Indian_Liver_Disease_final_edx_001

Frederic Ouedraogo

2022-10-10

INTRODUCTION

According to the study description available here: https://www.kaggle.com/datasets/uciml/indian-liver-patient-records?select=indian_liver_patient.csv, there has been a continuous rise in the prevalence for liver disease in a local community in India. A list of potential suspects has been established including excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. For decision makers to elaborate effective policies, they need to have data-driven evidence that the overmentioned suspects are in fact contributing factors. A dataset consisting of 583 hospital patients was gathered. The dataset contains some demographic information (gender and age) along with some biological test results (level of bilirubin, alkaline phosphotase, alamine aminotransferase, aspartate aminotransferase, proteins, albumin, and the ratio of albumin to globulin) and a diagnosis of liver disease. For the later, the outcome is a binary variable taking the value of 1 if the patient has a liver disease and 0 otherwise.

As a data scientist, I was given the opportunity to analyze the data and help the authorities handle the issue. My role is to come up with a model that is not only capable of identifying the driving factors, but more importantly, capable of predicting with high accuracy individuals within the community with higher probabilities of being diagnosed positive to the disease. That will allow decision makers to identify their high-risk individuals, to develop and deploy appropriate strategies and policies. Because of the nature of the response variable (binary variable), I relied on logit model which is more powerful in handling such type of response variable.

All analyses were performed in R studio. Machine learning was used for the statistical analyses. Dataset was divided into training set (75% or the data) and testing set (25%). Several models were estimated, and the AIC was used as a criterion for best model selection. Prediction was performed using the best model and the confusion matrix was used to test the performance of the prediction.

The results of the logistic regression suggest that only gender (p-value = 0.0171) and level of alamine_Aminotransferase (p-value = 0.0425) have significant effects on one likelihood of developing liver disease. We did not find enough evidence that albumin, age, albumin_and_globulin_Ratio, total_proteins, direct_bilirubin, alkaline_phosphotase, and total_bilirubin affect odds of being tested positive. We performed a prediction using the testing data and the accuracy result from the confusion Matrix suggests that our model is correct in 64% of the test cases. In other words, a true positive and true negative occur in 64% of cases.

METHODS

The dataset was downloaded from kaggle (https://www.kaggle.com/datasets/uciml/indian-liver-patient-records?select=indian_liver_patient.csv). A description of the variables in the dataset is presented in the Table 1. Most variables are positive numeric (Num) meaning that they take values of 0 or above. Gender is a binary (Female and Male) gender variable. The result from the liver diagnossis is a numeric variable taking the value of 1 if the result is positive and 0 otherwise.

Table 1. Description of the variables in the dataset.

(Table here)

Exploratory Data Analysis

To have a better understanding of the data, we performed a series of exploratory data analyses. A quick overview of the variables shows some concerning values that could potentially influence the results of the analyis. Total_Bilirubin, Direct_Bilirubin, Alkaline_phosphotase, Alamine_aminotransferase, Aspartate_aminotransferase carry some potential outlying values. Both visual and statistical tests were performed to identify and remove outliers. The visual tests include histogram and scatterplot. The Grubbs test was used as a statistical test for detecting outliers. In addition, Albumin_and_globulin_Ratio has 4 missing values that needs to be taken care of. Age of the patients ranges from 4 to 90 years with a typical patient being about 44 years of age (Table 2). About 3% of the sample are 15 years or below, 48.5% are between 16 and 45 years, 38.3% between 46 and 65 years, 10.1% between 66 and 85 years, and less than 1% are 86 years or older (Table 3). Women (43 years) are relatively younger than men (45 years).

Of the study sample, 456 (67%) are women and men represent 43% of the sample. The average level of total bilirubin, alkaline phosphotase, alamine aminotransferase, aspartate aminotransferase, total proteins, and albumin is 3.3, 290.6, 80.7, 109.9, 6.5, 3.1, and 0.9, respectively. The standard deviations, the minimum and the maximum values, as well as the inter quartile ranges (IQR) are also presented in the table below. Clearly, the large gap between mean and median values for some variables imply that the data are somehow skewed due to potential outliers.

Table 2. Summary statistics of the patients

(Table here)

Accessing the data from my github

```
data_stat <- read.csv("https://raw.githubusercontent.com/fdraogo/Indian-
Liver-Patient-Records-Analysis/main/Indian-Liver-Patient-Records-
Analysis.csv")
```

Installing packages for exploratory data analytics

The first sets of analysis performed are to explore the data. This will alllow a better understanding of the data, identify influencial values that could impact the results of the regression and pull insights that could lead to an efficient identification of the model that will be used to fit the data. Insights are presented in the form of table or chart that summarizes the key findings. The packages installed are tidyverse, ggplot2, dplyr, and stringr.

```
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.1
## ✓ readr   2.1.2     ✓ forcats 0.5.1
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()

library(ggplot2)
library(dplyr)
library(stringr)
```

Overview of the dataset

This will provide us with key indicators such as means, medians, first and third quartiles, minimum and maximum values of all numeric variables in the dataset. Data of 583 patients (Women = 142, Men = 441) were in the dataset.

```
summary(data_stat)

##       Age            Gender          Total_Bilirubin  Direct_Bilirubin
##  Min.   : 4.00   Length:583         Min.   : 0.400   Min.   : 0.100
##  1st Qu.:33.00   Class :character   1st Qu.: 0.800   1st Qu.: 0.200
##  Median :45.00   Mode  :character   Median : 1.000   Median : 0.300
##  Mean   :44.75                      Mean   : 3.299   Mean   : 1.486
##  3rd Qu.:58.00                      3rd Qu.: 2.600   3rd Qu.: 1.300
##  Max.   :90.00                      Max.   :75.000   Max.   :19.700
##
##  Alkaline_Phosphotase Alamine_Aminotransferase Aspartate_Aminotransferase
##  Min.   :  63.0       Min.   :  10.00          Min.   :  10.0
##  1st Qu.: 175.5       1st Qu.:  23.00          1st Qu.:  25.0
##  Median : 208.0       Median :  35.00          Median :  42.0
##  Mean   : 290.6       Mean   :  80.71          Mean   : 109.9
##  3rd Qu.: 298.0       3rd Qu.:  60.50          3rd Qu.:  87.0
##  Max.   :2110.0       Max.   :2000.00          Max.   :4929.0
##
##  Total_Protiens     Albumin      Albumin_and_Globulin_Ratio    Dataset
##  Min.   :2.700   Min.   :0.900   Min.   :0.3000             Min.   :1.000
##  1st Qu.:5.800   1st Qu.:2.600   1st Qu.:0.7000             1st Qu.:1.000
##  Median :6.600   Median :3.100   Median :0.9300             Median :1.000
##  Mean   :6.483   Mean   :3.142   Mean   :0.9471             Mean   :1.286
##  3rd Qu.:7.200   3rd Qu.:3.800   3rd Qu.:1.1000             3rd Qu.:2.000
##  Max.   :9.600   Max.   :5.500   Max.   :2.8000             Max.   :2.000
##                                  NA's   :4
##  liver_disease
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2864
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

Gender distribution of patients

