ST 558: Project 1

Kayla Kippes and Zack Rosen

2025-06-16

For this project, our goal was to manipulate and process data sets that came in a certain form. To start this process, we completed each individual step on one data set. This allowed us to ensure the content of our functions would be working properly. Then we added each of those steps into their respective functions. After that we created a wrapper function to pull everything into one place. From there we combined the necessary data sets and performed unique types of the plot function. The rest of this document will talk through each function and give examples of all of our functions coming together to be used on actual data.

Function 1: Read and Convert Data

We first started by preprocessing the read data. This involved selecting useful columns, namely, Area_name, STCOU and those that end with "D". The tidyverse package was extremely useful for data preprocessing techniques and manipulations such as these. We then renamed the column for consistency, and converted the data from a wide to a long format. To do this, we transformed the columns ending in "D" into a single column named "survey_value" and mapped the corresponding original value to these observations by adding a new column. This new column was named by the column_name variable which was included in the function signature as an optional parameter with the default value of "enrollment".

```
library(tidyverse)

read_and_preprocess <- function(data, column_name = "enrollment") {
  # Step 1
  ## select / rename columns
  EDU01a <- data |>
    select(Area_name, STCOU, ends_with("D")) |>
    rename("area_name" = "Area_name")
  ## print out the first 5 rows
  print("Preprocessed:")
```

```
  print(head(EDU01a, 5))

  # Step 2
  # pivot cols 3-12 into long format
  long_tibble <- EDU01a |>
    pivot_longer(cols = 3:12, names_to = "survey_value", values_to = column_name)
  ## print out the first 5 rows
  print("Long format:")
  print(head(long_tibble, 5))

  ##return long data
  return(long_tibble)
}
```

Function 2: Parsing the Data and Creating New Varaiables

In order to parse the data and create new variables, We figured "mutate" would have to be used. Since each year was embedded into the "survery_value" column and every value in that column was the same length, we were able to sub string the year out and make it a numeric. However, this only gave me two digits and we wanted four digit years. To solve for this, we added an "if" statement to add either 1900 or 2000 to my two digit year (this wouldn't have worked if the data includes years below 1925). Also, we had made a temporary column initially with the short year so we decided to select all other columns except for the one that wasn't needed.

```
parse_new_variables <- function(long_tibble) {
  long_updated <- long_tibble |>
  mutate(short_year = as.numeric(substr(survey_value, 8, 9)),
         year = ifelse(short_year > 25, 1900 + short_year, 2000 + short_year),
         measurement = substr(survey_value, 1, 7)) |>
  select(-short_year)
  ## print out the first 5 rows
  print("Updated:")
  print(head(long_updated, 5))

  ## returns long updated
  return(long_updated)
}
```

Function 3: County Level

Similar to the year scenario above, we had to use "substr" to create a state column for the county data. This was a bit trickier as the values in area_name were not all the same length. To solve for this, we need to grab the max number of characters in the string and pull the second to last and last one so we could get the two character state value.

```
## add state column
add_state_col_county <- function(county_tibble) {
  county_tibble <- county_tibble |>
  mutate(state = substr(area_name, nchar(area_name) - 1, nchar(area_name)))
  ## return the tibble
  return(county_tibble)
}
```

Function 4: Non-County Level

Similar to the above functions, we figured that "mutate" would be the best way to add a new division column. This new column's values were determined by a case_when statement that checked if the area_name of that observation was in a vector corresponding to one of the Census Bureau's designated divisions. After all of these divisions were checked, we added the value "ERROR" to the division column if none of the divisions were a match.

```
add_division_col_state <- function(state_tibble) {
  # Step 6
  ## create division variable and set division by state name, else ERROR
  state_tibble <- state_tibble |>
    mutate(division = case_when(
      area_name %in% c("CONNECTICUT", "MAINE", "MASSACHUSETTS", "NEW HAMPSHIRE",
                       "RHODE ISLAND", "VERMONT") ~ "New England",
      area_name %in% c("NEW JERSEY", "NEW YORK", "PENNSYLVANIA") ~ "Mid-Atlantic",
      area_name %in% c("ILLINOIS", "INDIANA", "MICHIGAN", "OHIO",
                       "WISCONSIN") ~ "East North Central",
      area_name %in% c("IOWA", "KANSAS", "MINNESOTA", "MISSOURI", "NEBRASKA",
                       "NORTH DAKOTA", "SOUTH DAKOTA") ~ "West North Central",
      area_name %in% c("DELAWARE", "DISTRICT OF COLUMBIA", "FLORIDA", "GEORGIA",
                       "MARYLAND", "NORTH CAROLINA", "SOUTH CAROLINA", "VIRGINIA",
                       "WEST VIRGINIA") ~ "South Atlantic",
      area_name %in% c("ALABAMA", "KENTUCKY", "MISSISSIPPI",
                       "TENNESSEE") ~ "East South Central",
      area_name %in% c("ARKANSAS", "LOUISIANA", "OKLAHOMA",
```

```
                        "TEXAS") ~ "West South Central",
      area_name %in% c("ARIZONA", "COLORADO", "IDAHO", "MONTANA", "NEVADA",
                        "NEW MEXICO", "UTAH", "WYOMING") ~ "Mountain",
      area_name %in% c("ALASKA", "CALIFORNIA", "HAWAII", "OREGON",
                        "WASHINGTON") ~ "Pacific",
      TRUE ~ "ERROR"))
  return(state_tibble)
}
```

Function 5: Returning Two Final Tibbles

This function filters the long format data into two tibbles: a county-level tibble and a state-level tibble. The county-level tibble corresponds to county entries, with area_name values identified by a comma and a two letter state abbreviation. The state-level tibble was simply all of the other entries that were not in the county-level tibble. Lastly, a county class was added to the county-level tibble and a state class was added to the state-level tibble.

```
create_datasets <- function(long_updated) {
  # Step 4
  ## get the county indices
  county_indices <- grep(pattern = ", \\w\\w", long_updated$area_name)
  ## create the non-county data
  state_tibble <- long_updated[-county_indices,]
  ## create the county data
  county_tibble <- long_updated[county_indices,]
  ## add a class to the county tibble
  class(county_tibble) <- c("county", class(county_tibble))
  ## add a class to the state tibble
  class(state_tibble) <- c("state", class(state_tibble))
  ## print out the first 10 rows
  print("State tibble:")
  print(head(state_tibble, 10))
  print("County tibble:")
  print(head(county_tibble, 10))

  final_county_tibble <- add_state_col_county(county_tibble)
  final_state_tibble <- add_division_col_state(state_tibble)
  return(list(county = final_county_tibble, state = final_state_tibble))
}
```

Wrapper Function

The outline for this one was very helpful as it pointed me to the format. Besides the initial csv read, we don't define any variables for my other functions because we assume the output of the previous function will be used as input for the next function. This makes it easier as their are less things to input.

```
my_wrapper <- function(url, default_var_name = "enrollment"){
  result <- read_csv(url) |>
    read_and_preprocess() |>
    parse_new_variables() |>
    create_datasets()
 ## return final result
  return(result)
 }
```

Combine Function

Here we are doing a simple combination of all the specific county and state data.

```
combine_results <- function(result1, result2) {
  list(county = dplyr::bind_rows(result1$county, result2$county),
       state = dplyr::bind_rows(result1$state, result2$state))
}
```

Custom Plot Function

We created our own classes by writing custom plot functions, unique to our data.

