Liver Disease in Indian Patients

Exploratory Data Analysis Project

Waseem Medhat

This project demonstrates the use of exploratory data analysis (EDA) to search for interesting information or insights within data. Data visualization and statistical summaries are the main tools in EDA, and together they can be very powerful in data analysis.

The data set used in this project contains data on liver disease in Indian patients[1]. It is a relatively small data set with 11 attributes (columns) on 583 patients (rows). The outcome variable is a binary categorical one indicating whether or not a particular case has liver disease, consequently this will be the center around which the exploration will be done. In other words, even when other relationships are considered, the primary goal here is finding any patterns or associations of liver disease with the other variables, which include:

- age,
- gender,
- total bilirubin (in mg/dL),
- direct, i.e. conjugated, bilirubin (in mg/dL),
- alkaline phosphatase (ALP, in IU/L),
- alanine aminotransferase (ALT, in IU/L),
- aspartate aminotransferase (AST, in IU/L),
- total proteins (in g/dL),
- albumin (in g/dL), and
- albumin/globulin (A/G) ratio.

There are usually preliminary data preparation/cleaning steps that should be done first. For this data set:

- There are spelling mistakes in the column names `Total_Protiens`, `Alamine_Aminotransferase`, and `Alkaline_Phosphotase`, which were corrected to `Total_Proteins`, `Alanine_Aminotransferase` and `Alkaline_Phosphatase`, respectively.
- For the sake of clarity, a new column named `Diseased` was created from the `Dataset` variable, which corresponds to the disease categories, with its levels being “Yes” for diseased cases and “No” for healthy ones. The original variable was removed from the data set.
- Other derived columns were created in the analysis process as transformations (base 10 log) of skewed variables to be able to analyze them.

---

1 Univariate Analysis

1.1 Data Types

```
## 'data.frame':    583 obs. of  11 variables:
```

[1] Data source: Kaggle

```
##  $ Age                       : int  65 62 62 58 72 46 26 29 17 55 ...
##  $ Gender                    : chr  "Female" "Male" "Male" "Male" ...
##  $ Total_Bilirubin           : num  0.7 10.9 7.3 1 3.9 1.8 0.9 0.9 0.9 0.7 ...
##  $ Direct_Bilirubin          : num  0.1 5.5 4.1 0.4 2 0.7 0.2 0.3 0.3 0.2 ...
##  $ Alkaline_Phosphatase      : int  187 699 490 182 195 208 154 202 202 290 ...
##  $ Alanine_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_Proteins            : num  6.8 7.5 7 6.8 7.3 7.6 7 6.7 7.4 6.8 ...
##  $ Albumin                   : num  3.3 3.2 3.3 3.4 2.4 4.4 3.5 3.6 4.1 3.4 ...
##  $ Albumin_and_Globulin_Ratio: num  0.9 0.74 0.89 1 0.4 1.3 1 1.1 1.2 1 ...
##  $ Diseased                  : chr  "Yes" "Yes" "Yes" "Yes" ...
```

Most of our variables are numeric, and they represent various blood concentrations of a particular compound or enzyme along with an age variable. There are two remaining categorical variables: disease category and gender, both of which are binary.

1.2 Missing Values

We can start by analyzing each variable on its own to take a closer look at the structure of the data set. But before starting the analysis, we should check the data set for missing values.

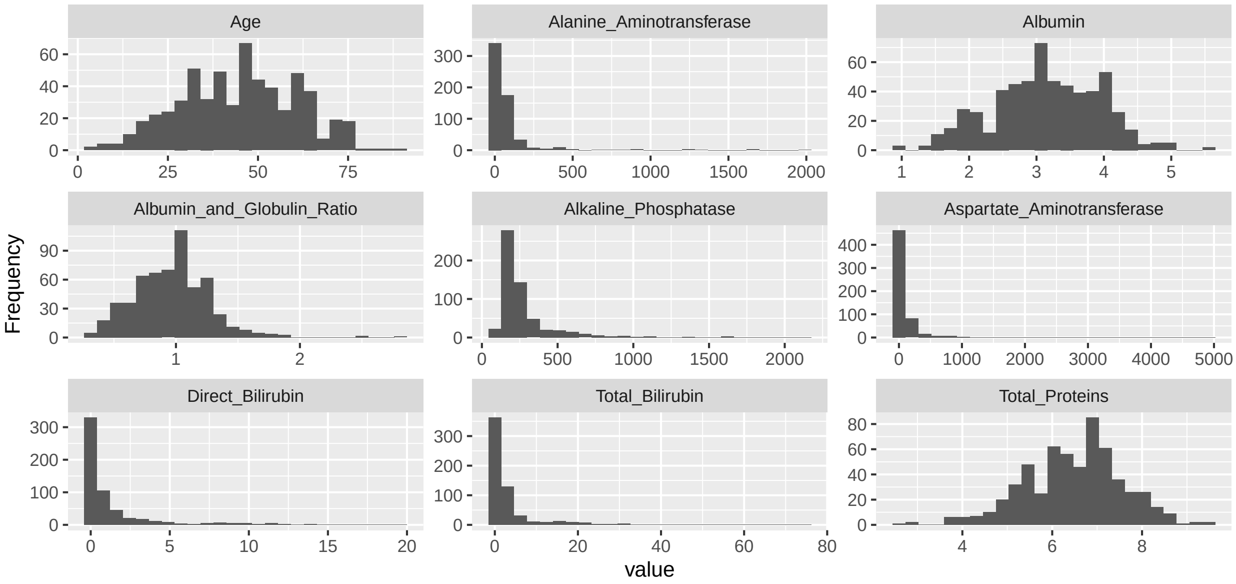
	number_missing	percent_missing
Age	0	0.00
Gender	0	0.00
Total_Bilirubin	0	0.00
Direct_Bilirubin	0	0.00
Alkaline_Phosphatase	0	0.00
Alanine_Aminotransferase	0	0.00
Aspartate_Aminotransferase	0	0.00
Total_Proteins	0	0.00
Albumin	0	0.00
Albumin_and_Globulin_Ratio	4	0.69
Diseased	0	0.00

There are no missing data in all variables except A/G ratio, and even the missing values in that variable constitute 0.69% of the data (4 cases). Therefore, it is safe to use all variables in the analysis and drop the missing values when necessary without affecting our analysis.

But it is still better to think about why these values were missing. One possible explanation comes from the fact that it is a ratio of albumin to globulin levels. We do not have the globulin data, but they could have been missing which resulted in A/G ratio being missing as well.

1.3 Numeric Variables

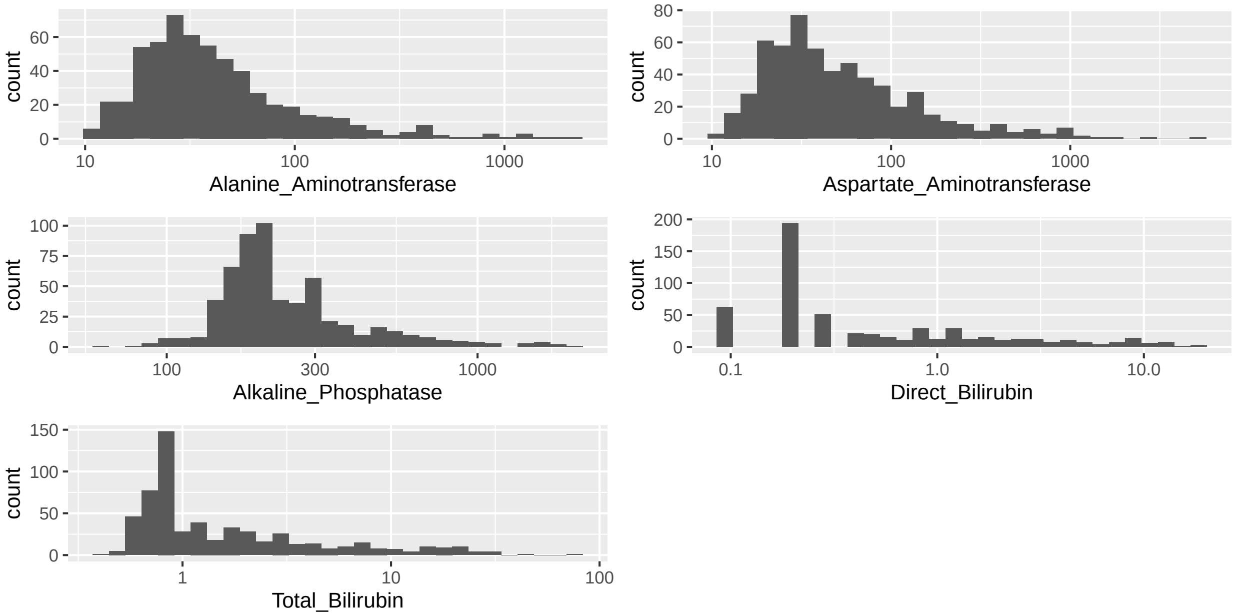
Histograms are a good method to examine the distributions of numeric variables.



