Project 1

Trever Yoder and Koji Takagi

Load Packages and Functions

In this section, we load all necessary libraries and our custom functions file.

```
library(tidyverse)
library(readr)
library(ggplot2)

# Load custom functions
source("functions.R")
```

Task 1: Data Processing

Question 1: Read in the dataset

We want to read in some of this Census data set, but not all of it. Here we specify which columns we want to read in and we named this data set: df_selected. We then slice the first 5 lines to display them to confirm we read the data in correctly.

```
#Read in the data while selecting specific columns
df_selected <- read_csv("https://www4.stat.ncsu.edu/~online/datasets/EDU01a.csv", show_col_ty
  select(Area_name, STCOU, ends_with("D")) %>% #select specified columns
  rename(area_name = Area_name) #rename "Area_name" as directed

#Display the first 5 lines
df_selected %>%
slice(1:5)
```

```
# 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>
```

Question 2

Now we want to convert the data into long format where each row has only one enrollment value for area_name. This converted data will be called df_long. We then display the first 5 rows to make sure everything looks as expected.

```r
df_long <- pivot_longer(
  df_selected,
  cols = ends_with("D"),
  names_to = "Survey",
  values_to = "Enrollment Value"
)

#Display the first 5 lines
df_long %>%
  slice(1:5)
```

```
# A tibble: 5 x 4
  area_name     STCOU Survey     `Enrollment Value`
  <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
```

Process the EDU Data Sets

We run our wrapper function on the two EDU datasets and inspect the results.

```
edu1 <- my_wrapper("https://www4.stat.ncsu.edu/~online/datasets/EDU01a.csv", value = "Enrollm
edu2 <- my_wrapper("https://www4.stat.ncsu.edu/~online/datasets/EDU01b.csv", value = "Enrollm

# Inspect to ensure correctness
head(edu1$county)
```

```
# A tibble: 6 x 7
  area_name   STCOU Survey     `Enrollment Value`  Year Measurement State
  <chr>       <chr> <chr>                   <dbl> <dbl> <chr>       <chr>
1 Autauga, AL 01001 EDU010187D               6829  1987 EDU0101     AL
2 Autauga, AL 01001 EDU010188D               6900  1988 EDU0101     AL
3 Autauga, AL 01001 EDU010189D               6920  1989 EDU0101     AL
4 Autauga, AL 01001 EDU010190D               6847  1990 EDU0101     AL
5 Autauga, AL 01001 EDU010191D               7008  1991 EDU0101     AL
6 Autauga, AL 01001 EDU010192D               7137  1992 EDU0101     AL
```

```
head(edu1$noncounty)
```

```
# A tibble: 6 x 7
  area_name     STCOU Survey     `Enrollment Value`  Year Measurement Division
  <chr>         <chr> <chr>                   <dbl> <dbl> <chr>       <chr>
1 UNITED STATES 00000 EDU010187D           40024299  1987 EDU0101     ERROR
2 UNITED STATES 00000 EDU010188D           39967624  1988 EDU0101     ERROR
3 UNITED STATES 00000 EDU010189D           40317775  1989 EDU0101     ERROR
4 UNITED STATES 00000 EDU010190D           40737600  1990 EDU0101     ERROR
5 UNITED STATES 00000 EDU010191D           41385442  1991 EDU0101     ERROR
6 UNITED STATES 00000 EDU010192D           42088151  1992 EDU0101     ERROR
```

Question 3: Combine EDU Data Sets

Here we use our combining function to merge the two processed data sets.

```
edu_combined <- combine_wrapper_results(edu1, edu2)
head(edu_combined$county)
```

```
# A tibble: 6 x 7
  area_name   STCOU Survey     `Enrollment Value`  Year Measurement State
  <chr>       <chr> <chr>                   <dbl> <dbl> <chr>       <chr>
1 Autauga, AL 01001 EDU010187D               6829  1987 EDU0101     AL
```

```
2 Autauga, AL 01001 EDU010188D               6900  1988 EDU0101     AL
3 Autauga, AL 01001 EDU010189D               6920  1989 EDU0101     AL
4 Autauga, AL 01001 EDU010190D               6847  1990 EDU0101     AL
5 Autauga, AL 01001 EDU010191D               7008  1991 EDU0101     AL
6 Autauga, AL 01001 EDU010192D               7137  1992 EDU0101     AL
```

