



FACTOR ANALYSIS ON AIR LINE DATA

Presented By :-
Utkarsh Mishra - 934

Agenda

- To identify the passenger satisfaction level
 - Identifying the hidden factor
 - Estimating the factor
- K-Means



Factor Analysis

- ❑ Factor analysis takes a large number of variables and reduces or summarizes it to represent them in different smaller factors, those factors are made up of the initial set of variables.
- ❑ A common usage of factor analysis is in developing scale/questionnaires for measuring constructs that are not directly observable in real life.

Variable

1. Gender: Gender of the passengers (Female, Male)
2. Customer Type: The customer type (Loyal customer, disloyal customer)
3. Age: The actual age of the passengers
4. Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
5. Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
6. Flight distance: The flight distance of this journey
7. **Inflight wifi service**: Satisfaction level of the inflight wifi service (0: Not Applicable; 1-5)
8. **Departure/Arrival time convenient**: Satisfaction level of Departure/Arrival time convenient
9. **Ease of Online booking**: Satisfaction level of online booking
10. **Gate location**: Satisfaction level of Gate location
11. **Food and drink**: Satisfaction level of Food and drink
12. **Online boarding**: Satisfaction level of online boarding
13. **Seat comfort**: Satisfaction level of Seat comfort
14. **Inflight entertainment**: Satisfaction level of inflight entertainment
15. **On-board service**: Satisfaction level of On-board service
16. **Leg room service**: Satisfaction level of Leg room service
17. **Baggage handling**: Satisfaction level of baggage handling
18. **Check-in service**: Satisfaction level of Check-in service
19. **Inflight service**: Satisfaction level of inflight service
20. **Cleanliness**: Satisfaction level of Cleanliness
21. Departure Delay in Minutes: Minutes delayed when departure
22. Arrival Delay in Minutes: Minutes delayed when Arrival
23. **Satisfaction: Airline satisfaction level (Satisfaction, neutral or dissatisfaction)**

Factorability

Bartlett's test of Sphericity

Bartlett's test of sphericity is a test statistic used to examine the hypothesis that the variables are uncorrelated in the population. In other words, the population correlation matrix is an identity matrix; each variable correlates perfectly with itself ($r = 1$) but has no correlation with the other variables ($r = 0$).

Hypothesis:-

H0:The correlation matrix is an identity matrix

H1:The correlation matrix is not an identity matrix

Chi-Square-Statistic : 601676.8938564031

P-value: 0.0

P-Value is < 0.05 , then we reject null hypothesis at 5% level of significance.

Factorability

Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is a statistics used to examine the appropriateness of factor analysis based on the sample of the study. A high value of statistic (from 0.5 – 1) indicates the appropriateness of the factor analysis for the data in hand, whereas a low value of statistic (below 0.5) indicates the inappropriateness of the factor analysis.

Hypothesis:

H₀: The variables are not related.

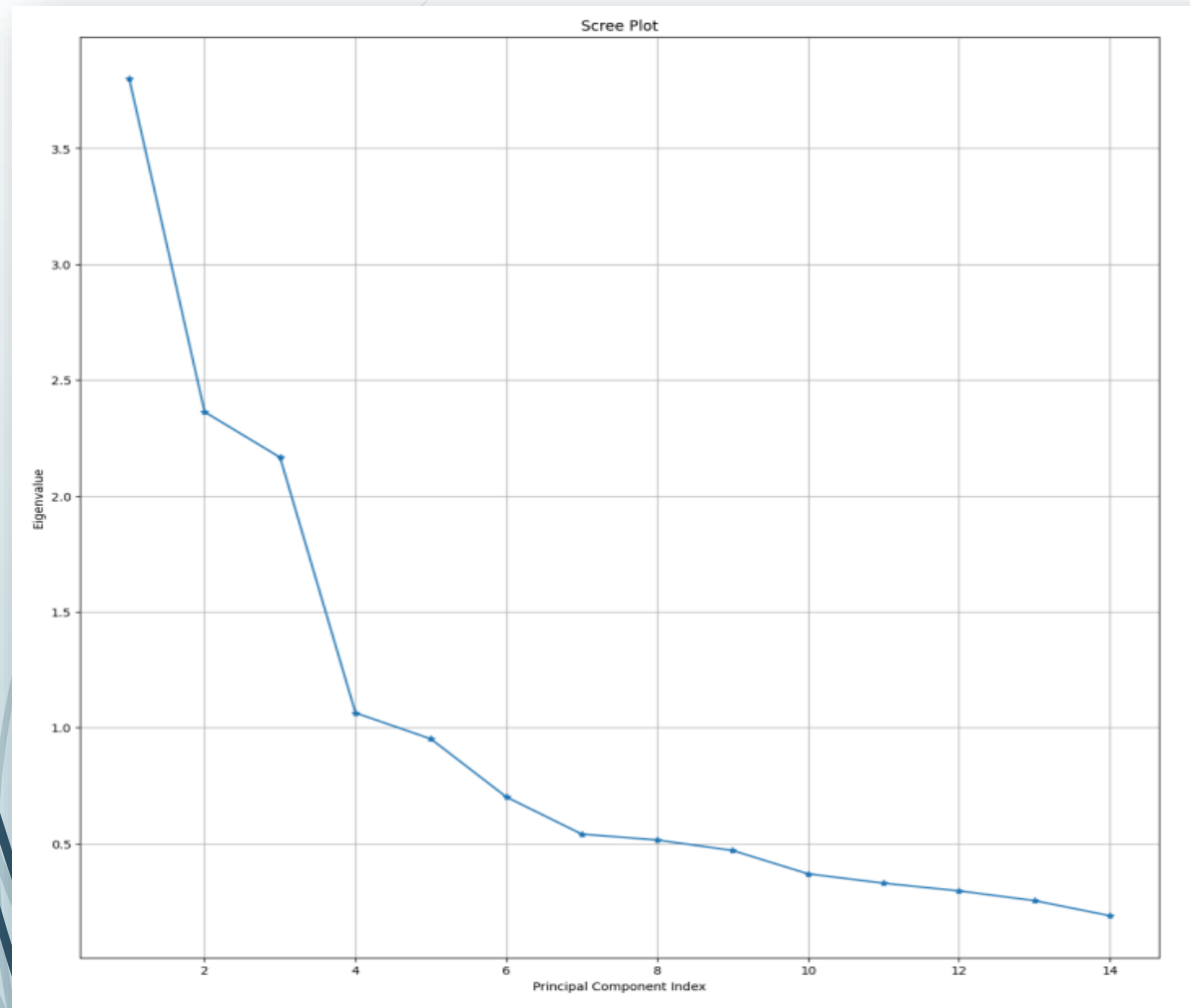
H₁: The variables are related.

