Data Cleaning

In November 2025, California will vote on Proposition 50, which would change the state's congressional district map. To see how this might affect elections, we use 2024 voting data, precinct maps, and the proposed new map from the California Secretary of State and the Statewide Database. This lets us see how the 2024 results would look under the new districts.

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```
Warning: package 'purrr' was built under R version 4.3.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
raw_df <- read_csv("data/g24_sov_by_g24_svprec.csv")
```

```
Rows: 51123 Columns: 76
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (49): FIPS, SVPREC, SVPREC_KEY, ELECTION, GEO_TYPE, ASSAIP01, ASSDEM01, ...
dbl (27): COUNTY, ADDIST, CDDIST, SDDIST, BEDIST, TOTREG, DEMREG, REPREG, AI...
```

```
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
cleaned_df <- raw_df %>%
  rename_with(~ str_replace_all(., " ", "_")) %>%
  rename_with(tolower) %>%
  mutate(total_votes = as.numeric(totvote),
         demvote = as.numeric(cngdem01),
         repvote = as.numeric(cngrep01)) %>%
  mutate(across(where(is.character), str_trim)) %>%
  filter(total_votes > 0)
```

```
Warning: There were 2 warnings in `mutate()`.
The first warning was:
i In argument: `demvote = as.numeric(cngdem01)`.
Caused by warning:
! NAs introduced by coercion
i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

```r
glimpse(cleaned_df)
```

```
Rows: 38,657
Columns: 77
$ county      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ fips        <chr> "06001", "06001", "06001", "06001", "06001", "06001", "060~
$ svprec      <chr> "200100", "200100A", "200200", "200200A", "201400", "20140~
$ addist      <dbl> 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14, 14~
$ svprec_key  <chr> "06001200100", "06001200100A", "06001200200", "06001200200~
$ election    <chr> "g24", "g24", "g24", "g24", "g24", "g24", "g24", "g24", "g~
$ geo_type    <chr> "svprec", "svprec", "svprec", "svprec", "svprec", "svprec"~
$ cddist      <dbl> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12~
$ sddist      <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7~
$ bedist      <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
$ totreg      <dbl> 3535, 0, 2442, 0, 3773, 0, 541, 0, 1105, 0, 948, 0, 2721, ~
$ demreg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ repreg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ aipreg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ grnreg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ libreg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
$ nlpreg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ refreg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ dclreg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ mscreg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ totvote     <dbl> 256, 2804, 262, 1816, 283, 2782, 89, 343, 394, 297, 837, 2~
$ demvote     <dbl> 102, 1668, 108, 1063, 139, 1688, 35, 192, 172, 150, 366, 1~
$ repvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, ~
$ aipvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ grnvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ libvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ nlpvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ refvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ dclvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ mscvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ prcvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ absvote     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ assaip01    <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"~
$ assdem01    <chr> "94", "444", "117", "348", "107", "588", "45", "105", "181~
$ assdem02    <chr> "110", "2023", "91", "1243", "128", "1841", "24", "172", "~
$ assrep01    <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"~
$ assrep02    <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"~
$ cngdem01    <chr> "102", "1668", "108", "1063", "139", "1688", "35", "192", ~
$ cngdem02    <chr> "102", "771", "99", "513", "98", "739", "38", "93", "143",~
$ cngind01    <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"~
$ cngrep01    <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"~
$ cngrep02    <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"~
$ prsaip01    <chr> "3", "10", "1", "6", "3", "13", "2", "2", "3", "4", "5", "~
$ prsdem01    <chr> "181", "2562", "207", "1647", "231", "2522", "73", "297", ~
$ prsgrn01    <chr> "9", "48", "13", "41", "9", "52", "0", "13", "6", "13", "2~
$ prslib01    <chr> "1", "10", "2", "3", "4", "13", "0", "0", "3", "2", "3", "~
$ prspaf01    <chr> "5", "17", "7", "12", "8", "32", "2", "4", "3", "7", "16",~
$ prsrep01    <chr> "51", "108", "26", "83", "17", "111", "11", "23", "55", "2~
$ pr_2_n      <chr> "58", "493", "45", "342", "39", "399", "17", "45", "52", "~
$ pr_2_y      <chr> "169", "2156", "196", "1385", "226", "2231", "66", "278", ~
$ pr_32_n     <chr> "78", "636", "55", "439", "74", "536", "14", "60", "81", "~
$ pr_32_y     <chr> "148", "1966", "187", "1261", "190", "2070", "68", "255", ~
$ pr_33_n     <chr> "136", "1774", "105", "1092", "124", "1509", "30", "127", ~
$ pr_33_y     <chr> "86", "784", "126", "584", "133", "1053", "49", "177", "23~
$ pr_34_n     <chr> "123", "1485", "121", "1027", "144", "1515", "46", "186", ~
$ pr_34_y     <chr> "98", "980", "105", "601", "96", "941", "33", "107", "174"~
$ pr_35_n     <chr> "54", "581", "45", "419", "58", "563", "20", "57", "61", "~
$ pr_35_y     <chr> "171", "2003", "188", "1261", "196", "1988", "58", "248", ~
$ pr_36_n     <chr> "106", "1356", "142", "888", "146", "1487", "49", "197", "~
```

```
$ pr_36_y     <chr> "118", "1223", "99", "786", "119", "1084", "31", "112", "1~
$ pr_3_n      <chr> "51", "133", "25", "116", "33", "152", "10", "26", "38", "~
$ pr_3_y      <chr> "183", "2553", "220", "1646", "240", "2508", "74", "295", ~
$ pr_4_n      <chr> "52", "381", "37", "271", "37", "330", "14", "41", "40", "~
$ pr_4_y      <chr> "181", "2294", "209", "1472", "231", "2316", "68", "279", ~
$ pr_5_n      <chr> "94", "961", "66", "605", "63", "742", "19", "76", "72", "~
$ pr_5_y      <chr> "132", "1660", "168", "1096", "197", "1862", "61", "240", ~
$ pr_6_n      <chr> "75", "607", "59", "407", "57", "532", "17", "53", "85", "~
$ pr_6_y      <chr> "143", "1958", "180", "1274", "196", "2029", "62", "257", ~
$ sendem01    <chr> "107", "1719", "101", "1102", "136", "1578", "37", "153", ~
$ sendem02    <chr> "103", "809", "114", "516", "105", "908", "34", "133", "17~
$ senrep01    <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"~
$ senrep02    <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"~
$ uspdem01    <chr> "172", "2461", "199", "1572", "217", "2444", "67", "285", ~
$ usprep01    <chr> "53", "155", "34", "111", "32", "153", "11", "23", "53", "~
$ ussdem01    <chr> "173", "2487", "207", "1593", "222", "2478", "67", "288", ~
$ ussrep01    <chr> "55", "155", "29", "109", "33", "151", "12", "23", "51", "~
$ total_votes <dbl> 256, 2804, 262, 1816, 283, 2782, 89, 343, 394, 297, 837, 2~
```

```
summary(cleaned_df$demvote)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0      33     111    1555     389 2273160    4684
```

```
summary(cleaned_df$repvote)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0      31      90    1008     240 1050936    4684
```

```
write_csv(cleaned_df, "data/cleaned_g24_sov_by_svprec.csv")
```