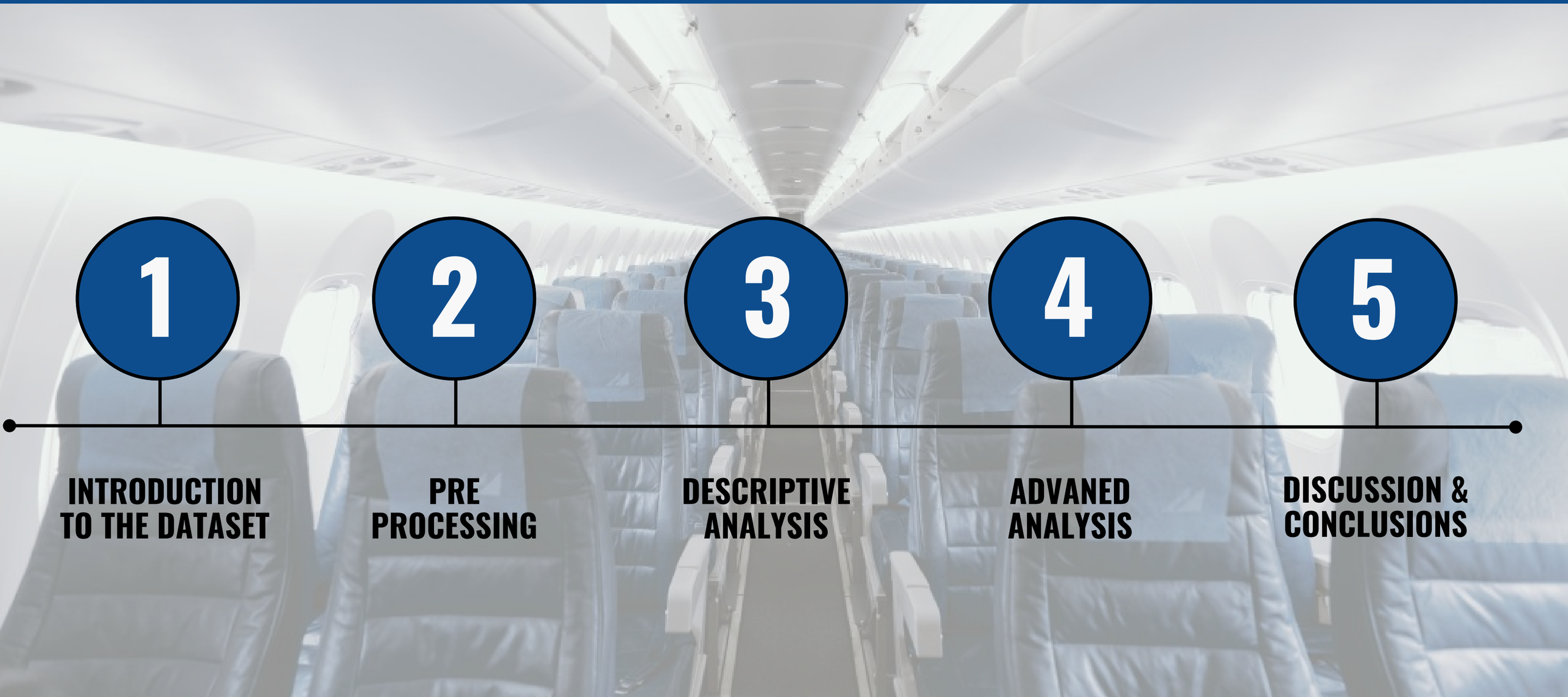
# Airline Passenger Satisfaction

# OBJECTIVES

- To gain a comprehensive understanding of the dataset's characteristics and underlying patterns by thoroughly exploring the dataset.

- To develop an accurate model to predict the satisfaction of an airline passenger, which will contribute in improving their services and encouraging higher levels of customer satisfaction.

# OVERVIEW



1. INTRODUCTION TO THE DATASET
2. PRE PROCESSING
3. DESCRIPTIVE ANALYSIS
4. ADVANED ANALYSIS
5. DISCUSSION & CONCLUSIONS

# About the Dataset

- The "Airline Passenger Satisfaction” Dataset is taken from the Kaggle website.
- It contains 103904 observations under 25 variables, out of which the response “satisfaction” is a categorical variable with two levels (‘Satisfied’ and ‘Neutral or dissatisfied’).

| Variable Name | Description |
|---|---|
| Gender | Gender of the passengers (Female, Male) |
| Customer Type | The customer type (Loyal customer, disloyal customer) |
| Age | The actual age of the passengers |
| Type of Travel | Type of Travel |
| Class | Travel class in the plane of the passengers (Business, Eco, Eco Plus) |
| Flight distance | The flight distance of this journey |
| Inflight wifi service | Satisfaction level of the inflight wifi service (0: Not Applicable;1-5) |
| Departure/Arrival time convenient | Satisfaction level of Departure/Arrival time convenient |
| Ease of Online booking | Satisfaction level of online booking |
| Gate location | Satisfaction level of Gate location |
| Food and drink | Satisfaction level of Food and drink |
| Online boarding | Satisfaction level of online boarding |
| Seat comfort | Satisfaction level of Seat comfort |
| Inflight entertainment | Satisfaction level of inflight entertainment |
| On-board service | Satisfaction level of On-board service |
| Leg room service | Satisfaction level of Leg room service |
| Baggage handling | Satisfaction level of baggage handling |
| Check-in service | Satisfaction level of Check-in service |
| Inflight service | Satisfaction level of inflight service |
| Cleanliness | Satisfaction level of Cleanliness |
| Departure Delay in Minutes | Departure Delay in Minutes |
| Arrival Delay in Minutes | Minutes delayed when Arrival |
| Satisfaction | Airline satisfaction level (Satisfied, ‘neutral or dissatisfied’) |

# Data Preprocessing

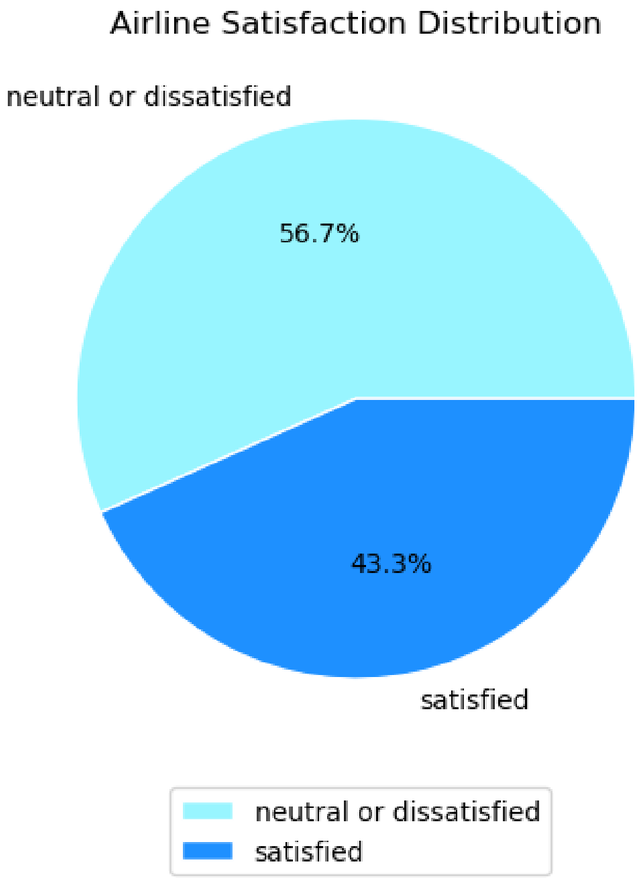
- It was observed that the presence of zero values, inconsistent with the ordinal scale ranging from 1 to 5. They were treated as an anomaly and replaced with 1, which represent the lowest level of satisfaction. This was done to ensure that the dataset aligns with the intended satisfaction ratings scale.

- There were 310 missing values in the variable, “Arrival Delay in Minutes” and they were imputed by using the median.

