UE20CS312 - Data Analytics - Worksheet 2b : Multiple Linear Regression

PES University

Reshmi Pradeep, Dept. of CSE - PES2UG20CS270

2022-09-18

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(InformationValue)

data<-read.csv('got_characters.csv')

###Problem 1
nrow(data) #no. of characters
```

```
## [1] 1946
```

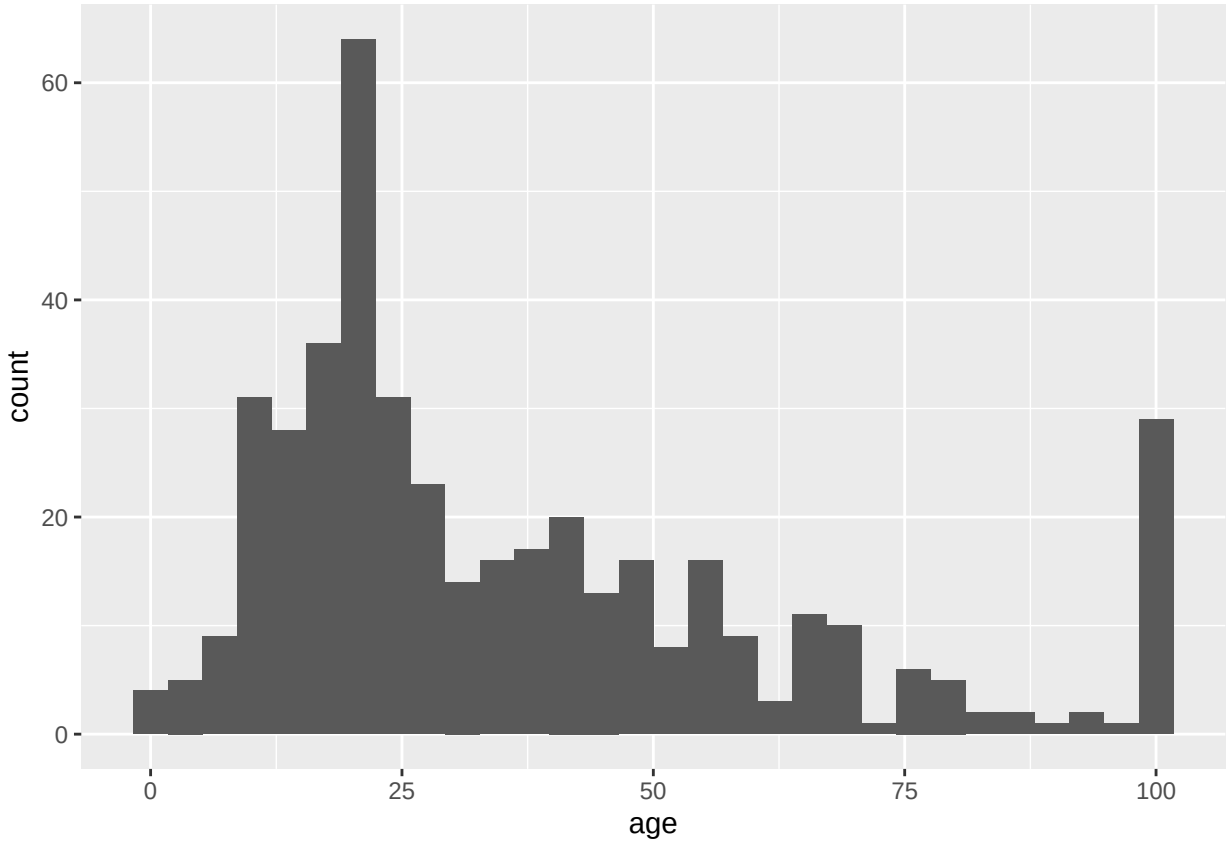
```
data[data==""] <- NA

naPercent<-(colMeans(is.na(data)))*100
coln<-colnames(data)
percentDf<-data.frame(coln,order(-naPercent))

