Productionize Your Open Source Project

Øredev 2023

Rami Krispin, Nov 8, 2023

About Me

- Data Science and Engineering Manager
- Forecasting
- MLOps
- Open Source
- Author
- ❤️ Data

About Me

- Data Science and Engineering Manager
- Forecasting
- MLOps
- Open Source
- Author
- ❤️ Data

About Me

- Data Science and Engineering Manager
- **Forecasting**
- MLOps
- Open Source
- Author
- ❤️ Data

Forecasting at Scale

Time Series Cluster Analysis

View By:

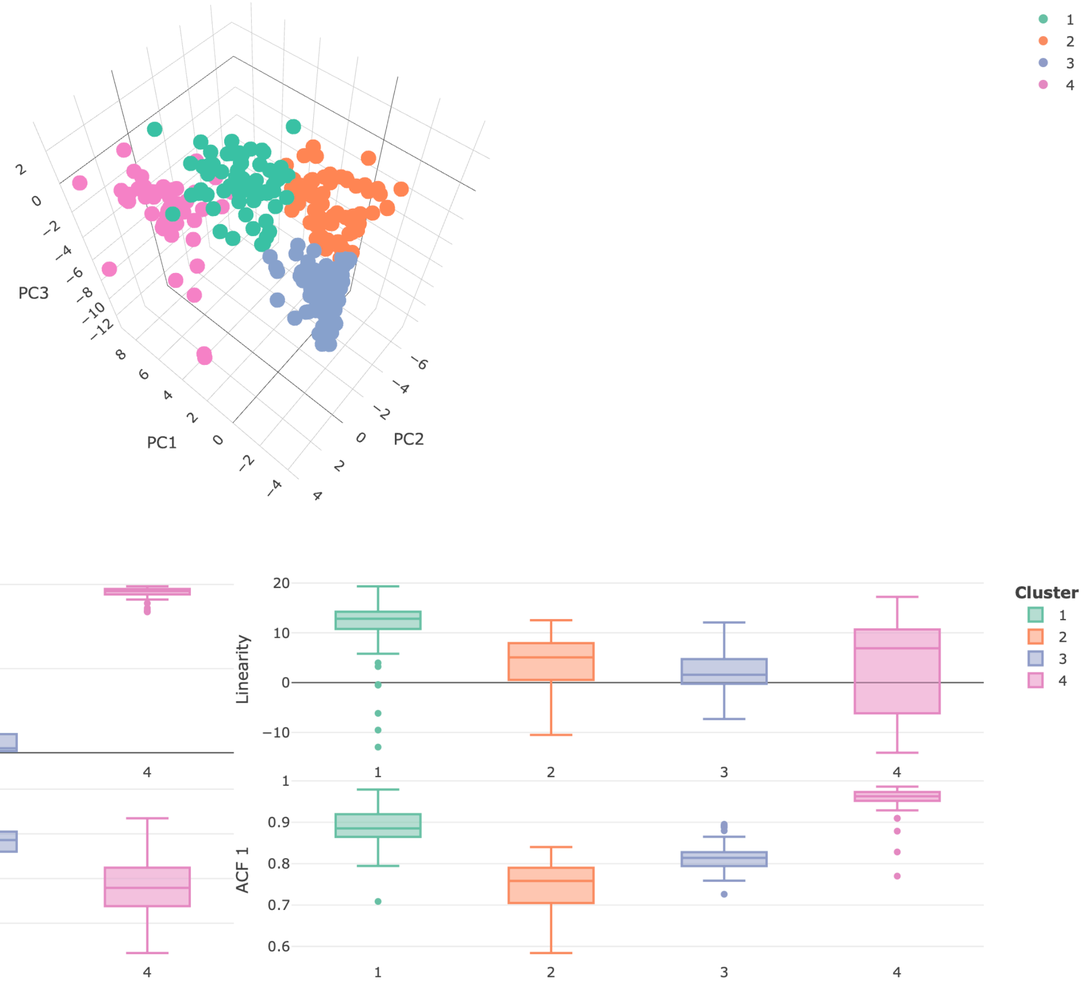
Cluster

Number of Clusters:

4

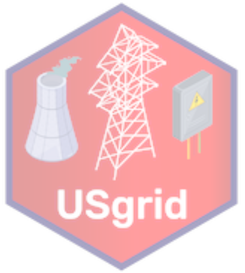
Number of PCs

Three



About Me

- Data Science and Engineering Manager
- Forecasting
- **MLOps**
- **Open Source**
- Author
- ❤️ Data



Productionize an Open Source Project

Why

How

Data Package

- Common method in R
- Store data
- Education - academia, research, papers
- Accessable
- Examples - iris, AirPassengers, CO2, mtcars

Data Package

The R Datasets Package

Documentation for package ‘datasets’ version 4.4.0

- DESCRIPTION file.

Help Pages

A B C D E F H I J L M N O P Q R S T U V W

datasets-package	The R Datasets Package
-- A --	
ability.cov	Ability and Intelligence Tests
airmiles	Passenger Miles on Commercial US Airlines, 1937-1960
AirPassengers	Monthly Airline Passenger Numbers 1949-1960
airquality	New York Air Quality Measurements
anscombe	Anscombe's Quartet of 'Identical' Simple Linear Regressions
attenu	The Joyner-Boore Attenuation Data
attitude	The Chatterjee-Price Attitude Data
austres	Quarterly Time Series of the Number of Australian Residents

Data Package

```
> data("coronavirus", package = "coronavirus")
> head(coronavirus)
        date province country     lat      long      type cases   uid iso2 iso3 code3    combined_key
1 2020-01-22  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN   124 Alberta, Canada
2 2020-01-23  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN   124 Alberta, Canada
3 2020-01-24  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN   124 Alberta, Canada
4 2020-01-25  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN   124 Alberta, Canada
5 2020-01-26  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN   124 Alberta, Canada
6 2020-01-27  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN   124 Alberta, Canada
  population continent_name continent_code
1    4413146  North America             NA
2    4413146  North America             NA
3    4413146  North America             NA
4    4413146  North America             NA
5    4413146  North America             NA
6    4413146  North America             NA
```

Motivation

February 2020

Chinese stocks plunged 8% as coronavirus fears took hold. It’s the worst day in years

By Laura He, CNN Business

Updated 9:15 AM EST, Mon February 3, 2020

Source: CNN

10:32 p.m. ET, February 2, 2020

The best photos from the game

Kyle Terada/USA Today Sports

Source: CNN

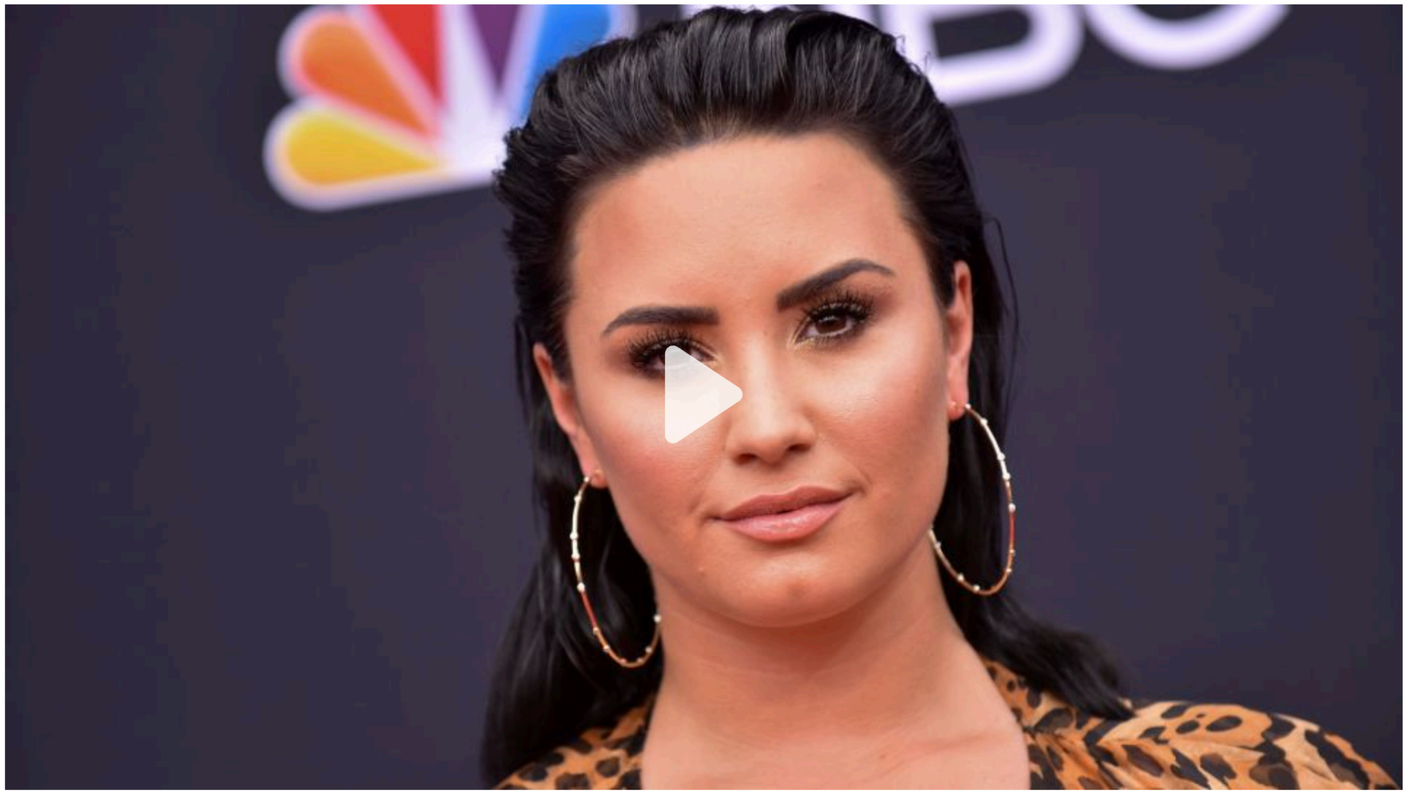
Jennifer Lopez's epic Super Bowl flag coat was custom Versace

