Name: Oneeka Medh
Roll No.: A017
SAP ID: 75252100020

Exploring Dimensions of Airline Passenger Satisfaction: A Factor Analysis Approach

Abstract

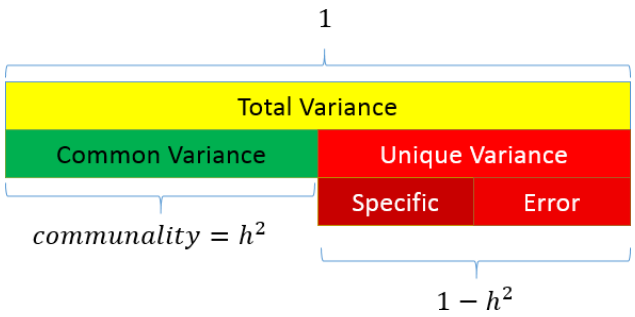
Factor Analysis is a powerful framework for demonstrating complex relationships and extracting meaningful insights from the data. In the realm of airline marketing, uncovering key dimensions that drive passenger perceptions and preferences is of paramount importance.

This paper aims to provide an intuitive understanding of factor analysis and its application to marketing research in the airline industry. By applying this statistical technique to customer satisfaction data, the objective is to extract critical factors of passenger experience that helps formulate marketing strategies and airlines can leverage such information to enhance passenger satisfaction and loyalty.

Introduction

Factor Analysis is a statistical method that estimates a model which explains the variance and covariance among a set of observed variables, also known as indicators by a set of fewer unobserved variables referred to as latent factors. Factor Analysis treats these indicators as linear combinations of the factors plus an error. The idea is that the latent factors create commonalities in the observed variables and the error term is the unique source of variance each indicator possesses.

For example, in a customer survey, a factor like 'Product Quality' can't be measured directly. However, we can access observed variables like product durability, design and value for money. These variables relate to Product Quality. The impact that product durability has on the factor is the 'weight 'or 'factor loading' and the weights will be different for other indicators. There is some proportion of product durability which is influenced by the product's overall quality, which is common among products. Yet, there are some other reasons unique to each product that also affect durability.



Factor analysis serves multiple purposes beyond explaining variance among observed variables by unobserved ones. It aids in dimensionality reduction by condensing numerous indicators into a smaller set of underlying factors, simplifying complex datasets. Additionally, it unveils the latent structure within the data, offering insights into its inherent organization. This process involves making theoretical and interpretive judgment calls on the researcher's end.

Types of Factor Analysis:

While factor analysis aims to find the latent factors in the model, researchers primarily use it for two goals: To discover the underlying structure of the data or to confirm the validity of their hypothesis. There are 2 types of factor analysis:

1. Exploratory Factor Analysis (EFA): We use EFA when we do not have a good understanding of factors present in the data. Hence, we do not form any prior hypothesis and depend on statistical outputs to determine the number of factors to extract.
2. Confirmatory Factor Analysis (CFA): We use CFA to confirm existing hypothesis developed. Hence, before we perform Factor Analysis, we must state the methodology, number of factors and other

determinants. Once FA is performed, we determine the model's goodness of fit to match those predicted by theory.

Theoretical Framework

The factor model can be described as a series of multiple regressions, predicting the indicators X_i 's from the values of the unobserved factors f_i

In the context of the example provided, our observed variables (Xi's), such as product durability, design, and value for money, are predicted through a series of multiple regressions. These indicators form a linear combination with underlying factors like product quality, customer service, etc. This relationship is captured by the equation:

$$\begin{aligned} X_1 &= \mu_1 + l_{11}f_{11} + l_{12}f_{12} + \cdots + l_{1m}f_m + \epsilon_1 \\ X_2 &= \mu_2 + l_{21}f_{21} + l_{22}f_{22} + \cdots + l_{2m}f_m + \epsilon_2 \\ &\vdots \\ X_p &= \mu_p + l_{p1}f_{p1} + l_{p2}f_{p2} + \cdots + l_{pm}f_m + \epsilon_p \end{aligned}$$

where,

μ = Intercept of the regression model

l_{ij} = The factor loading or the weight of i[th] indicator on the j[th] factor. It is a matrix of the form:

$$\mathbf{L} = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1m} \\ l_{21} & l_{22} & \dots & l_{2m} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & \dots & l_{pm} \end{pmatrix}$$

f_m= m unobserved factors that control the variation, as known as factor scores. It is a matrix of m factors on n respondents.

