НИУ ВШЭ, ОП «Политология», 2022
Курс «Введение в многомерный статистический анализ»

Домашнее задание 2
Рубанов Владислав БПТ 201

Теоретическая часть

Задача 1. Быть или не быть?

Какие из перечисленных ниже матриц могут быть матрицами расстояний? Обоснуйте свой ответ.

$$\text{A} = \begin{pmatrix} 0 & 2 & 3 \\ 3 & 0 & 5 \\ 3 & 5 & 0 \end{pmatrix} \text{B} = \begin{pmatrix} 0 & 2 & 4 \\ 2 & 1 & 5 \\ 4 & 5 & 0 \end{pmatrix} \text{C} = \begin{pmatrix} 0 & 1.5 & 6 \\ 1.5 & 0 & 5 \\ 6 & 5 & 0 \end{pmatrix} \text{D} = \begin{pmatrix} 0 & -1.5 & 6 \\ -1.5 & 0 & 5 \\ 6 & 5 & 0 \end{pmatrix}$$

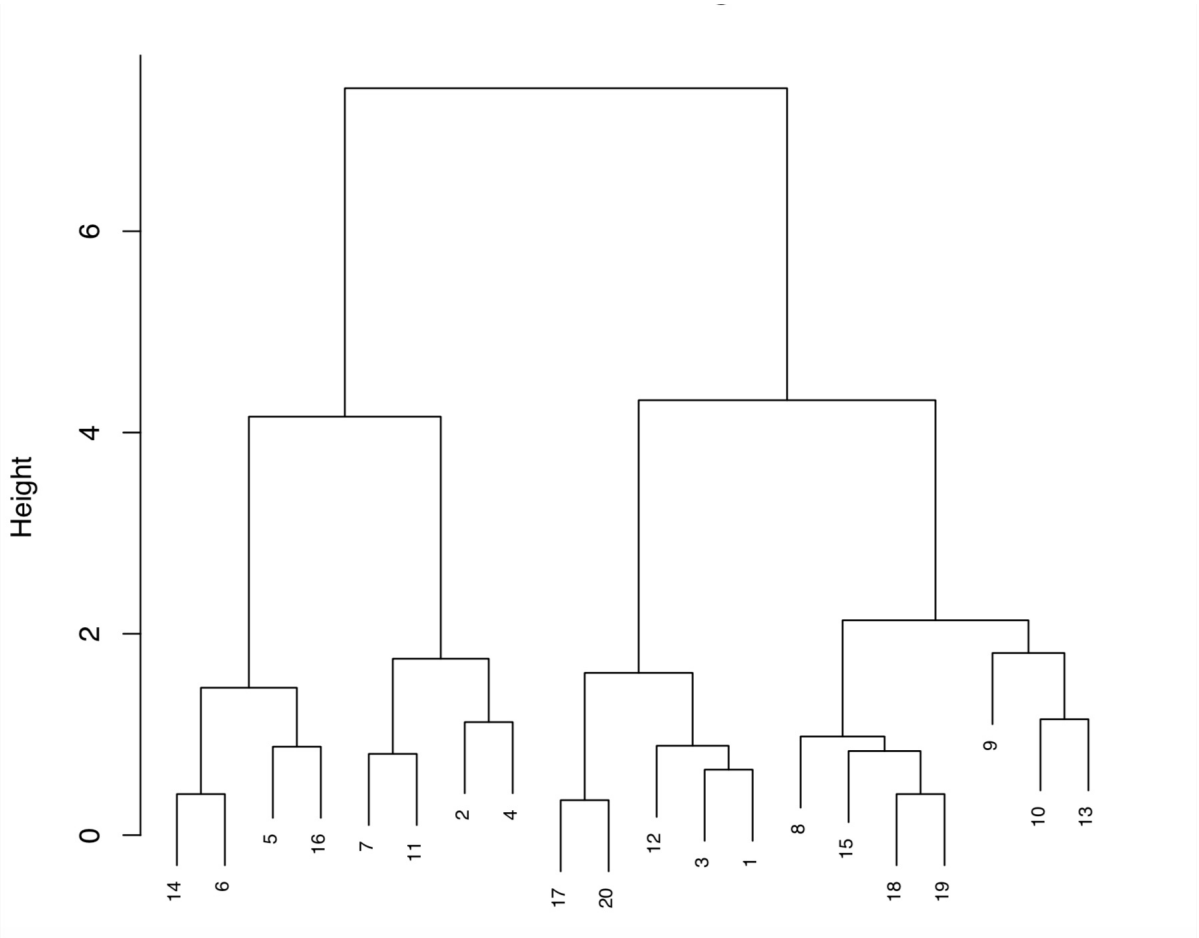
Из предложенных ниже матриц матрицей расстояния может быть только матрица $\text{C} = \begin{pmatrix} 0 & 1.5 & 6 \\ 1.5 & 0 & 5 \\ 6 & 5 & 0 \end{pmatrix}$, потому что остальные не соответствуют основным свойствам матриц расстояния и метрик, а именно:

- Матрица A: несимметричная (элемент $a_{1,2} \neq a_{2,1}$)
- Матрица B: по главной диагонали не везде находятся нули (ведь это расстояние от одной точки до самой себя: $b_{2,2} = 1$)
- Матрица D: метрика и, следовательно, расстояние не может быть отрицательным ($d_{1,2} = d_{2,1} = -1.5$)
- Матрица C соответствует всем свойствам метрик и матриц расстояния

К слову, правило треугольника выполняется во всех случаях, кроме отрицательного расстояния.

Задача 2. Выделяй и властвуй

Дана следующая дендрограмма:



Какое максимальное число кластеров наблюдений можно выделить на основании представленной ниже дендрограммы, если

a. *в каждом кластере должно быть не менее* 5 *наблюдений?*

   Максимум можно выделить 3 кластера для выполнения данного условия. Для этого нужно разрезать дендрограмму приблизительно на уровне 4.25: тогда получится три кластера, в которые войдет 8, 5 и 7 наблюдений.

b. *наблюдения номер* 1 *и* 17 *должны быть в одном кластере?*

   Максимум можно выделить 7 кластеров для выполнения данного условия. Для этого нужно разрезать дендрограмму приблизительно на уровне 1.7.

c. *наблюдения номер* 13 *и* 20 *должны быть в одном кластере?*

   Максимум можно выделить 2 кластера для выполнения данного условия, поскольку кластеры, в которых находятся наблюдения 13 и 20, объединяются последними перед общим объединением в один кластер. Для этого нужно разрезать дендрограмму приблизительно на уровне 4.3 и выше.

Задача 3. Чебышёв и много вычислений

Дан небольшой двумерный массив с данными по четырём наблюдениям:

a. *Сколько различных ненулевых расстояний необходимо посчитать для построения матрицы расстояний между наблюдениями?*

   Формально, число пар считается, как: $\frac{n \cdot (n-1)}{2}$. То есть в нашем случае необходимо посчитать $\frac{4 \cdot (4-1)}{2} = 6$ ненулевых расстояний. Именно столько расстояний будет записано в одном из „уголков“ матрицы расстояний.

id	X	Y
1	0	6
2	7	2
3	6	4
4	4	1

b. *Запишите матрицу расстояний для предложенного массива, используя в качестве метрики расстояние Чебышёва.*

Итак, нам необходимо рассчитать 6 расстояний, основываясь на метрике расстояния Чебышева.

$$
\begin{aligned}
d(1,2) &= max(|0-7|, |6-2|) = 7 \\
d(1,3) &= max(|0-6|, |6-4|) = 6 \\
d(1,4) &= max(|0-4|, |6-1|) = 5 \\
d(2,3) &= max(|7-6|, |2-4|) = 2 \\
d(2,4) &= max(|7-4|, |2-1|) = 3 \\
d(3,4) &= max(|6-4|, |4-1|) = 3
\end{aligned}
$$

То есть наша матрица расстояний будет выглядеть следующим образом:

$$
\mathrm{D} = \begin{pmatrix} 0 & 7 & 6 & 5 \\ 7 & 0 & 2 & 3 \\ 6 & 2 & 0 & 3 \\ 5 & 3 & 3 & 0 \end{pmatrix}
$$

Или

$$
D = \begin{bmatrix} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \mathbf{1} & 0 & 7 & 6 & 5 \\ \mathbf{2} & 7 & 0 & 2 & 3 \\ \mathbf{3} & 6 & 2 & 0 & 3 \\ \mathbf{4} & 5 & 3 & 3 & 0 \end{bmatrix}
$$

Проверим себя в R:

```
x <- c(0, 7, 6, 4)
y <- c(6, 2, 4, 1)

dat <- cbind.data.frame(x, y)

D <- dist(dat, method = 'maximum')
D
```

```
##   1 2 3
## 2 7
## 3 6 2
## 4 5 3 3
```

Мы можем увидеть, что матрица расстояний была найдено верно! В R она представлена в укороченном виде.

c. *Используя полученную на предыдущем шаге матрицу расстояний и метод дальнего соседа, реализуйте иерархический кластерный анализ и постройте дендрограмму.*

Шаг 1. На первом шаге иерархического кластерного анализа мы имеем 4 кластера, состоящих из каждого из 4-х наблюдений: *1, 2, 3, 4*. Перейдем к следующему шагу.

Шаг 2. Далее, вне зависимости от выбранного метода агломерации, на этом шаге необходимо соединить два наблюдения, которые **ближе всего друг к другу**. У нас это точки (кластеры) *2* и *3*, которые мы объединим на уровне 2 (см. матрицу расстояний). Получаем 3 кластера: *1, 2+3, 4*. Расположим их на графике сразу удобным образом:

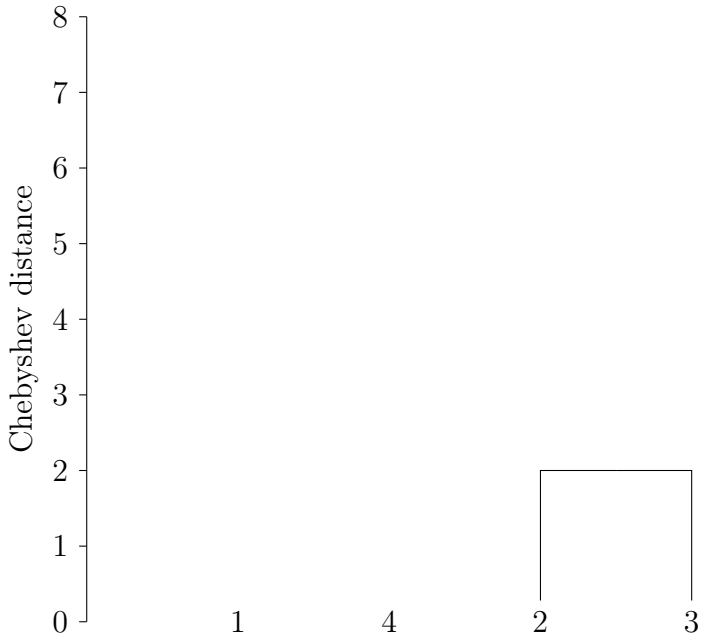


Рис. 1: Дендрограмма: шаг 2

Шаг 3. Для него нам необходимо построить обновленную матрицу расстояний, учитывающую новый кластер *2+3*. Для этого нужно найти расстояния между ним и остальными точками:

$$d(2+3,\ 1) = \begin{cases} \mathbf{d(2,1) = 7} \text{ выбираем этот вариант согласно методу дальнего соседа} \\ d(3,1) = 6 \end{cases}$$

$$d(2+3,\ 4) = \begin{cases} \mathbf{d(2,4) = 3} \text{ выбираем любой вариант согласно методу дальнего соседа} \\ \mathbf{d(3,4) = 3} \end{cases}$$

Таким образом, новая матрица расстояний:

$$D = \begin{bmatrix} & \mathbf{1} & \mathbf{2{+}3} & \mathbf{4} \\ \mathbf{1} & 0 & 7 & 5 \\ \mathbf{2{+}3} & 7 & 0 & 3 \\ \mathbf{3} & 5 & 3 & 0 \end{bmatrix}$$

То есть мы объединяем кластеры *2+3* и *4* на уровне 3. Теперь мы имеем 2 кластера: *1, 2+3+4*.

Шаг 4. Теперь нам осталось только объединить оставшиеся 2 кластера: *1* и *2+3+4* в один большой кластер — вопрос только на каком уровне.

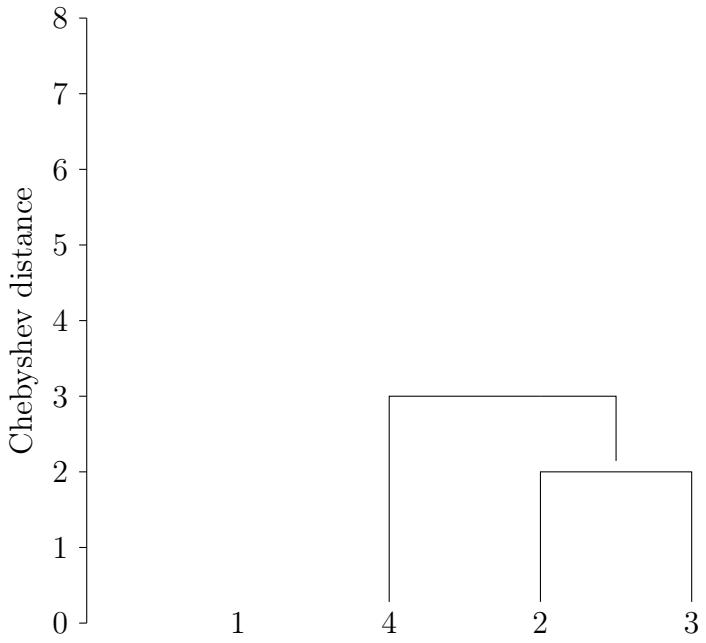


Рис. 2: Дендрограмма: шаг 3

$$d(2+3+4,\ 1) = \begin{cases} \mathbf{d(2,1) = 7} \text{ выбираем этот вариант согласно методу дальнего соседа} \\ d(3,1) = 6 \\ d(4,1) = 5 \end{cases}$$

Обновленная матрица расстояний:

$$D = \begin{bmatrix} & \mathbf{1} & \mathbf{2{+}3{+}4} \\ \mathbf{1} & 0 & 7 \\ \mathbf{2{+}3{+}4} & 7 & 0 \end{bmatrix}$$

Итак, мы объединяем два последних кластера в один большой кластер на уровне 7. Получается последний кластер *1+2+3+4*. Обновим дендрограмму:

Иерархический кластерный анализ реализован, дендрограмма построена. Теперь проверим себя в R:

```
hc <- hclust(D, method = "complete")
plot(hc, ylab = 'Chebyshev distance',
     main = 'Hierarchical cluster analysis', sub = '',
     xlab = 'complete linkage')
```

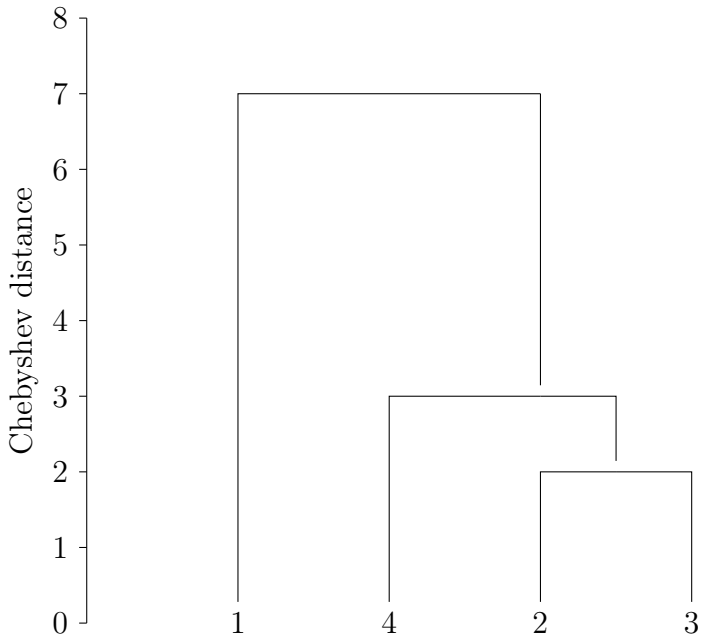
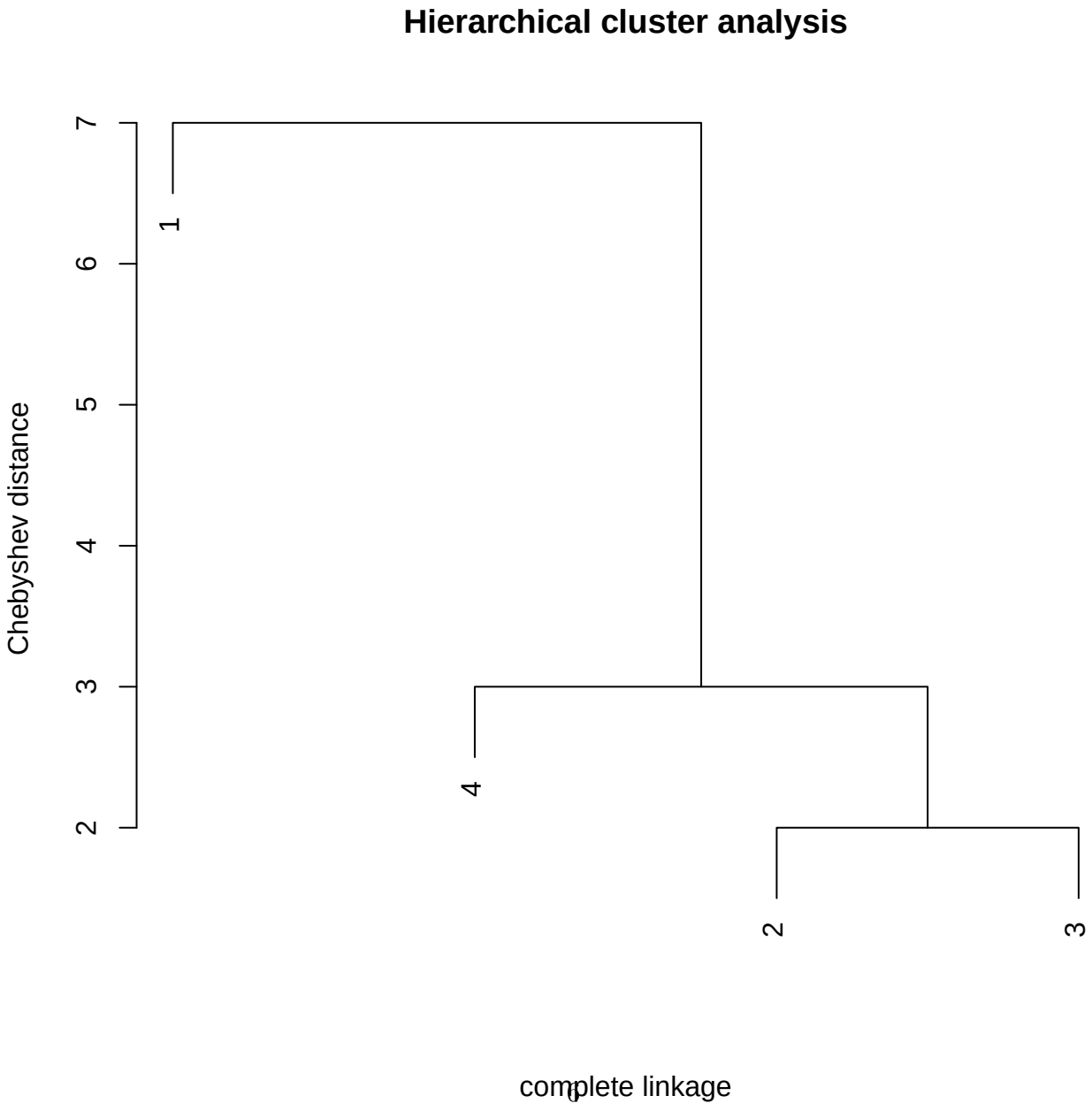


Рис. 3: Дендрограмма: шаг 4



Как мы можем увидеть, дендрограмма также была построена верно.

