# Airline Passenger Satisfaction Prediction

*Sheila Fitria Al'asqalani*

# Outline

**Step 1**

Business & Data Understanding

**Step 2**

Data Preparation

Data Preprocessing

**Step 3**

Data Analysis

**Step 4**

Machine Learning Model

**Step 5**

Conclusion & Business Recommendation
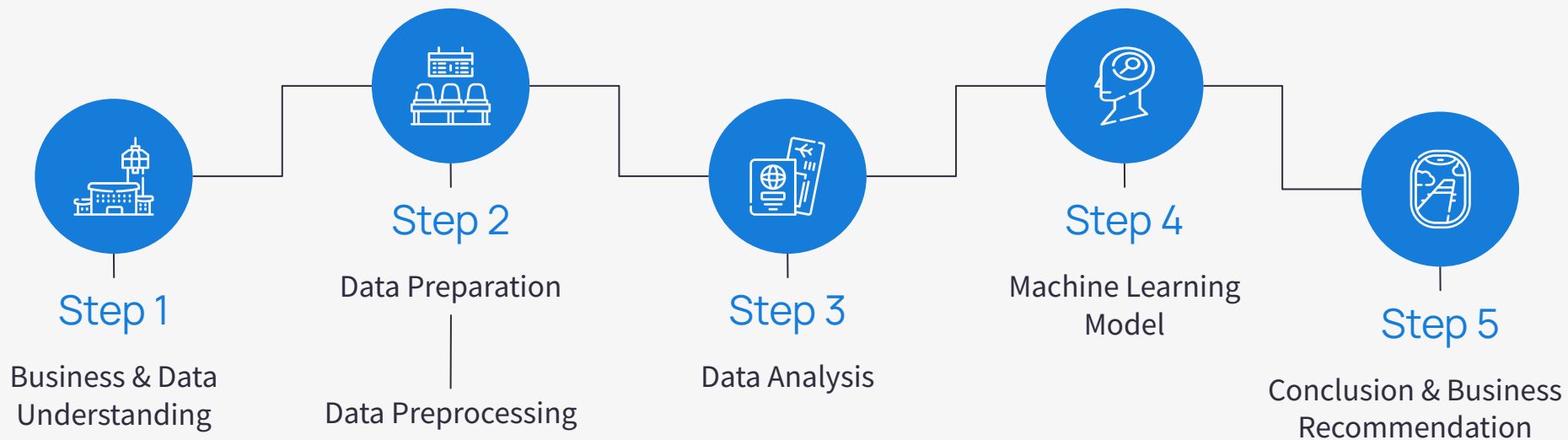
# Business Understanding

Predicting passenger satisfaction is important for several reasons:

**Customer Retention**. Satisfied customers are more likely to return and recommend the service to others, leading to increased business and revenue.

**Competitive Advantage.** Differentiate themselves from their competitors and attract new customers.

**Cost Reduction.** identify areas where customer experience can be improved, and take steps to address them before they become major problems. This can help reduce costs associated with customer complaints, negative reviews, and lost business.

**Reputation Management.** Positive customer experiences can lead to positive reviews, while negative experiences can lead to negative reviews and damage to the company's reputation

**Operational Efficiency.** By identifying areas of the travel experience that are most important to customers, companies can prioritize resources and make improvements that lead to increased efficiency and customer satisfaction.

# Dataset

Consist of customer satisfaction scores from 120,000+ airline passengers including additional information of the flight.

