# Lab4 Assignment: Wine Quality Data Mining

## *Overview*

You have been tasked by a company to analyze and process data related to Wine Quality (WineQT.csv) from various chemical parameters. Your role involves extracting key statistical insights from the dataset and visualizing the data to uncover patterns and trends. This project will test your ability to work with data mining, apply statistical methods, and create informative visualizations.

## *Objectives*

- Extract and calculate important statistical measures from the Wine dataset.
- Practice constructing data mining models with Python coding.
- Create visual representations of the data to highlight key findings.

## *Statistic Task*

1. Identify the three chemical features (attributes) that have the greatest impact on wine quality. You may use **heatmap** to visualize the correlation matrix from provided data. Use a **Markdown** cell to explain the purpose of your figure and the big takeaway.
2. Select any four attributes and appropriate statistical results, and use **bar** plots to present them. Use a **Markdown** cell to explain the purpose of your figure and the big takeaway.
3. Select any four attributes and appropriate statistical results, and use **line** plots to present them. Use a **Markdown** cell to explain the purpose of your figure and the big takeaway.

## *Data Mining Tasks*

You are required to analyze the data using different data mining techniques:

1. **K-Means**:
    - Can you classify wines into different price categories based on their chemical attributes ('fixed acidity', 'volatile acidity', 'citric acid', 'chlorides', 'free sulfur dioxide', 'sulphates', 'alcohol') using K-means clustering?
    - According to your results, how many different price categories should be divided into specifically? Please explain or prove your point with charts and data table.
    - Visualize your results using clear, readable font sizes for the x- and y-axis labels as well as the graph title.
    - Use a **Markdown** cell to explain the purpose of your figure and the big takeaway.

2. **K-NN**:
    - What is the optimal value of k (number of neighbors) for predicting wine quality accurately?
    - You may use cross-validation score ‘cross_val_score’ to verify your assumption.
    - Visualize your results using clear, readable font sizes for the x- and y-axis labels as well as the graph title.
    - Use a **Markdown** cell to explain the purpose of your figure and the big takeaway.
3. Please add necessary comments for each key function in your code.

### *Submission*

- Save your visualization results and upload them to GitHub.
- Code: Include the entire code used for the analysis and upload them to GitHub.
- Submit your GitHub link to Canvas.