National College of Ireland

Project Submission Sheet – 2019/2020

Student Name:	Antonio Caruso
Student ID:	19203608
Programme:	Higher Diploma in Science in Data Analytics **Year:** 2020
Module:	Advanced Business Data Analysis
Lecturer:	Theo Mendonca
Submission Due	26/08/2020
Date:	22/08/20
Project Title:	Principal Component Analysis on a Phobias Survey Dataset
Word Count:	2458

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: ………

Date:

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

Table of Figures

1 Introduction

This project reports a principal component analysis carried out on data extracted from a survey conducted on a group of 1110 people called “Young People Survey: explore the preferences, interests, habits, opinions, and fears of young people” (Sabo, 2016). The original survey contained 150 items/questions referring to a various number of topics categorised into groups such as music preferences, habits, demographics etc. For the purpose of this project, only elements of the survey referring to the topic group of “Phobias” were extracted and analysed.

The objective of this analysis was to apply a meaningful dimension reduction of the existing quantitative variables to discover latent information and further cluster the number of factors utilised in the survey referred to phobias as particular fears of something, in order to consolidate the survey data about phobias. The following research questions was set up before conducting the research;

"Is there an assumed set of underlying latent logical clusters among the variables classified in the survey questions as potential type or cause of Phobias?"

2 Background

The dataset fully explained in the following paragraph was extracted from a 2013 survey addressed to friends of students of a statistical class at Faculty of Social and Economic Sciences of Bratislava, Slovakia (Sabo, 2016). The original survey conducted in Slovak language, was addressed to Slovakian Nationals aged 15-30 and subsequently translated into English. As aforementioned, the original survey presented 150 questions whose topic allowed a categorisation into 9 groups as following;

Group	**Number of items/questions**
Music preferences	19
Movie Preferences	12
Hobbies & Interests	32
Phobias	10
Health Habits	2

Personality Traits, Views on life & opinions	57
Spending Habit	7
Demographics	10

Figure 1 "Information on Dataset". Source: Sabo (2016)

As previously mentioned, only the group of "Phobias" composed of 10 quantitative variables was chosen.

To answer the research question the technique of Factor analysis was used, being one of the best techniques applied in psychology and education. For the chosen dataset, it resulted to be a perfect suitable model to identify underlying dimension between measured variables and thus reveal grouping factors that underlie individual questions (Härdle & Simar, 2015). Factor analysis is known for splitting into the exploratory Factor Analysis, also called EFA, and Principal component analysis. Although sometimes applied or considered interchangeably, EFA aims to group together and summarise correlated variables that can identify possible underlying latent variables which cannot be measured directly, whereas Principal component analysis (Brown, 2009a) uses all the variance in the variables analysed to reduce the number of variables into smaller set of variables called components that accounts for most of the variance in the original variables .

Between the two, not having built any pre-determined theory on the variables, The PCA was applied on the dataset.

3 Data Description

The data structure presents 10 numerical variables containing the ideal Likert scale values for this research (Allen & Seaman, 2007) on a scale 1 to 5 and with 1110 observations as following;

Variable	Data Type	Variable Description	Value Description
Flying	Numerical Ordinal	**"Fear of flying"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Storm	Numerical Ordinal	**"Fear of Storm as Thunders and Lightning"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Darkness	Numerical Ordinal	**"Fear of darkness"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Heights	Numerical Ordinal	**"Fear of heights"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Spiders	Numerical Ordinal	**"Fear of spiders"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Snakes	Numerical Ordinal	**"Fear of Snakes"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Rats	Numerical Ordinal	**"Fear of Rats"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Ageing	Numerical Ordinal	**"Fear of Ageing"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Dangerous dogs	Numerical Ordinal	**"Fear of Dangerous dogs"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Fear of public speaking	Numerical Ordinal	**"Fear of public speaking"** ranging in Likert scale values from 1 to 5	Not afraid at all 1-2-3-4-5 Very afraid of (integer)

Figure 2" Young People survey dataset". Source: Sabo (2016)

1 and 5 represent the extreme of the values indicating respectively the participant not afraid at all and very afraid. According to Kelley, et al., (2003) 2 could be logically associated with "Little", 3 "some extent", and 4 "rather much". The dataset extracted from the group presented originally some missing values that were treated before starting the analysis.

