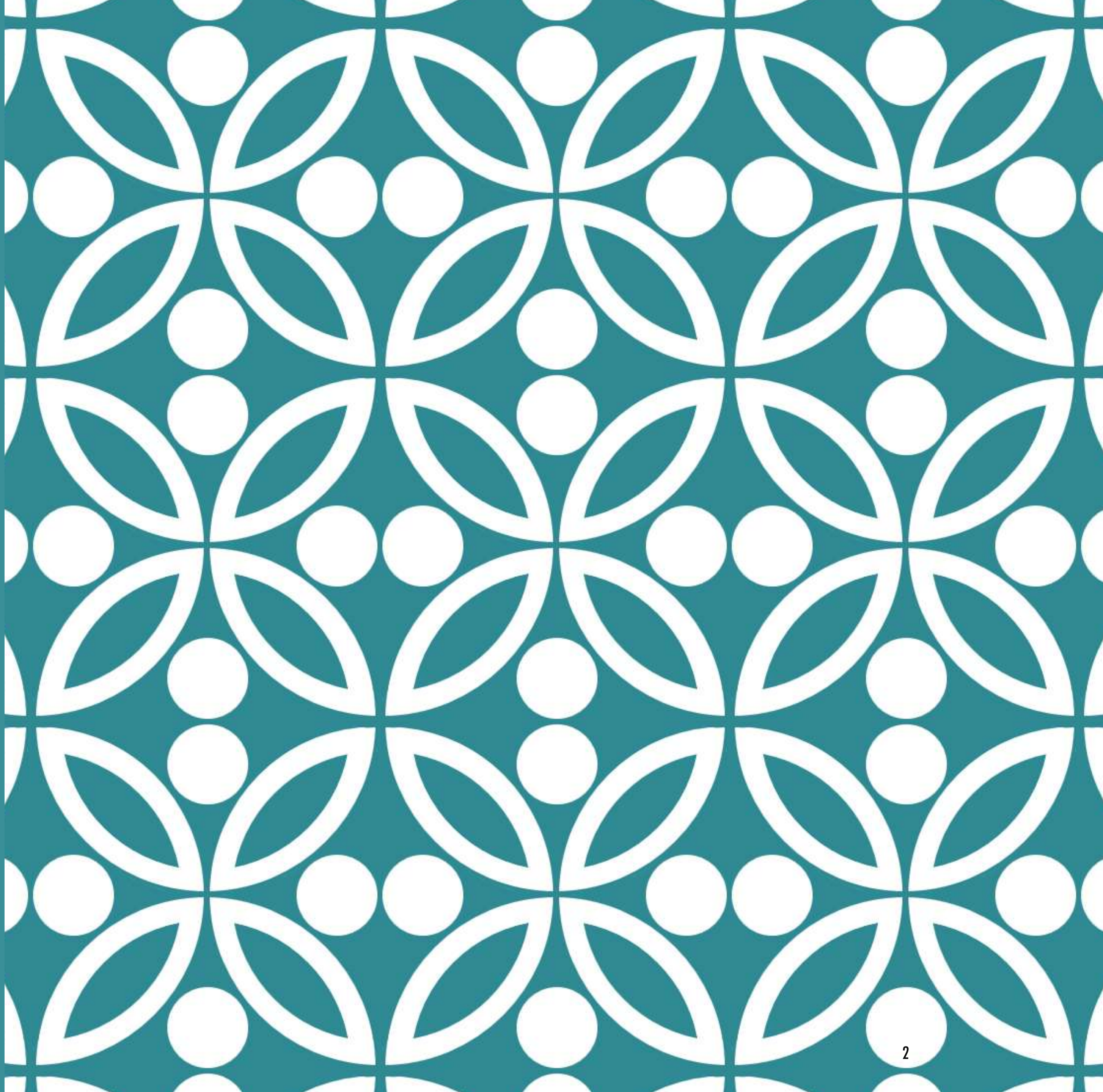
SESSION 2: TIDYVERSE

DR. SOFIA GIL-CLAVEL

- Session 1: Recap
- Ggplot2 & the Grammar of Graphs

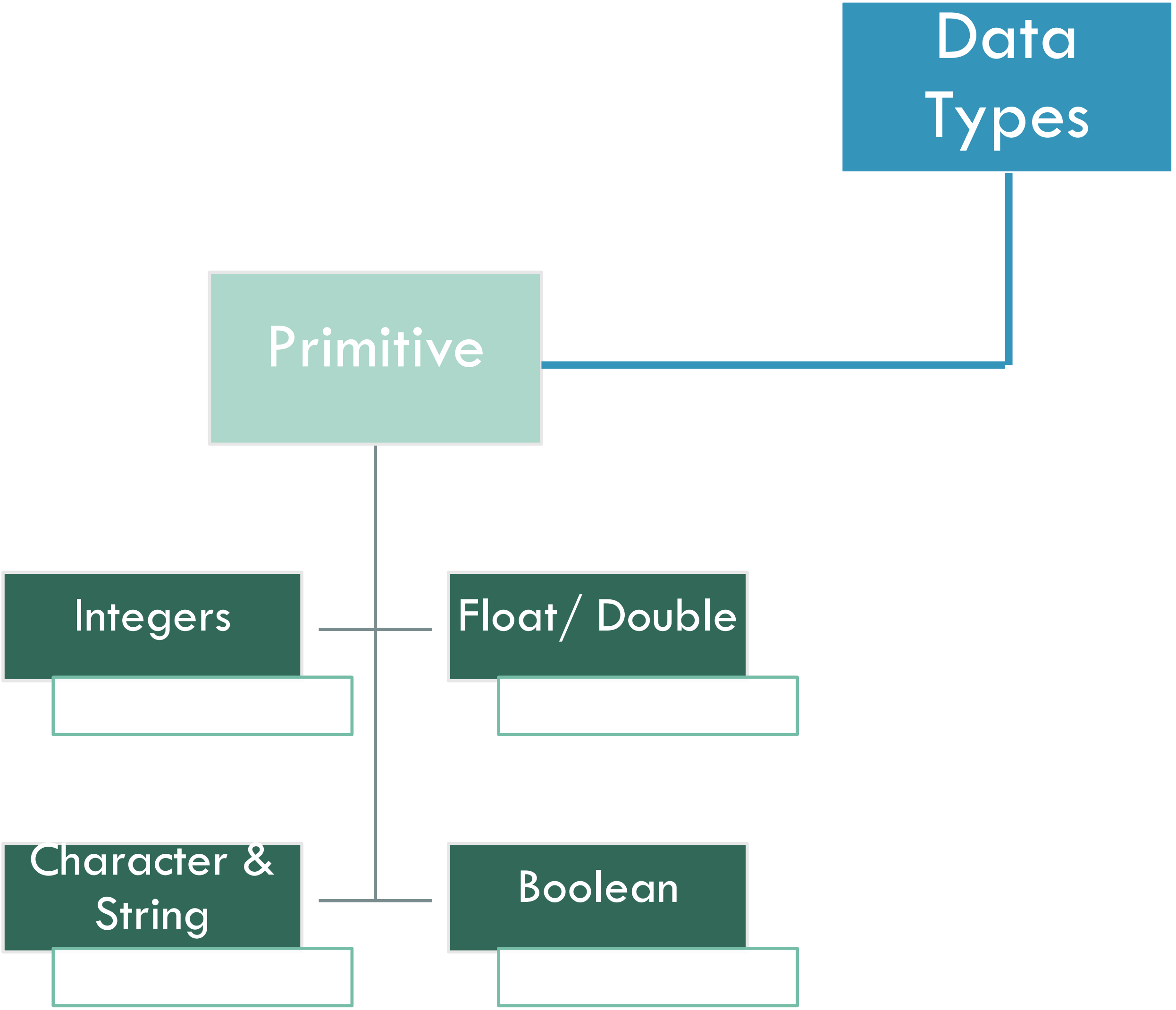
1. SESSION 1: RECAP

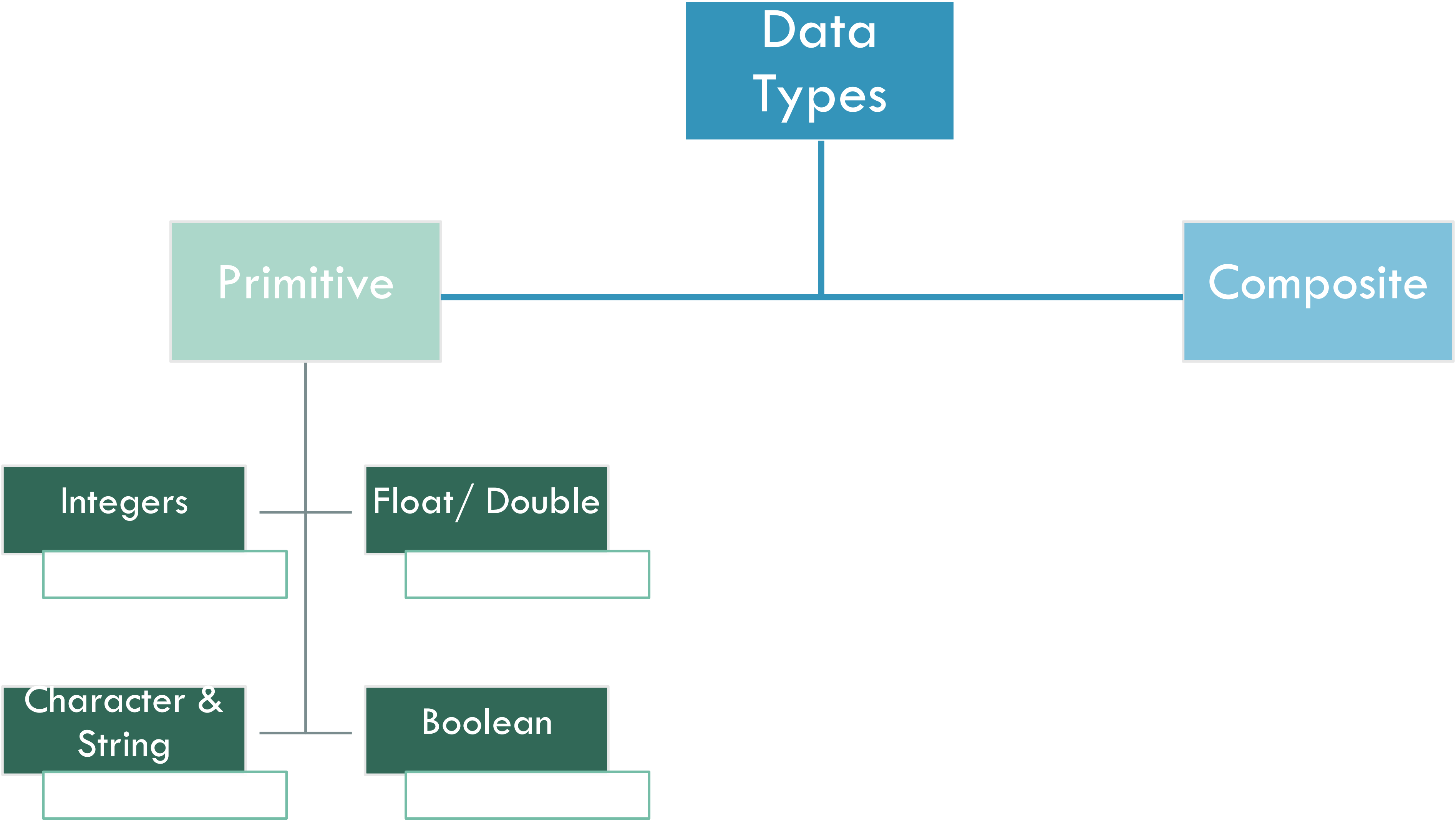


Data Types

Data Types

Primitive


Data Types
Primitive
Integers
Float/ Double
Character & String
Boolean



Data Types

Primitive

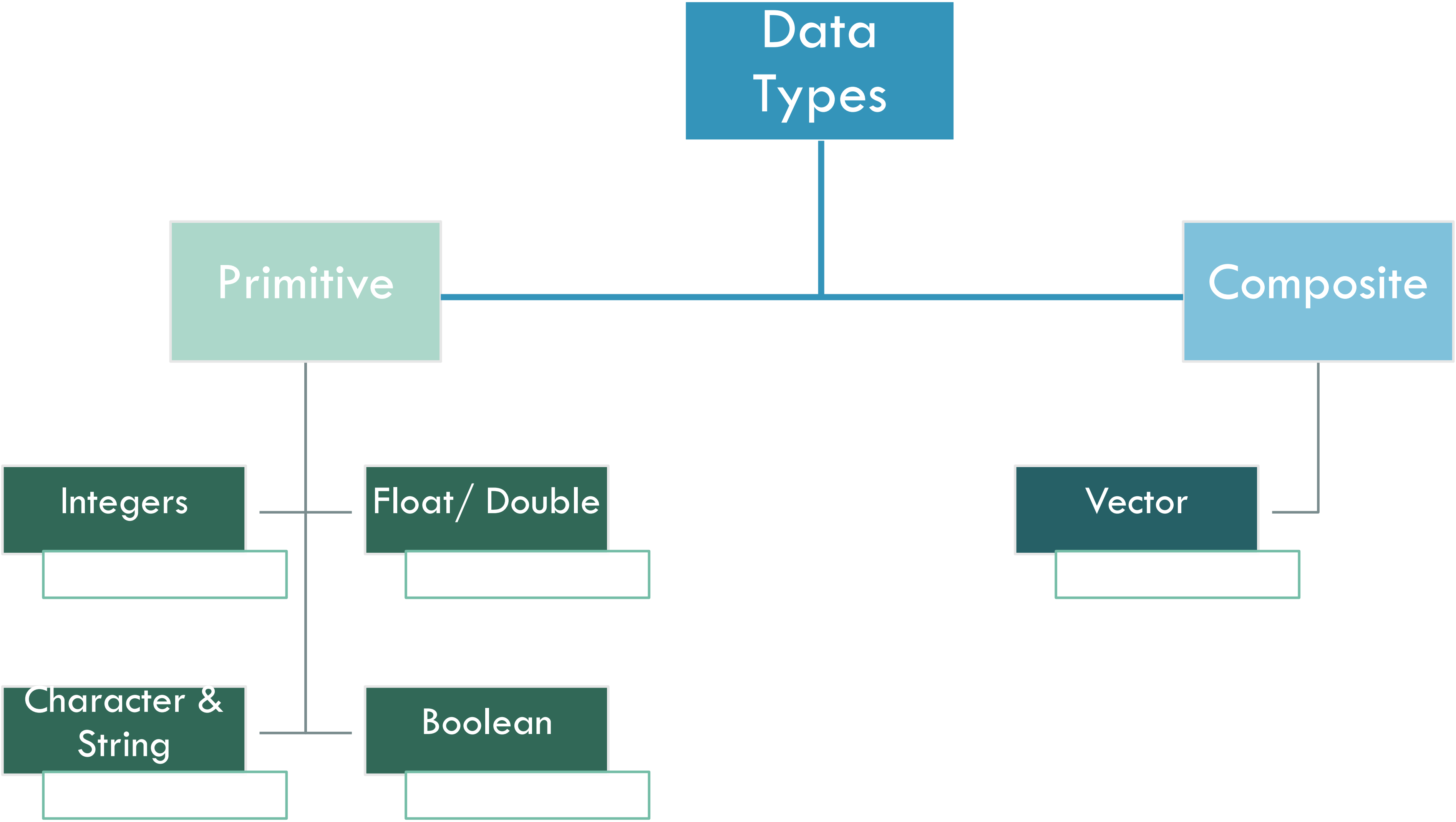
Composite

Integers

Float/ Double

Character & String

Boolean



Data Types

Primitive

Composite

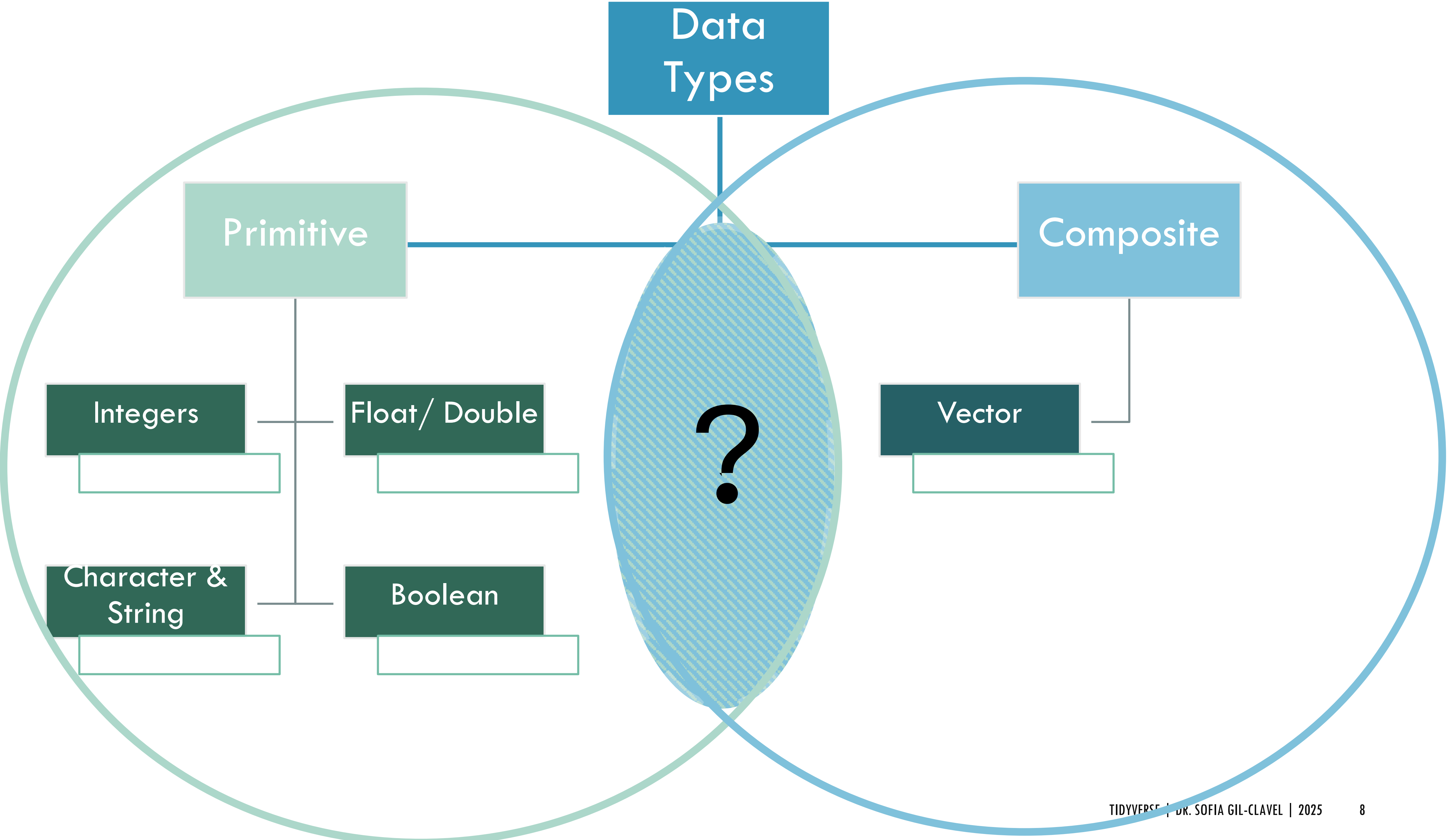
Integers

Float/ Double

