GROUP PRESENTATION

Andrea Baños, Alejandra Costa, Eric Hausken, María Jurado

SPANISH ELECTIONS ANALYSIS

This project will analyze **electoral and survey data**, from 2008 to 2019, from the Spanish Congress of Deputies.

The following R packages will be in our report:

Libraries:

```
library(tidyverse)
library(lubridate)
library(glue)
library(dplyr)
library(ggplot2)
library(corrplot)
library(forcats)
library(patchwork)
library(ghibli)
```

1. Tidy data

1.1 Tidy data: Election data

We first prepare and clean election dataset by applying some transformations.

```
election_pivot <- election_data |>
  pivot_longer(
    cols = `BERDEAK-LOS VERDES` : `COALICIÓN POR MELILLA` ,
    names_to = "party",
    values_to = "votos"
  ) |>
  drop_na(votos) |>
  select(-c(vuelta, tipo_eleccion, codigo_distrito_electoral)) |>
  mutate(
      date_elec_ym = lubridate::ym(paste(anno, mes)),
      .before = anno
  ) |>
  mutate(
    codigo_ccaa = as_factor(codigo_ccaa),
    codigo_municipio = as_factor(codigo_municipio),
    codigo_provincia = as_factor(codigo_provincia),
    municipio = as_factor(paste(codigo_ccaa,
                                codigo_provincia, codigo_municipio,
                        sep = "-")),
    party = as_factor(party)
  ) |>
   select(-c(mes,anno, numero_mesas,participacion_1, participacion_2))
```

1.1 Tidy data: Election data

We also grouped the parties into the main classifications.

```
election_pivot <- election_pivot |>

  mutate(party = case_when(

    str_detect(party, "PODEMOS")|
    str_detect(party, "PODEM") |
    str_detect(party, "VERDES") |
    str_detect(party, "IZQUIERDA UNIDA") |
    str_detect(party, "ESQUERRA UNIDA") |
    str_detect(party, "EZKER BATUA") ~ "PODEMOS",

    str_detect(party, "SOCIALISTA") |
    str_detect(party, "SOCIALISTES")  ~ "PARTIDO SOCIALISTA OBRERO ESPAÑOL",

    str_detect(party, "PARTIDO POPULAR") ~ "PARTIDO POPULAR",

    str_detect(party, "CIUDADANÍA") |
    str_detect(party, "CIUDADANIA")  ~ "CIUDADANOS-PARTIDO DE LA CIUDADANIA",

    str_detect(party, "NACIONALISTA VASCO") ~ "EUZKO ALDERDI JELTZALEA-PARTIDO NACIONALISTA VASCO",

    str_detect(party, "NACIONALISTA GALEGO") ~ "BLOQUE NACIONALISTA GALEGO",

    str_detect(party, "MÉS COMPROMÍS") ~ "MÉS COMPROMÍS",
```

1.1 Tidy data: Election data

After cleaning the dataset and selecting just the relevant information, the election data look as follows:

```
  1  head(election_pivot)
```

```
# A tibble: 6 × 11
  date_elec_ym codigo_ccaa codigo_provincia codigo_municipio censo votos_blancos
  <date>       <fct>       <fct>            <fct>             <dbl>         <dbl>
1 2008-03-01   14          01               001                1838            23
2 2008-03-01   14          01               001                1838            23
3 2008-03-01   14          01               001                1838            23
4 2008-03-01   14          01               001                1838            23
5 2008-03-01   14          01               001                1838            23
6 2008-03-01   14          01               001                1838            23
# ℹ 5 more variables: votos_nulos <dbl>, votos_candidaturas <dbl>, party <chr>,
#   votos <dbl>, municipio <fct>
```

1.2 Tidy data: Abbreviation data

The abbreviations are unified in the `abbrev` table, and those corresponding to non-relevant parties are categorized as “OTHERS”

As a result, we get a reference table containing unique parties, with their corresponding abbreviation.

```
1 print(abbrev)
```

```
# A tibble: 13 × 2
   denominacion                                          siglas
   <chr>                                                 <chr>
 1 OTHERS                                                OTHERS
 2 EUZKO ALDERDI JELTZALEA-PARTIDO NACIONALISTA VASCO    EAJ-PNV
 3 PODEMOS                                               PODEMOS
 4 PARTIDO SOCIALISTA OBRERO ESPAÑOL                     PSOE
 5 PARTIDO POPULAR                                       PP
 6 ESQUERRA REPUBLICANA DE CATALUNYA                     ERC
 7 CONVERGENCIA I UNIO                                   CiU
 8 BLOQUE NACIONALISTA GALEGO                            BNG
 9 CIUDADANOS-PARTIDO DE LA CIUDADANIA                   CS
10 EUSKAL HERRIA BILDU                                   EH-BILDU
11 VOX                                                   VOX
12 MÉS COMPROMÍS                                         COMPROMIS
13 MÁS PAÍS                                              M PAÍS
```

1.2 Tidy data: Abbreviation data

A vector is created with the final abbreviations, which will be useful later when cleaning and preparing the survey data.

```
 [1] "OTHERS"    "EAJ-PNV"   "PODEMOS"   "PSOE"      "PP"        "ERC"
 [7] "CiU"       "BNG"       "CS"        "EH-BILDU"  "VOX"       "COMPROMIS"
[13] "M PAÍS"
```

Lastly, we join the abbreviation data with the election data. By doing so, we get the correponding abbreviations of the parties in the election dataset and get rid of the variable “Party”, containing the whole name.

```
1 election_pivot <- election_pivot |>
2   left_join(abbrev, by = c("party" = "denominacion"))
```

1.3 Tidy data: Survey data

We apply the following transformations to satisfy the specified conditions:

```
surveys <- surveys |>
  filter(year(date_elec) >= 2008 ) |>
  filter(exit_poll == FALSE) |>
  filter(size > 750) |>
  mutate(fieldwork_days = field_date_to - field_date_from) |>
  filter(fieldwork_days > 1)

surveys <- surveys |>
  mutate(date_elec_ym = ym(paste(lubridate::year(surveys$date_elec),
                  lubridate::month(surveys$date_elec))),
         .after = date_elec)
```

Some other transformations are applied, so that we can work with a more organised dataset:

```
surveys_pivot <- surveys |>
  pivot_longer(cols = UCD:EV,
               names_to = "party",
               values_to = "votes_percent") |>
  drop_na(votes_percent) ## removed rows with zero (NA) votes for that party
```

1.3 Tidy data: Survey data

Survey data is aligned with the rest of the datasets.

```
surveys_pivot<- surveys_pivot |>
  mutate(party = ifelse(
    party %in% vector_abbrev, party, "OTHER")) |>
    select(-c(type_survey,exit_poll, id_pollster, media))
```

2. Questions

- How is the vote of national parties (PSOE, PP, VOX, CS, MP, UP - IU) distributed against regional or nationalist parties?
- Which party was the winner in the municipalities with more than 100,000 habitants (census) in each of the elections?
- Which party was the second when the first was the PSOE? And when the first was the PP?
- Who benefits from low turnout?
- How to analyze the relationship between census and vote? Is it true that certain parties win in rural areas
- How to calibrate the error of the polls (remember that the polls are voting intentions at national level)?
- In which election were the polls most wrong?
- How were the polls wrong in national parties (PSOE, PP, VOX, CS, MP, UP - IU)?
- Which polling houses got it right the most and which ones deviated the most from the results?

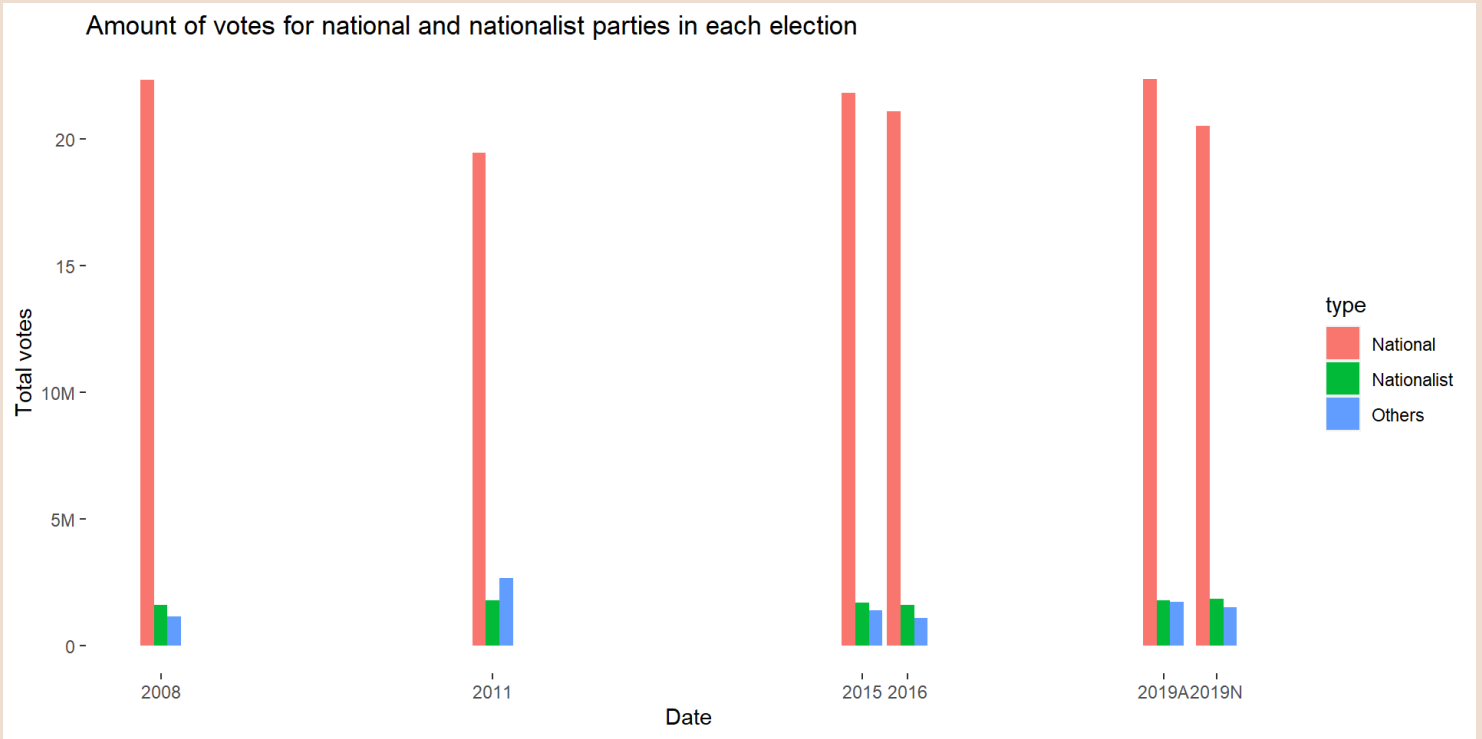
2.1 How is the vote of national parties distributed against regional or nationalist parties?

