ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 8.4
IJAR 2023; 9(2): 323-334
www.allresearchjournal.com
Received: 23-11-2022
Accepted: 16-01-2023

Joseph Justin Rebello
Department of Statistics, AC, Mahatma Gandhi University, Kottayam, Kerala, India

Shafna Noushad
Department of Statistics, AC, Mahatma Gandhi University, Kottayam, Kerala, India

Surya K S
Department of Statistics, AC, Mahatma Gandhi University, Kottayam, Kerala, India

Corresponding Author:
Joseph Justin Rebello
Department of Statistics, AC, Mahatma Gandhi University, Kottayam, Kerala, India

Some studies on principal component analysis, factor analysis and cluster analysis on a clubbed data

Joseph Justin Rebello, Shafna Noushad and Surya K S

DOI: https://doi.org/10.22271/allresearch.2023.v9.i2e.10635

Abstract
The objective of the study is to apply data reduction techniques PCA and FA on the Healthy Lifestyle Cities Report 2021 and also to cluster the same data using cluster Analysis. The data analysed 44 cities across the globe to uncover where it is easier to lead a well-rounded, healthy lifestyle. From obesity levels to pollution rates, each city has been scored across 6 healthy living metrics. Each of these metrics were awarded a weighted score and these were combined to give each city a total sore out of 100. This score was then used to rank the 33 cities to determine which were best for healthy living. For the analysis of the data, statistical packages “SPSS” and “R” are being applied.

Keywords: Principle component analysis, factor analysis, cluster analysis, complete linkage method, average linkage method, k means

1. Introduction
Multivariate analysis is a statistical procedure for the analysis of data involving more than one type of measurement or observation [1]. It may also mean solving problems where more than one dependent variable is analysed simultaneously with other variables. Multivariate analysis is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time. The main advantage of multivariate analysis is that since it considers more than one factor of independent variables that influence the variability of dependent variables, the conclusion drawn is more accurate. The conclusions are more realistic and nearer to the real-life situation [11].

1.1 Principal Component Technique
A Principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a *few linear* combinations of these variables [8]. Principle component may be useful to transform the original set of variables to a new set of *uncorrelated variables* [2]. These new variables are called Principal components which are the *normalized linear combination* of the original variables and are derived in the decreasing order of importance. PCA is mostly used as a tool in exploratory data analysis and for making predictive models [9]. It's often used to visualize genetic distance and relatedness between populations. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z scores) the data matrix for each attribute [3]. The results of a PCA are usually discussed in terms of *component scores*, *sometimes called factor scores* (the transformed variable values corresponding to a particular data point), and *loadings* (the weight by which each standardized original variable should be multiplied to get the component score).

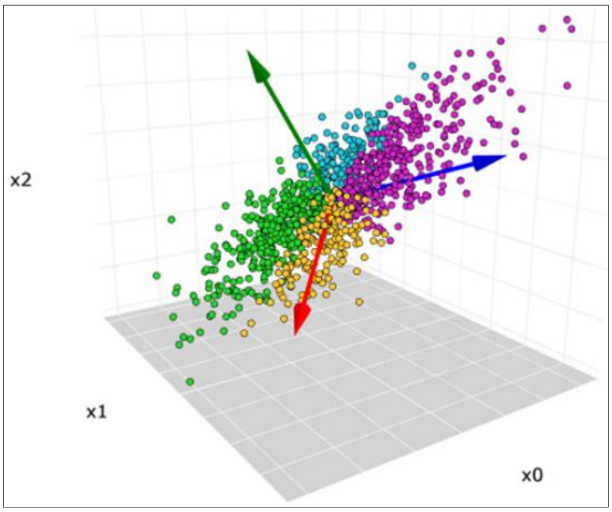

Fig 1: PCA

PCA is closely related to factor analysis. Factor analysis typically incorporates more domain specific assumptions about the underlying structure and solves eigenvectors of a slightly different matrix. PCA is also related to canonical correlation analysis (CCA)(6). CCA defines coordinate systems that optimally describe the cross-covariance between two datasets while PCA defines a new orthogonal coordinate system that optimally describes variance in a single dataset.

1.1.1 Scree Plot

Decision regarding the number of principle components to be taken in any data analysis is decided graphically by a scree plot. The term '*scree*' is taken from the word for the rubble at the bottom of the mountain (7).

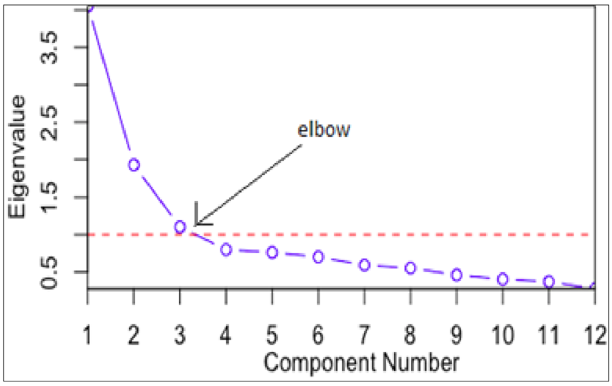

Fig 2: Scree Plot

1.2 Factor Analysis

Factor analysis was developed originally for the analysis of scores on mental tests; however, the methods are useful in a much wider range of situations such as analysing sets of tests of attitudes, sets of physical measurements and sets of economic quantities(4). Factor can be considered as an extension of Principal component analysis. Both can be viewed as attempts to approximate the covariance matrix $\sum$.

1.3 Cluster Analysis

Cluster analysis is multivariate method which aims to classify a set of objects in such a way that objects in the same group (called cluster) are more similar to each than to those in other groups (12). Grouping is done on the basis of similarities or distances (dissimilarities) (13).

