Data Science and Business Intelligence

BU.330.780

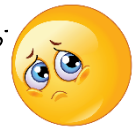
Course Administration

Instructor: Changmi Jung, Ph.D.

Welcome to DSBI

- Faculty Instructor: Changmi Jung
- E-mail: changmi@jhu.edu
- Office: HE 1331 / DC: Room 372H on Tue & Wed
- Ph.D. in Information Systems from Carnegie Mellon University
- How to contact me?
  - Before/after the class, during the break, or come visit Room 372H
  - Office hours: Thursdays 11:30 am – 1:30 pm
  - Set up an appointment: **Course Overview > Course Team**
  - Email: please allow 24 hours to respond

Note: There is only one of me, but there are many of you. Thus, please unders
that last minute questions via e-mail may not get answered on time.



Our Teaching Assistant

- Keran Jiang
  - Email: kjiang19@jh.edu
  - Carey MSIS ‘24
- Binjun Jiang
  - Email: bjiang13@alumni.jh.edu
  - Carey MSIS ‘24
- Please allow them to respond within 24 hours for the same reason
- Office hours + meeting links: Course Overview > Course Team

The Role of the Course Team

- Contact the instructor to discuss group project-related questions, course content, or general/personal guidance
- For grading-related and R coding-related questions, contact the TAs first
- Email rule
  - We appreciate shorter emails
  - Subject contains your section (T1, T2, or 51)

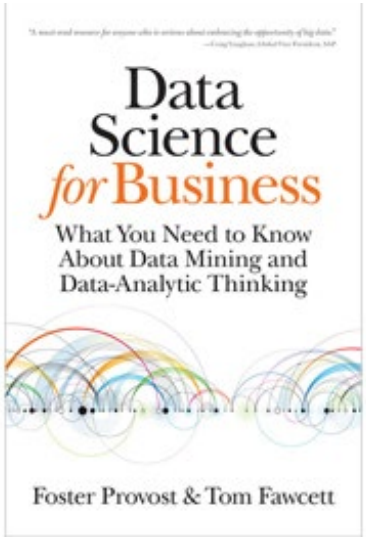
Class Policy

- Mutual respect applies to everyone, with no exception
- Join the class on time – otherwise, you are disrupting the class
- Attendance check-in: use the Qwickly App at random hours
- We have two breaks (please remind me if I forget…)
- Speak up – otherwise, cold calls await you….
- Turn off the local network: Whatever you speak must be shared with the entire class
- <u>This is an in-person class. Joining via Zoom is not allowed.</u>

Course Overview

Course Materials

- Main reference
  - Provost, F.; Fawcett, T.: Data Science for Business. What You Need to Know About Data Mining and Data Analytic Thinking

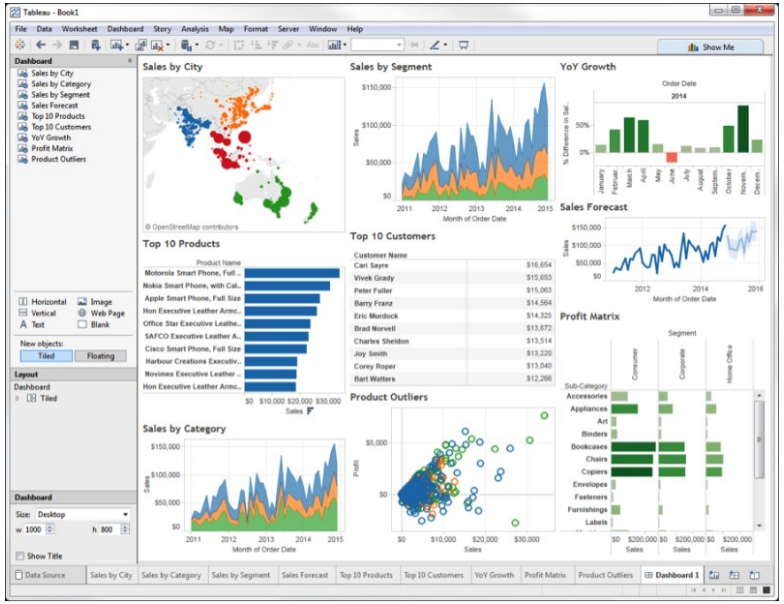


- Lecture notes: Modules > Week#
- R scripts used in class will be posted on the course Canvas before the class
- Readings (for week 7) will be posted on Canvas > e-Reserve
- You are responsible for frequently checking the course Canvas site for announcements, course materials, and updates

Software – R & Tableau

"The great beauty of R is that you can modify it to do all sorts of things, and you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants."

- Hal Varian, Chief Economist at Google



- More information about software installation and learning:
  - Canvas > R / Tableau

Source: Data Analysts Captivated by R's Power – http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all&_r=0

R vs. Python: Which one should I choose?

