PRA 2: Neteja, validació i anàlisi de les dades

Álvaro Díaz i David Leiva

13/12/2020

Contents

Detalls de l'activitat

Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes

Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració,transformació, neteja i validació) per al seu posterior anàlisi.

Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Resolució

Descripció del dataset

Es realitzarà l'anàlisi exploratori (EDA) del dataset https://archive.ics.uci.edu/ml/datasets/HCC+Survival on es recopila informació de pacients amb carcinoma hepatocelullar (HCC) i la seva supervivència a l'any.

L'HCC és el tumor hepàtic més comú en els pacients amb hepatopatia crónica. La supervivència d'aquests pacients no només depen de l'estadi tumoral si no que també depen de l'estat funcional del fetge.

El conjunt de dades HCC es va obtenir en l´'Hospital Universitari de Coïmbra (Portugal) i contenia diversos factors demogràfics, de risc, de laboratori i de supervivència global de 165 pacients reals diagnosticats de HCC. El conjunt de dades conté 49 funcions seleccionades segons les directrius de pràctica clínica EASL-EORTC (Associació Europea per a l'Estudi del Fetge - Organització Europea per a la Recerca i el Tractament del Càncer), que són els habituals en la gestió de HCC.

Es tracta d'un conjunt de dades heterogeni, amb 23 variables quantitatives i 26 variables qualitatives. La variable objectiu és la supervivència a 1 any i es va codificar com a variable binària: 0 (mor) i 1 (viu).

Les variables del dataset són:

- **Gender (Gen)**: [1=Home;0=Dona] Sexe del pacient
- **Symptoms (Sym)**:[1=Si;0=No] Sintomàtic
- **Alcohol (Alc)**: [1=Si;0=No] Hepatopatia alcohòlica
- **Hepatitis B Surface Antigen (HBS)**: [1=Si;0=No] Antigen de superfície de l'hepatitis B present a la sang
- **Hepatitis B e Antigen (HBe)**:[1=Si;0=No] Antigen e de l'hepatitis B present a la sang
- **Hepatitis B Core Antibody (HBC)**: [1=Si;0=No] Anticòs per l'hepatitis B present a la sang
- **Hepatitis C Virus Antibody (HCV)**: [1=Si;0=No] Anticòs per l'hepatitis C present a la sang
- **Cirrhosis (Cir)**: [1=Si;0=No] Estadio avançat d'hepatopatia crónica
- **Endemic Countries (End)**: [1=Si;0=No] Pacient procedent de països amb alta prevalença d'hepatitis vírica
- **Smoking (Smo)**: [1=Si;0=No] Fumador
- **Diabetes (Dia)**: [1=Si;0=No] Diabètic
- **Obesity (Obe)**: [1=Si;0=No] Obesitat
- **Hemochromatosis (Hem)**:[1=Si;0=No] Hemocromatosi
- **Arterial Hypertension (HyA)**: [1=Si;0=No] Hipertensió arterial
- **Chronic Renal Insufficiency(CRI)**: [1=Si;0=No] Insuficiència renal
- **Human Immunodeficiency Virus (HIV)**: [1=Si;0=No] Infecció per HIV
- **Nonalcoholic Steatohepatitis (Ste)**: [1=Si;0=No] Esteatosis hepàtica de origen no alcohòlic
- **Esophageal Varices (Eso)**: [1=Si;0=No] Presència de varius esofàgiges com indicador d'hipertensió portal
- **Splenomegaly (Spl)**: [1=Si;0=No] Augment del tamany de la melsa com indicador d'hipertensió portal
- **Portal Hypertension (PHT)**: [1=Si;0=No] Pacient amb hipertensió arterial coneguda
- **Portal Vein Thrombosis (PVT)**: [1=Si;0=No] Presència de trombosi venosa portal
- **Liver Metastasis (Met)**: [1=Si;0=No] Metàstasi hepàtica
- **Radiological Hallmark (Rad)**: [1=Si;0=No]Comportament radiològic típic per HCC
- **Age at diagnosis (Age)**: Anys d'edat al moment del diagnóstic de HCC
- **Grams of Alcohol per day (gAl)**: Grams d'alcohol ingerit de mitjana al dia
- **Packs of cigarets per year (PCi)**: Número de paquets de cigarrets consumits per any
- **Performance Status (PSt)**: [0=Activo;1=Restringit;2=Asistencia ocasional;3=Asistencia parcial;4=Asistencia total;5=Mort] Escala de l'estat general del pacient oncològic

- **Encephalopathy degree (Enc)**: [1=Cap;2=Grau I/II; 3=Grau III/IV] Grau d'afectació mental de l'hepatopatia
- **Ascites degree (Asc)**: [1=Cap;2=Lleu;3=Moderada a Severa] Grau d'ascitis com a indicador indirecte d'hipertensió portal
- **International Normalised Ratio (INR)**: Temps de protrombina
- **Alpha-Fetoprotein (ng/mL) (AFe)**: Nivells del marcador tumoral a la sang
- **Haemoglobin (g/dL) (Hae)**: Nivells de Hemoglobina a la sang
- **Mean Corpuscular Volume (MCV)**: Volum corpuscular mig dels eritrocits
- **Leukocytes(G/L) (Leu)**: Concentració de cél.lules blanques en sang
- **Platelets (Pla)**: Concentració de plaquestes en sang
- **Albumin (mg/dL) (Alb)**: Nivelss d'albumina en sang
- **Total Bilirubin(mg/dL) (BiT)**: Nivells de Bilirrubina Total en sang
- **Alanine transaminase (U/L) (ALT)**: Nivells d'ALT en sang
- **Aspartate transaminase (U/L) (AST)**: Nivells d'ASP en sang
- **Gamma glutamyl transferase (U/L) (GGT)**: Nivells gamma-GT en sang
- **Alkaline phosphatase (U/L) (ALP)**: Nivelss de fosfatasa alcalina en sang
- **Total Proteins (g/dL) (Pro)**: Concentració total de proteïnes en sang
- **Creatinine (mg/dL) (Crea)**: Concentració de creatinina en sang
- **Number of Nodules (Nod)**: Número de nóduls d'HCC visualitzats
- **Major dimension of nodule (cm) (DiN)**: Tamany major dels nóduls d'HCC
- **Direct Bilirubin (mg/dL) (BiD)**: Nivells de Bilirrubina Directe en sang
- **Iron (Iro)**: Concentració de ferro en sang
- **Oxygen Saturation (%) (OxS)**: Saturació d'oxigen de la sang
- **Ferritin (ng/mL) (Fer)**: Nivells de ferritina en sang
- **Class Attribute (Class)**:[0=Mort; 1=Viu] Supervivent a l'any del diagnóstic d'HCC

Importància i objectius de l'anàlisi

Amb les dades recopilades al dataset podem intentar saber quines variables estan més relacionades amb la supervivència a l'any. Podem conèixer el grau de correlació entre les variables independents per finalment definir un model de regressió logística per tal d'intentar predir la mortalitat a l'any amb les variables seleccionades.

Poder conèixer la probabilitat de supervivència d'un pacient amb diagnóstic recent d'HCC podrà fer que s'adaptin millors les opcions de tractament, sent més agressius en pacients amb alta probilitat de sobreviure, i en canvi, optant per terapies pal · liatives o de confort per pacients amb pitjor prosnóstic.

Neteja de les dades

Preparació del dataset

```
# Importació del dataset
download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/00423/hcc-survival.zip","mydata
unzip("mydata.zip")
hcc<-read.csv ("hcc-survival/hcc-data.txt",header = F, dec = ".", stringsAsFactors = F, na.strings = "?

# Noms de les variables
hcc_header<-c("Gender", "Symptoms", "Alcohol", "Hepatitis B Surface Antigen",
              "Hepatitis B e Antigen", "Hepatitis B Core Antibody",
              "Hepatitis C Virus Antibody",  "Cirrhosis",
              "Endemic Countries","Smoking",  "Diabetes", "Obesity",
              "Hemochromatosis", "Arterial Hypertension",
              "Chronic Renal Insufficiency", "Human Immunodeficiency Virus",
              "Nonalcoholic Steatohepatitis","Esophageal Varices",
              "Splenomegaly",  "Portal Hypertension",
              "Portal Vein Thrombosis", "Liver Metastasis",
              "Radiological Hallmark", "Age at diagnosis",
              "Grams of Alcohol per day", "Packs of cigarets per year",
              "Performance Status", "Encephalopathy degree", "Ascites degree",
              "International Normalised Ratio", "Alpha-Fetoprotein (ng/mL)",
              "Haemoglobin (g/dL)", "Mean Corpuscular Volume",
              "Leukocytes(G/L)", "Platelets", "Albumin (mg/dL)",
              "Total Bilirubin(mg/dL)",  "Alanine transaminase (U/L)"
              , "Aspartate transaminase (U/L)",
              "Gamma glutamyl transferase (U/L)",
              "Alkaline phosphatase (U/L)","Total Proteins (g/dL)",
              "Creatinine (mg/dL)","Number of Nodules",
              "Major dimension of nodule (cm)",
              "Direct Bilirubin (mg/dL)", "Iron", "Oxygen Saturation (%)",
              "Ferritin (ng/mL)", "Class Attribute"   )

hcc_header_cortos<-c("Gen", "Sym", "Alc", "HBS", "HBe", "HBC","HCV",  "Cir",
                     "End","Smo",  "Dia", "Obe", "Hem", "HyA", "CRI", "HIV",
                     "Ste","Eso", "Spl",  "PHT", "PVT", "Met", "Rad", "Age", "gAl",
                     "PCi", "PSt", "Enc", "Asc", "INR", "AFe", "Hae", "MCV",    "Leu",
                     "Pla", "Alb","BiT",  "ALT", "AST","GGT", "ALP","Pro",
                     "Crea","Nod",  "DiN", "BiD", "Iro", "OxS", "Fer", "Class"   )

colnames(hcc)<-hcc_header

tipVar <- c()
for (i in 1:ncol(hcc))  tipVar <- c(tipVar,is(hcc[,i])[1])
tipVar <- table(tipVar)
```

Al nostre dataset, tenim 32 variables de tipus integer i 18 variables de tipus numeric.

```
summary(hcc)
```

```
##      Gender          Symptoms         Alcohol       Hepatitis B Surface Antigen
```

```
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :0.0000
##  Mean   :0.8061   Mean   :0.6395   Mean   :0.7394   Mean   :0.1081
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##                   NA's   :18                        NA's   :17
##  Hepatitis B e Antigen Hepatitis B Core Antibody Hepatitis C Virus Antibody
##  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 :0.0000            Median :0.0000
##  Mean   :0.00794       Mean   :0.2695            Mean   :0.2179
##  3rd Qu.:0.00000       3rd Qu.:1.0000            3rd Qu.:0.0000
##  Max.   :1.00000       Max.   :1.0000            Max.   :1.0000
##  NA's   :39            NA's   :24                NA's   :9
##    Cirrhosis     Endemic Countries    Smoking          Diabetes
##  Min.   :0.000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.903   Mean   :0.07937   Mean   :0.5081   Mean   :0.3457
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##                  NA's   :39        NA's   :41       NA's   :3
##     Obesity      Hemochromatosis  Arterial Hypertension
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.129   Mean   :0.0493   Mean   :0.3642
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##  NA's   :10      NA's   :23       NA's   :3
##  Chronic Renal Insufficiency Human Immunodeficiency Virus
##  Min.   :0.0000              Min.   :0.00000
##  1st Qu.:0.0000              1st Qu.:0.00000
##  Median :0.0000              Median :0.00000
##  Mean   :0.1227              Mean   :0.01987
##  3rd Qu.:0.0000              3rd Qu.:0.00000
##  Max.   :1.0000              Max.   :1.00000
##  NA's   :2                   NA's   :14
##  Nonalcoholic Steatohepatitis Esophageal Varices  Splenomegaly
##  Min.   :0.00000               Min.   :0.0000     Min.   :0.00
##  1st Qu.:0.00000               1st Qu.:0.0000     1st Qu.:0.00
##  Median :0.00000               Median :1.0000     Median :1.00
##  Mean   :0.05594               Mean   :0.6106     Mean   :0.56
##  3rd Qu.:0.00000               3rd Qu.:1.0000     3rd Qu.:1.00
##  Max.   :1.00000               Max.   :1.0000     Max.   :1.00
##  NA's   :22                    NA's   :52         NA's   :15
##  Portal Hypertension Portal Vein Thrombosis Liver Metastasis
##  Min.   :0.0000      Min.   :0.0000         Min.   :0.0000
##  1st Qu.:0.0000      1st Qu.:0.0000         1st Qu.:0.0000
##  Median :1.0000      Median :0.0000         Median :0.0000
##  Mean   :0.7143      Mean   :0.2222         Mean   :0.2236
##  3rd Qu.:1.0000      3rd Qu.:0.0000         3rd Qu.:0.0000
##  Max.   :1.0000      Max.   :1.0000         Max.   :1.0000
```

```
##  NA's   :11          NA's   :3             NA's   :4
##  Radiological Hallmark Age at diagnosis Grams of Alcohol per day
##  Min.   :0.000         Min.   :20.00    Min.   :  0.00
##  1st Qu.:0.000         1st Qu.:57.00    1st Qu.:  0.00
##  Median :1.000         Median :66.00    Median : 75.00
##  Mean   :0.681         Mean   :64.69    Mean   : 71.01
##  3rd Qu.:1.000         3rd Qu.:74.00    3rd Qu.:100.00
##  Max.   :1.000         Max.   :93.00    Max.   :500.00
##  NA's   :2                              NA's   :48
##  Packs of cigarets per year Performance Status Encephalopathy degree
##  Min.   :  0.00             Min.   :0.000      Min.   :1.000
##  1st Qu.:  0.00             1st Qu.:0.000      1st Qu.:1.000
##  Median :  0.00             Median :1.000      Median :1.000
##  Mean   : 20.46             Mean   :1.018      Mean   :1.159
##  3rd Qu.: 30.50             3rd Qu.:2.000      3rd Qu.:1.000
##  Max.   :510.00             Max.   :4.000      Max.   :3.000
##  NA's   :53                                    NA's   :1
##  Ascites degree  International Normalised Ratio Alpha-Fetoprotein (ng/mL)
##  Min.   :1.000   Min.   :0.840                  Min.   :1.20e+00
##  1st Qu.:1.000   1st Qu.:1.170                  1st Qu.:5.20e+00
##  Median :1.000   Median :1.300                  Median :3.30e+01
##  Mean   :1.442   Mean   :1.422                  Mean   :1.93e+04
##  3rd Qu.:2.000   3rd Qu.:1.530                  3rd Qu.:6.15e+02
##  Max.   :3.000   Max.   :4.820                  Max.   :1.81e+06
##  NA's   :2       NA's   :4                      NA's   :8
##  Haemoglobin (g/dL) Mean Corpuscular Volume Leukocytes(G/L)
##  Min.   : 5.00      Min.   : 69.50          Min.   :    2.20
##  1st Qu.:11.43      1st Qu.: 89.78          1st Qu.:    5.10
##  Median :13.05      Median : 94.95          Median :    7.20
##  Mean   :12.88      Mean   : 95.12          Mean   : 1473.96
##  3rd Qu.:14.60      3rd Qu.:100.67          3rd Qu.:   19.52
##  Max.   :18.70      Max.   :119.60          Max.   :13000.00
##  NA's   :3          NA's   :3               NA's   :3
##    Platelets        Albumin (mg/dL) Total Bilirubin(mg/dL)
##  Min.   :     1.7   Min.   :1.900   Min.   : 0.300
##  1st Qu.:   255.8   1st Qu.:3.000   1st Qu.: 0.800
##  Median : 93000.0   Median :3.400   Median : 1.400
##  Mean   :113206.4   Mean   :3.446   Mean   : 3.088
##  3rd Qu.:171500.0   3rd Qu.:4.050   3rd Qu.: 2.925
##  Max.   :459000.0   Max.   :4.900   Max.   :40.500
##  NA's   :3          NA's   :6       NA's   :5
##  Alanine transaminase (U/L) Aspartate transaminase (U/L)
##  Min.   : 11.00             Min.   : 17.00
##  1st Qu.: 31.00             1st Qu.: 46.25
##  Median : 50.00             Median : 71.00
##  Mean   : 67.09             Mean   : 96.38
##  3rd Qu.: 78.00             3rd Qu.:110.25
##  Max.   :420.00             Max.   :553.00
##  NA's   :4                  NA's   :3
##  Gamma glutamyl transferase (U/L) Alkaline phosphatase (U/L)
##  Min.   :  23.00                  Min.   :  1.28
##  1st Qu.:  91.25                  1st Qu.:108.25
##  Median : 179.50                  Median :162.00
##  Mean   : 268.03                  Mean   :212.21
```

```
##  3rd Qu.: 345.25                 3rd Qu.:261.50
##  Max.   :1575.00                 Max.   :980.00
##  NA's   :3                       NA's   :3
##  Total Proteins (g/dL) Creatinine (mg/dL) Number of Nodules
##  Min.   :  3.900       Min.   :0.200      Min.   :0.000
##  1st Qu.:  6.300       1st Qu.:0.700      1st Qu.:1.000
##  Median :  7.050       Median :0.850      Median :2.000
##  Mean   :  8.961       Mean   :1.127      Mean   :2.736
##  3rd Qu.:  7.575       3rd Qu.:1.100      3rd Qu.:5.000
##  Max.   :102.000       Max.   :7.600      Max.   :5.000
##  NA's   :11            NA's   :7          NA's   :2
##  Major dimension of nodule (cm) Direct Bilirubin (mg/dL)      Iron
##  Min.   : 1.500                 Min.   : 0.10            Min.   :  0.0
##  1st Qu.: 3.000                 1st Qu.: 0.37            1st Qu.: 40.5
##  Median : 5.000                 Median : 0.70            Median : 83.0
##  Mean   : 6.851                 Mean   : 1.93            Mean   : 85.6
##  3rd Qu.: 9.000                 3rd Qu.: 1.40            3rd Qu.:118.0
##  Max.   :22.000                 Max.   :29.30            Max.   :224.0
##  NA's   :20                     NA's   :44               NA's   :79
##  Oxygen Saturation (%) Ferritin (ng/mL) Class Attribute
##  Min.   :  0.00        Min.   :   0     Min.   :0.0000
##  1st Qu.: 16.00        1st Qu.:  84     1st Qu.:0.0000
##  Median : 27.00        Median : 295     Median :1.0000
##  Mean   : 37.03        Mean   : 439     Mean   :0.6182
##  3rd Qu.: 56.00        3rd Qu.: 706     3rd Qu.:1.0000
##  Max.   :126.00        Max.   :2230     Max.   :1.0000
##  NA's   :80            NA's   :80
```

Descripció del dataset

El dataset està compost de 165 observacions de 49 atributs de pacients amb una variable de classe que registra la supervivència a l'any del diagnóstic. Dels atributs dels pacients, existeixen 26 categorics, 3 dels quals són ordinals (*Performance Status*, *Encephalopathy degree*, *Ascites degree*), sent la resta numérics. És pot veure que existeixen valors nuls codificats com *NA*.

```
# Definició de tipus de dades per columna
hcc_factor <-c(1:23,50)
hcc_order <- c(27:29)
hcc_factorT<-c(1:23,27:29,50)
hcc_num<-c(24:26,30:49)

# Factorització de les columnes categóriques
hcc <- hcc %>% mutate_at(vars(c(1:23,50)), as.factor)
hcc <- hcc %>% mutate_at(vars(c(27:29)), as.factor)
```

Valors nuls

La distribució de valors desconeguts per pacient és:

```
table(apply(hcc, 1, function(x) sum(is.na(x))))
```

```
##
```

```
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 16 18 22 23
##  8 16 15 28 19 21 15  7 11 12  4  2  1  1  1  1  1  1  1
```

El percentatge de NAs per variable és:

```
funNA <- function(a, n){
  a = round(100*a/n,1)
}

totNA <- hcc %>% select(everything()) %>%  #
  summarise_all(funs(sum(is.na(.))))
perNA <- totNA %>% mutate_all(funNA, n= nrow(hcc))
tauNA <- totNA %>% bind_rows(perNA)
tauNA <- as_tibble(t(tauNA), rownames = "Variable") %>%
  rename(`total NA` = V1, `%NA` = V2) %>%
  arrange(-`total NA`)

tau <- cbind(tauNA[1:25,],tauNA[26:50,])

kable(x = tau, format = "latex", caption = "Variables amb NA", booktabs = TRUE) %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
```

Table 1: Variables amb NA

Variable	total NA	%NA	Variable	total NA	%NA
Oxygen Saturation (%)	80	48.5	Total Bilirubin(mg/dL)	5	3.0
Ferritin (ng/mL)	80	48.5	Liver Metastasis	4	2.4
Iron	79	47.9	International Normalised Ratio	4	2.4
Packs of cigarets per year	53	32.1	Alanine transaminase (U/L)	4	2.4
Esophageal Varices	52	31.5	Diabetes	3	1.8
Grams of Alcohol per day	48	29.1	Arterial Hypertension	3	1.8
Direct Bilirubin (mg/dL)	44	26.7	Portal Vein Thrombosis	3	1.8
Smoking	41	24.8	Haemoglobin (g/dL)	3	1.8
Hepatitis B e Antigen	39	23.6	Mean Corpuscular Volume	3	1.8
Endemic Countries	39	23.6	Leukocytes(G/L)	3	1.8
Hepatitis B Core Antibody	24	14.5	Platelets	3	1.8
Hemochromatosis	23	13.9	Aspartate transaminase (U/L)	3	1.8
Nonalcoholic Steatohepatitis	22	13.3	Gamma glutamyl transferase (U/L)	3	1.8
Major dimension of nodule (cm)	20	12.1	Alkaline phosphatase (U/L)	3	1.8
Symptoms	18	10.9	Chronic Renal Insufficiency	2	1.2
Hepatitis B Surface Antigen	17	10.3	Radiological Hallmark	2	1.2
Splenomegaly	15	9.1	Ascites degree	2	1.2
Human Immunodeficiency Virus	14	8.5	Number of Nodules	2	1.2
Portal Hypertension	11	6.7	Encephalopathy degree	1	0.6
Total Proteins (g/dL)	11	6.7	Gender	0	0.0
Obesity	10	6.1	Alcohol	0	0.0
Hepatitis C Virus Antibody	9	5.5	Cirrhosis	0	0.0
Alpha-Fetoprotein (ng/mL)	8	4.8	Age at diagnosis	0	0.0
Creatinine (mg/dL)	7	4.2	Performance Status	0	0.0
Albumin (mg/dL)	6	3.6	Class Attribute	0	0.0

Només hi ha 8 pacients amb les dades completes, faltant a la majoria de pacients entre 2 i 9 dades. Fins i

tot hi ha pacients 13 pacients amb més de 10 dades desconegudes.