2. Methodology

The variables used for the analysis are the following:

- X_1 = Rank
- X_2 = Sunshine Hours(City)
- X_3 = Life Expectancy (Country)
- X_4 = Happiness Levels (Country)
- X_5 = Outdoor Activities (Cities)
- X_6 = Number of takeout places (City)

2.1 Principal Component Analysis

Let X be a P component random variable whose mean is assumed to be μ and dispersion matrix Σ, where Σ is a real positive matrix. The equation for the characteristic root and the corresponding characteristic vector is given by

$$\Sigma X = \Lambda X \ (1)$$

According to Hotelling's iterative procedure we start with an initial $P \times 1$ vector X_0 which is not orthogonal to e_1, the characteristic vector corresponding to the largest characteristic root λ_1 of Σ.

Define $X_i = \Sigma Z_{i-1}; i = 1,2,3, \dots \dots \dots, P$

$$Z_i = \frac{\mathrm{Xi}}{\sqrt{XiXi}}; i = 1,2,3, \dots \dots \dots, P \qquad (1.1)$$

It can be shown that,

$$\lim_{i\to\infty} z_{\mathrm{i}} = \pm e_i, \lim_{i\to\infty} X_i{}' X_i = \lambda_1{}^2 \qquad (1.2)$$

To find the second characteristic root and the corresponding characteristic vector we define, $\Sigma_2 = \Sigma - \lambda_1 e_1 e_1'$

Now to find λ_1 and e_2 we use the same iterative procedure to Σ_2. Repeat the steps (1.1) and (1.2)

Thus the non-zero Eigen values of Σ are $\lambda_1, \lambda_2, \lambda_3, \dots \dots \lambda_P$ with Eigen vectors $e_1, e_2, e_3, \dots \dots e_{P.}$

So we get the linear combination $Z_i = e_i' X$

That is,

$$Z_1 = e_{11}X_1 + e_{12}X_2 \dots \dots \dots + e_{1P}X_P$$

$$Z_2 = e_{21}X_1 + e_{22}X_2 \dots \dots \dots + e_{2P}X_P$$

$$Z_{\mathrm{P}} = e_{\mathrm{P}1}X_1 + e_{\mathrm{P}2}X_2 \dots \dots \dots + e_{\mathrm{PP}}X_P \qquad (1.3)$$

With the condition $V(Z_1) \geq V(Z_2) \dots \dots . V(Z_P) \geq 0$ (1.4)

The linear combination (1.3) is called principal component satisfying (1.4)

2.2 Factor Analysis

2.2.1 The Orthogonal Factor Model

The observable random vector X with p components has mean μ and covariance matrix $\sum$. The factor model postulates that X is linearly dependent upon a few unobservable random variables $F_1, F_2, \dots, F_m$ called common factors and p additional sources of variation $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \dots, \mathcal{E}_p$ called errors or specific factors. In matrix notation,

$$X - \mu = L\,F + \mathcal{E} \qquad (2)$$

where X- μ is a p x 1 vector, L is a p × m matrix, F is a m x 1 vector and $\mathcal{E}$ is a p × 1 vector.

The coefficient l_{ij} is called the loading of the i^{th}variable on the j^{th} factor, so that the matrix L is the matrix of factor

loadings. The i[th] specific factor ε_i is associated only with the i^{th}response X_i. The p deviations $X_1 - \mu_1$, $X_2 - \mu_2$, $\ldots X_p - \mu_p$, are expressed in terms of $p + m$ random variables $F_1, F_2, \ldots F_m, \varepsilon_1, \varepsilon_2 \ldots \varepsilon_p$ Which are *unobservable*. This distinguishes the factor model expressed in equation *(2)* from the regression model in the independent variables observed.
With so many unobservable quantities, a direct verification of the factor model from observations on $X_1, X_2, \ldots X_P$ is hope less. However, with some additional assumptions about the random vectors F and ε, the mode

$$X - \mu = L\,F + \varepsilon,$$

Implies certain covariance relationships can be checked.
The assumptions are:

$$E(F) = 0_{m\times 1}, Cov(F) = E(FF') = I_{m\times m}$$

$$E(\varepsilon) = 0_{p\times 1}, Cov(\varepsilon) = \varphi_{p\times p} = \begin{bmatrix} \varphi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi_p \end{bmatrix} \qquad (3)$$

F and ε are independent.

$$ie,\ Cov(F, \varepsilon) = E(\varepsilon F') = 0_{p\times m}$$

These assumptions and the relation $X - \mu = LF + \varepsilon$ constitute the *orthogonal factor model*. The factor analysis model with the above assumptions is called orthogonal Factor analysis model.

2.2.2 Methods of Estimation

Given observations $X_1, X_2, \ldots X_n$ on p generally correlated variables. The sample covariance matrix S is an estimator of the unknown population covariance matrix $\sum$. If the off-diagonal elements of S are small or those of the sample correlation matrix R essentially zero, the variables are not related, and a factor analysis will not be useful. In such circumstances, the specific factors play a dominant role, but the aim of factor analysis is to determine a few important common factors. If $\sum$ appears to deviate significantly from a diagonal matrix, then a factor model can be entertained and the initial problem is one of estimating the factor loadings I_{ij} and specific variance φ_i.

