

ADVANCED STATISTICS

Principal Component Analysis



A picture containing device

Description automatically generated[[1]](#footnote-1)

TERESA ventaja

Higher Diploma in Science in Data Analytics

INDEX

[1. Introduction 2](#_Toc48742466)

[2. Background 2](#_Toc48742467)

[3. Data Description 2](#_Toc48742468)

[4. Methodology & Calculations 3](#_Toc48742469)

[a. Exploring correlations 3](#_Toc48742470)

[b. Descriptive statistics 3](#_Toc48742471)

[c. Selection of rotation 3](#_Toc48742472)

[d. KMO & Bartlett’s Test 4](#_Toc48742473)

[5. Results 4](#_Toc48742474)

[a. Selecting number of components 4](#_Toc48742475)

[i. Total variance explained 4](#_Toc48742476)

[ii. Scree Plot 5](#_Toc48742477)

[b. Component matrix 5](#_Toc48742478)

[c. Commonalities 6](#_Toc48742479)

[d. Component plot 8](#_Toc48742480)

[6. Discussion & Analysis of Results 8](#_Toc48742481)

[Appendix 1 – Dataset Description 10](#_Toc48742482)

[Appendix 2 – Correlation matrix (partial) 13](#_Toc48742483)

[Appendix 3 – Oblique rotation results 14](#_Toc48742484)

[Appendix 4 – Varimax rotation results 15](#_Toc48742485)

[Appendix 5 – Quartimax rotation results 20](#_Toc48742486)

[Appendix 6 – Descriptive Statistics 26](#_Toc48742487)

[Appendix 7 – Total Variance Explained 27](#_Toc48742488)

[Appendix 8 – Rotated Component Matrix 28](#_Toc48742489)

[Bibliography 29](#_Toc48742490)

# Introduction

This project analyses a psychological survey, covering multiple aspects of young people’s interests, opinions, concerns, preferences, etc. The survey contains 150 variables, and I am going to perform a Factor Analysis (Principal Component Analysis - PCA) with the main objective of identifying if there are relevant clusters. If so, I will name the factors and assign variables to each of them.

# Background

In most cases, when used for psychology research, the interpretation of the results is not obvious. The pattern of correlations among them is very complex, with many medium-sized correlations and fewer correlations that are very large or very small. When this is the case, factor analysis can help the researcher, by taking a complicated pattern of correlations and reducing those variables to a small number of factors[[2]](#footnote-2). The main advantage of using PCA is that it reduces the dimensionality, but it preserves as much “*variability*” as possible, therefore, we keep the most relevant statistical information. The new factors found maximize the variance and are uncorrelated with each other[[3]](#footnote-3).

The dataset selected, “*Young People Survey*”[[4]](#footnote-4), explores the preferences, interests, habits, opinions, and fears of young people in Slovakia 2013. The participants’ age ranges from 15 to 30, they were all Slovakian nationals, and the survey was presented in both electronic and written form, in Slovak language originally, which was later translated into English.

# Data Description

Students enrolled in the Statistics class in Comenius University in Bratislava (Faculty of Social and Economic Sciences) invited their friends to participate in the survey. There is a file containing the description of the columns data, and the dataset itself:

* SAMPLE SIZE – 1010 observations
* # OF VARIABLES (in the original file) - 150 (139 integer and 11 categorical)

Since I am interested in “*Personality Traits, Views on Life & Opinions*”, I eliminated the columns related with other topics. Of the remaining 57 columns, I eliminated 3 of them because they did not match the categorical distribution 1 to 5 contained on the other columns, and that would have interfered negatively in the performance of calculations. I also eliminated “*I enjoy taking part in surveys*” column because I am not interested in survey participation behaviour. These are the columns I eliminated:

|  |
| --- |
| How much time do you spend online? |
| Timekeeping |
| Do you lie to others? |
| I enjoy taking part in surveys. |

I finally kept 53 variables, that are further described in ***Appendix 1***. As we can see, 184 missing values were found. The overall percentage of missing values per variable is no higher than 0.70% (less than 1%). I then decided to impute the median value into those (value 3) to perform the Principal Component Analysis. Being less than 1%, the ability of those values to change the overall result is extremely low, therefore in this case a regression model was not necessary for the imputation of values.

# Methodology & Calculations

The method used in this project is an Exploratory Factor Analysis (EFA). The decision on the number of factors used will be taken by examining the output of the PCA. It allows all items to load on every factor. The calculations will be performed in SPSS, using only the data included in the selected 53 variables, and processed so that there are 0 missing values.

## Exploring correlations

The first thing to check if a PCA is appropriate is to make sure that the variables are correlated. I created a bivariate correlation matrix to explore this, and I can see that the majority of variables are correlated to each other (see ***Appendix 2***), most of them with a significance level of 0.01% (2 tailed Pearson correlation). I can conclude that the data is suitable to PCA, because we aim to explain the correlation between the variables through a smaller number of components.

