ISSN (Print) 2307-6968, ISSN (Online) 2663-2209
Вчені записки Університету «КРОК» №2 (58), 2020

УДК: 334.012
DOI: 10.31732/2663-2209-2020-58-40-51

КЛАСТЕРНИЙ АНАЛІЗ ДАНИХ ЖИТТЯ В ЄВРОПІ

Taxa М.Д.
аспірант, ВНЗ «Університет економіки та права «КРОК», м. Київ, вул. Табірна, 30-32, 03113, Україна, тел.: (044)-455-69-82, e-mail: tahadaab@gmial.com , ORCID: https://orcid.org/0000-0001-6761-964X

CLUSTER ANALYSIS OF COST LIVING DATA IN EUROPE

Taha M.
postgraduate student, «KROK» University, Kyiv, st. Tabirna, 30-32, 03113, Ukraine, tel.: (044)-455-69-82, e-mail: tahadaab@gmial.com , ORCID: https://orcid.org/0000-0001-6761-964X

Анотація. *Досить багато досліджень, резюме, доповідей та статей про вартість життя у різних містах світу. Незважаючи на те, що ці дослідження багаті та потужні, публікуються лише остаточні результати, такі як рейтинг міст або зведення найважливіших та найменш оцінених міст. Ця робота спрямована на вивчення найбільш придатних для життя міст у Європі за допомогою кластерного аналізу даних про ціни на життя за 2019 рік. Я застосую метод групування за подібністю або різницею в алгоритмах кластерного аналізу, оскільки існують багато методів кластерної кластеризації, і вони є технічними або алгоритмом (Пов'язування), де результати різні, використовуючи найпоширеніші методи зв'язування для складання дендрограми, а саме: мінімальне або індивідуальне зв'язування, максимальне або повне зв'язування, середнє або середнє зв'язування , де ми помітимо, що міста можуть згрупуватися в інший кластер відповідно до методу зв'язування та географічного розташування, де цей тип блокового аналізу робиться, щоб допомогти компаніям або навіть людям визначити, яке місто переїхати, щоб зменшити вартість життя. Також уникати прийняття неправильного рішення у разі зміни наступного кварталу для конкретного міста. Результати дозволяють описати подібність та відмінності в структурі цін на товари та послуги в різних містах та європейських країнах. Результати аналізу показали, що в деяких європейських країнах існують певні подібності у вартості життя та об'єднання їх в один кластер з урахуванням різниці в їх географічному розташуванні. Це вказує на те, що склад кластерів залежить від методу зв'язування в аналізі. Ці результати можуть бути використані приватними особами чи установами для вибору найкращих європейських країн для життя шляхом порівняння вартості життя в них порівняно з наявним бюджетом.*

Ключові слова: *кластер, витрати на життя, місто, відстань, k-середні, k-медоїди та алгоритми ієрархічної кластеризації.*
Формули: 5, рис.:14, табл.: 2, бібл: 6

Annotation. *There are a lot of studies, research, summaries, reports, and articles on the cost of living in different cities around the world. Although these studies are rich and powerful, only final results are published, such as city rankings, or summaries of the most important and least ranked cities. This paper aims to study the most livable cities in Europe using the cluster analysis of cost-of-living price data for the year 2019. I will apply the method of grouping according to similarities or differences in distance and cluster analysis algorithms, as there are many methods of cluster clustering and they are technical or an algorithm (Linking), where the results are different using the most common linking methods to draw the dendrogram, namely: the minimum or individual linking, the maximum or full linking, average or average linking, where we will notice that cities can cluster in a different cluster according to the method of linking, and the geographical location, Where this type of block analysis is done to help companies or even people determine which city to move to reduce the cost of living. Where it is important to avoid making the wrong decision in case of changing the next block for a specific city. The results allow us to describe the similarities and differences in the price structure of products and services in different cities and European countries. The results of the analysis showed that there are some similarities in the cost of living in some European countries and grouping them into one cluster, taking into account the difference in their geographical location. This indicates that the composition of the clusters depends on the method of linking in the analysis. These results can be used by individuals or institutions to choose the best European countries to live by comparing the cost of living in them compared to the available budget.*

Key words: *cluster, living costs, city, distance, k-means, k-medoids, and hierarchical clustering algorithms.*
Formulas: 5; fig.: 14, tabl.: 2, bibl.: 6

Introduction. A lot of data can be gathered from different fields but this data is useless without proper analysis to obtain useful information. In this paper, we focus on one of the important techniques - Clustering. Data clustering is a method of grouping similar objects together. Thus the similar objects are clustered in the same group and dissimilar objects are clustered in different considered as an unsupervised learning technique in which objects are grouped in unknown predefined clusters. On the contrary, classification is supervised learning in which objects are assigned to predefined classes (clusters).

Data clustering is based on the similarity or dissimilarity (distance) measures between data points. Hence, these measures make the cluster analysis.

Also, when we use the dissimilarity (distance) concept, the latter sentence becomes: the high quality of clustering is to obtain low intra-cluster dissimilarity and high inter-cluster dissimilarity.

Literature review. There are a lot of researches and studies presented about the cost of living, methods of studying it, and how to classify it.for example “Rafaela Costa Martins de Mello Dourado and Alessandra de Ávila Montin “in the article [The cost of living in the best livable cities in the world: a brief predictive quantitative analysis], and “Ali Mohamed ALnayer “in the article [cost of living analysis in africa].

But through this paper, I will use cluster analysis to compile, organize and follow controls to find and group cities that are cheap or expensive to live in the same cluster even with different countries and their locations in the map, and also give a brief overview of the ways to use cluster analysis.

Aims. The main goal of this paper is to demonstrate how cluster analysis algorithms can be applied to the data on the cost of living in different European countries.

We introduce a general overview about Definition of Cost Living Data, and Description of data, and where we found the data. And we describe [Algorithms of Cluster Analysis, The K-means algorithm, k-medoids algorithm, Hierarchical Clustering, Principal Components] techniques and their application to the visual representation of cluster analysis results. This model could help companies or even people to decide which city to move to in order to decrease living costs

Results. In my research was used cluster analysis to compile, organize and follow controls to find and group cities that are cheap or expensive to live in the same cluster even with different countries and their locations in the map, and also give a brief overview of the ways to use cluster analysis.

Cost of living analysis by:

The definition of Cambridge Advanced Learner’s Dictionary the Cost of Living is "the amount of money that a person needs to live [Cambridge, D.2013] It depends on prices on such goods as housing, food, health care and so on. There can be two different approaches in the analysis and comparison of the cost of living in different cities or locations.

