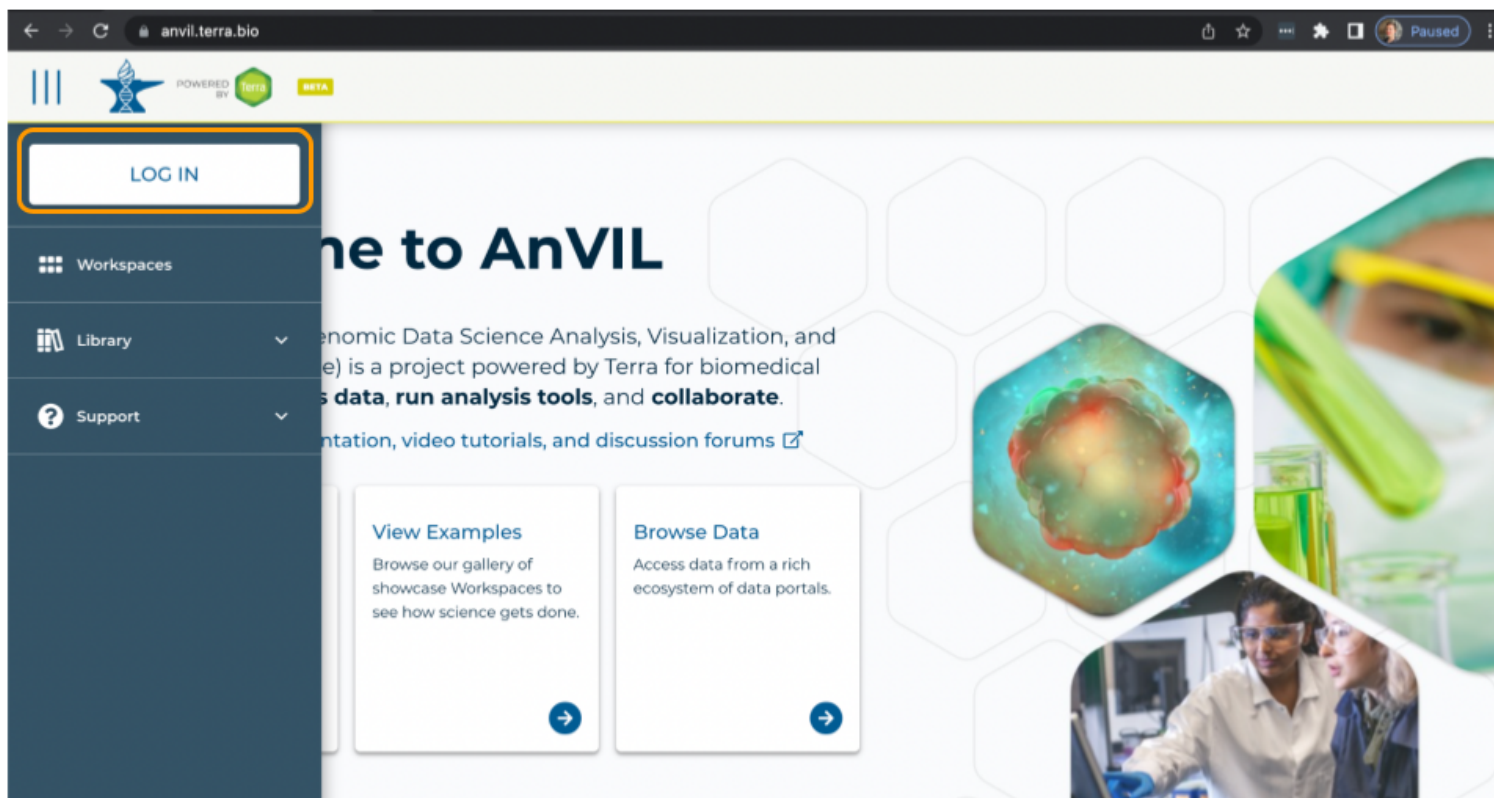


Setup on AnVIL

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', '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 side, there are expandable sections for '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

