Know your risk for Heart Dieases

Julia Capelli

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

##Introduction
I used the Heart Disease Data Set from UCI Machine Learning repository. This data set has 303 instances and 76 attributes. Additionally there are 14 variables which are age, sex, cp:chest pain, trestbps: resting blood pressure, chol:cholesterol levels, fbs: fasting blood sugar, restecg: resting electrocardiography results, thalach: maximum heart rate achieved, exang:excerise indcued angina, old peak:ST depression induced by exercise relative to rest, slope:The slope of the peak exercise ST segment, ca:number of major vessels, thal: 3 = normal; 6 = fixed defect; 7 = reversable defect, num: diagnosis of heart disease.

This is a multivariate data collection, that abstracted its data from 4 different sources. The four sources are: 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

In total, this data set contained 76 attributes and all of the of names and social security values in this data set were removed and replaced with a dummy values. This data set had to be tidy since there were a few NA variables in the data set. The changes done are clearly explained in the Data set portion of this report. This data set was used to see what factors lead to heart disease in patients. Heart Disease is a major reason for death in the world. Since it is such as serious disease I wanted to use this data in order to predict what health factors can lead to heart disease in order to hopefully reduce the number of Heart Disease Deaths in the world. After analyzing the data I predict to find that cholesterol resting blood pressure, age, and maxiumum heart rate will all have a significant relationship with heart disease. Additionally, I predict that sex will not be a factor in predicting heart disease.

##A.DataSet

```
url<- "http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.
cleveland.data"

#read the dataset from URL
data <- read.csv(url, header =FALSE)

#Name the columns
colnames(data) <- c("age", "sex", "chest.pain", "trestbps", "chol", "fb", "restecg", "th
alach", "exang", "oldpeak", "slope", "ca", "thal", "hd")
#Change ? to NA
data[data == "?"] <- NA

#View Data Set
data[data$sex == 0,]$sex <- "F"
data[data$sex == 1,]$sex <- "M"

#Convert Columns to factors
data$sex <- as.factor(data$sex)
data$chest.pain <- as.factor(data$chest.pain)
data$fb <- as.factor(data$fb)
data$restecg <- as.factor(data$restecg)
data$exang <- as.factor(data$exang)
data$slope <- as.factor(data$slope)
data$hd <- as.factor(data$hd)

#Tell R it is a column of integers
data$ca <- as.integer(data$ca)
#Convert to a factor
data$ca <- as.factor(data$ca)

data$thal <- as.integer(data$thal)
data$thal <- as.factor(data$thal)

#Make binary variable
data$hd <- ifelse(test=data$hd == 0, yes="Healthy", no="Unhealthy")
data$hd <- as.factor(data$hd)

#Remove the NA
data <- data[!(is.na(data$ca)| is.na(data$thal)),]

#Change Name of data set
Heart <- data
```

In order to clean up the data the first thing I did was change all of the 0's to NA. Then I changed all the of the 0 and 1 in the sex column to M and F so that it is easier to read. Then I converted all of the columns into factors. I also converted the ca and thal column to a columns of integers and then to a column of factors. In order to use the data in various statistical tests all of the NA values were removed. After removing all of the samples with NA values there were 297 samples remaining.

##B.EDA

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("ggplot2")
library("tidyverse")
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
```

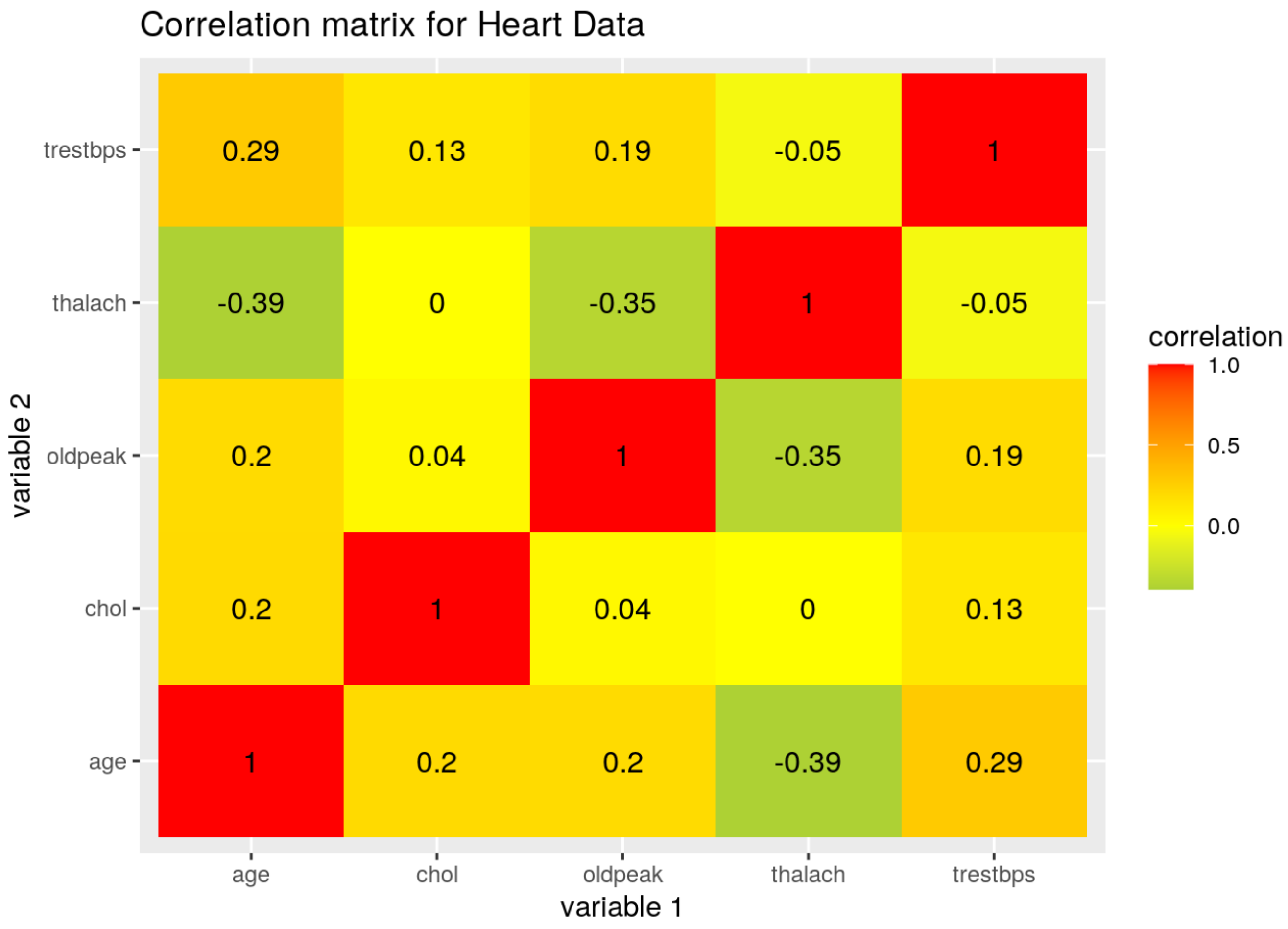
```
## ✓ tibble  3.0.4     ✓ purrr   0.3.4
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.0
```

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

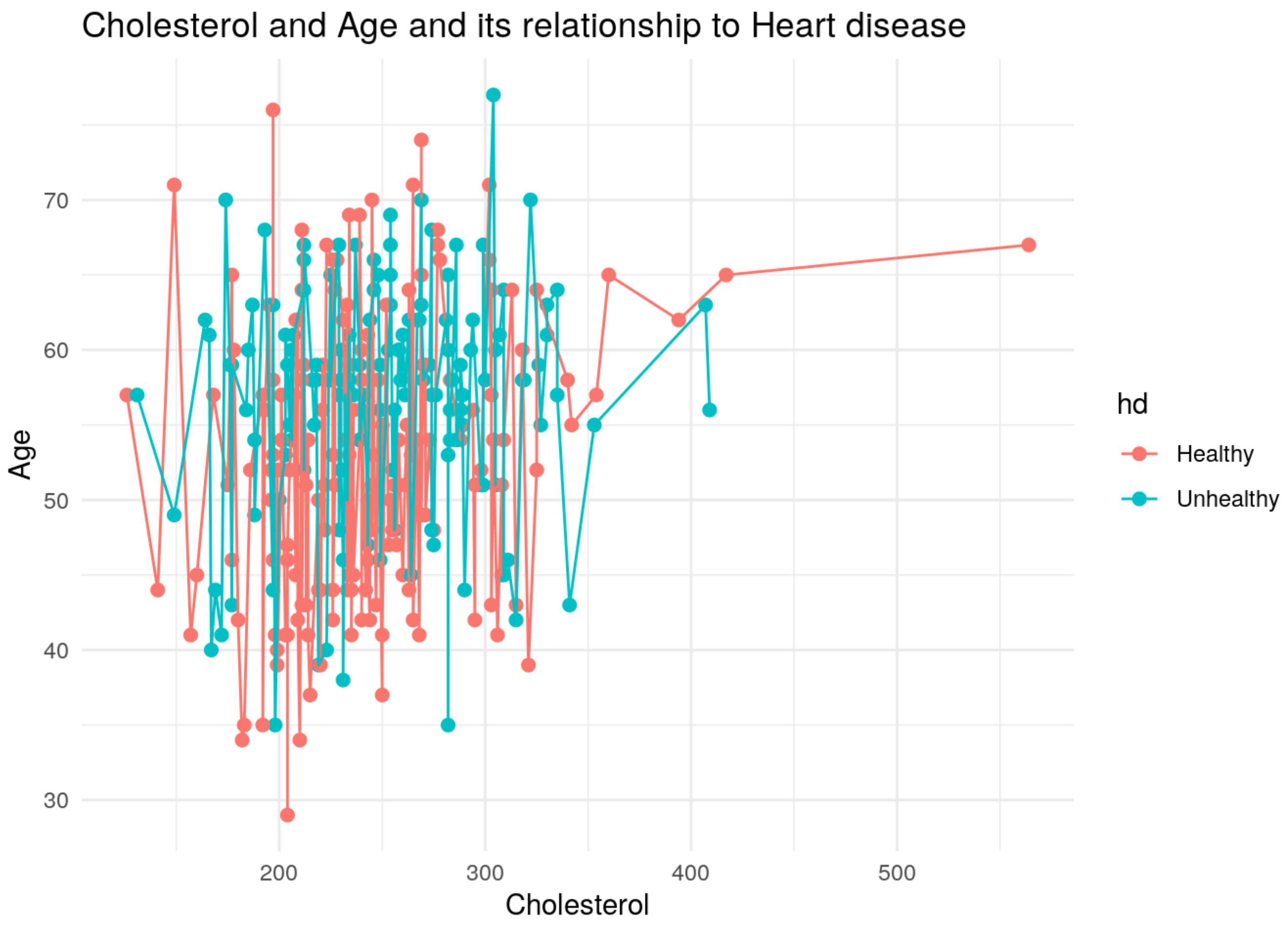
```
#Build a correlation matrix between all numeric variables
Heart_NUM <- Heart %>%
  select_if(is.numeric)

