

```

heart <- read.csv("C:/Users/teju1/OneDrive/Desktop/heart.txt", header=FALSE)

View(heart)

colnames(heart)<-
c('age','sex','cp','trestbps','chol','fbs','restecg','thalach','exang','oldpeak','slope','ca','thal','target')

head(heart)

plot(heart)

library(polycor)

corr <- hetcor(heart)

corr

hetcorr_corr <-corr$correlations

hetcorr_corr

#install.packages("corrplot")

library(corrplot)

corrplot(hetcorr_corr, type = "lower", order = "original",
         tl.col = "orange", tl.srt = 100, method = "number")

str(heart)


heart$sex <- as.factor(heart$sex)
heart$cp <- as.factor(heart$cp)
heart$fbs <- as.factor(heart$fbs)
heart$restecg <- as.factor(heart$restecg)
heart$exang <- as.factor(heart$exang)
heart$thal <- as.factor(heart$thal)
heart$slope <- as.factor(heart$slope)
heart$ca <- as.factor(heart$ca)
heart$target <- as.factor(heart$target)


#To check missing values in the dataset

sum(is.na(heart))

```

```
#install.packages("ggplot2")
```

```
library(ggplot2)
```

```
#install.packages("gridExtra")
```

```
library(gridExtra)
```

```
#Target and Thalach
```

```
thalach1<- ggplot(heart, aes(thalach, col=target, fill=target))+ geom_density(alpha=0.5)+  
guides(col=F)+ labs(fill="Target", x="Maximum heart rate achieved")
```

```
thalach1
```

```
thalach2<- ggplot(heart, aes(target, thalach, fill=target))+ geom_boxplot() + labs(y="Maximum  
heart rate achieved", x="Target", fill="Target")
```

```
thalach2
```

```
grid.arrange(thalach1, thalach2, nrow=1)
```

```
summary(heart$thalach)
```

```
#Target and chest pain (cp)
```

```
cp1<- ggplot(heart, aes(cp, fill=target)) + geom_bar(stat="count", position= "fill") +  
labs(x="Chest Pain Type", fill= "Target", y="Count")
```

```
cp1
```

#cp type 1 = Typical angina has the major cause of heart disease compared to the others, but cp type 2 and 3 has relationship to heart disease.

```
#Target and ca
```

```
ca1<- ggplot(heart, aes(ca, fill=target)) + geom_bar(stat="count", position= "fill") +  
labs(x="Number of vessels", fill= "Target", y="Count")
```

```
ca1
```

#0 and 4 vessels indicate cause of heart disease

#target and thal

```
thal1<- ggplot(heart, aes(thal, fill=target)) + geom_bar(stat="count", position= "fill") +  
labs(x="Thalium stress level test", fill= "Target", y="Count")
```

thal1

#The stress level of 2 indicates the major cause to have heart disease.

#Target and oldpeak

```
op1<- ggplot(heart, aes(oldpeak, col=target, fill=target))+ geom_density(alpha=0.5)+  
guides(col=F)+ labs(fill="Target", x="oldpeak")
```

op1

```
op2<- ggplot(heart, aes(target, oldpeak, fill=target))+ geom_boxplot() + labs(y="Oldpeak",  
x="Target", fill="Target")
```

op2

```
grid.arrange(op1, op2, nrow=1)
```

#Lower Oldpeak downsloping indicates heart disease

#Target and slope

```
slope1 <- ggplot(heart, aes(slope, fill=target)) + geom_bar(stat="count", position= "fill") +  
labs(x="Slope", fill= "Target", y="Count")
```

slope1

#slope of Level 2 = downsloping indicates heart disease.

#Target and exang

```
exang1<- ggplot(heart, aes(exang, fill=target)) + geom_bar(stat="count", position= "fill") +  
labs(x="Exercise Indiced angina", fill= "Target", y="Count")
```

exang1

#The level 0=no exercise angina indicate a count with patients with heart disease compared to level 1 = yes

#Target and sex

```
sex1<- ggplot(heart, aes(sex, fill=target)) + geom_bar(stat="count", position= "fill") +  
labs(x="Genre", fill= "Target", y="Count")
```

```
sex1
```

#women has more incidence of heart disease compared to men.

```
table(heart$sex)# In the dataset there are 713 men and 312 women
```

```
#Exang and cp
```

```
ex_cp <- ggplot(heart, aes(cp, fill=exang)) + geom_bar(stat="count", position= "fill") +  
labs(x="cp", fill= "exang", y="Count")
```

```
ex_cp
```

