Final Project

● Graded

Student

Sangwon Ji

Total Points

52 / 60 pts

Question 1

Question 1

5 / 5 pts

✔ **– 0 pts** Correct

– 1 pt Verifying: Code shows did some factors, but only print out to verify others.

– 2 pts Skipped stage for factor

– 3.5 pts Didn't address factors at all

– 2 pts Said changed to factors, but no code

– 3.5 pts Removed factors with more than 2 levels from entire analysis.

– 0.5 pts Didn't select pages on gradescope

Question 2

Question 2

14 / 15 pts

- ✔ **+ 1 pt** General numerical explorations (1)

Univariate Distribution EDA (3)

- ✔ **+ 3 pts** Comprehensive
- **+ 2 pts** Covers ground adequetely
- **+ 1 pt** Limited or Missing obvious possibilities: e.g. Doesn't compare between any different groups
- **+ 0 pts** Missing

Multivariate (4)

- **+ 0 pts** Missing
- **+ 1 pt** Minimal (e.g. Pairs/scatterplots only)
- **+ 2 pts** More than minimal, but no PCA (or no scatter)
- ✔ **+ 3 pts** Covers ground adequetely
- **+ 4 pts** Extensive

Code/Plots (3)

- ✔ **+ 3 pts** Accurate Code, plots with title, etc.
- **+ 2.5 pts** Small problem: Plotting with colored points is wrong, mostly plot titles, but some pretty off
- **+ 2 pts** Basic Problem: Poor plot labeling /
- **+ 2 pts** Basic Problem: Poor bin choice
- **+ 1 pt** Serious problem: Wrong plot from code.

General Interpretation / Comments (4)

- ✔ **+ 4 pts** Excellent commentary.
- **+ 3 pts** Good commentary
- **+ 2 pts** Low detail / Average commentary;
- **+ 1 pt** Minimal Commentary
- **+ 0 pts** Missing Commentary

Question 3

Question 3

8 / 10 pts

Evaluation outlier / transformation (5)

+ 5 pts Comprehensive

✔ + 4 pts Reasonable

+ 3 pts Missing key component (e.g. regression diagnostic plot) but otherwise okay.

+ 2 pts Limited evaluation (e.g. no regression diagnostic and not very good eval)

+ 1 pt Poor evaluation -- e.g. doesn't consider transform,

+ 0 pts Missing

Code / Plot evaluations (2)

✔ + 2 pts Good

+ 1 pt Problems in code/plots

+ 0 pts Missing

Commentary (3)

+ 3 pts Strong explanations and justifications

✔ + 2 pts Average

+ 1 pt Poor/limited justifications; problematic choices.

+ 0 pts Missing

Question 4

Question 4

8 / 10 pts

Correct implementation (6)

+ 6 pts Extensive

✔ + 5 pts Reasonable

+ 4 pts Relied on adjusted R2

+ 3 pts Manually tried step (with significance)

+ 0 pts Missing

Commentary (4)

+ 0 pts Click here to replace this description.

+ 4 pts Excellent

✔ + 3 pts Average

+ 2 pts Little

+ 0 pts Missing

Question 5

Question 5

9 / 10 pts

Address the main assumptions (7)

✔ + 7 pts All addressed appropriately

+ 6 pts Main assumptions, some mild confusion/lack of clarity

+ 5 pts Missing a main assumption entirely, but otherwise fine; or mix of clarity and not wrt to what supposed to fine.

+ 4 pts Addresses main assumptions, but doesn't explain why conclusions from plots (or serious confusion of what should find)

+ 2 pts Do not make explicit what should be assumptions checking and what should be finding

+ 1 pt Just plots

+ 0 pts Missing

General Conclusions -- general summary across (3)

+ 3 pts Excellent

✔ + 2 pts Average

+ 1 pt Poor

+ 0 pts Missing

Question 6

Question 6

8 / 10 pts

+ 10 pts Excellent discussion

✔ + 8 pts Reasonable discussion

+ 6 pts Only use variable selection to decide, limited justification

+ 4 pts Only p-value, no further; or confused

+ 2 pts No justification

+ 2 pts Thought wqs significnt

+ 0 pts Missing

Question assigned to the following page: 1

Final_Project

Sangwon Ji

2023-04-08

```
knitr::opts_chunk$set(echo = TRUE, tidy.opts=list(width.cutoff=60), tidy=TRUE)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggfortify)
library(patchwork)
```

Visualizing the Data

Importing the data

Changing data into factors

First step was to changing the data into factor, and take a look of the data if there is any values that has to be changed or removed to make the study better.

```
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
```

Question assigned to the following page: 1

```
## 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
```

Looking through the head function to display the first few rows of the dataset, where we can see that they are changed into factors !

```
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. Also, even though it's lot of observations that we are losing through dropping these variables however, it will occur a problem

Questions assigned to the following page: 1 and 2

in histogram where it can't read missing values, therefore I'm dropping the values prior in the EDA.

Basic exploratory data analysis

Following code is the codes of the dataframe that I have made prior to making a plots.

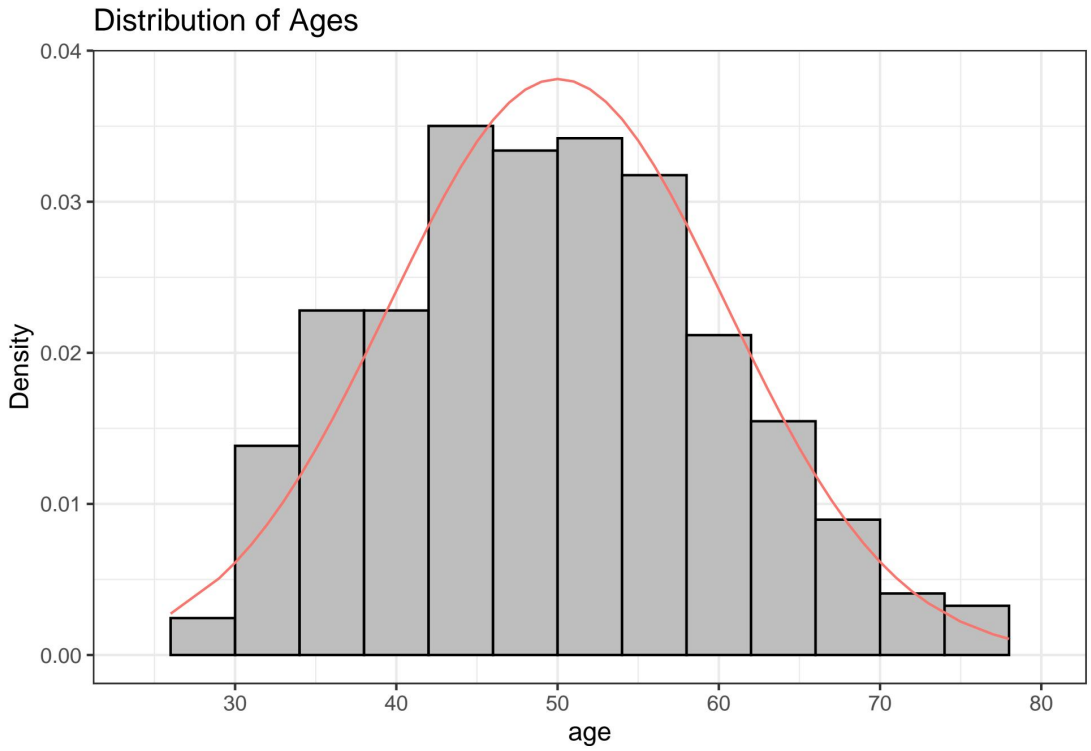
```
# general data frame that I've used throughout EDA.
#sex
df3 <- cholangitis %>%
  group_by(sex) %>%
  count()
#drug and stage
df <- cholangitis %>%
  group_by(drug, stage) %>%
  count()
#drug and status
df1 <- cholangitis %>%
  group_by(drug, status) %>%
  count()
#status and stage
df4 <- cholangitis %>%
  group_by(status, stage) %>%
  count()
#gender and status
df5 <- cholangitis %>%
  group_by(status, sex) %>%
  count()
#gender and drug
df6 <- cholangitis %>%
  group_by(drug, sex) %>%
  count()

dff <- cholangitis %>%
  group_by(n_days, age, bilirubin, cholesterol, albumin, copper, alk_phos, sgot,
           trygliceride, platelets, prothrombin)
```

This code chunk is the code for my dataframes, making it more easier to identify what data frame I've used to make following codes. It's all labeled with what kind of data the dataframe is consisted of.