cor(Heart_NUM, use="pairwise.complete.obs")
```

```
##                 age    trestbps          chol       thalach     oldpeak
## age       1.0000000  0.29047626  2.026435e-01 -3.945629e-01  0.19712262
## trestbps  0.2904763  1.00000000  1.315357e-01 -4.910766e-02  0.19124314
## chol      0.2026435  0.13153571  1.000000e+00 -7.456799e-05  0.03859579
## thalach  -0.3945629 -0.04910766 -7.456799e-05  1.000000e+00 -0.34763997
## oldpeak   0.1971226  0.19124314  3.859579e-02 -3.476400e-01  1.00000000
```

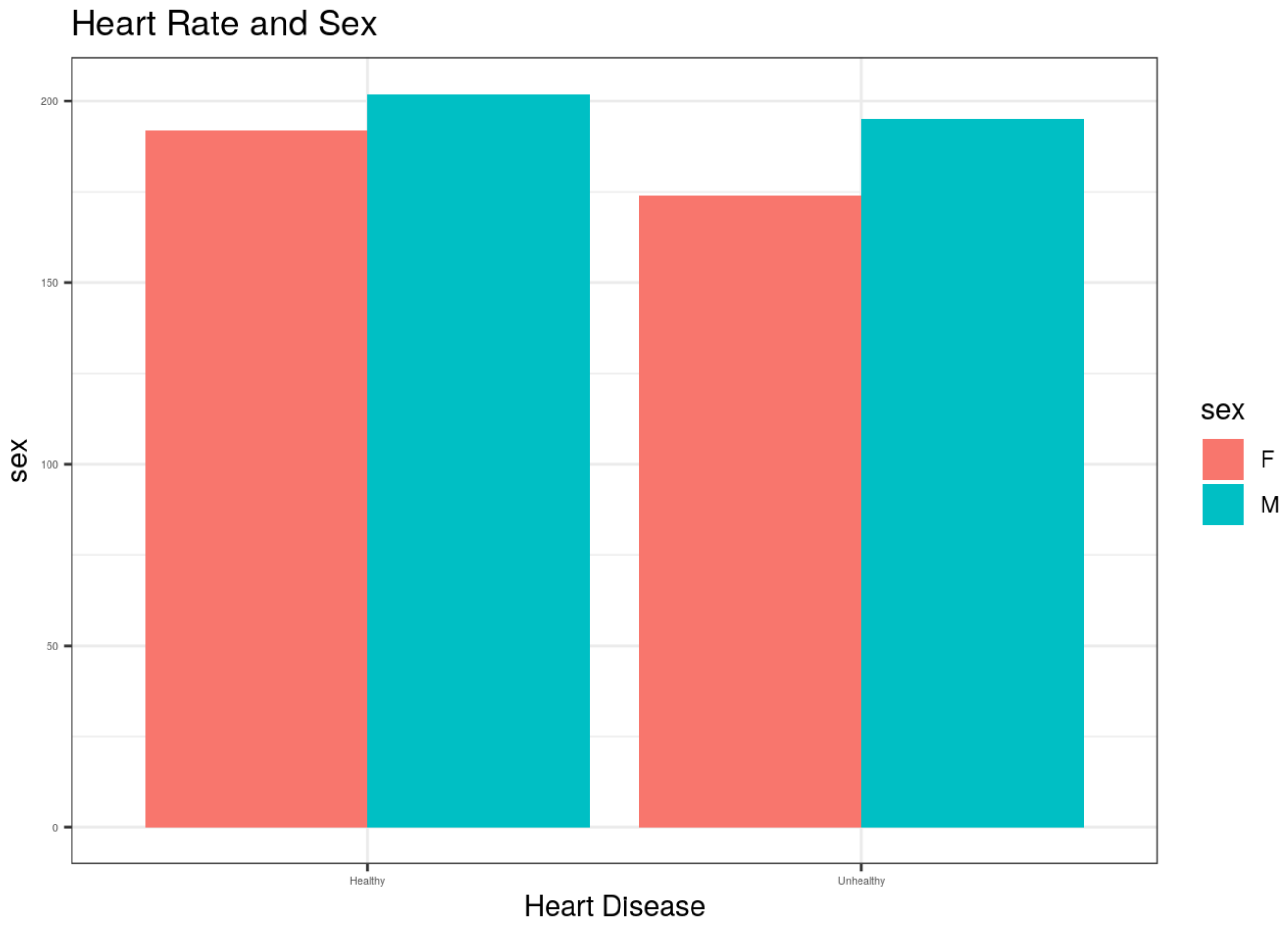
```
#We need to shape the correlation matrix as a date frame then we use manipulation to org
anize our Output using functions. use pivot longer to have all correlation as the same v
ariables. then have rows to column to represent first variable to second variable. Then
 use ggplot for color, this is using the heatmap.

