R_application_02

October 20, 2022

1 R applications – Part II Descriptive statistics

- 2.1 Simple descriptive statistics
- 2.2 Using factors to deal with complex datasets
- 2.3 Using factors for classification

1.1 Simple descriptive statistics

Early in the interpretation of a newly acquired geochemical (or any other) dataset it is handy to examine descriptive statistics for selected variables (here elements or oxides). R contains a plethora of statistical tools, either built in, or provided via additional packages. At this stage, however, simple functions such as `mean`, `median`, `sd` (standard deviation) and `summary` (a statistical overview) would suffice.

Revealing are also simple graphical tools such as boxplots (box-and-whiskers plots; function `boxplot`) and histograms (`hist`). Scatter matrices (`pairs`) serve to spot potentially significant correlations.

Let's have a look, for the last time, onto the file `sazava.data` in detail. First, we compute means for all columns (variables) in the data set. Then we shall display boxplot for strontium, and find out all the main statistical parameters characterizing distribution of this element (the range, median, number of observations and not determined cases…). Lastly, we plot all the possible combinations of binary diagrams (a scatterplot matrix) for the following oxides: SiO2, MgO, CaO, Na2O, K2O, and P2O5.

[35]:

```r
sazava <- read.table("data/sazava.data",sep="\t")
sazava <- sazava[,-(1:6)]
# geochemical data only (all but the first six columns)
#head(sazava)

result <- apply(sazava,2,mean,na.rm=TRUE)
# na.rm is important, if missing values are present
print(round(result,2))
```

```
    SiO2     TiO2    Al2O3      FeO    Fe2O3      MnO      MgO      CaO
   57.95     0.64    16.94     4.73     1.75     0.14     3.57     8.16
    Na2O      K2O     P2O5      CO2        F        S H2O_PLUS H2O_MINUS
    2.80     1.66     0.15     0.16     0.08     0.09     1.11     0.06
      Ba       Rb       Sr       Zr       Nb       Ni       Co       Zn
  883.25    51.50   443.00    94.67     6.67    11.17    18.80    61.08
```

```
     Cr      La      Ce      Pr      Nd      Sm      Eu      Gd 
  67.33   20.05   42.69    4.40   16.01    3.30    1.40    3.08 
     Tb      Dy      Ho      Er      Tm      Yb      Lu       Y 
   0.44    2.48    0.46    1.33    0.20    1.38    0.22   21.79 
     Cs      Ta      Hf 
   4.12    0.54    3.50 
```

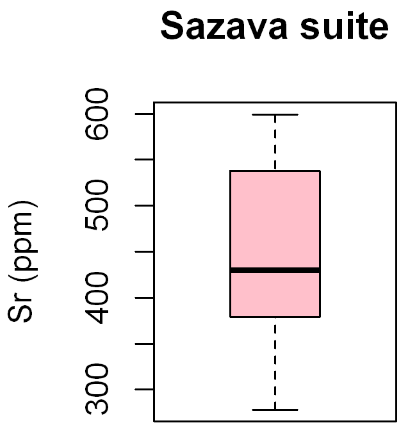
[36]:

```
options(repr.plot.width=2.5, repr.plot.height=3.5,repr.plot.res=300)
```

[37]:

```
boxplot(sazava[,"Sr"],main="Sazava suite", ylab="Sr (ppm)",col="pink")
summary(sazava[,"Sr"])
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  278.0   392.5   430.0   443.0   537.5   599.0       2 
```

