Data Acquisition

Day 2.
Git Branch
HTTP Requests/Responses
File-based Acquisition
Streamlit Basic

Diane Woodbridge, PH.D

UNIVERSITY OF SAN FRANCISCO

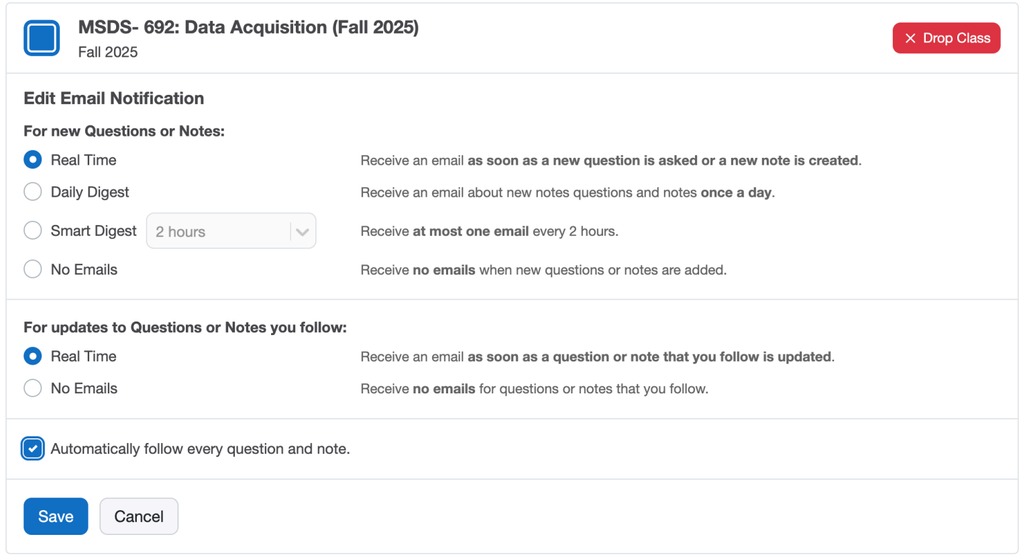
CHANGE THE WORLD FROM HERE

Announcement

HW 1 Due : Sept 1

- Individual Assignment

Don’t forget to join Piazza and change the settings to receive emails on the user setting.



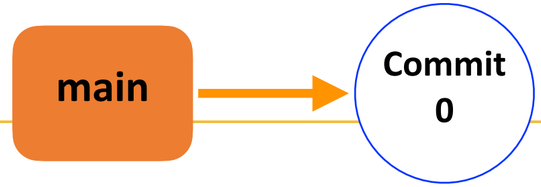
Office Hours : 12 PM - 12:45 PM on Mon/Thu

- If you’d need to schedule something outside of the office hours, please send me a slack message.
- If you want to keep your space/grab lunch till your turn, you can write your name on the whiteboard at my office.

Contents

Contents

Git Branch



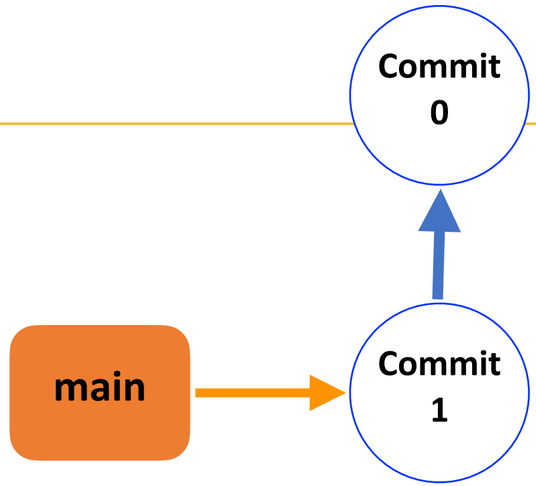
Branch

- Diversion from the main line of development, and work without messing with the main line.
  - You can think it as a parallel universe for your code.
    - main branch - default, stable and production-ready
    - feature branch - separate line where you can write code without messing up the main branch
- Benefits
  - Work in isolation : developers can work on new features or fix bugs simultaneously without affecting other’s code
  - Stable Codebase : The main branch remains stable and updates will be merged after being reviewed and tested.

Git Branch

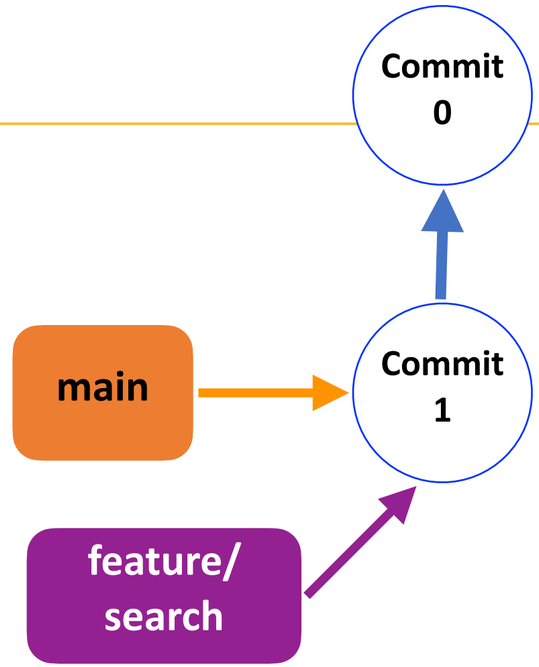
Branch

- Diversion from the main line of development, and work without messing with the main line.
  - You can think it as a parallel universe for your code.
    - main branch - default, stable and production-ready
    - feature branch - separate line where you can write code without messing up the main branch
- Benefits
  - Work in isolation : developers can work on new features or fix bugs simultaneously without affecting other's code
  - Stable Codebase : The main branch remains stable and updates will be merged after being reviewed and tested.



Note : Commit 0 is its Parent

Git Branch



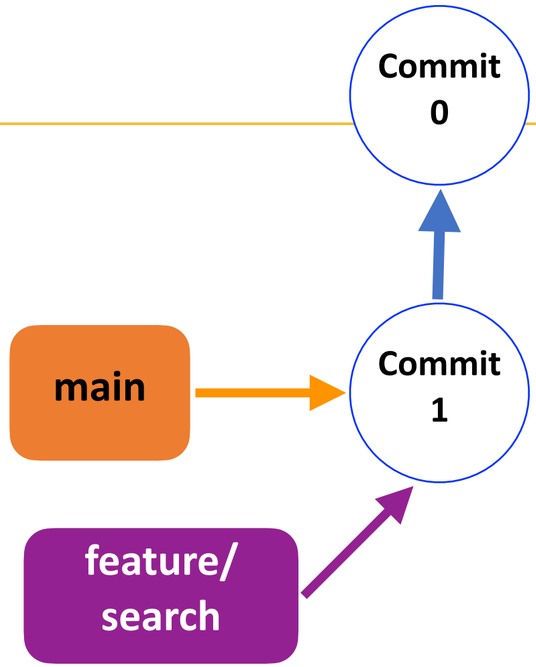
Branch

- Diversion from the main line of development, and work without messing with the main line.
  - You can think it as a parallel universe for your code.
    - main branch - default, stable and production-ready
    - feature branch - separate line where you can write code without messing up the main branch
- Benefits
  - Work in isolation : developers can work on new features or fix bugs simultaneously without affecting other's code
  - Stable Codebase : The main branch remains stable and updates will be merged after being reviewed and tested.

Git Branch

Create a branch

- `$ git branch new_branch_name`
- Best practice for naming a branch
  - You may have 1) main branch for production, 2) develop branch for the pre-production and testing stage.
  - Use a prefix for the type of work to indicate the purpose
    - Ex. feature/ (new feature), hotfix/ (urgent production fix), release/ (new production release)
  - Use short but descriptive names
    - Ex. feature/search, hotfix/login
  - Use lowercase and hyphens



