Data Analysis in the Social Sciences

Visualisation of relationships between variables

Alla Tambovtseva

Visualising the association between quantitative variables

Now we will work with the data set containing indices from Worldwide Governance Indicators and Freedom House Index (2016).

Variables:

- `va`: Voice & Accountability
- `ps`: Political Stability & Absence of Violence
- `ge`: Government Effectiveness
- `rq`: Regulatory Quality
- `rl`: Rule of Law
- `cc` : Control of Corruption
- `fh`: Freedom House Index (Freedom in the World)

For all indices higher values correspond to a worse situation: less accountability, less stability, less freedom, etc. As for the Freedom House Index, values from 1 to 2.5 correspond to free countries, values from 3 to 5.5 correspond to partly free countries, values greater than 5.5 – to not free countries.

Let's load these data specifying a non-default decimal separator (a comma), delete rows with missing values and make sure that types of variables are correct:

```
# dec - comma as a decimal separator
wgi <- read.csv("https://raw.githubusercontent.com/allatambov/cluster-analysis/master/clust1/wgi_fh.csv
wgi <- na.omit(wgi)
str(wgi)
```

```
## 'data.frame':    195 obs. of  11 variables:
##  $ X       : int  2 3 4 6 9 10 12 13 14 15 ...
##  $ country : Factor w/ 202 levels "Afghanistan",..: 4 1 5 2 7 8 6 10 11 12 ...
##  $ cnt_code: Factor w/ 202 levels "ABW","ADO","AFG",..: 2 3 4 5 7 8 9 10 11 12 ...
##  $ year    : int  2016 2016 2016 2016 2016 2016 2016 2016 2016 2016 ...
##  $ va      : num  1.2 -1.09 -1.17 0.16 0.54 -0.62 0.65 1.3 1.29 -1.6 ...
##  $ ps      : num  1.4 -2.75 -0.39 0.26 0.22 -0.6 1.01 0.96 0.82 -0.87 ...
##  $ ge      : num  1.86 -1.22 -1.04 0 0.18 -0.15 0.27 1.58 1.51 -0.16 ...
##  $ rq      : num  0.87 -1.33 -1 0.19 -0.47 0.25 0.34 1.9 1.44 -0.28 ...
##  $ rl      : num  1.56 -1.62 -1.08 -0.35 -0.35 -0.11 0.51 1.75 1.78 -0.57 ...
##  $ cc      : num  1.23 -1.56 -1.41 -0.4 -0.31 -0.57 0.69 1.77 1.54 -0.87 ...
##  $ fh      : num  1 6 6 3 2 4.5 2 1 1 6.5 ...
##  - attr(*, "na.action")= 'omit' Named int  1 6 44 73 75 113 196
##   ..- attr(*, "names")= chr  "1" "6" "44" "73" ...
```

Further we will try to evaluate the association between different variables for each type of country: free, partly free and not free. So as to make it possible, we have to create a new variable with types of countries based on the values of `fh`. We will need a built-in function `cut()`. First, let's see how it works on a simple example.

Suppose we have a numeric vector `v` and we want to create a new factor vector `w` that will contain labels "low", "moderate" and "high" depending on values in `v`. So, if a value is no greater than 2, it should be labeled as "low", if greater than 2 and no greater than 4 it is "moderate", and if greater than 4, it should be marked "high".

```
v <- c(2, 3, 3.5, 4.5, 1.5)
# (0, 2] - low
# (2, 4] - moderate
# (4, 6] - high
```

Thus, we have four breakpoints (numbers that divide the range of values of `v` into three groups): 0, 2, 4 and 6.

```
# indicate vector of interest first
# indicate breakpoints
# indicate labels for intervals
w <- cut(v, breaks = c(0, 2, 4, 6),
         labels = c("low", "moderate", "high"))
w
```

```
## [1] low      moderate moderate high     low
## Levels: low moderate high
```

