Shelby Data Wrangling

Jack Holland

2025-11-21

```
library(sf)
```

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

```
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(ggplot2)
library(stringr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v lubridate 1.9.4     v tibble    3.3.0
## v purrr     1.0.4     v tidyr     1.3.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(estimatr)
library(haven)
library(stargazer)
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
select <- dplyr::select
```

```r
s97_path  <- "/Users/jackholland/Downloads/97th cong/districts097.shp"
s103_path <- "/Users/jackholland/Downloads/districtShapes/districts103.shp"

s97  <- st_read(s97_path)
```

```
## Reading layer ‘districts097’ from data source
##   ‘/Users/jackholland/Downloads/97th cong/districts097.shp’ using driver ‘ESRI Shapefile’
## Simple feature collection with 435 features and 15 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -178.3381 ymin: 18.91071 xmax: 177.7388 ymax: 71.3857
## Geodetic CRS:  NAD83
```

```r
s103 <- st_read(s103_path)
```

```
## Reading layer ‘districts103’ from data source
##   ‘/Users/jackholland/Downloads/districtShapes/districts103.shp’
##   using driver ‘ESRI Shapefile’

## Warning in CPL_read_ogr(dsn, layer, query, as.character(options), quiet, : GDAL
## Message 1: /Users/jackholland/Downloads/districtShapes/districts103.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 436 features and 15 fields (with 1 geometry empty)
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -179.1473 ymin: 18.9177 xmax: 179.7785 ymax: 71.35256
## Geodetic CRS:  NAD83
```

```r
south_states <- c(
"Texas","Oklahoma","Arkansas","Louisiana","Mississippi","Tennessee",
"Georgia","Alabama","North Carolina","South Carolina","Virginia","Florida"
)

normalize_state_names <- function(x) {
str_replace_all(x, c("^Commonwealth of " = "", "^State of " = ""))
}

state_key <- tibble::tibble(
STATENAME  = state.name,
STATE_ABBR = state.abb
)

prep_south <- function(sf_obj) {
sf_obj %>%
mutate(STATENAME = normalize_state_names(STATENAME)) %>%
filter(STATENAME %in% south_states) %>%
left_join(state_key, by = "STATENAME") %>%
```

```
mutate(
DISTRICT_NUM = suppressWarnings(as.integer(as.character(DISTRICT))),
DISTRICT_NUM = ifelse(is.na(DISTRICT_NUM), as.integer(DISTRICT), DISTRICT_NUM),
district_id  = paste0(STATE_ABBR, "-", DISTRICT_NUM)
) %>%
st_make_valid() %>%
st_transform(2163)
}

south97  <- prep_south(s97)
```

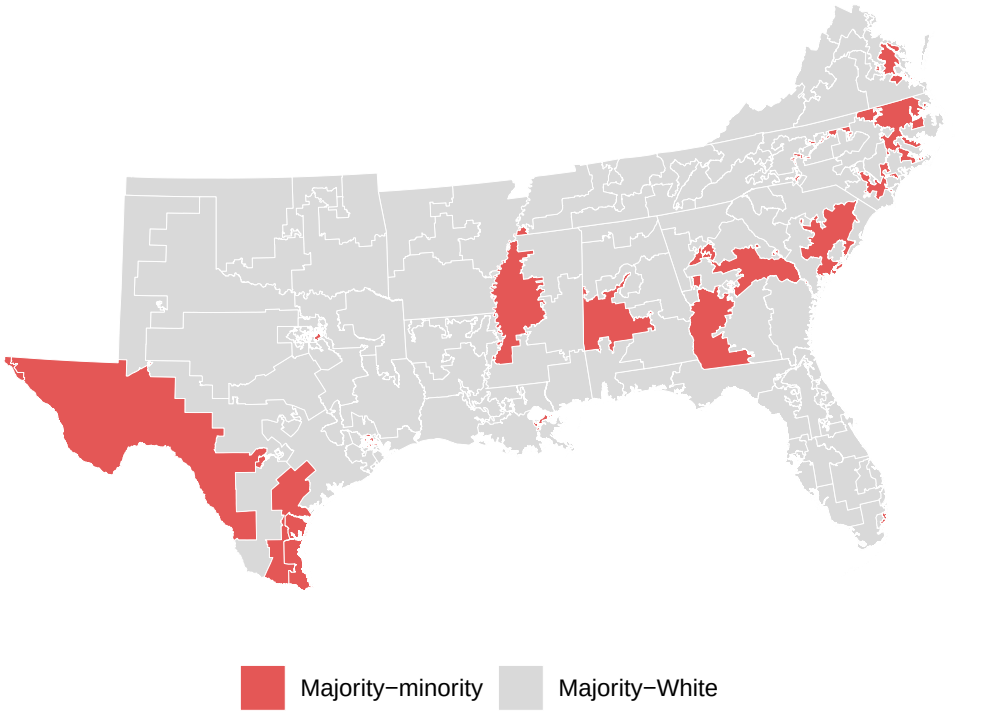
```
## Warning in CPL_crs_from_input(x): GDAL Message 1: CRS EPSG:2163 is deprecated.
## Its non-deprecated replacement EPSG:9311 will be used instead. To use the
## original CRS, set the OSR_USE_NON_DEPRECATED configuration option to NO.
```

```
south103 <- prep_south(s103)

mm_103_south <- c(
"AL-7",
"FL-17","FL-18",
"GA-2","GA-4","GA-5","GA-11",
"LA-2",
"MS-2",
"NC-1","NC-12",
"SC-6",
"TN-9",
"TX-15","TX-16","TX-18","TX-20","TX-23","TX-27","TX-29","TX-30",
"VA-3"
)

south97  <- south97  %>% mutate(maj_minority = FALSE)
south103 <- south103 %>% mutate(maj_minority = district_id %in% mm_103_south)
```

```
# 97th Congress map (no majority-minority districts)

p97 <- ggplot(south97) +
geom_sf(fill = "grey85", color = "white", linewidth = 0.15) +
labs(
title   = "97th Congress - South (1981-1983)",
caption = "Source: US Census Bureau, Data: Jeffrey B. Lewis"
) +
coord_sf(datum = NA) +
theme_minimal(base_size = 11) +
theme(
plot.title = element_text(hjust = 0.5, face = "bold"),
axis.text  = element_blank(),
axis.title = element_blank(),
panel.grid = element_blank()
)

# 103rd Congress map with majority-minority highlighted
```

```
p103 <- ggplot(south103) +
geom_sf(aes(fill = maj_minority), color = "white", linewidth = 0.15) +
scale_fill_manual(
values = c(`TRUE` = "#e45756", `FALSE` = "grey85"),
labels = c(`TRUE` = "Majority-minority", `FALSE` = "Majority-White"),
breaks = c(TRUE, FALSE),
name   = NULL
) +
labs(
title   = "103rd Congress - South (1993-1994)",
caption = "Source: US Census Bureau, Data: Jeffrey B. Lewis"
) +
coord_sf(datum = NA) +
theme_minimal(base_size = 11) +
theme(
plot.title      = element_text(hjust = 0.5, face = "bold"),
legend.position = "bottom",
axis.text       = element_blank(),
axis.title      = element_blank(),
panel.grid      = element_blank()
)

p97
```

