

冒険の続き ~ RedAmber ~

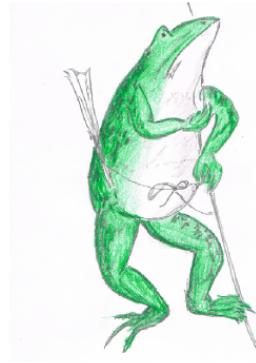
A DataFrame Library in Ruby

@heronshoes

2023-08-19/RubyKaigi 2023 follow up - 例のあれ、どうなりました？

self.introduction

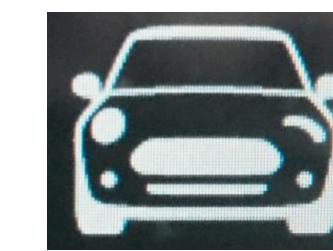
- 鈴木 弘一(Hirokazu SUZUKI)
- GitHub/ex-Twitter @heronshoes



Red Data Tools

- 広島県福山市から来ました
- Ruby愛好家を名乗っております
- コーヒーとクラフトビールとMINIが好き

今は短歌が
マイブームです



Gotham City in Batman movie
and Fukuyama city tied friendship city cooperation !





元ネタは？

May 11

May 12

May 13

| | Large Hall #rubykaigiA | Small Hall #rubykaigiB | Open Studio #rubykaigiC |
|---------------------|--|---|---|
| 09:00 - 09:40 | | Door Open | |
| 09:40 - 10:50 |  Ruby Committers and The World CRuby Committers @rubylangorg EN | | |
| 11:00 - 11:30 |  Build Your Own SQLite3 Hitoshi HASUMI @hasumikin JA |  Gradual typing for Ruby: comparing RBS and RBI/Sorbet Alexandre Terrasa @Morriar EN |  The Adventure of RedAmber - A data frame library in Ruby Hirokazu SUZUKI @heronshoes JA |
| 11:30 - 13:30 | | Lunch Break | |

これです！

RedAmber とは？

- RedAmber とはRubyで書かれたデータフレームライブラリ
 - データフレームは、列にラベルがある二次元のデータ構造
 - Pythonの**pandas**, Rのdplyr/tidyr, Rustの**Polars**
 - SQLのテーブルと共通する概念
- RedAmber は **Apache Arrow**を使っている
 - 列指向のメモリーフォーマットとそれを扱うライブラリ群
- RedAmber は2022年度 Rubyアソシエーション開発助成金対象プロジェクト

```
diamonds
  .slice { carat > 1 }
  .pick(:cut, :price)
  .group(:cut)
  .mean
  .sort('~-mean(price)')
```

| RedAmber::DataFrame <5 x 2 vectors> | |
|-------------------------------------|-----------------|
| | cut mean(price) |
| Ideal | 8674.23 |
| Premium | 8487.25 |
| Very Good | 8340.55 |
| Good | 7753.6 |
| Fair | 7177.86 |

8/28 成果報告会



私は次に何をすると言っていたか？

- ・ もっと多くの使用例を示す
- ・ 「データサイエンス100本ノック」を RedAmber で作る・・・②
- ・ より速い実装を目指す
- ・ RedAmber 自身・・・①
- ・ Query (Workload) を他のエンジンで処理できるようにする構想
- ・ (そのために) Substrait に貢献する

もっと多くのRubyistに
使ってみて欲しい

① 高速化

☑ DataFrame#join を速くした

- Apache Arrow の Aceroエンジンを使うようにした

Acero は
C++ で書かれたクエリの実行エンジン
Rubyから使える

☑ 部分文字列に対する操作を Vector#match_substring ファミリーを導入して速くした

- Arrow C Glib が MatchSubstringOptions に対応した成果を利用

☑ Vector#rank を速くした

- Arrow C GLib が RankOptionsに対応した成果を利用

Vector は
列を表すオブジェクト

- `100本ノック`もPure RubyでなくArrowを使って解けるので速くシンプルになる

RedAmber 0.5.1

- Ruby界で有名な締切効果により、昨日無事リリース
- RubyKaigi以降の成果を盛り込みました

Release note for v0.5.1 #279

heronshoes announced in Announcements

heronshoes 12 hours ago Maintainer ...