**129K**

Rows

**23**

Features

Encoded to

**28**

Features

*Kaggle Source*

# Data Understanding



```
#   Column                                   Non-Null Count     Dtype
---  ------                                   --------------     -----
0    ID                                       129880 non-null    int64
1    Gender                                   129880 non-null    object
2    Age                                      129880 non-null    int64
3    Customer Type                            129880 non-null    object
4    Type of Travel                           129880 non-null    object
5    Class                                    129880 non-null    object
6    Flight Distance                          129880 non-null    int64
7    Departure Delay                          129880 non-null    int64
8    Arrival Delay                            129487 non-null    float64
9    Departure and Arrival Time Convenience   129880 non-null    int64
10   Ease of Online Booking                   129880 non-null    int64
11   Check-in Service                         129880 non-null    int64
12   Online Boarding                          129880 non-null    int64
13   Gate Location                            129880 non-null    int64
14   On-board Service                         129880 non-null    int64
15   Seat Comfort                             129880 non-null    int64
16   Leg Room Service                         129880 non-null    int64
17   Cleanliness                              129880 non-null    int64
18   Food and Drink                           129880 non-null    int64
19   In-flight Service                        129880 non-null    int64
20   In-flight Wifi Service                   129880 non-null    int64
21   In-flight Entertainment                  129880 non-null    int64
22   Baggage Handling                         129880 non-null    int64
23   Satisfaction                             129880 non-null    object
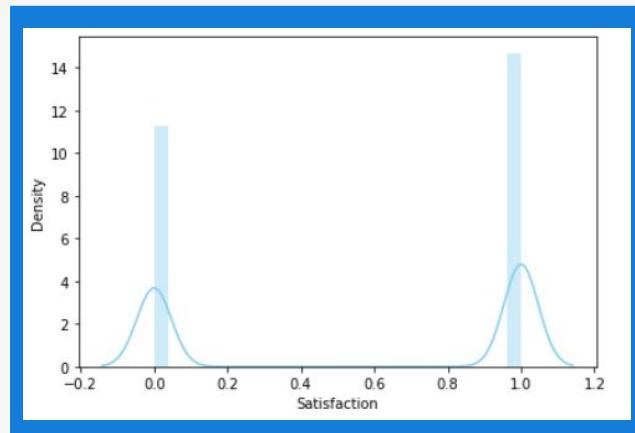```

## 0.3%
Missing Values

## 0%
Duplicated Data
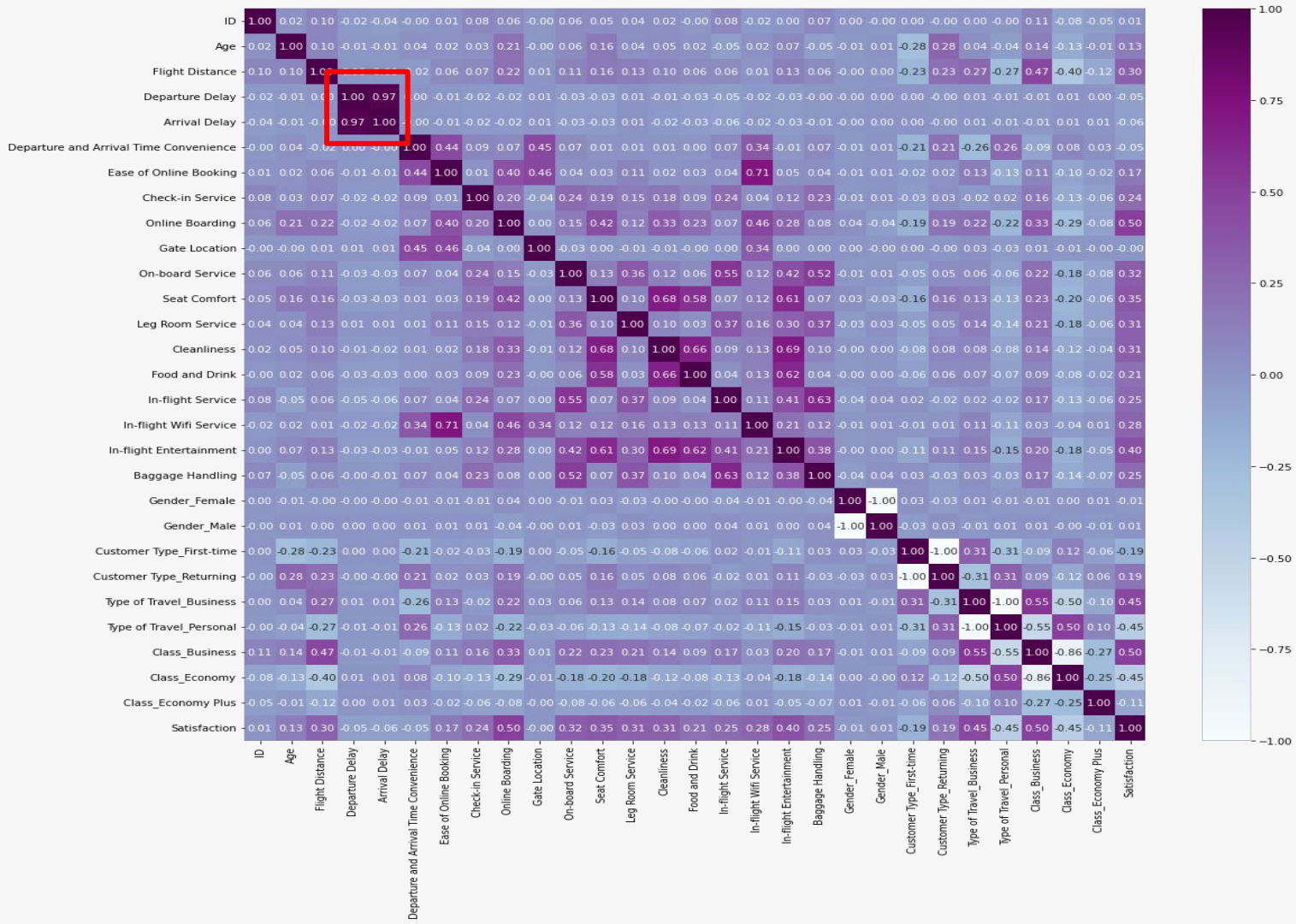
## 18
Numerical Columns