## Descriptive statistics

Mean ranges from 1.92 to 4.02, but it stays between 2-3 in most variables. In terms of data distribution, the standard deviation ranges from 0.7 to 1.5, so I can see that the data is approximately uniformly spread. The sample size is 1010, big enough to perform PCA, since we need more than 300 to reach a significant result (see ***Appendix 6***).

## Selection of rotation

Before performing the analysis itself, I am going to select a rotation. The reason why I do it in that order is because SPSS allow the selection of multiple analytical features that I will use to explore the data, including the PCA. But it requires the selection of a rotation, which is a mathematical procedure that rotates the factor axis in other to produce results that are more interpretable. It makes the loading patterns more clear, easier to identify and more pronounced. The purpose is to create a simple structure which will be easier to interpret. The two main rotation techniques are oblique and orthogonal. Orthogonal rotation is suitable when factors are independent, while oblique is more suitable for dependent factors. Following Brown, J. D. (2009), since I do not know about the dependency of variables, to select the optimal one, I am going to perform calculations and check which one performs the best.

I first checked Oblimin rotation in SPSS (Oblique) and found no significant values in the component correlation matrix, i. e. no values over 0.32 (see ***Appendix 3***). Therefore, I discarded the use of this method. Then I explored 2 orthogonal techniques, Varimax and Quartimax. To select which one performed the better, I found the significant loadings in the Rotated Component Matrix, i. e. those greater than 0.3, and selected the one that had fewer complex variables (variables with significant loadings in more than 1 factor). Since Varimax had 11 and Quartimax 13, I selected Varimax for my analysis (see ***Appendix 4*** and ***Appendix 5***).

## KMO & Bartlett’s Test

|  |  |  |
| --- | --- | --- |
| **KMO and Bartlett's Test** | | |
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .784 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 9397.323 |
| df | 1378 |
| Sig. | .000 |

*Figure 1: KMO & Bartlett’s Test*

Following Cerny & Kaiser (1977) and Kaiser (1974), this tests whether all the variables are significantly correlated than 0, but not in pairs but considering all the variables that are included in the correlation matrix we saw above. Since the P-value is lower than 0.05, we can conclude that there is a strong correlation. Again, this is an indicator that performing a PCA is suitable for this data.

# Results

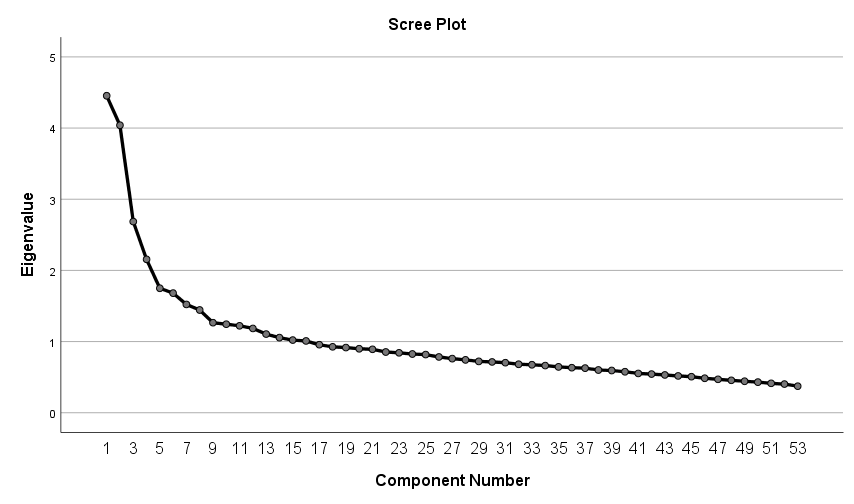
## Selecting number of components

In this section, I am going to analyse what is the best number of factors that I should create. I will compare techniques “*Total variance explained*” (TVE) and “*Scree Plot*” to make the most adequate decision. Since I have a large number of variables (53), the likelihood that both methods will agree is low, but I will still have relevant information to make my decision.

### Total variance explained

The rule that I set for selecting the eigenvalues in SPSS is keeping those that are greater than 1. In this case, 16 values meet that criteria (see ***Appendix 7***). These values represent a proportion of the total variance, and the other column the percentage of variance explained by the component. There is controversy over what percentage of variance explained by the component should we consider relevant to retain a variable. Maskey, R. Fei, J. and Nguyen, H. O. (2018) included in their research paper a table where we can compare different threshold used in a variety of papers. We can see that the criteria used in each of them varies enormously, but most would discard those that explain less than 40% of the variance. Since we have a large dataset with lots of variables, we are going to assume 40% as a good threshold in this case. In this case, SPSS takes as many factors as components have eigenvalues greater than 1, which is 16 (accounting for 54.41% of the variance).

