36-315 (Spring 2025)

Statistical Graphics and Visualization

Location and Times:
Spring 2025
Sun & Tue 8:30 - 9:45 AM
Location: CMB 1030

Instructor Information:
Taeyong Park, Ph.D.
✉taeyongp@andrew.cmu.edu
Office: CMB 2191
Office Hours: Sunday 10:00 - 11:00 AM
& Wednesday 11:30 AM - 12:30 PM
TA: Ishaq Ansari (✉iansari@andrew.cmu.edu)

1. Course Description

Graphical displays of quantitative information take on many forms, and they help us understand data and statistical methods by (hopefully) clearly communicating arguments, results, and ideas. This course introduces students to the most common forms of graphical displays and their uses and misuses. Students will learn how to create these displays and understand them from a statistical perspective. Furthermore, the course will consider complex data structures that are becoming increasingly common in data visualizations (temporal, spatial, and text data); we will discuss common ways to process these data that make them easy to visualize. In addition to two weekly lectures, there will be a weekly recitation session where students learn to use R to aid in the production of appropriate graphical displays, as well as weekly homework assignments that ask students to visualize and analyze real datasets. The course culminates in a final project, where students make public-facing data visualizations and analyses for a real dataset. All assignments will be in R; although this is not a programming class, using programming-based statistical software like R is essential to create modern-day graphics, and this class will give you practice using this kind of software. Throughout, communication skills (usually written or visual, but sometimes spoken) will play an important role. Indeed, if it's true that "a picture speaks a thousand words," then ideally the one thousand words you are communicating with your graphics are statistically correct, clear, and compelling.

2. Course Objectives

1. **Understand the Fundamentals of Data and Reproducible Data Analysis.**
    - Distinguish between data types and pinpoint which graphics and analyses are appropriate for a particular data type.
    - Write easily readable and reproducible code to explore datasets graphically.

- Master the use of R, RStudio, RMarkdown, and other tools to promote reproducible research and allow others to build from your work.

2. **Create Statistical Graphics.**

   - Master the use of R, RStudio, and RMarkdown to create statistical graphics that are easily readable and understandable for technical and non-technical audiences.
   - Incorporate statistical information (e.g. the results of statistical tests or uncertainty quantification) into elegant data visualizations.

3. **Write and Speak About Statistical Graphics and Data Visualizations.**

   - Describe graphics concisely and accurately to technical and non-technical audiences, both in writing and orally.
   - Incorporate appropriate statistical language in written and oral descriptions of graphics.

4. **Assess and Critique Statistical Graphics.**

   - Review others' statistical graphics objectively and academically.
   - Describe the pros and cons of a given graphical choice.
   - Give useful critiques, feedback, and suggestions for improvement on others' graphics.

3. Prerequisites

- 36-202 or 70-208

4. Textbooks

- *R for Data Science*, by Hadley Wickham, 2017. Click here to access its free on-line version.
- *ggplot2: Elegant Graphics for Data Analysis,* by Hadley Wickham. Click here to access its free on-line version.

5. Requirements and Evaluation

Attendance (5%): You are required to attend every class meeting on time and remain until the end. However, each student is allowed three free absences. Starting from the fourth absence, 0.5% will be deducted

from your final grade for each additional absence. No proof or documentation is required for absences, as I do not distinguish between excused and unexcused absences. It is advisable to save your free absences for emergencies or medical issues. The teaching assistant (TA) will take attendance. If you arrive more than 15 minutes late or leave before the end of the class (except for a 5-minute restroom break), you will be marked absent for the day. For any questions regarding your attendance record, please contact the TA or me.

Problem Sets (17.5%): Eight problem sets will be assigned throughout the semester as a homework assignment. Detailed schedules are in the course outline below.

- The problem set is designed to evaluate how well you understand the topic and motivate you to keep up with the material on a regular basis. Furthermore, it provides you with practice questions for the exams. Therefore, while you are allowed to work with your classmates to solve problem-set questions, I suggest that you ensure you are able to solve them independently.
- Of the eight problem sets, one lowest-scored problem set will be dropped at the end of the semester. Each of the remaining seven problem sets is worth 2.5% of the final grade, leading to 17.5% as total. You must submit your problem set in Canvas by the deadline specified in the course outline.
- Solution sets are provided in Canvas right after each problem set's deadline. Therefore, no extension is permitted.

