Preliminary work on the Chirossis Data set:

```
cir = read.csv("C:\\Users\\lazar\\Downloads\\cirrhosis.csv")

dim(cir)
```

```
## [1] 418  20
```

Starting with patients 313 - 418, there is a lack of information on categorical features such as ascites, hepmeg, spiders, and edema. As a result, I will be removing these patients. The reason I am not keeping them is because it will be hard to predict their categorical features, and other mreasurments such as Chol, cu, alkphos, sgot, trig, and plat. There are simply way too many missing patient variables for these

```
cir = cir[-c(313:418),]
```

Now we must convert all of these categorical variables. First, I would like to see what unique categories there are for the categorical variables

```
unique(cir$rx)
```

```
## [1] "D-penicillamine" "placebo"
```

```
unique(cir$sex)
```

```
## [1] "female" "male"
```

```
unique(cir$ascites)
```

```
## [1] "yes" "no"
```

```
unique(cir$hepmeg)
```

```
## [1] "yes" "no"
```

```
unique(cir$spiders)
```

```
## [1] "yes" "no"
```

```
unique(cir$edema)
```

```
## [1] "edema_despite_diuretic_therapy"
## [2] "no_edema_and_no_diuretic_therapy_for_edema"
## [3] "edema_present_without_diuretics_or_edema_resolved_by_diuretics"
```

Lets one hot encode the variables

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## 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
```

```
vars_to_encode <- c("rx", "sex", "ascites", "hepmeg", "spiders", "edema")

# Create a new data frame with the one hot encoded variables
one_hot_encoded <- cir %>%
  select(all_of(vars_to_encode)) %>%
  mutate_all(as.factor) %>%
  model.matrix(~.-1, data = .)

# Give meaningful column names to the one hot encoded variables
colnames(one_hot_encoded) <- gsub(".*\\.", "", colnames(one_hot_encoded))

# Combine the original data frame with the one hot encoded variables
cir_encoded <- cbind(cir %>% select(-all_of(vars_to_encode)), one_hot_encoded)
```

We will remove rxplacebo to remove colliniarities

```
cir_encoded <- cir_encoded[, -which(colnames(cir_encoded) == "rxplacebo")]
```

Transforming values from ? to NA

```
cir_encoded$chol <- as.integer(ifelse(cir_encoded$chol == "?", NA, cir_encoded$chol))
cir_encoded$cu <- as.integer(ifelse(cir_encoded$cu == "?", NA, cir_encoded$cu))
cir_encoded$alkphos <- as.integer(ifelse(cir_encoded$alkphos == "?", NA, cir_encoded$alkphos))
cir_encoded$sgot <- as.integer(ifelse(cir_encoded$sgot == "?", NA, cir_encoded$sgot))
cir_encoded$trig <- as.integer(ifelse(cir_encoded$trig == "?", NA, cir_encoded$trig))
cir_encoded$plat <- as.integer(ifelse(cir_encoded$plat == "?", NA, cir_encoded$plat))
cir_encoded$ptt <- as.integer(ifelse(cir_encoded$ptt == "?", NA, cir_encoded$ptt))
cir_encoded$stage <- as.integer(ifelse(cir_encoded$stage == "?", NA, cir_encoded$stage))
```

```
summary(cir_encoded)
```

```
##        id              time          event             age
##  Min.   :  1.00   Min.   :  41   Min.   :0.0000   Min.   : 9598
##  1st Qu.: 78.75   1st Qu.:1191   1st Qu.:0.0000   1st Qu.:15428
##  Median :156.50   Median :1840   Median :0.0000   Median :18188
##  Mean   :156.50   Mean   :2006   Mean   :0.8622   Mean   :18269
##  3rd Qu.:234.25   3rd Qu.:2697   3rd Qu.:2.0000   3rd Qu.:20715
##  Max.   :312.00   Max.   :4556   Max.   :2.0000   Max.   :28650
##
##       bili             chol             alb              cu
##  Min.   : 0.300   Min.   : 120.0   Min.   :1.96   Min.   :  4.00
##  1st Qu.: 0.800   1st Qu.: 249.5   1st Qu.:3.31   1st Qu.: 41.25
##  Median : 1.350   Median : 309.5   Median :3.55   Median : 73.00
##  Mean   : 3.256   Mean   : 369.5   Mean   :3.52   Mean   : 97.65
##  3rd Qu.: 3.425   3rd Qu.: 400.0   3rd Qu.:3.80   3rd Qu.:123.00
##  Max.   :28.000   Max.   :1775.0   Max.   :4.64   Max.   :588.00
##                   NA's   :28                      NA's   :2
##     alkphos             sgot            trig             plat
##  Min.   :  289.0   Min.   : 26.0   Min.   : 33.00   Min.   : 62.0
##  1st Qu.:  871.5   1st Qu.: 80.0   1st Qu.: 84.25   1st Qu.:199.8
##  Median : 1259.0   Median :114.0   Median :108.00   Median :257.0
##  Mean   : 1982.6   Mean   :122.2   Mean   :124.70   Mean   :261.9
##  3rd Qu.: 1980.0   3rd Qu.:151.0   3rd Qu.:151.00   3rd Qu.:322.5
##  Max.   :13862.0   Max.   :457.0   Max.   :598.00   Max.   :563.0
##                                    NA's   :30       NA's   :4
##       ptt           stage       rxD-penicillamine    sexmale
##  Min.   : 9.0   Min.   :1.000   Min.   :0.0000    Min.   :0.0000
##  1st Qu.:10.0   1st Qu.:2.000   1st Qu.:0.0000    1st Qu.:0.0000
##  Median :10.0   Median :3.000   Median :1.0000    Median :0.0000
##  Mean   :10.3   Mean   :3.032   Mean   :0.5064    Mean   :0.1154
##  3rd Qu.:11.0   3rd Qu.:4.000   3rd Qu.:1.0000    3rd Qu.:0.0000
##  Max.   :17.0   Max.   :4.000   Max.   :1.0000    Max.   :1.0000
##
##    ascitesyes        hepmegyes        spidersyes
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.07692   Mean   :0.5128   Mean   :0.2885
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##
##  edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.09295
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
##  edemano_edema_and_no_diuretic_therapy_for_edema
##  Min.   :0.0000
##  1st Qu.:1.0000
##  Median :1.0000
```

```
##  Mean   :0.8429
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

Usually in SAS, variables listed as …rxD-peni… mess up the reading of the data, thus I will change it to an underscore

```
names(cir_encoded)[names(cir_encoded) == "rxD-penicillamine"] <- "rxD_penicillamine"
```

event: censored = 0 liver transplant = 1 dead = 2

We will remove our liver transplant patients.

In addition, I will do the following censored = 0 dead = 1

```
# Drop value 1
cir_encoded <- cir_encoded[cir_encoded$event != 1,]

# Replace values 0 and 2
cir_encoded$event <- ifelse(cir_encoded$event == 0, 1, 0)
```

Hailey wanted this variable to be added back.

