

# R Notebook

Install and load helpful library for analysis.

```
# Install required library, need not to re-install if you already have  
# install.packages("dplyr")  
# install.packages("Hmisc")  
# install.packages("ggplot2")  
  
# Load the require library  
# library(dplyr)  
# library(Hmisc)  
# library(ggplot2)
```

## Data Loading & Exploratory data analysis

Load the data in csv file and store to variable name data.

```
path <- "https://raw.githubusercontent.com/pnhuy/datasets/master/heart_uci/heart.csv"  
data <- read.csv(path)
```

Show some rows of data to get some insight of data.

```
head(data)
```

```
##   X Age Sex   ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope  
## 1 1 63  1    typical    145  233  1         2   150    0     2.3    3  
## 2 2 67  1 asymptomatic    160  286  0         2   108    1     1.5    2  
## 3 3 67  1 asymptomatic    120  229  0         2   129    1     2.6    2  
## 4 4 37  1 nonanginal    130  250  0         0   187    0     3.5    3  
## 5 5 41  0 nontypical    130  204  0         2   172    0     1.4    1  
## 6 6 56  1 nontypical    120  236  0         0   178    0     0.8    1  
##   Ca      Thal AHD  
## 1  0      fixed  No  
## 2  3      normal Yes  
## 3  2 reversable Yes  
## 4  0      normal No  
## 5  0      normal No  
## 6  0      normal No
```

Remove the first columns because it is not useful for analysis.

```
data_col <- colnames(data)  
data <- data[data_col[2:length(data_col)]]  
head(data)
```

```
##   Age Sex   ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope  
## 1 63  1    typical    145  233  1         2   150    0     2.3    3  
## 2 67  1 asymptomatic    160  286  0         2   108    1     1.5    2  
## 3 67  1 asymptomatic    120  229  0         2   129    1     2.6    2  
## 4 37  1 nonanginal    130  250  0         0   187    0     3.5    3  
## 5 41  0 nontypical    130  204  0         2   172    0     1.4    1
```

```
## 6 56 1 nontypical 120 236 0 0 178 0 0.8 1
## Ca Thal AHD
## 1 0 fixed No
## 2 3 normal Yes
## 3 2 reversable Yes
## 4 0 normal No
## 5 0 normal No
## 6 0 normal No
```

What was the average age?

```
mean(data$Age)
```

```
## [1] 54.43894
```

From sex column, create new variable gender which only have 'male', 'female'?

```
gender <- factor(data$Sex, levels=c(0,1), labels=c('female', 'male'))
```

How many percent of patient who was male were there in the data?

```
sum(data$Sex == 1)/length(data$Sex)*100
```

```
## [1] 67.9868
```

How many percent of male patient who suffered heart disease were there in the data? What about female?

```
print(sum(data[data$Sex == 1, 'AHD'] == 'Yes')/length(data[data$Sex == 1, 'AHD']))
```

```
## [1] 0.5533981
```

```
print(sum(data[data$Sex == 0, 'AHD'] == 'Yes')/length(data[data$Sex == 0, 'AHD']))
```

```
## [1] 0.257732
```

What was the range of RestBP?

```
print(c(min(data$RestBP), max(data$RestBP)))
```

```
## [1] 94 200
```

What was the distribution of AHD?

```
print(table(data$AHD))
```

```
##
```

```
## No Yes
```

```
## 164 139
```

