Project: Capstone Project 1: Data Wrangling

This project is based on MALDI-TOF mass-spectrometry data. Information about this data set is provided in read-me text below. So far, I have not found any need to perform cleaning steps, address missing values, or handle outliers. I will update this document as needed if extra steps are found necessary

---

The dataset contained in this archive is related to bacterial identification using MALDI-TOF mass-spectrometry data.
It consists of 931 MALDI-TOF mass-spectra, divided as follows :

- 571 spectra belong to a "reference dataset". The bacterial panel considered consists of 20 Gram positive and negative bacterial species covering 9 genera among which several species are known to be hard to discriminate by mass spectrometry. Each species was represented by 11 to 60 mass spectra obtained from 7 to 20 bacterial strains.
- 360 spectra belong to a "mixture dataset". This dataset involves 10 pairs of species of various taxonomic proximity :
    - 4 mixtures, labelled A, B, C and D, involving species that belong to the same genus,
    - 2 mixtures, labelled E and F, involving species that belong to distinct genera, but to the same Gram type,
    - 4 mixtures, labelled G, H, I and J, involving species that belong to distinct Gram types.

Each mixture is represented by 2 pairs of strains, which were mixed according to the following 9 concentration ratios: 1:0, 10:1, 5:1, 2:1, 1:1, 1:2, 1:5, 1:10, 0:1. Two replicate spectra were acquired for each concentration ratio and each couple of strains, leading altogether to a dataset of 360 spectra.

These spectra were acquired by the VITEK-MS system (bioMerieux, France) and pre-processed by the standard algorithm used for routine (culture-based) bacterial identification, which provided a peak-list representation where a mass spectrum was represented by a vector of 1300 dimensions.

Further details about the dataset can be found in the supplementary materials of the paper "Automatic identification of mixed bacterial species fingerprints in a MALDI-TOF mass-spectrum" (http://www.ncbi.nlm.nih.gov/pubmed/24443381).

Content :

The archive is made of 7 files :

- this README file

- the file "pure_spectra_matrix.csv" :
    - plain-text file with semicolon-separated fields containing the peak-list representation of the reference dataset
    - 1 line = 1 spectrum ; 1 column = 1 variable
    - can be read with R by the command
        > tab = read.csv2("at-uci/pure_spectra_matrix.csv", header = F, dec = ".")
      which provides a [571 x 1300] data frame of numerical values
- the file "pure_spectra_metadata.csv" :
    - plain-text file with semicolon-separated fields containing the meta-data associated to the reference dataset
    - 1 line = 1 spectrum ; 1 column = 1 variable
    - variables :
        - Species : species identifier (6-letter codes in which the first 3 letters encore the bacteria genera)
        - Strain : strain identifier (integer from 1 to 213)
        - can be read with R with the command :
          > tab = read.csv2("at-uci/pure_spectra_metadata.csv")
      which provides a [571 x 2] data frame
- the file "pure_spectra_taxonomy.txt"
    - plain-text file providing the taxonomic definition of the reference panel
    - each line corresponds to a node of the taxonomy and provides :
    - the identifier of the node in the first column
    - the identifier of its parent in the second column
      (NB : for the root node, the second column corresponds to the first one)
- the file "pure_spectra_taxonomy.pdf"
    - a graphical representation of this taxonomy (generated by graphviz)
- the file "mixed_spectra_matrix.csv"
    - plain-text file with semicolon-separated fields containing the peak-list representation of the mixture dataset
    - 1 line = 1 spectrum ; 1 column = 1 variable
    - can be read with R by the command
        > tab = read.csv2("at-uci/mixed_spectra_matrix.csv", header = F, dec = ".")
      which provides a [360 x 1300] data frame of numerical values
- the file "mixed_spectra_metadata.csv"
    - plain-text file with semicolon-separated fields containing the meta-data associated to the mixture dataset
    - 1 line = 1 spectrum ; 1 column = 1 variable
    - variables :
        - Mixture_Label : species-level mixture identifier (a letter from A to J)
        - Mixture_Id : strain-level mixture identifier (1 or 2, corresponding to the couple of strains considered to represent a mixture)

- Species_1 : species identifier of the 1st species involved in the mixture (6-letter codes in which the first 3 letters encore the bacteria genera)
- Strain_1 : strain identifier of the 1st strain involved in the mixture(integer)
- Species_2 : species identifier of the 2nd species involved in the mixture (6-letter codes in which the first 3 letters encore the bacteria genera)
- Strain_2 : strain identifier of the 2nd strain involved in the mixture(integer)
- Replicate : integer (1 or 2) indicating the replicate index
- Ratio : relative Species_1:Species_2 concentration, stored as a string (1:0, 10:1, 5:1, 2:1, 1:1, 1:2, 1:5, 1:10, 0:1)
- Proportion : relative Species_1:Species_2 concentration, as numerical values
- can be read with R with the command :
  > tab = read.csv2("at-uci/mixed_spectra_metadata.csv")

which provides a [360 x 9] data frame