Five operations

Prof. Dr. Nicolas Meseth

Uncovering the value in the data

Data is the new oil, at least according to the mathematician Clive Humby:

> "Data is the new oil. Like oil, data is valuable, but if unrefined, it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity. So, must data be broken down, analysed for it to have value."

If we take this analogy seriously, the data, like oil, needs to be refined to turn it into something of value. Two important tools for refining data into a valuable output are *data transformation* and *data visualization*, both of which are the main focus of this book. In this part of the book, we first need to learn how to transform data from one form into another, so that we can apply visualization later on.

To master data transformation, we need to learn how to perform the following operations. We always start with a given data frame that we want to change into something else. In doing that, we typically want to …

1. … remove variables we don't currently need (or specify those we **do** need)
2. … remove any records we don't currently need (or specify those we **do** need)
3. … add new variables we need, but that don't exist yet
4. … summarize many records into one or a few numbers
5. … change the order of the records

The goal of the following chapters is to introduce means to perform theses five operations with R.

A helper in data transformation

To better understand what a transformation step does to our original data, there is a package called `{tidylog}` to help us. When the package is loaded, it overrides some of the `{dplyr}` functions and adds an extra output to the console. The output depends on the particular function, but in general, it gives us information about:

- How many columns where dropped by a `select` command
- How many rows where dropped by a `filter` command

```r
library(tidylog)
```