By Sandra Gonzalez, CNN
Updated 6:58 AM EST, Mon February 3, 2020

Source: CNN

Demi Lovato rocks the National Anthem at Super Bowl LIV

By Lisa Respers France, CNN
Updated 11:27 PM EST, Sun February 2, 2020



Source: CNN

Trump and Pelosi haven’t spoken in months

By Manu Raju, Senior Congressional Correspondent
Updated 8:49 AM EST, Mon February 3, 2020

Source: CNN

‘No reason for Americans to panic’: White House seeks to calm fears over coronavirus

By Chandelis Duster, CNN

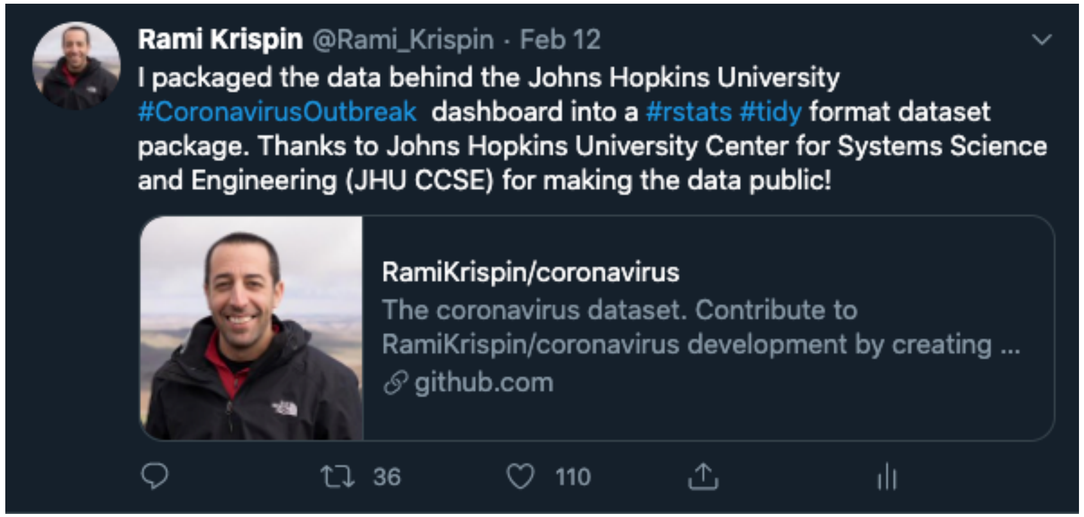
Updated 6:21 PM EST, Sun February 2, 2020

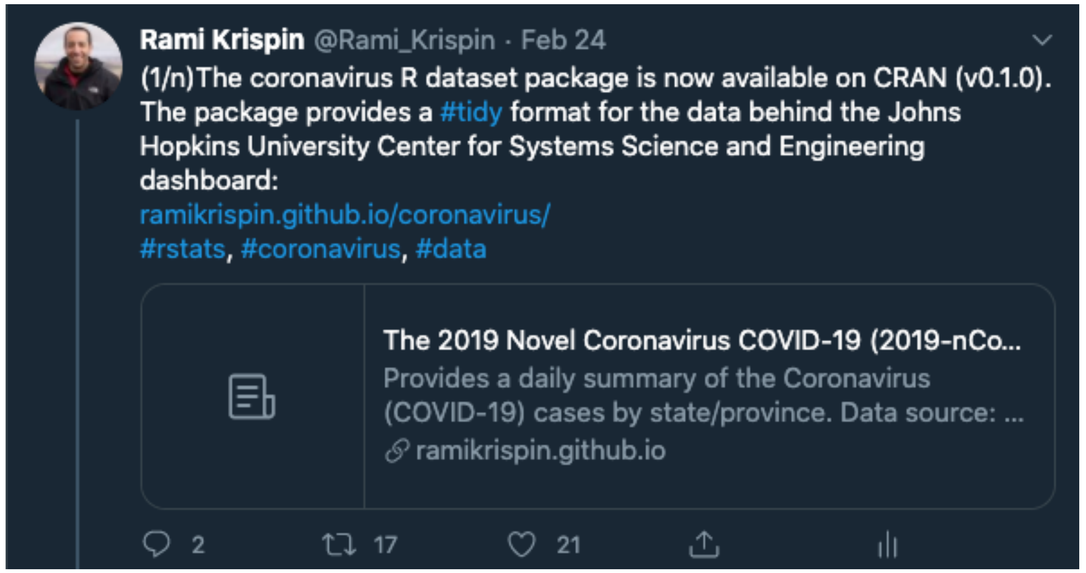

EVERYBODY PANIC!


TYRE CENTRE

Motivation


Rami Krispin @Rami_Krispin · Feb 8
Does anyone aware of a public dataset of the #coronavirus by case over time (e.g., time, city, country, lon, lat, etc...)?
2
1
1


Rami Krispin @Rami_Krispin · Feb 12
I packaged the data behind the Johns Hopkins University #CoronavirusOutbreak dashboard into a #rstats #tidy format dataset package. Thanks to Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) for making the data public!
RamiKrispin/coronavirus
The coronavirus dataset. Contribute to RamiKrispin/coronavirus development by creating ...
github.com
36
110


Rami Krispin @Rami_Krispin · Feb 24
(1/n)The coronavirus R dataset package is now available on CRAN (v0.1.0). The package provides a #tidy format for the data behind the Johns Hopkins University Center for Systems Science and Engineering dashboard:
ramikrispin.github.io/coronavirus/
#rstats, #coronavirus, #data
The 2019 Novel Coronavirus COVID-19 (2019-nCo...
Provides a daily summary of the Coronavirus (COVID-19) cases by state/province. Data source: ...
ramikrispin.github.io
2
17
21

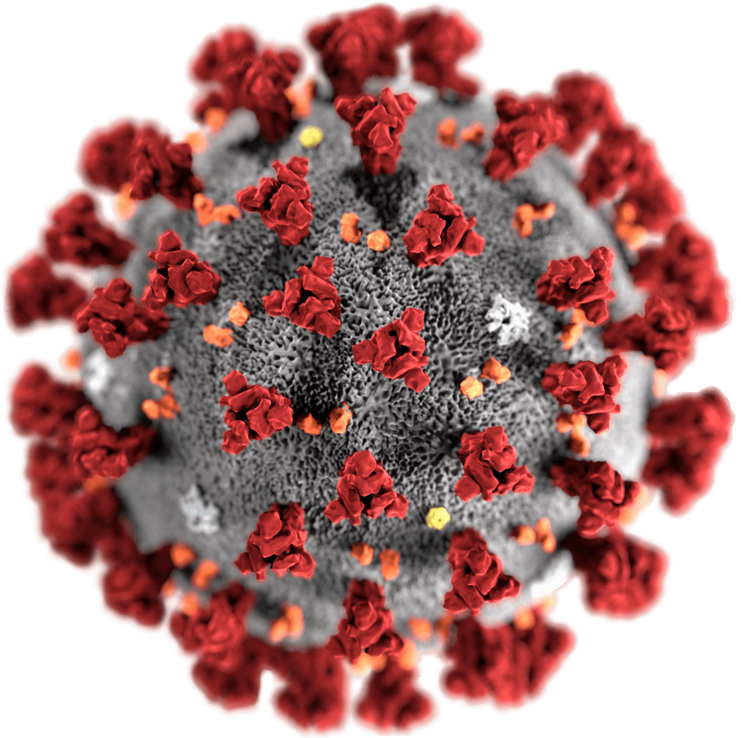
Motivation

coronavirus

The coronavirus package provides a tidy format for the COVID-19 dataset collected by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The dataset includes daily new and death cases between January 2020 and March 2023 and recovery cases until August 2022.

More details available here, and a `csv` format of the package dataset available here

