text(size=16,face="bold"),axis.text=element_text(size=12,face="bold"),legend.text = element_text(size=10,face="bold"),

      axis.ticks.x = element_blank(),panel.grid = element_blank(),

      plot.background = element_rect(fill = "#E3CDBF",color = "#E3CDBF"),

      legend.background = element_rect( fill = "#E3CDBF",color = "#E3CDBF"),

      legend.key = element_rect(fill = "#E3CDBF",color = "#E3CDBF"))+
```

```
  ggtitle('Effect of Alcohol in Cardiovascular Diseases')+
  labs(x="Alcohol",y="Proportion")


#smoke vs cardiovascular diseases
##### PLOT 8:E. two variables (both categorical)


df3<-
card2%>%group_by(smoke,cardio)%>%summarise(Count=n())
df4<-df3%>%group_by(smoke)%>%summarise(tot=sum(Count))
df3<-
merge(df3,df4,by="smoke")%>%mutate(Proportion=Count/tot)


#plotting the proportion


ggplot(df3,aes(x=factor(smoke),y=Proportion)) +
  geom_col(aes(fill=factor(cardio)),color="white",width = 0.5) +
  scale_fill_manual(name="Cardio",labels=c("No","Yes"),values =
c("#004C99","#A70042")) +

theme(title=element_text(size=16,face="bold"),axis.text=element_t
ext(size=12,face="bold"),legend.text =
element_text(size=10,face="bold"),
      axis.ticks.x = element_blank(),panel.grid = element_blank(),
      plot.background = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"),
      legend.background = element_rect( fill = "#E3CDBF",color =
"#E3CDBF"),
      legend.key = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"))+
  ggtitle('Effect of Smoking in Cardiovascular Diseases')+
```

```
  labs(x="Smoking",y="Proportion")


#active vs cardiovascular diseases
##### PLOT 9:E. two variables (both categorical)


df5<-
card2%>%group_by(active,cardio)%>%summarise(Count=n())
df6<-df5%>%group_by(active)%>%summarise(tot=sum(Count))
df5<-
merge(df5,df6,by="active")%>%mutate(Proportion=Count/tot)


#plotting the proportion


ggplot(df5,aes(x=factor(active),y=Proportion)) +
  geom_col(aes(fill=factor(cardio)),color="white",width=0.5) +
  scale_fill_manual(name="Cardio",labels=c("No","Yes"),values =
c("#004C99","#A70042")) +


theme(title=element_text(size=16,face="bold"),axis.text=element_t
ext(size=12,face="bold"),legend.text =
element_text(size=10,face="bold"),
      axis.ticks.x = element_blank(),panel.grid = element_blank(),
      plot.background = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"),
      legend.background = element_rect( fill = "#E3CDBF",color =
"#E3CDBF"),
      legend.key = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"))+
  ggtitle('Effect of Activity in Cardiovascular Diseases')+
  labs(x="Activity",y="Proportion")
```

```
#BMI vs cardiovascular diseases

##### PLOT 10:D. two variables (one categorical and one continuous)


ggplot(card2, aes(BMI, after_stat(density), color = factor(cardio))) +

  geom_freqpoly(size=1,binwidth = function(x) 2 * IQR(x) / (length(x)^(1/3))) +

  ggtitle('Effect of BMI in cardiovascular disease')+


theme(title=element_text(size=16,face="bold"),axis.text=element_text(size=12,face="bold"),legend.text = element_text(size=10,face="bold"),

      panel.grid = element_blank(), plot.background = element_rect(fill = "#E3CDBF",color = "#E3CDBF"),

      legend.background = element_rect( fill = "#E3CDBF",color = "#E3CDBF"),

      legend.key = element_rect(fill = "#E3CDBF",color = "#E3CDBF"))+

  labs(x="BMI",y="Density")+


scale_color_discrete(name="Cardio",labels=c("Absence","Presence"))


