Makeover Monday: 2016 US Household Income by State and Popular Election Results

Thinh T. Pham and Katherine M. Prioli

October 10, 2018

Objective 1: Reproduce Original Graph

The original graph (Fig. 1 below) was the 2018 Week 3 challenge on Makeover Monday. It is a stacked horizontal bar chart depicting US household income distribution by state, broken out into six categories, and organized by order of increasing percentage in the highest income category.

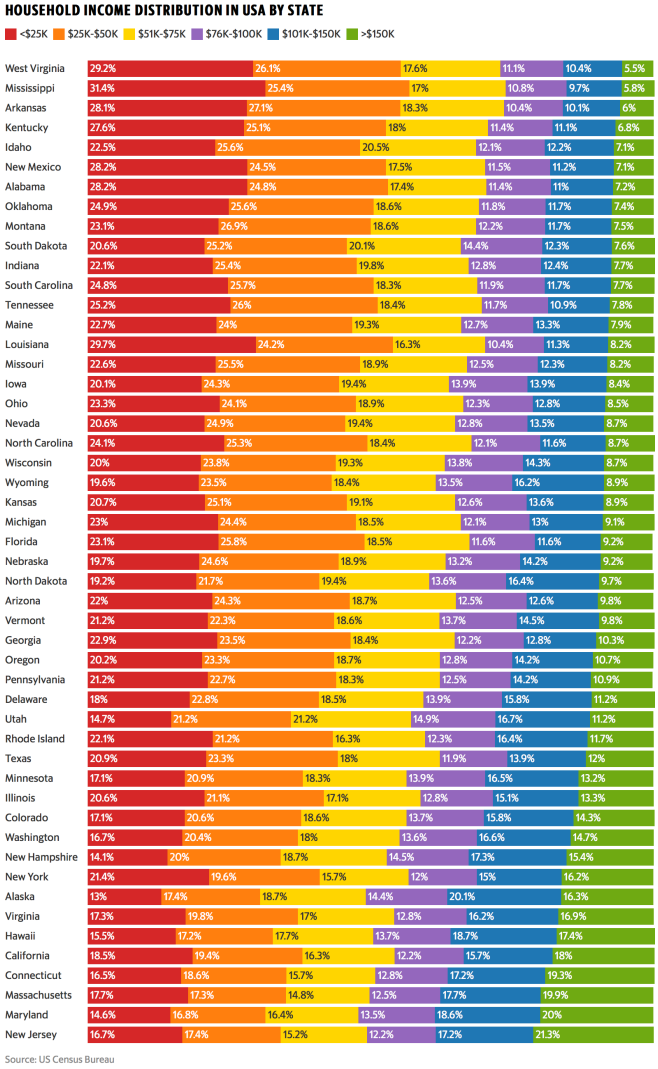

Figure 1. Original graph.

Loading Libraries

```
library(tidyverse)
library(readxl)
library(grid)
library(gridExtra)
```

Reading in and Subsetting Data

The dataset posted on the Makeover Monday website did not match the original graph - that is, the original graph had different underlying data. The author of the original graph directed us to the correct dataset for 2016 on the US Census Bureau's American FactFinder page (United States Census Bureau 2017). This data was downloaded as `ACS_16_1YR_S1901_with_ann.csv`.

The data was read in, subset to columns and rows of interest, and given sensible column names.

```
incomedata <- read_csv("data/ACS_16_1YR_S1901_with_ann.csv")
incomedata <- incomedata %>% select("GEO.display-label",     # Subset to only the variables of interest
                                    "HC01_EST_VC01",
                                    "HC01_EST_VC02",
                                    "HC01_EST_VC03",
                                    "HC01_EST_VC04",
                                    "HC01_EST_VC05",
                                    "HC01_EST_VC06",
                                    "HC01_EST_VC07",
                                    "HC01_EST_VC08",
                                    "HC01_EST_VC09",
                                    "HC01_EST_VC10",
                                    "HC01_EST_VC11")

colnames(incomedata) <- c("state",                           # GEO.display-label
                          "tot_households",                  # HC01_EST_VC01
                          "Less than $10,000",               # HC01_EST_VC02
                          "$10,000 to $14,999",              # HC01_EST_VC03
                          "$15,000 to $24,999",              # HC01_EST_VC04
                          "$25,000 to $34,999",              # HC01_EST_VC05
                          "$35,000 to $49,999",              # HC01_EST_VC06
                          "$50,000 to $74,999",              # HC01_EST_VC07
                          "$75,000 to $99,999",              # HC01_EST_VC08
                          "$100,000 to $149,999",            # HC01_EST_VC09
                          "$150,000 to $199,999",            # HC01_EST_VC10
                          "$200,000 or more")                # HC01_EST_VC11

incomedata <- incomedata %>%
  slice(2:n()) %>%
  filter(state != "Puerto Rico" & state != "District of Columbia")  # Omit PR and DC per original graph
```

