Conditions & Dates & Tables

ds4owd - data science for openwashdata

Lars Schöbitz

2023-11-28

Learning Objectives (for this week)

1. Learners can apply functions from dplyr to change specific values of a variable in a dataframe.
2. Learners can apply functions from the lubridate package to convert dates not in ISO 8601 to the date class in R.
3. Learners can apply functions from knitr and gt package to display and cross-reference tables in HTML output reports.

Homework module 4

Part 1: Reading

1. Read Broman and Woo (2018): "Data organization in spreadsheets".
2. Chose two of the recommendations and come up with real-world examples or scenarios where the recommendations could be applied in your work.
3. Be prepared to share these examples and explain how the recommendations would improve your workflows. This will be in a class setting as part of small discussion group (max 3 people).

Your turn: Discuss the reading

> In discussion groups of 3, share your examples and discuss how the recommendations would improve your workflows.

10:00

Rules for variable names

- avoid spaces
- avoid special characters
- use consistent naming conventions (e.g. snake_case)

use	avoid
• ts	• Total Solids (g/L)
• users	• Number of users
• system	• System (pit latrine / septic tank)

Data dictionary / codebook

- variable descriptions belong into a data dictionary / codebook (not into variable names)
- data dictionaries / codebooks are separate files

variable_name	description
ts	Total solids in g/L.
users	Number of users per system.
system	Sanitation system in use at sample location (septic tank / pit latrine).

Conditional statements

dplyr functions `mutate()` & `case_when()`

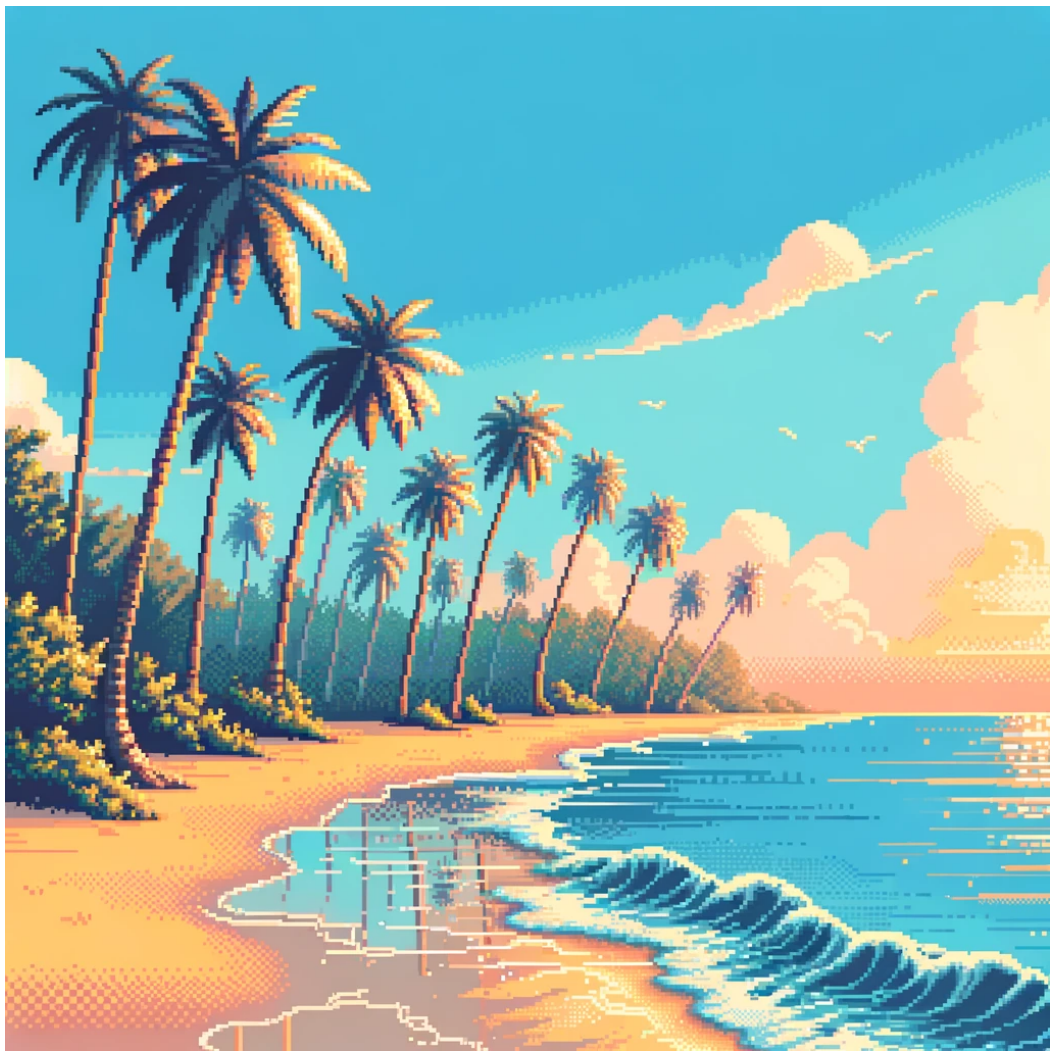
- `mutate()` adds new variables to a data frame
- `case_when()` is another form of an if-else statement
- combination of functions are used to create variables with new values or fix existing ones

