MA 5755: Data Analysis & Visualization in R/Python/SQL

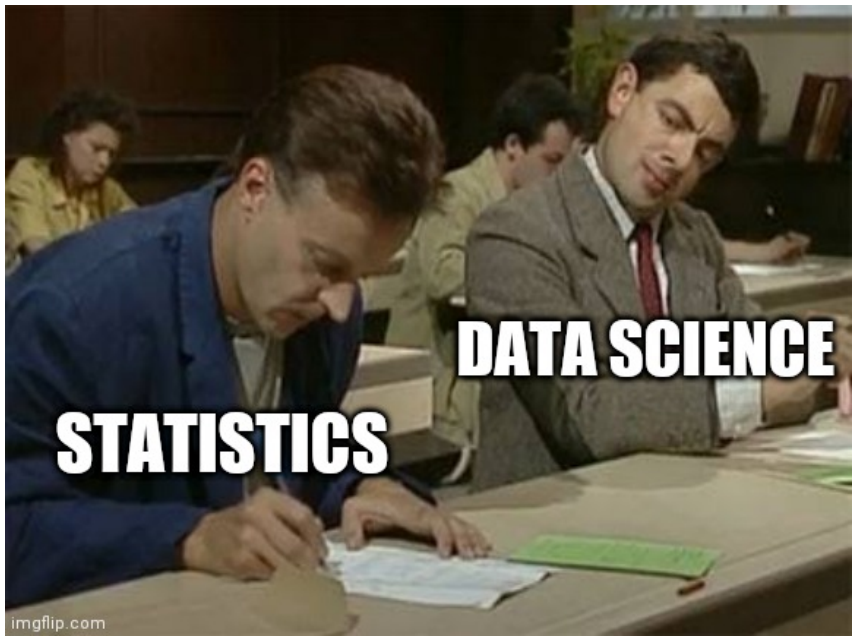
Week 1: Statistical Learning Perspective and Exploratory Data Analysis

Dr. Rakhi Singh
Department of Mathematics
IIT Madras

Spring 2026

What Is This Course About?

- Learning from data using computational and statistical methods
- Understanding structure, not just building models
- Using visualization as an analytical tool
- Emphasis on hands-on labs and interpretation



Source for the meme: web but unclear

Learning Outcomes

By the end of the course, you should expect [‡] to

- Independently perform exploratory data analysis on complex, real-world datasets using principled visualization techniques.
- Select, implement, and evaluate appropriate statistical and machine learning models for supervised and unsupervised learning tasks.
- Apply modern statistical learning methods, including ensemble models and neural networks, while accounting for generalization and overfitting.
- Interpret model outputs critically and communicate uncertainty, limitations, and insights effectively.
- Present analytical findings through clear, static, and animated visualizations suitable for technical and non-technical audiences.

[‡] granted that you work hard.

Prerequisites

There is a heavy dependency on laptops. If you don't have one or hesitate to bring it here, contact me early (that is, TODAY).

This is NOT an "Intro to Python" course, so I will not be teaching Python. If learning Python is your motivation, you could take assignments from this course (after the course is over) and do them yourself using a large language model. In this course, all grading will be "insight-centric", only 10-20% of marks will be attributed to "being able to code".

This is a statistical learning course, so I expect you to have completed at least one Statistics course prior to this one. Otherwise, you may lack the relevant background and struggle.

If you are ready to learn and be an active participant in the course, you are most welcome. I would be very honored to have contributed a little to your data science journey through statistical learning.