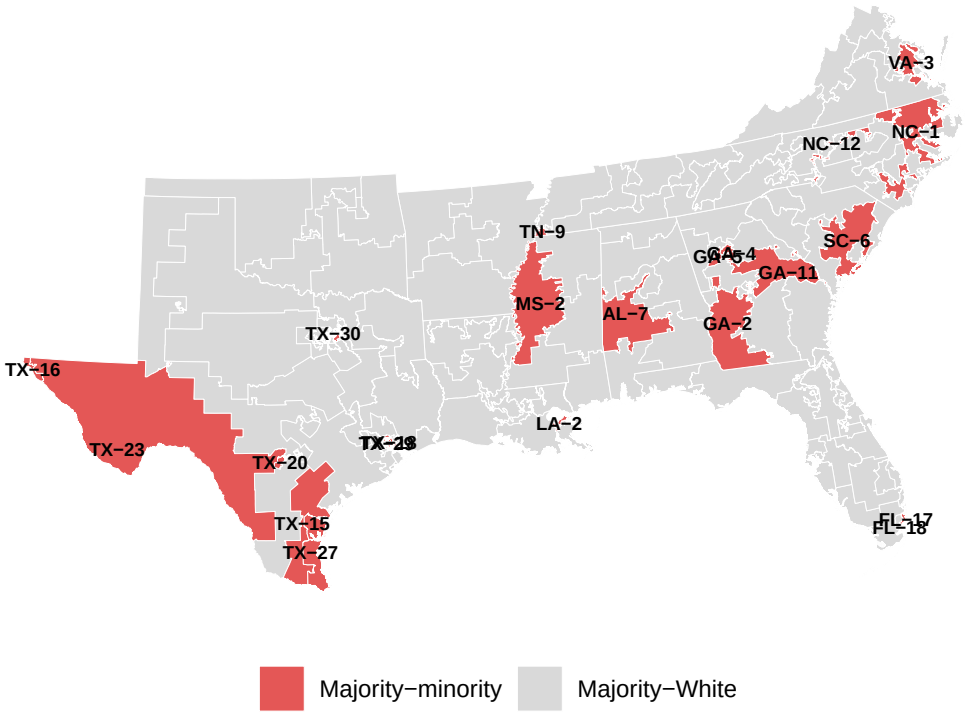
```
p103
```




```
# Label majority-minority districts in the 103rd Congress map

label_pts <- st_point_on_surface(south103 %>% filter(maj_minority))
```

```
## Warning: st_point_on_surface assumes attributes are constant over geometries
```

```
p103_labels <- p103 +
geom_sf_text(
data = label_pts,
aes(label = district_id),
size     = 2.5,
fontface = "bold",
color    = "black"
)

p103_labels
```