Calculate some basic descriptive statistics

```
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
```

```
summary(data)
```

```
##           Age           Sex           ChestPain           RestBP
##  Min.      :29.00   Min.      :0.0000   asymptomatic:144   Min.      : 94.0
##  1st Qu.:48.00   1st Qu.:0.0000   nonanginal    : 86   1st Qu.:120.0
##  Median :56.00   Median :1.0000   nontypical    : 50   Median :130.0
##  Mean   :54.44   Mean   :0.6799   typical       : 23   Mean   :131.7
##  3rd Qu.:61.00   3rd Qu.:1.0000                   3rd Qu.:140.0
##  Max.    :77.00   Max.    :1.0000                   Max.    :200.0
##
##           Chol           Fbs           RestECG           MaxHR
##  Min.      :126.0   Min.      :0.0000   Min.      :0.0000   Min.      : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :241.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.7   Mean   :0.1485   Mean   :0.9901   Mean   :149.6
##  3rd Qu.:275.0   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:166.0
##  Max.    :564.0   Max.    :1.0000   Max.    :2.0000   Max.    :202.0
##
##           ExAng           Oldpeak           Slope           Ca
##  Min.      :0.0000   Min.      :0.00   Min.      :1.000   Min.      :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :2.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.601   Mean   :0.6722
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.    :1.0000   Max.    :6.20   Max.    :3.000   Max.    :3.0000
##
##                                     NA's      :4
##           Thal           AHD
##  fixed      : 18   No :164
##  normal     :166   Yes:139
##  reversable:117
##  NA's       : 2
##
##
##
```

```
library(Hmisc)
```

```
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
describe(data)
```

```
## data
##
## 14 Variables      303 Observations
## -----
## Age
##      n missing distinct    Info    Mean    Gmd    .05    .10
##    303      0      41    0.999   54.44   10.3    40    42
##    .25    .50    .75    .90    .95
##    48     56     61     66     68
##
## lowest : 29 34 35 37 38, highest: 70 71 74 76 77
## -----
## Sex
##      n missing distinct    Info    Sum    Mean    Gmd
##    303      0      2    0.653    206   0.6799   0.4367
##
## -----
## ChestPain
##      n missing distinct
##    303      0      4
##
## Value      asymptomatic    nonanginal    nontypical    typical
## Frequency          144          86          50          23
## Proportion        