Lab 2: Julia Quickstart

Functions, Logic, and Packages

CEVE 421/521

Fri., Jan. 19

In this lab we will learn how to work with tabular data in Julia. Specifically, you will get some experience using:

1. `DataFrames.jl` to store tabular data as a DataFrame
2. `CSV.jl` to read CSV files and convert them to DataFrames
3. `DataFramesMeta.jl` to manipulate DataFrames
4. `Plots.jl` and `StatsPlots.jl` to create visualizations

For those of you who took CEVE 543, you'll find most of this familiar! If you find this challenging (e.g., if you're new to programming) please look at the Resources page for some tutorials.

> **!** Instructions
>
> Much of this lab is example code with some narration to give you a sense of what's going on. However, when you see a box like this it means you need to do something!

As with Lab 01, you should push your final code to GitHub and submit your rendered PDF or DOCX file to Canvas.

0.1 Setup

Here are some instructions for getting this lab working.

0.1.1 Clone the repository

First, you'll need to clone this repository to your computer. As with Lab 01, I recommend to use GitHub Desktop or the built-in Git support in VS Code. Remember to use the link from Canvas (`classroom.github.com/...`).

Next, open the repository in VS Code (you can do this directly from GitHub desktop if you'd like). All instructions from here assume you're in the *root directory* of the repository.

0.1.2 Install required packages

As we saw in Lab 01, Julia is a modular language with code in packages. Compared to a language like Python, the packages in Julia typically have a narrower scope (for example, instead of a single Pandas package that does everything, there are separate packages for reading CSV files, defining

dataframes, using clear syntax for data manipulation, etc.). When we're working with a new lab, we'll need to first install the packages we need.

1. Open the command palette and select `Julia: Start REPL`
2. In the Julia REPL, type `]` to enter package manager mode
3. Type `activate .` to activate the project environment
4. Type `instantiate` to install the packages listed in the `Project.toml` file. This may take a few minutes.[1]

0.1.3 Looking ahead

In the future, you'll repeat these steps *for every lab*:

1. `clone` the repository to your computer
2. `activate` the project environment
3. `instantiate` the packages
4. make your changes, saving and `commit`ing regularly as you go
5. `push` your changes to GitHub (you don't have to wait until the end for this – you can `push` multiple times)

0.2 Refresher: Quarto basics

Before diving in, let's quickly review some Quarto basics. As we saw in the last lab, Quarto is a program that lets you combine text, code, and output in a single document. Quarto files are just text files, typically with the file extension `.qmd`.

By default, all the text in a Quarto file is interpreted as Markdown, a simple markup language for formatting text. You've probably seen Markdown before. You can create headers with `##` (for a section), `###` (for a subsection), and so on. You can make text *italic* with `*italic*` and **bold** with `**bold**`. For more, you can learn more about it here.

When you're authoring your labs, you should take advantage of Markdown features!

0.2.1 Document metadata

If you open a Quarto file in your text editor (e.g., VS Code) or look at it on GitHub, you'll see that the file starts with some *metadata*. The metadata is a set of key-value pairs that tell Quarto how to render the document. In Lab 01, you edited the `author` field to include your name.

0.2.2 LaTeX Math

As in standard Pandoc markdown, you can use LaTeX math in Quarto. For example, `$\alpha$` yields α. You can also use `$$` to create a block equation:

```
$$
P(E) = \{ n \choose k \} p ^k (2-p) ^ {n-k}
$$
```

renders as

[1] Julia precompiles packages when they are installed, and (to a lesser extent) when they are first used. The first time you use a package it may take a moment to load. This is normal, nothing to worry about, and rapidly improving.

$$P(E) = \binom{n}{k} p^k (2-p)^{n-k}$$

For more, see the "Typesetting Math" section of the resources page.

0.2.3 Source code

Sometimes we want to provide example code in our documents. This is code that is not meant to be run, but is just there to illustrate a point. We do that by wrapping the code in ```. For example:

````
```
f(x) = 1.25 * sin(2  * x / 1.5 + 0.5) + 0.25
f(2.1)
```
````

yields

```
f(x) = 1.25 * sin(2  * x / 1.5 + 0.5) + 0.25
f(2.1)
```

You will typically want to specify the language of the code block, which will tell Quarto how to syntax highlight it. For example, see how the highlighting changes when we specify `julia`:

````
```julia
f(x) = 1.25 * sin(2  * x / 1.5 + 0.5) + 0.25
f(2.1)
```
````

```julia
f(x) = 1.25 * sin(2  * x / 1.5 + 0.5) + 0.25
f(2.1)
```

0.2.4 Code blocks

Often, we don't just want to show code, but we want to run it and show the output.

````
```{julia}
f(x) = 1.25 * sin(2  * x / 1.5 + 0.5) + 0.25
f(2.1)
```
````

which yields

```julia
f(x) = 1.25 * sin(2  * x / 1.5 + 0.5) + 0.25
f(2.1)
```

```
0.4099583491000567
```

You can run these blocks in Julia by clicking the "Run Cell" button, or by pressing the keyboard shortcut (to see it, open the command palette and search for "Run Cell"). For more on Julia, see here.

