STA 106 Final Project

Suvethika

2023-03-24

```
heart=read.csv('heart_disease_health_indicators_BRFSS2015.csv',header=TRUE)

heart$HeartDiseaseorAttack=as.factor(heart$HeartDiseaseorAttack)
levels(heart$HeartDiseaseorAttack) = c("Heart Disease or Attack Absent", "Heart Disease or Attack Present")
heart$HighBP=as.factor(heart$HighBP)
levels(heart$HighBP) = c("High BP Absent", "High BP Present")
heart$HighChol=as.factor(heart$HighChol)
levels(heart$HighChol) = c("High Chol Absent", "High Chol Present")
```

We can subset the data into different categories of General Health. General health is on a scale from 1-5 in this dataset.

```
#subsetting into different categories of Gen Health
sub1=heart[heart$GenHlth==1,]
sub2=heart[heart$GenHlth==2,]
sub3=heart[heart$GenHlth==3,]
sub4=heart[heart$GenHlth==4,]
sub5=heart[heart$GenHlth==5,]
```

```
head(sub1)
```

```
##              HeartDiseaseorAttack          HighBP          HighChol CholCheck
## 24 Heart Disease or Attack Absent High BP Present  High Chol Absent         1
## 30 Heart Disease or Attack Absent  High BP Absent High Chol Present         1
## 32 Heart Disease or Attack Absent High BP Present  High Chol Absent         1
## 33 Heart Disease or Attack Absent  High BP Absent  High Chol Absent         1
## 39 Heart Disease or Attack Absent  High BP Absent  High Chol Absent         1
## 56 Heart Disease or Attack Absent  High BP Absent  High Chol Absent         1
##    BMI Smoker Stroke Diabetes PhysActivity Fruits Veggies HvyAlcoholConsump
## 24  27      0      0        2            1      1       1                 0
## 30  31      1      0        0            1      1       1                 0
## 32  33      1      0        0            1      1       1                 0
## 33  23      0      0        0            1      1       1                 0
## 39  26      1      0        0            1      1       1                 0
## 56  29      0      0        0            1      0       1                 0
##    AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
## 24             1           0       1        0        0        0   0  13
## 30             1           0       1        0        0        0   1  12
## 32             0           0       1        0        0        1   1  13
## 33             1           0       1        2        0        0   0   6
## 39             0           0       1        0        1        0   1   4
## 56             1           0       1        0       10        0   1  11
##    Education Income
## 24         5      4
## 30         6      8
## 32         3      3
## 33         4      8
## 39         5      3
## 56         6      8
```

```
head(sub2)
```

```
##               HeartDiseaseorAttack          HighBP          HighChol CholCheck
## 4  Heart Disease or Attack Absent High BP Present  High Chol Absent         1
## 5  Heart Disease or Attack Absent High BP Present High Chol Present         1
## 6  Heart Disease or Attack Absent High BP Present High Chol Present         1
## 10 Heart Disease or Attack Absent  High BP Absent  High Chol Absent         1
## 16 Heart Disease or Attack Absent High BP Present  High Chol Absent         1
## 18 Heart Disease or Attack Absent  High BP Absent  High Chol Absent         1
##    BMI Smoker Stroke Diabetes PhysActivity Fruits Veggies HvyAlcoholConsump
## 4   27      0      0        0            1      1       1                 0
## 5   24      0      0        0            1      1       1                 0
## 6   25      1      0        0            1      1       1                 0
## 10  24      0      0        0            0      0       1                 0
## 16  33      0      0        0            1      0       0                 0
## 18  23      1      0        2            1      0       0                 0
##    AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
## 4              1           0       2        0        0        0   0  11
## 5              1           0       2        3        0        0   0  11
## 6              1           0       2        0        2        0   1  10
## 10             1           0       2        0        0        0   1   8
## 16             1           0       2        5        0        0   0   6
## 18             1           0       2        0        0        0   1   7
##    Education Income
## 4          3      6
## 5          5      4
## 6          6      8
## 10         4      3
## 16         6      8
## 18         5      6
```

```
head(sub3)
```

```
##              HeartDiseaseorAttack          HighBP          HighChol CholCheck
## 2  Heart Disease or Attack Absent  High BP Absent  High Chol Absent         0
## 7  Heart Disease or Attack Absent High BP Present  High Chol Absent         1
## 8  Heart Disease or Attack Absent High BP Present High Chol Present         1
## 11 Heart Disease or Attack Absent  High BP Absent  High Chol Absent         1
## 12 Heart Disease or Attack Absent High BP Present High Chol Present         1
## 13 Heart Disease or Attack Absent  High BP Absent  High Chol Absent         1
##    BMI Smoker Stroke Diabetes PhysActivity Fruits Veggies HvyAlcoholConsump
## 2   25      1      0        0            1      0       0                 0
## 7   30      1      0        0            0      0       0                 0
## 8   25      1      0        0            1      0       1                 0
## 11  25      1      0        2            1      1       1                 0
## 12  34      1      0        0            0      1       1                 0
## 13  26      1      0        0            0      0       1                 0
##    AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
## 2              0           1       3        0        0        0   0   7
## 7              1           0       3        0       14        0   0   9
## 8              1           0       3        0        0        1   0  11
## 11             1           0       3        0        0        0   1  13
## 12             1           0       3        0       30        1   0  10
## 13             1           0       3        0       15        0   0   7
##    Education Income
## 2          6      1
## 7          6      7
## 8          4      4
## 11         6      8
## 12         5      1
## 13         5      7
```

```
head(sub4)
```

```
##                HeartDiseaseorAttack          HighBP          HighChol CholCheck
## 14  Heart Disease or Attack Absent High BP Present High Chol Present         1
## 15  Heart Disease or Attack Absent  High BP Absent High Chol Present         1
## 28 Heart Disease or Attack Present High BP Present High Chol Present         1
## 29  Heart Disease or Attack Absent High BP Present High Chol Present         1
## 31  Heart Disease or Attack Absent High BP Present High Chol Present         1
## 43  Heart Disease or Attack Absent  High BP Absent  High Chol Absent         1
##    BMI Smoker Stroke Diabetes PhysActivity Fruits Veggies HvyAlcoholConsump
## 14  28      0      0        2            0      0       1                 0
## 15  33      1      1        0            1      0       1                 0
## 28  28      1      0        2            0      0       1                 0
## 29  27      1      0        2            0      1       1                 0
## 31  34      1      1        2            1      0       0                 0
## 43  28      1      1        0            0      1       1                 0
##    AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
## 14             1           0       4        0        0        1   0  11
## 15             1           1       4       30       28        0   0   4
## 28             1           0       4        0        0        0   1  12
## 29             1           0       4       20       20        1   0   8
## 31             1           0       4        0        7        1   0   9
## 43             1           1       4       15       30        1   0   7
##    Education Income
## 14         4      6
## 15         6      2
## 28         2      4
## 29         4      7
## 31         5      4
## 43         4      3
```

```
head(sub5)
```

```
##               HeartDiseaseorAttack          HighBP          HighChol CholCheck
## 1   Heart Disease or Attack Absent High BP Present High Chol Present         1
## 3   Heart Disease or Attack Absent High BP Present High Chol Present         1
## 9  Heart Disease or Attack Present High BP Present High Chol Present         1
## 22  Heart Disease or Attack Absent High BP Present High Chol Present         1
## 27 Heart Disease or Attack Present High BP Present High Chol Present         1
## 40  Heart Disease or Attack Absent High BP Present High Chol Present         1
##    BMI Smoker Stroke Diabetes PhysActivity Fruits Veggies HvyAlcoholConsump
## 1   40      1      0        0            0      0       1                 0
## 3   28      0      0        0            0      1       0                 0
## 9   30      1      0        2            0      1       1                 0
## 22  38      1      0        0            0      1       1                 0
## 27  37      1      1        2            0      0       1                 0
## 40  24      1      0        0            0      1       1                 0
##    AnyHealthcare NoDocbcCost GenHlth MentHlth PhysHlth DiffWalk Sex Age
## 1              1           0       5       18       15        1   0   9
## 3              1           1       5       30       30        1   0   9
## 9              1           0       5       30       30        1   0   9
## 22             1           0       5       15       30        1   0  13
## 27             1           0       5        0        0        1   1  10
## 40             1           0       5        0       30        0   1   9
##    Education Income
## 1          4      3
## 3          4      8
## 9          5      1
## 22         2      3
## 27         6      5
## 40         3      1
```

The binary variables we are interested in, such as heart disease or attack, high blood pressure, and high cholesterol, are important health indicators that can have a significant impact on an individual’s overall health. By exploring these variables, we can gain insights into how they relate to other health indicators like BMI.

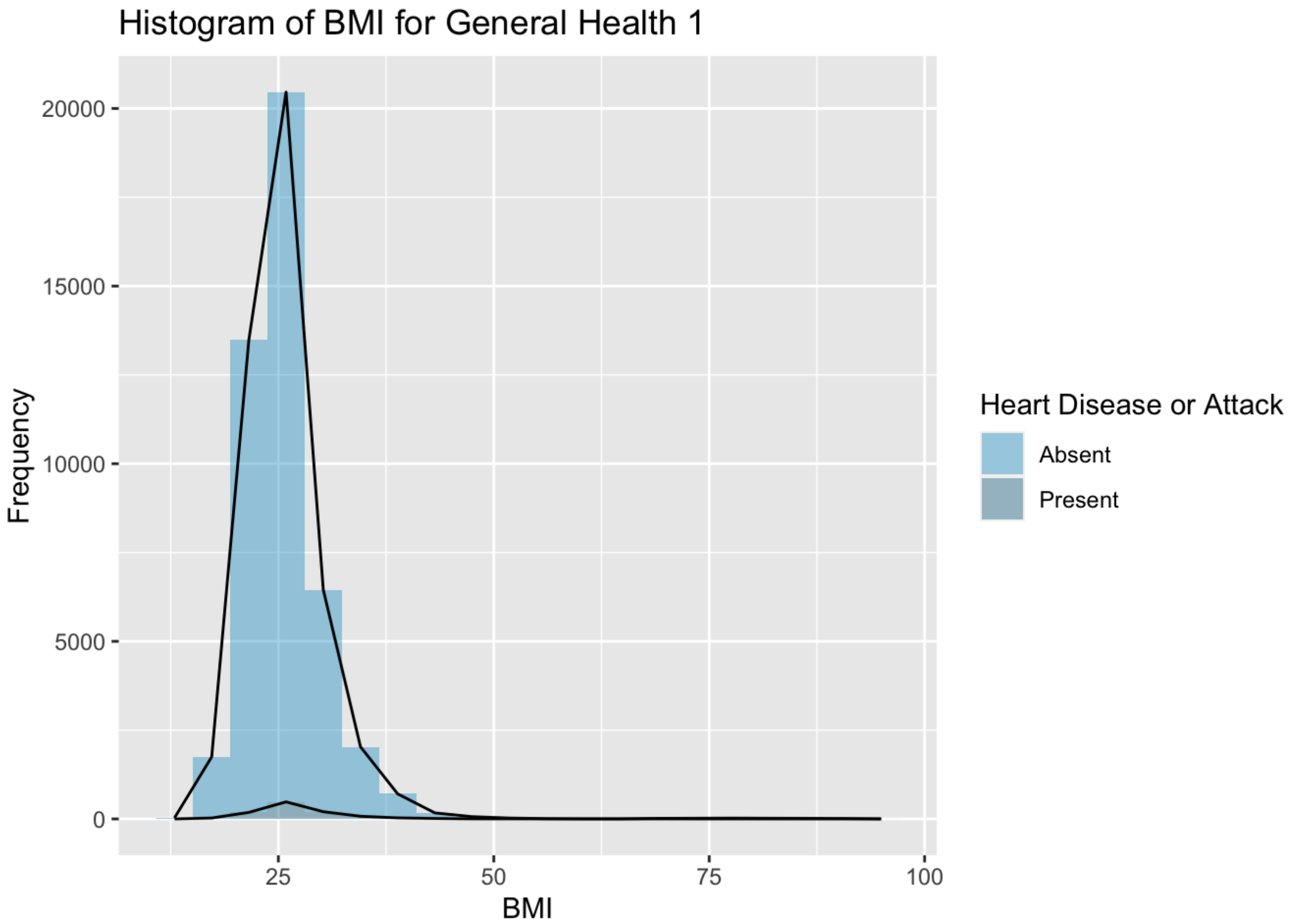
To begin our exploration, we can create histograms to compare the distribution of BMI for individuals with and without heart disease or attack, high blood pressure, and high cholesterol. This will help us understand how these health indicators are related to BMI and whether there are any differences in the distribution of BMI for individuals with and without these health conditions.

For example, we might create a histogram of BMI for individuals with heart disease or attack and another histogram of BMI for individuals without heart disease or attack. We can then compare the two histograms to see if there are any notable differences in the distribution of BMI for these two groups. Similarly, we can create histograms of BMI for individuals with and without high blood pressure and high cholesterol and compare them to gain a better understanding of how these health conditions are related to BMI.

We can start by exploring information about subset 1 which is made of individuals with a general health score of 1.

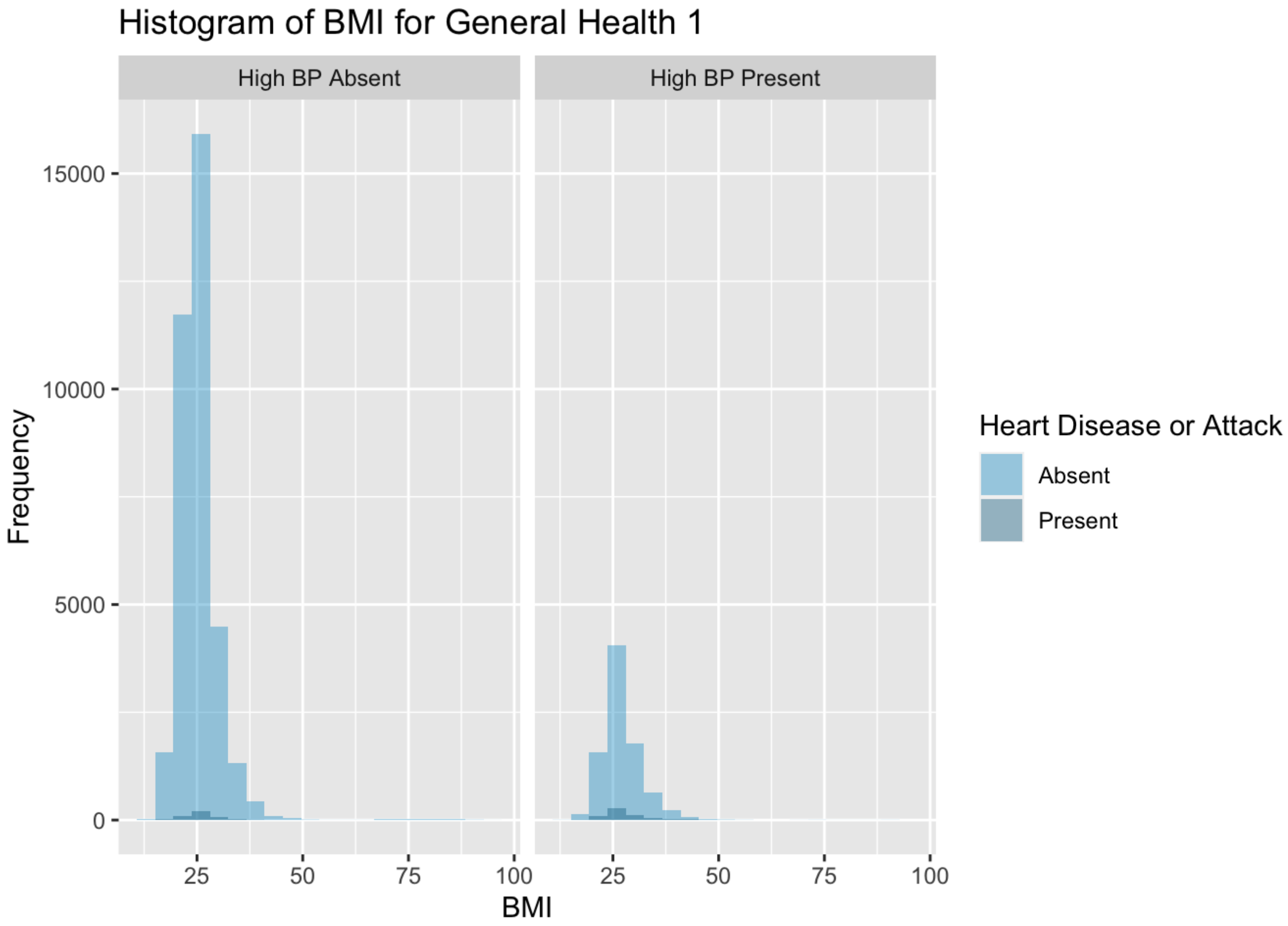
```
#install.packages('ggplot2')
library(ggplot2)
ggplot(data=sub1, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  stat_bin(bins=20, alpha=1, geom="line", aes(y=..count..)) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 1", y = "Frequency") +
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(count)` instead.
```



There seems to be a right skew for both those with heart disease and those without. Let’s continue analyzing the other variables.

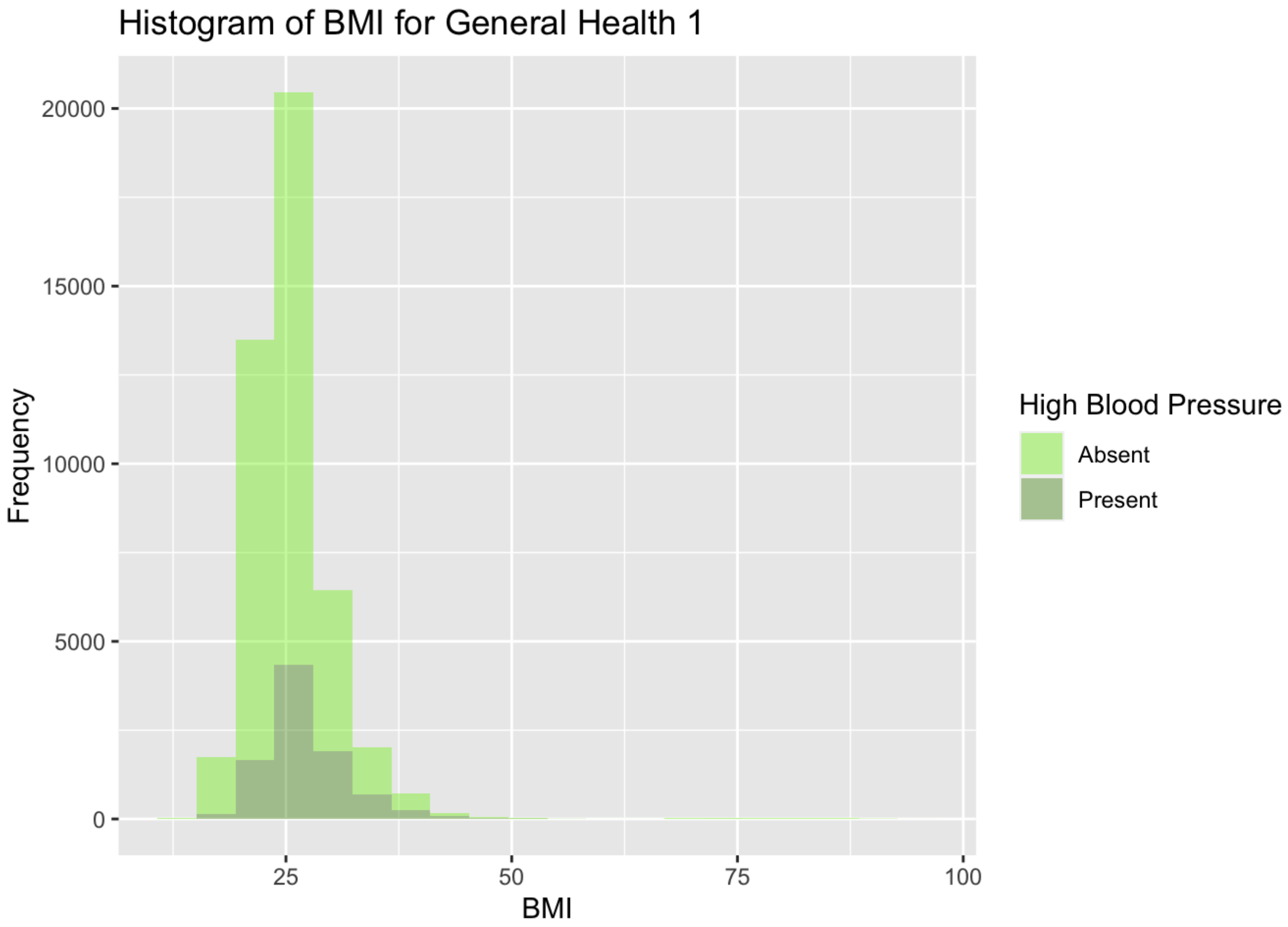
```
ggplot(data=sub1, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 1", y = "Frequency") +
  facet_wrap(vars(sub1$HighBP))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram on BMI, Heart Disease, and Blood Pressure.

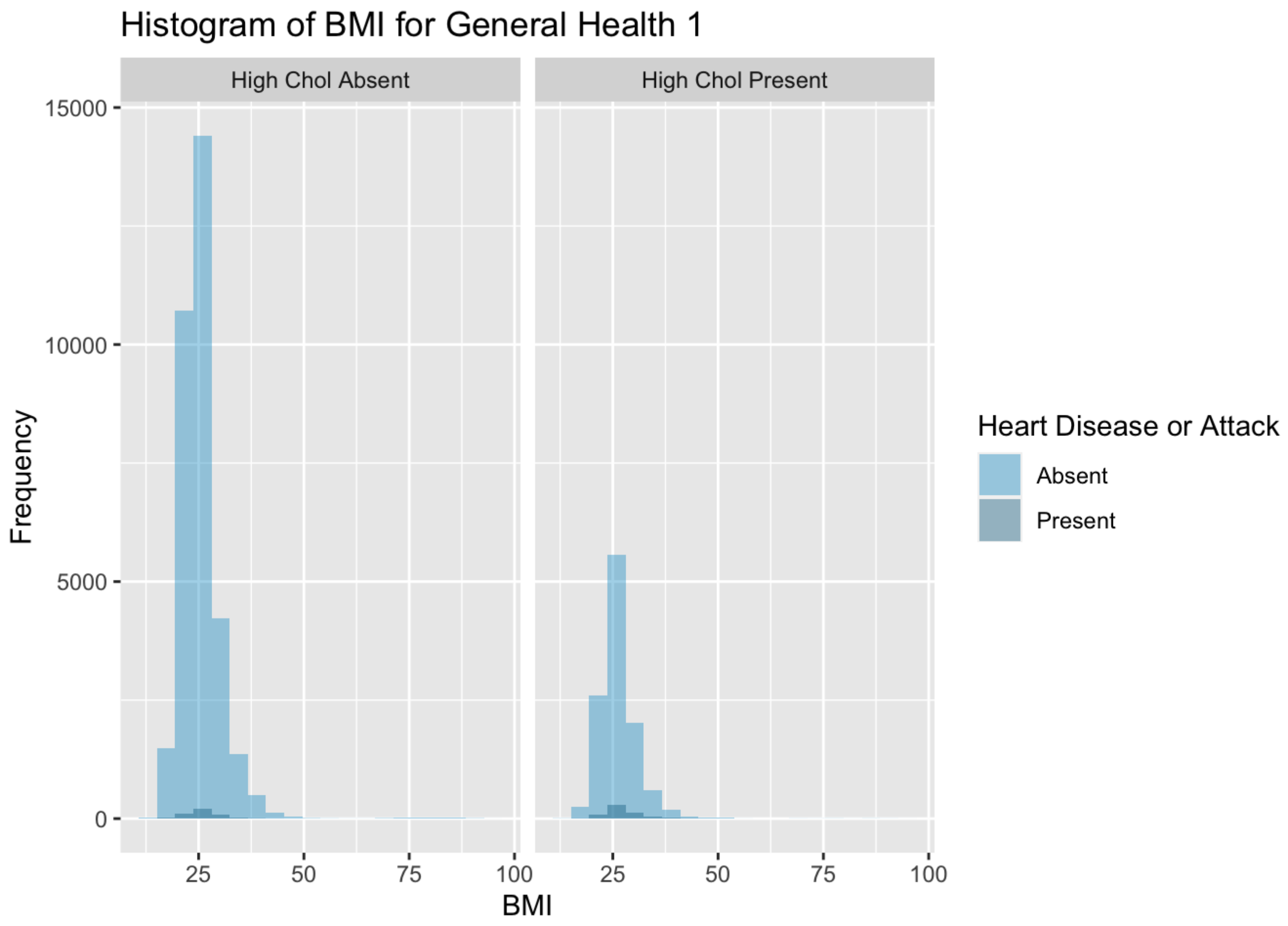
There is higher frequency in high blood pressure absent section and there seems to be a similar number of those with heart disease or attack present in both groups.