Tidying and Transforming the Data

Next, `gather()` was used to transform the data from wide to long format.

```
incomegat <- incomedata %>%
  gather(`Less than $10,000`,
         `$10,000 to $14,999`,
         `$15,000 to $24,999`,
         `$25,000 to $34,999`,
         `$35,000 to $49,999`,
         `$50,000 to $74,999`,
         `$75,000 to $99,999`,
         `$100,000 to $149,999`,
         `$150,000 to $199,999`,
         `$200,000 or more`,
         key = "incomelev", value = "pct")
```

Serial `mutate()` calls were used to cast existing variables as numeric values and create income categories consistent with those used in the original graph. Percentages in each income category were calculated using `group_by()` and `mutate()`, then the dataset was subset to only those variables and rows of interest.

```
incometbl <- incomegat %>%
  mutate(tot_households = as.numeric(tot_households)) %>%
  mutate(pct = as.numeric(pct)) %>%
  mutate(n_households = round((pct / 100) * tot_households, digits = 0)) %>%
  mutate(incomecat = case_when(
    incomelev == "Less than $10,000" ~ "<$25K",
    incomelev == "$10,000 to $14,999" ~ "<$25K",
    incomelev == "$15,000 to $24,999" ~ "<$25K",
    incomelev == "$25,000 to $34,999" ~ "$25K-$50K",
    incomelev == "$35,000 to $49,999" ~ "$25K-$50K",
    incomelev == "$50,000 to $74,999" ~ "$51K-$75K",
    incomelev == "$75,000 to $99,999" ~ "$76K-$100K",
    incomelev == "$100,000 to $149,999" ~ "$101K-$150K",
    incomelev == "$150,000 to $199,999" ~ ">$150K",
    incomelev == "$200,000 or more" ~ ">$150K")) %>%
  mutate(incomecat = factor(incomecat,
                        levels = c("<$25K",
                                   "$25K-$50K",
                                   "$51K-$75K",
                                   "$76K-$100K",
                                   "$101K-$150K",
                                   ">$150K"))) %>%
  group_by(state, incomecat) %>%
  mutate(catpct = sum(pct)) %>%
  select(state, catpct, incomecat) %>%
  distinct(state, catpct, incomecat) %>%
  arrange(desc(incomecat), catpct)
```

The `state` variable was converted to a factor to allow for ordering the graph in the desired manner.

```
incometbl <- within(incometbl,                      # Assign levels for state column
                    state <- factor(state,
                                    levels = rev(incometbl$state[1:50])))
```

Generating the Reproduced Plot

Finally, the original graph was recreated using the following key components (among others):

- `geom_bar()` to create the stacked bars
- `coord_flip()` to arrange them horizontally
- `scale_fill_manual()` to manually match the colors used in the original graph
- `geom_text()` to add labels indicating the percentage of state households in each income category
- `scale_y_reverse()` to order the income categories from lowest to highest

```
reprod_plot <- ggplot(incometbl, aes(x = state, y = catpct / 100,
                                     fill = incomecat, label = catpct)) +
  coord_flip() +    # Put states on vertical axis
  geom_bar(stat = "identity", aes(y = catpct * 100), position = "fill") +
  scale_fill_manual(values = c(     # Match color scheme of original graph
    "<$25K" = "#D62728",
    "$25K-$50K" = "#FF7F03",
    "$51K-$75K" = "#FFD500",
    "$76K-$100K" = "#9467BD",
    "$101K-$150K" = "#1F77B4",
    ">$150K" = "#6FAA12")) +
  geom_text(aes(label = paste0(catpct,"%")),    # Label stacked bars
            size = 2, color = "white", fontface = "bold",
            hjust = 0.5, position = position_stack(vjust = 0.5)) +
  scale_y_reverse() +    # Ensure <$25K category on left
  theme(axis.text.x = element_blank(),
```

```
        axis.ticks = element_blank(),
        axis.title = element_blank(),
        panel.background = element_blank(),
        legend.position = "top",
        legend.title = element_blank(),
        legend.key.width = unit(0.5, "cm"),
        legend.key.height = unit(0.5, "cm")) +
  guides(fill = guide_legend(nrow = 1)) +      # Force all legend elements into one row
  ggtitle("HOUSEHOLD INCOME DISTRIBUTION IN USA BY STATE")
reprod_plot
```