The first approach concentrates on the construction of the cost of living index which is a measure of differences in the price of goods and services. Such a measure represents the cost of living at a given location as one number of aggregating prices on different goods and services. It is easy then to compare and rank different locations by this measure. On the other hand, the amount of money that a person needs to live also depends on the living standards of the person. Different persons need different amounts of goods to be satisfied. So it is difficult to construct one measure which will be useful for any person.

Therefore we apply a second approach comparing the structures of prices on main goods and services at different locations. To do this we applied statistical techniques of cluster analysis and dimension reduction of multivariate data. In what follows data on the cost of living from the Numbeo [www.numbeo.com] website are used.

The description of Data :

Six European countries were selected for the study. Five most populated cities were taken in each state. For each city we obtained data on prices on twenty kinds of consumer

goods and services from the Numbeo website [www.numbeo.com], Twenty variables were selected about living costs prices by (USD) were collected in 2019 by each city which can be seen in Table 1.

The considered cities are divided into two groups according to their geographical location, as presented in Table 2.

The following different research methods are used in the article, including *Algorithms of Cluster Analysis.* Cluster analysis, also called data segmentation, has a variety of goals. All relate to grouping or segmenting a collection of objects into subsets or (clusters) such that those within each cluster are more closely related to one another than objects assigned to different clusters [Kassambara, 2017].

Table 1

Represents twenty variables of consumer goods and services for cost of living

	detailed living costs variables (USD)
Mealinexp	Meal, Inexpensive Restaurant
Meal2	Meal for 2 People, Mid-range Restaurant, Three-course
Beer	Domestic Beer
Milk	Milk (regular), (1 liter)
Bread	Loaf of Fresh White Bread (500g)
Eggs	Eggs (regular) (12)
Cheese	Local Cheese (1kg)
Beef	Beef Round (1kg) (or Equivalent Back Leg Red Meat)
Apples	(regular) (1kg)
Banana	(regular) (1kg)
Tomato	(regular) (1kg)
Potato	(regular) (1kg)
Taxi	Taxi 1km (Normal Tariff)
Utilities	Basic (Electricity, Heating, Cooling, Water, Garbage) for 85m2 Apartment
Gasoline	Gasoline (1 liter)
Mobile	1 min of Prepaid Mobile Tariff Local (No Discounts or Plans)
Internet	Internet (60 Mbps or More, Unlimited Data, Cable/ADSL)
Fitness	Fitness Club, Monthly Fee for 1 Adult
Jeans	1 Pair of Jeans (Levis 501 Or Similar)
Nike	1 Pair of Nike Running Shoes (Mid-Range)

The six countries are:{Ukraine, Poland, Czech Republic, Italy, France, Spain} And the cities being analyzed are: {Kyiv, Kharkov, Dnipro, Odessa, Lviv, Warsaw, Gdansk Krakov, Wroclaw, Katowice, Prague, Brno, Ostrava, Plzen, Olomouc, Rome, Milan, Naples, Turin, Palermo, Paris, Marseille, Lyon, Toulouse, Nice, Madrid, Barcelona, Valencia ,Seville ,Zaragoza}.

Table 2

30 cities by continent

Continent	Cities
Eastern Europe	15
western Europe	15
Total of	30

The considered goods and services can also be divided into groups with respect to their meaning in the life of typical consumer : (1) leisure time: {Mealinexp, Meal2, Beer, Fitness} (2) local food: {Milk, Bread ,Cheese ,Beef} (3) grocery: {Apples ,Banana ,Tomato ,Potato} (4) services: {Taxi ,Mobile, Internet}

Proximities and Dissimilarities. Cluster analysis begins with a quantitative measurement of similarity between Candidate objects among a large number of objects characterized by several variables.

Euclidean distance :

$$d(A,B) = \sqrt{(a1-b1)^2 + (a2-b2)^2 + \cdots + (ap-bp)^2} \quad (1)$$

$$d_{euc}(X,Y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (2)$$

Manhattan distance :

$$d_{man}(x,y) = \sum_{i=1}^{n}|(x_i - y_i)| \quad (3)$$

where the K-means algorithm.

The K-means algorithm is one of the most popular iterative descent clustering methods . and is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents the number of groups pre-specified by the analyst. It classifies objects in multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible (i.e., high intra-class similarity), whereas objects from different clusters are as dissimilar as possible (i.e., low inter-class similarity). In k-means clustering, each cluster is represented by its center (i.e, centroid) which corresponds to the mean of points assigned to the cluster [Kassambara,2017]. [Hastie et al., 2008]:

$$W(C_K) = \sum_{x_i \epsilon c_k} (x_i - \mu_k)^2 \quad (4)$$

We define the total within-cluster variation as follow :

$$\text{Tot.withinss} = \sum_{k=1}^{k} W(C_K) = \sum_{k=1}^{k} \sum_{x_i \epsilon c_k} (x_i - \mu_k)^2 \quad (5)$$

Estimating the optimal number of clusters: The k-means clustering requires to specify the number of clusters to be generated. To compute k-means clustering using different values of clusters k. the wss (within the sum of the square) is compute drawn according to the number of clusters. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters [Kassambara, 2017].

k-medoids algorithm: A k-medoids algorithm is a clustering approach related to k-means clustering for partitioning a data set into k groups or clusters. In k-medoids clustering, each cluster is represented by one of the data points in the cluster. These points are named cluster medoids. The term medoid refers to an object within a cluster for which average dissimilarity between it and all the other the members of the cluster is minimal. It corresponds to the most centrally located point in the cluster.

Hierarchical Clustering: Hierarchical clustering [or hierarchical cluster analysis (HCA)] is an alternative approach to partitioning clustering for grouping objects based on their similarity. In contrast to partitioning clustering, hierarchical clustering does not require to pre-specify the number of clusters to be produced. Hierarchical clustering can be subdivided into two types:

Similarity measures: To decide which objects/clusters should be combined or divided, we need methods for measuring the similarity between objects.

There are many methods to calculate the (dis)similarity information, including Euclidean and Manhattan distances to compute the distance between every pair of the object in a data set.

Linkage: The linkage function takes the distance information, and groups pairs of objects into clusters based on their similarity. formed clusters are linked to each other to create bigger clusters. This process is iterated until all the objects in the original data set are linked together in a hierarchical tree [Kassambara, 2017]**.**

Dendrogram: Dendrogram corresponds to the graphical representation of the hierarchical tree [Kassambara, 2017].

