Data science for everyone

Prof. Jones-Rooy & Prof. Policastro

Feb. 5, 2020

2.2: Course Technology & More Causality!

ANNOUNCEMENTS

1. Homework 1 out now; due Tue., Feb. 18, 8p
2. Lab 0 out tonight; due Wed., Feb. 12, 8p
3. Sections this week: Practice assignment completion & submission; Lab 0
   - Bring laptops to section
   - You should never miss section but this week you should **really really really (etc.)** not miss section

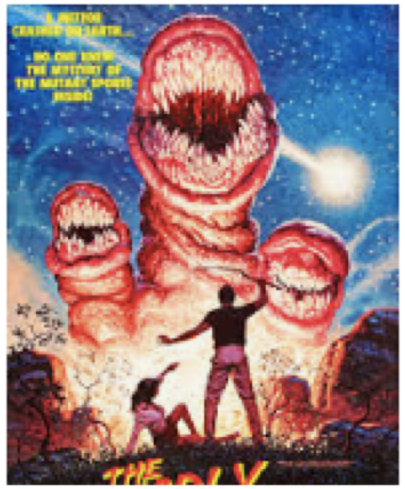
Outline

1. How to complete & submit all assignments in this course
2. Key concepts in causality
3. Principles & types of causality

HOW TO COMPLETE AND SUBMIT ASSIGNMENTS FOR THIS COURSE

1. Fetch assignment on JupyterHub
2. Complete assignment in that notebook on JupyterHub
3. Submit assignment on JupyterHub (NB-grader)
4. Submit assignment on Gradescope

1. Access the assignment by fetching it under the Assignments tab on JupyterHub
2. Complete the assignment in the spaces provided in the same document you accessed
3. Run tests within the assignment to check your formatting
4. Optional but recommended: **Validate** the entire assignment to again check formatting and ensure you haven't missed any questions
    1. You do not need to validate to submit, but it's a way to make sure you don't lose points for silly mistakes
5. **Submit** the assignment on JupyterHub
6. Go back into the assignment, download it as html, turn the html into a PDF
7. Upload the PDF to **Gradescope**, assign questions to pages, it will also automatically validate when you try to submit
8. Submit the assignment on **Gradescope**



The Deadly Spawn (198...
imdb.com

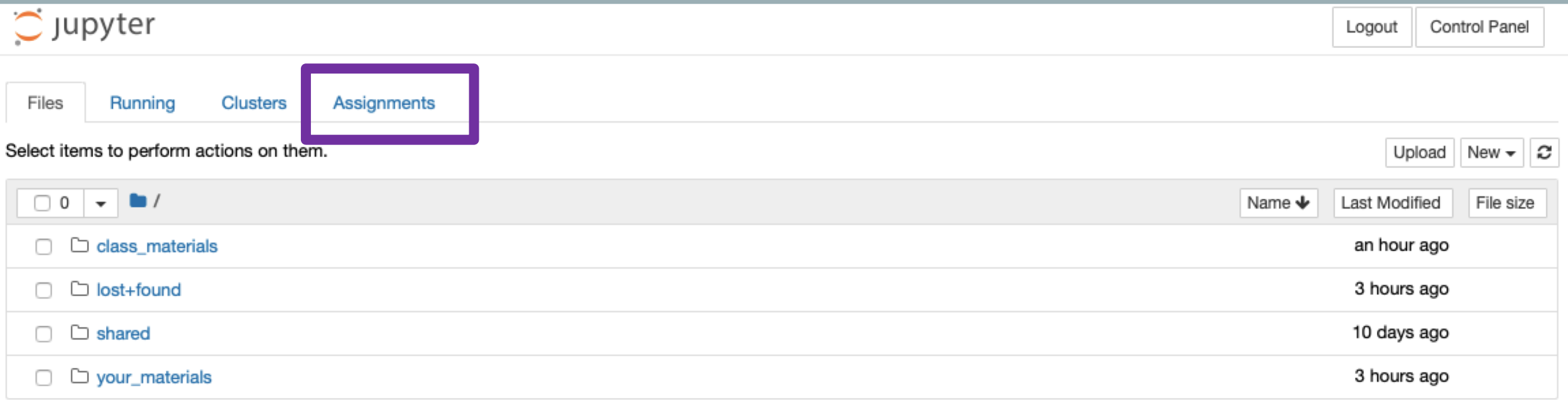
1. FETCH THE ASSIGNMENT

- All assignments (HWs, labs, project) will be completed in Jupyter notebooks on JupyterHub
- You can access JupyterHub through the tab on our Classes site or by going directly to: https://dsua-111.rcnyu.org/ (click SPAWN)
- To get the assignments, go to Assignments, then **fetch**.
  - If you want to ever re-fetch (e.g., you're working on it and want to start over with a clean slate), just delete the assignment under **Files** and then re-fetch
- All assignments will be handed out on JupyterHub, and you can complete them directly INSIDE the assignment we give you (there will be space for answers within the document) – so everything happens in the same place
  - The questions are all Read Only so you cannot accidentally delete one
- You're welcome to make a copy of the assignment to try things out separately before filling out the final version, or just creating your own file to use as scratch paper (I do this a lot!)
- But, you must submit the version we handed out on the **Assignments tab**. Ultimately, that's where you'll submit assignments, too

JupyterHub

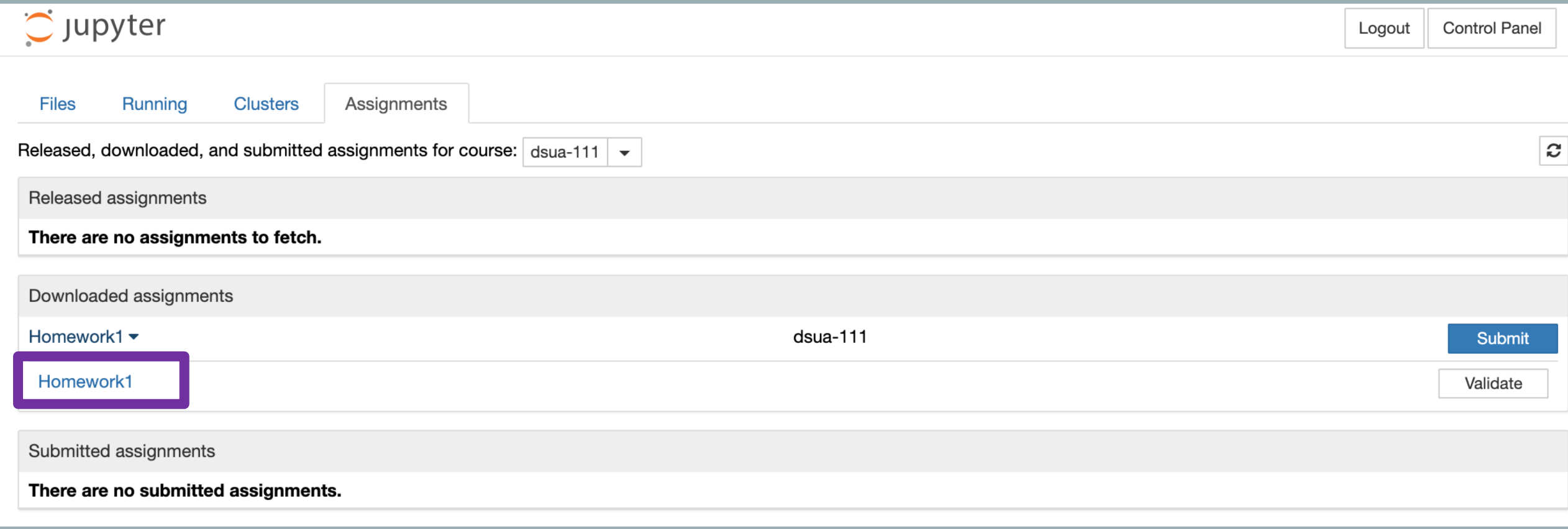
https://dsua-111.rcnyu.org

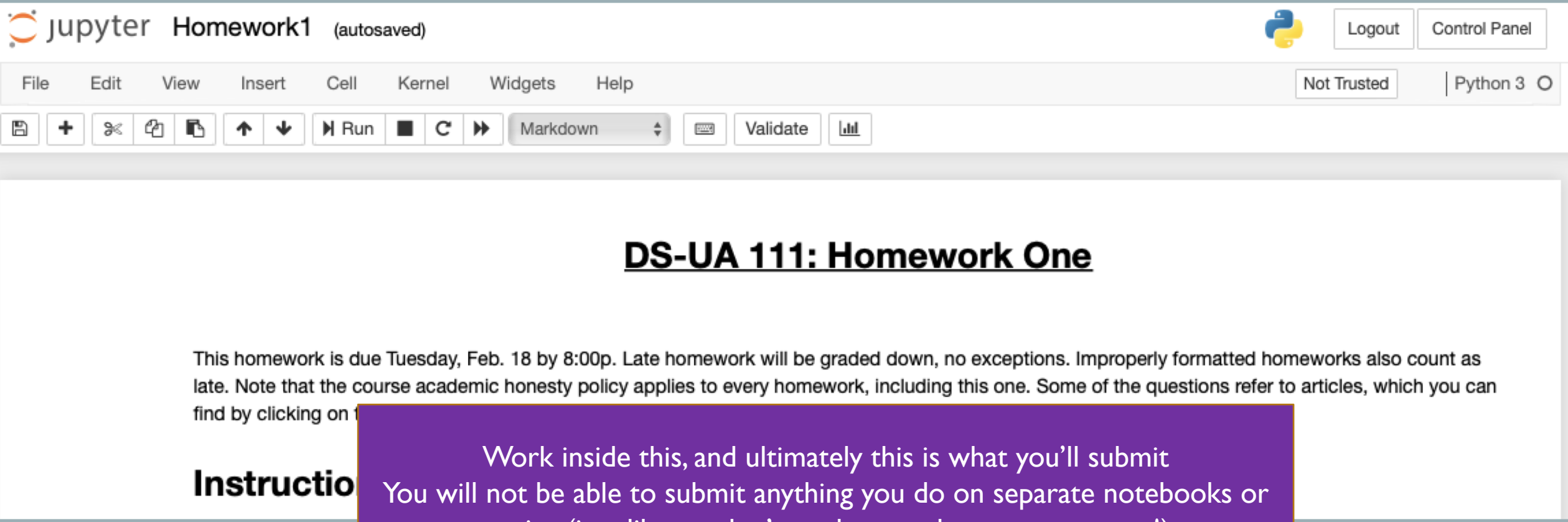
Where to find all assignments for this course




jupyter
Logout
Control Panel
Files
Running
Clusters
Assignments
Released, downloaded, and submitted assignments for course:
dsua-111
Released assignments
Homework1
dsua-111
Fetch
Downloaded assignments
There are no downloaded assignments.
Submitted assignments
There are no submitted assignments.

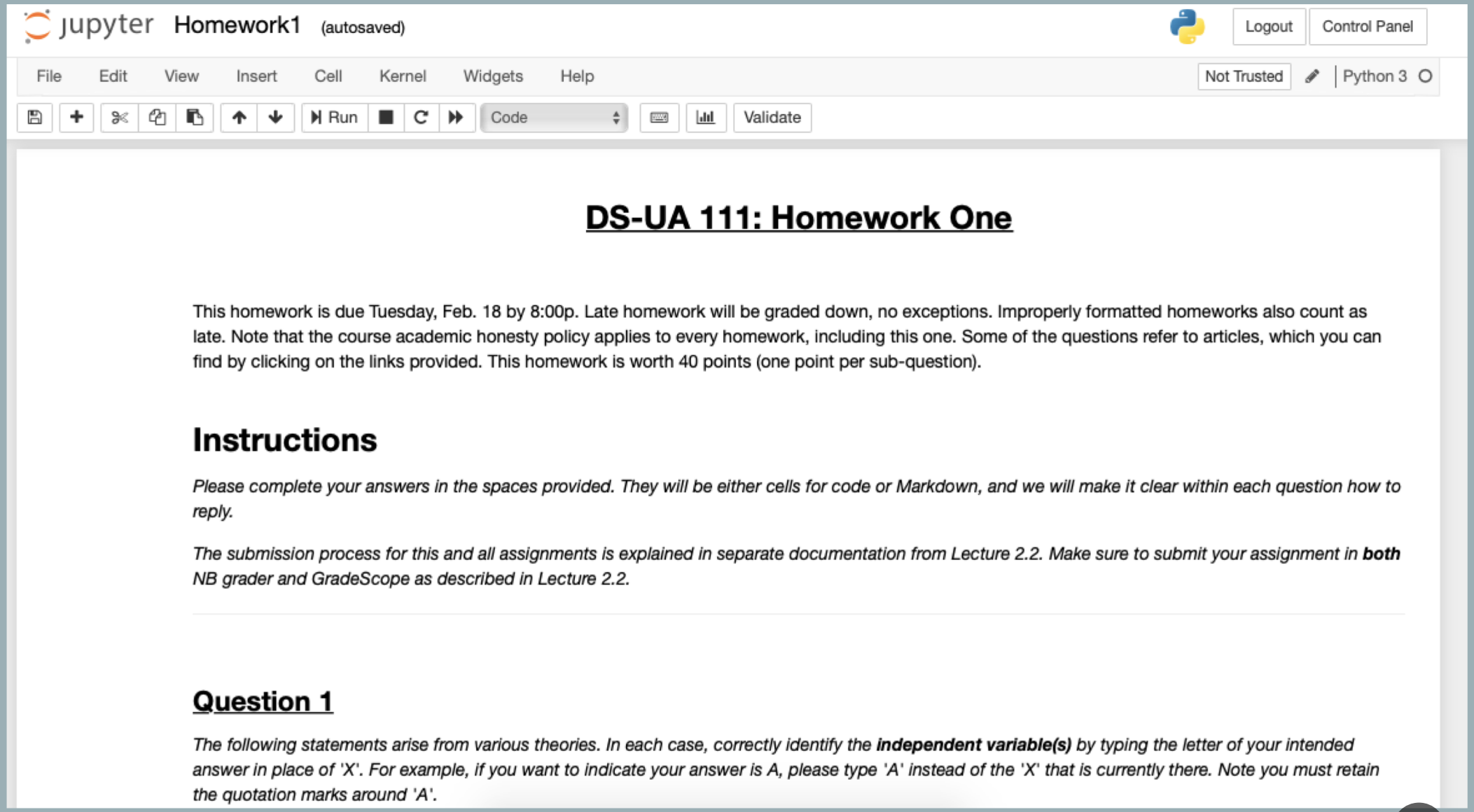

jupyter
Logout
Control Panel
Files
Running
Clusters
Assignments
Released, downloaded, and submitted assignments for course:
dsua-111
Released assignments
There are no assignments to fetch.
Downloaded assignments
Homework1
dsua-111
Submit
Submitted assignments
There are no submitted assignments.