## 5
Categorical Columns

The dataset is slightly imbalance with 56% of neutral or dissatisfied passengers and 44% satisfied passengers

Data Analysis

Departure delay and arrival delay are highly correlated with each other. 0.97 shows strong positive correlation this means that as departure delays increase, there is a very strong tendency for arrival delays to also increase. In other words, flights that experience longer departure delays are likely to arrive later than those that experience shorter departure delays.

In any case, this strong correlation highlights the importance of managing departure delays in order to minimize the impact on arrival times and improve overall airline efficiency.

# Departure & Arrival Delay Time



Satisfaction by Departure Delay Time

Satisfaction by Arrival Delay Time

1. Based on the correlation heatmap, departure delay and arrival delay is highly correlated to each other so when departure delay increase, arrival delay will increase too. as can be seen on the chart, the total passengers are the same for each time in departure and arrival delay.
2. This airline is most likely to experience delays of less than 15 minutes.
3. There is a decrease in passenger satisfaction as the duration of delays increases.

# Age



Satisfaction by Age Category

Satisfaction by Adult Category
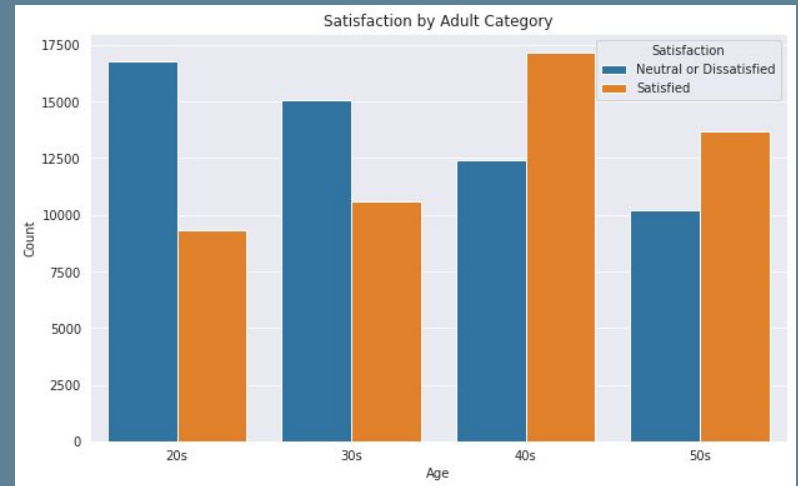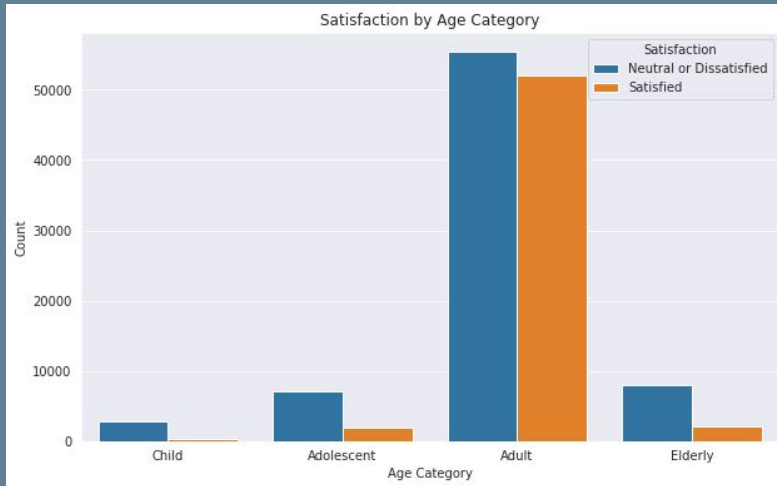
1. All age categories are mostly neutral or dissatisfied.
2. Most of the passenger are adult age 20-60. dominated by age 40s.
3. In adult passenger categories, the older the passengers, the total of neutral or dissatisfaction passengers decreases.
4. The most satisfied passengers are passengers in their 40s

# Flight Distance
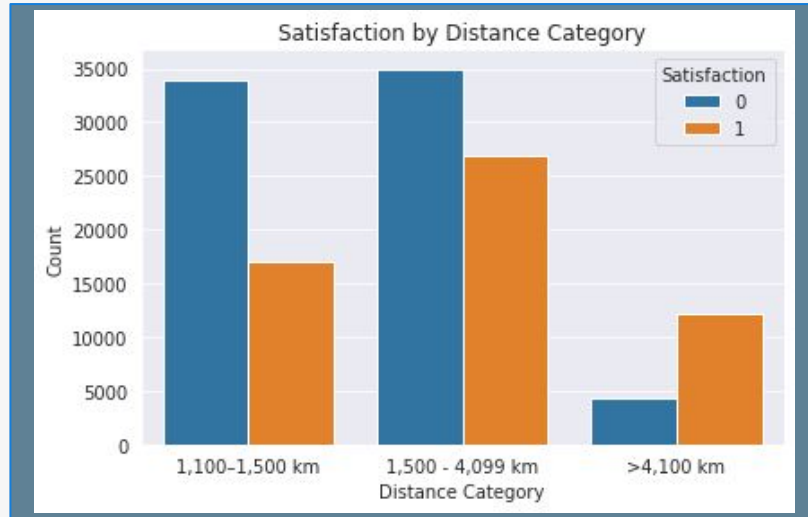


1. Flight distance categorized to short haul, medium haul. And long haul flight.
2. This airline passenger are mostly have medium haul flight (1,500 - 4100 KM)
3. Most of the passenger are neutral or dissatisfied but for the passenger who has long haul flight are satisfied with the airline service

