R Data Science Project

Farzad and Saurabh

04/03/2020

Indian Liver Patient Dataset

Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs.

This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors. This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records.

Any patient whose age exceeded 89 is listed as being of age "90".

Columns:

Age of the patient, Gender of the patient, Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Protiens, Albumin, Albumin, and Globulin Ratio, Dataset: field used to split the data into two sets (patient with liver disease, or no disease)

```
str(data)
```

```
## 'data.frame':    583 obs. of  11 variables:
##  $ Age                       : int  65 62 62 58 72 46 26 29 17 55 ...
##  $ Gender                    : Factor w/ 2 levels "Female","Male": 1 2
...
##  $ Total_Bilirubin           : num  0.7 10.9 7.3 1 3.9 1.8 0.9 0.9 0.9
...
##  $ Direct_Bilirubin          : num  0.1 5.5 4.1 0.4 2 0.7 0.2 0.3 0.3 .
..
##  $ Alkaline_Phosphotase      : int  187 699 490 182 195 208 154 202  ..
.
##  $ Alamine_Aminotransferase  : int  16 64 60 14 27 19 16 14 22 53 ...
##  $ Aspartate_Aminotransferase: int  18 100 68 20 59 14 12 11 19 58 ...
##  $ Total_Protiens            : num  6.8 7.5 7 6.8 7.3 7.6 7 6.7 7.4  ..
.
##  $ Albumin                   : num  3.3 3.2 3.3 3.4 2.4 4.4 3.5 3.6 ...
##  $ Albumin_and_Globulin_Ratio: num  0.9 0.74 0.89 1 0.4 1.3 1 1.1 1.2 .
..
##  $ Dataset                   : int  1 1 1 1 1 1 1 1 2 1 ...
```

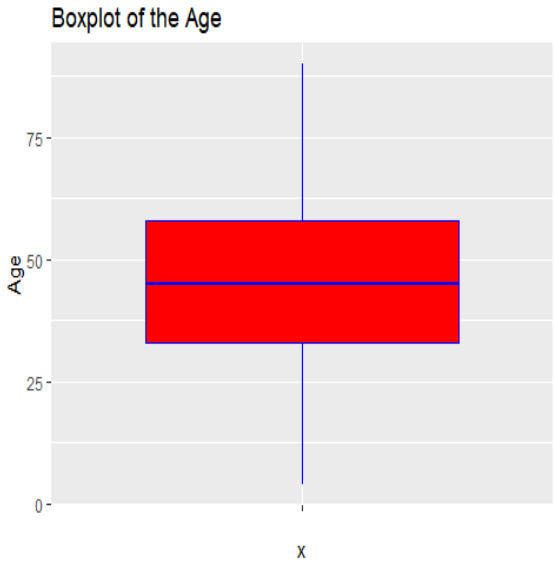
Looks like the data was read in well and we can see that all the variables but gender are numeric.

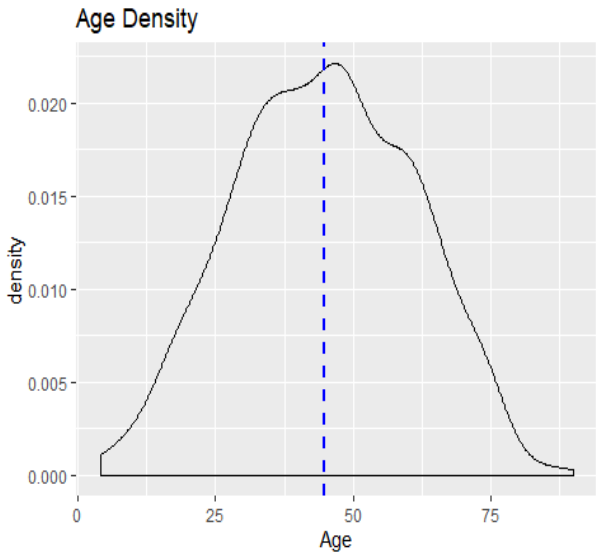
```
summary(data)
```

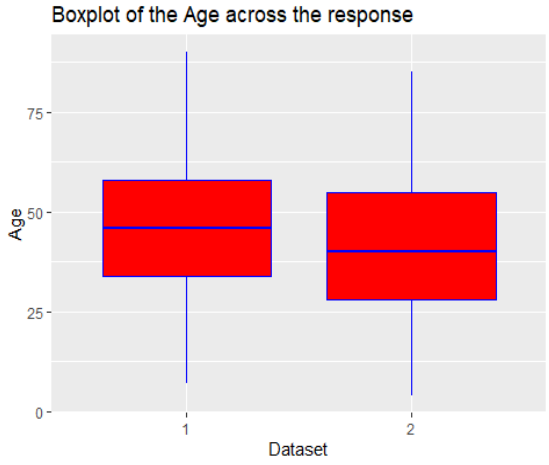
```
##       Age           Gender    Total_Bilirubin  Direct_Bilirubin
##  Min.   : 4.00   Female:142   Min.   : 0.400   Min.   : 0.100
##  1st Qu.:33.00   Male  :441   1st Qu.: 0.800   1st Qu.: 0.200
##  Median :45.00                Median : 1.000   Median : 0.300
##  Mean   :44.75                Mean   : 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_Aminotransfera
se
##  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.   :1.0
00
##  1st Qu.:5.800   1st Qu.:2.600   1st Qu.:0.7000             1st Qu.:1.0
00
##  Median :6.600   Median :3.100   Median :0.9300             Median :1.0
00
##  Mean   :6.483   Mean   :3.142   Mean   :0.9471             Mean   :1.2
86
##  3rd Qu.:7.200   3rd Qu.:3.800   3rd Qu.:1.1000             3rd Qu.:2.0
00
##  Max.   :9.600   Max.   :5.500   Max.   :2.8000             Max.   :2.0
00
##                                  NA's   :4
```

