Homework 2

Analysis

Your Name Here

Background

Gender achievement gaps in education have been well-documented over the years – studies consistently find boys outperforming girls on math tests and girls outperforming boys on reading and language tests. A particularly controversial article was published in Science in 1980 arguing that this pattern was due to an 'innate' difference in ability (focusing on mathematics rather than on reading and language). Such views persisted in part because studying systematic patterns in achievement nationwide was a challenge due to differential testing standards across school districts and the general lack of availability of large-scale data.

It is only recently that data-driven research has begun to reveal socioeconomic drivers of achievement gaps. The Standford Educational Data Archive (SEDA), a publicly available database on academic achievement and educational opportunity in U.S. schools, has supported this effort. The database is part of a broader initiave aiming to improve educational opportunity by enabling researchers and policymakers to identify systemic drivers of disparity.

> SEDA includes a range of detailed data on educational conditions, contexts, and outcomes in school districts and counties across the United States. It includes measures of academic achievement and achievement gaps for school districts and counties, as well as district-level measures of racial and socioeconomic composition, racial and socioeconomic segregation patterns, and other features of the schooling system.

The database standardizes average test scores for schools 10,000 U.S. school districts relative to national standards to allow comparability between school districts and across grade levels and years. The test score data come from the U.S. Department of Education. In addition, multiple data sources (American Community Survey and Common Core of Data) are integrated to provide district-level socioeconomic and demographic information.

A study of the SEDA data published in 2018 identified the following persistent patterns across grade levels 3 - 8 and school years from 2008 through 2015:

- a consistent reading and language achievement gap favoring girls;
- *no* national math achievement gap on average; and
- local math achievement gaps that depend on the socioeconomic conditions of school districts. You can read about the main findings of the study in this brief NY Times article.

Below, we'll work with selected portions of the database. The full datasets can be downloaded here.

Assignment objectives

In this assignment, you'll explore achievement gaps in California school districts in 2018, reproducing the findings described in the article above on a more local scale and with the most recent SEDA data. You'll practice the following:

- review of data documentation
- assessment of sampling design and scope of inference
- data tidying operations
  - slicing and filtering
  - merging multiple data frames
  - pivoting tables
  - renaming and reordering variables
- constructing exploratory graphics and visualizing trends
- data aggregations
- narrative summary of exploratory analysis

Import and assessment of datasets

You'll work with test data and socioeconomic covariates aggregated to the school district level. These data are stored in two separate tables. Here you'll examine them and review data documentation.

Test score data

The first few rows of the test data are shown below. The columns are:

Column name	Meaning
sedalea	District ID
grade	Grade level
stateabb	State abbreviation
sedaleaname	District name
subject	Test subject
cs_mn_...	Estimated mean test score
cs_mnse_...	Standard error for estimated mean test score
totgyb_...	Number of individual tests used to estimate the mean score

```
# import seda data
ca_main <- read_csv('data/ca-main.csv')
ca_cov <- read_csv('data/ca-cov.csv')

# preview test score data
head(ca_main, n=5)
```

The test score means for each district are named `cs_mn_...` with an abbreviation indicating subgroup (such as mean score for all `cs_mean_all`, for boys `cs_mean_mal`, for white students `cs_mn_wht`, and so on). Notice that these are generally small-ish: decimal numbers between -0.5 and 0.5.

These means are *estimated* from a number of individual student tests and *standardized* relative to national averages. They represent the number of standard deviations by which a district mean differs from the national average. So, for instance, the value `cs_mn_all = 0.1` indicates that the district average is estimated to be 0.1 standard deviations greater than the national average on the corresponding test and at the corresponding grade level.

Question 1: Interpreting test score values

Interpret the average math test score for all 4th grade students in Acton-Agua Dulce Unified School District (the first row of the dataset shown above). #Answer: The means that the estimated average math test score for 4th graders in this district is 0.367 standard deviations below the national average.

Covariate data

The first few rows of the covariate data are shown below. The column information is as follows:

Column name	Meaning
`sedalea`	District ID
`grade`	Grade level
`sedaleanm`	District name
`urban`	Indicator: is the district in an urban locale?
`suburb`	Indicator: is the district in a suburban locale?
`town`	Indicator: is the district in a town locale?
`rural`	Indicator: is the district in a rural locale?
`locale`	Description of district locale
Remaining variables	Demographic and socioeconomic measures

```
head(ca_cov, n=5)
```

You will only be working with a handful of the demographic and socioeconomic measures, so you can put off getting acquainted with those until selecting a subset of variables.