State

For plot state function, we had to filter out all observations that had a division value of "ERROR". We then had to figure out how to group by division across the year variable. This is easily done with a group_by statement that takes in division as the first argument and then year as the second. We then summarized by using the mean of the grouped var_name variable and, we decided that a line plot with many colored lines would be the best way to visualize this. Each line's color corresponds to a division.

```r
plot.state <- function(df, var_name = "enrollment") {
  df |>
    ## filter out ERROR entries and group by division across years
    filter(division != "ERROR") |>
    group_by(division, year) |>
    ## then find the mean of var_name (default is enrollment)
    summarize(mean_val = mean(get(var_name), na.rm = TRUE)) |>
    ## plot the statistic
    ggplot(aes(x = year, y = mean_val, color = division)) +
    geom_line() +
    labs(title = paste("Mean", var_name, "across years by division"),
         y = paste("Mean", var_name),
         x = "Year")
}
```

County

To start this plot county function, a certain state had to be filtered. This helped narrow down the data set. From there we had to group by area name in order to get our mean statistics. The difficult part about arranging these statistics was that it was dependent on an inputted value so we had to imply if else logic. After that we only choose the n number of specified rows. That was now considered our sorted data but we didn't want to only use that data for the plot. Instead we had to go back to our original filtered data and filter it again to only include the area names in the top or bottom n records. To view this neatly, we decided a box plot would be the best visualization.

```r
plot.county <- function(county_tibble, var_name = "enrollment", state = "NC",
                        direction = "top", n = 5) {
  ## filter for the selected state
  filtered_state <- county_tibble |>
    filter(state == state)

  ## find the mean by area_name and sort the data
  sorted_data <- filtered_state |>
    group_by(area_name) |>
    summarize(mean_val = mean(get(var_name), na.rm = TRUE)) |>
    arrange(if (direction == "top") {
      desc(mean_val)
    } else {
      mean_val
```

```
    }) |>
    slice_head(n = n)

  ## filter for state from above
  new_sorted_data <- filtered_state |>
    filter(area_name %in% sorted_data$area_name)

  ## plot the statistic
  ggplot(new_sorted_data, aes(x = area_name, y = get(var_name))) +
    geom_boxplot() +
    labs(title = paste(direction, n, "Counties in", state),
         y = var_name,
         x = "County") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}
```

Putting it All Together

Here we put it all together using two data sets and then using a different four data sets.

Two Enrollment Datasets

The goal here was to process two different data sets and save the results to their own respective variables. After that, we combined those results so we are left with a list that contains a combined state data frame and a combined tibble data frame. From there we used our state plot function to give us mean enrollment by division over time. Then we use the county plot function to retrieve a certain number of box plots of the enrollment data for the top or bottom area names in a specified state.

```
## using data processing on two enrollment datasets
result1 <- my_wrapper("data/EDU01a.csv")
```

```
[1] "Preprocessed:"
# A tibble: 5 x 12
  area_name     STCOU EDU010187D EDU010188D EDU010189D EDU010190D EDU010191D
  <chr>         <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1 UNITED STATES 00000   40024299   39967624   40317775   40737600   41385442
2 ALABAMA       01000     733735     728234     730048     728252     725541
3 Autauga, AL   01001       6829       6900       6920       6847       7008
4 Baldwin, AL   01003      16417      16465      16799      17054      17479
```

```
5 Barbour, AL   01005       5071       5098       5068       5156       5173
# i 5 more variables: EDU010192D <dbl>, EDU010193D <dbl>, EDU010194D <dbl>,
#   EDU010195D <dbl>, EDU010196D <dbl>
[1] "Long format:"
# A tibble: 5 x 4
  area_name     STCOU survey_value enrollment
  <chr>         <chr> <chr>             <dbl>
1 UNITED STATES 00000 EDU010187D     40024299
2 UNITED STATES 00000 EDU010188D     39967624
3 UNITED STATES 00000 EDU010189D     40317775
4 UNITED STATES 00000 EDU010190D     40737600
5 UNITED STATES 00000 EDU010191D     41385442
[1] "Updated:"
# A tibble: 5 x 6
  area_name     STCOU survey_value enrollment  year measurement
  <chr>         <chr> <chr>             <dbl> <dbl> <chr>
1 UNITED STATES 00000 EDU010187D     40024299  1987 EDU0101
2 UNITED STATES 00000 EDU010188D     39967624  1988 EDU0101
3 UNITED STATES 00000 EDU010189D     40317775  1989 EDU0101
4 UNITED STATES 00000 EDU010190D     40737600  1990 EDU0101
5 UNITED STATES 00000 EDU010191D     41385442  1991 EDU0101
[1] "State tibble:"
# A tibble: 10 x 6
   area_name     STCOU survey_value enrollment  year measurement
   <chr>         <chr> <chr>             <dbl> <dbl> <chr>
 1 UNITED STATES 00000 EDU010187D     40024299  1987 EDU0101
 2 UNITED STATES 00000 EDU010188D     39967624  1988 EDU0101
 3 UNITED STATES 00000 EDU010189D     40317775  1989 EDU0101
 4 UNITED STATES 00000 EDU010190D     40737600  1990 EDU0101
 5 UNITED STATES 00000 EDU010191D     41385442  1991 EDU0101
 6 UNITED STATES 00000 EDU010192D     42088151  1992 EDU0101
 7 UNITED STATES 00000 EDU010193D     42724710  1993 EDU0101
 8 UNITED STATES 00000 EDU010194D     43369917  1994 EDU0101
 9 UNITED STATES 00000 EDU010195D     43993459  1995 EDU0101
10 UNITED STATES 00000 EDU010196D     44715737  1996 EDU0101
[1] "County tibble:"
# A tibble: 10 x 6
   area_name   STCOU survey_value enrollment  year measurement
   <chr>       <chr> <chr>             <dbl> <dbl> <chr>
 1 Autauga, AL 01001 EDU010187D         6829  1987 EDU0101
 2 Autauga, AL 01001 EDU010188D         6900  1988 EDU0101
 3 Autauga, AL 01001 EDU010189D         6920  1989 EDU0101
 4 Autauga, AL 01001 EDU010190D         6847  1990 EDU0101
```

