Data Cleaning

```r
#load 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
```

```r
library(janitor)
```

```


Attaching package: 'janitor'

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

    chisq.test, fisher.test
```

```r
raw_data <- 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
head(raw_data)
```

```
# 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>, …
```

```r
#standardize column names
clean_data <- raw_data |>
    clean_names()
#vote counts columns
vote_cols <- clean_data |>
    select(matches("vote|ballot|tot|^d_|^r_")) |>
    names()
#make number cols numeric
clean_data <- clean_data |>
    mutate(across(all_of(vote_cols), as.numeric))
#remove empty cols
clean_data <- clean_data |>
    select(where(~ !all(is.na(.x))))
write_csv(clean_data, "data/g24_sov_by_g24_svprec_clean.csv")
```

###Part 5

```
library(sf)
```

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

```
sr_votes_raw <- read_csv("data/state_g24_sov_data_by_g24_srprec.csv")
```

```
Rows: 25245 Columns: 76
```

```
── Column specification
────────────────────────────────────────────────────────
Delimiter: ","
chr (49): FIPS, SRPREC, ELECTION, SRPREC_KEY, 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.
```

```
sr_votes <- sr_votes_raw|>
  clean_names()
#identifying key columns
precinct_col <- "srprec"
vote_cols <- c("uspdem01", "usprep01")

sr_votes <- sr_votes |>
  mutate(across(all_of(vote_cols), ~ as.numeric(.)))
```

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

```
sr_prec_shp <- st_read("data/shapefiles/srprec_state_g24_v01_shp/
srprec_state_g24_v01_shp.shp")
```

```
Reading layer `srprec_state_g24_v01_shp' from data source
  `/Users/molly/Desktop/stat-133/gerrymandering-mollyurfalian/data/shapefiles/
```

```
srprec_state_g24_v01_shp/srprec_state_g24_v01_shp.shp'
  using driver `ESRI Shapefile'
```

```
Warning in CPL_read_ogr(dsn, layer, query, as.character(options), quiet, :
GDAL
Message 1:
/Users/molly/Desktop/stat-133/gerrymandering-mollyurfalian/data/shapefiles/
srprec_state_g24_v01_shp/srprec_state_g24_v01_shp.shp
contains polygon(s) with rings with invalid winding order. Autocorrecting
them,
but that shapefile should be corrected using ogr2ogr for example.
```

```
Simple feature collection with 24224 features and 6 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: -124.482 ymin: 32.52883 xmax: -114.1312 ymax: 42.0095
Geodetic CRS:  NAD83
```

```
sr_shape <- sr_prec_shp |>
    clean_names()

sr_shp <- sr_shape |>
  st_transform(3310) |>
  st_set_precision(1) |>
  st_make_valid() |>
  st_collection_extract("POLYGON")

sr_geo <- sr_shp |>
  left_join(sr_votes, by = c("srprec" = "srprec")) |>
  mutate(sr_area = st_area(geometry))
```

```
Warning in sf_column %in% names(g): Detected an unexpected many-to-many
relationship between `x` and `y`.
i Row 11 of `x` matches multiple rows in `y`.
i Row 376 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

```
AB604_raw <- st_read("data/shapefiles/AB604/AB604.shp")
```

```
Reading layer `AB604' from data source
  `/Users/molly/Desktop/stat-133/gerrymandering-mollyurfalian/data/shapefiles/
AB604/AB604.shp'
```

```
  using driver `ESRI Shapefile'
Simple feature collection with 52 features and 15 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: -13857270 ymin: 3832931 xmax: -12705030 ymax: 5162404
Projected CRS: WGS 84 / Pseudo-Mercator
```

```
ab604 <- AB604_raw |>
    clean_names() |>
    st_transform(3310)

cd_id_col <- "district"

sr_cd <- st_intersection(
  sr_geo |> select(srprec, sr_area, all_of(vote_cols)),
  ab604 |> select(all_of(cd_id_col))
) |>
  mutate(
    inter_area = st_area(geometry),
    weight = as.numeric(inter_area / sr_area)
  )
```

```
Warning: attribute variables are assumed to be spatially constant throughout
all geometries
```

```
for (v in vote_cols) {
  sr_cd[[v]] <- sr_cd[[v]] * sr_cd$weight
}

ab604_results <- sr_cd |>
  st_drop_geometry() |>
  group_by(.data[[cd_id_col]]) |>
  summarise(across(all_of(vote_cols), sum, na.rm = TRUE)) |>
  ungroup()
```

```
Warning: There was 1 warning in `summarise()`.
ℹ In argument: `across(all_of(vote_cols), sum, na.rm = TRUE)`.
ℹ In group 1: `district = "01"`.
Caused by warning:
! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
Supply arguments directly to `.fns` through an anonymous function instead.

  # Previously
  across(a:b, mean, na.rm = TRUE)

  # Now
  across(a:b, \(x) mean(x, na.rm = TRUE))
```

```
ab604_results
```

```
# A tibble: 52 × 3
   district uspdem01 usprep01
   <chr>       <dbl>    <dbl>
 1 01        243745.  223427.
 2 02        317707.  302416.
 3 03        203879.  195560.
 4 04        205020.  163946.
 5 05        243996.  317274.
 6 06        175022.  154837.
 7 07        187744.  158232.
 8 08        218202.  123559.
 9 09        157030.  122739.
10 10        236048.  133233.
# ℹ 42 more rows
```

```
write_csv(ab604_results, "data/ab604_cd_results.csv")
```