Factor Analysis of Airline Passenger Satisfaction

A Comparative Study of Explanatory and Confirmatory Approaches

presented by

RATHNAWEERA, R.P.R.M. (S/17/465)

DEPARTMENT OF STATISTICS & COMPUTER SCIENCE

FACULTY OF SCIENCE

UNIVERSITY OF PERADENIYA

SRI LANKA

2023

Table Of Contents

Introduction

The airline industry is a highly competitive market, So Passenger satisfaction is an important factor when considering the success of an airline. Understanding the factors that contribute to passenger satisfaction is essential for airlines to improve their services. Factor analysis is a statistical technique that can be employed to identify the underlying dimensions of complex datasets, such as those related to customer satisfaction.

The objective of this project is to find out the factors influencing airline passenger satisfaction using two distinct approaches: Explanatory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). By comparing the results of these two methods, we aim to gain a deeper understanding of the key facts that related to satisfaction in the airline industry.

The major question addressed in this study is: What are the underlying factors that influence passenger satisfaction in the airline industry?
To address this question,

1. We hypothesize that multiple underlying factors contribute to passenger satisfaction in the airline industry.
2. We expect that the identified factors will have a significant impact on passenger satisfaction.

The importance of this study lies in its potential to provide a comprehensive understanding of the factors that drive airline passenger satisfaction. By comparing the results of EFA and CFA, we can identify the most robust and consistent factors, which can then be used to inform targeted service improvements.

Methodology

Dataset

The data used in this project is the Airline Passenger satisfaction dataset, which can be found here: https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction

The dataset contains an airline passenger satisfaction survey. It has 129,487 observations and 23 columns. 14 of those representing customers responses on a scale of 1 to 5. The survey conducted to evaluate different aspects of the flights (food and drink, online boarding, seat comfort, Leg room service, etc).

Data Pre-processing

Data Cleaning: Removing any incomplete or inconsistent records from the dataset.

Tests Used

a. Bartlett's test of sphericity:
   This is a statistical test that is used to test the null hypothesis that the correlation matrix is an identity matrix. An identity correlation matrix means that there is no correlation between any of the variables. If the null hypothesis is rejected, then it can be concluded that there is some correlation between the variables, and factor analysis can be used to identify the underlying factors.

b. The Kaiser–Meyer–Olkin (KMO) test
   This test is a statistical measure used to determine how suited data is for factor analysis. The statistic is a measure of the proportion of variance among variables that might be common variance. The higher the proportion, the higher the KMO-value, the more suited the data is to factor analysis.

Statistical Methods

The project uses two statistical methods, EFA and CFA, to analyse the dataset and identify the underlying factors contributing to airline passenger satisfaction.

a. Explanatory Factor Analysis (EFA): EFA is an exploratory technique used to identify the latent factors or dimensions underlying a set of observed variables. In this study, EFA was conducted using principal component analysis (PCA) as the extraction method and varimax rotation to simplify the factor structure. The number of factors to retain was determined using the Kaiser criterion (eigenvalues greater than 1) and the scree plot.

b. Confirmatory Factor Analysis (CFA): CFA is a hypothesis-driven technique used to test the fit of a predefined factor structure to the observed data. Here , CFA was conducted using structural equation modelling (SEM) to assess the fit of the factor structure identified through EFA. Model fit was evaluated using various indices, such as the chi-square statistic, comparative fit index (CFI), Tucker-Lewis index (TLI), and root mean square error of approximation (RMSEA).

Results and Discussion

Test Results

First, I apply Kaiser - Meyer- Olkin factor test for check whether this dataset can be used for this analysis. It gives Overall MSA value as 0.79 which is higher value. So according to kaiser – Meyer – Olkin test, this dataset can be used for factor analysis.

Then I performed Bartlett's test of Sphericity. The resulting p-value was 0, which is less than the significance level of 0.05. So, we can reject the null hypothesis that is no correlation between any of the variables. So, it can be concluded that there is some correlation between the variables, and factor analysis can be used to identify the underlying factors.

