Exploratory Data Analysis

The data set is called “heart” and was obtain from the following website https://web.stanford.edu/~hastie/ElemStatLearn//datasets/SAheart.data

The data set contains 462 observations and 10 variables. The response is “chd”, the presence (chd=1) or absence (chd=0) of coronary heart disease. The list of variables is shown in table 1.

Table 1. Variables in the Heart DataSet

Name	Description	Type	Missing (%)
sbp	Systolic Blood Pressure	Integer	0
tobacco	Cumulative tobacco (kg)	Numeric	0
ldl	Low Density Lipoprotein Cholesterol	Numeric	0
adiposity	Measure in fat percentage	Numeric	0
famhist	Family history of heart disease (Present, Absent)	Factor	0
typea	Type-A behavior	Integer	0
obesity	Measure in bmi	Numeric	0
alcohol	Current Alcohol Consumption	Numeric	0
age	Age at onset	Integer	0
chd	The presence (chd=1) or absence (chd=0) of Coronary Heart Disease.	Factor	0

The corresponding mean and standard deviation of the quantitative continuous variables in the data are shown in table 2.

Table 2. Continuous Variables – Data Description

Variable	Mean	Standard Deviation
sbp	138.33	20.49
tobacco	3.64	4.59
ldl	4.74	2.07
adiposity	25.41	7.78
typea	53.10	9.82
obesity	26.04	4.21
alcohol	17.04	24.48
age	42.82	14.61

In the Data exploration analysis is shown that the median Age of those who have a Coronary Heart Disease (chd=1) is higher than in those who do not have the condition (chd=0). Moreover, can be noted a more symmetrical distribution and higher Interquartile Range in the group with no coronary heart disease (chd=0) than on those with the condition (chd=1) Fig.1.

The distribution of Alcohol between the two groups of individuals with Coronary Heart Disease (chd=1) and individuals with no Coronary Heart Disease (chd=0). looks almost identical (Fig 8). Similarly, no changes can be notice in the distribution of obesity between the two groups (Fig 7). The median consumption of Tobacco in the group with Coronary Heart Disease (chd=1) is slightly higher than in those with no Coronary Heart Disease (chd=0) (Fig. 2). Similarly, the median ldl, typea behavior, adiposity and sbp is slightly higher than those without the condition (chd=0) (Fig 3-6).