Character & String

Boolean

Vector



Data Types

Primitive

Composite

Integers

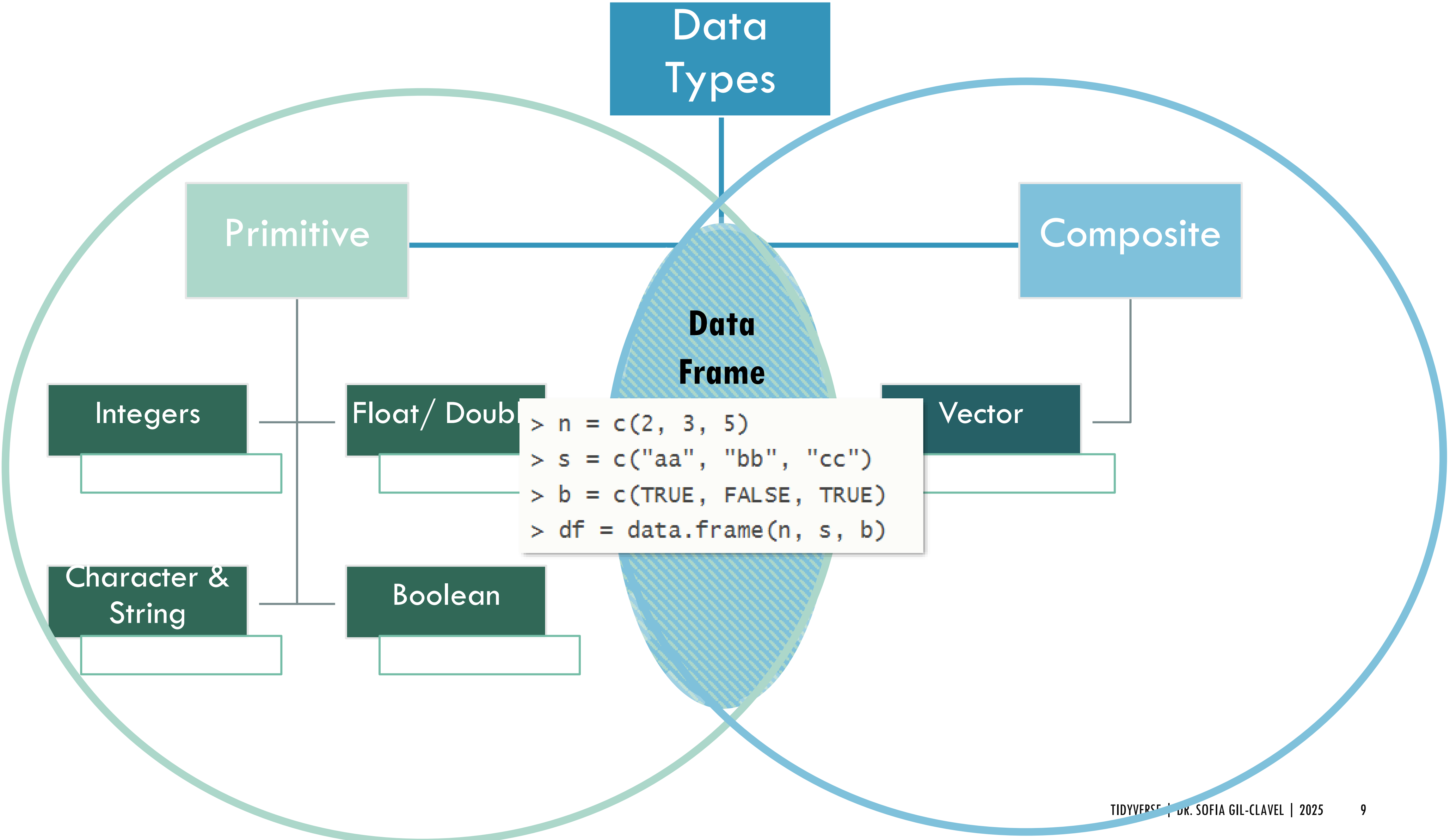
Float/ Double

Character & String

Boolean

Vector

?



Data Types

Primitive

Composite

Data Frame

Integers

Float/ Doubl

Vector

Character & String

Boolean

```
> n = c(2, 3, 5)
> s = c("aa", "bb", "cc")
> b = c(TRUE, FALSE, TRUE)
> df = data.frame(n, s, b)
```

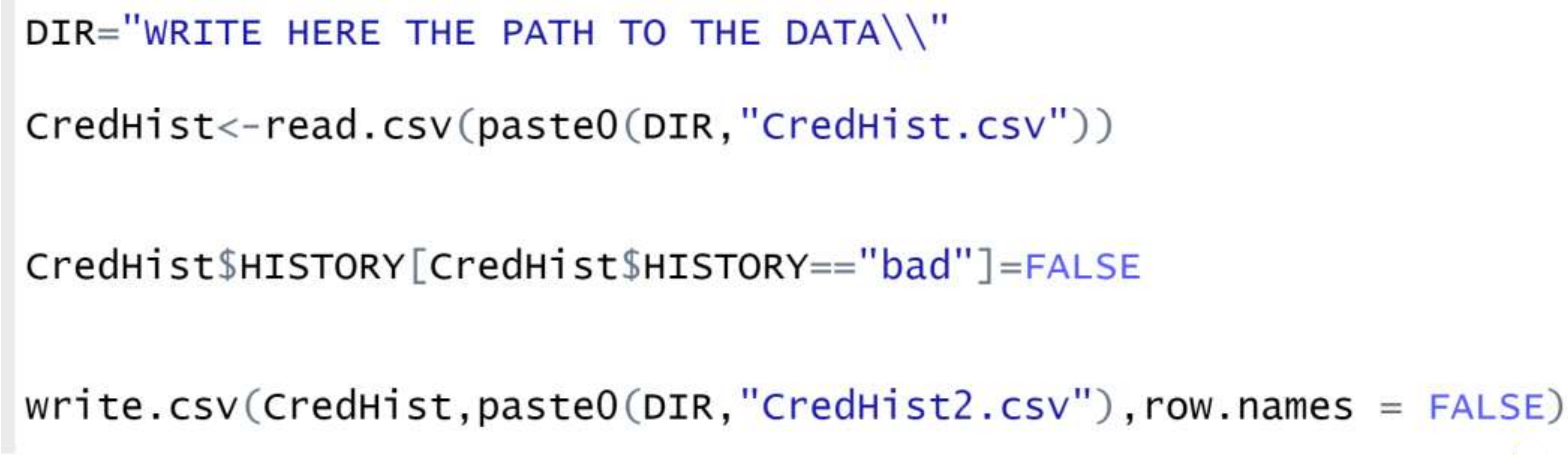
PRIMITIVE DATA AND ITS OPERATORS

Operators					
Arithmetic		**Comparative**		**Logic**	
+	addition	<	less than	! x	logic NO
