BUS 211F(1): Analyzing Big Data I
Course Syllabus (Draft – Subject to change)
Fall 2018

Course Instructor	**Course Logistics**
Bhoomija Ranjan	**Timing:** Mon/Wed, 3:30 PM – 4:50 PM
Assistant Professor of Marketing	**Class Location:** Pollack Fine Arts Auditorium 001
Email: branjan@brandeis.edu	**Office Location:** Sachar 214A
	Office Hours: Thu/Fri 3:00 PM – 4:30 PM, and by appointment
	Office Hour Location: Sachar 115
Course TA: Yaxian Lin	**TA Office Hours:** Tue 2PM – 3PM
TA Email: rin@brandeis.edu	

Course Information

We are now in an era of voluminous real-time data, fast affordable computing, and a data-driven decision making. Firms across the spectrum are gathering detailed real-world data on their customers, competitors and marketplace. This vast volume of information will provide significant competitive advantage to companies and managers who can leverage these data and derive meaningful managerial insights. It has been designed for managers who will be using analytics on large datasets, and so is intended for students wanting to go into marketing, finance, consulting, entrepreneurship, business strategy and operations management.

More specifically, the course provides experience with applications of data analytics specifically to business decisions. The learning objectives for this two-credit module are to learn how to:

- Define the decision problem and determine what information is needed
- Acquire trustworthy and relevant data and judge its quality. This includes-
    - An introduction to the data analysis process
    - The role of the data analyst, the value of big data, and the fundamental relationships between data and models
    - The R programming language so that students will have the tools needed to work complex data sets
- Analyze data to make strategic and tactical business decisions by
    - Assessing relationships between data entities via the SQL language

- Creating high-impact visualizations

Course Format

The course meets twice weekly and will rely on a blend of required readings, lectures, business cases, homework assignments and a course project to master the material. This course is designed to provide students some techniques for managing information abundance and for attending to the costs and benefits of information processing in decision contexts. As such, we start with typical decision-making situations in business settings and work towards making data-driven decisions. Readings focus on the theory of decision-making, data structure, and analytic models. In addition, articles and cases illustrate typical decision problems and the application of the techniques we will study.

Course Pre-Requisites

Students should have a background in statistics, such as ECON210f, FIN212a or equivalent. **Students not satisfying these requirements should contact the instructor.**

Learning Goals

Upon successful completion of this module, students will:

- Think of data as a strategic resource in business.
- Understand the logic of complex data queries in the context of on-line business research sources.
- Be familiar with current developments in Big Data, business intelligence, and competitive analytics.
- Be able to design a relational database structure suited to a business enterprise.
- Understand the underlying concepts for informational visualization.
- Be familiar with the R language and data-handling techniques

Course Materials

Required

1. Harvard Business Publishing course: articles and cases https://hbsp.harvard.edu/import/563532
2. Other course readings will be posted on the LATTE site

Recommended but Optional

1. UCLA Institute for Digital Research and Education Webpage: https://stats.idre.ucla.edu/r/
2. R Programming Tutorial: There are numerous R-tutorials available online. One I recommend is http://www.cyclismo.org/tutorial/R/ sections 1-7, 13 and14.

Course technology

In addition to software available on the IBS computer clusters, we will also use web-based resources, including some made available through the Teradata Student Network (TSN). This site, sponsored by Teradata, the Walton School of Business at the University of Arkansas, IBM, MicroStrategy, SAS and others, is a gateway to articles, cases, software tools and real corporate data. Details about use of TSN will be provided separately on our LATTE site.

Of particular importance are R, Teradata and DataCamp:

• **R**: R is a free software environment for statistical computing and graphics, and is widely used by both academia and industry. The advantage of the R software is that it can work on both Windows and Mac-OS. It is ranked no. 1 in the KDnuggets 2015 poll on top languages for analytics, data mining, and data science. **RStudio** is a user friendly environment for R that has become popular. ***Each student should download these two programs***: Please note: even if you already have them, please check for updates.