```
cholangitis$age <- round(cholangitis$age/365)
age <- cholangitis$age
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")
```

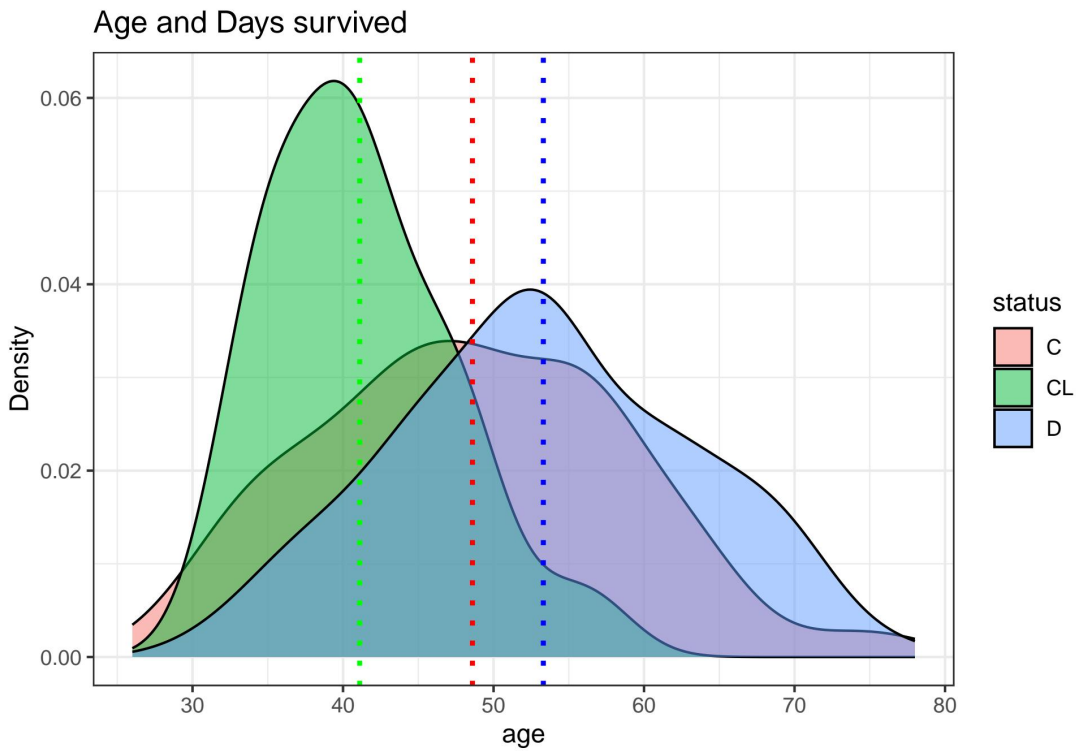
Question assigned to the following page: 2



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')$age),
             color = 'red', linetype="dotted", lwd= 1) +
  geom_vline(xintercept = mean(filter(cholangitis, status == 'CL')$age),
             color = 'green', linetype="dotted", lwd = 1) +
  geom_vline(xintercept = mean(filter(cholangitis, status == 'D')$age),
             color = 'blue', linetype="dotted", lwd =1) +
  theme_bw()
```

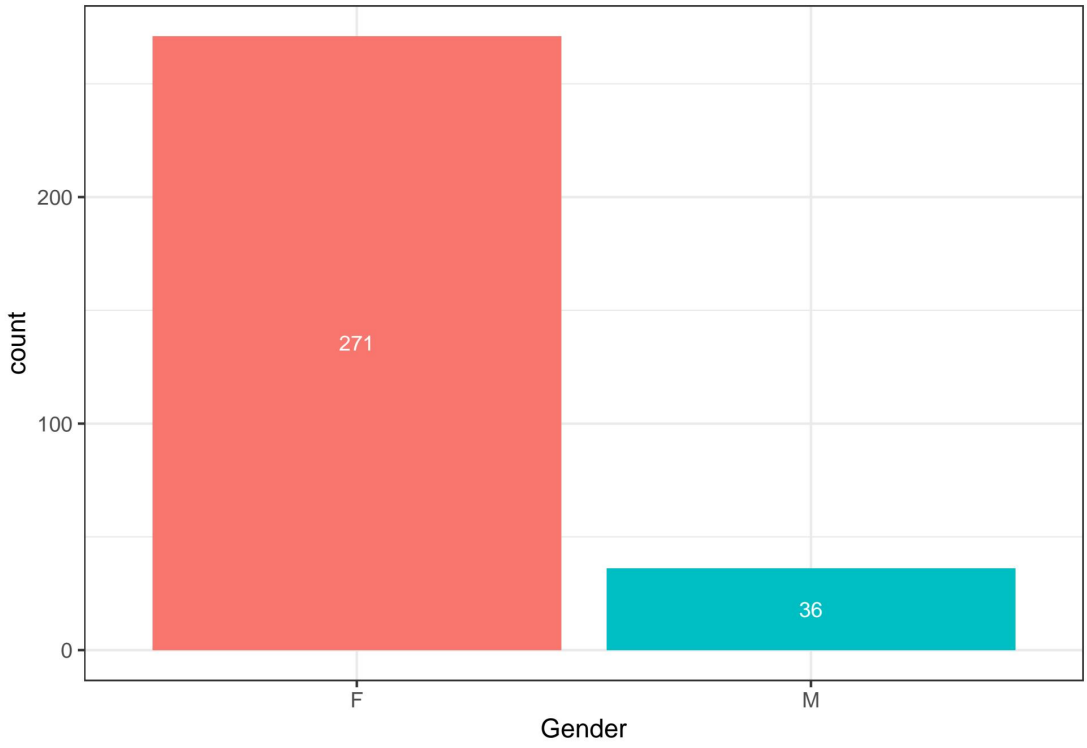
Question assigned to the following page: 2



Then to find out if there was any correlation among age and the days survived, I've used the density plot to compare them. The plot is filled, showing status for each based on the age. 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. The dotted lines indicates the mean of the data(age) with the following color indicated through legend.

```
sex_dist <- df3 %>%
  ggplot(aes(x = sex, y = n, fill = sex, label = n)) +
  geom_col() +
  theme_bw() +
  labs(x = "Gender", y = "count") +
   geom_text(position = position_stack(vjust = 0.5), size = 3,
             color = "#ffffff") +
  theme(legend.position = "none")
sex_dist
```

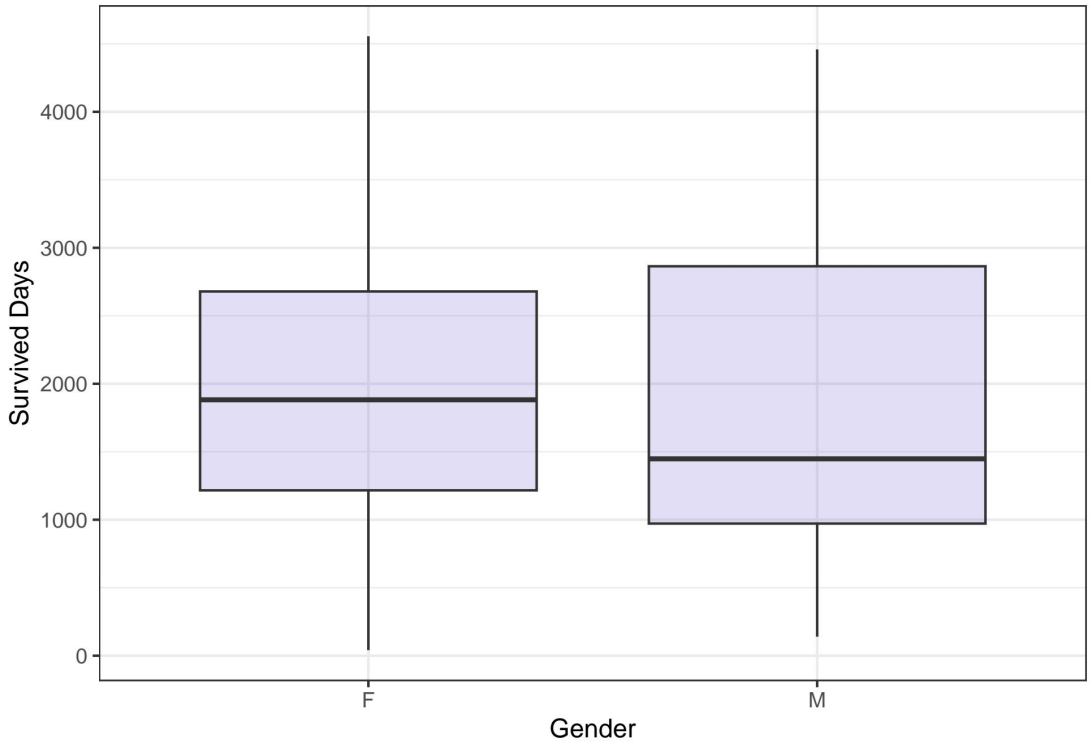
Question assigned to the following page: 2



The gender count seems a bit interesting, as there are about 9 more times of females in the experiement than the male participants. Just by looking at the bar, we are able to notice the difference.

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

Question assigned to the following page: 2



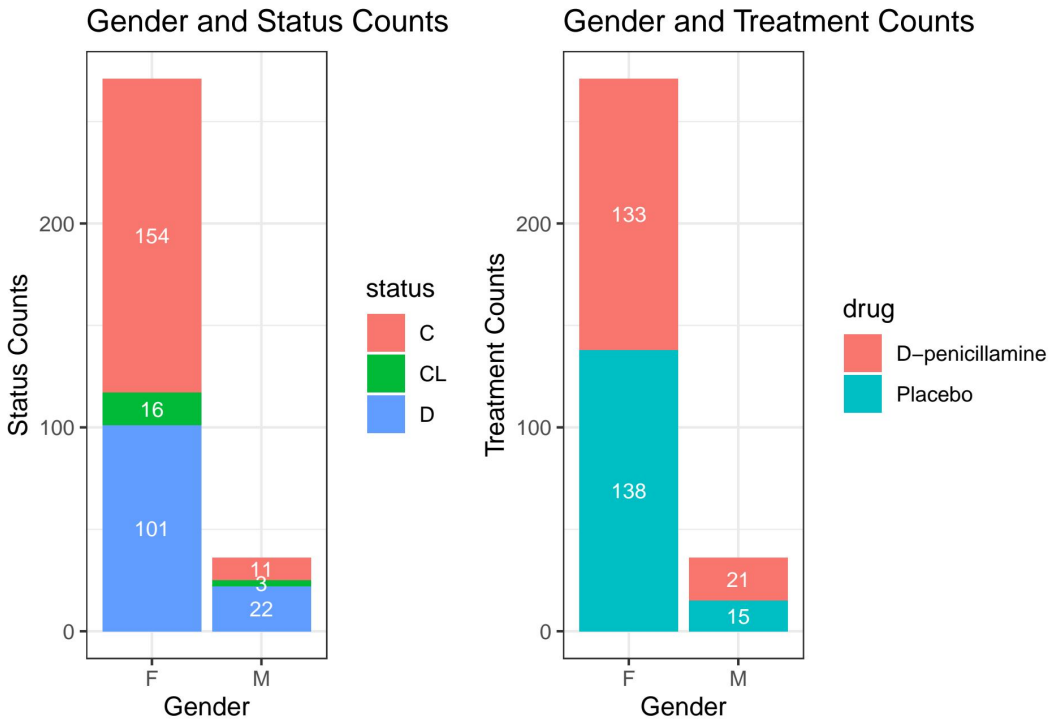
Adding on from gender, I was wondering if there was a difference among gender for the days of survival, therefore I plotted through a boxplot to see if there was a vast difference. The mean seemed to be lower for male, but they seemed not much of a difference in a whole.