-	subtraction	>	more than	x & y	element-wise AND
*	multiplication	<=	less or equal than	x && y	single comparison AND
/	division	>=	more or equal than	x | y	element-wise OR
^	power	==	equal than	x || y	single comparison OR
%/%	integer division	!=	different than	xor(x,y)	exclusive OR

DATA FRAMES: OPEN AND SAVE

With the *foreign* library you can manipulate databases other than ".csv"

Open



```
DIR="WRITE HERE THE PATH TO THE DATA\\"

CredHist<-read.csv(paste0(DIR,"CredHist.csv"))
```

Modify

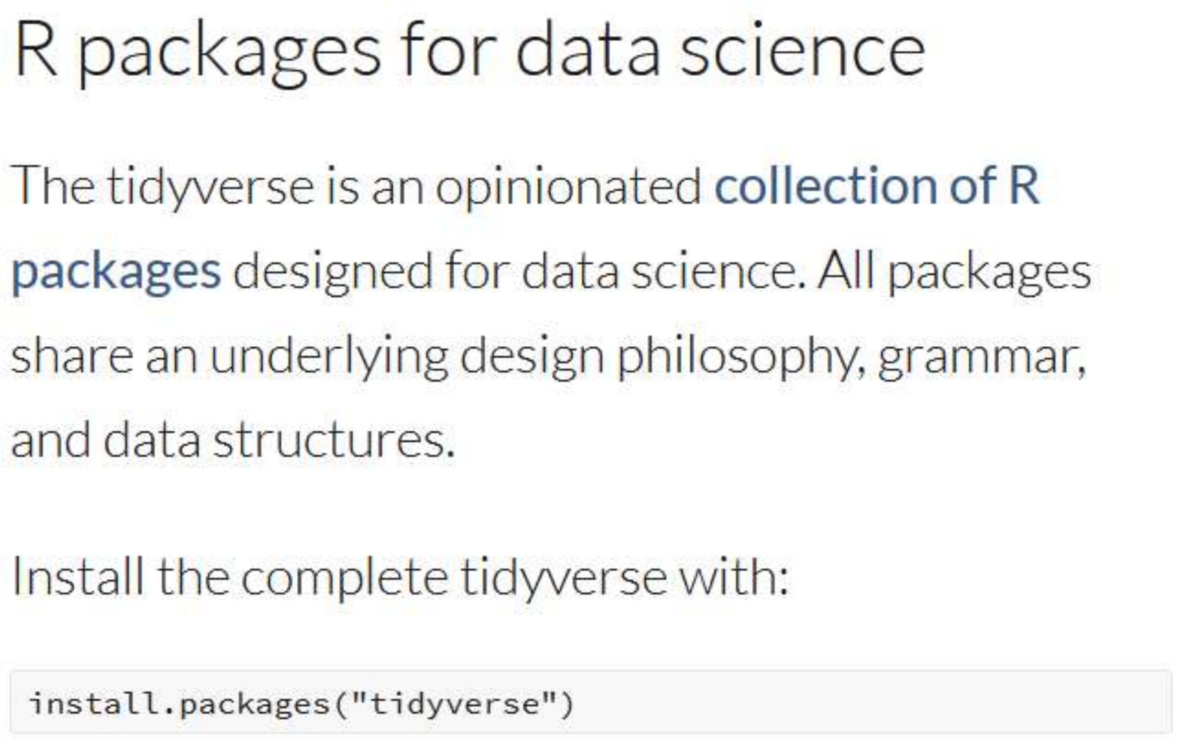
```
CredHist$HISTORY[CredHist$HISTORY=="bad"]=FALSE
```

