Heart Disease

Risk Analysis

By: Adrian Fils-Aime
Ajinkya Deshmukh
Sandhat Bylapudi
Silvia Chalkou

Introduction and Primary Purpose

Heart Disease is the leading cause of deaths in the US

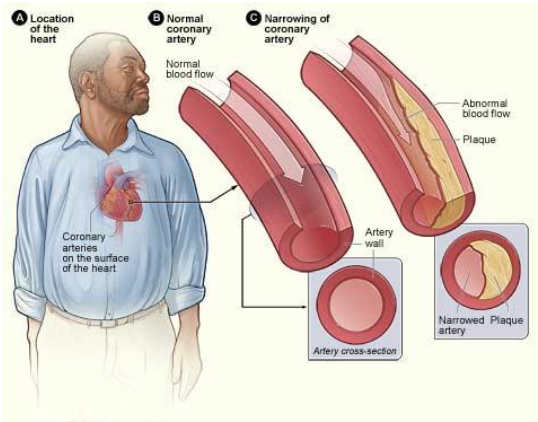
- Every **36 seconds** someone dies from heart disease
- **655,000 Americans die** each year
- World Health Organization predicts **12 million deaths** each year

How can heart risk be reduced?

1. **Do not smoke**
2. **Exercise often**
3. Maintain a **healthy body weight**
4. **Eat** a **healthy** diet

Two most common types of Heart Disease

1. Coronary Heart Disease
2. Heart Attack



Our Dataset

- A cardiovascular study on residents of Framingham, Massachusetts
- Source: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset
- There is over 4,000 Records and 15 Attributes
- 3749 observations after cleaning
- The dataset included the following variables:
    - Gender, age, current smoker, cigarettes smoked per day, Body mass index (BMI), systolic blood pressure, diastolic blood pressure, heart rate, glucose level , cholesterol level, diabetes, previous stroke

Preliminary Questions & Hypotheses

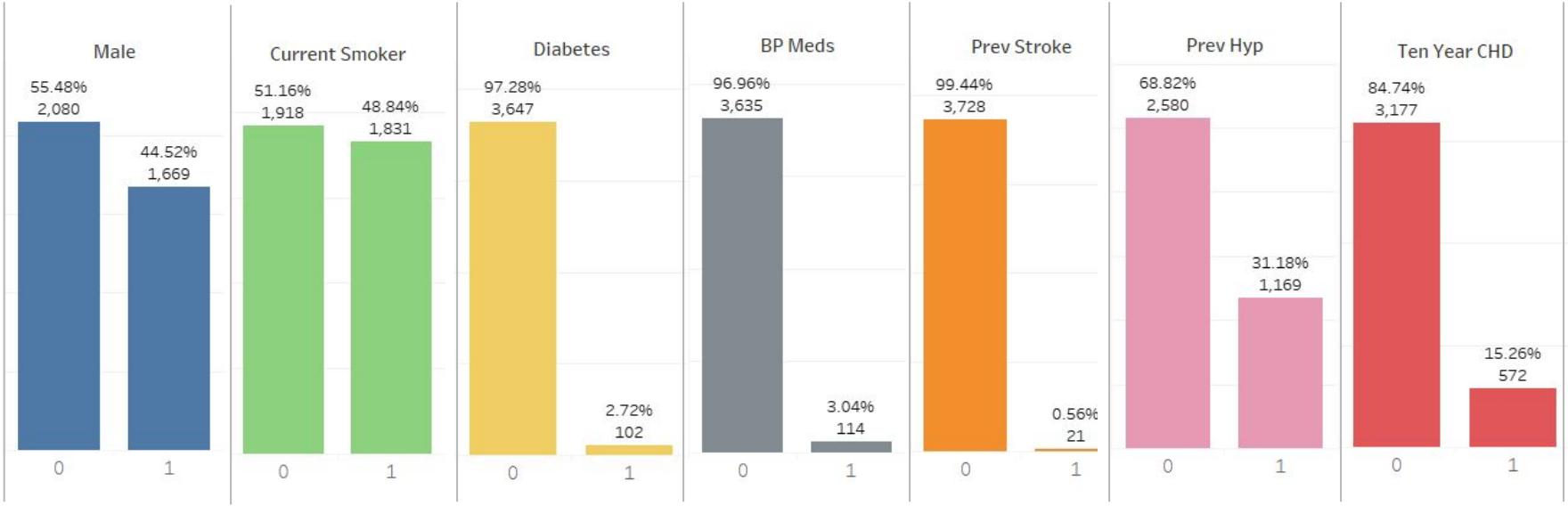
- What is the influence of various demographics on coronary heart disease ?
    - We assume that blood pressure, age, smoking, total cholesterol and BMI will have the most significant impact on ten-year risk of CHD.
- What influence do various demographics have on blood pressure?
    - We assume that age, BMI, smoking and total cholesterol will affect blood pressure.
- And finally, we intend to predict Systolic Blood Pressure and the ten-year risk of coronary heart disease (CHD).

