Heart Attack Prediction

Introduction:

```
#checking for missing data
sum(is.na(heart))
```

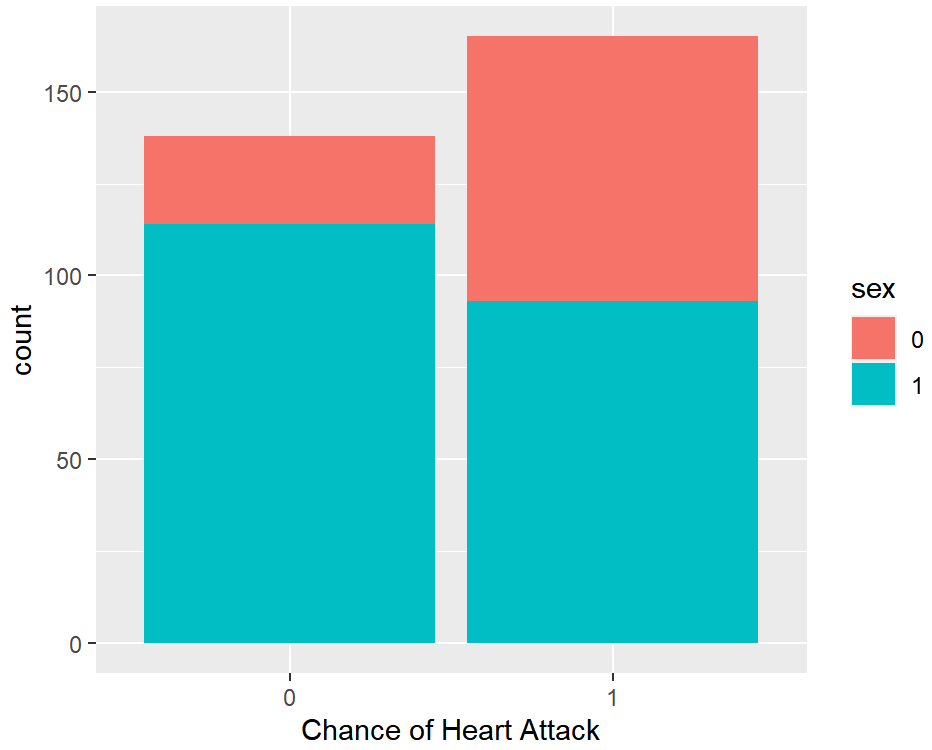
```
## [1] 0
```

```
#need to convert some columns into factors
factor_cols <- c("sex", "cp", "fbs", "restecg", "exng", "slp", "caa", "thall", "output")
heart[factor_cols] <- lapply(heart[factor_cols], factor)
summary(heart)
```