It seems that age, albumin, and total proteins have roughly symmetric distributions. A/G ratio has a few outliers. These will be worth looking into further in the analysis.

```
## Outliers
##  2.5: 2
##  2.8: 1
```

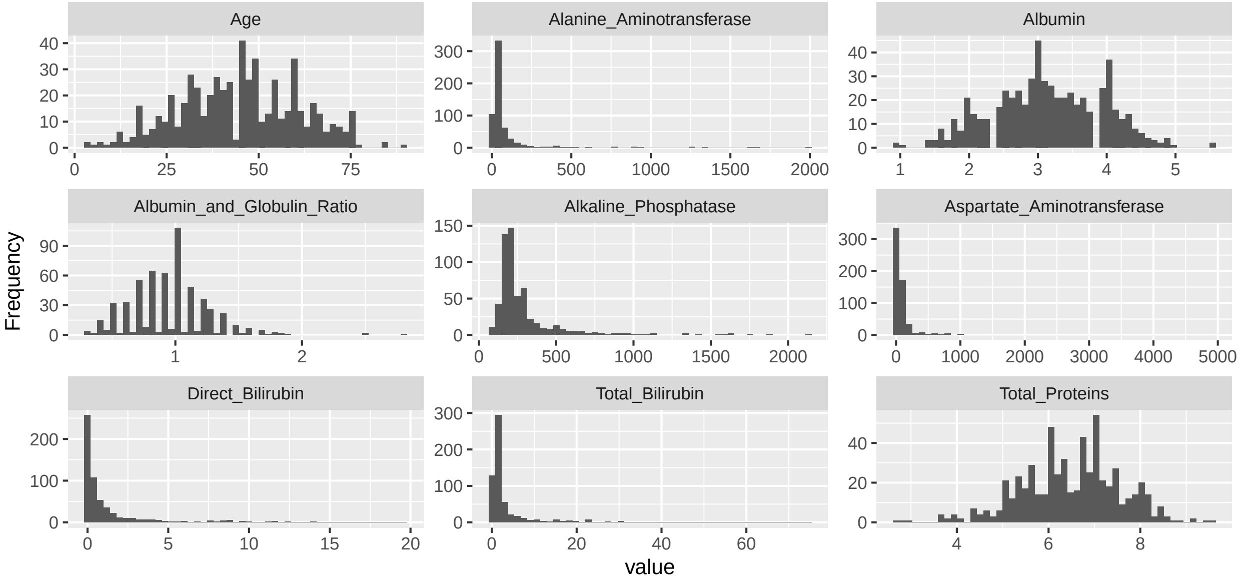
The rest of the variables, which are either bilirubin levels or liver enzymes, are heavily right skewed. We can recreate the plots of these variables with a log-transformed X axis to have a better look at their distribution.



Even with a log transformation, the histograms still show long tails. Also, total and direct bilirubin distributions look peculiar: each has a very high spike at a single bin which corresponds to a value of 0.8 mg/dL for total bilirubin and 0.2 mg/dL for direct bilirubin. The histogram of direct bilirubin shows some discreteness in the lower bin values which is only a result of the data being recorded/rounded with one decimal precision.

```
## Direct Bilirubin Frequencies (< 1 mg/dL)
##  0.1: 63
##  0.2: 194
##  0.3: 51
##  0.4: 21
##  0.5: 20
##  0.6: 16
##  0.7: 11
##  0.8: 22
##  0.9: 7
```

Before leaving the histograms, it might be a good idea to decrease the binwidth as it might uncover some hidden details.

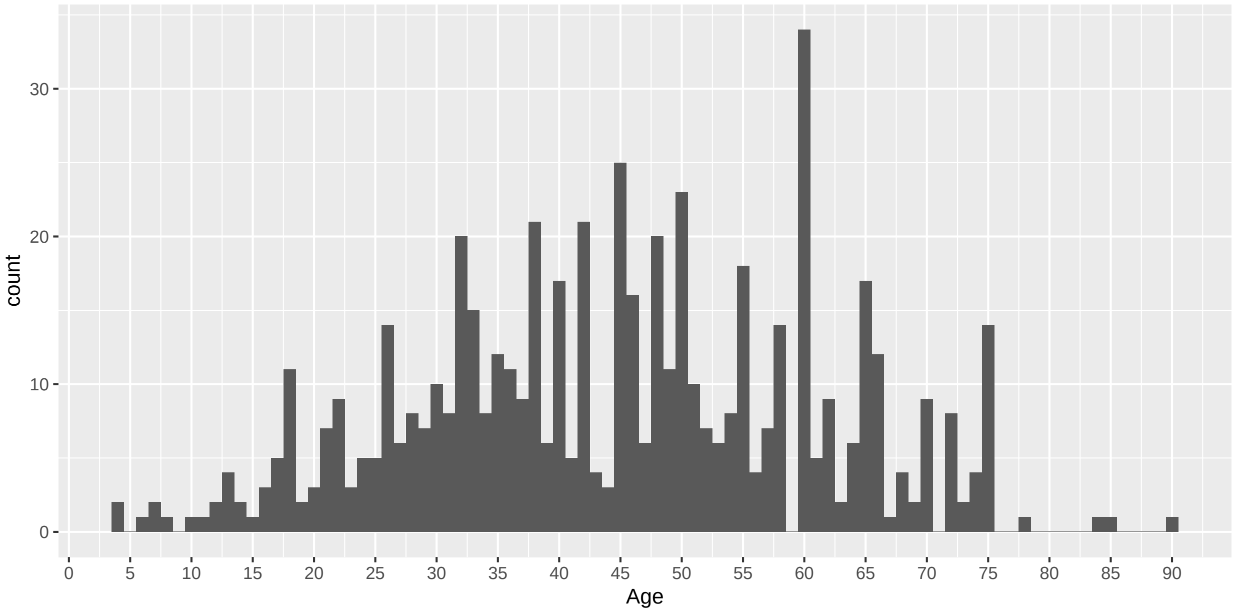


We can quickly notice the discreteness in the distribution of A/G ratio which results from the same cause as bilirubin levels. Moreover, for the distribution of albumin, the same peaks can be observed before and after decreasing the binwidth, which means that the most common albumin levels do not lie in ranges but specific values. Examining the exact numbers will give us a better idea.

```
## Albumin Frequencies (Top 10 Values)
##  3: 45
##  4: 37
##  2.9: 29
##  3.1: 28
##  3.2: 26
##  3.9: 25
##  2.5: 24
##  2.7: 24
##  3.5: 23
##  2: 21
```

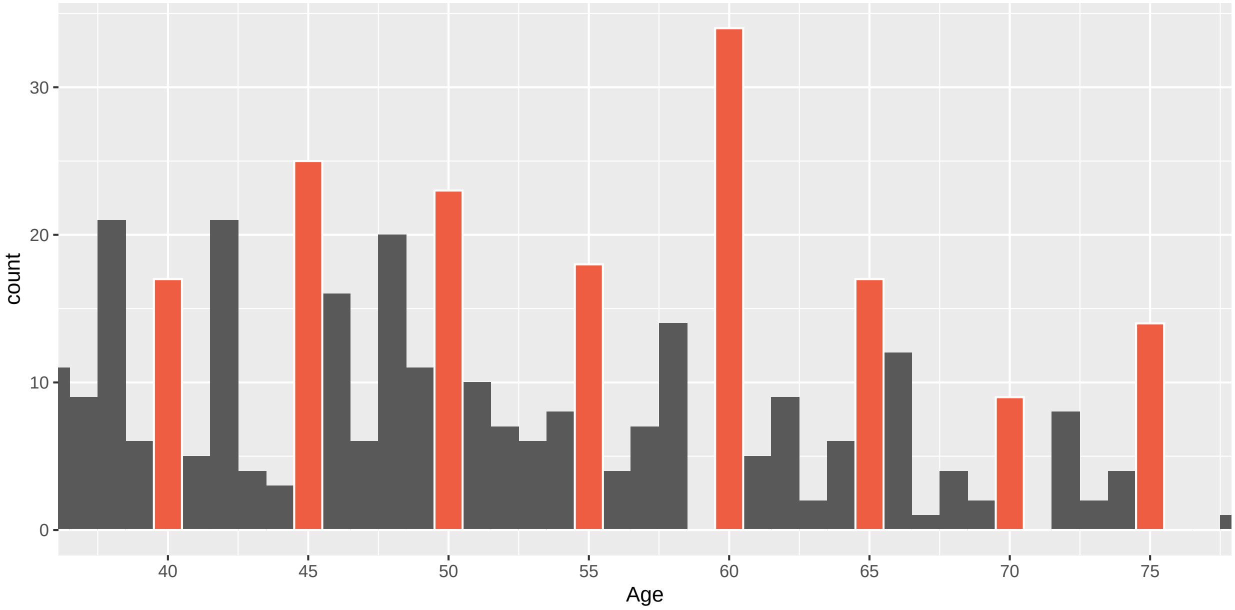
The most common values are 3 and 4 mg/dL. Even the next most common values lie near those values.