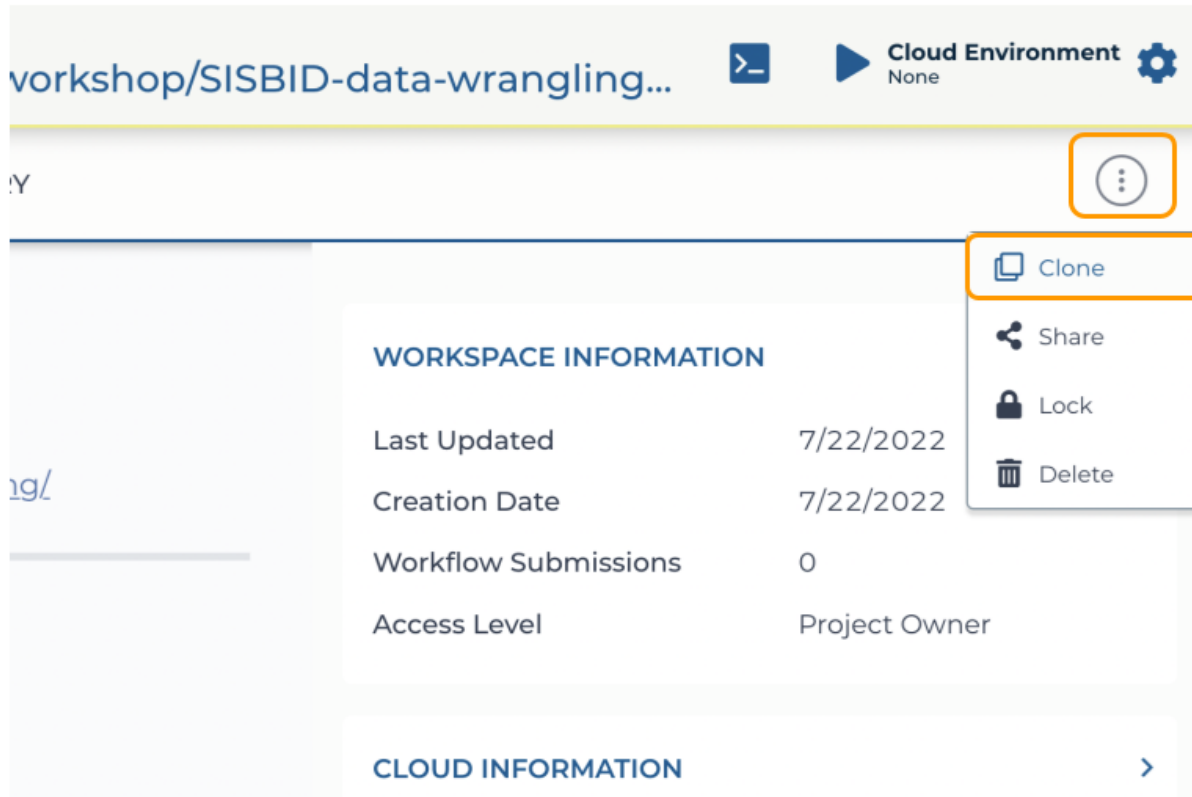
CLOUD INFORMATION

OWNERS

TAGS

Clone the Workspace

Click on the teardrop button and select “Clone”.



The screenshot shows a web interface for managing a workspace. At the top, there is a header bar with the text "workshop/SISBID-data-wrangling..." and a "Cloud Environment" section set to "None". Below the header, a sidebar on the left contains a search bar and a list of items. The main content area displays "WORKSPACE INFORMATION" with the following details:

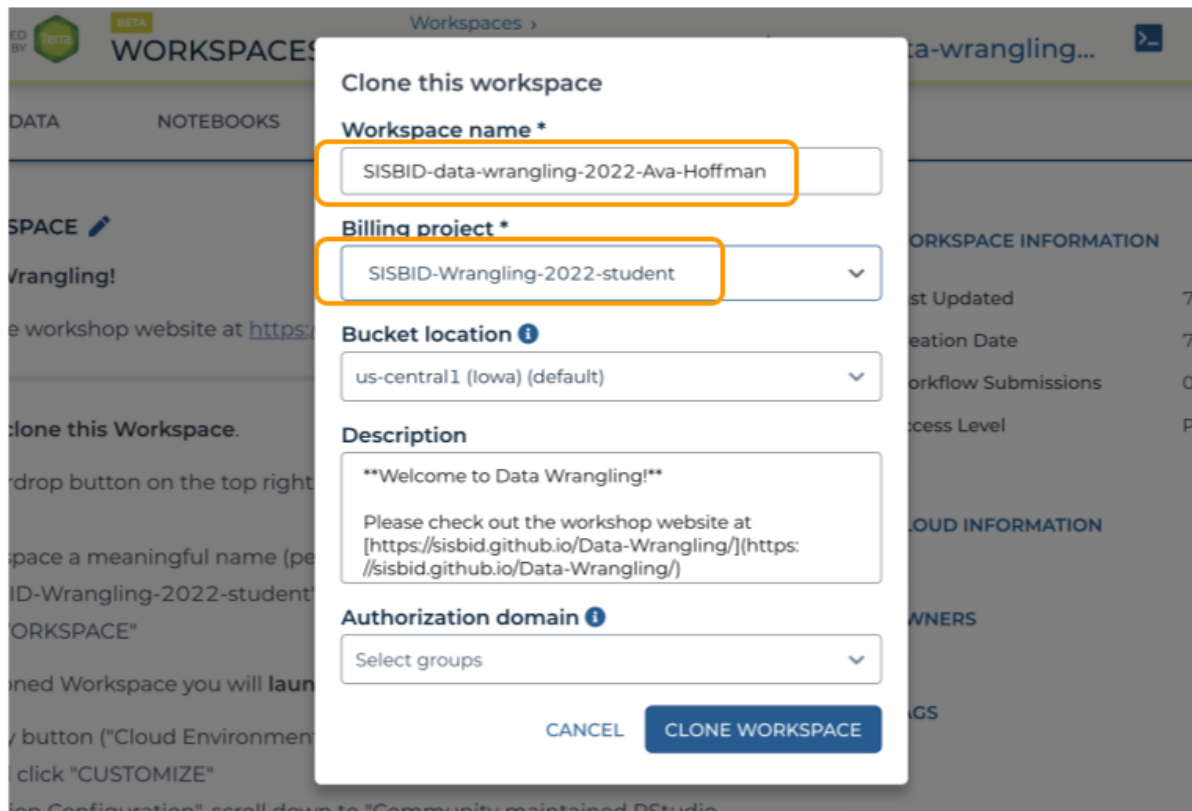
WORKSPACE INFORMATION	
Last Updated	7/22/2022
Creation Date	7/22/2022
Workflow Submissions	0
Access Level	Project Owner

Below the workspace information, there is a section for "CLOUD INFORMATION" with a right-pointing arrow. A dropdown menu is open, triggered by a teardrop button (three vertical dots) in the top right corner of the workspace card. The menu contains the following options:

- Clone
- Share
- Lock
- Delete

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 the Google Cloud Platform 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, there 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 top navigation bar of the Data Wrangling workspace. The 'Cloud Environment' button, located on the top right, is highlighted with an orange arrow. The button is labeled 'Cloud Environment' and 'None'.

The main content area on the left contains the following text:

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"

The right sidebar contains the following sections:

- WORKSPACE INFORMATION** (dropdown arrow)
 - Last Updated: 7/22/2022
 - Creation Date: 7/22/2022
 - Workflow Submissions: 0
 - Access Level: Project Owner
- CLOUD INFORMATION** (chevron right)
- OWNERS** (chevron right)
- TAGS** (chevron right)

Launch RStudio

Click “CUSTOMIZE”.

The screenshot shows the 'Data Wrangling Workspaces' interface. On the left, the 'ABOUT THE WORKSPACE' section provides instructions for cloning and launching the workspace. On the right, the 'Cloud Environment' panel is open, showing a 'Use default environment' option with a 'CREATE' button. Below this, a list of details about the default environment is provided, including default software versions, compute size, and disk size. A table at the bottom of the panel shows the costs for running, paused, and persistent disk. At the bottom right of the panel, there is a 'Create custom environment' section with a 'CUSTOMIZE' button highlighted by an orange border.