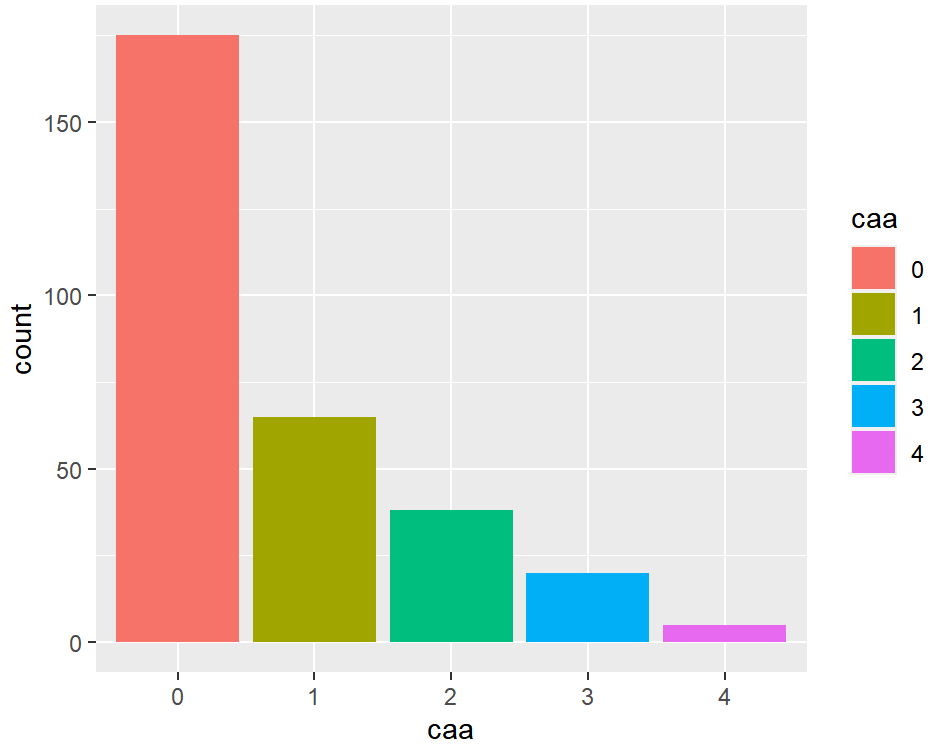
```
##       age        sex     cp          trtbps           chol       fbs
##  Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :55.00           2: 87   Median :130.0   Median :240.0
##  Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3
##  3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5
##  Max.   :77.00                   Max.   :200.0   Max.   :564.0
##  restecg    thalachh     exng       oldpeak     slp     caa     thall   output
##  0:147   Min.   : 71.0   0:204   Min.   :0.00   0: 21   0:175   0:  2   0:138
##  1:152   1st Qu.:133.5   1: 99   1st Qu.:0.00   1:140   1: 65   1: 18   1:165
##  2:  4   Median :153.0           Median :0.80   2:142   2: 38   2:166
##          Mean   :149.6           Mean   :1.04           3: 20   3:117
##          3rd Qu.:166.0           3rd Qu.:1.60           4:  5
##          Max.   :202.0           Max.   :6.20
```

Graphs

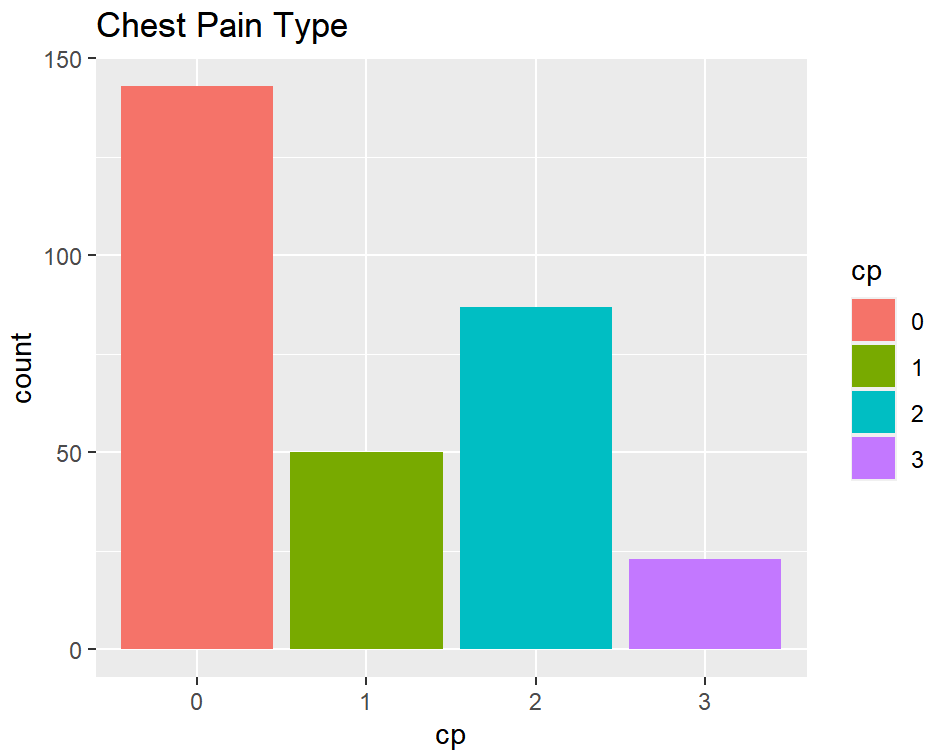
```
#data visualizations
ggplot(heart, aes(x = output, fill = sex)) + geom_bar() + labs(x = "Chance of Heart Attack")
```




```
ggplot(heart, aes(x = caa, fill = caa)) + geom_bar()
```