Задача 4. K-means, just k-means

На массиве данных из предыдущей задачи мы решили реализовать кластеризацию методом k-means с числом кластеров $K = 2$. На первом шаге алгоритма наблюдения были распределены на кластеры следующим образом (здесь указаны метки наблюдений, их id):

$$C_1 = \{2, 3, 4\}$$
$$C_2 = \{1\}$$

a. *Вычислите целевую функцию для первого кластера $W(C_1)$.*

$W(C_1) = \frac{1}{3} \cdot (((7-6)^2 + (7-4)^2 + (6-4)^2) + ((2-4)^2 + (2-1)^2 + (4-1)^2)) = \frac{1}{3} \cdot (14 + 14) = \frac{28}{3} \approx 9.33$

b. *Определите центроиды кластеров C_1 и C_2.*

$\bar{r}_1 = \left(\frac{7+6+4}{3}; \frac{2+4+1}{3}\right) = \left(\frac{17}{3}; \frac{7}{3}\right) \approx (5.67, 2.33)$

$\bar{r}_2 = (0, 6)$, поскольку кластер $C_2 = \{1\}$ состоит из одного наблюдения.

Проиллюстрируем центроиды:

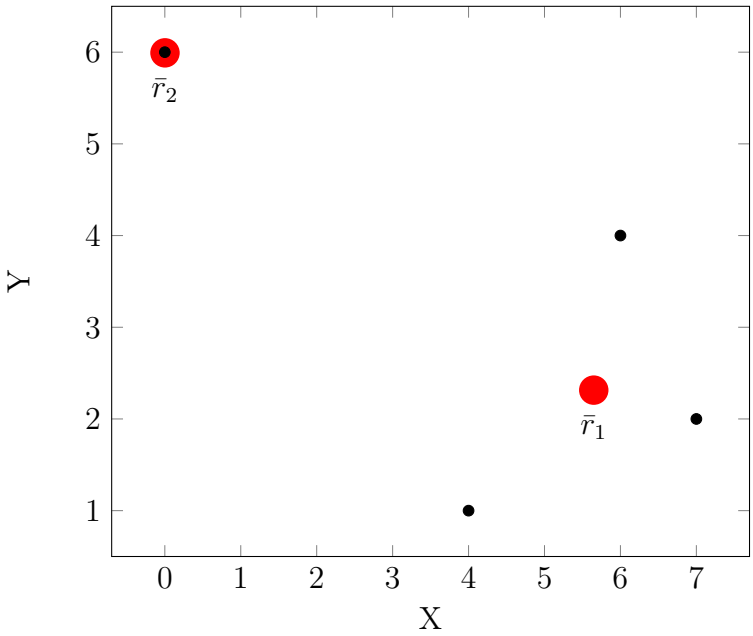


Рис. 4: Диаграмма рассеивания X vs Y

Задача 5. Бонусная задача

Когда идет речь о кластерном (в особенности иерархическом) анализе, я сразу же представляю корни дерева: именно корни, а не листва (несмотря на дендрограмму), поскольку корни сходятся в одно место именно сверху, это более интуитивно понятно.

Поэтому представляю рисунок своей ассоциации! Прошу прощения, на большее качество не хватило времени и большого набора профессиональных карандашей и других инструментов, которые остались дома (в другом регионе) :(

Практическая часть

Описание данных

Задача 1. Загрузка данных

Загрузим данные, которые содержатся в файле `cpds_2019.csv` и **посмотрим** на них, на их структуру и описательные статистики, чтобы представлять, с чем мы имеем дело.

```
dat <- read.csv('cpds_2019.csv')
head(dat)
```

```
##   X   country iso poco eu emu gov_party gov_type vturn gov_right1 gov_cent1
## 1 1 Australia AUS    0  0   0         1        5  91.8 100.000000   0.00000
## 2 2   Austria AUT    0  1   1         1        6  75.6  19.320000  21.29000
## 3 3   Belgium BEL    0  1   1         1        6  88.4  76.920000  23.08000
## 4 4  Bulgaria BGR    1  1   0         1        2  54.1 100.000000   0.00000
## 5 5    Canada CAN    0  0   0         1        1  67.0   0.000000 100.00000
## 6 6   Croatia HRV    1  1   0         1        5  52.6   4.761905  85.71429
##   gov_left1 leftsoc1 comm1 agrarian1 conserv1 relig1 liberal1 green1 ethnic1
## 1         0      0.0     0      4.51    36.66    0.0      0.0   10.4     0.0
## 2         0      0.0     0      0.00     0.00   37.5      0.0   13.9     0.0
## 3         0      8.6     0      0.00     0.00    0.0      0.0    6.1    16.0
## 4         0     27.9     0      0.00    33.50    0.0      0.0    0.0     9.2
## 5         0      0.0     0      0.00    34.30    0.0     33.1    6.5     7.6
## 6         0      0.0     0      0.00     0.00   19.0      4.4    0.0     0.0
```

```
str(dat)
```

```
## 'data.frame': 36 obs. of  20 variables:
##  $ X         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ country   : chr  "Australia" "Austria" "Belgium" "Bulgaria" ...
##  $ iso       : chr  "AUS" "AUT" "BEL" "BGR" ...
##  $ poco      : int  0 0 0 1 0 1 0 1 0 1 ...
##  $ eu        : int  0 1 1 1 0 1 1 1 1 1 ...
##  $ emu       : int  0 1 1 0 0 0 1 0 0 1 ...
##  $ gov_party : int  1 1 1 1 1 1 1 2 3 2 ...
##  $ gov_type  : int  5 6 6 2 1 5 1 5 4 2 ...
##  $ vturn     : num  91.8 75.6 88.4 54.1 67 52.6 66.7 60.8 84.6 63.7 ...
##  $ gov_right1: num  100 19.3 76.9 100 0 ...
##  $ gov_cent1 : num  0 21.3 23.1 0 100 ...
##  $ gov_left1 : num  0 0 0 0 0 ...
##  $ leftsoc1  : num  0 0 8.6 27.9 0 0 0 0 7.7 0 ...
##  $ comm1     : num  0 0 0 0 0 0 25.7 0 0 0 ...
##  $ agrarian1 : num  4.51 0 0 0 0 0 0 0 0 0 ...
##  $ conserv1  : num  36.7 0 0 33.5 34.3 ...
##  $ relig1    : num  0 37.5 0 0 0 19 0 5.8 0 0 ...
##  $ liberal1  : num  0 0 0 0 33.1 4.4 14.5 0 8.6 0 ...
##  $ green1    : num  10.4 13.9 6.1 0 6.5 0 4.8 0 3 0 ...
##  $ ethnic1   : num  0 0 16 9.2 7.6 0 0 0 0 0 ...
```

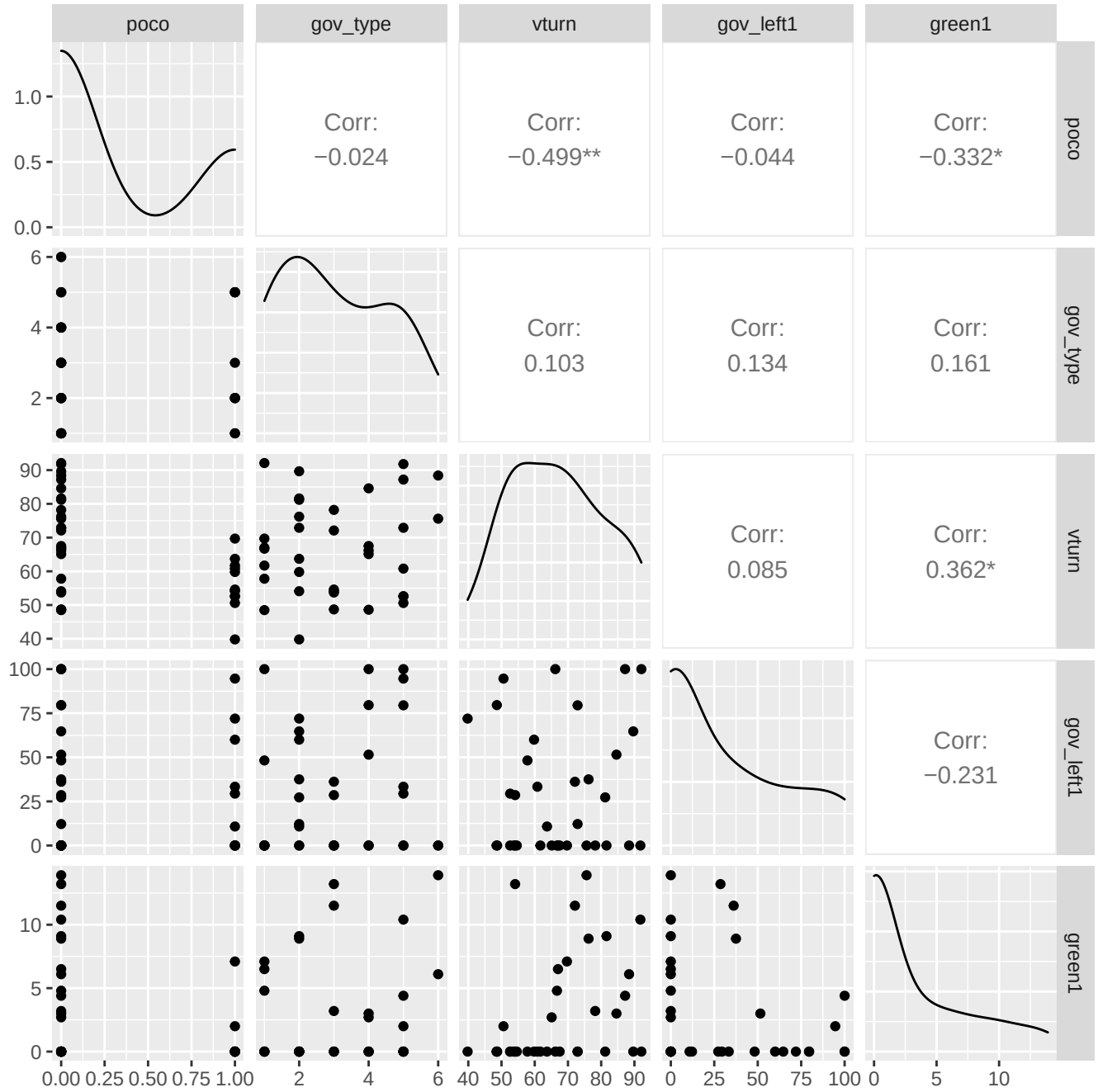
```
summary(dat)
```

```
##        X           country              iso                 poco
##  Min.   : 1.00   Length:36          Length:36          Min.   :0.0000
##  1st Qu.: 9.75   Class :character   Class :character   1st Qu.:0.0000
##  Median :18.50   Mode  :character   Mode  :character   Median :0.0000
##  Mean   :18.50                                         Mean   :0.3056
##  3rd Qu.:27.25                                         3rd Qu.:1.0000
##  Max.   :36.00                                         Max.   :1.0000
##        eu              emu        gov_party        gov_type         vturn
##  Min.   :0.0000   Min.   :0.0   Min.   :1.000   Min.   :1.000   Min.   :39.80
##  1st Qu.:1.0000   1st Qu.:0.0   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:54.10
##  Median :1.0000   Median :0.5   Median :2.000   Median :3.000   Median :66.45
##  Mean   :0.7778   Mean   :0.5   Mean   :2.167   Mean   :3.056   Mean   :66.90
##  3rd Qu.:1.0000   3rd Qu.:1.0   3rd Qu.:3.000   3rd Qu.:4.250   3rd Qu.:76.70
##  Max.   :1.0000   Max.   :1.0   Max.   :5.000   Max.   :6.000   Max.   :92.10
##    gov_right1       gov_cent1        gov_left1         leftsoc1
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.: 15.68   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.: 0.000
##  Median : 41.85   Median :  2.50   Median : 11.48   Median : 0.000
##  Mean   : 43.88   Mean   : 19.26   Mean   : 29.59   Mean   : 2.622
##  3rd Qu.: 71.86   3rd Qu.: 27.81   3rd Qu.: 53.63   3rd Qu.: 1.250
##  Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :27.900
##      comm1           agrarian1        conserv1         relig1
##  Min.   : 0.000   Min.   : 0.00   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 0.000   Median : 0.00   Median : 1.20   Median : 0.000
##  Mean   : 1.356   Mean   : 1.57   Mean   :12.46   Mean   : 6.328
##  3rd Qu.: 0.000   3rd Qu.: 0.00   3rd Qu.:25.05   3rd Qu.: 5.725
##  Max.   :25.700   Max.   :13.80   Max.   :44.50   Max.   :43.700
##     liberal1          green1          ethnic1
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 0.000   Median : 0.000
##  Mean   : 7.228   Mean   : 2.967   Mean   : 1.925
##  3rd Qu.:11.650   3rd Qu.: 5.125   3rd Qu.: 0.575
##  Max.   :52.900   Max.   :13.900   Max.   :16.000
```

Задача 2. Отбор переменных

Выберем из полученного датафрейма **переменные интереса** – переменные, по которым мы будем кластеризовать страны – и сохраним их в новый датафрейм `to_clust`. Кроме того, посмотрим на их распределение и корреляцию между ними.

```
library(tidyverse)
to_clust <- dat %>% select(poco, gov_type, vturn,
                           gov_left1, green1)
rownames(to_clust) <- dat$country

library(GGally)
ggpairs(to_clust)
```



Итак, предлагается взять для кластеризации **5 переменных**, представляющих интерес[1]:

1. `poco` — Дамми-переменная для посткоммунистических стран;
2. `gov_type` — Тип правительства на основе следующей классификации:
   - (1) Однопартийное правительство большинства: одна партия занимает все места в правительстве и имеет парламентское большинство [>50,0%].
   - (2) Минимальная победившая коалиция: все участвующие партии необходимы для формирования правительства большинства [>50,0%].
   - (3) Избыточная коалиция (Surplus coalition): коалиционные правительства, которые превышают критерий минимального выигрыша [>50,0%].
   - (4) Однопартийное правительство меньшинства: партия в правительстве не имеет большинства в парламенте [50,0%].

[1] Информация взята из https://www.cpds-data.org/images/Update2021/Codebook_CPDS_1960-2019_Update_2021.pdf.

- (5) Многопартийное правительство меньшинства: партии в правительстве не имеют большинства в парламенте [50,0%].
- (6) Правительство-смотритель (Caretaker government): правительства, которые должны просто поддерживать статус-кво.
- (7) Технократическое правительство: возглавляемое технократическим премьер-министром, состоящее из большинства технократических министров и обладающее мандатом на изменение статус-кво.

3. `vturn` — Явка избирателей на выборах,
4. `gov_left1` — Доля левых партий в процентах от общего числа должностей в кабинете правительства. Взвешено по количеству дней пребывания в должности в данном году;
5. `green1` — Доля „зеленых“ партий в процентах от общего числа должностей в кабинете правительства.

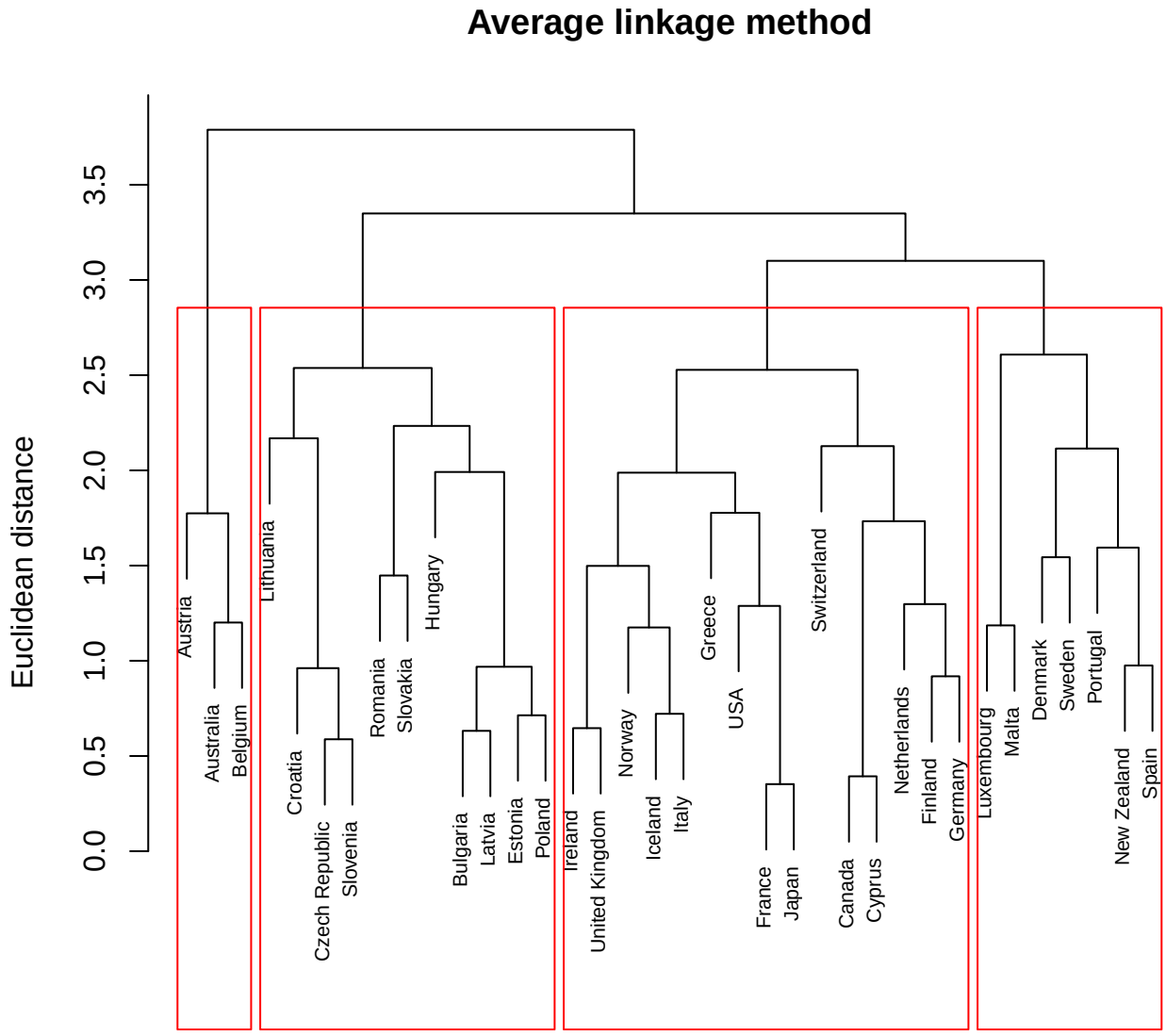
Можно отметить, что тот факт, что переменная `gov_type`, по сути закодирована как факторная переменная, создает некоторое **ограничение** для анализа, но сделаем допущение о том, что она представляет количественную шкалу: от отсутствия коалиционной политики в государстве ((1) — Однопартийное правительство большинства) до полной коалиционности ((7) — Технократическое правительство).