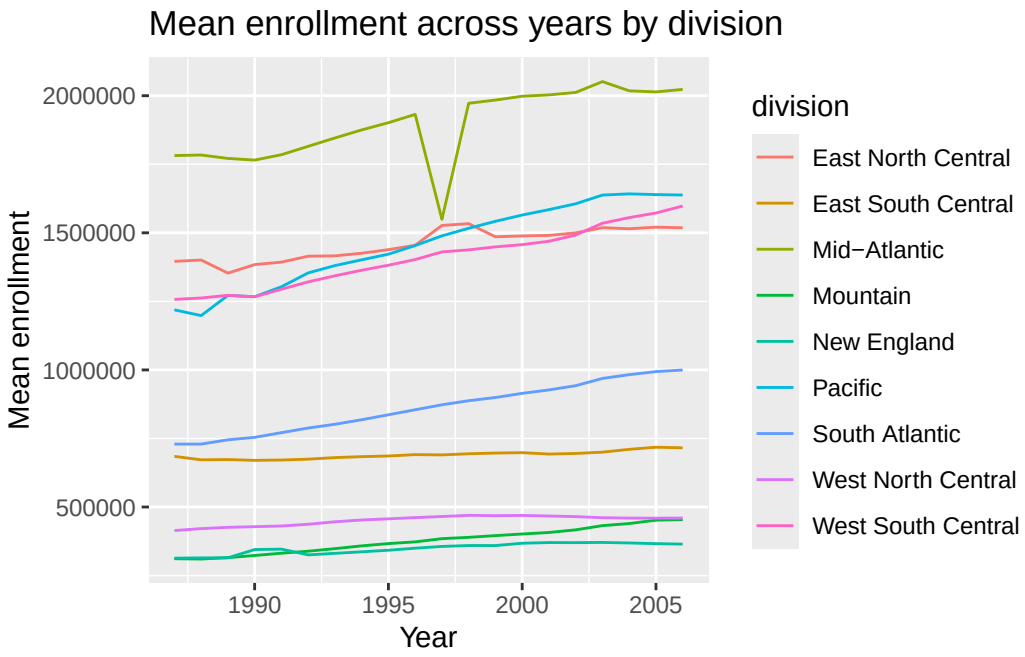
```
 5 Autauga, AL 01001 EDU010191D         7008  1991 EDU0101
 6 Autauga, AL 01001 EDU010192D         7137  1992 EDU0101
 7 Autauga, AL 01001 EDU010193D         7152  1993 EDU0101
 8 Autauga, AL 01001 EDU010194D         7381  1994 EDU0101
 9 Autauga, AL 01001 EDU010195D         7568  1995 EDU0101
10 Autauga, AL 01001 EDU010196D         7834  1996 EDU0101
```

```r
result2 <- my_wrapper("data/EDU01b.csv")
```

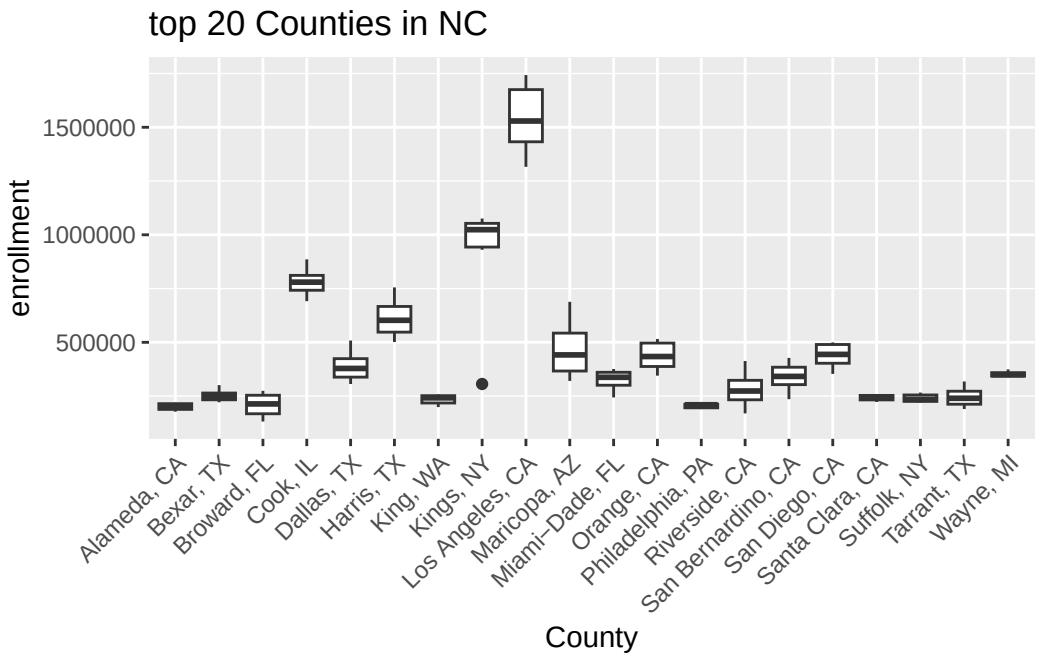
```
[1] "Preprocessed:"
# A tibble: 5 x 12
  area_name     STCOU EDU010197D EDU010198D EDU010199D EDU010200D EDU010201D
  <chr>         <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1 UNITED STATES 00000   44534459   46245814   46368903   46818690   47127066
2 ALABAMA       01000     737386     739321     737639     731613     730627
3 Autauga, AL   01001       8099       8211       8489       8912       8626
4 Baldwin, AL   01003      21410      21771      22176      22337      22656
5 Barbour, AL   01005       5100       5024       4906       4793       4671
# i 5 more variables: EDU010202D <dbl>, EDU015203D <dbl>, EDU015204D <dbl>,
#   EDU015205D <dbl>, EDU015206D <dbl>
[1] "Long format:"
# A tibble: 5 x 4
  area_name     STCOU survey_value enrollment
  <chr>         <chr> <chr>             <dbl>
1 UNITED STATES 00000 EDU010197D     44534459
2 UNITED STATES 00000 EDU010198D     46245814
3 UNITED STATES 00000 EDU010199D     46368903
4 UNITED STATES 00000 EDU010200D     46818690
5 UNITED STATES 00000 EDU010201D     47127066
[1] "Updated:"
# A tibble: 5 x 6
  area_name     STCOU survey_value enrollment  year measurement
  <chr>         <chr> <chr>             <dbl> <dbl> <chr>
1 UNITED STATES 00000 EDU010197D     44534459  1997 EDU0101
2 UNITED STATES 00000 EDU010198D     46245814  1998 EDU0101
3 UNITED STATES 00000 EDU010199D     46368903  1999 EDU0101
4 UNITED STATES 00000 EDU010200D     46818690  2000 EDU0102
5 UNITED STATES 00000 EDU010201D     47127066  2001 EDU0102
[1] "State tibble:"
# A tibble: 10 x 6
   area_name     STCOU survey_value enrollment  year measurement
```

```
   <chr>         <chr> <chr>             <dbl> <dbl> <chr>
 1 UNITED STATES 00000 EDU010197D     44534459  1997 EDU0101
 2 UNITED STATES 00000 EDU010198D     46245814  1998 EDU0101
 3 UNITED STATES 00000 EDU010199D     46368903  1999 EDU0101
 4 UNITED STATES 00000 EDU010200D     46818690  2000 EDU0102
 5 UNITED STATES 00000 EDU010201D     47127066  2001 EDU0102
 6 UNITED STATES 00000 EDU010202D     47606570  2002 EDU0102
 7 UNITED STATES 00000 EDU015203D     48506317  2003 EDU0152
 8 UNITED STATES 00000 EDU015204D     48693287  2004 EDU0152
 9 UNITED STATES 00000 EDU015205D     48978555  2005 EDU0152
10 UNITED STATES 00000 EDU015206D     49140702  2006 EDU0152
[1] "County tibble:"
# A tibble: 10 x 6
   area_name   STCOU survey_value enrollment  year measurement
   <chr>       <chr> <chr>             <dbl> <dbl> <chr>
 1 Autauga, AL 01001 EDU010197D         8099  1997 EDU0101
 2 Autauga, AL 01001 EDU010198D         8211  1998 EDU0101
 3 Autauga, AL 01001 EDU010199D         8489  1999 EDU0101
 4 Autauga, AL 01001 EDU010200D         8912  2000 EDU0102
 5 Autauga, AL 01001 EDU010201D         8626  2001 EDU0102
 6 Autauga, AL 01001 EDU010202D         8762  2002 EDU0102
 7 Autauga, AL 01001 EDU015203D         9105  2003 EDU0152
 8 Autauga, AL 01001 EDU015204D         9200  2004 EDU0152
 9 Autauga, AL 01001 EDU015205D         9559  2005 EDU0152
10 Autauga, AL 01001 EDU015206D         9652  2006 EDU0152
```