https://www.gitkraken.com/learn/git/git-flow

Git Branch

Create a branch

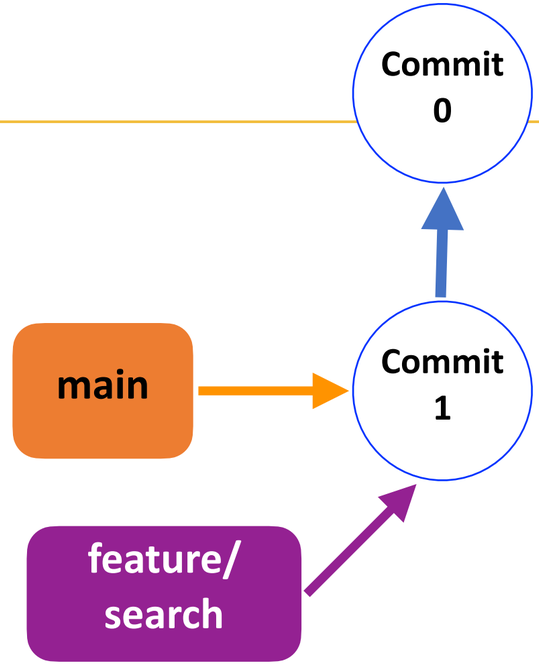
- `$ git branch new_branch_name`

Check the list of branches

- `$ git branch`
  - Your current branch will be indicated with *

Switch to a different branch

- `$ git checkout branch_name`
- `$ git switch branch_name`



Git Branch

Create a branch

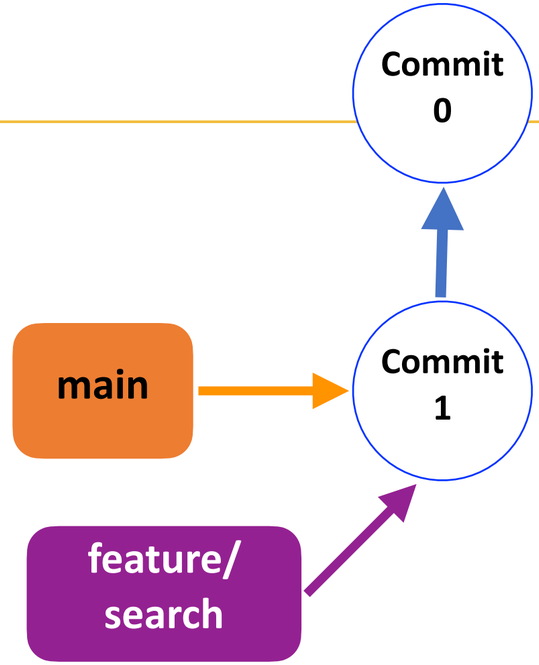
- `$ git branch new_branch_name`

Check the list of branches

- `$ git branch`
  - Your current branch will be indicated with *

Switch to a different branch

- `$ git checkout branch_name`
- `$ git switch branch_name`

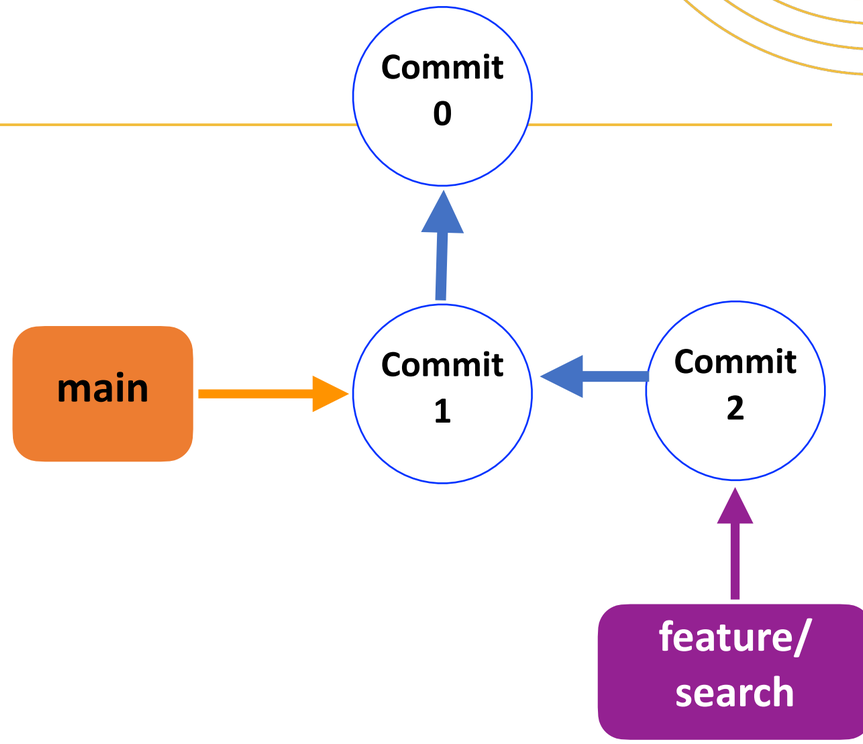



```
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git branch feature/search
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git branch
  feature/search
* main
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git checkout feature/search
Switched to branch 'feature/search'
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git branch
* feature/search
  main
```

Example 1

1. On feature/search branch, make some changes and check where main and feature/search branch are currently at.

2. Switch to the main branch, make some changes and check e main and feature/ search branch are currently at.

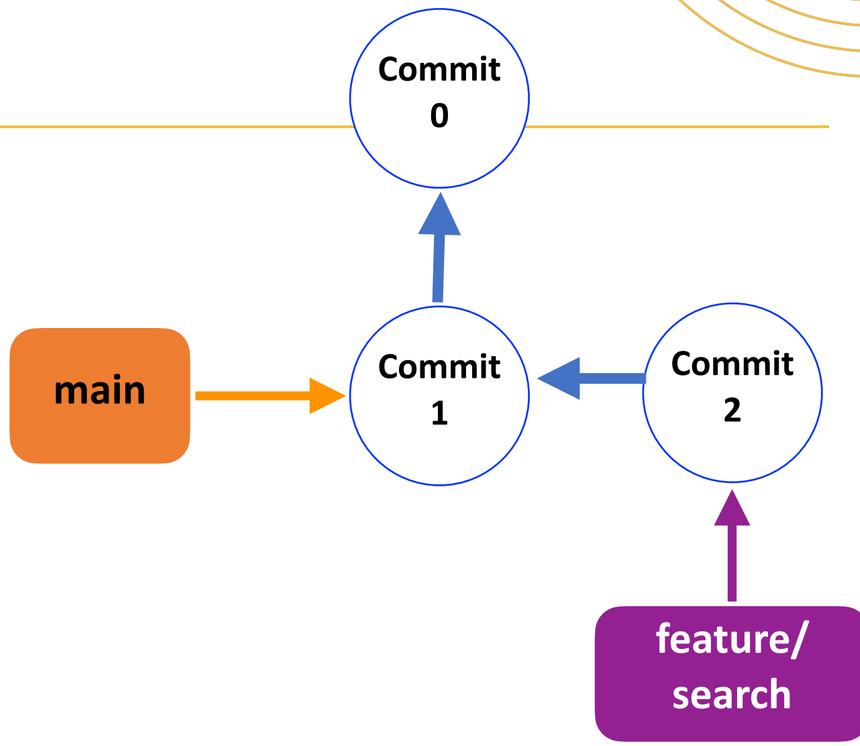



```
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % echo "# Example 1" > Day2/ex01.txt
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git add Day2/ex01.txt
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git commit -m "Added ex01.txt"
[feature/search 317bfa2] Added ex01.txt
 1 file changed, 1 insertion(+)
 create mode 100644 Day2/ex01.txt
```

Example 1

1. On feature/search branch, make some changes and check where main and feature/search branch are currently at.

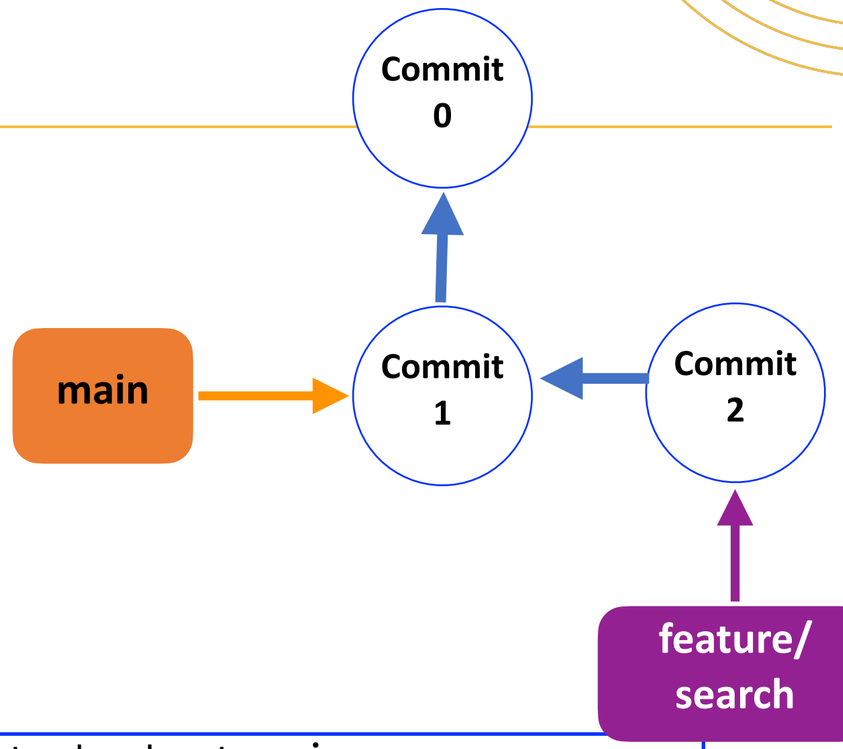
```
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git log --graph --all
* commit 317bfa228309e07eec8497a0d6d76037b793c629 (HEAD -> feature/search)
| Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| Date:   Sat Aug 23 11:26:33 2025 -0700
|
|     Added ex01.txt
|
* commit 3745965b5864e0729719c6f24acf7f04c8f90745 (origin/main, origin/HEAD, main)
| Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| Date:   Thu Aug 21 08:07:39 2025 -0700
|
|     added README.md
|
* commit 8e8d460060e451073c2c9db0a95d74a550995e19
| Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| Date:   Thu Aug 21 08:01:47 2025 -0700
|
|     created Day1 and environment file
|
* commit b648cca4076f63969dd38dfef61354892a7d8458
  Author: Diane Woodbridge <dwoodbridge@usfca.edu>
  Date:   Thu Aug 21 07:13:40 2025 -0700

      Initial commit
```



Example 1

1. On feature/search branch, make some changes and check where main and feature/search branch are currently at.

2. Switch to the main branch, make some changes and check e main and feature/ search branch are currently at.

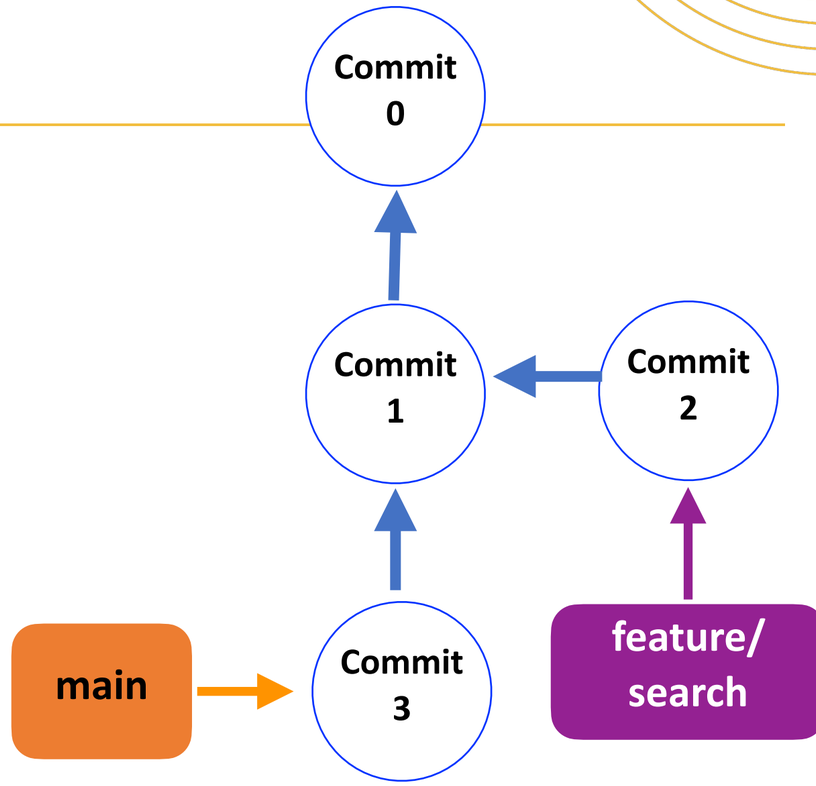



```
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git checkout main
Switched to branch 'main'
Your branch is up to date with 'origin/main'.
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % echo "hello" > Day2/temp.txt
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git add Day2/temp.txt
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git commit -m "added temp file"
[main 104b136] added temp file
 1 file changed, 1 insertion(+)
 create mode 100644 Day2/temp.txt
```

