THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Introduction to data wrangling in R

Matthew Curry & Tomasz Zając

Institute for Social Science Research

The University of Queensland

ARDC Summer School 2024 Day 1

Social Science Stream

Acknowledgment of **Country**

The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.

About the course

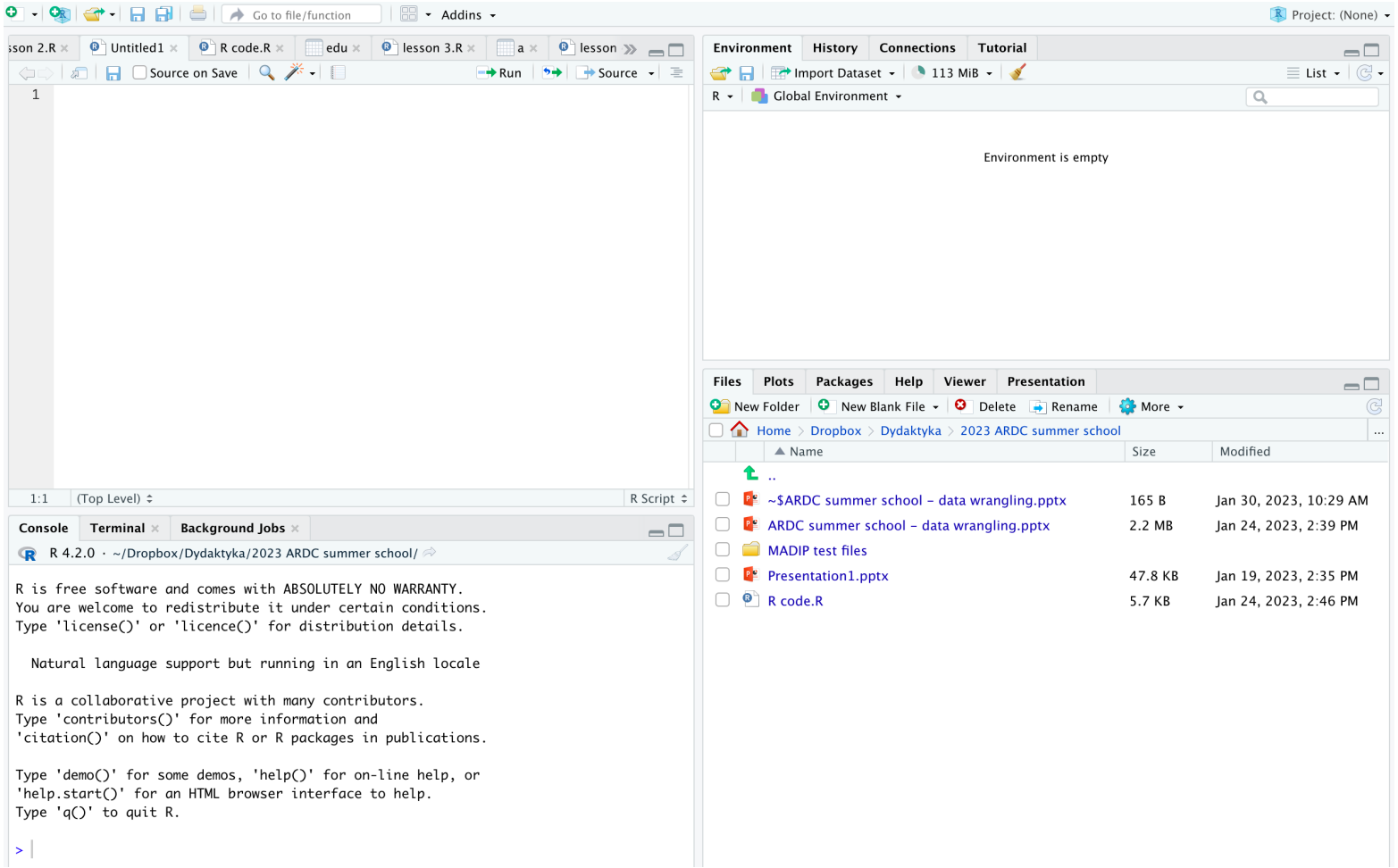
- Data manipulation methods in R
  - Subsetting data
  - Creating modifying variables
  - Merging datasets
  - Aggregating data

Why R?

