BUS 211A-3 Foundations of Data Analytics

Fall 2022 / T,TH 7:05 pm / Lemberg Academic Center 180
Yeabin Moon, Ph.D., Senior Lecturer in Data Analytics
yeabinmoon@brandeis.edu
Office Hour / locations: by appointment / Sachar International Center 209B

Teaching Assistant: Zilin Luo (zilinluo@brandeis.edu)
TA Office Hour : 3:00 pm – 4:00 pm Wednesday by appointment

version as of 10/11/22

General Information

The universe of data science is expanding persistently. The methods and practice have undergone a dramatic shift in recent years, so the menu of modern toolkits looks confusing, even to an experienced practitioner. On the plus side, the core data analytics methods remain unchanged, especially at the early stage of the data-analytics pipeline. Data preparation is arguably the most time-consuming, although it seems to be the least discussed. It is a necessary step prior to applying any advanced statistical algorithms. In this course, we will start from the dirty data and learn a series of steps that data would move through. Also, we will learn how to manage large data using R.

Course Format

This course is designed for business-oriented professionals or students who would face various data-driven challenges in a business context. I will provide clear explanations using less technical jargon and more intuitive examples wherever possible. Most classes can be understood without much coding experience.

I structure the course around the foundational data management activities and provide the tutorials accordingly. The purpose of examples is to teach how to use them in the general case so that you would directly apply the code examples or algorithms to their projects.

I will provide a set of lecture slides and code notebooks. A notebook interface is easy to execute and imitate code snippets. R is one of the most popular programming languages both in academia and practice. A set of R libraries, like tidyverse, are extremely handy for inexperienced learners. You are welcome to use another language but need to consult me in advance.

Many successful companies address complex data analytic tasks within a team environment. Hence, this course mandates collaboration among classmates. Git is a version control system. GitHub is a web-based service for collaborative projects built on top of Git and has a user-friendly platform. We will learn what a GitHub repository is and basic GitHub terminology like commit, issue, fork, branch, push, pull request, and merge, which are essential functions for the final project.

Course Requirements

There is no mandatory textbook for this course. Instead, I will use two online-learning platforms to help your learning process: *LinkedIn Learning* and *DataCamp*. Brandeis students have free academic access to both, and I will assign selected video lectures or exercises as a reading assignment. I put the readings on the Latte.

There will be problems sets every week. These will require writings on the corresponding course topics and computer exercises. Throughout the course, we will use mobile-phone tracking data providing point-of-interest and foot traffic data for various categories provided by SafeGraph. In class, we will discuss how to gain access to this commercial proprietary data, and the assignments are designed to provide hands-on experience on this data set.

For the final project, I want you to leverage the patterns of consumers in the Boston area, especially since January 2022 using SafeGraph. Each group will act as a data professional team and present their idea and workflows in class. In your presentation, you should describe:

1) What is your specific area of interest? Why is it more important than other areas?
2) Why is the cell phone tracking data fit for your interest?
3) What is the main challenge in answering the questions?
4) To answer the question, how do you manipulate/refine data?
5) How do you validate your data is properly processed?

Note that the focus is on the workflows, not your final answers. This project mimics the early state of the data analytics pipeline.

The project will involve four stages:

1. Submit the outline of the project by October 13th. It should contain some ideas on (1) and (2). Your team should contact me after. You can find me during office hours or through zoom. The question must be very specific but reasonable to achieve.
2. Submit 3 pages initial report by November 4th. The report must address (1)—(3).
3. Schedule meetings on November 21th—25th to discuss your progress during office hours or zoom.
4. Deliver group presentations in the last 2 class sessions.

Further guidelines will be provided once the class size is finalized. We will discuss access to DataCamp, and SafeGraph in class.

Please consult University Writing Center. Each report should be WELL prepared for submission. Good writing counts.

Learning Goals

With this course, you will

- Gain intermediate proficiency in programming
- Understand several aspects of the data analytics life cycle
- Evaluate various algorithms and approaches for the given task and data
- Develop collaborative skills
- Be familiar with real market data

Class Participation

There is no such thing as a stupid question. Dialogue is not only strongly encouraged, it is critical to your understanding of the material. Vocalizing your questions often helps you solidify what you do and do not understand. It also provides me important feedback on the areas in which we need to spend more time. During lectures, I will encourage questions, and I will solicit input. If I call on you, please relax, I am NOT trying to intimidate you or embarrass you in any way. I am trying to encourage active listening and keep you engaged in the course. This will greatly assist you in learning the material. If you do not know the answer, I will move on to another student. Hence, attendance is mandatory for this class.

Workload expectation

I expect you to study for about three hours for every hour of in-class time. However, it might vary depending on your prior proficiency in data analytics. I highly encourage you to try out more exercises on DataCamp or LinkedIn Learning, which might cost you significant time. Learning a data analytics or a programming language, in general, is like learning a foreign language. There is no shortcut to mastering it.