Estudiant la distribució dels valors desconeguts per variable veiem que només hi ha 6 variables amb totes les dades íntegres. Amb més de l'10% de dades desconegudes hi ha 16 de les 50 varibles (un 32%), destacant 9 varibles amb entre el 20 i el 50% de les seves dades desconegudes, com són la saturació d'oxigen o els nivells de ferritina en sang. Els valors NA poden seguir una distribució a l'atzar de manera que la proporció dels esperats en cada classe hauria de ser similar. En cas contrari, la correcció dels valors desconeguts podria provocar un biaix cap a un dels dos grups. Vegem com es distribueixen en les variables els valors missing i si hi ha difrències significatives depenent de la clase.

```
## REVISAR: Creo que así miramos si hay diferencias de la propia variable no de los valores NA de la va

# Valoració de la distribució dels NA a les variables amb >10% de NA amb
# respecte la variable classe
hcc_colNA<-names(hcc[hcc_factor][colSums(is.na(hcc[hcc_factor]))>10])
totFactor <- names(hcc[hcc_factor])
testSig <- tibble()

for (i in totFactor){
  test<-prop.test(table(hcc[,i],hcc$`Class Attribute`))
  if (test$p.value<0.05){
    testSig <- testSig %>% bind_rows(c(Clase = i, p_value = test$p.value))
  }
}

t

kable(x = testSig, format = "latex", caption = "", booktabs = TRUE, digits = 4) %>%
  kable_styling(latex_options = c("HOLD_position"))
```

```
# Valoració de la distribució dels NA a les variables amb respecte la variable classe
cero  <- hcc %>% filter(`Class Attribute` == 0)
uno  <- hcc %>% filter(`Class Attribute` == 1)

probTest <- tibble()
for (i in names(hcc[,hcc_factorT])) {
  if (sum(is.na(hcc[,i]))>0){
    casos<-c(sum(is.na(cero[,i])),sum(is.na(uno[,i])))
    long<-c(length(cero[,i]),length(uno[,i]))
    test<-prop.test(x=casos,n=long)
    probTest <- probTest %>%
      bind_rows(c(Clase = i, p_value = test$p.value,
                  prob_0=(casos/long)[1], prob_1=(casos/long)[2],
                  numNA_0=casos[1], numNA_1=casos[2]))
  }
}

testSig <- probTest %>% filter(p_value <= 0.05) %>%
  mutate_at(.vars = c("p_value", "prob_0","prob_1","numNA_0", "numNA_1"), as.numeric)

kable(x = testSig, format = "latex",
      caption = "Variables amb una difència significativa de NA entre els grups de la classe",
      booktabs = TRUE, digits = 4) %>%
  kable_styling(latex_options = c("HOLD_position"))
```

Table 2: Variables amb una difència significativa de NA entre els grups de la classe

Clase	p_value	prob_0	prob_1	numNA_0	numNA_1
Symptoms	0.0058	0.0159	0.1667	1	17
Hemochromatosis	0.0019	0.2540	0.0686	16	7
Esophageal Varices	0.0084	0.4444	0.2353	28	24

A la variable `Symptoms` s'observen molts NA entre els pacients que sobreviuen. Els pacients que no sobreviuen solen presentar molta simptomatologia i aquesta es registra. En canvi, els pacients que no presenten símptomes, poden no registrar-se com a negatiu a aquesta variable, existint un biaix d'informació.

Igualment succeix a les variables `Hemochromatosis` i `Esophageal Varices`. Els pacients afectats es registren i probablement presenten tases mes altes de mortalitat. En canvi, molts pacients es desconeixerà si presenten hemocromatosis o varius, però probablement no la patiran, i tindran tases de supervivència superiors.

Corregir aquest valors desconeguts cap a la moda condicionarà un biaix. Per tant, per corregir els valors NA es farà:

- A les variable qualitatives sense diferències significatives en el valor desconeguts, s'asignarà el valor més freqüent a cada variable.
- A les variables qualitatives amb diferències significatives entre els valors deconeguts, s'asignarà el valor més pròxim utilitzant l'algoritme kNN.
- A les variables quantitatives s'asignarà la mitjana de la variable. Per tal de no tenir una mitjana condicionada per valors erronis extrems, es corregiran abans de l'asignació del valor mitjà als valors desconguts.

Correcció valors nuls de variables categòriques

```
# Correcció de NA de variables categóriques amb la moda

# Funció moda
my_mode <- function(x) {
  unique_x <- unique(x)
  tabulate_x <- tabulate(match(x, unique_x))
  unique_x[tabulate_x == max(tabulate_x)]
}

# Correcció NA categórics moda
for (i in hcc_factorT)
{
  hcc[is.na(hcc[,i]),i]<-(my_mode(hcc[,i]))
}


# Correcció NA categòrics kNN

suppressWarnings(suppressMessages(library(VIM)))
hcc$Symptoms<-kNN(hcc)$Symptoms
hcc$Hemochromatosis<-kNN(hcc)$Hemochromatosis
hcc$`Esophageal Varices`<-kNN(hcc)$`Esophageal Varices`
```

Correcció valors nuls de variables quantitatives

Existeixen dues variables amb valors extranys, incompatibles amb la vida; son `Leukocytes` i `Platelets` La gran majoria dels valors a la variable `Leukocytes` esgan per sota de 100, que és l'esperat. Valors majors son pràcticament impossibles. Els valors d'aquesta variable es solen expressar sobre mm3 pel que solen tenir valors múltiples de 1000, d'aquí la probable confusió amb els valors extrems trobats. Es corregiran modificant les unitats d'aquest valors.

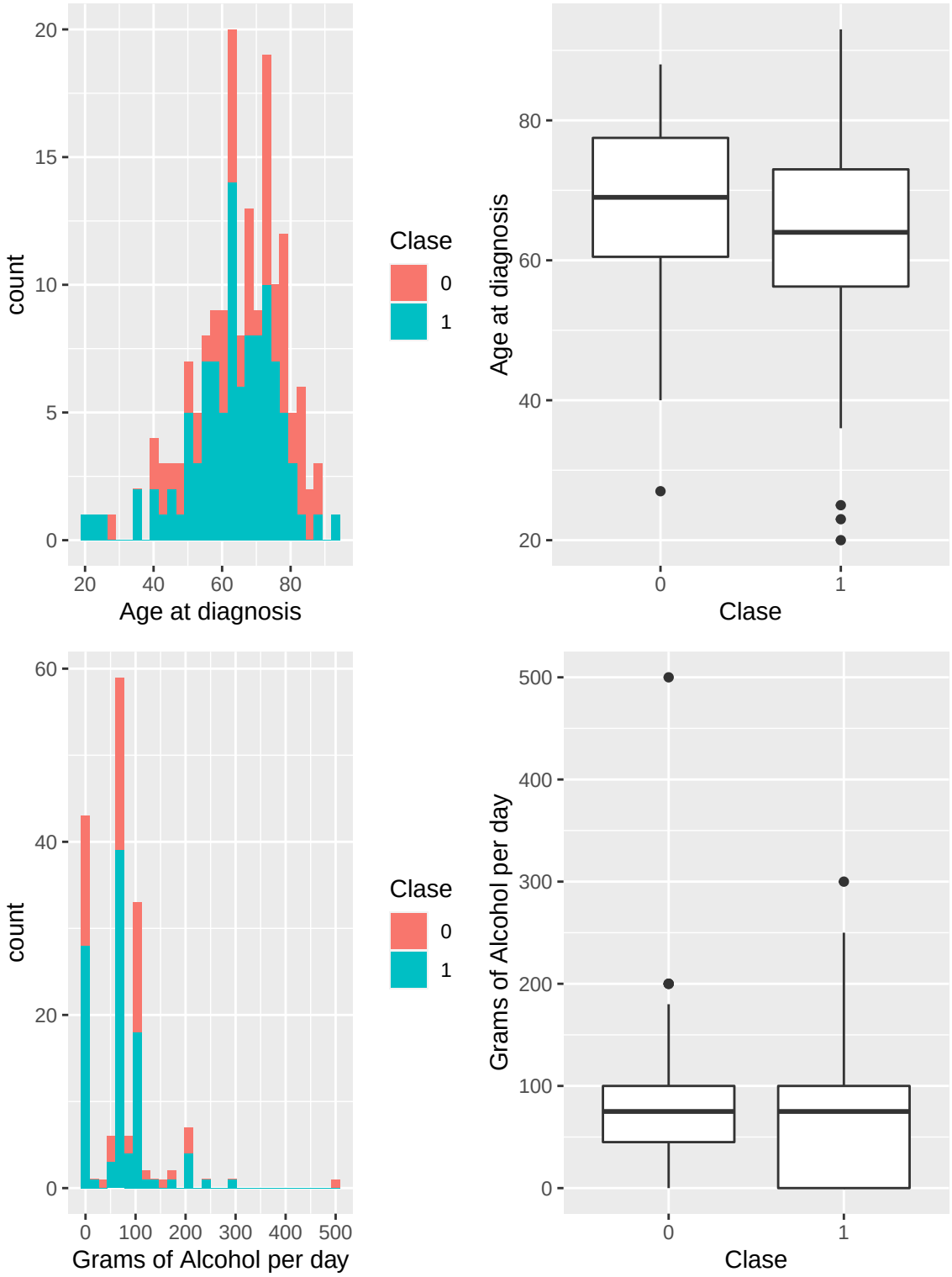
Amb respecte `Platelets`, l'error és similar al trobat a l'anterior variable.

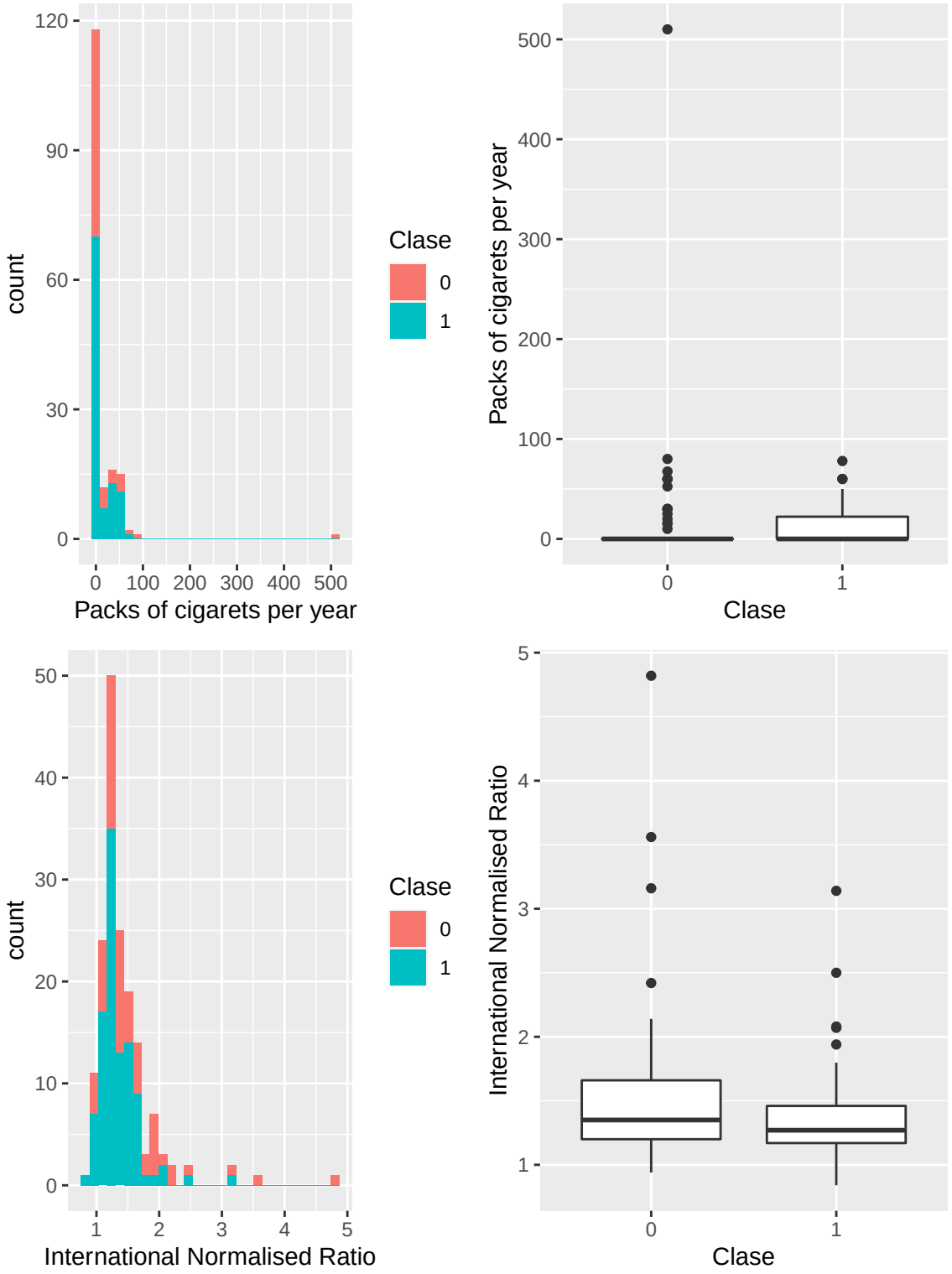
Es corregeix els errors i s'assigna la mitjana als valors desconeguts

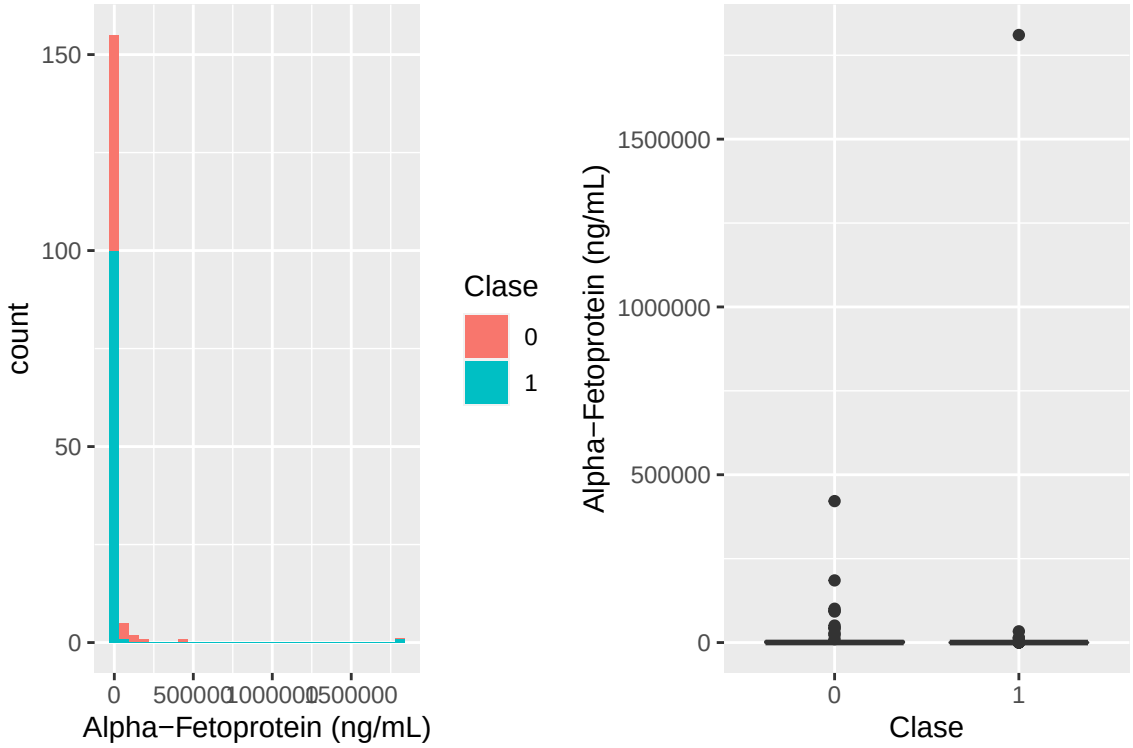
```
# Correcció valors leucocitosi i plaquetes
hcc$`Leukocytes(G/L)`[!is.na(hcc$`Leukocytes(G/L)`)&hcc$`Leukocytes(G/L)`>100]<-hcc$`Leukocytes(G/L)`[!

hcc$Platelets[!is.na(hcc$Platelets)&hcc$Platelets<1000]<-hcc$Platelets[!is.na(hcc$Platelets)&hcc$Platel

# Correcció NA amb mitjana
for (i in hcc_num){
  hcc[is.na(hcc[,i]),i]<-median(hcc[,i],na.rm = TRUE)
}
```

Anàlisi de les dades

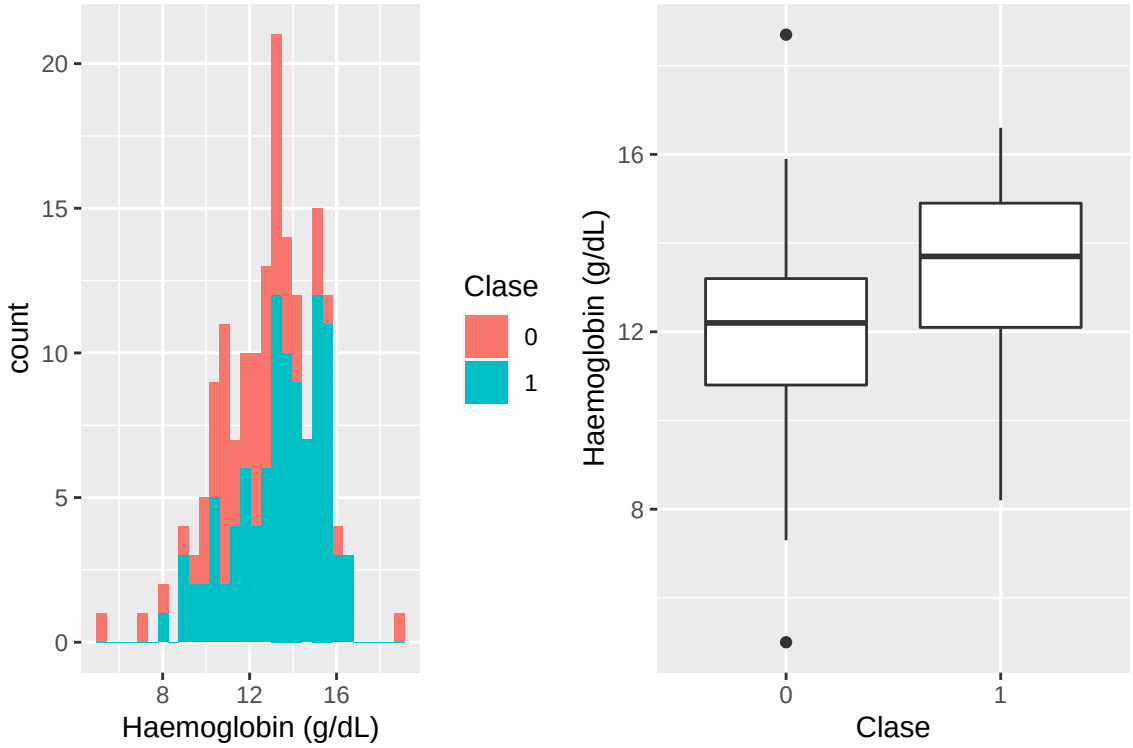
```
# Estudi distribució variables Numeriques
library(gridExtra)

for (i in hcc_num) {

  a1<-ggplot(hcc, aes(x=hcc[,i], fill=hcc$`Class Attribute`))+xlab(names(hcc)[i])+labs(fill="Clase")+ge
  a2<-ggplot(hcc, aes(x=hcc$`Class Attribute`,y=hcc[,i]))+ylab(names(hcc)[i])+xlab("Clase")+geom_boxplo

  grid.arrange(a1,a2,nrow=1)
}
```


count
Age at diagnosis
Clase
0
1
Age at diagnosis
Clase
count
Grams of Alcohol per day
Clase
0
1
Grams of Alcohol per day
Clase

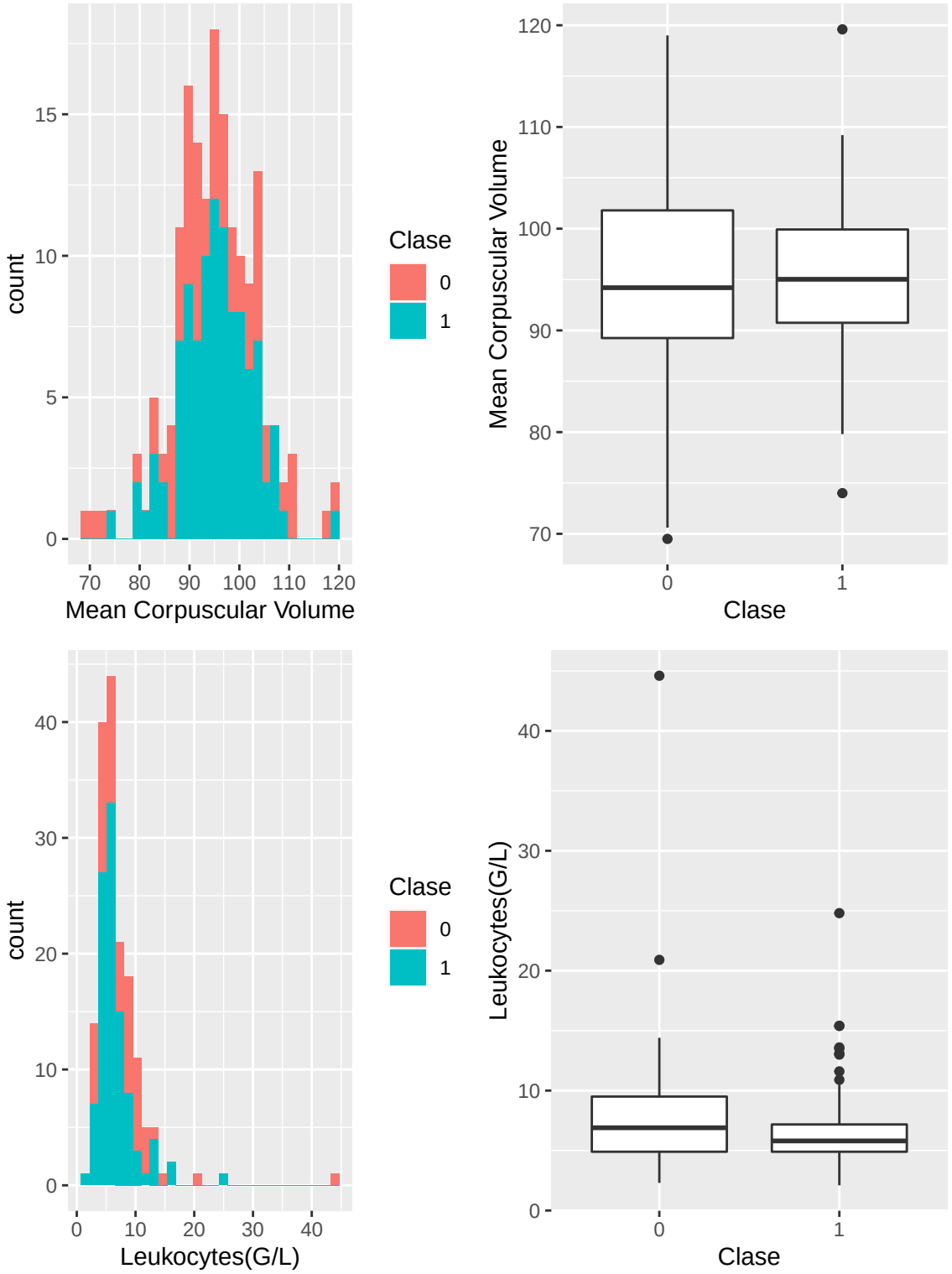

count
Packs of cigarets per year
Clase
0
1
Packs of cigarets per year
Clase
International Normalised Ratio
International Normalised Ratio
Clase

