Use of Time per country related to Happiness
A Cluster, PCA and FA analisys.

Frederico Vieira
Mestrado em Ciência de Dados
DETI
Universidade de Aveiro
Aveiro, Portugal
N.Mec. 98518

João António
Mestrado em Ciência de Dados
DETI
Universidade de Aveiro
Aveiro, Portugal
N.Mec. 76558

Luís Costa
Mestrado em Ciência de Dados
DETI
Universidade de Aveiro
Aveiro, Portugal
N.Mec. X

Tijan Bah
Mestrado X
DETI
Universidade de Aveiro
Aveiro, Portugal
N.Mec. X

***Abstract*—**

***Index Terms*—PCA, Clustering, FA, Time Use, Happiness**

December 12, 2022

Notes

The contribution of every user is already discriminated at the contributions, please try to have it ready by 20 December.

Whenever I need to write comments, I tend to use this violet color, please use a different color when doing your commentary. Please avoid writing on other user's contribution without checking with him.

Lastly, all that we have discussed in the meetings is a suggesting, if you want to suggested extra work, like more questions or compare it to anything feel free to suggest it.

1. Introduction/Motivation

2. Methodology

2.1. Clustering

2.2. Principal Component Analysis PCA

Principal component analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

2.3. Factor Analysis (FA)

Factor Analysis (FA) is also one of the tools from multivariate statistics tool-set. Can be interpret as a dimensionality reduction just like PCA, with the main difference that in PCA, the PC's are a linear combination of the observed variables of the dataset, while in FA, the observed variables of the dataset are themselves linear combination of the factors ("unobservable" variables), plus for each variable an error term (also known as specific factor) [1], [2]. Since this factors are not known we call them frequently latent.

This being said one of the most used libraries in R for FA, named 'factanal' requires us to place as input the number of factors that we want to use. To solve this we can start with an informed guess, with the number of factors that we have used in PCA to obtain a good initial estimation and lastly we have the option to do a Scree Test introduced in the article [3]. This type of plot find the optimal solution using a elbow technique under the hood. (And can be used for both FA and PCA, but it is more common on FA.)

Going back to the philosophy of the FA, the variability of our data X is given by Σ and it is estimated by $\hat{\Sigma}$. Which is composed by the variability explained by the factors explained as a linear combination of the factors (communality - $\hat{\Lambda}\hat{\Lambda}^T$) and the variability that cannot be explained by the combination of factors (uniqueness - $\hat{\Psi}$) [4]:

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}^T + \hat{\Psi}$$

We have decided to explain the Kaiser-Meyer-Olkin factor adequacy test and the Bartlett's test for sphericity on the results since we can explain them with the values that we obtain.

3. Dataset Introduction

We have used three main datasets, that we found online, the first is "Time Use", the second "Life Satisfaction vs GDP per capita" and the third is "Life Satisfaction vs Geography per year".

3.1. Time Use

We found and article [5] that makes a brief analysis of the time use per person, on particular perk of this journal, is that all graphs are interactive and all data can be downloaded for further analysis.

The first graph that they have is using data from the [6], they simply aggregate the results of [7] for the year of 2020. They bring together estimates from time diaries where users recorded where asked to recall the amount of time spend on different activities from one specific day of the previous week. It is worth noting that this could be a lot more efficient by in fact measuring, and not just asking. We humans, are notoriously bad at recalling what we did, and we are also notoriously bad at measuring time and on surveys we can always introduce our own bias. Yet the results were interesting enough to be credible.

So on this dataset we have a, average over multiple users, a regular working day for that country, with minute precision (note that all will sum to $1440 = 24 \times 60$). We encountered some minimal differences to some countries (4 of them), to fix this we just subtracted the extra minutes (or added the missing ones) from the sleep time, since it was just a small quantity of minutes, that this doesn't change the results in any significant way.

Also returning to the initial article he cited [8] with some interesting graphs that related geography, use of time. We used this dataset for questions one, two and three. This dataset also does not include any missing value which is very good since both PCA and FA don't work very well with missing values.