### Scree Plot



*Figure 2: Scree Plot*

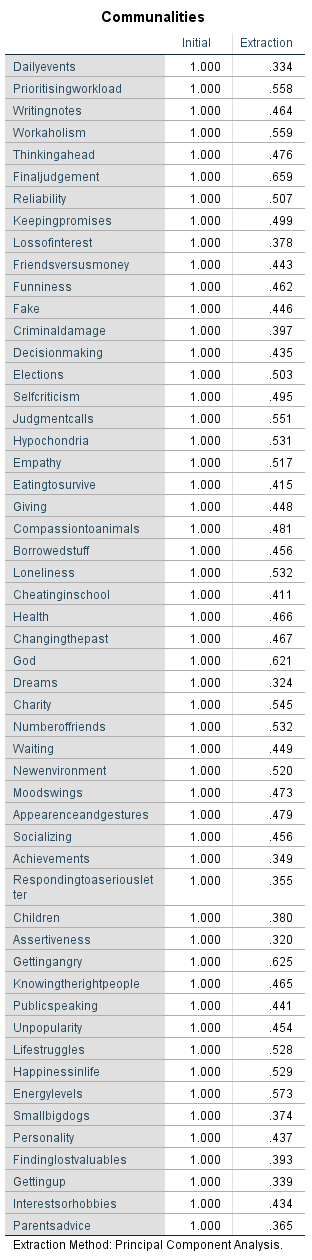
Following Hubert, M. (2009), the rule for retaining components in this case is keeping the number of components showing when the line tends to taper off gradually. We should keep the number of components showing in the graph before it drops off. In this case is 8, but the percentage of variance accounted for with 8 values is too low to make this selection (only 37%). There is another drop in value 12 to 13, with 12 accounting for 46.5% of the variance. Therefore, instead of keeping the defaulted 16 eigenvalues, I am going to select 12 components.

## Component matrix

The component matrix estimates the correlations between each variable and the estimated components before rotation[[5]](#footnote-5). But we will have a better estimation if we analyse the data after rotation. Therefore, since the number of factors in this case is greater than 1, I am going to focus only on the “*Rotated component matrix*” (see ***Appendix 8***). I will discuss the results in detail in the following chapter.

## Commonalities

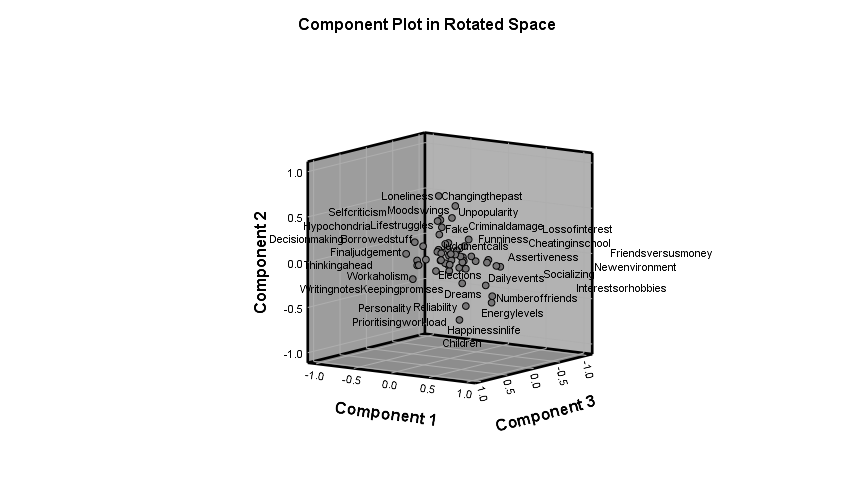
This table represents the amount of variance accounted for in every variable for all the factors that we retain[[6]](#footnote-6). The difference with the “*Total variance explained*” is that instead of checking the amount of variance that each factor accounts for, in this case we check the amount of variance that each variable account for considering all the components we retain. It is the sum of the squared component loadings. We can see that all the values are over 0.32, so the selection of 12 components is strong enough.



*Figure 3: Communalities*

## Component plot

Since we have 12 factors, we need to limit the plot to the 3 first components. We can still compare and see the special location of each variable, however, since we have 53 variables, extracting relevant information is complicated because the plot is too stacked.



