

# Deploy and Monitor ML Pipelines with Python, Docker, and GitHub Actions

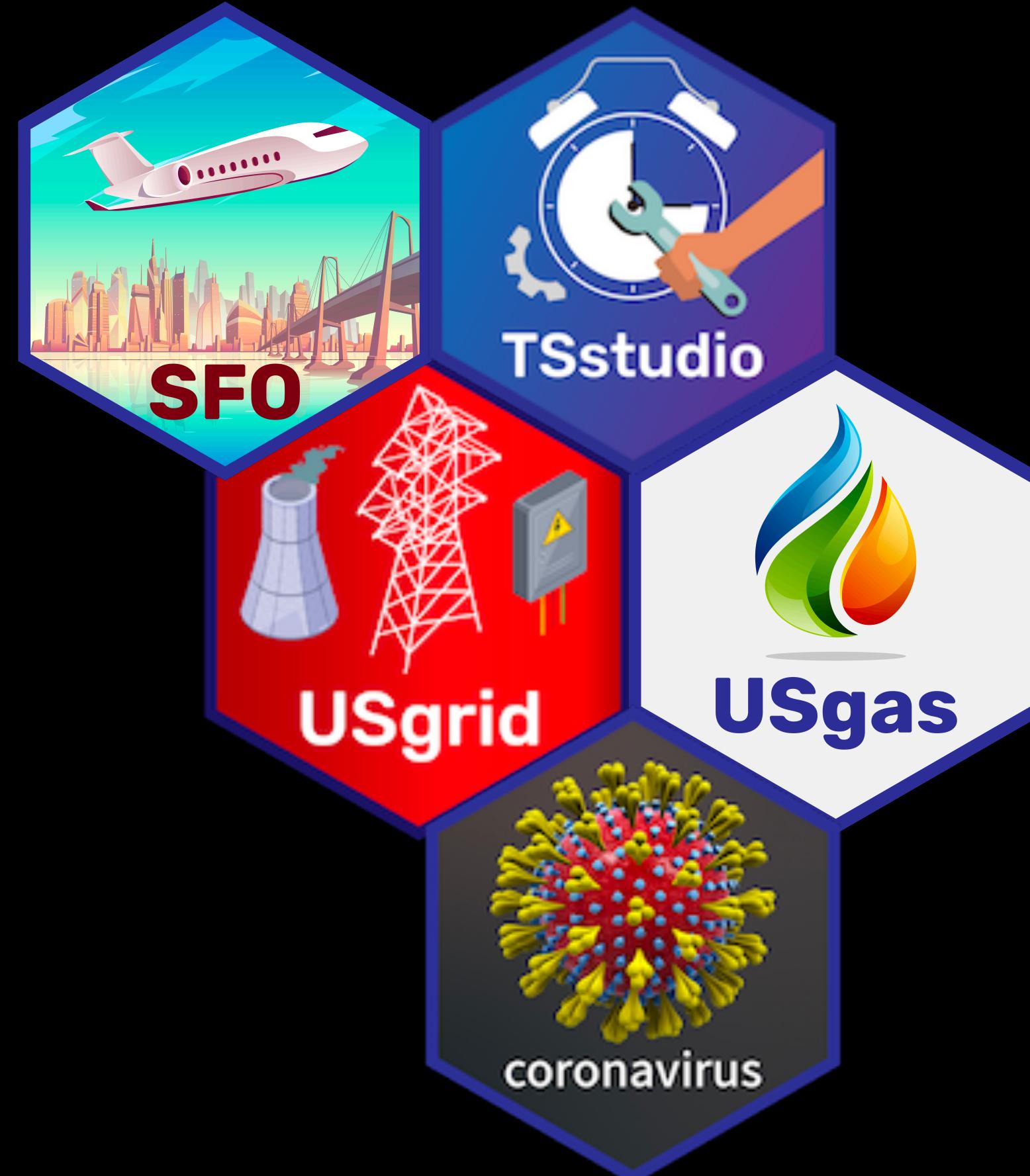


PyData NYC 2024 Conference

Rami Krispin, November 6th, 2024

# About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author
- Docker Captain 



# About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author
- Docker Captain 

Tutorials

 **vscode-python**

A Tutorial for Setting Python Development Environment with VScode and Docker

● Shell ⭐ 781 📂 69

 **vscode-r**

A Tutorial for Setting R Development Environment with VScode, Dev Containers, and Docker

● R ⭐ 224 📂 23

 **vscode-python-template** Template

A template for a dockerized Python development environment for VScode

● JavaScript ⭐ 83 📂 19

 **vscode-r-template** Template

A template for a dockerized R development environment for VScode

● R ⭐ 10 📂 3

 **ollama-poc**

Getting started with Ollama for Python - a short tutorial for setting up Ollama for Python

● Python ⭐ 80 📂 11

 **lang2sql**

A tutorial for setting an SQL code generator with the OpenAI API

● Jupyter Notebook ⭐ 238 📂 33

 **Introduction-to-Docker**

(WIP) Getting started with Docker - An introduction to Docker with data science and engineering applications

⭐ 124 📂 11

 **deploy-flex-actions**

Deploying flexdashboard on Github Pages with Docker and Github Actions

● HTML ⭐ 210 📂 28

 **shinylive-r**

A guide for deploying Shinylive R application into Github Pages

● R ⭐ 135 📂 22

 **shinylive**

A guide for deploying Shinylive Python application into Github Pages

● HTML ⭐ 129 📂 18

# About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author
- Docker Captain 

Add a note...

 Rami Krispin  in Towards Data Science

**Setting A Dockerized Python Environment —The Hard Way**

This post will review different methods to run a dockerized Python environment from the command line (CLI). Am I...

Feb 13 543 6



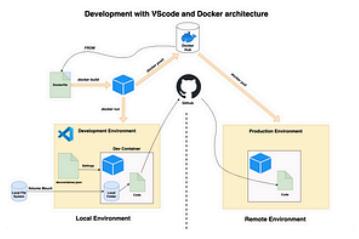
Add a note...

 Rami Krispin 

**Setting a Dockerized Python Development Environment Template**

In this post, we will review how to set, with a few simple steps, a dockerized Python development environment with VScode and...

Jan 13 116 2



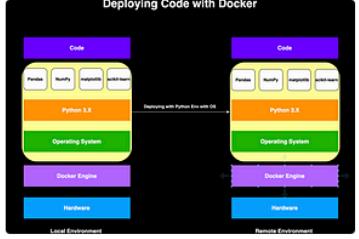
Add a note...

 Rami Krispin 

**Running Python/R with Docker vs. Virtual Environment**

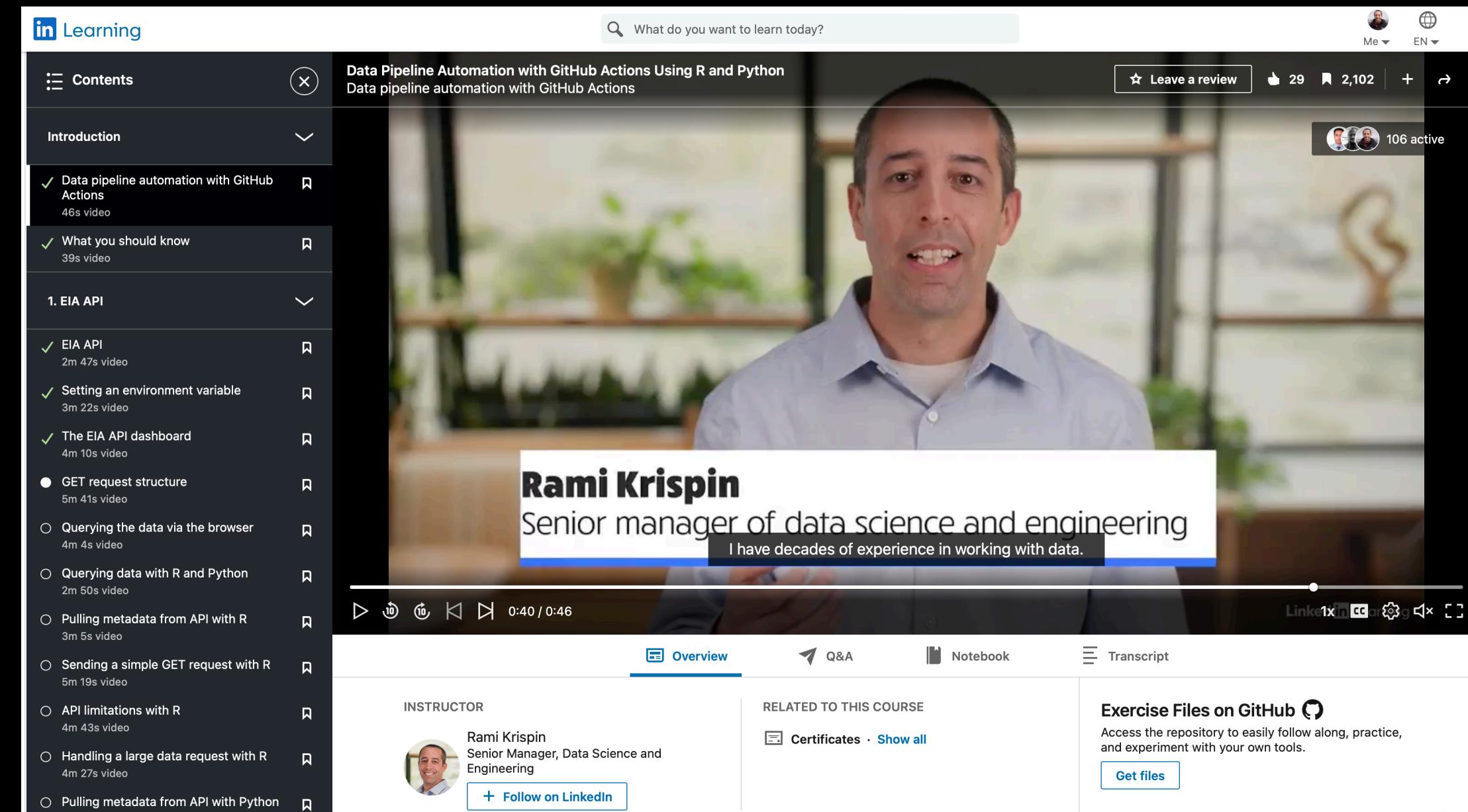
Someone asked me last week about the benefit of developing Python on a dockerized environment vs. using a virtual...

Jun 5, 2023 68 1



# About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author
- Docker Captain 



The screenshot shows a LinkedIn Learning course titled "Data Pipeline Automation with GitHub Actions Using R and Python". The course has 2,102 students and 106 active users. The instructor, Rami Krispin, is a Senior manager of data science and engineering with decades of experience in working with data. The course content includes an introduction and several video lessons on EIA API, setting environment variables, and automating data pipelines with GitHub Actions.

# About Me

- Senior Manager
- **Forecasting**
- **MLOps**
- **Open Source**
- Author
- Docker Captain 

[https://github.com/RamiKrispin/  
pydata-ny-ga-workshop/](https://github.com/RamiKrispin/pydata-ny-ga-workshop/)

# Poll

## Are You Familiar with?

- Docker
- GitHub Actions
- Quarto
- Forecasting

# Docker

Docs Get support Contact sales

Products Developers Pricing Support Blog Company

Sign In Get started

buildcloud

## Docker Builds: Now Lightning Fast

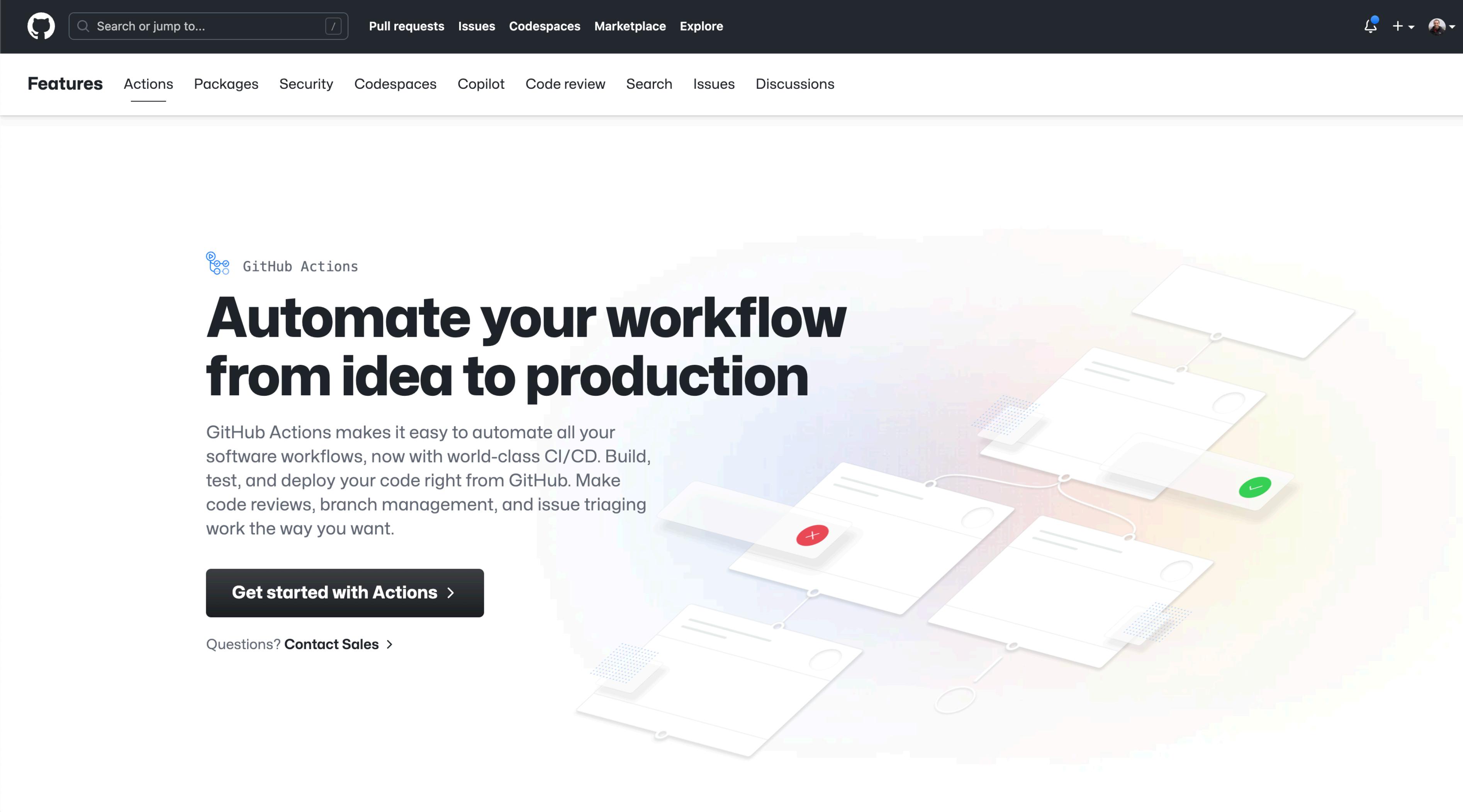
Announcing Docker Build Cloud general availability

Discover Docker Build Cloud

What is Docker?