Basic Visualizations and EDA

Age



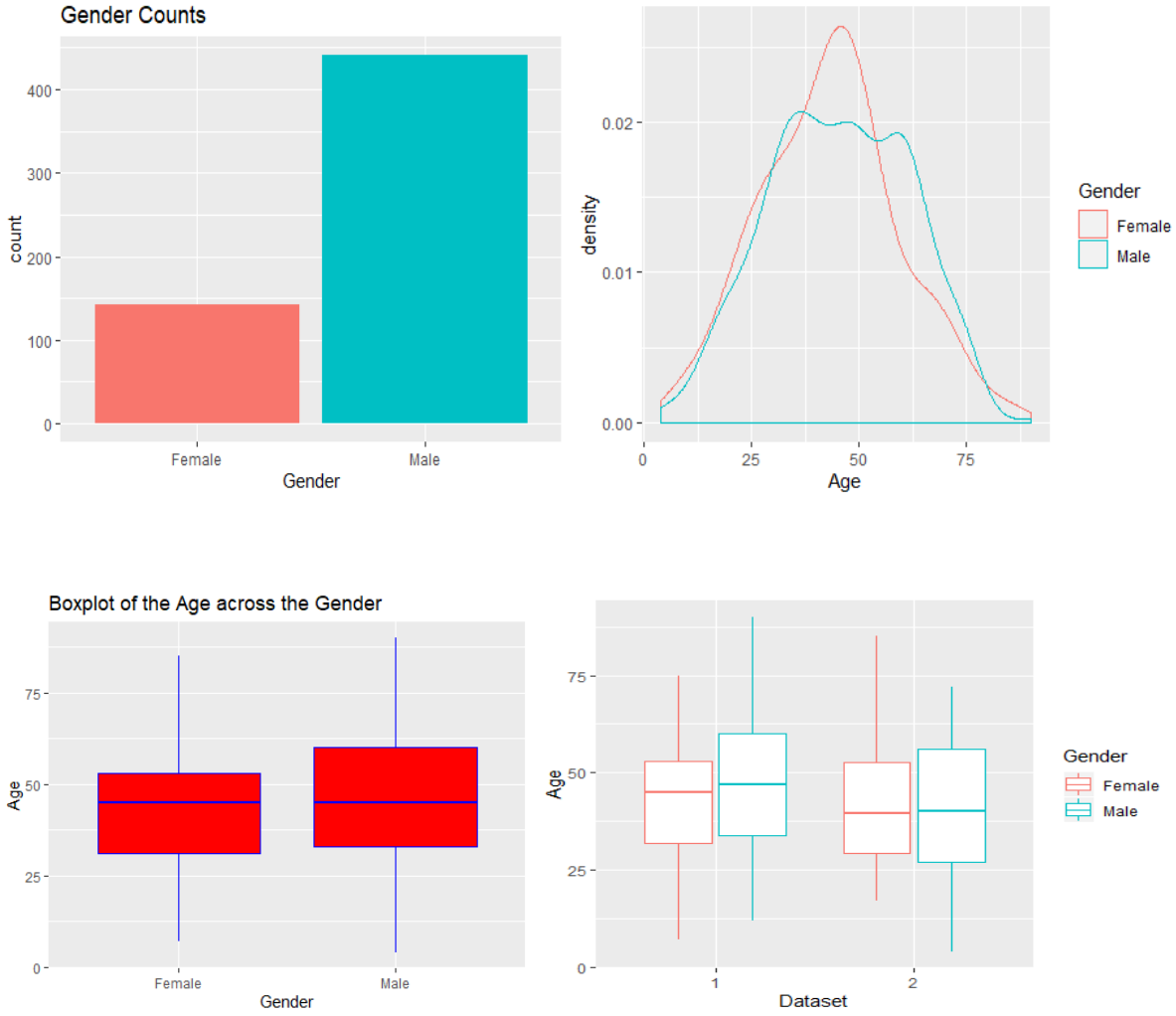





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   33.00   45.00   44.75   58.00   90.00
```

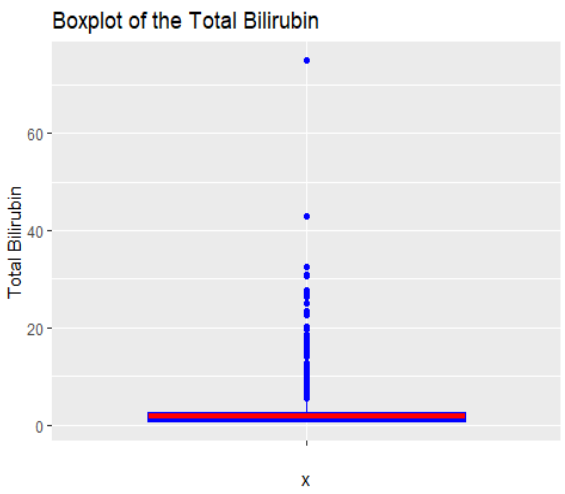
From the graphs we see what the data looks like and we also have a summary again for the age variable. The mean age is about 44 which is close to the median of 45. The oldest person in the data is 90 whilst the youngest is 4. The age variable seems to have a bell shaped curve and there seems to be a difference in the ages for the two different liver disease groups. Dataset = 1 has a higher mean than Dataset = 2.

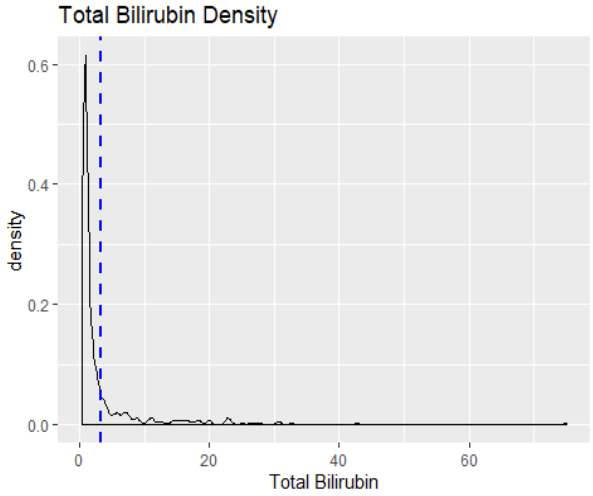
Gender

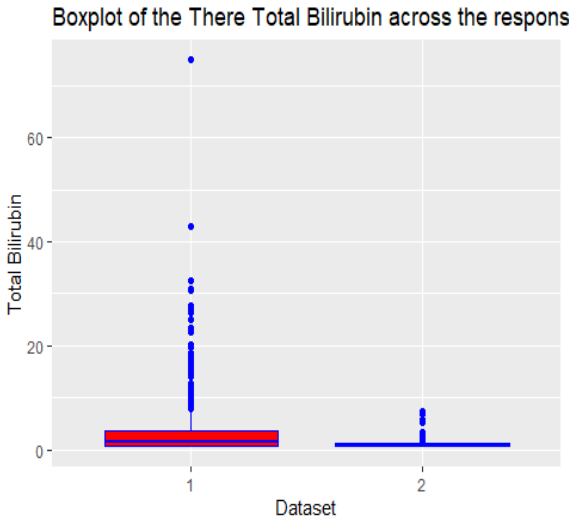


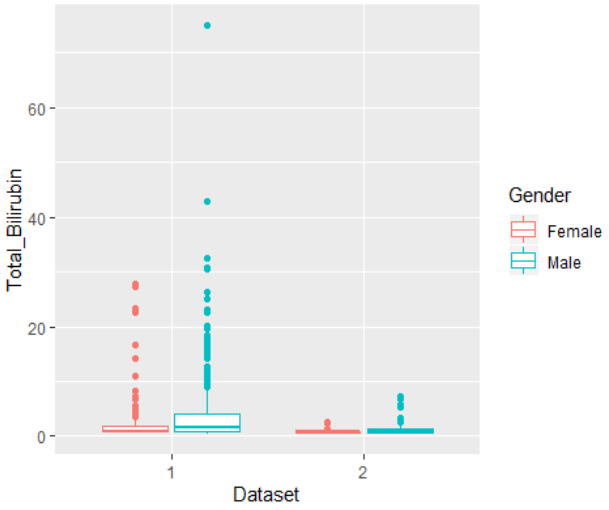
There are more males than females in the entire dataset. The mean age is higher for males than females in the data, both genders have a similar density shape when you look at the age. There are a few graphical ways of looking at this just to make sure we are doing the work right. When looking at the breakdown of the age, gender and response we see some interesting things. When we look at the females, the mean age is higher for the Dataset = 1 and the same is true for Dataset = 2. Within the Dataset = 2 group the mean ages are pretty close but in the Dataset = 1 ,the males have a higher mean age.

Total Bilirubin





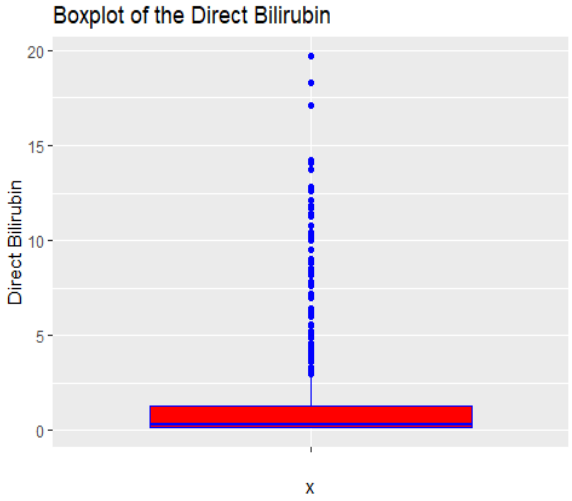


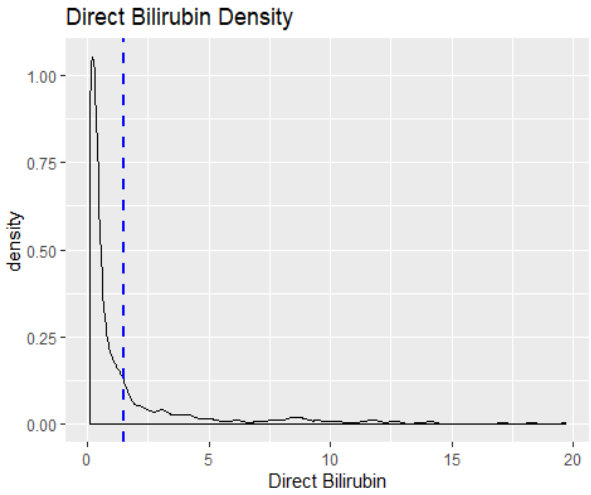


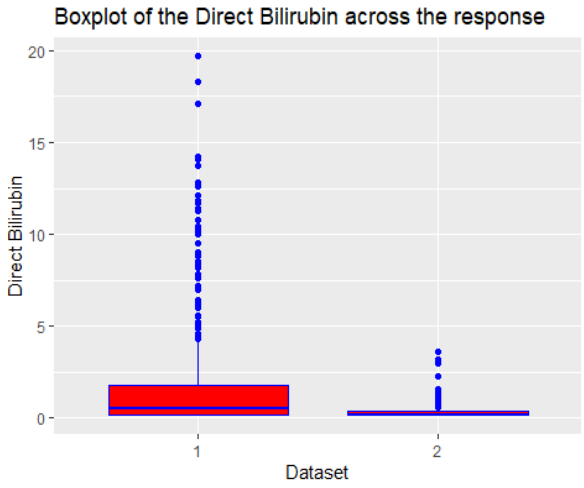

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.400   0.800   1.000   3.299   2.600  75.000
```

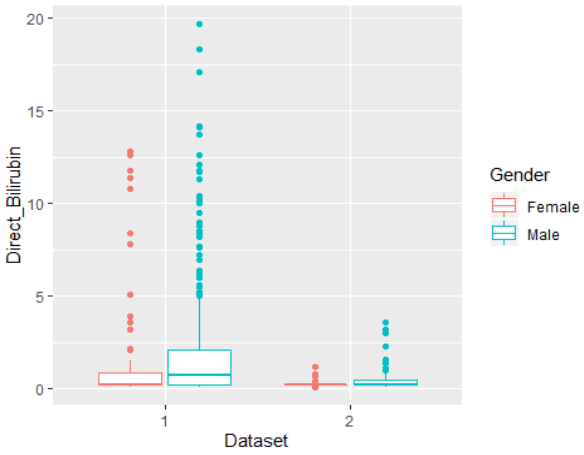
There is a wide range for the Total Bilirubin. The data is skewed, very skewed, we have a maximum value of 75 with the mean of 3.299, median of 1. There is a difference between the mean Total Bilirubin for the responses and this also seems to be true when you break it down further by gender.

