Bucharest University of Economic Studies

Faculty of Cybernetics, Statistics and Informatics

Quality of life analysis

using PCA

Student: Andreea Rizea

Groups: 1099

Series: G

Bucharest

2021

Table of Contents

[Introduction 3](#_Toc92651661)

[About the dataset 3](#_Toc92651662)

[Data analysis approach 4](#_Toc92651663)

[Interpretation of the results 7](#_Toc92651664)

[Correlogram of factor loadings 7](#_Toc92651665)

[Correlogram for the quality of points 7](#_Toc92651666)

[Correlogram of betas 8](#_Toc92651667)

[Correlogram of commonalities 9](#_Toc92651668)

[Correlation circle 10](#_Toc92651669)

[Conclusion 12](#_Toc92651670)

[Bibliography 13](#_Toc92651671)

# Introduction

Quality of life is defined by the World Health Organization as "an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns".

Standard indicators of the quality of life include wealth, employment, the environment, physical and mental health, education, recreation and leisure time, social belonging, religious beliefs, safety, security and freedom.

In this project, I referred to the quality of life given by 4 main aspects: healthy lifestyle (emphasized by smoking, doing sport, eating healthy or drinking alcohol), the quality of the environment (air pollution), interaction with others (participating to cultural events or even family meetings) and work satisfaction (the quality of interactions between individuals and their colleagues, how individuals can adapt their pace at work).

Further on, the aim of this analysis is to observe how these 4 aspects influence each other and, finally, how do they contribute to an equilibrate life.

# About the dataset

All the data found in the dataset of this project is extracted from Eurostat. The data was collected from some of the European countries during the 2014-2015 period. The unit of measurement is percentage and it reflects the proportion of the population of a country with a specific characteristic.

It is known that nowadays the quality of life is mainly influenced by a healthy lifestyle, social interactions, living environment and work quality. For this reason, I chose to analysis the dynamic of these factors and how do they relate to each other.

For each of these 4 categories, I selected some of the most representative variables.

A healthy lifestyle is characterized mostly by active movement of the body, eating high-nutritious food, not drinking alcohol as little as possible or, preferably, not smoking. For this, I extracted data related to these characteristics for computing the analysis. Relating to sports, I consider the individuals who practice aerobic sport weekly on an annually period. For proving that an individual has a healthy diet, he should consume, besides other aliments, fruits and vegetables. So, I take into consideration for each country the percentage of the population who eats between 1 to 4 servings of fruits and vegetables daily.

In order to ensure a healthy lifestyle, it is claimed that a person should avoid smoking and drinking alcohol as much as possible. As a consequence, I take into account the percentage of the population who smoked daily and also those who experiences heavy episodic drinking less then once a month during an year. Also, I consider the percentage of the population who self-evaluates its health status as good.

Regarding the living environment, I take into consideration the percentage of individuals who are exposed to air pollution considering the quantity of particulates. It is proven that the quality of the air that a person inhales regularly could influence his health status. Pollution could cause respiratory diseases or even mental health problems which would lead to a lower quality of life.

In order to keep a balanced life, it is recommended to involve ourselves in multiple recreative activities. This is one of the reason why I chose to integrate this part in the analysis. Here I consider the percentage of the population who participates in cultural activities at least once in the last 12 months and also the individuals who get together with their friends every week.

Lastly, a pleasurable daily experience at work could influence the overall well-being of an individual. Variables such as the percentage of persons who are in good relationships with their colleagues at work and those who can influence their pace making it more flexible would help the analysis regarding this topic.

