Final_Project

Sangwon Ji

2023-04-08

Visualizing the Data

Importing the data

Changing data into factors

```
cholangitis <- read.csv(file = "cholangitis.csv", header = TRUE, sep = ",")

cholangitis$status <- as.factor(cholangitis$status)
cholangitis$drug <- as.factor(cholangitis$drug)
cholangitis$sex <- as.factor(cholangitis$sex)
cholangitis$ascites <- as.factor(cholangitis$ascites)
cholangitis$hepatomegaly <- as.factor(cholangitis$hepatomegaly)
cholangitis$spiders <- as.factor(cholangitis$spiders)
cholangitis$edema <- as.factor(cholangitis$edema)
cholangitis$stage <- as.factor(cholangitis$stage)

head(cholangitis)
```

```
##   id n_days status            drug   age sex ascites hepatomegaly spiders edema
## 1  1    400      D D-penicillamine 21464   F       Y            Y       Y     Y
## 2  2   4500      C D-penicillamine 20617   F       N            Y       Y     N
## 3  3   1012      D D-penicillamine 25594   M       N            N       N     S
## 4  4   1925      D D-penicillamine 19994   F       N            Y       Y     S
## 5  5   1504     CL         Placebo 13918   F       N            Y       Y     N
## 6  6   2503      D         Placebo 24201   F       N            Y       N     N
##   bilirubin cholesterol albumin copper alk_phos   sgot tryglicerides platelets
## 1      14.5         261    2.60    156   1718.0 137.95           172       190
## 2       1.1         302    4.14     54   7394.8 113.52            88       221
## 3       1.4         176    3.48    210    516.0  96.10            55       151
## 4       1.8         244    2.54     64   6121.8  60.63            92       183
## 5       3.4         279    3.53    143    671.0 113.15            72       136
## 6       0.8         248    3.98     50    944.0  93.00            63       361
##   prothrombin stage
## 1        12.2     4
## 2        10.6     3
## 3        12.0     4
## 4        10.3     4
## 5        10.9     3
## 6        11.0     3
```

```
cholangitis <- filter(cholangitis, drug != "NA")
cholangitis <- na.omit(cholangitis)
```

```
summary(cholangitis)
```

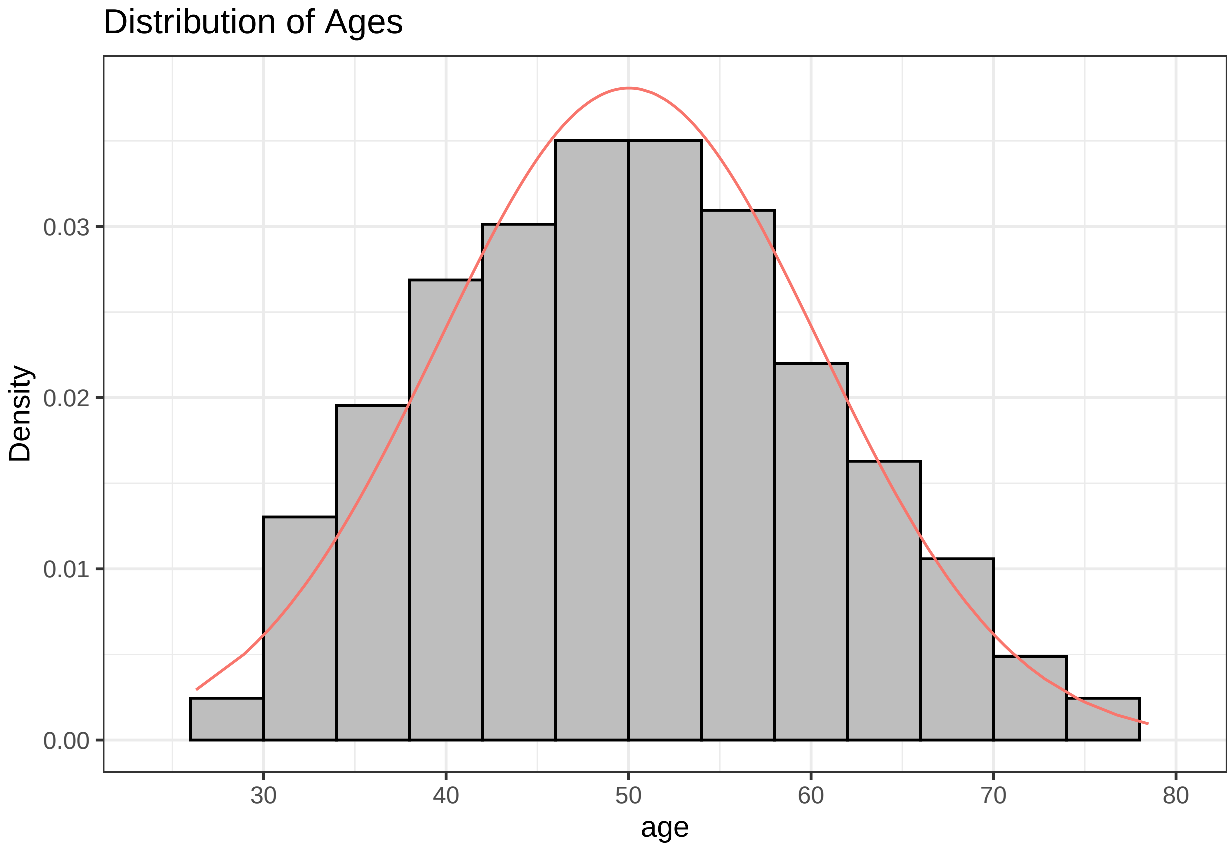
```
##        id            n_days     status                drug          age       
##   Min.   :  1.0   Min.   :  41   C :165   D-penicillamine:154   Min.   : 9598  
##   1st Qu.: 78.5   1st Qu.:1180   CL: 19   Placebo        :153   1st Qu.:15494  
##   Median :157.0   Median :1831   D :123                         Median :18176  
##   Mean   :156.9   Mean   :1999                                  Mean   :18257  
##   3rd Qu.:235.5   3rd Qu.:2702                                  3rd Qu.:20696  
##   Max.   :312.0   Max.   :4556                                  Max.   :28650  
##   sex     ascites hepatomegaly spiders edema     bilirubin       cholesterol    
##   F:271   N:284   N:149        N:218   N:259   Min.   : 0.300   Min.   : 120.0  
##   M: 36   Y: 23   Y:158        Y: 89   S: 28   1st Qu.: 0.800   1st Qu.: 248.0  
##                                        Y: 20   Median : 1.300   Median : 309.0  
##                                                Mean   : 3.267   Mean   : 367.4  
##                                                3rd Qu.: 3.450   3rd Qu.: 399.5  
##                                                Max.   :28.000   Max.   :1775.0  
##     albumin          copper           alk_phos           sgot       
##   Min.   :1.960   Min.   :  4.00   Min.   :  289   Min.   : 26.35  
##   1st Qu.:3.310   1st Qu.: 41.00   1st Qu.:  867   1st Qu.: 80.60  
##   Median :3.550   Median : 73.00   Median : 1260   Median :114.70  
##   Mean   :3.515   Mean   : 98.06   Mean   : 1995   Mean   :122.48  
##   3rd Qu.:3.790   3rd Qu.:123.50   3rd Qu.: 2002   3rd Qu.:151.90  
##   Max.   :4.640   Max.   :588.00   Max.   :13862   Max.   :457.25  
##   tryglicerides     platelets     prothrombin    stage  
##   Min.   : 33.0   Min.   : 62.0   Min.   : 9.00   1: 16  
##   1st Qu.: 85.0   1st Qu.:200.0   1st Qu.:10.00   2: 65  
##   Median :110.0   Median :258.0   Median :10.60   3:119  
##   Mean   :124.4   Mean   :262.3   Mean   :10.73   4:107  
##   3rd Qu.:151.5   3rd Qu.:323.0   3rd Qu.:11.10          
##   Max.   :598.0   Max.   :563.0   Max.   :17.10          
```

Looking through the dataset, for treatments, some patients have recieved none treatments, and when analyzing, we will drop those values of NA within drug.