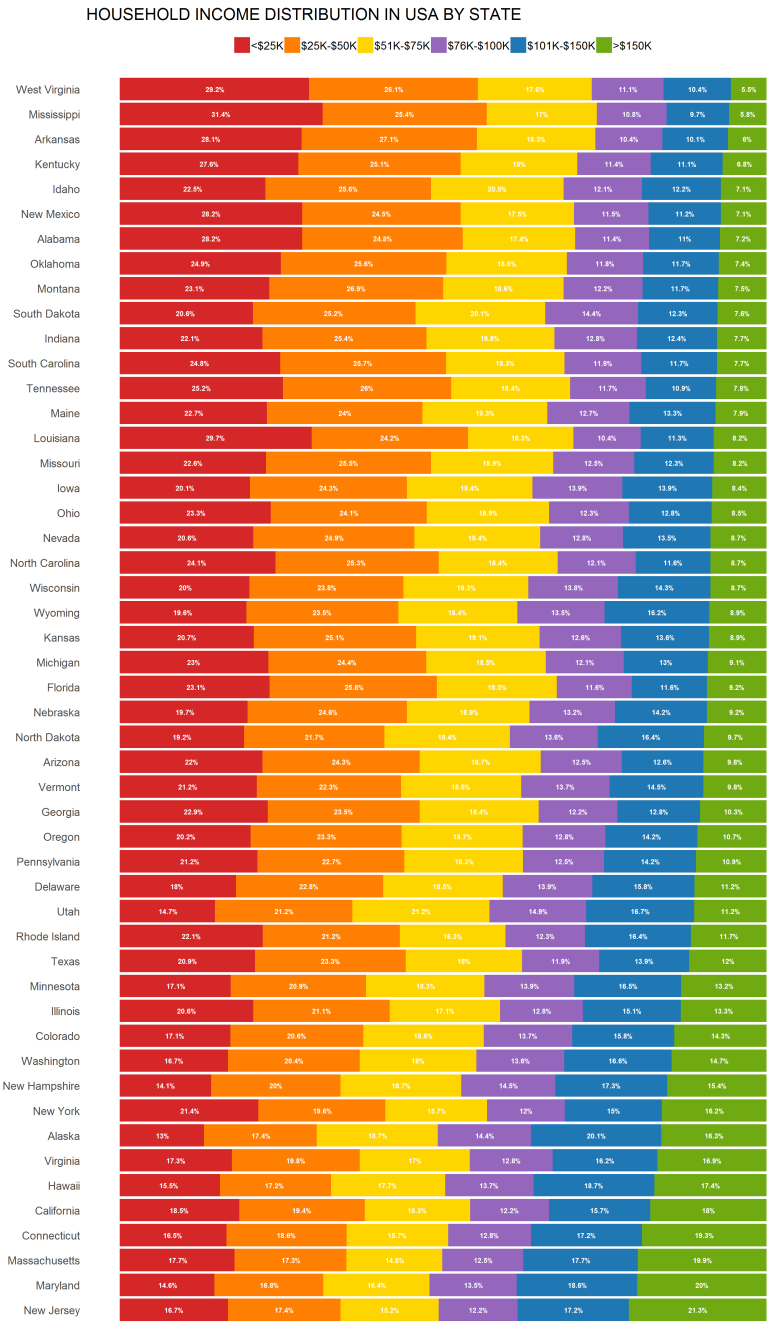

Figure 2. Reproduced graph.

