**Multivariate Statistical Analysis Final Report**

Group 8

2014150137 박정진

2015150026 임지원

2014100103 정석원

2017150428 김지영

2014130338 김자영

The report will be proceeded by the following order : Data Introduction, PCA and EFA with all variables, PCA and EFA with categories, Clustering with Raw Data, and Clustering with Factor Scores.

1. Data Introduction

The data that we have chosen is called Young People Survey. The survey consists of 150 questions, and explored the preferences of various fields like music, movies, and hobbies, interests, habits, opinions and fears of the young people. It was conducted by the students of the Statistics class at FSEV UK in 2013. All participants were of Slovakian nationality and aged between 15 and 30. The original questionnaire was in Slovakian language and was later translated into English.

The data file (responses.csv) consists of 1010 rows and 150 columns. The number of rows and columns means the number of participants and questions, respectively. For convenience, the original variable names were shortened in the data file. There are 139 integers and 11 categorical variables. The variables can be split into the following groups. Music preferences with 19 questions, Movie preferences with 12 questions, Hobbies and interests with 32 questions, Phobias with 10 questions, Health and habits with 3 questions, Personality traits, views on life and opinions with 57 questions, Spending habits with 7 questions and Demographics of the participants with 10 questions.

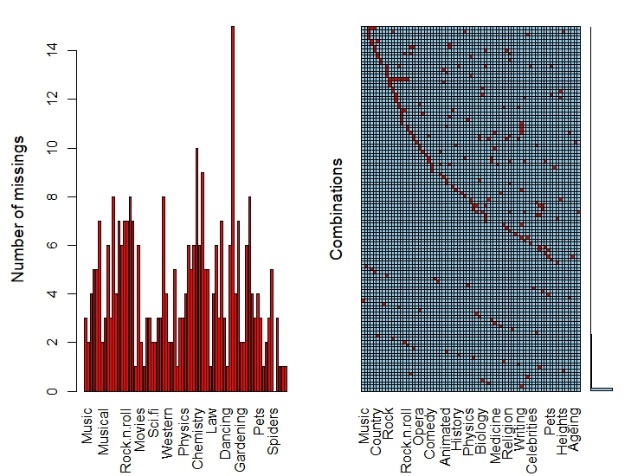
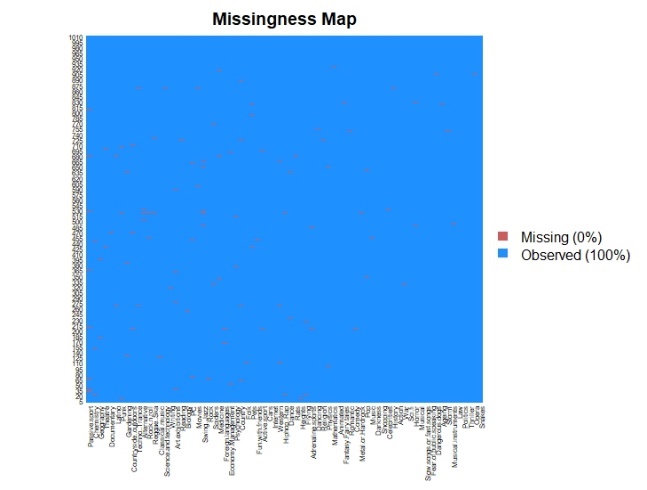
There are two reasons why we chose the data for Multivariate Data Analysis. First of all, there is no scaling issue when we analyzed the data because all the variables are equally measured from one to five. Second, as the data contains huge categories and variables, we believed that the data is appropriate to apply various methodologies like PCA, EFA and Clustering Analysis.

We had two objectives when we started to analyze the data. At first, we wanted to identify what kinds of variables group together. For example, what kind of preferences would a person who likes classical music show in other variables such as movies and hobbies. Next, we wanted to cluster the subjects with similar preferences and explain the groups’ characteristics based on demography.

