Use of Time per country related to Life satisfaction
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. 85044

Tijan Bah
Mestrado em
Matemática Aplicada
DMAT
Universidade de Aveiro
Aveiro, Portugal
N.Mec. 111910

***Abstract*—A study of life satisfaction per country, per GDP, per year and per geography. The idea behind the study was to to uncover habits, and patterns that we could serve as guide lines for Portuguese people to try to increase the life satisfaction.**

To better understand this dataset since it is multi dimensional, we have used clustering, principal components analysis, factorial analysis and time series. Despite being very ambitious on the tool set, was interesting that most of the tools reached different points for the conclusion.

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

January 8, 2023

1. Introduction

A good society is first of all a livable society and the livability of a society manifests in the life satisfaction of its members. Therefore improving society requires an understanding of life satisfaction which is Satisfaction with one's life implies a contentment with or acceptance of one's life circumstances, or the fulfillment of one's wants and needs for one's life as a whole. Life-satisfaction is about quality of life, yet not all quality of life is about life-satisfaction. One can have a good life but not be satisfied with that life or be satisfied with a not so good life. To help us understand the difference the four qualities of life are distinguish in the paper [1]. In a nutshell life-satisfaction is our subjective appreciation of our life as a whole. The synonyms are happiness and subjective well-being.

One of the ways life satisfaction is measure is using questioning. Questions on life-satisfaction can be asked in various contexts; in clinical assessments, in life review questionnaires and in common survey interviews. The questions can be posed in different ways; directly or indirectly, and by using single or multiple items where response on a single question is scale from 0 to 10 in most cases. The main contributing factors to life satisfaction are not completely understood yet, and the weight they are given by each individual varies which is a common problem to researcher in this domain.

Life satisfaction has been linked to many advantageous outcomes. For example, research consistently shows that individuals with high life satisfaction tend to have more positive social relationships, archived more occupational success and utilise their time to best of their interest for better life satisfaction [2] . If people are happier and more satisfactory with their life, they are willing to be more productive and spend more which boost gross domestic product of a country. According to [3] life satisfaction appears to be strictly monotonically increasing with GDP when one studies this relation at a point in time across nation.

1.1. Problem Motivation

Though developed and developing countries have developed their respective focuses pertaining to time use per country, there is a realization that an important contribution of a time use is that it gives a complete picture of the society by providing detailed information about how people spend their days (all 24 hours) on different economic and non-economic activities. In fact, it can be said that time use surveys is the only technique available to us at present that provides a comprehensive information on how individuals spend their time, on a daily or weekly basis. The motivation behind using the data set is to analyze the time use pattern of people in the selected countries in order to have a comprehensive information about the time spent by people on marketed and non-marketed economic activities.

The data sets used contains variables that have a high correlation among them which suited principal component analysis (PCA) and factor analysis (FA). However PCA allows us to compress our data and reduce the number of dimensions while retaining a lot of the information, the first principal component is always the direction that captures the most variance in the data, the second component capture the second highest variance and so on under the constraint that all components are always orthogonal and clustering is use to easily identify the similar countries. With FA we want to study the unobserved variables that may combine with observed variables to affect outcomes of the analysis and it will help us to to determine the number of factors that explain the correlations among a set of variables.

The economics and non economics contribution of man which depend on time, accounts for the gross domestic product (GDP) per capita of a country. GDP is the total value of all goods and services produced within a country's borders during a period of time, usually annually. According to [3] who claim that there is a positive relation between GDP and life satisfaction in developed countries and [4] who provide some evidence of no relationship between GDP and life satisfaction even for developing countries.

The scientific debate on the relation between Gross Domestic Product (GDP) and life satisfaction is still open but we are interested on the relationship between GDP per capita of a country and life satisfaction over certain period of time which have little literature's. The purpose of using this data set is to study how life satisfaction varies given different countries GDP per capita at a specific period of time.

Finally we want to study the trend of life satisfaction in Portugal and other countries with the motivation to find the countries that most change over time. In order to these we used the data set "Happiness.Cantril.Ladder" to compare countries life satisfaction over time and a special focus was given to Portugal to know how happiness evolve in Portugal over time.

2. Methodology

2.1. Clustering

Cluster analysis encompasses a set of exploratory data analysis techniques that try to find groups among a given set of data points. These techniques all have the same goal of finding groups whose members are the most similar to each other, all the while trying to be the most different from the members of other groups. Several types of clustering algorithms exist but we will only make use of Hierarchical Agglomerative and Centroid (k-means) clustering algorithms. Both algorithms are iterative, *ie*, the clusters are refined through several iterations of the algorithms.

In agglomerative hierarchical clustering, the algorithm starts from *n* clusters (*n* being the number of data points, every data point is treated as a cluster with one member) and groups the two closest ones on each iteration. The metric used to calculate the distance between clusters is chosen on a case by case scenario. Since we are trying to cluster individuals we opted to use the Euclidean distance as the distance metric. Another important part of this type of clustering is choosing the agglomeration criterion. Two criterions were considered, them being the single-linkage and the complete-linkage. When using single-linkage the algorithm aims to minimize the distance between clusters, thus creating a smaller number of clusters. Using complete-linkage leads the algorithm to maximize the distance between clusters, making the clusters found more compact.

The clusters produced by this algorithm can be visually evaluated through a graphic known as dendrogram.

Unlike hierarchical clustering, in the k-means algorithm the number of desired clusters is known beforehand. There are several tools that help finding the ideal number of clusters k for any given dataset. For this work we made use of the *Nbclust* function provided by R.

After knowing *k*, the algorithm initializes the centroids by choosing *k* random points from the dataset. Since the points are chosen randomly the results provided may not represent the optimal way of clustering the data. In order to minimize the effects of this randomness, most libraries offer the option to do several reruns of the algorithm with different starting points. Following the initialization, each point is assigned to the closest cluster and the centroids are recalculated. This step is repeated until no change in the centroid occurs between two consecutive iterations.

In order to evaluate the results provided by the algorithm we used an internal (average silhouette) and an external (ARI) validation metric. The average silhouette evaluates how well the data points fit into the cluster that they were assigned. This is done by calculating their average distance to their own cluster in relation to the average distance of the other points to their respective clusters. This value can range from [-1,1] and the higher the value the better. ARI (adjusted random index) measures the level of agreement between the clusters obtained and the clusters formed by creating random sets with the points. To put it simply, ARI measures how much better the obtained clusters are in relation to clusters formed by randomly picking points from the dataset. This metric usually ranges between 0 and 1, the higher the value the better.

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) [5], [6]. 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 [7]. 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 - $\hat{\Psi}$) [8]:

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

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 [9] 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 [10], they simply aggregate the results of [11] 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.

