Where to Find Public Data Sets

And a (brief) Introduction to Google Colab

codecademy

NYC

CHAPTER

What Are We Going to Do Today?

- Where to find public data sets that you can download
- Navigate downloading data, tips for cleaning the data
- Getting data into Google Colab/brief intro to using Colab
- At the end of today, my hope is that you will feel confident enough to venture out on your own to start your own projects.

Why Are We Here?

- You’ve taken a course or a bootcamp and done some Data Science or Data Analysis projects within a LE (learning environment) and you feel pretty confident there.

So, the question becomes: *Now what??*

- Well, don’t panic! You can take the skills you’ve learned and use them off -platform.

- And then you can start adding projects to your GitHub repository or your online portfolio (which potential employers like!)

Begin With: What Interests You?

It starts with ***you!***

- You have a general curiosity (or else you wouldn’t be involved with data)
- Is there a research question or a subject that you’re interested in investigating?
    - I find that we learn best when we focus on subjects that matter to us.
- Population demographics? Sports? Voting? CO2 Emissions? The possibilities are endless.

It’s Out There!

Seriously, there’s SO much public (& free) data out there.

Any basic Google search can be a bit overwhelming. Here are a few places where I look for inspiration or if I have a data question in mind:

Where to look - some ideas:

US Census.gov

NYCOpenData

Kaggle

Baseball Reference

Cleaning Data

- Once you’ve downloaded it you will probably have to clean it up a bit before you load it into Colab.

- You can do so however you’re most comfortable. In Sheets or Excel*, or with RegEx or Pandas in Python or with a free online tool called OpenRefine.

  *If you’re on a Mac & don’t have Excel, get a Office 365 subscription (~$7/mo)

- Once it’s cleaned up, you’re going to want to remember the file path where you saved it so you can load it into Colab.

What is Google Colab?

- It's an alternative to using Jupyter Notebook that is cloud-based and based on Jupyter. (You can write, execute & present Python code in it. (One isn't better than the other. It's good to have options & know what tools are available).

- Do you have a Google Drive? It's an app, just like Sheets, Slides, or Docs that you connect to your Drive

- I like Google Colab because there's little set up. It operates like Jupyter but it's cloud-based so you can access it anywhere. (No, I don't work for Google!).
-
- You don't have to worry about packages or libraries or their installation. If you find that a package isn't installed, then it's just a matter of using:

```
!pip install package_name
```

- (See: Colab FAQ and: Colab Intro notebook and Colab Documentation)

Add Colab to your Drive

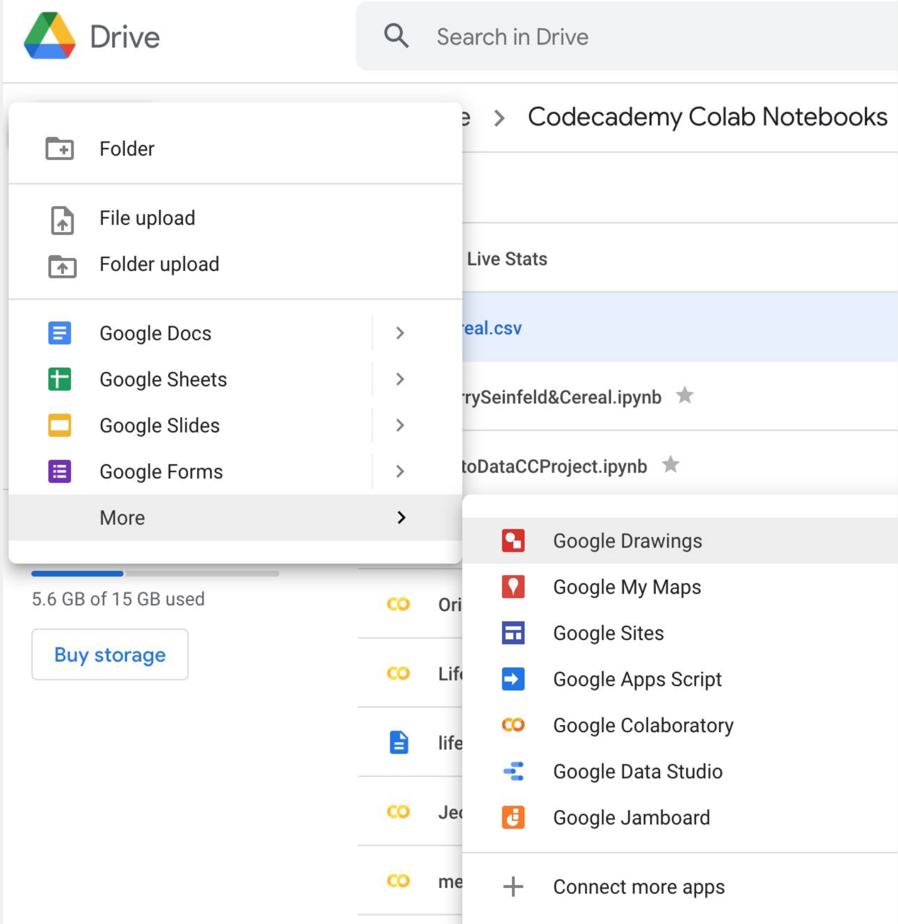
It’s an app!