cor(Heart_NUM, use="pairwise.complete.obs") %>%
  #Save as a data frame
  as.data.frame %>%
  #Convert to row naes to an explicity variable
  rownames_to_column %>%
  #Pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  #Heatmap with geom_tile
  geom_tile() +
  #Change the sclae to make the middle appear neutral. Neutral or zero is white, negativ
e is red and postive is blue
  scale_fill_gradient2(low="springgreen4", mid="yellow", high="red")+
  #Overlay Values
  geom_text(aes(label=round(correlation,2)), color = "black", size = 4) +
  #Give title and labels
  labs(title = "Correlation matrix for Heart Data", x= "variable 1", y= "variable 2")
```



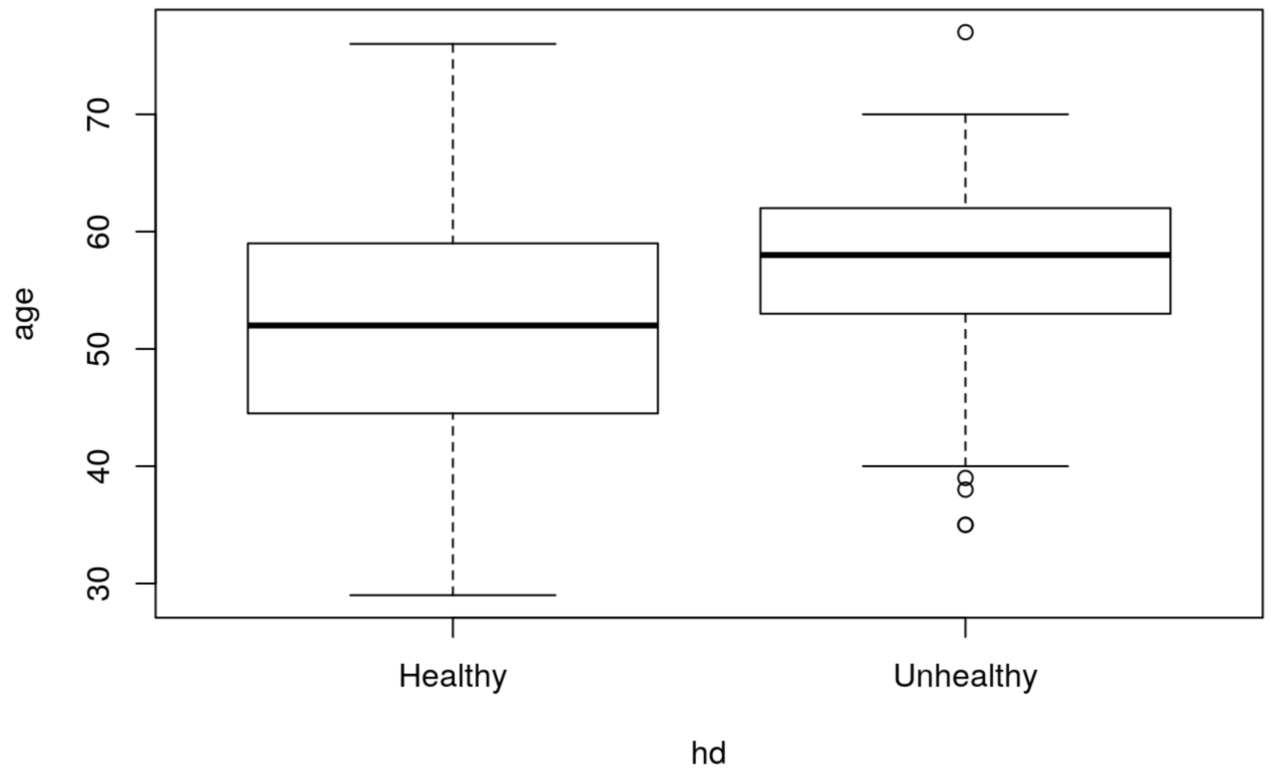
```
# Use the default theme_minimal()
ggplot(data = Heart, aes(x = `chol`, y = `age`, color = hd)) +
  geom_point(size=2) + geom_line() +
theme_minimal()+
ggtitle("Cholesterol and Age and its relationship to Heart disease")+
  ylab("Age")+
  xlab("Cholesterol")
```



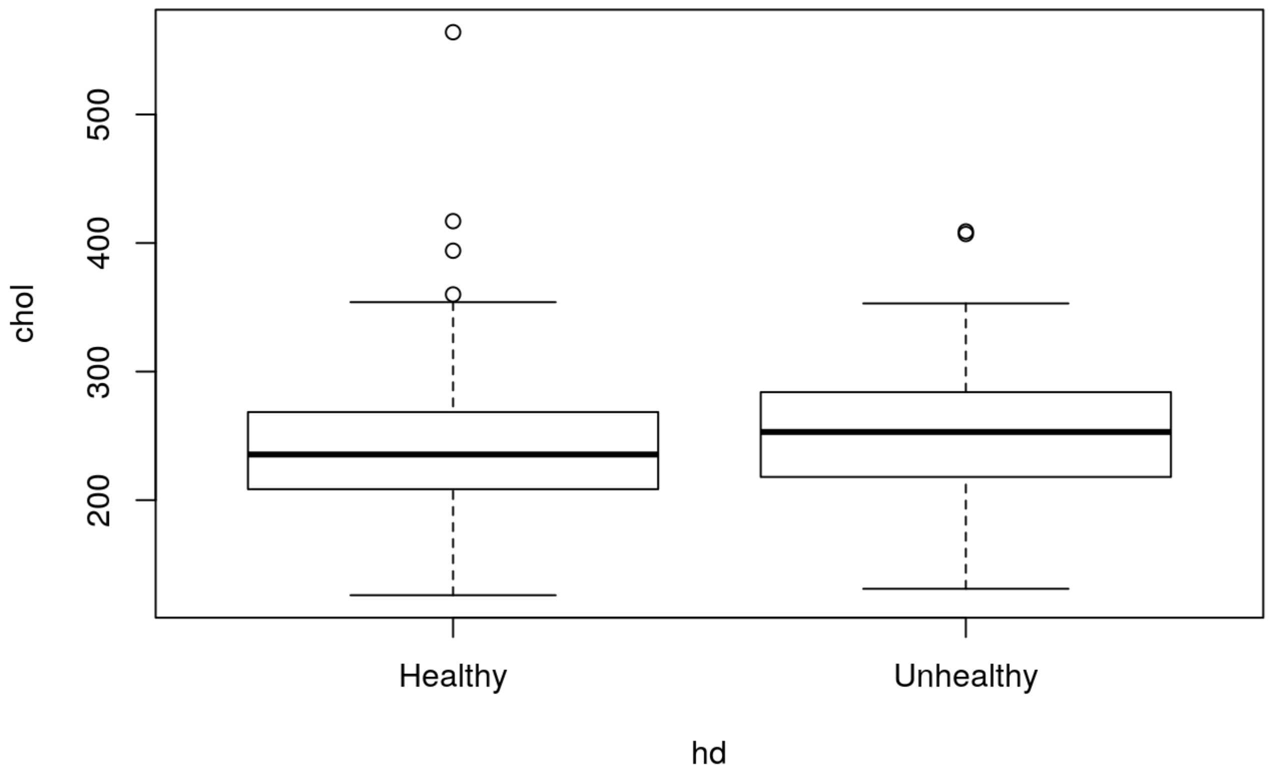
```
#Create Bar graph to see if there are differences among sex and Hearrt disease
Heart %>%
slice(1:300)%>%
ggplot(aes(hd,`thalach`, fill = sex)) +
geom_bar(position = "dodge", stat = "identity") +
theme_bw()+
theme(axis.text = element_text(size = 4))+
ggtitle("Heart Rate and Sex")+
xlab("Heart Disease")+
ylab("sex")
```



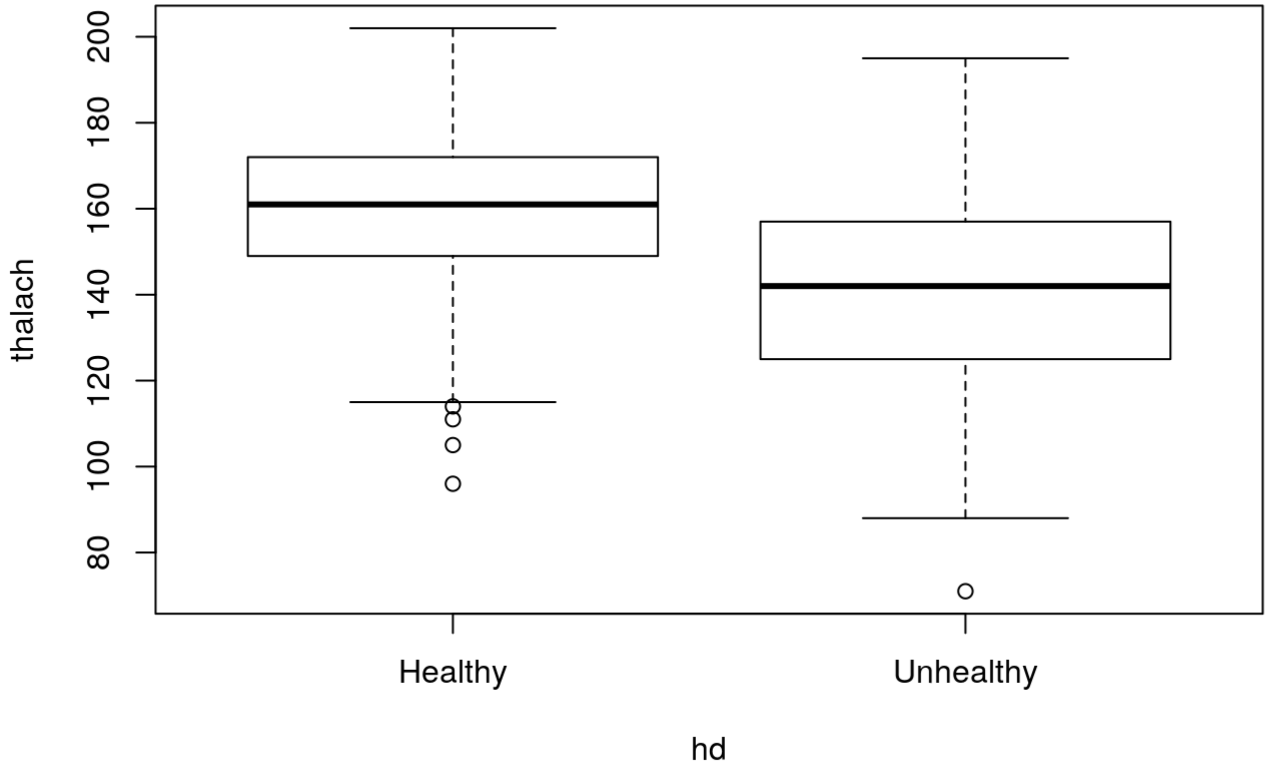
```
#Create Box plots in order to investigate the relationship between different variables and there relationship with Heart Disease.
boxplot(age ~ hd, data = Heart)
```



```
boxplot(chol ~ hd, data = Heart)
```



```
boxplot(thalach ~ hd, data = Heart)
```



According to the heat map, most of the variables have a low correlation. This indicates that most of the variables in the data set have no or little relationship to one another. Its important to note that Heart Condition is represented as a binary variable since it is a character and not a factor. Sex also was not calculated since it was not a numerical variable. In this heat map we are not looking for a correlation between heart disease and the variables but instead the relationship between all the other variables to each other. The scatter plot graph indicates that both cholesterol and ages have little relationship to heart disease since the there is a lot of variation. Therefore we can conclude that age is a not a good predictor of Heart disease. One reason might be that there is not a wide range of ages in this data set since the median was 56. Lastly, the bar graph shows that there is a relationship between sex and heart disease since many more males have an unhealthy heart when compared to males. The Box plots also show that the variables of thalch and sex are significantly associated with heart disease.

##C.Manova