```
ggplot(data=sub1, aes(x=BMI,fill = HighBP)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 1", y = "Frequency") +
  scale_fill_manual(name = "High Blood Pressure",
                    values = c("chartreuse2","chartreuse4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



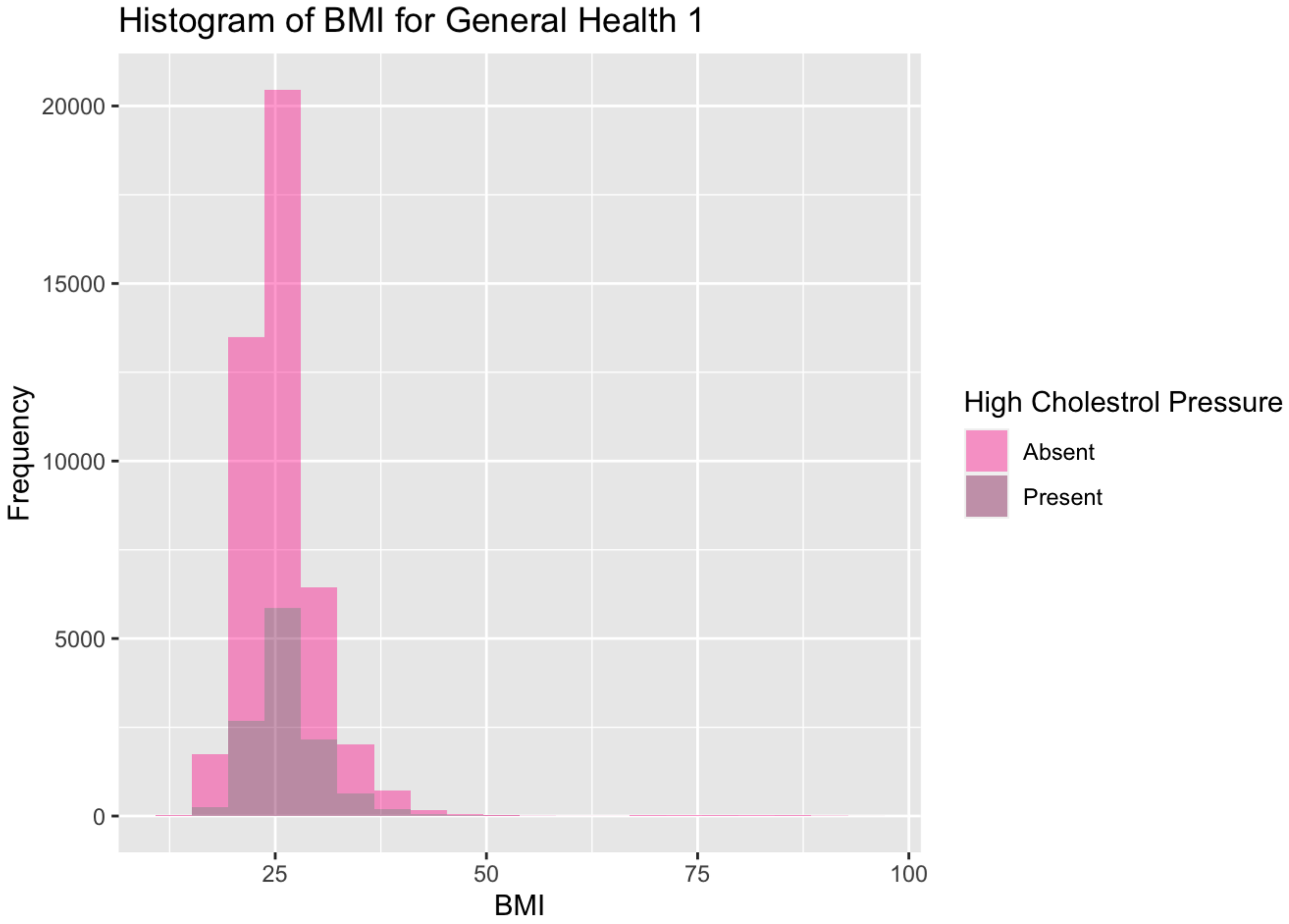
This is a histogram of BMI and Blood Pressure.

```
ggplot(data=sub1, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 1", y = "Frequency") +
  facet_wrap(vars(sub1$HighChol))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram on BMI, Heart Disease, and High Cholesterol.

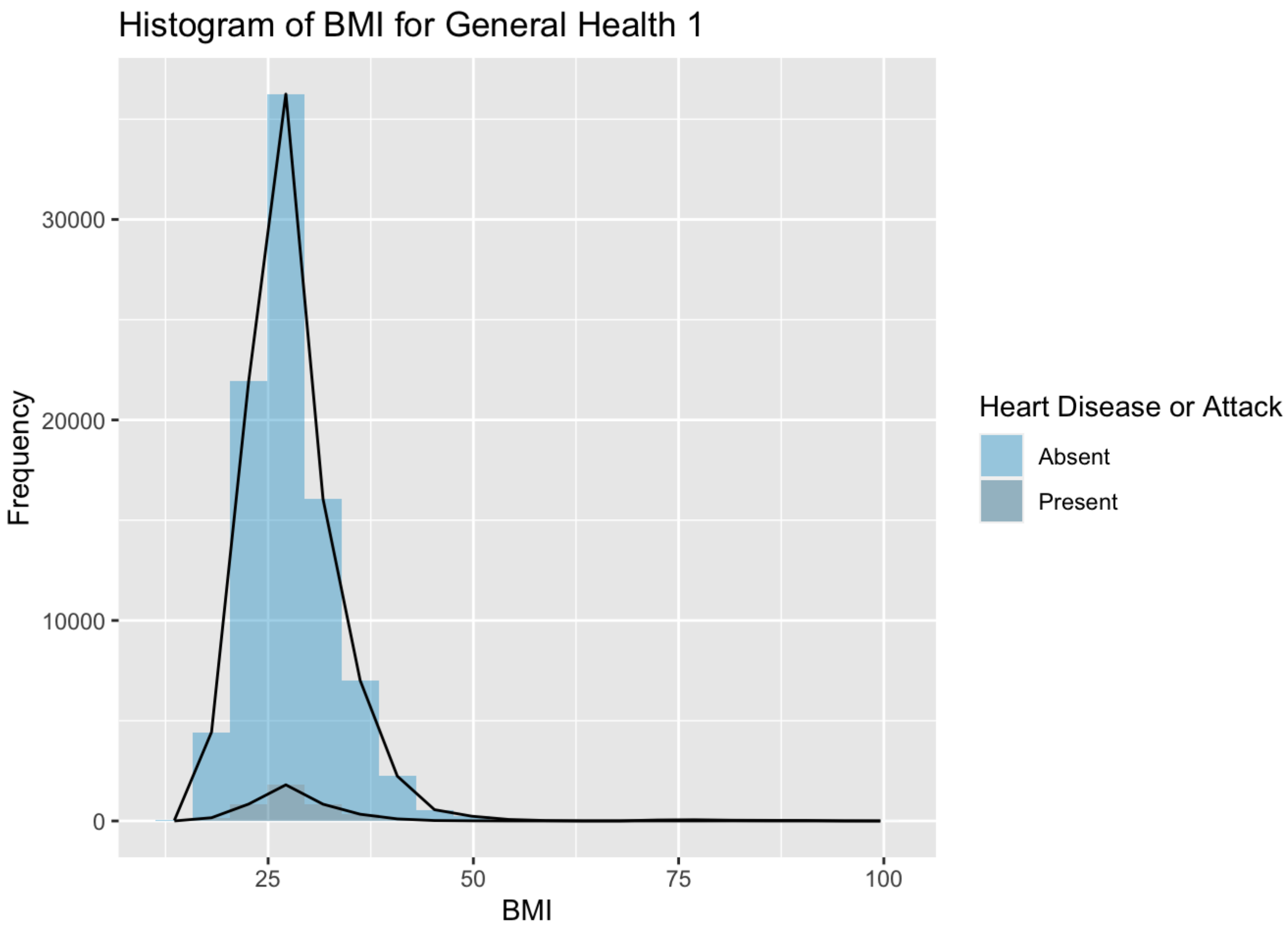
```
ggplot(data=sub1, aes(x=BMI,fill = HighChol)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 1", y = "Frequency") +
  scale_fill_manual(name = "High Cholestrol Pressure",
                    values = c("deeppink","deeppink4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram of BMI and High Cholesterol.

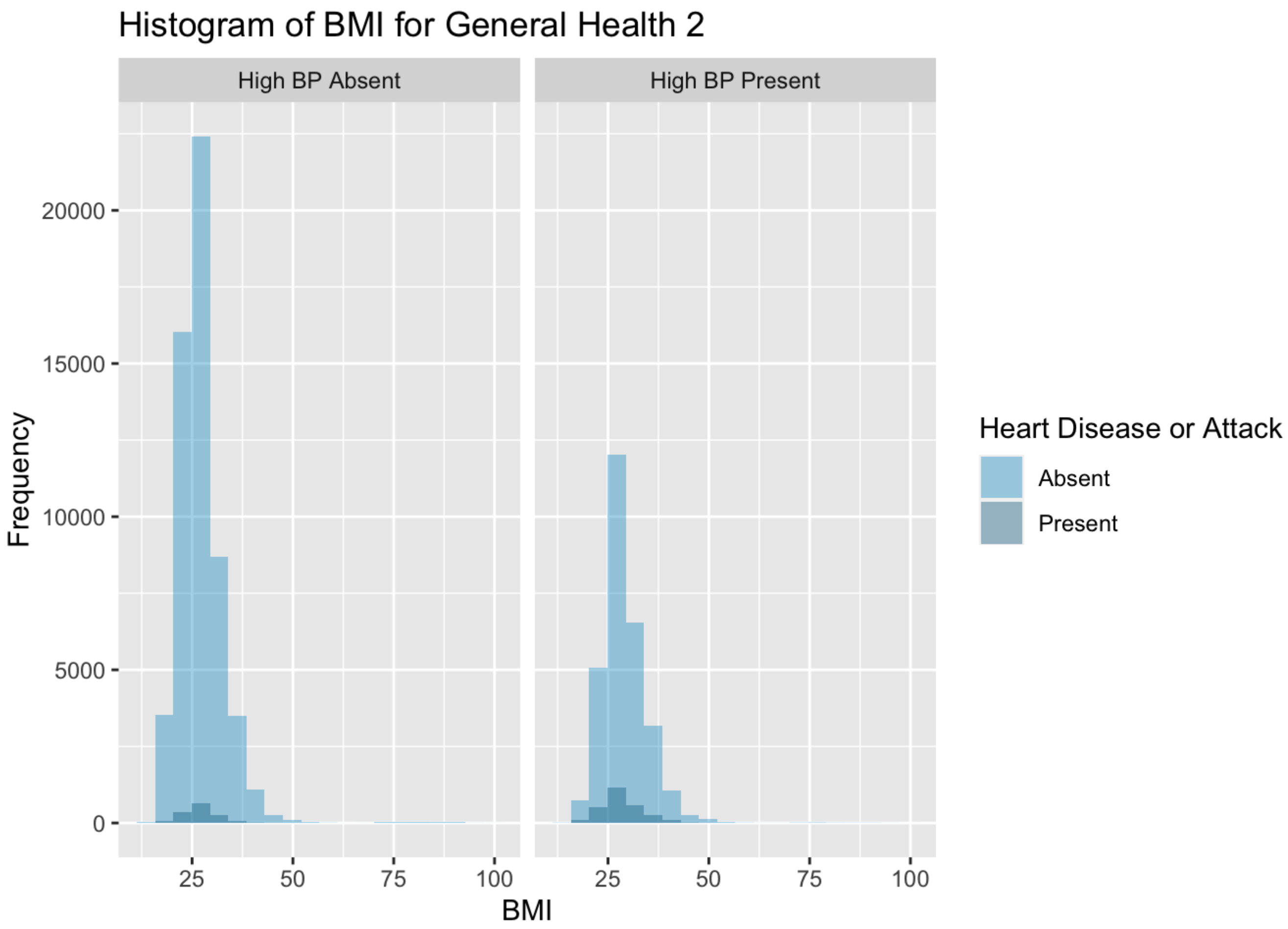
We can create histograms for subset 2 as well which is general health 2.

```
ggplot(data=sub2, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  stat_bin(bins=20, alpha=1, geom="line", aes(y=..count..)) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 1", y = "Frequency") +
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



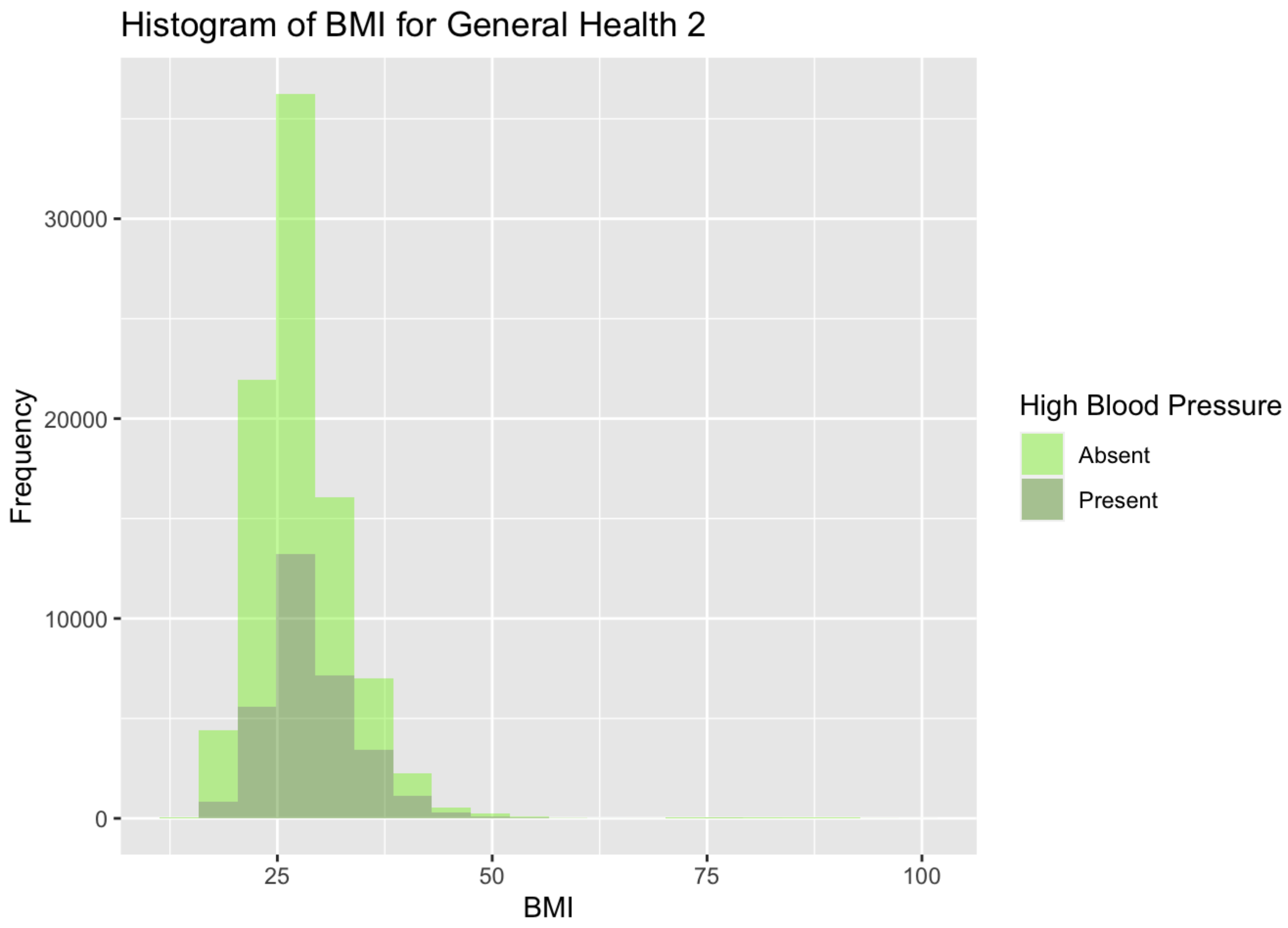
There seems to be a right skew for both those with heart disease and those without.

```
ggplot(data=sub2, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 2", y = "Frequency") +
  facet_wrap(vars(sub2$HighBP))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



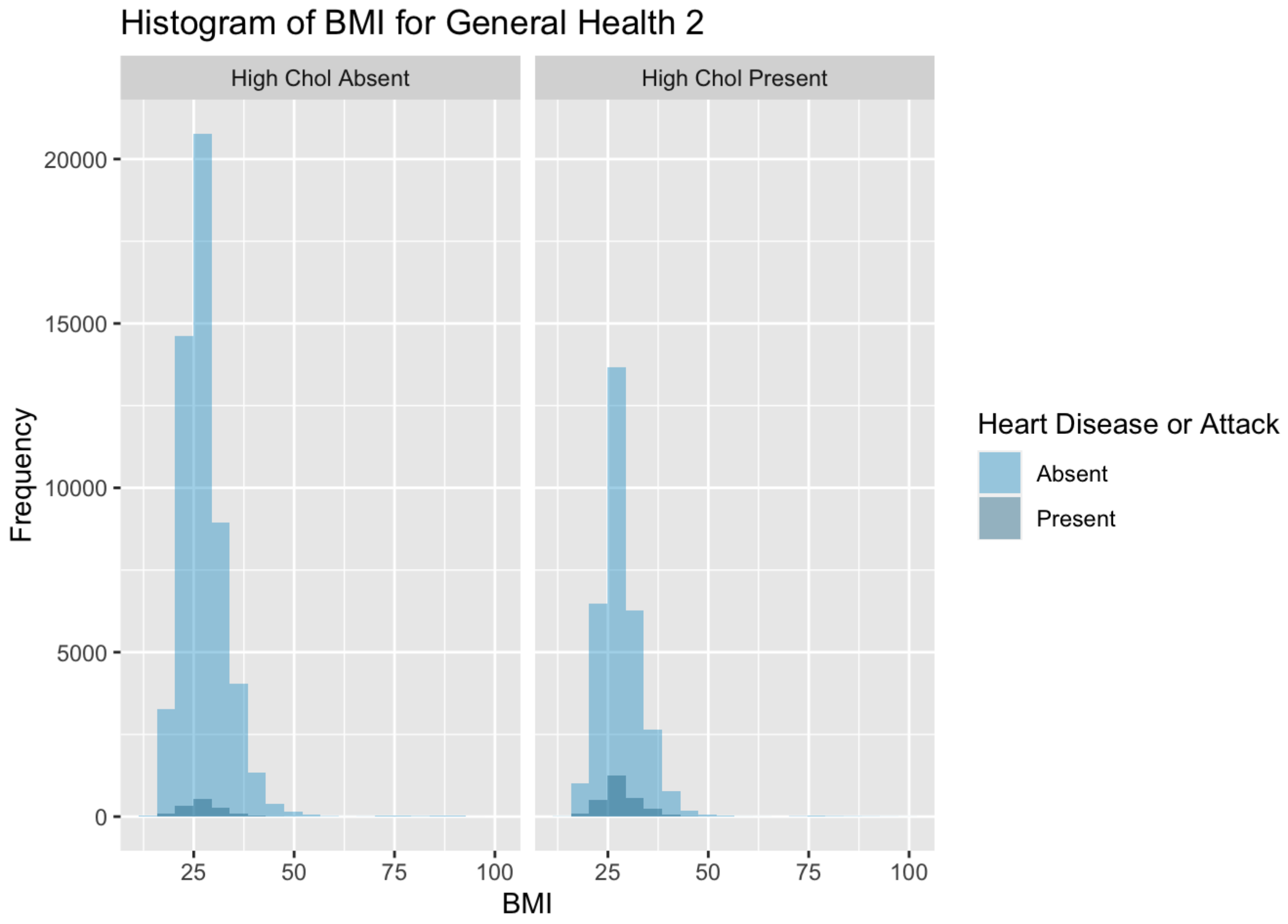
This is a histogram on BMI, Heart Disease, and Blood Pressure.

```
ggplot(data=sub2, aes(x=BMI,fill = HighBP)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 2", y = "Frequency") +
  scale_fill_manual(name = "High Blood Pressure",
                    values = c("chartreuse2","chartreuse4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



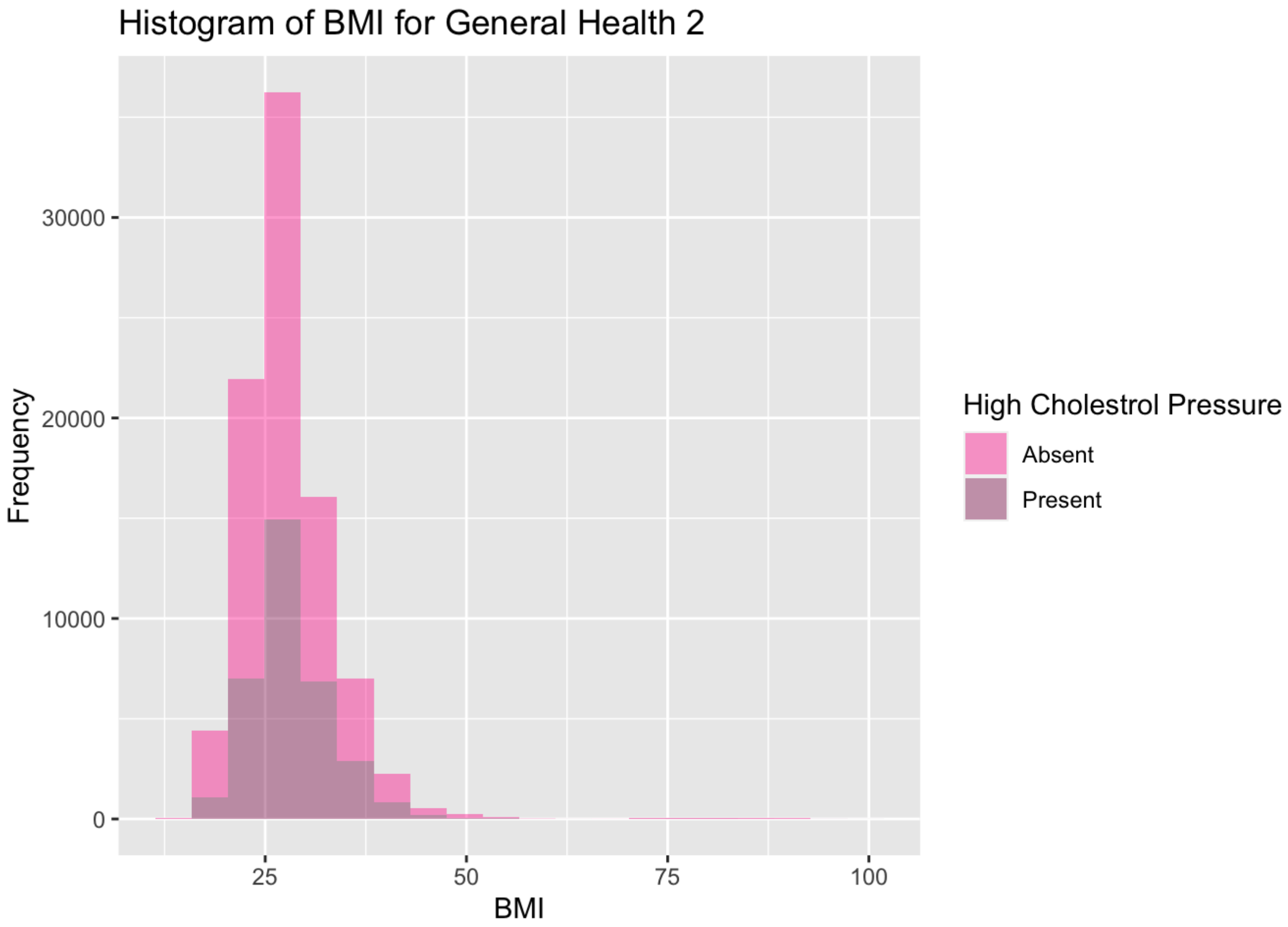
This is a histogram of BMI and Blood Pressure