R Software: http://www.r-project.org/index.html.
RStudio: http://www.rstudio.com/products/RStudio/#Desk

NOTE: IN GENERAL, YOU SHOULD NOT EXPECT TO RUN R OR RSTUDIO ON A MOBILE DEVICE, THOUGH YOU CAN ACCESS THE RSTUDIO SERVER THROUGH A MOBILE WEB BROWSER.

• **Github** is also a free environment that facilitates (a) collaborative work and (b) version control for software projects that are under development. It is very widely used by data scientists to manage and share their work.

• **DataCamp** is an online resource that offers interactive R and Python courses on topics in data science, statistics and machine learning (www.datacamp.com). During this course, you would be automatically enrolled in DataCamp (for the duration of the course) and would learn in the comfort of your browser with video lessons and coding challenges.

• **Teradata SQL** (pronounced "sequel") **Assistant** will provide exposure to writing Structured Query Language code to interrogate a large database

compiled by Nielsen, a publicly traded, multinational data measurement company.

- **LucidChart** is a cloud-based tool that enables users to design a coherent data structure as part of the process of designing a database. We will use a free version.
- **Interactive Visualization Tools (further instructions to be distributed in class):** These powerful applications (such as Tableau) can speed up some processes and allow much flexibility in depicting data visually. **More details will be provided later.**

Course Evaluation

Your final course grade will be computed on the following criteria:

Factor	% weight
Individual DataCamp assignments	20%
Contributions to Class Discussions	10%
Brief analyses (3—averaged together)	30%
Projects (2)	40%
TOTAL	100%

Class contributions. Class participation is important in this course both as a means of developing understanding and as an indicator of student progress. Students are expected to contribute actively, freely, and effectively to the classroom experience by raising questions, demonstrating preparedness and proficiency in the analysis of problems and cases in the class readings, and explaining the implications of particular analyses in context. Homework-based discussion is an important part of participation. ***To this end, regular class attendance is required, and students should use name cards.*** We meet only six weeks, so absence can become a serious problem. Even if you must arrive late or leave early, be here.

I will evaluate the quality of your contributions in class each session, as well as the quality of your contributions via email, LATTE discussion, etc. These will all be factored together in determining your ultimate Contributions grade. In general, absence from class reduces your contribution grade.

DataCamp, Written Assignments and Projects: Each student will *individually* complete 20 assigned chapters in our course group on the DataCamp website. The assignments will teach basic technical skills related to R and SQL, and will help to ensure that each student in the course really acquires a solid technology foundation.

Students will also complete five written assignments during the course. Three (3) of these will be brief analyses, requiring modest analysis and writing. Two other written assignments will be "Projects" requiring more significant time and analysis. The projects will include written and

computer-based elements. The analyses can be done in groups of 1-2 students, while the projects can be done in groups of 3-4.

This semester, I will assign each of you to a team of 3 to 4 people for the two projects, based on the skills you identify in our Data Science Profile. Early in the course I will create the teams, balancing the skills and experience of team members. Your project teams will remain together for the full module as well. All group members will receive the same grade for the project. At a minimum, all group members must participate fully in the project, including attending group meetings, preparing an analysis plan, conducting analyses, and writing up and presenting the project report.

DataCamp assignments are accessed through LATTE and are scored automatically. All other assignments should be submitted via LATTE upload prior to the start of class. ***Please check your DataCamp assignment grades in LATTE regularly and report any discrepancies.*** Papers should be professional in appearance and use clear, grammatically correct business English. Analytical work (graphs, tables, and other output) should be incorporated seamlessly into the written document, showing readers exactly and only what you want them to see.

Class Conduct

Use of technology in the classroom. You are allowed to use laptops and tablets during the lectures, which should only be used for coursework related activities and not for email, social media, or other activities not directly related to the course. Cell phones must be turned off or silenced during class. No photography or recording of any kind is allowed without explicit permission from the instructor.

