**Feature Selection - Wine Workshop**

**Objective:** To practice Feature Selection using Pearson Correlation

**Data**

The dataset provided is called ‘wine.csv’ and the description of the data can be found here - <https://archive.ics.uci.edu/ml/datasets/wine>

**Your Tasks**

1. Use Pandas to read in ‘wine.csv’.
2. Create a correlation matrix using Pandas’ *corr* function.

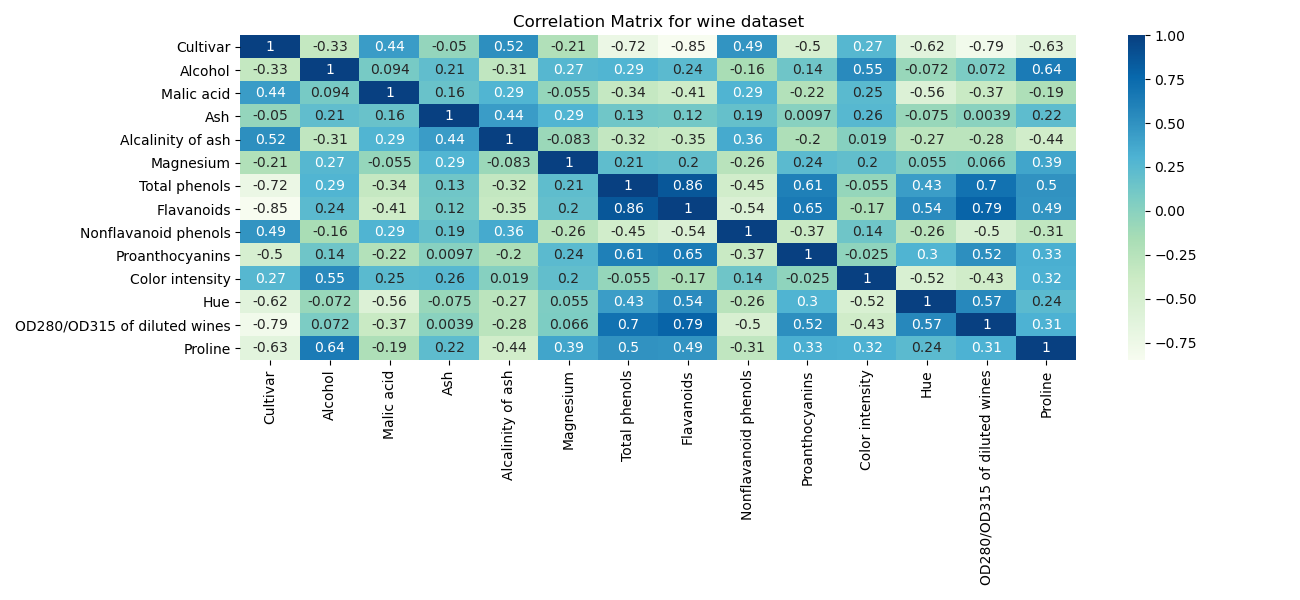


The corr function returns **corr\_mat**, a Pearson Correlation matrix, with values between -1 and 1.

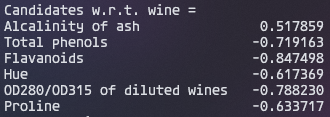
The documentation can be found here – <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html>

1. Display the heatmap of the Pearson Correlation matrix. Title the heatmap as ‘Correlation Matrix for wine dataset’.

Here is an example:



1. Using corr\_mat, compute the features that are either more than 0.5 (positively correlated) and less than -0.5 (negatively correlated) correlated to the column ‘Cultivar’ (our label/class in this dataset). Print out the potential features and their correlation like so:



1. For each potential feature (above), check that it is in turn not highly correlated with the rest of the potential features. Consider two or more potential features as highly correlated if the Pearson correlation coefficient between them is greater than 0.6 (positively correlated). If two potential features are highly correlated, select only the one with a higher correlation against our target ‘Cultivar’.

Print out the potential features that are removed and the final selected features like so:

