Term Project Step 3

Shashi Bhushan

2022-06-04

Introduction

This project Identify parameters that can be used to predict risk of heart attack utilizing existing data set from different hospital systems. Every year, more than 800,000 people have heart attack only in the U.S.A ("Centers for Disease Control and Prevention_2022" n.d.). Therefore, identification of certain body characteristics and health conditions will be helpful in understanding the risk of a heart attack and provide people a chance to remedy before it happens.Therefore, it is data science problem as it depends of variaous body attributes and other existing health conditions.

The problem statement you addressed

Following problem statements have been addressed utilzing three dataset as seen in the next section:

1. Identify relevant predictor variables such as age, sex, cholesterol, etc. from the dataset to predict heart disease diagnosis.
2. Identify relevant predictor variables to predict death due to heart attack.
3. Identify which predictor variables explain the variability the most.
4. Does the people who smoke are more suscepticle to heart disease / attack?
5. Are males or females more susceptible to heart disease / attack?

How you addressed this problem statement

Dataset Used

Following dataset have been used:

1. Heart Failure Prediction ("Heart Failure Prediction" n.d.)- Data has thirteen variables. It was used in the project, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." Dataset does not have missing values.

2. Heart Failure Prediction dataset ("Heart Failure Prediction Dataset," n.d.) - This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. It also has the predicted variable denoting heart disease.Dataset does not have missing values.

3. Heart Attack Prediction ("Heart Attack Prediction" n.d.)- This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. Missing values are denoted by a "?." This data was collected from four locations.

Methodology Used

Methodology includes the following:

1. Produce summary statistics of the data for the variables known to have direct correlation such as cholesterol and smoking with heart failure and/or heart attack.
2. Produce boxplots to illustrate association of heart disease with health parameters.
3. Create a logistic multiple regression model to identify relevant predictor variables.
4. Carry out hypothesis testing to test whether people who smoke and male or female are more susceptible to heart disease / attacks.

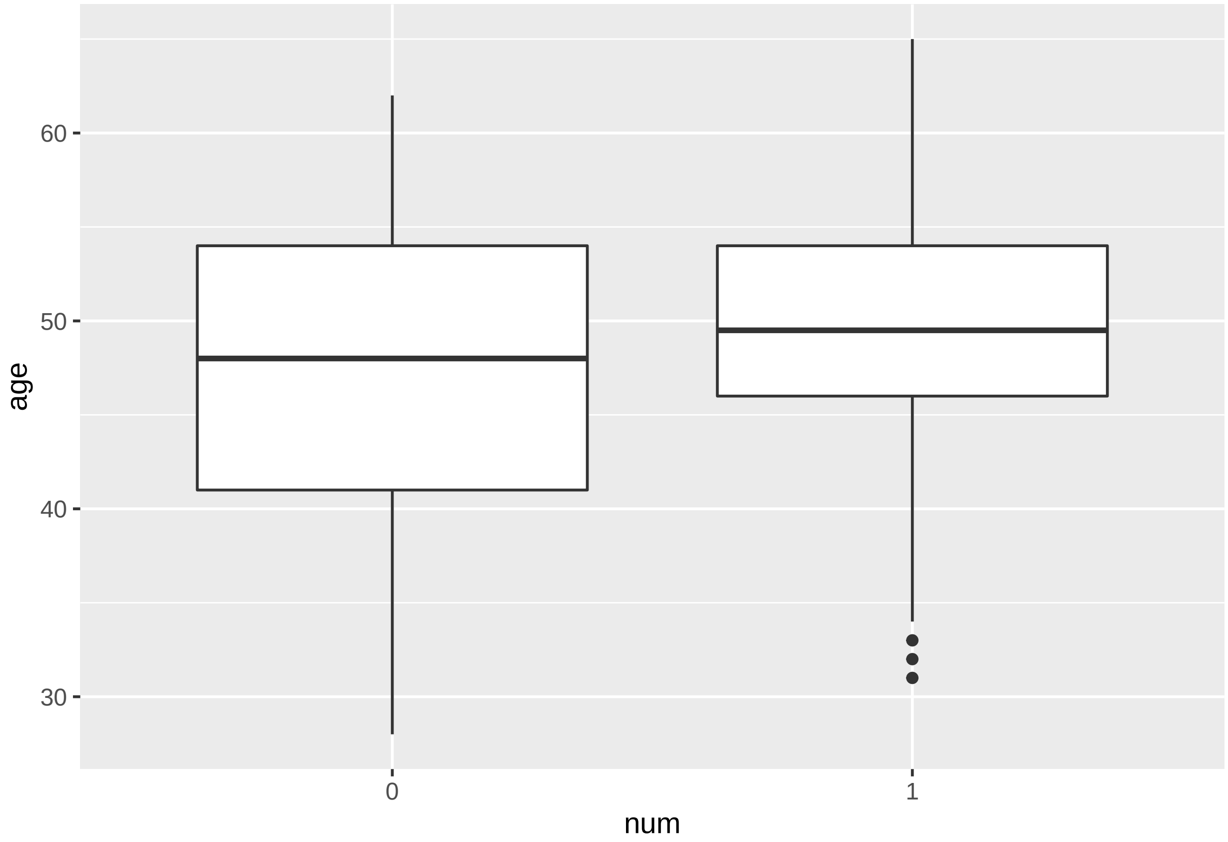
Model / Analysis Steps

```
setwd("/Users/sbhus/OneDrive/Documents/MSDS/DSC 520/Weekly Exercises/term project")
data1 <- read.csv("heart_attack_prediction.csv")
data2 <- read.csv("heart_failure_clinical_records_dataset.csv")
data3 <- read.csv("failure_prediction.csv")
library(lm.beta)
library(ggplot2)
library(dplyr)
```

