

Data Operations

Table of contents

1	Aims	1
2	Tools	2
3	Usage	2
3.1	Basics	2

1 Aims

The [R Learning Syllabus](#) is pretty much our first opportunity within [Exercism](#) to explore data science operations with dataframes. The traditional `data.frame` and newer `tibble` are include in the test runner, along with Tidyverse libraries such as `dplyr`, `readr`, `tidyr` and `forcats`.

In contrast, the [Python track](#) is beginner-biased (they get a *lot* of newbies) and limited to Core Python.

Hence this separate code playground, where it is possible to install extra packages and work across multiple languages (R, Python, SQL for starters).

The current ~~plan~~ vague aspiration cover all these combinations:

	R	Python	Julia
<code>data.frame</code>			
<code>tibble</code>			
<code>data.table</code>			
<code>pandas</code>			
<code>DataFrames</code>			
<code>Polars</code>			
<code>DuckDB</code>			
<code>PostgreSQL</code>			
<code>Arrow</code>			
<code>Dask</code>			

2 Tools

data.frame is part of Base R, and always available to R code.

tibble is a Tidyverse package. Because it is common for modern R code to start with `library(tidyverse)` (even within Exercism), it will often be available, as will packages such as `dplyr` to manipulate tibbles.

None of the remaining tools are available in the Exercism R or Python tracks. For local use, they must be installed separately.

data.table is an R package which (like tibbles) aims to improve on the traditional **data.frame**. Reputedly faster than tibbles with large datasets, though less integrated with the Tidyverse ecosystem.

pandas is the oldest type of Python dataframe, very widely used.

Polars is a newer Rust-based dataframe type, with binding for several languages including R, Python and Julia. Reputed to be very fast with larger datasets.

DuckDB is an in-process database (**OLAP**) server, based mainly (not exclusively) on SQL queries.

PostgreSQL is a powerful, open-source relational (**OLTP**) database server.

Apache Arrow is a language-independent columnar memory format for flat and nested data. Libraries to work with rrow data exist for many languages, including R, Python and Julia.

Dask is a pure Python package, pandas-like but optimized for parallel processing of large datasets.

3 Usage

For simplicity, examples are split into multiple language-specific documents. Each section should have a similar overall structure across languages.

3.1 Basics

- [R examples](#)
- [Python examples](#)
- [Julia examples](#)