Basic exploratory data analysis

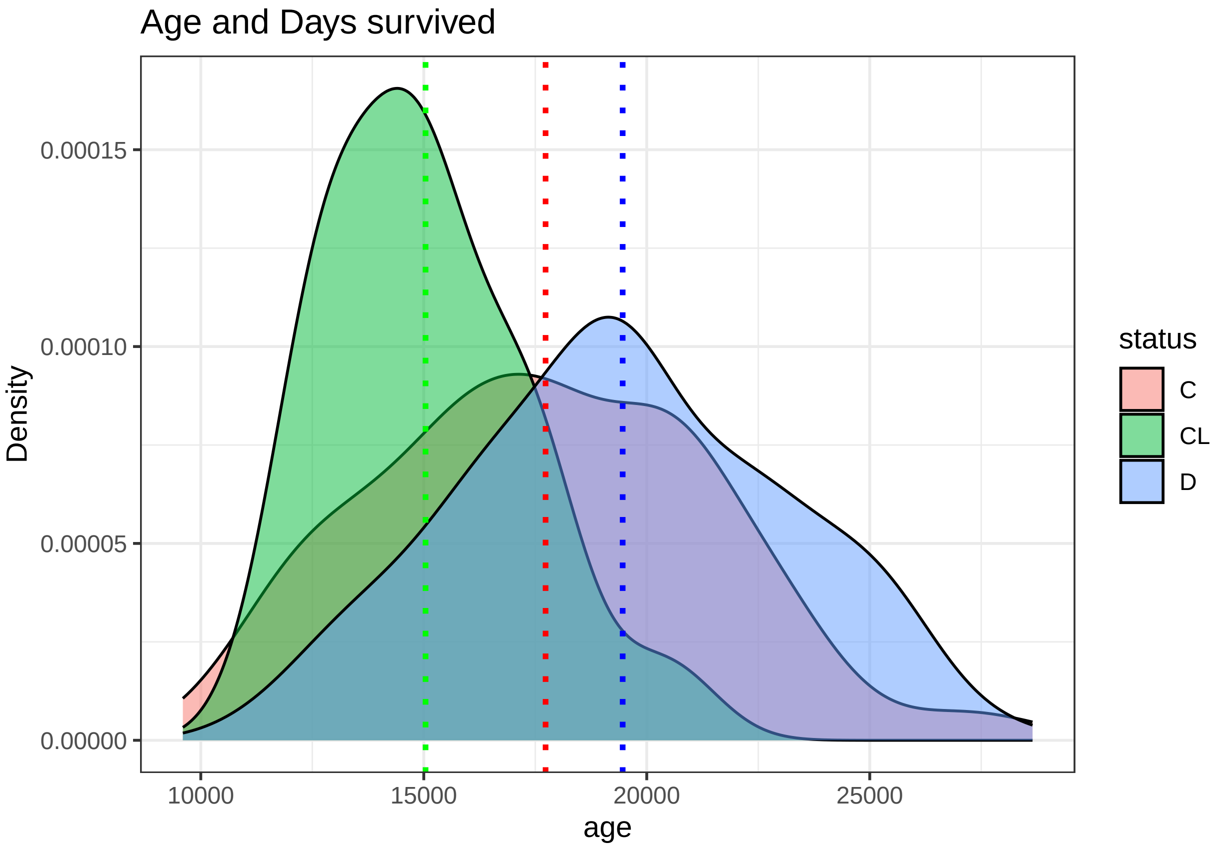
There are many variables in the dataset. To see the variables and it's correlation,

```
#Codes for Age
age <-cholangitis$age/365
Ages <- data.frame(age)
mean_age <- mean(age)
sd_age <- sd(age)
x.dens <- dnorm(age, mean = mean_age, sd = sd_age)
#Age histogram
ggplot(Ages, aes(age, y = ..density..)) +
geom_histogram(bins = 10, fill = 'grey', color = 'black', binwidth = 4) +
  scale_x_continuous(breaks = c(20,30,40,50,60,70,80), limits = c(24, 80)) +
geom_line(aes(x = age, y = x.dens, color = 'red'),data = Ages) +
labs(y = 'Density', title = 'Distribution of Ages') +
theme_bw() +
theme(legend.position = "none")
```



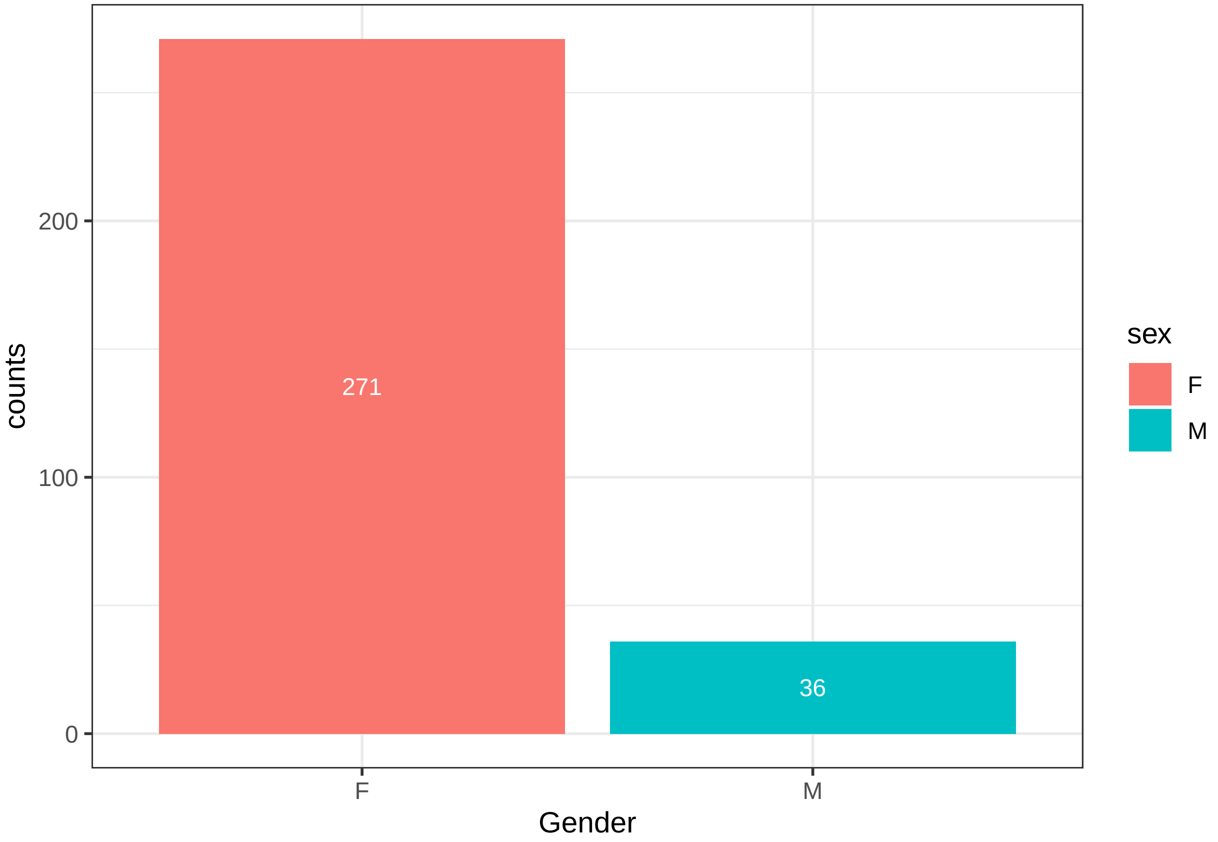
First looking at the age distribution of the whole group seems normally distributed ranged from 24 to 80, with age 50 being the most porportion of the whole group. Since it's a mediacal test, we can see that the general age is pretty high. Since the age is written in days, we would divide it by 365 days, and turn them into a year to make it look better visually.

```
ggplot(cholangitis, aes(x = age, fill = status)) + geom_density(alpha = 0.5) +
  labs(x= "age", y = "Density") + ggtitle("Age and Days survived") +
  geom_vline(xintercept = mean(filter(cholangitis, status == 'C')$age), color = 'red', linetype="dotted
  geom_vline(xintercept = mean(filter(cholangitis, status == 'CL')$age), color = 'green', linetype="dot
  geom_vline(xintercept = mean(filter(cholangitis, status == 'D')$age), color = 'blue', linetype="dotte
  theme_bw()
```