Question 2: Data semantics

In the non-public data, observational units are students – test scores are measured for each student. However, in the SEDA data you've imported, scores are *aggregated* to the district level by grade. Let's regard estimated test score means for each grade as distinct variables, so that an observation consists of a set of estimated means for different grade levels and groups. In this view, what are the observational units in the test score dataset? Are they the same or different for the covariate dataset?

#Answer: The observational units are district-grade-subject combinations. Each row represents an estimated test score mean for a specific school district, grade level, and subject. For the covariate dataset, the observational units appear to be district-grade combinations. The test score dataset is further broken down by subject, whereas the covariate dataset does not have this additional level of differentiation.

Question 3: Sample sizes

How many observational units are in each dataset? Count the number of units in the test dataset and the number of units in the covariate dataset separately. Store the values as `ca_cov_units` and `ca_main_units`, respectively.

(*Hint*: use `unique()`.)

```
ca_cov_units <- nrow(ca_cov)
ca_main_units <- nrow(ca_main)

print(paste('units in covariate data: ', ca_cov_units))
print(paste('units in test score data: ', ca_main_units))
```

Question 4: Sample characteristics and scope of inference

Answer the questions below about the sampling design in a short paragraph. You do not need to dig through any data documentation in order to resolve these questions.

- (i) What is the relevant population for the datasets you've imported? #Answer: The relevant population for the test score dataset(SEDA data) and covariate dataset consists of students enrolled in public school districts in California. Since the dataset is aggregated at the district and grade level, the population being studied includes students acroos multiple grade levels and districts, grouped by demographic and socioeconomic factors.
- (ii) About what proportion (to within 0.1) of the population is captured in the sample? (*Hint*: have a look at this website.)

#Answer:Based on data from the California Department of Education's Fingertip Facts, the total public school enrollment is approximately 5.84 million students. If the dataset includes most districts, it likely represents a large proportion, making it a near-census of district-level performance. The scope of inference is limited to district-level trends rather than individual student performance, allowing for analyses of educational disparities demographic effects, and overall academic achievement across districts.

- (iii) Considering that the sampling frame is not identified clearly, what kind of dataset do you suspect this is (*e.g.*, administrative, data from a 'typical sample', census, etc.)?

#Answer: I think this is an administrative dataset rather than data from a randomly selected sample. The data appears to be collected from official school records and standardized test results, making it a census-like dataset of district-level academic performance rather than a sample-based survey.

- (iv) In light of your description of the sample characteristics, what is the scope of inference for this dataset?

#Answer:The scope of inference for this dataset is district-level educational performance and demographic trends across California public schools. Since the dataset aggregates test scores and covariates at the district and grade level, it allows for inferences about district-wide academic achievement, disparities across demographic groups, and regional trends in education.

Data tidying

Since you've already had some guided practice doing this in previous assignments, you'll be left to fill in a little bit more of the details on your own in this assignment. You'll work with the following variables from each dataset:

- **Test score data**
  - District ID
  - District name
  - Grade
  - Test subject
  - Estimated male-female gap
- **Covariate data**
  - District ID
  - Locale
  - Grade
  - Socioeconomic status (all demographic groups)
  - Log median income (all demographic groups)
  - Poverty rate (all demographic groups)
  - Unemployment rate (all demographic groups)
  - SNAP benefit receipt rate (all demographic groups)

Question 5: Variable names of interest

Download the codebooks by opening the 'data' directory and downloading the codebook files. Identify the variables listed above, and store the column names in lists named `main_vars` and `cov_vars`.

```
# Store variable names of interest
main_vars <- c("sedalea", "sedaleaname", "grade", "subject",
               "cs_mn_mal", "cs_mn_fem") # For gender gap calculation

cov_vars <- c("sedalea", "locale", "grade",
              "sesall", "lninc50all", "povertyall",
              "unempall", "snapall")
```

Question 6: Slice columns

Use your result from above to slice the columns of interest from the covariate and test score data. Store the resulting data frames as `main_sub` and `cov_sub` (for 'subset').

```r
# Slice columns to select variables of interest
main_sub <- c("sedalea", "sedaleaname", "grade", "subject", "cs_mn_mal", "cs_mn_fem")
main_sub <- ca_main[, main_vars]

cov_sub <- c("sedalea", "locale", "grade",
              "sesall", "lninc50all", "povertyall",
              "unempall", "snapall")
cov_sub <- ca_cov[, cov_vars]
```

