

Session 2 - Data wrangling

R training

María Reyes Retana
The World Bank | January 2025





Reproducible Research Repository

Government Analytics and R Training:

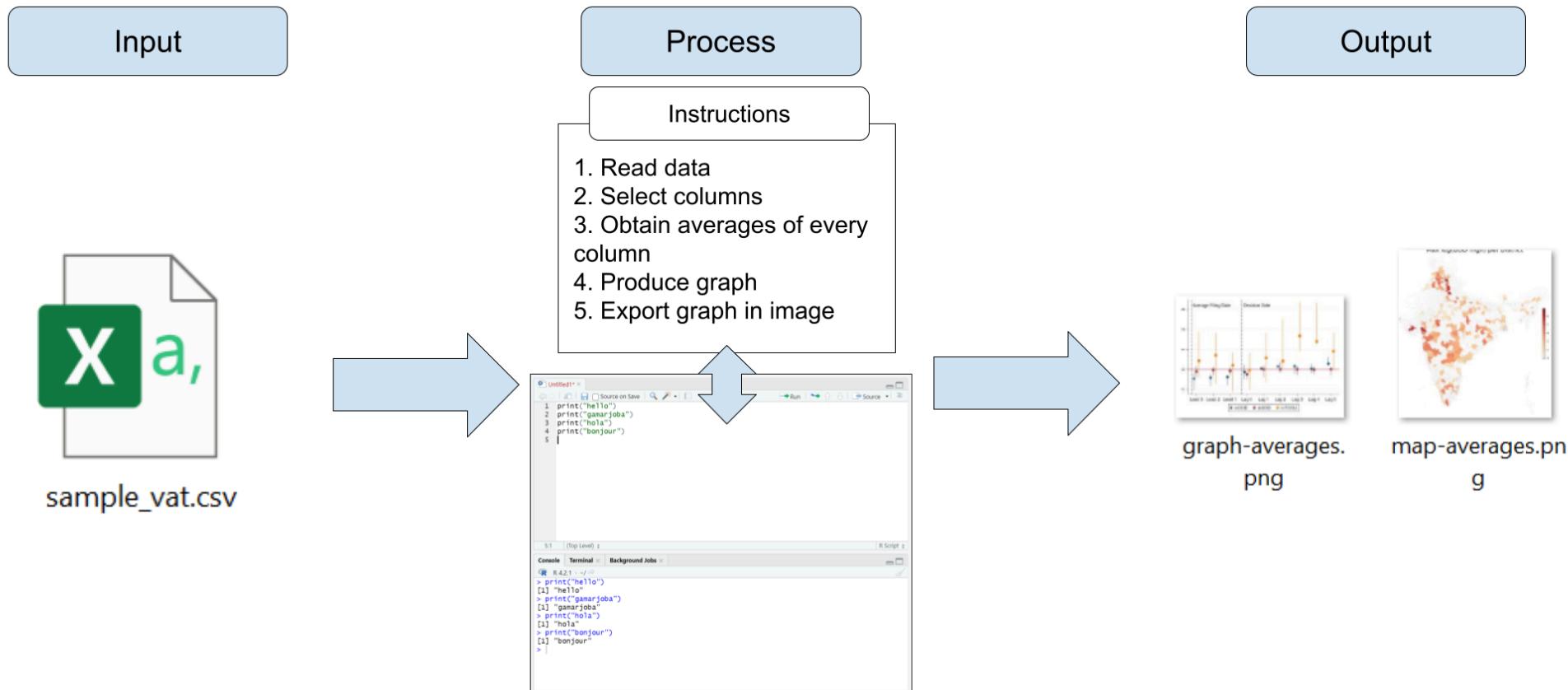
Strengthening Public Sector Reporting
and Data Analysis

January 13–17, 2024

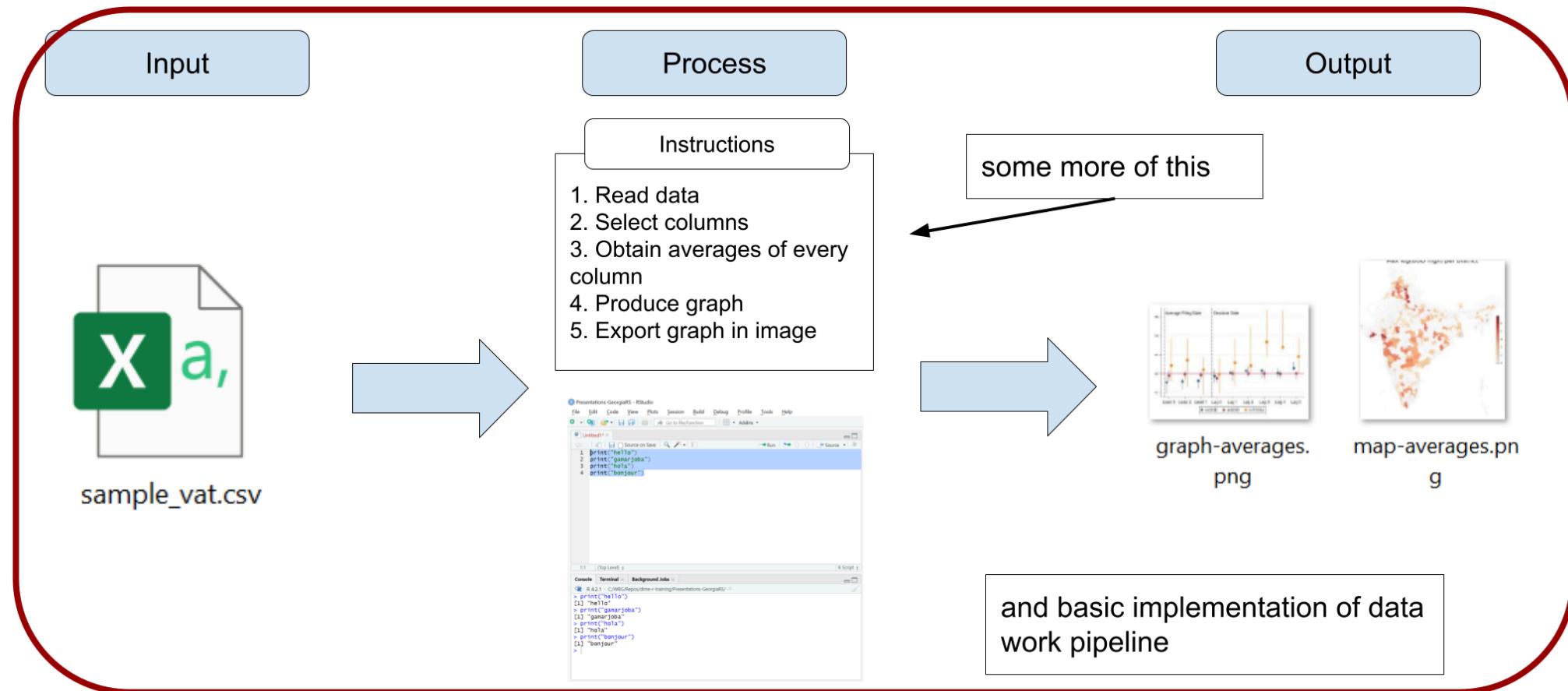


About this session

About this session



About this session



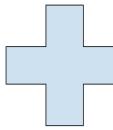
R Packages

R Packages

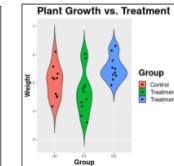
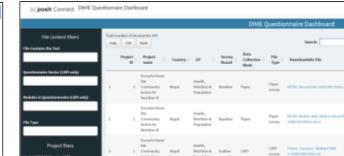
- Installing R in your computer gives you access to its basic functions
- Additionally, you can also install packages. Packages are a collection of R functions that allow you to do:
 - Operations that basic R functions don't do (example: work with geographic data)
 - Operations that basic R functions do, but easier (example: data wrangling)
- They contain code that other R users have prepared for the community.

R Packages

In a nutshell:



...