#tension_scale vs cardiovascular diseases

##### PLOT 11:E. two variables (both categorical)


df7<-
card2%>%group_by(Tension_class,cardio)%>%summarise(Count=n())
```

```
df8<-
df7%>%group_by(Tension_class)%>%summarise(tot=sum(Count))
df7<-
merge(df7,df8,by="Tension_class")%>%mutate(Proportion=Count/
tot)


#plotting the proportion


ggplot(df7,aes(x=Tension_class,y=Proportion)) +
  geom_col(aes(fill=factor(cardio)),color="white",width = 0.5) +
  scale_fill_manual(name="Cardio",labels=c("No","Yes"),values =
c("#004C99","#A70042")) +

theme(title=element_text(size=16,face="bold"),axis.text=element_t
ext(size=12,face="bold"),legend.text =
element_text(size=10,face="bold"),
      axis.ticks.x = element_blank(),panel.grid = element_blank(),
      plot.background = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"),
      legend.background = element_rect( fill = "#E3CDBF",color =
"#E3CDBF"),
      legend.key = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"))+
  ggtitle('Effect of Tension_class in Cardiovascular Diseases')+
  labs(x="Tension_class",y="Proportion")


#bmi vs tension_class
##### PLOT 12:D. two variables (one categorical and one
continuous)
ggplot(card2,aes(BMI,fill=Tension_class)) +
```

```
  geom_density(alpha=0.4) +

theme(title=element_text(size=16,face="bold"),axis.text=element_text(size=12,face="bold"),legend.text = element_text(size=10,face="bold"),

title=element_text(size=16,face="bold"),axis.text=element_text(size=12,face="bold"),legend.text = element_text(size=10,face="bold"),
      panel.grid = element_blank(), plot.background = element_rect(fill = "#E3CDBF",color = "#E3CDBF"),
      legend.background = element_rect( fill = "#E3CDBF",color = "#E3CDBF"),
      legend.key = element_rect(fill = "#E3CDBF",color = "#E3CDBF"))+
  labs(title='Change in BMI density grouped by Tension class')



#bmi vs tension_class
##### PLOT 13:F. three variables (two categorical and one continuous)
ggplot(card2,aes(BMI,fill=Tension_class))+
  geom_density(alpha=0.4)+
  facet_wrap(~cardio) + labs(title='Association between BMI and Tension class')+

theme(title=element_text(size=16,face="bold"),axis.text=element_text(size=12,face="bold"),legend.text = element_text(size=10,face="bold"),
  panel.grid = element_blank(), plot.background = element_rect(fill = "#E3CDBF",color = "#E3CDBF"),
  legend.background = element_rect( fill = "#E3CDBF",color = "#E3CDBF"),
```

```
  legend.key = element_rect(fill = "#E3CDBF",color = "#E3CDBF"))


#Age_group vs cardiovascular diseases
##### PLOT 14:E. two variables (both categorical)
ggplot(card2,aes(Age_group)) +
  geom_bar(color='black',aes(fill=Age_group),show.legend = F,alpha=0.7, position = 'dodge') +
  facet_wrap(~cardio) + ggtitle('Cardio Count in different age groups')+

theme(title=element_text(size=16,face="bold"),axis.text=element_text(size=12,face="bold"),legend.text = element_text(size=10,face="bold"),
      panel.grid = element_blank(), plot.background = element_rect(fill = "#E3CDBF",color = "#E3CDBF"),
      legend.background = element_rect( fill = "#E3CDBF",color = "#E3CDBF"))


#age vs tension_class
##### PLOT 15:D. two variables (one discrete and one categorical)
ggplot(card2,aes(age,fill=Tension_class)) +
  geom_density(alpha=0.4)+ labs(title='Association between age and Tension class')+

theme(title=element_text(size=16,face="bold"),axis.text=element_text(size=12,face="bold"),legend.text = element_text(size=10,face="bold"),
      panel.grid = element_blank(), plot.background = element_rect(fill = "#E3CDBF",color = "#E3CDBF"),
```

