Final Project Code

Aden Bhagwat, Emily Mingus, Erick Njue

```
library(here)
```

```
here() starts at C:/Users/adenb/OneDrive/Desktop/Git/EDLD_652_Final
```

```
library(edld652)
library(rio)
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.4.2
```

```
-- 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 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
```

```
-- 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 becom
```

```
library(janitor)
```

```
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```
library(knitr)
library(modelsummary)
```

```
Warning: package 'modelsummary' was built under R version 4.4.2
```

```
`modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing
  backend. Learn more at: https://vincentarelbundock.github.io/tinytable/

Revert to `kableExtra` for one session:

  options(modelsummary_factory_default = 'kableExtra')
  options(modelsummary_factory_latex = 'kableExtra')
  options(modelsummary_factory_html = 'kableExtra')

Silence this message forever:

  config_modelsummary(startup_message = FALSE)
```

```
options(modelsummary_factory_default =
          'kableExtra')
options(modelsummary_factory_latex =
          'kableExtra')
options(modelsummary_factory_html =
          'kableExtra')
```

```
get_documentation("EDFacts_rla_achievement_lea_2010_2019")
```

```
https://www2.ed.gov/about/inits/ed/edfacts/data-files/assessments-sy2018-19-public-file-docu
```

```
[1] "https://www2.ed.gov/about/inits/ed/edfacts/data-files/assessments-sy2018-19-public-file
```

```
# get_documentation("EDFacts_math_achievement_lea_2010_2019")

###same documentation for both

# rla_achieve <- get_data("EDFacts_rla_achievement_lea_2010_2019") %>%
```

```
#   clean_names()
#
# rla_sub <- rla_achieve %>%
#   select(leaid, matches("^(ecd|all).*pctprof$"))
#
# export(rla_sub, "Data/rla_sub.Rdata")
#
# math_achieve <- get_data("EDFacts_math_achievement_lea_2010_2019") %>%
#   clean_names()
#
# math_sub <- math_achieve %>%
#   select(leaid, matches("^(ecd|all).*pctprof$"))
#
# export(math_sub, "Data/math_sub.Rdata")

rla_sub<-import(here("Data/rla_sub.Rdata"))
```

```
Warning: Missing `trust` will be set to FALSE by default for RData in 2.0.0.
```

```
math_sub<-import(here("Data/math_sub.Rdata"))
```

```
Warning: Missing `trust` will be set to FALSE by default for RData in 2.0.0.
```

```
glimpse(rla_sub)
```

```
Rows: 15,717
Columns: 17
$ leaid            <chr> "0200001", "0200003", "0200004", "0200005", "0200006"~
$ all_rla00pctprof <chr> "46", "42", "26", "40-44", "75-79", "75-79", "60-79",~
$ all_rla03pctprof <chr> "34", "40-44", "30-39", "40-59", NA, "GE50", "PS", "6~
$ all_rla04pctprof <chr> "39", "40-44", "20-29", "21-39", NA, "GE80", "PS", "6~
$ all_rla05pctprof <chr> "50-54", "40-44", "40-59", "21-39", NA, "60-79", NA, ~
$ all_rla06pctprof <chr> "35-39", "30-34", "11-19", "40-59", NA, "GE50", "PS",~
$ all_rla07pctprof <chr> "45-49", "45-49", "20-29", "GE50", NA, "60-79", "PS",~
$ all_rla08pctprof <chr> "60-64", "45-49", "30-39", "60-79", NA, "60-79", "PS"~
$ all_rlahspctprof <chr> "54", "35-39", "11-19", "30-39", "75-79", "80-89", "P~
$ ecd_rla00pctprof <chr> "41", "35", "25-29", "40-44", "70-79", "70-74", "60-7~
$ ecd_rla03pctprof <chr> "29", "30-34", "30-39", "40-59", NA, "GE50", "PS", "6~
$ ecd_rla04pctprof <chr> "30-34", "35-39", "20-29", "21-39", NA, "60-79", "PS"~
$ ecd_rla05pctprof <chr> "45-49", "40-44", "40-59", "21-39", NA, "GE50", NA, "~
```

```
$ ecd_rla06pctprof <chr> "30-34", "25-29", "11-19", "40-59", NA, "GE50", "PS",~
$ ecd_rla07pctprof <chr> "45-49", "35-39", "20-29", "GE50", NA, "GE50", "PS", ~
$ ecd_rla08pctprof <chr> "55-59", "40-44", "30-39", "60-79", NA, "60-79", "PS"~
$ ecd_rlahspctprof <chr> "49", "30-34", "11-19", "30-39", "70-79", "60-79", "P~
```

```
glimpse(math_sub)
```

```
Rows: 15,747
Columns: 17
$ leaid            <chr> "0200001", "0200003", "0200004", "0200005", "0200006"~