View(percentDf)
View(data)

###Problem 2
#since columns with too many missing values are not useful, they are dropped, here by 80%
percentDf<-subset(percentDf,naPercent<80)

df<-subset(data, select=-c(mother, father, heir, spouse, isAliveMother, isAliveFather, isAliveHeir, isA
summary(is.na(df))
```

```
##      X              S.No            actual           name
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:1946      FALSE:1946      FALSE:1946      FALSE:1946
##
##    title            male          culture        dateOfBirth
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:938       FALSE:1946      FALSE:677       FALSE:433
##  TRUE :1008                      TRUE :1269      TRUE :1513
##    house            book1           book2           book3
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:1519      FALSE:1946      FALSE:1946      FALSE:1946
##  TRUE :427
##    book4            book5         isMarried        isNoble
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:1946      FALSE:1946      FALSE:1946      FALSE:1946
##
##     age           numDeadRelations boolDeadRelations isPopular
##  Mode :logical   Mode :logical    Mode :logical     Mode :logical
##  FALSE:433       FALSE:1946       FALSE:1946        FALSE:1946
##  TRUE :1513
##  popularity
##  Mode :logical
##  FALSE:1946
##
```

```
ggplot(df, aes(x=age)) + geom_histogram()
```

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

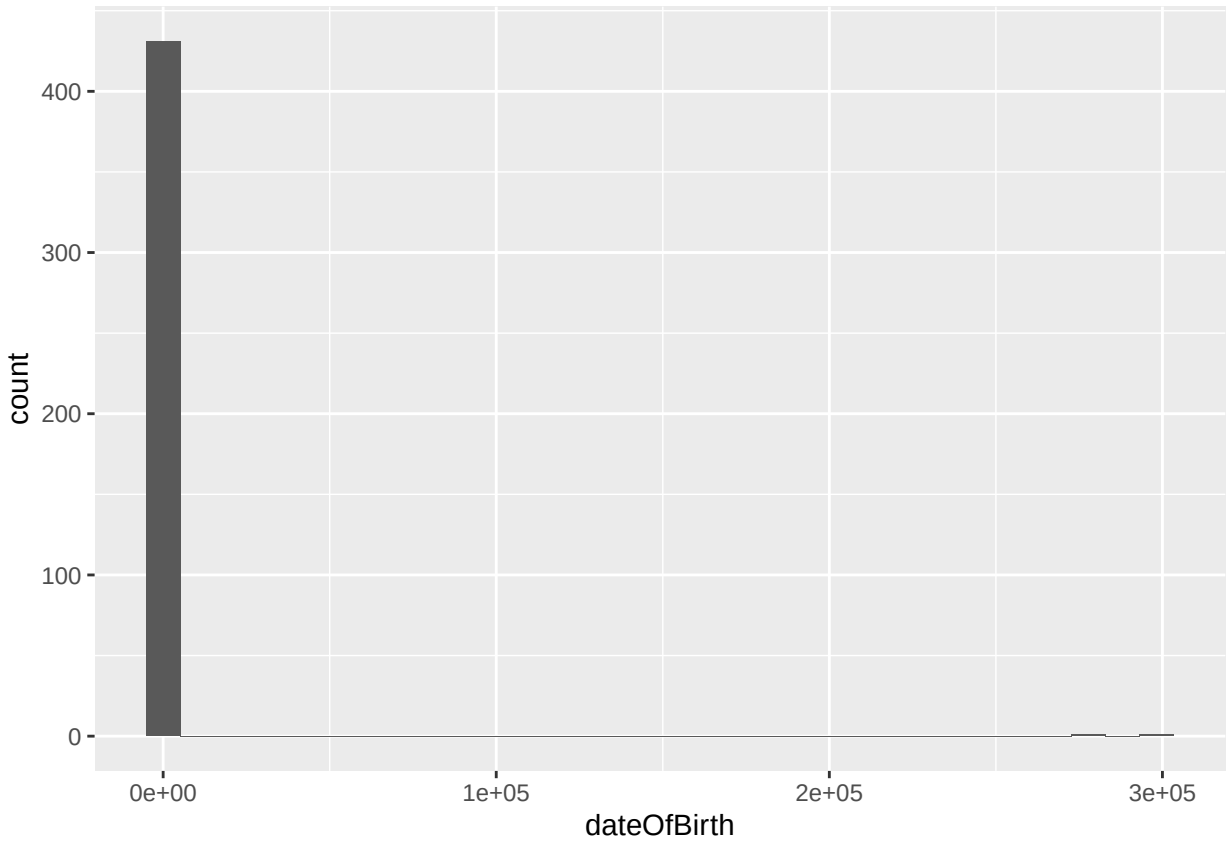
```
## Warning: Removed 1513 rows containing non-finite values (stat_bin).
```