3.2. Life Satisfaction vs GDP per capita

We wanted now to find some sources that related happiness with at least geography, this is a lot harder, since happiness is hard to measure. We found one related characteristic, life satisfaction, which is how happy the user see itself, so it is linked, but is not exactly the same, yet through this report we will considered them equal. Thought the text we will use both terms.

We found the article [9] that relates life satisfaction. Here they also gather a lot of information that is quite useful for a depth study on Happiness, namely they visited the [10] to obtain some of the data, among other data sets.

We were particular interested in the section "The link across countries", since it allows us to relate *GDP per capita*, life satisfaction and geography.

We used this dataset for the first question.

3.3. Life Satisfaction vs Geography per year

Lastly, and also from the article [9] we are also interested in how the life satisfaction evolved over time, so we obtained the dataset with the same name for our fourth question.

3.4. Assumptions

Like we stated previously this data sets have a few assumptions, namely they assume that

1) People will not have bias /tendencies to lie or that they will fade out with a large enough dataset;
2) People can remember previous working days (until one week away) activities with a minute precision;
3) People are not multitasking in any form;
4) That everyone considers the activities equally, for example, for me going taking a walk might just be leisure, but someone with a more sedentary life than mine might consider sport.
5) The people filling the surveys were taken at random without any type of bias.
6) The survey answers will not depend on the day/hour that the survey was made, or this fade out in the large sample size.

3.5. Problems with Our World in Data

We have a very mixed feeling about this website "Our World in Data", from one point of view we adore it since all the conclusions come with data that we can download and use, on the other side they have some controversy results that were dis-proven by independent researchers leading us to believe that at least in some articles they might be introducing some mistakes or assumptions.

Part of the problem is that Our World in Data is deeply funded by grants from the Quadrature Climate Foundation, the Bill and Melinda Gates Foundation, and a grant from the German entrepreneur, businesswoman and philanthropist Susanne Klatten [11]. All of which are interested in certain results from to justify certain products or actions.

To make matters worse, a quick research reveals that some scientist don't feel okay in talking about Bill & Melinda Gates our of fear in loosing support or funding [12], [13], [14]. This is not how we want science to evolve.

Still the website was a major source of inspiration and sources to obtain the data that we will use, all of it we were able to trace back to other sources but most was still funded by the same people.

4. Work Questions

5. Results

5.1. Q1:

What can we learn when compared to more happy places? (TRY to see what factors/PC most affect happiness) How does life satisfaction vary per GDP?

5.2. Q2: PCA on Time use

What are the countries that are more similar in terms of time spent? ([[Principal Component Analysis (PCA)—PCA]])?

TABLE 1. RESULTS OF THE KMO TEST FOR THE TIME USE DATASET, "RM" MEANS REMOVAL OF.

Category	MSA	Rm Paid Work	Rm MSA <0.4
Paid Word	0.07	-	-
Education	0.03	0.47	0.48
Household Members	0.05	0.51	0.49
Housework	0.07	0.59	0.71
Shopping	0.08	0.56	0.61
Unpaid work	0.06	0.53	0.61
Sleep	0.05	0.62	0.66
Eat	0.04	0.45	0.5
Person Care	0.05	0.40	0.68
Sports	0.04	0.34	-
Events	0.05	0.44	0.58
Friends	0.03	0.39	-
TV & Radio	0.03	0.38	-
Other leisure	0.07	0.62	0.66
Overall MSA	0.05	0.5	0.6

5.2.1. Cluster after PCA.

5.2.2. What can Portugal learn after the PCA.

5.3. Q3: FA on time use

Once more we will use the dataset "Time Use", like we have said previously we know that it doesn't have any missing value which is good. We decided to also test for correlation between variables, since if any variable was highly correlated to another we could remove from the start reducing the complexity of the analysis, and the correlation values were always above 0.6, so we will not remove any from the start.