# Descriptive Analysis

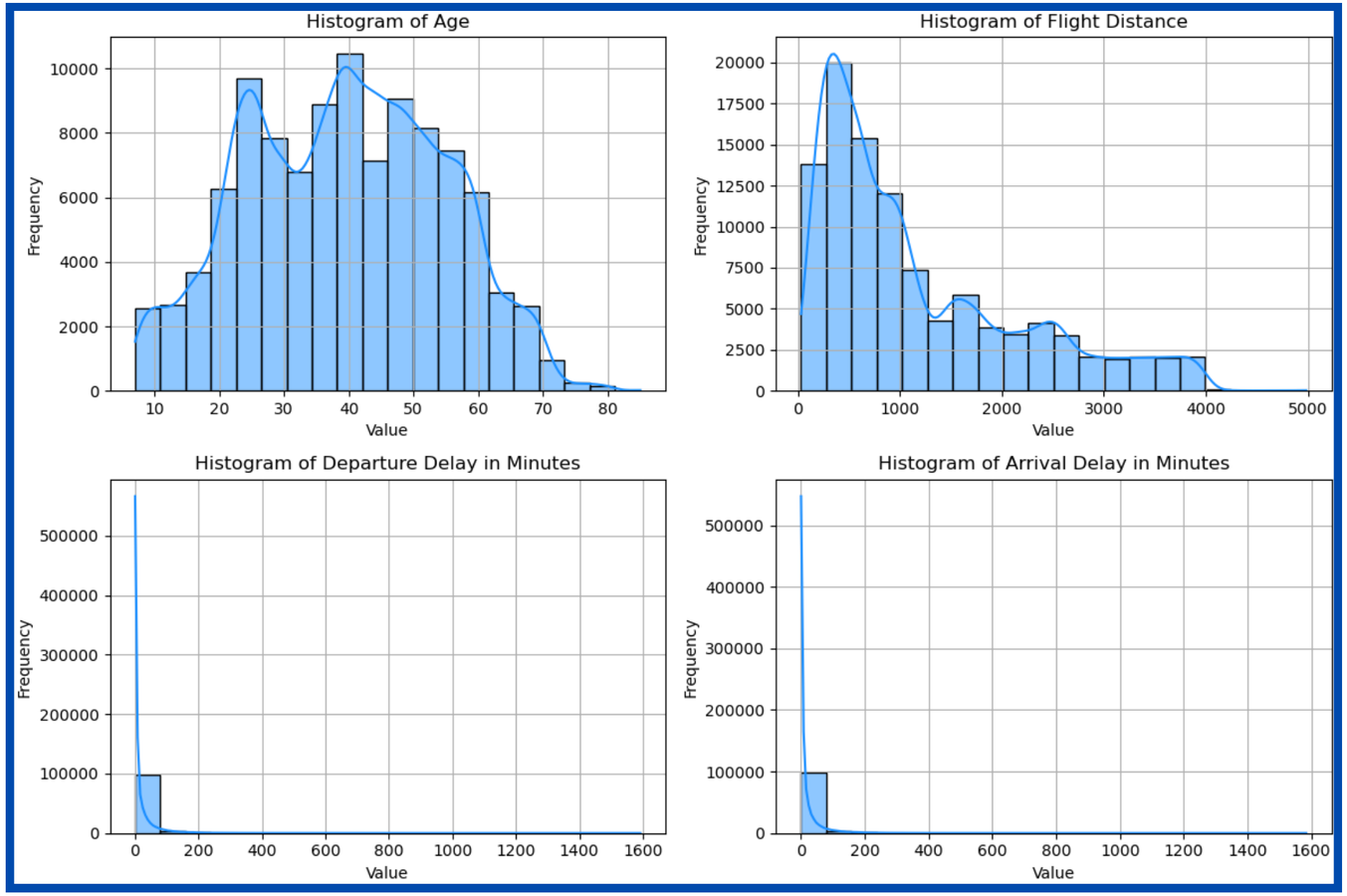
# Univariate Analysis

**Response Variable :**
**" Satisfaction"**




- **According to the 2019 Airline Satisfaction Survey conducted by Skytrax, the average percentage of satisfied air passengers in the world was 80%.**
- **Our data set does not represent the real-world scenario.**
- **This may lead to biased results in predictions.**

# Univariate Analysis




**The distribution of Age is approximately normal.**

**The Flight Distance distribution skews to the right**

"The distribution of flight distances traveled by air passengers is typically right-skewed.
The median flight distance is typically around 500-1000 kilometers. "

**- International Air Transport Association (IATA) and the Federal Aviation Administration (FAA) -**

**The departure and arrival delays exhibit a strong right skew in their distribution.**

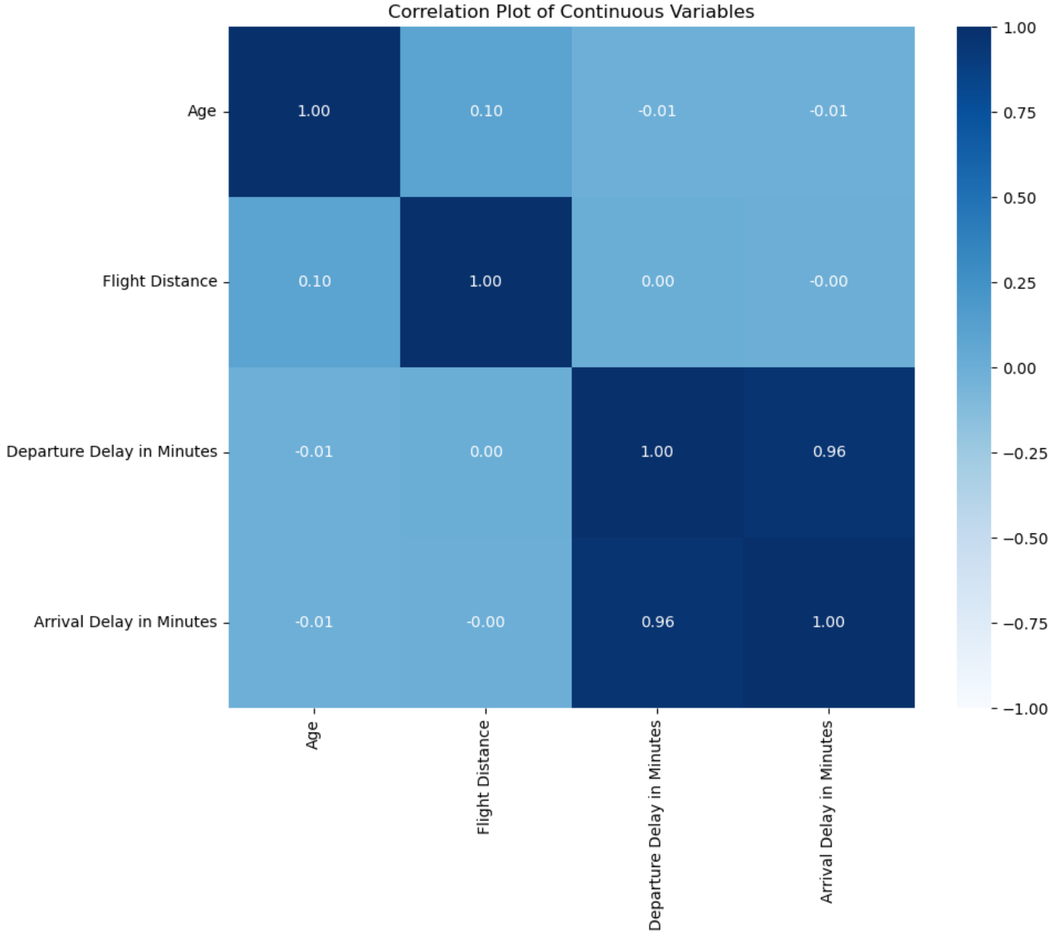
"The distribution of departure and arrival delays typically follows a right-skewed pattern, with most flights experiencing minimal delays and a smaller proportion encountering significant delays"

**- International Air Transport Association (IATA) and the Federal Aviation Administration (FAA) -**

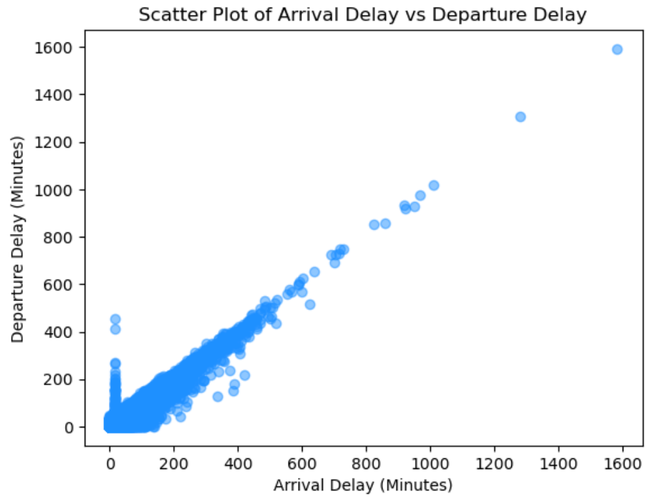
# Bivariate Analysis

## Correlation Between Continuous Variables

## Pearson Correlation




- A strong correlation of 0.96 exists between Departure Delay in Minutes and Arrival Delay in Minutes, which is expected.
  → A delay in departure time directly translates to a similar delay upon arrival.
- This relationship is visually apparent in the scatter plot of Departure Delay in Minutes versus Arrival Delay in Minutes.

