Project 1

Tamdan L, Alise M

Question 1 Reading in and selecting data

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.4.3
```

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```
Warning: package 'tibble' was built under R version 4.4.3
```

```
Warning: package 'tidyr' was built under R version 4.4.3
```

```
Warning: package 'readr' was built under R version 4.4.3
```

```
Warning: package 'purrr' was built under R version 4.4.3
```

```
Warning: package 'dplyr' was built under R version 4.4.3
```

```
Warning: package 'stringr' was built under R version 4.4.3
```

```
Warning: package 'forcats' was built under R version 4.4.3
```

```
Warning: package 'lubridate' was built under R version 4.4.3
```

```
-- 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     1.0.4
-- 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 becor
```

```
library(readr)
EDU01A <-read_csv("data/EDU01A.csv")
```

```
Rows: 3198 Columns: 42
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (22): Area_name, STCOU, EDU010187N1, EDU010187N2, EDU010188N1, EDU010188...
dbl (20): EDU010187F, EDU010187D, EDU010188F, EDU010188D, EDU010189F, EDU010...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
EDU01A |>
  select(Area_name, STCOU, ends_with("D")) |>
  rename(area_name = Area_name) |>
  head(EDU01A, n=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 Convert to long format

```
EDU01A_long<-
  EDU01A |>
  select(Area_name, STCOU, ends_with("D")) |>
  rename(area_name = Area_name) |>
  pivot_longer(cols= ends_with("D"),
               names_to = "EDU_combined",
               values_to = "enrollment_value"
               )
  head(EDU01A_long, n=5)
```

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

Question 3 Parsing

```
EDU01A_longer <-
  EDU01A_long |>
  mutate(two_digit_year= (substr(EDU_combined, start=8, stop = 9)) ,
  year_dbl = as.double(two_digit_year),
  year= if_else(year_dbl >25, 1900 + year_dbl, year_dbl <=25 + 2000),
  survey_value = substr(EDU_combined, start=1, stop = 7)
         )
EDU01A_longer
```

```
# A tibble: 31,980 x 8
   area_name   STCOU EDU_combined enrollment_value two_digit_year year_dbl  year
   <chr>       <chr> <chr>                   <dbl> <chr>             <dbl> <dbl>
 1 UNITED STA~ 00000 EDU010187D           40024299 87                   87  1987
 2 UNITED STA~ 00000 EDU010188D           39967624 88                   88  1988
 3 UNITED STA~ 00000 EDU010189D           40317775 89                   89  1989
 4 UNITED STA~ 00000 EDU010190D           40737600 90                   90  1990
```

```
 5 UNITED STA~ 00000 EDU010191D           41385442 91                    91  1991
 6 UNITED STA~ 00000 EDU010192D           42088151 92                    92  1992
 7 UNITED STA~ 00000 EDU010193D           42724710 93                    93  1993
 8 UNITED STA~ 00000 EDU010194D           43369917 94                    94  1994
 9 UNITED STA~ 00000 EDU010195D           43993459 95                    95  1995
10 UNITED STA~ 00000 EDU010196D           44715737 96                    96  1996
# i 31,970 more rows
# i 1 more variable: survey_value <chr>
```

```
long_updated <- select(EDU01A_longer, area_name,STCOU, enrollment_value,year, survey_value)

head(long_updated, n=5)
```

```
# A tibble: 5 x 5
  area_name     STCOU enrollment_value  year survey_value
  <chr>         <chr>            <dbl> <dbl> <chr>
1 UNITED STATES 00000         40024299  1987 EDU0101
2 UNITED STATES 00000         39967624  1988 EDU0101
3 UNITED STATES 00000         40317775  1989 EDU0101
4 UNITED STATES 00000         40737600  1990 EDU0101
5 UNITED STATES 00000         41385442  1991 EDU0101
```

Question 4 Two Tibbles