Save

```
write.csv(CredHist,paste0(DIR,"CredHist2.csv"),row.names = FALSE)
```

2. TIDYVERSE

❖ TIDYVERSE: HTTPS://WWW.TIDYVERSE.ORG/



R packages for data science

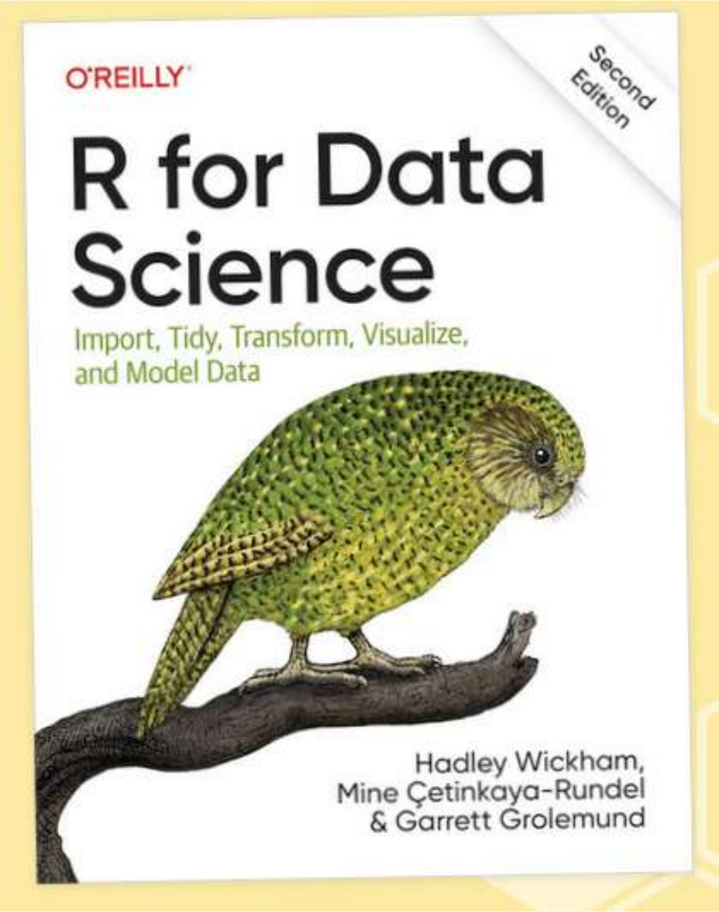
The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

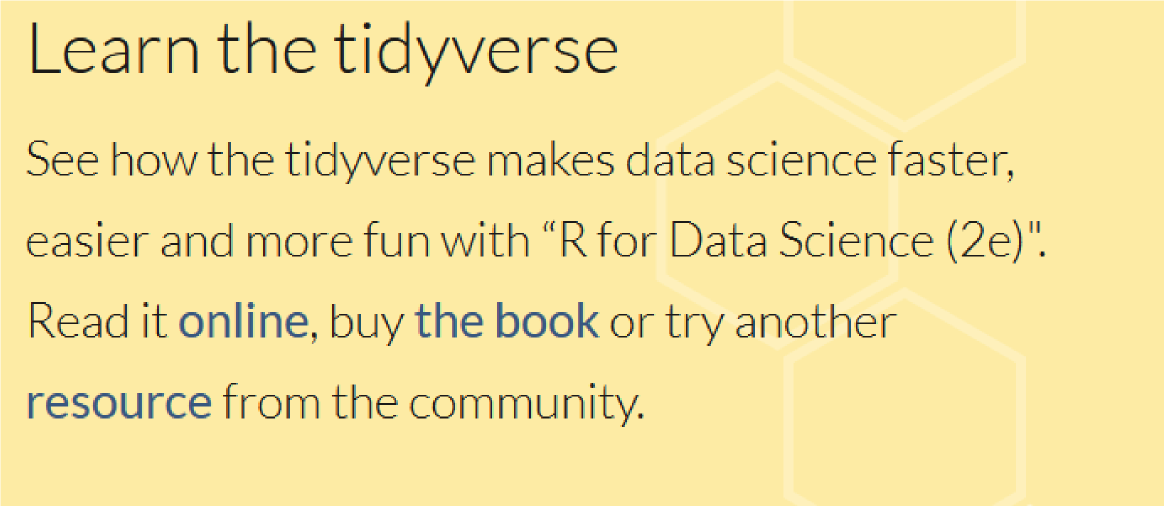
Install the complete tidyverse with:

```
install.packages("tidyverse")
```

10 MINS BREAK
WHILE TIDYVERSE INSTALLS

❖ TIDYVERSE: HTTPS://WWW.TIDYVERSE.ORG/

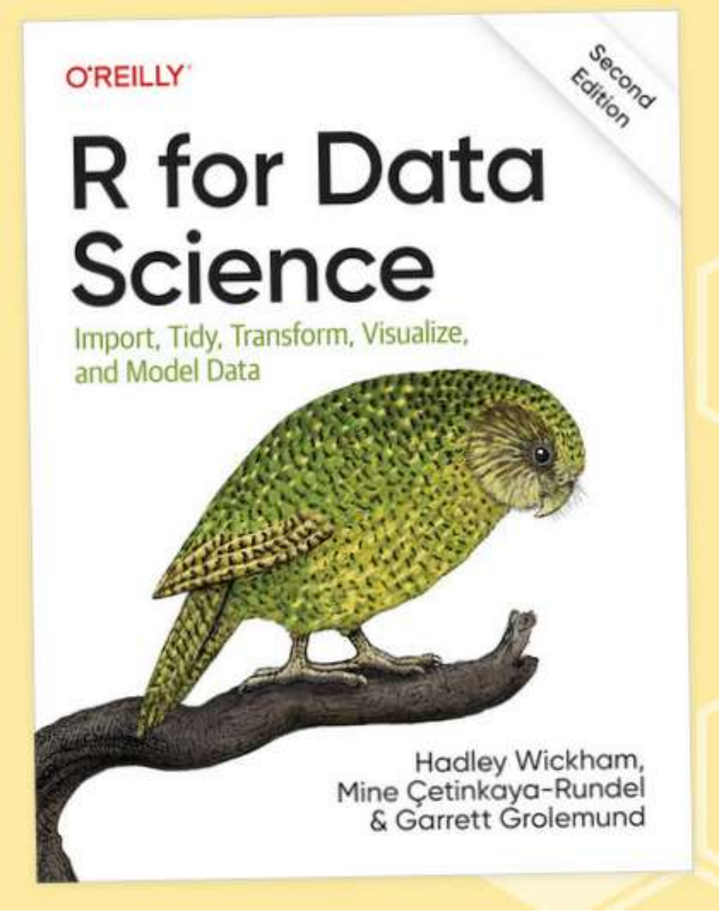

O'REILLY
Second Edition
R for Data Science
Import, Tidy, Transform, Visualize, and Model Data
Hadley Wickham, Mine Çetinkaya-Rundel & Garrett Grolemund



Learn the tidyverse

See how the tidyverse makes data science faster, easier and more fun with "R for Data Science (2e)". Read it online, buy the book or try another resource from the community.

