Applied Statistics Final Project: Heart Disease Health Indicators

Maria Varghese, Esther Aruti, Abhina Premachandran, Tyler Anatole

2023-12-11

Introduction

According to the CDC, one person dies every 33 seconds from heart disease, and is the leading cause of death in the United States. Additionally, heart disease costs the United States about $239.9 billion each year from 2018 to 2019. This includes the cost of health care services, medicines, and lost productivity due to death.

Heart disease is characterized by the buildup of plaque in coronary arteries, which is associated with aging, high blood pressure, and diabetes. These, and many more risk factors can result in typical heart disease symptoms such as chest pain, heart attack, or sudden cardiac arrest.

Because of the many risk factors associated with heart disease, it is possible to predict when a person is more likely to suffer from symptoms and take preventative measures against the disease.

This project will explore factors related to heart disease, and look at the strength of their relationships. The data set used was found on Kaggle (https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset (https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset)), and was collected by the Behavioral Risk Factor Surveillance System. This is a health related telephone survey by the CDC that collects responses every year. It contains over 244,000 survey responses for 22 questions regarding health.

The data in this set is written in such a way where a value of 0 represents ‘No’ to the question in the corresponding column, and 1 represents ‘Yes’. However, in the ‘Age’ column, the values 1-13 represent different age intervals. So is the case for the ‘Income’ and ‘Education’ columns.

Exploratory Data Analysis

Data Cleaning

First, we will begin by importing all necessary libraries.

```r
suppressWarnings(suppressMessages(library(dplyr)))
suppressWarnings(suppressMessages(library(ggplot2)))
suppressWarnings(suppressMessages(library(tidyr)))
suppressWarnings(suppressMessages(library(pROC)))
suppressWarnings(suppressMessages(library(boot)))
suppressWarnings(suppressMessages(library(caret)))
suppressWarnings(suppressMessages(library(ISLR)))
suppressWarnings(suppressMessages(library(scales)))
```

Reading in the data and taking a quick look at it.

```r
#setwd('~/DSE I1030_Applied Statistics/')
data <- read.table('heart_disease_health_indicators_BRFSS2015.csv', sep=",", header = 1)
head(data)
```

```
##   HeartDiseaseorAttack HighBP HighChol CholCheck BMI Smoker Stroke Diabetes
## 1                    0      1        1         1  40      1      0        0
## 2                    0      0        0         0  25      1      0        0
## 2                    0      0        0         0  25      1      0        0
## 3                    0      1        1         1  28      0      0        0
## 4                    0      1        0         1  27      0      0        0
## 5                    0      1        1         1  24      0      0        0
## 6                    0      1        1         1  25      1      0        0
##   PhysActivity Fruits Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost
## 1            0      0       1                 0             1           0
## 2            1      0       0                 0             0           1
## 3            0      1       0                 0             1           1
## 4            1      1       1                 0             1           0
## 5            1      1       1                 0             1           0
## 6            1      1       1                 0             1           0
##   GenHlth MentHlth PhysHlth DiffWalk Sex Age Education Income
## 1       5       18       15        1   0   9         4      3
## 2       3        0        0        0   0   7         6      1
## 3       5       30       30        1   0   9         4      8
## 4       2        0        0        0   0  11         3      6
## 5       2        3        0        0   0  11         5      4
## 6       2        0        2        0   1  10         6      8
```

Looking at the data types.

```
str(data)
```

```
## 'data.frame':    253680 obs. of  22 variables:
##  $ HeartDiseaseorAttack: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ HighBP              : num  1 0 1 1 1 1 1 1 1 0 ...
##  $ HighChol            : num  1 0 1 0 1 1 0 1 1 0 ...
##  $ CholCheck           : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ BMI                 : num  40 25 28 27 24 25 30 25 30 24 ...
##  $ Smoker              : num  1 1 0 0 0 1 1 1 1 0 ...
##  $ Stroke              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Diabetes            : num  0 0 0 0 0 0 0 0 2 0 ...
##  $ PhysActivity        : num  0 1 0 1 1 1 0 1 0 0 ...
##  $ Fruits              : num  0 0 1 1 1 1 0 0 1 0 ...
##  $ Veggies             : num  1 0 0 1 1 1 0 1 1 1 ...
##  $ HvyAlcoholConsump   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AnyHealthcare       : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ NoDocbcCost         : num  0 1 1 0 0 0 0 0 0 0 ...
##  $ GenHlth             : num  5 3 5 2 2 2 3 3 5 2 ...
##  $ MentHlth            : num  18 0 30 0 3 0 0 0 30 0 ...
##  $ PhysHlth            : num  15 0 30 0 0 2 14 0 30 0 ...
##  $ DiffWalk            : num  1 0 1 0 0 0 0 1 1 0 ...
##  $ Sex                 : num  0 0 0 0 0 1 0 0 0 1 ...
##  $ Age                 : num  9 7 9 11 11 10 9 11 9 8 ...
##  $ Education           : num  4 6 4 3 5 6 6 4 5 4 ...
##  $ Income              : num  3 1 8 6 4 8 7 4 1 3 ...
```

Now let’s look at a few summary statistics.

```
summary(data)
```

```
##  HeartDiseaseorAttack     HighBP         HighChol        CholCheck
##  Min.   :0.00000      Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000      1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.0000
##  Median :0.00000      Median :0.000   Median :0.0000   Median :1.0000
##  Mean   :0.09419      Mean   :0.429   Mean   :0.4241   Mean   :0.9627
##  3rd Qu.:0.00000      3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  3rd Qu.:0.00000      3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :1.00000      Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
##       BMI            Smoker           Stroke            Diabetes     
##  Min.   :12.00   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:24.00   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
##  Median :27.00   Median :0.0000   Median :0.00000   Median :0.0000  
##  Mean   :28.38   Mean   :0.4432   Mean   :0.04057   Mean   :0.2969  
##  3rd Qu.:31.00   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000  
##  Max.   :98.00   Max.   :1.0000   Max.   :1.00000   Max.   :2.0000  
##   PhysActivity        Fruits          Veggies       HvyAlcoholConsump
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000   
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :0.0000   
##  Mean   :0.7565   Mean   :0.6343   Mean   :0.8114   Mean   :0.0562   
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   
##  AnyHealthcare     NoDocbcCost         GenHlth         MentHlth     
##  Min.   :0.0000   Min.   :0.00000   Min.   :1.000   Min.   : 0.000  
##  1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 0.000  
##  Median :1.0000   Median :0.00000   Median :2.000   Median : 0.000  
##  Mean   :0.9511   Mean   :0.08418   Mean   :2.511   Mean   : 3.185  
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:3.000   3rd Qu.: 2.000  
##  Max.   :1.0000   Max.   :1.00000   Max.   :5.000   Max.   :30.000  
##     PhysHlth         DiffWalk           Sex              Age        
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   : 1.000  
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 6.000  
##  Median : 0.000   Median :0.0000   Median :0.0000   Median : 8.000  
##  Mean   : 4.242   Mean   :0.1682   Mean   :0.4403   Mean   : 8.032  
##  3rd Qu.: 3.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:10.000  
##  Max.   :30.000   Max.   :1.0000   Max.   :1.0000   Max.   :13.000  
##    Education        Income     
##  Min.   :1.00   Min.   :1.000  
##  1st Qu.:4.00   1st Qu.:5.000  
##  Median :5.00   Median :7.000  
##  Mean   :5.05   Mean   :6.054  
##  3rd Qu.:6.00   3rd Qu.:8.000  
##  Max.   :6.00   Max.   :8.000
```

Also, looking at the number of missing values.

```
sum(is.na(data))
```

```
## [1] 0
```

It looks like the data has already been very well cleaned, and has no missing values!

We will remove all rows where people haven't checked cholesterol in five years or more.

```
data1 <- data[(data$CholCheck != 0), ]
```

Now let's check if there is an even amount of people with and without heart disease sampled.

```
sum(data1$HeartDiseaseorAttack == 1)
```

```
## [1] 23622
```

```
sum(data1$HeartDiseaseorAttack == 0)
```

```
## [1] 220588
## [1] 220588
```

There are way more observations for people without heart disease than those with. It looks like we may need to do some stratification.

```
# even out number of people with and without heart disease
rows_to_remove <- data1 %>%
  filter(HeartDiseaseorAttack == 0) %>%
  sample_n(193000)

# Create a new data frame without the identified rows
filtered_data <- data1 %>%
  anti_join(rows_to_remove)
```

```
## Joining with `by = join_by(HeartDiseaseorAttack, HighBP, HighChol, CholCheck,
## BMI, Smoker, Stroke, Diabetes, PhysActivity, Fruits, Veggies,
## HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth,
## DiffWalk, Sex, Age, Education, Income)`
```

```
sum(filtered_data$HeartDiseaseorAttack == 1)
```

```
## [1] 23622
```

```
sum(filtered_data$HeartDiseaseorAttack == 0)
```

```
## [1] 23513
```

Now that they are of similar size, we can continue.

Let’s select data for all people that have heart disease.

```
heart_disease <- filtered_data[(filtered_data$HeartDiseaseorAttack == 1),]
```

Now for all people that do not have heart disease.

```
no_heart_disease <-filtered_data[(filtered_data$HeartDiseaseorAttack == 0),]
```

Hypothesis Testing

Testing to see if certain categorical variables impact our target variable. First we change our binary columns to yes or no depending on the number. In this case, Yes is 1 and No is 0. We exclude our target variable `HeartDiseaseorAttack` as we need our x and y to have two levels.

```
numerical_data <- filtered_data %>%
  mutate(across(c('HighBP', 'HighChol', 'Smoker', 'Stroke', 'PhysActivity', 'Fruits', 'V
eggies', 'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'DiffWalk', 'Sex'), ~ case
_when(.x == 1 ~ "Yes", .x == 0 ~ "No")))
```

```
for ( col in c('HighBP', 'HighChol', 'Smoker', 'Stroke', 'PhysActivity', 'Fruits', 'Vegg
ies', 'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'DiffWalk', 'Sex', 'Age')){
  # creating our table
  print(paste0("Table for Heart Disease or Attack and ", col, ":"))
  print(table(numerical_data$HeartDiseaseorAttack, numerical_data[[col]]))
  # running our chi squared test
  print(paste0("Chi Squared for Heart Disease or Attack and ", col, ":"))
  print(chisq.test(numerical_data$HeartDiseaseorAttack, numerical_data[[col]]))

  print("-------------------------------------------------------------")
}
```

```
## [1] "Table for Heart Disease or Attack and HighBP:"
## 
##         No   Yes
##   0 13169 10344
##   1  5856 17766
## [1] "Chi Squared for Heart Disease or Attack and HighBP:"
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 4769.2, df = 1, p-value < 2.2e-16
## 
## [1] "-------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and HighChol:"
## 
##         No   Yes
##   0 13396 10117
##   1  7013 16609
## [1] "Chi Squared for Heart Disease or Attack and HighChol:"
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 3571.9, df = 1, p-value < 2.2e-16
## 
## [1] "-------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and Smoker:"
## 
##         No   Yes
##   0 12824 10689
##   1  9011 14611
## [1] "Chi Squared for Heart Disease or Attack and Smoker:"
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 1272.9, df = 1, p-value < 2.2e-16
## 
## [1] "-------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and Stroke:"
## 
##         No   Yes
##   0 22733   780
##   1 19736  3886
## [1] "Chi Squared for Heart Disease or Attack and Stroke:"
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 2277.3, df = 1, p-value < 2.2e-16
## 
## [1] "-------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and PhysActivity:"
##    
##        No   Yes
##   0  6344 17169
##   1  8474 15148
## [1] "Chi Squared for Heart Disease or Attack and PhysActivity:"
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 431.9, df = 1, p-value < 2.2e-16
## 
## [1] "--------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and Fruits:"
##    
##        No   Yes
##   0  9146 14367
##   1  9312 14310
## [1] "Chi Squared for Heart Disease or Attack and Fruits:"
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 1.3323, df = 1, p-value = 0.2484
## 
## [1] "--------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and Veggies:"
##    
##        No   Yes
##   0  4802 18711
##   1  5570 18052
## [1] "Chi Squared for Heart Disease or Attack and Veggies:"
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 68.244, df = 1, p-value < 2.2e-16
## 
## [1] "--------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and HvyAlcoholConsump:"
##    
##        No   Yes
##   0 22009  1504
##   1 22795   827
## [1] "Chi Squared for Heart Disease or Attack and HvyAlcoholConsump:"
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 209.55, df = 1, p-value < 2.2e-16
## 
## [1] "--------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and AnyHealthcare:"
##    
##        No   Yes
##   0  1247 22266
##   1   801 22821
## [1] "Chi Squared for Heart Disease or Attack and AnyHealthcare:"
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 103.25, df = 1, p-value < 2.2e-16
##
## [1] "---------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and NoDocbcCost:"
##
##        No   Yes
##   0 21353  2160
##   1 21070  2552
## [1] "Chi Squared for Heart Disease or Attack and NoDocbcCost:"
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 34.068, df = 1, p-value = 5.323e-09
##
## [1] "---------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and DiffWalk:"
##
##        No   Yes
##   0 19494  4019
##   1 13809  9813
## [1] "Chi Squared for Heart Disease or Attack and DiffWalk:"
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 3396.1, df = 1, p-value < 2.2e-16
##
## [1] "---------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and Sex:"
##
##        No   Yes
##   0 13812  9701
##   1 10073 13549
## [1] "Chi Squared for Heart Disease or Attack and Sex:"
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 1221.3, df = 1, p-value < 2.2e-16
##
## [1] "---------------------------------------------------------------"
## [1] "Table for Heart Disease or Attack and Age:"
##
##        1    2    3    4    5    6    7    8    9   10   11   12   13
##   0  630  795 1058 1335 1558 1804 2441 2909 2995 2920 2111 1460 1497
##   1   27   54  116  186  334  693 1393 2219 3311 4162 3922 3075 4130
## [1] "Chi Squared for Heart Disease or Attack and Age:"
##
##  Pearson's Chi-squared test
##
## data:  numerical_data$HeartDiseaseorAttack and numerical_data[[col]]
## X-squared = 7073.7, df = 12, p-value < 2.2e-16
##
## [1] "---------------------------------------------------------------"
```

The tables printed show the distribution of our target variable `HeartDiseasesorAttack` and categorical variables: `HighBP`, `HighChol`, `Smoker`, `Stroke`, `PhysActivity`, `Fruits`, `Veggies`, `HvyAlcoholConsump`, `AnyHealthcare`, `NoDocbcCost`, `DiffWalk`, `Sex` along with its relationship. Let’s go through the results:

All categorical variables show a strong relationship with our target. Most categorical variables have p-values of 2.2e-16 or 0.00000000000000022 which is very close to zero indicating that those variables are considered statistically significant and do have an impact on our target variable!

