

# 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
```

## Histogram Plots

```

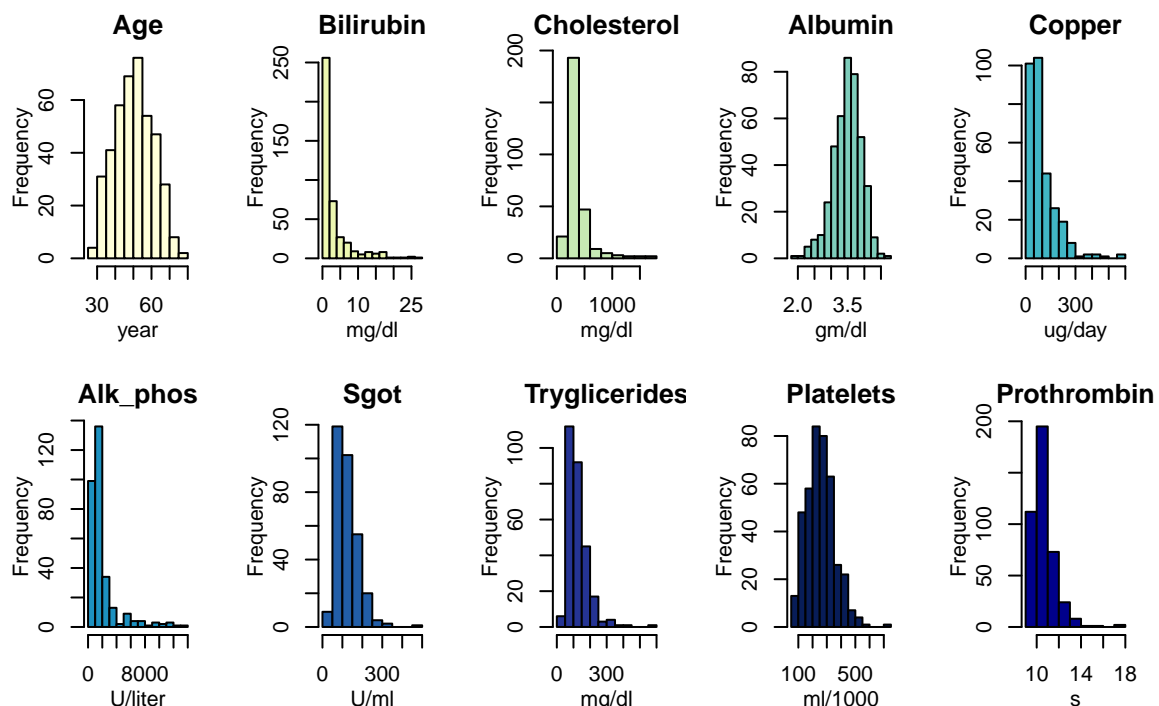
conti_vars = cirrhosis |>
  select(age, bilirubin, cholesterol, albumin, copper, alk_phos, sgot, tryglicerides, platelets, prothrombin)

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$sgot, main = "Sgot", xlab = "U/ml", ylab = "Frequency", col = colors[7])
hist(conti_vars$tryglicerides, main = "Tryglicerides", xlab = "mg/dl", ylab = "Frequency", col = colors[8])
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])

```

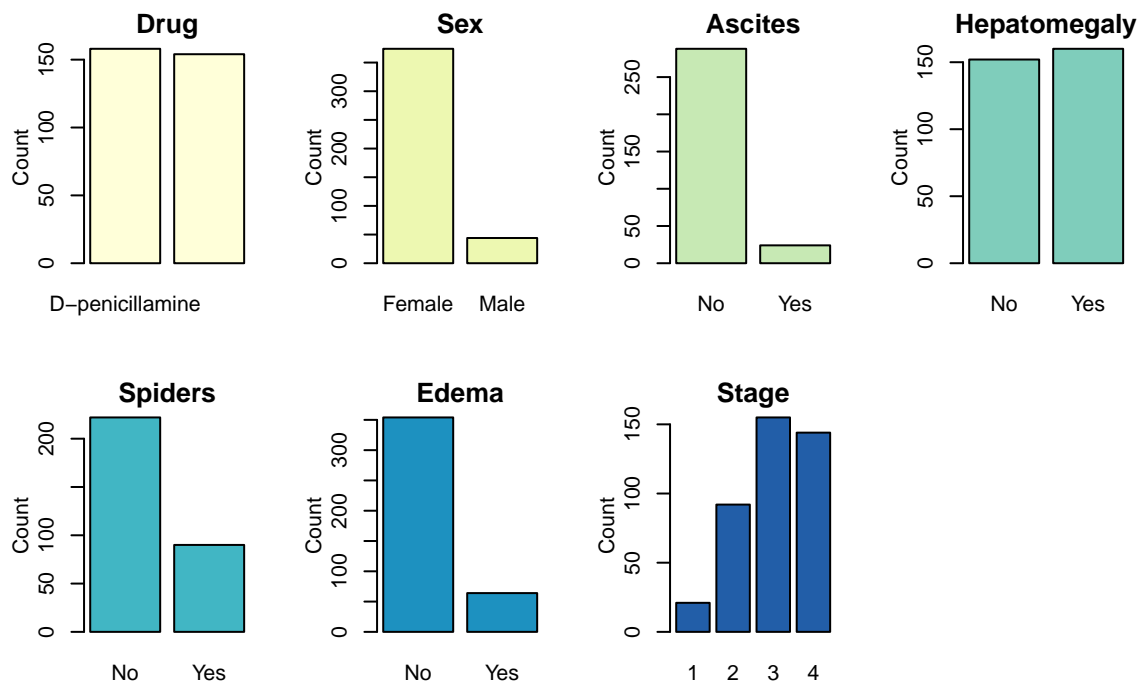


## Bar Plots

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

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

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

