Tidying US Census dataset

Farhod Ibragimov

2025-03-04

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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 error
```

```r
library(usmap)
```

```
## Warning: package 'usmap' was built under R version 4.4.3
```

```r
library(dplyr)
```

Tidying dataset

```r
census_data <- read.csv("https://raw.githubusercontent.com/farhodibr/CUNY-SPS-MSDS/refs/heads/main/DATA

census_data <- census_data |>
  select(-Fact.Note)
glimpse(census_data)
```

```
## Rows: 85
## Columns: 51
## $ Fact           <chr> "Population estimates, July 1, 2016,  (V2016)", "Popula~
## $ Alabama        <chr> "4,863,300", "4,780,131", "1.70%", "4,779,736", "6.00%"~
## $ Alaska         <chr> "741,894", "710,249", "4.50%", "710,231", "7.30%", "7.6~
## $ Arizona        <chr> "6,931,071", "6,392,301", "8.40%", "6,392,017", "6.30%"~
## $ Arkansas       <chr> "2,988,248", "2,916,025", "2.50%", "2,915,918", "6.40%"~
## $ California     <chr> "39,250,017", "37,254,522", "5.40%", "37,253,956", "6.3~
## $ Colorado       <chr> "5,540,545", "5,029,324", "10.20%", "5,029,196", "6.10%~
## $ Connecticut    <chr> "3,576,452", "3,574,114", "0.10%", "3,574,097", "5.20%"~
## $ Delaware       <chr> "952,065", "897,936", "6.00%", "897,934", "5.80%", "6.2~
## $ Florida        <chr> "20,612,439", "18,804,592", "9.60%", "18,801,310", "5.5~
## $ Georgia        <chr> "10,310,371", "9,688,680", "6.40%", "9,687,653", "6.40%~
## $ Hawaii         <chr> "1,428,557", "1,360,301", "5.00%", "1,360,301", "6.40%"~
## $ Idaho          <chr> "1,683,140", "1,567,650", "7.40%", "1,567,582", "6.80%"~
## $ Illinois       <chr> "12,801,539", "12,831,574", "-0.20%", "12,830,632", "6.~
## $ Indiana        <chr> "6,633,053", "6,484,136", "2.30%", "6,483,802", "6.40%"~
## $ Iowa           <chr> "3,134,693", "3,046,869", "2.90%", "3,046,355", "6.40%"~
## $ Kansas         <chr> "2,907,289", "2,853,129", "1.90%", "2,853,118", "6.70%"~
## $ Kentucky       <chr> "4,436,974", "4,339,344", "2.20%", "4,339,367", "6.20%"~
## $ Louisiana      <chr> "4,681,666", "4,533,479", "3.30%", "4,533,372", "6.60%"~
## $ Maine          <chr> "1,331,479", "1,328,364", "0.20%", "1,328,361", "4.90%"~
## $ Maryland       <chr> "6,016,447", "5,773,786", "4.20%", "5,773,552", "6.10%"~
## $ Massachusetts  <chr> "6,811,779", "6,547,813", "4.00%", "6,547,629", "5.30%"~
## $ Michigan       <chr> "9,928,300", "9,884,129", "0.40%", "9,883,640", "5.80%"~
## $ Minnesota      <chr> "5,519,952", "5,303,924", "4.10%", "5,303,925", "6.40%"~
## $ Mississippi    <chr> "2,988,726", "2,968,103", "0.70%", "2,967,297", "6.30%"~
## $ Missouri       <chr> "6,093,000", "5,988,928", "1.70%", "5,988,927", "6.10%"~
## $ Montana        <chr> "1,042,520", "989,414", "5.40%", "989,415", "6.00%", "6~
## $ Nebraska       <chr> "1,907,116", "1,826,334", "4.40%", "1,826,341", "7.00%"~
## $ Nevada         <chr> "2,940,058", "2,700,691", "8.90%", "2,700,551", "6.30%"~
## $ New.Hampshire  <chr> "1,334,795", "1,316,461", "1.40%", "1,316,470", "4.80%"~
## $ New.Jersey     <chr> "8,944,469", "8,791,953", "1.70%", "8,791,894", "5.80%"~
## $ New.Mexico     <chr> "2081015", "2059198", "0.011", "2059179", "0.062", "0.0~
## $ New.York       <chr> "19745289", "19378110", "0.019", "19378102", "0.059", "~
## $ North.Carolina <chr> "10146788", "9535688", "0.064", "9535483", "0.06", "0.0~
## $ North.Dakota   <chr> "757952", "672591", "0.127", "672591", "0.073", "0.066"~
## $ Ohio           <chr> "11614373", "11536727", "0.007", "11536504", "0.06", "0~
## $ Oklahoma       <chr> "3923561", "3751615", "0.046", "3751351", "0.068", "0.0~
## $ Oregon         <chr> "4093465", "3831072", "0.068", "3831074", "0.058", "0.0~
## $ Pennsylvania   <chr> "12784227", "12702857", "0.006", "12702379", "0.056", "~
## $ Rhode.Island   <chr> "1056426", "1052940", "0.003", "1052567", "0.052", "0.0~
## $ South.Carolina <chr> "4961119", "4625410", "0.073", "4625364", "0.059", "0.0~
## $ South.Dakota   <chr> "865454", "814195", "0.063", "814180", "0.071", "0.073"~
## $ Tennessee      <chr> "6651194", "6346298", "0.048", "6346105", "0.061", "0.0~
## $ Texas          <chr> "27,862,596", "25,146,100", "10.80%", "25,145,561", "7.~
## $ Utah           <chr> "3,051,217", "2,763,888", "10.40%", "2,763,885", "8.30%~
## $ Vermont        <chr> "624,594", "625,741", "-0.20%", "625,741", "4.90%", "5.~
## $ Virginia       <chr> "8,411,808", "8,001,041", "5.10%", "8,001,024", "6.10%"~
## $ Washington     <chr> "7,288,000", "6,724,545", "8.40%", "6,724,540", "6.20%"~
## $ West.Virginia  <chr> "1,831,102", "1,853,011", "-1.20%", "1,852,994", "5.50%~
## $ Wisconsin      <chr> "5,778,708", "5,687,289", "1.60%", "5,686,986", "5.80%"~
## $ Wyoming        <chr> "585,501", "563,767", "3.90%", "563,626", "6.50%", "7.1~
```

Why do I think this dataset is not tidy:

- **Multiple variables in one column**:
  "Fact" column contains several descriptions of the data, which needs to be separate to be tidy
- States supposed to be rows (observations), not columns.
- Each row represents multiple observations. In tidy dataset each row represents single observation, for example a specific state's demographic data for a single year.

Here I create separate tidy data table for populations of each state in years 2010 and 2016

