

Data transformation with dplyr

ds4owd - data science for openwashdata

Lars Schöbitz

2023-11-14

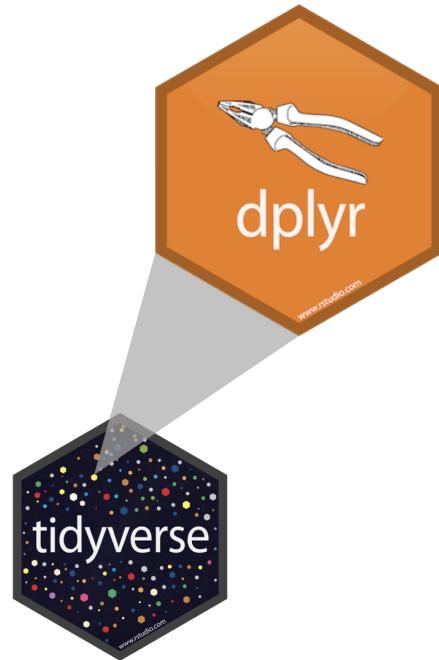
Learning Objectives (for this week)

1. Learners can apply ten functions from the dplyr R Package to generate a subset of data for use in a table or plot.

Data wrangling with dplyr

A grammar of data wrangling...

... based on the concepts of functions as verbs that manipulate data frames



- **select**: pick columns by name
- **arrange**: reorder rows
- **filter**: pick rows matching criteria
- **relocate**: changes the order of the columns
- **mutate**: add new variables
- **summarise**: reduce variables to values
- **group_by**: for grouped operations
- ... (many more)

dplyr rules

Rules of `dplyr` functions:

- First argument is always a data frame
- Subsequent arguments say what to do with that data frame
- Always return a data frame
- Don't modify in place

Functions & Arguments

```
1 library(dplyr)  
2  
3 filter(.data = gapminder,  
4         year == 2007)
```

- Function: `filter()`
- Argument: `.data =`
- Arguments following: `year == 2007` What to do with the data

Objects

```
1 library(dplyr)  
2  
3 gapminder_2007 <- filter(.data = gapminder,  
4                               year == 2007)
```

- Function: `filter()`
- Argument: `.data =`
- Arguments following: `year == 2007` What to do with the data
- Data (Object): `gapminder_2007`

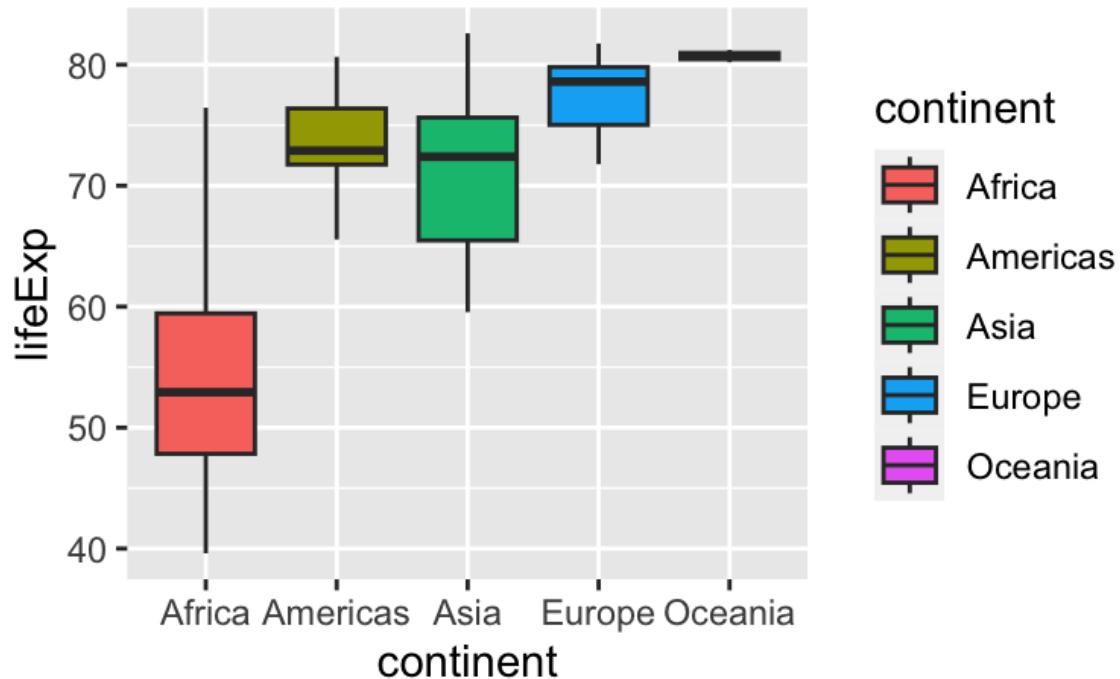
Operators

```
1 library(dplyr)  
2  
3 gapminder_2007 <- gapminder |>  
4   filter(year == 2007)
```

