BUSA3020 Advanced Analytics Techniques

Assignment 3: Clustering

Dataset: Young People Survey Data from Slovakia

Chosen Software: R Studio and SPSS

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) 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 be 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 be reduced after such a 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 an 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 data type 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 data type is nominal data as it cannot be ordered. (For instance: "Currently a primary school pupil" does not necessarily greater than a person with a doctorate.
4. Convert the variable '**Only Child**' from categorical to a dummy variable: 0 = No, 1 = Yes.
5. Convert the variable '**Village - Town** from categorical to numerical: 1 = Village,
   2 = Town. **Note**: The data type is nominal data as it cannot be ordered.
6. Convert the variable '**House – Block of Flats** from categorical to numerical:
   1 = Blocks of flats, 2 = House/ bungalow. **Note**: The data type is nominal data as it cannot be ordered.

Required Tasks

1. **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 the dataset has a large number of variables (41 variables). By conducting PCA, it groups similar variables to become a brand new variable in which we do not need to completely remove the variable from consideration. One of the aims of PCA is also to ensure the variables are independent of each other. This can be shown in a correlation matrix which will be shown in the later of the analysis.

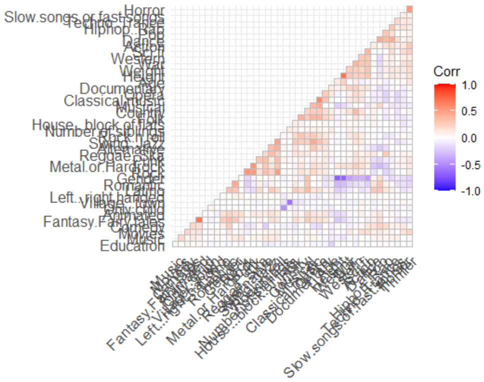


Figure 1: Correlation matrix of all variables

As shown in figure 1, a correlation matrix is created. Before conducting PCA, it is recommended to check the correlation between variables. If the correlation of some of the variables is high then we will consider removing them as they are essentially measuring the same thing. On the other hand, if the correlation is too low then it will not take those into account. In this instance, it seems that all variables are moderately correlated as a result it is feasible to conduct PCA.

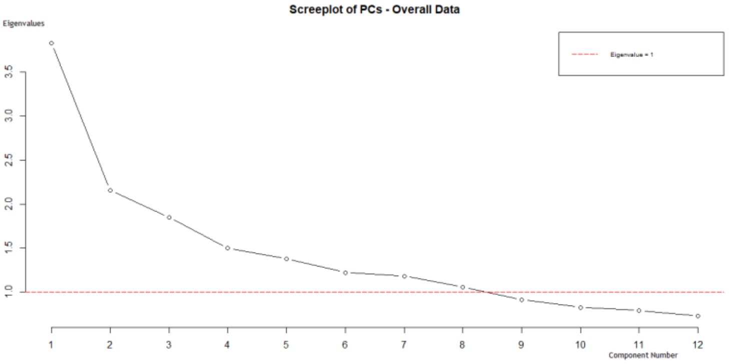


Figure 2: Scree Plot

A scree plot graphs the eigenvalues against the component number – this can be shown in figure 2. To interpret the scree plot, we shall look at the point where the “elbow” of the scree plot has flatten. In this instance, it starts to flatten out at component 4. Therefore we will use 4 principal components in this dataset.