Week-wise distribution

Week	Main Topic
1	Statistical Learning Perspective and Exploratory Data Analysis
2	Supervised learning and difference with unsupervised learning
3	Distance Geometry and K-means Clustering
4	Probabilistic Clustering and Gaussian Mixture Models
5	Hierarchical Clustering and Cluster Evaluation
6	Visualization of High-Dimensional Data
7	Regression as Prediction
8	Regularized Regression and Model Selection
9	Nonparametric Regression with Gaussian Processes
10	Distance- and Probabilistic-Based Classification
11	Tree-Based Methods and Support Vector Machines
12	Artificial Neural Networks
13	Deep Neural Networks and Convolutional Neural Networks
14	SQL and advanced visualization
15	Course Presentations

Primary reference: Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer. A free e-copy available at `https://hastie.su.domains/ElemStatLearn/`

Grading

Your grade will be based on your performance on

- about five small quizzes (10 percent) will be conducted,
- about five in-class homeworks (20 percent), you cannot miss these; if they happen on the day you miss the class, you will not be allowed to submit. (Dates can be worked out one week in advance)
- about six take-home homeworks (30 percent) for which about one week will be provided so that you can work at your own pace.
- final project (40 percent), more details in the first week of February, so that you can get started early. Group size: 2 or 3.

You are allowed to miss one quiz, one in-class homework, and one take-home homework. The best of $n-1$ exams in each of three categories (quiz, in-class homework, and take-home work) will account for the final score in that category.
All of this is subject to Class Committee approval.

Structure of the class

This is how the course will generally be structured.

- Weekly 3-hour laboratory session
- Each session consists of:
  - $\sim$ 50 minutes of deep-dive into the topic;
  - $\sim$ 25 minutes of guided lab;
  - $\sim$ 15 minutes break;
  - $\sim$ 75 minutes of guided lab.
- Tools include Jupyter, NumPy, pandas, scikit-learn, PyTorch/Keras, seaborn, ggplot2, and related libraries.
- We will use a combination of GitHub Classroom and Google Colab/your local files uploaded to the classroom. You will receive detailed instructions on how to do this by atmost Tuesday night.

Use of Generative AI/LLM (GAI)

You may find it very tempting to just ask an LLM to solve the problem on your behalf. However, we need to ensure that GAI is not used as a substitute or replacement for student learning.

You may use them

- as a reference tool, similar to looking up the documentation for a function or Googling your problem.
- to assist in writing code in this class. You are expected to understand how any/all submitted code works.

You may not make use of the technology as a substitute for critical thinking. For example, you may not upload your data file to a GAI platform and ask it to create charts and statistical models for you.

You may not use GAI to write narratives for your assignments.

More details at: `https://info2951.infosci.cornell.edu/course-syllabus.html`.

Statistician or data scientist perspective?

Fun (but true): `https://www.youtube.com/watch?v=uHGlCi9jOWY`

What sort of problems can we solve after this course?

All types where you (can) have structured data or can convert your data into a structured form.

One major consideration in today's world is that machine learning models are easy to apply. Who can't apply?

But if you don't know why these methods work, then it may be easy to misuse them, especially because modern datasets are large, complex, and messy.

Visualization and exploration are essential first steps.

Statistical Learning is broadly a framework for learning patterns from data, focusing on prediction, model structure, and generalization.

What is the difference between Prediction and Inference?

Exploratory Data Analysis (EDA)

EDA is really just an application of basic statistical concepts to each column in your data.

- Mean, median, mode, standard deviation
- Scatter plots
- Histograms to check the distributions
- Boxplots to identify potential outliers

We will apply these to one dataset today, but the class is also about how to generalize these to studying multivariable patterns.

EDA can

- reveal structure, patterns, and anomalies
- identify data quality issues early
- guide model choice and feature design

More on EDA

It is important to understand that visualization is part of the analysis process and that plots are tools for thinking, not just reporting. Poor visualization often leads to poor conclusions.

What can you hope to find in EDA?

- Hidden outliers
- Skewed distributions
- Class imbalance
- Misleading correlations

Typically, you should begin the analysis by

i. understanding the business perspective
ii. inspect the data
iii. visualize distributions and relationships
iv. decide what kind of problem this is and what methods need to be applied before choosing the desired model

Case study 1: Spam vs. regular email

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* `spam` *and* `email`.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

An objective is to *learn* how to differentiate in spam vs. regular email.

What are the potential *learners/decision rules* you will employ just looking at this data? No statistical tool required as of now.