Basically, if edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics & edemano_edema_and_no_diuretic_therapy_for_edema 0, then the new variable edema_despite_diuretic_therapy will be equal to 1.

```
library(dplyr)

cir_encoded <- cir_encoded %>%
  mutate(edema_despite_diuretic_therapy = if_else(edemaedema_present_without_diuretics_or_edema_
resolved_by_diuretics == 0 & edemano_edema_and_no_diuretic_therapy_for_edema == 0, 1, 0))
```

Now, I will factor the variable.

```
cir_encoded$rxD_penicillamine = as.factor(cir_encoded$rxD_penicillamine)
cir_encoded$sexmale = as.factor(cir_encoded$sexmale)
cir_encoded$ascitesyes = as.factor(cir_encoded$ascitesyes)
cir_encoded$hepmegyes = as.factor(cir_encoded$hepmegyes)
cir_encoded$spidersyes = as.factor(cir_encoded$spidersyes)
cir_encoded$edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics = as.factor(cir_
encoded$edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics)
cir_encoded$edemano_edema_and_no_diuretic_therapy_for_edema = as.factor(cir_encoded$edemano_edem
a_and_no_diuretic_therapy_for_edema)
```

```
sum(is.na(cir_encoded))
```

```
## [1] 62
```

We have a total of 62 mising values. Instead of dropping them, we will impute them by using a random forest algorithm.

I want to see what the distribution looks like before we impute the data

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
# Set the figure size and resolution
options(repr.plot.width=10, repr.plot.height=8, repr.plot.res=300)

# Define the list of excluded variables
excluded_vars <- c("rxD_penicillamine", "sexmale", "ascitesyes", "hepmegyes", "spidersyes",
                   "edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics",
                   "edemano_edema_and_no_diuretic_therapy_for_edema")

# Create a list of data frames, each containing one variable and its name, excluding the exclude
d variables
data_list <- lapply(names(cir_encoded[!(names(cir_encoded) %in% excluded_vars)]),
                    function(x) data.frame(variable = x, value = cir_encoded[,x]))

# Create a list of ggplot objects, one for each variable
plot_list <- lapply(data_list, function(x) ggplot(x, aes(x = value, fill = variable)) +
                    geom_density(alpha = 0.5) +
                    ggtitle(x$variable) +
                    theme(plot.title = element_text(hjust = 0.5)))

# Combine the ggplot objects into a single plot using the grid.arrange function from the gridExt
ra package
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

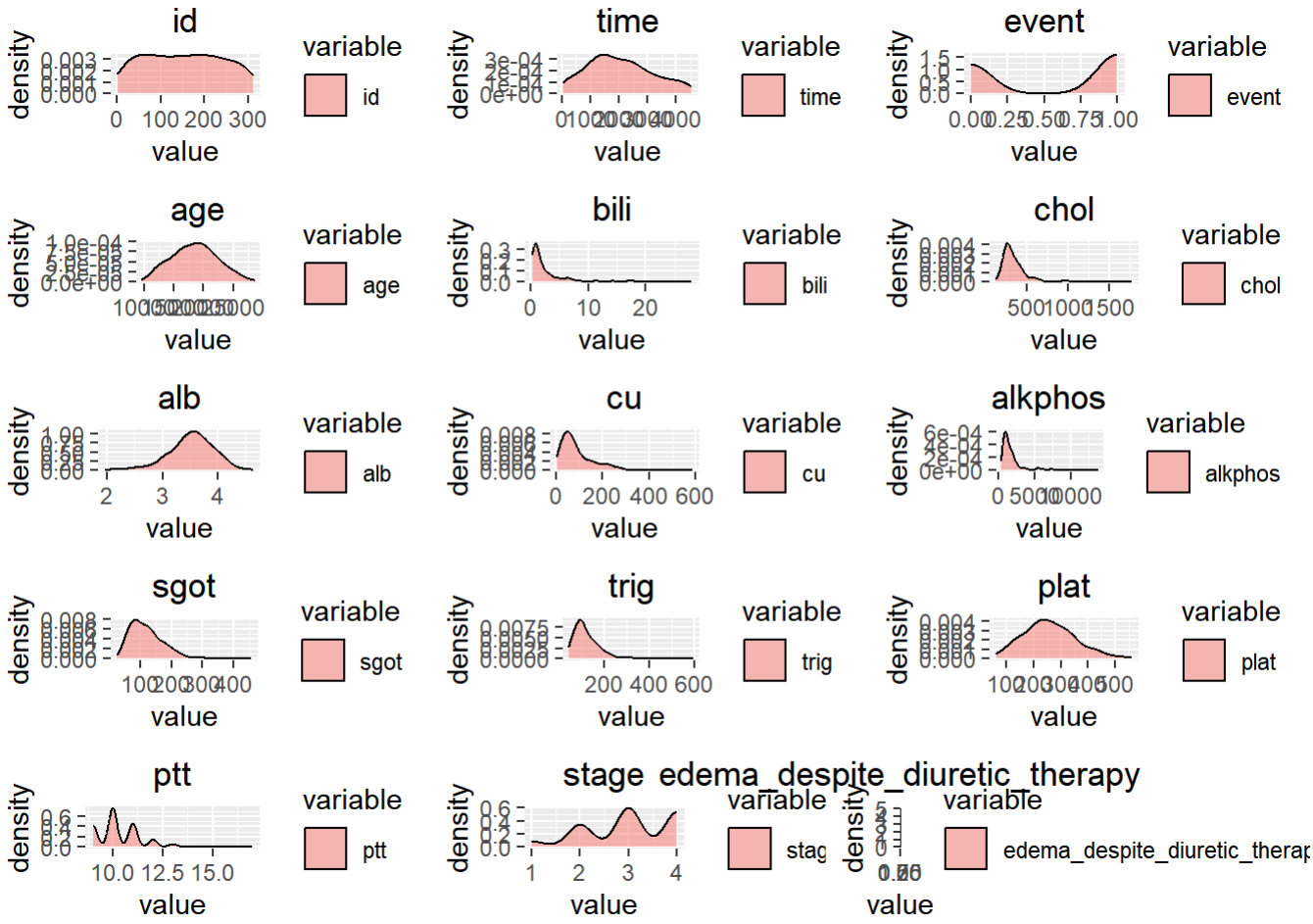
```
grid.arrange(grobs = plot_list, ncol = 3)
```

```
## Warning: Removed 27 rows containing non-finite values (`stat_density()`).
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_density()`).
```

```
## Warning: Removed 29 rows containing non-finite values (`stat_density()`).
```

```
## Warning: Removed 4 rows containing non-finite values (`stat_density()`).
```



We will now use the missForest algorithm to imputate the data

```
library(missForest)
```

```
## Warning: package 'missForest' was built under R version 4.2.3
```