```
table (data_stat$Gender)
```

```
##
## Female   Male
##    142    441
```

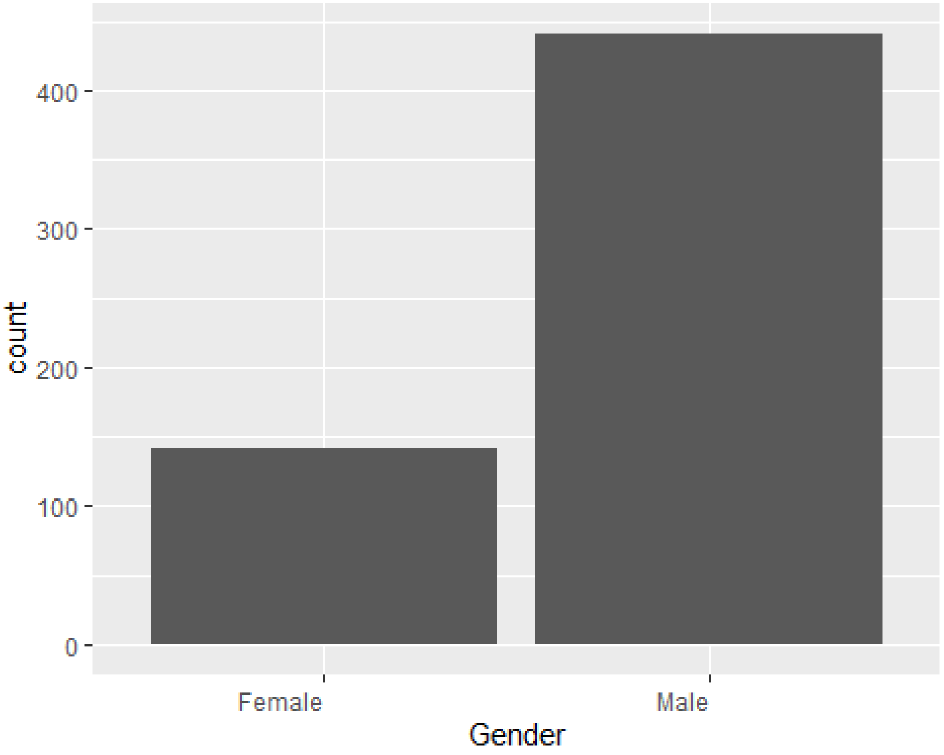
Distribution of patients by gender and liver disease status (positive = 1, negative = 0)

```
table (data_stat$Gender, data_stat$liver_disease)
```

```
##
##             0   1
##   Female  92  50
##   Male   324 117
```

Summary of the patients by gender

```
ggplot(data_stat, aes(x = Gender)) +
        geom_bar() +
        theme(axis.text.x = element_text(hjust = 1))
```



How many people were tested positive? by gender and age group

```r
data_stat <- data_stat %>%
  mutate(
    Age_group = dplyr::case_when(
      Age <= 15                 ~ "0-15",
      Age > 15 & Age <= 45 ~ "16-45",
      Age > 45 & Age <= 65 ~ "46-65",
      Age > 65 & Age <= 85 ~ "66-85",
      Age > 85                  ~ "> 85"
    ),

    Age_group = factor(
      Age_group,
      level = c("0-15", "16-45","46-65", "66-85","> 85")
    )
  )
```

Age distribution of each sub-population in the dataset

The distribution presented in the figure below suggests that both women and men populations have the same characteristics with the mode being at age group 16 to 45 years and the top three groups being “16-45”, “46-65”, and “66-85”, respectively.

```
ggplot(data_stat, aes(x = Age_group)) +
        geom_bar(aes(y = (..count..)/sum(..count..))) +
        xlab("Age groups of patients") +
        scale_y_continuous(labels = scales::percent, name = "Proportion") +
        facet_grid(~ Gender) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

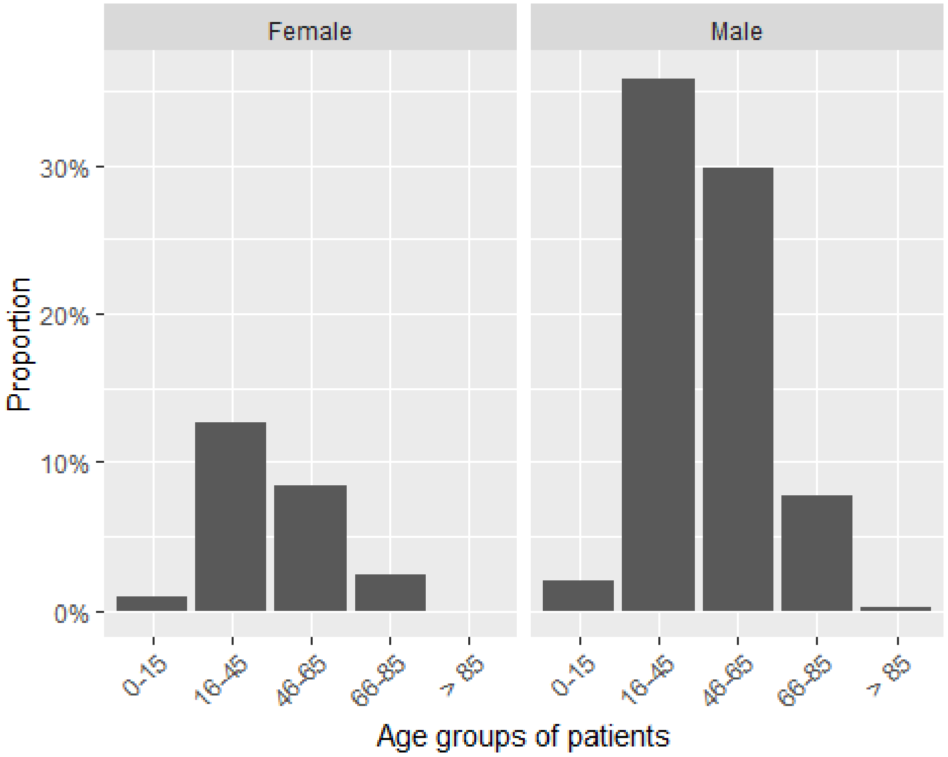


Table 3. Summarize the sample by age group

(Table here)

Mean age of the patients

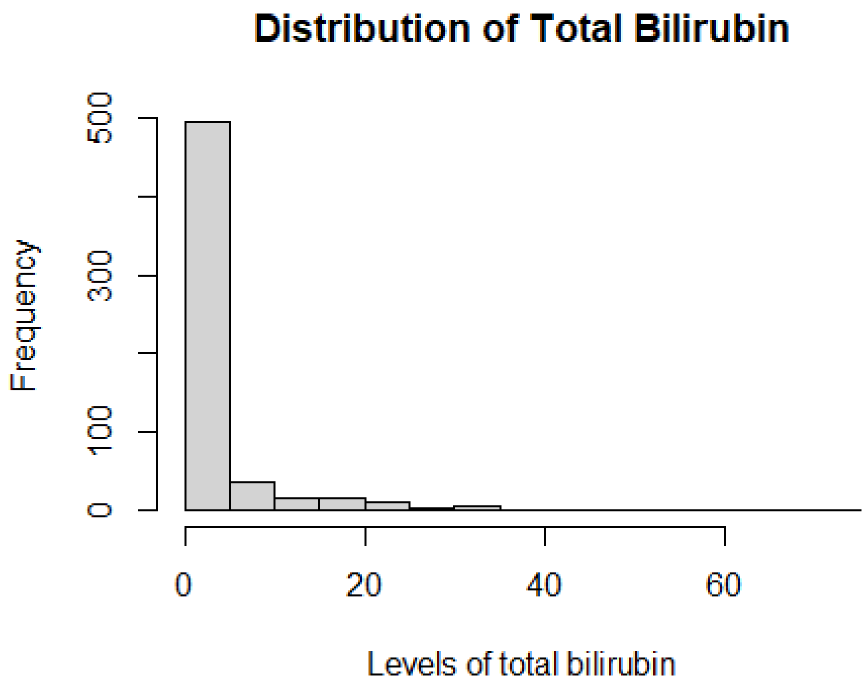
The mean age of the patients is 44 years with women (43) being relatively younger than men (45).

```
mean(data_stat$Age)

## [1] 44.74614

aggregate(x = data_stat$Age,
          by = list(data_stat$Gender),
          FUN = mean)
```