Direct Bilirubin





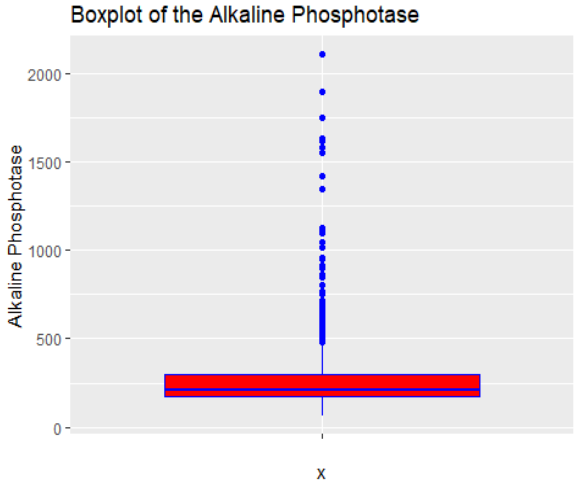


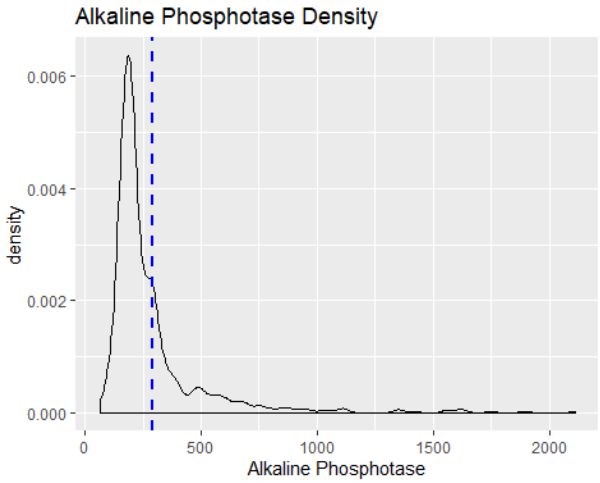


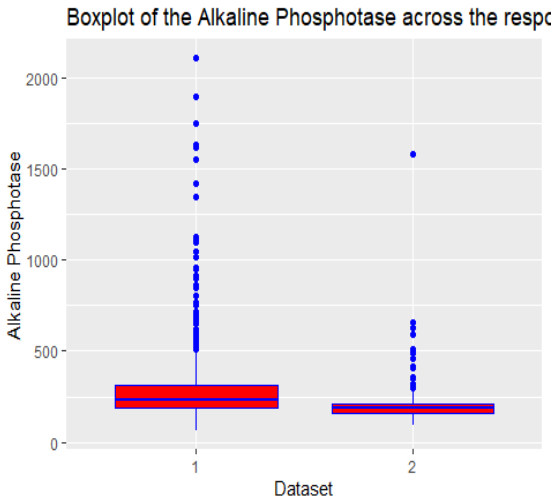

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.200   0.300   1.486   1.300  19.700
```

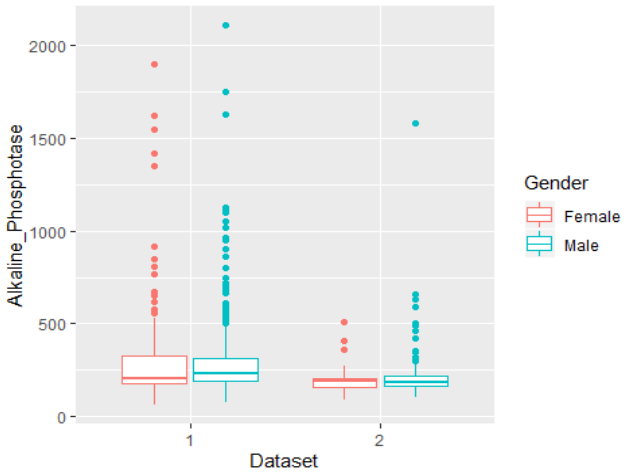
Similar to the Total Bilirubin we have skewed data for the Direct Bilirubin. The mean is 1.486, with a maximum being 19.7 and the median being 0.3. There is a difference in the mean and range of the Direct Bilirubin for the two responses. There are differences also when you look t it further by gender. The mean Direct Bilirubin is higher for the male group across the responses. There is a bigger range for both genders for Direct Bilirubin when the response is 1.

Alkaline Phosphotase





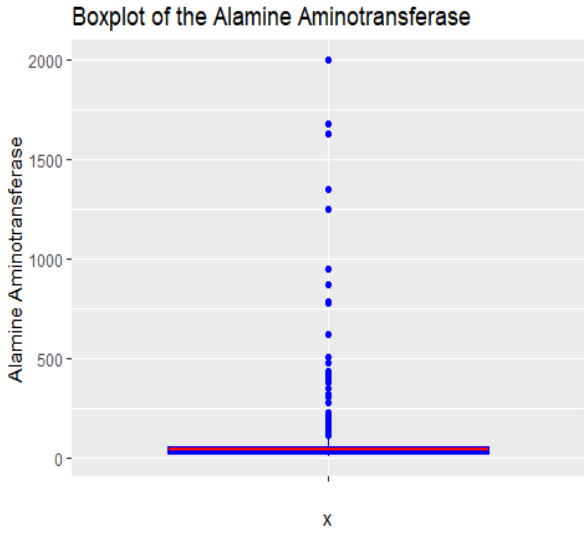


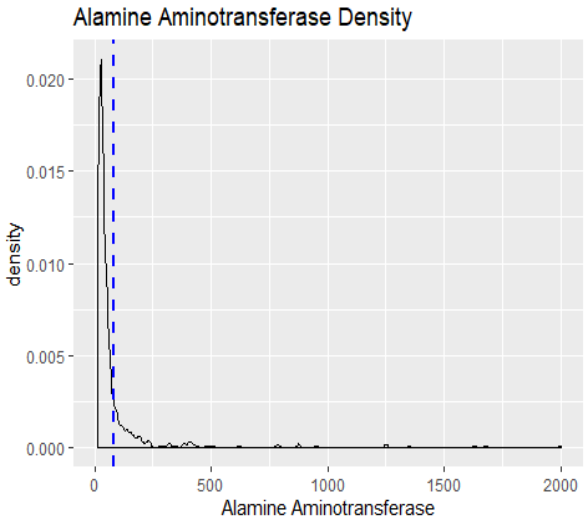


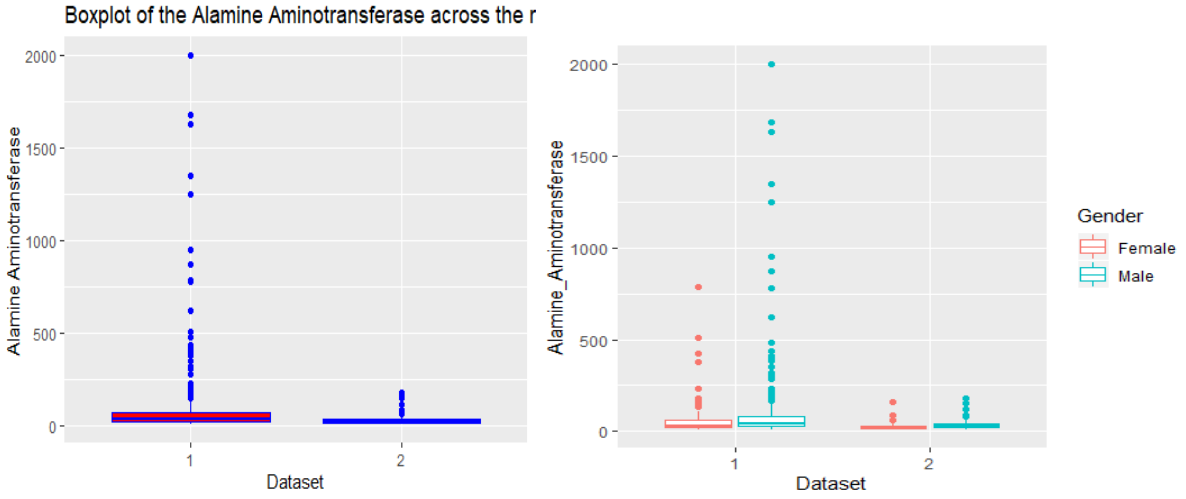

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    63.0   175.5   208.0   290.6   298.0  2110.0
```

The range for the Alkaline Phosphotase is wide, with a mean of 290.6, median or 208, minimum of 63 and maximum of 2110. The numbers for Alkaline Phosphotase are also skewed as shown in the density plot. When you look at the numbers across the response we see a difference across it. The mean is higher when the response is 1 and the range is also wider for response = 1. When you look further and compare by gender, there are differences in the mean within the Dataset = 1, but not much for the Dataset = 2 group.

