SYLLABUS

## Data Preparation and Analysis

# Instructor

Ming-Long Lam, Adjunct Faculty, Illinois Institute of Technology

[Ming-Long Lam | Illinois Institute of Technology (iit.edu)](https://www.iit.edu/directory/people/ming-long-lam)

# Course Description

This course introduces the necessary concepts and common techniques for analyzing data. The primary emphasis is on the process of data analysis, including data preparation, descriptive analytics, model training, and result interpretation. The process starts with removing distractions and anomalies, followed by discovering insights, formulating propositions, validating evidence, and finally building professional-grade solutions. Following the process properly, regularly, and transparently brings credibility and increases the impact of the results.

This course will cover topics including Exploratory Data Analysis, Feature Screening, Segmentation, Association Rules, Nearest Neighbors, Clustering, Decision Tree, Linear Regression, Logistic Regression, and Performance Evaluation. Besides, this course will review statistical theory, matrix algebra, and computational techniques as necessary.

This course prepares students ready for and capable of the data preparation and analysis process. Besides developing Python codes for carrying out the process, students will learn to tune the software tools for the most efficient implementation and optimal performance. At the end of this course, students will have built their inventory of data analysis codes and their confidence in advocating their propositions to the business stakeholders.

# Course Outcomes

Upon successful completion of this course, you will be able to:

* Aware of the process of data analysis and proficient in managing the process.
* Able to determine the appropriate techniques for generating various kinds of insights from data.
* Expand your horizon of the modeling algorithms and be able to deploy the appropriate algorithms for the problems.
* Elevate your skills in formulating propositions, training algorithms, and evaluating models.
* Present and advocate your actionable solutions with confidence to the business stakeholders.
* Develop your skills in developing software programs in the Python programming language.

# Course Materials

The link to reading materials and resources to learn on the topics can be found in each week’s learning module. All materials are available online for free, no required resources need to be purchased. There is no required textbook to supplement the course materials. Note: Be aware that some resources may open in a new tab.

**Optional Materials:** [A Practitioner's Guide to Machine Learning](https://he.kendallhunt.com/product/practitioners-guide-machine-learning) (abbreviated PGML for Reading)

**Software Requirements:** Python version 3.11 or above with the latest compatible versions of NumPy, SciPy, Pandas, Scikit-learn, and Statsmodels libraries.

# Course Outline

The course consists of 8 modules that focus on the following key areas:

## Module 1: Process of Preparing and Analyzing Data

**Key concepts**

* Why Do We Analyze Data?
* The Process of Data Analysis
* The First Step of Knowing Your Data

**Readings**

* PGML Chapter 1
* [What is Big Data?](https://www.oracle.com/big-data/what-is-big-data/)
* [IEEE\_754](https://en.wikipedia.org/wiki/IEEE_754)
* [CRISP-DM 2](https://www.datascience-pm.com/crisp-dm-2/)
* [A Method for Selecting the Bin Size of a Time Histogram](https://www.neuralengine.org/res/pdf/shimazaki_neco07.pdf)

## Module 2: Discover and Measure Associations

**Key concepts**

* Discover Associations
* Measure Associations (Part 1)
* Measure Associations (Part 2)

**Readings**

* [Use of Correlation Coefficient in Medical Research](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/)
* [Visit Chicago Taxi Trip Data](https://data.cityofchicago.org/Transportation/Taxi-Trips-2022/npd7-ywjz)
* [Who is Karl Pearson?](https://mathshistory.st-andrews.ac.uk/Biographies/Pearson/)
* [What are Statistical Tests?](https://www.itl.nist.gov/div898/handbook/prc/section1/prc13.htm)
* [Correlation with Python](https://realpython.com/numpy-scipy-pandas-correlation-python)
* [Who is Carl Harald Cramér?](https://mathshistory.st-andrews.ac.uk/Biographies/Cramer_Harald/)
* [What is Eta-squared?](https://www.statology.org/eta-squared/)

## Module 3: Market Basket Analysis

**Key concepts**

* What is in Your Basket?
* How are Association Rules Discovered?
* What Can Association Rules Tell Me?

**Readings**

* PGML Chapter 3
* [Cross-Selling](https://www.youtube.com/watch?v=g9cE9PHIcv8&t=36s)
* [Apriori Algorithm](https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/)
* [Association Rules](https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/)

## Module 4: Partitioning, Segmenting, and Clustering of Observations

**Key concepts**

* Partition Observations for Training Models
* Create Segments of Observations for Business Reasons
* Put Observations with Similar Feature Values in Clusters

**Readings**

* PGML Chapter 4
* [Linear Congruential Random Number Generator](https://www.youtube.com/watch?v=LUusa5Mhx_g)
* [How Stratified Random Sampling Works](https://www.investopedia.com/terms/stratified_random_sampling.asp)
* [What](https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp) are Recency, Frequency, and Monetary Values (RFM) in Marketing?
* [RFM analysis for Customer Segmentation](https://clevertap.com/blog/rfm-analysis)
* [What is Clustering According to Google?](https://developers.google.com/machine-learning/clustering/overview)

