ISS5066: Programming for Business Analytics (Graduate)

Jaewon Yoo

Assistant Professor
Institute of Service Science
National Tsing Hua University

Fall 2026

Course Duration: Sep. 2026 – Dec. 2026	*Classroom:* TSMC Bldg. R421
Class Time: Thursday, 14:20 pm- 17:20 pm	*TA:* Mave Alexander
Office number：(03) 5162543	
Instructor email: jaewon.yoo@iss.nthu.edu.tw	

COURSE DESCRIPTION:

This graduate-level course leverages R to equip students with the skills required for robust business data analysis. The course is designed to address both practical programming skills and the conceptual understanding necessary to apply data science effectively in business contexts.

Specifically, by the end of the semester you will be able to wrangle and visualize data, turning messy datasets into tidy, analyzable form and communicating them through clear graphics; evaluate causal claims and use simple and multiple linear regression to reach defensible conclusions; and quantify statistical uncertainty through sampling, the bootstrap, confidence intervals, and hypothesis testing, all with professional tools (R, RStudio, git, GitHub, and R Markdown or Quarto).

By the end of this course, students will not only be proficient in handling complex business datasets but will also appreciate the importance of meticulous data analysis. Even as agentic AI now writes much of the code, the essential skill has shifted from writing it to directing and verifying it: without the underlying building blocks, you cannot tell when an analysis is wrong, and you end up steered by the tool rather than steering it. The course thus builds programming and visualization foundations first, then turns to statistical inference for business analytics.

This course welcomes students from any background who are interested in developing data analysis skills applicable to the business sector. It is the graduate counterpart to the undergraduate section (ISS4066), with coursework and assessment held to a higher standard: the final project must use self-collected or non-trivial data, address identification and threats to inference, and reach greater analytical depth, and grading is calibrated to this higher bar. For graduate students, it also builds a foundation for more advanced business-analytics coursework, including causal-inference courses such as Experiments and Causal Inference (ECI; ISS5099). Students may not receive credit for both sections.

RECOMMENDED TEXTBOOKS:

We will use the following books in this class. Students are expected to complete the assigned reading before joining class, which will assist in understanding the lecture materials and foster in-class discussions. Students can purchase either a hard copy or e-book version of the books.

Be aware that it might take time for the textbooks to be delivered so if you plan to take the course, please make sure to place your order early enough so the books arrive in time.

1. **[MD]** Ismay, Chester and Albert Y. Kim. 2022. Statistical Inference via Data Science: A ModernDive into R and the Tidyverse.
2. **[QSS]** Imai, Kosuke and Nora Webb Willaims. 2022. Quantitative Social Science: An Introduction with Tidyverse, 2022. Princeton University Press.
3. **[IMS]** Mine Cetinkaya-Rundel and Johanna Hardin. 2021. Introduction to Modern Statistics. OpenIntro.
4. **[AAG]** R for Everyone: Advanced Analytics and Graphics, 2nd Edition by Jared P. Lander, O'Reilly Media, 2017.
5. **[VT]** Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics by Nathan Yau, John Wiley & Sons, 2011.

SOFTWARE TOOLS:

Students will use the R statistical programming language to write and run the codes for the assignments and in-class exercises. R is an open-source language that can be used free of charge and can be downloaded from https://www.r-project.org/. For improved productivity and convenience, students will also use RStudio, a widely used integrated development environment (IDE) for R which is available at https://www.rstudio.com/products/rstudio/ for download. You can also use Visual Studio Code to edit your code, which can be downloaded at https://code.visualstudio.com/.

COURSE EVALUATION:

Grading Policy

Performance in this course is assessed through a blend of practical assignments, participation, and in-class exercises, designed to gauge your proficiency in understanding of the materials we will cover throughout the semester. Here are how each portion of course contributes to the overall grade:

Category	Percent of Final Grade
Final Project	50%
Assignments	15%
Problem Sets	10%
Quizzes & In-Class Exercises	5%
Attitude/Participation	10%
Attendance	10%

Course Requirements

1. **AI Use Policy*:** Students are permitted to use AI tools, including large language models and agentic coding assistants (e.g., ChatGPT, Claude, Gemini), for assignments and the final project. You are fully responsible for everything you submit. The final project is designed so that a high-quality result requires genuine understanding; AI accelerates competent work but does not

substitute for it. However, it is imperative that the final submission predominantly reflects the student's understanding and personal input. To maintain academic integrity, students who choose to use AI tools must adhere to the following guidelines:

a. **Thorough Explanation:** Students must provide a detailed explanation of how the AI tool was used in the completion of the assignment.
b. **Original AI Responses:** Alongside their submission, students are required to submit the original responses generated by the AI tool.
c. **Critical Evaluation and Personal Input:** Students should critically evaluate the AI-generated content and clearly indicate the portions of the work that have been modified or expanded upon with their own insights and understanding.
d. **Proper Attribution:** Proper attribution should be given to the AI-generated content used, clearly indicating the sections that are AI-generated.

