Lecture 1: Setting our environment and parsing text files

Instructor: Michela Taufer

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Building our motivation

- Intel's Genevieve Bell shows that we have been dealing with big data for millennia, and that approaching big data problems with the right frame of reference is the key addressing many of the problems we face today from the keynote of Supercomputing 2013:

  https://youtu.be/CNoi-XqwJnA

- To-do task:
  - List three key concepts you learned by watching the video

Set up your environment

- Follow the steps in the file StartHere. We provide you with two different versions (i.e., Jupyter and pdf version)
  - **StartHere.ipynb**
  - **StartHere.pdf**
- We use git & GitHub to distribute & collect assignments as well as other class materials (e.g., slides, code, and datasets)
- We use Jupyter for our assignments and project
- We will use XSEDE Jetstream as our platform for assignments and project (we will introduce its use in the 3rd lecture)

GitHub and Git

- **GitHub:** web-based hosting service for version control used to distribute and collect assignments as well as other class materials (e.g., slides, code, and datasets)
- To-do list:
  - Create your own GitHub account
  - Send your GitHub username to taufer@utk.edu and jteague6@vols.utk.edu
- **Git:** software used by GitHub
- To-do list:
  - Install Git on your laptop

Open your assignment directory

Search or jump to... / Pull requests Issues Marketplace Explore

CISC879-BigData / 2018F-CS594-CS690 Private

Unwatch 5 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Branch: master 2018F-CS594-CS690 / Assignment01 / Create new file Upload files Find file History

imnasnainaec StartHere edits. Latest commit 7873682 3 days ago

..		
images	StartHere edits.	3 days ago
Assignment01.ipynb	Initial commit	4 days ago
StartHere.ipynb	StartHere edits.	3 days ago
StartHere.pdf	StartHere edits.	3 days ago
data.csv	Initial commit	4 days ago
data.tsv	Initial commit	4 days ago

Python and Anaconda

- **Python:** It is Python 3.6!
- **Anaconda:** Python distribution that includes many popular packages by default and makes installing additional packages easy
- To-do list:
  - Install Anaconda

Jupyter

- **Jupyter:** Our notebook for data analytics
- Programming in a browser
  - Create code in a cell – code in edit mode
  - Run code in a cell – code in command mode
  - Write text before and after code cells – markdown
- To-do list:
  - Start Jupyter from Anaconda GUI or command line

Open your assignment directory

jupyter

Quit | Logout

Files | Running | Clusters

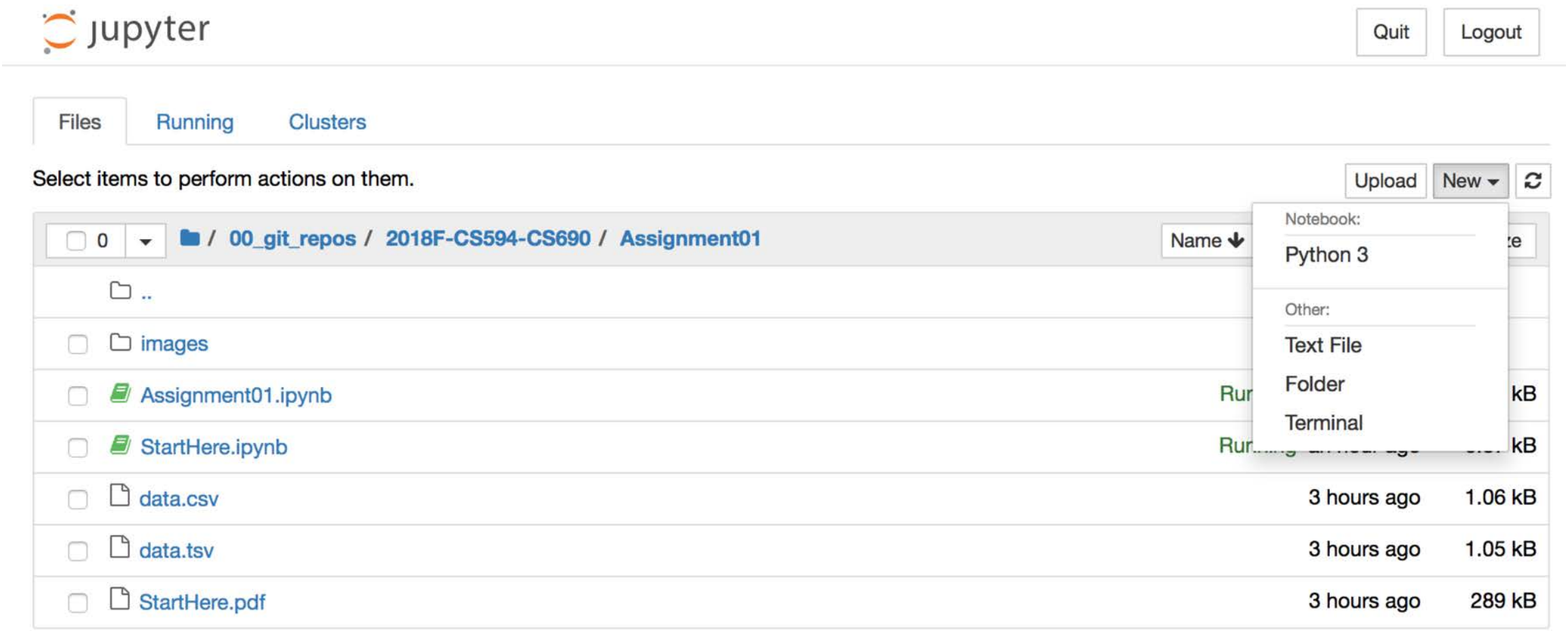
Select items to perform actions on them.

