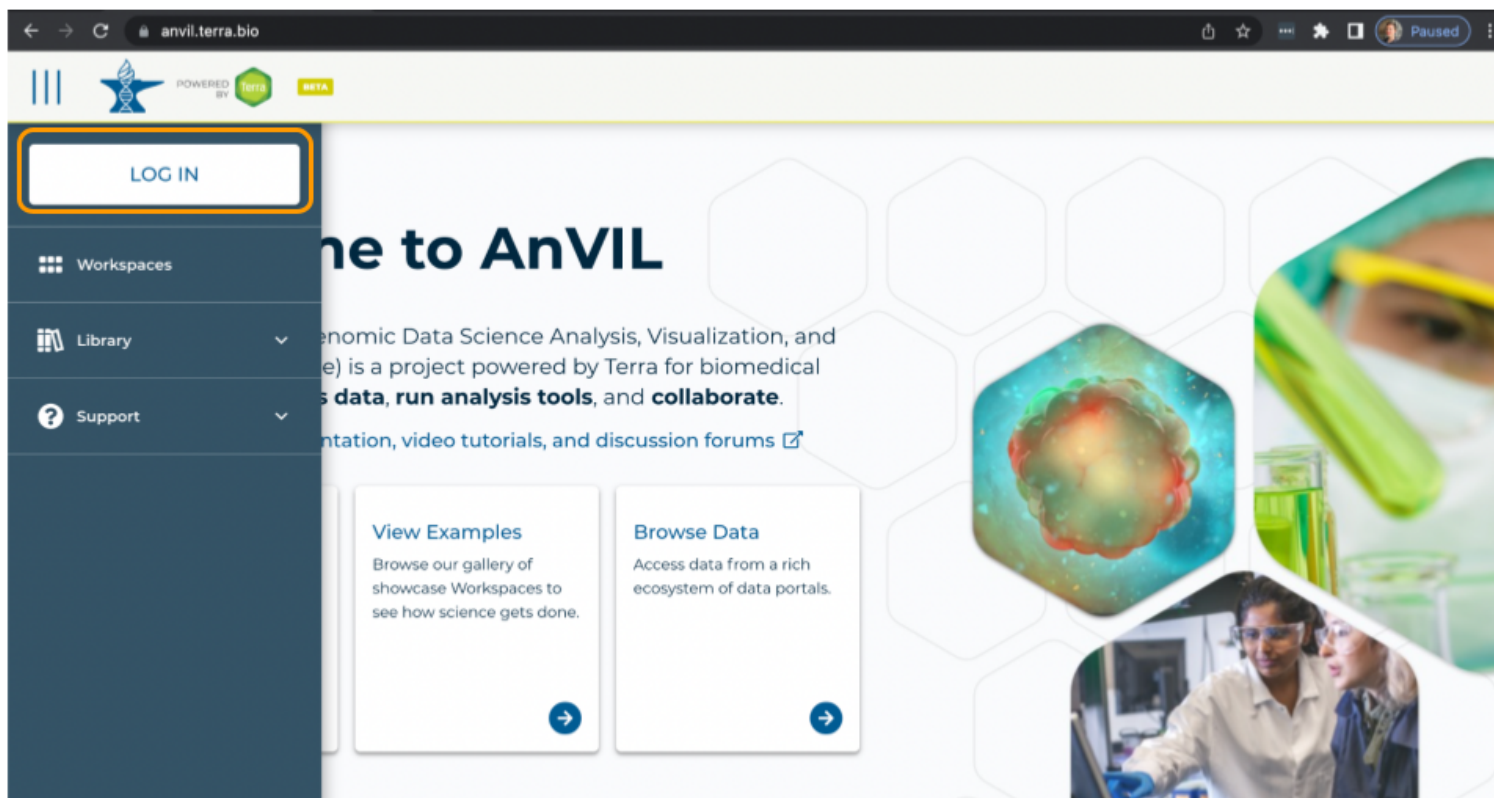


AnVIL Setup

Data Wrangling in R

Setup on AnVIL

1. You need to sign into Terra with your Google account. This is the only way you can launch applications and perform computations on AnVIL. Launch AnVIL at <https://anvil.terra.bio/>, and you should be prompted to sign in with your Google account.



Setup on AnVIL

⚠ Make sure you provide your Google login information to the instructor! ⚠

Setup on AnVIL

Go to the Class Workspace at <https://anvil.terra.bio/#workspaces/data-wrangling-workshop/SISBID-data-wrangling-2022>

The screenshot shows the AnVIL workspace interface. At the top, there's a header with the Terra logo, 'POWERED BY Terra', 'BETA WORKSPACES', and the workspace name 'data-wrangling-workshop/SISBID-data-wrangling-2022'. Below the header is a navigation bar with 'DASHBOARD', 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. The main content area is titled 'ABOUT THE WORKSPACE' and includes a welcome message and instructions on how to clone the workspace and launch an RStudio instance. On the right, there's a sidebar with 'WORKSPACE INFORMATION', 'CLOUD INFORMATION', 'OWNERS', and 'TAGS'.