Data Analysis Agenda

1. Exploratory Analysis
2. Logistic Regression and Classification Tree
3. Linear Regression and Regression Tree
4. Principal Component Analysis (PCA) and Principal Component Regression (PCR)

Exploratory Data Analysis

Categorical Variables



Exploratory Data Analysis

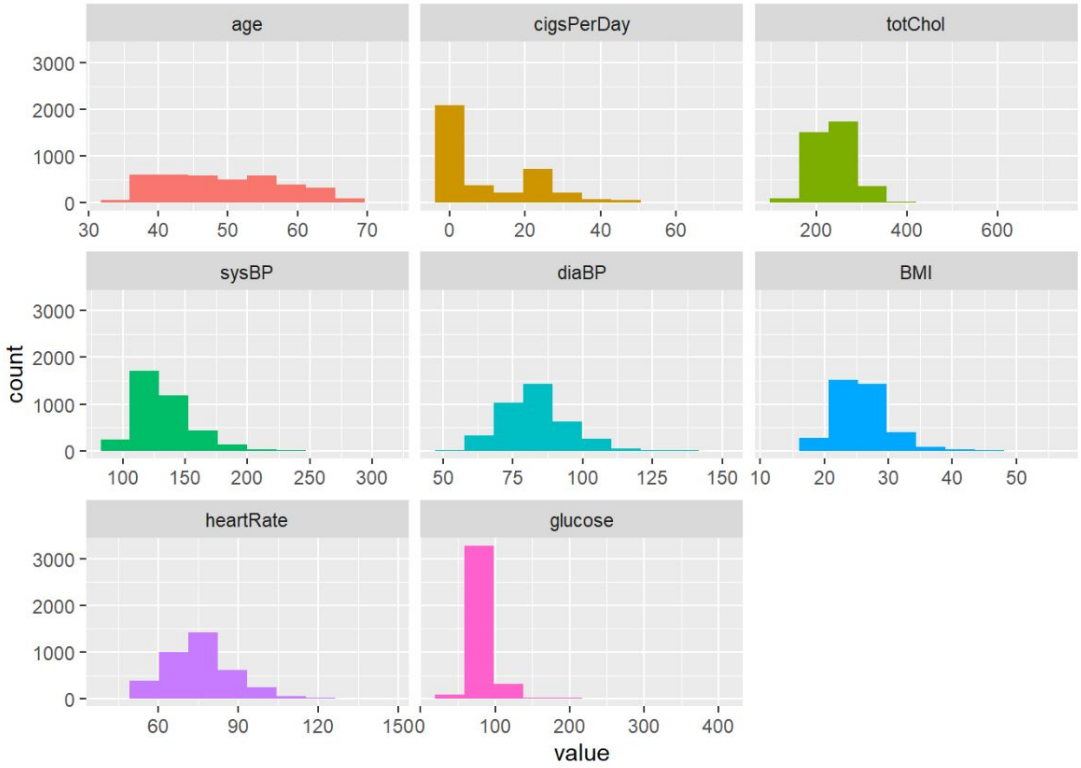
Numerical Variables

Age	Cigs_per_Day	Total_Cholesterol
Min. :32.00	Min. : 0.000	Min. :113
1st Qu.:42.00	1st Qu.: 0.000	1st Qu.:206
Median :49.00	Median : 0.000	Median :234
Mean :49.58	Mean : 9.005	Mean :237
3rd Qu.:56.00	3rd Qu.:20.000	3rd Qu.:264
Max. :70.00	Max. :70.000	Max. :696

Systolic_BP	Diastolic_BP	BMI
Min. : 83.5	Min. : 48.00	Min. :15.54
1st Qu.:117.0	1st Qu.: 75.00	1st Qu.:23.09
Median :128.0	Median : 82.00	Median :25.41
Mean :132.4	Mean : 82.93	Mean :25.81
3rd Qu.:144.0	3rd Qu.: 90.00	3rd Qu.:28.06
Max. :295.0	Max. :142.50	Max. :56.80

Heart_Rate	Glucose
Min. : 44.0	Min. : 40.00
1st Qu.: 68.0	1st Qu.: 71.00
Median : 75.0	Median : 78.00
Mean : 75.7	Mean : 81.88
3rd Qu.: 82.0	3rd Qu.: 87.00
Max. :143.0	Max. :394.00