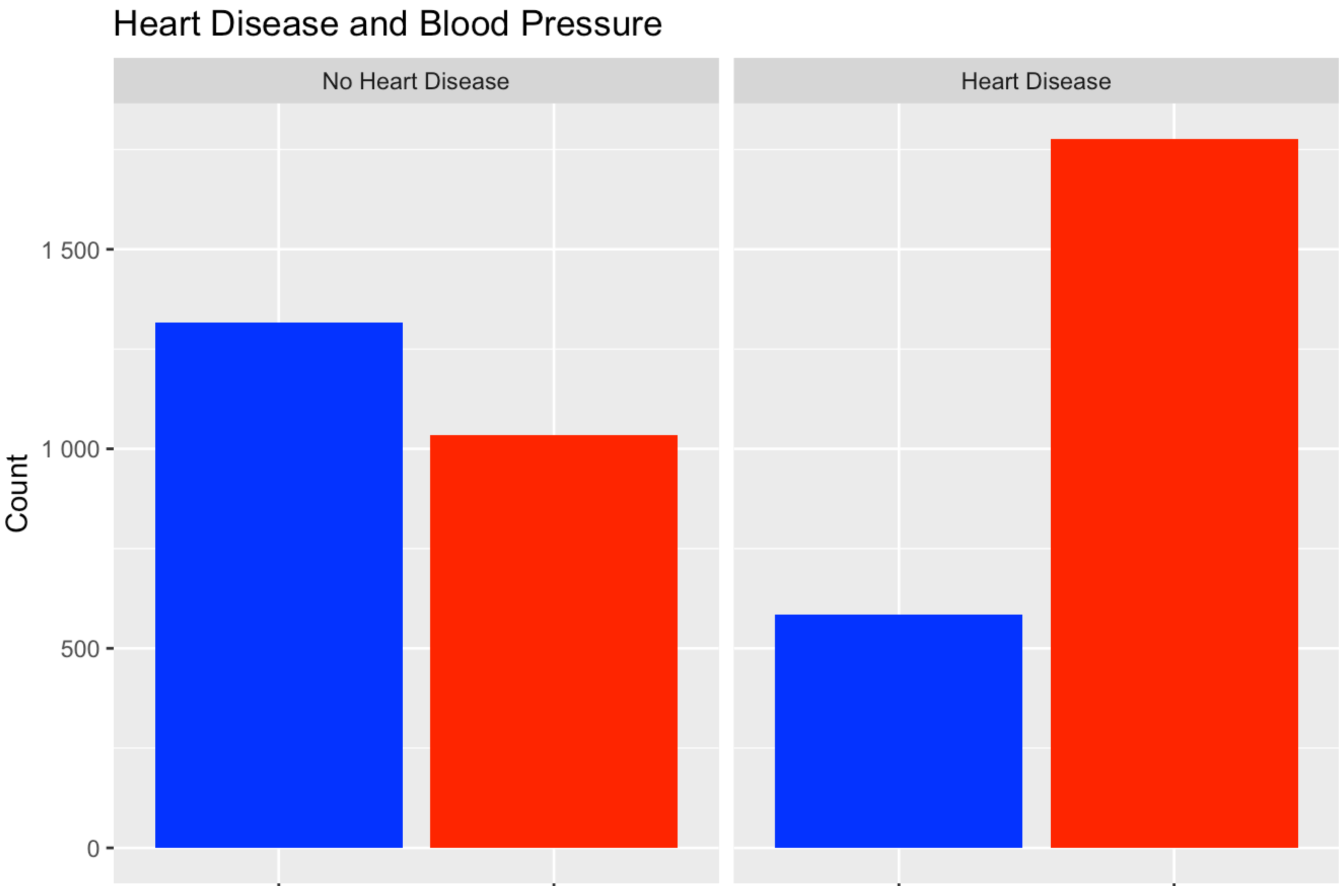
We can now begin plotting.

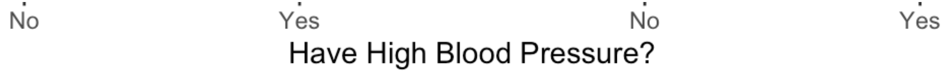
Plotting

High Blood Pressure

First, we will look at the differences in high blood pressure between people that have heart disease and people that don’t.

```
# change 0 to no and 1 to yes
highbp <- filtered_data %>%
  mutate(HighBP = ifelse(HighBP == 0, "No", "Yes"))

# number of people with high blood pressure
ggplot(highbp, aes(x = as.factor(HighBP), fill = as.factor(HighBP))) +
  geom_bar() +
  scale_fill_manual(values = c('blue', 'red')) +
  theme(legend.position = "none") +
  facet_wrap(~HeartDiseaseorAttack, labeller = labeller(HeartDiseaseorAttack = c('0' = '
No Heart Disease', '1' = 'Heart Disease')), strip.position = "top") +
  labs(title = "Heart Disease and Blood Pressure", x = 'Have High Blood Pressure?', y =
'Count') +
  scale_y_continuous(labels = label_number(scale = .1))
```



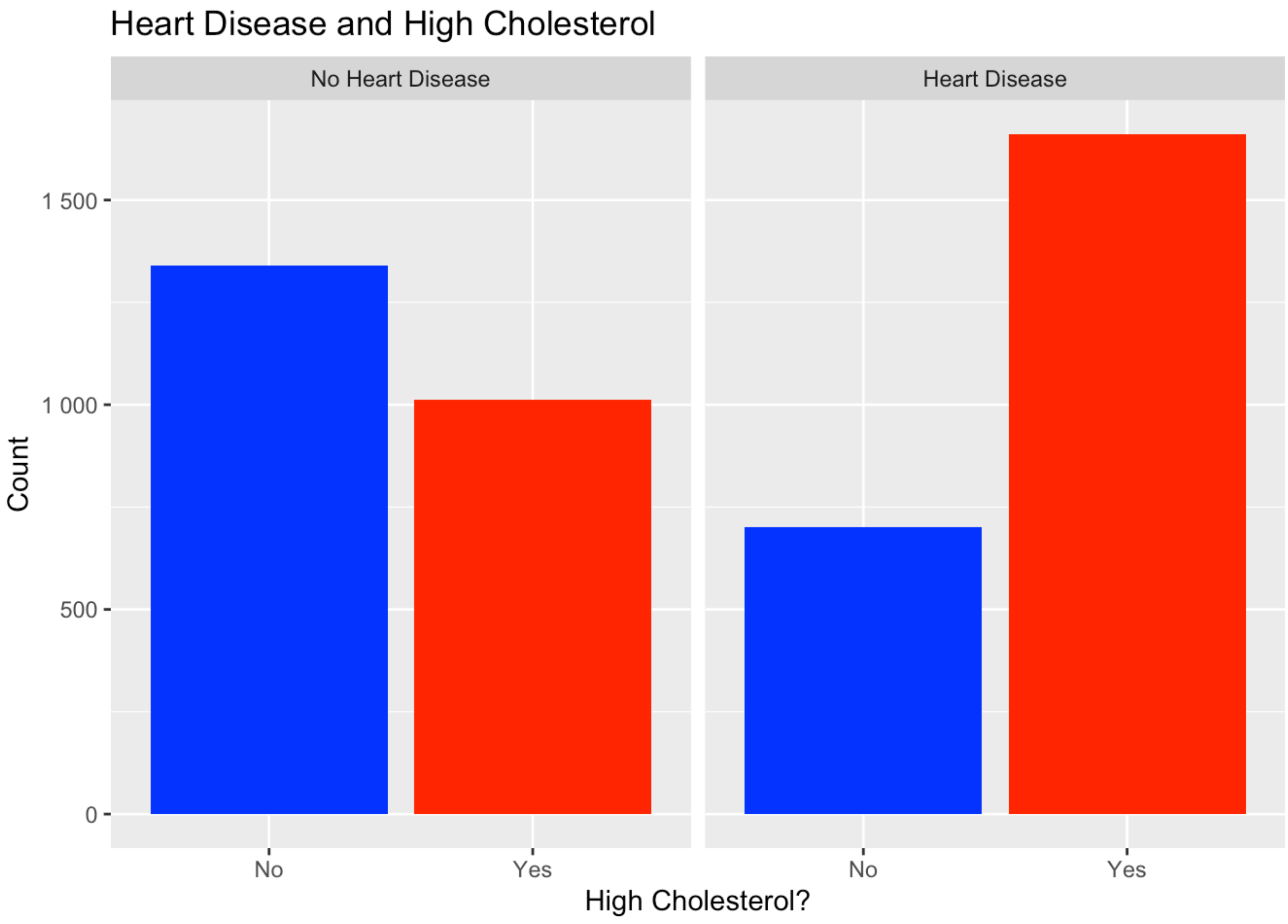


From this plot we can see that, people with heart disease tend to have high blood pressure compared to people without heart disease.

High Cholesterol

Here we will compare high cholesterol between people with and without heart disease.

```
# in high cholesterol column, change 0 to No, and 1 to Yes
hi_chol <- filtered_data %>%
  mutate(HighChol = ifelse(HighChol == 0, "No", "Yes"))

# bar plot for people with heart disease that have high cholesterol vs don't
ggplot(hi_chol, aes(x = as.factor(HighChol), fill = as.factor(HighChol))) +
  geom_bar() +
  scale_fill_manual(values = c('blue', 'red')) +
  theme(legend.position = "none") +
  facet_wrap(~HeartDiseaseorAttack, labeller = labeller(HeartDiseaseorAttack = c('0' = '
No Heart Disease', '1' = 'Heart Disease')), strip.position = "top") +
  labs(title = "Heart Disease and High Cholesterol", x = 'High Cholesterol?', y = 'Coun
t') +
  scale_y_continuous(labels = label_number(scale = .1))
```

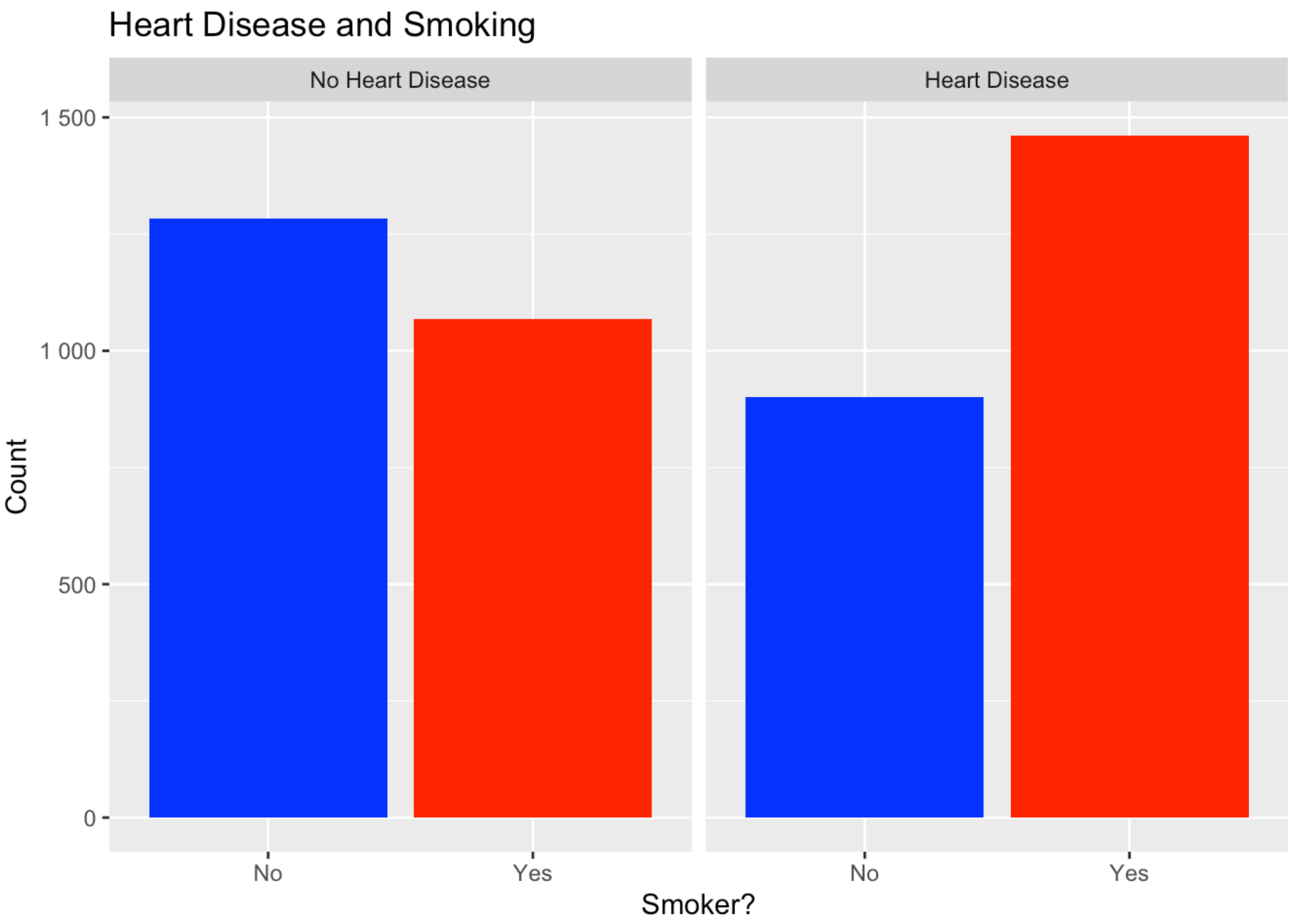


This plot shows a similar trend where people with heart disease tend to also have high cholesterol. Also, when compared to the previous plot, it seems that the same amount of people that have high blood pressure, also have high cholesterol.