3.1 Data Cleansing

Some Steps were applied in order to deal with missing values. The first step was to identify whether the values were missing in a random way with the MCAR test (Noruésis, 2002) carried out through Expectation Maximization function. The Null and alternative hypothesis were set as following with the assumption that data were missing completely at random if testing a significance value > 0.5;

H0 = missing values randomly > 0.5

H1 = missing values not randomly < 0.5

Missing Values

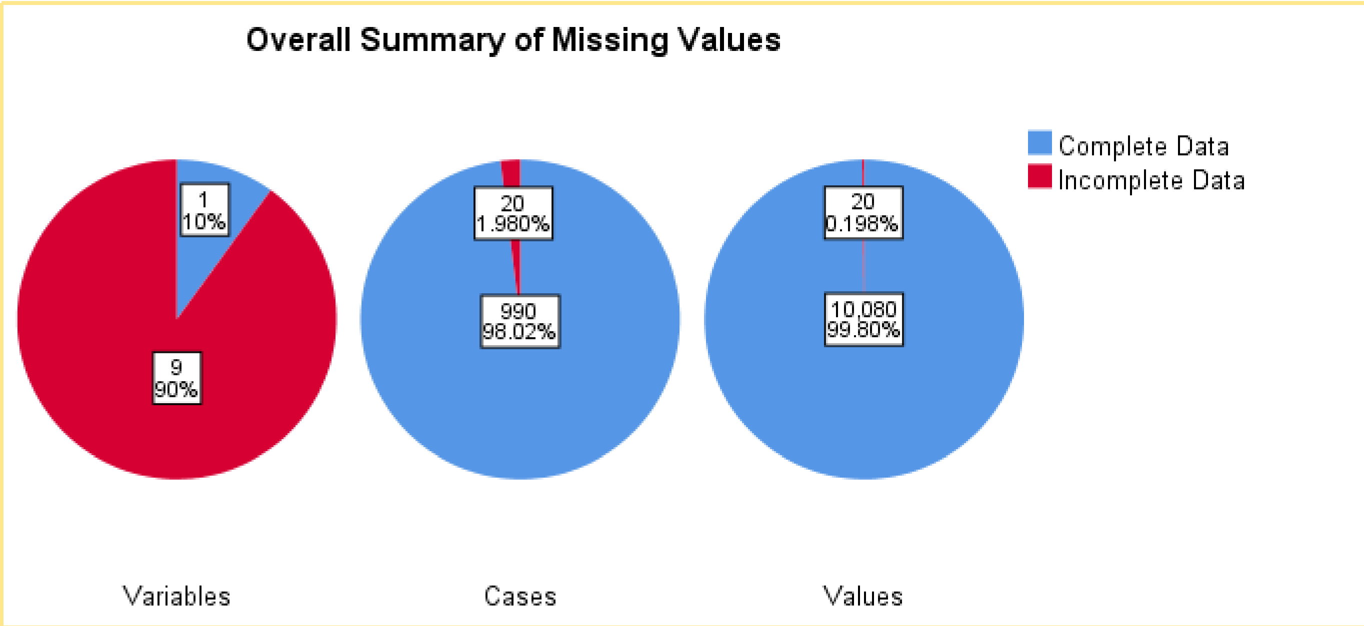

Figure 3: "Overall Summary of Missing Values", own results

The data analysis for multiple imputation (Noruésis, 2002) shows that 9 variables out of 10 had some missing data. The Central pie chart above showed that 990 respondents were missing one value at least, and the third pie chart showed that 0.20% of the total values were missing. Among the variables, Spiders seemed the variable with more missing values accounting for 0.5 of the total sample.

Variable Summary[a,b]

	Missing		Valid N
	N	Percent	
Spiders	5	0.5%	1005
Rats	3	0.3%	1007
Heights	3	0.3%	1007
Flying	3	0.3%	1007
Darkness	2	0.2%	1008
Fear of public speaking	1	0.1%	1009
Dangerous dogs	1	0.1%	1009
Ageing	1	0.1%	1009
Storm	1	0.1%	1009

a. Maximum number of variables shown: 25

b. Minimum percentage of missing values for variable to be included: 0.0%

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
Flying	1007	2.06	1.211	3	.3	0	0
Storm	1009	1.97	1.164	1	.1	0	0
Darkness	1008	2.25	1.255	2	.2	0	0
Heights	1007	2.62	1.295	3	.3	0	0
Spiders	1005	2.83	1.544	5	.5	0	0
Snakes	1010	3.03	1.501	0	.0	0	0
Rats	1007	2.41	1.401	3	.3	0	0
Ageing	1009	2.58	1.386	1	.1	0	0
Dangerousdogs	1009	3.04	1.367	1	.1	0	0
Fearofpublicspeaking	1009	2.80	1.215	1	.1	0	0

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Figure 4: "Variable Summary and Univariate Statistics", own results.