## Module 5: Linear Regression

**Key concepts**

* Linear Regression Model
* Feature Selection (a.k.a. Model Selection)
* Feature Importance

**Readings**

* [Linear Regression Analysis](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018/)
* [Least Squares Regression](https://www.mathsisfun.com/data/least-squares-regression.html)
* [Understand Forward and Backward Stepwise Regression](https://quantifyinghealth.com/stepwise-selection/)
* [How do Shapley's Values work?](https://www.aidancooper.co.uk/how-shapley-values-work/)

## Module 6: Binary Logistics Regression

**Key concepts**

* Logistic Regression
* Forward Selection
* Interpret Model and Assess Performance

**Readings**

* PGML Chapter 6
* [Maximum Likelihood, Clearly Explained!](https://www.youtube.com/watch?v=XepXtl9YKwc&t=20s)
* [Class Separation cannot be overlooked in Logistic Regression](https://medium.com/geekculture/class-separation-cannot-be-overlooked-in-logistic-regression-f20e58b203eb)
* [What is Forward Selection?](https://www.statisticshowto.com/forward-selection/)
* [What’s the Best R-squared for Logistic Regression?](https://statisticalhorizons.com/r2logistic/)

## Module 7: Decision Trees – The CART Algorithm

**Key concepts**

* Motivation of Decision Trees
* The CART Algorithm
* Cluster Profiling

**Readings**

* PGML Chapter 5
* [How are Variable Interactions Modeled in Decision Trees (CART)?](https://www.youtube.com/watch?v=NgwJWWlDdwk)
* [How to Write a Decision Tree (CART) as an Equation](https://www.youtube.com/watch?v=JdpZT9I0G5M&t=0s)
* [Decision Trees for Clustering](https://insidelearningmachines.com/decision_trees_for_clustering/)

## Module 8: Evaluating the Performance of Models

**Key concepts**

* Prediction Models
* Nominal Classification Models
* Charts for Classification Models

**Readings**

* PGML Chapter 7, 8
* [Identifying Outliers (Unusual y Values)](https://online.stat.psu.edu/stat501/lesson/11/11.3)
* [ROC curve and Area Under ROC Curve in Machine Learning](https://www.linkedin.com/pulse/roc-curve-area-under-machine-learning-infogen-labs-infogenlabsinc-1f)
* [Using Lift Analysis to Measure Campaign Metrics](https://heypoplar.com/articles/using-lift-analysis-to-measure-campaign-metrics)

# Course Structure and Learning Activities

There are 8 content modules in this course and each module may take about 9-12 hours to complete. The final module consists of your final exam for the course.

This course is comprised of the following elements:

* **Readings**: Each module may include several required and/or supplemental readings.
* **Video Lessons**: In each module, the concepts you need to know will be presented through a collection of short videos. You may stream these videos for playback within the browser by clicking on their titles.
* **In-Video Questions**. Some videos have questions associated with it to help verify your understanding of the topics. These questions will automatically appear while watching the video if you stream the video through your browser. These questions do not contribute toward your final score in the class.
* **Practice Quizzes**: Each module will include some practice quizzes, intended for you to assess your understanding of the topics. You will be allowed unlimited attempts at each practice quiz. There is no time limit on how long you take to complete each attempt at the quiz. These quizzes do not contribute toward your final score in the class.
* **Discussion Forum**: This course has a place for you to interact with other learners about class-related topics.
* **Assignments:** To provide students with opportunities to practice their data analysis skills, this course has two assignments with open-ended questions. Students’ answers will be manually read and graded. The first one helps you exercise what you have learned in Module 1 and 4 (Due at the end of Module 4). The second one is about topics in Module 6 and 7 (Due at the end of Module 7). You should expect to develop Python codes for finding the answers. However, you are not required to submit your Python codes. You will be able to submit one attempt for each of the open-ended assignments.
* **Summative Module Assessments:** Each module will include at least one summative module assessment. You will be allowed one attempt every eight hours for each assessment. There is no time limit on how long you take to complete each attempt at the assessment. Your highest grade will be recorded.
* **Final Assessment:** This course will contain one summative course assessment. You will be allowed one attempt for the assessment. Before taking the exam, please make sure you are in a place with reliable internet connection. No retakes will be granted for the lack of internet access. You are in an online program and the use of the Internet is a requirement.

# How to Pass This Course

Guidelines for completing and submitting each assigned course activity are posted along with the assignment. Assignments can be submitted at any time as you move through the module. Only those who complete and submit all assignments will receive a certificate of completion of this course. *No late assignments will be accepted.* In case of extenuating circumstances beyond your control that prevent an assignment or exam's submission, you must enter a request with the program advisor and the instructor.