This graph bears minor cosmetic differences from the original graph, but both graphs are functionally the same. The plot was saved using `ggsave()`.

```
ggsave("images/reprod_plot.png", height = 15, width = 9.1, units = "in")
```

Because we were working in parallel using GitHub, the dataset was exported to .csv for use in creating the improved graphs.

```
write_csv(incometbl, "data/incometbl.csv")
```

Objective 2: Create Improved Graphs

Merging in Election Results Data

Because the original data was from 2016, we were interested in how the 2016 presidential election popular vote results may be related to income distribution. Election results data, obtained from the Federal Election Commission website (Federal Election Commission 2018), were read in using `read_xlsx()`, and the popular vote `winner` was calculated and converted to a factor using several piped `mutate()` calls. Factors were also assigned to `state`, consistent with the `incometbl` data, and the percentages of states won by each candidate were calculated.

```
popvote_raw <- read_xlsx("data/federalelections2016.xlsx", sheet = "Appendix A", range = "A7:C59")

popvote <- as.tibble(popvote_raw) %>%
  filter(!is.na(STATE) & STATE != "D.C.") %>%
  mutate(gop = as.numeric(TRUMP)) %>%
  mutate(dem = as.numeric(CLINTON)) %>%
  mutate(winner = case_when(
    dem < gop ~ "R",
    dem > gop ~ "D"
  )) %>%
  mutate(winner = factor(winner, levels = c("R", "D"))) %>%
  rename(state = STATE) %>%
  mutate(state = factor(state, levels = rev(incometbl$state[1:50]))) %>%
  select(state, winner)

popvote_gop <- popvote %>% filter(winner == "R") %>% count()
weight_gop = popvote_gop / 50
popvote_dem <- popvote %>% filter(winner == "D") %>% count()
weight_dem = popvote_dem / 50
```

The data exported in Objective 1 above (`incometbl.csv`) was reimported and factors were reassigned to the income levels and states (it should be noted that, had we not been working in parallel, this step would not have been necessary).

```
incometbl_raw <- read_csv("data/incometbl.csv")

incometbl <- incometbl_raw %>%
  arrange(desc(incomecat), catpct) %>%
  mutate(incomecat = factor(incomecat,
                            levels = c("<$25K",
                                       "$25K-$50K",
                                       "$51K-$75K",
                                       "$76K-$100K",
                                       "$101K-$150K",
                                       ">$150K"))) %>%
  mutate(state = factor(state, levels = rev(state[1:50])))
```

The two datasets were joined, and the resulting dataset was exported.

```
incometbl <- full_join(incometbl, popvote, by = "state")

write_csv(incometbl, "data/incometbl_vote.csv")    # Export vote-augmented dataset in case Thinh wants it
```

Separating the Reproduced Graph by Popular Vote

Two graphs were produced in the same manner as in Objective 1, one for the states won by Trump (titled `reprod_plot_gop`), and the other for those won by Clinton (titled `reprod_plot_dem`). In the interest of brevity, code for these plots will not be

shown here, but is available in the repository for this project (Prioli and Pham 2018). In each plot, a `filter()` statement was used within the `ggplot()` call to subset the data to states won by the party of interest - for example, for Trump:

```
reprod_plot_gop <- ggplot(data = filter(incometbl, winner == "R"),
                          aes(...)) + ...
```

Next, `arrangeGrob()` in the `grid` package was used to annotate each plot with the winner using `textGrob()`, then `grid.arrange()` (from the `gridExtra` package) was used to arrange these two plots vertically and add a title.

```
plot_gop <- arrangeGrob(reprod_plot_gop, left = textGrob("TRUMP",
                                                         rot = 90,
                                                         gp = gpar(fontsize = 14, fontface = "bold")))
plot_dem <- arrangeGrob(reprod_plot_dem, left = textGrob("CLINTON",
                                                         rot = 90,
                                                         gp = gpar(fontsize = 14, fontface = "bold")))

grid.arrange(plot_gop, plot_dem, nrow = 2,
             heights = c(2 * weight_gop, 2 * weight_dem),
             top = textGrob("HOUSEHOLD INCOME DISTRIBUTION IN USA \n BY STATE AND 2016 POPULAR VOTE",
                            gp = gpar(fontsize = 16, fontface = "bold")))

# Note - grid.arrange disables ggsave(), so I manually exported the graph using the Plots tab
```