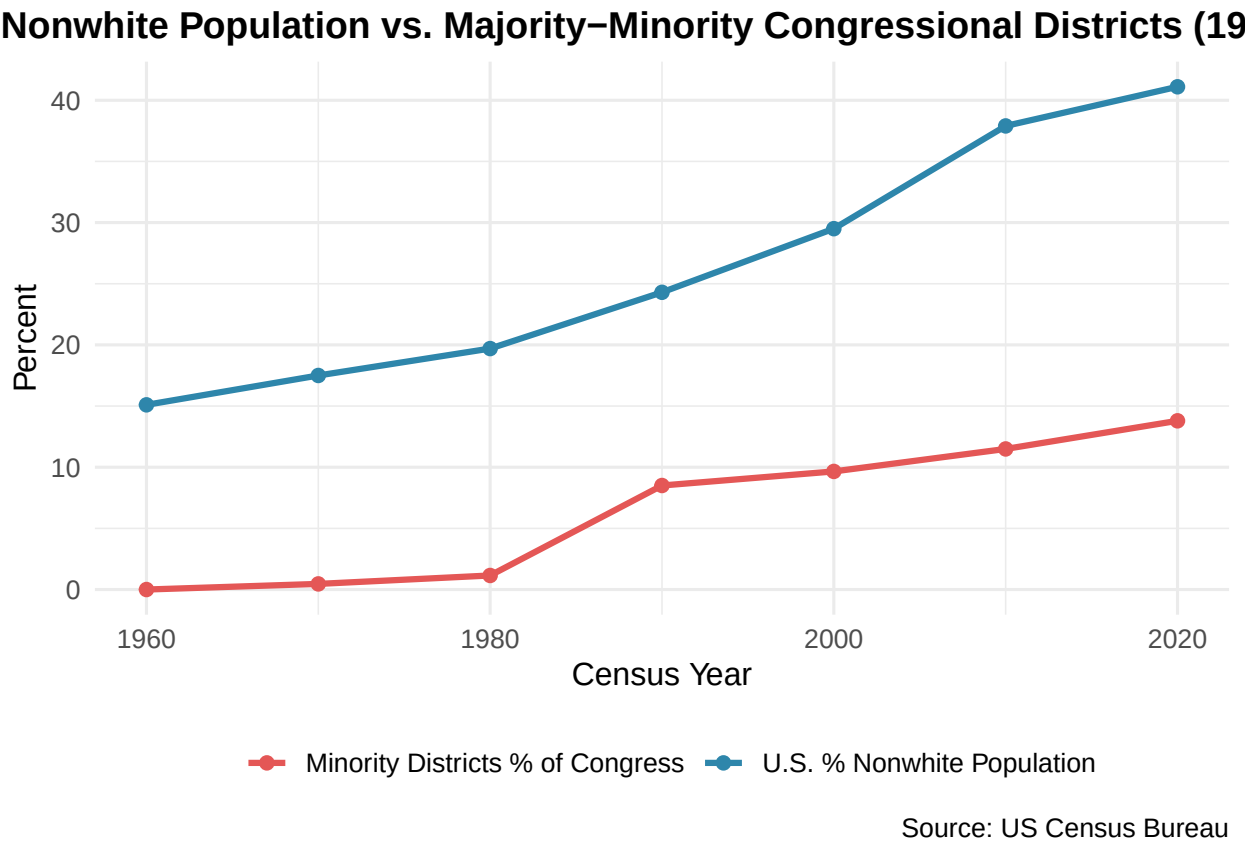



```
maj_min <- read.csv("/Users/jackholland/Downloads/majority_minority_districts_by_decade.csv") |>
mutate(Minority_Districts_Pct_of_Congress = (Majority_Minority_Districts / 435) * 100)

maj_min_long <- maj_min |>
pivot_longer(
cols      = c(Percent_Nonwhite_US, Minority_Districts_Pct_of_Congress),
names_to  = "Measure",
values_to = "Percent"
)

maj_min_chart <- ggplot(maj_min_long, aes(x = Census_Year, y = Percent, color = Measure)) +
geom_line(size = 1.1) +
geom_point(size = 2) +
scale_color_manual(
values = c(
"Percent_Nonwhite_US"                = "#2E86AB",
"Minority_Districts_Pct_of_Congress" = "#E45756"
),
labels = c(
"Percent_Nonwhite_US"                = "U.S. % Nonwhite Population",
"Minority_Districts_Pct_of_Congress" = "Minority Districts % of Congress"
)
) +
labs(
title   = "U.S. Nonwhite Population vs. Majority-Minority Congressional Districts (1960-2020)",
x       = "Census Year",
```

```
y       = "Percent",
color   = "",
caption = "Source: US Census Bureau"
) +
theme_minimal(base_size = 12) +
theme(
plot.title      = element_text(hjust = 0.5, face = "bold"),
legend.position = "bottom"
)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
maj_min_chart
```




```
maj_min_deriv <- maj_min_long |>
  group_by(Measure) |>
  arrange(Census_Year) |>
  mutate(
    dPercent        = Percent - lag(Percent),
    dYear           = Census_Year - lag(Census_Year),
```

