Data Cleaning

Imports

```
library(dplyr)
```

```

Attaching package: 'dplyr'
```

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

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(stringr)
library(sf)
```

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

```
library(ggplot2)
```

Cleaning sv vote count data by precinct

```
#read csv data
vote_counts = read.csv("data/g24_sov_by_g24_svprec.csv")

#remove rows with missing values and non precint data rows
clean_vote_counts = vote_counts |>
    filter(str_detect(CNGDEM01, "^\\d+$")) |>
    filter(ADDIST != 0)

#define function to check if column is numeric
is_numeric = function(x) {
    if (!is.character(x)) return(FALSE)
    sum(str_detect(x, "^\\d+$")) == length(x)
}
```

```
#convert rows that are fully numeric to numeric
clean_vote_counts = clean_vote_counts |>
    mutate(across(.cols = where(is_numeric), .fns = as.integer))

saveRDS(clean_vote_counts, "data/clean_vote_counts")
```

Cleaning sr vote count data

```
#read csv data
sr_votes = read.csv("data/state_g24_sov_data_by_g24_srprec.csv")

#remove rows with missing values and non precint data rows
clean_sr_votes = sr_votes |>
    filter(str_detect(CNGDEM01, "^\\d+$")) |>
    filter(ADDIST != 0)

#convert rows that are fully numeric to numeric
clean_sr_votes = clean_sr_votes |>
    mutate(across(.cols = where(is_numeric), .fns = as.integer))

saveRDS(clean_sr_votes, "data/clean_sr_votes")
```

Load geometries

```
#load geometries
prop_ca_prec = st_read("data/shapefiles/AB604/AB604.shp")
```

```
Reading layer `AB604' from data source
  `/Users/brycen/Documents/stat133/gerrymandering-brycenm7/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
```

```
ca_prec = 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/brycen/Documents/stat133/gerrymandering-brycenm7/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/brycen/Documents/stat133/gerrymandering-brycenm7/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
```

Join geometries

```
#correct errors
ca_prec <- ca_prec |>
 st_transform(crs = 3310) |> # switch to equal area projection first
 st_set_precision(1) |> # snap points to 1 m grid
 st_make_valid() |> # fix self-intersections/bow-ties
 st_collection_extract("POLYGON")

prop_ca_prec = prop_ca_prec |>
    st_transform(crs = 3310)

#join current precinct shape with vote data
sr_geom = ca_prec |>
    left_join(clean_sr_votes, by = "SRPREC_KEY")

#find intersection areas of current and proposed precincts
intersection = st_intersection(sr_geom, prop_ca_prec) |>
    mutate(areas = st_area(geometry))
```

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

```
weights = intersection |>
    group_by(SRPREC_KEY) |>
    mutate(original_area = sum(areas), area_weight = areas / original_area)
```

```
weighted_votes = weights |>
    mutate(CNGDEM01 = as.integer(CNGDEM01 * area_weight), CNGREP01 =
as.integer(CNGREP01 * area_weight))

saveRDS(weighted_votes, "data/weighted_votes")
```