DSCI-549: Introduction to Computational Thinking and Data Science

USC Viterbi School of Engineering

Homework 2

Communication

Please contact TA **Yashika Goyal** (yashikag@usc.edu) for this homework and include "**DSCI549**" in the subject line of your email.

Assignment

The purpose of this homework is to get familiar with Jupyter Notebooks and Google Colab, and run simple data analysis. You must complete the assignment using the Jupyter Notebook. Answer the questions in a separate word document, along with relevant screenshots. Then submit it on D2L.

For this assignment, you will be using four notebooks. You can access them here: https://github.com/doctorningwangusc/DSCI549/tree/main/Homework%20Assignments/Assignment2_CaesarCypher_SimpleMath%26Statistics

1. Getting data in and out of Colab: Notebook1_Importing_Data_to_Colab (25 points)
   a. Download the dataset from the internet using the wget function and take a screenshot of your Colab directory with the appropriate file. [Include two screenshots: One displaying the wget cell output & other of the Colab directory with the downloaded file present in it.]
   b. Follow the instructions to upload the Forest Fires dataset from your local machine to Colab and take appropriate screenshots of the process.

2. Running functions inside Jupyter: Notebook2_Running_Functions (25 points)
   a. Enter a text of your choosing when prompted. You may copy and paste any text. Take a screenshot of the resulting text with Line Breaks.
   b. Run the Caesar Cypher. Take a screenshot of the ciphered text.

3. Summarizing data: Notebook3_Summarizing _Data (25 points)
   a. When prompted, enter a list of 20 numbers. [Provide a screenshot]
   b. Take a screenshot of the summary statistics for your dataset. What do Q1 & Q3 represent?
   c. Generate a boxplot for your data and include the plot here. How do you interpret boxplots in the context of the summary statistics from part b?

4. Visualizing data: Notebook4_Visualiazing_Data (25 points)
   a. When prompted enter a list containing at least 20 scores for the "fictitious" class and 5 bins. Include the resulting histogram here.
   b. Rerun the cell with a varying number of bins. What's the optimal number of bins considering the size of your dataset? Why? Include the histogram here.

IMPORTANT NOTES

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in the Student Handbook https://policy.usc.edu/studenthandbook/. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, https://policy.usc.edu/research-and-scholarship-misconduct/.

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the American Language Institute http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students.

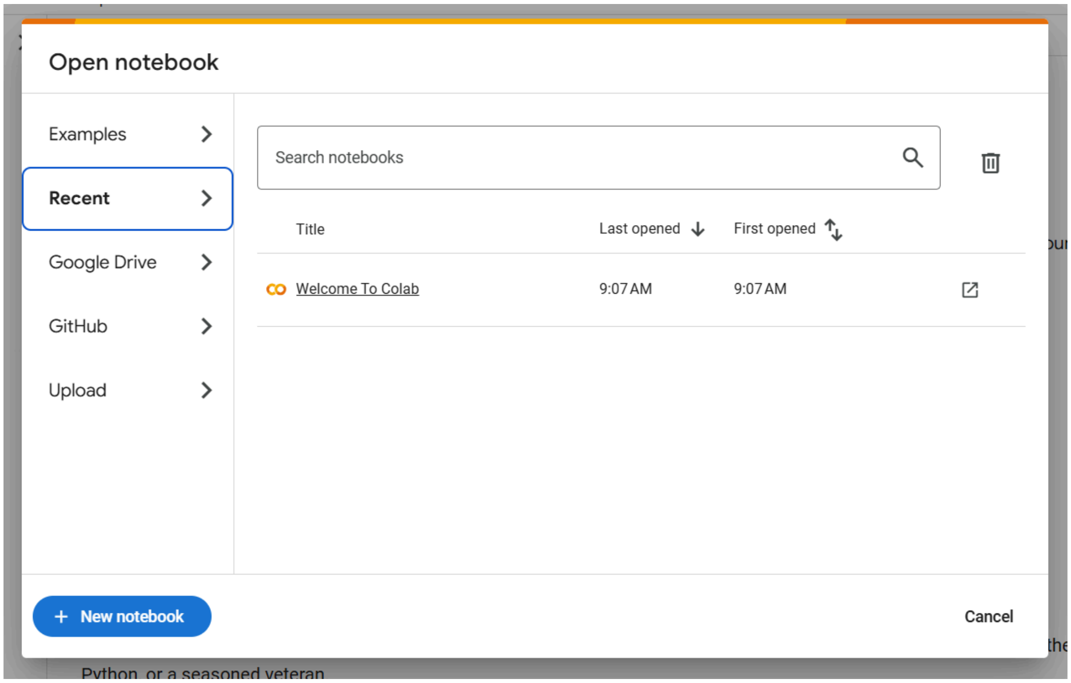
For more information, see the class syllabus and the USC web site.