KMO for All Variables: [0.74185992, 0.75273127, 0.68173114, 0.7066898, 0.84090996, 0.73646005, 0.8328631, 0.76790363, 0.82883039, 0.88966489, 0.8147844, 0.70010319, 0.78427772, 0.81687404]

KMO for Model: 0.78123271548

Based on these results, We concluded that dataset is appropriately suited for the Factor Analysis.

NUMBER OF FACTOR



- **Screen Plot** : It is a plot of eigenvalues and factor number according to the order of extraction. This plot is used to determine the optimal number of factors to be retained in the final solution.

NUMBER OF FACTOR

	Eigen	Variance_ratio	CumulativeVariance
comp 1	3.800153	27.143691	27.143691
comp 2	2.362009	16.871328	44.015020
comp 3	2.165913	15.470659	59.485678
comp 4	1.063284	7.594814	67.080493
comp 5	0.950940	6.792366	73.872859
comp 6	0.700342	5.002396	78.875255
comp 7	0.539962	3.856831	82.732086
comp 8	0.514660	3.676107	86.408194
comp 9	0.469479	3.353391	89.761585
comp 10	0.368664	2.633286	92.394871
comp 11	0.328411	2.345771	94.740641
comp 12	0.295098	2.107826	96.848467
comp 13	0.253173	1.808364	98.656831
comp 14	0.188045	1.343169	100.000000

- **Percentage of Variance Criteria** : The number of factors should be included in the model for which cumulative percentage of variance reaches a satisfactory level. The general recommendation is that the factors explaining 60%–70% of the variance should be retained in the model.
- **Eigen Value** : An eigenvalue is the amount of variance in the variable taken for the study that is associated with a factor. According to eigenvalue criteria, the factors having more than one eigenvalue are included in the model.

Rotation

•**Rotation:** solves this kind of interpretation difficulty. The main objective of rotation is to produce a relatively simple structure in which there may be a high factor loading on one factor and a low factor loading on all other factors.