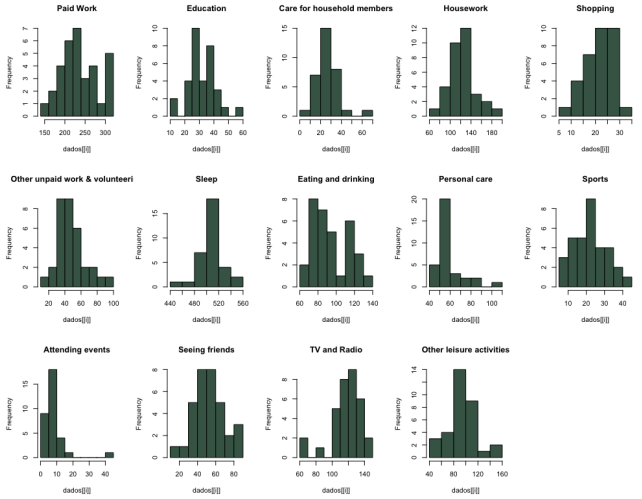

Figure 1. Distributions for every column of the data.

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. The histrogram of the Time used per category can be seen in 1, as we can see their distribution is close to a gaussian one.

Also returning to the initial article he cited [12] 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. Through the text we will use both terms.

We found the article [13] 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 [14] 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.

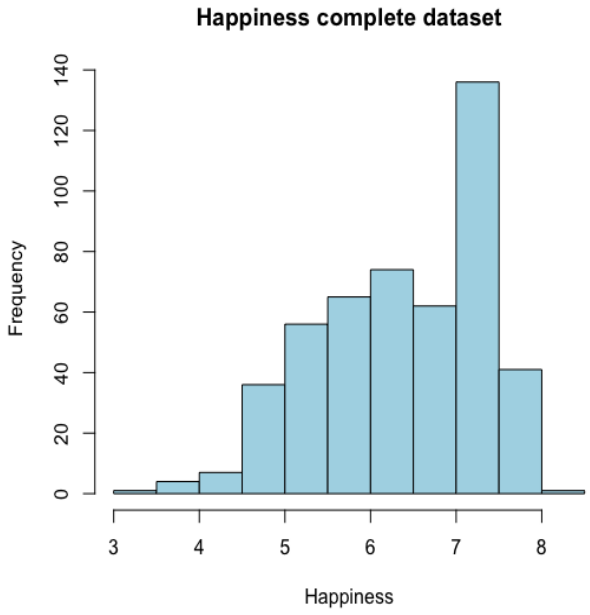

Figure 2. Distribution of the answers on happiness for the complete dataset

In order to try to draw conclusions from this dataset, we selected a subsection of it in order to only keep the countries that are present on the "Time Spent" dataset. From Figure 2 we can observe that the data contains a slight skew to reports of higher life satisfaction. This might be explained by the fact that the countries present on this dataset are mostly developed countries.

3.3. Life Satisfaction vs Geography per year

Lastly, and also from the article [13] 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 [15]. 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 [16], [17], [18]. 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. Results

4.1. Q1: How does life satisfaction vary per GDP?

To answer this question, we started by using the "Life Satisfaction vs GDP per capita" dataset. This dataset contains information for a time span of approximately 10 years.

Using the new dataset, we decided to first cluster the countries using k-means. In order to find out the ideal number of clusters we used *NbClust*, a function provided by R. This function runs every combination of number of clusters, distance measures and clustering methods and then returns the results of the majority vote on the optimal number of clusters for the provided dataset.

We tested for between 2 and 15 clusters and obtained that 3 would be the best number of clusters. The clusters were calculated using 20 different starting points to minimize the impact of the random aspect in choosing the starting points.

Both the ARI and the average silhouette metrics had a value of 0.62, meaning that the clusters found had acceptable structures. However, due to the huge amount of data points the visualizations produced were not legible. After analysing the clusters, we could conclude that countries tend to be assigned to the same cluster, independently of the year. This makes sense, since 10 years is a relatively short time span and we can't expect the GDP and life satisfaction to vary too much on this period of time.

As such, we represented each country by only one data point. In this case, we opted to use the data pertaining to 2020, as it was the most recent. After running the data through *NbClust*, the ideal number of clusters was once again 3. The new clusters have an ARI and average silhouette of 0.53, keeping them with acceptable structures.

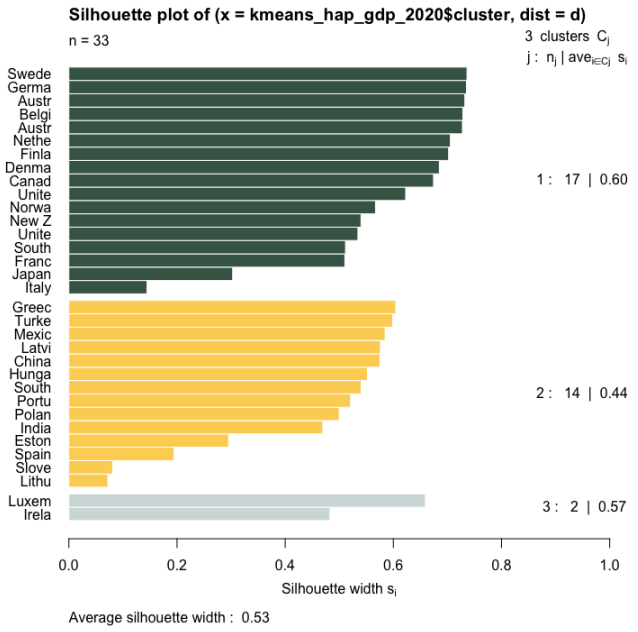

Figure 3. Silhouette plot of the clusters

On figure 3 we can see that cluster 3 only contains 2 countries, Luxembourg and Ireland, while clusters 1 and 2 contain the rest of the of the countries and are similar in size. Cluster 2 is mostly formed by eastern European countries plus the Iberian peninsula. Cluster 1 contains all of the central and north European countries.

In order to find out more about the relationship between life satisfaction and the GDP per capita we decided to plot the data points, while also representing the clusters found by k-means. Since there is a big difference between the GDP of some countries, this values were logged for easier interpretation. Figure 4 clearly shows that there is a linear relationship between life satisfaction and GDP, with the happiness increasing as the GDP increases. Cluster 2 is the exception, containing the two countries that despite having the biggest GDP are not the happiest. We can also see that cluster 2 is composed by the countries with the lowest GDP and cluster 1 by the countries with highest.

In order to see the distribution of life satisfaction, we decided to plot a histogram with the relative frequencies of the answers reported in each cluster, as seen on Figure 5. The results were consistent with the ones seen on Figure 4, with clusters 2 and 3 containing mostly positive answers, while cluster 1 has a concentration of answers on the lower and mid range of life satisfaction. Furthermore, we can see that cluster 3 concentrates the answers between happiness values of 6.5 and 7.5, meaning that the general population of this cluster feels really happy. However, it is cluster 1 that contains the happiest overall country, meaning that money is not the only ingredient for happiness.