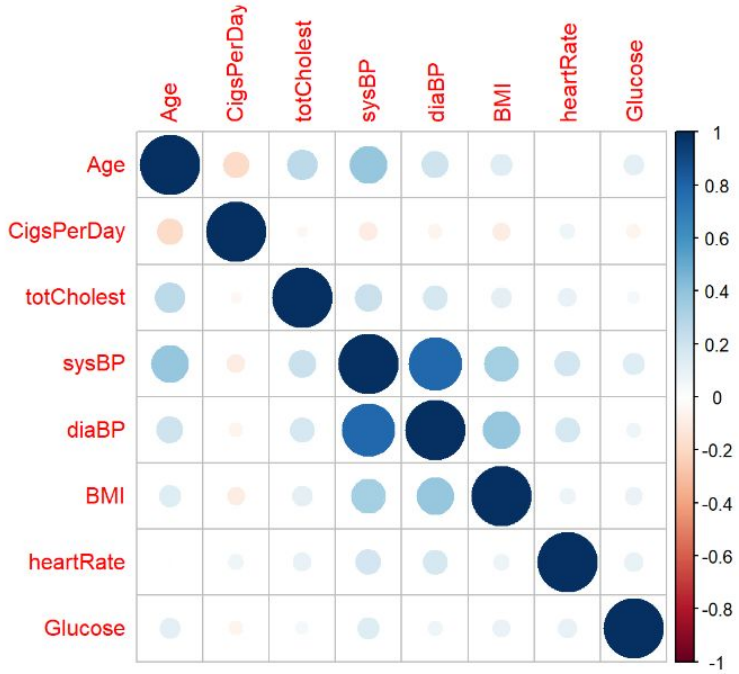
Histograms



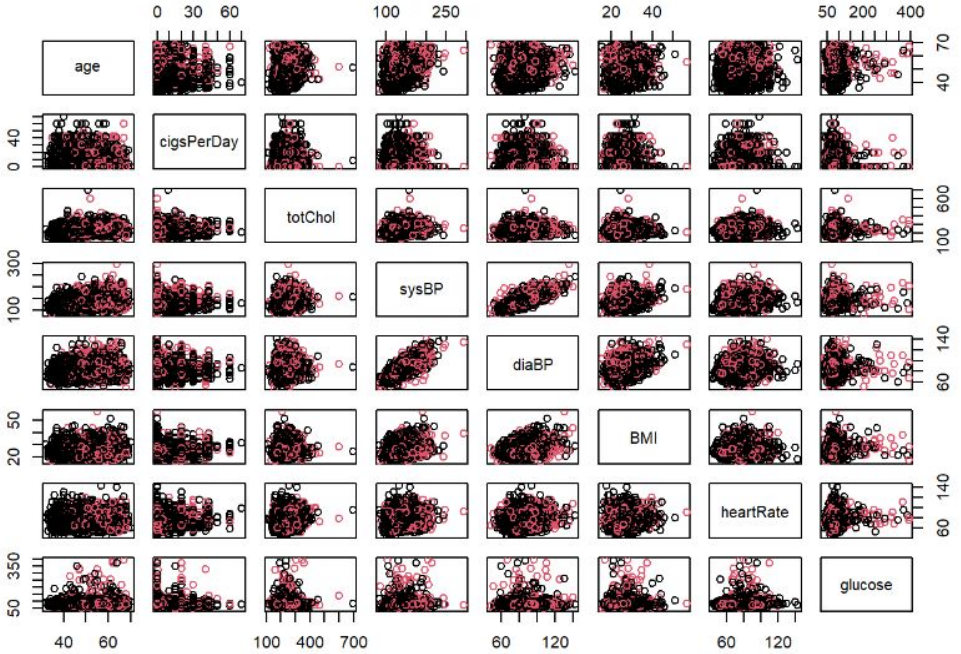
Exploratory Data Analysis

Numerical Variables

Correlation Matrix Plot

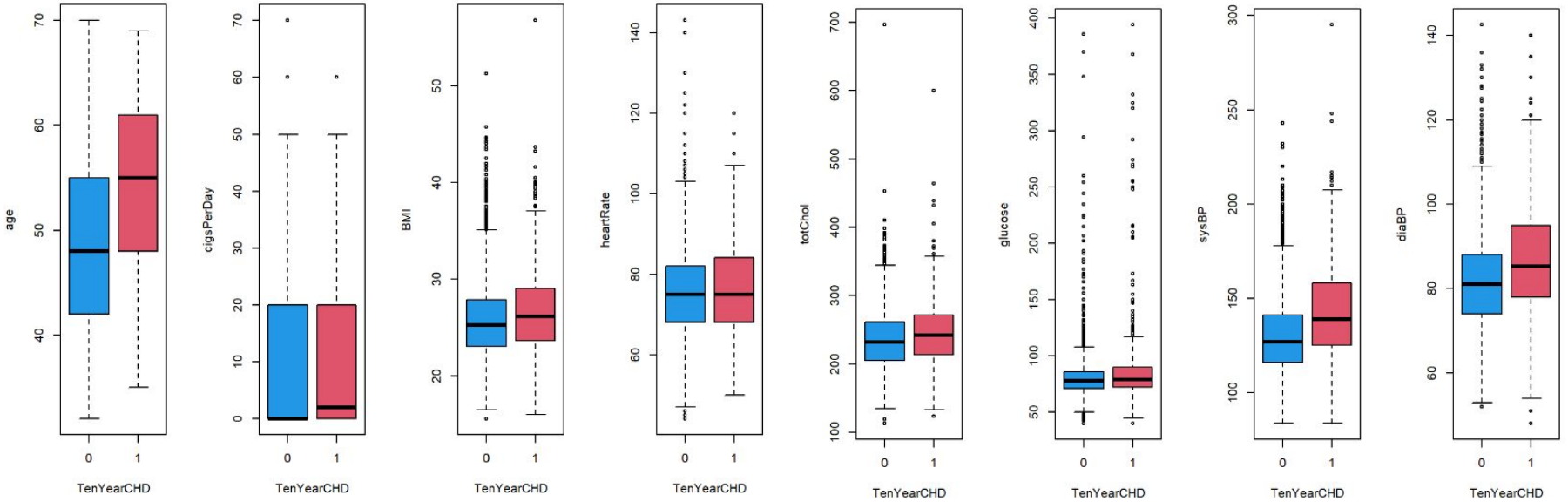


Scatterplots Matrix



EDA for Logistic Regression

Box Plots



TenYearCHD = Ten-year risk of Coronary Heart Disease

Logistic Regression to predict 10-year CHD

STEP 1: Select the best model by using Backward Addition AIC

- Selected predictors: **male** + **age** + **cigsPerDay** + **prevStroke** + **prevHyp** + **totChol** + **sysBP** + **glucose**

