COMP3425 and COMP8410 Data Mining 2022
Assignment 2 Description of Data

Data and Metadata

The data supplied for the assignment arises from The Australian Data Archive's ANU Poll Dataverse [1]. As a student of the course, you are assumed to accept the Terms and Conditions of Use reproduced below. Please read them carefully. The custodian of the data has requested you delete your data at the end of the course.

In particular the data captures the results of a survey poll conducted in 2020 on the topic of Data Governance. You can find a complete description of the purpose of the poll and coding ofthe data (metadata) and also a descriptive summary of the poll results here:

https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/ZFGFNE

The data is provided to you for the assignment in two forms. The first is the **original** dataset as download from the ADA called **2.ADA.ANUPoll35.CSV.01474**, in comma-separated-values format. This data is described by the metadata in **1.ADA.OTHER.01474**.

The second is a form derived from the original, **pre-processed** for the COMP3425 data mining assignment, in comma-separated-values format called **3425_data.csv**. Below you will find a description of the pre-processing undertaken and this, in addition to the original metadata, willbe needed to assist your understanding of the data.

If you are a COMP3425 (undergraduate) student, you must work with the pre-processeddataset 3425_data.csv.

If you are COMP8410 (postgraduate) student you may use either the original or the pre-processed data, or both. The original will give you more opportunity to show off your technical skills and creativity, while the pre-processed one is more constrained but may save time, requiring you to spend less effort understanding the data, and helping to avoid some data errors. The same rubric will be used for marking in both cases, but the original dataset provides an extended learning experience and better opportunity for higher marks. Even if you use the original data, you may find it useful to observe the pre-processing that has been undertaken to seed ideas or to solve problems you encounter.

Pre-processing applied to derive 3425_data.csv

- Only a small selection of the original attributes have been retained.
- The following columns have been added, based on respondent's answers to questions [A4s and F2s], which have answers that range from very negative to very positive.
  - *A4F2_agg*: A normalized number in the range [0,1] that shows how opinionated is the respondent on different parts of A4 and F2.
    A4F2_agg =AVERAGE((AVERAGE(ABS(A4i-2.5))/1.5),(AVERAGE(ABS(F2j-3)) /2))
  - *opinionated*: A Boolean version of A4F2_agg that expresses whether the respondent is opinionated or not.

opinionated =IF(AND(*A4F2_agg* >=0.5, *A4F2_agg* <=1),TRUE,FALSE)

- The *undecided_voter* variable was added based on the given answer to *A2*.
- For two categorical columns, A2 and StateMap, double quotations were added to all cells and the empty cells were filled with “NaN”. For the rest of categorical columns, you can use the same approach to help Rattle recognise the type of data in a column.

References

[1] Biddle, Nicholas; Edwards, Ben; Gray, Matthew; Sollis, Kate, 2020, "ANU Poll 35 (August 2020): COVID-19 attitudes and behaviours (Wave 3)", doi:10.26193/ZFGFNE, ADA Dataverse, V2

Terms and Conditions of Use

From https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/XHORAI

I acknowledge that:

Use of the material is restricted to use for analytical purposes and that this means that I can only use the material to produce information of an analytical nature.

Examples of such uses are: (a) the manipulation of data to produce means, correlations or other descriptive summary measures; (b) the estimation of population characteristics from sample data; (c) the use of data as input to mathematical models and for other types of analyses (e.g. factor analysis); and (d) to provide graphical and pictorial representation of characteristics of the population or sub-sets of the population.

The material is not to be used for any non-analytical purposes, or for commercial or financial gain, without the express written permission of the Australian Data Archive.

Outputs (such as statistics, tables and graphs) obtained from analysis of these data may be further disseminated provided that I:
(a) acknowledge both the original depositors and the Australian Data Archive; (b) acknowledge another archive where the data file is made available through the Australian Data Archive by another archive; and (c) declare that those who carried out the original analysis and collection of the data bear no responsibility for the further analysis or interpretation of it.

Use of the material is solely at my risk and I indemnify the Australian Data Archive and its host institution, The Australian National University.

The Australian Data Archive and its host institution, The Australian National University, shall not be held liable for any breach of this undertaking.

The Australian Data Archive and its host institution, The Australian National University, shall not be held responsible for the accuracy and completeness of the material supplied.