# DATA 606 Data Project Proposal

Michael Ippolito

### Data Preparation

```
# Load libraries
library(tidyverse)
library(psych)

# Load countries by region
# from https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-Codes/master/all/all.cs
regions_full <- read.csv("https://raw.githubusercontent.com/mmippolito/cuny/main/data606/project/country

# Rename some fields for standardization purposes
regions <- regions_full %>%
  rename(
    "country_numeric_code" = "country.code",
    "country_code" = "alpha.3",
    "subregion" = "sub.region"
  )

# Load govt spending on education data set
# from https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS
edspending_full <- read.csv("https://raw.githubusercontent.com/mmippolito/cuny/main/data606/project/gov

# Rename fields to something more useful
edspending <- edspending_full %>%
  rename(
    "country" = "Country.Name",
    "country_code" = "Country.Code"
  ) %>%
  select(-`Indicator.Name`, -`Indicator.Code`, -`X`)

# Remove Xs from years in column names
edspending <- rename_with(edspending, function(x) ifelse(substr(x, 1, 1) == "X", substr(x, 2, 5), x))

# Gather columns
edspending <- edspending %>% gather(3:63, key = "year", value = "pct_gdp")

# Convert years to numeric
edspending$year <- as.numeric(edspending$year)

# Load literacy rates
# from https://commons.wikimedia.org/wiki/Data:Cross-country_literacy_rates_-_World_Bank,_CIA_World_Fac
lit_full <- read.csv("https://raw.githubusercontent.com/mmippolito/cuny/main/data606/project/cross-coun
```

```r
# Rename fields to something more useful
lit <- lit_full %>%
  rename(
    "country" = "Entity",
    "country_code" = "Code",
    "year" = "Year",
    "lit_rate" = "Literacy.rates..World.Bank..CIA.World.Factbook..and.other.sources."
  )

# Calculate mean literacy rate per country
lit_sum <- lit %>% group_by(country, country_code) %>%
  summarize(lit_rate = mean(lit_rate, na.rm = T))

# Calculate mean education spending per country
ed_sum <- edspending %>% group_by(country, country_code) %>%
  summarize(pct_gdp = mean(pct_gdp, na.rm = T))

# Join the summarized tables
j <- ed_sum %>%
  inner_join(lit_sum, by = c("country_code")) %>%
  left_join(regions, by = c("country_code"))

# Filter out NaNs
j <- j %>%
  filter(!is.na(pct_gdp)) %>%
  filter(!is.na(lit_rate))

# Select only the fields we care about
j <- j %>%
  select(country_code, country = country.x, region, subregion, pct_gdp, lit_rate)
```

### Research question

**You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.**

The premise of my project will be to investigate how government spending on education affects a country's literacy rate. The presumption is that countries in richer regions of the world spend more on education and, therefore, enjoy higher literacy rates.

First, I'll pose the question of whether there truly is a statistically significant difference in mean educational spending and mean literacy rate across regions. Then I'll investigate whether the percentage of GDP spent on education influences the literacy rate of a country's population and, if so, what the quantitative impact of that influence is.

### Cases

**What are the cases, and how many are there?**

Each case represents a single country, and there are 191 obvservations in my dataset (after tidying, summarizing, and joining).

### Data collection

**Describe the method of data collection.**

The data comes from two separate data sources, as there wasn't a single source available (to my knowledge). Government spending on education is posted on worldbank.org and is collected annually from the UNESCO Institute for Statistics; data from the years 1960 to 2020 was available for download.

Literacy rates were obtained from an aggregated dataset on from wikimedia.org; the data was originally collected by worldbank.org, the CIA World Factbook, and other sources. The dataset includes some historical data as old as 1475 but with a majority of data points between 1960 and 2015.

### Type of study

**What type of study is this (observational/experiment)?**

This is an observational study.

### Data Source

**If you collected the data, state self-collected. If not, provide a citation/link.**

Educational spending rates were collected from https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS, while literacy data was obtained from https://commons.wikimedia.org/wiki/Data:Cross-country_literacy_rates_-_World_Bank,_CIA_World_Factbook,_and_other_sources_(OWID_2762).tab. A list of countries by global region was downloaded from here: https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-Codes/master/all/all.csv.