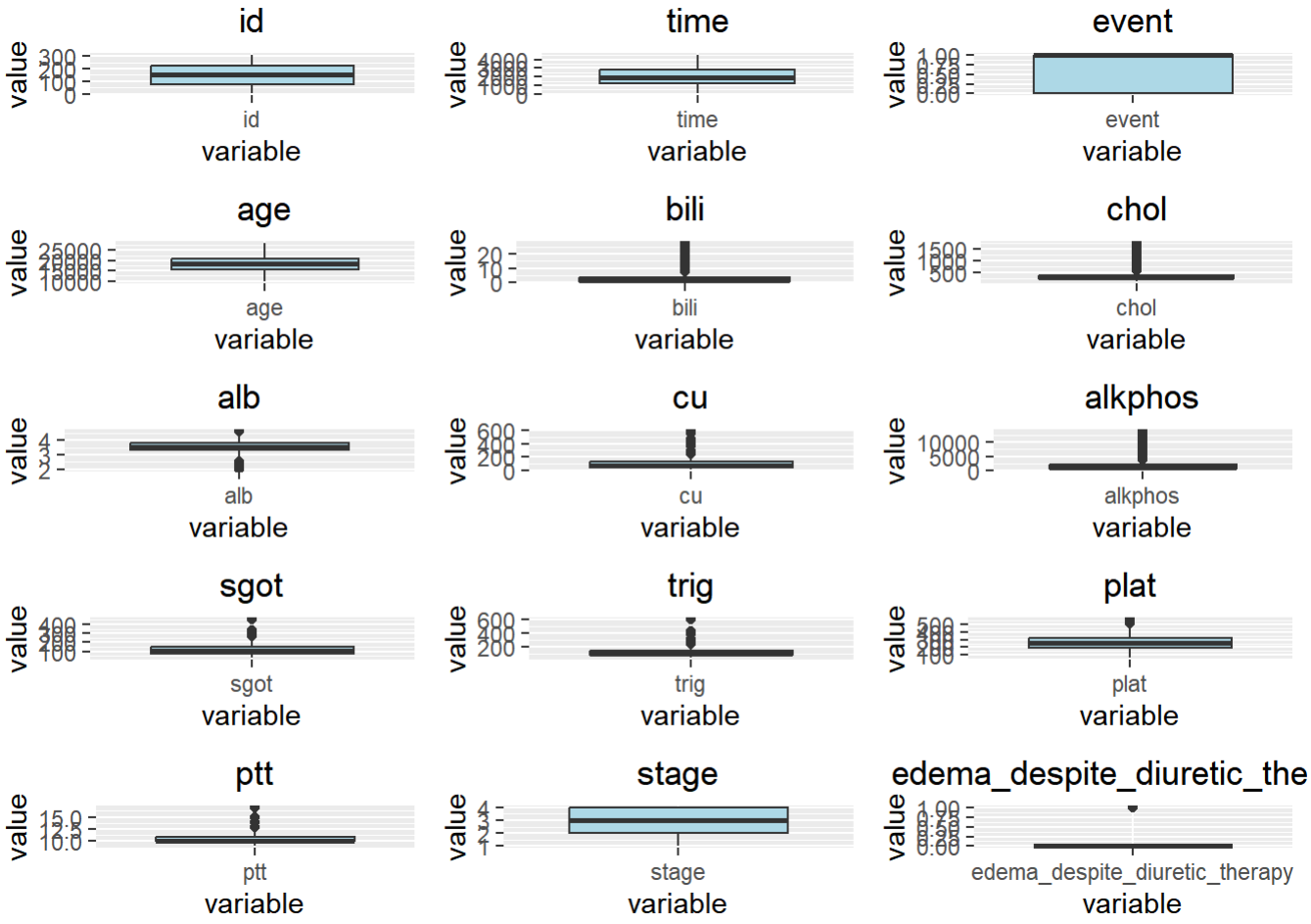
```
imputed_data = missForest(cir_encoded, maxiter = 10)


imp_randomForrest = imputed_data$ximp
```

```
summary(imp_randomForrest)
```

```
##        id              time          event              age
##  Min.   :  1.0   Min.   :  41   Min.   :0.0000   Min.   : 9598
##  1st Qu.: 75.0   1st Qu.:1216   1st Qu.:0.0000   1st Qu.:15694
##  Median :152.0   Median :1882   Median :1.0000   Median :18460
##  Mean   :152.9   Mean   :2039   Mean   :0.5734   Mean   :18479
##  3rd Qu.:227.0   3rd Qu.:2772   3rd Qu.:1.0000   3rd Qu.:20891
##  Max.   :312.0   Max.   :4556   Max.   :1.0000   Max.   :28650
##       bili             chol             alb              cu
##  Min.   : 0.300   Min.   : 120.0   Min.   :1.960   Min.   :  4.00
##  1st Qu.: 0.800   1st Qu.: 252.0   1st Qu.:3.310   1st Qu.: 41.00
##  Median : 1.300   Median : 309.0   Median :3.550   Median : 70.00
##  Mean   : 3.264   Mean   : 361.8   Mean   :3.517   Mean   : 95.86
##  3rd Qu.: 3.400   3rd Qu.: 395.0   3rd Qu.:3.800   3rd Qu.:123.00
##  Max.   :28.000   Max.   :1775.0   Max.   :4.640   Max.   :588.00
##     alkphos            sgot             trig             plat
##  Min.   :  289   Min.   : 26.0   Min.   : 44.0   Min.   : 62.0
##  1st Qu.:  858   1st Qu.: 79.0   1st Qu.: 87.0   1st Qu.:198.0
##  Median : 1258   Median :111.0   Median :111.0   Median :255.0
##  Mean   : 2012   Mean   :121.7   Mean   :123.8   Mean   :259.8
##  3rd Qu.: 2009   3rd Qu.:151.0   3rd Qu.:146.0   3rd Qu.:322.0
##  Max.   :13862   Max.   :457.0   Max.   :598.0   Max.   :563.0
##       ptt            stage       rxD_penicillamine sexmale ascitesyes hepmegyes
##  Min.   : 9.00   Min.   :1.000   0:145             0:260   0:269      0:145
##  1st Qu.:10.00   1st Qu.:2.000   1:148             1: 33   1: 24      1:148
##  Median :10.00   Median :3.000
##  Mean   :10.32   Mean   :3.017
##  3rd Qu.:11.00   3rd Qu.:4.000
##  Max.   :17.00   Max.   :4.000
##  spidersyes edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics
##  0:208      0:266
##  1: 85      1: 27
##
##
##
##
##  edemano_edema_and_no_diuretic_therapy_for_edema edema_despite_diuretic_therapy
##  0: 47                                           Min.   :0.00000
##  1:246                                           1st Qu.:0.00000
##                                                  Median :0.00000
##                                                  Mean   :0.06826
##                                                  3rd Qu.:0.00000
##                                                  Max.   :1.00000
```