0.2.5 Citations

You can add citations in Quarto. The easiest way is to export a bibliography from Zotero, and then add it to your Quarto document. You can use the Zotero Better BibTeX plugin to export a `.bib` file.

See here for instructions on using references with Quarto or see the website code for an example. I'll provide a template for your final project.

0.3 Julia Quickstart

0.3.1 Loading packages

In Julia we say `using` to import a package. By convention we'll put these at the top of our script or notebook in alphabetical order. When you run this cell, you'll see a bunch of activity in your REPL as Julia goes through the following steps:

1. Download a file from the internet that specifies which packages depend on which other packages
2. Solve an optimization problem to identify which versions of which packages (including dependencies, and their dependencies, and so on) are compatible with each other
3. Download the packages and compile them (this may take a few minutes)

```
using CSV
using DataFrames
using DataFramesMeta
using Dates
using Plots
using StatsBase: mean
using StatsPlots
using Unitful
```

0.3.2 Read in data

We will use the `CSV.jl` package to read in our data.

> **Tip**
>
> Hover over the numbers on the right of this code for explanations.

```
fname = "data/tidesandcurrents-8638610-1928-NAVD-GMT-metric.csv"  ①
df = CSV.read(fname, DataFrame)                                   ②
first(df, 5)                                                      ③
```

① We define a variable called `fname` that stores the path to our data file. The `data` folder is in the same directory as this notebook.

② We use the `CSV.read` function to read in the data. The first argument is the filename, and the second argument tells Julia to convert the data to a `DataFrame`. We store it as a variable called `df`.

③ We use the `first` function to show the first 5 rows of the DataFrame.

	Date Time	Water Level	Sigma	I	L
	String31	Float64	Float64	Int64	Int64
1	1928-01-01 00:00	-0.547	0.0	0	0
2	1928-01-01 01:00	-0.699	0.0	0	0
3	1928-01-01 02:00	-0.73	0.0	0	0
4	1928-01-01 03:00	-0.669	0.0	0	0
5	1928-01-01 04:00	-0.516	0.0	0	0

This data comes from the NOAA Tides and Currents website, specifically for a station at Sewells Point, VA for the year 1928. NAVD refers to the North American Vertical Datum, which is a reference point for measuring sea level, and GMT refers to Greenwich Mean Time, which is the time zone used in the data (rather than local time).

We can see that our DataFrame has five columns, the first of which is "Date Time". However, the "Date Time" column is being parsed as a `string`. We want it to be a `DateTime` object from the `Dates` package. To do that, we need to tell Julia how the dates are formatted. We could then manually convert, but `CSV.read` has a kewyord argument that we can use

```
date_format = "yyyy-mm-dd HH:MM"                                    ①
df = CSV.read(fname, DataFrame; dateformat=date_format)             ②
first(df, 3)
```

① This is a string that tells Julia how the dates are formatted. For example, `1928-01-01 00:00`. See the documentation for more information.

② `dateformat` is a *keyword argument* while `date_format` is a variable whose value is `"yyyy-mm-dd HH:MM"`. We could equivalently write `dateformat="yyyy-mm-dd HH:MM"`.

	Date Time	Water Level	Sigma	I	L
	DateTime	Float64	Float64	Int64	Int64
1	1928-01-01T00:00:00	-0.547	0.0	0	0
2	1928-01-01T01:00:00	-0.699	0.0	0	0
3	1928-01-01T02:00:00	-0.73	0.0	0	0

The next column is "Water Level", which is the height of the water above the reference point (NAVD) in meters. We can see that this is being parsed as a float, which is what we want . However, you have to *know* that the data is in meters rather than inches or feet or something else. To explicitly add information about the units, we can use the `Unitful` package.

```
df[!, " Water Level"] .*= 1u"m"                                     ①
first(df, 3)
```

① We select the column with water levels using its name. The `!` means "all rows". Thus, `df[!, " Water Level"]` is a vector of all the water levels stored. `*=` means to multiply in place. For example, if `x=2` then `x *= 2` is equivalent to `x = x * 2`. `.*=` is a vector syntax, meaning do the multiplication to each element of the vector individually. `1u"m"` is a `Unitful` object that

represents 1 meter. We multiply the water levels by this to convert them to meters.

	Date Time	Water Level	Sigma	I	L
	DateTime	Quantity…	Float64	Int64	Int64
1	1928-01-01T00:00:00	-0.547 m	0.0	0	0
2	1928-01-01T01:00:00	-0.699 m	0.0	0	0
3	1928-01-01T02:00:00	-0.73 m	0.0	0	0

0.3.3 Subsetting and renaming

We want to only keep the first two (for more on the other three, see here). We can also rename the columns to make them easier to work with (spaces in variable names are annoying). To do this, we use the `@rename` function:

```
df = @rename(df, :datetime = $"Date Time", :lsl = $" Water Level");          ①
```

① The `$` is needed here because the right hand side is a string, not a `Symbol`.

