90872 - Using R for Policy Data Analysis
(6 units)

Adjunct Professor: M William Sermons (msermons@andrew.cmu.edu , 301-499-5018)
Carnegie Mellon University
Heinz College - School of Public Policy and Management
Fall 2024 Mini 1, Tuesday, 6:00 – 8:50 p.m.

Office Hours:

Professor Sermons will create an online sign-up calendar that allows students to schedule assistance during weekends and weekdays. You can also request appointments outside these times as needed.

Course Description/Objective:

Data analysis is an essential component of quantitative policy analysis. However, the focused application of statistical methods extends beyond what is covered in core methods classes such as Cost–Benefit Analysis (CBA) and Program Evaluation. This course teaches students how to apply a variety of data analysis techniques using R, a free open-source statistical and graphical analysis environment increasingly used by data miners and analysts.

A key focus of this course is the use of categorical variables, with a specific emphasis on documenting and quantifying disparities across demographic lines such as race, gender, and other important factors. Students will learn to apply these techniques broadly, with disparities analysis serving as a real-world example.

Class sessions will combine instruction on data analysis techniques, in-class application using R, and presentations by practicing policy data analysts. Applications will focus on analysis relevant to the social safety net, including cases that emphasize consumer protection, affordable housing, homelessness, and quantifying disparities.

By the end of this course, students will:

- Master the basics of R programming, including data types, structures, and manipulation.
- Develop skills in data visualization and statistical analysis using R.
- Learn to identify and quantify disparities across various demographic lines or other categorical variables.
- Gain experience in using AI tools for code generation and structured prompt engineering.

- Complete an independent data analysis project demonstrating the application of course concepts.

Text Materials:

There is no required textbook for this course. However, links to relevant videos and Data Carpentry pages will be provided via Canvas. Additionally, a course page created by the library with links to further materials can be accessed here: Heinz College: 90-872 Using R for Public Data Analysis.

Prerequisite Skills:

Students are expected to have completed a course in statistics and have a basic understanding of R. Familiarity with basic statistical concepts such as measures of central tendency, t-tests, analysis of variance, and regression analysis is required. For students who have not taken a statistics course with R, a review of the Introduction to R recording (passcode: 7#wQBTWQ) is recommended.

Weekly Schedule:

Each week, the course will include three types of activities: asynchronous pre-work, in-class activities, and homework assignments.

- **Asynchronous Pre-Work:** Focused on applying course concepts using R, students will review R Notebooks with explanatory text and executable R code before class.
- **In-Class Activities:** Weekly sessions will include mini-lectures, guest speakers, discussions, and demonstrations of R application challenges. Class time is from 6:00 to 8:50 p.m., though we may routinely finish early.
- **Homework Assignments:** These include replicating analyses demonstrated in the asynchronous R Notebook, participating in online discussions, and progressing on the course project.

Course Requirements & Grading:

Percent (%)	Assignment
50 %	Assignments, including weekly data analysis assignments
50 %	Final Project

Grading Scale:

A+	99.0-100%	B+	88.0-90.9%	C+	78.0-80.9%

A	94.0-98.9%	B	84.0-87.9%	C	74.0-77.9%
A-	91.0-93.9%	B-	81.0-83.9%	C-	71.0-73.9%

Attendance Policy:

Students are expected to attend all classes. One excused absence may be granted for illness, personal emergencies, or apprenticeship-related travel with prior arrangement. Unexcused absences will result in a 5% deduction from the final grade. Assignments due on the day of a missed class must still be submitted on time.

Cheating & Plagiarism:

Students are expected to honor the letter and the spirit of the Carnegie Mellon University Policy on Cheating and Plagiarism. All activities cited in that policy will be treated as cheating in this course. Students are expected to familiarize themselves with this policy. Students are also encouraged to review the Carnegie Mellon University Academic Disciplinary Actions Overview for Graduate Students, which details penalties and sanctions, as well as students' rights. I will take a zero-tolerance policy on cheating and plagiarism and will consult with departmental leadership on appropriate action for all instances of cheating and plagiarism. As the aforementioned policies indicate, penalties can include course failure, suspension, and dismissal from the program.

- Carnegie Mellon University Policy on Cheating and Plagiarism
- Carnegie Mellon University Academic Disciplinary Actions Overview for Graduate Students

Use of Generative AI:

Generative AI tools, such as ChatGPT and DALL-E, can be valuable for learning and productivity, including completing assignments, generating ideas, and personalizing your learning experience. However, proper citation and adherence to academic integrity guidelines are required.

In this class, you may use generative AI programs to:

- Brainstorm ideas for your course project.
- Evaluate, debug, and improve your R code.
- Research topics or generate different ways to discuss a problem.
- Review and edit written text for course assignments.

You may not use generative AI programs to:

- Generate content for assignments without proper quotation and citation.

- Produce unresearched bibliographies or other uncited content.
- Use generative AI content as your own without appropriate acknowledgment.

Important Considerations: It is important to recognize that large language models frequently provide users with incorrect information, create professional-looking citations that are not real, generate contradictory statements, incorporate copyrighted material without appropriate attribution, and can sometimes integrate biased concepts. Code generation models may produce inaccurate outputs. Image generation models may create misleading or offensive content.

While you may use these tools in the work you create for this class, it is important to understand that you are ultimately responsible for the content you submit. Work that is inaccurate, biased, unethical, offensive, plagiarized, or incorrect will be penalized.

Data Sets for Exercises and Projects:

In-class demonstrations will use the American Housing Survey. Students are required to select a public microsample database, such as the American Community Survey or National Health Interview Survey, for their assignments. The selected data source must be publicly accessible, use individual or household-level data, and have no restrictions on circulation.

Course Schedule:

Week 1: Course Overview and Introduction to R

- Topics: Course Overview, R Basics, Data Preparation with DPLYR
- Assignments: R Notebook on R Syntax and Data Types, Project topic and dataset selection

Week 2: Descriptive Statistics and Data Manipulation

- Topics: Descriptive Statistics, Data Manipulation with DPLYR
- Assignments: R Application Assignment, Online Discussion: Peer feedback on project topics

Week 3: Understanding and Analyzing Categorical Variables

- Topics: Documenting Differences Across Groups, Introduction to Categorical Variables
- Assignments: R Application Assignment, Hypotheses about group differences

Week 4: Data Visualization

- Topics: GGPlot for Visualizing Group Differences

- Assignments: R Application Assignment: Visualizing differences across groups

Week 5: Inferential Statistics and Significance Testing

- Topics: Significance Testing, OLS and Logistic Regression for Categorical Variables
- Assignments: R Application Assignment, Testing hypotheses

Week 6: Advanced Disparities Analysis

- Topics: Disparities Indices, Visual Representation of Disparities
- Assignments: R Application Assignment

Week 7: Multivariate Analysis and Reporting Results

- Topics: Multivariate Analysis, Reporting Results
- Assignments: R Application Assignment, Final Project Submission