# Gender, Customer and Travel Type



1. Both gender are mostly Neutral or Dissatisfied. But the correlation coefficient shows 0.01 and -0.01 for satisfaction. it shows that there is almost no relationship between gender and satisfaction. This means that gender is not a strong predictor of customer satisfaction in this dataset.

2. Returning customer dominated the satisfaction categories but most of them are neutral or dissatisfied. the correlation coefficient of first time customer -0.19 for satisfaction and 0.19 for satisfaction returning customer suggest that there is a weak relationship between customer loyalty and satisfaction. This means that customers who have flown with the airline before tend to be more satisfied with their experience than those who are flying with the airline for the first time.

3. Passenger with the needs of business are mostly satisfied. The correlation coefficient that customers who are traveling for personal reasons tend to be less satisfied with their experience compared to those who are traveling for business.

# Seat Class, Time Convenience and Online Booking
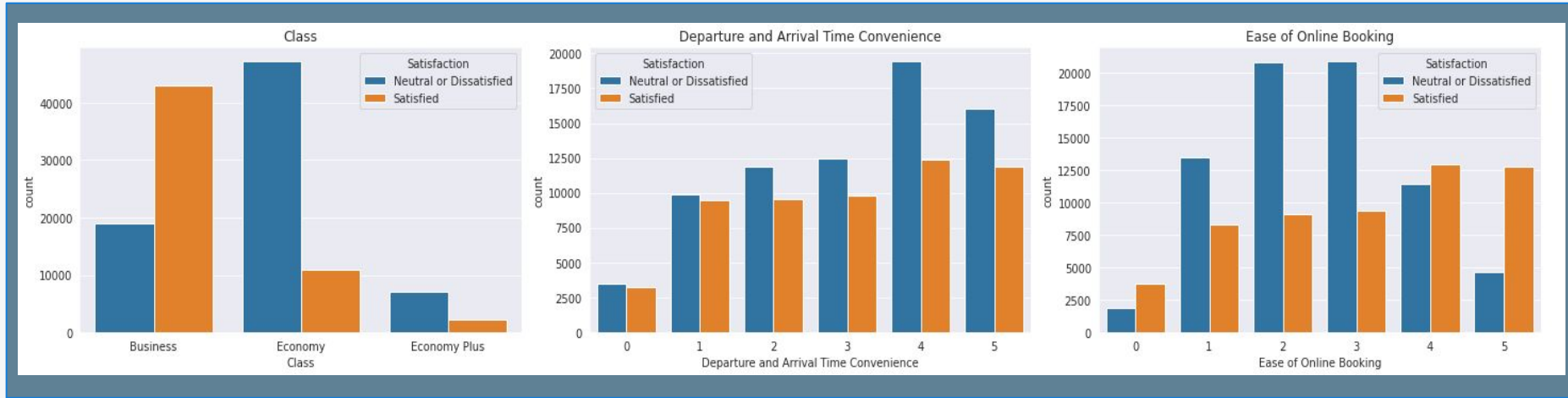


1. Passenger in business class are mostly satisfied compared to economy class. from the countplot and correlation shows that as the class of the flight moves from economy to higher classes, there is a tendency for satisfaction to increase. In other words, passengers who are flying in higher classes are generally more satisfied than those who are flying in economy class.
2. although the passenger are mostly satisfied with Departure and Arrival Time Convenience, but overall they are mostly neutral or dissatisfied. the correlation shows -0.05 which means that Departure and Arrival Time Convenience is not a strong predictor of customer satisfaction in this dataset.
3. As seen on the graph, passengers are mostly neutral or dissatisfied with the Ease of Online Booking. it means for most of them the online booking is still complicated or difficult. a correlation coefficient of 0.17 shows weak relationship between Ease of Online Booking and Satisfaction so it is not a strong predictor of customer satisfaction