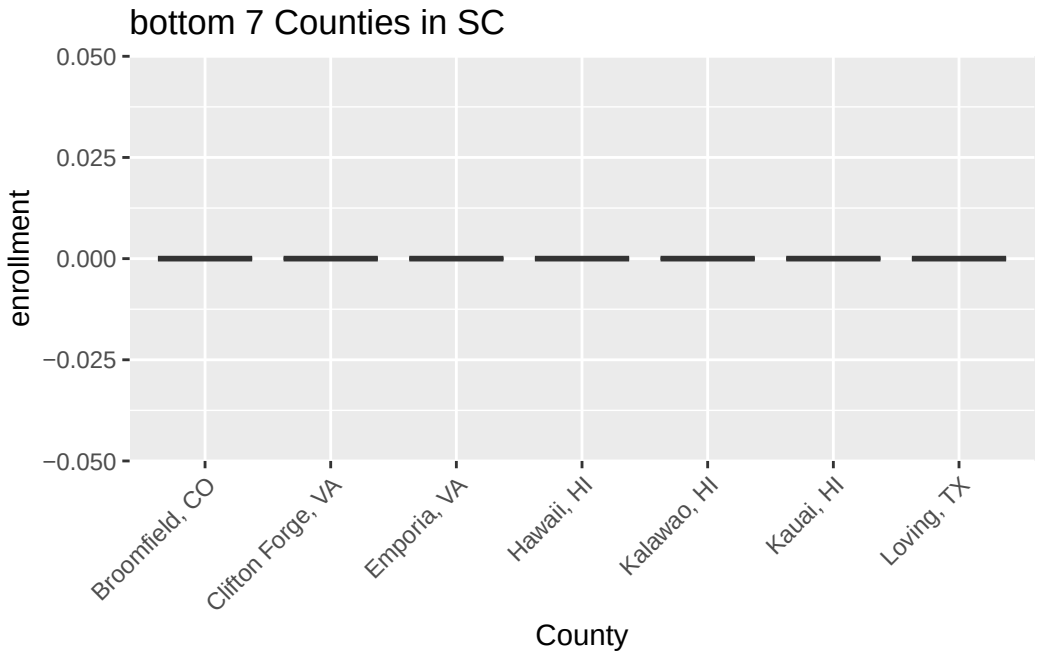
```r
## combining data sets
combined_results <- combine_results(result1, result2)

## use plot function on state
plot(combined_results$state)
```



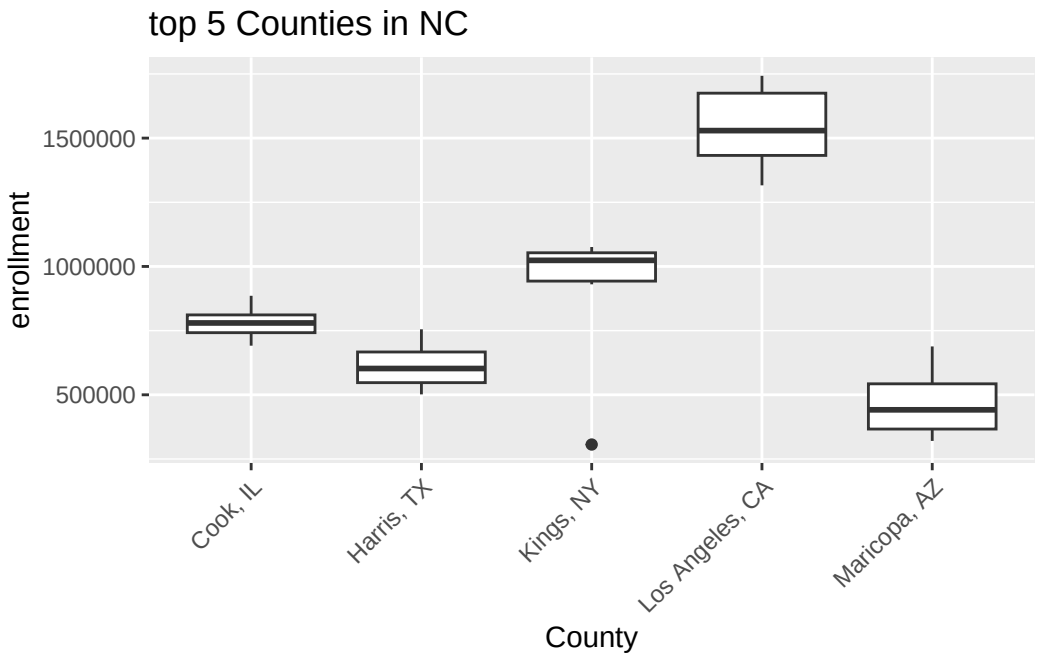

```
## use plot on county data
## scenario one
plot(combined_results$county, state = "NC", direction="top", n = 20)
```



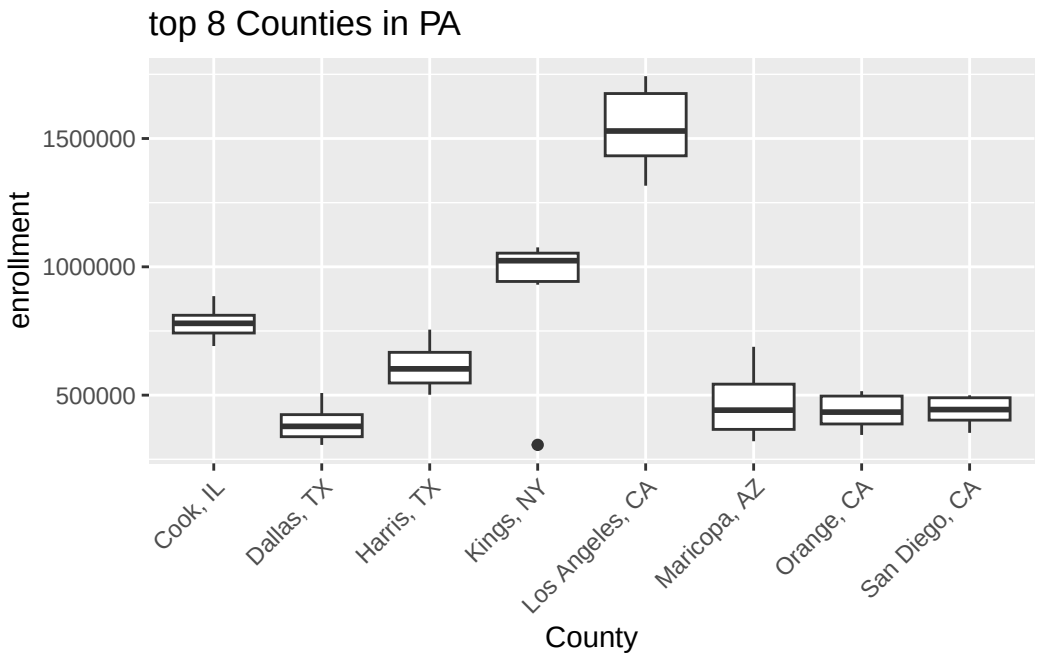

```
## scenario two
plot(combined_results$county, state = "SC", direction="bottom", n = 7)
```



```r
## scenario three
plot(combined_results$county)
```




```r
##scenario four
plot(combined_results$county, state = "PA", direction="top", n = 8)
```