The next step is to test if Factor analysis is a good idea on our dataset, for this we do a Kaiser-Meyer-Olkin factor adequacy on our dataset. The results of it can be seen in Table 1.

This results (Table 1) are very discouraging to use Factor analysis on this dataset, since most authors, for example [15], [16] suggest at least 0.50 MSA (measure of sampling adequacy) for every column used.

Since we still wanted to check if we could extract any type of insights from the Dataset still, so that we could compare we decided to remove the "Paid Work", without it we obtained a Overall MSA of around 50%, which is not great still, this is the third column of the Table 1. So we decided to remove the all the columns that had less than 0.40 of KMO this resulted the fourth column of the same Table 1.

Even though this changes ended up changing our original dataset a lot, reducing the number of observable occurrences to 10 from 14, it was a necessary step to increase the overall MSA. Still according to [16] the 0.6 results that we obtained is considered mediocre. We obtained better results of MSA by reducing the categories still but this would make this results from question even more difficult to compare to the ones above.

The next test to run is the Bartlett's test for sphericity, the null hypotheses of this test is:

$$H_0 : matrix = identity$$

And if we can't reject the null hypotheses, there is nothing for us to factor. By running the Bartlett's test we obtain a p-value of 0.00062 so we can reject the null value, so we can proceed with the Factorial Analysis test.

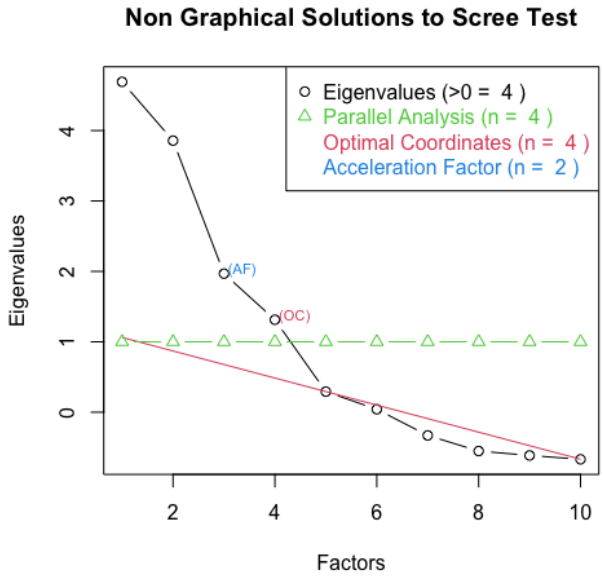

Figure 1. Scree Plot for the FA.

After this we do a plot a Scree plot to decide on the number of factors to use, this can be see in Figure 1. Here we can see that most of the tests recommend four factors for our data, with the Acceleration Factor recommended two.

After this decision we have one more, this time how do we want to extract and rotate the factors:

- promax - which is very popular due to fact that can deal with large datasets efficiently.
- oblimin - produces simpler factor structures
- varimax - optimized to reduce cross loading and minimize loading values
- quartimax - reduce the number of variables needed to explain one factor, making interpretation easier
- equamax - a compromise between quartimax and varimax

here it is worth noting that the first two options assume correlated factors (non-independent) while the latter three assume uncorrelated (independent) factors.

As such we decided to start with quartimax but we obtain lower loadings with it for some of the factors (we considered low when the $|\text{factor loadings}| < 0.4$, so we experimented with promax and we obtained all the loadings above 0.35 with only one of them lower than 0.4 but all a correlation matrix with relatively high correlation, so we decided to

try varimax, and again the loadings were quite high so we decide to keep it.

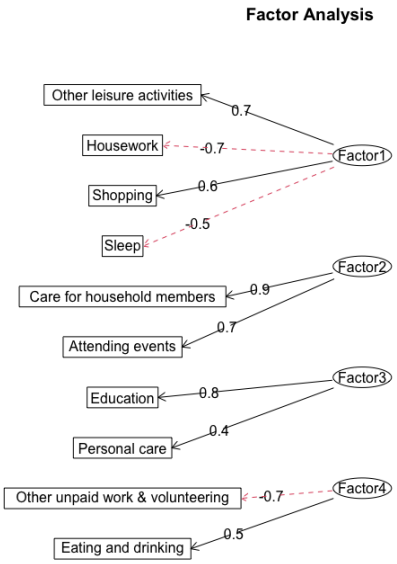

Figure 2. How every observable contributes to every factor using varimax rotation and 4 factors.