Question 7: Merge

Merge the covariate and test score data on both the ***district ID*** and ***grade level*** columns, and retain only the columns from the test score data (meaning, merge the covariate data *to* the test score data). You should use the `left_join` function with `main_sub` as the "left" table so all rows of `main_sub` are retained. Store the resulting data frame as `rawdata` and print the first four rows.

```r
rawdata <- left_join(main_sub, cov_sub, by = c("sedalea", "grade"))

# Print first four rows
head(rawdata, 4)
```

Question 8: Rename and reorder columns

Use `mutate` to create a new variable called `Gender gap` which is `cs_mn_mal - cs_mn_fem`. Use `rename()` to rename and rearrange the columns of `rawdata` so that they appear in the following order and with the following names:

- `District ID`, `District`, Locale, `Log(Median income)`, `Poverty rate`, `Unemployment rate`, `SNAP rate`, `Socioeconomic index`, Grade, Subject, `Gender gap`.

Select only the rows above, store the resulting data frame as `rawdata_mod1` and print the first four rows.

```r
# Define dictionary for renaming columns

name_dict <- c(
  "sedalea" = "District ID",
  "sedaleaname" = "District",
  "locale" = "Locale",
  "lninc50all" = "Log(Median income)",
```

```r
  "povertyall" = "Poverty rate",
  "unempall" = "Unemployment rate",
  "snapall" = "SNAP rate",
  "sesall" = "Socioeconomic index",
  "grade" = "Grade",
  "subject" = "Subject",
  "Gender_gap" = "Gender gap"
)


col_order <- c("District ID", "District", "Locale", "Log(Median income)",
               "Poverty rate", "Unemployment rate", "SNAP rate",
               "Socioeconomic index", "Grade", "Subject", "Gender gap")


rawdata_mod1 <- rawdata |>
  mutate(Gender_gap = cs_mn_mal - cs_mn_fem) |>
  rename(!!!setNames(names(name_dict), name_dict)) |>  # From Chatgpt
  select(all_of(col_order))

# Print first five rows
head(rawdata_mod1, n = 5)
```

Question 9: Pivot

Notice that the Gender gap column contains the values of two variables: the gap in estimated mean test scores for math tests, and the gap in estimated mean test scores for reading and language tests. To put the data in tidy format, use `pivot_longer` to pivot the table so that the gender gap column is spread into two columns `mth` and `rla`. Rename these columns `Math` and `Reading`. Store the result as `seda_data` and print the first five rows.

```r
# Pivot to unstack gender gap
seda_data <- rawdata_mod1 |>
  pivot_longer(
    cols = "Gender gap",
    names_to = "Subject_Type",
    values_to = "Gender_value"
  ) |>
  pivot_wider(
    names_from = Subject,
    values_from = Gender_value
```

```
  ) |>
  rename(
    Math = mth,
    Reading = rla
  )

# Print first five rows
head(seda_data, 5)
```

Your final dataset should match the dataframe below. You can use this to check your answer and revise any portions above that lead to different results.

```
# intended result
data_reference <- read_csv('data/tidy-seda-check.csv')
data_reference
```

Question 10: Sanity check

Ensure that your tidying did not inadvertently drop any observations: count the number of units in `seda_data`. Does this match the number of units represented in the original test score data `ca_main`? Store these values as `data_units` and `ca_main_units`, respectively.

```
# number of districts in tidied data compared with raw

data_units <- seda_data |> distinct(District) |> count()
ca_main_units <- ca_main |> distinct(sedaleaname) |> count()
cat("Number of units in tidied data:", data_units$n, "\n")
cat("Number of units in original data:", ca_main_units$n, "\n")
```

Question 11: Missing values

Gap estimates were not calculated for certain grades in certain districts due to small sample sizes (not enough individual tests recorded). Answer the following:

- (i) What proportion of rows are missing for each of the reading and math gap variables? Store these values as `math_missing` and `reading_missing`, respectively.
- (ii) What proportion of *districts* (not rows!) have missing gap estimates for one or both test subjects for at least one grade level? Store the value as `district_missing`.

```
# proportion of missing values
math_missing <- seda_data |> summarize(proportion_missing = mean(is.na(Math))) |> pull(propo
# proportion of districts with missing values
reading_missing <- seda_data |>
  summarize(proportion_missing = mean(is.na(Reading))) |>
  pull(proportion_missing)
```

Question 12: Missing mechanism

Do you expect that this missingness is more likely for some districts than for others? If so, explain; why is this, and is bias a concern if missing values are dropped?

