TP4 (FAS-1001)

Mohamed Chadli Bouzamondo

2024 – 25 – 03

1 – J'ai décidé de choisir la page Wikipédia concernant les sondages d'opinions par rapport aux différents partis présent au Royaume-Uni. Dû à l'impopularité grandissante du gouvernement conservateur de Rishi Sunak, actuel et de sa prédécesseure Liz Truss qui n'est restée qu'à peu près 1 mois de ses fonctions. (Webb et Louth, 2024) Avec la prochaine élection législative qui arrive en 2025, la question se pose : Quel parti est le plus enclin à avoir taux d'approbation le plus élevée et ce depuis le début de l'année 2024? Même s'il est trop tôt pour déterminer qui sera le vainqueur de l'élection à venir, il est intéressant d'observer les tendances actuelles. Elles permettent de savoir vers l'opinion publique se dirige de manière générale et vers quels partis ils sont plus enclins à supporter. Les données de sondage d'opinion de la page Wikipédia permettent justement d'observer les tendances actuelles de la population britanniques avec des résultats aussi récents que le 25 mars 2024. Même si la validité des sources de Wikipédia peuvent parfois être douteuse, il n'en reste pas moins un excellent outil d'information. J'ai décidé de scrapper les données du second tableau présent dans la page puisqu'il offre tous les éléments nécessaires pour répondre à ma question de recherche : les partis et leur taux d'approbations sur une période déterminée allant du début de l'année jusqu'à ces derniers jours. De plus, *scraper* les données de ce tableau seront plus simple à transmettre dans l'environnement R. Puisque les données sont déjà structurées d'une certaine manière, ce qui rendra le nettoyage et la sélection des variables plus simples.

```r
##Librairies
install.packages("rvest", repos = "http://cran.us.r-project.org")
```

```
le package 'rvest' a été décompressé et les sommes MD5 ont été
vérifiées avec succés

Les packages binaires téléchargés sont dans
    C:\Users\15150135\AppData\Local\Temp\RtmpgxBjyA\downloaded_packages
```

```r
library(rvest)
```

```
Warning: le package 'rvest' a été compilé avec la version R 4.3.3
```

```r
library(lubridate)
```

```
Warning: le package 'lubridate' a été compilé avec la version R 4.3.2


Attachement du package : 'lubridate'

Les objets suivants sont masqués depuis 'package:base':

    date, intersect, setdiff, union
```

```r
library(tidyverse)
```

```
Warning: le package 'tidyverse' a été compilé avec la version R 4.3.2

Warning: le package 'ggplot2' a été compilé avec la version R 4.3.3

Warning: le package 'tidyr' a été compilé avec la version R 4.3.2

Warning: le package 'readr' a été compilé avec la version R 4.3.2

Warning: le package 'purrr' a été compilé avec la version R 4.3.2

Warning: le package 'dplyr' a été compilé avec la version R 4.3.2

Warning: le package 'stringr' a été compilé avec la version R 4.3.2

Warning: le package 'forcats' a été compilé avec la version R 4.3.2

── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
✓ dplyr   1.1.4     ✓ readr   2.1.5
✓ forcats 1.0.0     ✓ stringr 1.5.1
✓ ggplot2 3.5.0     ✓ tibble  3.2.1
✓ purrr   1.0.2     ✓ tidyr   1.3.1

── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
✖ dplyr::filter()         masks stats::filter()
✖ readr::guess_encoding() masks rvest::guess_encoding()
```

```
✖ dplyr::lag()            masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force
all conflicts to become errors
```

```r
## Scraping ##
UK_2024_poll <-
read_html("https://en.wikipedia.org/wiki/Opinion_polling_for_the_next_U
nited_Kingdom_general_election") |>



## Sélection spécifique des tables du site web ##
html_elements("table") |>



## Chercher le deuxième table du site ##
pluck(2) |>



## Transformation en matrice de données ##
html_table(fill = T) |>
rename(Polling_firm = "Pollster",
Date = "Datesconducted",
Conservative = "Con",
Labour_party = "Lab",
Liberal_democrats = "Lib Dems",
Scottish_national_party = "SNP",
Green_Party = "Green",
Reform_UK = "Reform") |>



## Supression des colognes et des variables qui ne sont pas nécéssaires
##
slice(-c(1, 32, 49, 45)) |>

select(-c(Client, Area))

glimpse(UK_2024_poll)
```

```
Rows: 98
Columns: 11
$ Date                    <chr> "22–25 Mar", "24 Mar", "20–22 Mar",
"21–22 Mar…
$ Polling_firm            <chr> "Deltapoll", "Redfield & Wilton",
"Opinium", "…
$ Samplesize              <chr> "2,072", "2,000", "1,318", "1,270",
```

```
"1,632", "…
$ Conservative            <chr> "26%", "22%", "25%", "24%", "22%",
"25%", "19%…
$ Labour_party            <chr> "44%", "42%", "41%", "47%", "43%",
"43%", "44%…
$ Liberal_democrats       <chr> "9%", "12%", "10%", "10%", "10%",
"11%", "9%",…
$ Scottish_national_party <chr> "3%", "2%", "3%", "2%", "3%", "TBC",
"3%", "2%…
$ Green_Party             <chr> "6%", "6%", "8%", "6%", "6%", "5%",
"8%", "5%"…
$ Reform_UK               <chr> "11%", "14%", "11%", "11%", "13%",
"11%", "15%…
$ Others                  <chr> "2%", "2%", "2%", "2%", "3%", "TBC",
".mw-pars…
$ Lead                    <chr> "18", "20", "16", "23", "21", "18",
"25", "23"…
```