2.2.3 Factor Rotations

The results of factor extraction, unaccompanied by rotation are likely to be hard to interpret regardless, of which method of extraction is used. After extraction, rotation is used to improve the interpretability and scientific utility of solution. It is not used to improve the quality of mathematical fit between observed and reproduced correlation matrices because all orthogonally rotated solutions are equivalent to one another and to the solution before rotation. All factor loadings obtained from the initial loadings by an orthogonal transformation have the same ability to reproduce the covariance (or correlation) matrix. We know that, an orthogonal transformation corresponds to a rigid rotation of the co-ordinate axes. For this reason, an orthogonal transformation of the factor loadings, as well as the implied orthogonal transformations of the factors, is called *factor rotation*. Rotations are ordinarily used after extraction to maximize high correlations and minimize low ones.

If $\hat{L}$ is the $p \times m$ matrix of estimated factor loadings obtained by any method (principal component or maximum likelihood) then

$$L^{\wedge *} = L^{\wedge} T, \text{where } TT' = T'T = I \text{ (orthogonal)}.$$

Hence $L^{\wedge *} = L^{\wedge} T$ is a $p \times m$ matrix of "rotated" loadings. Moreover, the estimated covariance (or correlation) matrix remains unchanged since

$$L^{\wedge} L^{\wedge\prime} + \Psi = L^{\wedge} TT' L^{\wedge\prime} + \Psi^{\wedge} = L^{\wedge *} L^{\wedge *\prime} + \Psi^{\wedge}.$$

Hence the residual matrix

$$S_n - L^{\wedge} L^{\wedge\prime} - \Psi^{\wedge} = S_n - L^{\wedge *} L^{\wedge *\prime} + \Psi^{\wedge}$$

remains unchanged. Moreover, the specific variances $\Psi_i^{\wedge}$ and hence the communalities $h_i^{\wedge 2}$ are unaltered.

2.2.4 Factor Scores

Usually in Factor analysis, the interest in centered on the parameters in the factor model. We may also require the estimated values of the common factors called *factor scores*. These quantities are often used for diagnostic purposes, as well as inputs to a subsequent analysis. Factor scores are not estimates of unknown parameters in the usual sense. Rather, they are estimates of the values for the unobserved random vectors

$F_j, j = 1,2, \ldots, n$. That is, factor scores

$f_j^{\wedge}$ =Estimates of the values f_j attained by F_j (j^{th}case).
The estimation of factor scores is done using weighted least squares method as follows:
Suppose the mean vector μ, the factor loading L and specific variance $\Psi$are known for the factor model, $X - \mu = LF + \varepsilon$.
Further, regard the specific factor $\varepsilon' = (\varepsilon_1, \varepsilon_2, \ldots \ldots, \varepsilon_p)$ as errors. Since
$V(\varepsilon_i) = \Psi_i, i = 1,2, \ldots p$need not be equal. Bartlett suggested that weighted least squares can be used to estimate the common factor values.
The sum of squares of the errors weighted, by the reciprocal of their variance is

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\Psi_i} = \varepsilon'^{\Psi^{-1}} \varepsilon = (X - LF - \mu)' \Psi^{-1} (X - LF - \mu).$$

If we take $L^{\wedge}, \Psi^{\wedge}$and $\mu^{\wedge} = \bar{X}$, the estimates of $L, \Psi$and μ as the true values, then the factor scores for the j^{th}case is obtained by minimizing

$$(X_j - LF_j - \hat{\mu})' \Psi^{-1} (X_j - LF_j - \hat{\mu}).$$

The solution is given by

$$F_j^{\wedge} = (L^{\wedge\prime} \Psi^{-1} L^{\wedge})^{-1} L^{\wedge\prime} \Psi^{\wedge -1} (X_j - \bar{X}), \text{ j=1,2....n}$$

The factor scores generated have sample mean vector **o** and zero sample covariance matrix.

3. Analysis of Data
3.1 Principal component analysis
The result obtained by using the principal component analysis as the extraction method is given below.

Table 1: Communalities

Communalities		
	Initial	Extraction
Rank	1.000	.963
Sunshine Hours	1.000	.997
Life expectancy	1.000	.952
Happiness levels	1.000	.985
Outdoor activities	1.000	.963
Number of take out places	1.000	.931

Extraction

It indicates that proportion of variance that can be explained by the principal components. Now, a scree plot displays the eigen values associated with a component or factor in descending order versus the number of the components or factor. We use scree plots in principal components analysis and the factor analysis to visually assess which components or factors explain most of the variability in the data.