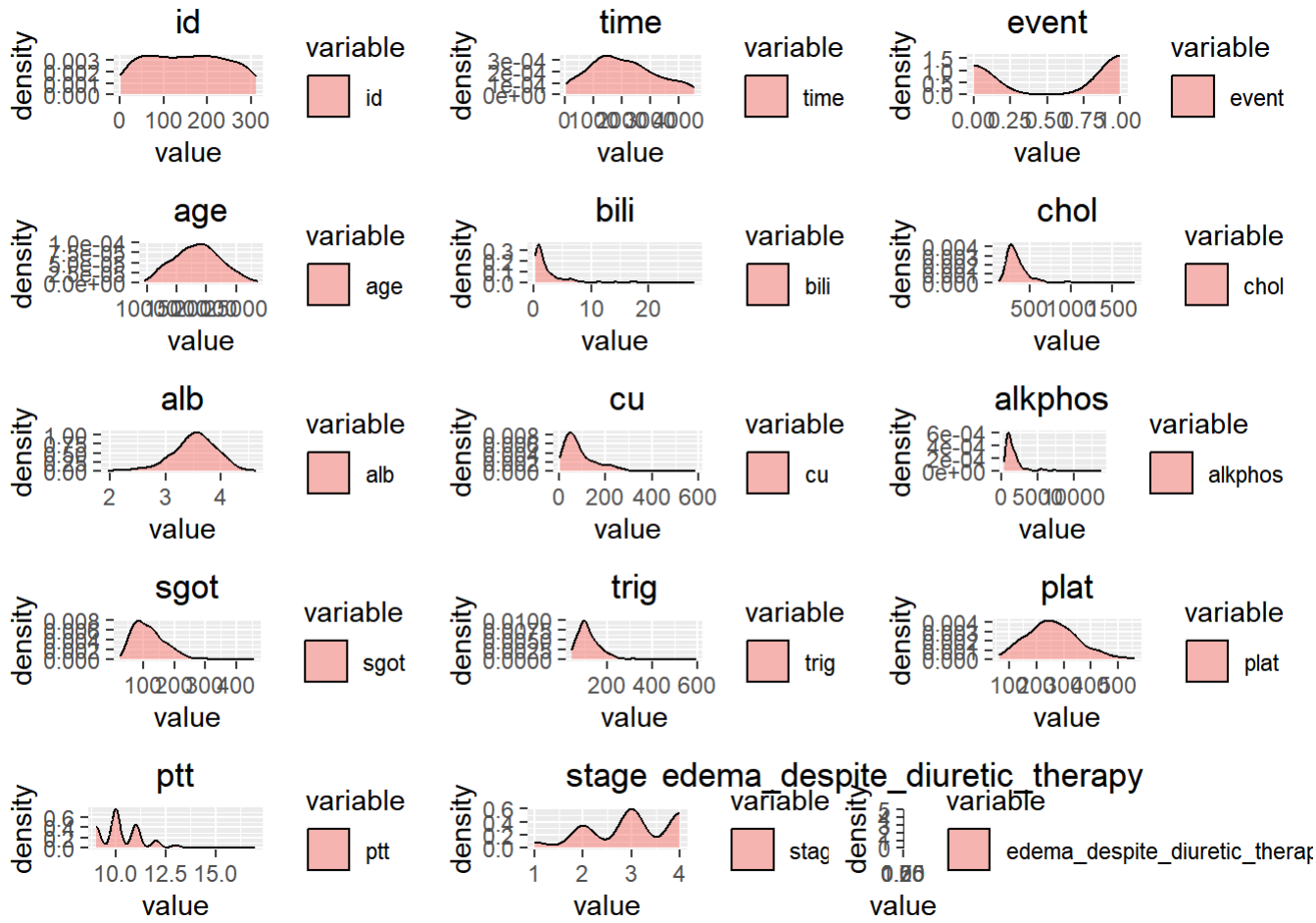
Box plot version of our pdf

```
library(ggplot2)

# Set the figure size and resolution
options(repr.plot.width=10, repr.plot.height=8, repr.plot.res=300)

# Define the list of excluded variables
excluded_vars <- c("rxD_penicillamine", "sexmale", "ascitesyes", "hepmegyes", "spidersyes",
                   "edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics",
                   "edemano_edema_and_no_diuretic_therapy_for_edema")

# Create a list of data frames, each containing one variable and its name, excluding the exclude
d variables
data_list <- lapply(names(imp_randomForrest[!(names(imp_randomForrest) %in% excluded_vars)]),
                    function(x) data.frame(variable = x, value = imp_randomForrest[,x]))

# Create a list of ggplot objects, one for each variable
plot_list <- lapply(data_list, function(x) ggplot(x, aes(x = variable, y = value)) +
                    geom_boxplot(fill = "lightblue") +
                    ggtitle(x$variable) +
                    theme(plot.title = element_text(hjust = 0.5)))

# Combine the ggplot objects into a single plot using the grid.arrange function from the gridExt
ra package
library(gridExtra)
grid.arrange(grobs = plot_list, ncol = 3)
```



pdf of the imputed data

```
library(ggplot2)

# Set the figure size and resolution
options(repr.plot.width=10, repr.plot.height=8, repr.plot.res=300)

# Define the list of excluded variables
excluded_vars <- c("rxD_penicillamine", "sexmale", "ascitesyes", "hepmegyes", "spidersyes",
                   "edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics",
                   "edemano_edema_and_no_diuretic_therapy_for_edema")

# Create a list of data frames, each containing one variable and its name, excluding the exclude
d variables
data_list <- lapply(names(imp_randomForrest[!(names(imp_randomForrest) %in% excluded_vars)]),
                    function(x) data.frame(variable = x, value = imp_randomForrest[,x]))

# Create a list of ggplot objects, one for each variable
plot_list <- lapply(data_list, function(x) ggplot(x, aes(x = value, fill = variable)) +
                    geom_density(alpha = 0.5) +
                    ggtitle(x$variable) +
                    theme(plot.title = element_text(hjust = 0.5)))

# Combine the ggplot objects into a single plot using the grid.arrange function from the gridExt
ra package
library(gridExtra)
grid.arrange(grobs = plot_list, ncol = 3)
```



now, I want to superimpose the graphs together to show before and after

```
library(ggplot2)
library(gridExtra)

# Set the figure size and resolution
options(repr.plot.width=10, repr.plot.height=8, repr.plot.res=300)

# Define the list of excluded variables
excluded_vars <- c("rxD_penicillamine", "sexmale", "ascitesyes", "hepmegyes", "spidersyes",
                   "edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics",
                   "edemano_edema_and_no_diuretic_therapy_for_edema")

# Create a list of data frames, each containing one variable and its name, excluding the exclude
d variables
data_list_1 <- lapply(names(cir_encoded[!(names(cir_encoded) %in% excluded_vars)]),
                      function(x) data.frame(variable = x, value = cir_encoded[,x], dataset = "o
riginal data set"))

data_list_2 <- lapply(names(imp_randomForrest[!(names(imp_randomForrest) %in% excluded_vars)]),
                      function(x) data.frame(variable = x, value = imp_randomForrest[,x], datase
t = "imputed random forest"))

# Combine the data frames
combined_data_list <- mapply(rbind, data_list_1, data_list_2, SIMPLIFY = FALSE)

# Create a list of ggplot objects, one for each variable
plot_list <- lapply(combined_data_list, function(x) ggplot(x, aes(x = value, fill = dataset, col
or = dataset)) +
                    geom_density(alpha = 0.5) +
                    ggtitle(x$variable[1]) +
                    theme(plot.title = element_text(hjust = 0.5)))

# Combine the ggplot objects into a single plot using the grid.arrange function from the gridExt
ra package
grid.arrange(grobs = plot_list, ncol = 3)
```

```
## Warning: Removed 27 rows containing non-finite values (`stat_density()`).
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_density()`).
```

```
## Warning: Removed 29 rows containing non-finite values (`stat_density()`).
```

```
## Warning: Removed 4 rows containing non-finite values (`stat_density()`).
```