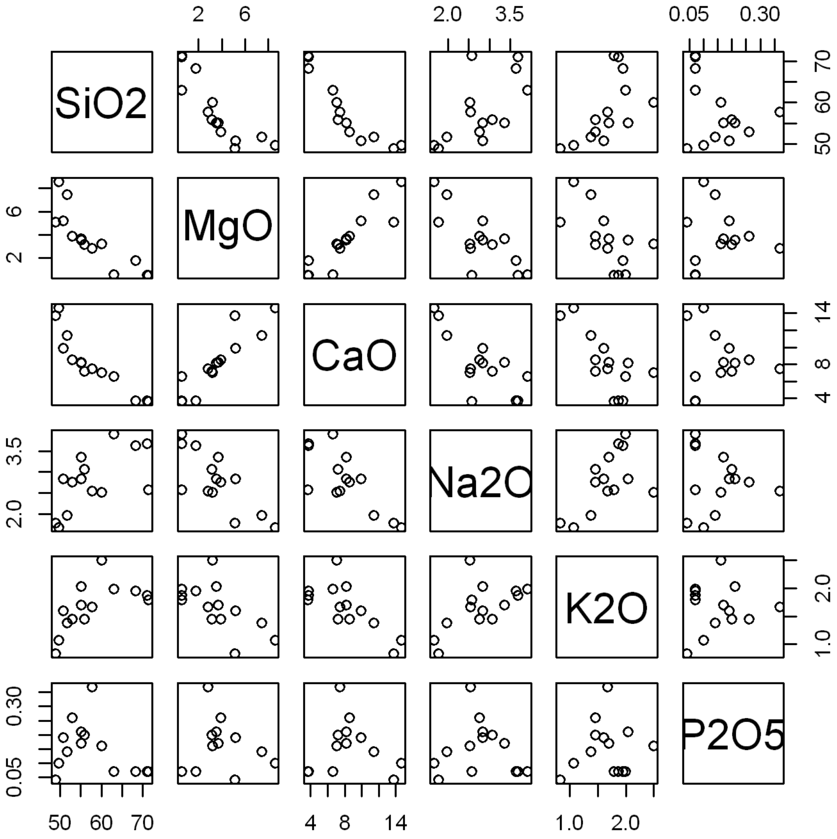


[38]:

```
options(repr.plot.width=5, repr.plot.height=5,repr.plot.res=300)
```

[39]:

```
oxides <- c("SiO2","MgO","CaO","Na2O","K2O","P2O5")
pairs(sazava[,oxides])
```



1.2 Using factors to deal with complex datasets

Statistical examination of complex geochemical data sets including, for instance, analyses for several intrusions, is tedious. Fortunately factors in R, in connection with the function `tapply`, offer a very flexible and elegant solution.

Using the factor `intrusion`, we will calculate the mean SiO2 and Ba contents for each of the pre-defined rock groups in the Sázava dataset.

[40]:

```r
sazava <- read.table("data/sazava.data",sep="\t")

# Defining the groups
intrusion <- factor(sazava[,"Intrusion"])
print(intrusion)
```

```
 [1] Sazava Sazava Sazava Sazava Sazava basic  basic  basic  basic  basic 
[11] Pozary Pozary Pozary Pozary
Levels: basic Pozary Sazava
```

[41]:

```
cat("Mean SiO2 contents in individual groups are (wt. %):\n")
ee <- tapply(sazava[,"SiO2"],intrusion,mean)
print(ee)
```

```
Mean SiO2 contents in individual groups are (wt. %):
 basic Pozary Sazava
51.778 68.440 55.738
```

[42]:

```
cat("Mean Ba contents in individual groups are (ppm):\n")
ee <- tapply(sazava[,"Ba"],intrusion,mean,na.rm=TRUE)
print(ee)
```

```
Mean Ba contents in individual groups are (ppm):
  basic  Pozary  Sazava
 676.25 1291.25  682.25
```

The R language provides additional, arguably even more powerful tools. For instance, `aggregate` applies a given function to each of the variables (columns) of a numeric matrix or data frame `x` respecting grouping (defined by a factor or list of factors). Analogous is the function `by`, which splits a data frame into several smaller ones based on a factor (or list of factors).

Utilizing the function `summary`, we shall calculate basic statistical parameters for SiO2 distribution in each of the rock groups of the Sázava suite (factor `intrusion`). What are the means for selected trace elements (Ba, Rb, Sr and Zr) in individual intrusions? Using the function `by`, we will display basic statistical summaries for major-element oxides in each of the rock groups.