Then, we can use the `@select` function to do select the columns we want. Notice how the first argument to `select` is the `DataFrame` and the subsequent arguments are column names. Notice also that our column names were strings (`"Date Time"`), but we can also use symbols (`:datetime`).

```
df = @select(df, :datetime, :lsl)
first(df, 3)
```

	datetime	lsl
	DateTime	Quantity…
1	1928-01-01T00:00:00	-0.547 m
2	1928-01-01T01:00:00	-0.699 m
3	1928-01-01T02:00:00	-0.73 m

For more on what `DataFramesMeta` can do, see this Tweet.

0.3.4 Writing a function

We have just done a lot of work to read in our data. However, this just gives us data for the year 1928. In fact, we have a CSV file for each year 1928-2021. To make sure we can read them each in excatly the same way, we want to write a function. This function will take in the year as an argument, and return a `DataFrame` with the data for that year.

Writing functions is an important part of programming effectively. Let's write a function that takes in a year and returns a DataFrame with the data for that year, following the steps we've explored above.

Before we do that, let's define a function that will return the filename for a given year. It's often valuable to stack several functions together.

```
get_fname(year::Int) = "data/tidesandcurrents-8638610-$(year)-NAVD-GMT-metric.csv"
```

```julia
function read_tides(year::Int)

    # define the CSV file corresponding to our year of choice
    fname = get_fname(year)

    # a constant, don't change this
    date_format = "yyyy-mm-dd HH:MM"

    # <YOUR CODE GOES HERE>
    # 1. read in the CSV file and save as a dataframe
    # 2. convert the "Date Time" column to a DateTime object
    # 3. convert the " Water Level" column to meters
    # 4. rename the columns to "datetime" and "lsl"
    # 5. select the "datetime" and "lsl" columns
    # 6. return the dataframe
end

# print out the first 10 rows of the 1928 data
first(read_tides(1928), 10)
```

! Instructions

Fill out this function. Your function should implement the six steps indicated in the instructions. When it's done, convert it to a live code block by replacing "`julia" with "`{julia}". When you run this code, it should print out the first 10 rows of the 1928 data. Make sure they look right!

0.3.5 Combining files

Now that we have the ability to read in the data corresponding to any year, we can read them all in and combine into a single `DataFrame`. First, let's read in all the data.

! Instructions

Update the code blocks below, then replace "`julia" with "`{julia}".

```julia
years = 1928:2021 # all the years of data
annual_data = # 1. call the read_tides function on each year
typeof(annual_data) # should be a vector of DataFrames
```

Next, we'll use the `vcat` function to combine all the data into a single `DataFrame`.

```julia
df = vcat(annual_data...) # don't change this
first(df, 5)
```

```
last(df, 5) # check the last 5 years and ensure they are the last 3 hours of 2021-12-31
```

Finally, we'll make sure we drop any missing data.

```
dropmissing!(df) # drop any missing data
```

	datetime	lsl
	DateTime	Quantity…
1	1928-01-01T00:00:00	-0.547 m
2	1928-01-01T01:00:00	-0.699 m
3	1928-01-01T02:00:00	-0.73 m
4	1928-01-01T03:00:00	-0.669 m
5	1928-01-01T04:00:00	-0.516 m
6	1928-01-01T05:00:00	-0.364 m
7	1928-01-01T06:00:00	-0.212 m
8	1928-01-01T07:00:00	-0.059 m
9	1928-01-01T08:00:00	-0.029 m
10	1928-01-01T09:00:00	-0.029 m
11	1928-01-01T10:00:00	-0.151 m
12	1928-01-01T11:00:00	-0.303 m
13	1928-01-01T12:00:00	-0.486 m
14	1928-01-01T13:00:00	-0.608 m
15	1928-01-01T14:00:00	-0.699 m
16	1928-01-01T15:00:00	-0.73 m
17	1928-01-01T16:00:00	-0.699 m
18	1928-01-01T17:00:00	-0.638 m
19	1928-01-01T18:00:00	-0.486 m
20	1928-01-01T19:00:00	-0.364 m
21	1928-01-01T20:00:00	-0.303 m
22	1928-01-01T21:00:00	-0.273 m
23	1928-01-01T22:00:00	-0.364 m
24	1928-01-01T23:00:00	-0.455 m
25	1928-01-02T00:00:00	-0.638 m
26	1928-01-02T01:00:00	-0.821 m
27	1928-01-02T02:00:00	-0.943 m
28	1928-01-02T03:00:00	-1.035 m
29	1928-01-02T04:00:00	-1.004 m
30	1928-01-02T05:00:00	-0.913 m
…	…	…

0.3.6 Time series plot

Now we're ready to make some plots of our data. Let's start with a simple time series plot of the water levels. Our data is collected hourly, so we have a lot of data points! Still, we can plot them all.