jupyter
Logout
Control Panel
Files
Running
Clusters
Assignments
Released, downloaded, and submitted assignments for course: dsua-111
Released assignments
There are no assignments to fetch.
Downloaded assignments
Homework1
dsua-111
Submit
Homework1
Validate
Submitted assignments
There are no submitted assignments.


jupyter Homework1 (autosaved)
Logout
Control Panel
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Markdown
Validate
DS-UA 111: Homework One
This homework is due Tuesday, Feb. 18 by 8:00p. Late homework will be graded down, no exceptions. Improperly formatted homeworks also count as late. Note that the course academic honesty policy applies to every homework, including this one. Some of the questions refer to articles, which you can find by clicking on t
Instructio
Work inside this, and ultimately this is what you'll submit
You will not be able to submit anything you do on separate notebooks or copies (just like we don't grade scratch paper on exams!)

2. COMPLETE THE ASSIGNMENT

- The cells where your answers should be typed are clearly marked and will always be one of two possible types: **code** or **Markdown**
  - These types will be pre-set for you in assignments, but your TAs can also show you how to switch between them in your own notebooks (very useful!)
  - **Code** cells are for code (nice) and for multiple choice questions
  - **Markdown** is for short answer questions
- Once you've input your answer, you can run the cell (click RUN or shift+enter while in the cell) to confirm it worked
  - Note: This will only catch blatant programming errors, it will not catch whether you've done it right or wrong
- For **code** questions, you can also run the next cell below it, which is a **test** to make sure your **formatting** is correct.
  - Again, this is not about substantive correctness. It just tells you whether you've answered the question the right way
  - This is a big advantage for you over not using this format as it means you won't lose points for formatting errors, which can be common in programming courses


jupyter Homework1 (autosaved)
Logout
Control Panel
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
Validate
DS-UA 111: Homework One
This homework is due Tuesday, Feb. 18 by 8:00p. Late homework will be graded down, no exceptions. Improperly formatted homeworks also count as late. Note that the course academic honesty policy applies to every homework, including this one. Some of the questions refer to articles, which you can find by clicking on the links provided. This homework is worth 40 points (one point per sub-question).
Instructions
Please complete your answers in the spaces provided. They will be either cells for code or Markdown, and we will make it clear within each question how to reply.
The submission process for this and all assignments is explained in separate documentation from Lecture 2.2. Make sure to submit your assignment in both NB grader and GradeScope as described in Lecture 2.2.
Question 1
The following statements arise from various theories. In each case, correctly identify the independent variable(s) by typing the letter of your intended answer in place of 'X'. For example, if you want to indicate your answer is A, please type 'A' instead of the 'X' that is currently there. Note you must retain the quotation marks around 'A'.

Question 1

*The following statements arise from various theories. In each case, correctly identify the **independent variable(s)** by typing the letter of your intended answer in place of 'X'. For example, if you want to indicate your answer is A, please type 'A' instead of the 'X' that is currently there. Note you must retain the quotation marks around 'A'.*

(a) Cholera infection is caused by drinking from the Broad Street pump.

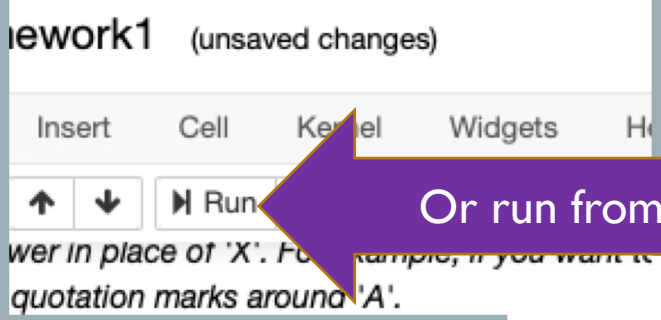
Type the letter of the option that represents the independent variable(s):
A. Cholera
B. Infection with cholera
C. The Broad Street pump
D. Drinking from the Broad Street pump
E. None of the above

In [ ]:
```
q1a_answer = 'X'
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q1a_answer' in locals()
assert q1a_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

(b) Successful people are successful because they wake up early!

Type the letter of the option that represents the independent variable(s):



Or run from menu

the quotation marks around 'A'.

(a) Cholera infection is caused by drinking from the Broad Street pump.

Type the letter of the option that represents the independent variable(s):

A. Cholera
B. Infection with cholera
C. The Broad Street pump
D. Drinking from the Broad Street pump
E. None of the above

In [3]:
```
q1a_answer = 'A'
```

Shift + Enter to run

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'q1a_answer' in locals()
assert q1a_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

E. None of the above

```
In [4]: q1a_answer = 'A'
        aa
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-4-7ba1507086c7> in <module>
      1 q1a_answer = 'A'
----> 2 aa

NameError: name 'aa' is not defined
```

This tells me my formatting is wrong

A. Cholera
B. Infection with cholera
C. The Broad Street pump
D. Drinking from the Broad Street pump
E. None of the above

In [5]:
```
q1a_answer = 'A'
```

Run

In [6]:
```
### BEGIN PUBLIC TESTS
assert 'q1a_answer' in locals()
assert q1a_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

Run

WE ARE GOOD ON FORMATTING!

But: Note this does **not** tell me if my answer is **correct**. It just helps as a failsafe against losing points for formatting/non-substantive or silly errors

Another failsafe!

Technically my formatting is correct

But it's not one of the possible options

```
In [7]: q1a_answer = 'X'

In [8]: ### BEGIN PUBLIC TESTS
        assert 'q1a_answer' in locals()
        assert q1a_answer.upper() in ['A','B','C','D','E']
        ### END PUBLIC TESTS
```

```
---------------------------------------------------------------------------
AssertionError                            Traceback (most recent call last)
<ipython-input-8-5d96dd7976bd> in <module>
      1 ### BEGIN PUBLIC TESTS
      2 assert 'q1a_answer' in locals()
----> 3 assert q1a_answer.upper() in ['A','B','C','D','E']
      4 ### END PUBLIC TESTS

AssertionError:
```

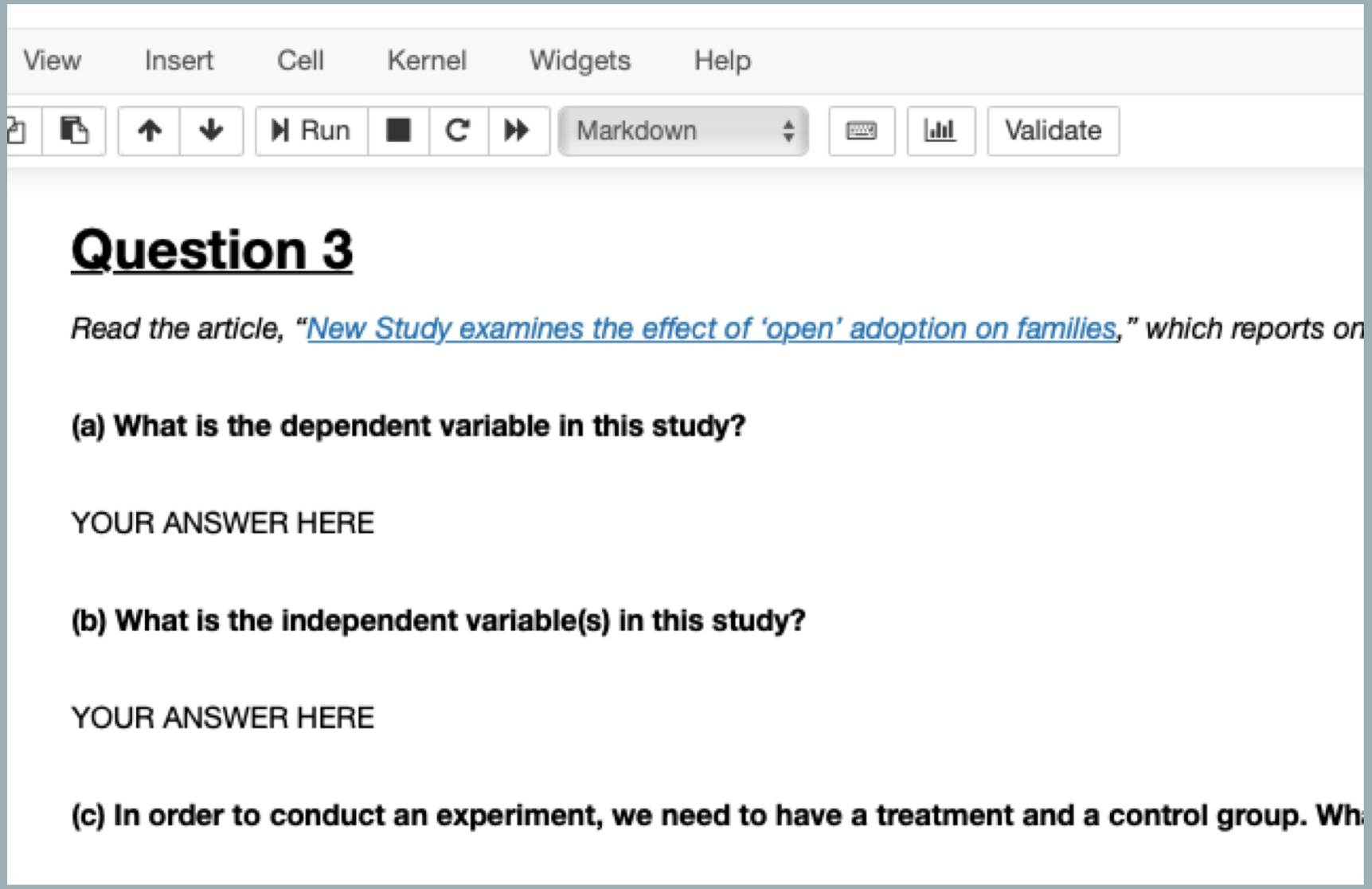
WE ARE GOOD ON TYPE OF ANSWER

```
In [9]: q1a_answer = 'A'

In [10]: ### BEGIN PUBLIC TESTS
         assert 'q1a_answer' in locals()
         assert q1a_answer.upper() in ['A','B','C','D','E']
         ### END PUBLIC TESTS
```

Again, this does not tell you whether the answer is correct.
Your TAs will grade and then you'll know if it's correct!

Markdown example



Question 3

Read the article, "New Study examines the effect of 'open' adoption on fam

(a) What is the dependent variable in this study?

```
YOUR ANSWER HERE
```

(b) What is the independent variable(s) in this study?

YOUR ANSWER HERE

(c) In order to conduct an experiment, we need to have a treatment an

Question 3

Read the article, "New Study examines the effect of 'open' ado

(a) What is the dependent variable in this study?

```
here is my answer
```

(b) What is the independent variable(s) in this study?

YOUR ANSWER HERE

Question 3

Read the article, "New Study examines the effect of 'open' .

(a) What is the dependent variable in this study?

here is my answer

Run, as before

(b) What is the independent variable(s) in this study?

YOUR ANSWER HERE

(g) We learned in lecture that you cannot manipulate lists as much as arrays in Python. Ye

YOUR ANSWER HERE

End of Homework

Please remember to submit it correctly -- as shown in Lecture 2.2 and practiced in Lab 0!