Finally, we decided to cluster the data with a hierarchical clustering algorithm, using the euclidean distance and a furthest neighbour linkage criterion. Through the analysis of figure 6, we can easily identify 4 clusters. However, this clusters mostly coincide with the ones found by k-means, with the difference that China, India, South Africa and Mexico were removed from cluster 2 to create a new cluster on their own.

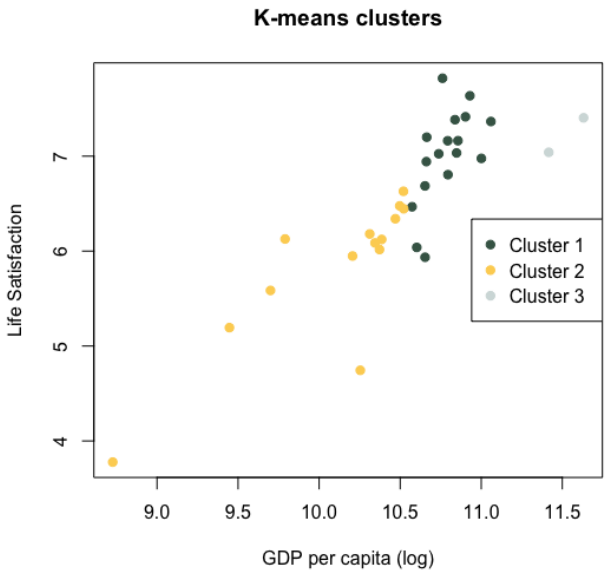

Figure 4. K-means clustering results

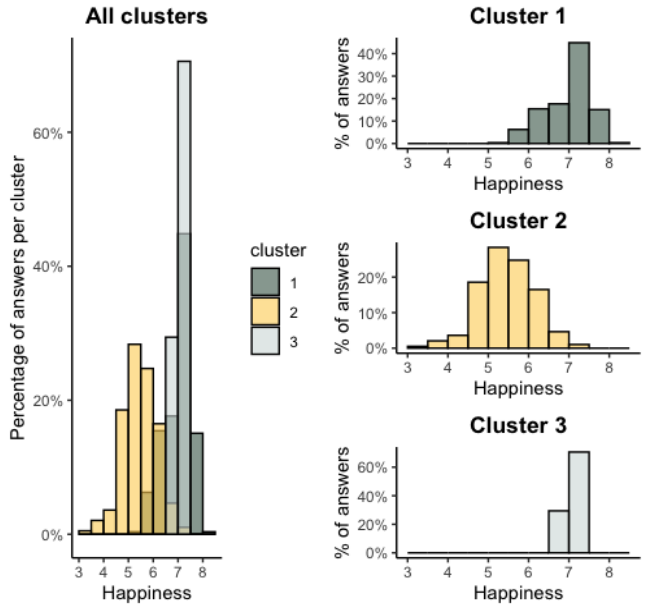

Figure 5. Distribution of the answers on happiness on the clusters

4.1.1. How is life satisfaction affected by how we spend time?. We were curious about a possible relationship between how the countries spend their time and their happiness and GDP. To study this possible relationship, we tried to cluster the countries based how they spent time, and plot them on the happiness versus GDP graph.

Through the *NbClust* function we found that 4 would be the ideal number of clusters and proceeded to cluster the dataset through k-means. However, the results were not very satisfactory, with the ARI and average silhouette values being 0.18. This means

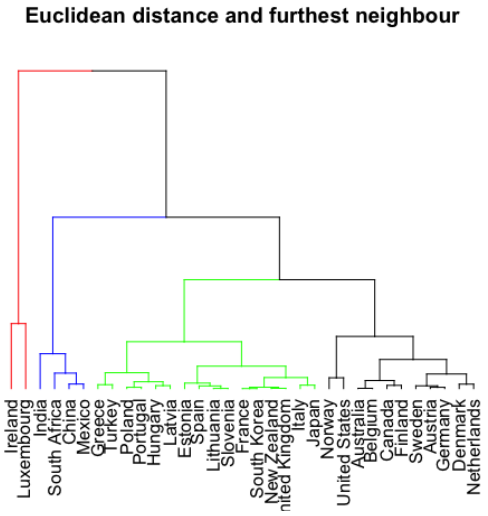

Figure 6. Hierarchical clustering based on Happiness and GDP

that the structure of the these clusters is very weak. Despite this, we plotted them on the happiness versus GDP graph. As expected, it was not possible to find any connection between how the countries spend time and their happiness and GDP.

Being faced with disappointing results, we tried to cluster variables of the "Time Use" dataset in order to compare them to the results from PCA and Factor Analysis. We opted by hierarchical clustering, using Pearson correlation as the distance metric and furthest neighbour as the linkage criterion. Similarly to what is discussed on the next sections, we could see 4 very distinct clusters on the dendrogram produced and a very small overlap between the clusters and the factors produced by FA. Therefore, we could not draw any conclusions from this clustering when comparing it to the results produced by FA.

4.2. Q2: PCA on Time use

For the use of PCA we will use the "Time Used" dataset already explained in the previous section. With the use of PCA, we wanted to compare and understand which countries were most similar in terms of how they spent their time. First, we did the PCA on the "Time Used" dataset, then we had to analyze how many components were needed to accurately study groups of countries with similar time spent.

As it is possible to see in the figure 7 second graph, as expected, the explained variance increases with the increase in the number of components. When looking at first graph in the figure 7, we can see that from component 5 the variance values, which the next components add, decrease in a different way. If we study about 5 components we can already obtain a variance of approximately 70%. With this we can say that studying 5 components would be enough to find solid conclusions. In the figure 8 it is possible to observe the components with greater importance, the blue numbers are the indices of the different countries and in red we have the different eigen vectors.