Four Additional Data Sets

The goal here was to process four additional data sets and save those into four respective variables. Then, two at a time, the results were combined into two new results called a_prime and b_prime. Lastly a_prime and b_prime were combined into one final result variable which contained all four additional data sets. Then we used the state plot function and the county plot function. The county plot function was called four times with four different combinations of arguments.

```
## using data processing on four additional datasets
a <- my_wrapper("data/PST01a.csv")
```

```
[1] "Preprocessed:"
# A tibble: 5 x 12
  area_name     STCOU PST015171D PST015172D PST015173D PST015174D PST015175D
  <chr>         <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1 UNITED STATES 00000  206827028  209283904  211357490  213341552  215465246
2 ALABAMA       01000    3497452    3540080    3580769    3627805    3680533
3 Autauga, AL   01001      25508      27166      28463      29266      29718
4 Baldwin, AL   01003      60141      62435      64195      66071      67860
5 Barbour, AL   01005      23092      22854      23457      23432      24869
```

```
# i 5 more variables: PST015176D <dbl>, PST015177D <dbl>, PST015178D <dbl>,
#   PST015179D <dbl>, PST025181D <dbl>
[1] "Long format:"
# A tibble: 5 x 4
  area_name     STCOU survey_value enrollment
  <chr>         <chr> <chr>             <dbl>
1 UNITED STATES 00000 PST015171D    206827028
2 UNITED STATES 00000 PST015172D    209283904
3 UNITED STATES 00000 PST015173D    211357490
4 UNITED STATES 00000 PST015174D    213341552
5 UNITED STATES 00000 PST015175D    215465246
[1] "Updated:"
# A tibble: 5 x 6
  area_name     STCOU survey_value enrollment  year measurement
  <chr>         <chr> <chr>             <dbl> <dbl> <chr>
1 UNITED STATES 00000 PST015171D    206827028  1971 PST0151
2 UNITED STATES 00000 PST015172D    209283904  1972 PST0151
3 UNITED STATES 00000 PST015173D    211357490  1973 PST0151
4 UNITED STATES 00000 PST015174D    213341552  1974 PST0151
5 UNITED STATES 00000 PST015175D    215465246  1975 PST0151
[1] "State tibble:"
# A tibble: 10 x 6
   area_name     STCOU survey_value enrollment  year measurement
   <chr>         <chr> <chr>             <dbl> <dbl> <chr>
 1 UNITED STATES 00000 PST015171D    206827028  1971 PST0151
 2 UNITED STATES 00000 PST015172D    209283904  1972 PST0151
 3 UNITED STATES 00000 PST015173D    211357490  1973 PST0151
 4 UNITED STATES 00000 PST015174D    213341552  1974 PST0151
 5 UNITED STATES 00000 PST015175D    215465246  1975 PST0151
 6 UNITED STATES 00000 PST015176D    217562728  1976 PST0151
 7 UNITED STATES 00000 PST015177D    219759860  1977 PST0151
 8 UNITED STATES 00000 PST015178D    222095080  1978 PST0151
 9 UNITED STATES 00000 PST015179D    224567234  1979 PST0151
10 UNITED STATES 00000 PST025181D    229466391  1981 PST0251
[1] "County tibble:"
# A tibble: 10 x 6
   area_name   STCOU survey_value enrollment  year measurement
   <chr>       <chr> <chr>             <dbl> <dbl> <chr>
 1 Autauga, AL 01001 PST015171D        25508  1971 PST0151
 2 Autauga, AL 01001 PST015172D        27166  1972 PST0151
 3 Autauga, AL 01001 PST015173D        28463  1973 PST0151
 4 Autauga, AL 01001 PST015174D        29266  1974 PST0151
 5 Autauga, AL 01001 PST015175D        29718  1975 PST0151
```

```
 6 Autauga, AL 01001 PST015176D        29896  1976 PST0151
 7 Autauga, AL 01001 PST015177D        30462  1977 PST0151
 8 Autauga, AL 01001 PST015178D        30882  1978 PST0151
 9 Autauga, AL 01001 PST015179D        32055  1979 PST0151
10 Autauga, AL 01001 PST025181D        31985  1981 PST0251
```

```
b <- my_wrapper("data/PST01b.csv")
```

```
[1] "Preprocessed:"
# A tibble: 5 x 12
  area_name     STCOU PST025182D PST025183D PST025184D PST025185D PST025186D
  <chr>         <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1 UNITED STATES 00000  231665106  233792697  235825544  237924311  240133472
2 ALABAMA       01000    3925328    3934100    3951766    3972539    3991552
3 Autauga, AL   01001      32038      32057      32130      32248      32895
4 Baldwin, AL   01003      82330      83980      86753      89403      91308
5 Barbour, AL   01005      24775      24796      24954      25001      24942
# i 5 more variables: PST025187D <dbl>, PST025188D <dbl>, PST025189D <dbl>,
#   PST030190D <dbl>, PST035190D <dbl>
[1] "Long format:"
# A tibble: 5 x 4
  area_name     STCOU survey_value enrollment
  <chr>         <chr> <chr>             <dbl>
1 UNITED STATES 00000 PST025182D    231665106
2 UNITED STATES 00000 PST025183D    233792697
3 UNITED STATES 00000 PST025184D    235825544
4 UNITED STATES 00000 PST025185D    237924311
5 UNITED STATES 00000 PST025186D    240133472
[1] "Updated:"
# A tibble: 5 x 6
  area_name     STCOU survey_value enrollment  year measurement
  <chr>         <chr> <chr>             <dbl> <dbl> <chr>
1 UNITED STATES 00000 PST025182D    231665106  1982 PST0251
2 UNITED STATES 00000 PST025183D    233792697  1983 PST0251
3 UNITED STATES 00000 PST025184D    235825544  1984 PST0251
4 UNITED STATES 00000 PST025185D    237924311  1985 PST0251
5 UNITED STATES 00000 PST025186D    240133472  1986 PST0251
[1] "State tibble:"
# A tibble: 10 x 6
   area_name     STCOU survey_value enrollment  year measurement
   <chr>         <chr> <chr>             <dbl> <dbl> <chr>
```

```
 1 UNITED STATES 00000 PST025182D    231665106  1982 PST0251
 2 UNITED STATES 00000 PST025183D    233792697  1983 PST0251
 3 UNITED STATES 00000 PST025184D    235825544  1984 PST0251
 4 UNITED STATES 00000 PST025185D    237924311  1985 PST0251
 5 UNITED STATES 00000 PST025186D    240133472  1986 PST0251
 6 UNITED STATES 00000 PST025187D    242289738  1987 PST0251
 7 UNITED STATES 00000 PST025188D    244499776  1988 PST0251
 8 UNITED STATES 00000 PST025189D    246819839  1989 PST0251
 9 UNITED STATES 00000 PST030190D    248790925  1990 PST0301
10 UNITED STATES 00000 PST035190D    249622814  1990 PST0351
[1] "County tibble:"
# A tibble: 10 x 6
   area_name   STCOU survey_value enrollment  year measurement
   <chr>       <chr> <chr>             <dbl> <dbl> <chr>
 1 Autauga, AL 01001 PST025182D        32038  1982 PST0251
 2 Autauga, AL 01001 PST025183D        32057  1983 PST0251
 3 Autauga, AL 01001 PST025184D        32130  1984 PST0251
 4 Autauga, AL 01001 PST025185D        32248  1985 PST0251
 5 Autauga, AL 01001 PST025186D        32895  1986 PST0251
 6 Autauga, AL 01001 PST025187D        33266  1987 PST0251
 7 Autauga, AL 01001 PST025188D        33637  1988 PST0251
 8 Autauga, AL 01001 PST025189D        33996  1989 PST0251
 9 Autauga, AL 01001 PST030190D        34222  1990 PST0301
10 Autauga, AL 01001 PST035190D        34353  1990 PST0351
```

```
c <- my_wrapper("data/PST01c.csv")
```

```
[1] "Preprocessed:"
# A tibble: 5 x 12
  area_name     STCOU PST035191D PST035192D PST035193D PST035194D PST035195D
  <chr>         <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1 UNITED STATES 00000  252980941  256514224  259918588  263125821  266278393
2 ALABAMA       01000    4099156    4154014    4214202    4260229    4296800
3 Autauga, AL   01001      35010      35985      36953      38186      39112
4 Baldwin, AL   01003     102420     106595     111416     116565     120896
5 Barbour, AL   01005      26506      26941      27371      27751      27854
# i 5 more variables: PST035196D <dbl>, PST035197D <dbl>, PST035198D <dbl>,
#   PST035199D <dbl>, PST040200D <dbl>
[1] "Long format:"
# A tibble: 5 x 4
  area_name     STCOU survey_value enrollment
```

