Graduate School of Management
University of California, Davis **Mehul Rangwala**

MGT-435 - Data Wrangling

TERM: Fall 2018

INSTRUCTOR: Mehul Rangwala 916.399.3271 mrangwala@ucdavis.edu

INSTRUCTOR OFFICE HOURS:
Please feel free to contact the instructor any time if you have questions. I will stay back and be available after each class meeting as long as the students need. I encourage you to use this time to ask questions or review any part of the material that you are having difficulty with. If the time after class is not convenient, then a separate appointment can be arranged for a meeting either in-person or over Zoom. You can also ask me questions anytime without appointment via email, text, or phone. In short, my office hours are almost when you need me. It is critical that you clearly understand the concepts covered in the course. Getting your questions answered and helping you understand the material, exercises, and homework are my topmost priorities.

CLASS SCHEDULE:

Date	Timing	Session Type
10/22/2018	2:10 – 4:00 pm	Lecture
10/29/2018	2:10 – 4:00 pm	Lecture
11/5/2018	2:10 – 4:00 pm	Lecture
11/19/2018	2:10 – 4:00 pm	Lecture
11/26/2018	4:10 – 6:00 pm	Lecture
12/3/2018	2:10 – 4:00 pm	Final

COURSE DESCRIPTION:
We live in an era of data deluge. There has been a dramatic increase in the amount of data, both structured and unstructured, obtained from plethora of sources including websites and devices. Steady decrease in the cost of storage will make this proliferation even more profound. "Big data" is not only about volume of data but also its velocity and variety. According to *Gartner Research*, big data and analytics are currently at the top of business agendas, with CxOs naming Business Intelligence (BI) and analytics as their top priorities.

All business initiatives entail performing some level of data analysis, obtaining actionable insights from the data, and making decisions. Visualizations and statistical modeling methods are just a couple of common ways to perform data analysis. Successful analysis relies upon accurate, well-structured data that has been formatted for the specific needs of the research questions to be addressed. Due to the increasing complexity of the available data, rarely is it available in the format that is ready for analysis. There is always a need to do preparatory work to make the data ready for analysis. It is commonly believed that analysts spend 50 – 80% of their time preparing the data for analysis. The *process* of taking the raw data from the source(s) (database, web, tweets, etc.) and transforming it into the *tidy* form useful for analysis is called *Data Wrangling*. Data Wrangling is a lot more than data cleaning. Specifically, the process[1] of data wrangling

[1] Source: https://www.trifacta.com/data-wrangling/

involves understanding data, structuring it, cleaning it, enriching it, validating it, and publishing it for downstream analysis.

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 (review for some) to R and introduce students to the RStudio computing environment. The course will then delve into importing data from text and Excel (CSV) files into R, scraping data and HTML text from the web (basic web scraping), processing strings with regular expressions, and using `dplyr` and `tidyr` packages in R to shape and transform your data. During the course, students will work on a group project that can entail wrangling data in datasets of their choice. If possible, students can also bring their data from work if it is non-sensitive in nature.

AUDIENCE:
Anyone with interest in analytics/data science who is currently performing data analysis work, or foresee performing such work will benefit from this course.

PREREQUISITES:
Since this is a one-credit (short) course, some understanding of R would be beneficial to save class time and gain a productive experience. You do not need to have a detailed understanding of R or need to be a programmer to take this course. The course will offer you opportunities to enhance your understanding in R. Students having no background in R can still take the course, provided they spend time outside the class going through the chapters 2 and 3 from the *Data Wrangling with R* required textbook **before** the first class session. There are plenty of other free resources available on the internet that you can use to supplement your learning the fundamentals of R.

TEXTBOOKS AND RESOURCES:
1. *Data Wrangling with R* by Bradley C. Boehmke (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. You may need to use the library VPN to access the book.

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.

TOPICS TO BE COVERED:
1. A refresher on R and RStudio.
2. Understanding Tidy data.
3. Importing data from text and Excel (CSV) files into R.
4. Data structures in R.
5. Web scraping.
6. Exporting data.
7. Processing Strings with Regular Expressions (regex).
8. Data shaping and transforming using R packages.

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

CLASS INSTRUCTION:
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.

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

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). For those who have completed the 203B course can go beyond this minimum requirement and perform some basic statistical analysis of the tidied data. However, this is optional. 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. This evaluation will provide you feedback and direction for your final project. Further instructions on which elements to include in this report will be provided in the class.

EVALUATION:
Your final course grade will be determined from the following:

Individual Homework Assignment1	25%
Individual Homework Assignment 2	25%
Midterm Project Evaluation	25%
Final Project	25%

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.

LEARNING OBJECTIVES:
Upon successful completion of the course, you will be able to:
1. Develop a better understanding of the R programming language and the RStudio programming environment.
2. Understand what tidy data means.
3. Import structured and unstructured data in R.
4. Export data into text and Excel files.
5. Reshape and transform your data with R packages.
6. Scrape HTML text and tables from the web.
7. Perform string processing using regular expressions (regex).

SCHEDULE (TENTATIVE):
This is a tentative schedule. Contents and sequence are subject to rearrangement. The chapter numbers correspond to those in the required textbook.

Date	Assignments Due	Topics Covered
Session 1 (10/22/18)		• The Basics (Chapter 3) • Data types in R • Introduction to Tidy Data • Importing Data (Chapter 15) • Data Frames versus Tibbles in R
Session 2 (10/29/18)		• Scraping Data (Chapter 16) • Subsetting and Filtering data • Dealing with Missing Values (Chapter 14)
Session 3 (11/5/18)	Homework 1	• Dealing with Numbers, Character Strings, Factors, and Dates (Chapters 4, 5, 7, and 8)
(11/9/18)	Mid-term Project Evaluation	**NO CLASS.** Only the Mid-term Project Evaluation Report due.
Session 4 (11/12/18)		• Shaping and Transforming Data with R (Chapters 21 and 22) • Combining Tables
Session 5 (11/26/18)	Homework 2	• Combining Tables (Continued) • Dealing with Regular Expressions (Chapter 6)
Final (12/3/18)	Final Project	Group Project Presentations.