StatsProj2WolfeRF

Michael J Wolfe

August 2, 2019

## Packages

## Loading required package: bitops

##   
## 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

## Registered S3 methods overwritten by 'ggplot2':  
## method from   
## [.quosures rlang  
## c.quosures rlang  
## print.quosures rlang

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

## Warning: package 'randomForest' was built under R version 3.6.1

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

## The following object is masked from 'package:dplyr':  
##   
## combine

## Warning: package 'rgl' was built under R version 3.6.1

## Warning: package 'tree' was built under R version 3.6.1

## Warning: package 'ISLR' was built under R version 3.6.1

## Warning: package 'ROCR' was built under R version 3.6.1

## Warning: package 'pheatmap' was built under R version 3.6.1

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

## Warning: package 'rpart.plot' was built under R version 3.6.1

## Warning: package 'ROSE' was built under R version 3.6.1

## Loaded ROSE 0.0-3

The city of Framingham is concerned about the long term health of some of its citizens. In this exercise we will test probability models on a variety of factors that lead to coronary heart disease and recommend one solution.

## Load Data

## male age education currentSmoker cigsPerDay BPMeds prevalentStroke  
## 1 1 39 4 0 0 0 0  
## 2 0 46 2 0 0 0 0  
## 3 1 48 1 1 20 0 0  
## 4 0 61 3 1 30 0 0  
## 5 0 46 3 1 23 0 0  
## 6 0 43 2 0 0 0 0  
## prevalentHyp diabetes totChol sysBP diaBP BMI heartRate glucose  
## 1 0 0 195 106.0 70 26.97 80 77  
## 2 0 0 250 121.0 81 28.73 95 76  
## 3 0 0 245 127.5 80 25.34 75 70  
## 4 1 0 225 150.0 95 28.58 65 103  
## 5 0 0 285 130.0 84 23.10 85 85  
## 6 1 0 228 180.0 110 30.30 77 99  
## TenYearCHD  
## 1 0  
## 2 0  
## 3 0  
## 4 1  
## 5 0  
## 6 0

Some of these data have categories that aren’t particularly intuitive. Let’s recode some of them.

## Rename columns

## [1] "male" "female"

## [1] "College" "FinishedHighSchool/GED"   
## [3] "SomeHighSchool" "SomeCollege/VocationalSchool"  
## [5] NA

## [1] "No" "Yes"

## [1] "No" "Yes" NA

## [1] 0 1

## [1] 0 1

## [1] 0 1

## [1] 0 1

## EDA

## Column analysis and cleanup

Let’s first confirm that our datatypes agree with the analysis we want to do.

## Gender age education currentSmoker   
## "character" "integer" "character" "character"   
## cigsPerDay BPMeds prevalentStroke prevalentHyp   
## "integer" "character" "character" "character"   
## diabetes totChol sysBP diaBP   
## "character" "integer" "numeric" "numeric"   
## BMI heartRate glucose TenYearCHD   
## "numeric" "integer" "integer" "character"

Those character columns would do better as factors, so let’s change them.

Now let’s take some sums and summaries.

## age cigsPerDay totChol sysBP diaBP BMI   
## 210220.0 NA NA 561183.5 351486.5 NA   
## heartRate glucose   
## NA NA

## Gender age education currentSmoker   
## 0 0 105 0   
## cigsPerDay BPMeds prevalentStroke prevalentHyp   
## 29 53 0 0   
## diabetes totChol sysBP diaBP   
## 0 50 0 0   
## BMI heartRate glucose TenYearCHD   
## 19 1 388 0

## Gender age education   
## female:2420 Min. :32.00 College : 473   
## male :1820 1st Qu.:42.00 FinishedHighSchool/GED :1253   
## Median :49.00 SomeCollege/VocationalSchool: 689   
## Mean :49.58 SomeHighSchool :1720   
## 3rd Qu.:56.00 NA's : 105   
## Max. :70.00   
##   
## currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp  
## No :2145 Min. : 0.000 No :4063 No :4215 No :2923   
## Yes:2095 1st Qu.: 0.000 Yes : 124 Yes: 25 Yes:1317   
## Median : 0.000 NA's: 53   
## Mean : 9.006   
## 3rd Qu.:20.000   
## Max. :70.000   
## NA's :29   
## diabetes totChol sysBP diaBP   
## No :4131 Min. :107.0 Min. : 83.5 Min. : 48.0   
## Yes: 109 1st Qu.:206.0 1st Qu.:117.0 1st Qu.: 75.0   
## Median :234.0 Median :128.0 Median : 82.0   
## Mean :236.7 Mean :132.4 Mean : 82.9   
## 3rd Qu.:263.0 3rd Qu.:144.0 3rd Qu.: 90.0   
## Max. :696.0 Max. :295.0 Max. :142.5   
## NA's :50   
## BMI heartRate glucose TenYearCHD  
## Min. :15.54 Min. : 44.00 Min. : 40.00 No :3596   
## 1st Qu.:23.07 1st Qu.: 68.00 1st Qu.: 71.00 Yes: 644   
## Median :25.40 Median : 75.00 Median : 78.00   
## Mean :25.80 Mean : 75.88 Mean : 81.96   
## 3rd Qu.:28.04 3rd Qu.: 83.00 3rd Qu.: 87.00   
## Max. :56.80 Max. :143.00 Max. :394.00   
## NA's :19 NA's :1 NA's :388

Looks like there’s some missing values in cigsPerDay, totChol, heartRate, and glucose. Let’s see how much leverage these values have.