Smoking

Here we will compare smoking in people with and without heart disease.

```r
# in smoke column, change 0 to No, and 1 to Yes
smoke <- filtered_data %>%
  mutate(Smoker = ifelse(Smoker == 0, "No", "Yes"))

# bar plot for people that smoke vs don't
  ggplot(smoke, aes(x = as.factor(Smoker), fill = as.factor(Smoker))) +
    geom_bar() +
    scale_fill_manual(values = c('blue', 'red')) +
    theme(legend.position = "none") +
    facet_wrap(~HeartDiseaseorAttack, labeller = labeller(HeartDiseaseorAttack = c('0' =
'No Heart Disease', '1' = 'Heart Disease')), strip.position = "top") +
    labs(title = "Heart Disease and Smoking", x = 'Smoker?', y = 'Count') +
    scale_y_continuous(labels = label_number(scale = .1))
```

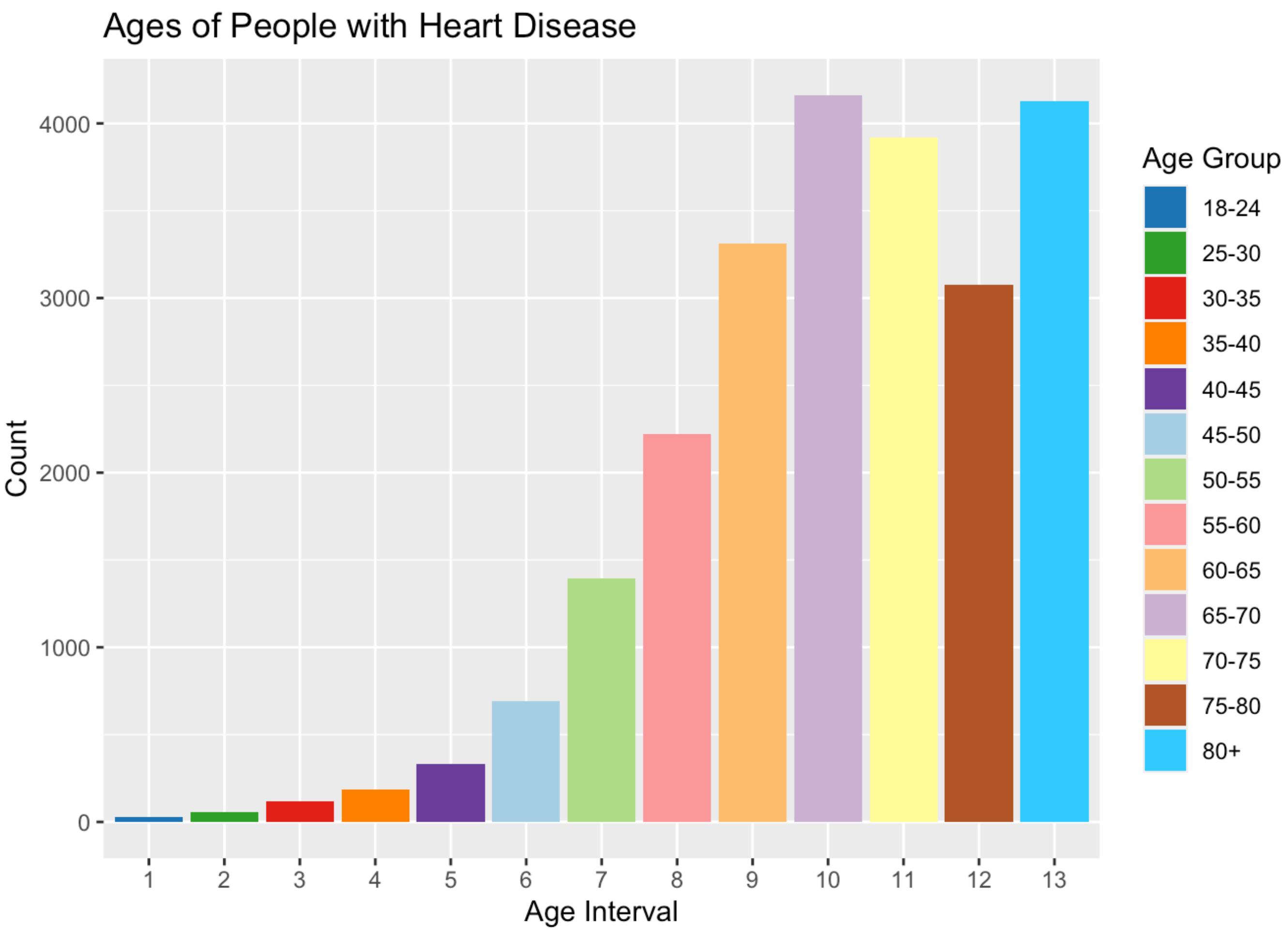


This plot shows that for people with heart disease, more people smoke than those who don’t, especially compared to people without heart disease. However this difference is only by a few hundred people. Perhaps smoking is not as strongly correlated to heart disease as other factors?

Age

Here we will compare the ages of people with heart disease.

```
# plot of ages of people that have heart disease
ggplot(heart_disease, aes(x = as.factor(Age), fill = as.factor(Age))) +
  geom_bar() +
  scale_fill_manual(
    values = c("#1f78b4", "#33a02c", "#e31a1c", "#ff7f00", "#6a3d9a", "#a6cee3", "#b2df8
a", "#fb9a99", "#fdbf6f", "#cab2d6", "#ffff99", "#b15928", "#33ccff"),
    labels = c('18-24', '25-30', '30-35', '35-40', '40-45', '45-50', '50-55', '55-60', '
60-65', '65-70', '70-75', '75-80', '80+')
  ) +
  labs(title = "Ages of People with Heart Disease", x = 'Age Interval', y = 'Count') +
  guides(fill = guide_legend(title = "Age Group"))
```

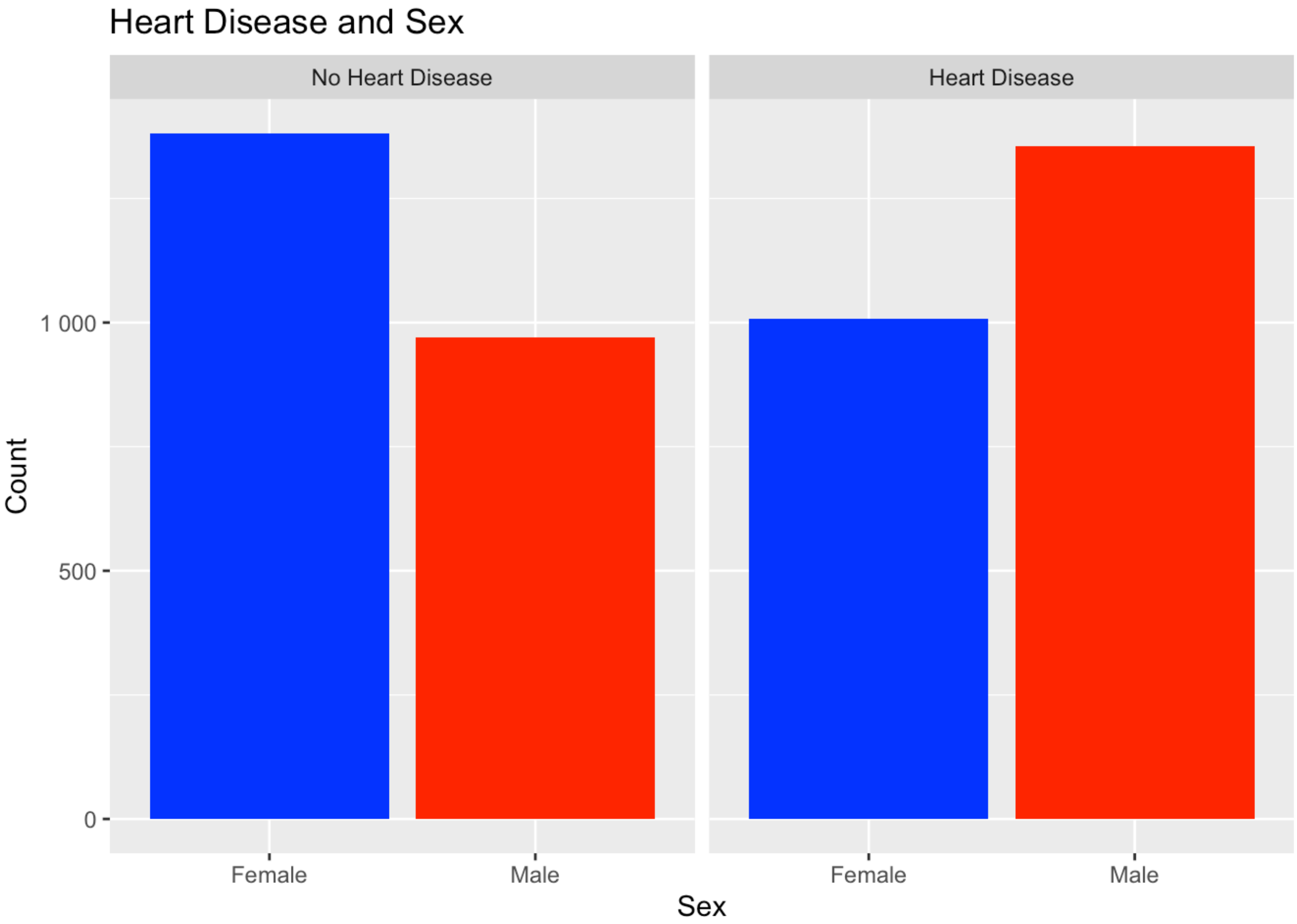


This plot shows the number of people in specific age brackets that have heart disease. From this, it is clear that there are greater amounts of older people with heart disease compared to younger people, especially from ages 60 to 80.

Sex

Here we will compare the prevalence of heart disease between the sexes.

```
# change 0 to no and 1 to yes
sex <- filtered_data %>%
  mutate(Sex = ifelse(Sex == 0, "Female", "Male"))

# plots for sex differences
ggplot(sex, aes(x = as.factor(Sex), fill = as.factor(Sex))) +
  geom_bar() +
  scale_fill_manual(values = c('blue', 'red')) +
  theme(legend.position = "none") +
  facet_wrap(~HeartDiseaseorAttack, labeller = labeller(HeartDiseaseorAttack = c('0' = '
No Heart Disease', '1' = 'Heart Disease')), strip.position = "top") +
  labs(title = "Heart Disease and Sex", x = 'Sex', y = 'Count') +
  scale_y_continuous(labels = label_number(scale = .1))
```

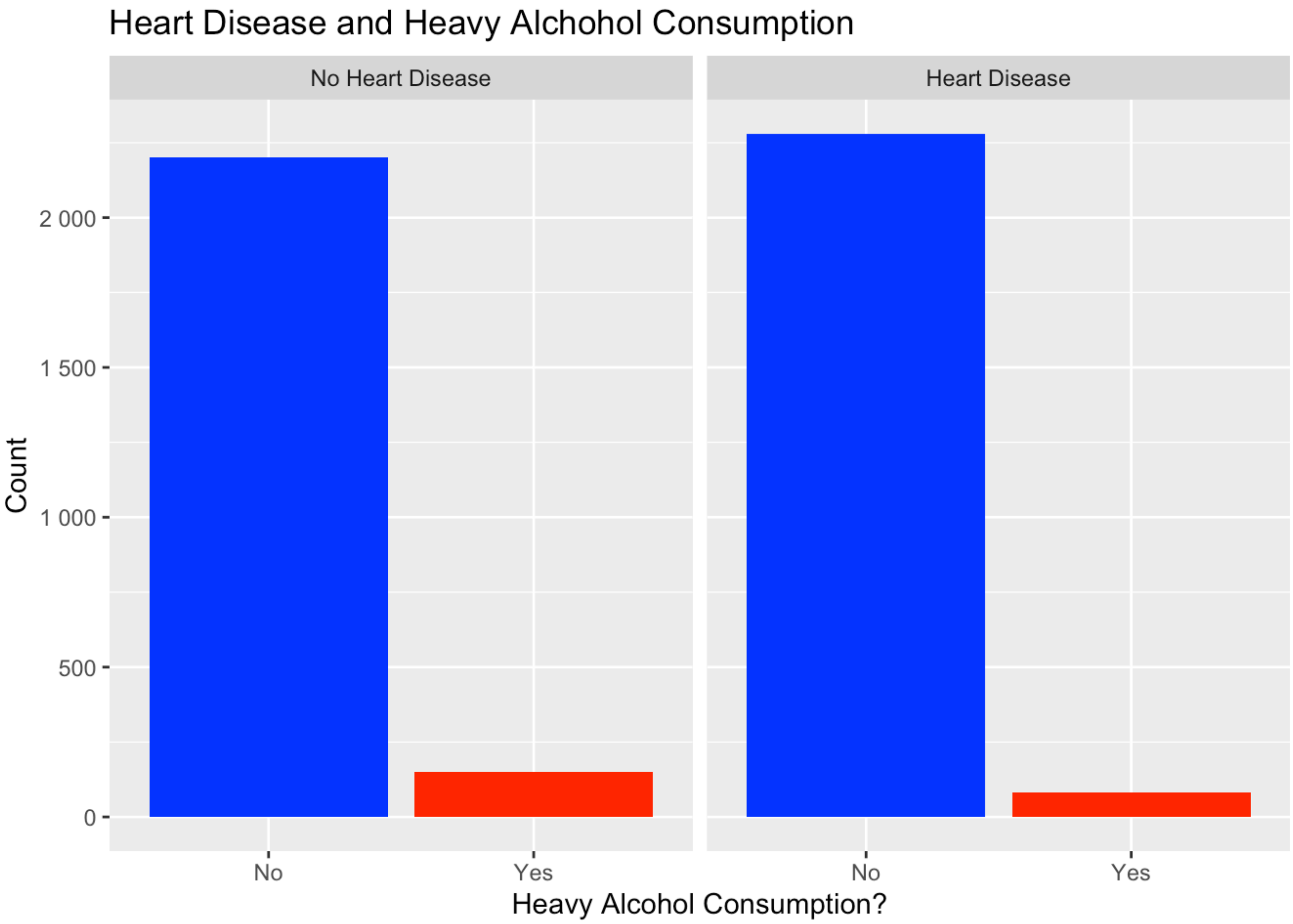


Here, it looks like there are more women without heart disease than men, and there are more men with heart disease than women, but only by a small amount.