Basic R

Libraries

Enhanced R capabilities

R Packages

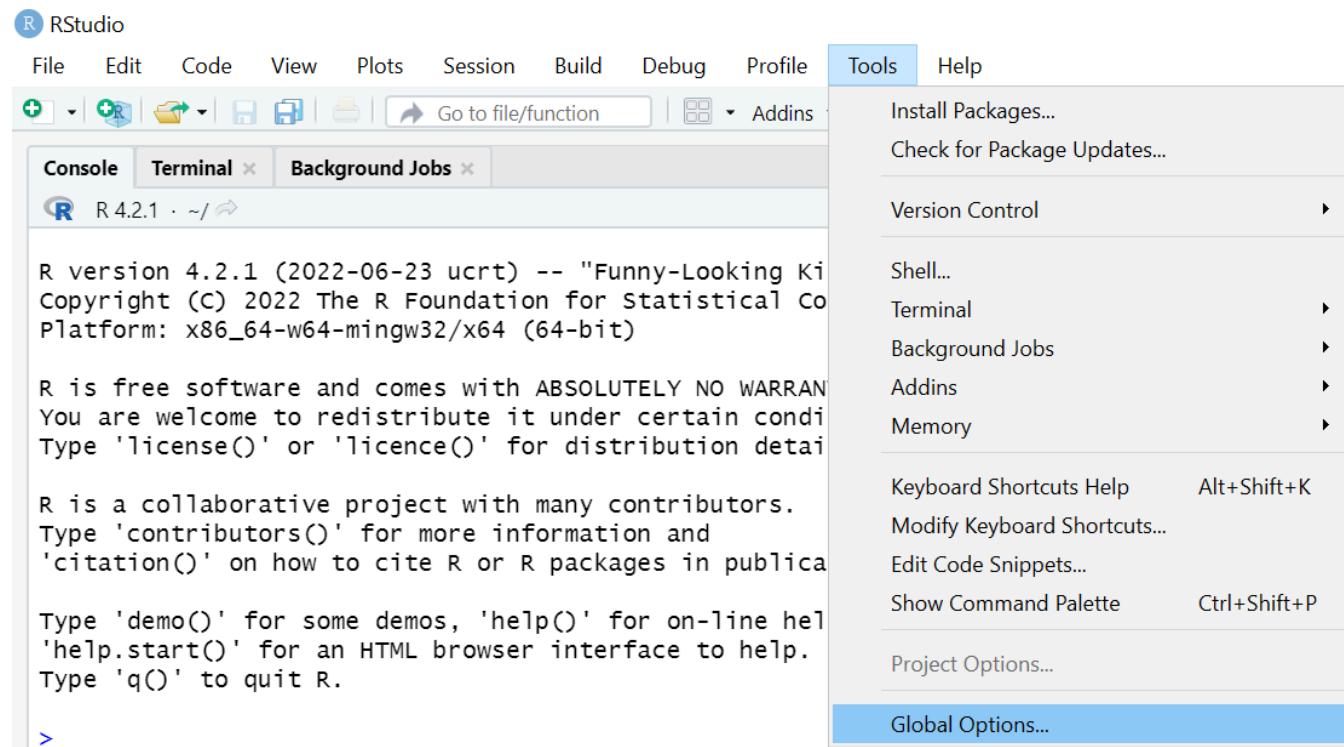
Installing R Packages

- Installing packages is usually simple, but it can be challenging in institutional network connections such as the World Bank.
- The next exercise will set up RStudio so that it can install R Packages without problems

R Packages

Exercise 1: Setting up the installation of packages

1 - In RStudio, go to **Tools** >> **Global Options...**



R Packages

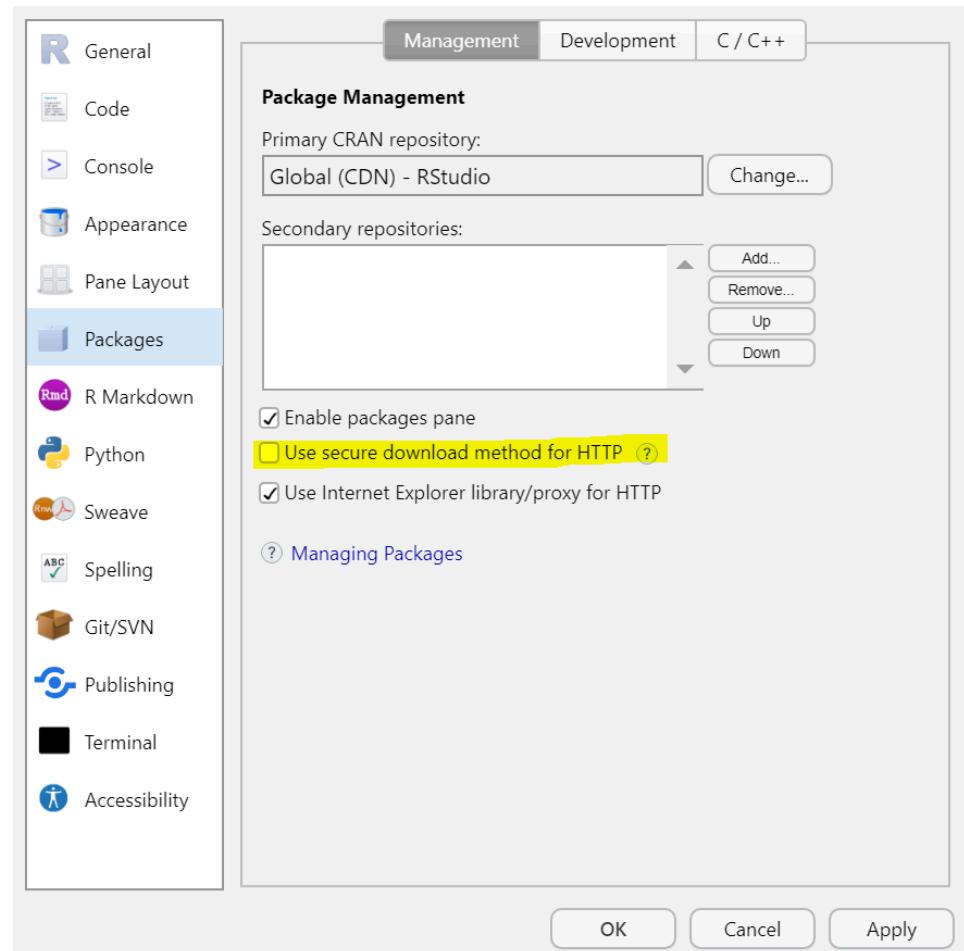
Exercise 1: Setting up the installation of R Packages

2 - Select **Packages** in the left pane

3 - Uncheck **Use secure download method for HTTP**

4 - Click **OK**

You will not see any changes in your RStudio window after this, but now you'll be able to install packages.



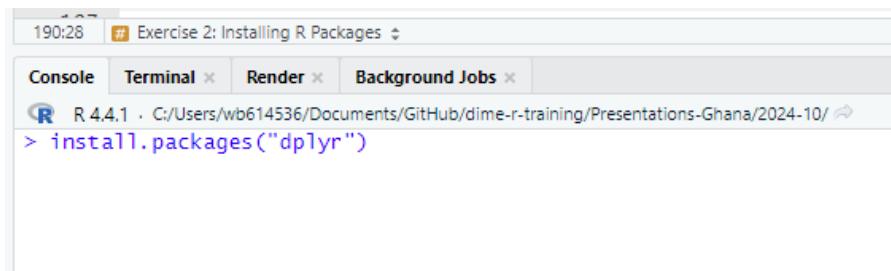
R Packages

We'll use one package in today's session: `dplyr` (really useful library for data wrangling).

Exercise 2: Installing R Packages

1. Install the R Packages by using `install.packages()`

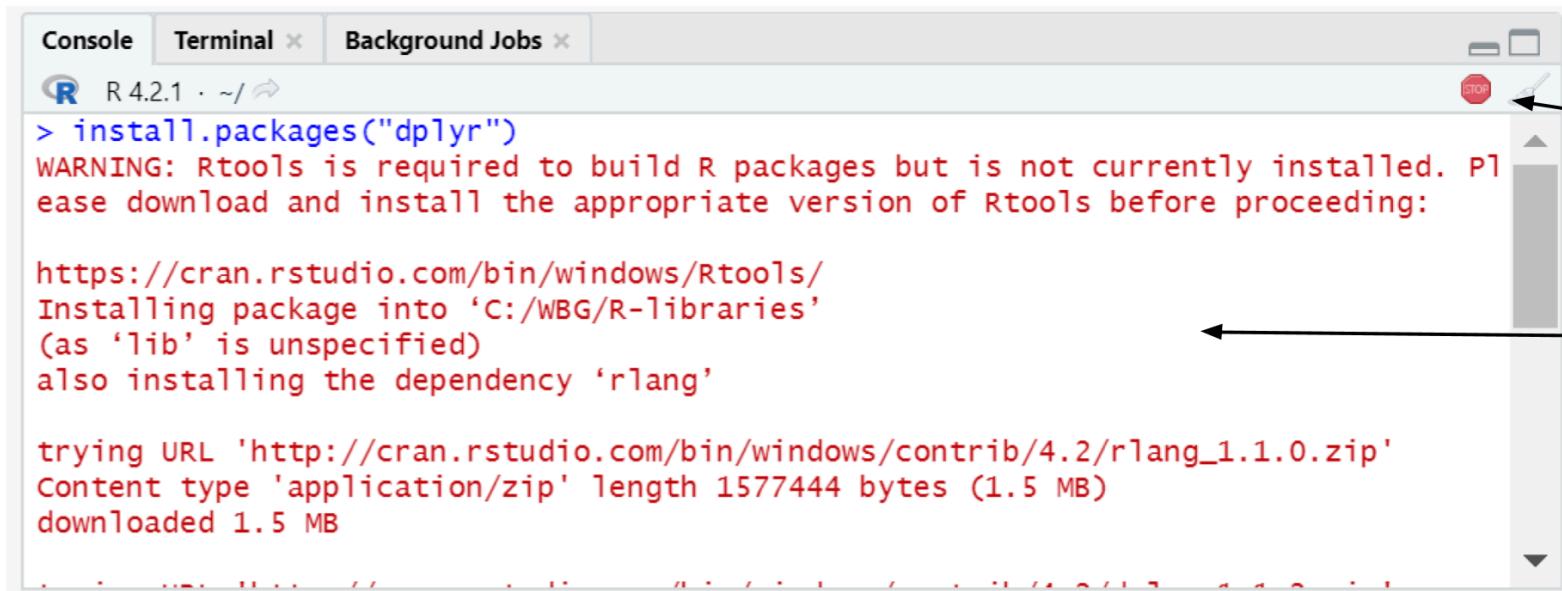
- o `install.packages("dplyr")`
- o Note the quotes (" ") in the packages names
- o **Introduce this code in the console**, not the script panel



A screenshot of the RStudio interface showing the Console tab selected. The title bar says "190:28 # Exercise 2: Installing R Packages". The console window shows the command `> install.packages("dplyr")` in blue, indicating it is being typed or has been run.