#Answer:Yes, missing data is more likely more common in some districts due to differences in size, testing policies and geographic location. Smaller or lower-income districts may have more missing values due to fewer students tested or reporting issues. Dropping missing values could introduce bias by overrepresenting heigher-performing districts and underestimating achievement gaps.

Question 13: Santa Barbara Unified

It's often helpful to build intuition and check your data by exploring observations for which you are (or might be familiar). Filter the district to "SANTA BARBARA UNIFIED", the district for the city of Santa Barbara. schools. Select `Grade`, math and reading gaps. Print all rows of the data frame for these 3 columns only. For which grades and subjects did boys outperform girls in Santa Barbara?

Optional: if you went to elementary school or Junior High in California, complete the exercise for your former school district instead.

```
seda_data |>
  filter(District == "SANTA BARBARA UNIFIED") |>
  select(Grade, Math, Reading)
```

Exploratory graphics

For the purpose of visualizing the relationship between estimated gender gaps and socioeconomic variables, you'll find it more helpful to work with a longer non-tidy version of the data. The cell below rearranges the dataset so that one column contains an estimated gap, one column contains the value of a socioeconomic variable, and the remaining columns record the gap type and variable identity.

Ensure that your results above match the reference dataset before running this cell.

```
# plot gap against socioeconomic variables by subject for all grades
plot_df <- seda_data |>
  pivot_longer(
    cols = c(Math, Reading),
    names_to = "Subject",
    values_to = "Gender gap"
  ) |>
  pivot_longer(
    cols = c(`Poverty rate`, `Log(Median income)`, `Unemployment rate`, `SNAP rate`, `Socioe
    names_to = "Socioeconomic Variable",
    values_to = "Socioeconomic Value"
  )

# preview
print(plot_df, n=5)
```

Gender gaps and socioeconomic factors

Question 14: Scatter Plot

Create a panel of scatterplots showing the relationship between estimated gender gap and socioeconomic factors, with points colored by test subject. Use `facet_wrap` to show a plot for each of the 5 socioeconimc variables. Adjust the size and transparency of the points to reduce the impact of overplotting. Which subject has the larger gap, reading or math? Which gender is performing better in this subject?

Hint: use `scales="free_x"` in the `facet_wrap` call to ensure that each faceted plot gets it's own x-limits.

```
fig1 <- plot_df |>
  ggplot(aes(x = `Socioeconomic Value`, y = `Gender gap`, color = Subject)) +
  geom_point(alpha = 0.6, size = 2) +
  facet_wrap(~ `Socioeconomic Variable`, scales = "free_x") +
  labs(
    title = "Gender Gap vs. Socioeconomic Factors",
    x = "Socioeconomic Value",
    y = "Gender Gap",
    color = "Test Subject"
  ) +
  theme_minimal() +
```

```
  theme(
    legend.position = "bottom"
  )

fig1
```

#Answer: The reading gender gap (blue) is larger than the Math gender gap (red). In Reading, girls tend to outperform boys more significantly, especially in districts with lower socioeconomic indices and higher unemployment rates. Girls perform better in Reading, as the gender gap is mostly negative, meaning boys score lower.

Question 15: linear fits.

You can tell from the previous plot which subject has the larger gender gap and who is outperforming, but it's hard to distinguish more subtle patterns about how these relationships depend on socioeconomic variables. Instead of a scatter plot, we'll use `geom_smooth` which can be used to to fit smooth functions to the data. We'll focus on linear fits to the data by specifying the `geom_smooth(method="lm")`, where here `lm` stands for "linear model". Again, color the lines by subject (setting the `col` aesthetic) and facet by socioeconomic variable (and set `scale="free_x"` again in the `facet_call`).

For what subjects and socioeconomic values do boys outperform girls? Is the relationship between soeconomic variables and gender gap the same for each subject?