Figure 1. Distribution of Age by Coronary Heart Disease Status

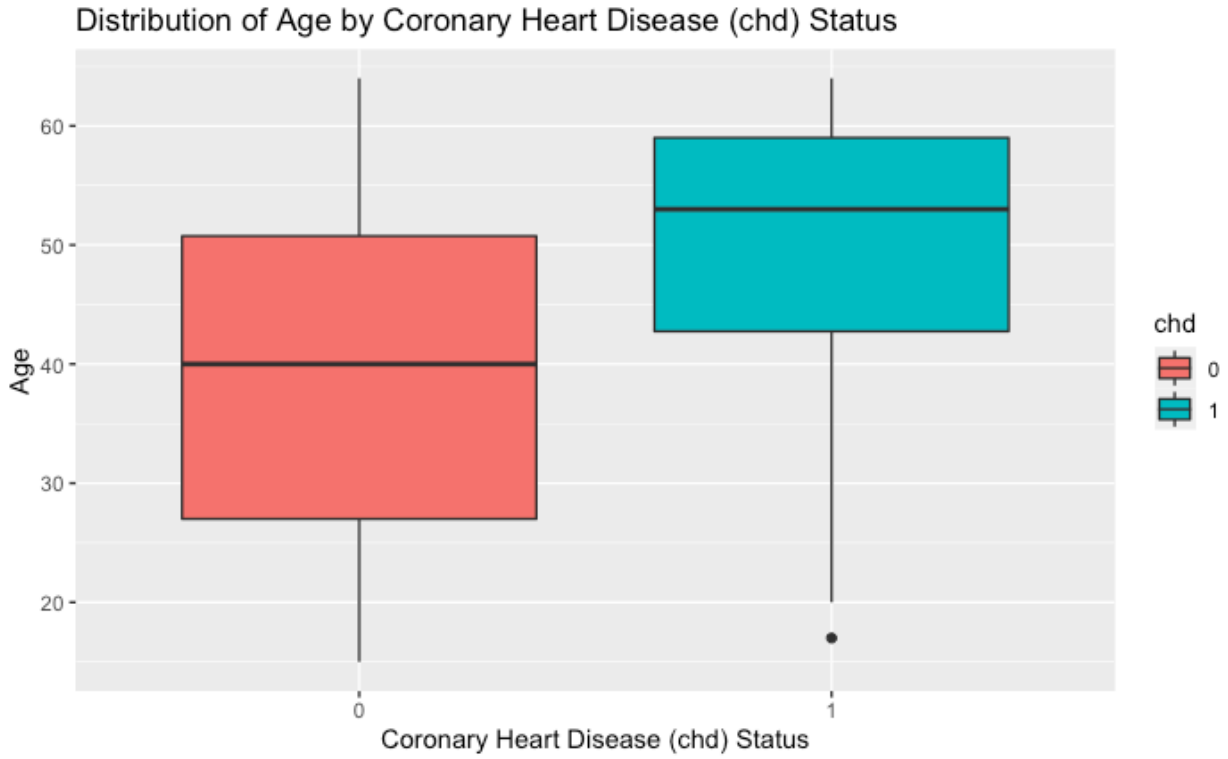