looking at placebo vs rxD

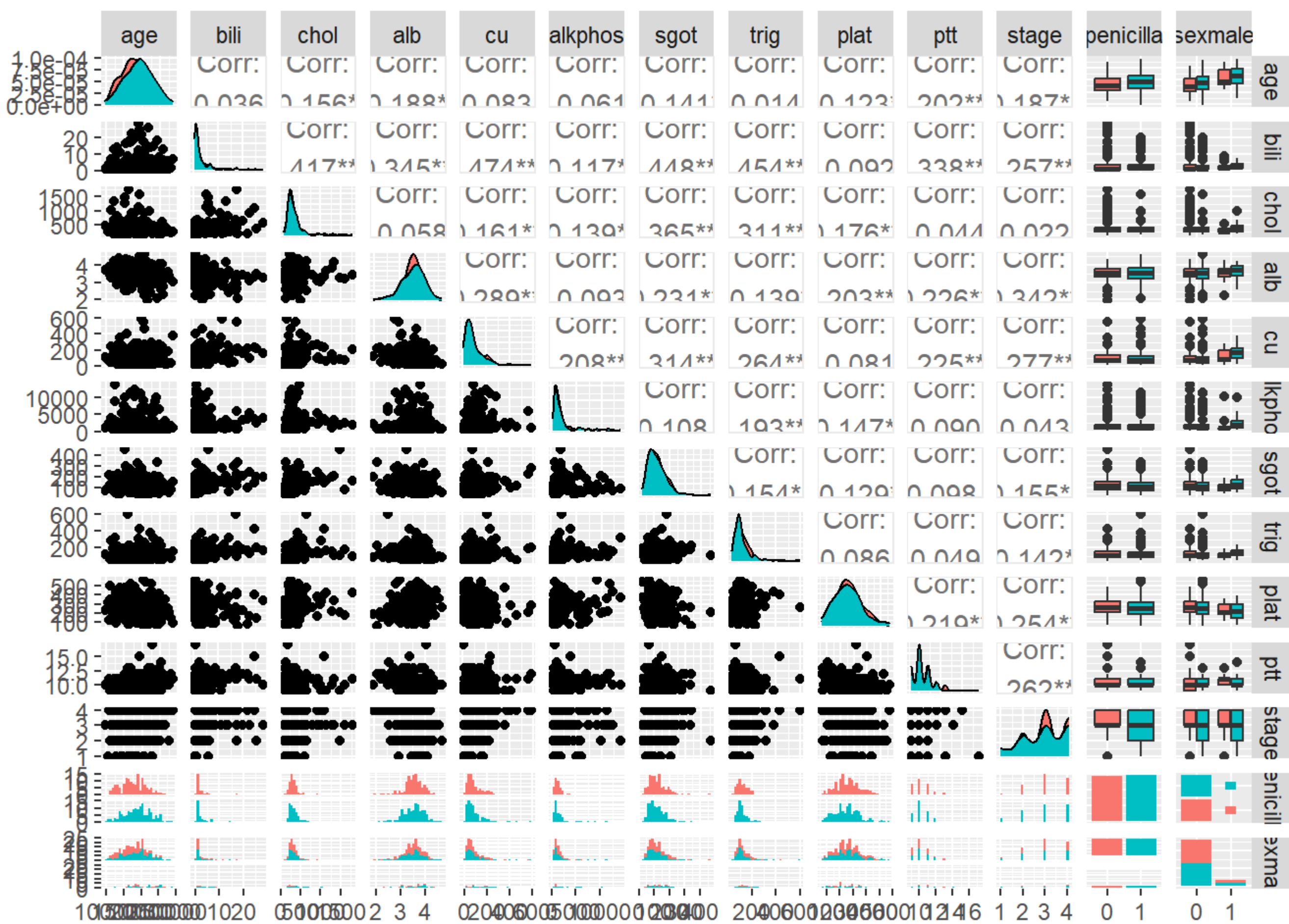
```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
imp_randomForrest %>%
  select(-id) %>%
  select(age, bili, chol, alb, cu, alkphos, sgot, trig, plat, ptt, stage, rxD_penicillamine, sex
male) %>%
  ggpairs(aes(fill = rxD_penicillamine))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

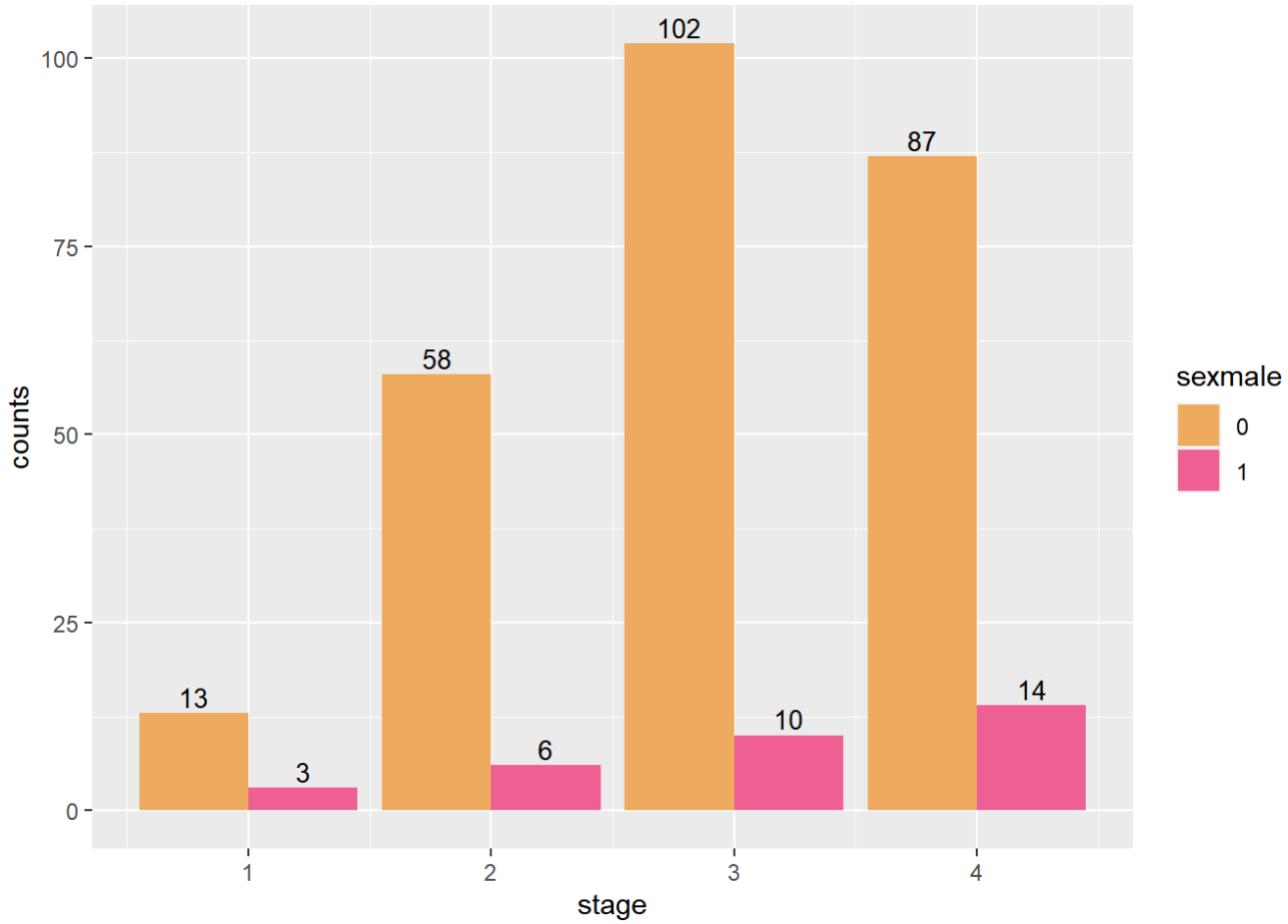


stage graph

```
df <- imp_randomForrest %>%
  group_by(sexmale, stage) %>%
  summarise(counts = n())
```

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

```
ggplot(df, aes(x = stage, y = counts)) +
  geom_bar(aes(fill = sexmale), stat = "identity", position = "dodge") +
  geom_text(aes(label = counts, group = sexmale), position = position_dodge(0.9), vjust = -.3, s
ize = 3.5) +
  scale_fill_manual(values = c("#EEAB5F", "#EE5F93"))
```



Survival Analysis

```
library(survival)
```

```
# Fit the Cox proportional hazards model with stratification by stage
fit <- coxph(Surv(age, event) ~ alb + alkphos + ascitesyes + bili + chol + cu + hepmegyes + plat
+ ptt + rxD_penicillamine +
               sexmale + sgot + spidersyes + strata(stage) + trig + edemaedema_present_without_
diuretics_or_edema_resolved_by_diuretics + edemano_edema_and_no_diuretic_therapy_for_edema + ede
ma_despite_diuretic_therapy, data = imp_randomForrest)
```

```
## Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
## Loglik converged before variable 3 ; coefficient may be infinite.
```