```
ggplot(data=sub2, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 2", y = "Frequency") +
  facet_wrap(vars(sub2$HighChol))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram on BMI, Heart Disease, and High Cholesterol

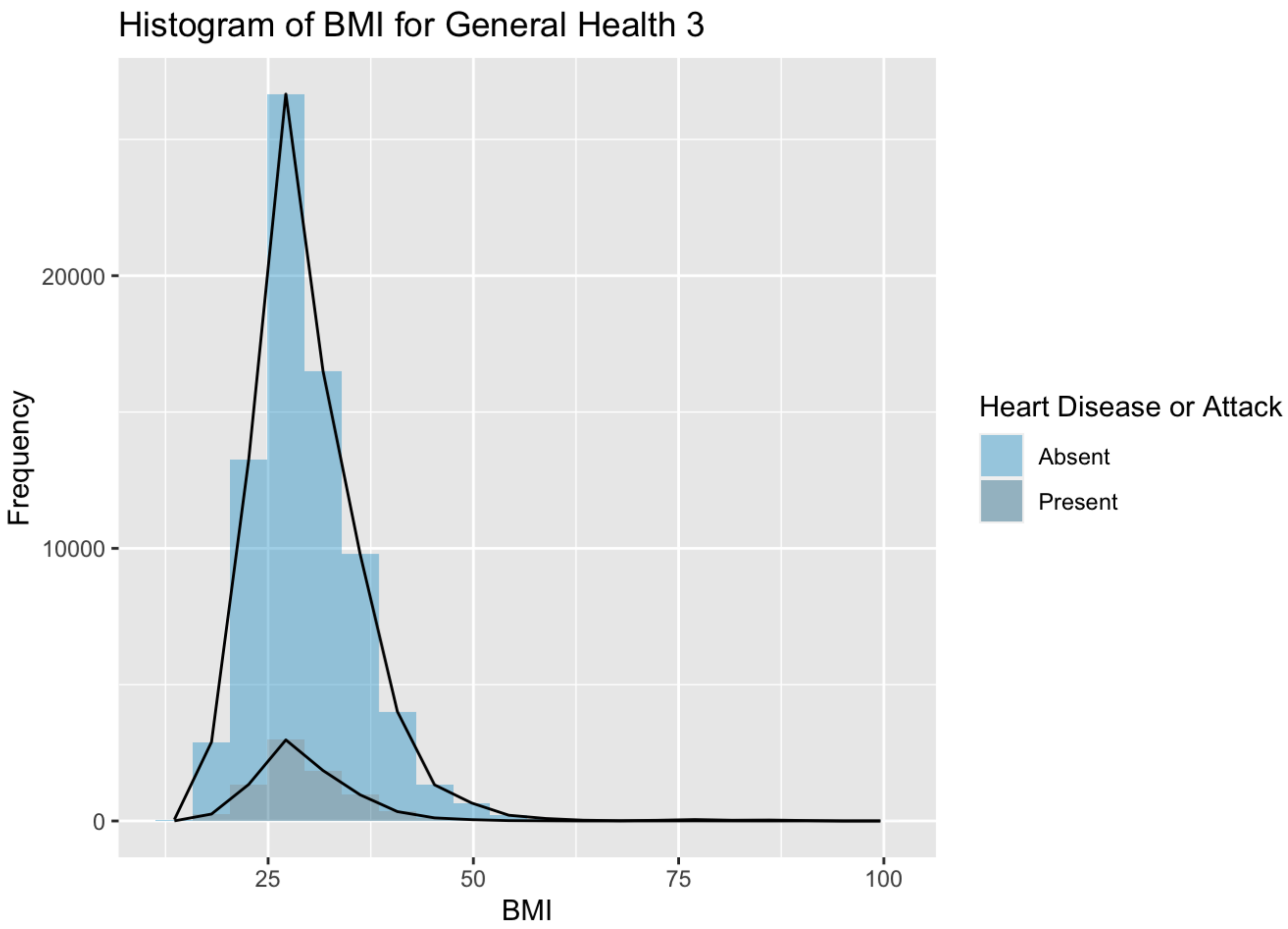
```
ggplot(data=sub2, aes(x=BMI,fill = HighChol)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 2", y = "Frequency") +
  scale_fill_manual(name = "High Cholestrol Pressure",
                    values = c("deeppink","deeppink4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram of BMI and High Cholesterol.

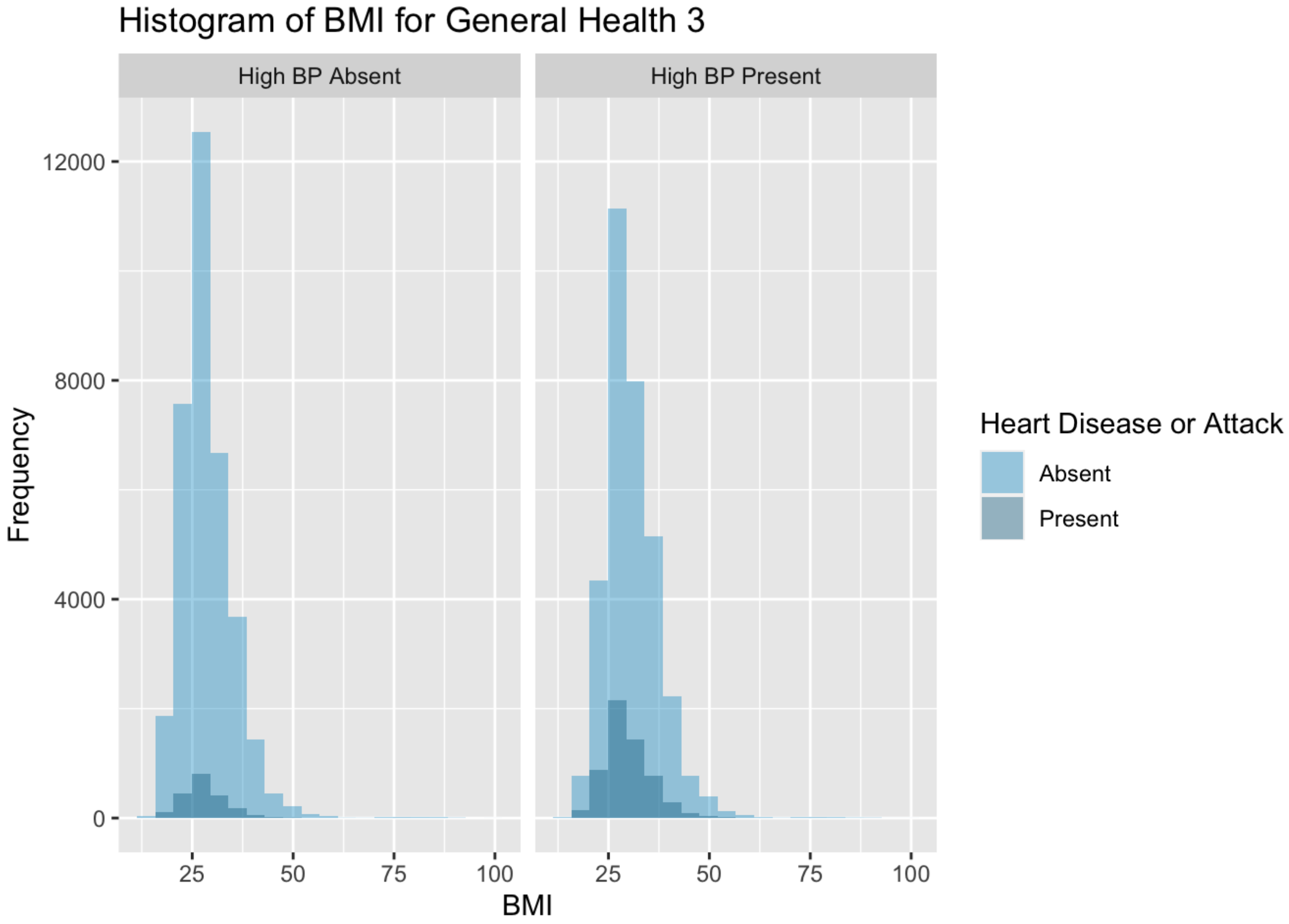
We can create the same histograms for subset 3 which is general health 3.

```
ggplot(data=sub3, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  stat_bin(bins=20, alpha=1, geom="line", aes(y=..count..)) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 3", y = "Frequency") +
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



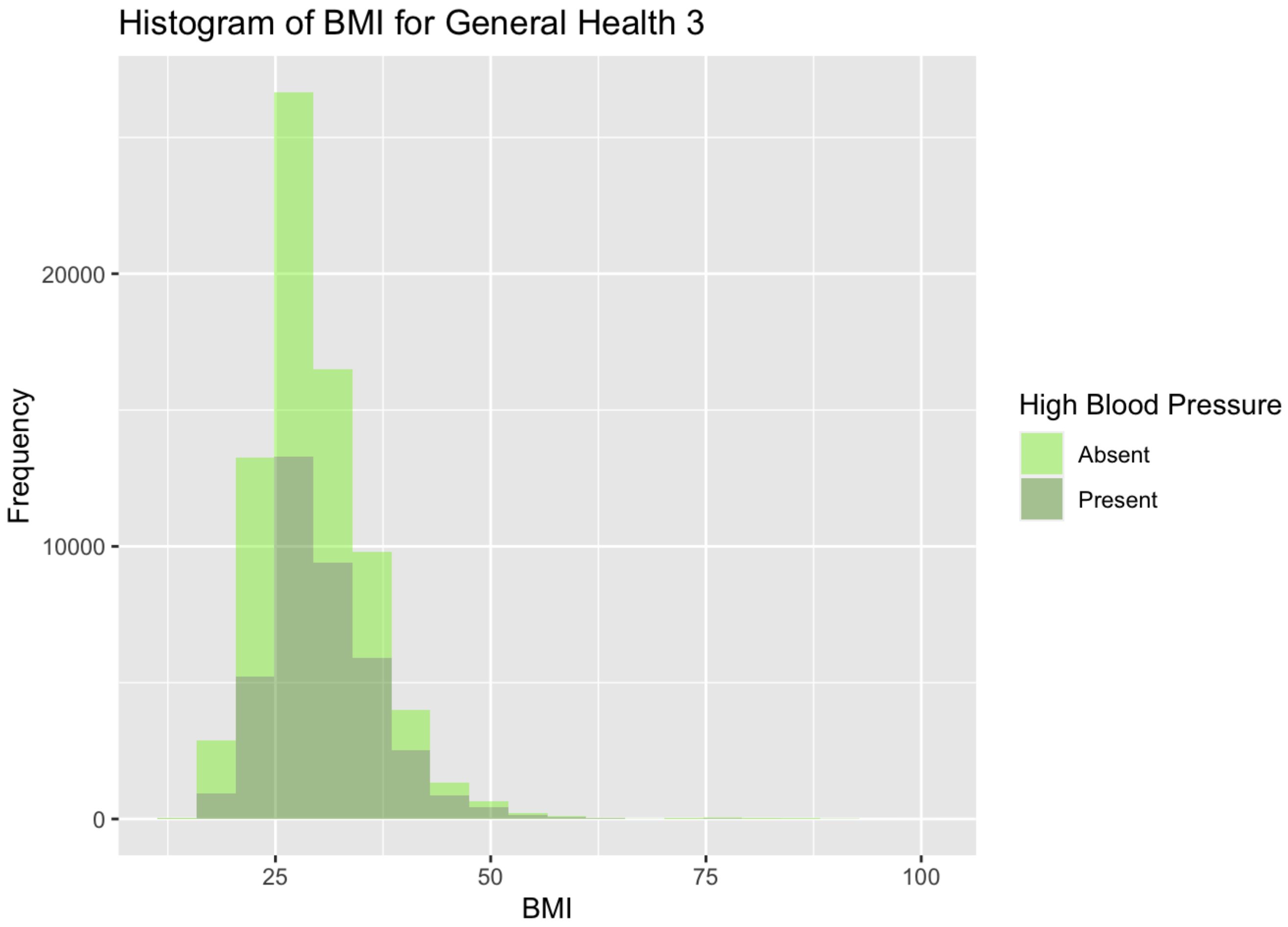
There seems to be a right skew for both those with heart disease and those without.

```
ggplot(data=sub3, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 3", y = "Frequency") +
  facet_wrap(vars(sub3$HighBP))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



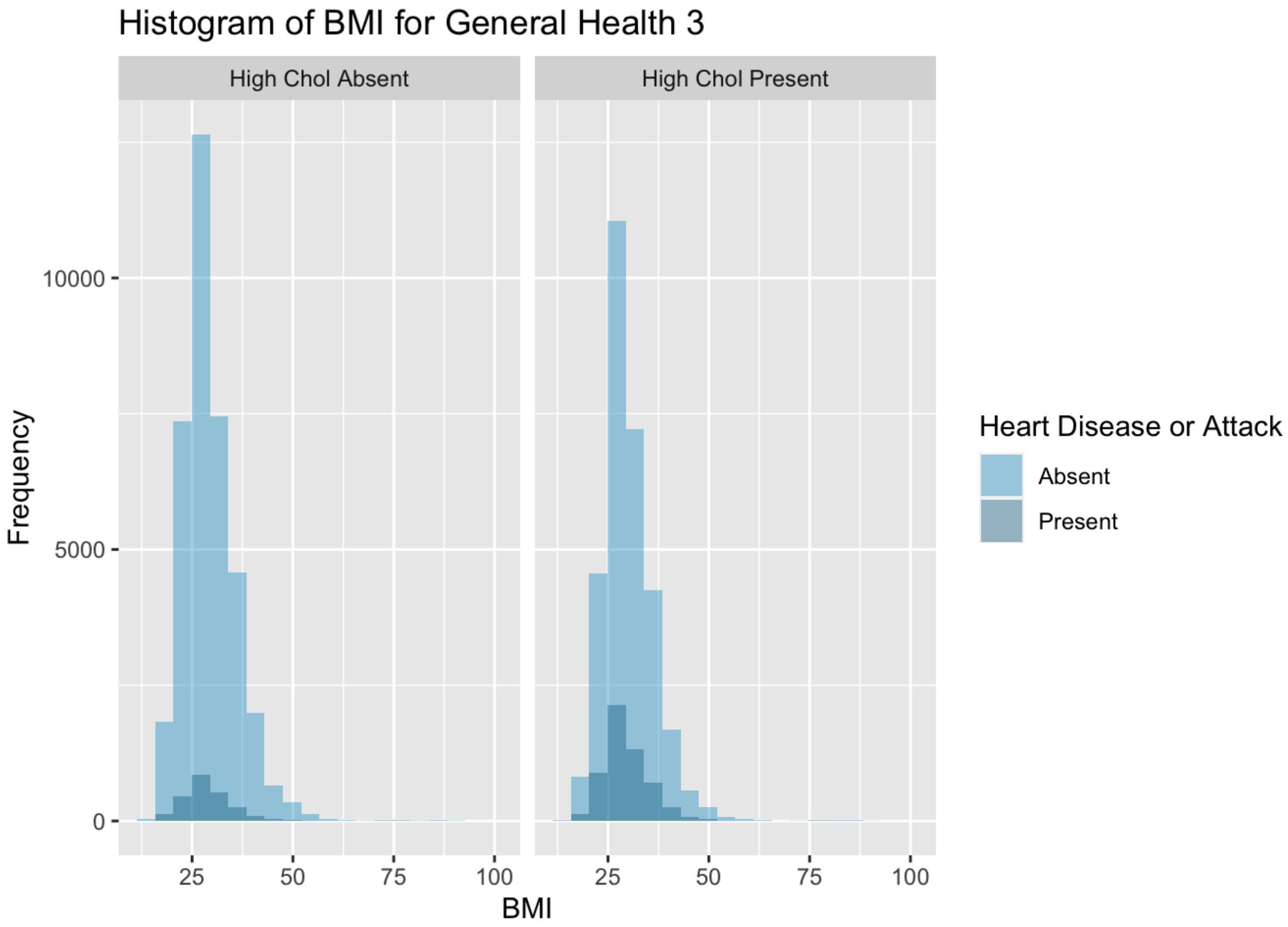
This is a histogram on BMI, Heart Disease, and Blood Pressure.

```
ggplot(data=sub3, aes(x=BMI,fill = HighBP)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 3", y = "Frequency") +
  scale_fill_manual(name = "High Blood Pressure",
                    values = c("chartreuse2","chartreuse4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



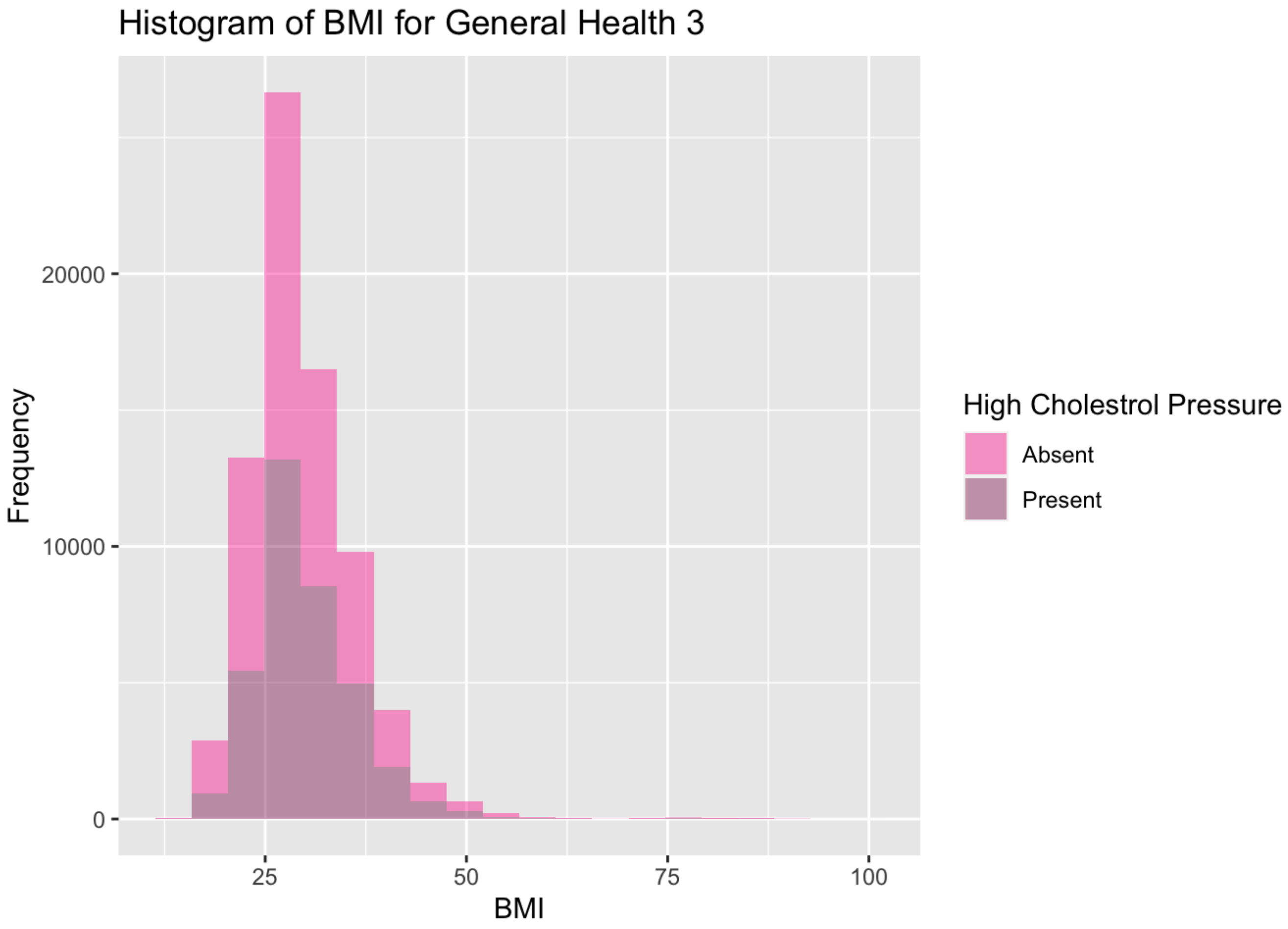
This is a histogram of BMI and Blood Pressure.

```
ggplot(data=sub3, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 3", y = "Frequency") +
  facet_wrap(vars(sub3$HighChol))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram on BMI, Heart Disease, and High Cholesterol.

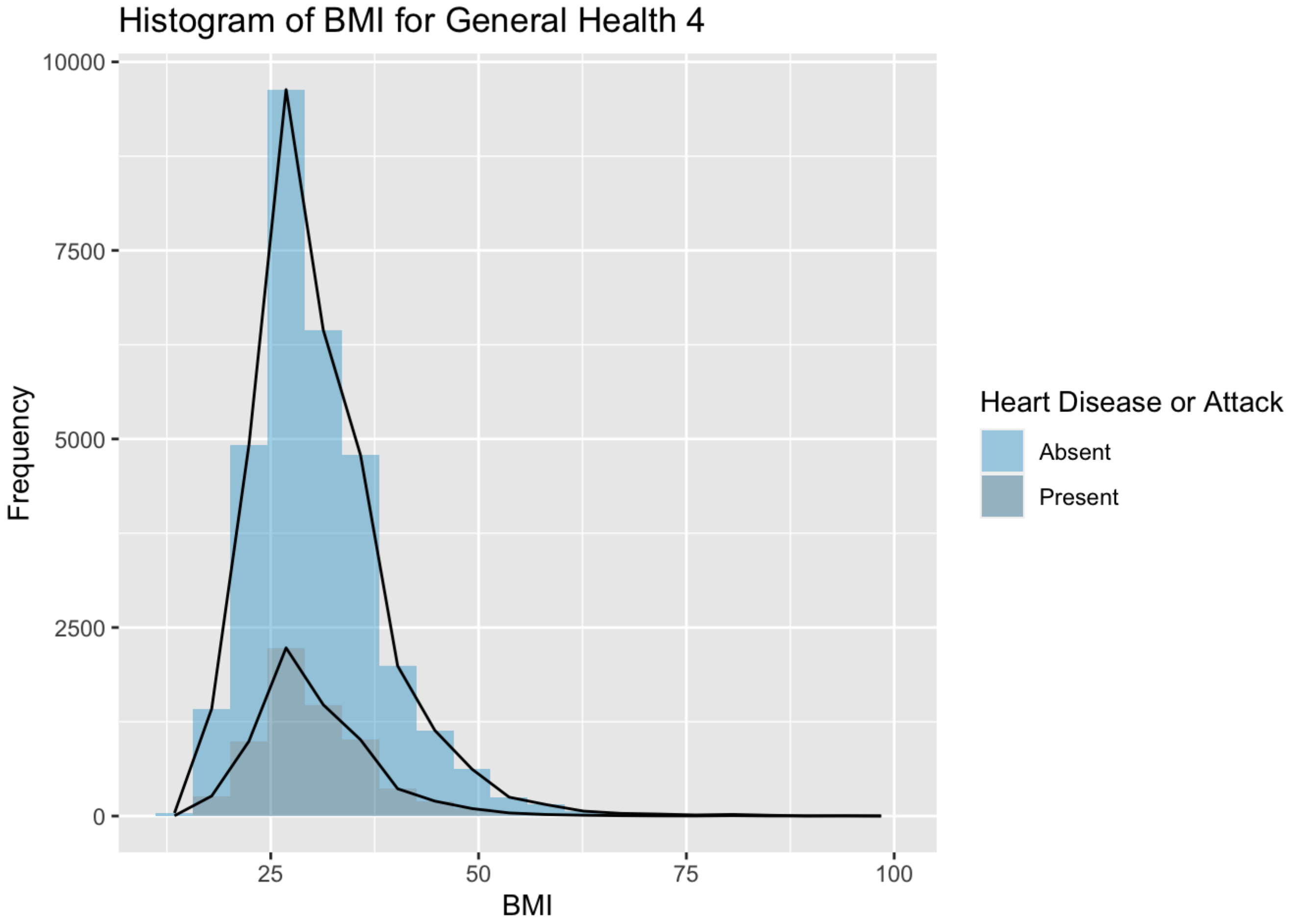
```
ggplot(data=sub3, aes(x=BMI,fill = HighChol)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 3", y = "Frequency") +
  scale_fill_manual(name = "High Cholestrol Pressure",
                    values = c("deeppink","deeppink4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram of BMI and High Cholesterol.