** Remember, attempting to cheat the system by heavily relying on AI-generated content without substantial personal input is ultimately the student's loss, as it undermines the learning process and personal growth. In line with NTHU's Implementation Guidelines on AI Ethics Statements and Learning Assessment (September 2026): if AI use is identified but not disclosed as required, the work may be re-evaluated or assigned no credit; in-class exercises are completed without generative AI; when the instructor uses AI to produce teaching materials, this is indicated in the materials; and by enrolling in this course, students are considered to have agreed to this statement.

2. **Academic Honesty and Plagiarism*:** All work submitted for academic evaluation must be the student's own. The penalty for violation of academic integrity will result in a zero for that assignment for the first time. Subsequent violation(s) will result in a failing grade for the course. Plagiarism will also not be tolerated.

   Academic dishonesty comprises of, but is not limited to, the following:

   - Cheating: Copying from other students' quizzes and assignments or allowing others to copy from one's own.
   - Plagiarism: Using other people's original work without giving appropriate credit or acknowledgment to the authors or sources.
   - Self-plagiarism: Submitting a piece of work in more than one course without the explicit permission of the instructors involved.
   - Misrepresentation of authorship: Submitting work as one's own, which has been prepared by or purchased from others.

   Students will be asked to upload their submission materials to Turnitin.com, an online plagiarism checker, to ensure academic integrity. Read more about online submission on http://learning.site.nthu.edu.tw/p/412-1319-7120.php?Lang=en.

3. **Attendance:** All students are expected to attend every class. Please bring your own hard copy of the course materials, which will be distributed by the instructor before class. If you have any urgent reason to miss a class, you are still responsible for the materials covered during the class and are expected to complete the required work. Attendance will be taken on a regular basis and will count towards your participation score (10%). Class missing will cause about 1% loss of final grade. Students who miss a class should inform me or the TA prior to the class via email or phone call.

4. **Final Project*:** In lieu of in-class examinations (typically two midterms and a final), the final research project is the primary summative assessment of this course. It is completed in teams and

includes a final in-class presentation with Q&A; a grading rubric will be provided. This is an open-ended data analysis project on a topic of your interest. Whether you are fascinated by voting trends in presidential elections, racial disparities in policing, NBA salary distributions, or the box-office economics of the Marvel Cinematic Universe, the core steps should involve: 1) develop a clear research question; 2) locate and prepare data that can address that question; 3) apply the analytical techniques covered in the course to answer the question. You should be able to present your findings in a format suitable for a general audience.

The aim is to produce a polished, portfolio-ready project that demonstrates the skills you've acquired to prospective employers. Your deliverable will be a publicly accessible article or webpage that includes: a concise overview of the research question and dataset; a visualization that directly addresses the question; a table or figure presenting a regression model and a plain-language explanation of the results; and a brief paragraph discussing the study's limitations, threats to inference, and how the analysis could be improved (for instance, by collecting better data). Finally, the piece must link to a GitHub repository containing all code used to load, clean, analyze, and visualize the data and to generate the article.

The data collection and cleaning must be meaningful -- it's not sufficient to simply use a pre-cleaned data from an *R* package. At the graduate level, self-collected or non-trivial data is required, though beware you may need to undergo the Institutional Review Board (IRB) review process to be able to use the data for your final project.

Milestone	Due
Creating a GitHub Repository	Sun Oct 11
Data and Proposal	Sun Oct 25
First Visualization	Sun Nov 15
First Analysis	Sun Dec 6
Final Report	Sun Dec 20
Presentation and Q&A	Thu Dec 24

5. **Assignments:** There will be a total of FIVE assignments throughout the semester. The assignments will be relevant to our class topic and most likely will require you to run and report a statistical method we will cover in class. For instance, after we cover ordinary least squares (OLS) estimation, students will be asked to run a dataset (to be provided by the instructor) and report the results in a simple article format. Students will be required to write a brief report using R Markdown/Quarto for the assignment. Each assignment is worth 3% of the final grade. Assignments are to be submitted on the due date. *Late submissions will be penalized 1 percentage point of the assignment's weight per day* (e.g., for an assignment worth 3% of the course grade, turning it in 2 days late caps the attainable score at 1%).

6. **Problem Sets:** Throughout the semester, you will be required to complete three problem sets that make up 10% of your final grade. These problem sets will consist of a mix of exercises taken from the recommended textbooks and are designed to reinforce your understanding of key concepts covered in class.

7. **In-class exercises: To keep the learning momentum going, students will complete five short in-class exercises (1% each) during class, without generative AI, and upload them to**

eeclass the same day. Each exercise ends with a verification step: check what your code actually did before you trust the output. There are no take-home quizzes.

8. **Attitude/participation:** In class, the most important thing for the students is to stay active and engaged about the topic being discussed. Positive contributions to class discussions will increase your score towards attitude. When we discuss a topic in class, effective discussions are only possible if everyone is well prepared. Please, be prepared to open and engage in discussions with your opinions and thoughts.

