EDA

Chen Liang

2024-11-16

Load data

```
cirrhosis <- read_csv("data/cirrhosis.csv")|>
  janitor::clean_names() |>
  mutate(age = round(age / 365))
```

```
## 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.
```

```
summary(cirrhosis)
```

```
##        id            n_days        status              drug
##  Min.   :  1.0   Min.   :  41   Length:418         Length:418
##  1st Qu.:105.2   1st Qu.:1093   Class :character   Class :character
##  Median :209.5   Median :1730   Mode  :character   Mode  :character
##  Mean   :209.5   Mean   :1918
##  3rd Qu.:313.8   3rd Qu.:2614
##  Max.   :418.0   Max.   :4795
##
##       age            sex              ascites          hepatomegaly
##  Min.   :26.00   Length:418         Length:418         Length:418
##  1st Qu.:43.00   Class :character   Class :character   Class :character
##  Median :51.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :50.77
##  3rd Qu.:58.00
##  Max.   :78.00
##
##    spiders             edema            bilirubin       cholesterol
##  Length:418         Length:418         Min.   : 0.300   Min.   : 120.0
##  Class :character   Class :character   1st Qu.: 0.800   1st Qu.: 249.5
##  Mode  :character   Mode  :character   Median : 1.400   Median : 309.5
##                                        Mean   : 3.221   Mean   : 369.5
##                                        3rd Qu.: 3.400   3rd Qu.: 400.0
```

```
##                                      Max.   :28.000   Max.   :1775.0
##                                                       NA's   :134
##     albumin          copper          alk_phos            sgot
##  Min.   :1.960   Min.   :  4.00   Min.   :  289.0   Min.   : 26.35
##  1st Qu.:3.243   1st Qu.: 41.25   1st Qu.:  871.5   1st Qu.: 80.60
##  Median :3.530   Median : 73.00   Median : 1259.0   Median :114.70
##  Mean   :3.497   Mean   : 97.65   Mean   : 1982.7   Mean   :122.56
##  3rd Qu.:3.770   3rd Qu.:123.00   3rd Qu.: 1980.0   3rd Qu.:151.90
##  Max.   :4.640   Max.   :588.00   Max.   :13862.4   Max.   :457.25
##                  NA's   :108      NA's   :106       NA's   :106
##  tryglicerides      platelets      prothrombin        stage
##  Min.   : 33.00   Min.   : 62.0   Min.   : 9.00   Min.   :1.000
##  1st Qu.: 84.25   1st Qu.:188.5   1st Qu.:10.00   1st Qu.:2.000
##  Median :108.00   Median :251.0   Median :10.60   Median :3.000
##  Mean   :124.70   Mean   :257.0   Mean   :10.73   Mean   :3.024
##  3rd Qu.:151.00   3rd Qu.:318.0   3rd Qu.:11.10   3rd Qu.:4.000
##  Max.   :598.00   Max.   :721.0   Max.   :18.00   Max.   :4.000
##  NA's   :136      NA's   :11      NA's   :2       NA's   :6
```

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

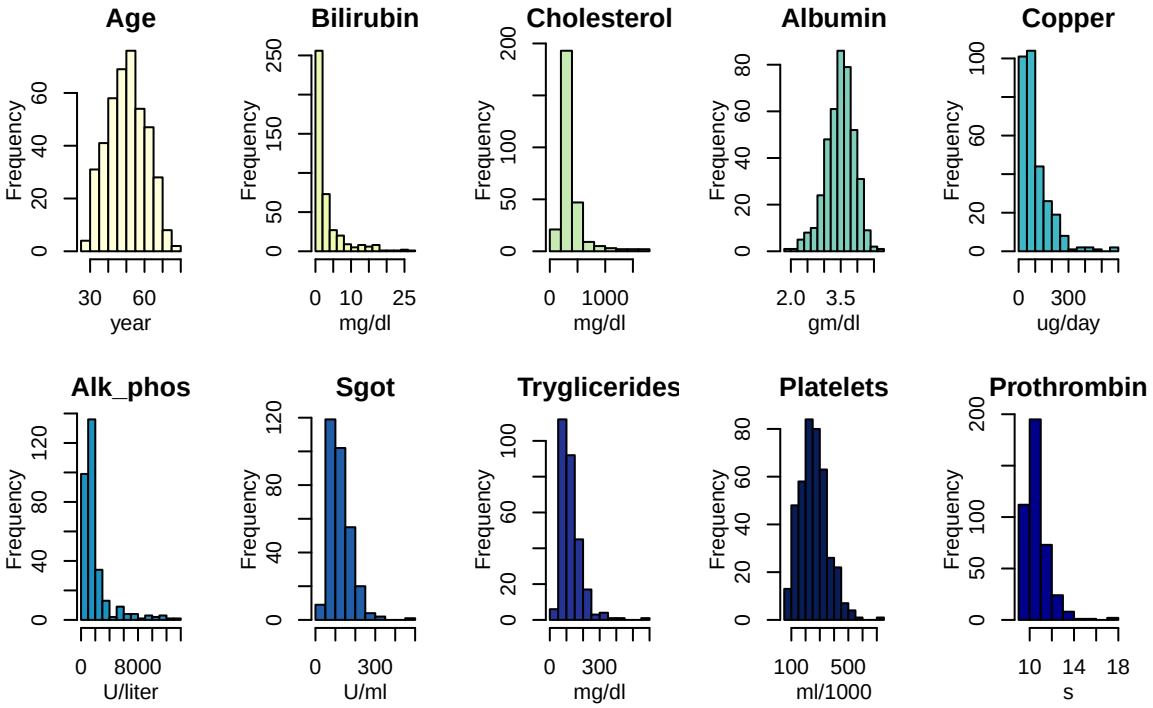
Historgram Plots

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


Age
Frequency
0
20
40
60
30
60
year
Bilirubin
Frequency
0
50
150
250
0
10
25
mg/dl
Cholesterol
Frequency
0
50
100
200
0
1000
mg/dl
Albumin
Frequency
0
20
40
60
80
2.0
3.5
gm/dl
Copper
Frequency
0
20
60
100
0
300
ug/day
Alk_phos
Frequency
0
40
80
120
0
8000
U/liter
Sgot
Frequency
0
40
80
120
0
300
U/ml
Tryglicerides
Frequency
0
20
60
100
0
300
mg/dl
Platelets
Frequency
0
20
40
60
80
100
500
ml/1000
Prothrombin
Frequency
0
50
100
200
10
14
18
s