## [1] 4240 16

## [1] 3658 16

## [1] 0.8627358

We can keep 86% of our values, so let’s go ahead and remove these rows.

## [1] 3658 16

## Gender age education   
## female:2035 Min. :32.00 College : 423   
## male :1623 1st Qu.:42.00 FinishedHighSchool/GED :1101   
## Median :49.00 SomeCollege/VocationalSchool: 608   
## Mean :49.55 SomeHighSchool :1526   
## 3rd Qu.:56.00   
## Max. :70.00   
## currentSmoker cigsPerDay BPMeds prevalentStroke prevalentHyp  
## No :1869 Min. : 0.000 No :3547 No :3637 No :2518   
## Yes:1789 1st Qu.: 0.000 Yes: 111 Yes: 21 Yes:1140   
## Median : 0.000   
## Mean : 9.025   
## 3rd Qu.:20.000   
## Max. :70.000   
## diabetes totChol sysBP diaBP   
## No :3559 Min. :113.0 Min. : 83.5 Min. : 48.00   
## Yes: 99 1st Qu.:206.0 1st Qu.:117.0 1st Qu.: 75.00   
## Median :234.0 Median :128.0 Median : 82.00   
## Mean :236.8 Mean :132.4 Mean : 82.92   
## 3rd Qu.:263.0 3rd Qu.:143.9 3rd Qu.: 90.00   
## Max. :600.0 Max. :295.0 Max. :142.50   
## BMI heartRate glucose TenYearCHD  
## Min. :15.54 Min. : 44.00 Min. : 40.00 No :3101   
## 1st Qu.:23.08 1st Qu.: 68.00 1st Qu.: 71.00 Yes: 557   
## Median :25.38 Median : 75.00 Median : 78.00   
## Mean :25.78 Mean : 75.73 Mean : 81.85   
## 3rd Qu.:28.04 3rd Qu.: 82.00 3rd Qu.: 87.00   
## Max. :56.80 Max. :143.00 Max. :394.00

Much better! However, it would be easier to interpret the model outcome as a person with \_ is \_ more likely to be diagnosed with heart disease, so let’s adjust the response variable accordingly.

## [1] "No" "Yes"

##   
## No Yes   
## 3101 557

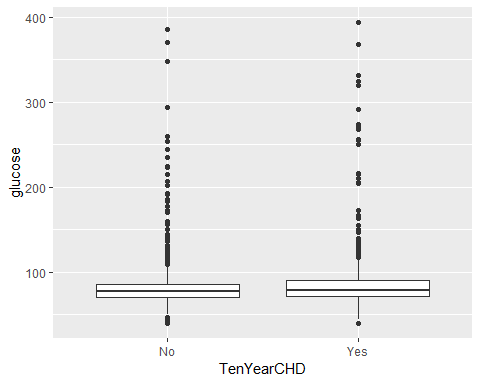
## [1] 0.152269

After re-factoring the response variable it appears to be an unbalanced binary, with only about 15% of residents developing coronary heart disease after 10 years. This will need to be considered when created confusion matrices and scoring accuracy. We will likely need to create several cross-validation iterations, and possibly downsample if it appears our train/test splits skew too heavily towards “No” and create a falsly inflated accuracy score.

## Continuous variables

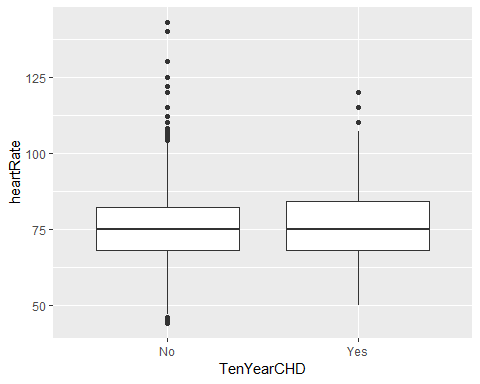
Let’s start by checking correlations.

## TenYearCHD glucose.Min. glucose.1st Qu. glucose.Median glucose.Mean  
## 1 No 40.00000 71.00000 78.00000 80.61722  
## 2 Yes 40.00000 72.00000 79.00000 88.73250  
## glucose.3rd Qu. glucose.Max.  
## 1 86.00000 386.00000  
## 2 90.00000 394.00000



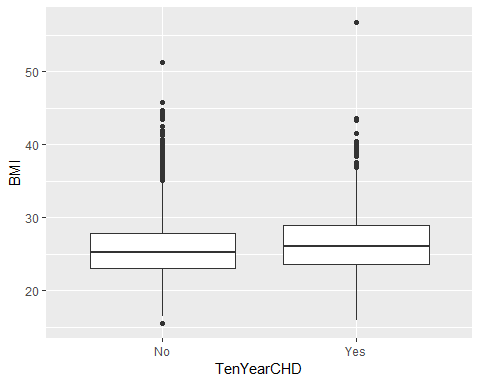
Lots of outliers here. Hard to say if there is correlation because of this.

## TenYearCHD heartRate.Min. heartRate.1st Qu. heartRate.Median  
## 1 No 44.00000 68.00000 75.00000  
## 2 Yes 50.00000 68.00000 75.00000  
## heartRate.Mean heartRate.3rd Qu. heartRate.Max.  
## 1 75.62657 82.00000 143.00000  
## 2 76.31059 84.00000 120.00000



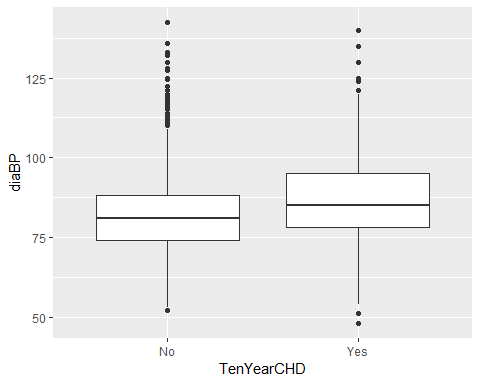
No correlation here.

