EDA

Chen Liang

2024-11-16

Load data

```
cirrhosis <- read_csv("data/cirrhosis.csv")|>
  janitor::clean_names() |>
  mutate(age = round(age / 365),
         sex = if_else(sex == "M", "Male", "Female"),
         ascites = if_else(ascites == "N", "No", "Yes"),
         hepatomegaly = if_else(hepatomegaly == "N", "No", "Yes"),
         spiders = if_else(spiders == "N", "No", "Yes"),
         edema = if_else(edema == "N", "No", "Yes"))
```

```
## Rows: 418 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): Status, Drug, Sex, Ascites, Hepatomegaly, Spiders, Edema
## dbl (13): ID, N_Days, Age, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Check for missing values
missing_data <- colSums(is.na(cirrhosis))
missing_data
```

```
##            id        n_days        status          drug           age
##             0             0             0           106             0
##           sex       ascites  hepatomegaly       spiders        edema
##             0           106           106           106             0
##     bilirubin   cholesterol       albumin        copper      alk_phos
##             0           134             0           108           106
##          sgot tryglicerides     platelets   prothrombin         stage
##           106           136            11             2             6
```

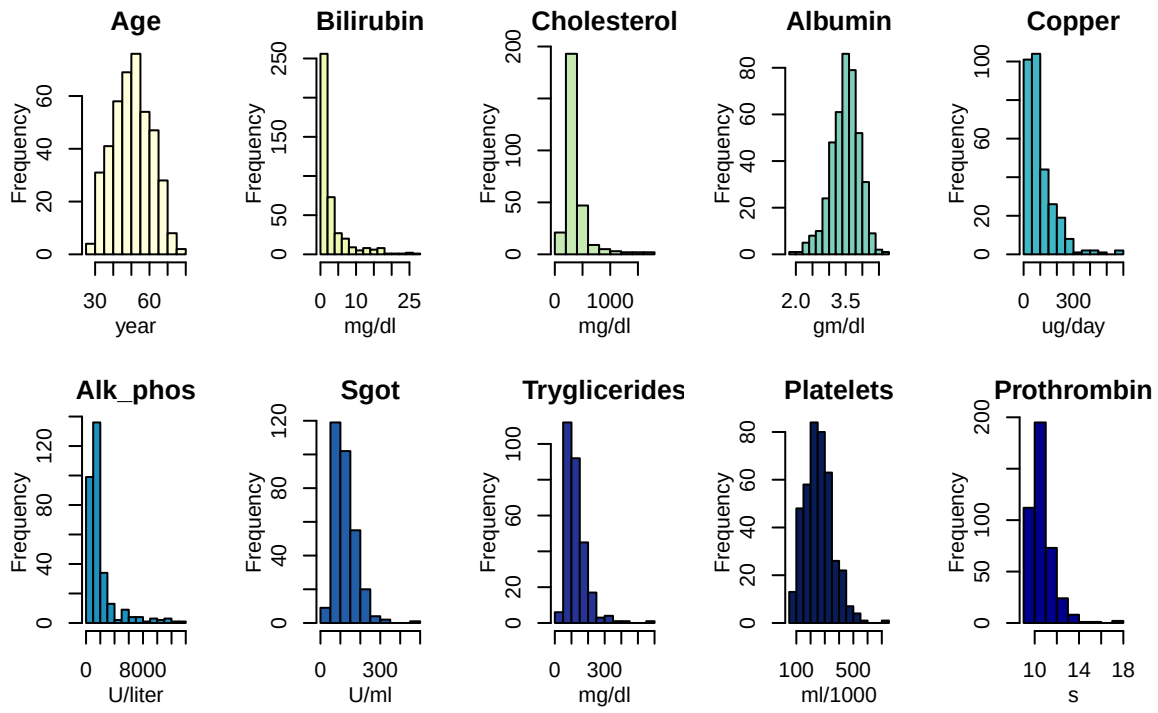
Historgram Plots for continuouse variables

```r
conti_vars = cirrhosis |>
  select(age, bilirubin, cholesterol, albumin, copper, alk_phos, sgot, trygliceride, platelets,prothro

par(mfrow = c(2, 5),  # 2 rows, 5 columns
    oma = c(2, 2, 3, 1),  # Outer margins
    mar = c(4, 4, 2, 1),  # Inner margins for individual plots
    mgp = c(2, 1, 0))     # Margins for axis labels and titles

colors <- c(brewer.pal(9, "YlGnBu"), "darkblue")

# Plot each histogram using a color from the Set3 palette
hist(conti_vars$age, main = "Age", xlab = "year", ylab = "Frequency", col = colors[1])
hist(conti_vars$bilirubin, main = "Bilirubin", xlab = "mg/dl", ylab = "Frequency", col = colors[2])
hist(conti_vars$cholesterol, main = "Cholesterol", xlab = "mg/dl", ylab = "Frequency", col = colors[3])
hist(conti_vars$albumin, main = "Albumin", xlab = "gm/dl", ylab = "Frequency", col = colors[4])
hist(conti_vars$copper, main = "Copper", xlab = "ug/day", ylab = "Frequency", col = colors[5])
hist(conti_vars$alk_phos, main = "Alk_phos", xlab = "U/liter", ylab = "Frequency", col = colors[6])
hist(conti_vars$sgot, main = "Sgot", xlab = "U/ml", ylab = "Frequency", col = colors[7])
hist(conti_vars$trygliceride, main = "Trygliceride", xlab = "mg/dl", ylab = "Frequency", col = colors
hist(conti_vars$platelets, main = "Platelets", xlab = "ml/1000", ylab = "Frequency", col = colors[9])
hist(conti_vars$prothrombin, main = "Prothrombin", xlab = "s", ylab = "Frequency", col = colors[10])
```