Alamine Aminotransferase



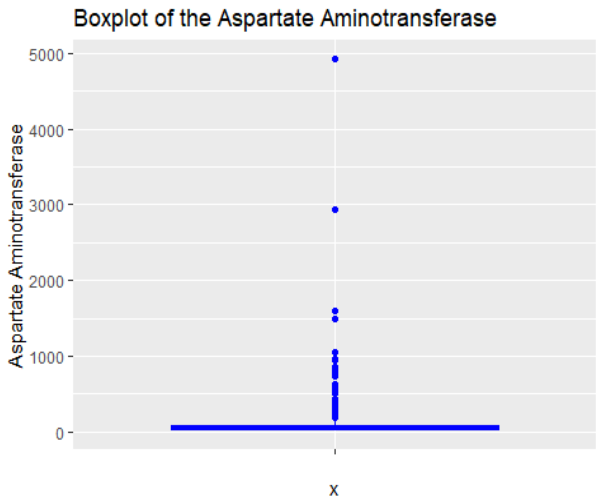


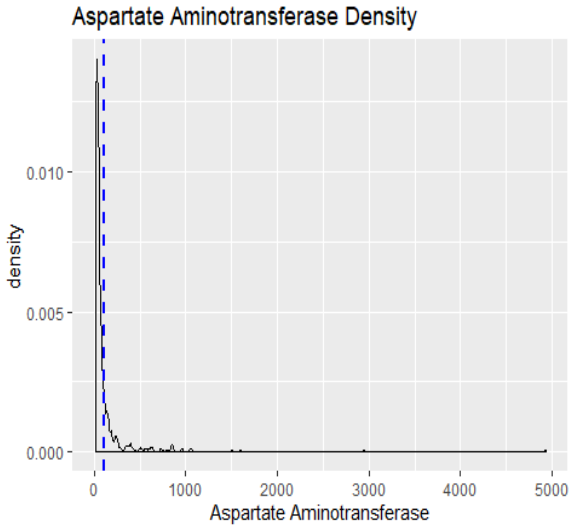


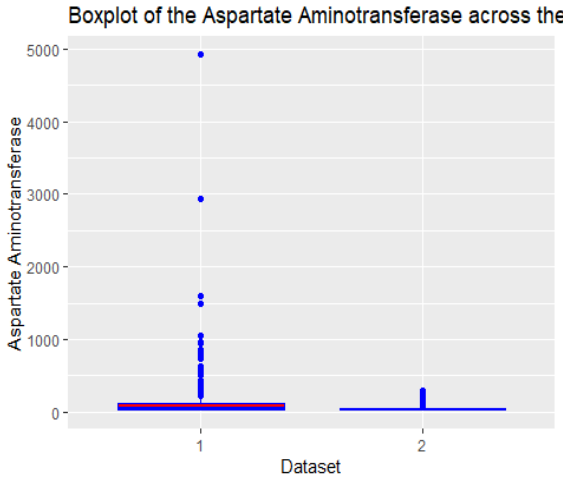

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   23.00   35.00   80.71   60.50 2000.00
```

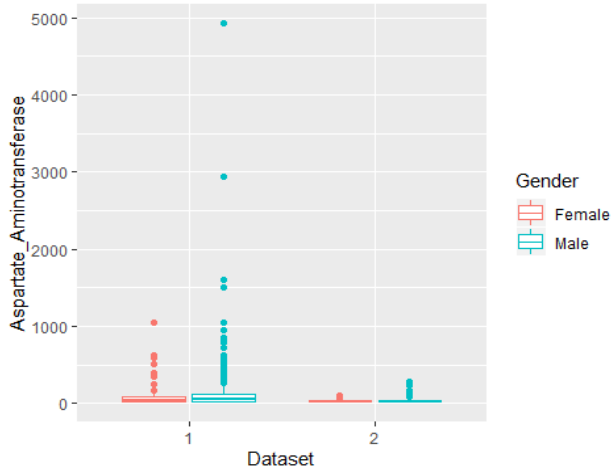
Another skewed variable. Range is from 10 to 2000, with a mean of 80.71, median of 35.00. There is a wider range within the Dataset = 1 group than the Dataset = 2 group, and the means are clearly different between the two groups. We again see a difference between the genders within the responses. The male groups within each response have a higher mean of Alamine Aminotransferase.

Aspartate Aminotransferase





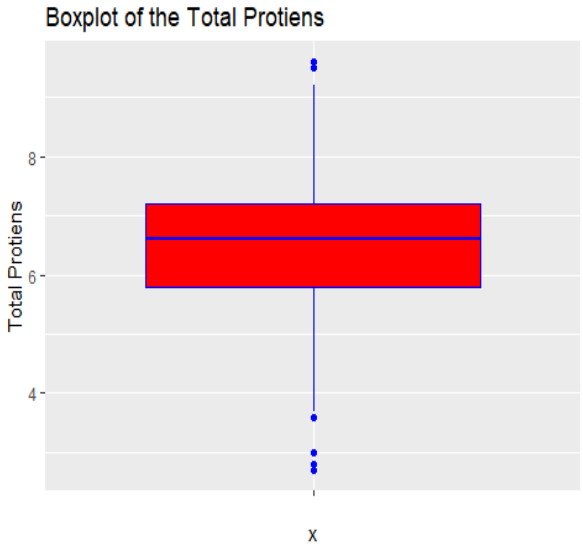


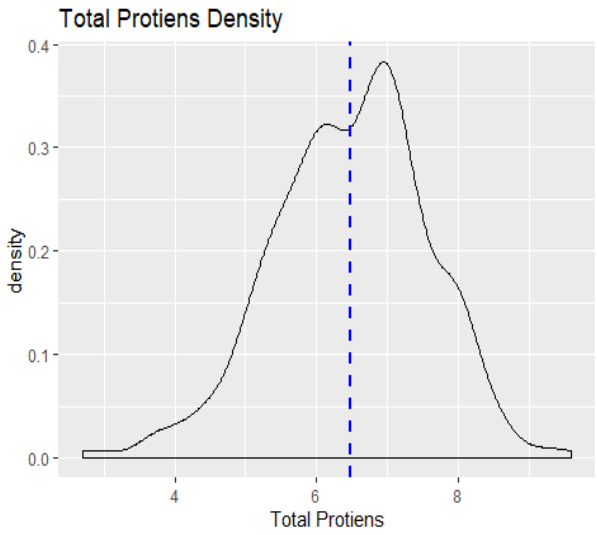


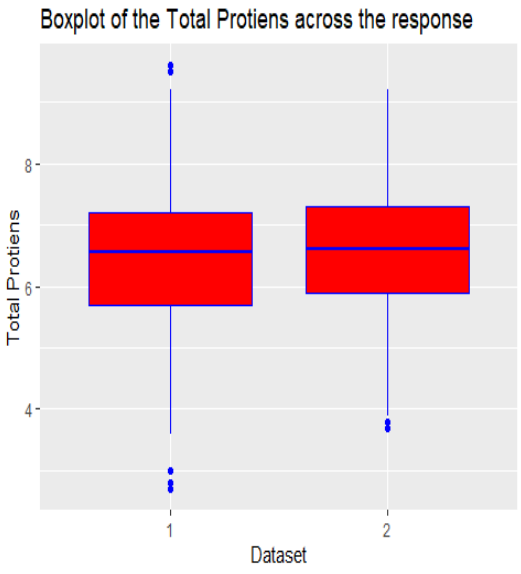

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0    25.0    42.0   109.9    87.0  4929.0
```

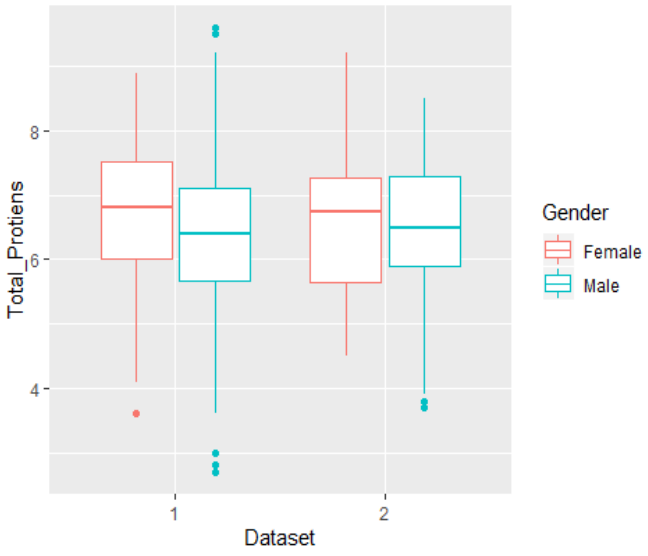
Another skewed variable with a range from 10 to 4929, with a mean of 109.9, and median of 42. The range is wider within the Dataset = 1 group in comparison to the Dataset = 2 group. The mean is also higher in the Dataset = 1 group. The box plots are hard to really read but they are there for you to look at.