Figure 2. Tobacco by CHD Status

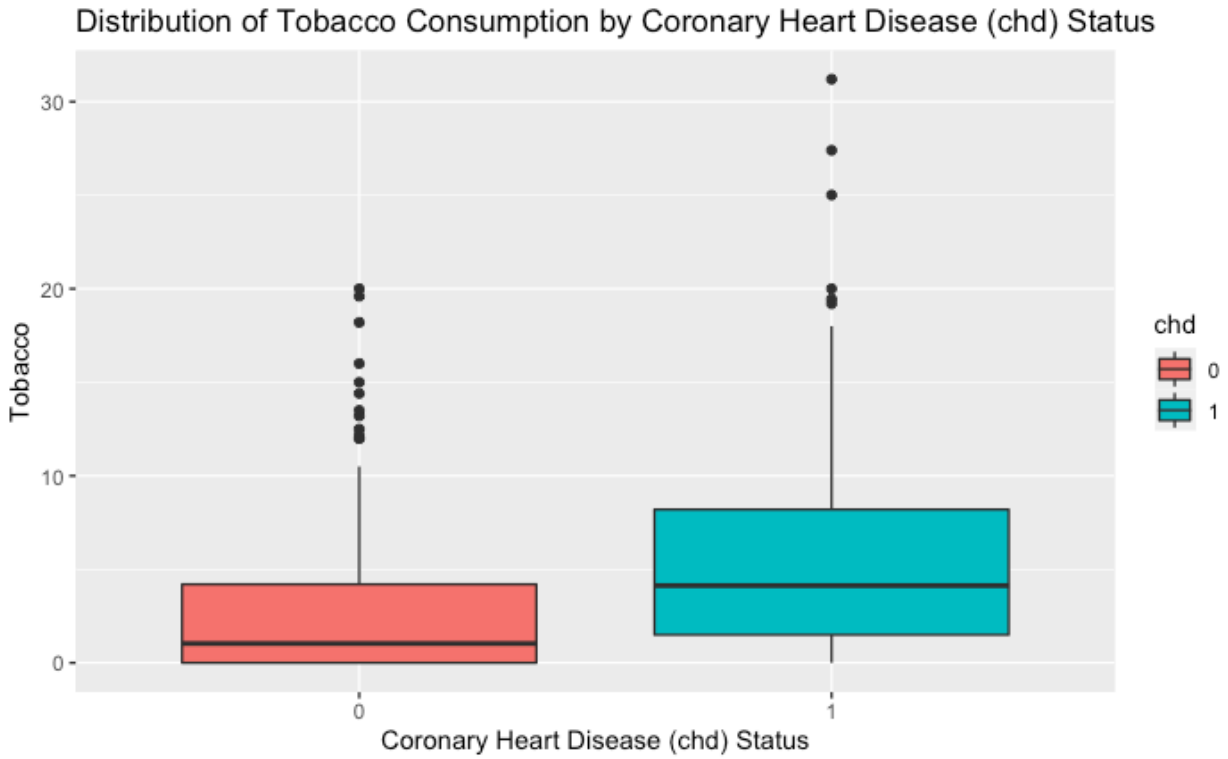


Figure 3.LDL Cholesterol by CHD Status

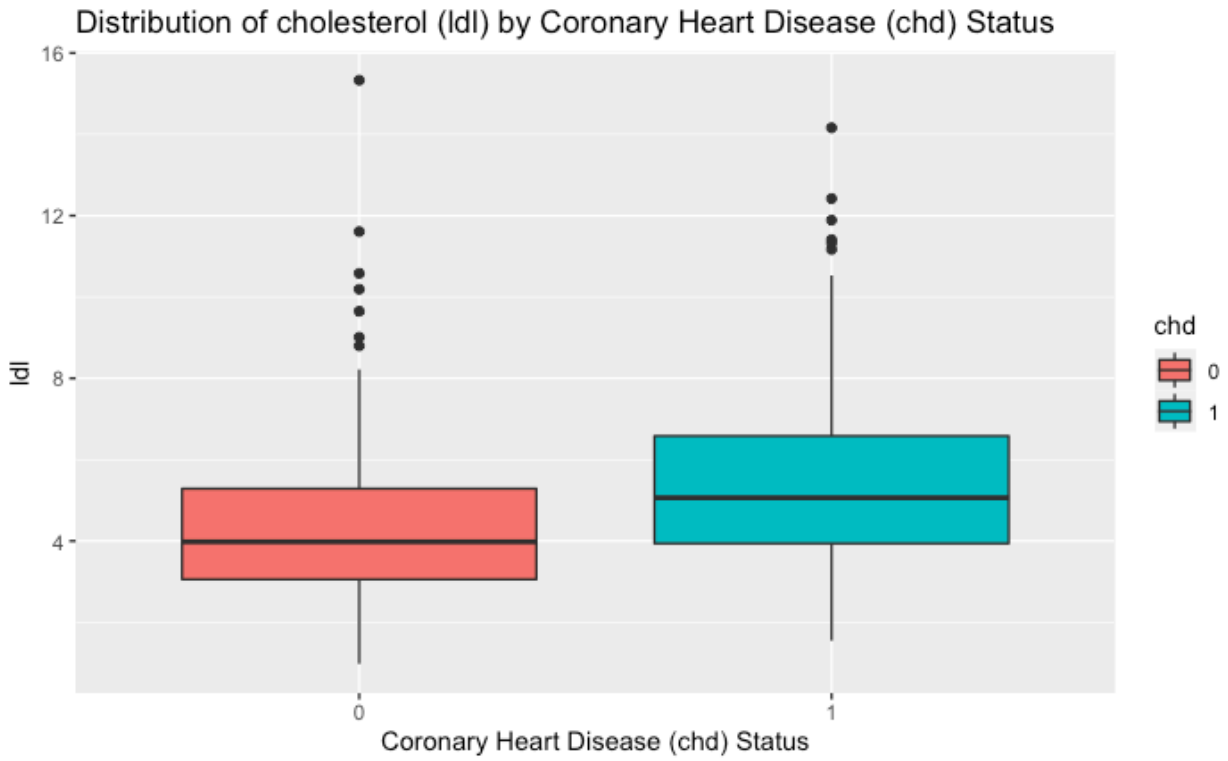