My turn: Conditional statements

Sit back and enjoy!

Take a break

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



ds4owd-001.github.io/website/

10:00

Your turn: md-05-exercises - conditions

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-05a-conditions-your-turn.qmd` file and click on it to open it in the top left window.
4. Follow instructions in the file

20:00

Task 1

1. A mistake happened during data entry for sample id 16. Use `mutate()` and `case_when()` to change the ts value of 0.72 to 8.72.

```
sludge |>
    mutate(ts = case_when(
        ts ==  0.72 ~ 8.72,
        .default = ts
    ))
```

id	date	system	location	users	ts
1	2023-11-01	pit latrine	household	5	136.24
2	2023-11-01	pit latrine	household	7	102.45
3	2023-11-01	pit latrine	household	NA	57.02
4	2023-11-01	pit latrine	household	6	27.03
5	2023-11-01	pit latrine	household	12	97.27
6	2023-11-02	pit latrine	household	7	78.21
7	2023-11-02	septic tank	household	14	15.24

id	date	system	location	users	ts
8	2023-11-02	septic tank	household	4	29.39
9	2023-11-02	septic tank	household	10	64.22
10	2023-11-02	septic tank	household	12	8.01
11	2023-11-03	pit latrine	public toilet	50	11.24
12	2023-11-03	pit latrine	public toilet	32	84.05
13	2023-11-03	pit latrine	public toilet	41	55.92
14	2023-11-03	pit latrine	public toilet	160	15.32
15	2023-11-03	pit latrine	public toilet	20	22.65
16	2023-11-04	septic tank	public toilet	26	8.72
17	2023-11-04	septic tank	public toilet	91	43.92
18	2023-11-04	septic tank	public toilet	68	10.37
19	2023-11-04	septic tank	public toilet	112	23.21
20	2023-11-04	septic tank	public toilet	59	15.64

Task 2

1. Another mistake happened during data entry for sample id 6. Use `mutate()` and `case_when()` to change the system value of id 6 from “pit latrine” to “septic tank”.

```
sludge |>
    mutate(system = case_when(
        id ==  6 ~ "septic tank",
        .default = system
    ))
```

id	date	system	location	users	ts
1	2023-11-01	pit latrine	household	5	136.24
2	2023-11-01	pit latrine	household	7	102.45
3	2023-11-01	pit latrine	household	NA	57.02
4	2023-11-01	pit latrine	household	6	27.03
5	2023-11-01	pit latrine	household	12	97.27
6	2023-11-02	septic tank	household	7	78.21
7	2023-11-02	septic tank	household	14	15.24

id	date	system	location	users	ts
8	2023-11-02	septic tank	household	4	29.39
9	2023-11-02	septic tank	household	10	64.22
10	2023-11-02	septic tank	household	12	8.01
11	2023-11-03	pit latrine	public toilet	50	11.24
12	2023-11-03	pit latrine	public toilet	32	84.05
13	2023-11-03	pit latrine	public toilet	41	55.92
14	2023-11-03	pit latrine	public toilet	160	15.32
15	2023-11-03	pit latrine	public toilet	20	22.65
16	2023-11-04	septic tank	public toilet	26	0.72
17	2023-11-04	septic tank	public toilet	91	43.92
18	2023-11-04	septic tank	public toilet	68	10.37
19	2023-11-04	septic tank	public toilet	112	23.21
20	2023-11-04	septic tank	public toilet	59	15.64

Task 3 (stretch goal)

1. Add a new variable with the name `ts_cat` to the dataframe. that categorises sludge samples into low, medium and high solids content.Use `mutate()` and `case_when()` to create the new variable.