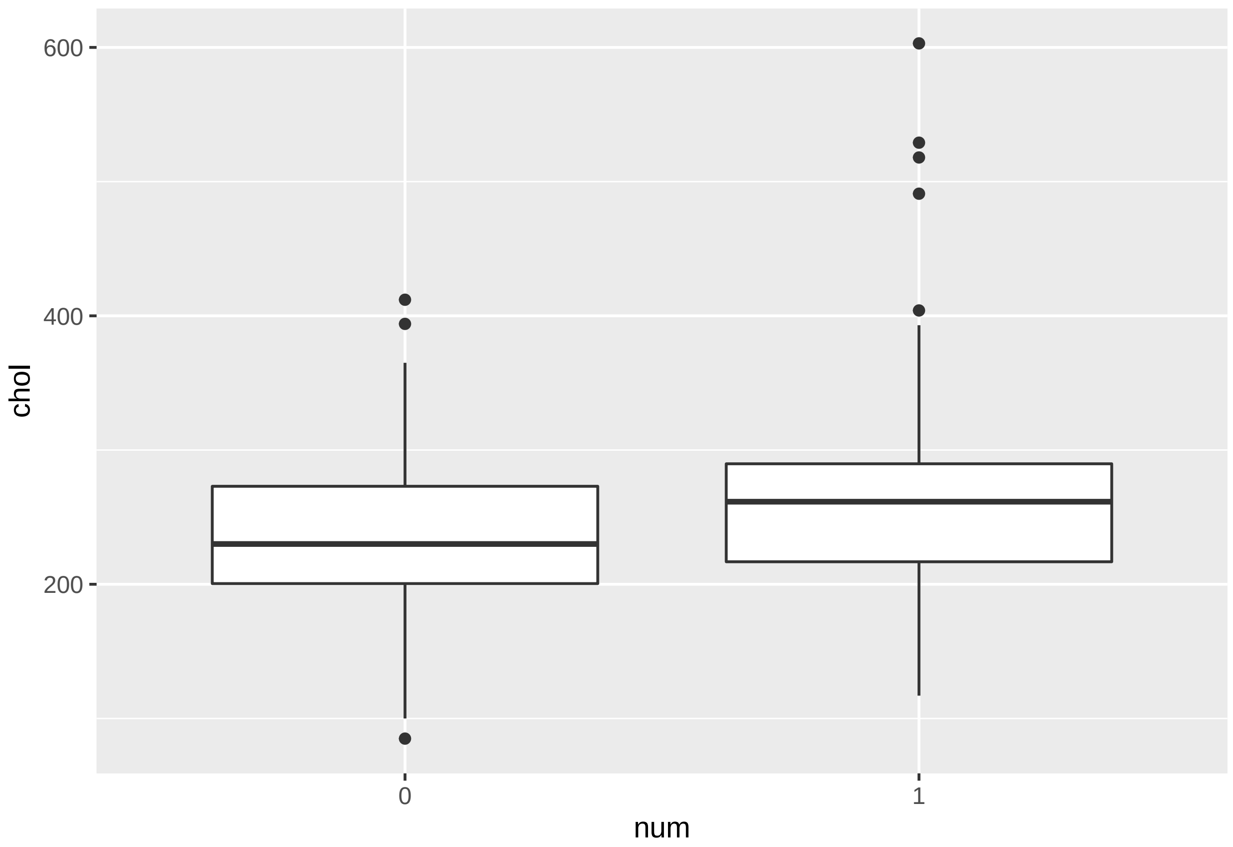
```
##
## Attaching package: ’dplyr’

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

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

```
library(caTools)

# Data Preparation
data1 <- select(data1, -slope, -ca, -thal)
data1[data1 == "?"] <- NA
data1 <- na.omit(data1)
data1$sex <-as.factor(data1$sex)
data1$cp <- as.factor(data1$cp)
data1$fbs <- as.factor(data1$fbs)
data1$restecg <- as.factor(data1$restecg)
data1$exang <- as.factor(data1$exang)
data1$num <- as.factor(data1$num)
data1$trestbps <- as.numeric(data1$trestbps)
data1$chol <- as.numeric(data1$chol)
data1$thalach <- as.numeric(data1$thalach)
data2$anaemia <- as.factor(data2$anaemia)
data2$diabetes <- as.factor(data2$diabetes)
data2$high_blood_pressure <- as.factor(data2$high_blood_pressure)
```

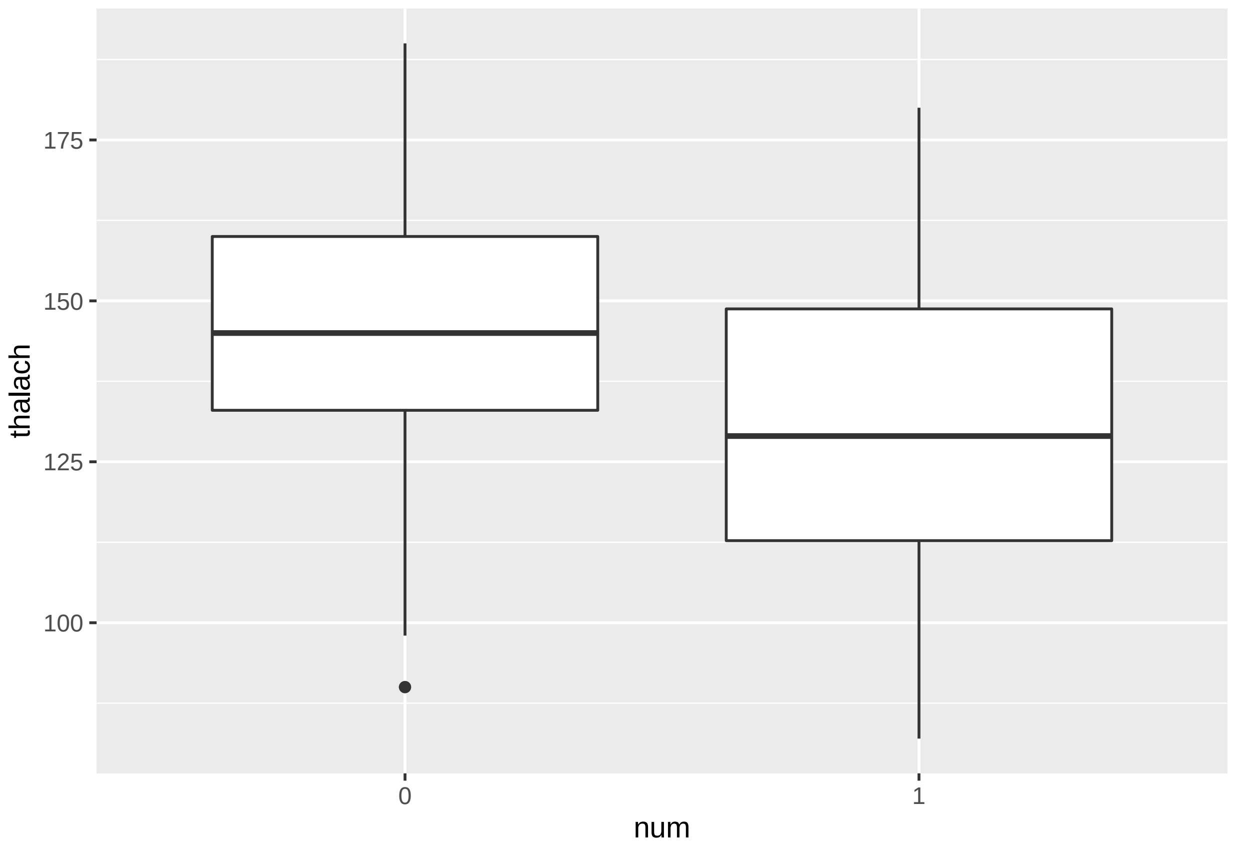
```
data2$sex <- as.factor(data2$sex)
data2$smoking <- as.factor(data2$smoking)
data2$DEATH_EVENT <- as.factor(data2$DEATH_EVENT)
data3$FastingBS <- as.factor(data3$FastingBS)
data3$HeartDisease <- as.factor(data3$HeartDisease)


