DS 5110: Introduction to Data Management and Processing
Khoury College, Northeastern University
Fall 2024
Course Syllabus

1. General Course Information

Course Description: Data science is the discipline concerned with transformation, processing, management, and modeling of data for the purpose of extracting knowledge from raw observations. This course discusses the practical issues and techniques for data importing, tidying, transforming, and modeling. Programming is a cross-cutting aspect of the course. Students will gain experience with data science tools through short assignments. The coursework includes a term project based on real-world data.

Number of Credit Hours: 4 Credit Hours.

Instructor and GTA Information

- **Dr. Fatema Nafa**
- **Office**: 140 Fenway
- **Schedule**: Tuesdays
- **Location**: West Village G 104(04) and Shillman Hall 320(07)
- **Telephone**: +1 (617) 373 -XXXX | E-mail: f.nafa@northeastern.edu
- **Office Hours:** Tuesdays 3:00 PM to 4:30:00 PM. or **by appointment.**
  - ✓ Please send me an email if you would like to meet at some other time.
  - ✓ During the office hours, I can meet with you either in-person or virtually through Zoom or Microsoft Teams.
  - ✓ Please note that the office hours may vary from week to week. Check the announcements posted on the course website on Canvas.

Graduate Teaching Assistant (GTA)

Email: deshpande.ana@northeastern.edu
Office Hours: Tuesdays and Thursdays
Zoom Meeting Link for Office Hours: https://northeastern.zoom.us/j/97610886053

Co-REQUISITE: None.

Recommended Textbook

The core concepts presented in this course are extracted from the following textbook, though the order of concepts and the depth of coverage may differ. Even though this book is good reference material for this class, the handouts given during lectures and posted on Canvas will be self-contained.

- **Database Systems:** The Complete Book, Hector Garcia-Molina, Jeffrey Ullman, Jennifer Widom Deitel, Paul, and Harvey M. Deitel**.**
- **Intro to Python for Computer Science and Data Science.** Pearson Education, **2020**.
- **Data Science for Business** by Foster Provost and Tom Fawcett
  Focuses on the practical application of data science concepts in business, including data cleaning, modeling, and visualization.

- **Python for Data Analysis** by Wes McKinney
  Essential for learning data manipulation, transformation, and analysis using Python, particularly with pandas, a key data analysis library.

REQUIRED SOFTWARE

- **Anaconda** - a comprehensive toolkit for data science and machine learning, providing an easy-to-use platform for Python and R languages.
- **Overleaf** - an online LaTeX document generator, ideal for crafting technical reports.
- **Google Colab** - an online Jupyter Notebook platform offering free GPU access.
- **GitHub:** a web-based platform for version control and collaboration, allowing users to manage and store their code projects efficiently.

1. COURSE OBJECTIVES AND LEARNING OUTCOMES

COURSE OBJECTIVES: The objective of this course is.

- **Understand Data Science Workflows:** Gain a clear understanding of the steps involved in managing and processing data within the lifecycle of a data science project.
- **Master Data Handling Techniques:** Develop proficiency in importing, cleaning, transforming, and preparing data for analysis.
- **Implement Data Modeling Practices:** Learn to apply statistical and machine learning models to data, interpreting and validating the outcomes effectively.
- **Enhance Programming Skills:** Improve programming abilities in a data science context, with an emphasis on using Python and R for data manipulation and modeling.
- **Utilize Data Science Tools:** Become familiar with key data science tools and libraries, such as pandas, NumPy, scikit-learn, and TensorFlow, to handle complex data science tasks.
- **Conduct a Data Science Project:** Execute a comprehensive data science project from start to finish, using real-world data to extract actionable insights.
- **Collaborative and Independent Learning:** Foster skills in both independent learning and teamwork, critical for tackling diverse challenges in the field of data science.

LEARNING OUTCOMES: Upon successful completion of the requirements of this course, students will be able to:

- **Demonstrate** the ability to import, clean, and prepare large datasets for analysis.
- **Use** various data transformation techniques to manipulate data and prepare it for modeling.
- **Build** predictive models and validate their effectiveness in solving real-world problems.
- **Show** proficiency in programming with Python and R, specifically for tasks related to data management and processing.
- **Employ** current data science tools and libraries proficiently in practical scenarios.
- **Complete** a term project that reflects comprehensive data management, processing, and modeling skills, resulting in actionable data-driven insights.
- **Communicate** technical results effectively to a non-technical audience, highlighting the insights and business impacts derived from data analyses.