Data source: https://github.com/CSSEGISandData/COVID-19



Source: Centers for Disease Control and Prevention's Public Health Image Library

Links

View on CRAN

Browse source code

Report a bug

License

Full license

MIT + file LICENSE

Citation

Citing coronavirus

Developers

Rami Krispin
Author, maintainer

Jarrett Byrnes
Author

Dev status

R-CMD

Data Pipeline passing

CRAN 0.4.1

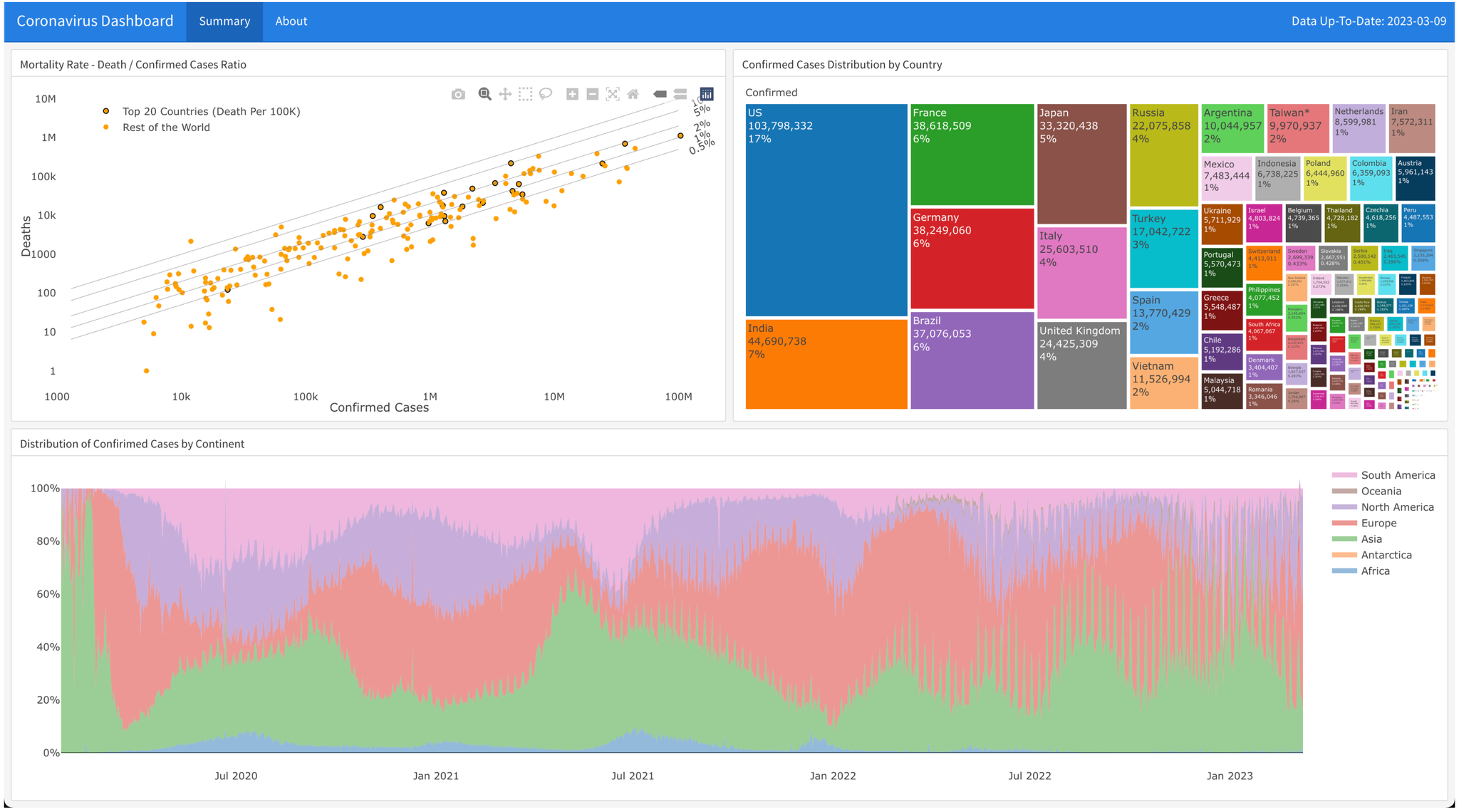
lifecycle stable

License MIT

last commit march

downloads 74K

Motivation

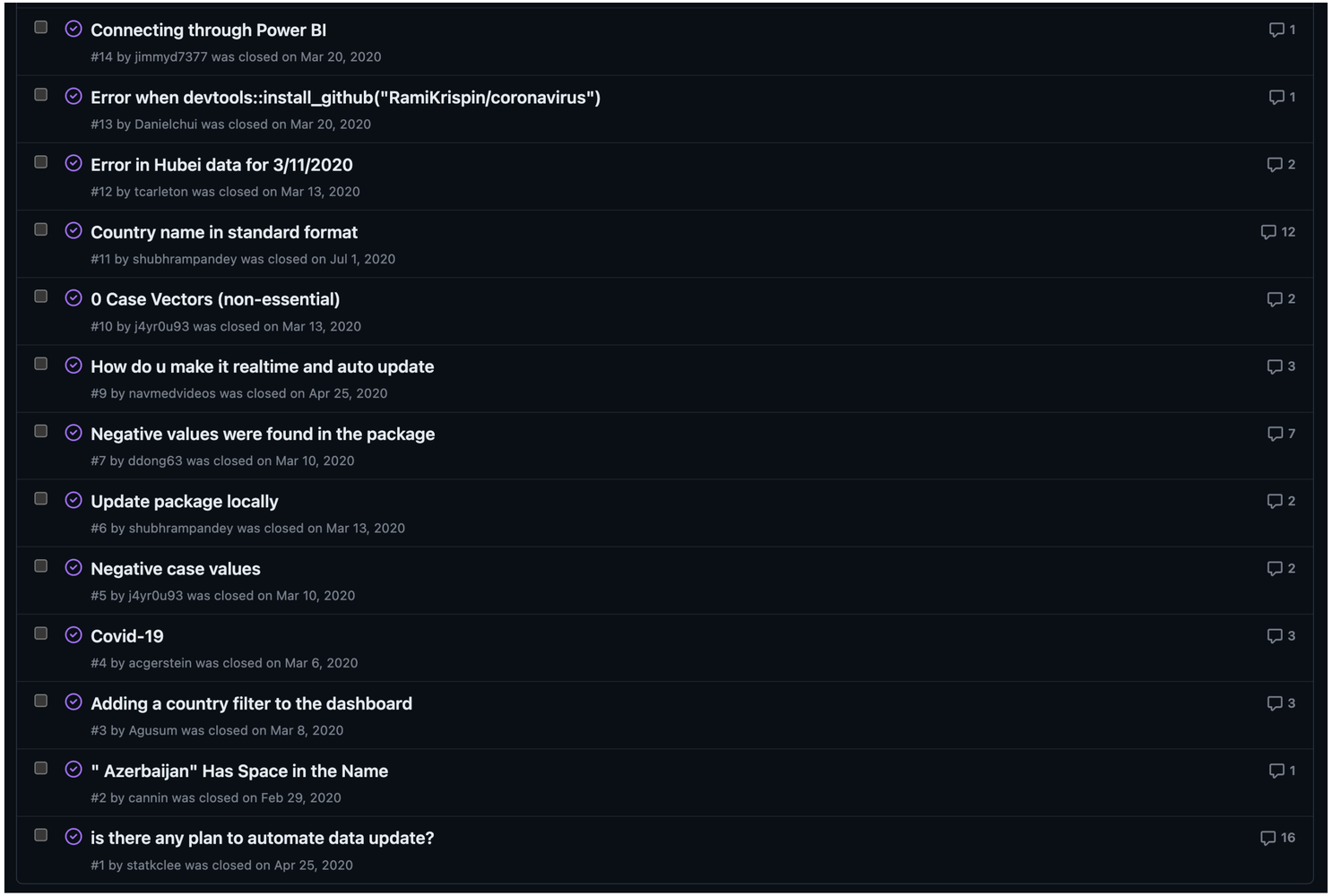

Coronavirus Dashboard
Summary
About
Data Up-To-Date: 2023-03-09
Mortality Rate - Death / Confirmed Cases Ratio
Top 20 Countries (Death Per 100K)
Rest of the World
Deaths
Confirmed Cases
10M
1M
100k
10k
1000
100
10
1
1000
10k
100k
1M
10M
100M
Confirmed Cases Distribution by Country
Confirmed
US 103,798,332 17%
India 44,690,738 7%
France 38,618,509 6%
Germany 38,249,060 6%
Brazil 37,076,053 6%
Japan 33,320,438 5%
Italy 25,603,510 4%
United Kingdom 24,425,309 4%
Russia 22,075,858 4%
Turkey 17,042,722 3%
Spain 13,770,429 2%
Vietnam 11,526,994 2%
Argentina 10,044,957 2%
Taiwan* 9,970,937 2%
Netherlands 8,599,981 1%
Iran 7,572,311 1%
Mexico 7,483,444 1%
Indonesia 6,738,225 1%
Poland 6,444,960 1%
Colombia 6,359,093 1%
Austria 5,961,143 1%
Ukraine 5,711,929 1%
Israel 4,803,824 1%
Belgium 4,739,365 1%
Thailand 4,728,182 1%
Czechia 4,618,256 1%
Peru 4,487,553 1%
Portugal 5,570,473 1%
Switzerland 4,413,911 1%
Sweden 2,699,339 0.433%
Slovakia 2,667,551 0.428%
Philippines 4,077,452 1%
Greece 5,548,487 1%
South Africa 4,067,067 1%
Chile 5,192,286 1%
Denmark 3,404,407 1%
Malaysia 5,044,718 1%
Romania 3,346,046 1%
Distribution of Confirimed Cases by Continent
100%
80%
60%
40%
20%
0%
Jul 2020
Jan 2021
Jul 2021
Jan 2022
Jul 2022
Jan 2023
South America
Oceania
North America
Europe
Asia
Antarctica
Africa

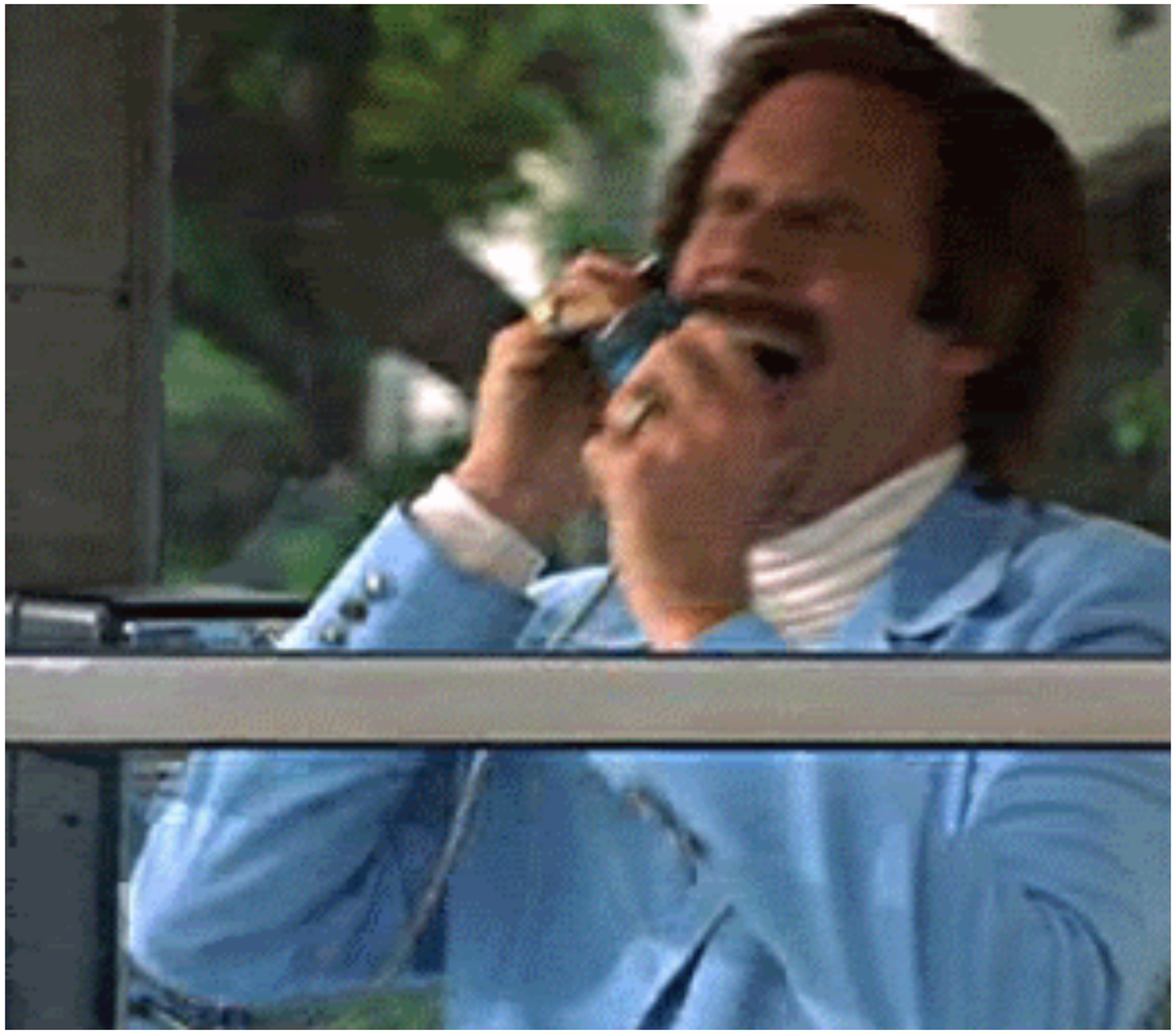
Why

Key Challenges

- Data refresh daily
- Data structure was not stable
- Dependency on external sources
- Lack of consistency

