Hepatisis data analysis

```
library(dplyr)
library(visdat)
library(caret)
library(RANN)
library(corrplot)
library(plotly)
library(ggplot2)
library(resample)
library(DataExplorer)
library(imputeMulti)
library(mice)
library(rmarkdown)
```

Introduction

First rows of our data frame.

```
df <- read.table("hepatitis.data", sep = ",")
colnames(df) <- c("Class","Age","Sex","Steroid","Antivirals","Fatigue","Malaise","Anorexia","LiverBig",
attach(df)
```

```
head(df)
```

```
##   Class Age Sex Steroid Antivirals Fatigue Malaise Anorexia LiverBig LiverFirm
## 1     2  30   2       1          2       2       2        2        1         2
## 2     2  50   1       1          2       1       2        2        1         2
## 3     2  78   1       2          2       1       2        2        2         2
## 4     2  31   1       ?          1       2       2        2        2         2
## 5     2  34   1       2          2       2       2        2        2         2
## 6     2  34   1       2          2       2       2        2        2         2
##   SpleenPalpable Spiders Ascites Varices Bilirubin AlkPhosphate Sgot Albumin
## 1              2       2       2       2      1.00           85   18     4.0
## 2              2       2       2       2      0.90          135   42     3.5
## 3              2       2       2       2      0.70           96   32     4.0
## 4              2       2       2       2      0.70           46   52     4.0
## 5              2       2       2       2      1.00            ?  200     4.0
## 6              2       2       2       2      0.90           95   28     4.0
##   Protime Histology
## 1       ?         1
## 2       ?         1
## 3       ?         1
## 4      80         1
## 5       ?         1
## 6      75         1
```

Preparing data

```
df[df == "?"] <- NA

df <- mutate_all(df, function(x) as.numeric(as.character(x)))
categorical <- c(1, 3:14, 20)
df[, categorical] <- replace(df[, categorical], df[, categorical] == 2, 0)
df[, categorical] <-  lapply(df[, categorical], as.factor)
```

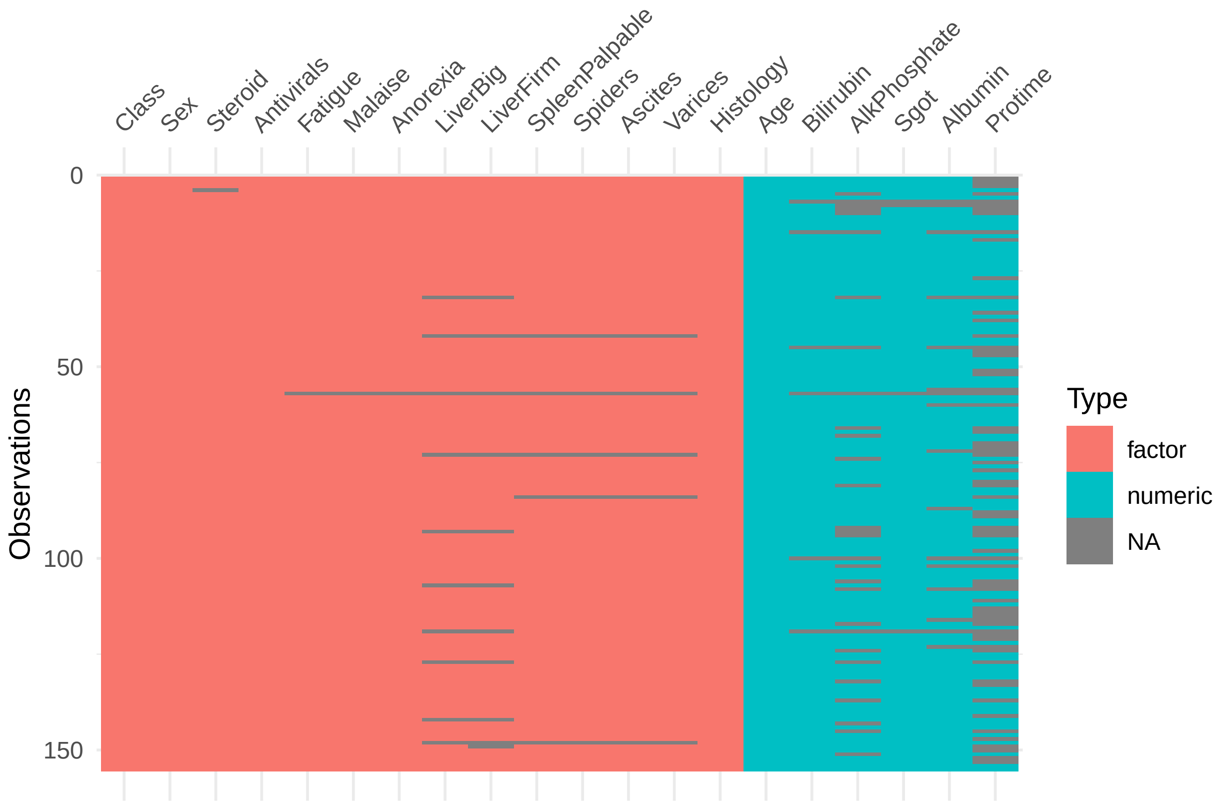
```
sum(rowSums(is.na(df)) != 0) # we cannot remove these rows!
```

```
## [1] 75
```

```
sum(is.na(df))
```

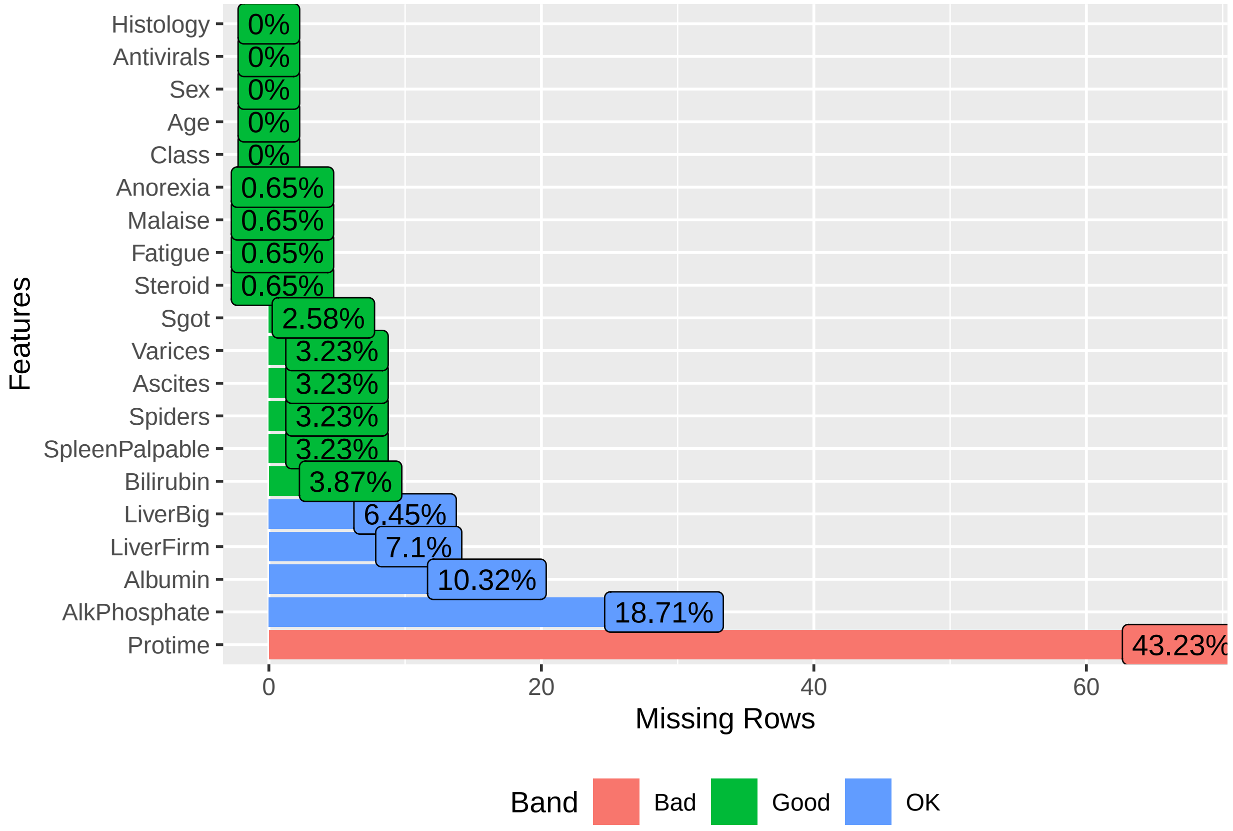
```
## [1] 167
```

```
vis_dat(df)
```

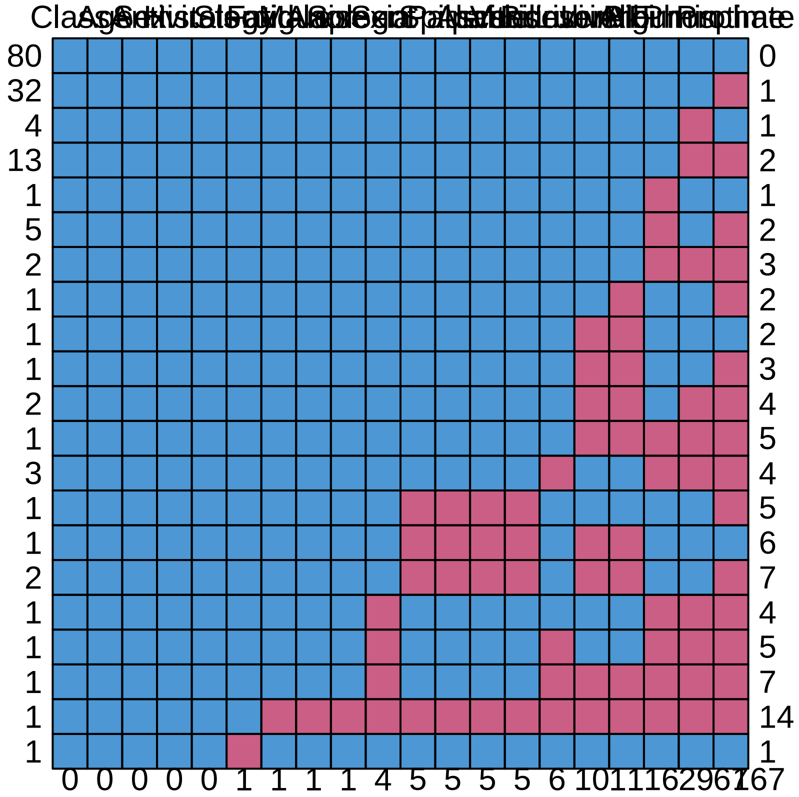


```
# vis_dat(df2)
```

```
plot_missing(df)
```



```
md.pattern(df)
```

