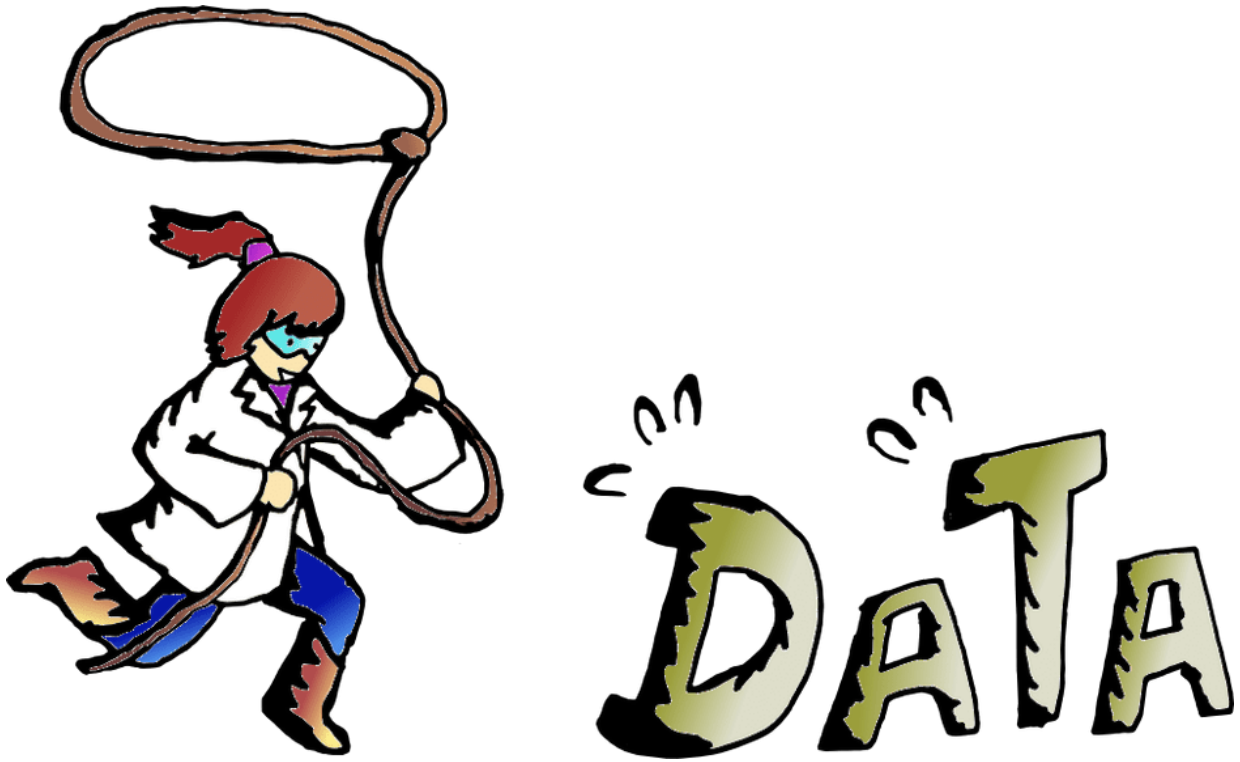
Data wrangling/manipulation



The tidyverse and dplyr

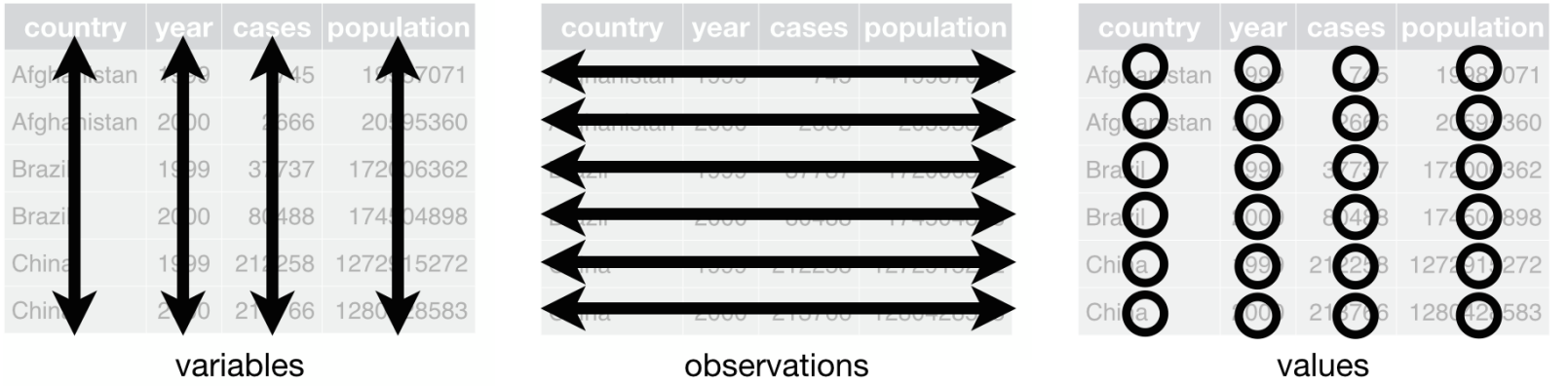
The ‘tidyverse’