Verify the cluster tree: After linking the objects in a data set into a hierarchical cluster tree, we want to see if the distances in the tree reflect the original distances accurately. the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the original distance matrix. The closer the value of the correlation coefficient is to 1 [Kassambara, 2017].

Principal Components: The principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. Its general objectives are (1) data reduction and (2) interpretation. Although p components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number k of the principal components. there is almost as much information in the k components as there is in the original p variables. The k principal components can then replace the initial p variables, and the original data set, consisting of n measurements on p variables, is reduced to a data set consisting of n

measurements on k principal components. Analyses of principal components are more of a means to an end rather than an end in themselves because they frequently serve as intermediate steps in much larger investigations [Johnson & Wichern, 1998].

With studied data that represent the cost of living according to the countries mentioned above, as well as through the cities mentioned.

We applied the statistical methods used to study mass analysis. We will review the results as follows:

The choice of distance measures is very important, as it has a strong influence on the clustering results.

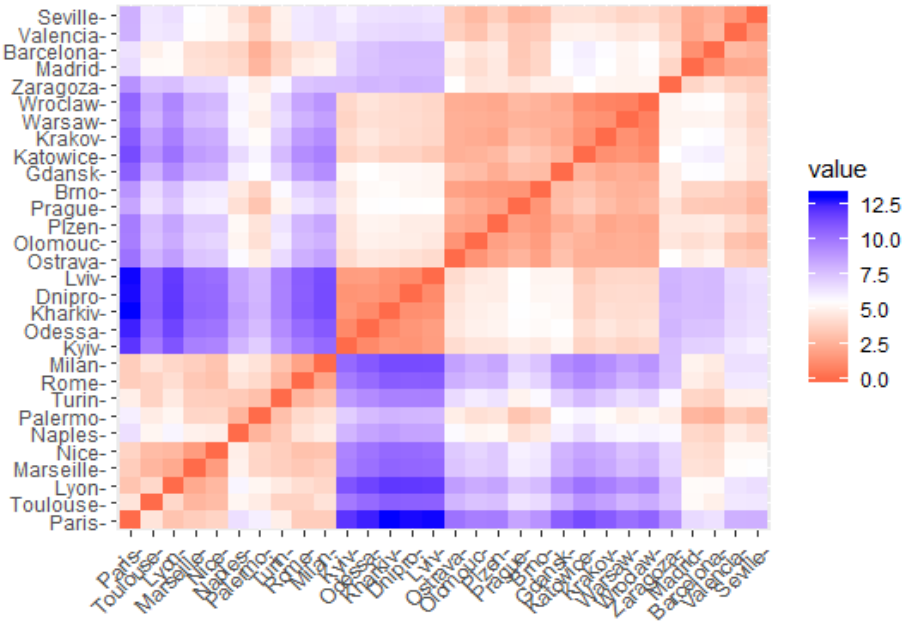

Figure 1. Visualize distance matrices by used Euclidean method from the cost of living dataset

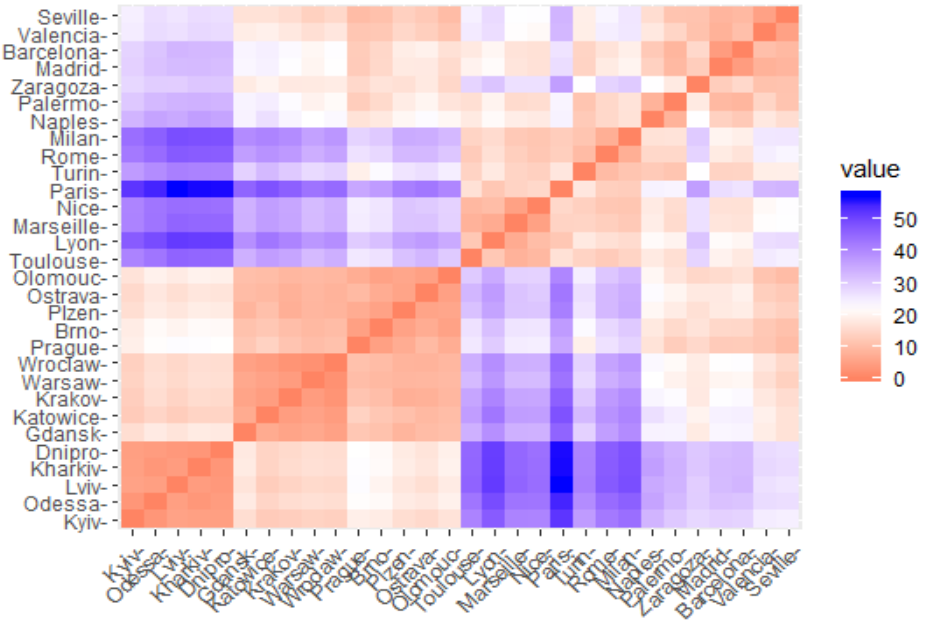

Figure 2. Visualize distance matrices by used Manhattan method from the cost of living dataset

Figure 1, 2: Visualize distance matrices by used Euclidean and Manhattan from the cost of living dataset. Red: high similarity (ie: low dissimilarity) | Blue: low similarity

The color level is proportional to the value of the dissimilarity between observations: pure red if dist.(xi, xj) = 0 and pure blue if dist.(xi, xj) = 1. Objects belonging to the

same cluster are displayed in consecutive order.

As we see From Fig. 3 & Fig. 4 Through the k-means algorithm results are the cities classify into k-groups or " cluster, based on similarities. Each cluster is represented by the mean value of points in the cluster known as the cluster centroid.

As we see from Fig. 5 the variance within the clusters. It decreases as k increases, but it can be seen as a bend (or "elbow") at k = 4. This bend indicates that additional clusters beyond the fourth have little value.