We got to the end!

3. SUBMIT THE ASSIGNMENT ON JUPYTERHUB

- Optional: Validate
  - Either within the notebook or under the Assignments tab
- Submit the assignment under the Assignments tab
- Observe the timestamp confirming you have submitted it
- You can submit many times, but we will **only** grade the most recent
- If the most recent is **after** the deadline, it counts as **late**
  - **NO exceptions**

jupyter Homework1 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

Run Markdown Validate

In [ ]:
```
ages_list = ...

ages_mean_list = ...
```

In [ ]:
```
### BEGIN PUBLIC TESTS
assert 'ages_list' in locals()
assert 'ages_mean_list' in locals()
### END PUBLIC TESTS
```

(g) We learned in lecture that you cannot manipulate lists as much as arrays in Python. Yet, the mean calculation worked for this list. Why?

YOUR ANSWER HERE

End of Homework

Please remember to submit it correctly -- as shown in Lecture 2.2 and practiced in Lab 0!

Validation Results

The following cell failed:

```
### BEGIN PUBLIC TESTS
assert 'q1b_answer' in locals()
assert q1b_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

```
---------------------------------------------------------------------------
AssertionError                            Traceback (most recent call last)
<ipython-input-4-0b9629723ec6> in <module>
      1 ### BEGIN PUBLIC TESTS
      2 assert 'q1b_answer' in locals()
----> 3 assert q1b_answer.upper() in ['A','B','C','D','E']
      4 ### END PUBLIC TESTS