❖ TIDYVERSE: HTTPS://WWW.TIDYVERSE.ORG/



O'REILLY

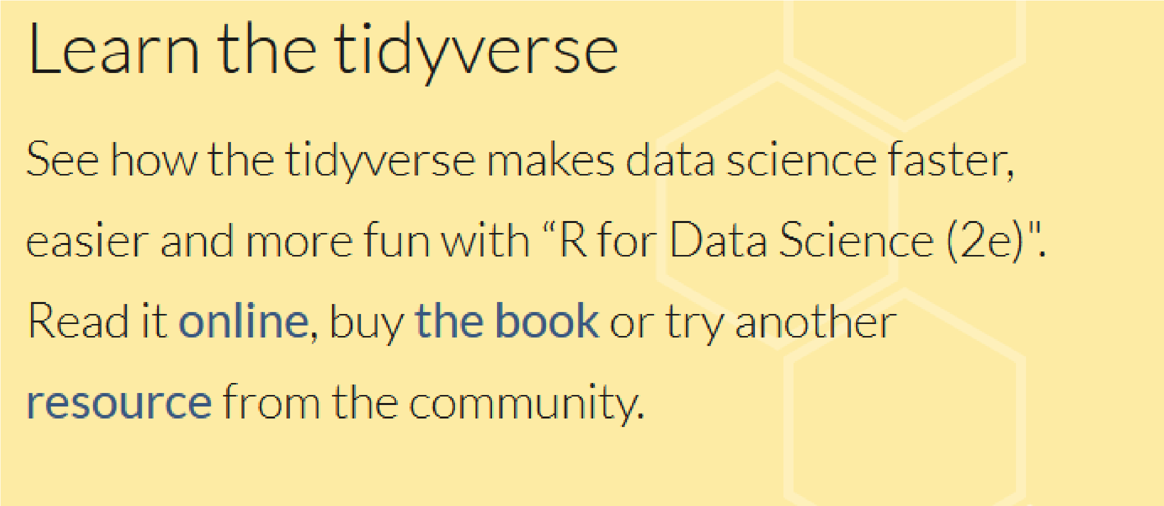
Second Edition

R for Data Science

Import, Tidy, Transform, Visualize, and Model Data

Hadley Wickham, Mine Çetinkaya-Rundel & Garrett Grolemund

You can check the bookdown version of the book in: https://r4ds.hadley.nz/

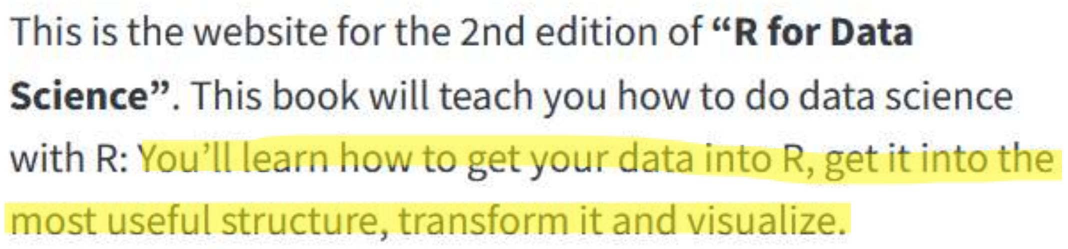


Learn the tidyverse

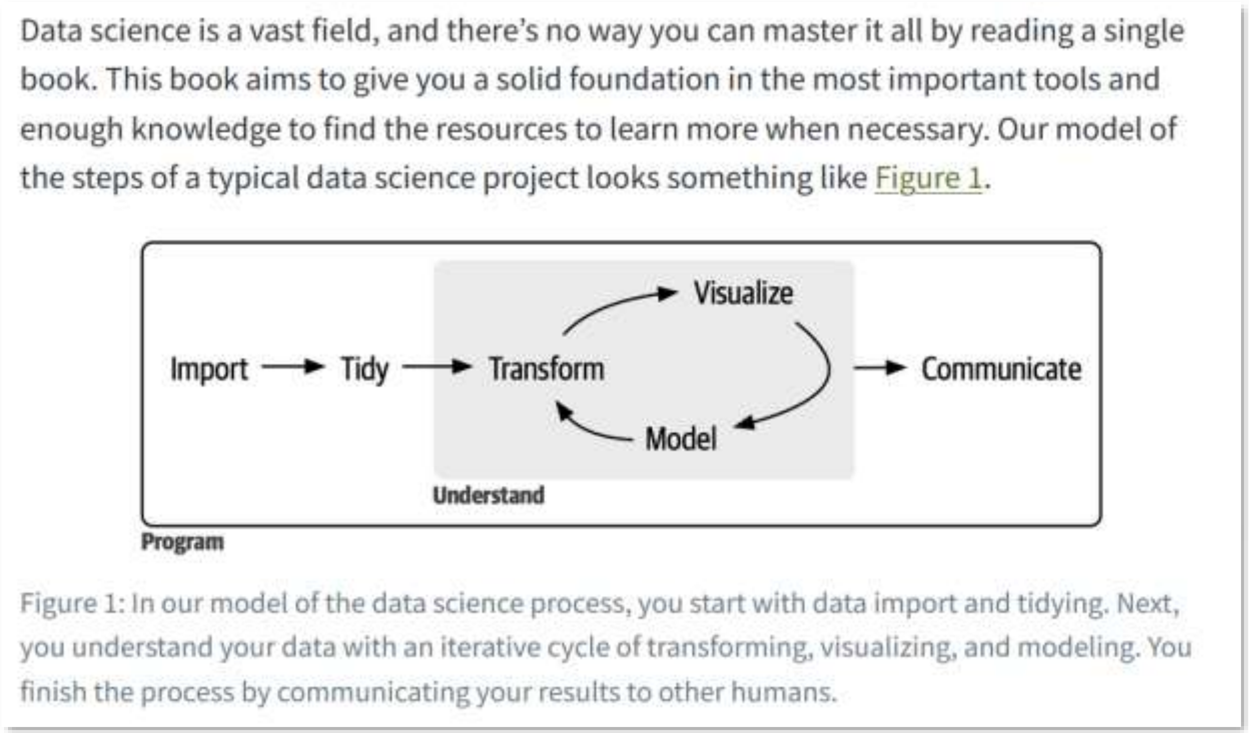
See how the tidyverse makes data science faster, easier and more fun with "R for Data Science (2e)". Read it online, buy the book or try another resource from the community.

❖ TIDYVERSE: [HTTPS://WWW.TIDYVERSE.ORG/](https://www.tidyverse.org/)

You can check the bookdown version of the book in: https://r4ds.hadley.nz/

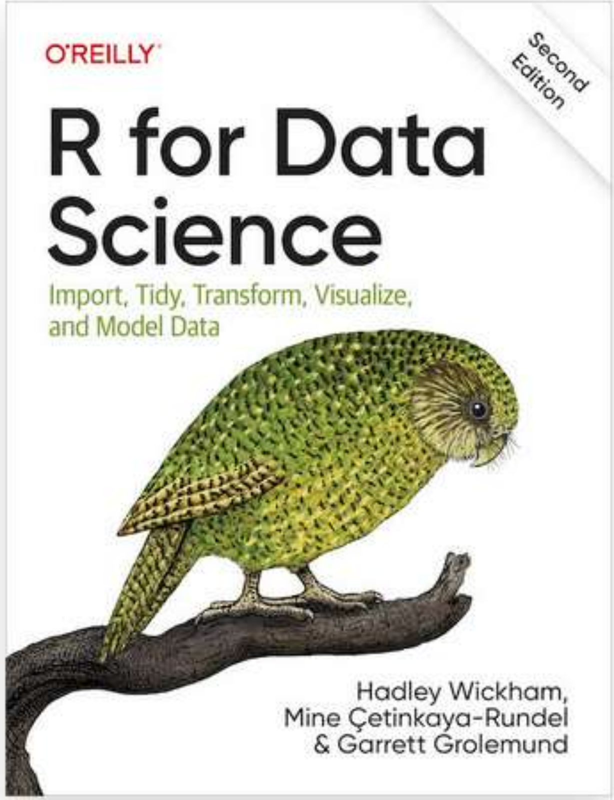


This is the website for the 2nd edition of **“R for Data Science”**. This book will teach you how to do data science with R: You’ll learn how to get your data into R, get it into the most useful structure, transform it and visualize.



Data science is a vast field, and there’s no way you can master it all by reading a single book. This book aims to give you a solid foundation in the most important tools and enough knowledge to find the resources to learn more when necessary. Our model of the steps of a typical data science project looks something like Figure 1.

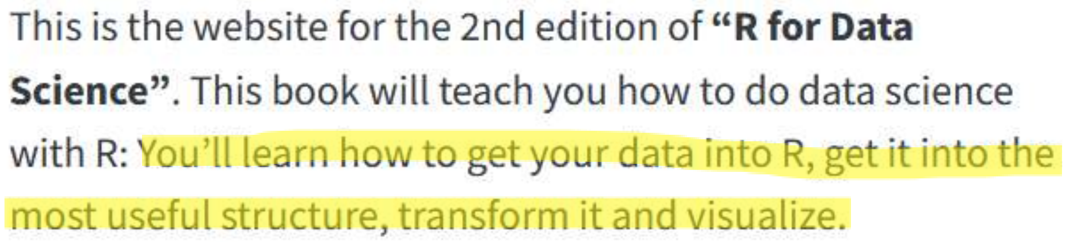
Figure 1: In our model of the data science process, you start with data import and tidying. Next, you understand your data with an iterative cycle of transforming, visualizing, and modeling. You finish the process by communicating your results to other humans.