$ all_mth00pctprof <chr> "37", "36", "20", "45-49", "60-64", "65-69", "60-79",~
$ all_mth03pctprof <chr> "31", "40-44", "20-29", "40-59", NA, "GE50", "PS", "5~
$ all_mth04pctprof <chr> "38", "45-49", "11-19", "40-59", NA, "GE80", "PS", "5~
$ all_mth05pctprof <chr> "40-44", "45-49", "21-39", "60-79", NA, "GE80", NA, "~
$ all_mth06pctprof <chr> "40-44", "40-44", "20-29", "40-59", NA, "GE50", "PS",~
$ all_mth07pctprof <chr> "30-34", "30-34", "11-19", "GE50", NA, "GE80", "PS", ~
$ all_mth08pctprof <chr> "40-44", "25-29", "20-29", "40-59", NA, "60-79", "PS"~
$ all_mthhspctprof <chr> "35", "20-24", "11-19", "20-29", "60-64", "50-59", "P~
$ ecd_mth00pctprof <chr> "32", "31", "20-24", "45-49", "50-59", "65-69", "60-7~
$ ecd_mth03pctprof <chr> "26", "35-39", "20-29", "40-59", NA, "GE50", "PS", "5~
$ ecd_mth04pctprof <chr> "35-39", "40-44", "11-19", "40-59", NA, "GE80", "PS",~
$ ecd_mth05pctprof <chr> "35-39", "40-44", "21-39", "60-79", NA, "GE50", NA, "~
$ ecd_mth06pctprof <chr> "35-39", "30-34", "20-29", "40-59", NA, "GE50", "PS",~
$ ecd_mth07pctprof <chr> "30-34", "25-29", "11-19", "GE50", NA, "GE50", "PS", ~
$ ecd_mth08pctprof <chr> "30-34", "15-19", "20-29", "40-59", NA, "60-79", "PS"~
$ ecd_mthhspctprof <chr> "30", "15-19", "11-19", "20-29", "50-59", "40-59", "P~
```

```
str(rla_sub)
```

```
'data.frame':   15717 obs. of  17 variables:
 $ leaid           : chr  "0200001" "0200003" "0200004" "0200005" ...
 $ all_rla00pctprof: chr  "46" "42" "26" "40-44" ...
 $ all_rla03pctprof: chr  "34" "40-44" "30-39" "40-59" ...
 $ all_rla04pctprof: chr  "39" "40-44" "20-29" "21-39" ...
 $ all_rla05pctprof: chr  "50-54" "40-44" "40-59" "21-39" ...
 $ all_rla06pctprof: chr  "35-39" "30-34" "11-19" "40-59" ...
 $ all_rla07pctprof: chr  "45-49" "45-49" "20-29" "GE50" ...
 $ all_rla08pctprof: chr  "60-64" "45-49" "30-39" "60-79" ...
 $ all_rlahspctprof: chr  "54" "35-39" "11-19" "30-39" ...
 $ ecd_rla00pctprof: chr  "41" "35" "25-29" "40-44" ...
 $ ecd_rla03pctprof: chr  "29" "30-34" "30-39" "40-59" ...
```

```
 $ ecd_rla04pctprof: chr  "30-34" "35-39" "20-29" "21-39" ...
 $ ecd_rla05pctprof: chr  "45-49" "40-44" "40-59" "21-39" ...
 $ ecd_rla06pctprof: chr  "30-34" "25-29" "11-19" "40-59" ...
 $ ecd_rla07pctprof: chr  "45-49" "35-39" "20-29" "GE50" ...
 $ ecd_rla08pctprof: chr  "55-59" "40-44" "30-39" "60-79" ...
 $ ecd_rlahspctprof: chr  "49" "30-34" "11-19" "30-39" ...
```

```r
table_rla <- as.data.frame(table(rla_sub$all_rla00pctprof))
colnames(table_rla) <- c("Percent of Students Proficient", "Count")
print(table_rla)
```

```
    Percent of Students Proficient Count
1                                10     1
2                              10-14    11
3                                11     1
4                              11-19    16
5                                12     1
6                                13     2
7                                14     1
8                                15     3
9                              15-19    22
10                               17     2
11                               18     2
12                               19     2
13                               20     5
14                             20-24    29
15                             20-29    28
16                               21     6
17                             21-39    50
18                               22     2
19                               23     2
20                               24     6
21                               25     7
22                             25-29    37
23                               26    12
24                               27     8
25                               28    10
26                               29    13
27                               30    20
28                             30-34    51
29                             30-39    65
30                               31    15
```