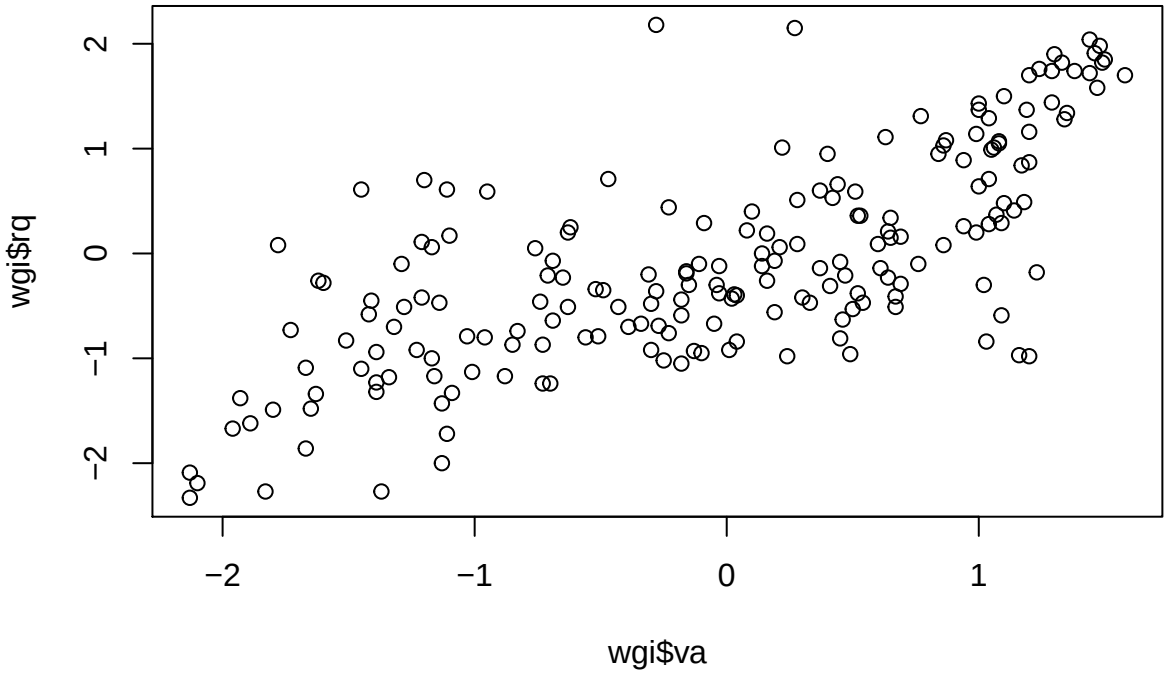
Now we can perform the same for our case (see info on breakpoints for Freedom House Index above):

```
wgi$fh_status <- cut(wgi$fh, breaks = c(0, 2.5, 5.5, 7),
    labels = c("not free", "partly free", "free"))
head(wgi$fh_status)
```

```
## [1] not free    free        free        partly free not free    partly free
## Levels: not free partly free free
```

Let's create a simple scatter plot for the following pair of variables: Voice & Accountability and Regulatory Quality:

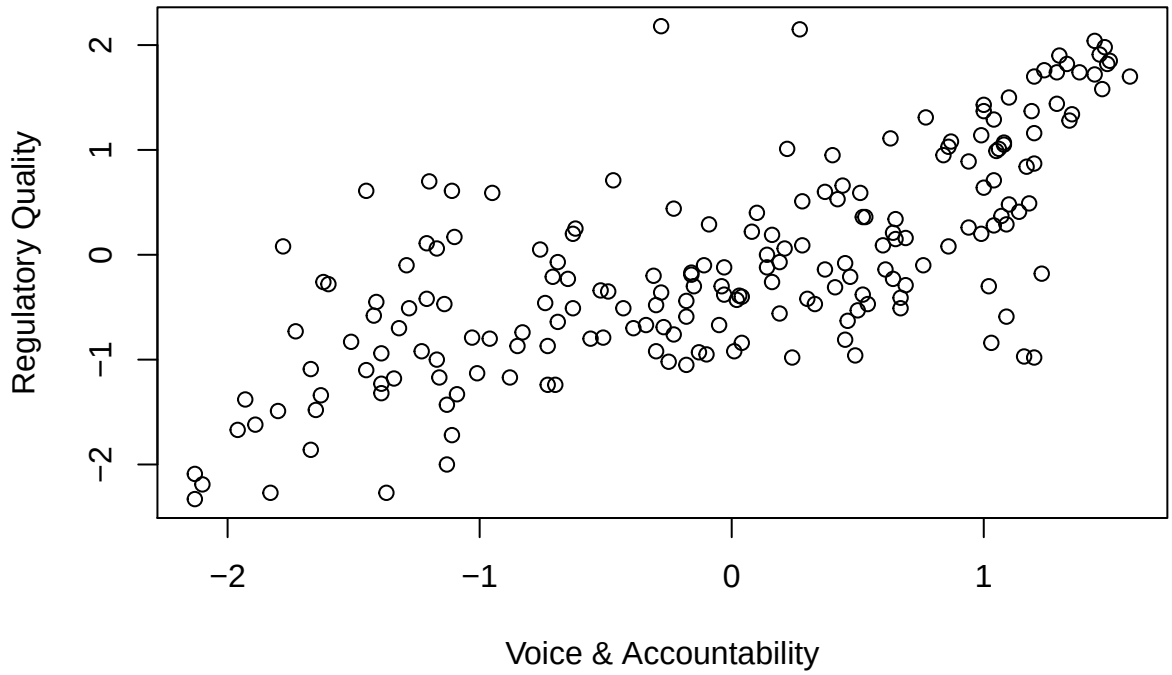
```
plot(wgi$va, wgi$rq)
```



We can add labels to axes so as to make our graph more informative:

```
plot(wgi$va, wgi$rq,
     xlab="Voice & Accountability",
```

```
     ylab="Regulatory Quality")
```