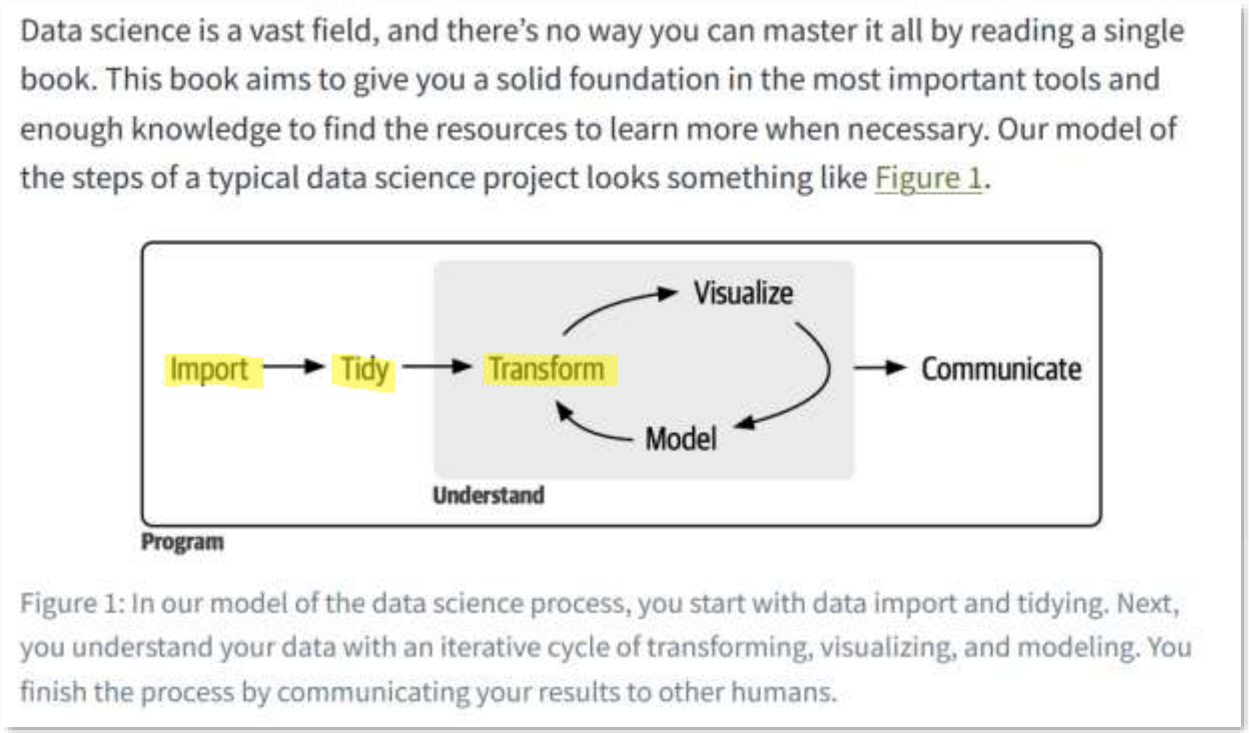
O'REILLY
Second Edition
R for Data Science
Import, Tidy, Transform, Visualize, and Model Data
Hadley Wickham, Mine Çetinkaya-Rundel & Garrett Grolemund

❖ TIDYVERSE: [HTTPS://WWW.TIDYVERSE.ORG/](https://www.tidyverse.org/)

You can check the bookdown version of the book in: https://r4ds.hadley.nz/

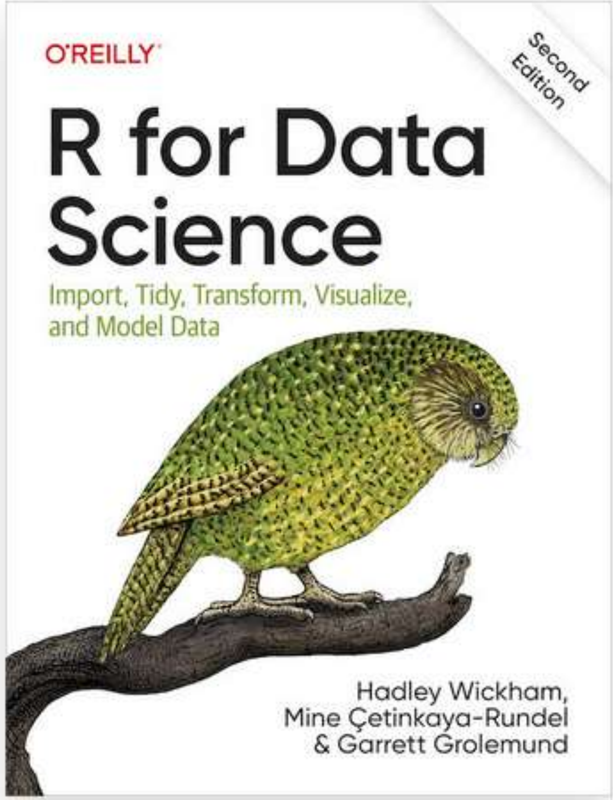


This is the website for the 2nd edition of **"R for Data Science"**. This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it and visualize.



Data science is a vast field, and there's no way you can master it all by reading a single book. This book aims to give you a solid foundation in the most important tools and enough knowledge to find the resources to learn more when necessary. Our model of the steps of a typical data science project looks something like Figure 1.

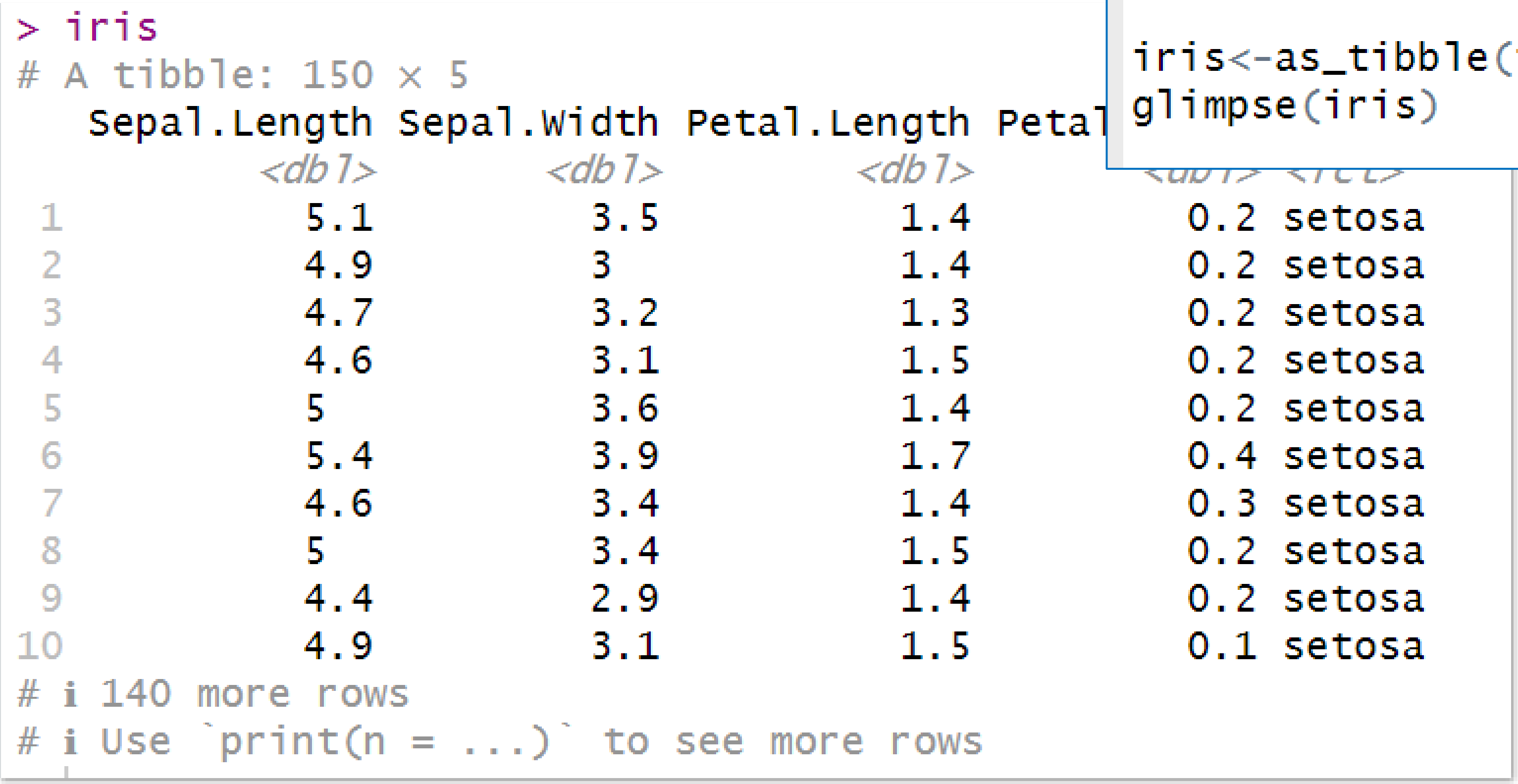
Figure 1: In our model of the data science process, you start with data import and tidying. Next, you understand your data with an iterative cycle of transforming, visualizing, and modeling. You finish the process by communicating your results to other humans.



O'REILLY
Second Edition
R for Data Science
Import, Tidy, Transform, Visualize, and Model Data
Hadley Wickham, Mine Çetinkaya-Rundel & Garrett Grolemund

