CHAPTER 1

INTRODUCTION

Statistics is the study of collection, analysis, interpretation, presentation and organization of data. In other words, statistics is the practice or science of collecting analyzing numerical data in large quantities especially for the purpose of inferring proportions in a whole from those in a representative sample.

Statistical methods can be used to summarize or describe a collection of data; this is called descriptive statistics. This is useful in research, when communicating the results of experiments. In additions patterns in the data may be modelled in a way that accounts for randomness and uncertainty in the observations and are used to draw the inferences about the processor population being studied; this is called inferential statistics.

Many problems are called upon to solve in our daily lives for which the observations are multivariate in nature. Data of this type arise in all branches of science. They often involve the consideration of several factors or variables for a number of individuals or objects. Sometimes, a distinction is made between univariate statistics and multivariate statistics, where univariate statistics have only one dependent variable, whereas multivariate statistics have two or more dependent variables.

There are many statistical techniques for conducting multivariate analysis and the most appropriate techniques for a given study varies with the type of study and the key research questions. Four of the most common techniques are multiple regression analysis, factor analysis, path analysis and multiple analysis of variable (MANOVA).

Many of the multivariate statistical methods require the assumption that the data being analyzed have multivariate normal distributions. One important aspect of multivariate normal distribution is that it is specified completely by a mean vector and a dispersion matrix.

Statistics form a key basic tool in business and manufacturing as well. It is used to understand measurement systems variability, control processes (as in statistical process control or SPC), for summarizing data, and to make data-driven decisions. In these roles, it is a key tool, and perhaps the only reliable tool.

Alcohol consumption is known to be one of the most important problems in today's society. The concerns are obvious: according to the World Health Organization, about 2.5 million people die every year because of alcohol abuse. Alcohol consumption has an impact on various aspects of a person's life: not only on one's health and mortality rate but also on social life, economic activity and subjective well-being. Numerous studies found ample evidence of the drinking habit impact on individual lives. It is accepted in economic literature to classify alcohol drinkers into two groups: heavy drinkers and moderate ones. Empirical research shows that the effect of alcohol consumption depends on the degree of drinking. For abstainers and heavy drinkers alcohol consumption leads to reductions in productivity and earnings and to decrease in subjective well-being. On the other hand, moderate drinkers increase their productivity and earnings and feel happier.

Alcohol consumption in Russia is known to be one of the most serious social problems. Struggle against alcoholism has a long history. Today the government is introducing special programs and measures to decrease alcohol drinking, particularly, among the youth. Alcohol abuse is believed to destroy human capital and considerably expand social costs.

Alcohol consumption is a field of research agenda for economists and social scientists. Drinking influences all aspects of an individual's life: health, social and economic life. From the economic point of view, it is interesting to study the correlation between alcohol consumption and labor market indices. Drinking alcohol can affect a worker's productivity and a person's earnings. Alcohol consumption also influences human capital.

One of the earlier investigations which estimated the consequences of alcohol consumption was the book by Irving Fisher published in 1926 (Fisher,

1926). The author studied the effects of alcohol consumption on a worker's productivity. The results showed that drinking alcohol has an unfavorable influence on productivity, it reduces proficiency and speed of some tasks fulfillment. Fisher, in particular, remarks that three glasses of beer per day lower productivity by about 10 percent.

In economic literature, the effects of alcohol consumption on different parts of personal and social life have been examined. Recent researches show that the impact of alcohol consumption depends on the frequency and quantity of drinking. For moderate drinking the effect is positive: the productivity increases, a person earns more and feels happier. But for abstainers and heavy drinkers the results are opposite. In any case alcohol consumption affects human capital and causes huge social costs.

There are several substantial caveats to the evidence of a pleasurable effect. There is a relatively small overspill of happiness into moments when people are not drinking (a difference of less than 0.5 points on a 0-to-100 scale between those weeks or months in which people drink more versus less often). What's more, looking at year-to-year changes, people are no more satisfied with life in heavier-drinking years than in lighter-drinking years. Indeed, if they develop a drink problem, then they become noticeably less satisfied with life (by about 0.2 points on a 0 to 10 scale).

Drinking makes hanging out with friends a little more pleasurable. But it can go a long way toward easing the pain of unpleasant activities, like the morning subway commute. The study was able to control for this, by looking at how happy people reported being earlier in the day. After controlling for prior happiness, the researchers found that drinking had a slightly smaller boost on overall happiness, but the effect was still significant. That suggests it's more the case that drinking makes you happy, rather than the other way around.

1.1 SOURCE OF DATA

The data is based on the study of Happiness and Alcohol Consumption of United Nations in the year 2016 and was collected from University of North Carolina at Charlotte – Data Science and Business Analytics website. The data used in the project is a secondary data and the source of the data is *https://data.world/uncc-dsba/dsba-6100-fall-2016* .

In the data there are 122 countries and 5 variables.

The variables under study are

- HDI
- GDP Per Capita
- Beer Per Capita
- Spirit Per Capita
- Wine Per Capita

1.1.1 ABOUT VARIABLES

HDI

Human Development Index (HDI) is a statistical tool used to measure a country's overall achievement in its social and economic dimensions. The social and economic dimensions of a country are based on the health of people, their level of education and their standard of living. HDI is used to measure the country's development by the United Nations Development Program (UNDP). Every year UNDP ranks countries based on HDI report released in their annual report.

HDI is one of the best tools to keep track of the level of development of a country, as it combines all major social and economic indicators that are responsible for economic development.

GDP Per Capita

GDP per capita is a measure of a country's economic output that accounts for its number of people. It divides the country's gross domestic product by its total population. That makes it the best measurement of a country's standard of living. It tells you how prosperous a country feels to each of its citizens.

Alcohol per capita is the total amount of alcohol consumed per adult (15+years) in a calendar year, in litres of pure alcohol. Various alcohol drinks include Beer, Spirit and Wine.

Beer Per Capita

It is the annual per capita consumption of beer of countries.

Spirit Per Capita

It is the annual per capita consumption of spirits of countries.

Wine Per Capita

It is the consumption of wine per capita of countries.

1.2 OBJECTIVES OF THE STUDY

The main objectives of the study are:

- To find the relationship between dependent variable and an independent variable.
- To group the countries based on their similarities.
- To find the variable which explains maximum information on the happiness of a country.
- To identify the country which is less happier due to alcohol consumption.

1.3 IMPORTANT TECHNIQUES USED

CLUSTER ANALYSIS

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is a main task of explanatory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics.

In other words, Cluster analysis is an explanatory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same groups and minimal otherwise. Also, it can be used to discover structures in data without providing an explanation or interpretation. In other words, it simply discovers structures in data without explaining why they exist.

MULTI DIMENSIONAL SCALING

Multidimensional Scaling is the term used to describe any procedure which starts with the 'distances' between a set of points (or individuals or objects) or information about those 'distances' and finds a configuration of the points, preferably in a small number of dimensions, usually 2 or 3. By configuration, we mean a set of coordinate values. For example; if we are given the road distances between all pairs of English towns, can we reconstruct a map of England? The map will of course be a two-dimensional configuration.

The main two types of scaling procedure are called 'classical scaling' and 'ordinal scaling'. The first of these two procedures is essentially an algebraic method of reconstructing the point coordinates assuming that the dissimilarities are Euclidian distances, although the method is robust to the situation where the distances are distorted by errors. Because classical scaling assumes the given 'distances' to be Euclidian, it is not suitable when the observed dissimilarities are

such that the actual numerical values are of little significance and the order of the dissimilarities through the only relevant information.

In other words, Multidimensional Scaling (MDS) is a mathematical tool that uses proximities between objects, subjects or stimuli to produce a special representation of these items. It is means of visualizing the level of similarities of individual cases of dataset. One important tool of visualizing data is to get a sense of how near or far points are from each other by using scatter plot.

