RMBI 3010 - Data Analytics with R (2020-21 Fall Term)

General Information:

- 2 Credits (Letter Graded)
- Lecture + Lab: (L1 + LA1) Mon 12:00PM - 2:50PM by Zoom
- Instructor: Dr. Jean WANG <jeanwang@ust.hk> Rm: LSK 5050A (office hour by appointment)
- TA: Mr. Samuel LAI <imsamuel@ust.hk> Rm: LSK 4065 (office hour by appointment)

Course Description

This course provides an introduction to R as a programming language and environment for data analytics and visualization. R is popular in many fields and industries for small and big data applications. It is open-source and backed by a huge community that creates new tools and packages every day.

The course will first cover the basic syntax of R language, including functions and flow control. Then, it will introduce some commonly used data structures, such as vectors, lists, matrices and data frames. Next, data importing and visualization in R will be presented. Furthermore, the course will also introduce a few primary data cleaning techniques in dealing with missing values, duplicates and inconsistency, and how to implement simple data transformation and normalization with R. Last, some classic data mining models and the corresponding packages in R will also be presented. Each session of the course will consist of presentations and demos on the topic and hands-on exercises for students to practice.

Teaching Schedule (tentative)

WK	Lecture Topic
1	*[Sep 7]* Basic Syntax of R • RStudio IDE • Basic data types • Read and write text files • Import and export files
2	*[Sep 14]* Flow of Control in R • Flow of control • Functions • Special values • Data cleaning basics
3	*[Sep 21]* Data Plotting in R • Basic plots • Plot options • Basic graphic functions • Intermediate plots

4	*[Sep 28]* Exploratory data analysis in R • Frequency count and aggregation • Normality test • Correlation plot • Linear regression
5	*[Oct 5]* Text Processing in R • String manipulation • Regular expression • Word frequency • Document-term matrix
6	*[Oct 12]* Web Scraping in R • HTML basics • CSS basics • *rvest* library
7	*[Oct 19]* Interactive Dashboard in R • *shiny* library • R Markup • Publish to Web
8	*[Oct 26]* *No Class (public holiday)*
9	*[Nov 2]* *Group Project Presentation*
10	*[Nov 9]* *Group Project Presentation*

Assessments and Weighting

- **Attendance and Class Participation (15%):** week 1 to week 7
  Students are required to attend all lectures, and are strongly encouraged to interact with the instructor and peers during the lectures.

- **Weekly Exercises (35%):** week 1 to week 7
  These are individual continuous assessments. Each week, students are given a real-world business data and a series of data analysis tasks. They are required to follow the instructions to complete an R script file, in order to accomplish a specific risk analysis task. After finishing, students need to submit their script file to present their findings.

- **Group Presentation (20%):** week 9 to week 10
  This is a group and individual assessment. 2-3 students will form a group to conduct a group presentation on a specific R library to build a statistical analysis, data mining or machine learning model: such as *Time Series Forecast, Bayesian Classification*, *Convolutional Neural Networks, etc*. Group needs to obtain the instructor's approval on their chosen topic.

- **Final Project (30%):** end of semester
  Each student needs to crawl and download a data set from a website, write R code to conduct data analysis on the set, and summarizes his/her findings. The submission includes the R code and a report.

Textbook and References:

- *[Textbook]* ***R and Data Mining:*** *Examples and Case Studies*
  Author: Yanchang Zhao Publisher: Elsevier Inc.
  ISBN-13: 978-0123969637 ISBN-10: 0123969638
  Preview available on Google Books https://books.google.com.hk/books?id=FEOh08LBD9UC

- *[References]* **RDataMining.com: R and Data Mining**
  http://www.rdatamining.com/

- *[Packages]* **Awesome R - Find Great R Packages**
  https://awesome-r.com/index.html

- *[Data Sets]* **Rdatasets: An archive of datasets distributed with R**
  https://vincentarelbundock.github.io/Rdatasets/

- *[Data Sets]* **World Bank Open Data**
  https://data.worldbank.org/

- *[Data Sets]* **Kaggle: Your Home for Data Science** (registration needed)
  https://www.kaggle.com/

- *[Data Sets]* **data.world: Datasets for Analysis & Download** (registration needed)
  https://data.world/