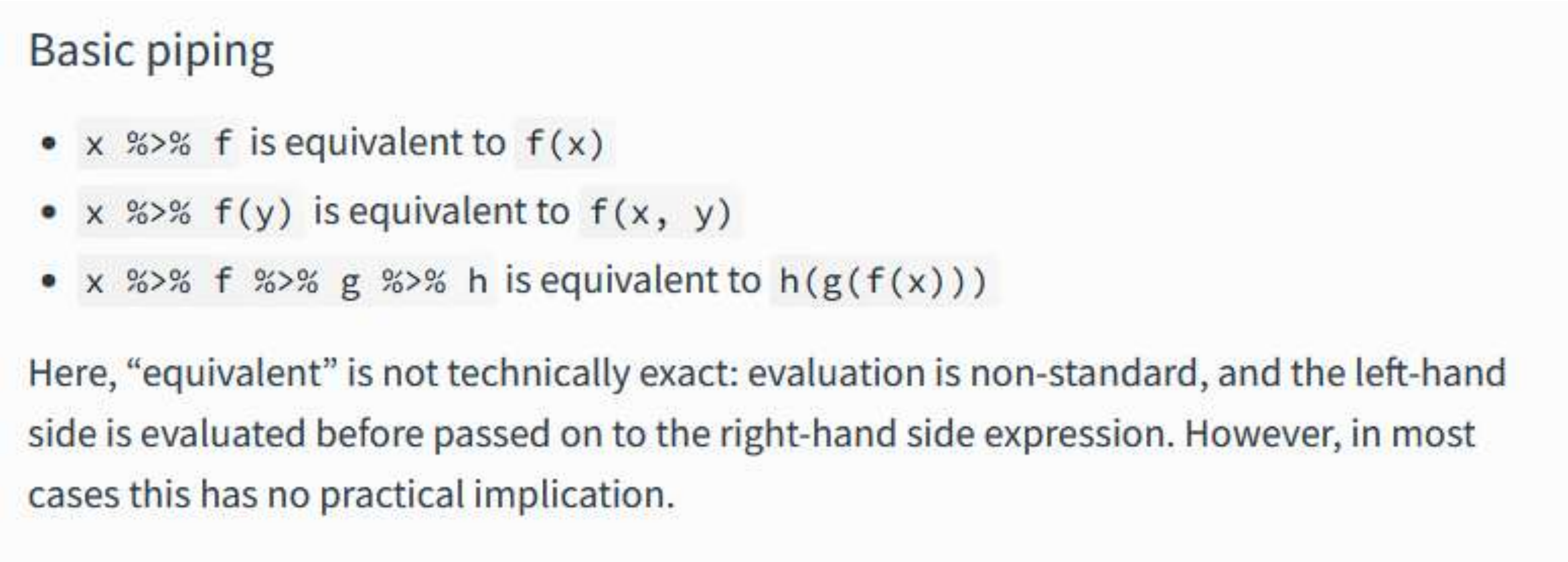
❖ TIBBLES: HTTPS://TIBBLE.TIDYVERSE.ORG/

```r
library(tidyverse)
?datasets

View(iris)
?iris
summary(iris)

iris
as_tibble(iris)

iris<-as_tibble(iris)
glimpse(iris)
```



```
> iris
# A tibble: 150 × 5
   Sepal.Length Sepal.Width Petal.Length Petal
          <dbl>       <dbl>        <dbl>
 1          5.1         3.5          1.4         0.2 setosa
 2          4.9         3            1.4         0.2 setosa
 3          4.7         3.2          1.3         0.2 setosa
 4          4.6         3.1          1.5         0.2 setosa
 5          5           3.6          1.4         0.2 setosa
 6          5.4         3.9          1.7         0.4 setosa
 7          4.6         3.4          1.4         0.3 setosa
 8          5           3.4          1.5         0.2 setosa
 9          4.4         2.9          1.4         0.2 setosa
10          4.9         3.1          1.5         0.1 setosa
# ℹ 140 more rows
# ℹ Use `print(n = ...)` to see more rows
```

❖ PIPES (%>%): [HTTPS://MAGRITTR.TIDYVERSE.ORG/](https://magrittr.tidyverse.org/)

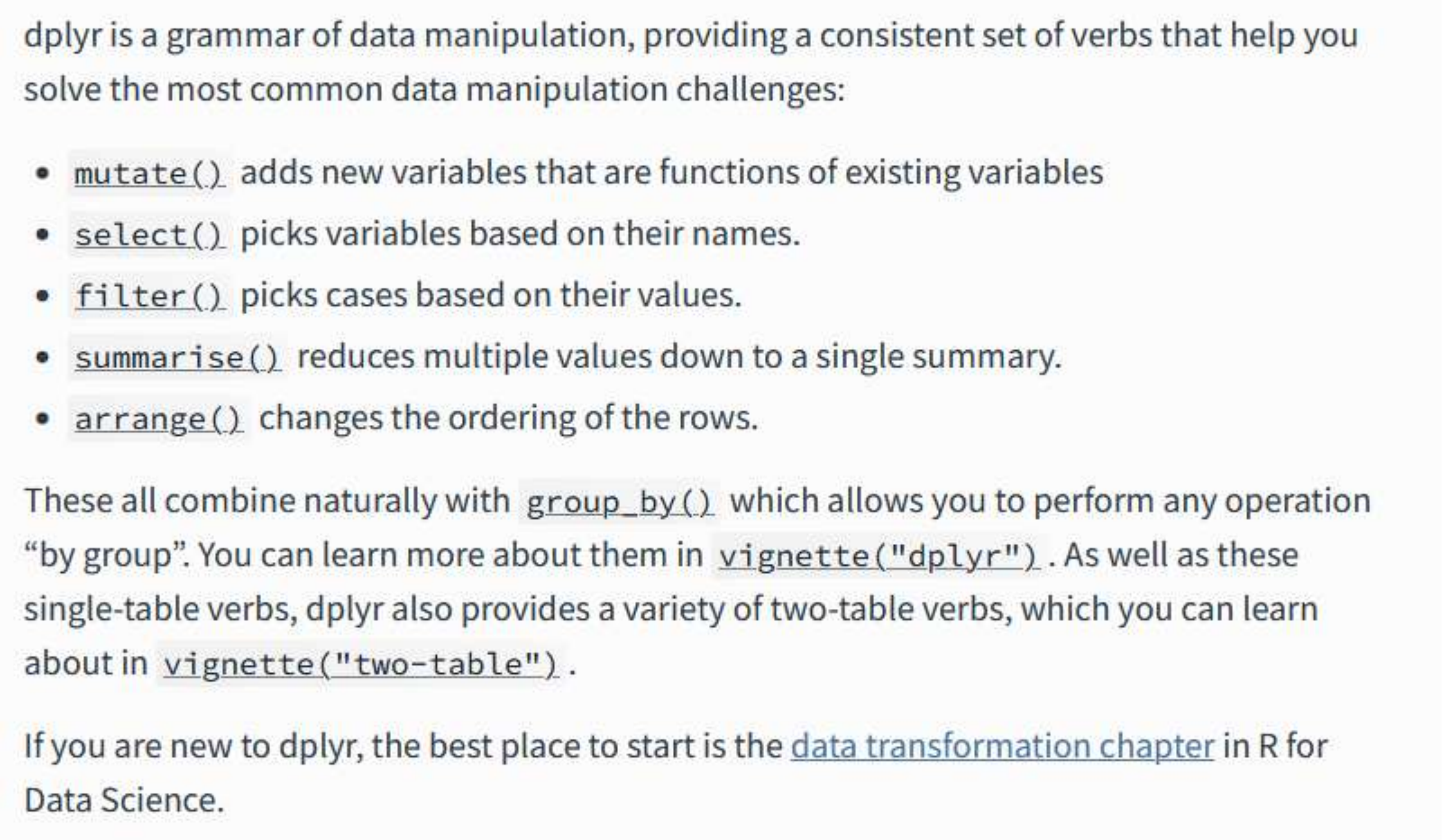


Basic piping

- `x %>% f` is equivalent to `f(x)`
- `x %>% f(y)` is equivalent to `f(x, y)`
- `x %>% f %>% g %>% h` is equivalent to `h(g(f(x)))`

Here, “equivalent” is not technically exact: evaluation is non-standard, and the left-hand side is evaluated before passed on to the right-hand side expression. However, in most cases this has no practical implication.

❖ DPLYR: HTTPS://DPLYR.TIDYVERSE.ORG/



dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges:

- `mutate()` adds new variables that are functions of existing variables
- `select()` picks variables based on their names.
- `filter()` picks cases based on their values.
- `summarise()` reduces multiple values down to a single summary.
- `arrange()` changes the ordering of the rows.

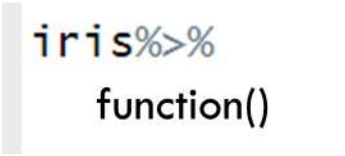
These all combine naturally with `group_by()` which allows you to perform any operation "by group". You can learn more about them in `vignette("dplyr")`. As well as these single-table verbs, dplyr also provides a variety of two-table verbs, which you can learn about in `vignette("two-table")`.

If you are new to dplyr, the best place to start is the data transformation chapter in R for Data Science.

EXERCISE 2.1

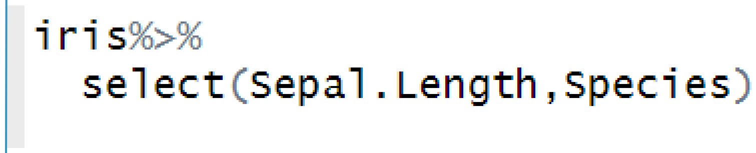
What do the following functions do?

pipe (%>%)



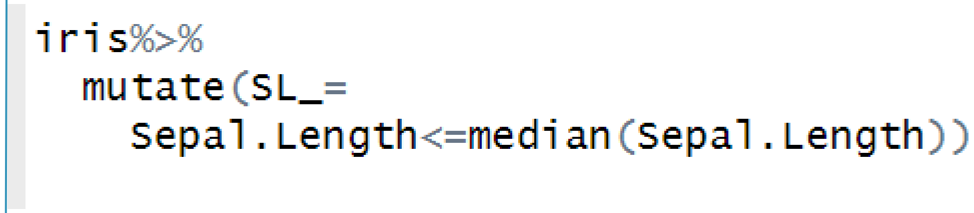
```
iris%>%
   function()
```

select



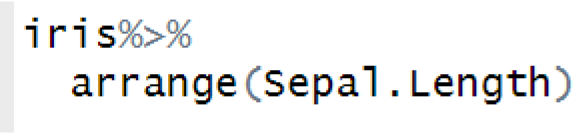
```
iris%>%
  select(Sepal.Length,Species)
```

mutate



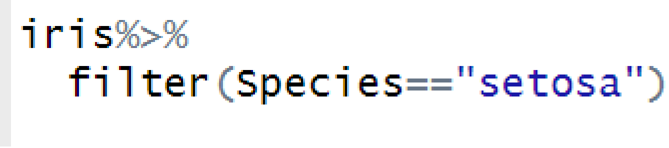
```
iris%>%
  mutate(SL_=
    Sepal.Length<=median(Sepal.Length))
```

arrange



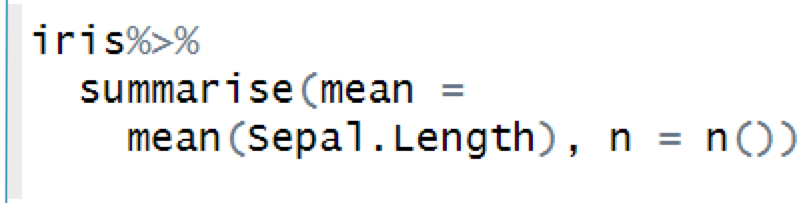
```
iris%>%
  arrange(Sepal.Length)
```

filter



```
iris%>%
  filter(Species=="setosa")
```

summarise



```
iris%>%
  summarise(mean =
    mean(Sepal.Length), n = n())
```

EXERCISE 2.2

In the Data folder you will find CredHist.csv, for this exercise you will have to create the code in R to be able to **open and find the number of people who can have a loan with the bank according to the following restrictions**:

a) The person does not have a bad credit history.