## TenYearCHD BMI.Min. BMI.1st Qu. BMI.Median BMI.Mean BMI.3rd Qu. BMI.Max.  
## 1 No 15.54000 23.01000 25.23000 25.64144 27.86000 51.28000  
## 2 Yes 15.96000 23.63000 26.11000 26.56984 28.94000 56.80000



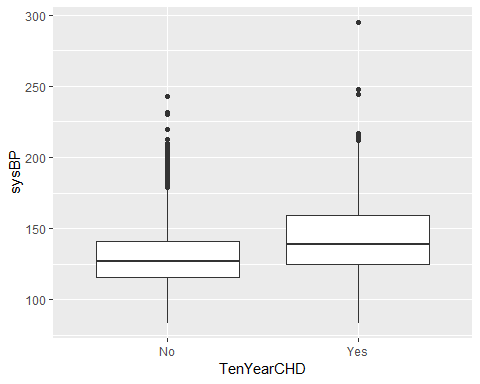
Looks like high BMI and heart disease might be correlated.

## TenYearCHD diaBP.Min. diaBP.1st Qu. diaBP.Median diaBP.Mean  
## 1 No 52.00000 74.00000 81.00000 82.15527  
## 2 Yes 48.00000 78.00000 85.00000 87.15799  
## diaBP.3rd Qu. diaBP.Max.  
## 1 88.00000 142.50000  
## 2 95.00000 140.00000



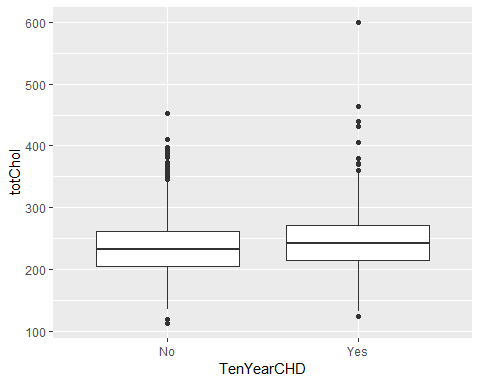
Similarly appears that high diastolic BP correlates to ten year CHD.

## TenYearCHD sysBP.Min. sysBP.1st Qu. sysBP.Median sysBP.Mean  
## 1 No 83.5000 116.0000 127.0000 130.2851  
## 2 Yes 83.5000 125.0000 139.0000 143.9811  
## sysBP.3rd Qu. sysBP.Max.  
## 1 141.0000 243.0000  
## 2 159.0000 295.0000



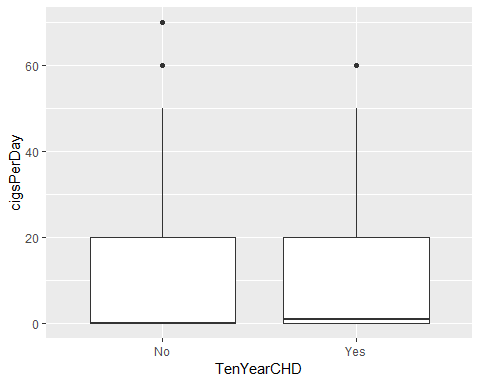
Same with systolic BP.

## TenYearCHD totChol.Min. totChol.1st Qu. totChol.Median totChol.Mean  
## 1 No 113.0000 205.0000 232.0000 235.1409  
## 2 Yes 124.0000 214.0000 243.0000 246.3501  
## totChol.3rd Qu. totChol.Max.  
## 1 261.0000 453.0000  
## 2 272.0000 600.0000



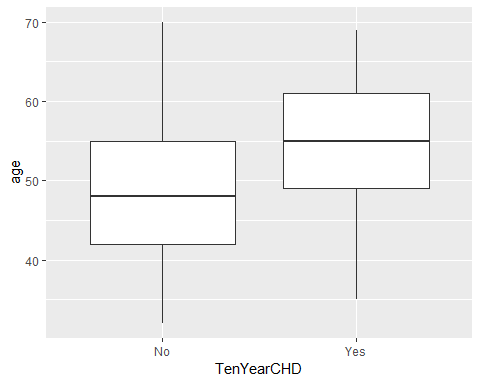
Again with cholesterol.

## TenYearCHD cigsPerDay.Min. cigsPerDay.1st Qu. cigsPerDay.Median  
## 1 No 0.000000 0.000000 0.000000  
## 2 Yes 0.000000 0.000000 1.000000  
## cigsPerDay.Mean cigsPerDay.3rd Qu. cigsPerDay.Max.  
## 1 8.762657 20.000000 70.000000  
## 2 10.488330 20.000000 60.000000