Heavy Alcohol Consumption

Here we will compare heart disease among people with and without heavy alcohol consumption.

```
# in alcohol column, change 0 to No, and 1 to Yes
alch <- filtered_data %>%
  mutate(HvyAlcoholConsump = ifelse(HvyAlcoholConsump == 0, "No", "Yes"))

# bar plot for people that smoke vs don't
  ggplot(alch, aes(x = as.factor(HvyAlcoholConsump), fill = as.factor(HvyAlcoholConsum
p))) +
    geom_bar() +
    scale_fill_manual(values = c('blue', 'red')) +
    theme(legend.position = "none") +
    facet_wrap(~HeartDiseaseorAttack, labeller = labeller(HeartDiseaseorAttack = c('0' =
'No Heart Disease', '1' = 'Heart Disease')), strip.position = "top") +
    labs(title = "Heart Disease and Heavy Alchohol Consumption", x = 'Heavy Alcohol Cons
umption?', y = 'Count') +
    scale_y_continuous(labels = label_number(scale = .1))
```



This is interesting. Although it is only by a small difference, it looks as though people with heart disease consume less alcohol than people without heart disease. Does this mean that there is a negative correlation between heavy drinking and heart disease?

In the next section, we will build a model to see which risk factors have the strongest relationships with heart disease, and check to see if they agree with our visualizations.

Model Building

From the hypothesis testings, `HighBP`, `HighChol`, `Smoker`, `Stroke`, `PhysActivity`, `Veggies`, `HvyAlcoholConsump`, `AnyHealthcare`, `NoDocbcCost`, `DiffWalk`, `Sex`, and `Age` have a very low p value and can be used as predictors in the model. But based on relevancy and from the plots, the list of predictors can be brought down to `HighBP`, `HighChol`, `Smoker`,`HvyAlcoholConsump`, `Sex` and `Age`.

General Logistic Regression Model

```
# create a general model
logistic_model <- glm(HeartDiseaseorAttack~HighBP+HighChol+Age+Sex+Smoker+HvyAlcoholCons
ump, family="binomial", data=filtered_data)
probabilities <- predict(logistic_model, type = "response")

predicted.classes <- ifelse(probabilities > 0.5, "pos", "neg")

head(predicted.classes)
```

```
##     1     2     3     4     5     6 
## "pos" "pos" "pos" "pos" "pos" "pos"
```

```
summary(logistic_model)
```

```
## 
## Call:
## glm(formula = HeartDiseaseorAttack ~ HighBP + HighChol + Age + 
##     Sex + Smoker + HvyAlcoholConsump, family = "binomial", data = filtered_data)
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)       -3.903990   0.047422  -82.33   <2e-16 ***
## HighBP             0.834735   0.022553   37.01   <2e-16 ***
## HighChol           0.726806   0.022003   33.03   <2e-16 ***
## Age                0.262807   0.004361   60.26   <2e-16 ***
## Sex                0.647022   0.021386   30.25   <2e-16 ***
## Smoker             0.510640   0.021444   23.81   <2e-16 ***
## HvyAlcoholConsump -0.633666   0.050883  -12.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 65343  on 47134  degrees of freedom
## Residual deviance: 52548  on 47128  degrees of freedom
## AIC: 52562
## 
## Number of Fisher Scoring iterations: 4
```

The summary of our model reveals interesting information. The performance of a logistic regression is evaluated with specific key metrics.

The low p-value for most explanatory variables indicates that there is a significant relationship between them and the response variable. This means that the null hypothesis can be rejected for these variables. ‘Age2’ have a p value higher than 0.05 indicating the irrelevance of the variable in predicting the target variable.

AIC (Akaike Information Criteria): This is the equivalent of R2 in logistic regression. It measures the fit when a penalty is applied to the number of parameters. Smaller AIC values indicate the model is closer to the truth. This model has an AIC value of 52451.

Null deviance: Explains how well the response is predicted by the model only with the intercept. Here, the null deviance value is 65284 on 47092 number of freedoms. The very high null deviance value indicates a lack of fit of the model . We can interpret it as a Chi-square value (fitted value different from the actual value hypothesis testing).

testing).

Residual Deviance: It explains how well the response is predicted by a model with all the variables. Similar to the null deviance value, a high value of residual deviance 52415 also indicates a lack of fit of the model. It is also interpreted as a Chi-square hypothesis testing.

Number of Fisher Scoring iterations: Number of iterations before converging. Here, there are 5 fisher scoring iterations.

From the plot of Male Vs Female, it looks like Females have less prevelance to Heart-Disease than Males. Also, the health factors are different for both sexes. Therefore, it is better to build a model for each sexes separately.

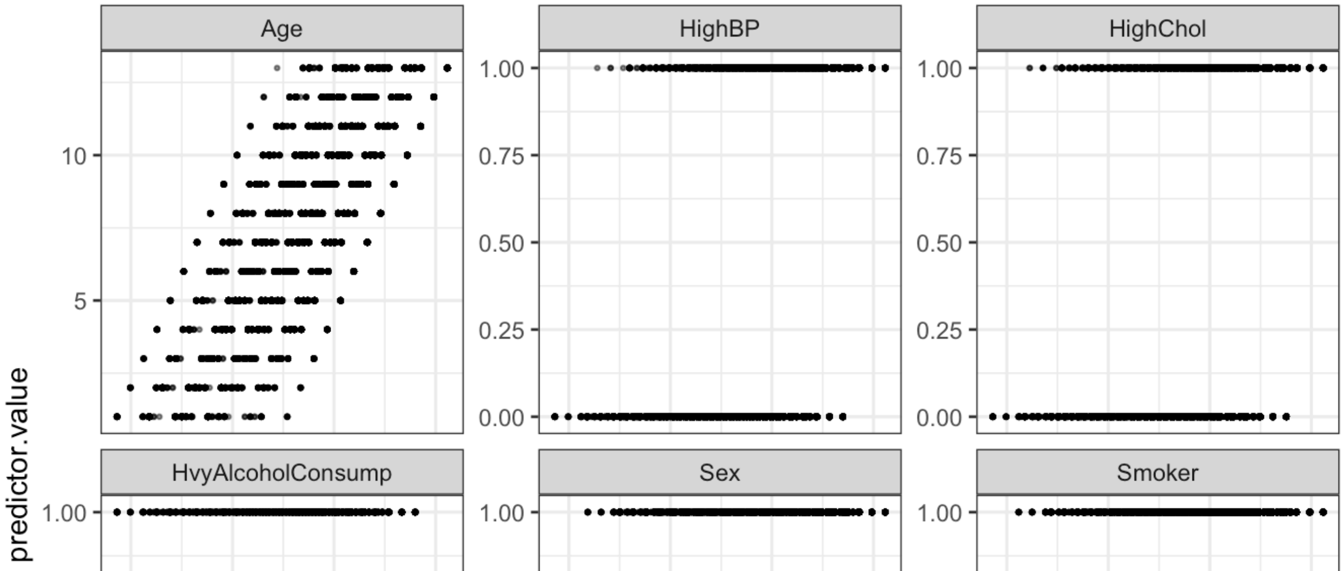
Diagnostic plot for General logistic regression

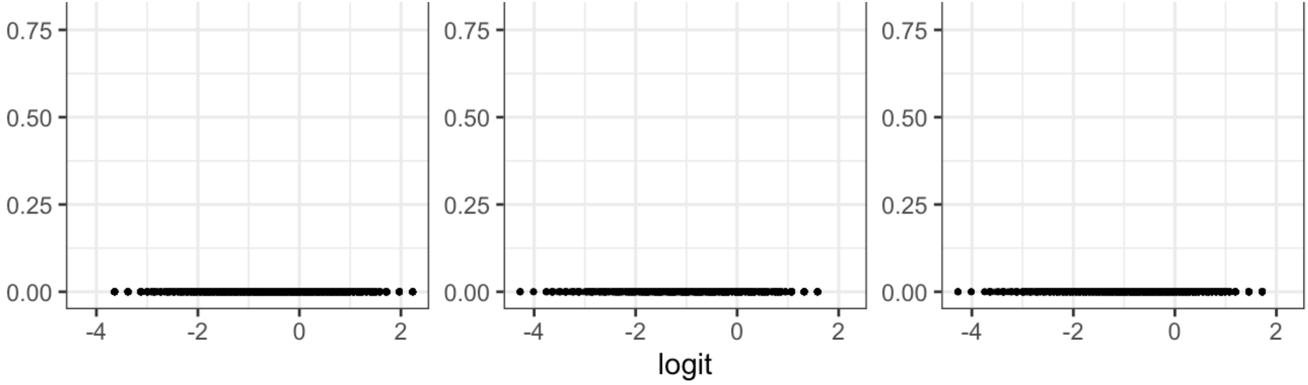
```
# tidying the data for plots
Columns1 <- c('HighChol','HighBP','Smoker','Age','HvyAlcoholConsump', 'Sex')
data_plot <- filtered_data[,(names(filtered_data) %in% Columns1)]
predictors <- colnames(data_plot)

data_plot <- data_plot %>%
  mutate(logit = log(probabilities/(1-probabilities))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)
```

```
ggplot(data_plot, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Computation failed in `stat_smooth()`
## Computation failed in `stat_smooth()`
## Computation failed in `stat_smooth()`
## Computation failed in `stat_smooth()`
## Computation failed in `stat_smooth()`
## Computation failed in `stat_smooth()`
## Caused by error in `predLoess()`:
## ! workspace required (3333045520) is too large probably because of setting 'se = TRU
E'.
```





This plots illustrates the relationship between the probability of heart disease and various predictors, including "HighBP," "HighChol," "Smoker," "Age," "Sex," and "HvyAlcoholConsump."

Analysis of the plots reveals compelling insights into how these predictors influence the likelihood of heart disease:

High Blood Pressure (HighBP): As the level of HighBP increases, there is a noticeable rise in the probability of heart disease. This suggests a positive correlation between high blood pressure and the likelihood of developing heart issues.

High Cholesterol (HighChol): Similarly, an escalation in High Cholesterol levels corresponds to an increased probability of heart disease. This underscores the significance of cholesterol levels as a potential risk factor for cardiovascular issues.

Smoking (Smoker): The predictor "Smoker" exhibits a discernible but relatively modest impact on the probability of heart disease. It suggests that smoking, while contributing to the likelihood of heart disease, may not be as potent a predictor as other factors.

Age: The plot for age demonstrates a clear trend – as age advances, the probability of heart disease rises significantly. This aligns with the well-established understanding that age is a crucial determinant in cardiovascular health.

Heavy Alcohol Consumption (HvyAlcoholConsump): Surprisingly, the plot indicates that an increase in heavy alcohol consumption is associated with a slight decrease in the probability of heart disease. This unexpected finding may warrant further investigation and consideration of potential confounding variables.

Sex: The plot for sex underscores a noteworthy gender-based difference. It indicates that the probability of heart disease is higher for females compared to males. This highlights the importance of considering gender as a relevant factor in assessing heart disease risk.

Male Logistic Regression Model

In this section, we will use a logistic regression to attempt to model the relationship between the response variable: heart disease, and the explanatory variables: high blood pressure, high cholesterol, smoking, age, and sex in Males.

Subsetting the data set for Modelling. As the health factors of males and females are different, it is better to subset the data based on sex.

```
# subsetting the male sex
data_male <- filtered_data[(filtered_data$Sex == 0), ]

# subsetting the female sex
data_female <- filtered_data[(filtered_data$Sex == 1), ]
```

```
# binomial glm with v as the predictor for data_male
data_male$HighBP <- as.factor(data_male$HighBP)
data_male$HighChol <- as.factor(data_male$HighChol)
data_male$Smoker <- as.factor(data_male$Smoker)
data_male$Age <- as.factor(data_male$Age)
data_male$HvyAlcoholConsump <- as.factor(data_male$HvyAlcoholConsump)

logistic_model1 <- glm(HeartDiseaseorAttack~HighBP+HighChol+Smoker+Age+HvyAlcoholConsum
p, data = data_male, family = 'binomial')

probabilities1 <- predict(logistic_model1, type = "response")

predicted1.classes <- ifelse(probabilities1 > 0.5, "pos", "neg")

head(predicted1.classes)
```

```
##     1     2     3     4    10    13 
## "pos" "pos" "pos" "pos" "pos" "pos"
```

```
summary(logistic_model1)
```

```
## 
## Call:
## glm(formula = HeartDiseaseorAttack ~ HighBP + HighChol + Smoker + 
##     Age + HvyAlcoholConsump, family = "binomial", data = data_male)
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        -3.46413    0.29542 -11.726  < 2e-16 ***
## HighBP1             0.99902    0.03208  31.139  < 2e-16 ***
## HighChol1           0.65372    0.03102  21.076  < 2e-16 ***
## Smoker1             0.51346    0.02990  17.175  < 2e-16 ***
## Age2                0.36690    0.35349   1.038  0.29930    
## Age3                0.64718    0.32465   1.993  0.04621 *  
## Age4                0.77882    0.31497   2.473  0.01341 *  
## Age5                0.84085    0.30936   2.718  0.00657 ** 
## Age6                1.35130    0.30326   4.456 8.35e-06 ***
## Age7                1.62441    0.30026   5.410 6.30e-08 ***
## Age8                1.68038    0.29916   5.617 1.94e-08 ***
## Age9                1.91614    0.29875   6.414 1.42e-10 ***
## Age10               1.96834    0.29865   6.591 4.37e-11 ***
## Age11               2.26234    0.29900   7.566 3.84e-14 ***
## Age12               2.46163    0.29968   8.214  < 2e-16 ***
## Age13               2.93100    0.29889   9.806  < 2e-16 ***
## HvyAlcoholConsump1 -0.77296    0.07939  -9.736  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 32524  on 23884  degrees of freedom
## Residual deviance: 26977  on 23868  degrees of freedom
## AIC: 27011
##
## Number of Fisher Scoring iterations: 5
```

The low p-value for most explanatory variables indicates that there is a significant relationship between them and the response variable. This means that the null hypothesis can be rejected for these variables. ‘Age2’ have a p value higher than 0.05 indicating the irrelevance of the variable in predicting the target variable.

logistic_model1 model have an AIC value of 26681 which is approximately half of the AIC score of logistic_model for which the sexes where not separated. This indicates that separating sexes makes the model better for predicting the reponse variable. The null deviance value is 32343 on 23720 number of freedoms, which is much lower than that of logistic_model indicating a better fit of the model. The lower residual deviance value of 26647 also indicates a better fit of the model. There are 5 fisher scoring iterations.