Check-in Service — Online Boarding — Gate Location
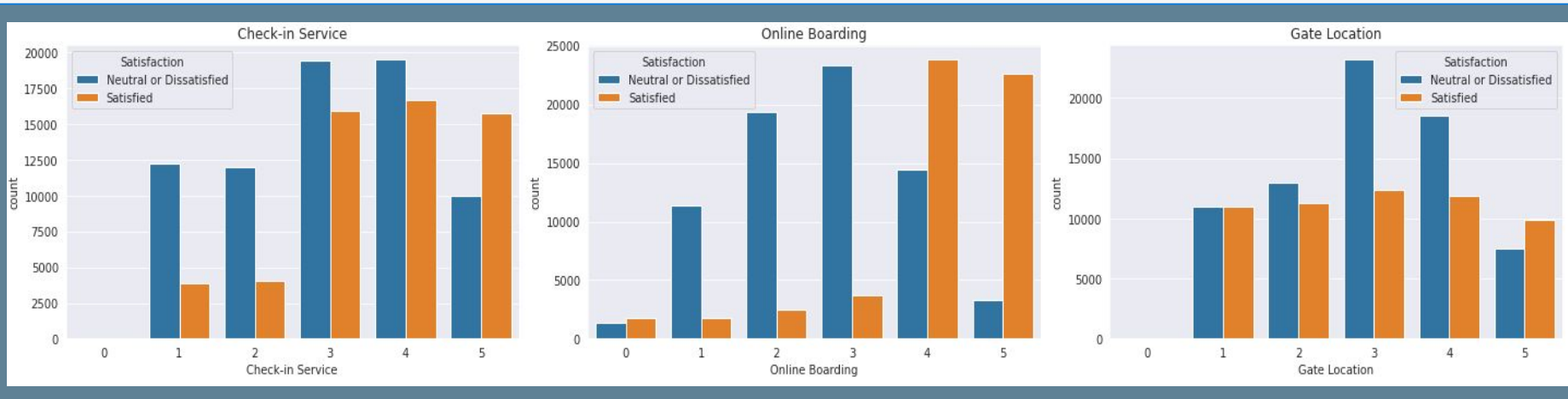
1. Passenger are mostly neutral to satisfied with the check-in service. a correlation coefficient of 0.24 suggests a weak positive relationship between the two variables. This means that as satisfaction with the check-in service increases, there is a slight tendency for overall passenger satisfaction to also increase.

2. Passengers are mostly satisfied with the online boarding, but passenger who are neutral or dissatisfied with the online boarding also feel neutral or dissatisfied for overall airline services. The correlation coefficient of 0.50 shows moderate positive relationship. This means that as satisfaction with the online booking process increases, there is a tendency for overall passenger satisfaction to also increase.

3. Passengers are mostly neutral to satisfied with the gate location of the airline. from the heatmap it shows no correlation at all with the overall satisfaction so the gate location is not a factor of passenger satisfaction.
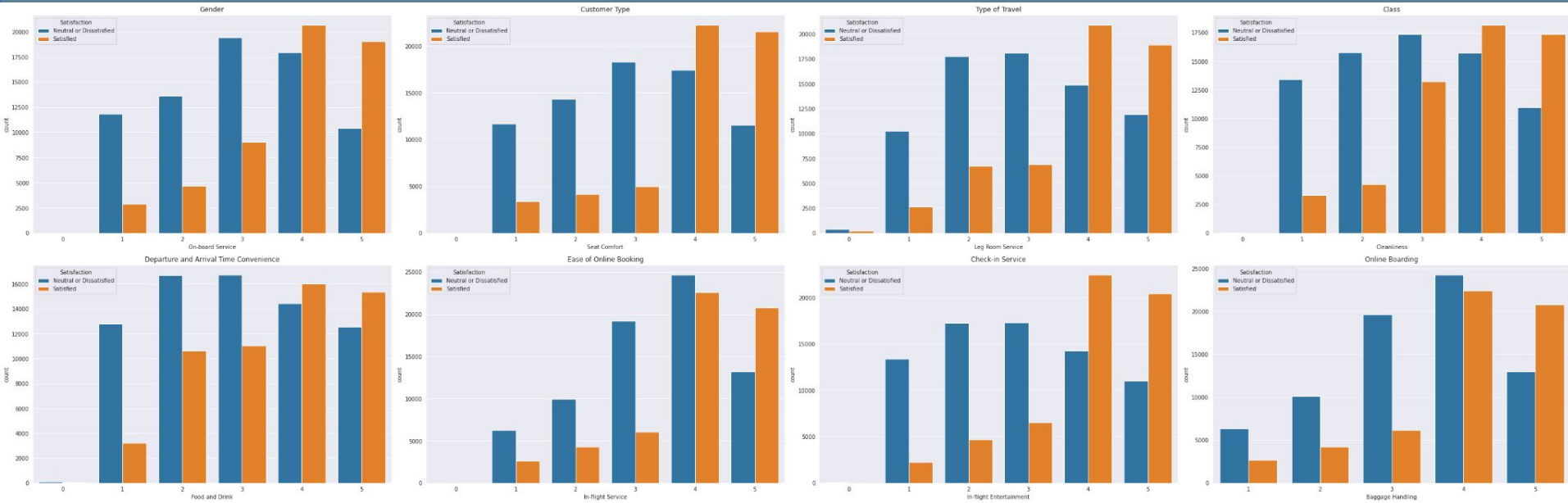
Passenger are satisfied with on-board service, Seat Comfort, Leg Room Service, Cleanliness, Food & Drink, In-Fligh Services, In-Flight Entertainment, and Baggage Handling. but for passenger who are dissatisfied with these services also most likely dissatisfied with overall airline service. the correlation heatmap shows positive correlation between satisfaction and these services, meaning that as these services increases, so does passenger satisfaction espesically for in-flight entertainment.

Modelling

# Data Preparation & Preprocessing
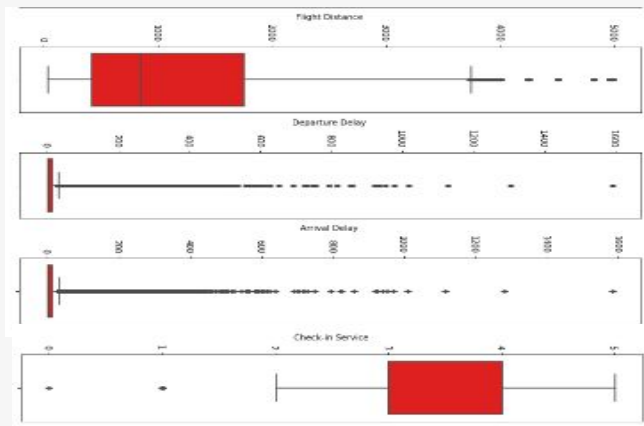
## 01

### One Hot Encoding

For all categorical columns except target columns

## 02

### Label Encoding

For target column *(Satisfaction)*

Data Preprocessing



There are several outliers detected on Flight Distance, Departure Delay, Arrival Delay, and Check in Service.

Outlier Detection

# 1st Model Evaluation

| Classifier | | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| KNeighborsClassifier | 0 | 0.731 | 0.674 | 0.702 | 11153.000 |