31	32	17
32	33	20
33	34	16
34	35	30
35	35-39	65
36	36	32
37	37	30
38	38	26
39	39	32
40	40	34
41	40-44	100
42	40-49	86
43	40-59	90
44	41	46
45	42	48
46	43	70
47	44	61
48	45	66
49	45-49	132
50	46	62
51	47	78
52	48	76
53	49	72
54	5	1
55	50	76
56	50-54	166
57	50-59	105
58	51	73
59	52	78
60	53	83
61	54	95
62	55	102
63	55-59	198
64	56	84
65	57	96
66	58	103
67	59	110
68	6	1
69	6-9	3
70	60	112
71	60-64	237
72	60-69	132
73	60-79	132

74	61	120
75	62	115
76	63	150
77	64	143
78	65	177
79	65-69	310
80	66	158
81	67	165
82	68	194
83	69	173
84	7	1
85	70	218
86	70-74	352
87	70-79	156
88	71	236
89	72	263
90	73	291
91	74	283
92	75	289
93	75-79	362
94	76	283
95	77	312
96	78	304
97	79	286
98	80	313
99	80-84	327
100	80-89	133
101	81	330
102	82	305
103	83	342
104	84	316
105	85	324
106	85-89	337
107	86	353
108	87	326
109	88	357
110	89	290
111	9	1
112	90	283
113	90-94	296
114	91	280
115	92	221
116	93	205

```
117                              94   192
118                              95   106
119                              96    84
120                              97    53
121                              98    32
122                            GE50   193
123                            GE80   125
124                            GE90   100
125                            GE95   162
126                            GE99    20
127                            LE10    10
128                            LE20    22
129                             LE5     3
130                            LT50    78
131                             n/a     9
132                              PS   142
```

```
table_math <- as.data.frame(table(math_sub$all_mth00pctprof))
colnames(table_math) <- c("Percent of Students Proficient", "Count")
print(table_math)
```

```
    Percent of Students Proficient Count
1                                10     5
2                              10-14    33
3                                 11     2
4                              11-19    56
5                                 13     2
6                                 14     2
7                                 15     4
8                              15-19    60
9                                 16     2
10                                17     2
11                                18     6
12                                19     5
13                                20     9
14                             20-24    45
15                             20-29    63
16                                21    11
17                             21-39    69
18                                22    10
19                                23     7
20                                24     7
```

21	25	18
22	25-29	55
23	26	15
24	27	12
25	28	21
26	29	22
27	30	28
28	30-34	85
29	30-39	72
30	31	24
31	32	26
32	33	23
33	34	33
34	35	27
35	35-39	104
36	36	34
37	37	45
38	38	43
39	39	50
40	4	1
41	40	57
42	40-44	128
43	40-49	98
44	40-59	99
45	41	48
46	42	64
47	43	61
48	44	63
49	45	80
50	45-49	169
51	46	77
52	47	100
53	48	82
54	49	90
55	5	2
56	50	94
57	50-54	203
58	50-59	109
59	51	96
60	52	111
61	53	102
62	54	97
63	55	108

64	55-59	207
65	56	120
66	57	99
67	58	125
68	59	112
69	6	3
70	6-9	13
71	60	114
72	60-64	216
73	60-69	113
74	60-79	91
75	61	137
76	62	134
77	63	156
78	64	148
79	65	155
80	65-69	229
81	66	156
82	67	173
83	68	156
84	69	165
85	7	2
86	70	211
87	70-74	260
88	70-79	106
89	71	194
90	72	223
91	73	231
92	74	238
93	75	266
94	75-79	325
95	76	273
96	77	305
97	78	291
98	79	321
99	8	1
100	80	321
101	80-84	310
102	80-89	101
103	81	335
104	82	335
105	83	326
106	84	371

107	85	302
108	85-89	304
109	86	307
110	87	321
111	88	280
112	89	261
113	9	3
114	90	245
115	90-94	249
116	91	219
117	92	177
118	93	189
119	94	151
120	95	127
121	96	99
122	97	90
123	98	40
124	GE50	152
125	GE80	77
126	GE90	81
127	GE95	170
128	GE99	19
129	LE10	45
130	LE20	79
131	LE5	18
132	LT50	119
133	n/a	28
134	PS	151