2. 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.**

Cluster Solutions for Music Preferences

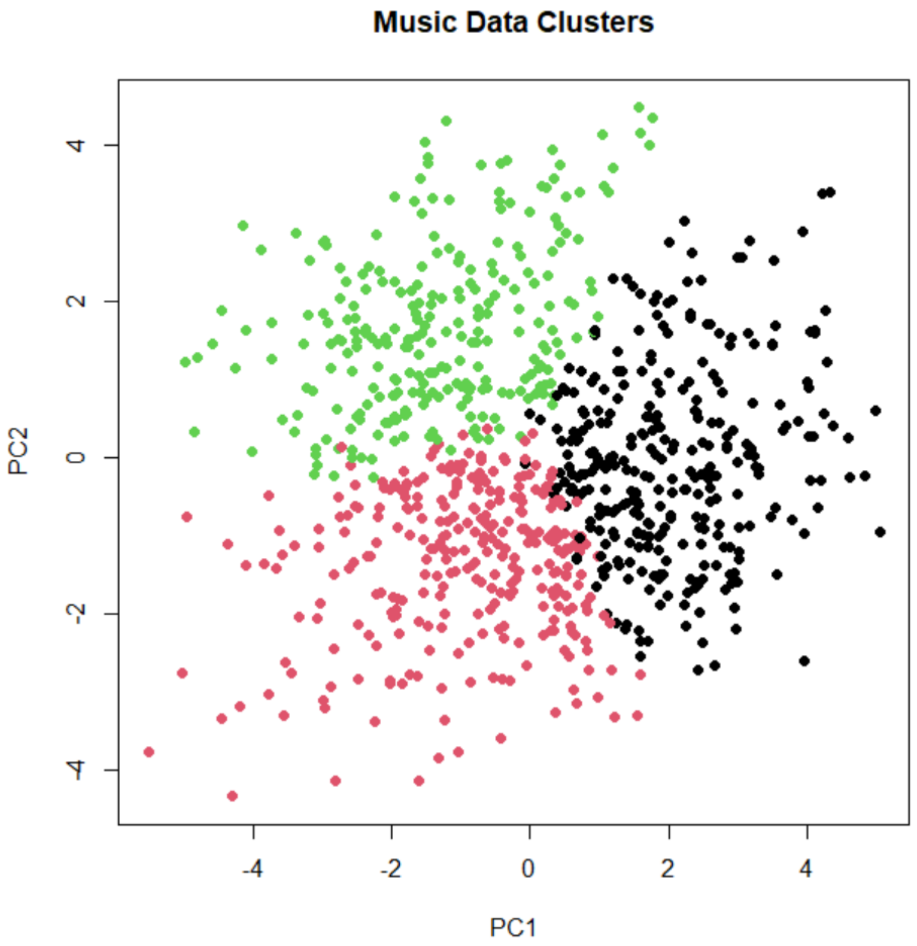


Figure 3: Scatterplot of the music data clusters components

In figure 3, it shows a scatterplot of 3 clusters.

To examine which variables are allocated into which cluster, a closer investigation to the dataset (as the clustering solutions are attached in the dataframe as a new column). For example, it seems that both people in cluster 1 and 2 like Pop music, while people in cluster 1 tend to dislike Metal.

On the other hand, people in cluster 3 do enjoy rock music but do not tend to enjoy Opera and Trance music.