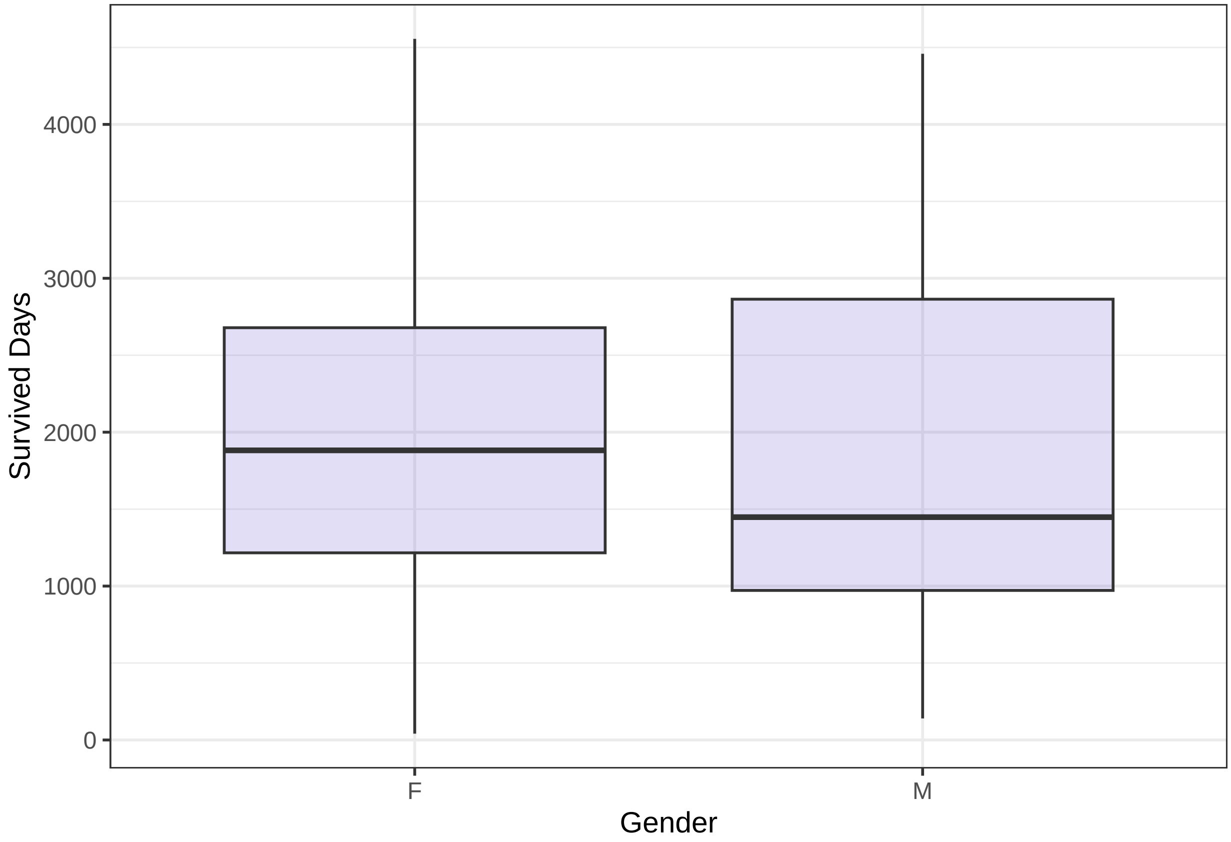
Then I wanted to explore if there was any correlation among age and the days survived. For those who died, is showing somewhat normal distribution, and those who are not dead and received liver transplant was right-skewed showing having the average age of lower. The groups died had the highest aveage among the groups regarding status.

```
sex_dist <- df3 %>%
  ggplot(aes(x = sex, y = n, fill = sex, label = n)) +
  geom_col() +
  theme_bw() +
  labs(x = "Gender", y = "counts") +
   geom_text(position = position_stack(vjust = 0.5), size = 3, color = "#ffffff")
sex_dist
```



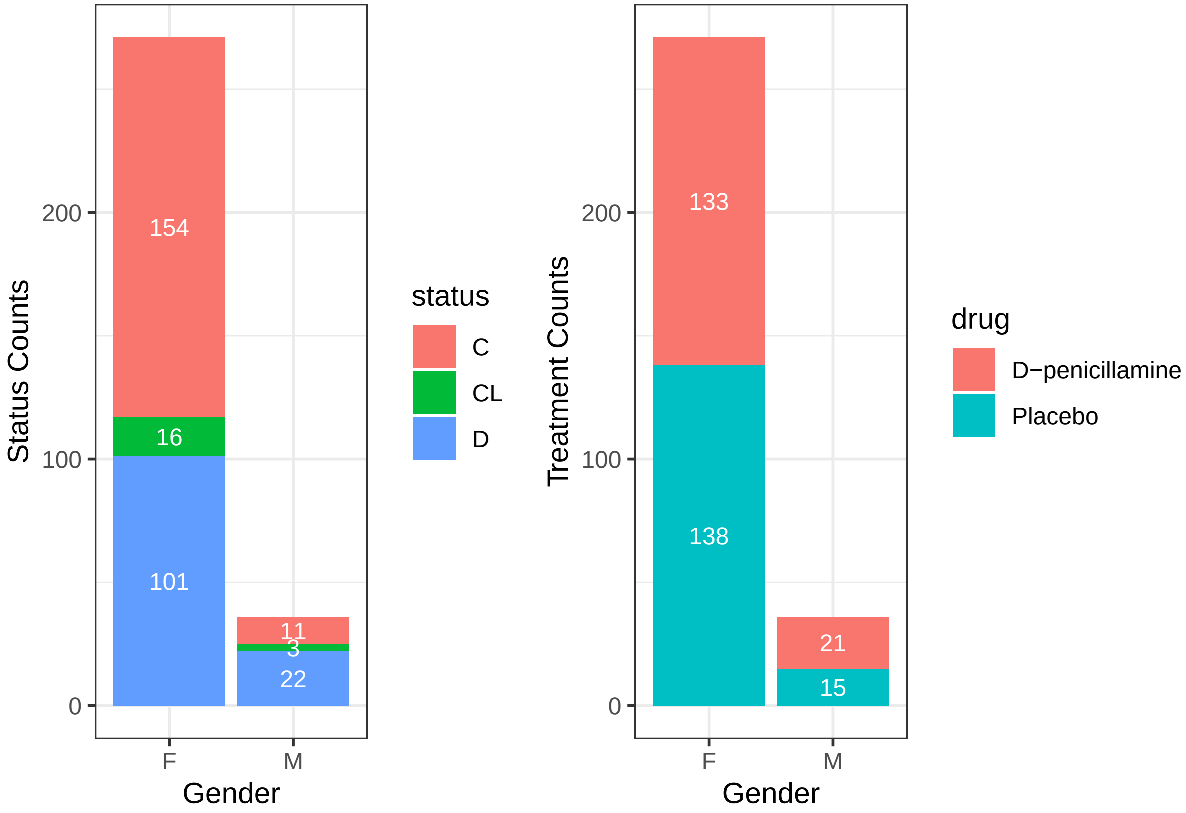
The gender count seems odd. There are females more than male by about nine times.

```r
ggplot(cholangitis, aes(x = sex, y = n_days)) +
  geom_boxplot(fill="slateblue", alpha=0.2) +
  labs(x = "Gender", y = "Survived Days") +
  theme_bw()
```



