Data & Source

The data used for this project come from the KIDS COUNT Study, which is conducted by the Annie E. Casey Foundation (https://datacenter.kidscount.org/). Data is collected at both the national level and the state level within the US. The KIDS COUNT project contains data related to a number of different aspects of child development and well-being, including education, health, risky behavior, and family/community relationships.

For the specific research questions I'm interested in looking at, achievement data are available for fourth and eighth graders. Data were collected every two years, from 2003-2022 (data were released in 2022 instead of 2021). Data are available per state, and nationally. Achievement data are represented using 4 categories (below basic, at or above basic, below proficient, at or above proficient). The proficient category builds on the basic category, so below basic and at or above basic add up to 100% of students, and below proficient and at or above proficient add up to 100% of students.

The data are available through a data portal, located at https://datacenter.kidscount.org/. The data are openly available, and don't require registration to be accessed. ## Research Questions 1. How has educational achievement in math and science changed over time? For this question, I'll look at math and science achievement separately, and look at the change over time. Additionally, I'll look at both basic and proficient levels of achievement. I'm not yet sure whether I'll look at state-level data, or national. I might use both. 2. How is educational achievement in math and science related to funding? For this question, I'm planning on looking at state-level data and achievement (both basic and proficient), as state-level educational expenditures (on a per-student basis) allows us to compare relationships between different states and funding types. I'm not sure whether I'll look at this cross-sectionally, or longitudinally.

Outline of Visualizations

Preliminary ideas

Question 1 To address this question, I'm planning on making a line plot/connected scatterplot, with time on the x-axis, and achievement on the y-axis. Achievement in each subject will be represented by a different line. From looking at the data, achievement is represented as 4 categories of achievement, and data for the proportion of students that are in each achievement category. I'm thinking of combining some of the categories to represent the data as the proportion of students that are at or above proficient.

I have a couple of ideas for how to deal with national vs state data. One option is to facet the plots, with a different panel for different regions. In this version, I'd probably have one panel for national-level data, and the panels for different regions in the US. Another option that I'm considering is to plot state or regions as different lines in the same plot. In this version, I'd use color to help separate regions or subjects. For example, using different shades of the same color for each subject, or using different shades/opacities of the same color for each region across subjects. In the latter case, that would look something like using darker versions of a color palette for math, and lighter versions for science. Alternately, I could use different line types instead of colors, or combine the two.

I'm not yet sure how this fits into the data, but I might also try to add vertical lines at years where new educational standards were introduced. I don't have enough data to really evaluate new educational standards, so I don't want to misrepresent what I'm showing. However, I think it might also be helpful to provide context to potential trends.

Question 2 As step 1 for this research question, I might include some plots just characterizing funding on a per-student basis between different states. To show funding alone, I could make a circular bar plot (https://r-graph-gallery.com/circular-barplot.html). I'm thinking of using this because it will allow me to include data for all 50 states in one graph, without making the graph too overwhelming. I'll either group regions together, or color-code by region. Alternately, I might make a map visualization of funding, which would use color to show funding levels. To compare funding and achievement, I'll make separate plots for each educational subject.

For both questions, I might also experiment with making the plots interactive, so that viewers could choose states/regions to include in the visualization.

Intended audience

Question 1 The intended audience for this visualization would be educational researchers, or policy makers evaluating educational needs.

Question 2 The intended audience for this set of visualizations is similar to the audience of Question 1. The audience would primarily be policy makers, or potentially parents/other people who might lobby or advocate for educational funding.

Intended message

Question 1 The intended message of this visualization is to illustrate both trends in achievement over time, and how achievement in different domains may track (or not) with each other.

Question 2 The intended message of the first plot (descriptive plot) is to show nationwide funding levels, and illustrate geographic trends in funding. The intended message of the second plot/set of plots is to show the impact of funding on educational achievement, and to show how the impact differs by subject.

Exploratory Data Analysis

Load data

```
math_achievement <- rio::import(here::here("finalproject", "Eighth-grade math achievement levels.xlsx"))
science_achievement_1 <- rio::import(here::here("finalproject", "Eighth-grade science achievement levels
science_achievement_2 <- rio::import(here::here("finalproject", "Eighth-grade science achievement levels
funding <- rio::import(here::here("finalproject", "Per-pupil educational expenditures.xlsx"))
```

Rearrange data

```
# Math
#Rename column and covert data to numeric
math_achievement <- math_achievement %>%
  rename("AchievementLevel" = "Achievement Level") %>%
  mutate(Data = as.numeric(Data))
```