Data Visualization Critique [Three critiques (4.5%) and Comments on classmates' critiques (3%)]. Data visualizations are found all over the place: News articles, advertisements, blogs, etc. We'll get practice discussing data visualizations we find in the wild, but from a statistical point of view. Once a month (January, February/March, April) you'll be expected to post on Piazza with a data visualization you found. Along with the data visualization, you must:

- *Describe the graph*. What does it show? What variables are plotted? What is the main result of the graph?
- *Critique the graph*. Does the graph do a good job of achieving its goals? What are the strengths and/or weaknesses of the graph? What would you change (if anything) about the graph?
- *Leave comments on at least two of your classmates' critiques*.

You'll get practice doing this on the first problem set. Then, you'll be expected to do this on Piazza. (You cannot use the graphic you found for the first problem set.) During lecture, we'll often discuss graphics we find in the wild, so one of your Piazza graphics may be featured in one of the lectures.

Take-home Exam 1 (15%): The purpose of the exam will be to assess students' ability to use material from the first half of the semester to answer open-ended questions about a real dataset. The problem sets and exercises are designed to prepare students for this exam, and we'll also have a midsemester review before the exam. The exam will be a "take-home" exam: Students will have from **8 AM - 11:59 PM on Thursday, February 20** (subject to change) to complete the exam and submit it via Canvas. Even though this is a 16-hour time frame, the exam will be designed to take approximately 3 - 4 hours; the 16-hour time frame is only meant to provide some flexibility and alleviate time pressure. More details will be provided closer to the exam date.

Take-home Exam 2 (15%): The purpose of the exam will be to assess students' ability to use material from the second half of the semester to answer open-ended questions about a real dataset. The problem sets and exercises are designed to prepare students for this exam, and we'll also have a midsemester review before the exam. The exam will be a "take-home" exam: Students will have from **8 AM - 11:59 PM on Thursday, March 27** (subject to change) to complete the exam and submit it via Canvas. Even though this is a 16-hour time frame, the exam will be designed to take approximately 3 - 4 hours; the 16-hour time frame is only meant to provide some flexibility and alleviate time pressure. More details will be provided closer to the exam date.

Term Project [Data and Research Idea (5%), Final Digital File (25%), and Presentation (10%)]: You will create a public-facing digital file (HTML via R Markdown) describing your work and give a presentation on the final exam day. The purpose of this project is to assess students' ability to communicate their work in the common formats seen in industry and graduate school: public-facing documents and oral presentations. These communication skills are important for data science in particular; the digital file can act as a piece of your "data science portfolio" as you apply for jobs, schools, and other opportunities. Detailed expectations about the project and some examples will be given later in the semester. In order to motivate you to start working on the term project early enough and allow myself to provide you with feedback, I set the following steps:

- Data and research idea submission due by 11:59 PM on Apr 12.
- Individual meeting on Apr 15.
- Final Digital File submission due by Presentation on the final exam day (TBD).
- Presentation on the final exam day (TBD).

Letter Grade Distribution:

>= 90.00	A
80.00 - 89.99	B
70.00 - 79.99	C
60.00 - 69.99	D
<= 59.99	R

6. Electronic Devices and Punctuality

I expect you to be respectful to me and your fellow students to create an environment that is most conducive to learning.

- You will often use your laptop or desktop during class this semester. However, this does not mean that you can feel free to use the computer for whatever you want. It is important to **use it only for class purposes** so that you will not distract yourself and you will not disrupt your classmates. Furthermore, **your cell phone must be turned off** during class. If there is an emergency that might oblige you to be contacted, please talk to me before class. I quote the following passage from the Qatar Business Administration Program Classroom Conduct, which, I believe, must apply to other programs as well:

  - *Laptops are to be closed. When class is in session, you may use your laptop only as directed by your professor. You should not check email, tweet, text, play games, or surf the Internet, any activity that diminishes your or your classmates' engagement with the classroom content and process. If you are unsure whether a given activity is appropriate, ask your professor. This policy extends to all electronic devices. Be sure that your phones and tablets are silenced and stowed before the class begins. Professors may add specific limits on the sharing or use of personal electronics in exam situations.*

- **You must come to class on time and remain in class** once the class has begun. I quote the following passage from the Qatar Business Administration Program Classroom Conduct, which, I believe, must apply to other programs as well:

  - *In common business culture, punctuality is an important part of showing respect for your colleagues and business partners. Showing up late for a meeting tells the others involved that you do not place much value on their time. QBA students will demonstrate respect for their courses, classmates and professors by arriving for class early enough to get settled and prepared before the scheduled meeting time.*

7. Office Hours and Appointments

I hold office hours: Sunday 10:00 - 11:00 AM & Wednesday 11:30 AM - 12:30 PM or by appointment. Please set up an appointment if you want to see me other than during my office hours.