```
# Create a binary variable coded as 0 and 1
Heart <- data %>%
  mutate(y = ifelse(hd == "Healthy",1, 0))

# Perform MANOVA with 2 response variables listed in cbind()
manova_Heart <- manova(cbind(`chol`,`sex`,`thalach`) ~ hd, data = Heart)

# OUtput of MANOVA
summary(manova_Heart)
```

```
##             Df  Pillai approx F num Df den Df    Pr(>F)    
## hd           1 0.26151   34.585      3    293 < 2.2e-16 ***
## Residuals 295                                              
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#If MANOVA is significant then we can perform one-way ANOVA for sex variable
summary(aov(y ~ sex, data = Heart))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## sex           1   5.72   5.723    24.8 1.09e-06 ***
## Residuals   295  68.08   0.231                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(y ~ thalach, data = Heart))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## thalach       1  13.26  13.257   64.59 2.24e-14 ***
## Residuals   295  60.55   0.205                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Since Anova is significant then we can perform post-hoc analysis for thickness
pairwise.t.test(Heart$y,Heart$sex, p.adj="none")
```

```
## 
##  Pairwise comparisons using t tests with pooled SD 
## 
## data:  Heart$y and Heart$sex 
## 
##   F      
## M 1.1e-06
## 
## P value adjustment method: none
```

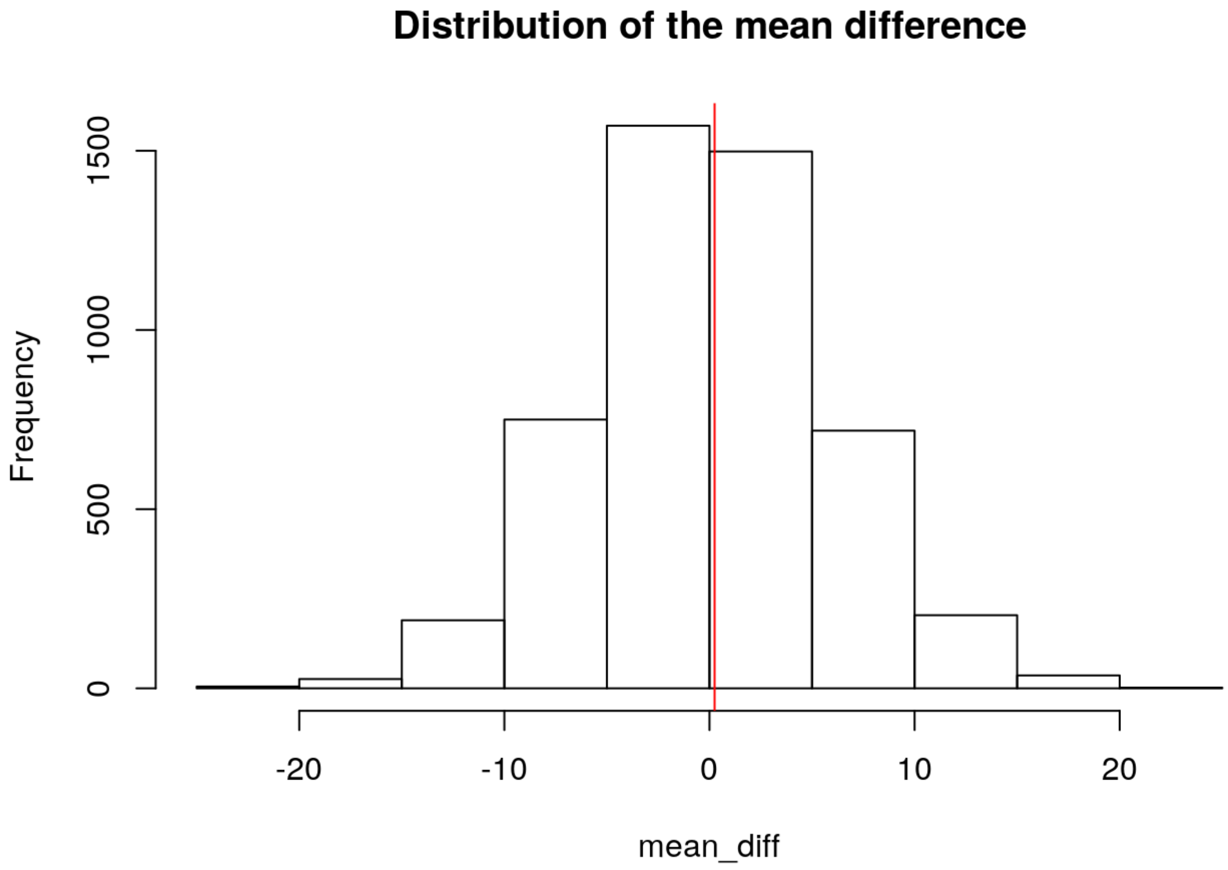
Based on the MANOVA Test we can conclude that cholesterol is not a good predictor of heart disease since its p-value is greater then 0.05. However, we can conclude that sex is has a significant value since its p-value is much greater then 0.05. An ANOVA test was computed since a Manova test resulted in a statistically significant value. After computing the Anova test for heart disease based on sex, we can conclude that sex is a good predictor of heart disease since the p-value is less then then 0.05 and therefore is statistically significant. After performing an anova test on the variable thalach:maximum heart rate achieved, we can conclude that it is statistically significant since its p-value is < 0.05. Therefore maximum heart rate is a good predictor of heart disease

##D.Randomization Test
This was used in order to increase the power in order to detect heart disease among patients with different cholesterol levels.

Null Hypothesis: There is not a significant relationship between heart disease and cholesterol.

Alternate Hypothesis: There is a significant relationship between heart disease and cholesterol levels.

```
#create an empty vector to store the difference
set.seed(348)
mean_diff <- vector ()


#Create many randomization with the loop function
for (i in 1:5000){
  chloes <- Heart %>%
    mutate(chol=sample(`chol`),.groups = 'drop')

  mean_diff[i] <- chloes %>%
    group_by(`hd`) %>%
    summarize(means = mean(`chol`),.groups = 'drop') %>%
    summarize(mean_diff = diff(means), .groups ='drop') %>%
    pull
}
#Represent the distribution of the mean differences
{hist(mean_diff, main="Distribution of the mean difference"); abline(v=0.25, col="red")}
```



```
#Calculate the corresponding p-value
mean(mean_diff > -.25| mean_diff < .25)
```

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

```
# Compare to a Welch's t-test
t.test(data = Heart, chol ~ hd, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  chol by hd
## t = -1.3834, df = 295, p-value = 0.1676
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -20.253644   3.533115
## sample estimates:
##   mean in group Healthy mean in group Unhealthy
##                243.4938                251.8540
```

After computing A Welch’s T-test the P-value was much larger then 0.05 which indicates that the results are not statistically significant. This means that we fail to reject our null hypothesis that cholesterol levels are a predictor of heart disease. Therefore I can conclude that the randomization tests do not conclude that there are no differences in heart disease between patients with high and low cholesterol.

##E.Linear Regression

```
# Include an interaction term in the regression model
levelss <- lm(chol ~ trestbps * hd, data = Heart)
summary(levelss)
```

```
##
## Call:
## lm(formula = chol ~ trestbps * hd, data = Heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124.57  -33.66   -3.67   27.69  325.32
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          199.61279   32.60955   6.121 2.97e-09 ***
## trestbps               0.33970    0.25045   1.356    0.176
## hdUnhealthy            2.03228   45.61771   0.045    0.964
## trestbps:hdUnhealthy   0.03322    0.34320   0.097    0.923
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.71 on 293 degrees of freedom
## Multiple R-squared:  0.02103,    Adjusted R-squared:  0.01101
## F-statistic: 2.098 on 3 and 293 DF,  p-value: 0.1006
```

```
# Center the data around the means (the intercept becomes more informative)
Heart$chol_c <- Heart$chol - mean(Heart$chol)

# Include an interaction term in the regression model with centered predictors
fit_c <- lm(trestbps ~ hd * chol_c, data = Heart)
summary(fit_c)
```