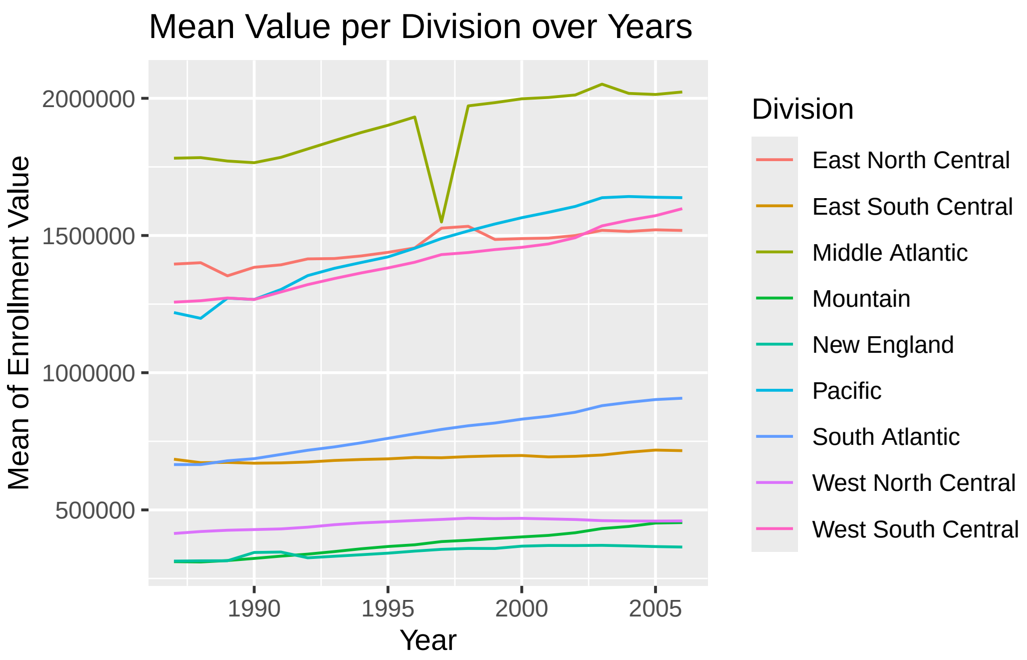
```r
head(edu_combined$noncounty)
```

```
# A tibble: 6 x 7
  area_name     STCOU Survey     `Enrollment Value`  Year Measurement Division
  <chr>         <chr> <chr>                   <dbl> <dbl> <chr>       <chr>
1 UNITED STATES 00000 EDU010187D           40024299  1987 EDU0101     ERROR
2 UNITED STATES 00000 EDU010188D           39967624  1988 EDU0101     ERROR
3 UNITED STATES 00000 EDU010189D           40317775  1989 EDU0101     ERROR
4 UNITED STATES 00000 EDU010190D           40737600  1990 EDU0101     ERROR
5 UNITED STATES 00000 EDU010191D           41385442  1991 EDU0101     ERROR
6 UNITED STATES 00000 EDU010192D           42088151  1992 EDU0101     ERROR
```

Question 4: State Plot for EDU Data

This plot shows the mean enrollment by Division across years.