STEP 2: Fit a logistic regression model on the **train** (75%) subset:

- **Response**: TenYearCHD (binary: “1” = “Yes”, “0” =“No”)
- **Predictors**: All demographic variables selected by AIC*
  - Deviance = 2139.9
  - AIC = 2157.9
  - Significant predictors (alpha = 5%):
    male + **age** + **cigsPerDay** + **prevStroke** + ~~**prevHyp**~~ + **totChol** + **sysBP** + **glucose**

STEP 3: Predict on the **test** (25%) subset:

- Build the confusion matrix
  - **Accuracy** of prediction = 86.14%
  - **Test Error** = 13.86%

pred	0	1
0	792	119
1	11	16

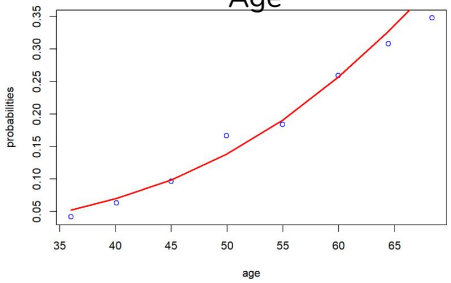
* BIC gives the same results

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.009410   0.600503 -15.003  < 2e-16 ***
male1        0.588799   0.121713   4.838 1.31e-06 ***
age          0.063265   0.007227   8.754  < 2e-16 ***
cigsPerDay   0.016361   0.004768   3.431 0.000601 ***
prevStroke1  1.171745   0.569384   2.058 0.039598 *
prevHyp1     0.208458   0.153636   1.357 0.174836
totChol      0.003206   0.001264   2.536 0.011207 *
sysBP        0.015333   0.003265   4.696 2.65e-06 ***
glucose      0.007781   0.002057   3.782 0.000156 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2429.1  on 2810  degrees of freedom
Residual deviance: 2139.9  on 2802  degrees of freedom
AIC: 2157.9

Number of Fisher Scoring iterations: 5
```

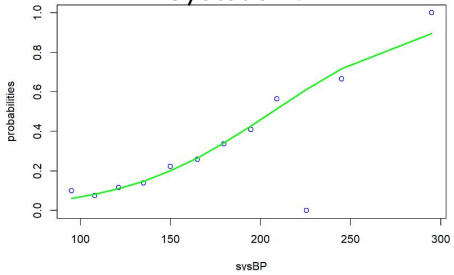
Visualizing Logistic Regression

Fitted and Observed Probabilities

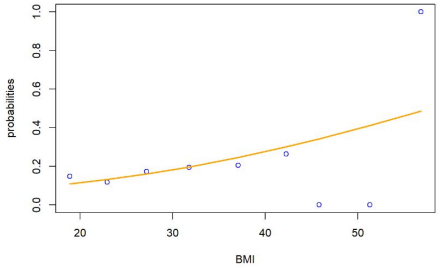
Age



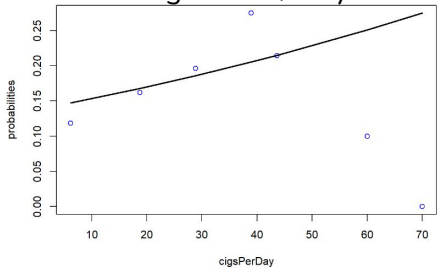
Systolic BP



BMI



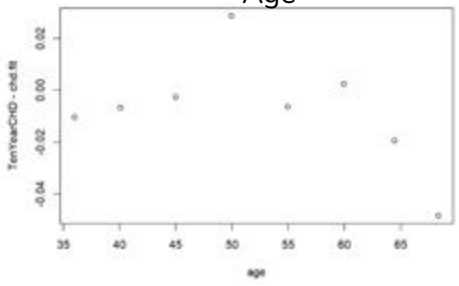
Cigarettes / Day



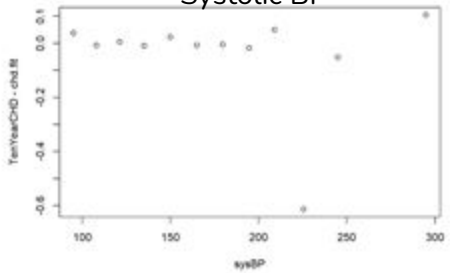
**Lines stands for fitted Probabilities, dots for observed Probabilities*

Residual Plots

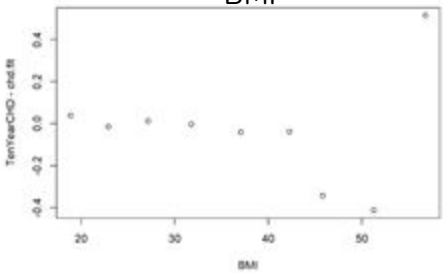
Age



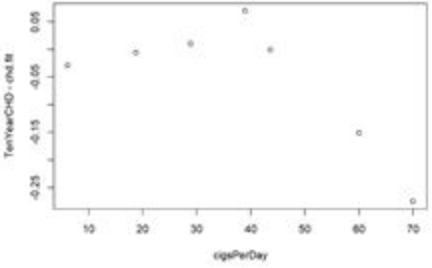
Systolic BP



BMI



Cigarettes / Day



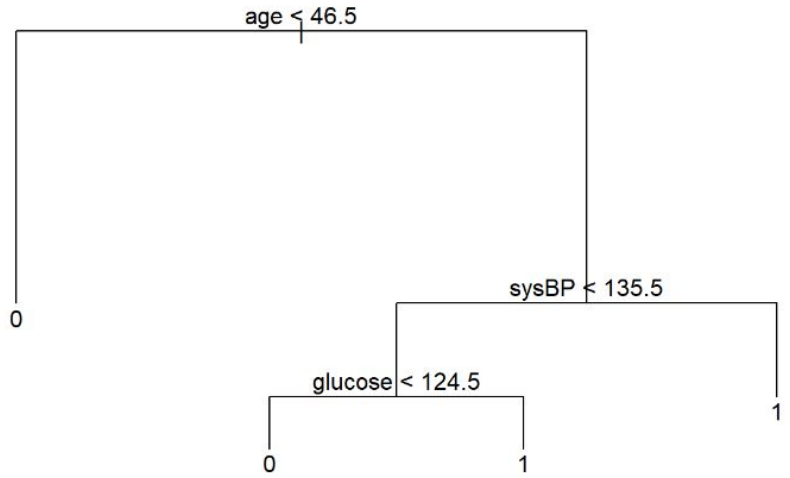
Classification Tree

STEP 1: Fit and plot the classification tree:

- **Response**: 10-year CHD
- **Predictors**: All the other variables
- **Splitting criterion**: Deviance
  - Number of terminal nodes: 4
  - Misclassification error rate: 0.3535

STEP 2: Cross Validation

STEP 3: Pruning



Sample:
50% with 1 for TenYearCHD and
50% with 0 for TenYearCHD