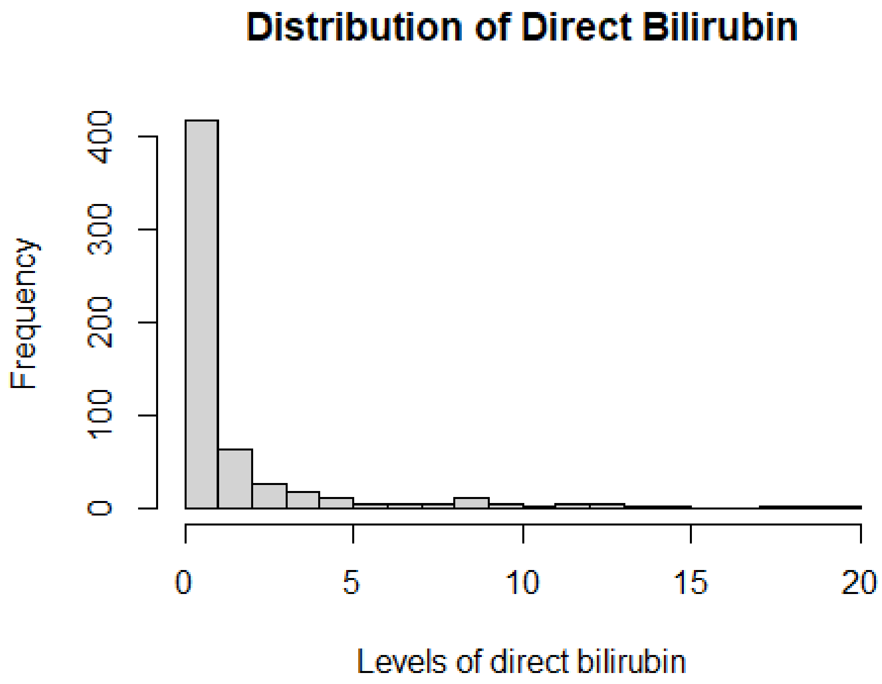
```
##   Group.1        x
## 1  Female 43.13380
## 2    Male 45.26531
```

Data visualization and identification of possible outlying values

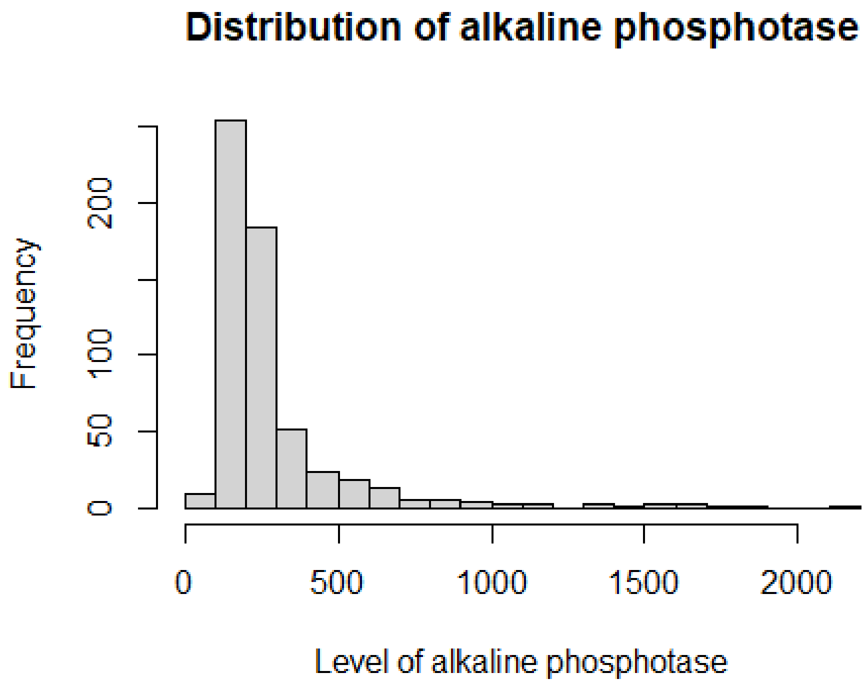
```
hist(data_stat$Total_Bilirubin,
  xlab = "Levels of total bilirubin",
  main = "Distribution of Total Bilirubin",
  border = "black",
  breaks = sqrt(nrow(data_stat))
)
```



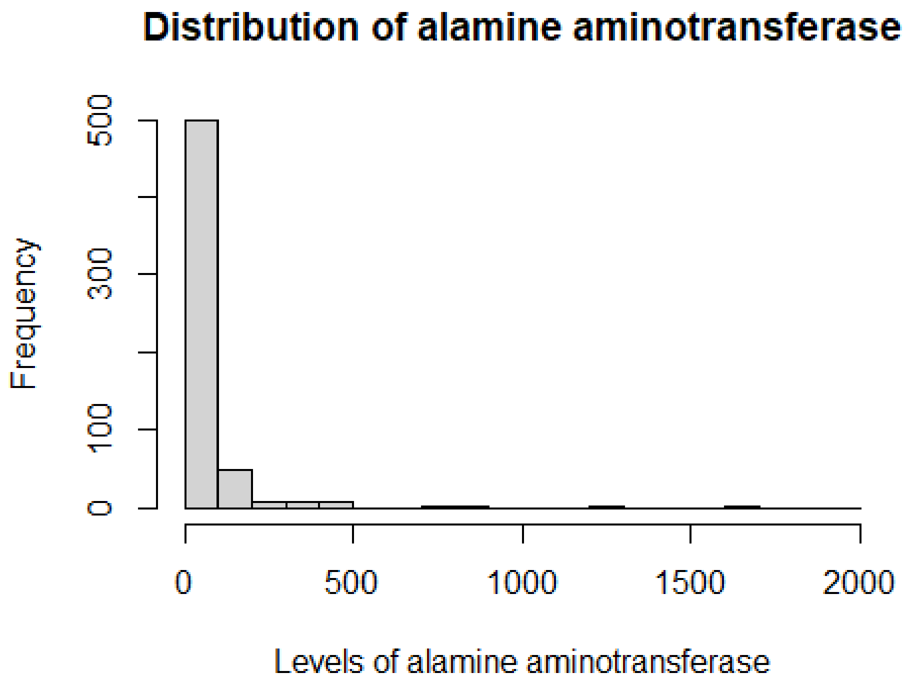
```
hist(data_stat$Direct_Bilirubin,
  xlab = "Levels of direct bilirubin",
  main = "Distribution of Direct Bilirubin",
  border = "black",
  breaks = sqrt(nrow(data_stat))
)
```



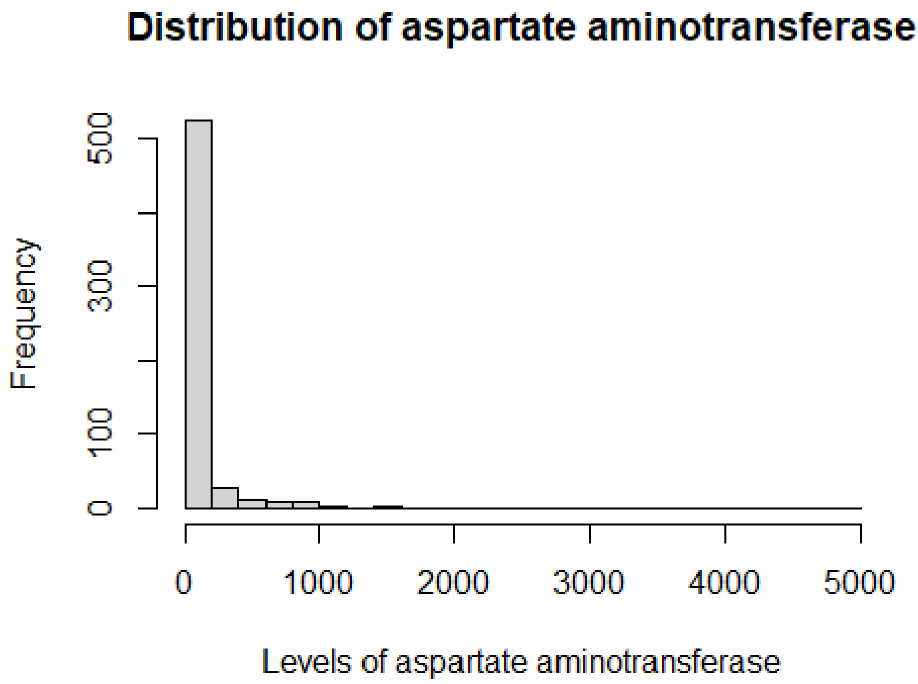
```
hist(data_stat$Alkaline_Phosphotase,
  xlab = "Level of alkaline phosphotase",
  main = "Distribution of alkaline phosphotase",
  border = "black",
  breaks = sqrt(nrow(data_stat))
)
```



```r
hist(data_stat$Alamine_Aminotransferase,
  xlab = "Levels of alamine aminotransferase",
  main = "Distribution of alamine aminotransferase",
  border = "black",
  breaks = sqrt(nrow(data_stat))
)
```



```r
hist(data_stat$Aspartate_Aminotransferase,
  xlab = "Levels of aspartate aminotransferase",
  main = "Distribution of aspartate aminotransferase",
  border = "black",
  breaks = sqrt(nrow(data_stat))
)
```