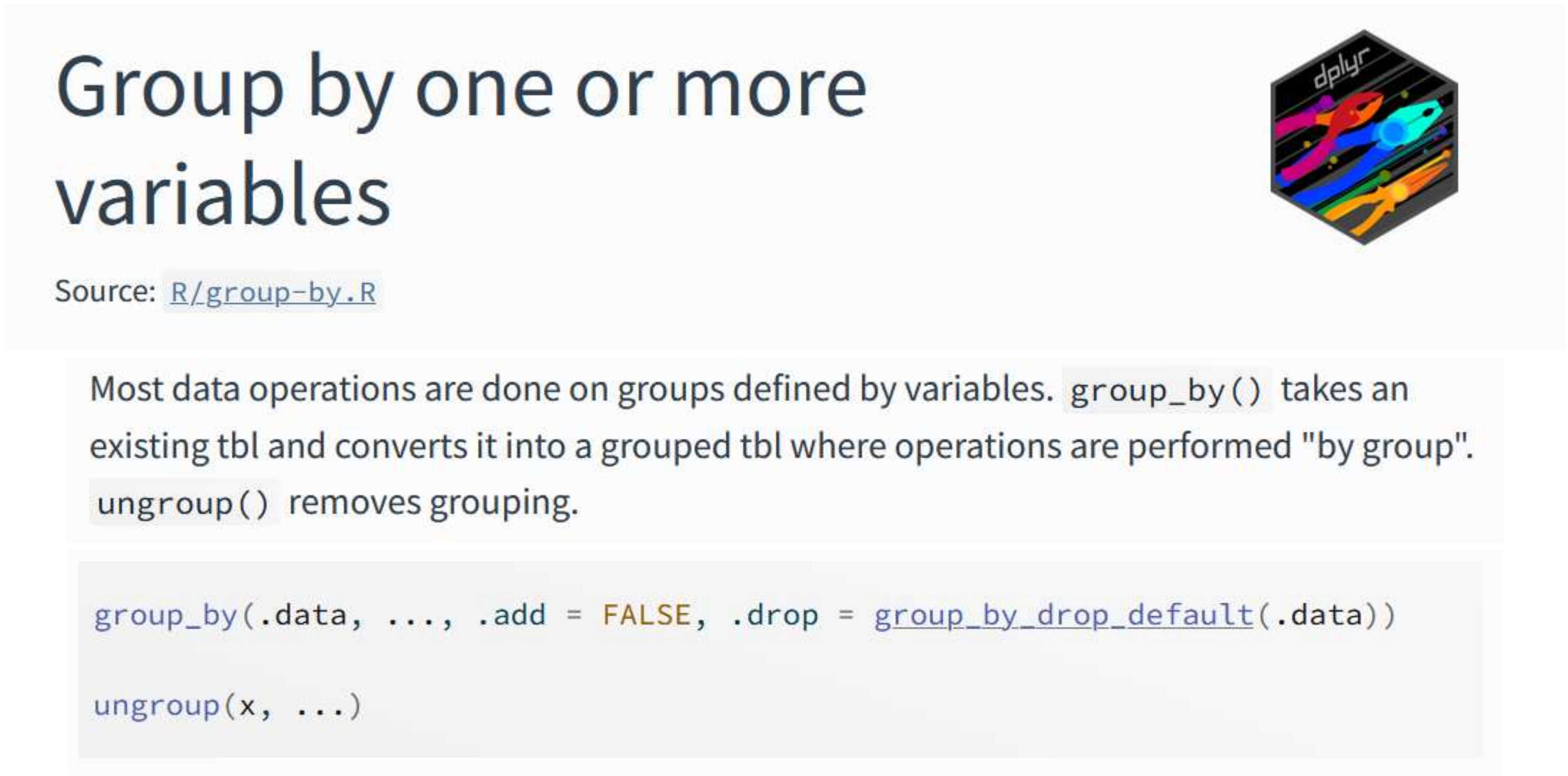
b) The person has a monthly income greater than 1000EUR.

c) The person is not older than 65 years old.

The bank applies the following exceptions:

- If you are an employee of the bank, you are granted.
- If you are over the age of 65, but your monthly income exceeds 5000EUR, you are awarded.

Once the people have been found, the subbase must be saved in a new ".csv". What are the dimensions of the resulting base? What kind of data did you have to work with?



Group by one or more variables

Source: R/group-by.R

Most data operations are done on groups defined by variables. `group_by()` takes an existing tbl and converts it into a grouped tbl where operations are performed "by group". `ungroup()` removes grouping.

```
group_by(.data, ..., .add = FALSE, .drop = group_by_drop_default(.data))

ungroup(x, ...)
```

Source: https://dplyr.tidyverse.org/reference/group_by.html

EXERCISE 2.3

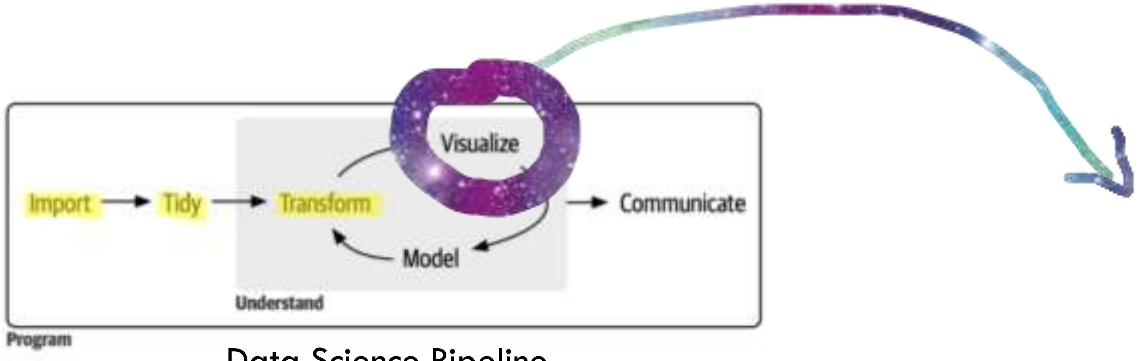
With the subbase that was found in "Exercise 2.2" find the following data:

1. Number of Men and Women Who Can Have the Credit.
2. Average Age of Men and Women.
3. Median monthly Income of Women and Men Over 25.
4. Number of Bank Female and Male employees in the database.
5. Average monthly Income of Female and Male Bank Employees.

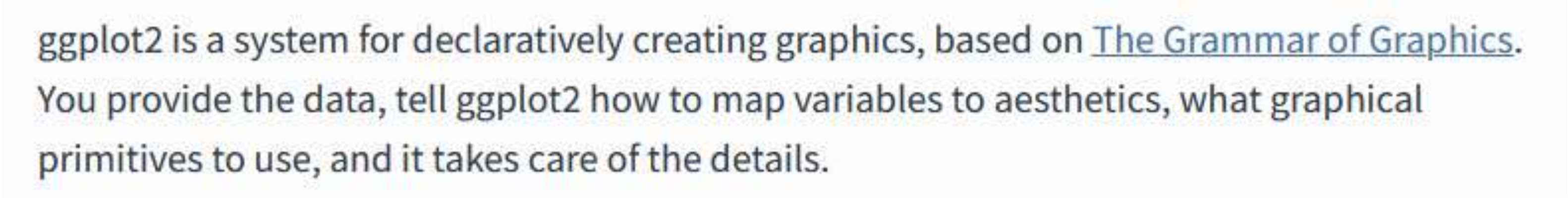
REFERENCES

- Albert, Jim, and Maria Rizzo. *R by Example: Concepts to Code*. Use R! New York, NY: Springer New York, 2012. https://doi.org/10.1007/978-1-4614-1365-3.
- Davies, Tilman M. *The Book of R: A First Course in Programming and Statistics*. San Francisco: No Starch Press, 2016. https://web.itu.edu.tr/~tokerem/The_Book_of_R.pdf
- Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolemund. *R for data science*. " O'Reilly Media, Inc.", 2023. Accessed May 7, 2024. https://r4ds.hadley.nz/.

SESSION 3: GGPLOT2 AND THE GRAMMAR OF GRAPHS



Data Science Pipeline



ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Source: https://ggplot2.tidyverse.org/

Join us!

https://forms.office.com/e/8Bgd2YsasJ

All about the lab:

https://societal-analytics.nl/

Contact us at:

analytics-lab.fsw@vu.nl.

https://sofiag1l.github.io/

THANKS!

Dr. Sofia Gil-Clavel