2 - Dans cette base de données que j'ai nommé « UK_2024_poll », Il y a 11 variables. La **variable « date »** montre à quelle date chaque sondage a été mené. Puisque j'ai scrappé le second tableau de la page, les dates vont du 2 janvier au 25 mars 2024. Certaines rangées couvrent plusieurs jours (de 2 à 3 jours habituellement) alors que d'autres ne couvrent qu'une journée spécifique. La **variable « Polling_firm »** représente les firmes de sondage qui ont effectué la recherche. La **variable « Samplesize »** est le nombre de personnes qui ont été sondé par chaque firme de sondage durant le laps de temps montré par la variable « date ». Les **variables « Conservatives, Labour_party, Liberal_democrats, Scottish_national_party, Green_Party, Reform_UK, Others »** sont des tous des parties politiques de la scène politique anglaises. La **variable « Others »** représente les partis qui n'ont pas eu assez d'importance/pas assez populaire auprès des répondants pour mériter leur propre catégorie. Les nombres exprimés pour chaque parti représente leur taux d'approbation auprès des répondants, selon une journée ou un groupe

de jours précis. La **variable « Lead »** représente la différence de pourcentage entre les deux partis les plus populaires auprès sur la période étudiée.

```
## Création d'une variable qui rassemble tout les partis politiques ##
Parties <- c("Conservative", "Labour_party", "Liberal_democrats",
"Scottish_national_party", "Green_Party", "Reform_UK")



## Convertion des colonnes de pourcentage en numérique ##
UK_2024_poll[Parties] <- lapply(UK_2024_poll[Parties], function(x)
as.numeric(gsub("%", "", x)))
```

```
Warning in FUN(X[[i]], ...): NAs introduits lors de la conversion
automatique
```