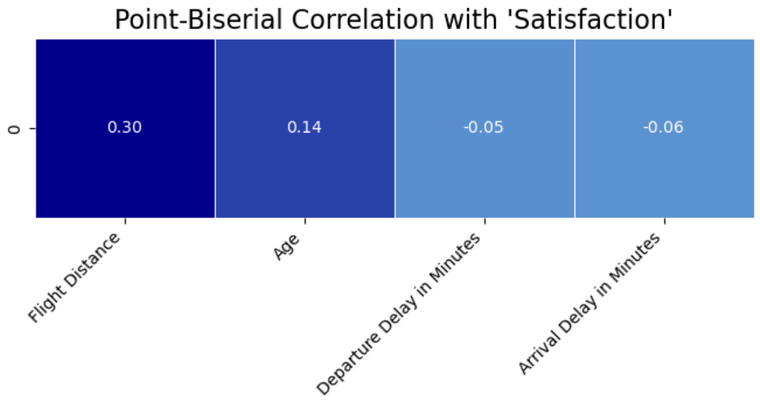



- All other variables display weaker correlations with each other

# Bivariate Analysis

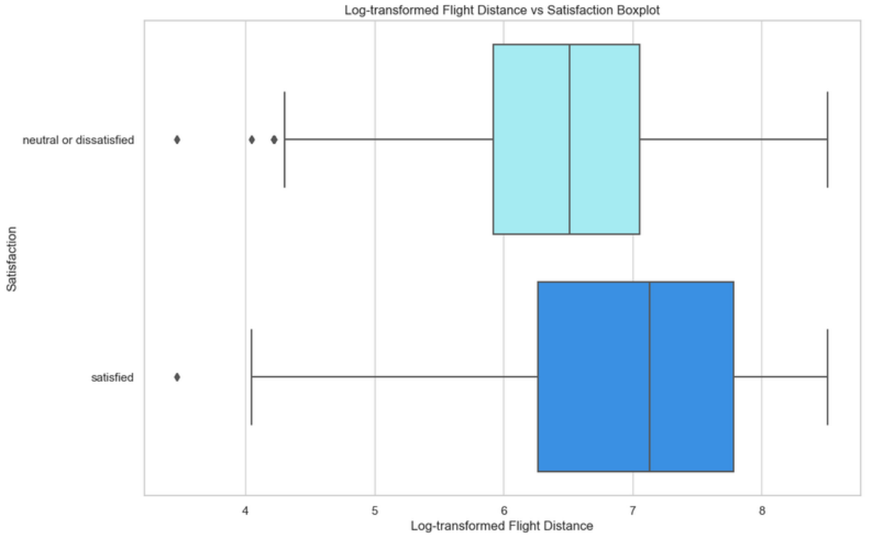
## Correlation Between Binary Response Variable and Continuous Variables

## Point-Biserial Correlation



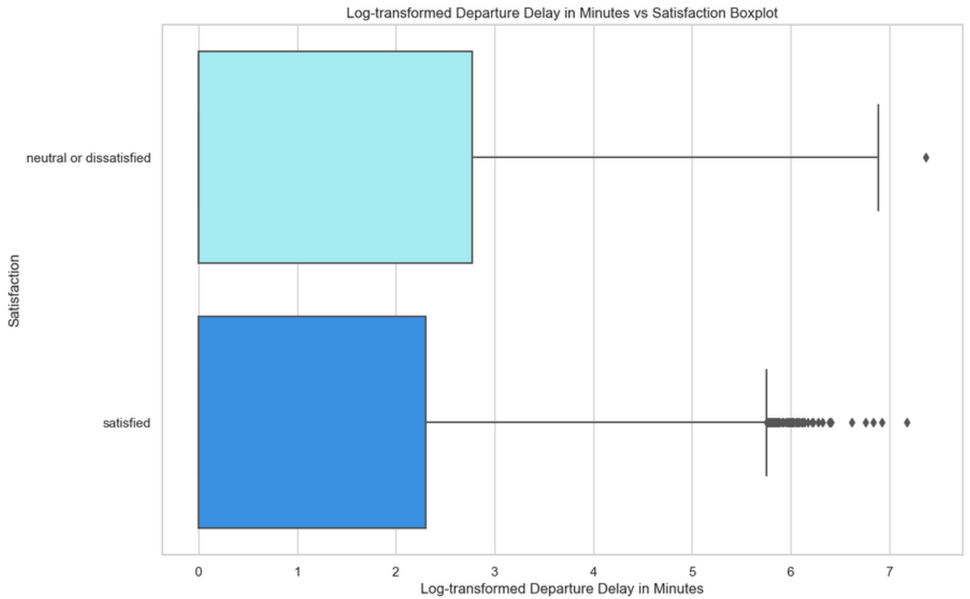

There is a positive correlation between flight distance and satisfaction, meaning that as flight distance increases, passenger satisfaction tends to increase as well.

**2019 Skytrax Survey support this, showing happier travelers on longer journeys**



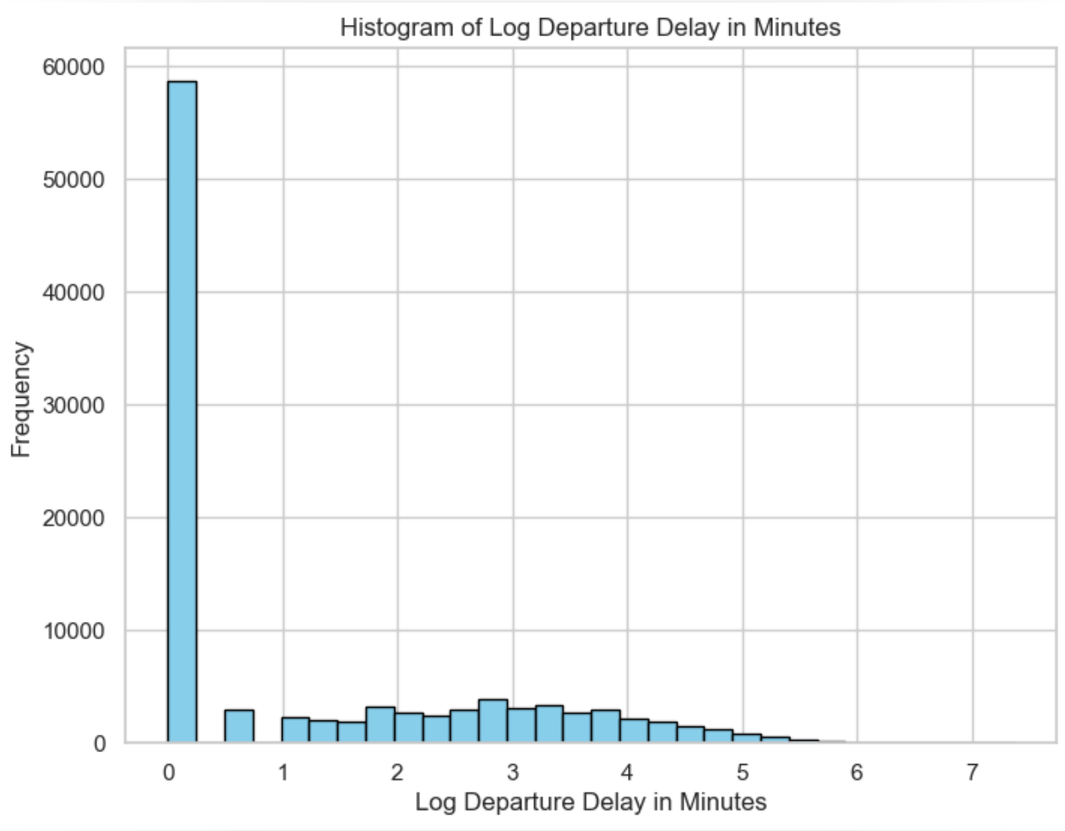

- Longer flights bring excitement with new experiences, better amenities like entertainment, and a sense of achievement.
- Travelers often see longer flights as more value for money due to the extended journey.

# Departure Delay in Minutes




**2022 J.D. Power North American Airline Satisfaction Study:** This study found that departure delays were among the top factors contributing to passenger dissatisfaction.

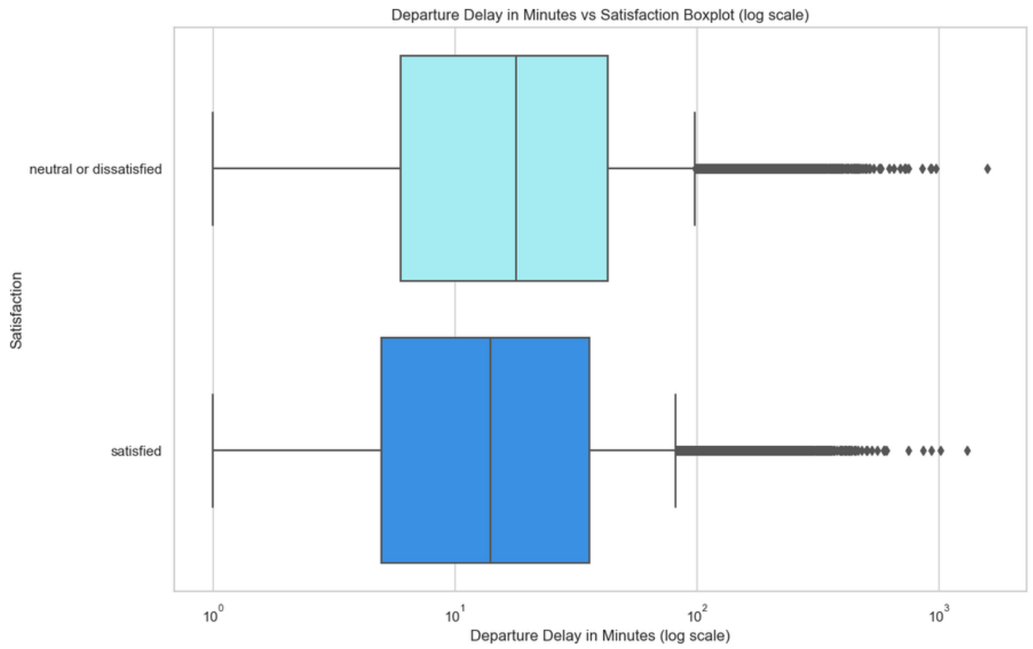
However, without a noticeable variance in the median departure delays between these categories, it's challenging to definitively state that higher departure delays consistently lead to more dissatisfaction.