# Boxplots

ggplot(data1, aes(x=num, y=age)) + geom_boxplot()
```



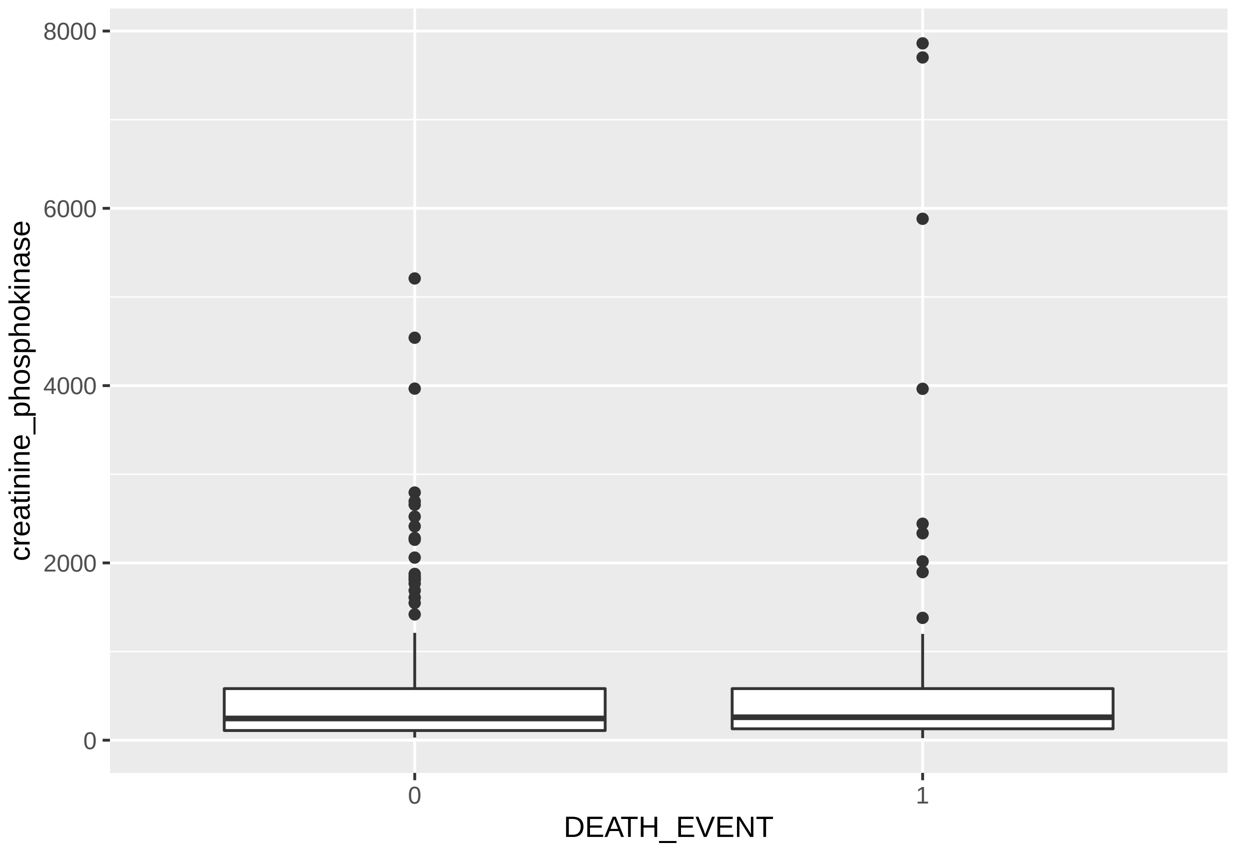
```
ggplot(data1, aes(x=num, y=chol)) + geom_boxplot()
```



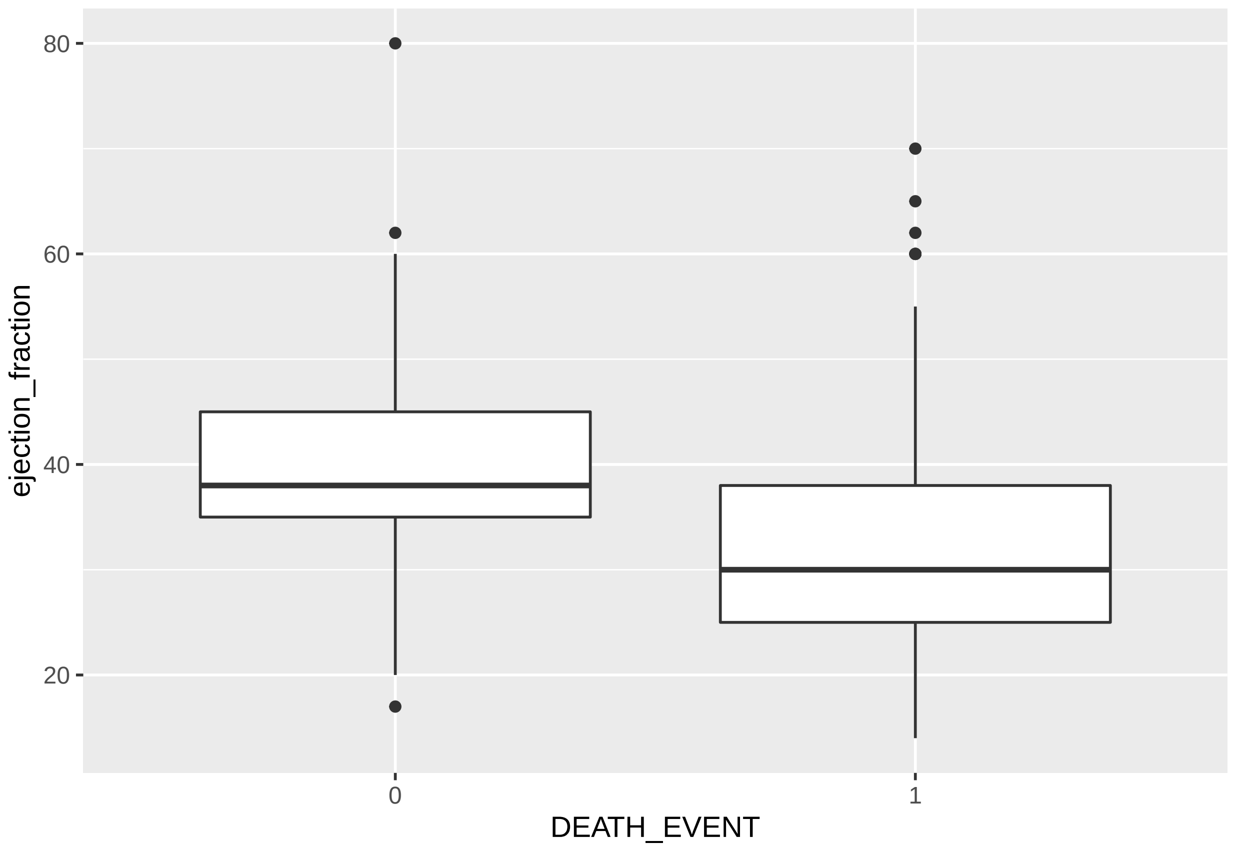
```