2.1 How is the vote of national parties distributed against regional or nationalist parties?

- First step is to group the parties into national and nationalists parties.
- NATIONAL: PSOE, PP, VOX, CS, Más País, Podemos
- NATIONALIST: PNV, Bloque Nacionalista Galego, Mès Compromís, Covergencia i Unió, Esquerra Republicana de Cataluña and Bildu.
- Total votes by type

```
data_new<-
  election_pivot |>
  group_by(type,date_elec_ym) |>
  summarise(total_votes = sum(votos, na.rm = TRUE)) |>
   ungroup()
```

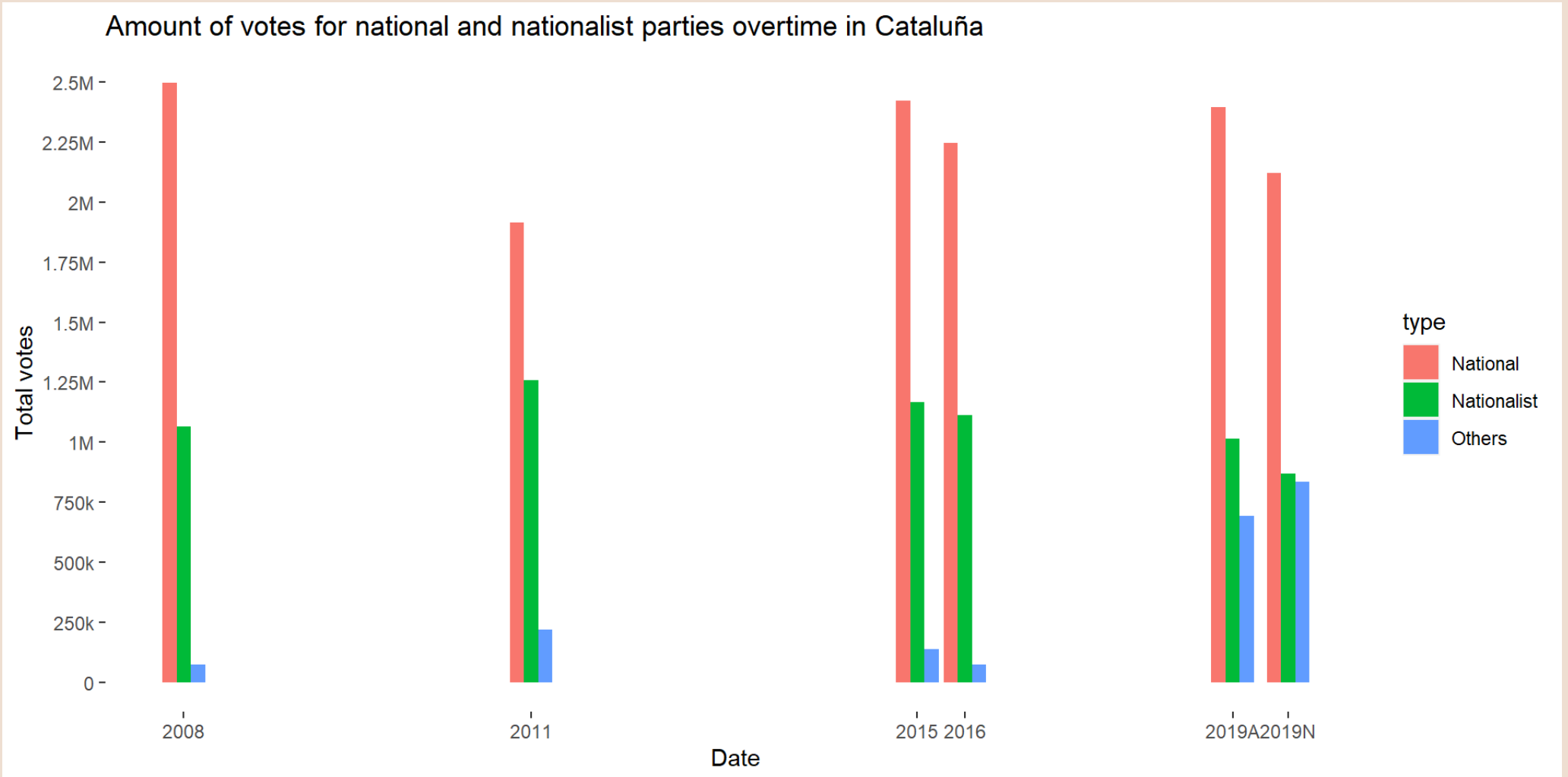
2.1 How is the vote of national parties distributed against regional or nationalist parties?



This graph is really informative.

- Total participation in each election.
- Evolution in votes for all 3 types of parties
- Compare among different elections.

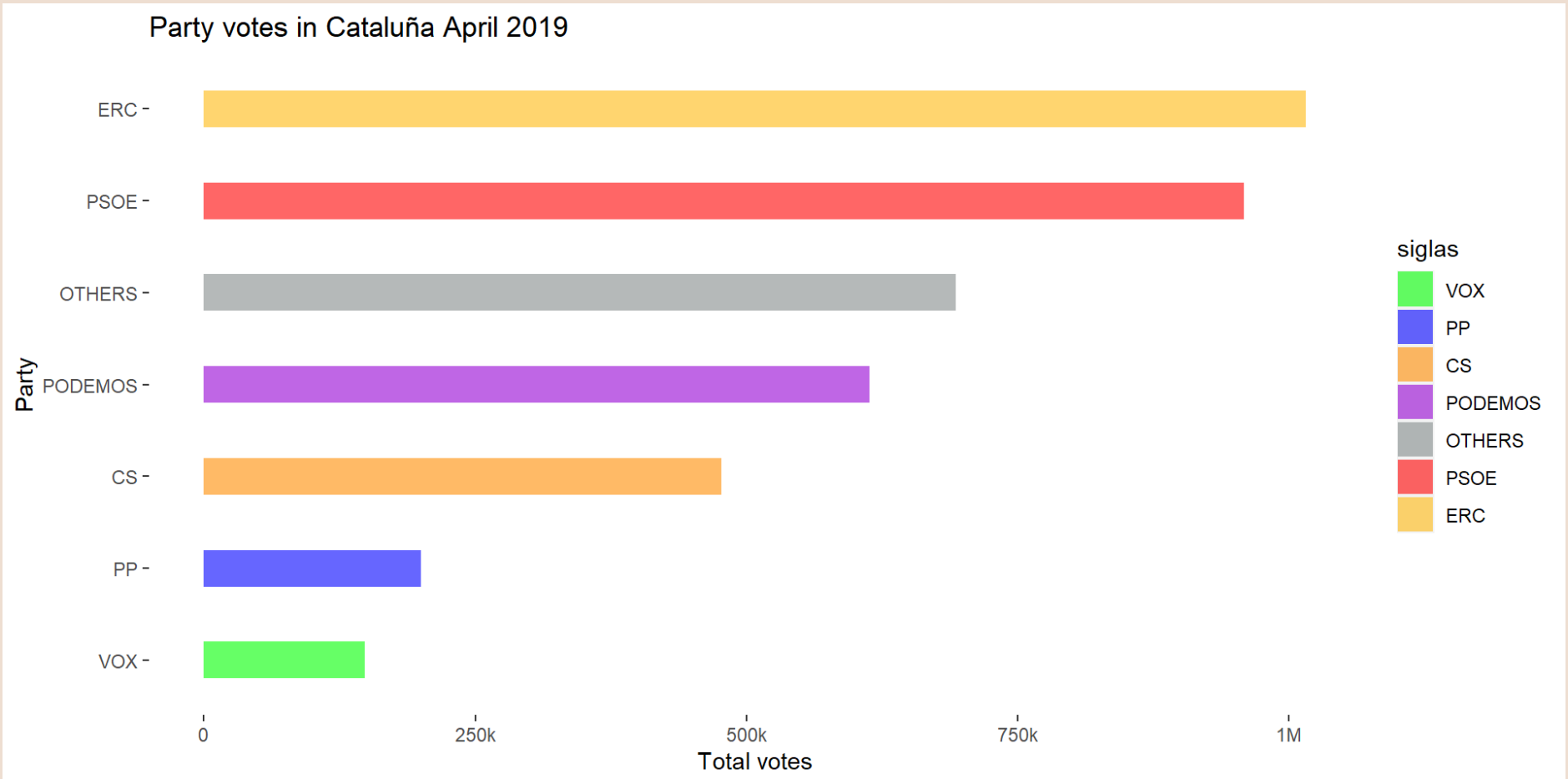
In Cataluña



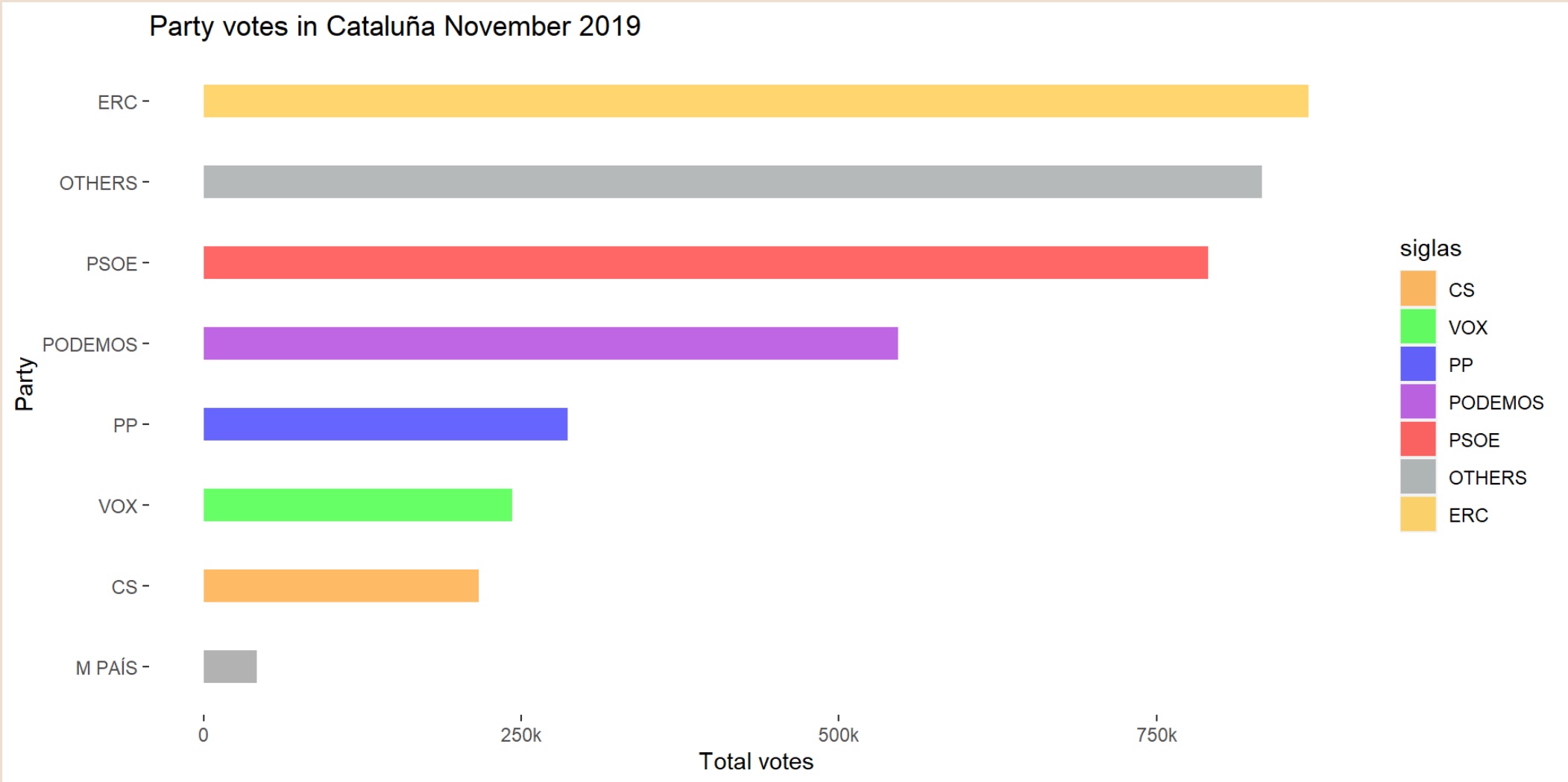
- Decrease in nationalism.
- Bias: Junts is not considered as nationalist –> could explain the increase in other parties.

