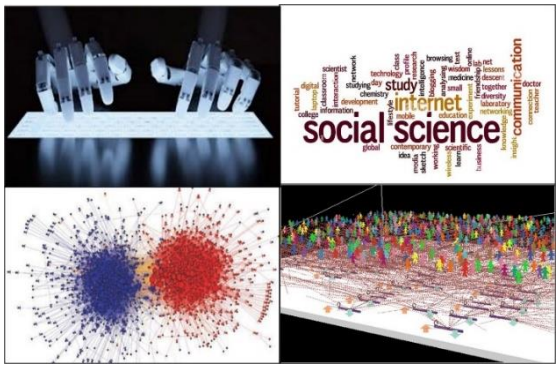
National Chengchi University, College of Social Sciences

IDAS, Course Number: 265773001

Data Science for Social Science Research



Office Hours	Tues & Thurs 2-4pm	**Class Time**	Wednesday 2-5pm
Office Location	綜北 13 樓	**Classroom #**	綜合院館 271215
Instructor	Dr. Jacob Reidhead	**Websites**	Moodle Google Drive
Email	reidhead@g.nccu.edu.tw		

1. Introduction

The explosion of digital data represents an unprecedented opportunity for social research! However, fullly leveraging the information revolution requires not only downstream analytical skills, such as data visualization and modeling, but also critical upstream skills of data collection, cleaning, curation and wrangling. This course will cover the full stack of tools in the data science toolkit, but with particular, hands-on emphasis on the upstream skills.

- **Coding Basics & Research Design**
- **Data Collection**
- **Data Wrangling & Curation**
- **Data Cleaning**
- **Data Visualization**
- **Data Analysis**
- **Presentation of Findings**

This course was created with IDAS students in mind. PhD and MA students from other graduate programs are welcome to join. No prerequisites and no prior coding experience are required.

2. Intended Learning Outcomes

In terms of practical data science skills, you will learn:

1. Upstream skills: data collection, cleaning, curation, wrangling
2. Downstream skills: data visualization, analysis, presentation

In terms of applying data science to social science research, you will learn how to:

3. Enhance a research design with creative data and high-validity measures
4. Design a series of visualizations and analyses that support a logical argument

3. Course Activities

Activities

- In class, students will actively participate in **Lectures**.
- Out of class, students will work individually on **Weekly Tasks** and **Research Posters**.
- Out of class, students will work individually or in pairs on **Datathons**.

ILO-Activity Matrix

Course activities are designed to satisfy the following Intended Learning Outcome (ILO).

Activity	**ILO1 Upstream Skills**	**ILO2 Downstream Skills**	**ILO3 Enhance Research Design**	**ILO4 Design Series of Analyses**
1. Weekly Tasks	x	x		
2. Datathon	x	x	x	x
3. Research Poster	x	x	x	x

4. Online Organization in Google Drive

Prepare Google Drive & Google Colab

1. Students will need to create a course folder on their NCCU Google Drive and share this with the instructor by the second week of class.
2. Students will also need to apply for Google Colab on their NCCU Google Drive.

Submit Assignments in Google Drive

1. Students will code their assignments in Google Colab,
2. Save all files in an assignment folder in their shared Google Drive,
3. Write up their assignments in a Google Doc,
4. Include links to all assignment files in that Google Doc,
5. And share this Google Doc with the instructor by each assignment deadline.
6. The instructor will need to be able acess to each student's shared Google Drive folder in order to open the links.

Access Course Materials in Google Drive

The instructor will also share a course folder with all students. The folder will include:

- Syllabus and Grades
- Lecture Slides and Worksheets
- Texts and Assigned Readings
- Sample Code and Data
- Instructions for all Assignments
- Links to Video Tutorials

5. Assessments & Grades

Grading Rubric

Task	Points per Assessment	Percent of Semester Grade
Attendance	After TWO FREE ABSENCES, each unexcused absence is -5%	0%
Tasks	10 tasks * 2% each	20%
Datathons	4 challenges * 10% each	40%
Research Poster	1 poster & presentation	40%

Weekly Tasks - 20%

A data science task will be assigned after each week's lecture. The instructor will provide sample code from the lecture and students will have one week to adapt the code and complete the task.