```
##
## Call:
## lm(formula = trestbps ~ hd * chol_c, data = Heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.179 -11.398  -1.050   9.565  63.415
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        129.29654    1.38729  93.201   <2e-16 ***
## hdUnhealthy          5.09549    2.04431   2.493   0.0132 *
## chol_c               0.03152    0.02582   1.221   0.2232
## hdUnhealthy:chol_c   0.02244    0.03974   0.565   0.5728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.5 on 293 degrees of freedom
## Multiple R-squared:  0.03891,    Adjusted R-squared:  0.02907
## F-statistic: 3.954 on 3 and 293 DF,  p-value: 0.008693
```

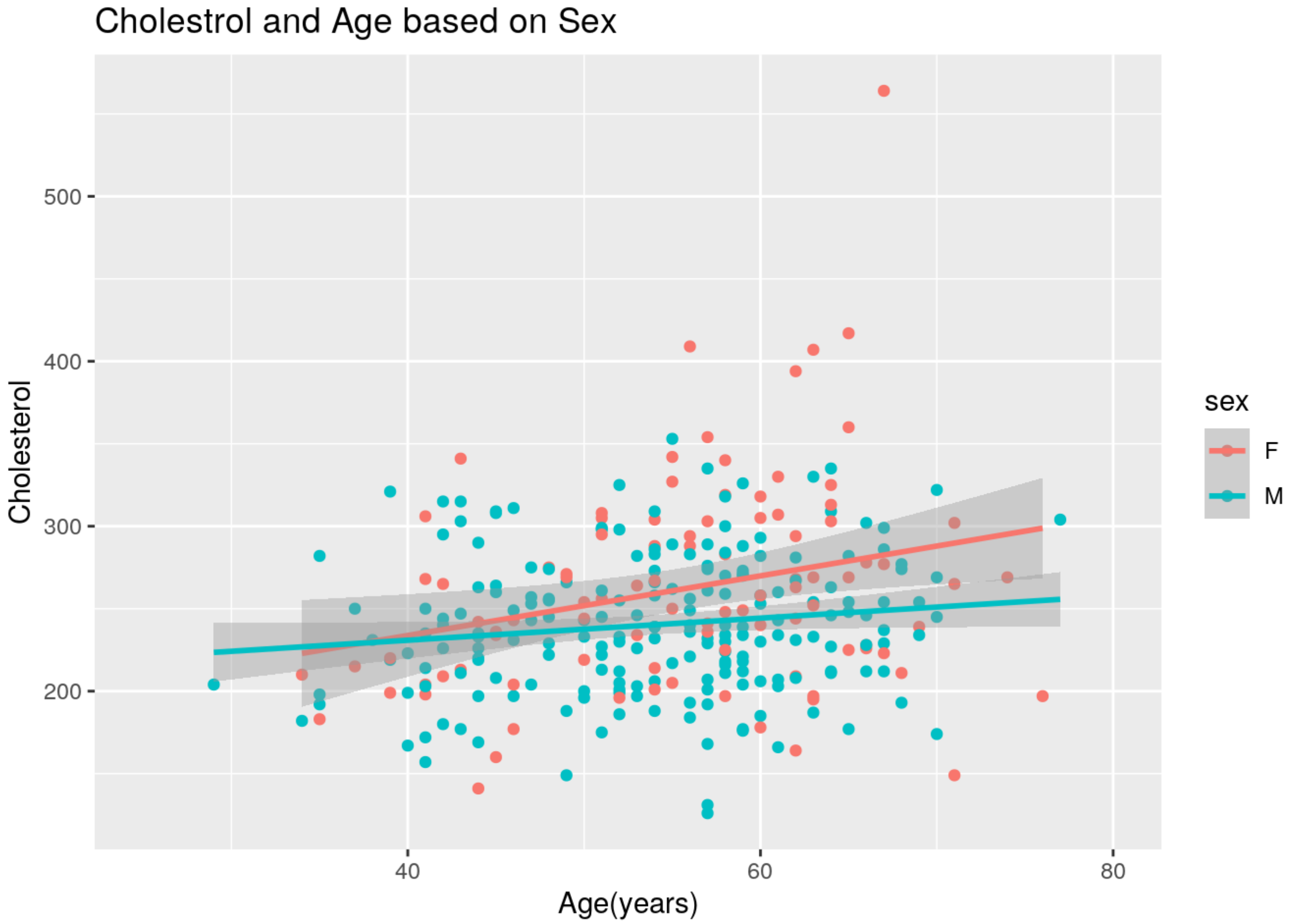
```
fit_e <- lm(chol ~ trestbps + hd, data = Heart)
summary(fit_e)
```

```
##
## Call:
## lm(formula = chol ~ trestbps + hd, data = Heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124.94  -33.48   -3.94   28.08  325.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 197.3272    22.4563   8.787   <2e-16 ***
## trestbps      0.3574     0.1709   2.091   0.0374 *
## hdUnhealthy   6.4089     6.0811   1.054   0.2928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.62 on 294 degrees of freedom
## Multiple R-squared:  0.021,  Adjusted R-squared:  0.01434
## F-statistic: 3.153 on 2 and 294 DF,  p-value: 0.04416
```

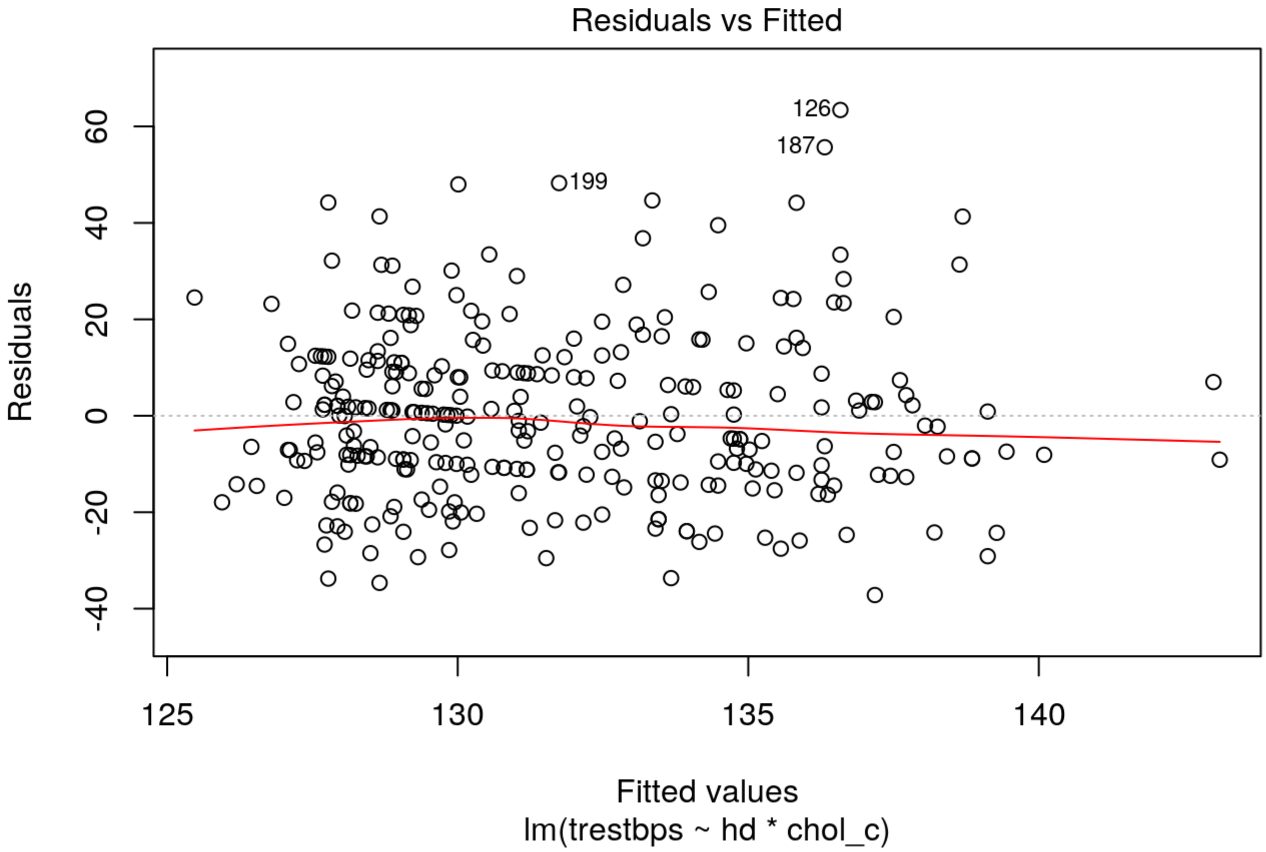
###F. model fit and assumptions

