A STUDY ON CUSTOMER SATISFACTION TOWARDS ONLINE SHOPPING

PROJECT REPORT

Submitted to

MAHATMA GANDHI UNIVERSITY, KOTTAYAM

In partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE IN STATISTICS (APPLIED)

By

SHALU M PILLAI

(Reg.No:**190011021916**)

Under the guidance of

Mr. TIJO MATHEWS

Assistant Professor

DEPARTMENT OF STATISTICS

KURIAKOSE ELIAS COLLEGE, MANNANAM

KOTTAYAM, KERALA

JULY – 2021

DEPARTMENT OF STATISTICS

KURIAKOSE ELIAS COLLEGE

MANNANAM P.O, KOTTAYAM, KERALA, 686561

PHONE: 0481 2597374

CERTIFICATE

I, hereby certify that the project report entitled **"A STUDY ON CUSTOMER SATISFACTION TOWARDS ONLINE SHOPPING"** is a bonafide work done by **SHALU M PILLAI** under my supervision and guidance in the department of statistics, Kuriakose Elias College, Mannanam during 2019-2021 towards the partial fulfilment of the requirements for the award of the degree of master of science in Statistics (Applied) by Mahatma Gandhi University, Kottayam and that no part of this project has been submitted earlier for the award of any degree.

Place: Mannanam

Date: ……………

TIJO MATHEWS

Assistant Professor

(Supervising Guide)

Forwarded By

Dr. SMITHA S

Head of Dept. of Statistics

Kuriakose Elias College, Mannanam

DECLARATION

I, **SHALU M PILLAI** declare that the project report entitled **"A STUDY ON CONSUMER ONLINE SHOPPING HABITS USING CLUSTER ANALYSIS"** submitted to the Mahatma Gandhi University, Kottayam in partial fulfilment of the requirements for the award of the Degree of Master of Science in Applied Statistics is a record of original research work done by me during April 2021 – July 2021 under the supervision and guidance of **Tijo Mathews**, Assistant Professor, Department of Statistics, Kuriakose Elias college, Mannanam and it has not formed the basis for the award of any Degree, Diploma, Associate ship, Fellowship or other similar title to any candidate in any University.

SHALU M PILLAI

ACKNOWLEDGEMENT

Working on this project "A STUDY ON CUSTOMER SATISFACTION TOWARDS ONLINE SHOPPING" was a source of immense knowledge to me. I would like to express my deep sense of gratitude to my guide Tijo Mathews, Department of Statistics for his guidance as well as ceaseless encouragement in suggesting me the outlines of this project and correcting my doubts throughout the project.

I am thankful to Dr Brigit Paul, Principal, Kuriakose Elias College, Mannanam for providing me the necessary facilities needed for the completion of this work.

Last, but not the least, I would like to thank God Almighty and my parents who helped me from this time to time in making this project unique.

SHALU M PILLAI

Contents

CHAPTER-1

INTRODUCTION

Generally speaking, the trend of e-commerce has been increased rapidly in the recent years. As the world responds to the coronavirus (COVID-19) pandemic, we are seeing a dramatic shift from in-persons to online shopping. Consumers are relying on the digital world more than ever and businesses are forced to adapt their strategies and shift towards digital transformation with much more urgency than before. Globally more than 627 million people have done online shopping so far, World's biggest online shoppers include Germans and British. Books, airline tickets/reservations, clothing/shoes videos/games and other electronic products are the most popular items purchased on the internet.

Through electronic marketing and internet communication business firms are coordinating different marketing activities such as market research, product development, inform customers about product features, promotion, customer services, customer feedback and so on. Online shopping is used as a medium for communication and electronic commerce, it is to increase or improve in value, quality and attractiveness of delivering customer benefits and better satisfaction, that is why online shopping is more convenience and day by day increasing its popularity.

Not only benefits but also risk is associated with online shopping. Moreover, internet users avert online shopping because of credit-card fraud, lack of privacy, non-delivery risk, lack of guarantee of quality of goods and services. Concerned authorities are devising policies to minimize the risk involved in e-business. On the other hand, E- commerce has been grown very fast because of many advantages associated with buying on internet because of lower transaction and search cost as compared to other types of shopping. Through online shopping consumers can buy faster, more alternatives and can order product and services with comparative lowest price. Therefore, Marketers have carefully analysed the consumers' attitude and behaviour towards the online shopping and spend billions of dollars to facilitate all the demographics of online shoppers.

In order to gain competitive edge in the market, marketers need to know the consumer behaviour in the field of online shopping. So, it is important to analyze and identify the factors which influence consumers to shop online in order to capture the demands of consumers.

Other than the factors which influence consumers to shop online, online shopper's demography in terms of Age, gender, income and education is equally important to define their strategies accordingly. As online shopping is a new medium so the consumer behaviour in the field of online shopping is also pretty diverse in nature compare to traditional consumer behaviour, so it is equally important for one to identify what factors influence consumers to shop online. In order to reach towards purchase decision, it consists of several factors which influence consumers to shop online. These factors are important for retailers to compete in the market and to make their product more compatible.

Objectives of the Study

i) To find out the satisfaction level of the customer for online purchase

ii) To classify the cases according to the online shopping habits.

iii) What are the factors influence consumer to shop online?

iv) How many factors are measured by questions?

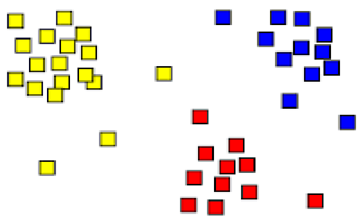
v) Which questions measure similar factors?

vi) Which satisfaction aspects are represented by which factor?

CHAPTER- 2
REVIEW OF LITERATURE

2.1 CLUSTER ANALYSIS

Grouping, or clustering, is distinct from the classification methods. 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 a 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.



Similarity Measures

Most efforts to produce a rather simple group structure from a complex data set require a measure of "closeness," or "similarity." There is often a great deal of subjectivity involved in the choice of a similarity measure. Important considerations include the nature of the variables (discrete, continuous, binary), scales of measurement (Nominal, ordinal, interval, ratio), and subject matter knowledge.

When *items* (units or cases) are clustered, proximity is usually indicated by some sort of distance. By contrast, *variables* are usually grouped on the basis of correlation coefficients or like measures of association.

Distances and Similarity Coefficients for Pairs of Items

Euclidean (straight-line) distance between two p-dimensional observations (items)
$\mathbf{x}' = [x_1, x_2, \ldots\ldots, xp]$ and $\mathbf{y}' = [y_1, y_2 \ldots y_p]$ is,

$$d\,(x, y) = \sqrt{(x_1 + y_1)^2 + (x_2 + y_2)^2 + \cdots . + (x_p + y_p)^2}$$

$$= \sqrt{(x - y)' A\, (x - y)}$$

The statistical distance between the same two observations is

$$d\,(x, y) = \sqrt{(x - y)' A\, (x - y)}$$

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

Another distance measure is the Minkowski metric

$$d(x, y) = \left[\sum_{i=0}^{n} |x_i - y_i|^m\right]^{1/m}$$

For $m = 1$, $d(x, y)$ measures the "city-block" distance between two points in p dimensions. For m = 2, $d(x, y)$ becomes the Euclidean distance.