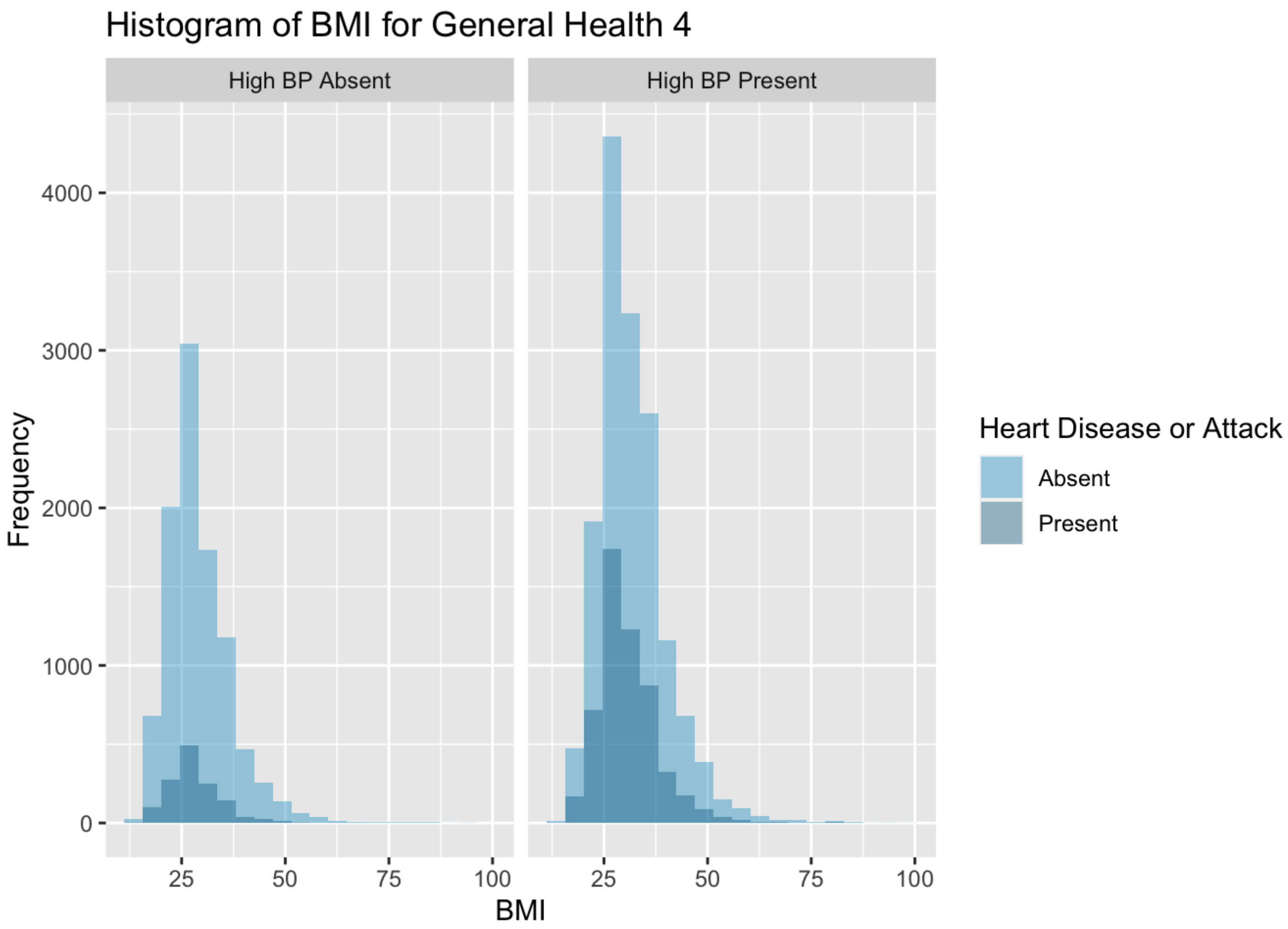
These are the histograms for subset 4 which is general health 4.

```
ggplot(data=sub4, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  stat_bin(bins=20, alpha=1, geom="line", aes(y=..count..)) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 4", y = "Frequency") +
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



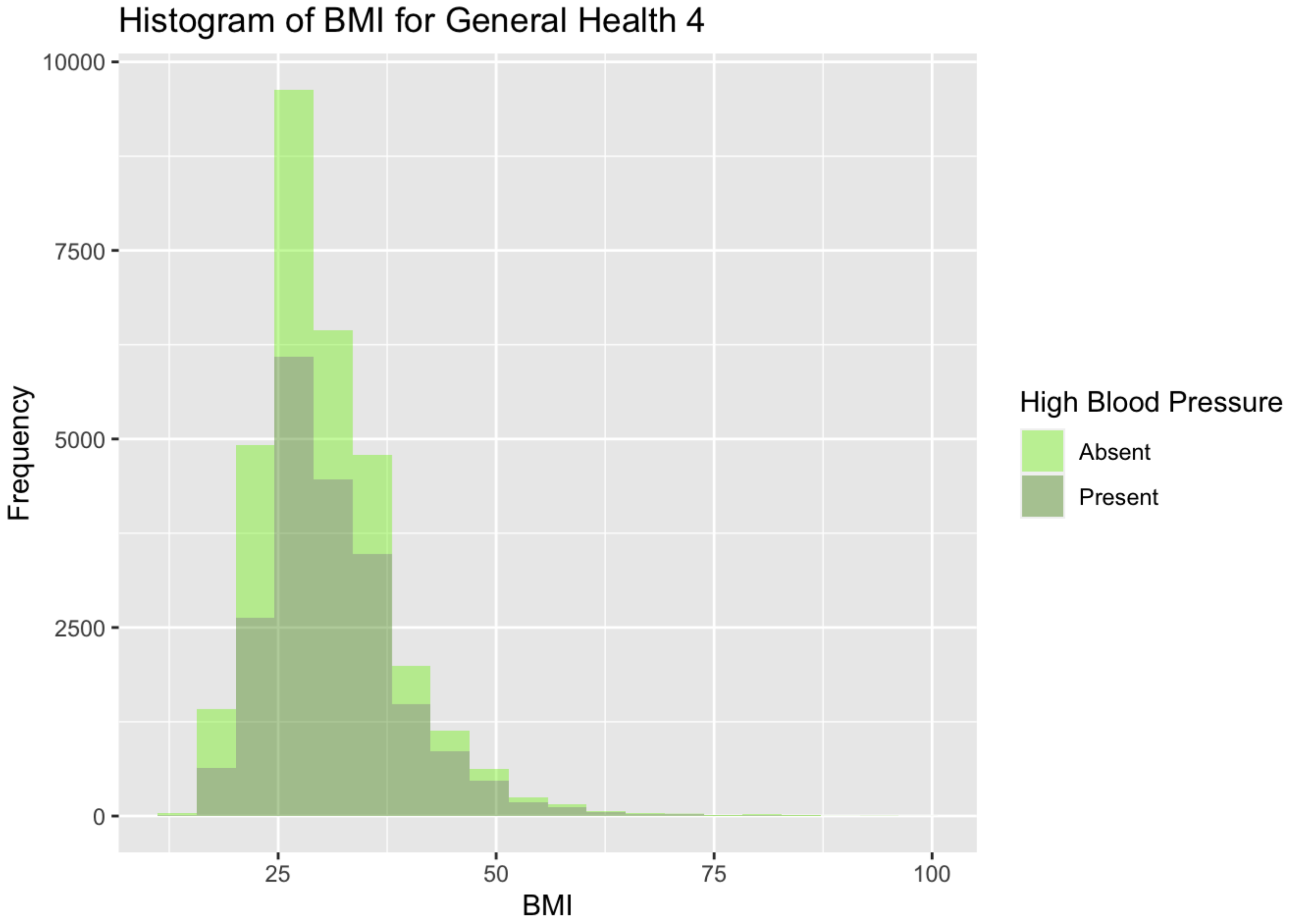
There seems to be a right skew for both those with heart disease and those without.

```
ggplot(data=sub4, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 4", y = "Frequency") +
  facet_wrap(vars(sub4$HighBP))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



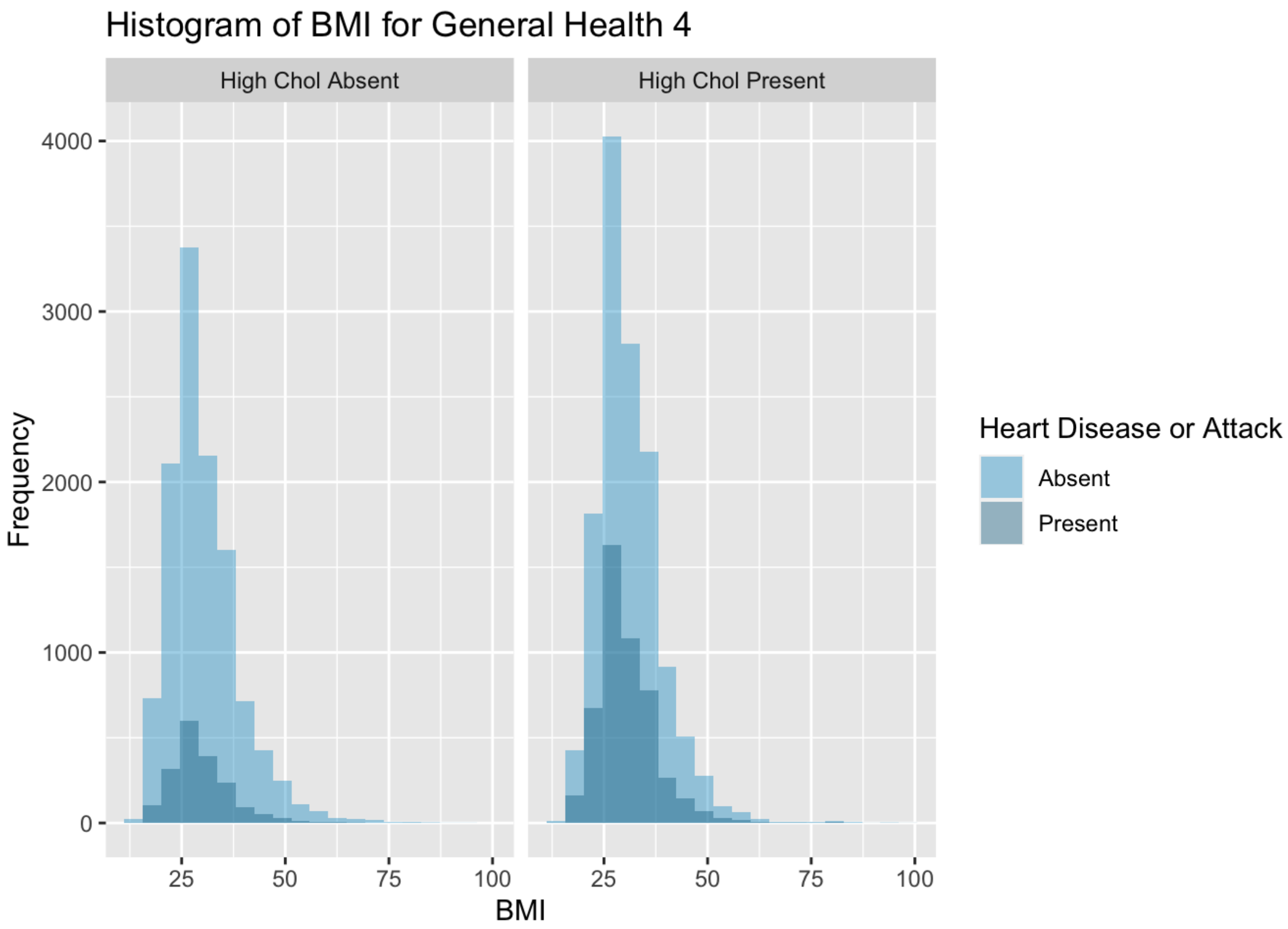
This is a histogram on BMI, Heart Disease, and Blood Pressure.

```
ggplot(data=sub4, aes(x=BMI,fill = HighBP)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 4", y = "Frequency") +
  scale_fill_manual(name = "High Blood Pressure",
                    values = c("chartreuse2","chartreuse4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



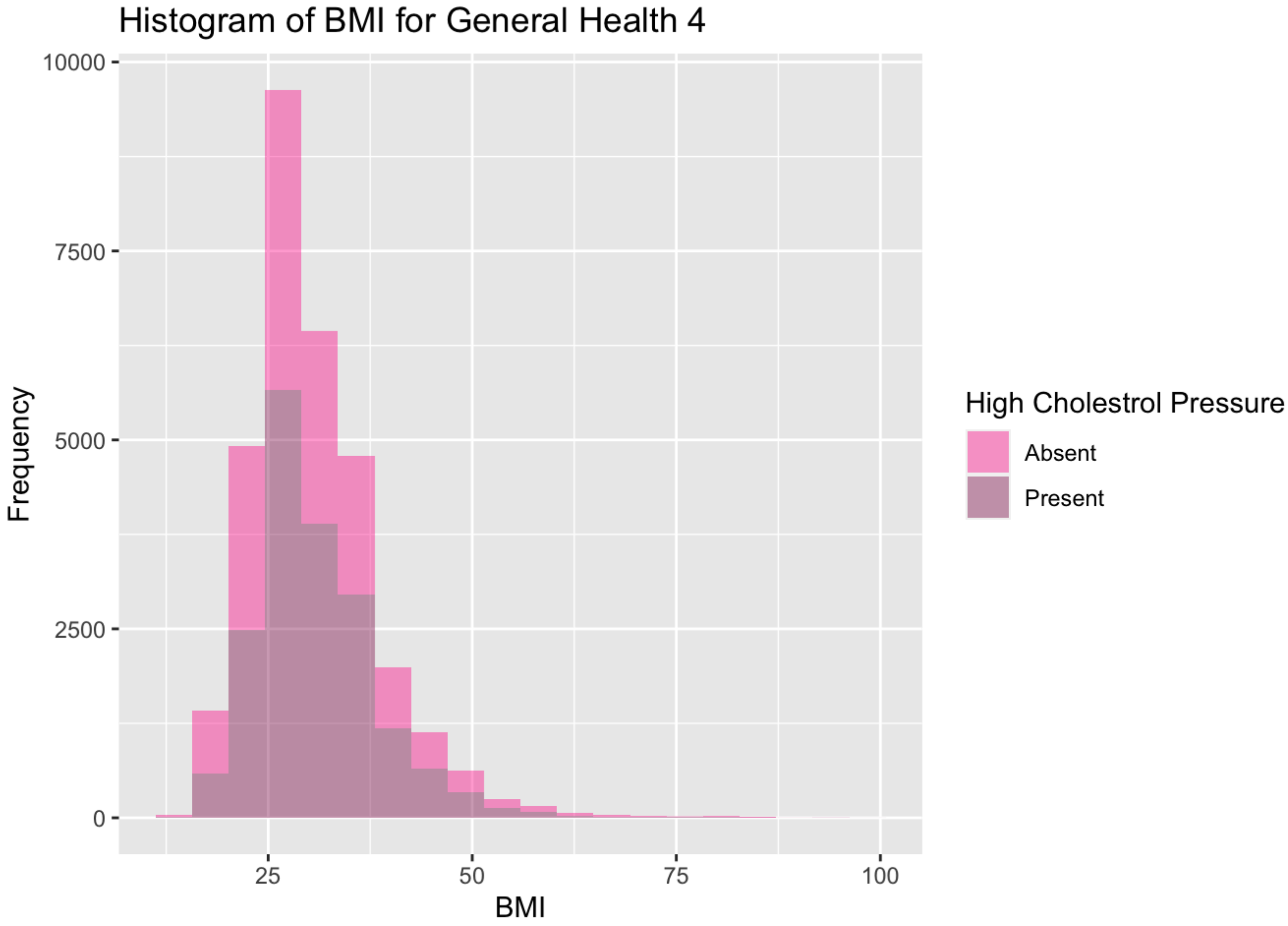
This is a histogram of BMI and Blood Pressure.

```
ggplot(data=sub4, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 4", y = "Frequency") +
  facet_wrap(vars(sub4$HighChol))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram on BMI, Heart Disease, and High Cholesterol.

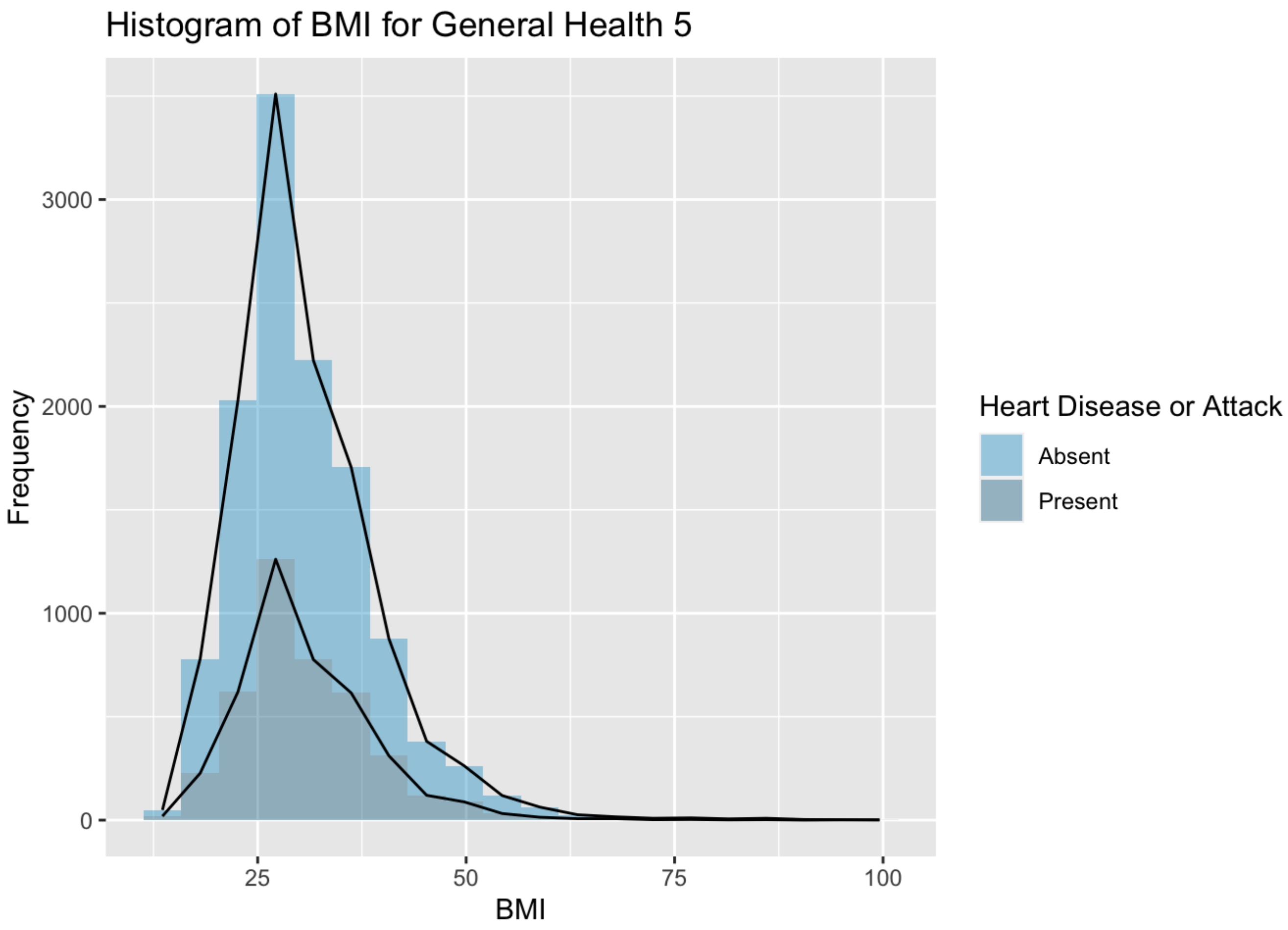
```
ggplot(data=sub4, aes(x=BMI,fill = HighChol)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 4", y = "Frequency") +
  scale_fill_manual(name = "High Cholestrol Pressure",
                    values = c("deeppink","deeppink4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram of BMI and High Cholesterol.

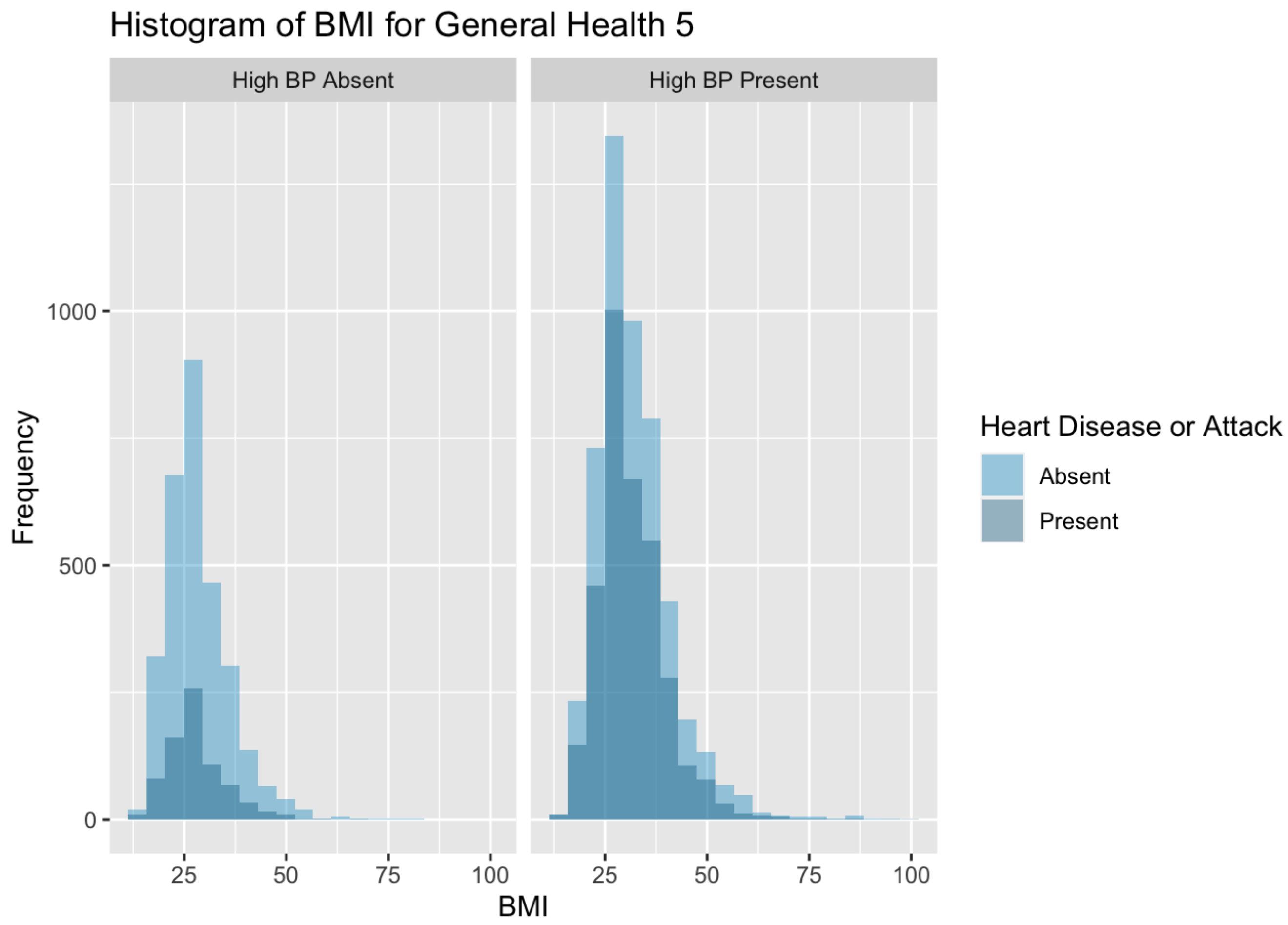
These are the histograms for general health category five.

