Lahore University of Management Sciences

CS 334/EE 402 – Principles and Techniques of Data Science

Spring 2024

Subject to Change

Instructor	Dr. Mobin Javed
Room No.	SBASSE 9-G10A
Email	mobin.javed@lums.edu.pk
Telephone	3338
TAs	TBA
TA Office Hours	TBA
Course URL (if any)	lms.lums.edu.pk

Course Basics				
Credit Hours	3			
Lecture(s)	Nbr of Lec(s) Per Week	2	Duration	75 mins
Recitation/Lab (per week)	Nbr of Lec(s) Per Week	--	Duration	
Tutorial (per week)	Nbr of Lec(s) Per Week	--	Duration	

Course Distribution	
Core	No
Elective	Yes
Open for Student Category	All
Close for Student Category	None

COURSE DESCRIPTION

This is an introductory-level Python-based course in data science to prepare students for scientific work, as well as advanced courses in data mining and machine learning. It is a hands-on course and involves data analysis work.

The first half will focus on the fundamentals. We will start with descriptive statistics, develop an understanding of the biases in working with data, and learn and practice exploratory data analysis. The second half will focus on drawing inferences from data -- we will talk about setting up controlled experiments, hypothesis testing, and the foundational concepts of statistical and machine learning. We will also spend a major part of the course learning about data engineering, i.e., tools and techniques for data collection, data storage and querying, and working with big data.

COURSE PREREQUISITE(S)

- CS100: Computational Problem Solving
- MATH120: Linear Algebra with Differential Equations
  (Exceptions possible with prior instructor approval)
- Basic knowledge of Probability and Statistics is assumed
- Basic knowledge of Python programming language is assumed

COURSE OBJECTIVES

The goal of this course is to train students to become good data scientists. The students will build a foundation in drawing inferences from data, which serves as a preparation for advanced courses in data mining and machine learning.

Learning Outcomes

- Learn how to conduct sound data analysis
- Learn how to describe a given dataset and assess its quality
- Understand issues in experiments involving active data collection and develop the discipline of maintaining meta-data
- Learn how to build data pipelines (collection, cleaning, EDA, modeling, evaluation, results) for “repeatable” work
- Become well-versed with tools and technologies for data analysis (e.g., Pandas, Spark, scitkit-learn, R)
- Learn the theory behind drawing inferences from data
- Learn how to communicate results effectively

Grading Breakup and Policy

HWs:	30%
Quizzes:	25%
Project:	15%
Final:	25%
Labs:	0%
Class Participation:	5%

HWs:
We will have four homeworks during the course of the semester, one corresponding to each of the first four modules.
Optional Labs:
Labs will be released to help you practice the concepts in the lecture. These labs are optional and will *not* count towards your grade.
Quizzes:
We will have a total of eight announced quizzes. Out of these best six will count towards your final grade (N-2 policy).
The quizzes will be held on LMS during lecture timings. The exact date of each quiz will be announced in lecture and on Piazza.
Projects:
Projects must be done in teams of 3-4.
The spirit behind the project is to give you some experience conducting a data science project end-to-end. This includes thinking about what question(s) to answer, gathering the right dataset, data quality assessment, EDA, drawing inferences from data, and building models. In addition, we want you to get some practice in storytelling and communicating your data science work clearly and crisply. Project is open ended by design, you are expected to come up with ideas on what to do and how to do. We have designed four check-points during the semester to help guide your exploration. More information on these will be released during the course of the semester.

Examination Detail

Midterm Exam	Yes/No: No Combine Separate: Duration: Preferred Date: Exam Specifications:

Final Exam	Yes/No: Yes Combine Separate: Duration: Exam Specifications:

Harassment Policy

SBASSE, LUMS and particularly this class, is a harassment free zone. There is absolutely zero tolerance for any behavior that is intended or has the expected result of making anyone uncomfortable and negatively impacts the class environment, or any individual's ability to work to the best of their potential. In case a differently-abled student requires accommodations for fully participating in the course, students are advised to contact the instructor so that they can be facilitated accordingly.

If you think that you may be a victim of harassment, or if you have observed any harassment occurring in the purview of this class, please reach out and speak to me. If you are a victim, I strongly encourage you to reach out to the Office of Accessibility and Inclusion at oai@lums.edu.pk or the sexual harassment inquiry committee at harassment@lums.edu.pk for any queries, clarifications, or advice. You may choose to file an informal or a formal complaint to put an end of offending behavior. You can find more details regarding the LUMS sexual harassment policy here. To file a complaint, please write to harassment@lums.edu.pk.

