

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

---

### General Information:

- Instructor: Dr. Jean Wang                      Email: [jeanwang@ust.hk](mailto:jeanwang@ust.hk)                      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 <ul style="list-style-type: none"> <li>• RStudio IDE</li> <li>• Basic data types</li> <li>• Flow of control</li> <li>• Functions</li> </ul>	Pokemon
2	Data Import and Export in R <ul style="list-style-type: none"> <li>• Vector, matrix, list structure</li> <li>• Data frame structure</li> <li>• Read and write text files</li> <li>• Import and export Excel files</li> </ul>	Airbnb in Seoul
3	Data Plot in R <ul style="list-style-type: none"> <li>• Graphics packages</li> <li>• Simple charts: histogram, bar/line Chart, scatter plot, box plot</li> <li>• Advanced charts: bubble chart, heat map, geography map</li> </ul>	Hong Kong Stocks
4	Data Cleaning and Transformation in R <ul style="list-style-type: none"> <li>• Deal with missing values</li> <li>• Detect inconsistencies</li> <li>• Remove duplicates and outliers</li> <li>• Simple transformation and normalization</li> </ul>	Shared Bike Rental

5	Text Processing in R <ul style="list-style-type: none"> <li>• Reading text data from files</li> <li>• Stemming words</li> <li>• Building a term-document matrix</li> </ul>	Donald Trump Tweets
6	<i>No class (consultation for group projects)</i>	
7	<i>No class (consultation for group projects)</i>	
8	<i>No class (consultation for group projects)</i>	
9	Decision Tree and Random Forest <ul style="list-style-type: none"> <li>• Decision / Regression Tree with package <i>party</i></li> <li>• Random Forest with package <i>randomForest</i> and <i>party</i></li> </ul>	<i>Group Project Presentation and Demo</i>
10	Clustering <ul style="list-style-type: none"> <li>• The k-Means Clustering with function <i>kmeans()</i></li> <li>• Hierarchical Clustering with function <i>hclust()</i></li> <li>• Density-based Clustering with package <i>fpc</i></li> </ul>	<i>Group Project Presentation and Demo</i>
11	Association Rules <ul style="list-style-type: none"> <li>• Association Rule Mining with package <i>arules</i></li> <li>• Association Rule Visualization with package <i>arulesViz</i></li> </ul>	<i>Group Project Presentation and Demo</i>
12	Network Analysis <ul style="list-style-type: none"> <li>• Network Visualization and Community Detection with package <i>igraph</i></li> <li>• Social network analysis with package suite <i>statnet</i></li> </ul>	<i>Group Project Presentation and Demo</i>
13	<i>To be determined based on enrollment</i>	

### 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=FE0h08LBD9UC>
  
- [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/>