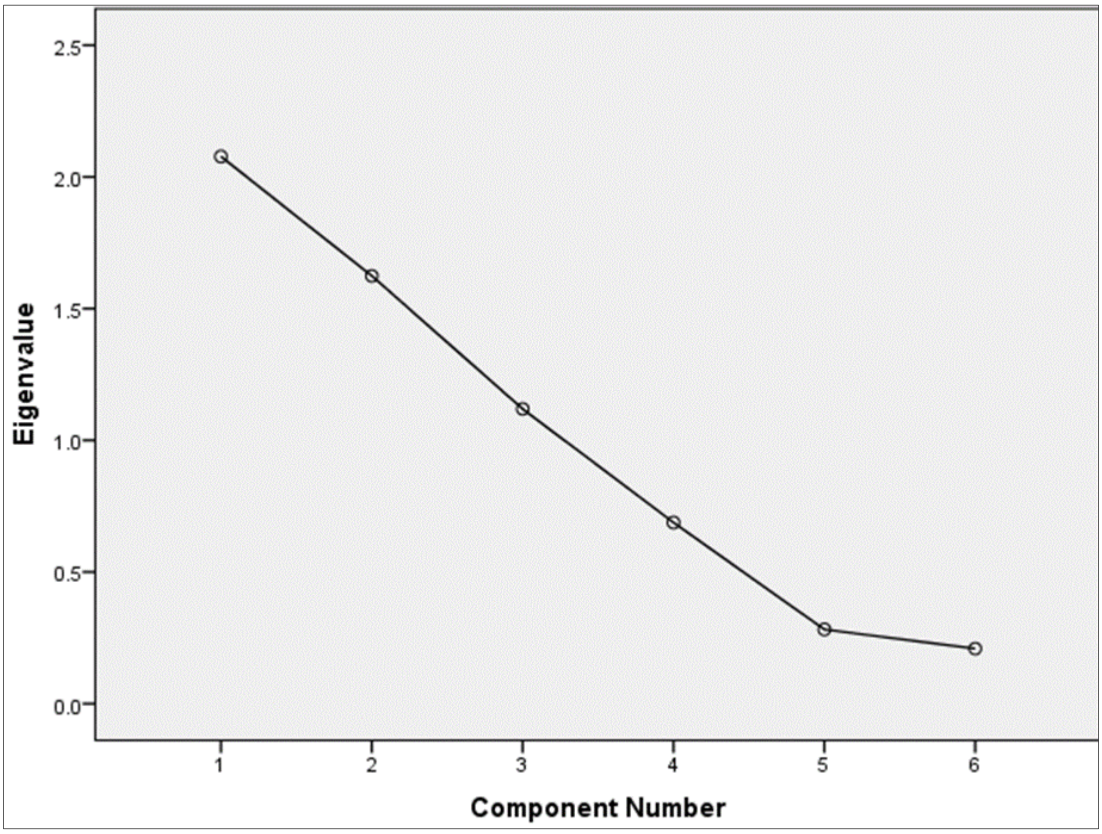

Fig 3: Scree plot

From the scree plot, it can be concluded that we can extract 5 principal components

Table 2: Total Variance Explained

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.078	34.625	34.625	2.078	34.625	34.625
2	1.624	27.071	61.697	1.624	27.071	61.697
3	1.119	18.654	80.350	1.119	18.654	80.350
4	.688	11.464	91.814	.688	11.464	91.814
5	.282	4.699	96.514	.282	4.699	96.514
6	.209	3.486	100.000			

The principal components with eigen values greater than 1 are normally considered. Here 5 components have eigen value greater than one. So these components are considered. Also from the scree plot, 5 components are retained. Therefore 5 components are extracted (14).

Table 3: Component Matrix

Component Matrix					
	Component				
	1	2	3	4	5
Rank	-.437	.401	-.699	.338	.089
Sunshine Hours	-.547	-.286	.455	.636	-.075
Life expectancy	.895	.108	.086	.213	-.292
Happiness levels	.872	.014	-.036	.331	.337
Outdoor activities	-.158	.692	.643	-.086	.196
Number of take out places	.021	.944	-.046	-0.83	-.176

It is seen that the 1st component explains 34.625% of variation of the data set, 2nd component explains 27.071% of data set, 3rd component explains 18.654% of variation of the data set, 4th component explains 11.464% of variation of the data set and 5th component explains 4.699% of data set. That is, first five component explain 96.514% of variation of data set. It is also from the component matrix that the first 5 components are highly influenced by all the 6 factors.

Also we get, the first component is highly influenced by the variables X_3followed by X_4 and X_6, second component is highly influenced by the variables X_6 followed by X_5 and X_1,3rd component is highly influenced by the variables X_5 followed by X_2 and X_3,4th component is highly influenced by the variables X_2 followed by X_1 and X_4 and the 5th component is highly influenced by the variables X_4 followed by X_5 and X_1.

From the component matrix,

Table 4: List of variables contributing more towards variability

Components				
1	2	3	4	5
X_3 (Life expectancy)	X_6 (No: of take out places)	X_5 (Outdoor Activities)	X_2 (Sunshine hours)	X_4 (Happiness levels)
X_4 (Happiness levels)	X_5 (Outdoor Activities)	X_2 (Sunshine hours)	X_1 (Rank)	X_5 (Outdoor Activities)
X_6 (No: of take out places)	X_1 (Rank)	X_3 (Life expectancy)	X_4 (Happiness levels)	X_1 (Rank)

Let the principal components be U_1, U_2, U_3, U_4 *and* U_5

From the score coefficient matrix we get,

$$U_1 = -.437X_1 - .547X_2 + .895X_3 + .872X_4 - .158X_5 + .021X_6$$

$$U_2 = .401X_1 - .286X_2 + .108X_3 + .014X_4 + .692X_5 + .944X_6$$

$$U_3 = -.699X_1 + .455X_2 + .086X_3 - .036X_4 + .643X_5 - .046X_6$$

$$U_4 = .338X_1 + .636X_2 + .213X_3 + 331X_4 - .086X_5 + .083X_6$$

$$U_5 = .089X_1 - .075X_2 - 292X_3 + .337X_4 + .196X_5 - .176X_6$$

3.2 Factor analysis

Table 5: Communalities

Communalities		
	Initial	Extraction
Rank	1.000	.729
Sunshine hours	1.000	.999
Life expectancy	1.000	.869
Happiness levels	1.000	.816
Number of take out places	1.000	.787

One or more communality estimates greater than 1 were encountered during iterations. The resulting solution should be interpreted with caution.

