# STA 6233: R Programming for Data Science – Spring 2022

**Instructor**: Matthew J. Martinez, Ph.D.
> **Office Hours:** Monday 11:00am – 11:50am
> **Office Phone**: 726.666.0385 (Text/Call) – This is the best way to reach me. Not e-mail.
> **Email**: matthew.martinez@utsa.edu
> **Availability**: I will make every effort to respond to texts within 24hours and e-mails as soon as I can. Admittedly you will receive a much faster response if you text me. I will not respond over the weekend. Do not text/call after 9:00pm CST on any day.
> **Website:** https://www.linkedin.com/in/mattdemography/
> **Course Site:** https://github.com/mattdemography/STA_6233_Spring2022
> **Github:** https://github.com/mattdemography

**Assistant:** Matthew Angotti
> **E-Mail:**
> **Office Hours:** Wednesday at 12:00 PM

**Assistant:** Austin Pesina
> **E-Mail:** apesina@gvec.org
> **Office Hours:** Wednesday at 6:00 PM

## Class Location/Meeting Times

Online Asynchronous

## Description of Course

This course is designed to introduce students to the statistical program R for data analysis and manipulation. Topics include preprocessing/manipulating/combining datasets, summarizing and visualizing data techniques, writing functions, object oriented programming, data simulation and resampling methods, and interfacing R with other programming languages such as SQL, Python, C++, and Hadoop. Techniques for efficient programming will be stressed. The concept of high-performance computing (multi-core/parallel-processing) is also demonstrated.

This course is designed to move forward your utilization of R in a practical environment. Although you may find the applications discussed in this course beneficial for research purposes, the main aim of this course will be to equip you to use R as one would at a daily job in a Fortune 500 company. In this sense, the course materials and assignments will focus more on practical application than theoretical understanding.

By completing this course, you will have the opportunity to:
1. Develop high-end graphical representations of data
2. Find ways to simplify your code for readability and performance
3. Learn how to interface R with SQL, Python, and Github
4. Use web and document scraping for innovation and curation of data
5. Learn how to present your highly technical work to non-technical audiences in a clear, concise, and brief manner

**Required Materials**

All required readings will be available on Blackboard Learn. You will have to incur costs of printing posters, but this is minimal compared to textbooks. Additionally you will have to have access to a laptop to use during the course.

**Course Policies**

*Projects and Presentations*
Assignments are due on the day indicated in the syllabus before the start of class at 5:30PM via Learn. **No late work will be accepted.** Additionally, assignments *will not* be accepted via email. *We will discuss all assignments in much more detail in class.

*Grades*
This is a graduate level course. I assume you are in this course to gain valuable information. I also assume you will put forth maximum effort and work during and outside of class. Failure to complete projects and presentations may result in an incomplete or failing grade.

*University Policies*
Please visit http://www.utsa.edu/syllabus for University policies including those related to accommodations and academic dishonesty.

*Public Health Considerations*
The health and safety of our campus community is a shared responsibility of all Roadrunners. It is important to note that none of us can guarantee a COVID-19 free environment. We all must, however, follow the guidelines outlined in the UTSA Public Health Task Force Report ("Report") and any other applicable policies as may be communicated by the University from time to time. This will include regulating behaviors outlined in the Report including:

- Encouraging the use of face coverings,
- Self-monitoring for symptoms using the Daily Health Check before coming to campus,
- Getting tested for COVID-19 if showing symptoms or after a close contact with a COVID-19 positive individual (if you are not already fully vaccinated and are not symptomatic),
- Following proper hygiene practices, including frequent hand sanitization, using cleansing wipes to disinfect surfaces, and minimizing the use of shared devices, tools and equipment,
- Avoid congregating (i.e. bottlenecking) near the entrances and exits before and after class – keeping your distance to reduce possible transmission from symptomatic or asymptomatic individuals.
- Communicating any COVID-19 related health concern to your supervisor or professor, and
- Submitting a self-report to report your positive test results or exposure (if not fully vaccinated and are also symptomatic for COVID-19 infection).