2. COURSE GRADING

TENTATIVE GRADING SCALE (PERCENTAGE %):

Letter Grade	A	A-	B+	B	B-	C+	C	C-	D+	D	D-	F
% Score	≥ 94	≥ 90	≥ 87	≥ 84	≥ 80	≥ 77	≥ 74	≥ 70	≥ 67	≥ 64	≥ 60	<60

Activity	Detail (*hours are on average*)	Estimated hours	% of grade
Readings	3.3 hours each x 14 weeks	46	N/A
Online/In-Lecture work, Attendance, Participation	10 hour each x 3	30	30%
Discussion	3 hour each x 4	12	N/A
Homework Assignments and Programming Mini-Projects	3 hours each x 11 weeks	33	30%
Quizzes	Prepare, complete, & submit	29	10%
Final projects	Prepare, complete, & submit	30	30%
	TOTAL	**180 hours**	**100%**

Note: At the instructor's sole discretion, the grades may be curved, the above decision boundaries may be adjusted in your favor, or the above tentative weights may be revised, but you are guaranteed at least the letter grade shown here if you obtain the corresponding score.

3. IMPORTANT COURSE POLICIES

This course encompasses a comprehensive structure comprising lectures, homework assignments, quizzes, examinations, and Final Projects.

1. **Lecture Arrangements:** The course lectures will be conducted fully in person adhering to the university's predetermined schedule. However, should any modifications arise in the class modularity, it will be the instructor's responsibility to promptly communicate these changes through Canvas, our primary online learning platform. Possible adjustments to class timings, content, or structure may occur due to various factors. Remaining well-informed is crucial to ensure a smooth learning journey.
2. **Effective Communication:** To ensure seamless communication, please inform the instructor promptly if you encounter any difficulties accessing Canvas or if you have not received the welcome email I sent before the first day of class. Kindly address this within the first week of the course.
3. **Canvas Notifications:** All notifications regarding alterations to the class modularity will be transmitted through your registered email address on Canvas. Consistently monitoring your email and staying updated with Canvas notifications will enable you to stay abreast of any changes or updates related to the course. Your attentiveness to these notifications is genuinely appreciated and will significantly contribute to your success in the course.
4. **Lectures and hands-on:** Lectures, conducted in person, will take place once a week. A corresponding hands-on session is associated with each lecture. The hands-on sessions will complement the lecture content and facilitate the practical application of the materials covered. Attendance and active participation in both lectures and hands-on are obligatory. Following the lectures, hands-on will be conducted to reinforce the learning process. The concluding 20 minutes

of each lecture will be dedicated to a review of the lecture content. These hands-on sessions will offer immediate opportunities to apply the concepts discussed and serve as a preliminary introduction to subsequent programming assignments.

5. **Written Assignments:** Every week concludes with a problem set or a programming assignment. These assignments are to be individually undertaken by the students.
6. **Quizzes**: Quizzes are scheduled during lecture times and are intended for individual completion by students and an announcement will be sent in advance via Canvas platform.
7. **Examinations**: Examinations will be administered in class. The assessment structure comprises two examinations. It is advised not to make any conflicting arrangements on the designated exam date.
8. **Projects**: As a vital component of this course, a final project serves as a significant assessment. The project can be a collaborative effort, accommodating teams consisting of a maximum of two individuals per group. You have the liberty to choose your project partner. However, both the submission of the final project and the presentation should occur only after reaching a mutual agreement on the collaborative work between the students. These guidelines and procedures provide the framework for an effective and structured learning experience throughout the course.
9. **Attendance and Participation:** Attendance and active participation in every lecture are mandatory to foster an effective learning environment. Students are expected to be punctual and engaged. Missing over three classes without a valid reason will result in official reporting to you and your advisor. The first three absences will be excused, subsequent absences will not. Each absence after the third incurs a five-point deduction from your course grade. If an absence is necessary, notify the instructor in advance if feasible.
10. **Academic Responsibility:** Students are accountable for all class materials, exams, and announcements. No excuses will be considered for incomplete assignments. Responsibility for obtaining missed class notes rests with the student, as the instructor does not provide slides or notes.
11. **Assignment Submissions:** All assignments must be submitted on time via the online Canvas system. Late submissions are not accepted unless they are due to family or medical emergencies. Files can be resubmitted multiple times before the deadline. Any submission past the deadline is marked as late. Email responses are generally within 24 hours on weekdays.
12. **ChatGPT**: The use of AI composition tools like ChatGPT is prohibited unless explicitly approved by me. Before using any AI-based writing software for assignments in this course, you must first get my consent. Failure to do so jeopardizes your academic integrity. Assignment instructions, due dates, and submission links will be on Canvas. Dishonesty or cheating is prohibited. While open discussions are encouraged, final work must be the student's own. Late submissions incur penalties.
13. **Study Groups and Extra Credit:** Study groups are welcomed, but individual answers should be unique. Extra credit can be added to assignments, exams, or lab work to uplift class averages. In case of disputes, extra credit may be revoked.
14. **Exam Policies:** Make-up exams are permitted in exceptional situations with valid documentation. The instructor may adjust the exam format based on student needs. Any changes will be communicated via email. Students are responsible for contacting the instructor to arrange rescheduled exams.