EM Estimated Statistics

EM Means[a]

Flying	Storm	Darkness	Heights	Spiders	Snakes	Rats	Ageing	Dangerousdogs	Fearofpublicspeaking
2.06	1.97	2.25	2.61	2.83	3.03	2.41	2.58	3.04	2.80

a. Little's MCAR test: Chi-Square = 110.208, DF = 81, Sig. = .017

Figure 5: "Estimated Statistics" Own results.

The expectation maximisation mean table (Field, 2013) with a significance of 0.17 < 0.5 , determined the rejection of the null hypothesis that the data were missing completely at random in favour of the alternate hypothesis. Following the result, the automatic multiple imputation with 5 standard iterations was applied to automatically impute values to missing values with a linear regression model due to non-response from survey participants, being t is not safe to listwise delete cases with missing values or singularly impute missing values.

4 Methodology and Calculations

The methodology used to obtain the calculation was a multivariate analysis (Härdle & Simar, 2015) used when several measurements, like in the case of the survey in questions, refer to each individual in the sample. The factor analysis helped understand which items of the survey were mostly correlated with each other to be grouped together. Unlike the Confirmatory Factor Analysis (Harrington, 2009) when the components to be extracted from the data are decided arbitrarily a priori , the Exploratory Factor Analysis was conducted by examining the output of the principal components analysis with a specific protocol. The aim of Principal Component Analysis was in fact to generate artificial variables called principal components which account for most of the variance in the original variables, therefore the technique was applied to interpret the output. Below are indicated the step for this methodology applied to the variables analysed in conjunction at the same time in synchrony;

1) Understand if the data was suitable for factor analysis.
2) Identify how to extract the factors.
3) Identify which criteria would assist in determining factor extraction.
4) Visualise the scree plot.
5) Analyse the communalities .
6) Select the ideal rotational method.
7) Interpretation and labelling of the obtained principal components.

Being the data Numerical Ordinal on Likert scale and being comprised exclusively within quantitative variables, the dataset could be considered suitable for factor analysis without further restrictions having the excellent number of >1000 observations (Allen & Seaman, 2007).

4.1 First step: Analysis of Correlation Matrix for factor extraction

The first step applied was to create a correlation matrix after an initial analysis of descriptive statistics. The correlation matrix for factor extraction came useful as it displayed each variable's loading on each component that contains estimates of the correlations between each of the variables and the estimated components (Pham-Gia & Choulakian, 2014). Since the numerical values taken into consideration were ordinal numerical values, in the correlation matrix was expected the same distance to identify how close these variables were to each other and which variables were highly correlated.

4.2 Second Step: Eigenvalues criterion for factor extraction

The Eigenvalues index was applied as it is an index of strengths of the components (Larsen & Warne , 2010)**,** useful to analyse how much components were relevant for the analysis. As a higher Eigenvalue corresponds to a higher percentage, 1 could be considered a strong Eigenvalue. In fact, each of the components obtained had a representation of the variables with higher loading.

4.3 Third step: Scree plot visualisation

The scree plot showed a visual representation of the eigenvalues (Frey, 2018). This elbow-shaped plot helped choose which components were relevant for this analysis by displaying the eigenvalues in a downward curve, ordering the eigenvalues from largest to smallest, shown in the calculation section. Because scree plots can be subjective and arbitrary to interpret, their primary utility is to provide the number of components useful for the analysis and identified by the elbow-shape, beyond which the other components would be useless.

4.4 Fourth Step: Communalities and total variance

The communality is the sum of the square component loadings up to the number of components extracted (Lewis-Beck, et al., 2004). It is a useful method for predicting the variables' value. A variable that doesn't have any unique variance at all - i.e. one with explained variance that is 100% a result of other variables - has a communality of 1, whereas a variable with variance that is completely unexplained by any other variables has a communality of zero (Lewis-Beck, et al., 2004).

4.5 Fifth Step: Selection of rotational method

Rotation is a method applied to further analyse the results of PCA with the goal of making the pattern of loading clearer, and represents the key output of principal component analysis (Brown, 2009b). According to Gorsuch (1983, pp. 203-204), cited in (Brown, 2009b) there are different type of rotation, among which the most famous are Orthogonal Varimax and Oblique rotation method. The ideal choice should lie in an initial applied oblique rotation, substituted by the Varimax Orthogonal in case factor correlation doesn’t look driven by the data, in which case should determine the choice of Varimax. By obtaining the principal final component, it was then possible to provide the final interpretation and conclude by answering the research question. (Brown, 2009a).