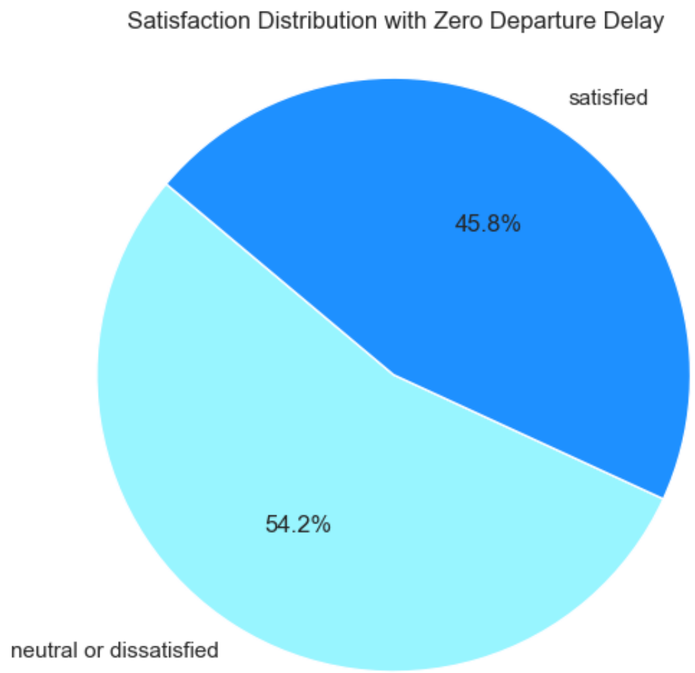

- Despite applying a log transformation, the departure delay distribution remains skewed due to the large number of zero delays.

- To study the association between departure delay and satisfaction, we can split the dataset into two subsets: those with non-zero delays and those with zero delays."

# Non zero group




# Zero group



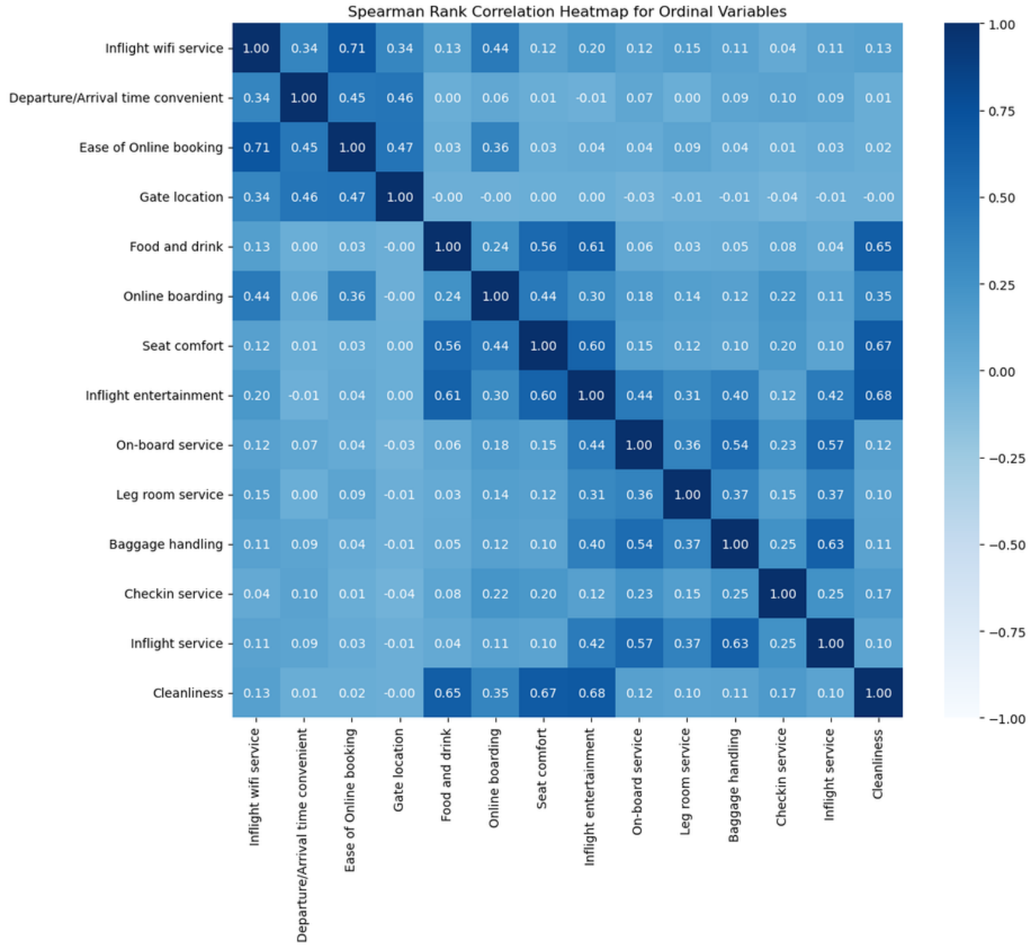

Looking at the box plot for non-zero departure delays, it is evident that the median departure delay is higher for non-satisfied passengers compared to other groups, as previous researchers have noted.

Despite having zero departure delays, a higher percentage of passengers still fall into the dissatisfied group. This suggests that other factors, strongly associated with customer satisfaction, are effecting these results.

# Bivariate Analysis

## Correlation Between Ordinal Variables

## Spearman Rank Correlation



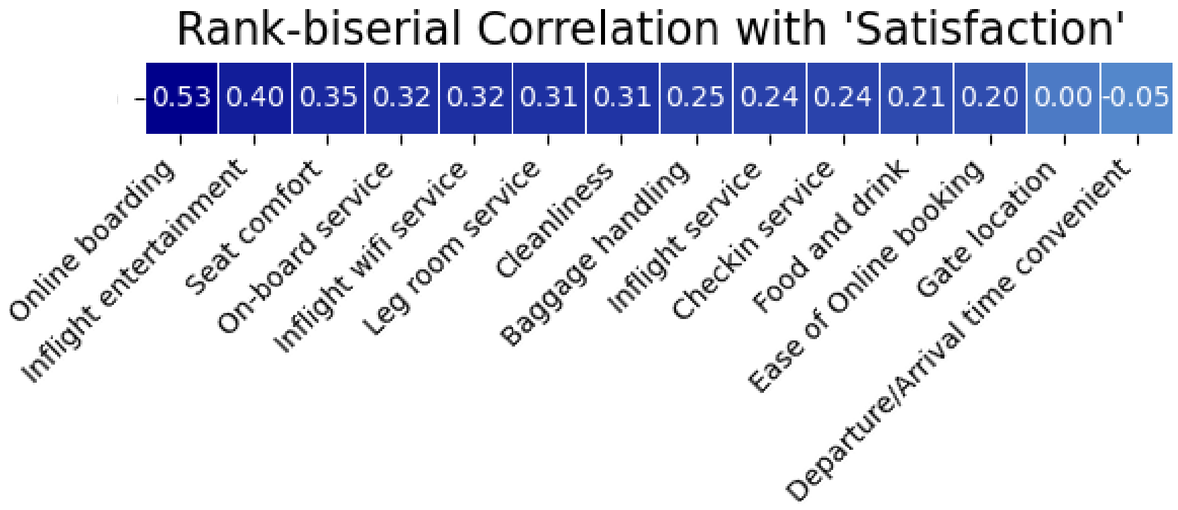

Ease of Online Booking, Inflight WiFi Service, Cleanliness, Inflight Entertainment, Seat Comfort, Food and Drink, Inflight Service, and Baggage Handling exhibit correlations as they collectively contribute to elevating the overall passenger experience during air travel.