Evolution of party votes in Cataluña in April 2019

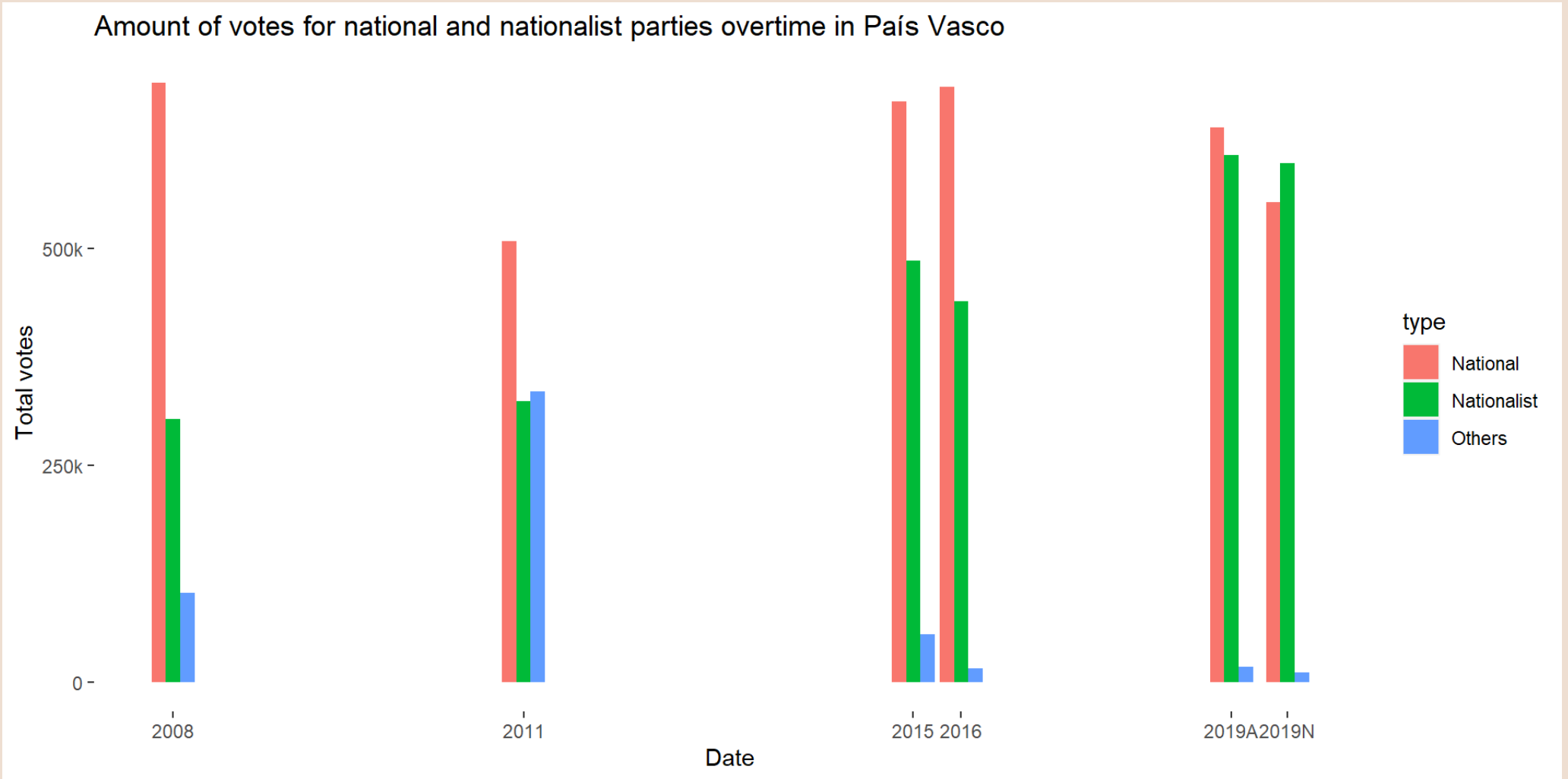
- Using library(forcats)
- Order from most voted to less voted party



Evolution of party votes in Cataluña in November 2019

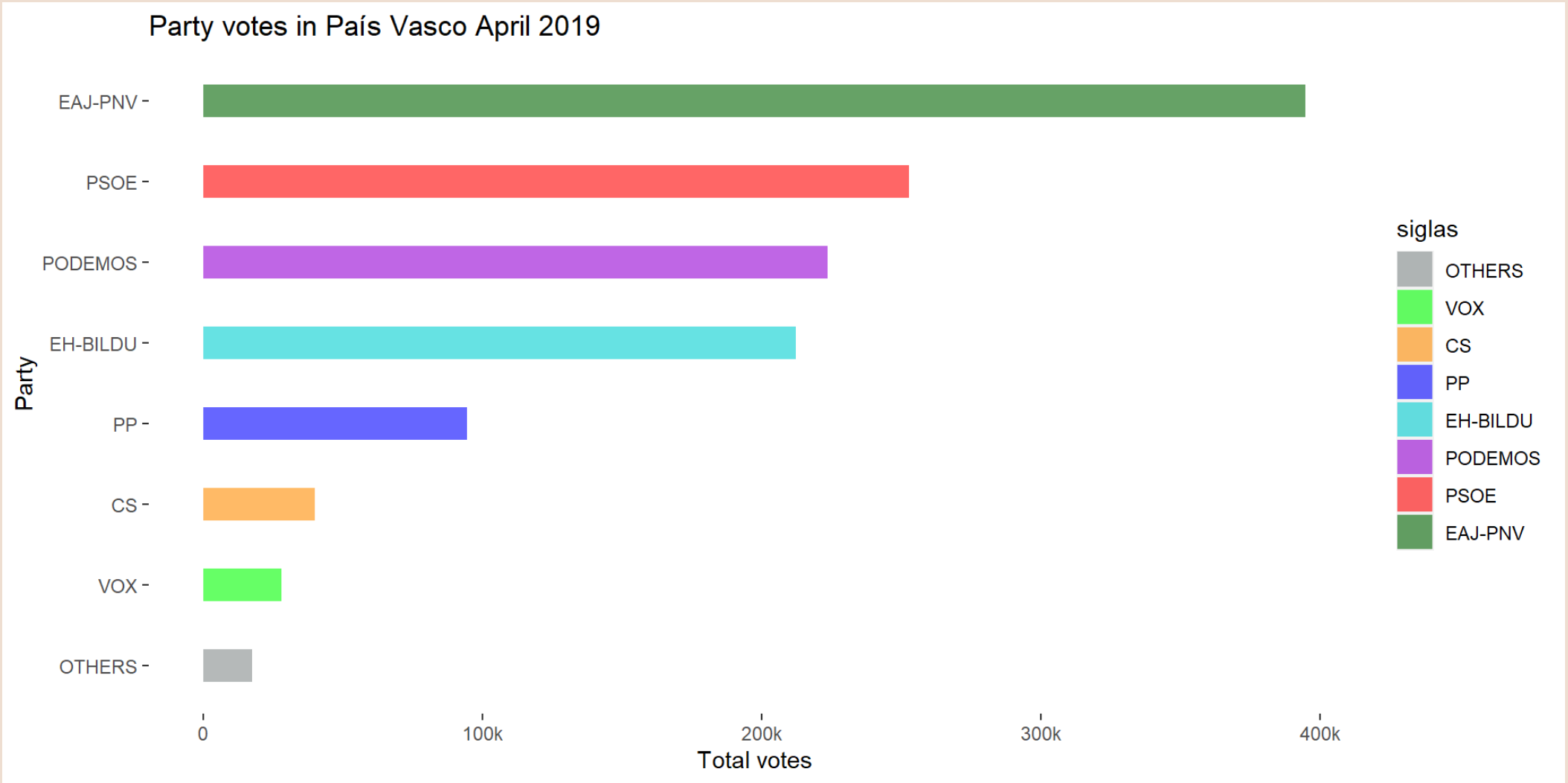


In País Vasco

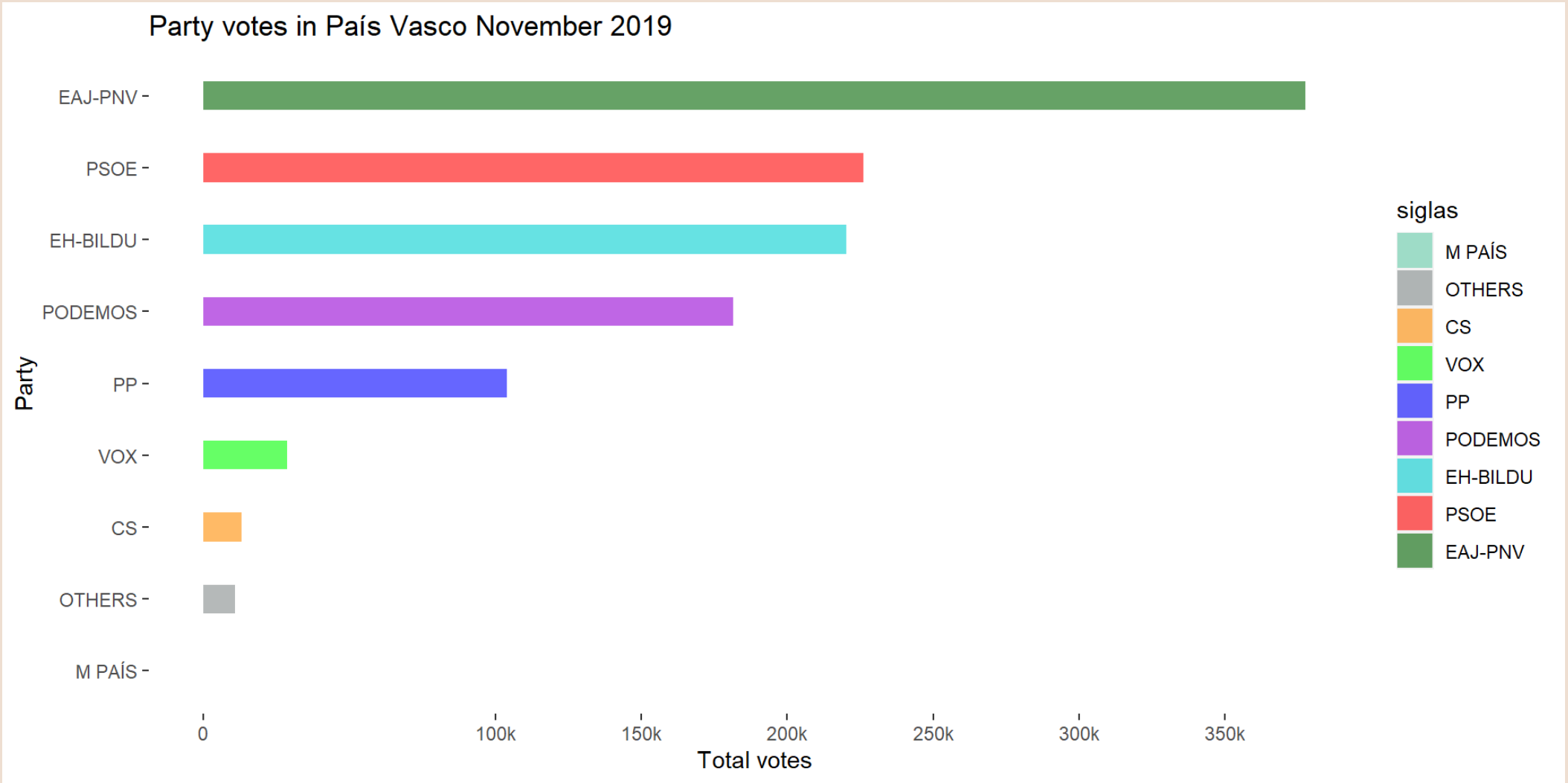


- Increase in nationalism overtime.
- Other parties have decreased significantly the amount of votes.