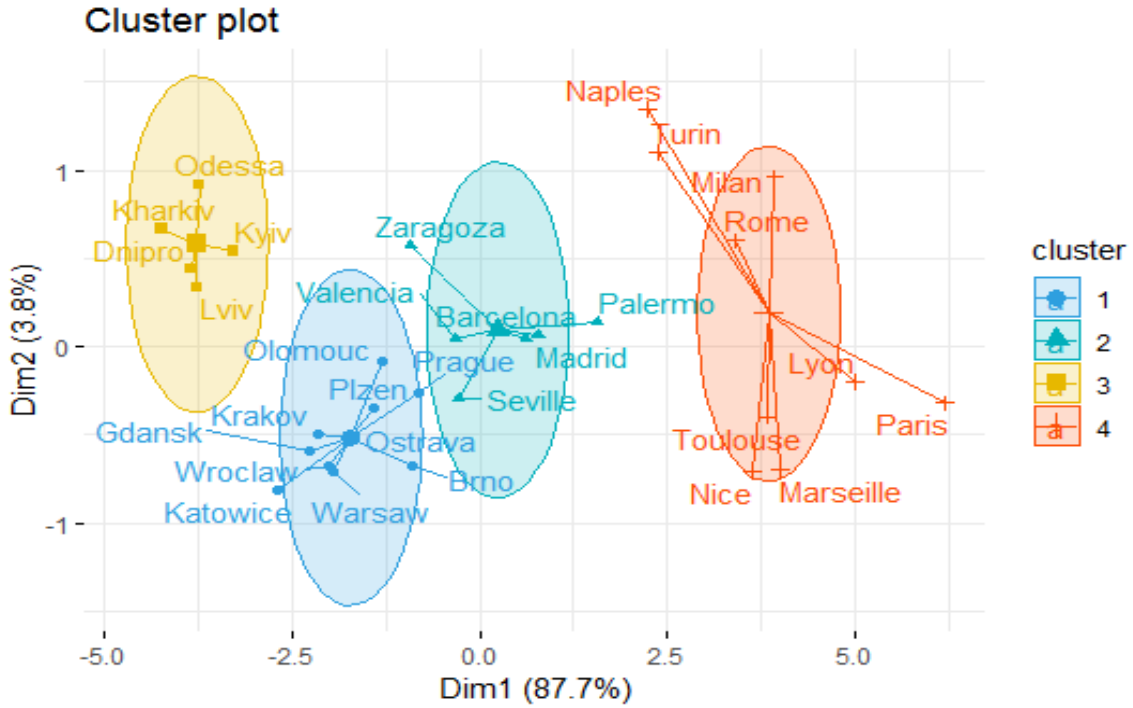

Fig. 3. Visualizing k-means clusters from the cost of living dataset

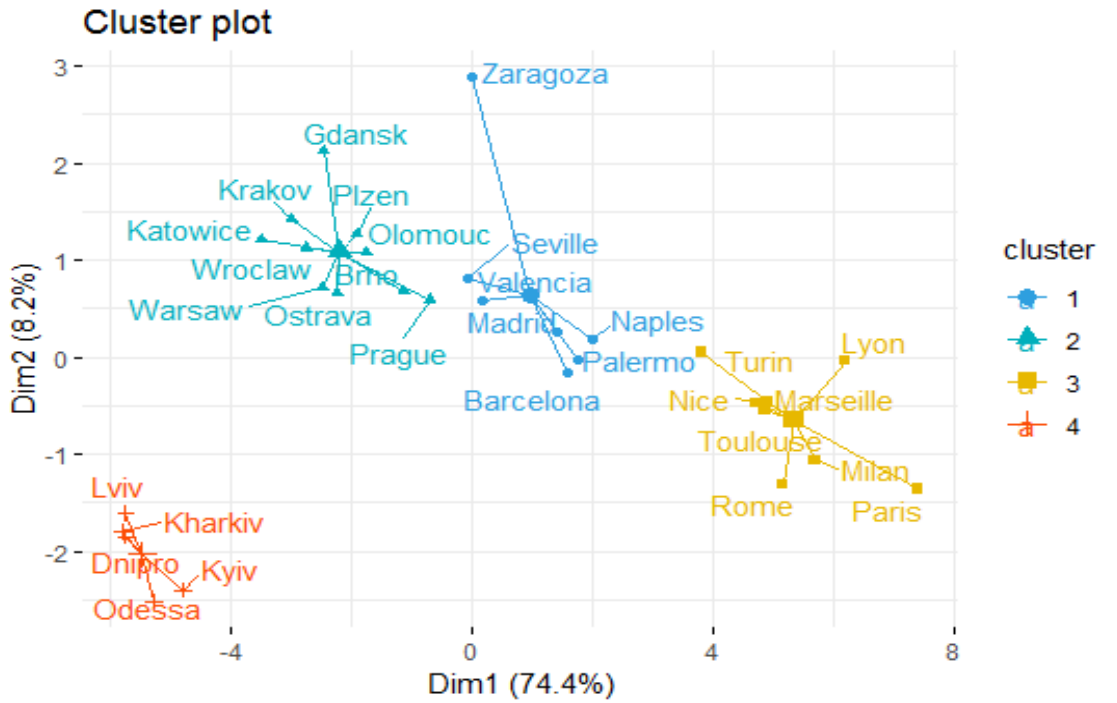

Fig. 4. The Determine number of clusters for the cost of living dataset

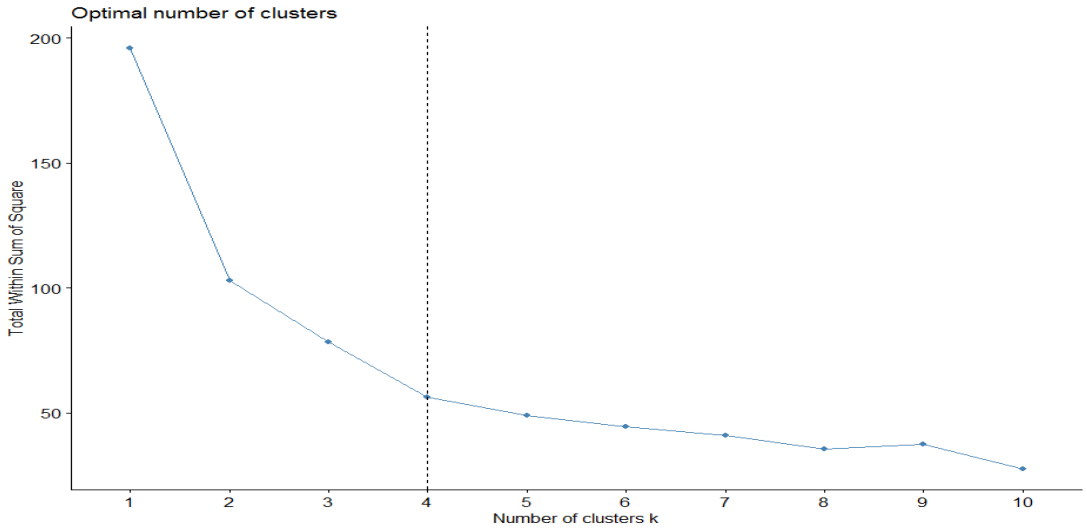

Fig. 5. Estimating the optimal number of clusters from the cost of living dataset