- samples with less than 15 g/L are categorised as low
- samples with 15 g/L to 50 g/L are categorised as medium
- samples with more than 50 g/L are categorised as high

Task 3 (stretch goal)

```
sludge |>
    mutate(ts_cat = case_when(
        ts < 15 ~ "low",
        ts >= 15 & ts <= 50 ~ "medium",
        ts > 50 ~ "high"
    ))
```

id	date	system	location	users	ts	ts_cat
1	2023-11-01	pit latrine	household	5	136.24	high
2	2023-11-01	pit latrine	household	7	102.45	high
3	2023-11-01	pit latrine	household	NA	57.02	high
4	2023-11-01	pit latrine	household	6	27.03	medium
5	2023-11-01	pit latrine	household	12	97.27	high
6	2023-11-02	pit latrine	household	7	78.21	high
7	2023-11-02	septic tank	household	14	15.24	medium
8	2023-11-02	septic tank	household	4	29.39	medium
9	2023-11-02	septic tank	household	10	64.22	high
10	2023-11-02	septic tank	household	12	8.01	low
11	2023-11-03	pit latrine	public toilet	50	11.24	low

id	date	system	location	users	ts	ts_cat
12	2023-11-03	pit latrine	public toilet	32	84.05	high
13	2023-11-03	pit latrine	public toilet	41	55.92	high
14	2023-11-03	pit latrine	public toilet	160	15.32	medium
15	2023-11-03	pit latrine	public toilet	20	22.65	medium
16	2023-11-04	septic tank	public toilet	26	0.72	low
17	2023-11-04	septic tank	public toilet	91	43.92	medium
18	2023-11-04	septic tank	public toilet	68	10.37	low
19	2023-11-04	septic tank	public toilet	112	23.21	medium
20	2023-11-04	septic tank	public toilet	59	15.64	medium

Task 3 (stretch goal)

```
sludge |>
    mutate(ts_cat = case_when(
        ts < 15 ~ "low",
        ts >= 15 & ts <= 50 ~ "medium",
        ts > 50 ~ "high"
    )) |>
  count(ts_cat)
```

ts_cat	n
high	8
low	4
medium	8

Dates and times

Dates and times in R

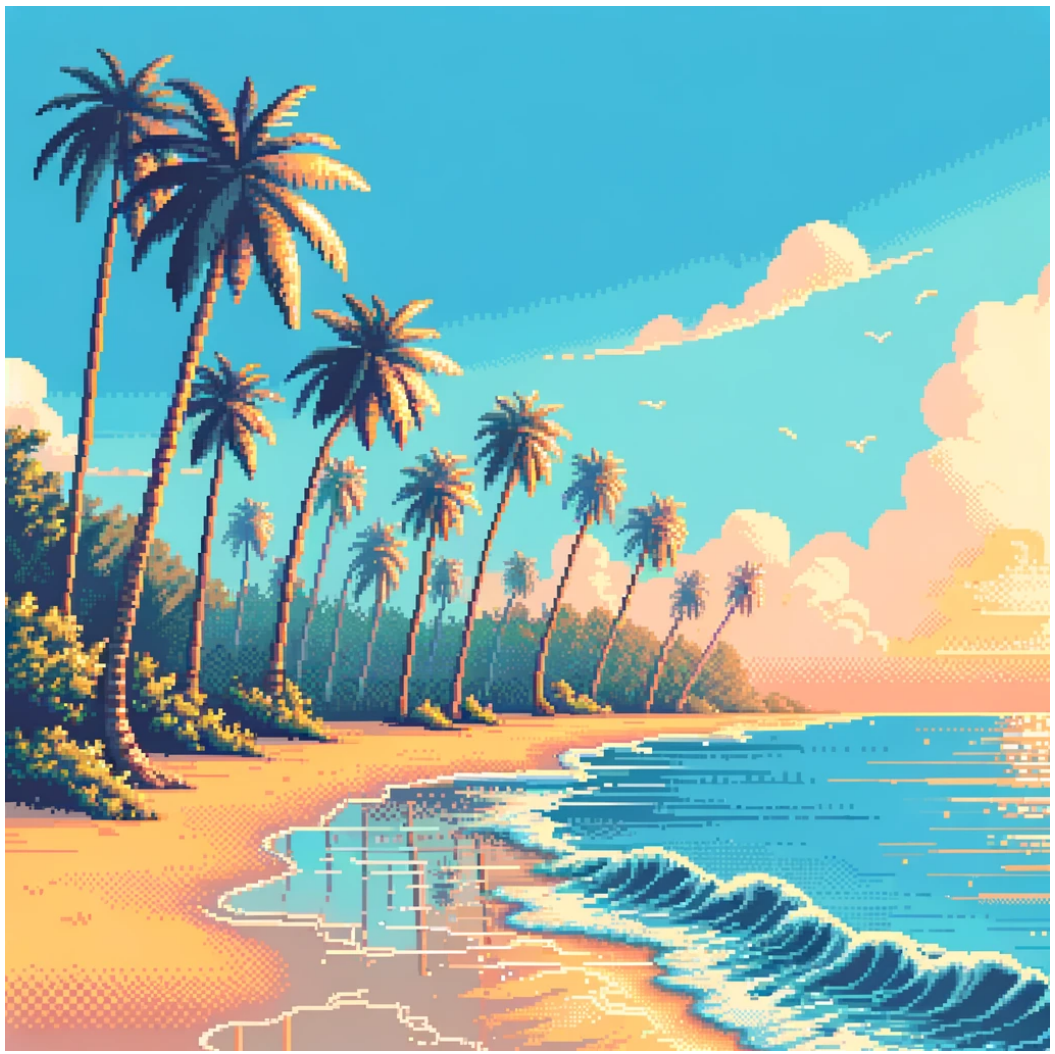
- Dates and times are stored as numbers in R
- Dates are stored as the number of days since 1970-01-01
- Times are stored as the number of seconds since 1970-01-01 00:00:00
- Dates and times are stored as numeric values, but can be formatted to look like dates and times