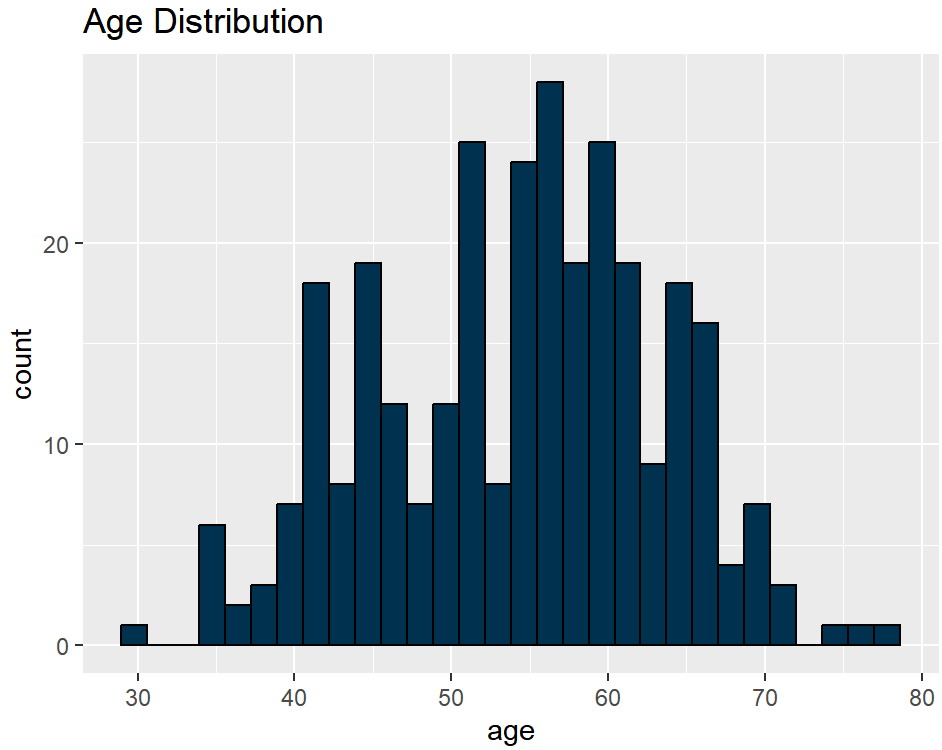



```
ggplot(heart, aes(x = cp, fill = cp)) + geom_bar() + ggtitle("Chest Pain Type")
```