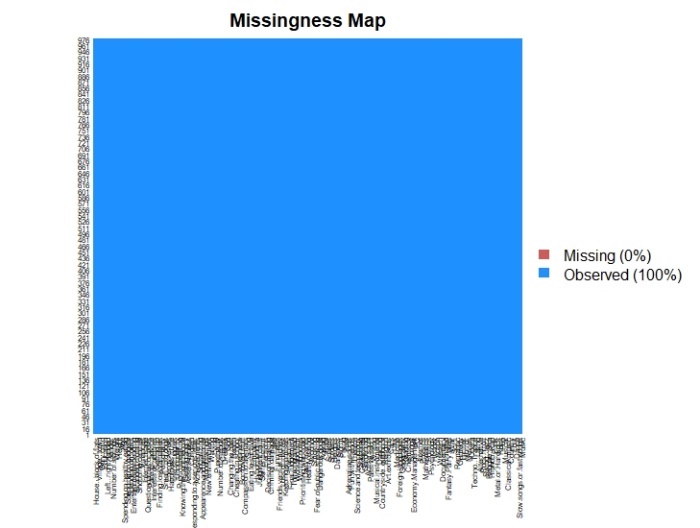
Before analyzing the data, we preprocessed the data. There are missing values in the data. According to the following image, there are about 200 rows containing NA’s out of 1010 rows, which is quite a lot.

https://lh4.googleusercontent.com/y3YsVM1XMCd7KyzckXidVy0o-8GU6ayfMVCEt_vybri5WLXOCfniGrX0HArFra3cjERjCmTiyRPHYBX2HEc5C_DJIBpuDm6W8q5XLDpDpNulrDrAYw5PdeE-HG21ZWY3fg

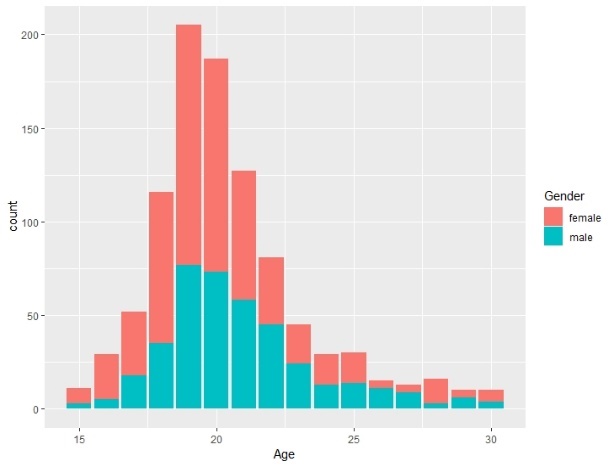
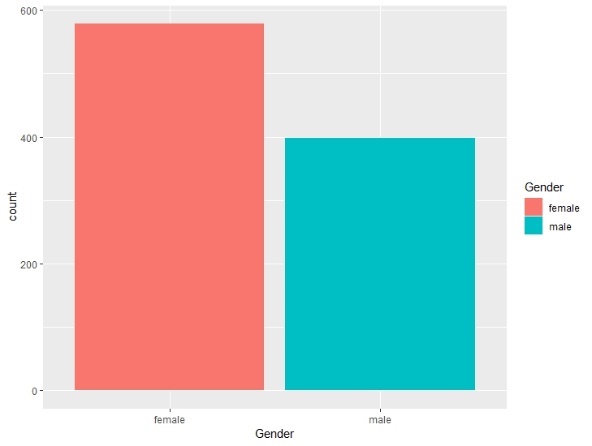
The following images show the missing values in data by graphs and missingness map.

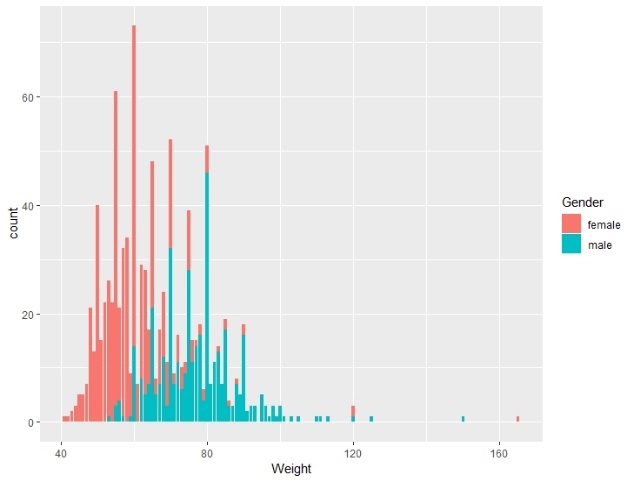
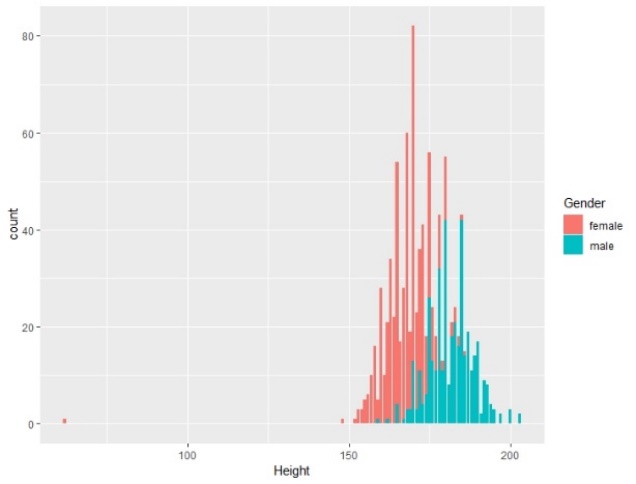
Therefore, we treated missing values using MICE package of R and deleted only 34 rows from the missing values that were in the demographics’ data set. We got the final data set which does not have missing values as the below missingness map. It consists of 976 rows and 150 columns.