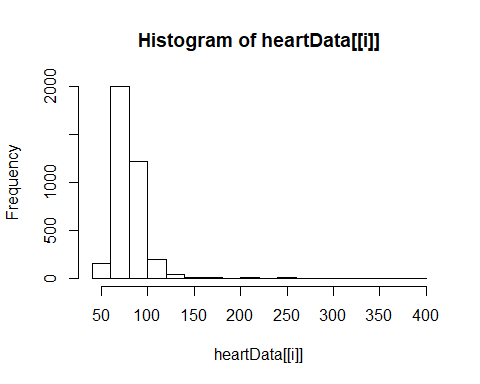
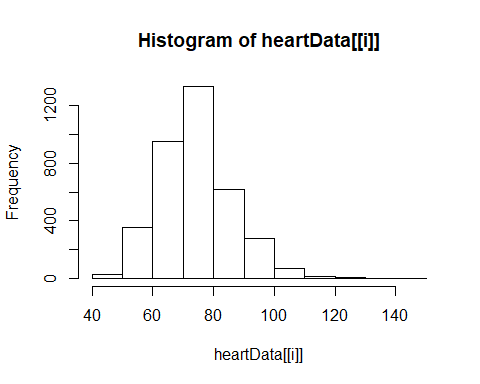
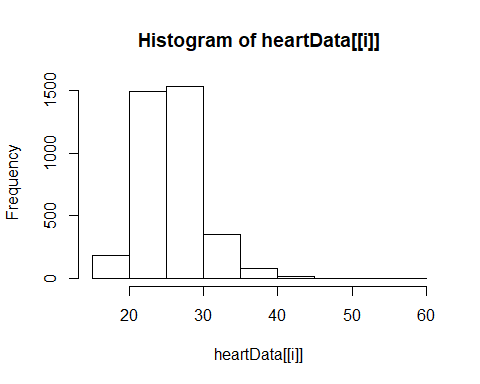
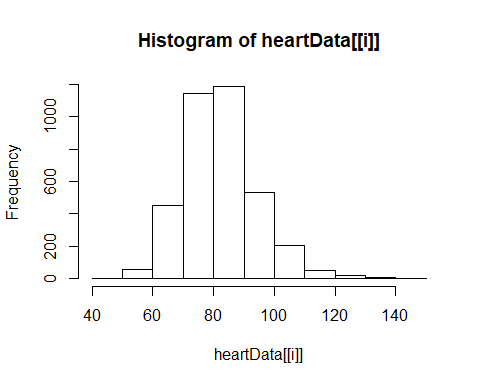
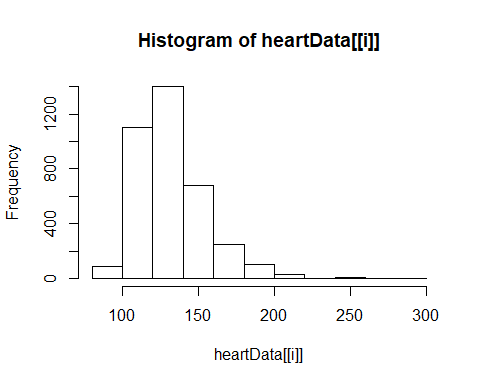
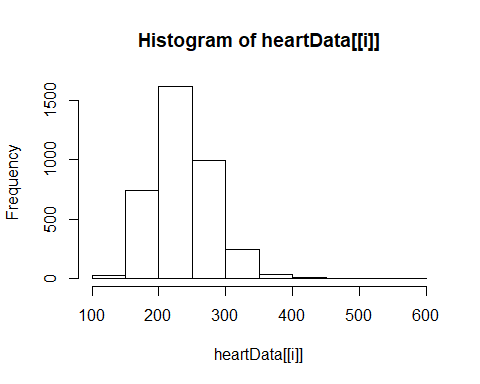
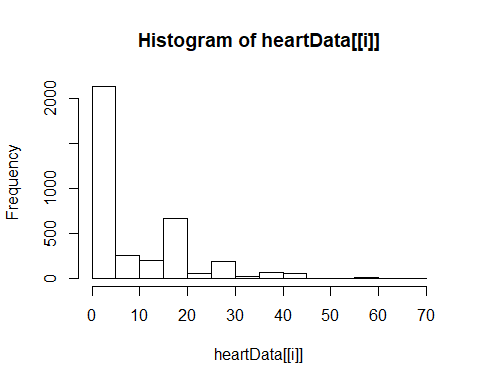
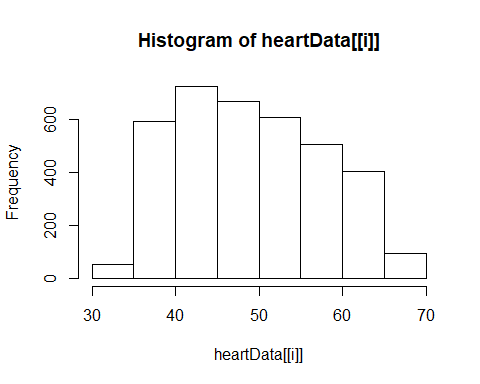
Yet again with cigarette use. Lots of outliers here too.

## TenYearCHD age.Min. age.1st Qu. age.Median age.Mean age.3rd Qu. age.Max.  
## 1 No 32.00000 42.00000 48.00000 48.70300 55.00000 70.00000  
## 2 Yes 35.00000 49.00000 55.00000 54.27828 61.00000 69.00000

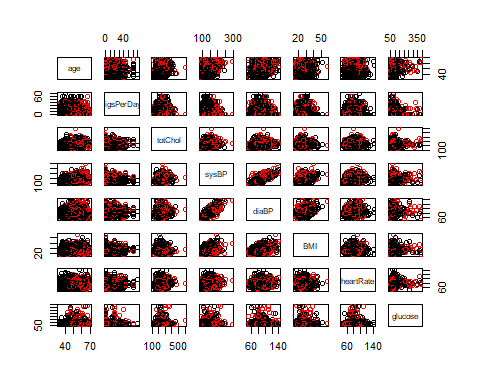
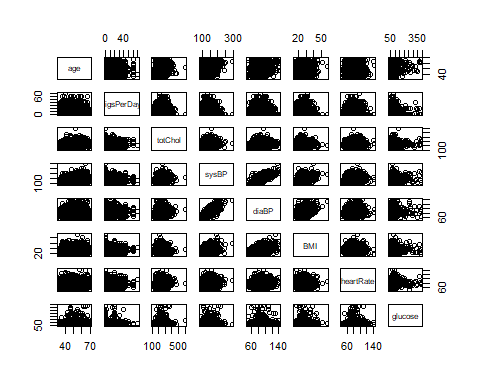


Higher age appears to be associated with CHD as well.