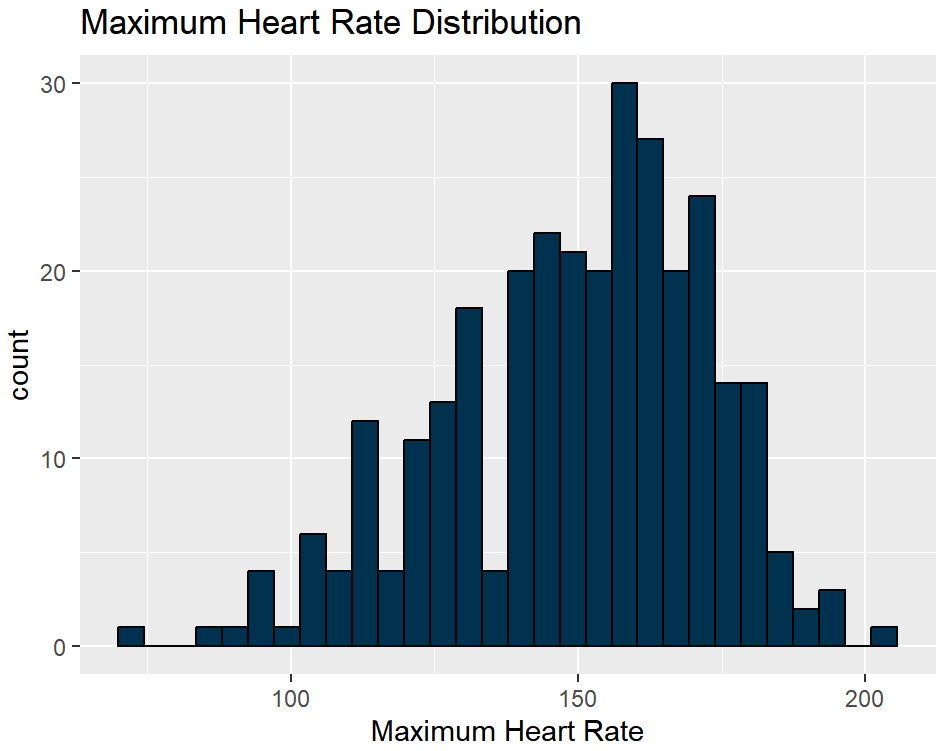

```
ggplot(heart, aes(x = age)) + geom_histogram(color = "black", fill = "#003554") + ggtitle("Age D
istribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(heart, aes(x = thalachh)) + geom_histogram(color = "black", fill = "#003554") + ggtitle
("Maximum Heart Rate Distribution") + xlab("Maximum Heart Rate")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Numerical Summary Analysis:

The numerical summary reveals that men are represented two times more than females. This could mean that this data set will be more accurate at predicting heart attack likelihood for males than females. The youngest participant is 29 while the oldest participant is 77. The mean age is 54.37 which is close to the median age of 55, so the data doesn't appear to be skewed in terms of age representation. The data is slightly left skewed in terms of maximum heart rate and right skewed in terms of oldpeak.

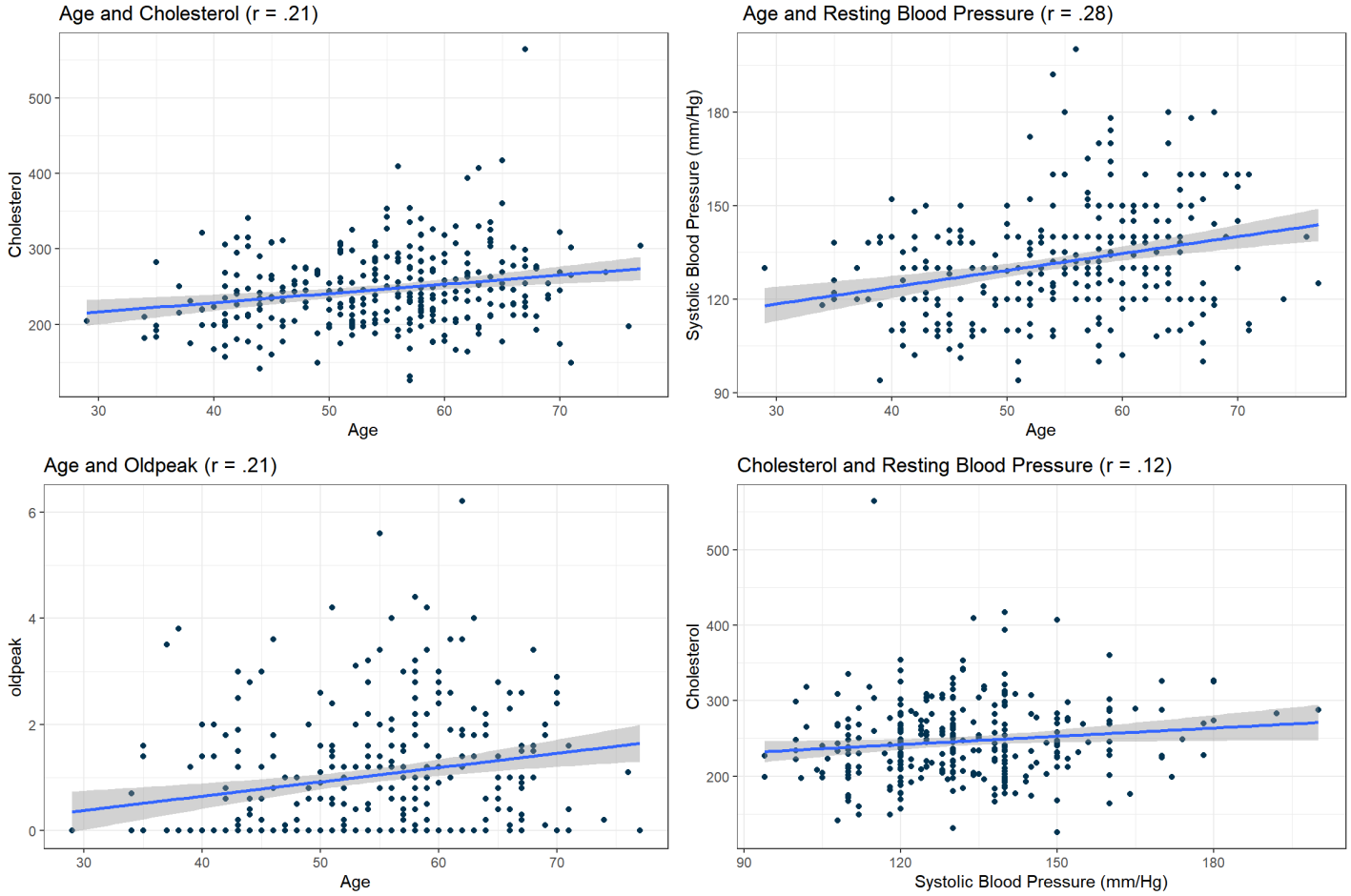
An output of 0 means there's a lower chance of a heart attack and an output of 1 means there's a higher chance of a heart attack.