```
# Print the model summary
summary(fit)
```

```
## Call:
## coxph(formula = Surv(age, event) ~ alb + alkphos + ascitesyes +
##     bili + chol + cu + hepmegyes + plat + ptt + rxD_penicillamine +
##     sexmale + sgot + spidersyes + strata(stage) + trig + edemaedema_present_without_diuretics
_or_edema_resolved_by_diuretics +
##     edemano_edema_and_no_diuretic_therapy_for_edema + edema_despite_diuretic_therapy,
##     data = imp_randomForrest)
##
##   n= 293, number of events= 168
##
##                                                                        coef
## alb                                                                4.008e-01
## alkphos                                                           -4.272e-05
## ascitesyes1                                                       -1.810e+01
## bili                                                              -1.025e-01
## chol                                                              -6.488e-05
## cu                                                                -1.767e-03
## hepmegyes1                                                        -2.782e-01
## plat                                                               6.197e-05
## ptt                                                               -3.997e-01
## rxD_penicillamine1                                                -4.298e-01
## sexmale1                                                          -7.539e-01
## sgot                                                               3.445e-03
## spidersyes1                                                        3.706e-01
## trig                                                              -1.541e-03
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1 -1.315e+00
## edemano_edema_and_no_diuretic_therapy_for_edema1                  -4.689e-01
## edema_despite_diuretic_therapy                                            NA
##                                                                   exp(coef)
## alb                                                               1.493e+00
## alkphos                                                           1.000e+00
## ascitesyes1                                                       1.376e-08
## bili                                                              9.025e-01
## chol                                                              9.999e-01
## cu                                                                9.982e-01
## hepmegyes1                                                        7.572e-01
## plat                                                              1.000e+00
## ptt                                                               6.705e-01
## rxD_penicillamine1                                                6.507e-01
## sexmale1                                                          4.705e-01
## sgot                                                              1.003e+00
## spidersyes1                                                       1.449e+00
## trig                                                              9.985e-01
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1  2.686e-01
## edemano_edema_and_no_diuretic_therapy_for_edema1                  6.257e-01
## edema_despite_diuretic_therapy                                           NA
##                                                                    se(coef)
## alb                                                               2.497e-01
## alkphos                                                           5.246e-05
## ascitesyes1                                                       2.395e+03
## bili                                                              6.380e-02
## chol                                                              6.708e-04
```

```
## cu                                                                 1.839e-03
## hepmegyes1                                                         1.961e-01
## plat                                                               1.072e-03
## ptt                                                                1.035e-01
## rxD_penicillamine1                                                 1.698e-01
## sexmale1                                                           3.781e-01
## sgot                                                               1.785e-03
## spidersyes1                                                        2.237e-01
## trig                                                               1.828e-03
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1  1.146e+00
## edemano_edema_and_no_diuretic_therapy_for_edema1                   1.090e+00
## edema_despite_diuretic_therapy                                     0.000e+00
##                                                                        z
## alb                                                                1.605
## alkphos                                                           -0.814
## ascitesyes1                                                       -0.008
## bili                                                              -1.607
## chol                                                              -0.097
## cu                                                                -0.961
## hepmegyes1                                                        -1.419
## plat                                                               0.058
## ptt                                                               -3.862
## rxD_penicillamine1                                                -2.531
## sexmale1                                                          -1.994
## sgot                                                               1.930
## spidersyes1                                                        1.657
## trig                                                              -0.843
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1 -1.147
## edemano_edema_and_no_diuretic_therapy_for_edema1                  -0.430
## edema_despite_diuretic_therapy                                       NA
##                                                                  Pr(>|z|)
## alb                                                              0.108457
## alkphos                                                          0.415433
## ascitesyes1                                                      0.993969
## bili                                                             0.107972
## chol                                                             0.922950
## cu                                                               0.336656
## hepmegyes1                                                       0.156038
## plat                                                             0.953922
## ptt                                                              0.000113
## rxD_penicillamine1                                               0.011372
## sexmale1                                                         0.046123
## sgot                                                             0.053645
## spidersyes1                                                      0.097591
## trig                                                             0.399037
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1 0.251387
## edemano_edema_and_no_diuretic_therapy_for_edema1                 0.666983
## edema_despite_diuretic_therapy                                         NA
##
## alb
## alkphos
## ascitesyes1
```

```
## bili
## chol
## cu
## hepmegyes1
## plat
## ptt                                                                    ***
## rxD_penicillamine1                                                     *
## sexmale1                                                               *
## sgot                                                                   .
## spidersyes1                                                            .
## trig
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1
## edemano_edema_and_no_diuretic_therapy_for_edema1
## edema_despite_diuretic_therapy
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                                    exp(coef)
## alb                                                                1.493e+00
## alkphos                                                            1.000e+00
## ascitesyes1                                                        1.376e-08
## bili                                                               9.025e-01
## chol                                                               9.999e-01
## cu                                                                 9.982e-01
## hepmegyes1                                                         7.572e-01
## plat                                                               1.000e+00
## ptt                                                                6.705e-01
## rxD_penicillamine1                                                 6.507e-01
## sexmale1                                                           4.705e-01
## sgot                                                               1.003e+00
## spidersyes1                                                        1.449e+00
## trig                                                               9.985e-01
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1 2.686e-01
## edemano_edema_and_no_diuretic_therapy_for_edema1                   6.257e-01
## edema_despite_diuretic_therapy                                            NA
##                                                                   exp(-coef)
## alb                                                                6.698e-01
## alkphos                                                            1.000e+00
## ascitesyes1                                                        7.268e+07
## bili                                                               1.108e+00
## chol                                                               1.000e+00
## cu                                                                 1.002e+00
## hepmegyes1                                                         1.321e+00
## plat                                                               9.999e-01
## ptt                                                                1.491e+00
## rxD_penicillamine1                                                 1.537e+00
## sexmale1                                                           2.125e+00
## sgot                                                               9.966e-01
## spidersyes1                                                        6.903e-01
## trig                                                               1.002e+00
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1  3.723e+00
## edemano_edema_and_no_diuretic_therapy_for_edema1                   1.598e+00
```

```
## edema_despite_diuretic_therapy                                                NA
##                                                                          lower .95
## alb                                                                        0.91521
## alkphos                                                                    0.99985
## ascitesyes1                                                                0.00000
## bili                                                                       0.79646
## chol                                                                       0.99862
## cu                                                                         0.99464
## hepmegyes1                                                                 0.51553
## plat                                                                       0.99796
## ptt                                                                        0.54737
## rxD_penicillamine1                                                         0.46647
## sexmale1                                                                   0.22427
## sgot                                                                       0.99995
## spidersyes1                                                                0.93440
## trig                                                                       0.99489
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1       0.02841
## edemano_edema_and_no_diuretic_therapy_for_edema1                           0.07393
## edema_despite_diuretic_therapy                                                 NA
##                                                                         upper .95
## alb                                                                        2.4359
## alkphos                                                                    1.0001
## ascitesyes1                                                                   Inf
## bili                                                                       1.0227
## chol                                                                       1.0013
## cu                                                                         1.0018
## hepmegyes1                                                                 1.1120
## plat                                                                       1.0022
## ptt                                                                        0.8213
## rxD_penicillamine1                                                         0.9076
## sexmale1                                                                   0.9871
## sgot                                                                       1.0070
## spidersyes1                                                                2.2457
## trig                                                                       1.0020
## edemaedema_present_without_diuretics_or_edema_resolved_by_diuretics1       2.5390
## edemano_edema_and_no_diuretic_therapy_for_edema1                           5.2956
## edema_despite_diuretic_therapy                                                 NA
##
## Concordance= 0.707  (se = 0.024 )
## Likelihood ratio test= 103.9  on 16 df,   p=6e-15
## Wald test            = 38.86  on 16 df,   p=0.001
## Score (logrank) test = 76.21  on 16 df,   p=8e-10
```

```
# Fit the Cox proportional hazards model with stratification by stage
fit <- coxph(Surv(age, event) ~  cu + hepmegyes + ptt + edemano_edema_and_no_diuretic_therapy_fo
r_edema, data = imp_randomForrest)