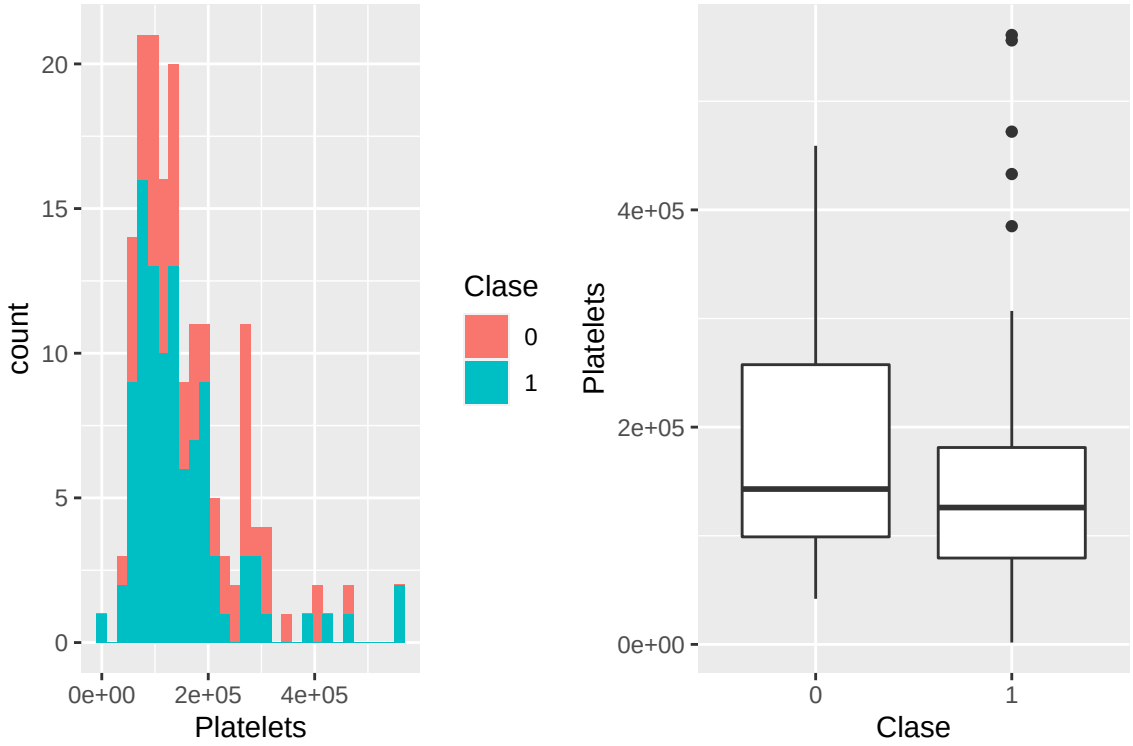

150
100
50
0
count
0
500000
1000000
1500000
Alpha−Fetoprotein (ng/mL)
Clase
0
1

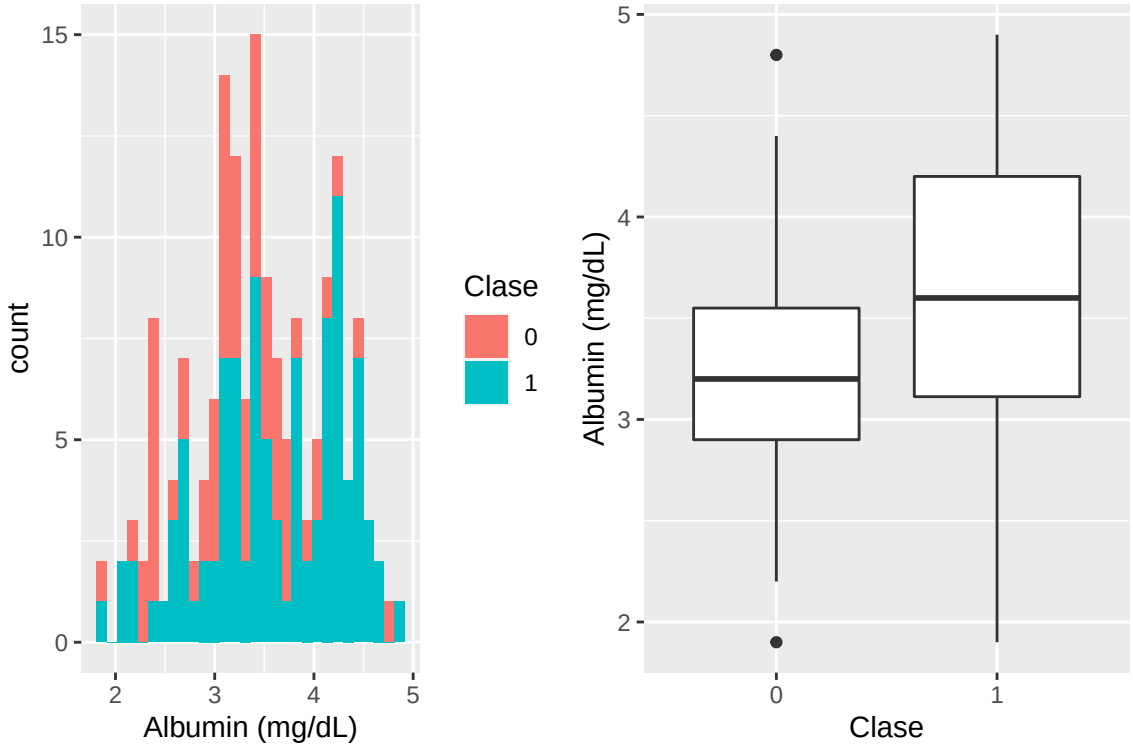
1500000
1000000
500000
0
Alpha−Fetoprotein (ng/mL)
0
1
Clase

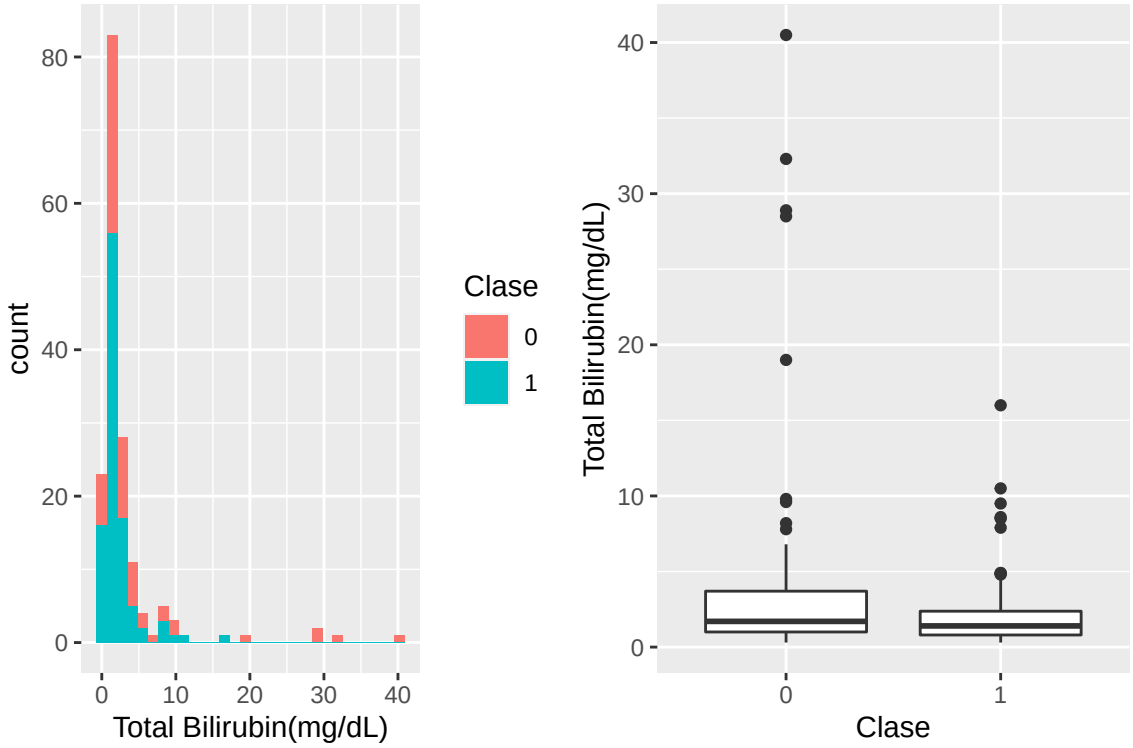

20
15
10
5
0
count
8
12
16
Haemoglobin (g/dL)
Clase
0
1

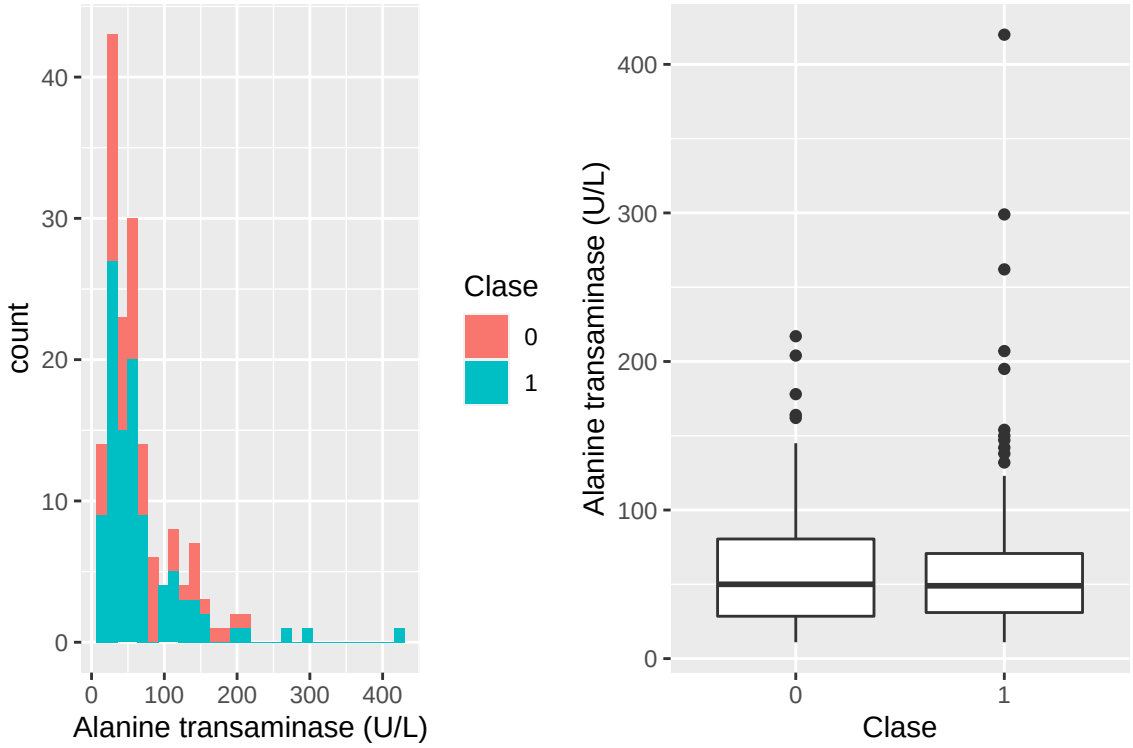
16
12
8
Haemoglobin (g/dL)
0
1
Clase

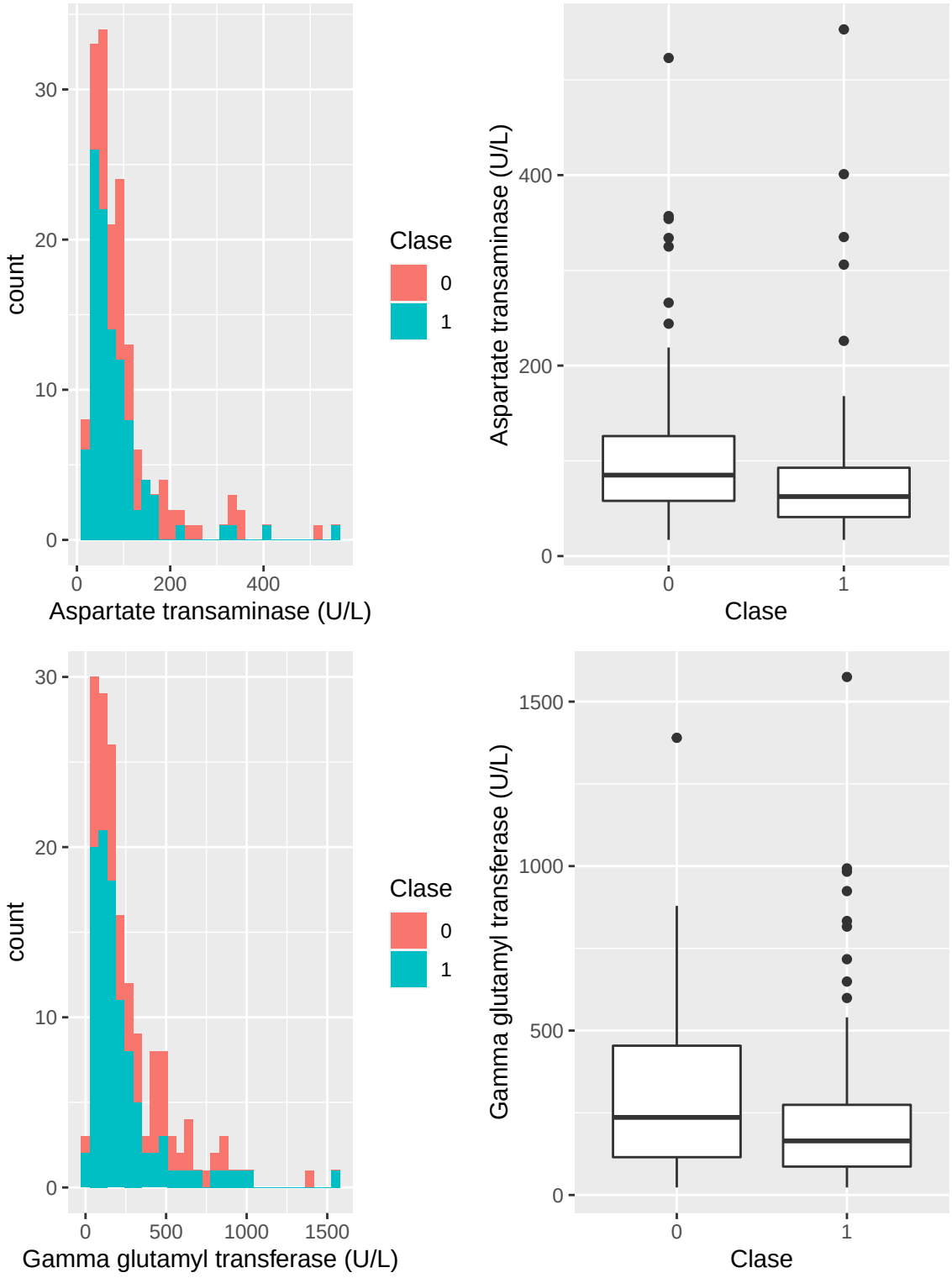

count
Mean Corpuscular Volume
Clase
0
1
Mean Corpuscular Volume
Clase
count
Leukocytes(G/L)
Clase
0
1
Leukocytes(G/L)
Clase