Through the graph in figure 9 we can see the most relevant eigen vectors for each component. When analyzing we see that for component 1 the most important vectors are "Shopping","House

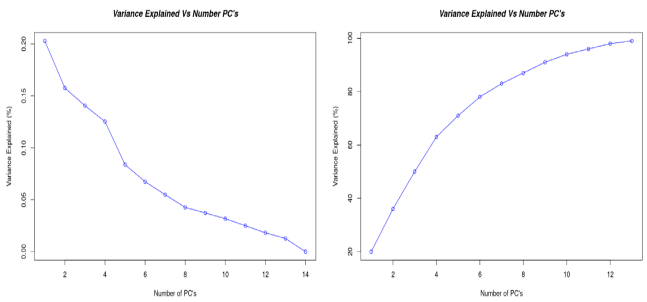

Figure 7. Variance Explained Vs Number of Principal Components

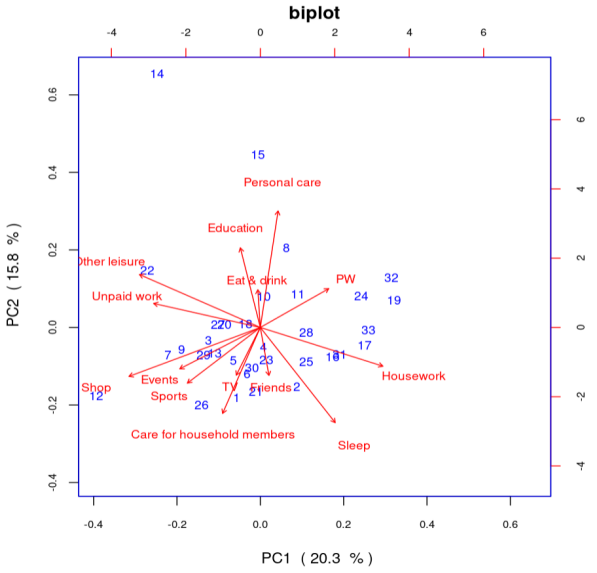

Figure 8. Biplot With Most Important Components

Work",”Other leisure activities” and ”Unpaid Work”. In component 2 ”Education”, ”Personal Care”,”Sleep”. For Component 3 we have ”Paid Work”, ”Eating and drinking” and ”Seeing friends”. In component 4 ”TV and Radio”, ”Attending Events” and Care for household members”. Finally in component 5 we have ”Sports”.

4.2.1. Cluster after PCA. In order to be able to see which groups of countries or which countries were most similar in terms of time spent, it was necessary, after using pca, to also perform clustering. To do the clustering it was necessary to know how many groups we should divide in kmeans (function used to make clusters). To find out how many groups we should make, we used an R function called NbClust, this function tests all the numbers of groups in a desired range, for the test we used values between 3 and 15, as you can see in figure 4. We can see the results obtained for the different values and the value 5 was the one that obtained the best results, which is what we will use in kmeans. Next, clustering was carried out with the kmeans function and 5 groups, as we can see in figure 10 that we have a graph with two components, in this case the most significant ones, as in figure 8 presented earlier, in this figure it is already possible to see the different groups created, these are presented with different colors thus distinguishing the groups.

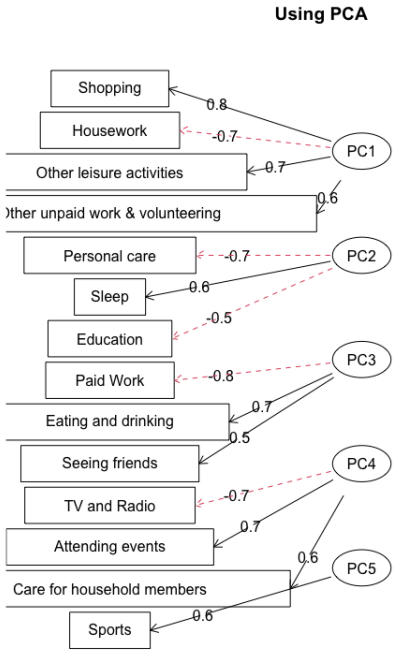

Figure 9. Graph with most important Eigen Vector for each PC

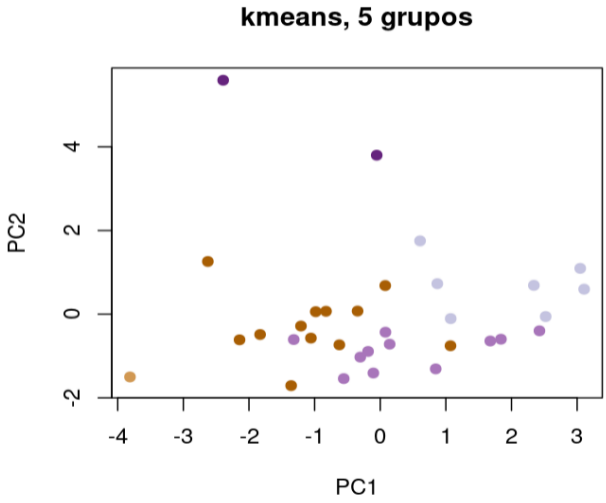

Figure 10. Biplot With 5 clusters in most important components

To obtain more information about the 5 clusters created, a silhouette plot presented in figure 11 was made, in this figure it is easier to see the different groups we can see that Latvia,USA,Canada,Lithuania, New Z, UK, Estonia, Poland, Austria, Australia are part of one group whe can see that 3 of the countries have negative values, which means they should or could be in another cluster. Ireland is part of other group which does make sense, as this cluster, in addition to having only 1 element, this element has value equal to 0.0.Japan and Korea are part of another group. Belgium, Greece, Spain, Denmark, Netherlands, Italy, Germany, Finland, Slovenia, Norway, Luxembourg, and Swe-

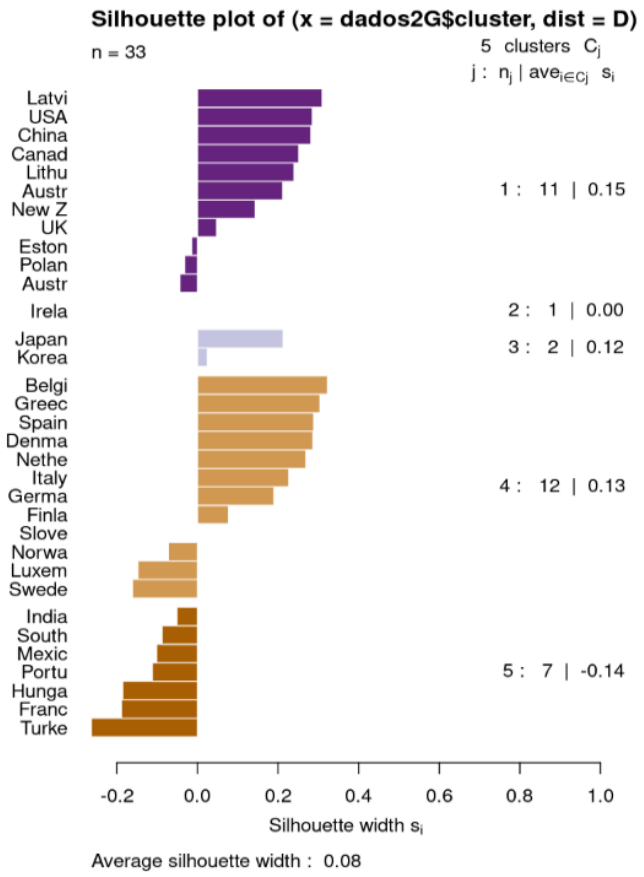

Figure 11. Silhoute Plot

den are part of another group, this group individually is the one that gets the best values, however, on average it doesn't get the highest values because some of the countries, in this case, Slovenia, Norway, Luxembourg, Sweden have negative values, meaning that they are missclassified and should be in another cluster, and even Finland, even if it does not have negative values, has a lower value compared to the rest cluster countries. And as the last group, we have the cluster where Portugal is located. This cluster is made up of India, South Africa, Mexico, Portugal, Hungary, France and Turkey. This is a group that makes no sense, because all its elements have negative values.