```
#correlations
cor(heart$age, heart$chol)
cor(heart$age, heart$trtbps)
cor(heart$age, heart$oldpeak)
cor(heart$trtbps, heart$chol)
```

```
a = ggplot(heart, aes( x = age, y = chol)) + geom_point(color = "#003554") + geom_smooth(method
= "lm") + theme_bw() + labs(title = "Age and Cholesterol (r = .21)", x = "Age", y = "Cholestero
l")
b = ggplot(heart, aes( x = age, y = trtbps)) + geom_point(color = "#003554") + geom_smooth(metho
d = "lm") + theme_bw() + labs(title = " Age and Resting Blood Pressure (r = .28)", y = "Systolic
Blood Pressure (mm/Hg)", x = "Age")
c = ggplot(heart, aes( x = age, y = oldpeak)) + geom_point(color = "#003554") + geom_smooth(meth
od = "lm") + theme_bw() + labs(title = "Age and Oldpeak (r = .21)", x = "Age")
d = ggplot(heart, aes( x = trtbps, y = chol)) + geom_point(color = "#003554") + geom_smooth(meth
od = "lm") + theme_bw() + labs(title = "Cholesterol and Resting Blood Pressure (r = .12)", y =
"Cholesterol", x = "Systolic Blood Pressure (mm/Hg)")

ggarrange(a,b,c,d, ncol = 2, nrow = 2)
```

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



The correlations between the continuous variables are all positive and relatively week. This could mean that there is a stronger non-linear relationship present, but the scatter plots of these interactions seems to support that there is simply a weak relationship between these pairs.

```
#building a logistic model
log_mod <- glm(output~., heart, family = "binomial")
summary(log_mod)
```

```
##
## Call:
## glm(formula = output ~ ., family = "binomial", data = heart)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.179045   3.705420   0.048 0.961461
## age          0.027819   0.025428   1.094 0.273938
## sex1        -1.862297   0.570844  -3.262 0.001105 **
## cp1          0.864708   0.578000   1.496 0.134645
## cp2          2.003186   0.529356   3.784 0.000154 ***
## cp3          2.417107   0.719242   3.361 0.000778 ***
## trtbps      -0.026162   0.011943  -2.191 0.028481 *
## chol        -0.004291   0.004245  -1.011 0.312053
## fbs1         0.445666   0.587977   0.758 0.448472
## restecg1     0.460582   0.399615   1.153 0.249089
## restecg2    -0.714204   2.768873  -0.258 0.796453
## thalachh     0.020055   0.011859   1.691 0.090820 .
## exng1       -0.779111   0.451839  -1.724 0.084652 .
## oldpeak     -0.397174   0.242346  -1.639 0.101239
## slp1        -0.775084   0.880495  -0.880 0.378707
## slp2         0.689965   0.947657   0.728 0.466568
## caa1        -2.342301   0.527416  -4.441 8.95e-06 ***
## caa2        -3.483178   0.811640  -4.292 1.77e-05 ***
## caa3        -2.247144   0.937629  -2.397 0.016547 *
## caa4         1.267961   1.720014   0.737 0.461013
## thall1       2.637558   2.684285   0.983 0.325808
## thall2       2.367747   2.596159   0.912 0.361759
## thall3       0.915115   2.600380   0.352 0.724901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 179.63  on 280  degrees of freedom
## AIC: 225.63
##
## Number of Fisher Scoring iterations: 6
```

7 out of the 12 predictors are categorical, so the logistic regression summary is not enough for us to exclude any of the categorical predictors. Some of the categorical predictors are not significant at when $\alpha = .05$, but this only means that there is not a significant different from their respective reference variable, not that the entire category is non-significant.

Random Forest