ϵ = The error term, which is specific to each observed variable. For the i[th] variable,

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}$$

We reduce this into matrix notation as:

$Y_{np} = f_{nm} L'_{mp} + \epsilon_{np}$

where,

$Y = X - \mu$

$n = no.of\ observations$

$m = no.of\ factors$

$p = no.of\ indicators$

We can rewrite the above equation as:

$\boldsymbol{X = \mu + Lf + \epsilon}$

Model Assumptions

The factor model relies on several key assumptions to accurately capture the relationship between the observed variables and the underlying factors. They are as follows:

1. The error terms (specific variance) have a mean of zero: $E(\epsilon_i) = 0; i = 1,2, \dots p$
2. The common factors have a mean of zero: $E(f_i) = 0; i = 1,2, \dots m$
3. The common factors have a variance of one: $var(f_i) = 1; i = 1,2, \dots m$
4. The common factors are uncorrelated with each other: $cov(f_i, f_j) = 0 \; for \; i \neq j$
5. The error terms are uncorrelated with each other: $cov(\epsilon_i, \epsilon_j) = 0 \; for \; i \neq j$

Using these assumptions, we can derive the variance of the i[th] observed variable as

$$var(X_i) = \sum_{j=1}^{m} l_{ij}^2 + \varphi_i$$

where,

l_{ij}^2 *is called the commonality for variable* i (*the sum of the squared factor loadings for each variable across all factors.*)

and φ_i *is the specific variance* (*diagonal matrix*)

$$\mathbf{\Psi} = \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p \end{pmatrix}$$

The covariance is derived as,

$$\sum = L\theta L' + \varphi$$

where,

θ *is diagonal matrix of factor variances*

Other Terminologies

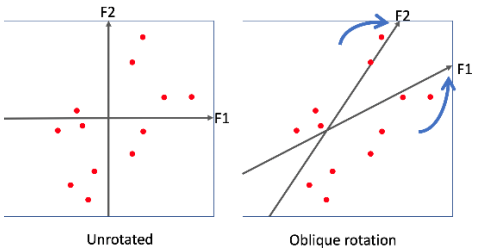
Kaiser- Meyer-Olkin (KMO): Kaiser- Meyer-Olkin is a statistical test used to determine if the data is suitable for factor analysis. KMO measures the sampling adequacy of the observed variables which is calculated based on partial correlation between variables. It ranges from 0 to 1 with values closer to 1 suggesting that factor analysis should yield distinct and reliable factors and values closer to 0 indicating that there may not be any common factor influencing the observed variables.

Barlett's test of sphericity: Bartlett's test of sphericity tests the null hypothesis that whether the correlation matrix is an identity matrix; meaning there is no significant correlation between the variables. The alternative hypothesis states that there is significant correlation among at least some of the variables. The test is considered significant when p-values are less than 0.05.

Factor Loadings: These correlations (l_{ij}) between factors and observed variables help identify which variables correspond to which factor, with coefficients ranging from -1 to 1 indicating direction and strength.

Factor Rotations: The initial set of factor loadings that we get are difficult to interpret because some variables can have high loadings for multiple factors. Rotating the factors simplifies interpretation by maximizing high loadings and minimizing low loadings. Imagine a scatterplot where each data point represents an observed variable, with X and Y coordinates indicating factor loadings. Initially, points are scattered, indicating variables aren't strongly related to specific factor or have high loadings for both

factors. Rotating the axes shifts these points to align more closely with one factor, making it easier to assign variables to specific factors based on their proximity to the axes.



Eigenvalue and Commonality: Eigenvalues represent the amount of variance in the observed variables explained by each factor. Suppose we have:

- Factor 1: X1 = 0.7, X2 = 0.5, X3 = 0.6 (loadings)
- Factor 2: X1 = 0.4, X2 = 0.6, X3 = 0.8

Eigenvalue of Factor 1= $(0.7)^2 + (0.5)^2 + (0.6)^2 = 1.1$

Whereas it is important to know that communalities will be represented as the sum of squared factor loadings for each variable represents the common variance explained by all factors combined.

Commonality of X1= $(0.7)^2 + (0.4)^2 = 0.65$

Methods of Factor Extraction

There are about three basic extraction methods, that help to estimate both the loading matrix, which captures the common variance and unique variance matrix, which accounts for variance specific to each variable. This helps us to construct and reconstruct variance- covariance matrix as mentioned above in the methodology.