As we know Norway and Finland are one of the happiest countries so let's consider their cluster the happiest.

4.2.2. What can Portugal learn after the PCA. On Figure 12, we can see a heatmap with dendograms both for the Principal Components and for the countries, the data was scaled by column to make it easier to understand [19] [19].Besides this we have included 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 -100% (brown) to 100% (purple)).

When looking at the dendrogram, and using figure 9, we can see that Portugal is very far from Norway and Finland, especially when looking at component 1 and 5, with this it is visible that Portugal spends more time at House Work, Shopping, Other leisure activities, Unpaid Work and also in Sports compared to the happiest

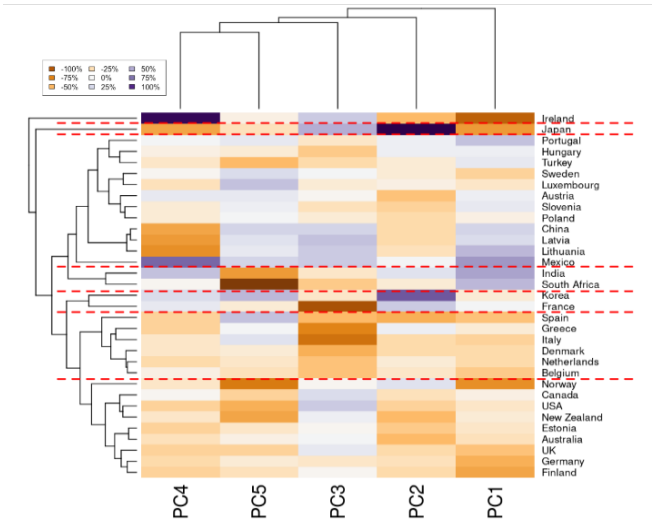

Figure 12. Heatmap of scores for every country and for every cluster with PCA

countries where they place a higher importance on Personal Care, Education, Paid Work and Seeing Friends.

With this, Portugal should dedicate less time to tasks, enjoy leisure activities as long as these are not excessive, as is the case in Portugal, and it should also change the way it thinks about education and work, it should set goals to achieve and try to learn always as much as possible. Since it's one of the best things we can do. Seeing friends often increases the feeling of sharing and unity in the different stages of life, which may also increase self-satisfaction/happiness.

4.3. 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. Yet since all the results are in minutes we decided to normalize the data, since the values where quite distinct.

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 [20], [21] 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, where we already have a 60%, which is acceptable.

TABLE 1. RESULTS OF THE KMO 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

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 [21] the 0.6 results that we obtained is considered mediocre. We obtained better results of MSA by reducing the categories even further 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:

$$H_0 : cor(matrix) = identity$$

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.

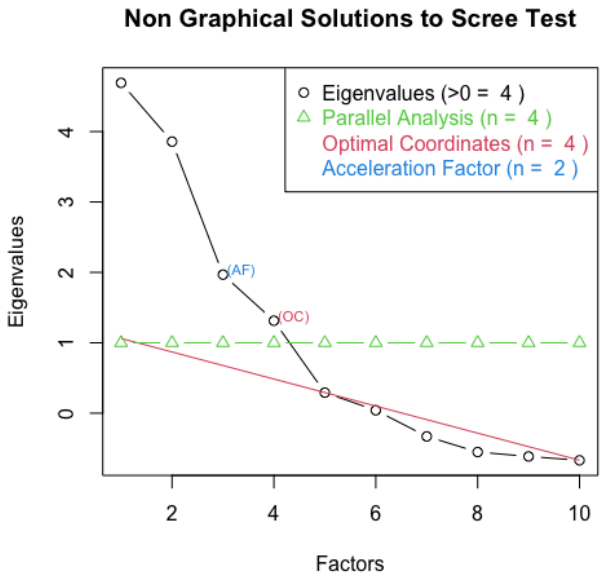

Figure 13. Scree Plot for the FA.

After this we do a plot a Scree plot to decide on the number of factors to use, this can be see in Figure 13. 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:

- Based on PCA - might result in similar results to what we have on the top section;
- Based on Maximum Likelihood methods - require a multi normal distribution for every column of the data;

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.

We decided to make both extraction, and choose one rotation for each. For PCA we did some tests but the rotations make little to no effect so we decided to stick to no rotations, it is possible to view the end diagram on Figure 14.

For the maximum likelihood, we haven't proven any Gaussian distribution on the data, yet experimentally it makes sense that the data will be Gaussian, or at least close to Gaussian as we can see in Figure 1.

This being said we decided to use as rotation for the maximum likelihood method, quartimax but we obtain lower loadings with it for some of the factors (we considered low when the $|$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.

In Figure 14 we can see both the components of the factors if we use PCA and if we use maximum likelihood methods (left and right respectively), as we can see that the right one is easier to explain, while the diagram on the left combines multiple data in the first factor, and leaves the last factor with just one component. As such we decided to use the one on the left, where it is easier to explain the factors, and the data columns are more distributed across the factors.

The graph on the left from the Figure 14 and on Table 2 we can see how the the various observable contribute to every factor. 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.

Note that could increase the cutoff of Tableloadingsfa in order to have a cleaner table, but we decided to use a cut of of 0.4 since more than that would remove the loading of personal care.

- Factor 1 - An isolated factor, here the important tasks that most of us think of as responsibilities almost are in

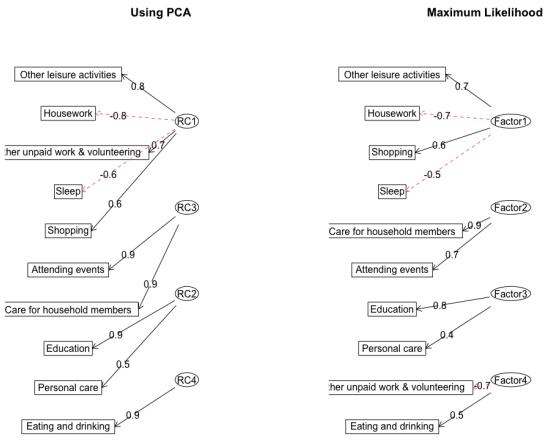

Figure 14. How every observable contributes to every factor using both a maximum likelihood extraction and a principal component extraction and 4 factors. Note: the red dotted line means a negative contribution.