In turn, faculty members or supervisors will submit a COVID Case Referral to alert the COVID Response Team about positive COVID-19 cases for operational action. Failure to abide by these guidelines and requirements may result in disciplinary action in accordance with the Student

[Code of Conduct](#) or applicable employment policies and procedures. Violations should be reported to the Office of Institutional Compliance via the [UTSA Hotline](#) for appropriate action.

*For Face-to-Face Classroom and Other Academic Sessions*
Face coverings/masks are recommended in indoor public and common spaces, especially for those individuals who are high risk and/or not vaccinated. For the latest information, please review [the Roadrunner Roadmap](#).

## PROJECTS

This course will have two main projects and presentations of those projects along with an evaluation of your peers' work.

*Document/Web-scrapping*
This project focuses on the creation of a unique dataset using data that are mined from documents and/or web sources. You will be responsible for coming up with a research question that you desire to answer using data that are previously unavailable. This task is designed to sharpen your skills in creative problem solving and ability to distill copious information into usable products using R packages centered on text manipulation, data frame management, feature creation, and regular expressions.

You will create a poster style presentation that will be showcased in a course symposium open to the department and the public. You will also evaluate your peer's work providing constructive criticism and encouragement.

*Interactive Project*
For this project you will have to design an interactive presentation using data to answer a different research question. This interactive project will either be a web document utilizing graphs and interactives from ggplot, Plotly, Google Vis, or d3 or you will create and deploy a Shiny Application. This task will allow you to use R packages centered on graphical representations of data including learning how to write and compile R Markdown files. Additionally, this project sharpens your ability to distill complex data to everyday audiences in a manner they find engaging and understandable.

You will publish your results online via Github and will be responsible for providing feedback to two of your classmates on their slide deck, use of graphics and interactivity, and your own ability to understand the main takeaway from their project.

| Summary of Assignments | Grading Scale |
|---|---|
| Poster Presentation | A+ = Successful Completion of All Projects |
| Evaluation of Peer's Posters | Incomplete/No-Credit = Unsatisfactory Work |
| Interactive Project | |
| Evaluation of Peer's Interactive Projects | |

**Tentative Class Schedule**

**Tentative Course Schedule of Assignments**

| Date(s) | Material to be Covered | Readings/Assignments/Exercises/Etc. |
|---|---|---|
| **Week of January 17th** | • Introduction to Course and Pre-Course Material | |
| **Week of January 24th** | • Introduction to R and R Studio | |
| **Week of January 31st** | • Data Manipulation: plyr, dplyr, Data Table | |
| **Week of February 7th** | • Data Readability and Performance: Functions, Loops, and Package Creation | Idea for Document/Web Mining Due |
| **Week of February 14th** | • Document and Web Mining | |
| **Week of February 21st** | • R Connections: Using and Learning Regular Expressions | |
| **Week of February 28th** | • Making Presentation Worthy Graphs: ggplot | |
| **Week of March 7th** | • **Course Symposium (Online)** | **Document/Web Mining Project Due** |
| **Week of March 14th** | • **Spring Break – No Class** | |
| **Week of March 21st** | • Github<br>• R Markdown | **Peer Evaluation of Document/Web Mining Projects Due Friday March 25th @ 11:59pm** |

| | | |
|---|---|---|
| **Week of March 28th** | • Making Interactive Presentations: Plotly and Google Vis | |
| **Week of April 4th** | • Shinny Apps Part 1 | Idea for Interactive Project Due |
| **Week of April 11th** | • Shinny Apps Part 2 | |
| **Week of April 18th** | • R Connections: SQL Coding Examples & Hadoop | |
| **Week of April 25th** | • **Working Week – No Class** | **Interactive Project Due May 1st @ 11:59pm** |
| **Week of May 2nd** | **All Peer Evaluations Due** Sunday, May 8th 11:59pm | |