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

gen_stat + gen_drug
```

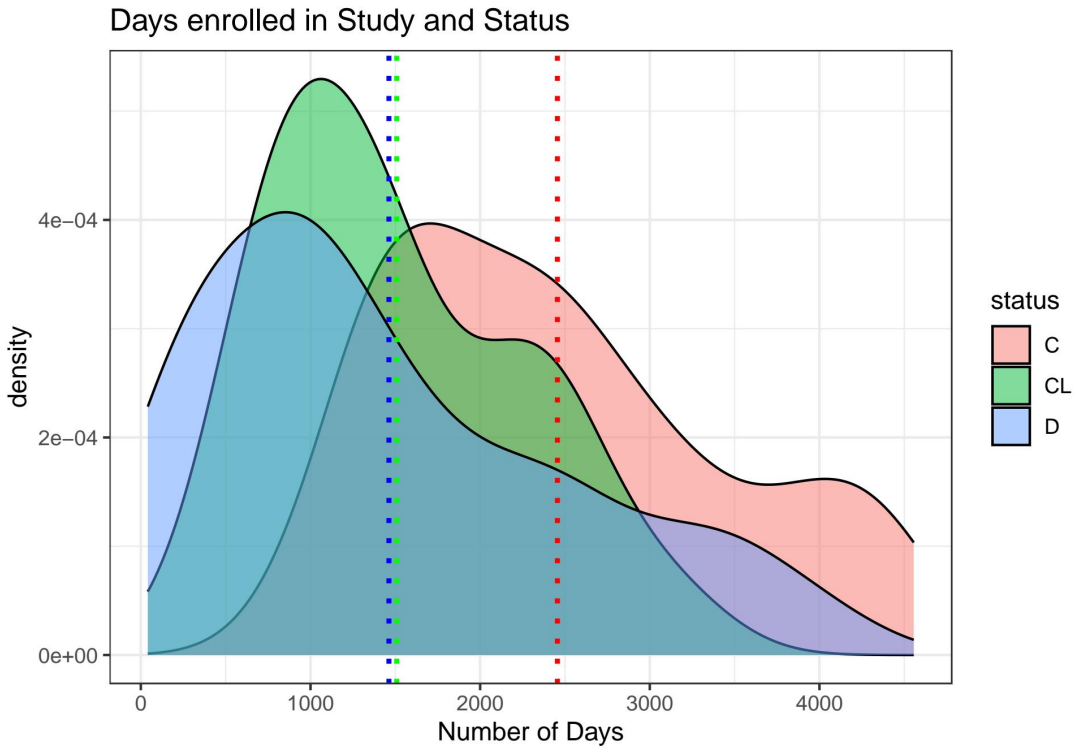
Question assigned to the following page: 2



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')$n_days),
             color = 'red', linetype="dotted", lwd = 1) +
  geom_vline(xintercept = mean(filter(cholangitis, status == 'CL')$n_days),
             color = 'green', linetype="dotted", lwd = 1) +
  geom_vline(xintercept = mean(filter(cholangitis, status == 'D')$n_days),
             color = 'blue', linetype="dotted", lwd = 1) +
  theme_bw()
```

Question assigned to the following page: 2



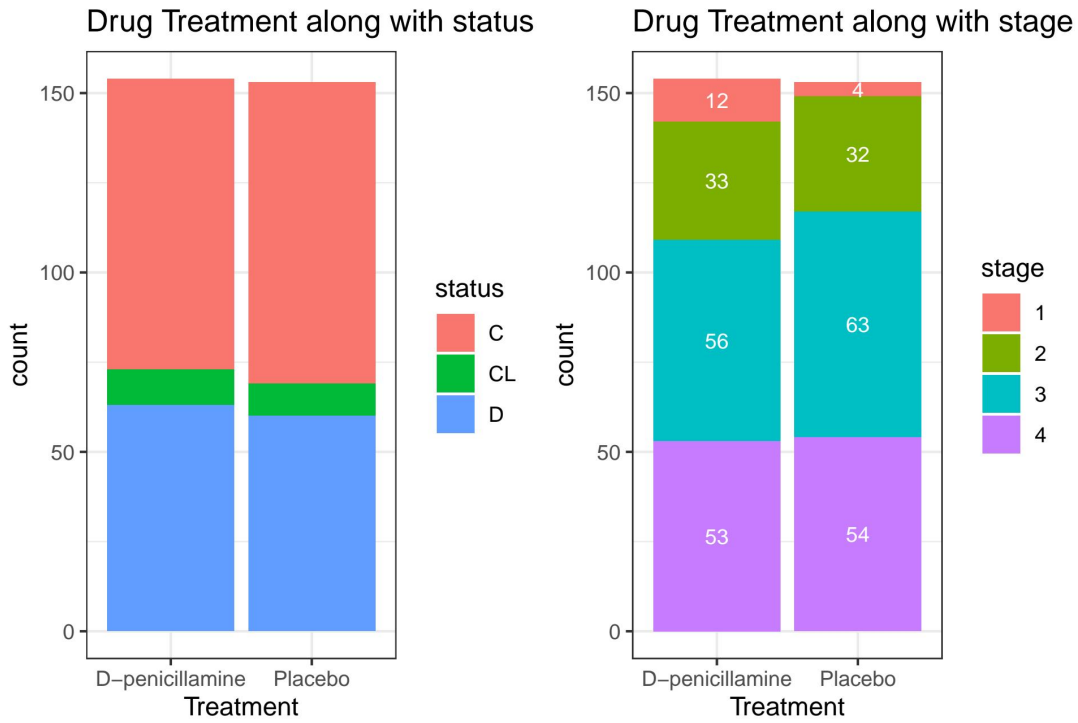
Moving on from gender, now looking at the number of days enrolled in the study and the status of them, the groups that died were skewed the most, having the lowest average of the days of survival. Those who didn't died with the highest average, but what was interesting is that it's kind of right skewed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

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

drug_stage <- df %>%
  ggplot(aes(x = drug, y = n, fill = stage, label = n)) +
  ggtitle("Drug Treatment along with stage") +
  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
```

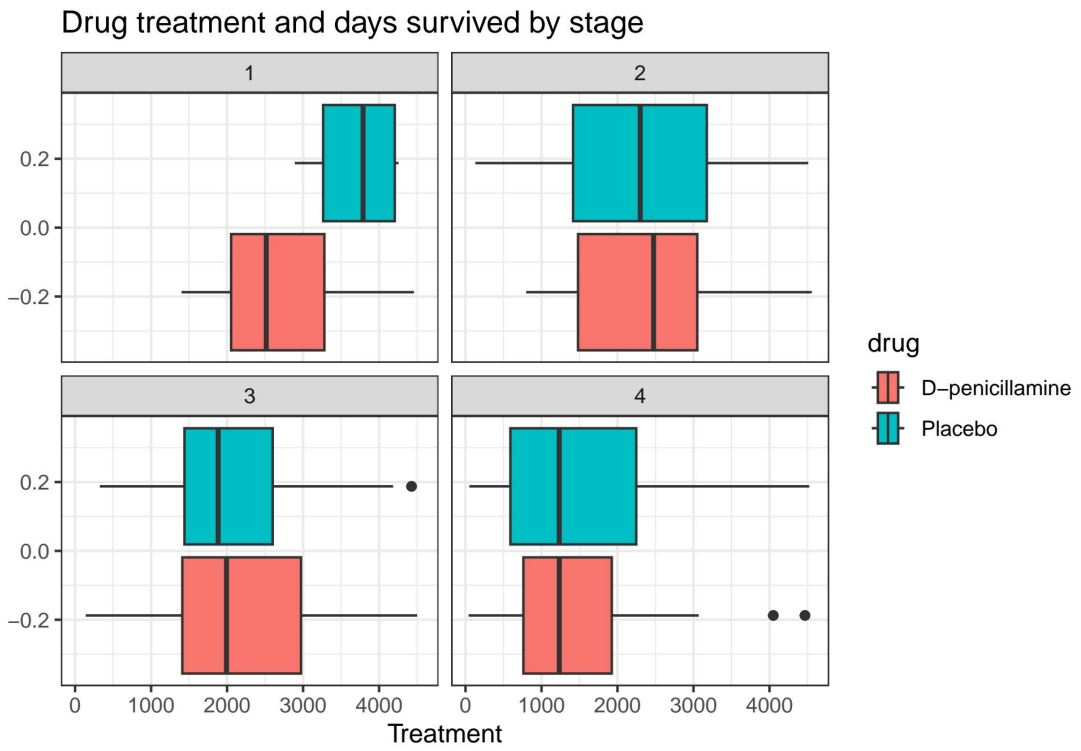
Question assigned to the following page: 2



Now comparing the treatment given with the current status and the stage, the left plot is representing the drug and it's count and each of the portion explains the status. The right plot is representing the drug given and the portion of 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 = n_days)) +
  theme_bw() + geom_boxplot(aes(fill = drug)) +
  xlab("Treatment") + ggtitle("Drug treatment and days survived by stage") +
  facet_wrap(~ stage)
```