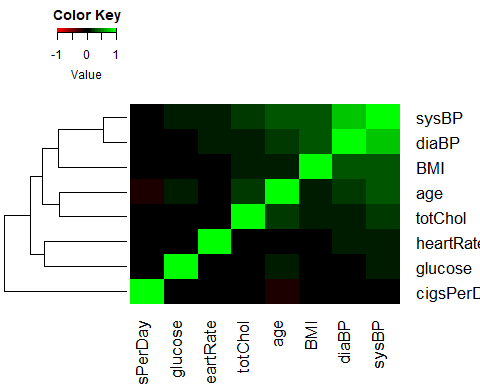
Let’s run some histograms.



Mostly normally distributed with a couple exceptions. Now let’s check mutlicolinearity and correlations.



## age cigsPerDay totChol sysBP diaBP  
## age 1.000000000 -0.18929528 0.26825160 0.38826667 0.20828288  
## cigsPerDay -0.189295281 1.00000000 -0.03040036 -0.09478131 -0.05674568  
## totChol 0.268251605 -0.03040036 1.00000000 0.21992459 0.17442197  
## sysBP 0.388266666 -0.09478131 0.21992459 1.00000000 0.78666936  
## diaBP 0.208282876 -0.05674568 0.17442197 0.78666936 1.00000000  
## BMI 0.137511104 -0.08739478 0.12105645 0.33091733 0.38534762  
## heartRate -0.002722424 0.06403043 0.09305284 0.18479691 0.17874359  
## glucose 0.118349131 -0.05372643 0.04988445 0.13465101 0.06354022  
## BMI heartRate glucose  
## age 0.13751110 -0.002722424 0.11834913  
## cigsPerDay -0.08739478 0.064030435 -0.05372643  
## totChol 0.12105645 0.093052845 0.04988445  
## sysBP 0.33091733 0.184796914 0.13465101  
## diaBP 0.38534762 0.178743586 0.06354022  
## BMI 1.00000000 0.074130514 0.08368315  
## heartRate 0.07413051 1.000000000 0.09707427  
## glucose 0.08368315 0.097074266 1.00000000

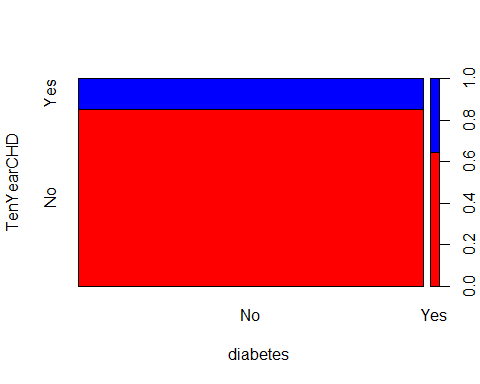


There’s a strong correlation between systolic blood pressure and diastolic blood pressure…which makes practical sense considering they are essentially the same measure. Age does a good job of stratifying the variables and reducing multicollinearity. Two takeaways from this - first might be to remove either systolic or diastolic BP, second is to perhaps create interactions on the age variable.

## Categorical variables

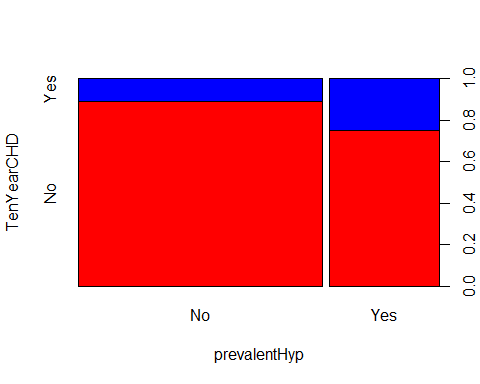
Next let’s check the categorical variables, first by measuring and plotting proportionality.

## diabetes  
## TenYearCHD No Yes  
## No 0.8533296 0.6464646  
## Yes 0.1466704 0.3535354



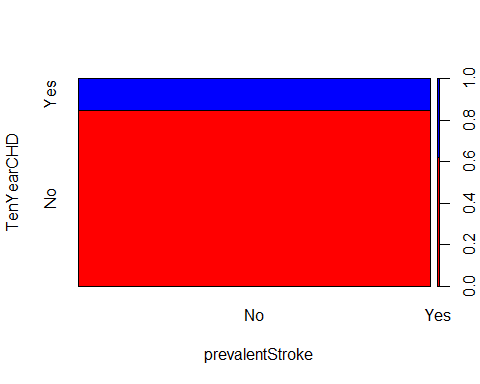
There appears to be an unbalanced response rate. We will need to keep this in mind when creating training data, and will likely necessitate a correction to the coefficient. It also appears that prevalent diabetes is associated with ten year CHD.