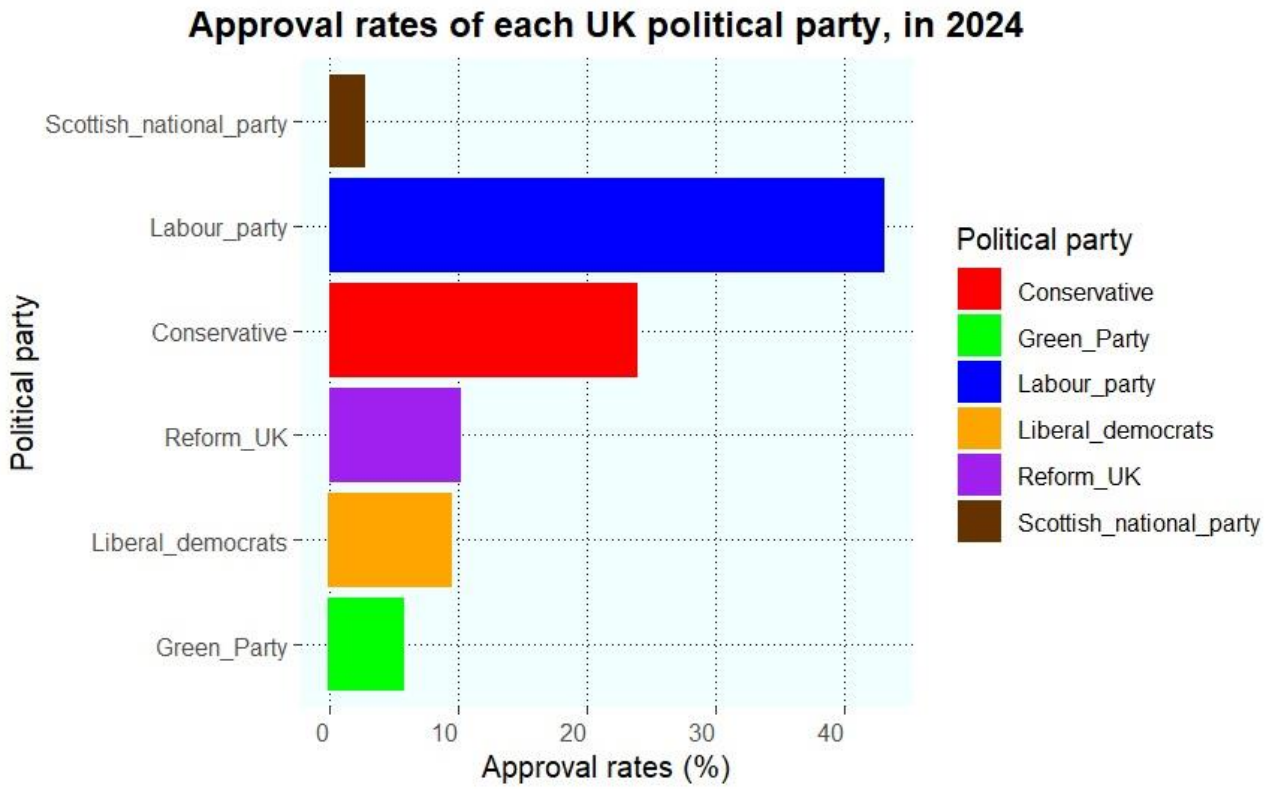
```
## Création d'une nouvelle base de donnée incluant les nouvelles
variables ##
UK_2024_poll_v2 <- pivot_longer(UK_2024_poll, cols = all_of(Parties),
names_to = "Political_parties", values_to = "Approval_rates")



## Création du graphique ##
ggplot(UK_2024_poll_v2, aes(x = reorder(Political_parties,
Approval_rates), y = Approval_rates/100, fill = Political_parties)) +
  geom_bar(stat = "identity") +
  labs(x = "Political party", y = "Approval rates (%)", fill =
"Political party") +
  scale_fill_manual(values = c("Conservative" = "red", "Labour_party" =
"blue", "Liberal_democrats" = "orange",
                               "Scottish_national_party" = "#663300",
"Green_Party" = "green", "Reform_UK" = "purple")) +
  theme(axis.text.x = element_text(hjust = 1)) + coord_flip() +
  ggtitle("Approval rates of each UK political party, in 2024") +
theme(plot.title = element_text(face = "bold", hjust = 0.5)) +
theme(panel.background = element_rect(fill = "azure2")) +
theme(panel.grid.major = element_line(color = "black", size = 0.5,
linetype = "dotted")) + theme(panel.grid.minor = element_blank())
```

```
Warning: The `size` argument of `element_line()` is deprecated as of
ggplot2 3.4.0.
ℹ Please use the `linewidth` argument instead.
```

```
Warning: Removed 2 rows containing missing values or values outside the
scale range
(`geom_bar()`).
```