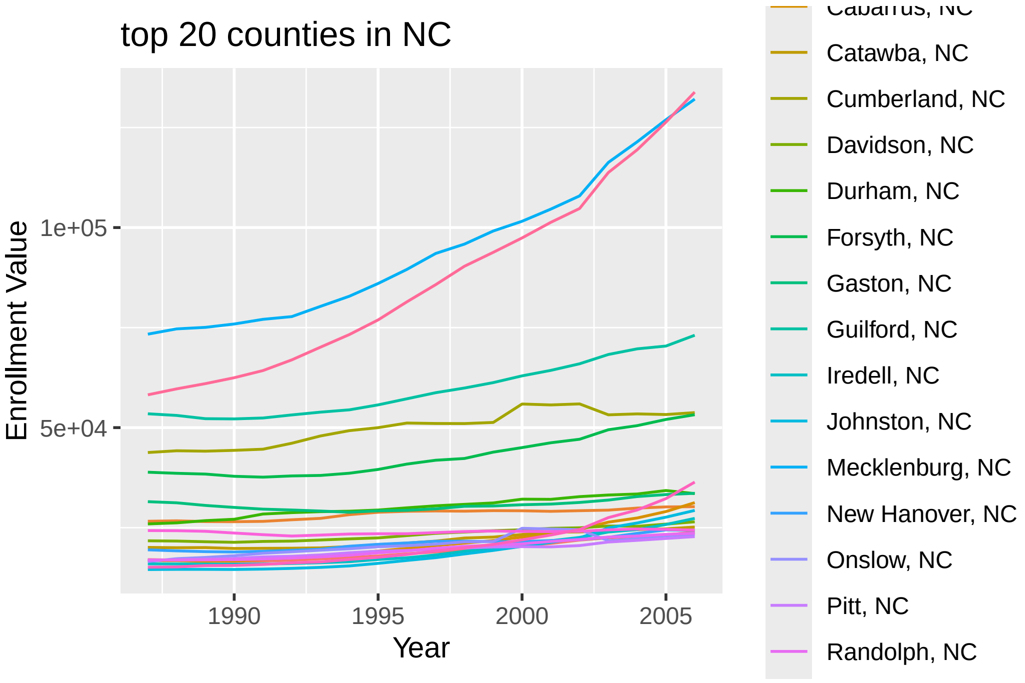
```r
plot(edu_combined$noncounty, var_name = "Enrollment Value")
```



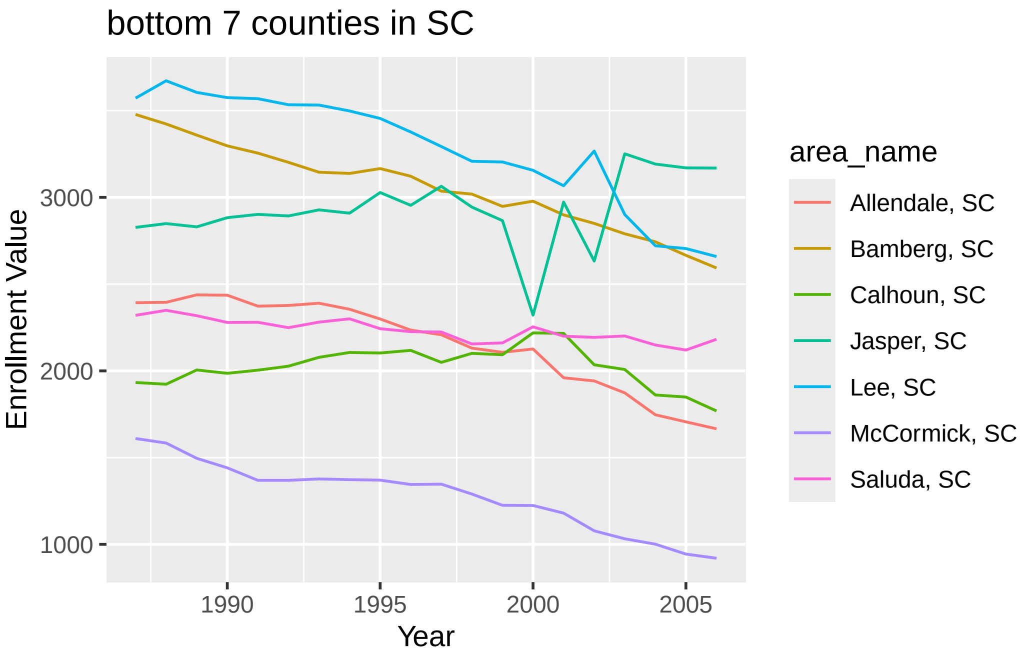
Question 5: County Plots for EDU Data

Below are various plots for county data, demonstrating flexibility in selecting state, top/bottom, and count.