```r
state_names <- colnames(census_data)[3:ncol(census_data)]
#state_names

population_data <- census_data |>
  filter(Fact %in% c("Population estimates, July 1, 2016,  (V2016)",
                     "Population, Census, April 1, 2010"))

long_population_data <- population_data |>
```

```
  pivot_longer(
    cols = -Fact,
    names_to = "state",
    values_to = "population"
  ) |>
  mutate(year = case_when(
    grepl("2016", Fact) ~2016,
    grepl("2010", Fact) ~ 2010,
    TRUE ~ NA_integer_
  )) |>
  select(-Fact) |>
  pivot_wider(
    names_from = year,
    values_from = population
   ) |>
  # |>
  # mutate(
  #   `2010` = as.numeric(`2010`),
  #   `2016` = as.numeric(`2016`)
  # )
  select(state, "2010", "2016")

long_population_data$`2010` <- gsub(",", "", long_population_data$`2010`)
long_population_data$`2010` <-  as.numeric(long_population_data$`2010`)

long_population_data$`2016` <- gsub(",", "", long_population_data$`2016`)
long_population_data$`2016` <-  as.numeric(long_population_data$`2016`)
print(long_population_data)
```

```
## # A tibble: 50 x 3
##    state         `2010`   `2016`
##    <chr>          <dbl>    <dbl>
##  1 Alabama      4779736  4863300
##  2 Alaska        710231   741894
##  3 Arizona      6392017  6931071
##  4 Arkansas     2915918  2988248
##  5 California  37253956 39250017
##  6 Colorado     5029196  5540545
##  7 Connecticut  3574097  3576452
##  8 Delaware      897934   952065
##  9 Florida     18801310 20612439
## 10 Georgia      9687653 10310371
## # i 40 more rows
```

This `long_population_data`data table is tidy and ready for analysis.

In this code I created `create_long_table`function which makes it easier to create different tidy data tables:

```
create_long_table <- function(column_name, rows) {
  result <- census_data |>
  slice(rows) |>
  pivot_longer(
    cols = !Fact,
    names_to = "state",
    values_to = column_name
  ) |>
   mutate(year = case_when(
    grepl("2016", Fact) ~2016,
    grepl("2010", Fact) ~ 2010,
    TRUE ~ NA_integer_
  )) |>
  select(-Fact) |>
  select(state, year, all_of(column_name)) |>
  mutate(
    !!column_name := round(as.numeric(gsub("%", "", !!sym(column_name))), 2)
    )

  return(result)
}
```

Here I create `gender_table_long`tidy data table which includes female population proportions for each state in years 2010 and 2016. I use `create_long_table`to create this table. Also I did data transformation because some states had proportions in percents, and some decimal values as actual proportions.