[43]:

```
sazava <- read.table("data/sazava.data",sep="\t")
intrusion <- factor(sazava[,"Intrusion"])
sio2 <- tapply(sazava[,"SiO2"],intrusion,summary)
print(sio2)
```

```
$basic
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  48.84   49.63   51.72   51.78   52.90   55.80

$Pozary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  62.95   66.96   69.69   68.44   71.17   71.42

$Sazava
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  50.72   55.09   55.17   55.74   57.73   59.98
```

[44]:

```
trace <- c("Rb","Sr","Ba","Zr")
print(aggregate(sazava[,trace],list(Rock=intrusion),
                mean,na.rm=TRUE))
```

```
    Rock    Rb       Sr       Ba      Zr
```

```
1  basic 34.5 346.25  676.25  65.75
2 Pozary 59.5 460.75 1291.25 157.25
3 Sazava 60.5 522.00  682.25  61.00
```

[45]:
```
loc<-factor(sazava[,"Locality"])
print(loc)
```

```
 [1] Mrac       Mrac       Mrac       Mrac       Teletín    Teletín   
 [7] Pecerady   Pecerady   Vavretice  Brtnice    Krhanice   Prosecnice
[13] Prosecnice Prosecnice
Levels: Brtnice Krhanice Mrac Pecerady Prosecnice Teletín Vavretice
```

[46]:
```
print(aggregate(sazava[,trace],list(Locality=loc,Rock=intrusion),mean,na.
 ↪rm=TRUE))
```

```
    Locality   Rock       Rb       Sr       Ba        Zr
1    Brtnice  basic 43.00000 325.0000  860.000  72.00000
2   Pecerady  basic 21.00000 352.0000  583.000  76.00000
3    Teletín  basic 43.00000 430.0000 1017.000  88.00000
4  Vavretice  basic 31.00000 278.0000  245.000  27.00000
5   Krhanice Pozary 51.00000 599.0000 1024.000 128.00000
6 Prosecnice Pozary 62.33333 414.6667 1380.333 167.00000
7       Mrac Sazava 61.66667 517.0000  669.000  62.33333
8    Teletín Sazava 57.00000 537.0000  722.000  57.00000
```

