Data Cleaning

```r
library(tidyverse)
```

```
── 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   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.4
── 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
```

```r
library(janitor)
```

```


Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```r
vote_prec <- read_csv("/Users/hibahalam/Desktop/stat133/gerrymandering-
hibahalam/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.
```

```
head(vote_prec)
```

```
# A tibble: 6 × 76
  COUNTY FIPS  SVPREC   ADDIST SVPREC_KEY  ELECTION GEO_TYPE CDDIST SDDIST
BEDIST
   <dbl> <chr> <chr>    <dbl> <chr>       <chr>    <chr>     <dbl>  <dbl>
<dbl>
1      1 06001 200100      14 06001200100 g24      svprec       12      7
2
2      1 06001 200100A     14 0600120010… g24      svprec       12      7
2
3      1 06001 200200      14 06001200200 g24      svprec       12      7
2
4      1 06001 200200A     14 0600120020… g24      svprec       12      7
2
5      1 06001 201400      14 06001201400 g24      svprec       12      7
2
6      1 06001 201400A     14 0600120140… g24      svprec       12      7
2
# ℹ 66 more variables: TOTREG <dbl>, DEMREG <dbl>, REPREG <dbl>, AIPREG <dbl>,
#   GRNREG <dbl>, LIBREG <dbl>, NLPREG <dbl>, REFREG <dbl>, DCLREG <dbl>,
#   MSCREG <dbl>, TOTVOTE <dbl>, DEMVOTE <dbl>, REPVOTE <dbl>, AIPVOTE <dbl>,
#   GRNVOTE <dbl>, LIBVOTE <dbl>, NLPVOTE <dbl>, REFVOTE <dbl>, DCLVOTE <dbl>,
#   MSCVOTE <dbl>, PRCVOTE <dbl>, ABSVOTE <dbl>, ASSAIP01 <chr>,
#   ASSDEM01 <chr>, ASSDEM02 <chr>, ASSREP01 <chr>, ASSREP02 <chr>,
#   CNGDEM01 <chr>, CNGDEM02 <chr>, CNGIND01 <chr>, CNGREP01 <chr>, …
```

```
vote_prec <- vote_prec |>
  clean_names()
vote_cols <- vote_prec |>
     select(matches("vote|ballot|tot|^d_|^r_")) |>
     names()
vote_prec <- vote_prec |>
    mutate(across(all_of(vote_cols), ~ gsub(",", "", .))) |>
    mutate(across(all_of(vote_cols), as.numeric))
```

```
if ("svprec" %in% names(vote_prec)) {
  vote_prec <- vote_prec |>
     mutate(svprec = as.character(svprec))
}
```

```
vote_prec <- vote_prec |>
     select(where(~ !all(is.na(.x))))
```

```
dup_ids <- vote_prec |>
     count(svprec) |>
     filter(n > 1)

print(dup_ids)
```

```
# A tibble: 1,923 × 2
   svprec       n
   <chr>    <int>
 1 0000001      2
 2 0000001A     2
 3 0000002      2
 4 0000002A     2
 5 0000003      2
 6 0000003A     2
 7 0000004      2
 8 0000004A     2
 9 0000005      2
10 0000005A     2
# ℹ 1,913 more rows
```

```
vote_prec <- vote_prec |>
    distinct(svprec, .keep_all = TRUE)
```

```
write_csv(
  vote_prec,
  "data/vote_prec_clean.csv"
)
```