Respect during class discussions: On some days we will discuss assigned cases. We have a shared responsibility to create a classroom environment where all voices and ideas can be expressed and different viewpoints can be exchanged. This often calls for active listening and for simply waiting to be recognized before you speak. Individuals vary widely in their willingness and desire to speak, as well as in their confidence with the English language. I will insist that everyone participate at some level—which means you should expect to be called upon, and no one will be allowed to dominate or monopolize class discussion.

Late assignments. Late assignments will not be accepted without my prior permission, and will incur a penalty unless the circumstances are exceptional.

Academic Honesty: You are expected to be honest in all of your academic work. Please consult Brandeis University ***Rights and Responsibilities*** for all policies and procedures related to

academic integrity. Students may be required to submit work to TurnItIn.com software to verify originality. Allegations of alleged academic dishonesty will be forwarded to the Director of Academic Integrity. Sanctions for academic dishonesty can include failing grades and/or suspension from the university. Citation and research assistance can be found at LTS - Library guides.

Disabilities: If you are a student with a documented disability on record at Brandeis and wish to have a reasonable accommodation made for you in this class, please see me immediately.

Workload Expectation

As this is a two-credit course, you are expected to spend a minimum of 9 hours of study time per week in preparation for class (readings, exercises, assignments, preparation for exams, research, etc.).

Communications and Getting Help

We'll make regular use of LATTE. All lecture notes, handouts, assignments, and supporting materials will be available via LATTE, and any late-breaking news will reach you via email. Please check your Brandeis email and the LATTE site regularly to keep apprised of important course-related announcements.

If you are hesitant to participate for any reason or if you have questions about anything, please come and see me. I am happy to help. Please contact me for assistance for any reason, or if you have questions, comments, or concerns about the course. All of my contact information is on the cover page of this syllabus

Course Plan

Session	Date	Topics and Readings	Complete and/or Upload to LATTE *before* class
1	08/29 Wed	**Finding Business Value in a Sea of Data** **BEFORE CLASS: Complete LATTE survey + "Data Science Profile"** *READINGS:* 1. *Davenport HBR Article "Competing on Analytics"* 2. *Thomas, Rob. "Transforming Farms with Data" blog post (LATTE)* 3. *"At Amadeus, Finding Data Science Talent…" (LATTE)* *AGENDA:* *a.* Course introduction and objectives	(LATTE survey and Data Science Profile)

		b. Competing on Analytics: Opportunities  **DataCamp: Begin 3-chapter course Introduction to the R Studio IDE (complete by Sep 5)**	
	09/03 No Classes – Labor Day		
2	09/05 Wed	**First Look at R and RStudio Environment** *READINGS:* *None, work on DataCamp* *REFERENCES:* Grolemund & Wickham, Chang *AGENDA:* *a.* Server vs. local installations *b.* R, R Packages, local and external data *REQUIRED:* ***Install R and RStudio before Class***	Install R and RStudio before Class; Complete DataCamp R chapters by now
3	09/06 Thu	**BRANDEIS MONDAY** **Preparing Data for Analysis** *READINGS:* 1. *HomeZilla: Attracting Homebuyers through Better Photos (HBP)* 2. Case Assignment: *Airbnb (HBP)* 3. *Davenport HBR Article, "Competing on Analytics" (revisited)* 4. *Wickham, Hadley. "Tidy Data" (LATTE)* *AGENDA:* *a.* Analysis 1 discussion *b.* Need for & Strategies for Data Cleaning *c.* Why isn't all data naturally tidy?  **DataCamp: Start 5 chapters from Data Manipulation from dplyr (complete by Sep 12).**	Analysis 1 due before class
	09/10 No Classes – Rosh Hashanah		
4	09/12 Wed	**Data Visualization—Finding Patterns in Voluminous Data** *READINGS:* 1. *HomeZilla: Attracting Homebuyers through Better Photos (HBP)* *AGENDA:* *a.* Visualizing data effectively  **DataCamp: Start 4 chapters from Data Visualization with ggplot2 (complete by Sep 17).**	Complete DataCamp Data Manipulation Chapters by now