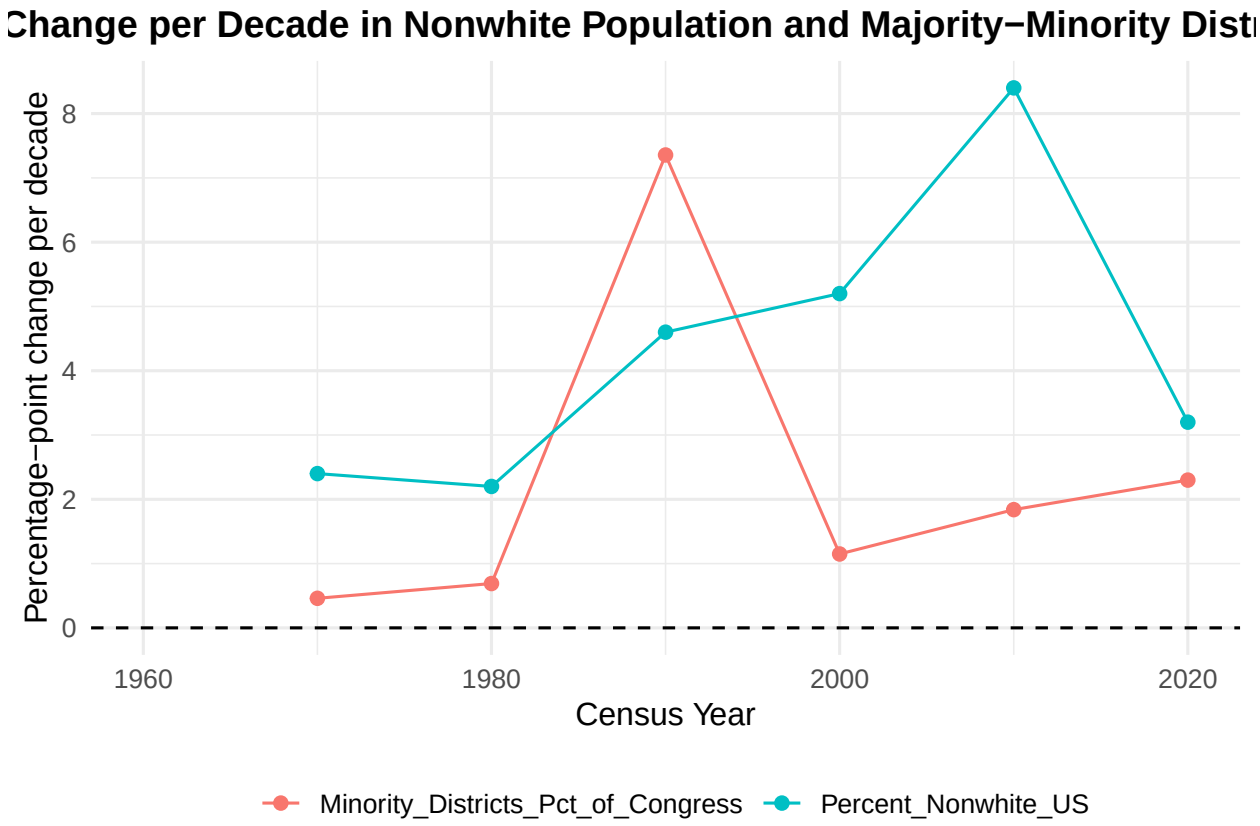
```
    slope_per_year  = dPercent / dYear,          # approximate derivative
    slope_per_decade = slope_per_year * 10       # change per 10 years
  )

maj_min_deriv
```

```
## # A tibble: 14 x 10
## # Groups:   Measure [2]
##    Census_Year Congress_Cycle  Majority_Minority_Districts Notes Measure Percent
##          <int> <chr>                                 <int> <chr> <chr>     <dbl>
##  1        1960 87th (1961-63)                            0 No m~ Percen~  15.1
##  2        1960 87th (1961-63)                            0 No m~ Minori~   0
##  3        1970 92nd (1971-73)                            2 A fe~ Percen~  17.5
##  4        1970 92nd (1971-73)                            2 A fe~ Minori~   0.460
##  5        1980 97th (1981-83)                            5 Firs~ Percen~  19.7
##  6        1980 97th (1981-83)                            5 Firs~ Minori~   1.15
##  7        1990 103rd (1993-95)                          37 Majo~ Percen~  24.3
##  8        1990 103rd (1993-95)                          37 Majo~ Minori~   8.51
##  9        2000 107th (2001-03)                          42 Stab~ Percen~  29.5
## 10        2000 107th (2001-03)                          42 Stab~ Minori~   9.66
## 11        2010 113th (2013-15)                          50 Rise~ Percen~  37.9
## 12        2010 113th (2013-15)                          50 Rise~ Minori~  11.5
## 13        2020 117th (2021-23)                          60 Nonw~ Percen~  41.1
## 14        2020 117th (2021-23)                          60 Nonw~ Minori~  13.8
## # i 4 more variables: dPercent <dbl>, dYear <int>, slope_per_year <dbl>,
## #   slope_per_decade <dbl>
```

```
ggplot(maj_min_deriv, aes(x = Census_Year, y = slope_per_decade, color = Measure)) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_point(size = 2) +
  geom_line() +
  labs(
    title = "Change per Decade in Nonwhite Population and Majority-Minority Districts",