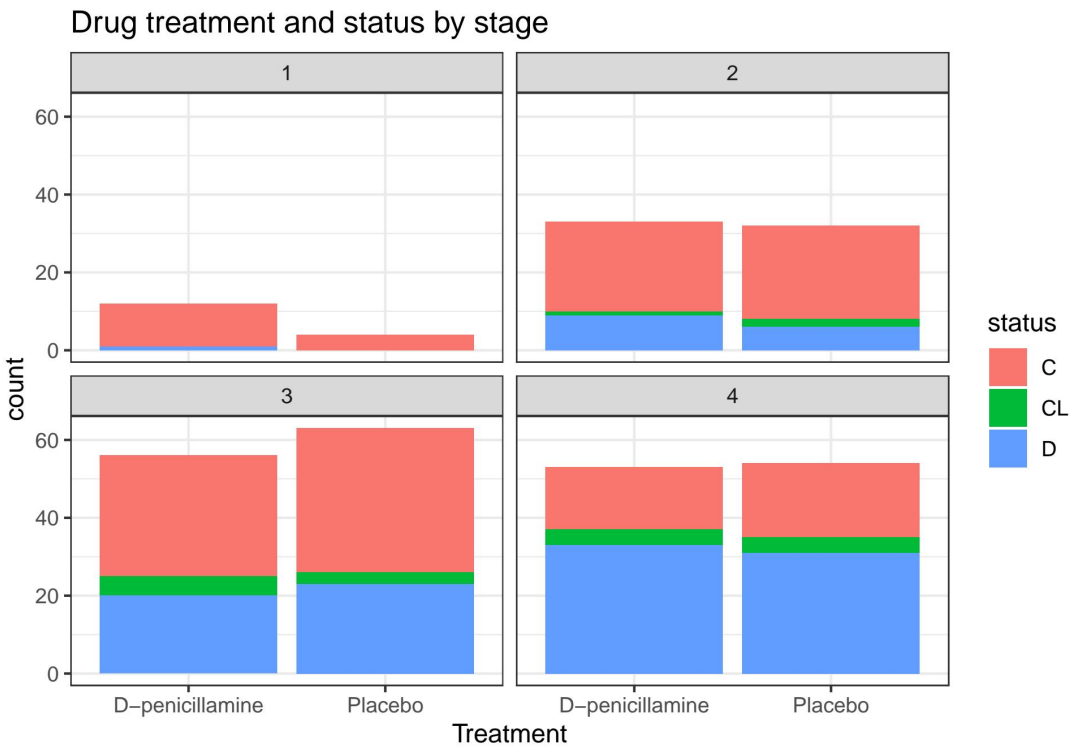
Question assigned to the following page: 2



To find the surived days of the patients, I wanted to find out if there would be a difference among the stage with the difference of treatment given with the days of survival. All except for the first stage, the means seem to be similar regardless of the treatment they've recieved but for the first stage, it kind of shows difference among the drug that had been applied, which was kind of intresting to find.

```
ggplot(cholangitis, aes(x = drug, fill = status)) +
  geom_bar() + xlab("Treatment") +
  ggtitle("Drug treatment and status by stage") +
  facet_wrap(~ stage) +
  theme_bw()
```

Question assigned to the following page: 2



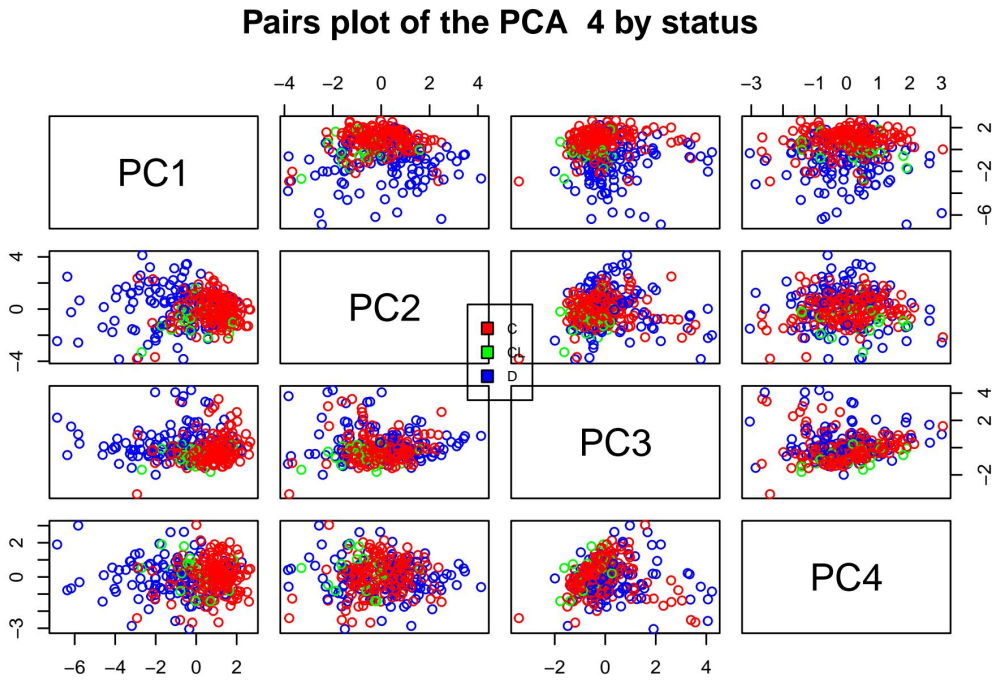
Now comparing the treatment with the treatments that were assigned on each stage and their survival, it shows that there are more and more deaths occured with the stage development, but since the proportion of treatments assigned seems mostly identical to each other. What's interesting is however, that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

```
cholangitis_explanatory <- cholangitis[,c(-1,-3,-4,-6:-10,-20)]
cholangitisSCALED = scale(cholangitis_explanatory)
cholangitis.pca = prcomp(cholangitisSCALED, scale = F)

colstatus <- c("red", "green", "blue")
plotPCA <- cholangitis.pca$x[,1:4]
names(colstatus) <- levels(cholangitis$status)

pairs(plotPCA, col = colstatus[cholangitis$status],
      main = "Pairs plot of the PCA  4 by status",
      pch = 21)
  legend("center",legend =names(colstatus), fill= colstatus, cex = 0.5)
```

Question assigned to the following page: 2

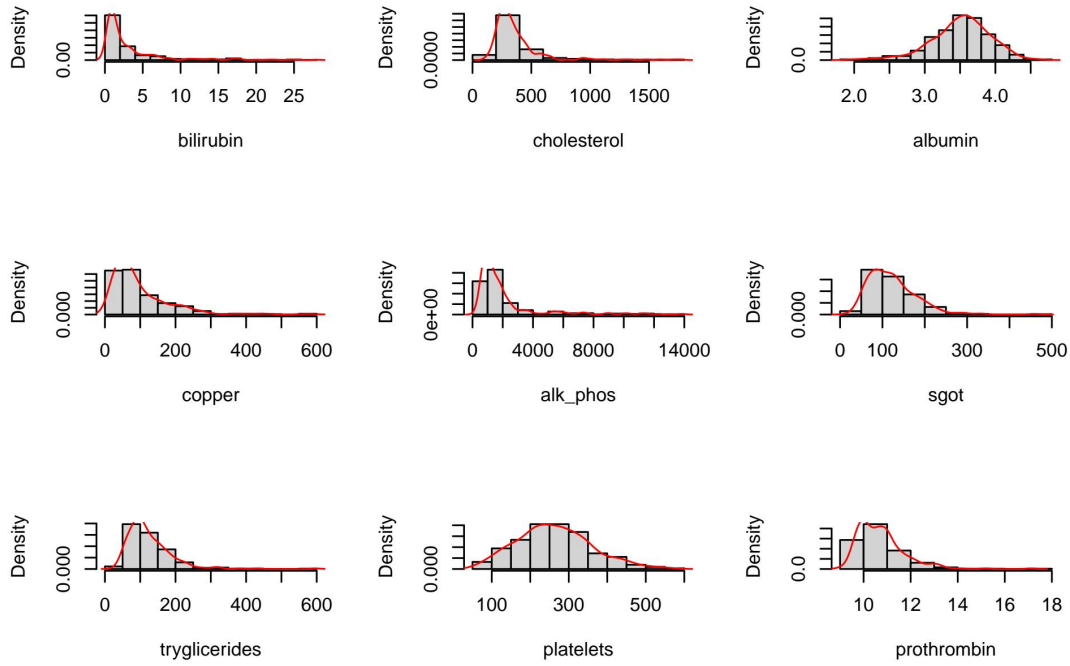


This is the PCA plotted by pairs of the explanatory data of the cholangitis dataset. Red, Green and Blue indicates the staus of lived, received liver transplant, and dead respectively in order. It is interesting to find, but in the same time kinds of make sense that the deads are showing difference that those dead are shown as outliers. They are the ones suffering dieases, showing outliers in the health data. The individual data seems to cluster.

```
# subsetting
subset_chol <- cholangitis[, names(which(sapply(cholangitis, is.numeric)))]
subset_chol <- subset(subset_chol, select = -c(1:3))

#code for histograms
par(mfrow =c(3,3))
for (i in names(subset_chol)) {
  hist(subset_chol[, i], freq = FALSE, xlab = i, main ="")
  lines(density(subset_chol[, i]), col ="red")
}
```

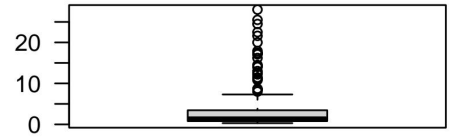
Question assigned to the following page: 2



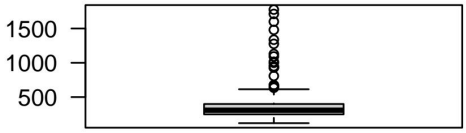
I plotted the histograms to find out the genral distribution of the genral health information, and surprisingly I was able to find out that most of the information were right-skewed, and were able to identify the outliers, except albumin and platelets. There were values far off from the where data is distributed. This later will be turned into log to see if they show a better distribution than they intially were.