The tidyverse is set of R packages that operate ‘tidy data’

- i.e., that operate on data frames (or tibbles)

Tidy data is data where:

- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell



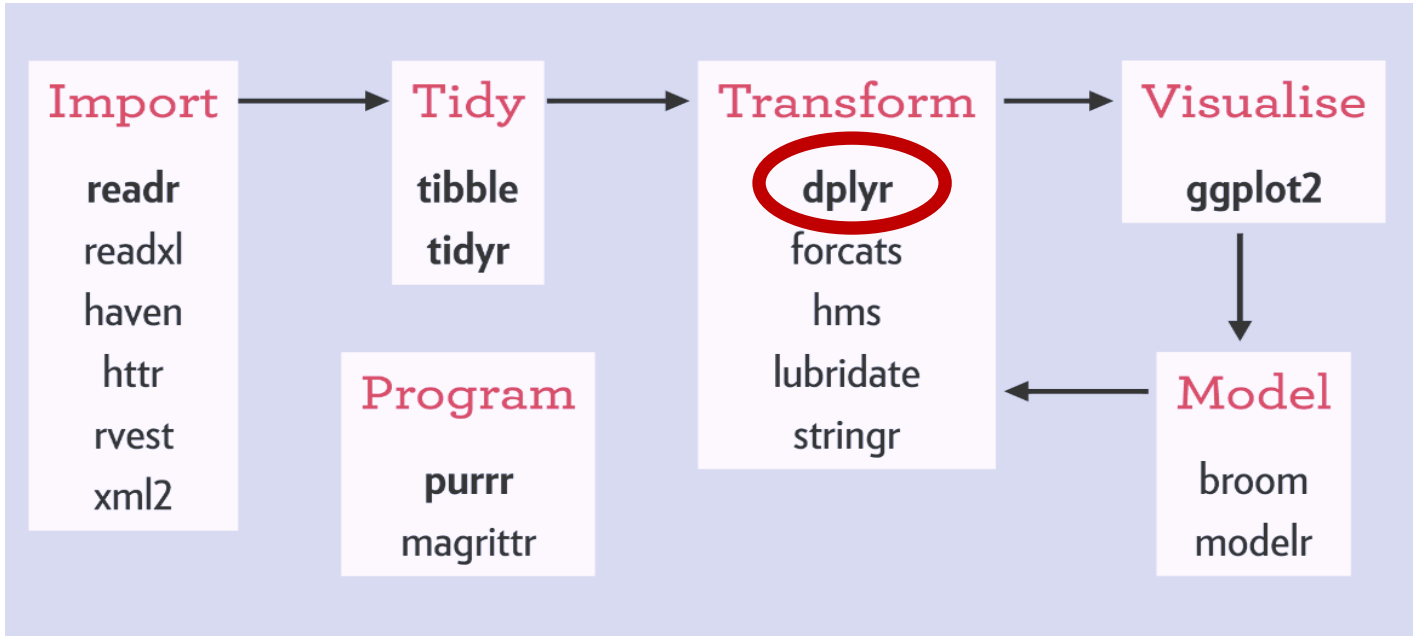
Messy data…

What would be an example of data that is not tidy?