Two additional popular measures of "distance" or dissimilarity are given by the Canberra metric and the Czekanowski coefficient. Both of these measures are defined for nonnegative variables only. We have

Canberra metric: $$d(x, y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}$$

Czekanowski coefficient: $$d(x, y) = 1 - \frac{2\sum_{i=1}^{p} min(x_i, y_i)}{\sum_{i=1}^{p}(x_i + y_i)}$$

2.12 Hierarchical Clustering Methods

Hierarchical clustering techniques proceed by either a series of successive mergers or a series of successive divisions. It can be divided into two main types: agglomerative and divisive. ***Agglomerative hierarchical methods*** start with the individual objects. Thus, there are initially as many clusters as objects. The most similar objects are first grouped, and these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster.

Divisive hierarchical methods work in the opposite direction. An initial single group of objects is divided into two subgroups such that the objects in one subgroup are "far from" the objects in the other. These subgroups are then further divided into dissimilar subgroups; the process continues until there are as many subgroups as objects-that is, until each object forms a group.

The results of both agglomerative and divisive methods may be displayed in the form of a two-dimensional diagram known as a *dendrogram.* As we shall see, the dendrogram illustrates the mergers or divisions that have been made at successive levels.

In this section we shall concentrate on agglomerative hierarchical procedures and, in particular, *linkage methods.* Linkage methods are suitable for clustering items, as well as variables. This is not true for all hierarchical agglomerative procedures**.** ***Single linkage*** (minimum distance or nearest neighbour), ***complete linkage*** (maximum distance or farthest neighbour), and ***average linkage*** (average distance).

Single linkage results when groups are fused according to the distance between their nearest members. Complete linkage occurs when groups are fused according to the distance between their farthest members.

For average linkage, groups are fused according to the average distance between pairs of members in the respective sets.

The following are the steps in the agglomerative hierarchical clustering algorithm for grouping N objects (items or variables):

1. Start with N clusters, each containing a single entity and an $N \times N$ symmetric matrix of distances (or similarities) $\mathbf{D} = \{d_{ik}\}$.

2. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters U and V be d_{uv}

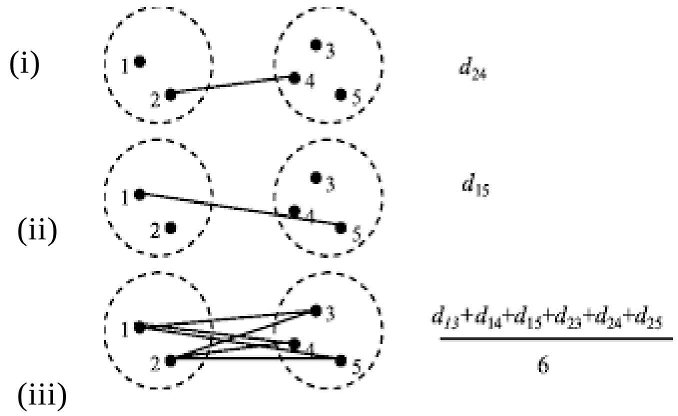


figure 1.1

for (i) single linkage, (ii) complete linkage, and (iii) average linkage.

3. Merge clusters U and V. Label the newly formed cluster (UV*)*. Update the entries in the distance matrix by (a) deleting the rows and columns corresponding to clusters U and V and (b) adding a row and column giving the distances between cluster (UV*)* and the remaining clusters.

4. Repeat Steps 2 and 3 a total. of N - 1 times. (All objects will be in a *single* cluster after the algorithm terminates.) Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take

Single linkage

The inputs to a single linkage algorithm can be distances or similarities between pairs of objects. Groups are formed from the individual entities by merging nearest neighbours, where the term *nearest neighbour* connotes the smallest distance or largest similarity.

Initially, we must find the smallest distance in **D** = { d_{ik} } and merge the corresponding objects, say, *U* and *V*, to get the cluster (UV*)*. the distances between (UV*)* and any other cluster *W* are computed by

$$d(uv)\ w = \min\{duw.dvw\}$$

Here the quantities *duw* and *dvw* are the distances between the nearest neighbors of clusters *U* and Wand clusters *V* and *W*, respectively. The+ results of single linkage clustering can be graphically displayed in the form of a *dendrogram*, or tree diagram. The branches in the tree represent clusters. The branches come together (merge) at nodes whose positions along a distance (or similarity) axis indicate the level at which the fusions occur.

Complete Linkage

Complete linkage clustering proceeds in much the same manner as single linkage clustering, with one important exception: At each stage, the distance (similarity) between clusters is determined by the distance (similarity) between the two elements, one from each cluster, that are *most distant*. Thus, complete linkage ensures that all items in a cluster are within some maximum distance (or minimum similarity) of each other.

The general agglomerative algorithm again starts by finding the minimum entry in **D** = {d_{ik} }and merging the corresponding objects, such as *U* and *V*, to get cluster (UV*)*. The distances between *(UV)* and any other cluster W are computed by

$$d(uv)w = \max\{duw, dvw\}$$

Here *duw* and *dvw* are the distances between the most distant members of clusters *U* and Wand clusters *V* and *W*, respectively.

Average Linkage

Average linkage treats the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster. Again, the input to the average linkage algorithm may be distances or similarities, and the method can be used to group objects or variables. We begin by searching the distance matrix **D** = {d_{ik}} to find the nearest (most similar) objects for example, *U* and *V*. These objects are merged to form the cluster (UV*)*. For Step 3 of the general agglomerative algorithm, the distances between (UV*)* and the other cluster W are determined by

$$d_{(u,v)w} = \frac{\sum i \sum k\ d_{ik}}{N_{(UV)}N_W}$$

where *d; k* is the distance between object *i* in the cluster *(UV)* and object *k* in the cluster *W*, and $N_{(UV)}$ and N_W are the number of items in clusters *(UV)* and *W*, respectively.

Ward's Hierarchical Clustering Method

Ward [32] considered hierarchical clustering procedures based on minimizing the 'loss of information' from joining two groups. This method is usually implemented with loss of information taken to be an increase in an error sum of squares criterion, ESS. First, for a given cluster *k*, let ESS_K be the sum of the squared deviations of every item in the cluster from the cluster mean (centroid). If there are currently K clusters, define ESS as the sum of the ESS_K or ESS = ESS_1 + ESS_2 + ... + ESS_K. At each step in the analysis, the union of every possible pair of clusters is considered, and the two clusters whose combination results in the smallest increase in ESS (minimum loss of information) are joined. Initially, each cluster consists of a single item, and, if there are N items, ESS_K = 0, k = 1, 2, ..., N, so ESS = 0. At the other extreme, when all the clusters are combined in a single group of N items, the value of

ESS is given by

$$\text{ESS} = \sum_{i=0}^{n} (x_i - \bar{x})'(x_j - \bar{x})$$

where x_j is the multivariate measurement associated with the jth item and $\bar{x}$ is the mean of all the items.

