Data Science and Statistics for Social Sciences I
CS&SS/SOC/STAT 321
Winter 2024
Section Syllabus

Ramses Llobet
Ph.D. Student
Department of Political Science, UW

Section	Office Hours
AC :: MW 08:30 – 09:20 am (LOW 102)	In zoom, by appointment.
AD :: MW 09:30 – 10:20 am (LOW 102)	Email: *rllobet@uw.edu*

Overview. Sections are designed to complement lectures by reviewing theoretical concepts and learn computational skills in R. We meet twice a week, on Monday and Wednesday. The section contents are divided into modules dedicated to "best practices" in R programming, theory review, data wrangling, visualization, and statistical analyses in R, consolidating techniques learned in lectures and QSS tutorials while introducing new skills relevant to the course contents. All lab materials can be found in my course website, including Zoom recordings of each lab section.

Office hours. I will offer office hours by appointment in Zoom. Please note that it *may* take me up to 24 hours to respond to a student's email, so it is a good idea to plan ahead and email me in advance. When you email me, please include (1) the topic you would like to discuss and (2) your time availability for scheduling a meeting.

In addition, I can also offer spontaneous assistance in Slack, but please email me if you see that I take more than 24 hours in replying you in Slack. I may not answer emails sent past regular working hours, and I recommend you email me rather

than send me messages on the Canvas messaging system.

Participation through Slack. A portion of your final grade depends on section participation, which I will monitor through Slack. Slack, designed for team communication, collaboration, and project management, organizes communication into channels, facilitating real-time messaging.

In some quiz sections, we will use Slack for sharing `R` code answers and collaborative problem-solving during in-section data analysis exercises. Participation is also considered when addressing questions or bug errors shared in Slack, either by you or your peers (see blow). For final projects, students are encouraged to create private Slack channels to share files, documents, and `R` code.

Programming Assistance. Slack is the most preferred communication channel, which allows you to insert code block in your messages. It has the added benefit of facilitating knowledge spillover through peer discussion and mutual assistance. Please post your questions on **Slack** related to **R** programming, graphic packages, or debugging. When you post a question, the best practice is to create a *minimal, reproducible example*, instead of taking a screenshot of a code snippet (see here and here).

Alternatively, please feel free to set an appointment for office hours for further consultation, or email/'Slack' me your questions. Please note that if you want me (or someone else) to debug your code, you should first create a minimal reproducible example of the error(s) along with the required data file for code execution.

Homework Submission. Please submit your homework in PDF. You must use **RMarkdown** to integrate plain text, graphic outputs, and code chunks which can then be rendered ("knitted") into a single PDF output. You will have to submit your homework PDF on the Canvas course website (in assignments).

Section schedule. Below is the tentative schedule of sections and the associated topics to cover during lab sections, which are subject to adjustment depending

on our progress and learning needs:

Week	Monday	Wednesday
1	*No section*	Introduction; R setup
2	Project and file management	Rmarkdown and knitting PDF
3	logicals, subsetting, and NA data	Getting help and debugging
4	quantiles and distributions	factors, ifelse, and case_when
5	Intro to ggplot2, I	Intro to ggplot2, II
6	Causation in science, I	Causation in science, II
7	Inference: samples and populations	Bivariate regression, I
8	Bivariate regression, II	Uncertainty and hypothesis test
9	Multivariate regression, I	Multivariate regression, II
10	Selection bias and missing data	Within-group regression

Module Outline

Week 1 - 3, Module 1: Getting started with R/RStudio. This module offers an overview of basic R functions and introduces the R-Studio interface. It covers installing R/ RStudio and relevant packages, creating R projects, managing working directories, introducing R Markdown, and creating and sharing minimal, reproducible examples for programming assistance via Slack.

Week 3 - 5, Module 2: Data management and exploratory visual analysis. This module equips students with essential computing skills for data science to successfully complete course assignments and their final projects. Some of the topics include creating and manipualting data frames, logical tests, subsetting, NA data, pivoting and merging datasets, and intro to visualization with ggplot2.

Week 6 - 8, Module 3: Introduction to causal inference and linear models. This module introduces students to causal inference and the linear model. Topics include the distinctions between experimental and observational designs, prediction, the method of least squares, standard errors, and the interpretation of con-

fidence intervals. While theory will be reviewed, the approach will be predominantly computational and visual.

Week 9 - 10, Module 4: Hypothesis tests and multivariate regression analysis. This module begins by explaining statistical inference and the importance of expressing uncertainty in our inferences. We will cover the construction and interpretation of hypothesis tests relevant to research questions, along with topics in multivariate regression analysis, including transformations, nonlinear relationships, interaction effects, and the interpretation of categorical predictors.

Additional Resources. If you're looking for a dataset for your project, find below several repositories and databases for ideas on research questions and datasets for your final projects. Remember always to download the codebook, if available, to know in more detail what variables are included in a dataset.

1. **Kaggle**: A classic in computer and data science, Kaggle offers a vast repository of datasets covering various topics and applications. Additionally, it hosts open competitions with prizes of up to $50K, providing opportunities for data scientists to solve problems and share code and replication materials.
2. **Harvard Dataverse** and **TidyTuesday**: These repositories focus on more academically driven data projects and include replication materials.
3. **Penn World Tables**: If you prefer a macroeconomics-oriented dataset, the Penn World Tables is a valuable resource. Additionally, consider consulting international agencies like the **World Bank**.
4. **Correlates of War**: A famous database for those interested in conflictology and international relations.
5. **General Social Survey**: For those interested in sociology, opinions, and values in the United States.
6. Epidemology and Biostatistics: Check out the WHO datasets to an external site. in these fields if your interest lies in these areas.
7. **Comparative Political Dataset**: A classic for those interested in comparative politics at the country level.