Задача 3. Иерархический кластерный анализ

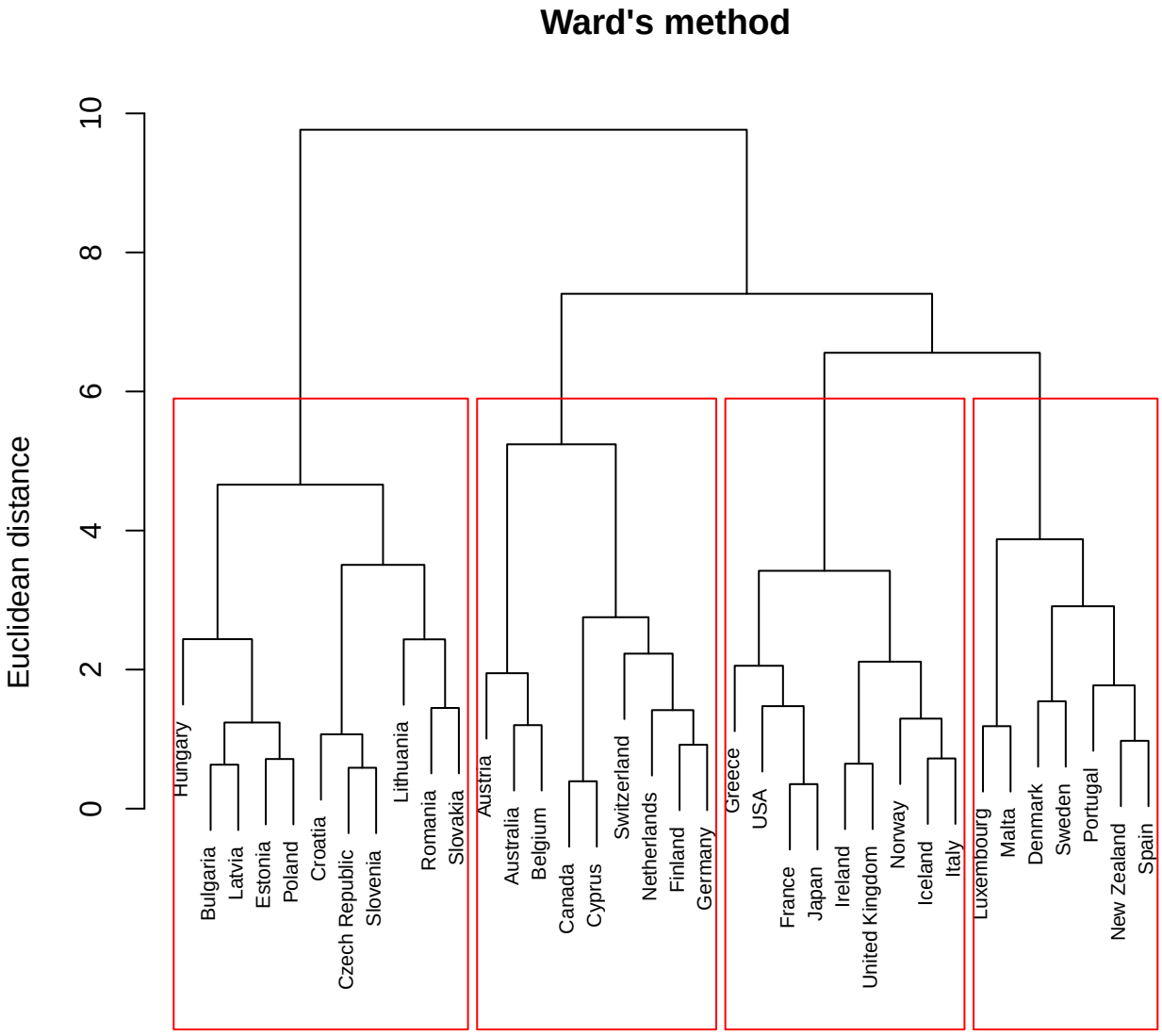
Реализуем иерархический кластерный анализ на основе датафрейма `to_clust`.

Выберем в качестве расстояния **евклидово расстояние** (L2), поскольку оно является самым распространенным и надежным в статистике. Также в качестве метода агрегирования выберем сразу два разных метода: более классический (и простой) **метод средней связи** (avereage linkage), как довольно надежный метод без ярких негативных сторон, и более продвинутый и распространенный **метод Уорда** (Ward's method), а затем — экспертно **выберем более подходящий** и качественно реализованный вариант иерархической кластеризации.

```
D <- dist(scale(to_clust), method = "euclidean")

hc <- hclust(D, method = "average")
plot(hc, main = "Average linkage method", cex = 0.7,
     ylab = 'Euclidean distance',
     xlab = '',
     sub = '')
rect.hclust(hc, k = 4, border = "red")
```




```
hc_ward <- hclust(D, method = "ward.D2")
plot(hc_ward, main = "Ward's method", cex = 0.7,
     ylab = 'Euclidean distance',
     xlab = '',
     sub = '')
rect.hclust(hc_ward, k = 4, border = "red")
```



На первый взгляд кластеры практически не отличаются друг от друга, однако это не так. С одной стороны, можно заметить, что метод средней связи формирует кластеры из одного наблюдения (**монокластеры**), но, с другой стороны, он выделяет *интересный кластер* из Австрии, Австралии и Бельгии (как мы потом выясним, не зря).

Метод Уорда же выделяет **кластеры** достаточно **маленького размера**, что также является небольшим недостатком, а также относит упомянутый выше кластер из трех стран к другому кластеру (а не соединяет его последним, как это делает метод средней связи).

Тем не менее, предлагается использовать именно метод Уорда, как более надежный и распространенный, однако *держать в голове информацию* про кластер Австрия-Австралия-Бельгия.

Хочется отметить, что в обоих случаях довольно очевидно **выделяются 4 основных кластера** стран: это, условно говоря, группа постсоветских стран, группа (вероятно) стран с „левым“ правительством, группа (вероятно) стран с „правым“ правительством и еще одна группа стран. На график нанесены **границы** данных кластеров. Кажется логичным, визуально и содержательно, выбрать именно 4 кластера для дальнейшей работы. К тому же, все они соединяются примерно на одинаковом (относительно небольшом) расстоянии.

„Разрежем“ дендрограмму примерно на уровне около 6 и **сохраним полученные метки** кла-

стеров в датафрейм `to_clust`, а также **представим их в факторном виде**.

```
clust <- cutree(hc_ward, k = 4)
to_clust$clust <- factor(clust)
```

Задача 4. Оценка качества кластеризации

Проведем ***проверку качества кластеризации***.

- **Выберем строки** в `to_clust`, соответствующие каждому полученному кластеру, и **прокомментируем**, какие страны входят в каждый кластер.

```
cluster01 <- to_clust %>% filter(clust == 1)
cluster01
```

```
##             poco gov_type vturn gov_left1 green1 clust
## Australia      0        5  91.8   0.00000   10.4     1
## Austria        0        6  75.6   0.00000   13.9     1
## Belgium        0        6  88.4   0.00000    6.1     1
## Canada         0        1  67.0   0.00000    6.5     1
## Cyprus         0        1  66.7   0.00000    4.8     1
## Finland        0        3  72.1  36.16000   11.5     1
## Germany        0        2  76.2  37.50000    8.9     1
## Netherlands    0        2  81.6   0.00000    9.1     1
## Switzerland    0        3  54.1  28.57143   13.2     1
```

Можно заметить, что в **первый кластер** входят станы, которые не являются посткоммунистическими, в каждом из их правительств есть зеленые партии (а вот левые — не везде), явка на выборах в них является достаточно высокой (кроме Швейцарии). Кроме того, можно заметить, что тип правительства у большинства данных стран достаточно „разношерстный“, в отличие от упомянутых выше Австралии, Австрии и Бельгии, которые выделяются схожестью практически во всех показателях, что было заметно еще по кластеризации методом средней связи.

```
cluster02 <- to_clust %>% filter(clust == 2)
cluster02
```

```
##                poco gov_type vturn gov_left1 green1 clust
## Bulgaria          1        2  54.1      0.00    0.0     2
## Croatia           1        5  52.6      0.00    0.0     2
## Czech Republic    1        5  60.8     33.33    0.0     2
## Estonia           1        2  63.7     10.78    0.0     2
## Hungary           1        1  69.7      0.00    7.1     2
## Latvia            1        3  54.6      0.00    0.0     2
## Lithuania         1        5  50.6     94.63    2.0     2
## Poland            1        1  61.7      0.00    0.0     2
## Romania           1        2  39.8     71.96    0.0     2
## Slovakia          1        2  59.8     60.00    0.0     2
## Slovenia          1        5  52.6     29.41    0.0     2
```

Во втором кластере находятся все страны, которые являются посткоммунистическими. Явка на выборах у них приблизительно низкая, зеленых партий в правительстве практически нет, а

вот левые правительства варьируются от 0% до ≈ 95%. Тип правительства у них также довольно сильно различается.

```
cluster03 <- to_clust %>% filter(clust == 3)
cluster03
```

```
##             poco gov_type vturn gov_left1 green1 clust
## Denmark        0        4 84.60     51.51    3.0     3
## Luxembourg     0        2 89.66     64.71    0.0     3
## Malta          0        1 92.10    100.00    0.0     3
## New Zealand    0        5 72.90     79.49    0.0     3
## Portugal       0        4 48.60     79.62    0.0     3
## Spain          0        4 66.20    100.00    0.0     3
## Sweden         0        5 87.20    100.00    4.4     3
```

В третьем кластере находятся страны, в которых бОльшую (кроме Португалии: 48,6 %) часть правительства занимают левые партии (при этом левых партий среди них практически нет). Явка на выборах у многих государств довольно высокая, однако есть исключения. Тип правительства снова сильно различается.

```
cluster04 <- to_clust %>% filter(clust == 4)
cluster04
```

```
##                poco gov_type vturn gov_left1 green1 clust
## France            0        3  48.7   0.00000    0.0     4
## Greece            0        1  57.8  48.23000    0.0     4
## Iceland           0        2  81.2  27.27273    0.0     4
## Ireland           0        4  65.1   0.00000    2.7     4
## Italy             0        2  72.9  12.18000    0.0     4
## Japan             0        3  53.7   0.00000    0.0     4
## Norway            0        3  78.2   0.00000    3.2     4
## United Kingdom    0        4  67.5   0.00000    0.0     4
## USA               0        1  48.5   0.00000    0.0     4
```

В последнем четвертом кластере находятся страны с преимущественно правыми и центристскими партиями в правительстве (почти) без зеленых партий. Явка на выборах у них варьируется от средне-низкой до средне-высокой. Тип правительства вновь различается.

Итак, можно подвести предварительный **вывод**: мы имеем кластер из посткоммунистических стран с низкой явкой на выборах практически без зеленых партий и большим разбросом по доле левых партий; кластер стран с высокой явкой и зелеными партиями в правительстве, но преимущественно без левых партий; кластер стран с высокой долей левых партий, но преимущественно без зеленых партий и высокой явкой на выборах; и кластер стран преимущественно без левых и зеленых партий в правительстве со средней явкой на выборах. К сожалению, по типу правительства (его структуре) сложно сделать какие-либо выводы, поскольку в большинстве случаев он смешан внутри каждого из кластеров.

- Выведем **описательные статистики** по каждому кластеру.

```
to_clust %>% group_by(clust) %>% tally
```

```
## # A tibble: 4 x 2
##   clust     n
##   <fct> <int>
## 1 1         9
## 2 2        11
## 3 3         7
## 4 4         9
```

```
to_clust %>% group_by(clust) %>% summarise_at(vars(poco:green1),
                                              .funs = median)
```

```
## # A tibble: 4 x 6
##   clust  poco gov_type vturn gov_left1 green1
##   <fct> <int>    <int> <dbl>     <dbl>  <dbl>
## 1 1         0        3  75.6       0      9.1
## 2 2         1        2  54.6      10.8    0
## 3 3         0        4  84.6      79.6    0
## 4 4         0        3  65.1       0      0
```

```
to_clust %>% group_by(clust) %>% summarise_at(vars(poco:green1),
                                                .funs = mean)
```

```
## # A tibble: 4 x 6
##   clust  poco gov_type vturn gov_left1 green1
##   <fct> <dbl>    <dbl> <dbl>     <dbl>  <dbl>
## 1 1         0     3.22  74.8     11.4   9.38
## 2 2         1     3     56.4     27.3   0.827
## 3 3         0     3.57  77.3     82.2   1.06
## 4 4         0     2.56  63.7      9.74  0.656
```

```
to_clust %>% group_by(clust) %>% summarise_at(vars(poco:green1),
                                                .funs = sd)
```

```
## # A tibble: 4 x 6
##   clust  poco gov_type vturn gov_left1 green1
##   <fct> <dbl>    <dbl> <dbl>     <dbl>  <dbl>
## 1 1         0     1.99 11.6       17.2   3.18
## 2 2         0     1.67  7.97      34.1   2.17
## 3 3         0     1.51 15.8       19.2   1.85
## 4 4         0     1.13 12.3       17.2   1.31
```

Несколько выводов на основе описательных статистик:

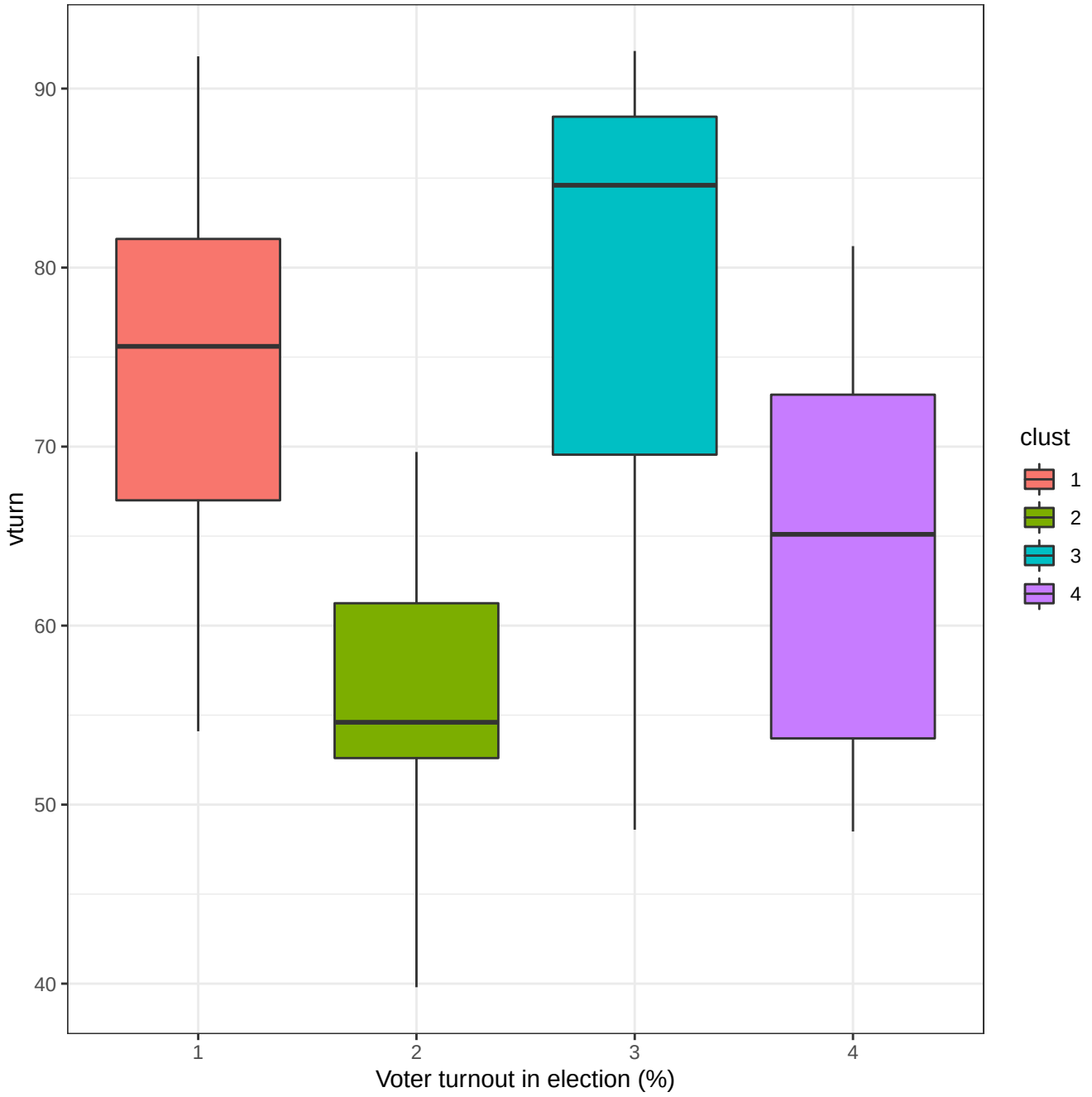
1. Прежде всего, заметно, что кластеры получились приблизительно одинакового размера (2-й немного больше и 3-й — меньше).
2. Видим еще одно подтверждение того, что все посткоммунистические страны находятся в одном кластере.
3. Страны с зелеными партиями в правительстве действительно сосредоточены в первом кластере — больше их практически нигде нет. К тому же, этот кластер занимает второе место по явке избирателей.

