MGP-435 – Data Wrangling

TERM: Spring 2020

LECTURES:
5/17 (9:30 AM – 12:00 PM),
5/24 (9:30 AM – 12:00 PM),
5/31 (9:30 AM – 12:00 PM),
6/7 (9:30 AM – 12:00 PM)
6/7 (1:00 PM – 3:00 PM) – Informal Discussion about Projects

INSTRUCTOR: Mehul Rangwala mrangwala@ucdavis.edu

OFFICE HOURS: Anytime I am not teaching.

COURSE DESCRIPTION: This hands-on one-credit course will introduce students to the essentials of preprocessing data using R programming language to turn noisy data into tidy one. The course will start with a brief introduction to R and introduce students to the RStudio computing environment. The course will then cover the concepts and mechanics of data wrangling and culminate with a group project.

REFERENCE TEXTBOOKS AND RESOURCES:

1. Data Wrangling with R by Bradley C. Boehmke (SUGGESTED, NOT REQUIRED)
Publisher: Springer
ISBN-10: 3319455982 | ISBN-13: 978-3319455983
The electronic copy of this book is available via our library. There is NO need to purchase it. Please follow this link to access the text.

2. R Studio's Data wrangling cheat sheet (free)
https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

3. R for Data Science (Available for free at the time of preparing this syllabus.) (SUGGESTED, NOT REQUIRED)
http://r4ds.had.co.nz/

4. Practical Data Wrangling (SUGGESTED, NOT REQUIRED)
Publisher: Packt Publishing
ISBN-10: 1787286134 | ISBN-13: 978-1787286139

NOTES AND HANDOUTS: I will upload notes and in-class exercise files to Canvas before each class meeting. My notes, exercises, homework, and solved examples are crucial to successful completion of the course.

COMPUTER PACKAGE: RStudio. You can download and install RStudio at no cost from https://www.rstudio.com/.

INSTRUCTION APPROACH: Each class session will be a blend of lectures and in-class exercises on the topics covered. This course is completely hands-on, as data wrangling cannot be learned just by listening to the instructor talk. Please bring your laptops to every class. Attending all class sessions and participating in all in-class exercises is essential to deriving maximum value from the course. In the class, we will run the code in the lecture notes and discuss the results.

GRADING:

Homeworks (Individual)	65%
Group Midterm Evaluation	10%
Group Final Project	25%

Course Objectives:

Upon successful completion of the course, you will be able to:

1. Develop an understanding of the R programming language and the RStudio programming environment.
2. Import structured and unstructured data into R.
3. Understand what tidy data means.
4. Reshape and transform your data with R packages.
5. Combine data sets using R packages.
6. Scrape HTML text and tables from the web.
7. Perform string processing using regular expressions (regex).

PREREQUISITES:
None. We will cover enough R needed for data wrangling in the first class. Please ensure you don't miss the first class.

WHY R, AND NOT EXCEL, FOR DATA WRANGLING?
I am sure this question will cross somebody's mind. Here are a few points that favor R (or any other programming language):

1. Excel cannot open very large datasets. It has limitations on the number of rows (and columns). Excel files with large datasets sometimes get unstable/corrupted and are very slow to open and load.

2. It is not very easy to plot some charts like a boxplot using Excel. We also need to use extensive formulas to identify outliers.
3. While using R packages and writing single lines of code in R may seem daunting at first because it is new to so many of you, but it is still comparatively easier than writing blocks and blocks of IF, TRIM, LEFT statements in Excel. It would not take very long for people to get lost in your "paragraph" of Excel formulas.
4. Third-party packages (like Trifacta) are available for data wrangling but still takes more time to learn it than a one-unit class can afford. Moreover, your knowledge of R can be applied to other areas of analytics.
5. R and Python are standard languages to perform data wrangling in the world of analytics.

HOMEWORKS:
There will be three homework assignments based on the topics covered in the class. These assignments will be identical to the in-class exercises and will need to be individually completed. The idea behind the individual homework assignments is to provide students some additional practice, assess their understanding, and prepare you to work on the final projects. The R-code should be submitted with the output.

WHY THREE HOMEWORKS FOR A ONE-UNIT COURSE?
Because data wrangling cannot be learned by reviewing someone's code or listening to lectures. You have to practice it yourself to understand the challenges involved with data wrangling.

The first homework will be on basic R skills.
The second homework will be on techniques of tidying data.
The third homework will be on dealing with missing values and data transformations.

As you see, three homeworks are necessary to obtain exposure to each area.

FINAL PROJECT:
Students will work in groups on the final project. The final project should entail taking an existing dataset (either publicly available or from your work), use the techniques from the class to tidy the data, and perform some exploratory data analysis/visualizations (learned in your 203A course). Further instructions on the final project will be provided in the class.

MID-TERM PROJECT EVALUATION:
Approximately midway through the course, you need to submit a written report summarizing your project. Further instructions on which elements to include in this report will be provided in the class.

WHY FINAL PROJECT IN ADDITION TO HOMEWORKS?
Because homeworks assess how you can apply the concepts learned from the class to scenarios I provide. Final project assesses how you can take a real-world dataset of your choice, identify what needs to be tidied, and apply the principles learned. Homeworks can be considered as practice drills while the project is a comprehensive start-to-finish data wrangling experience.

FINAL EXAM:
There is no final exam for the course. We may use this time to cover remaining topics and informally share with others what your project entailed.

Academic Honor Code:

All students are expected to adhere to the University of California, Davis' Code of Conduct as noted here: http://sja.ucdavis.edu/files/cac.pdf.

TOPICS SCHEDULE (TENTATIVE):
This is a tentative schedule. Contents and sequence are subject to rearrangement.

Date	Assignments Due	Topics Covered
5/17/2020		1. Introduction to R and RStudio 2. Collect and Import Data 3. Exploring Raw Data
5/24/2020	Homework 1	4. Dealing with Strings and Dates 5. Clean and Tidy Data
5/31/2020	Homework 2	6. Dealing with Missing Values and Outliers 7. Data Transformations with `dplyr`
6/7/2020 (AM)	Homework 3	8. Combining Data sets with `dplyr` 9. Data Visualization with Base R and `ggplot2`
6/7/2020 (PM)		10. Introduction to web scraping and regular expressions
6/10/2020	Final Project Due (submit if completed earlier)	**No need to come to the class. Submit online.**