[47]:
```
by(sazava[,7:17],list(Rock=intrusion),summary)
```

```
Rock: basic
      SiO2            TiO2            Al2O3            FeO             Fe2O3     
 Min.   :48.84   Min.   :0.340   Min.   :13.34   Min.   :2.740   Min.   :1.47  
 1st Qu.:49.63   1st Qu.:0.670   1st Qu.:14.17   1st Qu.:5.690   1st Qu.:2.44  
 Median :51.72   Median :0.760   Median :16.98   Median :6.220   Median :2.79  
 Mean   :51.78   Mean   :0.784   Mean   :16.87   Mean   :5.664   Mean   :2.64  
 3rd Qu.:52.90   3rd Qu.:0.800   3rd Qu.:18.23   3rd Qu.:6.430   3rd Qu.:3.22  
 Max.   :55.80   Max.   :1.350   Max.   :21.64   Max.   :7.240   Max.   :3.28  
      MnO             MgO              CaO             Na2O            K2O       
 Min.   :0.130   Min.   :3.160   Min.   : 7.22   Min.   :1.67   Min.   :0.830  
 1st Qu.:0.160   1st Qu.:3.890   1st Qu.: 8.55   1st Qu.:1.78   1st Qu.:1.070  
 Median :0.160   Median :5.110   Median :11.44   Median :1.97   Median :1.380  
 Mean   :0.174   Mean   :5.644   Mean   :11.12   Mean   :2.25   Mean   :1.236  
 3rd Qu.:0.170   3rd Qu.:7.470   3rd Qu.:13.75   3rd Qu.:2.76   3rd Qu.:1.450  
 Max.   :0.250   Max.   :8.590   Max.   :14.64   Max.   :3.07   Max.   :1.450  
      P2O5      
 Min.   :0.040  
 1st Qu.:0.100  
 Median :0.140  
 Mean   :0.148  
 3rd Qu.:0.200  
 Max.   :0.260
------------------------------------------------------------
Rock: Pozary
      SiO2            TiO2          Al2O3            FeO            Fe2O3
 Min.   :62.95   Min.   :0.28   Min.   :15.04   Min.   :1.650   Min.   :0.380
 1st Qu.:66.96   1st Qu.:0.28   1st Qu.:15.08   1st Qu.:2.002   1st Qu.:0.395
 Median :69.69   Median :0.29   Median :15.19   Median :2.120   Median :0.435
 Mean   :68.44   Mean   :0.29   Mean   :16.36   Mean   :2.075   Mean   :0.480
 3rd Qu.:71.17   3rd Qu.:0.30   3rd Qu.:16.47   3rd Qu.:2.192   3rd Qu.:0.520
 Max.   :71.42   Max.   :0.30   Max.   :20.02   Max.   :2.410   Max.   :0.670
      MnO              MgO             CaO             Na2O
 Min.   :0.0400   Min.   :0.520   Min.   :3.670   Min.   :2.580
 1st Qu.:0.0475   1st Qu.:0.520   1st Qu.:3.730   1st Qu.:3.368
 Median :0.0500   Median :0.535   Median :3.755   Median :3.655
 Mean   :0.0500   Mean   :0.840   Mean   :4.447   Mean   :3.450
 3rd Qu.:0.0525   3rd Qu.:0.855   3rd Qu.:4.473   3rd Qu.:3.737
 Max.   :0.0600   Max.   :1.770   Max.   :6.610   Max.   :3.910
      K2O            P2O5
 Min.   :1.79   Min.   :0.07
 1st Qu.:1.85   1st Qu.:0.07
 Median :1.91   Median :0.07
 Mean   :1.90   Mean   :0.07
 3rd Qu.:1.96   3rd Qu.:0.07
 Max.   :1.99   Max.   :0.07
------------------------------------------------------------
Rock: Sazava
      SiO2            TiO2           Al2O3            FeO
 Min.   :50.72   Min.   :0.630   Min.   :16.42   Min.   :5.260
 1st Qu.:55.09   1st Qu.:0.710   1st Qu.:17.00   1st Qu.:5.430
 Median :55.17   Median :0.750   Median :17.57   Median :5.460
 Mean   :55.74   Mean   :0.774   Mean   :17.48   Mean   :5.922
 3rd Qu.:57.73   3rd Qu.:0.830   3rd Qu.:17.59   3rd Qu.:5.810
 Max.   :59.98   Max.   :0.950   Max.   :18.82   Max.   :7.650
     Fe2O3            MnO             MgO            CaO            Na2O
 Min.   :1.000   Min.   :0.120   Min.   :2.82   Min.   :7.04   Min.   :2.520
 1st Qu.:1.350   1st Qu.:0.150   1st Qu.:3.21   1st Qu.:7.47   1st Qu.:2.540
 Median :2.130   Median :0.160   Median :3.52   Median :8.20   Median :2.830
 Mean   :1.866   Mean   :0.172   Mean   :3.68   Mean   :8.17   Mean   :2.816
 3rd Qu.:2.190   3rd Qu.:0.190   3rd Qu.:3.67   3rd Qu.:8.22   3rd Qu.:2.830
 Max.   :2.660   Max.   :0.240   Max.   :5.18   Max.   :9.92   Max.   :3.360
      K2O             P2O5
 Min.   :1.600   Min.   :0.16
 1st Qu.:1.670   1st Qu.:0.17
 Median :1.700   Median :0.19
 Mean   :1.902   Mean   :0.22
 3rd Qu.:2.040   3rd Qu.:0.21
 Max.   :2.500   Max.   :0.37
```

1.3 Using factors for classification

The function `cut` splits a numeric vector `x` into given number of intervals and codes its individual items according to the rank they fall into. So this function can be used for simple classification purposes.

We will classify samples in the Sázava set according to SiO2 contents (wt. %) in four groups, U ($<$ 45), B (45–52), I (52–63) and A ($>$ 63), i.e. in the geochemical jargon ultrabasic, basic, intermediate and acid rocks.

[48]:

```
sazava <- read.table("data/sazava.data",sep="\t")
silica <- cut(sazava[,"SiO2"],breaks=c(0,45,52,63,100),
              labels=c("U","B","I","A"))
acidity <- as.vector(silica)
names(acidity) <- rownames(sazava)
print(acidity)
```

```
  Sa-1    Sa-2    Sa-3    Sa-4    Sa-7   SaD-1   Gbs-1 Gbs-20  Gbs-2  Gbs-3    Po-1
   "I"     "I"     "I"     "B"     "I"     "I"     "B"    "B"    "B"    "I"     "I"
  Po-3    Po-4    Po-5
   "A"     "A"     "A"
```