The results of Ward's method can be displayed as a dendrogram. The vertical axis gives the values of ESS at which the mergers occur.

Ward's method is based on the notion that the clusters of multivariate observations are expected to be roughly elliptically shaped. It is a hierarchical precursor to non-hierarchical clustering methods that optimize some criterion for dividing data into a *given* number of elliptical groups.

2.13 Non-hierarchical Clustering Methods

Non-hierarchical clustering techniques are designed to group *items*, rather than *variables*, into a collection of K clusters. The number of clusters, K, may either be specified in advance or determined as a part of the clustering procedure. Because a matrix of distances (similarities) does not have to be determined, and basic data do not have to be stored during the computer run, non-hierarchical methods can be applied to much larger data sets than can hierarchical techniques.

Non-hierarchical methods start from either (1) an initial partition of items into groups or (2) an initial set of seed points, which will form the nuclei of clusters. Good choices for starting configurations should be free of overt biases. One way to start is to randomly select seed points from among the items or to randomly partition the items into initial groups.

In this section, we discuss one of the more popular non-hierarchical procedures, the K-means method.

K-means Method

MacQueen [25] suggests the term *K-means* for describing an algorithm of his that assigns each item to the cluster having the nearest centroid (mean). In its simplest version, the process is composed of these three steps:

1. Partition the items into K initial clusters.

2. Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. (Distance is usually computed using Euclidean distance with either standardized or unstandardized observations.) Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

3. Repeat Step 2 until no more reassignments take place.

Rather than starting with a partition of all items into K preliminary groups in Step 1, we could specify K initial centroids (seed points) and then proceed to Step 2.

The final assignment of items to clusters will be, to some extent, dependent upon the initial partition or the initial selection of seed points. Experience suggests that most major changes in assignment occur with the first reallocation step.

CHAPTER – 3

FACTOR ANALYSIS

Factor analysis is a method of data reduction. It does this by seeking underlying unobservable variables that are reflected in the observed variables. The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying. but unobservable, random quantities called *factors*. Basically, the factor model is motivated by the following argument: Suppose variables can be grouped by their correlations. That is, suppose all variables within a particular group are highly correlated among themselves, but have relatively small correlations with variables in a different group. Then it is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations. For example, correlations from the group of test scores in classics, French, English, mathematics, and music collected by Spearman suggested an underlying "intelligence" factor. A second group of variables, representing physical-fitness scores, if available, might correspond to another factor. It is this type of structure that factor analysis seeks to confirm.

Factor analysis can be considered an extension of principal component analysis. Both can be viewed as attempts to approximate the covariance matrix $\sum$. However, the approximation based on the factor analysis model is more elaborate. The primary question in factor analysis is whether the data are consistent with a prescribed structure.

3.1 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 $F_1, F_2 \ldots, F_m$, called *common factors,* and p additional sources of variation $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$, called *errors* or, sometimes, *specific factors.* In particular, the factor analysis model is

$$\begin{aligned}
X_1 - \mu_1 &= L_{11}F_1 + L_{12}F_2 + \ldots\ldots + L_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= L_{21}F_1 + L_{22}F_2 + \ldots\ldots + L_{2m}F_m + \varepsilon_2 \\
&\vdots \\
&\vdots \\
X_p - \mu_p &= L_{p1}F_1 + L_{p2}F_2 + \ldots\ldots + L_{pm}F_m + \varepsilon_p
\end{aligned}$$

or, in matrix notation,

$$\underset{(p\times 1)}{X - \mu} = \underset{(p\times m)}{L}\ \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\varepsilon}$$

The coefficient L_{ij} is called the *loading* of the ith variable on the jth factor, so the matrix **L** is the *matrix of factor loadings.* Note that the ith specific factor ε_i; is associated only with the ith response

X_i. The p deviations $X_1 - \mu_1, X_2 - \mu_2, \ldots\ldots, X_p - \mu_p$ 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*. With so many unobservable quantities, a direct verification of the factor model from observations on $X_1, X_2, \ldots, X_P$ is hopeless. However, with some additional assumptions about the random vectors **F** and ε, the model implies certain covariance relationships, which can be checked.

We assume that

$$\underset{(m\times 1)}{E(\boldsymbol{F})} = \mathbf{0}, \qquad \text{Cov}\,(\mathbf{F}) = E[\boldsymbol{FF}'] = \underset{(m\times m)}{\mathrm{I}}$$

$$E(\varepsilon) = 0, \qquad \text{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}\,[\varepsilon\,\varepsilon'] = \boldsymbol{\varphi} = \begin{bmatrix} \varphi_1 & 0 & \cdots & 0 \\ 0 & \varphi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varphi_p \end{bmatrix}$$

And that F and ε are independent, so

$$\text{Cov}\,(\boldsymbol{\varepsilon}, \mathbf{F}) = \underset{(p\times m)}{E\,(\boldsymbol{\varepsilon}, \boldsymbol{F}')} = 0$$

These assumptions and the relation in constitute the *orthogonal factor model*.

Orthogonal Factor Model with m Common Factors

$$\underset{(p\times 1)}{\mathrm{X}} = \underset{(p\times 1)}{\mu} + \underset{(p\times m)}{\mathrm{L}}\;\underset{(m\times 1)}{\mathrm{F}} + \underset{(p\times 1)}{\varepsilon}$$

μ_i = *mean* of variable i

ε_i = i^{th} *specific factor*

F_j = j^{th} *common factor*

L_{ij} = *loading* of the ith variable on the j^{th} factor

The unobservable random vectors F and $\boldsymbol{\varepsilon}$ satisfy the following conditions:

F and ε are independent

E(**F**) = 0, Cov(**F**) =**I**

$E\,(\varepsilon) = \mathbf{0}$, Cov (ε) = 'Ψ, where Ψ is a diagonal matrix

The orthogonal factor model implies a covariance structure for X. From the model,

$$(\mathrm{X}- \boldsymbol{\mu})\,(\mathrm{X} - \boldsymbol{\mu})' = (\mathbf{LF} + \varepsilon)\,(\mathbf{LF} + \varepsilon)'$$
$$= (\mathbf{LF} + \varepsilon)\,((\mathbf{LF})' + \varepsilon')$$
$$= \mathbf{LF}(\mathbf{LF})' + \varepsilon\,(\mathbf{LF})' + \mathbf{LF}\varepsilon' + \varepsilon\varepsilon'$$

So that,

$$\Sigma = \text{Cov}(\mathbf{X}) = E\,(\mathbf{X}- \boldsymbol{\mu})\,(\mathbf{X}- \boldsymbol{\mu})'$$
$$= \mathbf{L}\mathrm{E}(\mathbf{FF}')\,\mathbf{L}' + \mathrm{E}(\boldsymbol{\varepsilon}\mathbf{F}')\,\mathbf{L}' + \mathbf{L}\mathrm{E}(\mathrm{F}\boldsymbol{\varepsilon}') + E(\varepsilon\varepsilon')$$
$$= \mathbf{LL}' + \Psi$$

Also, by independence, Cov $(\boldsymbol{\varepsilon}, \mathbf{F}) = E\,(\varepsilon, \mathbf{F}') = \mathbf{0}$

Also, by the model $(\mathbf{X} - \boldsymbol{\mu})\,\mathbf{F}' = (\mathbf{LF} + \varepsilon)\,\mathbf{F}' = \mathbf{LFF}' + \varepsilon\mathbf{F}'$.