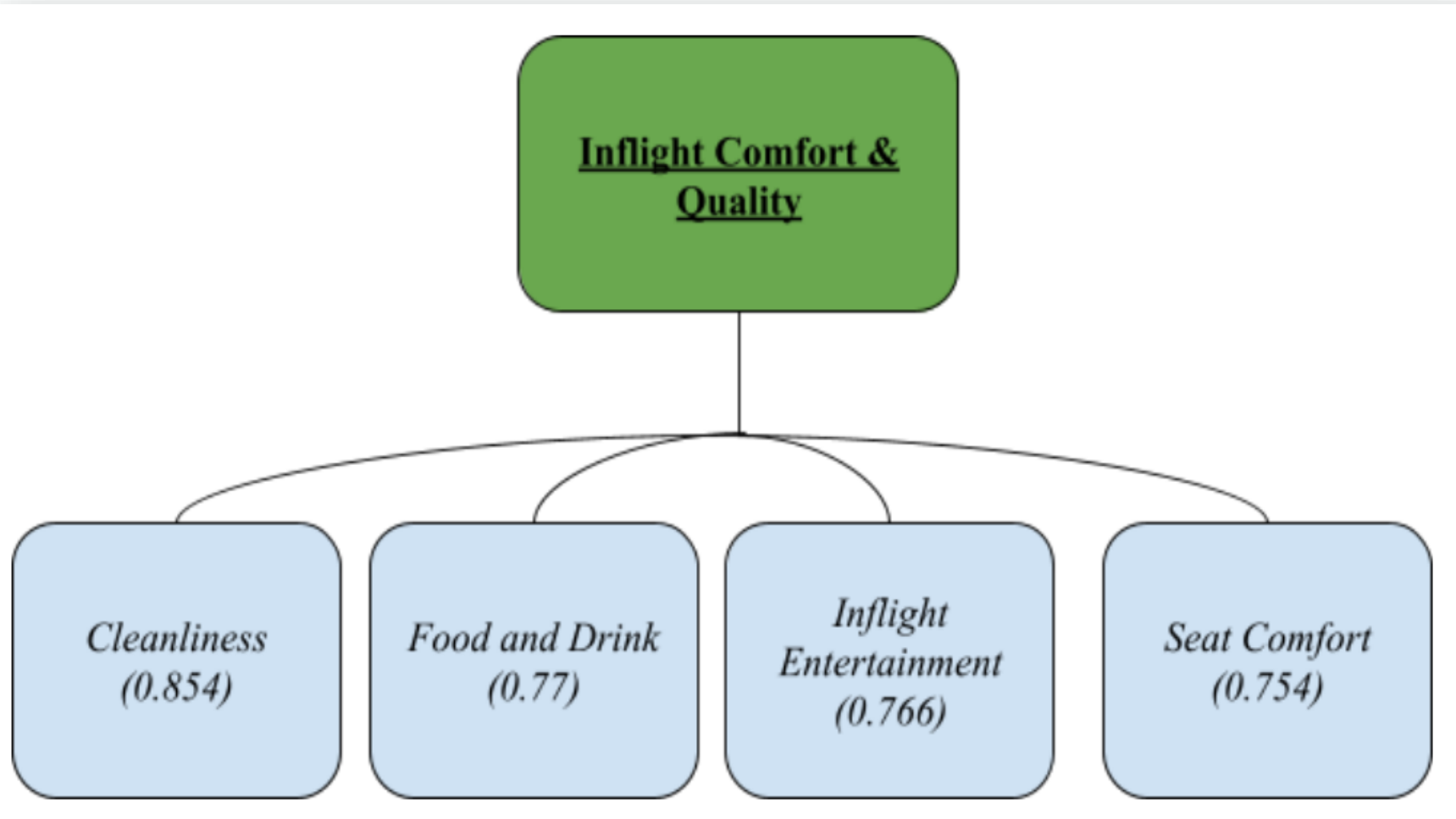
Varimax Rotation:

Minimizes the correlation between factors.
Makes it possible to identify a variable with a factor.
Components are always orthogonal—each component explains non-redundant information