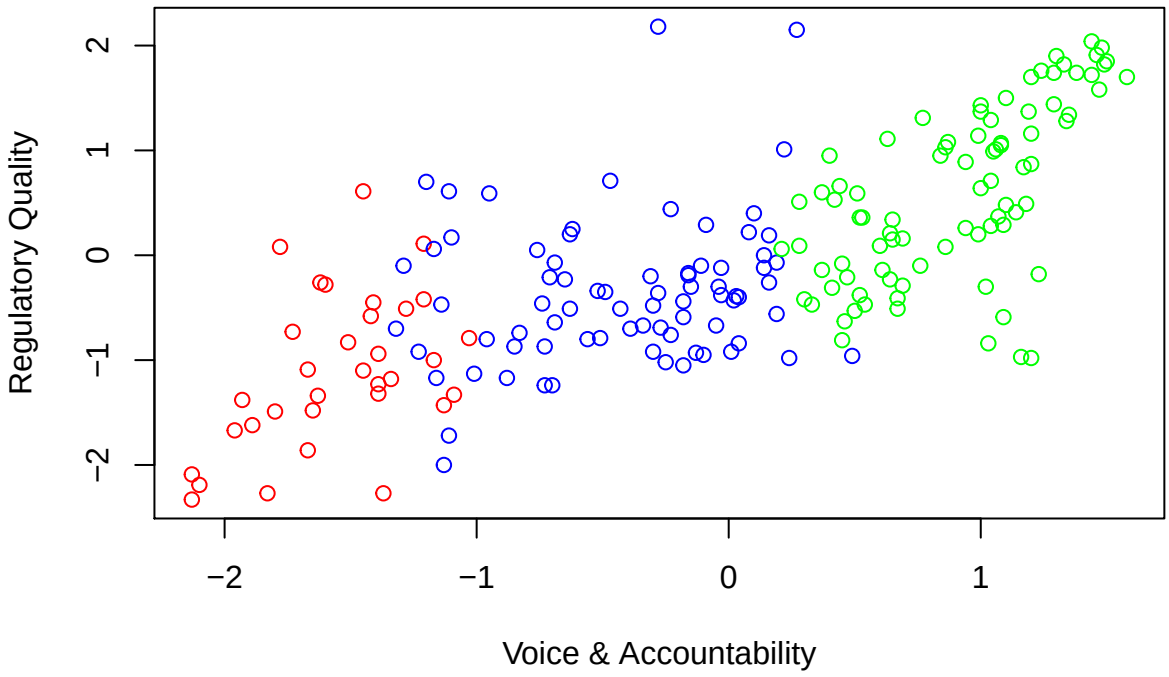
Now let's make our plot more interesting - color points by country type (free, partly free, not free). To do so we should choose three colors and then tell R to apply these colors to the variable `fh_status` (each factor level corresponds to one color).

```
cls <- c("green", "blue", "red") # choose colors
colors <- cls[wgi$fh_status] # apply to fh_status
head(colors) # look
```

```
## [1] "green" "red"   "red"   "blue"  "green" "blue"
```

Then we pass this vector of colors as a value of the argument `col`:

```
plot(wgi$va, wgi$rq,
     xlab="Voice & Accountability",
     ylab="Regulatory Quality",
     col=colors)
```