ggplot(data1, aes(x=num, y=thalach)) + geom_boxplot()
```



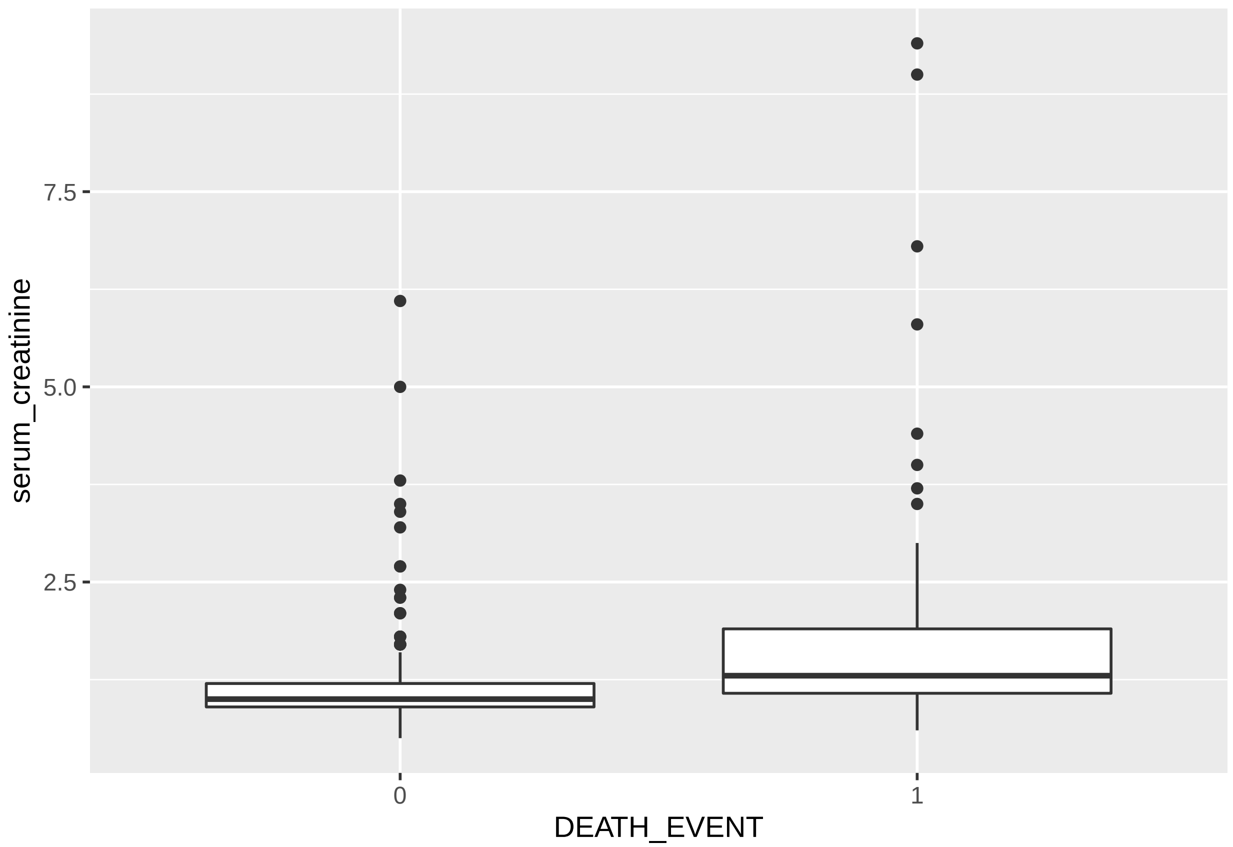
```
ggplot(data2, aes(x=DEATH_EVENT, y = creatinine_phosphokinase)) + geom_boxplot()
```



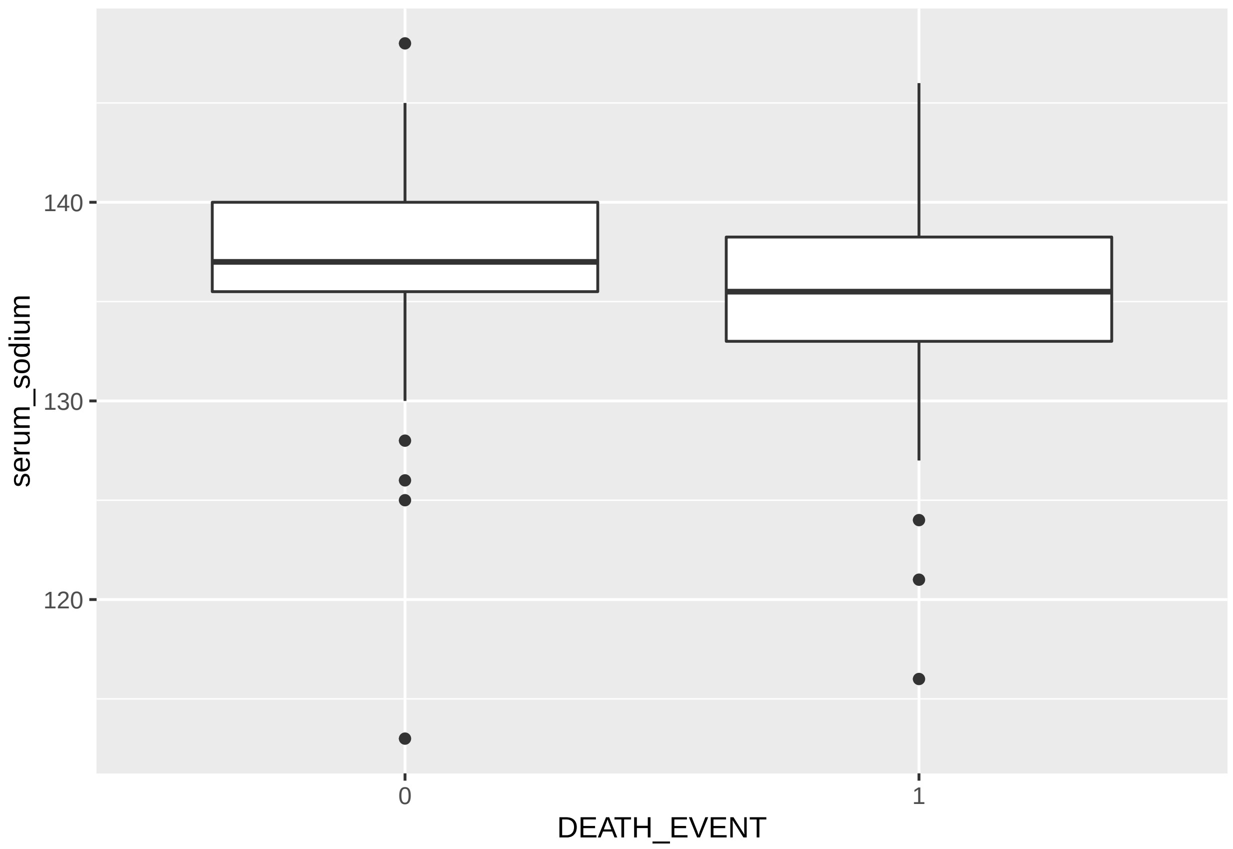
