

RMBI 3000B - Data Analytics with R (2017-18 Spring Term)

General Information:

- Instructor: Dr. Jean Wang Email: <u>jeanwang@ust.hk</u> Office: 4357 (Lift 17-18)

Lecture + Lab: (LA1) Thu 01:30PM - 04:20PM Rm 4402 (Lift 17-18)

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

WK	Lecture Topic	Dataset for Weekly
		Practice
1	Basic Syntax of R	Pokemon
	RStudio IDE	
	Basic data types	
	Flow of control	
	• Functions	
2	Data Import and Export in R	Airbnb in Seoul
	Vector, matrix, list structure	
	Data frame structure	
	Read and write text files	
	Import and export Excel files	
3	Data Plot in R	Hong Kong Stocks
	Graphics packages	
	• Simple charts: histogram, bar/line Chart, scatter plot, box plot	
	Advanced charts: bubble chart, heat map, geography map	
4	Data Cleaning and Transformation in R	Shared Bike Rental
	Deal with missing values	
	Detect inconsistencies	
	Remove duplicates and outliers	
	Simple transformation and normalization	





6	Text Processing in R Reading text data from files Stemming words Building a term-document matrix No class (consultation for group projects)	Donald Trump Tweets
8	No class (consultation for group projects) No class (consultation for group projects)	
9	Decision Tree and Random Forest Decision / Regression Tree with package party Random Forest with package randomForest and party	Group Project Presentation and Demo
10	Clustering The k-Means Clustering with function kmeans() Hierarchical Clustering with function hclust() Density-based Clustering with package fpc	Group Project Presentation and Demo
11	Association Rules	Group Project Presentation and Demo
12	Network Analysis Network Visualization and Community Detection with package <i>igraph</i> Social network analysis with package suite <i>statnet</i>	Group Project Presentation and Demo
13	To be determined based on enrollment	

Assessments and Weighting

- Weekly Exercises (50%): week 1 to week 5, week 9 to week 12

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 Project and Demo (40%):** week 9 to week 12

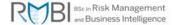
This is a group assessment. Students need to conduct search online to look for real business datasets and apply an analytic model to the data and prepare a set of data analysis related questions for peer students to practice and answer. The research include the applicability of the chosen model, data feature selection, model parameter fine-tuning and result interpretation.

Class Participation (10%)

This includes attendance and class participation in classes, as well as peer reviews in group presentations.

No Midterm Test and Exam

This course is purely project-based, without any midterm test and exam.





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/