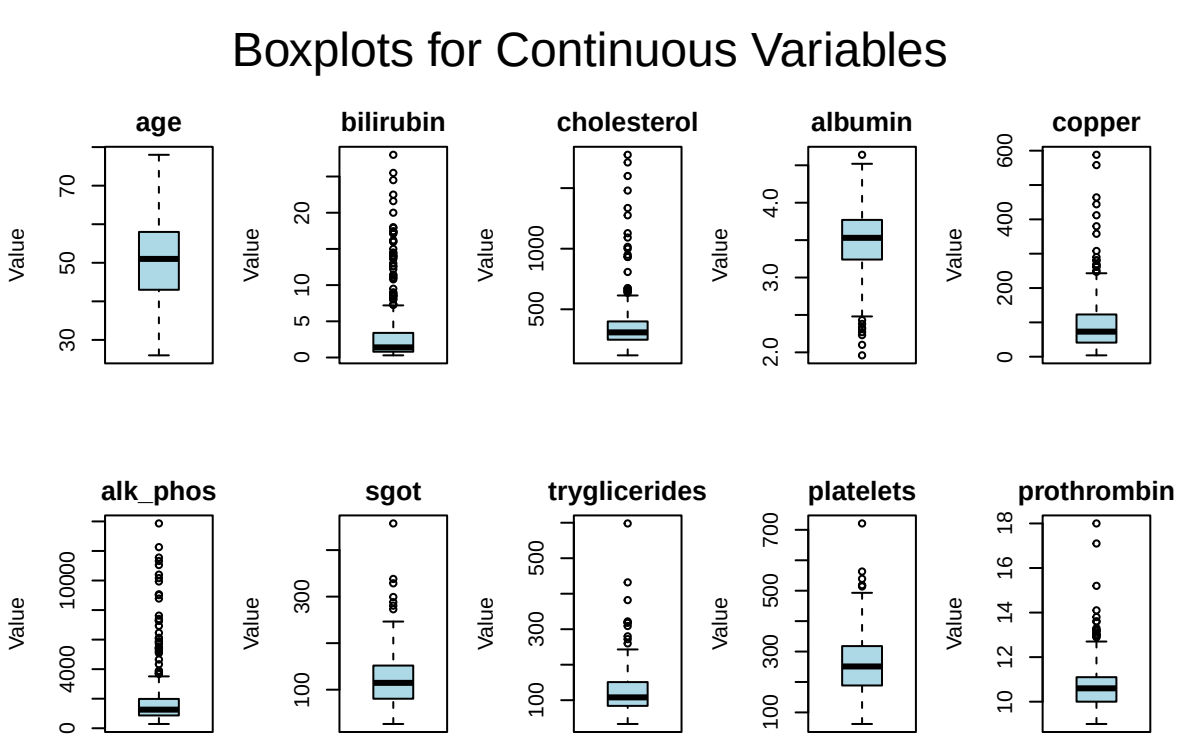


Boxplot for continuous variables

```
# Boxplot for all continuous variables
par(mfrow = c(2, 5), oma = c(2, 2, 3, 1), mar = c(4, 4, 2, 1))
conti_names <- names(conti_vars)

for (i in seq_along(conti_names)) {
  boxplot(conti_vars[[conti_names[i]]],
          main = conti_names[i],
          ylab = "Value",
          col = "lightblue",
          outline = TRUE)  # Show outliers
}

# Add an overall title
mtext("Boxplots for Continuous Variables", outer = TRUE, cex = 1.5, line = 1)
```

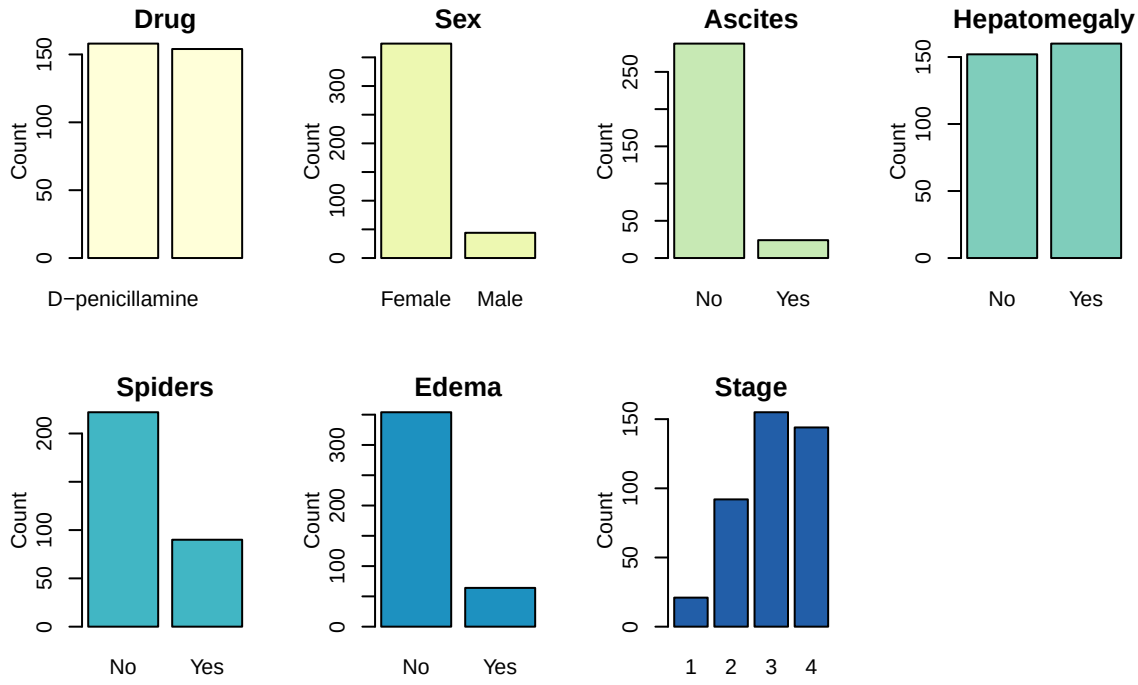


Bar Plots for categorical vairables

```
cate_vars = cirrhosis |>
  select(drug, sex, ascites, hepatomegaly, spiders, edema, stage)
```

```r
par(mfrow = c(2, 4),  # 2 rows, 5 columns
    oma = c(2, 2, 3, 1),  # Outer margins
    mar = c(4, 4, 2, 1),  # Inner margins for individual plots
    mgp = c(2, 1, 0))     # Margins for axis labels and titles

barplot(table(cate_vars$drug), main = "Drug", ylab = "Count", , col = colors[1])
barplot(table(cate_vars$sex), main = "Sex", ylab = "Count", , col = colors[2])
barplot(table(cate_vars$ascites), main = "Ascites", ylab = "Count", col = colors[3])
barplot(table(cate_vars$hepatomegaly), main = "Hepatomegaly", ylab = "Count", col = colors[4])
barplot(table(cate_vars$spiders), main = "Spiders", ylab = "Count", col = colors[5])
barplot(table(cate_vars$edema), main = "Edema", ylab = "Count", col = colors[6])
barplot(table(cate_vars$stage), main = "Stage", ylab = "Count", col = colors[7])
```