Party votes País Vasco April 2019



Party votes País Vasco November 2019



2.2 Which party was the winner in the municipalities with more than 100,000 habitants (census) in each of the elections?

2.2 Which party was the winner in the municipalities with more than 100,000 habitants (census) in each of the elections?

```r
winners <- election_pivot |>
  filter(censo>100000) |>
  group_by(codigo_municipio, date_elec_ym, party) |>
  summarise(total_votes = sum(votos, na.rm = TRUE)) |>
  slice(which.max(total_votes)) |>
  ungroup()

winners
```

```
# A tibble: 231 × 4
   codigo_municipio date_elec_ym party                              total_votes
   <fct>            <date>       <chr>                                    <dbl>
 1 003              2008-03-01   PARTIDO POPULAR                          49909
 2 003              2011-11-01   PARTIDO POPULAR                          55858
 3 003              2015-12-01   PARTIDO POPULAR                          36149
 4 003              2016-06-01   PARTIDO POPULAR                          38470
 5 003              2019-04-01   PARTIDO SOCIALISTA OBRERO ESPAÑOL        28729
 6 003              2019-11-01   PARTIDO SOCIALISTA OBRERO ESPAÑOL        27074
 7 013              2008-03-01   PARTIDO POPULAR                          49463
 8 013              2011-11-01   PARTIDO POPULAR                          53152
 9 013              2015-12-01   PARTIDO POPULAR                          34111
10 013              2016-06-01   PARTIDO POPULAR                          38809
# ℹ 221 more rows
```

2.3 Which party was the second when the first was the PSOE? And when the first was the PP?

2.3 Which party was the second when the first was the PSOE? And when the first was the PP?

```r
winners2 <- election_pivot |>
  group_by(date_elec_ym, party) |>
  summarise(total_votes = sum(votos, na.rm = TRUE)) |>
  slice_max(total_votes, n = 2) |>
  ungroup()

winners2
```

```
# A tibble: 12 × 3
   date_elec_ym party                              total_votes
   <date>       <chr>                                    <dbl>
 1 2008-03-01   PARTIDO SOCIALISTA OBRERO ESPAÑOL    11078605
 2 2008-03-01   PARTIDO POPULAR                      10171828
 3 2011-11-01   PARTIDO POPULAR                      10838951
 4 2011-11-01   PARTIDO SOCIALISTA OBRERO ESPAÑOL     6987723
 5 2015-12-01   PARTIDO POPULAR                       7114123
 6 2015-12-01   PODEMOS                               5640709
 7 2016-06-01   PARTIDO POPULAR                       7800328
 8 2016-06-01   PARTIDO SOCIALISTA OBRERO ESPAÑOL     5424130
 9 2019-04-01   PARTIDO SOCIALISTA OBRERO ESPAÑOL     7481667
10 2019-04-01   PARTIDO POPULAR                       4356714
11 2019-11-01   PARTIDO SOCIALISTA OBRERO ESPAÑOL     6752314
12 2019-11-01   PARTIDO POPULAR                       5021622
```

2.4 Who benefits from low turnout?

2.4 Who benefits from low turnout?

In order to calculate the low turnout in the surveys, we divided the turnout by its mean to generate a ratio. If this ratio exceeds 1, it indicates high turnout, and if it is less than 1, it suggests low turnout.

```
summary(surveys_pivot$turnout)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  59.40   66.50   70.00   69.84   72.70   79.90    8204
```

```
summary(surveys_pivot$votes_percent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.30    4.10   10.49   18.60   49.30
```

```
surveys_pivot$low_turnout <- surveys_pivot$turnout/68.38
```

2.4 Who benefits from low turnout?

Therefore, a data frame named `low_turnout` has been created, which is defined as turnout below the mean. The parties that benefit from it are those with more votes than the mean within the low turnout.

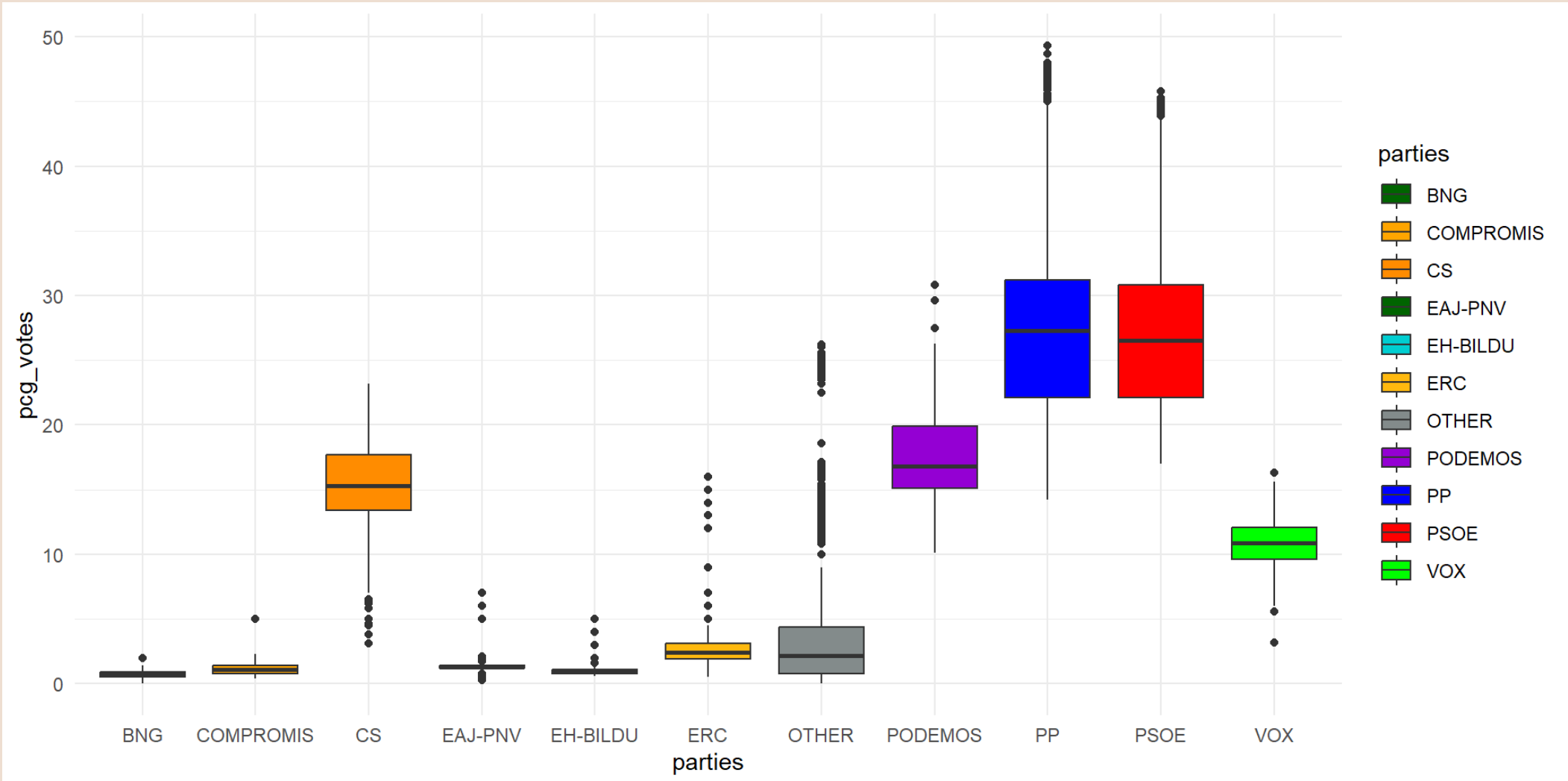
```
low_turnout <- surveys_pivot |> filter(low_turnout < 1 & votes_percent > 8.939)

low_turnout |>
  distinct(party)
```

```
# A tibble: 7 × 1
  party
  <chr>
1 PSOE
2 PP
3 CS
4 OTHER
5 PODEMOS
6 VOX
7 ERC
```

These are the parties who benefited from low turnout, because they have more votes than expected.

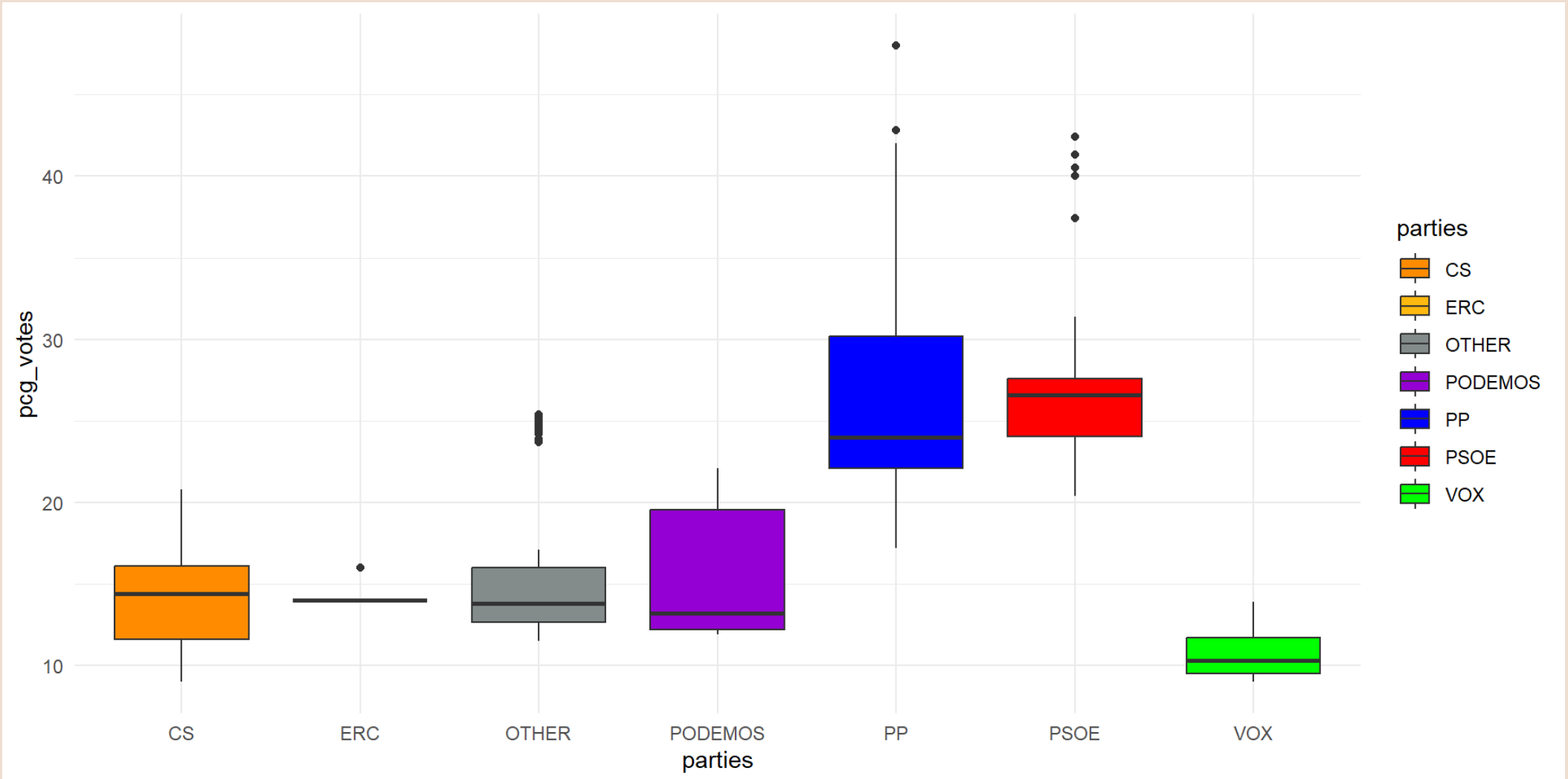
2.4 Who benefits from low turnout?

