Data Cleaning

```
── 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   3.5.2     ✔ 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

Attaching package: 'janitor'


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

    chisq.test, fisher.test
```

```
# Read in raw datasets
precinct_raw <- read_csv("data/g24_sov_by_g24_svprec.csv") |>
  clean_names()
```

```
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.
```

```
candidates <- read_csv("data/g24-candidates-by-district.csv") |>
```

```
  clean_names()
```

```
Rows: 313 Columns: 5
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (5): CONTEST, FIELD, DISTRICT_TYPE, DISTRICT, NAME

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.
```

```
# Clean candidates table
candidates <- candidates |>
  mutate(
    race_type = case_when(
      str_detect(field, "ASS") ~ "Assembly",
      str_detect(field, "CNG") ~ "Congress",
      TRUE ~ "Other"
    ),
    candidate_party = case_when(
      str_detect(contest, "DEM") ~ "DEM",
      str_detect(contest, "REP") ~ "REP",
      str_detect(contest, "AI")  ~ "AIP",
      str_detect(contest, "GRN") ~ "GRN",
      str_detect(contest, "LIB") ~ "LIB",
      TRUE ~ "Other"
    ),
    contest = str_to_upper(contest),
    district = as.numeric(district)
  ) |>
  select(contest, field, district_type, district, name, candida

# Clean precinct dataset
precinct <- precinct_raw |>
  mutate(
    addist = as.numeric(addist),
    totvote = as.numeric(totvote),
    totreg = as.numeric(totreg),
    turnout_rate = if_else(totreg > 0, totvote / totreg, NA_rea
    turnout_rate = if_else(turnout_rate > 1, NA_real_, turnout_
  )
```

```r
# Reshape precinct data to long format
precinct_long <- precinct |>
  pivot_longer(
    cols = matches("(ass|cng).*\\d{2}$", ignore.case = TRUE),
    names_to = "contest",
    values_to = "votes"
  ) |>
  mutate(
    contest = str_to_upper(contest),
    votes = suppressWarnings(as.numeric(votes)),
    race_type = case_when(
      str_detect(contest, "ASS") ~ "Assembly",
      str_detect(contest, "CNG") ~ "Congress",
      TRUE ~ "Other"
    ),
    district = addist
  )

# Fix contest codes to match candidate file (e.g., ASS01DEM01)
precinct_long <- precinct_long |>
  mutate(
    district_str = str_pad(district, 2, pad = "0"),
    contest = case_when(
      str_detect(contest, "ASS") ~ paste0("ASS", district_str, s
      str_detect(contest, "CNG") ~ paste0("CNG", district_str, s
      TRUE ~ contest
    )
  )

# Join with candidates
precinct_long <- precinct_long |>
  left_join(candidates, by = c("contest", "district", "race_type
  rename(candidate_name = name)

# Check join results
precinct_long |> count(race_type, !is.na(candidate_name))
```

```
# A tibble: 4 × 3
  race_type `!is.na(candidate_name)`      n
  <chr>     <lgl>                     <int>
1 Assembly  FALSE                    156481
2 Assembly  TRUE                      99134
3 Congress  FALSE                    185563
4 Congress  TRUE                      70052
```

```
# Preview matched rows
precinct_long |>
  filter(!is.na(candidate_name)) |>
  select(race_type, contest, district, candidate_name, candidate
  head(15)
```

```
# A tibble: 15 × 6
   race_type contest    district candidate_name
candidate_party votes
   <chr>     <chr>         <dbl> <chr>          <chr>
<dbl>
 1 Assembly  ASS14DEM01       14 Margot Smith    DEM
94
 2 Assembly  ASS14DEM02       14 Buffy Wicks*    DEM
110
 3 Congress  CNG14DEM01       14 Eric Swalwell*  DEM
102
 4 Congress  CNG14REP01       14 Vin Kruttiventi REP
0
 5 Assembly  ASS14DEM01       14 Margot Smith    DEM
444
 6 Assembly  ASS14DEM02       14 Buffy Wicks*    DEM
2023
 7 Congress  CNG14DEM01       14 Eric Swalwell*  DEM
1668
 8 Congress  CNG14REP01       14 Vin Kruttiventi REP
0
 9 Assembly  ASS14DEM01       14 Margot Smith    DEM
117
10 Assembly  ASS14DEM02       14 Buffy Wicks*    DEM
91