5 Results

5.1 Descriptive Statistics

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Flying	2.06	1.210	1010
Storm	1.97	1.164	1010
Darkness	2.25	1.255	1010
Heights	2.61	1.294	1010
Spiders	2.82	1.543	1010
Snakes	3.03	1.501	1010
Rats	2.41	1.400	1010
Dangerousdogs	3.04	1.367	1010
Ageing	2.58	1.386	1010
Fearofpublicspeaking	2.80	1.214	1010

Figure 6 "Descriptive Statistics", own results

The descriptive statistics shows different items whose scale of responses determined the level of fear of that item felt by the respondents. The mean represents the average of the answers for items on an analysis of 1010 observations. The output obtained shows the mean of the answers in the first column, in which dangerous dogs is the most feared item with an average of 3.04 points, followed by snakes with 3.03, where 3 could represents “some extent” in the grade of fear (Allen & Seaman, 2007), whereas storm with 1.97 seemed to be the lowest factor of fear for the surveyed people. It can be added that in general the animals have higher means that other items and so generate more fear in the respondents. As the Standard Deviation of the values is represented in the second column showing the highest dispersion from the mean (McClave, 2017), the highest relates to Spiders with a stand deviation of 1.54.

5.2 Interpretation of Component Matrix and Correlation Matrix

Correlation Matrix

		Flying	Storm	Darkness	Heights	Spiders	Snakes	Rats	Ageing	Dangerousdogs	Fearofpublicspeaking
Correlation	Flying	1.000	.333	.184	.240	.112	.209	.184	.135	.167	.132
	Storm	.333	1.000	.505	.273	.259	.324	.303	.176	.289	.087
	Darkness	.184	.505	1.000	.297	.312	.268	.308	.212	.211	.158
	Heights	.240	.273	.297	1.000	.173	.208	.190	.135	.215	.137
	Spiders	.112	.259	.312	.173	1.000	.432	.373	.146	.223	.155
	Snakes	.209	.324	.268	.208	.432	1.000	.570	.158	.381	.152
	Rats	.184	.303	.308	.190	.373	.570	1.000	.228	.418	.141
	Ageing	.135	.176	.212	.135	.146	.158	.228	1.000	.258	.109
	Dangerousdogs	.167	.289	.211	.215	.223	.381	.418	.258	1.000	.198
	Fearofpublicspeaking	.132	.087	.158	.137	.155	.152	.141	.109	.198	1.000

Figure 7: "Correlation Matrix", own results

The correlation Matrix shows the correlation between the 10 variables of the extracted Phobias dataset, where each variable represents a fear of something to which each respondent assigned a grade on a Likert Scale. The total amount of variance was 10 as it equals to the number of variables, as well as the initial number of factors extracted. The highlighted numbers represented the loading intended as the relationship of each variable on the factor and underlying dimensions summarizing the observed variables which is a linear combination of the original variables (Pham-Gia & Choulakian, 2014). Since most of the correlations above 0.3 are a good sign, the factors highlighted in red as Rats, Snakes, Spiders, Dangerous Dogs, Darkness, Storm, Flying listed from the highest to the lowest values of correlations, seemed initially to be the most relevant items to be separated into clean factors, whereas Ageing and Fear of Public speaking already did not result useful for the analysis.

Component Matrix[a]

	Component	
	1	2
Rats	.705	-.393
Snakes	.703	-.386
Storm	.655	.380
Darkness	.630	.315
Dangerousdogs	.609	
Spiders	.586	-.313
Heights	.484	.434
Ageing	.412	
Fearofpublicspeaking	.332	
Flying	.443	.488

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Figure 8: "Component Matrix", own results

The extraction process generated an initial component matrix with two component extracted that display the loading, intended as estimates of correlation between the variables extracted within each component before the rotation (UCLA: Statistical Consulting Group., 2020). In component 1 all the items had high loading, even Fear of public speaking within the correlation matrix showed a discrete value of correlation, whereas in component 2 it's clearly visible the high loading of Flying, Heights, Storm and Darkness. The key point is the fact that this matrix was temporary and could change after the rotation.