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

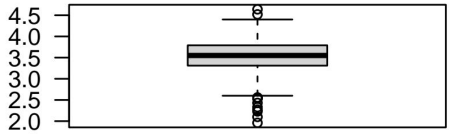
Question assigned to the following page: 2



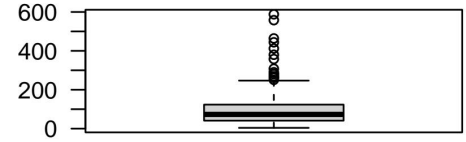
bilirubin



cholesterol

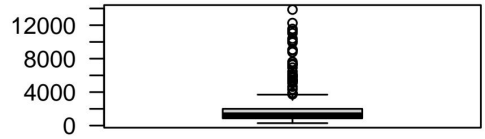


albumin

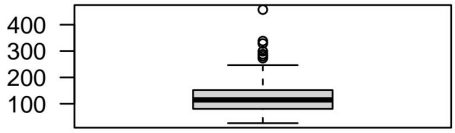


copper

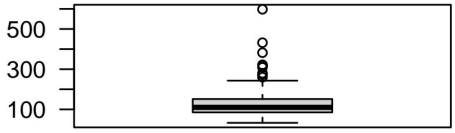
Question assigned to the following page: 2



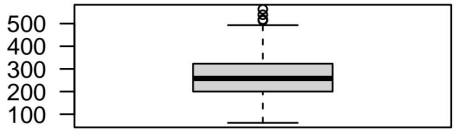
alk_phos



sgot

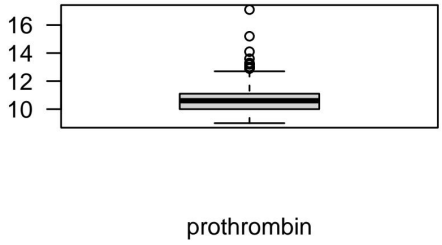


trygliceride s



platelets

Questions assigned to the following page: 2 and 3



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

Multivariate Regression

Data Processing

```
data <- subset(cholangitis[, -1])
data <- na.omit(data)
```

First for the data processing, I removed the 'id' column which I thought is not meaningful in the context of regression analysis. It is the unique number that is given to the individuals, which wouldn't make such differnece in the regression analysis. As we are trying to find the effect of the drug on the days of survival, it would make sense to remove the NA drugs. I also removed the rows containing missing values, as I've did previously.

Multivariate regression analysis

First, I'm going to run the regression model based on n_days as our response variable, and the summary of the linear model alongside regression diagnostics.

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

```
##
```

Question assigned to the following page: 3

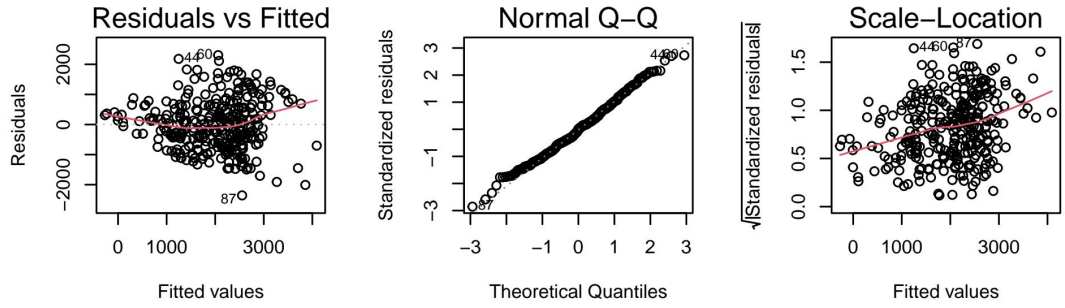
```
## Call:
## lm(formula = n_days ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2356.05  -609.43    11.36   566.25  2304.34
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.039e+03  9.498e+02  -1.094 0.274824
## statusCL      -6.669e+02  2.214e+02  -3.012 0.002833 **
## statusD       -5.872e+02  1.336e+02  -4.395 1.57e-05 ***
## drugPlacebo    3.967e+00  1.027e+02   0.039 0.969224
## age           -2.451e+00  5.583e+00  -0.439 0.660918
## sexM           8.876e+01  1.728e+02   0.514 0.607853
## ascitesY      -5.015e+01  2.639e+02  -0.190 0.849431
## hepatomegalyY -2.608e+01  1.205e+02  -0.216 0.828758
## spidersY      -9.592e+01  1.265e+02  -0.758 0.448961
## edemaS        -1.392e+02  1.861e+02  -0.748 0.454883
## edemaY        -3.489e+02  2.827e+02  -1.234 0.218170
## bilirubin     -5.307e+01  1.679e+01  -3.160 0.001747 **
## cholesterol   -1.930e-01  2.649e-01  -0.729 0.466906
## albumin        5.636e+02  1.454e+02   3.877 0.000131 ***
## copper        -1.756e+00  7.246e-01  -2.424 0.015993 *
## alk_phos       1.358e-01  2.459e-02   5.523 7.52e-08 ***
## sgot           7.256e-01  1.052e+00   0.690 0.491023
## tryglicerides  7.910e-01  8.978e-01   0.881 0.379042
## platelets      4.468e-01  5.851e-01   0.764 0.445742
## prothrombin    1.612e+02  6.109e+01   2.639 0.008775 **
## stage2        -2.256e+02  2.515e+02  -0.897 0.370373
## stage3        -3.448e+02  2.462e+02  -1.401 0.162427
## stage4        -5.255e+02  2.650e+02  -1.983 0.048301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 861.3 on 284 degrees of freedom
## Multiple R-squared:  0.4592, Adjusted R-squared:  0.4173
## F-statistic: 10.96 on 22 and 284 DF,  p-value: < 2.2e-16
```

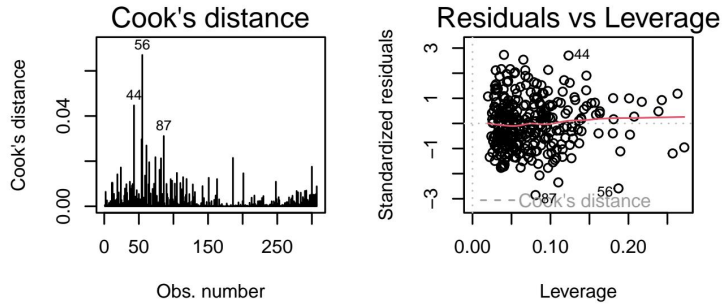
With this plot, we are able to find out that the values of R-square and Adjusted R-squared, along with the residual standard error and significant variables to compare with the transformed regression data. With the result we currently have, R-squared of 0.4592 and Adjusted R-squared of 0.4173, which is to say, it is hard to say we have a 'best' model so far.

Now, with the regression created, we are going to run a diagnostics test to see how the initial data would need transformation in any kind.

```
par(mfrow=c(2,3))
plot(initial_model, which = c(1:5))
```

Question assigned to the following page: 3





From the basic observation of the regression diagnostics, there seems to be few outliers that has to be identified for a better dataset. The residual mean isn't close to zero also the decrease and increase pattern in the residuals vs fitted plot shows heteroscedasticity, Normal Q-Q show that it's not following the straight line, which indicates that there might be a slight skew from the normal distribution, Scale-Location plot also indicates that there is heteroscedasticity. Cook's distance indicates the outliers that stands out.

```
identifying_outliers <- which(cooks.distance(initial_model) > 0.03)
identifying_outliers
```

```
## 44 56 87
## 43 55 86
```

Cooks Distance: The outliers were tried to defined thats above the 0.03 threshold, as those were the points that was spiking looking directly through the plot. 87 seems to be in the similar position to some of the points, which may seem that it isn't regarded as outliers, however if I plot them, I can identify with other plots that 87 is the point that's spiking out a lot of time, that's why I also decided to exclude it.

```
data = data[-c(44, 56, 87), ]
```

Next, we will try to transform the response variable into log, to see if it will show any outcomes that are interesting or approved from the original data. The reason we are trying to transform the data is because in the EDA part, we observed that the distribution of our response variable (n_days) is right-skewed. We are trying to find a better fit overall.

```
log_model <- lm(log(n_days) ~ ., data = data)
summary(log_model)
```

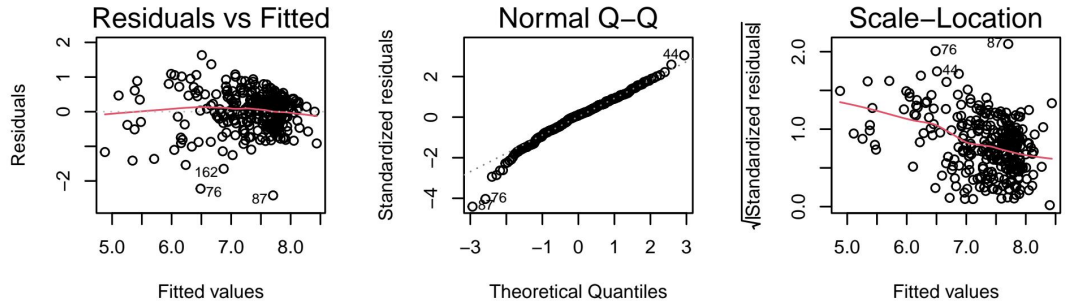
Question assigned to the following page: 3