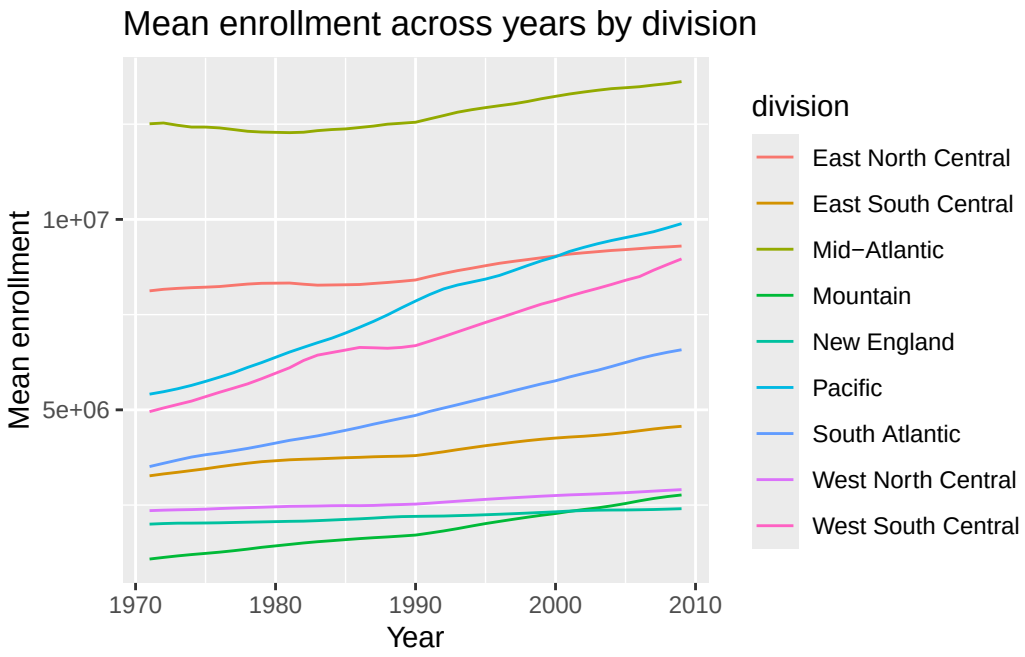
```
  <chr>         <chr> <chr>            <dbl>
1 UNITED STATES 00000 PST035191D    252980941
2 UNITED STATES 00000 PST035192D    256514224
3 UNITED STATES 00000 PST035193D    259918588
4 UNITED STATES 00000 PST035194D    263125821
5 UNITED STATES 00000 PST035195D    266278393
[1] "Updated:"
# A tibble: 5 x 6
  area_name     STCOU survey_value enrollment  year measurement
  <chr>         <chr> <chr>            <dbl> <dbl> <chr>
1 UNITED STATES 00000 PST035191D    252980941  1991 PST0351
2 UNITED STATES 00000 PST035192D    256514224  1992 PST0351
3 UNITED STATES 00000 PST035193D    259918588  1993 PST0351
4 UNITED STATES 00000 PST035194D    263125821  1994 PST0351
5 UNITED STATES 00000 PST035195D    266278393  1995 PST0351
[1] "State tibble:"
# A tibble: 10 x 6
   area_name     STCOU survey_value enrollment  year measurement
   <chr>         <chr> <chr>            <dbl> <dbl> <chr>
 1 UNITED STATES 00000 PST035191D    252980941  1991 PST0351
 2 UNITED STATES 00000 PST035192D    256514224  1992 PST0351
 3 UNITED STATES 00000 PST035193D    259918588  1993 PST0351
 4 UNITED STATES 00000 PST035194D    263125821  1994 PST0351
 5 UNITED STATES 00000 PST035195D    266278393  1995 PST0351
 6 UNITED STATES 00000 PST035196D    269394284  1996 PST0351
 7 UNITED STATES 00000 PST035197D    272646925  1997 PST0351
 8 UNITED STATES 00000 PST035198D    275854104  1998 PST0351
 9 UNITED STATES 00000 PST035199D    279040168  1999 PST0351
10 UNITED STATES 00000 PST040200D    281424602  2000 PST0402
[1] "County tibble:"
# A tibble: 10 x 6
   area_name   STCOU survey_value enrollment  year measurement
   <chr>       <chr> <chr>            <dbl> <dbl> <chr>
 1 Autauga, AL 01001 PST035191D        35010  1991 PST0351
 2 Autauga, AL 01001 PST035192D        35985  1992 PST0351
 3 Autauga, AL 01001 PST035193D        36953  1993 PST0351
 4 Autauga, AL 01001 PST035194D        38186  1994 PST0351
 5 Autauga, AL 01001 PST035195D        39112  1995 PST0351
 6 Autauga, AL 01001 PST035196D        40207  1996 PST0351
 7 Autauga, AL 01001 PST035197D        41238  1997 PST0351
 8 Autauga, AL 01001 PST035198D        42106  1998 PST0351
 9 Autauga, AL 01001 PST035199D        42963  1999 PST0351
10 Autauga, AL 01001 PST040200D        43671  2000 PST0402
```

```r
d <- my_wrapper("data/PST01d.csv")
```