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		0.44

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.

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

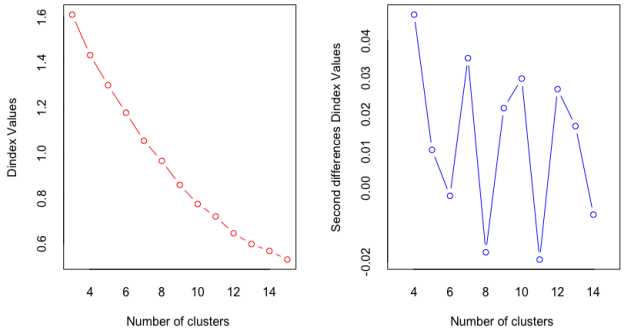

Figure 15. Determine the best number of clusters to use, using a majority rule.

4.3.1. 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 clusters that we want to use, for this we have used a majority rule [22]. The majority rule was studied from three clusters to 15, the results can be see in Figure 15, as we can see the recommendation for Kmeans was to use 3.

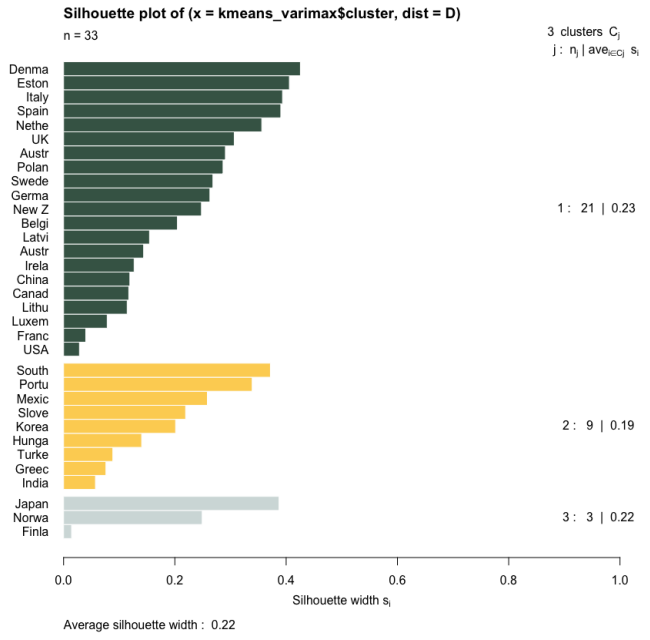

Figure 16. What countries belong to each cluster, using 3 clusters.

In Figure 24 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 (Norway and Finland are some of the happiest ones, while Japan has happiness values similar to the west of Europe, so on average they are a happy cluster).

4.3.2. 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.

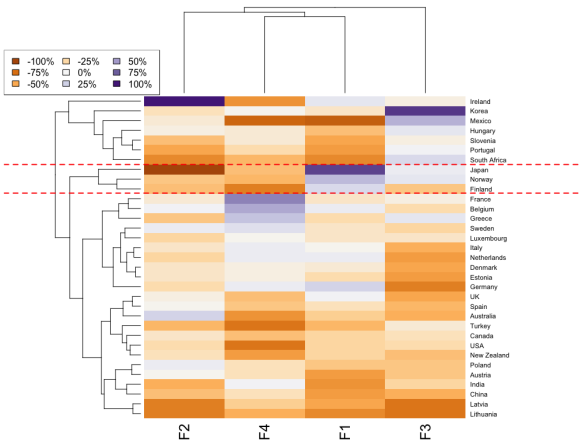

Figure 17. Heatmap of scores for every country and for every cluster after FA.

On Figure 17, 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 [19] [19]. 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 -100% (brown) to 100% (purple)).

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 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 [23], [24]. 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 (responsibilities) we can all increase our life satisfaction.

4.4. Q4: Evolution of Happiness for Portugal

From Figure 18 we can see that life satisfaction in Portugal exhibit a continual increase and decrease pattern. The maximum life satisfaction was attained in 2017 with an index of 6.342 and rank 13 in the world according to quality life index. In this year

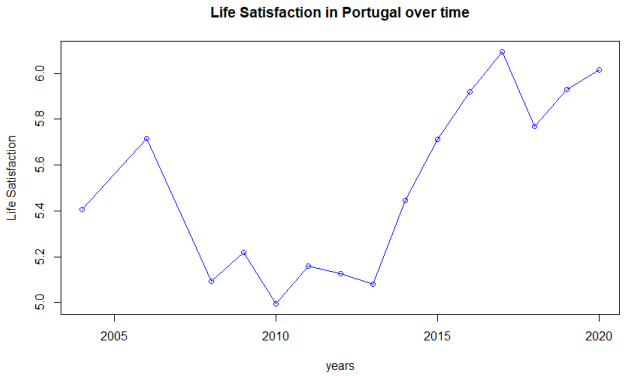

Figure 18. Life satisfaction over time for Portugal

the country’s GDP and GDP per capita was ranked 46 and 42 respectively according to IMF, in the same year there was many international festival events featuring world renounced musician’s like: Lil Wayne, Sean Paul, Mark Garr, Dua Lipa among others could be some reason for such index. The index was very low from 2007 to 2009 which can be associated to the world economic recession in 2007 [25], outbreak of Acute Norovirus Gastroenteritis in Porto in 2008 [26] and 2010 was the lowest index which can be associated with Portugal financial crises [27] at these period.

4.4.1. What are the countries that most changed. We answer this question on both descriptive and exploratory statistic method. In Figure 19 we see 6 countries from 6 continents respectively and the evolution of life satisfaction in each country. Canada and New Zealand dominating which can be associated to economical well being like: GDP, GDP per capita ranking while Ghana being the lowest. During 2008 global economics recession, all 6 countries have a decreasing trend except Ghana. This could be due to social factors like cultural activities because there was less work and more for fun. During Covid19 we can see all the 6 countries has a reduced index due to global viral outbreak.