To qualify for a Course Certificate, simply start verifying your coursework at the beginning of the course and pay the fee. Coursera [Financial Aid](https://learner.coursera.help/hc/en-us/articles/209819033-Apply-for-Financial-Aid) is available to offset the registration cost for learners with demonstrated economic needs. If you have questions about Course Certificates, [please see the help topics here](https://learner.coursera.help/hc/en-us/sections/201895943-Course-Certificates).

Also note that this course is part of the Master of Data Science program offered by Illinois Institute of Technology. By earning a Course Certificate in this course, you are on your way toward earning a Specialization Certificate in this topic.[See more information about the program here.](https://www.coursera.org/degrees/mas-data-science-illinois-tech)

If you choose not to pay the fee, you can still audit the course. You will still be able to view all videos, submit practice quizzes, and view required assessments. Auditing does not include the option to submit required assessments. As such, you will not be able to earn a grade or a Course Certificate.

The following table explains the breakdown for what is required to pass the class and qualify for a Course Certificate. You must pass all required activities to pass this course.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Activity** | **Required?** | **Number per Course** | **Estimated Time per Module** | **% Required to Pass** | **% of Total Grade** |
| Lecture Videos | Yes | 3-6 per module | .5-1 hour | N/A | N/A |
| Practice Quizzes | No | 3-6 per module | .5 hour | N/A | N/A |
| Discussions | No | 1 per course | 1 hour | N/A | N/A |
| Summative Module Assessments | Yes | 1 per module | .5 hour | 80% | 7.5%/each module (60%) |
| Summative Course Assessments | Yes | 1 -2 per course | 1-3 hours | 80% | 40% |

# Letter Grades

Letter grades are used for the final grade. Information about IIT grading system can be found in the [Graduate Student Handbook.](https://bulletin.iit.edu/grad-handbook/grading-system/)

|  |  |  |  |
| --- | --- | --- | --- |
| Letter Grade | Description | Points | Percentage |
| A | Excellent | 4.00 | 90-100 |
| B | Above Average | 3.00 | 80-89.99 |
| C | Average | 2.00 | 70-79.99 |
| E | Fail | 0.00 | Under 70% |

# Getting and Giving Help

* Use the [Learner Help Center](https://www.coursera.support/s/learner-help-center?language=en_US) to find information regarding specific technical problems. For example, technical problems would include error messages, difficulty submitting assignments, or problems with video playback. If you cannot find an answer in the documentation, you can also report your problem to the Coursera staff by clicking on the *Contact Us!* link available on each topic's page within the Learner Help Center.
* Use the flag icon under each item to report errors in lecture video content, assignment questions and answers, assignment grading, text and links on course pages, or the content of other course materials.
* Familiarize yourself with [Coursera’s policy on Accessibility](https://www.coursera.support/s/article/360050668591-Accessibility-Statement#-61?).

# Academic Integrity

Your attentiveness to academic integrity reflects the value you place on your own work and the work of others. In addition to [Coursera's Honor Code](https://www.coursera.support/s/article/209818863-Coursera-Honor-Code?language=en_US?#-5), we also have high expectations for conduct during course participation.

## Discussion Forums: Expectations

Sharing an online course with other avid learners like you gives you a unique opportunity to share, collaborate, and learn from others and their experiences, and helps you reinforce your understanding of the topics of the course. Interacting in the Discussion Forums is a great way to engage with your online community. We know that it is not possible to read every discussion forum post, so we recommend that you read those that interest you; and reply when you can contribute. The forum is part of your class activities and everybody is expected to act professionally and be civil and respectful of others in your class. Failure to meet these expectations may be considered a break in the Academic Code of Conduct and may result in your removal from the course. Please, check tips and helpful tools to [interact in discussion forums in this document](https://www.coursera.support/s/article/208279996-Get-help-with-course-content-in-the-discussion-forums?language=en_US).

## Academic Code of Conduct

Above all else, learners are expected to ensure that their conduct helps to create an atmosphere conducive to learning and the interchange of knowledge. While it is understood that some of these items are subject to interpretation, learners should nonetheless endeavor to:

* Be respectful of fellow learners.
* Not discriminate against fellow learners in any manner.
* Conduct peer reviews in a timely manner and give useful feedback on what was done well, helpful suggestions for how to improve, and specific comments about why you gave the grade you chose to assist peers in their learning.
* Turn assignments in on time and follow instructions on all assignments including those affecting the use of technology.
* Be truthful in all communication, which includes, but is not limited to, avoiding academic dishonesty.

# Illinois Institute of Technology Copyright Statement

This course material is copyrighted, and all rights are reserved by IIT. No part of this course material may be reproduced, transmitted, transcribed, stored in a retrieval system, or translated into any language or computer language, in any form or by any means, electronic, mechanical, magnetic, optical, chemical, manual, or otherwise, without the express prior written permission of the University.