- Powerful
- Designed for statistics and data science
- Free and open source
- Platform-independent
- Popular
- Great community support

R & RStudio

RStudio


Go to file/function
Addins
Project: (None)
Untitled1
R code.R
edu
lesson 3.R
a
lesson
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Environment
History
Connections
Tutorial
Import Dataset
113 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
New Folder
New Blank File
Delete
Rename
More
Home
Dropbox
Dydaktyka
2023 ARDC summer school
Name
Size
Modified
~$ARDC summer school – data wrangling.pptx
165 B
Jan 30, 2023, 10:29 AM
ARDC summer school – data wrangling.pptx
2.2 MB
Jan 24, 2023, 2:39 PM
MADIP test files
Presentation1.pptx
47.8 KB
Jan 19, 2023, 2:35 PM
R code.R
5.7 KB
Jan 24, 2023, 2:46 PM
Console
Terminal
Background Jobs
R 4.2.0 · ~/Dropbox/Dydaktyka/2023 ARDC summer school/
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>

Data types

- Booleans/logical (TRUE, FALSE)
- Numeric (1, 1.3, 5.9)
- Integer (1, 2, 10)
- Character – string (“a”, “abc”, “5”)

- Date
- Factor
- Polygons

- Missing: NA, NaN

Data objects

- Data.frame / tibble/ data.table
- Vector

…

- Matrix
- Array
- List

Some Useful Base R functions

- `<-` or `=`
- `view()`
- `head()`
- `tail()`
- `table()`
- `prop.table()`
- `summary()`

Open RStudio

"Melbournian hipster coding in R"

R packages

https://tidyverse.tidyverse.org

Loading data

Data format	Package	Command
csv	`base R`	`read.csv()` – special case of `read.table()`
	`readr`	`read_csv()` - special case of `read_delim()`
	`data.table`	`fread()`
	`arrow`	`read_csv_arrow()`
xls, xlsx	`readxl`	`read_excel()`
SPSS, Stata	`foreign`	`read.spss(), read.dta()`
	`haven`	`read_spss(), read_dta()`
rds	`base R`	`readRDS()`

Example data

Example datasets with random data:

File 1 – Data on higher education completions

File 2 – Data on income tax

Goal:

Compute the gender pay gap (GPG) for each broad field of education for graduates from 2011:

- Extract data on the 2011 cohort of graduates
- Create required variables (e.g. broad field of education), clean the data
- Extract tax data
- Transform tax data to the right format
- Merge both datasets
- Aggregate data and compute the GPG

Working with R packages and example data

"R packages being produced"