AssertionError:
```

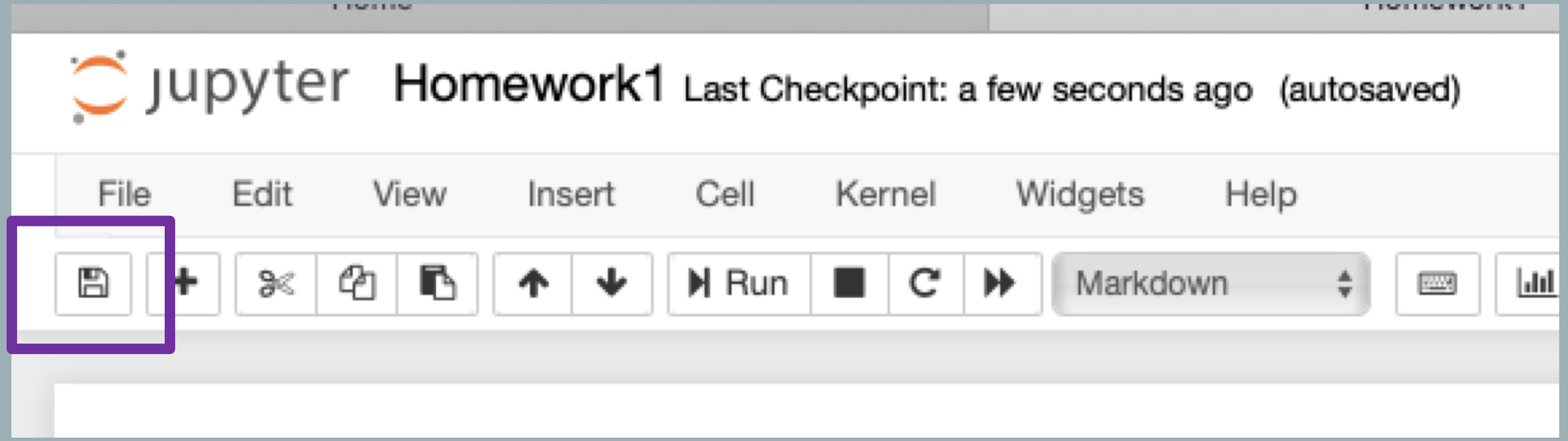
The following cell failed:

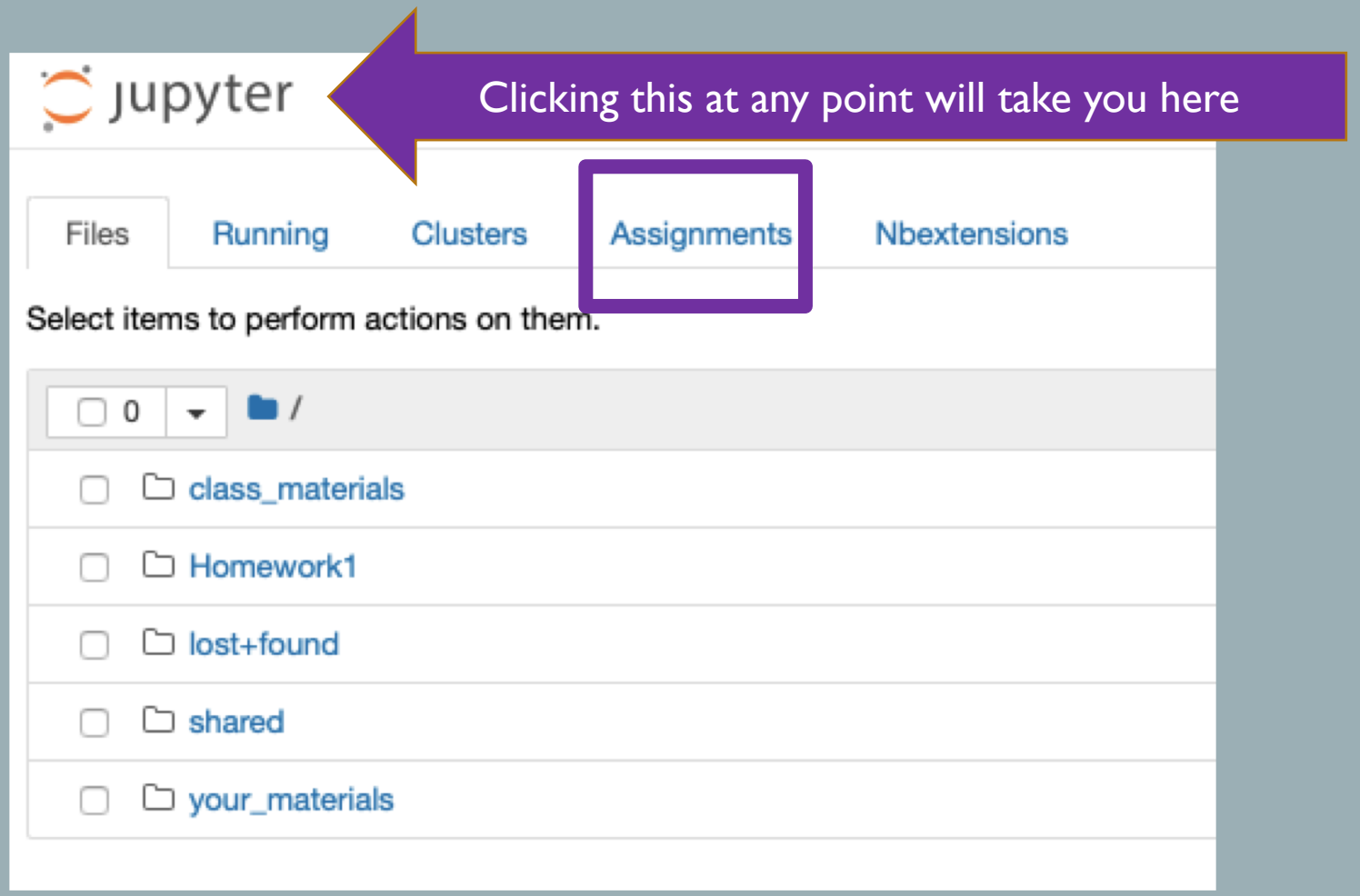
```
### BEGIN PUBLIC TESTS
assert 'q1c_answer' in locals()
assert q1c_answer.upper() in ['A','B','C','D','E']
### END PUBLIC TESTS
```

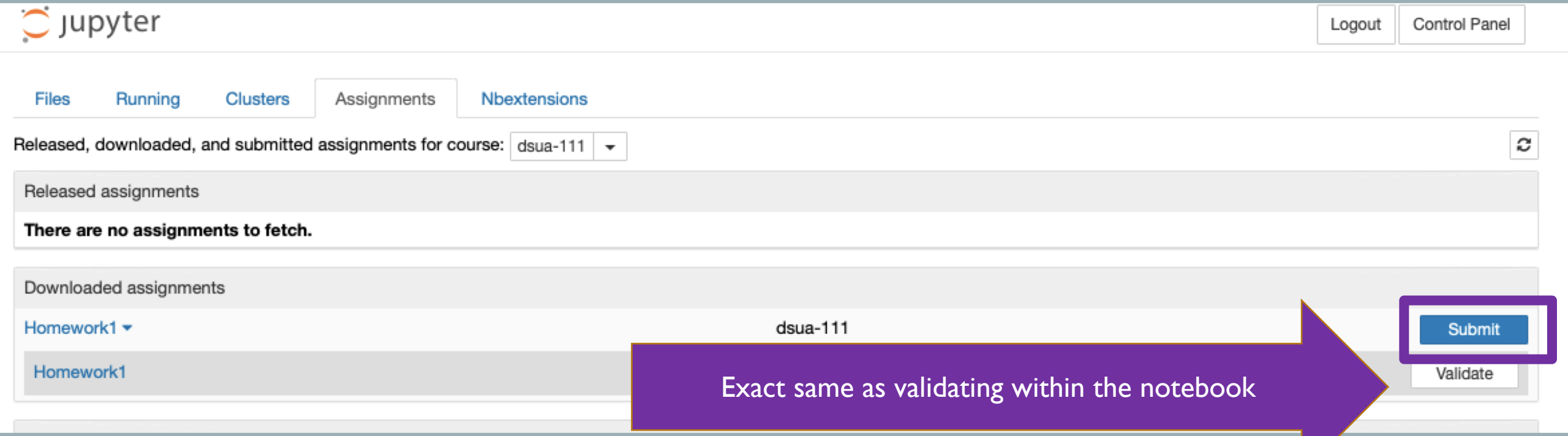
```
---------------------------------------------------------------------------
AssertionError                            Traceback (most recent call
<ipython-input-6-d795c9f86ee4> in <module>
```

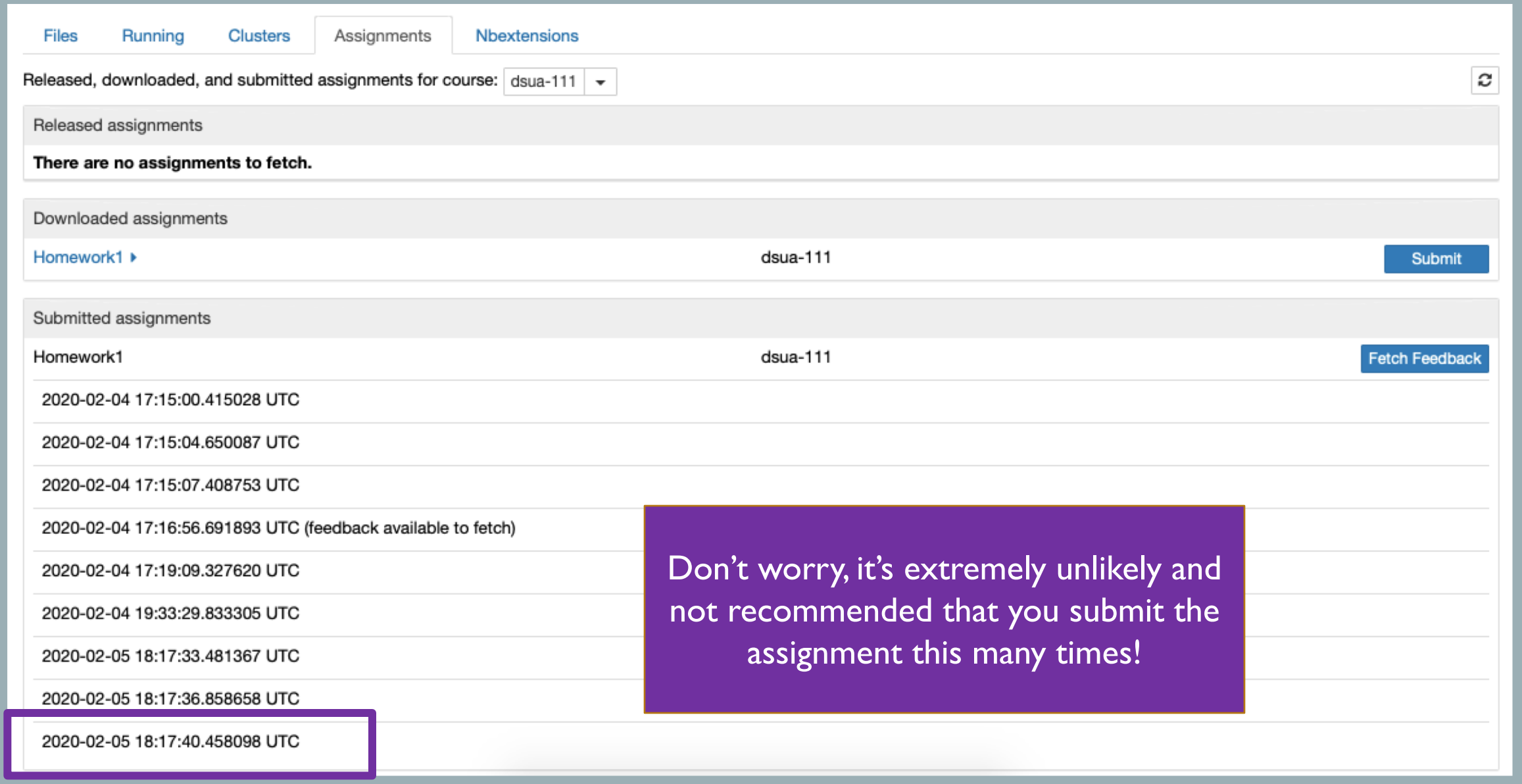
```
---------------------------------
    Traceback (most recent call last)
dule>
```

OK


jupyter
Homework1
Last Checkpoint: a few seconds ago
(autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Markdown

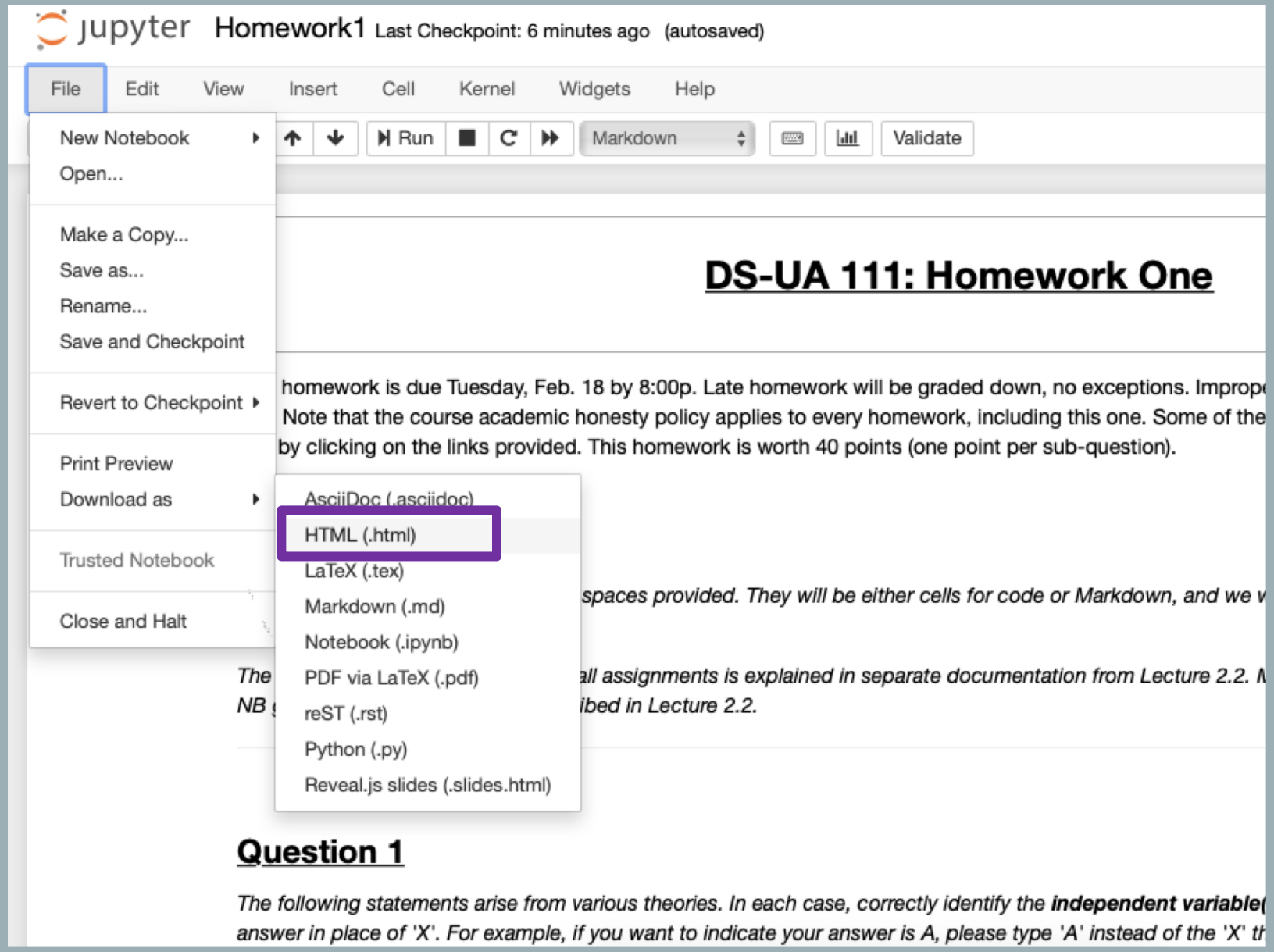

jupyter
Clicking this at any point will take you here
Files
Running
Clusters
Assignments
Nbextensions
Select items to perform actions on them.
0
/
class_materials
Homework1
lost+found
shared
your_materials

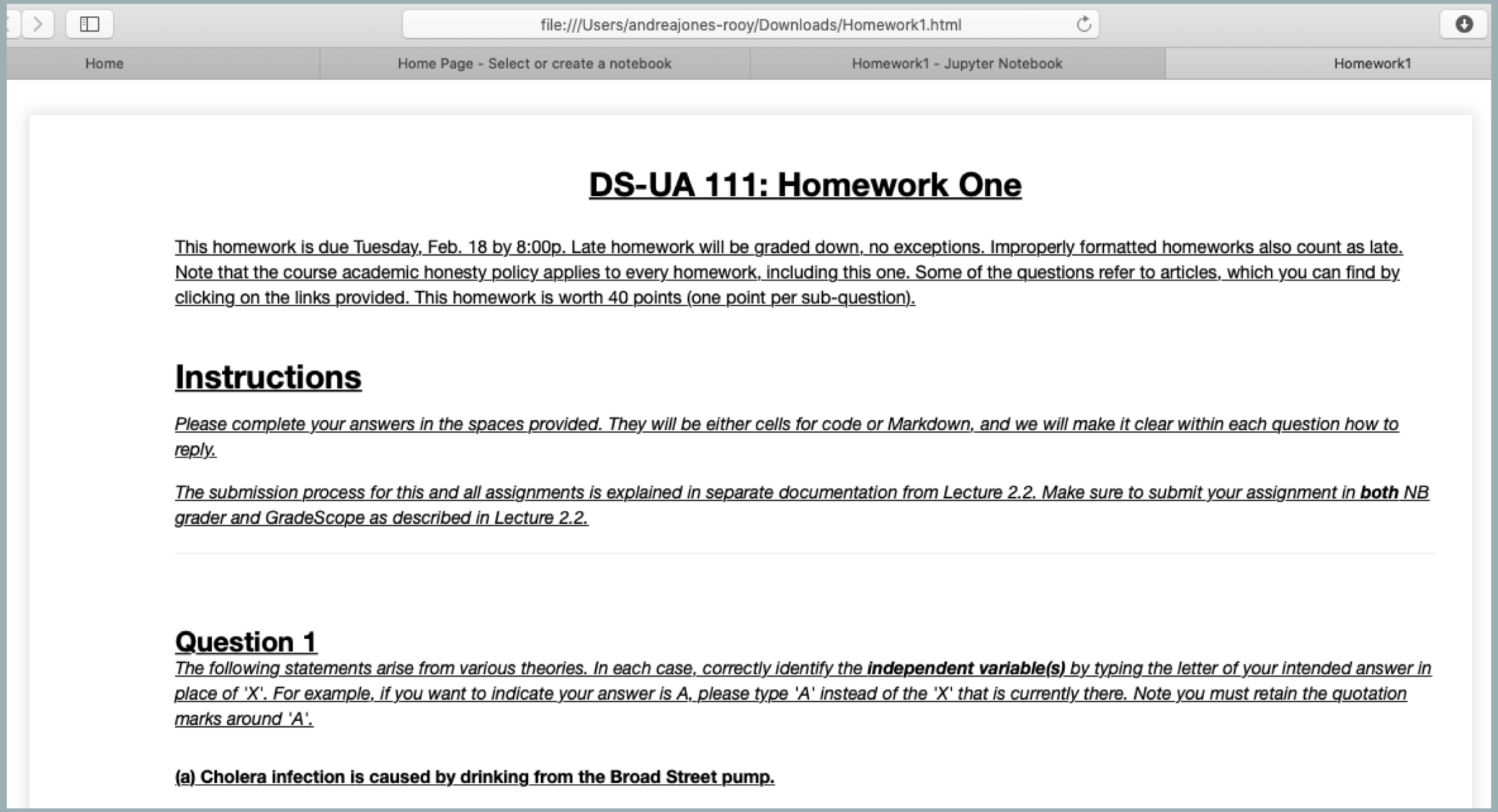

jupyter
Logout
Control Panel
Files
Running
Clusters
Assignments
Nbextensions
Released, downloaded, and submitted assignments for course:
dsua-111
Released assignments
There are no assignments to fetch.
Downloaded assignments
Homework1
dsua-111
Submit
Homework1
Validate
Exact same as validating within the notebook

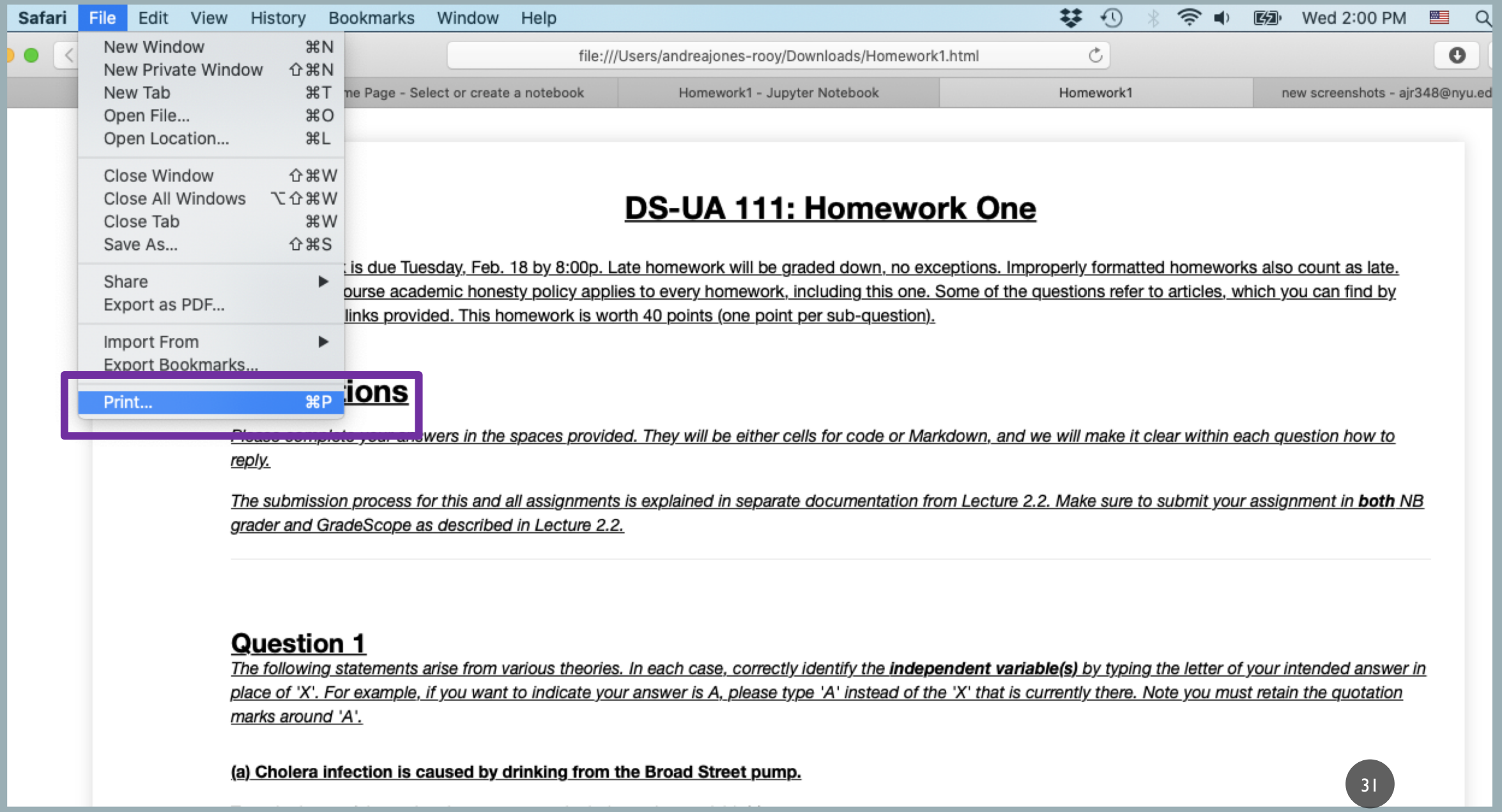

Files
Running
Clusters
Assignments
Nbextensions
Released, downloaded, and submitted assignments for course: dsua-111
Released assignments
There are no assignments to fetch.
Downloaded assignments
Homework1
dsua-111
Submit
Submitted assignments
Homework1
dsua-111
Fetch Feedback
2020-02-04 17:15:00.415028 UTC
2020-02-04 17:15:04.650087 UTC
2020-02-04 17:15:07.408753 UTC
2020-02-04 17:16:56.691893 UTC (feedback available to fetch)
2020-02-04 17:19:09.327620 UTC
2020-02-04 19:33:29.833305 UTC
2020-02-05 18:17:33.481367 UTC
2020-02-05 18:17:36.858658 UTC
2020-02-05 18:17:40.458098 UTC
Don't worry, it's extremely unlikely and not recommended that you submit the assignment this many times!

4. SUBMIT TO GRADESCOPE

- Go back into the assignment you just submitted
- Go to File > Download as > HTML (.html)
- Download as html
- Find the file on your computer and open it
- Log into Gradescope and find the assignment (e.g., Homework1)
- Manually assign questions to pages
  - E.g., Question 1a appears on page 1, so click page 1
  - If a question spans more than one page, click both pages
- Submit (noting unmatched pages and questions and checking if errors)
  - E.g., if instructions are on page 1, Gradescope will say "you haven't assignment any questions to page 1"
  - If you forgot to assign a question to a page, Gradescope will say "you haven't assigned a page to this question"
- Again, this is a check for your protection from silly/non-substantive mistakes!


jupyter Homework1 Last Checkpoint: 6 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
New Notebook
Open...
Make a Copy...
Save as...
Rename...
Save and Checkpoint
Revert to Checkpoint
Print Preview
Download as
Trusted Notebook
Close and Halt
Run
Markdown
Validate
AsciiDoc (.asciidoc)
HTML (.html)
LaTeX (.tex)
Markdown (.md)
Notebook (.ipynb)
PDF via LaTeX (.pdf)
reST (.rst)
Python (.py)
Reveal.js slides (.slides.html)
DS-UA 111: Homework One
homework is due Tuesday, Feb. 18 by 8:00p. Late homework will be graded down, no exceptions. Improp
Note that the course academic honesty policy applies to every homework, including this one. Some of the
by clicking on the links provided. This homework is worth 40 points (one point per sub-question).
spaces provided. They will be either cells for code or Markdown, and we
The
all assignments is explained in separate documentation from Lecture 2.2. N
NB
ibed in Lecture 2.2.
Question 1
The following statements arise from various theories. In each case, correctly identify the independent variable(
answer in place of 'X'. For example, if you want to indicate your answer is A, please type 'A' instead of the 'X' th


file:///Users/andreajones-rooy/Downloads/Homework1.html
Home
Home Page - Select or create a notebook
Homework1 - Jupyter Notebook
Homework1
DS-UA 111: Homework One