#Answer: Boys outperform girls in Math, in higher-income districts and Math gender gap increases in favor of boys, in low-income or high-poverty areas, boys tend to perform worse. No, the gender gap in Reading (blue) behaves differently from Math (red). The Reading gap is consistently negative, meaning girls outperform boys across all socioeconomic levels. The Math gap varies more, with boys outperforming firls in higher-income districts.

```
fig2 <- plot_df |>
  ggplot(aes(x = `Socioeconomic Value`, y = `Gender gap`, color = Subject)) +
  geom_smooth(method = "lm", se = TRUE, alpha = 0.2) +  # Linear regression with confidence i
  facet_wrap(~ `Socioeconomic Variable`, scales = "free_x") +  # Facet by socioeconomic varia
  labs(
    title = "Gender Gap vs. Socioeconomic Factors (Linear Fit)",
    x = "Socioeconomic Value",
    y = "Gender Gap",
    color = "Test Subject"
  ) +
  theme_minimal() +  # Clean theme
  theme(
```

```
    legend.position = "bottom"
  )

fig2
```

Question 16: Relationships by grade level

Modify the plot above to show these relationships by grade level: generate a panel of scatterplots of gap against socioeconomic measures by subject, coloring the lines according to `Grade`. Since `Grade` is a numeric variable by default, if we want separate colored lines for each grade, we need to specify that `Grade` should be treated as factor in the aesthetics. To do so, set `col=as.factor(Grade)`. Add `colorspace::scale_color_discrete_sequential(palette="Viridis")` to your `ggplot` to set the grades to the Viridis color palette.

Does the pattern shown in the plot above persist within each grade level? For which grades are the gender gaps the largest? Is this pattern consistent across subjects?

```
fig3 <- plot_df |>
  ggplot(aes(x = `Socioeconomic Value`, y = `Gender gap`, color = as.factor(Grade))) +
  geom_smooth(method = "lm", se = FALSE, alpha = 0.8) +  # Linear fit without confidence banc
  facet_wrap(~ `Socioeconomic Variable`, scales = "free_x") +  # Facet by socioeconomic facto
  labs(
    title = "Gender Gap vs. Socioeconomic Factors by Grade Level",
    x = "Socioeconomic Value",
    y = "Gender Gap",
    color = "Grade Level"
  ) +
  colorspace::scale_color_discrete_sequential(palette = "Viridis") +  # Apply Viridis color s
  theme_minimal() +  # Clean theme
  theme(
    legend.position = "bottom"  # Move legend below plot for clarity
  )

fig3
```

#Answer:Yes, the pattern persists across grades, but the gender gap widens in higher grades. Boys outperform girls more in wealthier districts, especially in grades 7 and 8. Girls consistently perform better across all grades, but the gap is slightly smaller in higher grades. Math gaps vary more with socioeconomic factors, while Reading gaps remain stable.

Question 17: Does locale matter?

Let's focus just on math scores and consider whether or not local matters. The chunk below adds a new variable, `locale2` which is a coarser summary of locale which inclues just "Rural", "Town", "Suburbu" and "City", but not the more detailed descriptions.

```
plot_df <- plot_df |> mutate(locale2 =
                    as_factor(case_when(
                      startsWith(Locale, "Rural") ~ "Rural",
                      startsWith(Locale, "Suburb") ~ "Suburb",
                      startsWith(Locale, "Sururb") ~ "Suburb",  ## There is a typo in the da
                      startsWith(Locale, "City") ~ "City",
                      startsWith(Locale, "Town") ~ "Town", .default=Locale)))
```

As you did above, make a plot showing the relationship between socioeconimc variables and the math gap. Do so by first filtering `plot_df` to only include math subjects and then use `geom_smooth(method="lm", se=FALSE)`. Set the color aesthetic to `locale2` so that a line is produced for each locale. Use any color palette you'd like for the lines. Does any locales have larger average gaps than others? Do some locales show a stronger or weaker relationship with the socioeconomic variables?

```
 math_df <- plot_df |> filter(Subject == "Math")

fig4 <- math_df |>
  ggplot(aes(x = `Socioeconomic Value`, y = `Gender gap`, color = locale2)) +
  geom_smooth(method = "lm", se = FALSE) +  facet_wrap(~ `Socioeconomic Variable`, scales = "
    title = "Math Gender Gap vs. Socioeconomic Factors by Locale",
    x = "Socioeconomic Value",
    y = "Math Gender Gap",
    color = "Locale"
  ) +
  scale_color_brewer(palette = "Dark2") +
  theme_minimal() +
  theme(
    legend.position = "bottom"
  )

fig4
```

#Answer:Suburban and city areas show the largest Math gender gaps, with boys outperforming girls as income increases and poverty decreases. In contrast, rural and town areas have smaller gaps and show weaker relationships between socioeconomic factors and gender disparities. The gap in math performance is more sensitive to economic conditions in suburban and city locales compared to rural areas.