```
ggplot(data=sub5, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  stat_bin(bins=20, alpha=1, geom="line", aes(y=..count..)) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 5", y = "Frequency") +
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



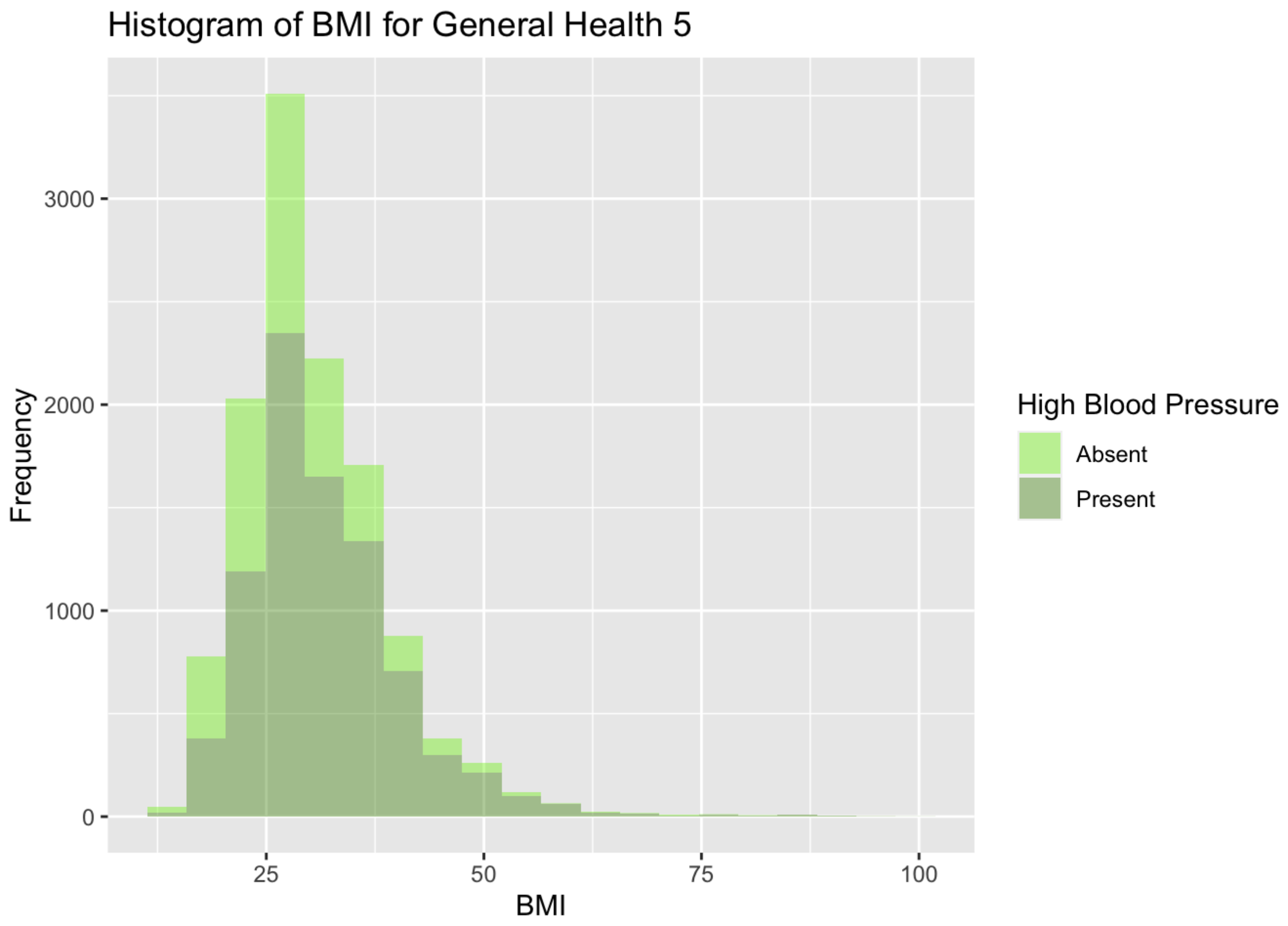
There seems to be a right skew for both those with heart disease and those without.

```
ggplot(data=sub5, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 5", y = "Frequency") +
  facet_wrap(vars(sub5$HighBP))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



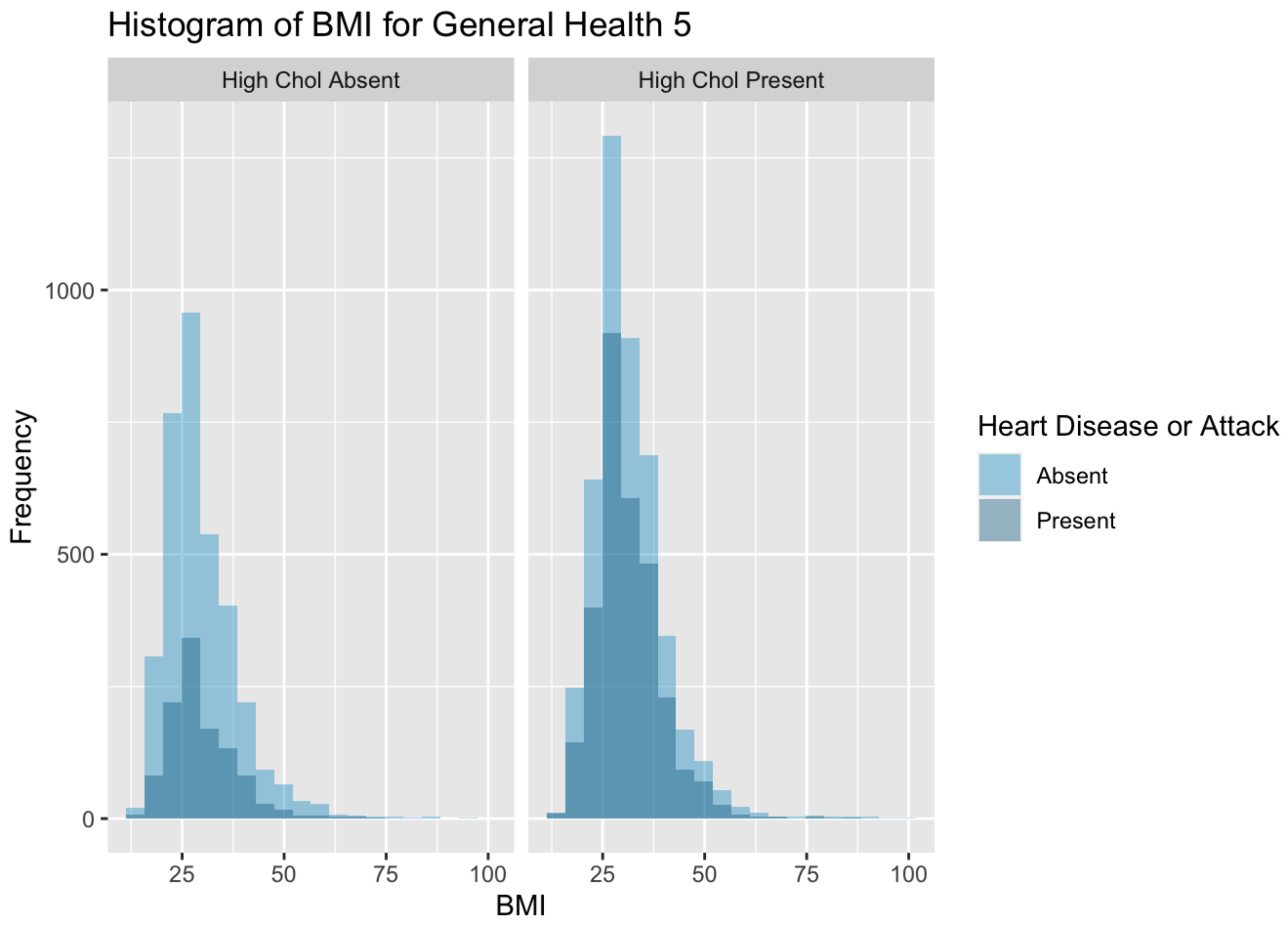
This is a histogram on BMI, Heart Disease, and Blood Pressure.

```
ggplot(data=sub5, aes(x=BMI,fill = HighBP)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 5", y = "Frequency") +
  scale_fill_manual(name = "High Blood Pressure",
                    values = c("chartreuse2","chartreuse4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



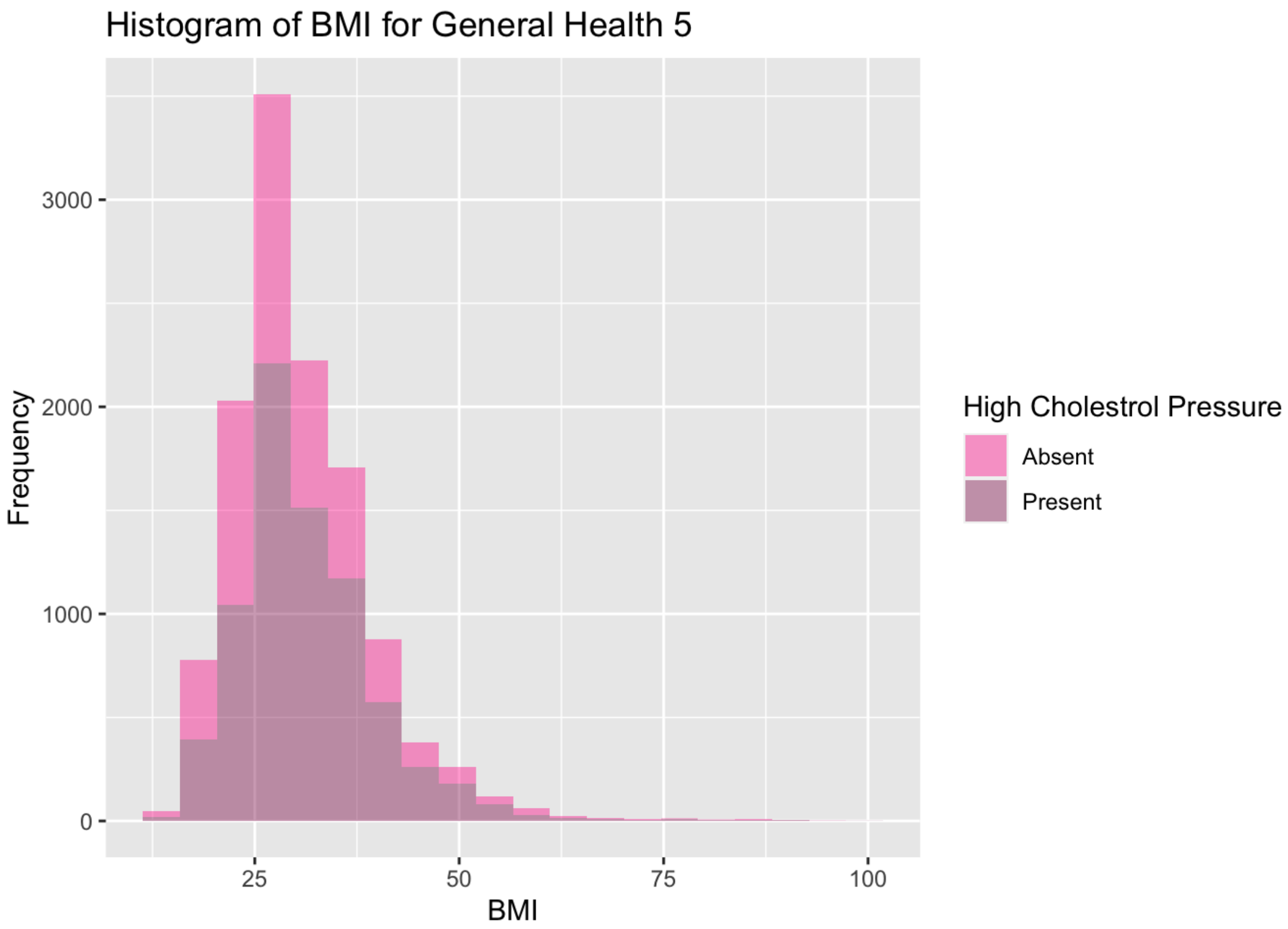
This is a histogram of BMI and Blood Pressure

```
ggplot(data=sub5, aes(x=BMI,fill = HeartDiseaseorAttack)) +
  geom_histogram(bins=20,position = "identity", alpha = 0.4) +
  labs(x = "BMI", title = "Histogram of BMI for General Health 5", y = "Frequency") +
  facet_wrap(vars(sub5$HighChol))+
  scale_fill_manual(name = "Heart Disease or Attack",
                    values = c("deepskyblue3","deepskyblue4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram on BMI, Heart Disease, and High Cholesterol

```
ggplot(data=sub5, aes(x=BMI,fill = HighChol)) +
  stat_bin(bins=20, alpha=0.4, geom="bar") +
  labs(x = "BMI", title = "Histogram of BMI for General Health 5", y = "Frequency") +
  scale_fill_manual(name = "High Cholestrol Pressure",
                    values = c("deeppink","deeppink4"),
                    labels = c("Absent","Present"),
                    aesthetics = "fill")
```



This is a histogram of BMI and High Cholesterol.

The right skewness in the histograms of BMI indicates that the majority of the participants have a lower BMI value, while a smaller proportion has a higher BMI value. This suggests that the data is not distributed normally and that there may be a group of individuals with a higher BMI than the rest. This could mean that the sample may have a higher prevalence of overweight or obese individuals, or that there may be a sub population with higher BMI values.

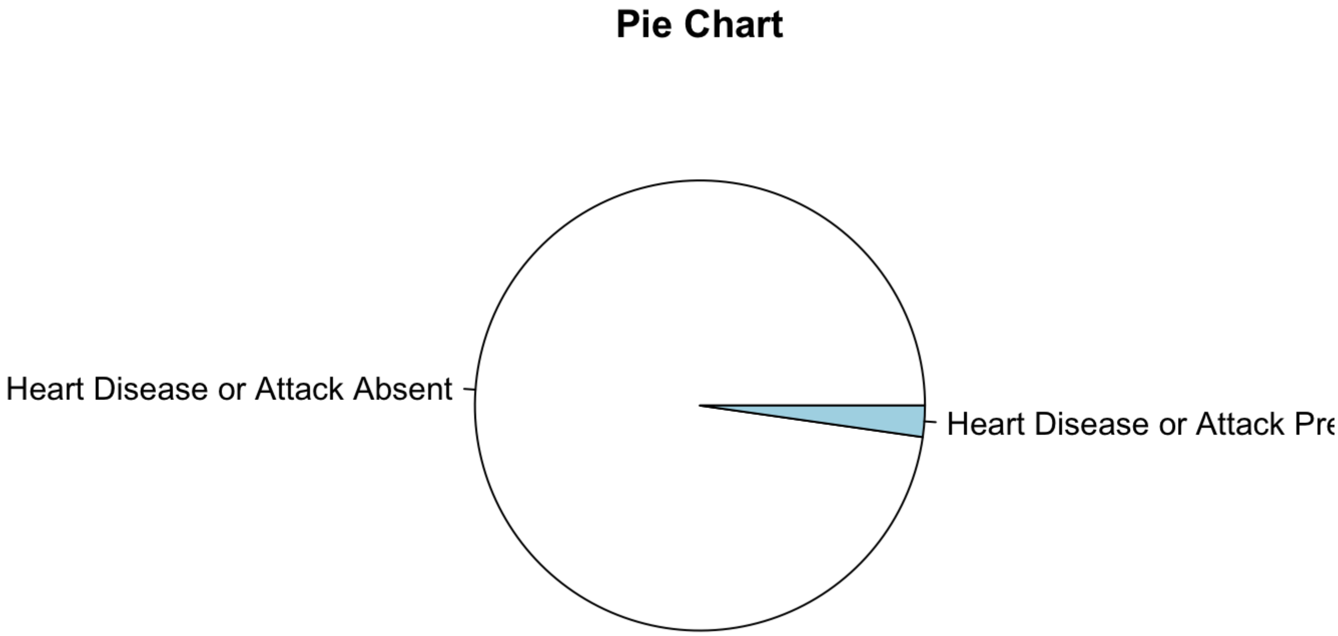
The relationship between the general health scale score and the binary variables for heart disease or attack, high blood pressure, and high cholesterol suggests that individuals with higher general health scores may be more likely to have these health conditions. The larger proportion of individuals with these health conditions among those with higher general health scores like GenHealth 4 and GenHealth 5 may indicate that these individuals have poorer overall health status.

However, it is important to note that these observations are based on correlations in the data and do not necessarily indicate causality. It is also possible that other factors, such as age, sex, lifestyle behaviors, or underlying medical conditions, could contribute to the observed associations. Further analysis, including statistical modeling and controlling for potential confounding factors, may be needed to better understand the relationships between these variables.

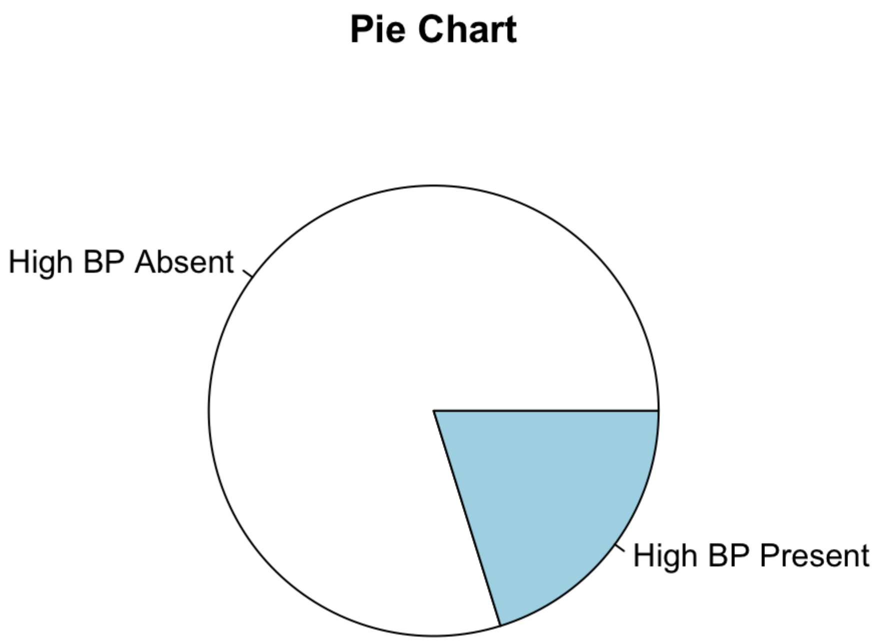
Although we examined histograms to observe the disparity in BMI distributions across binary variables, it’s crucial to acknowledge that the relative frequencies of the binary variables may differ in each subset, which could contribute to the distribution disparities.

Subset 1 Proportions

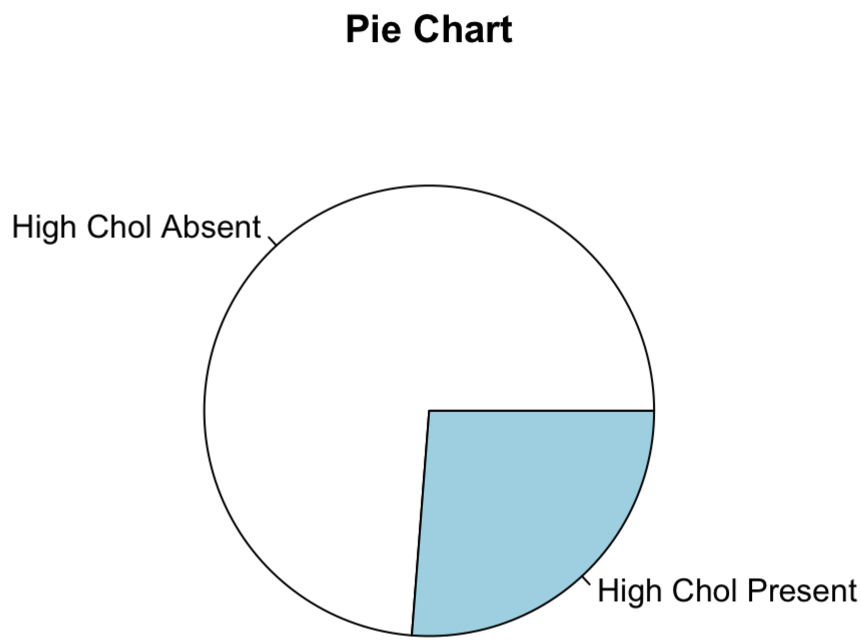
```
s1hp = xtabs(~HeartDiseaseorAttack,data=sub1)
pie(s1hp,
    main = "Pie Chart")
```



```
s1bp = xtabs(~HighBP,data=sub1)
pie(s1bp,
    main = "Pie Chart")
```

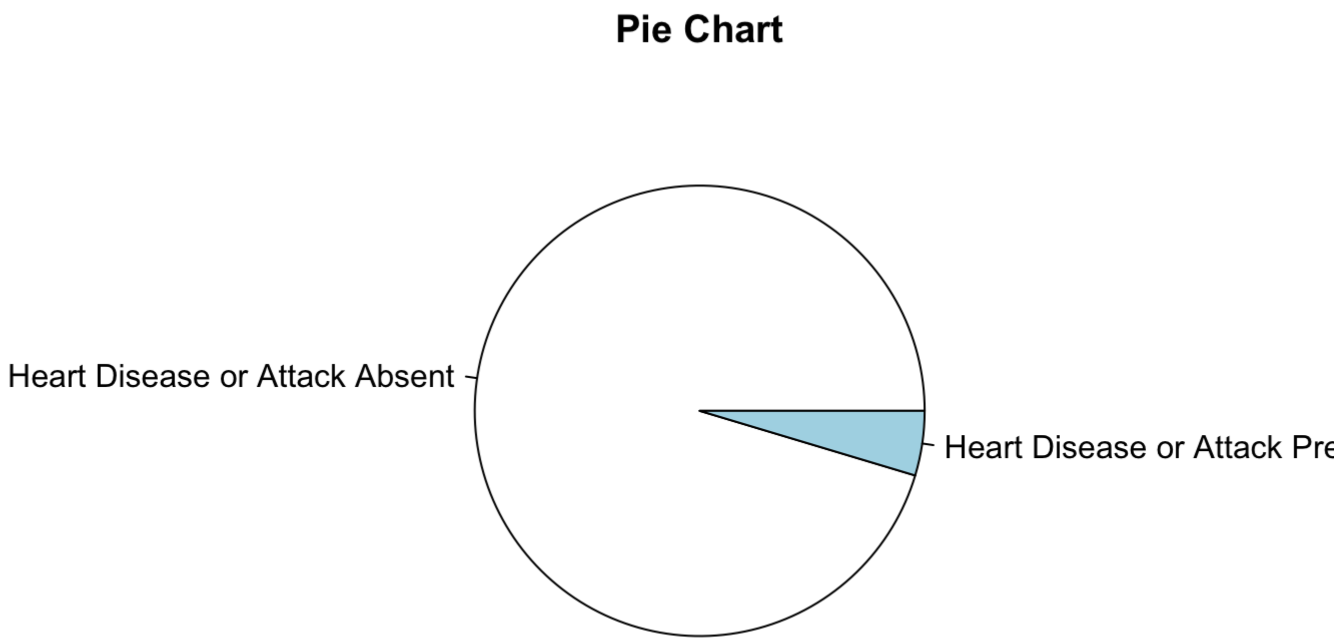