Key Challenges



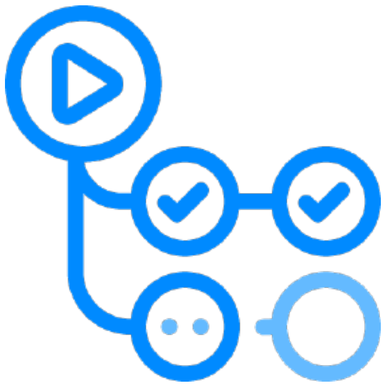


Different Approach

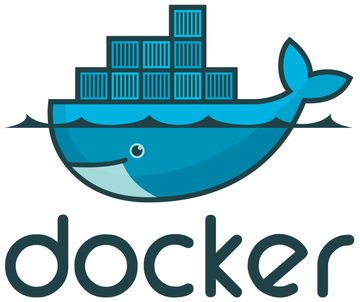
Different Approach

- Convert the data structure to an API
- Data automation/pipeline
- Add a robust unit testings
- Using open source and free tools

Solution

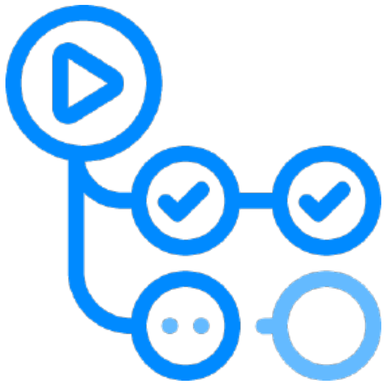


+

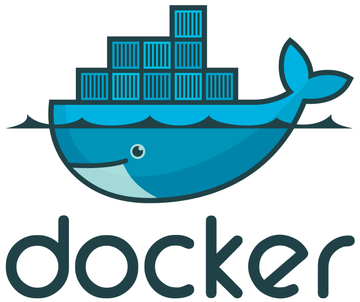


=

Solution

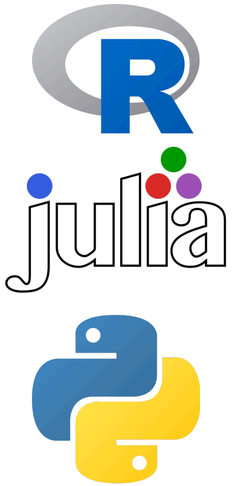


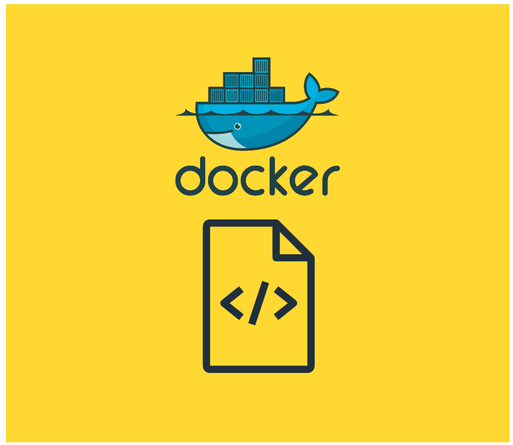
A Scheduler and CI/CD tool