The ‘tidyverse’

The packages share a common design philosophy

- Most written by Hadley Wickham



dplyr: A grammar for data wrangling

Grammar: a set of components that can be combined to achieve a goal

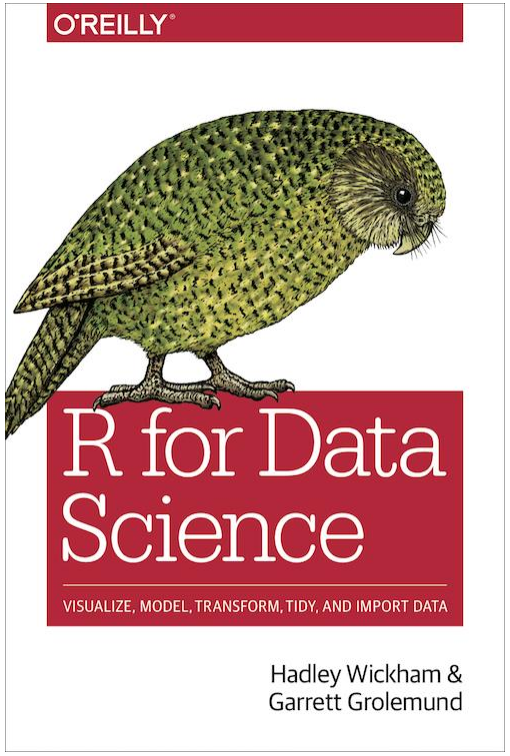
dplyr is a package that has a set of verbs that are useful for transformations data:

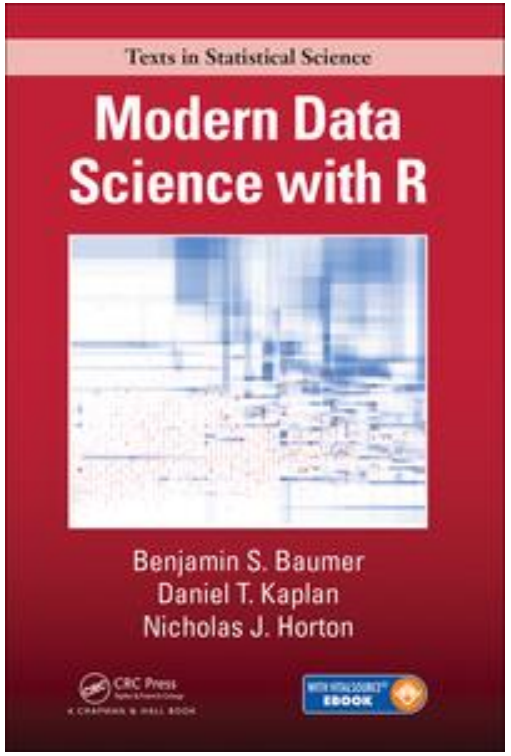
1. filter()
2. select()
3. mutate()
4. arrange()
5. group_by()
6. summarize()

All these function **take a data frame** and other arguments and **return a data frame**

```
> library(dplyr)   # load the dplyr package
```