Grades

Individual assignments (60%): You will have 7 days to complete each assignment. In almost every week, the problem sets will be assigned during the semester. The detailed guidelines will be provided. Collaboration is encouraged, but make sure that this is an individual assignment.

Final group project (30%): See the details above.

Class participation will constitute the remining 10%.

I follow the Brandeis grading system on a 4-point scale. You will find it in LATTE.

Accommodations

Brandeis seeks to create a learning environment that is welcoming and inclusive of all students, and I want to support you in your learning. Live auto transcription is available for all meetings or classes hosted on Zoom and you can turn it on or off to support your learning. Please check for Zoom updates to take advantage of this new feature. To learn more, visit the Zoom Live Transcription webpage. For questions, contact help@brandeis.edu

If you think you may require disability accommodations, you will need to work with Student Accessibility Support (SAS) (781-736-3470, access@brandeis.edu). You can find helpful student FAQs and other resources on the SAS website, including guidance on how to know whether you might be eligible for support from SAS. If you already have an

accommodation letter from SAS, please provide me with a copy as soon as you can so that I can ensure effective implementation of accommodations for this class.

Academic Integrity

Every member of the University community is expected to maintain the highest standards of academic integrity. A student shall not submit work that is falsified or is not the result of the student's own effort. Infringement of academic integrity by a student subjects that student to serious penalties, which may include failure on the assignment, failure in the course, suspension from the University or other sanctions. Please consult Brandeis University Rights and Responsibilities for all policies and procedures related to academic integrity. Students may be required to submit work via TurnItIn.com or similar software to verify originality. A student who is in doubt regarding standards of academic integrity as they apply to a specific course or assignment should consult the faculty member responsible for that course or assignment before submitting the work. Allegations of alleged academic dishonesty will be forwarded to the Department of Student Rights and Community Standards. Citation and research assistance can be found at Brandeis Library Guides - Citing Sources.

Classroom Health and Safety

- Register for the Brandeis Emergency Notification System. Students who receive an emergency notification while attending class should notify their instructor immediately. In the case of a life-threatening emergency, call 911. As a precaution, review this active shooter information sheet.
- Brandeis provides this shuttle service for traveling across campus or to downtown Waltham, Cambridge and Boston.
- On the Brandeis campus, all students, faculty, staff and guests are required to observe the university's policies on physical distancing and mask-wearing to support the health and safety of all classroom participants. Review up to date COVID-related health and safety policies regularly

Student Support

Brandeis University is committed to supporting all our students so they can thrive. If you want to learn more about support resources, the Support at Brandeis webpage offers a comprehensive list that includes these staff colleagues you can consult, along with other support resources:

- The Care Team
- Academic Services (undergraduate)
- Graduate Student Affairs
- Directors of Graduate Studies in each department, School of Arts & Sciences
- Program Administrators for the Heller School and International Business School
- University Ombuds
- Office of Equal Opportunity

Student Support

The following table shows the course-plan framework. It is subject to change if the surrounding circumstances dictate so. Please visit LATTE frequently for further announcements.

Class	Topics and Readings	References / requirements
1	Course Introduction - Course overview - Requirement reviews	DataCamp exercise - Introduction to R
2	Data analysis pipeline - Lifecycle of data science project - Data vs. Task Introduction to Algorithms I - How to approach tasks - Writing a GOOD code * SafeGraph introduction	
Fundamentals		
3	Introduction to Algorithms II - Big O notation - Role of Mathematics Communication in data science - Git, GitHub	Assignment 1
4	Introduction to Data structure - Data wrangling World of Computer Languages - Command line approaches - Cluster/cloud computing system - Entry problems - SQL	
5	Getting sense of R programming - Introduction to R analysis - R markdown	
6	Data Structures in R - Understanding tidyverse/tidyr	
7	Advanced features in R I - Control flow, iterations, and functions - vectorization	Team Assignment 1
8	Advanced features in R II - String analysis - Other types of data - Geocode	
9	Data storage I - Introduction to SQL database	
10	Data storage II - Intermediate SQL	
Theory and Practices		
11	Probability and Statistics I - Introduction to Sampling - Distributional analysis	

	- Measuring variability	
12	Probability and Statistics II - Introduction to Probabilities - Conditional probabilities - Bayes problems	
13	Data Preparation Technique Overview	
14	Imputation - Outliers / missing data - Methods of imputations	
15	Features selection - Common questions - Data type problems - Input/output problems - RFE / Feature importance	
16	Data scaling - Numerical scaling methods review - Robust scaling - Encode problems	
17	Challenge of Gaussian distribution - Power transform - Box-Cox transform - Quantile transform	
18	Introduction to high dimensional data - Big data? - Dimensionality reduction - Technique overview	
19	Introduction to statistical learning - Conceptual framework of learning problems - Roadmaps in our program	
20	Introduction to Natural language processing - Hello world? Hello world! - Machine problems vs. language problems - Application in academia / practices	
21	ENCORE	
22	Final Project showcase 1	
23	Final project showcase 2	
24	Epilogue	