DIMENSION REDUCTION METHODS

PRINCIPAL COMPONENT ANALYSIS AND FACTOR ANALYSIS

In many statistical studies, the number of variables under consideration is too large to handle. One way of reducing the number of variables to be treated is to consider only some of the linear combination which have small variance and consider only those with high variances.

The Principle Component Analysis is a technique to find a few linear combinations of original variables which can be used to summarize the data loosing as little information as possible. In other words, it is a technique to transform the original set of variables into a smaller set of linear combinations so that most of the variations in the original data set is explained by those linear combinations. The linear combinations so selected are called principal components. The objective of this analysis is to reduce the number of variables into a few factors that can explain the most of the variance of the original data set. The number of principal component is less than or equal to the number of original variables. This transformation is defined in such a way that the first PC has the largest possible variance and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to the preceding component. PCs are guaranteed to be independent only if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of original variables.

One of the important property of PC is that they can be considered as uncorrelated variables, obtained by an orthogonal relation of the coordinate system. Principal Components are linear combinations of random or statistical variables which have special properties in terms of the variances.

The first PC is the normalized linear combination with maximum variance. The second PC is the normalized linear combination which has the second largest variance and uncorrelated with the first PC. Thus the study of PC turns out to be that of the characteristic roots and vectors of the dispersion matrix which is either positive definite or positive semi definite.

The essential purpose of factor analysis is to describe if possible the covariance relationship among many variables in terms of few underlying, but unobservable, random quantities called factors. Factor Analysis is related to principal component analysis (PCA), but the two are not identical. Both technique attempts to approximate the covariance matrix $\sum$. First factors are those variables which has high correlation with first Principal Component.

1.4 PACKAGES USED

- **SPSS (Student's Version)**

SPSS (Statistical Package for the Social Sciences) is a software package used for statistical analysis. Long produced by SPSS Inc., it was acquired by IBM in 2009. The current versions (2015) are officially named IBM SPSS Statistics. Companion products in the same family are used for survey authoring and deployment (IBM SPSS Data Collection), data mining (IBM SPSS Modeler), text analytics, and collaboration and deployment (batch and automated scoring services).

The software name originally stood for Statistical Package for the Social Sciences (SPSS), reflecting the original market, although the software is now popular in other fields as well, including the health sciences and marketing.

SPSS is a widely used program for statistical analysis in social science. It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners, and others. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary was stored in the data file) are features of the base software.

1.5 DESCRIPTIVE STATISTICS

Descriptive statistics can be used to summarize the data when analyzing a dataset. One usually starts by determining some indices that give a global picture on where and how the data is concentrated and what is the shape of the distribution. Indices are useful for the purpose of summarizing the data. These indices are known as descriptive statistics.

Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include the standard deviation, variance, the minimum and maximum variables and the kurtosis and skewness. Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of descriptive statistics are the mean, median, and mode, which are used at almost all levels of math and statistics. However, there are less-common types of descriptive statistics that are still very important.

SCATTER DIAGRAM AND REGRESSION LINE

Scatter plots also called scatter graphs are similar to line graphs. A line graph uses a line on an X-Y axis to plot a continuous function, while a scatter plot uses dots to represent individual pieces of data. In statistics, these plots are useful to see if two variables are related to each other

Scatter plots are used when we want to show the relationship between two variables. Scatter plots are sometimes called correlation plots because they show how two variables are correlated.

A scatter plot gives you a visual idea of what is happening with the data. Scatter plots in statistics create the foundation for simple linear regression, where we take scatter plots and try to create a usable model using functions. In fact, all regression doing is trying to draw a line through all of those dots.

When line graphs are plotted on X axis and Y axis, the X variable is sometimes called the independent variable and Y variable is called the dependent variable. Linear Regression also called Simple linear regression plots one independent variable X against one dependent variable Y. Technically, in regression analysis ,the independent variable is usually called predictor variable and the dependent variable is called criterion variable. When we make a distribution in which there is an involvement of more than one variable, then such an analysis is Regression Analysis. It generally focuses on finding or rather predicting the value of the variable that is dependent on the other.

R–squared is a goodness of fit measure for linear regression models. R-squared value evaluates the scatter of the data points around the fitted regression line. For the same dataset, higher R-squared values represent smaller differences between the observed data and the fitted values. R-squared is the percentage of the dependent variable variation that a linear model explains.

Here the dependent variable HappinessScore is a metric measured in 2016 by asking the sampled people the question “How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest”.

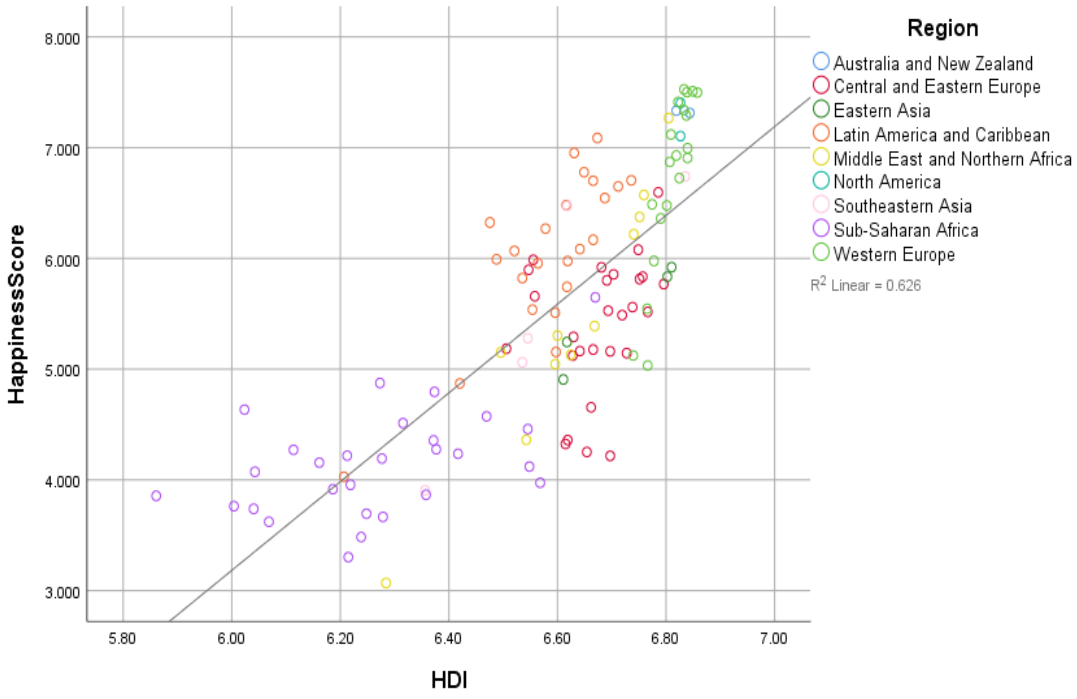

Figure 1.1 Scatter diagram with regression line

Figure 1.1 shows the relationship between the independent variable HDI and a dependent variable Happiness Score in different regions. From the figure it is clear that as HDI increases Happiness Score increases. Here R-squared value is 62.6%.This statistic measures the stength of the relationship between the dependent variable and the independent variable. Here it shows a smaller difference between observed data and the fitted values.

The present work is organized into four chapters.Introductory chapter deals with a brief description of data, source of data, objectives of study, list of variables, software used and important techniques used.Also a scatter plot diagram is depicted in this chapter.Chapter 2 deals with grouping of countries based on their similarities.In the third chapter we use multidimensional scaling to study the relative position of countries based on happiness and alcohol consumption.In the fourth chapter we examine the dimension reduction of data using principal component analysis and factor analysis. A summarized conclusion is given in the last chapter to discuss the findings of the analysis carried out in this work.

CHAPTER 2

CLUSTER ANALYSIS

2.1 INTRODUCTION

Clustering is different from classification. Classification pertains to a known number of groups and the operational objective is to assign new observations to one of these groups. Cluster analysis is more primitive technique in that no assumptions are made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities). The inputs required are similarity measures or data from which similarities can be computed. In most practical applications of Cluster analysis, the investigator knows enough about the problem to distinguish "good" grouping from the "bad" grouping.