Figure 4.Adiposity by CHD Status

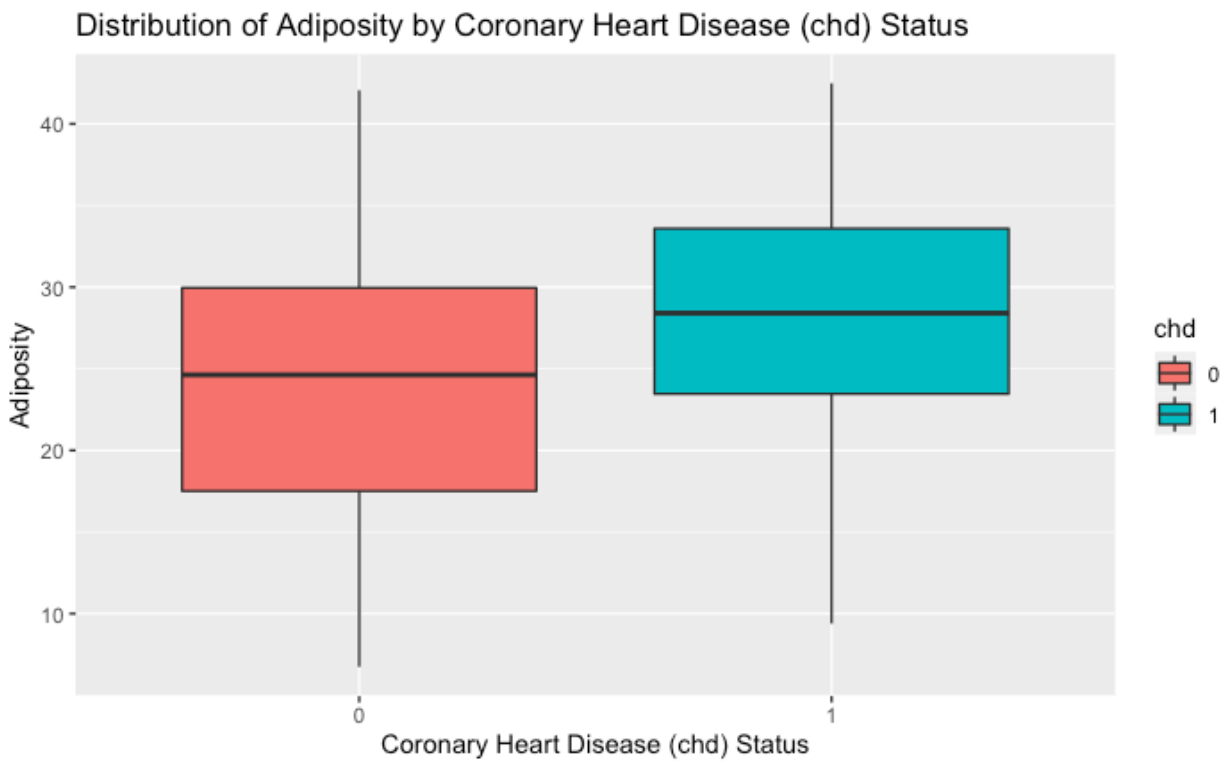


Figure 5. Type a Behavior by CHD Status