```
County_indices <- grep(pattern = ", [A-Z]{2}", long_updated$area_name)
noncounty_tibble <- long_updated [-County_indices, ]
county_tibble <- long_updated [County_indices, ]

class(county_tibble) <- c("county", class(county_tibble))
class(noncounty_tibble) <- c("state", class(noncounty_tibble))

head(county_tibble, n=10)
```

```
# A tibble: 10 x 5
   area_name   STCOU enrollment_value  year survey_value
   <chr>       <chr>            <dbl> <dbl> <chr>
 1 Autauga, AL 01001             6829  1987 EDU0101
 2 Autauga, AL 01001             6900  1988 EDU0101
 3 Autauga, AL 01001             6920  1989 EDU0101
```

```
 4 Autauga, AL 01001             6847  1990 EDU0101
 5 Autauga, AL 01001             7008  1991 EDU0101
 6 Autauga, AL 01001             7137  1992 EDU0101
 7 Autauga, AL 01001             7152  1993 EDU0101
 8 Autauga, AL 01001             7381  1994 EDU0101
 9 Autauga, AL 01001             7568  1995 EDU0101
10 Autauga, AL 01001             7834  1996 EDU0101
```

```
head(noncounty_tibble, n=10)
```

```
# A tibble: 10 x 5
   area_name     STCOU enrollment_value  year survey_value
   <chr>         <chr>            <dbl> <dbl> <chr>
 1 UNITED STATES 00000         40024299  1987 EDU0101
 2 UNITED STATES 00000         39967624  1988 EDU0101
 3 UNITED STATES 00000         40317775  1989 EDU0101
 4 UNITED STATES 00000         40737600  1990 EDU0101
 5 UNITED STATES 00000         41385442  1991 EDU0101
 6 UNITED STATES 00000         42088151  1992 EDU0101
 7 UNITED STATES 00000         42724710  1993 EDU0101
 8 UNITED STATES 00000         43369917  1994 EDU0101
 9 UNITED STATES 00000         43993459  1995 EDU0101
10 UNITED STATES 00000         44715737  1996 EDU0101
```

Question 5 County level new variable

```
county_tibble |>
mutate(state = substr(area_name, nchar(area_name) - 1, nchar(area_name))
)
```

```
# A tibble: 31,450 x 6
   area_name   STCOU enrollment_value  year survey_value state
   <chr>       <chr>            <dbl> <dbl> <chr>        <chr>
 1 Autauga, AL 01001             6829  1987 EDU0101      AL
 2 Autauga, AL 01001             6900  1988 EDU0101      AL
 3 Autauga, AL 01001             6920  1989 EDU0101      AL
 4 Autauga, AL 01001             6847  1990 EDU0101      AL
 5 Autauga, AL 01001             7008  1991 EDU0101      AL
 6 Autauga, AL 01001             7137  1992 EDU0101      AL
 7 Autauga, AL 01001             7152  1993 EDU0101      AL
```

```
 8 Autauga, AL 01001             7381  1994 EDU0101      AL
 9 Autauga, AL 01001             7568  1995 EDU0101      AL
10 Autauga, AL 01001             7834  1996 EDU0101      AL
# i 31,440 more rows
```

Question 6 Non-county "division"

```
noncounty_tibble |>
  mutate(division = if_else (area_name %in% state.name,
  state.division [match(area_name, state.name)], "ERROR")
  )
```

```
# A tibble: 530 x 6
   area_name     STCOU enrollment_value  year survey_value division
   <chr>         <chr>            <dbl> <dbl> <chr>        <chr>
 1 UNITED STATES 00000         40024299  1987 EDU0101      ERROR
 2 UNITED STATES 00000         39967624  1988 EDU0101      ERROR
 3 UNITED STATES 00000         40317775  1989 EDU0101      ERROR
 4 UNITED STATES 00000         40737600  1990 EDU0101      ERROR
 5 UNITED STATES 00000         41385442  1991 EDU0101      ERROR
 6 UNITED STATES 00000         42088151  1992 EDU0101      ERROR
 7 UNITED STATES 00000         42724710  1993 EDU0101      ERROR
 8 UNITED STATES 00000         43369917  1994 EDU0101      ERROR
 9 UNITED STATES 00000         43993459  1995 EDU0101      ERROR
10 UNITED STATES 00000         44715737  1996 EDU0101      ERROR
# i 520 more rows
```