### Accelerate how you build, share, and run applications

# GitHub Actions



The screenshot shows the GitHub Actions landing page. At the top, there's a navigation bar with the GitHub logo, a search bar, and links for Pull requests, Issues, Codespaces, Marketplace, and Explore. On the right side of the bar are icons for notifications, a plus sign, and a user profile. Below the navigation is a secondary navigation bar with tabs for Features, Actions (which is selected and underlined), Packages, Security, Codespaces, Copilot, Code review, Search, Issues, and Discussions. The main content area features a large, semi-transparent background image of several white cards connected by arrows, representing a workflow. In the upper left of this image, there's a small GitHub Actions icon and the text "GitHub Actions". The central text reads "Automate your workflow from idea to production" in a large, bold, dark font. Below this, a paragraph explains what GitHub Actions can do: "GitHub Actions makes it easy to automate all your software workflows, now with world-class CI/CD. Build, test, and deploy your code right from GitHub. Make code reviews, branch management, and issue triaging work the way you want." At the bottom left, there's a black button with white text that says "Get started with Actions >". Below that, smaller text says "Questions? Contact Sales >".

# Quarto

## Hello, Quarto

Python    R    Julia    Observable

Combine Jupyter notebooks with flexible options to produce production quality output in a wide variety of formats. Author using traditional notebook UIs or with a plain text markdown representation of notebooks.

The screenshot shows a Jupyter Notebook interface with a title cell "Palmer Penguins" and a code cell containing Python code to read a CSV file. Below the code is a section titled "Exploring the Data" with a note about a figure. A scatter plot of bill length vs bill depth for three penguin species is displayed.

```
author: Norah Jones
format:
  html:
    code-tools: true
    code-fold: true

[3]: #| echo: false
import pandas as pd
df = pd.read_csv("palmer-penguins.csv")
```

Exploring the Data

See @fig-bill-sizes for an exploration of bill sizes by species.

```
[6]: #| label: fig-bill-sizes
#| fig-cap: Bill Sizes by Species
import matplotlib.pyplot as plt
import seaborn as sns
g = sns.FacetGrid(df, hue="species", height=3, aspect=3.5/1.5)
g.map(plt.scatter, "bill_length_mm", "bill_depth_mm").add_legend()
```

[6]: <seaborn.axisgrid.FacetGrid at 0x2946720e0>

A scatter plot showing the relationship between bill length (mm) on the x-axis and bill depth (mm) on the y-axis. The data points are color-coded by species: Adelie (blue), Gentoo (orange), and Chinstrap (green). The x-axis ranges from approximately 35 to 60 mm, and the y-axis ranges from approximately 14 to 22 mm.

### Palmer Penguins

AUTHOR  
Norah Jones

PUBLISHED  
March 12, 2023

Show All Code    Hide All Code

Exploring the Data

See [Figure 1](#) for an exploration of bill sizes by species.

▼ Code

```
import matplotlib.pyplot as plt
import seaborn as sns
g = sns.FacetGrid(df, hue="species", height=3, aspect=3.5/1.5)
g.map(plt.scatter, "bill_length_mm", "bill_depth_mm").add_legend()
```

A faceted scatter plot titled "Palmer Penguins" showing bill length (mm) on the x-axis and bill depth (mm) on the y-axis. The facets represent different penguin species: Adelie (blue), Gentoo (orange), and Chinstrap (green). The x-axis ranges from 35 to 60 mm, and the y-axis ranges from 14 to 22 mm. The plot shows that Adelie penguins tend to have longer bills and deeper bills compared to Gentoo and Chinstrap penguins.

### Dynamic Documents

Generate dynamic output using Python, R, Julia,

### Beautiful Publications

Publish high-quality articles, reports,

### Scientific Markdown

Pandoc markdown has excellent support for

# Nixtla mlforecast



Search or ask...



Hyperparameter optimization

Transforming exogenous features

Cross validation

Probabilistic forecasting

Target transformations

Using scikit-learn pipelines

Analyzing the trained models

MLflow

Custom training

Training with numpy arrays

One model per step

Custom date features

Predict callbacks

Predicting a subset of ids

Transfer Learning

## Tutorials

### Electricity Load Forecast

Detect Demand Peaks

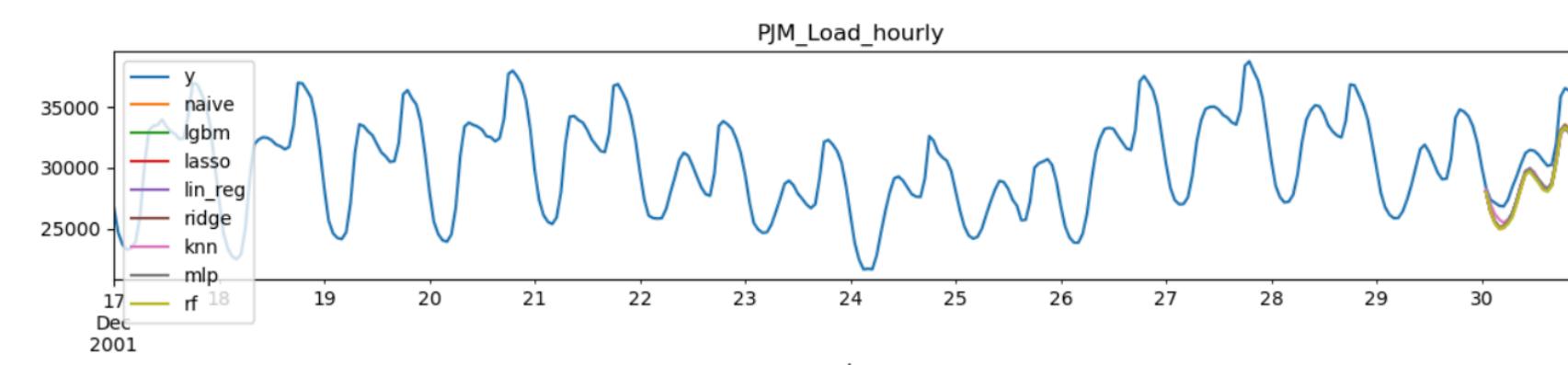
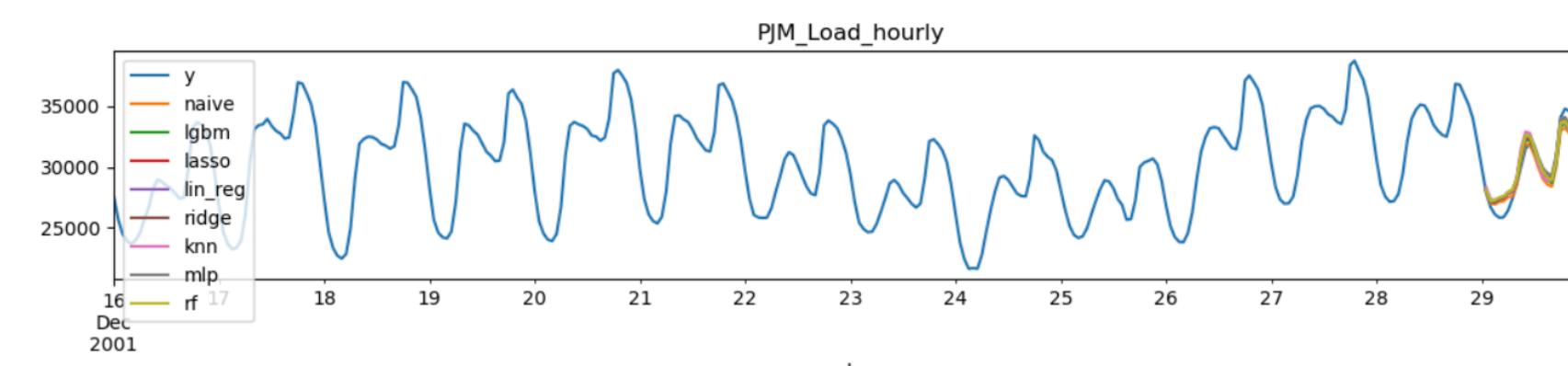
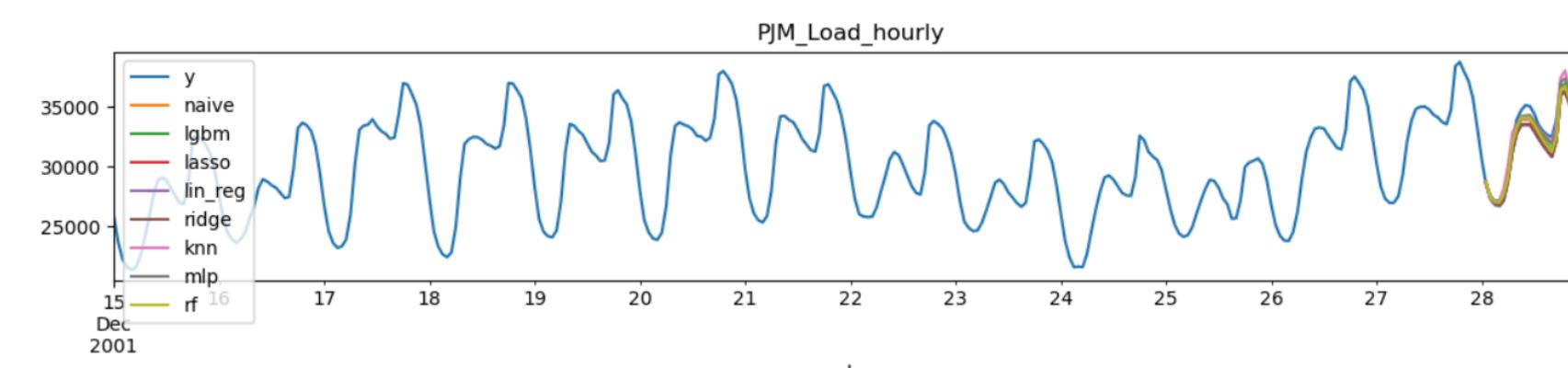
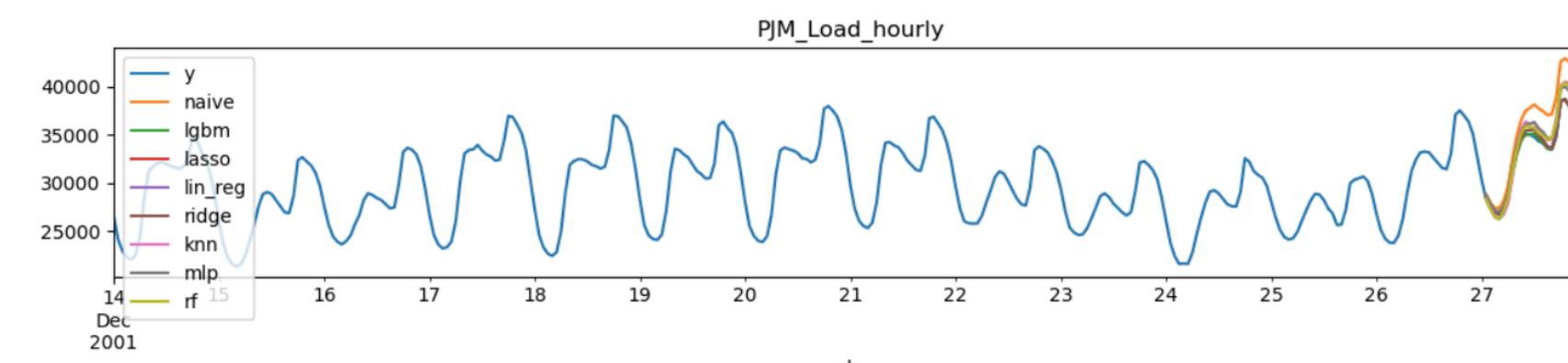
Prediction intervals

## API Reference

Local >

Distributed >

```
plot_cv(df_train, crossvalidation_df, 'PJM_Load_hourly', 'load_forecasting__pr [REDACTED]
```



≡ On this page

Introduction

Libraries

Forecast using Multiple Seasonalities

Electricity Load Data

Analizing Seasonalities

Model Selection with Cross-Validation

Test Evaluation

Comparison with Prophet

Ask AI

# Alternatives Tools

# Agenda

Motivation

Workflow

Design

Demo

# 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



Source:  
CNN

Kyle Terada/USA Today Sports

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



By Sandra Gonzalez, CNN

Updated 6:58 AM EST, Mon February 3, 2020



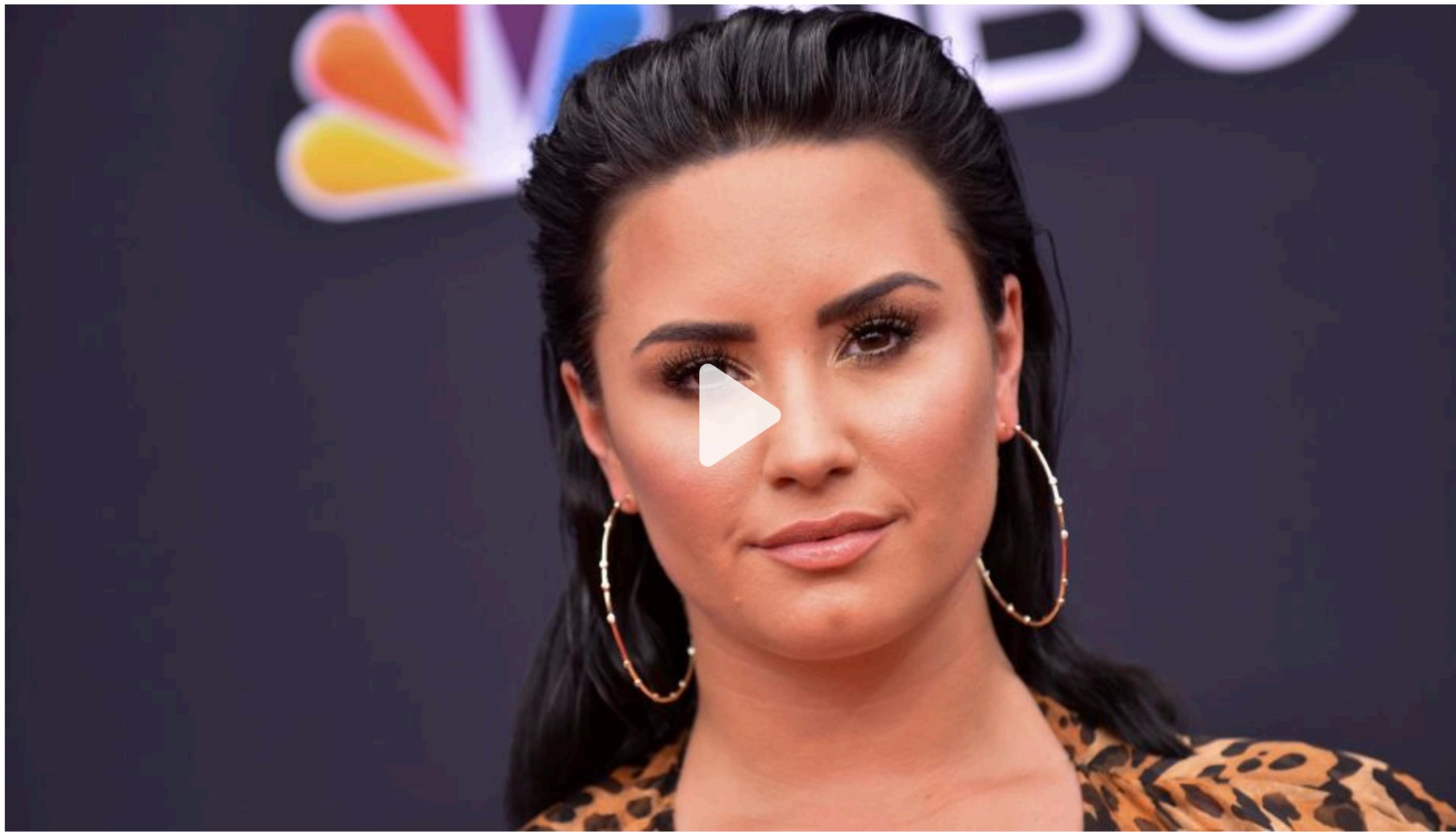
Source:  
CNN

# 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 Chadelis Duster, CNN

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



Source:  
CNN



**EVERYBODY PANIC!**



 **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 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](https://github.com/RamiKrispin/coronavirus)

