BAN 7100-80: Data Mining and Data Warehousing

Myles D. Garvey, Ph.D

Winter, 2020

E-mail: garveym2@wpunj.edu
Office Hours:By Appointment
Office: Valley Road Room 3070
Phone: 973-720-3928

In-Class Hours: Online
Class Room: Valley Road 2014

Contents

1 Required and Recommended Materials

Required Textbooks:

We will use an assortment of notes and papers that are posted weekly.

2 Course Description

This course covers two related topics in the age of big data and knowledge discovery: data warehousing and data mining - their concepts, principles, and techniques. Topics in data warehousing include data warehouse/data mart architecture; multi-dimensional model design; extracting, transforming, and loading strategies. Topics in data mining range from statistics to machine learning; including techniques of clustering, association rules, and classification. OLAP (On-Line Analytical Processing) applications and business intelligence are also introduced.

3 Prerequisites

You should have a solid understanding of Algebra. Calculus and probability theory is required. BAN 5000: Calculus for Business Analytics is a prerequisite for this course.

4 Course Format

This class will be conducted online and comprise of mostly readings that are posted on Blackboard. You will be required to complete the readings and complete assignments that test your knowledge on them. Our course will comprise of 8 modules. Each module is posted on the indicated days below at 12am on that day. The structure of each module will usually comprise of 1 - 3 academic articles and/or a small set of notes that I have written, a mini-workshop which will test your skill and application of knowledge in R, and finally a mini-assignment. Each module should take you at most 2 days to complete.

5 General Course Policies

1. Important announcements will be made in class and on Blackboard.
2. Final course grades are final. Let me repeat this. **Final course grades are final!** Changes will only be made if there is a mistake in the calculation of the final grade, but legitimate evidence suggesting the contrary must be presented to the professor. "Legitimate" constitutes the use of the professor's calculation in grade mismatching with the grade received. See below for more detail. It does NOT include a mistake made on a particular assignment or exam or project. Please keep in mind that grades are NOT rounded. So if you receive a 89.99, this constitutes a B+, not an A-. Do NOT request me to change a grade due to the closeness of a letter grade. I'm informing you right now, this will not happen! Same for other grading boundaries.

3. As you may know, it is against university policy to cheat. It is a very serious violation of academic integrity. Please note that if cheating of any kind is observed in/out of the class, you will be reported to a higher authority in accordance with university policy on academic dishonesty. Likewise, you will be given a 0 for cheating.

4. I do not give extra credit just because you are falling behind. Please do not request me to do so.

5. All course material is posted on Blackboard.

6. It is YOUR responsibility, not mine, to keep track of your grades. With that said, ensure that you use the formula indicated below to get an idea of your standing in my course. The "Total Score" grades on Blackboard **do not** properly reflect your grades. In order to determine your grade in the course, you must use the equation indicated below in this syllabus. Failure to keep track of your own grade is not an excuse for additional points, extra credit or additional revision on assignments outside the grace period for review for said assignments. It also is not a valid reason to contest a final course grade. If you receive a grade of F at the end of the semester, please keep in mind that this is not reason for me to change a final grade due to your lack of supervision of your own grade. If you need me to clarify or project what your final grade will be, or give you a comment on your progress in the course, I will be happy to do so, but YOU MUST first initiate that type of discussion with me, and do so **well before the semester ends**.

7. If you are having difficulty due to a death in the family, financial problems, or other personal issues, I MUST receive an email from you **in advance**. I do not accept such requests after due dates of assignments.

8. ALL submissions of anything in this course are digital. Furthermore, ALL documents MUST be of PDF file format upon submission as well as .R format. **DOC, TXT, DOCX, etc will NOT be accepted as valid submissions**. I have a hard absolute policy with this. So please, remember, don't let your hard work result in a 0 all due to you not submitting a PDF. We live in the 21^{st} century, and ALL modern operating systems have the easy ability to convert any of the aforementioned types into a PDF format. HENCE, please make sure that submissions are indeed in PDF format. Last, and most importantly, hard-submission (paper-format) WILL **NOT** be accepted as a valid form of submission.