$\text{Cov}(\mathbf{X}, \mathbf{F}) == E(\boldsymbol{X} - \boldsymbol{\mu})\,\mathbf{F}' = \mathbf{L}E(\mathbf{FF}') + E(\varepsilon \boldsymbol{F}') = \mathbf{L}$.

Covariance Structure for the Orthogonal Factor Model

1. $\text{Cov}(X) = \mathbf{LL}' + \boldsymbol{\Psi}$

Or

$$\text{Var}(X_i) = \ell_{i1}^2 + \cdots \ldots \ldots + \ell_{im}^2 + \Psi_i$$

$$\text{Cov}(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots \ldots \ldots + \ell_{im}\ell_{km}$$

2. $\text{Cov}(X, F) = \mathbf{L}$

or $\text{Cov}(X_i, F_j) = \ell_{ij}$

The model $X - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$ is *linear* in the common factors. If the p responses $\mathbf{X}$ are, in fact, related to underlying factors, but the relationship is nonlinear, such as, $\boldsymbol{X}_1 - \mu_1 = \ell_{11}F_1F_3 + \boldsymbol{\varepsilon}_1$, $\boldsymbol{X}_2 - \mu_2 = \ell_{21}F_2F_3 + \boldsymbol{\varepsilon}_2$, and so forth, then the covariance and structure $\mathbf{LL}' + \Psi$ given may not be adequate. The very important assumption of linearity is inherent in the formulation of the traditional factor model.

That portion of the variance of the ith variable contributed by them m common factors is called the i^{th} *communality*. That portion of $\text{Var}(X_i) = \sigma_{ii}$ due to the specific factor is often called the *uniqueness*, or *specific variance*. Denoting the i^{th} communality by h_i^2,

$$\underbrace{\sigma_{ii}}_{\text{Var}(X_i)} = \underbrace{\ell_{i1}^2 + \cdots \ldots \ldots + \ell_{im}^2}_{\textit{communality}} + \underbrace{\Psi_i}_{\textit{specific variance}}$$

Or

$$h_i^2 = \ell_{i1}^2 + \cdots \ldots \ldots + \ell_{im}^2$$

and

$$\sigma_{i1} = h_i^2 + \Psi_1 , \quad i = 1, 2, \ldots, p$$

The ith communality is the sum of squares of the loadings of the ith variable on the m common factors.

3.2 Methods of Estimation

Given observation $x_1, x_2, \ldots. x_n$ as ‘p’ generally correlated variables, factor analysis seeks to answer this question.

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

Factor model with small no. of factors adequately represented for we try to fixed the covariance structure.

$$\sum = \mathbf{LL}' + \Psi$$

3.21 The Principal Component Method

From the linear algebra, we have the spectral decomposition for $\sum$, we have the eigen value – eigen vector pair (λ_i e_i) with $\lambda_1 \geq \lambda_2 \ldots\ldots \lambda_p \geq 0$. Then $\sum = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \ldots\ldots + \lambda_p e_p e_p'$

$$= \begin{bmatrix} \sqrt{\lambda_1}\, e_1 & \sqrt{\lambda_2}\, e_2 & \cdots & \sqrt{\lambda_p}\, e_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\, e_1' \\ \sqrt{\lambda_2}\, e_2' \\ \vdots \\ \vdots \\ \sqrt{\lambda_p}\, e_p' \end{bmatrix}$$

This fits the prescribed covariance structure for the factor analysis model having many factors as variables (m = p) and specific variance $\Psi_i = 0$ for all i. The loading matrix has j^{th} column given $\sqrt{\lambda_j}\, e_j$. then we can write,

$$\sum = \mathbf{LL}' + 0 = \mathbf{LL}'$$

Apart from the scale factor $\sqrt{\lambda_j}$ the factor loading on the j^{th} factor are the coefficients for the j^{th} principal component of the population. We prefer models that explain the covariance structure in the terms of just a few common factor, one such approach is, when the last p – m eigen values are small is to be nested the combination of $\lambda_{m+1} e_{m+1} e_{m+1}' + \lambda_{m+2} e_{m+2} e_{m+2}' + \ldots\ldots + \lambda_p e_p e_p'$ to $\sum$ thus we obtain the approximation.

$$\sum = \begin{bmatrix} \sqrt{\lambda_1}\, e_1 & \sqrt{\lambda_2}\, e_2 & \cdots & \sqrt{\lambda_p}\, e_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\, e_1' \\ \sqrt{\lambda_2}\, e_2' \\ \vdots \\ \vdots \\ \sqrt{\lambda_p}\, e_p' \end{bmatrix}$$

$$= \mathbf{LL}'$$

In this representation assumes that specific factor $\sum$, in are of minor importance and can also be ignored in the factorising of $\sum$, if specific factors are individual in the model, their variances may be taken to be the diagonal elements of $\sum = \mathbf{LL}'$, then $\sum = \mathbf{LL}' + \Psi$

$$\begin{bmatrix} \sqrt{\lambda_1}\, e_1 & \sqrt{\lambda_2}\, e_2 & \cdots & \sqrt{\lambda_p}\, e_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\, e_1' \\ \sqrt{\lambda_2}\, e_2' \\ \vdots \\ \vdots \\ \sqrt{\lambda_p}\, e_p' \end{bmatrix} + \begin{bmatrix} \varphi_1 & 0 & \cdots & 0 \\ 0 & \varphi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varphi_p \end{bmatrix}$$

$\varphi_i = \sigma_{ii} - \sum_{j=1}^{m} \ell_{ij}^2$ i = 1,2…..,p

Principal Component Solutions of The Factor Model

The principal components factor analysis of the sample covariance matrix S,is specified in terms of its eigenvalues – eigenvector pairs ($\hat{\lambda}_1$,$\hat{e}_1$) ($\hat{\lambda}_2$,$\hat{e}_2$)….. ($\hat{\lambda}_p$,$\hat{e}_p$).

