Project

Claudio Previte and Ana -Maria Casian

automn 2019

Contents

```r
library(here)
library(Hmisc)
library(dplyr)
library(DataExplorer)
library(ggplot2)
library(lattice)
library(inspectdf)
```

Loading data

```r
bands <- read.csv2(file = here('bands3.csv'), sep = ';',na.strings = "?")
```

EDA

#Summary bands

```r
str(bands)
```

```
## 'data.frame':    540 obs. of  40 variables:
##  $ date             : int  19910108 19910109 19910104 19910104 19910111 19910104 19910111 1991011
##  $ cylinder_no      : Factor w/ 434 levels "1351","3","aa067",..: 297 338 22 249 176 255 332 401
##  $ customer         : Factor w/ 83 levels "ABBEY","ABBEYPRESS",..: 75 75 61 58 55 58 67 67 61 18
##  $ job_number       : int  25503 25503 47201 39039 37351 38039 35751 35751 47201 37000 ...
##  $ grain_screened   : Factor w/ 2 levels "NO","YES": 2 2 2 2 1 2 1 1 2 2 ...
##  $ ink_color        : Factor w/ 3 levels "key","KeY","KEY": 3 3 3 3 3 3 3 3 3 3 ...
##  $ proof_ink        : Factor w/ 2 levels "NO","YES": 2 2 2 2 2 2 2 2 2 2 ...
##  $ blade_mfg        : Factor w/ 2 levels "BENTON","UDDEHOLM": 1 1 1 1 1 1 1 1 1 1 ...
##  $ cylinder_division: Factor w/ 2 levels "gallatin","GALLATIN": 2 2 2 2 2 2 2 2 2 2 ...
##  $ paper_type       : Factor w/ 5 levels "coated","COATED",..: 5 5 5 5 5 5 2 2 5 5 ...
##  $ ink_type         : Factor w/ 6 levels "coated","COATED",..: 6 6 2 6 2 6 2 2 6 6 ...
##  $ direct_steam     : Factor w/ 3 levels "no","NO","YES": 2 2 2 2 2 2 2 2 2 2 ...
##  $ solvent_type     : Factor w/ 3 levels "LINE","NAPTHA",..: 1 1 1 1 1 1 1 1 3 1 ...
##  $ cylinder_type    : Factor w/ 4 levels "no","NO","yes",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ press_type       : Factor w/ 4 levels "Albert70","Motter70",..: 3 3 4 4 4 4 3 3 1 4 ...
##  $ press            : int  821 821 815 816 816 816 827 827 802 815 ...
##  $ unit_number      : int  2 2 9 9 2 2 2 9 7 2 ...
##  $ cylinder_size    : Factor w/ 6 levels "catalog","CATALOG",..: 6 6 2 2 6 2 6 6 2 2 ...
##  $ location         : Factor w/ 6 levels "CANAdiAN","CANADIAN",..: 4 4 4 4 NA 4 2 1 4 4 ...
##  $ plating_tank     : int  1911 NA NA 1910 1910 1910 1911 1911 1910 1911 ...
## $ proof_cut        : Factor w/ 27 levels "25","27.5","30",..: 18 18 22 16 15 15 15 15 15 24 ...
## $ viscosity        : int  46 46 40 40 46 40 46 46 45 43 ...
## $ caliper          : Factor w/ 20 levels ".200","0.133",..: 4 11 17 11 11 9 11 4 15 13 ...
## $ ink_temperature  : Factor w/ 66 levels "11.2","12","12.5",..: 55 29 42 42 55 54 50 50 2 42 ...
## $ humidity         : int  78 80 80 75 80 76 75 75 70 75 ...
## $ roughness        : Factor w/ 22 levels ".625","0.05625",..: 12 12 NA 5 12 7 12 12 12 19 ...
## $ blade_pressure   : int  20 20 30 30 30 28 30 28 60 32 ...
## $ varnish_pct      : Factor w/ 124 levels "0","0.5","1",..: 27 99 98 92 1 114 1 1 1 72 ...
## $ press_speed      : int  1700 1900 1850 1467 2100 1467 2600 2600 1650 1750 ...
## $ ink_pct          : Factor w/ 81 levels "41","41.3","41.7",..: 26 42 38 45 52 38 71 71 65 12 ..
## $ solvent_pct      : Factor w/ 116 levels "22","22.5","23.1",..: 43 62 74 64 96 54 53 53 74 10 .
## $ ESA_voltage      : Factor w/ 17 levels "0","0.5","0.75",..: 1 1 1 1 14 14 15 15 5 1 ...
## $ ESA_amperage     : Factor w/ 4 levels "0","0.5","4",..: 1 1 1 1 1 1 1 1 1 1 ...
## $ wax              : Factor w/ 32 levels "0","0.5","0.7",..: 23 23 27 23 21 23 23 23 30 30 ...
## $ hardener         : Factor w/ 32 levels ".8","0","0.2",..: 14 9 13 19 7 11 7 16 14 14 ...
## $ roller_durometer : Factor w/ 12 levels "28","30","32",..: 5 5 9 9 6 9 2 2 9 7 ...
## $ current_density  : int  40 40 40 40 40 40 40 40 40 40 ...
## $ anode_space_ratio: Factor w/ 81 levels "100","100.0",..: 21 21 20 41 34 20 34 34 15 33 ...
## $ chrome_content   : int  100 100 100 100 100 100 100 100 100 100 ...
## $ band_type        : Factor w/ 2 levels "band","noband": 1 2 2 2 2 2 2 2 1 2 ...
```

```
# summary(bands)
```