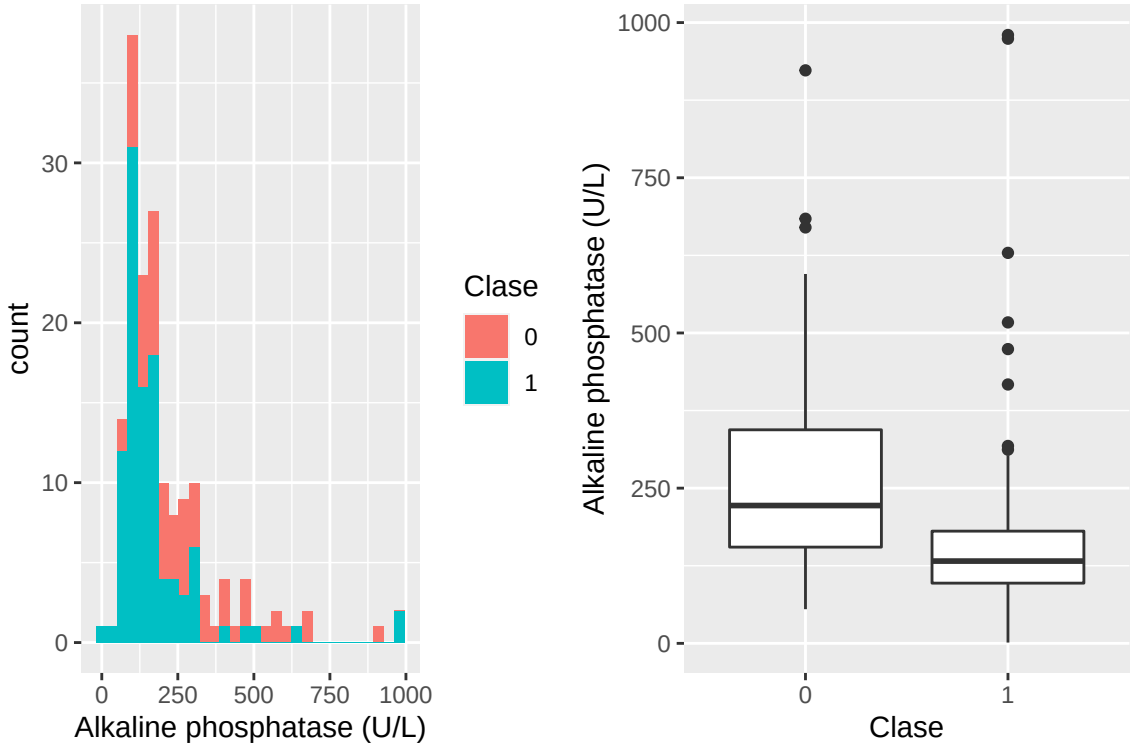

count
Platelets
0e+00
2e+05
4e+05
0
5
10
15
20
Clase
0
1
Platelets
Clase

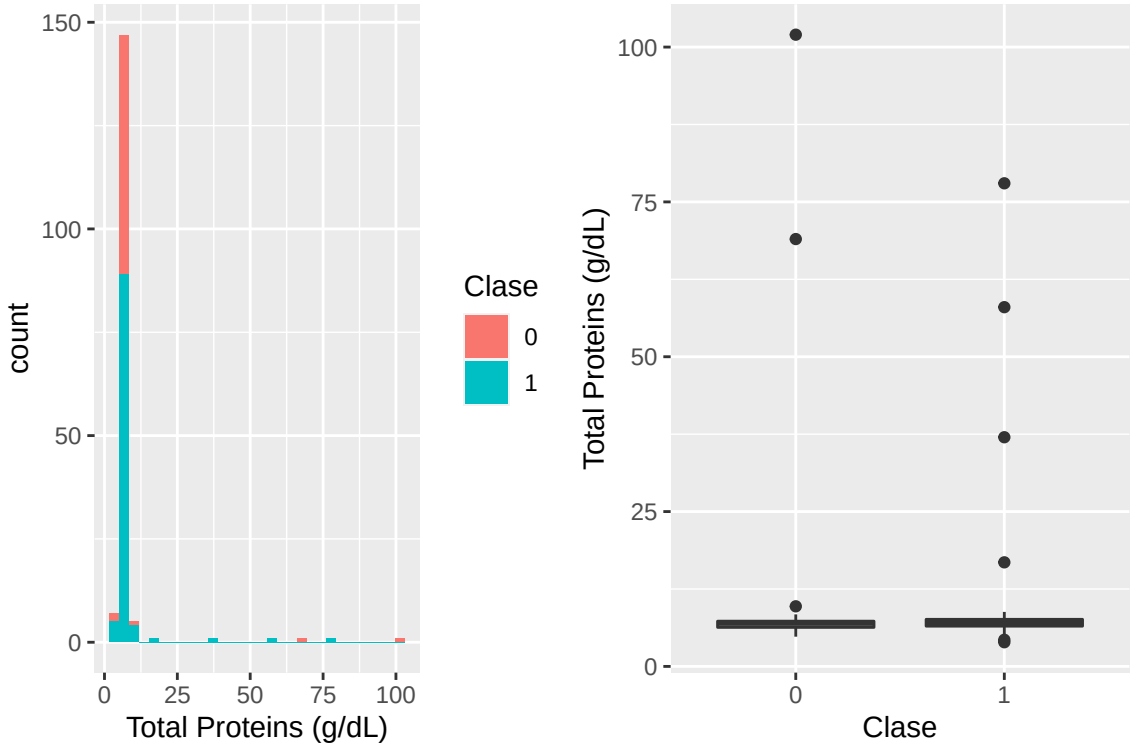

count
Albumin (mg/dL)
2
3
4
5
0
5
10
15
Clase
0
1
Albumin (mg/dL)
Clase

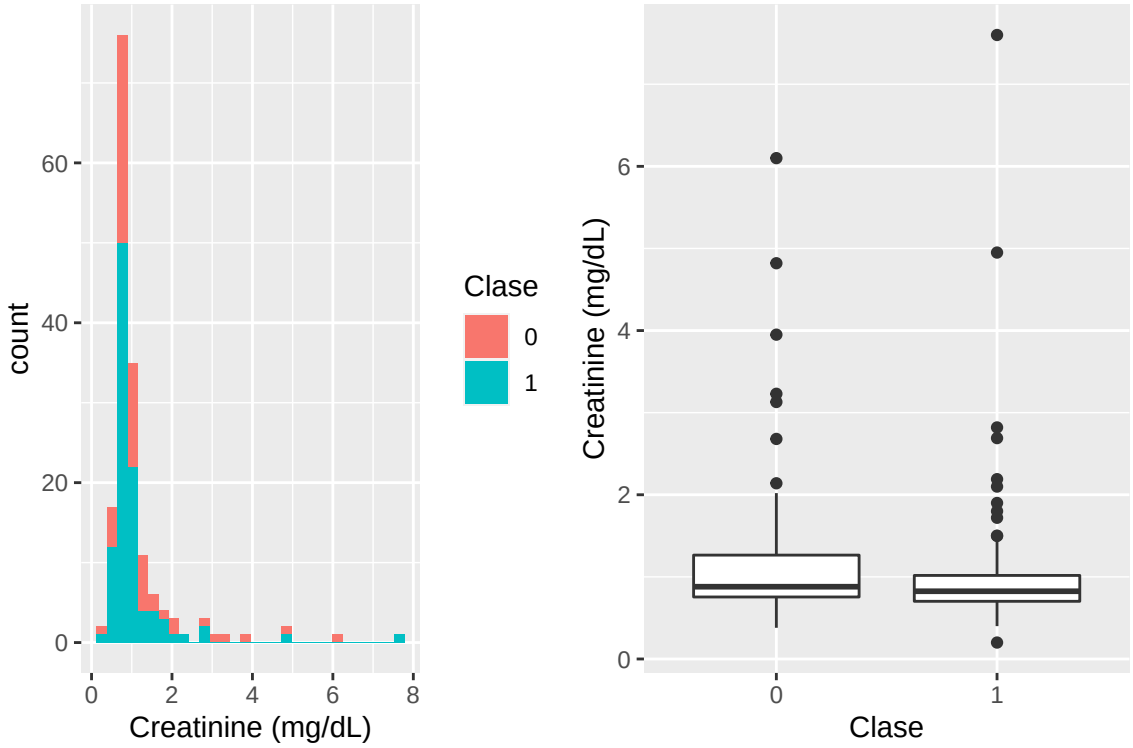

count
Total Bilirubin(mg/dL)
Clase
0
1
Total Bilirubin(mg/dL)
Clase

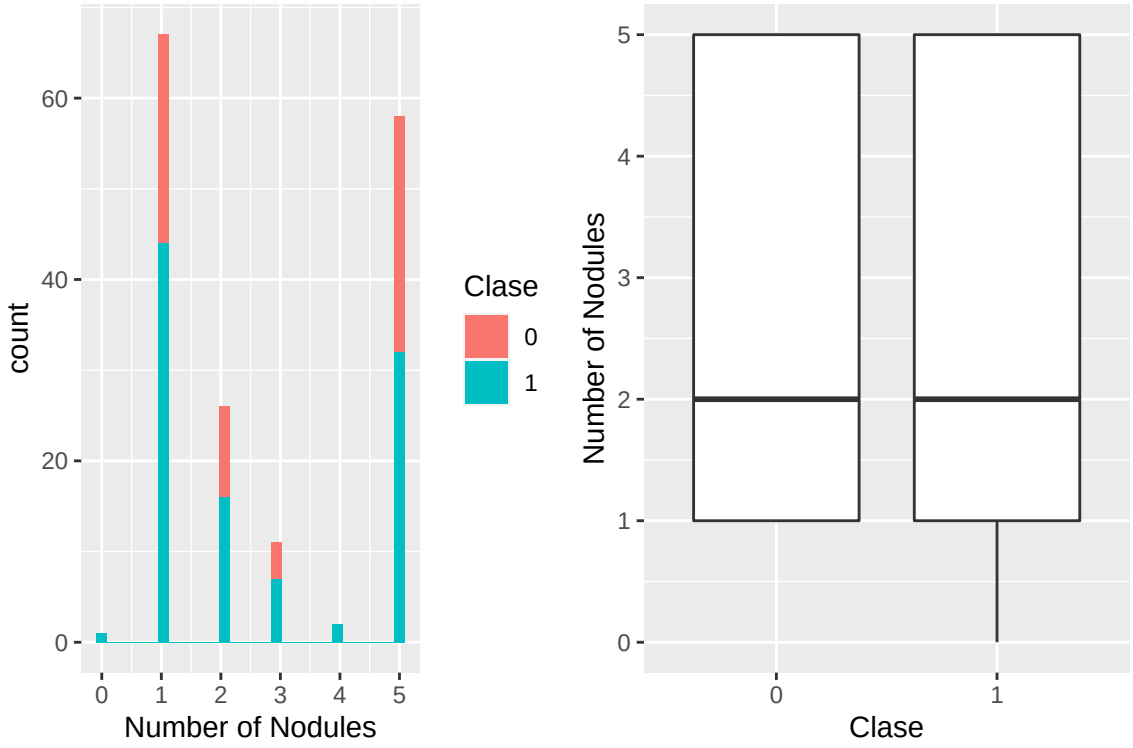

count
Alanine transaminase (U/L)
Clase
0
1
Alanine transaminase (U/L)
Clase

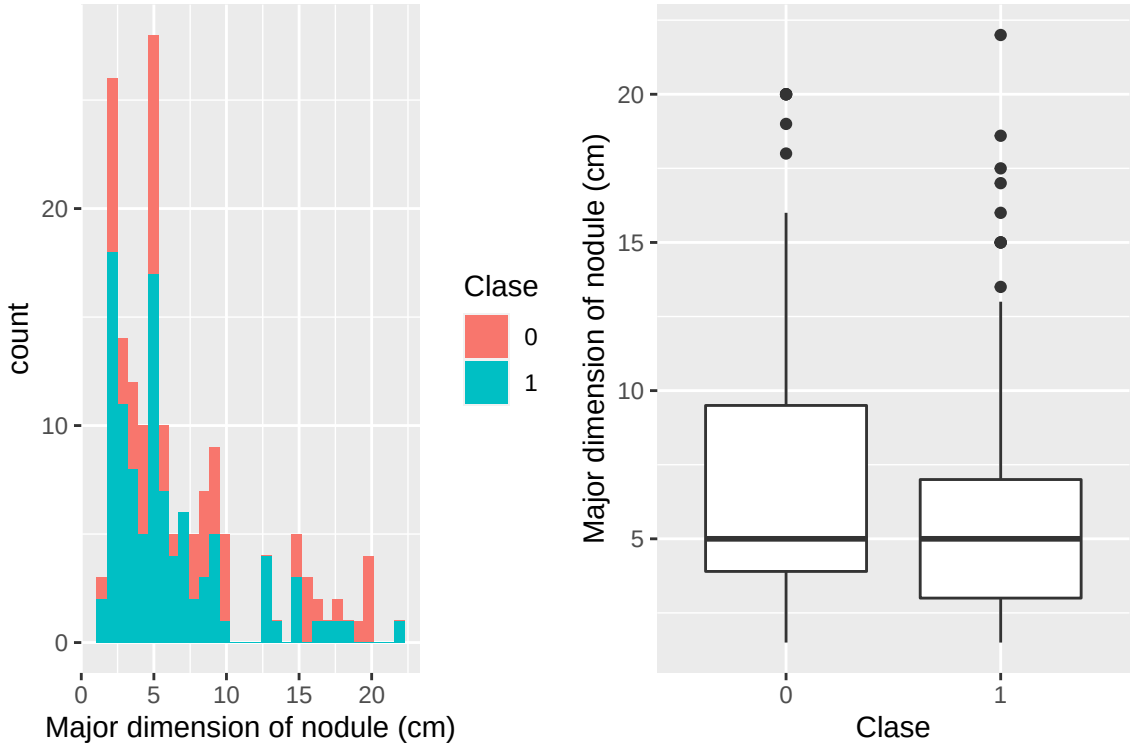

count
30
20
10
0
0
200
400
Aspartate transaminase (U/L)
Clase
0
1
Aspartate transaminase (U/L)
400
200
0
0
1
Clase
count
30
20
10
0
0
500
1000
1500
Gamma glutamyl transferase (U/L)
Clase
0
1
Gamma glutamyl transferase (U/L)
1500
1000
500
0
0
1
Clase

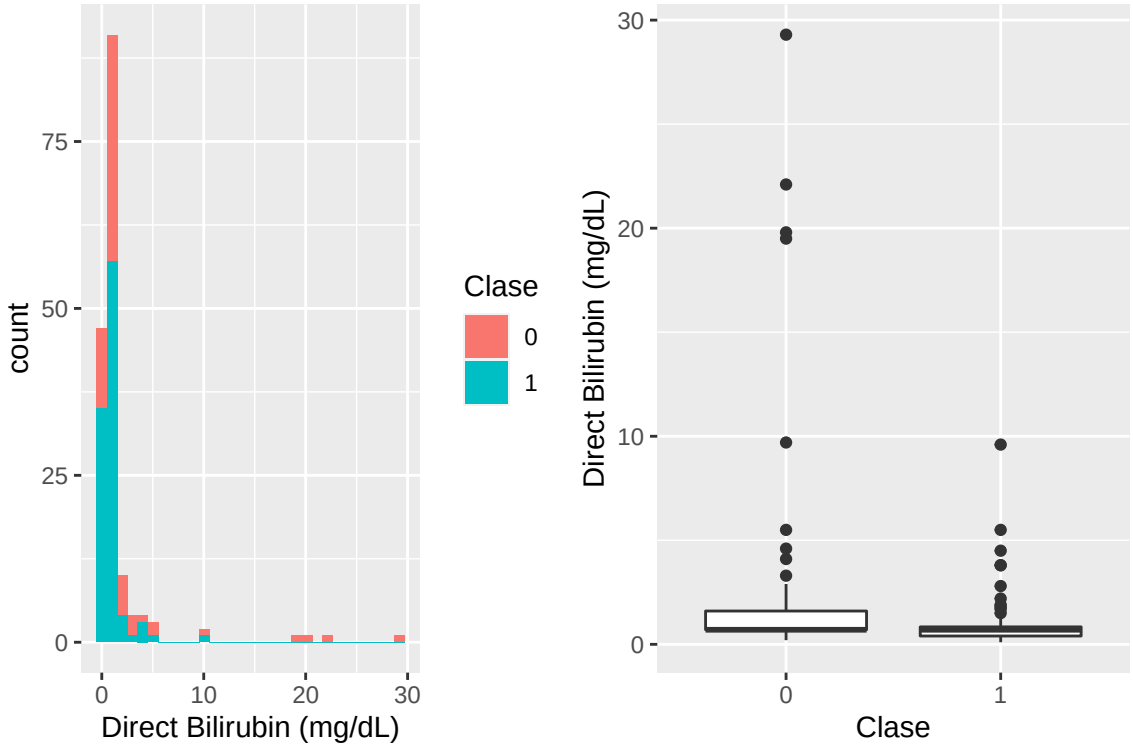

count
30
20
10
0
0
250
500
750
1000
Alkaline phosphatase (U/L)
Clase
0
1
Alkaline phosphatase (U/L)
1000
750
500
250
0
0
1
Clase

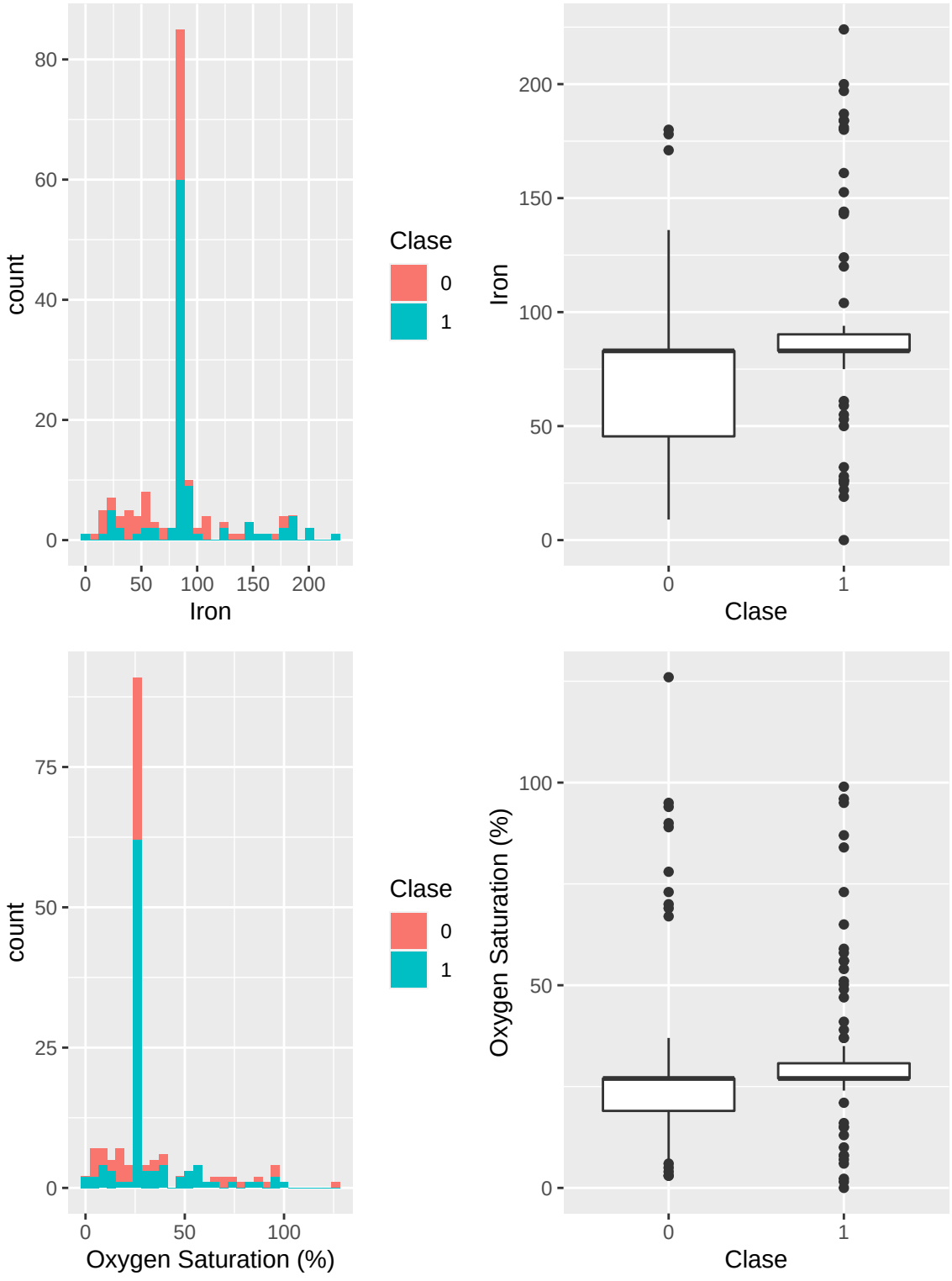

count
150
100
50
0
0
25
50
75
100
Total Proteins (g/dL)
Clase
0
1
Total Proteins (g/dL)
100
75
50
25
0
0
1
Clase

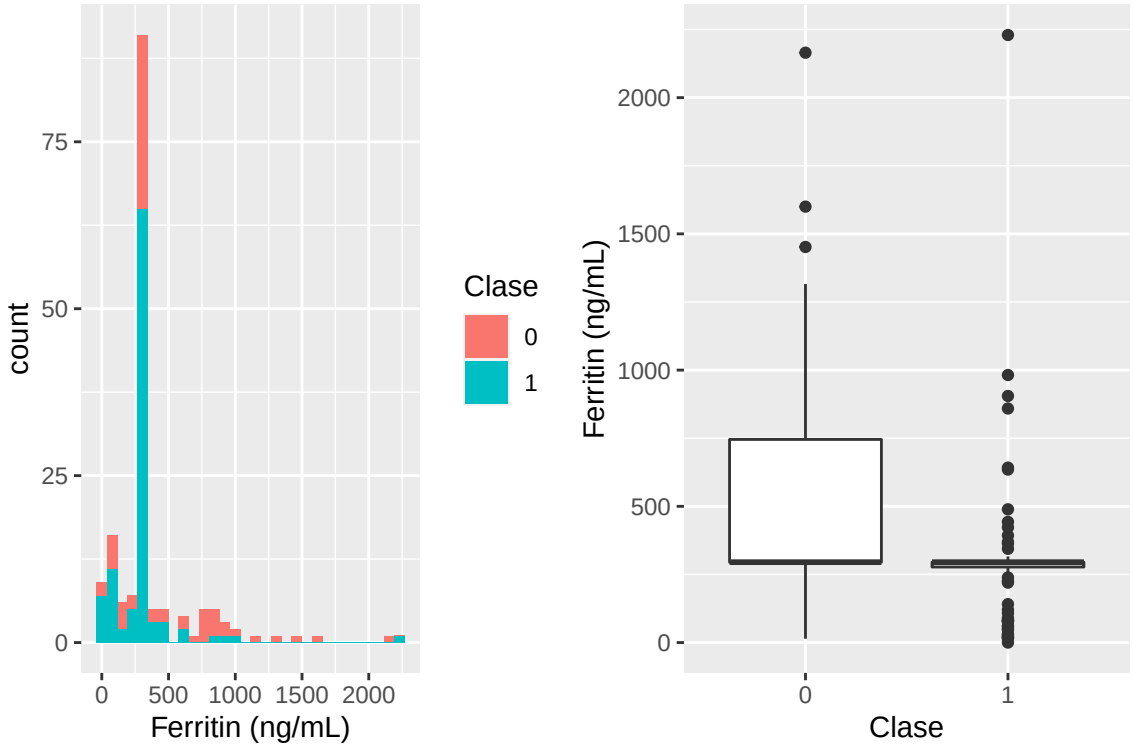

count
Creatinine (mg/dL)
Clase
0
1
Creatinine (mg/dL)
Clase


count
Number of Nodules
Clase
0
1
Number of Nodules
Clase


count
20
10
0
0
5
10
15
20
Major dimension of nodule (cm)
Clase
0
1
Major dimension of nodule (cm)
20
15
10
5
0
1
Clase


count
75
50
25
0
0
10
20
30
Direct Bilirubin (mg/dL)
Clase
0
1
Direct Bilirubin (mg/dL)
30
20
10
0
0
1
Clase


count
Iron
Clase
0
1
Iron
Clase
Oxygen Saturation (%)
Oxygen Saturation (%)

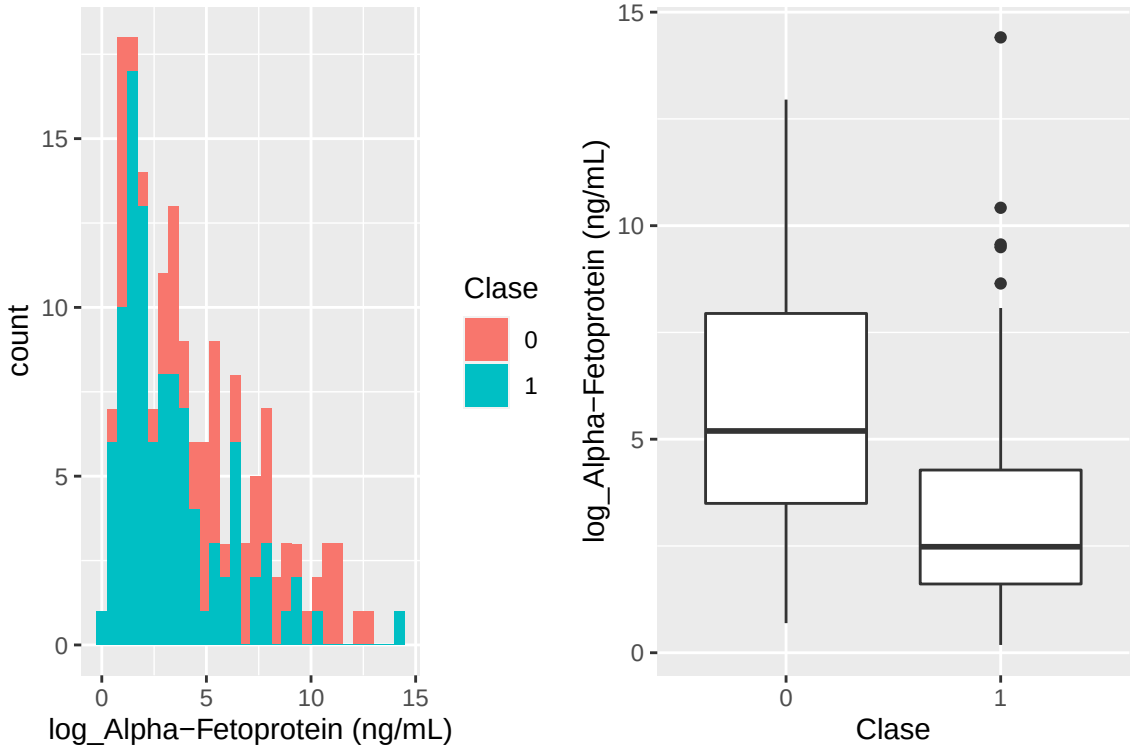


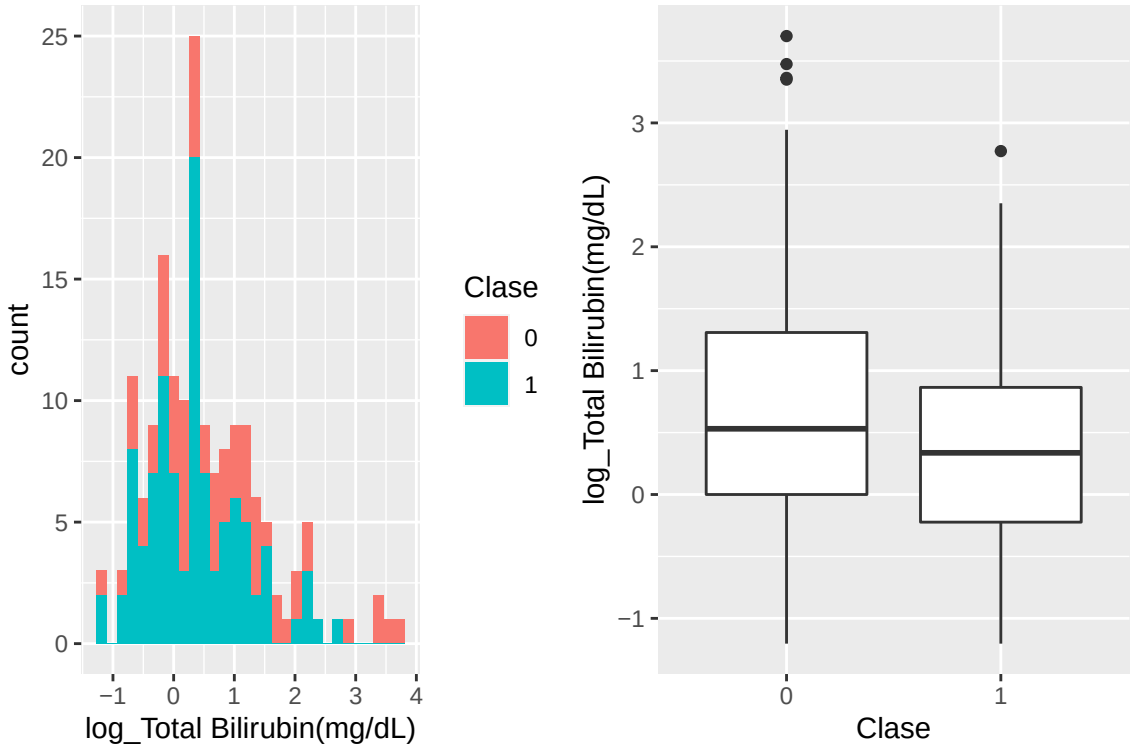
Crida l'atenció vàries variables amb una distribució molt desplaçada cap a valors baixos però amb valors extrems alts. Molts son valors de laboratori, i sembla que s'adapten més a distribucions logarítmiques, per el que es modificaran. Aquestes variables son:

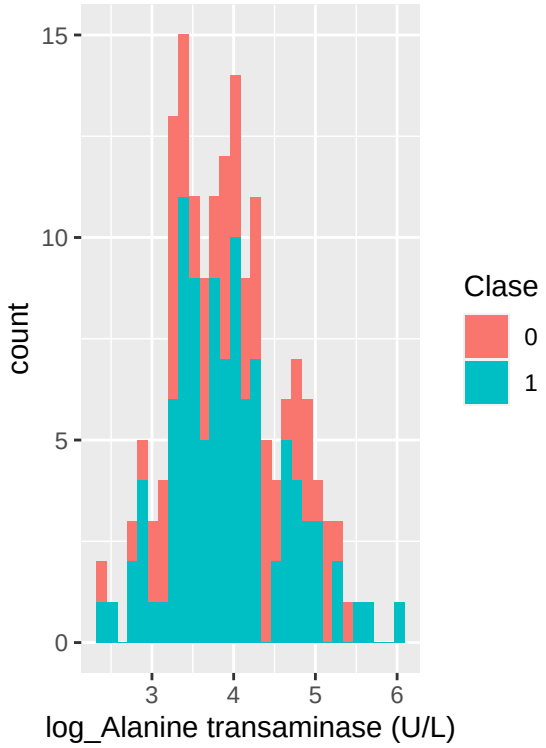
- Alpha-Fetoprotein (ng/mL)
- Total Bilirubin(mg/dL)
- Alanine transaminase (U/L)
- Aspartate transaminase (U/L)
- Gamma glutamyl transferase (U/L)
- Alkaline phosphatase (U/L)
- Total Proteins (g/dL)
- Creatinine (mg/dL)
- Direct Bilirubin (mg/dL)