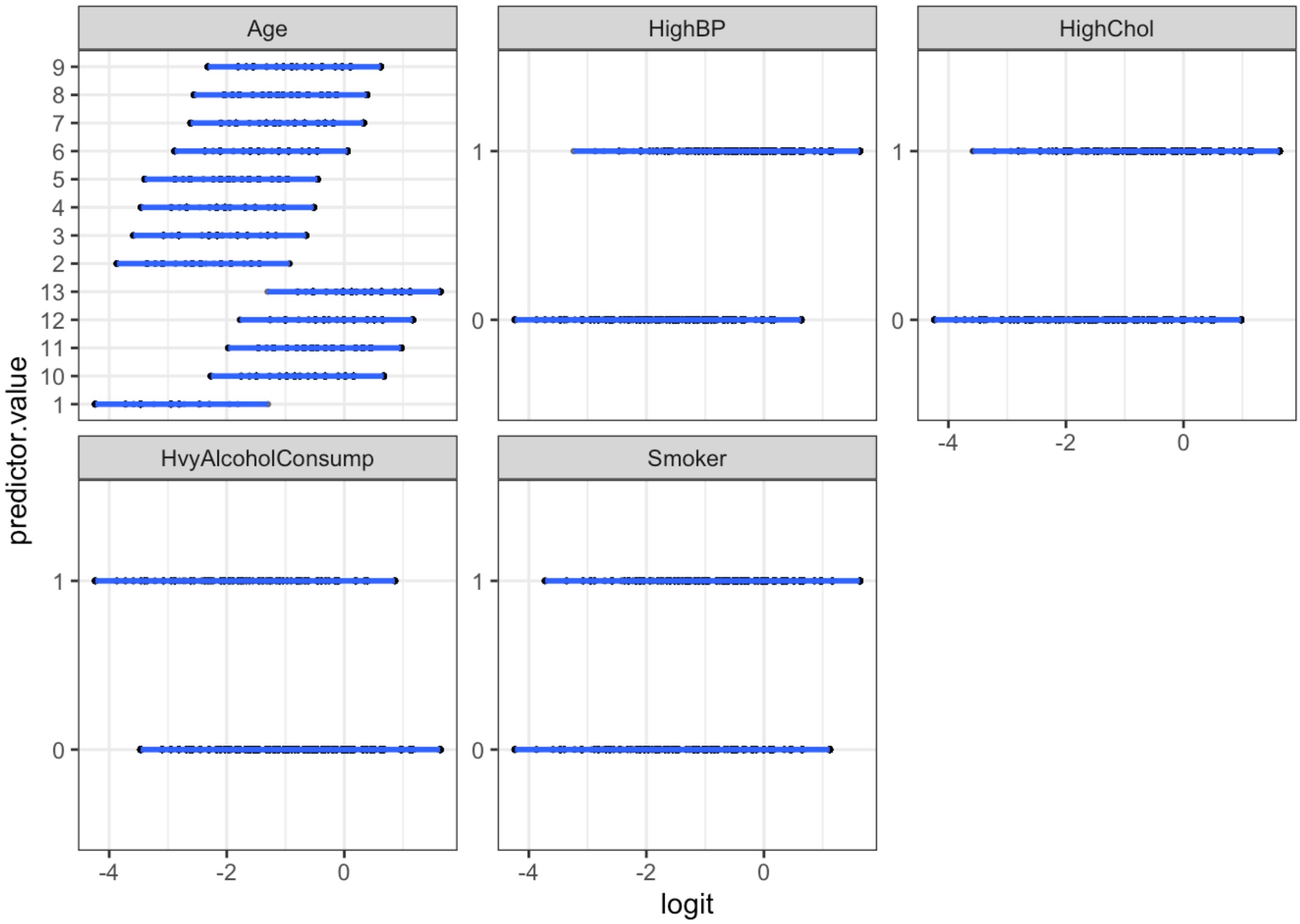
Diagnostic Plots for the Male Model

```
# tidying the data for plots
Columns1 <- c('HighChol','HighBP','Smoker','Age','HvyAlcoholConsump')
data_plot1 <- data_male[,(names(data_male) %in% Columns1)]
predictors1 <- colnames(data_plot1)

data_plot1 <- data_plot1 %>%
  mutate(logit = log(probabilities1/(1-probabilities1))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)
```

```
## Warning: attributes are not identical across measure variables; they will be
## dropped
```

```
ggplot(data_plot1, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This analysis explores the complex connection between various predictors and the likelihood of heart disease. The examination of critical factors like High Blood Pressure (HighBP) reveals a distinct positive association, signifying a significant contribution to the probability of cardiovascular issues. Likewise, an increase in High Cholesterol (HighChol) levels emphasizes its role as a noteworthy risk factor for heart disease. While the impact of smoking is noticeable, its predictive strength seems relatively moderate, suggesting that other factors may overshadow its influence. The age plot reaffirms a well-established pattern – a notable rise in the probability of heart disease as age advances, highlighting age as a crucial determinant in cardiovascular health. Unexpectedly, the plot for Heavy Alcohol Consumption (HvyAlcoholConsump) introduces a surprising discovery, indicating a slight reduction in the probability of heart disease with increased alcohol consumption, prompting further exploration of potential confounding variables. Together, these findings offer a comprehensive comprehension of the intricate relationship between predictors and the probability of heart disease.

Female Logistic Regression Model

In this section, we will use a logistic regression to attempt to model the relationship between the response variable: heart disease, and the explanatory variables: high blood pressure, high cholesterol, smoking, age, and sex in Females.

Prediction target - ‘HeartDiseaseorAttack’

Predictors - ‘HighBP’, ‘HighChol’, ‘Smoker’, ‘Age’,‘HvyAlcoholConsump’

```
# binomial glm with v as the predictor for data_female
data_female$HighBP <- as.factor(data_female$HighBP)
data_female$HighChol <- as.factor(data_female$HighChol)
data_female$Smoker <- as.factor(data_female$Smoker)
data_female$Age <- as.factor(data_female$Age)
data_female$HvyAlcoholConsump <- as.factor(data_female$HvyAlcoholConsump)

logistic_model2 <- glm(HeartDiseaseorAttack~HighBP+HighChol+Smoker+Age+HvyAlcoholConsum
p, data = data_female, family = 'binomial')
probabilities2 <- predict(logistic_model2, type = "response")
predicted2.classes <- ifelse(probabilities2 > 0.5, "pos", "neg")
head(predicted2.classes)
```

```
##     5     6     7     8     9    11
## "pos" "pos" "neg" "pos" "pos" "pos"
```

```
summary(logistic_model2)
```

```
## 
## Call:
## glm(formula = HeartDiseaseorAttack ~ HighBP + HighChol + Smoker + 
##     Age + HvyAlcoholConsump, family = "binomial", data = data_female)
## 
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         -3.43548    0.26743 -12.846  < 2e-16 ***
## HighBP1              0.68539    0.03242  21.142  < 2e-16 ***
## HighChol1            0.81079    0.03190  25.419  < 2e-16 ***
## Smoker1              0.49552    0.03137  15.798  < 2e-16 ***
## Age2                 0.20193    0.33906   0.596   0.5515    
## Age3                 0.58881    0.30592   1.925   0.0543 .  
## Age4                 0.69107    0.29277   2.360   0.0183 *  
## Age5                 1.23561    0.28115   4.395 1.11e-05 ***
## Age6                 1.68619    0.27547   6.121 9.29e-10 ***
## Age7                 1.98019    0.27198   7.281 3.33e-13 ***
## Age8                 2.28810    0.27082   8.449  < 2e-16 ***
## Age9                 2.66007    0.27004   9.851  < 2e-16 ***
## Age10                2.97738    0.26996  11.029  < 2e-16 ***
## Age11                3.20385    0.27073  11.834  < 2e-16 ***
## Age12                3.37939    0.27225  12.413  < 2e-16 ***
## Age13                3.69745    0.27198  13.595  < 2e-16 ***
## HvyAlcoholConsump1 -0.51616    0.06785  -7.607 2.80e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 31592  on 23249  degrees of freedom
## Residual deviance: 25329  on 23233  degrees of freedom
## AIC: 25363
## 
## Number of Fisher Scoring iterations: 5
```

Similar to logistic_model and logistc_model1, the low p-value for most explanatory variables indicates that there is a significant relationship between the explanatory and the response variables. This means that the null hypothesis can be rejected for these variables. `Age2` and `Age3` have a p value higher than 0.05 indicating the irrelevance of the variables in predicting the target variable.

logistic_model2 model have an AIC value of 25603 which is lower than both logistic_model and logistic_model1. This indicates that the model predicts the response variable better.

The null deviance value is 31804 on 23371 number of freedoms, which is much lower than that of logistic_model indicating a better fit of the model.

The lower residual deviance value of 25569 also indicates a better fit of the model.

There are 5 fisher scoring iterations.

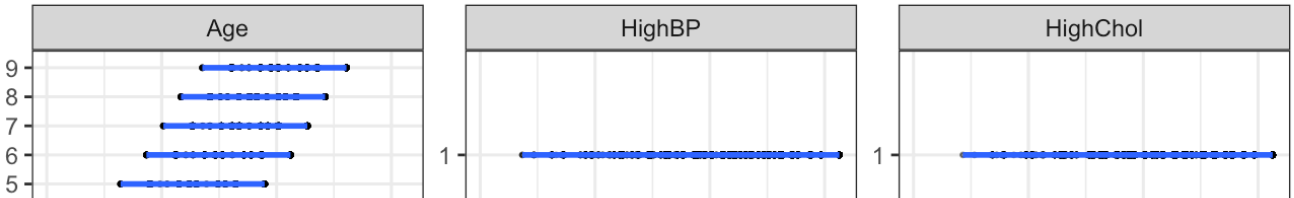
Diagnostic Plots of the Female Model

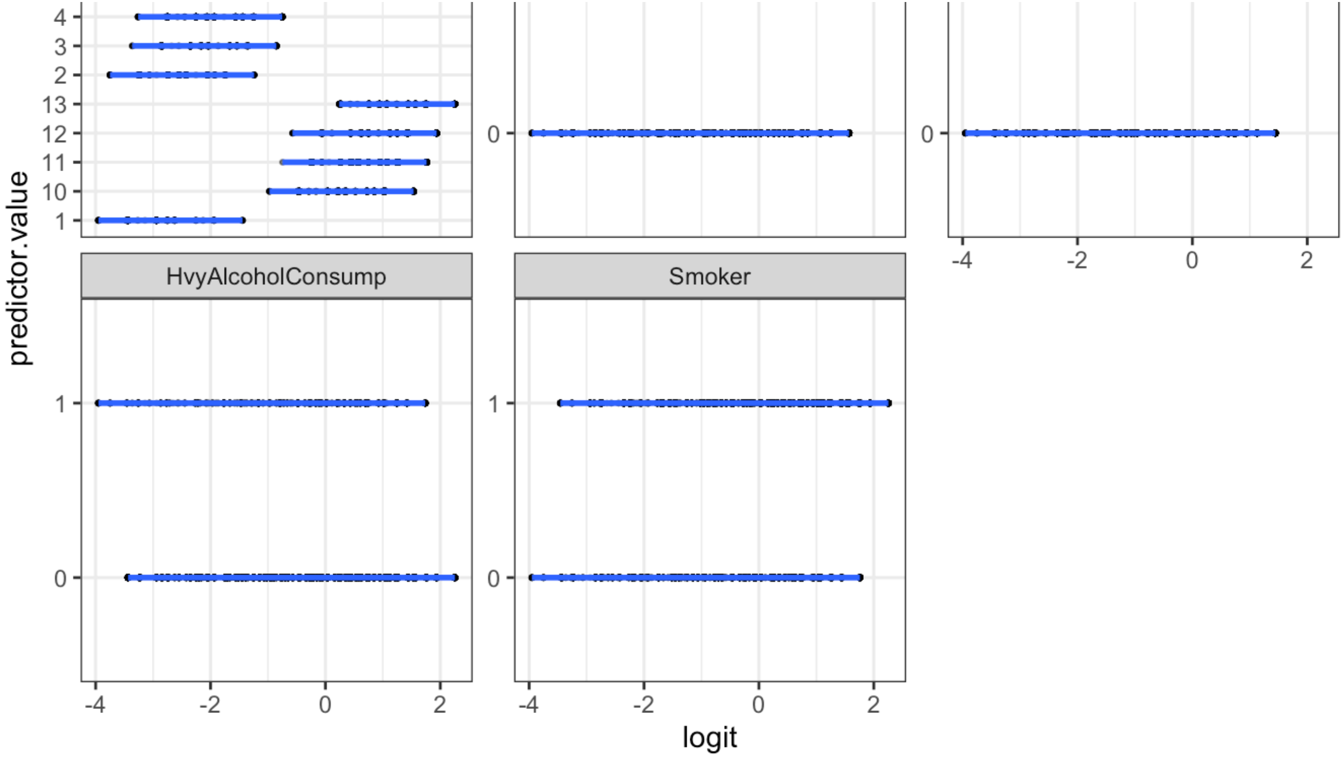
```
# tidying the data for plots
Columns1 <- c('HighChol','HighBP','Smoker','Age','HvyAlcoholConsump')
data_plot2 <- data_female[,(names(data_female) %in% Columns1)]
predictors2 <- colnames(data_plot2)

data_plot2 <- data_plot2 %>%
  mutate(logit = log(probabilities2/(1-probabilities2))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)
```

```
## Warning: attributes are not identical across measure variables; they will be
## dropped
```

```
ggplot(data_plot2, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```





Plots show that elevated High Blood Pressure (HighBP) and High Cholesterol (HighChol) significantly increase the likelihood. Smoking has a noticeable but moderate impact. Advancing age is a crucial factor, showing a substantial rise in heart disease probability. Unexpectedly, increased Heavy Alcohol Consumption (HvyAlcoholConsump) slightly reduces the likelihood, warranting further investigation. These insights offer a concise overview of the intricate dynamics influencing heart disease probability.

Step-wise Model Selection

The step() function can be used to iterate through predictor variables, adding them and removing them, in order to find the optimal combination that results in the best model.

