

# FinalProj184

2024-12-18

## **Exploring Government Expenditure on Education Impact on Education**

**Liam Geary, Josiah Harrison, Junwen Zeng**

With the recent election here in the US a major topic that is frequently brought up is how to improve the education in this country. A major issue with this debate is how drastically opinions can differ between people, with questions of how education has changed for better or worse, what is causing these changes, and even how to measure a good education. Here we want to focus on the topic that is discussed the most in this debate, which is how increasing funding impacts education.

This is relatively common belief held by many Americans, which is why in this report we will be using data from the World Education Data Set to explore the relationship between percent of GDP spent on education and education for that country. We will start by going over relevant background information, then go into how we created our study, what we are looking for, the data we are looking at, issues we ran into, then we will go in depth on our findings from the study and talk about what we see from the data

## **Background Information**

Our research question has been looked into before with some of the big examples being a 2017 study from the Learning Policy Institute, as well as a collection of studies from 2024 from Research.com which went over the states that spend the most on education, where they get the money, and how good their education is. The general consensus from both studies is that yes, the more money you spend the better the education will be. What is interesting is to see how they measure the quality of the education by looking at resources available to teachers and students, teacher student ratio, as well as success rate of the students which here means graduation rates, average GPA and so on. How their research was conducted did inspire us to choose certain categories for our research like teacher student ratio, but our main difference is we took less of an interest in how spending affects resources for teacher and student since

we mainly wanted to focus on variables that more directly correlate to a better education like literacy rate, and graduation rates since we found them more important and to the point.

## Methodology

The data as mentioned above, are from Kaggle, the World Education Data Set from the data source World Bank. The data meets the FAIR and CARE principles, it is findable, and easy to access because it is on the web page of Kaggle, and anyone can find it easily. It is accessible, there's no barrier preventing any user from downloading the dataset, therefore anyone can access it. It is interoperable, although the data does have some limitations, with proper piping and cleaning up the data, it can be integrated and used alongside other data. The data is reusable, in the webpage of Kaggle, there are written potential use cases, key questions, and some notes to guide the user of the data when the dataset might be in use. The data meets the collective benefit, it can be used to evaluate how well the government expenditure impacts several aspects of the country, beneficial in enhancing the economic outcomes. The data has the authority of control

For our study we will measure the quality of education by looking at the relationship between education percent spending with literacy rate, primary school completion percent, school enrollment rate, and teacher pupil ratio.

Install the "readr" Package and get the data from my .csv file.

```
options(repos = c(CRAN = "https://cloud.r-project.org/"))

library(readr)
data <- read_csv("/Users/junwen/Desktop/Stat 184/world-education-data.csv")
```

```
Rows: 5892 Columns: 11
```

```
-- Column specification -----
Delimiter: ","
```

```
chr (2): country, country_code
```

```
dbl (9): year, gov_exp_pct_gdp, lit_rate_adult_pct, pri_comp_rate_pct, pupil...
```

```
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.
```

Summarize the data to keep the values I need, and also remove the year to keep the mean value in terms of each country.

```
library(dplyr)
```

Attaching package: 'dplyr'

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

filter, lag

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

intersect, setdiff, setequal, union

```
summarized_data <- data %>%  
  group_by(country) %>%  
  summarize(  
    gov_exp_pct_gdp = mean(gov_exp_pct_gdp, na.rm = TRUE),  
    pri_comp_rate_pct = mean(pri_comp_rate_pct, na.rm = TRUE),  
    lit_rate_adult_pct = mean(lit_rate_adult_pct, na.rm = TRUE),  
    pupil_teacher_primary = mean(pupil_teacher_primary, na.rm = TRUE),  
    pupil_teacher_secondary = mean(pupil_teacher_secondary, na.rm = TRUE),  
    school_enrol_primary_pct = mean(school_enrol_primary_pct, na.rm = TRUE)  
  )
```