| | 1 | 0.767 | 0.812 | 0.789 | 14745.000 |
| | accuracy | 0.753 | 0.753 | 0.753 | 0.753 |
| | macro avg | 0.749 | 0.743 | 0.745 | 25898.000 |
| | weighted avg | 0.752 | 0.753 | 0.752 | 25898.000 |
| RandomForestClassifier | 0 | 0.971 | 0.941 | 0.956 | 11153.000 |
| | 1 | 0.957 | 0.978 | 0.967 | 14745.000 |
| | accuracy | 0.962 | 0.962 | 0.962 | 0.962 |
| | macro avg | 0.964 | 0.960 | 0.962 | 25898.000 |
| | weighted avg | 0.963 | 0.962 | 0.962 | 25898.000 |
| GradientBoostingClassifier | 0 | 0.943 | 0.921 | 0.932 | 11153.000 |
| | 1 | 0.941 | 0.958 | 0.949 | 14745.000 |
| | accuracy | 0.942 | 0.942 | 0.942 | 0.942 |
| | macro avg | 0.942 | 0.939 | 0.940 | 25898.000 |
| | weighted avg | 0.942 | 0.942 | 0.942 | 25898.000 |

KNN not perform really well in predicting the target variable and identifying the instances for this dataset.

# Model Evaluation after Data Normalization in KNN

| Classifier | | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| **KNeighborsClassifier** | **0** | 0.950 | 0.879 | 0.913 | 11153.000 |
| | **1** | 0.913 | 0.965 | 0.938 | 14745.000 |
| | accuracy | 0.928 | 0.928 | 0.928 | 0.928 |
| | macro avg | 0.931 | 0.922 | 0.925 | 25898.000 |
| | weighted avg | 0.929 | 0.928 | 0.927 | 25898.000 |
| **RandomForestClassifier** | **0** | 0.971 | 0.941 | 0.956 | 11153.000 |
| | **1** | 0.957 | 0.978 | 0.967 | 14745.000 |
| | accuracy | 0.962 | 0.962 | 0.962 | 0.962 |
| | macro avg | 0.964 | 0.960 | 0.962 | 25898.000 |
| | weighted avg | 0.963 | 0.962 | **0.962** | 25898.000 |
| **GradientBoostingClassifier** | **0** | 0.943 | 0.921 | 0.932 | 11153.000 |
| | **1** | 0.941 | 0.958 | 0.949 | 14745.000 |
| | accuracy | 0.942 | 0.942 | 0.942 | 0.942 |
| | macro avg | 0.942 | 0.939 | 0.940 | 25898.000 |
| | weighted avg | 0.942 | 0.942 | 0.942 | 25898.000 |

After adding standard scaler to KNN, the model performs really well in predicting the target variable and identifying the instances for this dataset. in terms of precision, recall, and F1-score, all models have higher performance metrics for class 1 (neutral or dissatisfied) than for class 0 (satisfied). Overall, Random Forest has the best performance for both classes based on the weighted average F1-score, with a value of **0.962.**