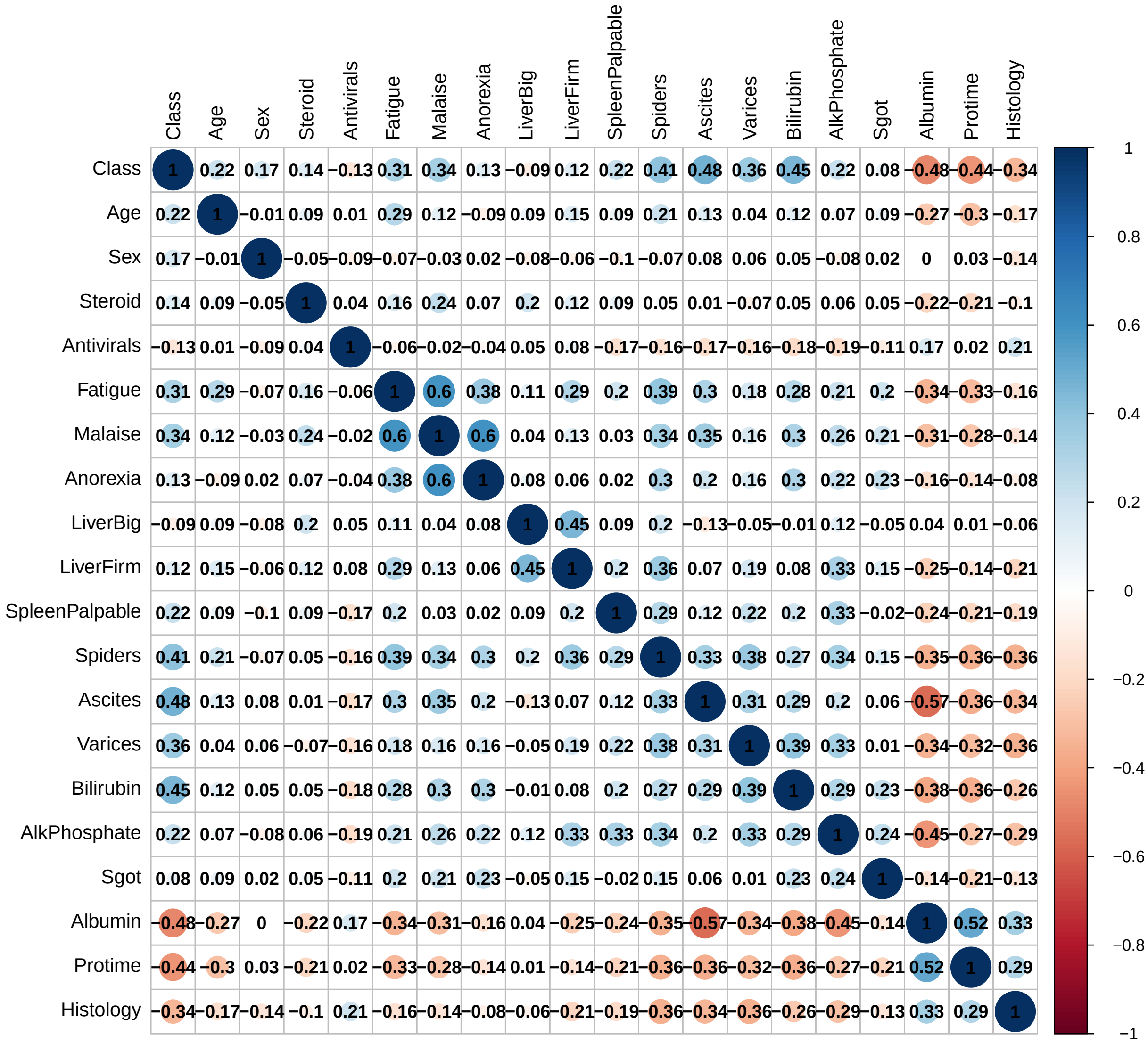


```
##    Class Age Sex Antivirals Histology Steroid Fatigue Malaise Anorexia Sgot
## 80     1   1   1          1         1       1       1       1        1    1
## 32     1   1   1          1         1       1       1       1        1    1
## 4      1   1   1          1         1       1       1       1        1    1
## 13     1   1   1          1         1       1       1       1        1    1
## 1      1   1   1          1         1       1       1       1        1    1
## 5      1   1   1          1         1       1       1       1        1    1
## 2      1   1   1          1         1       1       1       1        1    1
## 1      1   1   1          1         1       1       1       1        1    1
## 1      1   1   1          1         1       1       1       1        1    1
## 1      1   1   1          1         1       1       1       1        1    1
## 2      1   1   1          1         1       1       1       1        1    1
## 1      1   1   1          1         1       1       1       1        1    1
## 3      1   1   1          1         1       1       1       1        1    1
## 1      1   1   1          1         1       1       1       1        1    1
## 1      1   1   1          1         1       1       1       1        1    1
## 2      1   1   1          1         1       1       1       1        1    1
## 1      1   1   1          1         1       1       1       1        1    0
## 1      1   1   1          1         1       1       1       1        1    0
## 1      1   1   1          1         1       1       1       1        1    0
## 1      1   1   1          1         1       1       0       0        0    0
## 1      1   1   1          1         1       0       1       1        1    1
##         0   0   0          0         0       1       1       1        1    4
##    SpleenPalpable Spiders Ascites Varices Bilirubin LiverBig LiverFirm Albumin
## 80              1       1       1       1         1        1         1       1
## 32              1       1       1       1         1        1         1       1
## 4               1       1       1       1         1        1         1       1
## 13              1       1       1       1         1        1         1       1
## 1               1       1       1       1         1        1         1       0
## 5               1       1       1       1         1        1         1       0
## 2               1       1       1       1         1        1         1       0
## 1               1       1       1       1         1        1         0       1
## 1               1       1       1       1         1        0         0       1
## 1               1       1       1       1         1        0         0       1
## 2               1       1       1       1         1        0         0       1
## 1               1       1       1       1         1        0         0       0
## 3               1       1       1       1         0        1         1       0
## 1               0       0       0       0         1        1         1       1
## 1               0       0       0       0         1        0         0       1
## 2               0       0       0       0         1        0         0       1
## 1               1       1       1       1         1        1         1       0
## 1               1       1       1       1         0        1         1       0
## 1               1       1       1       1         0        0         0       0
## 1               0       0       0       0         0        0         0       0
## 1               1       1       1       1         1        1         1       1
##                 5       5       5       5         6       10        11      16
##    AlkPhosphate Protime
## 80             1       1   0
## 32             1       0   1
## 4              0       1   1
## 13             0       0   2
## 1              1       1   1
## 5              1       0   2
## 2              0       0   3
## 1              1       0   2
## 1              1       1   2
## 1              1       0   3
## 2              0       0   4
## 1              0       0   5
## 3              0       0   4
## 1              1       0   5
## 1              1       1   6
## 2              1       0   7
## 1              0       0   4
## 1              0       0   5
## 1              0       0   7
## 1              0       0  14
## 1              1       1   1
##                29      67 167
```

```r
df.new <- mutate_all(df, function(x) as.numeric(as.character(x)))
data_transform <- preProcess(df.new, method = "knnImpute")
data_transform2 <- preProcess(df.new, method = "bagImpute")
data_transform3 <- preProcess(df.new, method = "medianImpute")

df1 <- predict(data_transform, df.new)
df2 <- predict(data_transform2, df.new)
df3 <- predict(data_transform3, df.new)

df2[c(4:14, 16:17, 19)] <- round(df2[c(4:14, 16:17, 19)])
```

```
df3[c(4:14, 16:17, 19)] <- round(df3[c(4:14, 16:17, 19)])
```

```
cor_matrix <- cor(df2[, sapply(df2, is.numeric)], method = "pearson")
corrplot(cor_matrix, tl.col = "black", addCoef.col = 1, number.cex = 0.9)
```



```
methods = c(" "," "," ","logreg"," ","logreg","logreg","logreg","logreg","logreg","logreg","logreg","lo
imp_single <- mice(df, m = 1, method = methods) # Impute missing values
```

```
##
##  iter imp variable
##   1   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   2   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   3   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   4   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   5   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
```

```
df4 <- complete(imp_single)         # Store imputed data

imp_multi <- mice(df, method = methods)  # Impute missing values multiple times
```

```
##
##  iter imp variable
##   1   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   1   2  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   1   3  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   1   4  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   1   5  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   2   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   2   2  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   2   3  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   2   4  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   2   5  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   3   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   3   2  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   3   3  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   3   4  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   3   5  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   4   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   4   2  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   4   3  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   4   4  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   4   5  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   5   1  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   5   2  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   5   3  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   5   4  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
##   5   5  Steroid  Fatigue  Malaise  Anorexia  LiverBig  LiverFirm  SpleenPalpable  Spiders  Ascites
```

```
df5 <- complete(imp_multi, 1)
```

```
# pmm, predictive mean matching (numeric data)
# logreg, logistic regression imputation (binary data, factor with 2 levels)
```

```
summary(df)
```

```
##  Class        Age         Sex      Steroid    Antivirals Fatigue     Malaise
##  0:123   Min.   : 7.0   0: 16   0   :78   0:131      0   : 54   0   :93
##  1: 32   1st Qu.:32.0   1:139   1   :76   1: 24      1   :100   1   :61
##          Median :39.0           NA's: 1              NA's:  1   NA's: 1
##          Mean   :41.2
##          3rd Qu.:50.0
##          Max.   :78.0
##
##  Anorexia   LiverBig   LiverFirm SpleenPalpable Spiders   Ascites    Varices
##  0   :122   0   :120   0   :84   0   :120       0   :99   0   :130   0   :132
##  1   : 32   1   : 25   1   :60   1   : 30       1   :51   1   : 20   1   : 18
##  NA's:  1   NA's: 10   NA's:11   NA's:  5       NA's: 5   NA's:  5   NA's:  5
##
##
```

```
##
##
##    Bilirubin      AlkPhosphate         Sgot           Albumin
##  Min.   :0.300   Min.   : 26.00   Min.   : 14.00   Min.   :2.100
##  1st Qu.:0.700   1st Qu.: 74.25   1st Qu.: 31.50   1st Qu.:3.400
##  Median :1.000   Median : 85.00   Median : 58.00   Median :4.000
##  Mean   :1.428   Mean   :105.33   Mean   : 85.89   Mean   :3.817
##  3rd Qu.:1.500   3rd Qu.:132.25   3rd Qu.:100.50   3rd Qu.:4.200
##  Max.   :8.000   Max.   :295.00   Max.   :648.00   Max.   :6.400
##  NA's   :6       NA's   :29       NA's   :4        NA's   :16
##     Protime       Histology
##  Min.   :  0.00   0:70
##  1st Qu.: 46.00   1:85
##  Median : 61.00
##  Mean   : 61.85
##  3rd Qu.: 76.25
##  Max.   :100.00
##  NA's   :67
```

```
summary(df1)
```

```
##      Class             Age            Sex            Steroid
##  Min.   :0.0000   Min.   : 7.0   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:32.0   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0.0000   Median :39.0   Median :1.0000   Median :0.0000
##  Mean   :0.2065   Mean   :41.2   Mean   :0.8968   Mean   :0.4903
##  3rd Qu.:0.0000   3rd Qu.:50.0   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :78.0   Max.   :1.0000   Max.   :1.0000
##    Antivirals        Fatigue          Malaise          Anorexia
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1548   Mean   :0.6452   Mean   :0.3935   Mean   :0.2065
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     LiverBig        LiverFirm   SpleenPalpable    Spiders          Ascites
##  Min.   :0.0000   Min.   :0.0   Min.   :0.0    Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0   1st Qu.:0.0    1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0   Median :0.0    Median :0.0000   Median :0.000
##  Mean   :0.1677   Mean   :0.4   Mean   :0.2    Mean   :0.3419   Mean   :0.129
##  3rd Qu.:0.0000   3rd Qu.:1.0   3rd Qu.:0.0    3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.0   Max.   :1.0    Max.   :1.0000   Max.   :1.000
##     Varices         Bilirubin      AlkPhosphate        Sgot
##  Min.   :0.0000   Min.   :0.300   Min.   : 26.0   Min.   : 14.00
##  1st Qu.:0.0000   1st Qu.:0.800   1st Qu.: 75.0   1st Qu.: 32.50
##  Median :0.0000   Median :1.000   Median : 85.0   Median : 58.00
##  Mean   :0.1161   Mean   :1.415   Mean   :103.3   Mean   : 85.57
##  3rd Qu.:0.0000   3rd Qu.:1.500   3rd Qu.:126.0   3rd Qu.: 99.00
##  Max.   :1.0000   Max.   :8.000   Max.   :295.0   Max.   :648.00
##     Albumin         Protime         Histology
##  Min.   :2.100   Min.   :  0.00   Min.   :0.0000
##  1st Qu.:3.500   1st Qu.: 50.50   1st Qu.:0.0000
##  Median :4.000   Median : 64.00   Median :1.0000
##  Mean   :3.821   Mean   : 63.89   Mean   :0.5484
##  3rd Qu.:4.200   3rd Qu.: 78.50   3rd Qu.:1.0000
##  Max.   :6.400   Max.   :100.00   Max.   :1.0000
```

```r
summary(df2)
```

```
##      Class            Age             Sex            Steroid
##  Min.   :0.0000   Min.   : 7.0   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:32.0   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0.0000   Median :39.0   Median :1.0000   Median :0.0000
##  Mean   :0.2065   Mean   :41.2   Mean   :0.8968   Mean   :0.4903
##  3rd Qu.:0.0000   3rd Qu.:50.0   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :78.0   Max.   :1.0000   Max.   :1.0000
##    Antivirals        Fatigue          Malaise          Anorexia
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1548   Mean   :0.6452   Mean   :0.3935   Mean   :0.2065
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     LiverBig        LiverFirm      SpleenPalpable    Spiders
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0    Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0    1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0    Median :0.0000
##  Mean   :0.1613   Mean   :0.4129   Mean   :0.2    Mean   :0.3419
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0    3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0    Max.   :1.0000
##     Ascites          Varices          Bilirubin      AlkPhosphate
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.300   Min.   : 26.0
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.800   1st Qu.: 73.5
##  Median :0.0000   Median :0.0000   Median :1.000   Median : 85.0
##  Mean   :0.1419   Mean   :0.1161   Mean   :1.415   Mean   :104.2
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.500   3rd Qu.:129.5
##  Max.   :1.0000   Max.   :1.0000   Max.   :8.000   Max.   :295.0
##       Sgot           Albumin          Protime          Histology
##  Min.   : 14.00   Min.   :2.100   Min.   :  0.00   Min.   :0.0000
##  1st Qu.: 32.50   1st Qu.:3.500   1st Qu.: 47.00   1st Qu.:0.0000
##  Median : 59.00   Median :4.000   Median : 60.00   Median :1.0000
##  Mean   : 85.83   Mean   :3.829   Mean   : 61.85   Mean   :0.5484
##  3rd Qu.:100.50   3rd Qu.:4.200   3rd Qu.: 74.50   3rd Qu.:1.0000
##  Max.   :648.00   Max.   :6.400   Max.   :100.00   Max.   :1.0000
```

```r
summary(df3)
```

```
##      Class            Age             Sex            Steroid
##  Min.   :0.0000   Min.   : 7.0   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:32.0   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0.0000   Median :39.0   Median :1.0000   Median :0.0000
##  Mean   :0.2065   Mean   :41.2   Mean   :0.8968   Mean   :0.4903
##  3rd Qu.:0.0000   3rd Qu.:50.0   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :78.0   Max.   :1.0000   Max.   :1.0000
##    Antivirals        Fatigue          Malaise          Anorexia
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1548   Mean   :0.6516   Mean   :0.3935   Mean   :0.2065
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##     LiverBig        LiverFirm      SpleenPalpable      Spiders     
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000  
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000  
##  Mean   :0.1613   Mean   :0.3871   Mean   :0.1935   Mean   :0.329  
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000  
##     Ascites         Varices          Bilirubin      AlkPhosphate  
##  Min.   :0.000   Min.   :0.0000   Min.   :0.300   Min.   : 26.0  
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.800   1st Qu.: 78.0  
##  Median :0.000   Median :0.0000   Median :1.000   Median : 85.0  
##  Mean   :0.129   Mean   :0.1161   Mean   :1.411   Mean   :101.5  
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.500   3rd Qu.:119.5  
##  Max.   :1.000   Max.   :1.0000   Max.   :8.000   Max.   :295.0  
##       Sgot           Albumin         Protime         Histology     
##  Min.   : 14.00   Min.   :2.100   Min.   :  0.00   Min.   :0.0000  
##  1st Qu.: 32.50   1st Qu.:3.500   1st Qu.: 57.00   1st Qu.:0.0000  
##  Median : 58.00   Median :4.000   Median : 61.00   Median :1.0000  
##  Mean   : 85.17   Mean   :3.836   Mean   : 61.48   Mean   :0.5484  
##  3rd Qu.: 99.00   3rd Qu.:4.200   3rd Qu.: 65.00   3rd Qu.:1.0000  
##  Max.   :648.00   Max.   :6.400   Max.   :100.00   Max.   :1.0000  
```

```
summary(df4)
```

```
##  Class         Age        Sex     Steroid Antivirals Fatigue Malaise Anorexia
##  0:123   Min.   : 7.0   0: 16   0:78    0:131      0: 54   0:94    0:123   
##  1: 32   1st Qu.:32.0   1:139   1:77    1: 24      1:101   1:61    1: 32   
##          Median :39.0                                                      
##          Mean   :41.2                                                      
##          3rd Qu.:50.0                                                      
##          Max.   :78.0                                                      
##  LiverBig LiverFirm SpleenPalpable Spiders Ascites Varices   Bilirubin    
##  0:128    0:93      0:124          0:102   0:134   0:136   Min.   :0.300  
##  1: 27    1:62      1: 31          1: 53   1: 21   1: 19   1st Qu.:0.750  
##                                                            Median :1.000  
##                                                            Mean   :1.421  
##                                                            3rd Qu.:1.500  
##                                                            Max.   :8.000  
##   AlkPhosphate        Sgot           Albumin         Protime       Histology
##  Min.   : 26.0   Min.   : 14.00   Min.   :2.100   Min.   :  0.00   0:70     
##  1st Qu.: 75.0   1st Qu.: 31.50   1st Qu.:3.400   1st Qu.: 47.00   1:85     
##  Median : 85.0   Median : 58.00   Median :4.000   Median : 63.00            
##  Mean   :108.2   Mean   : 85.58   Mean   :3.814   Mean   : 63.34            
##  3rd Qu.:134.0   3rd Qu.:100.50   3rd Qu.:4.200   3rd Qu.: 82.00            
##  Max.   :295.0   Max.   :648.00   Max.   :6.400   Max.   :100.00            
```

```
summary(df5)
```

```
##  Class         Age        Sex     Steroid Antivirals Fatigue Malaise Anorexia
##  0:123   Min.   : 7.0   0: 16   0:79    0:131      0: 54   0:94    0:123   
##  1: 32   1st Qu.:32.0   1:139   1:76    1: 24      1:101   1:61    1: 32   
##          Median :39.0                                                      
##          Mean   :41.2                                                      
##          3rd Qu.:50.0                                                      
##          Max.   :78.0
##  LiverBig LiverFirm SpleenPalpable Spiders Ascites Varices   Bilirubin
##  0:127    0:88      0:122          0:102   0:134   0:136   Min.   :0.300
##  1: 28    1:67      1: 33          1: 53   1: 21   1: 19   1st Qu.:0.700
##                                                            Median :1.000
##                                                            Mean   :1.421
##                                                            3rd Qu.:1.500
##                                                            Max.   :8.000
##   AlkPhosphate        Sgot           Albumin         Protime      Histology
##  Min.   : 26.0   Min.   : 14.00   Min.   :2.100   Min.   :  0.0   0:70
##  1st Qu.: 74.5   1st Qu.: 31.50   1st Qu.:3.400   1st Qu.: 46.0   1:85
##  Median : 85.0   Median : 58.00   Median :4.000   Median : 60.0
##  Mean   :105.2   Mean   : 87.13   Mean   :3.833   Mean   : 61.6
##  3rd Qu.:131.5   3rd Qu.:105.50   3rd Qu.:4.200   3rd Qu.: 77.0
##  Max.   :295.0   Max.   :648.00   Max.   :6.400   Max.   :100.0
```

```
colVars(na.omit(df))
```

```
##          Class            Age            Sex        Steroid     Antivirals
##             NA    127.2390823             NA             NA             NA
##        Fatigue        Malaise       Anorexia       LiverBig      LiverFirm
##             NA             NA             NA             NA             NA
## SpleenPalpable        Spiders        Ascites        Varices      Bilirubin
##             NA             NA             NA             NA      0.7659984
##   AlkPhosphate           Sgot        Albumin        Protime      Histology
##   2882.0555380   5126.5563291      0.3321123    548.8606013             NA
```

```
colVars(df2)
```

```
##          Class            Age            Sex        Steroid     Antivirals
##   1.648932e-01   1.579013e+02   9.317134e-02   2.515291e-01   1.317134e-01
##        Fatigue        Malaise       Anorexia       LiverBig      LiverFirm
##   2.304147e-01   2.402178e-01   1.648932e-01   1.361542e-01   2.439883e-01
## SpleenPalpable        Spiders        Ascites        Varices      Bilirubin
##   1.610390e-01   2.264767e-01   1.225806e-01   1.033096e-01   1.418330e+00
##   AlkPhosphate           Sgot        Albumin        Protime      Histology
##   2.354365e+03   7.838145e+03   3.892900e-01   4.013265e+02   2.492669e-01
```

```
colVars(df3)
```

```
##          Class            Age            Sex        Steroid     Antivirals
##   1.648932e-01   1.579013e+02   9.317134e-02   2.515291e-01   1.317134e-01
##        Fatigue        Malaise       Anorexia       LiverBig      LiverFirm
##   2.284876e-01   2.402178e-01   1.648932e-01   1.361542e-01   2.387935e-01
## SpleenPalpable        Spiders        Ascites        Varices      Bilirubin
##   1.571010e-01   2.222036e-01   1.131127e-01   1.033096e-01   1.418905e+00
##   AlkPhosphate           Sgot        Albumin        Protime      Histology
##   2.216719e+03   7.848210e+03   3.834914e-01   2.957968e+02   2.492669e-01
```