To visualize the results more clearly, boxplots have been created.



2.4 Who benefits from low turnout?

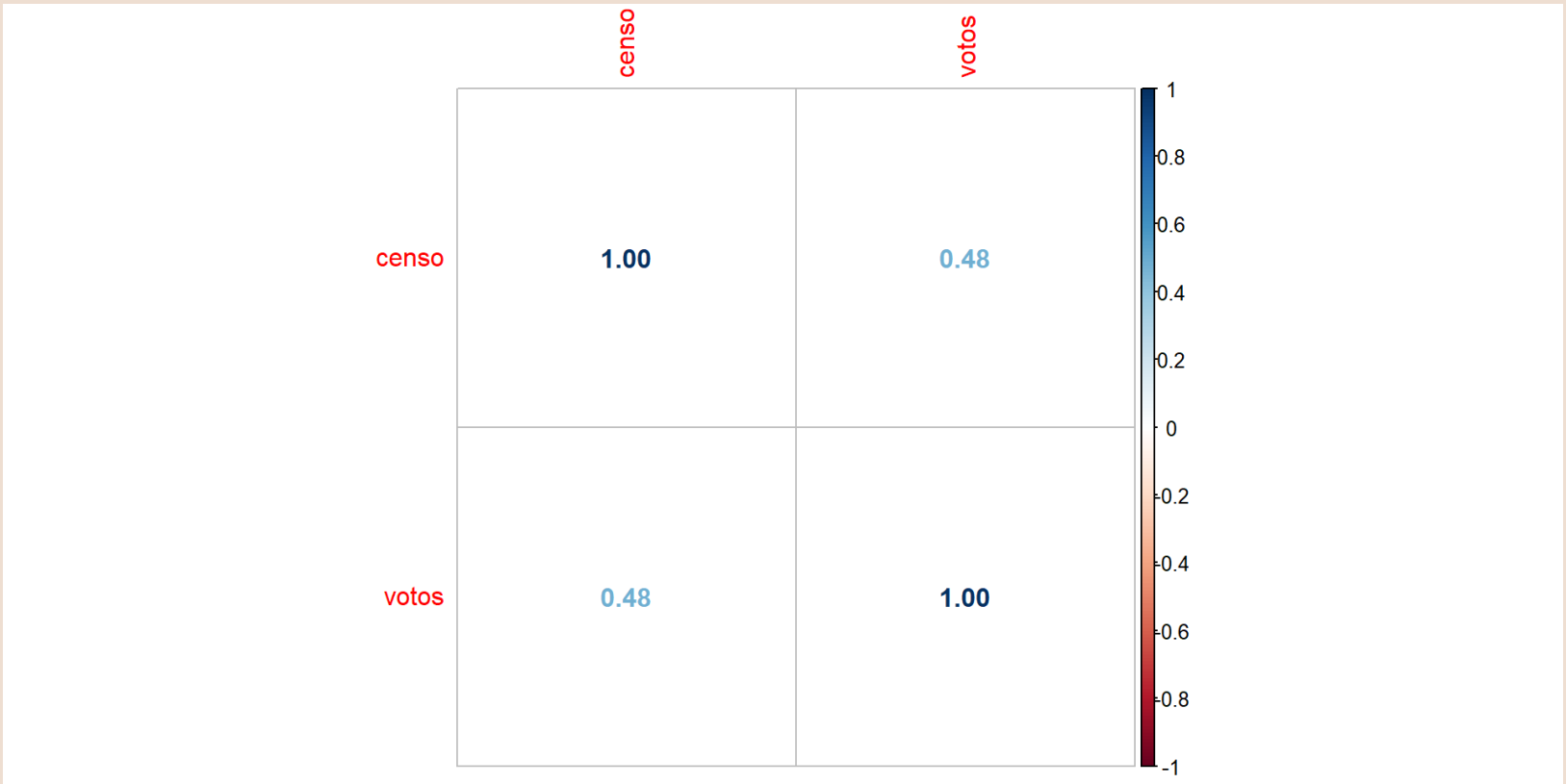
To visualize the results more clearly, boxplots have been created.



2.5 How to analyze the relationship between census and vote?

2.5 How to analyze the relationship between census and vote?

In order to analyze the relationship between the census and votes, a correlation plot has been generated to observe how these variables are correlated with each other. As we can see, they exhibit a moderate positive correlation of 0.48.



2.5 How to analyze the relationship between census and vote?

Additionally, a linear regression model has been created to determine if these variables are significant and may have an effect on each other.

```
Call:
lm(formula = censo ~ votos, data = election_pivot)

Residuals:
     Min       1Q   Median       3Q      Max
-2329601    -5130    -4688    -2521  2374697

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.188e+03  6.686e+01    77.6   <2e-16 ***
votos       4.988e+00  1.453e-02   343.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41970 on 396733 degrees of freedom
Multiple R-squared:  0.2289,    Adjusted R-squared:  0.2289
F-statistic: 1.178e+05 on 1 and 396733 DF,  p-value: < 2.2e-16
```

2.5 Is it true that certain parties win in rural areas?

As rural provinces, the following have been selected, as they are provinces in the depopulated Spain.

- 01-Álava 02-Albacete
- 05-Ávila 12-Castellón
- 13-Ciudad Real 16-Cuenca
- 19-Guadalajara 21-Huelva
- 22-Huesca 23-Jaén
- 25-Lleida, 26-La Rioja
- 27-Lugo 31-Navarra
- 32-Ourense 39-Cantabria,
- 40-Segovia 42-Soria
- 44-Teruel 49-Zamora

2.5 Is it true that certain parties win in rural areas?

```
depopulated_spain <- election_pivot |> filter(codigo_provincia == "01"|codigo_provincia == "02"|codig

win_depop_spain <- depopulated_spain |> select(c(codigo_provincia, date_elec_ym, siglas, votos))
win_depop_spain
```

```
# A tibble: 140,144 × 4
   codigo_provincia date_elec_ym siglas  votos
   <fct>            <date>       <chr>   <dbl>
 1 01               2008-03-01   PODEMOS     9
 2 01               2008-03-01   OTHERS     27
 3 01               2008-03-01   PSOE        1
 4 01               2008-03-01   OTHERS      1
 5 01               2008-03-01   OTHERS      2
 6 01               2008-03-01   PP        238
 7 01               2008-03-01   PODEMOS    61
 8 01               2008-03-01   OTHERS     85
 9 01               2008-03-01   OTHERS      4
10 01               2008-03-01   OTHERS     17
# ℹ 140,134 more rows
```

2.5 Is it true that certain parties win in rural areas?

Therefore, a new data frame named “maxvotes” has been created to consider the province code, election date, political party, and the number of votes each party received.

```r
maxvotes <- win_depop_spain |>
  filter(date_elec_ym %in% c("2008-03-01", "2011-11-01", "2015-12-01", "2016-06-01", "2019-04-01", "2
  group_by(date_elec_ym, codigo_provincia) |>
  slice_max(votos, n=1)
maxvotes
```

```
# A tibble: 120 × 4
# Groups:   date_elec_ym, codigo_provincia [120]
   codigo_provincia date_elec_ym siglas votos
   <fct>            <date>       <chr>  <dbl>
 1 01               2008-03-01   PSOE   56349
 2 02               2008-03-01   PP     49909
 3 05               2008-03-01   PP     20468
 4 12               2008-03-01   PP     42498
 5 13               2008-03-01   PP     23826
 6 16               2008-03-01   PP     15943
 7 19               2008-03-01   PP     23910
 8 21               2008-03-01   PSOE   37930
 9 22               2008-03-01   PSOE   13227
10 23               2008-03-01   PSOE   33232
# ℹ 110 more rows
```

2.5 Is it true that certain parties win in rural areas?

We can observe how in most provinces, the parties that won in the elections over the years have not always been the same. However, in Ávila, the PP wins over time.

```
alava_parties <- maxvotes |> filter(codigo_provincia == "01")
albacete_parties <- maxvotes |> filter(codigo_provincia == "02")
avila_parties <- maxvotes |> filter(codigo_provincia == "05")
avila_parties
```

```
# A tibble: 6 × 4
# Groups:   date_elec_ym, codigo_provincia [6]
  codigo_provincia date_elec_ym siglas votos
  <fct>            <date>       <chr>  <dbl>
1 05               2008-03-01   PP     20468
2 05               2011-11-01   PP     19391
3 05               2015-12-01   PP     14125
4 05               2016-06-01   PP     15553
5 05               2019-04-01   PP      9084
6 05               2019-11-01   PP      9260
```

2.5 Is it true that certain parties win in rural areas?

In Ciudad Real, the PP predominantly wins.

```
# A tibble: 6 × 4
# Groups:   date_elec_ym, codigo_provincia [6]
  codigo_provincia date_elec_ym siglas votos
  <fct>            <date>       <chr>  <dbl>
1 13               2008-03-01   PP     23826
2 13               2011-11-01   PP     24946
3 13               2015-12-01   PP     18080
4 13               2016-06-01   PP     19605
5 13               2019-04-01   PSOE   12838
6 13               2019-11-01   PP     12852
```

2.5 Is it true that certain parties win in rural areas?

In Cantabria, the PP predominantly wins.

```
# A tibble: 6 × 4
# Groups:   date_elec_ym, codigo_provincia [6]
  codigo_provincia date_elec_ym siglas votos
  <fct>            <date>       <chr>  <dbl>
1 39               2008-03-01   PP     58821
2 39               2011-11-01   PP     56866
3 39               2015-12-01   PP     39839
4 39               2016-06-01   PP     43755
5 39               2019-04-01   PSOE   25466
6 39               2019-11-01   PP     28019
```

2.5 Is it true that certain parties win in rural areas?

Lastly, in Segovia, the PP predominantly wins.

```
# A tibble: 6 × 4
# Groups:   date_elec_ym, codigo_provincia [6]
  codigo_provincia date_elec_ym siglas votos
  <fct>            <date>       <chr>  <dbl>
1 40               2008-03-01   PP     16555
2 40               2011-11-01   PP     15867
3 40               2015-12-01   PP     10669
4 40               2016-06-01   PP     12149
5 40               2019-04-01   PSOE    9286
6 40               2019-11-01   PP      9213
```

2.6 How to calibrate the error of the polls (remember that the polls are voting intentions at national level)?

2.6 How to calibrate the error of the polls?

We first need to convert the votes obtained by the parties in each election into percentages.