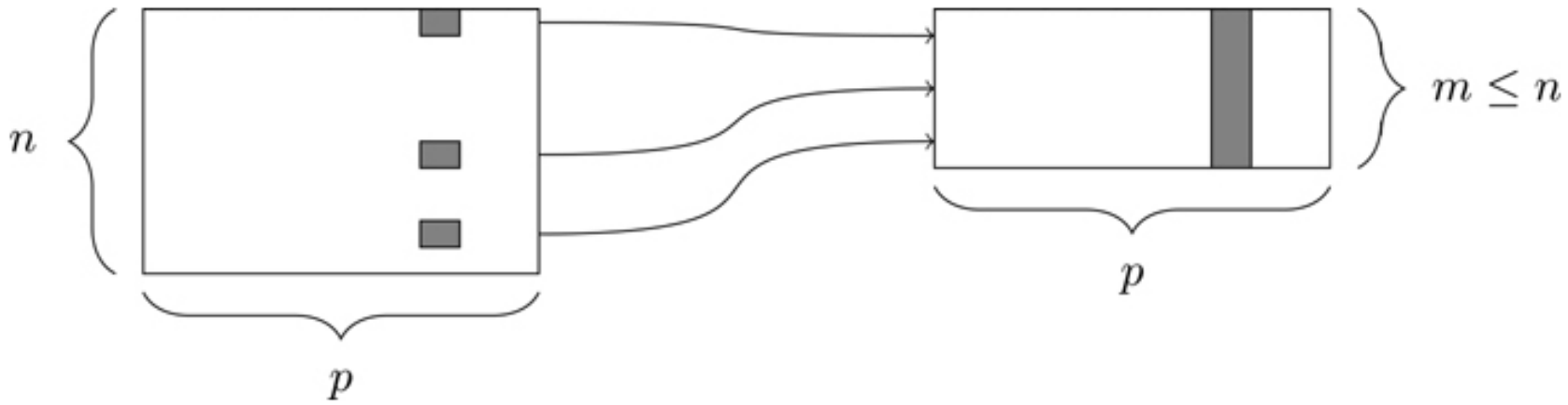
Quick overview of the dplyr functions





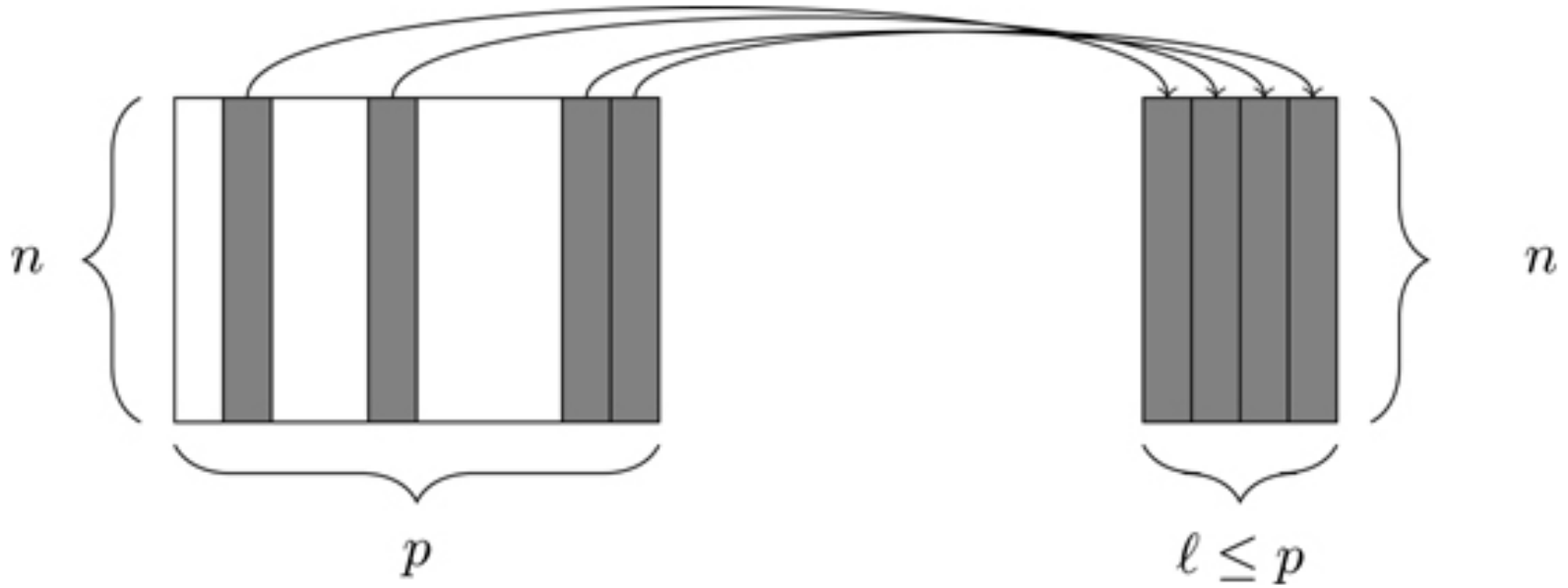
1. filter()