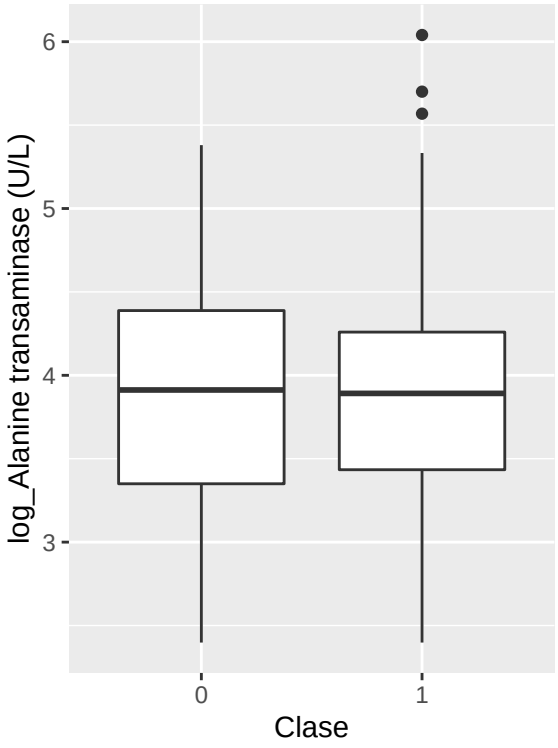
```
# Transformació logarítmica de variables numériques
hcc_log<-c(31,37:43, 46)
for (i in hcc_log){
  hcc[,i]<-log(hcc[,i])
  names(hcc)[i]<-paste("log_",names(hcc)[i], sep="" )
}
for (i in hcc_log) {

  a1<-ggplot(hcc, aes(x=hcc[,i], fill=hcc$`Class Attribute`))+xlab(names(hcc)[i])+labs(fill="Clase")+ge
  a2<-ggplot(hcc, aes(x=hcc$`Class Attribute`,y=hcc[,i]))+ylab(names(hcc)[i])+xlab("Clase")+geom_boxplo
```

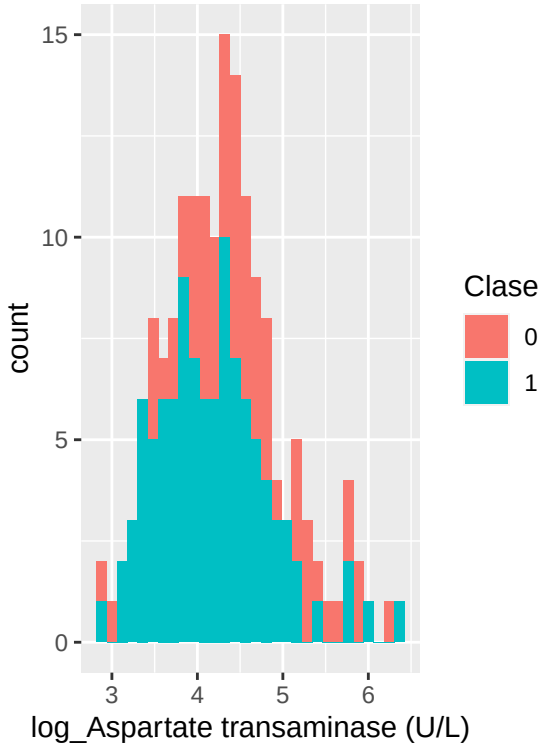
```
  grid.arrange(a1,a2,nrow=1)
}
```

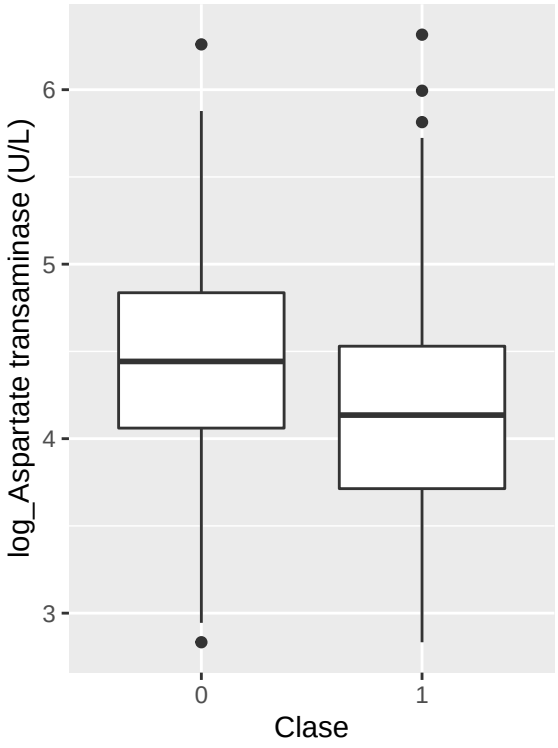


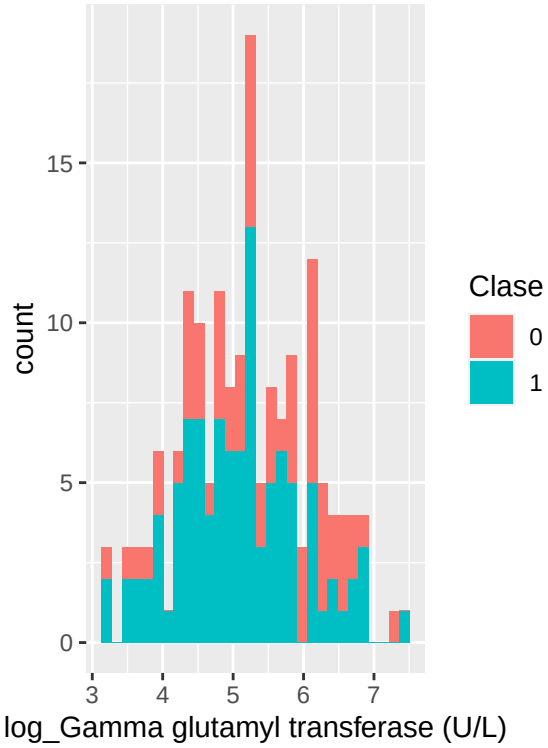


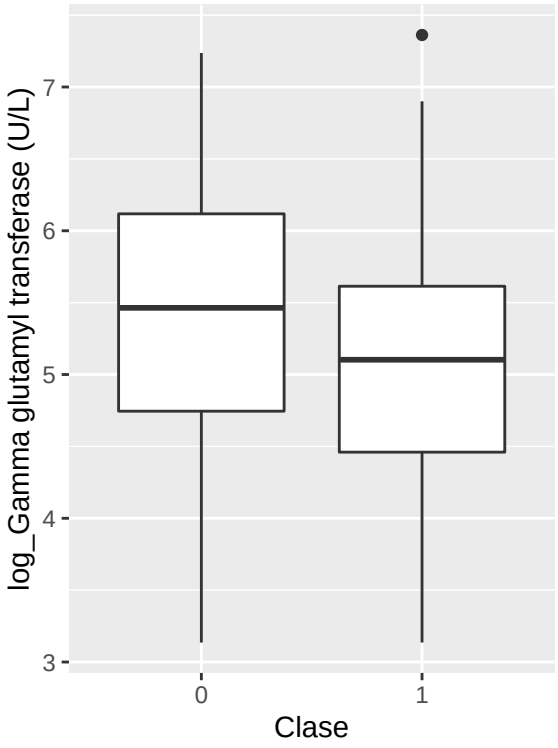

count
15
10
5
0
3
4
5
6
log_Alanine transaminase (U/L)
Clase
0
1


log_Alanine transaminase (U/L)
6
5
4
3
0
1
Clase

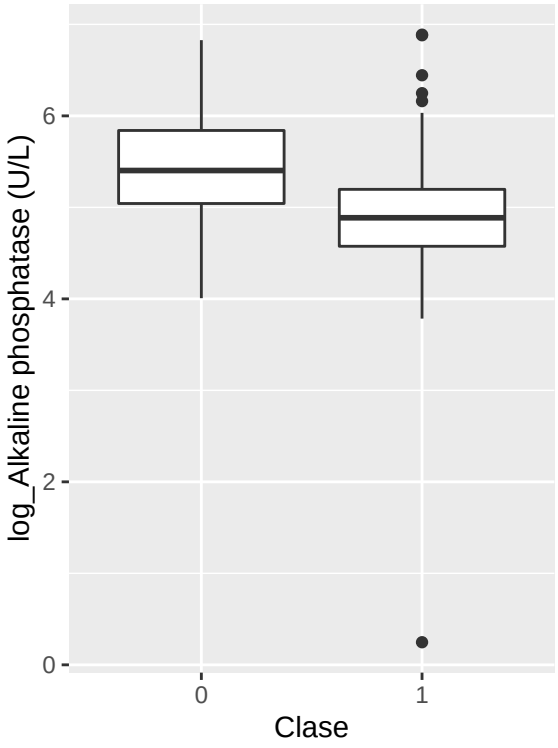

count
15
10
5
0
3
4
5
6
log_Aspartate transaminase (U/L)
Clase
0
1

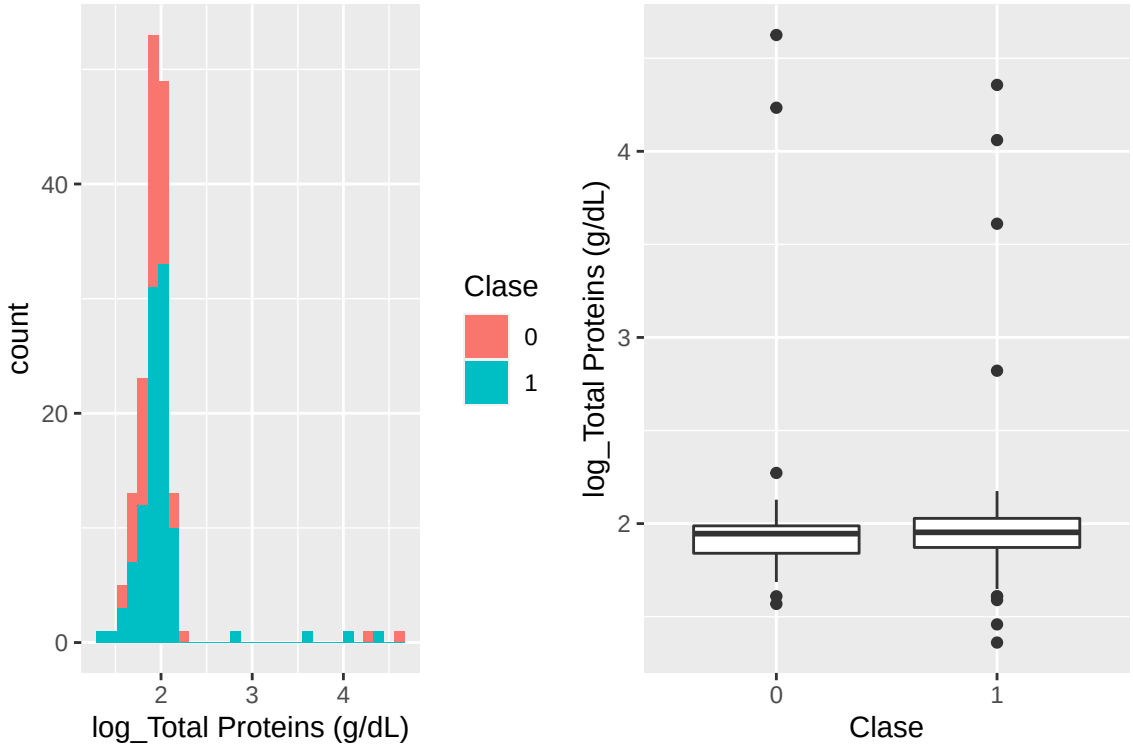

log_Aspartate transaminase (U/L)
6
5
4
3
0
1
Clase

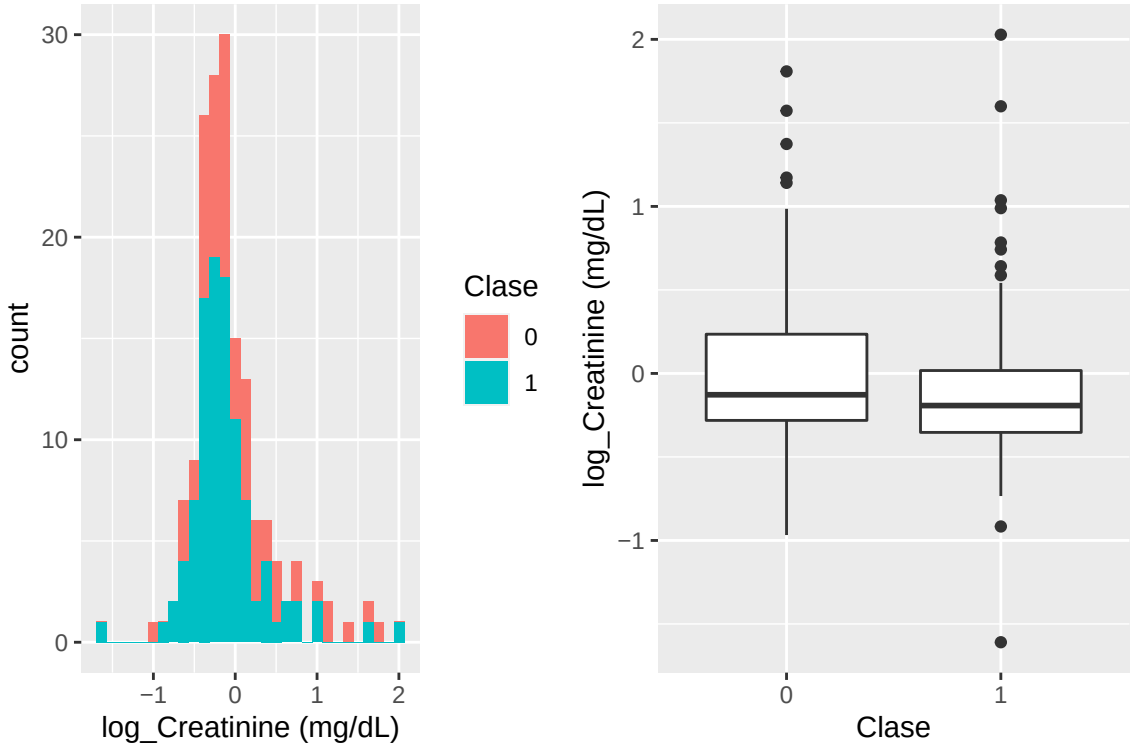

count
15
10
5
0
3
4
5
6
7
log_Gamma glutamyl transferase (U/L)
Clase
0
1

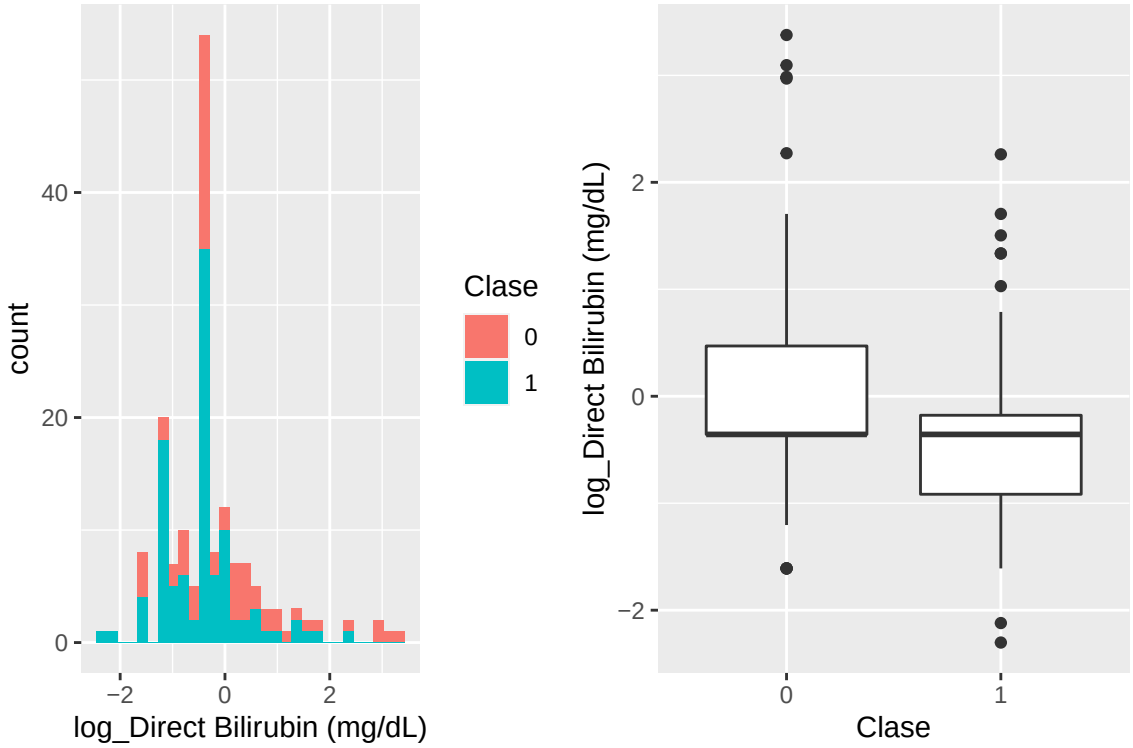

log_Gamma glutamyl transferase (U/L)
7
6
5
4
3
0
1
Clase


count
20
10
0
0
2
4
6
log_Alkaline phosphatase (U/L)
Clase
0
1

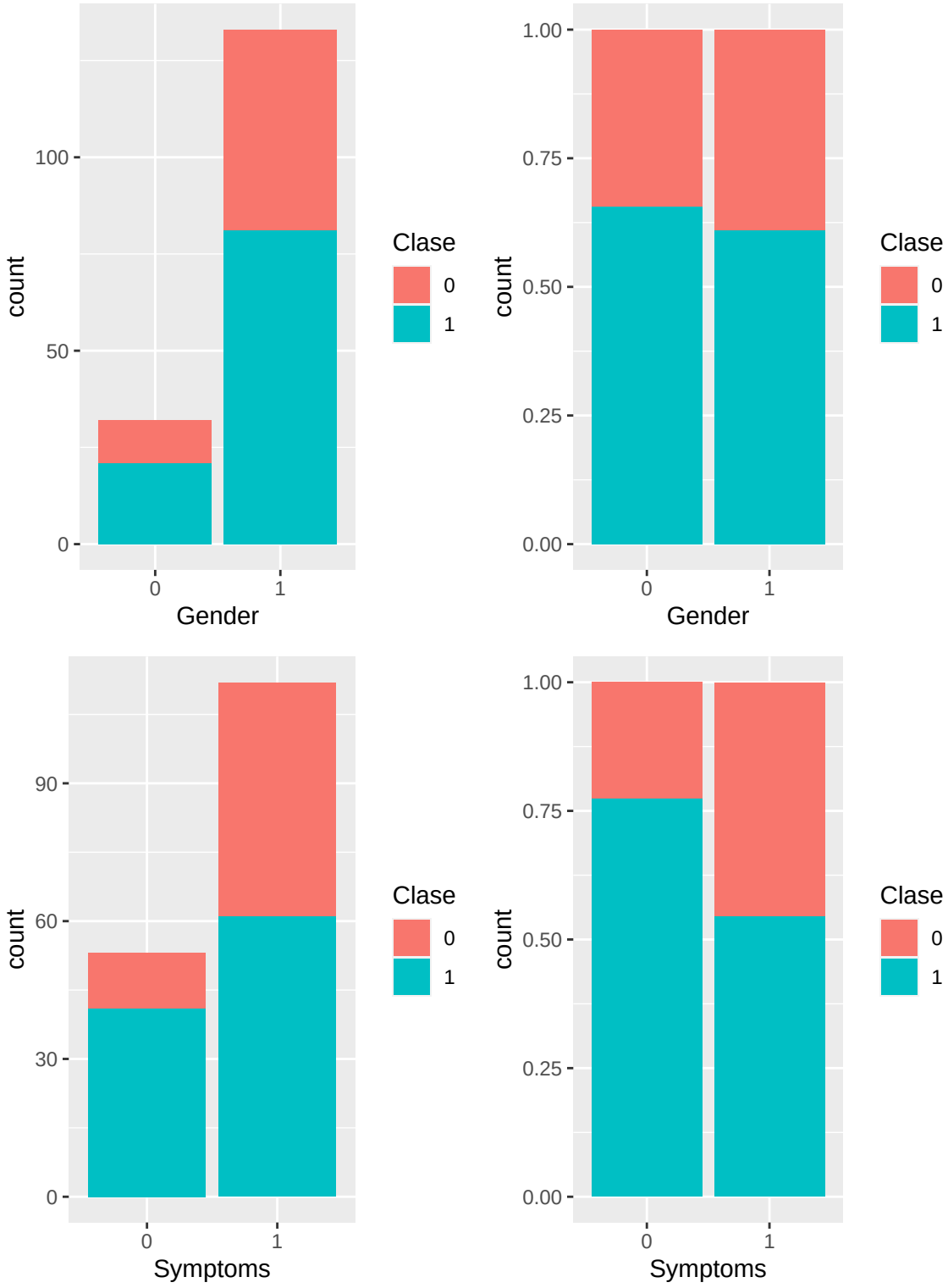

log_Alkaline phosphatase (U/L)
6
4
2
0
0
1
Clase

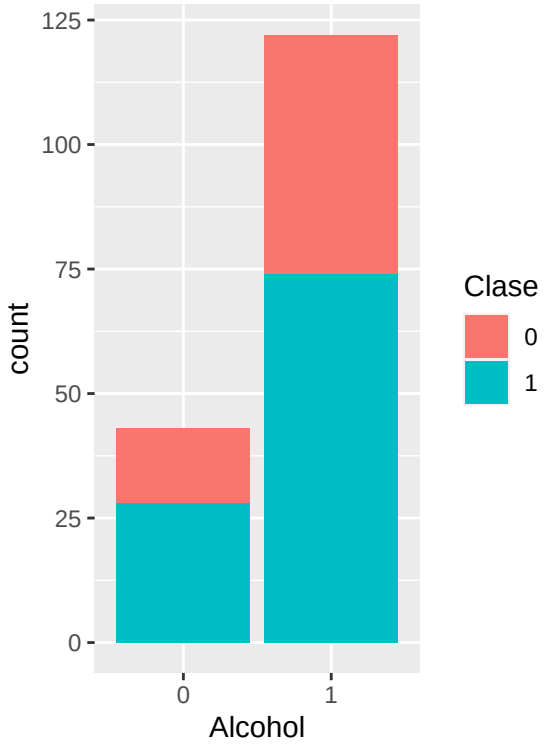

count
40
20
0
2
3
4
log_Total Proteins (g/dL)
Clase
0
1
log_Total Proteins (g/dL)
4
3
2
0
1
Clase