These factors emphasize convenience, comfort, and service quality,  enhancing the journey for travelers

# Bivariate Analysis

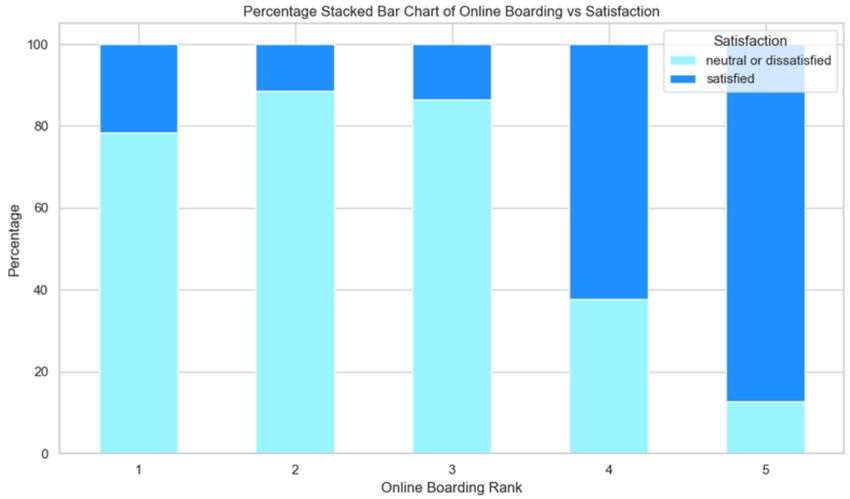
## Correlation Between Binary Response Variable and Ordinal Variables

## Rank-Biserial Correlation




Online boarding and inflight entertainment have been shown to have a higher correlation with passenger satisfaction compared to other ordinal variables.

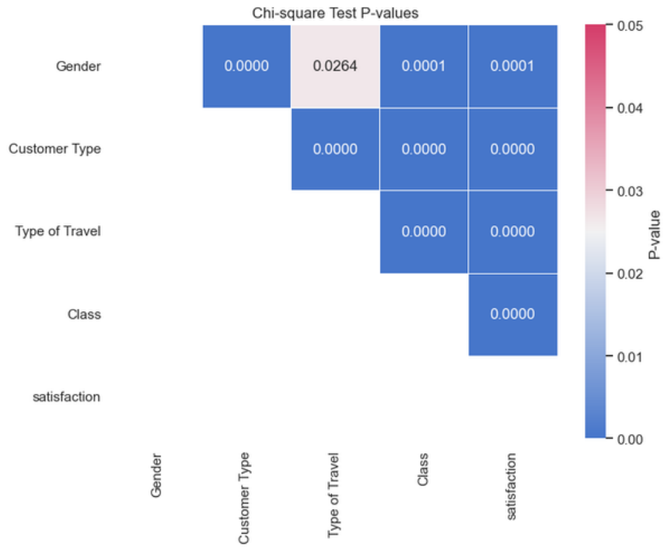
-Shafiei, A., et al. (2017). Passenger satisfaction in airline service: A study of the role of service quality and passenger characteristics. Journal of Hospitality and Tourism Management-

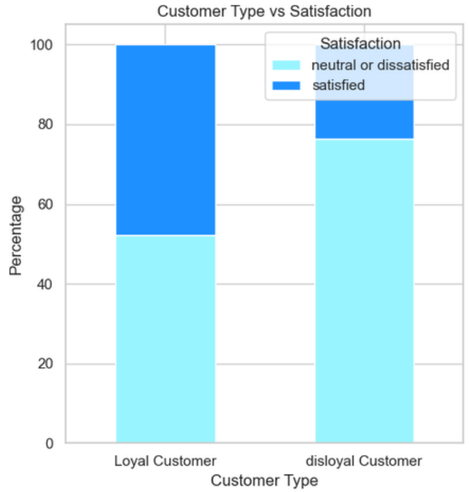


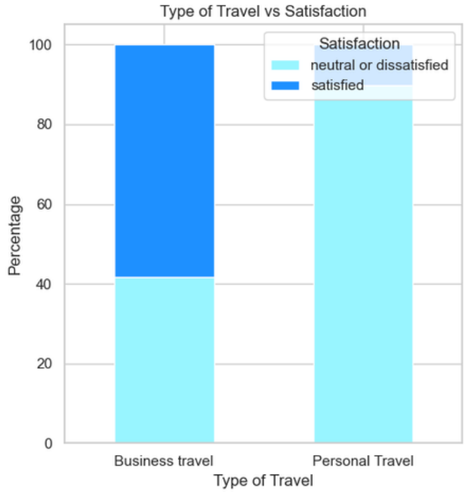
# Bivariate Analysis

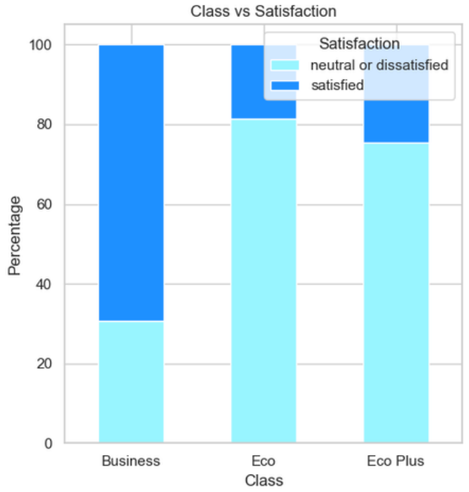
## Correlation Between Binary Response Variable and Categorical variables

## Chi-square Test



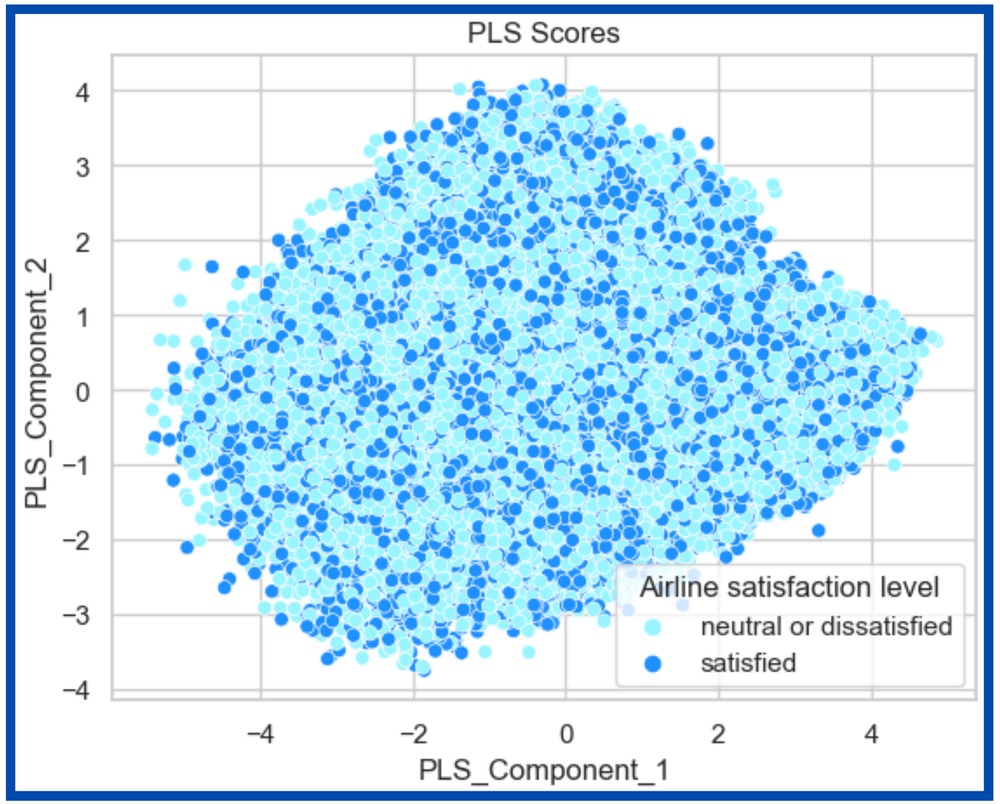

P-value < 0.05
Significant association between two variables

Factors such as customer type, type of travel, and class all have a significant impact on passenger satisfaction.