36 110 1 1 1

 **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/](https://ramikrispin.github.io/coronavirus/)  
#rstats, #coronavirus, #data



The 2019 Novel Coronavirus COVID-19 (2019-nCoV)  
Provides a daily summary of the Coronavirus (COVID-19) cases by state/province. Data source: ...  
[ramikrispin.github.io](https://ramikrispin.github.io/coronavirus/)

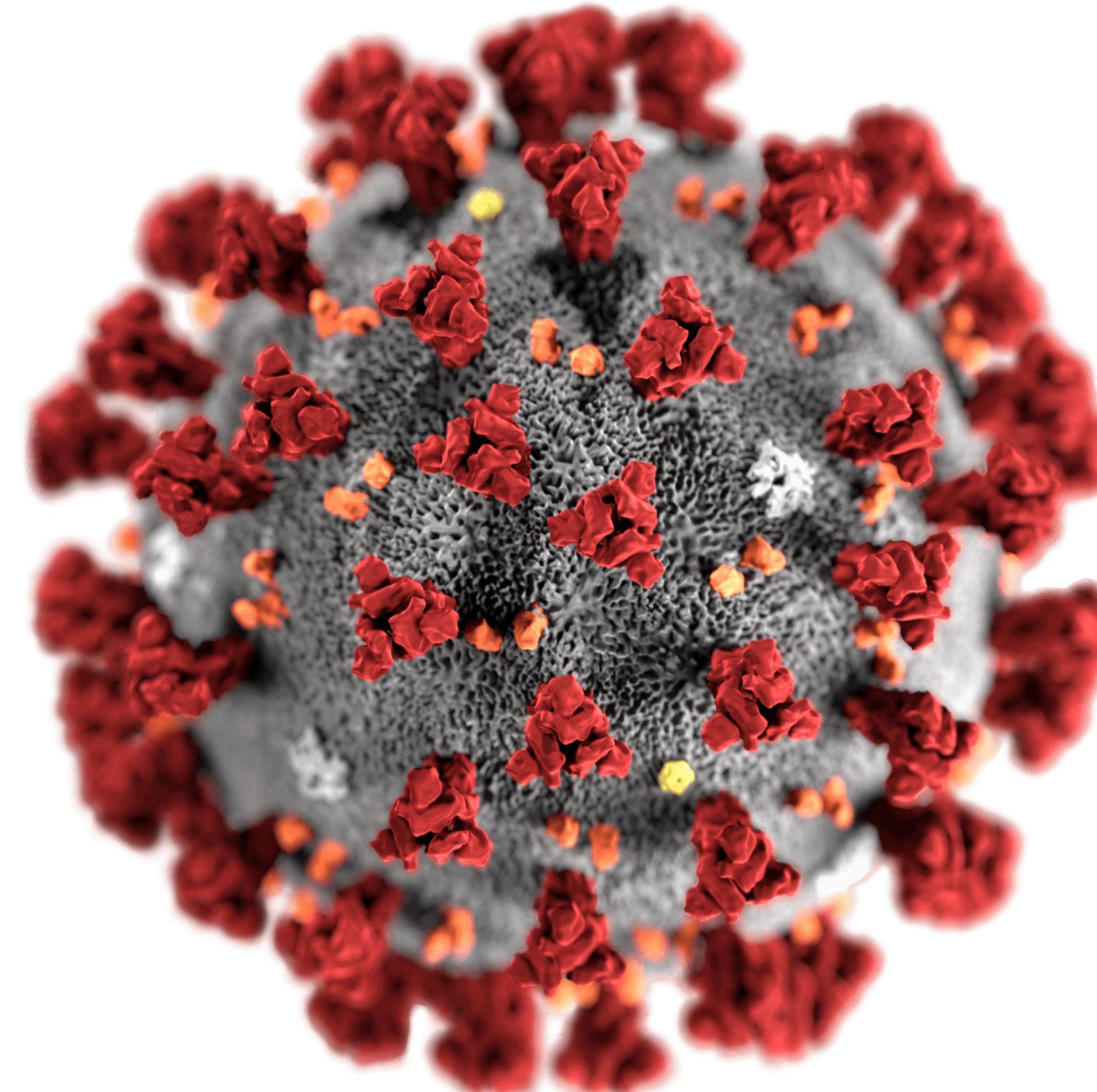
2 17 21 1 1

# coronavirus

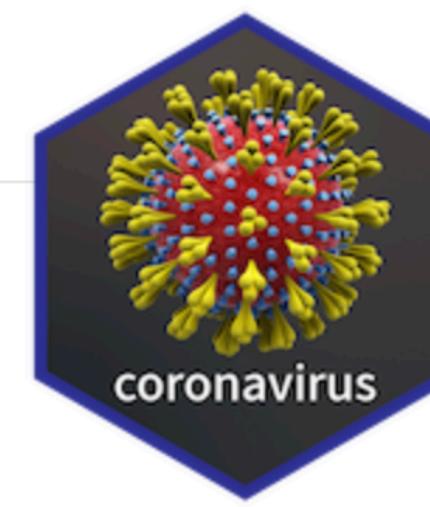
The coronavirus package provides a tidy format for the COVID-19 dataset collected by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The dataset includes daily new and death cases between January 2020 and March 2023 and recovery cases until August 2022.

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

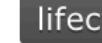
Author 

## Dev status



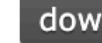
 Data Pipeline passing

 CRAN 0.4.1

 lifecycle stable

 License MIT

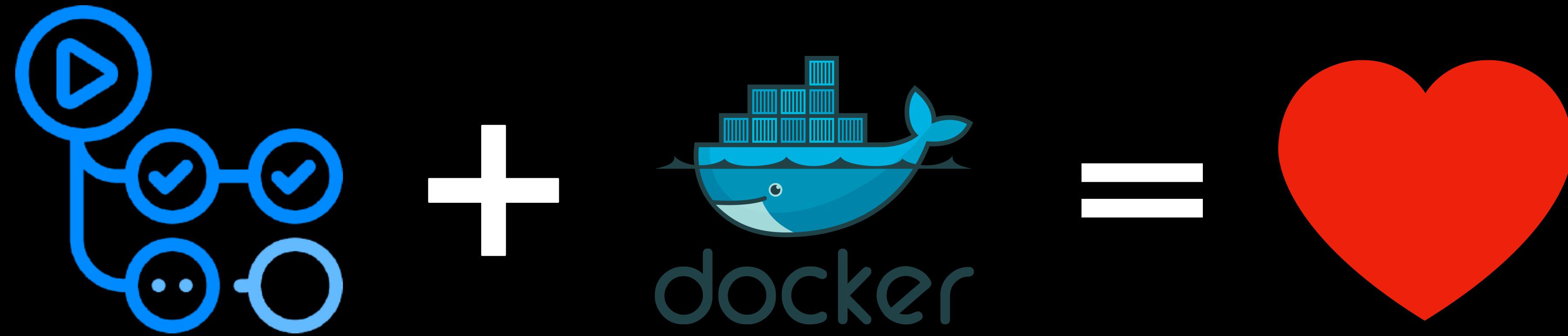
 last commit march

 downloads 74K

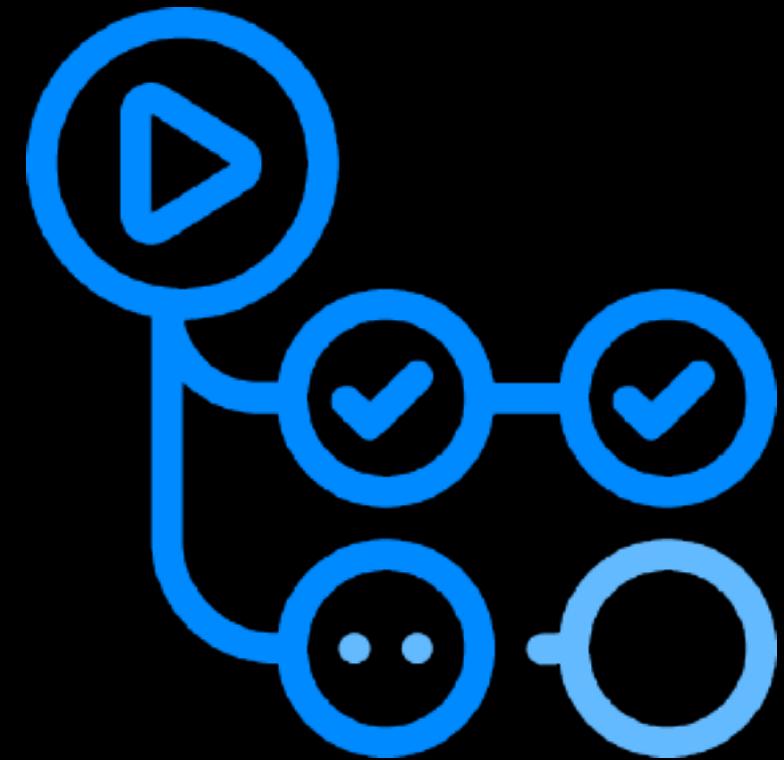
- Connecting through Power BI**  
#14 by jimmyd7377 was closed on Mar 20, 2020 1
- Error when devtools::install\_github("RamiKrispin/coronavirus")**  
#13 by Danielchui was closed on Mar 20, 2020 1
- Error in Hubei data for 3/11/2020**  
#12 by tcarleton was closed on Mar 13, 2020 2
- Country name in standard format**  
#11 by shubhrampandey was closed on Jul 1, 2020 12
- 0 Case Vectors (non-essential)**  
#10 by j4yr0u93 was closed on Mar 13, 2020 2
- How do u make it realtime and auto update**  
#9 by navmedvideos was closed on Apr 25, 2020 3
- Negative values were found in the package**  
#7 by ddong63 was closed on Mar 10, 2020 7
- Update package locally**  
#6 by shubhrampandey was closed on Mar 13, 2020 2
- Negative case values**  
#5 by j4yr0u93 was closed on Mar 10, 2020 2
- Covid-19**  
#4 by acgerstein was closed on Mar 6, 2020 3
- Adding a country filter to the dashboard**  
#3 by Agusum was closed on Mar 8, 2020 3
- " Azerbaijan" Has Space in the Name**  
#2 by cannin was closed on Feb 29, 2020 1
- is there any plan to automate data update?**  
#1 by statklee was closed on Apr 25, 2020 16



# Actions + Docker

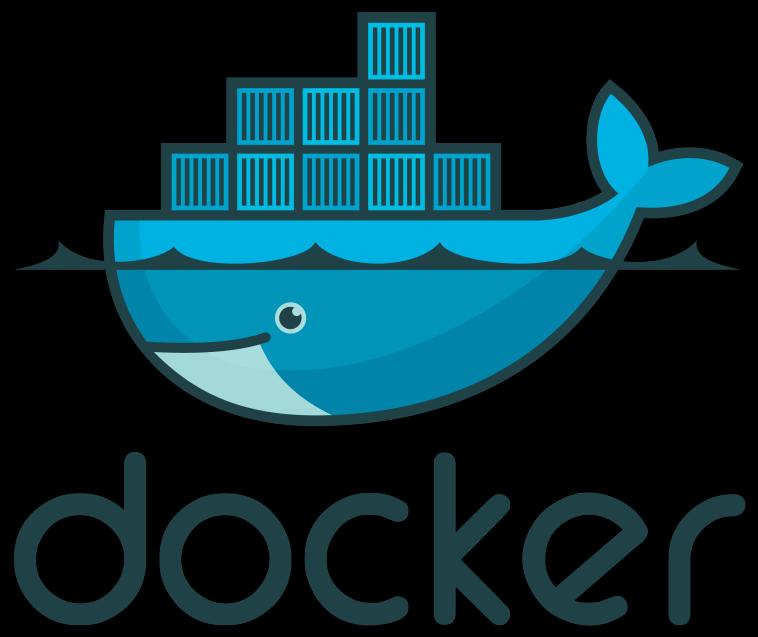


# Actions + Docker



## GitHub Actions

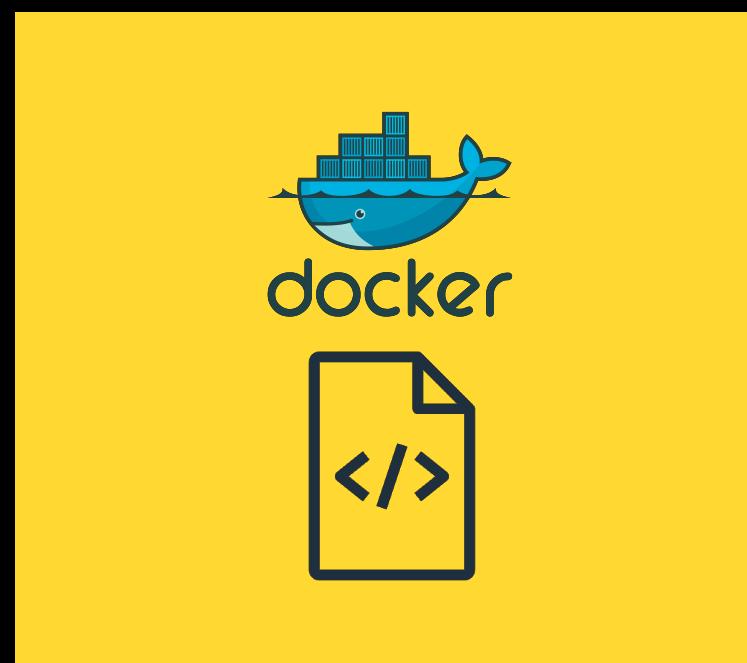
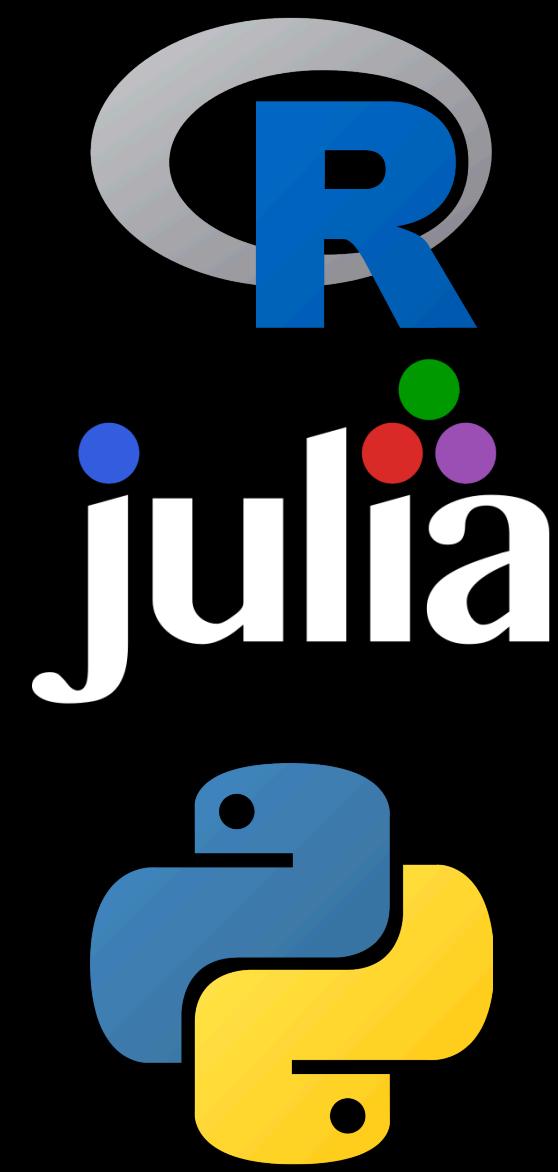
- Scheduler
- CI/CD