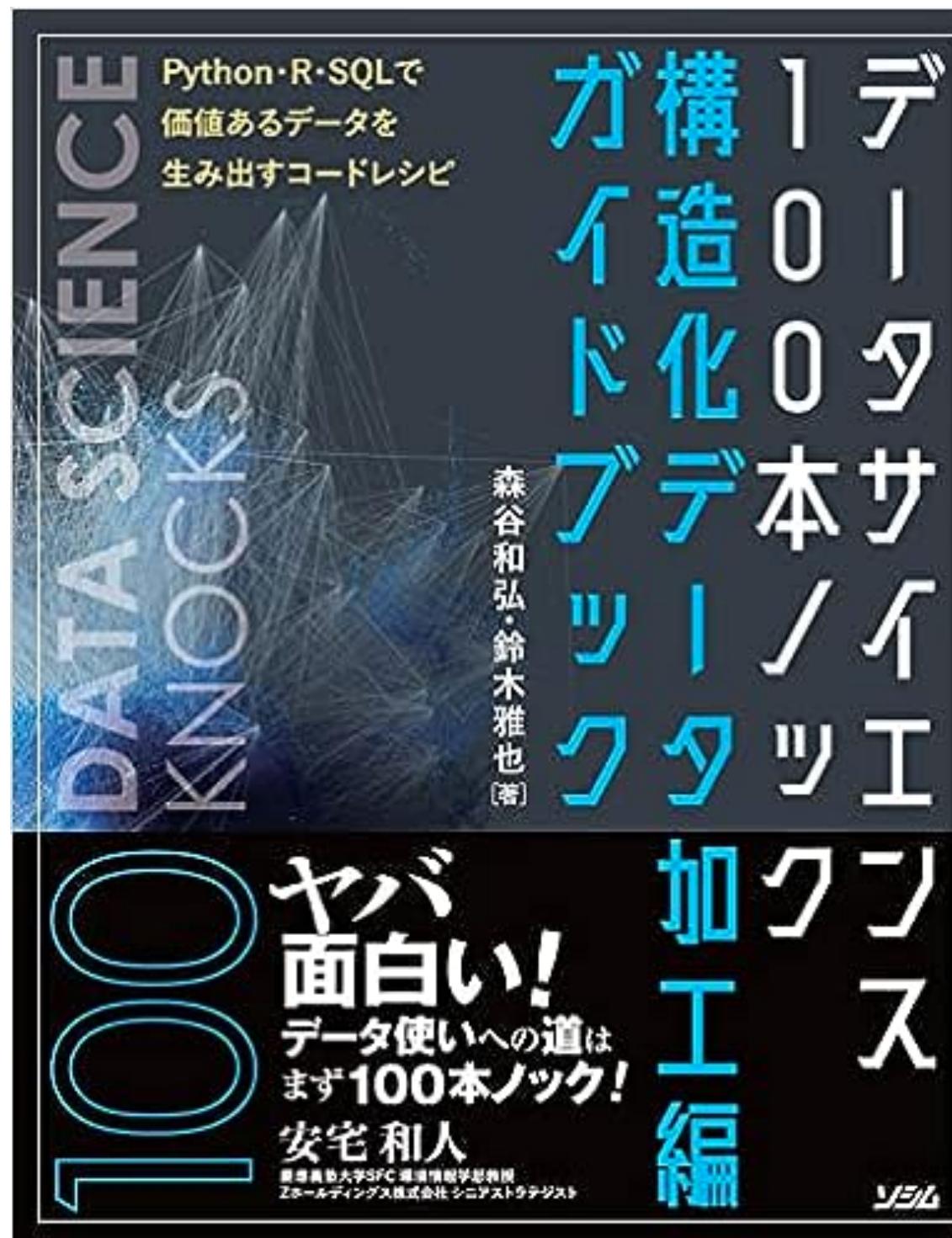
RedAmber v0.5.1 has been released 🎉

RedAmber now supports Dev Container. We can prepare full stack of development environment without changing local environment. Jupyter Notebooks for examples are created from qmd files. This enables simple source management of Notebooks.