Upload | New

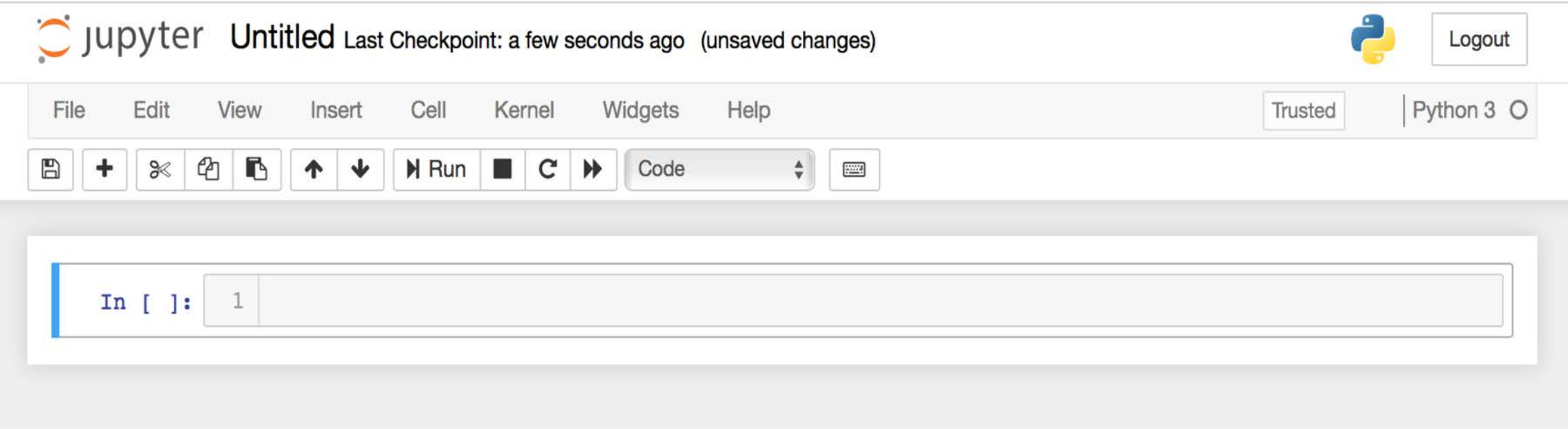
0 / 00_git_repos / 2018F-CS594-CS690 / Assignment01

	Name	Last Modified	File size
	..	seconds ago	
	images	3 hours ago	
	Assignment01.ipynb	Running 3 hours ago	11.1 kB
	StartHere.ipynb	Running an hour ago	6.87 kB
	data.csv	3 hours ago	1.06 kB
	data.tsv	3 hours ago	1.05 kB
	StartHere.pdf	3 hours ago	289 kB