Now, we take a closer look at the distribution of ages by using a bin for each single year (i.e. a binwidth of 1).



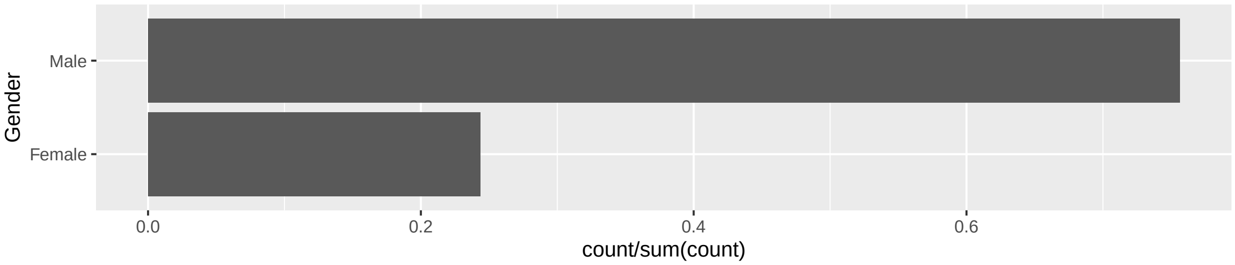
Interestingly, ages that are multiples of 5 are much more common than the adjacent ones. This pattern can be found starting from 40 years. Could it be that some old people report their ages rounded to the nearest 5? It might seem a trivial topic to investigate compared to liver disease, but it is interesting nonetheless.

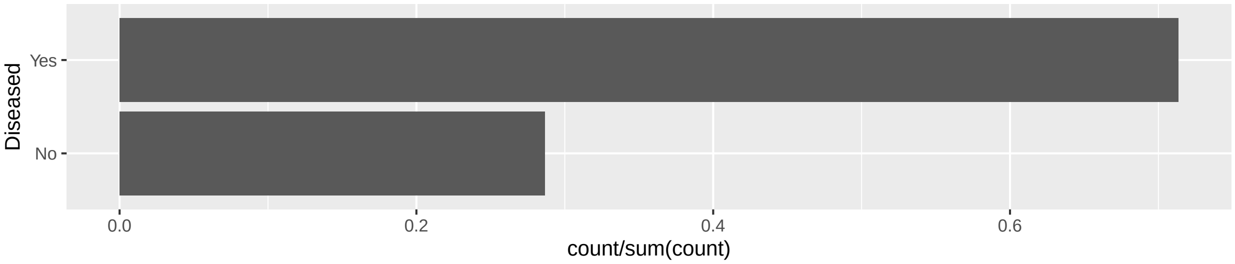
The next plot zooms in on the last one and adds color emphasis on multiples of 5 to make it easier to look at the pattern.



1.4 Categorical Variables

The data set has only two categorical variables: gender and disease state. We can look at frequencies and proportions of their categories using numerical summaries and bar charts.





```
## Gender
##  Female: 142
##  Male: 441
##
## Diseased
##  No: 167
##  Yes: 416
```

Looking at the gender data, there are more males than females. Similarly, there are more liver patients than healthy people.

Univariate Analysis Insights

This section serves as a reflection that connects findings in the data with their actual implications. We use what we have found to build some intuitions to pave the road for the next phase in the EDA process.

We saw that bilirubin and liver enzyme data have almost the same distribution pattern. This is due to a property they all have in common: their blood levels are very low in people with healthy livers but increase with liver disease. Bilirubin levels are kept low by the liver as it excreted by the liver, while the enzymes are normally present inside the liver cells. Either way, these compounds accumulate in the blood with liver damage and show high levels in lab tests[2]. Because there is no upper limit to these values, their frequencies are scattered across a very broad range, while normal values are condensed in a very narrow range, causing these skewed distributions even though we have many more patients than healthy people. Another thing to be noted is that the liver is the main source of albumin[3], but this is not the case for globulin (not every type of globulin is produced in the liver)[4]. This implies that low albumin levels, and consequently A/G ratio, correspond to liver damage.

---

[2] http://www.clevelandclinicmeded.com/medicalpubs/diseasemanagement/hepatology/guide-to-common-liver-tests/

[3] https://en.wikipedia.org/wiki/Albumin#Serum_albumin

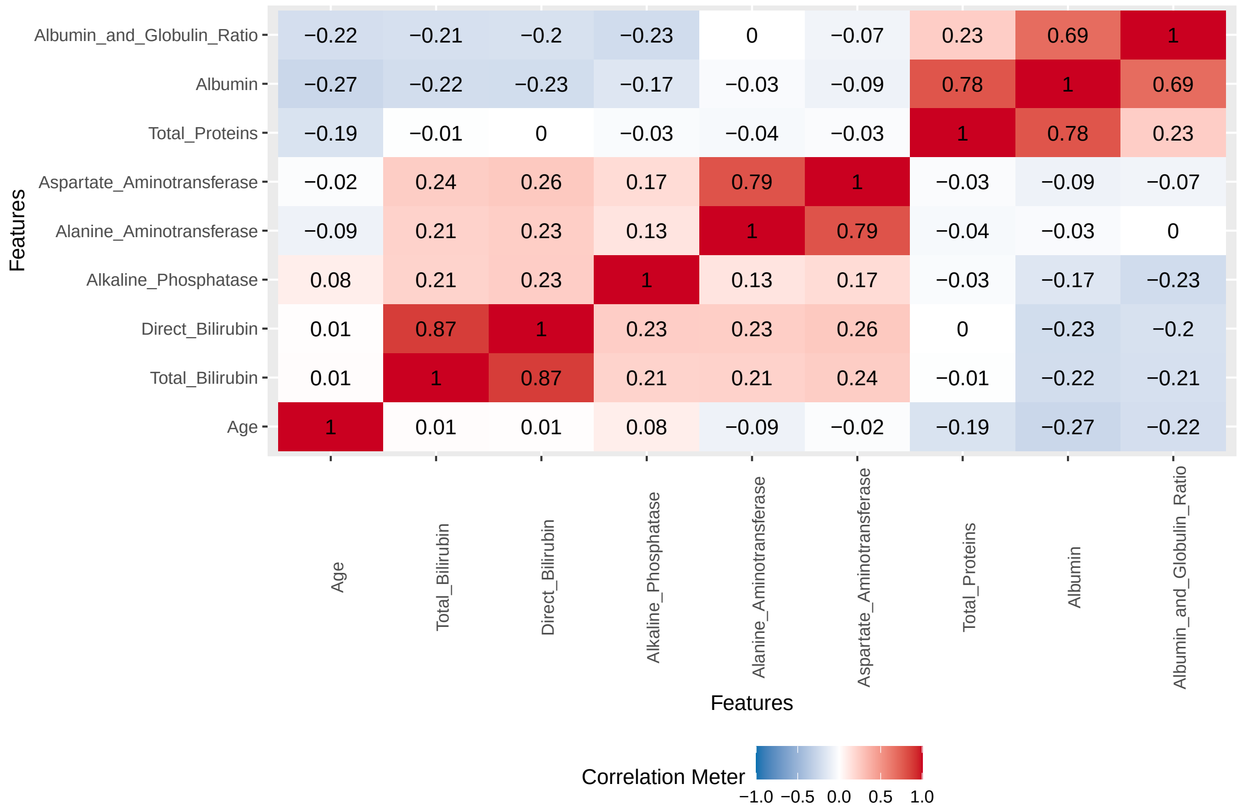
[4] https://en.wikipedia.org/wiki/Globulin

2 Bivariate Analysis

After using univariate analysis to get a sense of the structure of the data set and the information contained within each variable, it is time to dig deeper into the relationships between variables. As stated earlier, the main focus is on the disease state as an outcome variable, but other relationships between explanatory variables will be examined as well.

2.1 Correlations

A correlation matrix can be a very concise way to look at relationships between numeric variables.