```
    legend.background = element_rect( fill = "#E3CDBF",color = "#E3CDBF"),
    legend.key = element_rect(fill = "#E3CDBF",color = "#E3CDBF"))


#Cholesterol count grouped by presence/absence of cardiovascular diseases
##### PLOT 16:F. three variables (all categorical)
ggplot(card2) + geom_bar(aes(x = factor(cholesterol),fill=as.factor(cardio)),position = 'dodge') + scale_fill_discrete(name='Cardio') +
  facet_wrap(~gender)+
  labs(title='Cholesterol grouped by presence/absence of cardiovascular diseases')+

theme(title=element_text(size=16,face="bold"),axis.text=element_text(size=12,face="bold"),legend.text = element_text(size=10,face="bold"),
    panel.grid = element_blank(), plot.background = element_rect(fill = "#E3CDBF",color = "#E3CDBF"),
    legend.background = element_rect( fill = "#E3CDBF",color = "#E3CDBF"),
    legend.key = element_rect(fill = "#E3CDBF",color = "#E3CDBF"))+

scale_x_discrete(breaks=c("1","2","3"),labels=c("1"='Low',"2"='Medium',"3"='High'))+
  labs(x="Cholesterol",y="Proportion")
```

```
#glucose vs cardiovascular diseases
##### PLOT 17:E. two variables (both categorical)
df9<-card2%>%group_by(gluc,cardio)%>%summarise(Count=n())
df10<-df9%>%group_by(gluc)%>%summarise(tot=sum(Count))
df9<-merge(df9,df10,by="gluc")%>%mutate(Proportion=Count/tot)

#plotting the proportion

ggplot(df9,aes(x=factor(gluc),y=Proportion)) +
  geom_col(aes(fill=factor(cardio)),color="white",width = 0.5) +
  scale_fill_manual(name="Cardio",labels=c("No","Yes"),values =
c("#004C99","#A70042")) +

scale_x_discrete(breaks=c("1","2","3"),labels=c("1"='Low',"2"='Medi
um',"3"='High'))+

theme(title=element_text(size=16,face="bold"),axis.text=element_t
ext(size=12,face="bold"),legend.text =
element_text(size=10,face="bold"),
      panel.grid = element_blank(),
      plot.background = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"),
      legend.background = element_rect( fill = "#E3CDBF",color =
"#E3CDBF"),
      legend.key = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"))+
  ggtitle('Effect of Glucose in Cardiovascular Diseases')+
  labs(x="Glucose",y="Proportion")
```

```
#bmi_status vs cardiovascular diseases

##### PLOT 18:E. two variables (both categorical)

df11<-
card2%>%group_by(bmi_status,cardio)%>%summarise(Count=n()
)

df12<-
df11%>%group_by(bmi_status)%>%summarise(tot=sum(Count))

df11<-
merge(df11,df12,by="bmi_status")%>%mutate(Proportion=Count/
tot)


ggplot(df11,aes(x=bmi_status,y=Proportion)) +

  geom_col(aes(fill=factor(cardio)),color="white",width = 0.5) +

  scale_fill_manual(name="Cardio",labels=c("No","Yes"),values =
c("#004C99","#A70042")) +


theme(title=element_text(size=16,face="bold"),axis.text=element_t
ext(size=12,face="bold"),legend.text =
element_text(size=10,face="bold"),

      panel.grid = element_blank(),

      plot.background = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"),

      legend.background = element_rect( fill = "#E3CDBF",color =
"#E3CDBF"),

      legend.key = element_rect(fill = "#E3CDBF",color =
"#E3CDBF"))+

  ggtitle('Effect of Obesity in Cardiovascular Diseases')+

  labs(x="bmi_status",y="Proportion")
```