General

```
# Forward selection

model_start <- glm(HeartDiseaseorAttack~1,data = filtered_data, family = 'binomial')

forward_model <- step(model_start, scope = HeartDiseaseorAttack~HighBP+HighChol+Smoker+A
ge+HvyAlcoholConsump, direction = "forward")
```

```
## Start:  AIC=65344.73
## HeartDiseaseorAttack ~ 1
##
##                       Df Deviance   AIC
## + Age                  1    57824 57828
## + HighBP               1    60475 60479
## + HighChol             1    61720 61724
## + Smoker               1    64063 64067
## + HvyAlcoholConsump  1    65130 65134
## <none>                     65343 65345
##
## Step:  AIC=57828.18
## HeartDiseaseorAttack ~ Age
##
##                     Df Deviance   AIC
## + HighBP             1    55526 55532
## + HighChol           1    55819 55825
## + Smoker             1    56912 56918
## + HvyAlcoholConsump  1    57721 57727
## <none>                     57824 57828
##
## Step:  AIC=55531.57
## HeartDiseaseorAttack ~ Age + HighBP
##
##                     Df Deviance   AIC
## + HighChol           1    54323 54331
## + Smoker             1    54739 54747
## + HvyAlcoholConsump  1    55422 55430
## <none>                     55526 55532
##
## Step:  AIC=54330.67
## HeartDiseaseorAttack ~ Age + HighBP + HighChol
##
##                     Df Deviance   AIC
## + Smoker             1    53621 53631
## + HvyAlcoholConsump  1    54224 54234
## <none>                     54323 54331
##
## Step:  AIC=53631.5
## HeartDiseaseorAttack ~ Age + HighBP + HighChol + Smoker
##
##                     Df Deviance   AIC
## + HvyAlcoholConsump  1    53476 53488
## <none>                     53621 53631
##
## Step:  AIC=53487.76
## HeartDiseaseorAttack ~ Age + HighBP + HighChol + Smoker + HvyAlcoholConsump
```

```
summary(forward_model)
```

```
##
## Call:
## glm(formula = HeartDiseaseorAttack ~ Age + HighBP + HighChol +
##     Smoker + HvyAlcoholConsump, family = "binomial", data = filtered_data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.577717   0.045111  -79.31   <2e-16 ***
## Age                0.258094   0.004314   59.83   <2e-16 ***
## HighBP             0.839433   0.022282   37.67   <2e-16 ***
## HighChol           0.724097   0.021750   33.29   <2e-16 ***
## Smoker              0.576361   0.021153   27.25   <2e-16 ***
## HvyAlcoholConsump -0.599008   0.050153  -11.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 65343  on 47134  degrees of freedom
## Residual deviance: 53476  on 47129  degrees of freedom
## AIC: 53488
##
## Number of Fisher Scoring iterations: 4
```

The forward step-wise model function selected `HeartDiseaseorAttack ~ Age + HighBP + HighChol + Sex + Smoker + HvyAlcoholConsump` as the best model for filtered_data dataframe. The AIC score for this model is 52450.95. This indicates that out of the selected variables as predictors, logistic_model is the right choice of prediction model for filtered_data.

```
# Backward elimination
backward_model <- step(model_start, scope = HeartDiseaseorAttack~HighBP+HighChol+Smoker+
Age+HvyAlcoholConsump, direction = "backward")
```

```
## Start:  AIC=65344.73
## HeartDiseaseorAttack ~ 1
```

```
summary(backward_model)
```

```
##
## Call:
## glm(formula = HeartDiseaseorAttack ~ 1, family = "binomial",
##     data = filtered_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.004625   0.009212   0.502    0.616
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 65343  on 47134  degrees of freedom
## Residual deviance: 65343  on 47134  degrees of freedom
## AIC: 65345
##
## Number of Fisher Scoring iterations: 3
```

The backward step-wise model function also selected `HeartDiseaseorAttack ~ Age + HighBP + HighChol + Sex + Smoker + HvyAlcoholConsump` as the best model for filtered_data dataframe.

```
# Both forward and backward selection
both_model <- step(model_start, scope = HeartDiseaseorAttack~HighBP+HighChol+Smoker+Age+
HvyAlcoholConsump, direction = "both")
```

```
## Start:  AIC=65344.73
## HeartDiseaseorAttack ~ 1
##
##                      Df Deviance   AIC
## + Age                 1    57824 57828
## + HighBP              1    60475 60479
## + HighChol            1    61720 61724
## + Smoker             1    64063 64067
## + HvyAlcoholConsump  1    65130 65134
## <none>                    65343 65345
##
## Step:  AIC=57828.18
## HeartDiseaseorAttack ~ Age
##
##                     Df Deviance   AIC
## + HighBP             1    55526 55532
## + HighChol           1    55819 55825
## + Smoker             1    56912 56918
## + HvyAlcoholConsump  1    57721 57727
## <none>                    57824 57828
## - Age                1    65343 65345
##
## Step:  AIC=55531.57
## HeartDiseaseorAttack ~ Age + HighBP
##
##                     Df Deviance   AIC
## + HighChol           1    54323 54331
## + Smoker             1    54739 54747
## + HvyAlcoholConsump  1    55422 55430
## <none>                    55526 55532
## - HighBP             1    57824 57828
## - Age                1    60475 60479
##
## Step:  AIC=54330.67
## HeartDiseaseorAttack ~ Age + HighBP + HighChol
##
##                     Df Deviance   AIC
## + Smoker             1    53621 53631
## + HvyAlcoholConsump  1    54224 54234
## <none>                    54323 54331
## - HighChol           1    55526 55532
## - HighBP             1    55819 55825
## - Age                1    58657 58663
##
## Step:  AIC=53631.5
## HeartDiseaseorAttack ~ Age + HighBP + HighChol + Smoker
##
##                     Df Deviance   AIC
## + HvyAlcoholConsump  1    53476 53488
## <none>                    53621 53631
## - Smoker             1    54323 54331
## - HighChol           1    54739 54747
## - HighBP             1    55043 55051
## - Age                1    57817 57825
##
## Step:  AIC=53487.76
## HeartDiseaseorAttack ~ Age + HighBP + HighChol + Smoker + HvyAlcoholConsump
##
##                     Df Deviance   AIC
## <none>                    53476 53488
## - HvyAlcoholConsump  1    53621 53631
## - Smoker             1    54224 54234
## - HighChol           1    54584 54594
## - HighBP             1    54899 54909
## - Age                1    57577 57587
```

```
summary(both_model)
```

```
##
## Call:
## glm(formula = HeartDiseaseorAttack ~ Age + HighBP + HighChol +
##     Smoker + HvyAlcoholConsump, family = "binomial", data = filtered_data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.577717   0.045111  -79.31   <2e-16 ***
## Age                0.258094   0.004314   59.83   <2e-16 ***
## HighBP             0.839433   0.022282   37.67   <2e-16 ***
## HighChol           0.724097   0.021750   33.29   <2e-16 ***
## Smoker             0.576361   0.021153   27.25   <2e-16 ***
## HvyAlcoholConsump -0.599008   0.050153  -11.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 65343  on 47134  degrees of freedom
## Residual deviance: 53476  on 47129  degrees of freedom
## AIC: 53488
##
## Number of Fisher Scoring iterations: 4
```

The step-wise model function withboth direction also selected `HeartDiseaseorAttack ~ Age + HighBP + HighChol + Sex + Smoker + HvyAlcoholConsump` as the best model for filtered_data dataframe. Therefore implementing the step function concludes that logistic_model is the best possible predicting model with the chosen explanatory variables.

Males

```
# Forward selection

model_male_start <- glm(HeartDiseaseorAttack~1,data = data_male, family = 'binomial')

forward_model_Male <- step(model_male_start, scope = HeartDiseaseorAttack~HighBP+HighCho
l+Smoker+Age+HvyAlcoholConsump, direction = "forward")
```

```
## Start:  AIC=32525.92
## HeartDiseaseorAttack ~ 1
##
##                     Df Deviance   AIC
## + Age               12    29234 29260
## + HighBP             1    29607 29611
## + HighChol           1    30805 30809
## + Smoker             1    32148 32152
## + HvyAlcoholConsump  1    32360 32364
## <none>                    32524 32526
##
## Step:  AIC=29260
## HeartDiseaseorAttack ~ Age
##
##                     Df Deviance   AIC
## + HighBP             1    27845 27873
## + HighChol           1    28371 28399
## + Smoker             1    28892 28920
## + HvyAlcoholConsump  1    29136 29164
## <none>                    29234 29260
##
```

```
##
## Step:  AIC=27872.57
## HeartDiseaseorAttack ~ Age + HighBP
##
##                     Df Deviance   AIC
## + HighChol           1    27352 27382
## + Smoker             1    27532 27562
## + HvyAlcoholConsump  1    27762 27792
## <none>                    27845 27873
##
## Step:  AIC=27382.06
## HeartDiseaseorAttack ~ Age + HighBP + HighChol
##
##                     Df Deviance   AIC
## + Smoker             1    27079 27111
## + HvyAlcoholConsump  1    27275 27307
## <none>                    27352 27382
##
## Step:  AIC=27110.93
## HeartDiseaseorAttack ~ Age + HighBP + HighChol + Smoker
##
##                     Df Deviance   AIC
## + HvyAlcoholConsump  1    26977 27011
## <none>                    27079 27111
##
## Step:  AIC=27011.13
## HeartDiseaseorAttack ~ Age + HighBP + HighChol + Smoker + HvyAlcoholConsump
```

```
summary(forward_model_Male)
```

```
##
## Call:
## glm(formula = HeartDiseaseorAttack ~ Age + HighBP + HighChol +
##     Smoker + HvyAlcoholConsump, family = "binomial", data = data_male)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.46413    0.29542 -11.726  < 2e-16 ***
## Age2                0.36690    0.35349   1.038  0.29930
## Age3                0.64718    0.32465   1.993  0.04621 *
## Age4                0.77882    0.31497   2.473  0.01341 *
## Age5                0.84085    0.30936   2.718  0.00657 **
## Age6                1.35130    0.30326   4.456 8.35e-06 ***
## Age7                1.62441    0.30026   5.410 6.30e-08 ***
## Age8                1.68038    0.29916   5.617 1.94e-08 ***
## Age9                1.91614    0.29875   6.414 1.42e-10 ***
## Age10               1.96834    0.29865   6.591 4.37e-11 ***
## Age11               2.26234    0.29900   7.566 3.84e-14 ***
## Age12               2.46163    0.29968   8.214  < 2e-16 ***
## Age13               2.93100    0.29889   9.806  < 2e-16 ***
## HighBP1             0.99902    0.03208  31.139  < 2e-16 ***
## HighChol1           0.65372    0.03102  21.076  < 2e-16 ***
## Smoker1             0.51346    0.02990  17.175  < 2e-16 ***
## HvyAlcoholConsump1  0.77296    0.07939   9.736  < 2e-16 ***
## HvyAlcoholConsump1 -0.77296    0.07939  -9.736  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32524  on 23884  degrees of freedom
## Residual deviance: 26977  on 23868  degrees of freedom
## AIC: 27011
##
## Number of Fisher Scoring iterations: 5
```

The forward step function returns `HeartDiseaseorAttack ~ Age + HighBP + HighChol + Smoker + HvyAlcoholConsump` as the best fit model for data_male with the selected predictors. This is same as that of logistic_model1 defined above.

```
# Backward elimination
backward_model_male <- step(logistic_model1, direction = "backward")
```

```
## Start:  AIC=27011.13
## HeartDiseaseorAttack ~ HighBP + HighChol + Smoker + Age + HvyAlcoholConsump
##
##                     Df Deviance   AIC
## <none>                    26977 27011
## - HvyAlcoholConsump  1    27079 27111
## - Smoker             1    27275 27307
## - HighChol           1    27423 27455
## - HighBP             1    27969 28001
## - Age               12    28471 28481
```

```
summary(backward_model_male)
```

```
##
## Call:
## glm(formula = HeartDiseaseorAttack ~ HighBP + HighChol + Smoker +
##     Age + HvyAlcoholConsump, family = "binomial", data = data_male)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.46413    0.29542 -11.726  < 2e-16 ***
## HighBP1              0.99902    0.03208  31.139  < 2e-16 ***
## HighChol1            0.65372    0.03102  21.076  < 2e-16 ***
## Smoker1              0.51346    0.02990  17.175  < 2e-16 ***
## Age2                 0.36690    0.35349   1.038  0.29930
## Age3                 0.64718    0.32465   1.993  0.04621 *
## Age4                 0.77882    0.31497   2.473  0.01341 *
## Age5                 0.84085    0.30936   2.718  0.00657 **
## Age6                 1.35130    0.30326   4.456 8.35e-06 ***
## Age7                 1.62441    0.30026   5.410 6.30e-08 ***
## Age8                 1.68038    0.29916   5.617 1.94e-08 ***
## Age9                 1.91614    0.29875   6.414 1.42e-10 ***
## Age10                1.96834    0.29865   6.591 4.37e-11 ***
## Age11                2.26234    0.29900   7.566 3.84e-14 ***
## Age12                2.46163    0.29968   8.214  < 2e-16 ***
## Age13                2.93100    0.29889   9.806  < 2e-16 ***
## HvyAlcoholConsump1 -0.77296    0.07939  -9.736  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32524  on 23884  degrees of freedom
## Residual deviance: 26977  on 23868  degrees of freedom
## AIC: 27011
##
## Number of Fisher Scoring iterations: 5
```