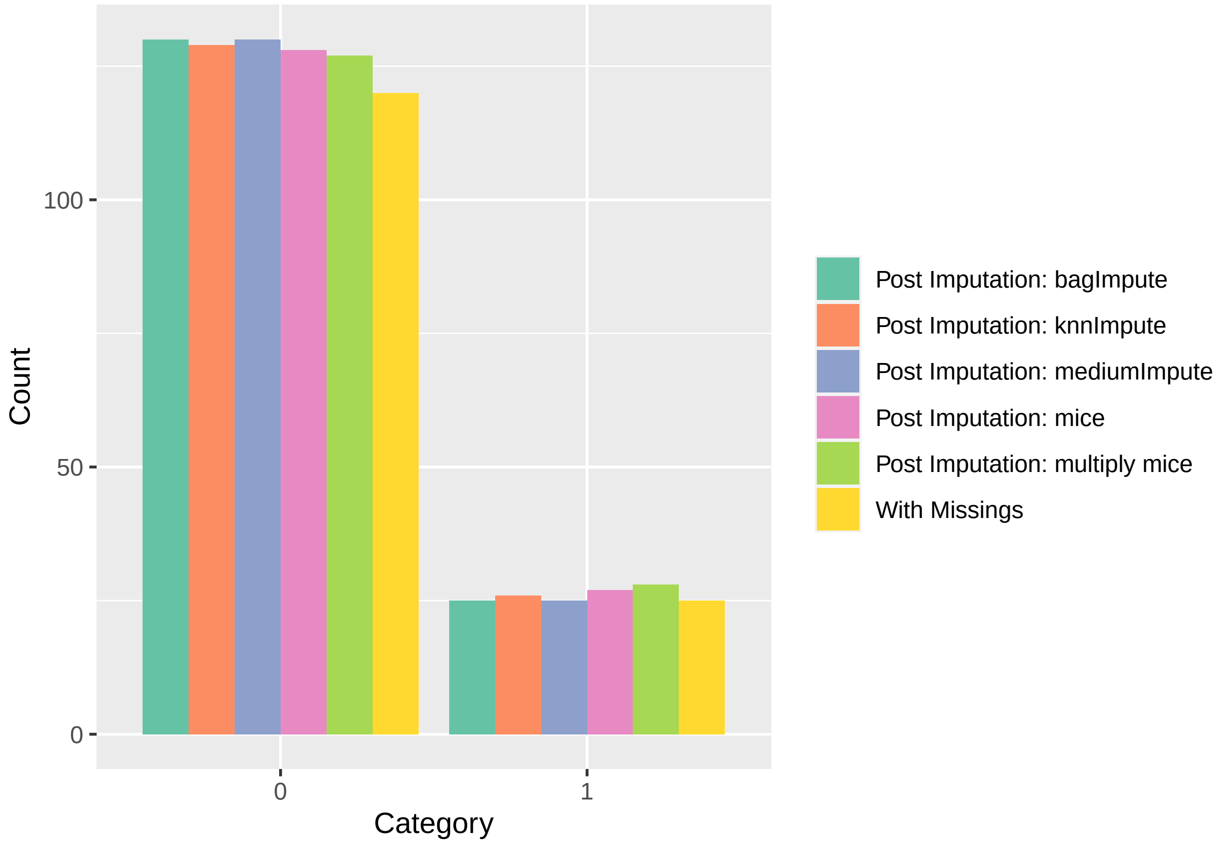
```
colVars(df4)
```

```
##          Class            Age            Sex        Steroid     Antivirals
##             NA     157.901299             NA             NA             NA
##        Fatigue        Malaise       Anorexia       LiverBig      LiverFirm
##             NA             NA             NA             NA             NA
## SpleenPalpable        Spiders        Ascites        Varices      Bilirubin
##             NA             NA             NA             NA       1.425842
##   AlkPhosphate          Sgot       Albumin       Protime     Histology
##    2802.348303   7924.738584      0.410057    516.161542            NA
```

