**Data Science for Social Science Research***

**Spring 2023**

Instructor: Anustubh Agnihotri

## Introduction:

Data science techniques are now central to quantitative research in social sciences. From academics to public policy practitioners to non-profit organizations, all increasingly rely on data science techniques for cleaning and gathering data, conducting data analysis, and presenting the results. For example, multiple datasets often need to be transformed, aggregated, and merged before a more comprehensive analysis can be carried out. Researchers might need to create datasets from scratch by converting a large number of PDFs into machine-readable texts or scraping data from a website, or interacting with an API. These tasks require knowing, among other things – basic data wrangling concepts, techniques for cleaning messy datasets, and the ability to write code that can be replicated and shared with multiple collaborators. This course will provide students with the skills necessary for successfully carrying out quantitative social science research by deploying data science techniques.

The research questions students will engage with will be related to **the political economy of development**. For example, we will try to understand the relationship between the level of socio-economic development and political outcomes like incumbency and competition, examine the determinants of differences in the quality of public service delivery and local institutions, or explore how the background of individual bureaucrats shapes their career trajectories. These broad sets of relationships will be examined from different perspectives throughout the course. We will begin with basic descriptive statistics, then visualize the data in different ways (cross-sectionally, over time changes, using maps, or different plotting methods) while also touching upon establishing causality. The course will rely on the R statistical software (open source) using a predefined set of freely available datasets (more details below). Students will learn how to gather, aggregate, analyze, and interpret datasets, as well as present findings in an effective manner.

The course can be divided into two broad parts that have substantial overlap. The **first** part of the course will focus on the basics of data science – data wrangling, visualization, and analysis. These skills will teach students how to aggregate, transform, and merge datasets for a richer analysis. The **second** part will focus on data gathering and cleaning. In this section, the students will be creating new datasets by gathering data from websites and APIs and then cleaning and processing the data. The aim here is to expose students to the challenges of creating new datasets.

## Requirements

- **Students are not required to have prior programming experience or a background in computer science to take this course. The course will enable students to become proficient with programming in the R software over the course of the semester. Thus, students should be willing to spend time learning how to code.**
- **Students should have taken an introductory course on quantitative social science research methodology (This requirement is waived for final year and ASP students)**

## Objectives

By the end of the course, students should be able to

- Understand basic programming terminologies, structures, and conventions
- Write, execute, and debug R code for data analysis, and visualization
- Be familiar with regular expressions and basic data cleaning techniques
- Be familiar with the basics of web scraping and APIs
- Be able to gather, analyze, visualize large scale datasets
- Be able to use understand and communicate relationship between variables
- Do preliminary geospatial analysis
- **Overall goal of the course is to enabled students with the ability to gather new sets of data and pursue new research ideas**

## Final Project

The final project will be group-based, and students will be asked to work on a research question that can use the tools and techniques being used in the class and is broadly related to the political economy of development. Students will be (randomly TBD) assigned to a group, and there will be accountability measures to ensure that all team members contribute to the best of their ability. Students will work toward their projects throughout the semester and will be asked to present their ideas to their peers in the form of multiple presentations. A final report will be turned in along with the code repository so that it can be replicated by the instructor.

## Course Requirements and Grades

This is a graded class based on the following:

- Completion of assigned homework (individual) +
    - Tutorials from Data Camp (10%)
    - Home works (3): (35%)
- Participation - attendance, engagement, and lightning presentations in front of peers (10%)
- Final project
    - Proposal : 10%
    - Final Presentation: 15%
    - Final project report 20%

## Datasets

The first half of the course will use a set of common datasets and research questions that allow for the exploration of different questions around the political economy of development. These datasets will be used across multiple weeks across different components of the course to explore similar requisitions from different perspectives - descriptive statistics, visualization, and mapping hotspots. Based on these datasets, we will identify a common set of variables that will be used to illustrate the linkages between the skills being taught in the course and their applicability to social science research questions. For example, predictors of the quality of public services (Rural employment, MLA/MP Local Development Funds, Road Construction) or patterns of political

incumbency would be recurring research themes in the course and will draw upon different datasets.

1. TCPD MLA Level and IAS Data
2. SHRUG Data (Socio-economic development indicators at MLA and District Level)
3. IHDS I/II
4. NFHS Rounds
5. Pew Survey on Religious Attitudes in India
6. Other datasets from NDAP (Especially Public Service Delivery Outcomes)

## Syllabus

### Week 1: Introduction to Social Science Research, Data Science, and R – Part 1

- The students will be introduced to the basics of social sciences research with a focus on the rapid expansion in data availability. For students who have taken an introductory course on social science research methodology, this week will act as a refresher. For others, we will cover the three ways in which quantitative social science research is conducted – a) Prediction b) Description c) Causal Inference. We will go over the basics of research design and how novel datasets are shaping the questions being asked by researchers. The students will be introduced to the basics of running commands and scripts in R.
- Readings on social science research methods (TBD)

### Week 2: Introduction to Social Science Research, Data Science, and R – Part 2

- This week will be a continuation of the first week but with an emphasis on the practical application of the concepts. Thus, we will discuss the datasets that will be used throughout the course along with 2-3 important social science research questions. The students will learn how to load the datasets and do preliminary checks on the data. We will also introduce major libraries being used in R and their applications. No prior knowledge of R will be assumed. The week will end with a demonstration of how the datasets can be used to engage with the research questions using analysis in R.