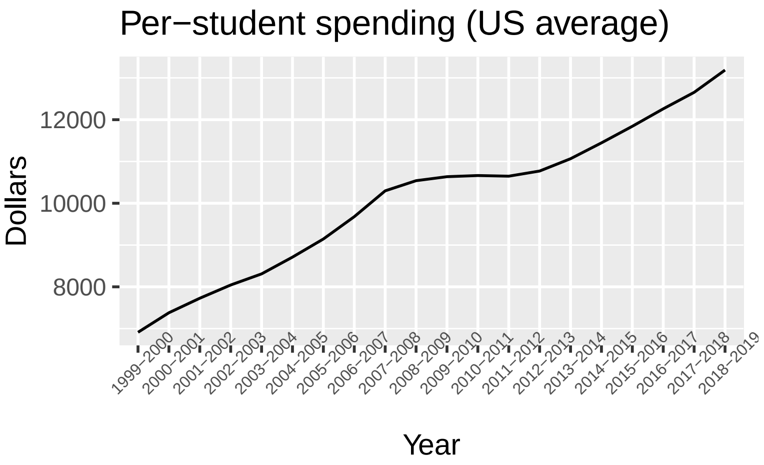
```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
# Science
#rbind both science achievement files to create one with all years, rename column, and convert data to
science_achievement <- rbind(science_achievement_1, science_achievement_2) %>%
  rename("AchievementLevel" = "Achievement Level") %>%
  mutate(Data = as.numeric(Data))
```

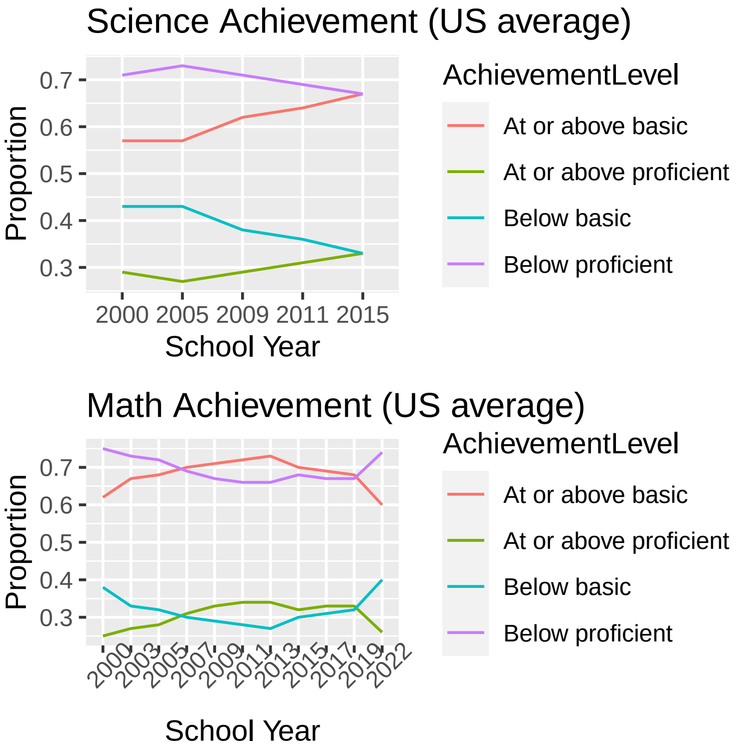
```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
# Funding
#Convert numbers from character to numeric
funding$Data <- as.numeric(funding$Data)
```

```
funding %>%
  filter(Location=="United States") %>%
  ggplot(aes(x=TimeFrame, y=Data)) +
  geom_line(aes(group=1)) +
  labs(x="Year", y = "Dollars", title = "Per-student spending (US average)")+
  theme(axis.text.x = element_text(size=6, angle = 45))
```



```
science_plot <-science_achievement %>%
  filter(Location=="United States") %>%
  ggplot(aes(x=TimeFrame, y=Data, group = AchievementLevel, color=AchievementLevel)) +
  geom_line() +
  labs(x="School Year", y = "Proportion", title = "Science Achievement (US average)")
math_plot <- math_achievement %>%
  filter(Location=="United States") %>%
  ggplot(aes(x=TimeFrame, y=Data, group = AchievementLevel, color=AchievementLevel)) +
  geom_line() +
  labs(x="School Year", y = "Proportion", title = "Math Achievement (US average)")+
  theme(axis.text.x = element_text( angle = 45))
grid.arrange(science_plot, math_plot)
```


Science Achievement (US average)
Proportion
School Year
AchievementLevel
At or above basic
At or above proficient
Below basic
Below proficient
Math Achievement (US average)
Proportion
School Year
AchievementLevel
At or above basic
At or above proficient
Below basic
Below proficient