Cluster Solutions for Movies Preferences

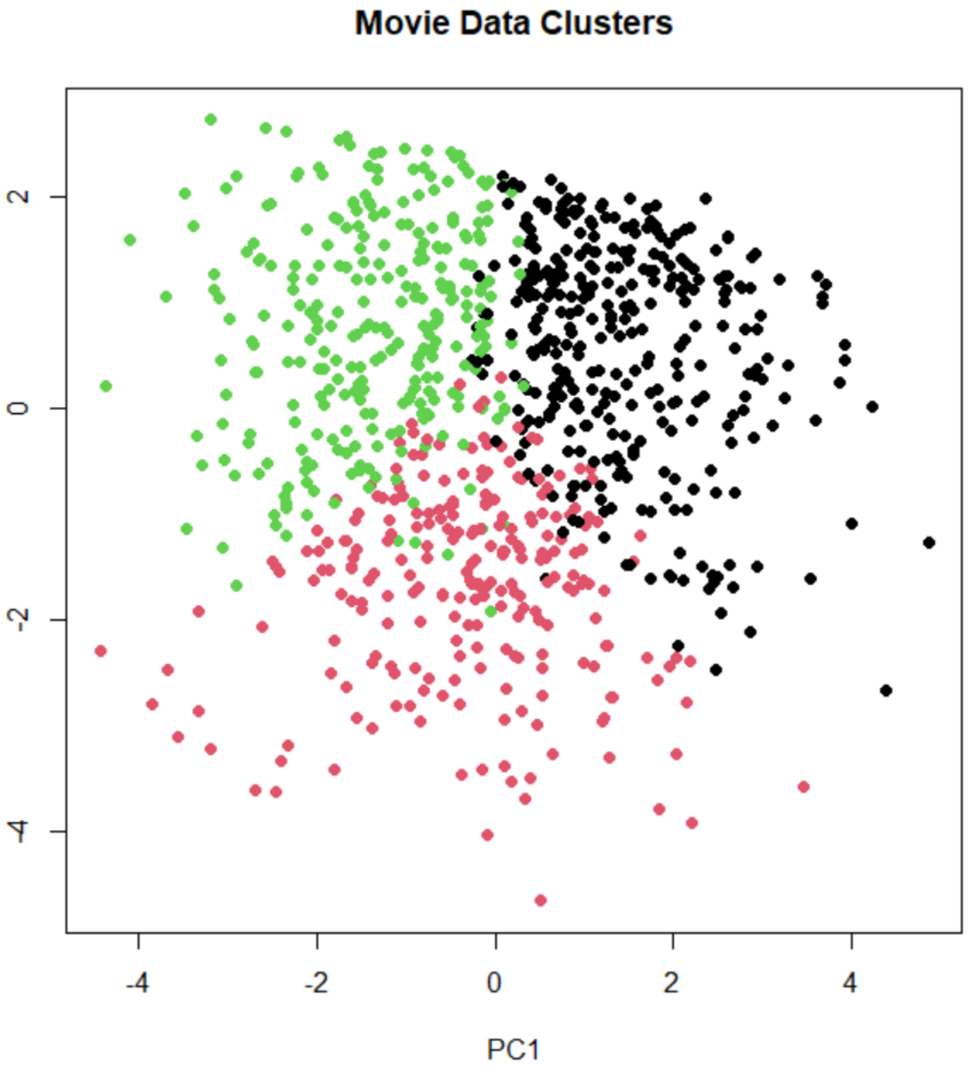


Figure 4: Scatterplot of the movie data clusters components

In figure 4 it shows a scatterplot of 3 clusters. It is interesting to see both music and movies dataset is usually recommended to use 3 clustering for their analysis.

In cluster 1, it contains respondents who enjoy fantasy, fairy tales and comedy movies; on the other hand, people do not seem to enjoy horror movies.

In cluster 2, it contains respondents who enjoy documentary and war movies.

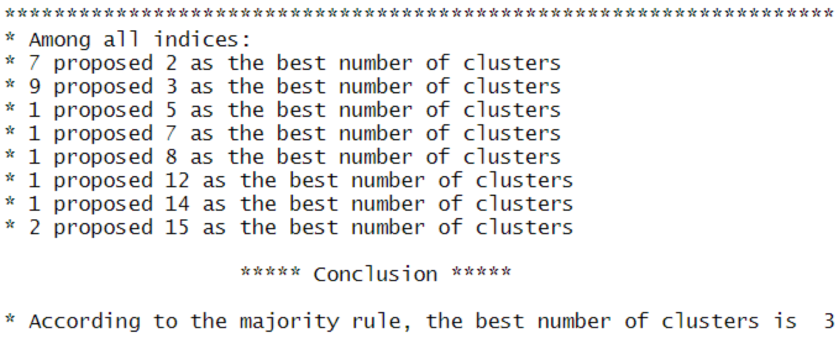
In cluster 3, most contain people enjoy horror movies.

3. 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.



```
*******************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 9 proposed 3 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 2 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```

Figure 5: NbClust output in R

K-means clustering: K-means clustering is a clustering algorithm which its goal is to assign each data point to a single cluster. It requires us to initialise the number of k cluster centres – which is called centroids. There are various ways to determine the optimal value of K. In R, there is a function called 'NbClust' and it will 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 above.

After initialising the number of k cluster centres, the algorithm will decide the cluster membership by assigning to the nearest cluster centroid. This can be done by calculating the distance (Euclidian distance) of each data point. We then re-estimate the k cluster centres, by assuming the memberships are correct.

Hierarchical clustering: Unlike k-means clustering, we do not pre-defined the number of clusters in hierarchical clustering. The hierarchy of clusters is presented as a tree – which is also known as a dendrogram. The root of the dendrogram is the single cluster which contains all the data points while the leaves are the cluster containing one sample each. The dendrogram of the music dataset (created by hierarchical clustering is provided below).

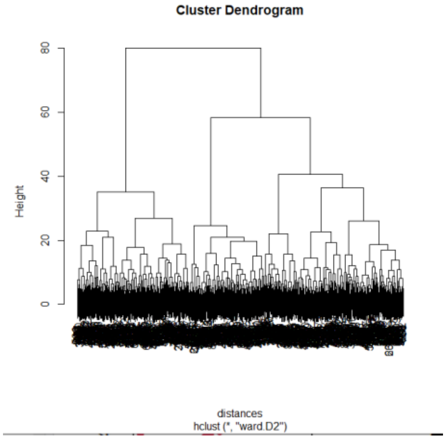


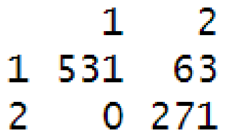
Figure 6: Cluster Dendrogram for Music data

Comparison of clustering solutions for two or more algorithms

A simple cross-tabulation of the clustering results allow us to compare solutions for both k-means and hierarchical clustering.

K-means clustering solution

Hierarchical clustering solution