Explanatory Factor Analysis

As the first step, I determined the number of factors to be considered. So here I used two methods. Scree plot and parallel Analysis scree plots. Parallel analysis compares the eigenvalues of the observed data to the eigenvalues of random data matrices of the same size. So, the Parallel analysis suggested that the number of factors are 5.

Then I performed factor analysis. This is the loading matrix I obtained.

```
Loadings:
                                  RC1    RC3    RC2    RC4    RC5
Seat.comfort                             0.125  0.539  0.675 -0.105
Departure.Arrival.time.convenient                0.840         0.108
Food.and.drink                                   0.664  0.614
Gate.location                                    0.863
Inflight.wifi.service             0.848
Inflight.entertainment            0.315                  0.826  0.196
Online.support                    0.814                  0.191  0.157
Ease.of.Online.booking            0.780  0.485
On.board.service                         0.774                  0.126
Leg.room.service                         0.681          0.191
Baggage.handling                         0.819                  0.106
Checkin.service                   0.103  0.200                  0.941
Cleanliness                              0.831                  0.101
Online.boarding                   0.878

                 RC1   RC3   RC2   RC4   RC5
SS loadings    2.900 2.731 2.198 1.618 1.036
Proportion Var 0.207 0.195 0.157 0.116 0.074
Cumulative Var 0.207 0.402 0.559 0.675 0.749
```

In factor analysis, Factor loadings represent the strength and direction of the relationship between each variable and the extracted factors. A high positive loading indicates that a variable has a strong positive relationship with a particular factor, meaning that as the factor increases, the variable also tends to increase. A high negative loading suggests a strong negative relationship, meaning that as the factor increases, the variable tends to decrease. A loading close to zero indicates a weak or no relationship between the variable and the factor. In This result we can see that factor 1(RC1) highly correlated with Inflight Wi-Fi service, Online support, Ease of online booking and Online Boarding.

Factor 2 (RC2) highly correlated with the variables Departure Arrival time convenient, Food and Drink, Gate location.

Factor 3 (RC3) highly correlated with the variables Cleanliness, Baggage handling, On board service and Leg room service.

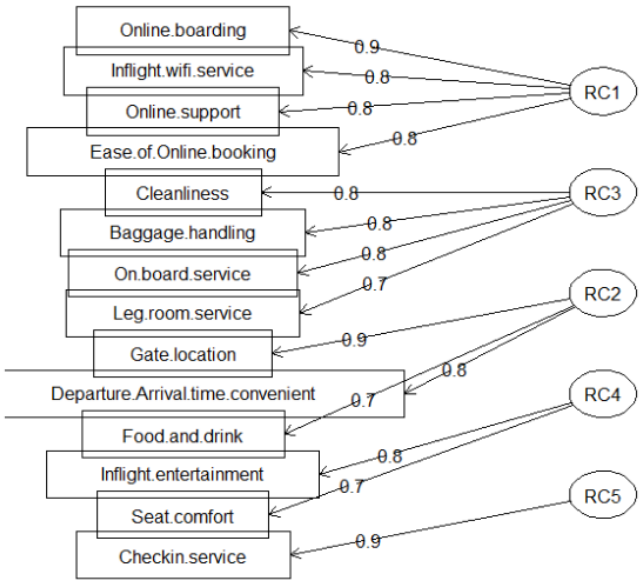
Factor 4 (RC4) highly correlated with the variables Inflight entertainment, Seat comfort and Factor 5 (RC5) highly correlated with the variable Check-in service.

Communality residuals

Seat.comfort	Departure.Arrival.time.convenient
0.22028543	0.27652054
Food.and.drink	Gate.location
0.17525341	0.25120912
Inflight.wifi.service	Inflight.entertainment
0.27779536	0.16799069
Online.support	Ease.of.Online.booking
0.27112670	0.14330493
On.board.service	Leg.room.service
0.37067193	0.49084761
Baggage.handling	Checkin.service
0.31179660	0.05772543
Cleanliness	Online.boarding
0.29131319	0.21040398

In factor analysis, uniqueness values represent the proportion of variance in a variable that is not explained by the extracted factors. A high uniqueness value for a variable suggests that the extracted factors do not explain much of its variance, meaning that the variable may not be strongly related to the underlying factors. Conversely, a low uniqueness value indicates that a large proportion of the variable's variance is explained by the factors, suggesting a strong relationship between the variable and the underlying factors. In this result it can see that all the uniqueness values are significantly low.