# Print the model summary
summary(fit)
```

```
## Call:
## coxph(formula = Surv(age, event) ~ cu + hepmegyes + ptt + edemano_edema_and_no_diuretic_thera
py_for_edema,
##     data = imp_randomForrest)
##
##   n= 293, number of events= 168
##
##                                                     coef exp(coef)  se(coef)
## cu                                                -0.005031  0.994981  0.001460
## hepmegyes1                                        -0.520310  0.594336  0.166832
## ptt                                               -0.389024  0.677718  0.095817
## edemano_edema_and_no_diuretic_therapy_for_edema1  1.262437  3.534025  0.345615
##                                                        z Pr(>|z|)
## cu                                                -3.445 0.000570 ***
## hepmegyes1                                        -3.119 0.001816 **
## ptt                                               -4.060 4.91e-05 ***
## edemano_edema_and_no_diuretic_therapy_for_edema1  3.653 0.000259 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                   exp(coef) exp(-coef) lower .95
## cu                                                   0.9950      1.005    0.9921
## hepmegyes1                                           0.5943      1.683    0.4286
## ptt                                                  0.6777      1.476    0.5617
## edemano_edema_and_no_diuretic_therapy_for_edema1    3.5340      0.283    1.7951
##                                                   upper .95
## cu                                                   0.9978
## hepmegyes1                                           0.8242
## ptt                                                  0.8177
## edemano_edema_and_no_diuretic_therapy_for_edema1    6.9576
##
## Concordance= 0.687  (se = 0.022 )
## Likelihood ratio test= 98.33  on 4 df,   p=<2e-16
## Wald test            = 67.26  on 4 df,   p=9e-14
## Score (logrank) test = 75.61  on 4 df,   p=1e-15
```

```
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 4.2.3
```

```
## Loading required package: ggpubr
```

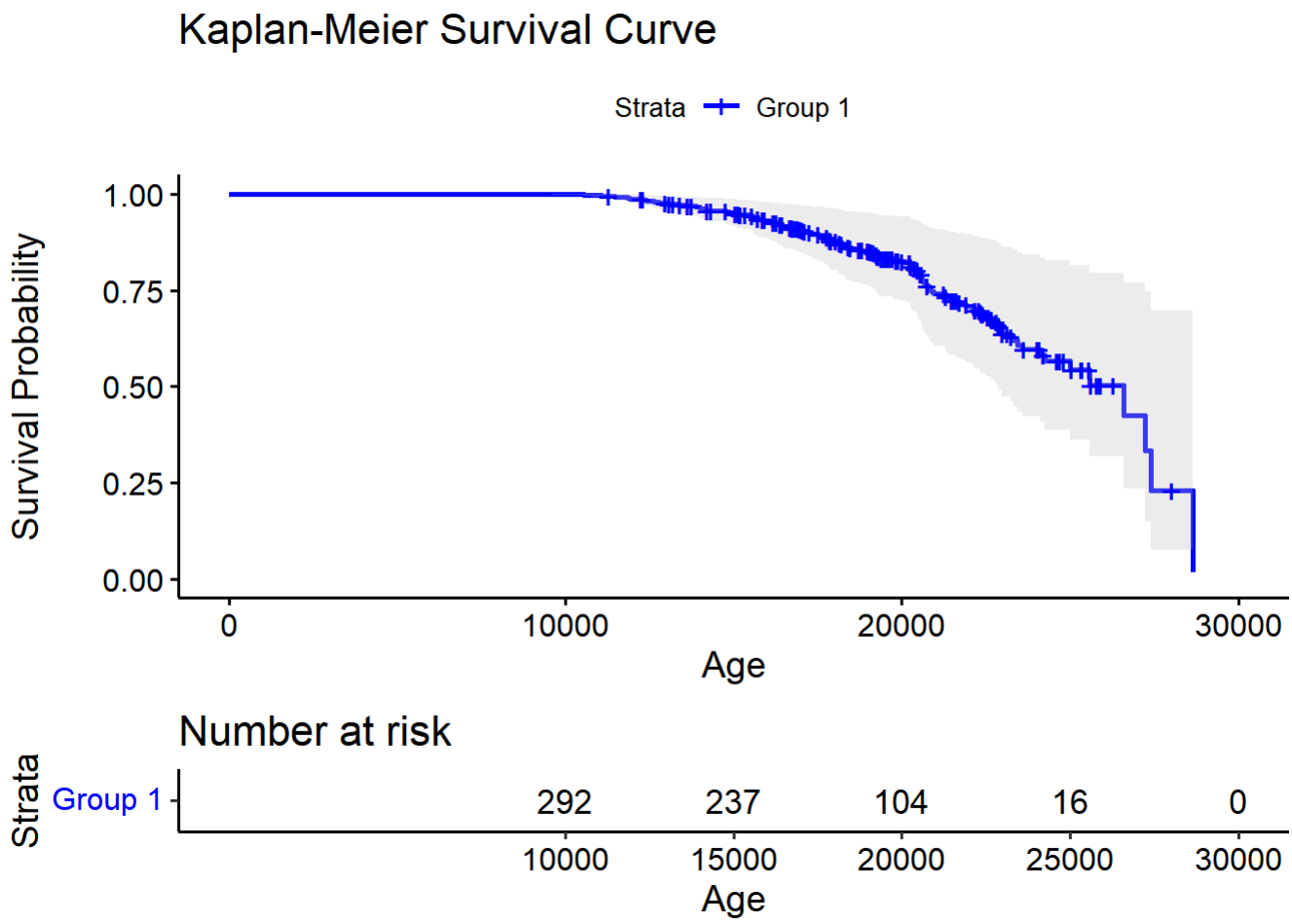
```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