Correlation Plot

```r
numeric_cirr <- cirrhosis |>
  select_if(is.numeric)

cor_matrix <- cor(numeric_cirr, use = "complete.obs")

corrplot(cor_matrix, method = "circle", type = "lower", order = "hclust")
```

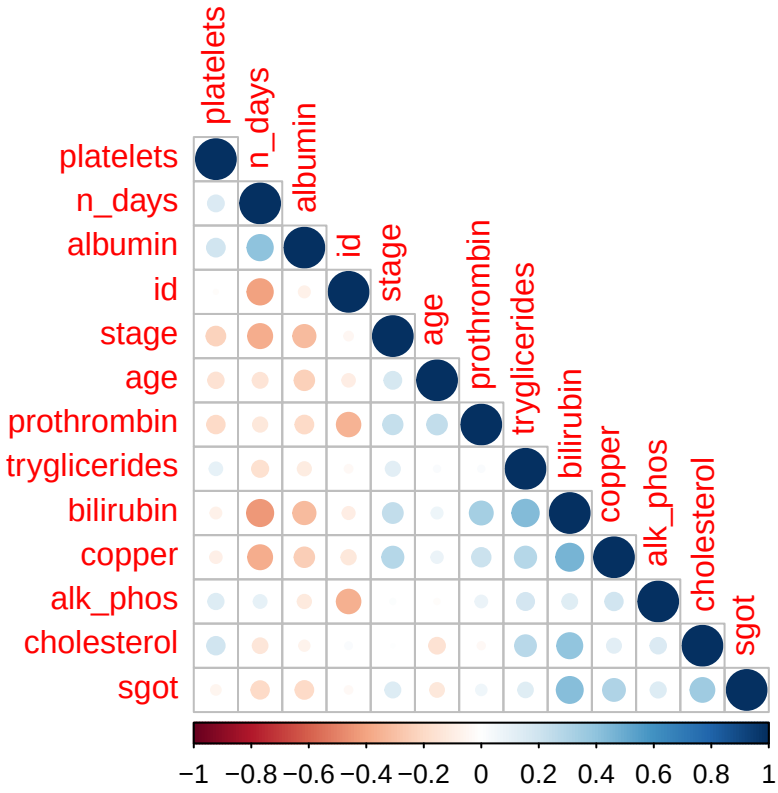


Table 1: Baseline Characteristics

```
theme_gtsummary_journal(journal = "nejm")
```

```
## Setting theme 'New England Journal of Medicine'
```

```
cirrhosis_df <- cirrhosis |>
  mutate(
    status = case_when(
      status == "C" ~ "Censored",
      status == "CL" ~ "Censored due to liver tx",
      status == "D" ~ "Death",
      TRUE ~ status))

table_1 <- cirrhosis_df |>
  select(-id) |>
  tbl_summary(
    by = status,
    statistic = list(
      all_continuous() ~ "{mean} / {median} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
```

```
    digits = all_continuous() ~ 1,
    missing = "no",
    label = list(
    n_days ~ "N_days",
    drug ~ "Drug",
    age ~ "Age",
    sex ~ "Sex",
    ascites ~ "Ascites",
    hepatomegaly ~ "Hepatomegaly",
    spiders ~ "Spiders",
    edema ~ "Edema",
    bilirubin ~ "Bilirubin",
    cholesterol ~ "Cholesterol",
    albumin ~ "Albumin",
    copper ~ "Copper",
    alk_phos ~ "Alk_phos",
    sgot ~ "SGOT",
    tryglicerides ~ "Trygliceride s",
    platelets ~ "Platelets",
    prothrombin ~ "Prothrombin",
    stage ~ "Stage"
  )) |>
  modify_caption("Baseline Characteristics") |>
  as_flex_table() |>
  line_spacing(space = 0, part = "body")

table_1
```

```
## Warning: fonts used in ‘flextable‘ are ignored because the ‘pdflatex‘ engine is
## used and not ‘xelatex‘ or ‘lualatex‘. You can avoid this warning by using the
## ‘set_flextable_defaults(fonts_ignore=TRUE)‘ command or use a compatible engine
## by defining ‘latex_engine: xelatex‘ in the YAML header of the R Markdown
## document.
```