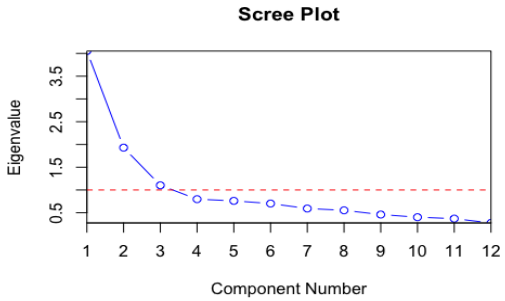
1. Maximum Likelihood Estimation (MLE): It is most used in Confirmatory Factor Analysis when the data is continuous and meets the normality assumption as it provides unbiased and precise estimates of the parameters.

2. Principal Axis Factoring (PAF): PFA is used when the data violates the normality assumption and while performing Exploratory Factor Analysis. However, it assumes that the amount of variance in each variable explained by the common factors (communality) is equal to the square of the strength of the relationship between that variable and the factors. It is commonly used when the goal is to identify underlying structure rather than precise parameter estimation.

3. Principal Component Analysis (PCA): In most statistical software packages, PCA is the default method for factor analysis. However, it is just a data reduction technique which doesn't assess the underlying commonalities caused by unobserved factors.

Deciding the number of factors

The typical method in determining the number of factors involves creating a scree plot. The scree plot is a two-dimensional graph that illustrates the factors on the x-axis and the Eigenvalues on the y-axis. The Eigenvalues relate to the amount of variance explained by the underlying factors. With the help of statistical software, we can produce a scree plot. It is important to identify the point in the data where the curve starts to bend. The number of factors before this bend is often considered the appropriate number to extract. Most software use Eigenvalue>1 to extract the number of factors.



Methodology

To implement this process with real data, we use the following procedure:

1. Select and measure variables
   - Gather data on relevant and clearly defined variables. The input data is of ordinal form which is collected through questionnaire, survey and other methods of study.

2. Data screening and adequacy
   Conduct correlation analysis which helps to examine relationship between variables.
   - Calculate the KMO statistic to examine the adequacy of data, aiming for values closer to 1.
   - Perform Bartlett's test to ensure the variables are correlated (p-value < .05) and suitable for factor analysis.

3. Data preparation
   - Centre the data by calculating the mean of each variable and subtract it from every observation.
   - Compute the sample variance-covariance matrix (S) to estimate variances and covariances between variables.

4. Factor analysis estimation
   - Utilize methods like Maximum Likelihood Estimation (MLE) or Principal Component Analysis (PCA) to estimate factor loadings (L) and unique variances (Ψ), assuming common factors have a variance one.
   - Reconstruct the variance-covariance matrix (Σ) using estimated factor loadings and unique variances, allowing insight into the underlying data structure.

5. Interpretation and model evaluation
   - Interpret the factor loading matrix to understand the relationship between observed variables and latent factors. Higher loadings indicate stronger associations between the variables and factors.
   - Examine the scree plot or the Kaiser criterion (Eigenvalue<1) to determine the appropriate number of factors.
   - Utilize indices such as Root Mean Square Error of Approximation (RMSEA) which provides information about how well the model reproduces observed covariance structure.

Case Study

The dataset contains survey data collected from airline passengers regarding their satisfaction with various aspects of their travel experience. These elements include a total of 16 observed variables measured on a Likert scale of 0(very dissatisfied) to 5(very satisfied). Based on their interactions with the airline, each respondent ranked these factors. The sample size is approximately 1.25 lakh observations since factor analysis requires sufficient sample size to produce reliable results.

Problem Statement: The aim of this analysis is to uncover key dimensions or factors that can group passenger satisfaction into fewer, more understandable components. By doing so, we seek to simplify the interpretation of passenger satisfaction patterns and identify critical factors influencing overall happiness with airline services. This insight can help airlines focus their efforts on improving specific service areas, ultimately enhancing total customer satisfaction and loyalty.

Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes
3	4	3	1	5	3	5	5	4	3	4	4	5	5	25	18
3	2	3	3	1	3	1	1	1	5	3	1	4	1	1	6
2	2	2	2	5	5	5	5	4	3	4	4	4	5	0	0
2	5	5	5	2	2	2	2	2	5	3	1	4	2	11	9
3	3	3	3	4	5	5	3	3	4	4	3	3	3	0	0
3	4	2	1	1	2	1	1	3	4	4	4	4	1	0	0
2	4	2	3	2	2	2	2	3	3	4	3	5	2	9	23
4	3	4	4	5	5	5	5	5	5	5	4	5	4	4	0
1	2	2	2	4	3	3	1	1	2	1	4	1	2	0	0

Given that we are conducting Exploratory Factor Analysis (EFA) on discrete data, our approach will involve utilizing Principal Axis Factoring (PAF). PAF is specifically suited for uncovering the underlying structure of the model.

To commence our analysis, we begin by scrutinizing the correlation matrix, depicted below.

Correlation Matrix[a]

		Inflightwifiservice	DepartureArrivaltimeconvenient	EaseofOnlinebooking	Gatelocation	Foodanddrink	Onlineboarding	Seatcomfort	Inflightentertainment	Onboardservice	Legroomservice	Baggagehandling	Checkinservice	Inflightservice	Cleanliness	DepartureDelayinMinutes	ArrivalDelayinMinutes
Correlation	Inflightwifiservice	1.000	.345	.715	.339	.132	.457	.121	.208	.120	.160	.121	.044	.110	.131	-.016	-.018
	DepartureArrivaltimeconvenient	.345	1.000	.438	.447	.001	.072	.009	-.008	.067	.011	.071	.091	.072	.010	.001	-.001
	EaseofOnlinebooking	.715	.438	1.000	.460	.031	.405	.029	.047	.039	.109	.039	.009	.035	.015	-.005	-.007
	Gatelocation	.339	.447	.460	1.000	-.003	.003	.002	.003	-.029	-.005	.001	-.039	.000	-.006	.006	.006
	Foodanddrink	.132	.001	.031	-.003	1.000	.234	.576	.623	.057	.033	.035	.085	.035	.658	-.029	-.032
	Onlineboarding	.457	.072	.405	.003	.234	1.000	.419	.284	.154	.123	.084	.204	.074	.329	-.019	-.023
	Seatcomfort	.121	.009	.029	.002	.576	.419	1.000	.612	.131	.104	.075	.190	.069	.680	-.028	-.031
	Inflightentertainment	.208	-.008	.047	.003	.623	.284	.612	1.000	.419	.301	.379	.120	.407	.692	-.027	-.030
	Onboardservice	.120	.067	.039	-.029	.057	.154	.131	.419	1.000	.358	.520	.245	.551	.122	-.030	-.035
	Legroomservice	.160	.011	.109	-.005	.033	.123	.104	.301	.358	1.000	.372	.153	.370	.097	.014	.011
	Baggagehandling	.121	.071	.039	.001	.035	.084	.075	.379	.520	.372	1.000	.235	.629	.097	-.004	-.008
	Checkinservice	.044	.091	.009	-.039	.085	.204	.190	.120	.245	.153	.235	1.000	.238	.177	-.019	-.022
	Inflightservice	.110	.072	.035	.000	.035	.074	.069	.407	.551	.370	.629	.238	1.000	.091	-.054	-.060
	Cleanliness	.131	.010	.015	-.006	.658	.329	.680	.692	.122	.097	.097	.177	.091	1.000	-.015	-.017
	DepartureDelayinMinutes	-.016	.001	-.005	.006	-.029	-.019	-.028	-.027	-.030	.014	-.004	-.019	-.054	-.015	1.000	.965
	ArrivalDelayinMinutes	-.018	-.001	-.007	.006	-.032	-.023	-.031	-.030	-.035	.011	-.008	-.022	-.060	-.017	.965	1.000

Interpretation: The correlation matrix shows how each of the 16 variables are associated with each other. We can observe that, some of the correlations are high (+/- 0.60 or greater) and some are low(i.e, near zero). Departure Delay and Arrival Delay have significantly low correlations with other variables. The positive (+) and negative (-) signs denote the direction whereas the coefficients denote the strength of association. Relatively high correlations indicate that two variables are associated and will most liekly be grouped together. Tems with low correlations (<0.20 for eg) will not have high loadings on the same factor.

Subsequently, we will proceed with KMO test to assess the saampling adequacy and Barlett's test of speherecity.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.735
Bartlett's Test of Sphericity	Approx. Chi-Square	1100454.346
	df	120
	Sig.	<.001

Interpretation: The KMO statistic is greater than 0.70 indicating that there might be common factors influencing the variables. The Barlett's test (significance <0.001) indicates that the correlation matrix is significantly different from an identity matrix, hence we can proceed further with our analysis.