```
set.seed(123)
cv.fram <- cv.tree(tree.fram, FUN = prune.misclass, K = 5)
cv.fram
```

```
## $size
## [1] 4 2 1
##
## $dev
## [1] 427 437 584
```

Linear Regression : Model Selection

Model Selection Method: Manual

Response Variable: sysBP (Systolic Blood Pressure)

Manually Predictor Variables:

1. Gender (Male: 1, Female: 0)
2. Age
3. Current Smoker (Yes: 1, No: 0)
4. Cigarettes Per Day
5. Blood Pressure Medication (Yes: 1, No: 0)
6. Previous Stroke(Yes: 1, No: 0)
7. Previous Hypertension (Yes: 1, No: 0)
8. Diabetes (Yes: 1, No: 0)
9. Total Cholesterol
10. BMI
11. Heart Rate
12. Glucose

```
Call:
lm(formula = sysBP ~ . - TenYearCHD - diaBP, data = Fram)

Residuals:
    Min      1Q  Median      3Q     Max
-56.926  -9.450  -0.712   8.038 126.794

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    66.706672   2.907549  22.943  < 2e-16 ***
male1          -1.598066   0.526293  -3.036 0.002410 **
age             0.441959   0.031144  14.191  < 2e-16 ***
currentSmoker1 -0.957474   0.774032  -1.237 0.216166
cigsPerDay      0.034494   0.033331   1.035 0.300773
BPMeds1         9.825855   1.469352   6.687 2.61e-11 ***
prevStroke1     0.637571   3.255726   0.196 0.844753
prevHyp1       27.030063   0.590304  45.790  < 2e-16 ***
diabetes1      -1.595370   1.885618  -0.846 0.397566
totChol         0.022453   0.005674   3.957 7.73e-05 ***
BMI             0.614743   0.063678   9.654  < 2e-16 ***
heartRate       0.147949   0.020883   7.085 1.66e-12 ***
glucose         0.043251   0.012893   3.355 0.000803 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.76 on 3736 degrees of freedom
Multiple R-squared:  0.5535,    Adjusted R-squared:  0.5521
F-statistic: 385.9 on 12 and 3736 DF,  p-value: < 2.2e-16
```

Linear Regression : Model Selection

Model Selection Method: Stepwise AIC

Response Variable: sysBP (Systolic Blood Pressure)

Manually Predictor Variables:
1. Gender (Male: 1, Female: 0)
2. Age
3. Blood Pressure Medication (Yes: 1, No: 0)
4. Previous Hypertension (Yes: 1, No: 0)
5. Total Cholesterol
6. BMI
7. Heart Rate
8. Glucose

Predictor Variables Removed: Current Smoker, Cigarettes Per Day, Previous Stroke & Diabetes

```
Call:
lm(formula = sysBP ~ male + age + BPMeds + prevHyp + totChol +
    BMI + heartRate + glucose, data = Fram)

Residuals:
    Min      1Q  Median      3Q     Max
-56.849  -9.442  -0.694   8.021 126.822