- Function: `filter()`
- Argument: `.data =`
- Arguments following: `year == 2007` What to do with the data
- Data (Object): `gapminder_2007`
- Assignment operator: `<-`
- Pipe operator: `|>`

Plot

```
1 library(dplyr)
2
3 gapminder_2007 <- gapminder |>
4   filter(year == 2007)
5
6 ggplot(data = gapminder_2007,
7         mapping = aes(x = continent,
8                         y = lifeExp,
9                         fill = continent)) +
10    geom_boxplot(outlier.shape = NA)
```



Our turn: SDG 6.2.1

Data

```
1 head(sanitation)
```

name	iso3	year	region_sdg	varname_short	varname_long	residence	percent
Afghanistan	AFG	2000	Central and Southern Asia	san_bas	basic sanitation services	national	21.9
Afghanistan	AFG	2000	Central and Southern Asia	san_bas	basic sanitation services	rural	19.3
Afghanistan	AFG	2000	Central and Southern Asia	san_bas	basic sanitation services	urban	30.9
Afghanistan	AFG	2000	Central and Southern Asia	san_lim	limited sanitation services	national	5.6
Afghanistan	AFG	2000	Central and Southern Asia	san_lim	limited sanitation services	rural	3.1
Afghanistan	AFG	2000	Central and Southern Asia	san_lim	limited sanitation services	urban	14.5

```
1 ncol(sanitation)
```

[1] 8

```
1 nrow(sanitation)
```

[1] 73710

Data

```
1 sanitation |>  
2 count(varname_short, varname_long)
```

varname_short	varname_long	n
san_bas	basic sanitation services	14742
san_lim	limited sanitation services	14742
san_od	no sanitation facilities	14742
san_sm	safely managed sanitation services	14742
san_unimp	unimproved sanitation facilities	14742

Our turn: md-03-exercises

1. Open [posit.cloud](#) in your browser (use your bookmark).
2. Open the [ds4owd workspace](#) for the course.
3. Click [Start](#) next to [md-03-exercises](#).
4. In the File Manager in the bottom right window, locate the [md-03a-data-transformation.qmd](#) file and click on it to open it in the top left window.

Your turn: md-03b-your-turn-filter.qmd

1. Open [posit.cloud](#) in your browser (use your bookmark).
2. Open the [ds4owd workspace](#) for the course.
3. In the File Manager in the bottom right window, locate the [md-03b-your-turn-filter.qmd](#) file and click on it to open it in the top left window.
4. Follow instructions in the file

Take a break

Please get up and move! Let your emails rest in peace.



R Terminology

```
1 library(dplyr)  
2  
3 sanitation_national_2020_sm <- sanitation |>  
4   filter(residence == "national",  
5         year == 2020,  
6         varname_short == "san_sm")
```

- Function: `filter()`
- Arguments following: `residence == "national"`, etc.
What to do with the data
- Data (Object): `sanitation_national_2020_sm`
- Assignment operator: `<-`
- Pipe operator: `|>`

Task 1.2

1. Use the `filter()` function to create a subset from the `sanitation` data containing urban and rural estimates for Nigeria.
2. Store the result as a new object in your environment with the name `sanitation_nigeria_urban_rural`

```
1 sanitation_nigeria_urban_rural <- sanitation |>  
2   filter(name == "Nigeria", residence != "national")
```