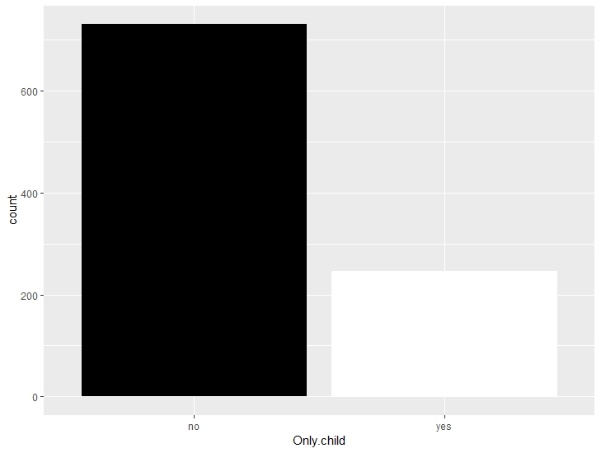
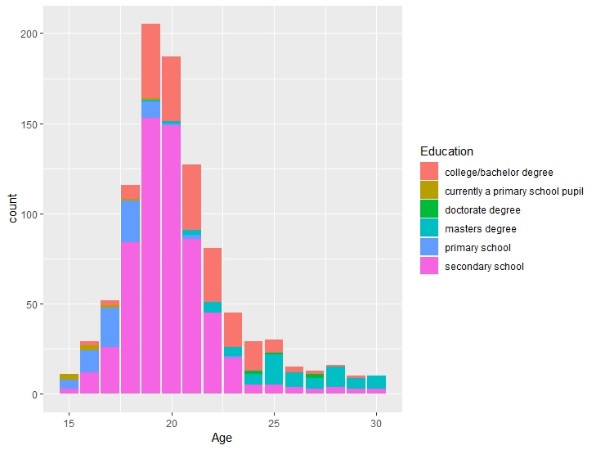
With the preprocessed data, we did Exploratory Data Analysis, called EDA. The data can be divided into two parts, one part consists of 140 subjective questionnaires out of 150 and the other part consists of 10 questionnaires about the demography of the respondents. At first, we overlooked the demographic features of the data. We could find that the data is well distributed regarding gender and age according to the below graphs.

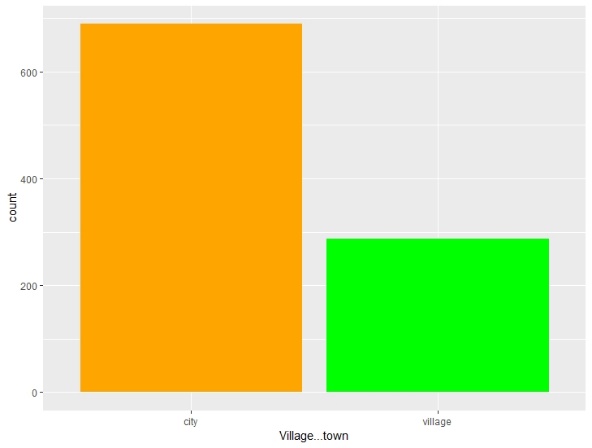


Height and weight appear to be highly correlated with gender as the below graphs show.