5	09/17 Mon	**Effective Data Visualization** *READINGS:* 1. *The Hidden Traps in Decision Making (HBR Classic)* 2. *Link to Peng's "Plotting Systems in R"* *AGENDA:* *a.* Graphing in R Demo	Complete DataCamp Data Visualization Chapters by now
	09/19, 09/24 – No Classes - Yom Kippur, Sukkot		
6	09/25 Tue	**BRANDEIS MONDAY** **Ethical Concerns for Analytic Competitors** *READINGS:* 1. *McKinsey Report Exec Summary (LATTE)* 2. *Barocas, S. and Nissenbaum, H. "Big Data's End Run Around Anonymity and Consent" (LATTE)* *AGENDA:* *a.* Analysis 2 discussion *b.* Interactivity in Visualizations *c.* Visualization for Business Intelligence	
7	09/26 Wed	**Data Organization** *READINGS:* 1. *HomeZilla: Attracting Homebuyers through Better Photos (HBP)* *REFERENCES:* 1. *G. Russell reading on Databases (LATTE; Chap 1 & 2)* *AGENDA:* *a.* Essential concepts: metadata, structure, data quality *b.* Structure of Relational Databases *c.* Entity-Relationship Diagramming	Analysis 2 due today
	10/01 – No Classes - Shmini Atzeret		
8	**10/03 Wed**	**SQL Part-1** *REFERENCES:* 1. *G. Russell reading on Databases (LATTE; Chap 3)* 2. *Nielsen Customer Data* *AGENDA:* *a.* Project 1 discussion *b.* Becoming bilingual: SQL and R c. Reproducible explorations d. SQL: one language, many dialects Fundamental commands: SELECT, FROM, WHERE, JOIN, computation of a new column (AS), case sensitivity,	Project 1 due today

		relational operators **DataCamp: Begin 4-chapter course on SQL for Data Science (complete by Oct 10)**	
9	10/08 Mon	**SQL Part-2** *REFERENCES:* 1. *G. Russell reading on Databases (LATTE; Chap 3)* 2. *Selected Teradata material about SQL* *AGENDA:* *a.* Teradata interface *b.* Nielsen Customer Data queries	Complete your Teradata registration (instructions on LATTE)
10	10/10 Wed	**SQL Part-3** *REFERENCES:* 1. *G. Russell reading on Databases (LATTE; Chap 3)* 2. *Selected Teradata material about SQL* *AGENDA:* *a.* More SQL commands – data joins **DataCamp: Begin 4-chapter course on Joining Data in PostgreSQL (complete by Oct 17)**	Complete DataCamp SQL Chapters by now
11	10/15 Mon	**Other issues in Data Preparation** *AGENDA:* *a.* Analysis 3 discussion *b.* Analysis-driven restructuring *c.* Transformations for analysis in R and SQL Missing data	Analysis 3
12	10/17 Wed	**Interactivity for Business Intelligence** *AGENDA:* *a.* Discussion of Random Assignment of visualization softwares	Complete DataCamp PostgreSQL Chapters by now
13	**10/22** **Mon**	**No Class Meeting today** • **Project due before midnight in lieu of final exam** • **Prof Ranjan will be in classroom for group assistance**	Project 2

Brief Description of Assignments (complete assignment details to be distributed in class):

Analysis 1	Brief analysis of “Airbnb” case in Davenport framework
Analysis 2	Visualization Exercises using Homezilla data
Analysis 3	ERD exercise for Homezilla Case
Project 1	Analyzing housing prices with Homezilla data using R
Project 2	Design and code database queries using SQL on a remote large-scale database

Useful technical references:

Russell, G. *Database eLearning*. online at https://db.grussell.org/index.html.

Grolemund, G. and Wickham, H. *R for Data Science* (2017). online at http://r4ds.had.co.nz/

Chang, W. *Cookbook for R.* Online at http://www.cookbook-r.com/

Zhang, Y. *R and Data Mining.* Online pdf through LATTE