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
```

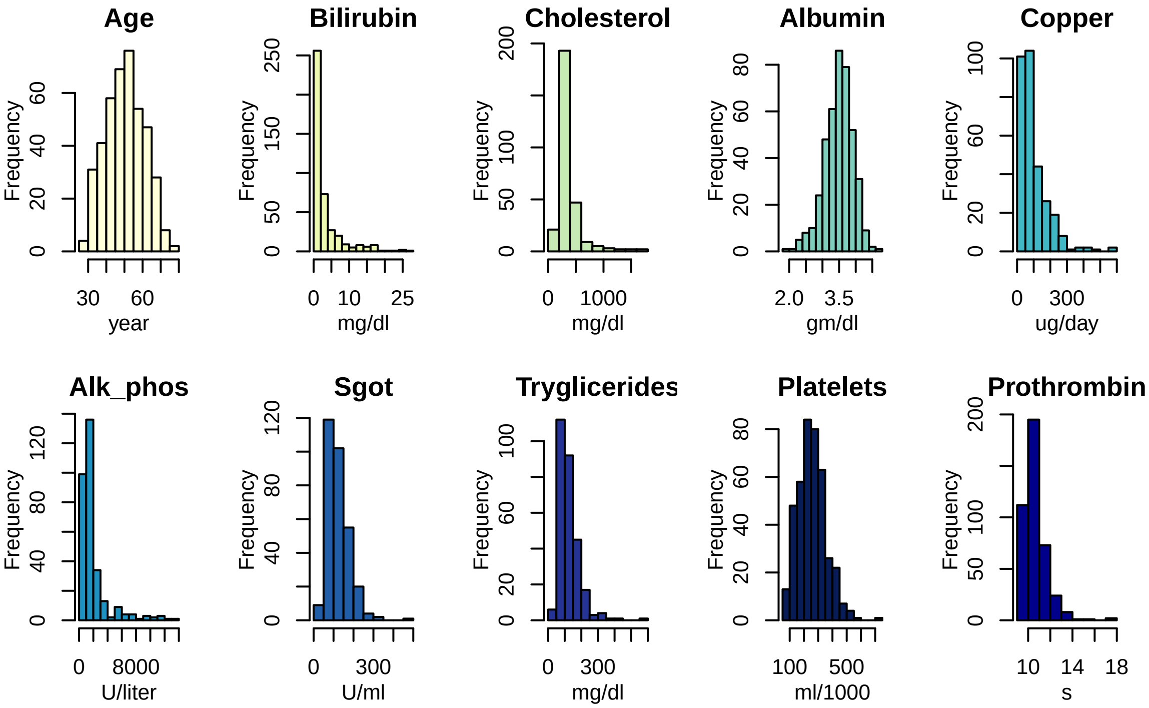
Historgram Plots

```
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])
```

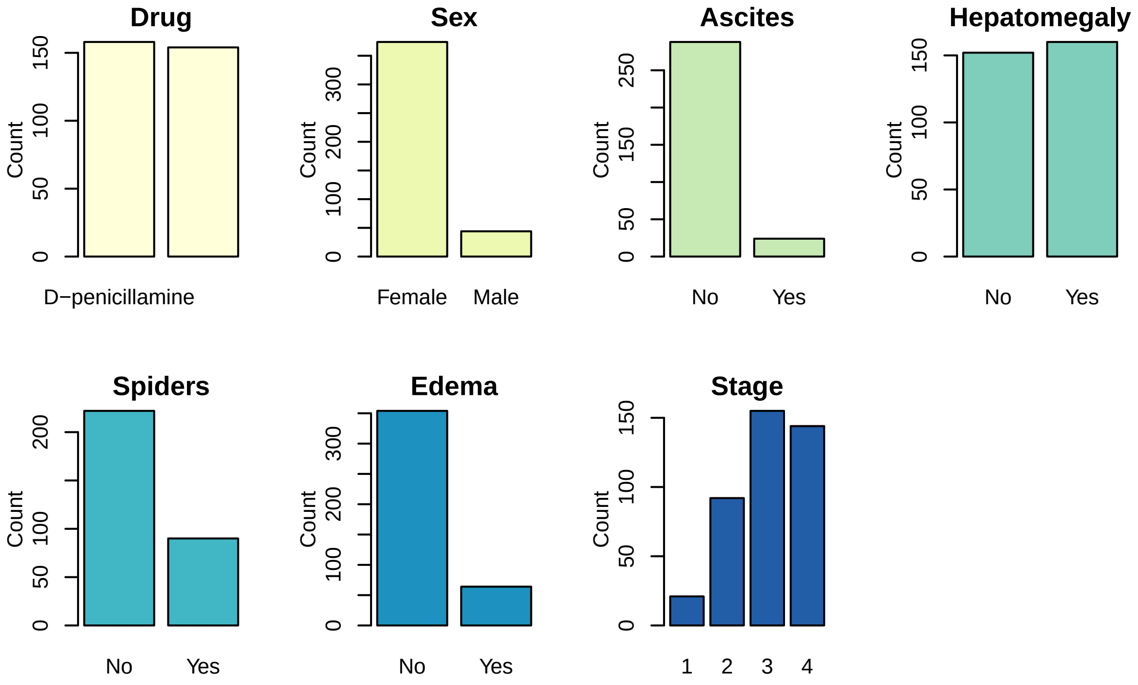


Bar Plots

```
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])
```


Drug
Count
D−penicillamine
Sex
Count
Female
Male
Ascites
Count
No
Yes
Hepatomegaly
Count
No
Yes
Spiders
Count
No
Yes
Edema
Count
No
Yes
Stage
Count