### Dependent Variable

**What is the response variable? Is it quantitative or qualitative?**

Literacy rate is the dependent variable; it is quantitative.

### Independent Variable

**You should have two independent variables, one quantitative and one qualitative.**

Government spending and region are the independent variables, and they are quantitative and qualitative, respectively.

### Relevant summary statistics

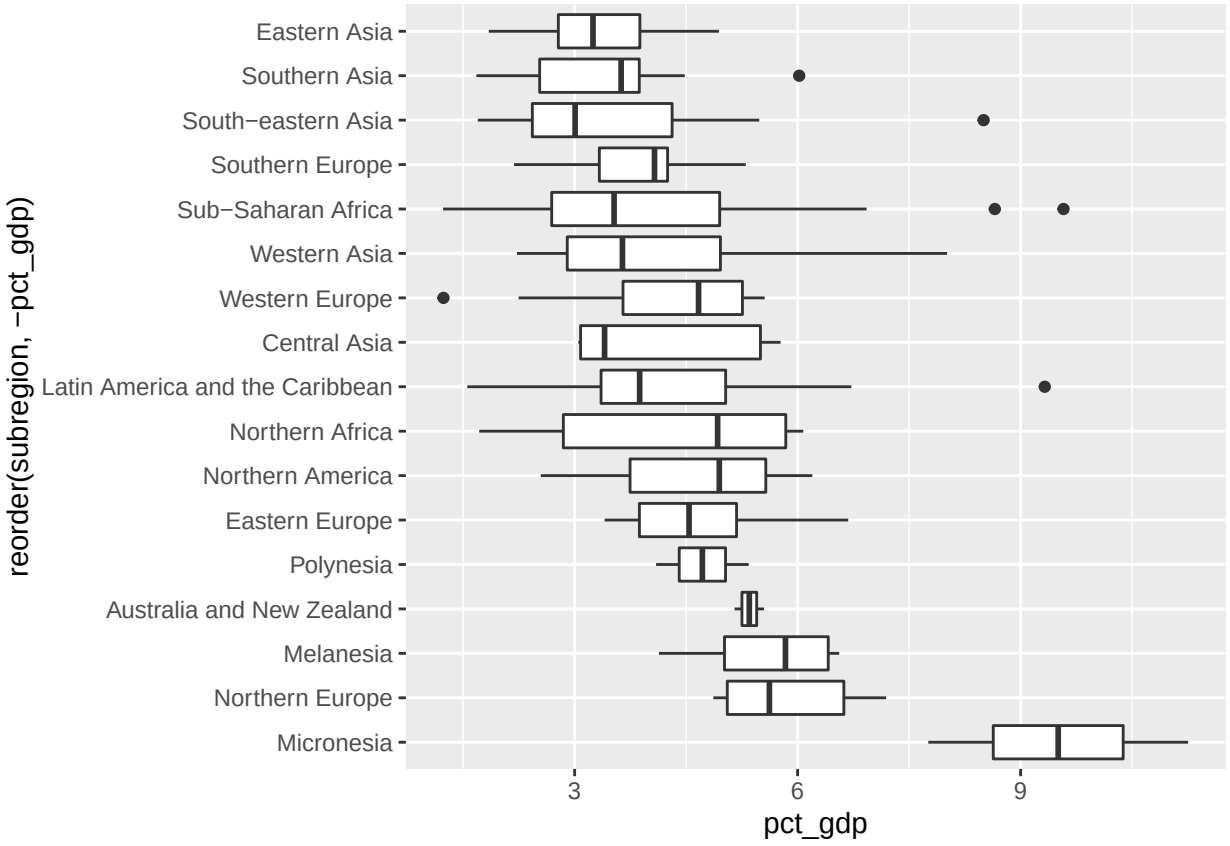
**Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.**

```
 # Summary stats for my categorical variable, region
# (actually I'm choosing subregion to be more granular)
(region_summ <- j %>%
  group_by(subregion) %>%
  summarize(
    n_pct_gdp = n(),
```