Table 1: Baseline Characteristics

Characteristic	**Censored**, N = 232[1]	**Censored due to liver tx**, N = 25[1]	**Death**, N =
N_days	2,333.2 / 2,186.5 (994.7)	1,546.2 / 1,435.0 (753.1)	1,376.9 / 1,083.0
Drug			
D-penicillamine	83 (49%)	10 (53%)	65 (52%)
Placebo	85 (51%)	9 (47%)	60 (48%)
Age	49.6 / 50.0 (10.4)	41.6 / 41.0 (6.3)	54.0 / 54.0 (
Sex			
Female	215 (93%)	22 (88%)	137 (85%
Male	17 (7.3%)	3 (12%)	24 (15%)
Ascites	1 (0.6%)	0 (0%)	23 (18%)

[1]Mean / Median (SD); n (%)

Table 1: Baseline Characteristics

Characteristic	**Censored**, N = 232[1]	**Censored due to liver tx**, N = 25[1]	**Death**, N =
Hepatomegaly	60 (36%)	12 (63%)	88 (70%)
Spiders	33 (20%)	5 (26%)	52 (42%)
Edema	16 (6.9%)	3 (12%)	45 (28%)
Bilirubin	1.6 / 0.9 (1.9)	3.6 / 3.1 (3.6)	5.5 / 3.2 (5.
Cholesterol	326.5 / 292.0 (165.8)	439.5 / 343.5 (335.5)	415.8 / 339.0 (
Albumin	3.6 / 3.6 (0.4)	3.5 / 3.5 (0.5)	3.4 / 3.4 (0.
Copper	66.6 / 52.0 (57.1)	124.0 / 102.0 (100.1)	135.4 / 111.0 (
Alk_phos	1,578.1 / 1,107.5 (1,633.1)	1,535.2 / 1,345.0 (837.7)	2,594.4 / 1,664.0 (
SGOT	107.3 / 94.6 (52.8)	130.1 / 127.0 (36.9)	141.9 / 134.9 (
Tryglicerides	111.8 / 104.0 (48.3)	133.9 / 124.0 (70.5)	140.5 / 122.0 (
Platelets	261.2 / 256.0 (88.6)	309.6 / 304.0 (102.7)	242.5 / 224.0 (
Prothrombin	10.5 / 10.4 (0.9)	10.4 / 10.3 (0.5)	11.2 / 11.0 (
Stage			
1	19 (8.3%)	0 (0%)	2 (1.3%)
2	64 (28%)	5 (20%)	23 (15%)
3	97 (42%)	10 (40%)	48 (31%)
4	50 (22%)	10 (40%)	84 (54%)

[1]Mean / Median (SD); n (%)

Multivariate analysis

```
cirrhosis <- cirrhosis |>
  mutate(
    status = case_when(
      status == "D" ~ 1,  # Event of interest (death)
      status == "C" | status == "CL" ~ 0,  # Censored data
      TRUE ~ as.numeric(status)))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `status = case_when(...)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Fit the Cox proportional hazards model
cox_model <- coxph(Surv(n_days, status) ~ age + sex + bilirubin + albumin + copper + prothrombin + stage
                   data = cirrhosis)
```

```r
# Summarize the results
cox_summary <- tbl_regression(cox_model, exponentiate = TRUE) %>%
  modify_caption("Multivariate Cox Proportional Hazards Analysis")

cox_summary
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Table 2: Multivariate Cox Proportional Hazards Analysis

Characteristic	**HR**	**95% CI**	**p-value**
age	1.02	1.00 to 1.04	0.019
sex			
Female	—	—	
Male	1.30	0.75 to 2.26	0.35
bilirubin	1.12	1.08 to 1.16	<0.001
albumin	0.35	0.22 to 0.56	<0.001
copper	1.00	1.00 to 1.01	0.002
prothrombin	1.32	1.12 to 1.57	0.001
stage	1.46	1.13 to 1.88	0.003