#Data Transformation

```
cols = c(21:39)
bands[,cols] = apply(bands[,cols], 2, function(x) as.numeric(as.character(x))) #change class to nume

#make sure that the variables are well defined
str(bands)
```

```
## 'data.frame':    540 obs. of  40 variables:
## $ date             : int  19910108 19910109 19910104 19910104 19910111 19910104 19910111 1991011
## $ cylinder_no      : Factor w/ 434 levels "1351","3","aa067",..: 297 338 22 249 176 255 332 401
## $ customer         : Factor w/ 83 levels "ABBEY","ABBEYPRESS",..: 75 75 61 58 55 58 67 67 61 18
## $ job_number       : int  25503 25503 47201 39039 37351 38039 35751 35751 47201 37000 ...
## $ grain_screened   : Factor w/ 2 levels "NO","YES": 2 2 2 2 1 2 1 1 2 2 ...
## $ ink_color        : Factor w/ 3 levels "key","KeY","KEY": 3 3 3 3 3 3 3 3 3 3 ...
## $ proof_ink        : Factor w/ 2 levels "NO","YES": 2 2 2 2 2 2 2 2 2 2 ...
## $ blade_mfg        : Factor w/ 2 levels "BENTON","UDDEHOLM": 1 1 1 1 1 1 1 1 1 1 ...
## $ cylinder_division: Factor w/ 2 levels "gallatin","GALLATIN": 2 2 2 2 2 2 2 2 2 2 ...
## $ paper_type       : Factor w/ 5 levels "coated","COATED",..: 5 5 5 5 5 5 2 2 5 5 ...
## $ ink_type         : Factor w/ 6 levels "coated","COATED",..: 6 6 2 6 2 6 2 2 6 6 ...
## $ direct_steam     : Factor w/ 3 levels "no","NO","YES": 2 2 2 2 2 2 2 2 2 2 ...
## $ solvent_type     : Factor w/ 3 levels "LINE","NAPTHA",..: 1 1 1 1 1 1 1 1 3 1 ...
## $ cylinder_type    : Factor w/ 4 levels "no","NO","yes",..: 4 4 4 4 4 4 4 4 4 4 ...
## $ press_type       : Factor w/ 4 levels "Albert70","Motter70",..: 3 3 4 4 4 4 3 3 1 4 ...
## $ press            : int  821 821 815 816 816 816 827 827 802 815 ...
## $ unit_number      : int  2 2 9 9 2 2 2 9 7 2 ...
## $ cylinder_size    : Factor w/ 6 levels "catalog","CATALOG",..: 6 6 2 2 6 2 6 6 2 2 ...
## $ location         : Factor w/ 6 levels "CANAdiAN","CANADIAN",..: 4 4 4 4 NA 4 2 1 4 4 ...
## $ plating_tank     : int  1911 NA NA 1910 1910 1910 1911 1911 1910 1911 ...
## $ proof_cut        : num  55 55 62 52 50 50 50 50 50 65 ...
## $ viscosity        : num  46 46 40 40 46 40 46 46 45 43 ...
## $ caliper          : num  0.2 0.3 0.433 0.3 0.3 0.267 0.3 0.2 0.367 0.333 ...
## $ ink_temperature  : num  17 15 16 16 17 16.8 16.5 16.5 12 16 ...
## $ humidity         : num  78 80 80 75 80 76 75 75 70 75 ...
##  $ roughness        : num  0.75 0.75 NA 0.312 0.75 ...
##  $ blade_pressure   : num  20 20 30 30 30 28 30 28 60 32 ...
##  $ varnish_pct      : num  13.1 6.6 6.5 5.6 0 8.6 0 0 0 22.7 ...
##  $ press_speed      : num  1700 1900 1850 1467 2100 ...
##  $ ink_pct          : num  50.5 54.9 53.8 55.6 57.5 53.8 62.5 62.5 60.2 45.5 ...
##  $ solvent_pct      : num  36.4 38.5 39.8 38.8 42.5 37.6 37.5 37.5 39.8 31.8 ...
##  $ ESA_voltage      : num  0 0 0 0 5 5 6 6 1.5 0 ...
##  $ ESA_amperage     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ wax              : num  2.5 2.5 2.8 2.5 2.3 2.5 2.5 2.5 3 3 ...
##  $ hardener         : num  1 0.7 0.9 1.3 0.6 0.8 0.6 1.1 1 1 ...
##  $ roller_durometer : num  34 34 40 40 35 40 30 30 40 38 ...
##  $ current_density  : num  40 40 40 40 40 40 40 40 40 40 ...
##  $ anode_space_ratio: num  105 105 104 108 107 ...
##  $ chrome_content   : num  100 100 100 100 100 100 100 100 100 100 ...
##  $ band_type        : Factor w/ 2 levels "band","noband": 1 2 2 2 2 2 2 2 1 2 ...
```