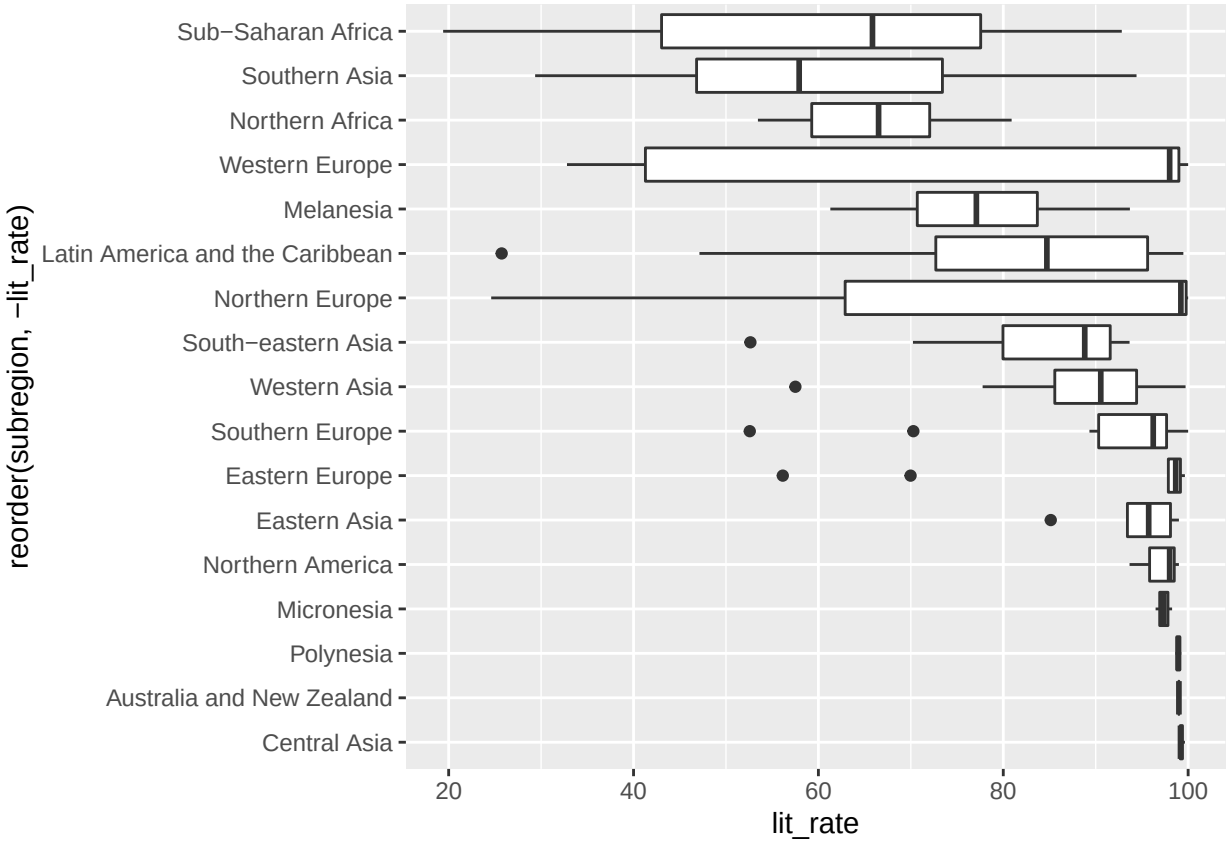
```
    sd_pct_gdp = sd(pct_gdp),
    mean_pct_gdp = mean(pct_gdp),
    n_lit_rate = n(),
    sd_lit_rate = sd(lit_rate),
    mean_lit_rate = mean(lit_rate)
  )
)
```

```
## # A tibble: 17 x 7
##    subregion            n_pct_gdp sd_pct_gdp mean_pct_gdp n_lit_rate sd_lit_rate
##  * <chr>                    <int>      <dbl>        <dbl>      <int>       <dbl>
##  1 Australia and New Z~         2      0.281         5.35          2       0
##  2 Central Asia                 5      1.36          4.16          5       0.221
##  3 Eastern Asia                 6      1.08          3.33          6       5.20
##  4 Eastern Europe              10      1.01          4.66         10      15.4
##  5 Latin America and t~        35      1.52          4.22         35      16.4
##  6 Melanesia                    4      1.12          5.59          4      13.5
##  7 Micronesia                   2      2.47          9.51          2       1.26
##  8 Northern Africa              6      1.90          4.32          6      10.1
##  9 Northern America             3      1.86          4.56          3       2.84
## 10 Northern Europe             10      0.865         5.85         10      29.2
## 11 Polynesia                    2      0.881         4.72          2       0.454
## 12 South-eastern Asia          11      1.99          3.66         11      12.8
## 13 Southern Asia                9      1.32          3.42          9      21.5
## 14 Southern Europe             12      0.855         3.83         12      14.3
## 15 Sub-Saharan Africa          48      1.81          3.92         48      20.8
## 16 Western Asia                17      1.59          4.14         17      10.5
## 17 Western Europe               9      1.50          4.15          9      30.9
## # ... with 1 more variable: mean_lit_rate <dbl>
```

```
# Box plots of ed spending by region
j %>%
  ggplot() +
  geom_boxplot(aes(x = reorder(subregion, -pct_gdp), y = pct_gdp)) +
  coord_flip()
```




```
# Box plots of literacy rate by region
j %>%
  ggplot() +
  geom_boxplot(aes(x = reorder(subregion, -lit_rate), y = lit_rate)) +
  coord_flip()
```