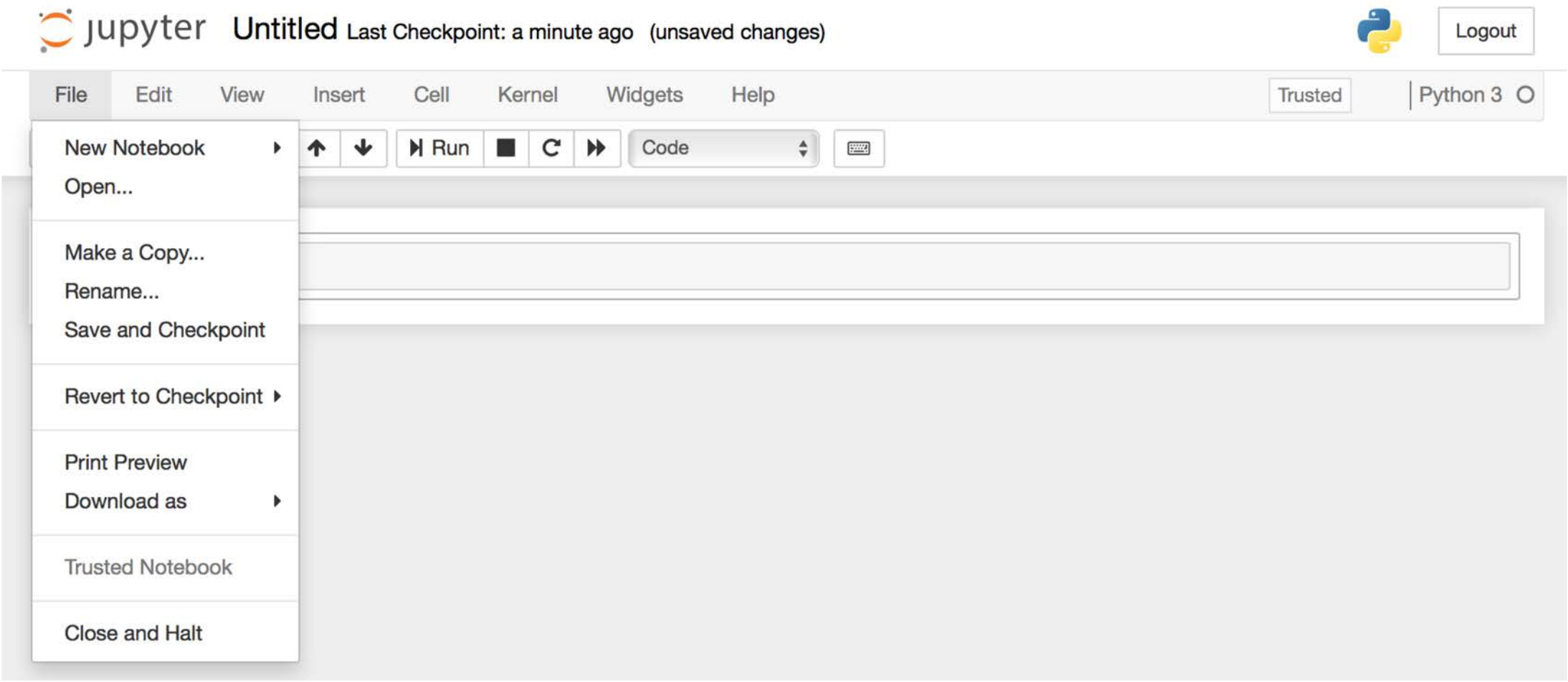
Create your code



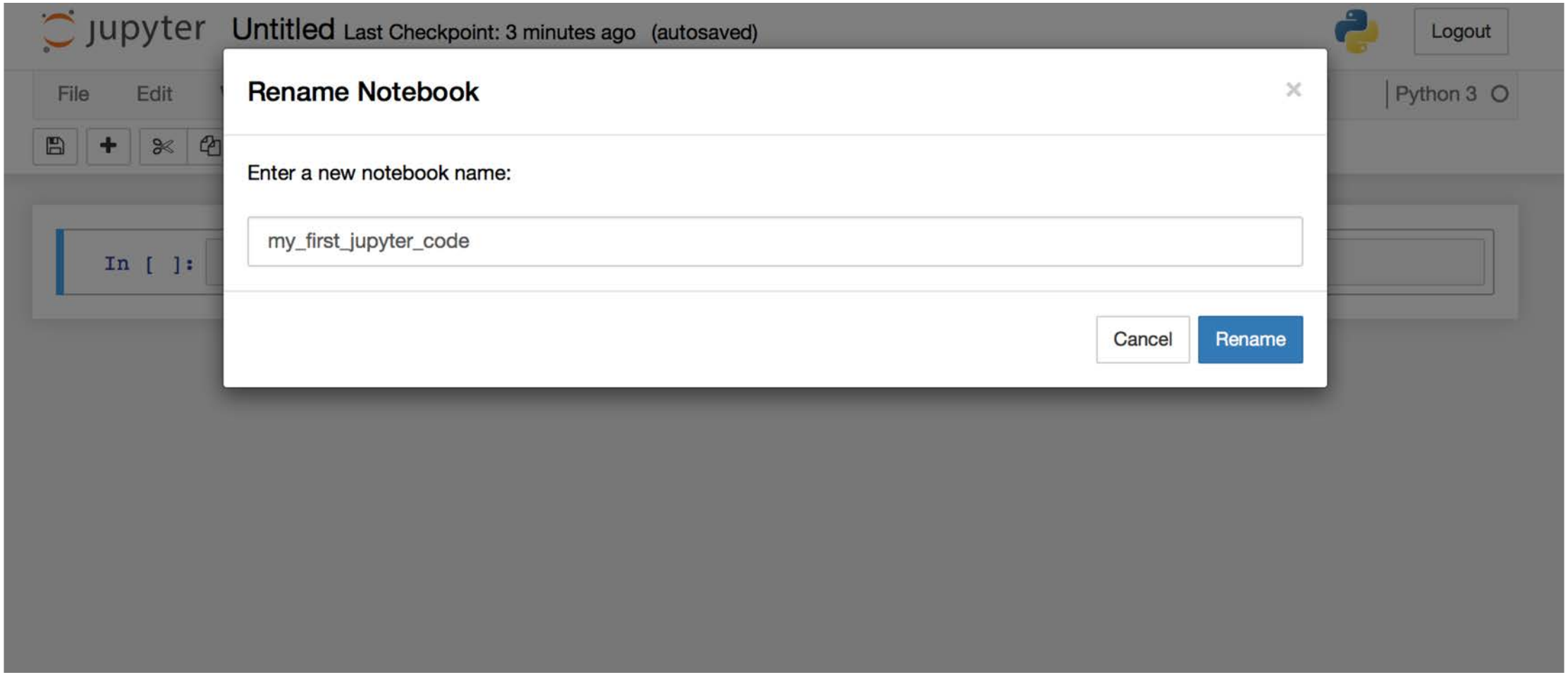
Create your code



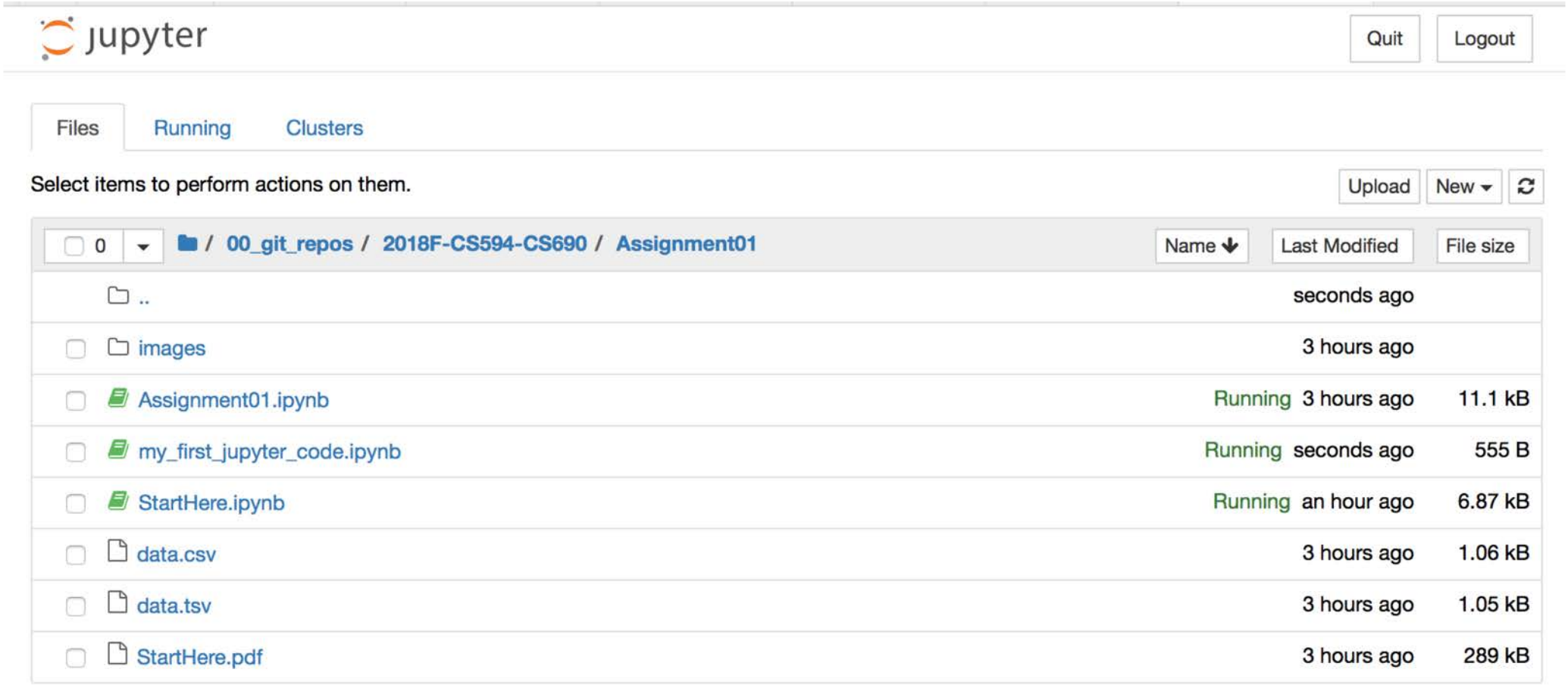
Create your code



Rename the file



Create your code



Create a cell

jupyter my_first_jupyter_code Last Checkpoint: a minute ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [ ]:

```
from datetime import datetime
print("Hello World! Right now, it is {}.".format(datetime.today().strftime("%c")))
```

Run your code

jupyter my_first_jupyter_code Last Checkpoint: 2 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [1]:
```
from datetime import datetime
print("Hello World! Right now, it is {}.".format(datetime.today().strftime("%c")))
```

```
Hello World! Right now, it is Sat Aug 25 13:47:49 2018.
```

In [ ]:
```

```

Propagations

jupyter my_first_jupyter_code Last Checkpoint: 28 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Notebook saved Trusted Python 3

Run Code

```
In [2]: a = 10

In [3]: print (a)
10

In [6]: b = 20

In [7]: print (a, b)
10 20
```

Add text to your notes

jupyter my_first_jupyter_code Last Checkpoint: 32 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

```
In [2]: a = 10

In [3]: print (a)
10

In [6]: b = 20