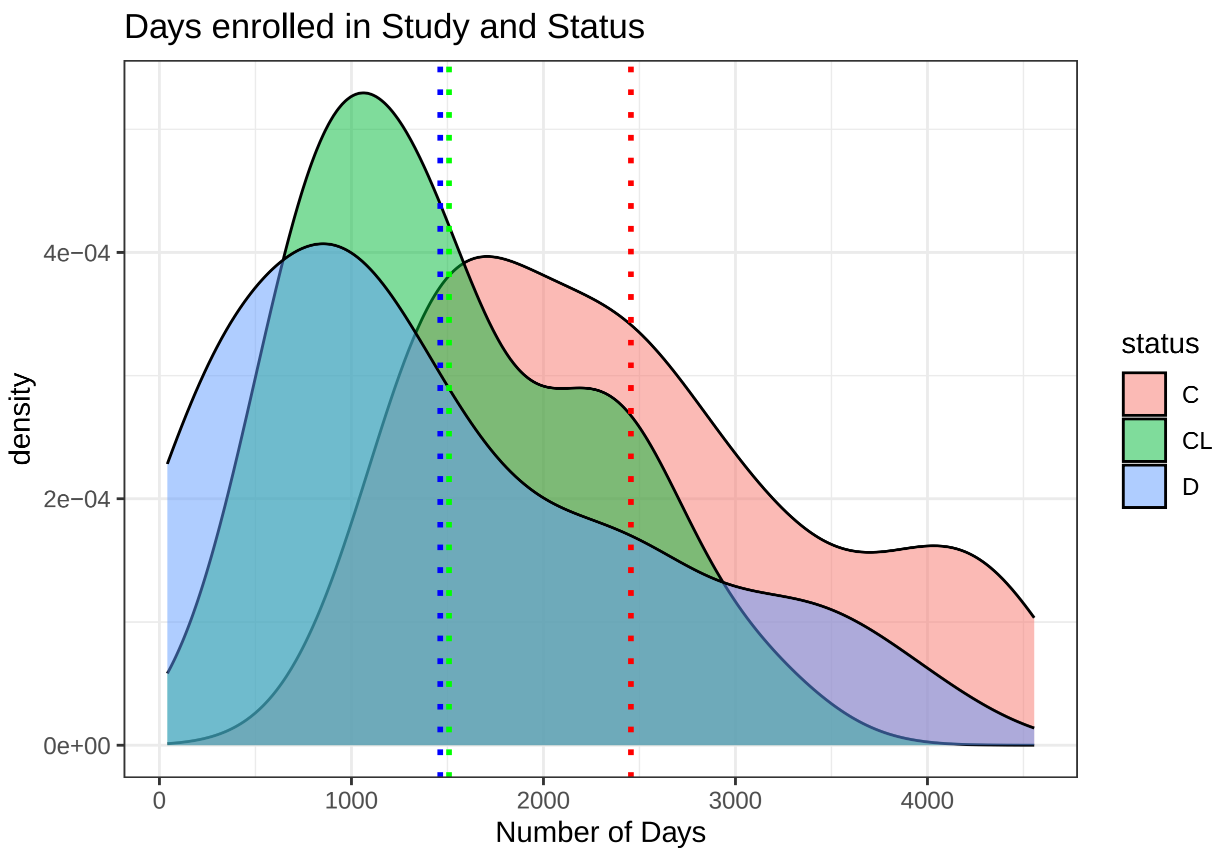
Adding on from gender, I was wondering if there was a difference among gender for the days of survival, therefore I plotted through a boxplot to see if there was a vast difference. The mean seemed to be lower for male, but they seemed not much of a difference in a whole.

```
gen_stat <- df5 %>%
  ggplot(aes(x = sex , y = n, fill = status, label = n)) +
  geom_col() +
  theme_bw() +
   labs(x= "Gender", y = "Status Counts") +
geom_text(position = position_stack(vjust = 0.5), size = 3, color = "#ffffff")

gen_drug <- df6 %>%
  ggplot(aes(x = sex, y = n, fill = drug, label = n)) +
  geom_col() +
  theme_bw() +
  labs(x= "Gender", y = "Treatment Counts") +
  geom_text(position = position_stack(vjust = 0.5), size = 3, color = "#ffffff")

gen_stat + gen_drug
```



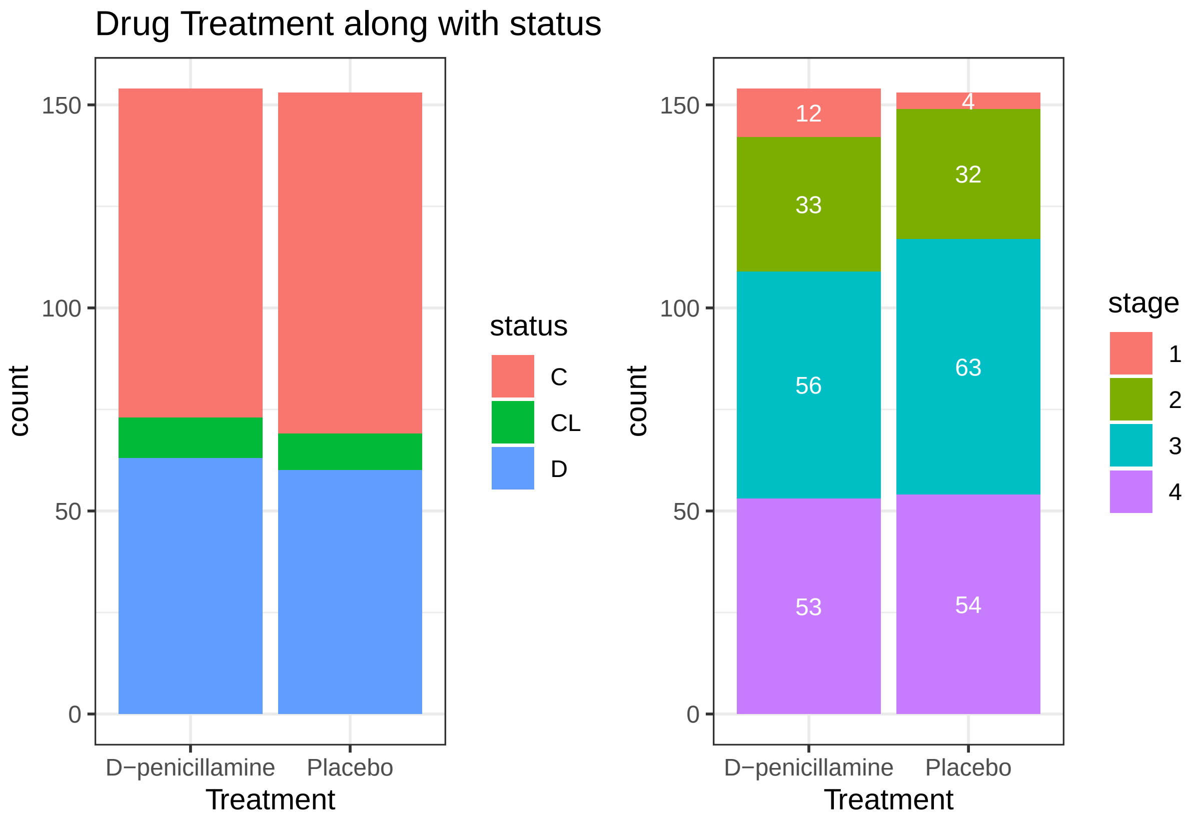
Even though there was a difference in the absolute number among gender, the proportion of status and treatment given was distributed in similar proportion to each other, which kind of relieved the thought that it might be biased somehow.

```
ggplot(cholangitis, aes(x = n_days, fill = status)) +
  geom_density(alpha = 0.5) + xlab("Number of Days") +
  ggtitle("Days enrolled in Study and Status") +
  geom_vline(xintercept = mean(filter(cholangitis, status == 'C')$n_days), color = 'red', linetype="dot
  geom_vline(xintercept = mean(filter(cholangitis, status == 'CL')$n_days), color = 'green', linetype="
  geom_vline(xintercept = mean(filter(cholangitis, status == 'D')$n_days), color = 'blue', linetype="do
  theme_bw()
```