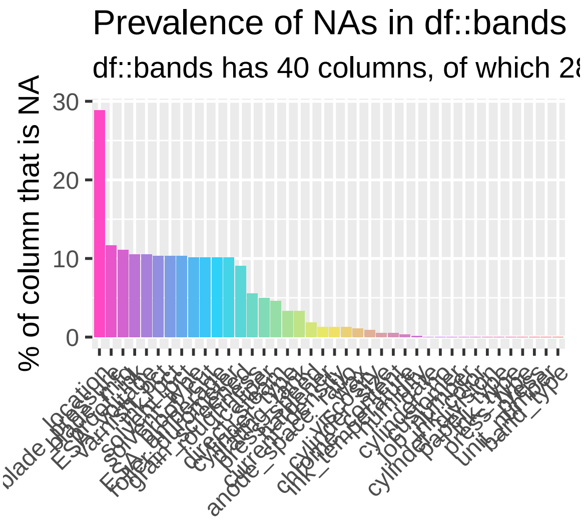
```
bands <- as.data.frame(lapply(bands,function(x)
  if(is.factor(x)) factor(toupper(x))
  else(x))) # uppercase for all the factor values

#there is a warning message!!!! I took it out for now, but to review it and understand it
```

```
knitr::kable(introduce(bands))
```

rows	columns	discrete_columns	continuous_columns	all_missing_columns	total_missing_values	complete_
540	40	16	24	0	999	

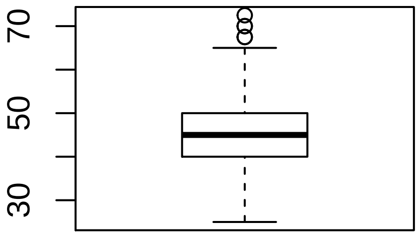
```
#describe(bands)
```

```
# managing missing values (19 rows with NAs)

na <- inspect_na(bands)
show_plot(na, col_palette=2)
```

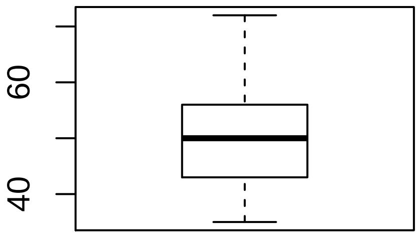


```r
for (i in 21:39) {
  print(summary(bands[i]))
  print (boxplot(bands[i])$out)
  bands[is.na(bands[,i]), i] <- mean(bands[,i], na.rm = TRUE)

}
```

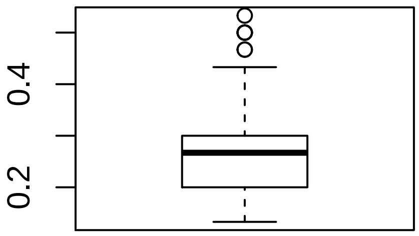
```
##    proof_cut
##  Min.   :25.00
##  1st Qu.:40.00
##  Median :45.00
##  Mean   :45.04
##  3rd Qu.:50.00
##  Max.   :72.50
##  NA's   :55
```



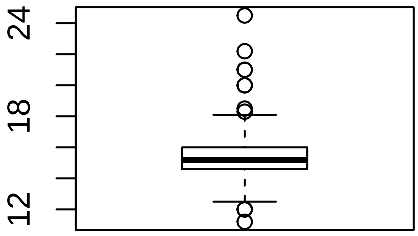
```
## [1] 67.5 72.5 70.0 70.0 67.5
##    viscosity
##  Min.   :35.00
##  1st Qu.:43.00
##  Median :50.00
##  Mean   :50.94
##  3rd Qu.:56.00
##  Max.   :72.00
##  NA's   :5
```



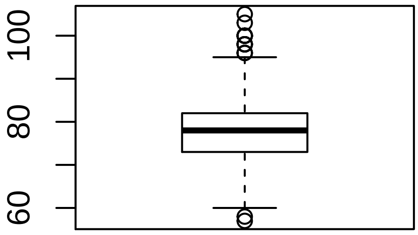
```
## numeric(0)
##     caliper
##  Min.   :0.1330
##  1st Qu.:0.2000
##  Median :0.2670
##  Mean   :0.2757
##  3rd Qu.:0.3000
##  Max.   :0.5330
##  NA's   :27
```