5.3 Total Variance explained: Eigenvalues and Communalities

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	3.242	32.425	32.425	3.242	32.425	32.425	2.871
2	1.126	11.263	43.688	1.126	11.263	43.688	2.410
3	.979	9.785	53.473				
4	.910	9.098	62.571				
5	.857	8.572	71.142				
6	.765	7.653	78.795				
7	.710	7.099	85.894				
8	.558	5.585	91.479				
9	.455	4.546	96.025				
10	.397	3.975	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Figure 9: "Total Variance explained table", own results

To generate the total variance explained table, the eigenvalues greater than 1 was firstly selected to extract the components > 1 that were used after the rotation, then a maximum number of iterations was set to find an optimal convergence (Larsen & Warne , 2010) . The table provides more information about the Eigenvalues for the components. Having selected only Eigenvalues > 1 for the extraction, only component 1 and 2 were extracted and considered for the analysis as the highlighted components in red represent less than 10% of the total variance, and therefore were too weak to be considered for the analysis. On the right of the table it was visible the Eigenvalues cumulative loadings representing 74% of the variance, and the rotation sum of the squared loadings.

Communalities

	Initial	Extraction
Flying	1.000	.434
Storm	1.000	.574
Darkness	1.000	.496
Heights	1.000	.423
Spiders	1.000	.441
Snakes	1.000	.643
Rats	1.000	.651
Dangerousdogs	1.000	.426
Ageing	1.000	.170
Fearofpublicspeaking	1.000	.110

Extraction Method: Principal Component Analysis.

Figure 10: "Communalities table", own results

The communalities table shows the result of 1 before the extraction as previously selected, and the values after the extraction always smaller than the total variance. Initially each variable represented 100% variance of that variables, then after the extraction it is evident how Rats, Snakes and Storm had the highest representation in the components respectively with 0.65, 0.64 and 0.57 which represent the sum of the squared loading.