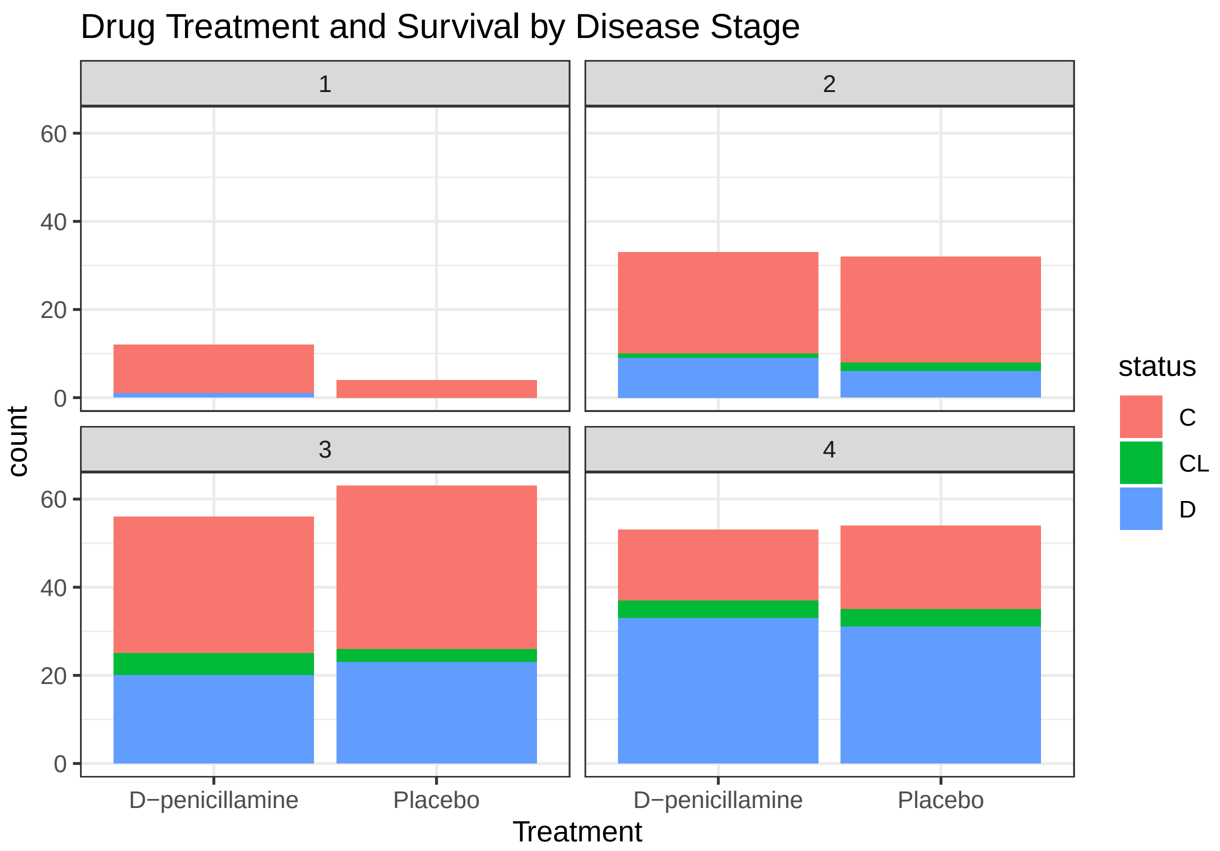
Moving on from gender, now looking at the number of days enrolled in the study and the status of them, the groups that diead were skewed the most, and those who didn't died with the highest average, but what was interesting is that it's kind of right skewed as well as other plots.

```
drug_status <- ggplot(cholangitis, aes(x = drug, fill = status)) +
  geom_bar() + xlab("Treatment") + ggtitle("Drug Treatment along with status and stage") + theme_bw()

drug_stage <- df %>%
  ggplot(aes(x = drug, y = n, fill = stage, label = n)) +
  geom_col() +
  theme_bw() +
  labs(x = "Treatment", y= "count") +
geom_text(position = position_stack(vjust = 0.5), size = 3, color = "#ffffff")

drug_status + drug_stage
```



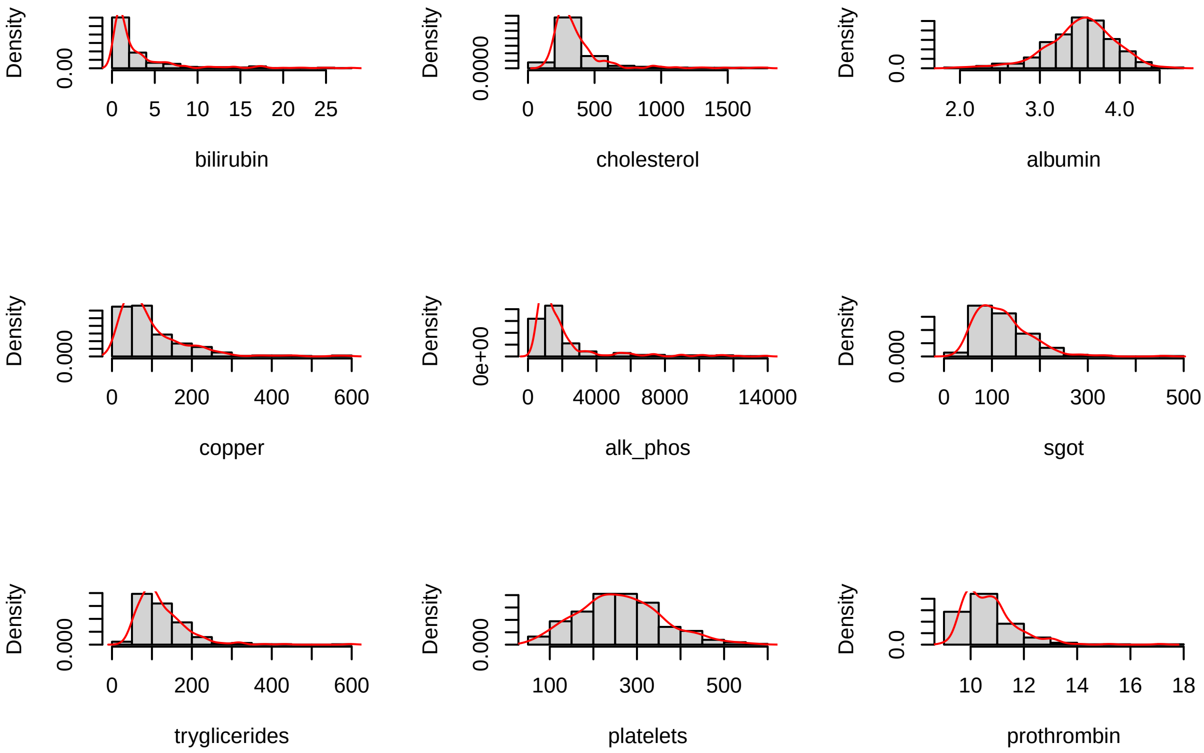
Now comparing the treatment given with the current status and the stage, for each of the treatments given, the status and stage seems to have similar proportions to each other, showing no big of a difference which I wondered if there would be a bias.

```
ggplot(cholangitis, aes(x = drug, fill = status)) + geom_bar() + xlab("Treatment") + ggtitle("Drug Trea
  theme_bw()
```



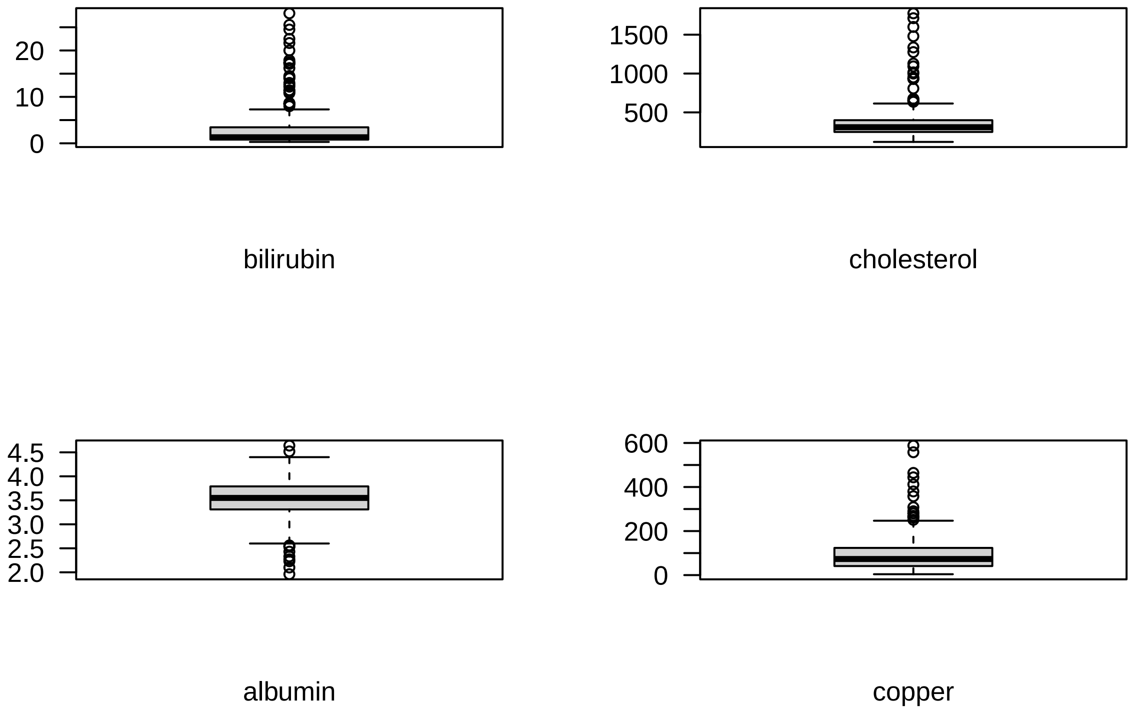
Now comparing the treatment with the treatments that were assigned on each stage and their survival, it shows that there are more and more deaths occured with the stage, but since the proportion of treatments assigned seems mostly identical to each other. What's interesting is however, that the stage one had the most living, regardless of the treatment given.