Topics for this release is below;

データサイエンス100本ノックとは？

<https://github.com/The-Japan-DataScientist-Society/100knocks-preprocess>



- データ分析の前段階で必要なデータ処理の実例を学ぶ
- Python, R, SQLの実行環境
- 各言語で同じ100問の設問と解答例
- GitHubのサイトと書籍の両方で学習できる

② 100本ノックを作るにあたって

- RedAmber を **Dev Container** 対応のリポジトリにして、簡単に開発環境が作成できるようにした
 - Apache Arrowをシステムにインストールしてもらうのは大変
 - ローカルの環境に影響を与えずにコンテナ内に環境を作れる
 - GitHub Codespaces を使えば、ブラウザ上のVS Codeでも動く
- RedAmberを使ってもらうハードルを低くできる！

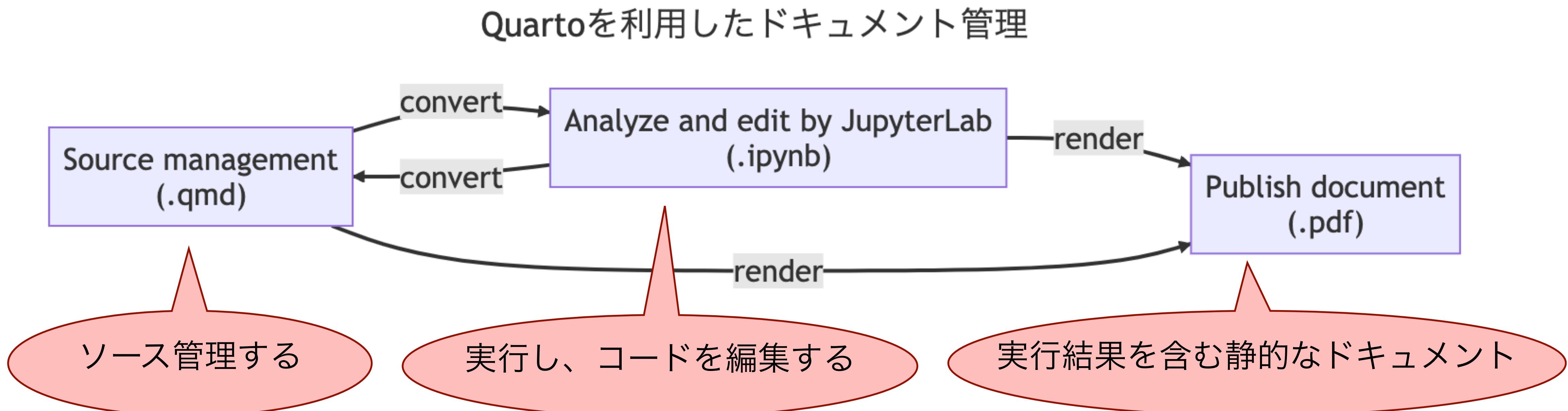
② 100本ノックを作るにあたって

- ・ オリジナルの「100本ノック」とは別のやり方で環境を作った
- ・ オリジナルはPythonのData Science Notebook環境
- ・ このDockerfileに必要なものを足していくやり方でいいの？
- ・ Dev Container Feature を使う
- ・ Python(+ Jupyter Lab) + Ruby + Quarto + Apache Arrow
- ・ 複数のコンポーネントが必要な開発環境を統一的に作るモデルケース