0.475        0.284        0.165        0.076
## -----
## RestBP
##      n missing distinct    Info    Mean    Gmd    .05    .10
##    303      0      50    0.995   131.7   19.41   108   110
##    .25    .50    .75    .90    .95
##    120    130    140    152    160
##
## lowest : 94 100 101 102 104, highest: 174 178 180 192 200
## -----
## Chol
##      n missing distinct    Info    Mean    Gmd    .05    .10
##    303      0      152     1    246.7   55.91   175.1   188.8
##    .25    .50    .75    .90    .95
##   211.0   241.0   275.0   308.8   326.9
##
## lowest : 126 131 141 149 157, highest: 394 407 409 417 564
## -----
## Fbs
##      n missing distinct    Info    Sum    Mean    Gmd
```

```

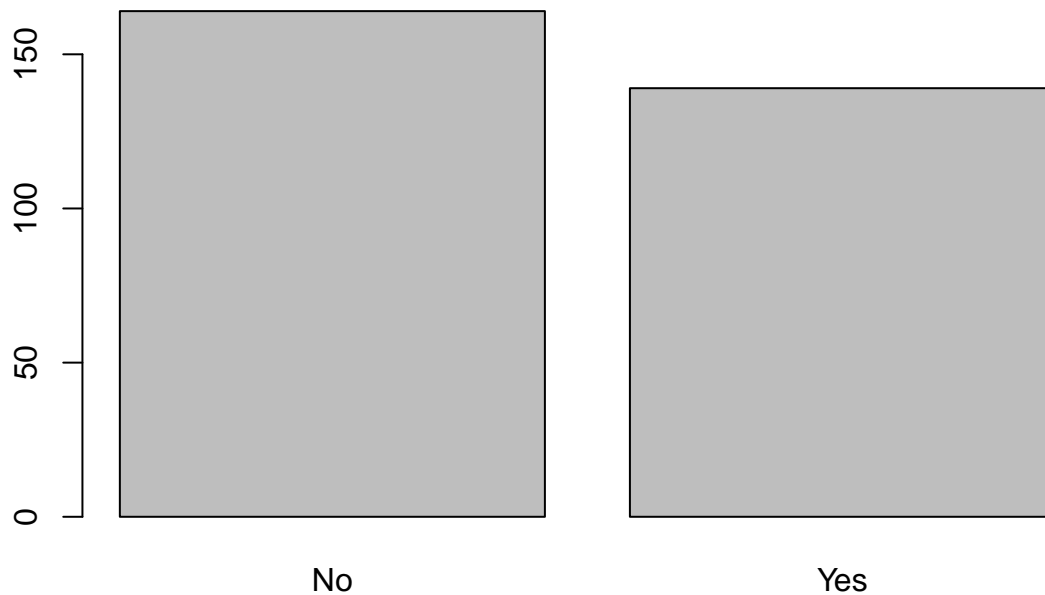
##      303      0      2    0.379      45    0.1485    0.2538
##
## -----
## RestECG
##      n missing distinct      Info      Mean      Gmd
##      303      0      3    0.76    0.9901    1.003
##
## Value      0      1      2
## Frequency   151      4    148
## Proportion 0.498 0.013 0.488
## -----
## MaxHR
##      n missing distinct      Info      Mean      Gmd      .05      .10
##      303      0      91      1    149.6    25.73    108.1    116.0
##      .25      .50      .75      .90      .95
##    133.5    153.0    166.0    176.6    181.9
##
## lowest : 71 88 90 95 96, highest: 190 192 194 195 202
## -----
## ExAng
##      n missing distinct      Info      Sum      Mean      Gmd
##      303      0      2    0.66      99    0.3267    0.4414
##
## -----
## Oldpeak
##      n missing distinct      Info      Mean      Gmd      .05      .10
##      303      0      40    0.964    1.04    1.225    0.0      0.0
##      .25      .50      .75      .90      .95
##      0.0      0.8      1.6      2.8      3.4
##
## lowest : 0.0 0.1 0.2 0.3 0.4, highest: 4.0 4.2 4.4 5.6 6.2
## -----
## Slope
##      n missing distinct      Info      Mean      Gmd
##      303      0      3    0.798    1.601    0.6291
##
## Value      1      2      3
## Frequency   142    140    21
## Proportion 0.469 0.462 0.069
## -----
## Ca
##      n missing distinct      Info      Mean      Gmd
##      299      4      4    0.783    0.6722    0.9249
##
## Value      0      1      2      3
## Frequency   176    65    38    20
## Proportion 0.589 0.217 0.127 0.067
## -----
## Thal
##      n missing distinct
##      301      2      3
##
## Value      fixed      normal reversable
## Frequency      18      166      117