High level of reproducibility

Docker in Nutshell

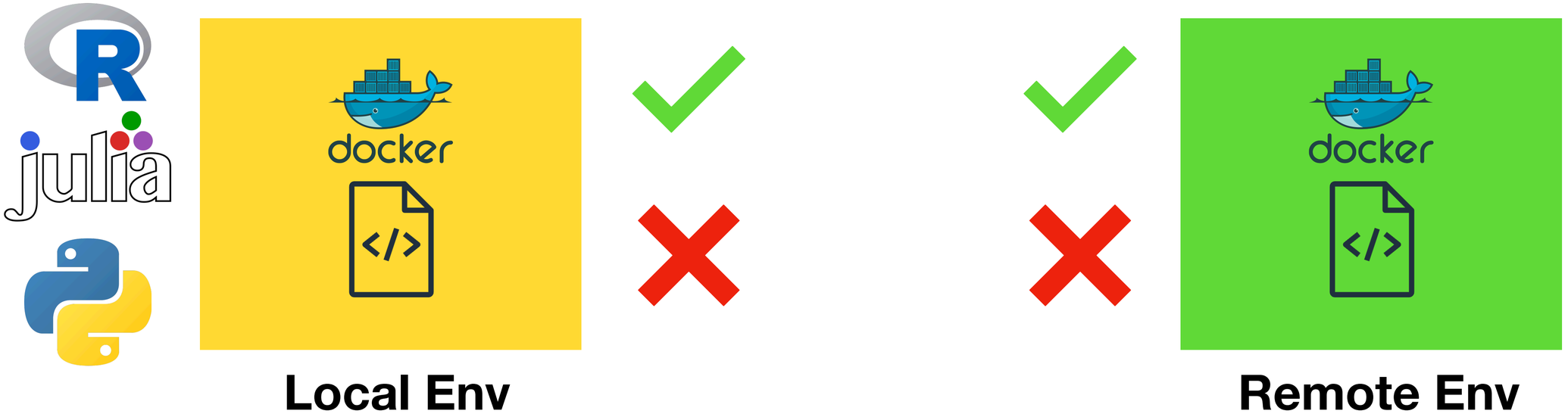




Local Env

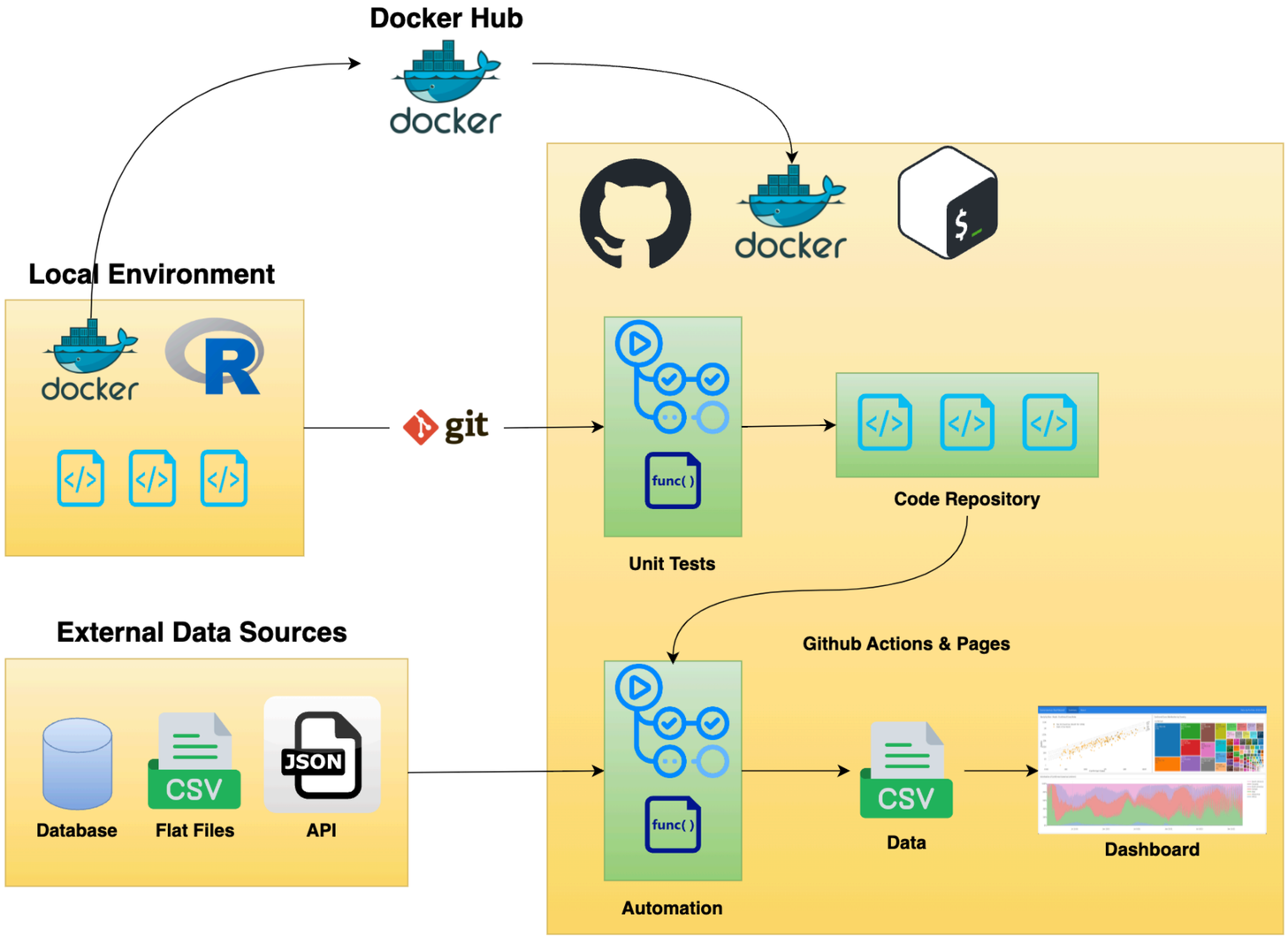
Remote Env

Docker in Nutshell

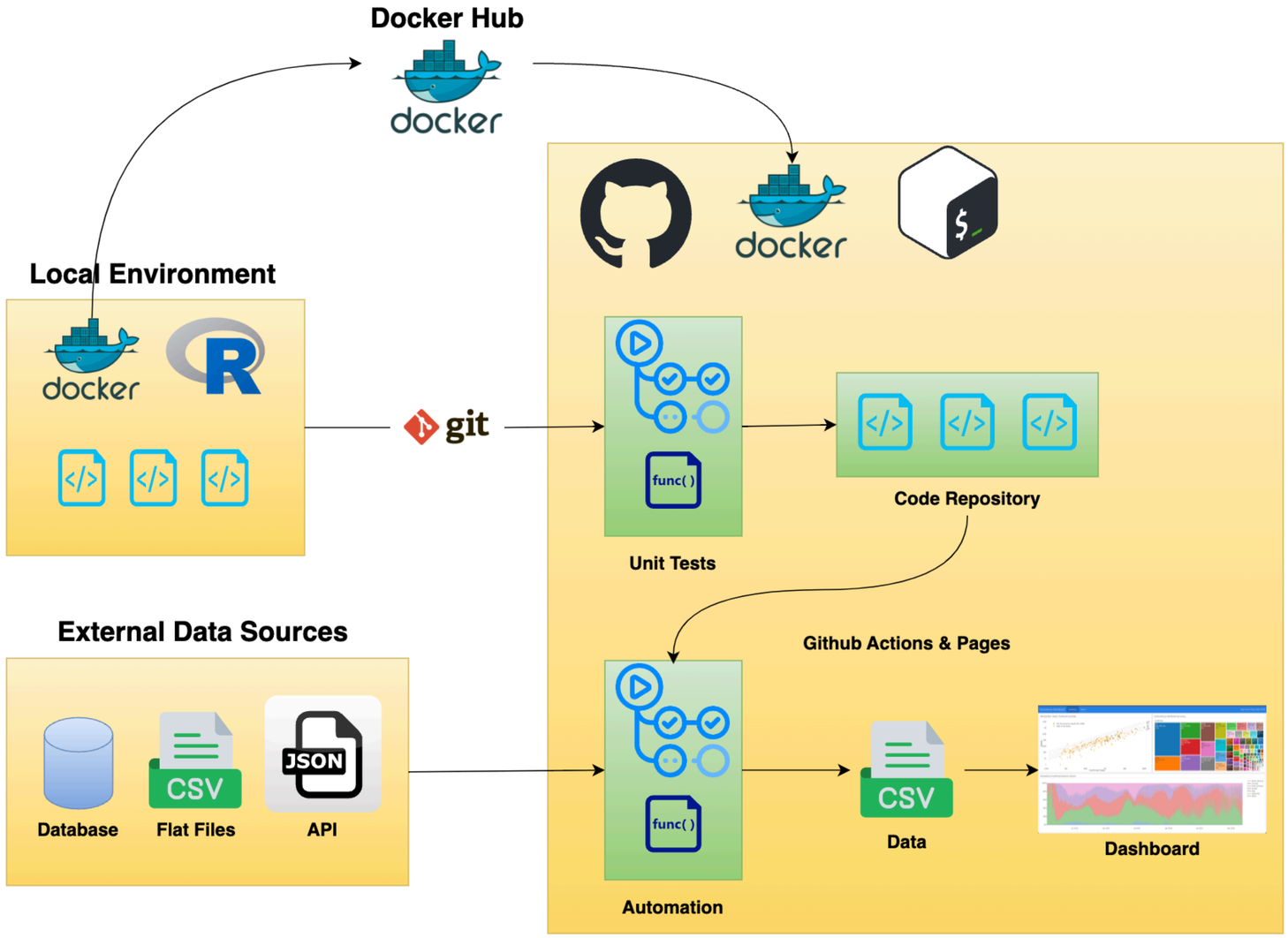


Reproducibility

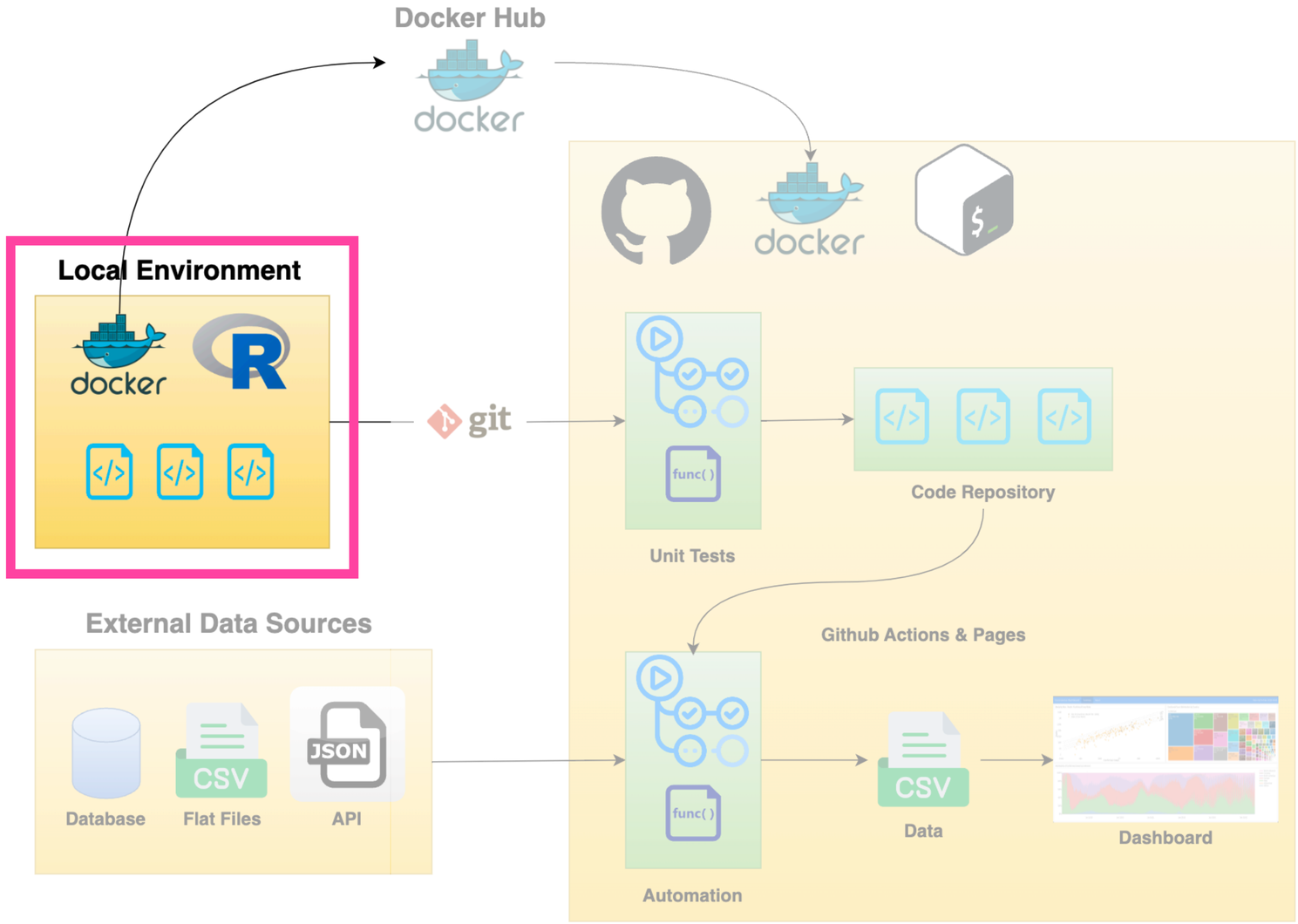
General Architecture



General Architecture

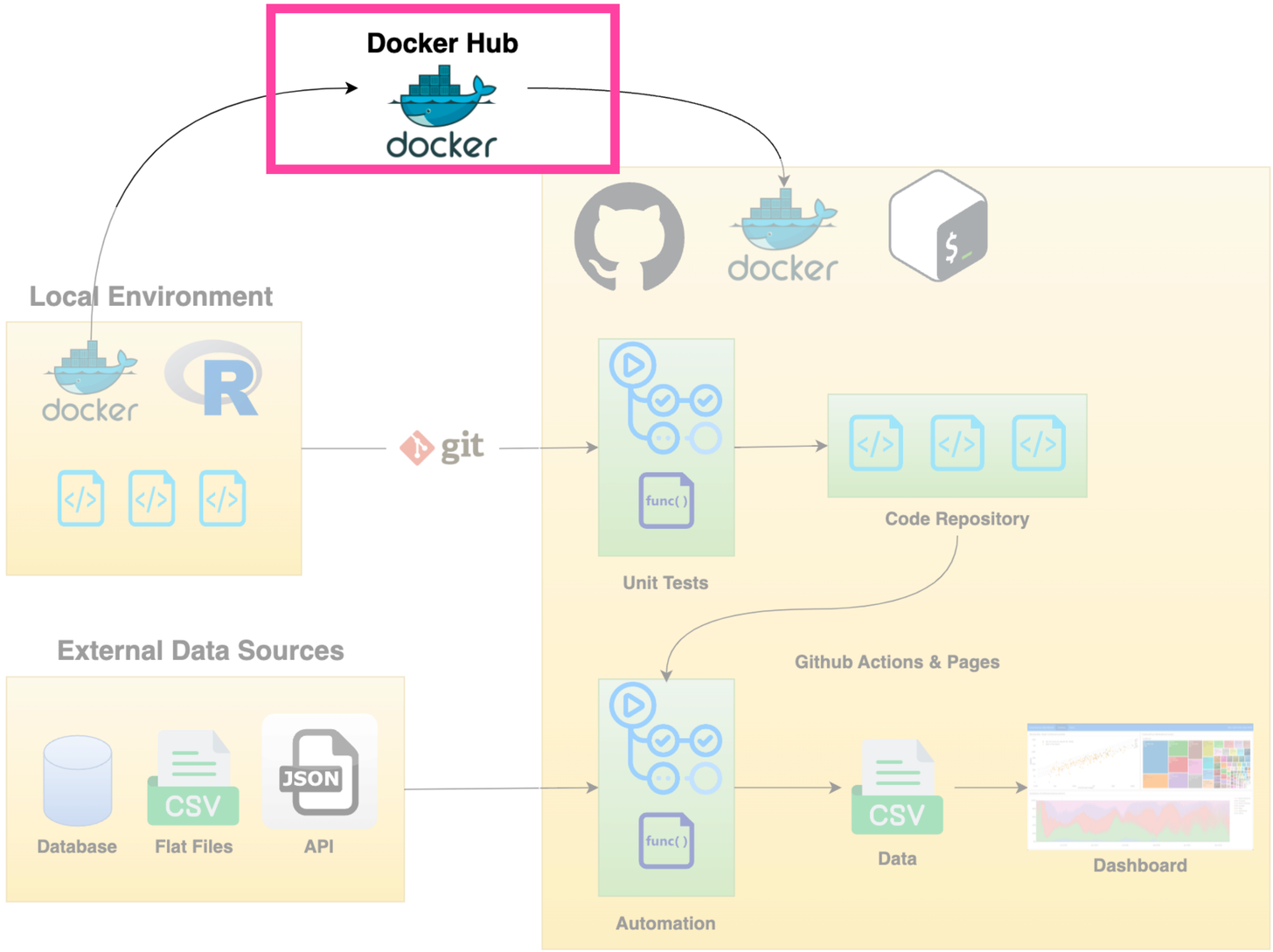


General Architecture



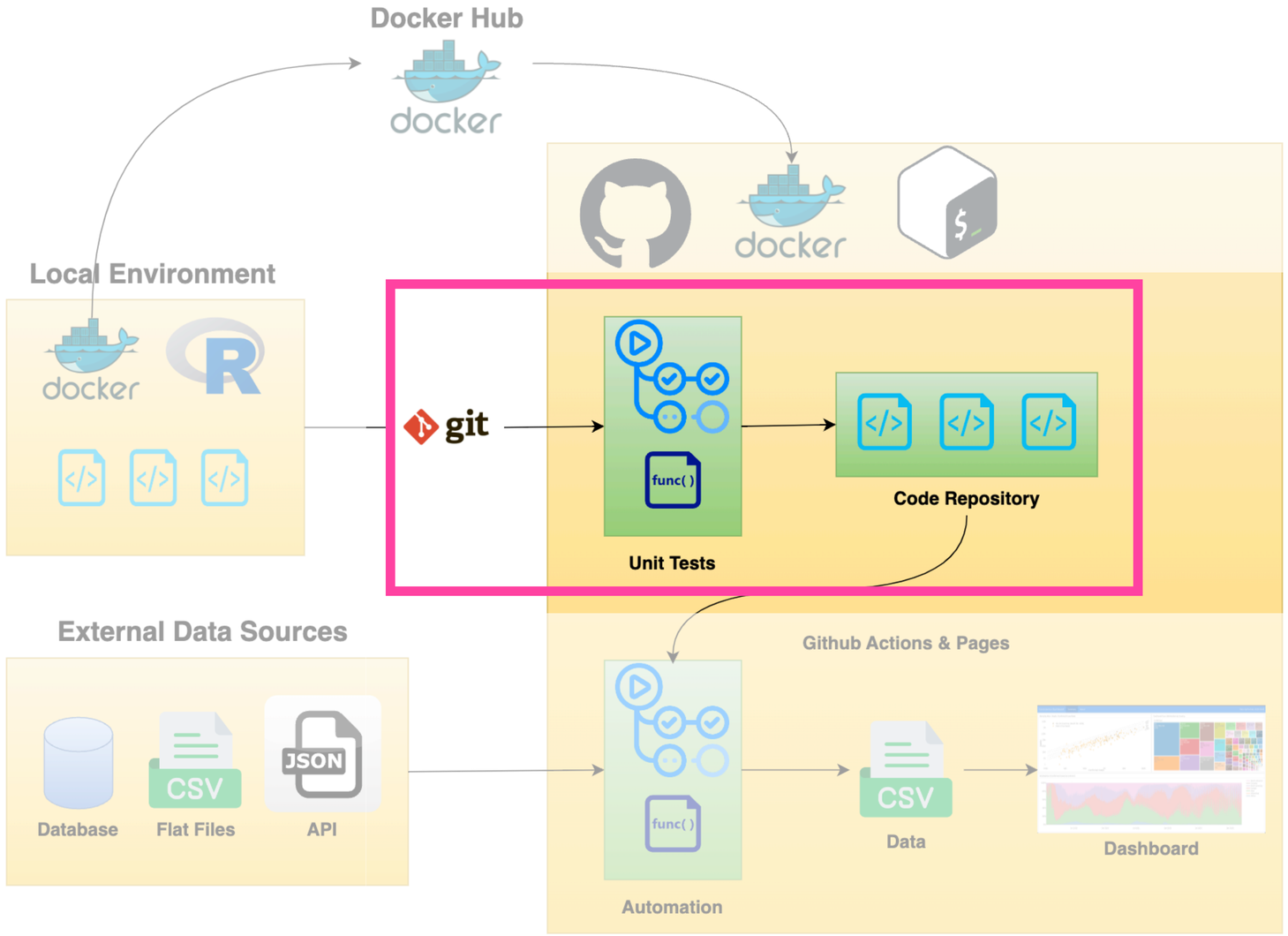
- Creating a docker image
- RStudio-Server
- VScode