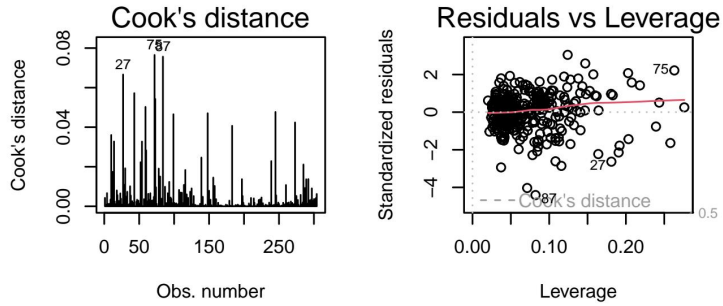
```
##
## Call:
## lm(formula = log(n_days) ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.41669 -0.32735  0.04674  0.35048  1.63170
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.673e+00  6.353e-01  10.504  < 2e-16 ***
## statusCL      -3.815e-01  1.473e-01  -2.590 0.010105 *
## statusD       -3.565e-01  8.885e-02  -4.013  7.7e-05 ***
## drugPlacebo   -2.964e-02  6.846e-02  -0.433 0.665422
## age           -2.759e-03  3.723e-03  -0.741 0.459214
## sexM           8.406e-02  1.152e-01   0.729 0.466400
## ascitesY      -2.906e-01  1.754e-01  -1.657 0.098671 .
## hepatomegalyY  6.630e-02  8.024e-02   0.826 0.409390
## spidersY      -5.857e-02  8.462e-02  -0.692 0.489432
## edemaS        -1.985e-01  1.289e-01  -1.540 0.124781
## edemaY        -6.055e-01  1.881e-01  -3.219 0.001435 **
## bilirubin     -4.064e-02  1.120e-02  -3.628 0.000339 ***
## cholesterol    1.463e-04  1.761e-04   0.831 0.406787
## albumin        3.209e-01  9.670e-02   3.318 0.001026 **
## copper        -1.366e-03  4.850e-04  -2.817 0.005195 **
## alk_phos       6.414e-05  1.679e-05   3.819 0.000165 ***
## sgot          -7.397e-05  6.997e-04  -0.106 0.915883
## tryglicerides  6.174e-04  5.980e-04   1.032 0.302793
## platelets      2.684e-04  3.947e-04   0.680 0.497033
## prothrombin    4.710e-03  4.110e-02   0.115 0.908835
## stage2        -1.170e-01  1.674e-01  -0.699 0.485099
## stage3        -1.423e-01  1.639e-01  -0.868 0.386089
## stage4        -2.688e-01  1.763e-01  -1.524 0.128509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5721 on 281 degrees of freedom
## Multiple R-squared:  0.5565, Adjusted R-squared:  0.5218
## F-statistic: 16.03 on 22 and 281 DF,  p-value: < 2.2e-16
```

It's interesting that the R-squared has been improved along with the lower residual standard error, which respectively explains that it explains more of the variance in the dependent variable and predictions are closer to the observed values on average. However, not just interpreting from this data, we will also have to compare the plots as well.

```
par(mfrow=c(2,3))
plot(log_model, which = c(1:5))
```

Question assigned to the following page: 3





The log transformed data wasn't effective. The Residuals vs Fitted models indicate that the resiudal mean is closer to zero compared to the initial model, but looking at the normal Q-Q, it was definately showing more skewness in graph, by looking at the fact that it's not closer to the line. In the Scale-Location plot, we can see a decrease pattern, where we can identify heteroscedasticity.

Next, we will look into account of the sqrt.

```
sqrt_model <- lm(sqrt(n_days) ~ ., data = data)
summary(sqrt_model)
```

```
##
## Call:
## lm(formula = sqrt(n_days) ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.064  -7.041   0.388   6.604  27.679
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.9767106 11.0523666   1.536 0.125657
## statusCL      -7.6863623  2.5632408  -2.999 0.002954 **
## statusD       -6.9021189  1.5458285  -4.465 1.16e-05 ***
## drugPlacebo   -0.2105645  1.1911694  -0.177 0.859815
## age           -0.0337549  0.0647754  -0.521 0.602703
## sexM           1.1940797  2.0051386   0.596 0.551982
```

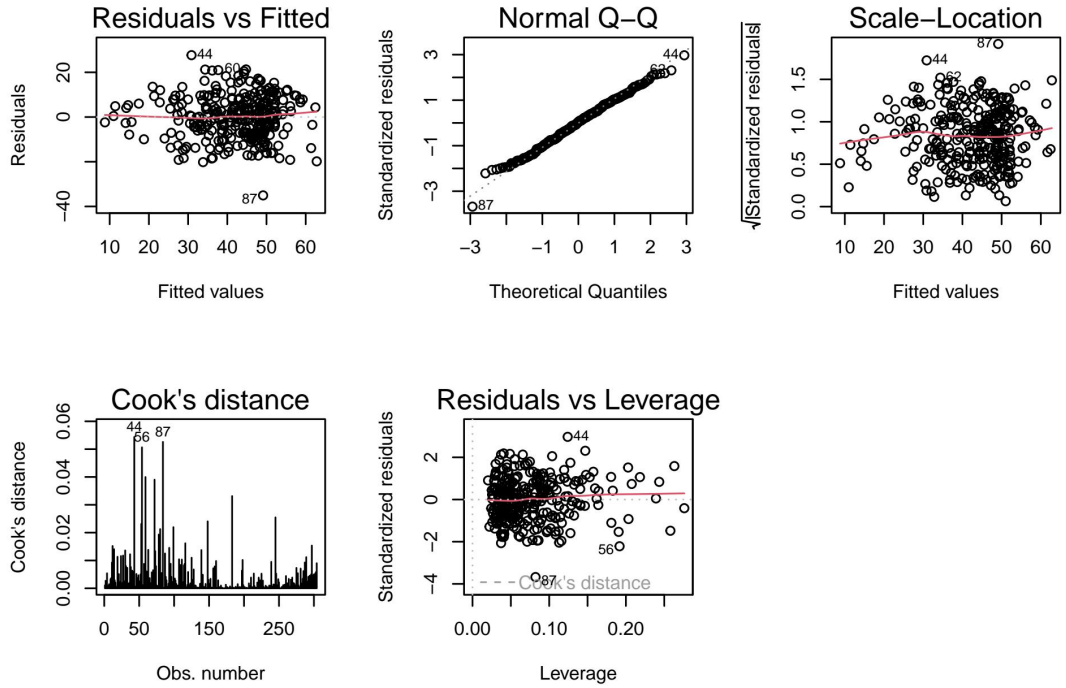
Question assigned to the following page: 3

```
## ascitesY      -2.2493320  3.0512562  -0.737 0.461627
## hepatomegalyY  0.1722153  1.3961008   0.123 0.901915
## spidersY      -0.9677032  1.4722451  -0.657 0.511528
## edemaS        -2.6094754  2.2430176  -1.163 0.245663
## edemaY        -6.7663312  3.2720665  -2.068 0.039564 *
## bilirubin     -0.6627395  0.1948734  -3.401 0.000769 ***
## cholesterol   -0.0005727  0.0030646  -0.187 0.851882
## albumin        6.3593466  1.6824950   3.780 0.000192 ***
## copper        -0.0239207  0.0084388  -2.835 0.004921 **
## alk_phos       0.0013680  0.0002922   4.682 4.43e-06 ***
## sgot           0.0058096  0.0121733   0.477 0.633558
## tryglicerides  0.0112431  0.0104044   1.081 0.280798
## platelets      0.0056005  0.0068678   0.815 0.415492
## prothrombin    1.0670172  0.7150430   1.492 0.136758
## stage2        -2.6162477  2.9121935  -0.898 0.369754
## stage3        -3.4950244  2.8522244  -1.225 0.221463
## stage4        -5.7884586  3.0677846  -1.887 0.060211 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.953 on 281 degrees of freedom
## Multiple R-squared:  0.5106, Adjusted R-squared:  0.4723
## F-statistic: 13.33 on 22 and 281 DF,  p-value: < 2.2e-16
```

Compared to the initial model, it has higher R-squared, however, lower than the log transformed model. The Residual are lower than the initial model, but higher than the log transformed data. As it's higher than the initial model it implies that transformation method would improve our regression model.

```
par(mfrow=c(2,3))
plot(sqrt_model,which= c(1:5))
```

Question assigned to the following page: 3

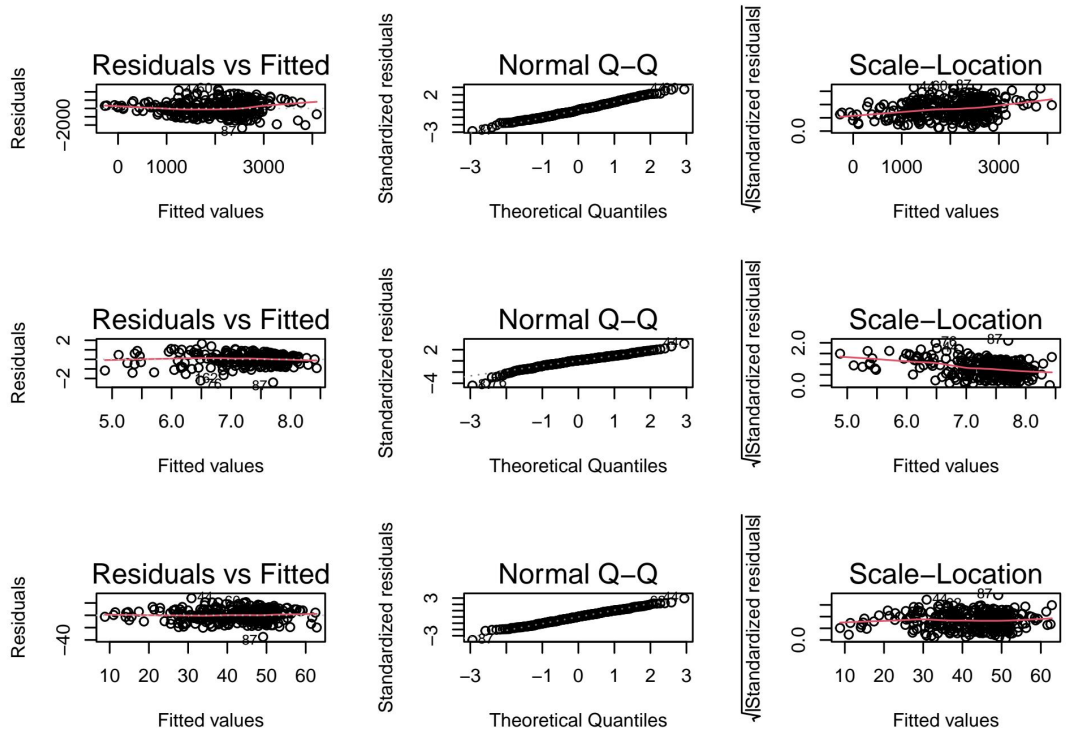


> Residuals are randomly distributed, and its mean are almost equal to zero. Besides, Q-Q plot shows that most points lies on the line, and for the Scale-Location as well, we can see comparably less increase and decrease pattern. So we can say that this model satisfies the assumptons of linearity.