    x = "Census Year",
    y = "Percentage-point change per decade",
    color = ""
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title      = element_text(hjust = 0.5, face = "bold"),
    legend.position = "bottom"
  )
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_line()`).
```




```
maj_min_filtered <- maj_min |>
filter(Census_Year >= 1980 & Census_Year <= 2000) |>
pivot_longer(
cols      = c(Percent_Nonwhite_US, Minority_Districts_Pct_of_Congress),
names_to  = "Measure",
values_to = "Percent"
)

maj_min_gingles_chart <- ggplot(maj_min_filtered, aes(x = Census_Year, y = Percent, color = Measure)) +
geom_point(size = 3) +
geom_smooth(method = "loess", se = FALSE, linetype = "solid", linewidth = 1) +
geom_vline(xintercept = 1986, color = "grey", linetype = "dashed", linewidth = 0.8) +
annotate(
"text",
x      = 1986,
y      = max(maj_min_filtered$Percent) + 1.5,
label  = "1986\nThornburg v. Gingles",
angle  = 90,
vjust  = -0.5,
hjust  = 0,
size   = 3.3,
color  = "black"
) +
scale_color_manual(
values = c(
"Percent_Nonwhite_US"                = "#2E86AB",
```

```
"Minority_Districts_Pct_of_Congress" = "#E45756"
),
labels = c(
"Percent_Nonwhite_US"                = "U.S. % Nonwhite Population",
"Minority_Districts_Pct_of_Congress" = "Minority Districts % of Congress"
)
) +
labs(
title = "Impact of Thornburg v. Gingles (1986) on Majority-Minority Representation",
x     = "Census Year",
y     = "Percent",
color = ""
) +
scale_x_continuous(breaks = seq(1980, 2000, by = 5)) +
theme_minimal(base_size = 12) +
theme(
plot.title      = element_text(hjust = 0.5, face = "bold"),
legend.position = "bottom"
)

maj_min_gingles_chart
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : span too small.  fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 1979.9
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 10.1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 102.01
```