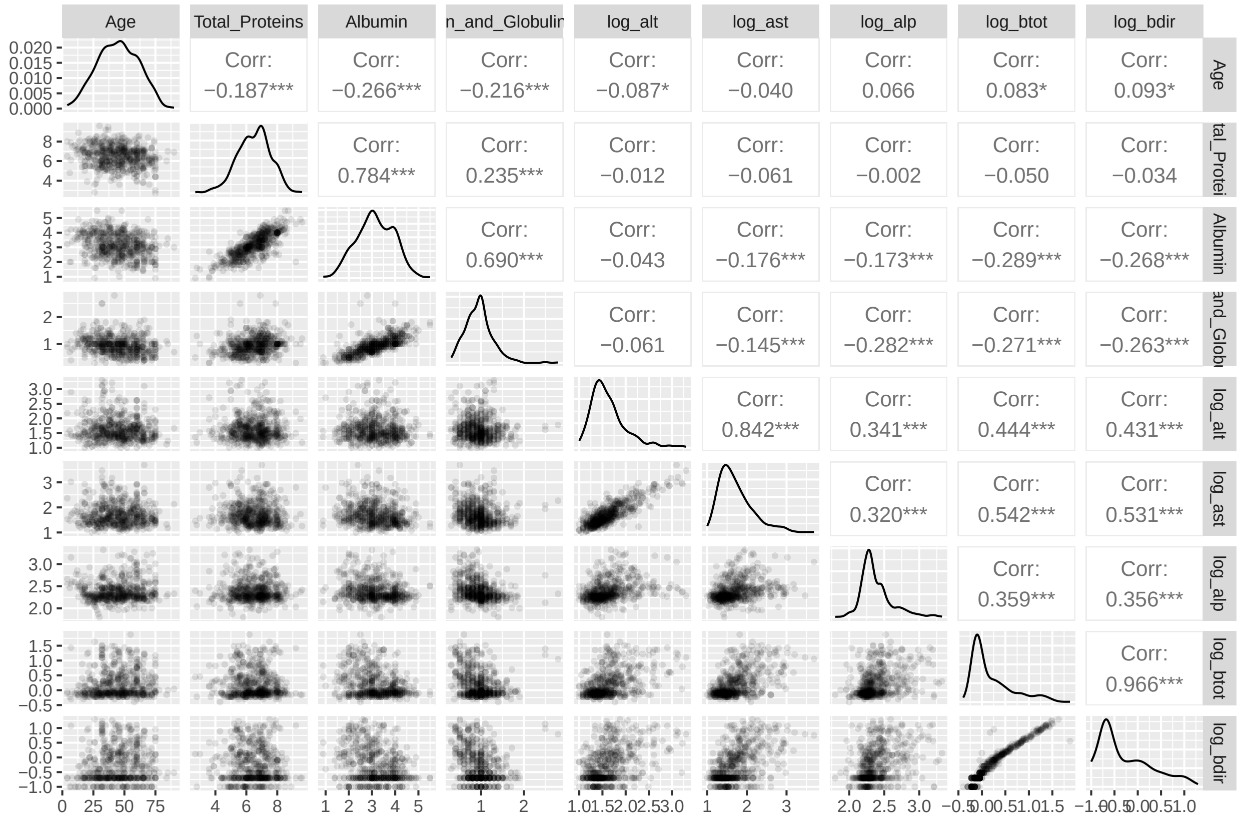


	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphatase	Alanine_Aminotransferase	Aspartate_Aminotransferase	Total_Proteins	Albumin	Albumin_and_Globulin_Ratio
Albumin_and_Globulin_Ratio	−0.22	−0.21	−0.2	−0.23	0	−0.07	0.23	0.69	1
Albumin	−0.27	−0.22	−0.23	−0.17	−0.03	−0.09	0.78	1	0.69
Total_Proteins	−0.19	−0.01	0	−0.03	−0.04	−0.03	1	0.78	0.23
Aspartate_Aminotransferase	−0.02	0.24	0.26	0.17	0.79	1	−0.03	−0.09	−0.07
Alanine_Aminotransferase	−0.09	0.21	0.23	0.13	1	0.79	−0.04	−0.03	0
Alkaline_Phosphatase	0.08	0.21	0.23	1	0.13	0.17	−0.03	−0.17	−0.23
Direct_Bilirubin	0.01	0.87	1	0.23	0.23	0.26	0	−0.23	−0.2
Total_Bilirubin	0.01	1	0.87	0.21	0.21	0.24	−0.01	−0.22	−0.21
Age	1	0.01	0.01	0.08	−0.09	−0.02	−0.19	−0.27	−0.22

Some unsurprisingly strong correlations are direct bilirubin with total bilirubin (the former is only a subset of the latter), AST with ALT (both are liver enzymes), and albumin (being the most abundant protein) with total proteins.

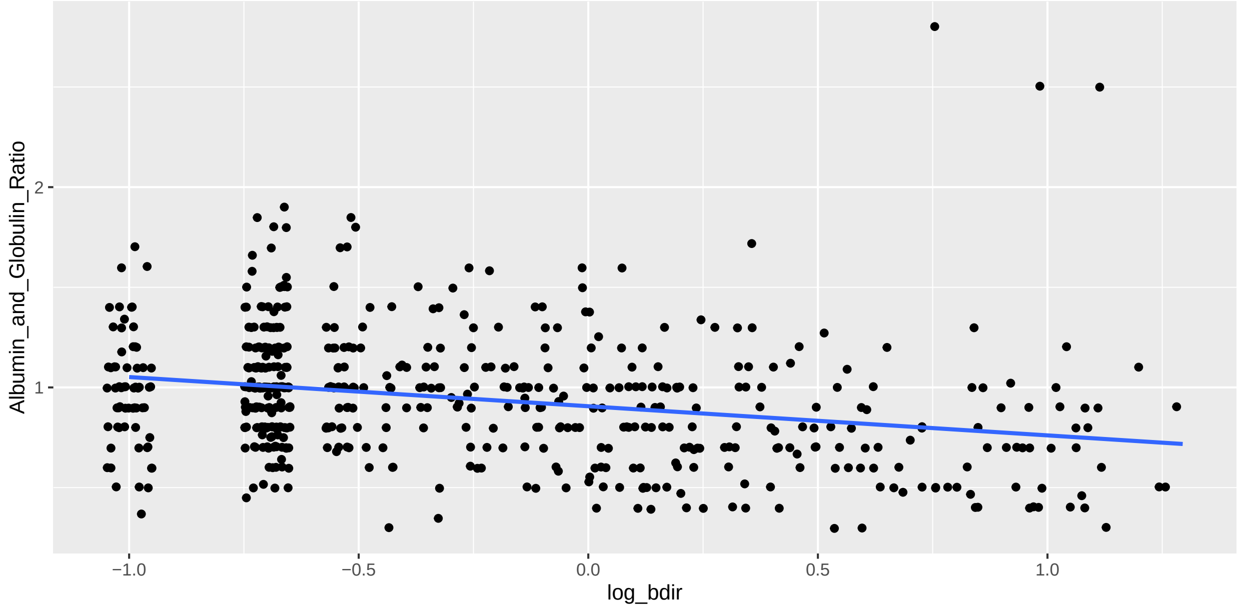
The correlation between albumin and A/G ratio is 0.69 which is also high but not as high as other strong correlations because globulin introduces some variability in the ratio that is not shared with albumin.

Although the correlation matrix is informative, it is better to look at scatterplots to examine the relationships in more detail or to ensure that no non-linear trends were missed. (Note: because bilirubin and liver enzymes are heavily skewed, they were log-transformed to clarify the scatterplots.)



In this particular case, no important insights can be gleaned from the scatterplots that the correlation coefficients did not show. However, there is something to notice: in the A/G ratio vs. (log) direct bilirubin, the outliers in A/G ratio lie in a position that would increase the slope of the fitted line (i.e. result in a more positive correlation). As a result, the correlation coefficient of -0.26 between A/G ratio and direct bilirubin could be biased by the outliers.

Here is a single scatterplot for clarification. (Note: the points were jittered to avoid overplotting.)