Git Branch

2. Switch to the main branch, make some changes and check e main and feature/ search branch are currently at.

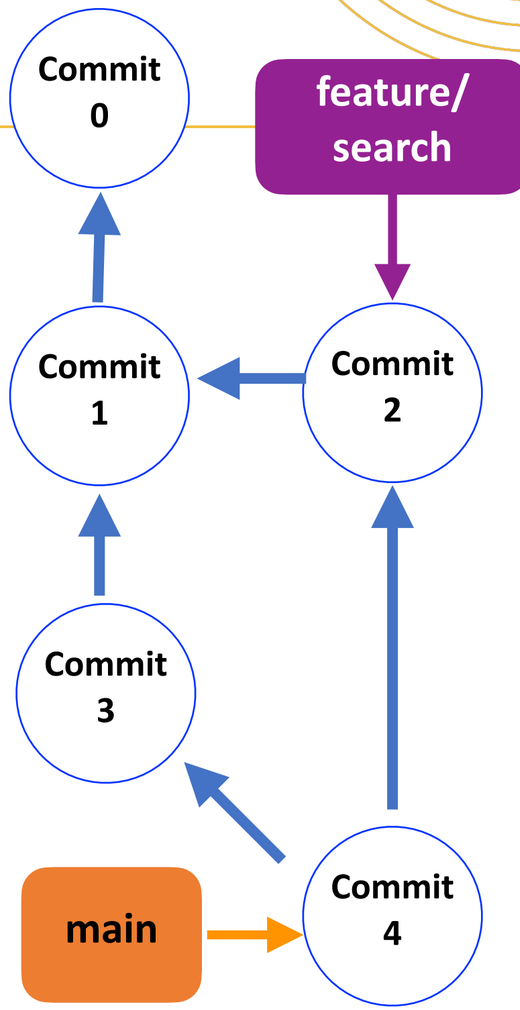
```
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git log --graph --all
* commit 104b136490784b56839776a314e788a17f5969af (HEAD -> main)
| Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| Date:   Sat Aug 23 12:58:44 2025 -0700
|
|     added temp file
|
| * commit 317bfa228309e07eec8497a0d6d76037b793c629 (feature/search)
|/  Author: Diane Woodbridge <dwoodbridge@usfca.edu>
|   Date:   Sat Aug 23 11:26:33 2025 -0700
|
|       Added ex01.txt
|
* commit 3745965b5864e0729719c6f24acf7f04c8f90745 (origin/main, origin/HEAD)
| Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| Date:   Thu Aug 21 08:07:39 2025 -0700
|
|     added README.md
|
* commit 8e8d460060e451073c2c9db0a95d74a550995e19
| Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| Date:   Thu Aug 21 08:01:47 2025 -0700
|
|     created Day1 and environment file
|
* commit b648cca4076f63969dd38dfef61354892a7d8458
  Author: Diane Woodbridge <dwoodbridge@usfca.edu>
  Date:   Thu Aug 21 07:13:40 2025 -0700

      Initial commit
```



Git Branch

Merging Branches

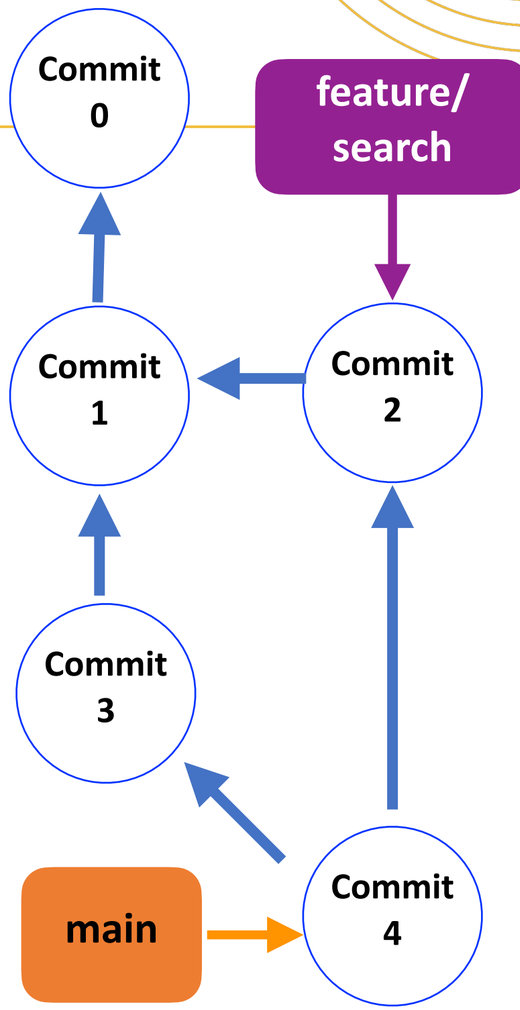
- Combining branches
  - If two branches have both have new commits from the common ancestor, it will combine changes and create a new merge commit.
- First, move to the branch that you'd like to combine the other branch into and then merge.
  - `$ git checkout branch_name_1`
  - `$ git merge branch_name_2`



Git Branch

Merging Branches

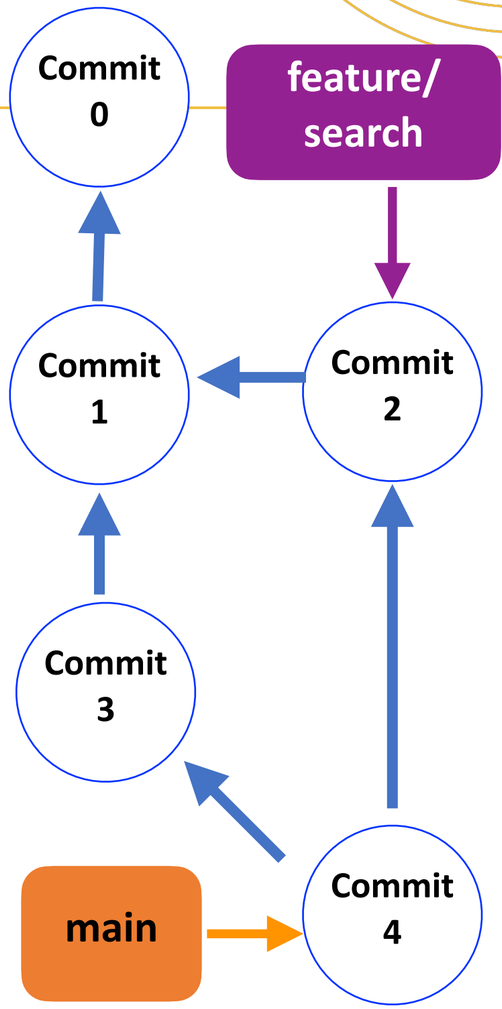
- Merge Conflict
  - When two branches both edited same lines, merge conflict can happen.
    - To resolve merge conflict, edit the file, add and commit to finish merge.



Example 2

Merge feature/search to main and check its log graph.

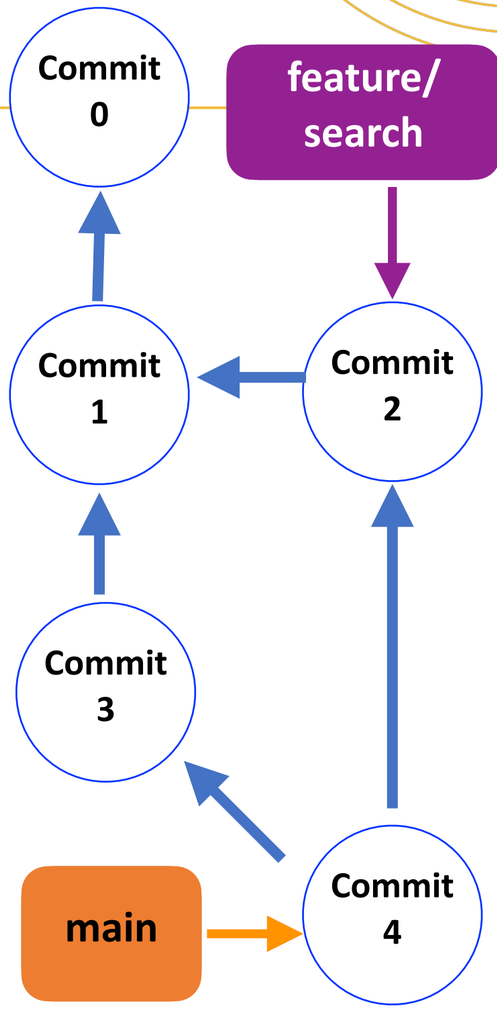
```
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git checkout main
Already on 'main'
Your branch is ahead of 'origin/main' by 1 commit.
  (use "git push" to publish your local commits)
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git merge feature/search
Merge made by the 'ort' strategy.
 Day2/ex01.txt | 1 +
 1 file changed, 1 insertion(+)
 create mode 100644 Day2/ex01.txt
```



Example 2

Merge feature/search to main and check its log graph.

```
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git log --graph --all
*   commit b1b7c651cc669752061640432l108a3fdb252511 (HEAD -> main)
|\  Merge: 104b136 317bfa2
| | Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| | Date:   Sat Aug 23 13:24:23 2025 -0700
| |     Merge branch 'feature/search'
| * commit 317bfa228309e07eec8497a0d6d76037b793c629 (feature/search)
| | Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| | Date:   Sat Aug 23 11:26:33 2025 -0700
| |     Added ex01.txt
* | commit 104b136490784b56839776a314e788a17f5969af
|/  Author: Diane Woodbridge <dwoodbridge@usfca.edu>
|   Date:   Sat Aug 23 12:58:44 2025 -0700
|       added temp file
* commit 3745965b5864e0729719c6f24acf7f04c8f90745 (origin/main, origin/HEAD)
| Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| Date:   Thu Aug 21 08:07:39 2025 -0700
|     added README.md
* commit 8e8d460060e451073c2c9db0a95d74a550995e19
| Author: Diane Woodbridge <dwoodbridge@usfca.edu>
| Date:   Thu Aug 21 08:01:47 2025 -0700
|     created Day1 and environment file
* commit b648cca4076f63969dd38dfef61354892a7d8458
  Author: Diane Woodbridge <dwoodbridge@usfca.edu>
  Date:   Thu Aug 21 07:13:40 2025 -0700

      Initial commit
```