There are a range of ways data is measured, from exact percentages, to ranges, to greater than/less than statements. We need to come up with rules to standardize and drop n/a and PS values. I think maybe we could just take the lower number in the range, and print whatever number is reported in greater than/less than statements. We would need to add a disclaimer.

```
rla_clean <- rla_sub %>%
  mutate(across(
    -leaid,
    ~ gsub("-.*", "", .))) %>%
  mutate(across(
    -leaid,
    ~ gsub("^[A-Za-z]{2}([0-9]{1}).*", "\\1", .))) %>%
  mutate(across(
```

```
    -leaid,
    ~ gsub("^[A-Za-z]{1}([0-9]{2}).*", "\\1", .))) %>%
  mutate(across(
    -leaid,
    ~ na_if(., "PS"))) %>%
  mutate(across(
    -leaid,
    ~ na_if(., "n/a"))) %>%
  mutate(across(
    everything(),
    as.numeric))

table(rla_clean$all_rla00pctprof, useNA = "ifany")
```

```
   1    2    5    6    7    8    9   10   11   12   13   14   15   17   18   19
  10   22  275    4    1  125  283   12   17    1    2    1   25    2    2    2
  20   21   22   23   24   25   26   27   28   29   30   31   32   33   34   35
  62   56    2    2    6   44   12    8   10   13  136   15   17   20   16   95
  36   37   38   39   40   41   42   43   44   45   46   47   48   49   50   51
  32   30   26   32  310   46   48   70   61  198   62   78   76   72  347   73
  52   53   54   55   56   57   58   59   60   61   62   63   64   65   66   67
  78   83   95  300   84   96  103  110  613  120  115  150  143  487  158  165
  68   69   70   71   72   73   74   75   76   77   78   79   80   81   82   83
 194  173  726  236  263  291  283  651  283  312  304  286  773  330  305  342
  84   85   86   87   88   89   90   91   92   93   94   95   96   97   98 <NA>
 316  661  353  326  357  290  579  280  221  205  192  106   84   53   32  151
```

```
mean(rla_clean$all_rla00pctprof, na.rm=T)
```

```
[1] 68.27984
```

```
math_clean <- math_sub %>%
  mutate(across(
    -leaid,
    ~ gsub("-.*", "", .))) %>%
  mutate(across(
    -leaid,
    ~ gsub("^[A-Za-z]{2}([0-9]{1}).*", "\\1", .))) %>%
  mutate(across(
```

```
    -leaid,
    ~ gsub("^[A-Za-z]{1}([0-9]{2}).*", "\\1", .))) %>%
  mutate(across(
    -leaid,
    ~ na_if(., "PS"))) %>%
  mutate(across(
    -leaid,
    ~ na_if(., "n/a"))) %>%
  mutate(across(
    everything(),
    as.numeric))

table(math_clean$all_mth00pctprof, useNA = "ifany")
```

```
   1    2    4    5    6    7    8    9   10   11   13   14   15   16   17   18
  45   79    1  291   16    2   78  273   38   58    2    2   64    2    2    6
  19   20   21   22   23   24   25   26   27   28   29   30   31   32   33   34
   5  117   80   10    7    7   73   15   12   21   22  185   24   26   23   33
  35   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50
 131   34   45   43   50  382   48   64   61   63  249   77  100   82   90  406
  51   52   53   54   55   56   57   58   59   60   61   62   63   64   65   66
  96  111  102   97  315  120   99  125  112  534  137  134  156  148  384  156
  67   68   69   70   71   72   73   74   75   76   77   78   79   80   81   82
 173  156  165  577  194  223  231  238  591  273  305  291  321  732  335  335
  83   84   85   86   87   88   89   90   91   92   93   94   95   96   97   98
 326  371  606  307  321  280  261  494  219  177  189  151  127   99   90   40
<NA>
 179
```

```
mean(math_clean$all_mth00pctprof, na.rm=T)
```

```
[1] 65.75976
```

Note: May need to pivot data longer

Below are our research questions:

Research Question #1: Does Socioeconomic Status (SES) affect educational proficiency sccores?

If SES affects proficiency scores, then:

#1a: Does Socioeconomic Status (SES) affect Math proficiency scores differently by grade level?

#1b: Does Socioeconomic Status (SES) affect Reading proficiency scores differently by grade level?

Visualization Ideas

Here are also some ideas related to what visualizations we want to create. This is district level data. We are interested in investigating/visualizing grade level data to show proficiency scores across general population compared to low SES status. The audiences will vary and we will want to customize the visuals based on who the intended target is.

