Session 3: Practice Problems

Andrew McCormack

2018-11-13

Review Questions

Problem 1

1) Using the dplyr functions from last week, create a new variable, gdp, in the civil war dataset by multiplying the existing GDP per capita and population variables together. Assign the modified dataset to a new object called cw3.
2) Next, filter out all countries that are not oil producers and then select only the gdp variable along with the country and year columns from the dataset.
3) Try to perform the same set of operations from above, this time by "piping" each operation together (%>%). Practice Problems

Practice problems

Problem 1

1) Using `dplyr`'s `filter()` function, filter the `civilwar` dataset to include observations from only one year. Save this as a new data frame.
2) Using `ggplot2`, create a scatter plot with two numeric variables that you think might be correlated. Select your two variables from among the following: GDP per capita, population, ethnic fractionalization, mountainous terrain, and polity (democracy) scores.
3) Sometimes it is difficult to see if there is a relationship with a scatterplot alone. To get a better sense of the relationship between the two variables, overlay a linear regression line on top of the scatterplot you just created (using `geom_smooth(method = "lm")`).
4) Describe the relationship based on what your graph tells you. Is it positive, negative, is there no relationship?

Problem 2

Let's visualize the trend of democracy over time for countries that export oil and countries that do not export oil.

1) To do this you first need to get the mean value of democracy for both groups (oil exporting and non-oil exporting countries) by year (`group_by()` and `summarize()` will come in handy here).
2) Select the appropriate `geom` to create a trend line and make sure that `ggplot2` knows to plot two separate lines. You will need to specify the variable `oil` as a factor (`factor(oil)`) inside `ggplot2` so that `ggplot2` knows these are categories and not continuous numeric values
3) What can we say about the evolution of democracy in these two countries? Do they follow a similar trend? Is one group more democratic than the other?

Problem 3

Let's compare the populations of the different regions across two different years.

1) First, restrict the data to include only two years (for example, 1960 and 1998). Next, get the sum of each region's population separately for both years (use `sum()` instead of `mean()` in the `summarise()` function). You will want to use `group_by()` and `summarise()` to do this.

2) Use the appropriate `geom` to create a bar/column chart where each region has two bars (one for each time period).

3) Flip the `x` and `y` axes so the region names do not overlap. Also, give the axes informative names (i.e. not just the default, which is the variable name), and title the plot.