Table 6: Total Variance Explained

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cum %	Total	% of Variance	Cum %	Total	% of Variance	Cum %
1	2.068	41.355	41.355	2.068	41.355	41.355	1.827	36.541	36.541
2	1.424	28.481	69.836	1.424	28.481	69.836	1.357	27.136	63.676
3	.707	14.149	83.985	.707	14.149	83.985	1.015	20.309	83.985
4	.553	11.064	95.049						
5	.248	4.951	100.000						

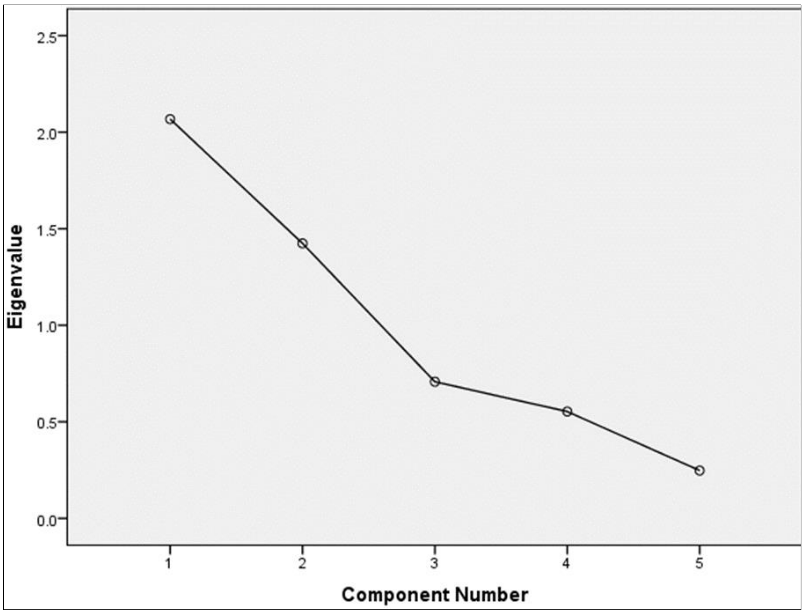

Fig 4: Scree plot

Table 7: Component Matrix

Component Matrix			
	Component		
	1	2	3
Rank	-.425	.728	.133
Sunshine hours	-.555	-.423	.715
Life expectancy	.903	.004	.231
Happiness levels	.867	-.022	.252
Number of take-out places	.106	.845	.249
Extraction Method: Principal Component Analysis.			
a. 3 components extracted.			

Table 8: Rotated Component Matrix

Rotated Component Matrix			
	Component		
	1	2	3
Rank	-.346	.780	.025
Sunshine hours	-.194	-.088	.976
Life expectancy	.916	-.038	-.168
Happiness levels	.893	-.050	-.126
Number of take-out places	.182	.858	-.131
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.			
a. Rotation converged in 4 iterations.			

SPSS calculates the factor loadings for each variable in the analysis. The loadings of a factor explain each variable. Large loadings (positive or negative) indicate the high influences the variable. Small loadings (positive or negative) indicate that the influence on the variable (5).

Unrotated factor loadings are often difficult to interpret. Factor rotations simplify structure, and make the factor loadings easier to interpret.

Table 9: Component Transformation Matrix

Component Transformation Matrix			
Component	1	2	3
1	.907	-.130	-.400
2	-.022	.935	-.354
3	.420	.330	.845
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.			

Interpret the output in the same way as PCA although the Component matrix is called the Factor Matrix if the extraction method has changed.

Look at the Rotated Factor matrix to see which variables contribute most to each factor (PC). Variables measuring the same underlying latent variable should all have high loadings on a particular factor and by looking at the raw variables, a sensible name can be given to the factor. The next factor should be measuring another latent variables etc.

The factor plot is useful for assessing grouping of variables on more than one factor. If there are two factors, the variables appear on a scatterplot.

Using the scree plot, it can be concluded that we can extract 3 Factors.

The Factors are

$$Y_1 = -.346X_1 - .194X_2 + .916X_3 + .893X_4 + .182X_5$$

$$Y_2 = .780X_1 - .088X_2 - .038X_3 - .050X_4 + .858X_5$$

$$Y_3 = .025X_1 + .976X_2 - .168X_3 - .126X_4 - .131X_5$$

It is seen that the 1^{st} component explains 36.541% of variation of the data set, 2^{nd} component explains 27.136% of data set and 3^{rd} component explains 20.309% of variation of the data set. That is, first 3 component explain 83.985% of variation of data set. It is also from the component matrix that the first 3 components are highly influenced by all the 5 factors.

3.3 Cluster analysis
Here we are comparing average linkage cluster and complete linkage cluster

3.3.1 Complete Linkage Method

Table 10: Cluster Membership

Cluster 1	Cluster 2	Cluster 3
Amsterdam	Vienna	Jakarta
Sydney	Stockholm	Cairo
Barcelona	Copenhagen	Mumbai
Tokyo	Helsinki	Johannesburg
Paris	Fukuoka	
London	Berlin	
New York	Vancouver	
	Melbourne	
	Beijing	
	Bangkok	
	Buenos Aires	
	Toronto	
	Madrid	
	Seoul	
	Frankfurt	
	Geneva	
	Tel Aviv	
	Istanbul	
	Taipei	
	Los Angeles	
	Boston	
	Dublin	
	Chicago	
	Hong Kong	
	Shanghai	
	Brussels	
	San Francisco	
	Sao Paulo	
	Zurich	
	Milan	
	Washington, D.C.	
	Moscow	
	Mexico City	