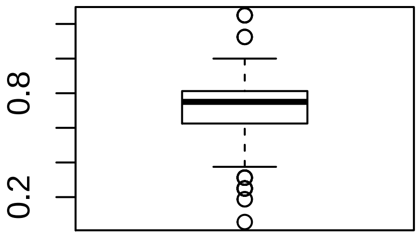
```
## [1] 0.500 0.467 0.467 0.500 0.500 0.500 0.500 0.533
##  ink_temperature
##  Min.   :11.20
##  1st Qu.:14.60
##  Median :15.20
##  Mean   :15.36
##  3rd Qu.:16.00
##  Max.   :24.50
##  NA's   :2
```



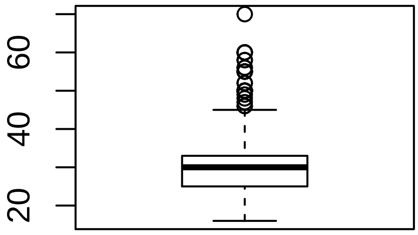
```
##  [1] 12.0 12.0 24.5 18.5 12.0 20.0 20.0 22.2 21.0 11.2 21.0 18.3
##     humidity
##  Min.   : 57.00
##  1st Qu.: 73.00
##  Median : 78.00
##  Mean   : 78.55
##  3rd Qu.: 82.00
##  Max.   :105.00
##  NA's   :1
```



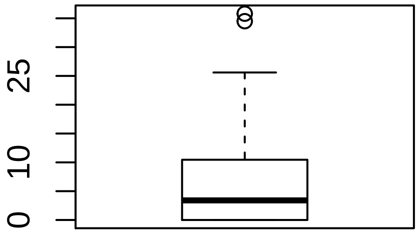
```
##  [1]  96  57  58 100 100  98  96  98  98 100 105  98  96 100 100 103 100
## [18] 100  98
##    roughness
##  Min.   :0.05625
##  1st Qu.:0.62500
##  Median :0.75000
##  Mean   :0.72451
##  3rd Qu.:0.81250
##  Max.   :1.25000
##  NA's   :30
```



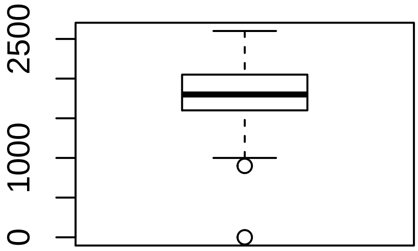
```
##  [1] 0.31250 1.25000 1.25000 1.25000 1.25000 1.12500 0.25000 0.31250
##  [9] 0.25000 0.25000 0.25000 0.31250 0.25000 0.25000 0.25000 0.25000
## [17] 0.25000 0.25000 1.25000 1.12500 0.31250 0.25000 0.31250 0.31250
## [25] 0.25000 0.05625 0.18750 0.25000 0.25000 0.31250 0.18750 0.25000
## [33] 0.25000 0.31250 1.12500
##  blade_pressure
##  Min.   :16.0
##  1st Qu.:25.0
##  Median :30.0
##  Mean   :30.9
##  3rd Qu.:33.0
##  Max.   :70.0
##  NA's   :63
```



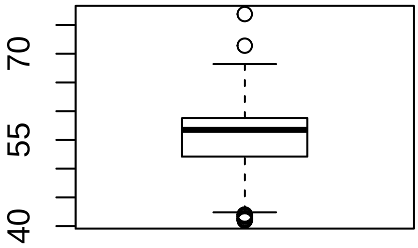
```
##  [1] 60 46 50 56 47 50 58 50 60 55 55 60 50 52 70 49 60 50 58 50 55 50 50
## [24] 55 55 50 50 52 55 50 46 48 50 50 50 56 50 50 50 50 50 55 50 50 50 50
## [47] 60 50
##   varnish_pct
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 3.400
##  Mean   : 5.781
##  3rd Qu.:10.425
##  Max.   :35.800
##  NA's   :56
```



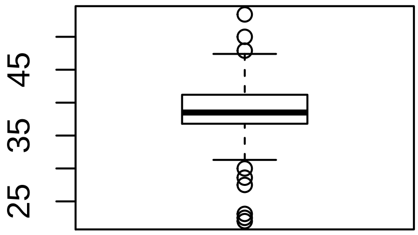
```
## [1] 35.8 34.5
##   press_speed
##  Min.   :   0
##  1st Qu.:1600
##  Median :1800
##  Mean   :1823
##  3rd Qu.:2042
##  Max.   :2600
##  NA's   :10
```



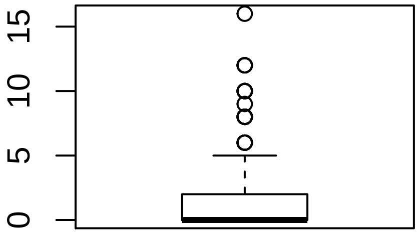
```
## [1] 900   0
##     ink_pct
##  Min.   :41.00
##  1st Qu.:52.10
##  Median :56.75
##  Mean   :55.64
##  3rd Qu.:58.80
##  Max.   :76.90
##  NA's   :56
```