Getting Started with Google Colab

1. To use Colab you should be logged in to your Google account. After this, click on the following link - https://colab.research.google.com

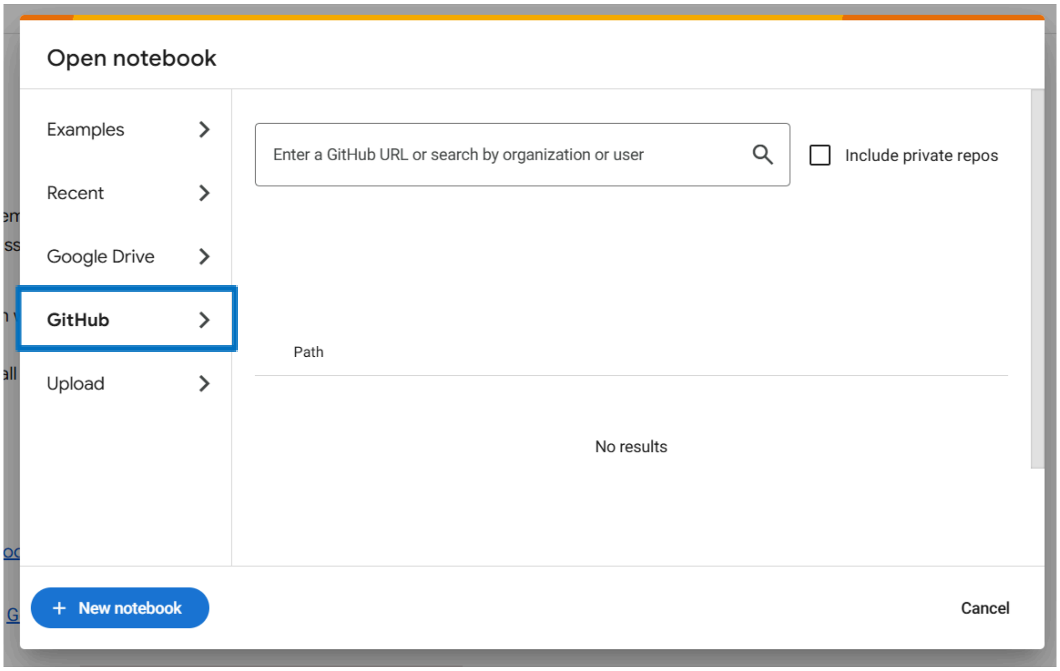
2. You will see the following tab pop up when you open Colab.

   [Note: If this tab doesn't pop up, go to File → Open Notebook]