The cluster plot is given in Figure 3

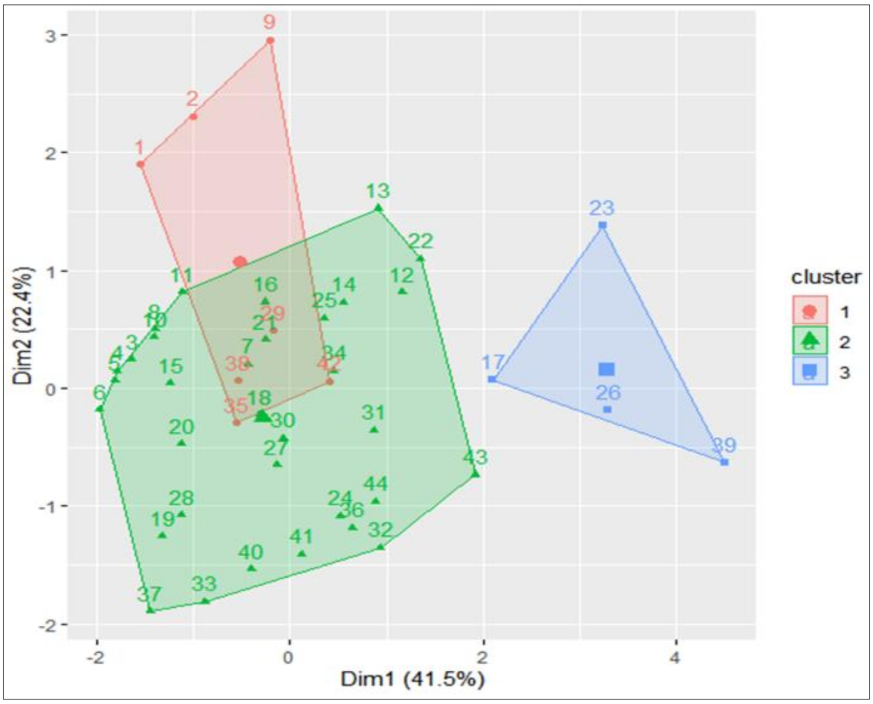

Fig 5: Cluster plot

3.3.2 Average Linkage method

Table 11: Cluster Membership

Cluster 1	Cluster 2	Cluster 3
Amsterdam	Vienna	Jakarta
Sydney	Stockholm	Cairo
Barcelona	Copenhagen	Mumbai
	Helsinki	Johannesburg
	Fukuoka	
	Berlin	
	Vancouver	
	Melbourne	
	Beijing	
	Bangkok	
	Buenos Aires	
	Toronto	
	Madrid	
	Seoul	
	Frankfurt	
	Geneva	
	Tel Aviv	
	Istanbul	
	Taipei	
	Los Angeles	
	Boston	
	Dublin	
	Tokyo	
	Chicago	
	Hong Kong	
	Shanghai	
	Brussels	
	San Francisco	
	Paris	
	Sao Paulo	
	Zurich	
	London	
	Milan	
	Washington, D.C.	
	New York	
	Moscow	
	Mexico City	

The cluster plot is given in Figure 4

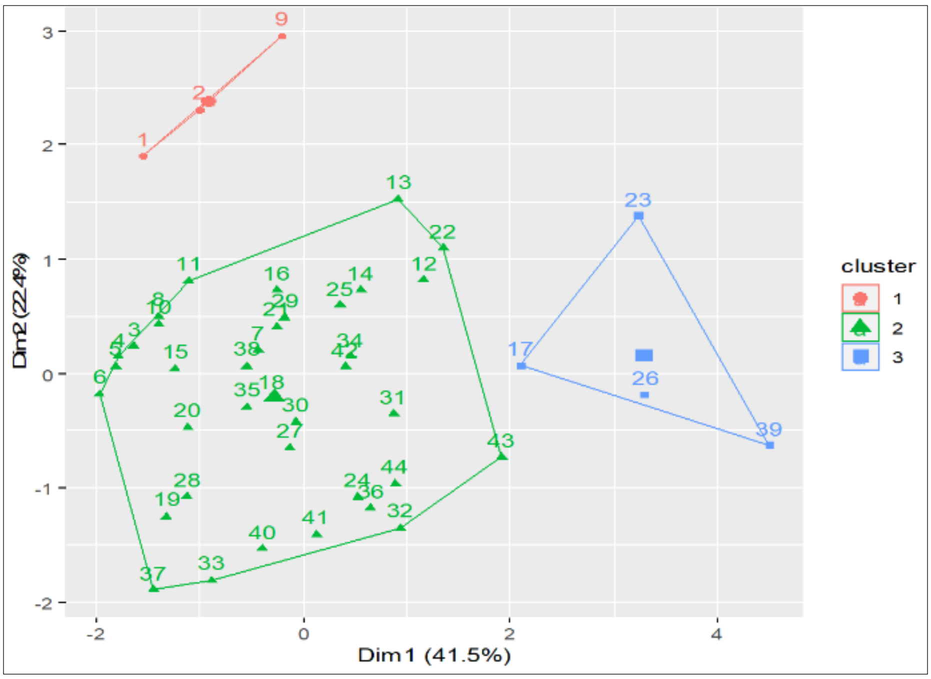

Fig 6: Cluster plot

Table 12: Final Clusters

Final clusters	
	Member Average Linkage
Member Complete Linkage	1 2 3
	1 3 4 0
	2 0 33 0
	3 0 0 4

This table tell us that using average linkage method, there are 3 observations belong to cluster 1. Four plus 33 observations belong to cluster 2 and 4 observations belong to cluster 3.

Using complete linkage method, there are 3 plus 4 Cities that belong to cluster 1, 33 belong to cluster 2 and 4 belong to cluster 3.

If we compare Average linkage method and Average linkage there are good match for 3 cities, both methods listed them as cluster one. Whereas, there are 33 cities both indicated belong to cluster 2 and also 4 cities belong to cluster 3.