In Cluster analysis multivariate technique are used to identify natural grouping among the units under consideration. Clustering is the grouping of similar objects using data. It is part of the general scientific process of searching for patterns in data and then trying to construct laws that explain the pattern. However, the goals of cluster analysis are varied and include widely different activities such as data snooping, looking for natural groups of like objects to form the first stage of a stratified sampling scheme, hypothesis generation, and searching for a suitable classification scheme, two notable examples being the classification of plants and animals and the classification of diseases.

Cluster analysis is also used to group variables into homogeneous and distinct groups. This approach is used, for example, in revising a questionnaire on the basis of responses received to a draft of the questionnaire. The grouping of questions by means of cluster analysis helps to identify redundant questions and reduce their number, thus improving the chances of a good response rate to the final version of the questionnaire. Cluster analysis goes under a number of names

including classification, pattern recognition, numerical taxonomy and morphometric. There are a countless number of examples in which clustering plays an important role. For instance, biologist has to organize the different species of animals before a meaningful description of the differences between animals is possible. Clustering techniques have been applied to a wide variety of research problems.

Cluster analysis is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy analysis. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification primarily their discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goals.

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection).

Possible Applications of clustering are,

Marketing: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records.

Biology: classification of plants and animals given their features.

Libraries: book ordering.

Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds.

City-planning: identifying groups of houses according to their house type, value and geographical location.

Earthquake studies: clustering observed earthquake epicenters to identify dangerous zones.

2.2 SIMILARITY MEASURES

The most important steps in classification procedure is to suggest a measure of "closeness" or "similarity" between the units based on the p variables considered in the study. When items are clustered, proximity is usually indicated by some sort of distance. On the other hand, variables are usually grouped on the basis of correlation coefficients or some measure of association.

Consider two p-dimensional observations $X' = [X_1, X_2, ..., X_P]$ and $Y' = [Y_1, Y_2, ..., Y_P]$. Then the Euclidean distance between X and Y is,

$$d(X,Y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + ... + (X_P - Y_P)^2}$$

$$= \sqrt{(X - Y)'(X - Y)}$$

The statistical distance between the same two observations is of the form

$$d(X, Y) = \sqrt{(X - Y)' A(X - Y)}$$

Ordinarily, A= S^{-1}, where S contain the sample variances and covariance. However, without any prior knowledge of this distinct groups, these sample quantities cannot be computed. For this reason Euclidean distance is often preferred for clustering.

Another distance measure is the Minkowski distance

$$d(X, Y) = \left[\sum_{i=1}^{p} | Xi - Yi |^m\right]^{1/m}$$

For m=1, d(X, Y) measures the "city-block" distance between the two points in p-dimension.

For m=2, d(X, Y) becomes the Euclidean distance. In general, varying m changes the weight given to the larger and smaller differences.

Two additional popular measures of "distance" are given by the Canberra metric and Czekanowski coefficient. Both of these measures are defined foe non negative variables only. Both are ratio type measures.

We have,

Canberra metric: $d(X, Y) = \sum_{i=1}^{p} \frac{| X_i - Y_i |}{(X_i + Y_i)}$

Czekanowski coefficient: $d(X, Y) = 1 - \frac{2\sum_{i=1}^{p} \min(X_i, Y_i)}{\sum_{i=1}^{p} (X_i + Y_i)}$

2.3 HIERARCHICAL CLUSTERING METHODS

Hierarchical clustering methods proceed by either a series of successive merges or a successive division. There are two types of Hierarchical clustering methods.

1) Agglomerative hierarchical methods

2) Divisive Hierarchical methods

Agglomerative hierarchical clustering

In agglomerative hierarchical methods start with individual objects. Thus there are initially as many clusters as objects. Suppose there are initially 'n' objects. In this case first we fuse 2 objects into a single cluster such that they are most near objects. So now we have (n-1) clusters. In the second step we fuse two nearest clusters among these (n-1) clusters. We continue this procedure until a single cluster is formed. Single linkage, complete linkage, average linkages are important agglomerative hierarchical methods. Single linkage occurs when groups are fused according as to the distance between their nearest members. Complete linkages occurs when groups are fused according as to the distance between there furthest members and in average linkage we consider the average distance between pair of members in the set.

Algorithm

Step 1

Start with N clusters. Each contain a single entity and NxN symmetric matrix of distances $D = \{d_{ik}\}$.

Step 2

Search the distance matrix for most similar pair of clusters. Let the distance between most similar clusters *u* and *v* be d (*u*, *v*).

Step 3

Merge the clusters *u* and *v* and label newly formed cluster as *uv*. Update the entries in the distance matrix by deleting the row and column corresponding to the clusters *u* and *v* and adding a new row and column for *uv*.

Step 4

Repeat step 2 and 3 (N-1) times so that all objects will form a single cluster.

- **Single Linkage**

Single linkage algorithm connects two clusters *u* and *v* by considering the nearest neighbor in this cluster. Initially we find the smallest distance in D, where $D = \{d_{ik}\}$ and merge the corresponding objects *u* and *v* to form new cluster *uv*. The distance between the new cluster *uv* and any other object w is given by,

$$d(uv)w = \min(d_{uw}, d_{vw}) \ ,$$

where d_{uw} and d_{vw} are the distance between objects in the cluster *uv* and *w*.

- **Complete Linkage**

Complete linkage clustering proceeds in the same manner as single linkage with one difference is that at each stage the distance between clusters is determined by the distance between two elements, one from each cluster that are most distinct. Thus complete linkage ensures that all items in a cluster are within some maximum distance of each other. The agglomerative algorithm start with finding minimum entity in the distance matrix $D = \{d_{ik}\}$ and merging the corresponding objects *u* and *v* to form new cluster *uv*. Then distance between new cluster *uv* and any other object *w* is computed as

$$d(uv)w = \min(d_{uw}, d_{vw}) \ ,$$

where d_{uw} and d_{vw} are the distance between objects in the cluster *uv* and *w*.

- **Average Linkage**

Average linkage procedure is different from single linkage and complete linkage. In this procedure the distance between clusters as the average distance between all pair of items where one member of a pair belongs to each cluster. The average linkage algorithm start with the same manner as that of general algorithm by searching the minimum distance in the distance matrix $D = \{d_{ik}\}$ and pooling the objects *u* and *v* together which have the minimum distance. So we get the new cluster namely *uv*. In the next step the distance between the cluster *uv* and *w* is computed as

$$\mathrm{d}(uv)w = \frac{\sum_i \sum_k d_{ik}}{N(uv) * N(w)} \ ,$$

where N(*uv*) and N(*w*) denote the number of objects in the cluster *uv* and *w* respectively.

- **Centroid Method**

This method is different from the above methods. In the centroid method the similarity between 2 clusters is the distance between the clusters centroid are the mean values of the observations on the variable in the cluster. In this method, every time individual is grouped a new centroid is computed. Cluster center changes when cluster merges take place.

These methods are the most popular in the physical and life science but sometime it may produce confusing results. The confusion occurs because of reversals this is instances when the distance between centroids of one pair may be less than he distances between the centroid of another pair merged at an earlier

combination. The advantage of this method like average linkage method is that it is less affected by outliers than the other two methods.

- **Ward's method**

This method is based on minimize the loss of information from joining two groups. This method is usually implemented with loss of information to be taken as increase in Error Sum of Squares (ESS). ESS is the sum of squared deviations of every item from the cluster centroid. Suppose there are k clusters then total error sum of squares is

$$ESS = ESS_1 + ESS_2 + ... + ESS_k \ ,$$

where ESS_i denote the ESS of the cluster i. i=1, 2,…, k.