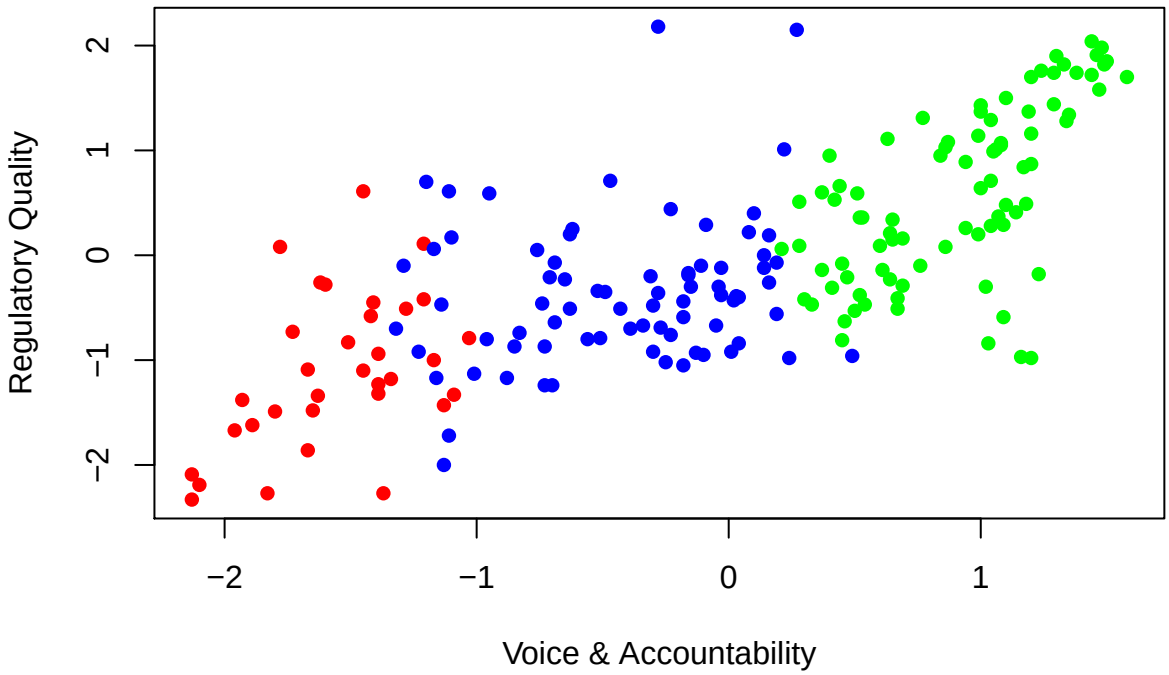


As we can see, points in our graph are not filled. This can impede understanding of this scatter plot. Let's choose another marker for points. We can look at the whole range of markers asking for help on the argument `pch`:

```
?pch
```

The 16th marker looks reasonable in our case, it is a filled point:

```
plot(wgi$va, wgi$rq,
     xlab="Voice & Accountability",
     ylab="Regulatory Quality",
     col=colors,
     pch = 16)
```

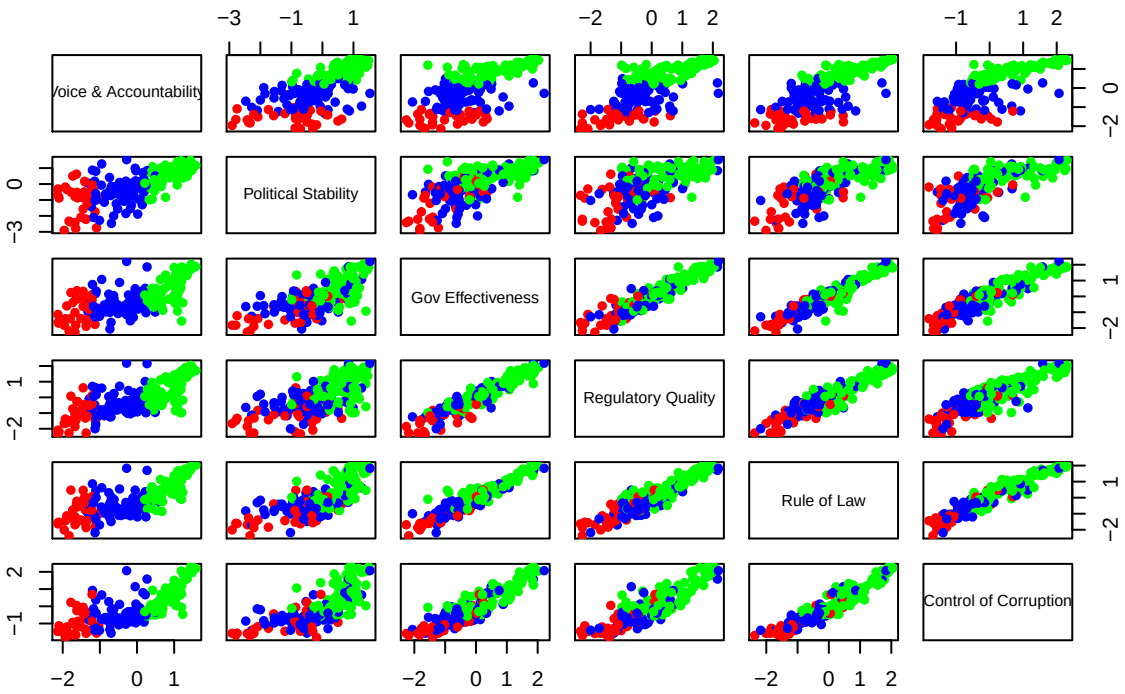


How can we interpret this? Why it might be sensible to color points? We can try to compare slopes of imaginary lines that show the trend of association. If we imagine these lines, the steepest line will be for free countries and the flattest - for partly free countries. So, judging by this graph we can say that the association between *Voice and Accountability* and *Regulatory Quality* is stronger for free countries and weaker for partly free countries.