This homework is due Tuesday, Feb. 18 by 8:00p. Late homework will be graded down, no exceptions. Improperly formatted homeworks also count as late. Note that the course academic honesty policy applies to every homework, including this one. Some of the questions refer to articles, which you can find by clicking on the links provided. This homework is worth 40 points (one point per sub-question).
Instructions
Please complete your answers in the spaces provided. They will be either cells for code or Markdown, and we will make it clear within each question how to reply.
The submission process for this and all assignments is explained in separate documentation from Lecture 2.2. Make sure to submit your assignment in both NB grader and GradeScope as described in Lecture 2.2.
Question 1
The following statements arise from various theories. In each case, correctly identify the independent variable(s) by typing the letter of your intended answer in place of 'X'. For example, if you want to indicate your answer is A, please type 'A' instead of the 'X' that is currently there. Note you must retain the quotation marks around 'A'.
(a) Cholera infection is caused by drinking from the Broad Street pump.


Safari File Edit View History Bookmarks Window Help
Wed 2:00 PM
New Window ⌘N
New Private Window ⇧⌘N
New Tab ⌘T
Open File... ⌘O
Open Location... ⌘L
Close Window ⇧⌘W
Close All Windows ⌥⇧⌘W
Close Tab ⌘W
Save As... ⇧⌘S
Share
Export as PDF...
Import From
Export Bookmarks...
Print... ⌘P
file:///Users/andreajones-rooy/Downloads/Homework1.html
Homework1 - Jupyter Notebook
Homework1
new screenshots - ajr348@nyu.ed
DS-UA 111: Homework One
is due Tuesday, Feb. 18 by 8:00p. Late homework will be graded down, no exceptions. Improperly formatted homeworks also count as late.
ourse academic honesty policy applies to every homework, including this one. Some of the questions refer to articles, which you can find by
links provided. This homework is worth 40 points (one point per sub-question).
ions
Please complete your answers in the spaces provided. They will be either cells for code or Markdown, and we will make it clear within each question how to reply.
The submission process for this and all assignments is explained in separate documentation from Lecture 2.2. Make sure to submit your assignment in both NB grader and GradeScope as described in Lecture 2.2.
Question 1
The following statements arise from various theories. In each case, correctly identify the independent variable(s) by typing the letter of your intended answer in place of 'X'. For example, if you want to indicate your answer is A, please type 'A' instead of the 'X' that is currently there. Note you must retain the quotation marks around 'A'.
(a) Cholera infection is caused by drinking from the Broad Street pump.

DS-UA 111: Homework One

This homework is due Tuesday, Feb. 18 by 8:00p. Late homework will be graded down, no exceptions. Improperly formatted homeworks also count as late. Note that the course academic honesty policy applies to every homework, including this one. Some of the questions refer to articles, which you can find by clicking on the links provided. This homework is worth 40 points (one point per sub-question).

Instructions

Please complete your answers in the spaces provided. They will be either cells for code or Markdown, and we will make it clear within each question how to reply.

Homework 1 | Assign Questions and Pages

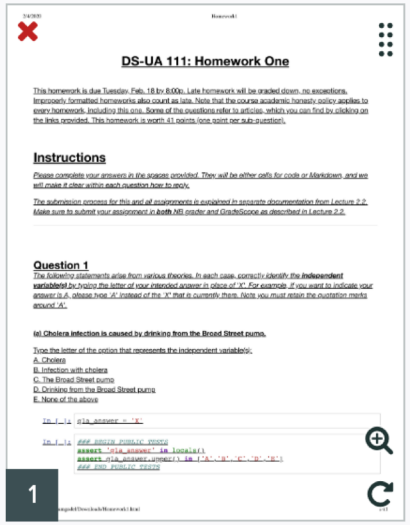
SUBMITTED AT: **FEBRUARY 5, 10:40 AM**

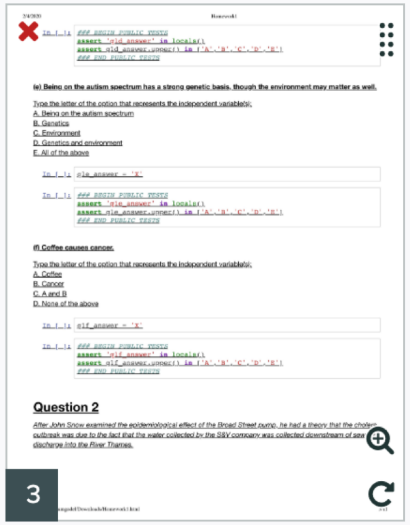
Select questions and pages to indicate where your responses are located. Use **esc** to deselect all items and hold **shift** to select multiple questions.

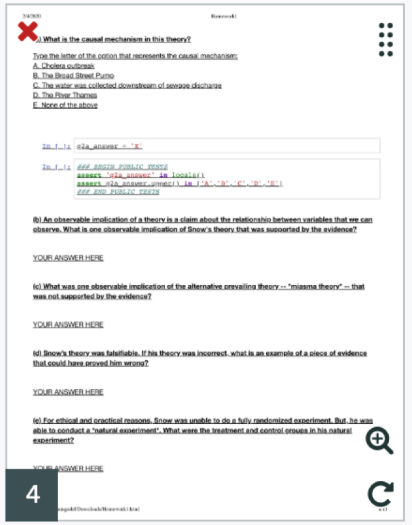
Question Outline

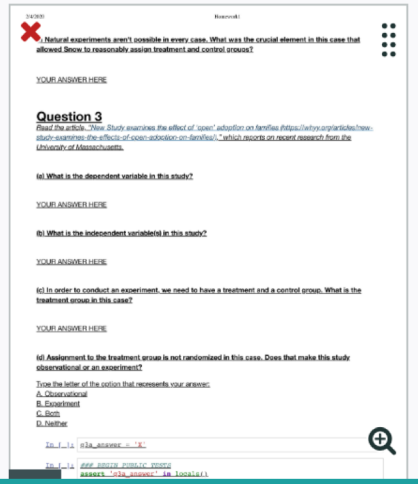
Select a question or a page.

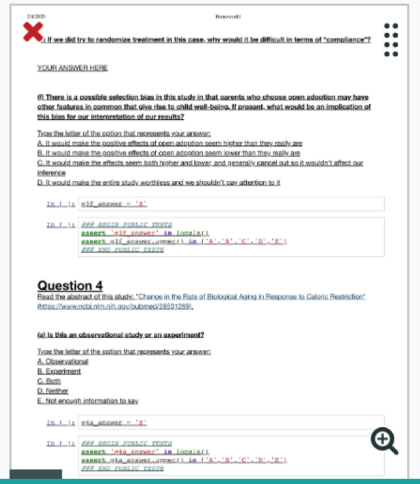
TITLE	POINTS
1 Question 1	0.0 pts
1.1 Q1A	0.0 pts
1.2 Q1B	0.0 pts
1.3 Q1C	0.0 pts
1.4 Q1D	0.0 pts
1.5 Q1E	0.0 pts
1.6 Q1F	0.0 pts
2 Question 2	5.0 pts
2.1 Q2A	0.0 pts
2.2 Q2B	1.0 pt
2.3 Q2C	1.0 pt
2.4 Q2D	1.0 pt

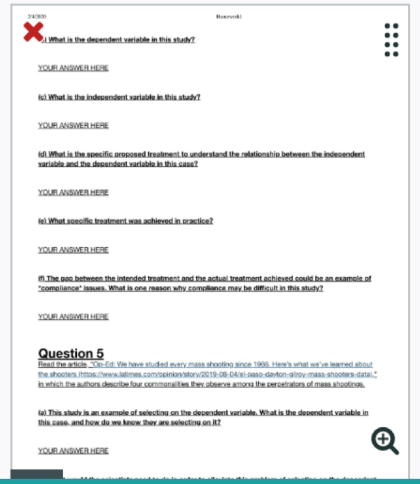


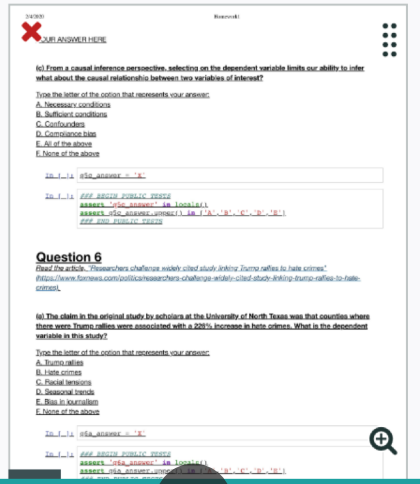