- Total votes by election

```
election_pivot_total <- election_pivot |>
  group_by(date_elec_ym) |>
  summarise(total_votes_election = sum(votos)) |>
  print()
```

```
# A tibble: 6 × 2
  date_elec_ym total_votes_election
  <date>                      <dbl>
1 2008-03-01                25069038
2 2011-11-01                23918052
3 2015-12-01                24895828
4 2016-06-01                23754401
5 2019-04-01                25877751
6 2019-11-01                23862002
```

2.6 How to calibrate the error of the polls (remember that the polls are voting intentions at national level)?

- Total votes obtained by each party in each election

```
election_pivot_votes <- election_pivot |>
  group_by(date_elec_ym, siglas) |>
  summarise(total_votes_party = sum(votos))
```

- Percentage

```
votes_election <-
  inner_join(election_pivot_total , election_pivot_votes, by="date_elec_ym") |>
  mutate(result_election = 100*(total_votes_party / total_votes_election))
```

2.6 How to calibrate the error of the polls (remember that the polls are voting intentions at national level)?

Since there are several polls for the same election, conducted by different media, we can compute the mean to obtain the average estimated result for each party and election.

```
votes_survey <- surveys_pivot |>
  group_by(date_elec_ym, party) |>
  summarize(result_poll = mean(votes_percent))
```

Once we have percentage in both election and poll data, we merge the results. Then, the difference between the poll estimation and the real results is computed in order to get the polling error.

```
merged_votes <-
  inner_join(votes_election, votes_survey,
             by = c("date_elec_ym" = "date_elec_ym",
                    "siglas" = "party")) |>
  mutate(poll_error = result_election - result_poll)
```

2.6 How to calibrate the error of the polls (remember that the polls are voting intentions at national level)?

```
print(merged_votes |>
        select(date_elec_ym,siglas, result_election, result_poll, poll_error))
```

```
# A tibble: 39 × 5
   date_elec_ym siglas  result_election result_poll poll_error
   <date>       <chr>             <dbl>       <dbl>      <dbl>
 1 2008-03-01   BNG               0.835       1.08      -0.248
 2 2008-03-01   EAJ-PNV           1.21        1.79      -0.581
 3 2008-03-01   ERC               1.18        2.33      -1.15
 4 2008-03-01   PP               40.6        38.4        2.18
 5 2008-03-01   PSOE             44.2        42.2        1.99
 6 2011-11-01   BNG               0.766       1.08      -0.316
 7 2011-11-01   EAJ-PNV           1.35        1.22       0.131
 8 2011-11-01   ERC               1.07        0.962      0.110
 9 2011-11-01   PP               45.3        43.3        2.05
10 2011-11-01   PSOE             29.2        36.0       -6.82
# ℹ 29 more rows
```

2.7 In which election were the polls most wrong?

2.7 In which election were the polls most wrong?

To analyze which election had the most inaccurate polls, we first compute the average error for each election and search for the date with the highest value.

```
# A tibble: 6 × 2
  date_elec_ym avg_poll_error
  <date>                <dbl>
1 2008-03-01             1.23
2 2011-11-01             1.89
3 2015-12-01             1.23
4 2016-06-01             1.04
5 2019-04-01             0.78
6 2019-11-01             1.41
```

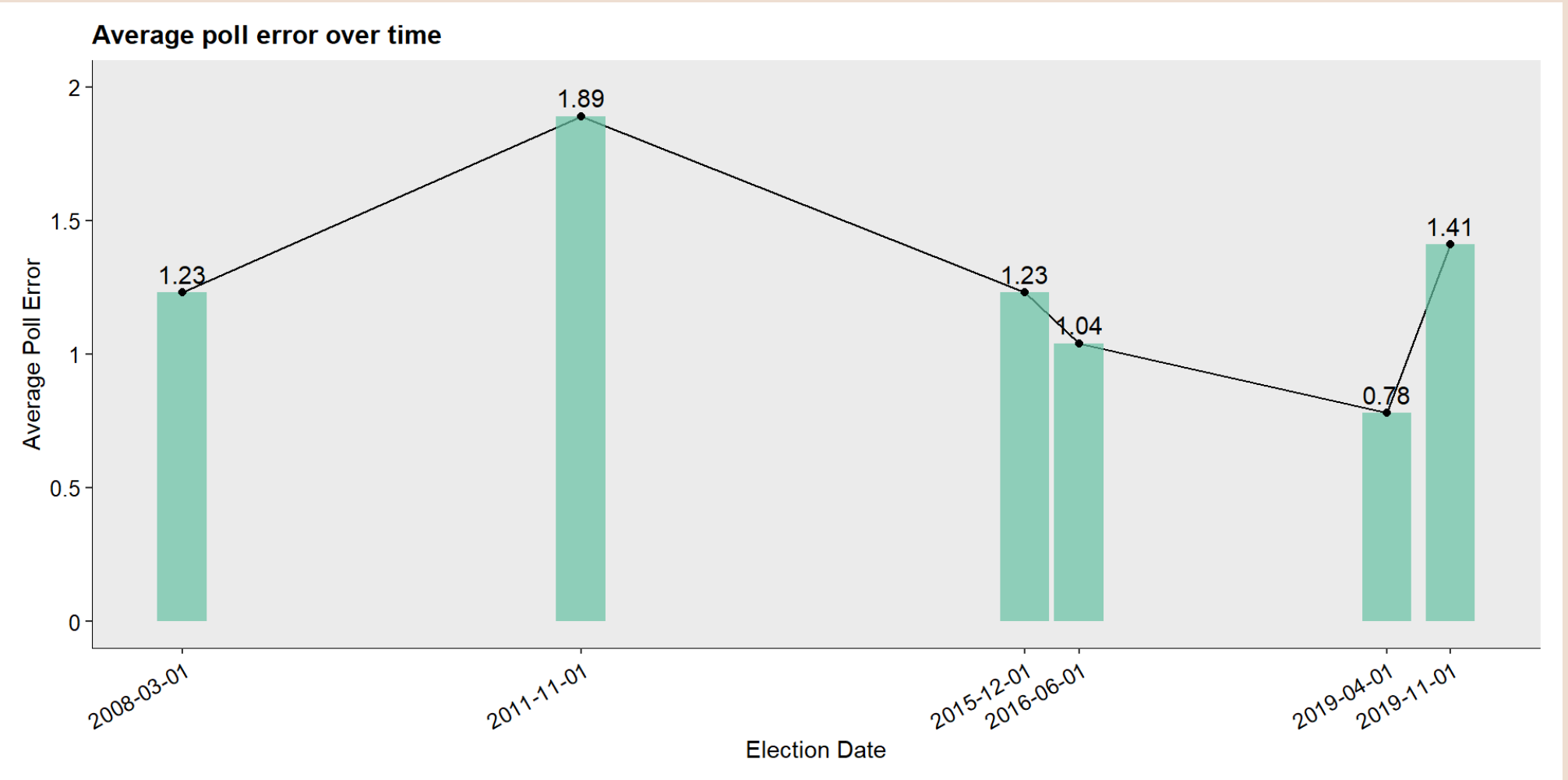
```
max_avg_error <- avg_error |>
  slice_max(order_by = abs(avg_poll_error), n = 1) |>
  print()
```

```
# A tibble: 1 × 2
  date_elec_ym avg_poll_error
  <date>                <dbl>
1 2011-11-01             1.89
```

The polls conducted on November 1, 2011, had the highest average error (1.89). Therefore, we can assume that these polls experienced the most significant inaccuracies, suggesting potential challenges or issues in their predictions.

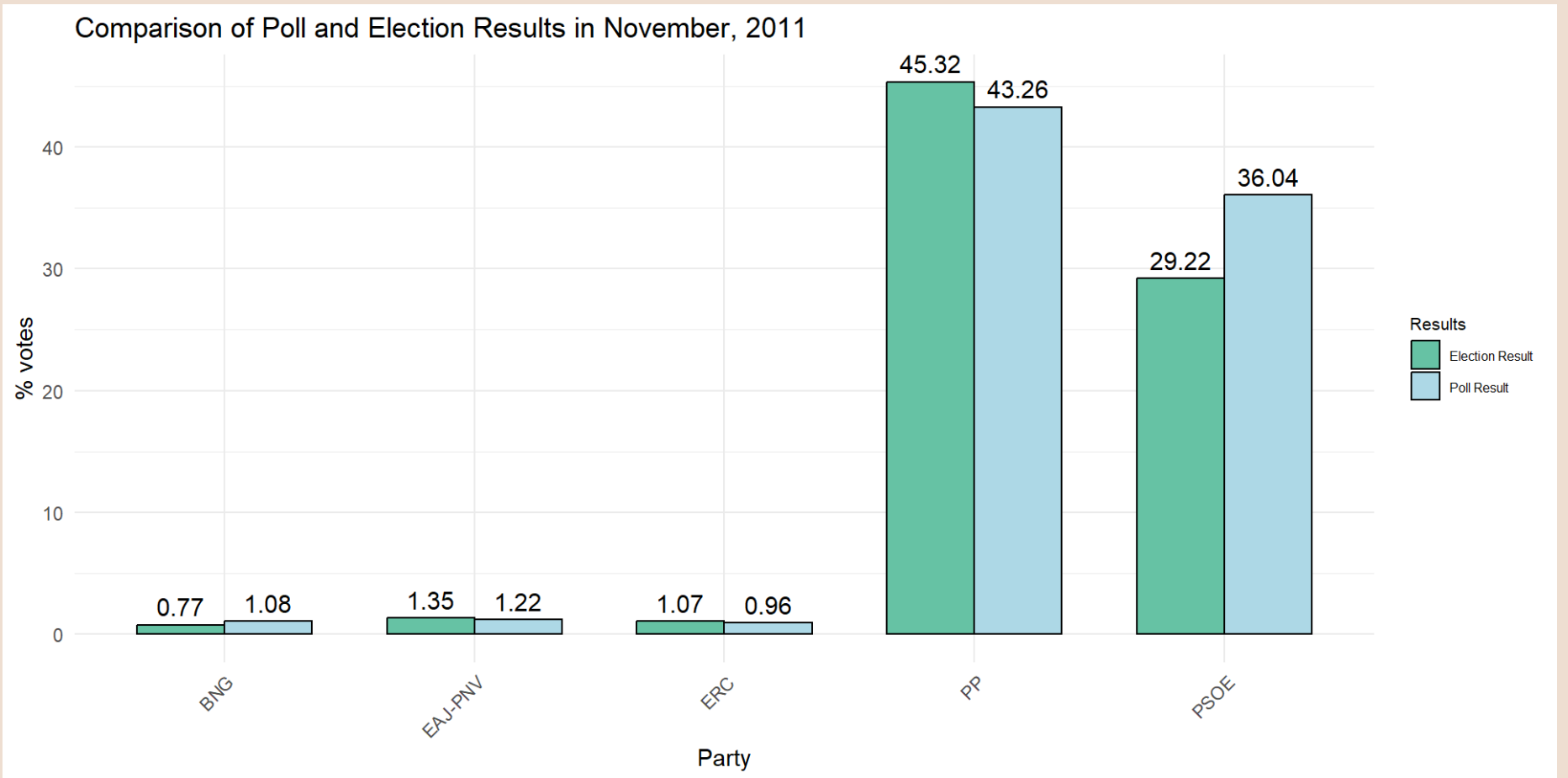
2.7 In which election were the polls most wrong?

The evolution of the average error can be represented for every election to observe the deviation in poll predictions over time.



2.7 In which election were the polls most wrong?

We further analyse the results obtained in these polls. By doing so, we can observe the deviation between the estimated results by the media and the real results obtained by the parties.