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**

- 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 environments" and select "RStudio 4.2.0, Bioconductor 3.15, Python 3.9"
- Leave everything else as-is, and scroll down and click "CREATE"

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 a navigation menu with 'DASHBOARD', 'DATA', 'NOTEBOOKS', 'WORKFLOWS', and 'JOB'. The main content area is titled 'ABOUT THE WORKSPACE' and includes a welcome message, a link to the workshop website, and instructions on how to clone the workspace and launch an RStudio instance. 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 set to 'Default: (GATK 4.2.4.0, Python 3.7.12, R 4.1.3)'. Below this, the 'COMMUNITY-MAINTAINED JUPYTER ENVIRONMENTS (VERIFIED PARTNERS)' section lists several options. The 'COMMUNITY-MAINTAINED RSTUDIO ENVIRONMENTS (VERIFIED PARTNERS)' section is highlighted with an orange box, and the option 'RStudio (R 4.2.0, Bioconductor 3.15, Python 3.8.10)' is selected.

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 displays the SISBIID-Workspaces interface. On the left, a sidebar contains navigation links: DASHBOARD, DATA, NOTEBOOKS, WORKFLOWS, and JOBS. The main content area is titled 'ABOUT THE WORKSPACE' and includes a welcome message and instructions for cloning and launching the workspace. On the right, a 'Cloud Environment' configuration panel is open. This panel shows the costs for running, paused, and persistent disk usage. It allows users to select a VM type (Standard VM), enable or disable autopause with a specified inactivity duration (30 minutes), and choose a location. Below these options, the 'Persistent disk' section lets users select a disk type (Standard) and set the disk size (50 GB). At the bottom right of the configuration panel, there are two buttons: 'Update Environment' and 'CREATE'. The 'CREATE' button is highlighted with an orange border.

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.

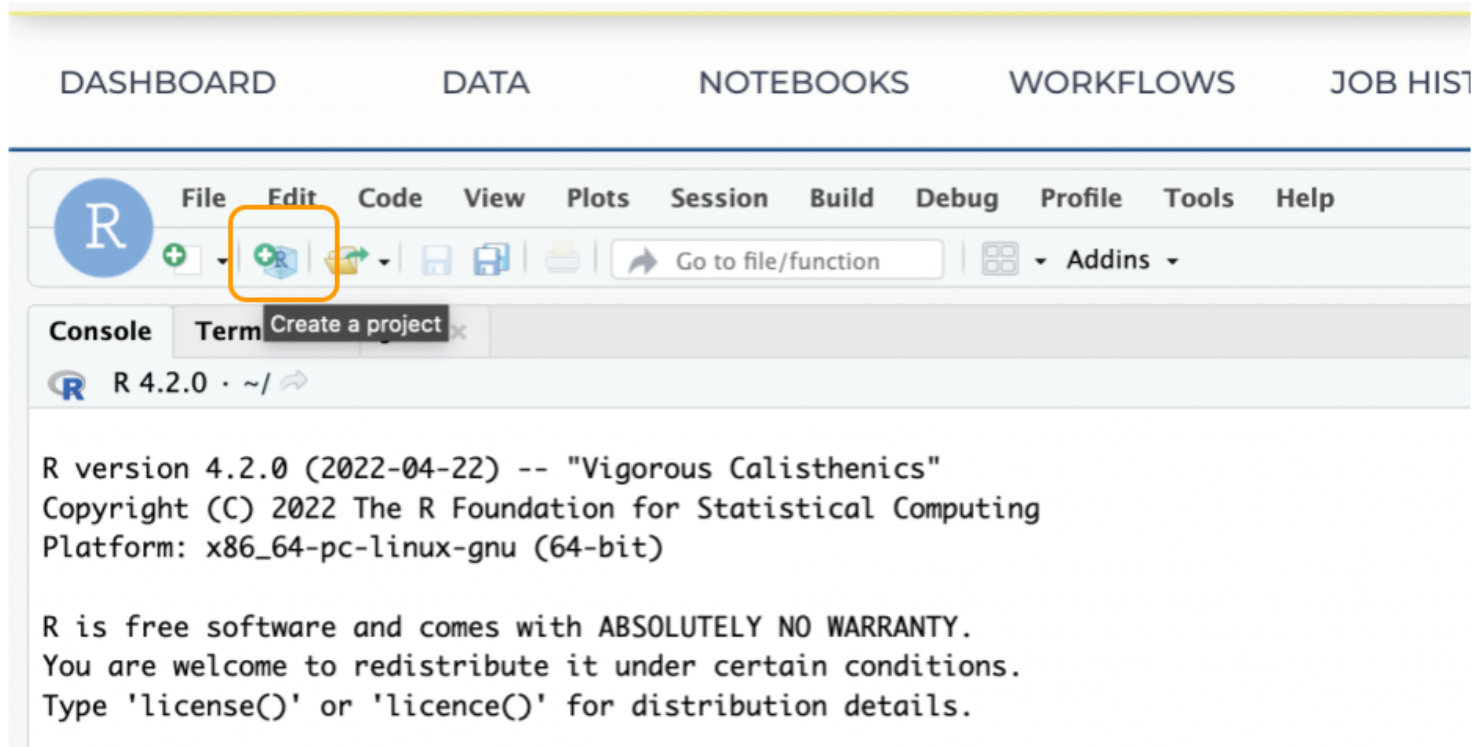
The screenshot shows a web interface with a notification banner at the top. The banner has a blue header with the text "Your cloud environment is ready." and a close button (X). Below the header, there are two buttons: "OPEN RSTUDIO" (highlighted with an orange border) and "Update cloud environment" (partially obscured by a black rectangle). The background of the page shows a URL "angling-2022-student/SISBID-d" and a section titled "3 HISTORY". Below this, there is a "WORKSPACE INFORMATION" section with a dropdown arrow. The workspace information table shows:

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

On the left side of the workspace information section, there is a link [Vrangling/](#).

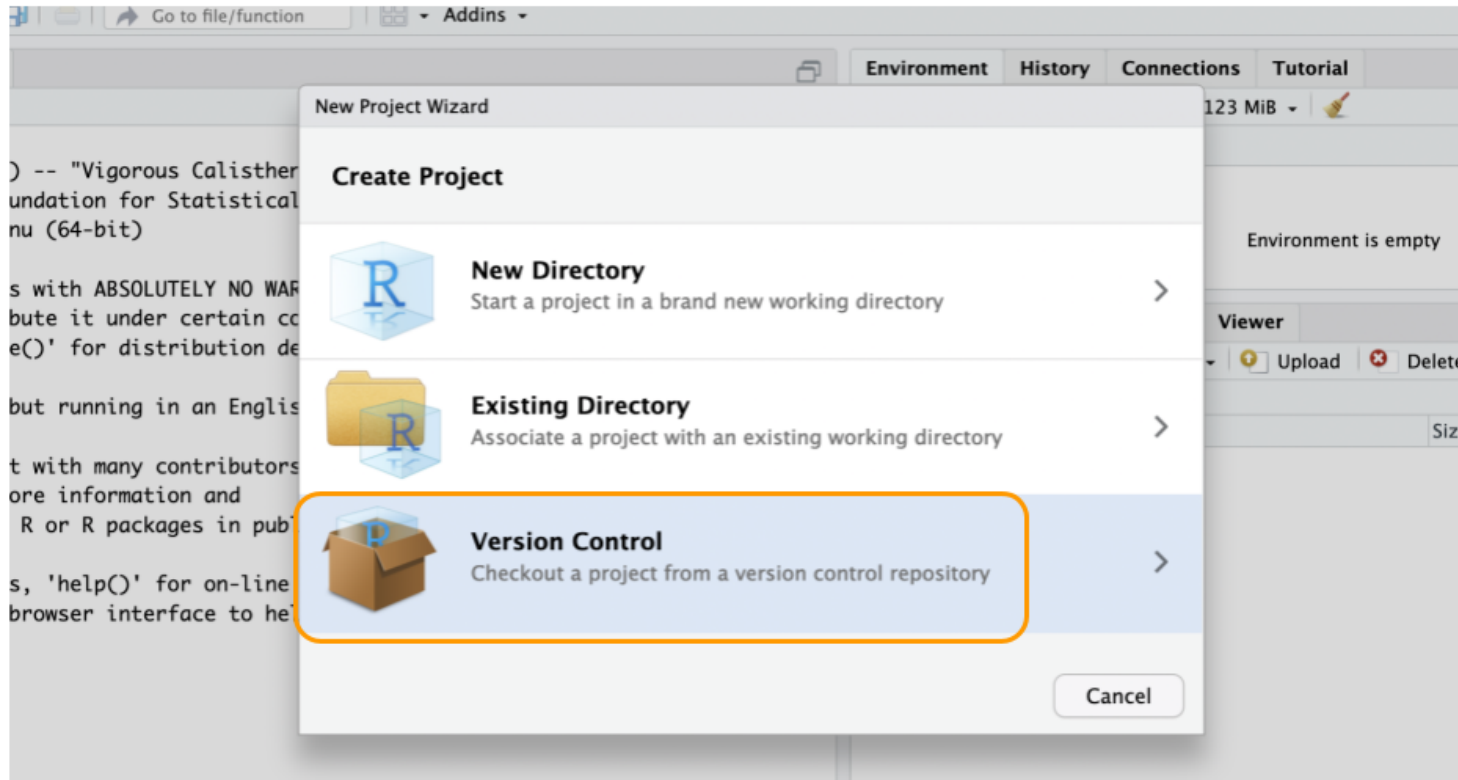
