Data Cleaning

```
# Libraries
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.1     ✔ stringr   1.5.2
✔ ggplot2   4.0.0     ✔ tibble    3.3.0
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.1.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(sf)
```

```
Linking to GEOS 3.13.0, GDAL 3.8.5, PROJ 9.5.1; sf_use_s2() is TRUE
```

```
library(units)
```

```
udunits database from /Library/Frameworks/R.framework/Versions/4.5-arm64/
Resources/library/units/share/udunits/udunits2.xml
```

```
# Read in 2024 General Election Data (precinct level)
raw_precinct <- 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.
```

```
# View raw data
View(raw_precinct)

# Columns that had * in them
cols_with_star <- c(
  "ASSAIP01", "ASSDEM01", "ASSDEM02", "ASSREP01", "ASSREP02",
  "CNGDEM01", "CNGDEM02", "CNGIND01", "CNGREP01", "CNGREP02",
  "PRSAIP01", "PRSDEM01", "PRSGRN01", "PRSLIB01", "PRSPAF01", "PRSREP01",
  "PR_2_N", "PR_2_Y", "PR_32_N", "PR_32_Y", "PR_33_N", "PR_33_Y",
  "PR_34_N", "PR_34_Y", "PR_35_N", "PR_35_Y", "PR_36_N", "PR_36_Y",
  "PR_3_N", "PR_3_Y", "PR_4_N", "PR_4_Y", "PR_5_N", "PR_5_Y",
  "PR_6_N", "PR_6_Y",
  "SENDEM01", "SENDEM02", "SENREP01", "SENREP02",
  "USPDEM01", "USPREP01",
  "USSDEM01", "USSREP01")

# Turn * into NA
precinct_election <- raw_precinct |>
  mutate(across(
    all_of(cols_with_star),
    ~ .x |>
      as.character() |>
      str_trim() |>
      na_if("*") |>        # replace * with NA
      as.numeric()))
```

```
Warning: There were 44 warnings in `mutate()`.
The first warning was:
ℹ In argument: `across(...)`.
Caused by warning:
! NAs introduced by coercion
ℹ Run `dplyr::last_dplyr_warnings()` to see the 43 remaining warnings.
```

```
View(precinct_election)
```

```
# Remove precincts with absurdly high number of voters (likely total county or
something instead of actual precinct numbers)
```

```r
# How many precincts will this filtering remove?
precinct_election |>
  mutate(
    precinct_size = PRSDEM01 + PRSREP01 + PRSAIP01 + PRSGRN01 + PRSLIB01 +
PRSPAF01) |>
  summarize(
    n_total = n(),
    n_over_10k = sum(precinct_size > 10000, na.rm = TRUE))
```

```
# A tibble: 1 × 2
  n_total n_over_10k
    <int>      <int>
1   51123        484
```

```r
# Filter out precincts with greater than 10,000 votes
precinct_election <- precinct_election |>
  mutate(
    precinct_size = PRSDEM01 + PRSREP01 + PRSAIP01 + PRSGRN01 + PRSLIB01 +
PRSPAF01) |>
  filter(precinct_size <= 10000)
```