At each step in the analysis the union of every possible pair of clusters is considered and two cluster whose combination results in the smallest increase in error sum squares i.e., minimum loss of information are joined. Initially each cluster consist of single item, and there are N clusters and ESS=0. At the other extreme when all objects are combined in a single cluster the value of error sum square is given by

$$ESS = \sum_{j=1}^{n} (X_i - \overline{X})'(\mathrm{X}_i - \overline{X}) \ ,$$

where X_i is the multivariate measurement associated with the i^{th} item and $\overline{X}$ is the mean of all items.

Divisive Hierarchical methods

Divisive hierarchical methods work in the opposite direction of agglomerative method. Initially there is a single group contain the entire item. First we divide it into two groups such that object of the one subgroup is far away from the objects in the other subgroup. These subgroups are again divided into dissimilar subgroups and the process is continued till there are as many subgroups

as objects. The results of both agglomerative and divisive method can be displayed in the form of a 2-D diagram called dendrogram. The diagram illustrates merges or divisions that have been made at successive levels.

2.4 NON HIERARCHICAL CLUSTERING METHODS

Nonhierarchical cluster analysis forms a grouping of set of units into a pre-determined number of groups, using an iterative algorithm that optimize a chosen criterion starting from an initial classification, units are transferred one group to another or swapped with units from other groups, until no further improvement can be made to the criterion value. There is no guarantee that the solution thus obtained will globally optimal by starting from a different initial classification it is sometimes possible to obtain a better classification. However, starting from a good initial classification much increase the chance of producing an optimator near optimal solution.

Most popular Nonhierarchical clustering method is K-means clustering.

K-means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed. At this point we need to re-calculate k new centroids. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no

more changes are done. So here we group the item into k clusters where k may be either specified in advance or determine as a part of clustering procedure.

Algorithm

1. Partition the items into K initial clusters.
2. Assign an item to the cluster whose centroid is nearest.
3. Re calculate the centroid for cluster receiving the new item and who lose the item.
4. Repeat step 2 until no more reassessment take place.

2.5 ANALYSIS

HIERARCHICAL CLUSTERING

In this section we cluster the countries based on their happiness Indices. Linkage method used here is single linkage and complete linkage. In order to measure the dissimilarity between countries squared euclidean distance measure is used.

Single Linkage

The Dendrogram for the clustering process is given in Figure 2.1.

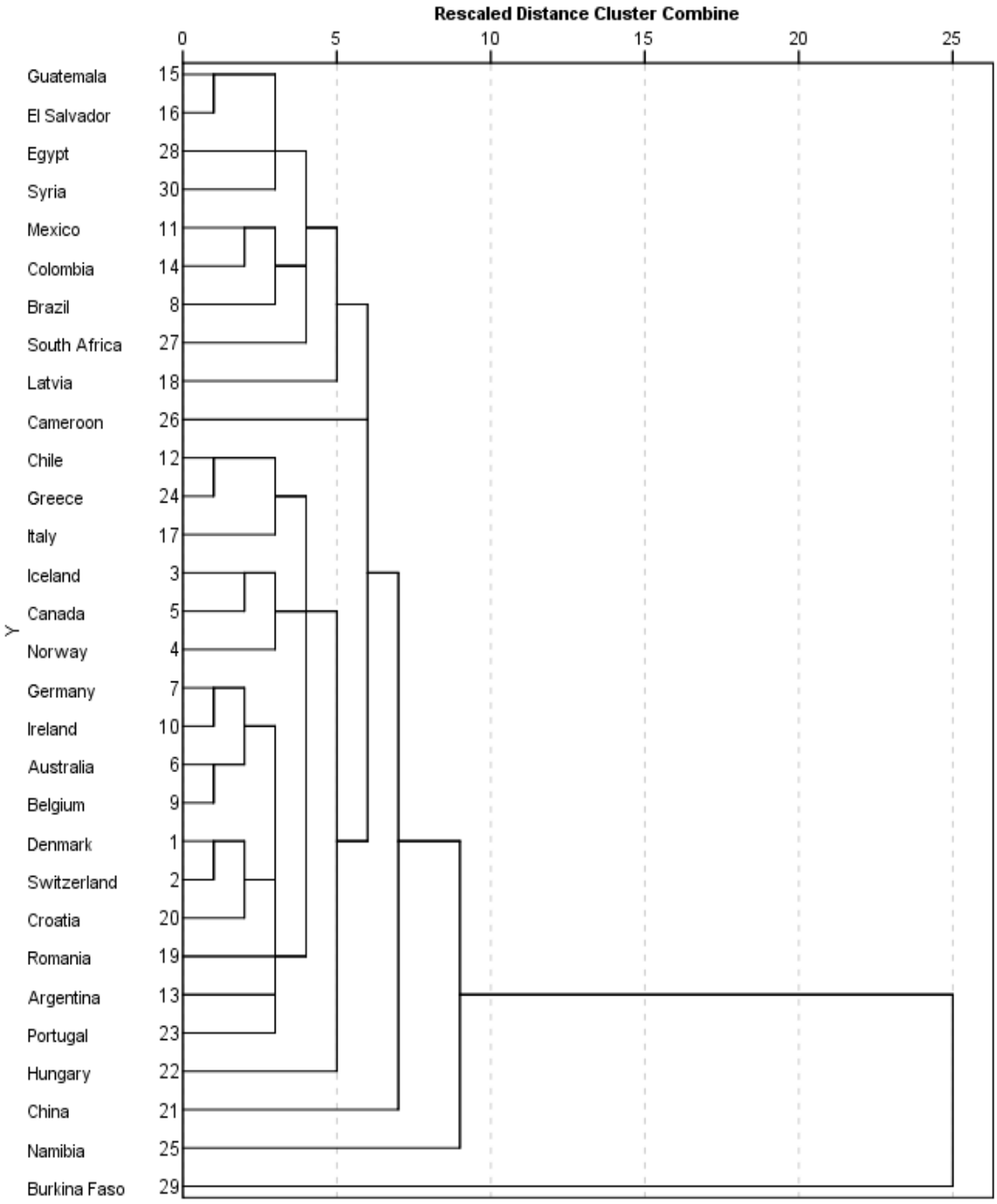

Figure 2.1 Dendrogram using single linkage

Complete linkage

The Dendrogram for the clustering process is given in Figure 2.2.