R packages

Installing packages



The screenshot shows the RStudio interface with the 'Console' tab selected. The console window displays the following text:

```
R 4.2.1 · ~/ 
> install.packages("dplyr")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/WBG/R-libraries'
(as 'lib' is unspecified)
also installing the dependency 'rlang'

trying URL 'http://cran.rstudio.com/bin/windows/contrib/4.2/rlang_1.1.0.zip'
Content type 'application/zip' length 1577444 bytes (1.5 MB)
downloaded 1.5 MB
```

The “STOP” sign means that the code is still running, just wait until it finishes

Note that this message is
not an error

R packages

Now that `dplyr` and `openxlsx` are installed, we only need to load them to start using the functions they have.

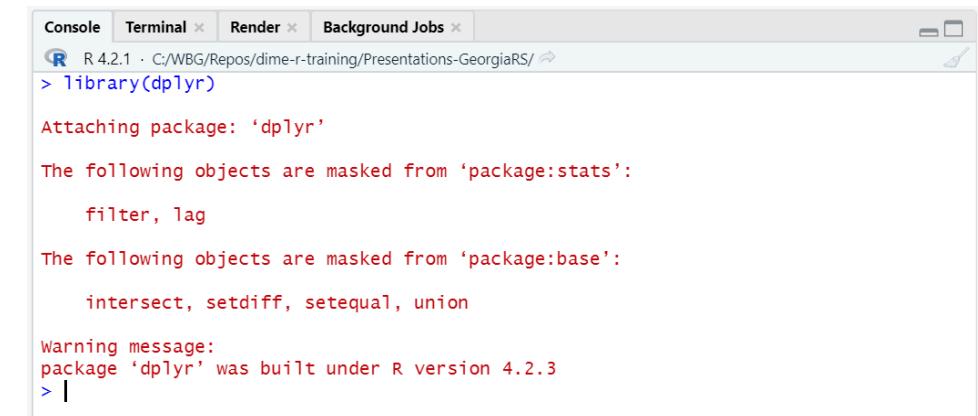
Exercise 3: Loading packages

1. Open a new script with `File` >> `New File` >> `R Script`

2. Load the packages with:

```
library(dplyr)
```

- Run this code from the new script you just opened
- Notice that we don't use quotes in the package names this time



The screenshot shows the RStudio interface with the 'Console' tab selected. The console window displays the following text:

```
Console Terminal × Render × Background Jobs ×
R 4.2.1 · C:/WBG/Repos/dime-r-training/Presentations-GeorgiaRS/
> library(dplyr)
Attaching package: 'dplyr'

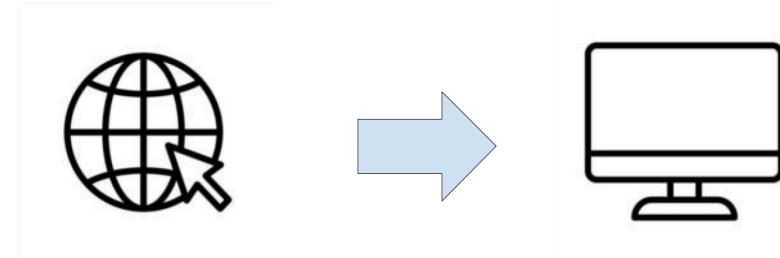
The following objects are masked from 'package:stats':
  filter, lag

The following objects are masked from 'package:base':
  intersect, setdiff, setequal, union

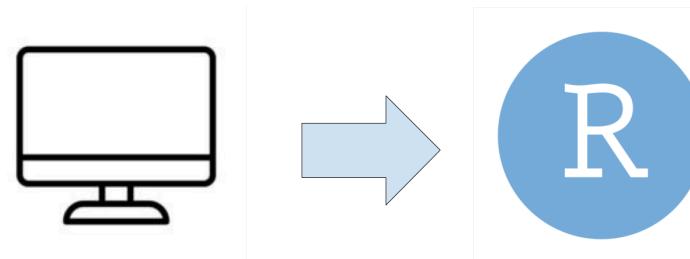
Warning message:
package 'dplyr' was built under R version 4.2.3
> |
```

R packages

- Library installation:



- Library loading:



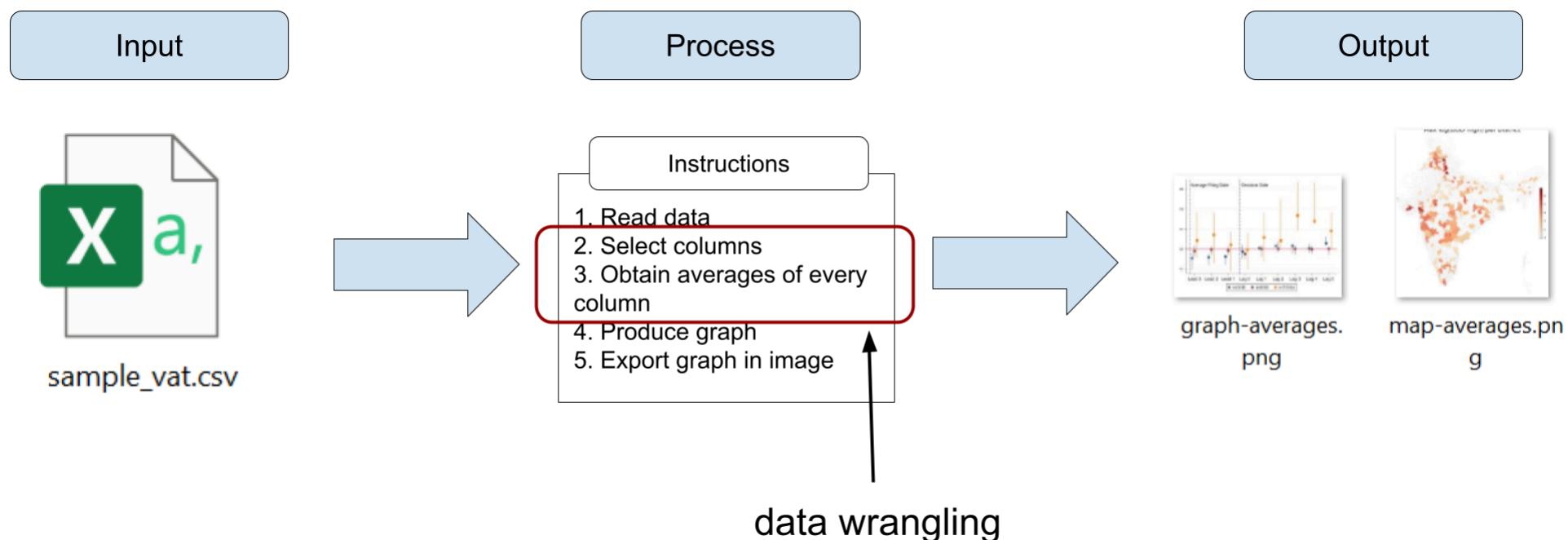
- You install R packages only once in your computer
- You load packages every time you open a new RStudio window (only load the packages you will use)

Data wrangling

Data wrangling

Getting your data ready

- Data is rarely in a format where it can be converted in an output right away
- In statistical programming, the process of transforming data into a condition where it's ready to be converted into an output is called **data wrangling**



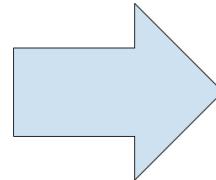
Data wrangling

Getting your data ready

- Data wrangling is one of the most crucial and time-consuming aspects of data work
- It involves not only coding, but also the mental exercise of thinking what is the shape and condition that your dataframe needs to have in order to produce your desired output