*Figure 4: Component Plot in Rotated Space*

# Discussion & Analysis of Results

As discussed in the previous chapter, ***Appendix 8*** contains the results on the analysis. Any variable correlating with the component with 0.3 or higher can be considered as part of the factor. The only exception is ***Parents Advice***, which correlates with component 9 the highest (close to 0.3) so in practice I am going to add it to this factor. However, it should be noticed that this variable is the one that correlates the least with the others, which means **its relevance in the overall topics of this psychological survey is lower**. The identified clusters are presented in the following diagram:

A screenshot of a map

Description automatically generated

*Figure 5: Factor Diagram*

The negative correlations are presented in colour blue, and the names that I selected for each factor in pink. As I mentioned, since Parents Advice is a special category (lowest correlation), I highlighted it in yellow. In terms of interpretation, cluster “*Sociability & Energy*” correlates negatively with public speaking and decision making. The reason why it makes sense is because the description of the variable talks about requiring preparation for it. We can conclude that sociable/energetic people need no preparation for public speaking and decision making, therefore it is a reasonable result.

In “*Positivity*”, variables related with a pessimistic viewpoint correlate inversely with happiness and satisfaction with personality. Similarly, “*Time Management*” is consistent, since highly skilled people would rank low in cheating in school (no need to cheat if you manage your time effectively) and difficulty to get up early. In “*Trust*”, it is also reasonable to assume that distrustful people would worry more about illnesses, so would have a higher score for hypochondria. As per “*Emotional Stability*”, unstable people tend to be inpatient, so the result is reasonable.

Lastly, “*Empathy*” component contains a variable that to me does not make sense. Finding lost values means handing in a missing item to the original owner, which would make sense to correlate positively with the other variables, not negatively as it is. I would recommend checking in the original source if it is either a translation issue (the survey description was translated from Slovakian to English), or an unmeaningful result.

To sum up, 12 relevant components have been identified, with 2 minor inconsistencies in “Finding lost values” and “Parents advice”. 2 out of 53 variables do not impede further psychological research, but they can be excluded instead. Therefore, the main objective of this project has been successfully met and factors have been given a name, indicating what they represent.

# Appendix 1 – Dataset Description

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column name | Column description | Data type | Range | % missing data | empty cells |
| Daily events | I take notice of what goes on around me. | numeric - ordinal - discrete | 5 categories | 0.69% | 7 |
| Prioritising workload | I try to do tasks as soon as possible and not leave them until last minute. | numeric - ordinal - discrete | 5 categories | 0.49% | 5 |
| Writing notes | I always make a list so I don't forget anything. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| Workaholism | I often study or work even in my spare time. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| Thinking ahead | I look at things from all different angles before I go ahead. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| Final judgement | I believe that bad people will suffer one day and good people will be rewarded. | numeric - ordinal - discrete | 5 categories | 0.69% | 7 |
| Reliability | I am reliable at work and always complete all tasks given to me. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Keeping promises | I always keep my promises. | numeric - ordinal - discrete | 5 categories | 0.10% | 1 |
| Loss of interest | I can fall for someone very quickly and then completely lose interest. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Friends versus money | I would rather have lots of friends than lots of money. | numeric - ordinal - discrete | 5 categories | 0.59% | 6 |
| Funniness | I always try to be the funniest one. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Fake | I can be two faced sometimes. | numeric - ordinal - discrete | 5 categories | 0.10% | 1 |
| Criminal damage | I damaged things in the past when angry. | numeric - ordinal - discrete | 5 categories | 0.69% | 7 |
| Decision making | I take my time to make decisions. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Elections | I always try to vote in elections. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| Self-criticism | I often think about and regret the decisions I make. | numeric - ordinal - discrete | 5 categories | 0.49% | 5 |
| Judgment calls | I can tell if people listen to me or not when I talk to them. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Hypochondria | I am a hypochondriac. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Empathy | I am emphatetic person. | numeric - ordinal - discrete | 5 categories | 0.49% | 5 |
| Eating to survive | I eat because I have to. I don't enjoy food and eat as fast as I can. | numeric - ordinal - discrete | 5 categories | 0% | 0 |
| Giving | I try to give as much as I can to other people at Christmas. | numeric - ordinal - discrete | 5 categories | 0.59% | 6 |
| Compassion to animals | I don't like seeing animals suffering. | numeric - ordinal - discrete | 5 categories | 0.69% | 7 |
| Borrowed stuff | I look after things I have borrowed from others. | numeric - ordinal - discrete | 5 categories | 0.20% | 2 |
| Loneliness | I feel lonely in life. | numeric - ordinal - discrete | 5 categories | 0.10% | 1 |
| Cheating in school | I used to cheat at school. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Health | I worry about my health. | numeric - ordinal - discrete | 5 categories | 0.10% | 1 |
| Changing the past | I wish I could change the past because of the things I have done. | numeric - ordinal - discrete | 5 categories | 0.20% | 2 |
| God | I believe in God. | numeric - ordinal - discrete | 5 categories | 0.20% | 2 |
| Dreams | I always have good dreams. | numeric - ordinal - discrete | 5 categories | 0% | 0 |
| Charity | I always give to charity. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| Number of friends | I have lots of friends. | numeric - ordinal - discrete | 5 categories | 0% | 0 |
| Waiting | I am very patient. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| New environment | I can quickly adapt to a new environment. | numeric - ordinal - discrete | 5 categories | 0.20% | 2 |
| Mood swings | My moods change quickly. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Appearence and gestures | I am well mannered and I look after my appearance. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| Socializing | I enjoy meeting new people. | numeric - ordinal - discrete | 5 categories | 0.49% | 5 |
| Achievements | I always let other people know about my achievements. | numeric - ordinal - discrete | 5 categories | 0.20% | 2 |
| Responding to a serious letter | I think carefully before answering any important letters. | numeric - ordinal - discrete | 5 categories | 0.59% | 6 |
| Children | I enjoy childrens' company. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Assertiveness | I am not afraid to give my opinion if I feel strongly about something. | numeric - ordinal - discrete | 5 categories | 0.20% | 2 |
| Getting angry | I can get angry very easily. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Knowing the right people | I always make sure I connect with the right people. | numeric - ordinal - discrete | 5 categories | 0.20% | 2 |
| Public speaking | I have to be well prepared before public speaking. | numeric - ordinal - discrete | 5 categories | 0.20% | 2 |
| Unpopularity | I will find a fault in myself if people don't like me. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| Life struggles | I cry when I feel down or things don't go the right way. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| Happiness in life | I am 100% happy with my life. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Energy levels | I am always full of life and energy. | numeric - ordinal - discrete | 5 categories | 0.49% | 5 |
| Small - big dogs | I prefer big dangerous dogs to smaller, calmer dogs. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Personality | I believe all my personality traits are positive. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Finding lost valuables | If I find something the doesn't belong to me I will hand it in. | numeric - ordinal - discrete | 5 categories | 0.40% | 4 |
| Getting up | I find it very difficult to get up in the morning. | numeric - ordinal - discrete | 5 categories | 0.49% | 5 |
| Interests or hobbies | I have many different hobbies and interests. | numeric - ordinal - discrete | 5 categories | 0.30% | 3 |
| Parents' advice | I always listen to my parents' advice. | numeric - ordinal - discrete | 5 categories | 0.20% | 2 |