4. Страны с преимущественно „левым“ правительством находятся в третьем кластере. В этом же кластере — страны с наивысшей явкой.
5. Разброс типа правительства везде примерно одинаковый и сосредоточен вокруг среднего ≈ 3, что соответствует избыточной коалиции (Surplus coalition).
6. В кластере посткоммунистических стран наблюдается очень сильный разброс параметра, отвечающего за долю левых партий в правительстве. При этом, явка в этом кластере имеет наименьший разброс (в этих странах она ниже всех).

- **Визуализируем** распределения выбранных показателей по кластерам.

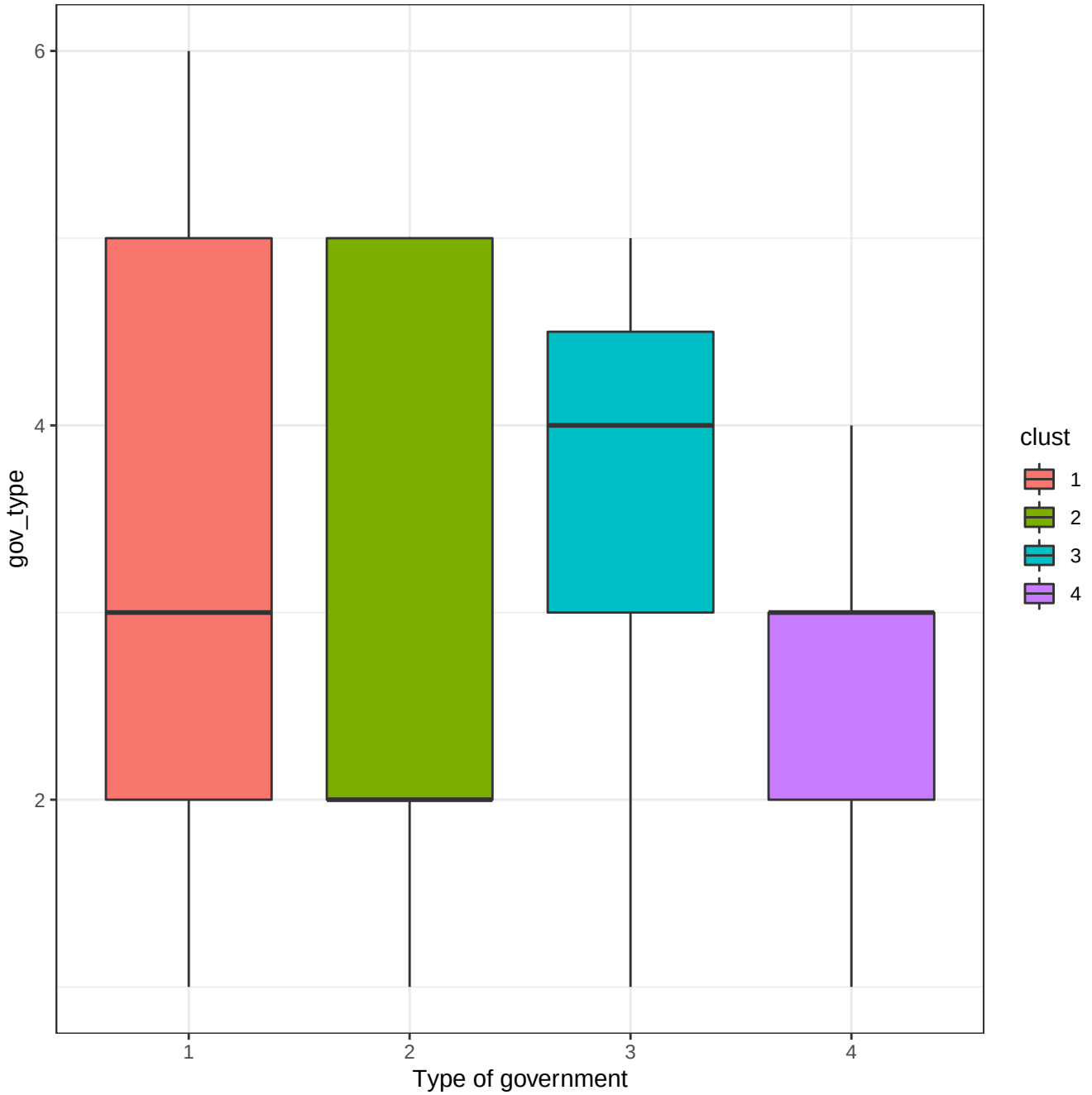
  В начале я представлю ряд ящиков с усами по отдельным переменным в разрезе по всем кластерам, а также ящики с усами по двум переменным, затем диаграммы рассеяния и в конце — трехмерные графики.

```
library(ggplot2)
ggplot(data = to_clust, aes(x = clust, y = vturn, fill = clust)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "Voter turnout in election (%)")
```



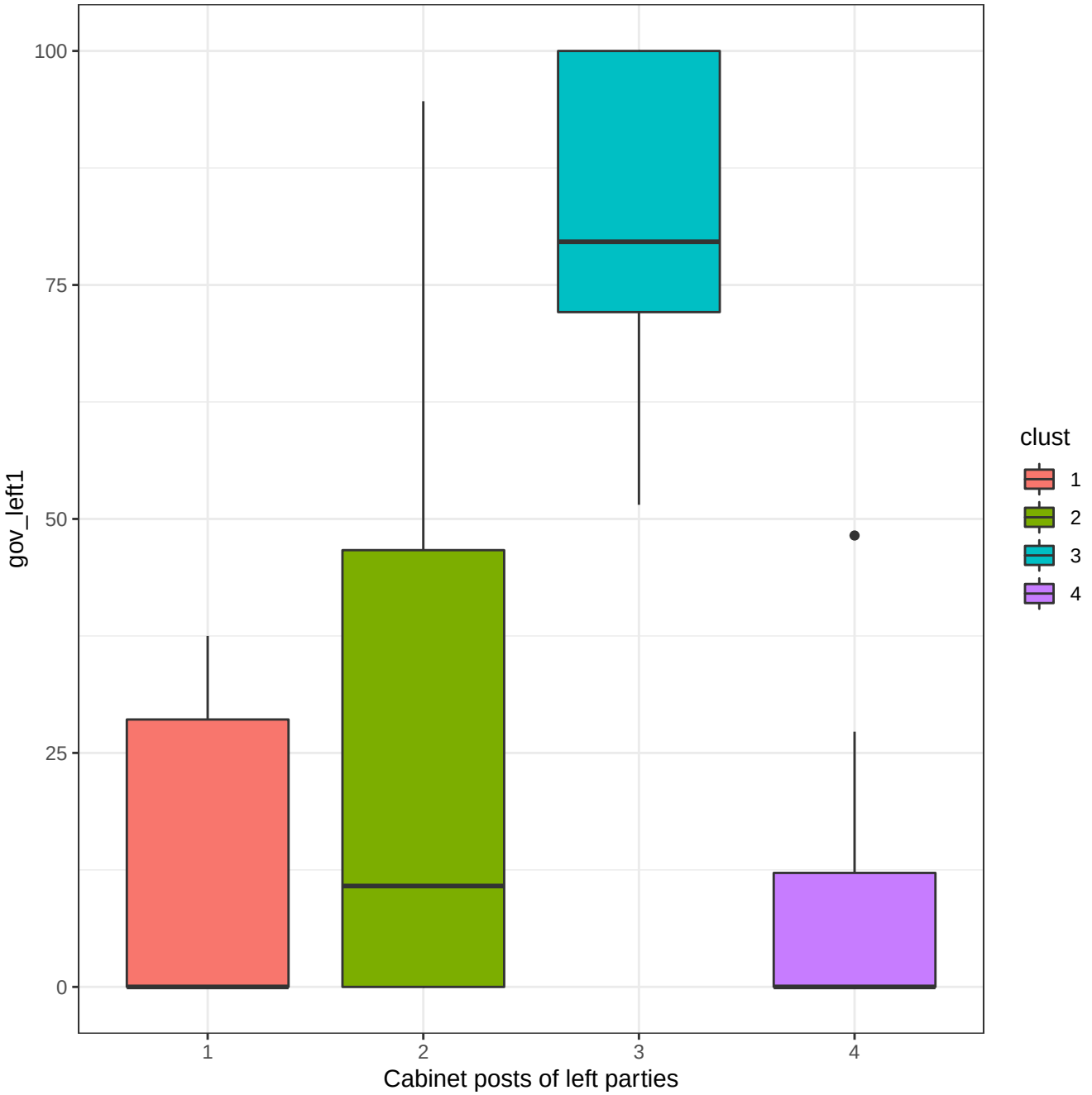
На данном графике можно увидеть разницу в распределении явки избирателей в кластерах: мы снова можем увидеть, что она минимальна в посткоммунистических странах, средняя для стран без левых и зеленых партий и максимальная для стран с левыми и зелеными партиями.

```
ggplot(data = to_clust, aes(x = clust, y = gov_type, fill = clust)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "Type of government")
```



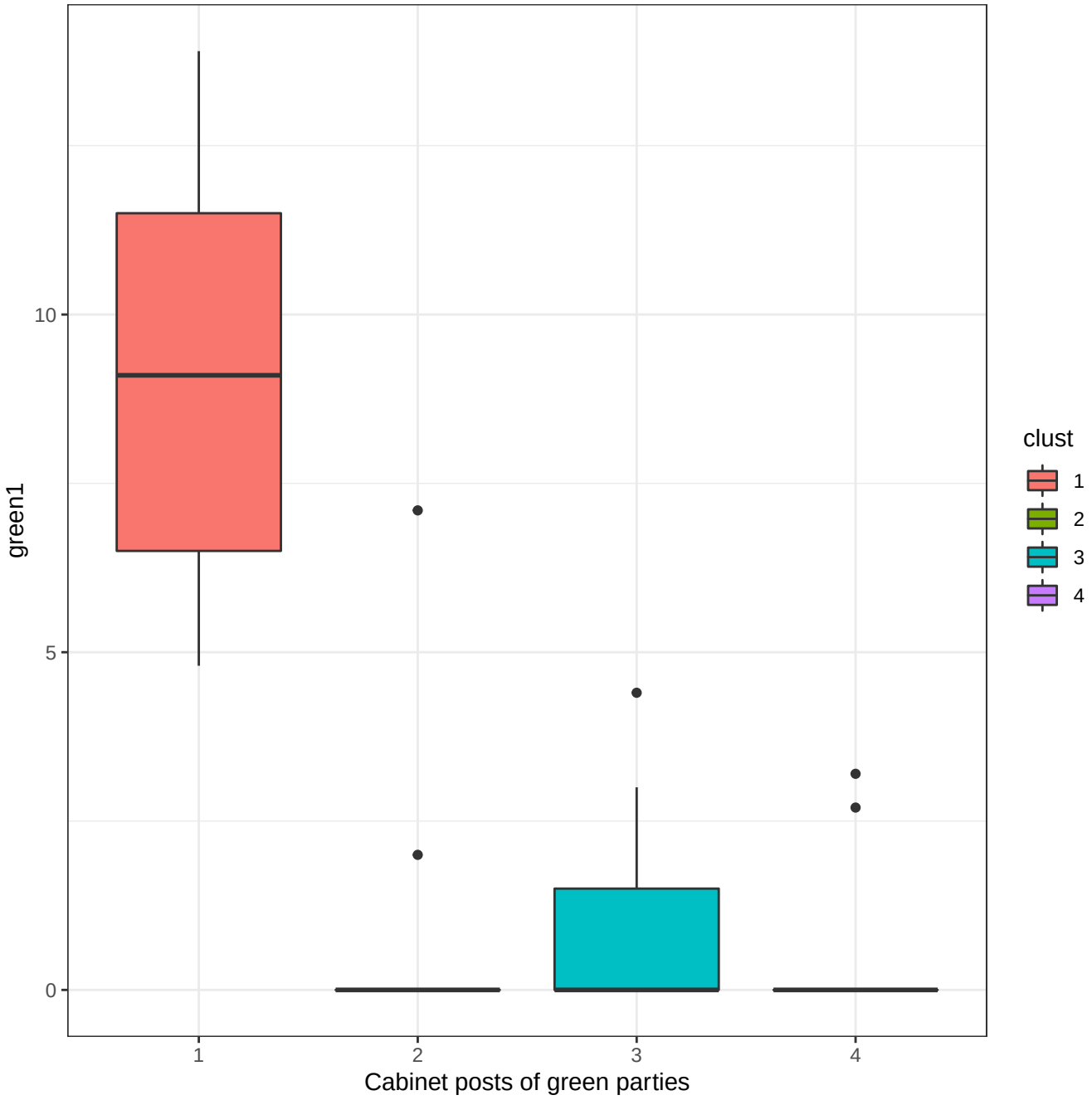
Из этого графика, как это не парадоксально, можно сделать предположение о том, что статистических различий между кластерами по типу (структуре) правительства практически нет.

```
ggplot(data = to_clust, aes(x = clust, y = gov_left1, fill = clust)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "Cabinet posts of left parties")
```



На этом графике заметно доминирование левых партий в правительстве для стран третьего кластера и их практически полное отсутствие в четвертом кластере („выбросом“ является Греция, имеющая 48.23% левых партий в правительстве).

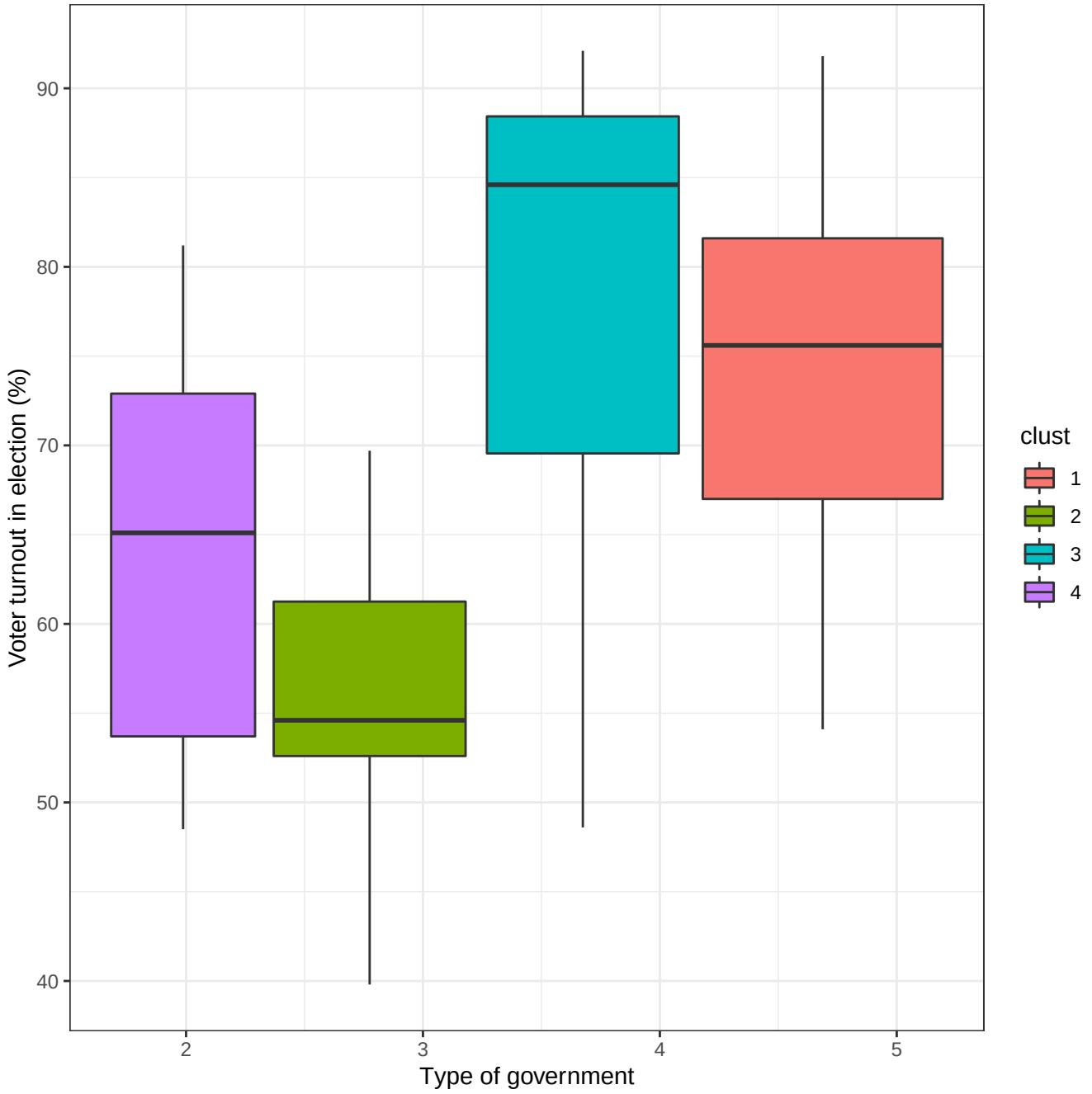
```
ggplot(data = to_clust, aes(x = clust, y = green1, fill = clust)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "Cabinet posts of green parties")
```



На приведенной визуализации можно увидеть заметное доминирование зеленых партий в первом кластере и их относительное большое присутствие в третьем кластере. Также заметен ряд выбросов.