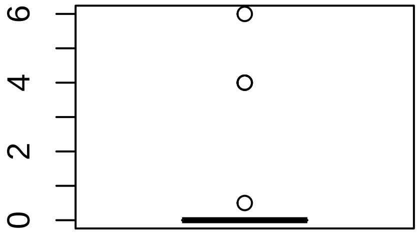
```
##  [1] 76.9 71.4 41.3 41.3 41.3 41.7 41.0 42.0 41.7 41.0 42.0
##   solvent_pct
##  Min.   :22.00
##  1st Qu.:36.80
##  Median :38.50
##  Mean   :38.57
##  3rd Qu.:41.20
##  Max.   :53.40
##  NA's   :56
```



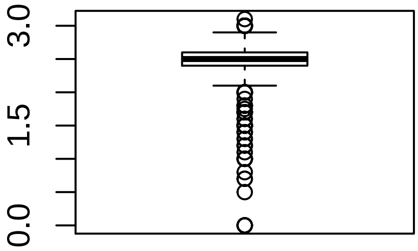
```
## [1] 50.0 23.1 28.6 47.9 22.5 22.0 30.0 27.5 53.4
##   ESA_voltage
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 1.319
##  3rd Qu.: 2.000
##  Max.   :16.000
##  NA's   :57
```



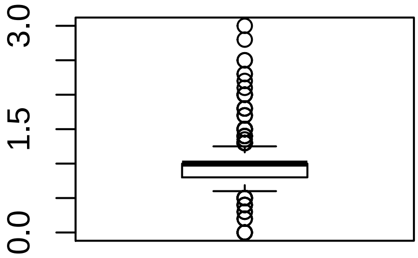
```
##  [1]  6  6 12 12 16  6  8  8  8  8 10 10 10  6  6 12 12 10 10  8  8  8  9
## [24]  8  8  6 10 10
##   ESA_amperage
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.03814
##  3rd Qu.:0.00000
##  Max.   :6.00000
##  NA's   :55
```



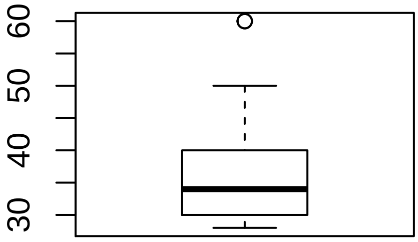
```
## [1] 0.5 4.0 4.0 6.0 4.0
##       wax
##  Min.   :0.000
##  1st Qu.:2.400
##  Median :2.500
##  Mean   :2.399
##  3rd Qu.:2.600
##  Max.   :3.100
##  NA's   :6
```



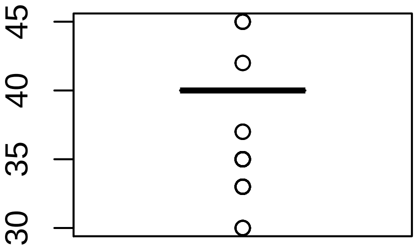
```
##   [1] 3.00 3.00 3.00 2.00 1.10 1.70 1.00 2.00 2.00 2.00 1.00 3.00 3.00 1.00
##  [15] 2.00 1.50 1.40 1.70 1.30 1.70 1.00 1.00 2.00 2.00 1.80 2.00 3.00 0.80
##  [29] 0.70 2.00 1.80 1.50 1.60 1.50 0.00 3.00 1.50 2.00 3.00 3.00 2.00 3.00
##  [43] 3.00 1.70 1.70 1.50 2.00 1.50 1.50 2.00 1.40 3.00 1.00 3.00 1.80 2.00
##  [57] 2.00 1.30 3.00 2.00 2.00 3.00 3.10 1.00 3.00 0.00 3.00 2.00 1.50 2.00
##  [71] 3.00 3.00 1.50 3.00 2.00 3.00 1.50 1.20 3.00 2.00 3.00 3.00 1.20 1.50
##  [85] 3.00 3.00 3.00 2.00 1.50 3.00 3.00 3.00 3.00 3.00 2.00 1.40 1.50 3.00
##  [99] 3.00 3.00 3.00 3.00 3.00 3.00 3.00 3.00 3.00 3.00 1.50 3.00 3.00 3.00
## [113] 1.70 2.00 3.00 3.00 3.00 3.00 1.70 2.00 3.00 3.00 3.00 2.00 2.00 1.70
## [127] 3.00 3.00 0.00 3.00 3.00 3.00 1.75 2.00 0.00 3.00 2.00 2.00 3.00 3.00
## [141] 3.00 3.00 3.00 3.00 1.50 3.00 1.90 3.00 3.00 3.00 3.00 3.00 3.00 3.00
## [155] 3.00 3.00 3.00 3.00 3.00 3.00 0.70 1.60 2.00 3.00 1.50 1.00 3.00 0.00
## [169] 3.00 0.00 3.00 3.00 3.00 2.00 1.50 1.50 0.00 2.00 1.80 2.00 1.50 3.00
## [183] 3.00 2.00 1.80 1.20 0.50 1.00 0.00 1.50
##     hardener
##  Min.   :0.0000
##  1st Qu.:0.8000
##  Median :1.0000
##  Mean   :0.9871
##  3rd Qu.:1.0000
##  Max.   :3.0000
##  NA's   :7
```