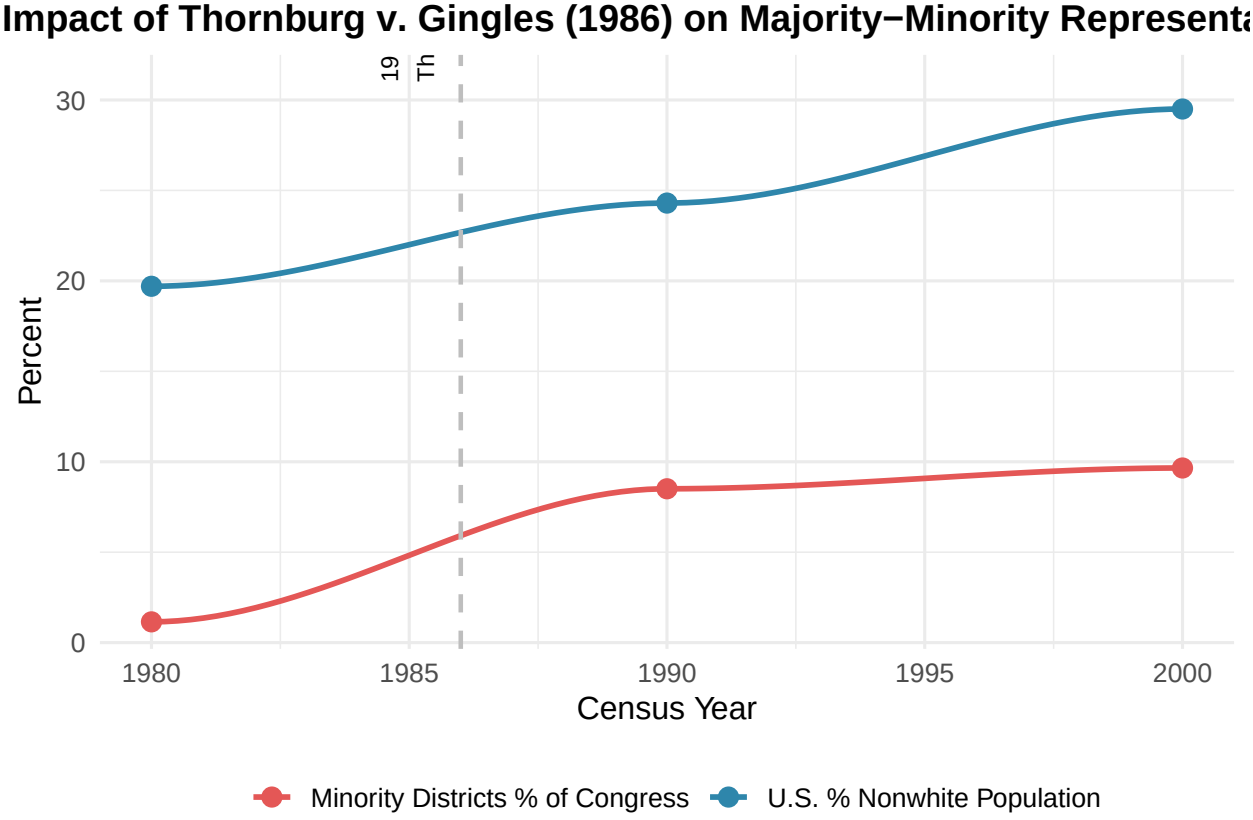
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : span too small.  fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 1979.9
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 10.1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 102.01
```

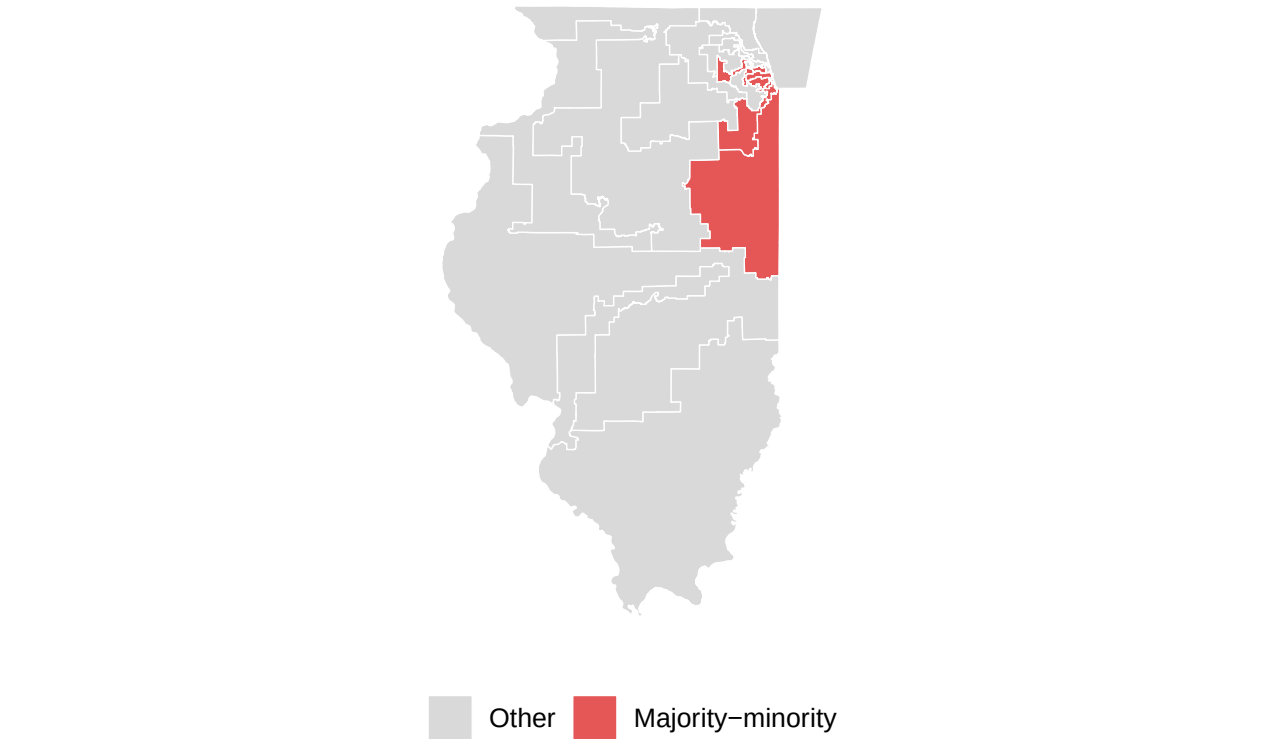