The first try of the table, I tried to color-code each country, but due to color-coding all of the countries, each country has a legend label, and the graph is nowhere to be seen, the legend labels even crashed the R program.

```
library(ggplot2)  
  
# 1st  
ggplot(summarized_data, aes(x = gov_exp_pct_gdp, y = lit_rate_adult_pct, color = country)) +  
  geom_point(size = 3) +  
  labs(  
    title = "Government Expenditure(In percentage of GDP) vs Literacy Rate by Country",  
    x = "Government Expenditure",  
    y = "Literacy Rate"  
  ) +  
  theme_minimal()
```

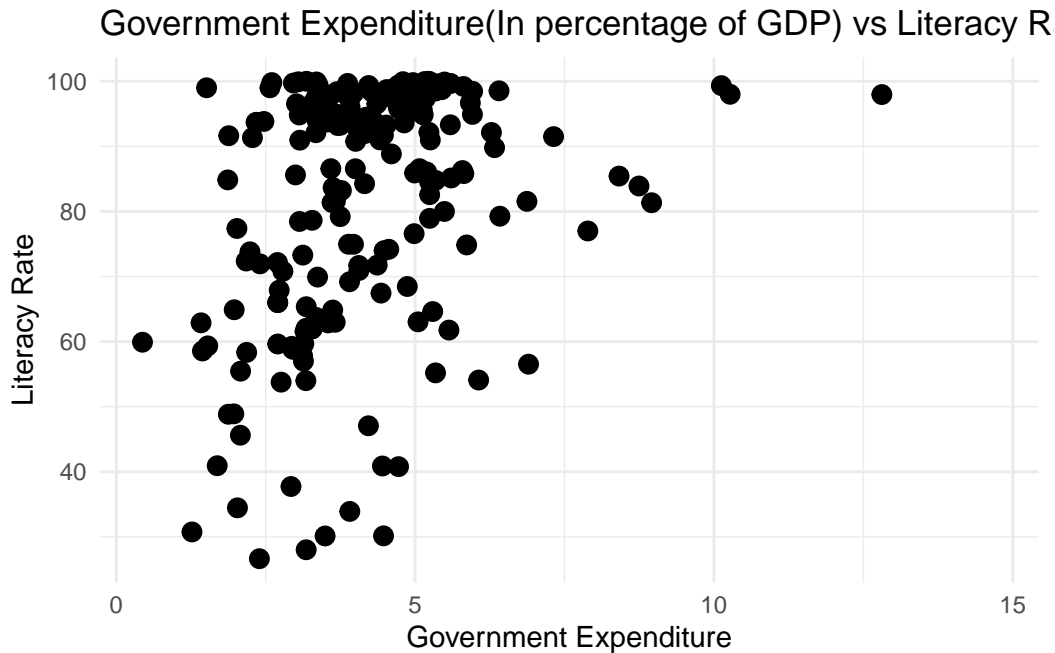
Warning: Removed 59 rows containing missing values or values outside the scale range (``geom_point()``).

Algeria	●	Ruanda	●	Madagascar
IDA & IBRD total	●	Kyrgyz Republic	●	Malawi
IDA blend	●	Lao PDR	●	Malaysia
IDA only	●	Late-demographic dividend	●	Maldives
IDA total	●	Latin America & Caribbean	●	Mali
India	●	Latin America & Caribbean (excluding high income)	●	Malta
Indonesia	●	Latin America & the Caribbean (IDA & IBRD countries)	●	Marshall Islands
Iran, Islamic Rep.	●	Latvia	●	Mauritania
Iraq	●	Least developed countries: UN classification	●	Mauritius
Ireland	●	Lebanon	●	Mexico
Israel	●	Lesotho	●	Micronesia, Fed.
Italy	●	Liberia	●	Middle East & North Africa
Jamaica	●	Libya	●	Middle East & North Africa (IDA & IBRD total)
Japan	●	Liechtenstein	●	Middle East & North Africa (IDA blend)
Jordan	●	Lithuania	●	Middle income