Case study 1: Spam vs. regular email

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* `spam` *and* `email`.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

An objective is to *learn* how to differentiate in spam vs. regular email.

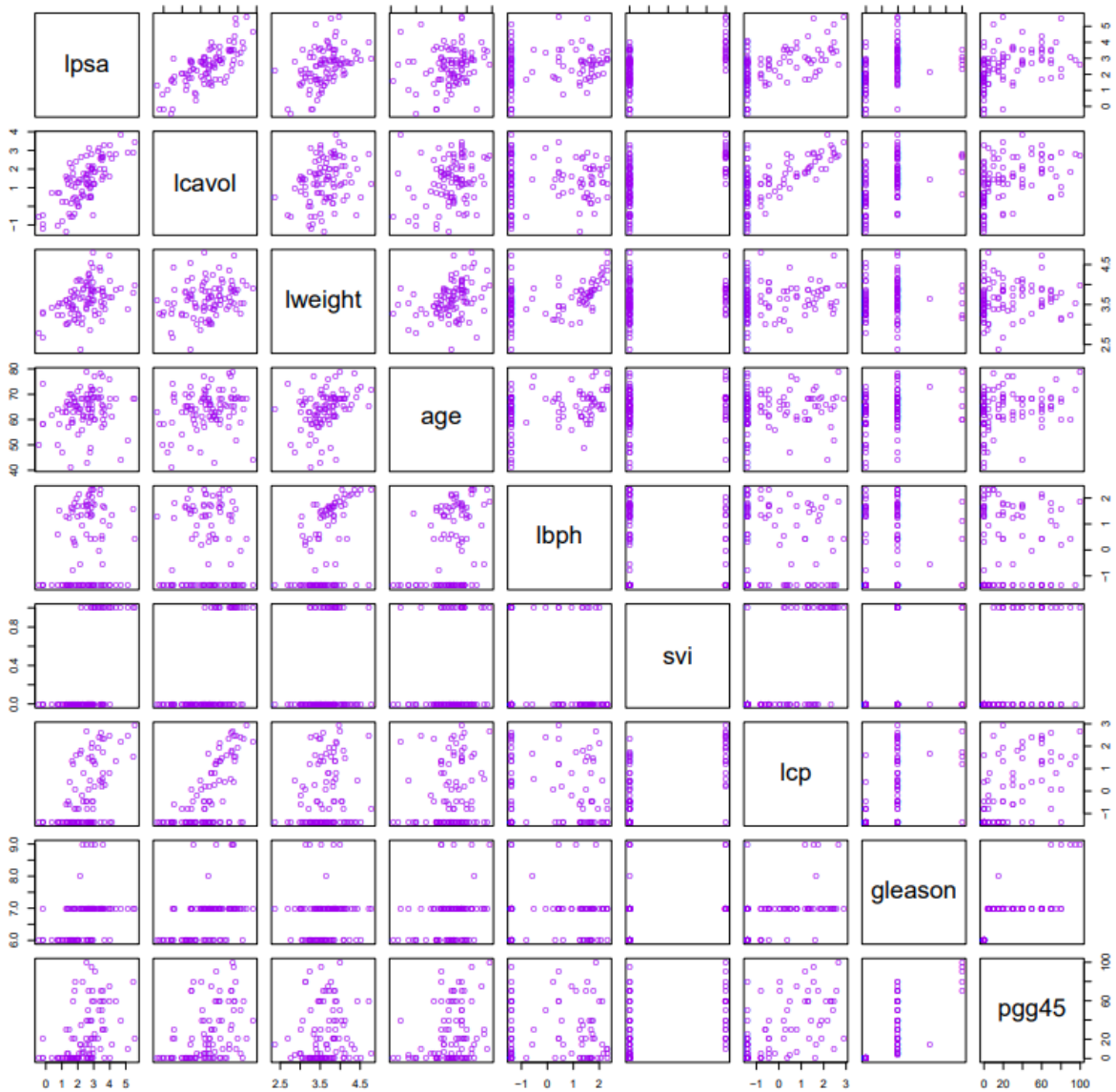
What are the potential *learners/decision rules* you will employ just looking at this data? No statistical tool required as of now.