More formal identification of possible outlying values

```
library(outliers)
```

identifying outliers using the Grubbs test statistics

The choice of Grubbs test is because of the large sample size that we are dealing with and also because the minimum values in our dataset do not seem to be of any concern. However, the maximum values are most often far away from the rest of the data points. In all of our suspect factors, the maximum values are in fact outliers as confirmed by the Grubbs' test.

```
test_Total_bilirubin  <- grubbs.test(data_stat$Total_Bilirubin)
test_Direct_bilirubin  <- grubbs.test(data_stat$Direct_Bilirubin)
test_Alkaline_phosphotase  <- grubbs.test(data_stat$Alkaline_Phosphotase)
test_Alamine_aminotransferase  <-
grubbs.test(data_stat$Alamine_Aminotransferase)
test_Aspartate_aminotransferase <-
grubbs.test(data_stat$Aspartate_Aminotransferase)


test_Total_bilirubin

##
##  Grubbs test for one outlier
```

```
##
## data:  data_stat$Total_Bilirubin
## G = 11.54698, U = 0.77051, p-value < 2.2e-16
## alternative hypothesis: highest value 75 is an outlier

test_Direct_bilirubin

##
##  Grubbs test for one outlier
##
## data:  data_stat$Direct_Bilirubin
## G = 6.48528, U = 0.92761, p-value = 1.161e-08
## alternative hypothesis: highest value 19.7 is an outlier

test_Alkaline_phosphotase

##
##  Grubbs test for one outlier
##
## data:  data_stat$Alkaline_Phosphotase
## G = 7.48925, U = 0.90346, p-value = 4.79e-12
## alternative hypothesis: highest value 2110 is an outlier

test_Alamine_aminotransferase

##
##  Grubbs test for one outlier
##
## data:  data_stat$Alamine_Aminotransferase
## G = 10.50971, U = 0.80989, p-value < 2.2e-16
## alternative hypothesis: highest value 2000 is an outlier

test_Aspartate_aminotransferase

##
##  Grubbs test for one outlier
##
## data:  data_stat$Aspartate_Aminotransferase
## G = 16.67975, U = 0.52115, p-value < 2.2e-16
## alternative hypothesis: highest value 4929 is an outlier
```

Cleaning the data and removing outliers and observations with missing values

A quantile approach was used to identify and remove influential values from the dataset.

```
Q1 <- quantile(data_stat$Total_Bilirubin, probs=c(.25, .75), na.rm = FALSE)
Q2 <- quantile(data_stat$Direct_Bilirubin, probs=c(.25, .75), na.rm = FALSE)
Q3 <- quantile(data_stat$Alkaline_Phosphotase, probs=c(.25, .75), na.rm =
FALSE)
```

```
Q4 <- quantile(data_stat$Alamine_Aminotransferase, probs=c(.25, .75), na.rm =
FALSE)
Q5 <- quantile(data_stat$Aspartate_Aminotransferase, probs=c(.25, .75), na.rm
= FALSE)


iqr1 <- IQR(data_stat$Total_Bilirubin)
iqr2 <- IQR(data_stat$Direct_Bilirubin)
iqr3 <- IQR(data_stat$Alkaline_Phosphotase)
iqr4 <- IQR(data_stat$Alamine_Aminotransferase)
iqr5 <- IQR(data_stat$Aspartate_Aminotransferase)


eliminated1<- subset(data_stat, data_stat$Total_Bilirubin > (Q1[1] -
1.5*iqr1) & data_stat$Total_Bilirubin < (Q1[2]+1.5*iqr1))
eliminated2<- subset(data_stat, data_stat$Direct_Bilirubin > (Q2[1] -
1.5*iqr2) & data_stat$Direct_Bilirubin < (Q2[2]+1.5*iqr2))
eliminated3<- subset(data_stat, data_stat$Alkaline_Phosphotase > (Q3[1] -
1.5*iqr3) & data_stat$Alkaline_Phosphotase < (Q3[2]+1.5*iqr3))
eliminated4<- subset(data_stat, data_stat$Alamine_Aminotransferase > (Q4[1] -
1.5*iqr4) & data_stat$Alamine_Aminotransferase < (Q4[2]+1.5*iqr4))
eliminated5<- subset(data_stat, data_stat$Aspartate_Aminotransferase > (Q5[1]
- 1.5*iqr5) & data_stat$Aspartate_Aminotransferase < (Q5[2]+1.5*iqr5))


dat<- data_stat %>% filter(Total_Bilirubin >= 0.4 & Total_Bilirubin <= 5.3,
                Direct_Bilirubin >= 0.1 & Direct_Bilirubin <= 4.2,
                Alkaline_Phosphotase >= 63.0 & Alkaline_Phosphotase <= 187.0,
                Alamine_Aminotransferase >= 10 & Alamine_Aminotransferase <=
58,
                Aspartate_Aminotransferase >= 10 & Aspartate_Aminotransferase
<= 95)

datfred <-na.omit(dat)
```

Overview of the new dataset (after removing outliers and NAs)

```
summary(datfred)

##       Age            Gender          Total_Bilirubin Direct_Bilirubin
##  Min.   :13.00   Length:166         Min.   :0.50    Min.   :0.1000
##  1st Qu.:31.25   Class :character   1st Qu.:0.70    1st Qu.:0.2000
##  Median :45.00   Mode  :character   Median :0.80    Median :0.2000
##  Mean   :44.48                      Mean   :1.08    Mean   :0.3735
##  3rd Qu.:58.00                      3rd Qu.:1.00    3rd Qu.:0.3000
##  Max.   :78.00                      Max.   :5.30    Max.   :2.3000
##  Alkaline_Phosphotase Alamine_Aminotransferase Aspartate_Aminotransferase
##  Min.   : 63.0        Min.   :10.00            Min.   :11.00
##  1st Qu.:146.2        1st Qu.:20.00            1st Qu.:21.00
##  Median :162.0        Median :25.00            Median :27.50  
##  Mean   :158.6        Mean   :26.81            Mean   :32.49  
##  3rd Qu.:175.0        3rd Qu.:32.00            3rd Qu.:40.00  
##  Max.   :187.0        Max.   :56.00            Max.   :92.00  
##  Total_Protiens     Albumin      Albumin_and_Globulin_Ratio    Dataset     
##  Min.   :2.800   Min.   :1.400   Min.   :0.400              Min.   :1.000  
##  1st Qu.:5.900   1st Qu.:2.825   1st Qu.:0.900              1st Qu.:1.000  
##  Median :6.550   Median :3.400   Median :1.000              Median :1.000  
##  Mean   :6.458   Mean   :3.304   Mean   :1.048              Mean   :1.452  
##  3rd Qu.:7.100   3rd Qu.:3.900   3rd Qu.:1.200              3rd Qu.:2.000  
##  Max.   :8.900   Max.   :4.900   Max.   :1.800              Max.   :2.000  
##  liver_disease    Age_group
##  Min.   :0.0000   0-15 : 1  
##  1st Qu.:0.0000   16-45:90  
##  Median :0.0000   46-65:59  
##  Mean   :0.4518   66-85:16  
##  3rd Qu.:1.0000   > 85 : 0  
##  Max.   :1.0000
```