Comparing the Models

```
par(mfrow = c(3,3))
plot(initial_model, which = c(1,2,3))
plot(log_model, which = c(1,2,3))
plot(sqrt_model, which = c(1,2,3))
```

Questions assigned to the following page: 3 and 4



> Comparing from the plots, First looking at the Residuals vs Fitted particaularly, the line of the sqrt plots seemed to be a straight line, the mean of residual is closer to zero compared to other grpahs. Looking at the nomal Q-Q graph, there isn't much of a difference between them. However, in the Scale - Location model, compared to the initial model, the sqrt model seems to have almsot no increase or decrease where we can say that almost no heteroscedasticity is found. Therefore we choose the third plot, the sqrt model plot.

Variable Selection

To find the ultimate model, we will use the function step for variable selection. Alongside looking at the plot itself, but also for the r-squared values, we can indicate it's a better plot to start at.

```
data_under_3 <- data %>% select(-c("status", "edema", "stage"), )
sqrt.model.for.step <- lm(sqrt(n_days) ~ ., data = data_under_3)
```

Since we were told to avoid using variables that have more than two levels, status, edema and stage were excluded from the data.

```
step(sqrt.model.for.step, trace = 1, direction = "both")
```

```
## Start:  AIC=1443.73
## sqrt(n_days) ~ drug + age + sex + ascites + hepatomegaly + spiders +
##     bilirubin + cholesterol + albumin + copper + alk_phos + sgot +
##     tryglicerides + platelets + prothrombin
##
##                 Df Sum of Sq   RSS    AIC
## - drug           1      0.10 31599 1441.7
## - sex            1      0.12 31599 1441.7
## - cholesterol    1      0.85 31600 1441.7
```

Question assigned to the following page: 4

```
## - prothrombin    1      2.06 31601 1441.8
## - sgot           1     18.22 31618 1441.9
## - platelets      1     86.81 31686 1442.6
## - trygliceride   1     88.86 31688 1442.6
## - age            1    159.15 31758 1443.3
## - spiders        1    207.84 31807 1443.7
## <none>                       31599 1443.7
## - hepatomegaly   1    221.64 31821 1443.9
## - ascites        1    659.85 32259 1448.0
## - copper         1   1446.90 33046 1455.3
## - bilirubin      1   1611.67 33211 1456.9
## - alk_phos       1   1885.70 33485 1459.3
## - albumin        1   2245.97 33845 1462.6
##
## Step:  AIC=1441.74
## sqrt(n_days) ~ age + sex + ascites + hepatomegaly + spiders +
##     bilirubin + cholesterol + albumin + copper + alk_phos + sgot +
##     trygliceride + platelets + prothrombin
##
##                 Df Sum of Sq   RSS    AIC
## - sex            1      0.11 31600 1439.7
## - cholesterol    1      0.89 31600 1439.7
## - prothrombin    1      2.17 31602 1439.8
## - sgot           1     18.22 31618 1439.9
## - platelets      1     87.32 31687 1440.6
## - trygliceride   1     88.82 31688 1440.6
## - age            1    162.81 31762 1441.3
## <none>                       31599 1441.7
## - spiders        1    209.18 31809 1441.7
## - hepatomegaly   1    224.16 31824 1441.9
## + drug           1      0.10 31599 1443.7
## - ascites        1    663.51 32263 1446.0
## - copper         1   1448.39 33048 1453.4
## - bilirubin      1   1620.06 33219 1454.9
## - alk_phos       1   1887.02 33486 1457.4
## - albumin        1   2246.35 33846 1460.6
##
## Step:  AIC=1439.74
## sqrt(n_days) ~ age + ascites + hepatomegaly + spiders + bilirubin +
##     cholesterol + albumin + copper + alk_phos + sgot + trygliceride +
##     platelets + prothrombin
##
##                 Df Sum of Sq   RSS    AIC
## - cholesterol    1      0.84 31600 1437.7
## - prothrombin    1      2.31 31602 1437.8
## - sgot           1     18.22 31618 1437.9
## - platelets      1     87.93 31687 1438.6
## - trygliceride   1     89.33 31689 1438.6
## - age            1    167.61 31767 1439.3
## <none>                       31600 1439.7
## - spiders        1    220.63 31820 1439.8
## - hepatomegaly   1    224.45 31824 1439.9
## + sex            1      0.11 31599 1441.7
## + drug           1      0.09 31599 1441.7
```

Question assigned to the following page: 4

```
## - ascites        1    663.95 32263 1444.1
## - copper         1   1600.41 33200 1452.8
## - bilirubin      1   1667.68 33267 1453.4
## - alk_phos       1   1888.44 33488 1455.4
## - albumin        1   2289.34 33889 1459.0
##
## Step:  AIC=1437.74
## sqrt(n_days) ~ age + ascites + hepatomegaly + spiders + bilirubin +
##     albumin + copper + alk_phos + sgot + tryglicerides + platelets +
##     prothrombin
##
##                 Df Sum of Sq   RSS    AIC
## - prothrombin    1      2.59 31603 1435.8
## - sgot           1     20.58 31621 1435.9
## - platelets      1     88.00 31688 1436.6
## - tryglicerides  1     88.62 31689 1436.6
## - age            1    166.88 31767 1437.3
## <none>                       31600 1437.7
## - spiders        1    220.23 31821 1437.9
## - hepatomegaly   1    227.60 31828 1437.9
## + cholesterol    1      0.84 31600 1439.7
## + drug           1      0.13 31600 1439.7
## + sex            1      0.06 31600 1439.7
## - ascites        1    668.92 32269 1442.1
## - copper         1   1606.18 33207 1450.8
## - bilirubin      1   1823.04 33423 1452.8
## - alk_phos       1   1887.86 33488 1453.4
## - albumin        1   2292.92 33893 1457.0
##
## Step:  AIC=1435.77
## sqrt(n_days) ~ age + ascites + hepatomegaly + spiders + bilirubin +
##     albumin + copper + alk_phos + sgot + tryglicerides + platelets
##
##                 Df Sum of Sq   RSS    AIC
## - sgot           1     21.15 31624 1434.0
## - platelets      1     85.46 31688 1434.6
## - tryglicerides  1     86.06 31689 1434.6
## - age            1    164.56 31768 1435.3
## <none>                       31603 1435.8
## - spiders        1    217.95 31821 1435.9
## - hepatomegaly   1    227.11 31830 1436.0
## + prothrombin    1      2.59 31600 1437.7
## + cholesterol    1      1.12 31602 1437.8
## + drug           1      0.26 31603 1437.8
## + sex            1      0.18 31603 1437.8
## - ascites        1    672.60 32276 1440.2
## - copper         1   1604.00 33207 1448.8
## - bilirubin      1   1906.27 33509 1451.6
## - alk_phos       1   1914.17 33517 1451.7
## - albumin        1   2297.30 33900 1455.1
##
## Step:  AIC=1433.97
## sqrt(n_days) ~ age + ascites + hepatomegaly + spiders + bilirubin +
##     albumin + copper + alk_phos + tryglicerides + platelets
```

Question assigned to the following page: 4

```
##
##                 Df Sum of Sq   RSS    AIC
## - tryglicerides  1     93.32 31717 1432.9
## - platelets      1     96.65 31721 1432.9
## - age            1    147.79 31772 1433.4
## <none>                       31624 1434.0
## - spiders        1    210.69 31835 1434.0
## - hepatomegaly   1    223.97 31848 1434.1
## + sgot           1     21.15 31603 1435.8
## + cholesterol    1      3.81 31620 1435.9
## + prothrombin    1      3.16 31621 1435.9
## + drug           1      0.34 31624 1436.0
## + sex            1      0.13 31624 1436.0
## - ascites        1    655.23 32279 1438.2
## - copper         1   1676.00 33300 1447.7
## - alk_phos       1   1895.74 33520 1449.7
## - bilirubin      1   2359.36 33983 1453.8
## - albumin        1   2388.86 34013 1454.1
##
## Step:  AIC=1432.87
## sqrt(n_days) ~ age + ascites + hepatomegaly + spiders + bilirubin +
##     albumin + copper + alk_phos + platelets
##
##                 Df Sum of Sq   RSS    AIC
## - platelets      1    119.71 31837 1432.0
## - age            1    144.40 31862 1432.2
## - hepatomegaly   1    208.90 31926 1432.9
## <none>                       31717 1432.9
## - spiders        1    218.46 31936 1433.0
## + tryglicerides  1     93.32 31624 1434.0
## + sgot           1     28.41 31689 1434.6
## + sex            1      0.60 31717 1434.9
## + cholesterol    1      0.42 31717 1434.9
## + prothrombin    1      0.08 31717 1434.9
## + drug           1      0.08 31717 1434.9
## - ascites        1    620.54 32338 1436.8
## - copper         1   1620.70 33338 1446.0
## - alk_phos       1   2000.20 33718 1449.5
## - bilirubin      1   2289.86 34007 1452.1
## - albumin        1   2486.11 34204 1453.8
##
## Step:  AIC=1432.01
## sqrt(n_days) ~ age + ascites + hepatomegaly + spiders + bilirubin +
##     albumin + copper + alk_phos
##
##                 Df Sum of Sq   RSS    AIC
## - age            1    165.18 32002 1431.6
## <none>                       31837 1432.0
## - hepatomegaly   1    245.75 32083 1432.3
## - spiders        1    258.25 32095 1432.5
## + platelets      1    119.71 31717 1432.9
## + tryglicerides  1    116.38 31721 1432.9
## + sgot           1     43.92 31793 1433.6
## + cholesterol    1      2.42 31835 1434.0
```