$$\text{if } (\texttt{\%george} < 0.6) \;\&\; (\texttt{\%you} > 1.5) \quad \text{then } \texttt{spam} \text{ else } \texttt{email}.$$

Another form of a rule might be:

$$\text{if } (0.2 \cdot \texttt{\%you} - 0.3 \cdot \texttt{\%george}) > 0 \quad \text{then } \texttt{spam} \text{ else } \texttt{email}.$$

Case study 2: Prostate Cancer



Case study 2: Prostate Cancer (contd.)

The data come from a study by Stamey et al. (1989), which examined the correlation between the level of prostate-specific antigen (PSA) and a number of clinical measures, in 97 men who were about to receive a radical prostatectomy.

The goal is to predict the log of PSA (*lpsa*) from a number of measurements.

The features include log cancer volume (*lcavol*), log prostate weight *lweight*, age, log of benign prostatic hyperplasia amount *lbph*, etc.

The previous page had a scatterplot matrix of the variables. Some correlations with *lpsa* are evident, but a good predictive model is difficult to construct by eye. *Supervised* problem, or, *regression* problem.

Case study 3: Handwritten digits

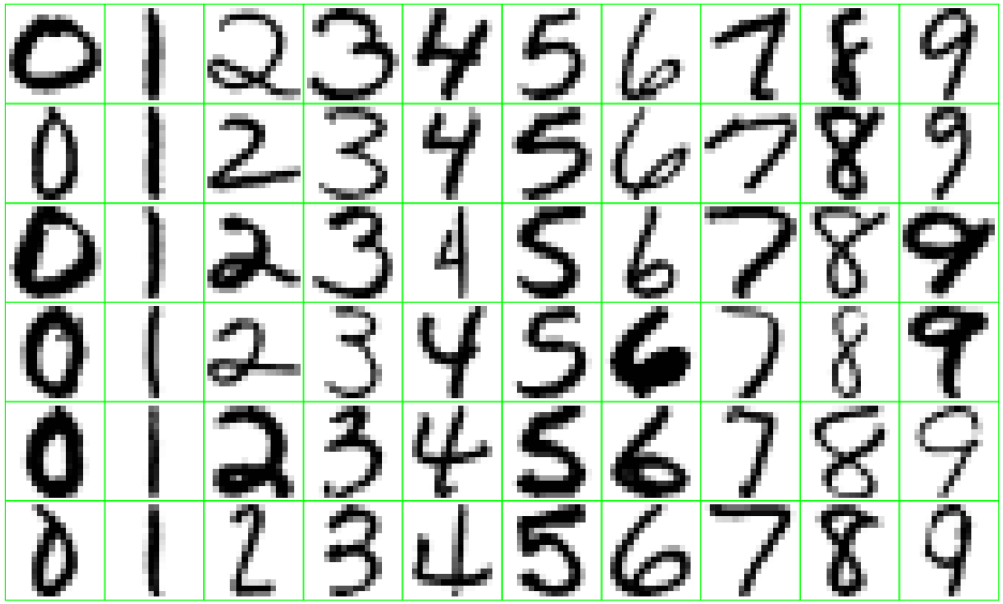


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

This is a classification problem for which the error rate must be kept very low to avoid misclassification.

In order to achieve this low error rate, some objects can be assigned to a "don't know" category and sorted instead by hand.

Variable Types and Terminology

- quantitative variable (continuous)
- qualitative variable (categorical, discrete)
- ordered categorical or ordinal
- non-ordinal (dummy variables)
- target (output, response, dependent variable)
- inputs (feature, predictor, variable, independent variable)
- regression vs. classification
- training vs test data

Let's code but remember!

Your plots should speak for themselves and should be complete in all regards. You should also care for the color-blind people.

Let's move to the code now: `https://colab.research.google.com/drive/15ixkCab9DfHtDuSRRslV8UJmg7i9IAKC?usp=sharing`