COURSE AND CONTACT INFORMATION

Course: **STAT 6289 Computing Environments**

Lectures: MON B33
Times: M 06:10PM - 08:40PM
Instructor: Dr. Tatiyana V. Apanasovich

Office Address: 729 Rome Hall
Telephone Number: NA
E-mail: apanasovich@gwu.edu

Office Hours: T, 1:00-2:00pm and by appointment.

COURSE DESCRIPTION

This course offers an in-depth exploration of computing environments tailored for statistical analysis, with a focus on leveraging the power of R and Python. Students will gain the skills necessary to efficiently handle a wide array of data analysis tasks within these programming environments. Throughout the course, students will engage with both foundational and advanced topics, starting from the basics of R and Python, progressing to specialized techniques such as spatial analysis, web development with Shiny and Dash, and advanced computations using TensorFlow and PyTorch.

The course emphasizes the critical role of computing environments and tools in the statistical analysis workflow. Students will be guided through real-world scenarios, highlighting the integration of statistical programming, data manipulation, and model development within robust computing frameworks. Special attention will be given to the tools and techniques that enhance computational efficiency, reproducibility, and scalability in statistical tasks.

<u>**Week 1:**</u>
Introduction to R and RStudio
Resources and basics of R/RStudio
Introduction to Version Control (GitHub)
Importance of statistical programming and computational environments

<u>**Week 2:**</u>
Workflow in R
Project-oriented workflows
Structuring R projects
Good practices in R programming
Command line usage
Introduction to literate programming in R

<u>**Week 3:**</u>

R Scripts and Memory Management
Input/Output operations, various data storing formats,
(CSV,JSON,SQL, etc)
R Markdown

Week 4:
Introduction to Tidyverse
Advanced R programming
Functional Programming with purrr, Error Handling and Debugging, Parallel
Programming

Week 5:
Python Basics
Introduction to Python for data analysis
Data shaping, JSON handling, and filtering

Week 6:
Advanced Python Topics
Exploratory Data Analysis (EDA) and descriptive statistics
Data Visualization

Week 7:
Python Regression Models
Multiple Linear Regression in Python using sklearn and statsmodels
More on regression and statistical models in Python, logistic regression,
Smoothing techniques

Week 8:
Data Analysis in R
Regression Models:
Linear and Logistic regression in R

Week 9:
Factors in R and using forcats
Web scraping techniques

Week 10:
Text mining in R
Working with time series data in R using lubridate

Week 11:
Spatial Analysis and Object-Oriented Programming in R
Introduction to spatial data analysis
Introduction to GeoPandas in Python for geospatial data analysis

Week 12:
Introduction to Shiny for interactive web applications in R
Flexdashboard for R
Dash in Python
Introduction to Dash for building web applications in Python

Week 13:
Interfacing Between Python and R
Using Python within R and vice versa (links and attached files provided)
Best Practices
Handling Large Data in R

Week 14:
Introduction to TensorFlow
Overview of TensorFlow for deep learning in Python
Introduction to PyTorch

COURSE PREREQUISITE(S)
Introductory Statistics: Familiarity with introductory statistics topics such as hypothesis testing, confidence intervals, and simple linear regression will provide a good foundation for the course.

Programming Fundamentals: Basic programming skills and familiarity with any programming language (e.g., Python, R, or similar) will be advantageous.

TEXTS

**There is no required textbook. We will make use of several
freely available textbooks and other materials. All course materials will be provided.
We will use the R and Python software for data analysis, which is freely available
for download. Some textbooks for R and Python**

1. "Python for Data Analysis" by Wes McKinney

This book provides a comprehensive guide to using Python for data analysis, data manipulation, and visualization. It covers essential libraries like NumPy, pandas, and matplotlib, which are commonly used in data analysis tasks.

2. "R for Data Science" by Hadley Wickham and Garrett Grolemund

This book is an excellent resource for learning data analysis using the R programming language. It covers data manipulation, visualization, and modeling with packages like tidyverse, dplyr, and ggplot2.

LEARNING OUTCOMES

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

- Explain the significance of computing environments in data analysis workflows and statistical modeling.
- Utilize programming languages such as Python, R, to handle diverse data formats and perform data manipulations.
- Apply statistical software environments for data cleaning, analysis, and visualization.
- Calculate summary statistics, create data visualizations, and perform exploratory data analysis (EDA) to comprehend data characteristics and patterns.
- Implement statistical models, regression analysis, ANOVA, and t-tests using programming languages for data-driven decision-making.
- Utilize distributed computing frameworks like Apache Spark for big data analysis and handling large datasets.
- Create reproducible analysis workflows using version control tools like Git and present results through R Markdown and Jupyter Notebooks.
- Apply computing environments to analyze real-world datasets from diverse domains such as healthcare, finance, and social media.

GRADING

Your final grade will be a weighted average of your homework average (20%), midterm take-home exam (40%), and take home project (40%).

CLASS POLICIES

Homework: There will be 6-8 homework assignments, with greater frequency in the first half of the course. No late homework will be accepted, but the lowest score will be dropped

Midterm Exam: There will be a take-home exam consisting of a few problems, where you will be using the software to do the problems(posted on Oct, 14 and due Oct, 21). Must work independently.

Take Home Project: There will be a take-home exam consisting of a few problems, where you will be using the software to do the problems (posted on December, 2 and due December, 16). Must work independently.

University policies:

University policy on observance of religious holidays

In accordance with University policy, students should notify faculty during the first week of the semester of their intention to be absent from class on their day(s) of religious observance. For details and policy, see: students.gwu.edu/accommodations-religious-holidays.

Academic integrity code

Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information. For details and complete code, see: studentconduct.gwu.edu/code-academic-integrity

Safety and security
In the case of an emergency, if at all possible, the class should shelter in place. If the building that the class is in is affected, follow the evacuation procedures for the building. After evacuation, seek shelter at a predetermined rendezvous location.

Support for students outside the classroom

Disability Support Services (DSS)
Any student who may need an accommodation based on the potential impact of a disability should contact the Disability Support Services office at 202-994-8250 in the Rome Hall, Suite 102, to establish eligibility and to coordinate reasonable accommodations. For additional information see: disabilitysupport.gwu.edu/

Mental Health Services 202-994-5300
The University's Mental Health Services offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems. Services for students include: crisis and emergency mental health consultations confidential assessment, counseling services (individual and small group), and referrals. For additional information see: counselingcenter.gwu.edu/