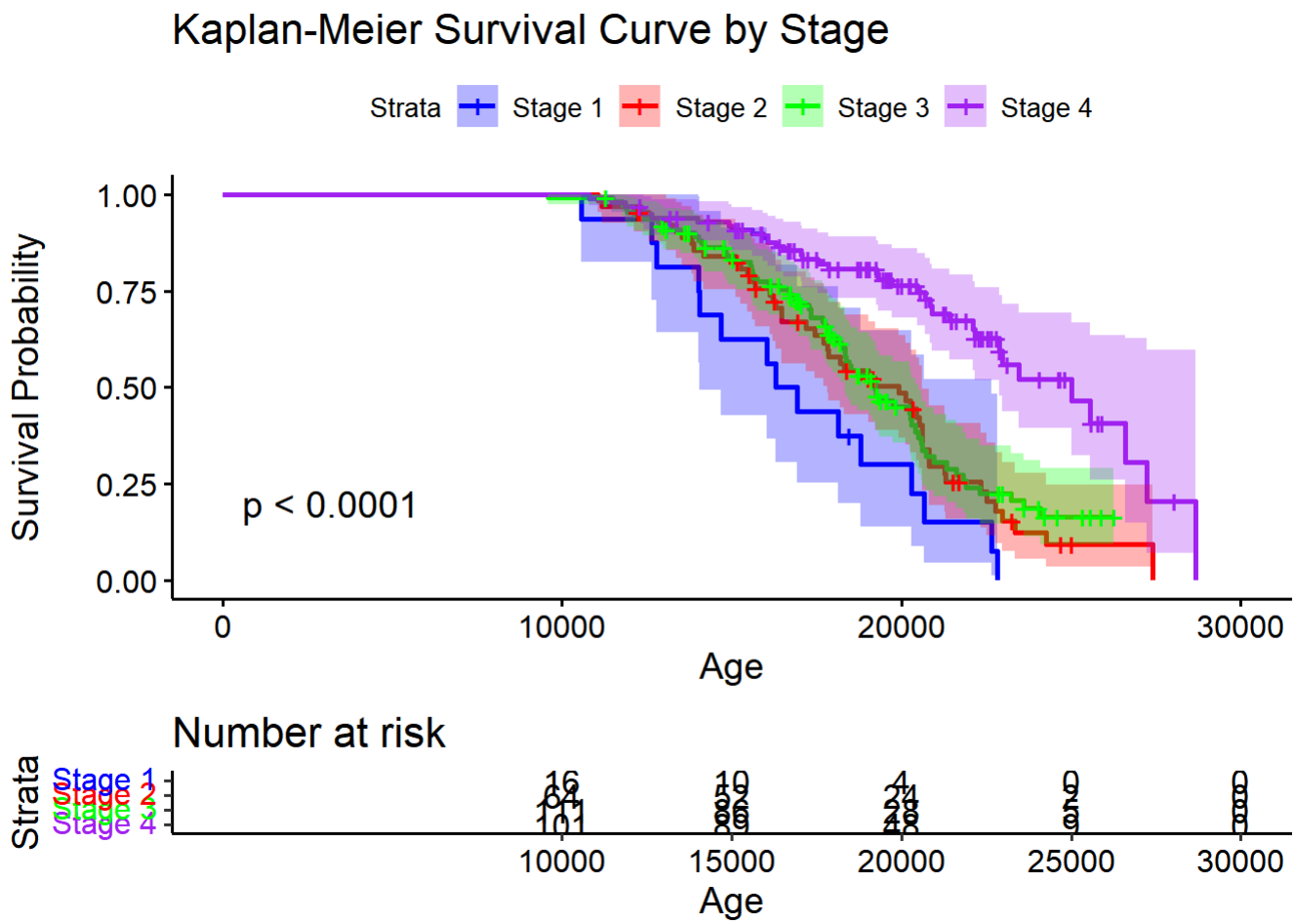
```
# Predict the survival probabilities
predicted_survival <- survfit(fit)

# Plot the survival curves
g <- ggsurvplot(
  predicted_survival,
  data = imp_randomForrest,
  pval = TRUE,           # Add p-value
  risk.table = TRUE,     # Add risk table
  conf.int = TRUE,       # Add confidence intervals
  legend.labs = c("Group 1"), # Change legend labels as per your groupings
  palette = c("blue"), # Change colors as desired
  xlab = "Age", # Customize x-axis label
  ylab = "Survival Probability", # Customize y-axis label
  title = "Kaplan-Meier Survival Curve" # Customize the title
)
```

```
## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.coord = pval.coord, :
There are no survival curves to be compared.
##  This is a null model.
```

```
# Print the plot
g
```




```
# Create separate survival objects for each stage
stage_survival <- survfit(Surv(age, event) ~ stage, data = imp_randomForrest)

# Plot the survival curves
g <- ggsurvplot(
  stage_survival,
  data = imp_randomForrest,
  pval = TRUE,           # Add p-value
  risk.table = TRUE,     # Add risk table
  conf.int = TRUE,       # Add confidence intervals
  legend.labs = c("Stage 1", "Stage 2", "Stage 3", "Stage 4"), # Change legend labels as per you
r groupings
  palette = c("blue", "red", "green", "purple"), # Change colors as desired
  xlab = "Age", # Customize x-axis label
  ylab = "Survival Probability", # Customize y-axis label
  title = "Kaplan-Meier Survival Curve by Stage" # Customize the title
)

# Print the plot
g
```




```
# Perform the log-rank test
log_rank_test <- survdiff(Surv(age, event) ~ stage, data = imp_randomForrest)

# Print the test results
log_rank_test
```

```
## Call:
## survdiff(formula = Surv(age, event) ~ stage, data = imp_randomForrest)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## stage=1  16       15      6.3     12.04     12.57
## stage=2  64       48     34.1      5.62      7.09
## stage=3 112       69     55.1      3.52      5.36
## stage=4 101       36     72.5     18.37     33.35
##
##  Chisq= 40.7  on 3 degrees of freedom, p= 8e-09
```

There are four stages (1, 2, 3, and 4), and their corresponding sample sizes are 16, 64, 112, and 101, respectively. For each stage, the observed and expected events are listed. For example, in stage 1, there were 15 observed events, whereas 6.3 events were expected under the null hypothesis. The test statistic contributions ((O-E)^2/E) and the variance contributions ((O-E)^2/V) are also listed for each stage. These values are used to compute the overall chi-squared statistic and its associated p-value. The overall chi-squared statistic is 40.7 with 3 degrees of freedom. The associated p-value is 8e-09 (which is extremely small). Since the p-value (8e-09) is much smaller

than a typical significance level (e.g., 0.05), you can reject the null hypothesis. This means that there is a statistically significant difference in the survival curves between the different stages. In other words, the survival probabilities are significantly different among the four stages.

```
# Perform pairwise log-rank tests
pairwise_tests <- pairwise_survdiff(Surv(age, event) ~ stage, data = imp_randomForrest, p.adjus
t.method = "bonferroni")

# Print the results
pairwise_tests
```

```
##
##  Pairwise comparisons using Log-Rank test
##
## data:  imp_randomForrest and stage
##
##   1       2       3
## 2 0.38    -       -
## 3 0.17    1.00    -
## 4 7.3e-08 2.8e-06 1.1e-05
##
## P value adjustment method: bonferroni
```

Stage 1 vs. Stage 2: The adjusted p-value is 0.38, which is greater than 0.05 (a typical significance level). Therefore, there is no statistically significant difference in the survival curves between Stage 1 and Stage 2.

Stage 1 vs. Stage 3: The adjusted p-value is 0.17, which is also greater than 0.05. This indicates that there is no statistically significant difference in the survival curves between Stage 1 and Stage 3.

Stage 1 vs. Stage 4: The adjusted p-value is 7.3e-08, which is much smaller than 0.05. This means that there is a statistically significant difference in the survival curves between Stage 1 and Stage 4.

Stage 2 vs. Stage 3: The adjusted p-value is 1.00, indicating no statistically significant difference in the survival curves between Stage 2 and Stage 3.

Stage 2 vs. Stage 4: The adjusted p-value is 2.8e-06, which is smaller than 0.05. This implies that there is a statistically significant difference in the survival curves between Stage 2 and Stage 4.

Stage 3 vs. Stage 4: The adjusted p-value is 1.1e-05, which is also smaller than 0.05. This means that there is a statistically significant difference in the survival curves between Stage 3 and Stage 4.

```
# library(ggridges)
# library(ggplot2)
# library(viridis)
# library(hrbrthemes)
# library(reshape2)
#
# # Select variables to plot
# vars_to_plot <- colnames(imp_randomForrest)[!(colnames(imp_randomForrest) %in% c("rxD-penicill
amine", "sexmale", "ascitesyes", "hepmegyes", "spidersyes", "edemaedema_present_without_diuretic
s_or_edema_resolved_by_diuretics", "edemano_edema_and_no_diuretic_therapy_for_edema"))]
#
# # Create plot
# ggplot(melt(imp_randomForrest[,vars_to_plot]), aes(x=value, y=variable, fill=value)) +
#   geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01) +
#   scale_fill_viridis(name = "Density", option = "C") +
#   labs(title = "Density distributions of variables in imp_randomForest") +
#   theme_minimal()
```

```
#write.csv(imp_randomForrest, file ="C:\\Users\\lazar\\Documents\\Spring2023\\Survival_Analysis
\\cirrhosisCLEANED.csv" )
```