```
ggplot(data2, aes(x=DEATH_EVENT, y = ejection_fraction)) + geom_boxplot()
```



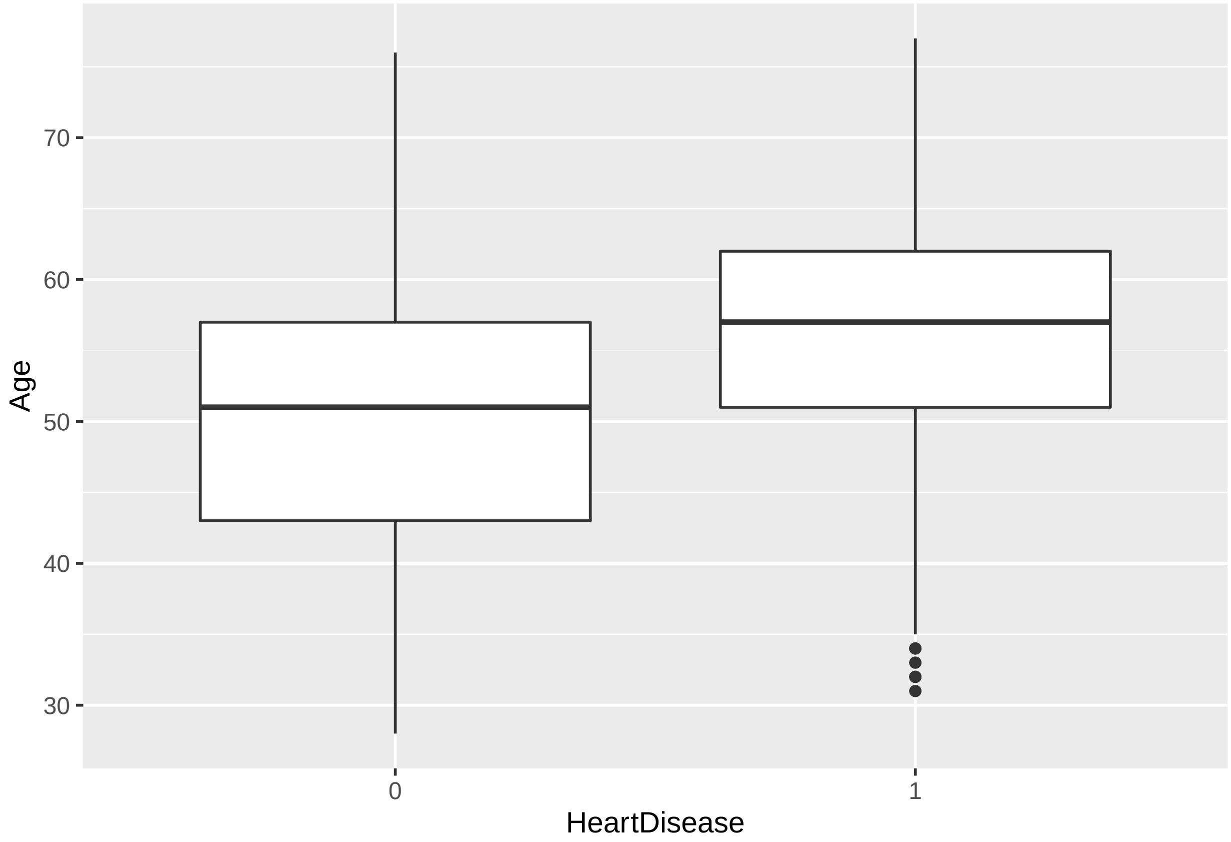
```
ggplot(data2, aes(x=DEATH_EVENT, y = serum_creatinine)) + geom_boxplot()
```