Total Protiens





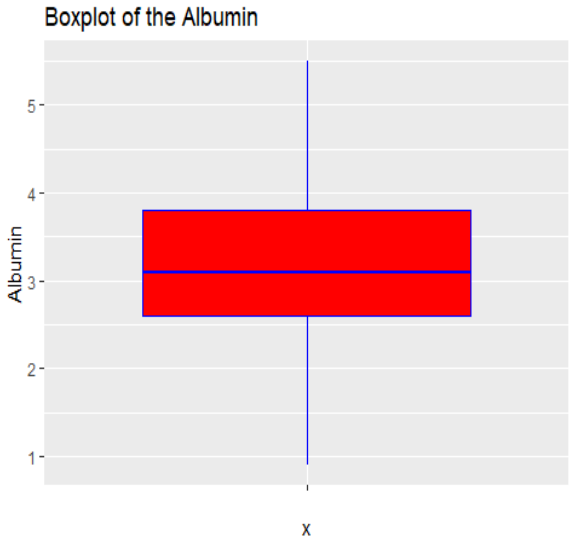


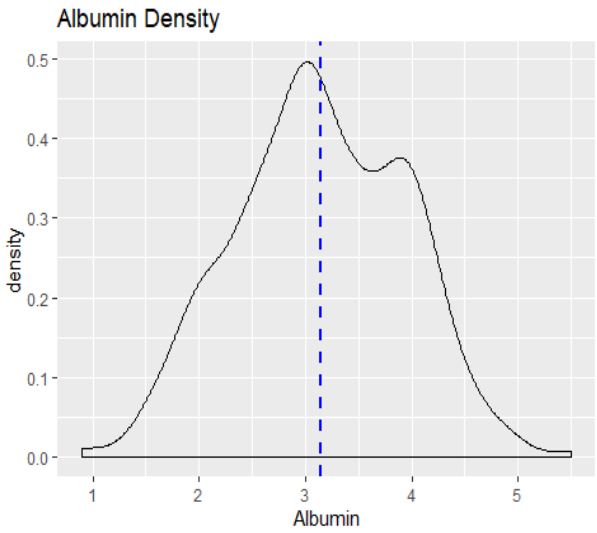


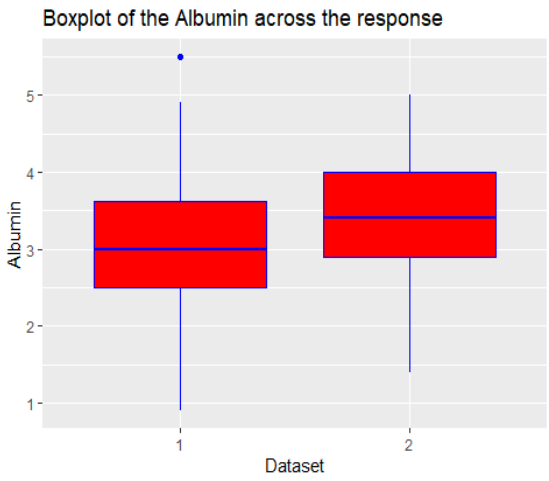

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.700   5.800   6.600   6.483   7.200   9.600
```

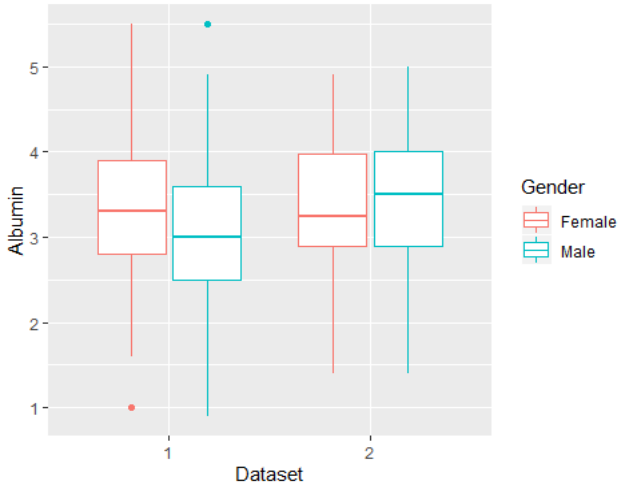
A variable that looks like it has a bell shaped curve. The range for the variable is from 2.7 to 9.6, with a mean of 6.483 and median of 6.6. There does not seem to be much of a difference between the means for the Total Proteins for each response. There is a difference between the responses when you look across the genders. The females have a higher mean for total proteins across the responses.

Albumin





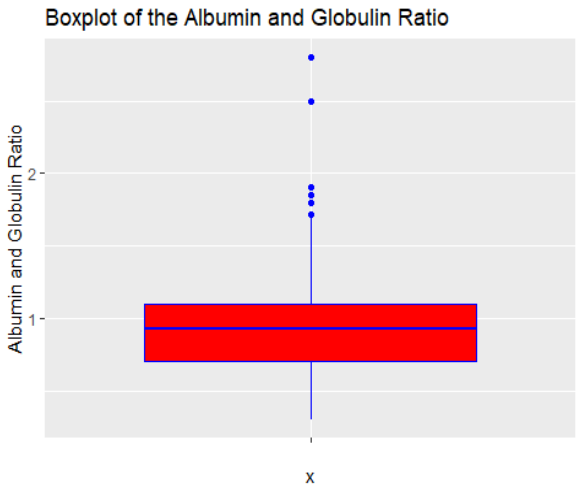


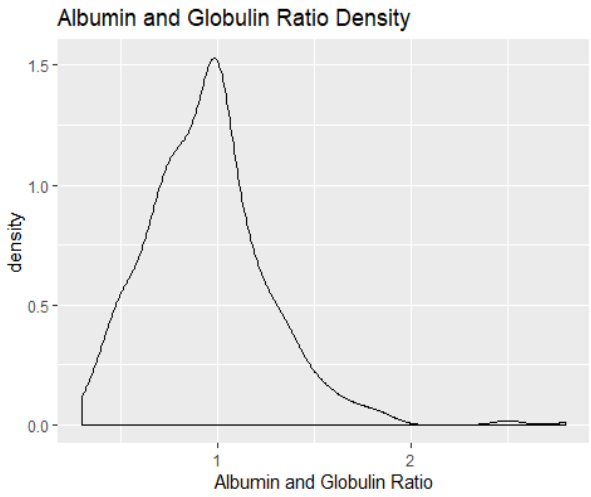


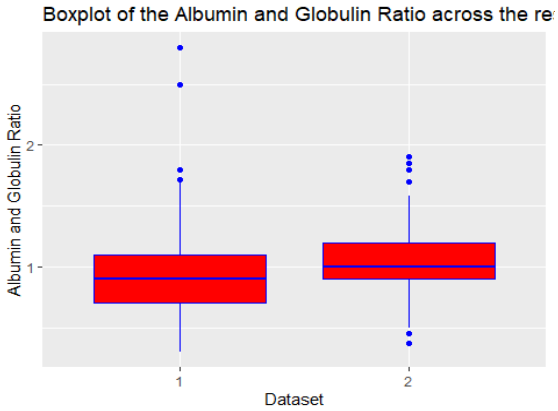

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   2.600   3.100   3.142   3.800   5.500
```

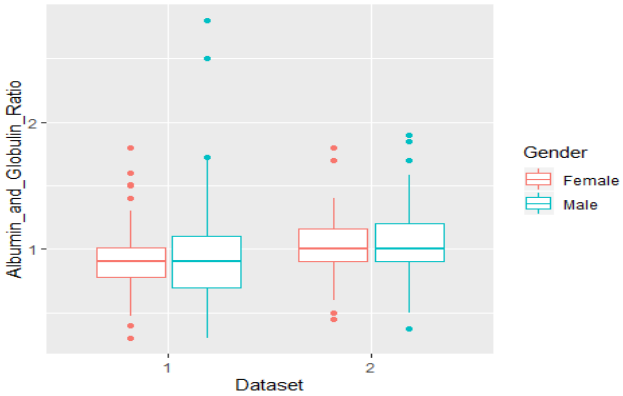
Range from 0.9 to 5.5, mean of 3.142, and median of 3.1. Bell shaped density for the Albumin values. The mean is higher for Dataset = 2 and the range is larger when the Dataset = 1. There is a clear difference in the means for the Albumin values across the genders and responses. The mean is higher for females when the response = 1 and is higher for males when the response = 2.

Albumin and Globulin Ratio





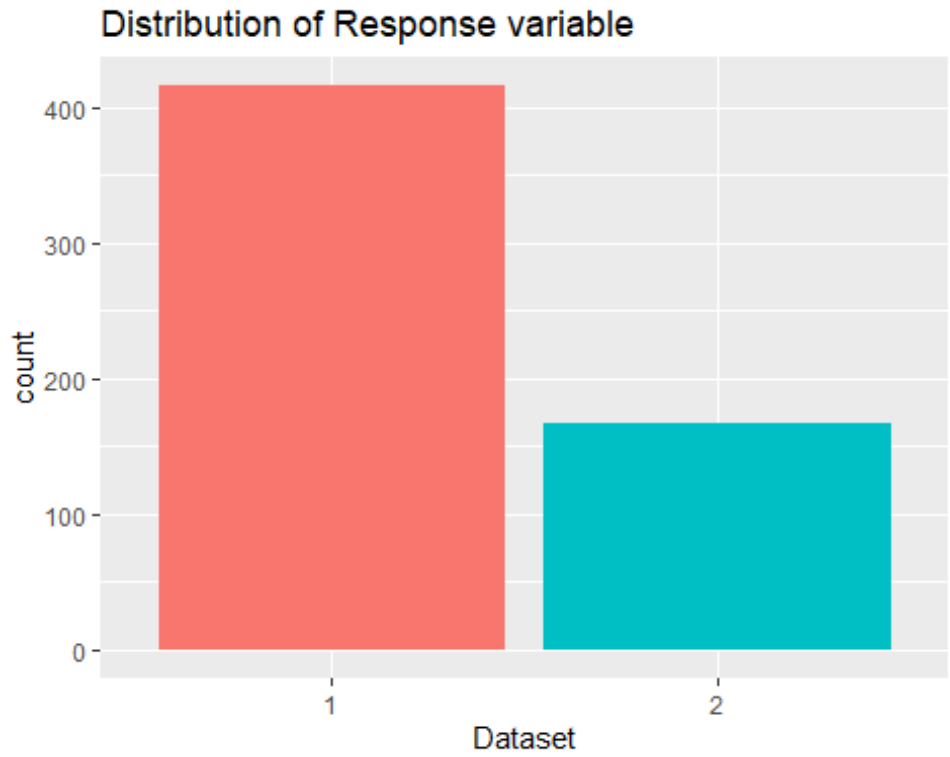





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.3000  0.7000  0.9300  0.9471  1.1000  2.8000       4
```

The mean for the ratio is 0.9471 and is very clos to the median of 0.93. The ratio has somewhat of a bell shaped curve. The mean for the ratio is higher in the Dataset = 2 response, with the range being larger in the Dataset = 1 group. There is no difference with the the responses when you compare the genders. The ranges vary though.

Response



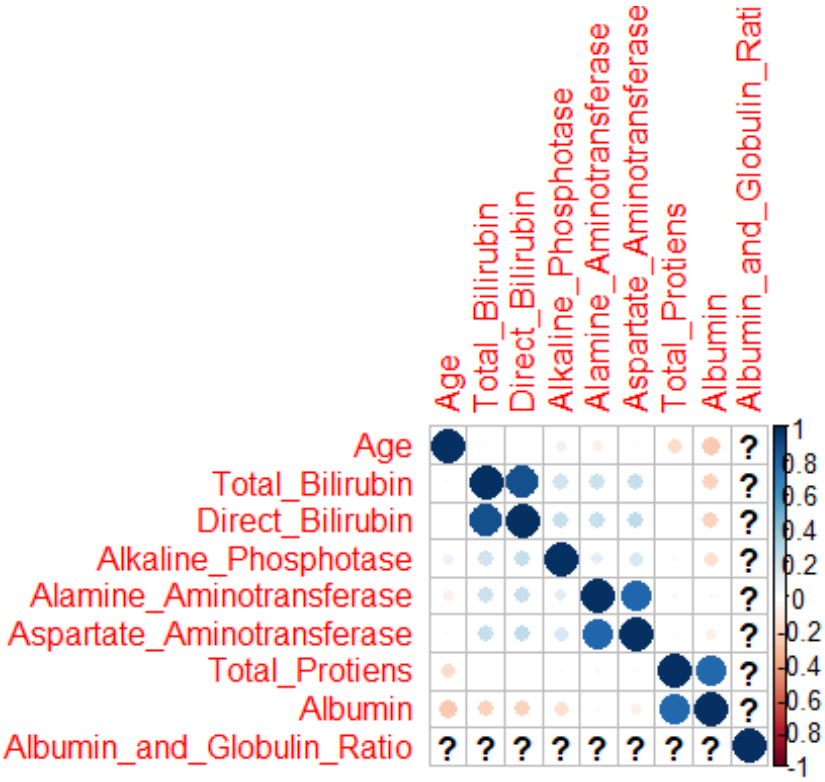
There are more Dataset = 1 than Dataset = 2 responses as shown in the bar chart above.

Correlations

Running a few things to visualize and calculate the correlations.