## prevalentHyp  
## TenYearCHD No Yes  
## No 0.8915806 0.7508772  
## Yes 0.1084194 0.2491228



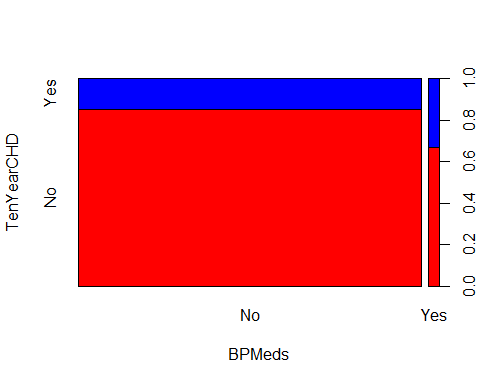
Again we see prevalent hypertension associated with ten year CHD.

## prevalentStroke  
## TenYearCHD No Yes  
## No 0.8490514 0.6190476  
## Yes 0.1509486 0.3809524



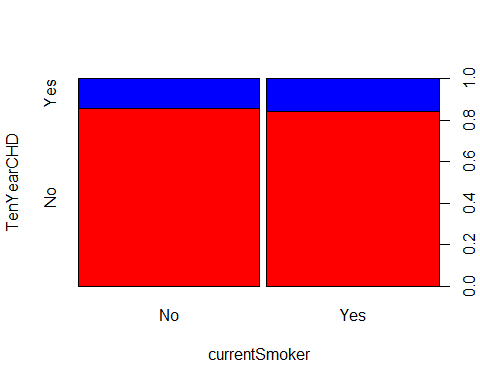
Another association

## BPMeds  
## TenYearCHD No Yes  
## No 0.8533972 0.6666667  
## Yes 0.1466028 0.3333333



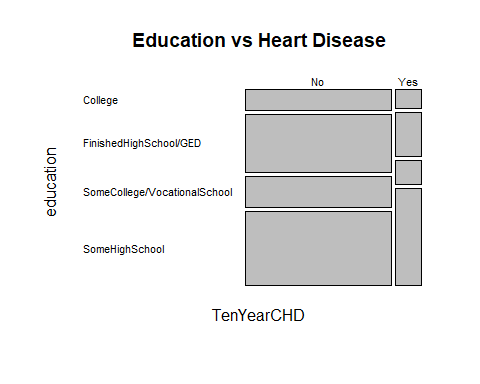
Another association, and class imbalance

## currentSmoker  
## TenYearCHD No Yes  
## No 0.8544676 0.8406931  
## Yes 0.1455324 0.1593069



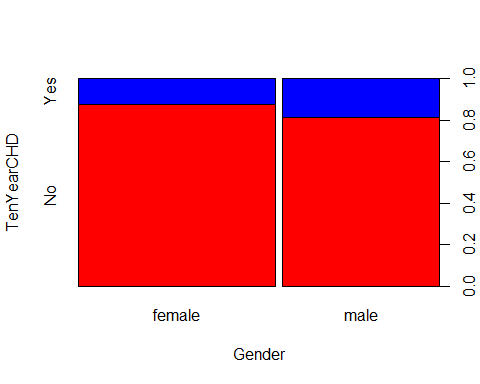
Oddly appears that there is no association between smoking and ten year CHD

## education  
## TenYearCHD College FinishedHighSchool/GED SomeCollege/VocationalSchool  
## No 0.8581560 0.8810173 0.8766447  
## Yes 0.1418440 0.1189827 0.1233553  
## education  
## TenYearCHD SomeHighSchool  
## No 0.8093054  
## Yes 0.1906946



Higher education appears to have lower rates of ten year CHD.

## currentSmoker  
## TenYearCHD No Yes  
## No 0.8544676 0.8406931  
## Yes 0.1455324 0.1593069



Males appear to develop CHD more often.

## No Yes Sum  
##   
## No 1597 1504 3101  
## Yes 272 285 557  
## Sum 1869 1789 3658

## female male Sum  
##   
## No 1785 1316 3101  
## Yes 250 307 557  
## Sum 2035 1623 3658

## College FinishedHighSchool/GED SomeCollege/VocationalSchool SomeHighSchool Sum  
##   
## No 363 970 533 1235 3101  
## Yes 60 131 75 291 557  
## Sum 423 1101 608 1526 3658

## No Yes Sum  
##   
## No 3027 74 3101  
## Yes 520 37 557  
## Sum 3547 111 3658

## No Yes Sum  
##   
## No 3088 13 3101  
## Yes 549 8 557  
## Sum 3637 21 3658

## No Yes Sum  
##   
## No 2245 856 3101  
## Yes 273 284 557  
## Sum 2518 1140 3658

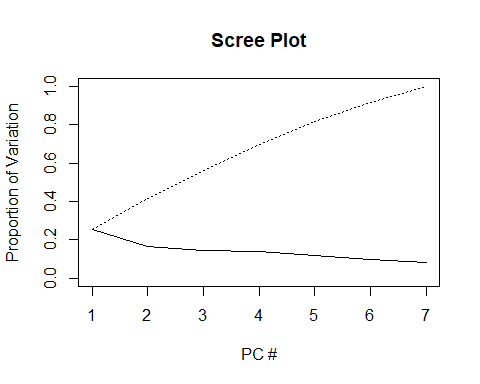
## No Yes Sum  
##   
## No 3037 64 3101  
## Yes 522 35 557  
## Sum 3559 99 3658