The backward model also returned logistic_model1 as the best fit model for predicting response variable for the data_male data frame.

```
# Both forward and backward selection
both_model_male <- step(model_male_start, scope = HeartDiseaseorAttack~HighBP+HighChol+S
moker+Age+HvyAlcoholConsump, direction = "both")
```

```
## Start:  AIC=32525.92
## HeartDiseaseorAttack ~ 1
##
##                     Df Deviance   AIC
## + Age               12    29234 29260
## + HighBP             1    29607 29611
## + HighChol           1    30805 30809
## + Smoker             1    32148 32152
## + HvyAlcoholConsump  1    32360 32364
## <none>                    32524 32526
##
## Step:  AIC=29260
## HeartDiseaseorAttack ~ Age
##

##                     Df Deviance   AIC
## + HighBP             1    27845 27873
## + HighChol           1    28371 28399
## + Smoker             1    28892 28920
## + HvyAlcoholConsump  1    29136 29164
## <none>                    29234 29260
## - Age               12    32524 32526
##
## Step:  AIC=27872.57
## HeartDiseaseorAttack ~ Age + HighBP
##
##                     Df Deviance   AIC
## + HighChol           1    27352 27382
## + Smoker             1    27532 27562
## + HvyAlcoholConsump  1    27762 27792
## <none>                    27845 27873
## - HighBP             1    29234 29260
## - Age               12    29607 29611
##
## Step:  AIC=27382.06
## HeartDiseaseorAttack ~ Age + HighBP + HighChol
##
##                     Df Deviance   AIC
## + Smoker             1    27079 27111
## + HvyAlcoholConsump  1    27275 27307
## <none>                    27352 27382
## - HighChol           1    27845 27873
## - HighBP             1    28371 28399
## - Age               12    28845 28851
##
```

```
## Step:  AIC=27110.93
## HeartDiseaseorAttack ~ Age + HighBP + HighChol + Smoker
##
##                     Df Deviance   AIC
## + HvyAlcoholConsump  1    26977 27011
## <none>                    27079 27111
## - Smoker             1    27352 27382
## - HighChol           1    27532 27562
## - HighBP             1    28082 28112
## - Age               12    28608 28616
##
## Step:  AIC=27011.13
## HeartDiseaseorAttack ~ Age + HighBP + HighChol + Smoker + HvyAlcoholConsump
##
##                     Df Deviance   AIC
## <none>                    26977 27011
## - HvyAlcoholConsump  1    27079 27111
## - Smoker             1    27275 27307
## - HighChol           1    27423 27455
## - HighBP             1    27969 28001
## - Age               12    28471 28481
```

```
summary(both_model_male)
```

```
##
## Call:
## glm(formula = HeartDiseaseorAttack ~ Age + HighBP + HighChol +
##     Smoker + HvyAlcoholConsump, family = "binomial", data = data_male)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.46413    0.29542 -11.726  < 2e-16 ***
## Age2                0.36690    0.35349   1.038  0.29930
## Age3                0.64718    0.32465   1.993  0.04621 *
## Age4                0.77882    0.31497   2.473  0.01341 *
## Age5                0.84085    0.30936   2.718  0.00657 **
## Age6                1.35130    0.30326   4.456 8.35e-06 ***
## Age7                1.62441    0.30026   5.410 6.30e-08 ***
## Age8                1.68038    0.29916   5.617 1.94e-08 ***
## Age9                1.91614    0.29875   6.414 1.42e-10 ***
## Age10               1.96834    0.29865   6.591 4.37e-11 ***
## Age11               2.26234    0.29900   7.566 3.84e-14 ***
## Age12               2.46163    0.29968   8.214  < 2e-16 ***
## Age13               2.93100    0.29889   9.806  < 2e-16 ***
## HighBP1             0.99902    0.03208  31.139  < 2e-16 ***
## HighChol1           0.65372    0.03102  21.076  < 2e-16 ***
## Smoker1             0.51346    0.02990  17.175  < 2e-16 ***
## HvyAlcoholConsump1 -0.77296    0.07939  -9.736  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32524  on 23884  degrees of freedom
## Residual deviance: 26977  on 23868  degrees of freedom
## AIC: 27011
##
```

```
## Number of Fisher Scoring iterations: 5
```

Step function in ‘Both’ directions also selected the logistic_model1 as the best model for the selected predictors.

Females

```
# Forward selection in data_female

model_female_start <- glm(HeartDiseaseorAttack~1,data = data_female, family = 'binomia
l')
forward_model_female <- step(model_female_start, scope = HeartDiseaseorAttack~HighBP+Hig
hChol+Smoker+Age+HvyAlcoholConsump, direction = "forward")
```

```
## Start:  AIC=31593.54
## HeartDiseaseorAttack ~ 1
## 
##                     Df Deviance   AIC
## + Age               12    27097 27123
## + HighBP             1    29742 29746
## + HighChol           1    29786 29790
## + Smoker             1    30925 30929
## + HvyAlcoholConsump  1    31506 31510
## <none>                    31592 31594
## 
## Step:  AIC=27123
## HeartDiseaseorAttack ~ Age
## 
##                     Df Deviance   AIC
## + HighChol           1    26093 26121
## + HighBP             1    26288 26316
## + Smoker             1    26793 26821
## + HvyAlcoholConsump  1    27060 27088
## <none>                    27097 27123
## 
## Step:  AIC=26120.99
## HeartDiseaseorAttack ~ Age + HighChol
## 
##                     Df Deviance   AIC
## + HighBP             1    25619 25649
## + Smoker             1    25826 25856
## + HvyAlcoholConsump  1    26057 26087
## <none>                    26093 26121
## 
## Step:  AIC=25649.16
## HeartDiseaseorAttack ~ Age + HighChol + HighBP
## 
##                     Df Deviance   AIC
## + Smoker             1    25387 25419
## + HvyAlcoholConsump  1    25578 25610
## <none>                    25619 25649
##
## Step:  AIC=25418.77
## HeartDiseaseorAttack ~ Age + HighChol + HighBP + Smoker
##
##                      Df Deviance   AIC
## + HvyAlcoholConsump  1    25329 25363
## <none>                    25387 25419
##
## Step:  AIC=25363.19
## HeartDiseaseorAttack ~ Age + HighChol + HighBP + Smoker + HvyAlcoholConsump
```

```
summary(forward_model_female)
```

```
##
## Call:
## glm(formula = HeartDiseaseorAttack ~ Age + HighChol + HighBP +
##     Smoker + HvyAlcoholConsump, family = "binomial", data = data_female)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.43548    0.26743 -12.846  < 2e-16 ***
## Age2                0.20193    0.33906   0.596   0.5515
## Age3                0.58881    0.30592   1.925   0.0543 .
## Age4                0.69107    0.29277   2.360   0.0183 *
## Age5                1.23561    0.28115   4.395 1.11e-05 ***
## Age6                1.68619    0.27547   6.121 9.29e-10 ***
## Age7                1.98019    0.27198   7.281 3.33e-13 ***
## Age8                2.28810    0.27082   8.449  < 2e-16 ***
## Age9                2.66007    0.27004   9.851  < 2e-16 ***
## Age10               2.97738    0.26996  11.029  < 2e-16 ***
## Age11               3.20385    0.27073  11.834  < 2e-16 ***
## Age12               3.37939    0.27225  12.413  < 2e-16 ***
## Age13               3.69745    0.27198  13.595  < 2e-16 ***
## HighChol1           0.81079    0.03190  25.419  < 2e-16 ***
## HighBP1             0.68539    0.03242  21.142  < 2e-16 ***
## Smoker1             0.49552    0.03137  15.798  < 2e-16 ***
## HvyAlcoholConsump1 -0.51616    0.06785  -7.607 2.80e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31592  on 23249  degrees of freedom
## Residual deviance: 25329  on 23233  degrees of freedom
## AIC: 25363
##
## Number of Fisher Scoring iterations: 5
```

The forward step function returned the logistic_model2 as the best model for predicting response variable for the data_female dataframe.

```
# Backward selection for data_female
backward_model_female <- step(logistic_model2, direction = "backward")
```

```
## Start:  AIC=25363.19
## HeartDiseaseorAttack ~ HighBP + HighChol + Smoker + Age + HvyAlcoholConsump
## 
##                     Df Deviance   AIC
## <none>                    25329 25363
## - HvyAlcoholConsump  1    25387 25419
## - Smoker             1    25578 25610
## - HighBP             1    25773 25805
## - HighChol           1    25977 26009
## - Age               12    28156 28166
```

```
summary(backward_model_female)
```

```
## 
## Call:
## glm(formula = HeartDiseaseorAttack ~ HighBP + HighChol + Smoker + 
##     Age + HvyAlcoholConsump, family = "binomial", data = data_female)
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        -3.43548    0.26743 -12.846  < 2e-16 ***
## HighBP1             0.68539    0.03242  21.142  < 2e-16 ***
## HighChol1           0.81079    0.03190  25.419  < 2e-16 ***
## Smoker1             0.49552    0.03137  15.798  < 2e-16 ***
## Age2                0.20193    0.33906   0.596   0.5515    
## Age3                0.58881    0.30592   1.925   0.0543 .  
## Age4                0.69107    0.29277   2.360   0.0183 *  
## Age5                1.23561    0.28115   4.395 1.11e-05 ***
## Age6                1.68619    0.27547   6.121 9.29e-10 ***
## Age7                1.98019    0.27198   7.281 3.33e-13 ***
## Age8                2.28810    0.27082   8.449  < 2e-16 ***
## Age9                2.66007    0.27004   9.851  < 2e-16 ***
## Age10               2.97738    0.26996  11.029  < 2e-16 ***
## Age11               3.20385    0.27073  11.834  < 2e-16 ***
## Age12               3.37939    0.27225  12.413  < 2e-16 ***
## Age13               3.69745    0.27198  13.595  < 2e-16 ***
## HvyAlcoholConsump1 -0.51616    0.06785  -7.607 2.80e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 31592  on 23249  degrees of freedom
## Residual deviance: 25329  on 23233  degrees of freedom
## AIC: 25363
## 
## Number of Fisher Scoring iterations: 5
```

```
# both selection in data_female

both_model_female <- step(model_female_start, scope = HeartDiseaseorAttack~HighBP+HighCh
ol+Smoker+Age+HvyAlcoholConsump, direction = "both")
```

```
## Start:  AIC=31593.54
## Start:  AIC=31593.54
## HeartDiseaseorAttack ~ 1
## 
##                      Df Deviance   AIC
## + Age               12    27097 27123
## + HighBP             1    29742 29746
## + HighChol           1    29786 29790
## + Smoker             1    30925 30929
## + HvyAlcoholConsump  1    31506 31510
## <none>                    31592 31594
## 
## Step:  AIC=27123
## HeartDiseaseorAttack ~ Age
## 
##                      Df Deviance   AIC
## + HighChol           1    26093 26121
## + HighBP             1    26288 26316
## + Smoker             1    26793 26821
## + HvyAlcoholConsump  1    27060 27088
## <none>                    27097 27123
## - Age               12    31592 31594
## 
## Step:  AIC=26120.99
## HeartDiseaseorAttack ~ Age + HighChol
## 
##                      Df Deviance   AIC
## + HighBP             1    25619 25649
## + Smoker             1    25826 25856
## + HvyAlcoholConsump  1    26057 26087
## <none>                    26093 26121
## - HighChol           1    27097 27123
## - Age               12    29786 29790
## 
## Step:  AIC=25649.16
## HeartDiseaseorAttack ~ Age + HighChol + HighBP
## 
##                      Df Deviance   AIC
## + Smoker             1    25387 25419
## + HvyAlcoholConsump  1    25578 25610
## <none>                    25619 25649
## - HighBP             1    26093 26121
## - HighChol           1    26288 26316
## - Age               12    28718 28724
## 
## Step:  AIC=25418.77
## HeartDiseaseorAttack ~ Age + HighChol + HighBP + Smoker
## 
##                      Df Deviance   AIC
## + HvyAlcoholConsump  1    25329 25363
## <none>                    25387 25419
## - Smoker             1    25619 25649
## - HighBP             1    25826 25856
## - HighChol           1    26037 26067
## - Age               12    28274 28282
## 
## Step:  AIC=25363.19
## HeartDiseaseorAttack ~ Age + HighChol + HighBP + Smoker + HvyAlcoholConsump
## 
##                      Df Deviance   AIC
## <none>                    25329 25363
## - HvyAlcoholConsump  1    25387 25419
## - Smoker             1    25578 25610
## - Smoker              1    25578 25610
## - HighBP              1    25773 25805
## - HighChol            1    25977 26009
## - Age                12    28156 28166
```

```
summary(both_model_female)
```

```
##
## Call:
## glm(formula = HeartDiseaseorAttack ~ Age + HighChol + HighBP +
##     Smoker + HvyAlcoholConsump, family = "binomial", data = data_female)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        -3.43548    0.26743 -12.846  < 2e-16 ***
## Age2                0.20193    0.33906   0.596   0.5515    
## Age3                0.58881    0.30592   1.925   0.0543 .  
## Age4                0.69107    0.29277   2.360   0.0183 *  
## Age5                1.23561    0.28115   4.395 1.11e-05 ***
## Age6                1.68619    0.27547   6.121 9.29e-10 ***
## Age7                1.98019    0.27198   7.281 3.33e-13 ***
## Age8                2.28810    0.27082   8.449  < 2e-16 ***
## Age9                2.66007    0.27004   9.851  < 2e-16 ***
## Age10               2.97738    0.26996  11.029  < 2e-16 ***
## Age11               3.20385    0.27073  11.834  < 2e-16 ***
## Age12               3.37939    0.27225  12.413  < 2e-16 ***
## Age13               3.69745    0.27198  13.595  < 2e-16 ***
## HighChol1           0.81079    0.03190  25.419  < 2e-16 ***
## HighBP1             0.68539    0.03242  21.142  < 2e-16 ***
## Smoker1             0.49552    0.03137  15.798  < 2e-16 ***
## HvyAlcoholConsump1 -0.51616    0.06785  -7.607 2.80e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31592  on 23249  degrees of freedom
## Residual deviance: 25329  on 23233  degrees of freedom
## AIC: 25363
##
## Number of Fisher Scoring iterations: 5
```

Similar to the forward step function, both backward and both directions in step function returned the logistic_model2 as the best model for prediction.

Predictions using the General Logistic Model

Next, we can use our model to predict the probability of success (heart disease) with different values for the explanatory variables.

Here are a few different combinations:

Let’s compare the results between smokers and non-smokers.

```
# smoker
predict(logistic_model, data.frame(HighBP=0, HighChol=0, Smoker=1, Age=7, HvyAlcoholCons
ump=0, Sex = 1))
```

```
##          1 
## -0.9066775
```

```
# non-smoker
predict(logistic_model, data.frame(HighBP=0, HighChol=0, Smoker=0, Age=7, HvyAlcoholCons
ump=0, Sex = 1))
```

```
##         1 
## -1.417317
```

The results show that when all other variables are the same, smoking does slightly increase the risk of having heart disease.

Let’s now compare the risk for people with and without high cholesterol.

```
# with
predict(logistic_model, data.frame(HighBP=1, HighChol=1, Smoker=1, Age=60, HvyAlcoholCon
sump=0, Sex = 1))
```

```
##        1 
## 14.58366
```

```
# without
predict(logistic_model, data.frame(HighBP=1, HighChol=0, Smoker=1, Age=60, HvyAlcoholCon
sump=0, Sex = 1))
```

```
##        1 
## 13.85685
```

These numbers indicate that people without high cholesterol have a significantly lower probability of having heart disease, which makes sense, as high cholesterol is highly correlated with it.