ABOUT THE WORKSPACE

Welcome to Data Wrangling!

Please check out the workshop website at <https://sisbid.github.io/Data-Wrangling/>

First, you'll need to **clone this Workspace**.

- Click on the teardrop button on the top right.
- Select "Clone"
- Give your Workspace a meaningful name (perhaps with your name)
- Select the "SISBID-Wrangling-2022-student" billing project
- Click "CLONE WORKSPACE"

From your newly cloned Workspace you will **launch your RStudio instance**. You should:

- Click on the play button ("Cloud Environment") on the top right of this page
- Scroll down and click "CUSTOMIZE"
- Under "Application Configuration", scroll down to "Community maintained RStudio environments" and select "RStudio 4.2.0, Bioconductor 3.15, Python 3.8.10"
- Leave everything else as-is, and scroll down and click "CREATE"

WORKSPACE INFORMATION

Last Updated	7/22/2022
Creation Date	7/22/2022
Workflow Submissions	0
Access Level	Project Owner

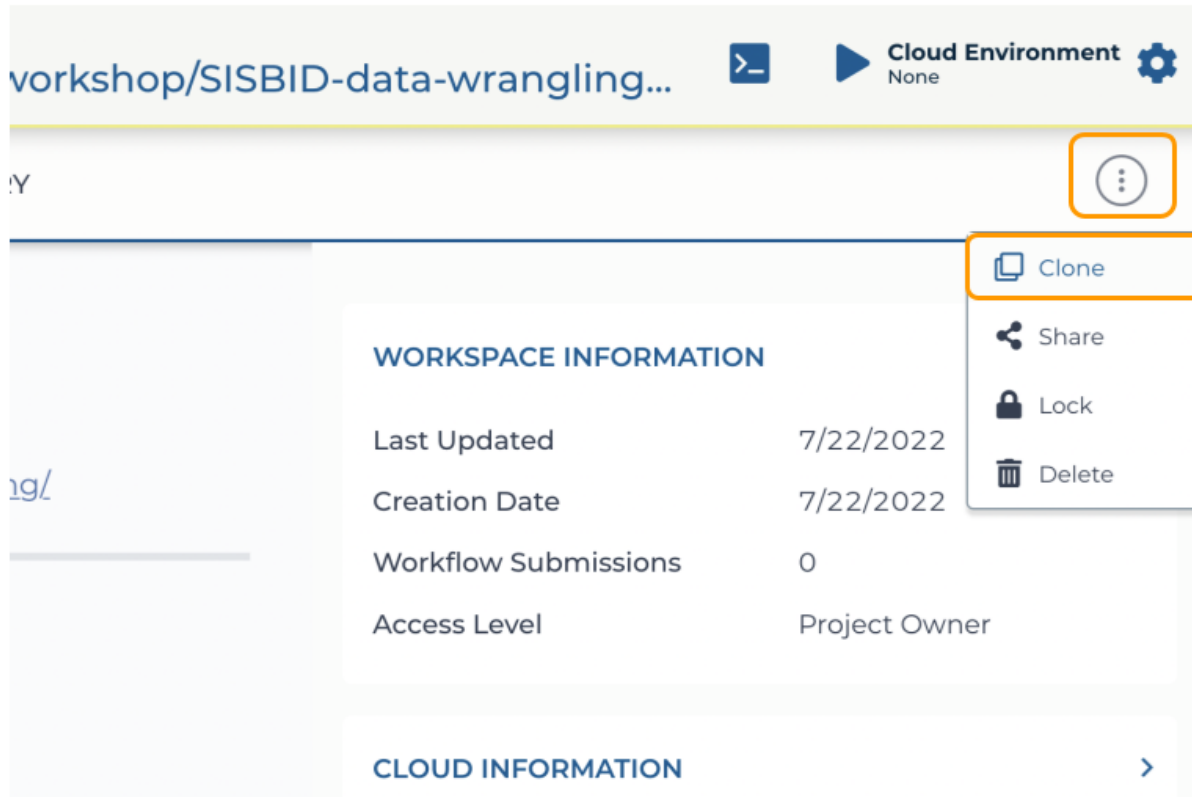
CLOUD INFORMATION

OWNERS

TAGS

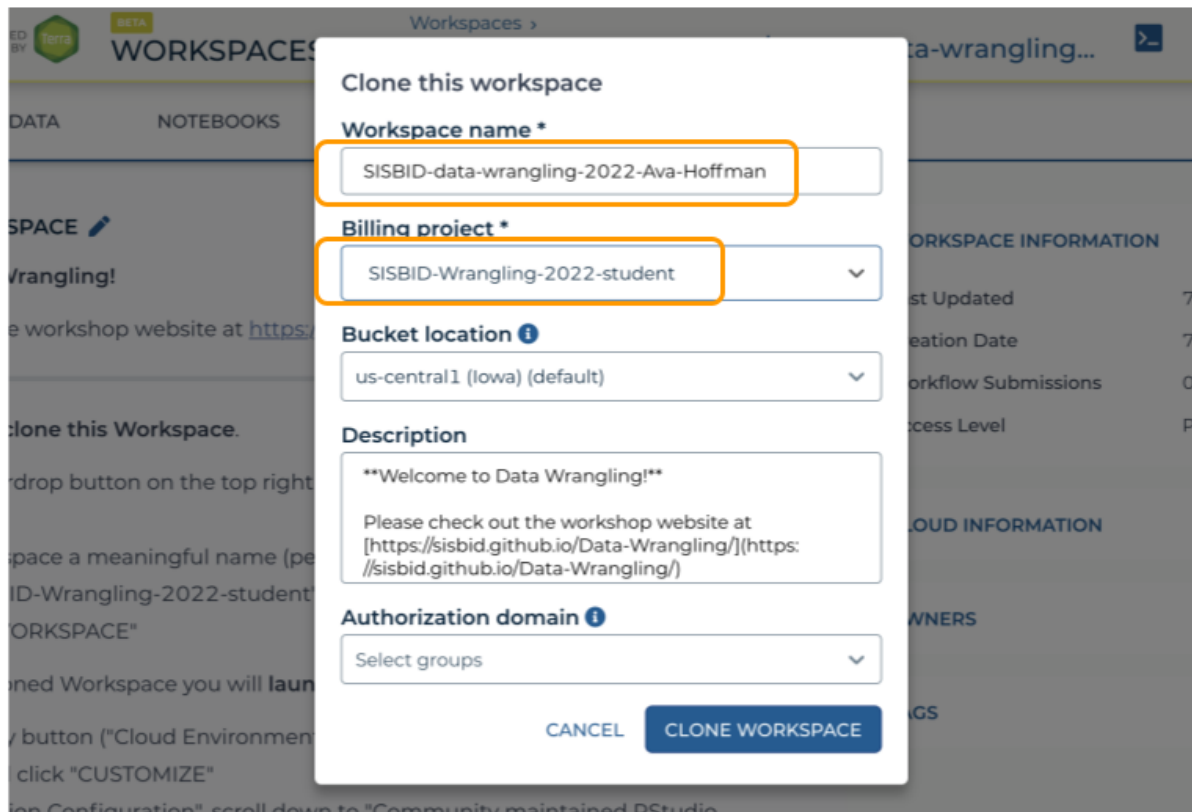
Clone the Workspace

Click on the teardrop button and select “Clone”.



Clone the Workspace

Name your Workspace (use your name!) and select the “SISBID-Wrangling-2022-student” billing project.



The screenshot shows a 'Clone this workspace' dialog box overlaid on a blurred background of the Google Cloud console. The dialog box contains the following fields and options:

- Workspace name ***: A text input field containing 'SISBID-data-wrangling-2022-Ava-Hoffman'.
- Billing project ***: A dropdown menu showing 'SISBID-Wrangling-2022-student'.
- Bucket location ⓘ**: A dropdown menu showing 'us-central1 (Iowa) (default)'.
- Description**: A text area containing the text:
Welcome to Data Wrangling!

Please check out the workshop website at [\[https://sisbid.github.io/Data-Wrangling/\]](https://sisbid.github.io/Data-Wrangling/)(<https://sisbid.github.io/Data-Wrangling/>)
- Authorization domain ⓘ**: A dropdown menu showing 'Select groups'.

At the bottom of the dialog box are two buttons: 'CANCEL' and 'CLONE WORKSPACE'.

Launch RStudio

Once in your newly cloned Workspace, you can launch the cloud instance! Click on “Cloud Environment” on the top right.

The screenshot shows the 'Data Wrangling' workspace interface. At the top, there's a navigation bar with a menu icon, a logo, 'POWERED BY Terra', 'BETA WORKSPACES', and the workspace name 'data-wrangling-workshop/SISBID-data-wrangling...'. On the right of the navigation bar, there's a 'Cloud Environment' button with a play icon and a gear icon. An orange arrow points to this button. Below the navigation bar, there's a tab bar with 'DASHBOARD', 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB HISTORY'. The main content area is divided into two columns. The left column has a section 'ABOUT THE WORKSPACE' with a welcome message and instructions on how to clone the workspace and launch the RStudio instance. The right column has a section 'WORKSPACE INFORMATION' with details like 'Last Updated', 'Creation Date', 'Workflow Submissions', and 'Access Level'. Below this, there are sections for 'CLOUD INFORMATION', 'OWNERS', and 'TAGS'.

WORKSPACES

Workspaces > data-wrangling-workshop/SISBID-data-wrangling...

Cloud Environment None

DASHBOARD DATA NOTEBOOKS WORKFLOWS JOB HISTORY

ABOUT THE WORKSPACE

Welcome to Data Wrangling!

Please check out the workshop website at <https://sisbid.github.io/Data-Wrangling/>

First, you'll need to **clone this Workspace**.