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.768977   2.757347  24.215  < 2e-16 ***
male1       -1.532504   0.491970  -3.115 0.001853 **
age          0.443284   0.030582  14.495  < 2e-16 ***
BPMeds1      9.827971   1.462006   6.722 2.06e-11 ***
prevHyp1    27.042339   0.589827  45.848  < 2e-16 ***
totChol      0.022501   0.005667   3.970 7.31e-05 ***
BMI          0.621546   0.062800   9.897  < 2e-16 ***
heartRate    0.148022   0.020722   7.143 1.09e-12 ***
glucose      0.036532   0.010243   3.567 0.000366 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.76 on 3740 degrees of freedom
Multiple R-squared:  0.5532,    Adjusted R-squared:  0.5523
F-statistic: 578.9 on 8 and 3740 DF,  p-value: < 2.2e-16
```

Linear Regression : Predictions

Mean Squared Error -

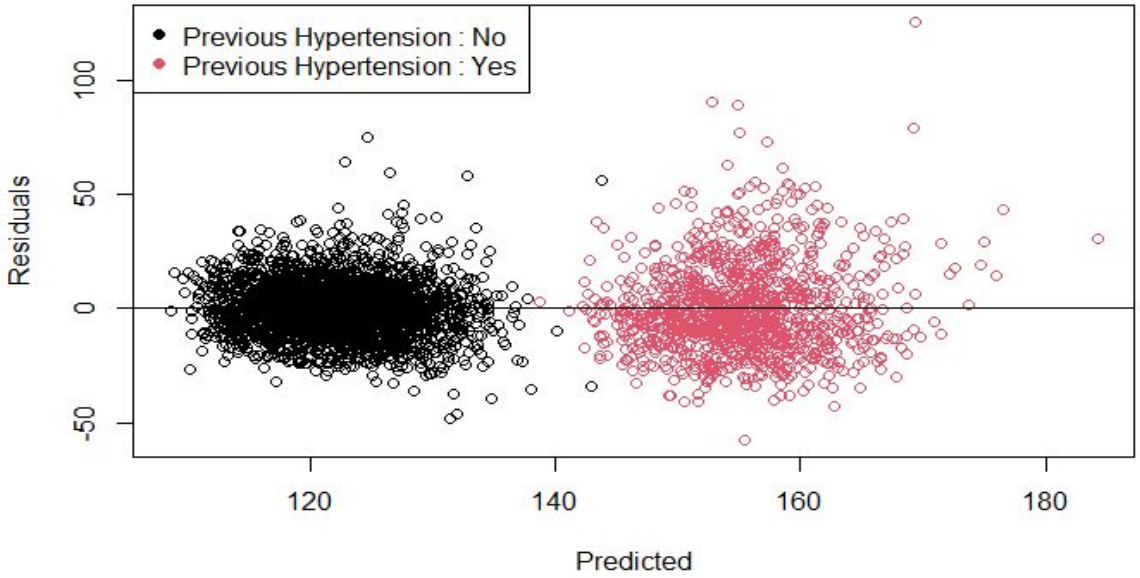
Full model - 211.9205

AIC model - 207.7618

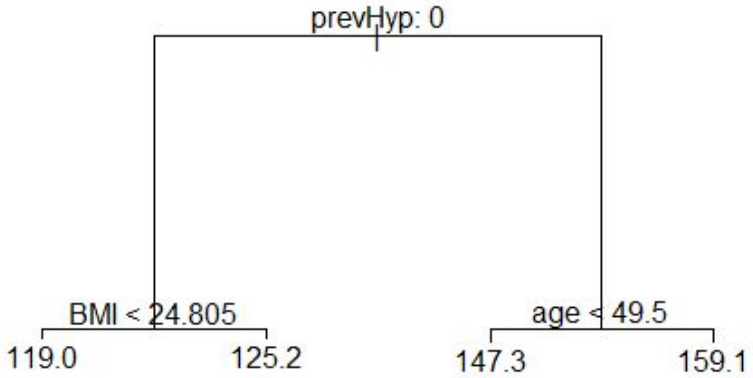
Predictions on random sample of 4 observations -

male	age	prevHyp	totChol	BMI	heartRate	glucose	sysBP (Observed)	sysBP(Predicted)
1	61	1	256	24.12	75	97	165	154.53
1	50	1	255	25.35	90	74	126	153.52
0	36	0	180	29.59	75	84	118	118.29
1	37	1	195	25.66	83	117	141	146.56

Residual Plot



Regression Tree