My turn: Dates

Sit back and enjoy!

Take a break

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



ds4owd-001.github.io/website/

10:00

Tables display

R packages for displaying tables

- Many packages for displaying tables in R
- `gt` package is one of the most popular and flexible
- `kable()` function of `knitr` package useful for simple tables

Our turn: md-05-exercises - tables

1. Open posit.cloud in your browser (use your bookmark).
2. Open the `ds4owd workspace` for the course.
3. Click `Start` next to `md-05-exercises`.
4. In the File Manager in the bottom right window, locate the `md-03c-tables.qmd` file and click on it to open it in the top left window.

Cross references

Cross references

- Help readers to navigate your document with numbered references and hyperlinks to entities like figures and tables.
- Cross referencing steps:
  - Add a caption to your figure or table.
  - Give an ID to your figure or table, starting with `fig-` or `tbl-`.
  - Refer to it with `@fig-...` or `@tbl-...`.

Table cross references

The presence of the caption (`A few penguins`) and label (`#tbl-penguins`) make this table referenceable:

See `@tbl-penguins` for data on a few penguins.

becomes:

See Table 1 for data on a few penguins.

```
```{r}
#| label: tbl-penguins
#| tbl-cap: A few penguins

head(penguins) |>
  gt()
```
```

Table 1:
A few penguins

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.3	20.6	190	3650	male	2007

Homework assignments module 5

Module 5 documentation

ds4owd-001.github.io/website/modules/md-05.html

Module 5

Data transformation, descriptive statistics and gt

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.
2. Learners can use functions of the gt R package to present summary tables in output formats.

Slides

- In preparation

Readings

1. Read R for Data Science - Section 4 - Data transformation
2. Read (Bryan 2018)

Assignments

- In preparation

Bring your own data for the capstone project

1. Find a dataset that you would like to work with.
2. Create a new repository on GitHub & clone to Posit Cloud.
3. Upload your data.
4. Create a new Quarto document an describe why you have chose this data.

Suitable datasets

- non-sensitive data that can be shared openly
- data that you have permission to use

Homework due date

- Homework assignment due: Monday, November 27th
- Correction & feedback phase up to: Thursday, November 30th

Wrap-up

ds4owd-001.github.io/website/

Thanks! 🌻

Slides created via revealjs and Quarto: https://quarto.org/docs/presentations/revealjs/ Access slides as PDF on GitHub

References

Broman, Karl W., and Kara H. Woo. 2018. “Data Organization in Spreadsheets.” *The American Statistician* 72 (1): 2–10. https://doi.org/10.1080/00031305.2017.1375989.