# Oil, Democracy, and Development

Researchers have theorized that natural resources may have an inhibiting effect on the democratization process. Although there are multiple explanations as to why this might be the case, one hypothesis posits that governments in countries with large natural resource endowments (like oil) are able to fund their operations without taxing civilians. Since representation (and other democratic institutions) are a compromise offered by governments in exchange for tax revenue, resource-rich countries do not need to make this trade. In this exercise, we will not investigate causal effects of oil on democracy. Instead, we examine whether the association between oil and democracy is consistent with the aforementioned hypothesis.

This exercise is in part based on Michael L. Ross. (2001). 'Does Oil Hinder Democracy?' *World Politics*,53:3, pp.325-361.

The data set is in the csv file `resources.csv`. The names and descriptions of variables are:

| Name | Description |
|---|---|
| `cty_name` | Country name |
| `year` | Year |
| `logGDPcp` | Logged GDP per capita |
| `regime` | A measure of a country's level of democracy: -10 (authoritarian) to 10 (democratic) |
| `oil` | Amount of oil exports as a percentage of the country's GDP |
| `metal` | Amount of non-fuel mineral exports as a percentage of the country's GDP |
| `illit` | Percentage of the population that is illiterate |
| `life` | Life expectancy in the country |

## Question 1

Use scatterplots to examine the bivariate relationship between logged GDP per capita and life expectancy as well as between logged GDP per capita and illiteracy. Be sure to add informative axis labels. Also, compute the correlation separately for each bivariate relationship. Briefly comment on the results. To remove missing data when applying the `cor` function, set `use` argument to `"complete.obs"`.

## Question 2

We focus on the following subset of the variables: `regime`, `oil`, `logGDPcp`, and `illit`. Remove observations that have missing values in any of these variables. Using the `scale()` function, scale these variables so that each variable has a mean of zero and a standard deviation of one. Fit the k-means clustering algorithm with two clusters. How many observations are assigned to each cluster? Using the original unstandardized data, compute the means of these variables in each cluster.

## Question 3

Using the clusters obtained above, modify the scatterplot between logged GDP per capita and illiteracy rate in the following manner. Use different colors for the clusters so that we can easily tell the cluster membership of each observation. In addition, make the size of each circle proportional to the `oil` variable so that oil-rich countries stand out. Briefly comment on the results.

## Question 4

Repeat the previous two questions but this time with three clusters instead of two. How are the results different? Which clustering model would you prefer and why?