Commonalities are squared estimated using the parameter estimation of the factor loadings matrix l_{ij}^2. We can verify our value by squaring the sum of factor loadings across each variable for all factors as mentioned above.
For Inflight wifi service: the 'Rotated Factor Matrix' has the following loadings across the factors: $(0.095)^2 + (0.135)^2 + (-0.009)^2 + (0.609)^2 + (0.474)^2 = 0.62288$

Communalities

	Initial	Extraction
Inflightwifiservice	.572	.622
DepartureArrivaltimeconvenient	.301	.351
EaseofOnlinebooking	.620	.801
Gatelocation	.327	.484
Foodanddrink	.537	.596
Onlineboarding	.434	.670
Seatcomfort	.575	.623
Inflightentertainment	.735	.808
Onboardservice	.429	.502
Legroomservice	.225	.251
Baggagehandling	.470	.589
Checkinservice	.177	.114
Inflightservice	.512	.648
Cleanliness	.647	.747
DepartureDelayinMinutes	.932	.964
ArrivalDelayinMinutes	.932	.966

Extraction Method: Principal Axis Factoring.

Interpretation: The commonalities quantify how well each variable is explained by the underlying factors. The initial commonalities serve as preliminary estimates and they denote the commonalities before the rotation. Variables with high commonalities (close to 1) are well explained by the identified factors, meaning a large portion of variability can be attributed to the latent factors, whereas commonalities close to 0 suggest that variability is mostly due to unique factors.

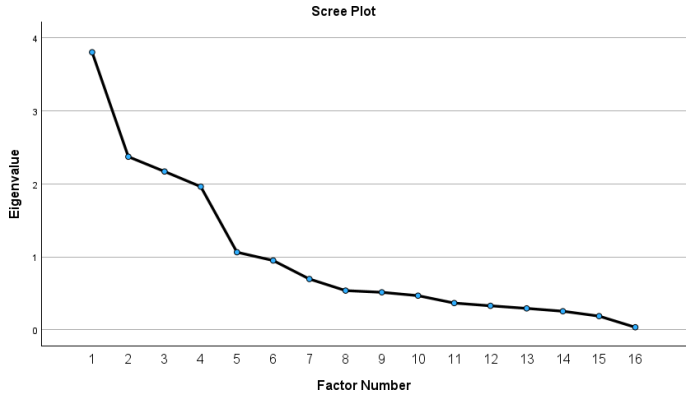
Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.804	23.774	23.774	3.444	21.528	21.528	2.606	16.287	16.287
2	2.372	14.826	38.600	2.000	12.502	34.030	2.305	14.408	30.695
3	2.170	13.560	52.160	1.925	12.030	46.060	1.933	12.083	42.778
4	1.962	12.261	64.421	1.745	10.907	56.967	1.799	11.246	54.023
5	1.063	6.644	71.065	.623	3.895	60.862	1.094	6.839	60.862
6	.950	5.935	77.000						
7	.696	4.350	81.351						
8	.537	3.358	84.708						
9	.514	3.211	87.919						
10	.468	2.924	90.843						
11	.366	2.290	93.133						
12	.329	2.055	95.188						
13	.293	1.833	97.022						
14	.254	1.590	98.612						
15	.187	1.171	99.783						
16	.035	.217	100.000						

Extraction Method: Principal Axis Factoring.

Interpretation: The distribution of the variation among the 16 potential factors is displayed in the Total Variance Explained table. It should be noted that five components meet the common requirement of having eigenvalues (a measure of explained variance) greater than 1.0 in order to be considered relevant. A factor explains less information than a single item would have when the eigenvalue is less than 1.0. The SS Loadings indicates the % of Variance explained by the common factors (derived from the commonalities table) before (Extraction table) and after rotation. Hence, we observe that 60% of the variance is explained by the first five factors which is a good indicator.

We can verify he number of factors extracted by examing the Scree Plot which yields the same output.



Factors are rotated for interpretation purpose.An orthogonal rotation (varimax) will be used for this analysis as well as future ones. This implies that the final factors will be perpendicular to one another. Consequently, we can presume that the information explained by a single factor is independent to the information in the other factors.