Select a question or a page.

Assign Pages Sequentially

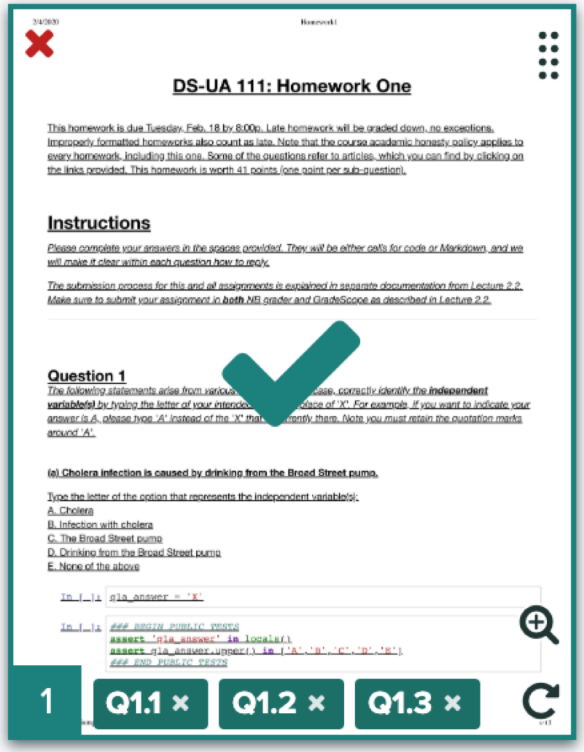
33

Submit

Question Outline

Select pages to assign to Question **1.3**.

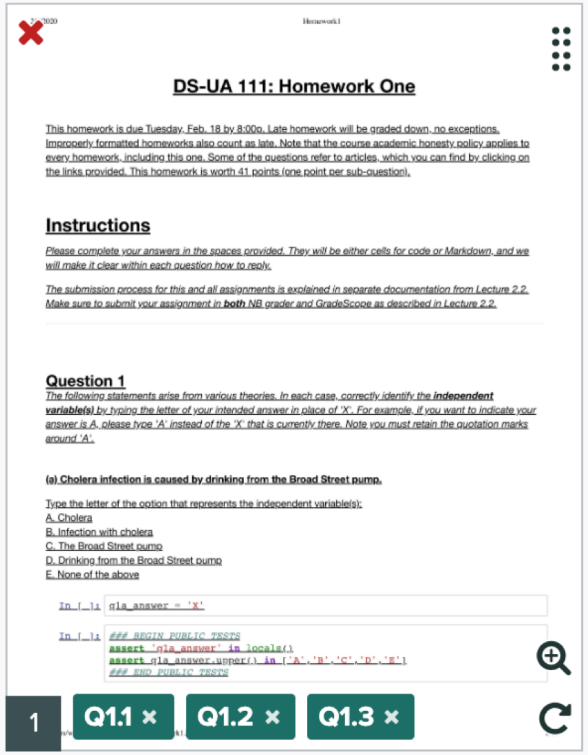
TITLE	POINTS
1 Question 1	0.0 pts
1.1 Q1A P1 ×	0.0 pts
1.2 Q1B P1 ×	0.0 pts
1.3 Q1C P1 ×	0.0 pts

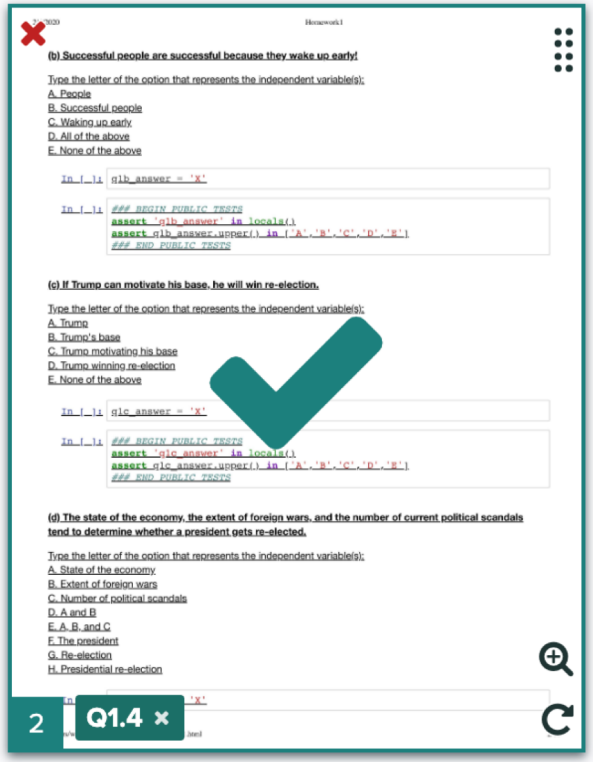


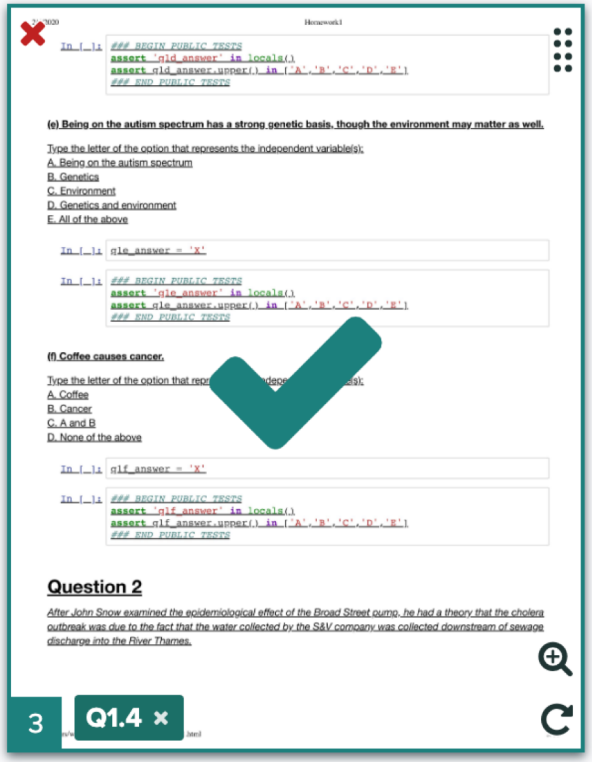
Question Outline

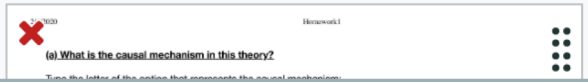
Select pages to assign to Question **1.4**.

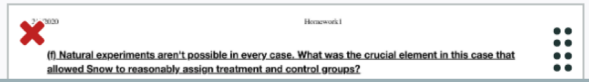
TITLE	POINTS
1 Question 1	0.0 pts
1.1 Q1A P1 ×	0.0 pts
1.2 Q1B P1 ×	0.0 pts
1.3 Q1C P1 ×	0.0 pts
1.4 Q1D P2 × P3 ×	0.0 pts
1.5 Q1E	0.0 pts

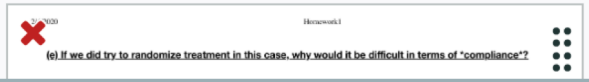


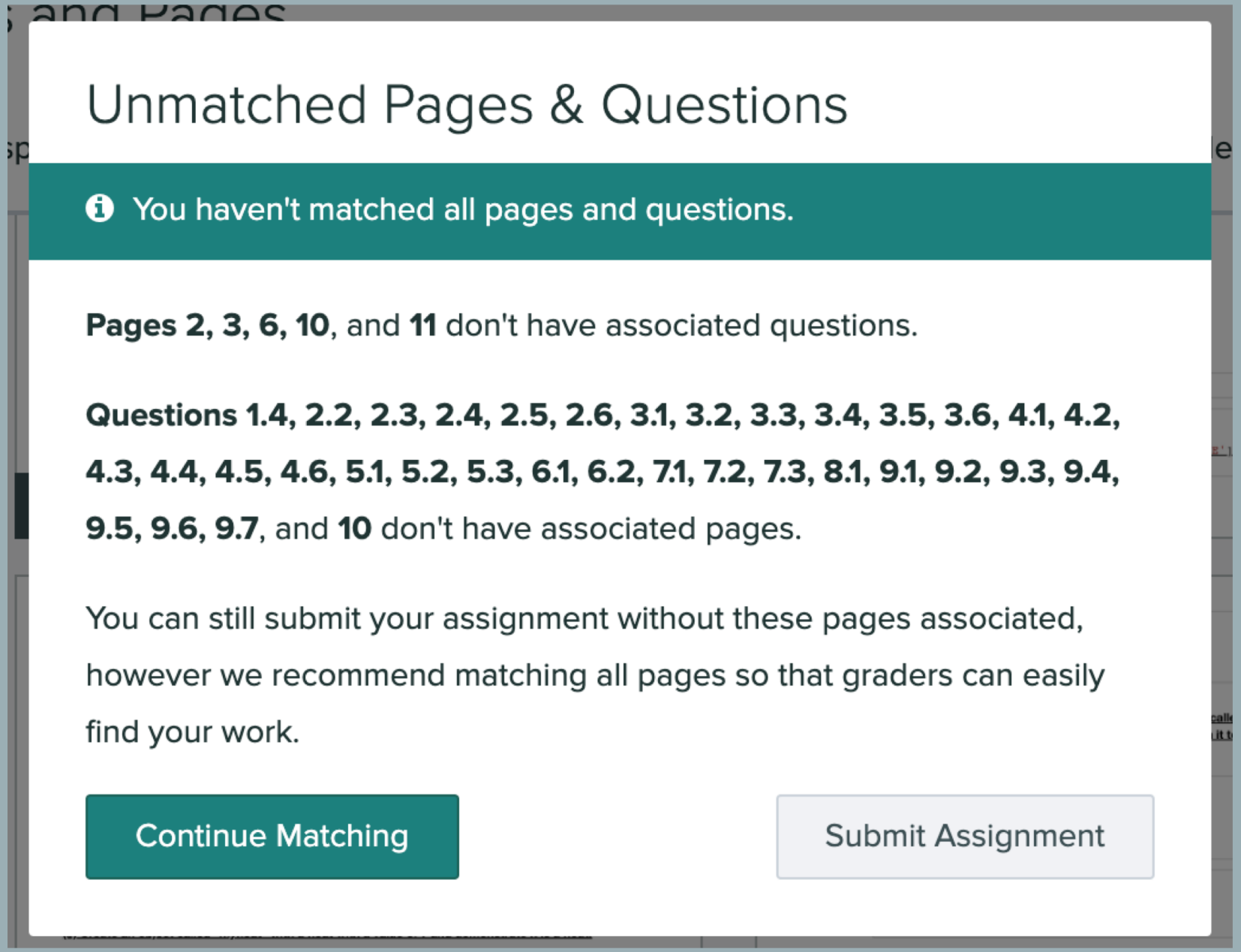











Unmatched Pages & Questions
You haven't matched all pages and questions.
Pages 2, 3, 6, 10, and 11 don't have associated questions.
Questions 1.4, 2.2, 2.3, 2.4, 2.5, 2.6, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 5.1, 5.2, 5.3, 6.1, 6.2, 7.1, 7.2, 7.3, 8.1, 9.1, 9.2, 9.3, 9.4, 9.5, 9.6, 9.7, and 10 don't have associated pages.
You can still submit your assignment without these pages associated, however we recommend matching all pages so that graders can easily find your work.
Continue Matching
Submit Assignment

SUMMARY OF THE ENTIRE PROCESS

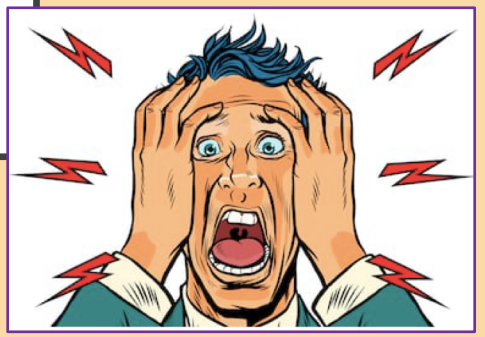
1. Fetch assignment on JupyterHub

2. Complete assignment in that notebook on JupyterHub

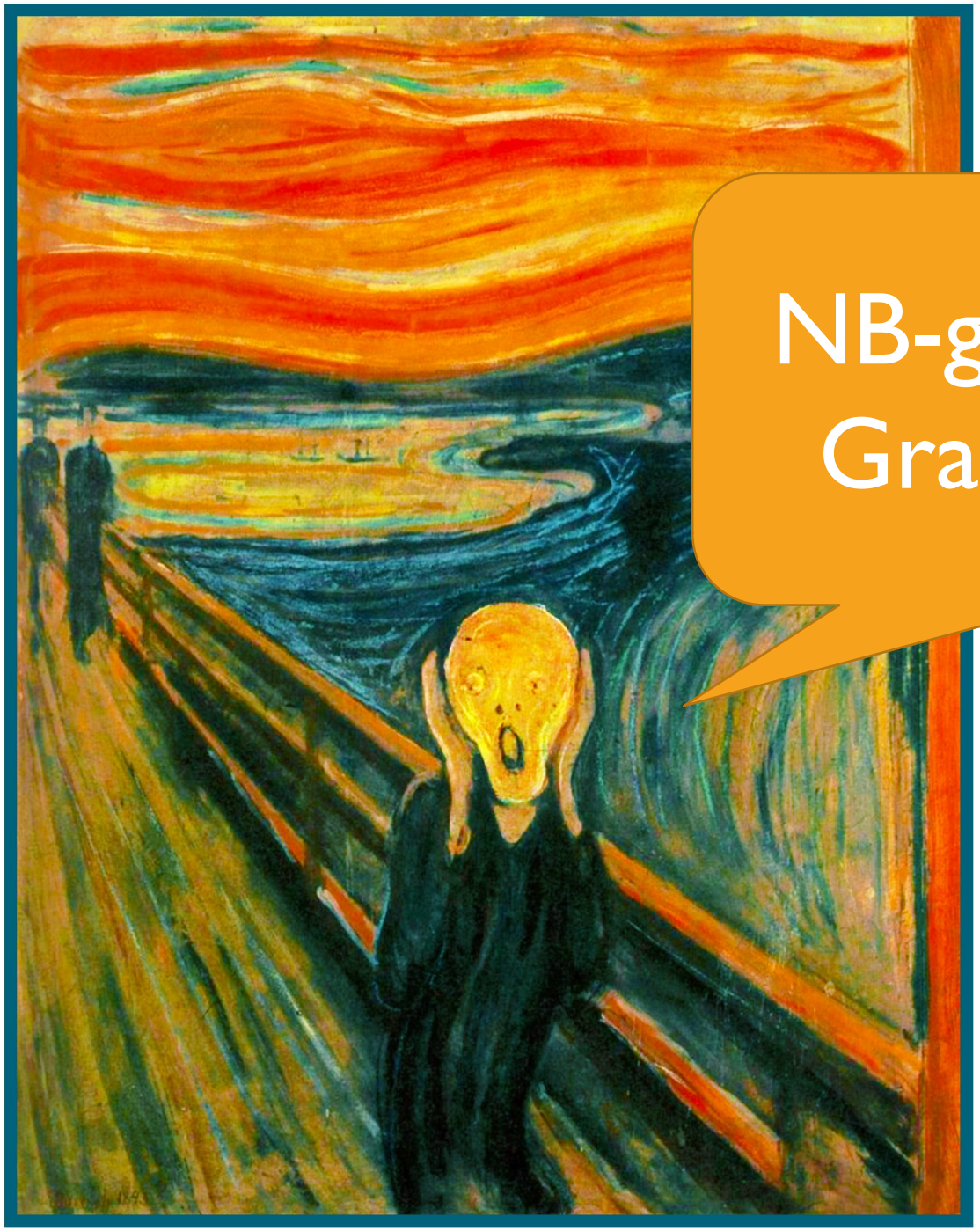
3. Submit assignment on JupyterHub (NB-grader)

4. Submit assignment on Gradescope

WHY ARE WE MAKING YOU DO ALL THIS CRAP?



1. You get more feedback more quickly
2. You get more coding practice
   - In a course this size, we cannot assign this many labs without these platforms, so you get more practice
3. You will encounter these platforms in other courses across NYU, and will definitely encounter them in the data science program (e.g., IDS uses the same workflow)
4. When you have jobs as data scientists (companies, government, universities, think tanks, non-profits): You **will** be navigating multiple platforms, some of which are public/you know and some of which are unique to that company
5. You can easily & quickly ask for question-specific regrades rather than have to wait for office hours or exchange a bunch of emails/meetings.
   - That said, do not abuse this power!
6. The alternative is just as complex and less useful to you
   - Write own notebook/doc, often take screenshots, create PDF, upload to Classes, resubmission is very hard, resubmission is very difficult on Classes, regrade requests are much more time-consuming
7. It's **exactly** the same for **all** assignments, and will get easy, we promise! (and TAs have done this in their own courses)



NB-grader **AND** Gradescope?!?!

Outline

1. How to complete & submit all assignments in this course
2. Key concepts in causality
3. Principles & types of causality

KEY CONCEPTS IN CAUSALITY

Lecture 2.1

- Theory
- Good theory
- Observable implication
- Hypothesis
- Causal mechanism
- Association
- Causation
- Experiment
- Treatment
- Control
- Randomization
- Natural experiment

Happening now!

- Independent variable
- Dependent variable
- Confounders
- Deterministic vs. Probabilistic causality
- Selecting on the dependent variable
- Endogeneity
- Necessary and. sufficient conditions

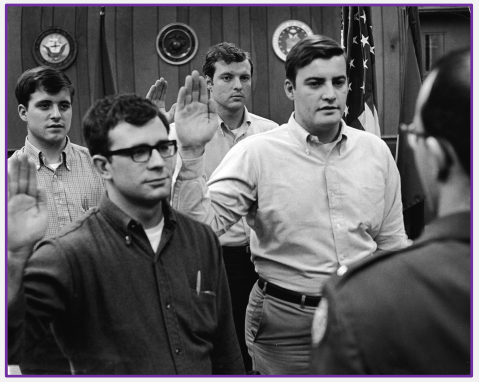
OBSERVATIONAL VS. EXPERIMENTAL STUDIES

- **Observational**: We look at data from the world and test hypotheses as best we can, but we cannot infer causality
  - Typically our inferences are limited to associations
- **Experimental**: Treatment and control groups where subjects have been randomized into the groups. Ideally double-blind
  - We still can't **conclude** causality, and we cannot **prove**, but we can make much stronger inferences
- We like experimental studies because they help us more rigorously explore the **counterfactual** (it isn't the counterfactual exactly, but it's as close as we can get)
- We like randomization because it protects us (hopefully) against selection bias
- When we can't do a randomized, controlled experiment, we can creatively think about natural experiments