**A Study of the Role of Service Quality and Passenger Characteristics" by Afsaneh Shafiei et al. (2017)**

# Further Analysis

# Partial Least Squares Analysis



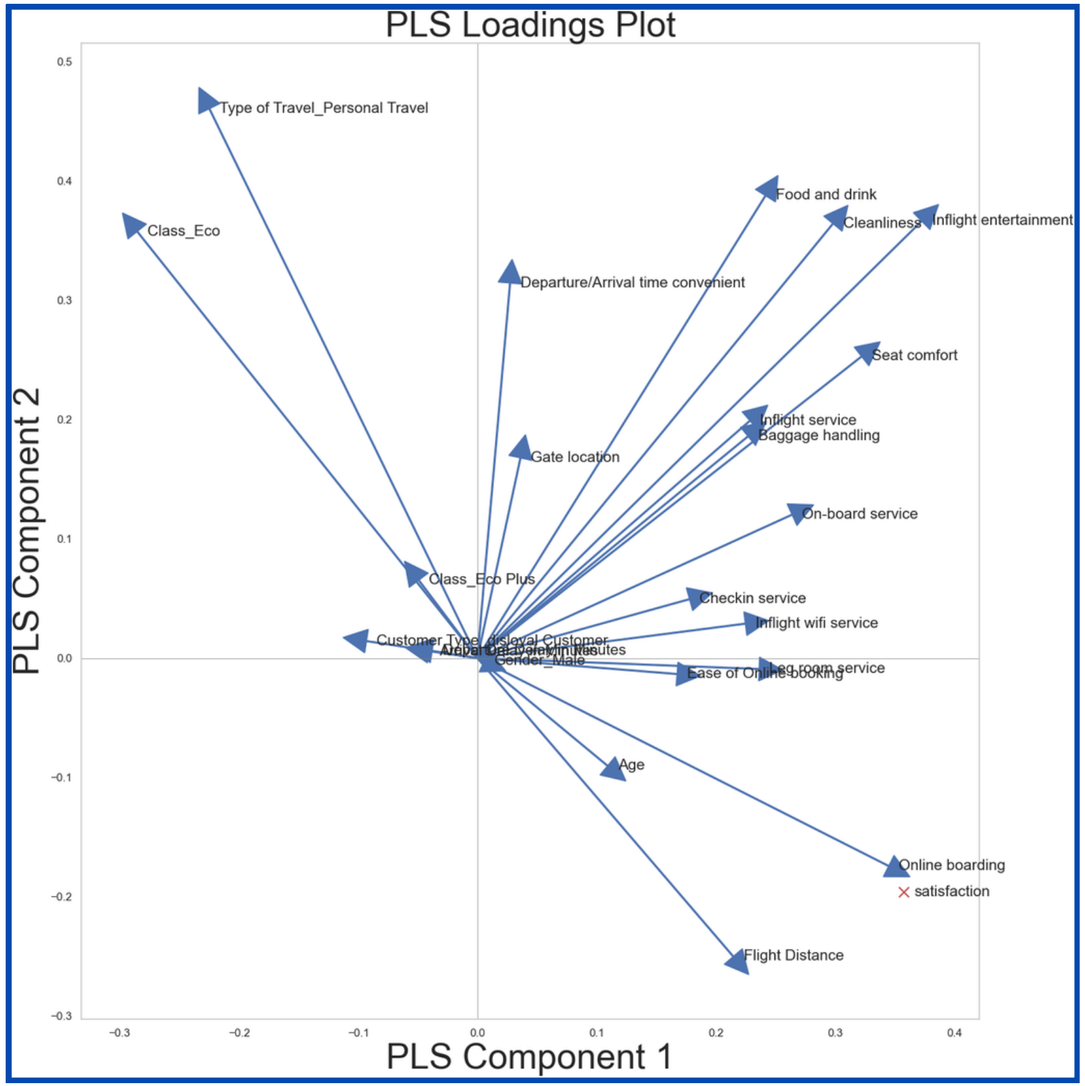

- **24.23% of the variance is explained by the first two PLS components.**
- **The observations in the score plot were colored with response variable**
- **No clear separation between the two classes, therefore, linear classification algorithms such as LDA cannot be used.**
- **the use of other techniques such as binary logistic regression, and support vector machines (SVM), may work well when predicting passenger satisfaction.**