Installing package for categorical data analysis

```
library(bitops)
```

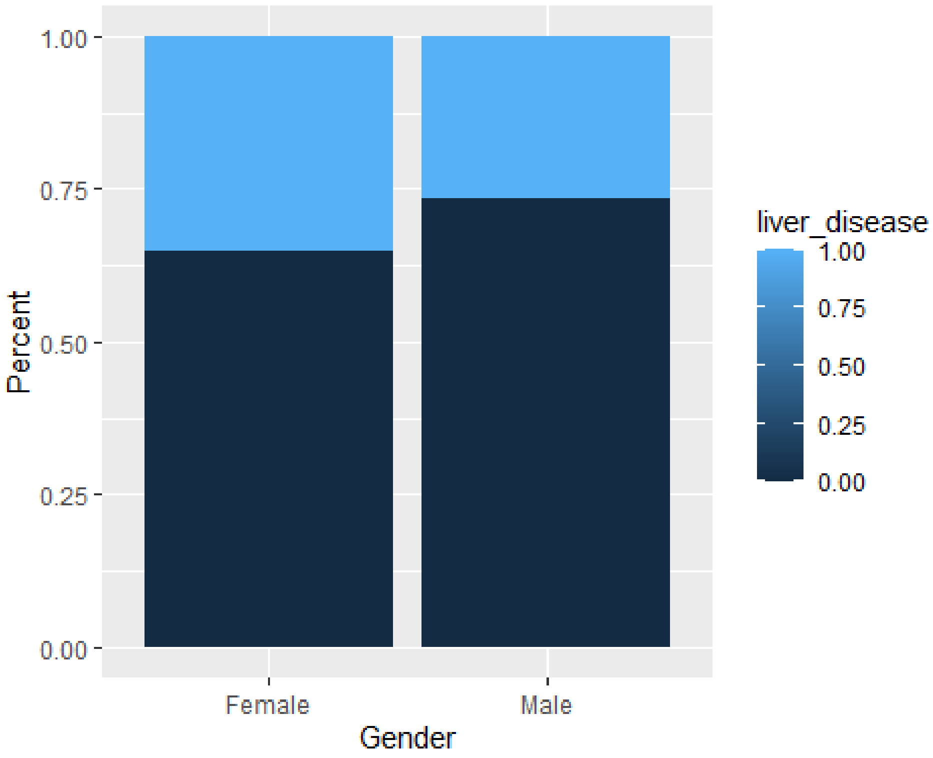
proportion of positive by gender

50 females (35.2%) were diagnosed with liver disease and 117 males (26.5%) were also found positive. From these statistics, females tend to have higher prevalence of liver disease than their male counterparts. However, given that we have more female in the study sample than male, a statistical test is required to test the hypothesis that women is more likely to be positive to liver disease than men. For this, we used the Chi-square test and the Fisher test. The Fisher test did not find enough evidence (p-value = 0.0546) that women have higher odds of being diagnosed with liver disease compared to men. After cleaning the data for potential outliers, a total of 166 patients (women = 53 and men = 113) were included in the analysis. The prevalence for women was 39.6% and that for men was 47.8%.

```
table(data_stat$liver_disease, data_stat$Gender)

## 
##     Female Male
##   0     92  324
##   1     50  117

table(datfred$liver_disease, datfred$Gender)

## 
##     Female Male
##   0     32   59
##   1     21   54
```

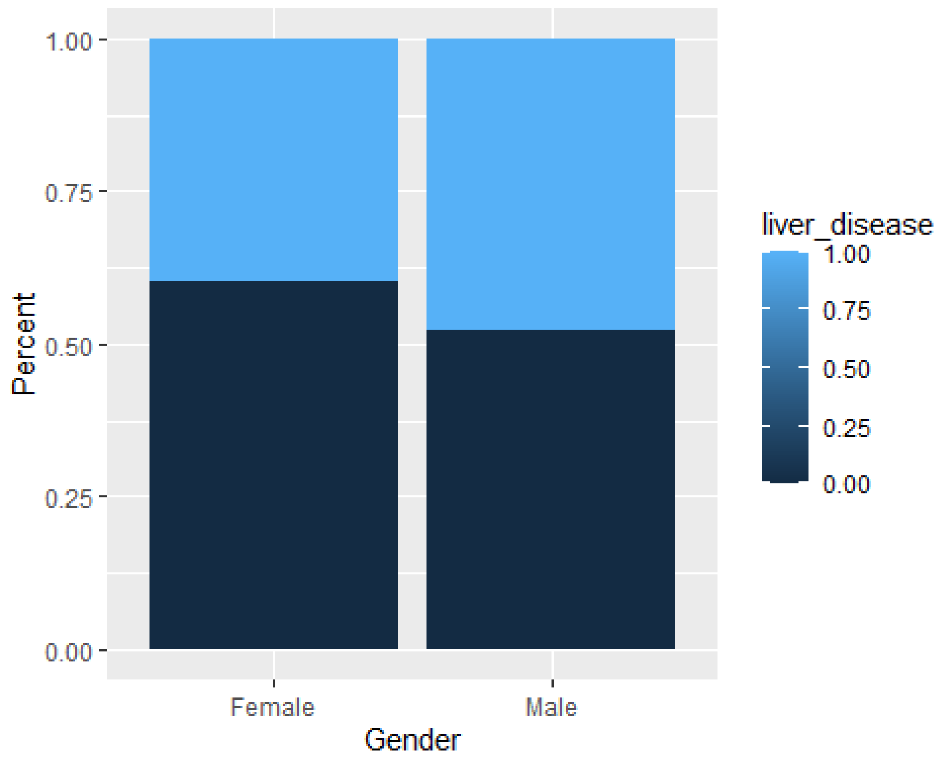
```
data_stat  %>%
  group_by(Gender, liver_disease)%>%
  summarize(n = n()) %>%
  mutate(Percent = n/sum(n)) %>%
  ggplot() +
  geom_col(aes(x = Gender, y = Percent, fill = liver_disease))
```

```
## `summarise()` has grouped output by 'Gender'. You can override using the
## `.groups` argument.
```



```
datfred  %>%
  group_by(Gender, liver_disease)%>%
  summarize(n = n()) %>%
  mutate(Percent = n/sum(n)) %>%
  ggplot() +
  geom_col(aes(x = Gender, y = Percent, fill = liver_disease))
```

```
## `summarise()` has grouped output by 'Gender'. You can override using the
## `.groups` argument.
```



Using a fisher test to see if women have high prevalence of liver disease

```
fisher.test(table(datfred$liver_disease, datfred$Gender), alternative =
c("two.sided"))

##
##  Fisher's Exact Test for Count Data
##
## data:  table(datfred$liver_disease, datfred$Gender)
## p-value = 0.4032
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.6844667 2.8702112
## sample estimates:
## odds ratio
##   1.391876
```

Using a fisher test to see if women have high prevalence of liver disease

```
fisher.test(table(datfred$liver_disease, datfred$Gender), alternative =
c("greater"))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(datfred$liver_disease, datfred$Gender)
## p-value = 0.2069
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.7598072       Inf
## sample estimates:
## odds ratio
##   1.391876
```