The scenario may be we are informing district administrators of disparities across schools/grade-level as it relates to proficiency scores. This information would be useful in determining what schools/grade levels may benefit from additional supports to increase proficiency. Based on our research questions above, here are some ideas generated during the team meeting.

For Question 1: Does SES affect proficiency scores?

- Boxplot: Compare Math and Reading proficiency scores across SES groups
- Bar Plot with Error Bars: Show mean proficiency scores (Math and Reading) by SES with error bar
- Violin Plot: Visualize the distribution of proficiency scores by SES

For Question 1a: Does SES affect Math proficiency scores by grade level?

- Line Plot: Show Math proficiency trends by grade level for each SES group
- Faceted Bar Plot: Display average Math scores by grade level and SES
- Heatmap: Visualize Math proficiency across SES and grade levels

For Question 1b: Does SES affect Reading proficiency scores?

- Side-by-Side Boxplots: Compare Reading proficiency across SES groups for each grade level
- Stacked Bar Chart: Show the distribution of proficiency levels (Math/Reading) by SES
- Grouped Bar Plot: Compare average Reading scores by SES and grade level

General Visualizations:

- Scatter Plot: Explore SES vs proficiency scores (Math/Reading)
- Density Plot: Compare distribution of proficiency scores by SES

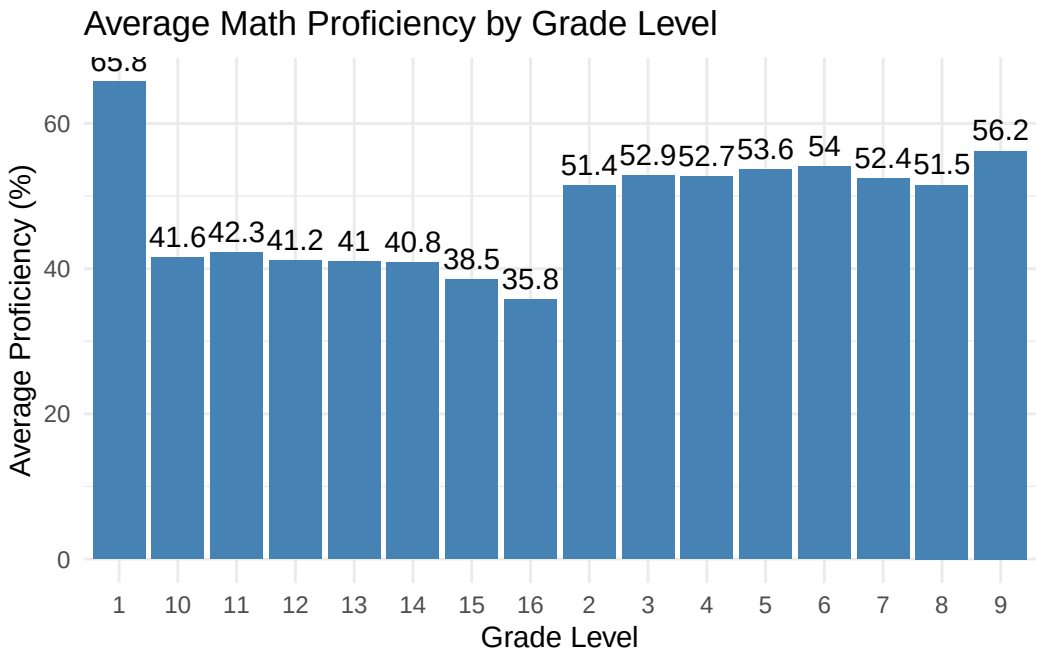
```
#histograms

#Excludes the 'leaid' column from the data and calculate averages
column_averages <- colMeans(math_clean[, -1], na.rm = TRUE)

#Creates the 'Grade_Level' vector with the correct length (16 grades)
Grade_Level <- c( "1",  "2",  "3",  "4",  "5",
                  "6",  "7",  "8", "9", "10",
                 "11", "12", "13", "14", "15", "16")

#Ensures the length of Grade_Level matches column_averages
averages_df <- data.frame(
  Grade_Level = Grade_Level,
  Average_Proficiency = column_averages
)

ggplot(averages_df, aes(x = Grade_Level, y = Average_Proficiency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(
    title = "Average Math Proficiency by Grade Level",
    x = "Grade Level",
    y = "Average Proficiency (%)"
  ) +
  geom_text(aes(label = round(Average_Proficiency, 1)), vjust = -0.5)
```




```
#second run the x-axis does not look as good, will need fixed
```