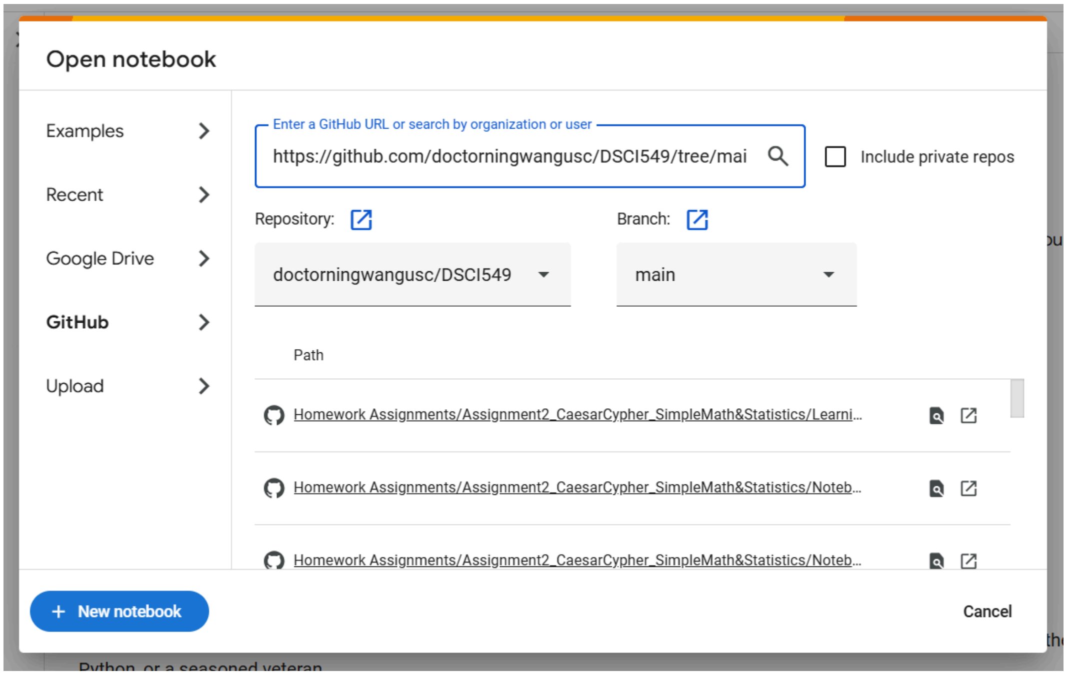


Title	Last opened ↓	First opened ⇅
Welcome To Colab	9:07 AM	9:07 AM

3. You will be using Notebooks stored on a GitHub repository.
   Click on the GitHub link.



4. Paste the following URL on the search bar: https://github.com/doctorningwangusc/DSCI549/tree/main/Homework%20Assignments/Assignment2_CaesarCypher_SimpleMath%26Statistics and click enter. This link should allow you access to all the notebooks on the repository. Scroll down and select the notebook you need to complete your homework. [You should be redirected to the Jupyter Notebook.]



Enter a GitHub URL or search by organization or user: https://github.com/doctorningwangusc/DSCI549/tree/mai

Repository: doctorningwangusc/DSCI549

Branch: main

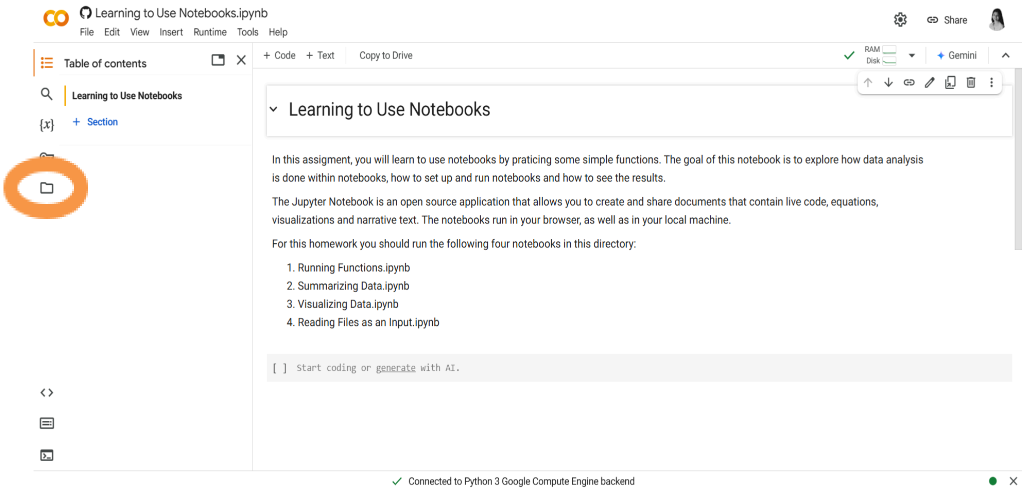
Path

Homework Assignments/Assignment2_CaesarCypher_SimpleMath&Statistics/Learni...

Homework Assignments/Assignment2_CaesarCypher_SimpleMath&Statistics/Noteb...

Homework Assignments/Assignment2_CaesarCypher_SimpleMath&Statistics/Noteb...

5. Click on the folder icon and wait a few seconds for the runtime to connect.


Learning to Use Notebooks.ipynb
File Edit View Insert Runtime Tools Help
Share
Table of contents
Learning to Use Notebooks
+ Section
+ Code + Text
Copy to Drive
RAM
Disk
Gemini

Learning to Use Notebooks

In this assigment, you will learn to use notebooks by praticing some simple functions. The goal of this notebook is to explore how data analysis is done within notebooks, how to set up and run notebooks and how to see the results.

The Jupyter Notebook is an open source application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. The notebooks run in your browser, as well as in your local machine.

For this homework you should run the following four notebooks in this directory:

1. Running Functions.ipynb
2. Summarizing Data.ipynb
3. Visualizing Data.ipynb
4. Reading Files as an Input.ipynb

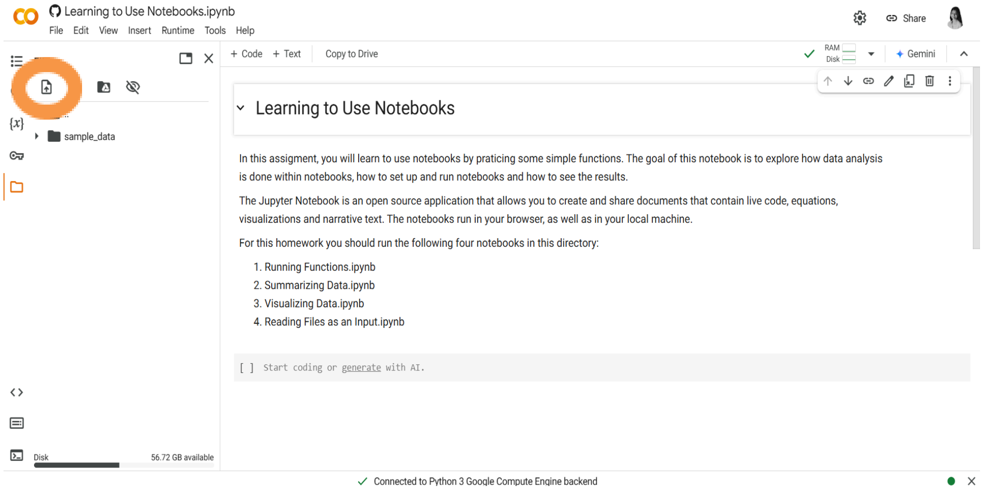
```
[ ] Start coding or generate with AI.
```