4. Student Responsibilities or Tips for Success in the Course

This course presents a rewarding yet programming-intensive challenge that accelerates in complexity throughout the semester. It centers on cultivating not just coding skills, but the art of crafting 'efficient' and 'elegant' code. Staying on track is crucial.

- Reach out for assistance. I'm dedicated to your success and value your inquiries. Don't hesitate to ask questions during class to grasp concepts in real-time. Your queries benefit both you and your peers.
- Patience over frustration. Programming demands patience. Investing hours in assignments/projects is common. Take brief breaks. If you're stuck on a bug, email me—I'm here to assist. Remember, practice enhances proficiency.
- Continual code testing. Regularly test and compile your code after a few lines. Errors are easier to spot within a smaller scope. Safeguard your work with frequent backups—cloud storage is recommended.
- Due to the demanding nature of the course and the time commitment it entails, consult your instructor, and seek peer assistance. However, remember that plagiarism is inexcusable.

5. COURSE TOPICS AND TENTATIVE SCHEDULE

Week	Topic	Assignment
1	**Lecture 01:** **Class Introduction and Syllabus Review** **Introduction to Data** Introduction to the course structure, expectations, and tools Overview of data science, its importance, and applications	**Read: Syllabus** **In Class work #SQLite Installation**
2	**Lecture 02:** **Basics of SQL and Relational Databases** SQL fundamentals: SELECT, INSERT, UPDATE, DELETE Understanding relational databases: tables, relationships, and keys	**Read:** **In Class work #02** **Homework 1 assigned**
3	**Lecture 03:** **Week 3: Data Importing Techniques** Importing data from various sources (CSV files, databases, APIs) Python libraries for data import (Pandas, SQLalchemy)	**Read:** **In Class work #03** **Homework 3 assigned**
4	**Lecture 04:** **Week 4: Data Cleaning and Preparation** Identifying and handling missing values Data type conversions, normalization, and error handling	**Read:** **In Class work #04** **Homework 4 assigned**
5	**Lecture 05:** **Week 5: Data Transformation** Transforming datasets using Pandas: merging, splitting, pivoting Techniques for data manipulation and preparation for analysis	**Read:** **In Class work #05** **Homework 5 assigned**
6	**Lecture 06:** **Week 6: Data Visualization** Introduction to data visualization tools (Matplotlib, Seaborn) Creating plots, histograms, and interactive visualizations	**Read:** **In Class work #06** **Homework 6 assigned**