```
     1   2
1  531  63
2    0 271
```

Figure 7: Cross-Tabulation Overall Data

The results of the cross-tabulation results can be interpreted in the following way:

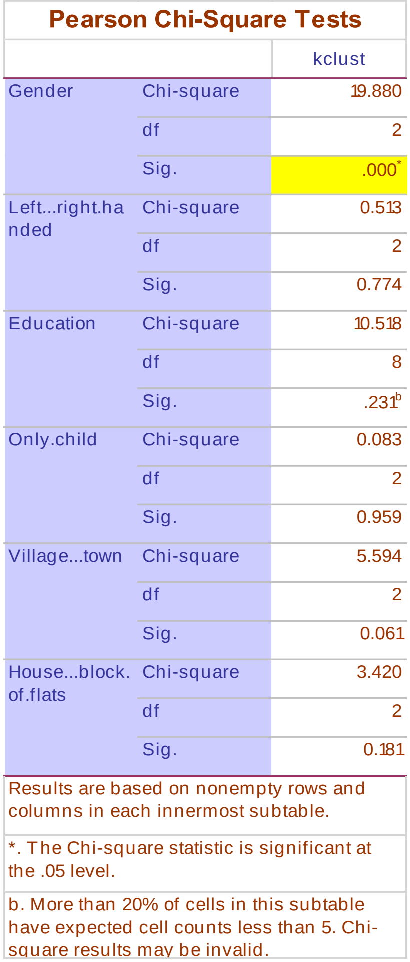
- The number refers to the cluster number (e.g. cluster 1, 2).
- If both cluster clustering algorithm has the same solution then the number will be 11 and 22.
- On the other hand, if both cluster clustering algorithm has a different solution then the number will be 12 and 21.

So it looks like, in this case, both clustering algorithms gave us similar results. Only 63 observations were allocated differently.

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

Table 6: Comparing music and movies preferences based on demographic traits

To compare the profile clusters on the respondent's music and movie preferences (based on demographics), a chi-squared test is conducted to determine whether which demographic trait is significant.



Pearson Chi-Square Tests

		kclust
Gender	Chi-square	19.880
	df	2
	Sig.	.000*
Left...right.handed	Chi-square	0.513
	df	2
	Sig.	0.774
Education	Chi-square	10.518
	df	8
	Sig.	.231b
Only.child	Chi-square	0.083
	df	2
	Sig.	0.959
Village...town	Chi-square	5.594
	df	2
	Sig.	0.061
House...block.of.flats	Chi-square	3.420
	df	2
	Sig.	0.181

Results are based on nonempty rows and columns in each innermost subtable.

*. The Chi-square statistic is significant at the .05 level.

b. More than 20% of cells in this subtable have expected cell counts less than 5. Chi-square results may be invalid.

Figure 8: Chi-Squared Test Results (Music)

Based on the results above, the significant components are highlighted in yellow. It appears that gender is significant (**The Chi-square statistic is significant at the .05 level**.). Bar charts will be created to demonstrate the requirements of the questions.