5.4 Scree plot visualisation

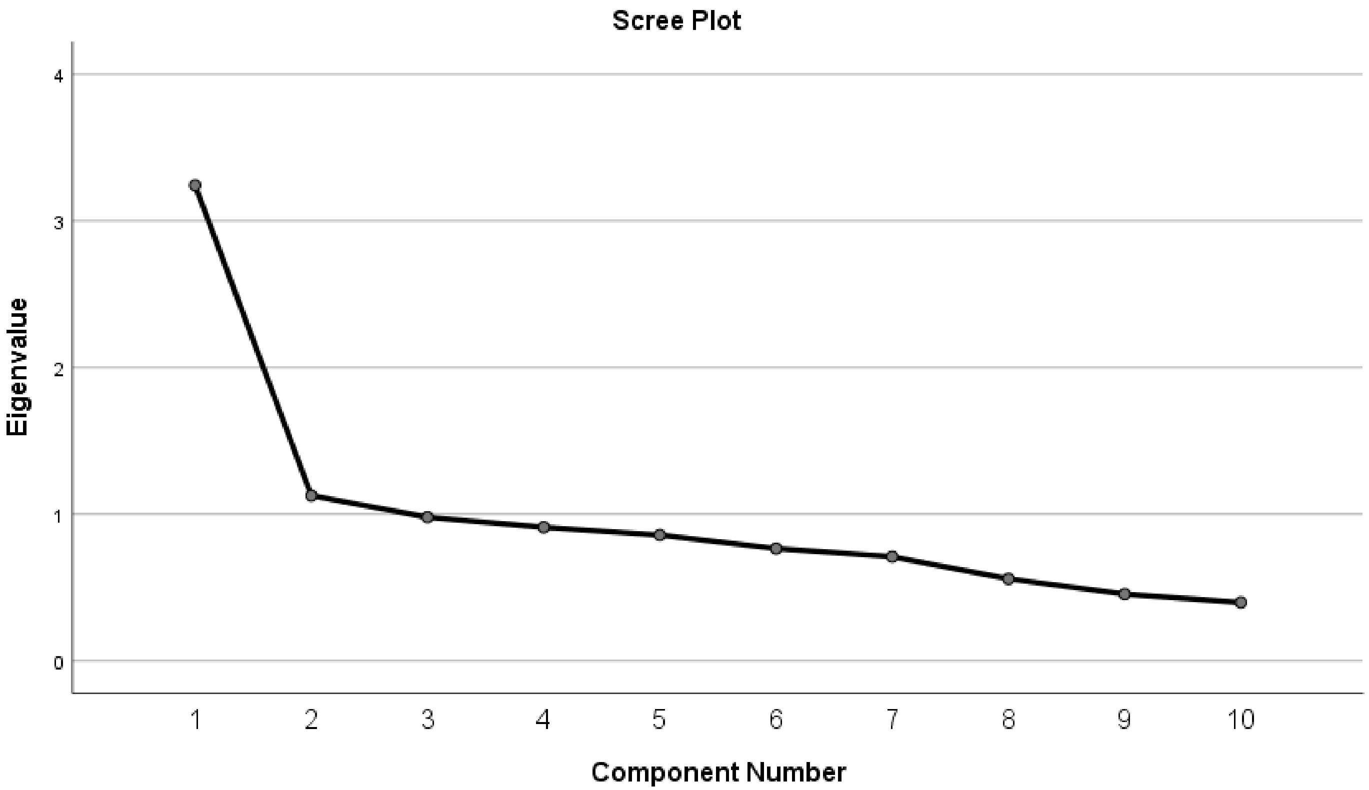

Figure 11: "Scree plot" own results

The scree plot (Frey, 2018) shows the Eigenvalues on the y-axis and the number of factors on the x-axis displaying a downward curve. The slope of the curve shaped like an elbow indicates the number of factors should be generated by the analysis, corroborating therefore the results shown previously and succeeding in finding a smaller number of interpretable factors that explain the maximum amount of variability in the survey data

5.5 Oblimin rotational method - Rotated component matrix & Component Plot in Rotated Space

The direct Oblimin with rotated solution was applied to obtain a non-orthogonal oblique solution in which the variables are allowed to be correlated and result in higher eigenvalues as previously selected, allowing also dependence of factors (Kirby, et al., 2015). The process resulted in iterations previously applied reorganising the items into new factors by rotating them away from each other's creating meaningful and separate factors. Although in some cases the rotation is not able to segregate the variables in different components, in this analysis it was possible and so the latent relationships could be named. The rotation helped segregate the components even more so no leakage of variables across the components after segregation would happen. Being segregated, it was easier to define and label the components.

Pattern Matrix[a]

	Component	
	1	2
Rats	.835	
Snakes	.828	
Spiders	.682	
Dangerousdogs	.630	
Ageing		
Fearofpublicspeaking		
Flying		.702
Storm		.692
Heights		.667
Darkness		.614

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Figure 12: "Rotation Table", own results

The Rotated Matrix represents the key outpuf of principal component analysis which corresponde directly to the factors representing groups for the two components. With no leakage of extra factors, the variables were grouped clearly into component 1 and 2 properly separated. Component 1 shows the highest loading for Rats, Snakes, Spiders and Dangerous dogs forming the first cluster, whereas the component 2 has Flying, Storms, Heights and Darkness forming a second cluster, so that could both be clearly distinguished and named.