Git Branch

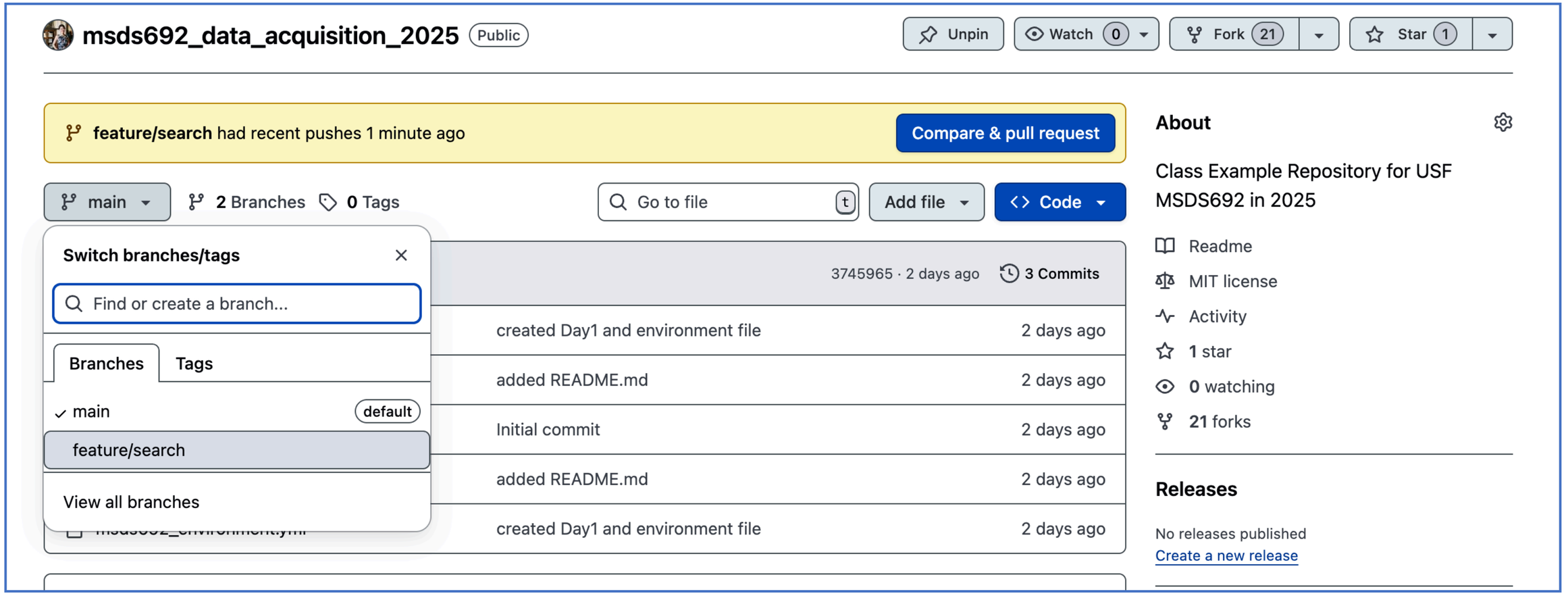
Remote branch

- So far, we talked about a local branch.
- If you'd want to track commits on a remote repository (Ex. GitHub), you'd need a remote branch.
- To create and push changes to a remote, use -u (set upstream) option with a remote name (typically origin).
  - `$ git push -u origin branch_name`
  - After that you can just push.

```
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git checkout feature/search
Switched to branch 'feature/search'
(msds692) dwoodbridge@ML-ITS-210588 msds692_data_acquisition_2025 % git push -u origin feature/search
Enumerating objects: 5, done.
Counting objects: 100% (5/5), done.
Delta compression using up to 10 threads
Compressing objects: 100% (2/2), done.
Writing objects: 100% (4/4), 331 bytes | 331.00 KiB/s, done.
Total 4 (delta 1), reused 0 (delta 0), pack-reused 0
```

Git Branch

Remote branch



Git Branch

Delete a branch

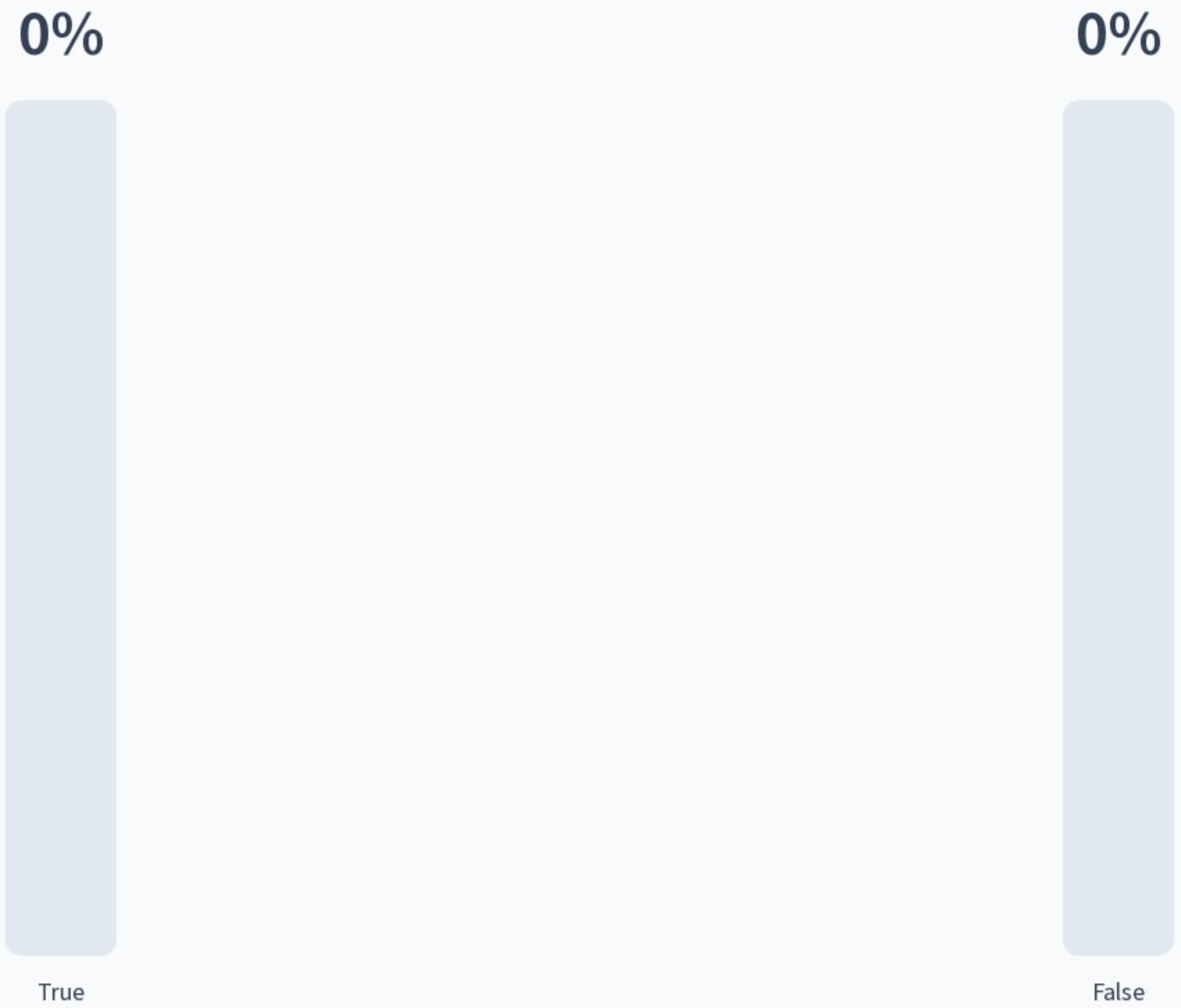
- If you no longer need a branch, you can delete a branch
- Delete a local branch
  - `$ git branch -d branch_name`
- Delete a remote branch
  - `$ git push origin --delete  branch_name`