- Click on the teardrop button on the top right.
- Select "Clone"
- Give your Workspace a meaningful name (perhaps with your name)
- Select the "SISBID-Wrangling-2022-student" billing project
- Click "CLONE WORKSPACE"

From your newly cloned Workspace you will **launch your RStudio instance**. You should:

- Click on the play button ("Cloud Environment") on the top right of this page
- Scroll down and click "CUSTOMIZE"
- Under "Application Configuration", scroll down to "Community maintained RStudio environments" and select "RStudio 4.2.0, Bioconductor 3.15, Python 3.8.10"
- Leave everything else as-is, and scroll down and click "CREATE"

WORKSPACE INFORMATION

Last Updated	7/22/2022
Creation Date	7/22/2022
Workflow Submissions	0
Access Level	Project Owner

CLOUD INFORMATION

OWNERS

TAGS

Launch RStudio

Click “CUSTOMIZE”.

The screenshot shows the 'Data Wrangling Workspaces' interface. The left sidebar contains a navigation menu with 'DASHBOARD', 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB'. The main content area is titled 'ABOUT THE WORKSPACE' and includes a welcome message and instructions for cloning the workspace. The right panel, titled 'Cloud Environment', provides details about the default environment and its costs. A 'CUSTOMIZE' button is highlighted with an orange border.

Cloud Environment

A cloud environment consists of application configuration, cloud compute and persistent disk(s).

Use default environment CREATE

- Default: (GATK 4.2.4.0, Python 3.7.12, R 4.1.3) What's installed on this environment?
- Default compute size of **1 CPU(s), 3.75 GB memory**, and a **50 GB persistent disk** to keep your data even after you delete your compute
- [Learn more about Persistent disks and where your disk is mounted](#)
- This cloud environment will be created in the region **us-central1**. Copying data from a bucket in a different region may incur network egress charges. Note that network egress charges are not accounted for in cost estimates. For more information, particularly if you work with data stored in multiple cloud regions, please read the [documentation](#).

Running cloud compute cost	Paused cloud compute cost	Persistent disk cost
\$0.06 per hr	\$0.01 per hr	\$2.00 per month

Create custom environment CUSTOMIZE

Launch RStudio

From the “Application configuration” menu, select “RStudio 4.2.0, Bioconductor 3.15, Python 3.8.10”.

The screenshot displays the Sisbid-Workspaces interface. On the left, a sidebar contains navigation links: DASHBOARD, DATA, NOTEBOOKS, WORKFLOWS, and JOB. The main content area is titled 'ABOUT THE WORKSPACE' and includes a welcome message and instructions for cloning the workspace. On the right, a 'Cloud Environment' panel is open, showing a table of costs and a list of application configurations. The 'Application configuration' dropdown is open, showing a list of environments. The 'RStudio (R 4.2.0, Bioconductor 3.15, Python 3.8.10)' option is highlighted with an orange box.

Running cloud compute cost	Paused cloud compute cost	Persistent disk cost
\$0.06 per hr	\$0.01 per hr	\$2.00 per month

Application configuration

Default: (GATK 4.2.4.0, Python 3.7.12, R 4.1.3)

Legacy R / Bioconductor (R 4.1.1, Bioconductor 3.13, Python 3.7.10)

COMMUNITY-MAINTAINED JUPYTER ENVIRONMENTS (VERIFIED PARTNERS)

Pegasus (Pegasuspy 1.6.0, Python 3.7.12, harmony-pytorch 0.1.7, nmf-torch 0.1.1, scVI-tools 0.16.0)

OpenVINO integration with Tensorflow (openvino-tensorflow 1.1.0, Python 3.7.12, GATK 4.2.4.1)

COMMUNITY-MAINTAINED RSTUDIO ENVIRONMENTS (VERIFIED PARTNERS)

RStudio (R 4.2.0, Bioconductor 3.15, Python 3.8.10)

OTHER ENVIRONMENTS

Custom Environment

Launch RStudio

Click “CREATE”. Your Cloud Environment will take a few minutes to spin up.

The screenshot shows the RStudio Workspaces interface. On the left, the 'DASHBOARD' tab is active, displaying 'ABOUT THE WORKSPACE' and 'Welcome to Data Wrangling!'. It includes instructions on how to clone the workspace and launch the RStudio instance. On the right, the 'Cloud Environment' configuration panel is open. It shows the costs for running, paused, and persistent disk. The configuration options include a 'Standard VM' type, 'Enable autopause' checked with a 30-minute inactivity period, and a 'Location' dropdown. The 'Persistent disk' section shows 'Standard' disk type and a 50 GB size. At the bottom right, there are two buttons: 'Update Environment' and 'CREATE', with the 'CREATE' button highlighted by an orange rectangle.

Cloud Environment

A cloud environment consists of application configuration, cloud compute and persistent disk(s).

Running cloud compute cost	Paused cloud compute cost	Persistent disk cost
\$0.06 per hr	\$0.01 per hr	\$2.00 per month

Standard VM

☒ Enable autopause [Learn more about autopause.](#)

30 minutes of inactivity

Location BETA

Select...