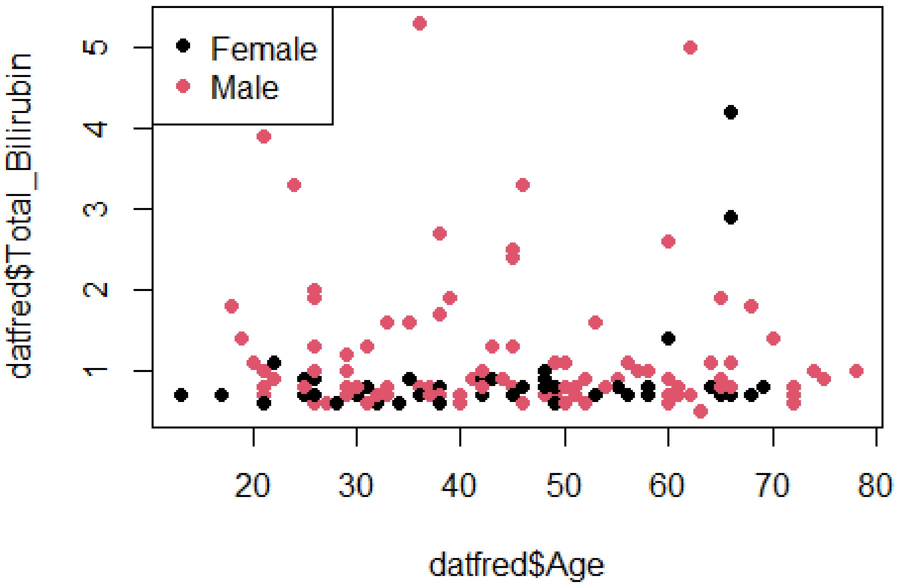
Using a Chi-square test to see if women have high prevalence of liver disease

```
chisq.test(table(datfred$liver_disease, datfred$Gender))
```

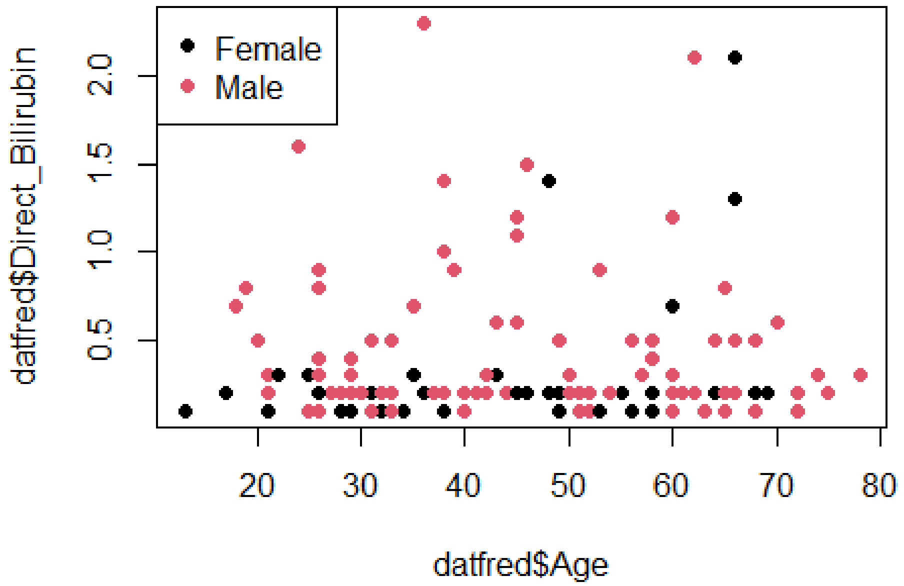
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(datfred$liver_disease, datfred$Gender)
## X-squared = 0.66943, df = 1, p-value = 0.4133
```

A scatterplot to see if outliers persisted and if gender and age could explain their presence.

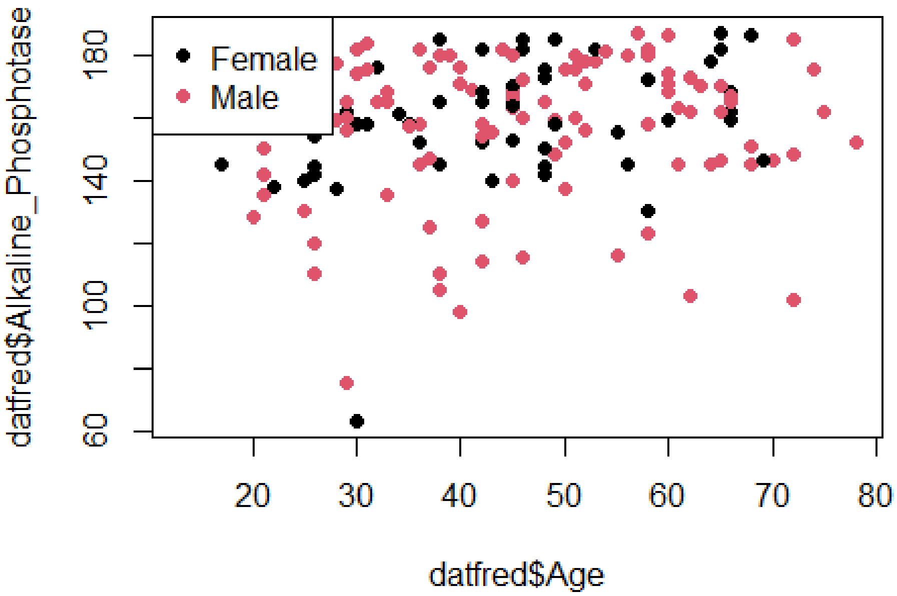
```
plot(datfred$Age, datfred$Total_Bilirubin,
     pch = 19,
     col = factor(datfred$Gender))

legend("topleft",
       legend = levels(factor(datfred$Gender)),
       pch = 19,
       col = factor(levels(factor(datfred$Gender))))
```



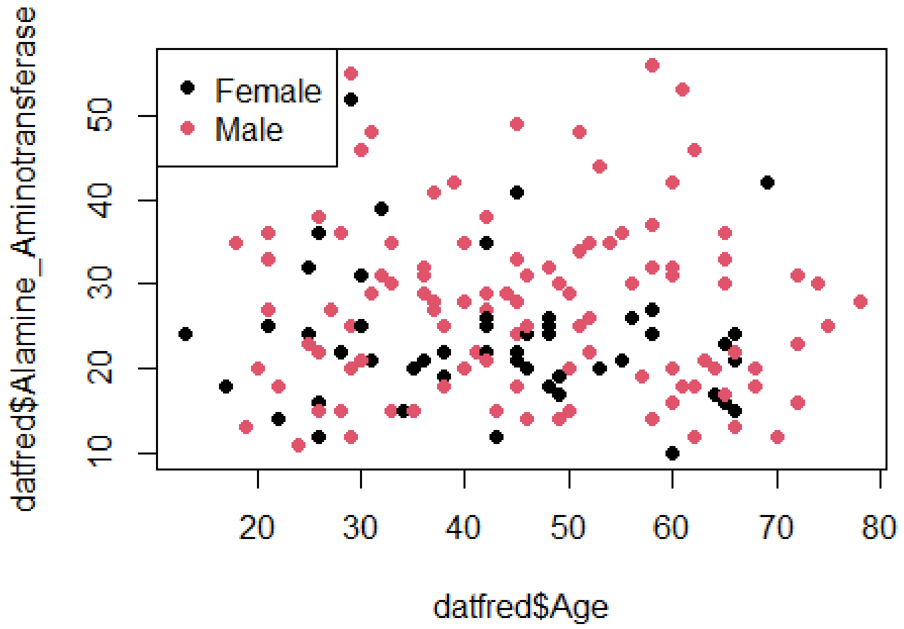
```
plot(datfred$Age, datfred$Direct_Bilirubin ,
     pch = 19,
     col = factor(datfred$Gender))