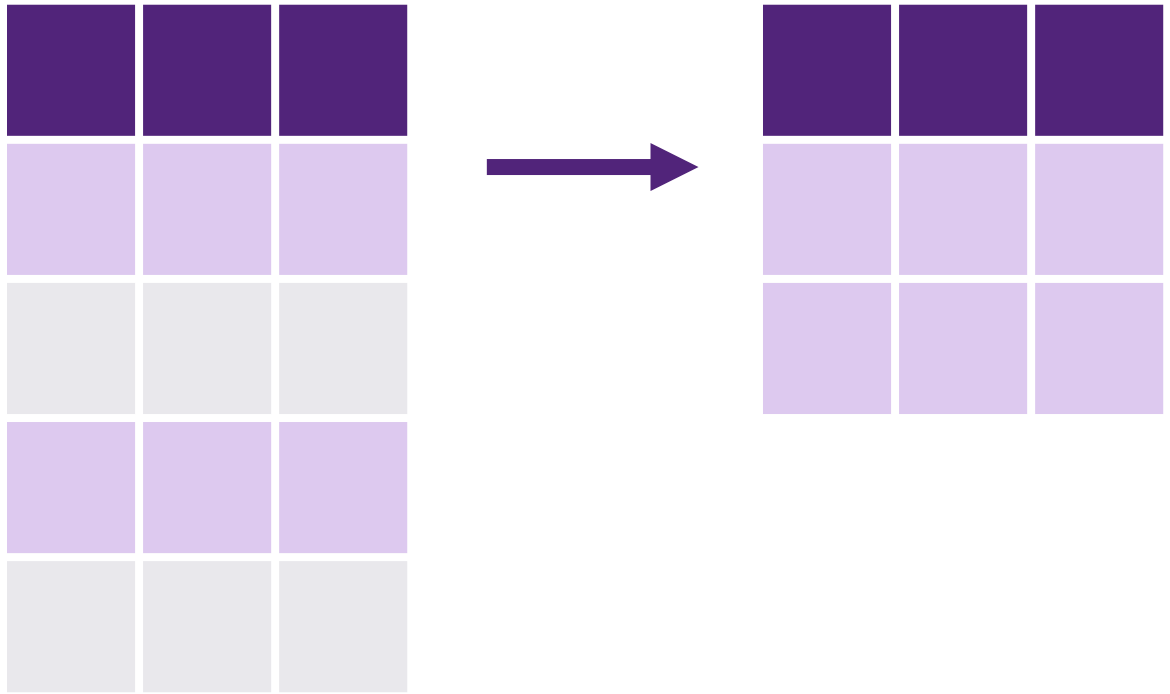
Filter rows/ extract cases

`filter(data, condition)`

Conditions:

- `x == y` equal to
- `x != y` is not equal to
- `x > y (>=, <=, <)` greater/smaller than
- `x %in% y` within
- `!is.na(x)` is missing

- `&` and
- `|` or



visualisations inspired by: https://github.com/rstudio/cheatsheets/blob/main/data-transformation.pdf

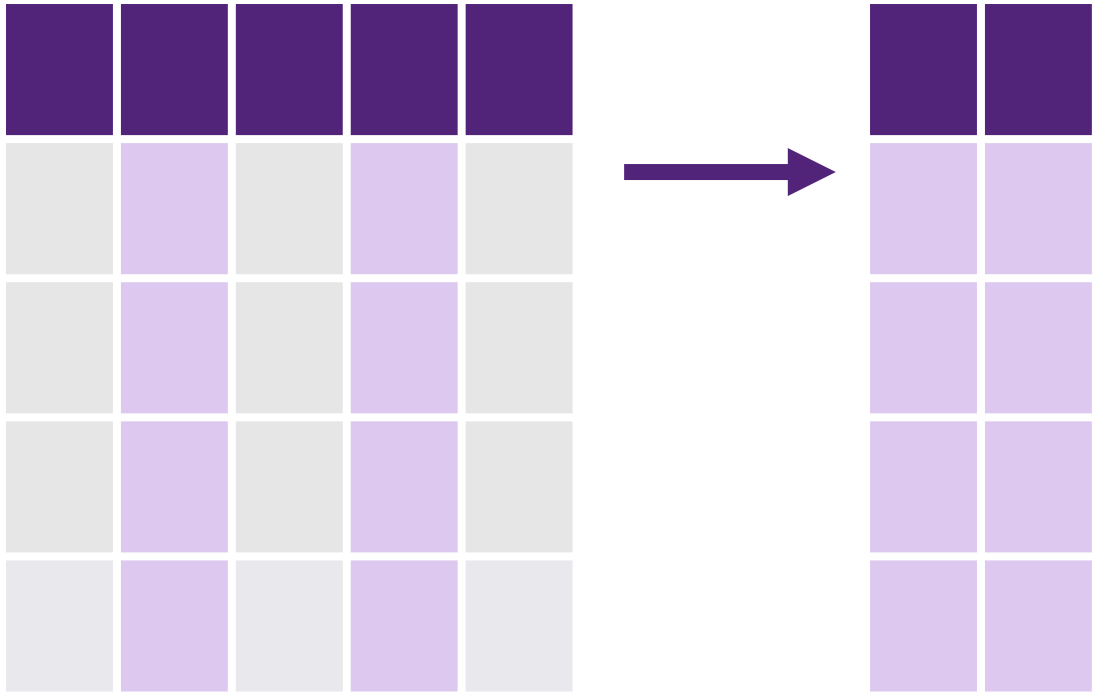
Select columns

```
select(data, columns)
```

tidy selection:

```
col1, col2, col3
col2:col5
-col3
col1 | col4:col8

starts_with()
ends_with()
contains()
where()
```