```
ggplot(df, aes(x=dateOfBirth)) + geom_histogram()
```

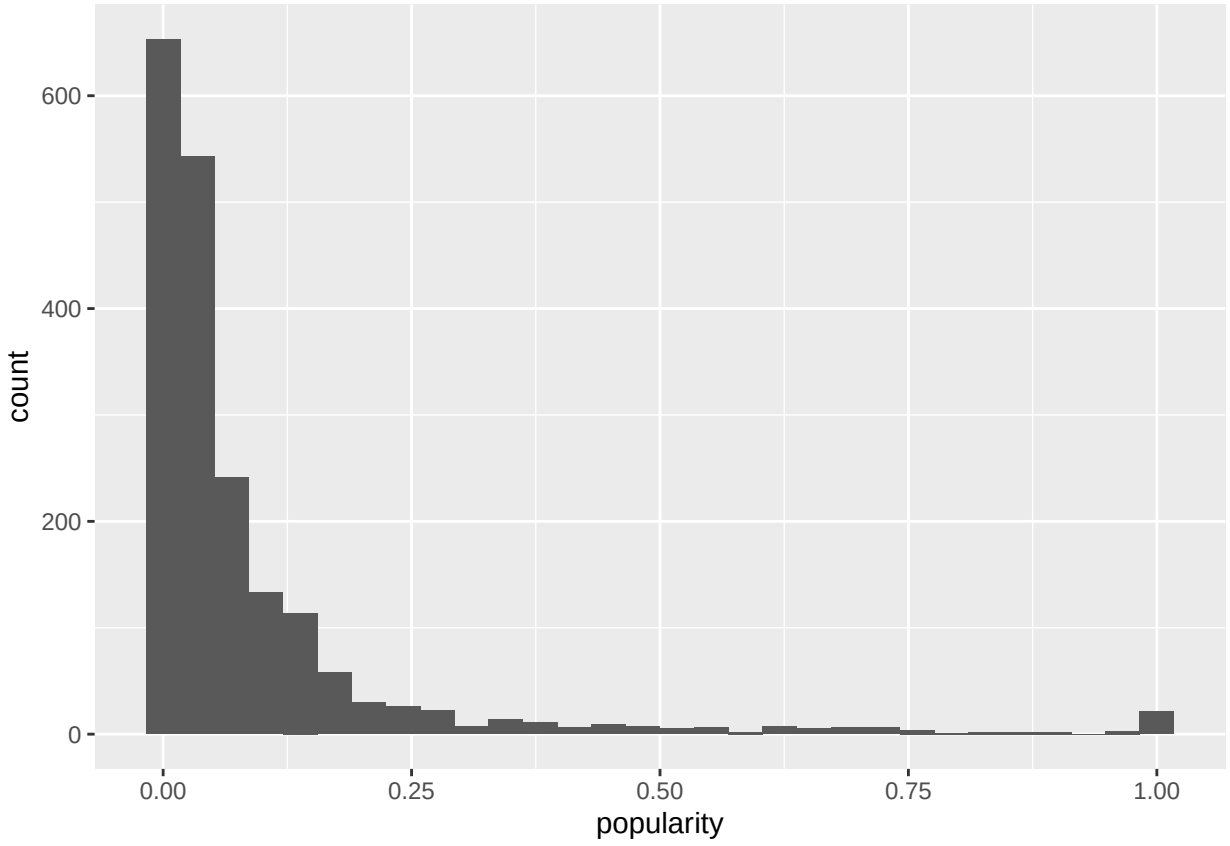
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1513 rows containing non-finite values (stat_bin).
```




```
ggplot(df, aes(x=popularity)) + geom_histogram()
```