Persistent disk

Persistent disks store analysis data. [Learn more about persistent disks and where your disk is mounted.](#)

Disk Type: Standard

Disk Size (GB): 50

Update Environment **CREATE**

Open RStudio

Click on “OPEN RSTUDIO” when the environment is ready.

The screenshot shows a web interface with a notification banner at the top. The banner contains the text "Your cloud environment is ready." and a close button (X). Below the banner, there is a button labeled "OPEN RSTUDIO" which is highlighted with an orange border. To the right of this button is a link labeled "Update cloud environment". Below the notification, there is a section titled "WORKSPACE INFORMATION" with a dropdown arrow. This section contains two rows of information: "Last Updated" with the value "7/22/2022" and "Creation Date" with the value "7/22/2022". On the left side of the interface, there is a link labeled "Vrangling/" and a section labeled "3 HISTORY".

angling-2022-student/SISBID-d

3 HISTORY

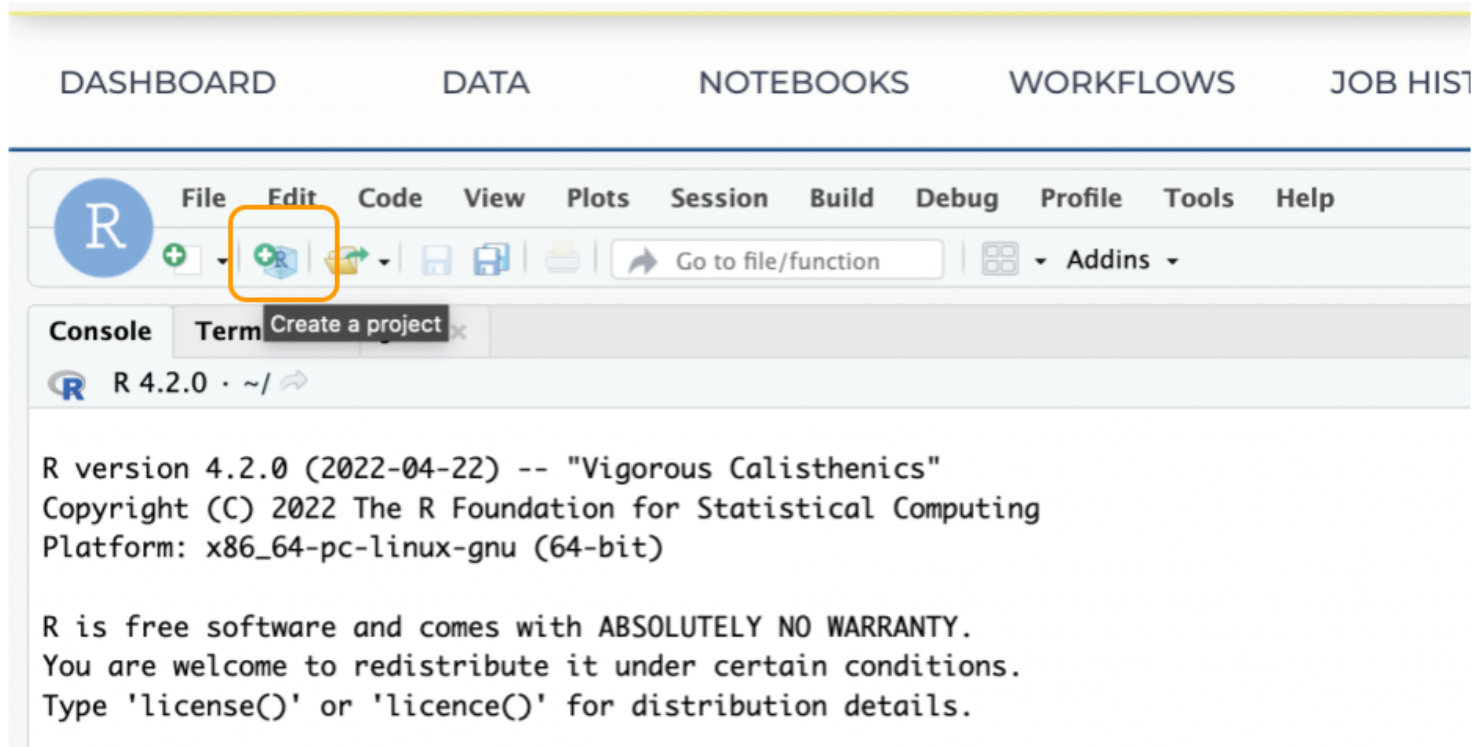
WORKSPACE INFORMATION

Last Updated	7/22/2022
Creation Date	7/22/2022

[Vrangling/](#)

