Project 1

Keshav Ramesh & Calista Harris

Load require packages

```
library(tidyverse)
```

Data Processing

Question 1: Reading in Data

```
edu01a <- read_csv("EDU01a.csv", show_col_types = FALSE) |>
  select(
    area_name = Area_name, #rename Area_name
    STCOU,
    ends_with("D") #select all columns ending in "D"
  )

#display the first 5 rows
edu01a |>
  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: Pivot Data

```
edu_long = edu01a %>%
  pivot_longer(
    cols = ends_with("D"),
    names_to = "surveyID_full", #store original column names (ex. "EST1234D")
    values_to = "enrollment"
  )

#display the first 5 rows
head(edu_long, 5)
```

```
# A tibble: 5 x 4
  area_name     STCOU surveyID_full 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
```

Question 3: Extracting the year

```
EDU01_long_updated = edu_long %>%
  mutate(
    surveyID_year = substr(surveyID_full, 8, 9)
    ) %>%
  mutate(
    year = as.numeric(paste0("19", surveyID_year))
  ) %>%
  mutate(
    surveyID = substr(surveyID_full, 1, 7)
  ) %>%
  select(-surveyID_year)



head(EDU01_long_updated, 5)
```

```
# A tibble: 5 x 6
  area_name     STCOU surveyID_full enrollment  year surveyID
  <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
```

Question 4: Identifying County Data

```
county_locations = grep(pattern = ", \\w\\w", EDU01_long_updated$area_name)

EDU01_long_updated_county = EDU01_long_updated[county_locations, ]

class(EDU01_long_updated_county) <- c("County", class(EDU01_long_updated_county))

EDU01_long_updated_non_county = EDU01_long_updated[-county_locations, ]

class(EDU01_long_updated_non_county) <- c("State", class(EDU01_long_updated_non_county))


head(EDU01_long_updated_county, 10)
```

```
# A tibble: 10 x 6
   area_name   STCOU surveyID_full enrollment  year surveyID
   <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
```

```
head(EDU01_long_updated_non_county, 10)
```

```
# A tibble: 10 x 6
   area_name     STCOU surveyID_full enrollment  year surveyID
   <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
```