If you create a local branch, it automatically creates a remote branch

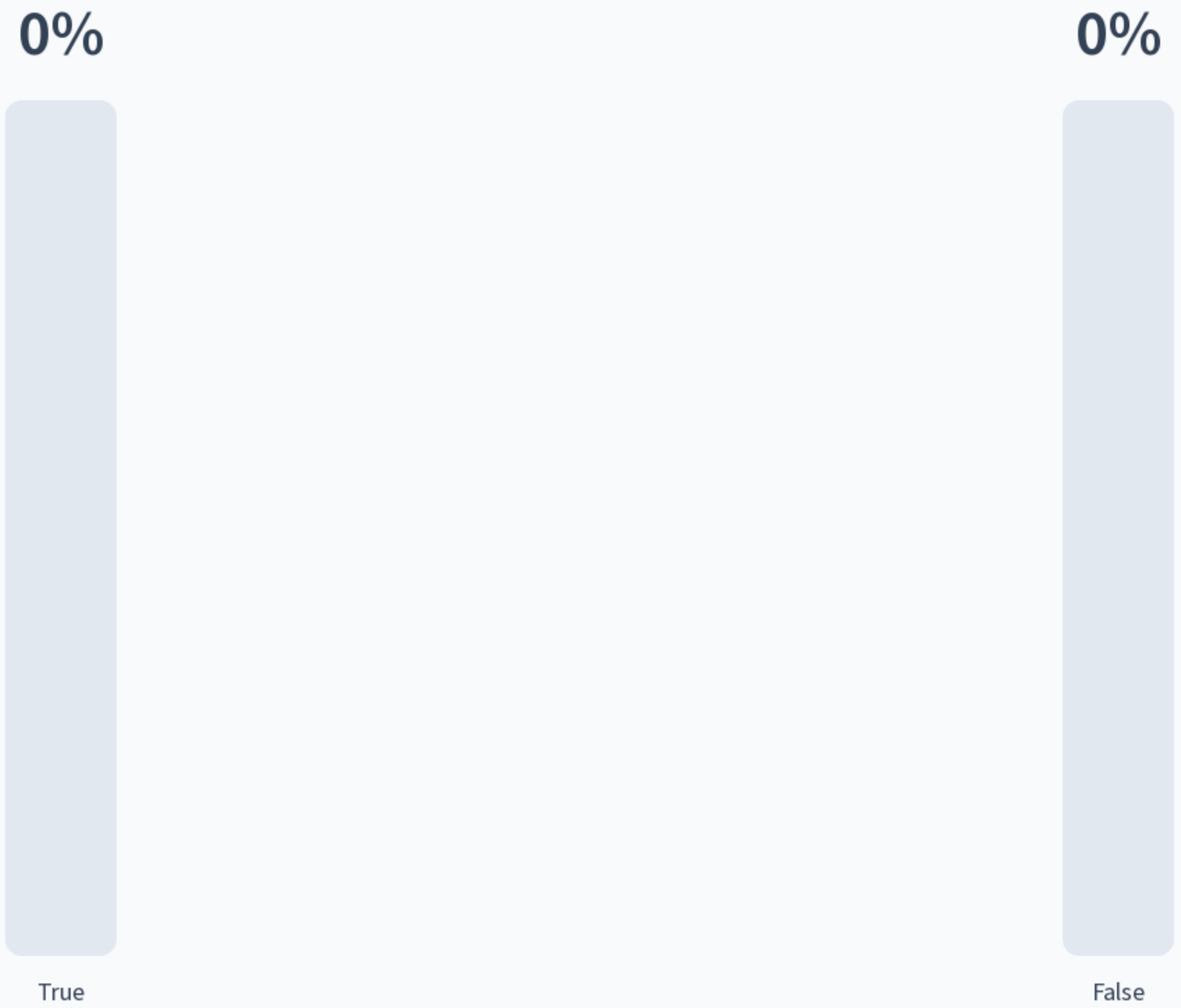
True

False

If you create a local branch, it automatically creates a remote branch



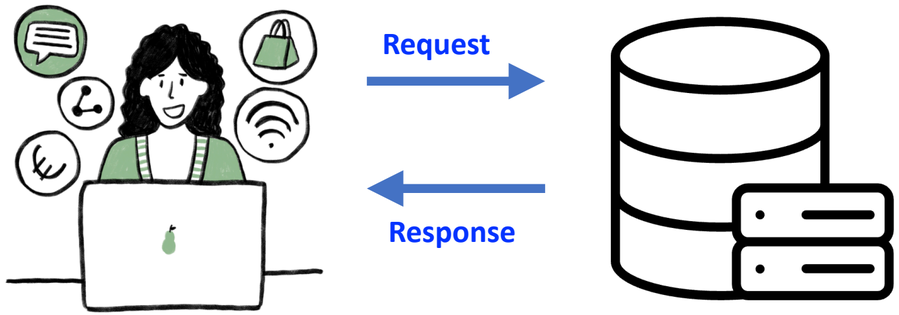
If you create a local branch, it automatically creates a remote branch



Contents

HTTP Requests and Responses

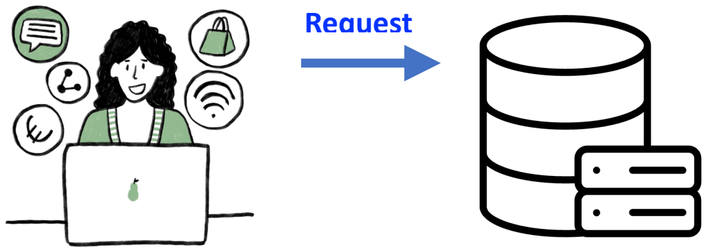
- HTTP (Hypertext Transfer Protocol) specifies various request methods to indicate the desired action to be performed on a given resource.
- It is a request-response model, where client (your browser or python script) sends a request and server processes it and sends a response back.



HTTP Requests and Responses

- **HTTP Requests**
  - Request Line : Define the action that you want to take
    - HTTP Methods: Types of action to take
      - GET(retrieve data), POST(send data), PUT (update data), DELETE (remove data), etc.
    - URI : Absolute or relative URL indicating the resource's location
    - HTTP Version : Protocol versions being used
    - Ex. GET /index.html HTTP/1.1
  - Headers : Metadata (extra info) about the request.
    - It may include Host (address), User-Agent (Info about your browser/app), Content-Type (format of the data you're sending (e.g., JSON, HTML, plain text)), Authorization (API Keys, Token), Cookie (For tracking session/user information previously set by a server)
  - Body (Optional) : For some HTTP methods including POST and PUT, a request may include a body containing data to be delivered to the server.

```
POST / HTTP/1.1
Host: developer.mozilla.org
User-Agent: curl/8.6.0
Accept: */*
Content-Type: application/json
Content-Length: 345

{
"data": "ABC123"
}
```

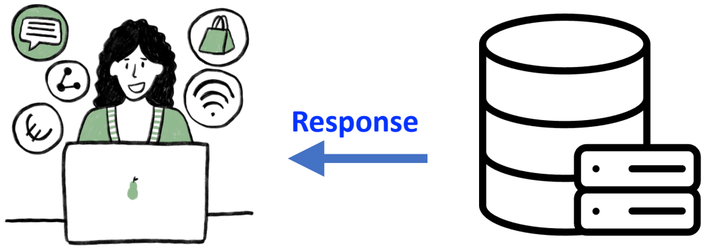


https://developer.mozilla.org/en-US/docs/Web/HTTP/Guides/Messages

HTTP Requests and Responses

- **HTTP Responses**
  - Status Line : HTTP Version, a three-digit status code and brief text message about the status
    - Status code : 2xx (Success, 200:OK), 4xx (Client Error, 404: Not Found, 401: Unauthorized), 5xx (Server Error, 500: Internal Server Error, 503 : Service Unavailable)
    - Ex. HTTP/1.1 200 OK
  - Headers : Metadata (extra info) about the response including content type, server, caching directives, etc.
  - Body : The actual content/data

```
HTTP/1.1 403 Forbidden
Server: Apache
Date: Fri, 21 Jun 2024 12:52:39 GMT
Content-Length: 678
Content-Type: text/html
Cache-Control: no-store

<!DOCTYPE html>
<html lang="en">
(more data…)
```



https://developer.mozilla.org/en-US/docs/Web/HTTP/Guides/Messages

HTTP Requests and Responses

Python's `requests`

- The most popular Python library for HTTP.
- It makes it simple to send requests (GET, POST, etc.) and handle responses.
  - Ex.

```
import requests
response = requests.get("https://example.com")
```

https://requests.readthedocs.io/en/latest/user/quickstart/#make-a-request

HTTP Requests and Responses

Python's `requests`

- Get
  - Syntax :

```
r = requests.get(url, header, params)
```

    - url: Resource's location
    - header : Optional
    - params : If you'd need to pass a parameter to the url (ex. url/something?key=val), you can set params in a dictionary format.
    - r: response
      - r.content : returns the response's content in bytes
      - r.headers : returns the response's headers

HTTP Requests and Responses

Python's requests

- Get
  - Example

```
url = "https://www.google.com/search"
payload = {"q":"USF Data Science"}
headers = ''

r = requests.get(url, params=payload, headers=headers)
print(r.url)
print(r.content)
```

https://requests.readthedocs.io/en/latest/user/quickstart/#make-a-request

HTTP Requests and Responses

Python's `requests`

- Post
  - Syntax :

```
r = requests.get(url, header, data)
```

    - url: Resource's location
    - header: Optional
    - data: A dictionary, list of tuples, bytes or a file object to send to the url
    - r: response
      - r.content : returns the response's content in bytes
      - r.headers : returns the response's headers

HTTP Requests and Responses

Python's requests

- Post
  - Example

```
url = 'https://www.w3schools.com/action_page.php'
data = {'key': 'value'}
header = ''

r = requests.post(url, data=data, headers=header)
print(r.headers)
print(r.content)
```

https://requests.readthedocs.io/en/latest/user/quickstart/#make-a-request

Example 3

For the given URL that host a data file, send a http request to retrieve the data.

- Can you guess the file type?

```
import requests

r = requests.get(url)
print(r.headers)
print(r.content)
```

```
{'x-amz-id-2':
'vuZEoyrK9Cj9v0gC1+rVbwPqGuE7gD8wF8YnudNK4fja+tibKWlaOxNvOFe42kOJXNT
PZ7tY570=', 'x-amz-request-id': 'H40V74XDJAB4TFVF', 'Date': 'Sun, 24
Aug 2025 22:42:27 GMT', 'Last-Modified': 'Fri, 01 Nov 2024 00:18:53
GMT', 'ETag': '"d70cd1b77fcc6fe6146f713e496a788b"', 'x-amz-server-
side-encryption': 'AES256', 'Accept-Ranges': 'bytes', 'Content-
Type': 'binary/octet-stream', 'Content-Length': '197639', 'Server':
'AmazonS3'}
b'linkUrl,customEvent:DATAGOV_dataset_organization,customEvent:DATAG
OV_dataset_publisher,fileExtension,fileName,eventCount\r\nhttps://
data.chhs.ca.gov/dataset/5a281abf-1730-43b0-b17b-ac6a35db5760/
resource/724c6fd8-a645-4e52-b6,State of California,ca-gov,csv,adult-
depression-lghc-indicator-24.csv,777\r\nhttps://data.chhs.ca.gov/
dataset/596b5eed-31de-4fd8-a645-249f3f9b19c4/resource/
57da6c9a-41a7-44b0-ab,State of California,ca-
gov,csv,cscpopendata.csv,690\r\nhttps://data.chhs.ca.gov/dataset/
bb703230-1f5f-44b5-8a90-55e45e08c452/resource/4a8dde27-
c4e1-4ca3-86,State of California,ca-gov,csv,sugar-sweetened-
beverage-consumption-in-california-
residents-20122013.csv,531\r\nhttps://data.chhs.ca.gov/dataset/
99bc1fea-c55c-4377-bad8-f00832fd195d/resource/bc09f211-200c-4c4c-
aa,State of California,ca-
gov,xlsx,hci_crime_752_pl_co_re_ca_2000-2013_21oct15-
ada.xlsx,398\r\n … }
```

Example 3

For the given URL that host a data file, send a http request to retrieve the data.