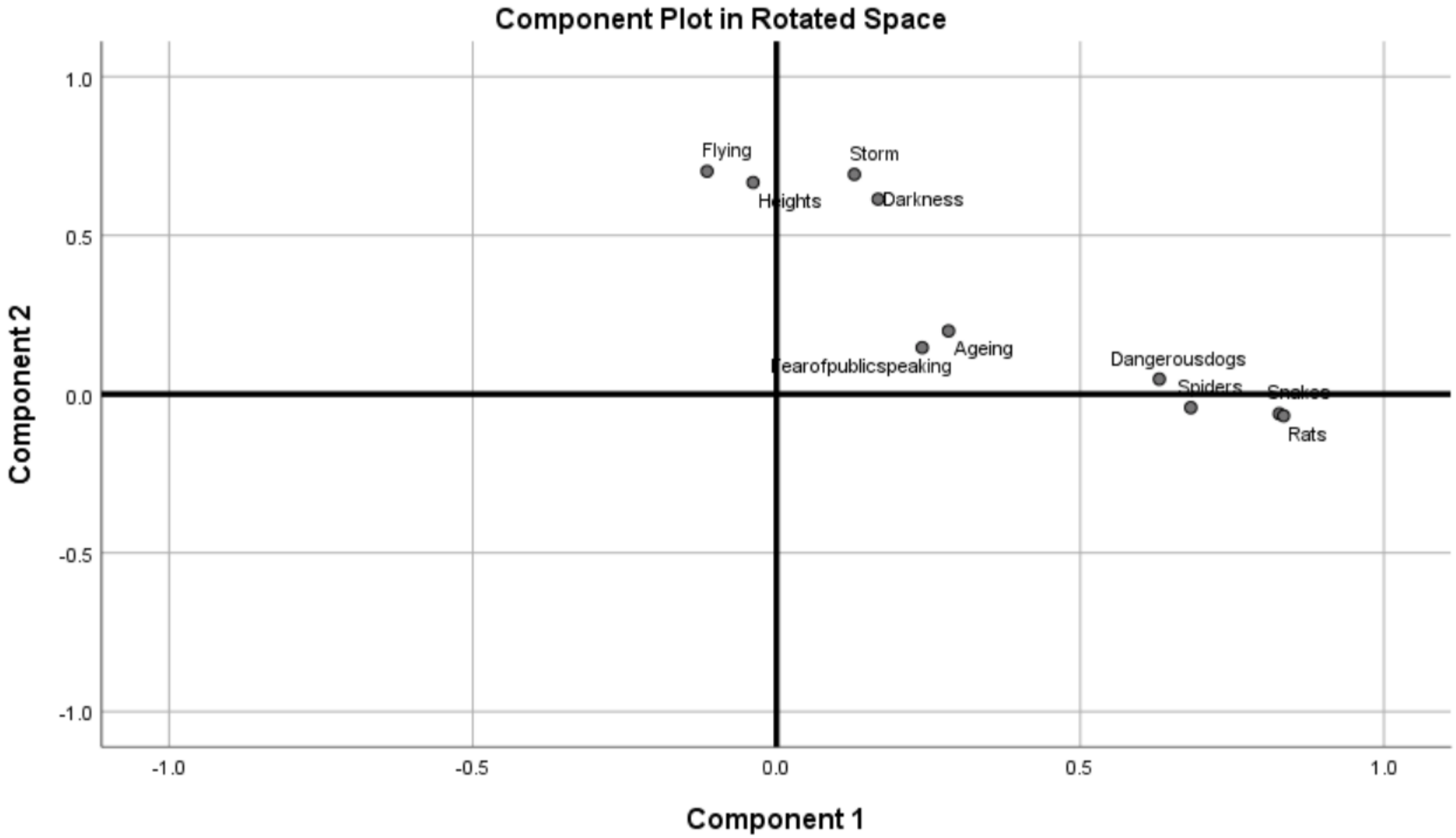

Figure 13: Component Plot in Rotated Space", own results

The Component Plot in rotated space visually displays the cluster obtained (Brown, 2009a). Component 1 is positioned on the horizontal axis whereas component two on the vertical axis. It clearly displays the group of variables Flying, Storm, Heights and Darkness with a high loading in component 2, whereas the group with the animals are placed at zero or in proximity of component 2, but have a high loading in component 1, forming the first cluster. The variable Ageing and Fear of Public speaking, as anticipated in the beginning, could not form any meaningful group being close to zero in the intersection between the two components.

