Machine Learning, January 2022 Assignment Task

1 Situation

You are working for a business that develops new perfumes for the cosmetic industry. The business has been doing experiments on mixtures of different chemicals that are used in perfumes. They have collected two data sets.

The first data set 'Perfume Score' contains the results of many experiments. In each experiment a different person was given a perfume with a random mixture of scent chemicals. Each person was asked to give a score for how nice the scent was to them. A higher (larger) score means better / nicer. The table includes data for the quantity in milligrams of 18 different scent chemicals. The names of the chemicals are in the first row of the table. The final column of the table includes the score given for that mixture of scent chemicals.

The second data set 'Perfume Preference' is data from a different set of experiments with different people. In this data set each customer is identified by a unique code. The customer code has the form 'C_00000n', where the 'n' is a number. All customer codes are unique. Each row of the 'Perfume Preference' table contains the recipe for the perfume that specific person liked best. Again, each of the numbers represents a quantity in milligrams.

2 Task

Your task is to analyse the two data tables both graphically and quantitatively using Machine Learning algorithms.

Your objective is to discover insights could help the company produce better perfume products. The company is interested to answer questions including:

- Predicting the Scent Quality Score for any new perfume based on the quantity of scent chemicals used
- Understanding if any scent chemicals are strongly related. That is, they have a similar response from customers
- Understanding if there are specific 'types' of customer who prefer specific kinds of mixture, and if so:
    - How many different groups are there?
    - What would be the best mixtures for each group?
    - For each group, how tightly clustered are they? That is, is every person in the group very similar for each of the chemicals? Or are they similar with regard to some chemicals but have a wide range of responses to other scent chemicals? Specifically, for any customer group identified indicate the spread (variance) for each group for each dimension (quantity of scent chemical)

The organisation needs to know how much they can rely on the insights you generate. Thus, you should provide a quality indicator for any supervised learning models you develop.

The organisation needs to ensure that they can understand and reproduce your analysis. Thus you should use Jupyter Notebooks to create your models. You notebooks should include both executable cells containing your modelling elements and 'markdown' cells. Markdown cells should:

- explain the models
- provide any rationale for why you have processed the data in a specific way, and
- document conclusions and insights about the data

You should submit a single Jupyter Notebook (a .ipynb file) containing the results of your analysis.

Version 3, 27/01/2022 17:14

3 Submission of Assignments

You are required to submit a .ipynb file with your analysis for the task 1.

Assignment submissions must be made by Monday 7th February at 12:00 mid-day China time (04:00 UK).

You will be informed separately by Oxford Study Abroad of how to submit your assignment.