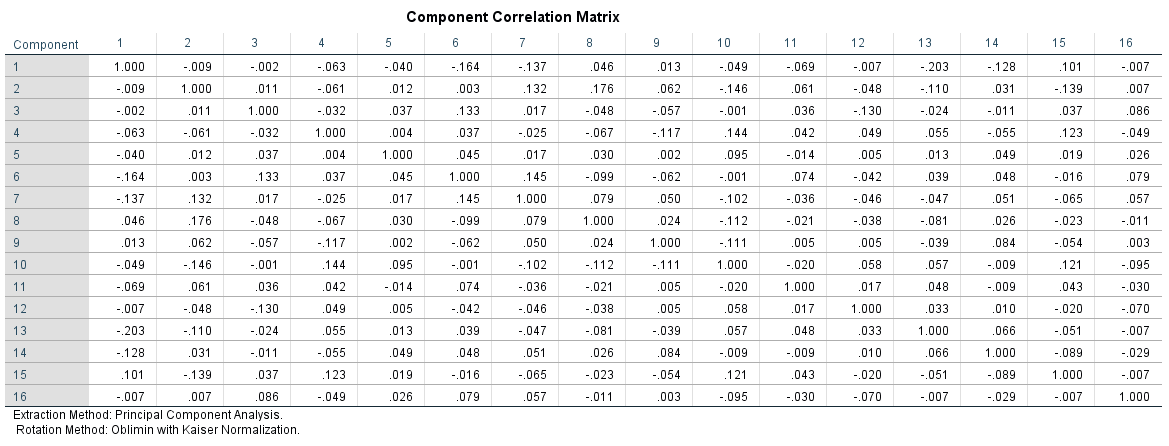
# Appendix 2 – Correlation matrix (partial)

Since a PDF exported document for this matrix is 84 pages long, I am just loading here a partial snippet of the matrix with the purpose of illustrating an example of how it looks like.