Chaining – Pipe operator

Requires `magritter` or `dplyr`

```
function4(function3(function2(function1(x, add_arg1), add_arg2), add_arg3), add_arg4)
```

```
x %>%
        function1(add_arg1) %>%
        function2(add_arg2) %>%
        function3(add_arg3) %>%
        function4(add_arg4)
```

Mutate – add/ modify variables

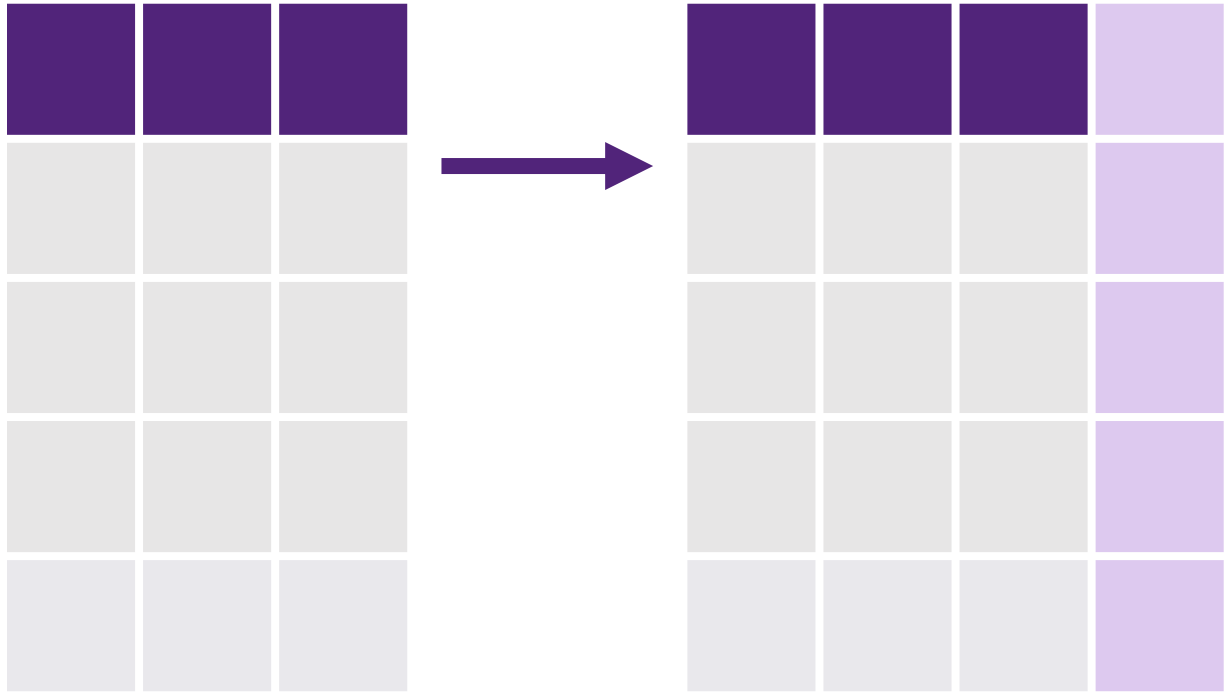
```
mutate(data, new_var = function(old_var))
```

`across()` – applies the same transformation to multiple columns

```
mutate(across(col1:col3, ~ .x^2)
```

```
mutate(across(where(is.numeric),
                    ~ .x^2))
```

```
mutate(across(where(is.numeric),
              list(sq = ~.x^2)))
```



Reshaping Data

`pivot_longer()`

wide → long transformations

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

`pivot_wider()`

long → wide transformations

Image source: https://www.garrickadenbuie.com/project/tidyexplain/#spread-and-gather