```
##                                    Age Total_Bilirubin Direct_Bilirubi
n
## Age                        1.000000000     0.011762651     0.007529138
1
## Total_Bilirubin            0.011762651     1.000000000     0.874617930
1
## Direct_Bilirubin           0.007529138     0.874617930     1.000000000
0
## Alkaline_Phosphotase       0.080424612     0.206668795     0.234938705
8
## Alamine_Aminotransferase  -0.086882759     0.214064740     0.233894054
5
## Aspartate_Aminotransferase -0.019909857     0.237831323     0.257543981
1
## Total_Protiens             -0.187461261    -0.008099343    -0.000138741
4
## Albumin                    -0.265924361    -0.222250406    -0.228530572
9
## Albumin_and_Globulin_Ratio           NA              NA               N
A
##                           Alkaline_Phosphotase Alamine_Aminotransferas
e
```

```
## Age                                0.08042461              -0.0868827
6
## Total_Bilirubin                    0.20666880               0.2140647
4
## Direct_Bilirubin                   0.23493871               0.2338940
5
## Alkaline_Phosphotase               1.00000000               0.1256799
5
## Alamine_Aminotransferase           0.12567995               1.0000000
0
## Aspartate_Aminotransferase         0.16719590               0.7919656
8
## Total_Protiens                    -0.02851436              -0.0425181
9
## Albumin                           -0.16545287              -0.0297416
7
## Albumin_and_Globulin_Ratio                NA                       N
A
##                           Aspartate_Aminotransferase Total_Protiens
## Age                                       -0.01990986  -0.1874612615
## Total_Bilirubin                            0.23783132  -0.0080993434
## Direct_Bilirubin                           0.25754398  -0.0001387414
## Alkaline_Phosphotase                       0.16719590  -0.0285143556
## Alamine_Aminotransferase                   0.79196568  -0.0425181903
## Aspartate_Aminotransferase                 1.00000000  -0.0256453651
## Total_Protiens                            -0.02564537   1.0000000000
## Albumin                                   -0.08529030   0.7840533354
## Albumin_and_Globulin_Ratio                          NA             NA
##                               Albumin Albumin_and_Globulin_Ratio
## Age                        -0.26592436                        NA
## Total_Bilirubin            -0.22225041                        NA
## Direct_Bilirubin           -0.22853057                        NA
## Alkaline_Phosphotase       -0.16545287                        NA
## Alamine_Aminotransferase   -0.02974167                        NA
## Aspartate_Aminotransferase -0.08529030                        NA
## Total_Protiens              0.78405334                        NA
## Albumin                     1.00000000                        NA
## Albumin_and_Globulin_Ratio          NA                         1
```



From the corrplot it seems there are a few variables that are highly correlated with each other and with that in mind will have to be careful when building the algorithms.

Model Building

To develop a model to predict liver disease and no liver disease based on patient records, we tried the following 3 dataset and 4 machine learning models:

-> ORIGINAL DATAFRAME: df.o
-> ORIGINAL DATAFRAME with EXTREME VALUES FITTED: df.o.wo
-> ORIGINAL DATAFRAME CATEGORIZATED BY CLUSTERING: df.cat.f

1. Logistic Regression:

ORIGINAL DATAFRAME:

```
## Call:
## glm(formula = Dataset ~ Total_Bilirubin + Alkaline_Phosphotase +
##     Alamine_Aminotransferase + Total_Protiens + Albumin + Age,
##     family = binomial(link = "logit"), data = df.Train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0560  -1.1370   0.4463   0.9076   1.4566
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.5301209  0.8769951  -1.745  0.08103 .
## Total_Bilirubin           0.2601250  0.1002763   2.594  0.00948 **
## Alkaline_Phosphotase      0.0009163  0.0008615   1.064  0.28752
## Alamine_Aminotransferase  0.0141188  0.0044700   3.159  0.00159 **
## Total_Protiens            0.4487603  0.2028569   2.212  0.02695 *
## Albumin                  -0.7214425  0.2883465  -2.502  0.01235 *
## Age                       0.0124358  0.0071520   1.739  0.08207 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 554.14  on 463  degrees of freedom
## Residual deviance: 467.20  on 457  degrees of freedom
## AIC: 481.2
##
## Number of Fisher Scoring iterations: 7
```

It can be seen that only 4 out of the 6 predictors are significantly associated to the outcome - liver_disease. These include: Total_Bilirubin, Alamine_Aminotransferase and Albumin.

The logistic regression coefficients give the change in the log odds of the outcome for a single unit increase in the predictor variable.

The coefficient estimate of the variable Alamine_Aminotransferase is b = 0.0141188, which is positive. This means that an increase in Alamine Aminotransferase lab test is associated with increase in the probability of having liver disease. This indicates that one unit increase in the Alamine_Aminotransferase unit will increase the odds of having liver disease by exp(0.0141188) 1.014219 times.

However, the coefficient for the variable Albumin is b = 0.7214425, which is negative. This means that an increase in Albumin lab test will be associated with a decreased probability of having liver disease.

```
## Confusion Matrix and Statistics
##
##                   Reference
## Prediction         no_liver_disease liver_disease
##   no_liver_disease                8             3
##   liver_disease                  25            79
##
##                Accuracy : 0.7565
##                  95% CI : (0.6677, 0.8317)
##     No Information Rate : 0.713
##     P-Value [Acc > NIR] : 0.1773
##
##                   Kappa : 0.257
##
##  Mcnemar's Test P-Value : 7.229e-05
##
##             Sensitivity : 0.24242
##             Specificity : 0.96341
##          Pos Pred Value : 0.72727
##          Neg Pred Value : 0.75962
##              Prevalence : 0.28696
##          Detection Rate : 0.06957
##    Detection Prevalence : 0.09565
##       Balanced Accuracy : 0.60292
##
##        'Positive' Class : no_liver_disease
##
```

ORIGINAL DATAFRAME CATEGORIZATED BY CLUSTERING:

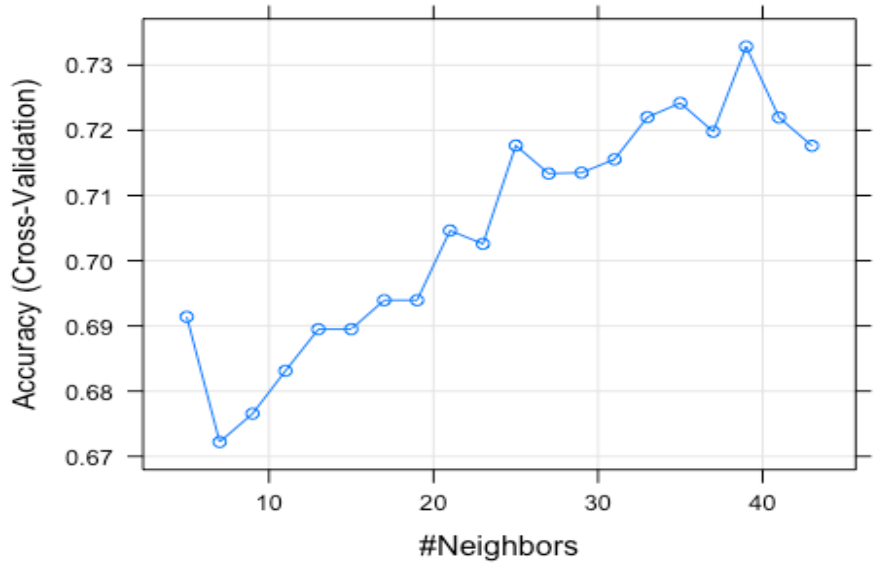
```
## Call:
## glm(formula = Dataset ~ Total_Bilirubin_ + Alkaline_Phosphotase_ +
##     Alamine_Aminotransferase_ + Total_Protiens_ + Albumin_ +
##     AST_ALT_ratio_ + Age, family = binomial(link = "logit"),
##     data = df.Train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4503  -0.9005   0.3412   0.8109   1.7160
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 554.14  on 463  degrees of freedom
## Residual deviance: 429.68  on 425  degrees of freedom
## AIC: 507.68
##
## Number of Fisher Scoring iterations: 15
```