A screenshot of a cell phone

Description automatically generated

# Appendix 3 – Oblique rotation results



# Appendix 4 – Varimax rotation results

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

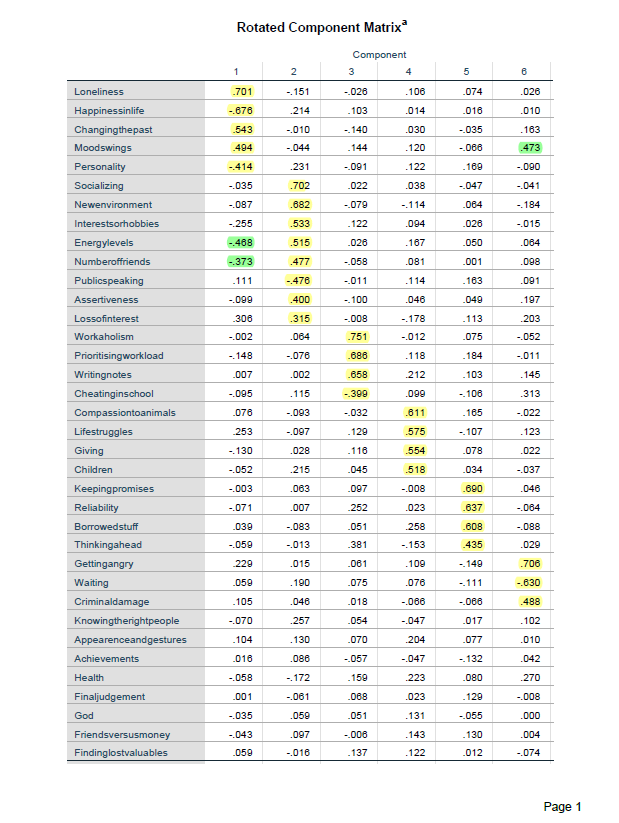
A screenshot of text

Description automatically generated

A screenshot of a cell phone

Description automatically generated

# Appendix 5 – Quartimax rotation results



A close up of text on a white background

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

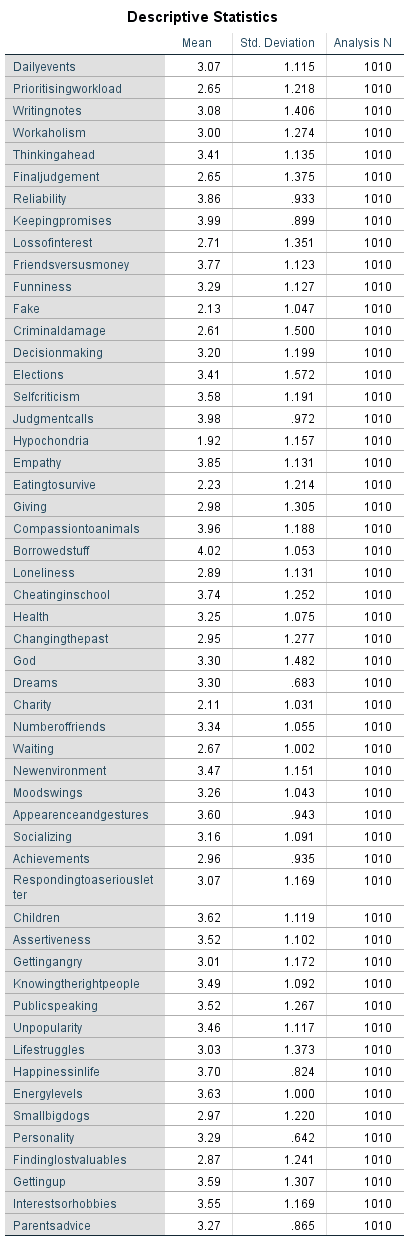
A screenshot of a cell phone

Description automatically generated

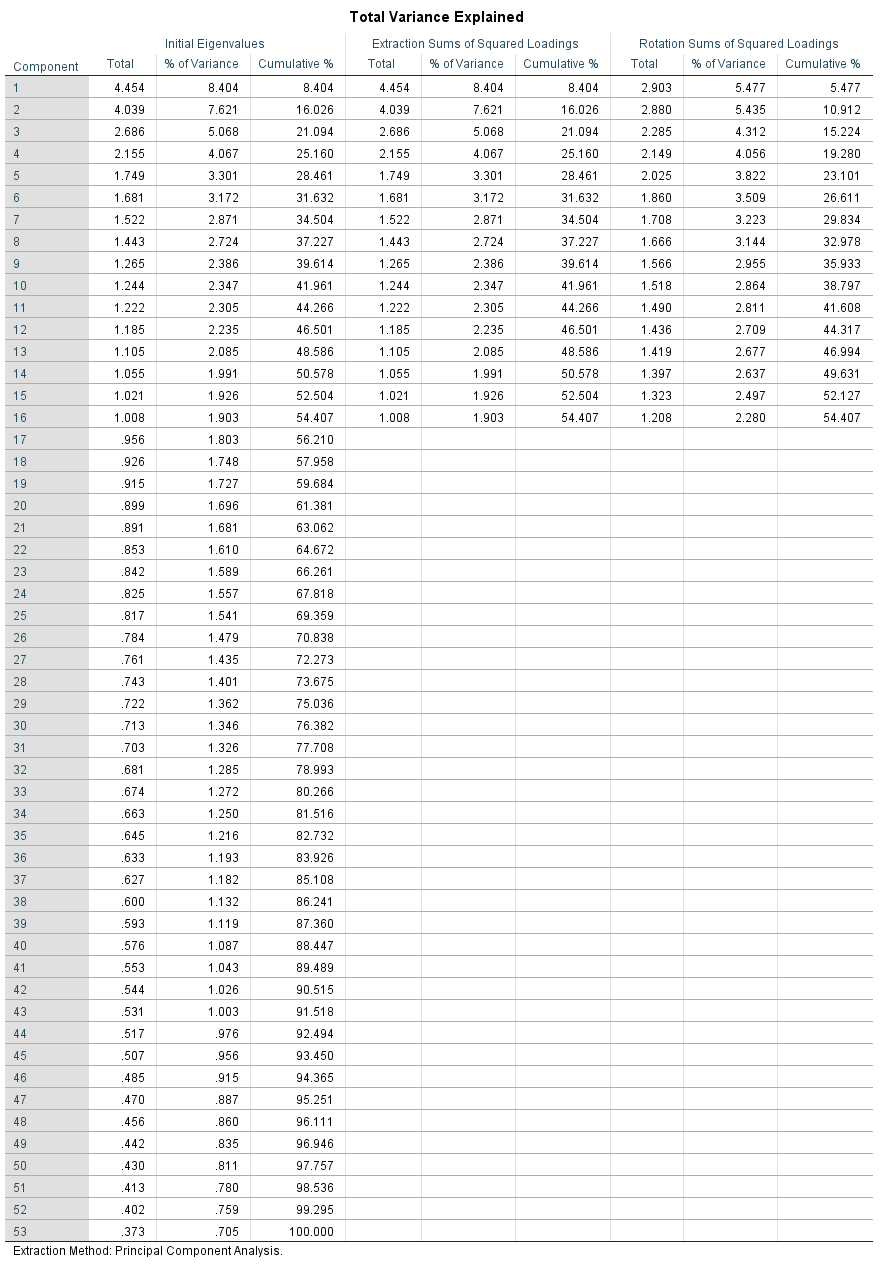
A screenshot of a cell phone

Description automatically generated

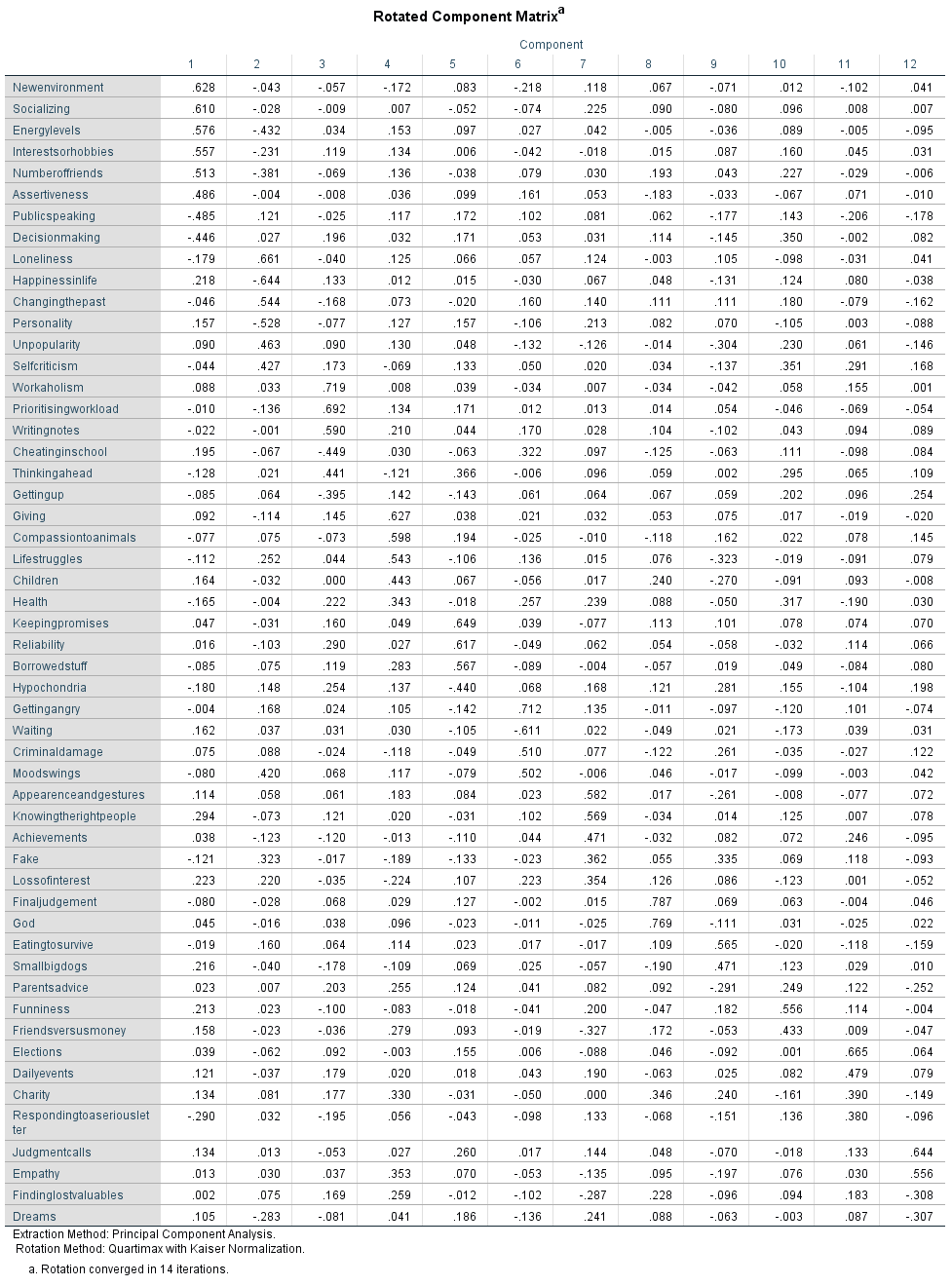
# Appendix 6 – Descriptive Statistics



# Appendix 7 – Total Variance Explained



# Appendix 8 – Rotated Component Matrix



# Bibliography

Ashton, M. C. (2018): *Individual Differences and Personality,* 3rd Edition, *ScienceDirect* [online]. Available at: <https://www.sciencedirect.com/science/article/pii/B9780128098455000032> (Accessed 18th of August 2020)