8. Academic Integrity

You must comply with the academic integrity policy. You are required to refer to CMU's general policies on cheating and plagiarism: `http://www.cmu.edu/academic-integrity/valuing/index.html`. Violations of CMU's general policies on cheating and plagiarism carry a range of consequences: `http://www.cmu.edu/academic-integrity/understanding/index.html`.

You may use generative AI programs like ChatGPT during the brainstorming and idea generation phase for assignments. However, doing so cannot be considered a substitute for traditional research. Generative AI programs rely on predictive models to generate content that may appear correct, but has been shown to sometimes be incomplete, inaccurate, taken without attribution from other sources, and / or biased. Any information generated by an AI program should be cited like any other reference material. You are ultimately responsible for the content of the information you submit. However, you may not attempt to pass off any work generated by an AI program as your own.

9. Disability Resources and Health and Well-being

You can find information about disability-related accommodations on `https://scotty.qatar.cmu.edu/health-and-wellness/medical-accommodations/`. You may also consult me or CMUQ staff (Office of Health and Wellness) regarding learning disabilities, health, and wellness.

10. Diversity, Equity, and Inclusion

It is critical for me to ensure that students from all diverse backgrounds and perspectives feel belonging to this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength and benefit. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally or for other students or student groups. In addition, if anything conflicts with your value based on gender, sexuality, disability, age, socioeconomic status, ethnicity, race, and culture, please let me know so that I can make arrangements for you.

11. Course Outline

Date	Topic
Jan 5, 7	Introduction and Principles for Statistical Graphics
	R Markdown, `tidyverse`, `ggplot2`
Jan 12, 14	Review of Data Types and Statistical Inference & Basics of `ggplot2`
Jan 13	Recitation 1
	Problem Set 1 Due by Jan 18 11:59 PM
Jan 19, 21, 26	1D and 2D Categorical Data
Jan 27	Recitation 2
	Problem Set 2 Due by Feb 1 11:59 PM
Jan 28, 2, 4	1D Continuous and 2D Categorical - Continuous Data
Feb 3	Recitation 3
	Data Visualization Critique 1 Due by Feb 6 11:59 PM
	Comments on Classmates' Critique 1 Due by Feb 8 11:59 PM
	Problem Set 3 Due by Feb 8 11:59 PM
Feb 9, 16, 18	2D Continuous Data, Linear Regression, and Nonlinear Regression
Feb 11	National Sports Day; No class
Feb 17	Recitation 4
	Problem Set 4 Due by Feb 19 11:59 PM
	Take-Home Exam 1 (8 AM - 11:59 PM, Feb 20)
Feb 23, 25	Spring Break; No Classes
Mar 2, 4, 9	Visualization of Inference with Multivariate Linear Regression
Mar 3	Recitation 5
	Data Visualization Critique 2 Due by Mar 6 11:59 PM
	Comments on Classmates' Critique 2 Due by Mar 8 11:59 PM
	Problem Set 5 Due by Mar 15 11:59 PM
Mar 11, 16, 18	Visualization of Inference with Logistic and Multinomial/Ordinal Logistic Regression
Mar 17	Recitation 6

Continued on next page

Date	Topic
	Problem Set 6 Due by Mar 24 11:59 PM
Mar 23, 25	Time Series and Longitudinal Data
Mar 24	Recitation 7
	Take-Home Exam 2 (8 AM - 11:59 PM, Mar 27)
Mar 30, Apr 1, 6	Eid al-Fitr; No Classes
Apr 8, 13	Text Data, Word Clouds, and Sentiment Analysis
Apr 14	Recitation 8
	Term Project Data and Research Idea Due by Apr 12 11:59 PM
	Data Visualization Critique 3 Due by Apr 17 11:59 PM
	Comments on Classmates' Critique 3 Due by Apr 19 11:59 PM
	Problem Set 8 Due by Apr 19 11:59 PM
Apr 15	Individual Meeting for the Term Project
Apr 20, 22	Spatial Data and Maps
Apr 21	Recitation 9
	Problem Set 9 Due by Apr 26 11:59 PM