```
gender_table_long <- create_long_table("female_prop", 11:12)
gender_table_long <- gender_table_long |>
  mutate(
    female_prop = if_else(
                          female_prop < 1,
                          round(female_prop * 100, 2),
                          round(female_prop, 2)
  ))
print(gender_table_long)
```

```
## # A tibble: 100 x 3
##    state        year female_prop
##    <chr>       <dbl>       <dbl>
##  1 Alabama      2016        51.6
##  2 Alaska       2016        47.7
##  3 Arizona      2016        50.3
##  4 Arkansas     2016        50.9
##  5 California   2016        50.3
##  6 Colorado     2016        49.7
##  7 Connecticut  2016        51.2
##  8 Delaware     2016        51.6
##  9 Florida      2016        51.1
## 10 Georgia      2016        51.3
## # i 90 more rows
```

This code creates new observations in `gender_table_long` for male population for each state which is better for analysis.

```
gender_table_long|>
  mutate(
    male_prop = 100 - female_prop
  ) |>
  pivot_longer(
    cols = contains("prop"),
    names_to = "gender",
    values_to = "value"
  ) |>
  pivot_wider(
    names_from = year,
    values_from = value,
    names_prefix = "X"
  ) |>
  select(state, gender, X2010, X2016) |>
  mutate(
    prop_change = X2016 - X2010
  )
```

```
## # A tibble: 100 x 5
##    state      gender      X2010 X2016 prop_change
##    <chr>      <chr>       <dbl> <dbl>       <dbl>
##  1 Alabama    female_prop  51.5  51.6       0.100
##  2 Alabama    male_prop    48.5  48.4      -0.100
##  3 Alaska     female_prop  48    47.7      -0.300
##  4 Alaska     male_prop    52    52.3       0.300
##  5 Arizona    female_prop  50.3  50.3       0
##  6 Arizona    male_prop    49.7  49.7       0
##  7 Arkansas   female_prop  50.9  50.9       0
##  8 Arkansas   male_prop    49.1  49.1       0
##  9 California female_prop  50.3  50.3       0
## 10 California male_prop    49.7  49.7       0
## # i 90 more rows
```

```
head(gender_table_long)
```

```
## # A tibble: 6 x 3
##   state       year female_prop
##   <chr>      <dbl>       <dbl>
## 1 Alabama     2016        51.6
## 2 Alaska      2016        47.7
## 3 Arizona     2016        50.3
## 4 Arkansas    2016        50.9
## 5 California  2016        50.3
## 6 Colorado    2016        49.7
```

From here I created few data tables for different age ranges:

```
under_5_proportions_long <- create_long_table("prop_under_5", 5:6) |>
  group_by(year, state)
print(under_5_proportions_long)
```

```
## # A tibble: 100 x 3
## # Groups:   year, state [100]
##    state        year prop_under_5
##    <chr>       <dbl>        <dbl>
##  1 Alabama      2016          6
##  2 Alaska       2016          7.3
##  3 Arizona      2016          6.3
##  4 Arkansas     2016          6.4
##  5 California   2016          6.3
##  6 Colorado     2016          6.1
##  7 Connecticut  2016          5.2
##  8 Delaware     2016          5.8
##  9 Florida      2016          5.5
## 10 Georgia      2016          6.4
## # i 90 more rows
```

```r
under_18_proportions_long <- create_long_table("prop_under_18", 7:8)
print(under_18_proportions_long)
```

```
## # A tibble: 100 x 3
##    state        year prop_under_18
##    <chr>       <dbl>         <dbl>
##  1 Alabama      2016          22.6
##  2 Alaska       2016          25.2
##  3 Arizona      2016          23.5
##  4 Arkansas     2016          23.6
##  5 California   2016          23.2
##  6 Colorado     2016          22.8
##  7 Connecticut  2016          21.1
##  8 Delaware     2016          21.5
##  9 Florida      2016          20.1
## 10 Georgia      2016          24.4
## # i 90 more rows
```