```
Regression tree:
tree(formula = full.model, data = Fram, subset = train)
Variables actually used in tree construction:
[1] "prevHyp" "BMI"     "age"
Number of terminal nodes:  4
Residual mean deviance:  233.9 = 656500 / 2807
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-61.110 -10.110  -1.113   0.000   8.522  95.660
```

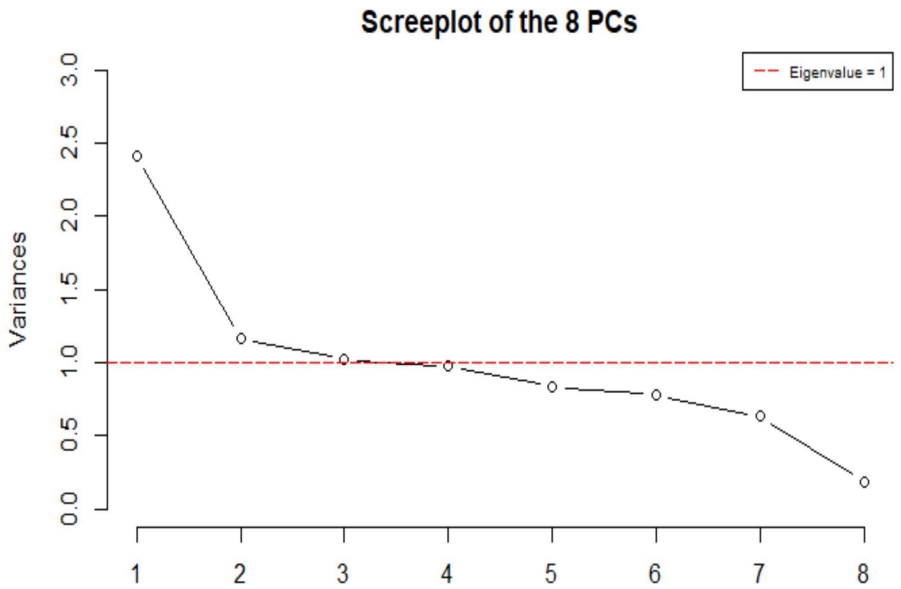
Mean Squared Error (Test)- 237.5884

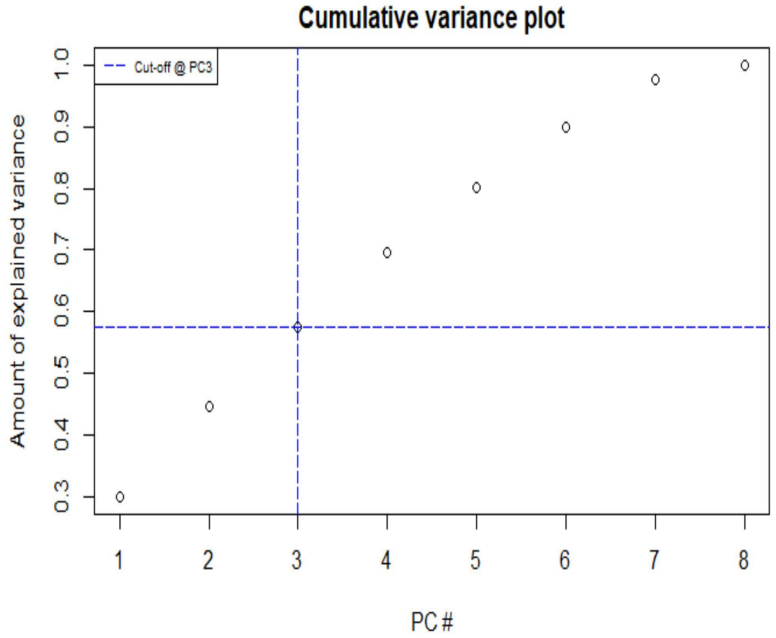
Principal Component Analysis

We notice that **the first 3 components** have an **Eigenvalue >1** and combinedly explain around **57% of variance.** (Eigenvalue here represents both Standard Deviation and Variance)

An **Eigenvalue<1** would mean that the component actually explains less than a single explanatory variable and we would like to discard those.

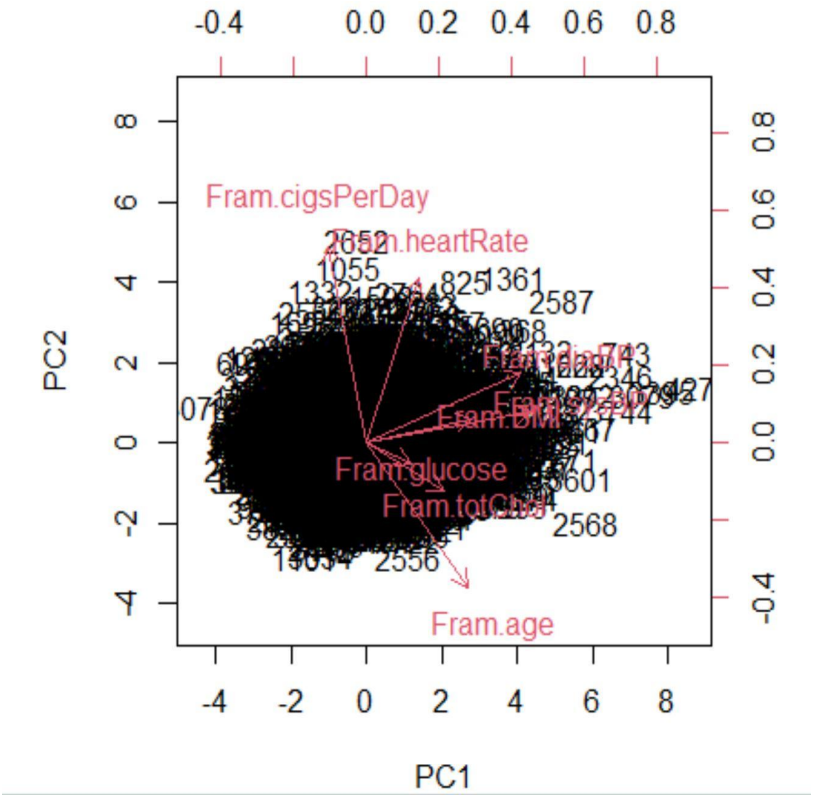
```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8
Standard deviation     1.5516 1.0785 1.0112 0.9876 0.9135 0.88321 0.79436 0.43133
Proportion of Variance 0.3009 0.1454 0.1278 0.1219 0.1043 0.09751 0.07888 0.02326
Cumulative Proportion  0.3009 0.4463 0.5741 0.6960 0.8004 0.89787 0.97674 1.00000
```

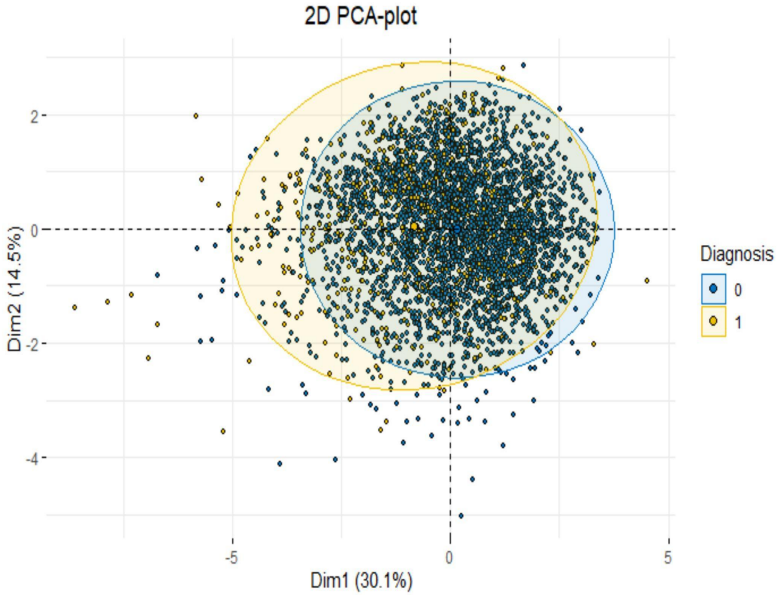

Screeplot of the 8 PCs
Eigenvalue = 1
Variances
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1
2
3
4
5
6
7
8


Cumulative variance plot
Cut-off @ PC3
Amount of explained variance
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
2
3
4
5
6
7
8
PC #