## Docker

- Container
- CI/CD

# Docker In A Nutshell

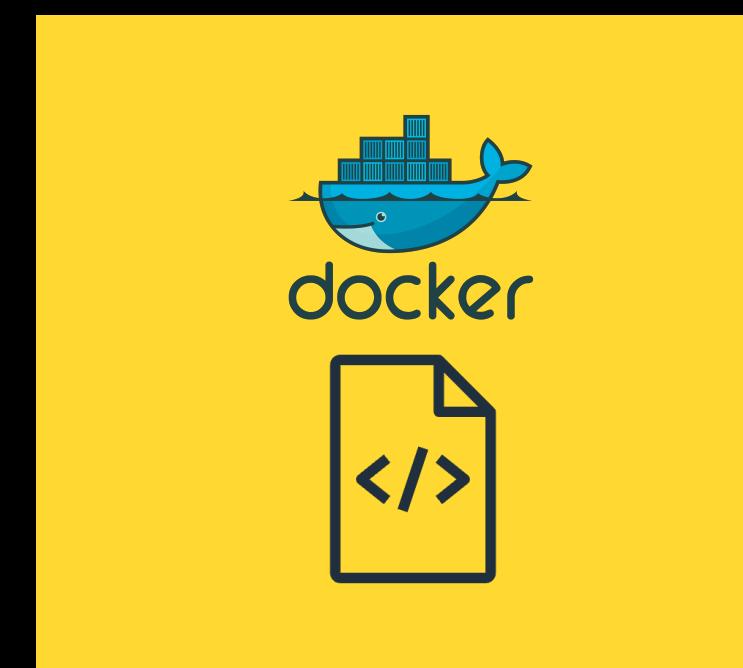
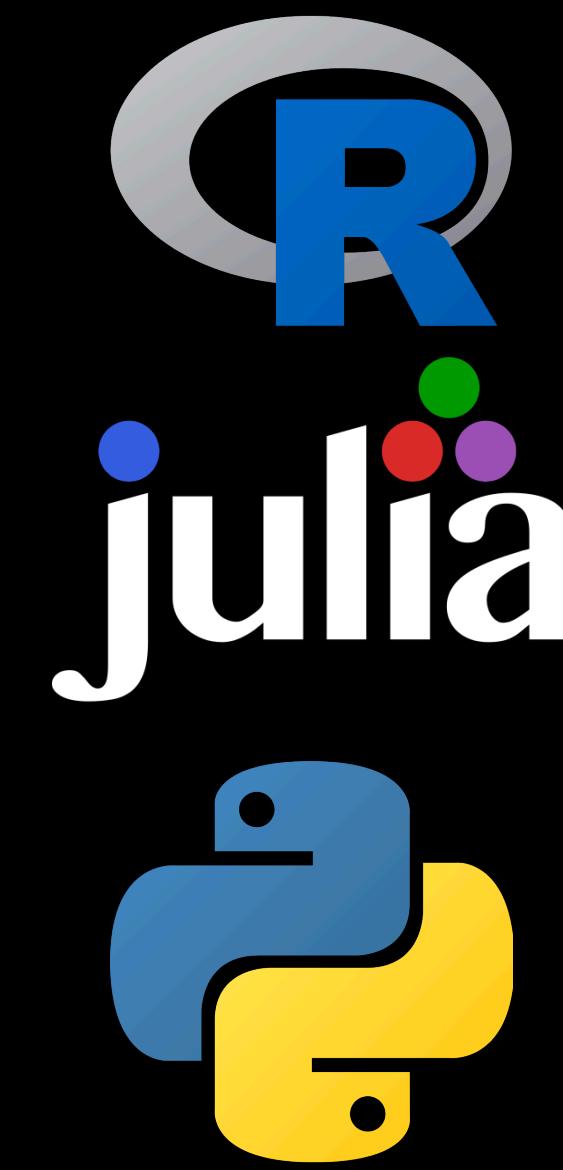


Local Env

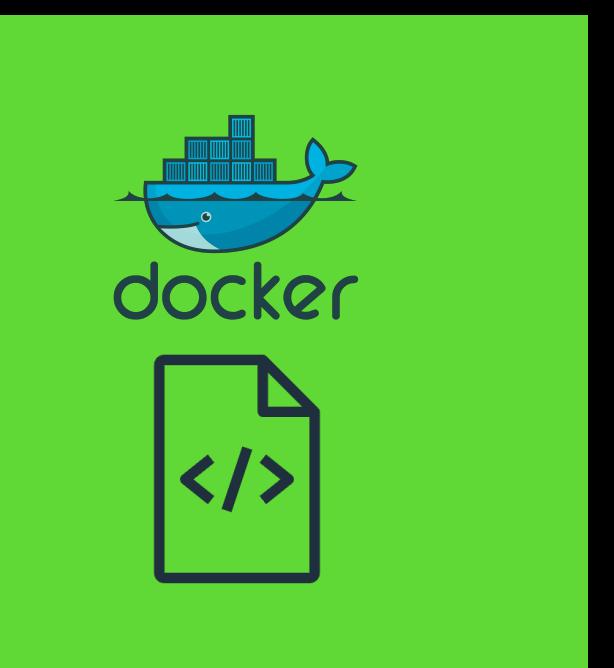


Remote Env

# Docker In A Nutshell



Local Env



Remote Env

# What Can You Build?

# CI/CD & Automation

The screenshot shows the GitHub Actions dashboard for the repository `RamiKrispin / eia-poc`. The `Actions` tab is selected. On the left, a sidebar lists actions: `Data Refresh`, `pages-build-deployment`, `Management`, `Caches`, `Deployments` (with a dropdown arrow), and `Runners`. The main area displays the `All workflows` section, which shows 1,488 workflow runs. The runs are listed in descending order of time, with the most recent at the top. Each run is identified by its name (e.g., `pages build and deployment` or `Data Refresh`), the event it triggered (`pages-build-deployment #644: by github-pages` or `Data Refresh #844: Scheduled`), the actor (`bot`), the branch it ran on (e.g., `main`), the time it started (e.g., 30 minutes ago, 32 minutes ago, etc.), and its status (e.g., 43s, 1m 11s). A search bar at the top right allows filtering of workflow runs.

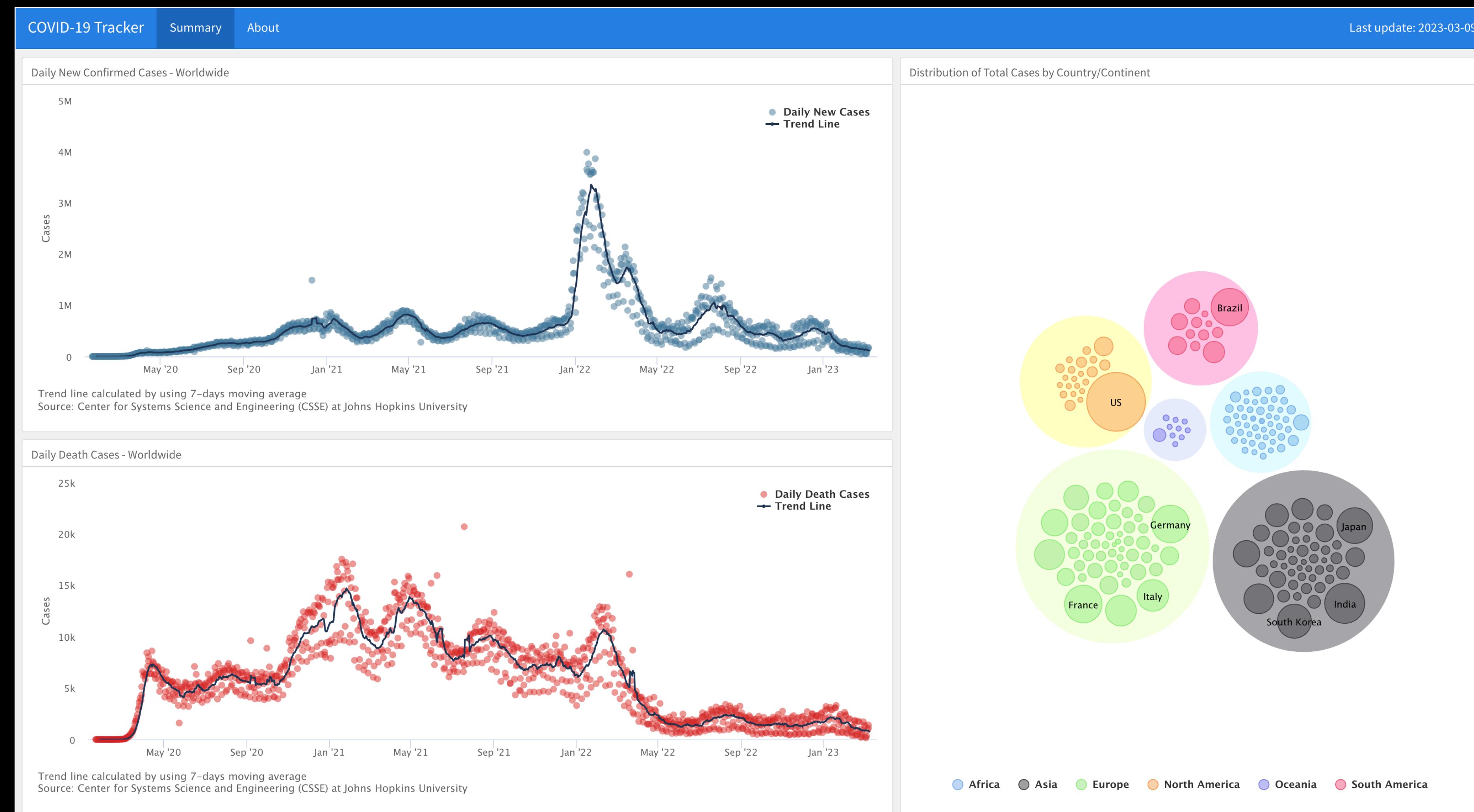
Workflow	Event	Status	Branch	Time Started	Actor
pages build and deployment	pages-build-deployment #644: by github-pages	bot	main	30 minutes ago	43s
Data Refresh	Data Refresh #844: Scheduled		main	32 minutes ago	1m 11s
pages build and deployment	pages-build-deployment #643: by github-pages	bot	main	2 hours ago	51s
Data Refresh	Data Refresh #843: Scheduled		main	2 hours ago	1m 11s
pages build and deployment	pages-build-deployment #642: by github-pages	bot	main	3 hours ago	49s
Data Refresh	Data Refresh #842: Scheduled		main	3 hours ago	1m 11s
pages build and deployment	pages-build-deployment #641: by github-pages	bot	main	4 hours ago	48s
Data Refresh	Data Refresh #841: Scheduled		main	4 hours ago	1m 15s

# Code Development and Testing

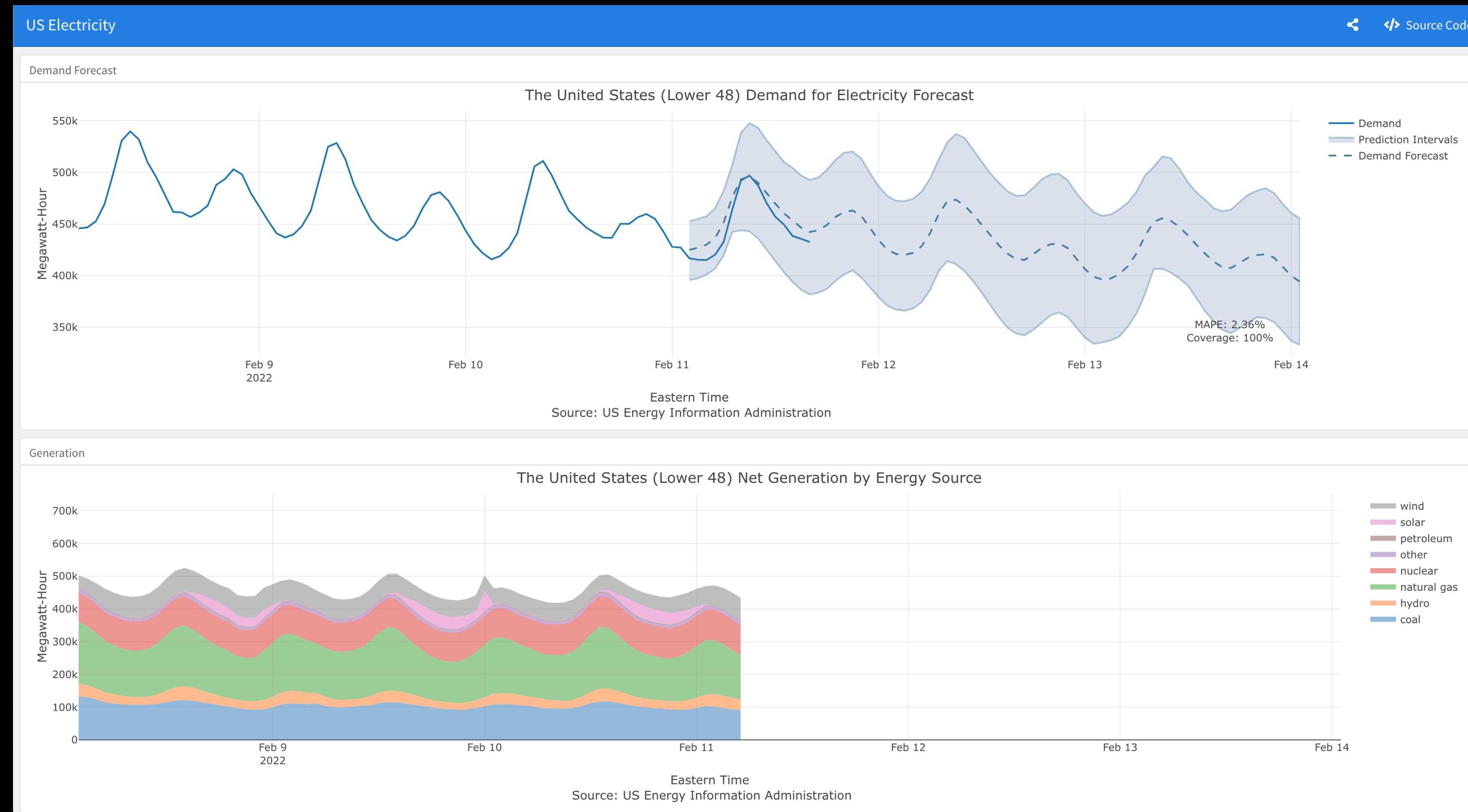
The screenshot shows a GitHub repository interface for the 'coronavirus' repository owned by 'RamiKrispin'. The 'Code' tab is selected in the navigation bar. On the left, the file tree shows various files and folders, with 'main.yml' currently selected. The main panel displays the content of 'main.yml':

```
coronavirus/.github/workflows/main.yml
rkrispin replacing branch, remove commit history ✓ aac85da · last year History
Code Blame 14 lines (13 loc) · 366 Bytes Code 55% faster with GitHub Copilot
1 on: [push, pull_request]
2
3 name: R CMD
4 jobs:
5   R-CMD-check:
6     name: R CMD check
7     runs-on: ubuntu-18.04
8     container:
9       image: docker.io/rkrispin/coronavirus:prod.0.3.31
10    steps:
11      - name: checkout_repo
12        uses: actions/checkout@v2
13      - name: Check
14        run: Rscript -e "rcmdcheck::rcmdcheck(args = '--no-manual', error_on = 'error')"
```

# Data Automation



# ML Applications



# Hugging Face 😊

Select the Space SDK

You can choose between Streamlit, Gradio and Static for your Space. Or pick Docker to host any other app.