```

```
## Proportion      0.060      0.551      0.389
## -----
## AHD
##      n missing distinct
##    303      0         2
##
## Value      No   Yes
## Frequency  164  139
## Proportion 0.541 0.459
## -----
```

Plot the distribution of AHD?

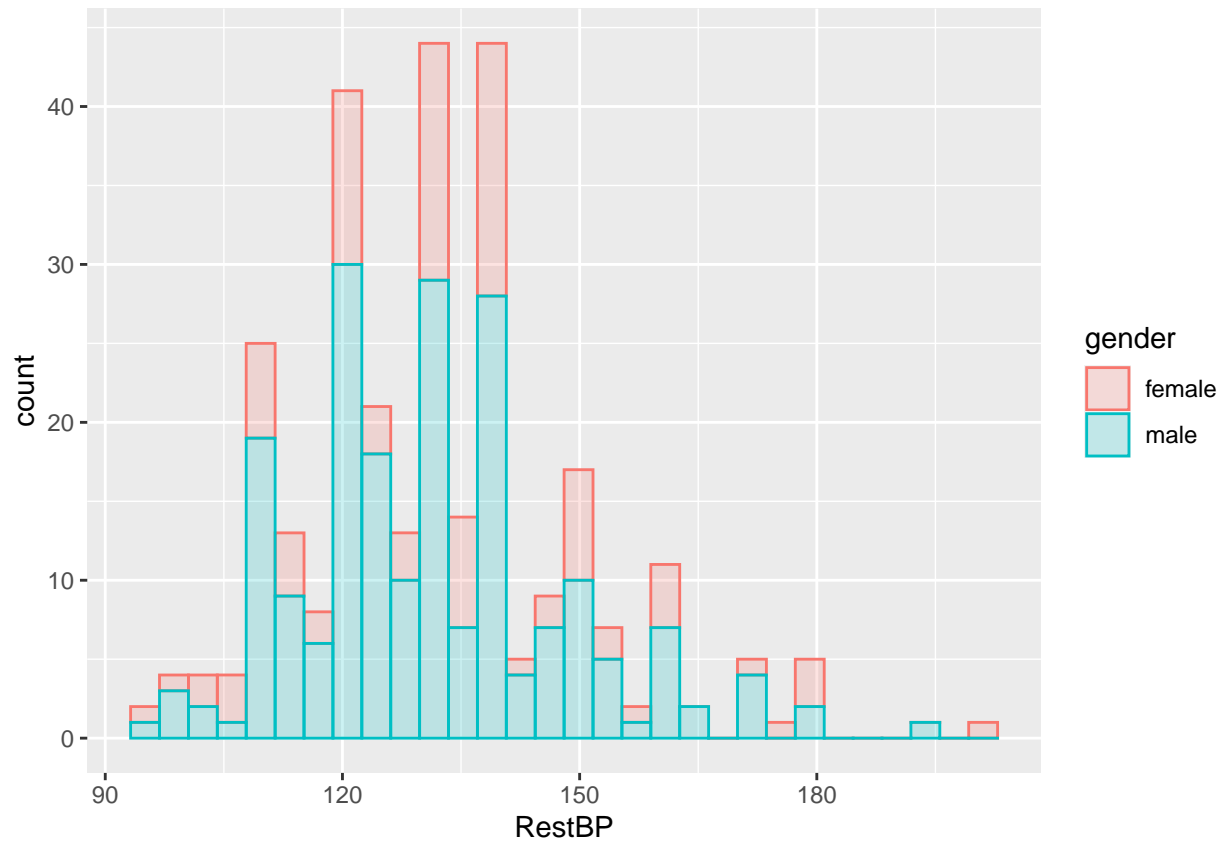
```
barplot(table(data$AHD))
```



Plot the distribution of RestBP per Sex?

```
# In this plot, the distribution need a Factor variable to split, so we input the `gender`
# instead of `Sex`
ggplot(data, aes(x=RestBP)) + geom_histogram(aes(color=gender, fill=gender), alpha=0.2)

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



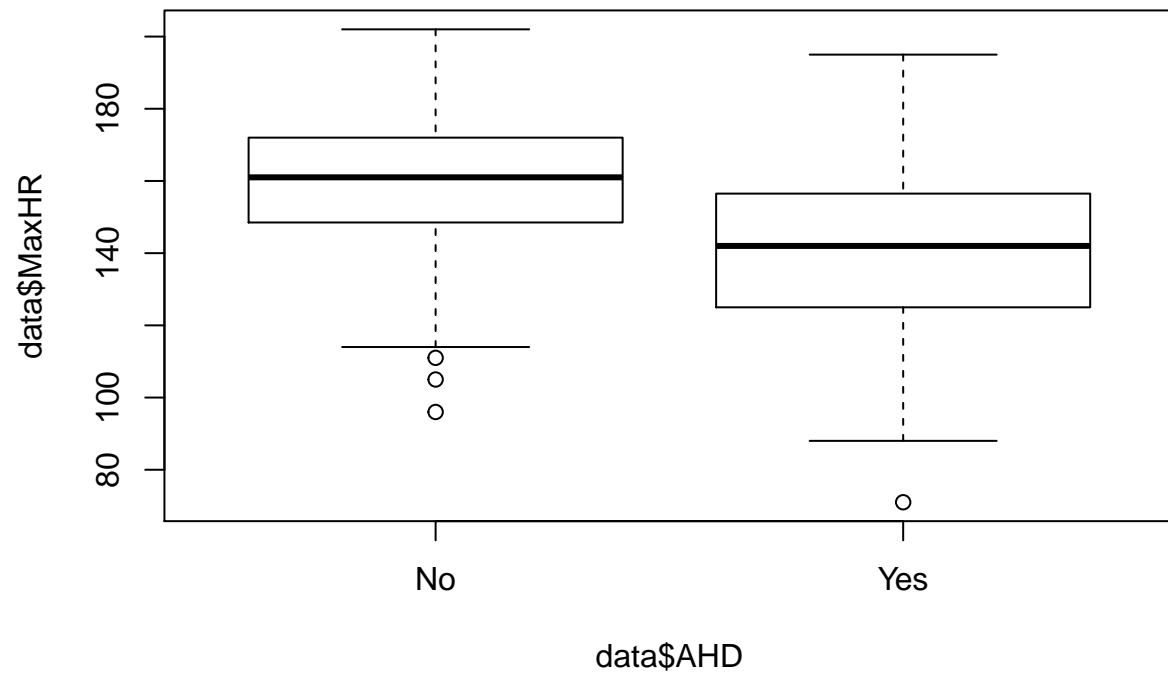
Illustrate the relationship between sex and AHD?