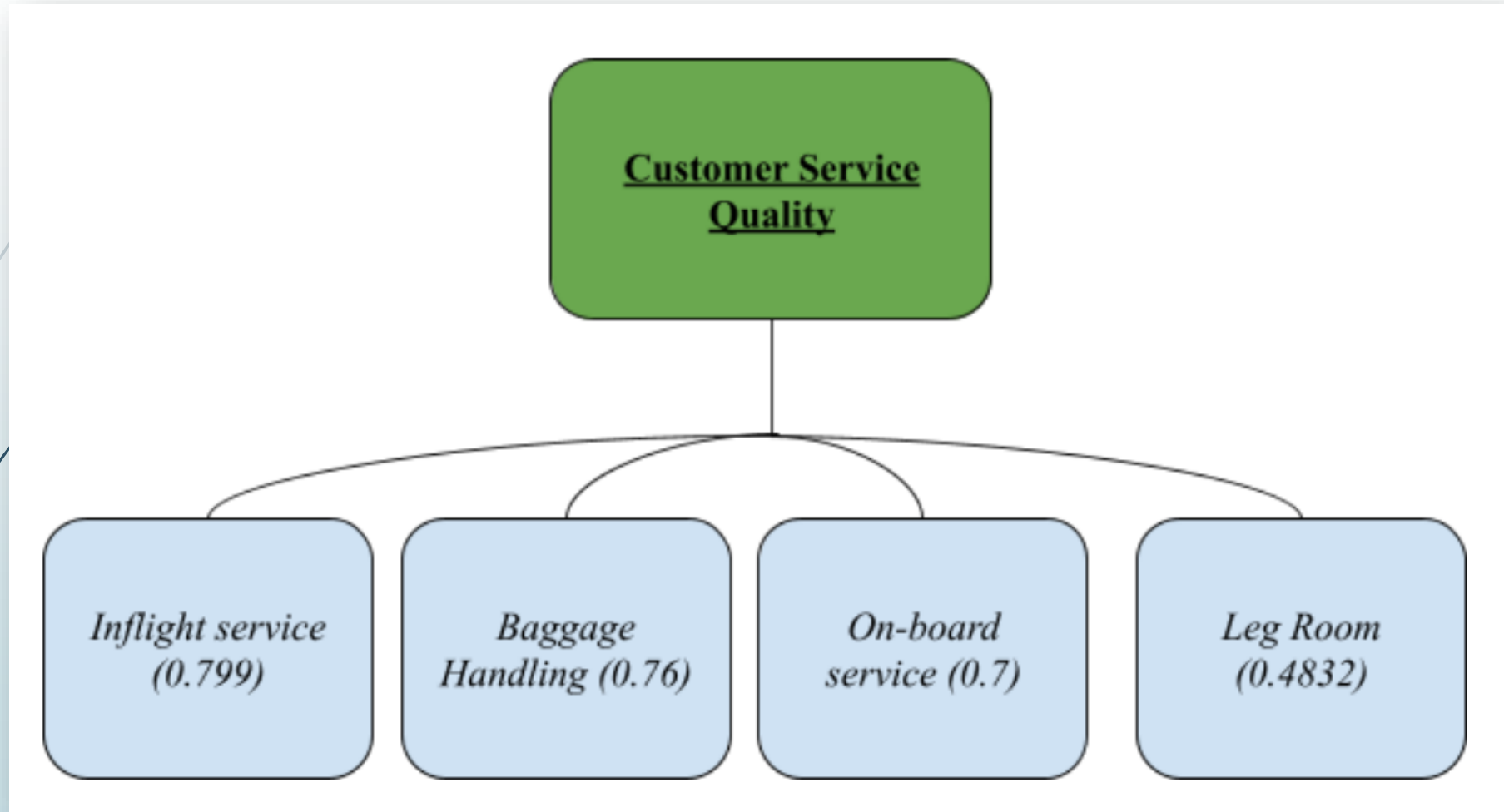
Varimax Rotation

	Factor1	Factor2	Factor3	Factor4
Inflight wifi service	0.0926063714	0.1323029101	0.06056295078	0.4780347071
Departure/Arrival time convenience	-0.006286899662	0.05716634016	0.5896429509	0.000283061667
Ease of Online booking	-0.03612383274	0.02750312899	0.7665089725	0.4633959034
Gate location	0.01309709553	-0.0451418857	0.6808125887	-0.1000437613
Food and drink	0.7701297431	0.002845115179	0.03298738325	0.04003352651
Online boarding	0.2868951661	0.1185121927	0.09430101417	0.7563822805
Seat comfort	0.754094308	0.07864630746	-0.02809953817	0.2136683889
Inflight entertainment	0.7662368339	0.4646659823	0.04122247048	0.03270846161
On-board service	0.08793107527	0.7004380526	0.01038486543	0.05184982528
Leg room service	0.05754094322	0.4832076846	0.0405585616	0.09748457034
Baggage handling	0.03673765319	0.7633841338	0.04761661197	-0.03058314085
Checkin service	0.1168161162	0.2857824714	-0.02550748074	0.1319143653
Inflight service	0.03596241233	0.7996406077	0.04759518138	-0.0515726381
Cleanliness	0.8543256811	0.08244969103	-4.59E-05	0.1031427739