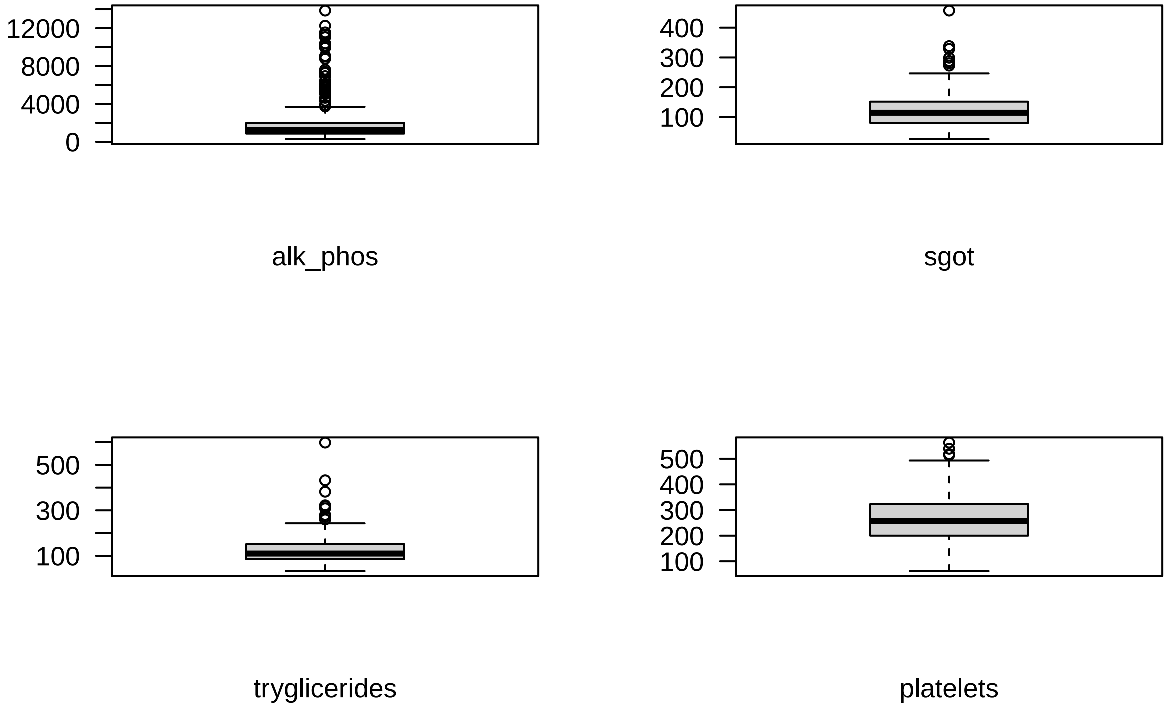
```
# subsetting
subset <- cholangitis[, names(which(sapply(cholangitis, is.numeric)))]
subset <- subset(subset, select = -c(1:3))
#code for histograms
par(mfrow =c(3,3))
for (i in names(subset)) {
  hist(subset[, i], freq = FALSE, xlab = i, main ="")
  lines(density(subset[, i]), col ="red")
}
```

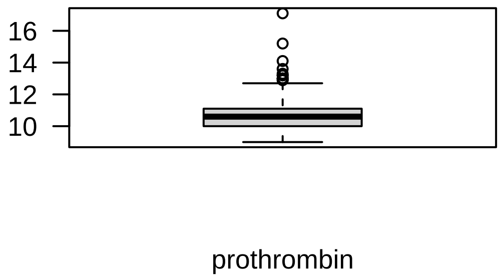


I plotted the histograms to find out the genral distribution of the genral health information, and surprisingly I was able to find out that most of the information were right-skewed, and had outliers, except albumin and platelets. There were values far off from the where data is distributed.

```r
par(mfrow = c(2,2))
for (i in names(subset)) {
  boxplot(subset[,i], xlab = i, las =2)
}
```


0
10
20
bilirubin
500
1000
1500
cholesterol
2.0
2.5
3.0
3.5
4.0
4.5
albumin
0
200
400
600
copper


12000
8000
4000
0
alk_phos
400
300
200
100
sgot
500
300
100
tryglicerides
500
400
300
200
100
platelets

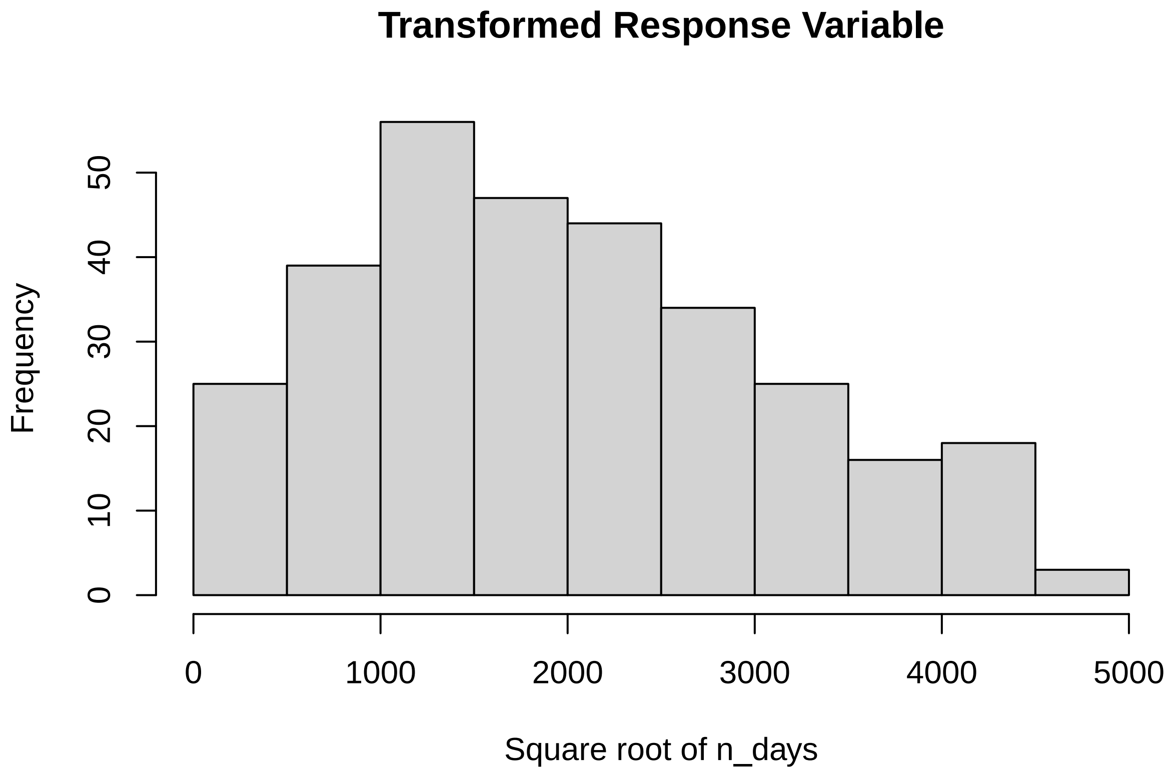


> As boxplots is an effective methods for showing the unusual or extreme points, I used them to see the overall general health information, and found out that there seems to be plenty of outliers that have to be discarded for a better dataset.

Multivariate Regression

Multivariate regression analysis

```
hist((cholangitis$n_days), xlab = "Square root of n_days",
     main = "Transformed Response Variable")
```



In the EDA part, we observed that the distribution of our response variable (n_days) is left-skewed. Therefore, I'm going to transform this data using square root before I perform multivariate regression analysis. As a result, the distribution is more likely normal distribution.