- `io.BytesIO(binary_data)`
  - Convert binary data directly in the program's memory
  - Useful for handling binary data to work as a Python object.


```
{'x-amz-id-2':
'vuZEoyrK9Cj9v0gC1+rVbwPqGuE7gD8wF8YnudNK4fja+tibKWlaOxNvOFe42kOJXNT
PZ7tY570=', 'x-amz-request-id': 'H40V74XDJAB4TFVF', 'Date': 'Sun, 24
Aug 2025 22:42:27 GMT', 'Last-Modified': 'Fri, 01 Nov 2024 00:18:53
GMT', 'ETag': '"d70cd1b77fcc6fe6146f713e496a788b"', 'x-amz-server-
side-encryption': 'AES256', 'Accept-Ranges': 'bytes', 'Content-
Type': 'binary/octet-stream', 'Content-Length': '197639', 'Server':
'AmazonS3'}
b'linkUrl,customEvent:DATAGOV_dataset_organization,customEvent:DATAG
OV_dataset_publisher,fileExtension,fileName,eventCount\r\nhttps://
data.chhs.ca.gov/dataset/5a281abf-1730-43b0-b17b-ac6a35db5760/
resource/724c6fd8-a645-4e52-b6,State of California,ca-gov,csv,adult-
depression-lghc-indicator-24.csv,777\r\nhttps://data.chhs.ca.gov/
dataset/596b5eed-31de-4fd8-a645-249f3f9b19c4/resource/
57da6c9a-41a7-44b0-ab,State of California,ca-
gov,csv,cscpopendata.csv,690\r\nhttps://data.chhs.ca.gov/dataset/
bb703230-1f5f-44b5-8a90-55e45e08c452/resource/4a8dde27-
c4e1-4ca3-86,State of California,ca-gov,csv,sugar-sweetened-
beverage-consumption-in-california-
residents-20122013.csv,531\r\nhttps://data.chhs.ca.gov/dataset/
99bc1fea-c55c-4377-bad8-f00832fd195d/resource/bc09f211-200c-4c4c-
aa,State of California,ca-
gov,xlsx,hci_crime_752_pl_co_re_ca_2000-2013_21oct15-
ada.xlsx,398\r\n … }
```

For receiving a text file from a url, which HTTP request should be used?

GET

POST

PUT

DELETE

For receiving a text file from a url, which HTTP request should be used?

GET

0%

POST

0%

PUT

0%

DELETE

0%

For receiving a text file from a url, which HTTP request should be used?

GET

0%

POST

0%

PUT

0%

DELETE

0%

Contents

File Acquisition

So far, we handled several file formats including .csv/.tsv, .txt, .json, .pickle, etc.

- We will cover other common file types including .pdf, .docx, and .xlsx.
  - .pdf: pypdf
  - .docx: python-docx
  - .xlsx: pandas.read_excel

About Data

- Microplastic - https://catalog.data.gov/dataset/microplastic-and-microparticle-data-from-surface-water-san-francisco-bay-and-adjacent-sanc

- Fishery- https://www.fisheries.noaa.gov/inport/item/30650

Potential Project - Analysis on microplastics in fishery products and potential risks.

File Acquisition

PDF

- **pypdf** allows reading text, extracting images/metadata, splitting/ merging, and writing PDFs.

- Extracting texts

```
from pypdf import PdfReader

reader = PdfReader(file_path) # Initialize a PdfReader object.

for page in reader.pages:  # Emulates a list of PageObject.
  page.extract_text() # Extract texts
```

https://pypdf.readthedocs.io/en/stable/index.html

Example 4

From url, return a list of dictionary of `{page_number: text}` all the pages including `anchovies`

[{8: 'Executive Summary \n viii “margins” of the Bay, in open portions of the Bay, and in a less urban reference area (Tomales Bay). Microparticles were identified in sediment from all 20 sites. Fibers, followed by fragments, were the most abundant type of microparticles in Bay sediment, with detected concentrations ranging between 1 and 49 microfibers per gram dry weight (dw), and between 0.1 and 11 non-fiber microparticles (including fragments, films, spheres, and foams) per gram dw. The highest concentrations of microparticles were measured in Lower South Bay, which is strongly influenced by wastewater and urban stormwater discharges. Concentrations at the reference site, Tomales Bay, were among the lowest observed in the study. Black fragments that had a rubbery texture were frequently detected in sediment samples. Spectroscopy was unonclusively as plastic.'},
 {214: 'Chapter 6 – Prey Fish \n 201 Highlights ¨ Northern anchovy (Engraulis mordax) and topsmelt (Atherinops affinis) were collected at multiple sites in San Francisco Bay and at a reference area with minimal urban influence. Fish guts were digested whole in potassium hydroxide (KOH), filtered through a 10 µm polycarbonate filter, and analyzed for microparticles and microplastics down to 20 µm. ¨ These two prey confirmed to be plastic, while 60% were classified as anthropogenic unknown because dyes embedded in the microfibers interfered with the laboratory’s ability to identify the composition. Twenty-one percent of non-fiber microparticles analyzed by spectroscopy were confirmed to be plastic. ¨ Particles smaller than 150 µm represented 16% of fragments and 6% of fibers observed in fish samples. Particles in this size fraction have the potential to translocate out of the gut and bioaccumulate.'},
 {215: 'Chapter 6 – Prey Fish \n 202 Objectives The goal of this element of the San Francisco Bay Microplastics Project was to characterize microparticles and microplastics in the digestive tracts of prey fish collected in and around San Francisco Bay. Microplastics are a subset of microparticles and microfibers that have been definitively determined as plastic through spectroscopy or other means. Many studies in the literature identify microparticles that appear to be plastic using only visual techniques, such as microscopy. Prey fish serve as they age. Anchovies and topsmelt also have different feeding habits, providing an opportunity to test the hypothesis that fish at the same trophic level with different foraging strategies ingest different amounts and types of microplastics.‘},….]

Example 2

From file_path, return a list of dictionary of `{page_number: text}` all the pages including `anchovies`

```
response = requests.get(url)
pdf_data = BytesIO(response.content)
reader = PdfReader(pdf_data)

output = []
page_no = 1
for page in reader.pages: # Emulates a list of PageObject.
  page_text = page.extract_text()
  if 'anchovies' in page_text.lower():
     output.append({page_no : page_text}) # Extract texts
  page_no += 1
```

File Acquisition

DOCX

- **python-docx** allows reading/writing/editing Word files programmatically.

```
from docx import Document

# Document() constructor, which returns DocumentObject (This takes str, bytesIO, etc.)
document = Document(file_path)

# For texts
for paragraph in document.paragraphs:
    print(pragraph.text)

# For tables
tables = []
for table in document.tables:
    output = []
    for row in table.rows:
        row_content = [cell.text for cell in row.cells]  # Collect cell content in a list
        output.append(row_content)
    tables.append(output)
```

Example 5

Load from a .docx located at `url`, read text paragraphs and tables.

```
[[['#', 'Variable', 'Type', 'Len'],
 ['1', 'DATE_GMT', 'Num', '8'],
 ['2', 'TIME_GMT', 'Num', '8'],
 ['3', 'CRUISE', 'Num', '8'],
 ['4', 'VESSEL', 'Num', '8'],
 ['5', 'STATION', 'Char', '15'],
 ['6', 'LAT', 'Num', '8'],
 ['7', 'LON', 'Num', '8'],
 ['8', 'TOD', 'Char', '15'],
 ['9', 'SEASON', 'Char', '150']]…]
```

Example 5

Load from a .docx located at `url`, read text paragraphs and tables.

```
response = requests.get(url)
data = BytesIO(response.content)
document = Document(data)

# For text
text_output = []
for paragraph in document.paragraphs:
    text_output.append(paragraph.text)

# For tables
for table in document.tables:
    output = []
    for row in table.rows:
        row_content = [cell.text for cell in row.cells]  # Collect cell content in a list
        output.append(row_content)
    tables += output
```

File Acquisition

XLSX

- While you can read .xlsx using `openpyxl`, we will cover `pandas`' `read_excel`. (Note - read_excel uses openpyxl)
- **`pandas.read_excel`**

`pandas.read_excel(`*`io, sheet_name=0, header=0, names=None,…`*`)`

- **`io`** : string (path, url), bytes, file object.
- **`sheet_name`** : strings(sheet names), integer(zero-indexed sheet positions), list of strings and/or integers for multiple sheets, None for all worksheets.
- **`header`** : int (the row # to be used for column names)
- **`names`** : List of column names to use. If file contains no header row, then you should explicitly pass header=None.

Example 6

Read .xlsx's `fish` sheet from the URL.

	SampleID	StationType_Final	Count	StationCode	SampleDate	Latitude	Longitude	Datum	ProjectCode	EventCode	...	PartsComments	CompositeID	CompositeType	CompositeReplicate	CompositeWeight
0	17MMP-AN-CB010-MP_1	Field	33	CB10	2017-07-05	37.906718	-122.346692	WGS84	17-18_MP_Moore	TI	...	NaN	17MMP-AN-CB010-MP_1	Normal	1	-88
1	17MMP-AN-CB010-MP_10	Field	11	CB10	2017-07-05	37.906718	-122.346692	WGS84	17-18_MP_Moore	TI	...	NaN	17MMP-AN-CB010-MP_10	Normal	1	-88
2	17MMP-AN-CB010-MP_2	Field	21	CB10	2017-07-05	37.906718	-122.346692	WGS84	17-18_MP_Moore	TI	...	NaN	17MMP-AN-CB010-MP_2	Normal	1	-88
3	17MMP-AN-CB010-MP_3	Field	18	CB10	2017-07-05	37.906718	-122.346692	WGS84	17-18_MP_Moore	TI	...	NaN	17MMP-AN-CB010-MP_3	Normal	1	-88
4	17MMP-AN-CB010-MP_4	Field	33	CB10	2017-07-05	37.906718	-122.346692	WGS84	17-18_MP_Moore	TI	...	NaN	17MMP-AN-CB010-MP_4	Normal	1	-88
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
162	Blank 29-Dec-17	LABQA	8	LABQA	1950-01-01	NaN	NaN	NaN	17-18_MP_Moore	TI	...	NaN	Blank 29-Dec-17	LABQA	1	-88
163	Blank 30-Dec-17	LABQA	2	LABQA	1950-01-01	NaN	NaN	NaN	17-18_MP_Moore	TI	...	NaN	Blank 30-Dec-17	LABQA	1	-88
164	Blank 4-Dec-17	LABQA	3	LABQA	1950-01-01	NaN	NaN	NaN	17-18_MP_Moore	TI	...	NaN	Blank 4-Dec-17	LABQA	1	-88
165	Blank 5-Dec-17	LABQA	6	LABQA	1950-01-01	NaN	NaN	NaN	17-18_MP_Moore	TI	...	NaN	Blank 5-Dec-17	LABQA	1	-88

Example 6

Read .xlsx's `fish` sheet from the URL.