```
## Confusion Matrix and Statistics
##
##                   Reference
## Prediction         no_liver_disease liver_disease
##   no_liver_disease               16             9
##   liver_disease                  17            73
##
##                Accuracy : 0.7739
##                  95% CI : (0.6865, 0.8467)
##     No Information Rate : 0.713
##     P-Value [Acc > NIR] : 0.08794
##
##                   Kappa : 0.4044
##
##  Mcnemar's Test P-Value : 0.16981
##
##             Sensitivity : 0.4848
##             Specificity : 0.8902
##          Pos Pred Value : 0.6400
##          Neg Pred Value : 0.8111
##              Prevalence : 0.2870
##          Detection Rate : 0.1391
##    Detection Prevalence : 0.2174
##       Balanced Accuracy : 0.6875
##
##        'Positive' Class : no_liver_disease
##
```

This model performed better with the ORIGINAL DATAFRAME CATEGORIZATED BY CLUSTERING - The overall accuracy was 77,39%. The accuracy represents the correct classification of patients with and without the condition (prediction x reference).

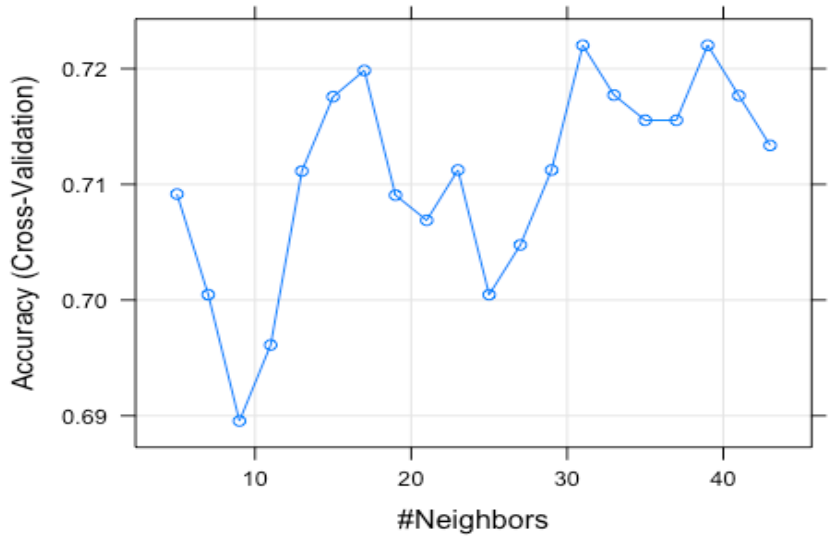
2. KNN (K-Nearest Neighbors):

ORIGINAL DATAFRAME:




```
## Confusion Matrix and Statistics
##
##                   Reference
## Prediction         no_liver_disease liver_disease
##   no_liver_disease                5             5
##   liver_disease                  28            77
##
##                Accuracy : 0.713
##                  95% CI : (0.6212, 0.7935)
##     No Information Rate : 0.713
##     P-Value [Acc > NIR] : 0.5467811
##
##                   Kappa : 0.1144
##
##  Mcnemar's Test P-Value : 0.0001283
##
##             Sensitivity : 0.15152
##             Specificity : 0.93902
##          Pos Pred Value : 0.50000
##          Neg Pred Value : 0.73333
##              Prevalence : 0.28696
##          Detection Rate : 0.04348
##    Detection Prevalence : 0.08696
##       Balanced Accuracy : 0.54527
##
##        'Positive' Class : no_liver_disease
##
```

ORIGINAL DATAFRAME CATEGORIZATED BY CLUSTERING:

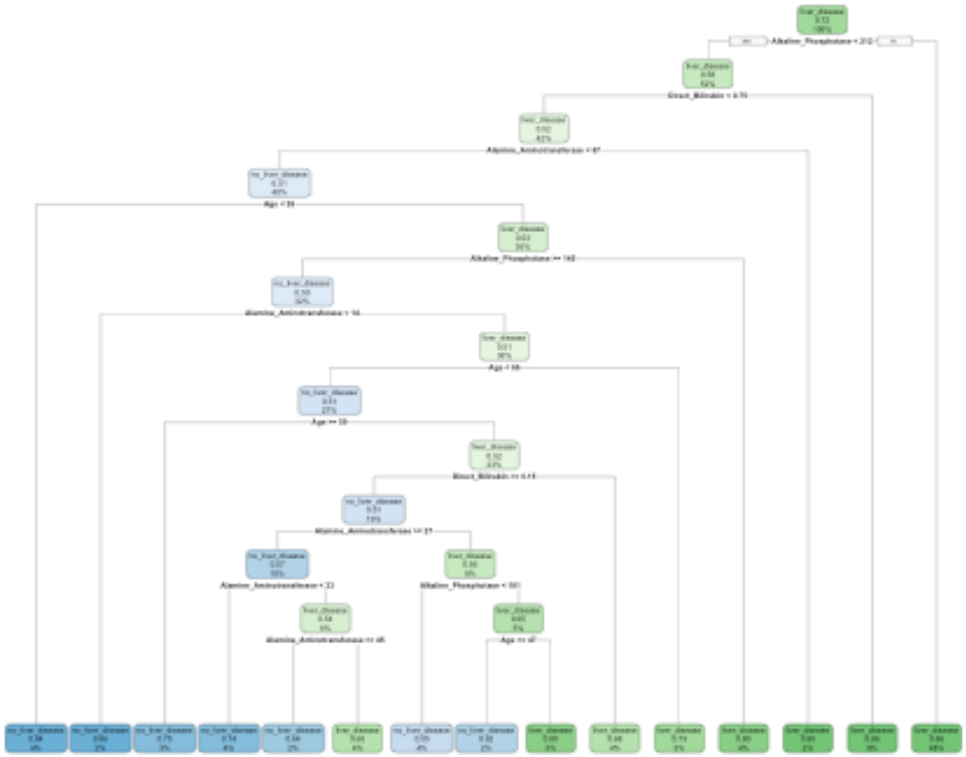



```
## Confusion Matrix and Statistics
##
##                   Reference
## Prediction         no_liver_disease liver_disease
##   no_liver_disease                1             3
##   liver_disease                  32            79
##
##                Accuracy : 0.6957
##                  95% CI : (0.6029, 0.778)
##     No Information Rate : 0.713
##     P-Value [Acc > NIR] : 0.7005
##
##                   Kappa : -0.0085
##
##  Mcnemar's Test P-Value : 2.214e-06
##
##             Sensitivity : 0.030303
##             Specificity : 0.963415
##          Pos Pred Value : 0.250000
##          Neg Pred Value : 0.711712
##              Prevalence : 0.286957
##          Detection Rate : 0.008696
##    Detection Prevalence : 0.034783
##       Balanced Accuracy : 0.496859
##
##        'Positive' Class : no_liver_disease
##
```

This model performed better with the ORIGINAL DATAFRAME - The overall accuracy was 71.30%.

3. Classification and Regression Trees (CART):

ORIGINAL DATAFRAME:

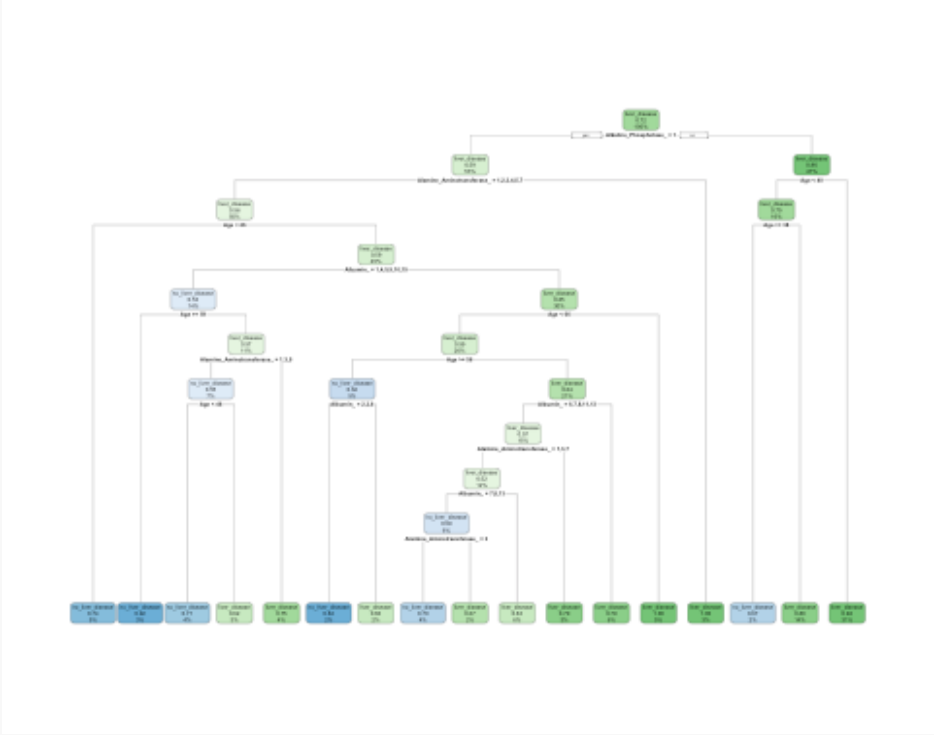


```
## Classification tree:
## rpart(formula = Dataset ~ Direct_Bilirubin + Alkaline_Phosphotase +
##     Alamine_Aminotransferase + Age, data = df.Train, method = "class")
##
## Variables actually used in tree construction:
## [1] Age                        Alamine_Aminotransferase
Alkaline_Phosphotase
## [4] Direct_Bilirubin
##
## Root node error: 132/464 = 0.28448
##
## n= 464
##         CP nsplit rel error xerror     xstd
## 1 0.024621      0   1.00000 1.0000 0.073625
## 2 0.018939     10   0.75000 1.1061 0.075780
## 3 0.015152     12   0.71212 1.1061 0.075780
## 4 0.010000     14   0.68182 1.0682 0.075055
```