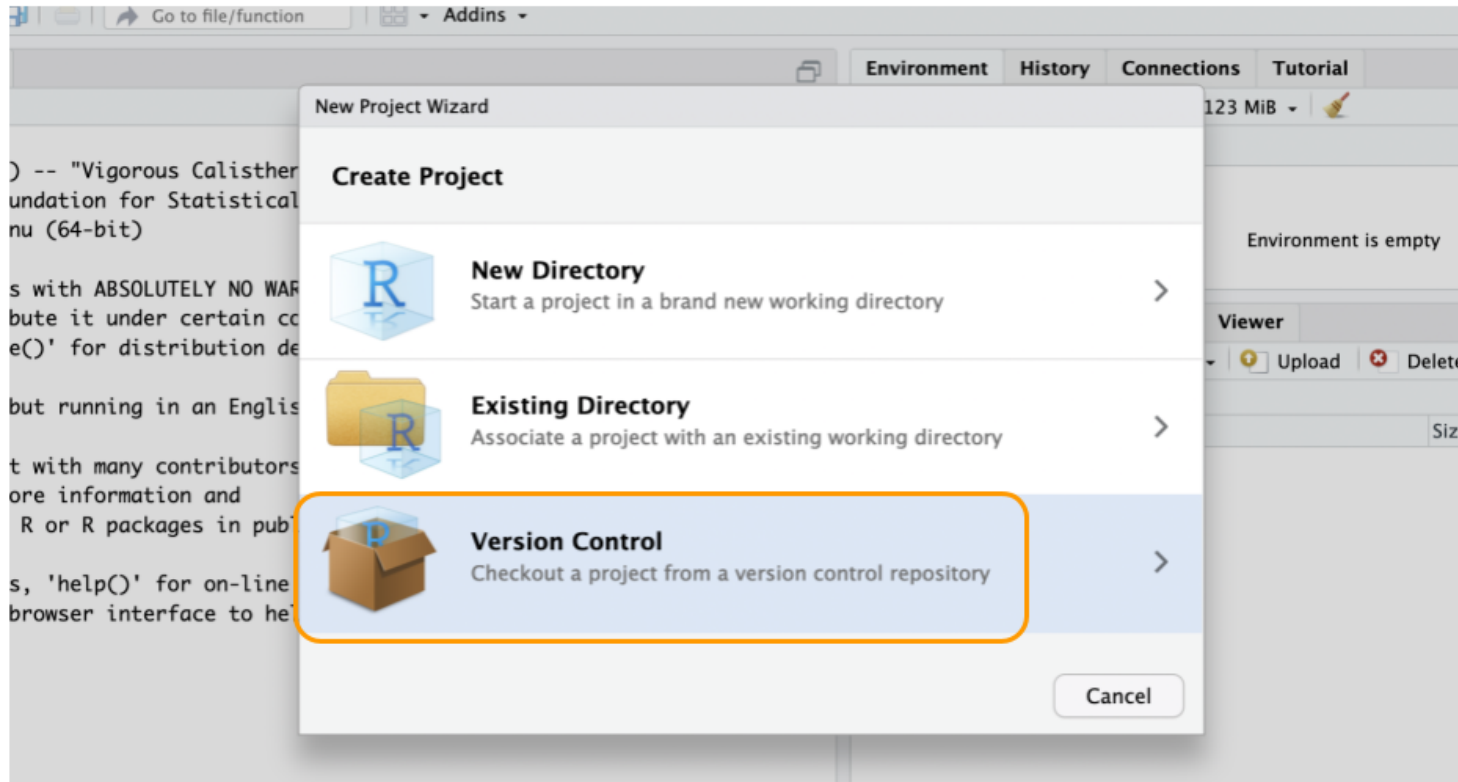
Create the Project

Once in RStudio, select the “New Project” button.



