* **Once you start, you have 120 minutes to complete.**
* **You cannot search over the Internet. Any use (a part or whole) of online materials found from the Internet will be considered as cheating.  Suspicious cheating cases will be reported to the College;**
* **You can only use the files, if any, provided in the question. You are not allowed to change the filename unless instructed. Otherwise, zero marks will be awarded to the whole exam.**
* **You are not allowed to share this paper or your work, in any form, to others. Violation will be reported to College.**
* **Save files, if needed, frequently when working with your PC;**
* **Answers will be automatically submitted at the end of the exam; Unsaved and un-uploaded files may be lost;**
* **No (re-)submission will be accepted after the Exam;**
* **Instructors have no responsibility for any network interruption, PC failure, file corruption, etc.**

**Dataset description:**

The dataset is related to variants of red wine. The following shows the variable information:

1 - fixed acidity  
2 - volatile acidity  
3 - citric acid  
4 - residual sugar  
5 - chlorides  
6 - free sulfur dioxide  
7 - total sulfur dioxide  
8 - density  
9 - pH  
10 - sulphates  
11 - alcohol  
12 - quality (score between 0 and 10)

**Goal:**

You goal is to present the variable 12 - quality, given all other variables.

**Tasks:**

**Part 1: Data Pre-Processing (3 Points)**

1. Check and drop all the rows that are empty, if any;
2. Check for missing/null value. Fill them, if any, the mean of the corresponding column;
3. Check for duplicate records and only retain the first one;
4. Perform any necessary pre-preprocessing (This is optional if you think it is necessary.)

**Part 2: Data Exploration (8 Points)**

1. Display a summary statistic of the data.
2. Show correlations among features
3. Display some scatter plot of some interesting features (based on the correlation)
4. Display the heatmap of those interesting features.
5. Select a subset of features for the subsequent process.
6. Create a cell (a Markdown cell) in the Jupyter notebook and explain your choice

**Part 3: Linear Regression (6 Points)**

1. Build a linear regression model using 90:10 training-testing split.
2. Evaluate the model by calculating the RMSE and R2 metrics, plot the predicted vs actual and print the model coefficients.

**Part 4: Linear Regression (4 Points)**

1. Apply standard scaling and redo Part 3.
2. Create a cell (a Markdown cell) in the Jupyter notebook and write down your observation and conclude if scaling is necessary.

**Part 5: Linear Regression with Ridge using Scaled Dataset (6 Points)**

1. Based on the scaled dataset you created in Part 4, create linear regression models with Ridge as outline below:
   1. Use the following alpha values: [100 10 1 0.1 1e-2 1e-4 1e-6 1e-8].
   2. Use the same training-testing split as stated in Part 3.
   3. Build linear regression models with Ridge
2. Evaluate these models by calculating the RMSE and R2 on testing data.
3. Create a cell (a Markdown cell) in the Jupyter notebook. Compare the performance of 1) linear regression, 2) linear regression with scaling, and 3) linear regression with Ridge.

**Part 6: Discussion (8 Points)**

1. Create a cell (a Markdown cell) in the Jupyter notebook and answer: From managerial perspective, how can one improve the quality of the wine? Explain your answer.
2. Create a cell (a Markdown cell) in the Jupyter notebook and answer: What is overfitting? How linear regression with Ridge solve the problem of overfitting?

**Re-run the whole notebook to make sure there is no error. Any error will result in signficant mark deduction.**