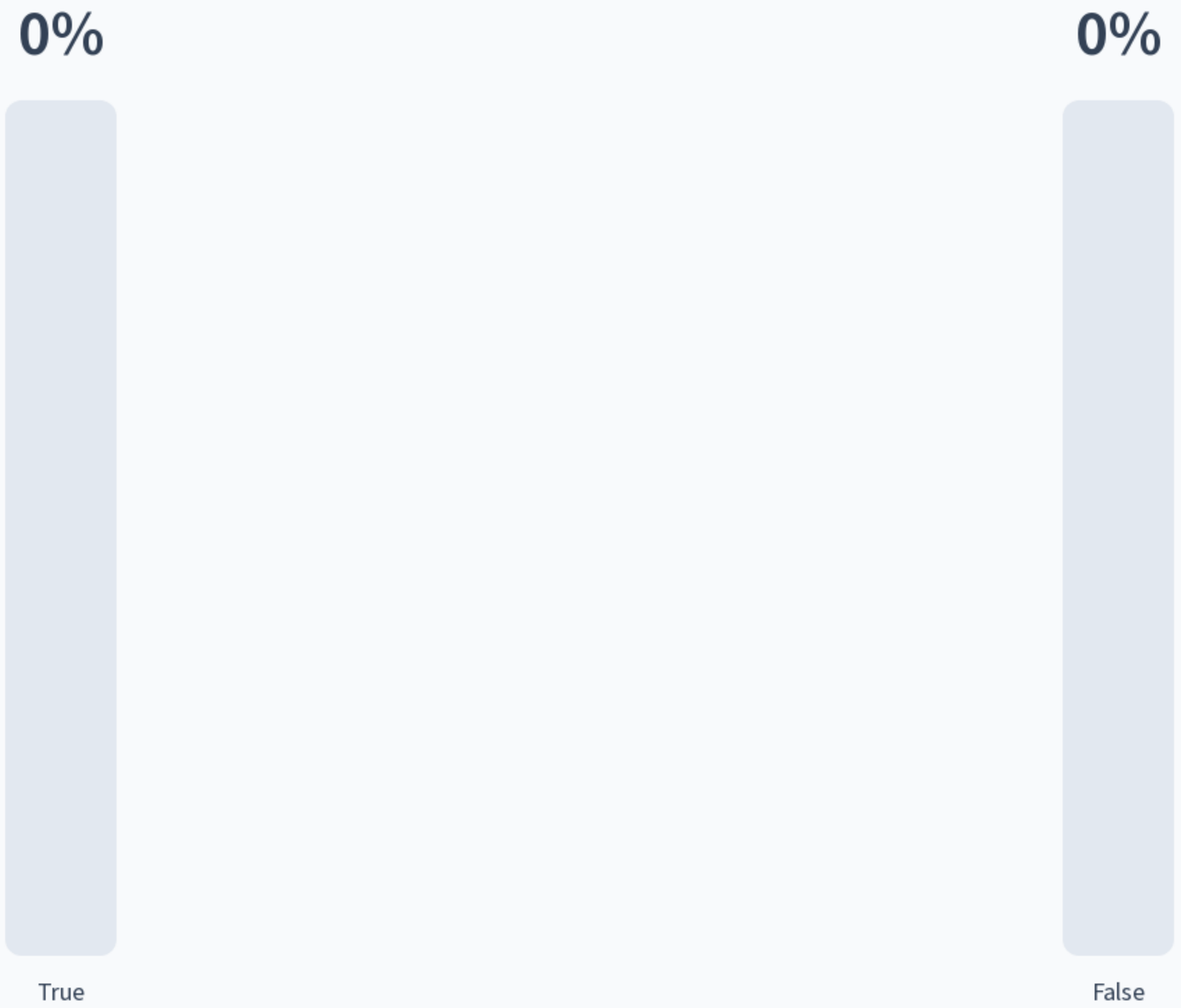
```
response= requests.get(url)
data = io.BytesIO(response.content)
pd.read_excel(data, sheet_name="fish")
```

You can retrieve by the page numbers using python-docx

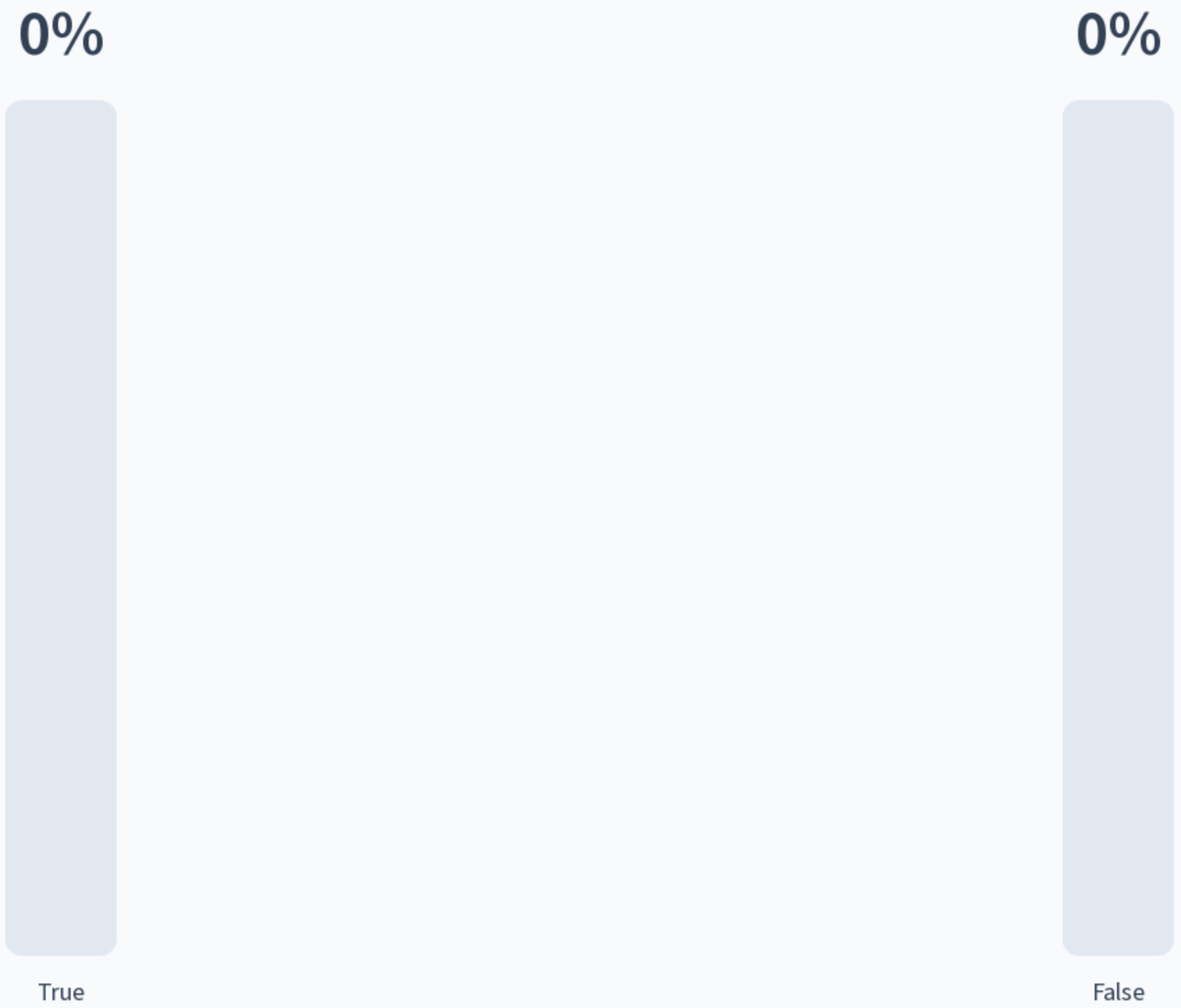
True

False

You can retrieve by the page numbers using python-docx



You can retrieve by the page numbers using python-docx



Contents

Streamlit

Streamlit is a Python framework for building interactive web apps for data science

- Runs in the browser, but you code in Python only (no HTML/JS needed).
  - By default it runs on port 8501.
- Great for quick dashboards, data visualization, and demos.
- You can run a streamlit application by
  - `$ streamlit run app_name.py`

Streamlit

Some basic APIs

- There will be more APIs that you should use for assignments, and I added URLs that you can refer. Also, we will cover more APIs and examples as the course progresses.

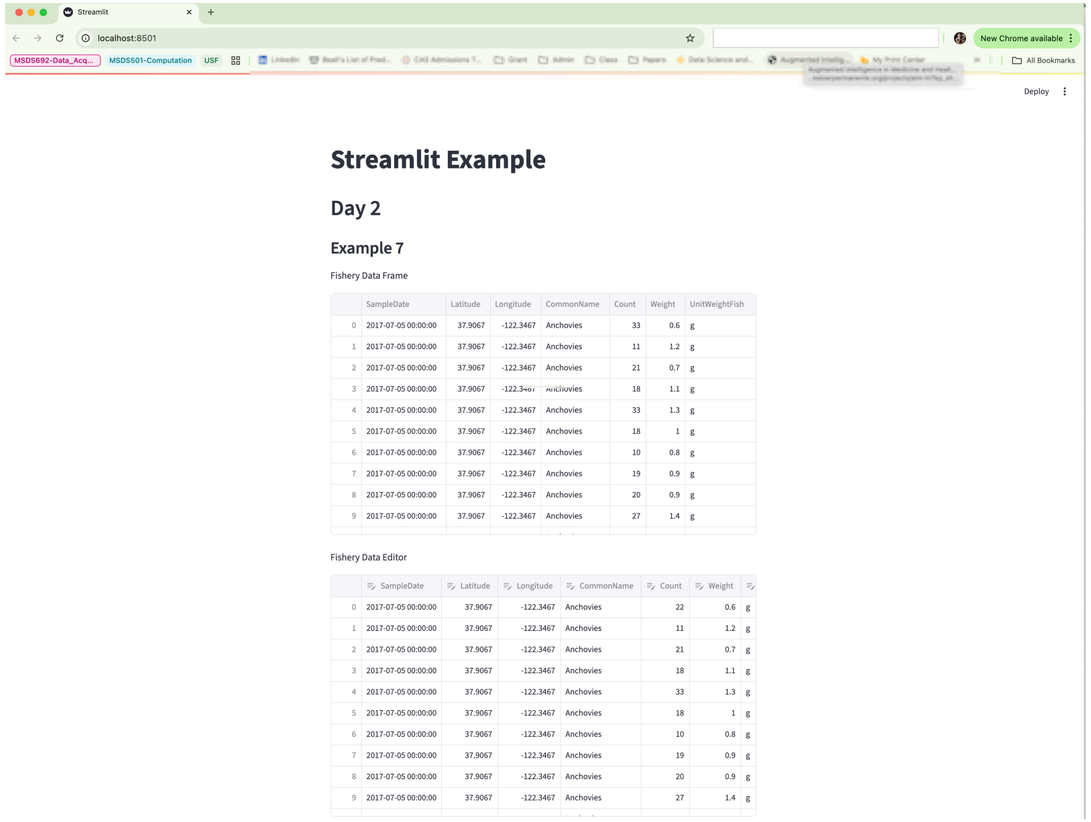
```
import streamlit as st
```

  - Text
    - `st.title(text), st.header(text), st.subheader(text)`
  - Data
    - `st.dataframe(pandas_df), st.data_editor(pandas_df)`
      - You can store updated data via

```
edited_df = st.data_editor(df)
edited_df.to_excel(path)
```

Example 7

Update and run the given ex07.py, which is a streamlit application.



