

BUSA3020 Advanced Analytics Techniques

Assignment 3: Clustering

Dataset: Young People Survey Data from Slovakia

Chosen Software: R Studio and Python

Student Number: 45197083

## Introduction (Data Cleansing and Feature Engineering)

In 2013, the students of the Statistics class at [Faculty of Social and Economic Sciences, Comenius University in Bratislava, in Slovakia (FSEV UK)](https://fses.uniba.sk/en/) were asked to invite their friends to participate in this survey. After collecting all the results, the data was presented in comma-separated values (CSV) file.

The original dataset consists of 1010 rows and 150 columns, nevertheless we will be only dealing with 1010 rows and 41 columns in this assignment. The variables can categorised into three major groups: movie preferences, movie preferences and demographics.

Like other projects, data cleansing is required and it allows us to remove the missing values. It looks like there are missing values for every column, as a result the dimension will reduced after such process. The dimension of the data after the missing values is 865 rows x 41 columns.

After completing data cleansing, feature engineering is needed (such as changing the data type, changing the value of the data) before we conduct any analysis. The steps will be shown in the table below.

|  |
| --- |
| Table 1: Feature Engineering |
| The aim of feature engineering (such as converting categorical variables to numeric variables) allows us to apply Principal Component Analysis (PCA) in our dataset. Such algorithm is designed for continuous variables and tries to minimise our variables. Therefore, feature engineering is needed before our data analysis.   1. Convert the variable ‘**Gender**’ from categorical to a dummy variable: 0 = Male, 1 = Female 2. Convert the variable ‘**Left – Right Handed**’ from categorical to numerical. **Note**: The datatype is nominal data as there cannot be ordered. 3. Convert the variable ‘**Education**’ from categorical to numerical:   1 = College/Bachelor Degree, 2 = Currently a primary school pupil, 3 = Doctorate degree, 4 = Masters school, 5 = primary school, 6 = secondary school. **Note**: The datatype is nominal data as it cannot be ordered. (For instance: “Currently a primary school pupil” does not necessary greater than a person with a doctorate degree.   1. Convert the variable ‘**Only Child**’ from categorical to a dummy variable: 0 = No, 1 = Yes. 2. Convert the variable ‘**Village - Town** from categorical to numerical: 1 = Village,   2 = Town. **Note**: The datatype is nominal data as it cannot be ordered.   1. Convert the variable ‘**House – Block of Flats** from categorical to numerical:   1 = Blocks of flats, 2 = House/ bungalow. **Note**: The datatype is nominal data as it cannot be ordered. |

**Required Tasks**

## See if it is feasible to reduce the data to fewer variables using Principal Components Analysis (PCA) or similar.

|  |  |
| --- | --- |
| Table 2: Principal Component Analysis (PCA) | |
| **Introduction to Principal Component Analysis (PCA)**  It is feasible to use PCA in this instance as it has a large number of variables (41 variables) and allows us to reduce the dimension of a problem and provide new variables.  The new variables are the linear combinations of the original variables and they should be uncorrelated. In order to determine the principal components used in this dataset. A scree plot will be used.  A series of images are used to demonstrate the process of PCA. | |
| Figures | Interpretation: |
| Figure 1: Correlation matrix of all variables | The first step of the PCA is to generate a correlation matrix against all the variables. Ideally, the all the variables should be uncorrelated against each other and this is shown in this is instance. |
| Figure 2: Scree plot of the all principal components | We then generate a scree plot to determine whether how many principals’ components are needed to summarise the overall dataset.  By observing at the scree plot, it seems that it starts to ‘flatten out’ when there are 4 principal components (x axis represents to the number of principal components) |
|  |  |

## Find clusters of people based on their music and movie preferences

|  |  |
| --- | --- |
| Table 3: Clustering solutions based on music and movie preferences. | |
| **Background Information:**  Two separate datasets were created to answer this question: music (which contains all 19 variables) and movies (which contains all 12 variables). We will be also using PCA results (from each dataset) to conduct our cluster analysis.  **\*Note: the following solutions are interpreted based on k-means clustering solutions.** | |
| Diagram | Interpretation |
| Figure 4: Scatterplot of the music data clusters components | In figure 4, it shows a scatterplot of the cluster components by splitting the music variables from the main dataset. It is apparent that there are 3 clusters in the plot.  By examining which traits are allocated into which cluster, we can examine the music.data dataframe as the clustering solutions is provided. For instance, cluster 1 and 2 contain people who enjoy both music (in general) and pop music. On the other hand, people in cluster 3 do enjoy rock music but do not tend to enjoy Opera and Trance music. |
| Figure 5: Scatterplot of the movie data clusters components | In figure 5, it shows how the cluster solutions for the movies. 3 clusters are also shown in the scatterplot.  In cluster 1, it contains of interviewers who enjoy fantasy, fairy tales and comedy movies; on the other hand people do not seem enjoy horror movies. Additionally, in cluster 2, it contains of interviewers who enjoy documentary and war movies. Lastly, cluster 3 most contains people enjoy horror movies. |

## Compare cluster solutions for two (2) or more different algorithms

|  |
| --- |
| Table 4: The comparison of two different clustering algorithms |
| **Introduction of clustering algorithms**  Two different clustering algorithms were used in this assignment. They are: k-means clustering and hierarchical clustering. K-means clustering: It is a clustering method where we need to pre-define the number of K (number of clusters) before running the algorithm. There are various different algorithms to determine the optimal value of K. In R, there is a function called ‘NbClust’ where we can determine to the best number of clusters by stating the distance metric (in this case, it is Euclidian distance) and the clustering method (which is k-means). An example of the ‘NbClust’ output, it is provided below. Apart from k-means clustering, hierarchical clustering is also used in this assignment. Unlike k-means clustering, we are not able to determine the number of clustering before the conducting clustering. Instead, we need to find out the number by looking at a dendrogram.  **Comparison of clustering solutions for two or more algorithms**  After selecting the clustering algorithms, a cross-tabulation of results can allow us to compare solutions. Cluster analysis is conducted within the whole dataset.    We can interpret the table in the following: 11, 22, and 33 means that both clustering algorithms are allocating the data to the same clusters respectively. From the results from the cross-tabulation, it seems that both clustering algorithm did an alright job in allocating values in cluster 1. On the other hand, the solutions in cluster 2 and 3 look are poor. Based on the results above, we can concluded that both of the clustering are different towards each other. |

## Compare/profile clusters on their music, movie preferences and on demographics.