In Figure 2 and on Table 2 we can see how the the various observable contribute to every factor, the red dotted line means a negative contribution. Here we can do some subjective work trying to find a possible cause for every factor that we obtained. On the diagram, every observable only belongs to one factor (the one where it has the higher modular value) while on the Table we can see the full loading.

- Factor 1 - An isolated factor, here the important tasks that most of us think of as responsibilities almost are in red (contributing negatively), while the black activities are more hobbies.
- Factor 2 - A social factor, with a great sense of belonging, sense of community.
- Factor 3 - Another isolated factor, this time both activities contribute a lot for how other people see us.
- Factor 4 - Another social factor, Egoism or reversed altruism seem to be the general trend, since volunteering contributes negatively and eating contributes positively.

Lastly from this study we can see on Table 3 the SS loadings, the proportional variance and the cumulative variance, it is rather odd that we only obtained 0.58 % of the variance explained by this variables yet we were already able to make one possible description of the latent factors.

5.3.1. Cluster after FA. So we want to do cluster after the FA and using the results that we obtained in FA, like we have discussed before we have to decide on the number of

TABLE 2. LOADINGS AND UNIQUENESS (REPRESENTED AS $\hat{\Psi}$) OF THE FA FOR THE VARIOUS CATEGORIES OF ACTIVITIES.

Loadings	Factor 1	Factor 2	Factor 3	Factor 4	$\hat{\Psi}$
Housew.	-0.68				0.30
Shopp.	0.56	0.40	-0.45		0.15
Sleep	-0.52				0.52
Other	0.72				0.31
Care		0.86			0.00
Events		0.68			0.65
Educat.			0.81		0.73
Volut.	0.69			-0.71	0.62
Eat				0.51	0.50
Pers.Care			0.42	0.39	0.44

TABLE 3. SS LOADINGS, PROPORTION OF VAR AND CUMULATIVE VAR FOR THE FA STUDY.

Loadings	Factor 1	Factor 2	Factor 3	Factor 4
SS Loadings	2.16	1.44	1.12	1.05
Proportional Var	0.22	0.14	0.11	0.10
Cumulative Var	0.22	0.36	0.47	0.58

clusters that we want to use, for this we have used a majority rule [17]. The majority rule was studied from three clusters to 15, the results can be see in Figure 3, as we can see the recommendation for Kmeans was to use 3.

In Figure 4 we can see how the various clusters are composed, as we can also see the clusters sizes are not uniform at all, with the first cluster having the most countries, the second group seems to be the more diverse one (includes Portugal) while the last group seems to be the happiest including Japan, Norway and Finland.

5.3.2. What can Portugal learn after the FA. Lastly we want to know how does every single cluster and country performs scores on every factor, to see if we can extract any type of insight from it.

On Figure 5, we can see a heatmap with dendograms both for the Factors and for the countries, the data was scaled by column to make it easier to understand [18] [18]. Besides this we have included two red lines to differentiate between the clusters.

We have also included a legend, with the values just to be easier to compare the values (the scale goes from brow to white to dark blue).

Since our idea is to generate insights in how Portugal can become closer to the ”happy cluster”, the main differences is that they have higher values than Portugal for Factor 1 and lower values for Factor 4.