```
#splitting the data into train and test set
set.seed(123)
heart_0 = which(heart$output == 0)
heart_1 = which(heart$output == 1)
train_id = c(sample(heart_0, size = trunc(0.70 * length(heart_0))),
sample(heart_1, size = trunc(0.70 * length(heart_1))))
heart_train = heart[train_id, ]
heart_test = heart[-train_id, ]

nrow(heart_train)
```

```
## [1] 211
```

```
nrow(heart_test)
```

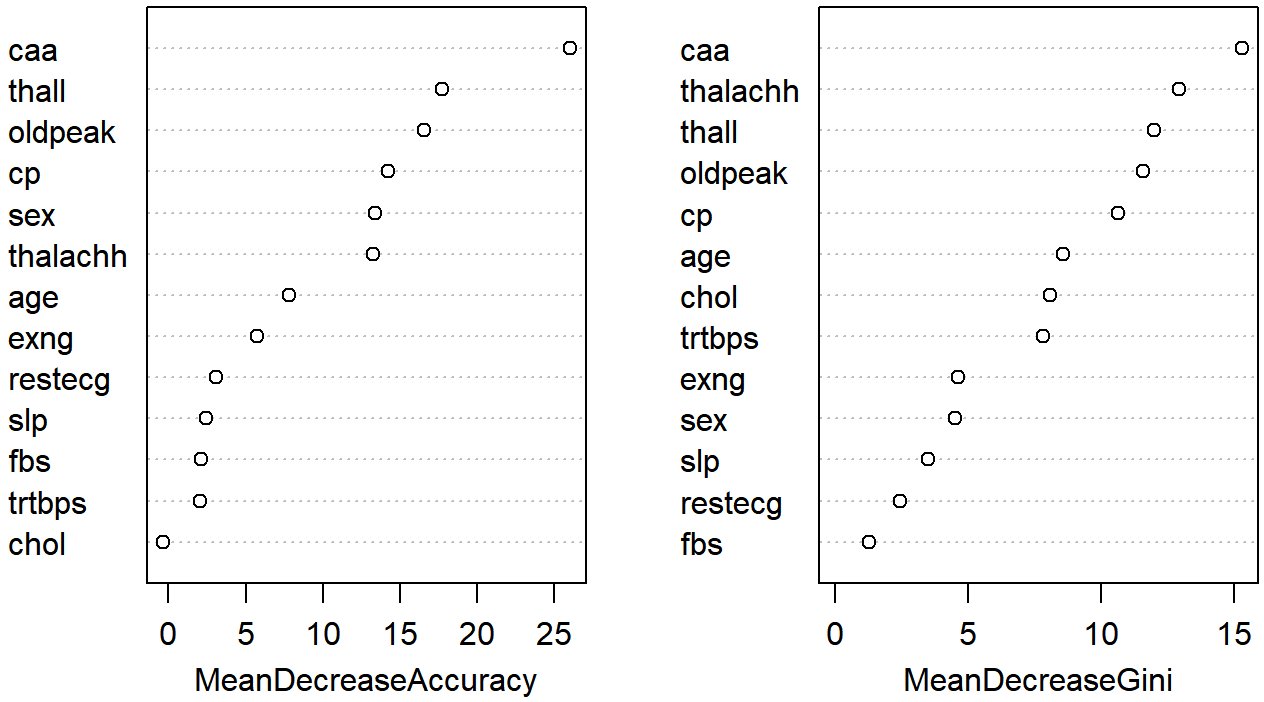
```
## [1] 92
```