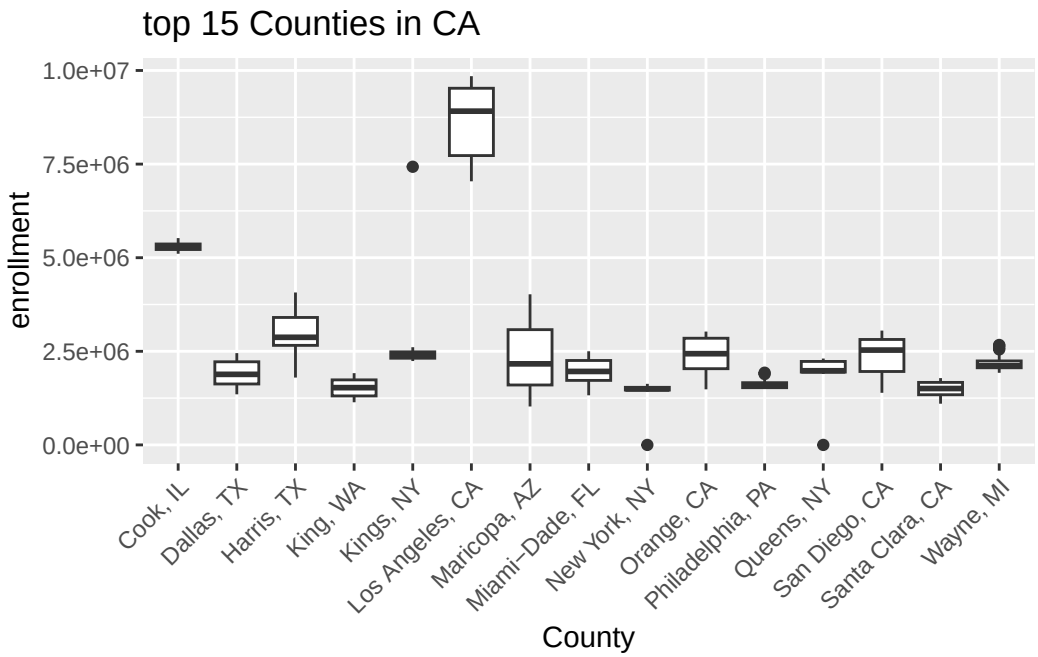
```
[1] "Preprocessed:"
# A tibble: 5 x 12
  area_name     STCOU PST045200D PST045201D PST045202D PST045203D PST045204D
  <chr>         <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1 UNITED STATES 00000  282171957  285081556  287803914  290326418  293045739
2 ALABAMA       01000    4451849    4464034    4472420    4490591    4512190
3 Autauga, AL   01001      43872      44434      45157      45762      46933
4 Baldwin, AL   01003     141358     144988     148141     151707     156573
5 Barbour, AL   01005      29035      29223      29289      29480      29458
# i 5 more variables: PST045205D <dbl>, PST045206D <dbl>, PST045207D <dbl>,
#   PST045208D <dbl>, PST045209D <dbl>
[1] "Long format:"
# A tibble: 5 x 4
  area_name     STCOU survey_value enrollment
  <chr>         <chr> <chr>             <dbl>
1 UNITED STATES 00000 PST045200D    282171957
2 UNITED STATES 00000 PST045201D    285081556
3 UNITED STATES 00000 PST045202D    287803914
4 UNITED STATES 00000 PST045203D    290326418
5 UNITED STATES 00000 PST045204D    293045739
[1] "Updated:"
# A tibble: 5 x 6
  area_name     STCOU survey_value enrollment  year measurement
  <chr>         <chr> <chr>             <dbl> <dbl> <chr>
1 UNITED STATES 00000 PST045200D    282171957  2000 PST0452
2 UNITED STATES 00000 PST045201D    285081556  2001 PST0452
3 UNITED STATES 00000 PST045202D    287803914  2002 PST0452
4 UNITED STATES 00000 PST045203D    290326418  2003 PST0452
5 UNITED STATES 00000 PST045204D    293045739  2004 PST0452
[1] "State tibble:"
# A tibble: 10 x 6
   area_name     STCOU survey_value enrollment  year measurement
   <chr>         <chr> <chr>             <dbl> <dbl> <chr>
 1 UNITED STATES 00000 PST045200D    282171957  2000 PST0452
 2 UNITED STATES 00000 PST045201D    285081556  2001 PST0452
 3 UNITED STATES 00000 PST045202D    287803914  2002 PST0452
 4 UNITED STATES 00000 PST045203D    290326418  2003 PST0452
 5 UNITED STATES 00000 PST045204D    293045739  2004 PST0452
 6 UNITED STATES 00000 PST045205D    295753151  2005 PST0452
 7 UNITED STATES 00000 PST045206D    298593212  2006 PST0452
```

```
 8 UNITED STATES 00000 PST045207D    301579895  2007 PST0452
 9 UNITED STATES 00000 PST045208D    304374846  2008 PST0452
10 UNITED STATES 00000 PST045209D    307006550  2009 PST0452
[1] "County tibble:"
# A tibble: 10 x 6
   area_name   STCOU survey_value enrollment  year measurement
   <chr>       <chr> <chr>             <dbl> <dbl> <chr>
 1 Autauga, AL 01001 PST045200D        43872  2000 PST0452
 2 Autauga, AL 01001 PST045201D        44434  2001 PST0452
 3 Autauga, AL 01001 PST045202D        45157  2002 PST0452
 4 Autauga, AL 01001 PST045203D        45762  2003 PST0452
 5 Autauga, AL 01001 PST045204D        46933  2004 PST0452
 6 Autauga, AL 01001 PST045205D        47870  2005 PST0452
 7 Autauga, AL 01001 PST045206D        49105  2006 PST0452
 8 Autauga, AL 01001 PST045207D        49834  2007 PST0452
 9 Autauga, AL 01001 PST045208D        50354  2008 PST0452
10 Autauga, AL 01001 PST045209D        50756  2009 PST0452
```