In [8]: print (a, b)
10 20

I cannot run!
```

Add text to your notes

jupyter my_first_jupyter_code Last Checkpoint: 33 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

```
In [2]: a = 10

In [3]: print (a)
10

In [6]: b = 20

In [8]: print (a, b)
10 20
```

I cannot run!

```
In [ ]:
```

What we will learn today

- Reading in, parsing, and processing delimiter-separated values stored in files – **comma-separated values (csv)** and **tab-separated values** (**tsv**)
  - Count (and print) the number of rows of data (header is excluded) in the csv file
  - Count (and print) the number of columns of data in the csv file
  - Calculate (and print) the average of the values that are in the "age" column - You can assume each age in the file is an integer, but the average should be calculated as a float

What we will learn today

- Converting the unicode-formatted names into ascii-formatted names
  - Use this dictionary to convert the unicode strings to ascii

The first assignment

jupyter Assignment01 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Markdown

CS 594 / CS 690 - Assignment 01

August 27, 2018

For this assignment, you must work in groups of one or two students. Each person is responsible to write their own code, but the group will (together) discuss their solution. In this notebook, we provide you with basic functions for completing the assignment. *You will need to modify existing code and write new code to find a solution*. Each member of the group must upload their own work to GitHub (which we will cover in the next lecture).

Problem 1

In this problem we will explore reading in and parsing delimiter-separated values stored in files. We will start with comma-separated values and then move on to tab-separated values.

Problem 1a: Comma-Separated Values (CSV)

From Wikipedia: In computing, a comma-separated values (CSV) file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.

If you were to consider the CSV file as a matrix, each line would represent a row and each comma would represent a column. In the provided CSV file, the first row consists of a header that "names" each column. In this problem, ...

For the next week ….

- Get your solution done before our next lecture
  - This week we are not pushing the solution into your own repos yet

- Next week ….
  - More about private GitHub repos
  - Pull your solution into your GitHub
  - More practice with Jupyter, Python, and problem solving