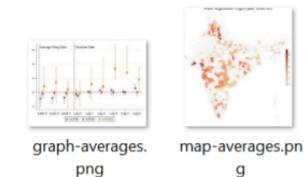
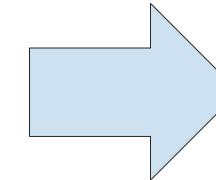
	Modified.ID	TaxPeriod	turnover_taxable18	turnover_exempt	turnover_exports	vat_liability
1	22172	201901	13444.64	0.00	0	2420.04
2	4592952	201906	0.00	0.00	0	0.00
3	10046564	201905	13237.30	0.00	0	2382.71
4	4797328	201912	164477.88	0.00	0	33206.02
5	5889624	201904	0.00	0.00	0	0.00
6	12233616	201912	0.00	0.00	0	0.00
7	6248240	201904	2416.93	0.00	0	435.05
8	538420	201905	7090.50	0.00	0	1276.29
9	10072288	201905	8039.20	0.00	0	1447.06

initial dataframe



	Modified.ID	TaxPeriod	turnover_taxable18	turnover_exempt	turnover_exports	vat_liability
1	22172					2420.04
2	4592952					0.00
3	10046564					2382.71
4	4797328					33206.02
5	5889624					0.00
6	12233616					0.00
7	6248240					435.05
8	538420	201905	7090.50	0.00	0	1276.29
9	10072288	201905	8039.20	0.00	0	1447.06

dataframe needed for outputs



outputs

Data wrangling

Getting your data ready

- As we said before we'll use `dplyr` and `tidyverse` for data wrangling in this training
- You can also use basic R, but we recommend these packages because its functions are easier to use



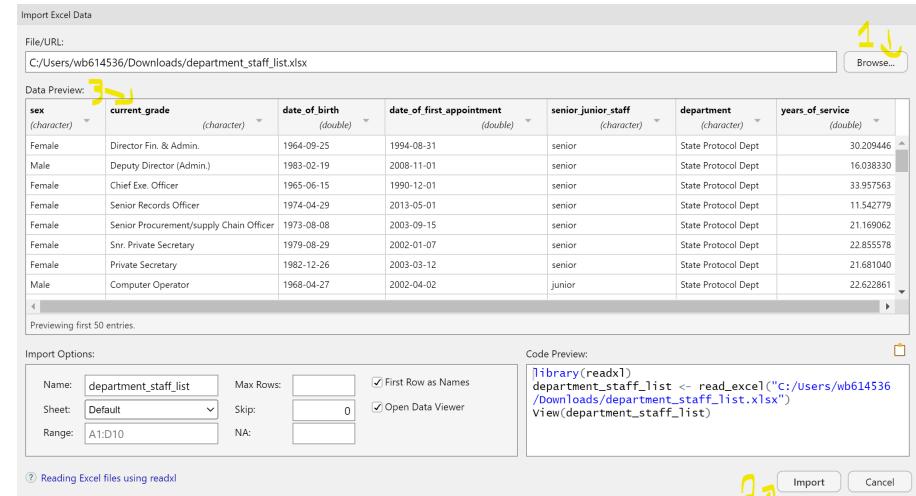
Data wrangling

Exercise 4: Loading data

Note that this part of this is the same exercise we did in session 1, but it's okay to repeat it in order to start using a new RStudio session. **If you have RStudio open, start by closing the window and opening RStudio again.**

1. In your new RStudio window, go to **File > Import Dataset > From Text(base)** and select again the file **department_staff_list.csv**

- o if you don't know where the file is, check in the **Downloads** folder
- o if you need to download it again, it's here:
<https://osf.io/2apht>



sex	current_grade	date_of_birth	date_of_first_appointment	senior_junior_staff	department	years_of_service
(character)	(character)	(double)	(double)	(character)	(character)	(double)
Female	Director Fin. & Admin.	1964-09-25	1994-08-31	senior	State Protocol Dept	30.209446
Male	Deputy Director (Admin.)	1983-02-19	2008-11-01	senior	State Protocol Dept	16.038330
Female	Chief Exec. Officer	1965-06-15	1990-12-01	senior	State Protocol Dept	33.957563
Female	Senior Records Officer	1974-04-29	2013-05-01	senior	State Protocol Dept	11.542779
Female	Senior Procurement/supply Chain Officer	1973-08-08	2003-09-15	senior	State Protocol Dept	21.169062
Female	Snr. Private Secretary	1979-08-29	2002-01-07	senior	State Protocol Dept	22.855578
Female	Private Secretary	1982-12-26	2003-03-12	senior	State Protocol Dept	21.681040
Male	Computer Operator	1968-04-27	2002-04-02	junior	State Protocol Dept	22.622861

2. Make sure to select **Heading > Yes** in the next window
3. Select **Import**
4. Download this new file: <https://osf.io/v6psa> and repeat steps 1-3 with it

Data wrangling

The screenshot shows the RStudio interface. The top navigation bar includes tabs for Environment, History, Connections, Git, and Tutorial. Below the tabs, there are icons for file operations (Import Dataset, 289 MiB), and dropdown menus for R and Global Environment. The main workspace is titled "Data" and lists two datasets: "department_staff_age" (8714 obs. of 2 variables) and "department_staff_list" (8714 obs. of 6 variables).

department_staff_age	8714 obs. of 2 variables
department_staff_list	8714 obs. of 6 variables

Data wrangling

Note: loading data with a function

- You can also load Excel files with the function `read.csv()` instead of using this point-and-click approach
- The **argument** of `read.csv()` is the path in your computer where your data is. For example

```
department_staff_list <- read.csv("C:/Users/wb614536/Downloads/department_staff_list.csv")
```

- As usual, you need to save the result of `read.csv()` into a dataframe object with the arrow operator (`<-`) for it to be stored in the environment

Data wrangling

Recap: knowing your data

- Dataframe `department_staff_list` is the same dataframe we used last session that contains the data from your department.

```
glimpse(department_staff_list)
```

```
## Rows: 8,714
## Columns: 6
## $ ID                  <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## $ sex                 <chr> "Female", "Male", "Female", "Female", "Female", "F...
## $ current_grade        <chr> "Director Fin. & Admin.", "Deputy Director (Admin...
## $ senior_junior_staff <chr> "senior", "senior", "senior", "senior", "senior", ...
## $ department           <chr> "State Protocol Dept", "State Protocol Dept", "Sta...
## $ years_of_service     <dbl> 30.220397, 16.049281, 33.968515, 11.553730, 21.180...
```

Data wrangling

Recap: knowing your data

- Now we are introducing a second dataset that only has the ID number (one that I invented), and the age of that person.

```
glimpse(department_staff_age)
```

```
## Rows: 8,714  
## Columns: 2  
## $ ID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,...  
## $ age <dbl> 60.15058, 41.74949, 59.43053, 50.55989, 51.28268, 45.22656, 41.900...
```

Data wrangling

- We will only use this second dataframe in one of the next exercises, but we load it now because it's in general a good practice to have data loaded into the memory so it's ready to be used.
- For the next exercises, we will face mention scenarios that could show up in doing your annual reports on in day-to-day operations.
- We will do everything using **functions** from the `dplyr` package, that is already in our environment.

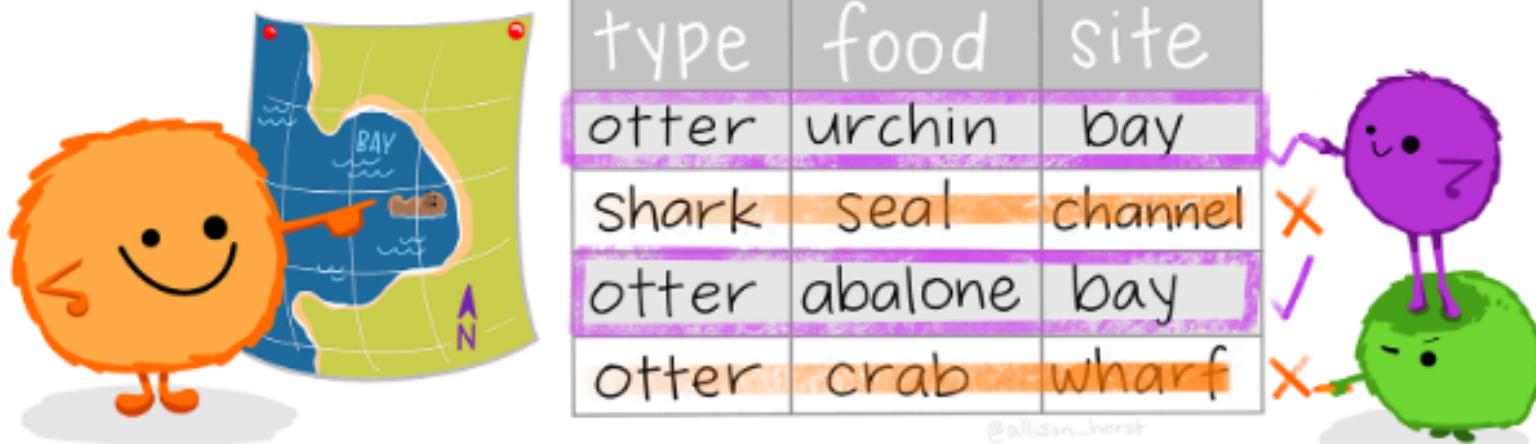
Filtering and sorting

Filtering and sorting

dplyr::filter()