```
#Visualize normal
ggplot(Heart, aes(age, chol, color=sex))+
  #Fitted Value
  geom_point()+
#Regression line with no confidence interval
geom_smooth(method="lm")+
  xlim(25,80)+
  ggtitle("Cholestrol and Age based on Sex")+
xlab("Age(years)")+
  ylab("Cholesterol")
```

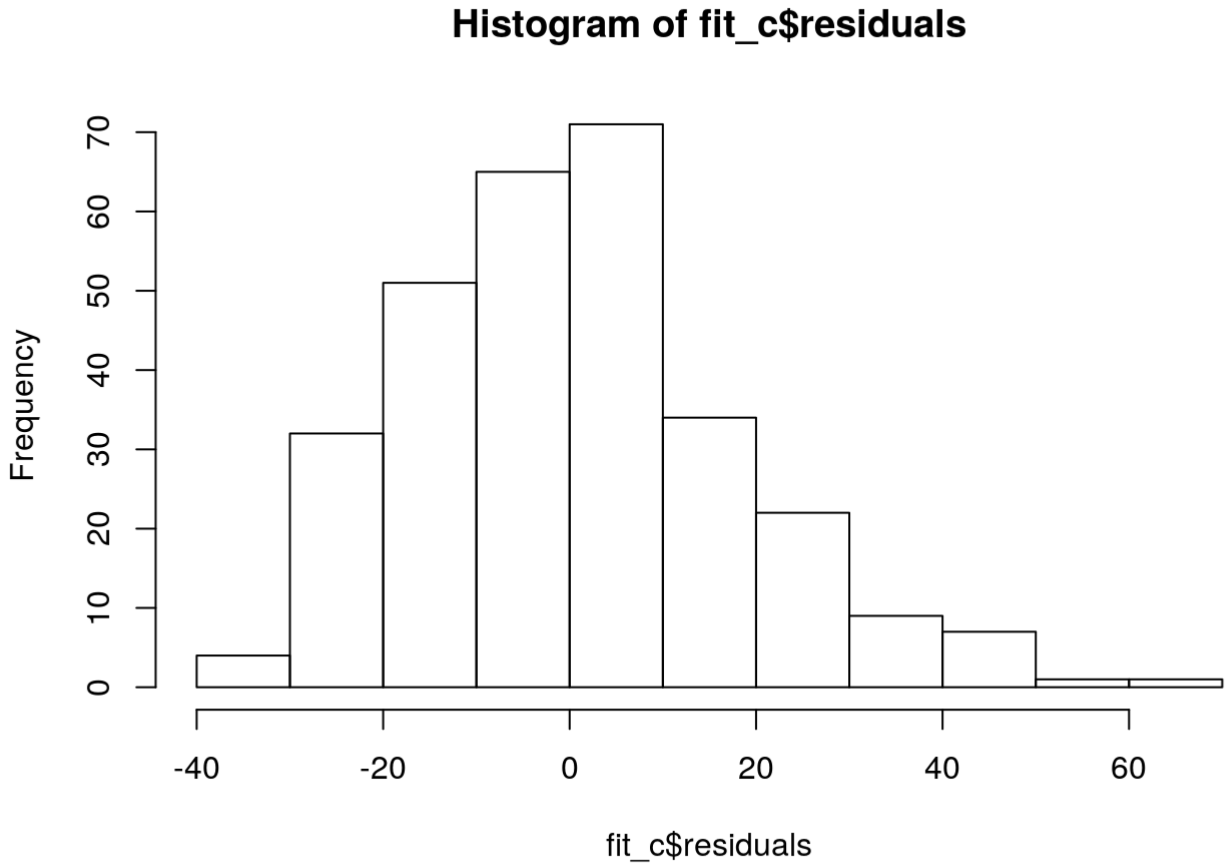
```
## `geom_smooth()` using formula 'y ~ x'
```



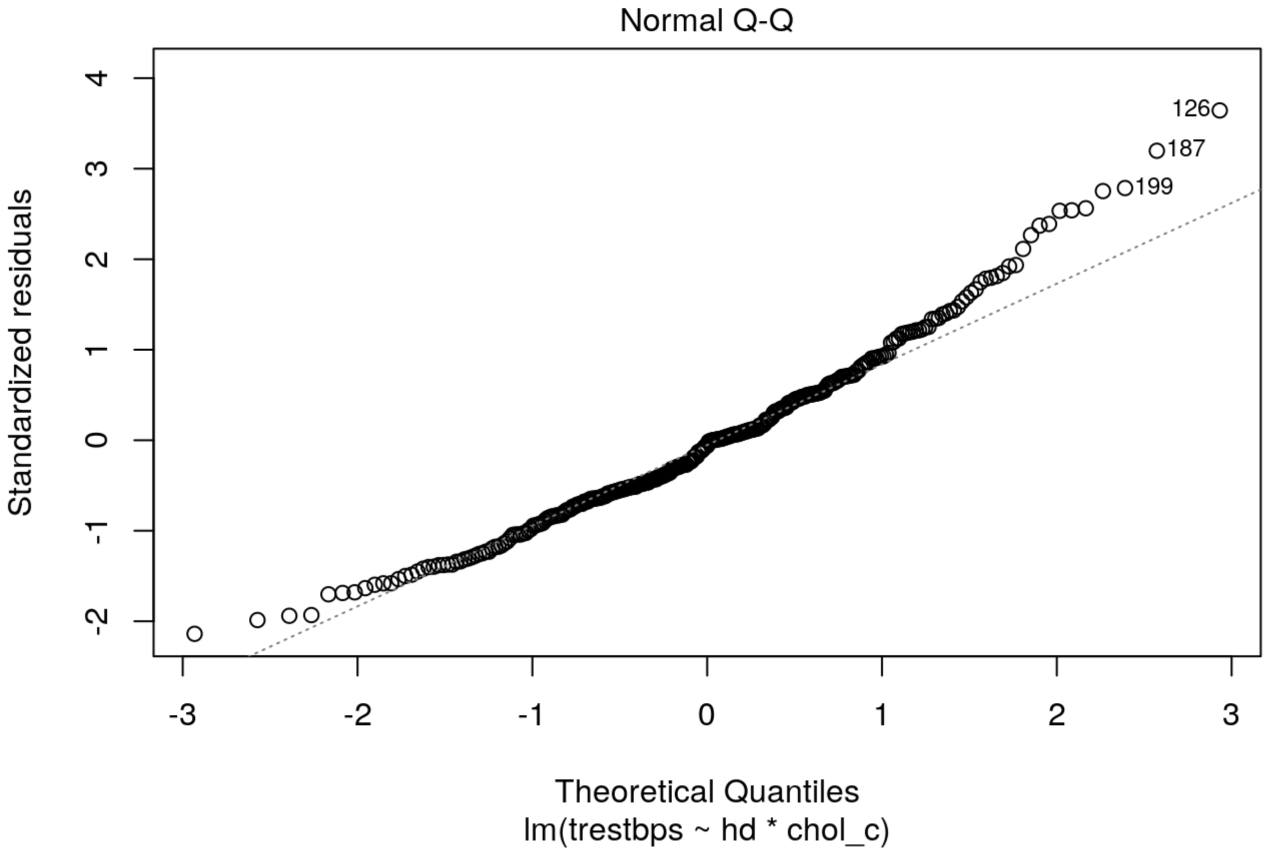
```
#Check for assumptions
plot(fit_c, which =1)
```



```
#Histogram of residuals '
hist(fit_c$residuals)
```



```
#Q-Q plot for the residuals
plot(fit_c, which = 2)
```



```
#Compute Bootstrapping
samp.Heart <- replicate(5000, {
  boot_data <- sample_frac(Heart, replace = TRUE)
  fitboot <- lm(chol ~ age + sex, data=boot_data)
  new <- data.frame(age = 40, sex = 'F')
  predict(fitboot, newdata = new, interval = "prediction")
})
head(samp.Heart)
```

```
## [1] 236.9967 136.9931 337.0003 254.4126 155.6553 353.1699
```

```
quantile(samp.Heart, .025)
```

```
##     2.5%
## 134.0412
```

```
quantile(samp.Heart, .975)
```

```
##    97.5%
## 360.3093
```

```
# Remove variation in thalach shared with Heart Disease
reschol <- lm(thalach ~ hd, data = Heart)$residuals
# Remove variation in trestbps shared with Heart Disease
restrest <- lm(trestbps ~ hd, data = Heart)$residuals

# Regress the residuals BMI on residuals BP (Glucose removed), show the coefficients
coef(lm(reschol ~ restrest))
```

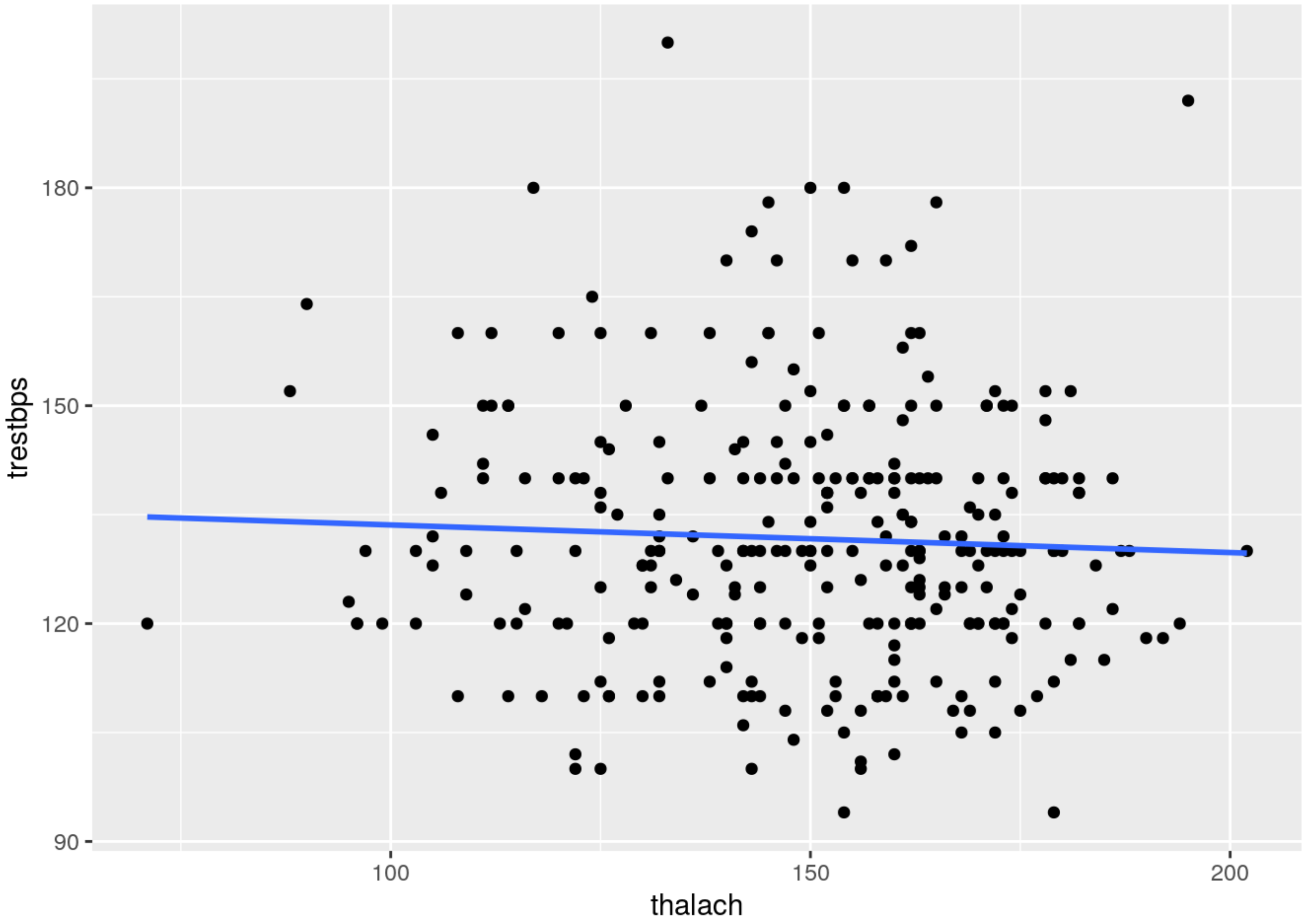
```
##  (Intercept)     restrest
## 1.052485e-15 2.108937e-02
```