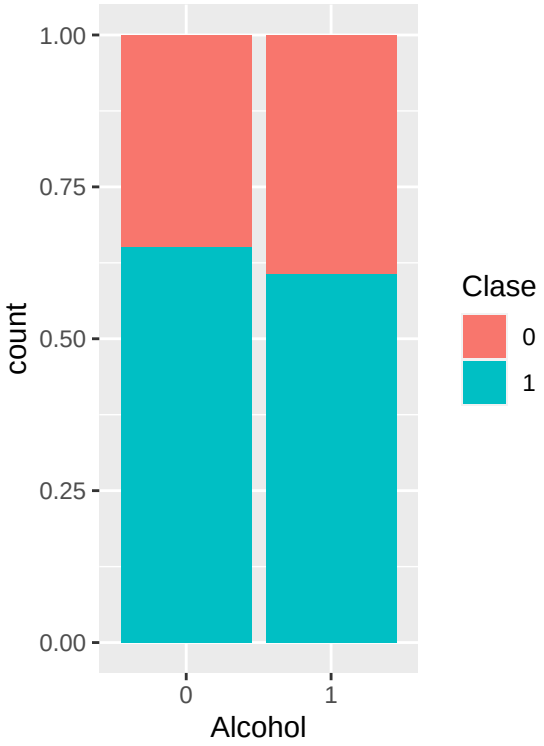

count
30
20
10
0
−1
0
1
2
log_Creatinine (mg/dL)
Clase
0
1
log_Creatinine (mg/dL)
2
1
0
−1
0
1
Clase

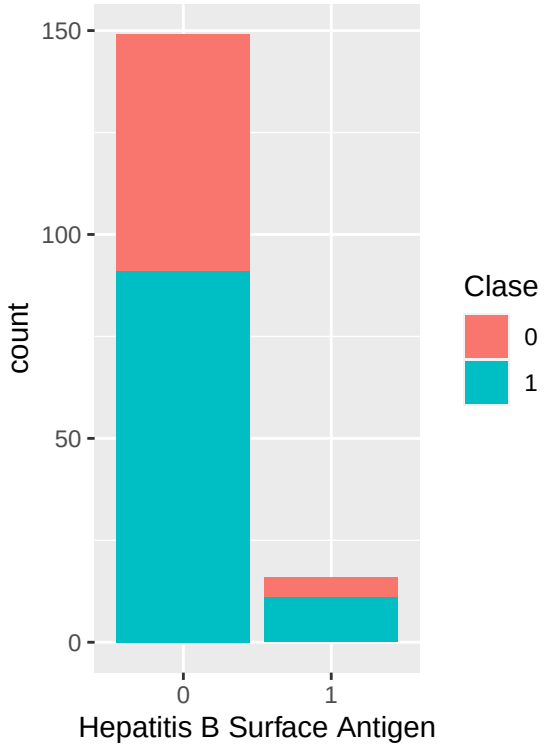


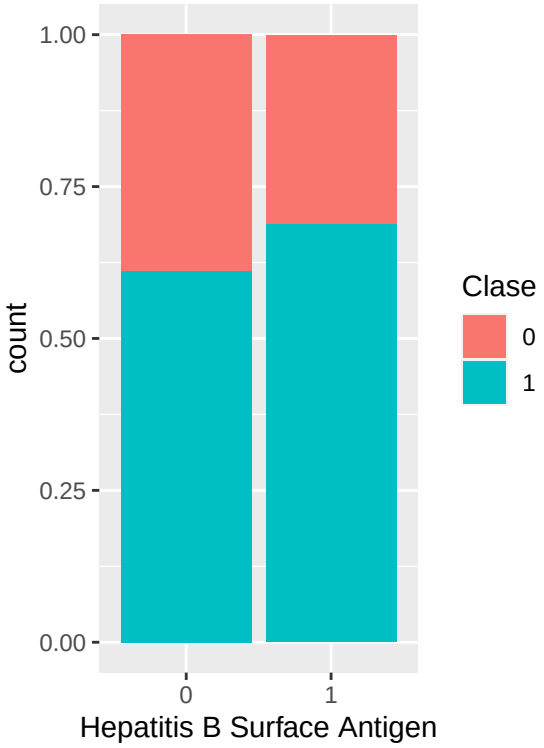

```
# Distribució variables quantitatives
hcc_factorT<-c(1:23,27:29)


for (i in hcc_factorT)
{
  a1<-ggplot(hcc, aes(x=hcc[,i],fill=hcc$`Class Attribute`))+
    xlab(names(hcc)[i])+
    labs(fill="Clase") +
    geom_bar()
  a2<-ggplot(hcc, aes(x=hcc[,i],fill=hcc$`Class Attribute`))+
    xlab(names(hcc)[i])+
    labs(fill="Clase") +
    geom_bar(position = "fill")

  grid.arrange(a1,a2,nrow=1)
}
```

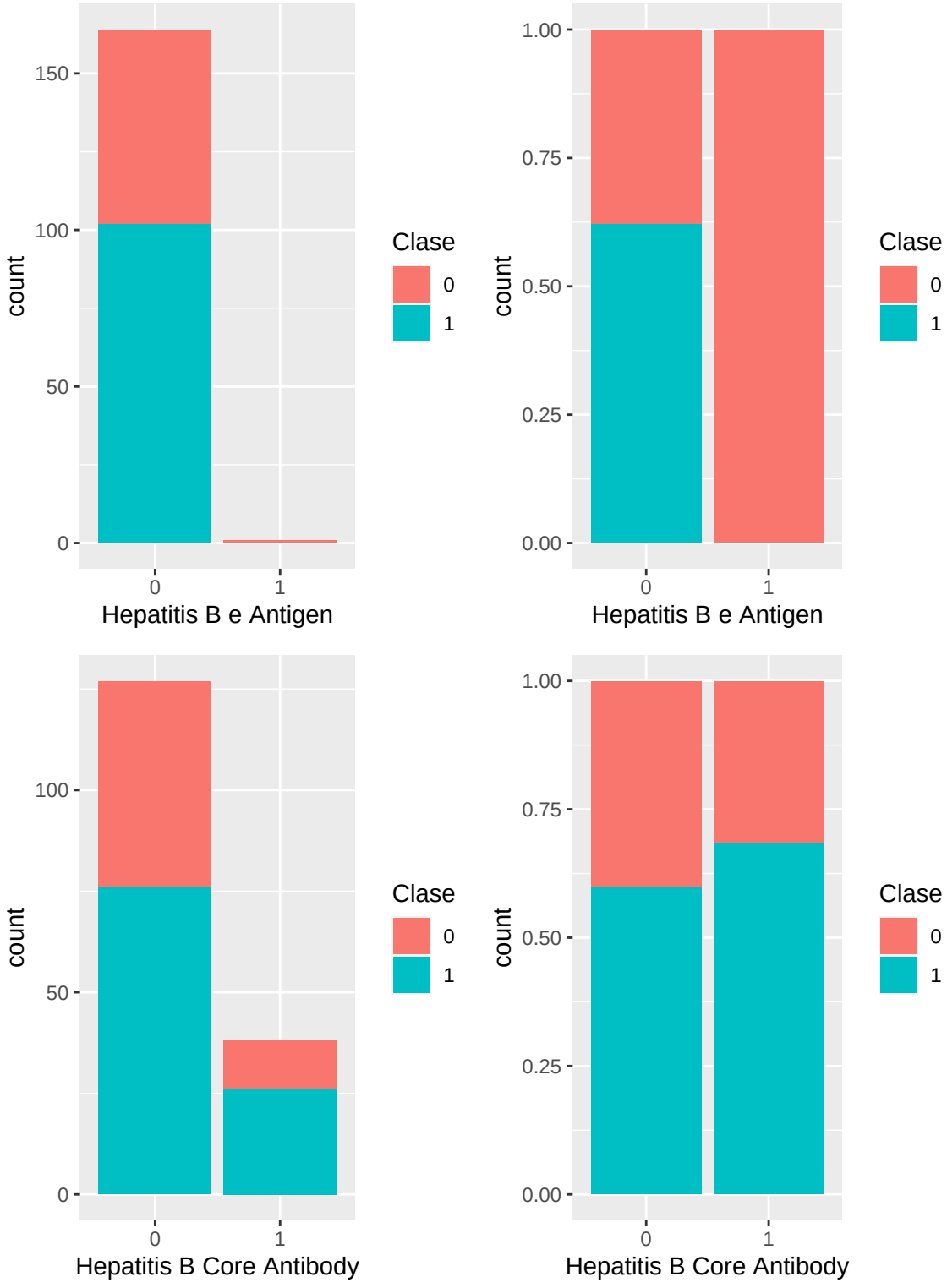

count
100
50
0
0
1
Gender
Clase
0
1
count
1.00
0.75
0.50
0.25
0.00
0
1
Gender
Clase
0
1
count
90
60
30
0
0
1
Symptoms
Clase
0
1
count
1.00
0.75
0.50
0.25
0.00
0
1
Symptoms
Clase
0
1

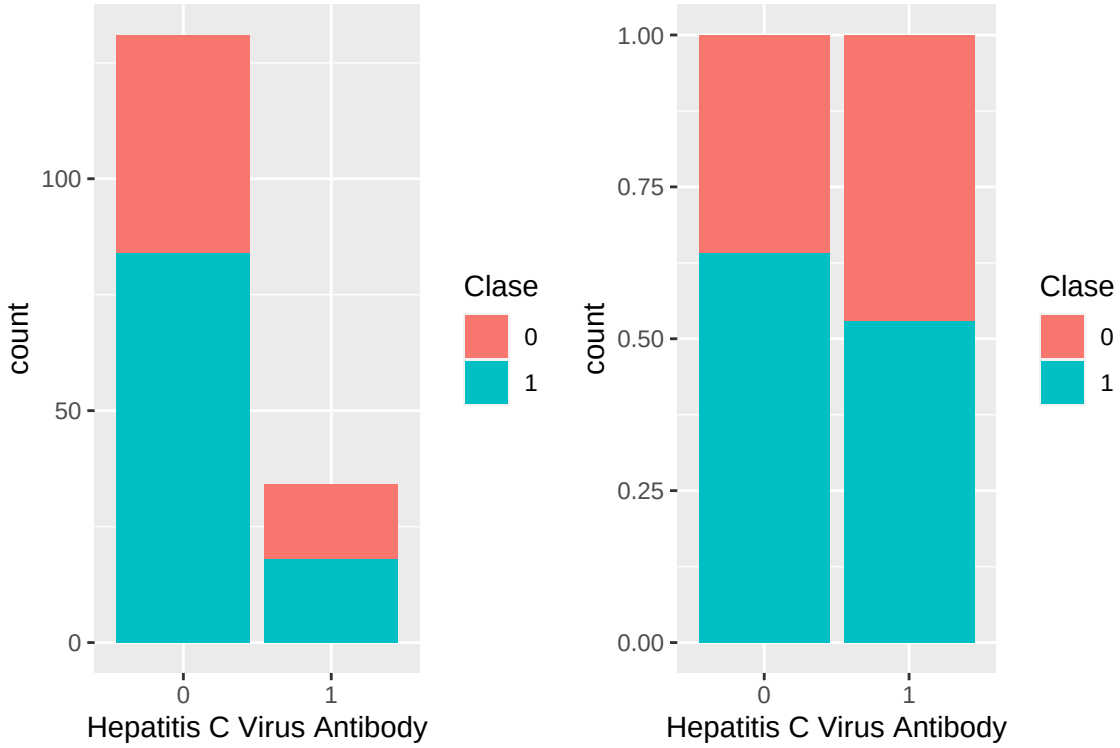

125
100
75
50
25
0
count
0
1
Alcohol
Clase
0
1


1.00
0.75
0.50
0.25
0.00
count
0
1
Alcohol
Clase
0
1

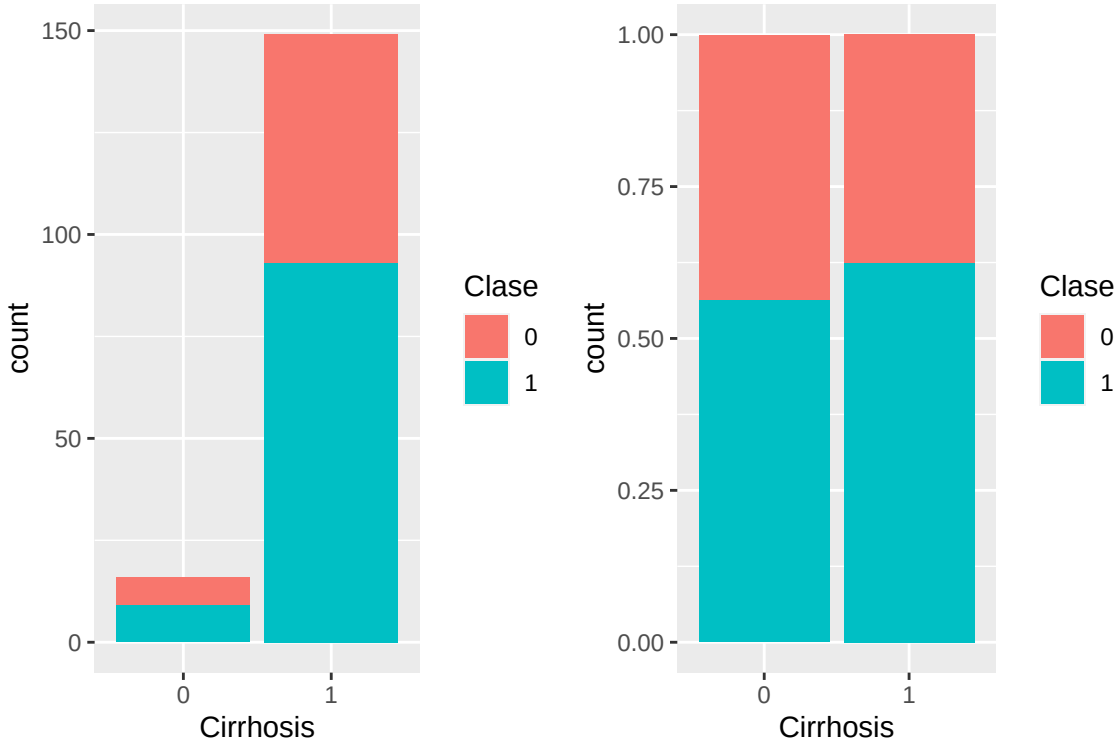

150
100
50
0
count
0
1
Hepatitis B Surface Antigen
Clase
0
1


1.00
0.75
0.50
0.25
0.00
count
0
1
Hepatitis B Surface Antigen
Clase
0
1

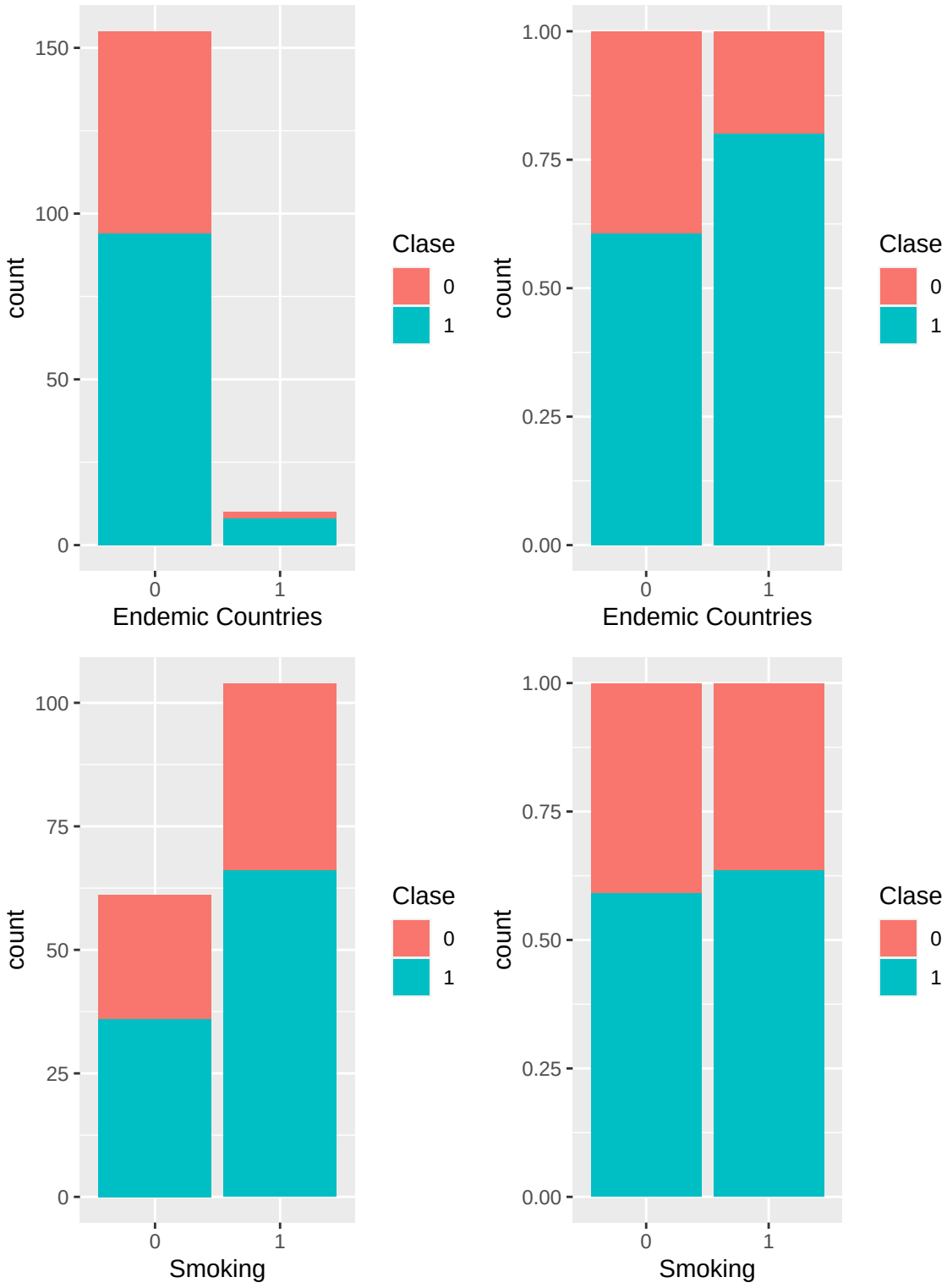

150
100
50
0
count
0
1
Hepatitis B e Antigen
Clase
0
1
1.00
0.75
0.50
0.25
0.00
count
0
1
Hepatitis B e Antigen
Clase
0
1
100
50
0
count
0
1
Hepatitis B Core Antibody
Clase
0
1
1.00
0.75
0.50
0.25
0.00
count
0
1
Hepatitis B Core Antibody
Clase
0
1

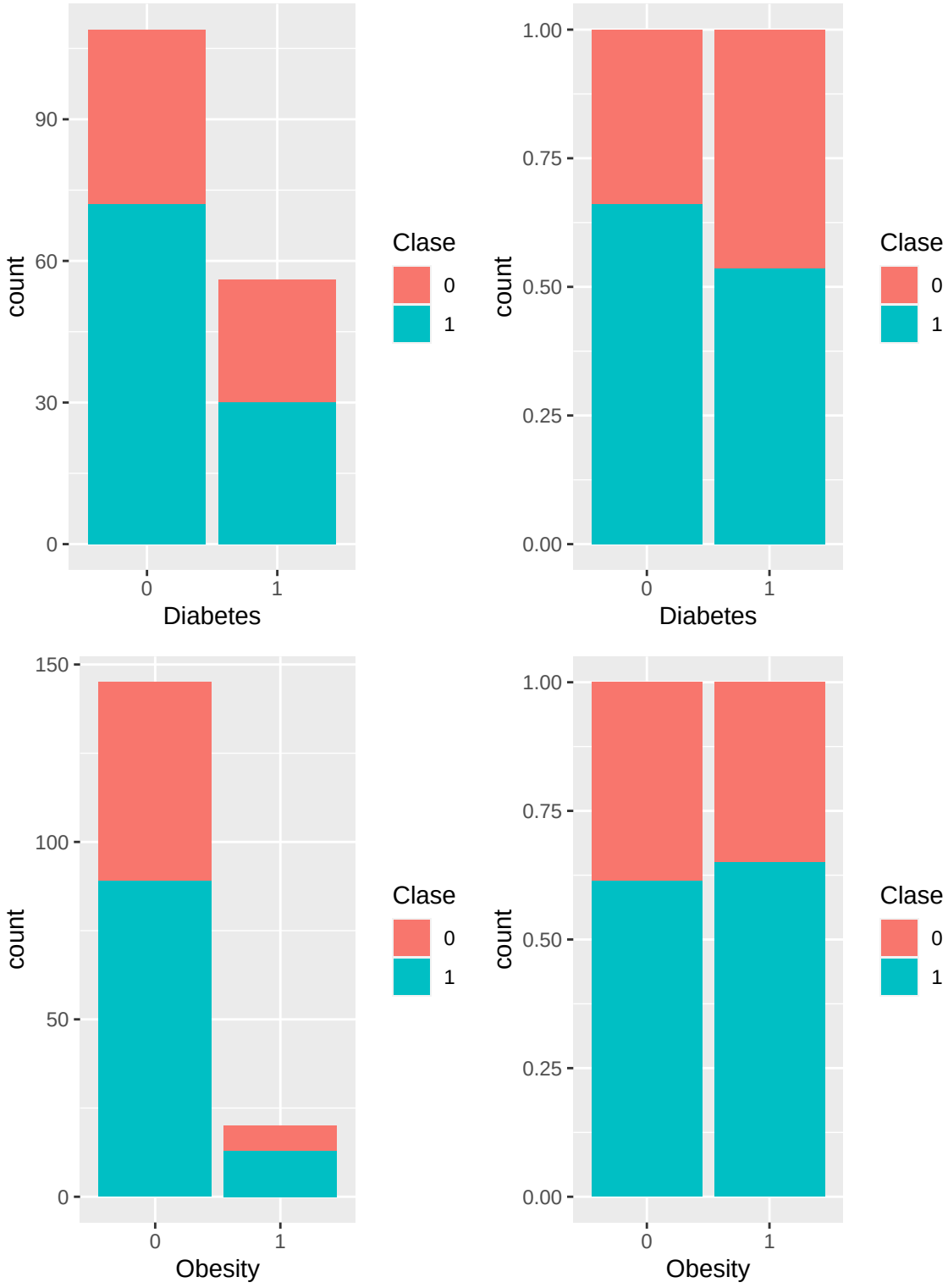

count
100
50
0
0
1
Hepatitis C Virus Antibody
Clase
0
1

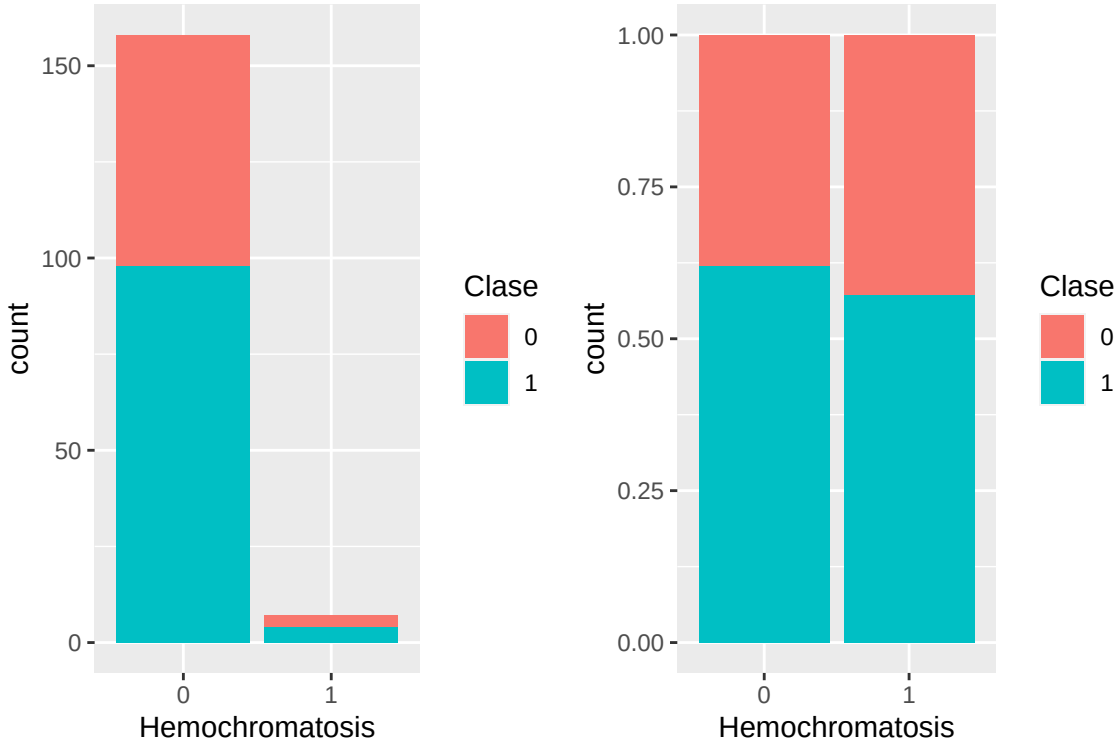
count
1.00
0.75
0.50
0.25
0.00
0
1
Hepatitis C Virus Antibody
Clase
0
1

