**Question 1 (80%)**

The focus of this assignment is clustering. You will be using some data of expression levels of proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex. In the original data there are 77 proteins/protein modifications. There are 38 control mice and 34 trisomic mice (Down syndrome), for a total of 72 mice. In the experiments, 15 measurements were registered of each protein per sample/mouse. Therefore, for control mice, there are 38x15, or 570 measurements, and for trisomic mice, there are 34x15, or 510 measurements. The dataset contains a total of 1080 measurements per protein.

 Each measurement can be considered as an independent sample/mouse.   
  
The eight classes of mice are described based on characteristics such as genotype, behavior and treatment. According to genotype, mice can be control or trisomic. According to behavior, some mice have been stimulated to learn (context-shock) and others have not (shock-context) and in order to assess the effect of the drug memantine in recovering the ability to learn in trisomic mice, some mice have been injected with the drug and others have not.   
  
*Classes:*c-CS-s: control mice, stimulated to learn, injected with saline (9 mice)   
c-CS-m: control mice, stimulated to learn, injected with memantine (10 mice)   
c-SC-s: control mice, not stimulated to learn, injected with saline (9 mice)   
c-SC-m: control mice, not stimulated to learn, injected with memantine (10 mice)   
  
t-CS-s: trisomy mice, stimulated to learn, injected with saline (7 mice)   
t-CS-m: trisomy mice, stimulated to learn, injected with memantine (9 mice)   
t-SC-s: trisomy mice, not stimulated to learn, injected with saline (9 mice)   
t-SC-m: trisomy mice, not stimulated to learn, injected with memantine (9 mice)

*Feature information in columns*

1 Mouse ID   
2- 78 Values of expression levels of 77 proteins (their names are given as names of the columns)

79 Genotype: control (c) or trisomy (t)   
80 Treatment type: memantine (m) or saline (s)   
81 Behavior: context-shock (CS) or shock-context (SC)   
82 Class: c-CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m

The data have been placed as a separate module at quercus.

**Task:** Using any of the methods and techniques you learned under unsupervised learning (clustering, PCA, dissimilarity metrics, heatmap plots, biplots etc.) investigate any patterns in the data in order to answer questions like the following:

1. Are there any distinct clusters in the measurements?
2. Are these clusters associated with any of the different genotype, behaviour and/or treatment? E.g. you may find that some of these clusters are more representative for some of the classes described above.
3. Do we see any particular set of proteins that exhibit a distinct expression pattern (profile) in any or all of these clusters? (You can use different plotting methods with/without clustering of the proteins in order to identify the patterns.)

Make sure you clearly explain what you are doing and why, at every step.

**Bonus question:** In the data we consider the samples as independent. The reality is that most likely they are not since multiple measurements are coming from the same mouse. Are there any implications about this? Discuss possible ways *if any* that you could improve the overall analysis plan by taking into consideration this potential correlatedness in the data. *Note:* I am not looking for any particular *right* answer here. You could do some research in order to answer this question.

**Missing data:** There is significant amount of missing values in the data. We have not discussed any formal way to deal with missing data, so you are not expected to apply any particular method. Some practical suggestions are to remove columns and/or rows with large percentage of missing values and to try to impute the remaining missing values. Again, for the imputation, there are plenty sophisticated methods you could use but here for simplicity you could use the mean value of the rest of the samples in the same class as defined above. For example, if the sample i has a missing value for protein j, impute that missing value by the average value of the rest of the samples **that belong to the same class.**

Since the nature of this assignment is more exploratory, you are allowed to use freer format than the one before. However, code needs to be provided separately as a separate R script file, properly organized and documented.  Tables and figures need to be appropriately numbered, labeled and explained. There is no limit in the number of tables and figures but they need to be meaningful and add something in the paper. You could place the tables and figures in the paper or at the end, but make sure you reference them properly and clearly.

Using the typical sections of Introduction, Methods, Results, Conclusions/Discussion is strongly recommended for better readability.

**Question 2 (20%)**

Think of ways that you could use some of the unsupervised learning methods in problems you are currently facing or have recently faced in your research or work. Explain how and why they can be helpful in investigating the problems and answering the questions you are dealing with.