KEEP ROWS THAT
s-a-t-i-s-f-y
your CONDITIONS

keep rows from... this data... ONLY IF... type is "otter" AND site is "bay"
filter(df, type == "otter" & site == "bay")



A cartoon illustration featuring three anthropomorphic shapes: an orange circle, a purple circle, and a green circle. The orange character is pointing at a map of a coastal area with a blue bay labeled "BAY". The purple character is holding a pink checkmark and pointing at a table row. The green character is holding a red X and pointing at another table row. The table contains the following data:

	type	food	site
1	otter	urchin	bay
2	shark	seal	channel
3	otter	abalone	bay
4	otter	crab	wharf

Attributed to @allisonhorst

Filtering and sorting

Data work request

We will use the data from our dataframe
department_staff_list

For this we can filter the Female rows from the sex category

ID	sex	current_grade
1	Female	Director Fin. & Admin.
2	Male	Deputy Director (Admin.)
3	Female	Chief Exe. Officer
4	Female	Senior Records Officer
5	Female	Senior Procurement/supply
6	Female	Snr. Private Secretary
7	Female	Private Secretary
8	Male	Computer Operator
9	Male	Asst. Director IIA
10	Female	Private Secretary
11	Male	IT/IM Officer



ID	sex	current_grade
1	Female	Director Fin. & Admin.
2	Female	Chief Exe. Officer
3	Female	Senior Records Officer
4	Female	Senior Procurement/supply Chain Officer
5	Female	Snr. Private Secretary
6	Female	Private Secretary
7	Female	Private Secretary
8	Female	Stenographer Gd I
9	Female	Records Officer
10	Female	Principal Exe. Officer
11	Female	Steno Secretary
12	Female	senior Technical Asst.

Filtering and sorting

Data work request

Now let's say that we are also interested at a first glance of the females that recently joined the department. We would have to do the following

1. Keeping only the female employees
2. Sorting by years of service

Filtering and sorting

1. Keeping only the female employees

Use `filter()` for this:

Remember how we use functions in this case 1st argument is **data**, second argument is **filter** (sex takes the value female in this case)

```
temp1 <- filter(department_staff_list, sex == "Female")
```

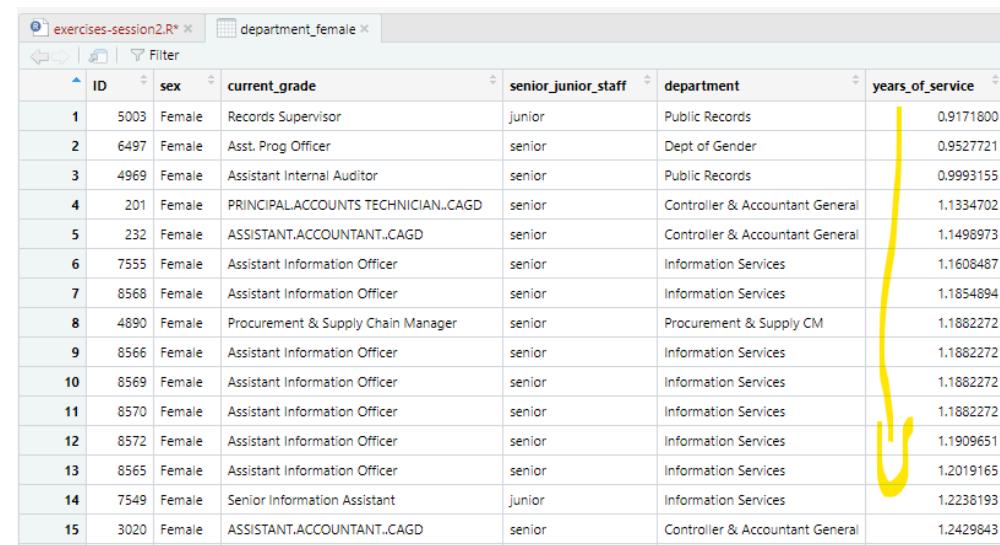
ID	sex	current_grade	senior_junior_staff	department	years_of_service
1	1	Female Director Fin. & Admin.	senior	State Protocol Dept	30.220397
2	3	Female Chief Exe. Officer	senior	State Protocol Dept	33.968515
3	4	Female Senior Records Officer	senior	State Protocol Dept	11.553730
4	5	Female Senior Procurement/supply Chain Officer	senior	State Protocol Dept	21.180014
5	6	Female Snr. Private Secretary	senior	State Protocol Dept	22.866530
6	7	Female Private Secretary	senior	State Protocol Dept	21.691992
7	10	Female Private Secretary	senior	State Protocol Dept	5.056810
8	12	Female Stenographer Gd I	junior	State Protocol Dept	16.249144
9	14	Female Records Officer	senior	State Protocol Dept	6.428474
10	15	Female Principal Exe. Officer	senior	State Protocol Dept	4.908966
11	16	Female Steno Secretary	senior	State Protocol Dept	15.134839
12	17	Female senior Technical Asst.	junior	State Protocol Dept	32.134155
13	22	Female Principal Protocol Officer	senior	State Protocol Dept	22.962355
14	30	Female State Protocol Officer	senior	State Protocol Dept	5.464750
15	31	Female State Protocol Officer	senior	State Protocol Dept	5.464750
16	32	Female State Protocol Officer	senior	State Protocol Dept	5.464750
17	33	Female Assistant Protocol Officer	senior	State Protocol Dept	5.286790

Filtering and sorting

2. Sorting by years of service

Use the function `arrange()` to sort. Sortings are ascending by default in R (this will get us from less years of service to more years of service).

```
department_female <- arrange(temp1, years_of_service) # arrange by years of service
```



The screenshot shows a Jupyter Notebook interface with two tabs: 'exercises-session2.R* x' and 'department_female x'. The 'department_female' tab is active, displaying a table with 15 rows of data. The columns are labeled: ID, sex, current_grade, senior_junior_staff, department, and years_of_service. A yellow vertical arrow points down the 'years_of_service' column, indicating the direction of sorting. The data rows are as follows:

	ID	sex	current_grade	senior_junior_staff	department	years_of_service
1	5003	Female	Records Supervisor	junior	Public Records	0.9171800
2	6497	Female	Asst. Prog Officer	senior	Dept of Gender	0.9527721
3	4969	Female	Assistant Internal Auditor	senior	Public Records	0.9993155
4	201	Female	PRINCIPAL ACCOUNTS TECHNICIAN..CAGD	senior	Controller & Accountant General	1.1334702
5	232	Female	ASSISTANT ACCOUNTANT..CAGD	senior	Controller & Accountant General	1.1498973
6	7555	Female	Assistant Information Officer	senior	Information Services	1.1608487
7	8568	Female	Assistant Information Officer	senior	Information Services	1.1854694
8	4890	Female	Procurement & Supply Chain Manager	senior	Procurement & Supply CM	1.1882272
9	8566	Female	Assistant Information Officer	senior	Information Services	1.1882272
10	8569	Female	Assistant Information Officer	senior	Information Services	1.1882272
11	8570	Female	Assistant Information Officer	senior	Information Services	1.1882272
12	8572	Female	Assistant Information Officer	senior	Information Services	1.1909651
13	8565	Female	Assistant Information Officer	senior	Information Services	1.2019165
14	7549	Female	Senior Information Assistant	junior	Information Services	1.2238193
15	3020	Female	ASSISTANT ACCOUNTANT..CAGD	senior	Controller & Accountant General	1.2429843

Filtering and sorting

Exercise 5: Now let's do it.. filter and sort your data

We can write the whole code for this in our exercise script. (You can copy and paste the code below)

1.- Filter by `Female`:

2.- Sort by `years_of_service`:

```
temp1 <- filter(department_staff_list, sex == "Female") # filter by female  
department_female <- arrange(temp1, years_of_service) # order by years of service
```

Filtering and sorting

Some notes:

- `filter()` and `arrange()` are all functions from `dplyr`. Remember you have to always load `dplyr` first with `library(dplyr)` to be able to use them
- The resulting dataframe is `department_female` (and we also have `temp1`)

Data	
▶ <code>department_female</code>	3559 obs. of 6 variables
▶ <code>department_staff_age</code>	8714 obs. of 2 variables
▶ <code>department_staff_list</code>	8714 obs. of 6 variables
▶ <code>temp1</code>	3559 obs. of 6 variables

Filtering and sorting

- Filtering and sorting are two very common data wrangling operations in statistical programming
- Now we'll review a new data wrangling operation that is also quite common and useful: **mutate**, which basically means creating new variables

Mutate

Mutate

- `mutate` will take a statement like this:

```
mutate(variable_name = some_calculation)
```

- And attach variable_name at the end of the dataset.