6 Evaluation

Workshop Sets (8 Total)	3.125% each
Mini-Assignments (8 Total)	9.375% each

Numerical Grade	Letter Grade
$[94, 100]$	A
$[90, 94)$	A-
$[87, 90)$	B+
$[83, 87)$	B
$[80, 83)$	B-
$[77, 80)$	C+
$[70, 77)$	C
$[68, 70)$	C-
$[65, 68)$	D
$[0, 65)$	F

7 Blackboard

All submissions for everything are conducted through blackboard. Make SURE you have access to this! Everything will equally be posted on blackboard. Another side-note, please DO NOT EMAIL ME via blackboard. If you need to email me, please do so DIRECTLY from your WPUNJ email account.

8 Workshops

There will be a total of 8 Workshops. Each workshop is comprised of a combination of reading and manipulating provided R-Code to solve problems in data warehousing. Let me emphasize that all you are basically doing is typing in code, making very small changes, and reporting to me the output. I want to see the result of each workshop IN PDF format. Failure to do so WILL result in 0. If you complete all questions in the workshop, as well as all code, you receive full points. Each response to each question of each workshop must have:

- The original question
- The code for the question number
- The output for the code that was run

Do NOT submit to me a .R file. Each file must be a PDF. My recommendation: type all of your workshop code in a file, run the entire file in R, copy and paste the entire console input/output into Word, number the questions in word.

9 Assignments

You will also be required to complete 8 mini-assignments. Each assignment is NOT long, and it should only take you at MOST 1-hour to complete. They will range from 3 - 10 questions. Assignments are to be written in Word and submitted in PDF format. Just like the workshops, I will ask that you submit your code and output in PDF format. My suggestion: type all your code in a .R file, run it, copy/paste the input/output from the console to word, organize each one by question number. Some questions will necessitate a response other than code! Each response to each question for each assignment should be as follows:

10 Course Schedule and General Syllabus Changes

10.1 Policy Regarding Changes to the Syllabus

I try my best to stick to this schedule and grade distribution. With that said, if I see that you need more time to absorb the material, some homework assignments or topics may be extended. If I begin to see that our grading distribution in regards to assignments becomes too much, I will need to redistribute points. Again, **this is a very rare situation that happens**. However, I understand that some of you may have more difficulty in absorbing the material than others. With that said, I want to ensure we spend the proper amount of time on each topic. If we need to change anything about this course, I will update the syllabus and post it.

If I need to make changes to the syllabus, you will be notified, on Blackboard, and via email. Changes to the syllabus may come as a consequence of in class discussions. I always consult with my students first before making any necessary changes. **Please note that if you miss class, especially on a regular basis, then your voice on such matters may not be heard**. I only take into consideration comments made within a class session. Moral of the story: attend class and you will have your voice heard if such a need for change arises.

10.2 Tentative Course Schedule

- **Part 1: What is Data?**
  - Module 1 (12/28): A Review of R
  - Module 2 (12/30): An Overview of Data, Data Structures, and Extract-Transform-Load (ETL)
- **Part 2: How and Where Do We Get Data?**
  - Module 3 (1/6): Application Programming Interfaces (APIs)
  - Module 4 (1/9): Fundamentals of Web Scrapping
- **Part 3: How Do We Mine Data?**
  - Module 5 (1/12): What Does it Mean to "Analyze" Data?
  - Module 6 (1/15): Descriptive and Correlation Analysis, Clustering, Classification, and Association Learning
  - Module 7 (1/18): Fundamentals of Textual Analytics
- **Part 4: Databases**
  - Module 8 (1/21): Fundamentals of Unstructured and Structured Database Systems, Fundamentals of Structured Query Language (SQL)