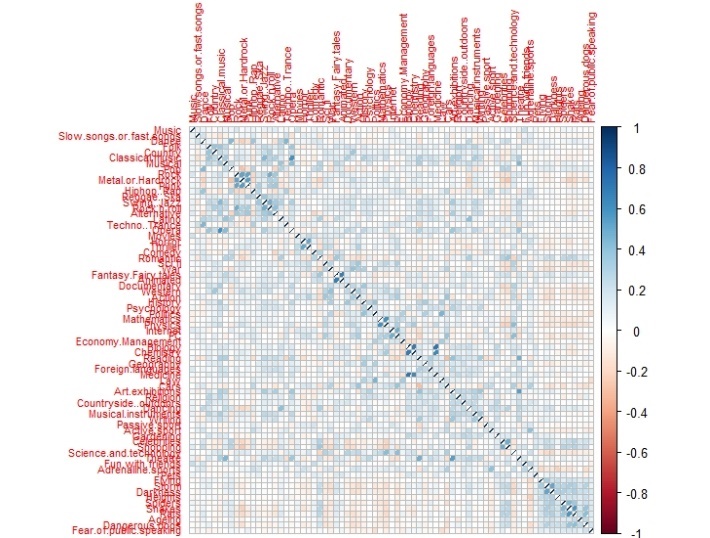


The following graphs show that the older subjects showed higher education level, which all seem to be reasonable. The most subjects resided in urban areas and had no siblings.

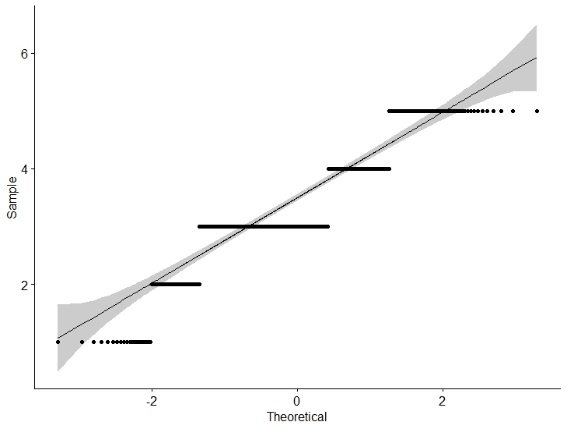
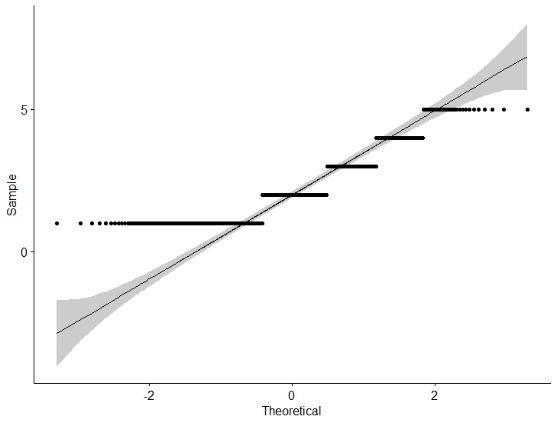
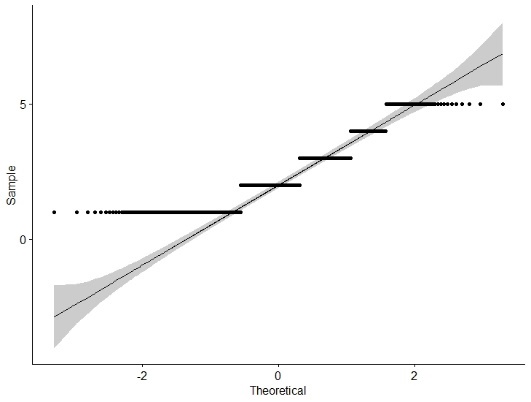
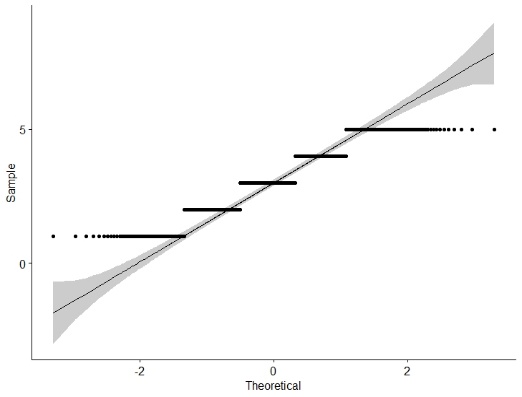




Next, as we can see in the below image, the data showed low correlations between variables.