Connected to Python 3 Google Compute Engine backend

6. The sample data folder contains datasets populated automatically by Google Colab. **DO NOT USE FOR YOUR HOMEWORK.**

   a. Some datasets are included as examples. To access them and download them into your Google Colab directory, run the corresponding cells in your notebook.

   b. If you want to upload a file from your local computer, click on ‘Upload to session storage’ in the menu bar.



Learning to Use Notebooks

In this assigment, you will learn to use notebooks by praticing some simple functions. The goal of this notebook is to explore how data analysis is done within notebooks, how to set up and run notebooks and how to see the results.

The Jupyter Notebook is an open source application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. The notebooks run in your browser, as well as in your local machine.

For this homework you should run the following four notebooks in this directory:

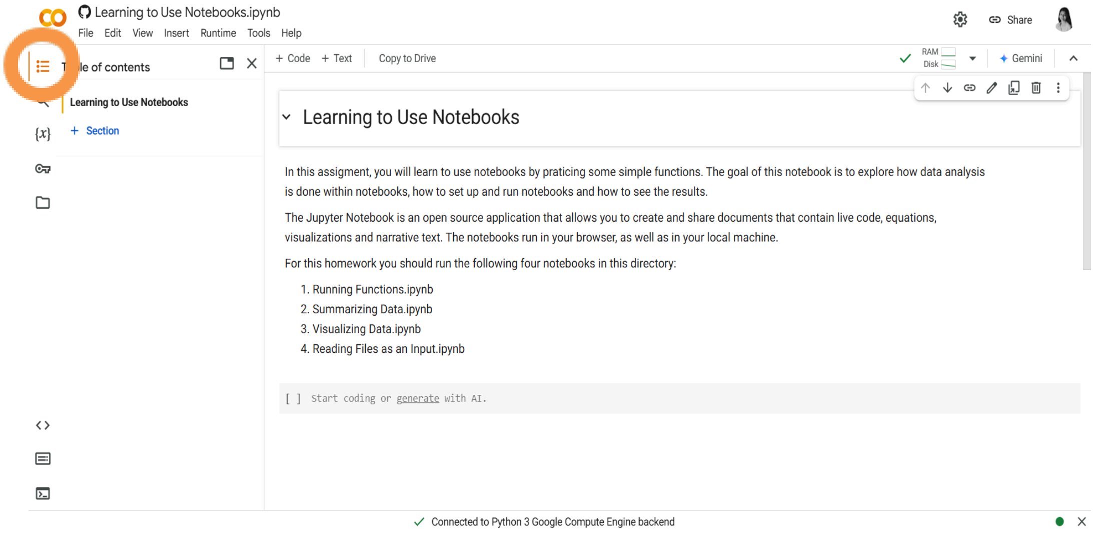
1. Running Functions.ipynb
2. Summarizing Data.ipynb
3. Visualizing Data.ipynb
4. Reading Files as an Input.ipynb

```
[ ] Start coding or generate with AI.
```

Connected to Python 3 Google Compute Engine backend

Note: Uploaded files are deleted after 12 hours. You might need to upload it again to use the file after 12 hours.

7. You can click on the Contents button to view and access any part of the notebook.


Learning to Use Notebooks.ipynb
File Edit View Insert Runtime Tools Help
Share
Table of contents
Learning to Use Notebooks
+ Section
+ Code + Text
Copy to Drive
RAM
Disk
Gemini

Learning to Use Notebooks

In this assigment, you will learn to use notebooks by praticing some simple functions. The goal of this notebook is to explore how data analysis is done within notebooks, how to set up and run notebooks and how to see the results.

The Jupyter Notebook is an open source application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. The notebooks run in your browser, as well as in your local machine.

For this homework you should run the following four notebooks in this directory:

1. Running Functions.ipynb
2. Summarizing Data.ipynb
3. Visualizing Data.ipynb
4. Reading Files as an Input.ipynb

```
[ ] Start coding or generate with AI.
```

Connected to Python 3 Google Compute Engine backend

8. Use the notebooks to answer the questions given in the homework assignment.