# Feature Importance



Online boarding, In-Flight Wifi Service, and Business Class are the most influential features that contribute the most to the prediction accuracy.
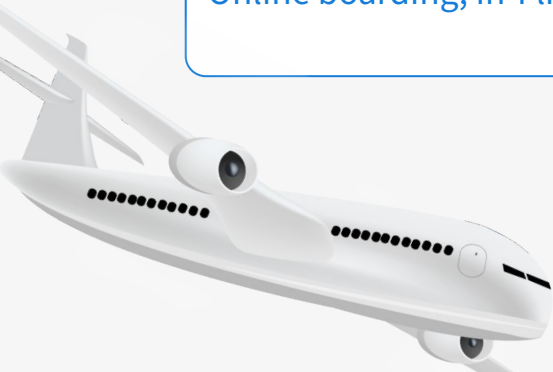
# Conclusion

The airline passengers satisfaction dataset is consist of 129,880 rows and 23 columns. The dataset is slightly imbalanced and has outliers in some features. To transform the non-numerical columns into numerical, I did one-hot encoding for the features and label encoding for the target variable. After encoding process, the dataset has 28 features. I split the dataset into 80% training data and 20% test data. I choose KNN, Random Forest, and Gradient boosting as the model to predict the airline passengers satisfaction since the dataset is a high-dimensional data, has outliers, and the target variable is binary (1: neutral or dissatisfied, 0: satisfied).

Before doing the data normalization, KNN performance is not relly good comparing to the other models. Then I add standard scaler to improve KNN performance with the end result Random Forest has the best performance for both classes based on the weighted average F1-score, with a value of 0.962, KNN and Gradient Boosting have slightly lower weighted average F1-scores of 0.927 and 0.942, respectively.

**From this research we can conclude that tree based models performance is better than non tree based models. The tree based models performs really well without any normalization. It proves that tree based model is robust with outlier. Meanwhile KNN needs data normalization to handle the outliers before achieving the best evaluation score. It is because the behaviour of KNN is sensitive with outlier and the scale of the input features. For this three models, the slightly imbalanced dataset is not affecting too much to their performance.