The filter() function allows you to select a subset of rows in data frame



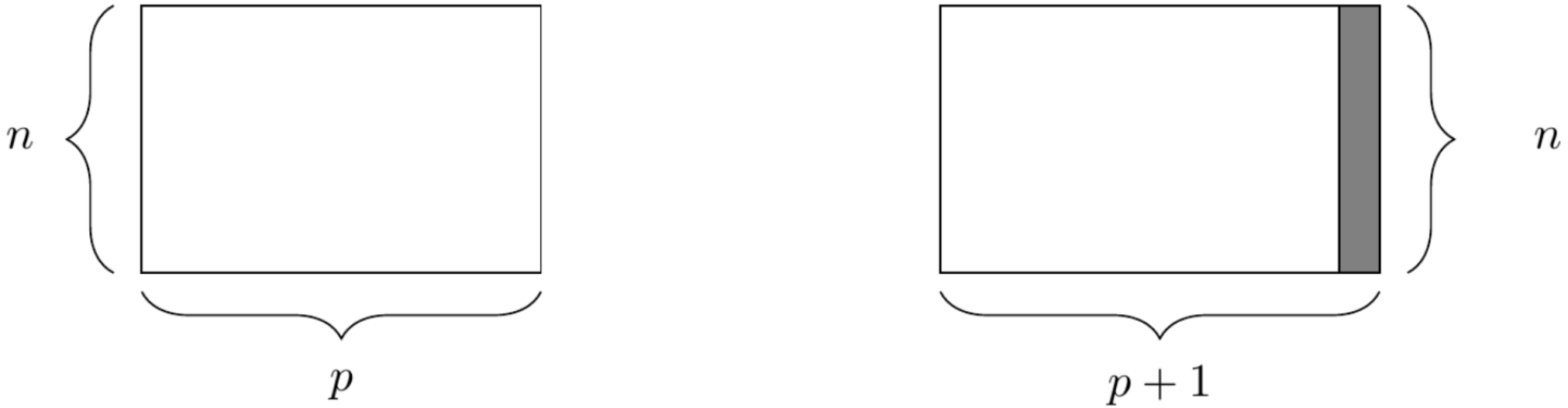
2. select()

The select() function allows you to select a subset of columns



3. mutate()

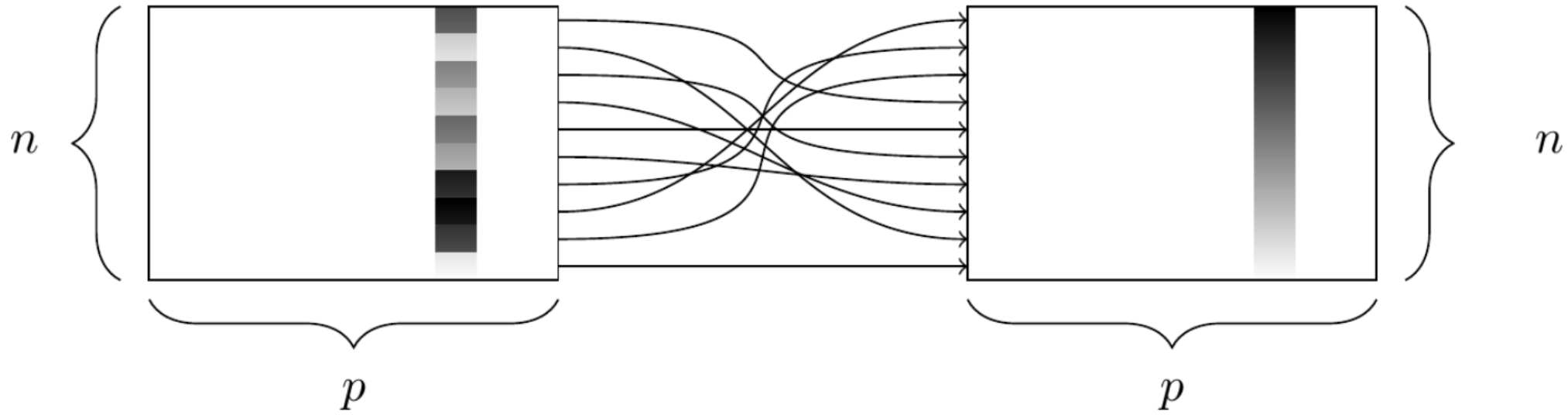
The mutate() function allows you to create new columns that are functions of existing columns



4. arrange()