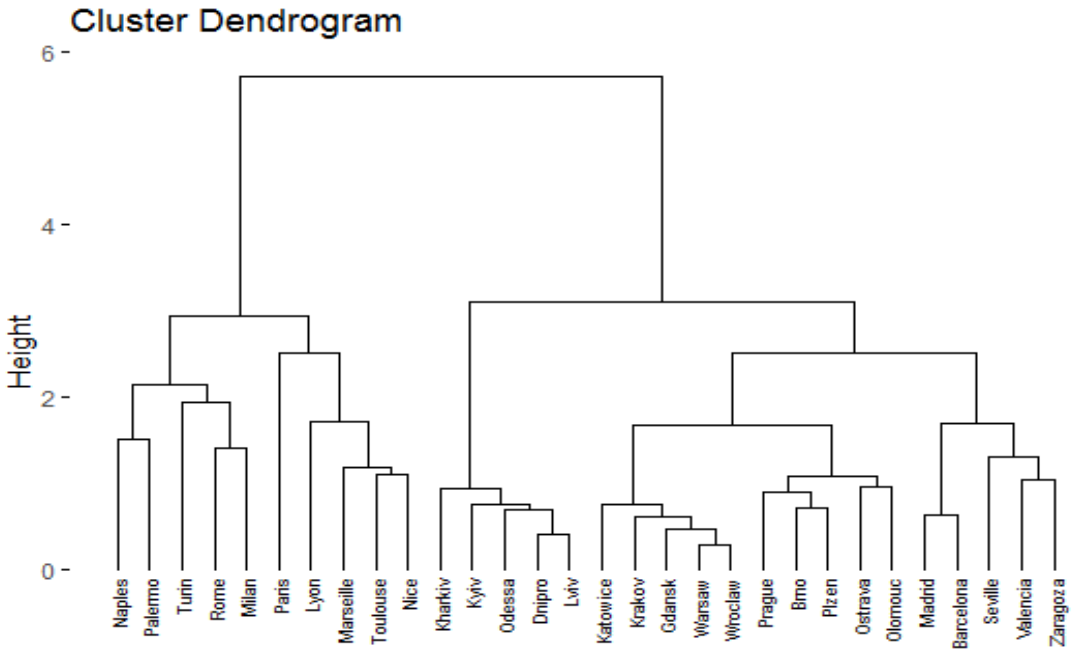

Fig. 6. Hierarchical Clustering for the cost of living dataset

As we see from Fig 6, the cities grouping in a cluster based on the similarity in the first stage, and in the second stage the clusters are merged between them based on the similarity, and the process continues between them tell give us dendrogram like above.

We also see from Fig 7 & 8, the results of Principal Components analysis the main component explaining the structure of variance and variance of a group of variables through a few linear groups of these variables.

As we see from the result here are many cluster agglomeration methods (i.e, linkage methods). The most common linkage methods are described below as.

From Figure 9, the result is by the lowest or singular linkage. That means that the distance between two clusters is defined as the minimum value of all marital distances between the elements by Euclidean in cluster 1 and the elements in cluster 2, and cluster 3, and cluster 4. It is combined with the linking between them consisting of the dendrogram.