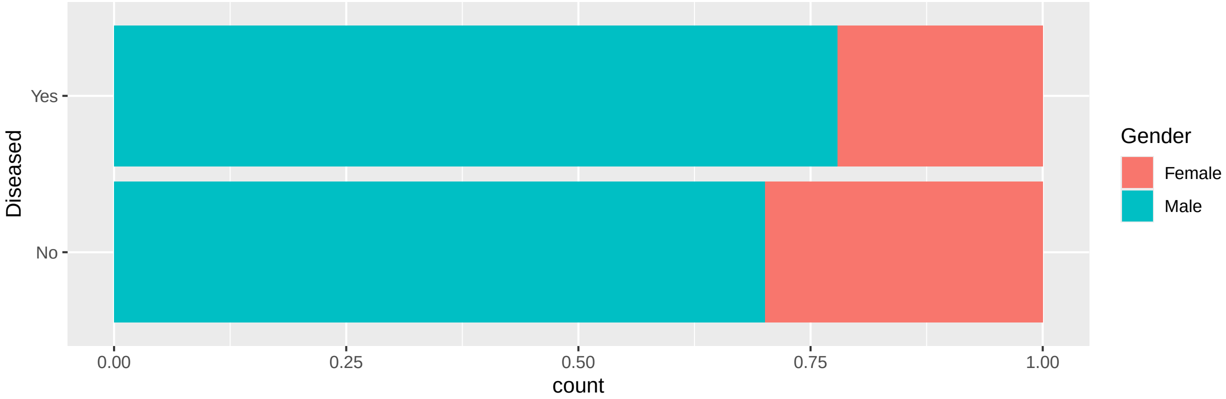


We can see a slightly negative trend. Higher values of direct bilirubin are associated, even if weakly, with lower A/G ratio. This makes sense since low A/G ratio and high direct bilirubin are both indicators of liver disease.

If we recalculated the correlation coefficient between these 2 variables after removing the outliers, we would get a value of -0.35, which indicates a stronger correlation (to the negative side). It is important, however, to carefully examine those outliers before deciding to remove them. Outliers are not necessarily erroneous; maybe they are a special group of observations that should be analyzed "in isolation" instead of completely dismissing them.

2.2 Gender vs. Disease

We might want to know if liver disease is more common in males or females. Also, in later steps we may analyze variables that are grouped by both gender and disease state. So we might first examine the distribution of genders across the two disease states.

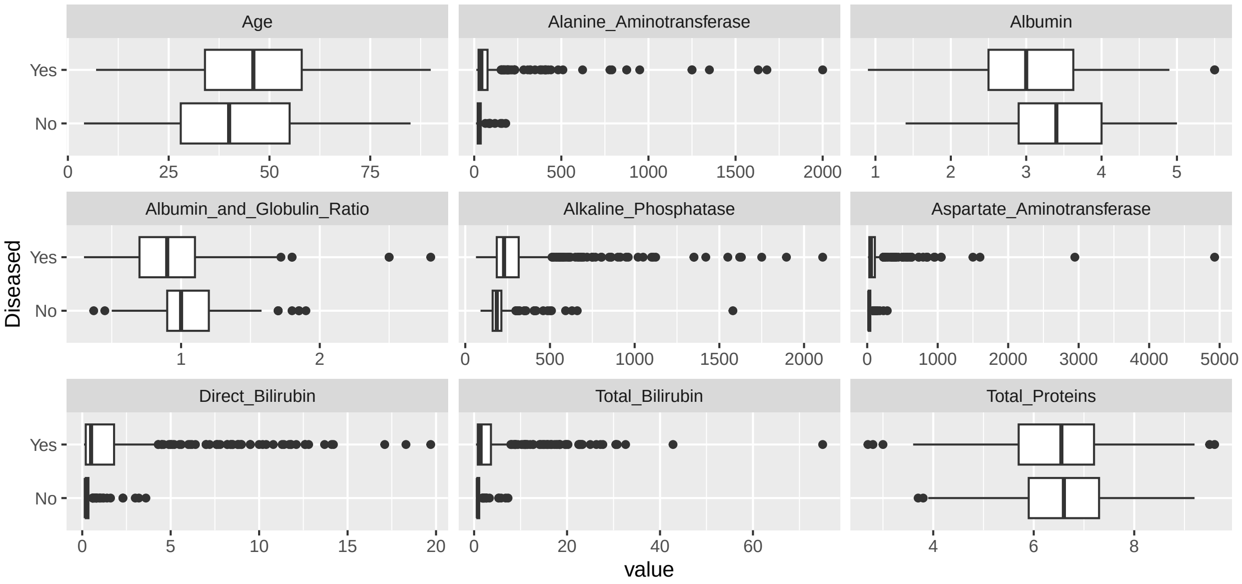


The proportion of males is not substantially different across the two states. Both roughly reside near 0.75. We can also look at counts using a crosstab.

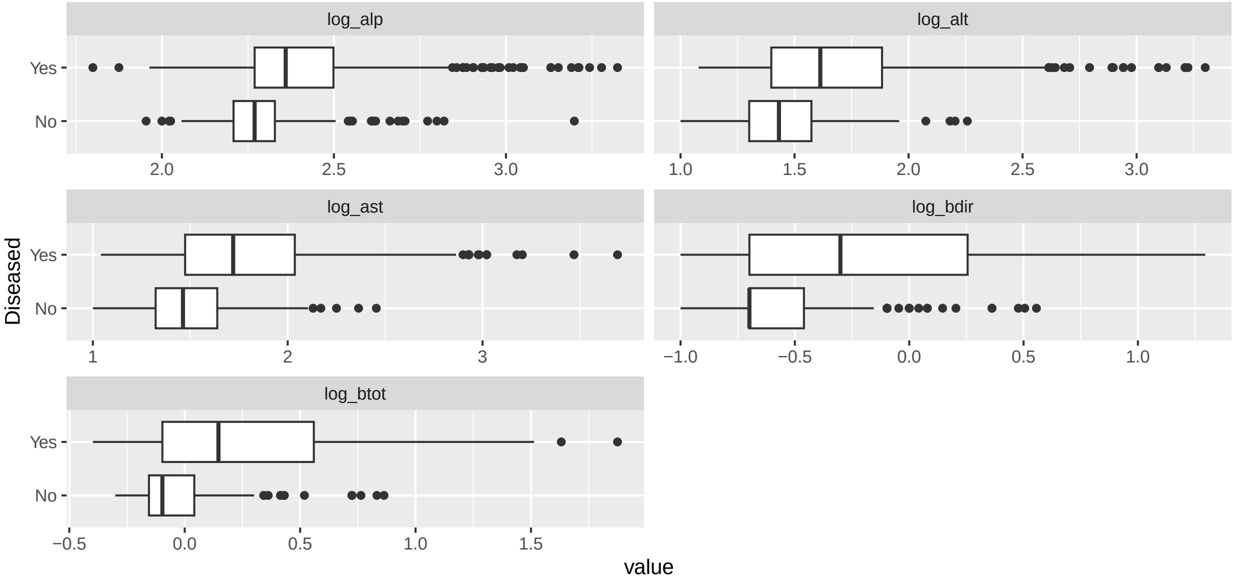
```
##         Gender
## Diseased Female Male
##      No      50  117
##      Yes     92  324
```

2.3 Numeric Variables vs. Disease

Here we examine age and lab tests and see how their values differ between people who have liver disease and those who do not. This is where we test if the intuitions developed in the univariate analysis phase are correct.



- **Age:** Liver disease patients are typically older than healthy people. This can be mainly due to the fact that liver diseases are chronic.
- **Albumin:** Since, as mentioned above, albumin is produced by the liver, a decrease in its levels is expected to be associated with the presence of liver disease.
- **A/G ratio:** The same concept with albumin applies here. However, the outliers spotted earlier seem to represent liver disease patients. This does not align with our intuition or the rest of the data.
- **Total proteins:** There is only a slight change in total proteins between diseased and healthy people.
- **Bilirubin and liver enzymes:** It is clear that their values stretch to a high range in diseased patients. But in the case of ALP, there is one outlier in the healthy people with an unusually high value.