Mutate

Example

Using our data frame let's say that we want to include a variable that instead of years of service we want days of service. We would do this by doing the following:

```
example_mutate <- mutate(department_staff_list, days_of_service = years_of_service*365)
```

We will not use that variable for our datawork examples, but this is a really useful data wrangling function.

Questions?



Merging dataframes

Merging dataframes

Merging data is a common task in data analysis, especially when working with data from multiple departments.

Let's see how it would apply to our dataframes.

Data work request

Scenario 2:

Let's imagine that for our annual report we are also interested in the age distribution from the employees but these are in different datasets

Merging dataframes

Data work request

Use the data `department_staff_list` that you already know with the `department_staff_age`

We want to include the `age` variable to our dataframe

ID	sex	current_grade	senior_junior_staff	department	years_of_service
1	Female	Director Fin. & Admin.	senior	State Protocol Dept	30.220397
2	Male	Deputy Director (Admin.)	senior	State Protocol Dept	16.049281
3	Female	Chief Exe. Officer	senior	State Protocol Dept	33.968515
4	Female	Senior Records Officer	senior	State Protocol Dept	11.553730
5	Female	Senior Procurement/supply Chain Officer	senior	State Protocol Dept	21.180014
6	Female	Snr. Private Secretary	senior	State Protocol Dept	22.866530
7	Female	Private Secretary	senior	State Protocol Dept	21.691992
8	Male	Computer Operator	junior	State Protocol Dept	22.633812
9	Male	Asst. Director IIA	senior	State Protocol Dept	5.379877
10	Female	Private Secretary	senior	State Protocol Dept	5.056810
11	Male	IT/IM Officer	senior	State Protocol Dept	5.043121
12	Female	Stenographer Gd I	junior	State Protocol Dept	16.249144
13	Male	Records Supervisor	junior	State Protocol Dept	12.919918
14	Female	Records Officer	senior	State Protocol Dept	6.428474
15	Female	Principal Exe. Officer	senior	State Protocol Dept	4.908966
16	Female	Steno Secretary	senior	State Protocol Dept	15.134839



ID	age
1	60.15058
2	41.74949
3	59.43053
4	50.55989
5	51.28268
6	45.22656
7	41.90007
8	56.56400
9	29.09788
10	38.76523
11	32.14784
12	38.01506
13	34.16564
14	43.42505
15	44.99932
16	43.50171

Merging dataframes

To do this we will use `left_join()` to merge the dataframes:

- The arguments of the function are 1. the "principal" dataset and 2. the dataframe we want to merge (paste) into that one

```
department_age <- left_join(department_staff_list, department_staff_age) # Our original data frame
```

`left_join(x, y)`

x1	1
x2	2
x3	3

NA	4	2	1
NA	y3	y2	y1



key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA

Merging dataframes

Exercise 6: Now let's do it.

- You can copy and paste the code to your exercise script.

```
deparment_age <- left_join(deparment_staff_list, department_staff_age) # Our original data frame
```

The screenshot shows the RStudio interface. On the left, the 'Environment' tab is selected in the top navigation bar. Below it, the 'Data' section lists four datasets:

Dataset	Observations	Variables
deparment_age	8714 obs.	7 variables
department_female	3559 obs.	6 variables
department_staff_age	8714 obs.	2 variables
department_staff_list	8714 obs.	6 variables

On the right, the 'Global Environment' pane shows the same four datasets. In the center, the 'deparment_age' data frame is displayed as a table with columns: current_grade, senior_junior_staff, department, years_of_service, and age. The table contains 15 rows of data.

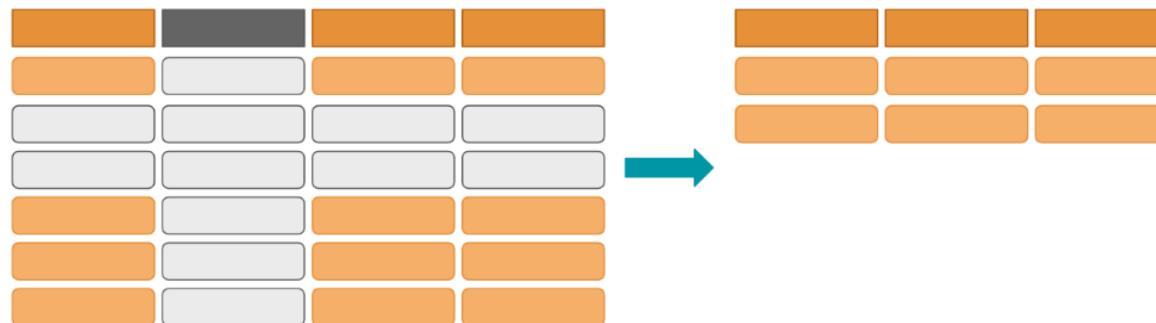
current_grade	senior_junior_staff	department	years_of_service	age
Director Fin. & Admin.	senior	State Protocol Dept	30.220397	60.15058
Deputy Director (Admin.)	senior	State Protocol Dept	16.049281	41.74949
Chief Exe. Officer	senior	State Protocol Dept	33.968515	59.43053
Senior Records Officer	senior	State Protocol Dept	11.553730	50.55989
Senior Procurement/supply Chain Officer	senior	State Protocol Dept	21.180014	51.28268
Snr. Private Secretary	senior	State Protocol Dept	22.866530	45.22656
Private Secretary	senior	State Protocol Dept	21.691992	41.90007
Computer Operator	junior	State Protocol Dept	22.633812	56.56400
Asst. Director IIA	senior	State Protocol Dept	5.379877	29.09768
Private Secretary	senior	State Protocol Dept	5.056810	38.76523
IT/IM Officer	senior	State Protocol Dept	5.043121	32.14784
Stenographer Gd I	junior	State Protocol Dept	16.249144	38.01506
Records Supervisor	junior	State Protocol Dept	12.919918	34.16564

group_by() and summarize()

We use this when we want to aggregate your data (by groups).

This is one of the most commons operations. Sometimes we want to calculate statistics by groups

Customize with **group_by()** and **summarize()**



Grouping and summarizing

Data work request

Let's say that we are interested again in the number of employees by department

We will use our department_staff_list dataset to do this.

1. The first step is to define the group, in this case `department`

```
employees_by_department <- group_by(department_staff_list, department)
```

If we **only** do this, this won't do anything, to complete the function we need to use this with `summarise()`

Grouping and summarizing

Data work request

summarize works in a similar way to mutate:

```
variable_name = some_calculation
```

In this case the `some_calculation` will be to count the number of employees

```
temp1 <- group_by(department_staff_list, department)
employees_by_department <- summarise(temp1, number = n())
```

Grouping and summarizing

Data work request

This will create the following dataframe/table:

```
## # A tibble: 23 × 2
##   department      number
##   <chr>          <int>
## 1 Births & Deaths     172
## 2 Bureau of Languages    58
## 3 Co-operatives       293
## 4 Controller & Accountant General 3776
## 5 Department of Comm Devt    109
## 6 Dept of Children        101
## 7 Dept of Gender         100
## 8 Dept of Labour        290
## 9 Factories Inspectorate 121
## 10 Feeder Roads        102
## # i 13 more rows
```

More wrangling operations

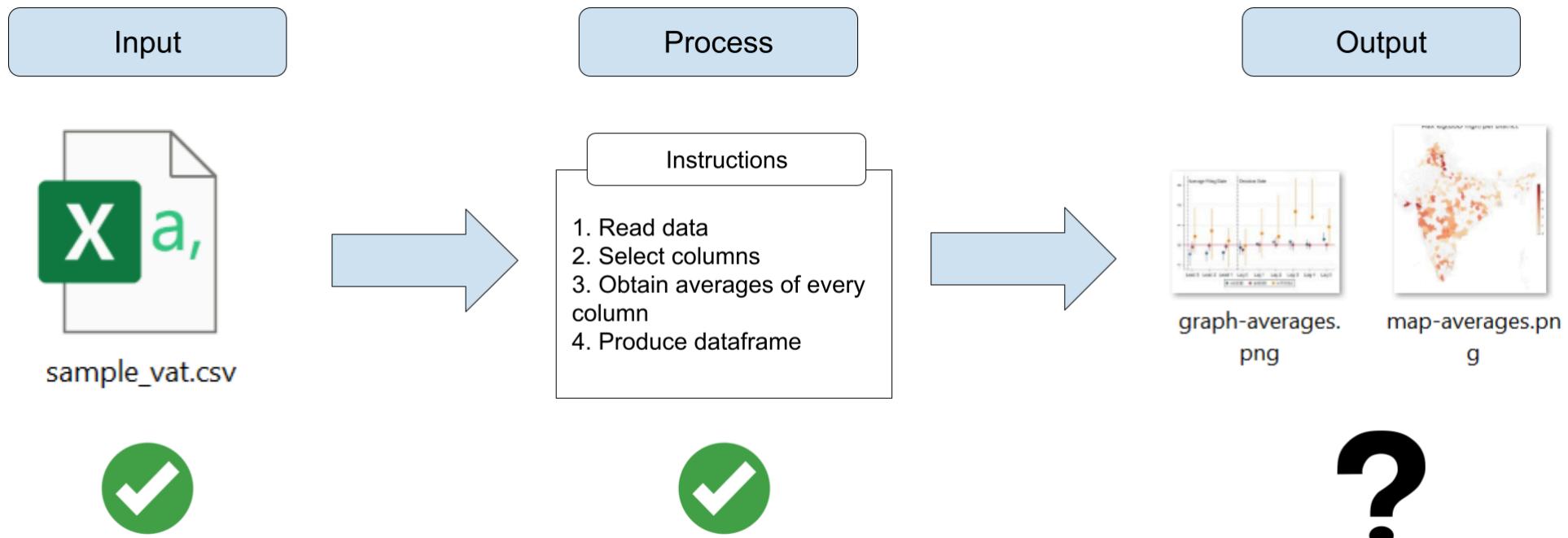
These were two examples we chose to show different possible data wrangling operations. A summary of these and other common operations are:

Operation	Function in <code>dplyr</code>
Subset columns	<code>select()</code>
Subset rows (based on condition)	<code>filter()</code>
Create new columns	<code>mutate()</code>
Create new columns based on condition	<code>mutate()</code> and <code>case_when()</code>
Create new rows	<code>add_row()</code>
Merge dataframes	<code>inner_join()</code> , <code>left_join()</code> , <code>right_join()</code> , <code>full_join()</code>
Append dataframes	<code>bind_rows()</code>
Deduplicate	<code>distinct()</code>
Collapse and create summary indicators	<code>group_by()</code> , <code>summarize()</code>
Pass a result as the first argument for the next function	<code>%>%</code> (operator, not function (tomorrow))

Exporting outputs

Exporting outputs

- Until now, we've seen full examples of part 1 and 2 of the data work pipeline
- What about exporting outputs?



- We'll see this in the next exercise

Exporting outputs

Exporting dataframes

- We can export it as a csv file with the function `write.csv()`
- `write.csv()` creates a csv file with the dataframe
- It takes two basic arguments:
 1. The name of the object you want to export
 2. A file path to export the object to

Exporting outputs

Exercise 7: Export `employees_by_department`

1. Use this code to export the results of the last two exercises:

```
write.csv(employees_by_department, "employees_by_department", row.names = FALSE)
```

Exporting outputs

Now `employees_by_department.csv` (probably in your `Documents` folder).

Name	Date modified ^	Type	Size
.Rproj.user	9/19/2023 2:37 AM	File folder	
1-introduction-to-r_cache	9/19/2023 3:30 AM	File folder	
img	9/19/2023 3:43 PM	File folder	
2-data-wrangling_cache	9/19/2023 3:46 PM	File folder	
data	9/20/2023 12:11 AM	File folder	
libs	9/20/2023 5:09 AM	File folder	
.Rhistory	9/19/2023 4:20 PM	RHISTORY File	1 KB
1-introduction-to-r.Rmd	9/20/2023 1:34 AM	RMD File	24 KB
1-introduction-to-r.html	9/20/2023 1:34 AM	Chrome HTML Docu...	30 KB
exercises-session1.R	9/20/2023 1:41 AM	R File	1 KB
202309.Rproj	9/20/2023 1:58 AM	R Project	1 KB
exercises-session2.R	9/20/2023 4:37 AM	R File	1 KB
2-data-wrangling.Rmd	9/20/2023 5:09 AM	RMD File	25 KB
2-data-wrangling.html	9/20/2023 5:09 AM	Chrome HTML Docu...	30 KB
total_income.csv	9/20/2023 5:11 AM	Microsoft Excel Com...	1 KB
df_tbilisi_50.csv	9/20/2023 5:12 AM	Microsoft Excel Com...	2 KB

Exporting outputs

Some notes on file paths

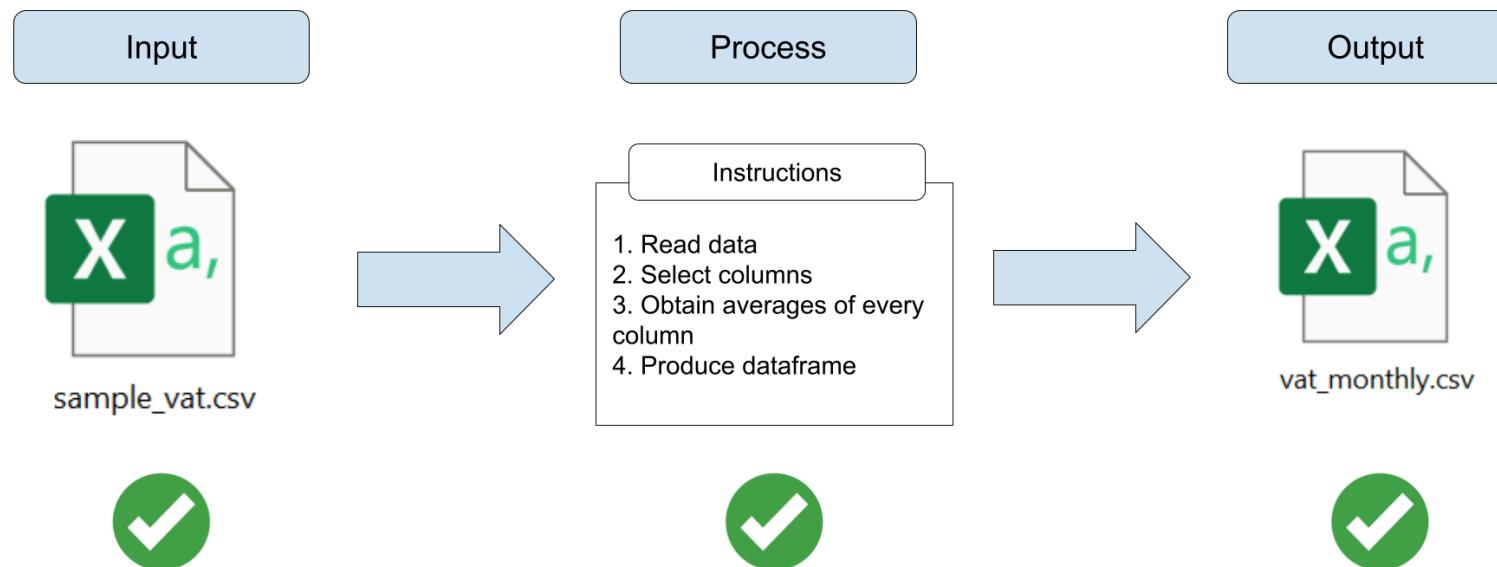
- The second argument of `write.csv()` specifies the file path we export the dataframe to

```
write.csv(employees_by_department, "employees_by_department", row.names = FALSE)
```

- You can include any path in your computer and R will write the file in that location
 - For example: `"C:/Users/wb614536/OneDrive - WBG/Desktop"` exports the file to the desktop of my computer (this will not work in other computers)
 - Note that file paths in R use forward slashes (`/`). Back slashes (`\`) **do not work in R**

Exporting outputs

Our data pipeline has been fully implemented at this point. Great!

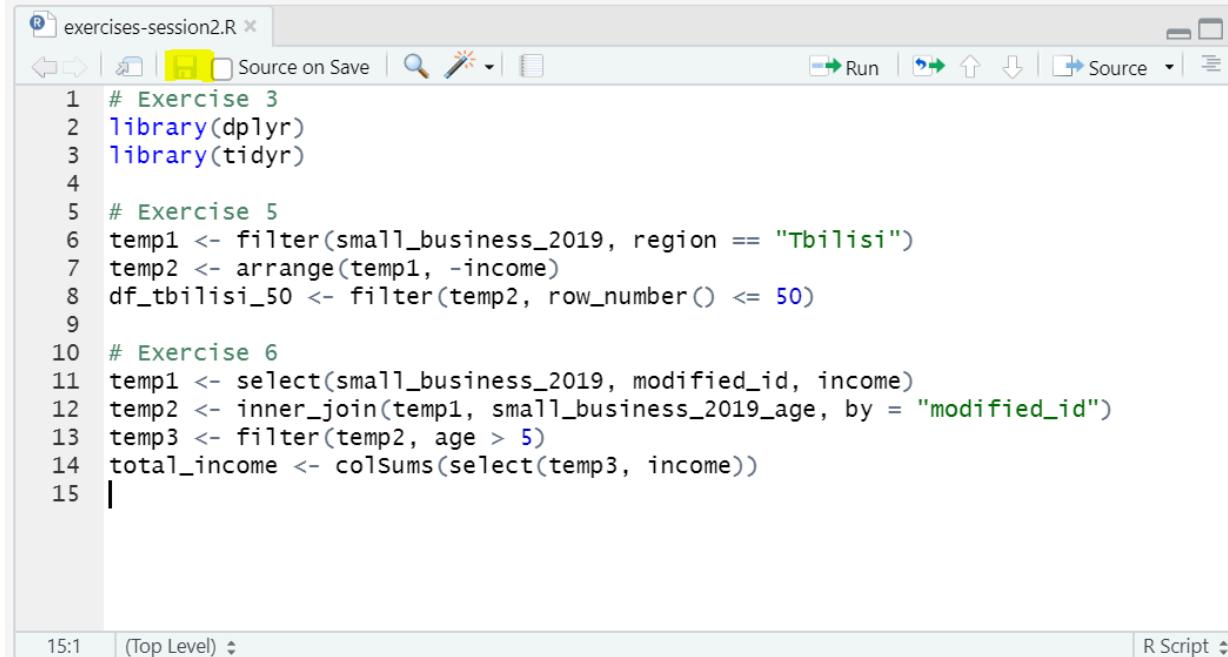


Wrapping up

Wrapping up

Don't forget to save your work!