legend("topleft",
       legend = levels(factor(datfred$Gender)),
       pch = 19,
       col = factor(levels(factor(datfred$Gender))))
```



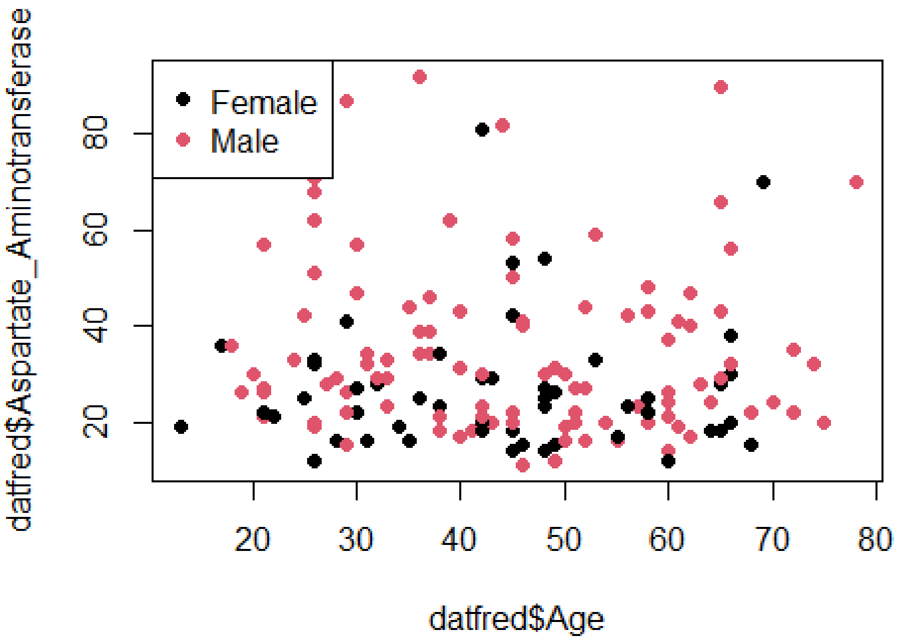
```
plot(datfred$Age, datfred$Alkaline_Phosphotase,
     pch = 19,
     col = factor(datfred$Gender))

legend("topleft",
       legend = levels(factor(datfred$Gender)),
       pch = 19,
       col = factor(levels(factor(datfred$Gender))))
```



```
plot(datfred$Age, datfred$Alamine_Aminotransferase,
     pch = 19,
     col = factor(datfred$Gender))

legend("topleft",
       legend = levels(factor(datfred$Gender)),
       pch = 19,
       col = factor(levels(factor(datfred$Gender))))
```



```
plot(datfred$Age, datfred$Aspartate_Aminotransferase,
     pch = 19,
     col = factor(datfred$Gender))

legend("topleft",
       legend = levels(factor(datfred$Gender)),
       pch = 19,
       col = factor(levels(factor(datfred$Gender))))
```



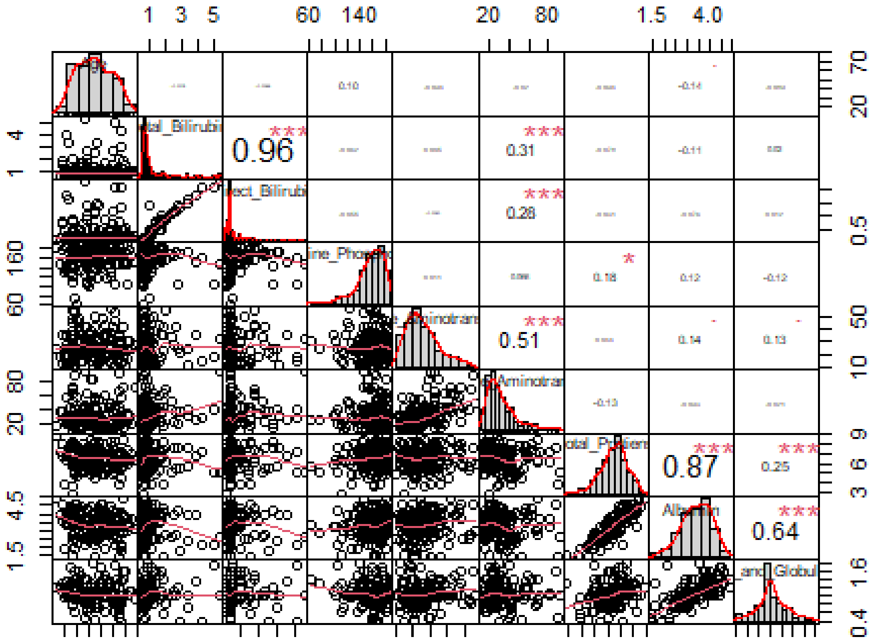
Exploratory data analysis on new dataset

```
library(ggstatsplot)

## You can cite this package as:
##      Patil, I. (2021). Visualizations with statistical details: The
'ggstatsplot' approach.
##      Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167

##
## Attaching package: 'ggstatsplot'

## The following object is masked from 'package:bitops':
##
##     %|%

library(caTools)
library(cowplot)
library(PerformanceAnalytics)

## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric

## 
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
## 
##     first, last

## 
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
## 
##     legend
```

Correlation analytics for a better understanding of the relationship between variables

Only few significant correlations were found among the factors suspected to influence one odds of being diagnosed with liver disease. There was a strong positive correlation between total and direct bilirubin and of course this was expected. Then there was a significant positive correlation between direct bilirubin and aminotransferase, and between total bilirubin and aminotransferase. In addition, alamine aminotransferase and aspartate aminotransferase had significant positive correlation, total protein and albumin, total proteins and ratio of albumin to globulin were also correlated with one another.

```r
my_data <- datfred[, c(1,3,4,5,6,7,8,9,10)]
chart.Correlation(my_data, histogram=TRUE, pch=19)
```



Final preparation of data for machine learning - converting the binary response variable to factor variable

```
datfred$liver_disease <- factor(datfred$liver_disease)
datfred$Gender <- factor(datfred$Gender)
```

Installing packages for ML

```
library(caret)

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

#use 70% of dataset as training set and 30% as test set 75% of the data were used as training set and the remainder(25%) are kept as testing set.

```
set.seed(1, sample.kind="Rounding")
```

```
## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding'
sampler
## used
```

```
test_index <- createDataPartition(y = datfred$liver_disease, times = 1, p =
0.25, list = FALSE)
train <- datfred[-test_index,]
test <- datfred[test_index,]
```