Merging data – mutating joins

x - left dataset

ID	V1	V2
1		
2		
3		

ID	V3	V4
1		
2		
4		

y - right dataset

`left_join(x,y)`

ID	V1	V2	V3	V4
1				
2				
3			NA	NA

`right_join(x,y)`

ID	V1	V2	V3	V4
1				
2				
4	NA	NA		

`inner_join(x,y)`

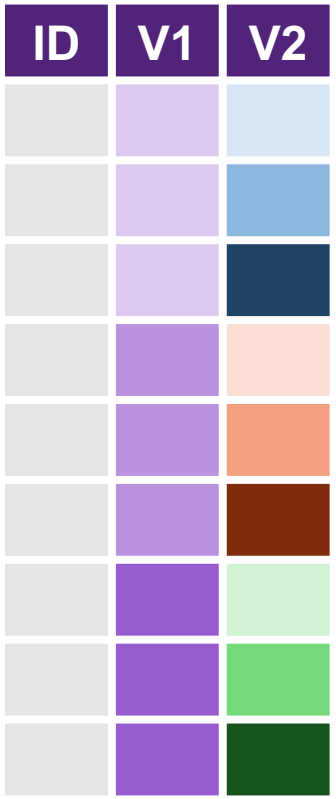
ID	V1	V2	V3	V4
1				
2				

`full_join(x,y)`

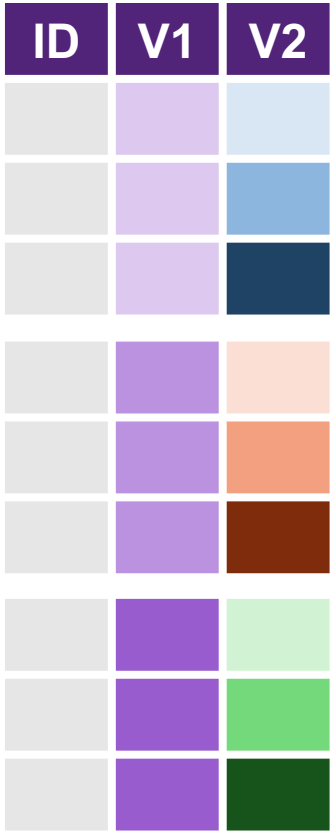
ID	V1	V2	V3	V4
1				
2				
3			NA	NA
4	NA	NA		

Collapsing data

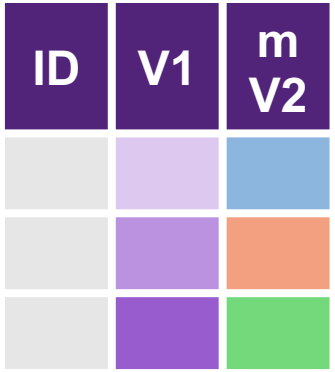
original data



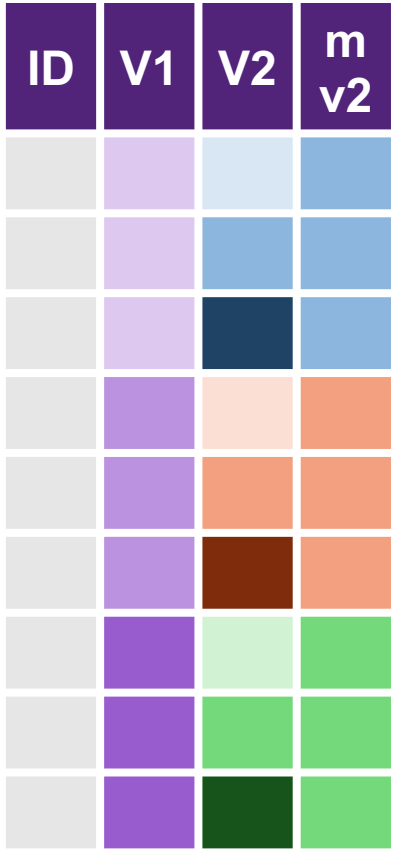
grouped data
`group_by()`



collapsed data
`summarise()`



aggregated value as new column - `mutate()`



Additional resources

https://github.com/AURIN-OFFICE/HASS_Summer_School

https://cran.r-project.org

https://posit.co/resources/cheatsheets/

https://stackoverflow.com

Google/ Bing/ DuckDuckGo/…

ChatGPT