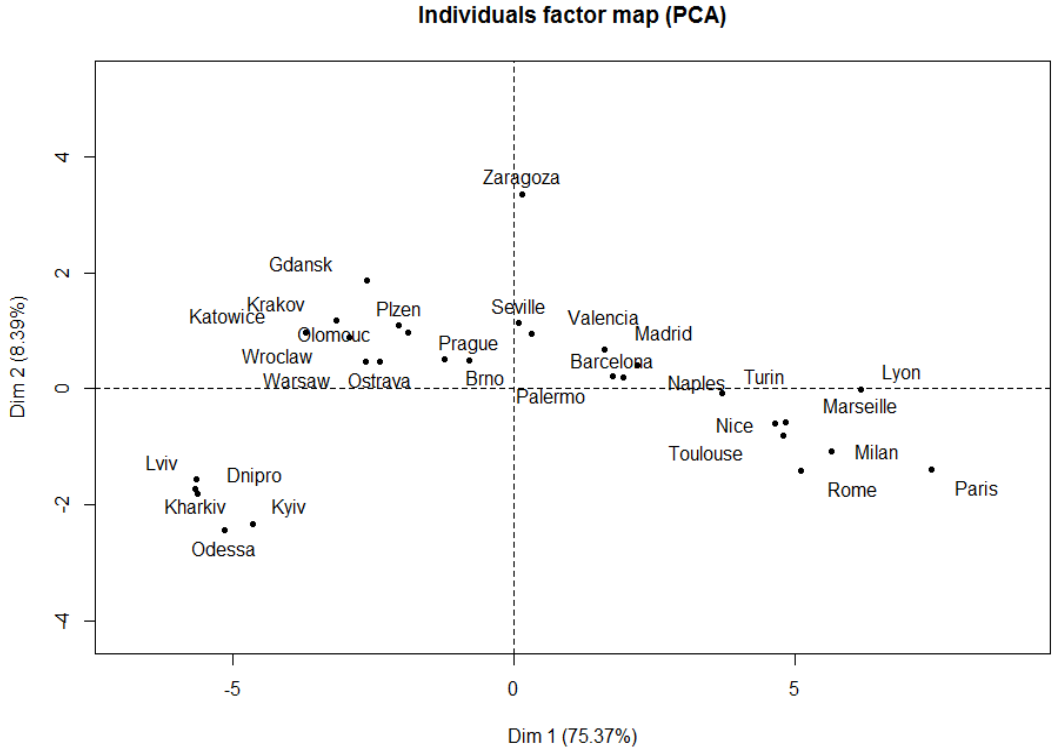

Figure 7. Principal Components from the cost of living dataset

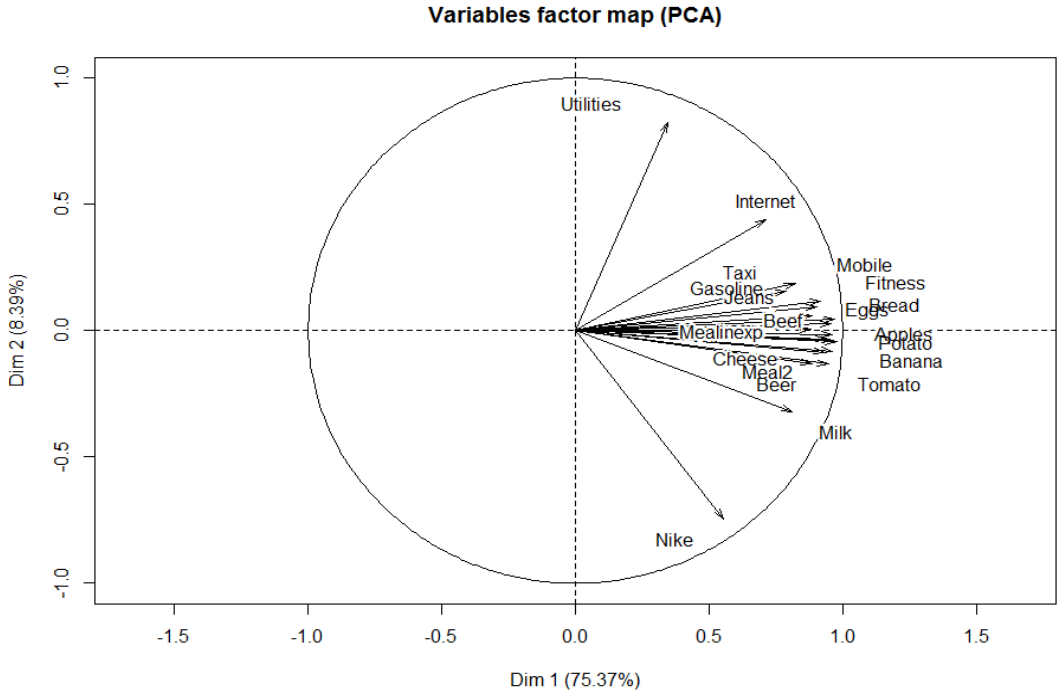

Figure 8. Principal Components from the cost of living dataset

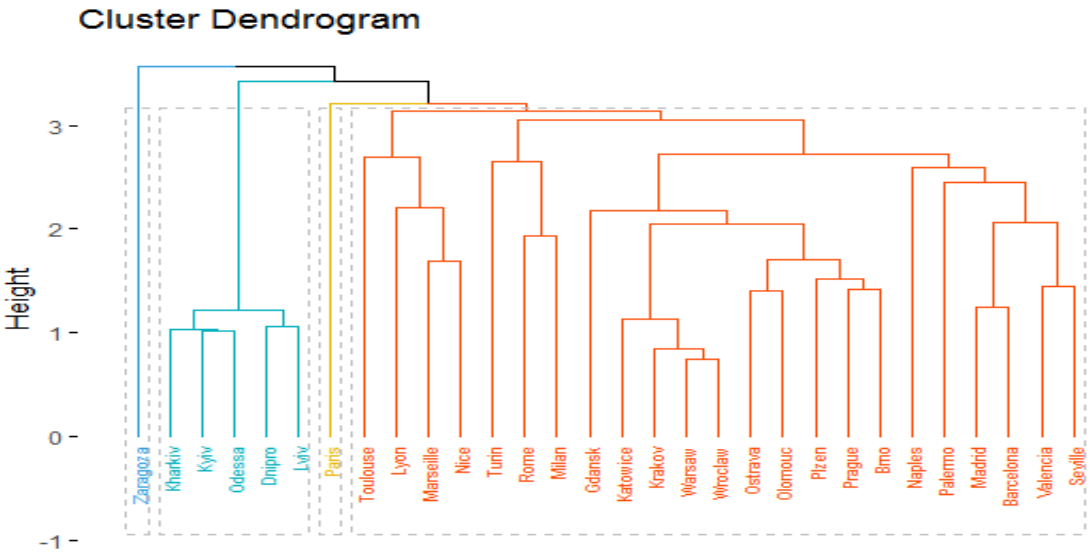

Figure 9. Dendrogram by Euclidean distance and Single Linkage from the cost of living dataset

From Fig 10, the result is by the lowest or singular linkage. That means that the distance between two clusters is defined as the minimum value of all marital distances between the elements by Manhattan in cluster 1 and the elements in cluster 2, and cluster 3, and cluster 4. It is combined with the linking between them consisting of the dendrogram.

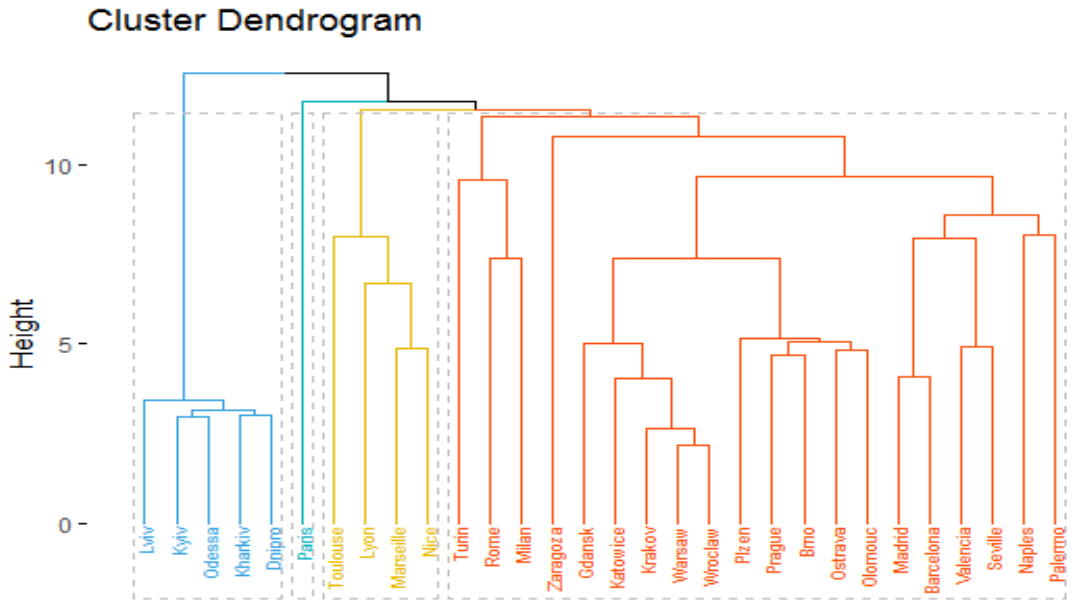

Figure. 10. Dendrogram by Manhattan distance and Single Linkage from the cost of living dataset