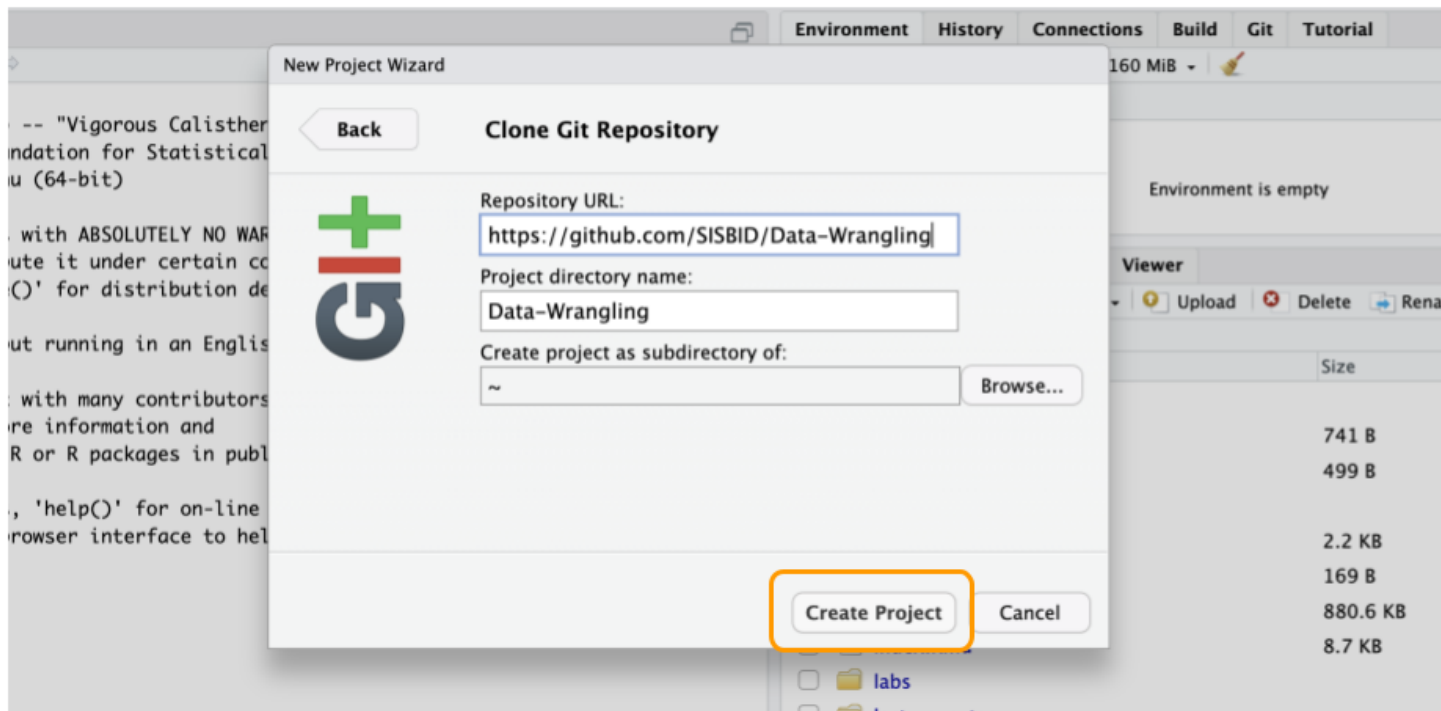
Create the Project

Select “Version Control”.



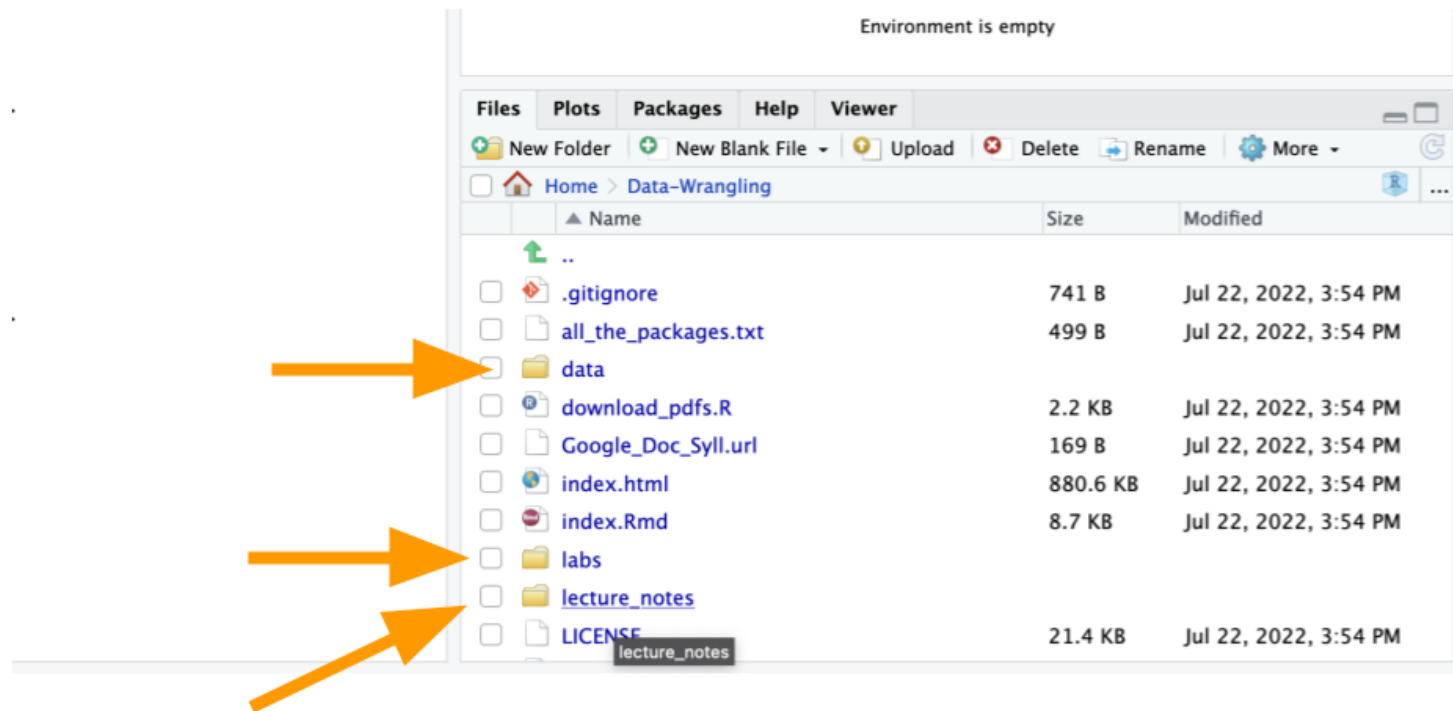
Create the Project

Enter the URL : <https://github.com/SISBID/Data-Wrangling>. Make sure the Project is a subdirectory of ~. Click "Create Project".