```
# Update the path to your TL shapefile as needed

il_cong <- st_read("/Users/jackholland/Downloads/tl_2022_17_cd118/tl_2022_17_cd118.shp")
```

```
## Reading layer 'tl_2022_17_cd118' from data source
##   '/Users/jackholland/Downloads/tl_2022_17_cd118/tl_2022_17_cd118.shp'
##   using driver 'ESRI Shapefile'
## Simple feature collection with 18 features and 12 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: -91.51308 ymin: 36.9703 xmax: -87.01994 ymax: 42.50848
## Geodetic CRS:  NAD83
```

```
il <- il_cong %>%
mutate(
DISTRICT_NUM = suppressWarnings(as.integer(as.character(CD118FP))),
maj_minority = DISTRICT_NUM %in% c(1, 2, 3, 4, 7),
district_lab = paste0("IL-", DISTRICT_NUM)
)

ggplot(il) +
geom_sf(aes(fill = maj_minority), color = "white", linewidth = 0.25) +
```

```
scale_fill_manual(
values = c(`TRUE` = "#e45756", `FALSE` = "grey85"),
labels = c(`TRUE` = "Majority-minority", `FALSE` = "Other"),
name   = NULL
) +
coord_sf(datum = NA) +
labs(
title = "Illinois Congressional Districts (118th) - Majority-Minority Highlighted"
) +
theme_minimal(base_size = 12) +
theme(
plot.title      = element_text(hjust = 0.5, face = "bold"),
legend.position = "bottom"
)
```




```
# If you want labels on majority-minority districts:

# library(ggrepel)

# lab_pts <- st_point_on_surface(il %>% filter(maj_minority))

# last_plot() +

# geom_sf_text(data = lab_pts, aes(label = district_lab), size = 3, fontface = "bold")
```

```r
library(sf)
library(dplyr)
library(readr)
library(stringr)
library(tidyr)
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
il_precincts <- st_read('/Users/jackholland/Downloads/School/Thesis/Section 5 Preclearance/Shelby/il_sha
```

```
## Reading layer `tl_2020_17_vtd20' from data source
##   `/Users/jackholland/Downloads/School/Thesis/Section 5 Preclearance/Shelby/il_shapefile/tl_2020_17_
##   using driver `ESRI Shapefile'
## Simple feature collection with 10084 features and 14 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -91.51308 ymin: 36.9703 xmax: -87.01993 ymax: 42.50848
## Geodetic CRS:  NAD83
```

```r
biden_precincts <- read_csv(
  "biden_precincts.csv"
)
```

```
## Rows: 63063 Columns: 12
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): JurisName, CandidateName, ContestName, PrecinctName, PartyName
## dbl (7): JurisdictionID, JurisContainerID, EISCandidateID, EISContestID, Reg...
##
## 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
biden_agg <- biden_precincts %>%
  filter(ContestName == "PRESIDENT AND VICE PRESIDENT") %>%
  mutate(
    prec_key = PrecinctName %>%
      str_to_upper() %>%
      str_squish(),
```

```
    cand_std = case_when(
      str_detect(CandidateName, "BIDEN") ~ "BIDEN",
      str_detect(CandidateName, "TRUMP") ~ "TRUMP",
      TRUE                               ~ "OTHER"
    )
  ) %>%
  group_by(prec_key, cand_std) %>%
  summarise(votes = sum(VoteCount, na.rm = TRUE), .groups = "drop") %>%
  pivot_wider(
    names_from  = cand_std,
    values_from = votes,
    values_fill = 0
  ) %>%
  mutate(
    total_votes_all   = BIDEN + TRUMP + OTHER,
    total_two_party   = BIDEN + TRUMP,
    biden_share       = if_else(total_two_party > 0, BIDEN / total_two_party, NA_real_)
  )

il_precincts_clean <- il_precincts %>%
  mutate(
    prec_key = NAME20 %>%
      str_to_upper() %>%
      str_squish()
  )

il_precincts_joined <- il_precincts_clean %>%
  left_join(biden_agg, by = "prec_key")

il_precincts_joined %>%
  summarise(
    n_geoms          = n(),
    n_with_data      = sum(!is.na(biden_share)),
    n_missing_data   = sum(is.na(biden_share))
  )
```

```
## Simple feature collection with 1 feature and 3 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: -91.51308 ymin: 36.9703 xmax: -87.01993 ymax: 42.50848
## Geodetic CRS:  NAD83
##   n_geoms n_with_data n_missing_data                       geometry
## 1   10084        4155           5929 POLYGON ((-89.52156 37.5721...
```

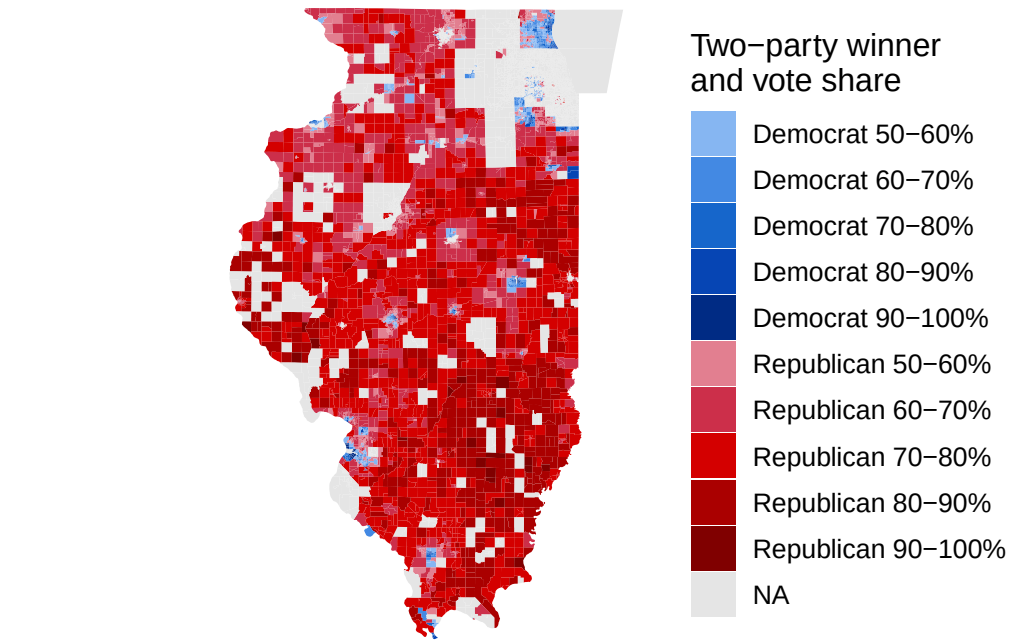
```
library(dplyr)
library(stringr)