```r
over_65_proportions_long <- create_long_table("over_65", 9:10)
print(over_65_proportions_long)
```

```
## # A tibble: 100 x 3
##    state        year over_65
##    <chr>       <dbl>   <dbl>
##  1 Alabama      2016    16.1
##  2 Alaska       2016    10.4
##  3 Arizona      2016    16.9
##  4 Arkansas     2016    16.3
##  5 California   2016    13.6
##  6 Colorado     2016    13.4
##  7 Connecticut  2016    16.1
##  8 Delaware     2016    17.5
##  9 Florida      2016    19.9
## 10 Georgia      2016    13.1
## # i 90 more rows
```

```r
population_proportions <- under_5_proportions_long |>
  left_join(under_18_proportions_long, by = c("state", "year")) |>
  left_join(over_65_proportions_long, c("state", "year")) |>
  mutate(
    prop_18_to_65 = 100 - prop_under_18 - over_65
  )
```

This code creates data table for population proportions in 18-65 ages range for each state

```r
population_18_65_long <- population_proportions |>
  select(state, year, prop_18_to_65)
print(population_18_65_long)
```

```
## # A tibble: 100 x 3
## # Groups:   year, state [100]
##    state        year prop_18_to_65
##    <chr>       <dbl>         <dbl>
##  1 Alabama      2016          61.3
##  2 Alaska       2016          64.4
##  3 Arizona      2016          59.6
##  4 Arkansas     2016          60.1
##  5 California   2016          63.2
##  6 Colorado     2016          63.8
##  7 Connecticut  2016          62.8
##  8 Delaware     2016          61
##  9 Florida      2016          60
## 10 Georgia      2016          62.5
## # i 90 more rows
```

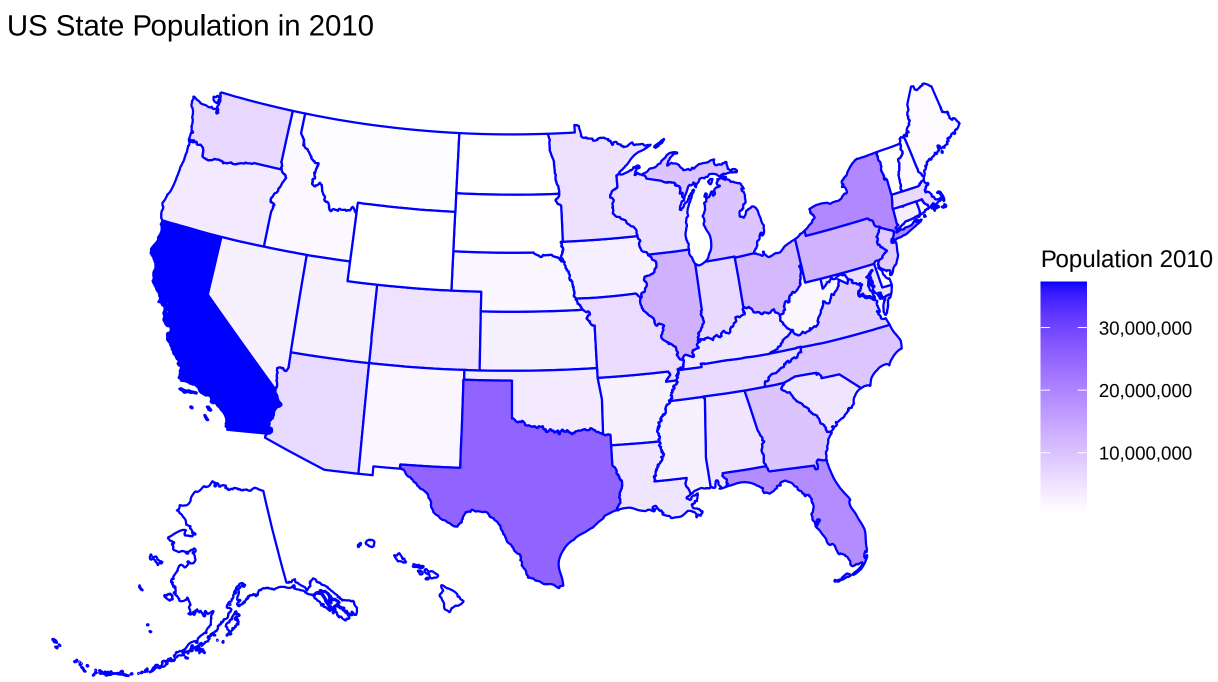
Analysis

This plot shows analysis for each state's population in 2010.