$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots\ldots \hat{\lambda}_{p.\text{ Let}}$ m < p be the no. of common factors then the matrix of estimated factors loadings

$\{\hat{e}_{ij}\}$ given by

$$\tilde{L} = \sqrt{\hat{\lambda}_1}\,\hat{e}_1\ \sqrt{\hat{\lambda}_2}\,\hat{e}_2 \ldots\ldots \sqrt{\hat{\lambda}_m}\,\hat{e}_m$$

The estimated specific variance is provided the diagonal elements of the matrix S - $\tilde{L}\tilde{L}'$ so

$$\tilde{\varphi} = \begin{bmatrix} \widetilde{\varphi 1} & 0 & \cdots & 0 \\ 0 & \tilde{\varphi}2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varphi_p \end{bmatrix}, \text{ with } \tilde{\varphi}_I = S_i - \sum_{j=1}^{n} \tilde{\ell}_{ij}^{\,2}$$

3.3 DATA DESCRIPTIVE

When it comes to data collection there are two methods in general used by researchers to collect data, primary and secondary method. Primary data, it includes observation method, Interview/ questionnaire method, case study method. Whereas, secondary data is one which is already collected by some other researcher not for the reason for particular study or research.

We would like to go for primary data collection method that will include questionnaire from consumers as what are the factors that influence consumers to purchase online. A sample of 320 customers were taken.

Questionnaire design

The questionnaire is carefully designed to meet the requirements of the research. The questionnaire consists of two main parts, first part includes personal and sensitive question. This section includes questions pertaining to Gender, Age, Income and Education. Second part of the questionnaire will cover the questions relating to factors influencing consumers to shop online. All questions in this section are constructed with 5-point Likert scale ranging from 1 (strongly disagree), 3 (Neutral) to 5 (strongly agree).

First of all, the data set can be imported into SPSS. We use **SPSS 26 version** for analysing the result that we are obtained.

CHAPTER- 3

DATA ANALYSIS AND INTERPRETATION

STEPS TO CONDUCT A CLUSTER ANALYSIS

1. Hierarchical clustering method to define the number of clusters.

- Squared Euclidean Distance used.
- Using Ward's clustering method**.**
- Agglomerative clustering procedure

Case Processing Summary[a,b]

	Cases		
	Valid	Missing	Total
N	320	0	320
Percent	100.0	.0	100.0

a. Squared Euclidean Distance used

b. Ward Linkage

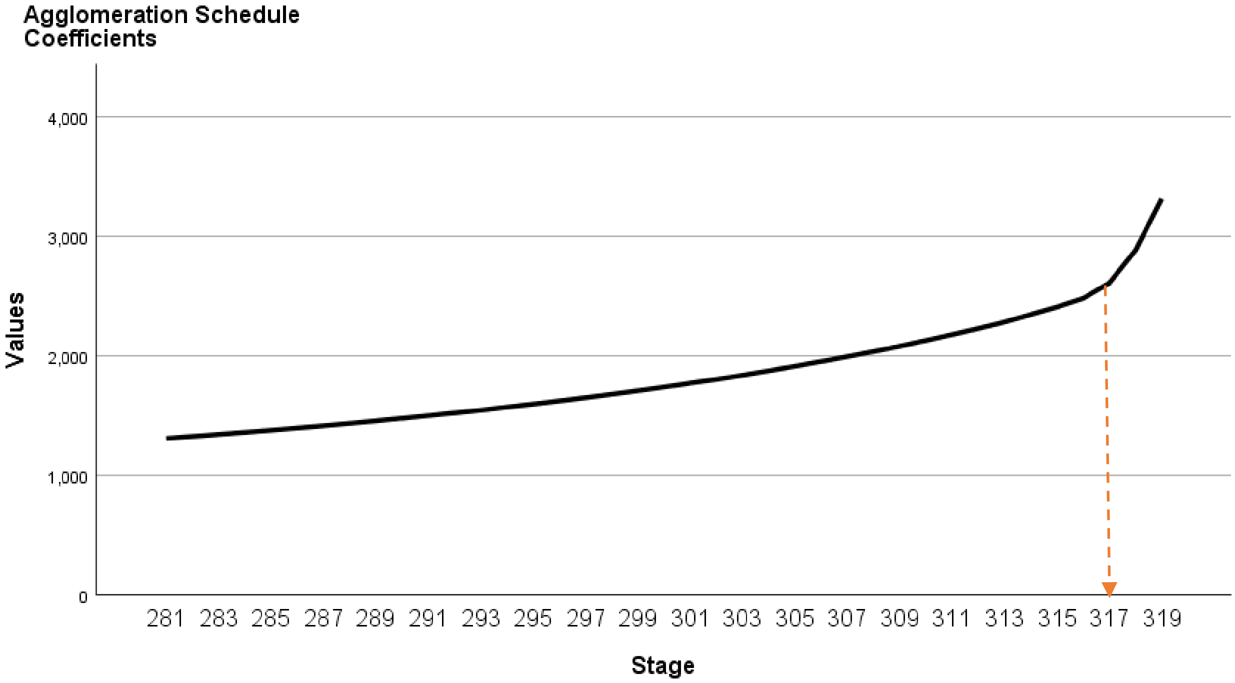
This table tells us that 100% of the 320 cases have been processed.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	261	278	.000	0	0	14
2	133	163	.000	0	0	112
3	98	105	.000	0	0	15
4	33	34	.000	0	0	245
5	9	29	.000	0	0	258
6	234	247	.500	0	0	43
7	144	246	1.000	0	0	40
8	110	229	1.500	0	0	213
9	117	223	2.000	0	0	81
10	187	216	2.500	0	0	16
11	129	200	3.000	0	0	82
12	147	198	3.500	0	0	45
13	170	191	4.000	0	0	44
14	261	299	4.667	1	0	57
15	60	98	5.333	0	3	17
16	184	187	6.167	0	10	34
17	37	60	7.000	0	15	80
18	270	314	8.000	0	0	92
298	10	11	1680.006	278	280	310
299	40	112	1709.221	291	247	312
300	228	257	1740.467	264	277	308
301	114	207	1772.154	290	273	314
302	6	52	1803.917	284	275	306
303	2	4	1838.507	268	260	305
304	118	123	1875.802	272	276	309
305	2	15	1914.129	303	295	316

306	1	6	1954.044	297	302	310
307	3	19	1994.215	296	288	316
308	188	228	2036.671	292	300	314
309	113	118	2081.426	286	304	312
310	1	10	2129.652	306	298	313
311	5	30	2178.852	279	293	313
312	40	113	2229.883	299	309	317
313	1	5	2283.507	310	311	315
314	114	188	2346.040	301	308	318
315	1	44	2409.886	313	287	319
316	2	3	2481.243	305	307	317
317	2	40	2609.192	316	312	318
318	2	114	2883.203	317	314	319
319	1	2	3313.900	315	318	0

When we plotted Agglomeration Schedule coefficient with stages, we get **Scree diagram**



Here we can find a drastic change will happen in the stage 317(elbow)