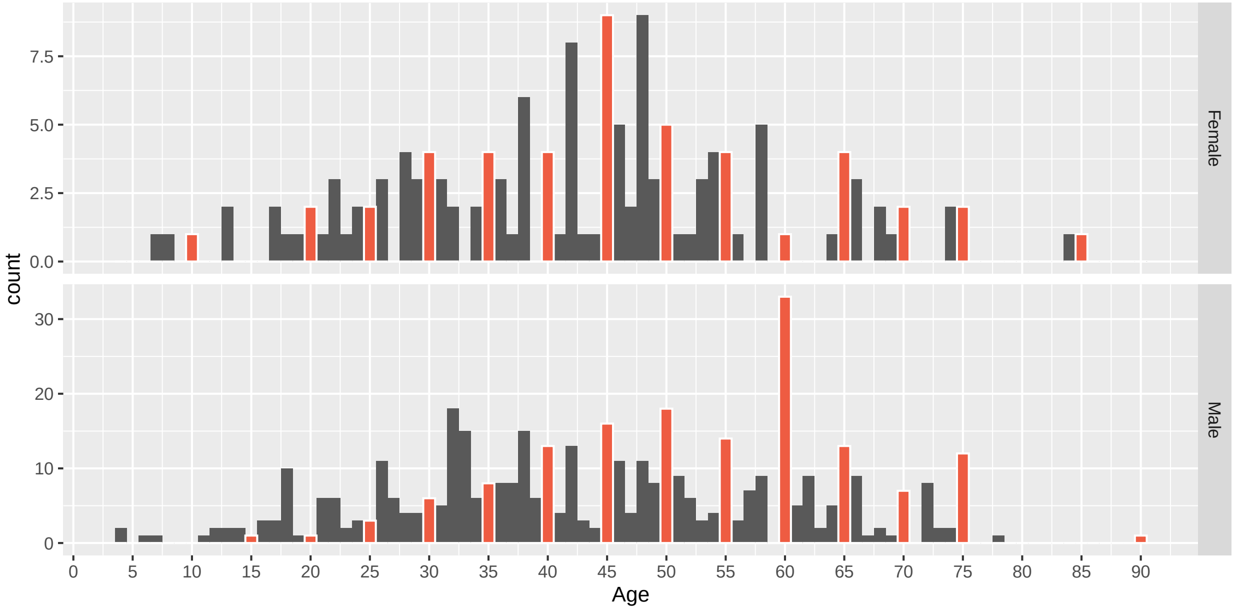


Since boxplots do not show the means, we can separately summarize them into a table.

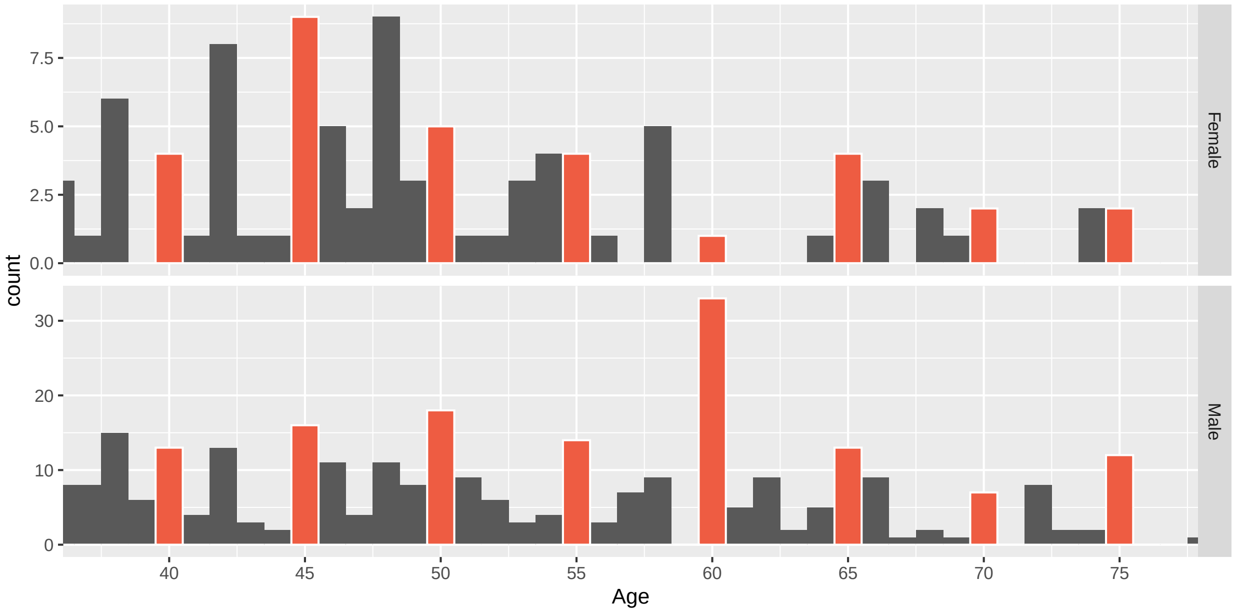
	No	Yes
Age	41.2395210	46.1538462
Total_Bilirubin	1.1425150	4.1644231
Direct_Bilirubin	0.3964072	1.9235577
Alkaline_Phosphatase	219.7544910	319.0072115
Alanine_Aminotransferase	33.6526946	99.6057692
Aspartate_Aminotransferase	40.6886228	137.6995192
Total_Proteins	6.5431138	6.4591346
Albumin	3.3443114	3.0605769
Albumin_and_Globulin_Ratio	1.0295758	0.9141787

2.4 Age vs. Gender

We found in the univariate analysis that ages that are multiples of 5 are particularly common. We can further refine this by splitting the distribution across genders and see how they differ.



Zooming in. . .



We can see that the pattern we found in the univariate histogram is more visible in males than in females.

Bivariate Analysis Insights

Through bivariate analysis, we started understanding relationships between variables in the data set and testing our intuitions as we found with the values of blood tests. We also refined findings of univariate analysis when we found that the age pattern is more visible in 40+ year old males.

We also discovered an interesting outlier in which a non-diseased case had high ALP level. The most likely explanation is that we do not know if the cases that have no liver disease are actually free from other conditions, and ALP level is not associated only with liver diseases[5]. Besides the new outlier, we found new information about the previously discovered ones. Similar to the ALP outlier, there were 3 unexpectedly high A/G ratios in patients with no liver disease. Again these unusual values could be attributed to other conditions[6].

A word of caution: all theses speculations assume that there were no errors in measurement or data entry. In situations where it is feasible to consult those responsible for data collection, it would be advisable to do so.

---

3 Multivariate Analysis

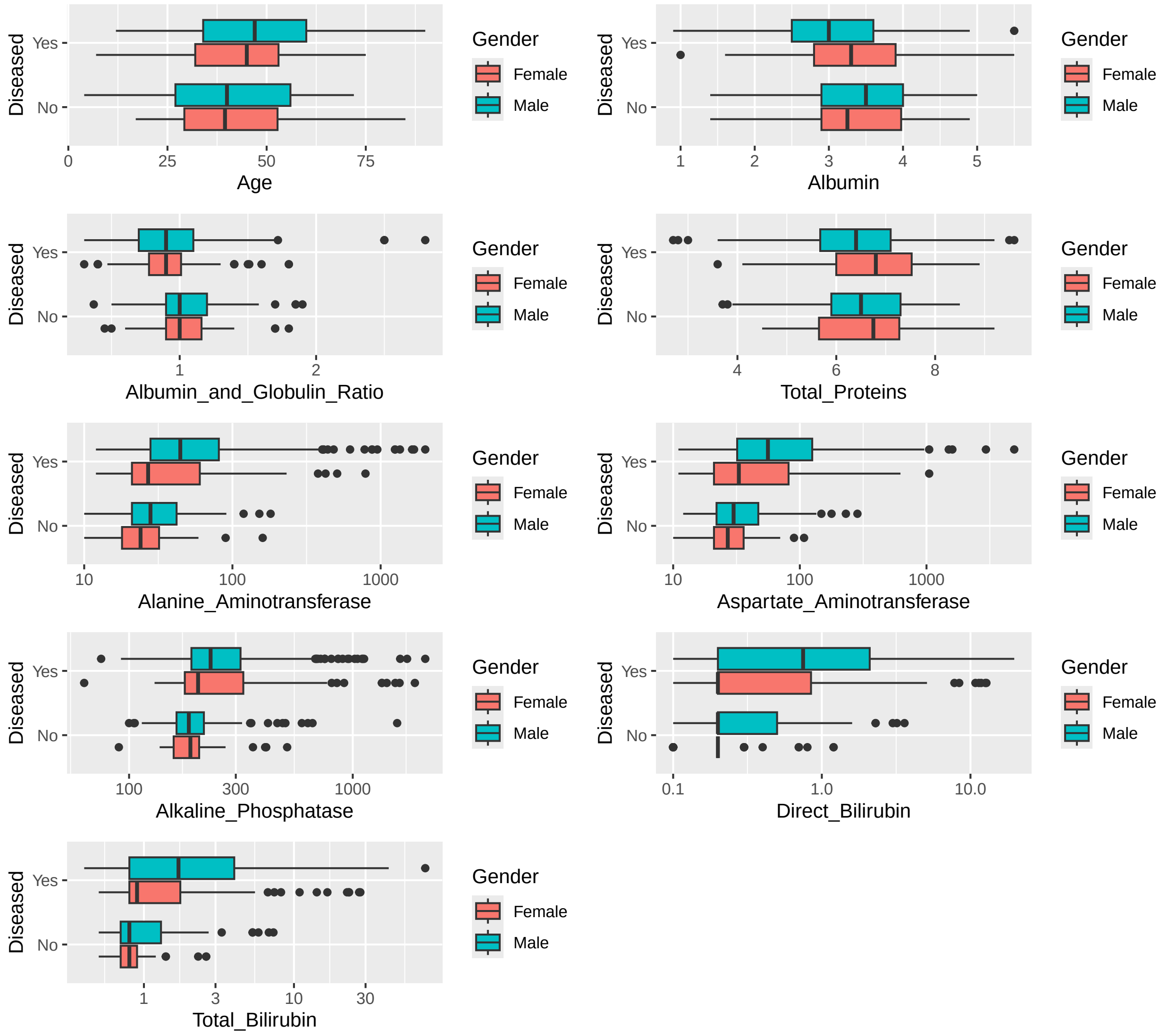
At this point of the EDA process, important relationships and features of the data have been discovered. Next, we explore more complex relationships and interactions that involve more than two variables.

