Exercice 1

R course

Master 2 Statistics and Econometrics

Summary

We are interested in the results of the first round of the French presidential election that took place in 2022.

We will work with the raw data produced by the French Ministry of the Interior.

The objective is to see what is the influence of a voting system on the results of an election.

Notation

You will have to return the exercice in *.pdf* or *.html* format, which would have been done with **R** Markdown if possible. It should contain the lines of code used to answer the questions, but you should also explain what you are doing.

Remark: try to use the least amount of command lines per question asked. For the first 5 questions, it is *a priori* possible to deal with only one command line per question. If you use several lines, it does not really matter, but the idea of this exercice is also to make you look for and find the most "elegant" and simple solutions to answer a given question.

The election

The raw data are given at:

https://www.data.gouv.fr/fr/datasets/election-presidentielle-des-10-et-24-avril-2022-resultats-definitifs-du-1er-tour/#/resources/79b5cac4-4957-486b-bbda-322d80868224

First and last names of the 11 candidates at the presidential election in 2007 are ranked according to their appearance at the electoral boards :

- Madame Nathalie ARTHAUD
- Monsieur Fabien ROUSSEL
- Monsieur Emmanuel MACRON
- Monsieur Jean LASSALLE
- Madame Marine LE PEN
- Monsieur Eric ZEMMOUR
- Monsieur Jean-Luc MELENCHON
- Madame Anne HIDALGO
- Monsieur Yannick JADOT
- Madame Valerie PECRESSE
- Monsieur Philippe POUTOU
- Monsieur Nicolas DUPONT-AIGNAN

In the file, a row crresponds to the results of the election in one vote place.

For each of the 69682 vote places, we observe 105 variables:

- Code departement

- Name departement
- Code circonscription
- Name circonscription
- Code commune
- Name commune
- Code vote place
- Number of people registered (i.e. number of people who can vote in the vote place)
- Number of abstention (i.e. number of people who were registered but who did not vote)
- % Abs/Reg (ratio "abstention" on "registered")
- Voters (i.e. number of people who did vote)
- % Vot/Ins (ratio "Voters" on "registered")
- Number of white vote (a white vote is a voter who did not select anyone among the candidates)
- % white/Reg (ratio "white" on "registered")
- % white/Vot (ratio "white" on "Voters")
- Number of null votes (a null vote is a voter who did not respect the procedure of voting)
- % null/Reg (ratio "nuls" on "registered")
- % null/Vot (ratio "nuls" on "Voters")
- Number of valid votes (i.e. number of voters minus number of null votes minus number of white votes)
- % valid/Reg (ratio "valid" on "registered")
- % valid/Vot (ratio "valid" on "Voters")

Then, we have for each of the 11 candidates:

- number of bilboard
- Sexe of the candidate i, $i = 1, ..., 11$
- Last name of the candidate i
- First name of the candidate i
- Number of votes obtained by candidate i
- % (votes for i)/Reg (ratio "(votes for i)" on "registered")
- % (votes for i)/valid (ratio "(votes for i)" on "Voters")

Q1. Import the data

Import the data by using one of the method seen in the course.

Remark: the 1st line of the file is supposed to correspond to the names of the columns, but only the first 28 columns are filled in the header. To import the data, this information will have to be taken into account.

Solution 1 : we use *read.table()* function.

```
link <- "https://www.data.gouv.fr/fr/datasets/r/79b5cac4-4957-486b-bbda-322d80868224"
```

If we import the first lines, some information can be deduced about the options to use.

```
readLines(con = link, n = 2)
```

```
## [1] "Code du d\xe9partement;Libell\xe9 du d\xe9partement;Code de la circonscription;Libell\xe9 de la
## [2] "01;Ain;04;4\xe8me circonscription;001;L'Abergement-Cl\xe9menciat;0001;645;108;16,74;537;83,26;1
```

Note that some communes are using ' like L'Abergement-Clémenciat. We have to indicate that this character does not mean the end of a string, that is why we use option **quote = " "**. The encoding for French files is "latin-1" or "ISO_8859-1". It can be also sometimes "utf-8".

solution 2: use **data.table** package

Solution 3: with **readr** package, we first define the names of the variables and their types. To define a type, "c" means character, "n", numeric and "i", integer and we concatenate in a string all these values.

How to improve solution 1 ?

Q2. Select variables

- Keep only the following variables: 2, 4, 6, 7, 8, 26, 33, 40, 47, 54, 61, 68, 75, 82, 89, 96, 103
- Give the following names:
  - "departement",
  - "circonscription",
  - "commune",
  - "bureau_vote",
  - "inscrits",
  - "arthaud"
  - "roussel"
  - "macron"
  - "lassalle"
  - "le_pen"
  - "zemmour"
  - "melenchon"
  - "hidalgo"
  - "jadot"
  - "pecresse"
  - "poutou"
  - "dupont_aignan"

We use the tidyverse syntax

We use *names()*

Q3. Messy to tidy data

Transform the data so that one raw corresponds to the result of a candidate in a vote place. We use *pivot_longer()* which is very useful

Q4. Analysis of a first voting method

Among the 12 candidates, calculate the percentage of votes obtained by each candidate across the country. Who has the most votes?

We use the **dplyr** syntax

Compare your results with: https://www.interieur.gouv.fr/Elections/Les-resultats/Presidentielles/elecresult___presidentielle-2022/(path)/presidentielle-2022/FE.html

Q5. Analysis of a 2nd voting method

solution 1: Create a table where a row corresponds to a department and gives the number of registered as well as the name of the candidate who won the most votes within the department.

Solution 2: we use the messy data

- We will assume that we have 578 representants in France who are allocated to the departments of France. Propose a method of re-allocation of these 578 representants in these departments.

We allocate proportionnaly to the number of registered people:

We check that the sum is equal to 578:

- who is the candidate who has won the biggest number of departements?
- Assuming that the candidate who has won in a department wins all the representants of the department, which candidate won the largest number of representants?

Q6 Ranking

Finally, we will assign the following notes, by vote place:

- the score of 10 to the candidate who came first,
- the score of 5 to the candidate who is second,
- the score of 2 to the candidate arriving third.

Who is the candidate with the highest score on all vote places?

We first create a function which attributes the number of points according to the rankings:

Then we apply this function to each voting place:

We transpose the results:

Note that we do not have taken into account the ex-aquo. It could have be done by using *sampling()* when there are the same number of voters.