```
table(data$Sex, data$AHD)
```

```
##
##      No Yes
## 0    72  25
## 1    92 114
```

Illustrate the relationship between MaxHR and AHD?

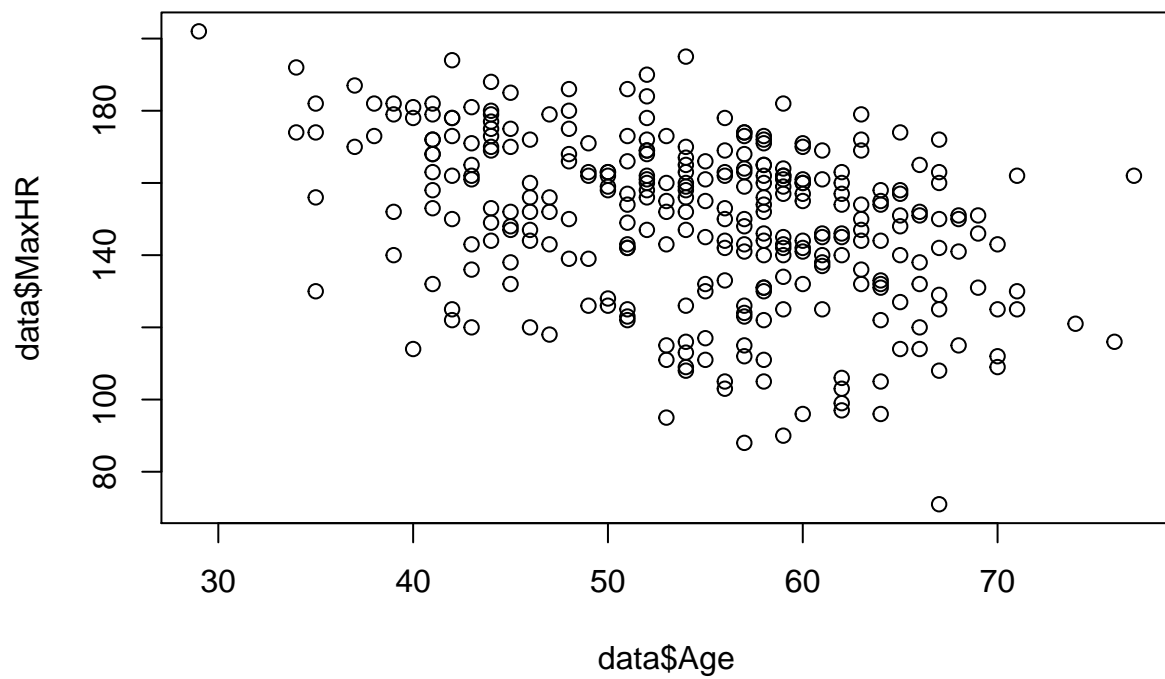
```
boxplot(data$MaxHR ~ data$AHD)
```



Illustrate the relationship between Age and MaxHR?

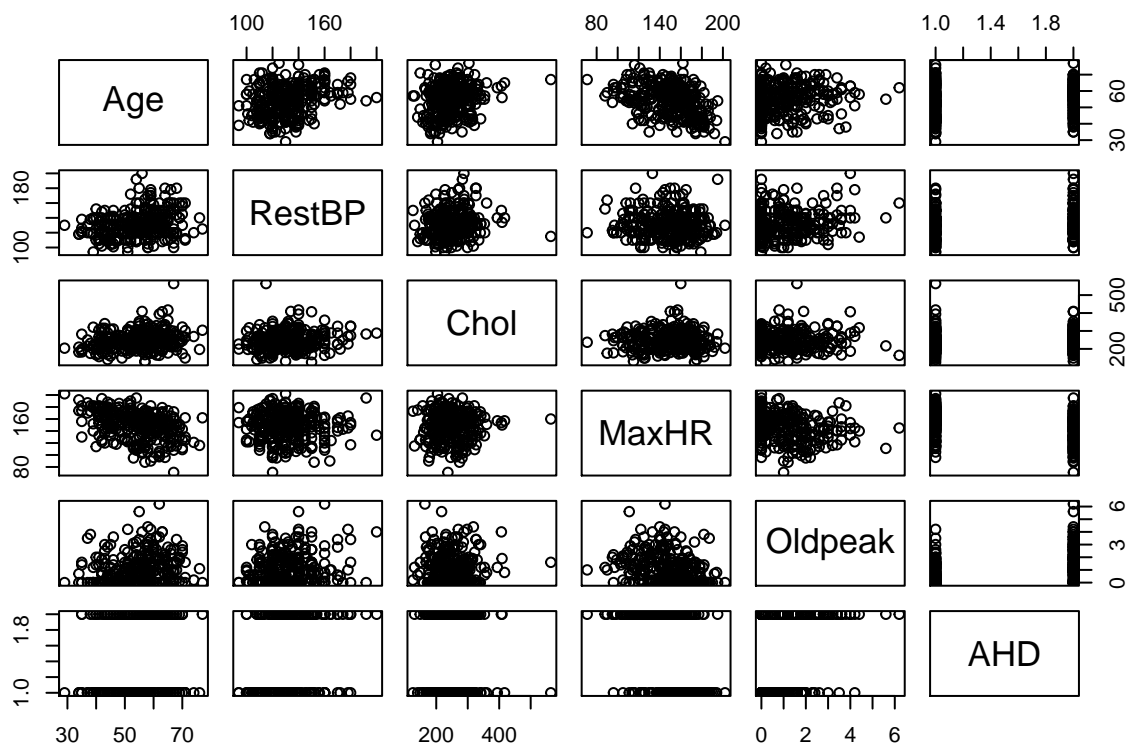
```
plot(data$Age, data$MaxHR)
```





Illustrate the relationship between the continuous variables and the target?