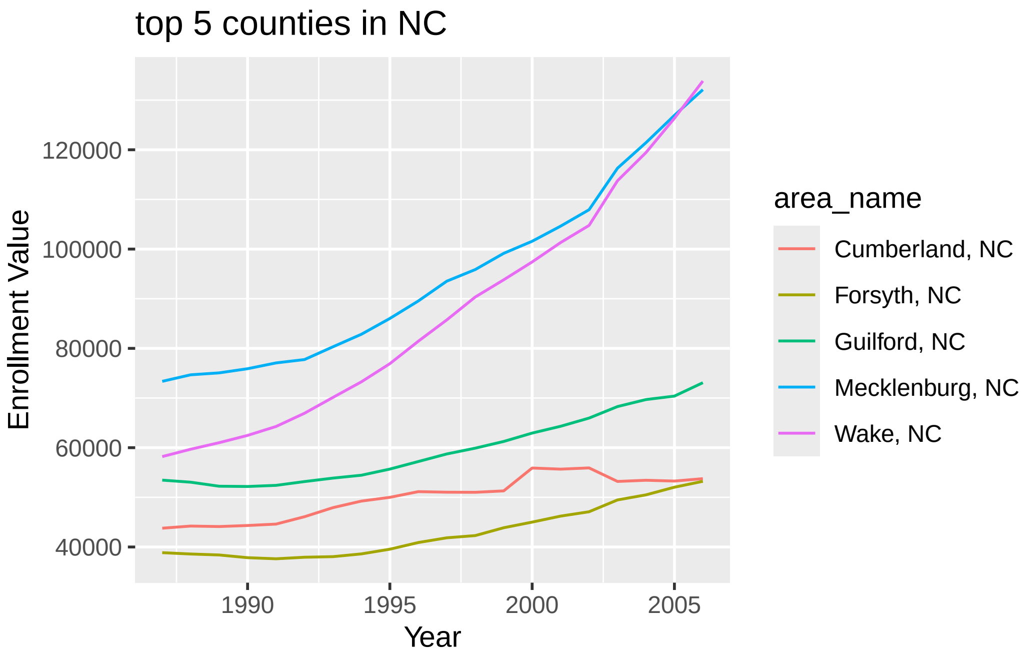
```
# NC, top 20
plot(edu_combined$county, var_name = "Enrollment Value", state = "NC", top_or_bottom = "top"
```



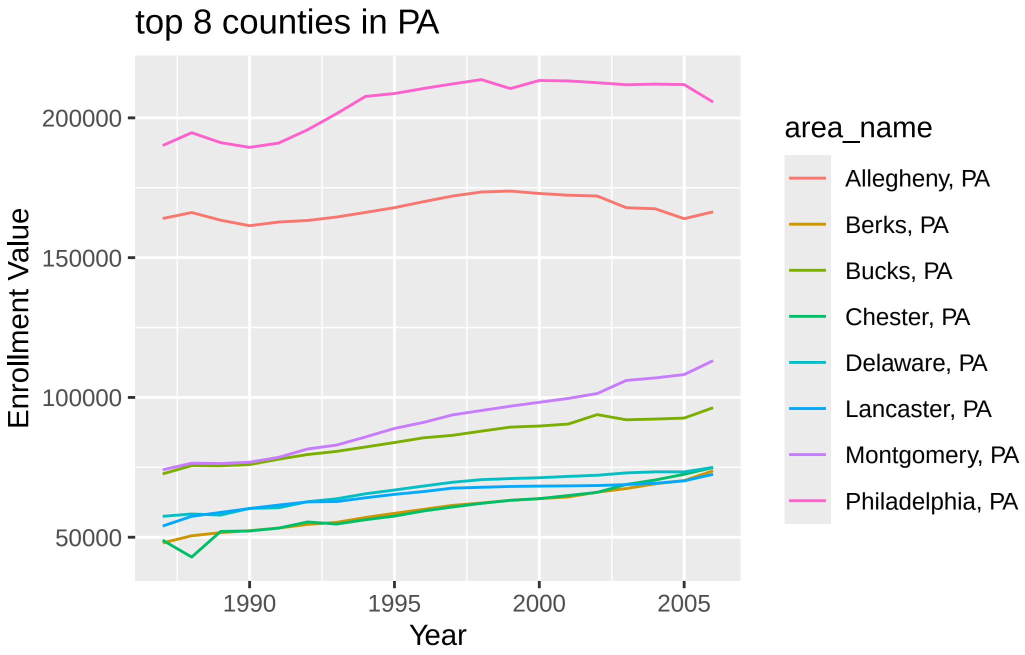
```
# SC, bottom 7
plot(edu_combined$county, var_name = "Enrollment Value", state = "SC", top_or_bottom = "botto
```



```
# Default (uses NC top 5)
plot(edu_combined$county, var_name = "Enrollment Value")
```



```
# PA, top 8
plot(edu_combined$county, var_name = "Enrollment Value", state = "PA", top_or_bottom = "top"
```