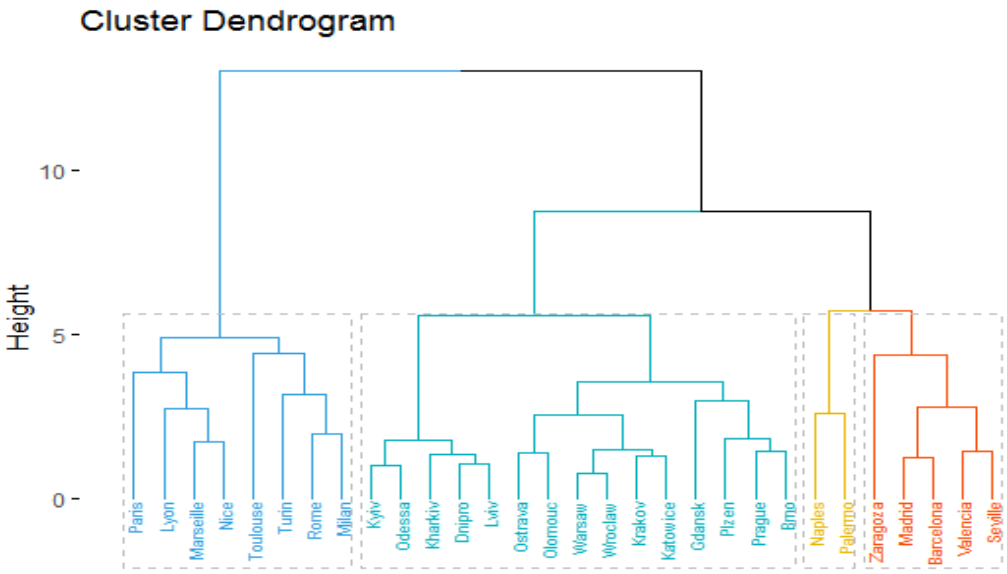

Figure 11. Dendrogram by Euclidean distance and complete Linkage from the cost of living dataset

From Fig 11, the result is by the Maximum or complete linkage. That means that The distance between two clusters is defined as the maximum value of all pairwise distances between the elements by Euclidean in cluster 1 and the elements in cluster 2, and cluster 3, and cluster 4. It tends to produce more compact clusters and It is combined with the linking between them consisting of the dendrogram. From Fig 12, the result is by the Maximum or complete linkage. That means that The distance between two clusters is defined as the maximum value of all pairwise distances between the elements by Manhattan in cluster 1 and the elements in cluster 2, and cluster 3, and cluster 4. It tends to produce more compact clusters and It is combined with the linking between them consisting of the dendrogram.

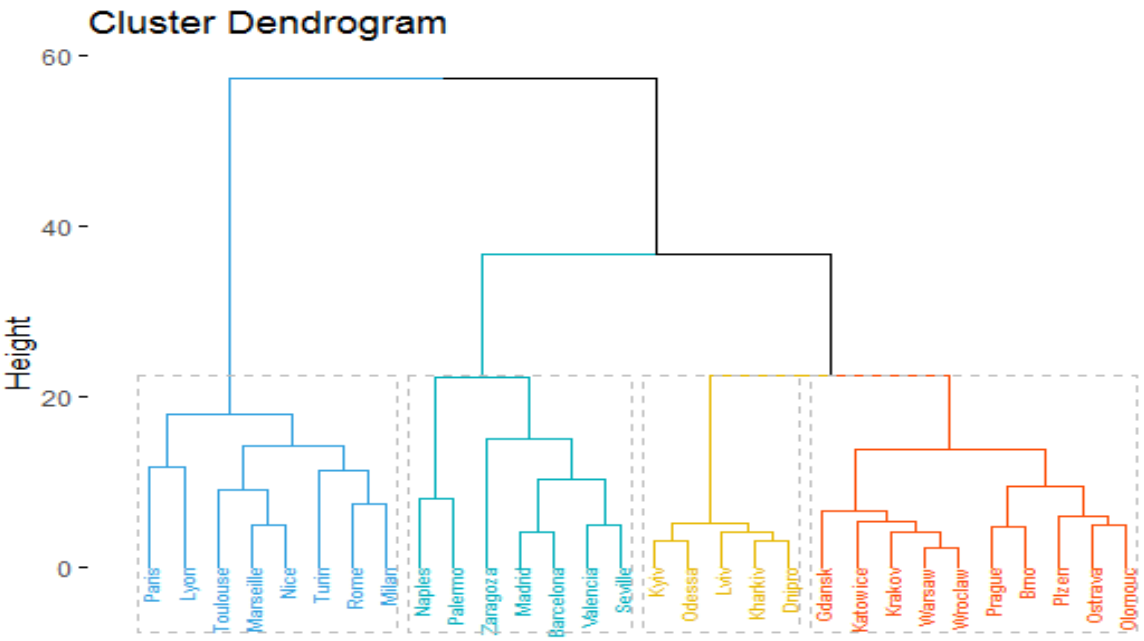

Figure 12. Dendrogram by Manhattan distance and Complete Linkage from the cost of living dataset

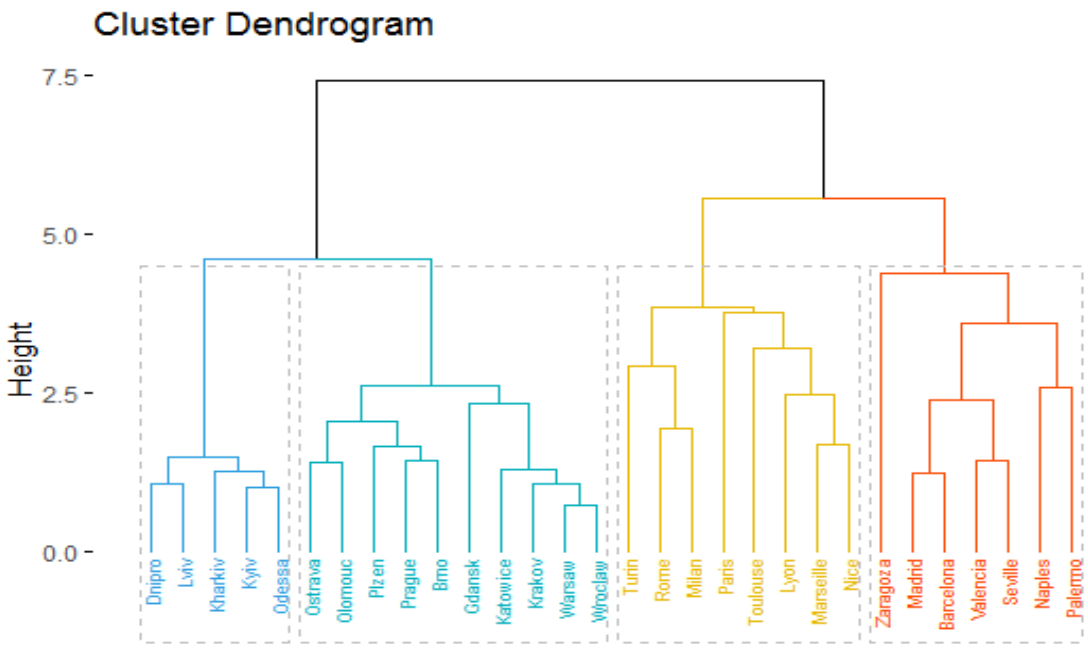

Figure 13. Dendrogram by Euclidean distance and Average Linkage from the cost of living dataset

From Fig 13, the result is by the Mean or average linkage. That means that The distance between two clusters is defined as the average distance between elements by Euclidean in cluster 1 and the elements in cluster 2, and cluster 3, and cluster 4. It tends to produce more compact clusters and It is combined with the linking between them consisting of the dendrogram.