**

# Business Recommendation

Most of the passengers are adults age 20-60. The most satisfied passengers are passengers in their 40s. Although the neutral or dissatisfied passengers is more than the satisfied passengers, The gap between both satisfaction in all age categories are not too big so the airline company can improve the service to retain existing passengers and reduce the number of neutral or dissatisfied passenger. The majority of the passengers type is returning passengers with business needs and most of them choose business class. People tend to choose this airline for medium to long haul flight.

from this insight company can focus on :

1. maintaining high standards of service in business class: Since most of the returning passengers choose business class, the airline company should focus on maintaining high standards of service in this class. This includes providing comfortable seats, high-quality food and drinks, and excellent customer service.
2. Offer loyalty programs and incentives for frequent business travelers: To retain their loyal customers, the airline company could offer loyalty programs and incentives such as discounts on future flights, complimentary upgrades, or exclusive access to airport lounges.
3. Enhance the in-flight entertainment options: Since many business travelers use their flight time for work or productivity, the airline company could consider enhancing the in-flight entertainment options to include more business-oriented content such as access to online productivity tools, business news, and webinars.
4. Improve the booking and check-in process: To make the travel experience more seamless and efficient for business travelers, the airline company could consider improving the booking and check-in process. This could include offering online check-in, mobile boarding passes, and expedited security lines for business class passengers.
5. Expand routes and frequencies to key business destinations: To attract more business travelers, the airline company could consider expanding its routes and frequencies to key business destinations. This could help the airline company capture a larger share of the corporate travel market and increase its revenue.

# Thank You!

Do you have any questions?

Sheilaftria@gmail.com