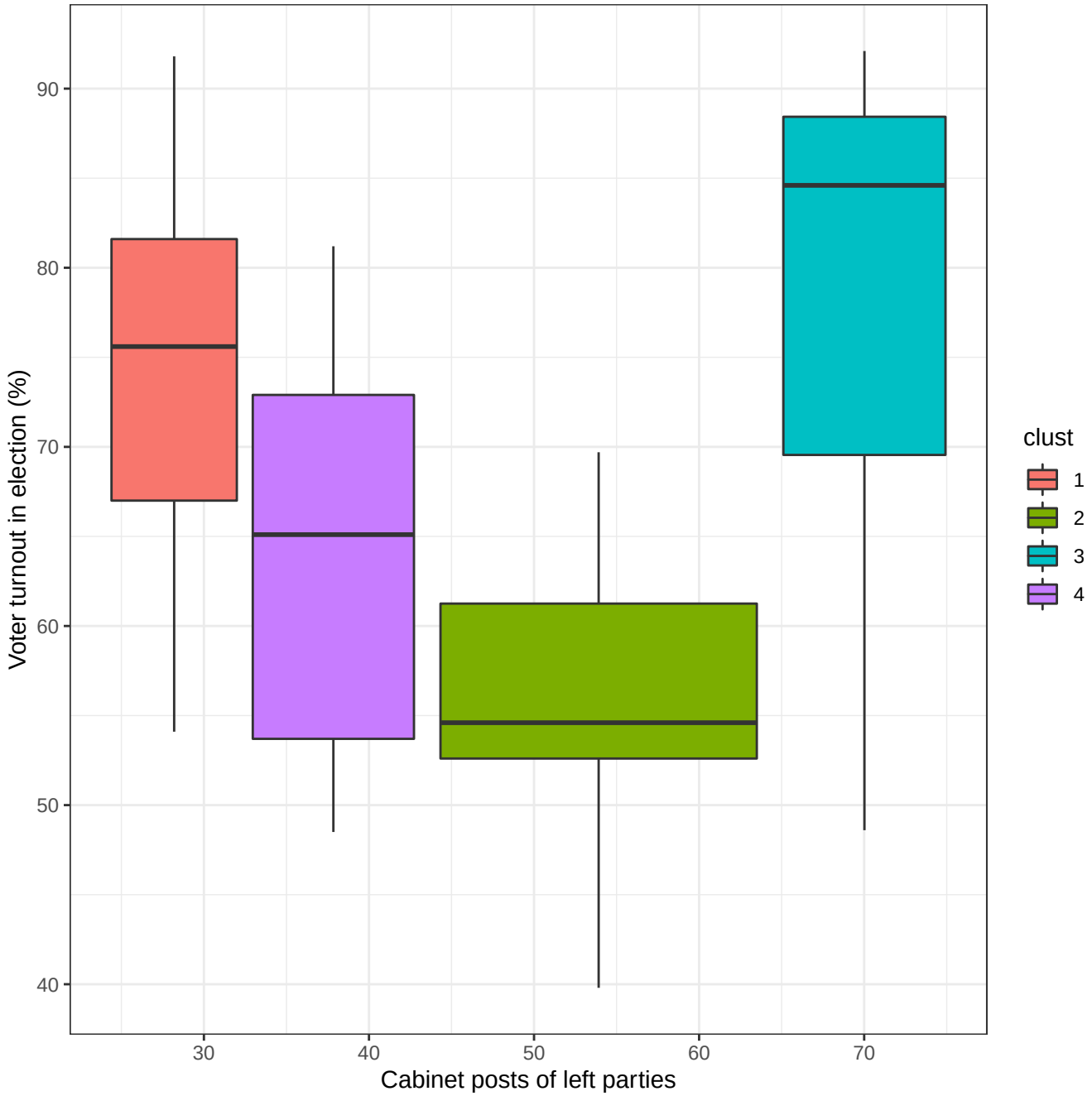
Далее следуют графики, имеющие **две разных переменных** по оси „x“ и „y“.

```
ggplot(data = to_clust, aes(x = gov_type, y = vturn, fill = clust)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "Type of government", y = 'Voter turnout in election (%)')
```



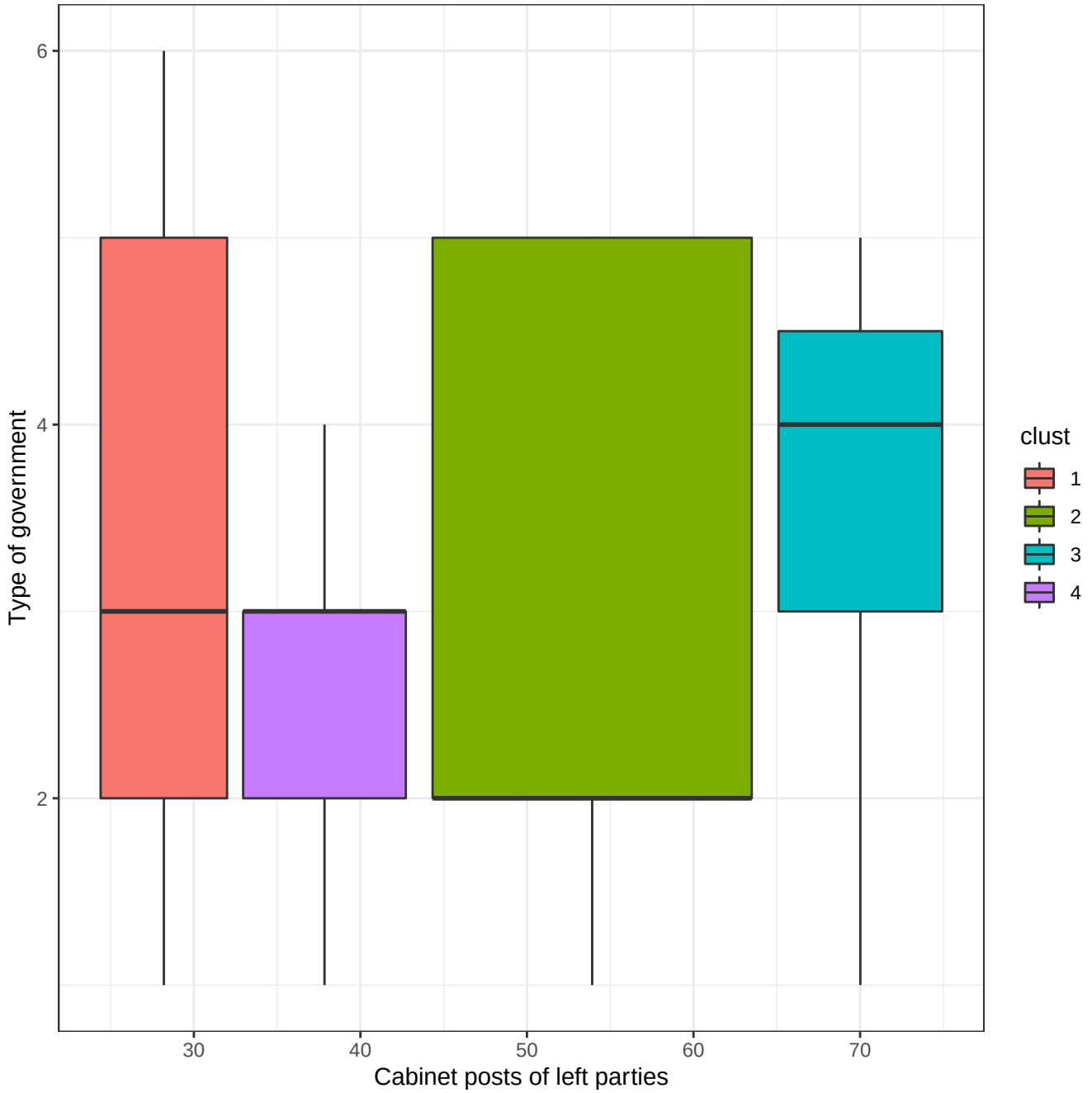
Из данного графика нельзя сделать однозначный вывод о наличии взаимосвязи между явкой избирателей и типом правительства, но можно заметить, что страны из третьего кластера (с левыми правительствами) скорее можно отнести к однопартийному правительству меньшинства и многопартийному правительству меньшинства (закодировано как 4 и 5 соответственно), т.е. признаку наличия сложных коалиционных правительств.

```
ggplot(data = to_clust, aes(x = gov_left1, y = vturn, fill = clust)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "Cabinet posts of left parties",
       y = 'Voter turnout in election (%)')
```



На данном графике можно обратить внимание на взаимосвязь явки на выборах и доли левых партий. Однозначный вывод относительно нее сделать сложно, но можно заметить, что, когда явка высока, левых партий в правительстве может быть как заметно больше, так и меньше. Когда их довольно среднее количество: то и явка также средняя.

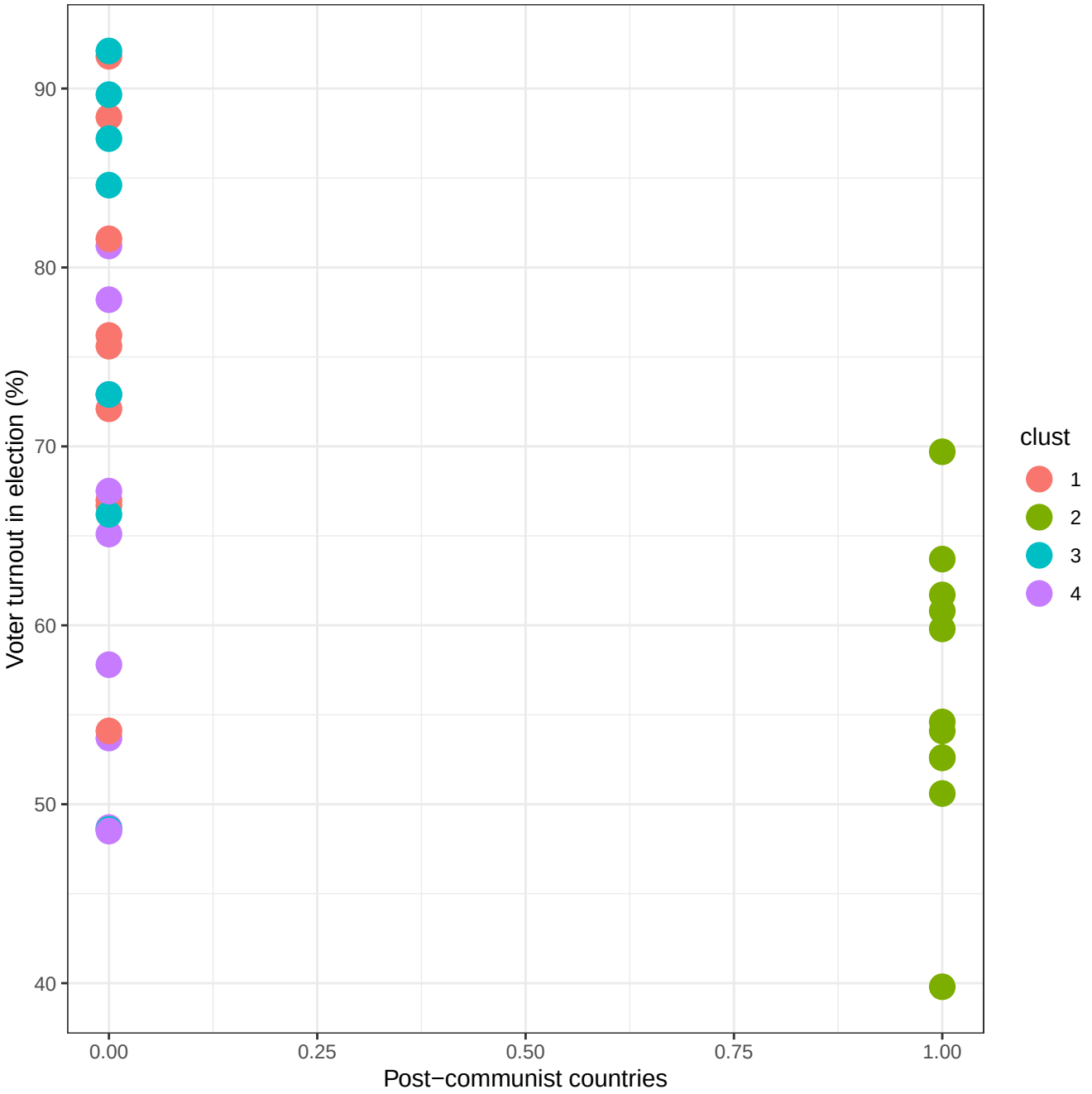
```
ggplot(data = to_clust, aes(x = gov_left1, y = gov_type, fill = clust)) +
  geom_boxplot() +
  theme_bw() +
  labs(x = "Cabinet posts of left parties",
       y = 'Type of government')
```



На данном графике можно заметить отсутствие связи между долей левых партий и типом правительства. Можно лишь отметить, что правительство редко бывает коалиционным в кластере стран без левых и зеленых партий.

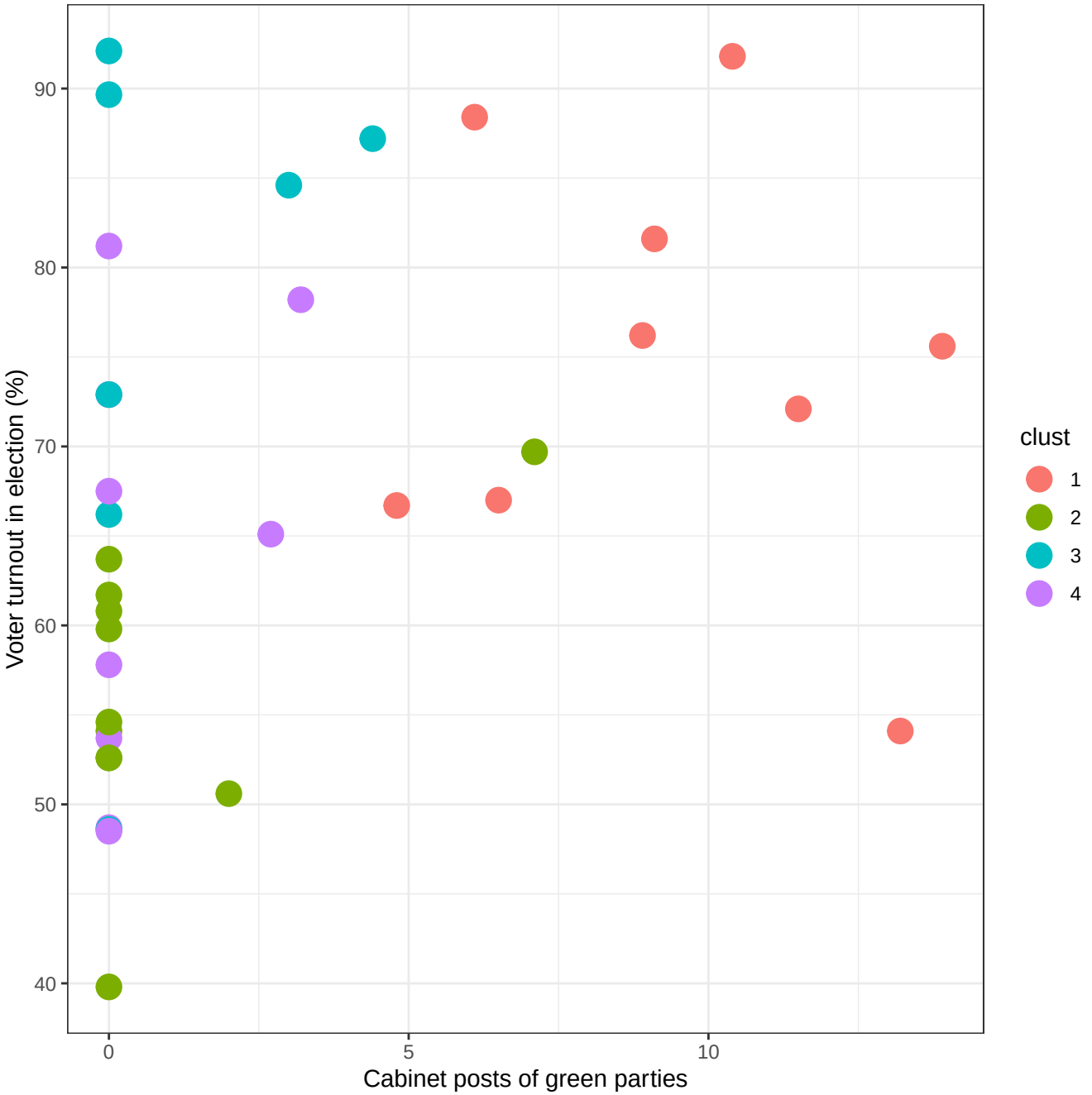
Далее будут представлены диаграммы рассеяния.

```
ggplot(data = to_clust, aes(x = poco, y = vturn, color = clust)) +
  geom_point(size = 5) +
  theme_bw() +
  labs(x = "Post-communist countries",
       y = 'Voter turnout in election (%)')
```



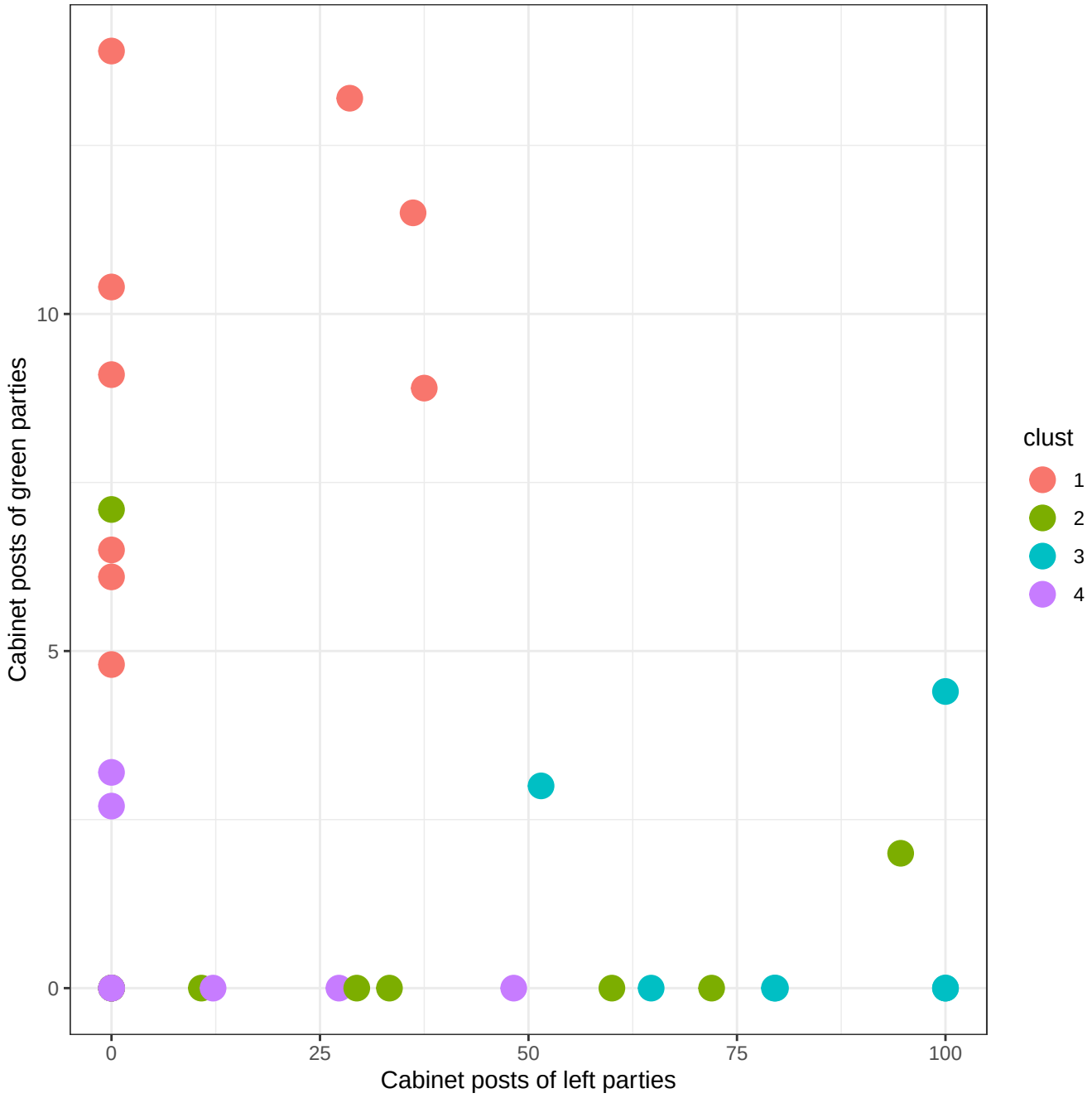
На данной визуализации можно более явно заметить, что посткоммунистическим странам характера более низкая явка на выборах (а странам с левыми и зелеными партиями — наибольшая).

```
ggplot(data = to_clust, aes(x = green1, y = vturn, color = clust)) +
  geom_point(size = 5) +
  theme_bw() +
  labs(x = "Cabinet posts of green parties",
       y = 'Voter turnout in election (%)')
```