```
## Confusion Matrix and Statistics
##
##                   Reference
## Prediction         no_liver_disease liver_disease
##   no_liver_disease               13            16
##   liver_disease                  20            66
##
##                Accuracy : 0.687
##                  95% CI : (0.5938, 0.7702)
##     No Information Rate : 0.713
##     P-Value [Acc > NIR] : 0.7668
##
##                   Kappa : 0.2063
##
##  Mcnemar's Test P-Value : 0.6171
##
##             Sensitivity : 0.3939
##             Specificity : 0.8049
##          Pos Pred Value : 0.4483
##          Neg Pred Value : 0.7674
##              Prevalence : 0.2870
##          Detection Rate : 0.1130
##    Detection Prevalence : 0.2522
##       Balanced Accuracy : 0.5994
##
##        'Positive' Class : no_liver_disease
##
```

ORIGINAL DATAFRAME CATEGORIZATED BY CLUSTERING:

```
## Classification tree:
## rpart(formula = Dataset ~ Alkaline_Phosphotase_ + Alamine_Aminotransfer
ase_ +
##     Albumin_ + Age, data = df.Train, method = "class")
## Variables actually used in tree construction:
## [1] Age                       Alamine_Aminotransferase_
## [3] Albumin_                  Alkaline_Phosphotase_
##
## Root node error: 132/464 = 0.28448
## n= 464
##         CP nsplit rel error xerror     xstd
## 1 0.027778      0   1.00000 1.0000 0.073625
## 2 0.026515      5   0.83333 1.0682 0.075055
## 3 0.018939      7   0.78030 1.0909 0.075496
## 4 0.015152      9   0.74242 1.1136 0.075920
## 5 0.012626     10   0.72727 1.1288 0.076193
## 6 0.011364     14   0.66667 1.1742 0.076968
## 7 0.010000     16   0.64394 1.1742 0.076968
```



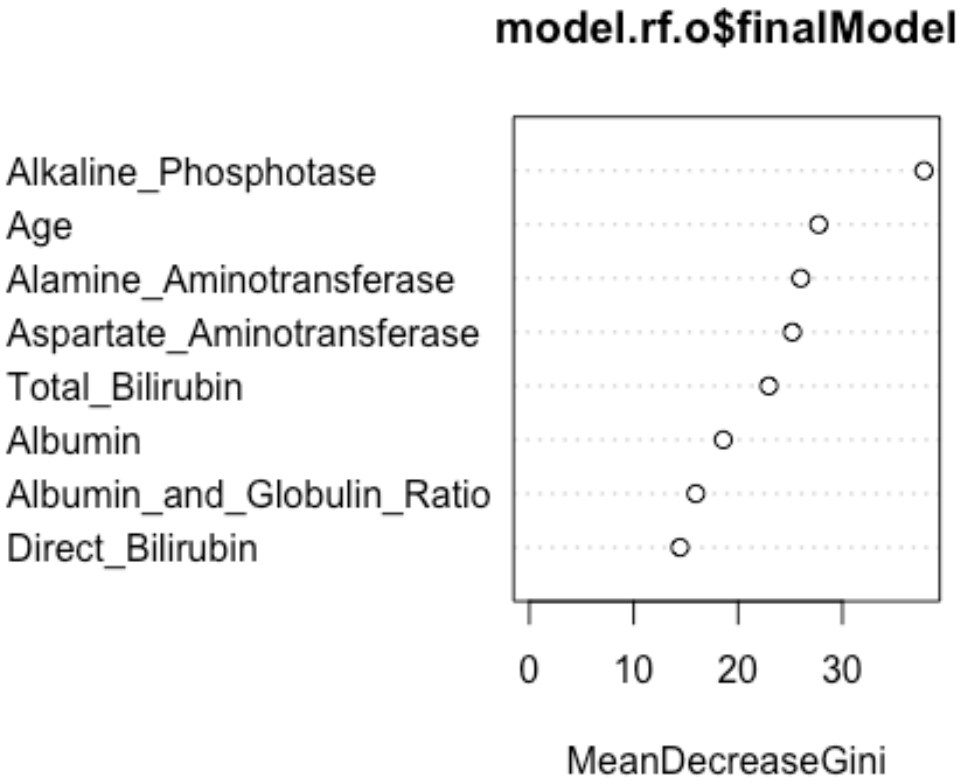
```
Confusion Matrix and Statistics
##                    Reference
## Prediction          no_liver_disease liver_disease
##   no_liver_disease                13            21
##   liver_disease                   20            61
##
##                Accuracy : 0.6435
##                  95% CI : (0.5488, 0.7306)
##     No Information Rate : 0.713
##     P-Value [Acc > NIR] : 0.9579
##
##                   Kappa : 0.1366
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.3939
##             Specificity : 0.7439
##          Pos Pred Value : 0.3824
##          Neg Pred Value : 0.7531
##              Prevalence : 0.2870
##          Detection Rate : 0.1130
##    Detection Prevalence : 0.2957
##       Balanced Accuracy : 0.5689
##
##        'Positive' Class : no_liver_disease
##
```

This model performed better with the ORIGINAL DATAFRAME - The overall accuracy was 68.70%.

4. Random Forest:

ORIGINAL DATAFRAME:

```
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry, importance = FALSE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 31.25%
## Confusion matrix:
##                  no_liver_disease liver_disease class.error
## no_liver_disease               44            88   0.6666667
## liver_disease                  57           275   0.1716867
```




```
## rf variable importance
##
##                            Overall
## Alkaline_Phosphotase       100.000
## Age                         56.839
## Alamine_Aminotransferase    49.489
## Aspartate_Aminotransferase  46.045
## Total_Bilirubin             36.432
## Albumin                     17.744
## Albumin_and_Globulin_Ratio   6.483
## Direct_Bilirubin             0.000
```

The results show that across all of the trees considered in the random forest, the Alkaline_Phosphotase, Age and Alamine_Aminotransferase variables are the three most important variables.

```
## Confusion Matrix and Statistics
##
##                   Reference
## Prediction         no_liver_disease liver_disease
##   no_liver_disease               14            11
##   liver_disease                  19            71
##
##                Accuracy : 0.7391
##                  95% CI : (0.649, 0.8166)
##     No Information Rate : 0.713
##     P-Value [Acc > NIR] : 0.3071
##
##                   Kappa : 0.3127
##
##  Mcnemar's Test P-Value : 0.2012
##
##             Sensitivity : 0.4242
##             Specificity : 0.8659
##          Pos Pred Value : 0.5600
##          Neg Pred Value : 0.7889
##              Prevalence : 0.2870
##          Detection Rate : 0.1217
##    Detection Prevalence : 0.2174
##       Balanced Accuracy : 0.6450
##
##        'Positive' Class : no_liver_disease
##
```

ORIGINAL DATAFRAME CATEGORIZATED BY CLUSTERING:

```
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry, importance = FALSE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 27.8%
## Confusion matrix:
##                  no_liver_disease liver_disease class.error
## no_liver_disease                3           129   0.9772727
## liver_disease                   0           332   0.0000000
```

```
## Confusion Matrix and Statistics
##
##                   Reference
## Prediction         no_liver_disease liver_disease
##   no_liver_disease                1             0
##   liver_disease                  32            82
##
##                Accuracy : 0.7217
##                  95% CI : (0.6305, 0.8013)
```

```
##     No Information Rate : 0.713
##     P-Value [Acc > NIR] : 0.4648
##
##                   Kappa : 0.0427
##
##  Mcnemar's Test P-Value : 4.251e-08
##
##             Sensitivity : 0.030303
##             Specificity : 1.000000
##          Pos Pred Value : 1.000000
##          Neg Pred Value : 0.719298
##              Prevalence : 0.286957
##          Detection Rate : 0.008696
##    Detection Prevalence : 0.008696
##       Balanced Accuracy : 0.515152
##
##        'Positive' Class : no_liver_disease
##
```

This model performed better with the ORIGINAL DATAFRAME - The overall accuracy was 73.91%.

Summary

```
##                 MLA              Dataframe Accuracy Number.of.Predictors
## 1  Log. Regression               ORIGINAL   0.7565               6 of 11
## 2  Log. Regression ORIGINAL CATEGORIZATED   0.7739               7 of 11
## 3               KNN               ORIGINAL   0.7130               6 of 11
## 4               KNN ORIGINAL CATEGORIZATED   0.6957               6 of 11
## 5              CART               ORIGINAL   0.6870               4 of 11
## 6              CART ORIGINAL CATEGORIZATED   0.6435               4 of 11
## 7   Random Forest               ORIGINAL   0.7391              8 of 11
## 8   Random Forest ORIGINAL CATEGORIZATED   0.7217              5 of 11
```

Considering this dataset with few observations and some extreme values, the Logistic Regression model performed better: accuracy value of 77.39% with 7 of 11 predictors.

In general, the models performed with a range accuracy of 64.35% - 77.39% and a number of predictors between 4 and 8.

The data transformation (dealing with extreme values and categorizing data) improved reasonably the performance of some prediction models.

As a tool for doctors diagnosis, we recommend the Logistic Regression model, which showed a good accuracy (73.91% - 77.39%) and a good balance of sensitivity and specificity. By balanced we mean similar levels of performance. Furthermore, this model works well with the original dataframe without any transformation (accuracy = 75,65%). The data transformation can be complex (like the categorization by clustering) and could lead to interpretation difficulties.

References

[1]. Dataset Obtained from UCI, Machine Learning. Our sincere thanks to them.

[2]. Model Building – Obtained from the class labs and rpubs.com.

[3]. EDA – Referenced from the blog of *bdetanico* .

[4]. Inferences – Referenced from towardsdatascience.com and r-gallery.