Category	R	Python
Ease for Statisticians	Built for stats, very intuitive	Slightly steeper for pure statisticians
Machine Learning & AI	Has tools, but less popular	Industry standard, cutting-edge libraries
Data Visualization	ggplot2, shiny, rich visualization tools	Great options (matplotlib, seaborn, plotly)
Big Data / Pipelines	Limited scalability	Strong integration with Spark, Airflow, etc.
Academic Research	Preferred in many fields	Gaining traction but still catching up in some
Industry Use	Niche (e.g., bio, pharma, finance)	Widely used across domains
Community Support	Strong in academia and data science	Huge, active, diverse community
Deployment / APIs	Harder to deploy into production	Easier to use with Flask, FastAPI, cloud
Learning Curve	Easier for statistical tasks	Easier for general-purpose programming

Course Schedule

Week	Topic	Reading & In-class Exercise	Deliverables (due before each class)
1	Introduction & Data Preparation	• Chapter 1. Introduction: Data-Analytic Thinking • Chapter 2. Business Problems and Data Science Solutions	
2	Data Visualization	• No book chapter	• Assignment 1
3	Supervised Segmentation (I)	• Chapter 3. Introduction to Predictive Modeling: From correlation to supervised segmentation	• Assignment 2 • Project proposal
4	Supervised Segmentation (II)	• Chapter 4. Fitting a Model to Data	**Quiz 1**
5	Model Validation	• Chapter 5. Overfitting and Its Avoidance	• Assignment 3
6	Model Evaluation	• Chapter 7. Decision Analytic Thinking	• Assignment 4 • Project status report
7	Data-Driven Decision Making	• Please check the course syllabus	**Quiz 2**
8	Team project presentation (Final Report submission)		

Grading Policy

- Class attendance + participation: 10%
  - Attendance
  - In-class and online participation via Canvas Discussion
- Individual assignments: 25%
- Quizzes: 30%
  - Closed book, closed note
- Group Project: 35%
  - Peer evaluation affects the score

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Attendance?

- All students are expected to attend all classes
  - Attendance will be checked at random hours
  - We use the Quickly App to read a QR code.

- A deduction system is applied
  - Each absence will result in 2 points deduction from the participation point (this will result in 2 points deduction from your total scores)
  - However, up to one absence is allowed if preapproved (sister's wedding, the birth of baby, etc.) I may disapprove due to lack of proof.
  - Notify me at least 48 hours before the class to get permission.

Participation

- In-class participation

- Online participation via Canvas > Discussion Board
  - Ask questions or answer your peers – crowdsource the ideas and solutions
  - Share tips and troubleshoot
  - Based on the degree of effort and how informative it is, your online presence will be considered class participation

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Assignments (25%)

- Four assignments in total (each one equally accounts)
- Assignments will be posted on Canvas > Modules (due weeks) or under Assignments
  - They become available approximately 6 days before the due date
  - Must complete it on Canvas **BEFORE** the class on the due date.
- Late submissions will **NOT** be accepted.
- All submitted work should be your own.
- Contact TAs for clarifications if you are ever in doubt.

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Quizzes

- There will be two quizzes (approx. <= 1 hour each): **Week 4** & **Week 7**
- It will be administered via Canvas with Respondus Lockdown Browser
- We will use the first hour of our class time. Our normal class meeting will resume after the quiz.
- There are no questions about R syntax or Tableau functions. You should have a conceptual understanding and be able to ***interpret*** results.
- Everyone must complete them **in person** – online quiz taking is not allowed
- Let me know ASAP if you require any special accommodation.

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Group Project (1/3)

- Form a group of max 4 members (by week 2)
  - Canvas > People > Team tab
  - Each team must elect a team leader (only one person)
  - Teams cannot close recruitment – should be open to all
  - Do not add your name to an existing team without notifying the members
- Solve a business problem of your interest by using data
  - Choose a problem of classifying each individual into a smaller set of mutually exclusive classes; all teams must choose a **classification** task, but there is no restriction to the business domain.
- Instruction Details and Rubrics
  - **Modules > Project >** Team Project Instructions

Group Project (2/3)

- Where to get data for your project?
  - Choose from the pre-selected datasets: sign up on the Google sheet that includes a list of data for a classification project

    Canvas > Project > Project Data

  - Or choose from the Potential Data Source
  - Or data from your work that you can share with team members (no NDA, IRB, etc. involved) → must provide data to the instructor and TAs

- Goal of the project:
  - Get a realistic hands-on experience, given the constraints of what we have learned
  - NOT developing a fancy model or algorithms with high accuracy
  - Don't worry about coming up with a novel idea
  - The important part is getting things done and understanding the whole process

Group Project (3/3)

- Key deadlines for group projects (refer to Team Project Instructions)
  - **Week 2**: Form a team, indicate the team leader
  - **Week 3**: Submit proposal (max 1 page) – not graded
  - **Week 6**: Submit 1-page status report (max 3 pages) – not graded
  - **Week 8**: Present your work and submit the final report

- Final Deliverables:
  - Presentation (time limit is TBA) – every member participates
  - Presentation slides (either PPT or PDF) – no restriction on the slides
  - Report: max 10 pages (including all appendices and references), 11 font size, single space
  - R script: submit your R markdown file (including your comments for the details)
  - Please refer to the rubrics for the expectation

How to Form a Team

- Maximum 4 members in one team (no negotiation). Do not ask me to allow you to include more members, and do not ask me to add you to a team at its maximum capacity.
  - Canvas > People > ‘Team’ tab, then choose a team and click on ‘Join’