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




```
#to fix discrepancy in age
agemedian<- median(df$age, na.rm=TRUE)
df$age[is.na(df$age)]<- agemedian
df$dateOfBirth[is.na(df$dateOfBirth)]<- -1

#converting categorical variables to numerical
df$house[is.na(df$house)]<- -1
df$title[is.na(df$title)]<- -1
df$culture[is.na(df$culture)]<- -1

x<-as.factor(df$house)
df$house<-unclass(x)

x<-as.factor(df$title)
df$title<-unclass(x)

x<-as.factor(df$culture)
df$culture<-unclass(x)


###Problem 3

table(df$actual)
```

```
##
##    0    1
```

```
##  495 1451
```

```
#it's not the same

ones<-df[which(df$actual== 1),]
zeros<-df[which(df$actual== 0),]

set.seed(123)
ones_trsample<- sample(1:nrow(ones), 0.7*nrow(zeros))
zeros_trsample <- sample(1:nrow(zeros), 0.7*nrow(zeros))

#training
trainOnes<-ones[ones_trsample,]
trainZeros<-zeros[zeros_trsample,]
trainDf<-rbind(trainOnes,trainZeros)
num1<-nrow(trainDf)
trainDf<-trainDf[sample(1:num1),]

#testing
testOnes<- ones[-ones_trsample,]
testZeros<-zeros[-zeros_trsample,]
testDf<-rbind(testOnes,testZeros)
num2<-nrow(testDf)
testDf<-testDf[sample(1:num2),]

#check
table(trainDf$actual)
```

```
##
##   0   1
## 346 346
```

```
table(testDf$actual)
```

```
##
##    0    1
##  149 1105
```

```
### Problem 4
lrm<-glm(actual ~ age + culture + male + book1 + isMarried + boolDeadRelations + isPopular + popularity
summary(lrm)
```

```
##
## Call:
## glm(formula = actual ~ age + culture + male + book1 + isMarried +
##     boolDeadRelations + isPopular + popularity, family = binomial(link = "logit"),
##     data = trainDf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7066  -1.1326   0.2118   1.0843   2.2258
##
```

```
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.547089   0.239597   6.457 1.07e-10 ***
## age                -0.024912   0.006215  -4.008 6.12e-05 ***
## culture            -0.023358   0.007438  -3.140 0.001687 **
## male               -0.618398   0.176944  -3.495 0.000474 ***
## book1              -0.496420   0.204061  -2.433 0.014986 *
## isMarried          -0.067317   0.249135  -0.270 0.787005
## boolDeadRelations -0.624110   0.378678  -1.648 0.099326 .
## isPopular           0.435407   0.647427   0.673 0.501253
## popularity         -0.727168   1.091560  -0.666 0.505300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 959.32  on 691  degrees of freedom
## Residual deviance: 884.41  on 683  degrees of freedom
## AIC: 902.41
##
## Number of Fisher Scoring iterations: 4
```

```
predicted<-plogis(predict(lrm, testDf))

cutoff<-optimalCutoff(testDf$actual, predicted)[1]
cutoff
```

```
## [1] 0.07380093
```

```
### Problem 5
misClassError(testDf$actual, predicted, threshold=cutoff)
```

```
## [1] 0.118
```

```
confusionMatrix(testDf$actual, predicted, threshold=cutoff)
```

```
##     0    1
## 0   3    2
## 1 146 1103
```