Question assigned to the following page: 4

```
## + prothrombin    1      2.12 31835 1434.0
## + sex            1      0.37 31837 1434.0
## + drug           1      0.34 31837 1434.0
## - ascites        1    700.64 32538 1436.6
## - copper         1   1611.53 33449 1445.0
## - alk_phos       1   2240.91 34078 1450.7
## - bilirubin      1   2252.13 34089 1450.8
## - albumin        1   2602.34 34439 1453.9
##
## Step:  AIC=1431.59
## sqrt(n_days) ~ ascites + hepatomegaly + spiders + bilirubin +
##     albumin + copper + alk_phos
##
##                 Df Sum of Sq   RSS    AIC
## <none>                       32002 1431.6
## - spiders        1    212.83 32215 1431.6
## + age            1    165.18 31837 1432.0
## - hepatomegaly   1    268.68 32271 1432.1
## + platelets      1    140.48 31862 1432.2
## + tryglicerides  1    114.54 31888 1432.5
## + sgot           1     16.55 31986 1433.4
## + prothrombin    1     14.05 31988 1433.5
## + cholesterol    1     12.83 31989 1433.5
## + sex            1     11.40 31991 1433.5
## + drug           1      4.43 31998 1433.5
## - ascites        1    842.03 32844 1437.5
## - copper         1   1651.51 33654 1444.9
## - bilirubin      1   2184.40 34187 1449.7
## - alk_phos       1   2335.52 34338 1451.0
## - albumin        1   2866.52 34869 1455.7

##
## Call:
## lm(formula = sqrt(n_days) ~ ascites + hepatomegaly + spiders +
##     bilirubin + albumin + copper + alk_phos, data = data_under_3)
##
## Coefficients:
##   (Intercept)       ascitesY  hepatomegalyY       spidersY      bilirubin
##      17.73932       -7.23596       -2.06956       -2.00749       -0.72299
##       albumin         copper       alk_phos
##       8.45130       -0.03145        0.00135
```

Using the stepwise method, we can indicate given the lowest AIC of 1431.59, find the optimal model to predict n_days would be in this case ascites, hpatomegaly, spiders, bilirubin, albumin, copper and alk_phos.

```
AIC_data <- lm(formula = sqrt(n_days) ~ ascites + hepatomegaly + spiders +
    bilirubin + albumin + copper + alk_phos + drug, data = data_under_3)
summary(AIC_data)
```

```
##
## Call:
## lm(formula = sqrt(n_days) ~ ascites + hepatomegaly + spiders +
##     bilirubin + albumin + copper + alk_phos + drug, data = data_under_3)
##
## Residuals:
```

Questions assigned to the following page: 4 and 5

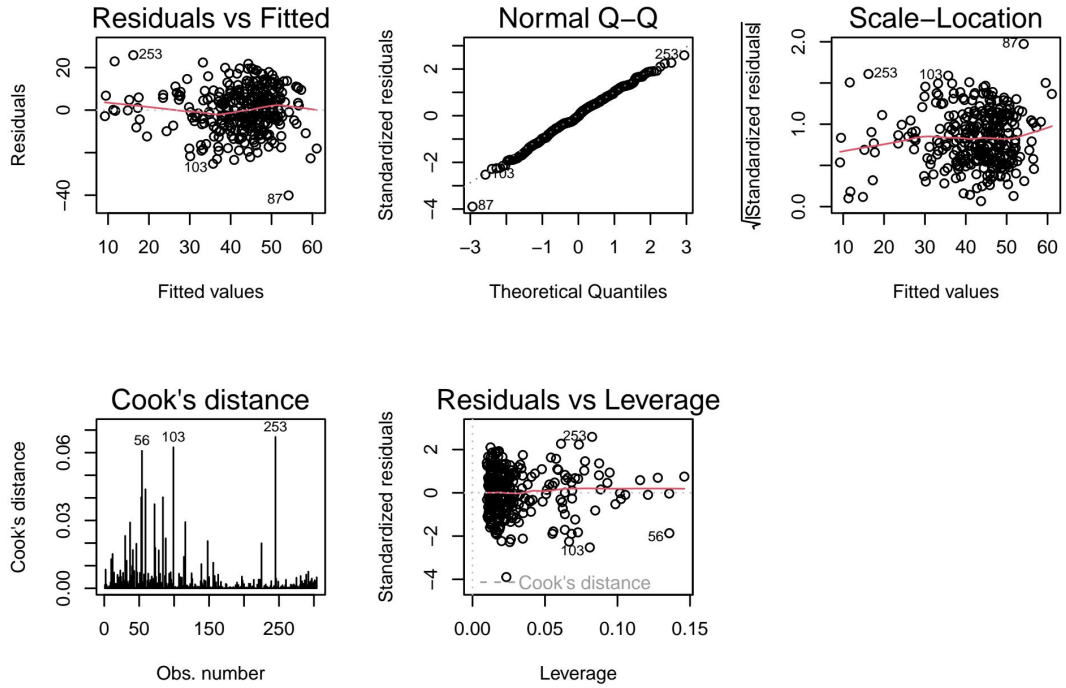
```
##     Min      1Q  Median      3Q     Max
## -40.097  -6.330  -0.174   7.052  25.825
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.6620525  6.2308039   2.835 0.004905 **
## ascitesY      -7.2014969  2.6026517  -2.767 0.006015 **
## hepatomegalyY -2.0996462  1.3233545  -1.587 0.113672
## spidersY      -1.9996340  1.4336720  -1.395 0.164136
## bilirubin     -0.7263365  0.1619566  -4.485 1.05e-05 ***
## albumin        8.4423317  1.6445753   5.133 5.17e-07 ***
## copper        -0.0313822  0.0080651  -3.891 0.000123 ***
## alk_phos       0.0013508  0.0002909   4.643 5.16e-06 ***
## drugPlacebo    0.2450471  1.2119674   0.202 0.839908
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.41 on 295 degrees of freedom
## Multiple R-squared:  0.4375, Adjusted R-squared:  0.4222
## F-statistic: 28.68 on 8 and 295 DF,  p-value: < 2.2e-16
```

Now putting a summary of our best prediction, I have added the drugs since it was not included as the outcome, but since it's our primary interest of the project, I've added them for building the regression. To further elaborate, even our goal in this variable section is to take the best features and make the best model, and it has given the variables to use, but adding drug shouldn't have an major impact, since randomized coefficience on other varaible, it will not be correlated with other variables. The r-squared seems to be lower than the initial data, but I think this it's due to excluding the variables that are higher than level 2. Even though the number of variables are different,looking at the adjusted R-squared, we can still find that it's higher than the initial data.

Regression Diagnostics

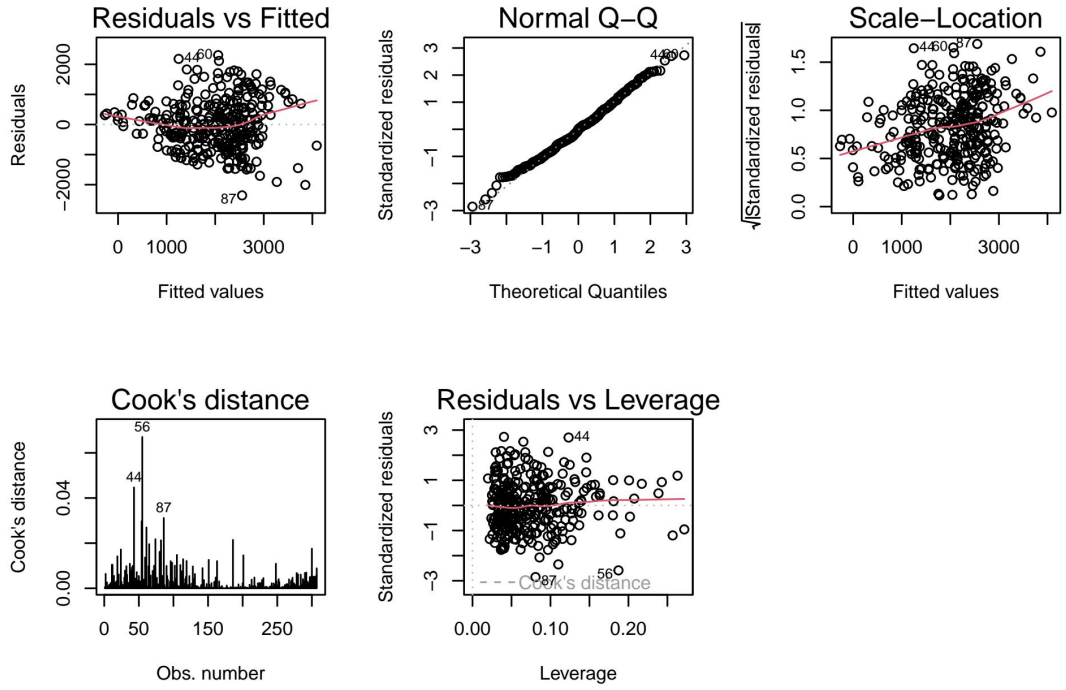
```
data_final <- AIC_data
par(mfrow=c(2,3))
plot(data_final, which = c(1:5))
```

Question assigned to the following page: 5




```
par(mfrow=c(2,3))
plot(initial_model, which = c(1:5))
```

Questions assigned to the following page: 5 and 6



> These are the plots of the final model, and the initial model in order respectively. Compared to the inital model, we can say that the final model has improved significantly in terms of plots. For Residuals vs Fitted, even though it shows some decrease and increase pattern, the residual mean is much closer to zero, The Q-Q plot doesn't show much of a difference, yet are still in the range of normality, and can show some skewness, and for the Scale-Location plot, we can say that it got much better based on the fact that it refined the increasing pattern a bit, showing somewhat straight line. It was overall stablizing the variance and leading to more reliable estimates and standard errors .

Conclusion : What conclusion can you draw about the effect of the drug D-penicillamine on the number of days a patient survived?

The p-value of our final model came out to be 0.839908 regarding drugs. Without the sqrt function to the response variable, it still came out to be pretty high, way higher than 0.05. However, the way it doesn't really matter is that even the p-value indicates that there isn't much of the effect of the drug D-penicillamine on the number of days a patient survived, we cannot conclude it since we don't have enough evidence to suggest that particualr drug has a significant impact. This is because the p-value indicates that the difference in D-penicillamine and Placebo could be due to random chance. The reason the drug is randomly assigned is crucial because otherwise, the affect of the variables are interacting with each other.