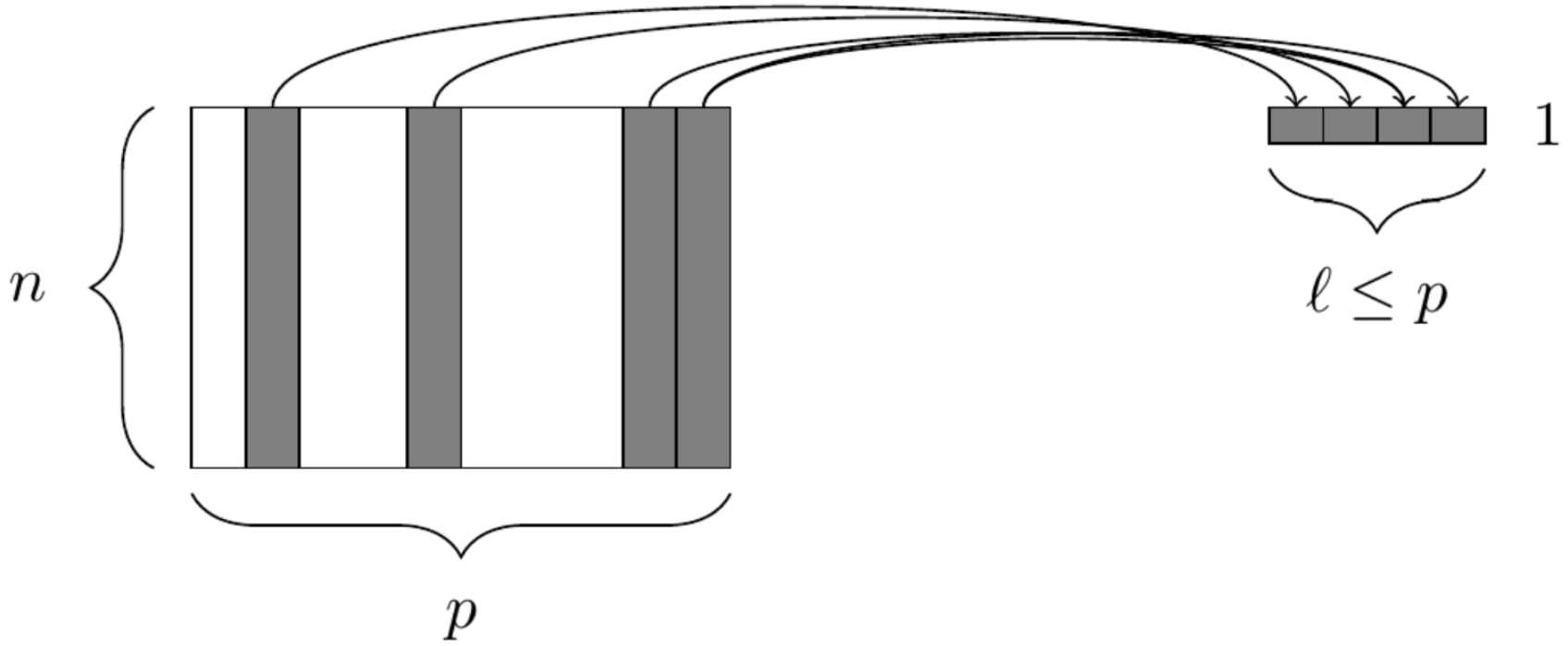
The arrange() function arranges the rows based values in a column

