Project Work 2

Cluster analysis

1.11.2024

In this project we will explore and analyze the Candidate Selector dataset (vaalikone_questions_all.csv) with the aim of clustering candidates based on their responses. Investigate whether these clusters align with political parties or other factors present in the Vaalikone Profiles (vaalikone_profiles_all.csv).

Data and Question Text:

The complete question text is available in both English (vaalikone_questions_eng.rtf) and Finnish (vaalikone_questions_fi.txt). You can uniquely identify candidates through the "ID" column in both vaalikone_questions_all.csv and vaalikone_profiles_all.csv, facilitating the merging of datasets.

The data set consists of answers to questions (along with their importance), with the full question text available in the rtf/txt files. Check the questions to become familiar with the data.

In overall, there are 31 questions (Q1-Q31), and the answers for two of them (Q21 and Q31) are binarized, allowing candidates to select multiple options. The columns starting with Q21.Fb and Q31.GVTPrt are related to these two questions.

Additionally, there are 5 count columns ("CountAnsw," "CountQ21.Fb," "CountQ31.GvtPrt," "CountNegW," "CountPosW") that count the candidate answers. As the answers are included as variables, those columns may contain redundant information, making them candidates for removal during data cleaning.

The columns starting from W1:IncomeDiff, W2:GayAdoption, and so on until the last one represent the candidate's opinions on the questions. Candidates can add comments and rank the importance of each question (high-medium-low).

Thus, the dataset (vaalikone_questions_all.csv) consists of two types of data: factor (all columns starting with Q) representing answers to questions and ordered factor with three levels for columns starting with W. As a tip for distance calculation, Gower distance could be a good choice since the dataset consists of factor and ordered factor variables.

Here is a link to an example of clustering with mixed data types in R, which might be helpful:

https://www.r-bloggers.com/2016/06/clustering-mixed-data-types-in-r/

Tasks:

1. Data Loading and Preprocessing:
    - Load the Vaalikone dataset (vaalikone_questions_all.csv) and understand the dataset's structure and variables.

- Handle missing values, outliers, or any data quality issues.
- Covert different variables in the dataset to a suitable data type (hint: all columns starting with Q to the factor and all the columns starting with W to the ordered factor)

2. Clustering Candidates:
    - Select a suitable distance measure for clustering.
    - Apply an appropriate clustering algorithm to group candidates based on their answers.
    - Justify the choice of the distance measure and clustering approach.
3. Visualize Clusters:
    - Create visualizations to represent the clusters.
    - Load the Vaalikone Profiles (vaalikone_profiles_all.csv) dataset and use the "ID" column to merge with Vaalikone dataset and results (clusters).
    - Check if the identified clusters align with political parties or other relevant factors (from vaalikone_profiles_all.csv).
4. Analysis and Interpretation:
    - Analyze the results and interpret the meaning of the clusters.
    - Assess the degree to which clusters correspond to political parties or other factors.
5. Documentation and Reporting:
    - Document the entire process, including data preprocessing, clustering techniques, and results (code and the output).
    - Prepare a concise yet complete report highlighting key findings and insights