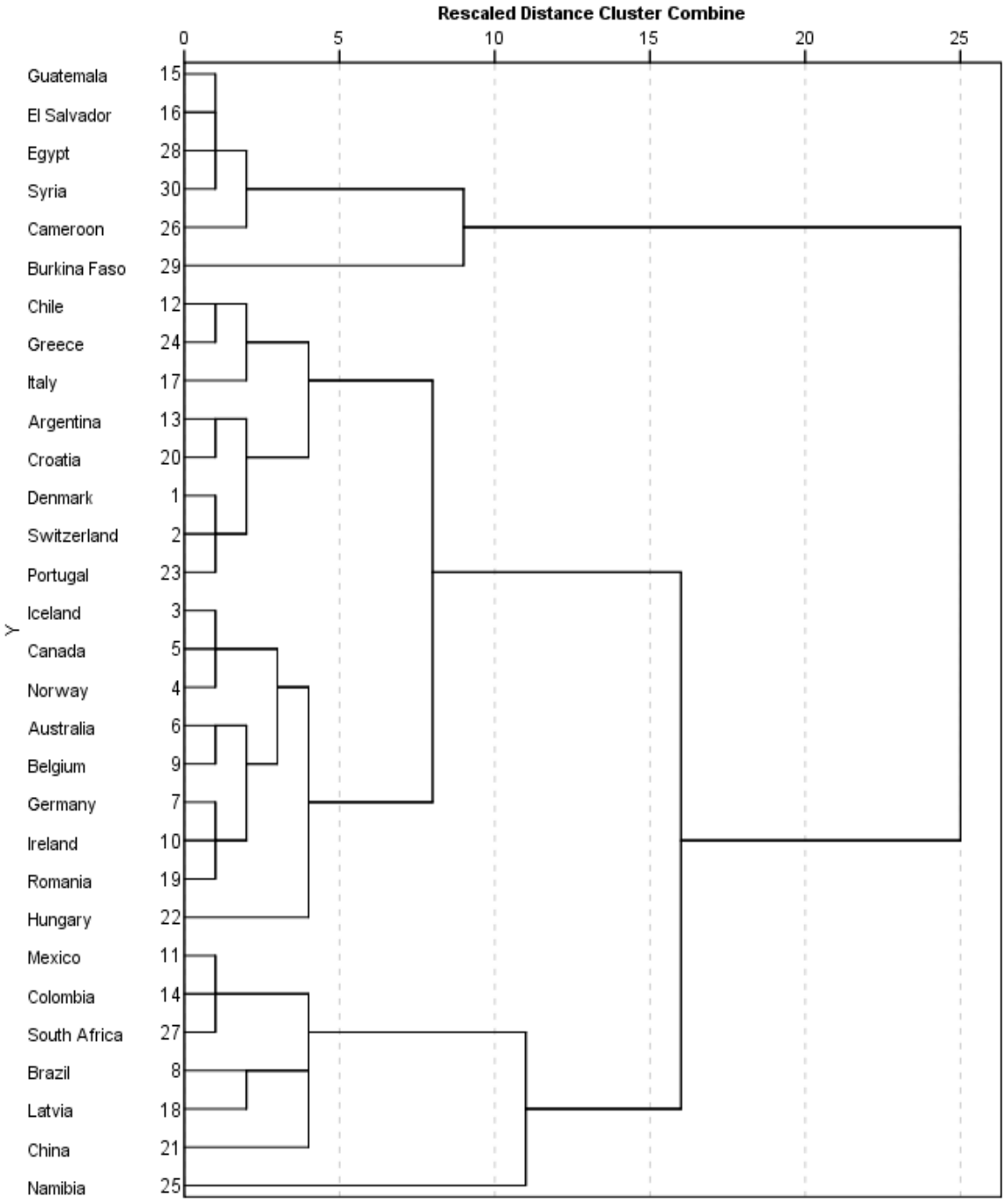

Figure 2.2 Dendrogram using complete linkage

From Figure 2.1 we can see that Guatemala and El Salvador form a cluster which are of similar nature. When we increase the allowable distance, these countries together with Egypt and Syria form a cluster. These clusters are at minimum distance. Moving on we obtain all these countries into a single cluster at a maximum distance 25.

From Figure 2.2 we can see Guatemala, El Salvador, Egypt and Syria form a cluster. Chile, Greece and Italy forms another cluster. These clusters are at minimum distance. At maximum distance 25, all countries form a single cluster.

NON HIERARCHIAL CLUSTERING

K-Means clustering is one of the most important nonhierarchical clustering method. For getting output SPSS (IBM SPSS Version 25) software is used.

K-Means Clustering

Let us take k=3

TABLE 2.1: INITIAL CLUSTER CENTERS

	Cluster		
	1	2	3
Z(HDI)	0.77816	-2.07308	-0.56360
Z(GDP Per Capita)	-0.34823	2.38690	-0.41872
Z(Beer Per Capita)	0.59143	-1.03322	2.34105
Z(Spirit Per Capita)	-0.33926	-1.05470	-1.10240
Z(Wine Per Capita)	3.28121	-0.61565	-0.68607

This Table gives initial cluster centers of standardized variables. Here first we divide the items into 3 clusters. After that if an item in one cluster is more close to the centroid of another cluster then we allocate item into that cluster. Also centroid of the cluster who lose the item and who receive the item are recalculated. This process is continued until no more reallocation take place.

TABLE 2.2: FINAL CLUSTER CENTERS

	Cluster		
	1	2	3
Z(HDI)	1.04339	-0.94265	0.37563
Z(GDP Per Capita)	-0.27659	0.04281	-0.35903
Z(Beer Per Capita)	0.83433	-0.81211	0.94392
Z(Spirit Per Capita)	0.00416	-0.77051	0.13241
Z(Wine Per Capita)	1.84063	-0.63521	-0.23614

TABLE 2.3: NUMBER OF CASES IN EACH CLUSTER

Cluster	1	15.000
	2	6.000
	3	9.000
Valid		30.000
Missing		.000

TABLE 2.4: GROUPING OF COUNTRIES BASED ON VARIABLES

Cluster 1	**Cluster 2**	**Cluster 3**
Denmark	Guatemala	Iceland
Switzerland	El Salvador	Canada
Norway	Cameroon	Brazil
Australia	Egypt	Mexico
Germany	Burkina Faso	Colombia
Belgium	Syria	Latvia
Ireland		China
Chile		Namibia
Argentina		South Africa
Italy		

Romania		
Croatia		
Hungary		
Portugal		
Greece		

Therefore, by k means clustering method we can group 30 countries based on 5 variables into 3 clusters. The first cluster consists of 15 countries. The second cluster consists of 6 countries and third cluster consists of 9 countries.

Countries like Denmark, Switzerland, Norway and Australia belongs to the first cluster. These countries are the happiest countries due to alcohol consumption.

Iceland, Canada, Brazil and China belongs to the third cluster. These Countries are considerably happier due to alcohol consumption.

Underdeveloped countries like Burkina Faso, Cameroon and Guatemala belongs to the second cluster. These Countries have less happiness due to alcohol consumption.

CHAPTER 3

MULTIDIMENSIONAL SCALING

3.1 INTRODUCTION

One major aim of multivariate data analysis is dimension reduction. Multidimensional scaling is a dimension reduction technique in multivariate analysis. MDS is similar to principal component analysis. But in principal component analysis the data is measured in Euclidean coordinates. Also in principal component analysis we use linear transformation technique to reduce dimension.

Multidimensional scaling techniques deal with the following problem: For a set of observed similarities (or distances) between every pair of N items, find a representation of the items in few dimensions such that the interitem proximities "nearly match" the original similarities (or distances).

It may not be possible to match exactly the ordering of the original similarities (distances). Consequently, scaling techniques attempt to find configurations in $q \leq (N-1)$ dimensions such that the match is as close as possible. The numerical measure of closeness is called the stress.

It is possible to arrange the N items in a low-dimensional coordinate system using only the rank orders of the $\frac{N(N-1)}{2}$ original similarities (distances), and not their magnitudes. When only this ordinal information is used to obtain a geometric representation, the process is called nonmetric multidimensional scaling. If the actual magnitudes of the original similarities (distances) are used to obtain a geometric representation in q dimensions, the process is called metric multidimensional scaling. Metric multidimensional scaling is also known as principal coordinate analysis.

That is MDS is a mathematical tool that use proximities between objects, subjects or stimulate to produce geometrical representation for items. Proximities are defined as any set of numbers that express the amount of similarity or dissimilarity between the pair of items. Consider the observations Y_i ; i =1, 2,….,n. Let δ_{ij} represents the distance between Y_i and Y_j in p-dimension. So we wish to represent these n items in a low dimensional coordinate system in which the distance d_{ij} between items closely match the original distance δ_{ij} . That is, $d_{ij} \cong \delta_{ij} \ \forall i, j.$ MDS is classified as metric MDS and nonmetric MDS. If observational vectors Y_i 's are not available, but we have actual distance between items then the process of reduction is known as metric multidimensional scaling. If the original distances are only similarities, then the process is called nonmetric multidimensional scaling.

3.2 METRIC MULTIDIMENSIONAL SCALING

Metric MDS is also known as classical solution or principal coordinates analysis. We begin with a $(n \times n)$ distance matrix $D = (\delta_{ij})_{n \times n}$. Our goal is to find n points in k dimension such that the inter point distance d_{ij} in k dimension are approximately equal to the values of δ_{ij} in D. Typically we use k=2 for plotting purpose. But k=1 or k=3 is also useful.