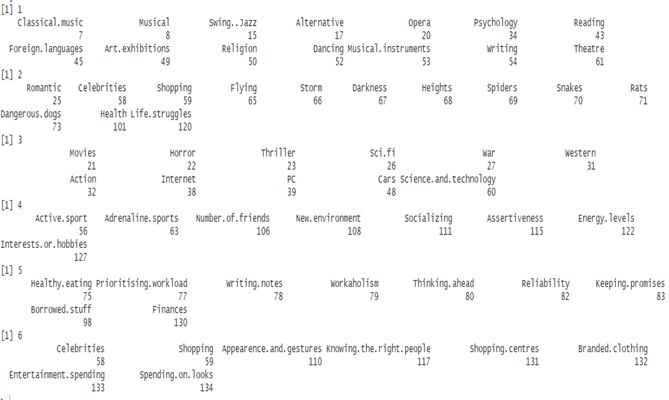


In addition, it is hard to say that the distribution of data is normal so we knew that it is hard to use Linear Discriminant Analysis and Quadratic Discriminant Analysis according to the below images which show the differences between theoretical and actual data.



2. PCA, EFA with all variables

First, we analyzed the whole data set without demography by using PCA and EFA methods. We found couple of problems as we can find by the below image which shows the EFA result of whole data. There is not much part of the data set which are accountable with the factors derived from the EFA method. In addition, we could not explain each factor properly because each factor did not contain variables from a variety of categories.

 We believed that this result is due to the characteristics of each category. As each category consists of related variables, we decided to apply PCA and EFA to categories respectively. Then, we grouped the similar variables of each category and analyzed the data.

3. PCA, EFA in each category

First, we did the PCA in each category. We decided to choose the number of principal components that can explain more than 50% of the variables because there are too many variables and it is not meaningful to analyze data if we contain all the variables of which the variation is over one. Next, we did the EFA. We chose the number of factors in EFA as same as the number of the principal components. As a result, we could reduce 140 variables to 35 variables. For example, the original data set of music category consists of 19 items including classical music, alternative, opera, swing jazz, rock, punk, and etc. We could reduce 19 items into 4 factors. The below table shows the result of EFA.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Music | Movies | Hobbies | Phobia | Health | Personality | Spending |
| Quiet  Rock  Noisy  Pop | Fantasy Action Horror | Liberal arts Science Society  IT  Gossip Sports Math | Wild Animal  Dog | Smoking  Alcohol  Healthy eating | Social  Lonely  Work  Kind Reliability  HR  Angry  Decisive & Religious Hypochondria  Family  Self-criticism  Charity  Judgement | Spending on  Spending at |

In the Music category with 4 factors, the factor ‘Quite’ consists of Folk, Country, Classical music, Musical, Swing Jazz, Latino, and Opera variables. The factor ‘Rock’ consists of Rock, Metal or Hardrock, Punk, Rock n roll, and Alternative music variables. ‘Noisy’ factor consists of Dance, Hip hop or Rap, Reggae ska, and Techno Trance music variables. Last ‘Pop’ factor consists of Pop Music variable.

In the Movie category with 3 factors, ‘Fantasy’ factor contains Animation, Fantasy, Fairy tales, and Romantic movies variables. ‘Action’ factor contains the science fiction, War, Documentary, and Action movies variables. ‘Horror’ factor contains Horror and Thriller movies variables.

In the Hobbies category with 7 factors, ‘Liberal arts’ factor consists of Art exhibitions, Writing, Reading, Musical instruments, and Theatre variables. ‘Science’ factor consists of Biology, Chemistry, and Medicine variables. ‘Society’ factor consists of Politics, Law, History, and Economy Management variables. ‘IT’ factor consists of PC, Science and Technology variables. ‘Gossip’ factor consists of Celebrities and Shopping variables. ‘Sports’ factor consists of Cars, Passive sport, Active sports, and Adrenaline sports variables. ‘Math’ factor consists of Mathematics and Physics variables.

In Phobias category with 3 factors, ‘Wild’ contains Storm, Darkness, and Heights variables. ‘Animal’ factor contains Spiders, Snakes, and Rats variables. ‘Dog’ factor contains Dangerous dogs variables. In Health category, there are three factors ‘Smoking’, ‘Alcohol’ and ‘Healthy eating’.