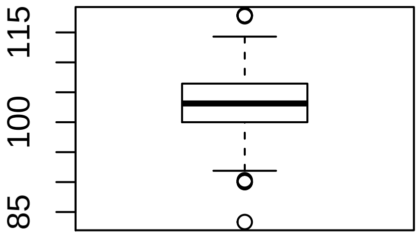
```
##   [1] 1.30 1.80 1.40 1.30 1.50 1.30 1.50 1.70 1.30 1.50 1.40 1.40 1.30 0.20
##  [15] 1.50 0.50 1.30 1.70 0.00 0.20 1.50 1.50 2.00 1.50 0.50 1.70 1.30 1.30
##  [29] 2.00 1.35 0.40 1.50 0.50 2.50 1.50 1.50 1.50 1.30 1.80 1.50 0.50 1.30
##  [43] 0.50 1.30 0.30 1.80 0.40 0.50 0.00 0.20 0.40 0.50 2.30 1.50 3.00 0.50
##  [57] 0.50 0.50 0.50 0.50 0.50 1.30 0.50 0.50 0.40 1.30 1.50 1.30 1.40 0.50
##  [71] 0.50 1.50 1.30 0.50 0.50 0.40 0.50 1.50 0.00 1.30 1.30 1.30 0.00 2.00
##  [85] 1.40 0.50 2.00 2.00 1.50 1.30 1.40 0.50 1.50 1.80 1.30 1.50 1.50 2.50
##  [99] 1.80 2.00 1.50 1.50 0.00 1.50 0.00 0.30 1.30 0.00 1.50 1.50 2.00 1.70
## [113] 1.50 1.50 1.50 1.50 2.20 2.10 0.50 2.00 1.50 1.30 1.30 0.50 0.00 2.80
## [127] 2.30
##  roller_durometer
##  Min.   :28.00
##  1st Qu.:30.00
##  Median :34.00
##  Mean   :34.78
##  3rd Qu.:40.00
##  Max.   :60.00
##  NA's   :55
```



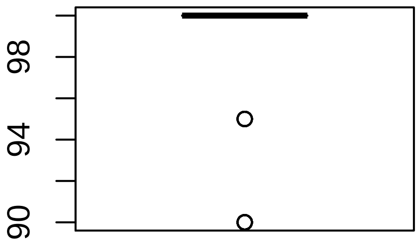
```
## [1] 60 60
##  current_density
##  Min.   :30.00
##  1st Qu.:40.00
##  Median :40.00
##  Mean   :39.06
##  3rd Qu.:40.00
##  Max.   :45.00
##  NA's   :7
```



```
##    [1] 33 33 33 33 35 35 33 33 33 35 33 33 33 33 33 33 33 33 33 30 33 33 33
##   [24] 33 33 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 37 37 35 42 37
##   [47] 42 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35 35
##   [70] 35 35 35 35 35 35 45 35 35 35 35 35 35 35 35 35 35 35 35 35 30 30 30
##   [93] 45 45 30 45 45 45 35 35 35 30
##  anode_space_ratio
##  Min.   : 83.33
##  1st Qu.:100.00
##  Median :103.13
##  Mean   :103.04
##  3rd Qu.:106.45
##  Max.   :117.86
##  NA's   :7
```

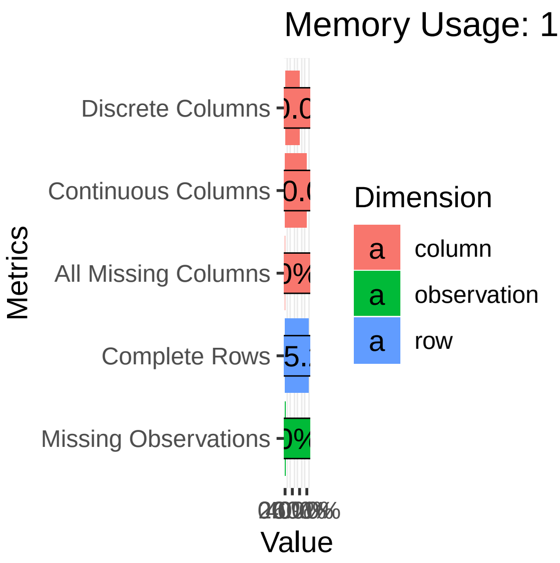


```
## [1]  90.00  90.00  90.30 117.85 117.85 117.85 117.86  83.33 117.70
##  chrome_content
##  Min.   : 90.00
##  1st Qu.:100.00
##  Median :100.00
##  Mean   : 99.59
##  3rd Qu.:100.00
##  Max.   :100.00
##  NA's   :3
```