- Does this mean observational studies are useless?
  - No, but it means being cautious and humble when interpreting results
  - And following good principles, like not selecting on the dependent variable

Uncertainty is a big part of this process!

And that's ok! Often good, even!

EXAMPLES OF NATURAL EXPERIMENTS



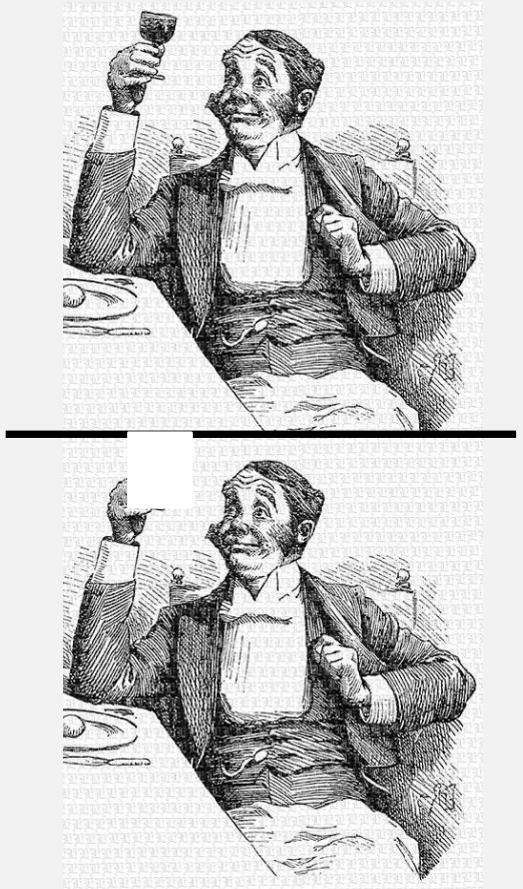
- The US military service draft lottery
  - During the Vietnam war, young men were drafted by a lottery that pulled birthdays
  - This means it was random and presumably unrelated to relevant features of who ended up serving in the army
  - This means after the war, we can study things like the effect of war on, e.g., mental health, income, family status, etc. etc.
- A new highway in India
  - Does living within easy access to a hospital lead to better health outcomes?
- The *NYTimes* science section
  - Does publication in the *New York Times* science section increase citations of scientific papers?
- Air pollution effects on health in Beijing
  - Does exposure to air pollution from vehicle traffic lead to more trips to the hospital with lung issues?

CONFOUNDING AND CONFOUNDERS

- Snow *couldn't control* the assignment of the treatment.
  - E.g., he couldn't randomize
- So, he had to worry about ***confounders***: factors that are different between treatment and control groups that could be affecting their outcomes (separate to the treatment).
- E.g., we observe that people who attend church regularly tend to be more optimistic about life than those that don't. Is this the ***causal effect*** of church? What might be a ***confounder***?

SAME RULES APPLY FOR OBSERVATIONAL STUDIES

- To test a ***hypothesis***, we need to compare units with different values on the ***independent variables*** (treatment status) in terms of whether this results in different values on their ***dependent variables*** (outcomes)
- Obviously, the variables must ***vary***! People need to have different values of X (and Y).
- Less obviously, ***don't select on the dependent variable*** and try to draw conclusions about what generates that outcome!

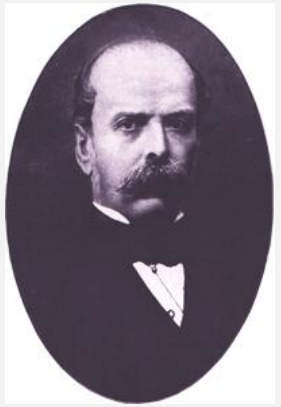


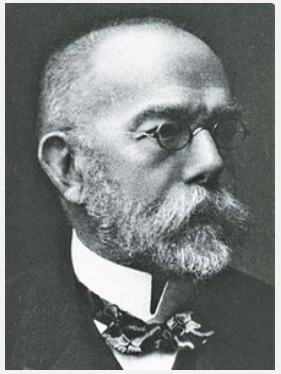
INDEPENDENT & DEPENDENT VARIABLES

- **Dependent variable**: The thing we are trying to explain
- **Independent variable(s)**: The thing(s) we think explain the dependent variable
- Example: My grandmother used to tell me that if I went outside in the winter with wet hair I would catch a cold
  - **Independent variable (IV)**: Going outside with wet hair
  - **Dependent variable (DV)**: Catching a cold
    - Catching a cold **depends** on me going outside with wet hair
    - Is it the **only** way to catch a cold? Hold that thought for necessary & sufficient conditions!
  - **Causal mechanism**: I'm not sure, but I think it was something like it weakens my immune system

REMINDER: CAUSAL MECHANISMS

- Snow didn't know *exactly* how consuming bad water lead to cholera, but he understood that fecal-oral transmission was part of the problem. This was not politically popular at the time!
- So, Snow had a **causal mechanism** in mind: ***a sequence of processes leading from X to Y***.
- We ***need mechanisms***, so we know what to test.
- Later, scientists (Pacini, Koch) would ***refine*** the mechanism, and understand the importance of the cholera bacterium.





WHAT'S WRONG WITH THIS DESIGN?

The New York Times

I've Interviewed 300 High Achievers About Their Morning Routines. Here's What I've Learned.

Your morning routine should suit your needs, but there are some habits everyone should try.

Basically, they get up early.

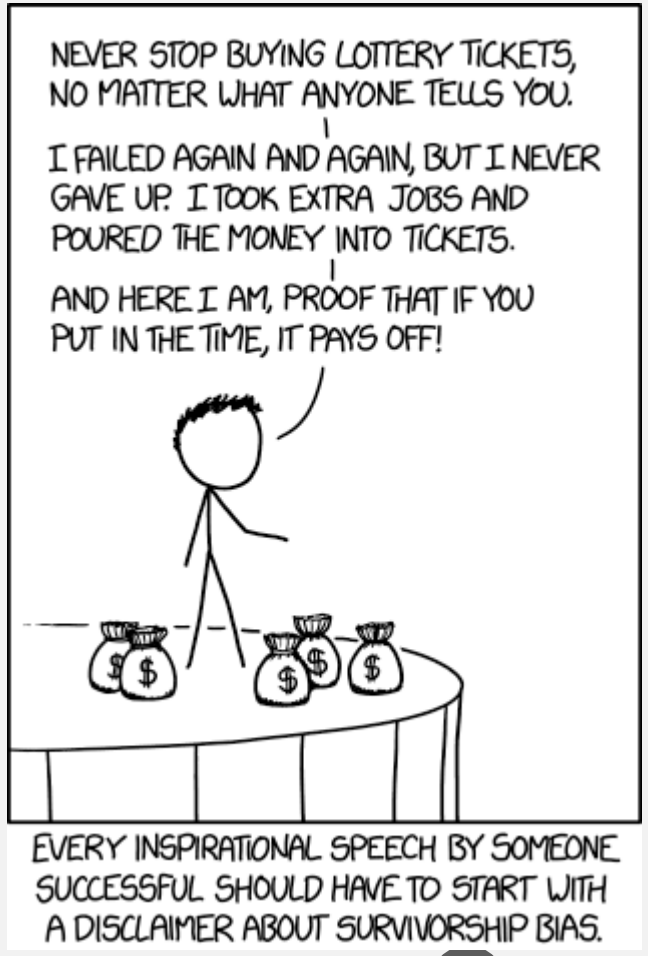
But this is *selecting on the dependent variable* Y="successful". Maybe unsuccessful people *also* get up early? (maybe they get up earlier)

We cannot know if this is a causal factor or not!

WHAT'S WRONG WITH THIS DESIGN?

SURVIVOR BIAS

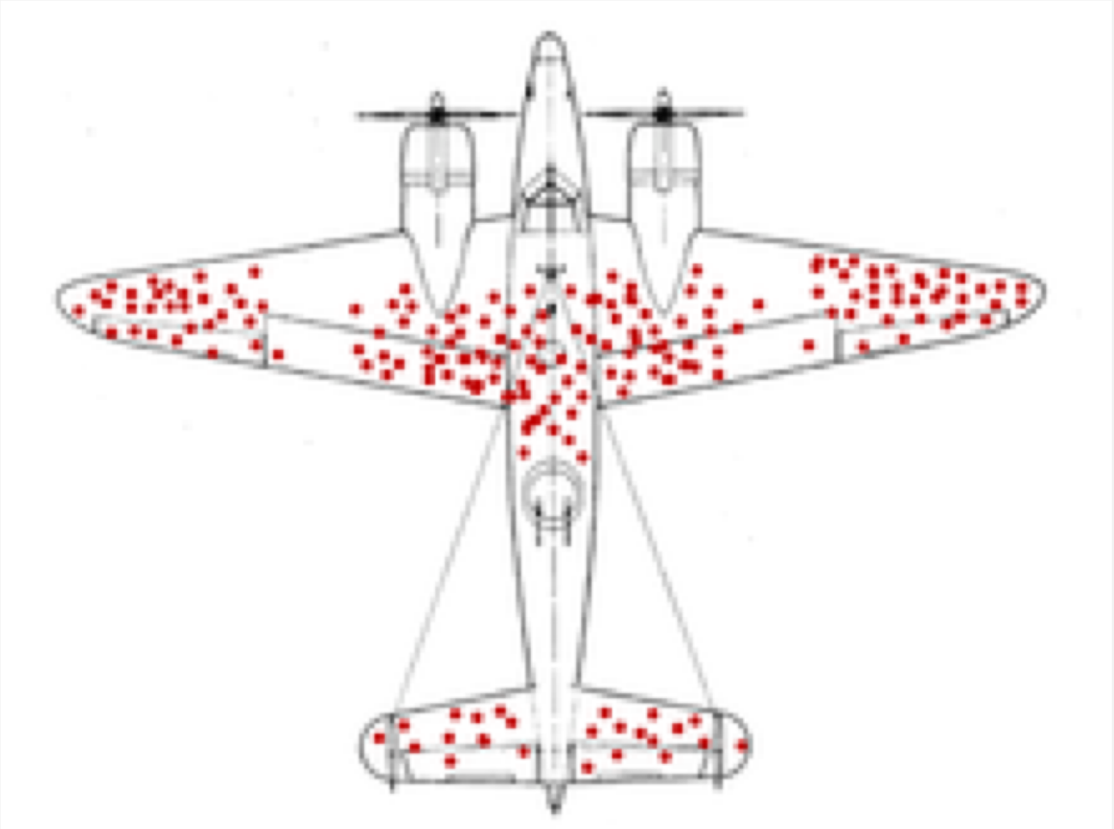
- Looking only at units who ***survived*** a certain process leads to biased inferences. We don't get to see those who didn't survive this process.
- what do very old people have in common (what about those who didn't survive to old age)?