```
pairs(data[c('Age', 'RestBP', 'Chol', 'MaxHR', 'Oldpeak', 'AHD')])
```



## Hypothesis testing

Compare the mean RestBP with normal BP (120)?

```
t.test(data$RestBP, mu=120)
```

```
##
## One Sample t-test
##
## data: data$RestBP
## t = 11.562, df = 302, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 120
## 95 percent confidence interval:
## 129.7001 133.6794
## sample estimates:
## mean of x
## 131.6898
```

Compare mean RestBP by AHD?

```
RestBP_Yes = data[data$AHD == "Yes", "RestBP"]
RestBP_No = data[data$AHD == "No", "RestBP"]
t.test(RestBP_Yes, RestBP_No)
```

```
##
```

```
## Welch Two Sample t-test
##
## data: RestBP_Yes and RestBP_No
## t = 2.6152, df = 274.64, p-value = 0.009409
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.314915 9.321775
## sample estimates:
## mean of x mean of y
## 134.5683 129.2500
```

## Test the independence between AHD and Sex?

```
chisq.test(data$Sex, data$AHD)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: data$Sex and data$AHD
## X-squared = 22.043, df = 1, p-value = 2.667e-06
```

## Compare mean MaxHR by Thal

### ANOVA

```
anova <- aov(MaxHR ~ Thal, data=data)
summary(anova)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Thal           2  14396     7198   15.06 5.86e-07 ***
## Residuals    298 142389       478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

### Pairwise T-Test

```
pairwise.t.test(data$MaxHR, data$Thal)
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: data$MaxHR and data$Thal
##
##           fixed    normal
## normal      0.00036 -
## reversible 0.13359 1.5e-05
##
## P value adjustment method: holm
```

### Tukey multiple pairwise-comparisons

```
TukeyHSD(anova)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = MaxHR ~ Thal, data = data)
##
## $Thal
##              diff          lwr          upr          p adj
## normal-fixed    20.587684    7.810658  33.364710  0.0005225
## reversible-fixed  8.324786   -4.711349  21.360922  0.2903956
## reversible-normal -12.262898 -18.478131 -6.047665  0.0000150
```

## Linear Regression

### Build a model to predict MaxHR by Age?

```
model_1 <- lm(MaxHR ~ Age, data=data)
summary(model_1)

##
## Call:
## lm(formula = MaxHR ~ Age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -66.088 -12.040   3.965  15.937  44.955
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  203.8634     7.3991  27.553  < 2e-16 ***
## Age          -0.9966     0.1341  -7.433  1.11e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.06 on 301 degrees of freedom
## Multiple R-squared:  0.1551, Adjusted R-squared:  0.1523
## F-statistic: 55.25 on 1 and 301 DF, p-value: 1.109e-12
```

### Build a model to predict MaxHR by Age & RestBP & Thal?

```
model_2 <- lm(MaxHR ~ Age + RestBP + Thal, data=data)
summary(model_2)

##
## Call:
## lm(formula = MaxHR ~ Age + RestBP + Thal, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -70.046 -12.269   3.579  14.086  52.824
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  172.57393    11.48221  15.030  < 2e-16 ***
```

```

## Age          -0.98287    0.13478   -7.292  2.8e-12 ***
## RestBP       0.13314    0.06932    1.921  0.055720 .
## Thalnormal   18.34722    5.04476    3.637  0.000325 ***
## Thalreversible 7.69466    5.11709    1.504  0.133721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.19 on 296 degrees of freedom
## (2 observations deleted due to missingness)
## Multiple R-squared:  0.2301, Adjusted R-squared:  0.2197
## F-statistic: 22.12 on 4 and 296 DF,  p-value: 5.425e-16

```