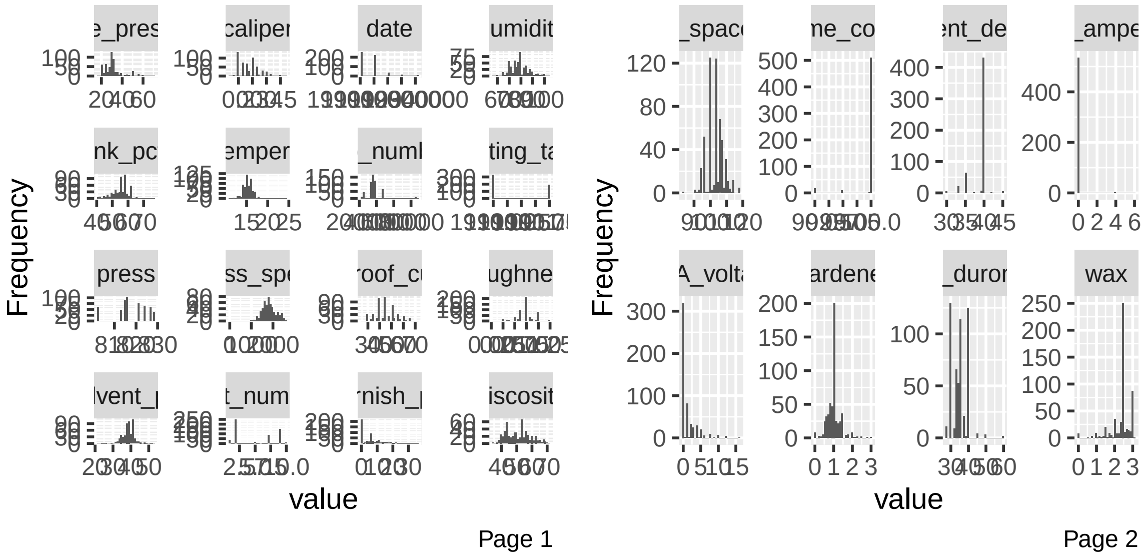


```
##  [1] 90 90 90 90 90 90 90 90 90 90 90 90 90 90 90 90 90 95 95 95 95 95 95
## [24] 95 95 95 95
```

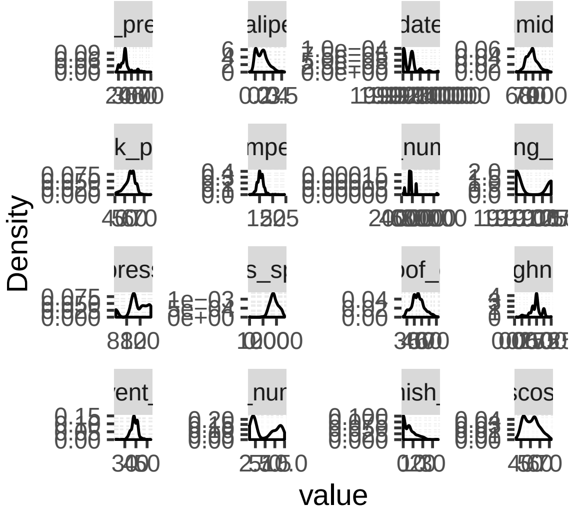
```
plot_intro(bands)
```



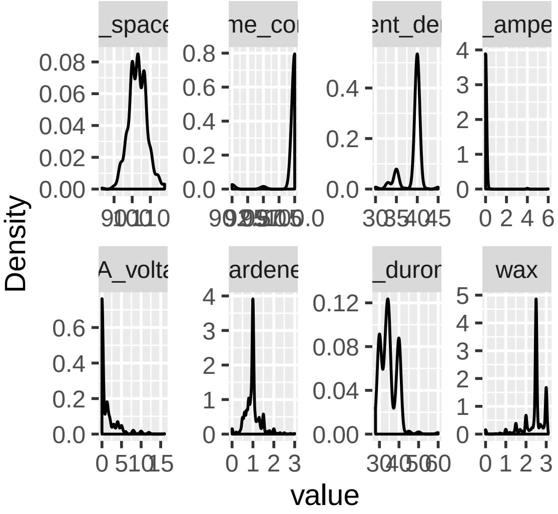
```r
plot_histogram(bands)
```



```r
plot_density(bands)
```