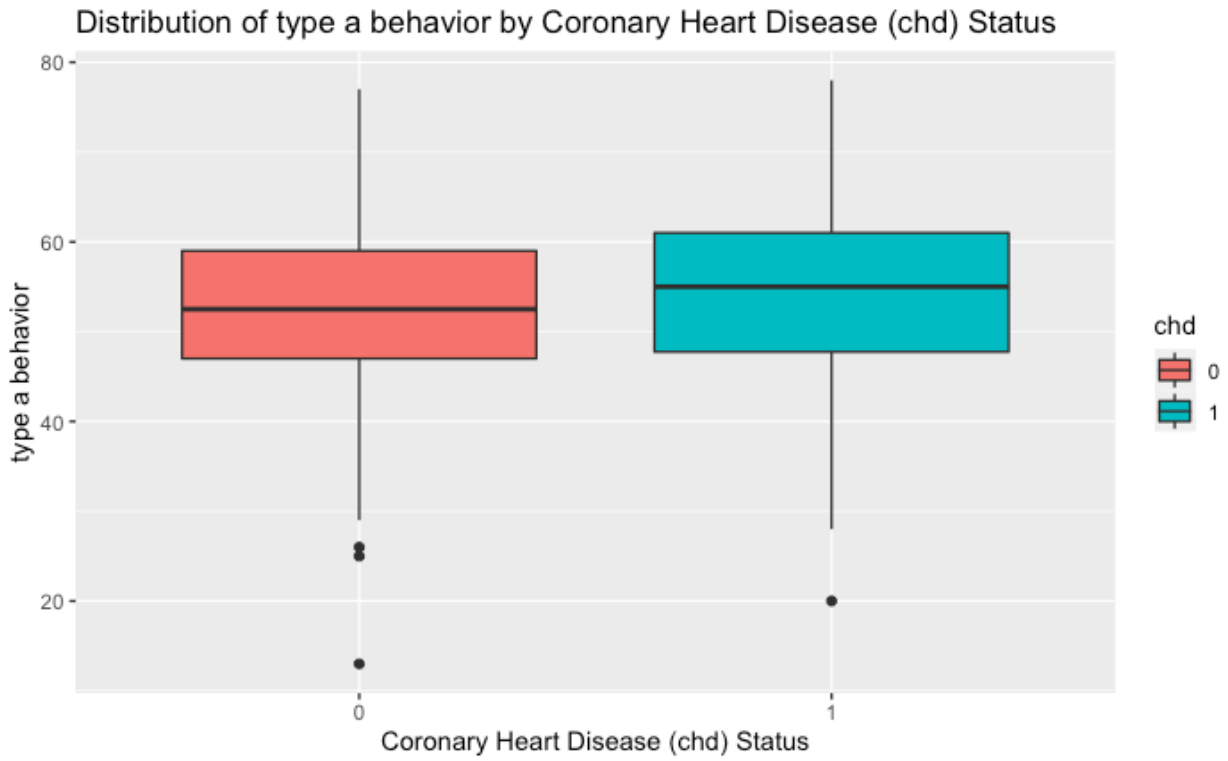


Figure 6. SBP by CHD Status

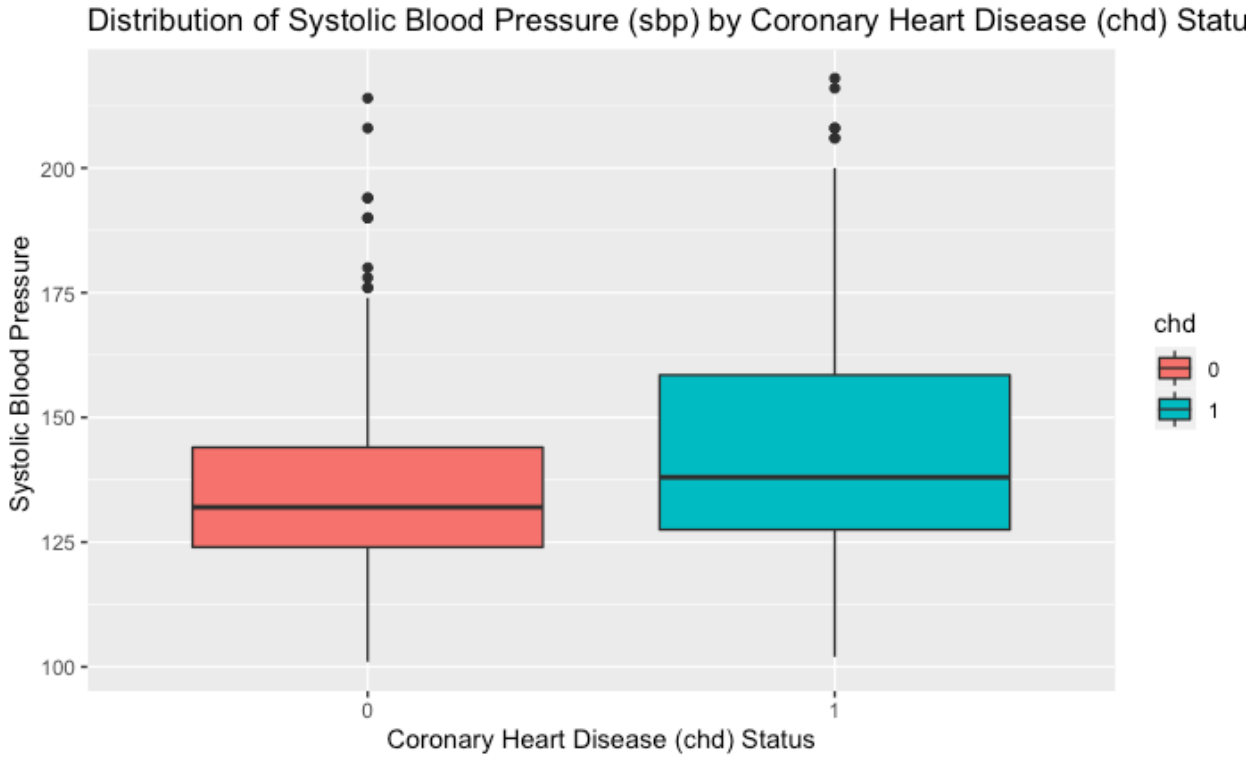