```
rf_mod <- randomForest(
  output~.,
  data=heart_train,
  importance=TRUE)

rf_mod
```

```
##
## Call:
##  randomForest(formula = output ~ ., data = heart_train, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 18.01%
## Confusion matrix:
##    0  1 class.error
## 0 74 22   0.2291667
## 1 16 99   0.1391304
```

```
varImpPlot(rf_mod)
```




```
rf_test_preds <- predict(rf_mod, newdata=heart_test, type = "class")
roc(response= rf_test_preds, predictor= factor(heart_test$output,
ordered = TRUE))
```

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

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

```
##
## Call:
## roc.default(response = rf_test_preds, predictor = factor(heart_test$output,     ordered = TRU
E))
##
## Data: factor(heart_test$output, ordered = TRUE) in 40 controls (rf_test_preds 0) < 52 cases
(rf_test_preds 1).
## Area under the curve: 0.8702
```

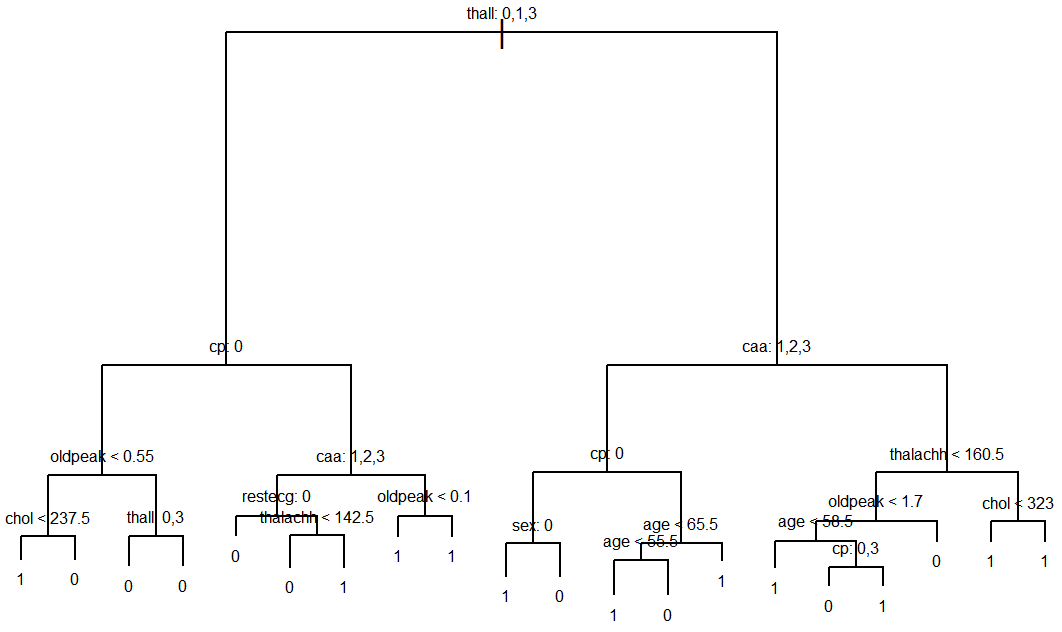
For both measurements, caa is the most important variable for the random forest. caa causes the greatest mean decrease in accuracy and decreases in the Gini score when it is excluded from the model.

Tree Diagrams

```
tree.heart=tree(output~., heart)
summary(tree.heart)
```

```
##
## Classification tree:
## tree(formula = output ~ ., data = heart)
## Variables actually used in tree construction:
## [1] "thall"    "cp"       "oldpeak"  "chol"     "caa"      "restecg"  "thalachh"
## [8] "sex"      "age"
## Number of terminal nodes:  20
## Residual mean deviance:  0.4802 = 135.9 / 283
## Misclassification error rate: 0.1122 = 34 / 303
```

```
plot(tree.heart)
text(tree.heart,pretty=0,cex=0.5)
```



The tree diagram supports thall being the most influential predictor of heart attack risk because it is the first root node. However, caa and cp are also strong predictors which matches the results from the random forest.