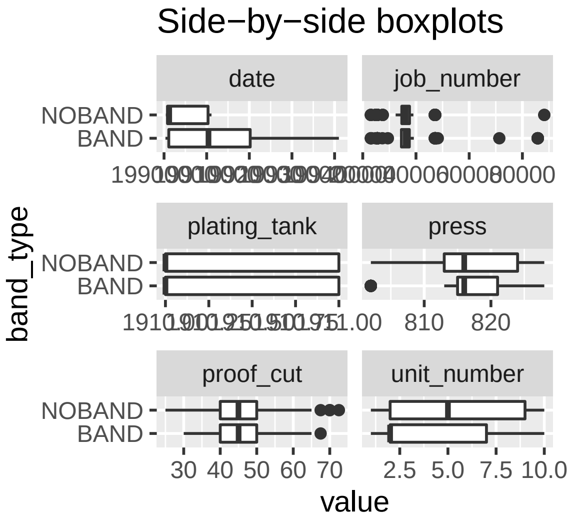
Page 1



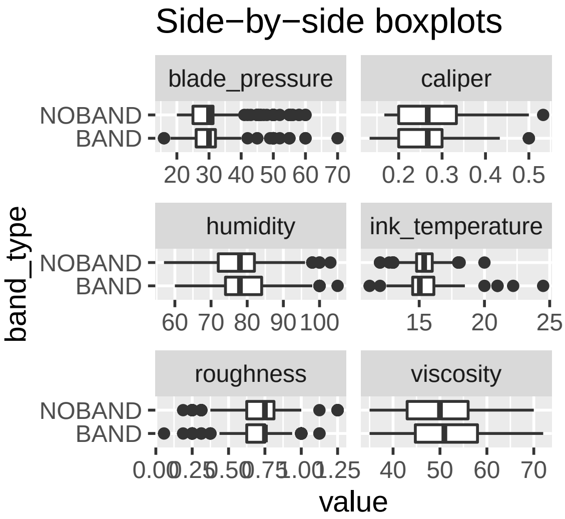
Page 2

```r
plot_boxplot(bands, by= 'band_type',  ncol = 2, title = "Side-by-side boxplots")
```

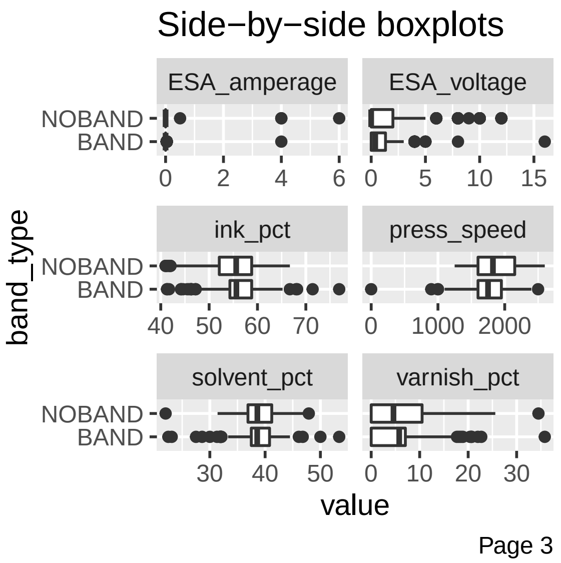
```
## Warning: Removed 18 rows containing non-finite values (stat_boxplot).
```

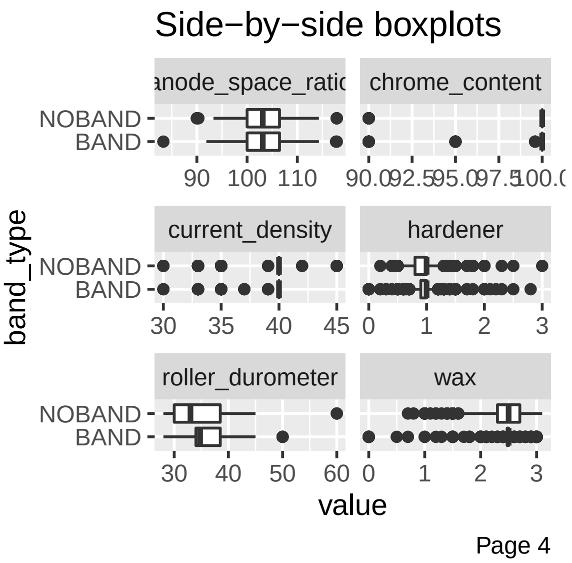


Page 1

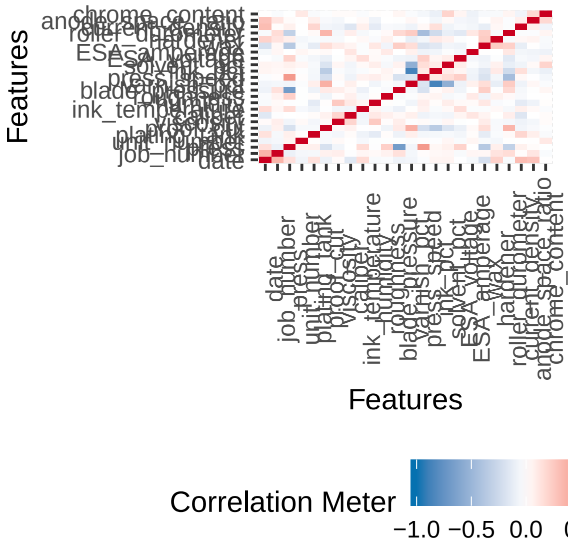


Page 2





```
plot_correlation(bands, type= 'c', cor_args = list( 'use' = 'complete.obs'))
```



```
#ggpairs(bands[,-40], ggplot2::aes(colour=band_type))

# split data in 2
bands.band <- filter(bands, bands$band_type == 'BAND')

bands.noband <- filter(bands, bands$band_type == 'NOBAND')

# summary(bands.band)
# summary(bands.noband)
```