## PCA

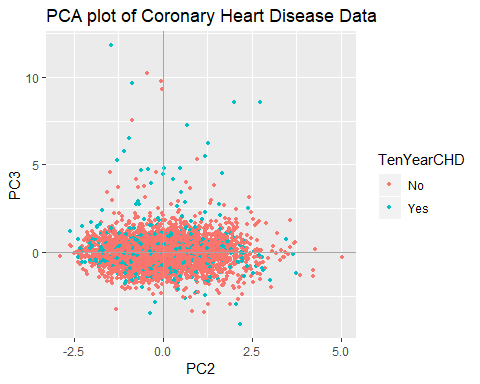
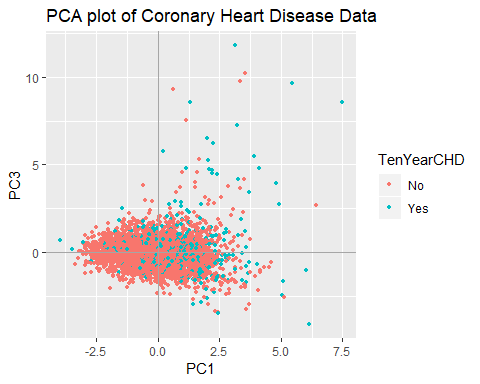
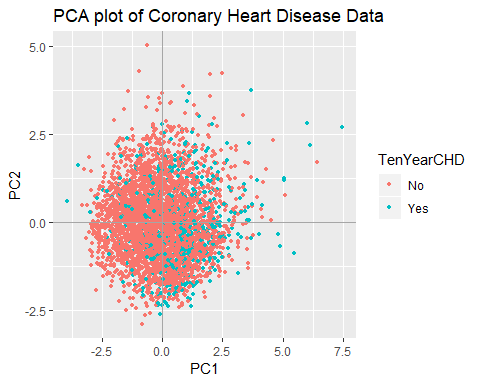
Let’s see if we can reduce this dataset down to a few principle components. Let’s also remove Systolic BP, since there is a lot of correlation between SysBP and DiaBP.

## PC1 PC2 PC3 PC4 PC5  
## age 0.4458879 -0.41179614 0.16397106 -0.25991763 -0.04778866  
## cigsPerDay -0.2126565 0.60962829 0.03634309 -0.37136454 -0.59766524  
## totChol 0.3945294 -0.06865090 0.17573516 -0.68123108 -0.05901458  
## diaBP 0.5262090 0.24519375 -0.32673903 0.09984649 -0.03126212  
## BMI 0.4705536 0.15313349 -0.43046647 0.31561046 -0.25964735  
## heartRate 0.2224436 0.60844262 0.31356081 -0.01355406 0.64857950  
## glucose 0.2291502 0.01709042 0.74197687 0.46977912 -0.38468214  
## PC6 PC7  
## age 0.66598065 0.3022420  
## cigsPerDay 0.27593052 0.1027459  
## totChol -0.57604231 -0.0967858  
## diaBP 0.23428163 -0.7002793  
## BMI -0.27040820 0.5741783  
## heartRate 0.08613164 0.2318152  
## glucose -0.11436887 -0.1221286

Looks like fairly even variance. It’s possible using PC’s won’t make much of a difference. Let’s check the Scree plots and clustering.



If we were extremely conservative and used 0.1 as a cutoff we would still likely have 7 components. Looks like our initial assumption that using PC’s is not the way to go might have been correct. Let’s do some clustering around the first few PC’s and confirm.

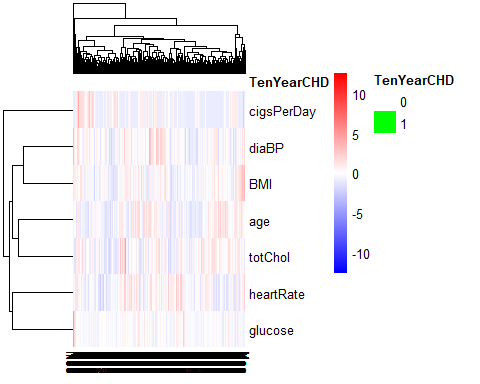


Separation isn’t great in any of these. We will not use PCA for this data. Since the continuous As this is a predictive model, it does not make pratical sense

## Random Forest Model

Let’s start by splitting our data into training and test data. We will use this split for all future tasks.

Let’s try some clustering.



Similar to the earlier heatmap we are seeing a lot of clustering around blood pressure. Now to create and train the model.

##   
## fit.pred No Yes  
## No 762 131  
## Yes 10 12

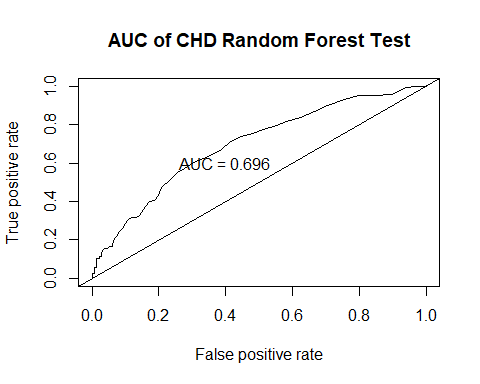
Lots of false positives. Let’s try an undersampled model.

