**DATA2001 Sem 2 2024 - Assignment 1 (Weight: 20%)**

**Due: 02/Sep/2024 3pm**

The aim of this assignment is to gain practical experience in analysing structured data. You must complete this in Python using a Jupyter notebook. You will need to submit a single Jupyter notebook (.ipynb file) via Blackboard.

**Dataset**:

The dataset for this assignment is provided in blackboard. The dataset contains results from the chemical analysis of different wines. These wines are grown in the same region in Italy but by 3 different cultivators. The analysis determined the quantity of 13 components found in each of the wine samples. The dataset has 178 samples and 14 attributes.

1. Wine (3 different cultivators of wine are represented by the three integers: 1 to 3).
2. Alcohol
3. Malic acid
4. Ash
5. Alcalinity of ash
6. Magnesium
7. Total phenols
8. Flavanoids
9. Nonflavanoid phenols
10. Proanthocyanins
11. Color intensity
12. Hue
13. OD280/OD315 of diluted wines
14. Proline

More information on dataset can be accessed from here: [Wine - UCI Machine Learning Repository](http://archive.ics.uci.edu/dataset/109/wine) . Note: Different versions of this dataset that can be found online should **not** be used for this assignment.

The submitted notebook should address 6 tasks (see marking grid for mark allocation):

1. Data Preparation: Read the dataset using the “pandas” library. Can you identify the missing data both row- and column-wise in the dataset? Handle data quality issues you found in an appropriate way. Explain how you did it along with the reasons of your choice.
2. Exploratory Data Analysis (EDA): Perform a detailed univariate and bivariate EDA on the columns in the dataset. Produce plots and report your observation for each plot clearly. In case the given dataset has many attributes, you can focus on performing EDA and reporting on just the most important attributes.
3. Find the mean and standard deviation for each type of component for each cultivator of wine and report your findings in a table. Comment on apparent differences between the cultivators of wine (i.e., vignerons).
4. Find the correlation among the numerical columns for each cultivator. Produce visualisations for the correlations and explain the observed results.
5. Perform k-means clustering on the data. Comment on the number of clusters chosen, on possible limitations, and on any form of uncertainty about the results. Are the results in agreement with what you observed in the EDA?
6. Perform principal component analysis on the data. Comment on the results, plot the percentage of variance explained by each principal component. Also plot the principal components which you think are of interest, report your observations and limitations.

**Note**: The submitted Jupyter notebook should be commented properly and written in a way that makes it easy for the reader to understand. For marking purpose, your code may be rerun to verify the results.