**Streamlit**



**Gradio**  
3 templates



**Docker**  
13 templates



**Static**  
3 templates

Choose a Docker template:



**Blank**



**JupyterLab**



**Argilla**



**Livebook**



**LabelStudio**



**AimStack**



**AutoTrain**



**Shiny (R)**



**Shiny (Python)**



**ZenML**



**ChatUI**



**Panel**



**Giskard**



**Quarto**

**Open and Free To Use**

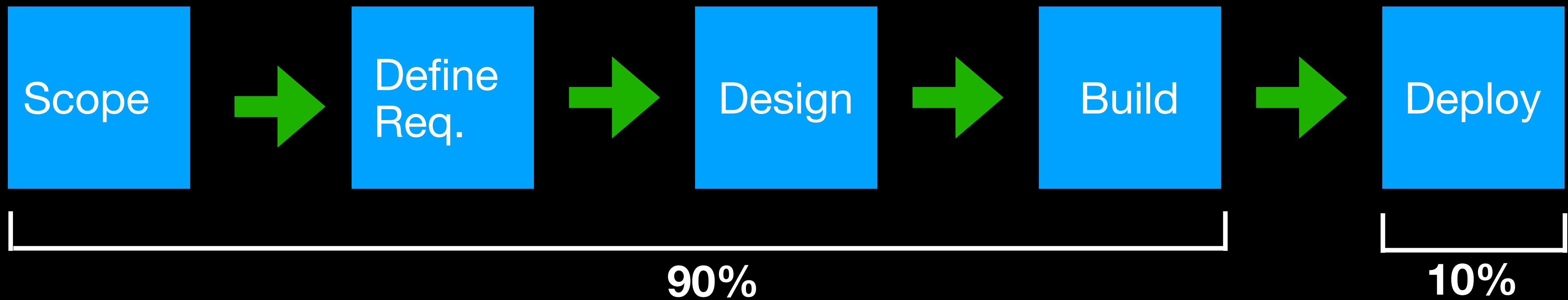
# Limitations

# Limitations

- Not secure
- Compute power
- Availability
- Data storage

# Deploy and Monitor ML and Data Pipelines with GitHub Actions

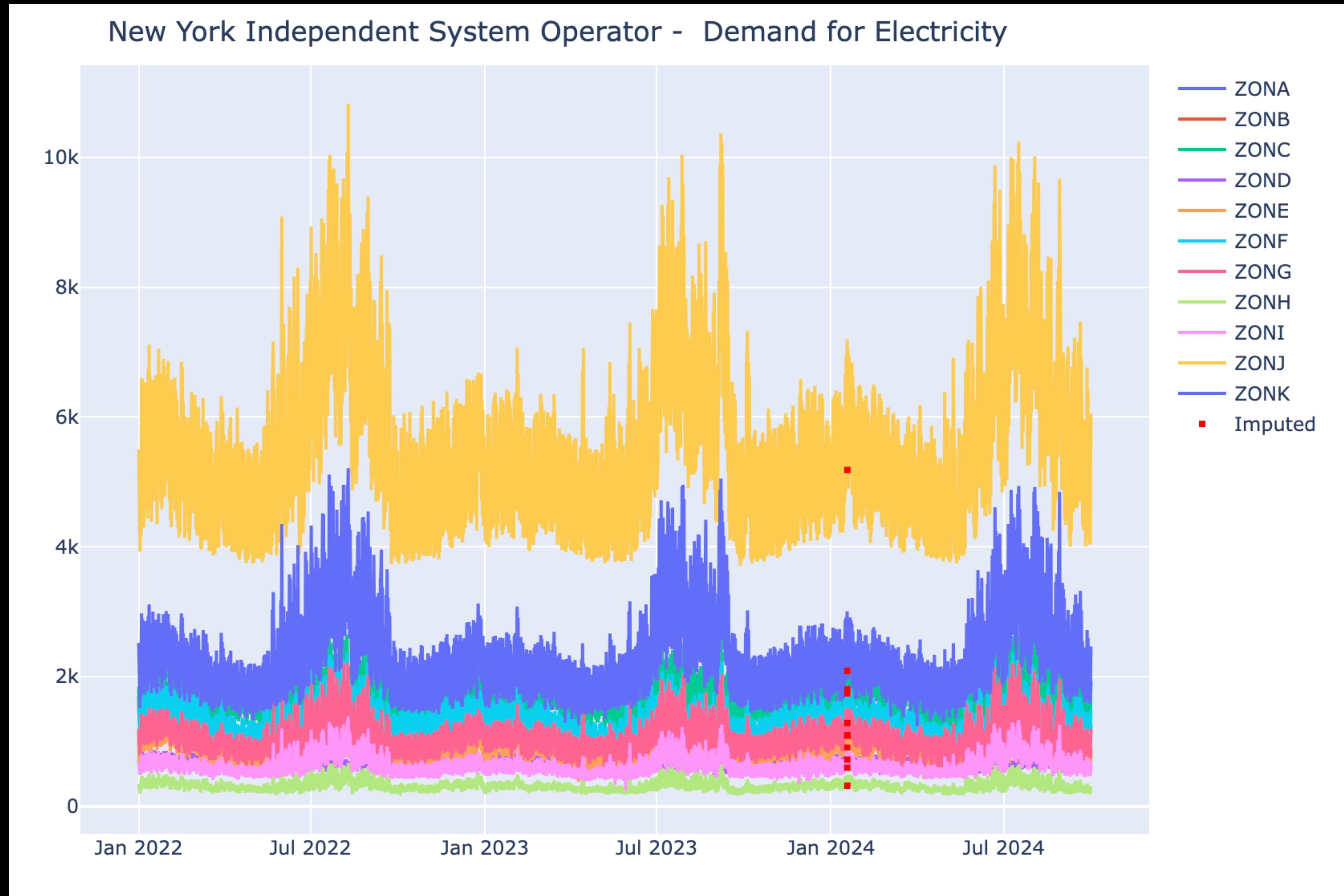
# Building Pipeline - Workflow



# Scope

- Forecast the hourly demand for electricity in New York by sub-region
- This includes the following 11 providers:
- Refresh the data daily
- Update the forecast every 24 hours

# Scope

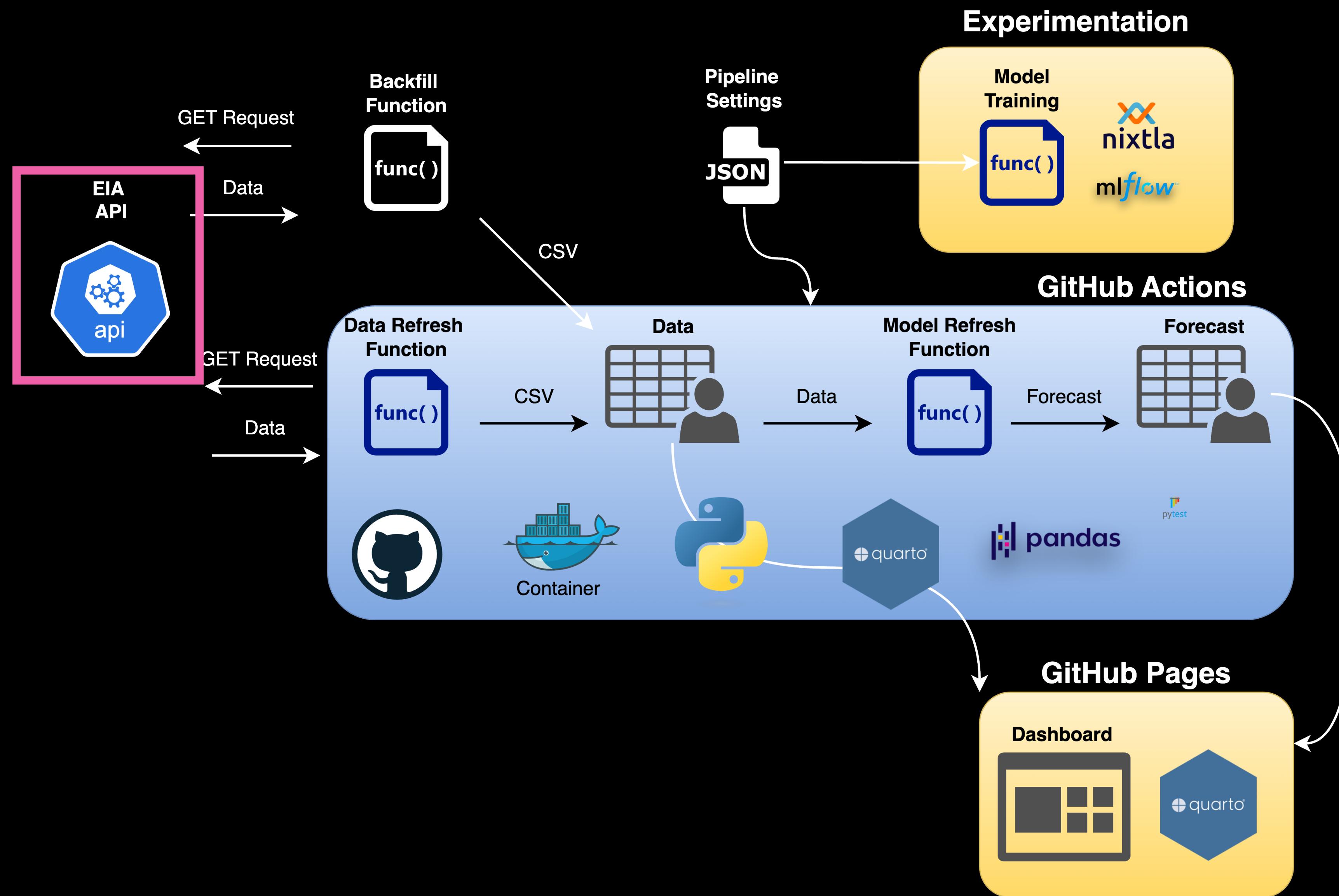


# Requirements

- Environment - Docker image
- Orchestration - GitHub Actions
- Data pipeline - keep the data up-to-date
- ML pipeline - refresh the forecast
- Dashboard - GitHub Pages

# Design

# Pipeline Design



# Data Source - EIA API

The screenshot shows the official website of the U.S. Energy Information Administration (EIA). The header features the EIA logo and the text "Independent Statistics and Analysis U.S. Energy Information Administration". Navigation links include "+ Tools", "+ Learn About Energy", "+ News", "+ Sources & Uses", "+ Topics", and "+ Geography". A search bar says "Search eia.gov" and there are social media icons for X and Facebook.

A large banner image shows a road leading through a landscape with power transmission towers. Overlaid on the banner is a dark box containing the text "Short-Term Energy Outlook June 2024" and "Energy projections for supply, demand, and prices →". Below the banner, a horizontal navigation bar includes "What's New", "Today in Energy", and "Data Highlights".

**What's New**

- Refinery Capacity Report  
Jun 14, 2024
- Wholesale Electricity Market Data  
Jun 13, 2024
- Short-Term Energy Outlook  
Jun 11, 2024
- [More >](#)

**Today in Energy** Posted June 18, 2024

**U.S. refinery capacity increased 2% in 2023** [>](#)

Operable atmospheric crude oil distillation capacity, our primary measure of refinery capacity in the United States, increased by 2%, or 324,000 barrels per day (b/d), in 2023, according to our recently published Refinery Capacity Report. [More >](#)

**U.S. refinery atmospheric distillation capacity on Jan 1 (2017–2024)**  
million barrels per day

Year	Capacity (million barrels per day)
2017	18.0
2018	18.0
2019	18.2
2020	18.2
2021	17.8
2022	17.8
2023	17.8
2024	18.0

idle operating

**Data Highlights**

- WTI crude oil futures price**  
6/17/2024: NA/barrel  
NA from week earlier  
NA from year earlier
- Natural gas futures price**  
6/17/2024: NA/MMBtu  
NA from week earlier  
NA from year earlier
- Retail gasoline price**  
6/17/2024: \$3.435/gal  
▲ \$0.006 from week earlier  
▼ \$0.142 from year earlier
- Retail diesel price**  
6/17/2024: \$3.735/gal  
▲ \$0.077 from week earlier  
▼ \$0.080 from year earlier
- Weekly coal production**

# Data Source - EIA API

The screenshot shows a search interface for the EIA API. At the top, there is a navigation bar with the EIA logo and links for Sources & Uses, Topics, and Geography. A search bar is also present. Below the navigation bar, there are two main sections: 'API URL:' and 'HEADER:'.