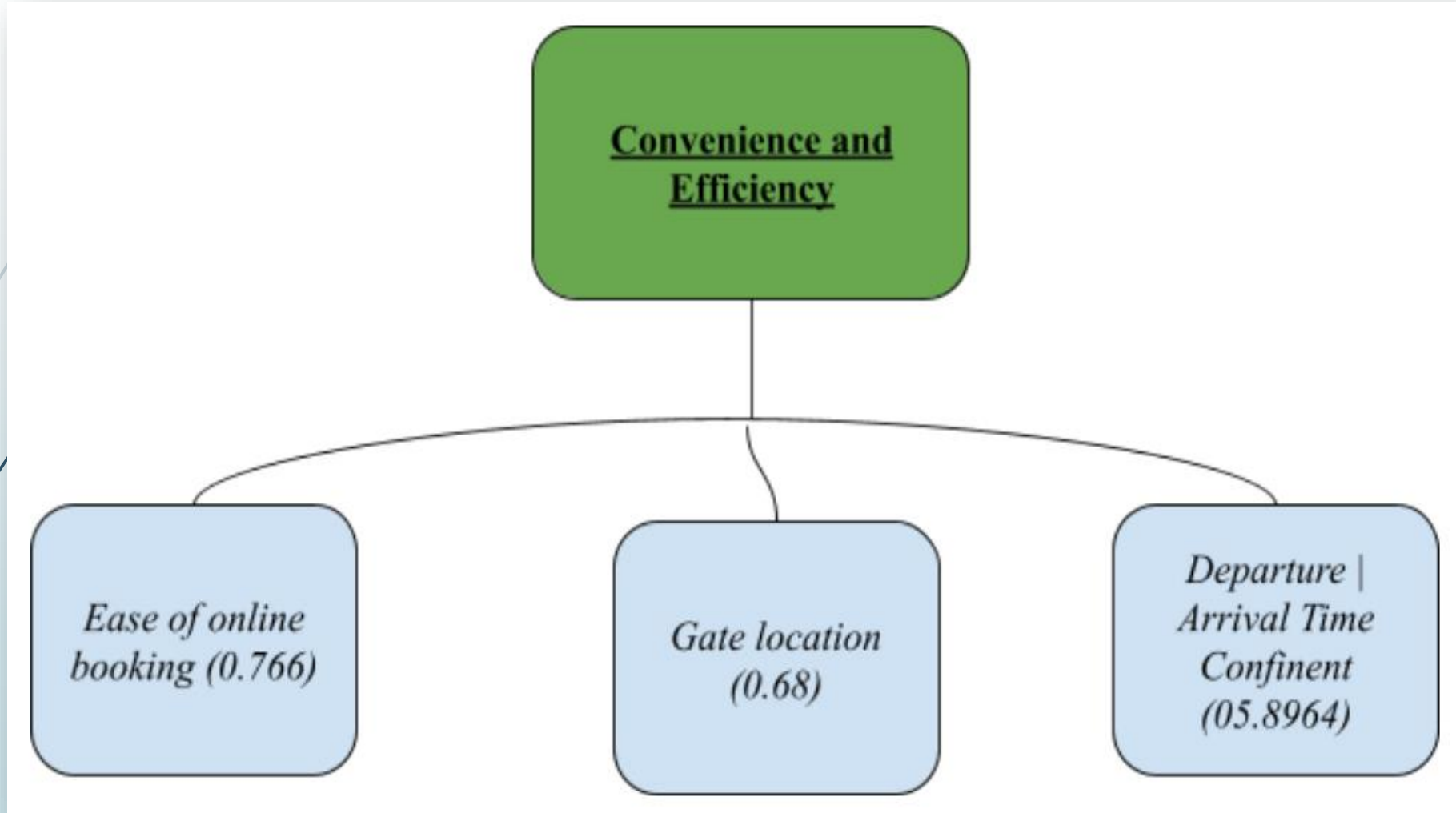
Factor Fitting



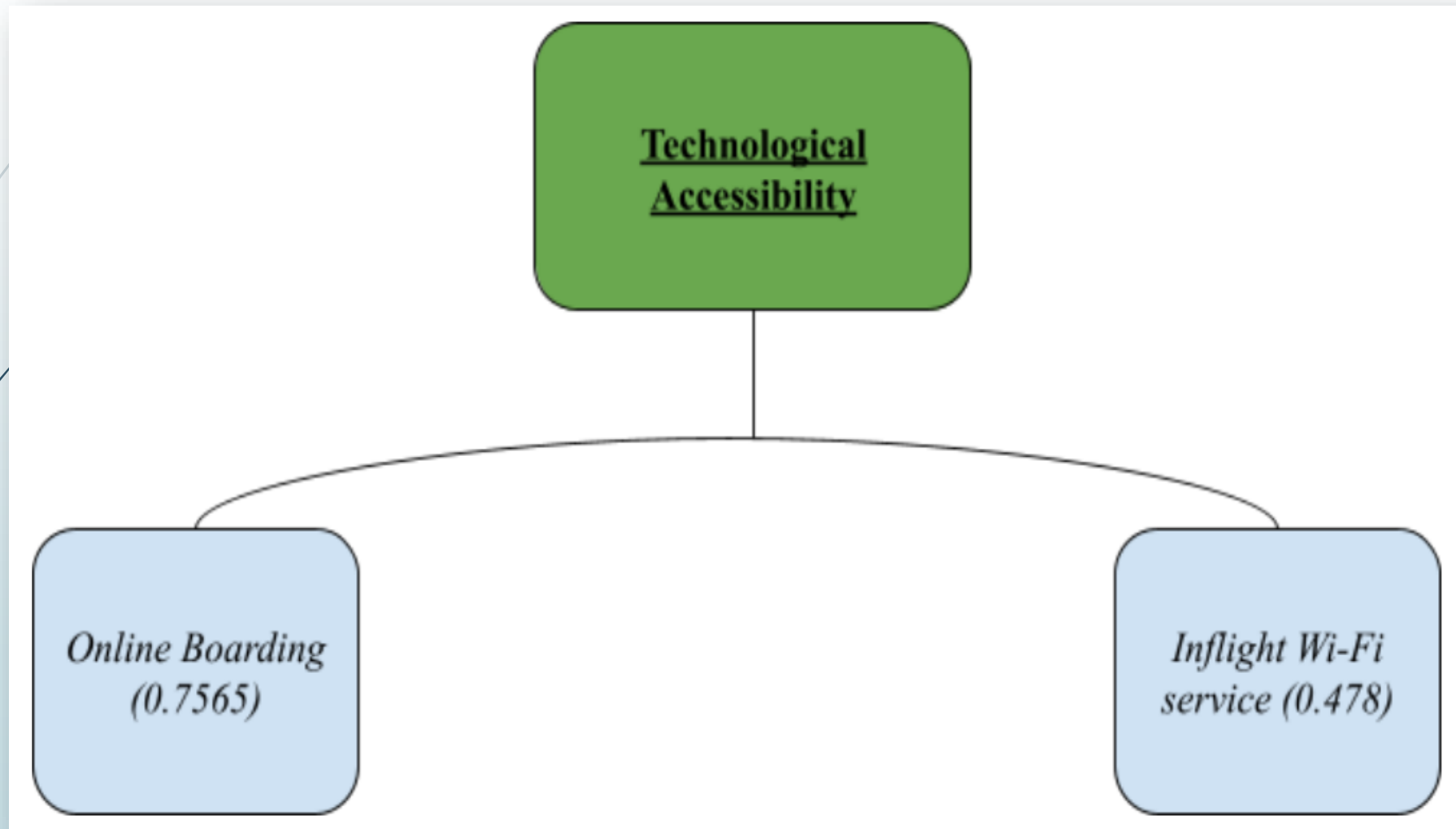
Factor Fitting



Factor Fitting



Factor Fitting

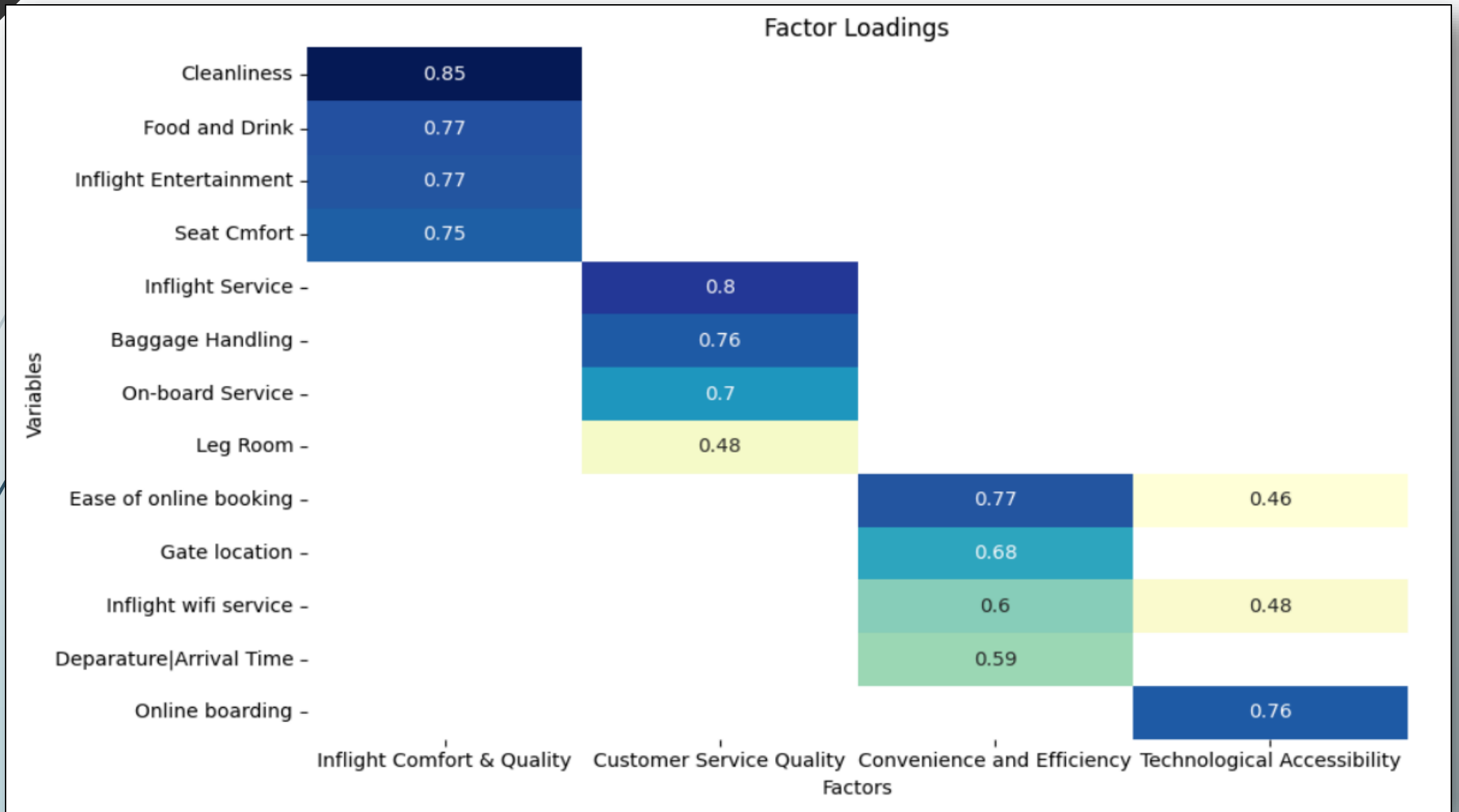


Communalities

	Communalities
Inflight wifi service	0.621384
Departure/Arrival time convenient	0.350986
Ease of Online booking	0.804333
Gate location	0.475724
Food and drink	0.595799
Online boarding	0.677361
Seat comfort	0.621287
Inflight entertainment	0.805802
On-board service	0.501142
Leg room service	0.247949
Baggage handling	0.587308
Checkin service	0.113370
Inflight service	0.645643
Cleanliness	0.747309

- **Communality** : is the amount of variance a variable shares with all the other variables being considered.