Question 6: Process PST Data Sets

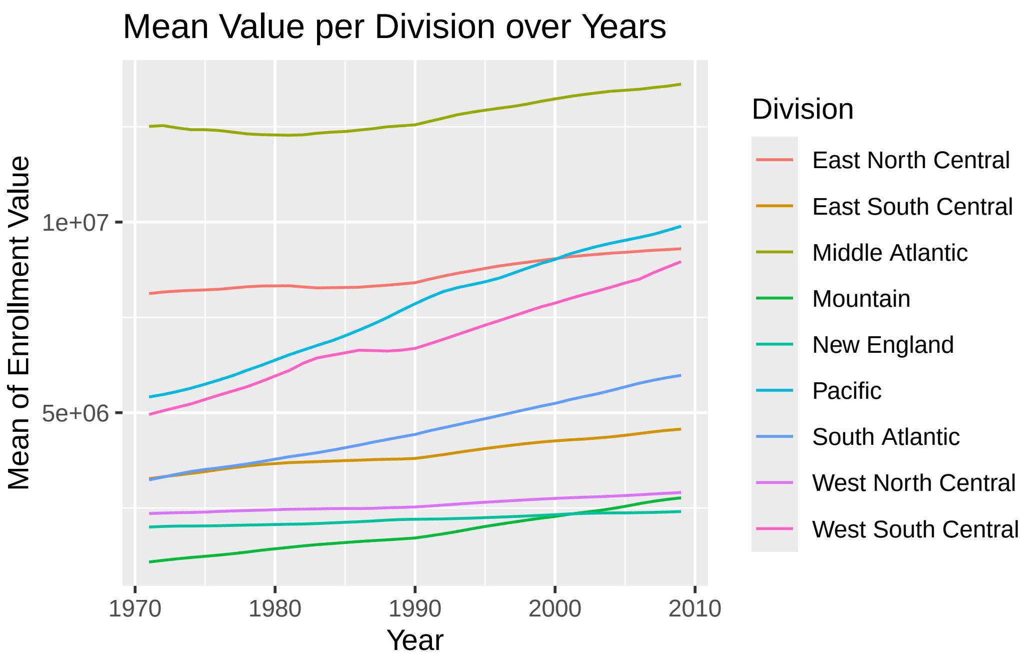
We repeat the same workflow for the four PST datasets.

```
pst1 <- my_wrapper("https://www4.stat.ncsu.edu/~online/datasets/PST01a.csv", value = "Enroll
pst2 <- my_wrapper("https://www4.stat.ncsu.edu/~online/datasets/PST01b.csv", value = "Enroll
pst3 <- my_wrapper("https://www4.stat.ncsu.edu/~online/datasets/PST01c.csv", value = "Enroll
pst4 <- my_wrapper("https://www4.stat.ncsu.edu/~online/datasets/PST01d.csv", value = "Enroll

# Combine step by step
pst12 <- combine_wrapper_results(pst1, pst2)
pst34 <- combine_wrapper_results(pst3, pst4)
pst_combined <- combine_wrapper_results(pst12, pst34)
```

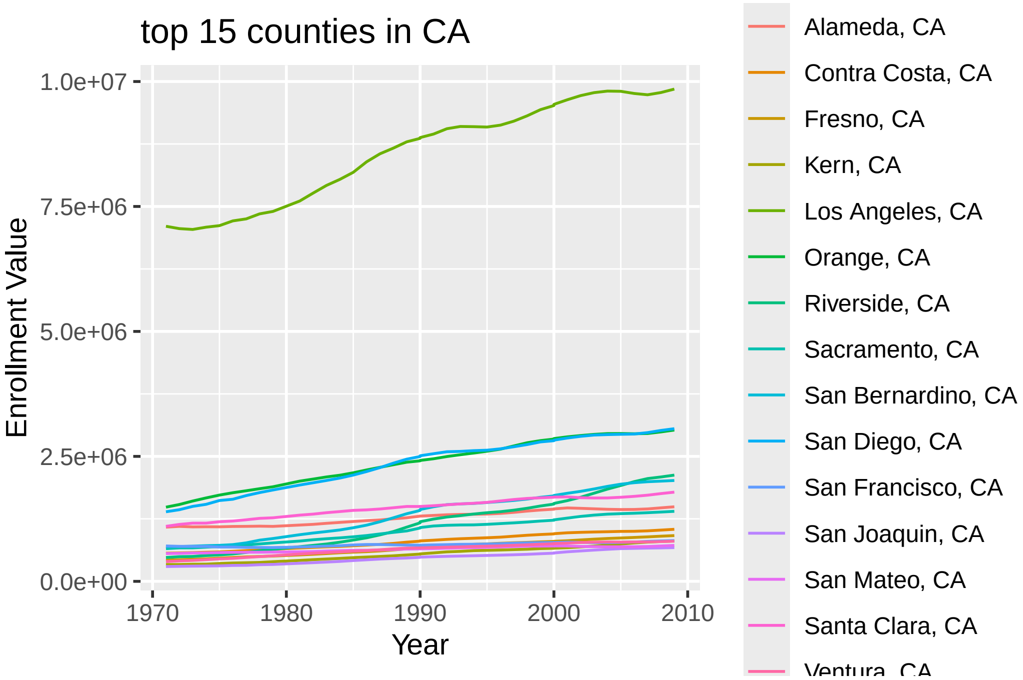
Question 7: State Plot for PST Data

```
plot(pst_combined$noncounty, var_name = "Enrollment Value")
```