General Architecture



- Docker Hub

General Architecture

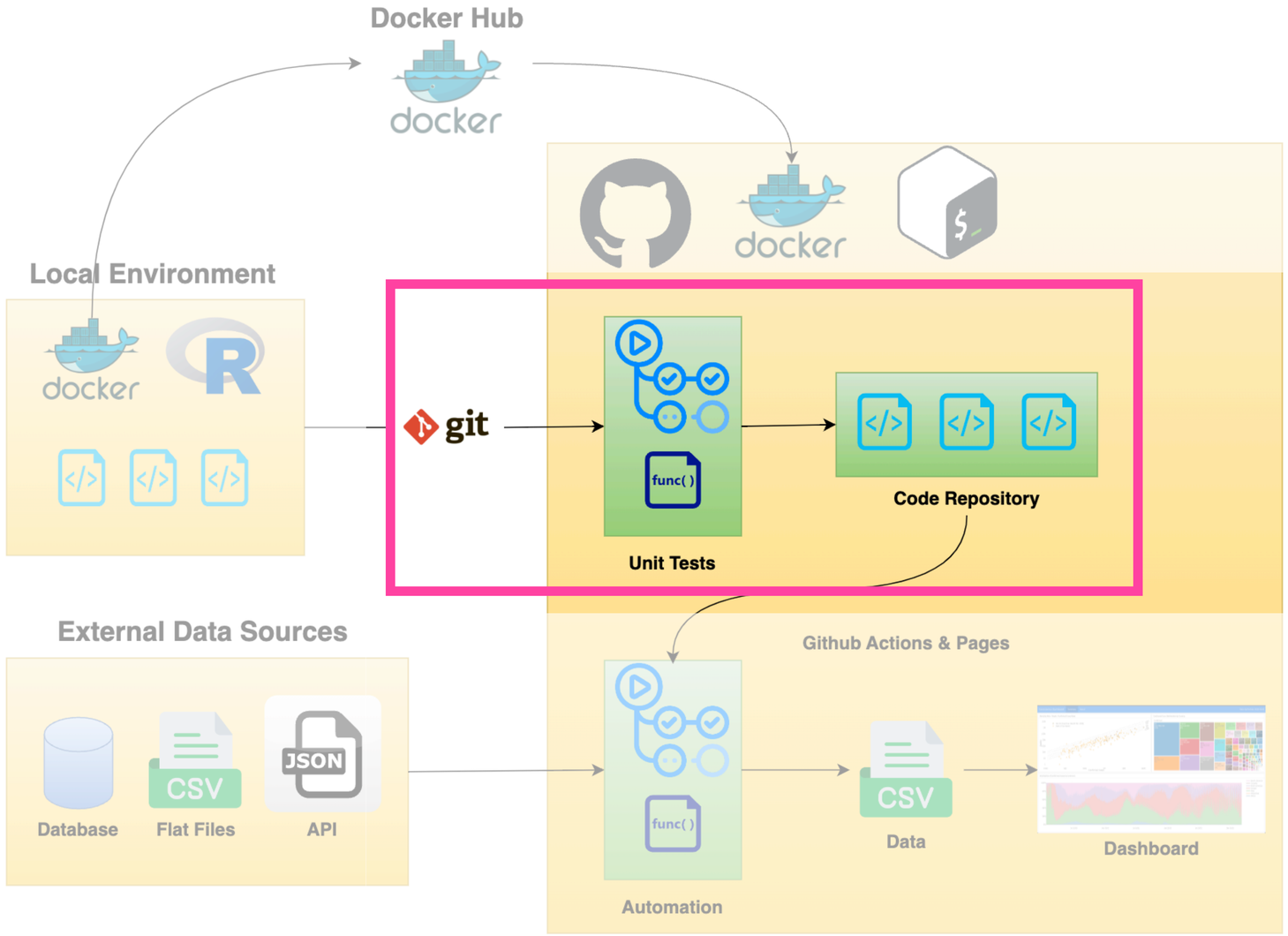


- Github Actions
- Unit tests

```yaml
on: [push, pull_request]

name: R CMD
jobs:
  R-CMD-check:
    name: R CMD check
    runs-on: ubuntu-18.04
    container:
        image: docker.io/rkrispin/coronavirus:prod.0.3.31
    steps:
      - name: checkout_repo
        uses: actions/checkout@v2
      - name: Check
        run: Rscript -e "rcmdcheck::rcmdcheck(args = '--no-manual', error_on = 'error')"
```

