Data Cleaning

`Read in the vote count data from the 2024 General Election at the precinct level. Clean the c`

- Read the website of the statewide database carefully for disclaimers about the state of the data.
- Inspect columns to ensure they're the correct type. If they're not, investigate why not and correct it.
- Inspect the values in each column to ensure they have the values that you would expect.
- Check that columns that you expect to have unique values indeed don't have replicates.

Tips

#read in relevant libraries library(dplyr) library(readr) library(sf) statement_of_vote2024 = read.csv('g24_sov_by_g24_svprec.csv') votes_by_sr_2024 = read.csv('state_g24_sov_data_by_g24_srprec. g24_shp <- st_read("/Users/yeonwoo/Downloads/stat133/gerrymandering-yeonwoolee-sudo-main/data/mprec_state_g24_v01_shp") sr_prec_shp <- st_read("/Users/yeonwoo/Downloads/stat133/gerry yeonwoolee-sudo-main/data/srprec_state_g24_v01_shp/") proposed_map <- st_read("/Users/yeonwoo/Dow yeonwoolee-sudo-main/data/AB604/") congress_statement_of_vote2024 <- statement_of_vote2024 |> select(SVPREC, SVPREC_KEY, CDDIST, TOTREG, CNGDEM01, CNGDEM02, CNGREP01, CNGREP02) total_congress_statement_of_vote2024 <- congress_statement_of_vote2024 |> group_by(CDDIST) |> # group by congressional district summarize(CNGDEM01 = sum(as.numeric(CNGDEM01), na.rm = TRUE), CNGDEM02 = sum(as.numeric(CNGDEM02), na.rm = TRUE), CNGREP01 = sum(as.numeric(CNGREP01), na.rm = TRUE), CNGREP02 = sum(as.numeric(CNGREP02), na.rm = TRUE), total_registered_voters = sum(as.numeric(TOTREG), na.rm = TRUE), total_cngdem_votes = CNGDEM01 + CNGDEM02, total_cngrep_votes = CNGREP01 + CNGREP02) |> # seeing total votes per party to assign a party affiliation per voting district mutate(winning_party = ifelse(total_cngdem_votes > total_cngrep_votes, "DEM", "REP") # classify each district per party)

cleaned_sr_votes_2024 <- votes_by_sr_2024 |> select(SRPREC, SRPREC_KEY,CDDIST, TOTREG, CNGDEM01, CNGDEM02, CNGREP01, CNGREP02) |> mutate(CNGDEM01 = as.numeric(CNGDEM01), CNGDEM02 = as.numeric(CNGDEM02), CNGREP01 =

as.numeric(CNGREP01), CNGREP02 = as.numeric(CNGREP02), total_cngdem_votes = CNGDEM01 + CNGDEM02, total_cngrep_votes = CNGREP01 + CNGREP02)