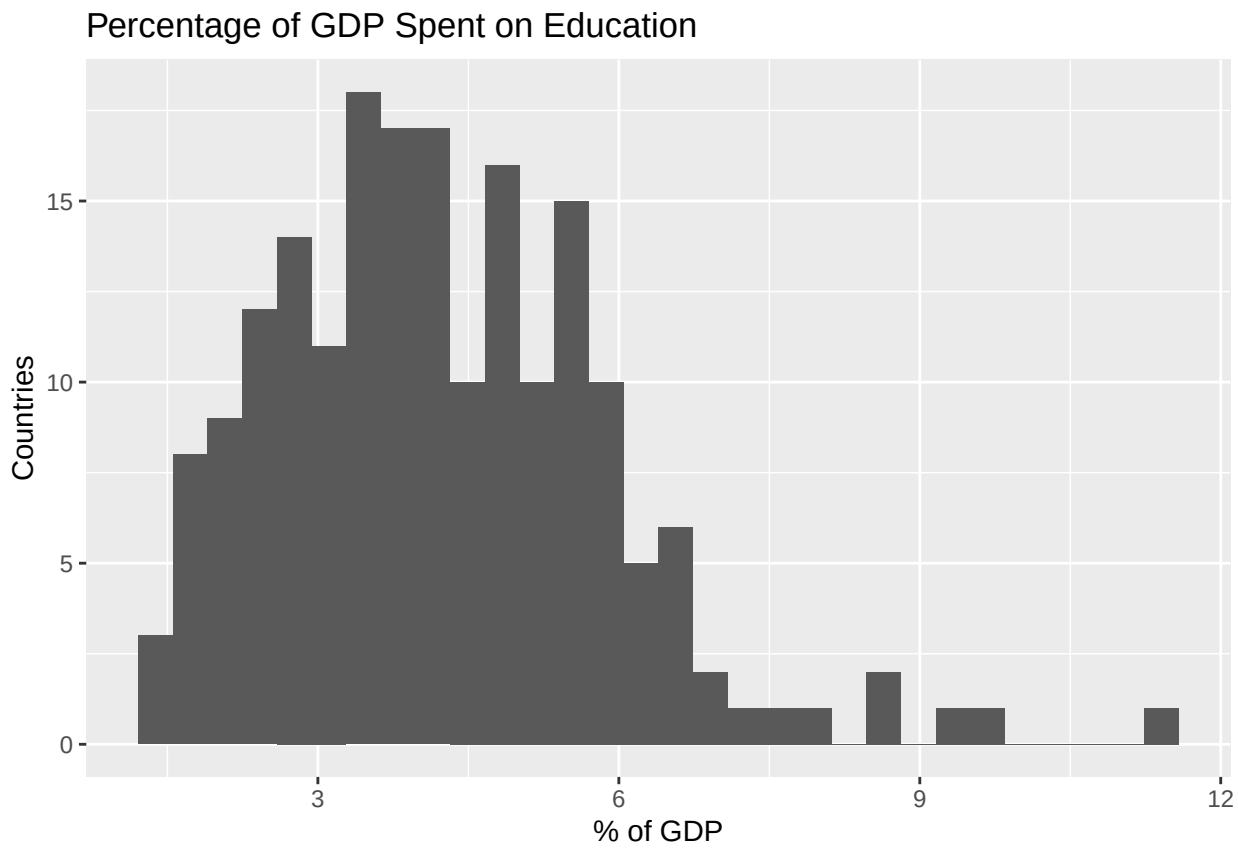

```
# Summary statistics for percent GDP education spending
describe(j$pct_gdp)
```

```
##    vars   n mean   sd median trimmed  mad  min   max range skew kurtosis   se
## X1    1 191 4.23 1.68   4.05    4.12 1.72 1.23 11.25 10.02 0.83     1.36 0.12
```