After removing the legend labels, the color code became meaningless.

```
# 2nd
ggplot(summarized_data, aes(x = gov_exp_pct_gdp, y = lit_rate_adult_pct)) +
  geom_point(size = 3) +
  labs(
    title = "Government Expenditure(In percentage of GDP) vs Literacy Rate by Country",
    x = "Government Expenditure",
    y = "Literacy Rate"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

Warning: Removed 59 rows containing missing values or values outside the scale range (``geom_point()``).



The third visualization looks the best, it is quite tidy. This is the visualization of Government Expenditure(In percentage of GDP) vs Literacy Rate by Country. We focused on the numbers, not paying too much attention to which countries were attributed to each dot. Diving into which countries have highest literacy rates with respect to their government expenditure is a great further analysis of this project and data. A comforting statistic found is that most countries are above the 60% literacy rate, with a large amount above 80%. Most of the government expenditures across countries are in the same percentage of their respective GDP, so the spread of literacy rates is fascinating.

```
# 3rd
ggplot(summarized_data, aes(x = gov_exp_pct_gdp, y = lit_rate_adult_pct)) +
  geom_point(color = "black", size = 3) +
  geom_smooth(method = "lm", color = "blue", se = TRUE) +
  labs(
    title = "Government Expenditure(In percentage of GDP) vs Literacy Rate by Country",
    x = "Government Expenditure",
    y = "Literacy Rate"
  ) +
  theme_minimal() +
  scale_y_continuous(limits = c(30, 100))
```

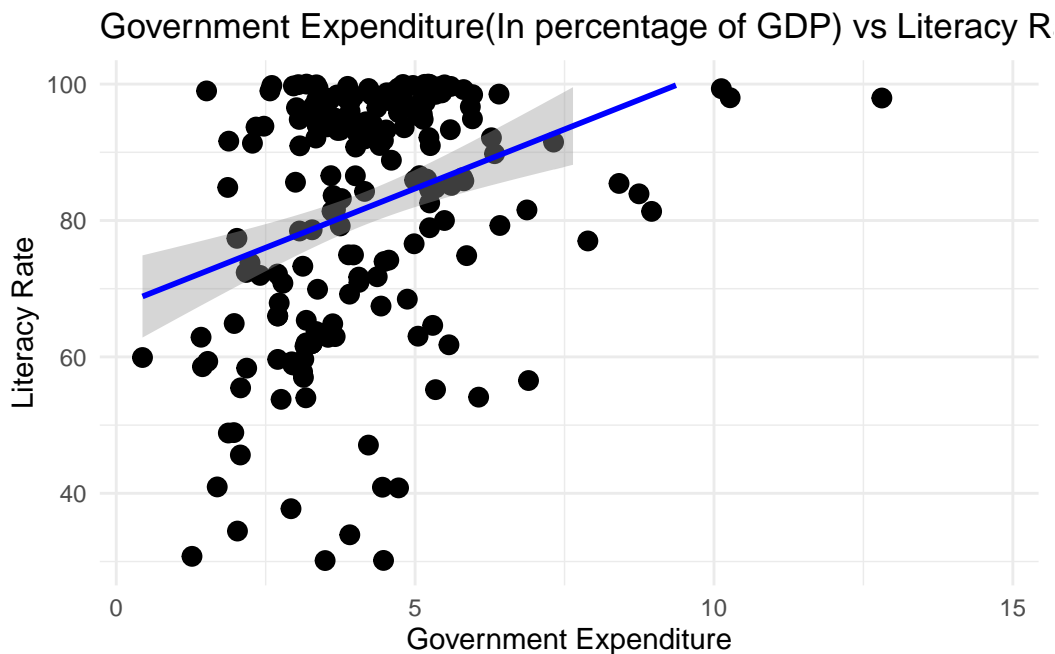
`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 61 rows containing non-finite outside the scale range