**API URL:** `https://api.eia.gov/v2/electricity/rto/region-sub-ba-data/data/?frequency=hourly&data[0]=value&facets[parent][]=CISO&facets[subba][]=PGAE&facets[subba][]=SCE&facets[subba][]=SDGE&facets[subba][]=VEA&sort[0][column]=period&sort[0][direction]=desc&offset=0&length=5000`

**METHOD:** GET

**SERIES DESCRIPTION:** Hourly demand by balancing authority subregion. Source: Form EIA-930  
Product: Hourly Electric Grid Monitor

**HEADER:**

```
X-Params: {  
    "frequency": "hourly",  
    "data": [  
        "value"  
    ],  
    "facets": [  
        "parent": [  
            "CISO"  
        ],  
        "subba": [  
            "PGAE",  
            "SCE",  
            "SDGE",  
            "VEA"  
        ]  
    ],  
    "start": null,  
    "end": null,  
    "sort": [  
        {  
            "column": "period",  
            "direction": "desc"  
        }  
    ],  
    "offset": 0,  
    "length": 5000  
}
```

# Pipeline Design

EIA  
API



```
FROM python:3.10-slim AS builder

ARG QUARTO_VER="1.5.56"
ARG VENV_NAME="my_project"
ENV QUARTO_VER=$QUARTO_VER
ENV VENV_NAME=$VENV_NAME
RUN mkdir requirements

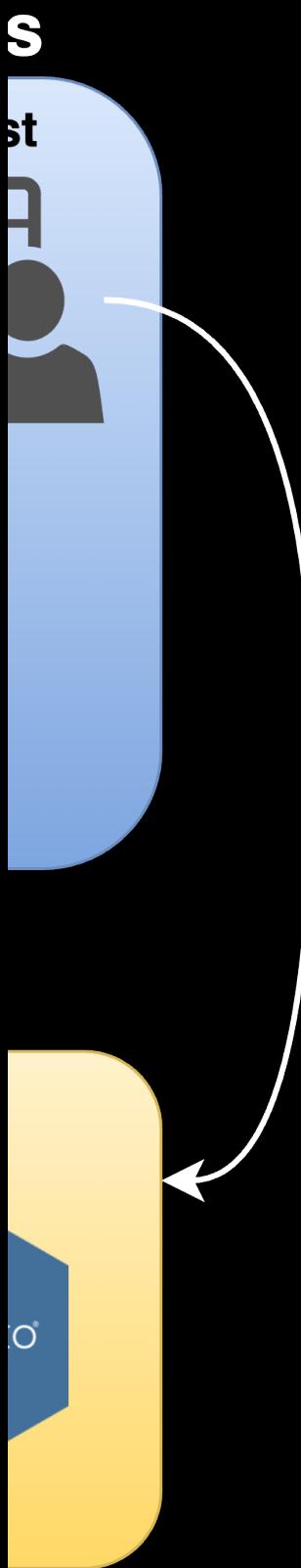
COPY install_requirements.sh requirements/
COPY requirements.txt requirements/
RUN bash ./requirements/install_requirements.sh $VENV_NAME

FROM python:3.10-slim

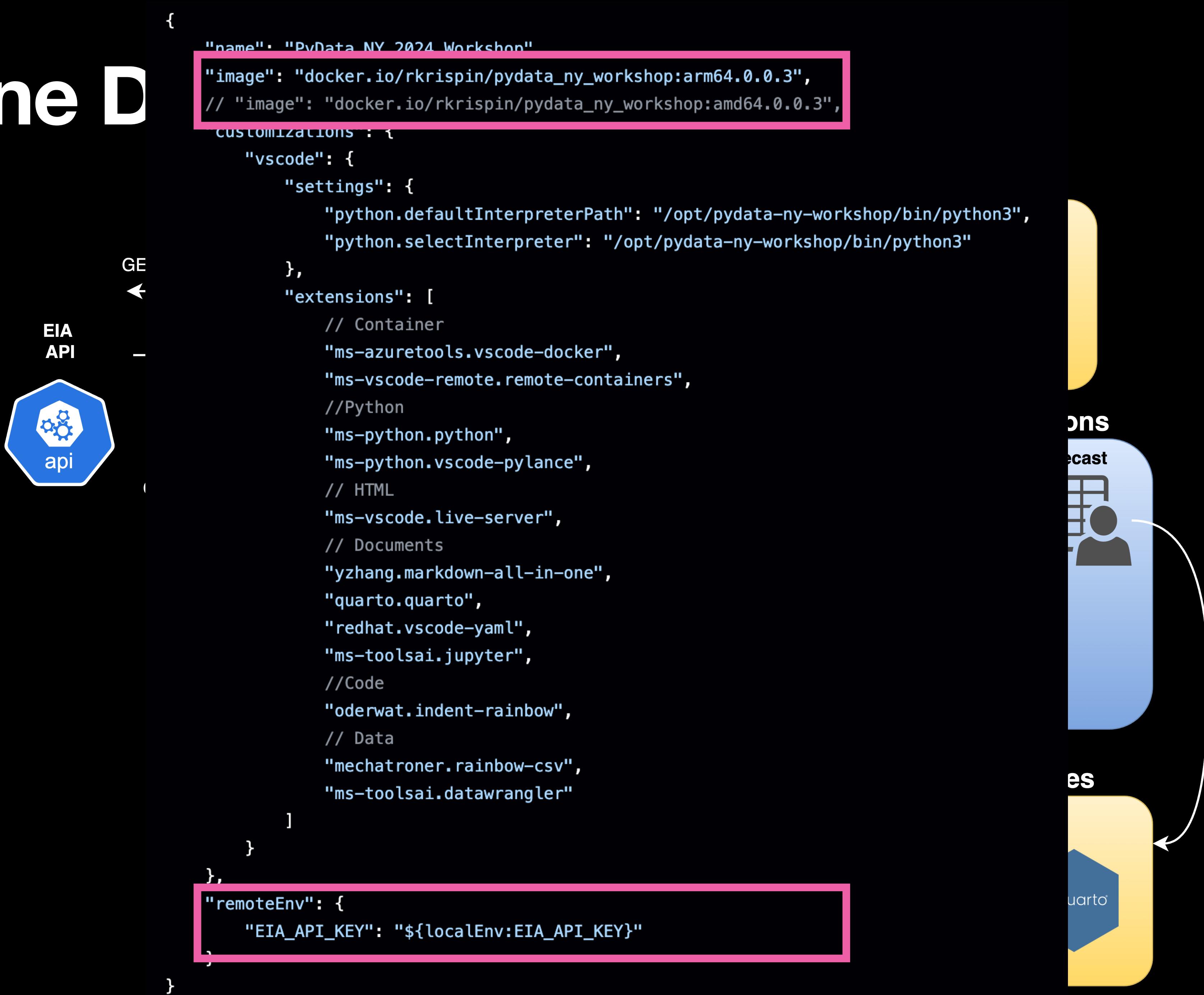
ARG QUARTO_VER="1.5.56"
ARG VENV_NAME="my_project"
ENV QUARTO_VER=$QUARTO_VER
ENV VENV_NAME=$VENV_NAME

COPY --from=builder /opt/$VENV_NAME /opt/$VENV_NAME
COPY install_requirements.sh install_quarto.sh requirements.sh requirements/
RUN bash ./requirements/install_dependencies.sh $QUARTO_VER
RUN bash ./requirements/install_quarto.sh $QUARTO_VER

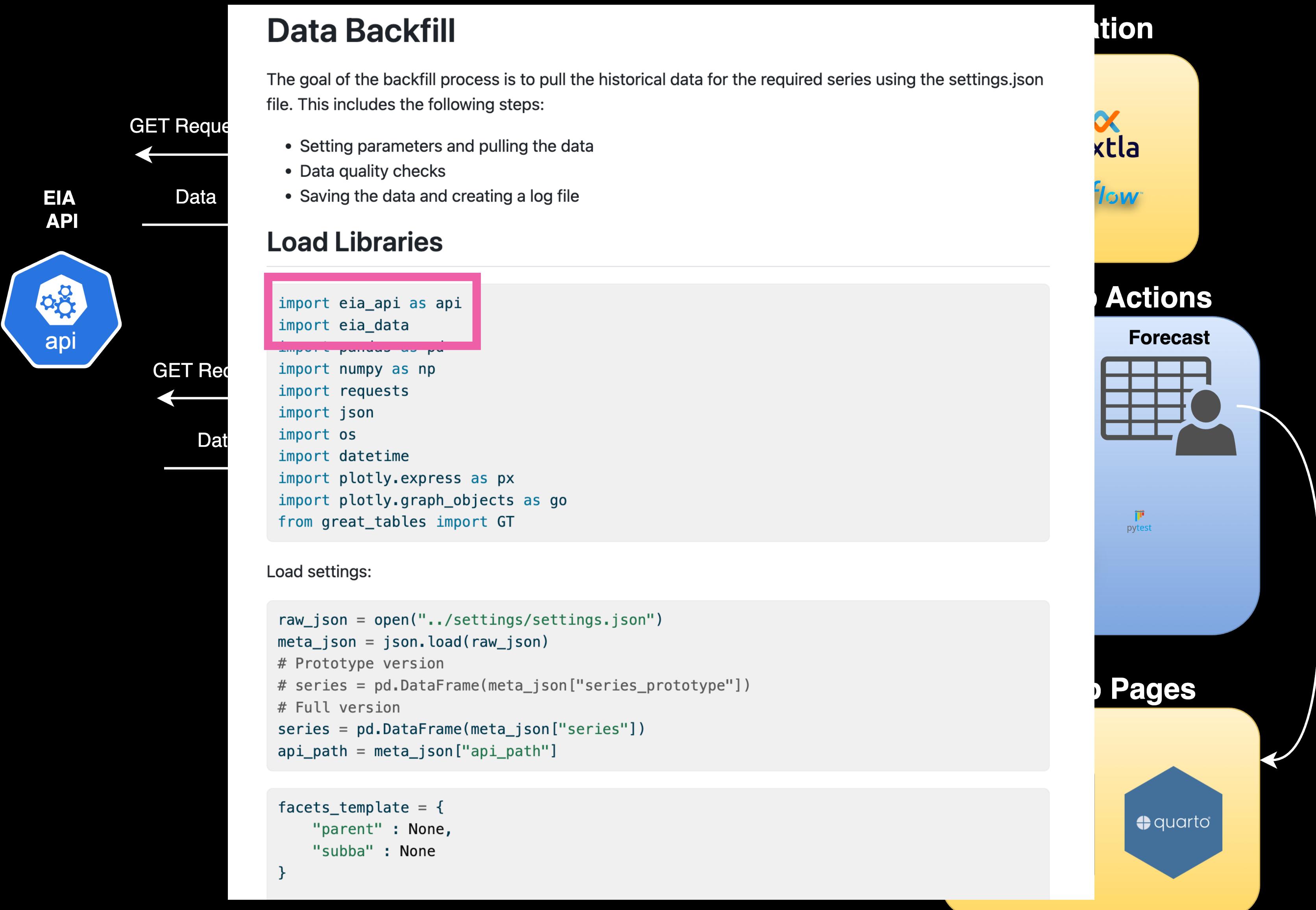
RUN echo "source /opt/$VENV_NAME/bin/activate" >> ~/.bashrc
```