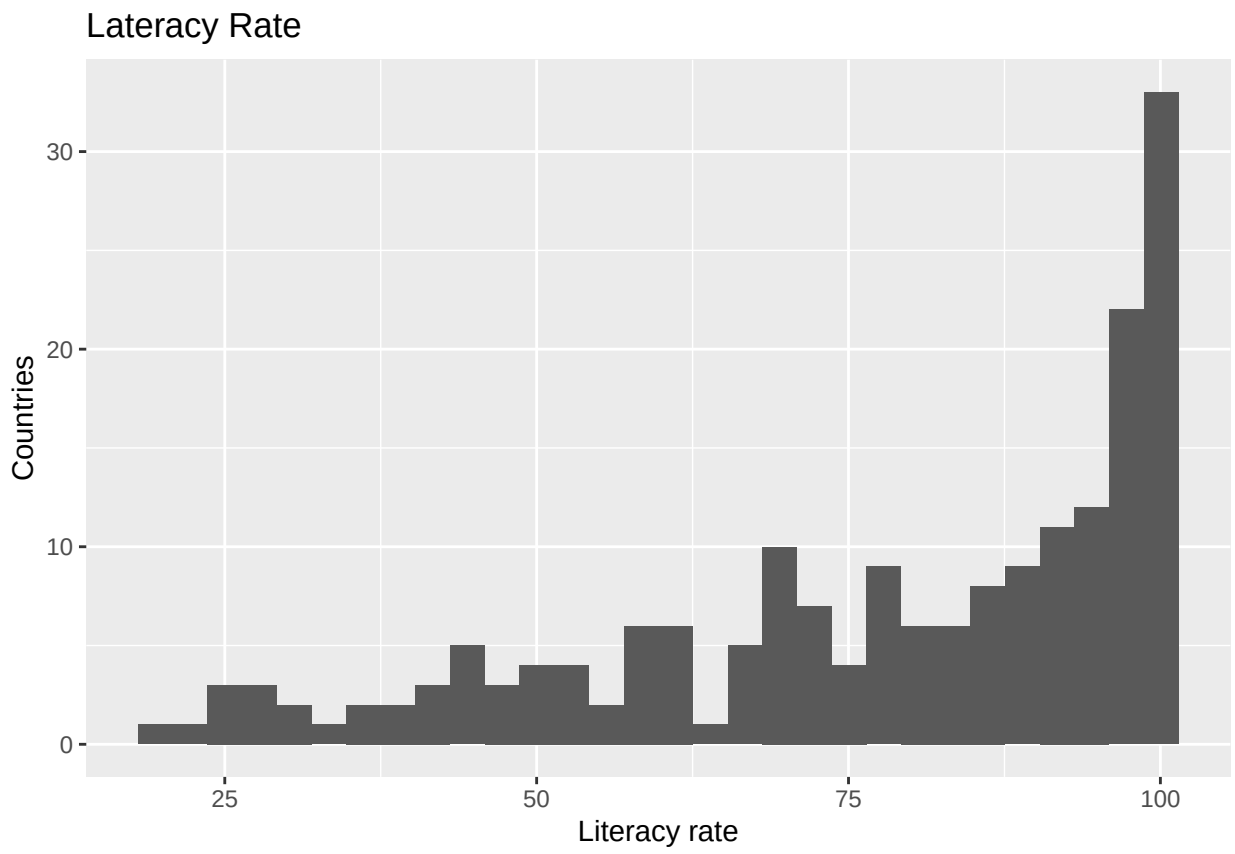
```
# Summary statistics for literacy rate
describe(j$lit_rate)
```

```
##    vars   n  mean    sd median trimmed   mad  min max range  skew kurtosis   se
## X1    1 191 77.53 21.95  84.72   80.43 21.17 19.4 100  80.6 -0.87    -0.32 1.59
```