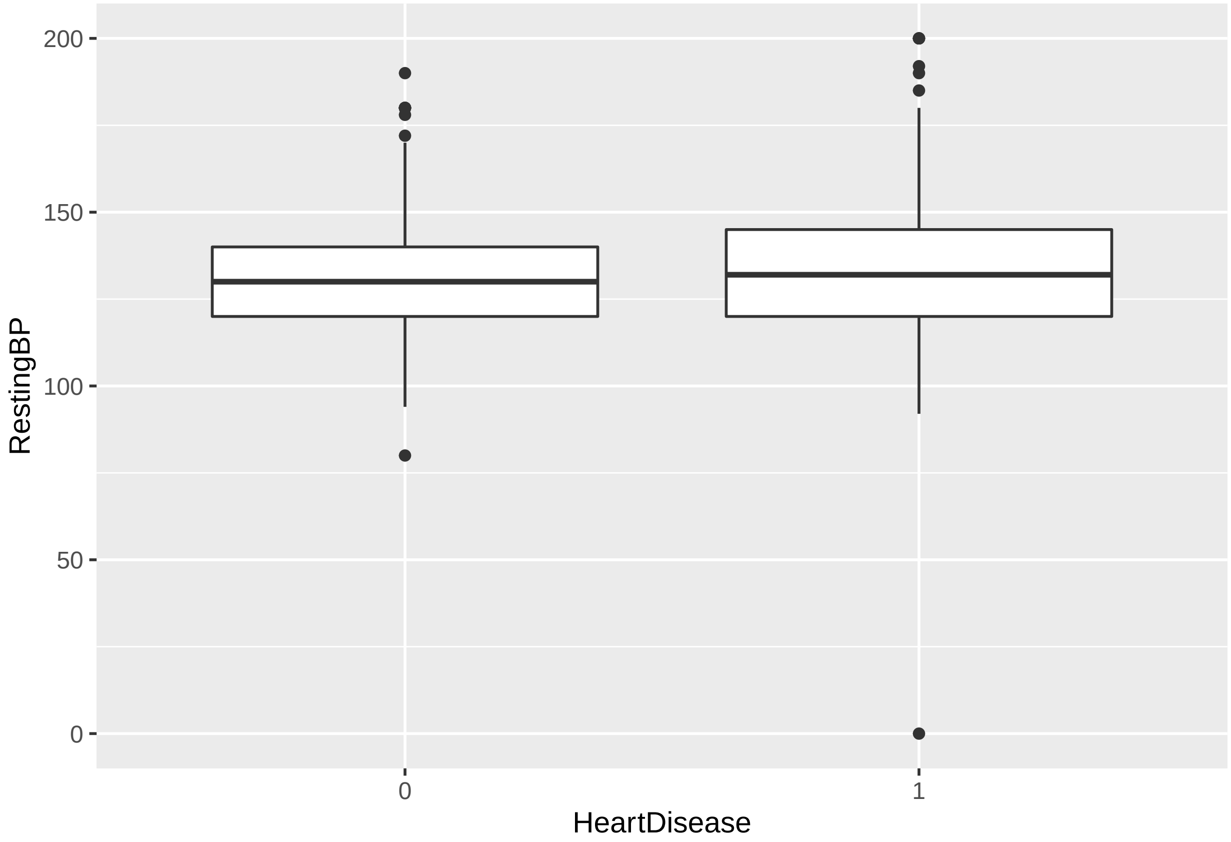
```
ggplot(data2, aes(x=DEATH_EVENT, y = serum_sodium)) + geom_boxplot()
```



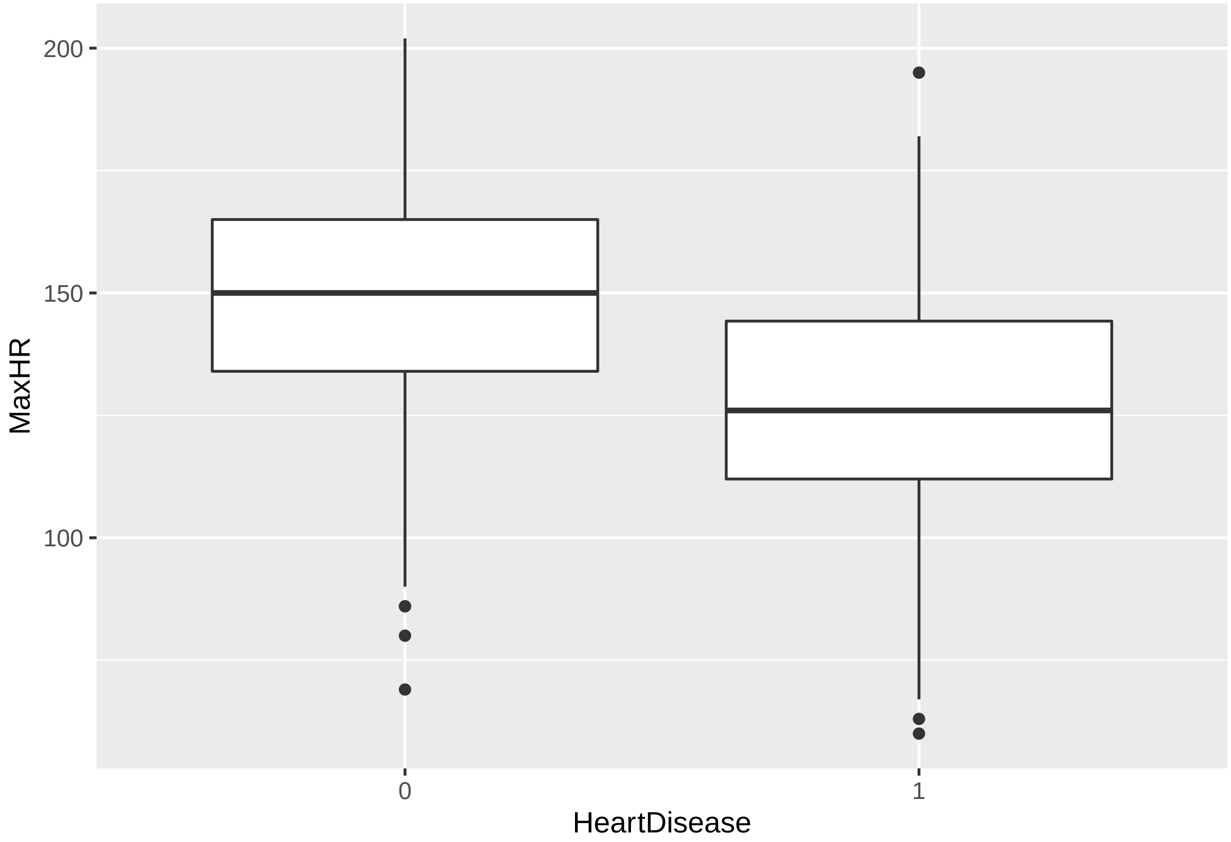
```
ggplot(data3, aes(x=HeartDisease, y = Age)) + geom_boxplot()
```