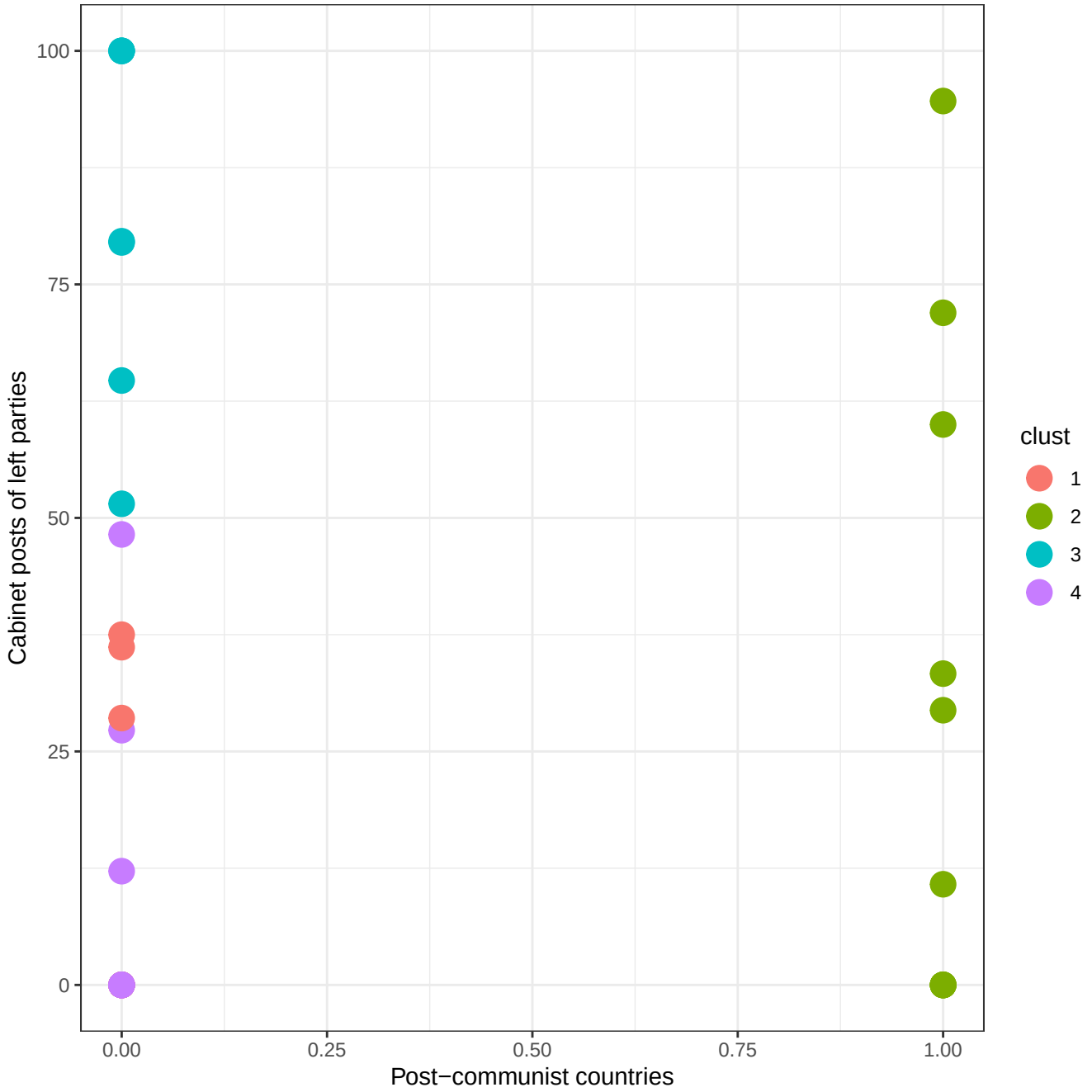
На данном графике можно обратить внимание на отсутствие взаимосвязи между явкой избирателей и долей зеленых партий (однако максимальная доля зеленых партий наблюдается в странах с довольно низкой явкой).

```
ggplot(data = to_clust, aes(x = gov_left1, y = green1, color = clust)) +
  geom_point(size = 5) +
  theme_bw() +
  labs(x = "Cabinet posts of left parties",
       y = 'Cabinet posts of green parties')
```



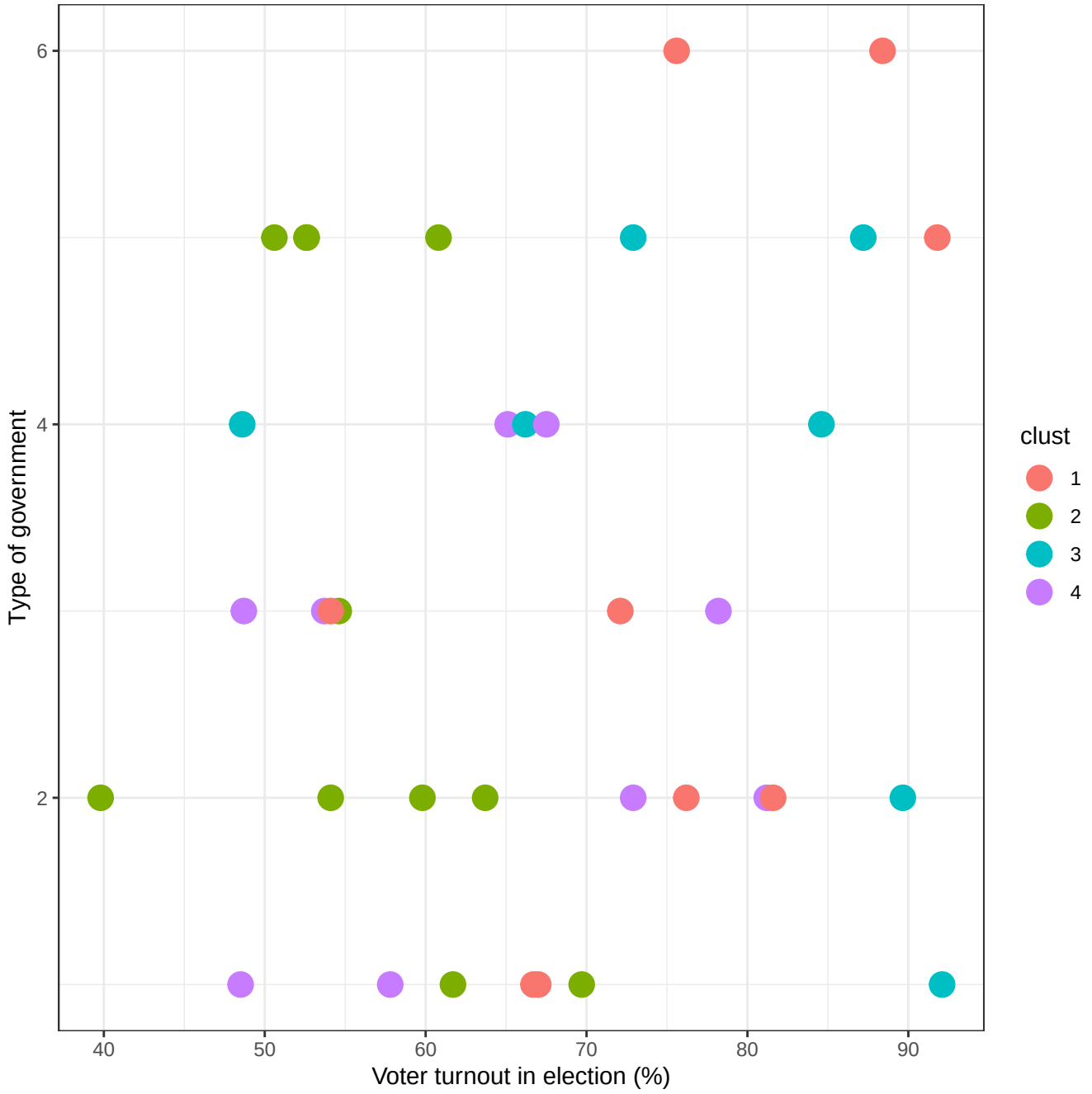
На данном графике можно увидеть отрицательную взаимосвязь между долями левых и зеленых партий в правительстве.

```
ggplot(data = to_clust, aes(x = poco, y = gov_left1, color = clust)) +
  geom_point(size = 5) +
  theme_bw() +
  labs(x = "Post-communist countries",
       y = 'Cabinet posts of left parties')
```



На данном графике можно явно заметить, что посткоммунистическим странам в равной степени характерны как низкие доли левых партий в правительстве, так и высокие.

```
ggplot(data = to_clust, aes(x = vturn, y = gov_type, color = clust)) +
  geom_point(size = 5) +
  theme_bw() +
  labs(x = "Voter turnout in election (%)",
       y = 'Type of government')
```



На данном графике можно вновь заметить (1) отсутствие взаимосвязи между типом правительства и явкой на выборах и (2) в целом тот факт, что кластеризация слабо заметна относительно переменной типа правительства.

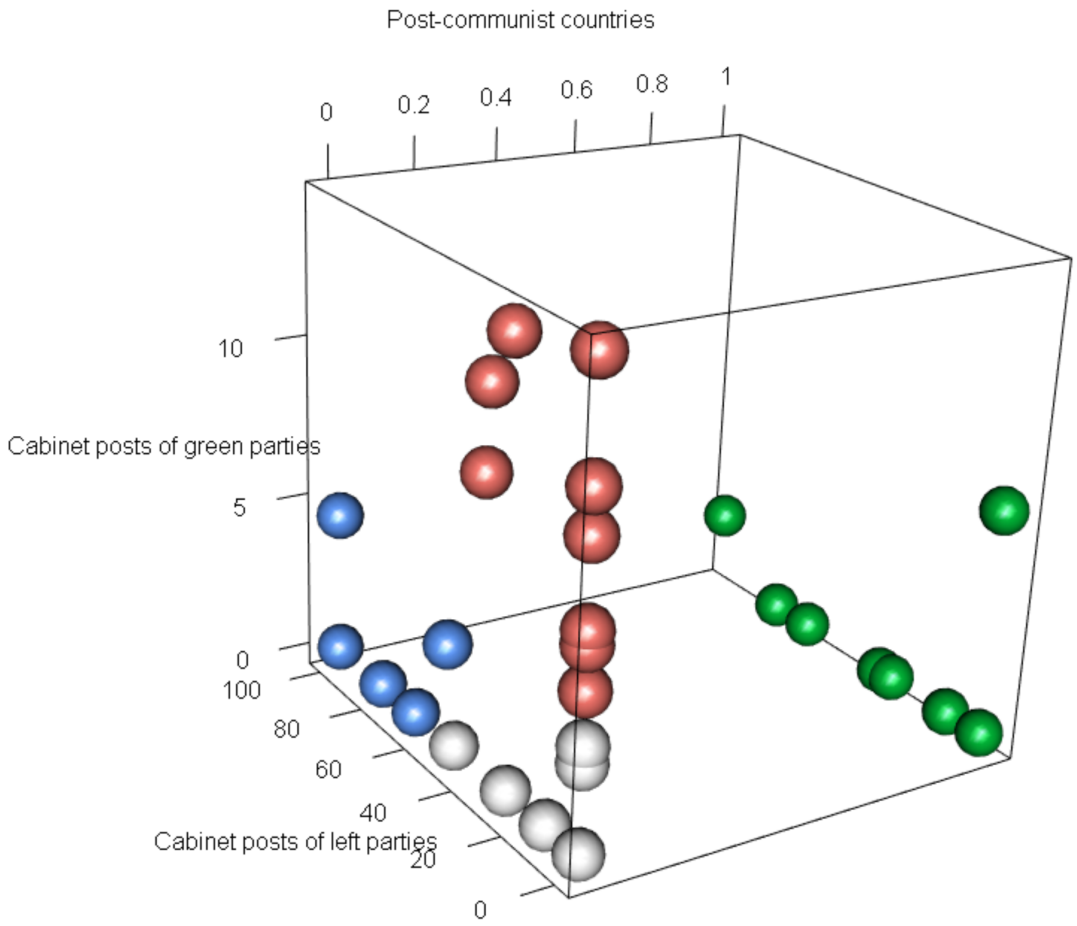
```
mycolors <- c('#F8766D', '#00BA38', '#619CFF')
to_clust$color <- mycolors[as.numeric(to_clust$clust)]

library(rgl)
setupKnitr()
plot3d(
  x = to_clust$poco,
  y = to_clust$gov_left1,
  z = to_clust$green1,
  col = to_clust$color,
  type = 's',
  radius = 3,
```

```
  xlab="Post-communist countries",
  ylab="Cabinet posts of left parties",
  zlab="Cabinet posts of green parties")
rglwidget()
```

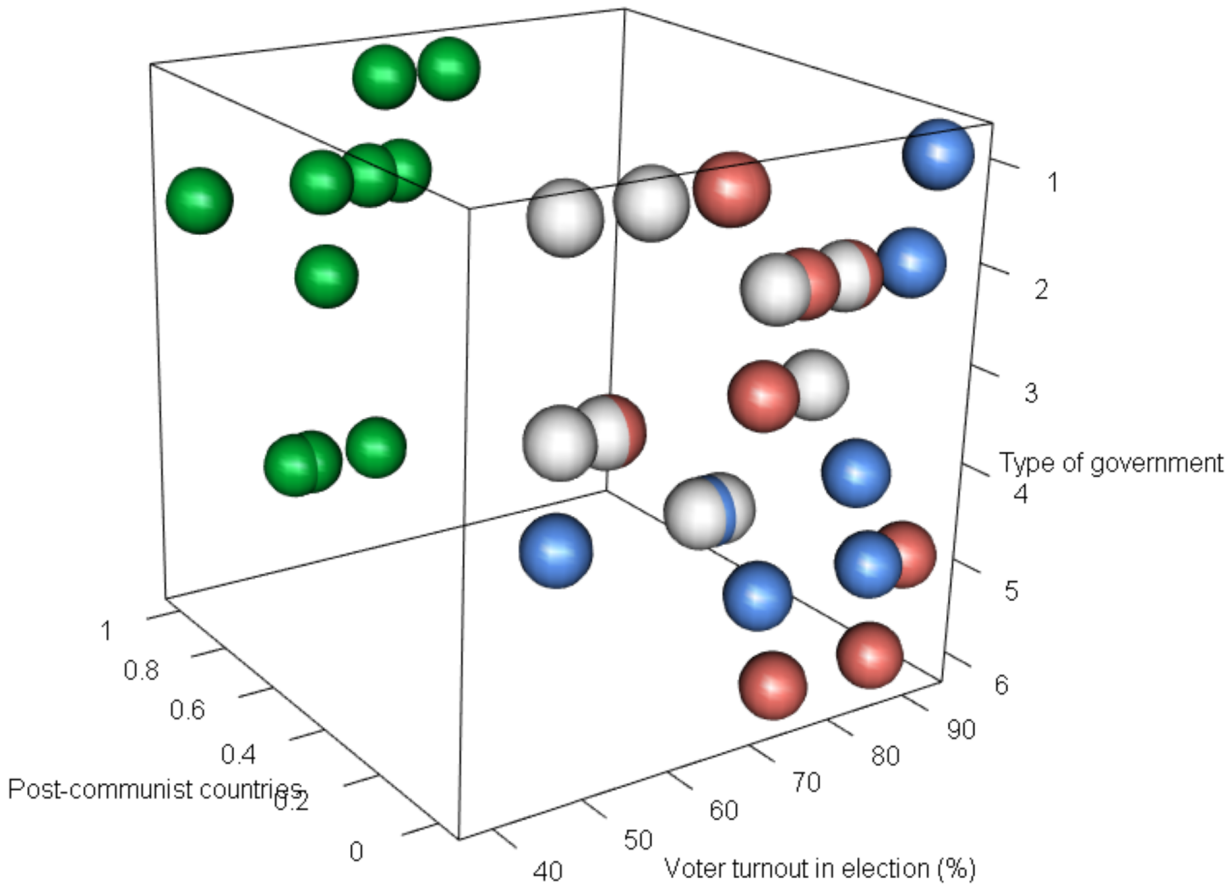
```
setupKnitr()

## NULL

plot3d(
  x = to_clust$vturn,
  y = to_clust$gov_type,
  z = to_clust$poco,
  col = to_clust$color,
  type = 's',
  radius = 2,
  xlab="Voter turnout in election (%)",
  ylab="Type of government",
  zlab="Post-communist countries")
rglwidget()
```

```
setupKnitr()

## NULL

plot3d(
  x = to_clust$vturn,
  y = to_clust$gov_type,
  z = to_clust$gov_left1,
  col = to_clust$color,
  type = 's',
  radius = 3,
  xlab="Voter turnout in election (%)",
  ylab="Type of government",
  zlab="Cabinet posts of left parties")
rglwidget()
```