But also we can see there is some mismatch, 4 cities have membership in cluster 2 based on Average method. But this cities have membership in cluster 1 if you use complete linkage method.

So this table allow us to compare these two different methods, Complete linkage and Average Linkage Method.

We can also calculate Cluster means,

Average Value of Complete linkage

Table 13: Average value of Complete Linkage

Group	Rank	Sunshine hours	Life Expectancy	Happiness levels	Outdoor Activities
1	-1.44022305	0.1867216	0.6896830	0.5901925	2.02077308
2	0.08521407	-0.1686275	0.1983321	0.1615732	-0.15579106
3	0.29193710	1.4197628	-2.3518346	-2.0000969	-0.07451255

The average value help us to find out which is important variable. For example in case of outdoor activities 2.02077308 is the highest value which means most of outdoor activities occurs at cluster one and -0. 15579106 the lowest value which belongs to cluster 2 indicates minimal outdoor activities occurs at cluster 2. These Averages indicate which variables are really playing an important role in characterizing the clusters.

Average Value of Average linkage method

Table 14: Average value of Average Linkage

Group	Rank	Sunshine hours	Life Expectancy	Happiness levels	Outdoor Activities
1	-0.01668212	-0.2531972	0.6052961	0.381211	1.60070509
2	-0.03184768	-0.1183840	0.1566747	0.161573	-0.33051168
3	0.29193710	1.4197628	-2.3518346	-2.000097	-0.07451255

The average value help us to find out which is important variable. For example in case of outdoor activities 1.60070509 is the highest value which means most of outdoor activities occurs at cluster one and -0. 33051168 the lowest value which belongs to cluster 2 indicates minimal outdoor activities occurs at cluster 2. These Averages indicate which variables are really playing an important role in characterizing the clusters.

K-mean Cluster

The important step in K-means clustering technique is to decide the number of clusters.

So here we choose k=4

Table 15: Initial Clusters Centers

Initial Cluster Centers				
	Cluster			
	1	2	3	4
Rank	38.0000	36.0000	23.0000	37.0000
Sunshine hours	1633.0000	2003.0000	3542.0000	1566.0000
Life expectancy	80.40000	73.90000	70.70000	82.60000
Happiness levels	7.1600	6.3700	4.1500	7.5600
Outdoor activities	433.0000	158.0000	323.0000	69.0000
Number of take out places	6417.00000	3355.00000	250.00000	538.00000

Table 16: Iteration History

Iteration History				
Iteration	Change in Cluster Centers			
	1	2	3	4
1	331.650	400.172	723.467	596.835
2	.000	.000	78.415	51.946
3	.000	.000	49.468	34.404
4	.000	.000	146.533	146.671
5	.000	.000	26.163	33.417
6	.000	.000	.000	.000

Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is. 000. The current iteration is 6. The minimum distance between initial centers is 2013.054.

Table 17: Cluster Membership

Cluster Membership			
Case Number	**City**	**Cluster**	**Distance**
1	Amsterdam	4	344.920
2	Sydney	3	341.566
3	Vienna	4	222.179
4	Stockholm	4	239.395
5	Copenhagen	4	332.622
6	Helsinki	4	530.842
7	Fukuoka	3	386.907
8	Berlin	4	917.351
9	Barcelona	2	845.329
10	Vancouver	4	181.812
11	Melbourne	3	444.772
12	Beijing	3	641.573
13	Bangkok	3	941.124
14	Buenos aires	3	612.666
15	Toronto	4	882.962
16	Madrid	2	793.143
17	Jakarta	3	209.812
18	Seoul	4	531.111
19	Frankfurt	4	330.863
20	Geneva	3	480.639
21	Tel aviv	3	697.151
22	Istanbul	4	525.190
23	Cairo	3	988.630
24	Taipei	4	378.890
25	Los angeles	3	714.559
26	Mumbai	3	364.089
27	Boston	3	359.783
28	Dublin	4	354.732
29	Tokyo	1	331.650
30	Chicago	3	520.336
31	Hong kong	4	448.820
32	Shanghai	4	484.849
33	Brussels	4	298.762
34	San francisco	3	303.000
35	Paris	2	1433.152
36	Sao paulo	2	400.172
37	Zurich	4	366.522
38	London	1	331.650
39	Johannesburg	3	514.118
40	Milan	2	721.890
41	Washington,d.c.	3	327.673
42	New york	2	347.401
43	Moscow	2	343.325
44	Mexico city	3	489.337

Table 18: Final Cluster Centers

Final Cluster Centers				
	Cluster			
	1	**2**	**3**	**4**
Rank	33.5000	31.5714	22.5556	17.4118
Sunshine hours	1755.0000	2196.5714	2798.5000	1765.4706
Life expectancy	81.80000	78.72857	76.13889	79.67647
Happiness levels	6.5150	6.3843	6.1689	6.7282
Outdoor activities	410.0000	297.2857	196.1111	175.5294
Number of take-out places	6109.50000	3033.71429	889.11111	825.76471

Table 19: Distance between Final Centers

Distances between Final Cluster Centers				
Cluster	**1**	**2**	**3**	**4**
1		3109.367	5327.969	5288.970
2	3109.367		2229.790	2252.979
3	5327.969	2229.790		1035.193
4	5288.970	2252.979	1035.193	

The initial cluster centres are given in the table 20 followed by the changes to clusters centres in the iteration history. The last row should show negligible change, The final cluster centres show how the variables differ in each cluster. It should be clear which variables are most different and therefore define each cluster but ANOVA table shows which variables contribute most to the separation(Highest F-statistics) and least.