So if we combine this with our interpretation of the factors we obtain: - That Portugal should increase the amount of activities that give us pleasure, while reducing the one that we consider responsibilities, or at least change our mindset about how we view our responsibilities. In these three ”happy” countries the traditions there is great respect and joy associated with the traditions, and most of this traditions translate also to responsibilities. This would increase the Factor 1. - And reduce the Factor 4, we previously interpret

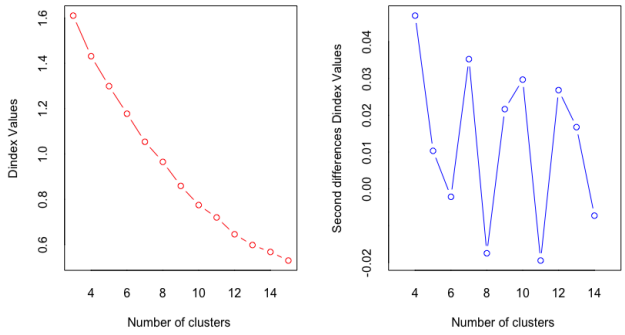

Figure 3. Determine the best number of clusters to use, using a majority rule.

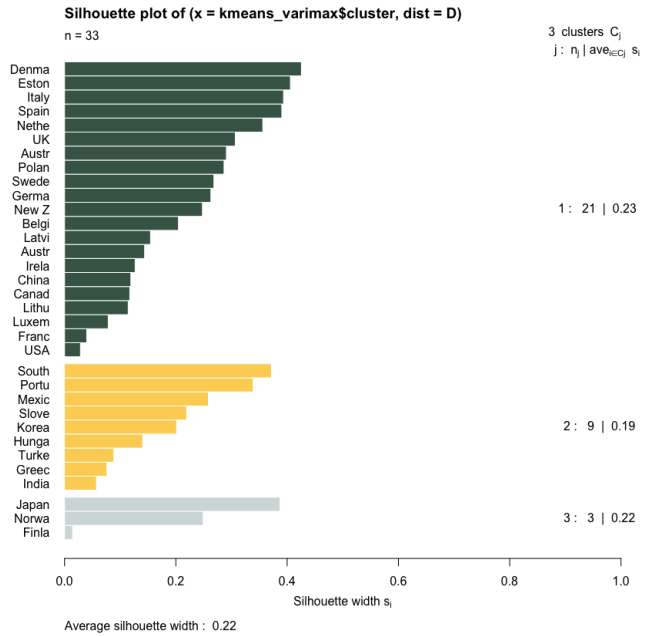

Figure 4. What countries belong to each cluster, using 3 clusters.

it as egoism, so reducing our egoism tendencies as even increasing our altruism habits can increase our happiness overall according to this comparison.

None of this two insights is specially new for any of us, a lot of examples can be found just by doing a quick search for the second insight, for example [19], [20]. The first is something that we all know that by spending more time in what makes us actually happy (note: TV and Radio were excluded), and reducing the negative images that we have about what we have to do (responsabilities) we can all increase our life satisfaction.

It is missing the rest of the discussion of the cluster graph and it is also missing what can we learn from more happy countries, like cluster 2.

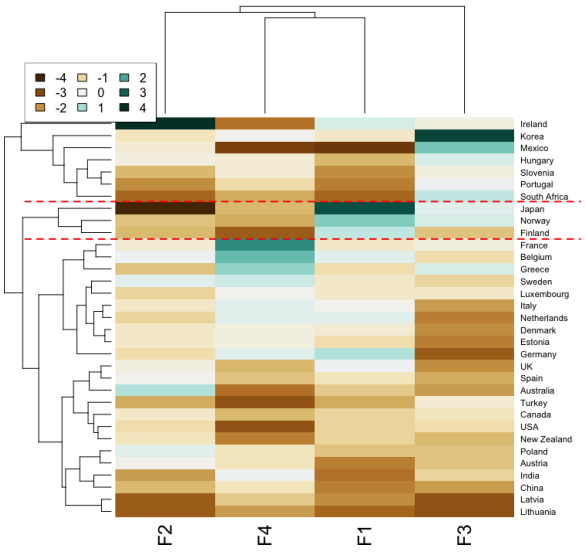

Figure 5. Heatmap of scores for every country and for every cluster.

5.4. Q4: Evolution of Happiness for Portugal

How does happiness evolve for Portugal over time? what what are the countries that most changed?