- If you haven't, add code comments with `#` to differentiate your solutions for each exercise
- Click the floppy disk to save your work
- Make sure to remember where you're saving your file



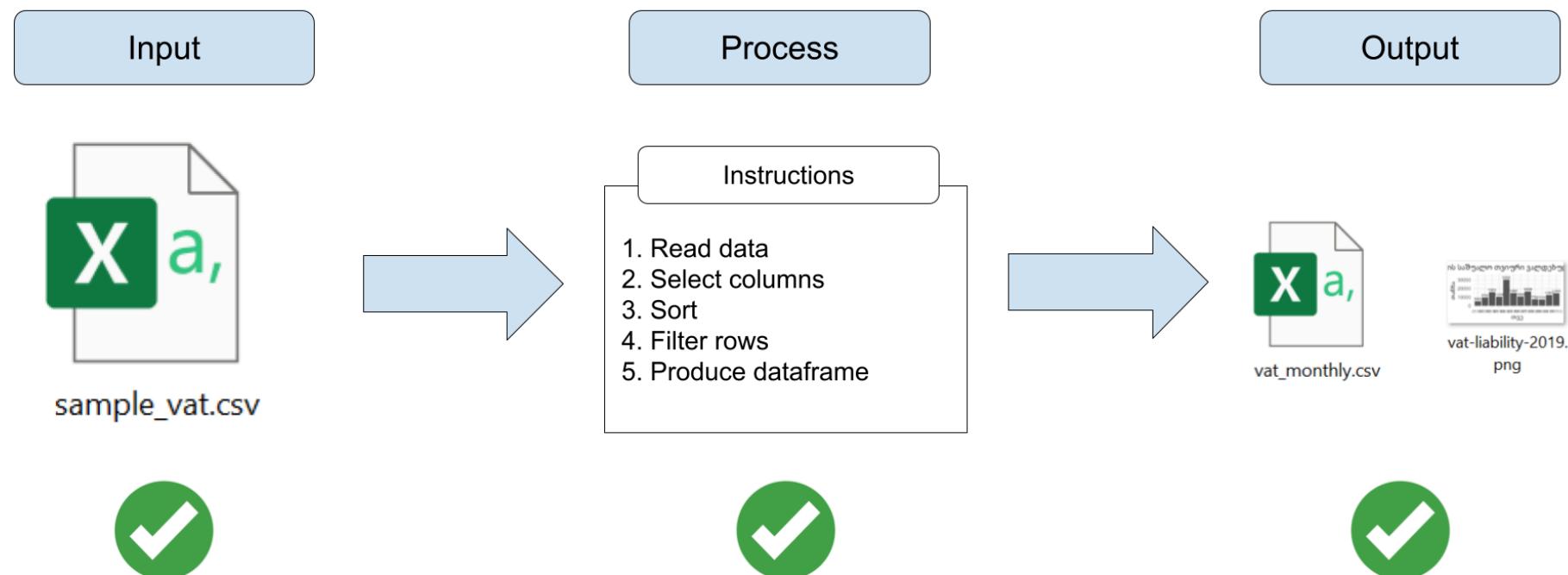
The screenshot shows an RStudio interface with an R script file open. The title bar says "exercises-session2.R x". The code in the editor is:

```
1 # Exercise 3
2 library(dplyr)
3 library(tidyr)
4
5 # Exercise 5
6 temp1 <- filter(small_business_2019, region == "Tbilisi")
7 temp2 <- arrange(temp1, -income)
8 df_tbilisi_50 <- filter(temp2, row_number() <= 50)
9
10 # Exercise 6
11 temp1 <- select(small_business_2019, modified_id, income)
12 temp2 <- inner_join(temp1, small_business_2019_age, by = "modified_id")
13 temp3 <- filter(temp2, age > 5)
14 total_income <- colsums(select(temp3, income))
15 |
```

The status bar at the bottom left shows "15:1 (Top Level)". The status bar at the bottom right shows "R Script".

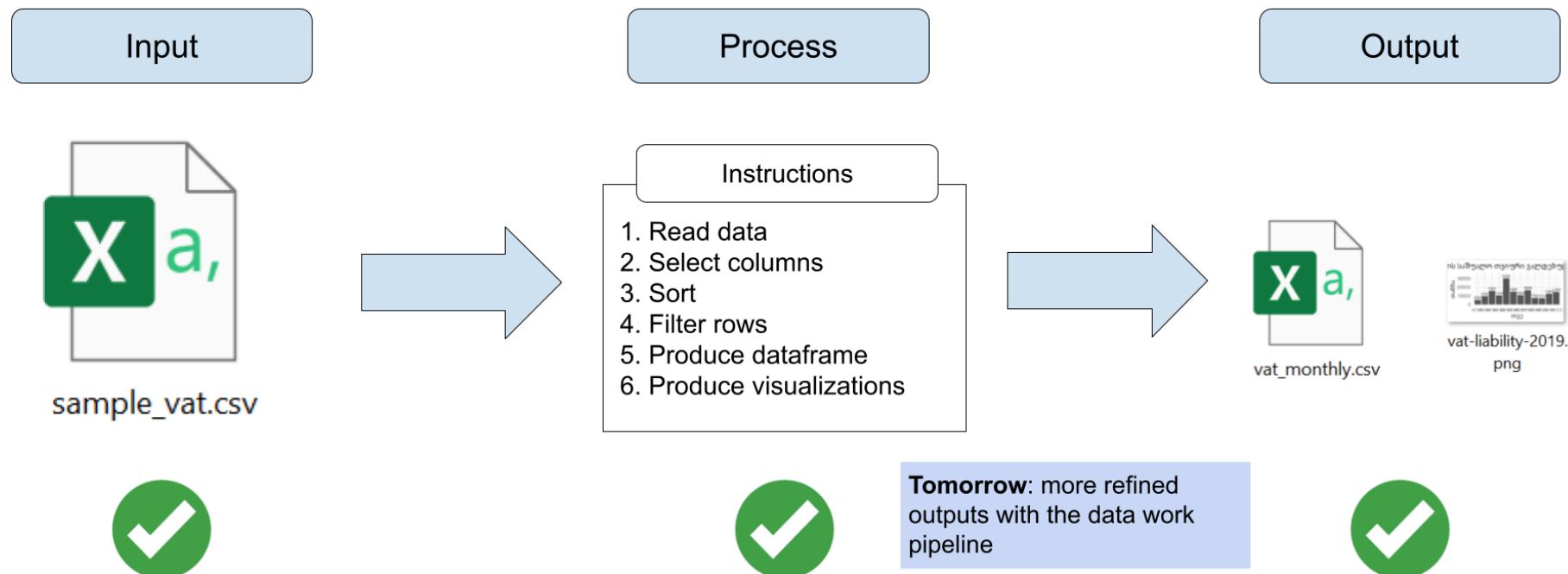
Wrapping up

Data work pipeline



Wrapping up

Data work pipeline



Thanks! // ¡Gracias! // Obrigado!



Appendix

Appendix

Keeping only relevant columns

Imaging I only want a list with ID and department.

Use `select()` for this:

```
temp1 <- select(department_staff_list, ID, department)
```

Appendix

Calculating aggregated columns: total and average income

Use `summarize()` combined with `sum()` and `mean()`:

For our current data this does not apply, but this will be really useful for budget dataframes.

This would look something like this

```
budget_2024 <- group_by(budget_data, year, product)

budget_2024 <- summarise(budget_2024,
                          total = sum(income),
                          average = mean(income))
```

Useful links

- **R for Data Science** by Hadley Wickham and Garrett Grolemund
 - Comprehensive introduction to data wrangling and visualization.
 - [Read it online for free.](#)
- **Tidyverse Cookbook**
 - Practical solutions for common data wrangling tasks.
 - [Tidyverse Cookbook GitHub.](#)
- **Tidyverse Cheat Sheets**
 - Official cheat sheets for `dplyr`, `tidyverse`, and other Tidyverse packages.
 - [Tidyverse Resources.](#)