	SampleDate	Latitude	Longitude	CommonName	Count	Weight	UnitWeightFish
0	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	33	0.6	g
1	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	11	1.2	g
2	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	21	0.7	g
3	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	18	1.1	g
4	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	33	1.3	g
5	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	18	1	g
6	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	10	0.8	g
7	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	19	0.9	g
8	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	20	0.9	g
9	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	27	1.4	g

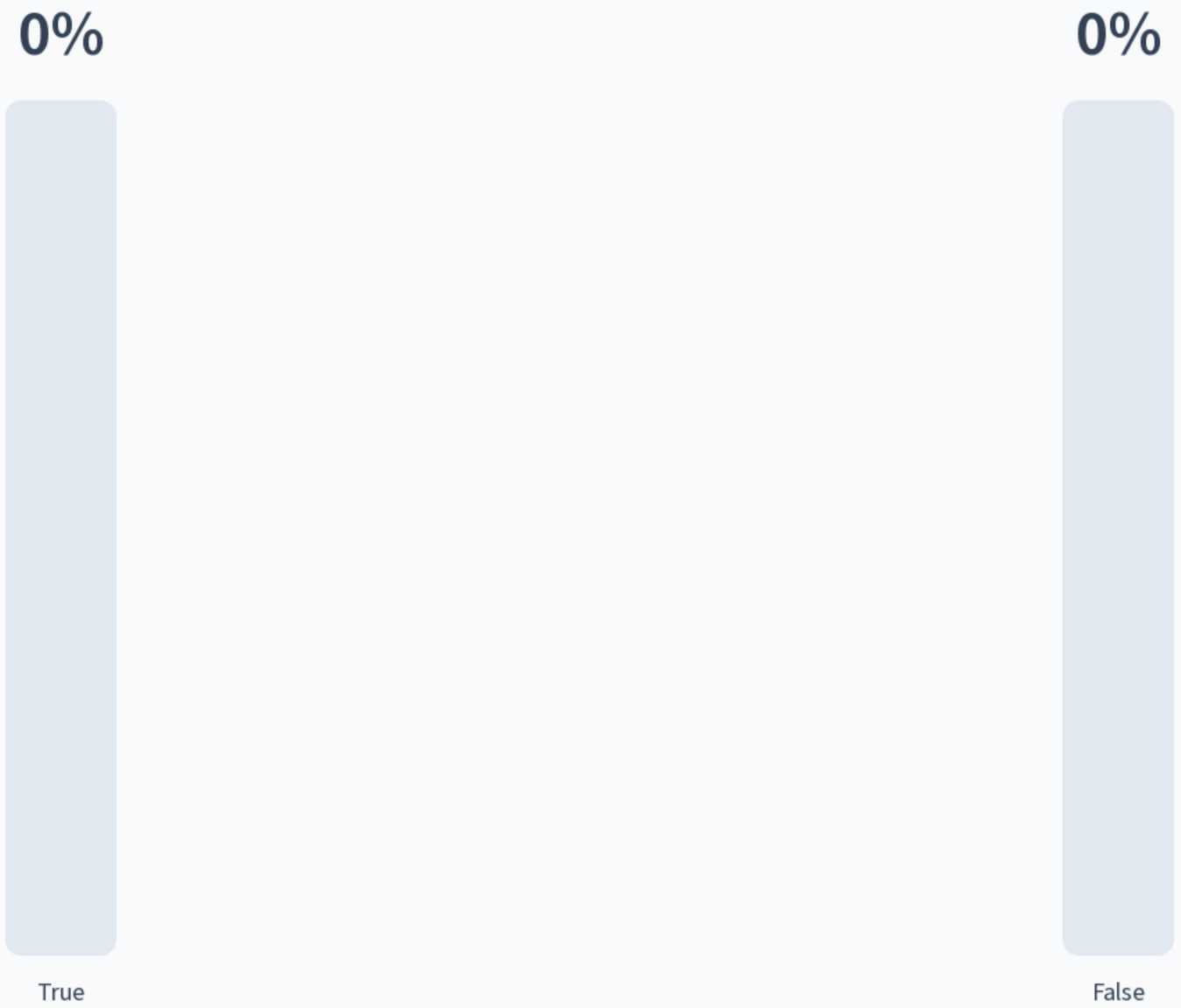
	SampleDate	Latitude	Longitude	CommonName	Count	Weight	
0	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	22	0.6	g
1	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	11	1.2	g
2	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	21	0.7	g
3	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	18	1.1	g
4	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	33	1.3	g
5	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	18	1	g
6	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	10	0.8	g
7	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	19	0.9	g
8	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	20	0.9	g
9	2017-07-05 00:00:00	37.9067	-122.3467	Anchovies	27	1.4	g

You can start a streamlit application by $ python app_name.py

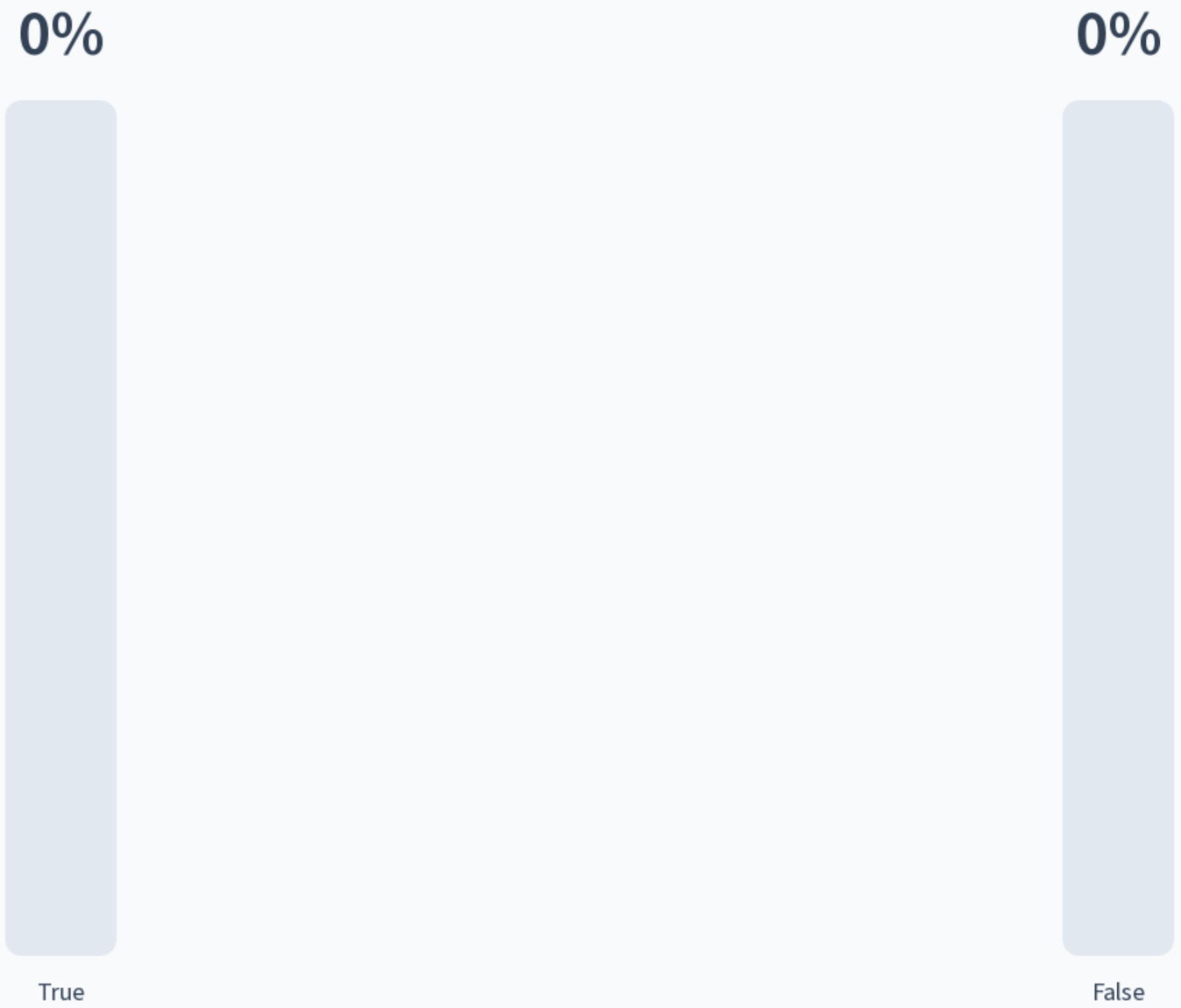
True

False

You can start a streamlit application by $ python app_name.py



You can start a streamlit application by $ python app_name.py



Contents

Your feedback is important and will help shape the rest of the course. What aspects of the course have been most helpful? Do you have suggestions for changes that could make the course more engaging or effective for you?

Nobody has responded yet.

Hang tight! Responses are coming in.

References

Git Branching, https://git-scm.com/book/en/v2/Git-Branching-Branches-in-a-Nutshell

HTTP Messages, https://developer.mozilla.org/en-US/docs/Web/HTTP/Guides/Messages

Requests: HTTP for Humans, https://requests.readthedocs.io/en/latest/

Pypdf, https://pypdf.readthedocs.io/en/stable/index.html

Python-docx, https://python-docx.readthedocs.io/en/latest/

pandas.read_excel, https://pandas.pydata.org/docs/reference/api/pandas.read_excel.html

Streamlit, https://streamlit.io/