Factor Fitting



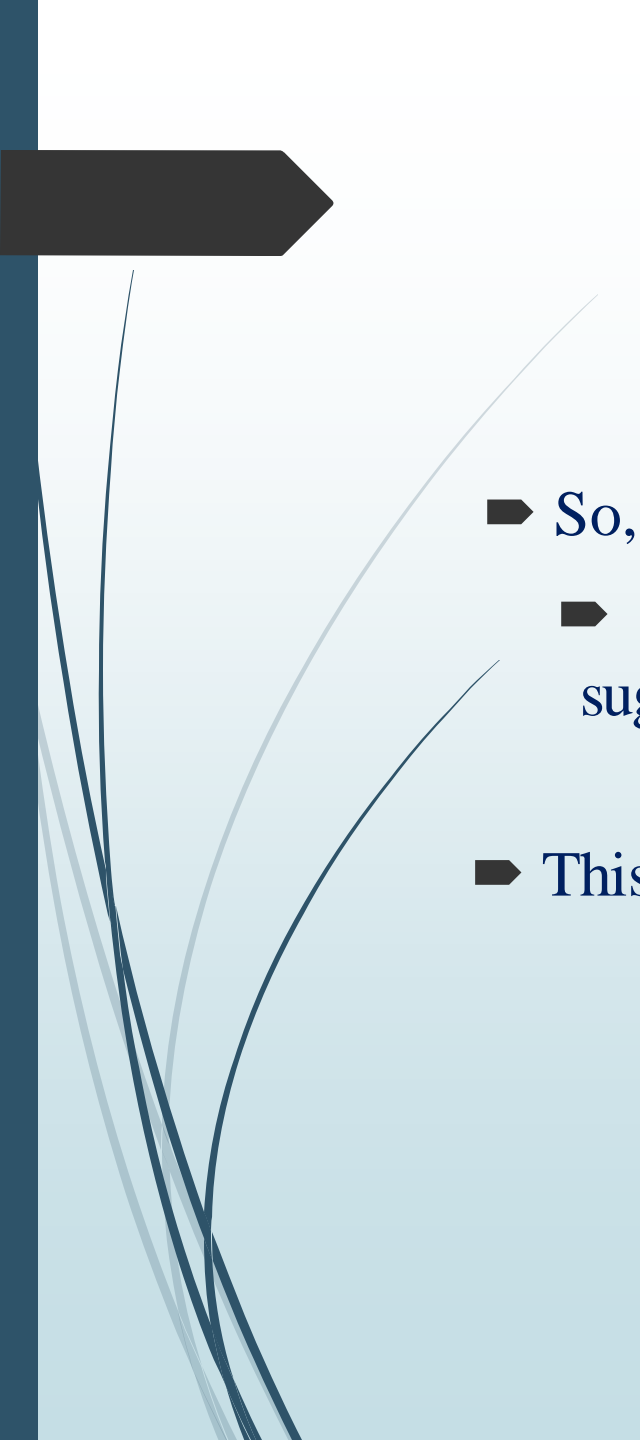
Logistic Regression

Logistic Regression is a statistical method used for analyzing a dataset in which there are one or more independent variables that determine an outcome. It's mainly used for binary classification problems, where the outcome is a categorical variable with two possible values, like yes/no or 1/0.

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i})}}$$

Where,

- p_i is the probability of the event occurring (e.g, the probability of being "Satisfied").
- β_0 is the intercept term.
- $\beta_1, \beta_2, \beta_3, \dots, \beta_k$ are the coefficients associated with each independent variable $X_1, X_2, X_3, \dots, X_k$

- 
- All the variables have $p\text{-value} < 0.05$
 - So, we Reject H_0 and conclude that the coefficient β_i is not equal to zero.
 - Hence, We reject H_0 , it typically indicates that there is evidence to suggest that the model coefficients are not equal to zero for at least one predictor variable.
 - This suggests that the predictor variables have some explanatory power in predicting the outcome variable.

K-MEANS

It aims to group data points into clusters based on their similarity, with the objective of minimizing the within-cluster variance.