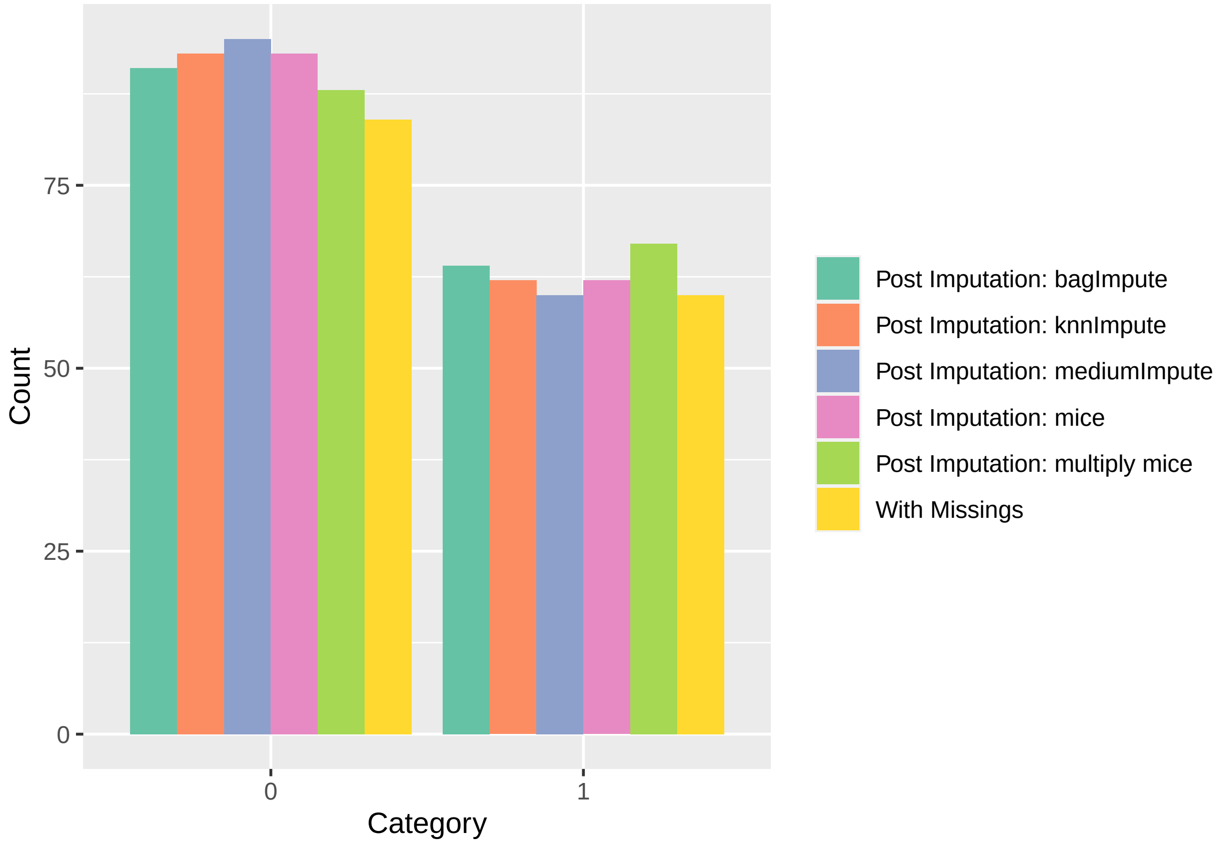
```
colVars(df5)
```

```
##          Class           Age           Sex       Steroid    Antivirals
##             NA   157.9012987            NA            NA            NA
##        Fatigue       Malaise      Anorexia      LiverBig     LiverFirm
##             NA            NA            NA            NA            NA
## SpleenPalpable       Spiders       Ascites       Varices     Bilirubin
##             NA            NA            NA            NA     1.4251554
##   AlkPhosphate          Sgot       Albumin       Protime     Histology
##   2455.2959363  8155.9832426     0.4631311   519.1246753            NA
```