7	**Lecture 07:** **Week 7: Mid-Course Project** Application of learned skills in a mini project Data import, cleaning, transformation, and preliminary analysis	**Read:** **In Class work #07** **Homework 7 assigned**
8	**Lecture 08:** **Week 8: Advanced SQL and Big Data** Advanced SQL queries and optimizations Integrating Python with SQL for dynamic data manipulation	**Read:** **In Class work #08** **Homework 8 assigned**
9	**Lecture 09:** **Week 9: Building Data Processing Pipelines** Automating data workflows using Python Tools and libraries for pipeline creation (Luigi, Airflow)	**Read:** **In Class work #09** **Homework 9 assigned**
10	**Lecture 10:** **Week 10: Repeatable and Reproducible Analysis** Techniques for ensuring analysis reproducibility (version control, Docker) Documenting and sharing analysis workflows.	**Read:** **In Class work # 10** **Homework 10 assigned**
11	**Lecture 11:** **Week 11: Data Modeling** Basics of statistical modeling and hypothesis testing	**Read:** **In Class work # 11**
12	**Lecture12:** **Week 12: Text Mining and Natural Language Processing** Techniques for text extraction, tokenization, and sentiment analysis Using libraries like NLTK and spaCy for NLP tasks	
13	**Lecture13:** Week 13: **Final Project Workshop** presentations of project	

IMPORTANT UNIVERSITY POLICIES

ACADEMIC INTEGRITY, PLAGIARISM, DISHONESTY, AND CHEATING POLICY: A commitment to the principles of academic integrity is essential to the mission of Northeastern University. The promotion of independent and original scholarship ensures that students derive the most from their educational experience and their pursuit of knowledge. Academic dishonesty violates the most fundamental values of an intellectual community and undermines the achievements of the entire University.

As members of the academic community, students must become familiar with their rights and responsibilities. In each course, they are responsible for knowing the requirements and restrictions regarding research and writing, examinations of whatever kind, collaborative work, the use of study aids, the appropriateness of assistance, and other issues. Students are responsible for learning the conventions of documentation and acknowledgment of sources in their fields. Northeastern University expects students to complete all examinations, tests, papers, creative projects, and assignments of any kind according to the highest ethical standards, as set forth either explicitly or implicitly in this Code or by the direction of instructors. The University is committed to investigating any allegation of violations of academic integrity against a student. Violations include, but are not limited to, plagiarism, cheating, fabrication, unauthorized

collaboration, and academic misconduct. Sanctions for violations of academic integrity are administered through the Office of Student Conduct and Conflict Resolution (OSCCR) in conjunction with other University offices as deemed appropriate. It is generally the responsibility of the faculty member overseeing the academic activity to report the violation to OSCCR and to determine the appropriate sanction. A student who believes they have been wrongly sanctioned has a right to an appeals process. Go to http://www.northeastern.edu/osccr/academic-integrity-policy/ to access the full academic integrity policy.

STUDENT ACCOMMODATIONS: Northeastern University and the Disability Resource Center (DRC) are committed to providing disability services that enable students who qualify under Section 504 of the Rehabilitation Act and the Americans with Disabilities Act Amendments Act (ADAAA) to participate fully in the activities of the university. To receive accommodations through the DRC, students must provide appropriate documentation that demonstrates a current substantially limiting disability. For more information, visit http://www.northeastern.edu/drc/getting-started-with-the-drc/.

DIVERSITY AND INCLUSION: Northeastern University is committed to equal opportunity, affirmative action, diversity, and social justice while building a climate of inclusion on and beyond campus. In the classroom, members of the University community work to cultivate an inclusive environment that denounces discrimination through innovation, collaboration and an awareness of global perspectives on social justice. It is my intention that students from all backgrounds and perspectives will be well served by this course, and that the diversity that students bring to this class will be viewed as an asset. I welcome individuals of all ages, backgrounds, beliefs, ethnicities, genders, gender identities, gender expressions, national origins, religious affiliations, sexual orientations, socioeconomic background, family education level, ability, and other visible and nonvisible differences. All members of this class are expected to contribute to a respectful, welcoming, and inclusive environment for every other member of the class. Your suggestions are encouraged and appreciated. Please visit http://www.northeastern.edu/oidi/ for complete information on Diversity and Inclusion.

TITLE IX: Title IX of the Education Amendments of 1972 protects individuals from sex or gender-based discrimination, including discrimination based on gender-identity, in educational programs and activities that receive federal financial assistance. Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, faculty, and staff. In case of an emergency, please call YJJ. Please visit www.northeastern.edu/titleix for a complete list of reporting options and resources both on- and off-campus.