Here are the abbreviations used for each variable in the input data file:

* AEROBIC\_S – percentage of the population performing aerobic sports annually
* DAILY\_FV – percentage of the population who consumes daily from 1 to 4 servings of fruits and vegetables
* SPH\_GOOD – percentage of the population who self-perceives its health status as good
* DAILY\_SMOKE – percentage of the population who smokes daily
* HEAVY\_DRINKING – percentage of the population who experiences heavy episodic drinking less than once a month per year
* AIR\_POLL – percentage of the population who is exposed to air pollution by particulate matter (Particulates < 2.5µm)
* CULT\_ACTIVITIES – percentage of the population who participates in cultural and sportive activities at least once in the last 12 months
* FRIENDS\_MEET – percentage of the population (over 16 years old) who gets together with friends every week
* GR\_WORK – percentage of the employed persons who have are in good relationships with their colleagues at work
* CHOOSE\_WORK – percentage of the employed individuals who are able to choose their methods of work or to influence their pace of work

# Data analysis approach

As I mentioned earlier, I chose Principal Component Analysis (PCA) as type of analysis for this project.

Principal Component Analysis 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.

There are several steps which should be done when performing this type of analysis.

Firstly, I started by standardizing the values. The aim of this step is to avoid big differences between the values of the variables because it could lead into biased results and also could affect the computation of the variance.

Text

Description automatically generated

Figure 1: Standardization of the data

Further on, I continued with computing the covariance matrix. This procedure is used to understand how the variables of the dataset are varying from the mean with respect to each other and, moreover, to see of there is any relationship between them.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 2: Covariance matrix

The main focus should be on the darker areas. There, the correlation is higher.

For example, the percentage of individuals who smoke is correlated in proportion of 65% with the percentage of people exposed to air pollution. It can be said that if the percentage of smokers would increase with 1%, the percentage of those who are exposed to air pollution would also increase with 1%. One of the reason behind this could be that the cigarette smoke pollutes the air.

On the opposite, there is a weak correlation between those who practice aerobic sports and those who smoke daily. This means that the 2 variables are inverse correlated. Most of the time, people decide to start having a healthy lifestyle by stop smoking and start going to gym, for example.

In order to move on with the analysis, I computed the eigenvalues and eigenvectors.

Eigenvectors and eigenvalues are the linear algebra concepts that should be computed from the covariance matrix in order to determine the principal components of the data. The eigenvalues help to determine the number of principal components to be considered.

Geometrically speaking, principal components represent the directions of the data that explain a maximal amount of variance, so, the lines that capture most information of the data.

According to Kaiser criterion, when the data is standardized, the principal components who have the variance higher then 1 should be kept.

Chart, line chart

Description automatically generated

Figure 3: Explained variance of the principal components

In this case, here are the first 4 principal components should be kept in order to perform the analysis. So, most of the information contained in the data is retained by the first 4 principal components.

# Interpretation of the results

## Correlogram of factor loadings

Factor loadings refer to the correlation coefficients between principal components and the initial variables.

The highest the factor loading for each variable indicates the strongest relationship between the variable and the factor. In other words, the higher the factor loading, the better that variable is at measuring that factor.

Graphical user interface, application

Description automatically generated

Figure 4: Correlogram of factor loadings

The correlogram of factor loadings shows that the first principal components is computed by the participation in cultural events, doing sports, not drinking too much alcohol, having a flexible program at work or going out with friends.

At the same time, the second principal component is mainly composed by eating fruits and vegetables daily, whereas the third component is focused on the percentage of individuals who are in good relation with their work colleagues.

The forth principle component is represented by the percentage of the population who self-perceives its health status as good.

## Correlogram for the quality of points

The quality of representation of the variables on factor map is called square cosine. It indicates the contribution of a component to the squared distance of the observation to the origin. In other terms, it shows the importance of a component for a given observation.

Table

Description automatically generated

Figure 5: Correlogram for the quality of points representation on axis of principal components

It is clear that the first component is important for Finland, Luxembourg, Sweden, Hungary, Croatia, Bulgaria or Denmark. This means that most individuals in these countries practice aerobic sports, go to cultural events or are not having major episodic drinking in the last months.

In addition, the second principal component which is related mostly to eating healthy is considered in countries like Belgium, Czechia or Norway.

The most people who are in good relations with their colleagues at work are from Italy and Ireland, while the individuals who self-perceive their health status as good are from Iceland or Spain.