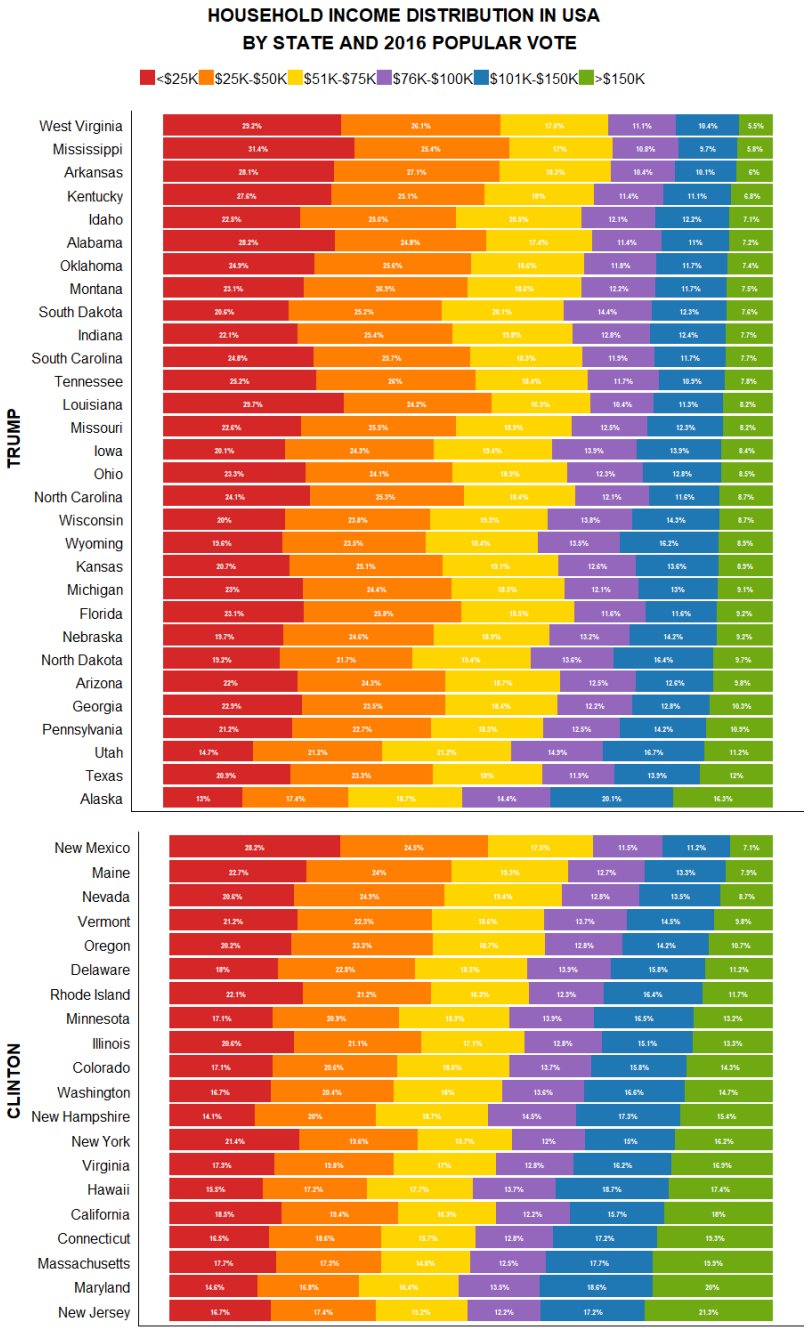

Figure 3. Reproduced graph segmented by 2016 popular vote results.

The most noteworthy trend in this graph was that the states won by Hillary Clinton had a noticeably higher percentage of households in the >$150K income category, whereas those won by Donald Trump had greater percentages in the low- and middle-ranges (<$25K and $25-$50K).

Choropleths

Another way to visualize this data is by choropleths. We incorporated the data from the original graph with the map of the US.