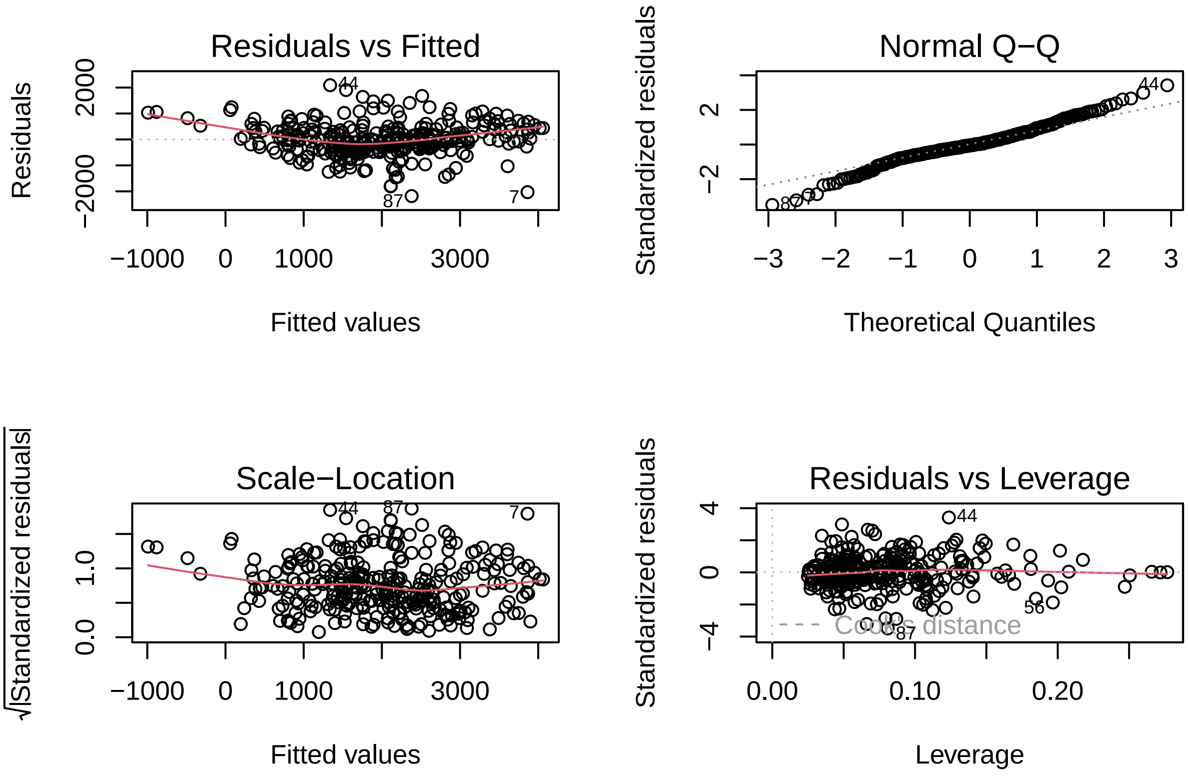
First, I'm going to use all predictor variables.

```
full_model <- lm(n_days ~ ., data = cholangitis)
summary(full_model)
```

```
##
## Call:
## lm(formula = n_days ~ ., data = cholangitis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2182.88  -311.10   -35.55   343.92  2089.75
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.510e+03  7.844e+02   4.475 1.11e-05 ***
## id              -7.649e+00  5.263e-01 -14.535  < 2e-16 ***
## statusCL        -5.823e+02  1.678e+02  -3.469 0.000603 ***
## statusD         -1.144e+03  1.083e+02 -10.559  < 2e-16 ***
## drugPlacebo     -6.939e+01  7.803e+01  -0.889 0.374580
## age             -8.051e-03  1.158e-02  -0.695 0.487604
## sexM             1.316e+02  1.310e+02   1.004 0.316026
## ascitesY        -6.190e+01  2.001e+02  -0.309 0.757231
## hepatomegalyY -4.288e-01  9.133e+01  -0.005 0.996257
## spidersY      -1.635e+02  9.600e+01  -1.703 0.089593 .
## edemaS        -2.304e+02  1.412e+02  -1.632 0.103810
## edemaY        -7.074e+02  2.157e+02  -3.280 0.001169 **
## bilirubin     -3.554e+01  1.279e+01  -2.779 0.005812 **
## cholesterol   -5.561e-02  2.010e-01  -0.277 0.782258
## albumin        1.201e+02  1.143e+02   1.050 0.294490
## copper        -1.531e+00  5.494e-01  -2.787 0.005680 **
## alk_phos       4.170e-02  1.974e-02   2.112 0.035524 *
## sgot           1.119e+00  7.984e-01   1.401 0.162304
## tryglicerides  7.626e-01  6.805e-01   1.121 0.263407
## platelets      9.003e-02  4.442e-01   0.203 0.839517
## prothrombin    1.829e+01  4.735e+01   0.386 0.699533
## stage2        -6.910e+01  1.910e+02  -0.362 0.717779
## stage3        -1.788e+02  1.870e+02  -0.956 0.339911
## stage4        -3.380e+02  2.013e+02  -1.679 0.094267 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 652.8 on 283 degrees of freedom
## Multiple R-squared:  0.6904, Adjusted R-squared:  0.6652
## F-statistic: 27.44 on 23 and 283 DF,  p-value: < 2.2e-16
```

Diagnostic Plots

```
par(mfrow=c(2,2))
plot(full_model)
```



```r
# Part 1
studentized_residuals <- rstudent(full_model)
outliers_sr <- which(abs(studentized_residuals) > 3)

# Part 2
cooks_d <- cooks.distance(full_model)
threshold_cooks <- 4 / length(cooks_d)
outliers_cooks <- which(cooks_d > threshold_cooks)

# Part 3
leverage <- hatvalues(full_model)
threshold_leverage <- 2 * (length(coef(full_model)) - 1) / nrow(data)
outliers_leverage <- which(leverage > threshold_leverage)

outliers_all <- unique(c(outliers_sr, outliers_cooks, outliers_leverage))
print(outliers_all)
```

```
##  [1]   7  43  86 110   3   5  12  22  27  30  37  54  55  56  61  65  80  81  82
## [20]  90  96 112 200 248 263 276 288
```

Part 1 (Studentized Residuals): A common threshold for identifying outliers using studentized residuals is an absolute value greater than 2 or 3. Therefore, I found some rows whose studentized residuals are greater than 3.

Part 2 (Cooks Distance): A common threshold for identifying outliers using Cook's distance is 4/n, where n is the number of observations.

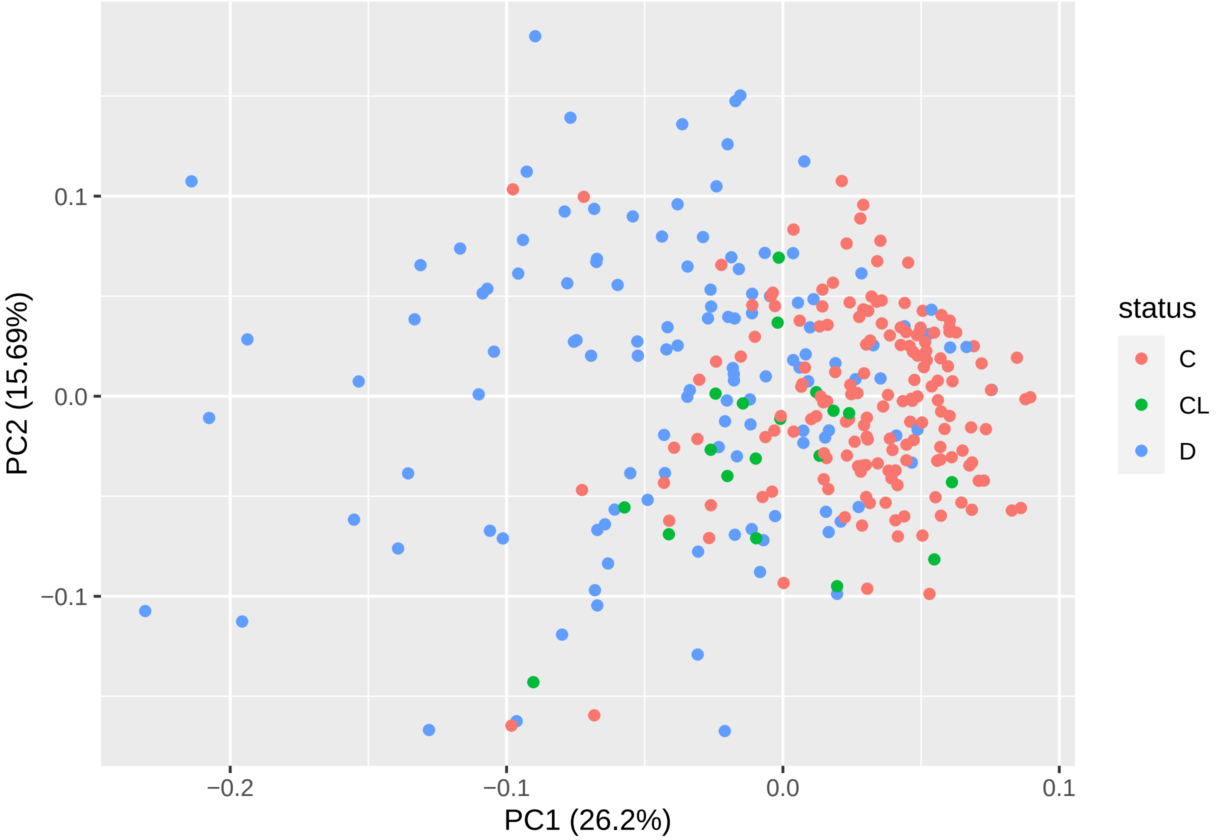
Part 3 (Leverage): A common threshold for identifying outliers using leverage is 2 * (p+1) / n, where p is the number of predictor variables and n is the number of observations.

```
data_final <- cholangitis[-outliers_all, ]
```