Training the data - initial model

An initial model was used to train the data. This initial model encompasses all the variables in the dataset. The AIC was used to select the model that fits best the data. The selection criteria is that the model with the smallest AIC is the best. The fith model was chosen since it yielded the lowest AIC (170.65). Only gender (p-value = 0.0171) and level of alamine aminotransferase (p-value = 0.0425) have significant impact on liver disease. Increasing level of alamine aminotransferase is associated with higher probabilities of being tested positive to liver disease.

```
initial_model <- glm(liver_disease ~ Gender + Age + Direct_Bilirubin +
Alkaline_Phosphotase +
                       Alamine_Aminotransferase + Total_Bilirubin +
Total_Protiens + Albumin + Albumin_and_Globulin_Ratio,
                       data = train,
                       family = binomial(link="logit"))

second_model <- glm(liver_disease ~ Gender + Age + Direct_Bilirubin +
Alkaline_Phosphotase +
                       Alamine_Aminotransferase + Total_Protiens + Albumin +
Albumin_and_Globulin_Ratio,
                       data = train,
                       family = binomial(link="logit"))

third_model <- glm(liver_disease ~ Gender + Age + Direct_Bilirubin +
Alkaline_Phosphotase +
                       Alamine_Aminotransferase + Total_Protiens,
                       data = train,
                       family = binomial(link="logit"))

fourth_model <- glm(liver_disease ~ Gender + Age + Alkaline_Phosphotase +
Alamine_Aminotransferase + Total_Protiens,
                       data = train,
                       family = binomial(link="logit"))

fifth_model <- glm(liver_disease ~ Gender + Age + Alkaline_Phosphotase +
Alamine_Aminotransferase,
                       data = train,
                       family = binomial(link="logit"))
```

```
summary(initial_model)
```

```
##
## Call:
## glm(formula = liver_disease ~ Gender + Age + Direct_Bilirubin +
##     Alkaline_Phosphotase + Alamine_Aminotransferase + Total_Bilirubin +
##     Total_Protiens + Albumin + Albumin_and_Globulin_Ratio, family =
binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6326  -1.0685  -0.5796   1.1068   1.7234
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 5.880939   4.184877   1.405   0.1599
## GenderMale                  1.027135   0.473030   2.171   0.0299 *
## Age                        -0.016639   0.013433  -1.239   0.2155
## Direct_Bilirubin            0.032424   1.577816   0.021   0.9836
## Alkaline_Phosphotase        0.008033   0.009622   0.835   0.4038
## Alamine_Aminotransferase   -0.048311   0.021610  -2.236   0.0254 *
## Total_Bilirubin            -0.018177   0.834514  -0.022   0.9826
## Total_Protiens             -2.033132   1.218701  -1.668   0.0953 .
## Albumin                     3.987697   2.356832   1.692   0.0907 .
## Albumin_and_Globulin_Ratio -5.806939   3.413344  -1.701   0.0889 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 170.74  on 123  degrees of freedom
## Residual deviance: 155.73  on 114  degrees of freedom
## AIC: 175.73
##
## Number of Fisher Scoring iterations: 5
```

```
summary(second_model)
```

```
##
## Call:
## glm(formula = liver_disease ~ Gender + Age + Direct_Bilirubin +
##     Alkaline_Phosphotase + Alamine_Aminotransferase + Total_Protiens +
##     Albumin + Albumin_and_Globulin_Ratio, family = binomial(link =
"logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6314  -1.0689  -0.5796   1.1077   1.7246
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 5.8693303  4.1497141   1.414   0.1572
## GenderMale                  1.0263460  0.4716104   2.176   0.0295 *
## Age                        -0.0166087  0.0133600  -1.243   0.2138
## Direct_Bilirubin           -0.0003665  0.4737180  -0.001   0.9994
## Alkaline_Phosphotase        0.0080165  0.0095932   0.836   0.4034
## Alamine_Aminotransferase   -0.0483242  0.0216033  -2.237   0.0253 *
## Total_Protiens             -2.0321083  1.2173821  -1.669   0.0951 .
## Albumin                     3.9876194  2.3561017   1.692   0.0906 .
## Albumin_and_Globulin_Ratio -5.8068843  3.4123849  -1.702   0.0888 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 170.74  on 123  degrees of freedom
## Residual deviance: 155.73  on 115  degrees of freedom
## AIC: 173.73
##
## Number of Fisher Scoring iterations: 5
```

```
summary(third_model)
```

```
##
## Call:
## glm(formula = liver_disease ~ Gender + Age + Direct_Bilirubin +
##     Alkaline_Phosphotase + Alamine_Aminotransferase + Total_Protiens,
##     family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6050  -1.0318  -0.7224   1.1703   1.8489
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.722433   1.801967  -0.401   0.6885
## GenderMale                1.087149   0.452991   2.400   0.0164 *
## Age                      -0.020818   0.012909  -1.613   0.1068
## Direct_Bilirubin         -0.184461   0.433415  -0.426   0.6704
## Alkaline_Phosphotase      0.012880   0.009273   1.389   0.1648
## Alamine_Aminotransferase -0.040390   0.020473  -1.973   0.0485 *
## Total_Protiens           -0.028465   0.187244  -0.152   0.8792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 170.74  on 123  degrees of freedom
## Residual deviance: 160.45  on 117  degrees of freedom
## AIC: 174.45
##
## Number of Fisher Scoring iterations: 4
```

```
summary(fourth_model)
```

```
##
## Call:
## glm(formula = liver_disease ~ Gender + Age + Alkaline_Phosphotase +
##     Alamine_Aminotransferase + Total_Protiens, family = binomial(link =
"logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5918  -1.0204  -0.7346   1.1832   1.8671
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.868926   1.771091  -0.491   0.6237
## GenderMale                1.069566   0.450977   2.372   0.0177 *
## Age                      -0.021001   0.012903  -1.628   0.1036
## Alkaline_Phosphotase      0.013474   0.009211   1.463   0.1435
## Alamine_Aminotransferase -0.041091   0.020483  -2.006   0.0448 *
## Total_Protiens           -0.025237   0.186471  -0.135   0.8923
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 170.74  on 123  degrees of freedom
## Residual deviance: 160.63  on 118  degrees of freedom
## AIC: 172.63
##
## Number of Fisher Scoring iterations: 4
```

```
summary(fifth_model)
```

```
##
## Call:
## glm(formula = liver_disease ~ Gender + Age + Alkaline_Phosphotase +
##     Alamine_Aminotransferase, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5688  -1.0187  -0.7396   1.1819   1.8910
##
## Coefficients:
```

```
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.990845   1.522623  -0.651   0.5152
## GenderMale                1.073700   0.450154   2.385   0.0171 *
## Age                      -0.020797   0.012812  -1.623   0.1045
## Alkaline_Phosphotase      0.013190   0.008961   1.472   0.1410
## Alamine_Aminotransferase -0.041372   0.020395  -2.029   0.0425 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 170.74  on 123  degrees of freedom
## Residual deviance: 160.65  on 119  degrees of freedom
## AIC: 170.65
##
## Number of Fisher Scoring iterations: 4
```

Let’s test our results on unused data (testing set)

We calculated the prediction probabilities and predicted classes on top of those probabilities. We picked 0.5 to be our threshold so that patients will be declared negative if their probabilities fall under 0.5 and positive otherwise.

```
probabs <- predict(fifth_model, test, type = 'response')
preds <- ifelse(probabs > 0.5, 1,0)
```

Getting the confusion matrix

So, overall, our model is correct in 64% of the test cases.

```
confusionMatrix(factor(preds), factor(test$liver_disease))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 17  9
##          1  6 10
##
##                Accuracy : 0.6429
##                  95% CI : (0.4803, 0.7845)
##     No Information Rate : 0.5476
##     P-Value [Acc > NIR] : 0.1387
##
##                   Kappa : 0.2691
##
##  Mcnemar's Test P-Value : 0.6056
##
```

```
##             Sensitivity : 0.7391
##             Specificity : 0.5263
##          Pos Pred Value : 0.6538
##          Neg Pred Value : 0.6250
##              Prevalence : 0.5476
##          Detection Rate : 0.4048
##    Detection Prevalence : 0.6190
##       Balanced Accuracy : 0.6327
##
##        'Positive' Class : 0
##
```

RESULTS

The results of the logistic regression suggest that only gender (p-value = 0.0171) and level of alamine aminotransferase (p-value = 0.0425) have significant effects on one likelihood of developing liver disease. We did not find enough evidence that albumin, age, ratio of albumin to globulin, total proteins, direct bilirubin, alkaline phosphotase, and total bilirubin affect odds of being tested positive. We performed a prediction using the testing data and the accuracy result from the confusion Matrix suggests that our model is correct in 64% of the test cases. In other words, a true positive and true negative occur in 64% of cases.

CONCLUSION

Data were collected on 583 patients to determine the drivers of liver disease. Data collected include demographic information (gender and age) along with some biological test results (level of bilirubin, alkaline phosphotase, alamine aminotransferase, aspartate aminotransferase, proteins, albumin, and the ratio of albumin to globulin). Each patient was then diagnosed for liver disease and the outcome for this diagnosis is either positive (has liver disease) or negative (does not have liver disease). To help decision makers develop and deploy policies to fight the increasing prevalence of liver disease, I used a ML algorithm and a logistic regression to test whether each of the factors suspected is in fact a significant driver. After selecting the best model that fits the data based on the AIC, I used the best model to predict and the confusion matrix was used to test the performance of the prediction. The results of the logistic regression suggest that (1) efforts should first target women who were found to have higher prevalence of liver disease. In addition, local authorities should work on reducing level of alamine aminotransferase that has significant effect on likelihood of developing liver disease. One of the biggest limitations of this study is that most of the real contributing factors were not included in the data.

REFERENCES

1. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.