1. Go to your Drive
2. Select “New”

If Colab isn’t in the dropdown menu, select

“Connect More Apps” and search for it

in their Marketplace & then add to your Drive.



Getting Data into Colab- a few ways

- Because Colab is in the Cloud there is no direct access to files located on your local drive.

- Upload from local drive, import the csv from GitHub or Mount your Drive to Colab

Resource1 GeeksforGeeks

Resource2 NeptuneAI

Resource3 Towards DS

There are a few ways and how you do it will affect who else can see your notebook if you share it on GitHub or in an online portfolio.

For ex:

```
from google.colab import files

uploaded = files.upload()
```

Getting data into Colab con’t.

Grab the file from a GitHub repo:

1. Click on the file in the repo
2. View raw file
3. Copy the raw file address
4. Use that URL as the location of your file

Ex: df =
pd.read_csv(“https://raw.githubusercontent.com/maeve70/random_stuff/main/cereal.csv”)

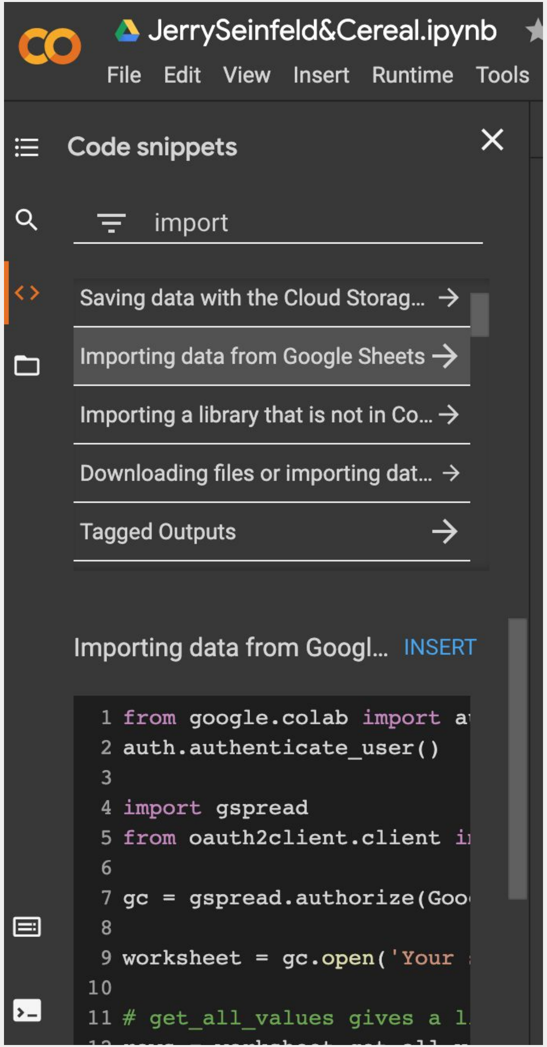
Google Colab

Colab documentation is pretty thorough.