Create the Project

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



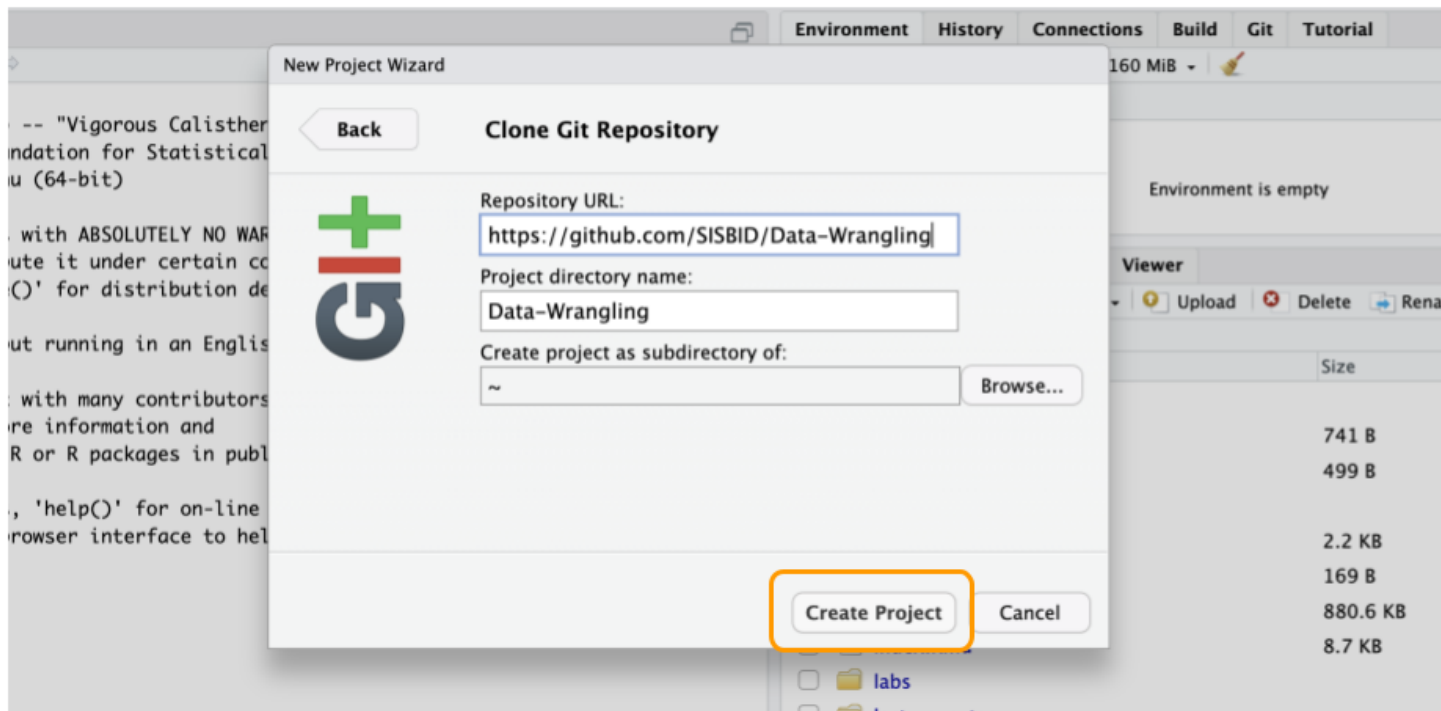
Create the Project

Select “Version Control”.