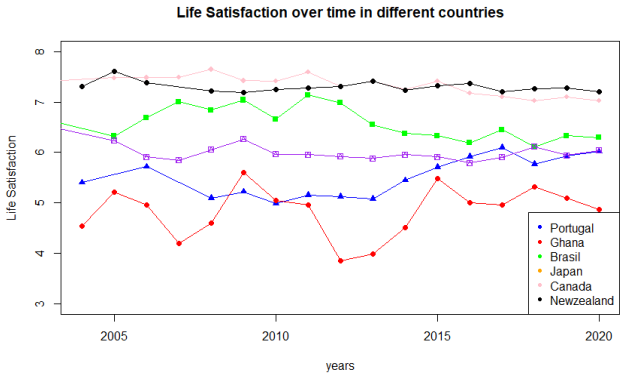

Figure 19. Life satisfaction of different countries

In order to see how most countries will change in relation to life satisfaction over time we extracted 5 countries from each continent from the data set available. Given the new data frame obtained we checked the correlation between the variables and computed the

average correlation which is 88% which show that the variables are highly correlated and hence PCA can be used. In Figure 20 we can see new variables are form from the highly correlated variables as principal components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Ghana	-0.7530616	2.43096040	2.724018976	-0.686538652	-0.24721737	0.0652249111	0.04936803	0.184998998
Brasil	2.6923930	0.72711857	-0.129017456	0.276995272	0.52640071	0.1160612317	-0.09892531	-0.129398427
New_Zealand	5.4802339	-0.26527379	-0.238240580	-0.105075839	-0.43907634	-0.1340408111	0.11669426	-0.111981947
Canada	5.6504965	0.23406011	-0.346762575	0.158591240	-0.21390553	-0.0004055851	-0.08993447	0.087983670
Japan	0.2262433	0.05840455	-0.142639401	-0.250709870	0.03615535	-0.2768290522	-0.18581154	-0.019382549
Portugal	-1.7163074	-1.22217486	-0.266209884	-0.388428603	-0.12234752	0.2238524675	0.11733263	-0.052436230
Benine	-7.3763276	-2.20980097	0.468581089	0.283439483	-0.29621390	0.7206848576	-0.02191419	-0.106916826
Denmark	6.8834586	-0.40557391	-0.459370627	-0.112123877	-0.33880566	0.0320624316	-0.20837601	-0.012215588
India	-6.2871133	1.83072846	-1.157546095	-0.025582191	-0.05836681	-0.4978631230	-0.14201966	0.247886304
USA	4.5806584	0.04555959	-0.520426791	-0.194827344	-0.18106248	-0.0746322045	0.07733837	-0.098412878
Argentina	1.4224850	0.85434136	0.328761991	0.236992088	0.44031230	-0.3255614801	-0.19563824	-0.172674263
Austrial	5.3711354	-0.13787320	-0.141342416	0.079743246	-0.08932403	-0.0997452699	-0.08788396	-0.002201273
Bostuwana	-7.8212562	1.96315232	-1.772122193	-0.952965984	-0.18336334	0.2879561329	0.25520680	-0.105764381
Estonia	-0.9561589	-1.29364767	0.456709748	-0.621596316	-0.02740471	-0.2318705134	0.24435363	0.135536963
Indonisia	-2.6447006	0.04915698	-0.005016349	0.183348082	0.63120263	-0.0844915126	0.32711495	-0.243037267
Mexico	2.9479348	0.94502096	-0.227641706	0.273015987	0.33810483	0.7804012823	-0.14008467	0.196983258
Colombia	1.2172214	0.62576397	0.067077577	0.194359256	0.22405115	0.1883972391	-0.06332960	-0.232816272
Burkinafasso	-1.6908584	-1.15645775	0.410447458	-0.068799952	-0.14047814	-0.2625546042	-0.80833337	-0.239386452
Franch	2.8555805	-0.23187905	-0.235296018	-0.392622468	-0.16056975	0.1573351514	-0.29116570	0.155905245
Kazakstan	-0.2395665	-0.60681979	0.226299203	-0.708583598	0.04256606	-0.0124948298	0.24648164	-0.256579713
Costarica	4.8326245	0.05376933	-0.262403599	0.534212119	-0.14386551	-0.0435703690	0.09409213	-0.154847240
Peru	-0.9975628	-0.14282229	0.590809400	0.009357139	0.48646769	-0.0027559889	0.08365552	-0.117272387
Cameroon	-4.8611739	-1.05311268	0.303111158	0.165075319	0.02377842	-0.3954809957	0.08418937	0.110517428
Germany	3.8037344	-0.83532420	0.506196141	-0.135526440	-0.01438770	-0.1552874272	0.40834200	0.126097359
Lebanon	-4.6353433	1.45124756	0.370621933	1.661969635	-0.61392211	-0.0260029534	0.24617396	-0.071840634
Hondarus	-1.8038802	-1.69462150	-0.571226871	0.538761486	0.24802349	-0.1779785195	0.21765976	0.430975941
Paraguag	-1.1811258	-0.01390251	0.022627886	0.047520783	0.27324926	0.2295895345	-0.23458632	0.450279162

Figure 20. Principal Components

In Figure 21 we can see see all the chosen countries and the length of each variables. Variable 11 has the highest length which show that it is the most important eigenvectors while variable 8 has the lowest length which makes it it the lest important variable. The angle between variable 7 and 10 is the highest which is an acute angle and it show less correlation between the variables.

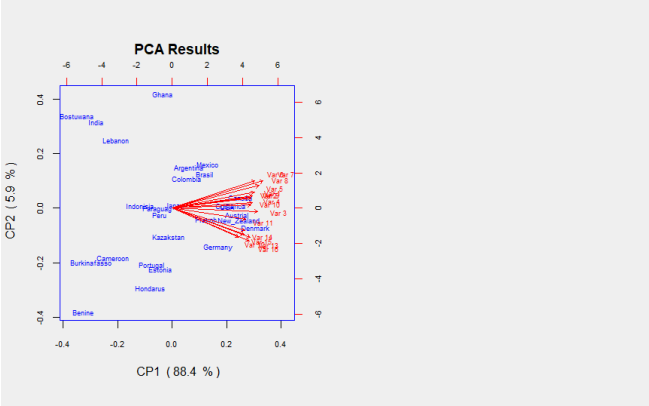

Figure 21. Bi-plot

In Figure 22 we can see the optimum number of cluster as 6. This is obtained using NbClust package which provide 30 indices for determining the number of clusters and proposes to the user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods. Preference are given to k means clusters and dendrogram clusters because those are the technique done in class.

Since the optimum number of cluster obtained from the NbCluster package is 6 we decided to used 6 clusters for the k means clustering. In Figure 23 we can see 6 clusters with different colors and each cluster is initialise by a unique centroid . As mentioned earlier ”what are the countries that most change ?” will be address by discussing the changes that can be seen from this results:

- In cluster 1 we can see that most of these countries are highly industrialised with a very high life satisfaction index, GDP and GDP per capita. People in these countries are more satisfied with economical activities and hence need to change to social,cultural and religious activities for better diversity.
- In cluster 2 and 3 most of these countries have a large number of the population following a specific religion which prompt us to say they are satisfied due to religious activities. The life satisfaction of most of these countries is high and need to change to social activities.
- In cluster 4 all those countries are battling with environmental hazards during this period and life satisfaction index is mostly on average compare to others countries. These countries should address environmental problem for better life satisfaction.
- In cluster 5 all those countries are under a dictatorship leader. These countries have a very low index over these period and hence changing the leadership style can be a way to improved life satisfaction.
- Cluster 6 has only one country and a low index. These people might be more satisfied with cultural and social activities distancing them from most of the clusters. The country can invest more economic and industrial activities to have a better utility.