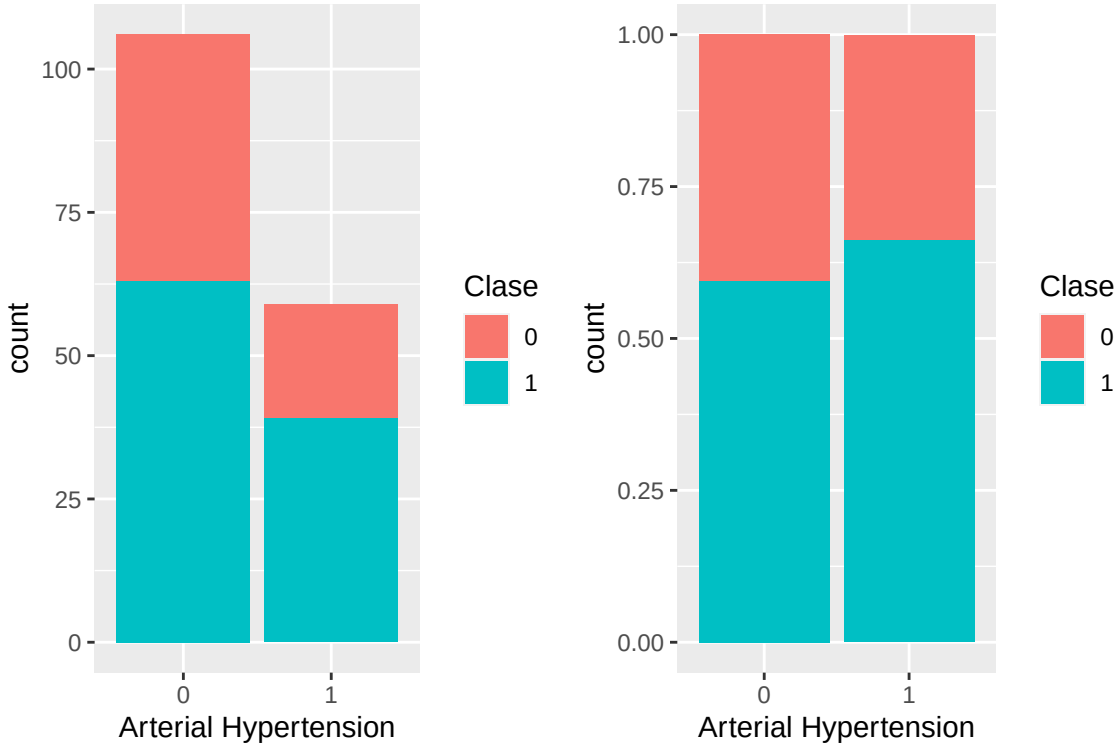

count
150
100
50
0
0
1
Cirrhosis
Clase
0
1

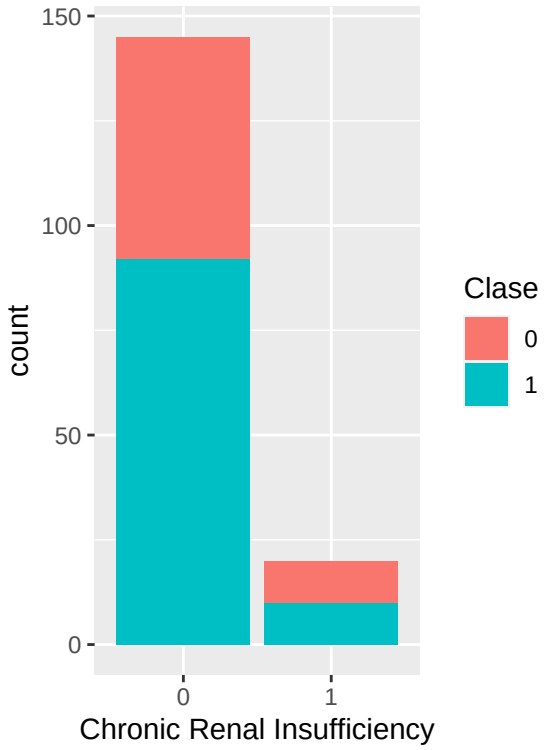
count
1.00
0.75
0.50
0.25
0.00
0
1
Cirrhosis
Clase
0
1

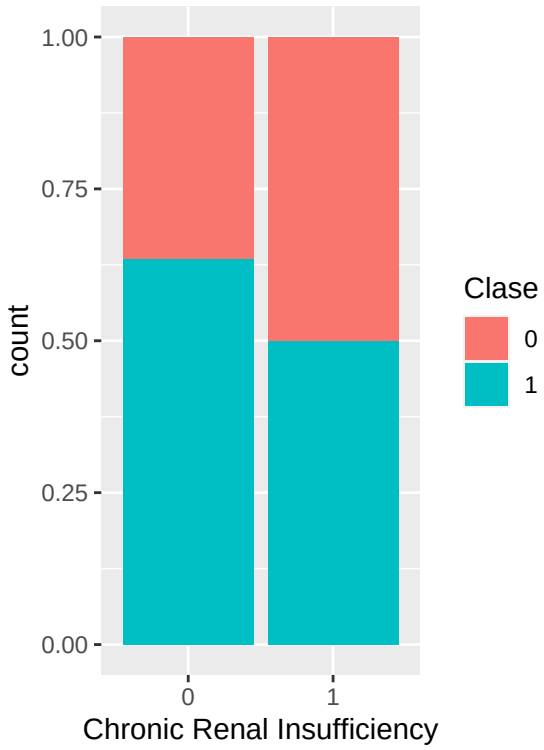

150
100
50
0
count
0
1
Endemic Countries
Clase
0
1
1.00
0.75
0.50
0.25
0.00
count
0
1
Endemic Countries
Clase
0
1
100
75
50
25
0
count
0
1
Smoking
Clase
0
1
1.00
0.75
0.50
0.25
0.00
count
0
1
Smoking
Clase
0
1

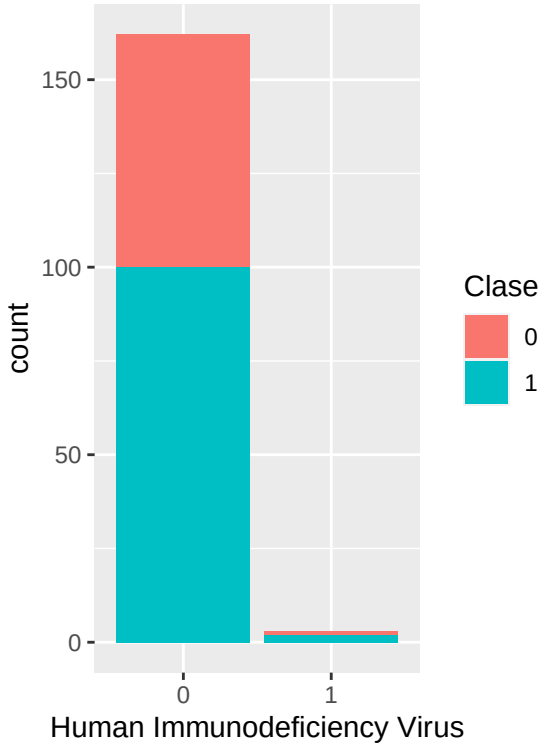

count
90
60
30
0
0
1
Diabetes
Clase
0
1
1.00
0.75
0.50
0.25
0.00
0
1
Diabetes
Clase
0
1
150
100
50
0
0
1
Obesity
Clase
0
1
1.00
0.75
0.50
0.25
0.00
0
1
Obesity
Clase
0
1

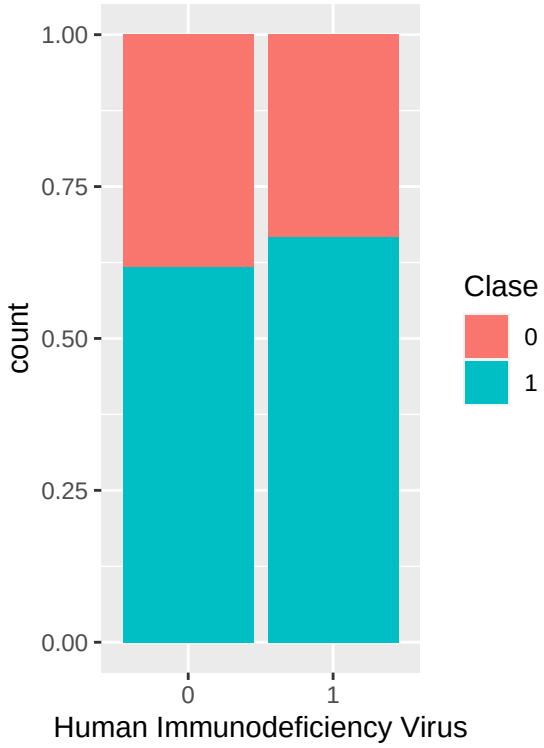

150
100
50
0
count
0
1
Hemochromatosis
Clase
0
1
1.00
0.75
0.50
0.25
0.00
count
0
1
Hemochromatosis
Clase
0
1

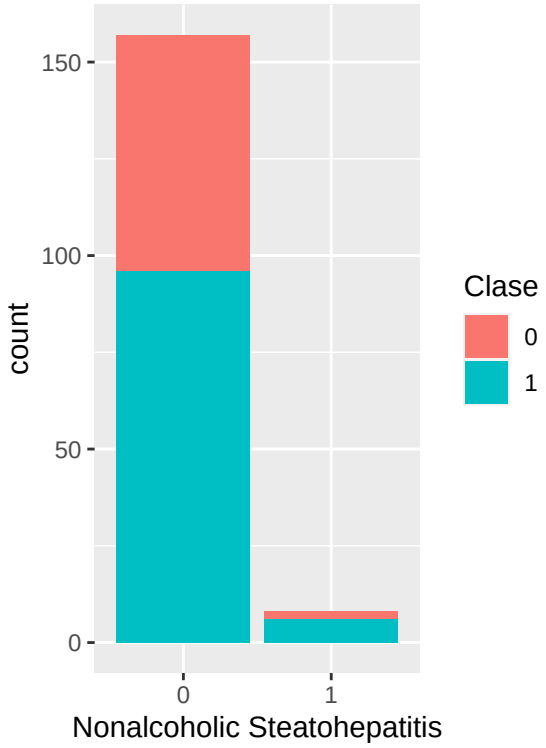

100
75
50
25
0
count
0
1
Arterial Hypertension
Clase
0
1
1.00
0.75
0.50
0.25
0.00
count
0
1
Arterial Hypertension
Clase
0
1

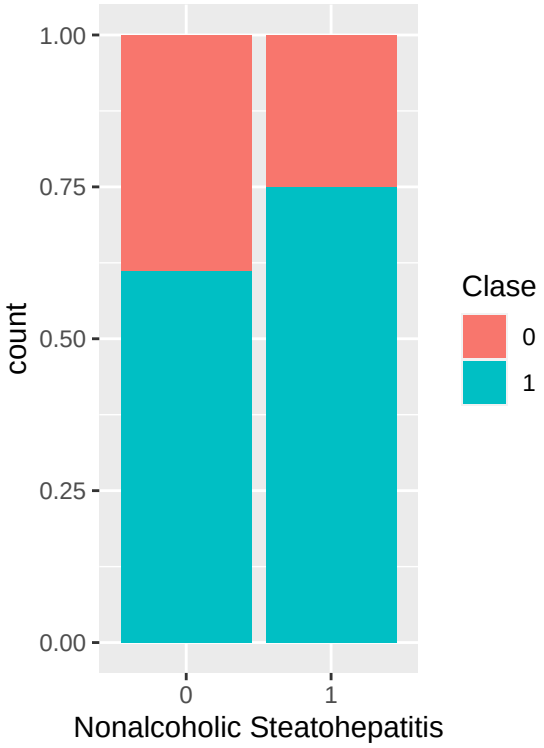

150
100
50
0
count
0
1
Chronic Renal Insufficiency
Clase
0
1

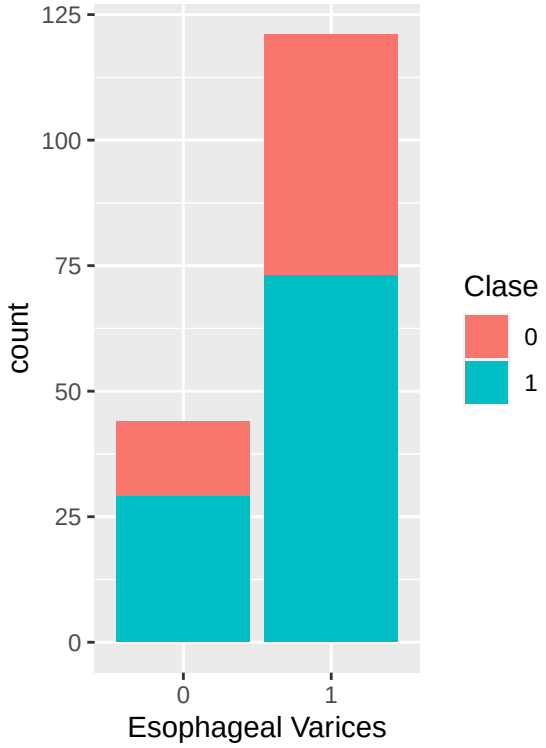

1.00
0.75
0.50
0.25
0.00
count
0
1
Chronic Renal Insufficiency
Clase
0
1

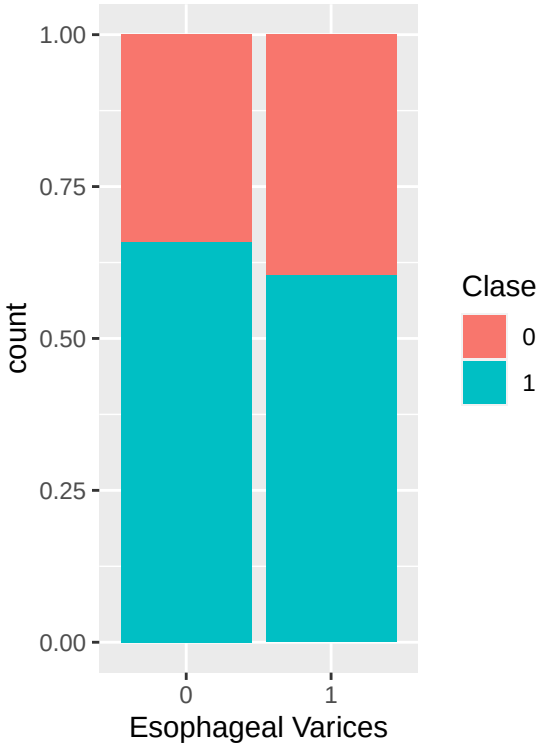

150
100
50
0
count
0
1
Human Immunodeficiency Virus
Clase
0
1

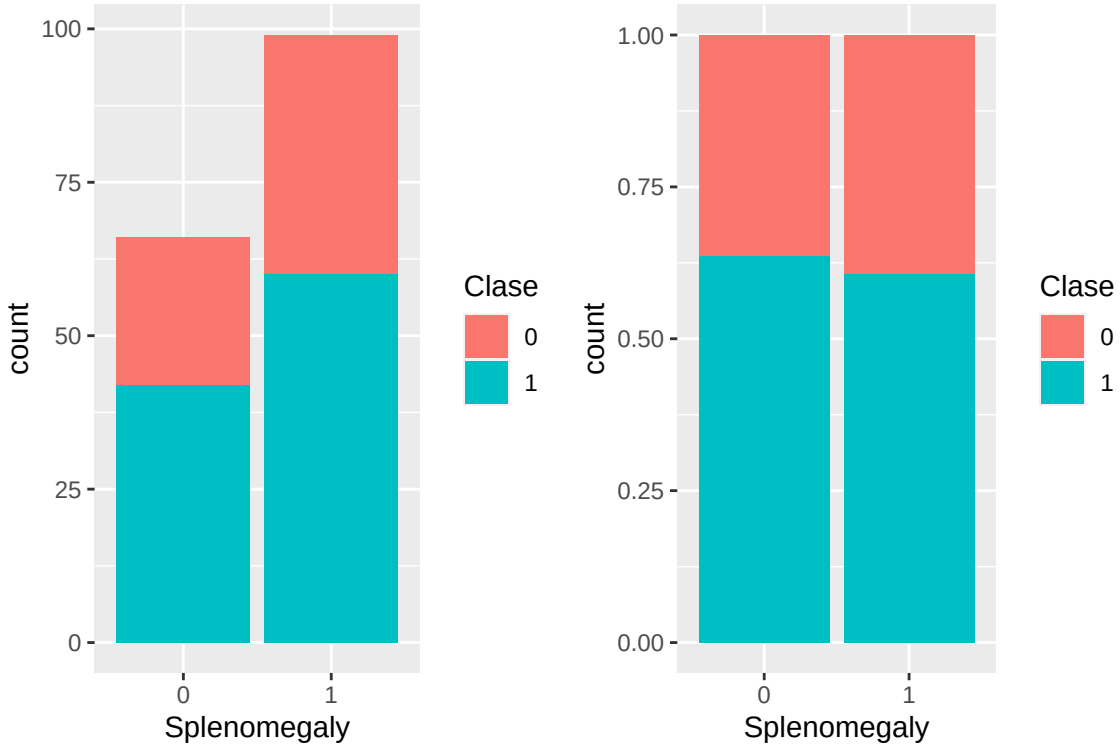

1.00
0.75
0.50
0.25
0.00
count
0
1
Human Immunodeficiency Virus
Clase
0
1


150
100
50
0
count
0
1
Nonalcoholic Steatohepatitis
Clase
0
1

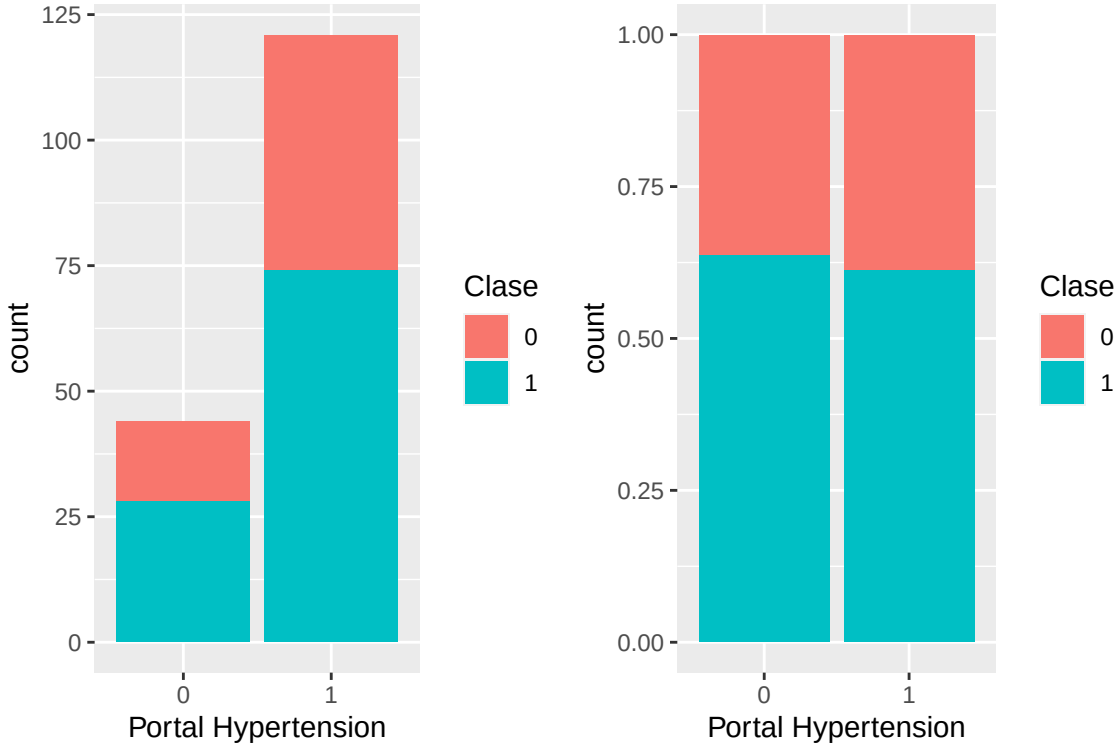

1.00
0.75
0.50
0.25
0.00
count
0
1
Nonalcoholic Steatohepatitis
Clase
0
1


125
100
75
50
25
0
count
0
1
Esophageal Varices
Clase
0
1


1.00
0.75
0.50
0.25
0.00
count
0
1
Esophageal Varices
Clase
0
1


100
75
50
25
0
count
0
1
Splenomegaly
Clase
0
1

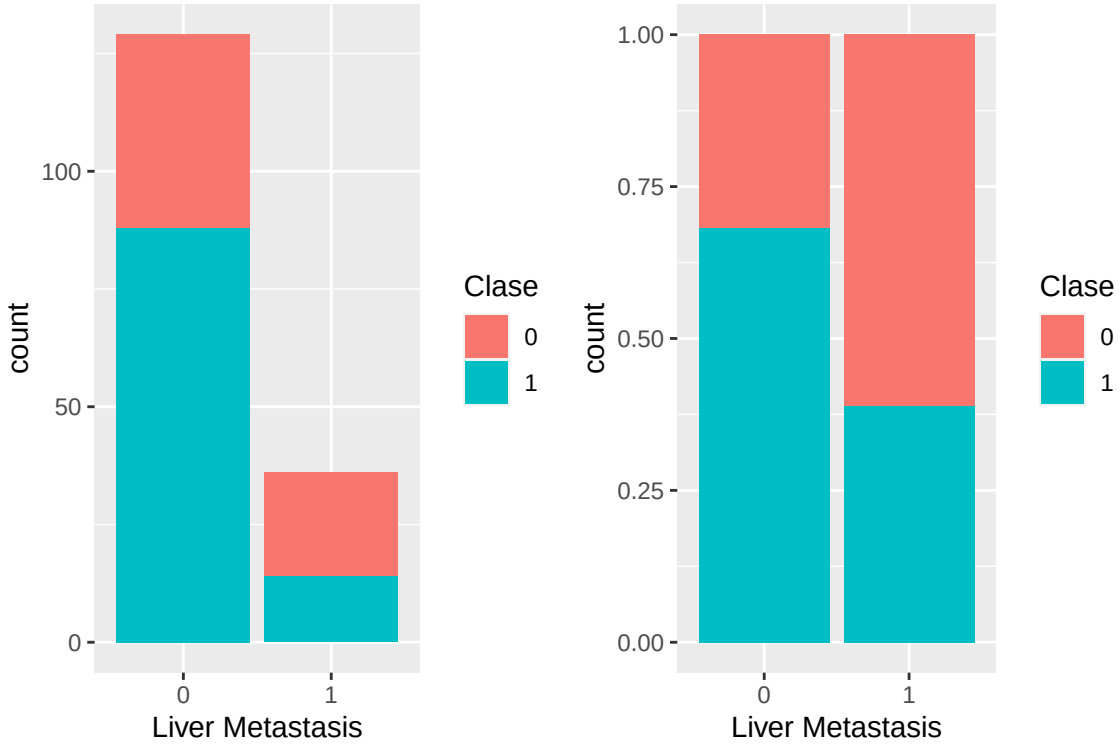
1.00
0.75
0.50
0.25
0.00
count
0
1
Splenomegaly
Clase
0
1


