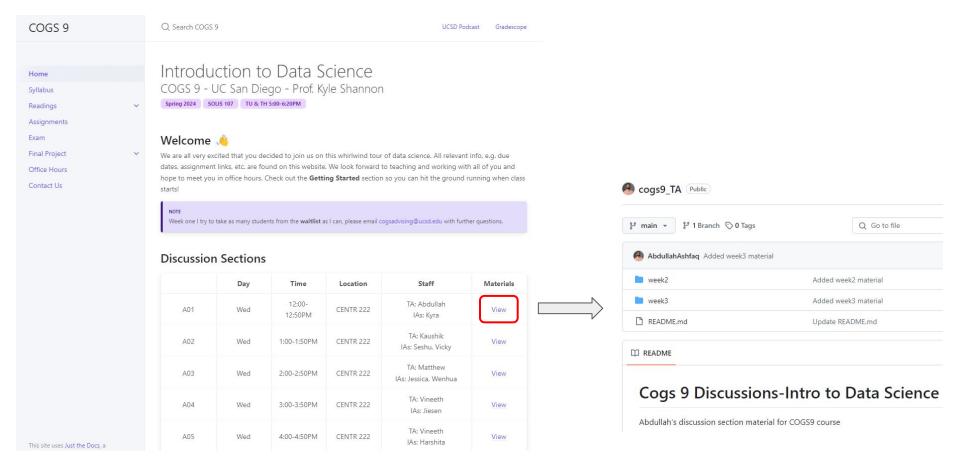
COGS9-Intro to Data Science

Spring24 - Prof. Kyle Shannon

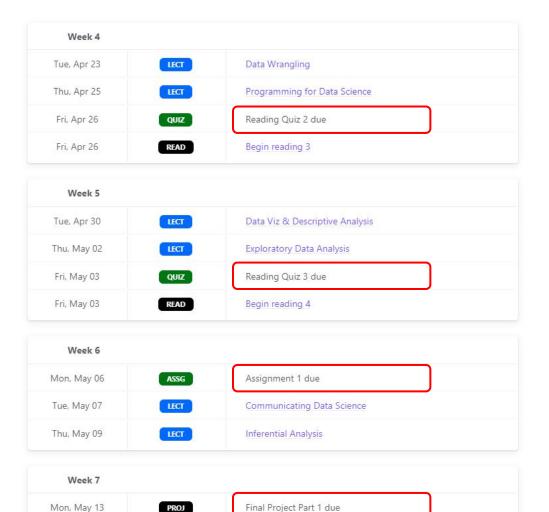
Discussion Section A01
Week 4

Teaching Assistant (TA): Abdullah Instructional Assistant (IA): Kyra

Where to find all material



Upcoming Deadlines



Discussion Sections Outline: Mostly Hands-on

- Week 2: Introductions, Making teams, Reading 1 (Part 1)
- Week 3: Reading 1 (Part 2), Python Basics with Jupyter Notebook
- Week 4: Reading 2, Getting data and wrangling it using Pandas
- Week 5: Reading 3, Assignment 1, Basics of SQL
- Week 6: Reading 4, Final Project Part 1 reviews/discussions
- Week 7: Assignment 2, Data Visualization and EDA demo
- Week 8: Assignment 3, Machine Learning demo
- Week 9: Reading 5, Closing thoughts
- Week 10: Final Project Part 2 reviews/discussions

Today's Outline

Participation = Extra Credit 😃

- Reading 2
- Getting Data
 - Downloading
 - o API
 - web scraping
- Data Wrangling in Python
 - Subset dataset
 - Change order
 - Add or modify column
 - Summarize data

Reading 2

Ethics, Privacy, Security

What is "PII"

"Any information relation to an ... natural person ... who can be identified, directly or indirectly, in particular by reference ... to one or more factors specific to his physical, physiological, mental, economic, cultural, or social identity."

Data Protection Detective

PII Privacy Protection Technologies

Examples: k-anonymity, I-diversity

"These methods aim to make joins with external datasets harder by anonymizing the identifying attributes."

The methods modifies quasi-identifiers to satisfy various syntactic properties to prevent.

Problem: Simply not enough to do the job

Re-identification without PII

It turns out there's a wide spectrum of human characteristics that enable re identification as long as they satisfy the following key properties:

- They are reasonably stable across time and contexts
- The corresponding data attributes are sufficiently numerous and fine-grained that no two people are similar, except with a small probability.

Re-identification algorithms take advantage of such properties to re-identify individuals using other attributes

Differential Privacy

Differential privacy (DP) is a system for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset.



Go to Notebooks