TENTATIVE WEEKLY SCHEDULE:

Week	Topics	Prepare before class	In-class deliverable
#1 (Sep 10)	Course Introduction and R Programming Basics: • How the course works • A first look at R and RStudio; Quarto • Getting started: programming errors, operators, variables	Read: MD Ch 1 AAG Ch 4.3, 5, and 6 VT Ch 1 and 2 Install R & RStudio	Self-introduction on eeclass discussion board (by Mon Sep 14) Complete Swirl tutorials 1–3 (by Week 2)
#2 (Sep 17)	Data Types, Data Structures, and Visualization I: • Numeric, character, logical, factor, dates • Vectors, lists, data frames; reading data • Building plots by layers with ggplot2	Read: MD Ch 2 AAG Ch 3, 4, 7.1, 9, 10, 27, 28 VT Ch 3 and 4	In-class exercise E1 AS1 released Group formation survey (Sun Sep 20)
#3 (Sep 24)	Functions and Visualization II: • Writing and reading functions • Histograms, facets, boxplots; principles of analytic graphics	Read: AAG Ch 7.2 and 11 VT Ch 5 and 6	In-class exercise E2 (gapminder) GitHub setup guide posted
#4 (Oct 1)	Data Wrangling with dplyr: • Operating on rows, columns, and groups • Bar plots	Read: AAG Ch 14, 15, and 16 VT Ch 7 MD Ch 3	In-class exercise E3
#5 (Oct 8)	Causality: • What is a causal effect? • Randomized experiments; calculating effects • Observational studies	Read: QSS Ch 2.1–2.5	AS1 due AS2 released Milestone 1: GitHub repository (Sun Oct 11)

#6 (Oct 15)	Individual or team meetings by appointment (no class meeting): • Discuss your research question and data before the proposal is due • Book a slot via Google Calendar; meetings at TSMC Bldg. R828B		
#7 (Oct 22)	Relationships, Importing, and Tidying Data: • Z-scores and correlation • Pivoting longer; joining datasets	Read: AAG Ch 11 MD Ch 4 QSS Ch 3.5–3.6	Milestone 2: data and proposal (Sun Oct 25)
#8 (Oct 29)	Prediction and Iteration: • Predicting election outcomes • Loops; evaluating predictions; time-series plots	Read: AAG Ch 4.7 and 12 VT Ch 8 and 9	
#9 (Nov 5)	Regression and Model Fit: • Modeling with a line; linear regression in R • Model fit	Read: MD Ch 5 QSS Ch 4.1 and 4.2.1–4.2.4	AS2 due AS3 and Problem Set 1 (IMS Ch 7) released (Course withdrawal period Nov 2–20)
#10 (Nov 12)	More on Regression: • Multiple regression • Categorical independent variables	Read: MD Ch 6.1–6.2 QSS Ch 4.2.6–4.3.2	In-class exercise E4 (NYC condominiums) Problem Set 2 (IMS Ch 8) released Problem Set 1 (Sun Nov 15) Milestone 3: first visualization (Sun Nov 15)
#11 (Nov 19)	Sampling and Sampling Distributions: • Sampling framework • Random variables; the normal and the central limit theorem	Read: MD Ch 7	AS3 due Problem Set 2 (Sun Nov 22)
#12 (Nov 26)	The Bootstrap and Confidence Intervals: • Resampling; bootstrap CIs for means and ATEs • Computing and interpreting CIs	Read: AAG Ch 18 and 19 MD Ch 8 IMS Ch 12	AS4 released

#13 (Dec 3)	Hypothesis Testing: • Hypothesis tests with infer; two-sample and permutation tests • Issues with hypothesis testing; power analysis	Read: MD Ch 9 IMS Ch 11	In-class exercise E5 Problem Set 3 (IMS Ch 11–12) released Milestone 4: first analysis (Sun Dec 6) Team consultations Nov 30–Dec 4
#14 (Dec 10)	Models of Uncertainty: • Using the normal for inference • CIs for experiments	Read: IMS Ch 13	AS4 due AS5 released Problem Set 3 (Sun Dec 13)
#15 (Dec 17)	Inference for Regression: • Uncertainty for regression coefficients • Presenting OLS regressions	Read: QSS Ch 7.3	
#16 (Dec 24)	Team Presentations and Q&A: • Each team presents for 15 minutes, followed by Q&A • Peer evaluation	Presentation materials ready	Final report (Sun Dec 20) Peer evaluations AS5 (Sun Dec 27)

Notes: The course schedule is subject to change if necessary (AAG, VT, MD, QSS, and IMS refer to 'Recommended Textbooks'; AS: assignment; E: in-class exercise). Assignments are due at the beginning of class on the due date; Sunday deadlines are 23:59. Week 6 has no class meeting; teams meet the instructor by appointment.