INSTRUCTIONAL CONTINUITY AND CLASS CANCELLATION: Instructional continuity refers to the continuation of instruction during unforeseen campus closure or instructor absence. Should the need to cancel a class session occur, students will be notified through Canvas and/or their Northeastern email address. Students are responsible for all course material provided through this instructional continuity plan. The following procedures will be in place to ensure continuity of instruction in this course:

- The lecture scheduled to be delivered in a cancelled class session will be recorded either synchronously or asynchronously and posted on Canvas for the students to watch within Db hours of the cancelled class session. In lieu of the recorded lecture, students might be asked to watch one or multiple videos from reputable online resources. A reading or an online problem-solving practice session with the instructor or teaching assistant for the course might also replace the lecture of a cancelled class session. All instructional materials (e.g., readings, lecture notes, slides, handouts, homework assignments, lab manuals, source codes, etc.) will be distributed to students either online through Canvas or in the subsequent class session.

- If a homework assignment was due on a day on which a class session is cancelled, students are asked to upload a legible, high-quality scanned copy of their own homework solutions in PDF format on Canvas by midnight on the same day.
- If a quiz or exam was scheduled to be held on a day on which a class session is cancelled, the quiz or exam will be either postponed to the next regularly scheduled class session or held online through Canvas and Respondus Lockdown Browser and Monitor on a later day, giving students enough time to prepare the necessary software and hardware to be able to take the quiz or exam on the aforementioned online platforms.

RECORDING OF CLASS SESSIONS: Video/audio of class sessions (face-to-face or online synchronous or asynchronous) may be recorded for the benefit of students in the class. Recordings will be shared via platforms with access limited to other members of the class. The instructor will attain consent from students if recordings of student comments or images are shared with a broader audience.

COURSE COPYRIGHT: All course materials that students receive or to which students have online access are protected by copyright laws. Students may use course materials and make copies for their own use as needed, but unauthorized distribution and/or uploading of materials without the instructor's express permission is strictly prohibited. Students who engage in the unauthorized distribution of copyrighted materials may be held in violation of the Student Code of Conduct, and/or be liable under Federal and State laws. In addition, distributing completed essays, labs reports, homework, quizzes, exams, reports, or other assignments created for this course constitutes a violation of the Student Code of Conduct policy.

USE OF CANVAS: In this course we will be using Canvas to distribute and collect lecture notes, handouts, assignments, labs, quizzes, exams, activities, and/or discussions. Students should have regular access to Canvas and their Northeastern email. Downloading the Canvas mobile app will also allow you to view content and participate in courses on an iOS or Android mobile device.

IMPORTANT ANNOUNCEMENTS: Important messages will be communicated through Canvas and/or emailed to students' Northeastern email addresses.

COURSE EVALUATION: Your feedback about the course and instructor is the only way instructors and academic units can improve the quality of a course and its content. Courses administered by the Department of Electrical and Computer Engineering are evaluated electronically via the Online Course Evaluation System. Students will receive all necessary information via email at the end of the semester. The evaluations are entirely confidential and will preserve your anonymity.

LEARNING ENVIRONMENT: Universities provide a safe haven for multiple perspectives and for disagreement and dissent. However, all of our conversations should be pursued in the spirit of mutual respect and civility. Together, we will work to create an environment in which every voice and perspective is heard and respected. The use of harmful or exclusionary language, including language that is racist, sexist, homophobic, or transphobic, would erode what we are trying to accomplish in our course and is not acceptable in the University classrooms.

UNIVERSAL DESIGN FOR LEARNING: I am committed to the principle of universal learning. This means that our classroom, virtual spaces, practices, and interactions have been designed to be as inclusive as possible. If you have a particular need, please email me or arrange a meeting with me so I can help you learn in this course. I will treat any information that you share as private and confidential. Contact the Disability Resource Center to seek official accommodations due to a disability, pregnancy, or emergency medical condition.

RELIGIOUS OBSERVANCES STATEMENT: It is my personal goal to respect the faith and religious obligations of each student. Students with exams and classes that conflict with their religious observances should notify their instructor at the beginning of the semester in order to work out a mutually agreeable alternative. Please note that, regardless of whether an absence is "excused" or "unexcused," the student is responsible for all missed course content and activities.

SYLLABUS SUBJECT-TO-CHANGE STATEMENT: This syllabus is a guide for the course. Students will be notified of any changes made by the instructor at his sole discretion with the advice of the students.