In Personality category with 13 factors, ‘Social’ factor consists of number of friends, new environment, socializing, energy levels, and interests variables. ‘Lonely’ factor contains loneliness, and changing the past variables. ‘Work’ factor consists of prioritising workload, writing notes, and workaholism variables. ‘Kind’ factor consists of giving and empathy variables. ‘Reliability’ factor consists of thinking ahead, reliability, keeping promises, and borrowed stuff variables. ‘HR’ factor consist of knowing the right people variable. ‘Angry’ factor consists of mood swings and getting angry variables. ‘Decisive & religious’ factor contains final judgement and god variables. ‘Hypochondria’ consists of hypochondria variable. ‘Family’ factor consists of children, life struggles, and parents advice variable. ‘Self criticism’ factor consists of self criticism variable. Charity consists of charity variable. ‘Judgement’ factor consists of judgment calls variable.

In Spending Category with two factors, ‘Spend on’ factor contains branded clothing, entertainment, looks, gadgets, and healthy eating variables. ‘Spend at’ factor contains shopping centers variable.

We could find that each factor has its own unique characteristic so we assigned the reasonable names to each factor based on the factor loadings. However, there was problem that the uniqueness of each factor was too large so it was hard to think that the result of EFA is significant. To solve the problem, we decided to compare the result of clustering analysis with raw data in each category to the result of EFA in each category.

4. Cluster with Raw Data

We conducted the Clustering Analysis in each category to compare to EFA. We chose the number of clusters in each category as same as the number of factors in EFA because we wanted to confirm that the result of EFA is significant even if its uniqueness is too large. As a result, there were not many differences between the clusters from Clustering Analysis and the factors from EFA but they had a lot of similar features. Therefore, we think the result of EFA is significant .

Next, we conducted the Clustering Analysis using the whole data. We used the two clusters resulting from NBCluster function of R. After clustering, we applied each cluster to demographics, we found that the clusters were mainly grouped according to the gender. The below table shows the gender distribution in the clusters. According to the table, we thought that first cluster consists of most women and the second cluster consists of more men than women.

|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| 1 | 491 | 51 |
| 2 | 81 | 347 |

However, we had a same problem when we cluster the raw data as we saw when we conducted PCA and EFA. Therefore, we decided to transform the raw data to the factor score data and conduct Clustring Analysis using the factor scores in the data.

5. Cluster with Factor Score

When we conduct the Clustering Analysis with factor score data, we also applied K-means clustering with the two clusters in order to match the consistency with the raw data. As a result, we came up with two clusters as follows. The below table shows the factors in the Cluster 1. Cluster 1 has a tendency to prefer quiet, pop music, and fantasy movies. They have a personality of workaholic, kind, family friendly, and have a phobia of wild environment and animals.

|  |  |
| --- | --- |
| Categories | Factors |
| Music | Quiet, Pop |
| Movies | Fantasy |
| Hobbies | Liberal arts, Science, Gossip |
| Phobia | Wild, Animal |
| Health | Healthy lifestyle |
| Personality | Work, Kind, Angry, Decisive and Religious, Family |
| Spend | Spend at |

The below table shows the factors in Cluster 2. Cluster 2 prefers rock music, action and horror movies, and they have a hobby about IT and sports related fields.