3.3 NON METRIC MULTIDIMENSIONAL SCALING

If there are n items then there are m = $\frac{n(n-1)}{2}$ similarities. Here δ_{ij} cannot be measured but can be ranked order.

$$\text{ie}, \delta_{i_1 j_1} < \delta_{i_2 j_2} < .... < \delta_{i_m j_m} \rightarrow (A)$$

where (i_1, j_1) indicate the pair of items with the smallest similarity and (i_m, j_m) indicate the pair of item with the greatest similarity. In nonmetric MDS we seek a low dimensional representation of the items such that the ranking distances

$$d_{i_1 j_1} < d_{i_2 j_2} < .... < d_{i_m j_m} \rightarrow (B)$$

Match exactly with the ordering of similarities as in (A). In sometime the ordering of (B) may not match exactly with ordering in (A), i.e.; when we consider a plot of d_{ij} against δ_{ij} may not be monotonic. For this problem Kruskal proposed a measure and is known as stress which is given as

$$\text{Stress} = \left[\frac{\sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2} \right]^{\frac{1}{2}}$$

The stress should be small as much as possible for good MDS.

Stress (%)	**Goodness of fit**
>20	Poor
10	Fair
5	Good
2.5	Excellent
0	Perfect

3.4 APPLICATION OF MDS

Applications include scientific visualization and data mining in fields as cognitive science, information science, psychophysics, marketing and ecology. New applications arise in the scope of autonomous wireless nodes that postulate a space or an area. MDS may apply as a real time enhanced approach to monitoring and managing such populations.

3.5 ANALYSIS

When the variables are plotted on two dimensions using the distance measure “Squared Euclidean Distance” the figure obtained is given in the Figure 3.1.

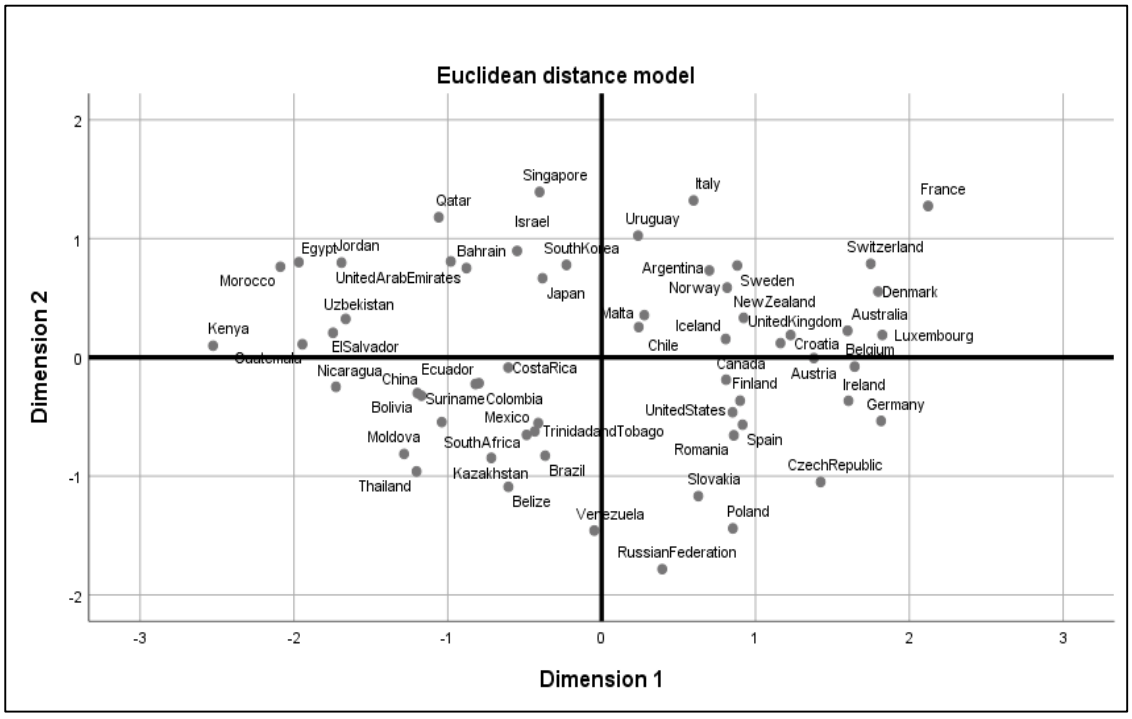

Figure 3.1 Multidimensional Scaling approach

By 2D iteration with ‘Squared Euclidean Distance’ in MDS, we get

Stress = 0.12626 RSQ = 0.92009

From the Figure 3.1 it is clear that Bolivia, Moldova, Suriname, South Africa, Mexico and Costa Rica shows similar nature in the countries happiness. Iceland, Chile, Croatia, United Kingdom, Sweden, Switzerland are also similar. Singapore, Qatar, Italy, France are different from all other countries according to the variable under consideration. So more the distance between countries means the more different these countries are.

CHAPTER 4

DIMENSION REDUCTION METHODS

4.1 INTRODUCTION

Principal component analysis amongst the oldest and most widely used multivariate technique. The basic idea of this technique is to describe the variation in a set of multivariate data in terms of a set of new, uncorrelated variables, each of which is defined to be a particular linear combination of the original variables. In other words principal component analysis is a transformation from the observed variables $x_1,x_2,\ldots,x_p$ to variables $y_{1,}y_2,\ldots,y_{p.}$ In many statistical studies the number of variables under consideration is too large to carry out one way of reducing the number of variables is to consider some of the linear combinations of these variables only, we can discard those linear combinations which have smaller variances and consider only those combinations which have high variances.

Principal component analysis (PCA) is a Mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called Principal components. PCA was invented in 1901, by Karl Pearson. Now it is mostly used as a tool in explanatory data analysis and for making predictive models. PCA can be done by Eigen value decomposition of a data covariance or singular value decomposition of a data matrix, usually after mean centering the data for each attribute.

PCA was invented by 1901 by Karl Pearson; it is mostly commonly used as a tool in explanatory data analysis and for making predicting models. The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores and loadings.

A PCA is concerned with explaining the variance-covariance structure through linear combination of the original variables. In general objectives are:

- Data reduction
- Interpretation

Principal component analysis is available reduction technique that shares a many similarities to exploratory factor analysis. It aims is to reduce a large set of variables into a smaller set of 'artificial' variables called Principal components. The PCA is a method which involves finding the linear combination of a set of variables that has maximum variance and removing its effect, repeating this successively.

Principal component analysis is a linear combination of random or statistical variables which have special properties in terms of variances. The number of principal component is less than or equal to the number of original variables. For example, the first principal component is the normalized linear combination with maximum variance. The second principal component is the normalized linear combination which has second largest variance and uncorrelated with the first principal component. Thus the study of the principal components turns out to be that of characteristic roots and the vectors of the dispersion matrix which either positive definite or positive semi-definite. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables. PCA is a powerful, well-established technique for data reduction and visualization.

Data reduction of a principal component analysis is a reduction technique used extensively in remote sensing studies. PCA is in fact a linear transformation applied on highly correlated multidimensional data. The purpose of PCA is to derive a small number of independent linear combinations (PC) of a set variables that retain as much of the information in the original variables as possible.

Interpretation of a PCA can be interpreted as 3 different ways:

- Maximize the variance of projection along each component.
- Minimize the reconstruction error (i.e., the squared distance between the original data and its estimate).
- Some MLE of a parameter in a probabilistic model.