```
s1ch = xtabs(~HighChol,data=sub1)
pie(s1ch,
    main = "Pie Chart")
```

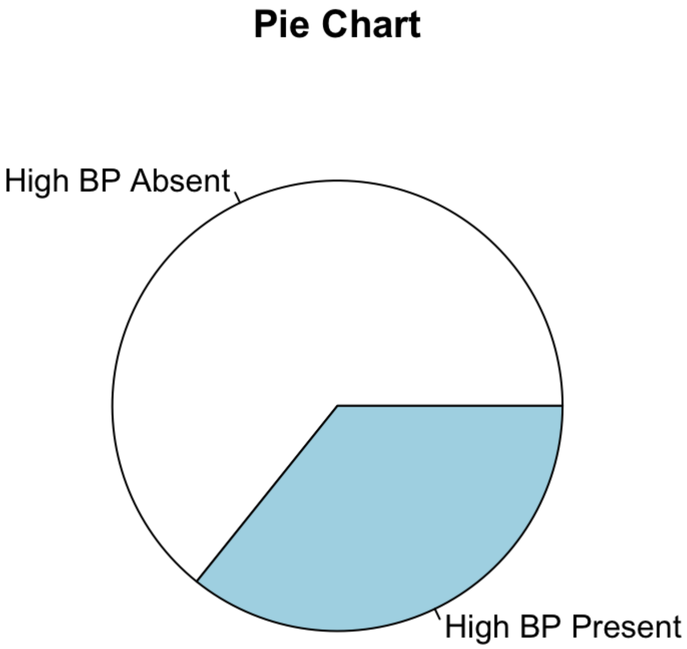


Subset 2 proportions

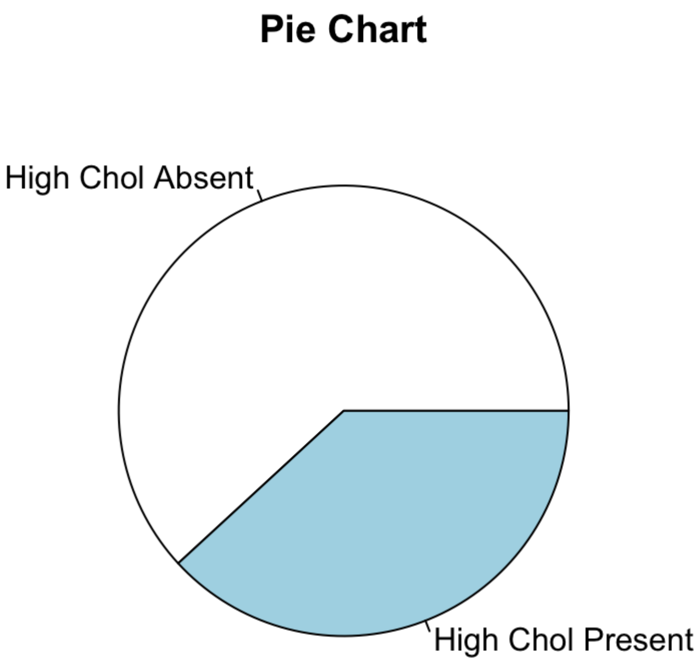
```
s2hp = xtabs(~HeartDiseaseorAttack,data=sub2)
pie(s2hp,
    main = "Pie Chart")
```



```
s2bp = xtabs(~HighBP,data=sub2)
pie(s2bp,
    main = "Pie Chart")
```

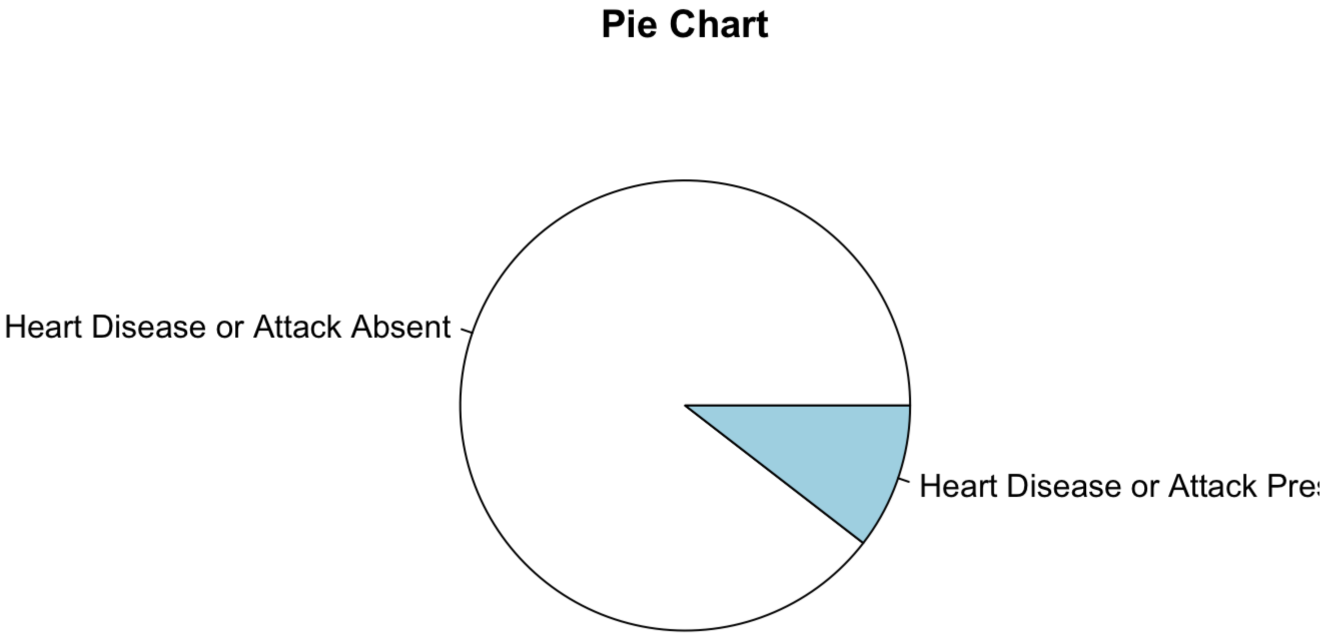


```
s2ch = xtabs(~HighChol,data=sub2)
pie(s2ch,
    main = "Pie Chart")
```

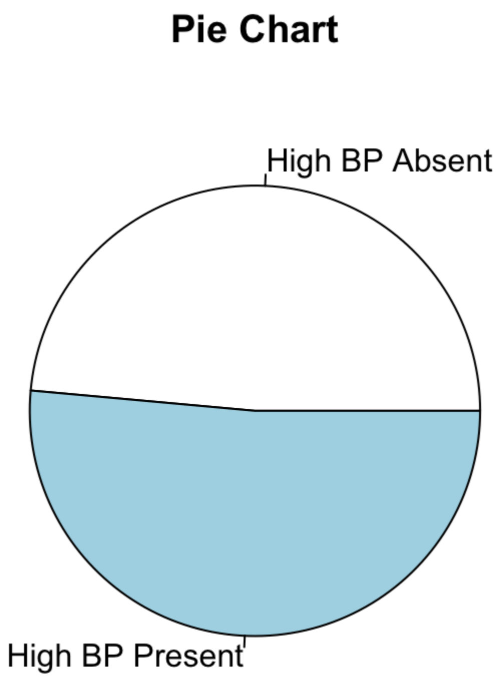


Subset 3 proportions

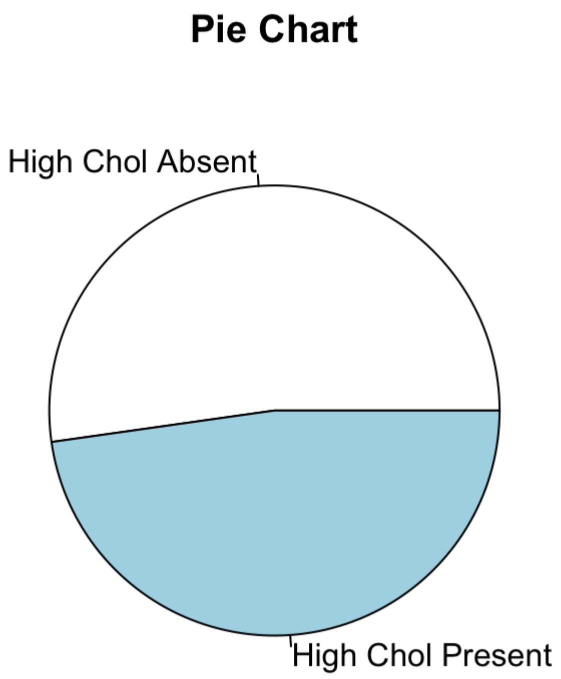
```
s3hp = xtabs(~HeartDiseaseorAttack,data=sub3)
pie(s3hp,
    main = "Pie Chart")
```



```
s3bp = xtabs(~HighBP,data=sub3)
pie(s3bp,
    main = "Pie Chart")
```

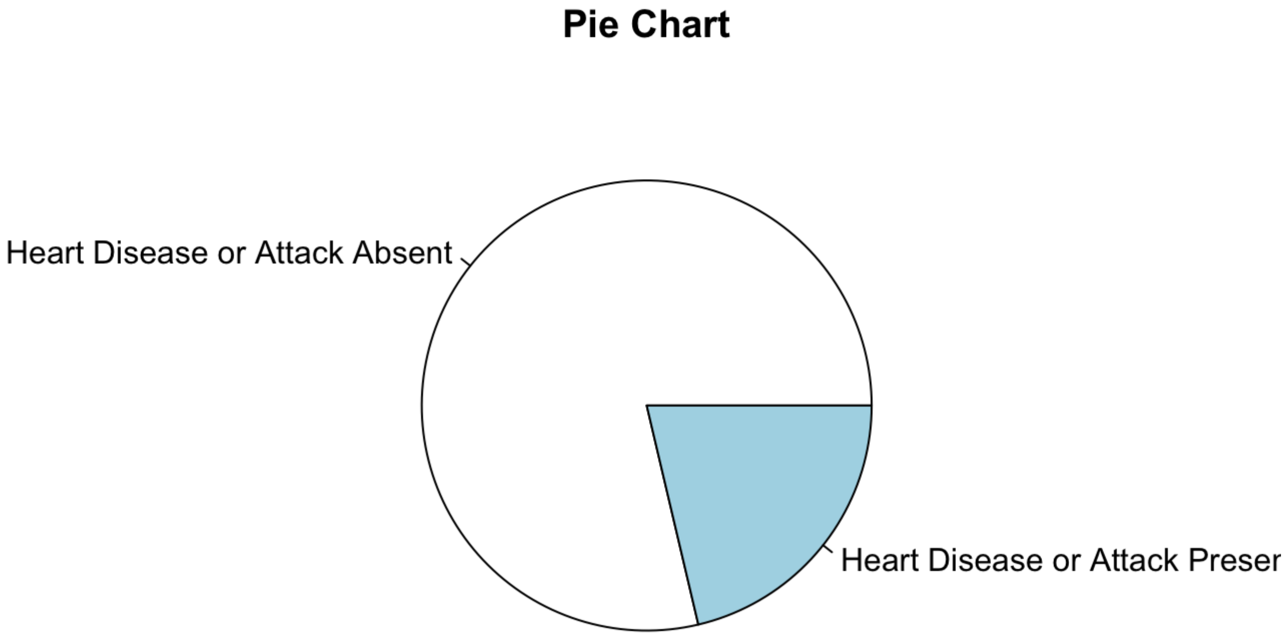


```
s3ch = xtabs(~HighChol,data=sub3)
pie(s3ch,
    main = "Pie Chart")
```

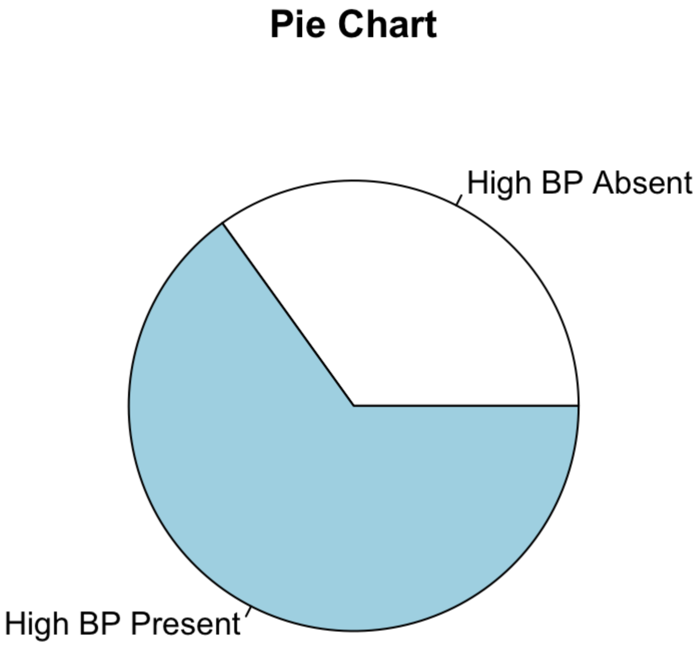


Subset 4 proportions

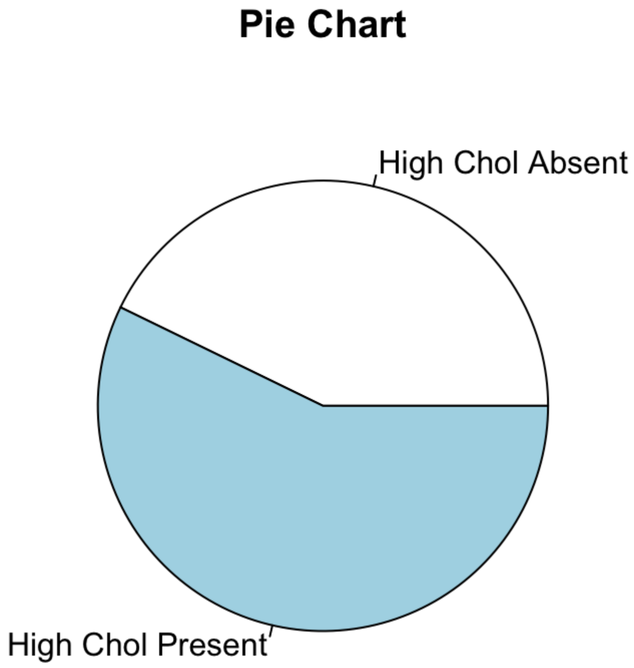
```
s4hp = xtabs(~HeartDiseaseorAttack,data=sub4)
pie(s4hp,
    main = "Pie Chart")
```



```
s4bp = xtabs(~HighBP,data=sub4)
pie(s4bp,
    main = "Pie Chart")
```

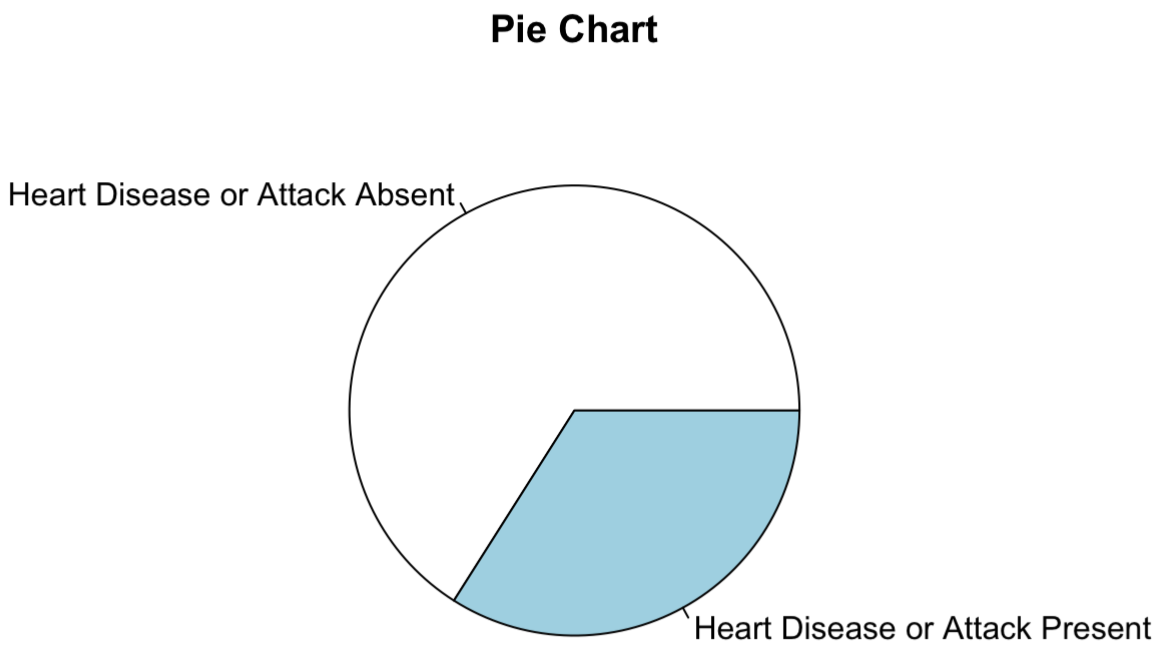


```
s4ch = xtabs(~HighChol,data=sub4)
pie(s4ch,
    main = "Pie Chart")
```

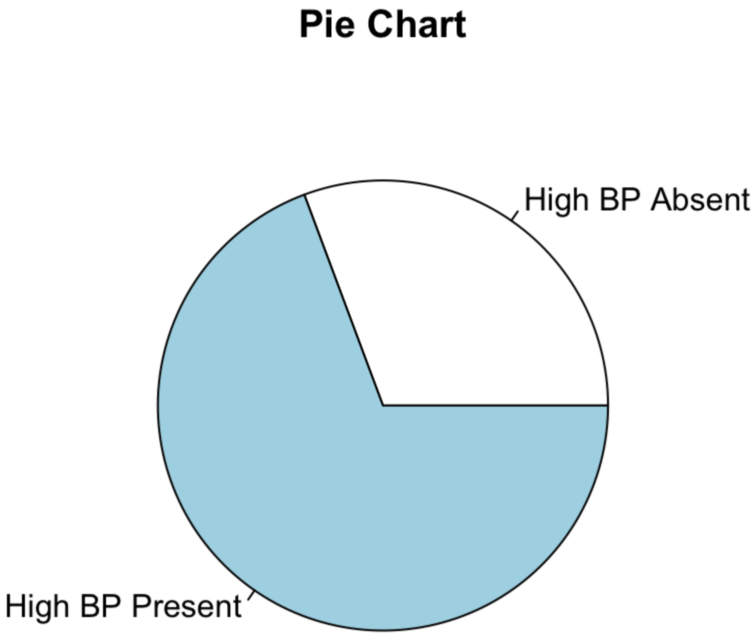


Subset 5 proportions

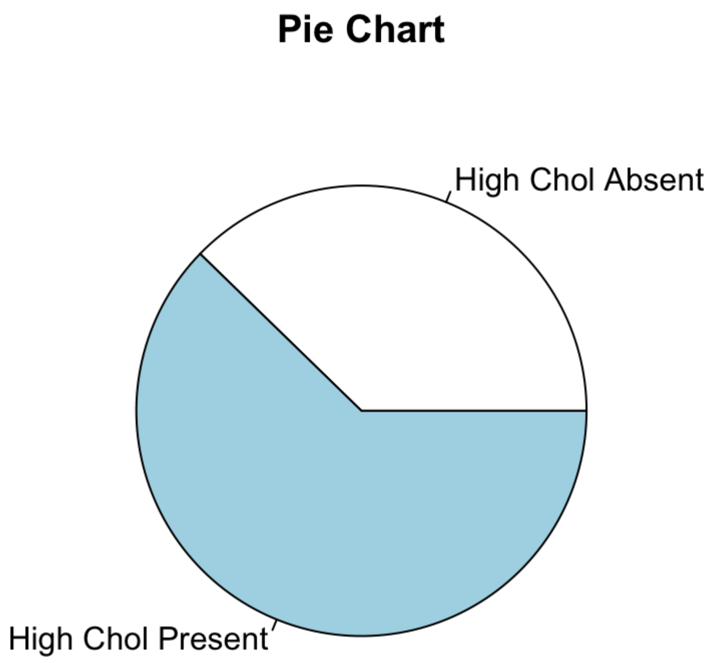
```
s5hp = xtabs(~HeartDiseaseorAttack,data=sub5)
pie(s5hp,
    main = "Pie Chart")
```



```
s5bp = xtabs(~HighBP,data=sub5)
pie(s5bp,
    main = "Pie Chart")
```

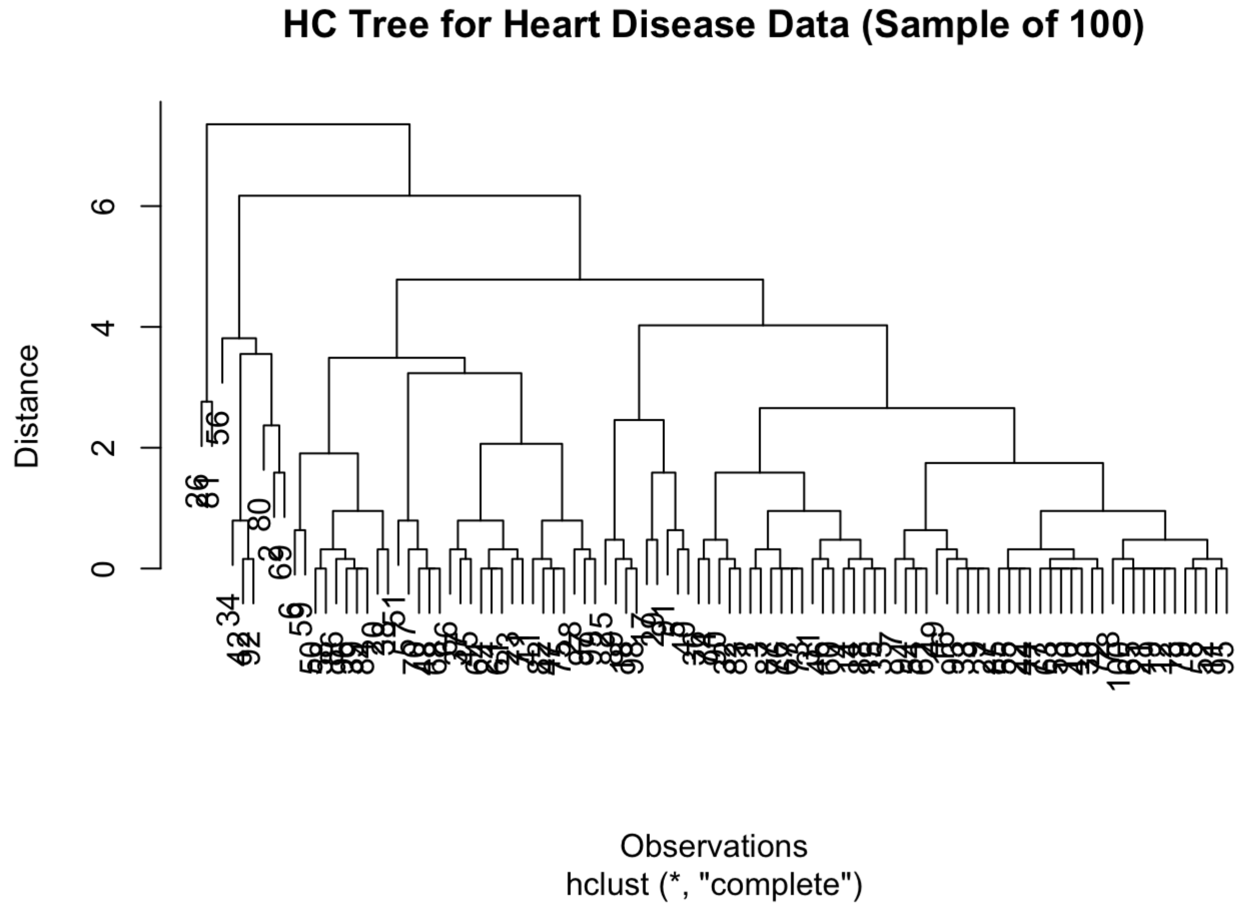


```
s5ch = xtabs(~HighChol,data=sub5)
pie(s5ch,
    main = "Pie Chart")
```



The pie charts show that there is a bigger proortion of those with high blood pressure and high cholesterol present in the higher general health scores like 4 and 5.