6. Conclusions

References

[1] I. Joliffe and B. Morgan, "Principal component analysis and exploratory factor analysis," *Statistical Methods in Medical Research*, vol. 1, no. 1, pp. 69–95, 1992, pMID: 1341653. [Online]. Available: https://doi.org/10.1177/096228029200100105

[2] I. T. Joliffe and B. Morgan, "Principal component analysis and exploratory factor analysis," *Statistical methods in medical research*, vol. 1, no. 1, pp. 69–95, 1992.

[3] R. B. Cattell, "The scree test for the number of factors," *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276, 1966, pMID: 26828106. [Online]. Available: https://doi.org/10.1207/s15327906mbr0102_10

[4] "A simple example of factor analysis in r," May 2017, [Online; accessed 11. Dec. 2022]. [Online]. Available: https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/factor-analysis/A-simple-example-of-FA/index.html

[5] E. Ortiz-Ospina, C. Giattino, and M. Roser, "Time use," *Our World in Data*, 2020, https://ourworldindata.org/time-use.

[6] "Oecd-2020." [Online]. Available: https://stats.oecd.org/Index.aspx?DataSetCode=TIME_USE

[7] "Multinational time use study." [Online]. Available: https://www.timeuse.org/mtus

[8] R. C. Feenstra, R. Inklaar, and M. P. Timmer, "The next generation of the penn world table," *American Economic Review*, vol. 105, no. 10, pp. 3150–82, October 2015. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/aer.20130954

[9] E. Ortiz-Ospina and M. Roser, "Happiness and life satisfaction," *Our World in Data*, 2013, https://ourworldindata.org/happiness-and-life-satisfaction.

[10] “The world happiness report is a publication of the sustainable development solutions network, powered by the gallup world poll data.” [Online]. Available: https://worldhappiness.report/

[11] “How we’re funded.” [Online]. Available: https://ourworldindata.org/funding

[12] “The media loves the gates foundation. these experts are more skeptical.” [Online]. Available: https://www.vox.com/2015/6/10/8760199/gates-foundation-criticism

[13] “Who official complains about gates foundation’s dominance in malaria fight.” [Online]. Available: https://www.nytimes.com/2008/02/17/world/americas/17iht-gates.4.10120087.html

[14] “Not many speak their mind to gates foundation.” [Online]. Available: https://www.seattletimes.com/seattle-news/not-many-speak-their-mind-to-gates-foundation/

[15] Stephanie, “Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy,” *Statistics How To*, May 2021. [Online]. Available: https://www.statisticshowto.com/kaiser-meyer-olkin

[16] “RPubs - Exploratory Factor Analysis in R,” Dec. 2022, [Online; accessed 11. Dec. 2022]. [Online]. Available: https://rpubs.com/pjmurphy/758265

[17] D. Singhvi, “Cluster Analysis Cluster analysis has a vital role in numerous fields we are going to see it in the banking...” Aug. 2018, [Online; accessed 12. Dec. 2022]. [Online]. Available: https://www.linkedin.com/pulse/20141010060815-25435731-cluster-analysis-using-r-banking-insight-study

[18] Y. Holtz, “Building heatmap with R,” Dec. 2022, [Online; accessed 12. Dec. 2022]. [Online]. Available: https://r-graph-gallery.com/215-the-heatmap-function.html

[19] S. G. Post, “Altruism, happiness, and health: it’s good to be good,” *Int. J. Behav. Med.*, vol. 12, no. 2, pp. 66–77, 2005.

[20] K. Baka, “The Impact of Altruism On Overall Happiness and Compassion,” *ResearchGate*, Apr. 2019.

Contributions

Order by the question number:

- Conceptualization, work questions - all authors;
- Q1, Clustering - L.C.;
- Q2, PCA - F.V.;
- Q3, FA, Dataset Introduction - J.A.;
- Q4, [Introduction or Time Series] - T.B;

All authors have read and agreed to the published version of the manuscript.