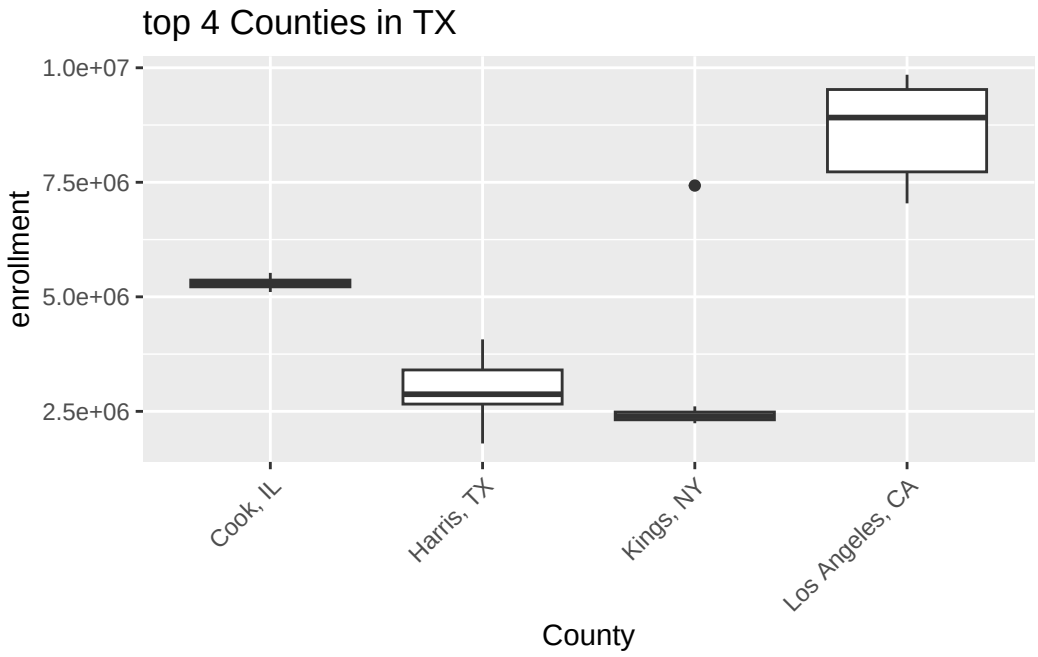
```r
## combine four datasets into one
a_prime <- combine_results(a, b)
b_prime <- combine_results(c, d)
four_combined_results <-combine_results(a_prime, b_prime)

## use plot function on state
plot(four_combined_results$state)
```



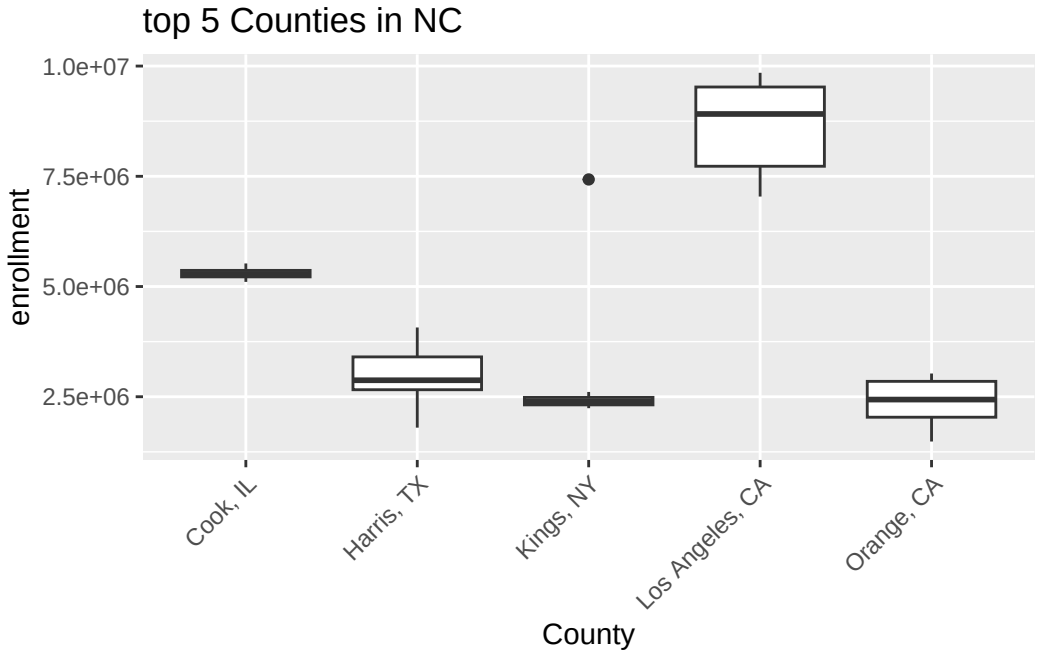

```
## use plot on county data
## scenario one
plot(four_combined_results$county, state = "CA", direction="top", n = 15)
```



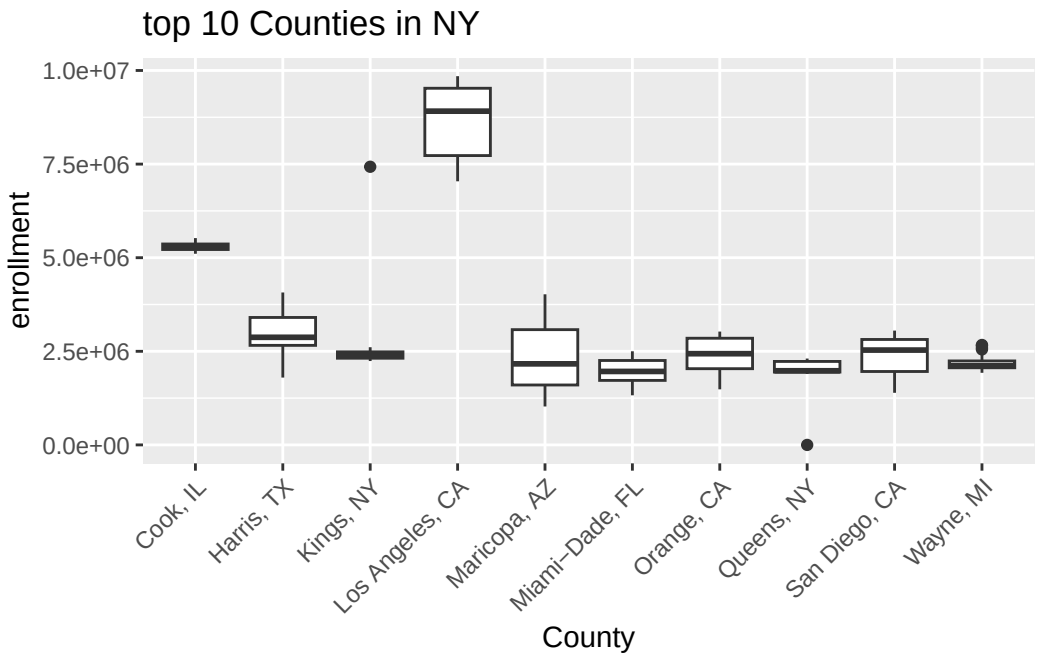

```
## scenario two
plot(four_combined_results$county, state = "TX", direction="top", n = 4)
```



```
## scenario three
plot(four_combined_results$county)
```




```
##scenario four
plot(four_combined_results$county, state = "NY", direction="top", n = 10)
```


top 10 Counties in NY
1.0e+07
7.5e+06
5.0e+06
2.5e+06
0.0e+00
enrollment
Cook, IL
Harris, TX
Kings, NY
Los Angeles, CA
Maricopa, AZ
Miami−Dade, FL
Orange, CA
Queens, NY
San Diego, CA
Wayne, MI
County