AS.230.160.12 (Intersession 2021)
Rhiannon Miller
Office Hours: Immediately following class and by appointment
Email: rhiannon.miller@jhu.edu
Lecture/Lab: M-F 1:30pm – 4:15pm EST, January 4 – 15, 2021
Zoom Info:
https://JHUBlueJays.zoom.us/j/97904713310?pwd=cDlXMkdZaTdhTmk5Ump4QWpTbFErUT09
Meeting ID: *979 0471 3310*
Passcode: *190623*

Dealing with Dirty Data: An Introduction to Data Cleaning and Analysis in R for the Social Sciences

Course Description: Data cleaning is the art of taking raw data and turning it into comparable files for analysis. Students will gain an introduction to data acquisition, cleaning, and analysis. Lectures and assignments will be grounded in common problems encountered by researchers working with new data. This course will provide an essential foundation for students preparing for careers in policy analysis, marketing research, academia, public health, or any other field that relies on quantitative data.

By the end of this course you will be able to find and download data, read it into R, create new variables, merge data to other sources, apply simple debugging strategies, and present data through basic analysis. Students' final project will produce a programming sample required by many job applications.

Attendance: Session attendance and participation is required. Please contact the instructor to discuss any barriers to either.

Logistics:

Course Sessions*:* The course will meet via synchronous Zoom Session every day, beginning at 1:30pm, Monday through Friday, January 4, 2021 through January 15, 2021. The instructor will remain available until 4:15pm during each session. Most lectures will be pre-recorded and made available 24 hours in advance of the beginning of each session. Students are asked to watch lectures before logging on to Zoom at 1:30pm each day. Zoom sessions will be used to work through exercises, talk through solutions as a group, and help students as they work on final projects. Students will need to work from a laptop to be able to complete tasks and exercises during course time.

Office Hours: The instructor will stay online in each Zoom session to answer any questions after class. You may also make an appointment to meet with the instructor.

Troubleshooting Teams: Learning how to work with data while also learning a new programming language can be challenging. Your peers are your best resource and first line of defense. Class participants will be organized into Troubleshooting Teams of four to six students by the instructor. When questions arise, as they inevitably will, you should contact your Troubleshooting Team first before posting on Blackboard or contacting the Instructor.

Course Texts (Available for free online):
R for Data Science, by Garrett Grolemund and Hadley Wickham
Advanced R, by Hadley Wickham

Course Requirements and Grading:

Daily Exercises: Most course sessions will include exercises during the "lab portion" of the class. Sometimes you will need to finish these exercises on your own. Support will be provided during lab time for any questions. Answers to daily exercises will be reviewed in class through participatory instruction. Daily exercises will not be graded, but completion and participation are strongly encouraged to facilitate progress on the Midterm and Final projects.

Extra Practice: Learning R is like learning any new language and the best way to make progress is to practice. Additional opportunities will be recommended after most sessions. Students are strongly encouraged to allocate time each day to work with provided datasets and practice the skills reviewed in the previous course. Students will have the opportunity to share any fun experiments, break throughs, or problems they encounter during the discussion portion of each session.

Midterm check in: You may submit rough draft of your final project to the instructor for feedback. This is optional. To receive feedback, please submit no later than Saturday, January 9, 2021 at 11am EST.

Final Project (60%): Final projects will require students to apply skills learned in class to their own dataset. Students are encouraged to select a dataset to work with by the beginning of the fourth course.

Troubleshooting Team Support (20%): Students will be placed into groups of four to six based on their time zone. Groups will be balanced to include diversity in previous experience with data and majors. Some of the best resources in learning to program and work with data are others who are working through the same problems. Most data problems can be solved in multiple ways and being exposed to a variety of ways of thinking about data and programming is one of the best ways to improve quickly. You should use your team as a line of first defense in fielding questions, checking code, and working through daily exercises. If you are stuck on something for more than 30 minutes (after checking documentation and doing a google search), then you should reach out to your team. At the end of the term, all team members will be asked to rate their fellow team members on the responsiveness, support, and contributions to the team. These ratings will be averaged to form this portion of the grade.

Class Attendance and Participation (20%): The best way to learn to work with data is to practice, troubleshoot, and ask lots of questions. Class time will alternate between lecture, demonstration, and lab time. Thus, it is very important to attend class and participate actively. Participation may vary according to each student but can take the form of asking questions (on Zoom, Blackboard, or by email); submitting alternate ways of doing examples in class or daily exercises; sharing found resources with the class; or sending the instructor feedback. Due to the condensed nature of the course, it will be easy to fall behind, as each session will build on the previous session. For this reason, students are required to keep the pace with the course and should discuss any barriers with the instructor as soon as they may present themselves.

Course Plan:
Session 1 (January 4, 2021) – Basics of R
Session 2 (January 5, 2021) – Getting Data
Session 3 (January 6, 2021) – Documentation and Basic Data Manipulation
Session 4 (January 7, 2021) – Strings, Dates, and Tidyverse
Session 5 (January 8, 2021) – Merging and Debugging
Session 6 (January 11, 2021) – Loops and Data Checking
Session 7 (January 12, 2021) – Inheriting Code
Session 8 (January 13, 2021) – Creating Tables
Session 9 (January 14, 2021) – Data Visualization
Session 10 (January 15, 2021) – Odds, Ends, Q&A