```
> pr.out$rotation
                    PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8
Fram.age      0.3483543 -0.46808701 -0.16782879  0.24772953 -0.09456600 -0.336624367 -0.6451368 -0.177615500
Fram.cigsPerDay -0.1314999  0.63513638 -0.06115451  0.34596336 -0.62370192 -0.109580918 -0.2342353 -0.000880487
Fram.totChol  0.2667795 -0.15551075 -0.31711325  0.68232534 -0.02287316  0.448094939  0.3704059  0.001914912
Fram.sysBP    0.5656778  0.09862733  0.14419223 -0.01503757 -0.04898593 -0.322599264  0.1686621  0.717179861
Fram.diaBP    0.5321469  0.22641409  0.28063398 -0.06940721 -0.03022298 -0.189909842  0.3112349 -0.669430505
Fram.BMI      0.3617057  0.07564255  0.25916033 -0.26828940 -0.14945375  0.730466070 -0.4058815  0.060390578
Fram.heartRate  0.1789360  0.53100251 -0.44218060  0.00759733  0.64874207  0.003992231 -0.2631773 -0.017098964
Fram.glucose  0.1554509 -0.06809255 -0.71084332 -0.52563405 -0.39371850  0.012423083  0.1800436 -0.044997135
```

```
> head(pr.out$x)
            PC1        PC2        PC3        PC4         PC5         PC6        PC7         PC8
[1,]  1.6989678 -0.1073280 -0.1354449  1.08199538 -0.971582204 -0.87583097  0.1611673 -0.10648890
[2,] -0.1643343 -0.5107193  0.4485884  0.19720452 -1.570533962 -1.07949935  0.3432320  0.15890198
[3,]  0.5226138 -0.5603516 -0.1662180 -0.70888541  0.368700805 -0.06887914  0.1712490 -0.05499442
[4,] -1.3753648 -0.3600646 -0.0966786 -0.02010806  2.318534152  0.70655323  0.8037954  0.32550262
[5,]  0.1073137 -1.1366856  0.9428487 -1.14905902  0.162124865 -0.03013400 -0.4925264  0.12096607
[6,] -2.7374830 -0.7297138 -0.9155957  1.45045673  0.002930809  0.06024944 -1.3203461 -0.20005315
```


Fram.cigsPerDay
Fram.heartRate
Fram.glucose
Fram.age
PC1
PC2


2D PCA-plot
Dim1 (30.1%)
Dim2 (14.5%)
Diagnosis
0
1

Conclusion & Insight

Conclusions from Logistic Regression:

1. Older people are more likely to have a CHD in ten years if all the other features remain constant.
2. Males are more likely to have a CHD than women if all the other variables remain the same
3. People with previous stroke are more likely to have a CHD in ten years.
4. People who smoke more have a higher chance of CHD in ten years
5. People with higher systolic BP or higher glucose or higher total cholesterol have a higher CHD risk in ten years.

Our initial assumption was that the BMI affects the CHD risk, but it was not chosen by our model. We assume that the inclusion of the other variables such as sysBP, age, gender, glucose, and totChol has made up for the impact of BMI.

The linear regression model to predict the systolic BP proved that systolic BP is positively influenced by hypertension occurrence in the past (prevHyp), total cholesterol, heart rate, glucose, BMI and age.

Conclusion & Insight

Based on linear regression as well as the regression tree, hypertension history affects the systolic BP the most. According to the regression tree, if the residents have no previous hypertension then BMI influences their systolic BP and if there the residents have previous hypertension then their age affects their systolic BP.

Data is not suitable for conducting PCA Analysis!

Insights:

As you grow older the risk of CHD goes up. Take more care of your health and get regular checkups done to avoid the risk of CHD or to be able to detect it before it gets fatal. Maintain a healthy lifestyle. Exercise regularly, do not over indulge in activities such as smoking, eating fast food and sugary food that will affect your health negatively.

Males seem to be more susceptible to CHD as they age and hence need to take more care.

Thank you

Questions ?