Choosing Data Verbs

What's your input and your intended output?

Questions about your input and output data frames:

- For the input data table(s):
  1. What is the meaning of a case?
  2. How many cases are there?
  3. Is there more than one input data table?
- For the output data table:
  1. What is the meaning of a case?
  2. How many cases are there?
  3. Are there more cases in the input than the output?

Choosing a Data Verb

Walk through these questions. Stop when you hit the first "yes" answer.

1. Are there two input tables? Use `join()`

   Otherwise ...

2. Is the "physical" meaning of a case different from input to output? Example: cases are people in input, but cases are age groups in the output. Use `group_by()` and `summarise()`

   Otherwise ...

3. Are there fewer cases in the output than in the input? Use `filter()` or `head()`, `sample_n()`, etc.

   1. Does the output have the first few cases in the input? Use `head()` (or `tail()` for the last few cases).

      Otherwise ...

   2. Are the cases in the output randomly selected from the input? Use `sample_n()`

      Otherwise ...

   3. Are the cases in the output a subset from the input of cases that meet one or more criteria? Use `filter()`

   Otherwise ...

4. There is a variable in the output that is not in the input? Example, there's population and area in the input. The output includes population density. Use `mutate()`

   Otherwise ...

5. Are there some variables in the input that are not in the output? Use `select()`

   Otherwise ...

6. Are the cases are the same in the input and the output, but in a different order. Use `arrange()`

   Otherwise

7. It sounds like the input is the same as the output. If so, you don't need any data verb. If you don't think this is compelling, Start over and reconsider the above questions.

Adverbs

The arguments to a data verb include a data table and one or more "adverbs" that describe how the verb is to operate.

Join

There are two data-verb arguments. No adverbs are needed, unless you want to specify what cases to keep. See `merge()` in that situation.

Arrange

The quantity or quantities by which to arrange the cases.

```
HappinessIndex %>% arrange( gdpPerCapita )
HappinessIndex %>% arrange( region, gdpPerCapita )
# Sort by region, within each region break ties by GDP.

# Descending order, pipe the quantity to `desc()`
HappinessIndex %>% arrange( region, gdpPerCapita %>% desc() )
```