Graph Factor Loading Matrices



Comfimetry Factor analysis

```
User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.834
  Tucker-Lewis Index (TLI)                       0.777
```

The values in the "User Model versus Baseline Model" section are fit indices that are used to evaluate the fit of the user model to the data. The Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) are both incremental fit indices, which means that they compare the fit of the user model to the fit of a baseline model. The baseline model is a model that does not have any constraints on the relationships between the observed and latent variables.

A CFI of 0.834 and a TLI of 0.777 are both considered to be good fits. In general, a CFI of 0.90 or greater and a TLI of 0.90 or greater are good fits.

```
  RMSEA                                          0.122
  90 Percent confidence interval - lower         0.121
  90 Percent confidence interval - upper         0.122
  P-value H_0: RMSEA <= 0.050                    0.000
  P-value H_0: RMSEA >= 0.080                    1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.085
```

The Root Mean Square Error of Approximation (RMSEA) is a measure of how well a model fits the data. In this case, the RMSEA value is 0.122. The 90% confidence interval for the RMSEA ranges from 0.121 to 0.122. The p-value for the null hypothesis that the RMSEA is less than or equal to 0.05 is 0.000, indicating that this hypothesis can be rejected.

The Standardized Root Mean Square Residual (SRMR) is another measure of how well a model fits the data. In this case, the SRMR value is 0.085.

These results suggest that the model does not fit the data very well, as indicated by the relatively high values of RMSEA and SRMR.

Since the model fit is not satisfactory, we can allow the error terms to correlate and refitting the model. So I get these results for the refitted model

```
User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.931
  Tucker-Lewis Index (TLI)                       0.895
```

In here CFI of 0.931 and a TLI of 0.895 are both considered to be good fits.

```
Root Mean Square Error of Approximation:

  RMSEA                                          0.084
  90 Percent confidence interval - lower         0.083
  90 Percent confidence interval - upper         0.084
  P-value H_0: RMSEA <= 0.050                    0.000
  P-value H_0: RMSEA >= 0.080                    1.000

Standardized Root Mean Square Residual:

  SRMR                                           0.056
```

In this case, the RMSEA value is 0.084. The 90% confidence interval for the RMSEA ranges from 0.083 to 0.084. The p-value for the null hypothesis that the RMSEA is less than or equal to 0.05 is 0.000, indicating that this hypothesis can be rejected. In this case, the SRMR value is 0.056.

These results suggest that the model fits the data reasonably well, as indicated by the relatively low values of RMSEA and SRMR.