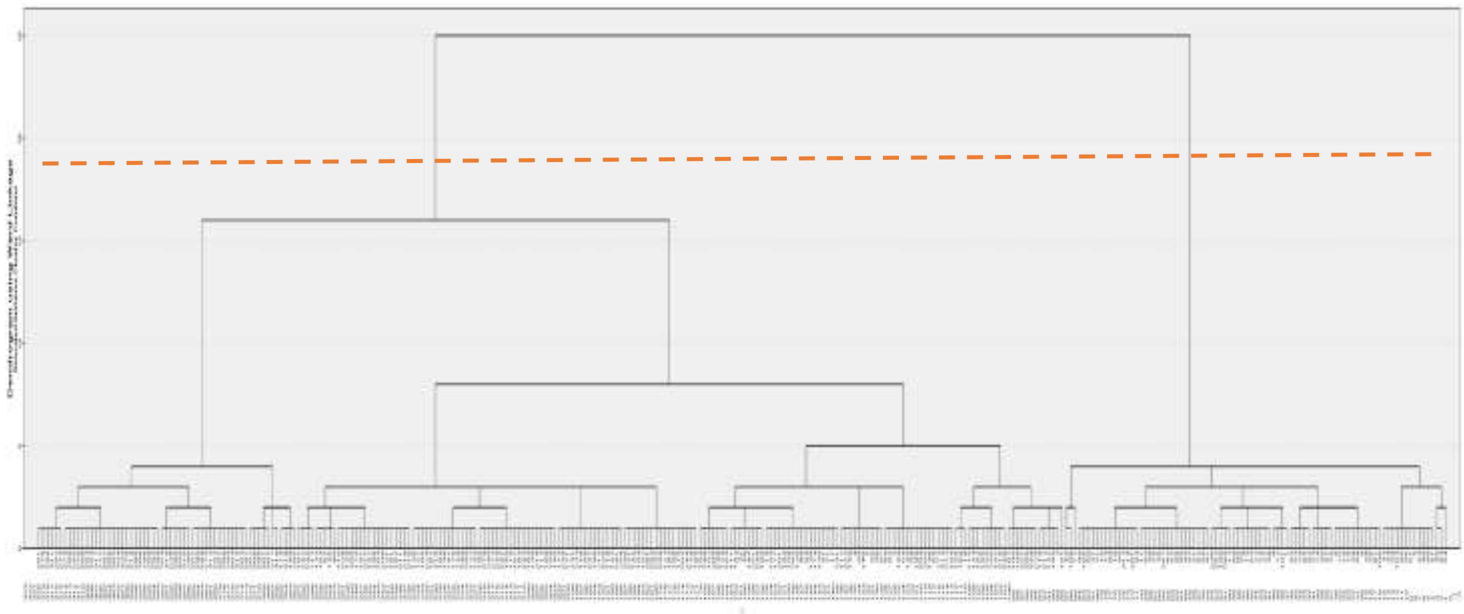
Number of cases = 320

Step of 'elbow' = 317

Number of clusters = 320 – 317

= 3

Dendrogram using Ward Linkage



2) Apply k-mean method

Iteration History[a]			
	Change in Cluster Centers		
Iteration	1	2	3
1	13.828	66.546	53.494
2	16.509	3.001	13.503
3	7.501	2.501	10.004
4	4.501	3.001	7.503
5	3.001	3.001	6.002

6	2.002	2.001	4.002
7	1.501	1.501	3.002
8	1.501	1.001	2.501
9	.501	1.001	1.501
10	1.001	.501	1.501

a. Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum absolute coordinate change for any center is 1.500. The current iteration is 10. The minimum distance between initial centers is 53.498.

FINAL CLUSTER CENTERS

	CLUSTERS		
	1	2	3
I think shopping on internet saves time.	3.5	3.9	4.0
It is great advantage to be able to shop at any time of the day on the internet.	3.9	4.3	4.3
I prefer traditional or conventional shopping than online shopping.	3.5	3.1	4.4
Shopping online is risky.	3.2	2.7	2.7
A long time is required for the delivery of products and services.	3.1	2.5	2.8
I satisfied with the choice of products available in online shopping	3.1	3.9	3.8
The description of products shown on the websites are very informative	3.3	3.9	3.9
Online shopping is as secure as traditional shopping.	2.4	3.1	3.3
While shopping online, I hesitate to give my credit card number.	3.7	2.5	3.3
I prefer cash on delivery than payment via credit / debit card	4.0	3.0	4.1
I am satisfied with the price among these online shopping	3.2	3.9	3.8
I never felt any problem while conducting online purchase.	2.4	3.4	3.7
Website layout helps in searching the products easily.	3.6	4.1	4.0
You are overall satisfied with your experience of shopping online.	3.1	4.1	4.1
Gender			
Male	36.5	31.8	34.0
Female	63.5	68.2	66.0
Age Group			
18 - 30	76.9	70.0	80.2
31 - 39	10.6	13.6	10.4
40 - 49	8.7	13.6	6.6
50 or over	3.8	2.7	2.8
Designation			
Student	1.0	20.0	25.5
Government Employee	14.4	18.2	14.2
Private Employee	60.6	42.7	37.7
Other	24.0	19.1	22.6

Monthly income			
Below Rs. 5000	17.3	20.0	19.8
Rs. 5000 to Rs. 10,000	16.3	9.1	14.2
15,000 to Rs. 20,000	24.0	13.6	20.8
Above Rs.20,000	12.5	29.1	23.6
Have you ever shop online			
Yes	95.2	97.3	99.1
No	4.8	2.7	0.9
which of the following websites do you more prefer			
Flipkart	51.0	45.5	52.8
Amazon	44.2	48.2	41.5
Myntra	3.8	6.4	4.7
Snapdeal	1.0		0.9

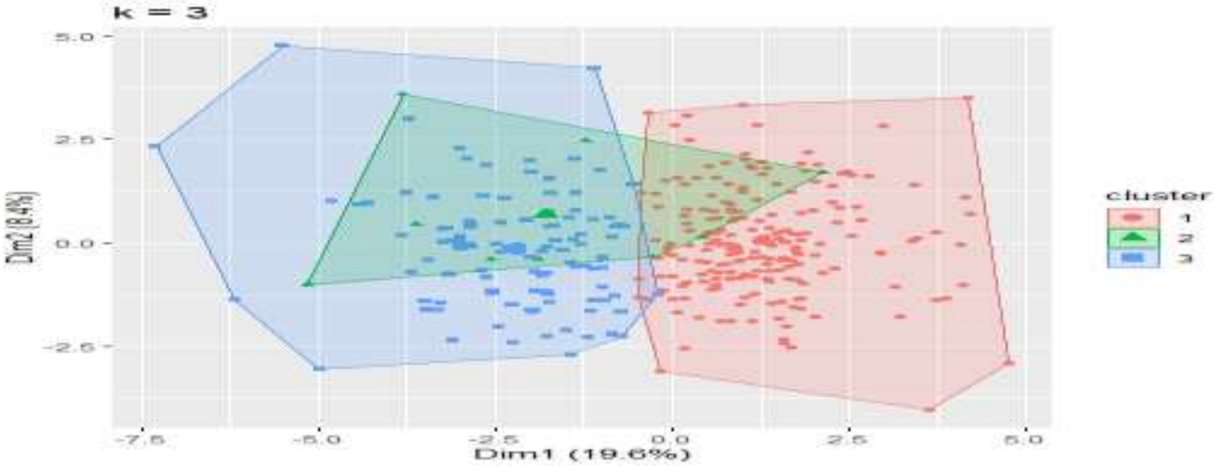
SD	D	N	A	SA
1	2	3	4	5
	<2.5		>3.5	

The scale that we will use for our data analysis is a 5 Point Likert Scale (1=Strongly Disagree, 2**&**<2.5=Disagree, 3=Neutral, 4& > 3.5=agree, 5=Strongly Agree). By using cluster analysis, the number of cases has been classified in to 3 clusters.