Merge Data Tables

We merged data from "incometbl_vote.csv" with states data from package `maps`. `inner_join()` is used to eliminate non-matching rows. Washington D.C. is therefore not included in the map because it was not in the original data.

```
states <- map_data("state")
incometbl <- read_csv("data/incometbl_vote.csv")

### Prepare to merge incometbl and state data
incometbl <- incometbl %>%
                  mutate(region = tolower(state)) %>%
                  select(-state)

income_data <- states %>%                                  #Inner_join() used to eliminate D.C
                  inner_join(incometbl, by = "region")     #(not listed in the original graph)
```

Convert Income Category into Factor Variable

Originally, a choropleth graph with income category as a facet was employed, so we create a factor levels for `incomecat` variable using the code below. We ultimately choose not to follow this path, but preserve the code in case such need comes up.

```
# Set up factor levels for organization of graph
incomecat_levels <- factor(c("<$25K", "$25K-$50K", "$51K-$75K",
                             "$76K-$100K", "$101K-$150K", ">$150K"))
income_data[,"incomecat"] <- factor(income_data[,"incomecat"],
                                    levels = incomecat_levels)
```

Create Graphing Function

Since a choropleth with income category as a facet clutters the entire graph and makes it hard to digest the information, we decided to create an income distribution map for each income category separately. The function belows takes in an income cateogry and a desired color theme (1 or 2) as inputs and outputs a graph showing the distribution of that income group. The range of the income distribution for each income category is divided into 6 equal intervals; each interval corresponds to the color of the fill of a state on the map. In addition, the border colors of the states represent their political leaning in the 2016 election. The color schemes are chosen from colorbrewer2.org to be both colorblind and LCD friendly (Brewer and Harrower 2018).

```
#### Create plotting function with diverging color schemes ---
plot_category_scheme1 <- function(cat, scheme = 1) {
    scheme1 <- c('#8c510a','#d8b365','#f6e8c3',
                 '#c7eae5','#5ab4ac','#01665e')
    scheme2 <- c('#b35806','#f1a340','#fee0b6',
                 '#d8daeb','#998ec3','#542788')
    color_scheme <- if(scheme == 1) {scheme1} else {scheme2}

    result <- income_data %>% filter(incomecat == cat) %>%
    mutate(percentage = cut_interval(catpct, n = 6)) %>%  #Divide into 6 equal intervals
       ggplot() +
       geom_polygon(aes(x = long, y = lat,
                    fill = percentage,
                    group = group,
                    color = winner)) +                    #Border color by political party
       scale_fill_manual(values = color_scheme) +
       scale_color_manual(values = c("black", "red")) +
       coord_fixed(1.3) +
       ggtitle(paste("Distribution of income group", cat))

    result
}
```

Create Cholopleths

Finally, we created all choropleths and saved the files using the snippet of code below.

```
# Create and save plots ----
cat_names <- c("below_25k", "25_to_50k", "51_to_75k",
               "76_to_100k", "101_to_150k", "above_150k")
for (i in 1:6) {
    cat <- as.character(incomecat_levels)[i]
    name <- cat_names[i]
    plot_category_scheme1(cat,scheme = 1)
    ggsave(filename = paste0("choropleth\\scheme1\\choropleth1_",name,".png"),
           height = 4, width = 8)
    plot_category_scheme1(cat,scheme = 2)
    ggsave(filename = paste0("choropleth\\scheme2\\choropleth2_",name,".png"),
           height = 8, width = 8)
}
```

For brevity purpose, we only show two graphs of the lowest and the highest income groups. The rest are easily generated from our code or can be accessed from the repository associated with this report.