```r
library(usmap)
long_population_data$state <- gsub("\\.", " ", long_population_data$state)

names(long_population_data)[names(long_population_data) == "2010"] <- "pop_2010"
names(long_population_data)[names(long_population_data) == "2016"] <- "pop_2016"

data_for_map <- long_population_data |>
  left_join(usmap::statepop, by = c("state" = "full"))

mismatches <- setdiff(long_population_data$state, usmap::statepop$full)
print(mismatches)
```

```
## character(0)
```

```r
# long_population_data$state <- trimws(long_population_data$state)
# usmap::statepop$full <- trimws(usmap::statepop$full)
# head(data_for_map)
# plot_usmap(data_for_map, values = "pop_2010", color = "blue") +
#   scale_fill_continuous(name = "Population 2010",
#                         low = "white",
#                         high = "blue") +
#   theme(legend.position = "right") +
#   labs(title = "US State Population in 2010")

plot_usmap(data = long_population_data, values = "pop_2010", color = "blue") +
  scale_fill_continuous(name = "Population 2010",
                        low = "white",
                        high = "blue",
                        labels = scales::comma) +
  theme(legend.position = "right") +
  labs(title = "US State Population in 2010")
```

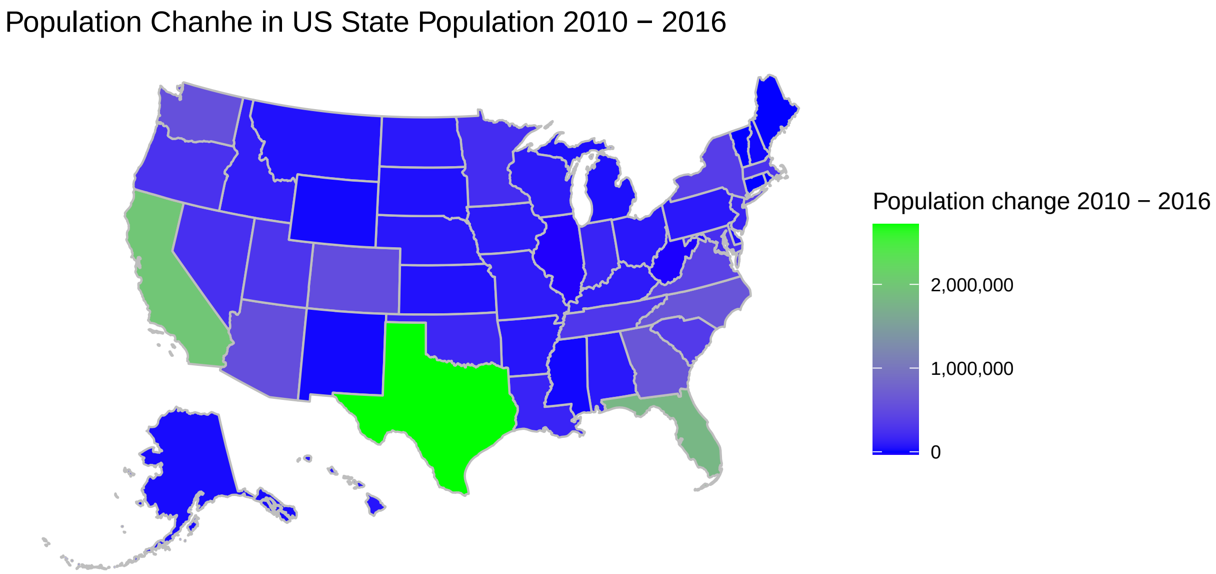


As we see on the plot most populated states in 2010:

1. California
2. Texas
3. Florida
4. New York

This plot shows how population changed in states in period of 2010 - 2016

```r
long_population_data <- long_population_data |>
  mutate(
    pop_change = pop_2016 - pop_2010
  )

plot_usmap(data=long_population_data, values = "pop_change", color = "gray") +
  scale_fill_gradient2(name = "Population change 2010 - 2016",
                       low = "red", mid = "blue", high = "green",
                       midpoint = 0,
                       labels = scales::comma
                       ) +
  theme(legend.position = "right") +
  labs(title = "Population Chanhe in US State Population 2010 - 2016")
```


Population Chanhe in US State Population 2010 − 2016
Population change 2010 − 2016
2,000,000
1,000,000
0

From this plot we see states which had most increases in population:

1. Texas
2. California
3. Florida