il_precincts_colored <- il_precincts_joined %>%
  mutate(
    biden_pct = biden_share * 100,
    trump_pct = 100 - biden_pct,
    winner    = case_when(
      is.na(biden_pct)           ~ NA_character_,
```

```
      biden_pct > 50              ~ "Democrat",
      biden_pct < 50              ~ "Republican",
      TRUE                        ~ "Tie"
    ),
    pct_for_bins = case_when(
      winner == "Democrat"   ~ biden_pct,
      winner == "Republican" ~ trump_pct,
      TRUE                   ~ NA_real_
    ),
    # bin into 20-30, 30-40, ... 90-100
    bin = case_when(
      pct_for_bins >= 20 & pct_for_bins < 30 ~ "20-30%",
      pct_for_bins >= 30 & pct_for_bins < 40 ~ "30-40%",
      pct_for_bins >= 40 & pct_for_bins < 50 ~ "40-50%",
      pct_for_bins >= 50 & pct_for_bins < 60 ~ "50-60%",
      pct_for_bins >= 60 & pct_for_bins < 70 ~ "60-70%",
      pct_for_bins >= 70 & pct_for_bins < 80 ~ "70-80%",
      pct_for_bins >= 80 & pct_for_bins < 90 ~ "80-90%",
      pct_for_bins >= 90 & pct_for_bins <= 100 ~ "90-100%",
      TRUE ~ NA_character_
    ),
    fill_cat = case_when(
      is.na(winner) | is.na(bin) ~ NA_character_,
      TRUE                       ~ paste(winner, bin)
    )
  )
```

```
dem_palette <- c(
  "Democrat 20-30%" = "#E1EFFF",
  "Democrat 30-40%" = "#D3E7FF",
  "Democrat 40-50%" = "#B9D7FF",
  "Democrat 50-60%" = "#86B6F2",
  "Democrat 60-70%" = "#4389E3",
  "Democrat 70-80%" = "#1666CB",
  "Democrat 80-90%" = "#0645B4",
  "Democrat 90-100%" = "#002B84"
)

rep_palette <- c(
  "Republican 20-30%" = "#FFDFE1",
  "Republican 30-40%" = "#FFCCD0",
  "Republican 40-50%" = "#F2B3BE",
  "Republican 50-60%" = "#E27F90",
  "Republican 60-70%" = "#CC2F4A",
  "Republican 70-80%" = "#D40000",
  "Republican 80-90%" = "#AA0000",
  "Republican 90-100%" = "#800000"
)

party_palette <- c(dem_palette, rep_palette)
```

```
library(ggplot2)
library(scales)
```

```r
ggplot(il_precincts_colored) +
  geom_sf(aes(fill = fill_cat), color = NA) +
  scale_fill_manual(
    values   = party_palette,
    na.value = "grey90",
    name     = "Two-party winner\nand vote share"
  ) +
  coord_sf(datum = NA) +
  labs(
    title    = "2020 Presidential Election in Illinois by Precinct",
    subtitle = "Biden vs. Trump, binned by two-party vote share",
    caption  = "Source: 2020 precinct-level election returns"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title      = element_text(hjust = 0.5, face = "bold"),
    legend.position = "right",
    axis.text       = element_blank(),
    axis.title      = element_blank(),
    panel.grid      = element_blank()
  )
```




```r
library(sf)
library(ggplot2)
library(paletteer)
```

```
nc_1992_senate_shp <- st_read('~/Downloads/School/Thesis/Section 5 Preclearance/Shelby/NC State Senate
```

```
## Reading layer ‘1992_Senate_Base_Plan_6’ from data source
##   ‘/Users/jackholland/Downloads/School/Thesis/Section 5 Preclearance/Shelby/NC State Senate 1992/199
##   using driver ‘ESRI Shapefile’
## replacing null geometries with empty geometries
## Simple feature collection with 43 features and 4 fields (with 1 geometry empty)
## Geometry type: GEOMETRY
## Dimension:     XY
## Bounding box:  xmin: 123998.5 ymin: 822.5378 xmax: 935803.8 ymax: 318095.3
## Projected CRS: NAD83 / North Carolina
```

```
ggplot(nc_1992_senate_shp) +
  geom_sf() +
  theme_void() +
  labs(
    title = "North Carolina State Senate Districts, 1992"
  )
```