## Correlogram of betas

One of the reason for computing the betas here was to see how each country contributed to the variance of each variable. In more details, it can be seen which country is mostly characterized by each aspect of life.

Table

Description automatically generated

Figure 6: Correlogram of betas

The correlogram of betas shows that Bulgaria is one of the countries with the highest presence of the first principal component in its characteristics. It is followed by Finland, Sweden, Croatia or Denmark. In concrete terms, these countries have a strong combination of the principal components’ variables: participation in cultural events, doing sports, not drinking too much alcohol, having a flexible program at work or going out with friends.

The same applies to Poland or Norway which are highly described by the healthy daily diet. In addition, Italy seems to be the country with the best relationships at work, while Iceland self-perceives its health status as good.

## Correlogram of commonalities

In Principal Component Analysis, the commonalities represent the proportion in which each variable’s variance is explained by the principal components.

Graphical user interface, application

Description automatically generated

Figure 7: Correlogram of commonalities

So, the first principal component seems to explain the variances of doing sports, daily smoking, going out to cultural events or being able to make a flexible work program. Regarding the second principal component, it contributes to the variances of practicing sport, daily smoking, drinking too much alcohol rarely, participating to cultural activities, meeting friend or having a good work program.

The third principal component is responsible of the variances of doing aerobic sports, eating healthy, daily smoking, drinking alcohol rarely, good relationships with work colleagues and exposure to polluted air. Lastly, the forth main principal component explains variances for self-perceiving a good health status, drinking too much alcohol rarely, meeting friends, good relationships with work colleagues and eating healthy.

## Correlation circle

The correlation circle is used to represent the cloud of points of variables in order to prove the relationship between them.

If the variable has a low square cosine, it means that variable is represented close to the center of the circle. At the same time, variables that are closed to the center are less important to the first principal components.

Chart

Description automatically generated

Figure 8: Correlation circle for the initial variables in the space of C1 and C2

As the positively correlated variables are grouped together, it can be said that doing sports and going to cultural activities are influenced one by another. At the same time, there is a correlation between drinking alcohol rarely and being able to have a flexible program at work.

The variables who are positioned on opposite quadrans are negatively correlated. Those are eating healthy and daily smoking with drinking alcohol rarely, having a flexible working program and self-perceiving as healthy individual. In the same manner, exposing to polluted air has no correlation with meeting friends or having good relations with colleagues at work.

# Conclusion

Each individual is unique and for this reason the meaning of a high-quality life could differ from person to person. The characteristics that would contribute to an increased life satisfaction, physically and mentally speaking, could be living in an non-polluted area, quite smoking, eating more fruits and vegetables or even having the possibility to establish its own pace at work.

The principal component analysis technique helped in finding the main habits for increasing the quality of life among a series of different activities. Doing sport, daily smoking, not drinking alcohol too frequently, having a good relationship with colleagues at work or spending time with friends are one of the most valuable attributes of a high-quality life.

The countries in which these habits are prevalent for defining a high-quality lifestyle are Finland, Luxembourg, Sweden, Hungary, Croatia, Denmark, Belgium, Czechia, Norway, Italy, Ireland, Iceland or Spain. The majority of the these countries are positioned in the North part of Europe (Finland, Sweden, Denmark, Norway, Iceland) where it is known that the standard if living is higher than the one in the other parts of the continent.

To sum up, principle component analysis (PCA) was useful for identifying which factors represent the most valuable influence for a high-quality life and also in which countries the population is already guided by these habits.

# Bibliography

* <https://en.wikipedia.org/wiki/Quality_of_life>
* <https://ec.europa.eu/eurostat/web/main/data/database?p_p_id=NavTreeportletprod_WAR_NavTreeportletprod_INSTANCE_nPqeVbPXRmWQ&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view>
* <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
* <https://towardsdatascience.com/principal-component-analysis-algorithm-in-real-life-discovering-patterns-in-a-real-estate-dataset-18134c57ffe7>
* <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/>