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  F1 =~
    Online.boardng    1.000
    Es.f.Onln.bkng    0.991    0.003  343.614    0.000
    Online.support    0.916    0.003  308.642    0.000
    Inflght.wf.srv    0.877    0.003  287.499    0.000
  F2 =~
    Cleanliness       1.000
    Baggage.hndlng    1.084    0.005  221.177    0.000
    On.board.servc    0.968    0.005  198.447    0.000
    Leg.room.servc    0.841    0.005  166.270    0.000
  F3 =~
    Gate.location     1.000
    Food.and.drink    1.886    0.010  189.753    0.000
    Dprtr.Arrvl.t.    1.188    0.006  200.030    0.000
  F4 =~
    Inflght.ntrtnm    1.000
    Seat.comfort      0.494    0.007   75.006    0.000
  F5 =~
    Checkin.servic    1.000
```

The estimates in the table represent the factor loadings of the observed variables on their respective latent variables. A factor loading represents the relationship between an observed variable and its corresponding latent variable. Higher factor loadings indicate a stronger relationship between the observed variable and the latent variable.

Conclusion and recommendation

Conclusion

The results of the exploratory factor analysis (EFA) suggest that there are five latent factors that underlie the airline passenger satisfaction data. These factors are:

RC1 represents a latent variable related to online services (such as online boarding and online booking, Online support, Inflight wi-fi service),

RC3 represents a latent variable related to in-flight services (such as cleanliness and baggage handling, Lag room service, On board service),

RC2 represents a latent variable related to airport services (such as gate location and food and drink, Departure Arrival time convenient),

RC4 represents a latent variable related to in-flight entertainment and seat comfort,

RC5 represents a latent variable related to check-in service.

The five latent factors explain 74.9% of the variance in the airline passenger satisfaction data. This suggests that the EFA was successful in identifying the key factors that influence passenger satisfaction.

Recommendation

The results of the EFA suggest that airlines can improve passenger satisfaction by focusing on the following areas:

- **Convenience**: Airlines can improve the convenience of the flight by making sure that flights depart and arrive on time, that gate locations are convenient, and that food and drink are available.
- **Inflight experience**: Airlines can improve the inflight experience by providing good wifi service, entertainment, and support.
- **Ease of booking and boarding**: Airlines can improve the ease of booking and boarding by making the online booking process easy to use and by providing a convenient way to board the plane.
- **Cleanliness:** Airlines can improve the cleanliness of the aircraft by cleaning the aircraft more frequently and by training employees on how to keep the aircraft clean.
- **Onboard service:** Airlines can improve the onboard service by providing more legroom and by providing friendly and helpful flight attendants.

By focusing on these areas, airlines can improve passenger satisfaction.

References:

Brown, T. (2006). Confirmatory Factor Analysis for Applied Research. *The American Statistician*, *62*(1), 91–92. https://doi.org/10.1198/tas.2008.s98

Goldberg, L. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Kline, R. (2005). Response to Leslie Hayduk's Review of Principles and Practice of Structural Equation Modeling, 4th Edition. *Canadian Studies in Population*, *45*(3–4), 188. https://doi.org/10.25336/csp29418