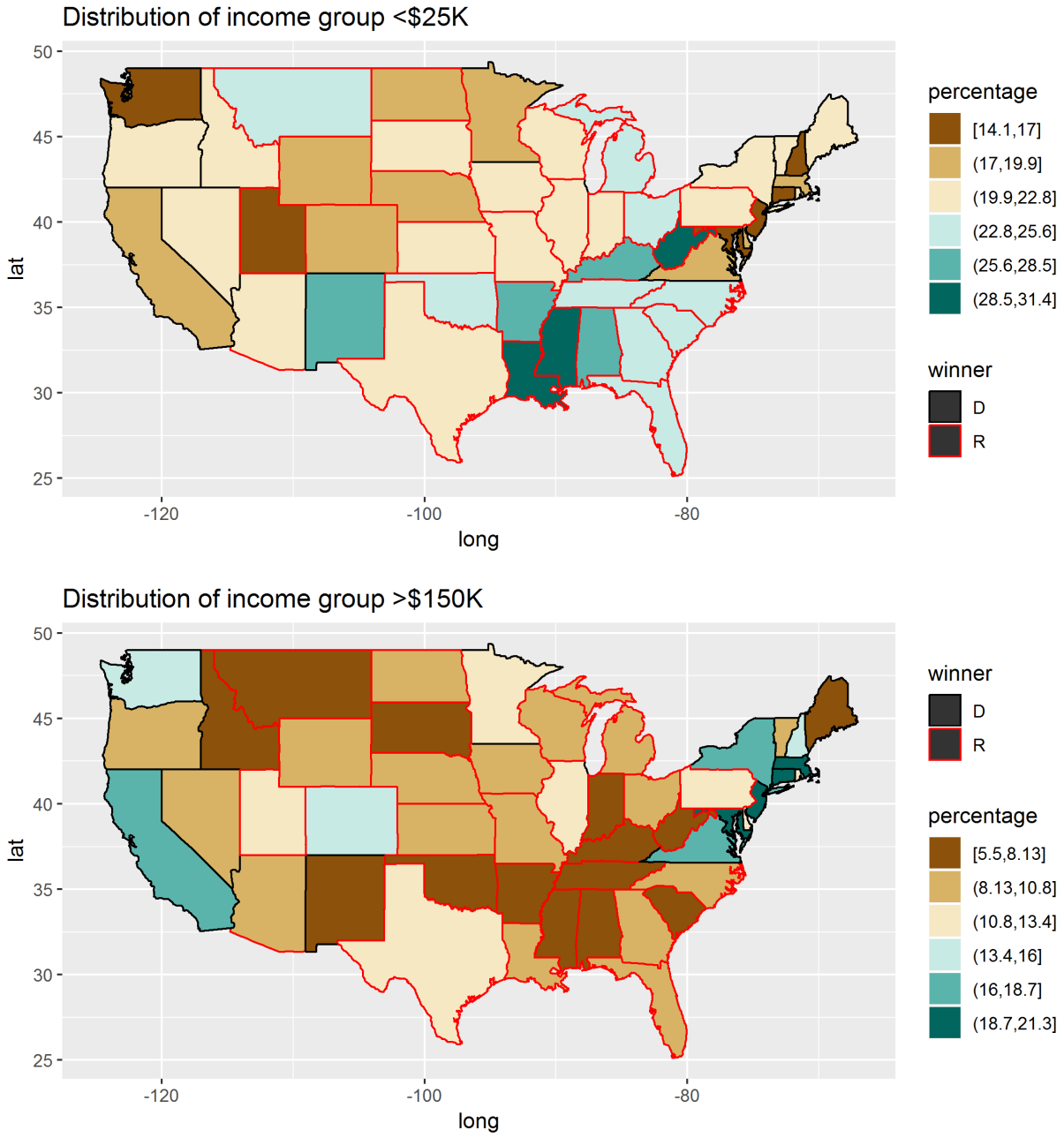

Figure 4. Choropleths for highest and lowest income categories.

States over the east coast and California have higher concentration of rich earners. These states were also Democratic leaning in the 2016 race, as illustrated by their black borders. The rest of the country have lower portion of top income earners, with the southern states having the lowest share. These states leaned towards the GOP in the 2016 election, as indicated by the red borders.

The reverse is true for income group earning below $25,000 a year. States with moderate to high concentration of low earners mostly are in the south and around the great lakes. States with the lowest portion of low income earners are in the northest and along the west coast.

Overall, the message brought by the choropleths is the same as the stacked bar graph presented earlier: states where Clinton won tend to have higher portion of high income earners and low portion of low income earners, while states where Trump won have higher portion low earners and lower portion of high earners.

Full Code

Full code is available in a public Github repository (Prioli and Pham 2018).

References

Brewer, Cynthia, and Mark Harrower. 2018. "Color Brewer 2.0 - Color Advice for Cartography." http://colorbrewer2.org/#type=diverging&scheme=BrBG&n=6.

Federal Election Commission. 2018. "Election Results for the U.S. President, the U.S. Senate, and the U.S. House of Representatives." https://transition.fec.gov/general/FederalElections2016.shtml.

Prioli, Katherine M., and Thinh T. Pham. 2018. "Makeover_Monday." https://github.com/kmprioliPROF/Makeover_Monday.

United States Census Bureau. 2017. "Fact Finder." https://factfinder.census.gov/bkmk/table/1.0/en/ACS/16_1YR/S1901/0100000US.04000.