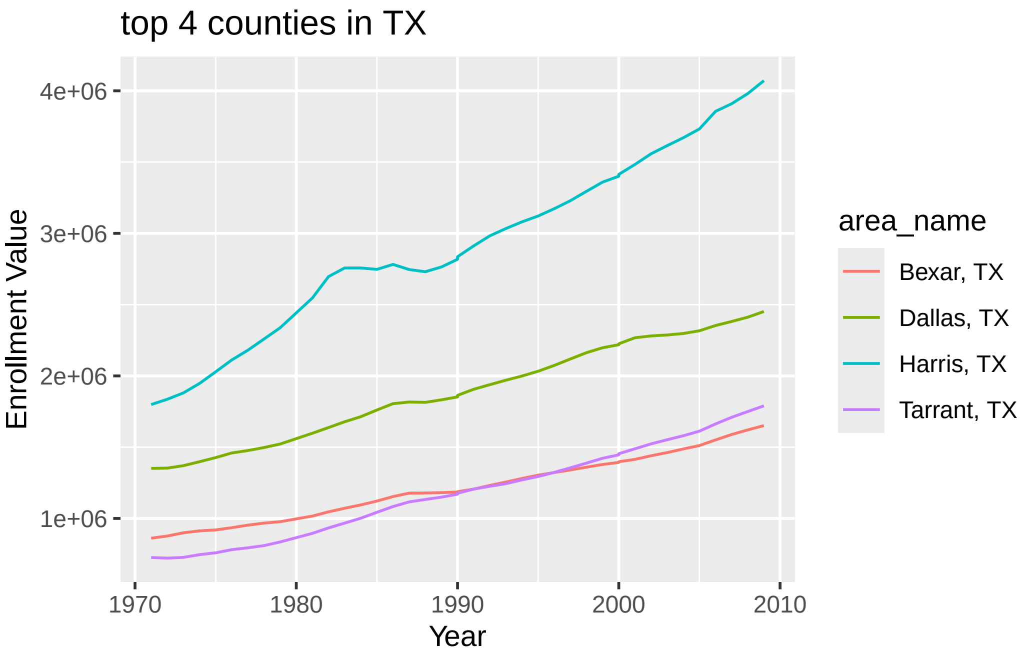


Question 8: County Plots for PST Data

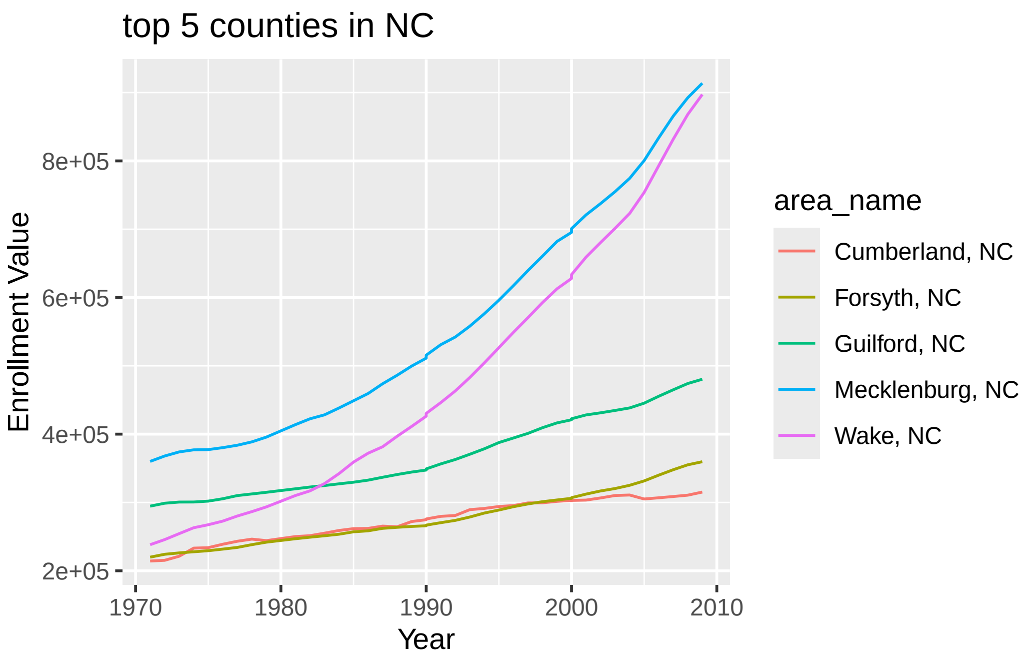
```
# CA, top 15
plot(pst_combined$county, var_name = "Enrollment Value", state = "CA", top_or_bottom = "top"
```



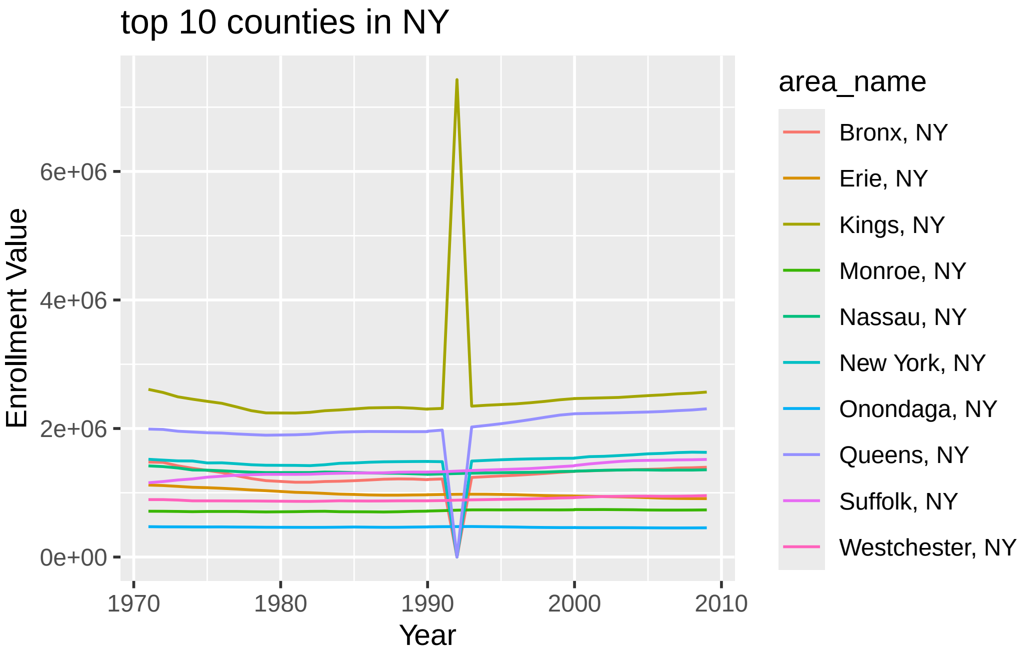
```
# TX, top 4
plot(pst_combined$county, var_name = "Enrollment Value", state = "TX", top_or_bottom = "top"
```



```
# Default
plot(pst_combined$county, var_name = "Enrollment Value")
```



```
# NY, top 10
plot(pst_combined$county, var_name = "Enrollment Value", state = "NY", top_or_bottom = "top"
```


top 10 counties in NY
Enrollment Value
6e+06
4e+06
2e+06
0e+00
1970
1980
1990
2000
2010
Year
area_name
Bronx, NY
Erie, NY
Kings, NY
Monroe, NY
Nassau, NY
New York, NY
Onondaga, NY
Queens, NY
Suffolk, NY
Westchester, NY