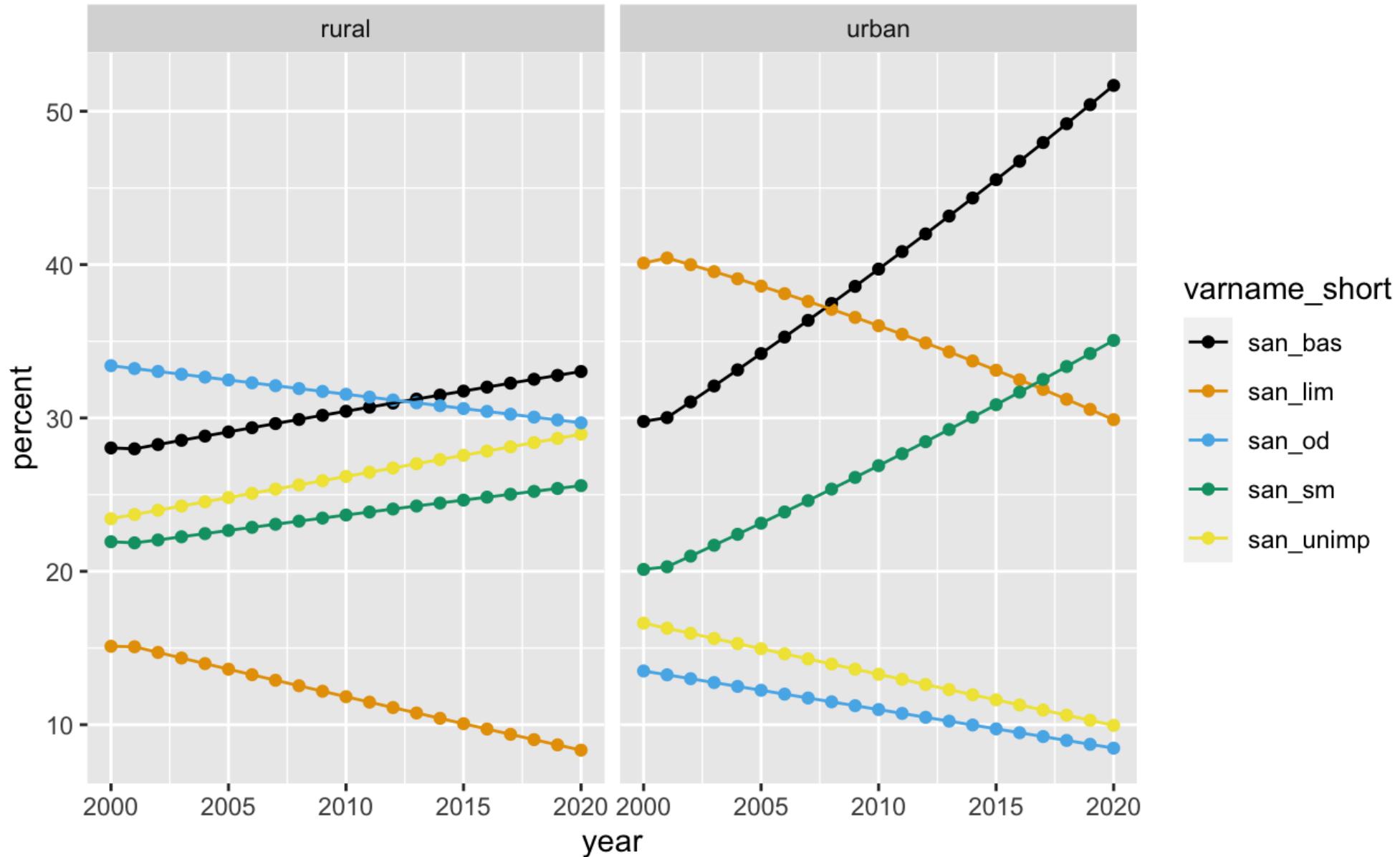
Task 1.3 - Connected scatterplot

Great for timeseries data 📅

1. Use the `ggplot()` function to create a connected scatterplot with `geom_point()` and `geom_line()` for the data you created in Task 1.2.
2. Use the `aes()` function to map the year variable to the x-axis, the `percent` variable to the y-axis, and the `varname_short` variable to color and group aesthetic.
3. Use `facet_wrap()` to create a separate plot urban and rural populations.
4. Change the colors using `scale_color_colorblind()`.

```
1 ggplot(data = sanitation_nigeria_urban_rural,  
2         mapping = aes(x = year,  
3                             y = percent,  
4                             group = varname_short,  
5                             color = varname_short)) +  
6   geom_point() +  
7   geom_line() +  
8   facet_wrap(~residence) +  
9   scale_color_colorblind()
```

Task 1.3 - Connected scatterplot



Our turn: back to md-03a-data-transformation.qmd

1. Open [posit.cloud](#) in your browser (use your bookmark).
2. Open the [ds4owd workspace](#) for the course.
3. Click [Start](#) next to [md-03-exercises](#).
4. In the File Manager in the bottom right window, locate the [md-03a-data-transformation.qmd](#) file and click on it to open it in the top left window.

Take a break

Please get up and move! Let your emails rest in peace.



Your turn: md-03c-your-turn-summarise.qmd

1. Open [posit.cloud](#) in your browser (use your bookmark).
2. Open the [ds4owd workspace](#) for the course.
3. In the File Manager in the bottom right window, locate the [md-03c-your-turn-summarise.qmd](#) file and click on it to open it in the top left window.
4. Follow instructions in the file

Homework assignments

module 3

Module 3 documentation

ds4owd-001.github.io/website/modules/md-03.html

Module 3

Data transformation with dplyr

⌚ Learning Objectives

1. Learners can apply ten functions from the dplyr R Package to generate a subset of data for use in a table or plot.

💻 Slides

[View slides in full screen](#) | [Download slides as PDF](#)

Data transformation with dplyr

ds4owd - data science for openwashdata

Lars Schöbitz

2023-11-14
✉️ ds4owd-001.github.io/website/

Homework due date

- Homework assignment due: Monday, November 20th
- Correction & feedback phase up to: Thursday, November 23rd

Wrap-up

Thanks! 🌻

Slides created via revealjs and Quarto:

<https://quarto.org/docs/presentations/revealjs/> Access slides as
[PDF on GitHub](#)

All material is licensed under [Creative Commons Attribution Share Alike 4.0 International.](#)