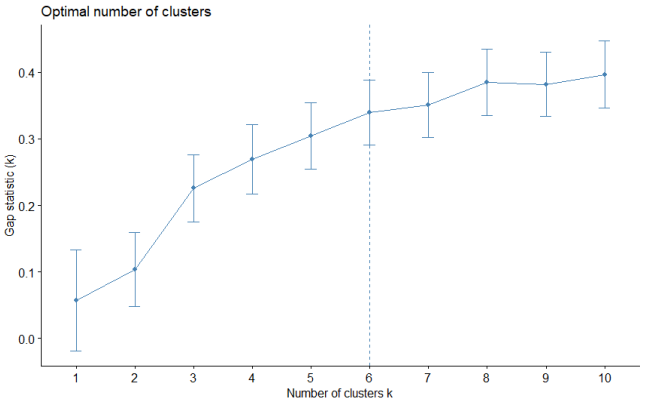

Figure 22. Neighbour Cluster Optimisation

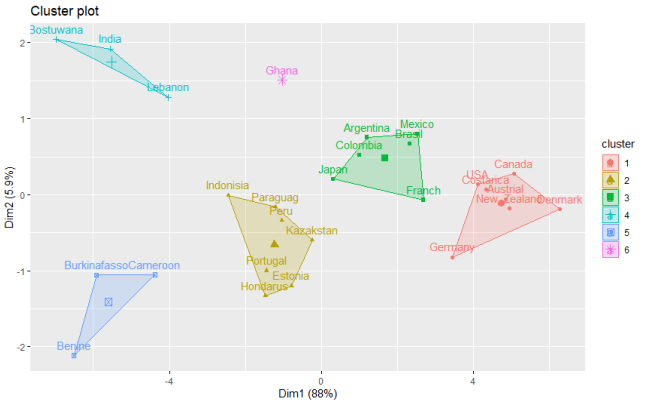

Figure 23. K Means Cluster

In Figure 24 the diagram represents the hierarchical relationship between countries displaying the Pythagorean distance

between each pair of sequentially merged countries. The distance between countries represent how dissimilar those countries are and the height of the block represent the distance between clusters while the clades branches are arrange according to how similar or dissimilar countries are. Given the same data we are able to obtained the same clusters with no miss-classification.

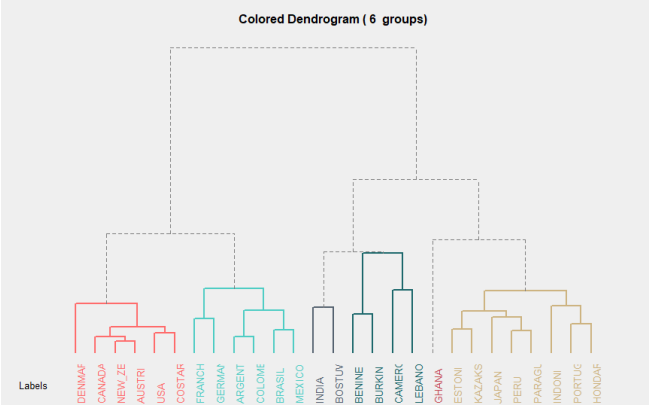

Figure 24. Dendrogram Cluster

5. Conclusions

After studying the effects of the GDP per capita on life satisfaction (Q1), we reached the same conclusion as in [3], ie, there is a connection between the increase in the GDP per capita and the perceived life satisfaction of the population of a country. Through clustering, we were able to distinctly group the countries and found the GDP was a dominant variable in creating the clusters, as is represented on Figure 4. Although the clusters obtained by k-mean and hierarchical clustering were slightly different, we could still see a very clear division in the self reported happiness of the countries of central and north Europe and the rest of the continent. In both algorithms, Luxembourg and Ireland proved to be the exception to the rule, adding strength to the argument that after a certain threshold money stops radically increasing happiness. The attempts to find a connection between how countries spend their time and how they fit into the GDP vs Happiness relationship proved to be fruitless. In the future, a systematic and more in-depth study might uncover a stronger connection between the two.

Using PCA in Q2 it was possible to conclude that Portugal has to change the time he spends on tasks and leisure activities and try to take advantage of part of that time to dedicate to new goals and continuous learning. Seeing friends often increases the feeling of sharing and unity in the different stages of life, which may also increase self-satisfaction/happiness.

From Q3 we can see that this data, suggests that Portugal should increase the altruism habits, while reducing the sense of responsibilities, i.e., trying to change our mindset about our tasks and errands. Despite none of this findings being groundbreaking it is interesting that this emerged from the data as the first two points to work from.

In Q4 we learn that recession, financial crisis and global disease outbreak could be some of the reason why life satisfaction index is low. Moreover we can see that life satisfaction assessment is based on different activities and a change toward those activities can improved their happiness but the yardstick depend on the individual self assessment.

Contributions

Order by the question number:

- Conceptualization, Work Questions, Conclusions - all authors;
- Q1, Clustering - L.C.;
- Q2, PCA - F.V.;
- Q3, FA, Dataset Introduction, Abstract - J.A.;
- Q4, Introduction - T.B;

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

References

[1] R. Veenhoven, "The four qualities of life," *Journal of Happiness Studies*, vol. 1, pp. 1–39, 02 2000.

[2] P. William and D. Ed, "The satisfaction with life scale and the emerging construct of life satisfaction."

[3] B. Stevenson and J. Wolfers, "Economic growth and subjective well-being: Reassessing the easterlin paradox," National Bureau of Economic Research, Working Paper 14282, August 2008. [Online]. Available: http://www.nber.org/papers/w14282

[4] M. L. A. S. M. S. O. Easterlin, R. A. and S. Zweig, "The happiness-income paradox revised."

[5] 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

[6] 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.

[7] 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

[8] "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

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

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

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

[12] 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

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

[14] "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/

[15] "How we're funded." [Online]. Available: https://ourworldindata.org/funding

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

[17] "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

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

[19] 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

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

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

[22] 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

[23] 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.

[24] K. Baka, "The Impact of Altruism On Overall Happiness and Compassion," *ResearchGate*, Apr. 2019.

[25] E. Musvoto and D. Boshoff, "The 2008 economic recession," 10 2015, pp. 65–78.

[26] J. Mesquita and M. Nascimento, "A foodborne outbreak of norovirus gastroenteritis associated with a christmas dinner in porto, portugal, december 2008," *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin*, vol. 14, p. 19355, 10 2009.

[27] K. Hausken and J. Welburn, "Assessing the 2010–2018 financial crisis in greece, portugal, ireland, spain, and cyprus," *Journal of Economic Studies*, vol. ahead-of-print, 12 2020.