Kline, R. (2016). Response to Leslie Hayduk's Review of Principles and Practice of Structural Equation Modeling, 4th Edition. *Canadian Studies in Population*, *45*(3–4), 188. https://doi.org/10.25336/csp29418

(N.d.). https://towardsdatascience.com/exploratory-factor-analysis-in-r-e31b0015f224

Appendices:

Dataset

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
satisfaction	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Seat comfort	Departure/ Arrival time	Food and drink	Gate location	Inflight wifi service	Inflight entertainment	Online support	Ease of Online booking	On-board service	Leg room service	Baggage handling	Checkin service	Cleanliness	Online boarding	Departure Delay in Minutes	Arrival Delay in Minutes
satisfied	Female	Loyal Custo	65	Personal Tri	Eco	265	0	0	0	2	2	4	2	3	3	0	3	5	3	2	0	0
satisfied	Male	Loyal Custo	47	Personal Tri	Business	2464	0	0	0	3	0	2	2	3	4	4	4	2	3	2	310	305
satisfied	Female	Loyal Custo	15	Personal Tri	Eco	2138	0	0	0	3	2	0	2	2	3	3	4	4	4	2	0	0
satisfied	Female	Loyal Custo	60	Personal Tri	Eco	623	0	0	0	3	3	4	3	1	1	0	1	4	1	3	0	0
satisfied	Female	Loyal Custo	70	Personal Tri	Eco	354	0	0	0	3	4	3	4	2	2	0	2	4	2	5	0	0
satisfied	Male	Loyal Custo	30	Personal Tri	Eco	1894	0	0	0	3	2	0	2	2	5	4	5	5	4	2	0	0
satisfied	Female	Loyal Custo	66	Personal Tri	Eco	227	0	0	0	3	2	5	5	5	5	0	5	5	5	3	17	15
satisfied	Male	Loyal Custo	10	Personal Tri	Eco	1812	0	0	0	3	2	0	2	2	3	3	4	5	4	2	0	0
satisfied	Female	Loyal Custo	56	Personal Tri	Business	73	0	0	0	3	5	3	5	4	4	0	1	5	4	4	0	0
satisfied	Male	Loyal Custo	22	Personal Tri	Eco	1556	0	0	0	3	2	0	2	2	2	4	5	3	4	2	30	26
satisfied	Female	Loyal Custo	58	Personal Tri	Eco	104	0	0	0	3	3	3	3	3	3	0	1	2	3	5	47	48
satisfied	Female	Loyal Custo	34	Personal Tri	Eco	3633	0	0	0	4	2	0	2	2	3	2	5	2	5	2	0	0
satisfied	Male	Loyal Custo	62	Personal Tri	Eco	1695	0	0	0	4	5	0	5	5	1	3	2	2	4	5	0	0
satisfied	Male	Loyal Custo	35	Personal Tri	Eco	1766	0	1	0	1	4	0	4	4	3	5	2	3	2	4	0	0
satisfied	Female	Loyal Custo	47	Personal Tri	Eco	84	0	1	0	1	5	2	1	5	5	0	5	2	5	2	40	48
satisfied	Male	Loyal Custo	60	Personal Tri	Eco	1373	0	1	0	1	1	0	1	1	3	4	1	4	2	1	0	0
satisfied	Female	Loyal Custo	13	Personal Tri	Eco	3693	0	1	0	2	4	0	4	4	4	4	1	3	1	4	5	0
satisfied	Female	Loyal Custo	52	Personal Tri	Business	2610	0	1	0	2	1	2	2	1	1	0	1	2	1	3	0	0
satisfied	Female	Loyal Custo	55	Personal Tri	Eco	2554	0	1	0	2	0	1	1	2	1	1	2	1	3	1	0	0
satisfied	Female	Loyal Custo	28	Personal Tri	Eco	3095	0	1	0	2	3	0	3	3	2	5	2	3	2	3	0	0
satisfied	Female	Loyal Custo	9	Personal Tri	Eco	3305	0	1	0	2	3	0	5	3	1	1	1	3	3	3	0	0
satisfied	Female	Loyal Custo	10	Personal Tri	Eco	2090	0	1	0	2	1	0	1	1	3	5	1	4	2	1	0	0
satisfied	Female	Loyal Custo	25	Personal Tri	Eco	2122	0	1	0	2	2	0	4	2	4	1	3	1	3	2	0	0
satisfied	Male	Loyal Custo	53	Personal Tri	Business	1099	0	1	0	2	1	3	3	1	1	0	1	3	1	1	0	0
satisfied	Female	Loyal Custo	16	Personal Tri	Eco Plus	1747	0	1	0	2	2	0	2	2	3	3	2	4	3	2	0	0
satisfied	Male	Loyal Custo	30	Personal Tri	Eco	1817	0	1	0	2	4	0	4	4	2	1	3	3	2	4	0	0
satisfied	Male	Loyal Custo	64	Personal Tri	Eco	1707	0	1	0	2	5	0	3	5	4	4	2	3	2	5	0	0
satisfied	Female	Loyal Custo	42	Personal Tri	Eco	470	0	1	0	2	3	2	2	3	3	0	3	1	3	4	2	23
satisfied	Male	Loyal Custo	9	Personal Tri	Eco	972	0	1	0	2	4	0	4	4	4	3	3	1	3	4	0	0
satisfied	Female	Loyal Custo	35	Personal Tri	Eco	3695	0	1	0	3	0	4	4	2	2	3	4	4	3	4	0	0
satisfied	Male	Loyal Custo	62	Personal Tri	Eco Plus	2948	0	1	0	3	5	0	5	5	4	1	2	2	2	5	34	19
satisfied	Female	Loyal Custo	21	Personal Tri	Eco	2823	0	1	0	3	2	0	2	2	2	2	2	2	3	2	4	0
satisfied	Female	Loyal Custo	20	Personal Tri	Eco	2485	0	1	0	3	2	0	2	2	2	3	3	4	3	2	0	0
satisfied	Female	Loyal Custo	26	Personal Tri	Eco	2408	0	1	0	3	4	0	4	4	1	4	4	2	3	4	0	0

Codes