4.2 PRINCIPLE COMPONENT ANALYSIS

Let $X = \begin{bmatrix} X1 \\ X2 \\ . \\ . \\ . \\ Xp \end{bmatrix}$ is a random vector with E[X] =0 and covariance matrix $\sum$. Let the eigenvalue–eigenvector pairs of $\sum$ be (λ_1, e_1) (λ_2, e_2) ... (λ_p, e_p) such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

Now consider the linear transformations

$$Y_1 = a_1^T x = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_P$$

$$Y_2 = a_2^T x = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_P$$

$$\vdots$$

$$Y_p = a_p^T x = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_P$$

where $a_i^T a_i = 1;\ i = 1,2,..,p$

Thus $V(Y_i) = a_i^T \sum a_i;\ \ i = 1, 2,...,p$

$Cov(Y_i, Y_j) = a_i^T \sum a_j;\ i \neq j$

$i = 1, 2\ldots p;\quad j = 1, 2...p$

The Principle components are those uncorrelated linear combinations Y_1, Y2, ... , Yp whose variance $V(Y_i) = a_i^T \sum a_i$ is a maximum. The 1st principle component is the linear combination $Y_i = a_i^T x$ with maximum $V(Y_i) = a_i^T \sum a_i$.

Thus the first principle component is the linear combination $a_1^T x$that maximize

V ($a_1^T x$) with $a_1^T a_1 = 1$.

The second principle component is the linear combination $a_2^T x$ that maximize V ($a_2^T x$) with $a_2^T a_2 = 1$ and Cov ($a_1^T x$, $a_2^T x$) = 0.

The i^{th} principle component is the linear combination $a_i^T x$ that maximize V ($a_i^T x$) with $a_i^T a_i = 1$ and Cov ($a_k^T x$, $a_i^T x$) =0; k<i.

4.2.2 DETERMINATION OF PRINCIPLE COMPONENTS

Let $X = \begin{bmatrix} X1 \\ X2 \\ \cdot \\ \cdot \\ \cdot \\ Xp \end{bmatrix}$ is a random vector with E[X] =**0** and covariance matrix $\sum$. Let the eigenvalue – eigenvector pairs of $\sum$ be (λ_1, e_1) (λ_2, e_2)... (λ_p, e_p) such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

Then the i^{th} principle component is given by

$$Y_i = e_i^T X,\ i = 1, 2, ..., p \quad \text{and} \quad V(Y_i) = e_i^T \sum ei = \lambda i.$$

4.3 FACTOR ANALYSIS

The main purpose of the Factor Analysis is to describe if possible the covariance relationship between the variables in terms of few underlying unobservable random quantities called factors. For example, a basic desire of obtaining a certain social level might explain most of the consumption behavior. These unobserved factors are more interesting to the social scientist than the observed quantitative measurements.

Factor Analysis is a method for data reduction. It does this by seeking underlying unobservable (latent) variables that are reflected in the observed variables (manifest variables). There are many different methods that can be used to conduct a factor such as principal axis factor, maximum likelihood, generalized

least squares, and weighted least squares. There are also many different types of rotations that can be done after the initial extraction of factors, including orthogonal rotations such as varimax and equimax, which impose the restriction that the factors cannot be correlated. You need to determine the number of factors that you want to extract. As a "rule of thumb", a bare minimum of 10 observations per variable is necessary to avoid computational difficulties. Factor analysis can be considered as an extension of principal component analysis. Both can be viewed as attempt to approximate the covariance matrix $\sum$.

The Orthogonal Factor Model

The observable random vector X with p component has mean μ and covariance matrix $\sum$. The factor model postulate that X is a linearly dependent upon a few unobservable random variables $F_1, F_2, \ldots, F_m$ called common factors and p additional source of variations $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ called error or specific factors. In particular, the factor analysis model in matrix notation is

$$\mathbf{X}_{(px1)} - \boldsymbol{\mu}_{(pX1)} = \mathbf{L}_{(pxm)}\mathbf{F}_{(mx1)} + \boldsymbol{\varepsilon}_{(px1)} \qquad \text{---- (1)}$$

where $\mathbf{L}=(l_{ij})$, the matrix of factor loading of order (pxm). Also l_{ij} is the loading of the i^{th} variable on the j^{th} factor. Note that the i^{th} specific factor ε_i associated only with the ith response X_i.

The p deviations $X_i-\mu_i$, i = 1,2..,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. However some additional assumptions about **F** and **ε,** the model (1) implies certain covariance relationship.

We assume that,

$$E(\mathbf{F}) = \mathbf{0}_{(mx1)} \qquad \text{Cov}(\mathbf{F}) = E(\mathbf{FF'}) = \mathbf{I}_{(mxm)}$$

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}_{(pxm)} \qquad \text{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Psi}_{(pxp)} = \begin{pmatrix} \Psi_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Psi_m \end{pmatrix}$$

Also $\text{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = \text{E}(\boldsymbol{\varepsilon},\mathbf{F}') = \mathbf{0}_{(pxm)}$

So we get $\text{Cov}(\mathbf{X}) = \sum = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'$

$$= \mathbf{L}\, E(\mathbf{F}\mathbf{F}')\mathbf{L}' + E(\boldsymbol{\varepsilon}\mathbf{F}')\mathbf{L}' + \mathbf{L}E(\mathbf{F}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')$$

$$= \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$$

4.4 KMO AND BARTLETT'S TEST OF SPHERICITY

KMO (Kaiser-Meyer-Olkin) and Bartlett's measure test the null hypothesis that the original correlation matrix (R-matrix) is an identity matrix. For factor analysis we need some relationships between variables and if the R-matrix were an identity matrix then all correlation coefficient should be zero. Therefore we want this test to be significant (that is have a significant value less than 0.05). A significant test tell us that the R-matrix is not an identity matrix; therefore there are some relationship between the variables we hope to include in the analysis. The KMO statistic varies between 0 and 1. A value of 0 indicates that the sum of partial correlations is large relative to the sum of correlations, indicating diffusion in the pattern of correlations. A value close to 1 indicates that patterns of correlation are relatively compact and so factor analysis distinct and reliable factors. Kaiser recommends values greater than 0.5 as acceptable. Furthermore values between 0.5 and 0.7 are mediocre, values between 0.7 and 0.8 are good, values between 0.8 and 0.9 are great and values above 0.9 are superb.

SCREE PLOT OF THE VARIABLES

Scree plot is a single line segment plot that shows the fraction of total variance in the data as explained or represented by each principal component. The principal components are ordered, and by definition are therefore assigned a number label, by decreasing order of contribution to total variance. The principal component with largest fraction contribution is labelled with the label name from the preference file, such a plot read left to right. We can often see a clear separation in fraction of total variance where the most important components cease and the least important components begin. In PCA literature, the plot is called a Scree plot because it often looks like a scree slope, where rocks have

fallen down and accumulated on the side of a mountain. A scree plot shows the sorted Eigen values from large to small, as a function of the Eigen value index.

4.5 ANALYSIS

PRINCIPAL COMPONENT ANALYSIS

The results of Principal Component Analysis are given in Table 4.1.

TABLE 4.1: TOTAL VARIANCE EXPLAINED BY THE VARIABLES

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.638	52.761	52.761	2.638	52.761	52.761	1.955	39.106	39.106
2	0.914	18.281	71.042	0.914	18.281	71.042	1.597	31.936	71.042
3	0.742	14.837	85.879						
4	0.468	9.355	95.234						
5	0.238	4.766	100.00						

The table gives the variance of all the principal components obtained. There will be 2 components generated after the principal component analysis. The total of eigenvalues is the sum of values in the column total. It will always equal to the number of variables that were used in the principal component analysis. This is because each variable is assumed to be measured without error, and hence all of the variance is included in the analysis. The average variance will always be one because the total will always sum to the number of variables used, and the average

is found by dividing the total by the number of variables used. Eigen values are the variance of the principle components. Because we conduct our principal component analysis on the correlation matrix, the variables are standardized, which means that each variable has a variance of one, and the total variance is equal to the number of variables used in the analysis in this case it is 1.The minimum eigenvalue criterion states that only components with eigenvalue above one should be retained.