- arrange(desc()) arranges from largest to smallest



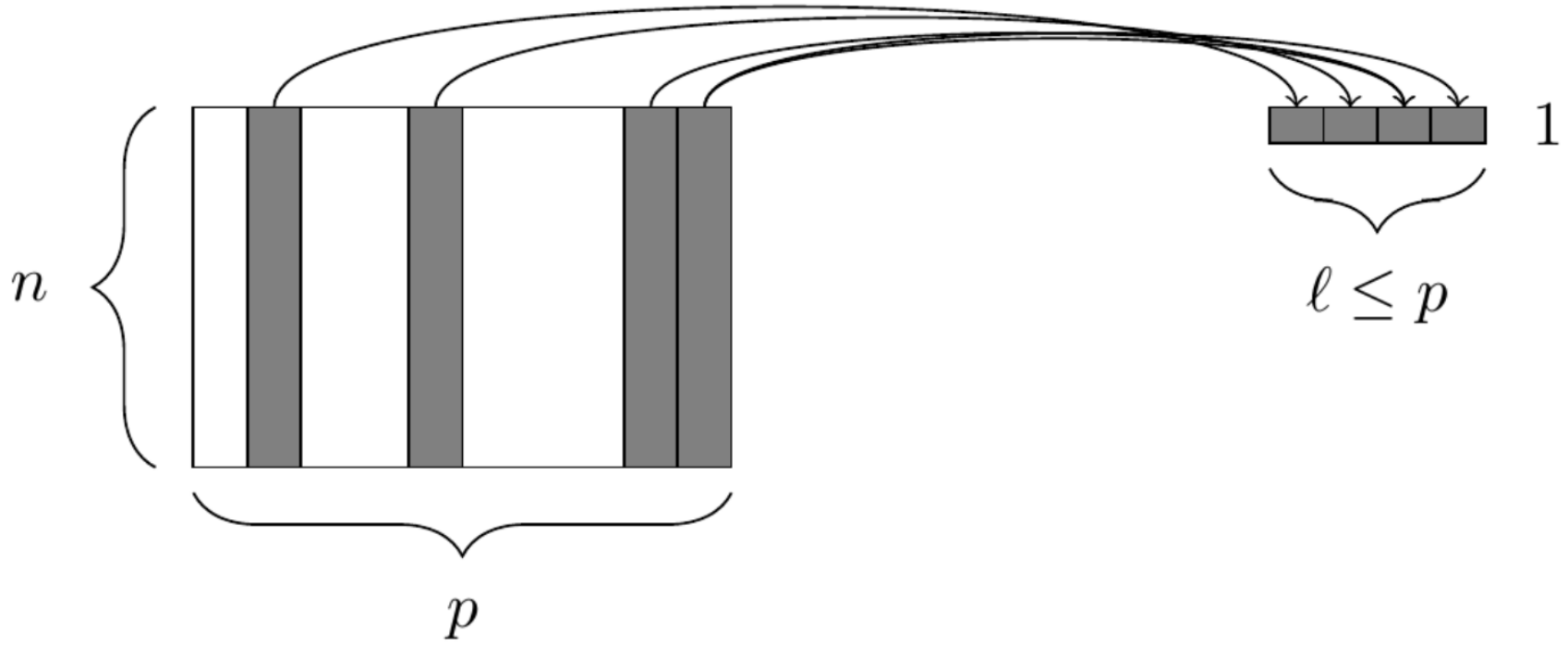
5. summarize()