```
library("psych")
library("corrplot")
library("psych")
library("ggplot2")
library("stats")

df = read.csv('E:/Uni Docs/4th year/1st Sem/ST405/Mini Project/Invistico_Airline.csv',
              header = T)

describe(df)

colnames(df)
```

Data Cleaning

```
df <- na.omit(df)
```

Extract Survey Results

```
clean_df <- df[8:21]
clean_df

library(corrplot)
cor_matrix <- cor(clean_df)
corrplot(cor_matrix, method = "circle")
```

Calculate Eigan Values and Eigan Vectors

```
df.eigen <- eigen(cor_matrix)
```

Kaiser-Meyer-Olkin factor adequacy test

```
KMO(cor_matrix)
```

The total KMO is `0.79`, indicating that, based on this test, we can probably conduct a factor analysis.

Bartlett's Test of Sphericity

```
cortest.bartlett(cor_matrix, n = 384)
```

No of factors to extract

```
fafitfree <- fa(clean_df,nfactors = ncol(X), rotate = "none")
n_factors <- length(fafitfree$e.values)
scree     <- data.frame(
  Factor_n =  as.factor(1:n_factors),
  Eigenvalue = fafitfree$e.values)

ggplot(scree, aes(x = Factor_n, y = Eigenvalue, group = 1)) +
  geom_point() + geom_line() +
  xlab("Number of factors") +
  ylab("Initial eigenvalue") +
  labs( title = "Scree Plot",
        subtitle = "(Based on the unreduced correlation matrix)")

parallel <- fa.parallel(cor_matrix ,n.obs=129487)
```

Conducting the Factor Analysis

```
fa_2 <- principal(clean_df,nfactors = 5,rotate = "varimax", covar = FALSE)
fa_2$values

fa_2$loadings

fa_2$uniquenesses

fa.diagram(fa_2$loadings)
```

Comfemetry Factor analysis

```
library(lavaan)

# Define the CFA model
cfa_model <- '
  # Factor 1
  F1 =~ Online.boarding + Ease.of.Online.booking  + Online.support + Infli
ght.wifi.service

  # Factor 2
  F3 =~ Cleanliness  + Baggage.handling + On.board.service + Leg.room.serv
ice

  # Factor 3
  F2 =~ Gate.location  + Food.and.drink  + Departure.Arrival.time.convenie
nt

  # Factor 4
  F4 =~ Inflight.entertainment + Seat.comfort
```

```
  # Factor 5
  F5 =~ Checkin.service
'

# Fit the CFA model
fit <- cfa(cfa_model, data = clean_df)

# Check the model fit
summary(fit, fit.measures = TRUE)

# Check the standardized loadings and thresholds
inspect(fit, what = "std")
```

Model is not satisfactory. So allow the error terms to correlate and refitting the model

```
# Display top 20 modification indices
mod_indices <- modificationIndices(fit)
head(mod_indices[order(-mod_indices$mi),], n = 20)

cfa_model_refined <- '
  # Factor 1
  F1 =~ Online.boarding + Ease.of.Online.booking + Online.support + Inflig
ht.wifi.service

  # Factor 2
  F2 =~ Cleanliness + Baggage.handling + On.board.service + Leg.room.servi
ce

  # Factor 3
  F3 =~ Gate.location + Food.and.drink + Departure.Arrival.time.convenient

  # Factor 4
  F4 =~ Inflight.entertainment + Seat.comfort

  # Factor 5
  F5 =~ Checkin.service

  # Correlate error terms based on modification indices
  Gate.location ~~ Departure.Arrival.time.convenient
  Online.support ~~ Inflight.entertainment
  Gate.location ~~ Inflight.entertainment
  Food.and.drink    ~~  Inflight.entertainment
  Ease.of.Online.booking    ~~  Checkin.service
  Departure.Arrival.time.convenient ~~  Inflight.entertainment
  Ease.of.Online.booking    ~~  Cleanliness
  Ease.of.Online.booking    ~~  On.board.service

'

cfa_fit_refined <- cfa(cfa_model_refined, data = clean_df)
summary(cfa_fit_refined, fit.measures = TRUE)
```