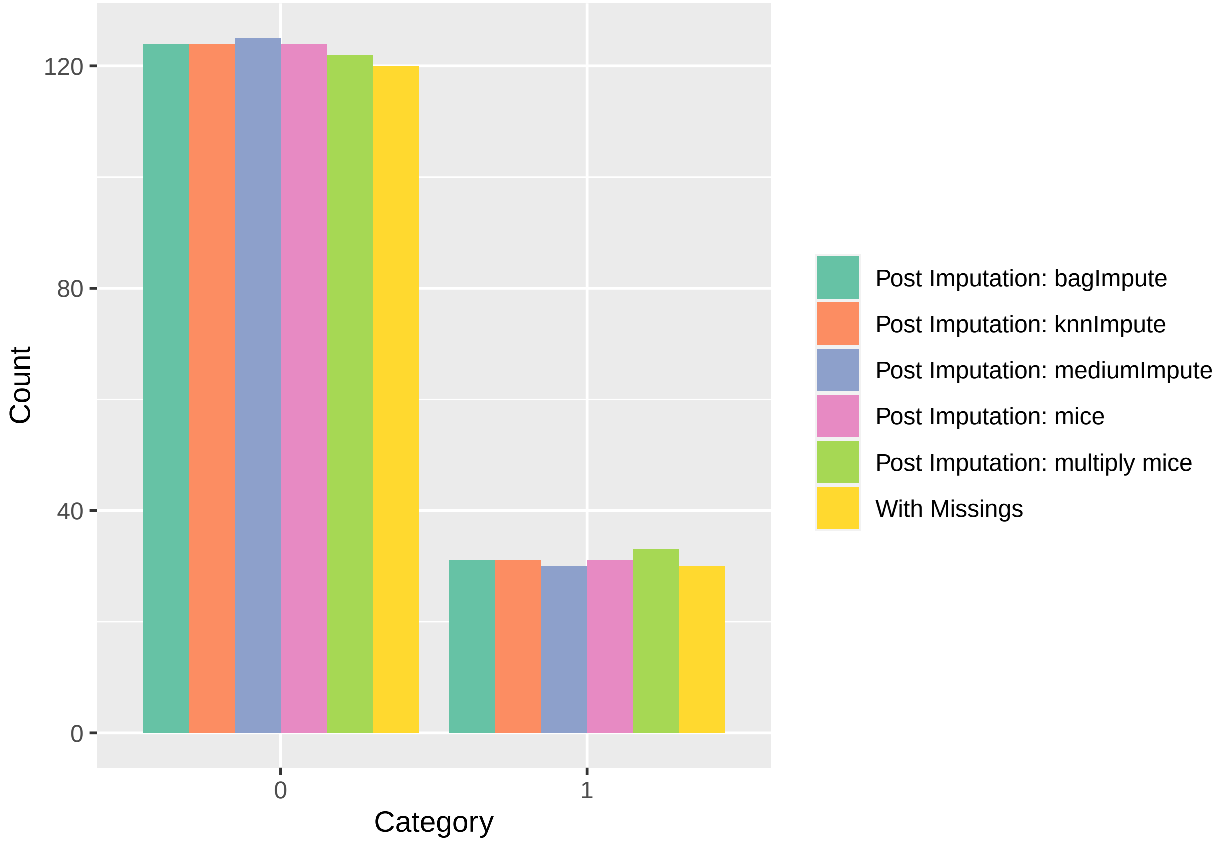
```
n <- 9
missingness <- c(rep("With Missings", 2), rep("Post Imputation: knnImpute", 2), rep("Post Imputation: ba
Category <- as.factor(rep(names(table(df[n])), 6))                       # Categories
Count <- c(as.numeric(table(df[n])), as.numeric(table(df1[n])), as.numeric(table(df2[n])), as.numeric(ta

data_barplot <- data.frame(missingness, Category, Count)            # Combine data for plot

ggplot(data_barplot, aes(Category, Count, fill = missingness)) +    # Create plot
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set2") +
  theme(legend.title = element_blank())
```



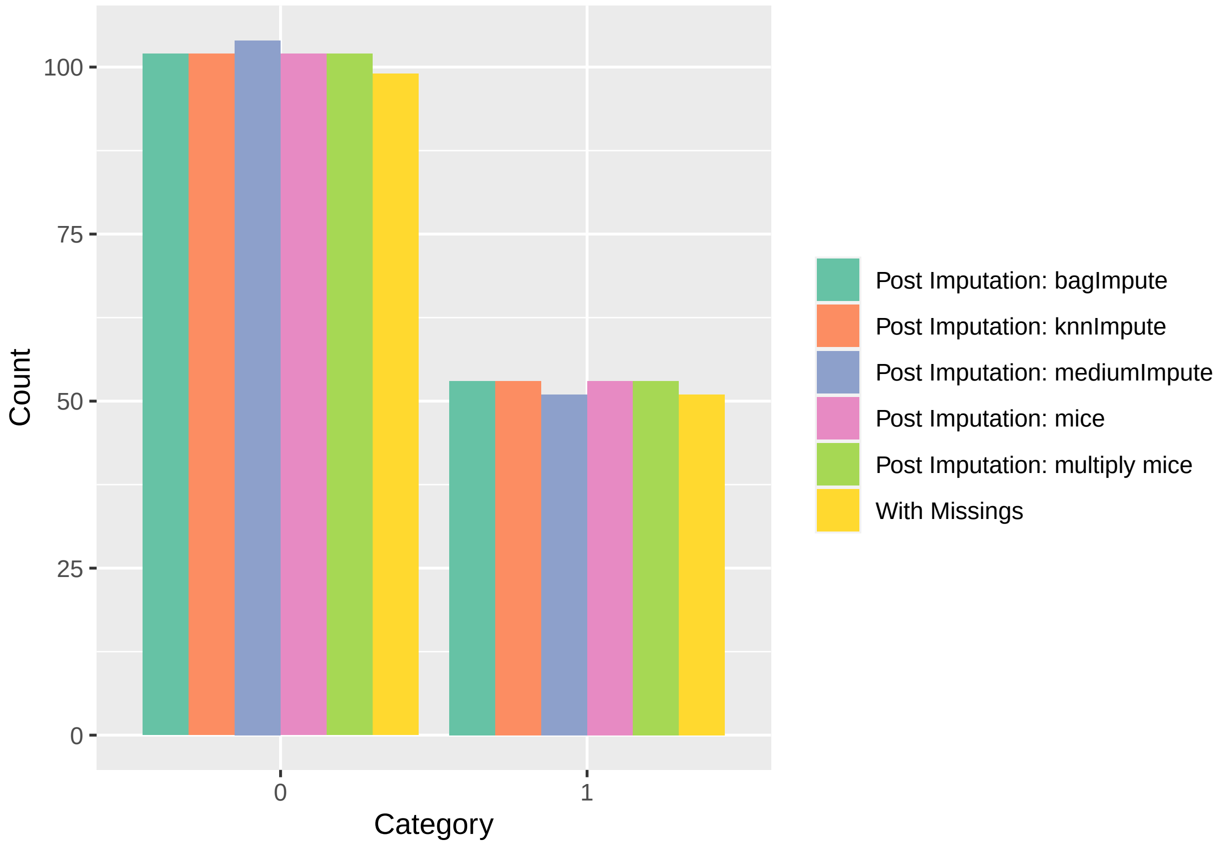
```
n <- 10
missingness <- c(rep("With Missings", 2), rep("Post Imputation: knnImpute", 2), rep("Post Imputation: ba
Category <- as.factor(rep(names(table(df[n])), 6))                   # Categories
Count <- c(as.numeric(table(df[n])), as.numeric(table(df1[n])), as.numeric(table(df2[n])), as.numeric(ta

data_barplot <- data.frame(missingness, Category, Count)            # Combine data for plot

ggplot(data_barplot, aes(Category, Count, fill = missingness)) +   # Create plot
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set2") +
  theme(legend.title = element_blank())
```