Figure 7 Obesity by CHD Status

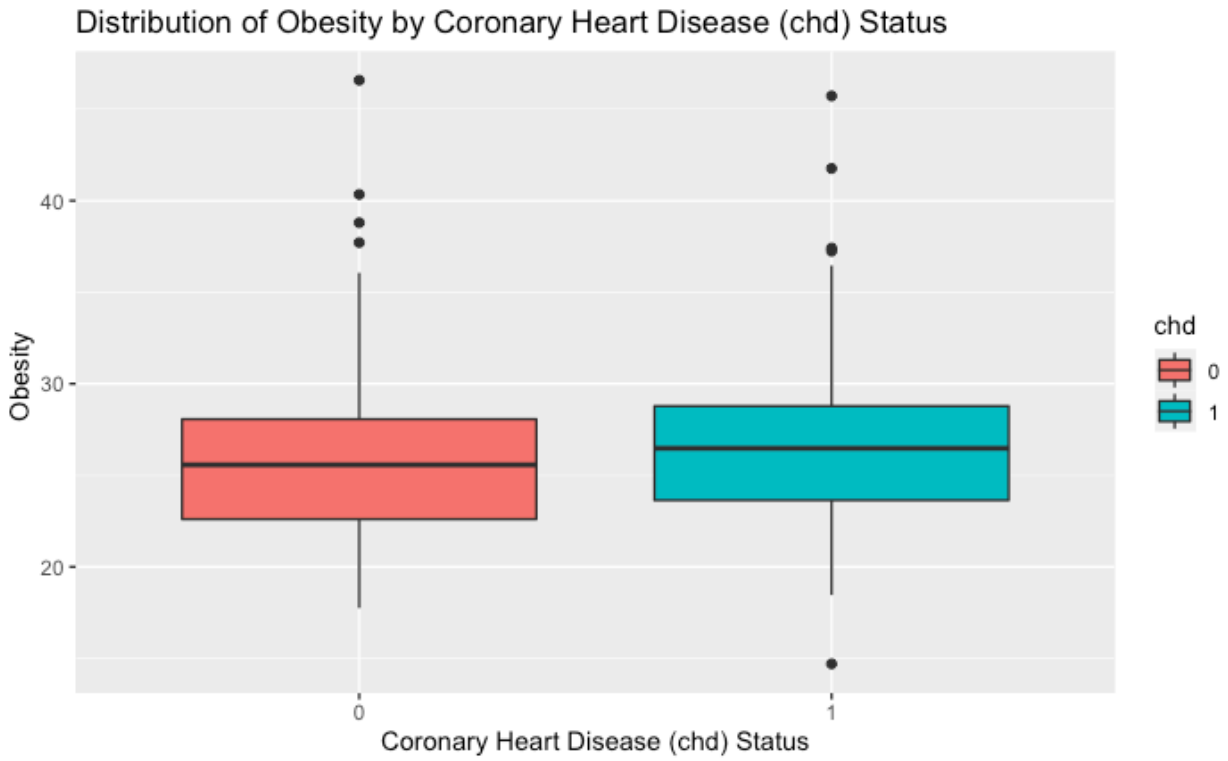
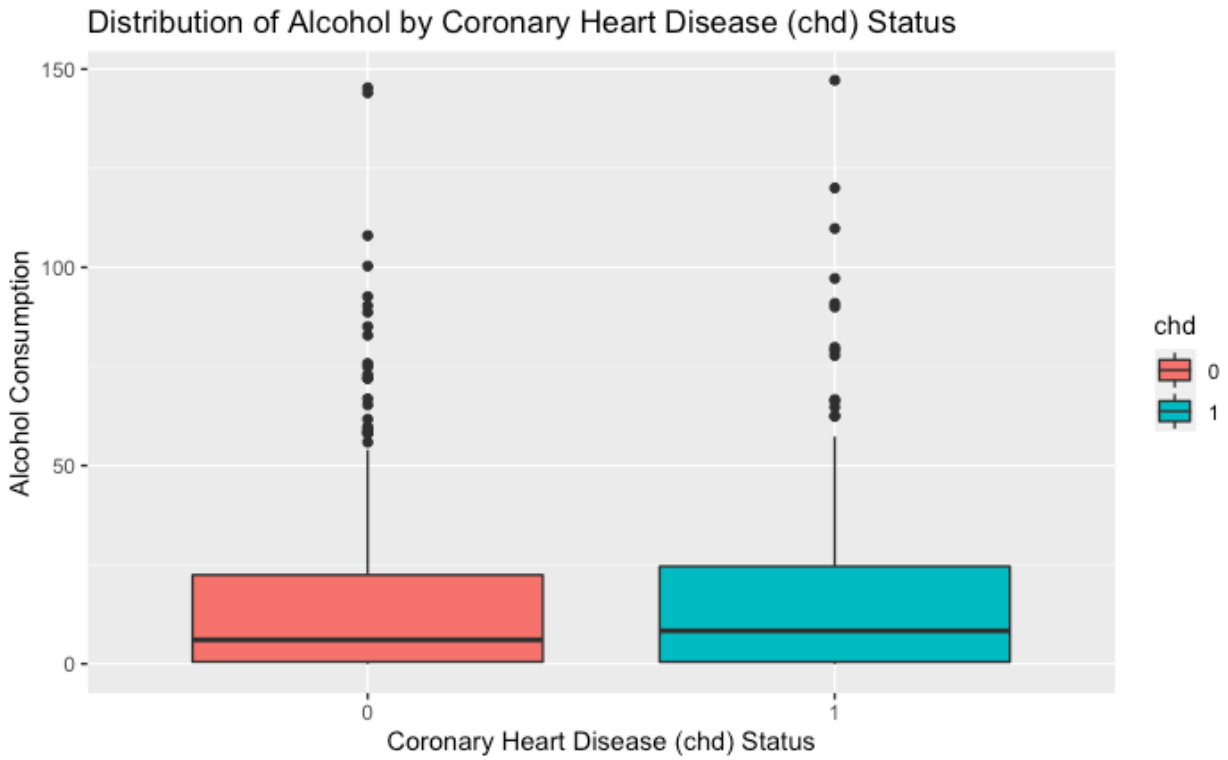