Rotated Factor Matrix[a]

	Factor				
	1	2	3	4	5
Cleanliness	.854	.085	.001	-.002	.098
Foodanddrink	.771	.004	-.018	.031	.035
Inflightentertainment	.767	.466	-.008	.041	.024
Seatcomfort	.757	.080	-.014	-.028	.209
Inflightservice	.036	.799	-.044	.047	-.058
Baggagehandling	.036	.765	.007	.047	-.035
Onboardservice	.085	.701	-.019	.010	.047
Legroomservice	.058	.486	.023	.042	.094
Checkinservice	.114	.288	-.013	-.028	.132
ArrivalDelayinMinutes	-.017	-.020	.983	-.001	-.008
DepartureDelayinMinutes	-.016	-.014	.982	.000	-.006
EaseofOnlinebooking	-.033	.031	-.002	.768	.458
Gatelocation	.013	-.047	.005	.686	-.107
Inflightwifiservice	.095	.135	-.009	.609	.474
DepartureArrivaltimeconvenient	-.010	.056	.000	.590	-.001
Onlineboarding	.290	.122	-.010	.103	.749

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Interpretation: The Rotated Factor Matrix is key for understanding the analysis. The analysis has sorted out 16 variables into 5 overlapping factors. The variables are sorted so that the variables that have the highest loading (not considering whether the correlation is positive or negative) from factor 1 (four variables in this analysis) are listed first, and they are sorted from the one with the highest factor weight or loading (i.e., Cleanliness, with a loading of .854) to the one with the lowest loading from that first factor (Seat Comfort). The next four variables have highest loading on factor 2 (In Flight Service with a loading of .799 and Leg Room Service .486). Arrival and Departure Delay have a loading of .983 and .982 respectively on factor 3. Ease of Online Booking, Gate Location, Inflight wifi, Departure/Arrival Time Convenience make up factor four. Factor 5 consists of Online Boarding with a loading of .749.

Results: Based on the observed grouping of variables, the factors can be names as follows:

- *Comfort and Amenities:* Cleanliness, Food and Drink, Inflight Entertainment, Seat Comfort
- *Service Quality:* Inflight Services, Baggage Handling, Onboard Services, Leg Room
- *Punctuality:* Arrival and Departure Delay
- *Facilities:* Online Booking, Gate Location, Inflight Wi-Fi, Departure/Arrival Time Convenience
- *Boarding Process:* Online Boarding

Conclusion

The goal of this study was to discover the underlying dimensions of airline passenger satisfaction. The suitability of factor analysis is demonstrated by using threshold values such as Kaiser-Meyer-Olkin, Barlett's test of Sphericity but we further confirm internal consistency by calculating Cronbach's alpha and discriminant score. Our findings reveal that factors such as comfort, service quality, punctuality, boarding process and facilities emerge as pivotal factors in measuring passenger satisfaction, as identified by principal axis factoring and varimax orthogonal rotation. These insights provide valuable guidance to the decision markers, directing their attention to manageable set of factors critical for enhancing passenger satisfaction.

In conclusion, Factor Analysis is a powerful tool for uncovering underlying structures within a dataset, offering valuable insights across various domains. Despite its assumptionsof linearity which may not always hold true in real- world and its limitations such as difficulty in handling categorical variables, a relatively large sample size and sensitivity to data quality, its applications are widespread and impactful, ranging from market research to healthcare analysis.

References

- *Hervé Abdi (2003), "Factor Rotations in Factor Analyses," In: Lewis-Beck M., Bryman, A., Futing T. (Eds.) (2003). Encyclopaedia Research Methods. Thousand Oaks (CA): Sage.*
- *Brown, Michael W., (2001) "An Overview of Analytic Rotation in Exploratory Factor Analysis," Multivariate Behavioural Research, 36 (1), 111-150.*
- *Costello, Anna B. and Osborne, Jason (2005) "Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis," Practical Assessment, Research, and Evaluation: Vol. 10 , Article 7.*
- *Mohsen Tavakol and Angela Wetzel (2020) "Factor Analysis: a means for theory and instrument development in support construct validity" International Journal of Medical Education.*
- *Syed Mohammad Ather (2009) "Factor Analysis: Nature, Mechanism & Uses in Social and management Researchers" Journal of Cost and Management Accountant.*
- *AA Jowkar (2014) "A Factor Analysis of Identifying The Customer Behaviour Patterns" European Online Journal of Natural and Social Sciences.*
- *Noora Shrestha (2021) "Factor Analysis as a Tool for Survey Analysis" American Journal of Applied Mathematics and Statistics.*