6 Discussion and Analysis of Results

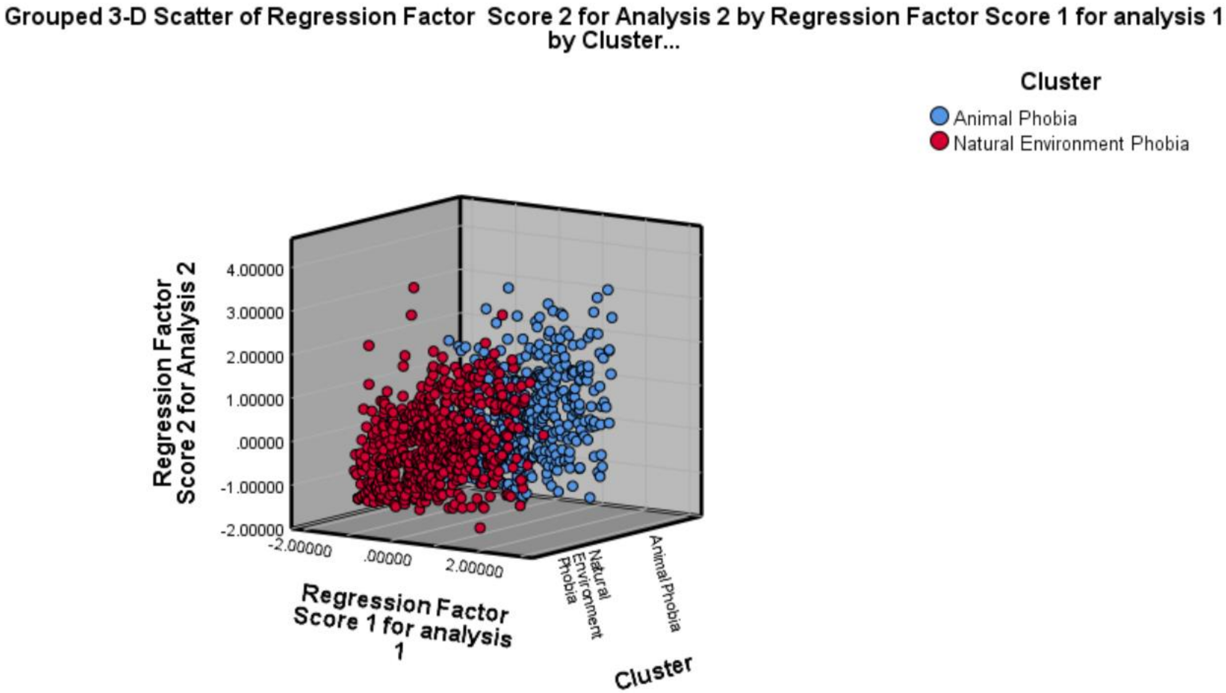

Figure 14: "Grouped 3D Scatter plot", own creation

The 3D scatterplot above displays the conclusion of the analysis and the response to the research question the graph clearly shows the obtained two latent groups and separated into two clusters , labelled “Animal Phobia” for the variables Rats, Snakes, Spiders and Dangerous dogs presenting a high loading in component one, and “Natural Environmental Phobia “ which clearly resonates with the items Darkness, Storms, Heights and Flying with a high loading in component 2. The separated Clusters give visual proof of how strong is the relationship between two groups, and provides a valid answer to the research question showing that a set of underlying latent logical clusters among the variables classified exists. It can be therefore stated that the classified two type of Phobias for Animals and Natural environment would highly be specific phobias suffered by the survey respondents based on their responses as proven by the principal component analysis conducted.

7 References

Allen, E. & Seaman, C. A., 2007. *Statistics Roudtable: Likert Scales and Data Analyses.* [Online]
Available at: http://rube.asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html
[Accessed 10th August 2020].

Brown, J. D., 2009a. Principal components analysis and exploratory factor analysis &ndash. *Shiken: JALT Testing & Evaluation SIG Newsletter,* 13(1), pp. 26 - 30.

Brown, J. D., 2009b. Choosing the Right Type of Rotation in PCA and EFA. *Shiken:JALT Testing & Evaluation SIG Newsletter,* 13(3), pp. 20 - 25.

Central Statistics Office, 2016. *Census 2016 Small Area Population Statistics.* [Online]
Available at: https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/
[Accessed 24 July 2020].

Chen, S.-Y., Feng, Z. & Yi, X., 2017. A general introduction to adjustment for multiple comparisons. *Journal of Theoracic Disease,* Volume 9, p. 1725–1729.

Field, A. P., 2013. *Discovering statistics using IBM SPSS.* 4th ed. London: SAGE.

Frey, B. B., 2018. *Scree Plot: The SAGE Encyclopedia of Social Science Research Methods.* [Online]
Available at: https://methods.sagepub.com/reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i18507.xml
[Accessed 14th August 2020].

Härdle, W. & Simar, L., 2015. *Applied multivariate statistical analysis.* 4th ed. Berlin :: Springer.

Harrington, D., 2009. *Confirmatory Factor Analysis.* New York: Oxford University Press.

Kelley, K., Clark, B., Brown, V. & Sitzia, J., 2003. Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care,,* 15(3), p. 261–266.

Kirby, J. R., Kim, H.-J. & Silvestri, R., 2015. Chapter 11 - Cognitive Constructs and Individual Differences Underlying ADHD and Dyslexia: A Cognitive Mosaic Approach. In: *Cognition, Intelligence, and Achievement: A Tribute to J. P. Das.* Ontario: Academic Press, pp. 197-223.

Larsen, R. & Warne , R. T., 2010. Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods*, August, p. 871–876.

Lewis-Beck, M. S., Bryman, A. & Liao, T. F., 2004. *Communality: The SAGE Encyclopedia of Social Science Research Methods.* [Online]
Available at: https://methods.sagepub.com/reference/the-sage-encyclopedia-of-social-science-research-methods/n139.xml
[Accessed 14th August 2020].

McClave, J. &. S. T., 2017. *Statistics, Global Edition.* 13th ed. Harlow: Person.

McClave, J. & Sincich, T., 2017. *Statistics, Global Edition..* 13th ed. Harlow: Person.

Noruésis, M. J., 2002. *SPSS 11.0 guide to data analysis.* New Jersey: Prentice Hall.

Pham-Gia, T. & Choulakian, V., 2014. Distribution of the Sample Correlation Matrix and Applications. *Open Journal of Statistics,* Volume 4, pp. 330-344.

Sabo, M., 2016. *Young People Survey - Explore the preferences, interests, habits, opinions, and fears of young people.* [Online]
Available at: https://www.kaggle.com/miroslavsabo/young-people-survey?select=responses.csv
[Accessed 13th August 2020].

UCLA: Statistical Consulting Group., 2020. *Principal Components (PCA) and Exploratory factor analysis (EFA) with SPSS.* [Online]
Available at: https://stats.idre.ucla.edu/spss/seminars/efa-spss/#s2
[Accessed 13th August 2020].