##   
## fit.pred.under No Yes  
## No 512 55  
## Yes 260 88

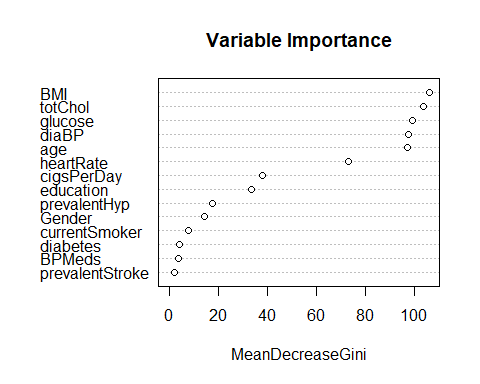
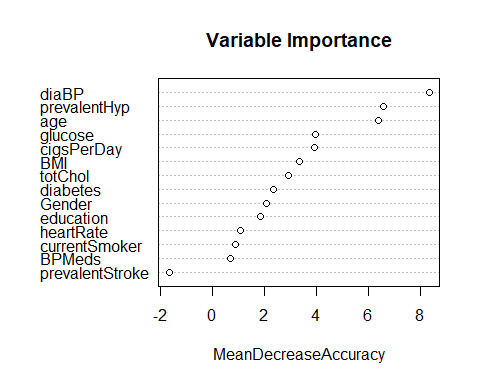
Now too many false negatives! Let’s try an oversampled model now.

##   
## fit.pred.over No Yes  
## No 737 129  
## Yes 35 14

Not much improvement. Not likely we will get much accuracy out of this model but let’s test anyway.



As expected, not a great accuracy score, but perhaps we can get some insight from the variable importance analysis.



## The following objects are masked from heartData (pos = 3):  
##   
## age, BMI, BPMeds, cigsPerDay, currentSmoker, diabetes, diaBP,  
## education, Gender, glucose, heartRate, prevalentHyp,  
## prevalentStroke, TenYearCHD, totChol

This affirms what we saw in model selection: blood pressure/hypertension, age, gender, and cigarette usage have the greatest impact on development of coronary heart disease.

## QDA

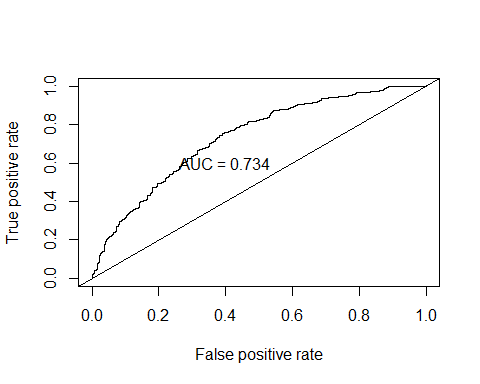
Given that several of the continuous variables were critical for model accuracy according to the EDA, let’s try some discriminant analysis. Since we same some non-normal distributions (lots of 0’s in cigarette usage), let’s use QDA so we can remain robust to violations in assumptions.

## heartData.train.y  
## No Yes  
## No 2253 357  
## Yes 76 57

Again it appears we are going to need to undersampled to account for the unbiased response

## heartData.train.under.y  
## No Yes  
## No 389 226  
## Yes 85 188

Still not very accurate but some improvement. Let’s try a test fit and see how accurate it is.

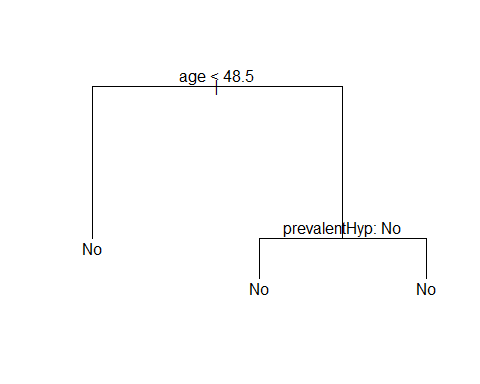


## heartData.train.under.y  
## No Yes  
## No 389 226  
## Yes 85 188

Pretty good results for a basic QDA model. Most of the inaccuracy seems to come from false positives.

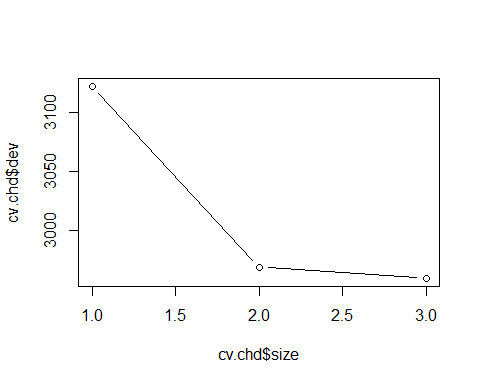
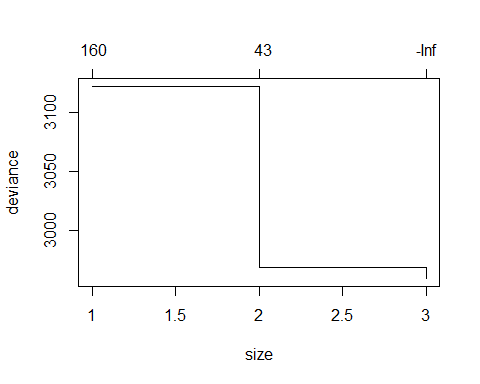
Great ROC score, and the confusion matrix looks balanced.

##   
## Classification tree:  
## tree(formula = TenYearCHD ~ ., data = heartData)  
## Variables actually used in tree construction:  
## [1] "age" "prevalentHyp"  
## Number of terminal nodes: 3   
## Residual mean deviance: 0.7979 = 2916 / 3655   
## Misclassification error rate: 0.1523 = 557 / 3658



## [1] "size" "dev" "k" "method"

## $size  
## [1] 3 2 1  
##   
## $dev  
## [1] 2959.655 2969.133 3122.424  
##   
## $k  
## [1] -Inf 42.53676 162.16706  
##   
## $method  
## [1] "deviance"  
##   
## attr(,"class")  
## [1] "prune" "tree.sequence"



## Warning in prune.tree(tree.deep, best = 8): best is bigger than tree size