R由来の出版システム

② 100本ノックを作るにあたって

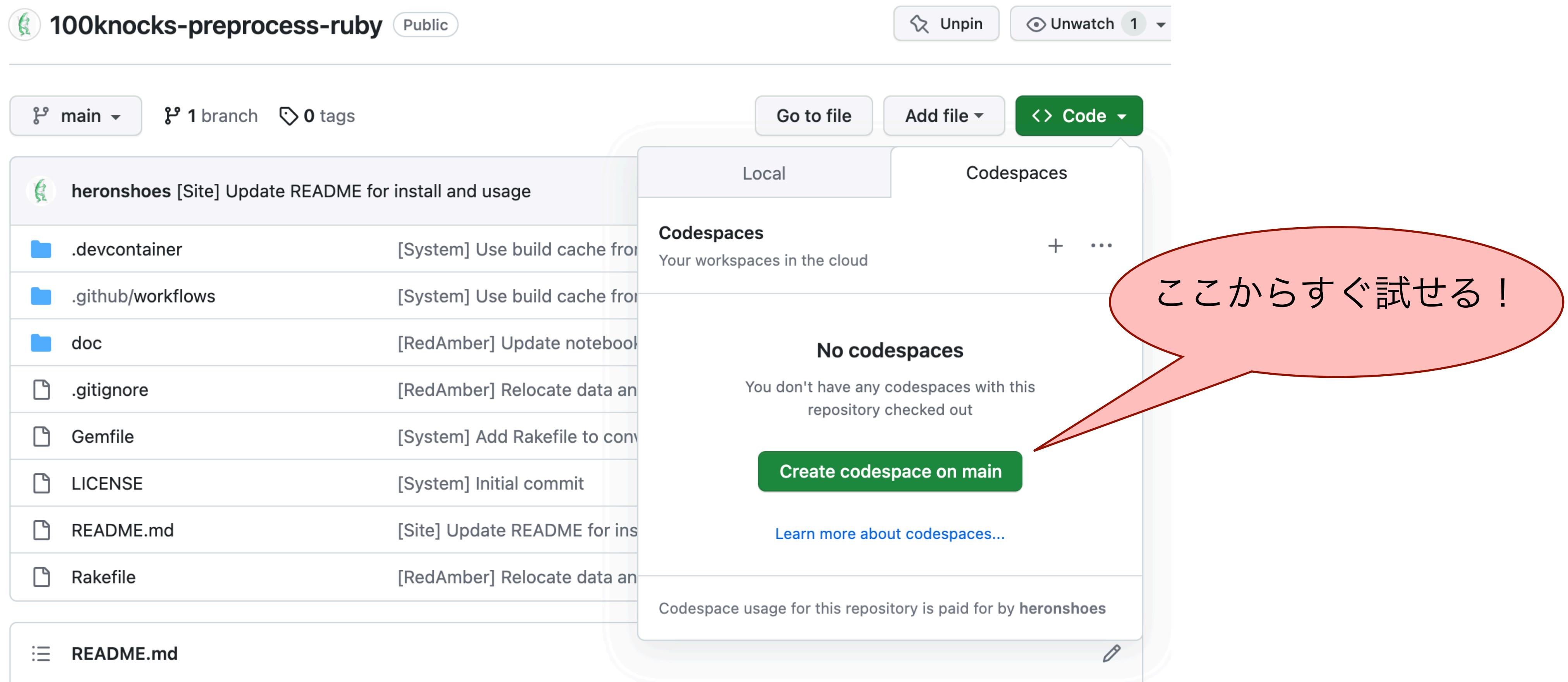
- Jupyter Notebook (.ipynb) を Quarto で .qmd ファイルから生成するようにして、シンプルなソース管理を可能にした