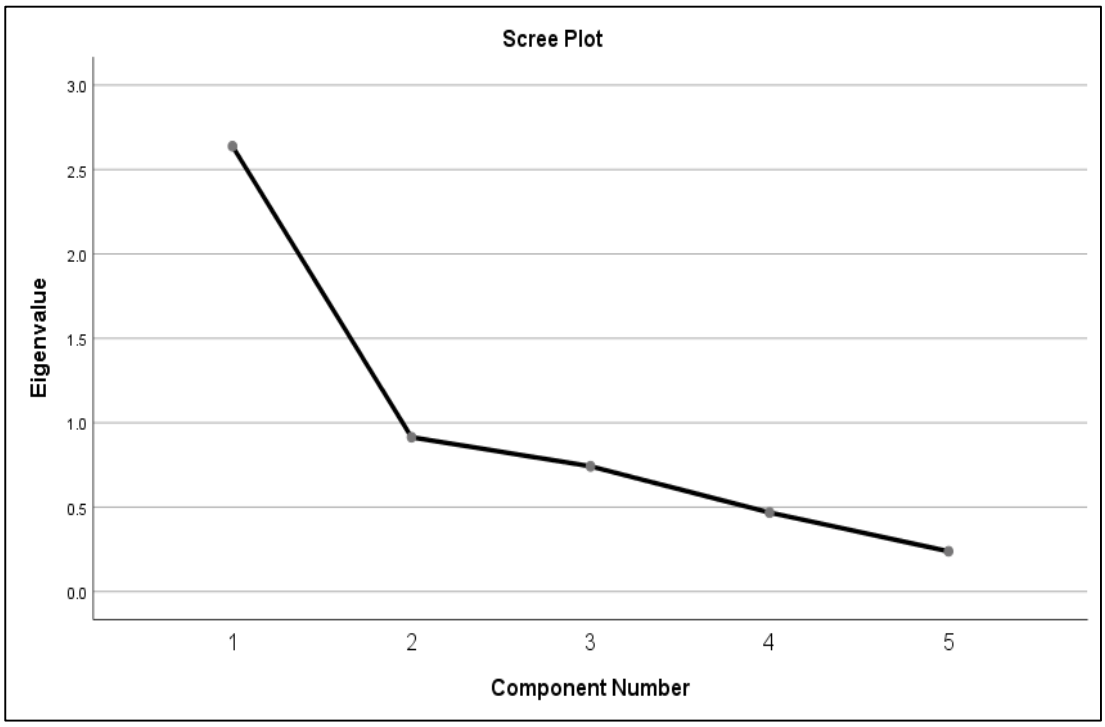

Figure 4.1 Scree Plot

From the Scree plot we can conclude that the first principal components give the most part of the total variability in the data. The remaining principal components account for a very small proportion of the variability. The Scree plot is obtained by plotting the eigenvalues against the corresponding component numbers.

In general, we are interested in keeping only those principal components whose values greater than one. Components with an eigen value less than one account for less variance than did the original variable and so are of little use. Hence, we can

see that the point of principal component is to redistribute the variance in the correlation matrix to redistribute the variance to first components extracted.

Component loadings of different principal components.

TABLE 4.2: COMPONENT MATRIX

	Component	
	1	2
Z(HDI)	0.899	0.079
Z(GDP Per Capita)	-0.664	0.287
Z(Beer Per Capita)	0.792	0.074
Z(Spirit Per Capita)	0.547	-0.659
Z(Wine Per Capita)	0.681	0.621

The first principal component is

0.899X1-0.664X2+0.792X3+0.547X4+0.681X5

The second principal component is

0.079X1+0.287X2+0.074X3-0.659X4+0.621X5

The first principal component accounts for 52.76 % of total variance and second principal component accounts for 18.28 % variance of the total data. We retain the first two principal component since the have variance greater than one. They explain 71.042% of the total variation. The variable which explains maximum information in this data is HDI and GDP Per Capita explains relatively least information.

By using the first principal component scores, we can suggest a new index to the countries. So we can see that Senegal and Mali are the countries which is less happier due to alcohol consumption. Czech Republic and Germany are the happiest countries subjected to alcohol consumption.

FACTOR ANALYSIS

TABLE 4.3: KMO AND BARTLETT'S TEST

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.678
Bartlett's Test of Sphericity	Approx. Chi-Square	191.072
	Df	10
	Sig.	.000

Here these data the value of statistic is 0.678, which falls into the range of mediocre. So we should be confident that factor analysis is acceptable for these data. Also for these data Bartlett's test is highly significant ($p<0.001$) and therefore factor analysis is acceptable.

TABLE 4.4: COMMUNALITIES

	Initial	**Extraction**
Z(HDI)	1.000	0.814
Z(GDP Per Capita)	1.000	0.523
Z(Beer Per Capita)	1.000	0.632
Z(Spirit Per Capita)	1.000	0.734
Z(Wine Per Capita)	1.000	0.848

We know that communalities is the proportion of each variable's variance that can be explained by the factors. The above table shows the communalities before and after extraction. Principal component analysis works on the initial assumption that all variance is common. Therefore before extraction the communalities are all one. The communalities in the column labelled extraction reflect the common variance in the data structure. The values in this column indicate the proportion of each

variable's that can be explained by the retained principal components. Variable with high values are we represented in the common factor space, while variables with low values are not well represented. Here we can say that for example percentage of variance of Wine Per Capita variable that can be explained by the retained principal component is 85% and it is the variable which is well represented by the common factor space.

TABLE 4.5: ROTATED COMPONENT MATRIX

	Component	
	1	2
Z(HDI)	0.748	0.505
Z(GDP Per Capita)	-0.335	-0.641
Z(Beer Per Capita)	0.662	0.441
Z(Spirit Per Capita)	0.010	0.857
Z(Wine Per Capita)	0.919	-0.054

This table shows the component matrix after rotation. This matrix contains the loadings of each variable onto each factor. (Note: If only one component was extracted, the solution cannot be rotated).

In this data HDI, Beer Per Capita and Wine Per Capita forms the first factor. Among these variables, Wine Per Capita is more important. Second factor includes GDP Per Capita and Spirit Per Capita.

CHAPTER 4

CONCLUSION

In this project we got an idea about various multivariate techniques like Cluster Analysis, Multidimensional Scaling, Dimension Reduction Methods like Principal Component Analysis and Factor Analysis.

From descriptive statistics, we obtain the relationship between the dependent variable and an independent variable using scatter plot diagram.

Grouping of data classifies the sets of items into two or more groups in such a way that items within each group are similar than items in different group. So we can find natural groupings among them. By using k-means clustering (k=3), we can divide the items into three clusters, each countries can be grouped based on the happiness due to alcohol consumption.

Multidimensional Scaling attempts to arrange objects in a space with a particular number of dimensions so as to reproduce the observed distance. From Multidimensional Scaling, we can see that the countries are how distinct based on their happiness score in a lower dimensional plane.

From the Principal Component Analysis, we obtain two components explaining 71.042% of the total variation. The first principal component account for 52.76% of the total variability. Also, HDI is the variable which explains maximum information in the data with 89.9 percentage.

Using principal component score it is observed that among 122 countries in United Nations, Senegal is ranked in the first position. Hence Senegal is the country which is less happier due to alcohol consumption and most happiest country is Germany.

In Factor Analysis we found which variables have most influence on the happiness of a country. Here it is found that Wine Per Capita is the most influencing factor.

REFERENCES

1. Anderson T.W. (1984). *An Introduction to Multivariate Statistical Analysis,* 2nd edition, John Wiley, New York, United States.
2. HO R. (2006). *Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS,* Chapman & Hall, New York.
3. Johnson R.A. and Wichern D.W. (1990). *Applied Multivariate Statistical Analysis,* 6th edition, Pearson Prentice Hall, United Kingdom.
4. Sharma S. (1996*). Applied Multivariate Techniques,* Wiley, New York.
5. World Health Organization (2018) - *Global Status report on alcohol and health,* Geneva, Switzerland.
6. World Happiness Report 2016 published by United Nations