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 20: Number of cases in each Cluster

Number of Cases in each Cluster		
Cluster	1	2.000
	2	7.000
	3	18.000
	4	17.000
Valid		44.000
Missing		1.000

4. Summary and Conclusions

4.1 Principal component

Using the scree plot, it can be concluded that we can extract three principal components.
The components are,

$$U_1 = -.437X_1 - .547X_2 + .895X_3 + .872X_4 - .158X_5 + .021X_6$$

$$U_2 = .401X_1 - .286X_2 + .108X_3 + .014X_4 + .692X_5 + .944X_6$$

$$U_3 = -.699X_1 + .455X_2 + .086X_3 - .036X_4 + .643X_5 - .046X_6$$

$$U_4 = .338X_1 + .636X_2 + .213X_3 + 331X_4 - .086X_5 + .083X_6$$

$$U_5 = .089X_1 - .075X_2 - 292X_3 + .337X_4 + .196X_5 - .176X_6$$

- It is seen that the 1st component explains 34.625% of variation of the data set, 2nd component explains 27.071% of data set, 3rd component explains 18.654% of variation of the data set, 4th component explains 11.464% of variation of the data set and 5th component explains 4.699% of data set. That is, first five component explain 96.514% of variation of data set
- Also we get, the first component is highly influenced by the variables X_3followed by X_4 and X_6, second component is highly influenced by the variables X_6 followed by X_5 and X_1,3rd component is highly influenced by the variables X_5 followed by X_2 and X_3,4th component is highly influenced by the variables X_2 followed by X_1 and X_4 and the 5th component is highly influenced by the variables X_4 followed by X_5 and X_1.

4.2 Factor Analysis

Using the scree plot, it can be concluded that we can extract 3 Factors.
The Factors are,

$$Y_1 = -.346X_1 - .194X_2 + .916X_3 + .893X_4 + .182X_5$$

$$Y_2 = .780X_1 - .088X_2 - .038X_3 - .050X_4 + .858X_5$$

$$Y_3 = .025X_1 + .976X_2 - .168X_3 - .126X_4 - .131X_5$$

- It is seen that the 1st component explains 36.541% of variation of the data set, 2nd component explains 27.136% of data set and 3rd component explains 20.309% of variation of the data set. That is, first 3 component explain 83.985% of variation of data set. It is also from the component matrix that the first 3 components are highly influenced by all the 5 factors.

4.3 Cluster Analysis

Complete Linkage and Average Linkage Method

Table 21: Final Clusters

Final clusters	
Member Average Linkage	
Member Complete Linkage	1 2 3 1 3 4 0 2 0 33 0 3 0 0 4

This table tell us that using average linkage method, there are 3 observations belong to cluster 1. Four plus 33 observations belong to cluster 2 and 4 observations belong to cluster 3.

Using complete linkage method, there are 3 plus 4 Cities that belong to cluster 1, 33 belong to cluster 2 and 4 belong to cluster 3.

If we compare Average linkage method and Average linkage there are good match for 3 cities, both methods listed them as cluster one. Whereas, there are 33 cities both indicated belong to cluster 2 and also 4 cities belong to cluster 3.

But also we can see there is some mismatch, 4 cities have membership in cluster 2 based on Average method. But this cities have membership in cluster 1 if you use complete linkage method.

So this table allow us to compare these two different methods, Complete linkage and Average Linkage Method.

K-means

Table 22: The final clusters are,

Final Cluster Centers				
	Cluster			
	1	**2**	**3**	**4**
Rank	33.5000	31.5714	22.5556	17.4118
Sunshine	1755.0000	2196.5714	2798.5000	1765.4706
Life expectancy	81.80000	78.72857	76.13889	79.67647
Happiness levels	6.5150	6.3843	6.1689	6.7282
Outdoor activities	410.0000	297.2857	196.1111	175.5294
Number of take out places	6109.50000	3033.71429	889.11111	825.76471

5. References

1. Bartlett MS. Multivariate Analysis, Journal of Royal Statistics Society. 1947;2:176-197.
2. Cadima J, Cerdeira JO, Minhoto M. Computational aspects of algorithms for variable selection in the context of principal components. Comp. Stat. Data Anal. 2004;47:225-236.
3. Cadima J, Jolliffe IT. On relationships between uncentred and column-centred principal component analysis. Pak. J Stat. 2009;25:473-503.
4. Gorsuch RL. Factor Analysis (2nd ed). Hillsdae, NJ: Erlbaum; c1983.
5. Hallin M, Paindaveine D, Verdebout T. Efficient R-estimation of principal and common principal components. J Am. Stat. Assoc. 2014;109:1071-1083.
6. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ. Psychol. 1933;24:417-441, 498-520.
7. Jackson JE. A user’s guide to principal components. New York, NY: Wiley; c1991.
8. Jolliffe IT. Principal component analysis, 2nd edn. New York, NY: Springer-Verlag; c2002.
9. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments, Phil. Trans. R. Soc. A. 2016;374:20150202.
10. Li Y, Wang N, Carroll RJ. Selecting the number of principal components in functional data. J Am. Stat. Assoc. 2013;108:1284-1294.
11. Rao CR. Tests of significance in Multivariate Analysis, Biometrika. 1948;35:58-79.
12. Romsberg H. Cluster Analysis for Researchers, Lulu Press; c2004.
13. Sharma M, Wadhawan P. A Cluster Analysis Study of Small and Medium Enterprises, The IUP Journal of Management Research. 2009, 8(10).
14. SPSS Inc. SPSS 16.0 [Computer software]. Chicago: Author; c2007.