3.1 Numeric Variables vs. Gender and Disease

The boxplots created earlier are split (using color) by gender, and scales were log-transformed when necessary.

[5] https://www.healthline.com/health/alp#uses

[6] https://www.idexx.eu/globalassets/documents/parameters/8090-us-agratio-interpretive-summary.pdf

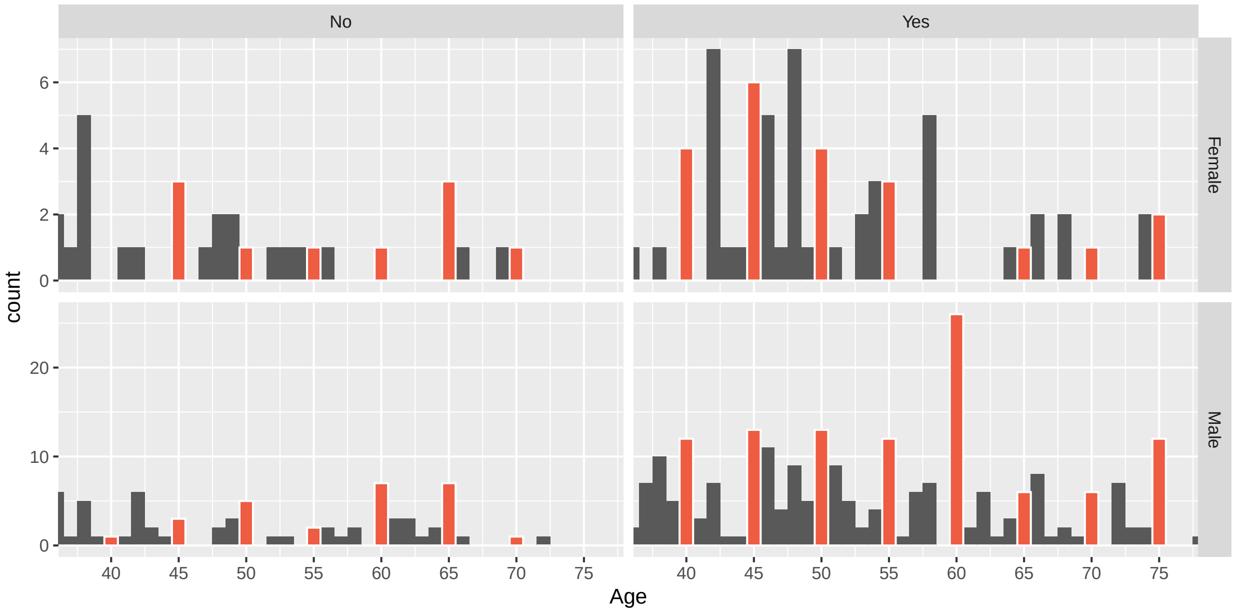


In most cases, non-diseased men and women are not substantially different. However, diseased males have typically have lower levels of albumin and higher levels of bilirubin and enzymes than diseased females. Also, it seems as if there is no box in the direct bilirubin distribution for healthy females. This indicates that the 3 quartiles (1st, median, and 3rd) are equal. We can back this up by summary statistics.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.200   0.200   0.268   0.200   1.200
```

3.2 Age vs. Gender and Disease

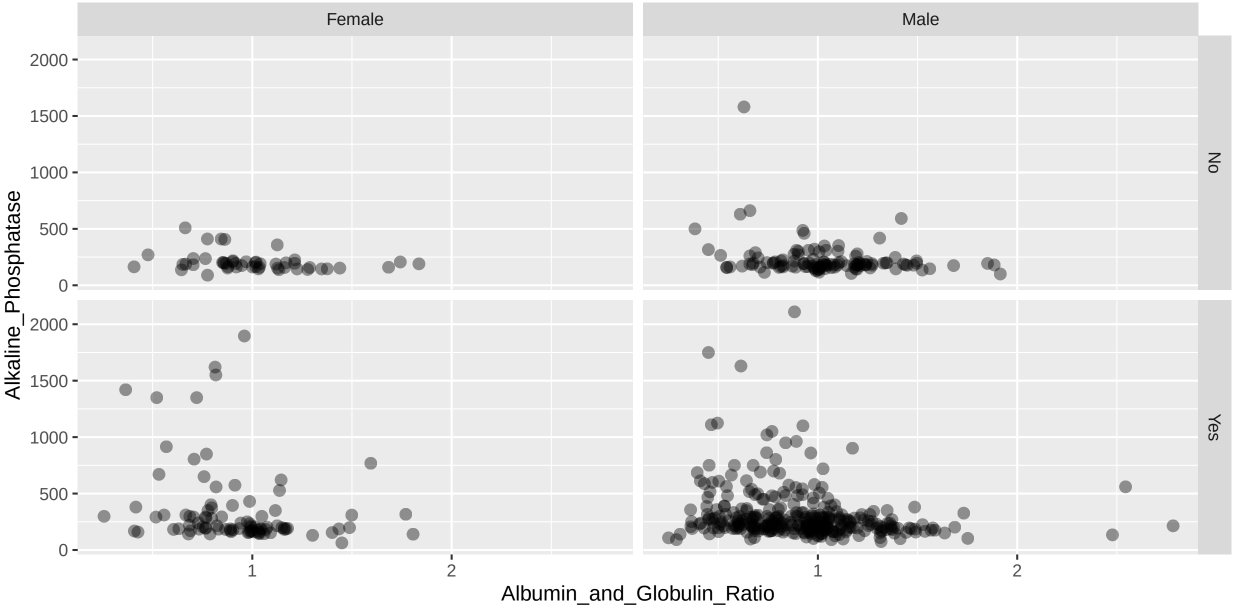
We can build up on the age vs. gender histograms by faceting by the disease variable as well.



Unfortunately, for a data set this small, we over-split the age variable. Frequencies in the non-diseased cases are all below 10 which is too low to make a solid judgment, so we cannot go beyond what we found in the bivariate analysis.

3.3 Final Look at Outliers

Previously we found outliers in A/G ratio and another one in ALP. The next plot is a scatterplot of the two variables faceted by both gender and disease. This will give us an idea if the outliers are related in any way. (This time ALP was used without log-transformation because we are interested in the outliers.)



The 3 A/G ratio outliers are males with liver disease and have seemingly typical ALP values. The ALP outlier is a male with no liver disease and also has a typical A/G value. This gives us no evidence that these outliers are related.

Multivariate Analysis Insights

We discovered that blood test results might differ between males and females that do have liver disease. This could point out, without being a solid proof, to a conclusion that males are affected by liver disease more severely than females or that there is a confounder that is associated with gender (e.g. physiological differences or behavioral risk factors) and cause this difference.

Other multivariate explorations were dead ends. The data set was too small to explore age across two factors, and the suspicion that different outliers are related turned out to be false.

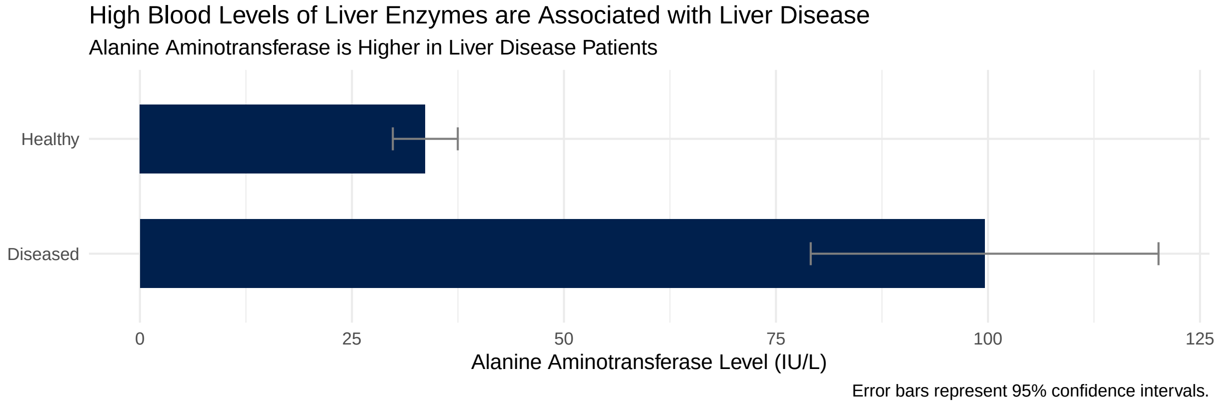
---

4 Final Plots and Summary

In this section the paradigm switches from exploratory, where we search for information, to explanatory, where we present information. This switch will have an effect on the structure and form of the plots we create which will be the final presentables.

4.1 First Plot

This plot shows the mean ALT level in patients vs. healthy people. It summarizes the findings we found in the boxplots of liver enzymes grouped by disease category.

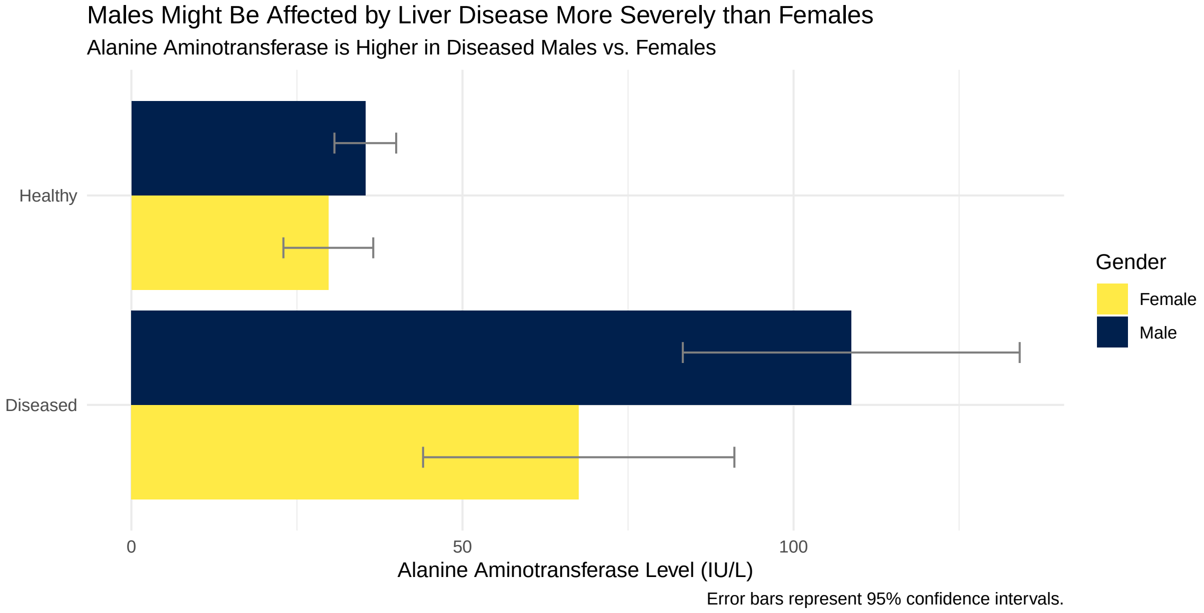


The following points are the choices made in the plot with and the reasons behind those choices:

- Instead of boxplots, like the exploratory plots, bar charts are easier to look at and interpret as it is known by a larger number of people from different domains and levels of expertise.
- Only a single plot of a single enzyme was used. Since the plot is meant to deliver a single piece of information, using multiple plots for each enzyme is redundant.