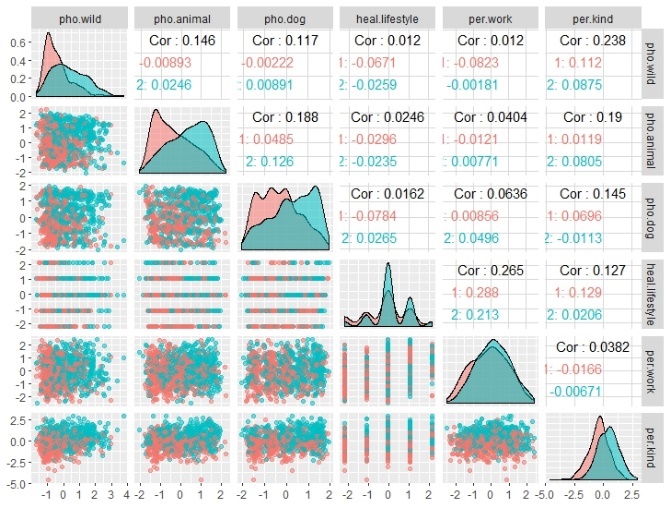
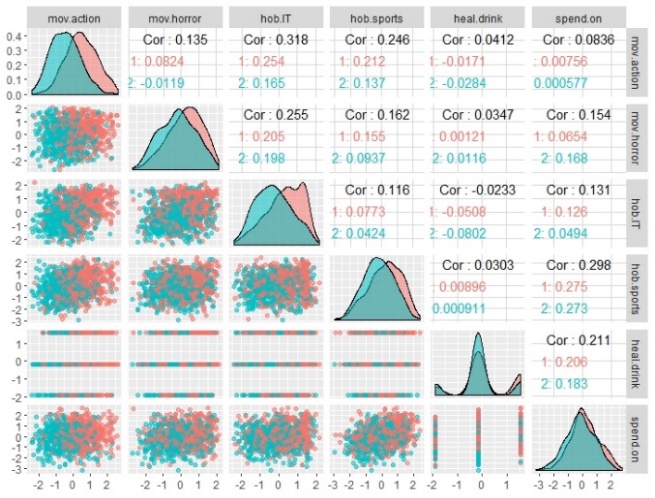
|  |  |
| --- | --- |
| Categories | Factors |
| Music | Rock |
| Movies | Action, Horror |
| Hobbies | Society, IT, Sports |
| Phobia |  |
| Health | Drink |
| Personality |  |
| Spend | Spending on |

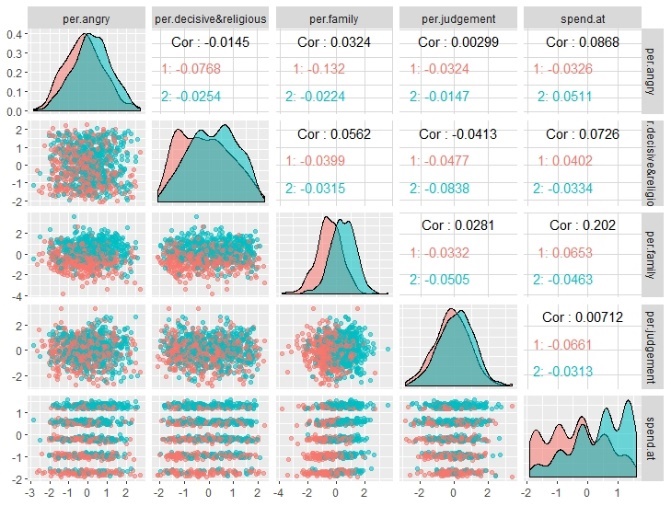
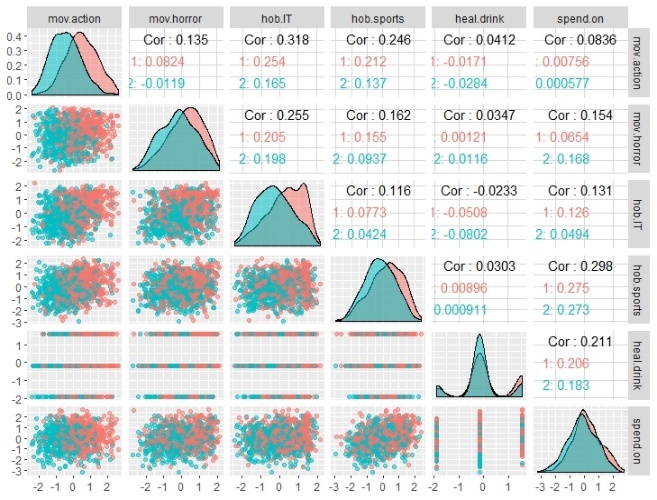
When we apply these clusters to demographic features, there was a big difference regarding gender. The below table shows the gender distribution in the two clusters. Cluster 1 significantly consists of female subjects, and cluster 2 has higher proportion of male. When we think of social conventions, we might be able to say that the cluster is distinguished accordingly.

