# Data Science - Assignment 2

## Submission Requirement

- Submit **one `.ipynb` notebook**
- Use:
    - Code cells for implementation
    - Markdown cells for explanations and answers
- Your notebook should be well-organized and readable

## Background

You are working as a data scientist for a wine company.

Your task is to analyze the **Wine Quality Dataset** to understand:

- How different chemical properties behave
- Whether there are unusual observations (outliers)
- How the data is distributed

Your analysis will help the company better understand wine characteristics and quality patterns.

You can download the dataset using this link (CSC-405-605-705-Data-Science/upload slides/WineQT.csv at main · XiaochenLi-w/CSC-405-605-705-Data-Science) or from Kaggle (https://www.kaggle.com/datasets/yasserh/wine-quality-dataset).

## Part 1: Data Exploration (Descriptive Statistics) — 50 points

### Task 1: Load the Dataset

- Load the **Wine Quality Dataset** using `pandas`

---

### Task 2: Statistical Summary

Select **three variables of your choice** from the dataset.

For each variable, compute:

- mean
- median
- variance
- standard deviation
- minimum
- maximum

---

### Write Code and Answer Questions

1. Which variable has the largest variance? What does this imply about the data?
2. Which variable appears the most stable? Why?

---

### Task 3: Z-score & Outlier Detection

Select **one variable** (from the three you chose).

1. Compute the z-score for every data point:

$$z = \frac{x - \mu}{\sigma}$$

1. Apply the 3σ rule to identify potential outliers.

---

### Write Code and Answer Questions

1. How many potential outliers did you detect? Please include some explanations.
2. Do you think these outliers should be removed in a data science pipeline? Briefly explain.

## Part 2: Kernel Density Estimation (KDE) — 50 points

You will now analyze how a variable is distributed using Kernel Density Estimation (KDE).

---

### Task 1: KDE Modeling

- Select **one continuous variable** from Wine Quality Dataset.
- Use a Python library (e.g., `sklearn.neighbors.KernelDensity`) to:
  - Fit a KDE model
  - Choose a kernel function (e.g., Gaussian, Epanechnikov, uniform)
  - Choose a bandwidth

---

### Task 2: Evaluate Density

Compute the estimated density at the following points:

$$x = [\mu - 2\sigma,\ \mu - \sigma,\ \mu,\ \mu + \sigma,\ \mu + 2\sigma]$$

Report your results in a table:

| x value | Estimated density |
|---|---|
| … | … |

---

### Write Code and Answer Questions

1. What kernel function did you choose?
2. How does the density change as $x$ moves away from the mean?
3. Based on your results, where is the data most concentrated?

---

### Bonus (Optional, +10 points)

- Try a different bandwidth
- Compare the density values

Question:

- What effect does bandwidth have on the results?

---

## Academic Integrity

Each student is expected to complete this assignment independently.

- You may choose different variables, which is encouraged.
- If multiple students select the same variables for all tasks, submissions may be compared line-by-line.