Let’s compare the differences between males and females.

```
# female
predict(logistic_model, data.frame(HighBP=1, HighChol=1, Smoker=0, Age=25, Sex=0, HvyAlc
oholConsump = 0))
```

```
##        1 
## 4.227735
```

```
# male
predict(logistic_model, data.frame(HighBP=1, HighChol=1, Smoker=0, Age=25, Sex=1, HvyAlc
oholConsump = 0))
```

```
##        1 
## 4.874757
```

By comparing these two results, we can see that males have a higher probability of having heart disease.

Now let's try using explanatory variables that should ideally have the lowest risk of heart disease vs ones with the highest

```
# least risk
predict(logistic_model, data.frame(HighBP=0, HighChol=0, Smoker=0, Age=1, Sex=0, HvyAlcoholConsump = 0))
```

```
##         1
## -3.641183
```

```
# highest risk
predict(logistic_model, data.frame(HighBP=1, HighChol=1, Smoker=1, Age=30, Sex=1, HvyAlcoholConsump = 1))
```

```
##        1
## 6.065768
```

As expected, when a person has high blood pressure, high cholesterol, smokes, is older in age, and is male, there is a high probability that they will have heart disease compared to a person that doesn't have high blood pressure, high cholesterol, doesn't smoke, is young, and is female.

Assessing Model Performance

Confidence Intervals

Confidence intervals are used to measure the accuracy of the coefficients of the predictor variables in the model. Here, we specified that the confidence level should be 99%.

```
confint(logistic_model, level = 0.99)
```

```
## Waiting for profiling to be done...
```

```
##                        0.5 %     99.5 %
## (Intercept)       -4.0268160 -3.7825050
## HighBP             0.7766682  0.8928582
## HighChol           0.6701504  0.7835040
## Age                0.2516133  0.2740817
## Sex                0.5919807  0.7021571
## Smoker             0.4554269  0.5659004
## HvyAlcoholConsump -0.7651537 -0.5029751
```

The output shows that we are 99% confident that the coefficients for high blood pressure in the model are between 0.77 and 0.89, and so on for the remaining variables. Another thing to note is that the variable for high blood pressure also has the widest confidence interval out of all variables, and age has the narrowest.

Confusion Matrix

A confusion matrix is another way of assessing the performance of a logistic regression. It provides the number of true positives, true negatives, false positives, and false negatives. Let's give it a try.

```
#split data set into training and testing set
set.seed(1)

sample <- sample(c(TRUE, FALSE), nrow(filtered_data), replace=TRUE, prob=c(0.7,0.3))

train <- filtered_data[sample, ]
test <- filtered_data[!sample, ]
```

```
# create a general model from the trained data
g_model <- glm(HeartDiseaseorAttack~HighBP+HighChol+Age+Sex+Smoker, family="binomial", d
ata=train)
```

```
# Make predictions
predictions <- predict(g_model, newdata = test, type = "response")
```

```
# Convert the true labels to a factor with the same levels
predictions <- factor(ifelse(predictions > 0.5, 1, 0), levels = c(0, 1))
test$HeartDiseaseorAttack <- factor(test$HeartDiseaseorAttack, levels = c(0, 1))

# create the confusion matrix
confusionMatrix(test$HeartDiseaseorAttack, predictions)
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 4663 2403
##          1 1634 5432
## 
##                Accuracy : 0.7143
##                  95% CI : (0.7068, 0.7218)
##     No Information Rate : 0.5544
##     P-Value [Acc > NIR] : < 2.2e-16
## 
##                   Kappa : 0.4287
## 
##  Mcnemar's Test P-Value : < 2.2e-16
## 
##             Sensitivity : 0.7405
##             Specificity : 0.6933
##          Pos Pred Value : 0.6599
##          Neg Pred Value : 0.7688
##              Prevalence : 0.4456
##          Detection Rate : 0.3300
##    Detection Prevalence : 0.5000
##       Balanced Accuracy : 0.7169
## 
##        'Positive' Class : 0
## 
```

The results of the confusion matrix show that the model correctly predicted 5,424 instances of heart disease, and 4,602 instances of no heart disease. However, it also incorrectly predicted 2,447 instances of heart disease, and 1,655 of no heart disease. Furthermore, the accuracy was found to be 0.71, that is, the proportion of correctly classified instances out of all. This indicates that we have a decent classifier, that doesn’t perform poorly, but also not as best as it could be.

ROC Curve

The Receiver Operating Characteristic plot is used to visualize the trade-off between sensitivity and specificity in our model. A perfect model would display an ROC curve that is very close to the top left corner, indicating a high true positive rate, and a low false positive rate.

Let’s start by creating some predictions for the model.

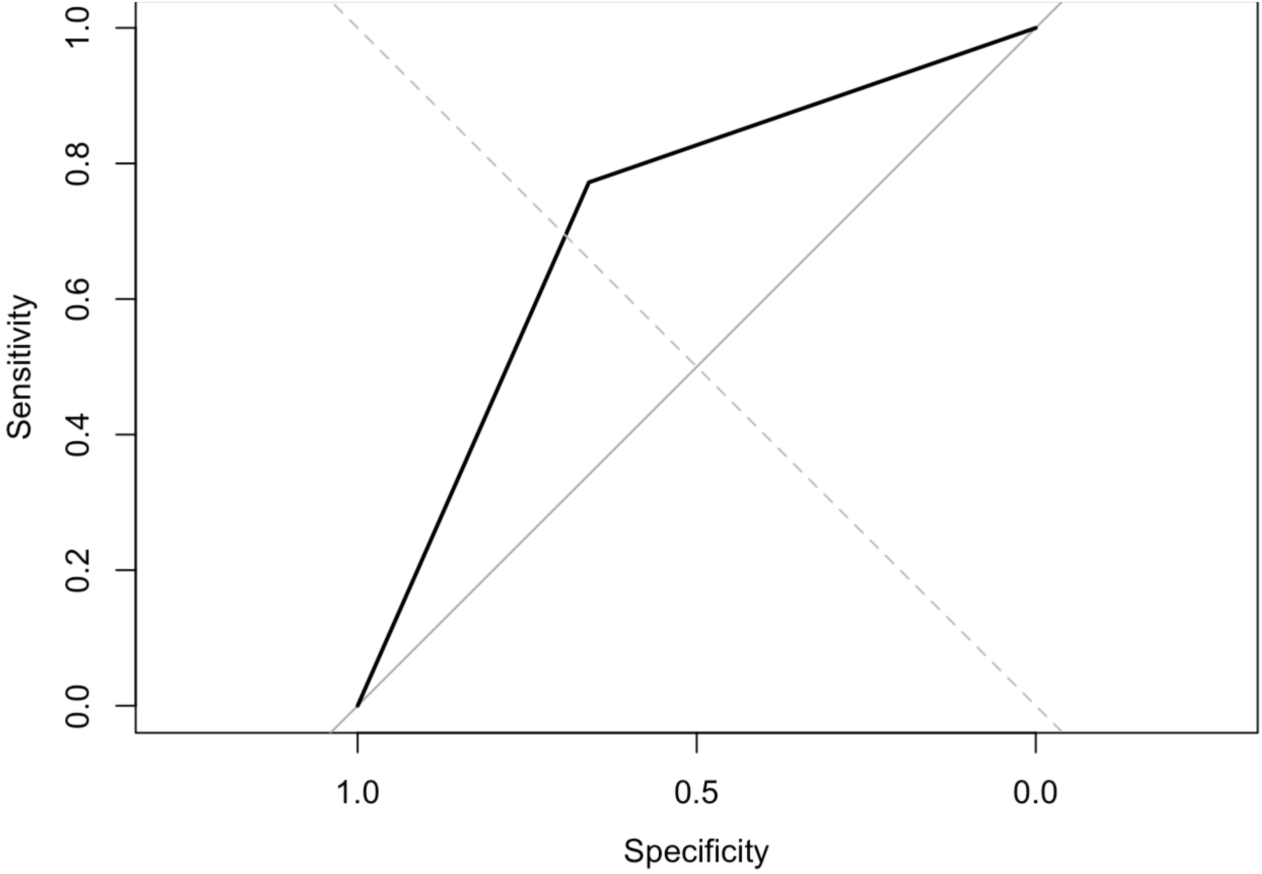
```r
# create predictions
predictions <- predict(logistic_model, newdata = test, type = "response")

# get predictions and response variable to be same size
predictions <- factor(ifelse(predictions > 0.5, 1, 0), levels = c(0, 1))

test$HeartDiseaseorAttack <- factor(test$HeartDiseaseorAttack, levels = c(0, 1))

# turn predictions into a numeric type
predictions <- as.numeric(predictions)
```

```r
# create roc curve
roc_curve <- roc(test$HeartDiseaseorAttack, predictions)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
# Plot the ROC curve
plot(roc_curve, main = "ROC Curve for the Logistic Regression Model")

# Add a reference line for a random classifier
abline(0, 1, lty = 2, col = "gray")
```



In comparison to the diagonal line, which represents a model with no discrimination power, this ROC curve is much closer to the top left, indicating some amount of discrimination power, but perhaps not the best.

The AUC, or “area under the curve” is a value that can be used to summarize the performance of the model.

```
# calculating the AUC
auc(roc_curve)
```

```
## Area under the curve: 0.7155
```

We can see here that the area under the curve is 0.71. A perfect classifier would have an AUC of 1.0. This value, along with the ROC plot, further confirm that our classifier is decent, but not perfect.

Leave One Out Cross-Validation

LOOCV is a form of cross-validation where the model is trained on all data except one point, which them model then attempts to accurately predict. This process is repeated for every point in the data set.

```
# create a model with no predictors
nopred_logistic_model <- glm(HeartDiseaseorAttack~1, family="binomial", data=filtered_da
ta)
```

```
set.seed(5)

# model with predictors
cv.glm(filtered_data, logistic_model, K = 5)$delta[1]
```

```
## [1] 0.1887334
```

```
# model with no predictors
cv.glm(filtered_data, nopred_logistic_model, K = 5)$delta[1]
```

```
## [1] 0.2500025
```

The results for the LOOCV show that the model with multiple predictors had a prediction error of 0.19, significantly better than the model with no predictors, which had a prediction error of 0.25.

Root Mean Squared Residual

The RMSR is a way of measuring the goodness of fit of a model. The residuals from the model are extracted and squared to prevent positive and negative values from canceling out. Then, the mean of the squared residuals is calculated and the square root of this value is taken. This is a measure of the average size of the residuals. The lower the RMSR, the closer the predicted probabilities are to the actual values.

```
# general model
sqrt(mean(resid(logistic_model) ^ 2))
```

```
## [1] 1.055858
```

```
# general model with no predictors
sqrt(mean(resid(nopred_logistic_model) ^ 2))
```

```
## [1] 1.177408
```

```
# male model
sqrt(mean(resid(logistic_model1) ^ 2))
```

```
## [1] 1.06276
```

```
# female model
sqrt(mean(resid(logistic_model2) ^ 2))
```

```
## [1] 1.043756
```

These results show that out of the three models, the RMSR value is the least for logistic_model2. Therefore, this model predicts better than the other two.

Conclusion

Heart disease is an illness that can be predicted by many different risk factors. In this project, we began by cleaning the data to suit our task. Next, we plotted the relationships of several variables of interest. Briefly, our plots showed that in general, people with heart disease tended to have high cholesterol, high blood pressure, smoked, were older, and were male.

In order to find the predictors of Heart disease, hypothesis testing was done. For this a filtered data frame was defined…

For predicting Heart Disease or Attack for people with specific values for different health factors, a generalized linear binomial family model, `logistic_model` is defined using HighBP, HighChol, Age, Sex, Smoker, and HvyAlcoholConsump as explanatory variables. The null deviance, AIC, and residual null deviance values were much higher. The diagnostic plots provide valuable insights into how various predictors interact to influence the probability of heart disease. These observations underscore the importance of a detailed understanding of multiple risk factors to enhance the accuracy of predictive models and guide targeted interventions for cardiovascular health. The plots indicate that the probability of heart disease tends to rise with higher levels of high blood pressure, elevated cholesterol, and advancing age. Additionally, the habit of smoking is associated with an increased probability of heart disease.

In order to build a better fit model, the sexes were defined separately as data_male and data_female data frames. A generalized binomial model(logistic_model1 and logistic_model2 for data_male and data_female respectively) is defined for both data frames using HighBP, HighChol, Age, Smoker, and HvyAlcoholConsump as predictors. The AIC, null deviance, and residual null deviance scores were much lower for both models compared to the logistic_model defined without separating the sexes. Moreover, step-wise selection function in all 'forward', 'backward', and 'both' directions were used to find the best fit model and logistic_model, logistic_model1, and logistic_model2 turned out to the best fit models using the selected features as predictors.

When we attempted to use the model to predict the probability of heart disease based on different factors, the results were consistent with our previous knowledge. That is, people without high blood pressure, high cholesterol, young of age, non-smokers, that were female were about five times less likely to have heart disease in comparison to their counterparts.

The performance of our model was assessed using confidence intervals, a confusion matrix, and ROC curve, LOOCV, and the RMSR. The confidence intervals were determined in order to help reassure us of the coefficients obtained from the model. It was determined that we were 99% confident that the coefficients we obtained were within the resulting interval. Next, we used a confusion matrix to find the specificity and sensitivity of our model. The results showed that our classifier had am accuracy of 0.71 which is decently close to 1.0, but not amazing. To further test the specificity and sensitivity, we plotted an ROC curve that further confirmed our results - our classifier had decent predictive power. This was further confirmed by the AUC. LOOCV was also another method we used. We used a 5-fold cross validation on our main logistic model with predictors and one without. The results showed that the model with predictors was better than the one without, as it had a lower prediction error. The lowest RMSR is the model subsetted to only include female, indicating that it performs the best out of all.

As evident from the predictions and the model assessments, it can be concluded that the logistic_model is not the best, but did a great job in identifying some of the relevant features that can be used to predict the risk of Heart disease or attack in a person. More analysis could be done to identify the best predicting model.

From this analysis, we were able to confirm the heart disease predictors identified by the CDC. High blood pressure and high cholesterol are highly correlated to heart disease, and age, smoking, and sex have a significant, albeit weaker correlation to it as well. One surprising thing to note is that, although the CDC determined heavy alcohol consumption to be a possible explanatory variable for heart disease, our studies showed that they were in fact slightly negatively correlated. Our study only looked at a handful of factors suspected to be related to heart disease. This data set has the potential to allow for much more deep statistical testing to extract possible causes for heart disease and provide further insight on how to detect it and ultimately provide treatment for patients.