```
# Compare to the coefficients obtained in the MLR
coef(lm(chol ~ trestbps + hd, data = Heart))
```

```
## (Intercept)    trestbps hdUnhealthy
## 197.3272417   0.3573951   6.4088745
```

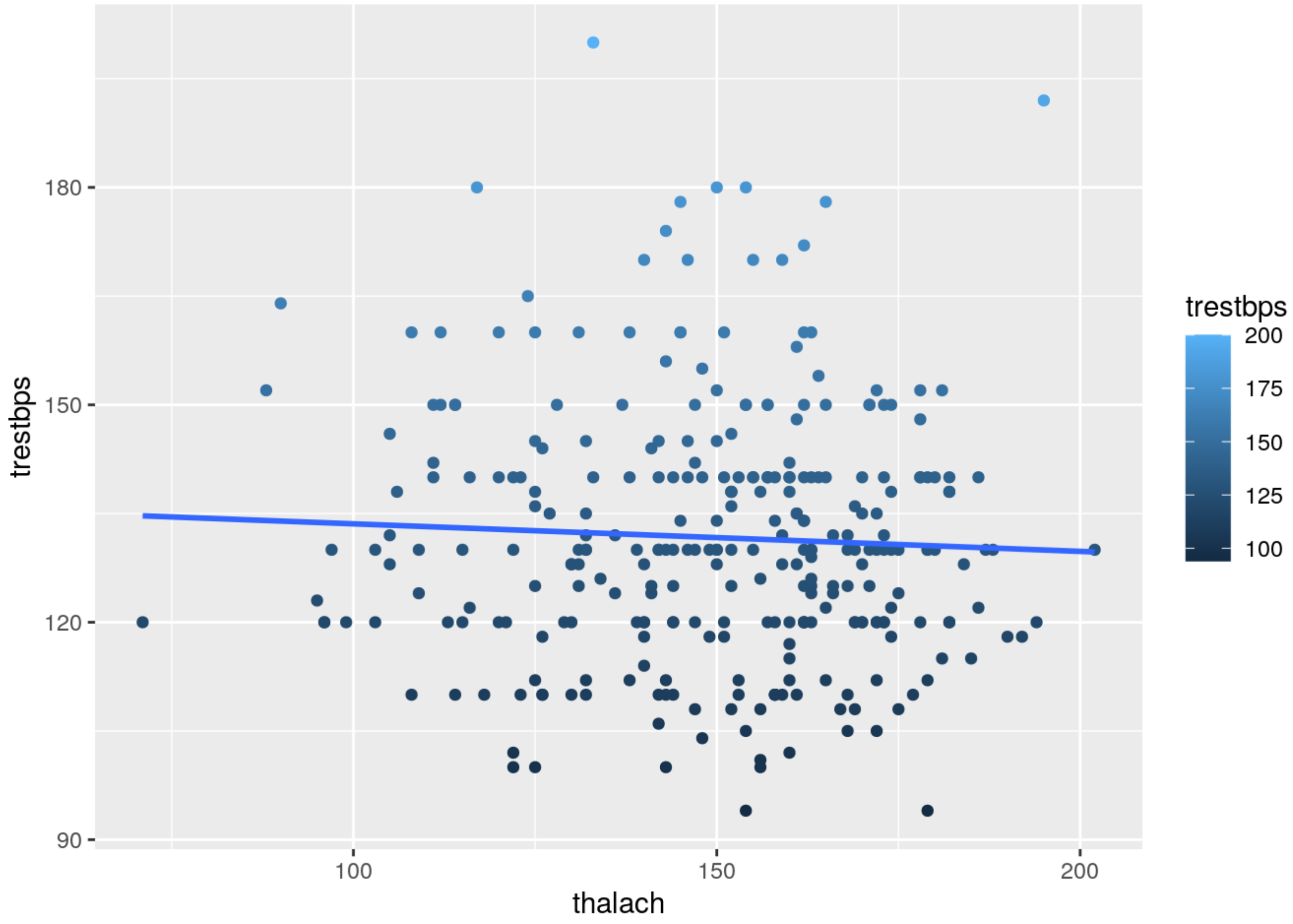
```
# Visualize the relationship with the regression line
Heart %>%
  ggplot(aes(thalach,trestbps)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE)
```

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



```
# Visualize the relationships between the three variables
ggplot(Heart, aes(x = thalach, y = trestbps, color = trestbps)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
```

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



```
#Tidy the dataset by considering character variables as factors
Heart <- Heart %>%
  mutate_if(is.character, as.factor)
```

The Q-Q plot is approximately linear so the variable of normal distribution is met. The histogram however is a little skewed to the right, with small numbers around 50 The difference may be due to the outlier in row 152 so if that was removed it seems that normal distribution would be have been met. Heteroskedasticity also seems to have been met since the observations are randomly scattered. After analyzing the residual plot one can conclude that the random observation assumption can also be met. The residuals vs. fitted graph indicates that a quadratic fit is not better then a linear fit. After computing a linear regression model our estimates state that on average males are more likely to get heart disease over women. Controlling for resting heart rate there is not a significant effect of cholesterol on Heart Disease. After controlling for cholesterol there is not difference on Heart diseases between patients with high and low Resting heart rate. Therefore we can conclude that both cholesterol and Trestbps are not a good predictor of heart disease. Sex is a good predictor of heart disease.

##G.Logistic Regression

```
# Create a binary variable coded as 0 and 1
Heart <- data %>%
  mutate(y = ifelse(hd == "Healthy",1, 0))

# Fit a new regression model
Log_heart <- glm(y ~ age + sex + thalach, data = Heart, family = binomial(link="logit"))
summary(Log_heart)
```

```
##
## Call:
## glm(formula = y ~ age + sex + thalach, family = binomial(link = "logit"),
##     data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1055  -0.8980   0.4231   0.8418   2.2358
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.193343   1.633897  -1.954   0.0507 .
## age         -0.032046   0.016630  -1.927   0.0540 .
## sexM        -1.466780   0.308529  -4.754 1.99e-06 ***
## thalach      0.040926   0.007172   5.706 1.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 325.12  on 293  degrees of freedom
## AIC: 333.12
##
## Number of Fisher Scoring iterations: 4
```

```
# Interpret the coefficients by considering the odds (inverse of log(odds))
exp(coef(Log_heart))
```

```
## (Intercept)         age        sexM     thalach
##  0.04103445  0.96846248  0.23066703  1.04177534
```

```
#
```

In this model we are using sex, age, and max heart rate in order to predict heart disease. rate. According to the result both sex and max heart rate are statistically significant since they are less then a p-value of 0.05. If we are predicting heart disease for a female patient we would 4.03 + (-1.38) x 0 therefore the log(odds) that a female would have heart disease is 4.03. If were were predicting if a female would have heart disease we would use the equation 4.03 + (-1.38) x 1 which shows that males are more likely to get heart disease then a female. We can conclude that sex and thalach are a good predictor of heart disease since it they both have a small p-value.

4. H Logstic Regression

```
#Interpret the Coefficents

# Write the model with odds-coefficients (multiplicative)
coef(Log_heart) %>% exp %>% round(5) %>% data.frame
```

```
##                     .
## (Intercept) 0.04103
## age         0.96846
## sexM        0.23067
## thalach     1.04178
```

```
# Add predicted value to the data set
Heart$prob <- predict(Log_heart, type = "response")

#Predicted outcomes based ont he porbablity of unhealthy

#if the probability is greater then .5, then the sex is found to be unhealthy
Heart$predicted <- ifelse(Heart$prob > .5, "Unhealthy", "Healthy")


# Confusion matrix: compare true to predicted condition
conf_matrix_objc <- table(truth = Heart$hd, predicted_condition = Heart$predicted) %>%
  addmargins

