Syllabus for MAT 328: Techniques in Data Science

Course Description MAT 328: *4 hours, 3 credit*s. Analyzing data sets to extract new insights. Acquisition, data mining, storage, and visualization of real world data using scripting and statistical programming languages. Application of standard statistical tools including hypothesis testing, Bayesian analysis, bootstrapping and regression. Classifying and clustering multidimensional data sets via dimensionality reduction and machine learning techniques.

Textbooks:
See course readings and references in schedule on course webpage.

Supplemental:
• *Computational and Inferential Thinking: The Foundations of Data Science* by Ani Adhikari and John DeNero. **https://www.inferentialthinking.com/chapters/intro.html**
• *Online Statistics: An Interactive Multimedia Course of Study*. Project leader: David M. Lane. http://onlinestatbook.com/2/index.html
• *How to Think Like a Computer Scientist (Python)* by Jeffrey Elkner, Allen B. Downey, and Chris Meyers. http://interactivepython.org/runestone/static/thinkcspy/index.html

Grading: The grading for the course will be based on:

Assignments	20%
Project	25%
Midterm	20%
Final exam	35%

You must take and pass the final exam to pass the course.

Assignments: All assignments should be submitted on Blackboard. **No late assignments are accepted.** Solutions to each assignment will be posted immediately after the due date. You will then have an additional week to correct your submitted assignment, and the corrected assignment will be graded. You can miss up to 2 assignments without affecting your grade (if you turn in all assignments, we will drop the lowest 2 scores).

Project: In the semester-long project, you will analyze a data set of your choice. 10-15% of the project grade will be distributed during the semester for meeting milestones, while the remaining 10-15% of the project grade will be based on a presentation and evaluation of the final project.

Midterm: The midterm is in class on Tuesday October 29.

Final exam: The final exam is required and will be on Tuesday December 17 from 3:45-5:45pm. You must pass the final exam to pass the course.

Jupyter Hub: Labs and assignments will be done in a Jupyter Notebook, which can be accessed through Jupyter Hub (under Lehman Apps on **Lehman One Access**). To access Jupyter Hub off

campus (ex. at home), you will have to connect to the Lehman Virtual Private Network (VPN) first (see instructions on course webpage).

Use of Technology & Blackboard: Homework solutions and grades will be posted on Blackboard. The Blackboard system is provided by CUNY to all enrolled students. If you have not accessed Blackboard or are having difficulties, contact Blackboard Support in the Information Technology Division. You can also visit the Help Desk in the Computer Center (first floor, Carman Hall) in person. They can reset passwords and help with simple Blackboard issues.

Honor Code: You are encouraged to work together on solving the homework problems. However, you should write up the solutions on your own. You are responsible for knowing and following Lehman's academic integrity code (available from the Undergraduate Bulletin, Graduate Bulletin, or the Office of Academic Standards and Evaluations). All incidents of cheating and plagiarism will receive a 0 on the test or assignment, and will be reported to the Office of Student Affairs.

Accommodating Disabilities: Lehman College is committed to providing access to all programs and curricula to all students. Students with disabilities who may need classroom accommodations are encouraged to register with the Office of Student Disability Services. For more info, please contact the Office of Student Disability Services, Shuster Hall, Room 238, phone number, 718-960-8441.

Learning Objectives:
At the end of the course, students should be able to:

1. Acquire data sets from multiple sources.
2. Use data mining to extract new insights about the data.
3. Understand basic storage techniques and constraints.
4. Analyze data using standard techniques from statistics and linear algebra.
5. Visualize data using standard packages.