```
ggplot(data3, aes(x=HeartDisease, y = RestingBP)) + geom_boxplot()
```



```
ggplot(data3, aes(x=HeartDisease, y = MaxHR)) + geom_boxplot()
```



```
# Data 1 Analysis
split <- sample.split(data1, SplitRatio = 0.8)
train <- subset(data1, split == "TRUE")
validate <- subset(data1, split == "FALSE")
lgm_data1 <- glm(num ~ age+sex+cp+trestbps+
                   chol+fbs+restecg+thalach+exang+oldpeak,
                 data=train, family=binomial())
summary(lgm_data1)
```

```
##
## Call:
## glm(formula = num ~ age + sex + cp + trestbps + chol + fbs +
##     restecg + thalach + exang + oldpeak, family = binomial(),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3698  -0.5191  -0.2503   0.4677   2.3959
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.502525   3.563658  -1.263 0.206425
## age          0.007374   0.033070   0.223 0.823559
## sex1         0.848648   0.545283   1.556 0.119627
## cp2         -1.448804   1.319383  -1.098 0.272164
## cp3         -0.172144   1.292906  -0.133 0.894079
## cp4          1.080182   1.266396   0.853 0.393683
## trestbps     0.010950   0.012484   0.877 0.380403
## chol         0.005049   0.003309   1.526 0.126981
## fbs1         1.681779   0.844271   1.992 0.046372 *
## restecg1    -1.049184   0.646776  -1.622 0.104766
## restecg2    -0.362185   2.028955  -0.179 0.858324
## thalach     -0.006167   0.011959  -0.516 0.606076
## exang1       0.458949   0.559378   0.820 0.411953
## oldpeak      1.179658   0.317122   3.720 0.000199 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 250.20  on 188  degrees of freedom
## Residual deviance: 141.04  on 175  degrees of freedom
## AIC: 169.04
##
## Number of Fisher Scoring iterations: 6
```

```
lgm_data1 <- glm(num ~ sex+exang+oldpeak, data=train, family=binomial())
summary(lgm_data1)
```

```
##
## Call:
## glm(formula = num ~ sex + exang + oldpeak, family = binomial(),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6252  -0.6186  -0.3857   0.5999   2.2959
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5613     0.4791  -5.346 8.99e-08 ***
## sex1          1.0048     0.4716   2.131  0.03311 *
## exang1        1.3907     0.4334   3.209  0.00133 **
## oldpeak       1.1931     0.2869   4.159 3.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 250.20  on 188  degrees of freedom
## Residual deviance: 173.82  on 185  degrees of freedom
## AIC: 181.82
##
## Number of Fisher Scoring iterations: 5
```

```
res <- predict(lgm_data1, validate, type = "response")
res2 <- predict(lgm_data1, train, type = "response")
confmatrix <- table(Actual_Value=train$num, Predicted_Value = res2 >0.5)
(confmatrix[[1.1]] +confmatrix[[2,2]])/sum(confmatrix)
```

```
## [1] 0.7989418
```

```
# Data 2 Analysis
split <- sample.split(data2, SplitRatio = 0.8)
train <- subset(data2, split == "TRUE")
validate <- subset(data2, split == "FALSE")
lgm_data2 <- glm(DEATH_EVENT ~ age+anaemia+creatinine_phosphokinase+
                   diabetes+ejection_fraction+high_blood_pressure+
                   platelets+serum_creatinine+serum_sodium+
                   sex+smoking+time, data=train, family=binomial())
summary(lgm_data2)
```

```
##
## Call:
## glm(formula = DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase +
##     diabetes + ejection_fraction + high_blood_pressure + platelets +
##     serum_creatinine + serum_sodium + sex + smoking + time, family = binomial(),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3954  -0.5386  -0.1878   0.3558   2.3906
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.484e+01  7.722e+00   1.922 0.054646 .
## age                       7.520e-02  2.092e-02   3.594 0.000325 ***
## anaemia1                  3.427e-01  4.268e-01   0.803 0.422085
## creatinine_phosphokinase  3.250e-04  2.214e-04   1.468 0.142108
## diabetes1                -6.823e-03  4.194e-01  -0.016 0.987019
## ejection_fraction        -1.085e-01  2.190e-02  -4.953 7.30e-07 ***
## high_blood_pressure1     -6.945e-02  4.324e-01  -0.161 0.872382
## platelets                -2.535e-06  2.491e-06  -1.018 0.308731
## serum_creatinine          8.233e-01  2.145e-01   3.838 0.000124 ***
## serum_sodium             -1.024e-01  5.508e-02  -1.859 0.063035 .
## sex1                     -9.477e-01  5.219e-01  -1.816 0.069410 .
## smoking1                  4.837e-01  5.136e-01   0.942 0.346288
## time                     -2.219e-02  3.643e-03  -6.092 1.11e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 293.26  on 229  degrees of freedom
## Residual deviance: 154.27  on 217  degrees of freedom
## AIC: 180.27
##
## Number of Fisher Scoring iterations: 6
```

```
lgm_data2 <- glm(DEATH_EVENT ~ age+ejection_fraction+serum_creatinine
                 +time, data=train, family=binomial())
summary(lgm_data2)
```

```
##
```

```
## Call:
## glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
##     time, family = binomial(), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3034  -0.5674  -0.2133   0.3790   2.2216
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.480678   1.198500   0.401 0.688371
## age                0.062921   0.018190   3.459 0.000542 ***
## ejection_fraction -0.101709   0.020310  -5.008 5.50e-07 ***
## serum_creatinine   0.857261   0.208276   4.116 3.86e-05 ***
## time              -0.021763   0.003429  -6.347 2.20e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 293.26  on 229  degrees of freedom
## Residual deviance: 163.45  on 225  degrees of freedom
## AIC: 173.45
##
## Number of Fisher Scoring iterations: 6
```

```
res <- predict(lgm_data2, validate, type = "response")
res2 <- predict(lgm_data2, train, type = "response")
confmatrix <- table(Actual_Value=train$DEATH_EVENT, Predicted_Value = res2 >0.5)
(confmatrix[[1.1]] +confmatrix[[2,2]])/sum(confmatrix)
```

```
## [1] 0.8434783
```

```
# Data 3 Analysis
split <- sample.split(data3, SplitRatio = 0.8)
train <- subset(data3, split == "TRUE")
validate <- subset(data3, split == "FALSE")
lgm_data3 <- glm(HeartDisease ~ Age+Sex+ChestPainType+RestingBP+
                   Cholesterol+FastingBS+RestingECG+MaxHR+ExerciseAngina+
                   Oldpeak+ST_Slope, data=train, family=binomial())
summary(lgm_data3)
```

```
##
## Call:
## glm(formula = HeartDisease ~ Age + Sex + ChestPainType + RestingBP +
##     Cholesterol + FastingBS + RestingECG + MaxHR + ExerciseAngina +
##     Oldpeak + ST_Slope, family = binomial(), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6910  -0.3179   0.1483   0.4204   2.6861
##
## Coefficients:
```

```
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.6549982  1.7244893  -0.380 0.704078
## Age                0.0084104  0.0161671   0.520 0.602910
## SexM               1.7035230  0.3499553   4.868 1.13e-06 ***
## ChestPainTypeATA -2.0238082  0.4104293  -4.931 8.18e-07 ***
## ChestPainTypeNAP -2.0454947  0.3317495  -6.166 7.01e-10 ***
## ChestPainTypeTA  -1.8083115  0.5186958  -3.486 0.000490 ***
## RestingBP         -0.0004637  0.0073620  -0.063 0.949779
## Cholesterol       -0.0043411  0.0012894  -3.367 0.000760 ***
## FastingBS1         1.3670436  0.3286284   4.160 3.18e-05 ***
## RestingECGNormal -0.2502107  0.3208049  -0.780 0.435422
## RestingECGST     -0.3564143  0.4104469  -0.868 0.385199
## MaxHR             -0.0023195  0.0061119  -0.380 0.704310
## ExerciseAnginaY   0.9953086  0.2915358   3.414 0.000640 ***
## Oldpeak            0.3724326  0.1392300   2.675 0.007474 **
## ST_SlopeFlat       1.7891805  0.5315252   3.366 0.000762 ***
## ST_SlopeUp        -0.9555825  0.5566211  -1.717 0.086024 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 946.64  on 687  degrees of freedom
## Residual deviance: 413.46  on 672  degrees of freedom
## AIC: 445.46
##
## Number of Fisher Scoring iterations: 6
```