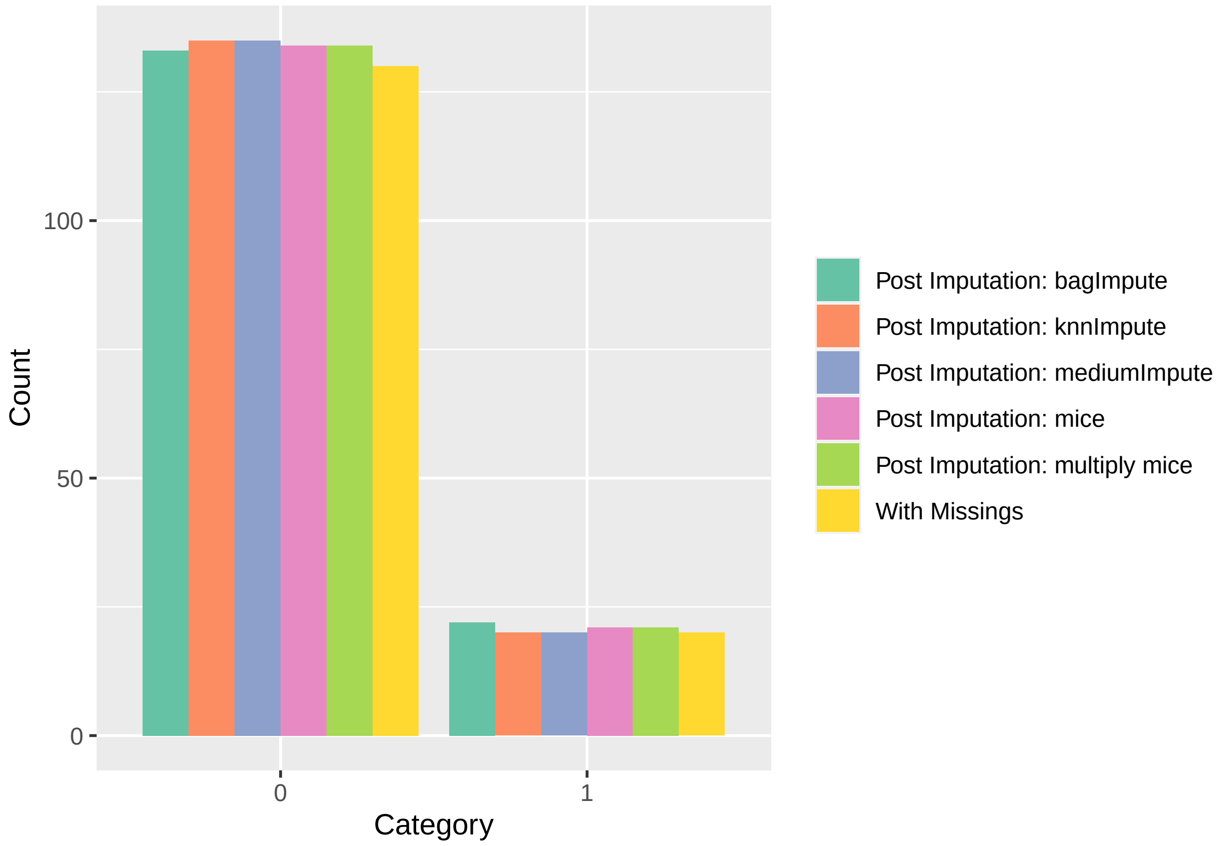
```
n <- 11
missingness <- c(rep("With Missings", 2), rep("Post Imputation: knnImpute", 2), rep("Post Imputation: ba
Category <- as.factor(rep(names(table(df[n])), 6))                   # Categories
Count <- c(as.numeric(table(df[n])), as.numeric(table(df1[n])), as.numeric(table(df2[n])), as.numeric(ta

data_barplot <- data.frame(missingness, Category, Count)            # Combine data for plot

ggplot(data_barplot, aes(Category, Count, fill = missingness)) +   # Create plot
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set2") +
  theme(legend.title = element_blank())
```



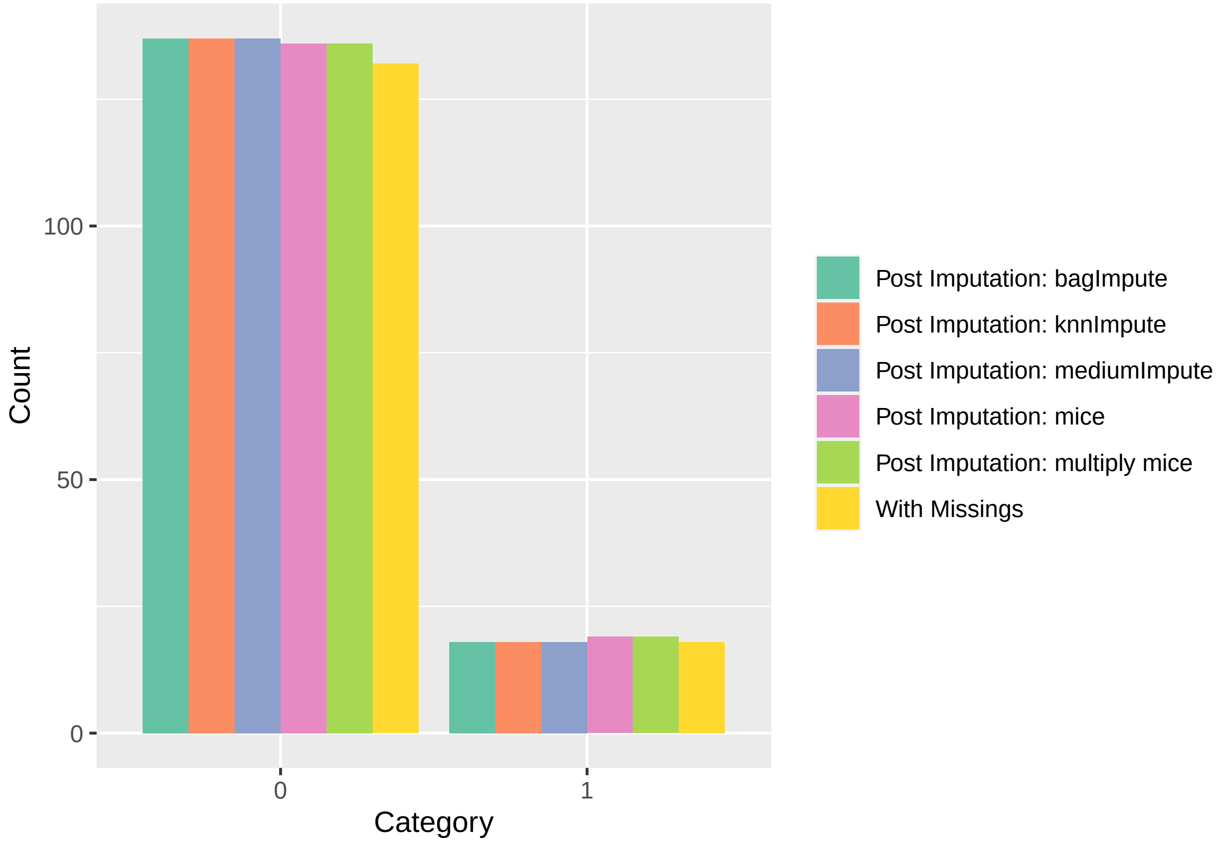
```
n <- 12
missingness <- c(rep("With Missings", 2), rep("Post Imputation: knnImpute", 2), rep("Post Imputation: ba
Category <- as.factor(rep(names(table(df[n])), 6))                    # Categories
Count <- c(as.numeric(table(df[n])), as.numeric(table(df1[n])), as.numeric(table(df2[n])), as.numeric(ta

data_barplot <- data.frame(missingness, Category, Count)           # Combine data for plot

ggplot(data_barplot, aes(Category, Count, fill = missingness)) +   # Create plot
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set2") +
  theme(legend.title = element_blank())
```



```
n <- 13
missingness <- c(rep("With Missings", 2), rep("Post Imputation: knnImpute", 2), rep("Post Imputation: ba
Category <- as.factor(rep(names(table(df[n])), 6))                   # Categories
Count <- c(as.numeric(table(df[n])), as.numeric(table(df1[n])), as.numeric(table(df2[n])), as.numeric(ta

data_barplot <- data.frame(missingness, Category, Count)           # Combine data for plot

ggplot(data_barplot, aes(Category, Count, fill = missingness)) +   # Create plot
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set2") +
  theme(legend.title = element_blank())
```