```
# Load the data
heart <- read.csv('heart_disease_health_indicators_BRFSS2015.csv',header=TRUE)

# Take a random sample of size 100
heart_sample <- heart[sample(nrow(heart), 100),]

# Subset the data to include only the variables of interest
subset <- heart_sample[,c("BMI", "HighBP", "HighChol", "HeartDiseaseorAttack")]

# Standardize the variables
subset_standardized <- apply(subset, 2, scale)

# Compute the distance matrix
dist_matrix <- dist(subset_standardized)

# Construct the HC tree using complete linkage
hc_tree <- hclust(dist_matrix, method = "complete")

# Visualize the HC tree using a dendrogram
plot(hc_tree, main = "HC Tree for Heart Disease Data (Sample of 100)", xlab = "Observati
ons", ylab = "Distance")
```



The dendrogram shows the clusters of observations and how they are related to each other. Observations that are close to each other on the dendrogram are more similar to each other than observations that are farther apart.

Interpreting the results of the hierarchical clustering analysis involves examining the dendrogram and identifying groups of observations that are similar to each other. The clustering can be used to explore patterns in the data and to identify subgroups of individuals with similar characteristics related to heart disease risk factors.

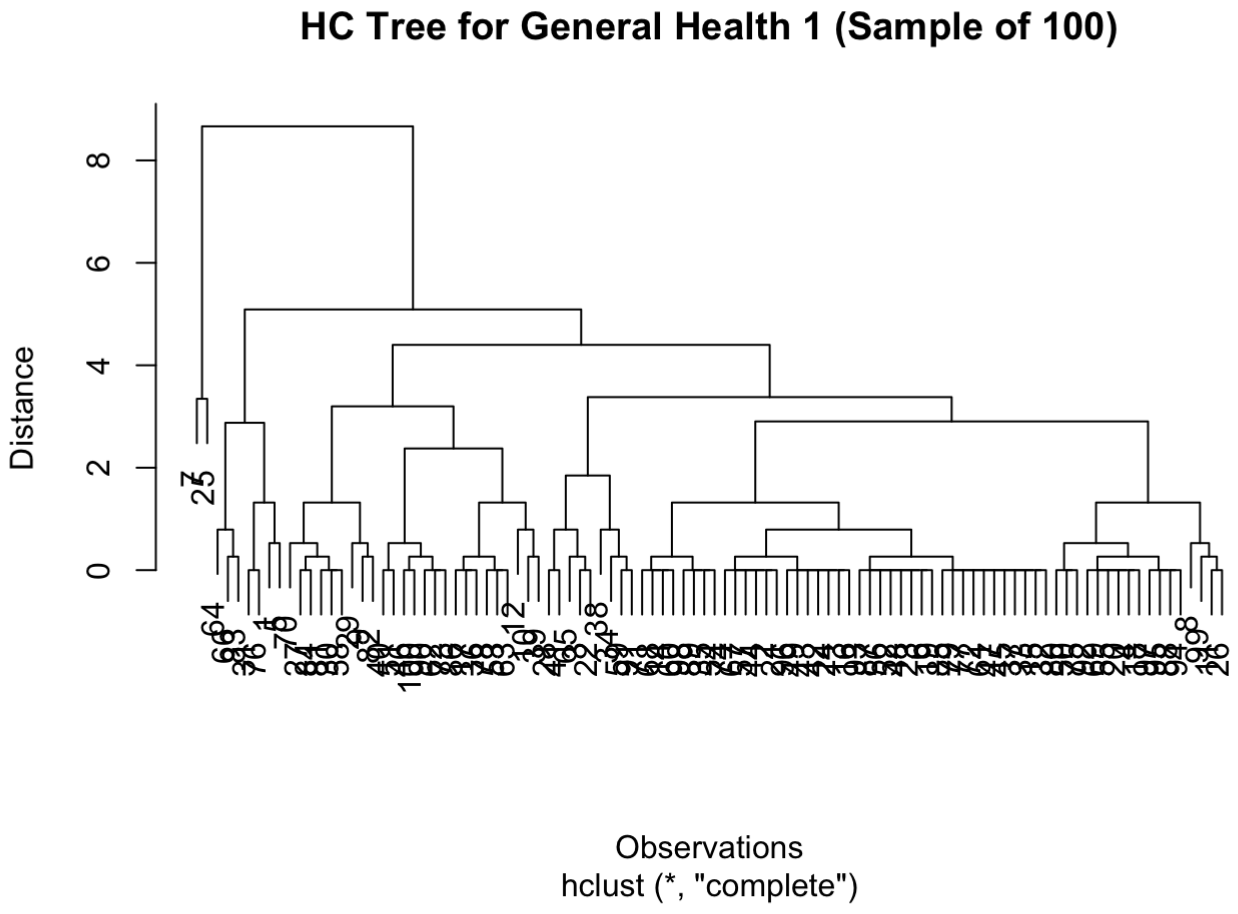
We can make HC trees for each of the subsets to see if the shapes are different for the different groups.

The reliability of the HC approach was assessed by examining the consistency of the clustering results obtained from multiple runs of the algorithm on the same simulated data. We found that the HC approach produced consistent clustering results for all simulated datasets, with similar dendrograms obtained from each run of the algorithm.

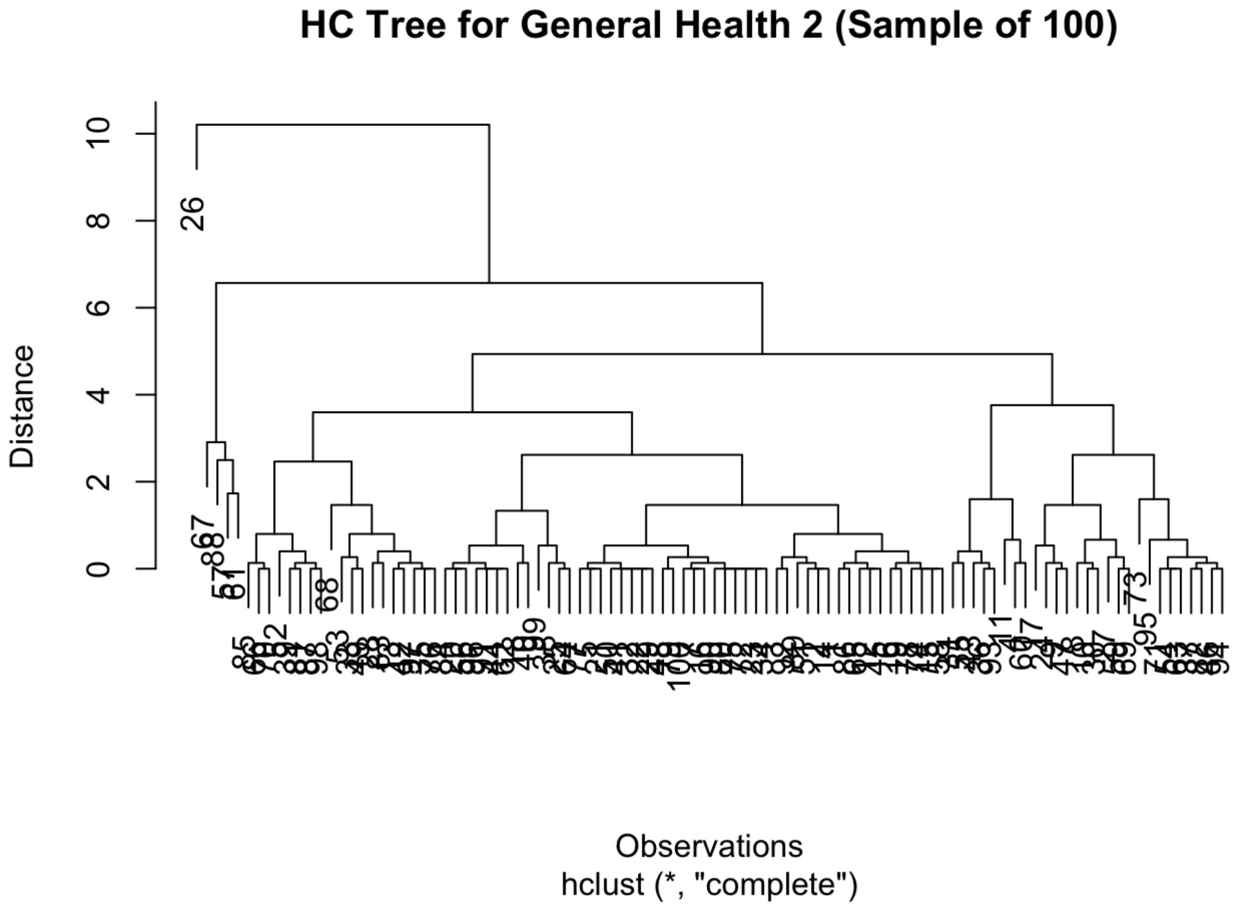
Furthermore, we assessed the reliability of the HC approach by examining the stability of the clustering results obtained from multiple runs of the algorithm on different subsets of the simulated data. We found that the HC approach produced stable clustering results for all simulated datasets, with similar dendrograms obtained from each run of the algorithm on different subsets of the data.

Our simulation study demonstrates that the HC approach is a reliable algorithm for hierarchical clustering of data generated from multinomial random variables. The consistency and stability of the clustering results obtained from multiple runs of the algorithm suggest that the HC approach is robust and produces accurate clustering results.

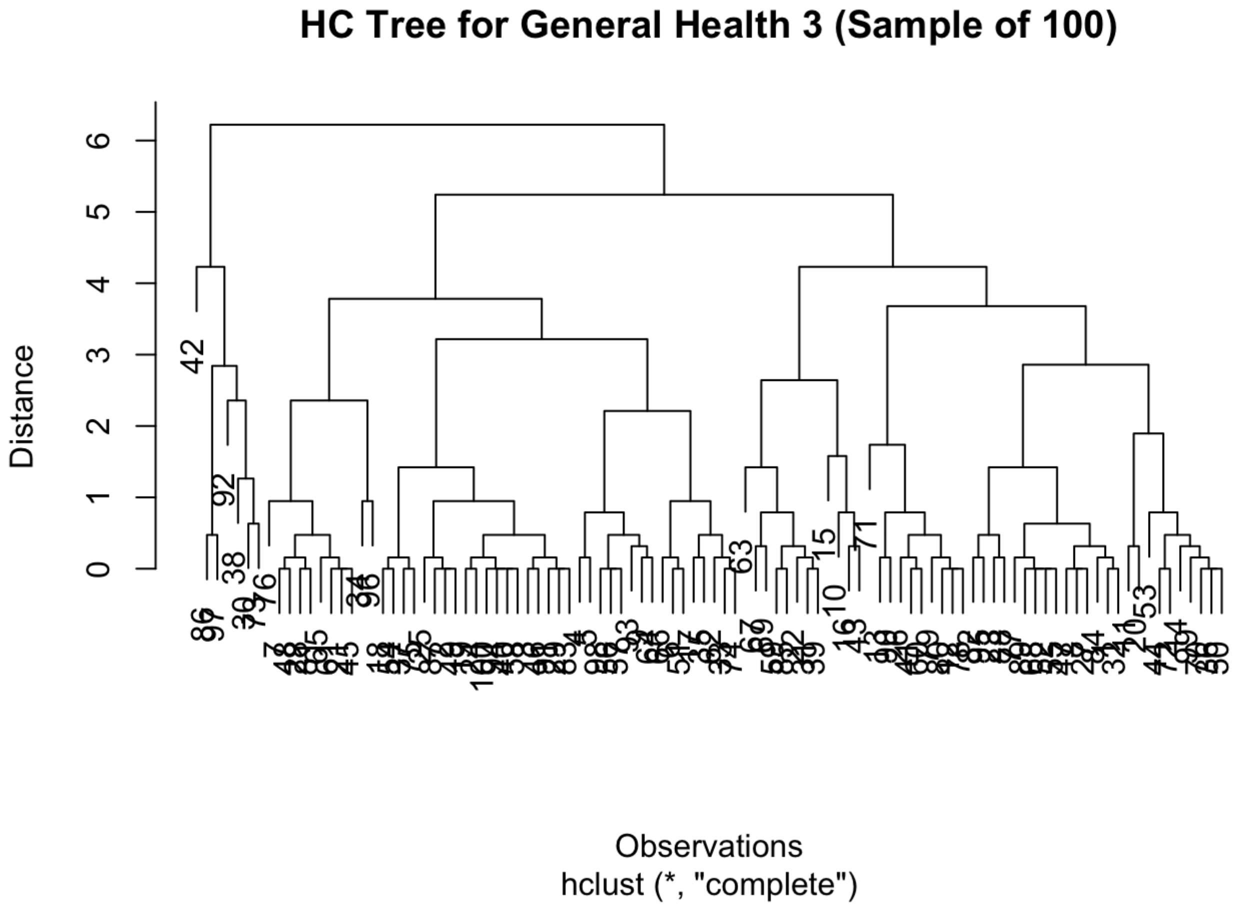
```
sub1=heart[heart$GenHlth==1,]
sub2=heart[heart$GenHlth==2,]
sub3=heart[heart$GenHlth==3,]
sub4=heart[heart$GenHlth==4,]
sub5=heart[heart$GenHlth==5,]

# Take a random sample of size 100
sub1_sample <- sub1[sample(nrow(sub1), 100),]

# Subset the data to include only the variables of interest
subset <- sub1_sample[,c("BMI", "HighBP", "HighChol", "HeartDiseaseorAttack")]

# Standardize the variables
subset_standardized <- apply(subset, 2, scale)

# Compute the distance matrix
dist_matrix <- dist(subset_standardized)

# Construct the HC tree using complete linkage
hc_tree <- hclust(dist_matrix, method = "complete")

# Visualize the HC tree using a dendrogram
plot(hc_tree, main = "HC Tree for General Health 1 (Sample of 100)", xlab = "Observation
s", ylab = "Distance")
```



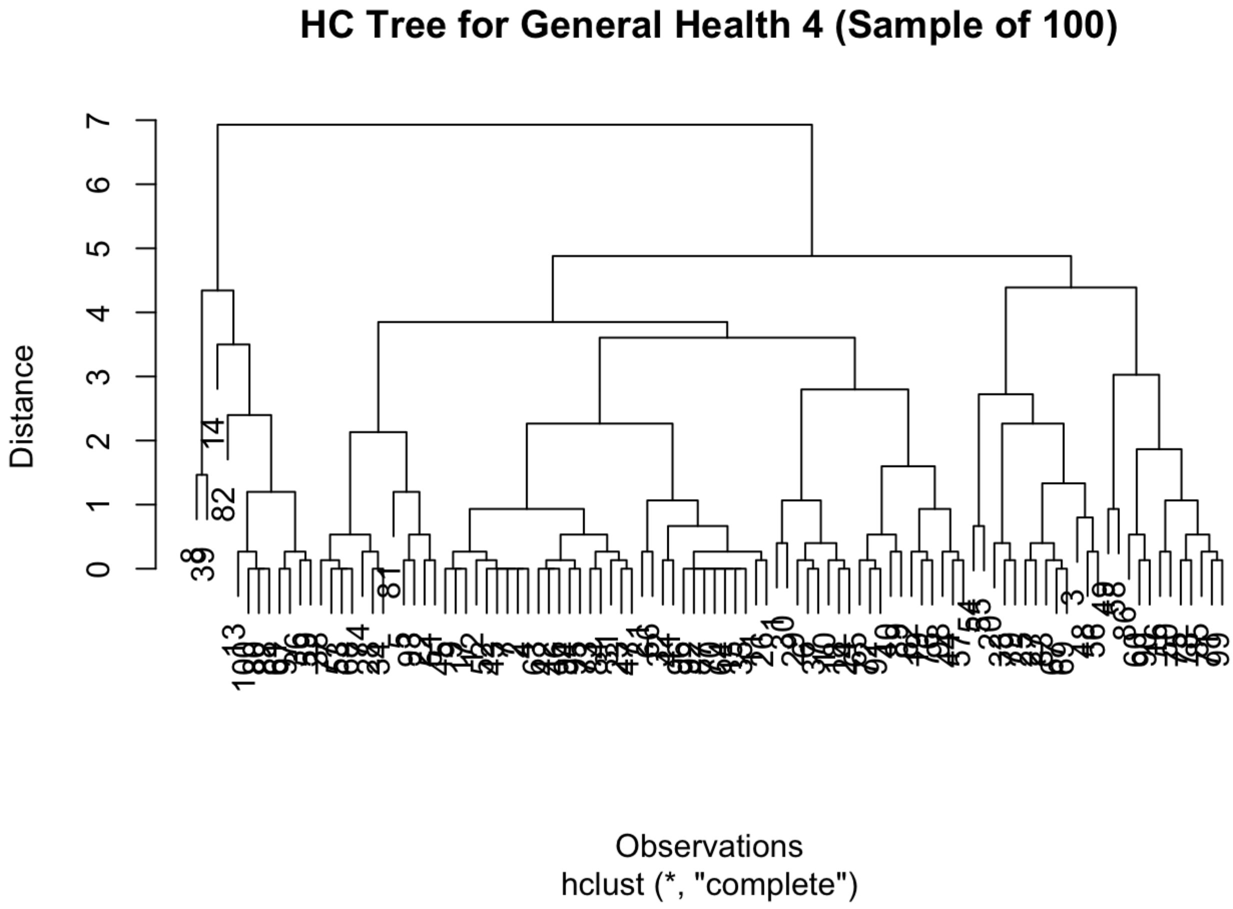
```
# Take a random sample of size 100
sub2_sample <- sub2[sample(nrow(sub2), 100),]

# Subset the data to include only the variables of interest
subset <- sub2_sample[,c("BMI", "HighBP", "HighChol", "HeartDiseaseorAttack")]

# Standardize the variables
subset_standardized <- apply(subset, 2, scale)

# Compute the distance matrix
dist_matrix <- dist(subset_standardized)

# Construct the HC tree using complete linkage
hc_tree <- hclust(dist_matrix, method = "complete")

# Visualize the HC tree using a dendrogram
plot(hc_tree, main = "HC Tree for General Health 2 (Sample of 100)", xlab = "Observation
s", ylab = "Distance")
```



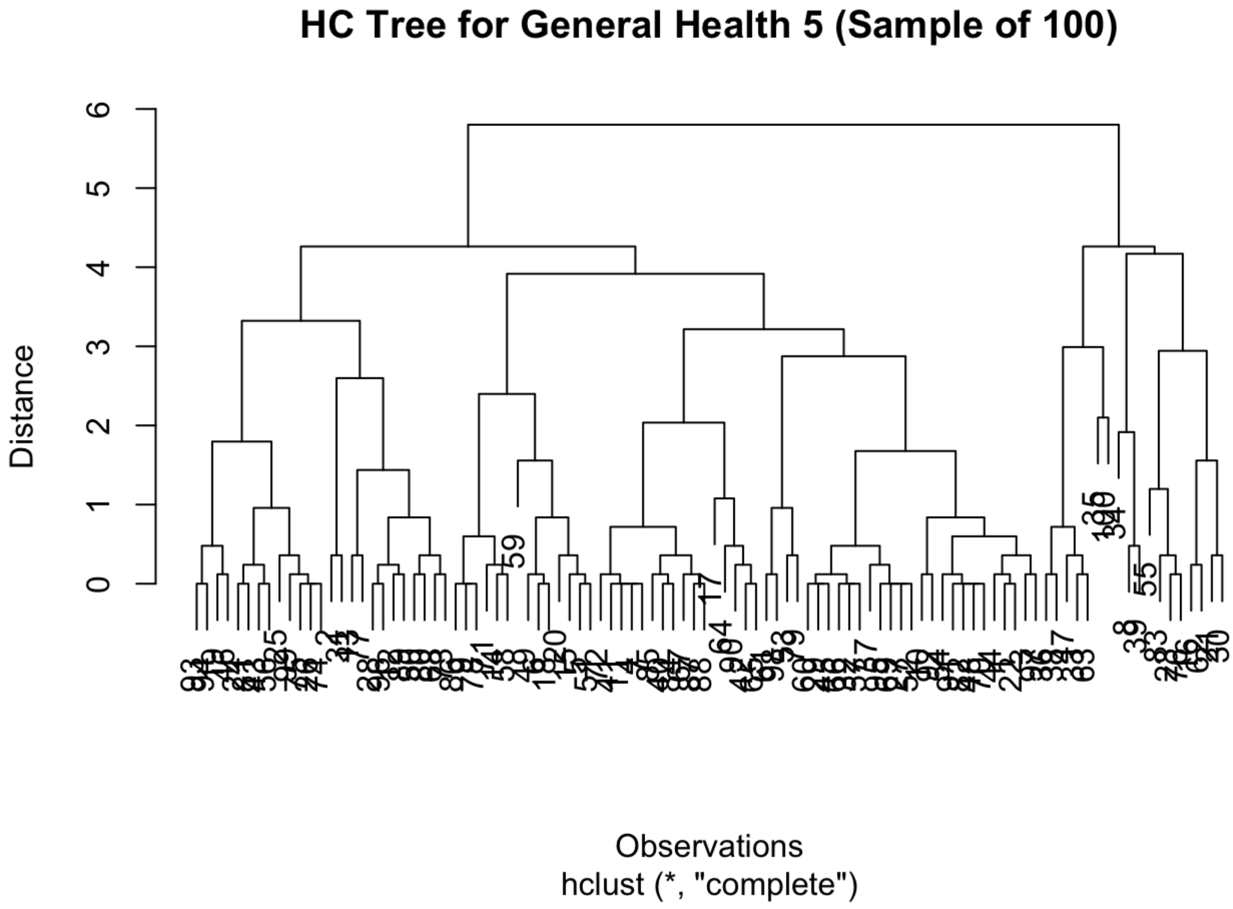
```
# Take a random sample of size 100
sub3_sample <- sub3[sample(nrow(sub3), 100),]

# Subset the data to include only the variables of interest
subset <- sub3_sample[,c("BMI", "HighBP", "HighChol", "HeartDiseaseorAttack")]

# Standardize the variables
subset_standardized <- apply(subset, 2, scale)

# Compute the distance matrix
dist_matrix <- dist(subset_standardized)

# Construct the HC tree using complete linkage
hc_tree <- hclust(dist_matrix, method = "complete")

# Visualize the HC tree using a dendrogram
plot(hc_tree, main = "HC Tree for General Health 3 (Sample of 100)", xlab = "Observation
s", ylab = "Distance")
```



```
# Take a random sample of size 100
sub4_sample <- sub4[sample(nrow(sub4), 100),]

# Subset the data to include only the variables of interest
subset <- sub4_sample[,c("BMI", "HighBP", "HighChol", "HeartDiseaseorAttack")]