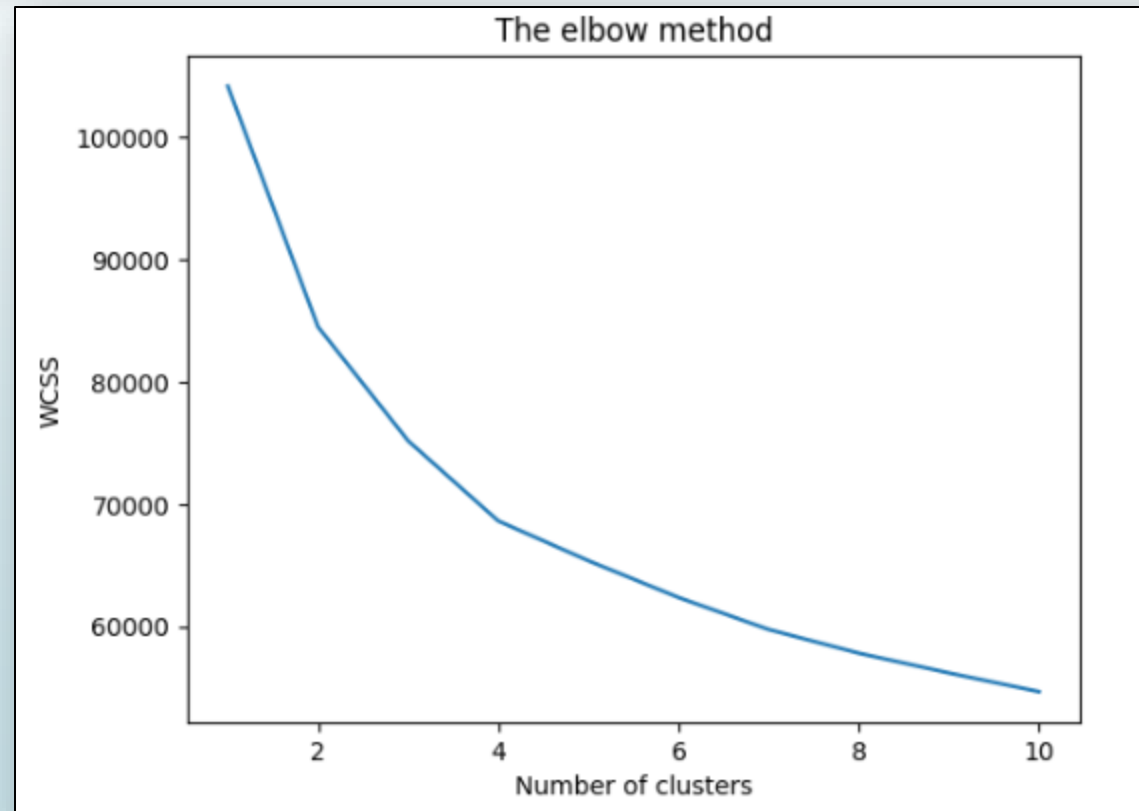
```
data_06 = data_02.iloc[:, 8:22]
data_06["Result"] = data_02['satisfaction']

X = data_06.drop(columns=["Result"])
y = data_06["Result"]

data_06
```

[illegible]

```
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter = 300, n_init = 10, random_state = 0)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.title('The elbow method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS') #within cluster sum of squares
plt.show()
```





```
kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=10, random_state=0)
```

```
kmeans.fit(X_scaled)
```

```
KMeans  
KMeans(n_clusters=4, n_init=10, random_state=0)
```

```
cluster_centers = kmeans.cluster_centers_
```

```
labels = kmeans.labels_
```

```
correct_labels = sum(y_encoded == labels)  
print("Result: %d out of %d samples were correctly labeled." % (correct_labels, y_encoded.size))
```

```
Result: 40398 out of 103904 samples were correctly labeled.
```

```
inertia = kmeans.inertia_
```

```
silhouette_score = metrics.silhouette_score(X_scaled, labels, metric='euclidean')
```



```
silhouette_score = metrics.silhouette_score(X_scaled, labels, metric='euclidean')
```

```
print("Silhouette Score:", silhouette_score)
```

```
Silhouette Score: 0.1500143336478502
```

```
print('Accuracy score: {0:0.2f}'.format(correct_labels/float(y.size)))
```

```
Accuracy score: 0.39
```

CONCLUSION

Factor Analysis

Principal Component Analysis (PCA), identifies 4 principal components that explain a significant portion of the variance, representing key dimensions of airline satisfaction.

- ☐ **Inflight Comfort & Quality**
- ☐ **Customer Service Quality**
- ☐ **Convenience and Efficiency**
- ☐ **Technological Accessibility**

CONCLUSION

Logistic regression

The coefficients obtained from the logistic regression model provide insights into the magnitude and direction of these effects.

Using the coefficients obtained from the logistic regression analysis, we can write the full logistic regression model as :

$$\log\left(\frac{p}{1-p}\right) = -6.6653 + 0.7408 x_1 + 0.7172 x_2 + 0.4942 x_3 + 0.9806 x_4$$

- p is the probability of customer satisfaction.
- x_1 represents "Inflight Comfort and Quality".
- x_2 represents "Customer Service Quality".
- x_3 represents "Convenience and Efficiency".
- x_4 represents "Technological Accessibility".

CONCLUSION

K-Means Clustering

- ❑ The clustering results, as measured by the silhouette score, suggest meaningful separation between clusters and potential actionable insights for targeted interventions.
- ❑ These clusters offer valuable insights into customer preferences and behavior, enabling targeted marketing strategies, personalized service offerings, and tailored interventions to enhance overall customer satisfaction and loyalty.
- ❑ By understanding the unique needs and preferences of each cluster, airlines can optimize resource allocation, improve customer engagement, and drive business growth.



THANK YOU