```

# Compute the correlation matrix
cor_matrix <- cor(numeric_cirr, use = "complete.obs")

# Plot the correlation matrix
corrplot(cor_matrix, method = "circle", type = "upper", order = "hclust")

```

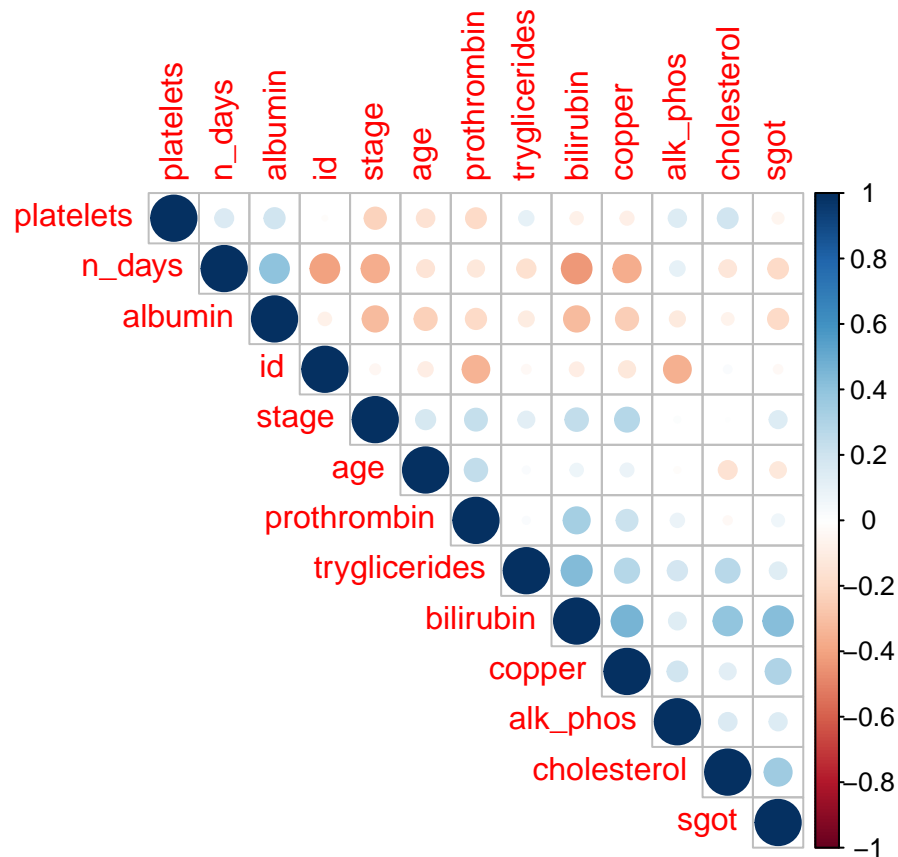


Table 1: Baseline Characteristics

```

theme_gtsummary_journal(journal = "nejm")

```

```

## Setting theme 'New England Journal of Medicine'

```

```

table_1 <- cirrhosis |>
  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 ~ "Tryglicerides",
  platelets ~ "Platelets",
  prothrombin ~ "Prothrombin",
  stage ~ "Stage"
)) |>
modify_caption("Table 1: 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: Table 1: Baseline Characteristics			
Characteristic	C, N = 232 <sup>1</sup>	CL, N = 25 <sup>1</sup>	D, N = 161 <sup>1</sup>
N_days	2,333.2 / 2,186.5 (994.7)	1,546.2 / 1,435.0 (753.1)	1,376.9 / 1,083.0 (1,049.2)
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 (9.8)
Sex			
Female	215 (93%)	22 (88%)	137 (85%)
Male	17 (7.3%)	3 (12%)	24 (15%)
Ascites	1 (0.6%)	0 (0%)	23 (18%)
Hepatomegaly	60 (36%)	12 (63%)	88 (70%)
<sup>1</sup> Mean / Median (SD); n (%)			

Table 1: Table 1: Baseline Characteristics

Characteristic	C, N = 232 <sup>1</sup>	CL, N = 25 <sup>1</sup>	D, N = 161 <sup>1</sup>
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 (5.8)
Cholesterol	326.5 / 292.0 (165.8)	439.5 / 343.5 (335.5)	415.8 / 339.0 (275.0)
Albumin	3.6 / 3.6 (0.4)	3.5 / 3.5 (0.5)	3.4 / 3.4 (0.5)
Copper	66.6 / 52.0 (57.1)	124.0 / 102.0 (100.1)	135.4 / 111.0 (98.5)
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 (2,677.1)
SGOT	107.3 / 94.6 (52.8)	130.1 / 127.0 (36.9)	141.9 / 134.9 (58.4)
Tryglicerides	111.8 / 104.0 (48.3)	133.9 / 124.0 (70.5)	140.5 / 122.0 (79.3)
Platelets	261.2 / 256.0 (88.6)	309.6 / 304.0 (102.7)	242.5 / 224.0 (107.9)
Prothrombin	10.5 / 10.4 (0.9)	10.4 / 10.3 (0.5)	11.2 / 11.0 (1.0)
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%)

<sup>1</sup>Mean / Median (SD); n (%)

## Boxplot for all 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)
```

## Boxplots for Continuous Variables