```
(`stat_smooth()`).
```

```
Warning: Removed 61 rows containing missing values or values outside the scale range
(`geom_point()`).
```

```
Warning: Removed 22 rows containing missing values or values outside the scale range
(`geom_smooth()`).
```



This is the visualization of Government Expenditure(In percentage of GDP) vs Primary School Completion Rate Percentage by Country. The literacy rate and completion rate graphs were almost identical. This makes sense because those that complete school, would contribute to the literacy rate. Some countries had higher expenditure than most, nearing or surpassing 10% but marginal increases, if any, in completion rate.

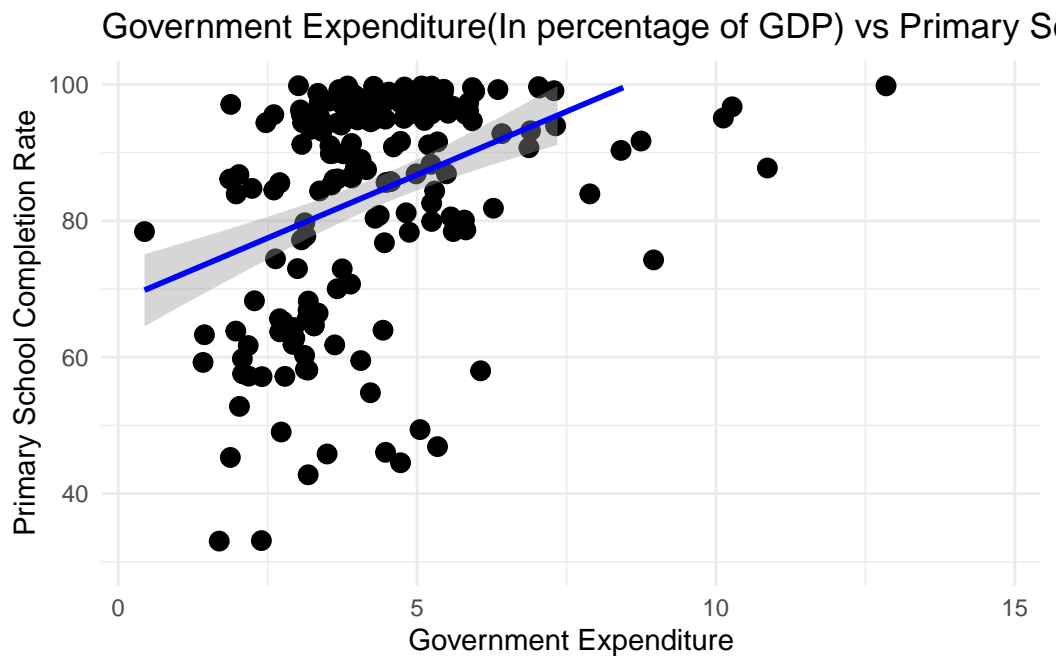
```
ggplot(summarized_data, aes(x = gov_exp_pct_gdp, y = pri_comp_rate_pct)) +  
  geom_point(color = "black", size = 3) +  
  geom_smooth(method = "lm", color = "blue", se = TRUE) +  
  labs(  
    title = "Government Expenditure(In percentage of GDP) vs Primary School Completion Rate",  
    x = "Government Expenditure",  
    y = "Primary School Completion Rate"  
  ) +  
  theme_minimal() +  
  scale_y_continuous(limits = c(30, 100))
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 63 rows containing non-finite outside the scale range  
(`stat_smooth()`).
```

```
Warning: Removed 63 rows containing missing values or values outside the scale range  
(`geom_point()`).
```

