# INFO523 Sec 004 Data Mining and Discovery –

# Guidelines for final project

# Fall 2025

The goals are twofold: a) to utilize what you have learned from the course; b) to give you an opportunity to work on your own data.

## Dataset

Choosing a dataset is something you should do carefully but also relatively quickly. Choose a dataset from any online repository e.g. Kaggle, TidyTuesday etc., or from any research paper that you might have read recently, or you can use your own data.

## Questions

You have to think of a **minimum of 2 questions** that you can address following the steps of a data mining analysis and using the different methods learnt in this course. Here are a few ideas to help you get started:

Time Series Analysis

- Predict stock market prices using historical data

Anomaly Detection

- Detect credit card frauds using transaction data.
- Identify network intrusions from server logs.

Association Rule Mining

- Find patterns from a supermarket transaction dataset (like market basket analysis).

Social Network Analysis

- Identify influential nodes in a social network.
- Study the spread of information (or misinformation) in a network.

Multimedia Mining

- Develop a music recommendation system based on user listening history.

- Use data mining techniques to automatically tag and categorize videos.

Interactive Data Visualization

- Develop an interactive dashboard that allows users to explore a large dataset, finding patterns or outliers.

Ensemble Methods

- Compare the performance of various ensemble techniques (like boosting, bagging) on a challenging classification problem.

## Proposal writing and submission – 2 page max

The proposal should contain the following;

1. Dataset selected
2. Project goals and a minimum of two questions based on the dataset selected
3. Tentative plan of analysis e.g. methods that will be used for data analysis, visualization etc.
4. Expected outcomes (what results you think you will get after the analysis)

Keep in mind the proposal is a rough idea of what analysis you will perform; the plans can change down the road when you actually start analyzing the data. However, please put in some effort to write a detailed proposal as it will help you get started.

***PS:** If possible, avoid changing the question completely or picking up a different question after proposal is submitted.

## Report Write-up and Submission

(**Do not write a lengthy project report, keep it concise**)

The project report should contain the following parts;

1. **Introduction (1-2 paragraphs):** Brief introduction to the dataset. Write it imagining that your project is a standalone document and the grader has no prior knowledge of the dataset.

2. **Question 1**
    a. **Introduction (1-2 paragraphs):** Introduction to the question and what parts of the dataset are necessary to answer the question. Also discuss why you're interested in this question.
    b. **Approach/Methodology (1-2 paragraphs):** Describe what types of methods you used to address your question. For each method, provide a clear explanation as to why this method (e.g. imputation, transformation, correlation analysis, etc.) is best for providing the information you are asking about.
    c. **Data Analysis and results:** In this section, provide the code that generates the results for your analysis. Show plots, figures and tables wherever necessary
    d. **Discussion (1-3 paragraphs):** In this section, interpret the results. Identify any trends revealed (or not revealed) by the data analysis from previous step. Speculate about why the data looks the way it does.

3. **Question 2**

    Same structure outlined for Question 1, but for your new question.

*Note - *The project report should be submitted as a PDF document, any codes generated should be submitted as .RMD, .R or .ipynb (if you are using python). It doesn't matter if you are using R or Python, but the codes should be properly annotated and easy to follow, and reproducible.*

## Grading (Total project points = 100 pts*)

| | | |
|---|---|---|
| Proposal | - | 30 pts |
| Report writing | - | 70 pts |

***Detailed grading scheme**

Proposal - 30 pts

Report writing - 70 pts

Introduction – 4 pts

Question 1:

Introduction – 3 pts

Methods – 10 pts

Data Analysis and result – 15 pts

Discussion – 5 pts

Question 2:

Introduction – 3 pts

Methods – 10 pts

Data Analysis and result – 15 pts

Discussion – 5 pts

**All the group members will receive the same base marks as the group project. (So, make sure that everyone contributes equally and nobody lags behind)**

## Deadlines

Proposal submission – **November 3, 2025, 11:59 pm**

Final Project report submission – **December 10, 2025, 11:59 PM**