You can customize it (auto-complete code, dark/light mode)

They even have popular code snippets/shortcuts

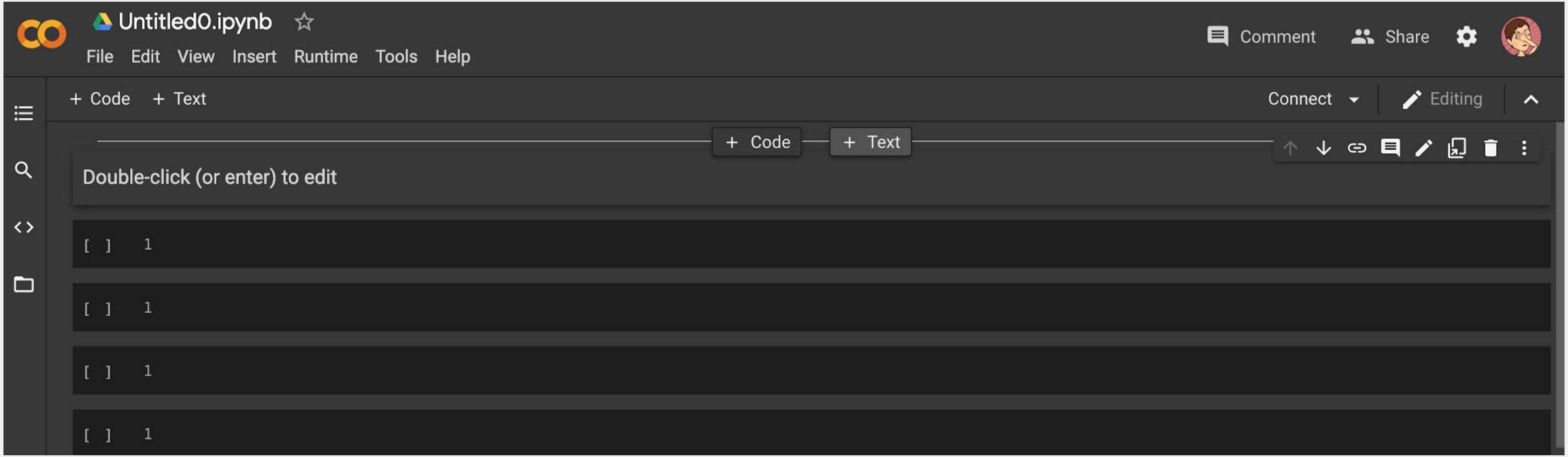
(left pane, under </>) you can use in your notebook.


JerrySeinfeld&Cereal.ipynb
File
Edit
View
Insert
Runtime
Tools
Code snippets
import
Saving data with the Cloud Storag...
Importing data from Google Sheets
Importing a library that is not in Co...
Downloading files or importing dat...
Tagged Outputs
Importing data from Googl...
INSERT

```
from google.colab import a
auth.authenticate_user()

import gspread
from oauth2client.client i

gc = gspread.authorize(Goo

worksheet = gc.open('Your 

# get_all_values gives a l
```

New Notebook

Woah! A blank notebook!



Double-click (or enter) to edit

Notebook Example Time

<do a demo with a data set!>

- I downloaded a .csv from Kaggle (cereal.csv) and loaded it into Colab

Wrap Up

- Start with a data question
- Find a data set that interests **you**
- Do some exploratory data analysis & go from there to do hypothesis testing, predictive analytics, machine learning, create a web scraper, and more

I hope that this has made it less scary/intimidating!

I always love to see what other people are working on!

Send me links to your projects! Or, if you want feedback, let us know! We will be happy to help. You can contact us via our Community Page or on Discord.

Thank you!!! :)

Additional Links, Classes, Webinars

The Census is a great repository of information. They also have webinars on how to navigate their datasets:

U.S. Census Webinars

There's also U.S. government data here:

data.gov

If you're interested in learning how to use Data Viz software, like Tableau, they also have (free) courses and a *free* public version:

Tableau

Public Tableau

Even More Links!

Colab Docs

public APIs

European Union Open Data Portal

U.S. Energy Information & Administration

California Open Data Portal

CA HHS Open Data Portal

Open Data UK

Covid Tracking Project

MTA Subway Ridership data

BART Ridership Reports

San Francisco Open Data

London Open Data Portal

The World Bank Open Data

Open Data India

United Nations Open Data

Towards Data Science Blog on Medium

Met Museum API Docs

Random Items

Excel & Census Data Webinar

Microsoft 365 Subscription (for Mac users)

SQLAlchemy- DB toolkit for Python

Psycopg2 (use with SQLAlchemy)

Python In Plain English- blog on Medium

DISCOUNT CODE (Codecademy PRO): chapter-member

Thank you!

You have reached the end of the line, please disembark the subway..

Thank you for coming to my Ted Talk. :)

Now, go out there and make something!!