In addition to LUMS resources, SSE's Council on Belonging and Equity is committed to devising ways to provide a safe, inclusive and respectful learning environment for students, faculty and staff. I serve as a member of CBE. To seek counsel related to any issues, please feel free to approach any member of the council or email at cbe.sse@lums.edu.pk

Lec #	Topics	Assessments
1	Overview of Data Science: Untangling the Data Science Process - What is Data Science? Why Data Science? Why Now? - Data Science Lifecycle - Overview of course modules	**Lab-0 Release (Setup)**
Module 1: Descriptive Statistics, Data Acquisition, and Tools		**Quiz Dates TBA**
2	Descriptive Statistics, Deceptive Descriptions, Important Distributions, Choosing Unit of Analysis, Data Acquisition, Sampling, and Sources of Bias	
3	Data Manipulation Using Pandas – I (Lecture + Lab) - Jupyter Notebooks - Introduction to Pandas	**Lab-1 Release (Optional)**
4	Data Manipulation Using Pandas – II (Lecture + Lab) - Data Aggregation - Case Study	**HW-1 Release**
Module 2: Data Cleaning, Exploratory Data Analysis, and Visualization		**Quiz Dates TBA**
5	Data Cleaning and Exploratory Data Analysis (EDA) – I - Transforming Data for Ease of Analysis - Common Data Anomalies	
6	EDA – II and Data Visualization and Transformations – I - Structure, Granularity, Scope, Temporality, Faithfulness of Data - Visual Representations of Data: A Way of Amplifying Cognition	
7	Data Visualization and Transformations – II - Principles of Sound Visualizations - Smoothing and Transformations	
8	Analyzing Text Data - Text Mining and String Manipulation - Regular Expressions	
9	Databases and SQL - Relational Databases - SQL Queries	**HW-2 Release**
Module 3: Experiments, Causality, and Foundations of Statistical Inference		**Quiz Dates TBA**
10	Experiments, Observational Studies, and Causal Inference – I - Association, Causality, and Data - Randomized Control Trials and Case Study	
11	Causal Inference – II - Estimating the Counterfactual - Rubin's Causal Model of Potential Outcomes	
12	Causal Inference – III - Causal Inference from Observational Studies (RDD, IV etc) - Causal DAGs	

13	Statistical Inference and Hypothesis Testing • Sampling, Assessing Models, and Comparing Distributions • Hypothesis Testing and P-values • Case Study	
14	A/B Testing and Permutation Tests • Comparing Samples and Case Study	**HW-3 Release**
15	Bootstrap Sampling, Central Limit Theorem, Confidence Intervals • Repeated Random Sampling • Samples Averages • Variability and Bounds	
Module 4: Machine Learning		**Quiz Dates TBA**
16	Models & Estimation • What is a Model? • Modeling Process and Loss Functions	
17	Optimization and Gradient Descent • Convexity and Gradient Descent • Stochastic Gradient Descent	
18	Regression and Linear Models • Simple Linear Regression • Ordinary Least Squares	
19	**Pakistan Day Holiday (Mar 23)** Class Rescheduled to Apr 25: Industry Guest Lecture	
20	Feature Engineering • What is Feature Engineering? • Pitfalls	
21	Fundamental Challenges in Learning: Bias-Variance Tradeoff • Risk and Cost Minimization • Models Bias and Variance • Bias-Variance Tradeoff	**HW-4 Release**
22	Regularization and Cross-Validation • Train-Test Split	
23	Classification and Logistic Regression • Logistic Function • Stochastic Gradient Descent	
Module 5: Big Data and Ethics		**Quiz Dates TBA**
24	Big Data Processing - I • Storage: Distributed File Systems • MapReduce and Spark	**Lab-2 Release (Optional)**
25	Big Data Processing – II & Ethics in Data Science	

26	Guest Lecture: Ethics in Data Science	

Textbook(s)/Supplementary Readings

There is no specific textbook for this course. Following are recommended readings:

(i) Principles and Techniques of Data Science
The textbook for DS100 at UC Berkeley

(ii) Doing Data Science

(iii) Naked Statistics: Stripping the Dread from the Data
A short good read on the fundamentals of Probability and Stats