2.8 How were the polls wrong in national parties (PSOE, PP, VOX, CS, MP, UP - IU)?

2.8 How were the polls wrong in national parties (PSOE, PP, VOX, CS, MP, UP - IU)?

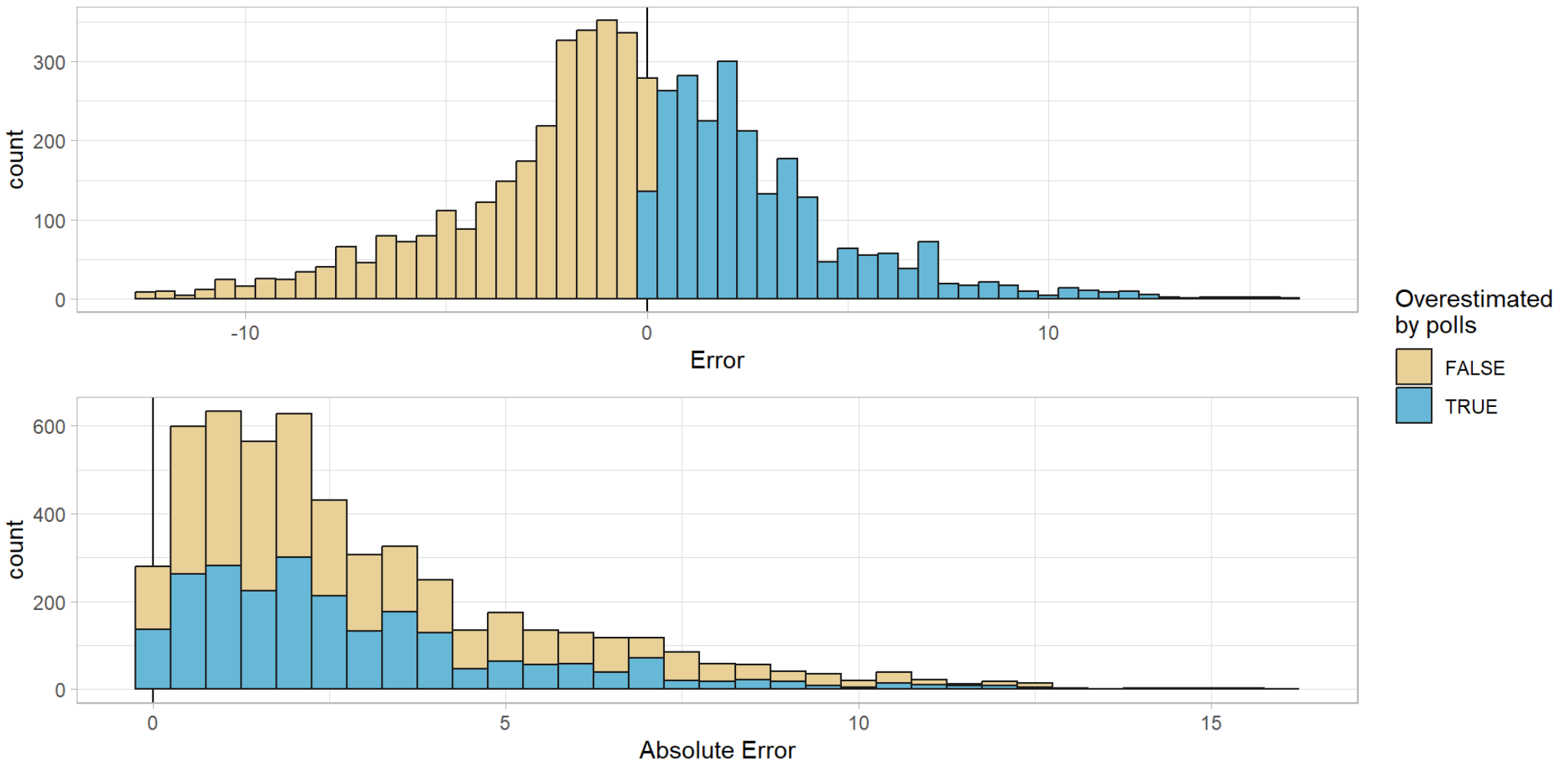
PSOE in 2011 was very over estimated by the polls

```
national_parties <- c("PSOE", "PP", "VOX", "CS", "M País", "PODEMOS")

merged_votes <- inner_join(votes_election, surveys_pivot,
                           by = c("date_elec_ym" = "date_elec_ym",
                  "siglas" = "party"),
                  multiple = "all") |>
  mutate(error = votes_percent - result_election,
         abs_error = abs(error),
         overestimated = if_else(error > 0, T, F)
         ) |>
  filter(siglas %in% national_parties) |>
  select(-c(low_turnout))
```

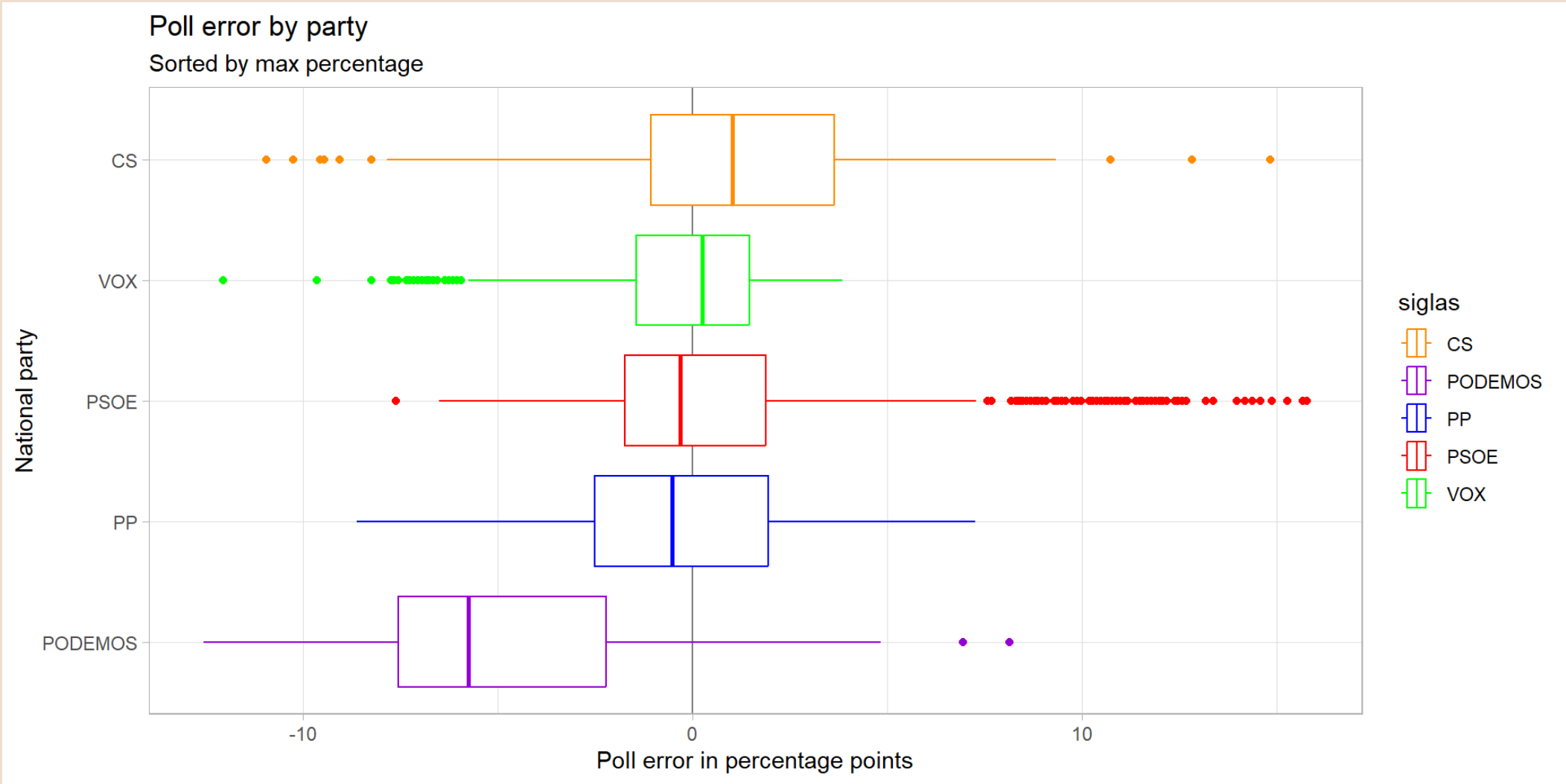
2.8 How were the polls wrong in national parties (PSOE, PP, VOX, CS, MP, UP - IU)?

```
# A tibble: 10 × 6
   date_elec  siglas result_election pollster      votes_percent error
   <date>     <chr>            <dbl> <chr>                 <dbl> <dbl>
 1 2011-11-20 PSOE              29.2  GESOP                  45    15.8
 2 2011-11-20 PSOE              29.2  CELESTE-TEL            44.9  15.7
 3 2011-11-20 PSOE              29.2  CELESTE-TEL            44.5  15.3
 4 2011-11-20 PSOE              29.2  ASEP                   44.1  14.9
 5 2019-11-10 CS                 6.86 SIMPLE LÓGICA          21.7  14.8
 6 2011-11-20 PSOE              29.2  ASEP                   43.8  14.6
 7 2011-11-20 PSOE              29.2  CIS                    43.6  14.4
 8 2011-11-20 PSOE              29.2  CELESTE-TEL            43.4  14.2
 9 2011-11-20 PSOE              29.2  SIMPLE LÓGICA          43.2  14.0
10 2011-11-20 PSOE              29.2  SIGMA DOS              42.6  13.4
```