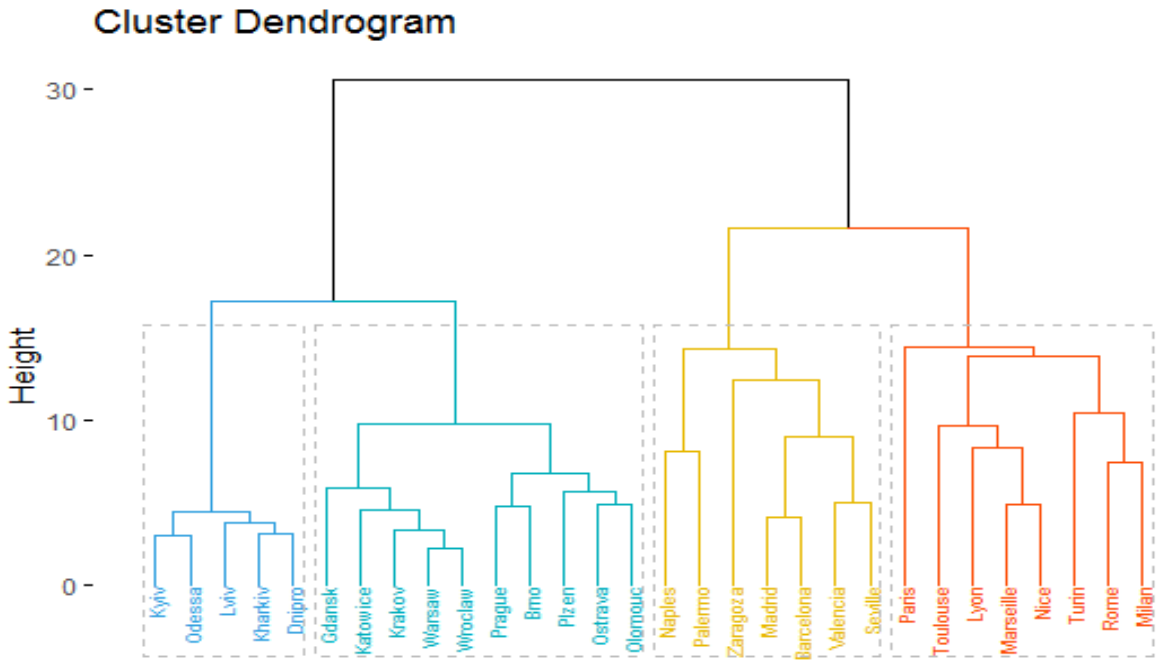

Figure 14. Dendrogram by Manhattan distance and Average Linkage from the cost of living dataset

From Fig 14, the result is by the Mean or average linkage. That means that The distance between two clusters is defined as the average distance between elements by Manhattan in cluster 1 and the elements in cluster 2, and cluster 3, and cluster 4. It tends to produce more compact clusters and It is combined with the linking between them consisting of the dendrogram.

From the previous results of analyzing the cost of living data in Europe, it is possible to use the comprehensive analysis to clarify that it is possible to conclude that cities close to each other in cost converge in one group regardless of their geographical location in terms of proximity or distance, regardless of the difference in language or Currency or internal system, surprisingly, there are cities from Western Europe that fall in the same group as cities from Eastern Europe in terms of cheap living costs, and we can notice the difference by looking at the charts shown earlier.

This is because there are countries that have a local currency such as Ukraine and countries that have the euro currency, such as Italy and France, for this purpose in this paper I used the US dollar as a general measure.

In this paper, we consider k-mean and k-medoids methods and hierarchical aggregation algorithms and their applications for the problem of cost-of-living data analysis. Ironing processes obtained through grouping techniques are visualized through dimensional reduction techniques such as major components and multidimensional scaling. The results obtained through different algorithms are different, but they are consistent with each other in their basic features. Whereas, as we note from the graphs in the results, the grouping of cities in each cluster differs according to the method of connection used, as in this way we have been able to compile and direct them in drawing a dendrogram clear fundament that differs from what is common in other research that reflects different opinions in the way of output and analysis The results, as these results allow us to describe the similarities and differences in the price structure of products and services in different cities and European countries.

Conclusions. The results of the analysis showed that there are some similarities in the cost of living in some European countries and grouping them into one cluster, taking into account the difference in their geographical location. This indicates that the composition of the clusters depends on the method of linking in the analysis.

These results can be used by individuals or institutions to choose the best European countries to live by comparing the cost of living in them compared to the available budget.

References:

1. Aboukadel, K. (2017), “Practical Guide To Cluster Analysis in R”. *Unsupervised Machine Learning*. Sthda Press. Edition 1.

2. Trevor, H. Robert, T. and Jerome, F. (2008), “The Elements of Statistical Learning”, *Data Mining, Inference, and Prediction*. Springer Press. Second Edition.

3. Sadanori, K. (2014), “Introduction to Multivariate Analysis”. Linear and Nonlinear Modeling Chapman & Hall Press.

4 . Richard, A. J. & Dean, W. W. (1998), “Applied Multivariate Statistical Analysis”, *Prentice Hall Press.* Four Edition.

5. Rafaela, C., D., M. & Alessandra, D., M. (2016), “The cost of living in the best livable cities in the world: a brief predictive quantitative analysis”. Int. J. Multivariate Data Analysis, Vol. 1, №1.

6. Numbeo (2019), “Is the world’s largest database of user contributed data about cities and countries worldwide”, Numbeo provides current and timely information on world living conditions including the cost of living, housing indicators, health care, traffic, crime, and pollution. 5,025,770 prices in 8,855 cities entered by 422,329 contributors and the information. Retrieved from: https://www.numbeo.com.

Стаття надійшла до редакції 01.05.2020 р.