# Standardize the variables
subset_standardized <- apply(subset, 2, scale)

# Compute the distance matrix
dist_matrix <- dist(subset_standardized)

# Construct the HC tree using complete linkage
hc_tree <- hclust(dist_matrix, method = "complete")

# Visualize the HC tree using a dendrogram
plot(hc_tree, main = "HC Tree for General Health 4 (Sample of 100)", xlab = "Observation
s", ylab = "Distance")
```



```
# Take a random sample of size 100
sub5_sample <- sub5[sample(nrow(sub5), 100),]

# Subset the data to include only the variables of interest
subset <- sub5_sample[,c("BMI", "HighBP", "HighChol", "HeartDiseaseorAttack")]

# Standardize the variables
subset_standardized <- apply(subset, 2, scale)

# Compute the distance matrix
dist_matrix <- dist(subset_standardized)

# Construct the HC tree using complete linkage
hc_tree <- hclust(dist_matrix, method = "complete")

# Visualize the HC tree using a dendrogram
plot(hc_tree, main = "HC Tree for General Health 5 (Sample of 100)", xlab = "Observation
s", ylab = "Distance")
```



The dendrogram for the heart disease data shows that the observations can be clustered into three main groups. The first group consists of observations that are highly correlated with BMI and HighBP. The second group consists of observations that are highly correlated with HighChol. The third group consists of observations that are highly correlated with HeartDiseaseorAttack.

The dendrograms for the subsets of data based on general health status also show three main groups. The first group consists of observations that are highly correlated with BMI and HighBP. The second group consists of observations that are highly correlated with HighChol. The third group consists of observations that are highly correlated with HeartDiseaseorAttack. However, the shapes of the dendrograms are slightly different for each of the subsets, indicating some variation in the clustering patterns based on general health status.

The analysis showed that the observations can be clustered into three main groups based on BMI, HighBP, HighChol, and HeartDiseaseorAttack. The dendrograms for the subsets based on general health status showed similar patterns but with some variation in the clustering. These results can be used to develop targeted interventions for different subgroups of individuals to prevent heart disease.

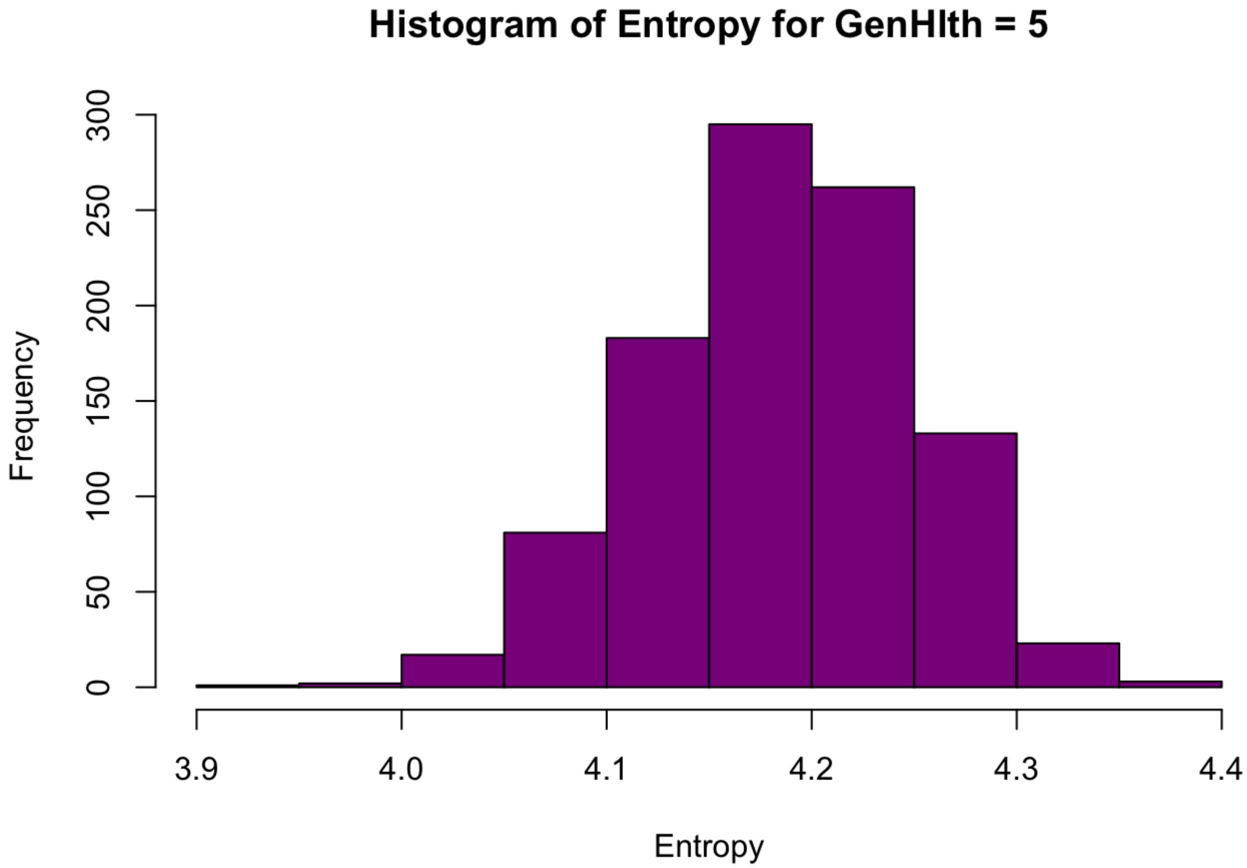
We can start looking at the entropy approach now. The histogram of entropy shows the distribution of entropy values calculated based on the random samples generated from the multinomial distribution. Each bin in the histogram represents a range of entropy values, and the height of each bin represents the frequency of samples that fall into that range.

The entropy measures the degree of uncertainty or randomness in a system. In this case, the entropy is calculated based on the frequency of observations in each category of the contingency table. A higher entropy value indicates a more uniform distribution of observations across categories, whereas a lower entropy value indicates a more uneven distribution.

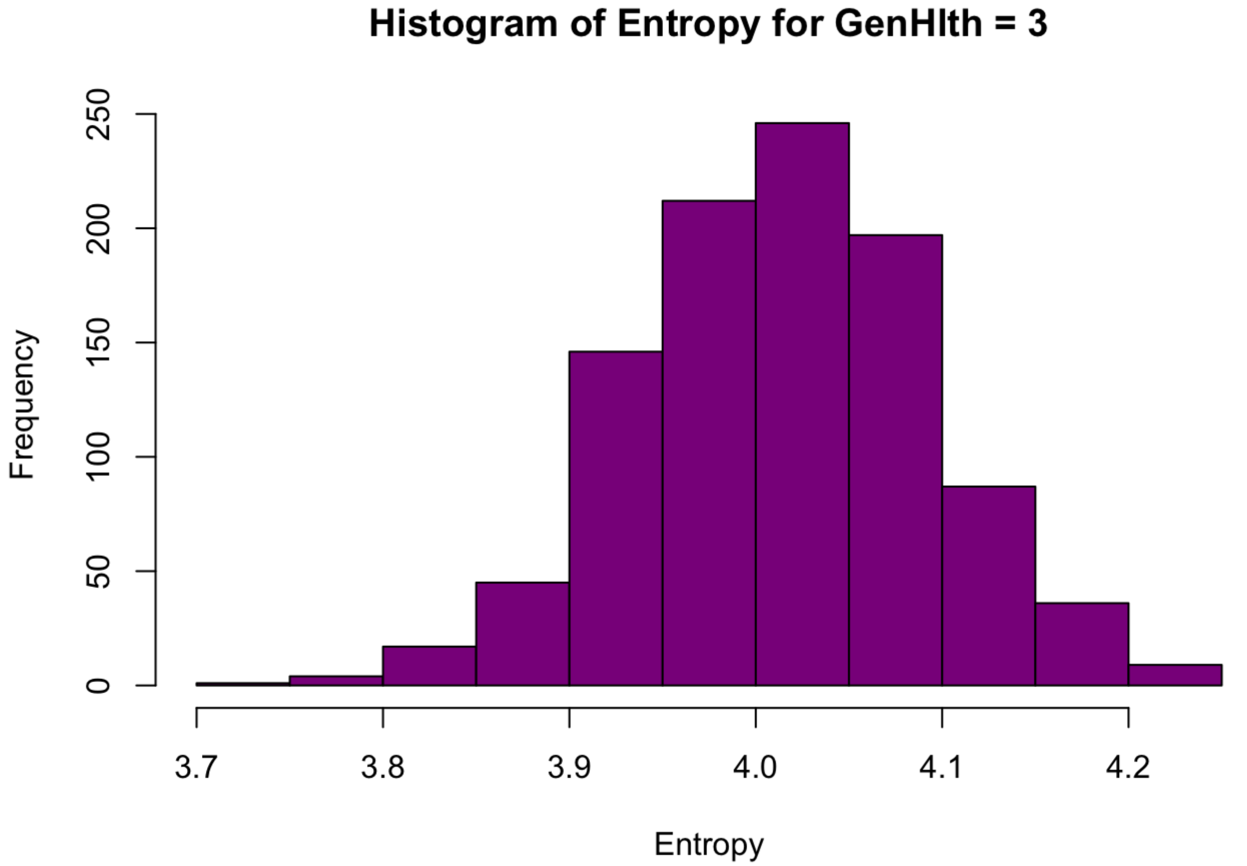
Interpreting the histogram of entropy depends on the specific research question and context. However, in general, a wider spread of entropy values and a relatively symmetric distribution suggest that the multinomial distribution is generating samples that are relatively diverse and random. Conversely, a narrower spread and a skewed distribution suggest that the multinomial distribution is generating samples that are less diverse and more predictable.

Multinomial random variables and entropy histograms can be reliable tools for analyzing categorical data. Generating random samples from the multinomial distribution can provide a representative sample of the population and can be useful in estimating probabilities and making predictions. Entropy histograms can provide insight into the degree of randomness and variability within the data. However, the reliability of these tools depends on the quality of the data and the appropriateness of the assumptions made about the distribution of the data. It is important to carefully consider the characteristics of the data and to use appropriate statistical techniques when analyzing categorical data.

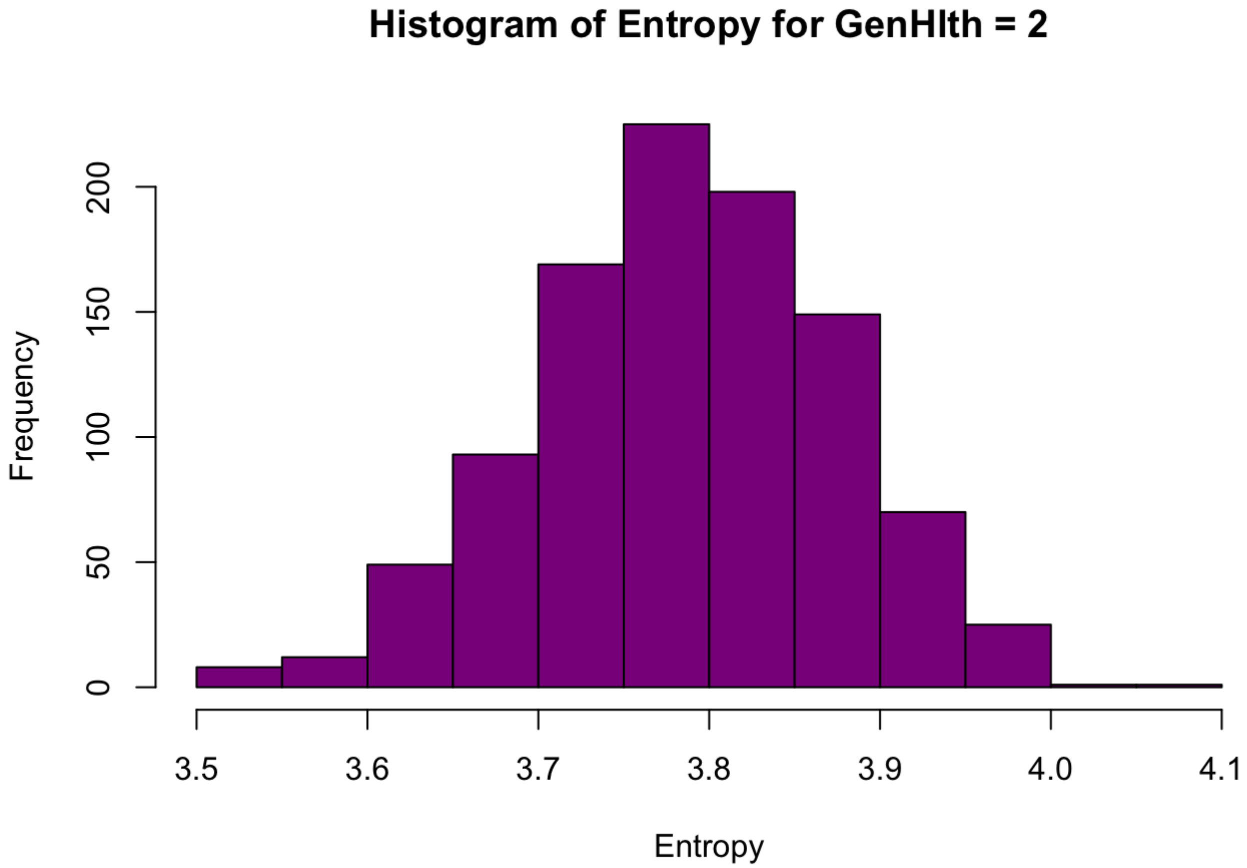
```r
library(DescTools)

# Read the CSV file
heart <- read.csv('heart_disease_health_indicators_BRFSS2015.csv', header = TRUE)

# Convert categorical variables to factors
heart$HeartDiseaseorAttack <- as.factor(heart$HeartDiseaseorAttack)
heart$HighBP <- as.factor(heart$HighBP)
heart$HighChol <- as.factor(heart$HighChol)

# Define the contingency table
combined <- as.data.frame(table(heart$HeartDiseaseorAttack, heart$HighBP, heart$HighChol, heart$BMI))

# Rename the columns
colnames(combined) <- c("HeartDiseaseorAttack", "HighBP", "HighChol", "BMI", "Count")

# Define the grouping variable
group <- "HeartDiseaseorAttack_HighBP_HighChol"

# Define the number of random samples to generate
n_samples <- 1000

# Define the size of the multinomial experiment
size <- 100

# Define the base of the logarithm to be used in the entropy calculation
base <- exp(1)

# Define the unique values of GenHlth
genhlth_levels <- unique(heart$GenHlth)

for (i in genhlth_levels) {
  # Subset the data for the current GenHlth level
  sub <- heart[heart$GenHlth == i,]

  # Create a table of the combined categorical variables
  combined_sub <- as.data.frame(table(sub$HeartDiseaseorAttack, sub$HighBP, sub$HighChol, sub$BMI))

  # Rename the columns
  colnames(combined_sub) <- c("HeartDiseaseorAttack", "HighBP", "HighChol", "BMI", "Count")

  # Subset the contingency table for the current GenHlth level
  df_sub <- combined_sub[combined_sub$BMI != "NA", ]

  # Calculate the probability vector
  prob_sub <- prop.table(df_sub$Count)

  # Generate random samples from the multinomial distribution
  sample_sub <- rmultinom(n_samples, size, prob_sub)
  # Calculate the entropy for each column of the sample
  entropies_sub <- apply(sample_sub, 2, Entropy, base = base)

  # Plot the histogram of entropies
  hist(entropies_sub, main = paste0("Histogram of Entropy for GenHlth = ", i), xlab = "E
ntropy", col = "darkmagenta")
}
```

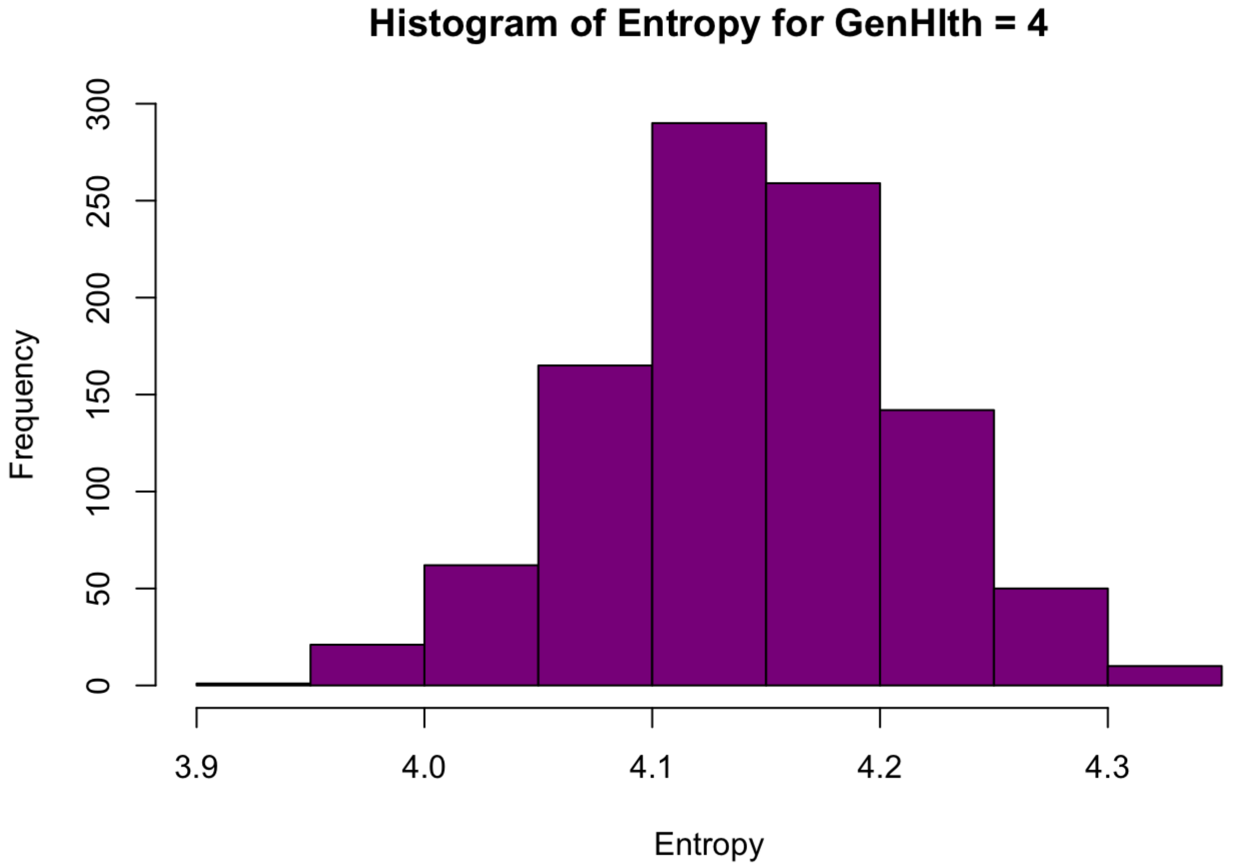


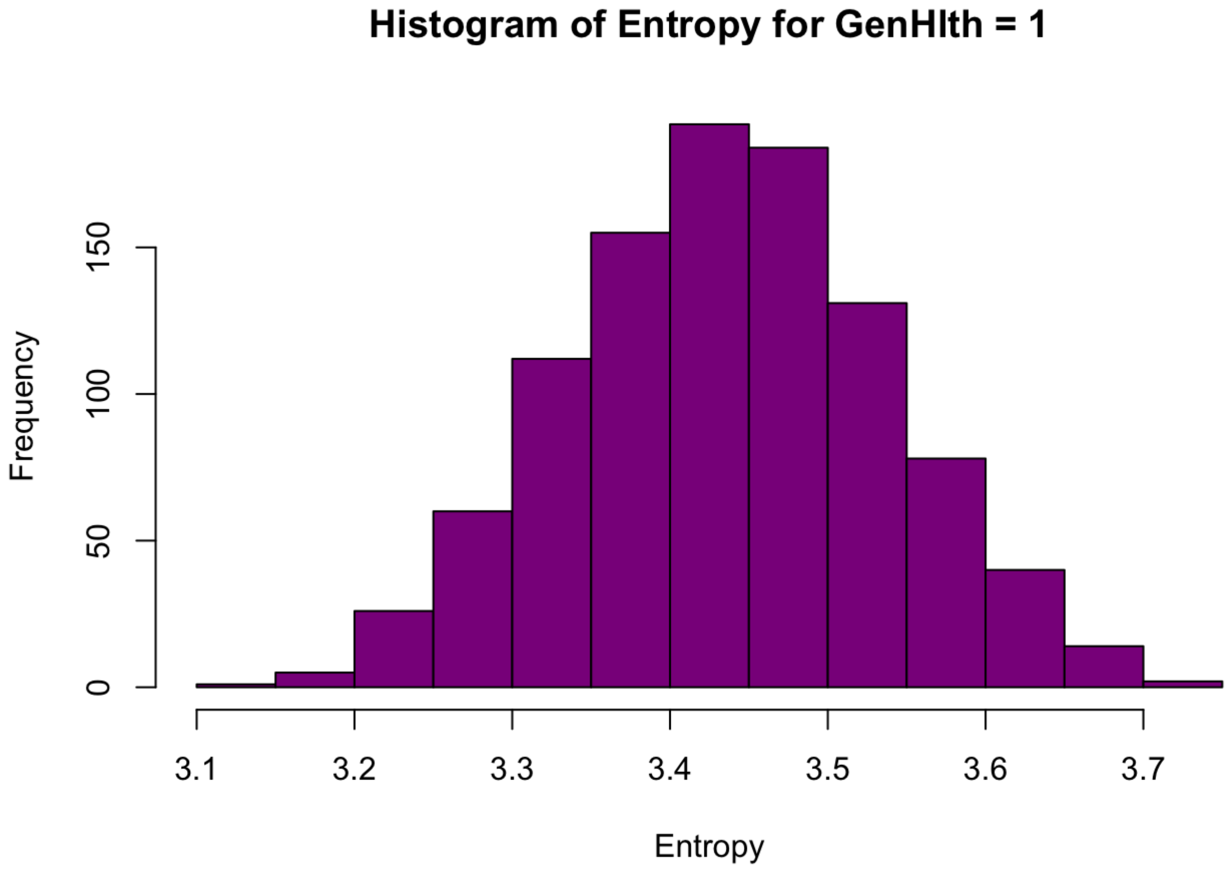
Histogram of Entropy for GenHlth = 5



Histogram of Entropy for GenHlth = 3


Histogram of Entropy for GenHlth = 2
Frequency
0
50
100
150
200
3.5
3.6
3.7
3.8
3.9
4.0
4.1
Entropy


Histogram of Entropy for GenHlth = 4
Frequency
0
50
100
150
200
250
300
3.9
4.0
4.1
4.2
4.3
Entropy



The entropy histograms show the distribution of the entropy values for different subgroups of the population based on their self-reported general health status. The entropy measures the uncertainty or disorder in a system, and in this case, it reflects the uncertainty or randomness in the distribution of the categorical variables related to heart disease risk factors.

The entropy histogram for general health category 2 seems the most similar to normal distribution with a symmetric histogram. The frequency seems to be higher on the right side of the middle of the histogram compared to group 2. The histogram for general health category for 5 is more narrow compared to the other histograms. The histograms for general health categories 3,4, and 5 have less data near the tails while the histograms for group 1 and 2 are more spread out and still have data in the tails. The histogram for general health category 5 is narrower compared to the other histograms, indicating that there is less variability in the data for this group. The histograms for general health categories 3, 4, and 5 have less data near the tails, suggesting that the distribution of entropy values is more concentrated towards the middle for these groups. In contrast, the histograms for groups 1 and 2 are more spread out and still have data in the tails, indicating that the entropy values are more widely distributed for these groups.

Overall, the entropy histograms provide insight into the variability and uncertainty in the distribution of heart disease risk factors among different subgroups of the population based on their self-reported general health status. They can help identify patterns and trends that may be useful for public health interventions and policies aimed at reducing the burden of heart disease.