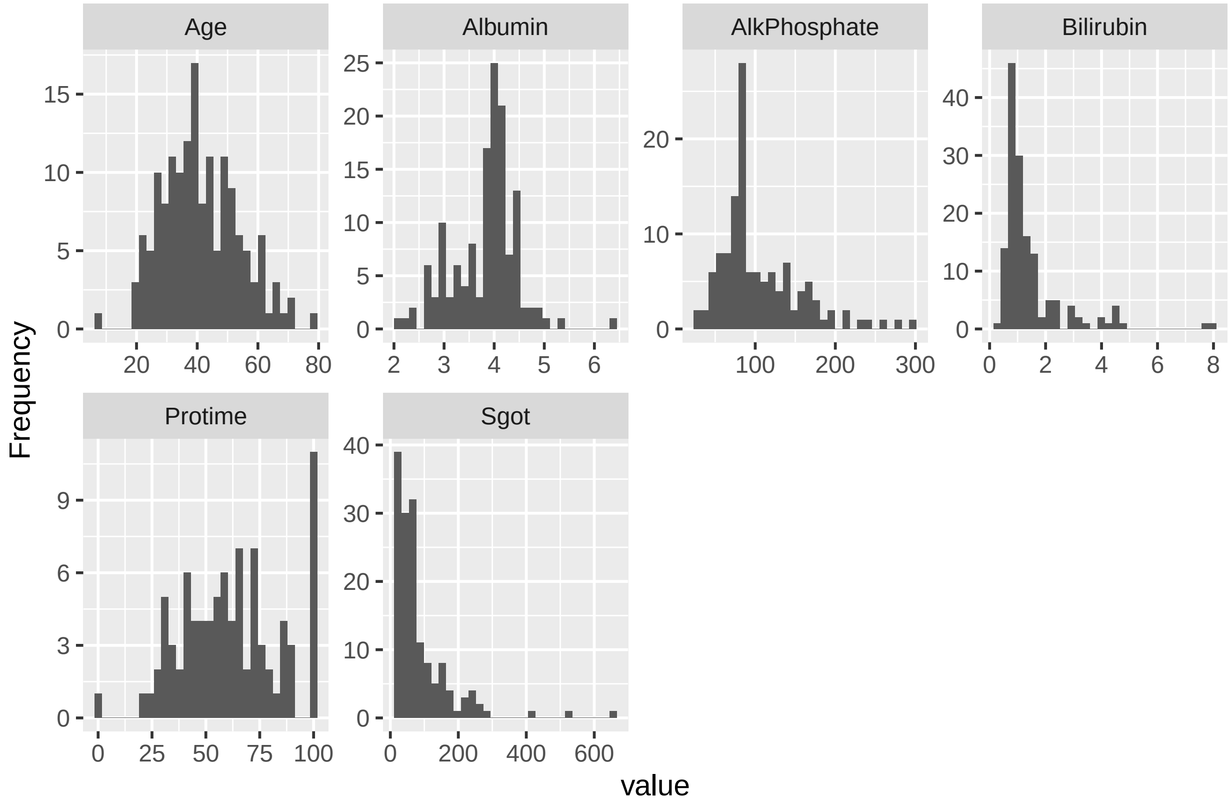


```
n <- 14
missingness <- c(rep("With Missings", 2), rep("Post Imputation: knnImpute", 2), rep("Post Imputation: ba
Category <- as.factor(rep(names(table(df[n])), 6))                    # Categories
Count <- c(as.numeric(table(df[n])), as.numeric(table(df1[n])), as.numeric(table(df2[n])), as.numeric(ta

data_barplot <- data.frame(missingness, Category, Count)          # Combine data for plot

ggplot(data_barplot, aes(Category, Count, fill = missingness)) +   # Create plot
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_brewer(palette = "Set2") +
  theme(legend.title = element_blank())
```

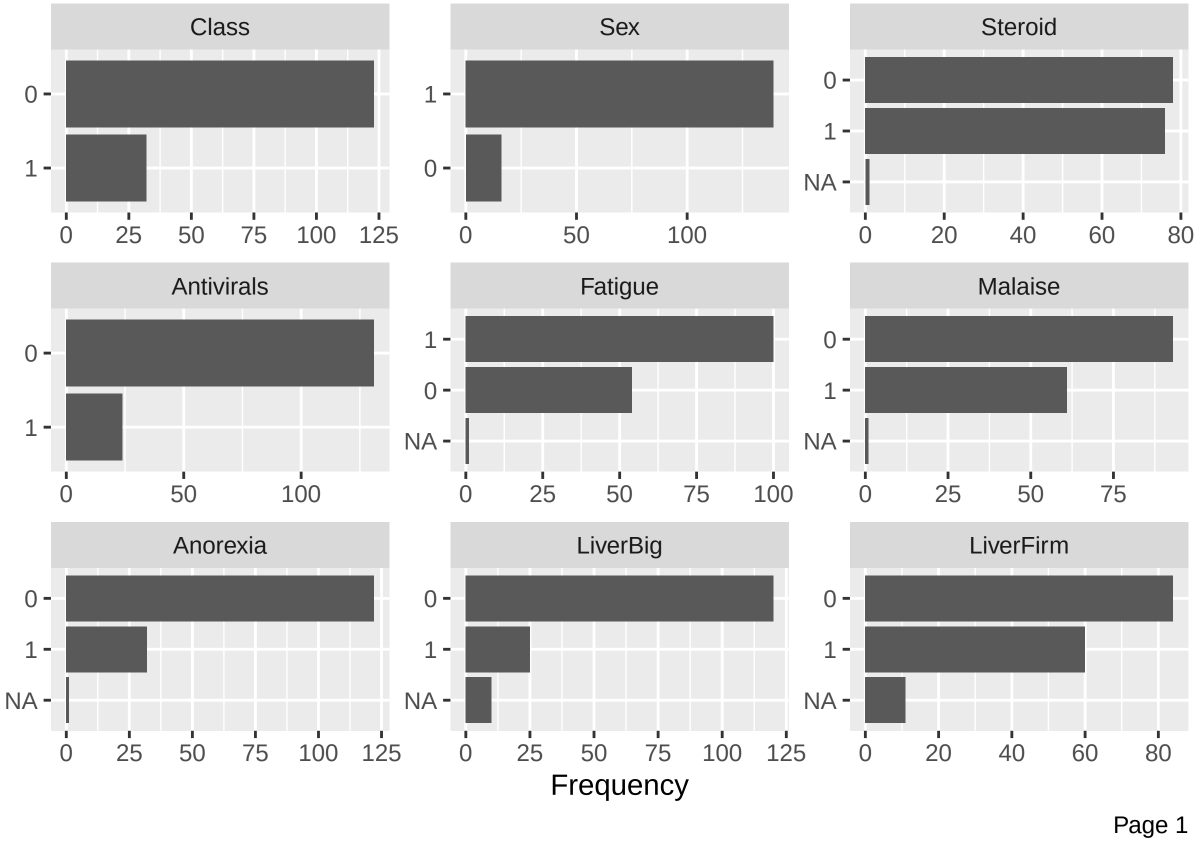


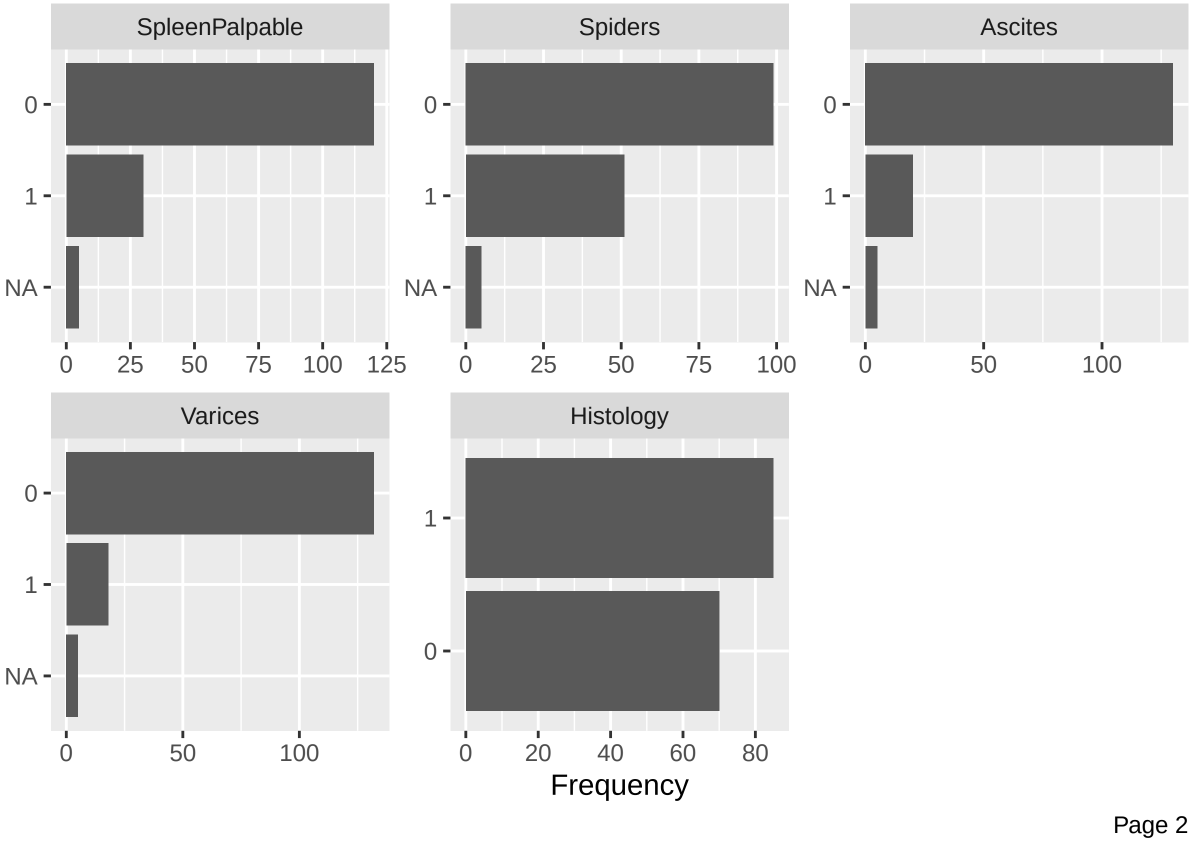
```
{r echo=FALSE} # col.cat.with.mv <- c(4) # imputeEM_none <-
multinomial_impute(df[col.cat.with.mv], method = "EM", conj_prior
= "non.informative", verbose = TRUE) #
```

```r
plot_histogram(df)
```