The summarize() function reduces values in many rows into single values



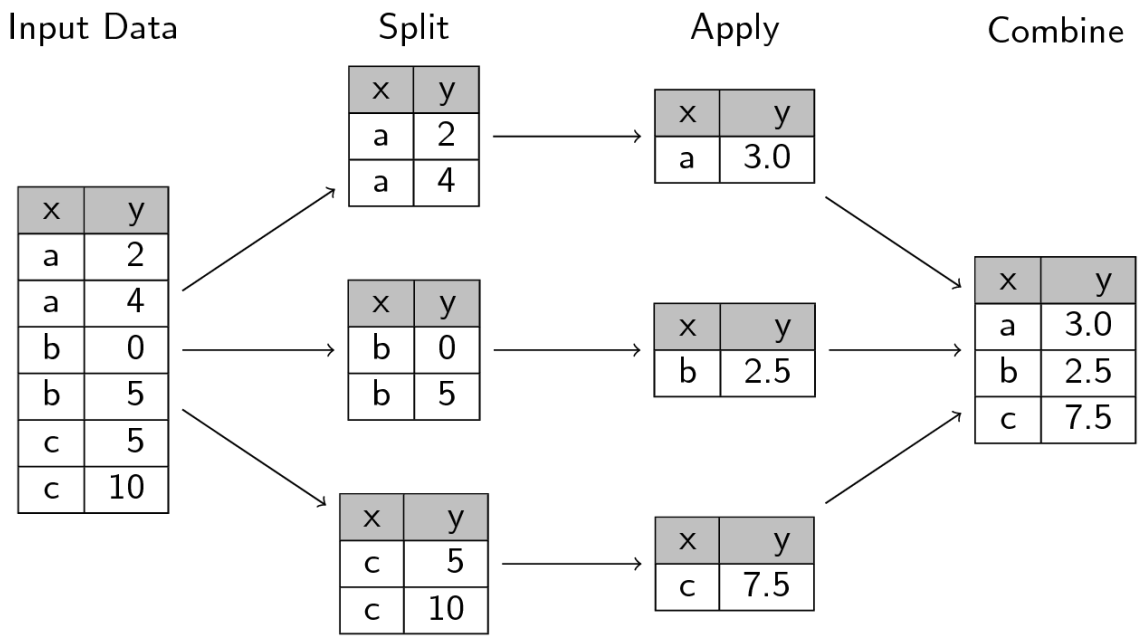
6. The group_by() function

The group_by() function groups variables for future operations



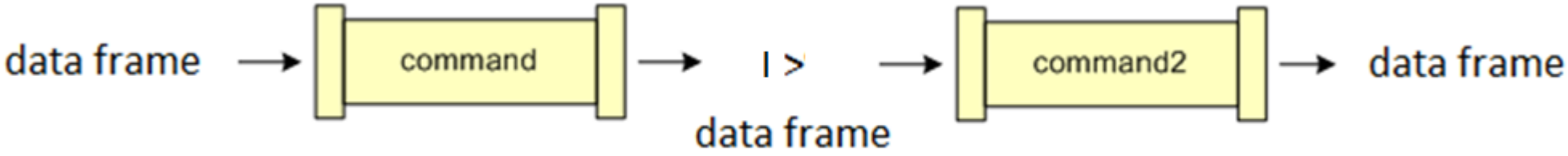
6. The group_by() function

group_by: split, apply, combine



The pipe operator

The pipe operator |> allows us to chain commands together



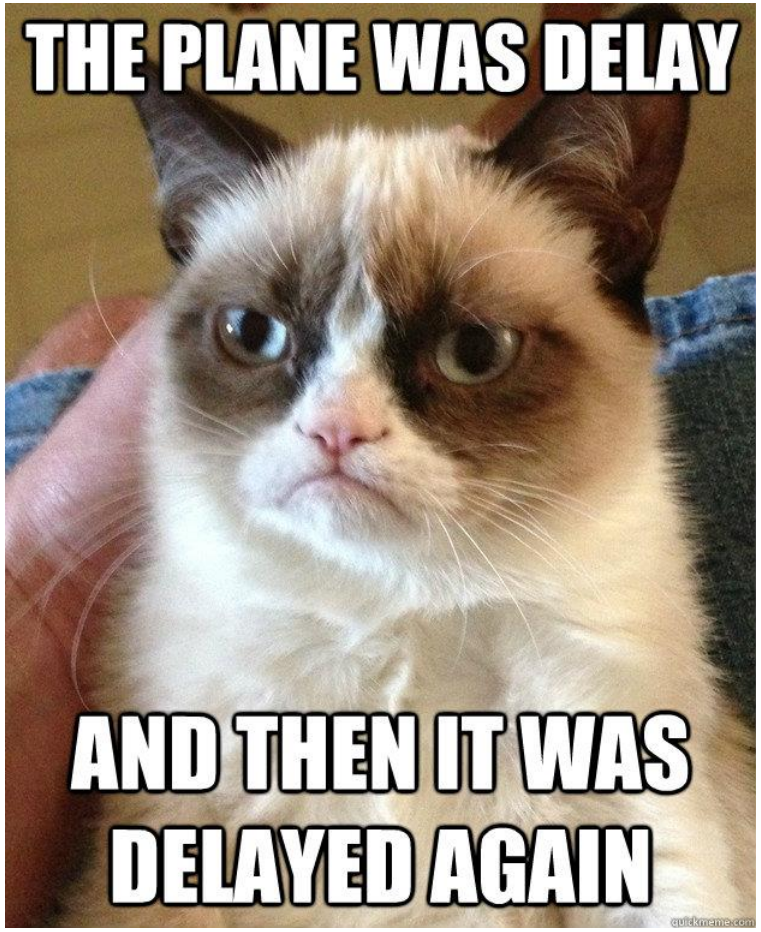
Let’s try it out!

Exercises: Flight delays

If there is time, let's try some exercises to practice using dplyr

```
> library("nycflights13")
> data(flights)
```

Exercises: Flight delays



Steps:

1. What result do I want?

2. What steps can I take to get the result?

3. How can I implement these steps using dplyr?