3 - J'ai créé un graphique qui croise deux variables : le « Political party » sur mon axe x et les « Approval rates in (%) ». Depuis que j'ai utilisé coord_flip(), les deux variables apparaissent sur des axes opposés à ceux décrits précédemment. J'ai réalisé un graphique à bandes qui permettrait de voir plus facilement le résultat du croisement de ces deux variables. Pour commencer, j'ai dû créer une nouvelle base de données appelée « UK_2024_poll_v2 » pour pouvoir rassembler tous les partis politiques en une seule variable, tout comme les taux d'approbation. Étant donné que ces deux variables n'ont pas été créées dans l'ensemble de données précédent, il aurait été impossible de les croiser avec succès et d'en créer un graphique. Le diagramme à bandes montre que le Labour party est le parti le plus populaire au cours d'une période déterminée, de début janvier

jusqu'à fin mars environ, avec environ plus de 40 % de suffrages. Battant largement les autres partis. Le deuxième candidat le plus populaire était le Conservative party, avec un taux d'approbation légèrement inférieur à 25 %. Les Reform UK et les Liberal democrats semblent être proches en termes de taux d'approbation, le premier ne battant le second que de très peu. Alors que le Scottish national party affiche les taux d'approbation les plus bas, soit moins de 5 %, sur la même période. Seulement battu de quelques pourcentages par le Green Party, à environ 5 %. Quelque chose peut paraître bizarre dans ce graphique et c'est l'omission des partis « Others ». On pourrait supposer que leurs marges étant si faibles, ils n'ont pas pu accéder au graphique. Bien que cela puisse être une explication, je pense que cela a à voir avec un problème concernant la suppression du tableau du site Web. Sur Wikipédia, de nombreux pourcentages dans la variable « Others » avaient une note de bas de page, que R-Studio n'était pas en mesure d'interpréter comme des chiffres. Cela a probablement amené à les considérer comme NA et à ne pas les utiliser lors de la production de ce graphique, étant donné qu'il y a une quantité excessive de chiffres avec une note de bas de page dans cette colonne.

4 - Lorsqu’on réfléchit aux aspects éthiques du scraping, de nombreuses questions viennent à l’esprit. Pourquoi est-ce que je scrappe, comment vais-je utiliser les données? Ai-je l'autorisation d'utiliser ces données? De quel type de données s'agit-il, des données publiques ou liées à la vie privée? Je pense que toutes ces questions sont valables lorsqu'on travaille sur le scraping. De nombreuses considérations éthiques doivent être prises en compte lors du scraping de sites web. Être capable de réfléchir aux défis éthiques contribue à apporter une nouvelle perspective sur ce que nous faisons

réellement. Après tout, il n'y a pas de réglementation spécifique, n'importe qui a les bons outils et beaucoup de temps libre peut y arriver. Même avec les limitations imposées par les API et les « paywalls », il est toujours assez facile d'obtenir des informations sensibles. Lorsque j'ai collecté les données des sondages britanniques, la plupart de ces questions ne m'ont pas traversé l'esprit. Après tout, c'est une page Wikipédia, elle est publique donc ça doit vouloir dire que j'ai le droit d'utiliser ces données ? Bien que cela puisse être une opinion valable, je pense toujours qu'il est important de garder à l'esprit que le fait de pouvoir récupérer ces données sans être limité est un privilège et doit être traité comme tel. Bien sûr, je travaille sur des données de sondage, donc il n'y a pas de dilemme éthique énorme comme ce serait le cas si je scrappais, la page privée d'une personne sur un média social. L'argument pourrait être fait en faveur de cela, car je ne sais pas comment la personne qui a publié les données voudrait qu'elles soient utilisées. Mais même dans ce cas, il est toujours bon de garder à l'esprit nos intentions et l'intention de ceux qui rendent les données accessibles à tous. Peut-être qu'ils ne veulent pas que leurs données soient utilisées de telle ou telle manière. Il est difficile de se conformer aux besoins de tous ceux qui mettent des données sur Internet. Même s'il s'agit d'un lieu public avec beaucoup de liberté, cela ne veut pas dire que les considérations éthiques ne doivent pas venir à l'esprit.

Bibliographie

Webb, Paul David et Louth, Lord Norton of. 2024. « Conservative. », *Party Encyclopedia Britannica*, https://www.britannica.com/topic/Conservative-Party-political-party-United-Kingdom.

Wikipedia contributors. 25 mars 2024. « Opinion polling for the next United Kingdom general election. », *Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Opinion_polling_for_the_next_United_Kingdom_general_election&oldid=1215534365