# Pipeline D



# Pipeline Design

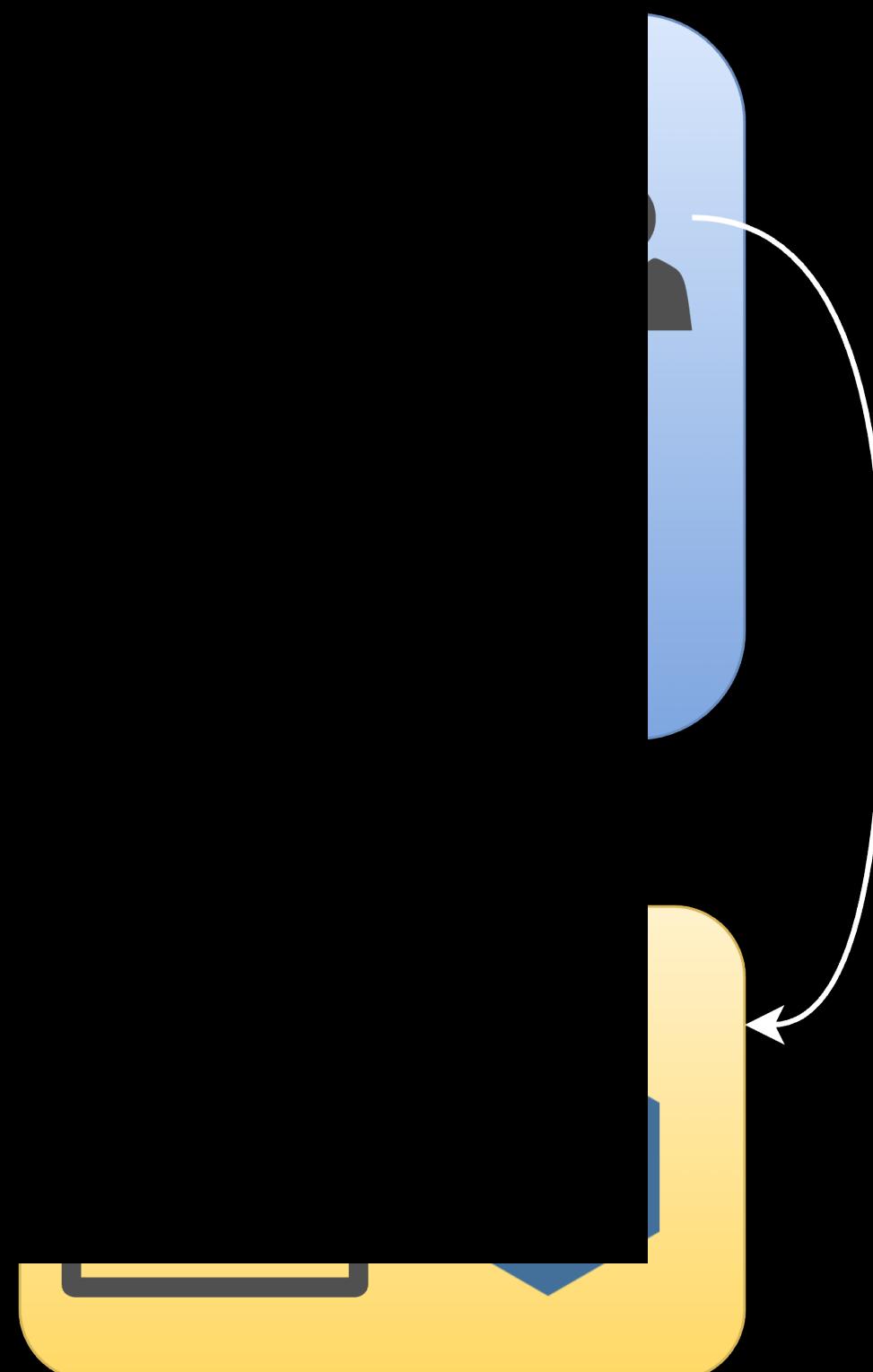


# Pipeline Design

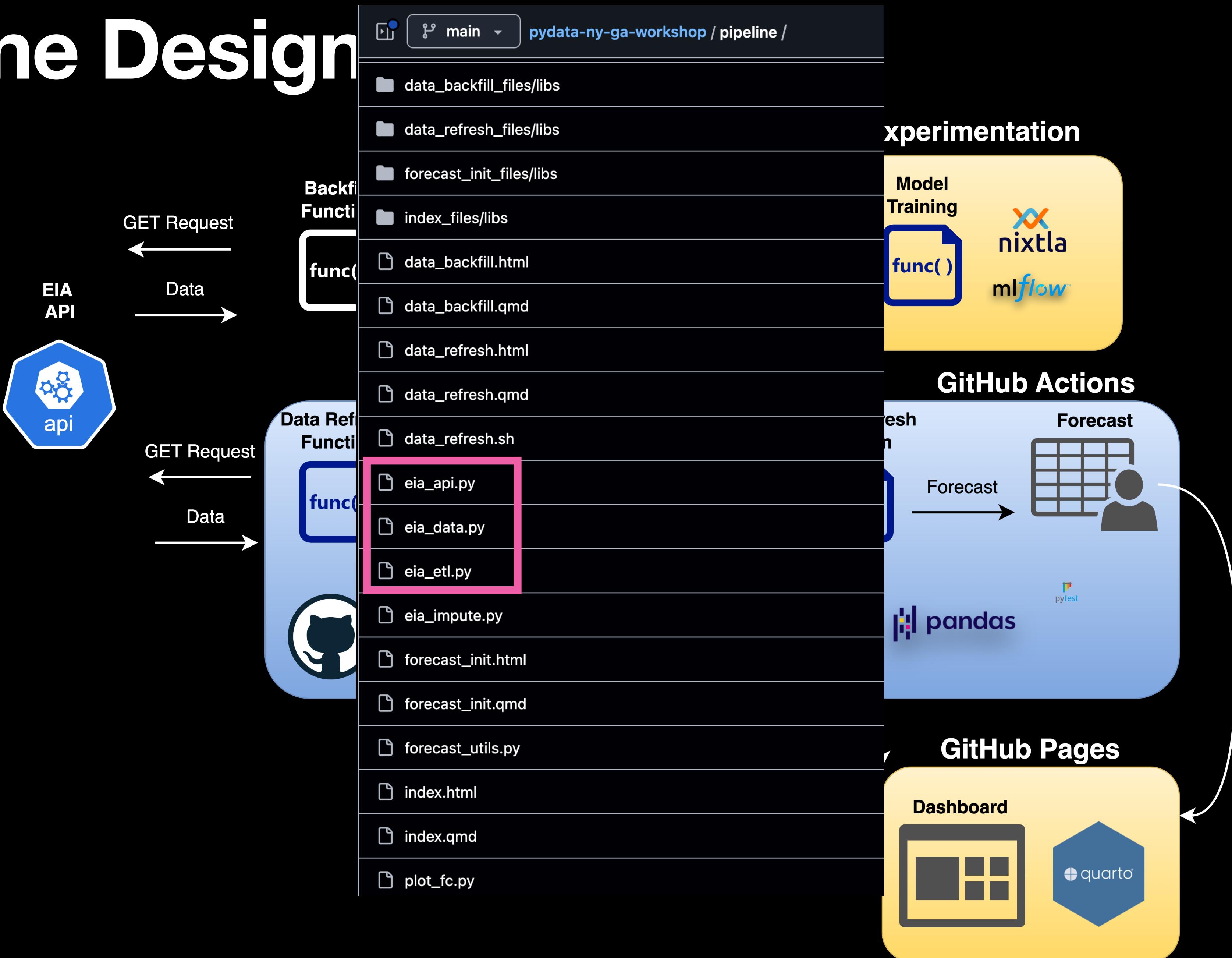
## Experimentation



```
1 period,subba,impute,y,subba-name,parent,parent-name,value,value-units
2 2022-01-01 00:00:00,ZONA,,1707.0,....
3 2022-01-01 01:00:00,ZONA,,1673.0,....
4 2022-01-01 02:00:00,ZONA,,1644.0,....
5 2022-01-01 03:00:00,ZONA,,1605.0,....
6 2022-01-01 04:00:00,ZONA,,1550.0,....
7 2022-01-01 05:00:00,ZONA,,1487.0,....
8 2022-01-01 06:00:00,ZONA,,1422.0,....
9 2022-01-01 07:00:00,ZONA,,1373.0,....
10 2022-01-01 08:00:00,ZONA,,1336.0,....
11 2022-01-01 09:00:00,ZONA,,1317.0,....
12 /data/data.csv 01-01 10:00:00,ZONA,,1307.0,....
13 2022-01-01 11:00:00,ZONA,,1315.0,....
14 2022-01-01 12:00:00,ZONA,,1343.0,....
15 2022-01-01 13:00:00,ZONA,,1373.0,....
16 2022-01-01 14:00:00,ZONA,,1395.0,....
17 2022-01-01 15:00:00,ZONA,,1436.0,....
18 2022-01-01 16:00:00,ZONA,,1488.0,....
19 2022-01-01 17:00:00,ZONA,,1526.0,....
20 2022-01-01 18:00:00,ZONA,,1553.0,....
21 2022-01-01 19:00:00,ZONA,,1565.0,....
22 2022-01-01 20:00:00,ZONA,,1574.0,....
23 2022-01-01 21:00:00,ZONA,,1592.0,....
```



# Pipeline Design



# Pipeline Design

## Experimentation

parent	subba	time		start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00		2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found in the data.
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00		2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were found in the data.
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00		2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were found in the data.
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00		2024-11-04 03:00:00	2024-11-04 02:00:00							0	refresh	False	True	No new data is available.



# Pipeline Design

## Experimentation

parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were found.
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were found.
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00	2024-11-04 03:00:00	2024-11-04 02:00:00							0	refresh	False	True	No new data is available.



# Pipeline Design

## Experimentation

parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were found.
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were found.
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00	2024-11-04 03:00:00	2024-11-04 02:00:00						0	refresh	False	True	No new data is available.	



# Pipeline Design

## Experimentation

parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00	2024-11-04 03:00:00	2024-11-04 02:00:00							0	refresh	False	True	No new data is available



# Pipeline Design

## Experimentation

parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values were found.
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were found.
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values were found.
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00	2024-11-04 03:00:00	2024-11-04 02:00:00							0	refresh	False	True	No new data is available.



# Pipeline D



EIA  
API

## Data Refresh

### Load Libraries

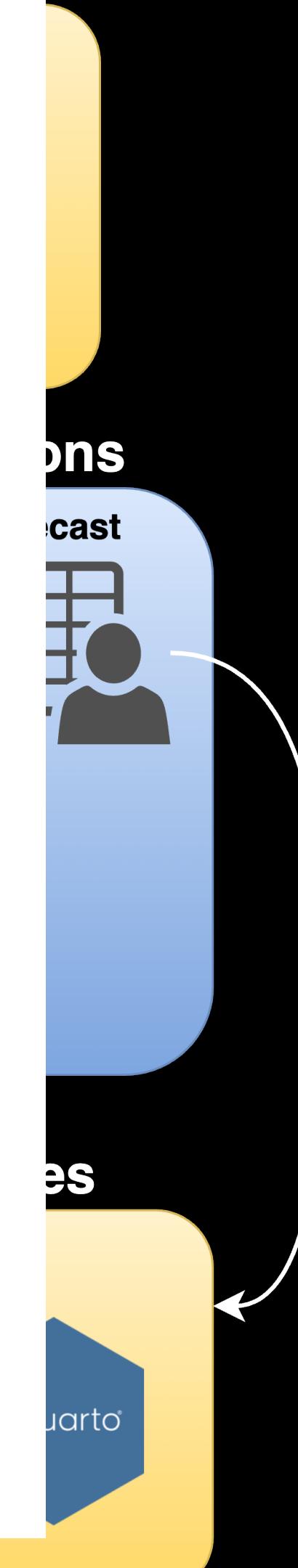
```
import eia_api as api
import eia_data
import eia_impute as impute
import pandas as pd
import numpy as np
import requests
import json
import os
import datetime
import plotly.express as px
import plotly.graph_objects as go
from great_tables import GT
```

### Settings

Load settings:

```
raw_json = open("../settings/settings.json")
meta_json = json.load(raw_json)
# Prototype version
# series = pd.DataFrame(meta_json["series_prototype"])
# Full version
series = pd.DataFrame(meta_json["series"])
api_path = meta_json["api_path"]
meta_path = meta_json["meta_path"]
data_path = meta_json["data_path"]
# Should be setting as false during development
save_data = True
save_meta = True

eia_api_key = os.getenv('EIA_API_KEY')
forecast_settings = meta_json["forecast"]
leaderboard_path = meta_json["backtesting"]["leaderboard_path"]
```



# Pipeline D



EIA  
API

GET



## Experimentation

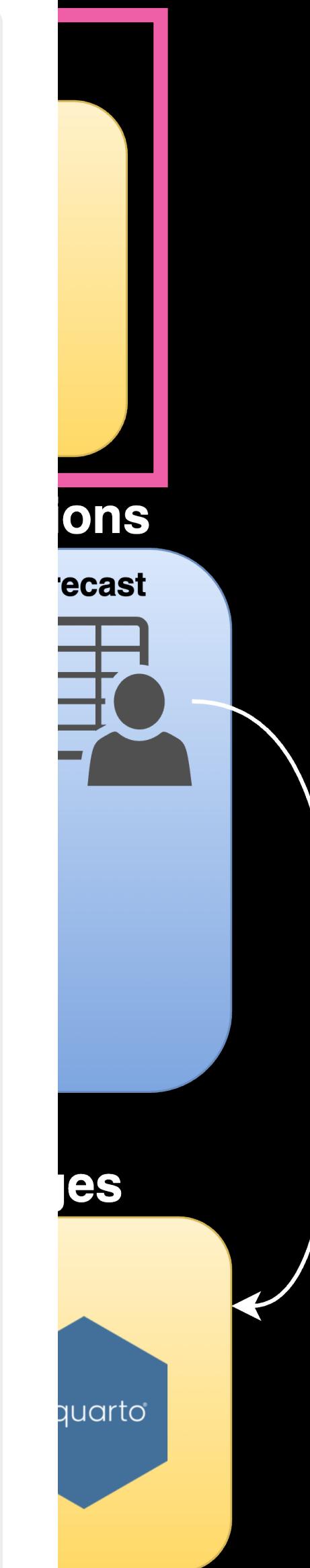
### Loading Required Libraries

```
import pandas as pd
import numpy as np
import requests
import json
import os
import mlflow
import datetime
import plotly.graph_objects as go
from great_tables import GT

from statsforecast import StatsForecast
from statsforecast.models import (
    HoltWinters,
    CrostonClassic as Croston,
    HistoricAverage,
    DynamicOptimizedTheta,
    SeasonalNaive,
    AutoARIMA,
    AutoETS,
    AutoTBATS,
    MSTL
)

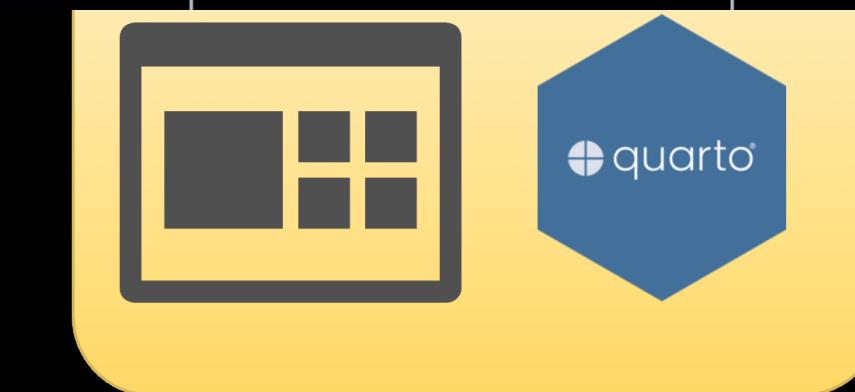
from mlforecast import MLForecast
from mlforecast.target_transforms import Differences
from mlforecast.utils import PredictionIntervals
from window_ops.expanding import expanding_mean
from lightgbm import LGBMRegressor
from xgboost import XGBRegressor
from sklearn.linear_model import LinearRegression
from utilsforecast.plotting import plot_series
from statistics import mean

import backtesting
```

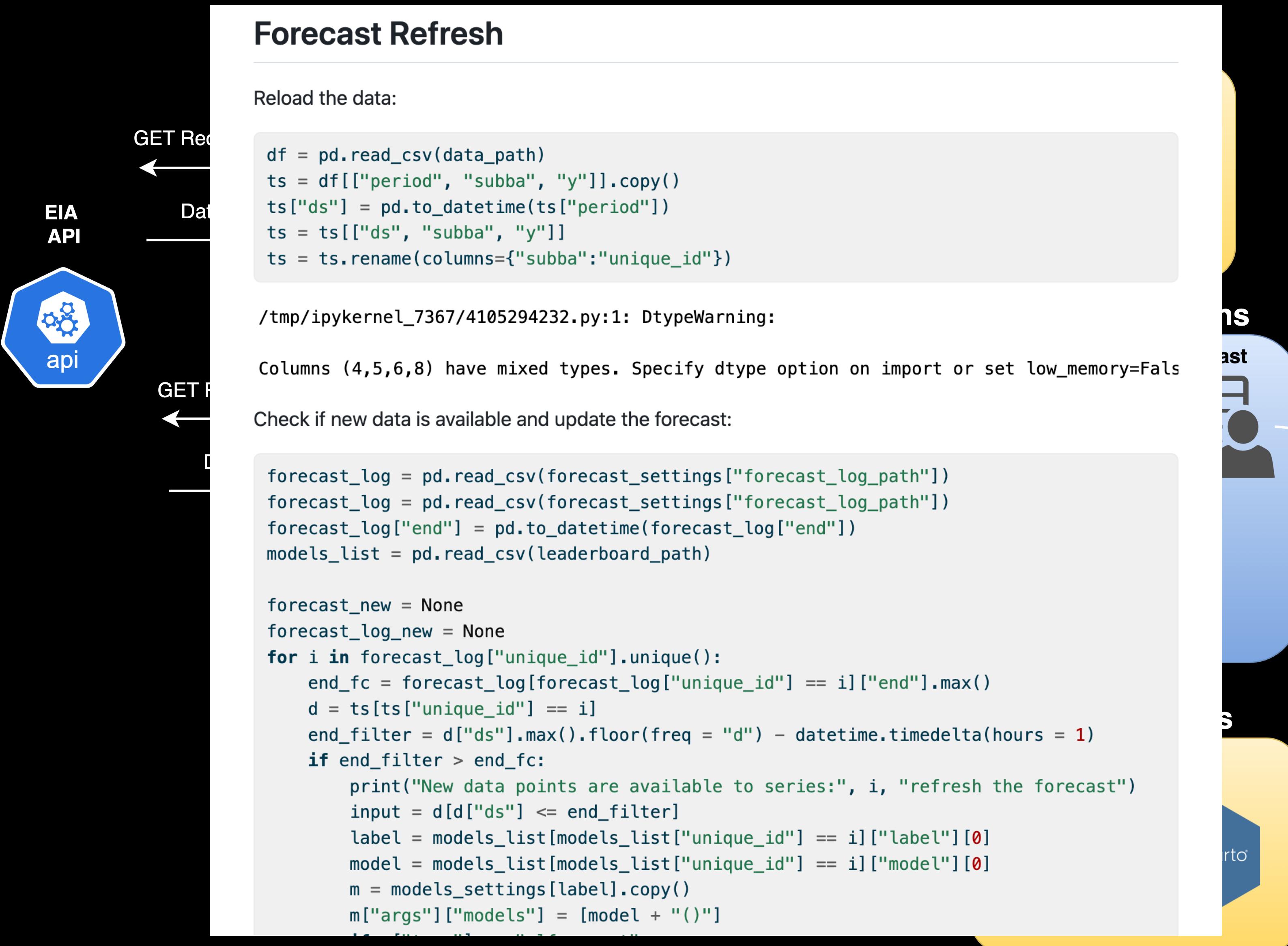


# Pipeline Design

GET Request		Backfill Function		Pipeline Settings		Experimentation		
		unique_id	label	model	type	partitions	avg_mape	avg_rmse
							Model Training	XX
model_unique_id		unique_id	label	model	type	partitions	avg_mape	avg_rmse
model4_LinearRegression	ZONA	model4	LinearRegression	mlforecast	20		0.031451369703253865	63.70007892927969
model4_XGBRegressor	ZONB	model4	XGBRegressor	mlforecast	20		0.04990164204689953	61.50748573552928
model4_LinearRegression	ZONC	model4	LinearRegression	mlforecast	20		0.05883017394892269	109.75741267896122
model4_LGBMRegressor	ZOND	model4	LGBMRegressor	mlforecast	20		0.04610411734303354	32.59497249269365
model4_LinearRegression	ZONE	model4	LinearRegression	mlforecast	20		0.0745053032568893	67.76425655956902
model4_LinearRegression	ZONF	model4	LinearRegression	mlforecast	20		0.05212821927381432	75.84396938968229
model4_LinearRegression	ZONG	model4	LinearRegression	mlforecast	20		0.048855291350719504	55.6707171450734
model4_XGBRegressor	ZONH	model4	XGBRegressor	mlforecast	20		0.05390935709492725	16.710052356332614
model4_XGBRegressor	ZONI	model4	XGBRegressor	mlforecast	20		0.03842874906638785	27.07394224935995
model4_LGBMRegressor	ZONJ	model4	LGBMRegressor	mlforecast	20		0.025910214240455358	163.31549830977897
model4_LinearRegression	ZONK	model4	LinearRegression	mlforecast	20		0.040560537435138336	95.77579962905139