1. Barrett, K., Leech, B. & Morgan, G. A. (2013): *SPSS for Intermediate Statistics: Use and Interpretation*, 3rd Edition, US: Taylor & Francis.

Brown, J. D. (2009): *Choosing the Right Type of Rotation in PCA and EFA*, University of Hawai‘i at Manoa Shiken: JALT Testing & Evaluation SIG Newsletter (p. 20 - 25) [online] Available at: <http://hosted.jalt.org/test/PDF/Brown31.pdf> (Accessed 19th of August 2020)

Cadima, J. & Jolliffe, I. T. (2016): *Principal component analysis: a review and recent developments*, *Philosophical Transactions of the Royal Society A* [online]. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792409/> (Accessed 18th of August 2020)

1. Cerny, C.A., & Kaiser, H.F. (1977): “A study of a measure of sampling adequacy for factor-analytic correlation matrices”, *Multivariate Behavioral Research*, 12(1), p. 43-47.
2. Fei, J., Maskey, R. and Nguyen, H. O. (2018): *Use of exploratory factor analysis in maritime research*, Volume 34, Issue 2, June 2018, Pages 91-111, *The Asian Journal of Shipping and Logistics* [online]. Available at: <https://www.sciencedirect.com/science/article/pii/S2092521218300245>
3. Hubert, M. (2009): *Comprehensive Chemometrics*, Volume 3, *Chemical and Biochemical Data Analysis* [online]. Pages 315-343. Available at: <https://www.sciencedirect.com/topics/mathematics/scree-plot> (Accessed 19th of August 2020)
4. Kaiser, H. F. (1974): “An index of factor simplicity”, *Psychometrika* 39, p. 31–36.
5. Kaiser, H. F. (1960): “The application of electronic computers to factor analysis”. *Educational and Psychological Measurement*, 20, p. 141-151.
6. Miroslav Sabo (2017): “Young People Survey”, *Faculty of Social and Economic Sciences (Comenius University in Bratislava)* [online]. Available at: <https://www.kaggle.com/miroslavsabo/young-people-survey> (Accessed 19th of August 2020)

1. Image source: <https://community.jmp.com/t5/JMP-Blog/Principal-components-or-factor-analysis/ba-p/38347> [↑](#footnote-ref-1)
2. Ashton, M. C. (2018) [↑](#footnote-ref-2)
3. Cadima, J. & Jolliffe, I. T. (2016) [↑](#footnote-ref-3)
4. Data source: <https://www.kaggle.com/miroslavsabo/young-people-survey> [↑](#footnote-ref-4)
5. Barrett, K., Leech, B. & Morgan, G. A. (2013), page 80 [↑](#footnote-ref-5)
6. Barrett, K., Leech, B. & Morgan, G. A. (2013), page 73 [↑](#footnote-ref-6)