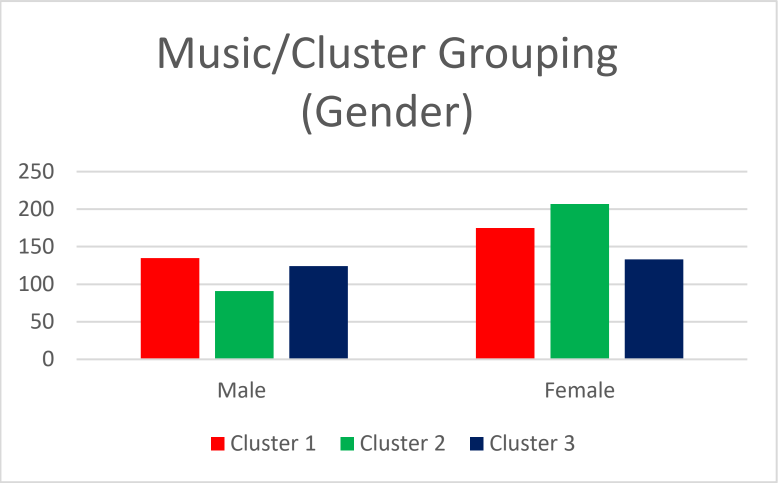
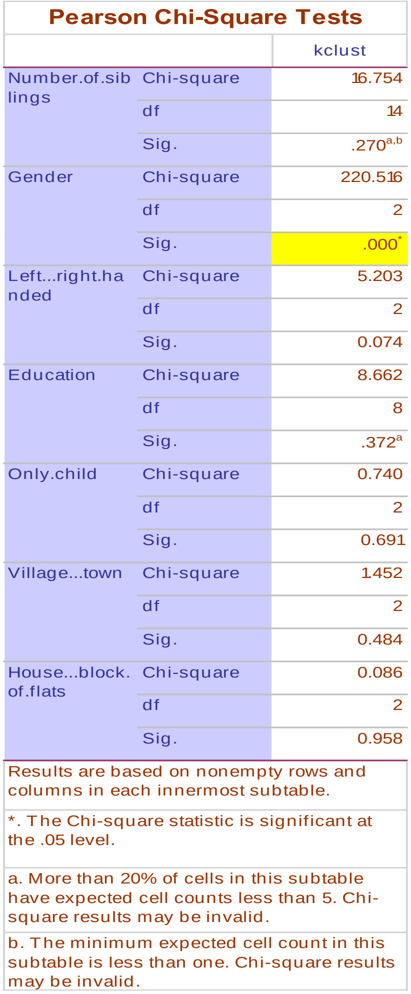


Figure 9: Music/Cluster Grouping based on Gender

Interpretation:

- There are more females over males in cluster 1 but there is the only difference of 40 between two genders.
- There are more females than males in cluster 2.
- There are more females over males in cluster 3 but there is the only difference of 9 between two genders.



Pearson Chi-Square Tests

		kclust
Number.of.siblings	Chi-square	16.754
	df	14
	Sig.	.270[a,b]
Gender	Chi-square	220.516
	df	2
	Sig.	.000*
Left...right.handed	Chi-square	5.203
	df	2
	Sig.	0.074
Education	Chi-square	8.662
	df	8
	Sig.	.372[a]
Only.child	Chi-square	0.740
	df	2
	Sig.	0.691
Village...town	Chi-square	1.452
	df	2
	Sig.	0.484
House...block.of.flats	Chi-square	0.086
	df	2
	Sig.	0.958

Results are based on nonempty rows and columns in each innermost subtable.

*. The Chi-square statistic is significant at the .05 level.

a. More than 20% of cells in this subtable have expected cell counts less than 5. Chi-square results may be invalid.

b. The minimum expected cell count in this subtable is less than one. Chi-square results may be invalid.

Figure 11: Chi-Squared Test Results (Movies)

Interpretation:
Based on the results above, the significant components are highlighted in yellow. It appears that gender is significant (**The Chi-square statistic is significant at the .05 level**.). Bar charts will be created to demonstrate the requirements of the questions.

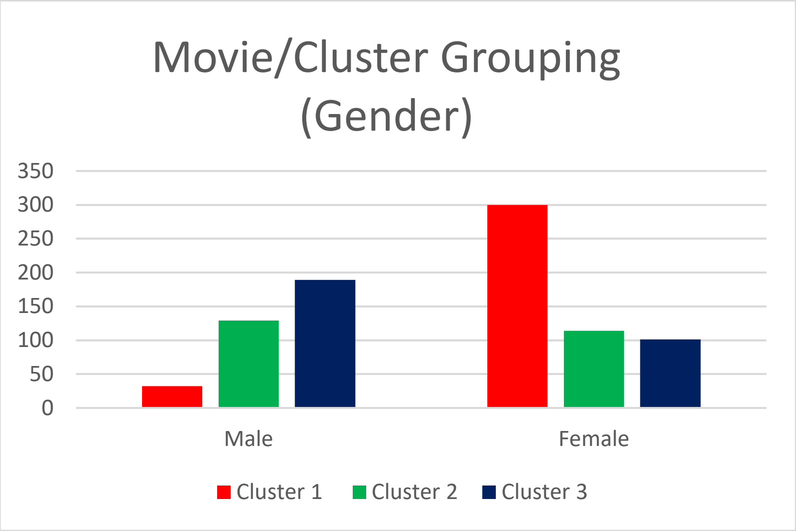


Figure 12: Movie/Cluster Grouping based on Gender

Interpretation:

- There are more females over males in cluster 1.
- There is a fair share of males and females in cluster 2.
- There are more males over females in cluster 3.