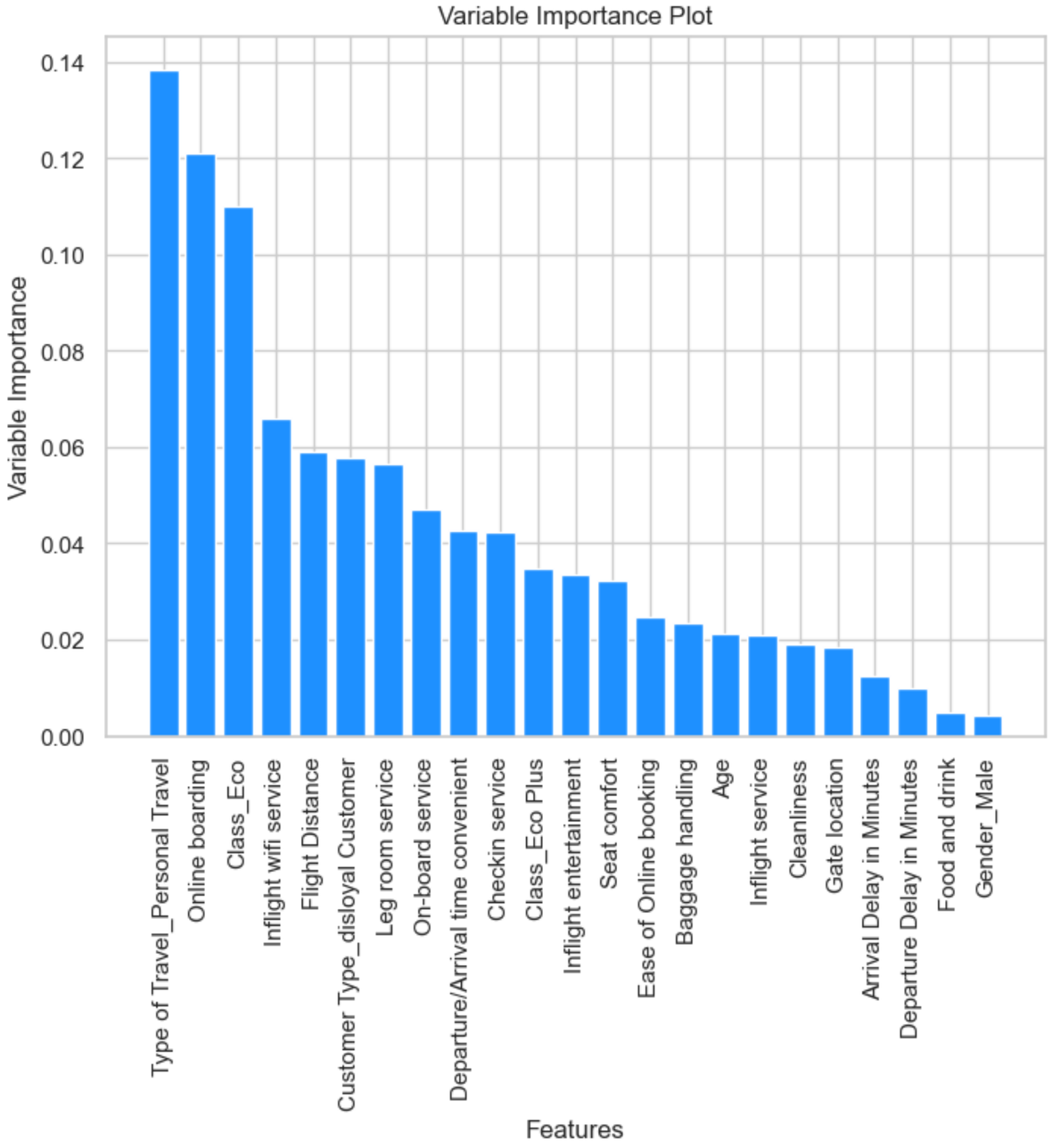
# Partial Least Squares Analysis

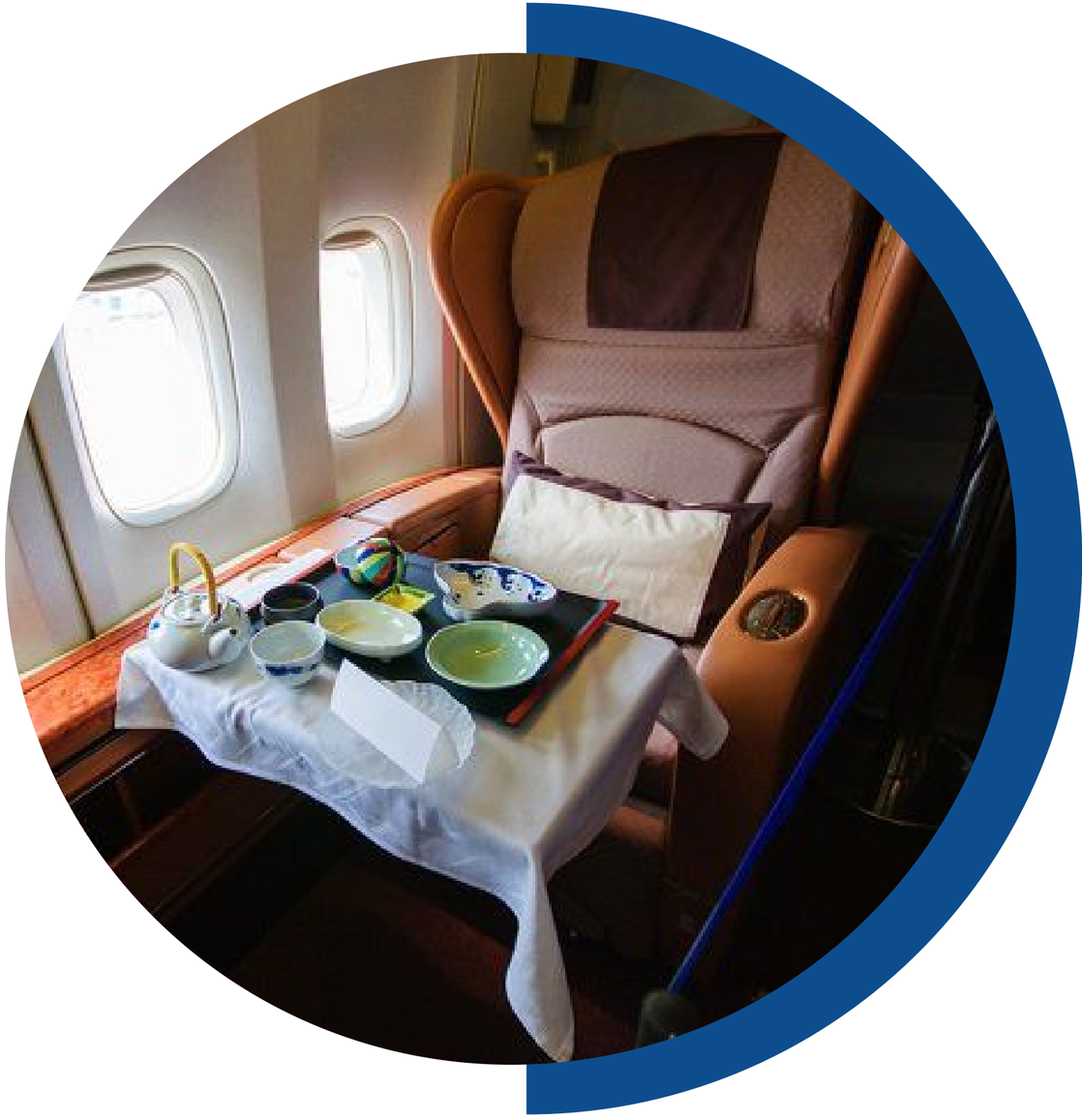
**Predictors which show a significant association with satisfaction**

- Age
- Online Boarding
- Flight Distance
- Type of Travel_PersonalTravel (negatively)

The loadings plot suggests potential variable clusters with closely positioned predictor points indicating strong correlations.
High predictor correlation might lead to multicollinearity issues in logistic regression models, possibly requiring the dimensionality reduction techniques for model improvement.




Variable Importance Plot
Variable Importance
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Type of Travel_Personal Travel
Online boarding
Class_Eco
Inflight wifi service
Flight Distance
Customer Type_disloyal Customer
Leg room service
On-board service
Departure/Arrival time convenient
Checkin service
Class_Eco Plus
Inflight entertainment
Seat comfort
Ease of Online booking
Baggage handling
Age
Inflight service
Cleanliness
Gate location
Arrival Delay in Minutes
Departure Delay in Minutes
Food and drink
Gender_Male
Features



# Important Results of the Advanced Analysis

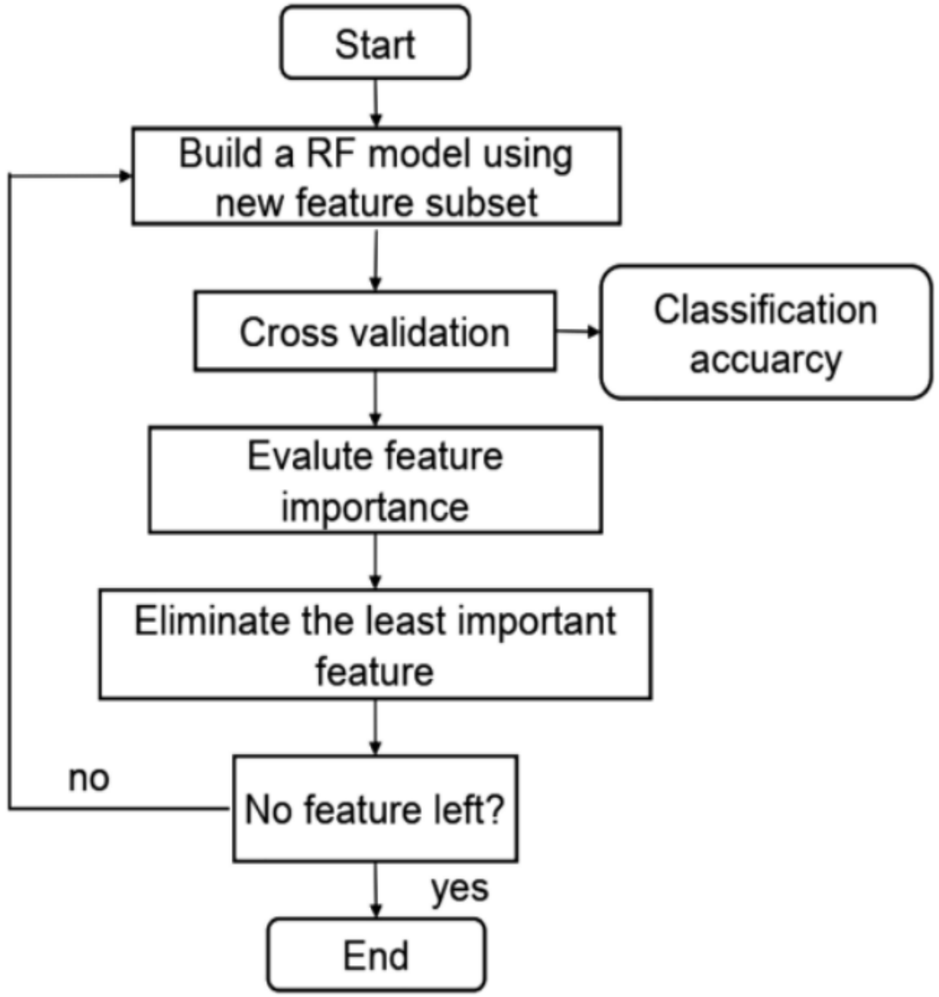
# RF-RFE

- **Recursive Feature Elimination based on Random Forest (RF-RFE)** was used to identify a subset of the most informative features.

- RF-RFE is a potent approach within statistical learning, blending the ensemble power of Random Forest with iterative feature selection.

- It iteratively refines the feature subset by systematically eliminating the least informative features based on their importance scores. Through cross-validation at each iteration, the methodology ensures robust evaluation, leading to a focused set of features that contribute most to model interpretability and predictive performance.

- This technique provides a robust means of identifying the most impactful features for classification accuracy.




*RF-RFE method*

# Logistic Regression

- Two scenarios were considered:
  - a logistic regression model with a ridge penalty was fitted using the features passively selected by RF-RFE
  - a separate model was fitted using all the features initially considered

| Using RF-RFE feature selection | | Without using RF-RFE feature selection | |
|---|---|---|---|
| | Accuracy | | Accuracy |
| Training Set | 0.89 | Training Set | 0.92 |
| Test Set | 0.89 | Test Set | 0.92 |

- As seen from the table above, the full model performs better than the reduced model. Nevertheless, adherence to the parsimony principle prompts a preference for the reduced model, given its only marginal sacrifice in performance.

# KNN Classifier and SVC

- A K-Nearest Neighbors (KNN) classifier was employed; however, its performance fell considerably short when compared to the preceding logistic regression model.

| Using RF-RFE feature selection | | Without using RF-RFE feature selection | |
|---|---|---|---|
| | Accuracy | | Accuracy |
| Training Set | 0.86 | Training Set | 0.80 |
| Test Set | 0.86 | Test Set | 0.70 |

- A Support Vector Classifier was employed. The results were as follows:

| Using RF-RFE feature selection | | Without using RF-RFE feature selection | |
|---|---|---|---|
| | Accuracy | | Accuracy |
| Training Set | 0.87 | Training Set | 0.88 |
| Test Set | 0.86 | Test Set | 0.88 |

# Random Forest Classifier

- Hyperparameter tuning on the Random Forest model using a randomizedCV grid search strategy, and the results were promising.
- Here, two cases were considered. The full model and the reduced model.

| With All Features + hyperparameter tuning | | With only the important features (10) + hyperparameter tuning | |
|---|---|---|---|
| 'n_estimators':10<br>'max_depth':30<br>'min_samples_split':5<br>'min_samples_leaf':4<br>'max_features': None | | 'n_estimators':10<br>'max_depth':30<br>'min_samples_split':5<br>'min_samples_leaf':2<br>'max_features': None | |
| | Accuracy | | Accuracy |
| Training Set | 0.90 | Training Set | 0.93 |
| Test Set | 0.89 | Test Set | 0.93 |

- Notable improvement in overall performance for the [reduced + hyperparameter tuned] RF model.

# Voting classifier

- A Voting Classifier is constructed by combining the predictions of four different classification models: Random Forest, Logistic regression, XG Boost, and K-Nearest Neighbors (KNN).

- Hard Voting was used. In hard voting, each individual classifier in the ensemble "votes" for a class, and the class that receives the majority of votes is selected as the final prediction.

