

1. Download Wine dataset (wine.data) from UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/wine>).

Note: This website is a good resource for open data

2. Read the wine.data file as data frame (name the dataframe as wine_data). Don't be concerned that it is not a .csv file. Open the document in text editor and see that it indeed has comma separated values. You can read this file using regular csv reading functions. Note that the file does not have header (column headings)

Additional Note: You could directly read the data from the web by passing url as the path argument for csv reading function

3. Names of 13 attributes can be found in the description. It is a labelled data set with 3 classes of wines; note that 1st column in the data corresponds to wine class and has to be removed. Add column names to the data frame.
4. Do a descriptive analytics by comparing mean, median etc. of 13 attributes
5. Create scatter plots for some attribute combinations (trying out for all combinations would be exhaustive – dimensionality is a curse!)
6. Run k means clustering on the attributes with $k = 3$.
7. Visualize scatter plots generated in Step 5 with colors from cluster numbers
8. What would have been your suggestion for number of clusters (k) if it was not provided as input?
9. Rerun the clustering with just Proline attribute (last column in data).
10. Compare the cluster outcome of just Proline attribute (Q 11) vs All attributes (Q 7). If they are comparable, what do you think is the value addition from other 12 attributes? What is special about the values in Proline column?