```
Warning: Removed 28 rows containing missing values or values outside the scale range  
(`geom_smooth()`).
```



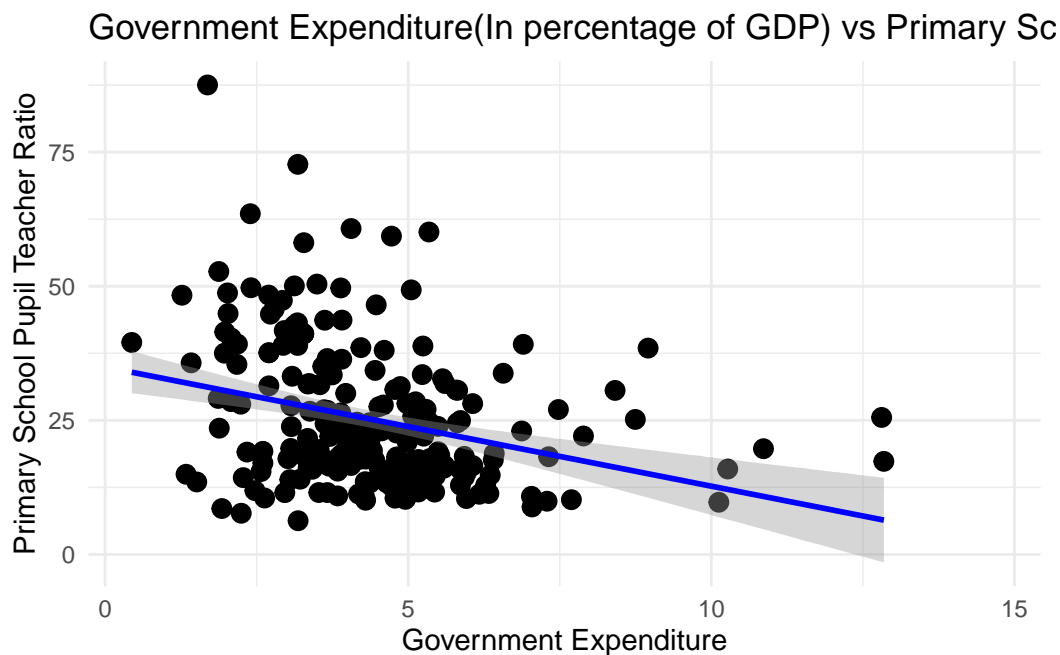
This is the visualization of Government Expenditure(In percentage of GDP) vs Primary School Pupil Teacher Ratio by Country. This visualization looks at Government Expenditure on education (as a percentage of GDP) vs Primary School Pupil Teacher Ratio by Country. Pupil Teacher Ratio is one of the strongest indicators of student success and engagement, according to the Hun School of Princeton. The less number of students assigned to a professor, the more time and attention are given to students. Government expenditure plays an important role in this ratio because of the funds devoted to hiring teachers. The dots represent countries of the study, with most around the 20-1 pupil-teacher ratio. This is a valuable insight because our entire analysis is a broad look at educations systems across the world. Seeing many places have a low ratio is very good for the well-being of the students in those areas.

```
ggplot(summarized_data, aes(x = gov_exp_pct_gdp, y = pupil_teacher_primary)) +
  geom_point(color = "black", size = 3) +
  geom_smooth(method = "lm", color = "blue", se = TRUE) +
  labs(
    title = "Government Expenditure(In percentage of GDP) vs Primary School Pupil Teacher Ra",
    x = "Government Expenditure",
    y = "Primary School Pupil Teacher Ratio"
  ) +
  theme_minimal()
```

`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 19 rows containing non-finite outside the scale range  
(`stat\_smooth()`).

Warning: Removed 19 rows containing missing values or values outside the scale range  
(`geom\_point()`).



This is the visualization of Government Expenditure(In percentage of GDP) vs Secondary School Pupil Teacher Ratio by Country. Secondary School pupil-teacher ratio is very similar to the primary school ratio. Primary school is typically ages 5-11, while secondary school is the next bracket of ages 12-18. There are some outliers that have 30+ ratio, but most fall in the 10-25 ratio.

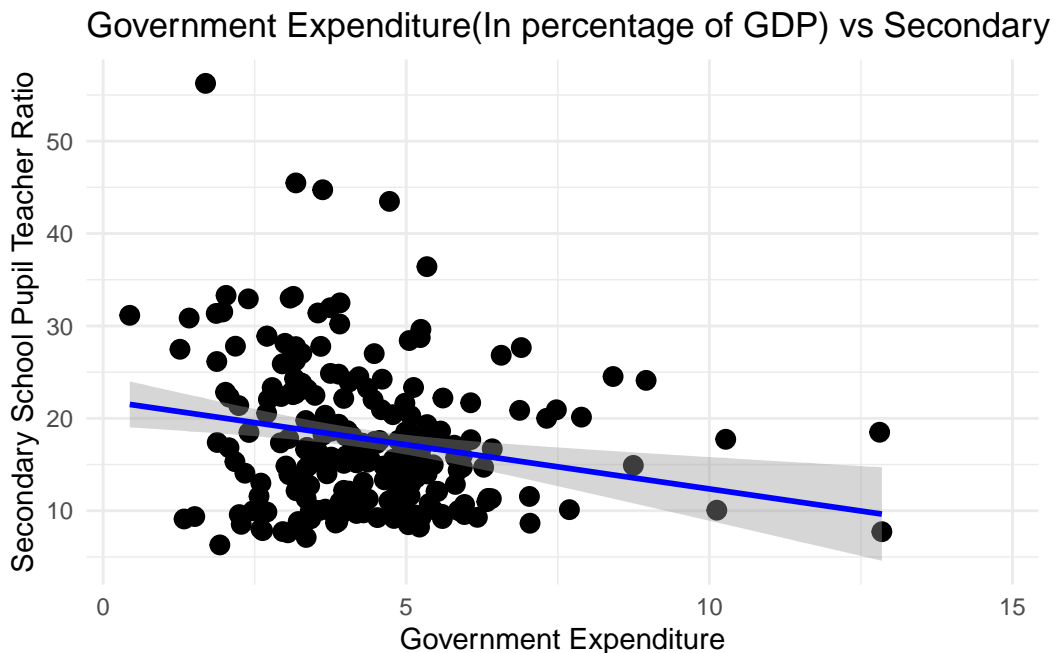


```
ggplot(summarized_data, aes(x = gov_exp_pct_gdp, y = pupil_teacher_secondary)) +
  geom_point(color = "black", size = 3) +
  geom_smooth(method = "lm", color = "blue", se = TRUE) +
  labs(
    title = "Government Expenditure(In percentage of GDP) vs Secondary School Pupil Teacher Ratio",
    x = "Government Expenditure",
    y = "Secondary School Pupil Teacher Ratio"
  ) +
  theme_minimal()
```

`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 29 rows containing non-finite outside the scale range  
(`stat\_smooth()`).

Warning: Removed 29 rows containing missing values or values outside the scale range  
(`geom\_point()`).



This is the visualization of Government Expenditure(In percentage of GDP) vs Primary School Enrollment Rate by Country. We explored government expenditure versus Primary School Enrollment Rate by Country to see any relationship. We found many countries at or above the 100 rate. This is because the enrollment rate of analysis is the gross enrollment. So it

takes people outside of the age group because they enrolled at primary school. So the rate can be 100. If it were the net enrollment rate, than it would not be able to surpass 100%.

```
ggplot(summarized_data, aes(x = gov_exp_pct_gdp, y = school_enrol_primary_pct)) +  
  geom_point(color = "black", size = 3) +  
  geom_smooth(method = "lm", color = "blue", se = TRUE) +  
  labs(  
    title = "Government Expenditure(In percentage of GDP) vs Primary School Enrollment Rate",  
    x = "Government Expenditure",  
    y = "Primary School Enrollment Rate"  
  ) +  
  theme_minimal()
```

`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 13 rows containing non-finite outside the scale range  
(`stat\_smooth()`).

Warning: Removed 13 rows containing missing values or values outside the scale range  
(`geom\_point()`).