- The outcomes indicate that the Voting Classifier model yielded lower performance when compared to the hyperparameter-tuned Random Forest model.

| | Accuracy |
|---|---|
| Training set | 0.90 |
| Testing set | 0.88 |

# Thus, the Best Model ...

- Conclude that the best model out of the models that we have fitted is:

**The hyperparameter-tuned random forest model with only the most important variables**

| With only the important features (10) + hyperparameter tuning | |
|---|---|
| 'n_estimators':10<br>'max_depth':30<br>'min_samples_split':5<br>'min_samples_leaf':2<br>'max_features': None | |
| | Accuracy |
| Training Set | 0.93 |
| Test Set | 0.93 |

# Model Interpretation

- By uncovering the features and decision-making processes within the model, model interpretation allows airlines to identify specific service aspects that significantly impact passenger satisfaction.

- This knowledge empowers airlines to make informed enhancements, ensuring a more tailored and satisfactory travel experience for their passengers.

# Global Model-Agnostic Methods

- An essential aspect of our predictive model's interpretability is the examination of feature importance.

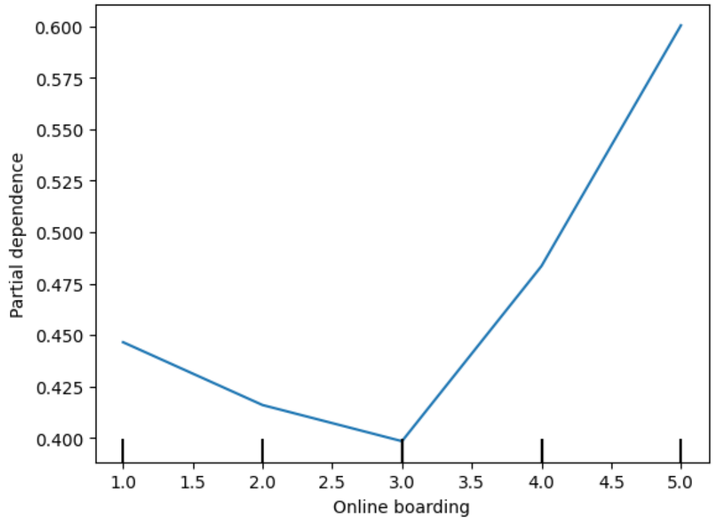
- The Random Forest model highlights that the Online_boarding, Personal Travel, Class_Economy, Inflight_wifiservice are the top four important variables.

- The top features identified by the Random Forest model all contribute to an enhanced passenger experience.
  - Online boarding streamlines the pre-flight process,
  - personal travel considerations tailor services to individual needs,
  - economy class satisfaction caters to a significant passenger demographic, and
  - reliable inflight Wi-Fi keeps passengers connected and entertained.

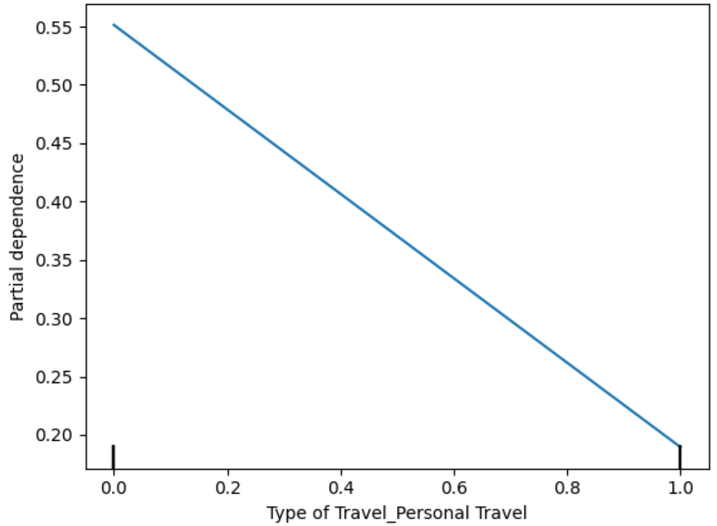
- Airlines that prioritize these features are likely adopting a customer-centric approach and are more likely to receive positive reviews, recommendations, and build a strong brand image.

| feature | importance |
|---|---|
| Online boarding | 0.213679 |
| Type of Travel_Personal Travel | 0.161772 |
| Class_Eco | 0.136956 |
| Inflight wifi service | 0.119648 |
| Inflight entertainment | 0.076509 |
| Seat comfort | 0.051292 |
| Leg room service | 0.040596 |
| On-board service | 0.031635 |
| Customer Type_disloyal Customer | 0.027649 |
| Ease of Online booking | 0.026662 |
| Cleanliness | 0.022469 |
| Flight Distance | 0.020639 |

# Global Model-Agnostic Methods

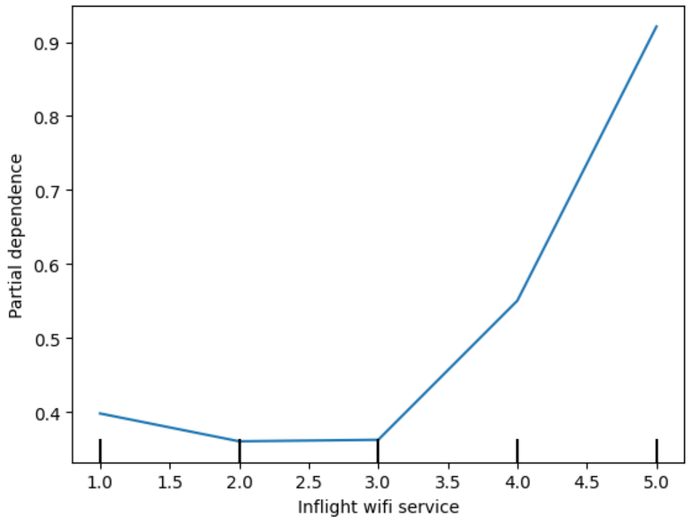



- The predicted probability slightly decreased as the rating increased from 1 to 3. This suggests that, within this range, there may be factors or interactions influencing the online boarding experience.
- The predicted probability then increased significantly as the rating increased from 3 to 5 as expected, because better the online boarding experience, better the satisfaction would be.




- Higher Value when Type_of_travel_Personal is 0, and Very Low Value when Type_of_travel_Personal is 1:
  - This could be because business travelers might have different expectations, priorities, or experiences.
  - Airlines could consider tailoring services, communication, or features to better meet the expectations and preferences of business travelers, as indicated by the higher predicted probability.

# Global Model-Agnostic Methods



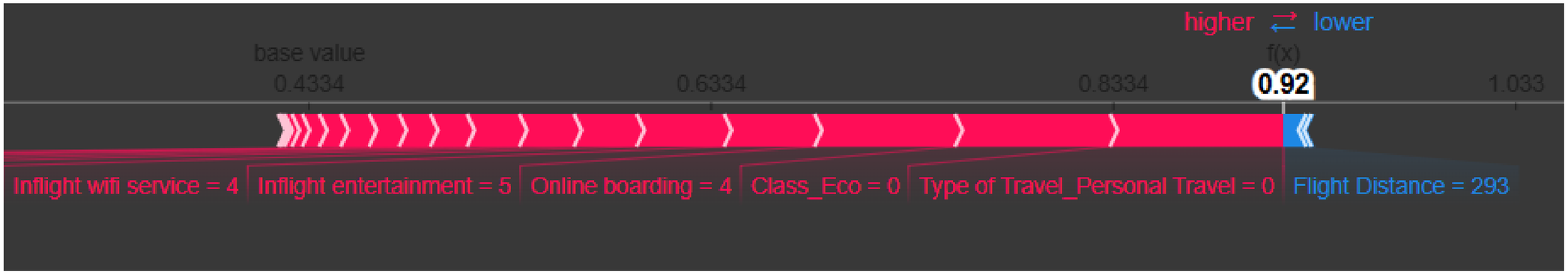

- The increase in the predicted probability as the rating of inflight wifi service moves from 3 to 5 suggests that passengers who rate the inflight wifi service more favorably (closer to 5) are associated with a higher likelihood of satisfaction or positive outcomes.

- Airlines may consider investing in and improving their inflight wifi services to meet or exceed passenger expectations.

# Local Model-Agnostic Methods

- Local interpretation methods explain individual predictions.
- SHAP (SHapley Additive exPlanations) is a method to explain individual predictions.
- TreeSHAP, a variant of SHAP for tree-based machine learning models is used in this analysis.

- The below force plot is for an observation where: Online boarding rating is 4, Class is not Economy, The inflight entertainment rating is 5, the inflight wifi service is 4.
- The random forest model predicts the outcome as satisfied with a probability of 0.92




- from the base value of 0.4334, the predicted probability has been increased due to the forces such as high ratings for inflight_wifi_service, Online boarding, etc

# Issues Encountered and Proposed solutions

- Considering that the initial dataset comprised 14 ordinal variables and most of these variables exhibited correlations, conducting a **factor analysis** could be more effective than just doing a feature elimination.

# Thank You!