- **Cluster 1** shows that the respondents **prefer cash** on delivery than payment via credit / debit.
- On the other hand, **Cluster 2** respondents **enjoy online shopping**. They agreed that they can shop at any time of the day on internet.
- In **Cluster 3**, respondents **prefer traditional shopping** than online shopping.

- In our data, most of the respondents are females. They are spending more time on online shopping.
- Most of the respondents are in the age group 18 – 30.
- 60% of the respondents are private employees in cluster 1.
- There is an association between monthly income and cluster no of cases. Most of the respondents of cluster 1 have annual income about 150,00 – 20,000 and most of the respondents of cluster 2 and 3have income above Rs.20,000.
- There are plenty of websites are available for online shopping. Most of the respondents in cluster 1 and cluster 3 prefer Flipkart and in cluster 2, they prefer Amazon.
- Majority of respondents in three clusters are agreed with these statements**:**
    - I think shopping on internet saves time.
    - It is great advantage to be able to shop at any time of the day on the internet.
    - Website layout helps in searching the products easily

We can also view our results by using R programming, this provides a nice illustration of the clusters.



ANOVA						
	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
I think shopping on internet saves time.	5.654	2	.724	317	7.812	.000
It is great advantage to be able to shop at any time of the day on the internet.	7.712	2	.543	317	14.208	.000
I prefer traditional or conventional shopping than online shopping.	3.789	2	.754	317	5.027	.007
Shopping online is risky.	10.121	2	.576	317	17.574	.000
A long time is required for the delivery of products and services.	9.767	2	.766	317	12.752	.000
I satisfied with the choice of products available in online shopping	19.166	2	.510	317	37.593	.000
The description of products shown on the websites are very informative	11.466	2	.463	317	24.792	.000
Online shopping is as secure as traditional shopping.	25.640	2	.591	317	43.386	.000
While shopping online, I hesitate to give my credit card number.	44.235	2	.809	317	54.687	.000
I prefer cash on delivery than payment via credit / debit card	44.316	2	.924	317	47.970	.000
I am satisfied with the price among these online shopping	16.471	2	.469	317	35.135	.000
I never felt any problem while conducting online purchase.	47.970	2	.747	317	64.200	.000
Website layout helps in searching the products easily.	9.369	2	.370	317	25.327	.000
You are overall satisfied with your experience of shopping online.	33.143	2	.388	317	85.479	.000

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.

Factor Analysis Output I –

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.804
Bartlett's Test of Sphericity	Approx. Chi-Square	813.283
	df	91
	Sig.	.000

Kaiser-Meyer-Olkin Measure of Sampling Adequacy – This measure varies between 0 and 1, and values closer to 1 are better. High values (close to 1) indicates that a factor analysis may be useful with our data. If the value is less than 0.50, the results of the analysis probably won't be useful.

Bartlett's Test of Sphericity – Tests the hypothesis that correlation matrix is an identity matrix, which would indicate that our variables are unrelated and therefore unsuitable for structure detection. Small values (less than 0.05) of the significance level indicate that a factor analysis may be useful with our data

Output II –

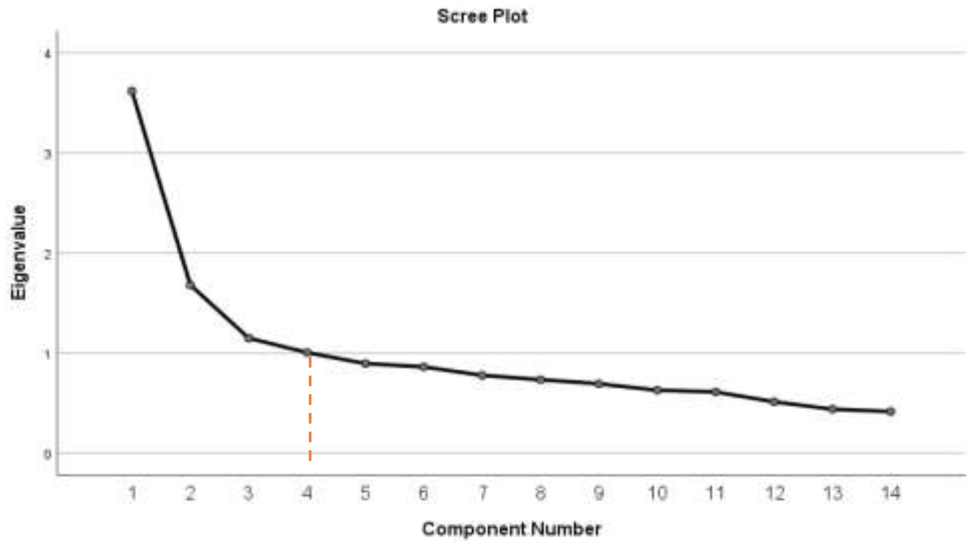
Total Variance Explained

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	3.617	25.832	25.832	3.617	25.832	25.832	2.480	17.716	17.716
2	1.680	11.997	37.829	1.680	11.997	37.829	2.137	15.266	32.982
3	1.148	8.198	46.028	1.148	8.198	46.028	1.518	10.840	43.823
4	1.004	7.171	53.198	1.004	7.171	53.198	1.313	9.376	53.198
5	.893	6.381	59.580						
6	.860	6.146	65.726						
7	.776	5.540	71.266						
8	.732	5.231	76.497						
9	.691	4.938	81.435						
10	.627	4.478	85.914						
11	.610	4.357	90.271						
12	.511	3.653	93.924						
13	.437	3.123	97.047						
14	.413	2.953	100.000						

Extraction Method: Principal Component Analysis.

Here, there are 14 input variables, PCA initially extracts 14 factors (or "components"). Each components have a **quality score** called **Eigenvalues**. Only components with high eigen values are likely to represent a real underlying factor. Here only first 4 components have an eigenvalue of at least 1. The other components - having low quality scores- are not assumed to represent real traits underlying our 14 questions.

Output II - Scree Plot



A scree plot visualizes the Eigenvalues (quality scores). Again, we see that the first 4 components have Eigenvalues over 1. We consider these "strong factors". After that component 5 and onwards- the Eigenvalues **drop off dramatically**. There are 4 strong factors.