# Accuracy (correctly classified cases)
(27 + 98)/297
```

```
## [1] 0.4208754
```

```
# Sensitivity (True Positive Rate, TPR)
98/137
```

```
## [1] 0.7153285
```

```
# Specificity (True Negative Rate, TNR)
27/160
```

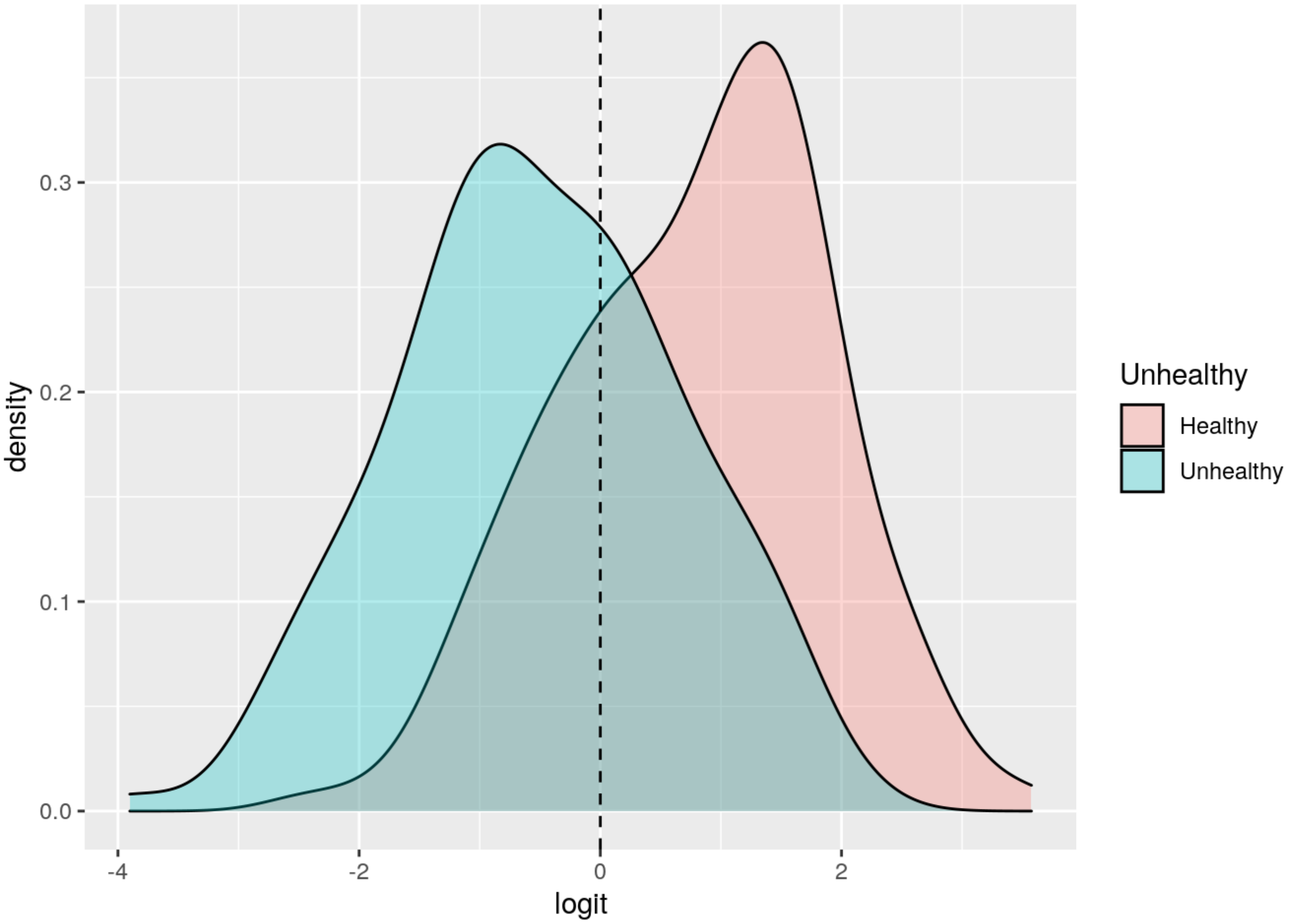
```
## [1] 0.16875
```

```
# Precision (Positive Predictive Value, PPV)
98/297
```

```
## [1] 0.3299663
```

The Accuracy or the proportion of correctly classified cases is 42%. The sensitivity and proportion of true positive cases is 71%. The specificity or proportion of true negative cases is 16% and the precision is 33%

```
# Save the predicted log-odds in the dataset
Heart$logit <- predict(Log_heart)

# Compare to the outcome in the dataset with a density plot
ggplot(Heart, aes(logit, fill = as.factor(hd))) +
  geom_density(alpha = .3) +
  geom_vline(xintercept = 0, lty = 2) +
  labs(fill = "Unhealthy")
```



```
# Confusion matrix: compare true to predicted condition
table(true_condition = Heart$hd, predicted_condition = Heart$predicted) %>%
  addmargins
```

```
##                predicted_condition
## true_condition Healthy Unhealthy Sum
##      Healthy        40       120 160
##      Unhealthy      92        45 137
##      Sum           132       165 297
```

```
# Predicted log odds
Heart$logit <- predict(Log_heart, type = "link")
```

Since the darker area represents the clumps that were misclassified, it seems like there were alot of cases that were misclassified.

5. I. ROC curve

```
# Call the library plotROC
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

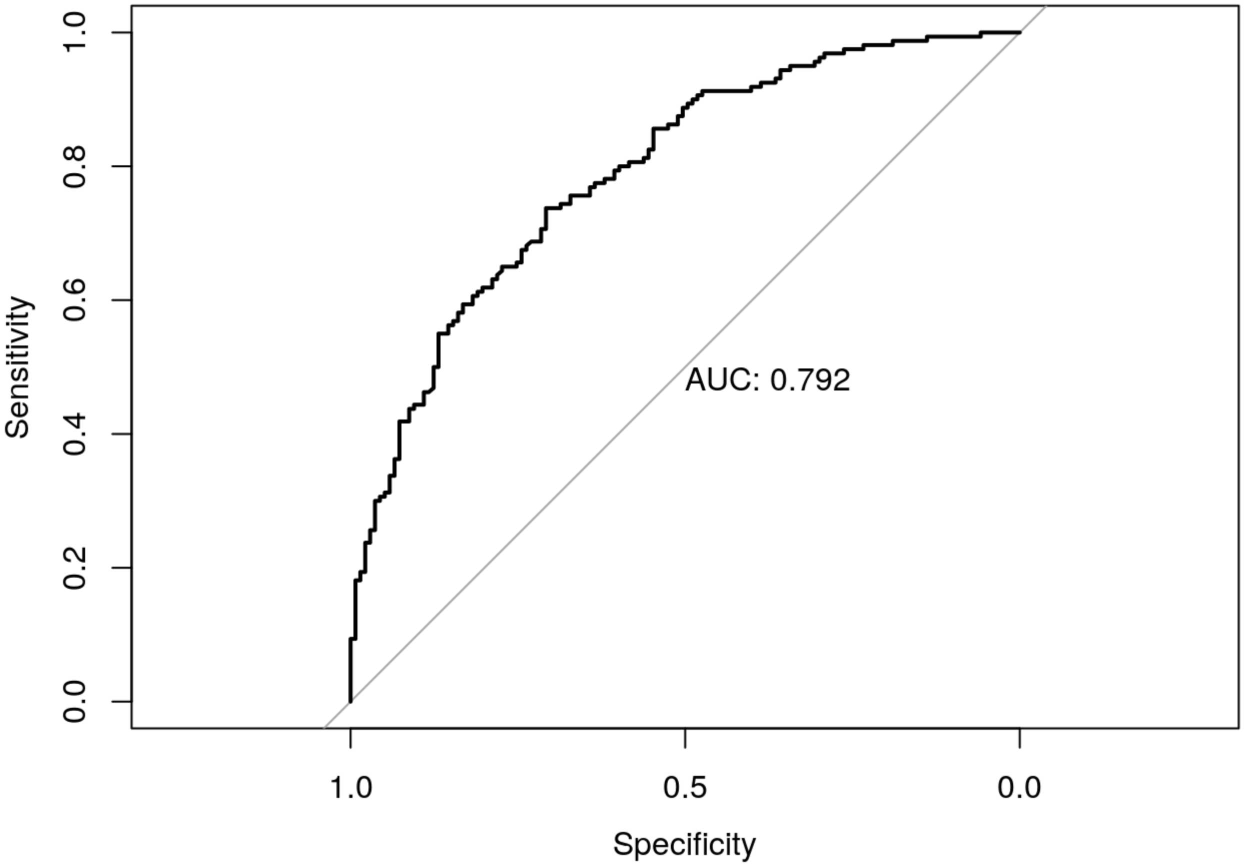
```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
test_prob = predict(Log_heart, newdata = Heart, type = "response")
test_roc = roc(Heart$y ~ test_prob, plot = TRUE, print.auc = TRUE)
```

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

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



```
as.numeric(test_roc$auc)
```

```
## [1] 0.7919024
```

From the output, we can see that the model's accuracy is 0.79. Since the AUC is extremely low we do not have a good model. Therefore the curve does not indicate a great performance since the curve is not close to the top left corner. In conclusion since the AUC is 0.6-0.7 it is poor and the curve is not good it may be hard to predict heart disease based on sex and resting heart rate

##Conclusion
In conclusion we can use the variables of Sex and Max heaet rate achieved in order to predict Heart Disease. Males are more likely to be diagnosed with Heart Disease over females and the lower ones heart rate achieved is then the more likely they are to have Heart Disease problems. Contrary to my prediction both age and cholesterol were not good predictors of Heart Disease.

##Citations *1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.* Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.