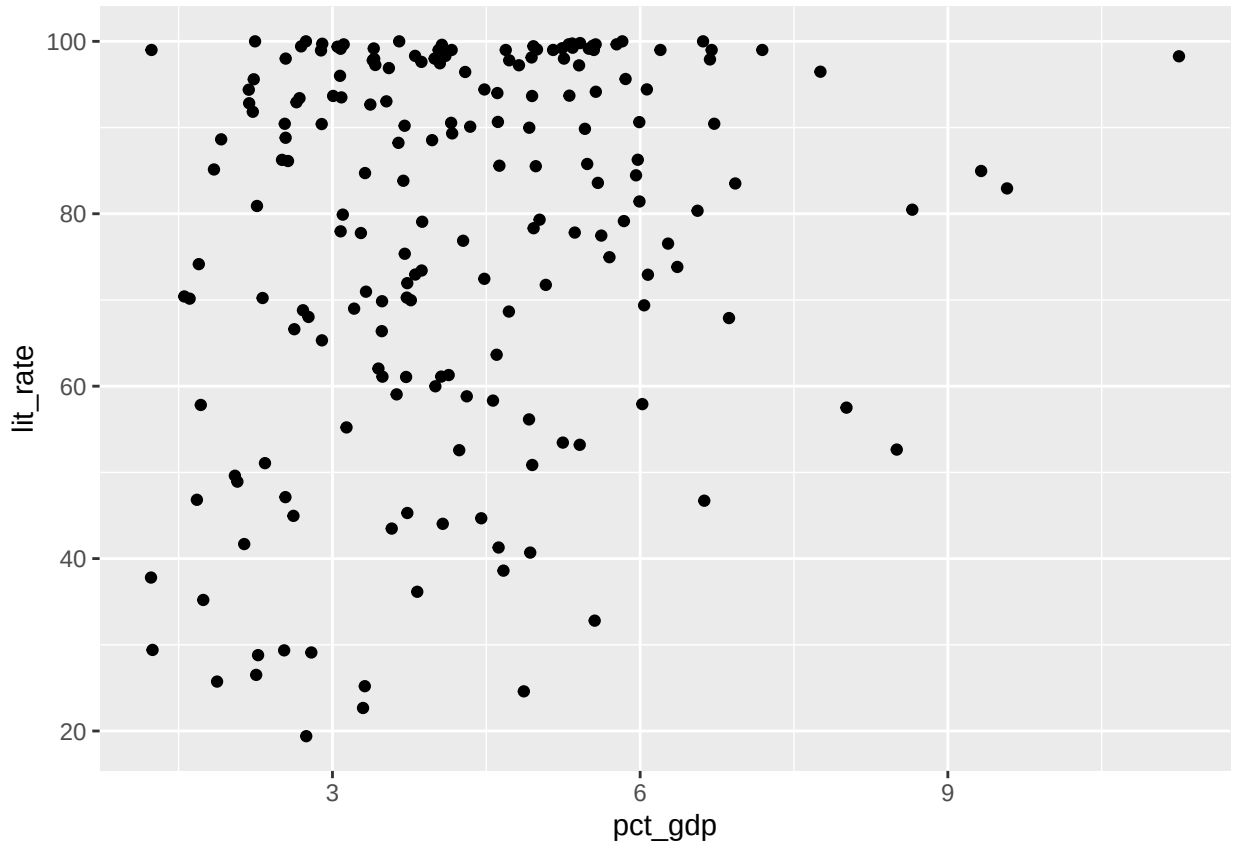
```
# Histogram of GDP education spending
j %>%
  ggplot() +
  geom_histogram(aes(x = pct_gdp), bins = 30) +
  xlab("% of GDP") + ylab("Countries") +
  ggtitle("Percentage of GDP Spent on Education")
```




```
# Histogram of literacy rate
j %>%
  ggplot() +
  geom_histogram(aes(x = lit_rate), bins = 30) +
  xlab("Literacy rate") + ylab("Countries") +
  ggtitle("Lateracy Rate")
```




```
# Scatter plot of literacy rate vs education spending
j %>% ggplot() +
  geom_point(aes(x = pct_gdp, y = lit_rate))
```


100
80
60
40
20
lit_rate
3
6
9
pct_gdp