If we want to draw several scatterplots at the same time, we can create a *scatterplot matrix*, a table of scatter plots for every pair of variables we are interested in. To do it we should use a built-in function `pairs()` and specify columns of interest inside:

```
# columns from 5 to 10
# col = colors, pch - point marker as above
# labels - texts in the rectangles
# cex.labels - labels size

pairs(wgi[5:10], col = colors, pch=16,
      labels = c("Voice & Accountability", "Political Stability",
                 "Gov Effectiveness", "Regulatory Quality",
                 "Rule of Law", "Control of Corruption"),
      cex.labels = 0.7,
      main = "World Governance Indicators")
```



Visualising the association between qualitative (nominal) variables

Now let's proceed to visualising relationships between nominal variables. We can load the data set on Titanic passengers we discussed before and look at the association between some variables.

```
tit <- read.csv("http://math-info.hse.ru/f/2018-19/pep/r/Titanic.csv")
tit <- na.omit(tit)  # delete rows with NA's
```

Let's see whether the class of a passenger (`PClass`) depends on the gender (`Sex`). In other words, is it true that in particular classes female or male passengers prevail. First, we should make the variable `PClass` factor. In fact, it is ordinal as we can sort values in an ascending order, but here we will consider it as a nominal one.

```
tit$Pclass <- factor(tit$Pclass)
```

We can create a contingency table:

```
tab <- table(tit$Pclass, tit$Sex)
tab
```

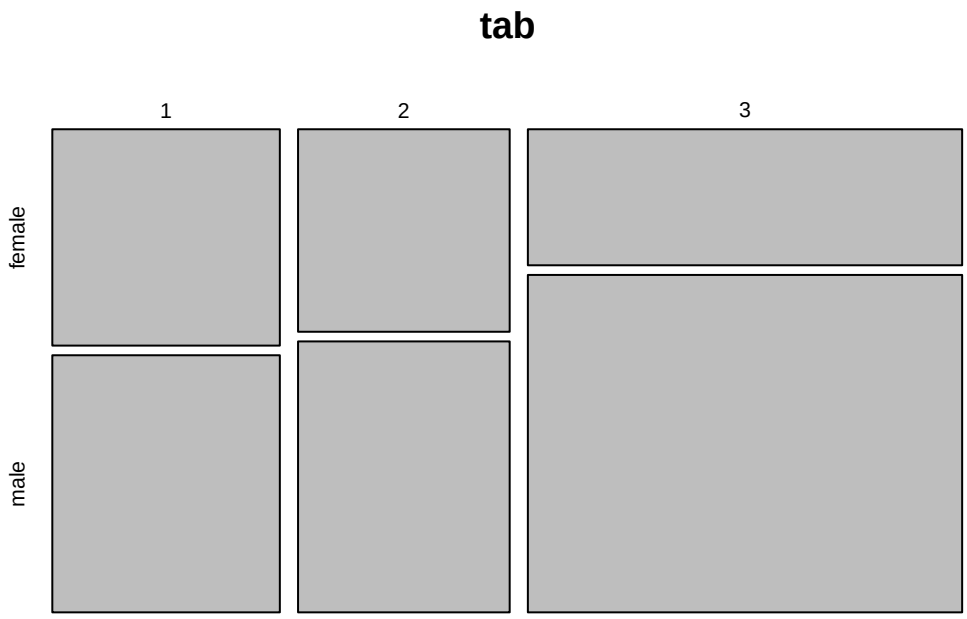
```
##
##     female male
##   1     85  101
##   2     74   99
##   3    102  253
```

Now let's pass this table to the function used for creating a *mosaic plot*, a graph used for visualising the frequencies. We will need the library `vcd` (*vcd* from *visualising categorical data*):

```
install.packages("vcd")
```

Create a mosaic plot:

```
library(vcd)
mosaicplot(tab)
```



Here we can see that in the third class male passengers prevail while in the first and the second class the ratio of males and females is approximately the same.