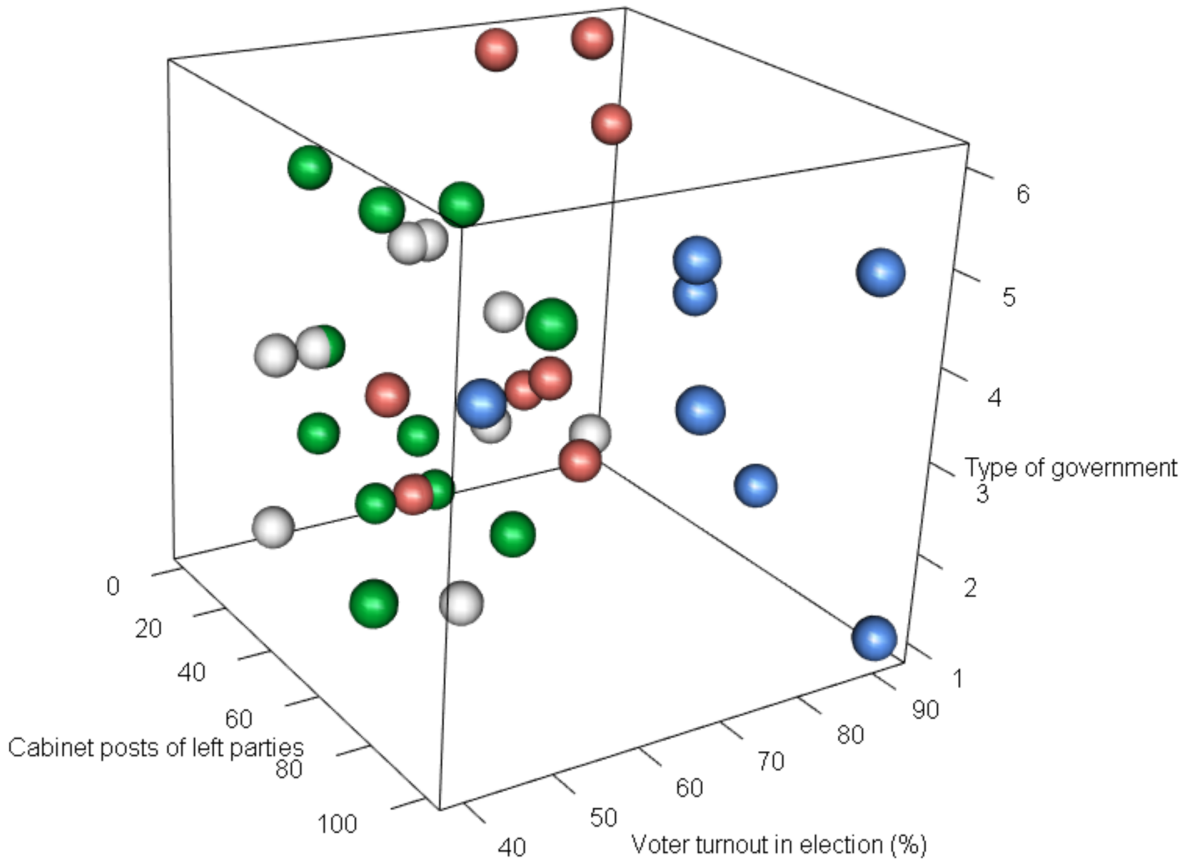
Далее будут вставлены статичные картинки, поскольку LaTeXне выводит данные графики сам:



В отношении посткоммунистических стран можно заметить, что, помимо того, что им характерна низкая явка на выборах, в их правительствах также практически нет зеленых партий. Также можно более явно заметить различия между другими кластерами относительно отрицательной взаимосвязи зеленых и левых партий. Видно, по какой „линии“ прошла кластеризация.



На данном графике можно увидеть, как „перемешаны“ между собой наблюдения из разных кластеров в отношении типа устройства парламента, о чем говорилось ранее. Также вновь заметно, что среди посткоммунистических стран нет государств с высокой явкой на выборах.



Данный график сложно представить в виде статичной картинки, но можно заметить очень слабую связь (если она вообще есть) между типом правительства и долей левых партий. Также вновь бросается в глаза высокая явка для стран с большими долями левых и зеленых партий.

- Используем **формальные статистические тесты** для определения того, отличаются ли средние значения/распределения показателей по кластерам.

  Начнем с рангового критерия Краскела—Уоллиса, предназначенного для проверки равенства медиан и распределений нескольких выборок. Из матрицы в начале документа можно заметить, что нормальности распределения не наблюдалось.

```
kruskal.test(to_clust$vturn ~ to_clust$clust)

##
##  Kruskal-Wallis rank sum test
##
## data:  to_clust$vturn by to_clust$clust
## Kruskal-Wallis chi-squared = 12.454, df = 3, p-value = 0.005978

kruskal.test(to_clust$gov_type ~ to_clust$clust)

##
##  Kruskal-Wallis rank sum test
```

```
##
## data:  to_clust$gov_type by to_clust$clust
## Kruskal-Wallis chi-squared = 1.4984, df = 3, p-value = 0.6826
```

```
kruskal.test(to_clust$gov_left1 ~ to_clust$clust)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  to_clust$gov_left1 by to_clust$clust
## Kruskal-Wallis chi-squared = 17.414, df = 3, p-value = 0.0005808
```

```
kruskal.test(to_clust$green1 ~ to_clust$clust)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  to_clust$green1 by to_clust$clust
## Kruskal-Wallis chi-squared = 23.46, df = 3, p-value = 3.238e-05
```

Можно заметить, что **нулевая гипотеза отвергается во всех случаях, кроме типа правительства**. До этого из визуализаций было неоднократно замечено, что в кластеризации эта переменная учтена плохо, многие наблюдения из разных кластеров буквально „сливались“ на графиках, связанных с ней. Остальные медианы же можно назвать различающимися во всех кластерах на очень высоком уровне доверия.

Далее представим **критерий Хи-квадрат** для таблицы сопряженности, которая описывает распределение дамми-переменной для посткоммунистических стран (*ведь, по сути, это выборочная доля*):

```
tab_poco <- table(to_clust$clust, to_clust$poco)
tab_poco
```

```
##
##      0  1
##   1  9  0
##   2  0 11
##   3  7  0
##   4  9  0
```

```
chisq.test(tab_poco)
```

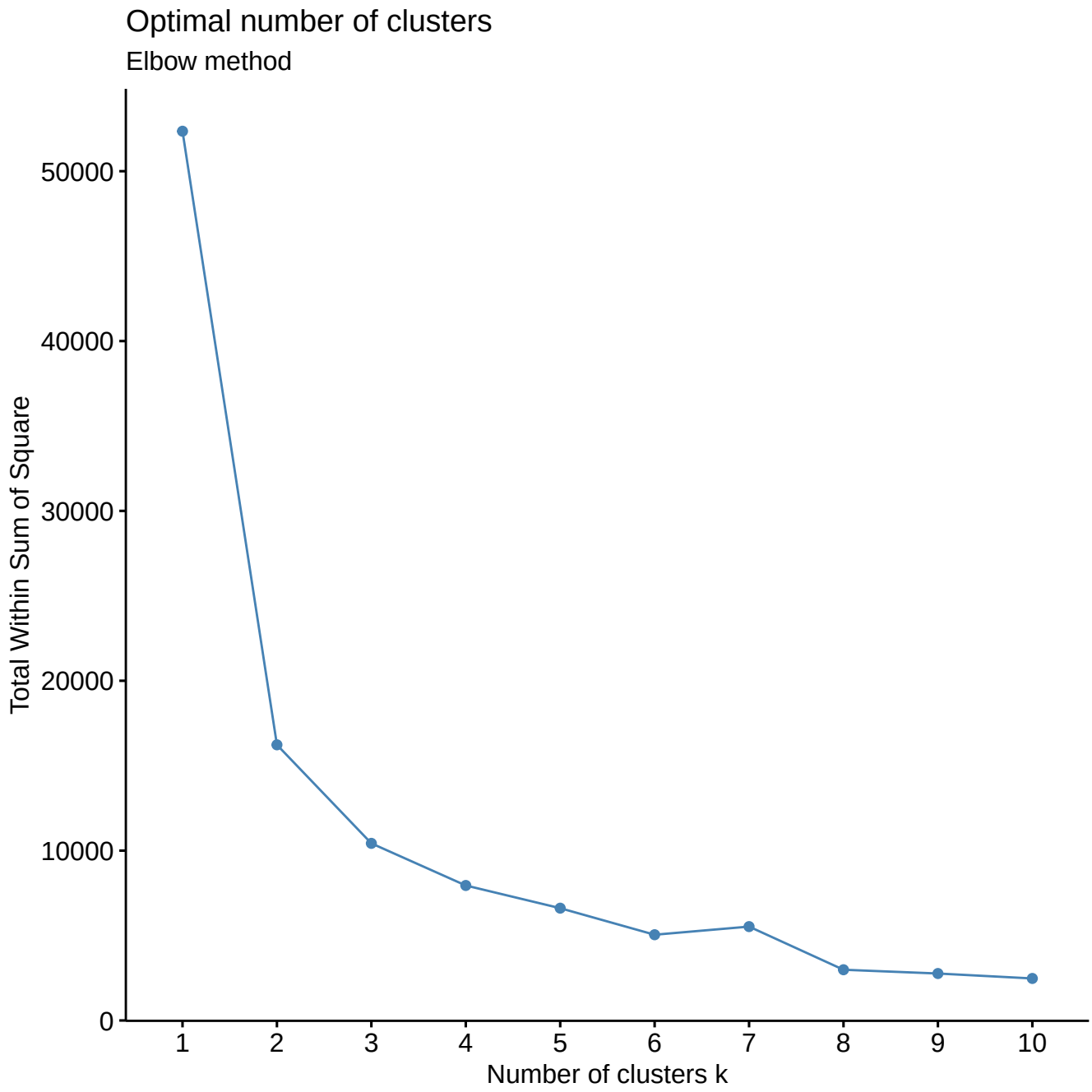
```
##
##  Pearson's Chi-squared test
##
## data:  tab_poco
## X-squared = 36, df = 3, p-value = 7.488e-08
```

Очевидно, поскольку все страны из бывшего советского блока попали в один кластер, p-value для данного теста очень маленький — **нулевая гипотеза отвергается на любом уровне доверия.**

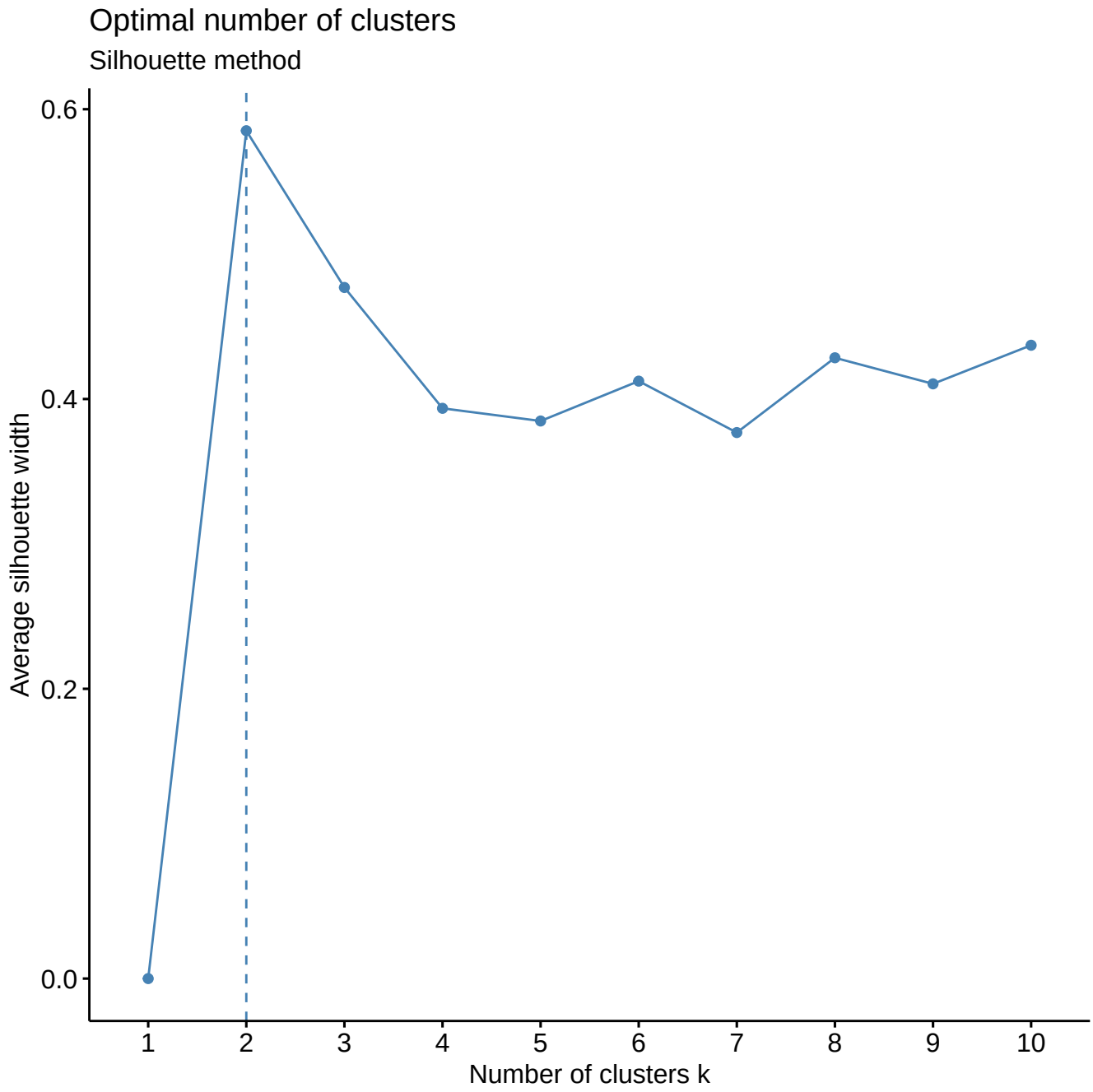
Задача 5. Уточнение числа кластеров

Проверим, используя **метод согнутого локтя и силуэтный метод**, какое число кластеров нужно выбрать, исходя из статистических соображений и *сравним* его с числом, выбранным нами.

```
library(factoextra)
fviz_nbclust(to_clust[1:5], kmeans, method = "wss") +
  labs(subtitle = "Elbow method")
```




```
fviz_nbclust(to_clust[1:5], kmeans, method = "silhouette") +
  labs(subtitle = "Silhouette method")
```

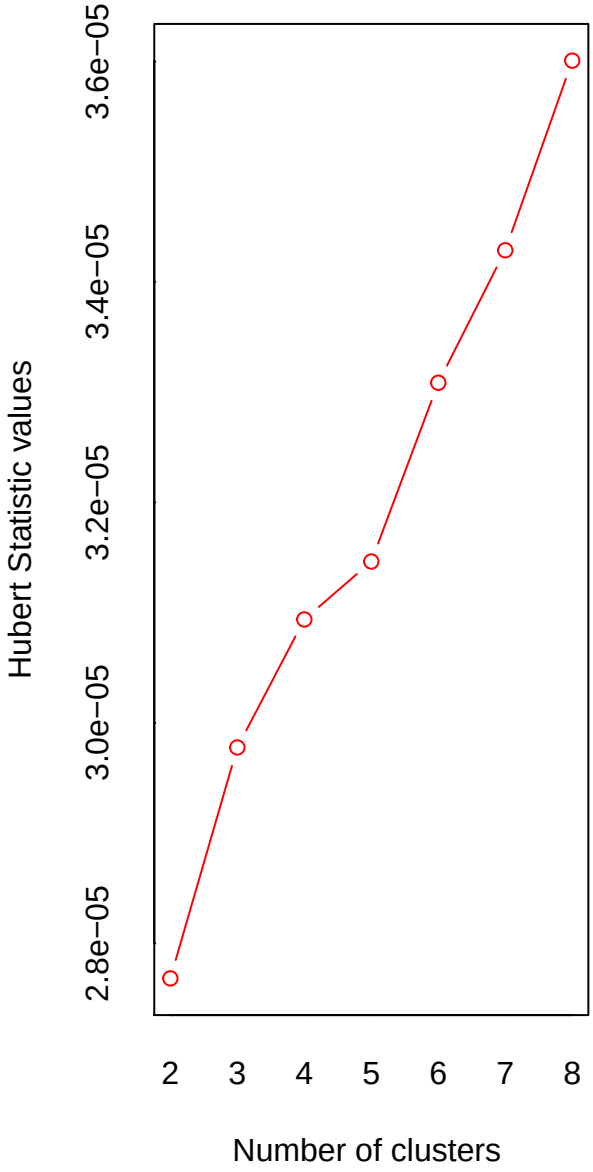


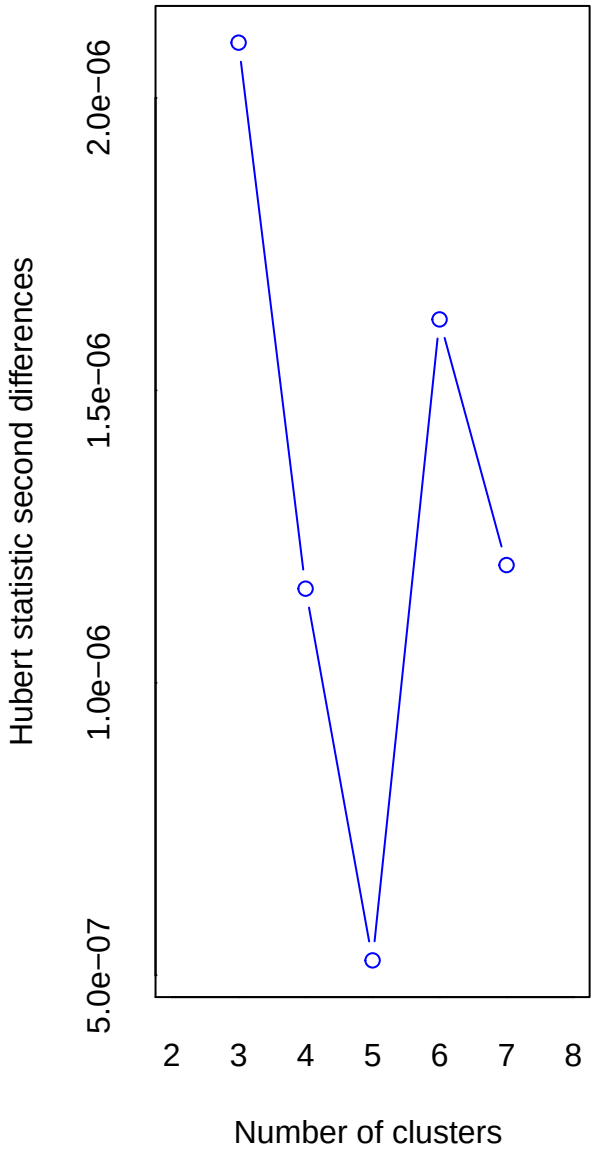
Итак, можно отметить, что:

- **Метод согнутого локтя** подсказывает, что нам необходимо взять от 3 до 4 кластеров — именно после этих значений внутригрупповой разброс начинает уменьшаться незначительно.
- Согласно **силуэтному методу**, максимальный зазор („силуэт“) достигается, когда мы берем 2 кластера (так происходят чаще всего для данного метода). Кроме того, 4 кластера является локальным пиком, после которого идет незначительное уменьшение, что свидетельствует в пользу этого выбора.