```
lgm_data3 <- glm(HeartDisease ~ Sex+ChestPainType+Cholesterol+
                   FastingBS+ExerciseAngina+
                   Oldpeak+ST_Slope, data=train, family=binomial())
summary(lgm_data3)
```

```
##
## Call:
## glm(formula = HeartDisease ~ Sex + ChestPainType + Cholesterol +
##     FastingBS + ExerciseAngina + Oldpeak + ST_Slope, family = binomial(),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7508  -0.3207   0.1464   0.4192   2.7822
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.769743   0.702994  -1.095 0.273539
## SexM               1.695619   0.347220   4.883 1.04e-06 ***
## ChestPainTypeATA -2.065955   0.405898  -5.090 3.58e-07 ***
## ChestPainTypeNAP -2.067013   0.324248  -6.375 1.83e-10 ***
## ChestPainTypeTA  -1.808224   0.501215  -3.608 0.000309 ***
## Cholesterol       -0.004301   0.001222  -3.518 0.000435 ***
## FastingBS1         1.370031   0.325165   4.213 2.52e-05 ***
## ExerciseAnginaY   1.017197   0.282168   3.605 0.000312 ***
## Oldpeak            0.386095   0.134930   2.861 0.004217 **
## ST_SlopeFlat      1.777226   0.531017   3.347 0.000817 ***
## ST_SlopeUp       -1.011658   0.549069  -1.842 0.065403 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 946.64  on 687  degrees of freedom
## Residual deviance: 415.01  on 677  degrees of freedom
## AIC: 437.01
##
## Number of Fisher Scoring iterations: 6
```

```
res <- predict(lgm_data3, validate, type = "response")
res2 <- predict(lgm_data3, train, type = "response")
confmatrix <- table(Actual_Value=train$HeartDisease, Predicted_Value = res2 >0.5)
(confmatrix[[1,1]] +confmatrix[[2,2]])/sum(confmatrix)
```

```
## [1] 0.880814
```

```
data2$smoking <- as.numeric(data2$smoking)
data2$DEATH_EVENT <- as.numeric(data2$DEATH_EVENT)
data2$sex <- as.numeric(data2$sex)

# Hypothsis testing for correlation using Data 2
cor.test(data2$smoking, data2$DEATH_EVENT)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data2$smoking and data2$DEATH_EVENT
## t = -0.21756, df = 297, p-value = 0.8279
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1258732  0.1009517
## sample estimates:
##         cor
## -0.01262315
```

```
cor.test(data2$sex, data2$DEATH_EVENT)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data2$sex and data2$DEATH_EVENT
## t = -0.074388, df = 297, p-value = 0.9408
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1176891  0.1091674
## sample estimates:
##          cor
## -0.004316376
```

Analysis

Data 1 Results

Boxplots

Reference to the boxplots, following variables appear to be determining factors in predicting heart disease:

1. Age as the median and lower quartile numbers are higher for people with heart disease.
2. Cholesterol with lower, median, and upper quartile numbers being higher for people with evidence of the heart disease.

Regression

Logistic regression indicates that sex (Male), exang (exercise induced angina), and oldpeak(ST depression induced by exercise relative to rest) are the statistically relevant variables with p-values less than 0.05. Regression equation with these three variables had an accuracy rate of ~83%.

Data 2 Results

Boxplots

Reference to the boxplots, following variables appear to be determining factors in predicting death event due to heart failure:

1. Serum Creatinine with ower, median, and upper quartile numbers being higher for people with evidence of death event.

Regression

Logistic regression indicates that age, ejection_fraction(Percentage of blood leaving the heart at each contraction), serum_creatinine (Level of serum creatinine in the blood), time (Follow-up period in days) are the statistically relevant variables with p-values less than 0.05. Regression equation with these three variables had an accuracy rate of ~83% in predicting heart failure related death event.

Hypothesis Testing to indentigy significant correlation between variables

Hypothesis test to check if true correlation is not equal to zero for Death_Event vs sex and smoking variables. In both cases, null hypothsis that true correlation is equal to zero could not be rejected based on p-values.

Data 3 Results

Boxplots

Reference to the boxplots, following variables appear to be determining factors in predicting heart disease :

1. Age as the lower, median and lower quartile numbers are higher for people with heart disease.

Regression

Logistic regression indicates that sex, chestpaintype, cholesterol, fastingBS (fasting blood sugar), ExerciseAngina (exercise induced angina), oldpeak (ST depression induced by exercise relative to rest),ST_Slope (the slope of the peak exercise ST segment) are the statistically relevant variables with p-values less than 0.05. Regression equation with these three variables had an accuracy rate of ~87%.

Implications

Main takeaways from the analysis of three dataset are:

1. Age 50-60 appear to be most vulnerable to heart disease / failure based on the boxplot.
2. Exercise test / Stress test helps in predicting heart disease or death event associated with heart disease.
3. Cholesterol, blood sugar are indicator of heart disease.
4. Smoking and gender do not appear to play a role in heart disease/failure.

Limitations

1. Age was found to be statistically relevant in 1 of the 3 dataset used. More comprehensive data collection and analysis is required to check a statistically significant connection.
2. Did not check for normalcy of the dataset and therefore, it is not confirmed that the data set are representing population.Check for normalcy can help in confirming the results applicable to entire population.

Concluding Remarks

Age, cholesterol, blood sugar, stress test, etc. are widely used for identifying heart disease. therefore, results are not surprising. However, since the data set are only ranging from 261 to 918 observations, collection of unbiased data from a larger sample may help in making robust conclusions.

References

"Centers for Disease Control and Prevention_2022." n.d. www.cdc.com. Accessed February 17, 2022. https://www.cdc.gov/heartdisease/facts.htm.

"Heart Attack Prediction." n.d. www.kaggle.com. Accessed May 17, 2022. https://www.kaggle.com/datasets/imnikhilanand/heart-attack-prediction.

"Heart Failure Prediction." n.d. www.kaggle.com. Accessed May 17, 2022. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data?resource=download.

"Heart Failure Prediction Dataset." n.d. www.kaggle.com. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.