Output III – Communalities

Communalities

	Initial	Extraction
I think shopping on internet saves time.	1.000	.322
It is great advantage to be able to shop at any time of the day on the internet.	1.000	.634
I prefer traditional or conventional shopping than online shopping.	1.000	.709
Shopping online is risky.	1.000	.546
A long time is required for the delivery of products and services.	1.000	.395
I satisfied with the choice of products available in online shopping	1.000	.388
The description of products shown on the websites are very informative	1.000	.399
Online shopping is as secure as traditional shopping.	1.000	.598
While shopping online, I hesitate to give my credit card number.	1.000	.725
I prefer cash on delivery than payment via credit / debit card	1.000	.716
I am satisfied with the price among these online shopping	1.000	.431
I never felt any problem while conducting online purchase.	1.000	.460
Website layout helps in searching the products easily.	1.000	.540
You are overall satisfied with your experience of shopping online.	1.000	.584

Extraction Method: Principal Component Analysis.

This is the proportion of each variable's variance that can be explained by the factors. It is also noted as h^2 and can be defined as the sum of squared factor loadings for the variables. The values in this column extraction indicate the proportion of each variable's variance that can be explained by the retained factors. Variables with high values are well represented in the common factor space, while variables with low values are not well represented. we don't have any particularly low values. They are the reproduced variances from the factors that you have extracted.

Output IV - Component Matrix

The component matrix shows the Pearson correlation between the items and the components. For some dumb reason, these correlations are called **factor loadings**.

Component Matrix[a]

	Component			
	1	2	3	4
You are overall satisfied with your experience of shopping online.	.746			
I am satisfied with the price among these online shopping	.649			
Website layout helps in searching the products easily.	.643			
The description of products shown on the websites are very informative	.613			
I satisfied with the choice of products available in online shopping	.610			
Online shopping is as secure as traditional shopping.	.606		-.406	
I never felt any problem while conducting online purchase.	.575			
I think shopping on internet saves time.	.412			
I prefer cash on delivery than payment via credit / debit card		.708	-.429	
While shopping online, I hesitate to give my credit card number.		.643	-.426	
It is great advantage to be able to shop at any time of the day on the internet.	.427		.565	
I prefer traditional or conventional shopping than online shopping.		.418		.677
A long time is required for the delivery of products and services.				.767

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Here, marked statements correlates with more than one component. If a variable has more than 1 substantial factor loading, we call those **cross loadings**. They complicate the interpretation of our factors. The solution for this is **rotation**: we'll **redistribute the factor loadings** over the factors according to some mathematical rules that we'll leave to SPSS. This redefines what our factors represent. Now, there's different rotation methods but the most common one is the **varimax rotation**, short for "**vari**able **max**imization. It tries to redistribute the factor loadings such that each variable measures precisely one factor -which is the ideal scenario for understanding our factors

Output V - Rotated Component Matrix

Rotated Component Matrix[a]

	Component			
	1	2	3	4
It is great advantage to be able to shop at any time of the day on the internet.	.761			
Website layout helps in searching the products easily.	.687			
You are overall satisfied with your experience of shopping online.	.630			
I think shopping on internet saves time.	.654			
I am satisfied with the price among these online shopping	.697			
The description of products shown on the websites are very informative		.449		
I satisfied with the choice of products available in online shopping		.420		
I prefer cash on delivery than payment via credit / debit card			.840	
While shopping online, I hesitate to give my credit card number.			.822	
I prefer traditional or conventional shopping than online shopping.				.836
A long time is required for the delivery of products and services.				.795

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 6 iterations.

A factor (or component) represents whatever its variables have in common. Here, rotated component matrix (above) shows that first component is measured by

- Website layout helps in searching the products easily.
- It is great advantage to be able to shop at any time of the day on the internet.
- You are overall satisfied with your experience of shopping online.
- I think shopping on internet saves time
- I am satisfied with the price among these online shopping.

These variables all relate to the respondent's satisfaction towards online shopping. Therefore, we interpret component 1 as **"customer satisfaction**". This is the **underlying trait** measured by above . After interpreting all components in a similar way, we get

- Component 1 - "**Customer satisfaction"**
- Component 2 - "**Product information"**
- Component 3 - "**Mode of payment"**
- Component 4 - "**Lack of interest"**

Adding Factor Scores to Data

Factor score are often used as predictors in regression analysis or drivers in cluster analysis. They are **z-scores** their mean is **0** and their standard deviation is **1**. This complicates their interpretation.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Customer satisfaction	320	1.60	5.00	3.8719	.50894
Product information	320	1.00	5.00	3.4312	.70819
Mode of payment	320	1.00	5.00	3.4250	.91264
Lack of interest	320	1.00	5.00	3.0062	.68570
Valid N (listwise)	320				

This descriptive table shows how we interpreted our factors. Because we computed them as means, they have the same 1 - 5 scales as our input variables. This allows us to conclude that

- **"Customer satisfaction"** is rated **best** (roughly 3.8 out of 5 points) and
- **"Lack of interest"** is rated **worst** (roughly 3.0 out of 5 points).

CONCLUSION

Online shopping is becoming more popular day by day. Understanding customer's need for online selling has become challenge for marketers. Specially understanding the consumer's attitudes towards online shopping. Here, the consumer behaviour on online shopping habits is studied. The respondents are classified as the cases according to the online shopping habits by cluster analysis. It is found that using cluster analysis there are three types of customers namely, customers who prefer cash on delivery than payment via credit / debit. Customers enjoy online shopping. They agreed that they can shop at any time of the day on internet and customers who prefer traditional shopping than online shopping. Next, the questions(variables) are classified by factor analysis. Then we get, 4 underlying factors. Customer satisfaction, product information, Mode of payment, Lack of interest. The 14 questions are reduced to 4 factors. Finally, adding the factor score to the data. Then it is found that "Customer satisfaction" is rated best (roughly 3.8 out of 5 points) and "Lack of interest" is rated worst (roughly 3.0 out of 5 points).

Bibliography

- Anderson T. W. (2010) An Introduction to Multivariate Statistical Analysis (3rd ed.) John Wiley.

- Seber G. F. (2004) Multivariate Observations, John Wiley.
- Rencher,A. C. (2012) Methods of Multivariate Analysis.(3rd ed.) John Wiley.
- Johnson R.A. and Wichern D.W. (2008) Applied Multivariate Statistical Analysis. (6th ed.) Pearson education.
- Johnson, D. E. (1998): Applied Multivariate methods for Data Analysts, Duxbury Press, USA-An International Thomson Publishing Company.
- Morrison, F (2003): Multivariate Statistical Methods, Brooks/Cole, 4th Revisededn. McGraw Hill Book Company.
- Kshirsagar A.M. (1972): Multivariate Analysis, M.Dekker.
- RE Anderson, et.al., Multivariate Data Analysis, Pearson Prentice Hall,2006