Age
Albumin
AlkPhosphate
Bilirubin
Protime
Sgot
Frequency
value
15
10
5
0
20
40
60
80
25
20
15
10
5
0
2
3
4
5
6
20
10
0
100
200
300
40
30
20
10
0
0
2
4
6
8
9
6
3
0
0
25
50
75
100
40
30
20
10
0
0
200
400
600

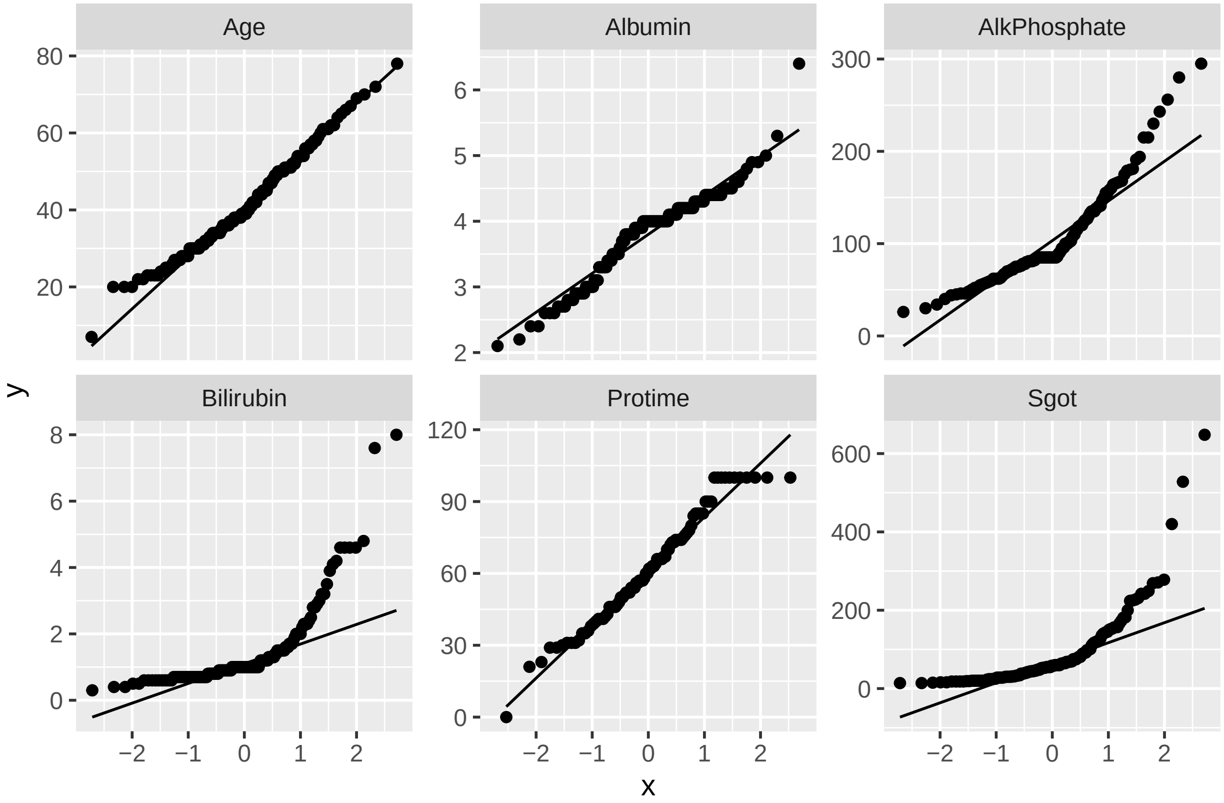
```
plot_bar(df)
```


Class
0
1
0
25
50
75
100
125
Sex
1
0
0
50
100
Steroid
0
1
NA
0
20
40
60
80
Antivirals
0
1
0
50
100
Fatigue
1
0
NA
0
25
50
75
100
Malaise
0
1
NA
0
25
50
75
Anorexia
0
1
NA
0
25
50
75
100
125
LiverBig
0
1
NA
0
25
50
75
100
125
LiverFirm
0
1
NA
0
20
40
60
80
Frequency
Page 1



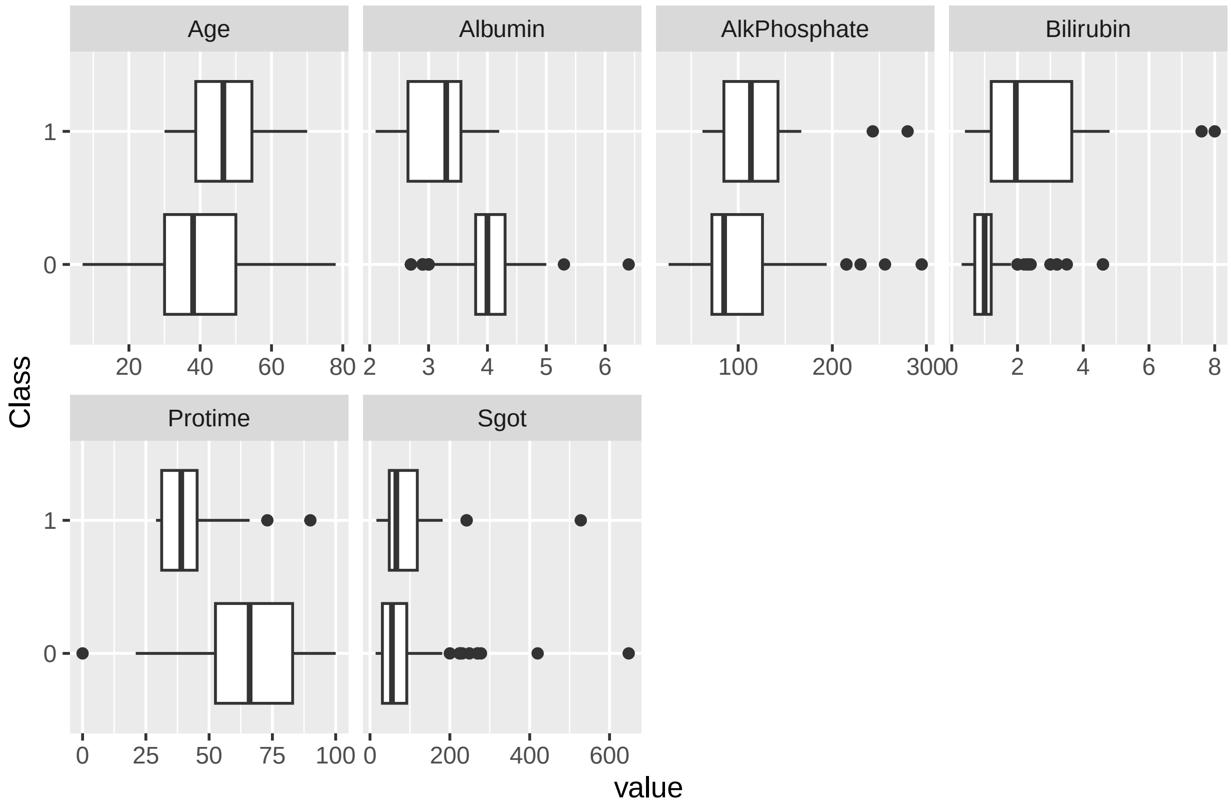
```r
plot_qq(df)
```

```
## Warning: Removed 122 rows containing non-finite values (`stat_qq()`).

## Warning: Removed 122 rows containing non-finite values (`stat_qq_line()`).
```



```
plot_boxplot(df, by="Class")
```

```
## Warning: Removed 122 rows containing non-finite values (`stat_boxplot()`).
```



```
## Charts: continuous quantitative data
# Typical questions:
  #
  # - Uni- or multi-modal distribution?
  # - Symmetric or left- / right-skewed distribution?
  # - Mode / modal interval?
  # - Does the distribution resemble a normal, uniform distribution?

## Charts: discrete quantitative data
# Typical questions:
  #
  #   - The most frequent value?
  #   - min / max?
  #   - Frequencies of consecutive values?

## Charts: qualitative data
  # Analysis within groups: Why do customers leave?
```