- There are a title and a subtitle in order to make the plot informative on its own without having to refer to any text. The title shows the main conclusion, and the subtitle tells the finding which led to that conclusion.
- Instead of putting the disease categories as Yes/No, names were changed to “Diseased” and “Healthy” which is a slight addition to both overall polish and ease of interpretation.
- Error bars were added to remedy the loss of information that is due to switching from a boxplot to a bar chart. A caption was added to inform that the error bars represent 95% confidence intervals.

4.2 Second Plot

The next plot shows how in liver disease the mean ALT increases more significantly in males than females.

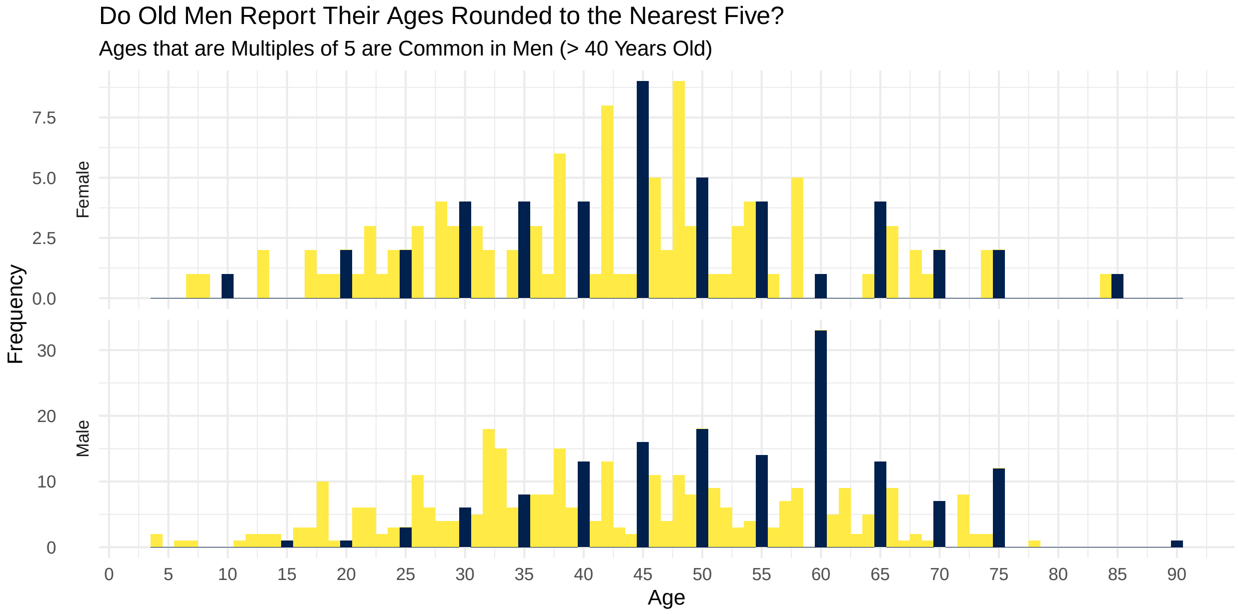


For this plot:

- Most points considered in the first plot apply here, including the choice of a bar chart, adding error bars, a title, a subtitle, and a caption.
- The color scale now encodes gender, and a legend was added accordingly.

4.3 Third Plot

The final plot shows the distribution of ages split across gender, same as the one explored earlier, but refined visually.



For this plot:

- A title and a subtitle were added.
- The same color scale is used for consistency. But in this plot, it has an important contribution to the plot's effectiveness: the blue color is much stronger than yellow, and this makes the blue bars look more prominent at the first glance because those are were the emphasis should lie.
- Breaks of 5 were used in the exploratory phase and kept here as they are.
- Although the main question of the plot involves males only, the female part of the plot was kept to show that the pattern is less pronounced there.
- There is one tiny detail: switching the Male/Female labels from the right to the left. I found it being on the right makes it slightly less visible "in the spotlight." Usually it is the left (axis) position that draws the attention first, and the frequency axis is not critically important in this particular case, so putting the gender labels on the left is a better choice by that logic.

---

5 Reflection

EDA has an iterative, brainstorming-like quality to it, and it does not always have strict directions. One question can be posed based on ideas already formed before analysis, but another can be posed after noticing a pattern or an outlier. This means that the project by no means provides an exhaustive list of available approaches in EDA for a data set, not even for *this* data set, but it is merely an example of few thought processes to follow in order to draw information from data.

Most of the conclusions that we drew from analyzing this data set confirm (or are confirmed by) available research and other resources. The age pattern, on the other hand, is an example of a peculiar finding that comes from data analysis but could be just a myth where one draws a line around a random set of points and call it a shape. This is why finding and collecting more data will be an improvement. First, such patterns need to be confirmed in multiple sets for the same (Indian) or different populations. Also, more data will allow us to zoom in on groups and facet by multiple factors.

Finally, since it has the word "exploratory" in its name, EDA is merely a tool that introduces the analyst to the nature of the data. There are stronger statistical procedures like inference and modeling that can help us make better use and draw more solid conclusions from data.