General Architecture

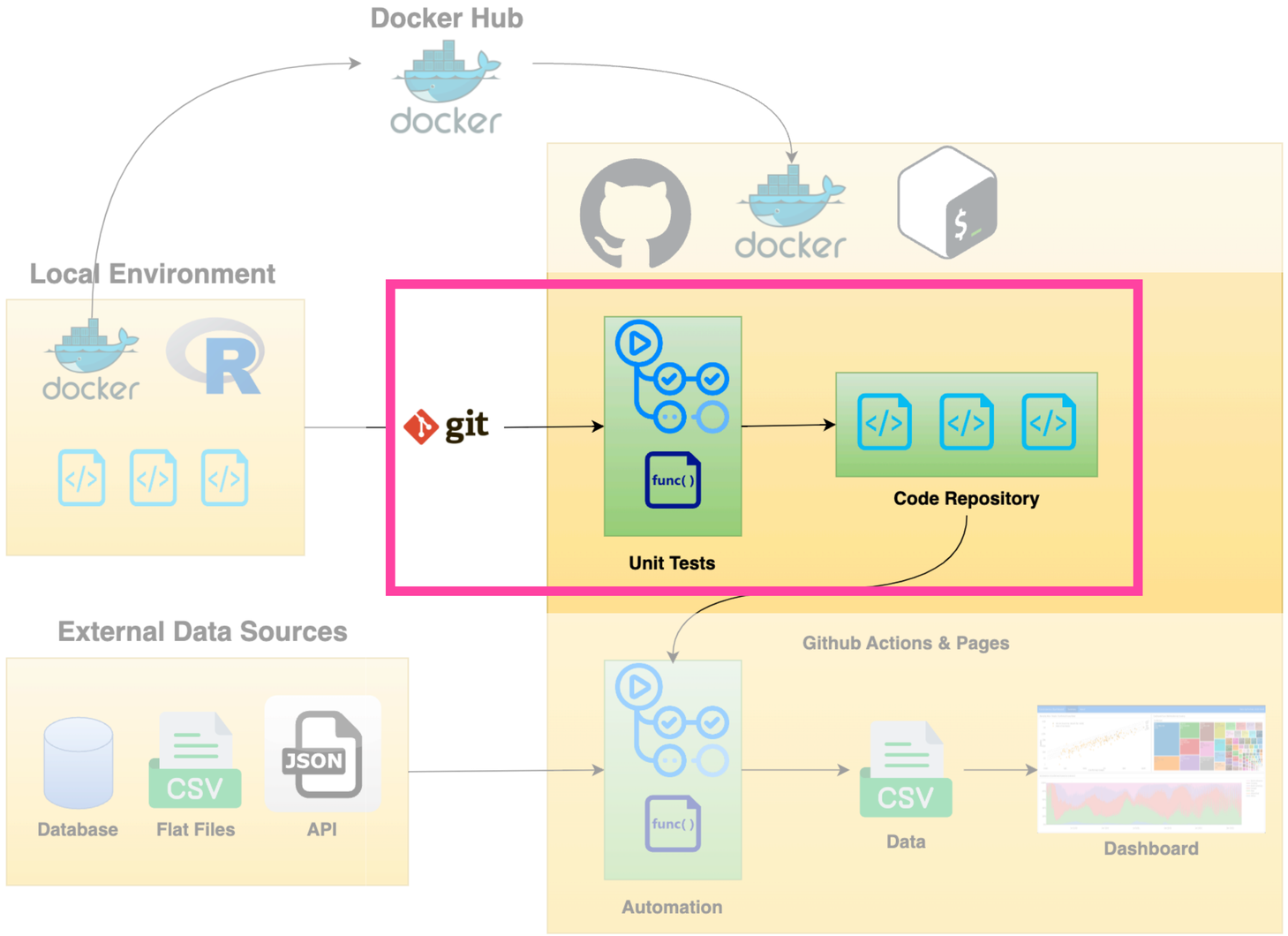


- Github Actions
- Unit tests

```yaml
on: [push, pull_request]

name: R CMD
jobs:
  R-CMD-check:
    runs-on: ubuntu-18.04
    container:
        image: docker.io/rkrispin/coronavirus:prod.0.3.31
    steps:
      - name: checkout_repo
        uses: actions/checkout@v2
      - name: Check
        run: Rscript -e "rcmdcheck::rcmdcheck(args = '--no-manual', error_on = 'error')"
```

General Architecture

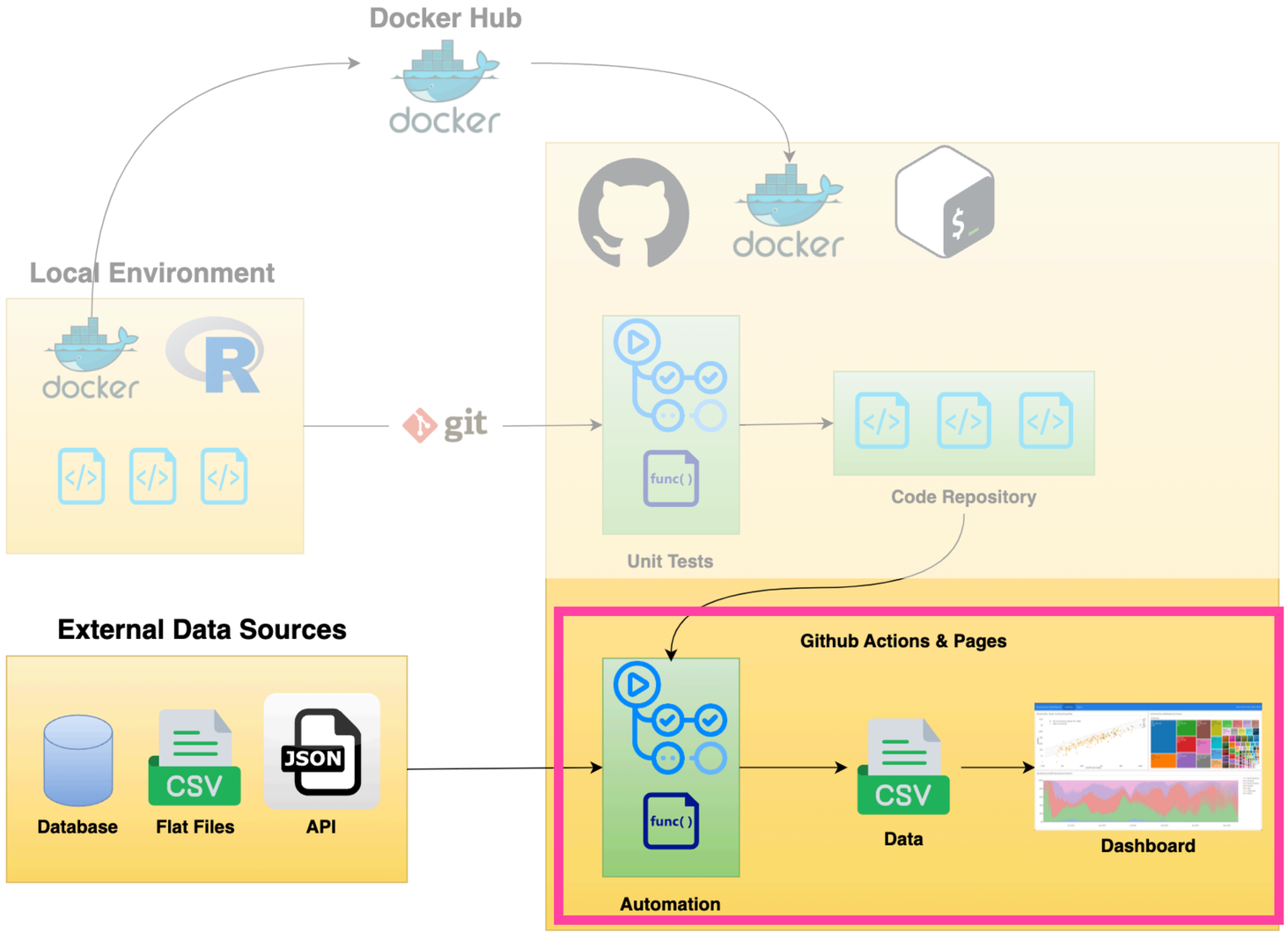


- Github Actions
- Unit tests

```yaml
on: [push, pull_request]

name: R CMD
jobs:
  R-CMD-check:
    name: R CMD check
    runs-on: ubuntu-18.04
    container:
        image: docker.io/rkrispin/coronavirus:prod.0.3.31
    steps:
       - name: checkout_repo
         uses: actions/checkout@v2
         name: Check
         run: Rscript -e "rcmdcheck::rcmdcheck(args = '--no-manual', error_on = 'error')"
```