PCA

Before performing PCA, we need to standardize our predictor variable to ensure that all variables can be equally treated. Therefore, I'm going to create a new data frame containing only numeric columns.

```
cholangitis_explanatory <- cholangitis[,c(-1,-3,-4,-6:-10,-20)]

cholangitisSCALED = scale(cholangitis_explanatory)

cholangitis.pca = prcomp(cholangitisSCALED, scale = F)

autoplot(cholangitis.pca, data= cholangitis, colour = "status")
```



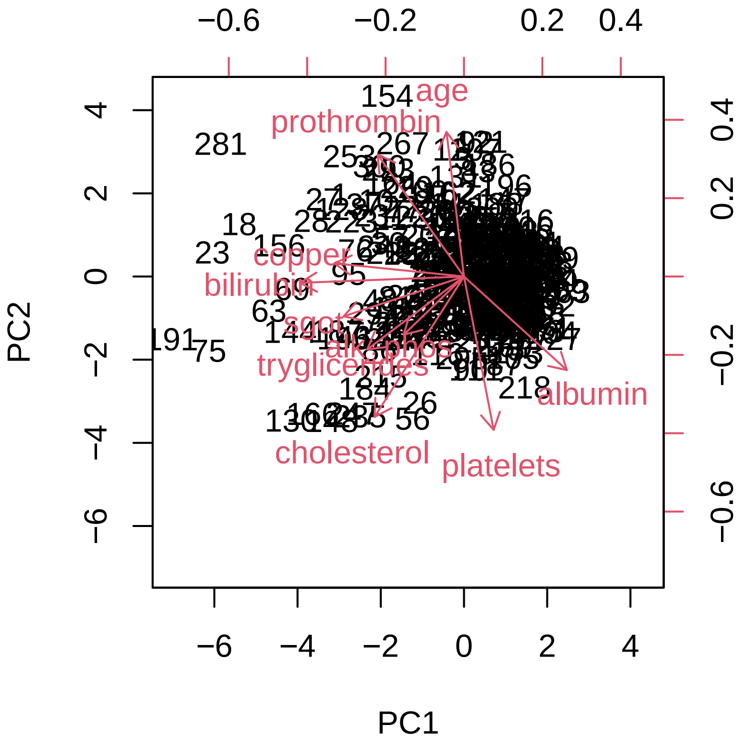
```
colss <- c("age", "bilirubin", "cholesterol", "albumin", "copper", "alk_phos",
           "sgot", "tryglicerides", "platelets", "prothrombin")
data_std <- scale(cholangitis[, colss])
pca_result <- prcomp(data_std, center = FALSE, scale = FALSE)
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.6071 1.2910 1.0437 0.97029 0.89485 0.87610 0.83146
## Proportion of Variance 0.2583 0.1667 0.1089 0.09415 0.08008 0.07676 0.06913
## Cumulative Proportion  0.2583 0.4250 0.5339 0.62803 0.70810 0.78486 0.85399
##                            PC8     PC9    PC10
## Standard deviation     0.78064 0.70954 0.58927
## Proportion of Variance 0.06094 0.05034 0.03472
## Cumulative Proportion  0.91493 0.96528 1.00000
```

```
pca_loadings <- pca_result$rotation
print(pca_loadings)
```

```
##                       PC1         PC2          PC3         PC4         PC5
## age           -0.05572478  0.46116899 -0.457674900  0.32366552 -0.33093007
## bilirubin     -0.52251116 -0.02010152  0.053357753  0.17328938  0.13262223
## cholesterol   -0.28514649 -0.44805760  0.133940177  0.21262980 -0.28941853
## albumin        0.32795986 -0.29844875  0.005869899  0.05834725  0.65401307
## copper        -0.41351301  0.04212311 -0.119966147 -0.09416073  0.24940947
## alk_phos      -0.19199035 -0.18540993 -0.459904159 -0.71649834 -0.07531862
## sgot          -0.38371833 -0.12710646  0.492412936 -0.16925305 -0.14715087
## tryglicerides -0.30854767 -0.23255881 -0.352623170  0.47367018  0.27916700
## platelets      0.09469014 -0.48925327 -0.414309658 -0.01153470 -0.17411505
## prothrombin   -0.27522665  0.38901441 -0.073091758 -0.20301818  0.40449824
##                        PC6          PC7         PC8         PC9        PC10
## age            0.149394403  0.052027996 -0.57612604 -0.05754845  0.01479478
## bilirubin      0.161188368  0.024800169  0.08533964 -0.07588449  0.79779387
## cholesterol    0.360394199 -0.214563790 -0.14066892  0.57866895 -0.20113249
## albumin        0.172427886  0.003858128 -0.57310383  0.02939398  0.11463180
## copper        -0.353005314  0.649875636 -0.10344735  0.38778672 -0.18368028
## alk_phos      -0.127915982 -0.356952027 -0.18984058  0.04008303  0.12955660
## sgot          -0.009489488  0.082743134 -0.41163758 -0.55836863 -0.23595809
## tryglicerides -0.331607219 -0.359179525  0.13175265 -0.28258407 -0.29645941
## platelets      0.378585282  0.505600694  0.19792250 -0.33185589 -0.03609574
## prothrombin    0.630657438 -0.096871171  0.19960414 -0.03827740 -0.33811339
```

```
biplot(pca_result, scale = 0)
```



```
# 1:5 Specifies the Number of top variables to select
selected_vars <- names(sort(abs(pca_loadings[, 1]), decreasing = TRUE))[1:5]
print(selected_vars)
```

```
## [1] "bilirubin"     "copper"        "sgot"          "albumin"
## [5] "tryglicerides"
```

Final Model

```
model_final <- lm(sqrt(n_days) ~ status + drug + sex + ascites +
                    hepatomegaly + spiders + stage + edema +
                    bilirubin + copper + sgot + albumin + tryglicerides
                  , data = data_final)
summary(model_final)
```

```
##
## Call:
## lm(formula = sqrt(n_days) ~ status + drug + sex + ascites + hepatomegaly +
##     spiders + stage + edema + bilirubin + copper + sgot + albumin +
##     tryglicerides, data = data_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.4950  -6.8731  -0.1978   6.5232  24.2727
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)   28.872063   7.049925   4.095 5.62e-05 ***
## statusCL      -7.490279   2.512761  -2.981 0.003144 ** 
## statusD       -4.745756   1.552299  -3.057 0.002465 ** 
## drugPlacebo    0.286140   1.195302   0.239 0.810993    
## sexM           0.147658   2.055389   0.072 0.942784    
## ascitesY      -4.192221   3.983190  -1.052 0.293550    
## hepatomegalyY -0.404462   1.448536  -0.279 0.780295    
## spidersY      -0.898949   1.494586  -0.601 0.548047    
## stage2        -3.438685   2.799874  -1.228 0.220490    
## stage3        -4.891253   2.772145  -1.764 0.078825 .  
## stage4        -6.907058   3.008595  -2.296 0.022479 *  
## edemaS        -2.478835   2.341235  -1.059 0.290679    
## edemaY        -7.287995   4.010210  -1.817 0.070304 .  
## bilirubin     -0.640598   0.190501  -3.363 0.000887 ***
## copper        -0.019924   0.008859  -2.249 0.025342 *  
## sgot           0.011151   0.012310   0.906 0.365862    
## albumin        6.548347   1.707802   3.834 0.000158 ***
## tryglicerides  0.014729   0.010926   1.348 0.178807    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.733 on 262 degrees of freedom
## Multiple R-squared:  0.4885, Adjusted R-squared:  0.4553 
## F-statistic: 14.72 on 17 and 262 DF,  p-value: < 2.2e-16
```

```
mean(full_model$residuals^2)
```

```
## [1] 392825
```

```
mean(model_final$residuals^2)
```

```
## [1] 88.63567
```