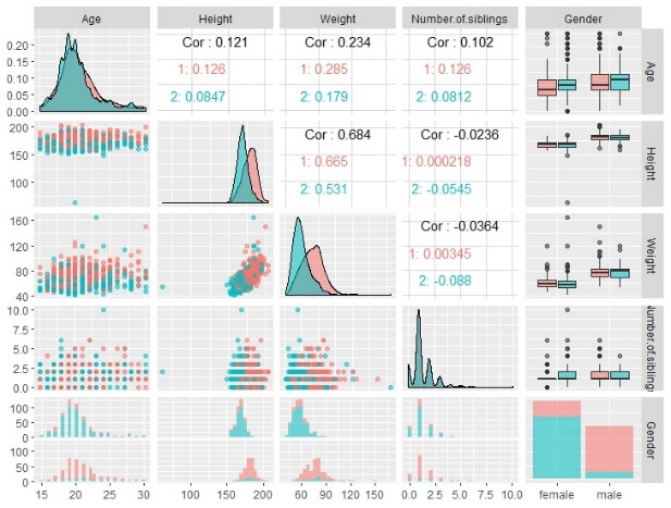
|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| 1 | 466 | 51 |
| 2 | 112 | 347 |

Compared the gender distributions of clusters from the raw data and from the factor score data, both appear to be similar. In addition, when we conduct the confusion matrix, we can find that the accuracy is about 89%.

There are visualization plots of the two clusters from the factor score data as follows. According to the visualization plots, we can find that there are distinctions of clusters among variables.

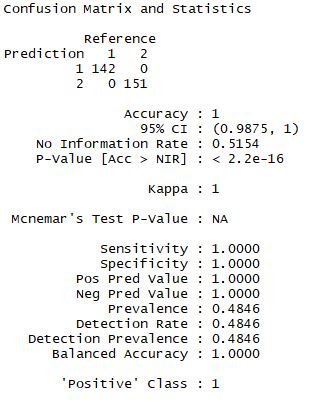
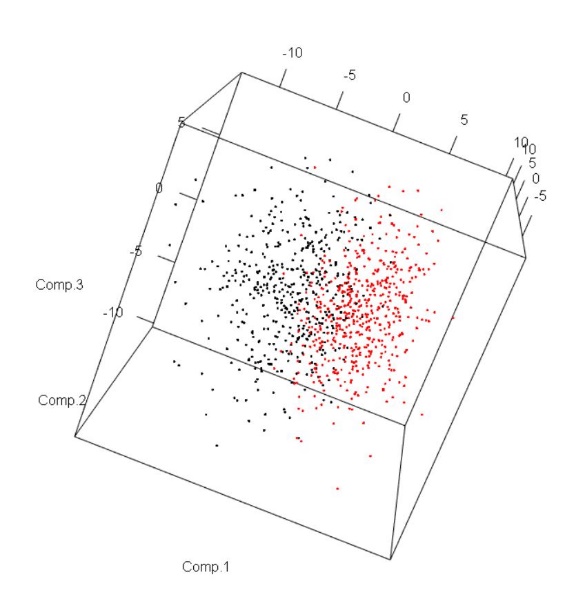
****

****

****

To summarize each cluster, cluster 1 prefers quiet music, fantasy movies, enjoy recreational life, fear of natural disasters, animals, maintain healthy life style, high level of empathy, and tend to be moody. Also, care a lot about families, and resolute in decision making. On the other hand, cluster 2 prefers rock music, action and horror movies, interests in politics, law, history, economics, IT, sports, enjoy drinking, entertainment, spend a lot on appearances, and tends to be active and extroverted.

Finally, we validate the clustering method following the steps for Cluster Validation. We divided the data set into training data and test data at proportion of 7:3. Accodring to the following confusion matrix, the accuracy appears to be 100%. The cluster seemed to be valid when we saw the clustering plots and the confusion matrix.

****

6. Conclusion

In conclusion, we analyzed the data using Principal Component Analysis, Exploratory Factor Analysis, and Clustering methods. We were able to reduce the variables from 140 to 35, and then using the factor scores, we could divide the whole data into 2 clusters. When we applied the clusters into demography, there were no significant differences in terms of age, number of siblings, left-right handed, education and etc, but only in gender.

For the further research, as the data is based on the youths in Slovakia, we thought that it would be interesting to collect similar data from Korean young people and compare the results across countries. In addition, as the clusters were mainly grouped based on gender, we could develop the research with searching for other variables that can affect other factors. Also, it would be useful to conduct this survey with people who are in 40s and apply to other fields like marketing because the people who are in 40s have high purchasing power.