- what do very successful people have in common (what about those who weren't successful)?
- ***Wald's problem***: WW2, bombers return from raids. Where should we put extra armor?



SURVIVOR BIAS

We (only) get to look at the bombers that survived the raid. **Red spots** are damage.

Where should we put the extra armor, next time?



Outline

1. How to complete & submit all assignments in this course
2. Key concepts in causality
3. Principles & types of causality

GENERAL PRINCIPLES OF CAUSALITY

COVARIATION

TEMPORAL PRECEDENCE

CONTROL FOR THIRD VARIABLES

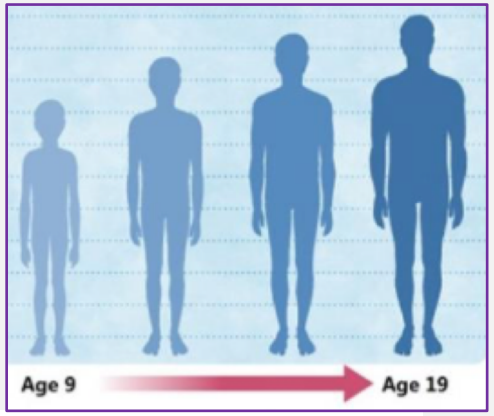
NECESSARY BUT NOT SUFFICIENT

COVARIATION

- generally, we can ***objectively determine*** the statistical association between variables

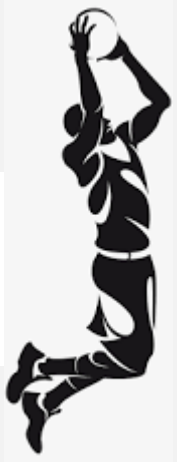
- but as we've seen, association ***does not*** imply causation

- and, actually, we can sometimes have a causal link ***without*** obvious association

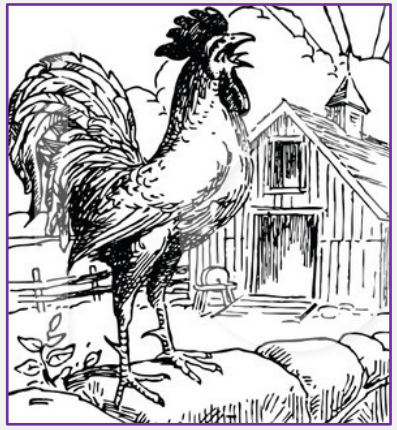


Height **is** correlated with spelling ability. Why?

Height does **not** predict salary in the NBA. Does that mean height doesn't matter for ability?

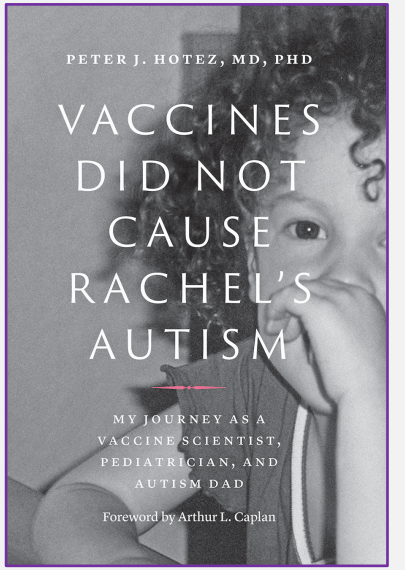


TEMPORAL PRECEDENCE



- we want X (treatment) to ***precede*** the Y (outcome)

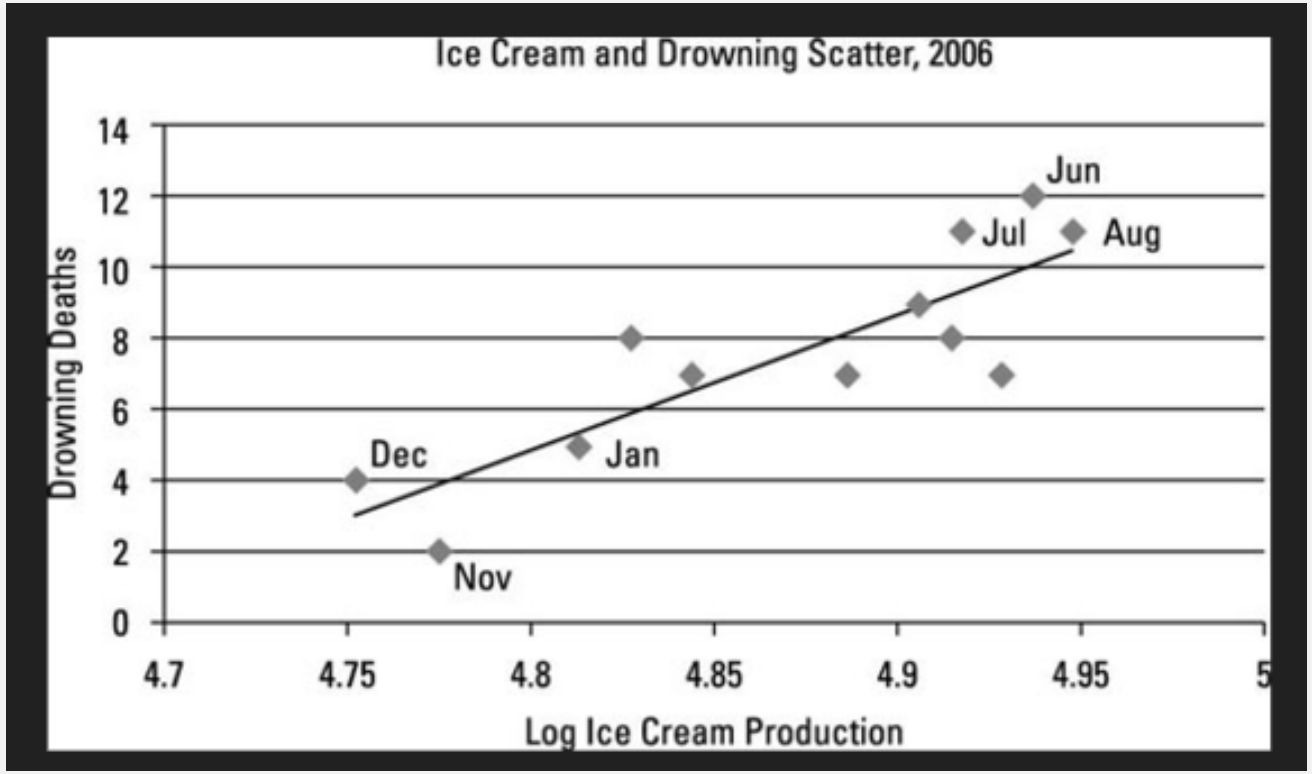
- but just because X came before Y does ***not*** mean it X caused Y!



CONFOUNDERS AND CONTROLS

What should we control for?

- like Snow, we must be careful about **confounders**.
- if we can't randomize (and it's implausible), we may need to "**control**" for that Z



TALKING ABOUT CAUSALITY

1. Deterministic vs. probabilistic causality
2. Endogeneity (reciprocal causation)
3. Necessary and sufficient conditions

DETERMINISTIC V PROBABILISTIC

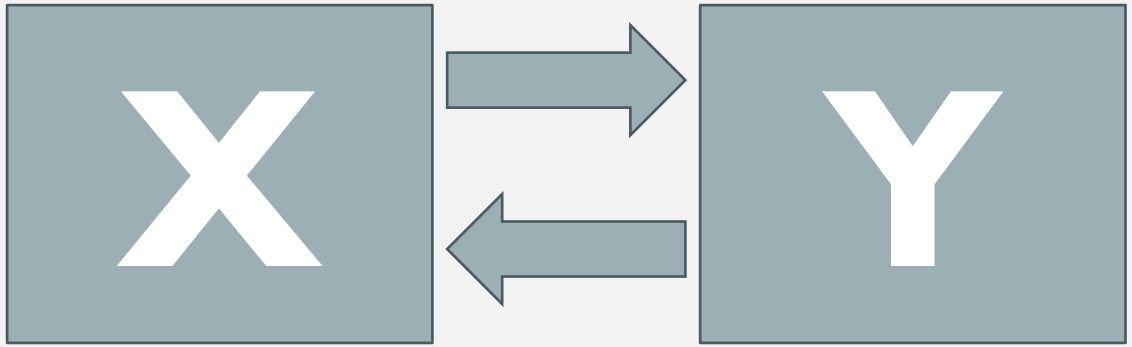
Deterministic

- if X occurs, Y occurs with certainty
- common in physical sciences (e.g. laws of gravity)

Probabilistic

- if X occurs, Y is *more likely* to occur
- arguably true of all causal relations: we simply can't observe all the factors, so it "looks" probabilistic

ENDOGENEITY



X and Y cause each other, though may not be simultaneous in practice

- Democracy and development
- Stress and illness
- Grade and how much you like a course

"Reciprocal" causality

CAUSAL CONDITIONS

Necessary Condition

- X is necessary for Y
- X must be there for Y to happen
- If X is not there, you will not see Y

Sufficient Condition

- X is sufficient for Y
- Every time you see X, you will see Y
- If X then Y, but Y can occur without X

Being a US citizen is a *necessary* but not sufficient condition for registering to vote.

Birth in the US is a *sufficient* but not necessary condition for being a US citizen.

NECESSARY VS. SUFFICIENT: EXAMPLES

Showing up for the exams in this course is a necessary but not sufficient condition for earning an A

Getting a perfect score on all assignments is a sufficient but not necessary condition for earning an A

Completing all of NYU's requirements is a necessary and sufficient condition for getting a degree

A college degree is neither necessary nor sufficient for getting a job

A job is neither necessary nor sufficient for happiness

NECESSARY VS. SUFFICIENT: EXAMPLES

Showing up for the exams in this course is a [ ] but not [ ] condition for earning an A

Getting a perfect score on all assignments is a [ ] but not [ ] condition for earning an A

Completing all of NYU's requirements is a [ ] and [ ] condition for getting a degree

A college degree is neither [ ] nor [ ] for getting a job

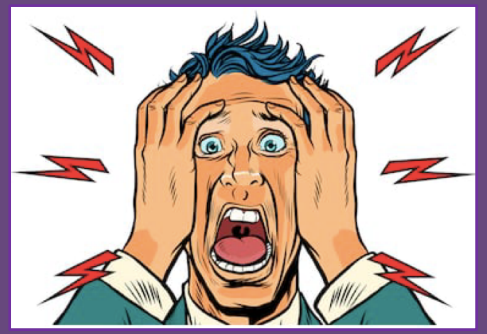
A job is neither [ ] nor [ ]for happiness

SUMMARY

- Thinking like a scientist:
  - Asking good questions
  - Following the scientific method: theory, hypothesis, data, conclusions.
  - Iterative process
- Causality: The goal of most scientific exercises
  - is hard! Think counterfactually.
  - Experiments good, but often not available.
  - Many kinds and directions and causal relationships
  - Thinking of conditions of causality can be useful

Outline

1. How to complete & submit all assignments in this course
2. Key concepts in causality
3. Principles & types of causality



Thank you!