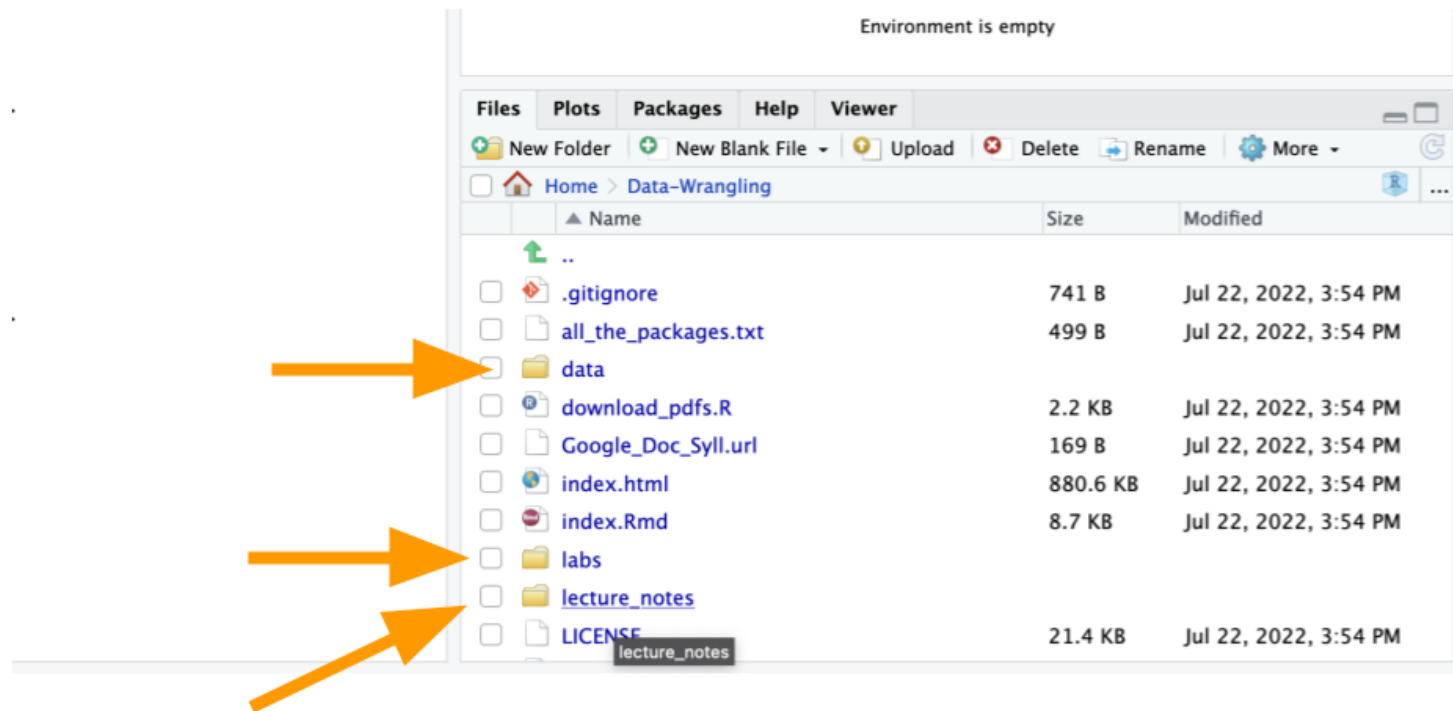
Create the Project

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



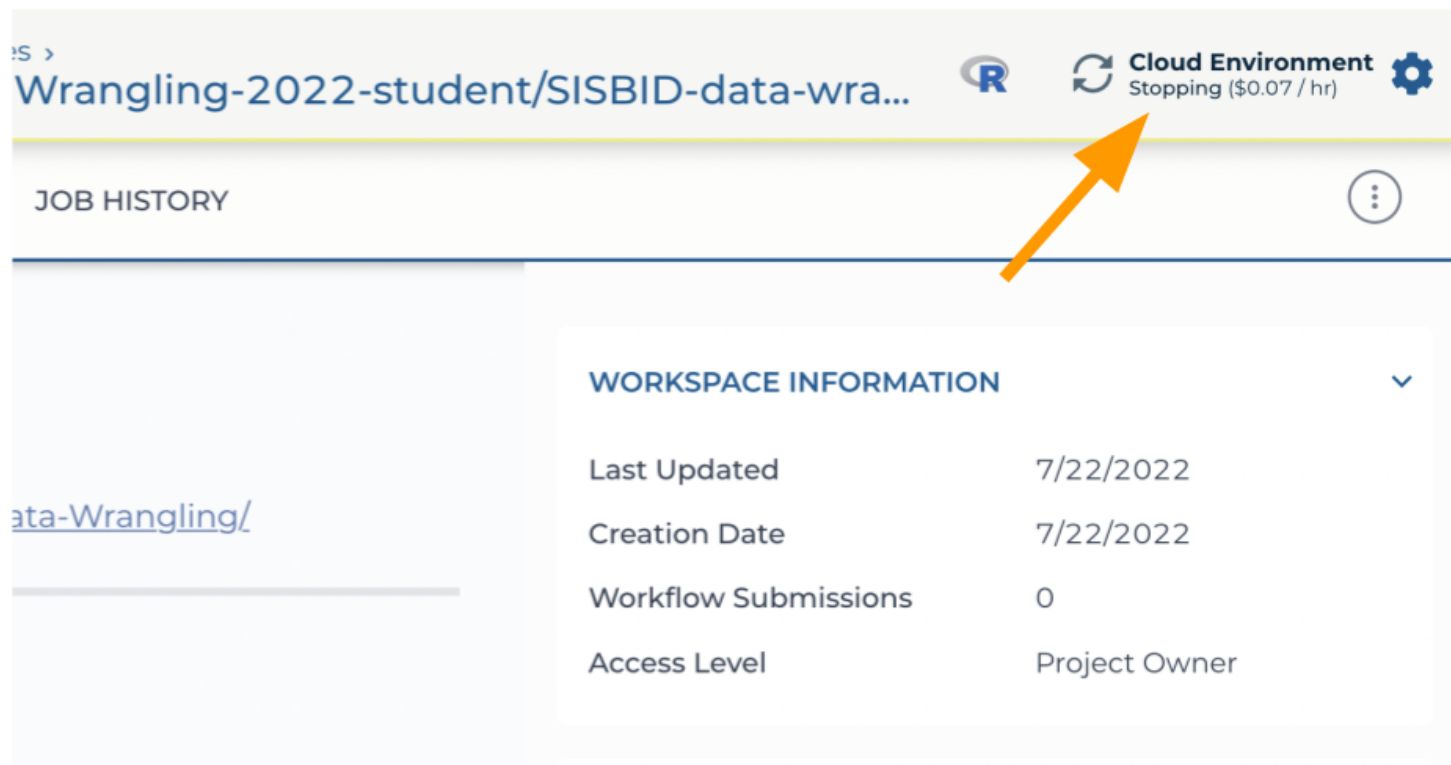
Create the Project

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 a blue "R" logo, a circular refresh icon, and a "Cloud Environment" button with a gear icon. Below the button, it says "Stopping (\$0.07 / hr)". An orange arrow points to the "Cloud Environment" button. Below the top bar, there is a "JOB HISTORY" section. A "WORKSPACE INFORMATION" panel is open, showing the following details:

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