#For a typical angina there is more incidence to have an exercise induced angina

```
#Target and restecg
```

```
rest1<- ggplot(heart, aes(restecg, fill=target)) + geom_bar(stat="count", position= "fill") +  
labs(x="Resting ecg measurement", fill= "Target", y="Count")
```

```
rest1
```

#1 indicates having ST-T wave abnormality.

```
#oldpeak and slope
```

```
old<-ggplot(heart, aes(x=slope, y=oldpeak))+ geom_point(aes(colour=target)) +  
labs(y="Oldpeak", x="Slope", fill="Target")
```

```
old
```

#incidence of heart disease for patients with a slope of 2 and low oldpeak is more

```
grid.arrange(cp1,thalach1, thalach2, thal1, ca1, op1, op2, slope1, rest1, exang1, sex1, ex_cp, old,  
nrow=5)
```

```
grid.arrange
```

```
heart_mod1 <- glm(target ~., family = "binomial", data = heart)
```

```
summary(heart_mod1)
```

```
#reduced model
```

```
heart_mod_red <- step(heart_mod1, direction = "backward")
```

```
new_heart <- data.frame(exang=heart$exang, slope=heart$slope, trestbps=heart$trestbps,  
oldpeak=heart$oldpeak, thal=heart$thal, sex=heart$sex, cp= heart$cp, ca=heart$ca,  
target=heart$target)
```

```
heart_mod2 <- glm(target ~., family = "binomial", data = new_heart)
```

```
summary(heart_mod2)
```

```
#install.packages("MASS")
```

```
library(MASS)
```

```
exp(coef(heart_mod2))
```

```
exp(-1.63154) #if all regressors are held at a fixed value, the odds of getting heart disease for  
males (sex=1) over the odds of getting heart disease for females is  $\exp(-1.63154) = 0.1956$ 
```

```
lr.anova <- anova(heart_mod1, test="Chisq")
```

```
lr.anova
```

```
#In this model we see that the variables chol, fbs, and restecg are not significant for the model.
```

```
lr.anova_red <- anova(heart_mod2, test="Chisq")
```

```
lr.anova_red
```

```
#All the variables look significant for the model.
```

```
anova(heart_mod1, heart_mod2, test="Chisq")
```

```
na <- sum(is.na(heart))
```

```
na
```

```
library(caret)
```

```
set.seed(100)
```

```
train_data <- createDataPartition(new_heart$target, p=0.8, list=FALSE)
```

```
trainData<- new_heart[train_data,]
```

```
testData <- new_heart[-train_data, ]
```

```
trainData
```

```
testData
```

```
table(testData$target)
```

```
table(trainData$target)
```

```
set.seed(100)
```

```
heart_mod_train <- train(target ~., method='glm', family = "binomial", data = trainData )
```

```
heart_pred <- predict(heart_mod_train, testData)
```

```
heart_pred_prob <- predict(heart_mod_train, testData, type = "prob")
```

```
head(heart_pred_prob)
```

```
confMatr <- confusionMatrix(heart_pred, testData[,9])
```

```
confMatr
```

```
set.seed(50)
```

```
train_data2 <- createDataPartition(heart$target, p=0.8, list=FALSE)
```

```
trainData2<- heart[train_data2,]
```

```
testData2 <- heart[-train_data2, ]
```

```

table(testData2$target)
table(trainData2$target)

set.seed(10)
heart_mod_train2 <- train(target ~., method='glm', family = "binomial", data = trainData2 )
heart_pred2 <- predict(heart_mod_train2, newdata=testData2)
heart_pred_prob2 <- predict(heart_mod_train2, newdata=testData2, type = "prob")
head(heart_pred_prob2)

confMatr2 <- confusionMatrix(heart_pred2, testData2[,14])
confMatr2
table(heart$target) #in this data there are 1025 observation out of which 526 of them are
suffering from heart disease, and 499 of them are healthy.

library(pROC)

area1 <- roc(response=testData2$target, predictor=heart_pred_prob2$`1`)
plot(area1, legacy.axes=TRUE)
auc(area1)
area2 <- roc(response=testData$target, predictor=heart_pred_prob$`1`)
plot(area2, legacy.axes=TRUE)
auc(area2)

```