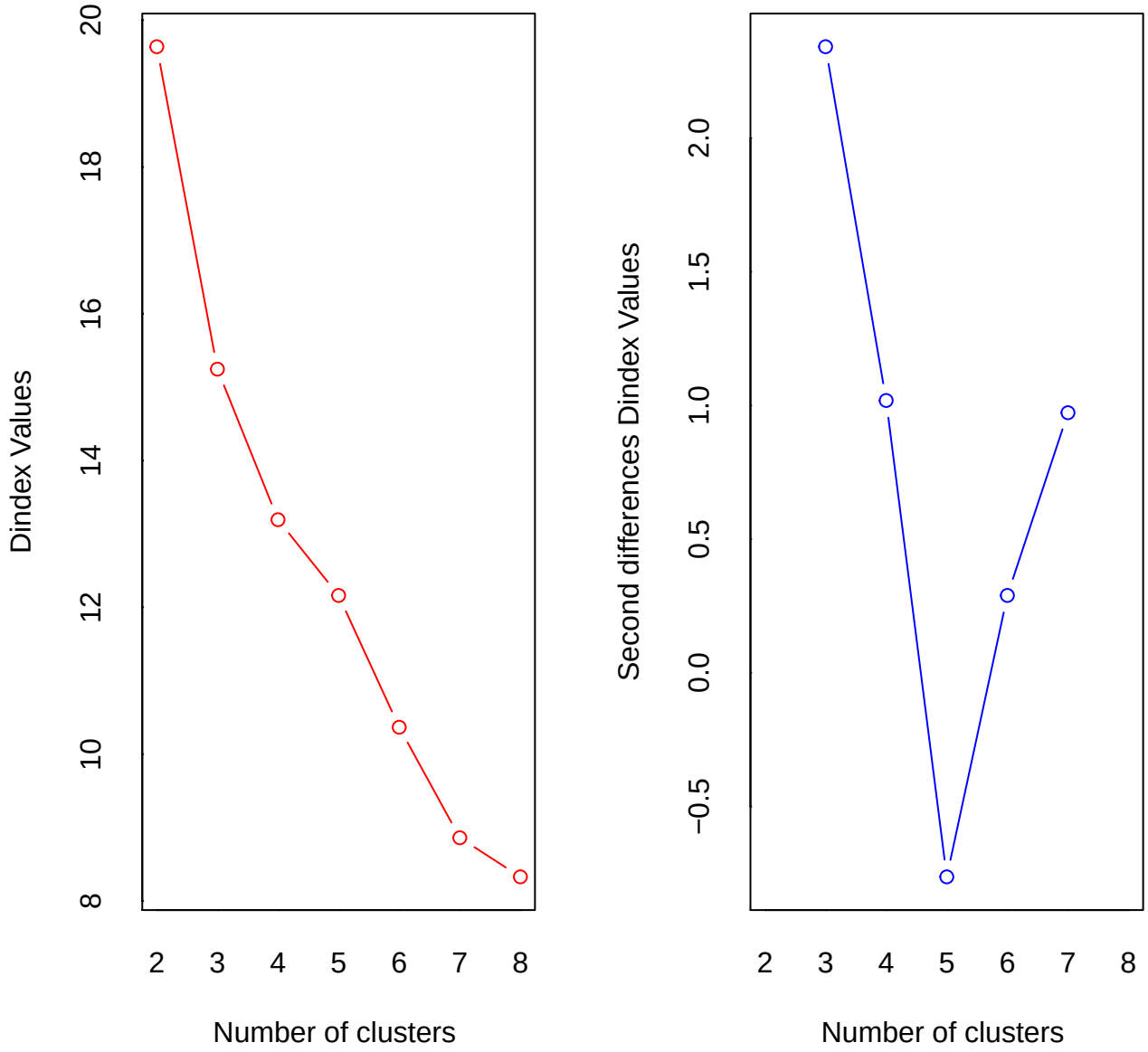
Также хотелось бы представить менее классический (и не очень устойчивый) способ выявления числа кластеров из библиотеки `NbClust`.

```
library(NbClust)
res <- NbClust(to_clust[1:5], min.nc = 2, max.nc = 8,
               method = "kmeans")
```






```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to
##                 significant increase of the value of the measure i.e the significant peak i
##                 index second differences plot.
##
```

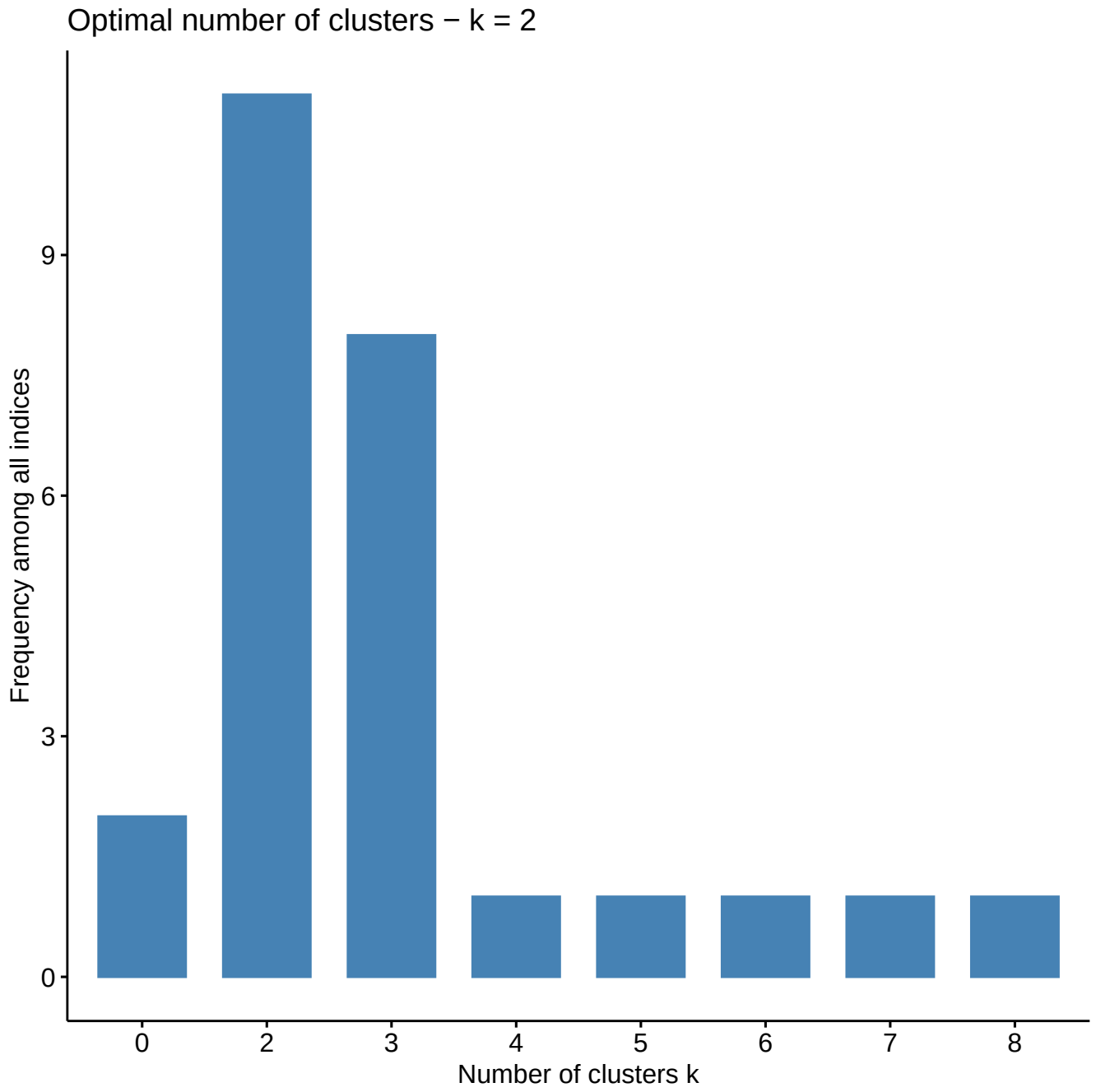



```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in
##                 second differences plot) that corresponds to a significant increase of the
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 11 proposed 2 as the best number of clusters
## * 8 proposed 3 as the best number of clusters
## * 1 proposed 4 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 1 proposed 8 as the best number of clusters
##
```

```
##                   ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```

```
fviz_nbclust(res)
```

```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 11 proposed  2 as the best number of clusters
## * 8 proposed  3 as the best number of clusters
## * 1 proposed  4 as the best number of clusters
## * 1 proposed  5 as the best number of clusters
## * 1 proposed  6 as the best number of clusters
## * 1 proposed  7 as the best number of clusters
## * 1 proposed  8 as the best number of clusters
##
## Conclusion
## ========================
## * According to the majority rule, the best number of clusters is  2 .
```



Разные алгоритмы, зашитые в данный инструмент, свидетельствуют, что **выбор происходил между 2 и 3 кластерами**, но хотелось бы отметить, что выбор как 2, так и 3 кластеров привел бы к затруднению содержательной интерпретации самих кластеров, поскольку он бы соединил уже слишком разные страны по выбранным переменным интереса.

Итак, можно сделать вывод о том, что **выбранное число кластеров соответствует методу согнутого локтя и силуэтному методу** и не соответствует алгоритму `NbClust`, *результаты которого выглядят странно.*

Задача 6. K-means

Реализуем кластерный анализ методом k-средних с выбранным окончательным числом кластеров (*4*) и **сохраним метки кластеров**, полученные в результате процедуры k-means, в датафрейм `to_clust`, а также проведем окончательную **содержательную интерпретацию** по каждому кластеру. Кроме того, выведем описательные статистики по ним.

```
to_clust$color <- NULL
kclust <- kmeans(to_clust[1:5], 4)
```

```
to_clust$k <- factor(kclust$cluster)

to_clust %>% group_by(clust) %>% summarise_at(vars(poco:green1),
                                              .funs = c(mean))

## # A tibble: 4 x 6
##   clust  poco gov_type vturn gov_left1 green1
##   <fct> <dbl>    <dbl> <dbl>     <dbl>  <dbl>
## 1 1         0     3.22  74.8     11.4   9.38
## 2 2         1     3     56.4     27.3   0.827
## 3 3         0     3.57  77.3     82.2   1.06
## 4 4         0     2.56  63.7      9.74  0.656

to_clust %>% group_by(k) %>% summarise_at(vars(poco:green1),
                                          .funs = c(mean))

## # A tibble: 4 x 6
##   k      poco gov_type vturn gov_left1 green1
##   <fct> <dbl>    <dbl> <dbl>     <dbl>  <dbl>
## 1 1     0.333     3     66.6    39.1     4.07
## 2 2     0.417     2.5   58.7     0.898   1.17
## 3 3     0.25      3.5   68.4    86.3     0.8
## 4 4     0.143     3.57  79.7     1.74    7.11

to_clust %>% group_by(clust) %>% summarise_at(vars(poco:green1),
                                                .funs = c(median))

## # A tibble: 4 x 6
##   clust  poco gov_type vturn gov_left1 green1
##   <fct> <int>    <int> <dbl>     <dbl>  <dbl>
## 1 1         0        3  75.6       0      9.1
## 2 2         1        2  54.6      10.8    0
## 3 3         0        4  84.6      79.6    0
## 4 4         0        3  65.1       0      0

to_clust %>% group_by(k) %>% summarise_at(vars(poco:green1),
                                          .funs = c(median))

## # A tibble: 4 x 6
##   k      poco gov_type vturn gov_left1 green1
##   <fct> <dbl>    <dbl> <dbl>     <dbl>  <dbl>
## 1 1         0      3    60.8      36.2    0
## 2 2         0      2.5  58.2       0      0
## 3 3         0      4    69.6      87.1    0
## 4 4         0      3    78.2       0      7.1

cluster11 <- to_clust %>% filter(k == 1)
cluster11

##                poco gov_type vturn gov_left1 green1 clust k
## Czech Republic    1        5  60.8  33.33000    0.0     2 1
## Denmark           0        4  84.6  51.51000    3.0     3 1
```

```
## Finland           0        3  72.1  36.16000   11.5     1 1
## Germany           0        2  76.2  37.50000    8.9     1 1
## Greece            0        1  57.8  48.23000    0.0     4 1
## Iceland           0        2  81.2  27.27273    0.0     4 1
## Slovakia          1        2  59.8  60.00000    0.0     2 1
## Slovenia          1        5  52.6  29.41000    0.0     2 1
## Switzerland       0        3  54.1  28.57143   13.2     1 1
```

```
cluster12 <- to_clust %>% filter(k == 2)
cluster12
```

```
##                poco gov_type vturn gov_left1 green1 clust k
## Bulgaria          1        2  54.1      0.00    0.0     2 2
## Canada            0        1  67.0      0.00    6.5     1 2
## Croatia           1        5  52.6      0.00    0.0     2 2
## Cyprus            0        1  66.7      0.00    4.8     1 2
## Estonia           1        2  63.7     10.78    0.0     2 2
## France            0        3  48.7      0.00    0.0     4 2
## Ireland           0        4  65.1      0.00    2.7     4 2
## Japan             0        3  53.7      0.00    0.0     4 2
## Latvia            1        3  54.6      0.00    0.0     2 2
## Poland            1        1  61.7      0.00    0.0     2 2
## United Kingdom    0        4  67.5      0.00    0.0     4 2
## USA               0        1  48.5      0.00    0.0     4 2
```

```
cluster13 <- to_clust %>% filter(k == 3)
cluster13
```

```
##             poco gov_type vturn gov_left1 green1 clust k
## Lithuania      1        5 50.60     94.63    2.0     2 3
## Luxembourg     0        2 89.66     64.71    0.0     3 3
## Malta          0        1 92.10    100.00    0.0     3 3
## New Zealand    0        5 72.90     79.49    0.0     3 3
## Portugal       0        4 48.60     79.62    0.0     3 3
## Romania        1        2 39.80     71.96    0.0     2 3
## Spain          0        4 66.20    100.00    0.0     3 3
## Sweden         0        5 87.20    100.00    4.4     3 3
```

```
cluster14 <- to_clust %>% filter(k == 4)
cluster14
```

```
##             poco gov_type vturn gov_left1 green1 clust k
## Australia      0        5  91.8      0.00   10.4     1 4
## Austria        0        6  75.6      0.00   13.9     1 4
## Belgium        0        6  88.4      0.00    6.1     1 4
## Hungary        1        1  69.7      0.00    7.1     2 4
## Italy          0        2  72.9     12.18    0.0     4 4
## Netherlands    0        2  81.6      0.00    9.1     1 4
## Norway         0        3  78.2      0.00    3.2     4 4
```

Несмотря на то, что медиана для дамми-переменной, отвечающей за посткоммунистические государства, равна 1 только в одном кластере от k-means, по средним значениям и самим выдачам

кластеров видно, что **теперь данные страны не сгруппированы в одном кластере** — по моему мнению, это ***упущение для анализа и интерпретации***.

Также можно отметить, что в новой кластеризации также выделяется кластер стран со средней явкой и низкими значениями доли левых и зеленых партий. При этом, явного кластера, в котором были бы страны с высоким значением доли зеленых партий **нет**: они также рассредоточены по всем кластерам, как и постукоммунистические страны. Все так же есть кластер стран с высокой (еще большей) долей левых партий, без зеленых партий и очень (еще большей) высокой явкой на выборах. При этом у двух кластеров: с левыми и зелеными партиями и без них уже примерно одинаковая явка на выборах.

Тем не менее, по моему мнению, из-за применения метода k-means **уровень интерпретации кластеров упал**, по сравнению с методом Уорда из-за того, что теперь постукоммунистические страны и страны с большой долей зеленых партий рассредоточены по всем кластерам, а не собраны в отдельных кластерах. В том числе из-за этого упала интерпретируемость доли явки на выборах, которая „рассосредоточилась“. Кроме того, кластеры получились гораздо менее сбалансированными по объему: 8, 4, 19 и 5 вместо 9, 11, 7 и 9, соответственно. Структура правительства все так же осталась слабоинтерпретируемой.

Интерпретация.
Итак, мы можем выделить 4 следующих кластера после кластеризации методом k-means (**примечание**: очередность кластеров может быть нарушена из-за последующей повторной компиляции перед выгрузкой):

1. В первом кластере находятся страны с относительно высокими долями левых и зеленых (не у всех наблюдений) партий в правительстве (2 место по этим показателям), а также со средней явкой на выборах.

2. Во втором кластере находятся страны с низким уровнем левых партий и (не всегда) зеленых партий в правительстве, а также со средним уровнем явки на выборах (который сильно варьируется).

3. В третьем кластере находятся страны с преимущественно очень высоким уровнем левых партий и без зеленых партий в правительстве, а также с высокой явкой на выборах.

4. В четвертом кластере находятся страны с очень высокой долей зеленых партий, но низкой долей левых партий в правительстве, а также с высокой явкой на выборах.

Относительно типа правительства и посткоммунистических стран никаких выводов сделать нельзя: они равномерно сосредоточены во всех кластерах.

Также посмотрим на соответствие двух методов.

```
library(fossil)

rand.index(as.integer(to_clust$clust),
           as.integer(to_clust$k))

## [1] 0.6873016
```

Можно заметить, что доля совпадений по итогам реализации двух методов составляет примерно 69%, что довольно много для кластерного анализа. Но в данной ситуации я бы отдал предпочтение методу Уорда из-за явного выделения постсоветских стран в отдельный кластер с низкой явкой на выборах.