- Readings on social science research methods (TBD)

**Week 3: Introduction to Data Wrangling**

- The week will introduce students to the commonly used techniques for merging, aggregating, and reshaping data. The students will be used to tidyverse (library in R) functions that allow students to easily combine datasets and aggregate them. We will combine different publicly available datasets together. A common set of research questions will guide the exercises. For example, we might combine the TCPD and SHRUG datasets to understand how economic growth impacts incumbent vote share.
- Readings
    - Cheatsheets Dplyr/Lubridate

**Week 4: Introduction to Data Analysis**

- The students will be introduced to commonly used data analysis techniques (regression models, frequency tables, summary statistics of variables) as well as ways to produce outputs that can be used to present this analysis. The same set of research questions from the previous week will be carried forward. The goal for this week would be to teach students how to conduct quantitative analysis and create a data flow that allows for easy replicability of results across multiple collaborators.
- Readings on political economy of development
    - Iyer and Mani
    - Pritchett

**Week 5: Introduction to Data Visualization – Part 1**

- The part of the two-week lecture will build upon the data wrangling and analysis exercises. The datasets that were created in the previous week will be visually represented by using the ggplot library in R. The students will be taught ways to visually describe variables as well as the relationship between multiple variables. The week will involve doing bar plots, box plots, and density plots (and other descriptive visualizations)

on the existing datasets. The emphasis will be on being able to meaningfully articulate the visualizations.

- Readings on data visualization (TBD)

**Week 6: Introduction to Data Visualization – Part 2**

- In the second part of the data visualization, we will look at describing more complex relationships and focus on visually representing bivariate and multivariable relationships. We will also explore how heterogeneity across subgroups can be visually represented. The exercises will have the same set of research questions from the previous week, but the plotting techniques will be more complex in terms of programming requirements.

**Week 7: Introduction to Data Cleaning**

- This week shifts the focus towards creating new datasets rather than using readily available datasets. Students will learn how to convert files from different sources (images and pdf) into readable text and use regular expressions to clean data. The focus will also be on automating the data cleaning to a large number of files to achieve scaling up of datasets. The final cleaned dataset will be analyzed and visualized.

**Week 8: Introduction to Web Scarping/APIs**

- Publicly available data often resides on websites. The students will engage with R programming techniques used to scrape data from websites. The scripts will enable students to loop over multiple elements of a website and create a dataset. The introduction to web scraping will be accompanied by a light introduction to the basics of html. A government website will be scraped and the data will be cleaned, merged with existing datasets, and analyzed. We will also briefly cover APIs or Application programming interfaces, which allow researchers to collect data as well as use existing resources to create new measures. For example, the Twitter API can be used to access

text-based data and the Google Maps API can be used to compute spatial distances between two points. This week will introduce students to APIs and explain how to engage with them.

**Week 9: Presentations: Project Proposals**

- The students will present their proposals (separately submitted) to their peers in the form of lighting talks.

**Week 10: Introduction to Geospatial Analysis I**

- The focus for this week will be on learning the basics of representing data spatially using the libraries in R. The students will learn how to project data onto maps and handle shp files within the R software. The maps will be used to understand both distribution of a particular variable as well as bivariate relationships.
- Readings on political economy of development (TBD)

**Week 11: Introduction to Geospatial Analysis II**

- The second well will involve adding new covariates to spatial datasets from other datasets. We will also use R to do spatial merges and aggregation of data. We will briefly touch upon using satellite imagery to add new covariates for social science research. These new covariates obtained from satellite images will be merged with existing data for analysis.
- Readings on political economy of development (TBD)

**Week 12: Concluding Week: Brief Introduction to Big Data and Machine Learning**

- This week will provide the students with an overview of how digital traces (mobile phones, sensors) are being used by researchers to create large datasets and the role of machine learning in enabling the interpretation of these large datasets. The lectures will

introduce students to the basics of machine learning algorithms and their applications. We will also cover some examples from text-as -data analysis and satellite image analysis that use supervised and unsupervised machine learning techniques.

- Readings on Big Data and Machine Learning (TBD)

**Week 13: Final Project Presentations**

- The final week will be dedicated to the presentations by students. Based on the feedback from presentations students will submit final project reports along with a reproducible code repository.

*This is a draft version. Final syllabus will be shared with students on the first day of class.