A complete submission will include (1) the student's code, (2) any output or analysis, (3) a one-paragraph summary of how the task was executed, and when applicable, (4) one or more paragraphs summarizing findings. Tasks will be submitted via Moodle.

Each task will be graded out of 2 points. The final task score will sum the top 10 scores out of a total of 20 points.

Datathons – 40%

Four datathon will be conducted throughout the semester. Each datathon will test a set of skills learned over the prior 3-4 weeks. The parameters of each datathon will be introduced at the end of a class period, and students will have one week to complete the challenge. Students may complete the datathon individually or in groups of two.

A complete submission will include (1) the student's code, (2) any output or analysis, (3) a one-paragraph summary of how the datathon was executed, and when applicable, (4) one or more paragraphs summarizing findings. All materials will be submitted via Moodle.

Each datathon will be scored out of 10 points.

Research Poster & Poster Session – 40%

Over the course of the semester, students will collect, clean, curate and analyze a dataset of their choosing in order to examine a social science research question. At the end of the semester, students will present their data and findings in a poster. Posters will have sections on background, the student's theory and hypotheses, data collection process, descriptive analysis, inferential analysis to test hypotheses, discussion and conclusion.

Three weeks before the poster session, students will submit a draft of their poster and receive feedback from the instructor. Students will revise the poster based on feedback. They will then print their final poster and present it in an end-of-semester poster session.

Research posters and participation in the poster session will be scored out of 40 points.

6. Course Materials

Texts & Assigned Readings

- Required Text
    - Molin, S. (2021). *Hands-On Data Analysis with Pandas: A Python data science handbook for data collection, wrangling, analysis, and visualization*. Packt Publishing Ltd.
- Optional texts will be recommended in advance of specific tasks
- Weekly readings will be provided in advance of each week's lecture.

Course Google Drive

All course materials will be made available on a course Google Drive. Course materials include:

- Syllabus, Academic Calendar, Records of attendance and grades
- Lecture slides
- Texts and Assigned Readings
- Worksheets and URLs used for all in-class activities
- Video tutorials, sample code, data and instructions for all weekly tasks
- Instructions for datathons and research poster

7. Academic Policies

Grading Scale

The grading scale for this course follows the system typically used at NCCU.

Extra Credit & Revisions

I rarely offer extra credit. However, if extra credit is offered, it will not be arbitrarily offered to individual students, but systematically offered to all students equally.

If a class collectively performs poorly on a particular assignment, I may extend the deadline and offer students the opportunity to revise and resubmit their assignments.

Academic Integrity

NCCU requires all students to adhere to high standards of integrity in their academic work. No type of academic misconduct (including but not limited to plagiarism, cheating, or lying to the professor) will be tolerated in this class and may result in penalties including but not limited to scores of 0 on assignments and forfeiture of extra credit points. Instances of academic misconduct will be referred directly to the appropriate disciplinary committee. For full information on these matters, please refer to the NCCU catalog or official website.

Generative AI

Students are encouraged to use generative AI to augment any aspects of all assignments including literature reviews, coding, team videos and the research poster. If AI-generated results do not fully satisfy assignment criteria, some human intervention may be required in order to complete the assignment and receive full credit.

8. Course Schedule

Week	Date	Title	Topics
1		Role of Data Science in Social Science Research	
2		Coding I	data types, control structures
3		Coding II	functions, classes
4		File Management & I/O	os, dir, paths, file types: xlsx, csv, json, xml, txt
5		Data Collection	online APIs, web-scraping
Datathon I: Collect Raw Data Files from Internet			
6		Wrangling & Curation I	selection queries, simple covariates
7		Wrangling & Curation II	grouping, aggregate fxns and agg covariates
8		Wrangling & Curation III	joins, melt & pivot, indexes & look-up tables
9		Data Cleaning	recoding, missing values, errors, disambig.
Datathon II: Produce Clean, Structured Data Table from Raw Data			
10		Visualization I	univariate & bivariate graphs
11		Visualization II	multivariate graphs
12		Visualization III	dynamic graphs, dashboards
Datathon III: Produce Graphs from Clean, Structured Data			
13		Analysis I	descriptive statistics, t-tests of means & props
14		Analysis II	clustering, unsupervised labels as covariates
15		Analysis III	linear and logistic regression
Datathon IV: Describe and Model Clean, Structured Data			
16		Research Poster Session	