Assignment 2: Predicting Wine Quality with Linear Regression

- **Introduction**
    - The goal of this assignment is to apply the basic concepts of data science to a real-world problem.
    - The dataset you will be using is the Wine Quality dataset: https://archive.ics.uci.edu/ml/datasets/wine+quality, which contains information about red wine quality.
    - Your task is to answer the following questions:
        - Can you build a model to predict the quality of wine?
        - What are the most important factors that influence the quality of wine?
        - What are the limitations of your model?
- Data Preparation:
    - The first step is to load the dataset into a data frame.
    - You will need to clean the data by removing missing values and outliers.
    - You may also need to transform some of the data, such as converting categorical variables to numerical values.
- Exploratory Data Analysis (EDA):
    - Once the data is clean, you can start exploring it.
    - This involves creating visualizations and performing statistical analyses to understand the data.
    - Some of the questions you might want to answer during EDA include:
        - What is the distribution of the wine quality scores?
        - What are the relationships between the different features?
        - Are there any outliers in the data?
- Model Building:
    - Once you have a good understanding of the data, you can start building a model to predict wine quality.
    - There are many different machine learning algorithms that you can use.
    - Some of the most common algorithms for regression problems (like predicting wine quality) include linear regression, decision trees, and random forests.

- Model Evaluation:
    - Once you have built a model, you need to evaluate its performance.
    - This involves using a holdout dataset to test the model on unseen data.
    - You can use metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared to evaluate the model's performance.
- Conclusion:
    - In the conclusion, you should summarize your findings and discuss the limitations of your work.
    - You should also discuss the implications of your findings for the real-world problem.

Dataset Download Link:

The Wine Quality dataset can be downloaded from the UCI Machine Learning Repository website: https://archive.ics.uci.edu/ml/datasets/wine+quality

The 10 questions that students need to solve are:

1. What is the distribution of the wine quality scores?
2. What are the relationships between the different features?
3. Are there any outliers in the data?
4. What is the accuracy of the linear regression model?
5. What are the most important features for the linear regression model?
6. What is the MSE of the linear regression model?
7. What is the R-squared of the linear regression model?
8. How can you improve the performance of the linear regression model?
9. What are the limitations of the linear regression model?
10. What are the implications of your findings for the real-world problem?