Question 18: Aggregation across grade levels

Compute the mean estimated achievement gap in each subject, averaged across grade levels by district using `District` and retain the district-level socioeconomic variables. Store the resulting data frame as `seda_data_agg`.

Note: best practice here would be to aggregate just the test scores by district and then re-merge the result with the district-level socioeconomic variables. However, since the district-level socioeconomic variables do not differ by grade within a district, averaging them across grade levels by district together with the test scores will simply return their unique values; so the aggregation can be applied across *all* columns for a fast-and-loose way to obtain the desired result.

```r
# aggregate across grades
seda_data_agg <- seda_data |>
  group_by(District) |>
  summarise(across(everything(), mean, na.rm = TRUE))

head(seda_data_agg)
```

The cell below adds an `Income bracket` variable by cutting the median income into 8 contiguous intervals using `cut()`.

```r
## Filter nans
seda_data_agg <- seda_data_agg |>
  drop_na(Math, Reading, `Log(Median income)`) |>
  mutate(
    `Income bracket` = cut(
      exp(`Log(Median income)`),
      breaks = 8,
    )
  )
```

As an example, the next cell tabulates the average socioeconomic measures and estimated gaps across districts by income bracket. Does the data in this table roughly match the trends in the plots you saw above?

```
seda_data_agg |>
  group_by(`Income bracket`) |>
  summarise(
    across(-c(`District ID`, District, Locale, `Log(Median income)`), # The minus indicates '
      \(x) mean(x, na.rm=TRUE)
    )
  ) |>
  arrange(desc(`Income bracket`)) |> ## arrange from highest income to lowest
  select(`Income bracket`, Math, Reading)
```

Question 19: Proportion of districts with a math gap

What proportion of districts in each income bracket have an average estimated math achievement gap favoring boys? Answer this question by performing the following steps:

- Append an indicator variable `Math gap favoring boys` to `seda_data_agg` that records whether the average estimated math gap favors boys *by more than 0.1 standard deviations relative to the national average.*
- Compute the proportion of districts in each income bracket for which the indicator is true: group by bracket and summarize `Math gap favoring boys` by taking the mean. Store the resulting data frame as `income_bracket_boys_favored`. What fraction of districts with a log income of $95,000-$119000 have a math gap favoring boys by more than 0.1 standard deviations?

```
# define indicator
income_bracket_boys_favored <- seda_data_agg |> mutate(`Math gap favoring boys` = ifelse(Mat

# print result
print(income_bracket_boys_favored, n=10)
```

#Answer: It looks like in the dataset provided, none of the districts in the income bracket of $95,000–$119,000 have a math gap favoring boys by more than 0.1 standard deviations. The "Math gap favoring boys" column contains only zeros, indicating that in this income range, the estimated math gaps do not significantly favor boys.

Question 20: Statewide averages

To wrap up the exploration, calculate a few statewide averages to get a sense of how some of the patterns above compare with the state as a whole.

- (i) Compute the statewide average estimated achievement gaps. Store the result as `state_avg`.
- (ii) How many districts in the state have a reading gap favoring boys? Store this result as `reading_boys_number`

```
# statewide average
state_avg <- seda_data_agg |>
  summarise(
    avg_math_gap = mean(Math, na.rm = TRUE),
    avg_reading_gap = mean(Reading, na.rm = TRUE)
  )

# proportion of districts in the state with a math gap favoring boys
math_boys_proportion <- seda_data_agg |>
  summarise(
    proportion = mean(Math > 0, na.rm = TRUE)
  )

# proportion of districts in the state with a math gap favoring girls
reading_boys_number <- seda_data_agg |>
  summarise(
    count = sum(Reading > 0, na.rm = TRUE)
  )

print(state_avg)
print(math_boys_proportion)
print(reading_boys_number)
```

Communicating results

Take a moment to review and reflect on your findings and consider what you have learned from the analysis.

Question 21: Summary

Write a brief summary of your exploratory analysis. What have you discovered about educational achievement gaps in California school districts? Aim to answer in 3-5 sentences or less.

#Answer: The exploratory analysis revealed significant gender gaps in educational achievement across California school districts. Girls consistently outperform boys in reading, while boys have a slight advantage in math, with variations influenced by socioeconomic factors. Higher-income districts tend to show smaller gender disparities, whereas lower-income and higher-poverty districts exhibit larger gaps. Locale also plays a role, with urban, suburban, and rural districts displaying different trends in gender gaps. These findings suggest that both socioeconomic conditions and geographic factors contribute to disparities in student performance, highlighing some important areas fro targeted policy interventions.