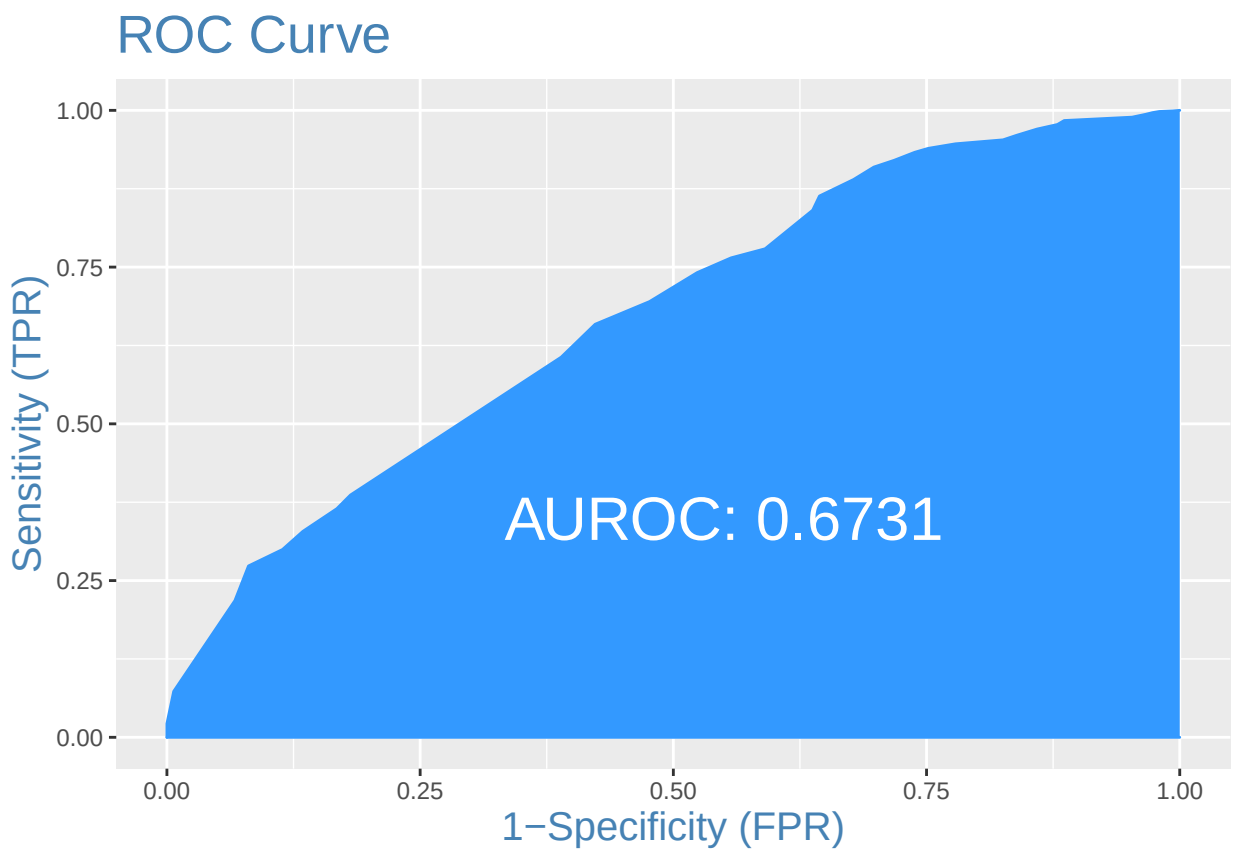
```
sensitivity(testDf$actual, predicted, threshold=cutoff)
```

```
## [1] 0.99819
```

```
specificity(testDf$actual, predicted, threshold=cutoff)
```

```
## [1] 0.02013423
```

```
plotROC(testDf$actual, predicted)
```




```
#area under the curve is 0.6731
```