# Pipeline Design



# Pipeline Design

## Experimentation

unique_id	ds	forecast	lower	upper	model	label	forecast_label
ZONA	2024-11-02 00:00:00	1800.517800565871	1773.3704623536762	1827.665138778066	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 01:00:00	1780.5180835427157	1755.1184614882377	1805.9177055971936	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 02:00:00	1730.5068915232632	1693.8877826409214	1767.126000405605	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 03:00:00	1654.7358302549685	1603.9235353809265	1705.5481251290105	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 04:00:00	1584.9708637502977	1525.4609120298428	1644.4808154707525	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 05:00:00	1525.9659106867446	1456.1828503187191	1595.74897105477	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 06:00:00	1476.4784068292424	1405.0116680784008	1547.945145580084	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 07:00:00	1447.0144869963324	1381.9366979035979	1512.092276089067	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 08:00:00	1440.470086812012	1376.3634214344702	1504.5767521895536	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 09:00:00	1444.5006873888851	1377.8258236132153	1511.175551164555	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 10:00:00	1477.0962733051292	1410.6160902619663	1543.5764563482921	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 11:00:00	1538.3728728922558	1455.8060075247724	1620.9397382597392	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 12:00:00	1600.5665991741648	1475.2608582003436	1725.872340147986	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 13:00:00	1648.9442846152804	1520.1779561471667	1777.710613083394	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 14:00:00	1688.4020113226267	1572.7049079504051	1804.000124916849	LinearRegression	model4	2024-11-01

# Pipeline Design

unique_id	model	label	forecast_label	start	end	n_obs	h	refresh_time	success	mape	rmse	coverage	score
ZONA	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:09	True	0.04365224383524807	95.44281545897655	0.708333333333334	True
ZONB	XGBRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:10	True	0.04897028797549269	55.111854724106706	0.6666666666666666	True
ZONC	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:10	True	0.02942917761312673	54.89107179763394	0.958333333333334	True
ZOND	LGBMRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:11	True	0.022954528805170193	18.555757547214046	1.0	True
ZONE	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:12	True	0.022176471731920985	21.898557451386342	0.958333333333334	True
ZONF	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:12	True	0.026830806664641407	40.537501009876465	1.0	True
ZONG	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:13	True	0.1035057362451028	105.17885990069209	0.75	True
ZONH	XGBRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:14	True	0.1064112030516249	25.454014229573012	0.458333333333333	True
ZONI	XGBRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:15	True	0.05056286468047634	33.89463797766538	0.833333333333334	True
ZONJ	LGBMRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:16	True	0.06566055892878288	314.97418947984215	0.083333333333333	True
ZONK	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:16	True	0.13673648504648134	270.2176330551692	0.333333333333333	True
ZONA	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:17	True	0.04069130036037995	71.60527947080918	0.708333333333334	True
ZONB	XGBRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:18	True	0.06327243341515874	60.72460597309195	0.75	True
ZONC	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:18	True	0.10937874297241784	180.95626539971647	0.708333333333334	True
ZOND	LGBMRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:19	True	0.030088697473674635	23.02897283243766	0.9166666666666666	True
ZONE	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:19	True	0.12608821062538283	97.36083704095965	0.75	True
ZONF	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:20	True	0.07354674230651301	93.86685356486336	0.7916666666666666	True
ZONG	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:20	True	0.07126757093240858	65.10305390846375	0.708333333333334	True
ZONH	XGBRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:21	True	0.07273225940792014	20.817034743164285	0.5	True
ZONI	XGBRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:23	True	0.02257437803721225	13.925164828678495	1.0	True
ZONJ	LGBMRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:23	True	0.015217619426924294	78.30077282677799	1.0	True

# New York Independent System Operator Hourly Demand PyData NYC 2024

Hourly Demand By Provider

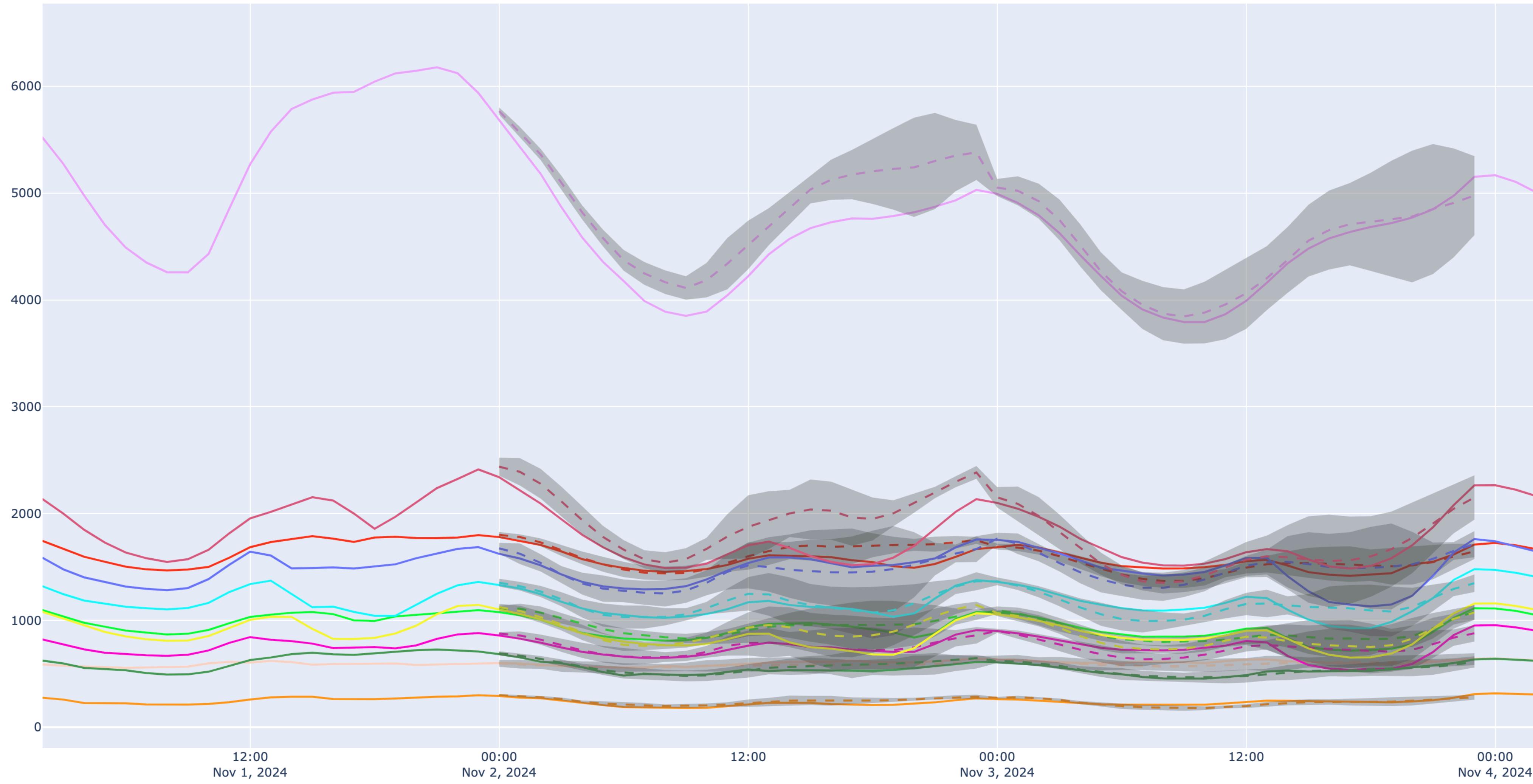
Metadata

Forecast Leaderboard

7d 1m 6m YTD 1y all



- ZONA Forecast
  - Actual
  - Forecast
- ZONB Forecast
  - Actual
  - Forecast
- ZONC Forecast
  - Actual
  - Forecast
- ZOND Forecast
  - Actual
  - Forecast
- ZONE Forecast
  - Actual
  - Forecast
- ZONF Forecast
  - Actual
  - Forecast
- ZONG Forecast
  - Actual
  - Forecast
- ZONH Forecast
  - Actual
  - Forecast
- ZONI Forecast
  - Actual
  - Forecast
- ZONJ Forecast
  - Actual





RamiKrispin / pydata-ny-ga-workshop

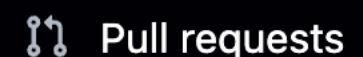
Type / to search



Code



Issues 6



Pull requests



Discussions



Actions



Projects 1



Wiki



Security



Insights



Settings

## Actions

[New workflow](#)

All workflows

### Data Refresh

pages-build-deployment

Management

Caches

Deployments



Attestations



Runners

## Data Refresh

pipeline\_refresh.yml

[Filter workflow runs](#)

### 61 workflow runs

Event ▾ Status ▾ Branch ▾ Actor ▾

#### ✓ Data Refresh

Data Refresh #61: Scheduled

main

39 minutes ago

1m 55s



#### ✓ Data Refresh

Data Refresh #60: Scheduled

main

1 hour ago

1m 55s



#### ✓ Data Refresh

Data Refresh #59: Scheduled

main

2 hours ago

1m 53s



#### ✓ Data Refresh

Data Refresh #58: Scheduled

main

3 hours ago

1m 51s



#### ✓ Data Refresh

Data Refresh #57: Scheduled

main

4 hours ago

1m 47s



#### ✓ Data Refresh

Data Refresh #56: Scheduled

main

5 hours ago

1m 52s



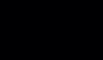
#### ✓ Data Refresh

Data Refresh #55: Scheduled

main

6 hours ago

1m 49s



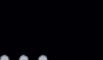
#### ✓ Data Refresh

Data Refresh #54: Scheduled

main

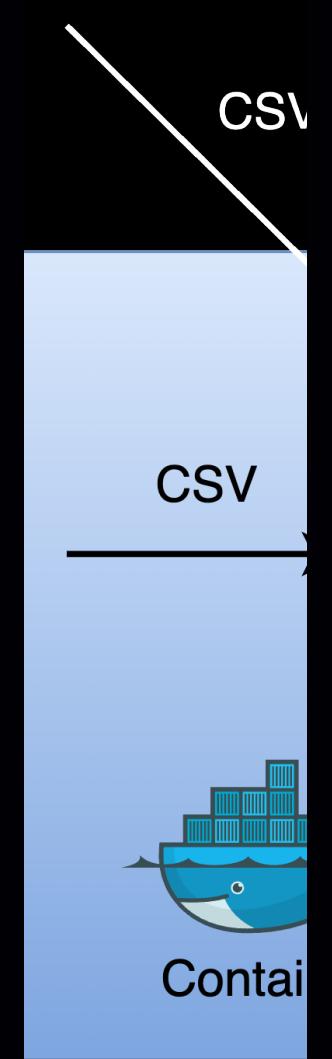
7 hours ago

2m 5s



# Pipeline Design

```
"series": [
  {
    "parent_id": "NYIS",
    "parent_name": "New York Independent System Operator",
    "subba_id": "ZONA",
    "subba_name": "West - NYIS"
  },
  {
    "parent_id": "NYIS",
    "parent_name": "New York Independent System Operator",
    "subba_id": "ZONB",
    "subba_name": "Genesee - NYIS"
  },
  {
    "parent_id": "NYIS",
    "parent_name": "New York Independent System Operator",
    "subba_id": "ZONC",
    "subba_name": "Central - NYIS"
  },
  {
    "parent_id": "NYIS",
    "parent_name": "New York Independent System Operator",
    "subba_id": "ZOND",
    "subba_name": "North - NYIS"
  },
  {
    "parent_id": "NYIS",
    "parent_name": "New York Independent System Operator",
    "subba_id": "ZONE",
    "subba_name": "Mohawk Valley - NYIS"
  },
  {
    "parent_id": "NYIS",
    "parent_name": "New York Independent System Operator",
    "subba_id": "ZONF",
    "subba_name": "Capital - NYIS"
  }
],
```



```
  "model3": {
    "label": "model3",
    "type": "mlforecast",
    "args": {
      "models": [
        "LGBMRegressor()", "XGBRegressor()", "LinearRegression()"
      ],
      "lags": [
        1,
        24,
        48
      ],
      "date_features": [
        "day",
        "dayofweek",
        "hour"
      ],
      "n_windows": 5,
      "freq": "h",
      "comments": "ML models with seasonal features and short 3 lags (1, 24, 48)"
    }
  },
```



# Deployment

# GitHub Actions Requirements

- Workflow file
- Environment
- Script to execute
- Trigger

## name: Data Refresh

### Data Refresh

pipeline\_refresh.yml

Filter workflow runs



63 workflow runs

Event ▾ Status ▾ Branch ▾ Actor ▾

#### 🟡 Data Refresh

Data Refresh #63: Scheduled

main

1 minute ago

In progress



#### 🟢 Data Refresh

Data Refresh #62: Scheduled

main

1 hour ago

2m 4s



#### 🟢 Data Refresh

Data Refresh #61: Scheduled

main

1 hour ago

1m 55s



#### 🟢 Data Refresh

Data Refresh #60: Scheduled

main

2 hours ago

1m 55s



#### 🟢 Data Refresh

Data Refresh #59: Scheduled

main

4 hours ago

1m 53s



#### 🟢 Data Refresh

Data Refresh #58: Scheduled

main

5 hours ago

1m 51s



#### 🟢 Data Refresh

Data Refresh #57: Scheduled

main

6 hours ago

1m 47s



#### 🟢 Data Refresh

Data Refresh #56: Scheduled

main

7 hours ago

1m 52s



USER\_NAME: \${{ secrets.USER\_NAME }}

```
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"

jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
    env:
      EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
      USER_EMAIL: ${{ secrets.USER_EMAIL }}
      USER_NAME: ${{ secrets.USER_NAME }}
```

```
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

```
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

```
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
    env:
      EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
      USER_EMAIL: ${{ secrets.USER_EMAIL }}
      USER_NAME: ${{ secrets.USER_NAME }}
```

```
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
    env:
      EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
      USER_EMAIL: ${{ secrets.USER_EMAIL }}
      USER_NAME: ${{ secrets.USER_NAME }}
```

# Unit Tests!!!

What else can we do?

Many other cool things!

# Use Cases

- Alerts
- Social
- Data visualization
- Automation

# Best Practices

- Set prototype version
- Know the tools limitations
- Using settings files and environment variables
- Unit tests
- Interactive documents
- GitHub Templates!

# Summary

- GitHub Actions enables to schedule workflows
- Docker provides a reproducible environment
- Great use cases - open source projects
- Quarto documents to host the pipeline
- Leverage great open source libraries (MLflow, Y-Data Profile, etc.)
- Not enterprise ready

# Questions?

# Get In Touch