General Architecture

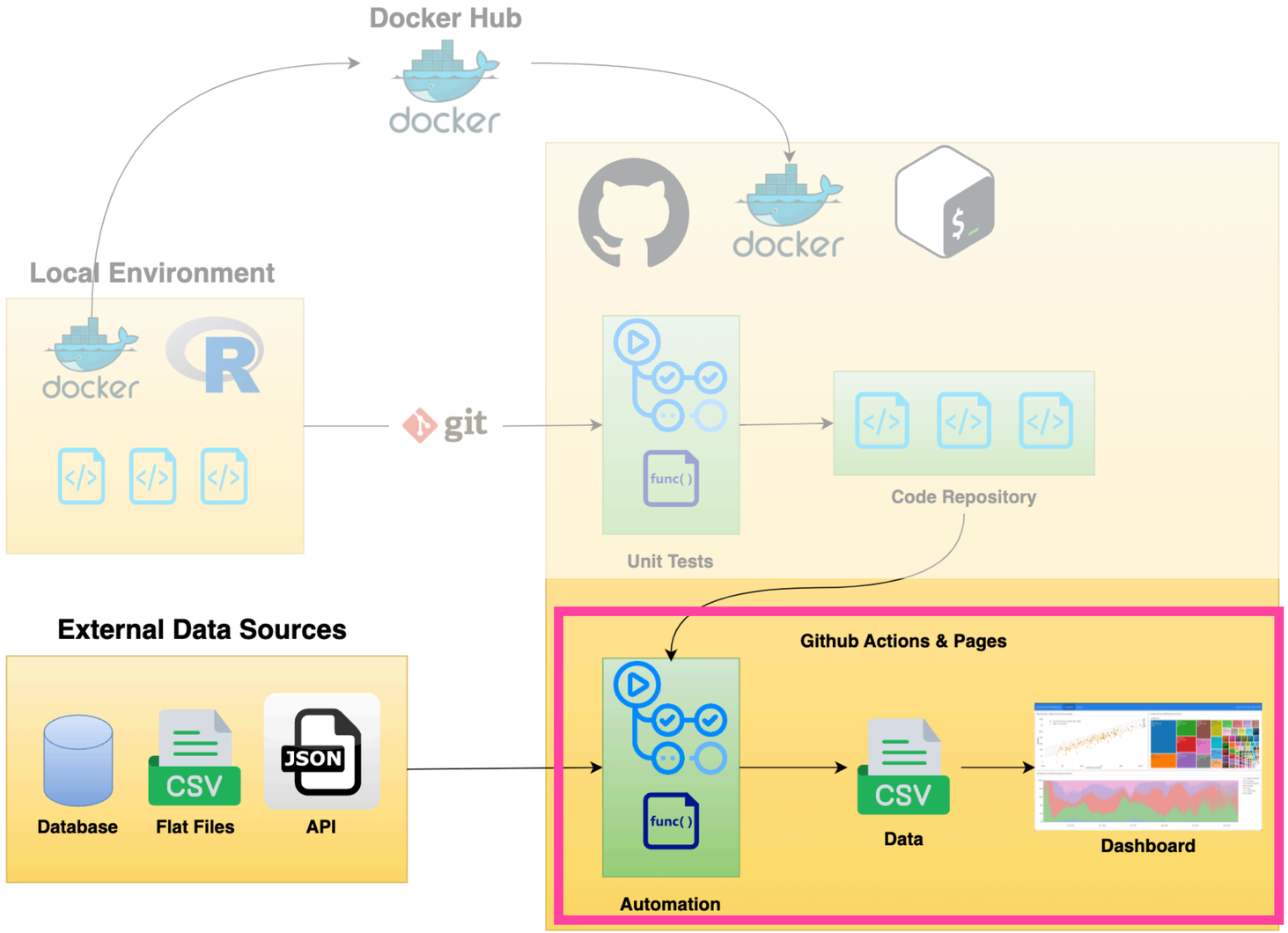


- Github Actions
- Bash helper script
- Github Pages
- Collecting Metadata

```yaml
name: Data Pipeline

on:
  schedule:
    - cron:  '0 */8 * * *'
jobs:
  data_refresh_staging:
    name: coronavirus dataset refresh main
    runs-on: ubuntu-18.04
    container:
     image: rkrispin/coronavirus:dev.0.3.34
    steps:
      - name: checkout_repo
        uses: actions/checkout@v2
        with:
          ref: 'main'
      - name: Refresh the data
        run: bash ./data_raw/data_refresh.sh "main"
```

General Architecture

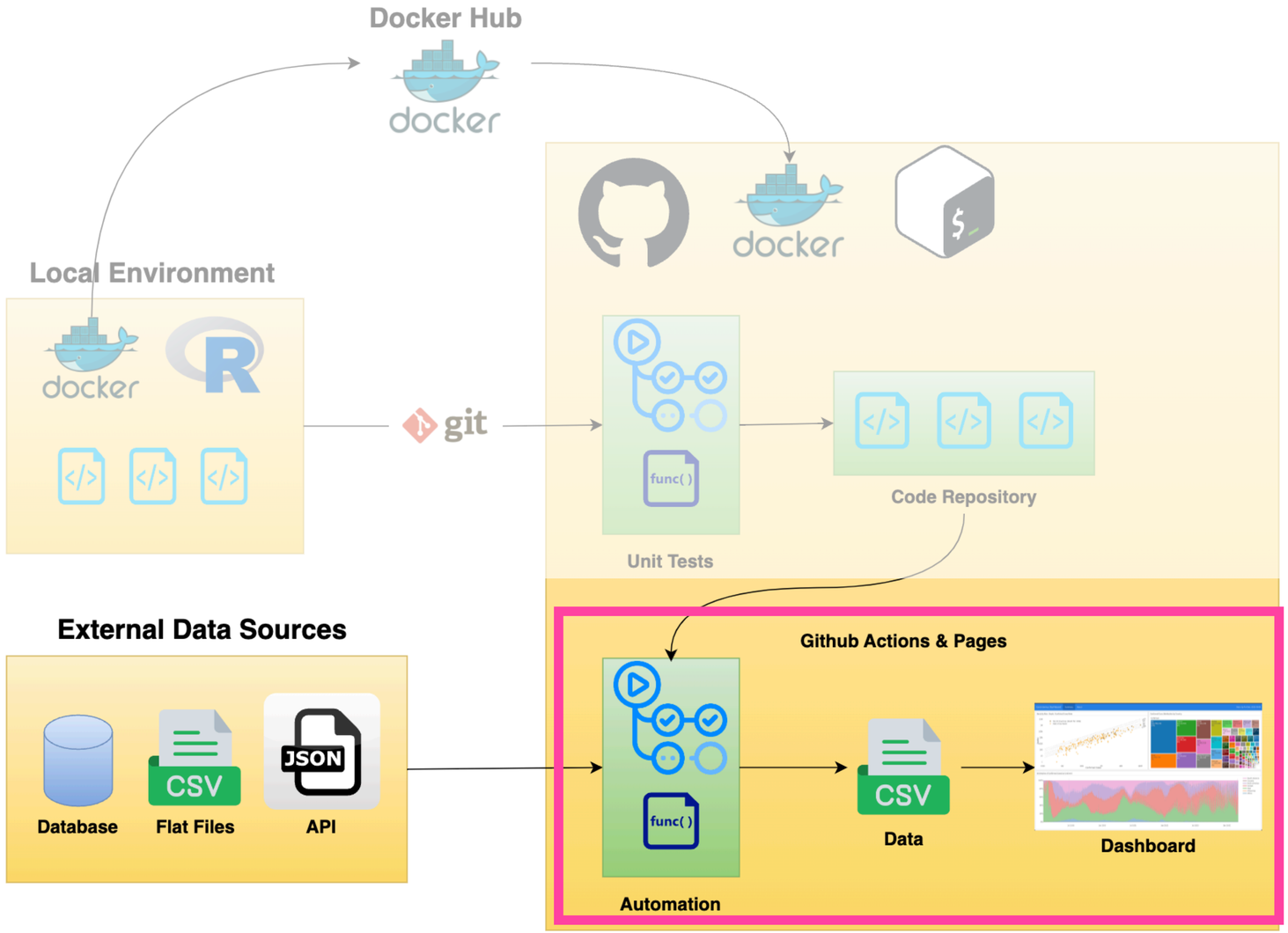


- Github Actions
- Bash helper script
- Github Pages
- Collecting Metadata

```yaml
name: Data Pipeline

on:
  schedule:
    - cron:  '0 */8 * * *'
jobs:
  data_refresh_staging:
    name: coronavirus dataset refresh main
    runs-on: ubuntu-18.04
    container:
     image: rkrispin/coronavirus:dev.0.3.34
    steps:
      - name: checkout_repo
        uses: actions/checkout@v2
        with:
          ref: 'main'
      - name: Refresh the data
        run: bash ./data_raw/data_refresh.sh "main"
```

General Architecture



- Github Actions
- Bash helper script
- Github Pages
- Collecting Metadata

```yaml
name: Data Pipeline

on:
  schedule:
    - cron:  '0 */8 * * *'
jobs:
  data_refresh_staging:
    name: coronavirus dataset refresh main
    runs-on: ubuntu-18.04
    container:
     image: rkrispin/coronavirus:dev.0.3.34
    steps:
      - name: checkout_repo
        uses: actions/checkout@v2
        with:
          ref: 'main'
      - name: Refresh the data
        run: bash ./data_raw/data_refresh.sh "main"
```

Unit Tests

Table Info - coronavirus

Number of columns: 15
Number of rows: 973836
Duplicated rows: 0

cols_name	cols_class	cols_NAs	cols_min	cols_max	cols_unique
date	Date	0			1143
province	character	680085			92
country	character	0			201
lat	numeric	5715	-71.9499	71.7069	289
long	numeric	5715	-178.1165	178.065	290
type	character	0			3
cases	numeric	0	-6298082	1354505	21969
uid	numeric	36576	4	15699	280
iso2	character	46863			223
iso3	character	46863			223

Unit Tests

Data validation

```r
# Checking merge
if(nrow(coronavirus) != nrow(coronavirus_temp)){
  s <- FALSE
  msg <- c(msg, "Merge fail - the number of rows of the coronavirus_temp is differe
}

# Checking the table dimensions
if(nrow(coronavirus) < 900000){
  s <- FALSE
  msg <- c(msg, "The number of rows of the coronavirus table is too small")
}

if(ncol(coronavirus) != 15){
  s <- FALSE
  msg <- c(msg, "The number of columns of the coronavirus table is invalid")
}

if(min(coronavirus$date) != as.Date("2020-01-22")){
  s <- FALSE
  msg <- c(msg, "The starting date is invalid")
}
```

Metadata

Saving the data

```
branch <- system(command = "git rev-parse --abbrev-ref HEAD", intern = TRUE)
load(sprintf("../data_pipelines/log_%s.RData", branch))
tail(log)
```

```
                  time         dataset  nrows  last_date update success
410 2023-03-19 16:24:45     coronavirus 973836 2023-03-09  FALSE    TRUE
411 2023-03-19 16:24:57 covid19_vaccine 142597 2023-03-09  FALSE    TRUE
412 2023-03-20 01:37:01     coronavirus 973836 2023-03-09  FALSE    TRUE
413 2023-03-20 01:37:11 covid19_vaccine 142597 2023-03-09  FALSE    TRUE
414 2023-03-20 08:25:07     coronavirus 973836 2023-03-09  FALSE    TRUE
415 2023-03-20 08:25:16 covid19_vaccine 142597 2023-03-09  FALSE    TRUE
    backfile branch
410    FALSE   main
411    FALSE   main
412    FALSE   main
413    FALSE   main
414    FALSE   main
415    FALSE   main
```

No Free Lunches

Limitations

- Resources
- Support
- Uncertainty

Summary

- Treat your open-source projects like your work projects
- Docker - steep learning curve, long term benefits
- Unit tests - insurance your work
- Make it visual!
- Knowledge sharing!

Tutorials

- **Setting a dockerized Python development environment with VScode -** https://github.com/RamiKrispin/vscode-python
- **Deploy a flexdashboard on Github Pages with Github Actions and Docker -** https://github.com/RamiKrispin/deploy-flex-actions
- **Setting a dockerized R development environment with VScode (WIP) -** https://github.com/RamiKrispin/vscode-r

Get in Touch

Questions?

Thank You!