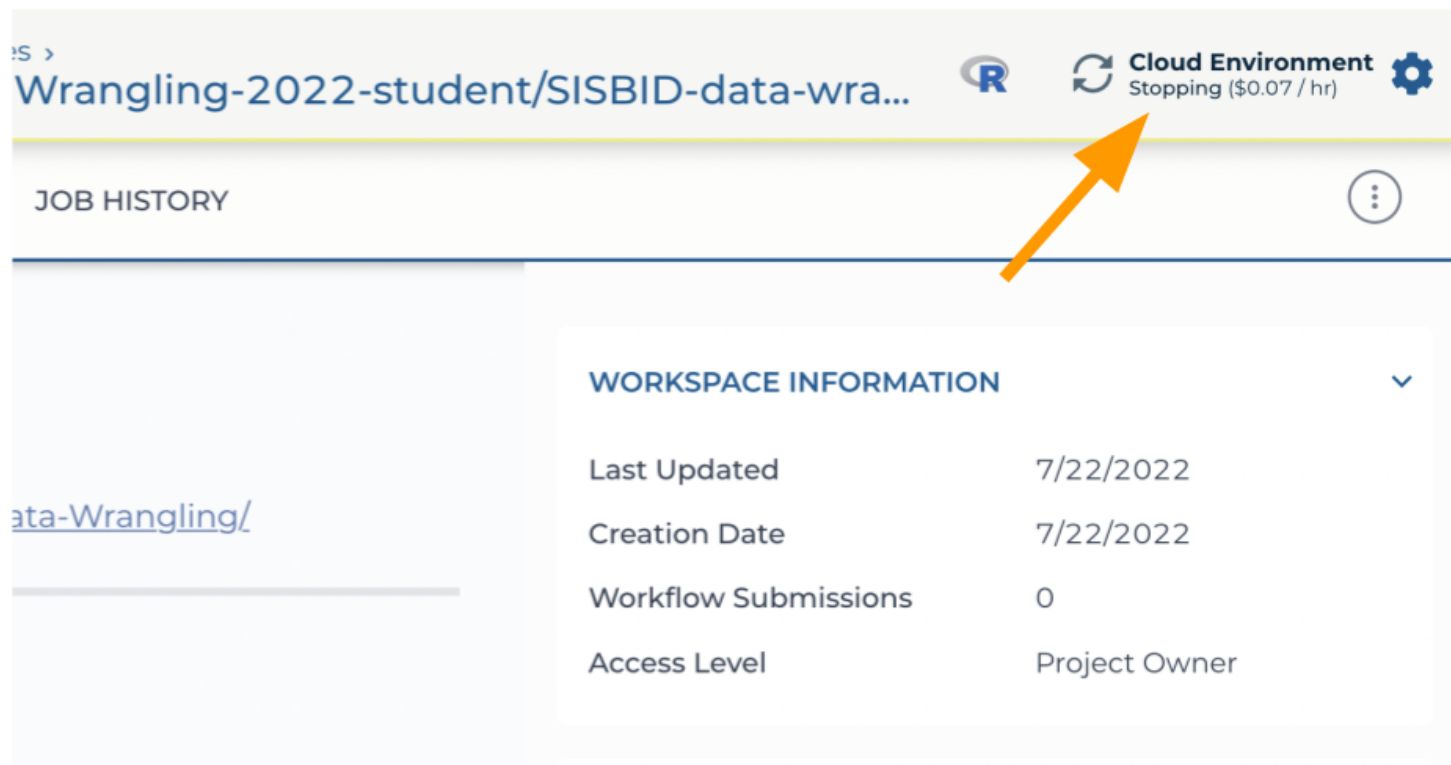
Create the Project

You should now see files listed in your workspace (including some datasets we'll be using). Feel free to change any files here - they are yours now! For the most up-to-date versions of files, please visit the website: <http://sisbid.github.io/Data-Wrangling/>.



Pause the Environment

It's very important to stop your cloud environment when you are done so you don't accumulate too many charges. Between class sessions, you can click the pause button on the top right. This frees up resources for others!



The screenshot shows the AWS SageMaker console interface. At the top, the breadcrumb navigation reads "Wrangling-2022-student/SISBID-data-wra...". To the right of the breadcrumb is the SageMaker logo and a "Cloud Environment" button with a circular arrow icon. Below the button, it says "Stopping (\$0.07 / hr)". An orange arrow points to this button. Below the breadcrumb, there is a "JOB HISTORY" section. On the left side, there is a link "ata-Wrangling/". In the center, there is a "WORKSPACE INFORMATION" panel with a dropdown arrow. The panel contains the following information:

WORKSPACE INFORMATION	
Last Updated	7/22/2022
Creation Date	7/22/2022
Workflow Submissions	0
Access Level	Project Owner