SPEC Lab REU R Resources: Data Management 3

Alix Ziff and Miriam Barnum

Summer 2021

Data Management for Visualization

In this module, we will continue working with our IDC powersharing data to hone our data management skillset. We will use the packages `dplyr` and `countrycode` to work with country-year data. While group_by() and summarise() are introduced in Data Management II, this training goes more in depth.

```
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
```

```
library(countrycode)
```

```
idc_controls <- readRDS("IDC_analysis_master_MB_20210414.rds")
```

##Collapsing Country-Year Data ###univariate First, we want to collapse the country-year data down to global averages. In other words, we need to `groupby(year)`. We'll start by focusing on a single variable: resource rents. *Helpful Hint: make sure you save your steps as an object*

```
rents_global <- idc_controls %>%
  group_by(year) %>%
  summarise(natresource_rents_mean = mean(natresource_rents_WDI, na.rm = T))
View(rents_global)
```

###multivariate Let's get global averages for multiple variables at once. We may want to compare resource rents with trade. To take a look, we'll do the same thing as above, but will use the `summarise_at()` function to bring in our second variable. We'll call it trade_rents.

```
trade_rents_global <- idc_controls %>%
  group_by(year) %>%
  summarise_at(vars(natresource_rents_WDI, trade_WDI), mean, na.rm = T)
View(trade_rents_global)
```

As social scientists, we know that we can expect lots of regional variation. This is especially likely when looking at things like resources and trade. In order to look at our regional averages, we need to `mutate()` our data and a region variable and then pipe together our code.

```
rents_regional <- idc_controls %>%
  mutate(region = countrycode(gwno, "gwn", "region")) %>%
  group_by(region, year) %>%
  summarise(natresource_rents_mean = mean(natresource_rents_WDI, na.rm = T))
View(rents_regional)
```

Summarizing & Simplifying

###univariate Let's try to simplify our data a bit so we can identify broad patterns. We'll start by collapsing our country-year data on resource rents down to country-decade data. Then, we'll take a look at our averages, minimum, maximum, and median. *Helpful Hint* we'll add a decade variable by mutating our dataset and subtracting ten from our start date.

```
rents_decade <- idc_controls %>%

  # add decade variable (year %% 10 gives us the last digit, then we subtract from the year to get the
  mutate(decade = year - year %% 10) %>%

  group_by(gwno, decade) %>%
  dplyr::summarise(rents_mean = mean(natresource_rents_WDI, na.rm = T),
                   rents_med = median(natresource_rents_WDI, na.rm = T),
                   rents_max = max(natresource_rents_WDI, na.rm = T),
                   rents_min = min(natresource_rents_WDI, na.rm = T))
```

###multivariate We'll do the same thing with our global_rents data on trade and resource rents.

```
trade_rents_decade <- idc_controls %>%
  mutate(decade = year - year %% 10) %>%
  group_by(gwno, decade) %>%
  summarise_at(vars(natresource_rents_WDI, trade_WDI), tibble::lst(mean, median, max, min), na.rm = T)
```

##before we can make pretty pictures... we prep! ###making maps In order to make map figures, we need a single value per country. To do this, we'll pull out a single cross-sectional year either using base R or dplyr.

```
#base R
idc_2006 <- idc_controls[idc_controls$year == 2006,]
#dplyr
idc_2006 <- idc_controls %>%
  filter(year == 2006)
```

We may also want a dataset that includes the annual values for some individual countries, one region, and the global average. We already have our global averages, we just need to add a region name to it.

```
trade_rents_global$region <- "Global"
```

Say we want to compare countries in North America. We'll first need to get regional averages by continent and then filter so we just keep the Americas.

```
trade_rents_americas <- idc_controls %>%
  mutate(region = countrycode(gwno, "gwn", "continent")) %>%
  group_by(region, year) %>%
  summarise_at(vars(natresource_rents_WDI, trade_WDI), mean, na.rm = T)
  filter(region == "Americas")
```

Next, we want to pull out annual data for the US, Canada, and Mexico. Since variables should have the same names across dataframes we'll rbind them.

```
idc_sub <- idc_controls %>%
  filter(country %in% c("United States of America", "Canada", "Mexico")) %>%
  select(region = country, year, natresource_rents_WDI, trade_WDI)
```

Let's finish by binding it all into one dataframe.

```
df <- idc_sub %>%
  bind_rows(trade_rents_americas) %>%
  bind_rows(trade_rents_global)
```

And now we have a dataframe prepared for visualization. In this case, we could make a nice line plot with year on teh X axis and either natural resource rents or trade volume on the Y axis. We could plot,for example, how trade volume has evolved over time in the americas compared to the world overall.