125
100
75
50
25
0
count
0
1
Portal Hypertension
Clase
0
1

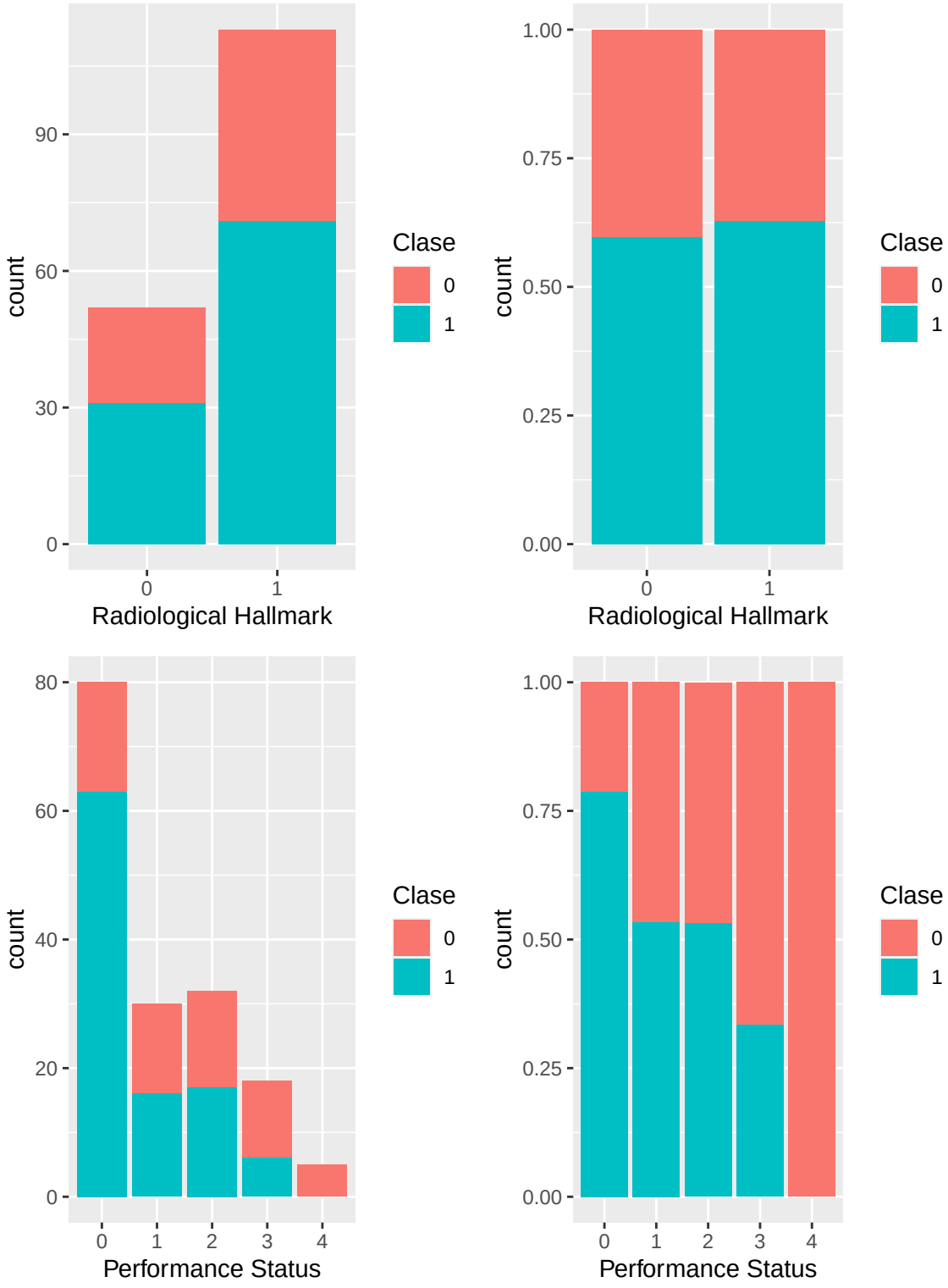
1.00
0.75
0.50
0.25
0.00
count
0
1
Portal Hypertension
Clase
0
1

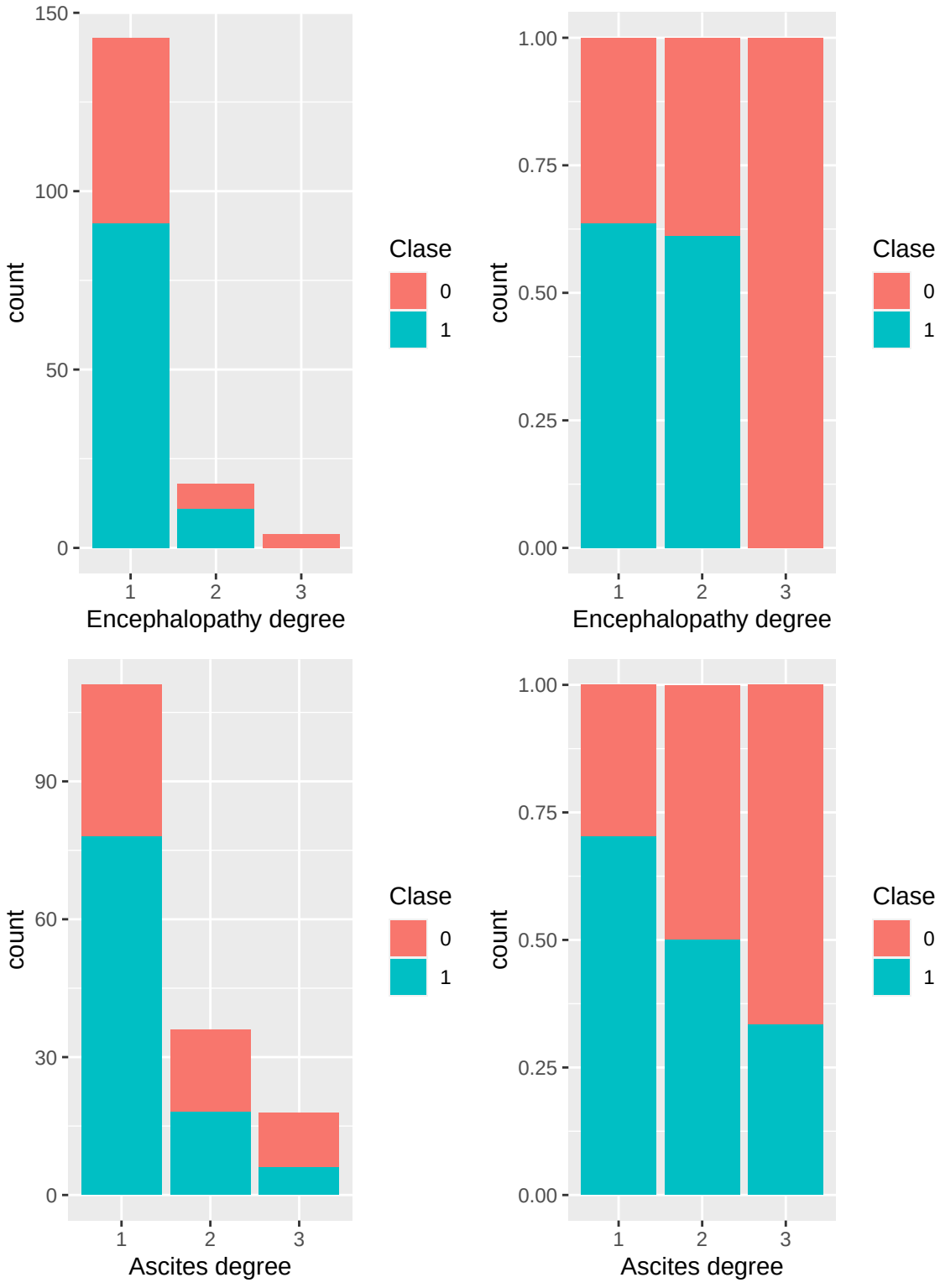

count
100
50
0
0
1
Portal Vein Thrombosis
Clase
0
1

count
1.00
0.75
0.50
0.25
0.00
0
1
Portal Vein Thrombosis
Clase
0
1


count
100
50
0
0
1
Liver Metastasis
Clase
0
1

count
1.00
0.75
0.50
0.25
0.00
0
1
Liver Metastasis
Clase
0
1


count
90
60
30
0
0
1
Radiological Hallmark
Clase
0
1
count
1.00
0.75
0.50
0.25
0.00
0
1
Radiological Hallmark
Clase
0
1
count
80
60
40
20
0
0
1
2
3
4
Performance Status
Clase
0
1
count
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
Performance Status
Clase
0
1



FALTA DESCRIPCIÓN DE LAS VARIABLES!!!!

Comprobació de la normalitat

```
library(nortest)
alpha = 0.05
col.names = colnames(hcc)
for (i in 1:ncol(hcc)) {
  if (i == 1) cat("Variables que no segueixen una distribució normal:\n")
  if (is.integer(hcc[,i]) | is.numeric(hcc[,i])) {
    p_val = ad.test(hcc[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(hcc) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no segueixen una distribució normal:
## Age at diagnosis,
## Grams of Alcohol per day, Packs of cigarets per year, International Normalised Ratio,
## log_Alpha-Fetoprotein (ng/mL), Leukocytes(G/L), Platelets, Albumin (mg/dL),
## log_Total Bilirubin(mg/dL), log_Alanine transaminase (U/L), log_Aspartate transaminase (U/L),
## log_Alkaline phosphatase (U/L), log_Total Proteins (g/dL),
## log_Creatinine (mg/dL), Number of Nodules, Major dimension of nodule (cm),
## log_Direct Bilirubin (mg/dL), Iron, Oxygen Saturation (%),
## Ferritin (ng/mL)
```

Com es pot veure, hi ha moltes variables numériques que es distancien significativament de la distribució normal, per el que s'usaran test no paramétrics (Mann–Whitney–Wilcoxon) per la comparativa en relació a la supervivència.

Proves estadístiques

Comparació entre grups de la classe

Per tal de valorar quines variables es comporten diferents entre en que sobreviuen i els que no, es realitzarà els diferents test estadístics:

- Per les variables quantititatives, donada les seves distribucións majoritària diferent a la normalitat, es realitzara el test no paramètric de Mann–Whitney–Wilcoxon
- Per les variables qualitatives es realitzarà un test chi-quadrat.

```
testSig <- tibble()


cat("Variables categóriques amb p<0.10 entre la classe:\n")
```

```
## Variables categóriques amb p<0.10 entre la classe:
```

```r
for (i in hcc_factorT) {
  tabla=table(hcc[,i], hcc$`Class Attribute`)
  chi=chisq.test(tabla)
  if (chi$p.value<0.1){
    cat(names(hcc)[i],":\n")
    cat("p=",chi$p.value, "\n")
    testSig <- testSig %>% bind_rows(c(Clase = i, Name=names(hcc[i]),Categorica="1",p_value = chi$p.valu
  }
}
```

```
## Symptoms :
## p= 0.007933578
## Portal Vein Thrombosis :
## p= 0.008776766
## Liver Metastasis :
## p= 0.002624464
## Performance Status :
## p= 3.254868e-05
## Encephalopathy degree :
## p= 0.03543194
## Ascites degree :
## p= 0.002912547
```

```r
cat("\n")
```

```r
cat("Variables numériques amb p<0.10 entre la classe:\n")
```

```
## Variables numériques amb p<0.10 entre la classe:
```

```r
for (i in hcc_num) {
  wil=wilcox.test(hcc[hcc$`Class Attribute`==1,i],hcc[hcc$`Class Attribute`==0,i], mu = 0,paired = FALSE

  if (wil$p.value<0.1){
    cat(names(hcc)[i],":\n")
    cat("p=",wil$p.value, "\n")
    testSig <- testSig %>% bind_rows(c(Clase = i, Name=names(hcc[i]),Categorica="0",p_value = wil$p.valu
  }
}
```

```
## Age at diagnosis :
## p= 0.03568324
## International Normalised Ratio :
## p= 0.03140037
## log_Alpha-Fetoprotein (ng/mL) :
## p= 2.831168e-06
## Haemoglobin (g/dL) :
## p= 5.968093e-05
## Platelets :
## p= 0.03501771
## Albumin (mg/dL) :
## p= 0.000223525
```

```
## log_Total Bilirubin(mg/dL) :
## p= 0.03695266
## log_Aspartate transaminase (U/L) :
## p= 0.001587891
## log_Gamma glutamyl transferase (U/L) :
## p= 0.01862733
## log_Alkaline phosphatase (U/L) :
## p= 6.898549e-07
## log_Creatinine (mg/dL) :
## p= 0.09840758
## Major dimension of nodule (cm) :
## p= 0.03803416
## log_Direct Bilirubin (mg/dL) :
## p= 0.003341791
## Iron :
## p= 0.003808222
## Ferritin (ng/mL) :
## p= 0.007306848
```

Aquestes variables seran les que es seleccionarà per a la creació d'un model de regressió logística, però abans, valorarem les correlacions entre elles per tal de seleccionar variables que estiguin poc relacionades entre elles.

Correlació entre les variables seleccionades

Amb respecte a la corrrelació entre les variables seleccionades, veiem la seva matriu de correlacions. Donada l'existència de variables categóriques, aquestes es consideradan numériques i usarem la correlació no paramétrica de Spearman per a la seva valoració. Previament es normalitzaran totes les variables quantitatives.

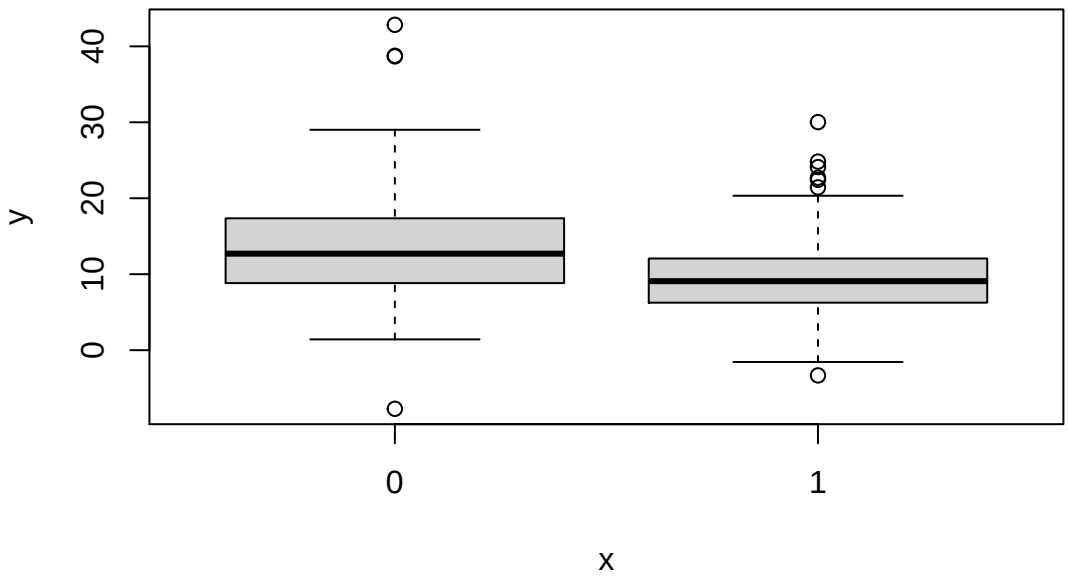
```
# Correlación variables independientes.
library(corrplot)
hcc_norm<- hcc
atr<-names(hcc)
hcc_norm[,hcc_num]<-scale(hcc_norm[,hcc_num])

hcc2<-as.data.frame(lapply(hcc_norm,as.numeric))

names(hcc2)<-atr

hcc_cor<-cor(hcc2[,testSig$Name], method = "spearman")
corrplot(hcc_cor, cl.pos='n',tl.srt = 45, tl.cex = 0.5, type="upper",method = "circle", order="FPC", di
```

Per intentar reduïr el número de variables del model, d'entre les que han tingut difències significatives entre els que sobreviuen i els que no, seleccionarem les que tingin poca correlació amb altres variables, i amb les que tenen molta correlació amb d'altres, només seleccionarem una com a representiva.

MELD

Hi ha un valor extès per a valorar la probabilitat de mort als tres mesos de pacients amb hepatopatia que depend de la creatinina, la bilirrubina total i del INR, amb un valor calculat denominat MELD. Aquest valor és una estimació de probabilitat de fallida hepàtica. A majors valors major probabilitat de mort. Aquest valor està validat pels 3 mesos, no per a l'any, com es el nostre cas.

```
# MELC calculat vs clase
meld<-3.78*hcc$`log_Total Bilirubin(mg/dL)`+ 11.2*log(hcc$`International Normalised Ratio`)+9.57*hcc$`l

hcc$MELD<-meld

plot(hcc$`Class Attribute`, meld)
```




```r
# Ho no hi ha diferències en la supervivència depenent del MELD
wilcox.test(meld[hcc$`Class Attribute`==1],meld[hcc$`Class Attribute`==0], mu = 0,paired = FALSE, conf.
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  meld[hcc$`Class Attribute` == 1] and meld[hcc$`Class Attribute` == 0]
## W = 2170, p-value = 0.0004713
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -5.524871 -1.604940
## sample estimates:
## difference in location
##               -3.56166
```

És pot veure que els valors de MELD són significativament diferents entre el grup de pacients que sobreviuen i els que no. Amb aquesta combinació lineal agrupem en una única varible la bilirrubina total, l'INR i la creatinina.

Per tant, les variables seleccionades per estudiar amb regressió logística seran:

- Symptoms
- Portal Vein Thrombosis
- Liver Metastasis
- Age at diagnosis
- Encephalopathy degree
- Ascites degree
- log_Alpha-Fetoprotein (ng/mL)

- Haemoglobin (g/dL)
- log_Aspartate transaminase (U/L)
- log_Gamma glutamyl transferase (U/L)
- Major dimension of nodule (cm)
- Iron
- Ferritin (ng/mL)
- MELD

Model amb regressió logística

```
selecc<-c(2, 21, 22, 24, 28, 29, 31, 32, 39, 40, 45, 47, 49, 51,50)

hcc_sel<-hcc[selecc]

#hcc_sel<-as.data.frame(lapply(hcc_sel,as.numeric))
#names(hcc_sel)<-names(hcc[selecc])

modelo <- glm(`Class Attribute` ~ ., data = hcc_sel, family = "binomial")
summary(modelo)
```

```
##
## Call:
## glm(formula = `Class Attribute` ~ ., family = "binomial", data = hcc_sel)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2021  -0.5747   0.3268   0.6910   2.6484
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          3.642e+00  2.443e+00   1.491  0.13603
## Symptoms1                           -5.761e-01  4.887e-01  -1.179  0.23841
## `Portal Vein Thrombosis`1           -6.680e-01  5.335e-01  -1.252  0.21059
## `Liver Metastasis`1                 -4.365e-01  5.355e-01  -0.815  0.41498
## `Age at diagnosis`                  -2.719e-02  1.756e-02  -1.549  0.12149
## `Encephalopathy degree`2             1.385e-01  7.752e-01   0.179  0.85816
## `Encephalopathy degree`3            -1.462e+01  9.961e+02  -0.015  0.98829
## `Ascites degree`2                   -2.173e-01  5.791e-01  -0.375  0.70748
## `Ascites degree`3                   -5.911e-01  6.799e-01  -0.869  0.38459
## `log_Alpha-Fetoprotein (ng/mL)`     -1.867e-01  7.806e-02  -2.391  0.01678
## `Haemoglobin (g/dL)`                 2.361e-01  1.161e-01   2.034  0.04196
## `log_Aspartate transaminase (U/L)`  -4.453e-01  3.819e-01  -1.166  0.24364
## `log_Gamma glutamyl transferase (U/L)` 1.723e-02  2.768e-01   0.062  0.95037
## `Major dimension of nodule (cm)`    -6.818e-02  4.763e-02  -1.432  0.15227
## Iron                                 1.323e-02  7.251e-03   1.824  0.06811
## `Ferritin (ng/mL)`                  -2.445e-03  9.004e-04  -2.716  0.00661
## MELD                                -5.389e-02  3.797e-02  -1.419  0.15579
##
## (Intercept)
## Symptoms1
## `Portal Vein Thrombosis`1
## `Liver Metastasis`1
```

```
## `Age at diagnosis`
## `Encephalopathy degree`2
## `Encephalopathy degree`3
## `Ascites degree`2
## `Ascites degree`3
## `log_Alpha-Fetoprotein (ng/mL)`        *
## `Haemoglobin (g/dL)`                   *
## `log_Aspartate transaminase (U/L)`
## `log_Gamma glutamyl transferase (U/L)`
## `Major dimension of nodule (cm)`
## Iron                                   .
## `Ferritin (ng/mL)`                     **
## MELD
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 219.43  on 164  degrees of freedom
## Residual deviance: 141.62  on 148  degrees of freedom
## AIC: 175.62
##
## Number of Fisher Scoring iterations: 15
```