データサイエンス100本ノック（構造化データ加工編）

Ruby(RedAmber)版

- <https://github.com/heronshoes/100knocks-preprocess-ruby>



RedAmber自身もDev Container化

- すぐ試せます！開発もすぐできます！

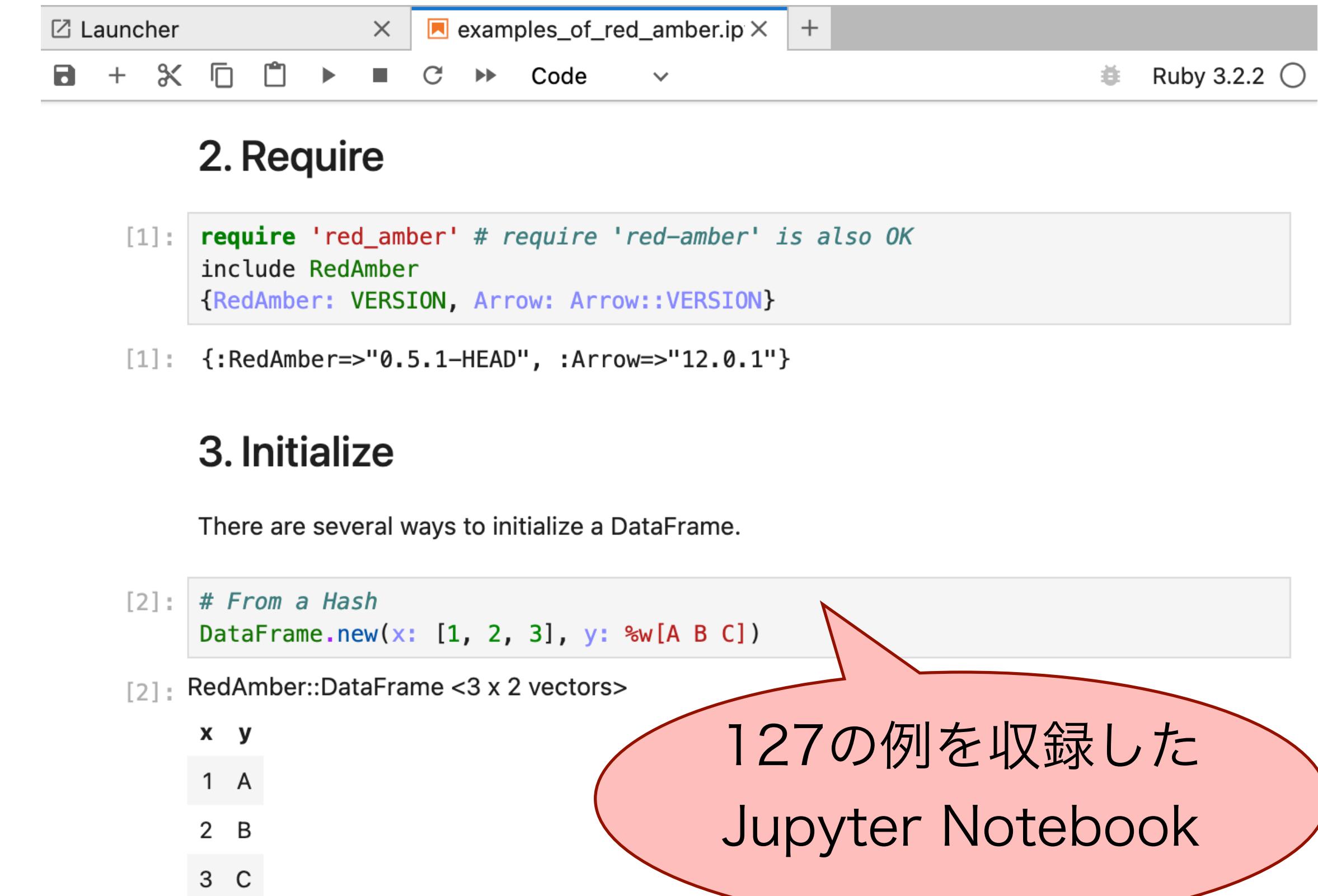
```
bundle exec --gemfile=bin/Gemfile bin/example
reading general dataframe and subframes...
From: bin/example @ line 82 :

77: # Welcome to RedAmber example!
78: # This environment will offer these pre-loaded datasets:
79: #   penguins, diamonds, iris, starwars, simpsons_paradox_covid,
80: #   mtcars, band_members, band_instruments, band_instruments2
81: #   import_cars, comecome, rubykaigi, dataframe, subframes
=> 82: binding.irb

irb(main):001:0> dataframe
=>
#<RedAmber::DataFrame : 6 x 3 Vectors, 0x000000000000c760>
      x   y     z
  <uint8> <string> <boolean>
0      1 A    false
1      2 A    true
2      3 B    false
3      4 B    (nil)
4      5 B    true
5      6 C    false

irb(main):002:0>
irb(main):003:0>
```

irbを使った
REPL環境。
サンプルデータセット付き



The screenshot shows a Jupyter Notebook interface with the title 'examples_of_red_amber.ipynb'. The notebook contains the following content:

2. Require

```
[1]: require 'red_amber' # require 'red-amber' is also OK
include RedAmber
{RedAmber: VERSION, Arrow: Arrow::VERSION}
```

```
[1]: {:RedAmber=>"0.5.1-HEAD", :Arrow=>"12.0.1"}
```

3. Initialize

There are several ways to initialize a DataFrame.

```
[2]: # From a Hash
DataFrame.new(x: [1, 2, 3], y: %w[A B C])
```

```
[2]: RedAmber::DataFrame <3 x 2 vectors>
      x   y
      1   A
      2   B
      3   C
```

127の例を収録した
Jupyter Notebook

関連するリソース

- **RedAmber GitHub Home**
 - https://github.com/red-data-tools/red_amber
- **データサイエンス100本ノック（構造化データ加工編）Ruby(RedAmber)版**
 - <https://github.com/heronshoes/100knocks-preprocess-ruby>
- **RubyKaigi 2023の発表スライドと今日のスライド in rabbit環境**
 - https://github.com/heronshoes/rubykaigi2023-red_amber

使ってみてください！

Red Data Toolsでお会いしましょう!



写真：福山市鞆 仙酔島 五色岩にて

Powered by Rabbit 3.0.1