11 Congress  CNG14DEM01       14 Eric Swalwell*  DEM
108
12 Congress  CNG14REP01       14 Vin Kruttiventi REP
0
13 Assembly  ASS14DEM01       14 Margot Smith    DEM
348
14 Assembly  ASS14DEM02       14 Buffy Wicks*    DEM
1243
15 Congress  CNG14DEM01       14 Eric Swalwell*  DEM
1063
```

```
# Save cleaned wide and long datasets
write_csv(precinct, "data/precinct_clean_wide.csv")
```

```
write_csv(precinct_long, "data/precinct_clean_long.csv")
```

```
library(sf)
```

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

```
# Load SR precinct-level voting data
votes_sr <- read.csv("data/shapefiles/state_g24_sov_data_by_g24_
  clean_names()

# Load SR precinct shapefile
sr_shp <- st_read("data/shapefiles/srprec_state_g24_v01_shp/srp
```

```
Reading layer `srprec_state_g24_v01_shp' from data source

`/Users/timroth/Documents/UC_Berkeley/Cal_Senior/Senior_Fall/st
at133/gerrymandering-
timrothtr/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/timroth/Documents/UC_Berkeley/Cal_Senior/Senior_Fall/sta
t133/gerrymandering-
timrothtr/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 24145 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
```

```
# Load proposed AB 604 district shapefile
ab604_shp <- st_read("data/shapefiles/AB604/AB604.shp")
```

```
Reading layer `AB604' from data source

`/Users/timroth/Documents/UC_Berkeley/Cal_Senior/Senior_Fall/st
at133/gerrymandering-timrothtr/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
```

```
# Transform to same projection and fix geometry issues
sr_shp <- sr_shp |>
  st_transform(3310) |>
  st_set_precision(1) |>
  st_make_valid() |>
  st_collection_extract("POLYGON")

ab604_shp <- ab604_shp |>
  st_transform(3310) |>
  st_make_valid()

# Check column names for join key
names(votes_sr)
```

```
 [1] "county"     "fips"       "srprec"     "election"
"srprec_key"
 [6] "geo_type"   "addist"     "cddist"     "sddist"
"bedist"
[11] "totreg"     "demreg"     "repreg"     "aipreg"
"grnreg"
[16] "libreg"     "nlpreg"     "refreg"     "dclreg"
"mscreg"
[21] "totvote"    "demvote"    "repvote"    "aipvote"
"grnvote"
[26] "libvote"    "nlpvote"    "refvote"    "dclvote"
"mscvote"
[31] "prcvote"    "absvote"    "assaip01"   "assdem01"
"assdem02"
[36] "assrep01"   "assrep02"   "cngdem01"   "cngdem02"
"cngind01"
[41] "cngrep01"   "cngrep02"   "prsaip01"   "prsdem01"
"prsgrn01"
```

```
[46] "prslib01"   "prspaf01"   "prsrep01"   "pr_2_n"
"pr_2_y"
[51] "pr_32_n"    "pr_32_y"    "pr_33_n"    "pr_33_y"
"pr_34_n"
[56] "pr_34_y"    "pr_35_n"    "pr_35_y"    "pr_36_n"
"pr_36_y"
[61] "pr_3_n"     "pr_3_y"     "pr_4_n"     "pr_4_y"
"pr_5_n"
[66] "pr_5_y"     "pr_6_n"     "pr_6_y"     "sendem01"
"sendem02"
[71] "senrep01"   "senrep02"   "uspdem01"   "usprep01"
"ussdem01"
[76] "ussrep01"
```

```
# Join votes to SR precinct geometry
sr_geo <- sr_shp |>
  left_join(votes_sr, by = c("SRPREC" = "srprec")) |>
  clean_names() |>
  select(!matches("\\.1$"))
```

```
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.
```

```
# Save cleaned versions
st_write(sr_geo, "data/shapefiles/srprecinct_cleaned.shp", dele
```

```
Warning in abbreviate_shapefile_names(obj): Field names
abbreviated for ESRI
Shapefile driver

Deleting source `data/shapefiles/srprecinct_cleaned.shp' using
driver `ESRI Shapefile'
Writing layer `srprecinct_cleaned' to data source
  `data/shapefiles/srprecinct_cleaned.shp' using driver `ESRI
Shapefile'
Writing 26993 features with 81 fields and geometry type Multi
Polygon.
```

```
st_write(ab604_shp, "data/shapefiles/ab604_cleaned.shp", delete_
```

```
Deleting source `data/shapefiles/ab604_cleaned.shp' using
driver `ESRI Shapefile'
Writing layer `ab604_cleaned' to data source
  `data/shapefiles/ab604_cleaned.shp' using driver `ESRI
Shapefile'
Writing 52 features with 15 fields and geometry type Unknown
(any).
```

```
write_csv(votes_sr, "data/srprecinct_votes_cleaned.csv")
```