Figure 8. Alcohol by CHD Status

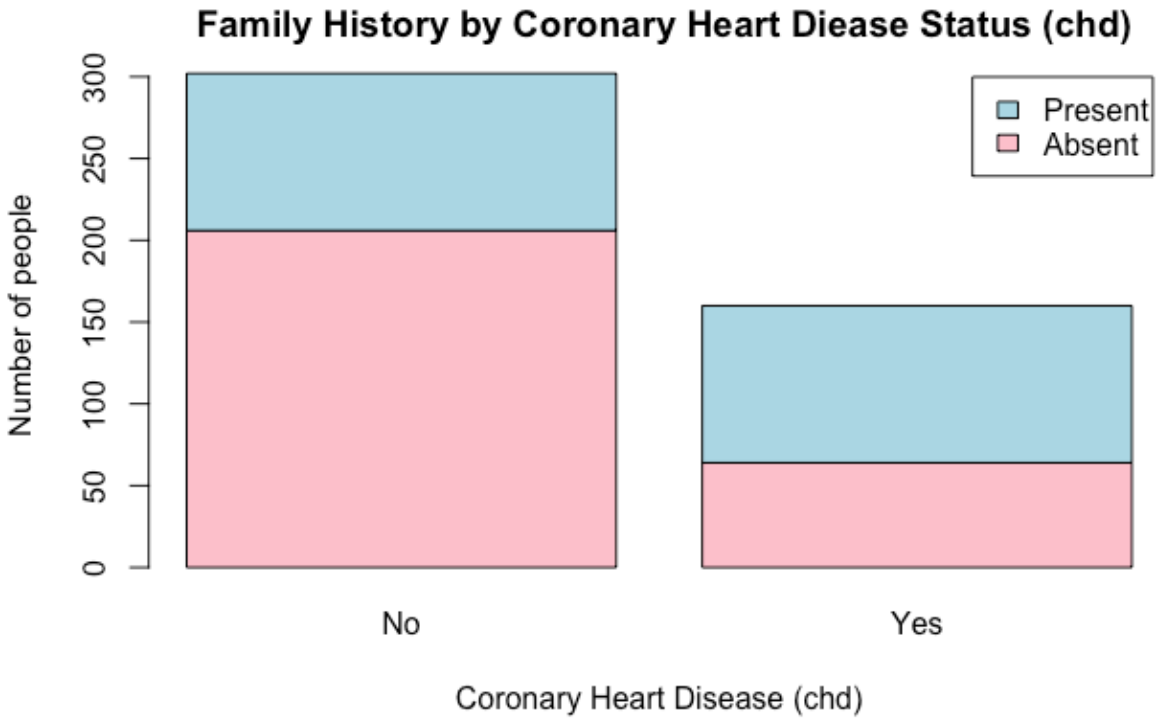


The exploration of the categorical variable Family History of heart disease is summarized in table 3. The proportion of a present family history on individuals with coronary heart disease is higher than in the individuals without the condition fig 9.

Table 3. Family History of Heart Disease vs Coronary Heart Disease Status

	Coronary Heart Disease (chd)		
Fam History	No	Yes	Totals
Absent	206	64	270
Present	96	96	192
Totals	302	160	462

Figure 9. Family History by Coronary Heart Disease Status (chd)



Setting up a training and testing data set

A training data set was created called" train" which contain 75% of the observation from the original data set. The rest of the data was allocated for a test data set "Test".

```
set.seed(3456)
trainIndex <- createDataPartition(heart$chd, p = .75, list = FALSE)


Train <- heart[ trainIndex,]
Test <- heart[ -trainIndex,]
```

Full logistic regression model

A full logistic regression model was built which includes all the predictors and fit using the training data.

```
mylogitfull <- glm (chd ~., data = Train, family = "binomial")
summary(mylogitfull)


Call:
glm(formula = chd ~ ., family = "binomial", data = Train)

Deviance Residuals:
  Min       1Q   Median       3Q      Max
 -1.7928  -0.7847  -0.4213   0.8599   2.5008

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
 (Intercept)    -5.148747   1.480688  -3.477 0.000507 ***
sbp              0.003835   0.006449   0.595 0.552105
tobacco          0.082248   0.030124   2.730 0.006326 **
ldl              0.204374   0.075954   2.691 0.007129 **
adiposity        0.031659   0.034743   0.911 0.362178
famhistPresent   0.902090   0.269484   3.347 0.000816 ***
typea            0.040143   0.014145   2.838 0.004542 **
obesity         -0.109172   0.054591  -2.000 0.045522 *
alcohol         -0.001097   0.005377  -0.204 0.838374
age              0.046491   0.014065   3.306 0.000948 ***
```

```
     Null deviance: 447.51 on 346 degrees of freedom
  Residual deviance: 346.37  on 337  degrees of freedom
AIC: 366.37
```