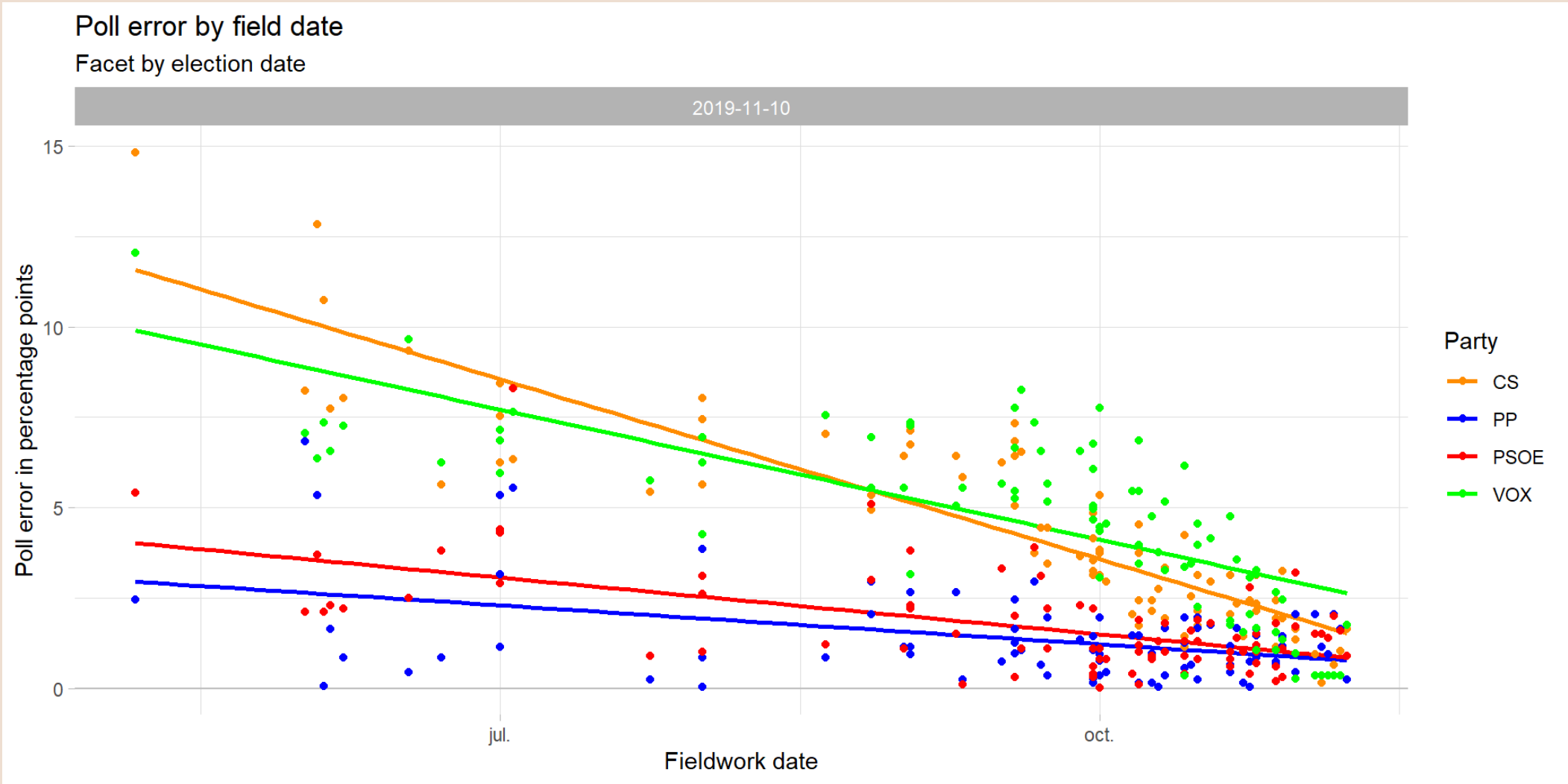
Poll error distribution



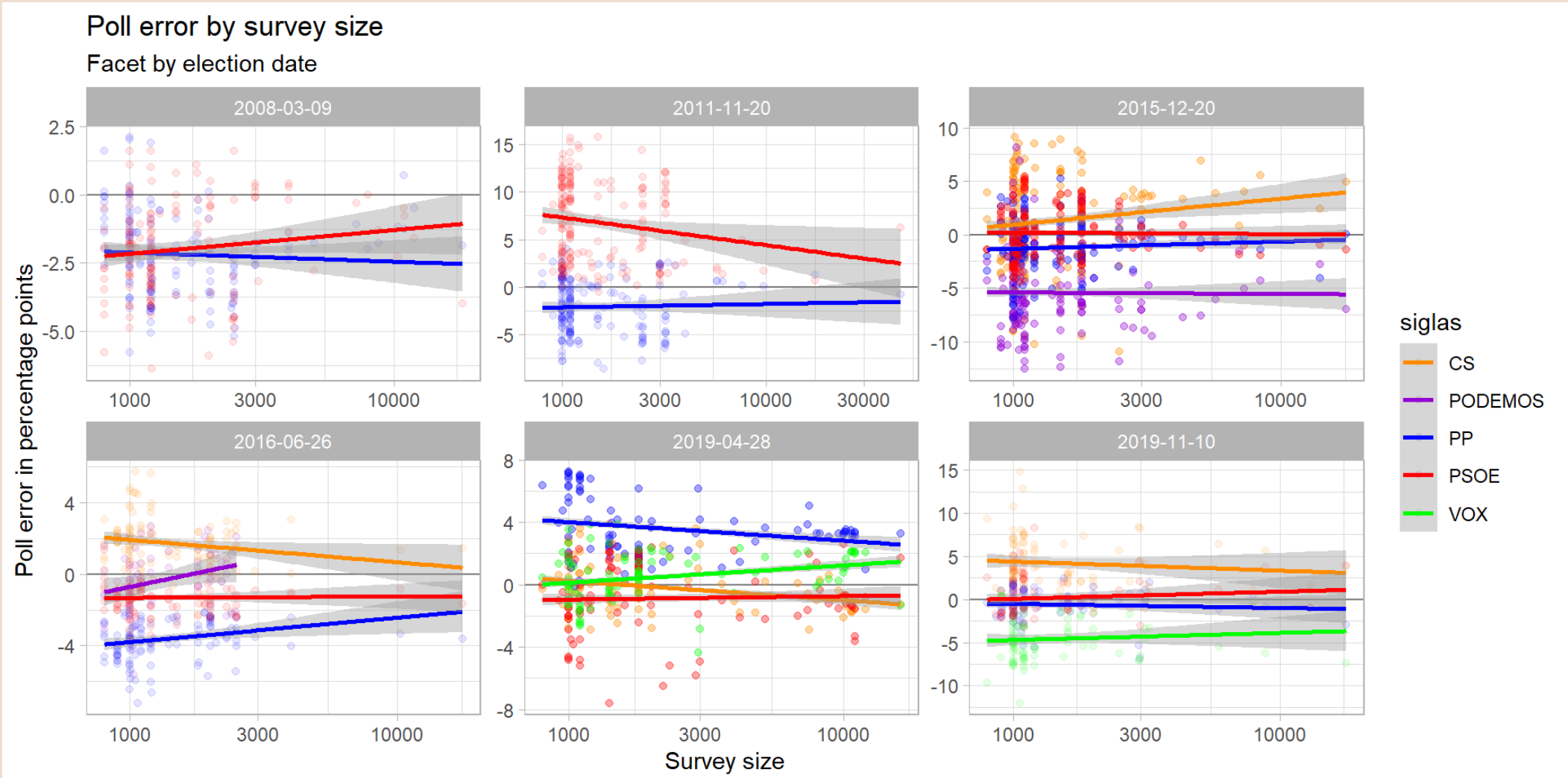
Most polls underestimated PODEMOS



Fieldwork date affected accuracy of Nov 2019 election



Did size have an affect on poll error?



2.9 Which polling houses got it right the most and which ones deviated the most from the results?

Each polling house surveyed a different number of parties

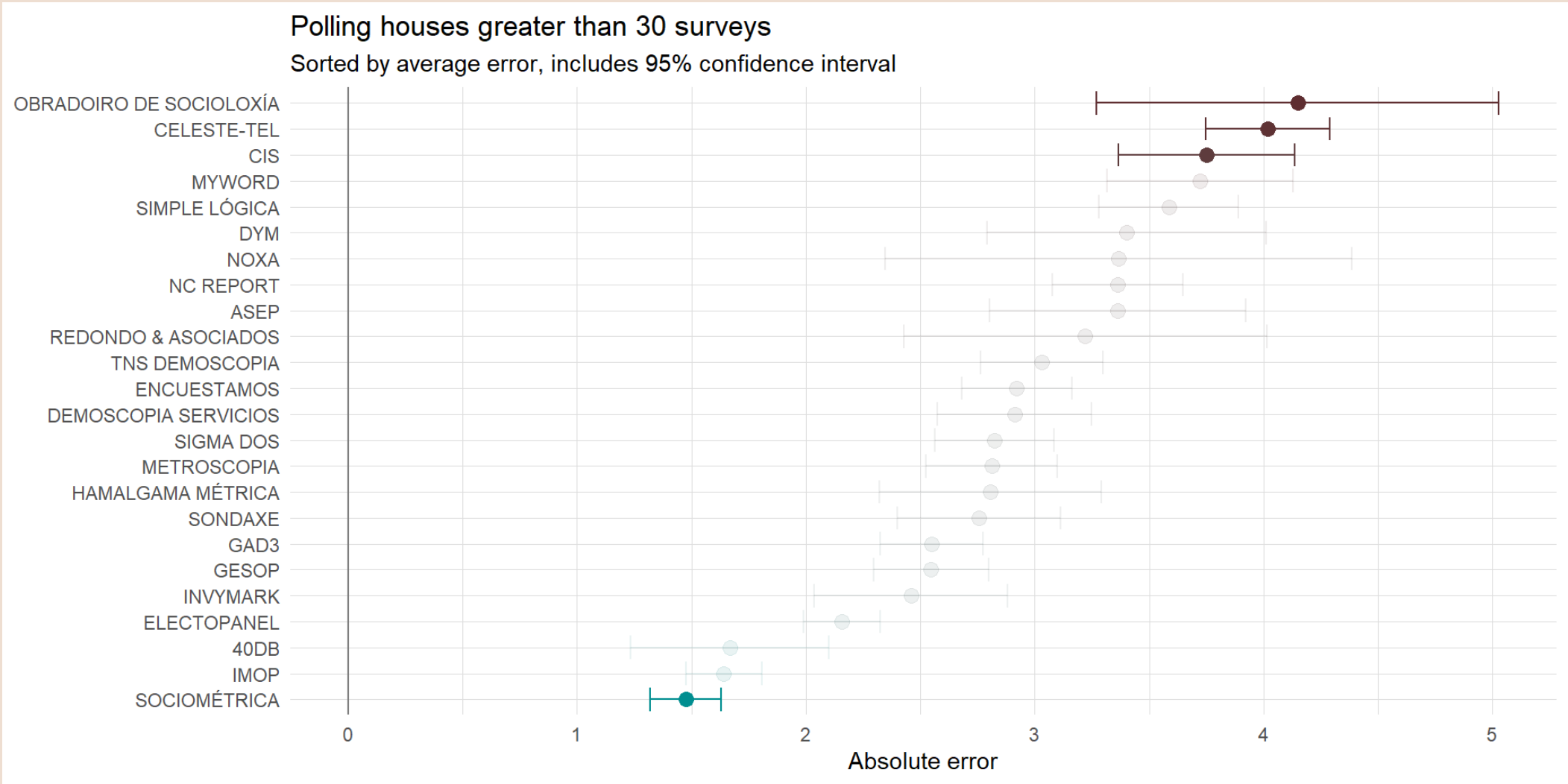
	CS	PODEMOS	PP	PSOE	VOX
40DB	10	0	10	10	10
A+M	5	4	5	5	0
ADVICE STRATEGIC	3	1	3	3	0
APPEND	0	0	1	1	0
ASEP	0	0	39	39	0
CELESTE-TEL	117	68	164	164	43
CEMOP	0	0	1	1	0
CIS	39	22	70	70	16
DEMOMÉTRICA	0	0	6	6	0
DEMOSCOPIA SERVICIOS	45	32	45	45	13
DYM	16	10	28	28	6
ELECTOPANEL	80	0	80	80	80
ENCUESTAMOS	56	53	56	56	0
ESTUDIO DE SOCIOLOGÍA CONSULTORES	3	3	3	3	0
GAD3	107	52	109	109	50
GALLUP	0	0	7	7	0
GESOP	51	21	75	75	16
GIPEYOP	7	5	7	7	0
HAMALGAMA MÉTRICA	13	4	13	13	8

Obradoiro de Socioloxía have the worst average error

```r
subgroup <- merged_votes |>
  group_by(pollster) |>
  summarise(
    avg_error = mean(abs_error),
    var_error = var(abs_error),
    sd_error = sd(abs_error),
    total_surveys = n(),
    num_parties = length(unique(siglas))
  )

subgroup |> select(-var_error) |>
  slice_max(avg_error, n = 5)
```

```
# A tibble: 5 × 5
  pollster                avg_error sd_error total_surveys num_parties
  <chr>                       <dbl>    <dbl>         <int>       <int>
1 OBRADOIRO DE SOCIOLOXÍA      4.15     3.64            66           2
2 CELESTE-TEL                  4.02     3.25           556           5
3 INTERCAMPO                   4.01     3.36            12           2
4 A+M                          3.84     3.11            19           4
5 CIS                          3.75     2.89           217           5
```

2.9 Which polling houses got it right the most and which ones deviated the most from the results?



Which polling houses had the best estimates?

```
subgroup |> select(-var_error) |>
  slice_min(avg_error, n = 7)
```

```
# A tibble: 7 × 5
  pollster                          avg_error sd_error total_surveys num_parties
  <chr>                                 <dbl>    <dbl>         <int>       <int>
1 TÁBULA V                               1.10    1.39              2           2
2 SW DEMOSCOPIA                          1.11    0.784             4           4
3 SYM CONSULTING                         1.15    1.72              3           3
4 DEMOMÉTRICA                            1.21    1.11             12           2
5 OPINA                                  1.38    1.14              2           2
6 ESTUDIO DE SOCIOLOGÍA CONSULTORES      1.45    1.36             12           4
7 SOCIOMÉTRICA                           1.47    1.24            248           4
```