Stepwise variable selection

Select the most contributive variables:

```
#Stepwise model selection

library (MASS)
library(dplyr)

step.model<-mylogitfull %>% stepAIC(trace=FALSE)
coef(step.model)
```

```
    (Intercept)   tobacco            ldl famhistPresent          typea
    -5.20391216     0.08171878     0.22508537     0.89113868     0.03885501
        obesity            age
   -0.07082302     0.05464758
```

```
summary(step.model)
```

```
 Call:
 glm(formula = chd ~ tobacco + ldl + famhist + typea + obesity +
    age, family = "binomial", data = Train)

 Deviance Residuals:
    Min       1Q   Median       3Q      Max
 -1.7810  -0.7714  -0.4065   0.8623   2.5230

 Coefficients:
                Estimate Std. Error z value Pr(>|z|)
 (Intercept)    -5.20391    1.24004  -4.197 2.71e-05 ***
 tobacco         0.08172    0.02943   2.777 0.005485 **
 ldl             0.22509    0.07310   3.079 0.002075 **
 famhistPresent  0.89114    0.26825   3.322 0.000893 ***
 typea           0.03886    0.01407   2.762 0.005744 **
 obesity        -0.07082    0.03577  -1.980 0.047686 *
 age             0.05465    0.01204   4.541 5.61e-06 ***

     Null deviance: 447.51  on 346  degrees of freedom
 Residual deviance: 347.60  on 340  degrees of freedom
AIC: 361.6
```

After stepwise variable selection the reduced model named “step. model” removed the predictors adiposity and alcohol from the full model. Both models were used to predict the response on the test data set.

Full vs Reduced model

The performance of the full and the stepwise (reduced) logistic model were made. The best model is defined as the model that has the lowest classification error rate in predicting the class of new test data. The results are summarized in table 4.

```
#Prediction accuracy of the full logistic regression model:

# Make predictions
probabilities <- mylogitfull %>% predict(Test, type = "response")

predicted.classes <- ifelse(probabilities > 0.5, "1", "0")


# Prediction accuracy

observed.classes <- Test$chd
mean(predicted.classes == observed.classes)

[1] 0.6956522

#Prediction accuracy of the stepwise logistic regression model:

# Make predictions
probabilitiesstepmodel <- step.model %>% predict(Test, type = "response")

predicted.classesstepmodel <- ifelse(probabilitiesstepmodel > 0.5, "1", "0")


# Prediction accuracy

observed.classesstepmodel <- Test$chd
mean(predicted.classesstepmodel == observed.classesstepmodel)

[1] 0.7217391
```

Table 4. Comparison Full vs Reduced Model

	Full Model	Reduced Model
Prediction Accuracy	69.56%	72.17%
AIC	366.37	361.6

Results

The stepwise logistic regression model was chosen.

$\text{logit}(p) = \log(p/(1-p)) = \beta0 + \beta1*\text{tobacco} + \beta2*\text{ldl} + \beta3*\text{famhistPresent} + \beta1*\text{typea} + \beta2*\text{ldl} + \beta3*\text{famhistPresent} + \beta4*\text{typea} + \beta5*\text{obesity} + \beta6*\text{age}$

The result of the model chosen can be seen below in the table 5:

Table 5. Results from multivariate logistic regression

Term	β estimate	Standard error	P value
Intercept (β0)	-5.20391	1.24004	<0.001
tobacco (β1)	0.08172	0.02943	0.005485
ldl (β2)	0.22509	0.07310	0.002075
famhistPresent (β3)	0.89114	0.26825	<0.001
Type (β4)	0.03886	0.01407	0.005744
obesity (β5)	-0.07082	0.03577	0.047686
age (β6)	0.05465	0.01204	<0.001

Individuals who do not have a family history of coronary heart disease have an 8.5 % increase in the odds of having a Coronary Heart Disease for each one -unit increase in cumulative tobacco (kg). Since exp (0.08172) = 1.08515. Similarly, those individuals have an 25.24 % increase in the odds of having a Coronary Heart Disease for each one unit increase in ldl. Since, exp (0.22509) =1.25243. Similarly, Individuals who do not have a family history of coronary heart disease have an % 3.96 % increase in the odds of having a Coronary Heart Disease for each one -unit increase in typea behavior. Since exp (0.03886) = 1.03962. Individuals who do not have a family history of coronary heart disease have an 5.6% increase in the odds of having a Coronary Heart Disease for each one -unit increase in age. Since exp (0.05465) = 1.05617.

While holding the other predictors at a fixed value, the odds of having a Coronary Heart Disease is 143.79 % higher for individuals with a family history of Coronary Heart Disease than for individuals who do not have a family history. Since exp(0.89114)=2.4379.

Surprisingly, in this population individuals with no family history of heart disease the odds of having a Coronary Heart Disease decreases by 6.8 % for each one unit increase in obesity. Since exp (-0.07082) = 0.9312.

Conclusion

In the analysis was found that the factors that increase the risk of having a Coronary Heart Disease in this population are a family history of Coronary Heart Disease, Cumulative tobacco (kg), Low Density Lipoprotein Cholesterol (ldl), type A Behavior and Age at onset.